Introduction to Data Science and Analytics

Amar Sahay, Ph.D.

A Data Driven Decision-Making Approach for Business and Engineering

Data Science

- *Data Science is about making decisions using enormous amounts of data in business, engineering, research, finance, and health science.*
- *Data Science is a* ***data driven decision making approach.*** *It* ***uses several different areas, and disciplines including:***

  Statistics, Data Analysis, Mathematics, Information Technology Programming and Computer Science, and others.

- ***Its purpose is to extracts insights and knowledge from structured and unstructured data to analyze big data to draw conclusions, make predictions to drive business.***
- ***It is data-driven approach to decision making***

Need For Data Science

- Companies now collect massive amounts of data from exabytes to zettabytes which are both structured and unstructured.
- The advancement in technology and the computing capabilities have made it possible to store, process, and analyze this huge data (***Big Data***) with smarter storage spaces.
- The stored data in company's data bases must be processed and analyzed to make sense.
- **Data Science is a dynamic approach to help business and research and other fields to process, analyze, and make decisions**

What does Data Science Do?

- **At the core of Data Science is Data.** Most businesses, research and other areas rely and use data to get insights and make decisions and draw conclusions. Data Science is about making sense from huge amounts of data.
- **Data Science uses data for analysis, modeling, drawing inferences and making predictions for the future outcomes. The models and analysis help businesses make effective and timely decisions.**

Purpose

- This paper looks at the current state of the field of Data Science and its applications.
- Given the booming interest in Data Science and analytics, this research is timely and informative.
- The paper brings many terms, tools and methods used in Data Science together in a meaningful way.
- It is common for practitioners and even scholars to conflate terms such as data sciece, business intelligence, data analytics and data mining. This paper clarifies such terms and helps differentiate their meanings and provides the details on the tools a Data Scientist or analyst uses.

Overview

This paper explores the key topics in Data Science & Analytics and provides an overview of:

(1) Data Science – scope, its evolution, and its relationship with other disciplines,

(2) An overview of the field of Business Intelligence (BI) (**investigates historical data to better understand business performance; thereby, improving performance, and creating new strategic opportunities for growth**)

Overview…cont.

(3) Business analytics (BA) and the three major categories of business analytics – the descriptive, predictive, and prescriptive analytics along with advanced analytics tools

(4) Overview of most widely used Predictive Analytics models including **regression, classification, forecasting, data** mining and machine learning based models and applications,

Overview…cont.

(5) Business Analytics, Business Intelligence (BI) and their relation to Data Science,

(6) Data Science and Software Tools in Data Science and Analytics,

(7) Data Visualization in Data Science

(8) Statistical Tools in Data Science – Inferential and Probability Concepts

(9) Overview of Machine Learning and R-statistical Software in Data Science.

Need For Data Science

- Companies now collect massive amounts of data from exabytes to zettabytes which are both structured and unstructured. [1 zetaabyte= 1 trillion Gigabytes]

- The advancement in technology and the computing capabilities have made it possible to store, process, and analyze this huge data with smarter storage spaces.

Data Science Defined

Data Science is a ***data driven decision-making approach*** *that* ***uses several different areas, methods, algorithms, models, and disciplines with a purpose of extracting insights and knowledge from structured and unstructured data.*** *These insights are helpful in applying algorithms and models to make decisions. The models in Data Science are used in predictive analytics and machine learning to solve many different types of problems and predict future outcomes.*

Another Look at Data Science

- Data science can be viewed as a ***multidisciplinary field*** focused on finding actionable insights from large sets of raw, structured, and unstructured data.
- It uses several different **areas** including **statistical analysis**, **programming and computer science**, **predictive analytics**,

statistical and mathematical modeling, and **machine learning** to use massive datasets in an effort to find solutions to problems that haven't been thought of yet.

Data Science…cont.

- Data science techniques have the ability to understand, process, and visualize data in the initial stages.
- Utilize statistics, modelling, mathematics, and technology, machine learning techniques to address and solve analytically complex problems using structured and unstructured data.
- **At the core of data science is data**. It is about using this data in creative and effective ways to help businesses in making data-driven business decisions.
  https://en.wikipedia.org/wiki/Data_science

Structured and Unstructured Data

- Data science is applied to extract information from both structured and unstructured data.[1][2]
- **Unstructured data** is usually not organized in a structured manner and may contain qualitative or categorical elements, such as dates, categories, etc. and are text heavy
- Compared to structured data, the ***unstructured data contain irregularities***. The ambiguities in unstructured data make it difficult to apply traditional tools of statistics and data analysis.

Structured and Unstructured Data

Structured data are usually stored in clearly defined fields in databases. The software applications and programs are designed to process such data.

In recent years, a number of newly developed tools and software programs have emerged that are capable of analyzing big and unstructured data. One of the earliest applications of unstructured data is in analyzing text data using text -mining and other methods.

Some Facts and Predictions About Data

In 1998, Merrill Lynch said "unstructured data comprises the vast majority of data found in an organization, some estimates run as high as 80%."[1]

Here are some other predictions: As of 2012[update], IDC and Dell EMC project that data will grow to **40 zettabytes by 2020**, resulting in a 50-fold growth from the beginning of 2010.[6] More recently, IDC and Seagate predict that the global datasphere will grow to **163 zettabytes by 2025** [7] and majority of that will be unstructured. The Computer World magazine states that unstructured information might account for more than 70%–80% of all data in in organizations.[1.]. [https://en.wikipedia.org/wiki/Unstructured_data]

Specific Areas of Data Science and Applications

Data Science draws from different areas. Some are:

- visualization techniques,
- statistical modeling and analysis,
- statistical programming language, such as R programming,
- a knowledge of Data Bases (SQL or MySQL) or other data base management system.
- One major application of Data Science is in the area of Machine Learning (ML) and Artificial Intelligence. It uses Programming languages and statistical models (Python is a popular program for solving Machine Learning Problem)

Data science and its relation to other areas

We will provide a detailed overview of Data Science by defining, outlining and reviewing the tools and techniques.

Explain the differences and similarities between **Data Science and Data Analytics.**

Relationship of Data Science to **Analytics, Business Analytics, and Business Intelligence (BI)**

The figures in the following slides show the field of data science and associated areas.

Figure 1. Broad View of Data Science with associated areas

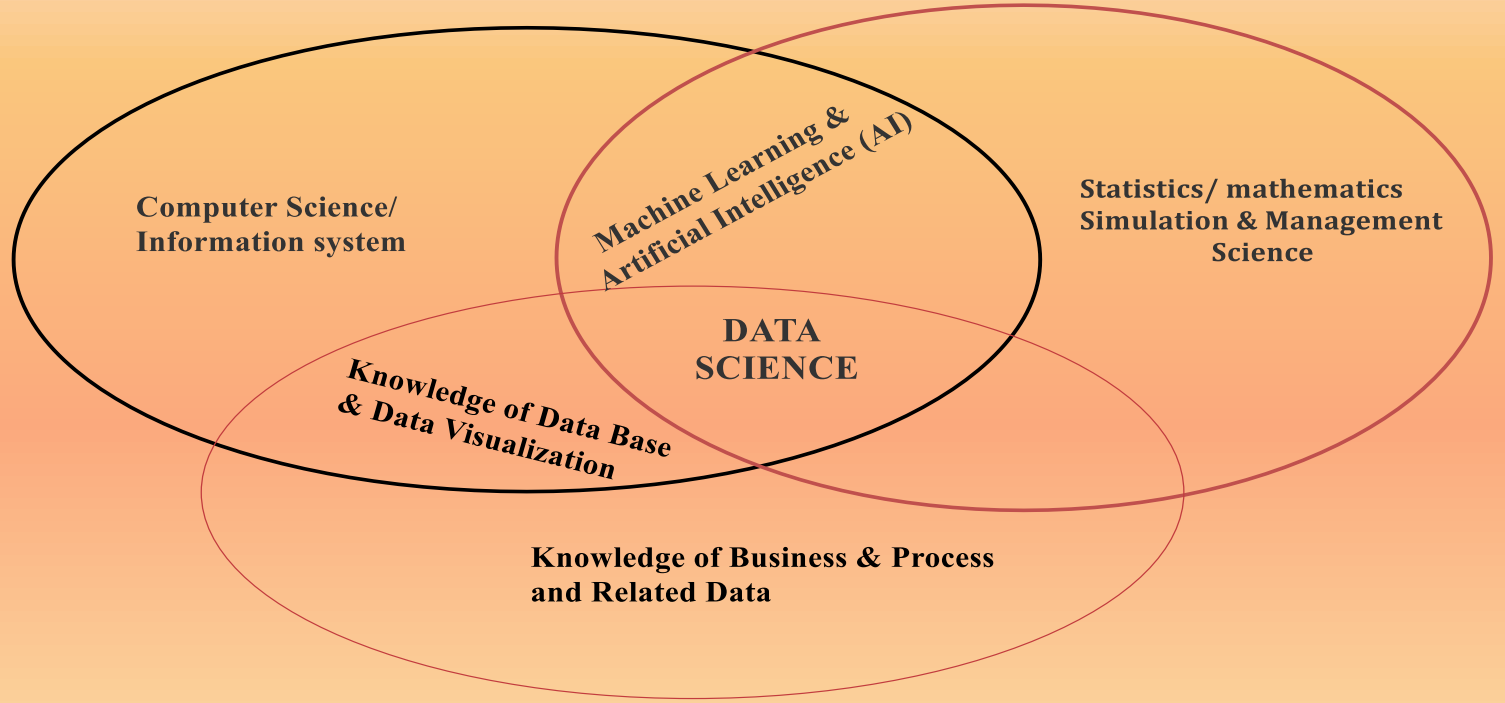
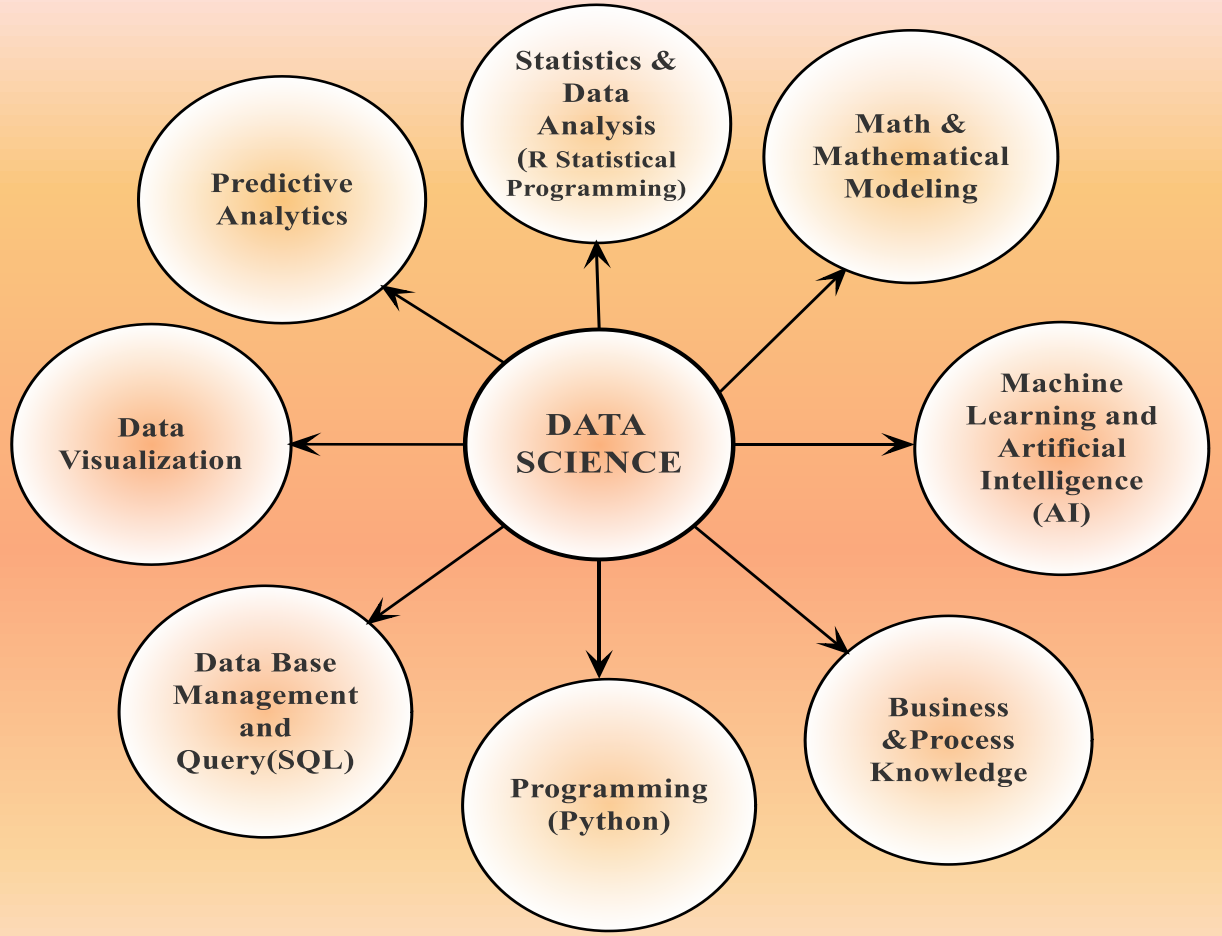


Figure 2. Data Science Body of Knowledge



Data Science and Data Analytics

[https://www.sisense.com/blog/data-science-vs-data-analytics/]

Data analytics focuses on processing and performing statistical analysis on existing datasets or data bases of companies using different tools and methods to capture, process, and organize, and perform data analysis to data to uncover actionable insights from data and find ways to present this data. More simply, the field of data and analytics is directed toward solving problems for questions we know but we don't know the answers to. More importantly, it's based on producing results that can lead to immediate improvements.

Data analytics also encompasses a few different branches of broader statistics

Difference between Data Science and Data Analytics

- The terms **data science** and *data analytics* are used ***interchangeably***, but data science and big data analytics are unique fields, with the major difference being the scope.
- **Data science is an umbrella term for a group of fields that are used to mine large datasets.**
- ***Data Science has much broader scope compared to data analytics, analytics, and business analytics.***
- **Data analytics is a more focused version of data science** and focuses more on data analysis and statistics and can even be considered part of the larger process that uses simple to advanced statistical tools.

Difference between Data Science & Data Analytics…Cont.

- Data Science uses Big Data that are useful for solving problems.
- It uses modeling, **machine learning**, and used to enhance **AI algorithms** as it can improve how information is sorted and understood.
- ***Machine learning (ML) is widely used to model, train, and solve a number of problems (supervised and unsupervised).***

Statistics in Data Science

Data Science professionals and Data Scientists should have a strong background in statistics, mathematics, and computer applications. Good analytical and statistical skills are a pre-requisite to successful application and implementation of data science projects. Some key statistical concepts that every data scientist should know:

- Descriptive Statistics and Data Visualization
- Inferential Statistics Concepts and tools of inferential statistics

  (Estimation theory, Regression from simple to logistics regression, classification, clustering, and more)
- Concepts of probability and probability distributions
- Concepts of Sampling and Sampling Distribution/ Over and Under-Sampling
- Bayesian Statistics/ Dimensionality Reduction and more

Career Path for Data Science Professional and Data Scientist

- A data scientist requires knowledge/ expertise from varied fields.
- The field of data science provides a unifying approach by combining varied areas ranging from statistics, mathematics, analytics, business intelligence, computer science, programming and information systems.
- It is rare to find a data science professional with knowledge and background in all these areas. It is often the case that a data scientist has specialization in a sub-field.
- The minimum education requirement for a data science professional is a bachelors' degree in mathematics, statistics, IT, computer science, or business with analytical background . A master' or PhD is preferred.

Figure 1: Broad View of Data Science with associated areas

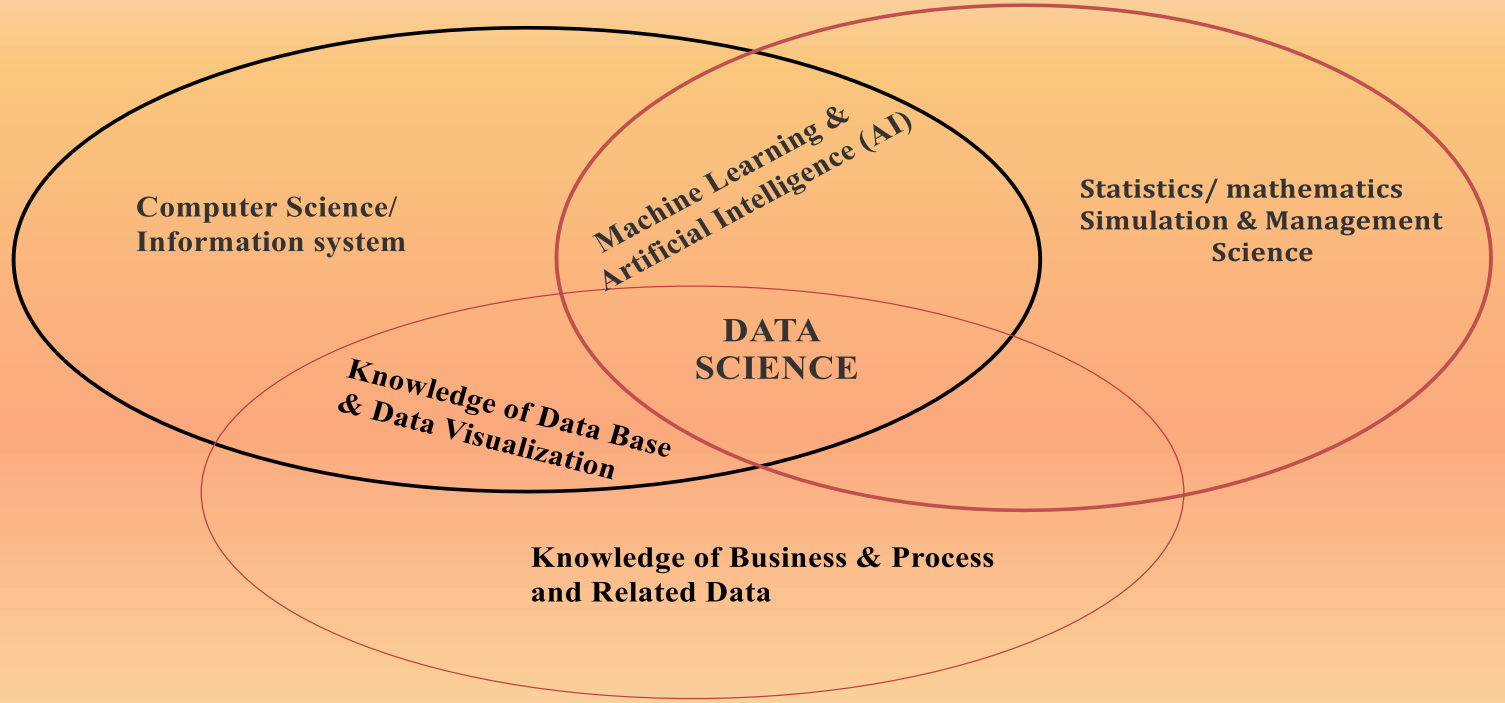
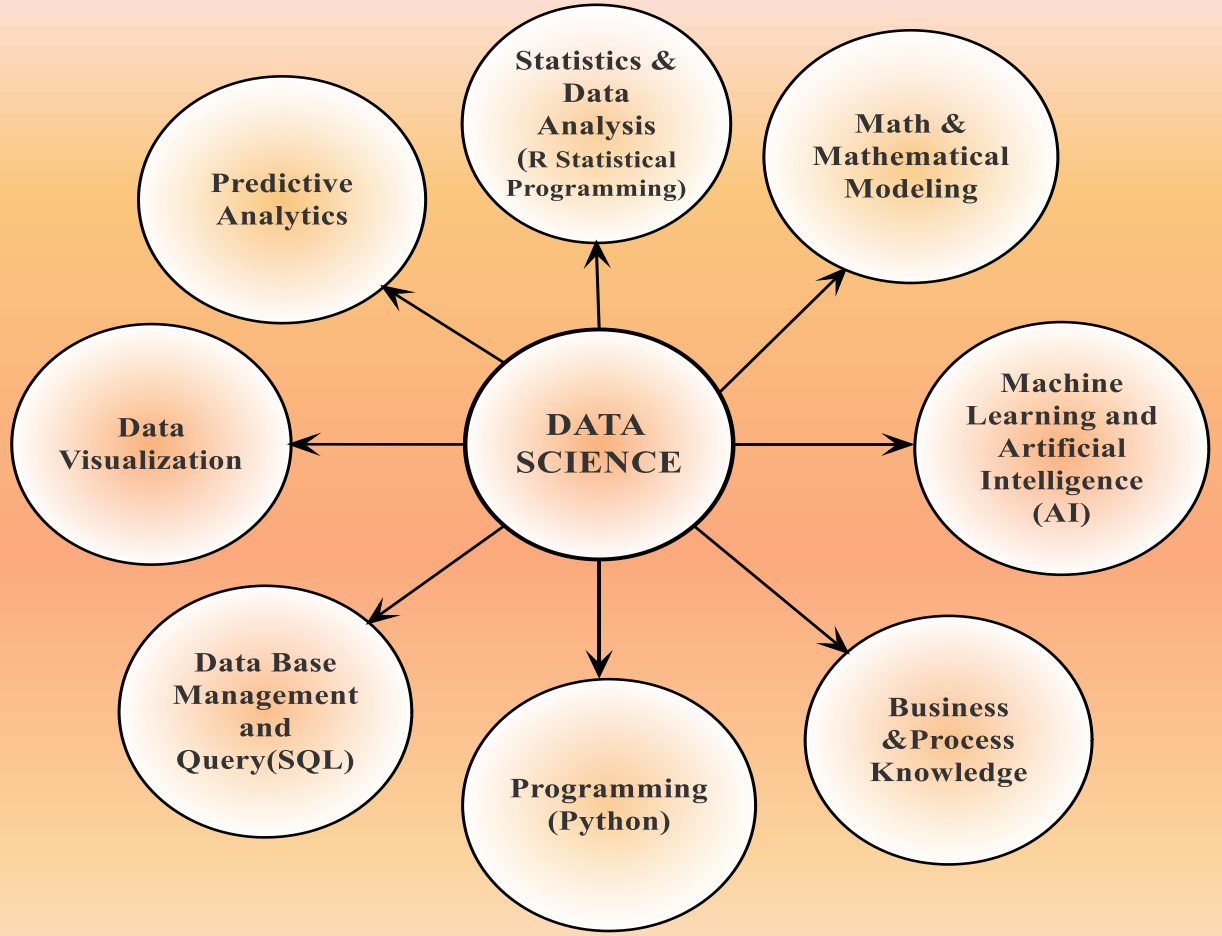


Figure 2: Data Science Body of Knowledge



Data Science Career

- *Data science continues to evolve as one of the most sought-after areas by companies.* A career in data science is ranked at the third best job in America for 2020 by Glassdoor, and was ranked the number one best job from 2016-2019.[29]
- Data scientists have a median salary of $118,370 per year or $56.91 per hour.[30] Job growth in this field is also above average, with a projected increase of 16% from 2018 to 2028.[30]
- The largest employer of data scientists in the US is the federal government, employing 28% of the data science workforce.[30] Other large employers of data scientists are computer system design services, research and development laboratories, and colleges and universities.[30] The future outlook for data science field looks promising. *It is estimated that 1.5 to 2.0 million jobs will be created in this area in the next ten years*.

Data Science, Analytics and Business Analytics (BA)

- The field of data science is vast, and it requires the knowledge and expertise from diverse fields ranging from statistics, mathematics, data analysis, machine learning/artificial intelligence as well as computer programming and data base management skills.
- *One of the major areas of data science is analytics and business analytics.* These terms are often used interchangeably with data science.
- There is a clear distinction between data science and analytics. we discuss the area of analytics and business analytics.

Business Analytics- What is it?

- Here we discuss the broad meaning of the terms – **analytics, business analytics, different types of analytics, the tools of analytics** and how they are used in business decision making.
- The companies now use massive amount of data referred to as *big data The tools of analytics are used to analyze big data*
- The other area is Data Mining. We briefly discuss data mining and the techniques used in data mining to extract useful information from huge amounts of data.

Business Analytics

The essence of analytics lies in the application – making sense from data.

Business analytics is a decision-making approach that uses data visualization techniques, statistical and quantitative analysis, information technology, management science (mathematical modeling, simulation), along with data mining and fact-based data to measure past business performance to guide an organization in business planning and effective decision making.

Need for Business Analytics

- Big data analysis is now becoming an integral part of business analytics. The organizations use business analytics as an organizational commitment to data-driven decision making. Business Analytics helps businesses in making informed business decisions and also in automating and optimizing business processes.
- The analyses are needed to explore, investigate, draw conclusions, and predict and optimize business outcomes.

How are Analytics used?

- **Business Analytics (BA) combines advanced statistical analysis and predictive modeling to give us an idea of what to expect so that one can anticipate developments or make changes now to improve outcomes.**

- Business analytics is more about anticipated future trends of the key performance indicators. This is about using the past data, models to learn from the existing data (descriptive analytics), make predictions (predictive analytics), and optimize business processes (prescriptive analytics).

- Analytics models use the data with a view to drawing out new, useful insights to improve business planning and boost future performance.

Another Sub-field of Analytics_ Data Mining

- In business, Data Mining is used to analyze huge amount of business data. Business transaction data along with other customer and product related data are continuously stored in the databases.
- The data mining software are used to analyze the vast amount of customer data to reveal hidden patterns, trends, and other customer behavior. Data Mining uses the process of discovering knowledge from the data (KDD).

Data Mining

Businesses use data mining to perform market analysis to identify and develop new products, analyze their supply chain,

find the root cause of manufacturing problems, study the customer behavior for product promotion, improve sales by understanding the needs and requirements of their customer,

prevent customer attrition and acquire new customers.

Data Mining Application

- **Wal-Mart** collects and processes over 20 million point-of-sale transactions every day. Data mining software are used to understand and determine customer behavior, needs and requirements. The data are analyzed to determine sales trends and forecasts, develop marketing strategies, and predict customer-buying habits [http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/].
- Data and and information about products, companies and individuals are available through **Google**, **Facebook**, **Amazon**, and several other sources. Data mining and analytics tools are used to extract meaningful information and pattern to learn customer behavior. **Financial institutions analyze data of millions of customers to assess risk and customer behavior. Data mining techniques are also used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering.**

Type of Business Analytics

Business Analytics has three broad categories

1. Descriptive analytics,
2. Predictive analytics, and
3. Prescriptive analytics.

Each type of analytics uses a number of tools that may overlap depending on the applications and problems being solved.

The **descriptive analytics** tools are used to visualize and explore the patterns and trends in the data.

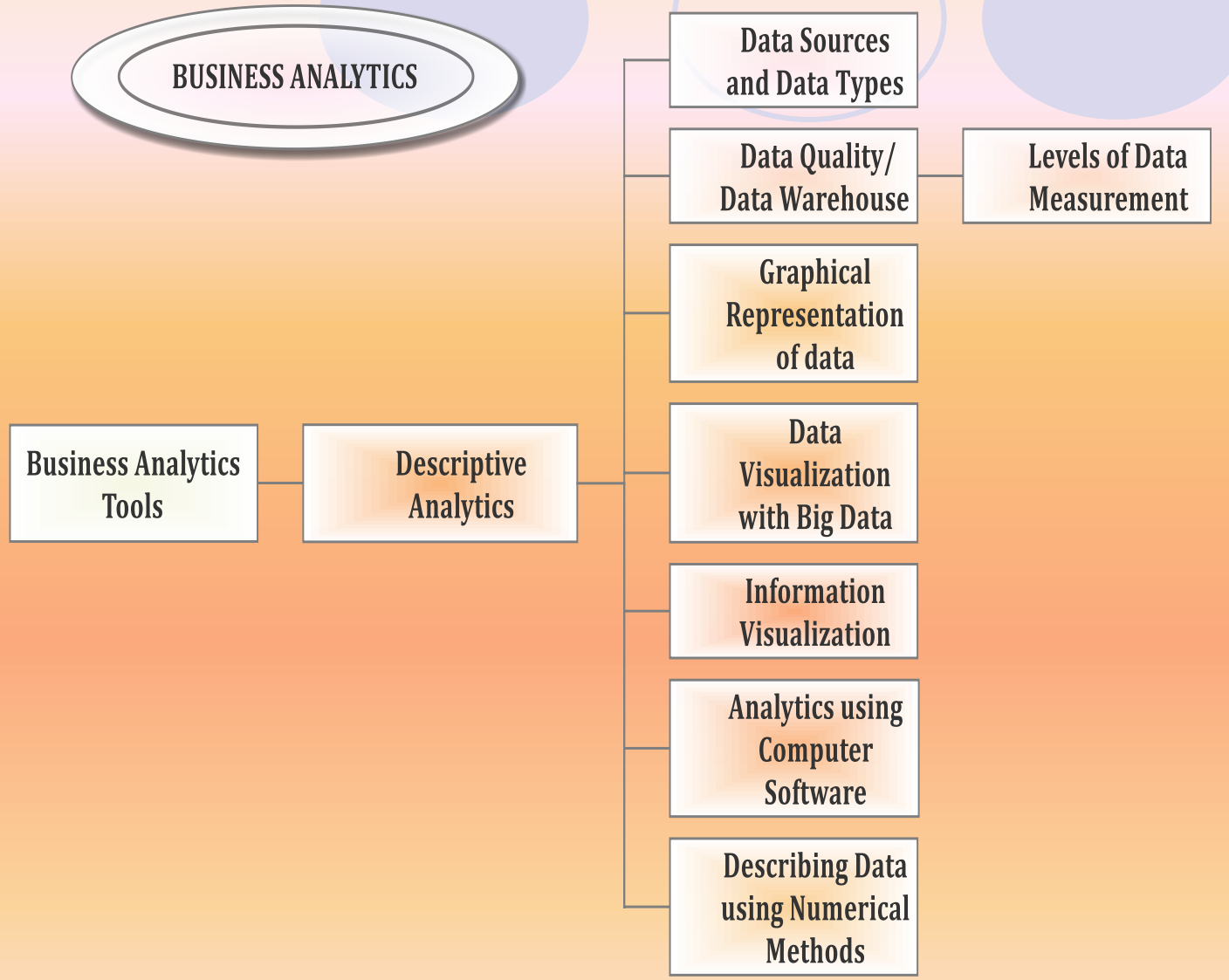
Predictive analytics uses the information from descriptive analytics to model and predict future business outcomes with the help of regression (different types), classification and clustering, forecasting and predictive modeling.

Descriptive Analytics: Data Visualization, Numerical Methods and Tools

Use of descriptive statistics, Data Visualization techniques, and Numerical analysis.

Many of the hidden patterns and features not apparent through mere examination of data can be exposed through graphical and numerical analysis. Descriptive analytics uses simple tools to uncover many of the problems quickly and easily. The results enable us question many of the outcomes so that corrective actions can be taken.

Successful use and implementation of descriptive analytics requires the understanding of types of data (structured vs. unstructured data), graphical/visual representation of data, and graphical techniques using specialized computer software capable of handling *big data*. Big data analysis is integral part of business analytics.



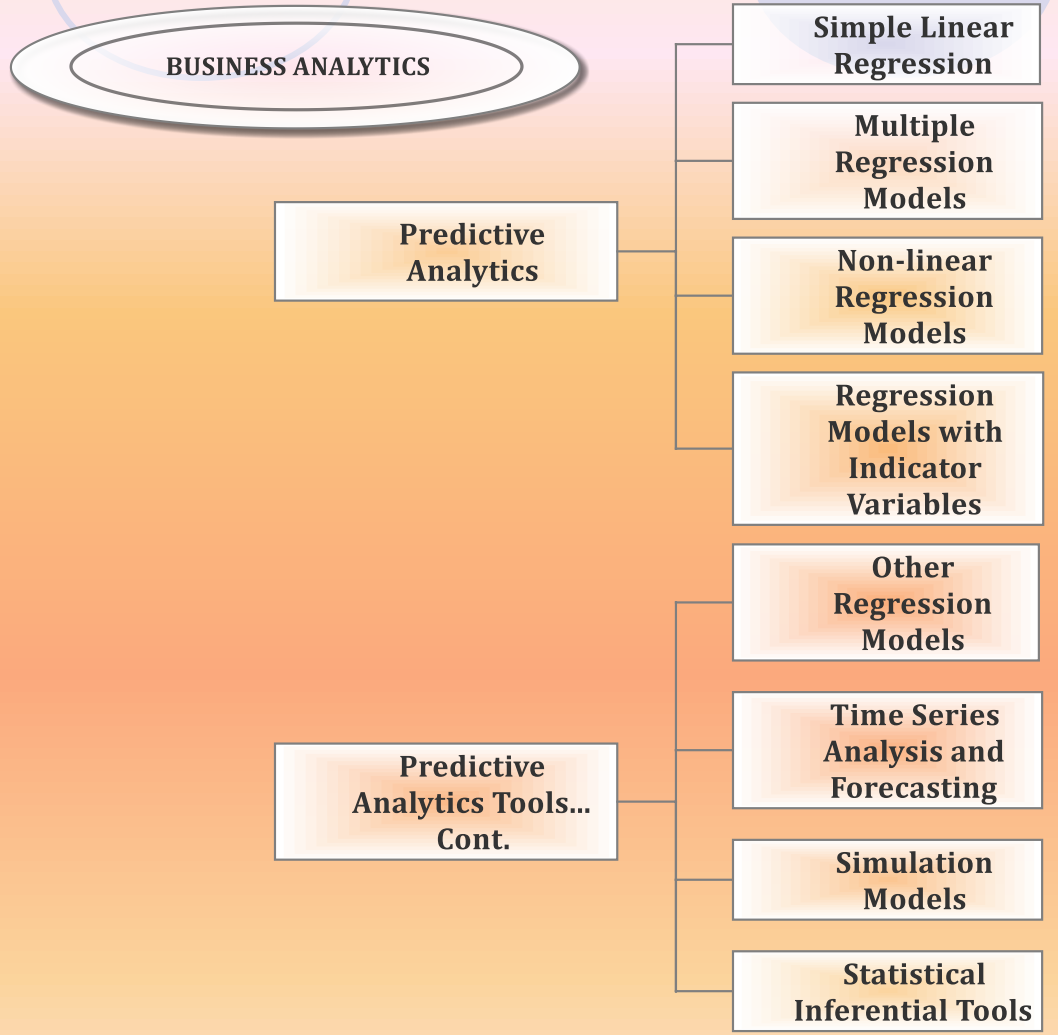
Descriptive analytics…cont.

- Recently, interconnections of the devices in **IOT** (Internet of Things) generate huge amounts of data providing opportunities for big data applications.
- The tools of descriptive analytics are helpful in understanding the data, identifying the trend or patterns in the data, and making sense from the data contained in the databases of companies. The understanding of databases, data warehouse, web search and query, and Big Data concepts are important in extracting and applying descriptive analytics tools. A number of statistical software are used for statistical analysis. Widely used software are **SAS, MINITAB**, and **R**- programming language for statistical computing, **Tableau** – a big data visualization software..

Predictive Analytics

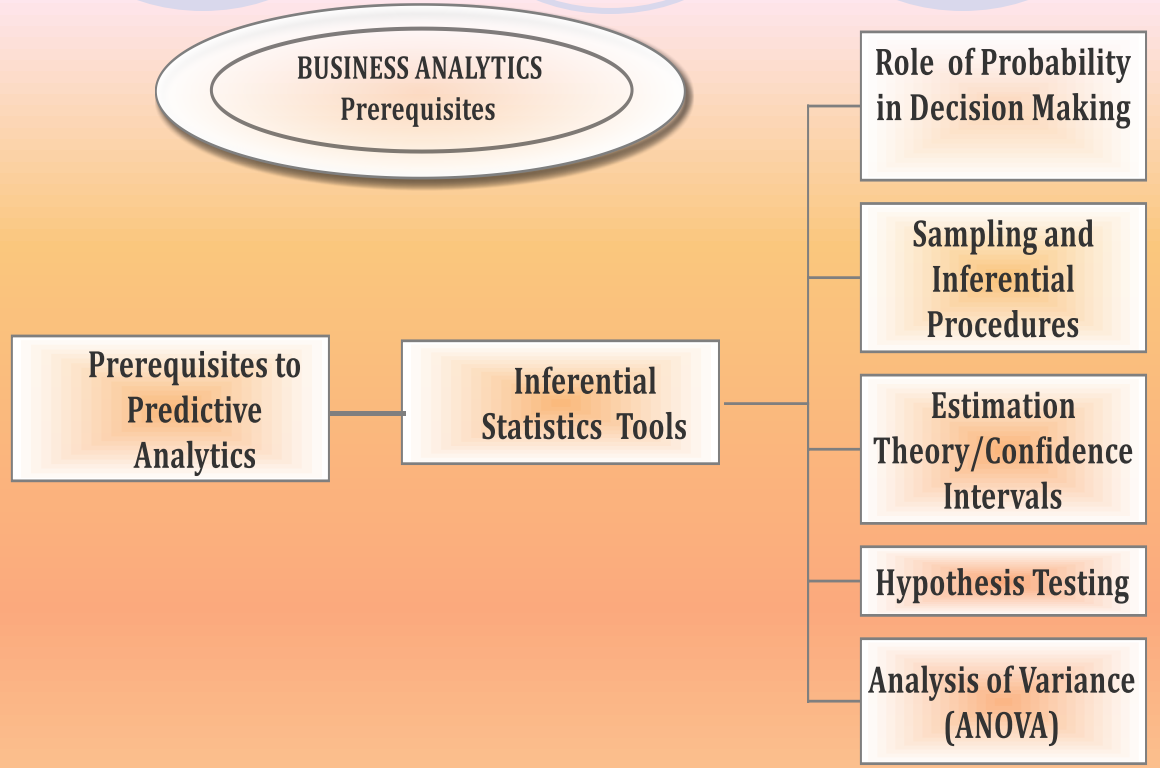
- **Predictive Analytics** is the application of predictive models to predict future business outcomes and trends
- ***Data mining*** techniques are used to extract useful information from huge amounts of data using predictive analytics, computer algorithms, software, mathematical, and statistical tools.
- ***Regression models*** are used for predicting the future outcomes. Variations of regression models include: (a) Simple regression models, (b) Multiple regression models, (c) Non-linear regression models including the quadratic or second-order models, and polynomial regression models (d) Regression models with indicator or qualitative independent variables, and (e) Regression models with interaction terms or interaction models, (f) Logistic Regression.
- ***Time Series Analysis and Forecasting models-*** *various types.*

Figure 2.2: Tools of Predictive Analytics



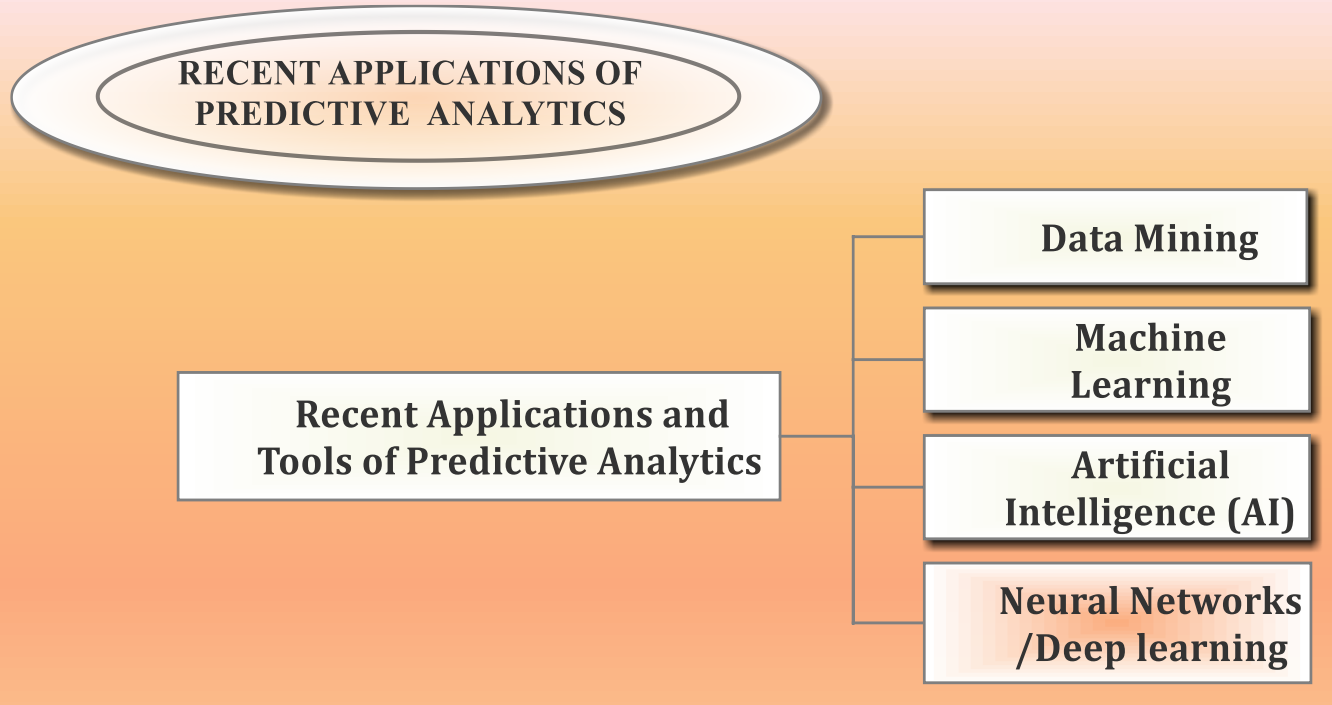
Background and Pre-requisites to Predictive Analytics Tools

- Probability theory, probability distributions and their role in decision making, (b) Sampling and inference Procedures, *(c)* Estimation and Confidence Intervals, (d) Hypothesis testing/Inference procedures for one and two population parameters, and (e) Analysis of Variance (ANOVA) and Experimental Designs. The understanding of these tools is critical in understanding and applying predictive analytics.



Recent applications and Tools of predictive analytics

Extensive applications have emerged in recent years using methods (in flow disgram) which are hot topics of research. A number of applications in business, engineering, manufacturing, medicine, signal processing and computer engineering using machine learning, neural networks, and deep learning are being reported



Prescriptive Analytics Tools

Prescriptive analytics is concerned with optimal allocation of resources in an organization. A number of operations research and management science tools have been applied for allocating the limited resources in the most effective way. The operations management tools that are derived from Management Science and Industrial Engineering including the simulation have also been used to study different types of manufacturing and service organizations.

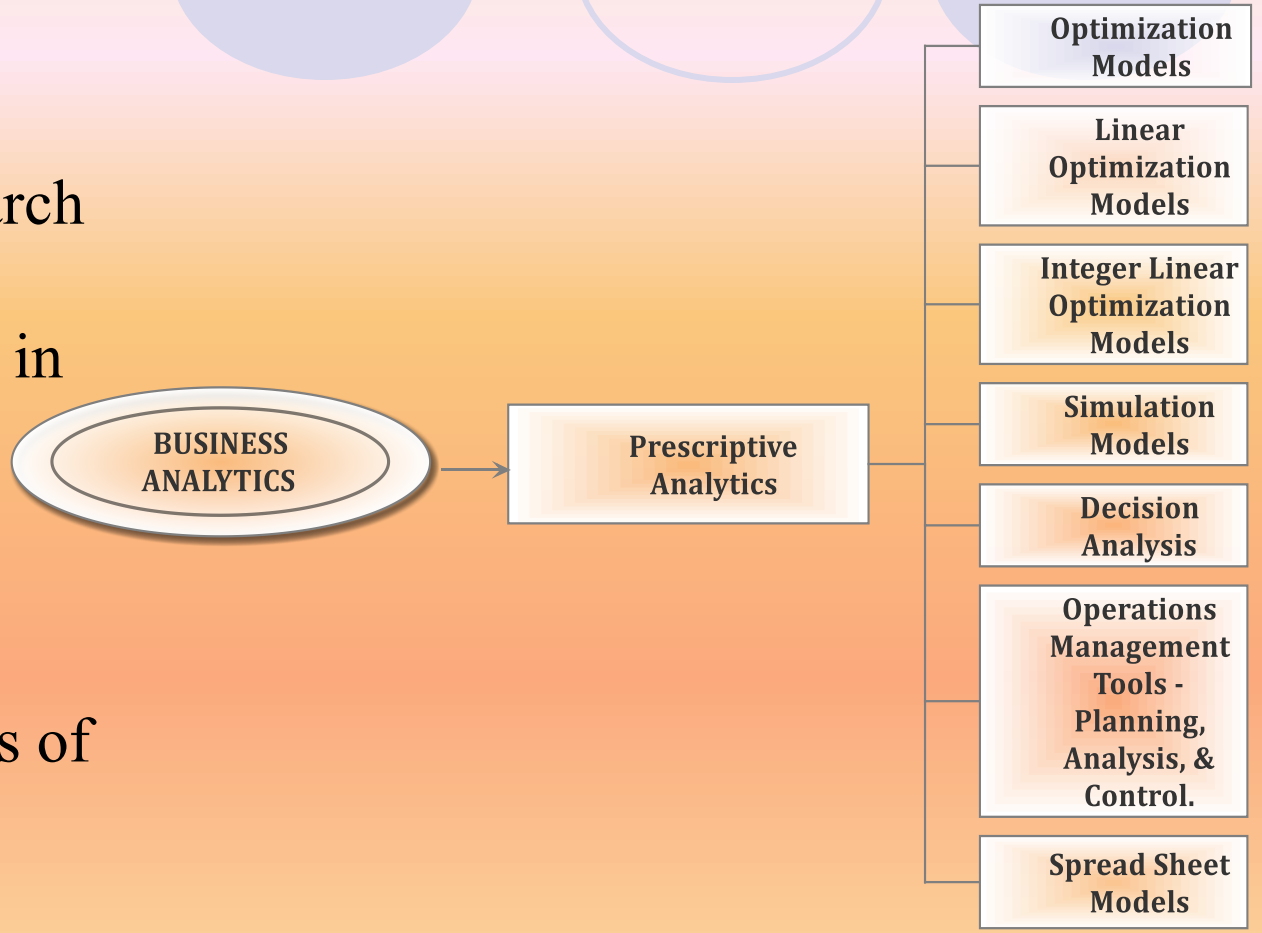
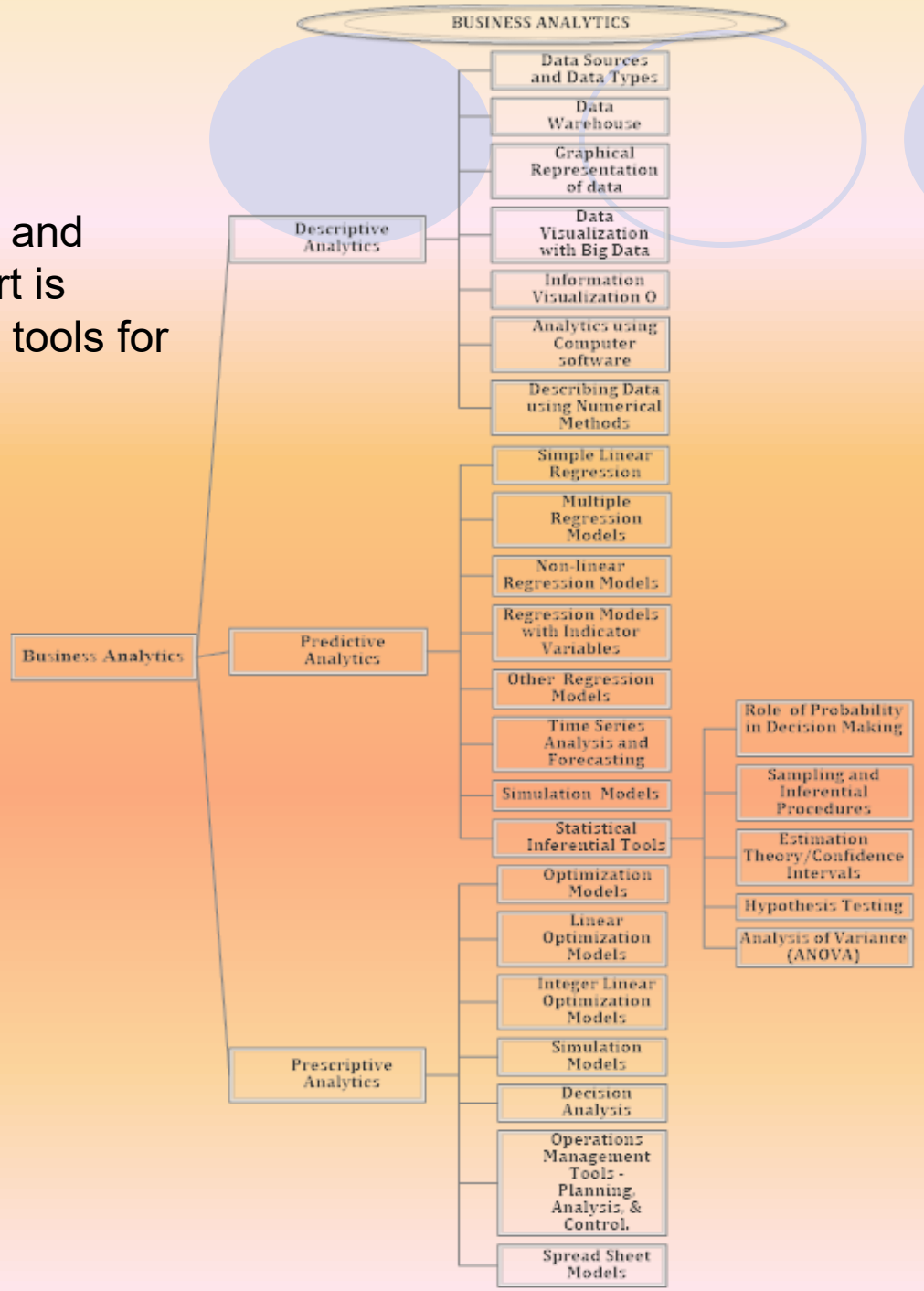


Figure 2.6: Descriptive, Predictive, and Prescriptive Analytics Tools

Figure outlines the tools of Descriptive, Predictive, and Prescriptive Analytics tools together. This flow chart is helpful in outlining the difference and details of the tools for each type of analytics.



Objective of each of the analytics

Type of Analytics	*Objectives*
Descriptive	Use graphical and numerical methods to describe the data. The tools of descriptive analytics are helpful in understanding the data, identifying the trend or pattern in the data, and making sense from the data contained in the databases of companies
Predictive	Predictive analytics is the application of predictive models that are used to predict future trends.
Prescriptive	Prescriptive analytics is concerned with optimal allocation of resources in an organization using a number of operations research, management science, and simulation tools.

Figure 3.1: Input to the Business Analytics process, types of analytics, and description of tools in each type of analytics

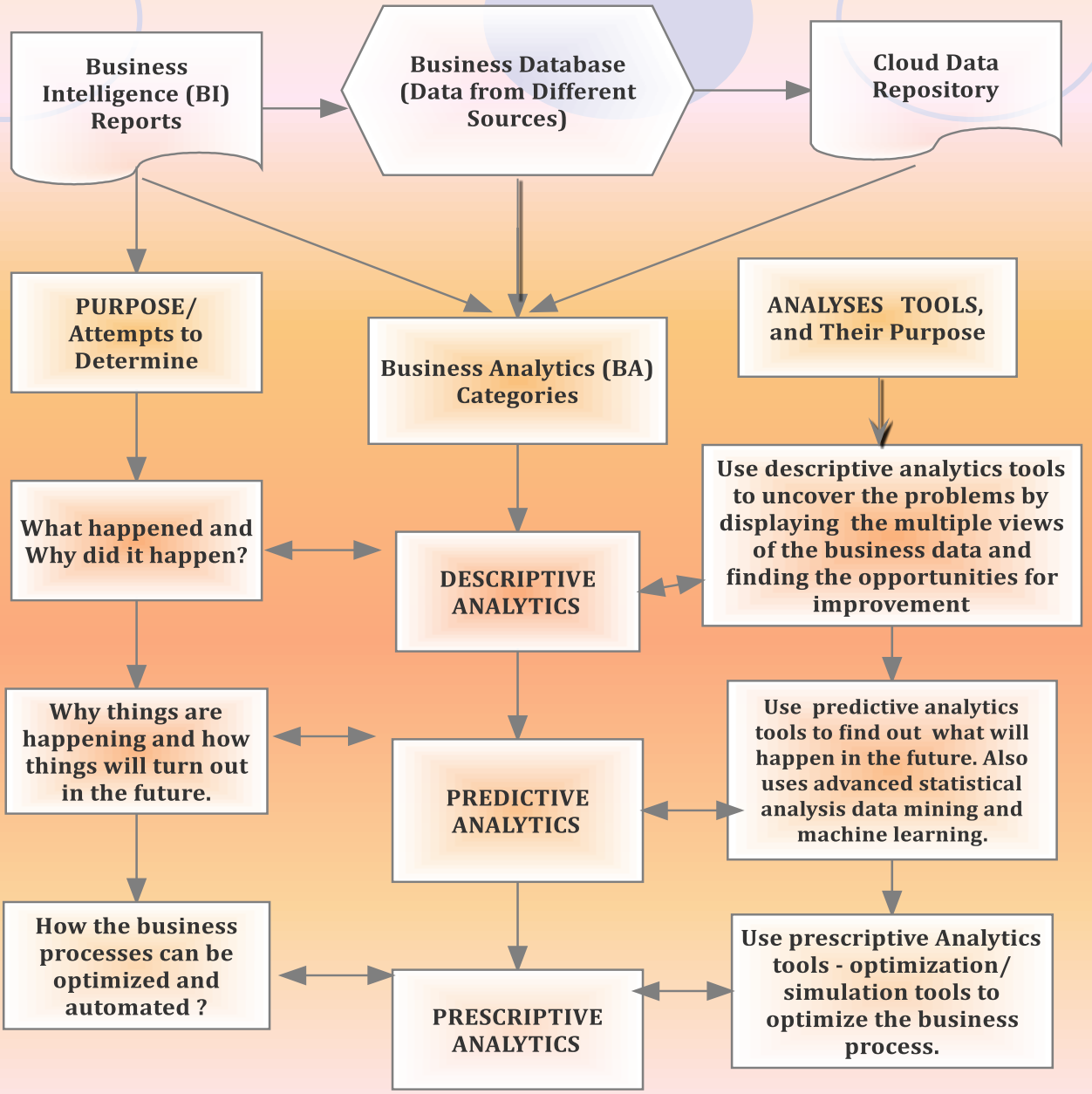
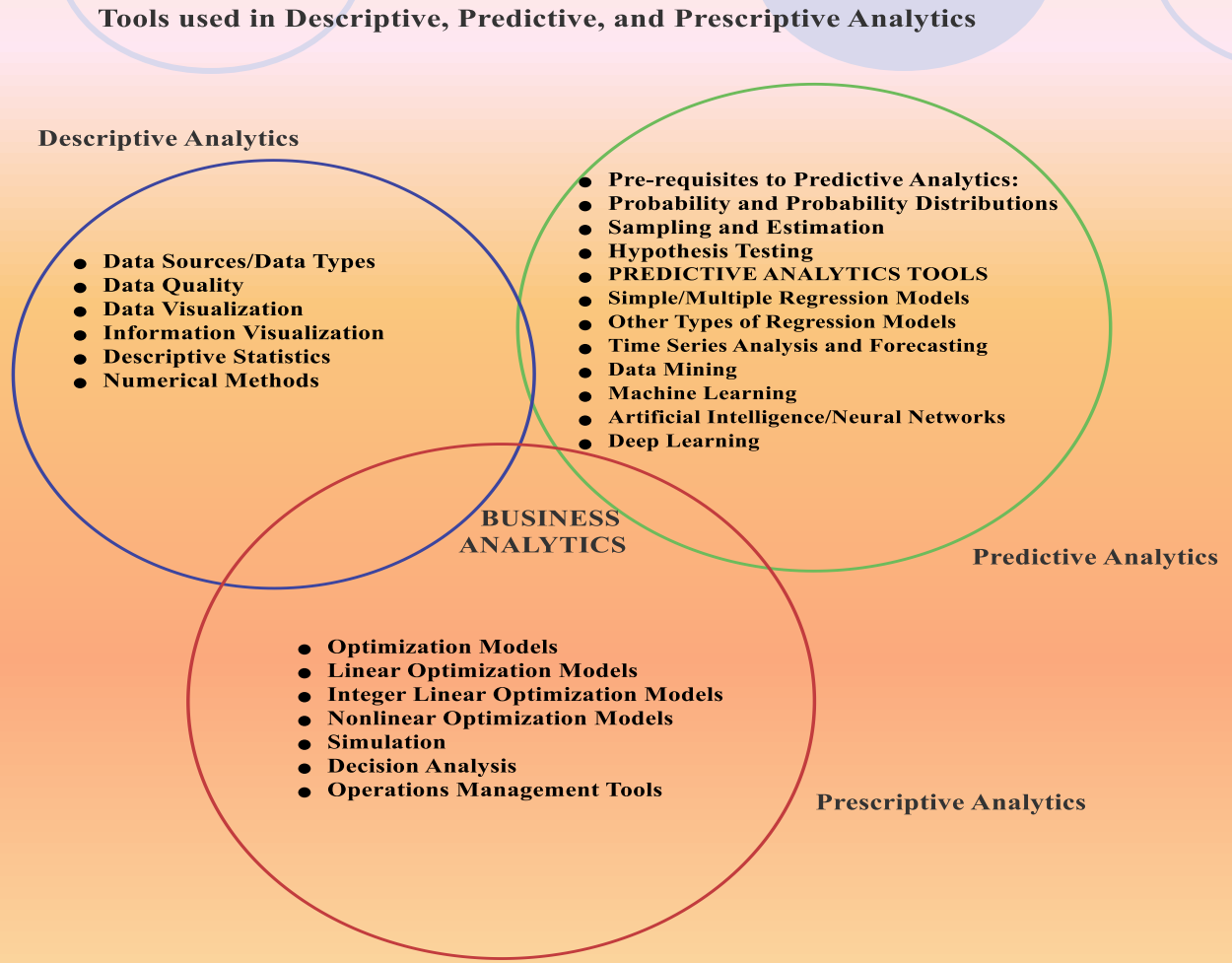


Figure 3.2: Interconnection between the tools of different types of analytics

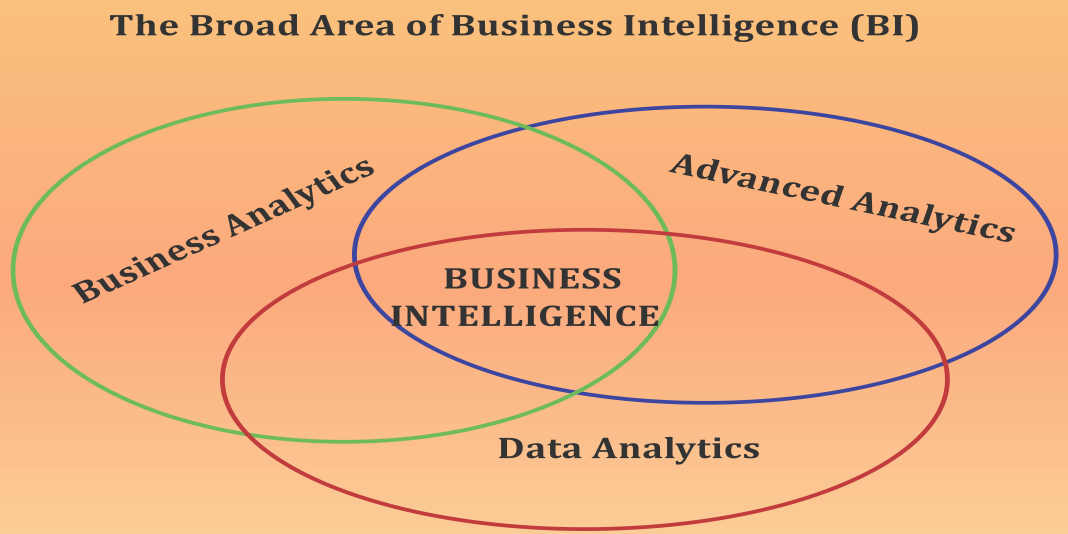


Business Intelligence (BI) and Business Analytics (BA): Differences

- **Business intelligence refers to collecting business data to find information primarily through asking questions, reporting, and online analytical processes (OLAP).**
- Business analytics, on the other hand, uses statistical and quantitative tools and models for explanatory, predictive, and prescriptive modelling.[15]
- Business Intelligence (BI) is the "descriptive" part of data analysis; whereas, Business Analytics (BA) means BI plus the predictive and prescriptive elements, and all the visualization tools and extra bits and pieces that make up the way we handle, interpret visualize, and analyze data.

Figure 3.3: The Broad Area of Business Intelligence (BI)

Broad area of Business Intelligence (BI) that comprises of business analytics, advanced analytics, and data analytics.



Business Intelligence and Business Analytics: A comparison

- The flow chart (next slide) compares the Business Intelligence (BI) to Business Analytics (BA). The overall objectives and functions of a BI program are outlined.
- The Business Intelligence originated from reporting but later emerged as an overall business improvement process that provides the current state of the business. The information about what went wrong or what is happening in the business provides opportunities for improvement.

Business Intelligence and Business Analytics: A comparison…cont.

- BI may be seen as the descriptive part of data analysis but when combined with other areas of analytics - predictive, advanced analytics, and data analytics provides a powerful combination of tools.

- These tools enable the analyst and data scientists to look into the business data, the current state of the business, and make use of predictive, prescriptive, data analytics tools as well as the powerful tools of data mining to guide an organization in business planning, predicting the future outcomes, and make effective data driven decisions.

Figure 3.4: Comparing Business Intelligence (BI) and Business Analytics

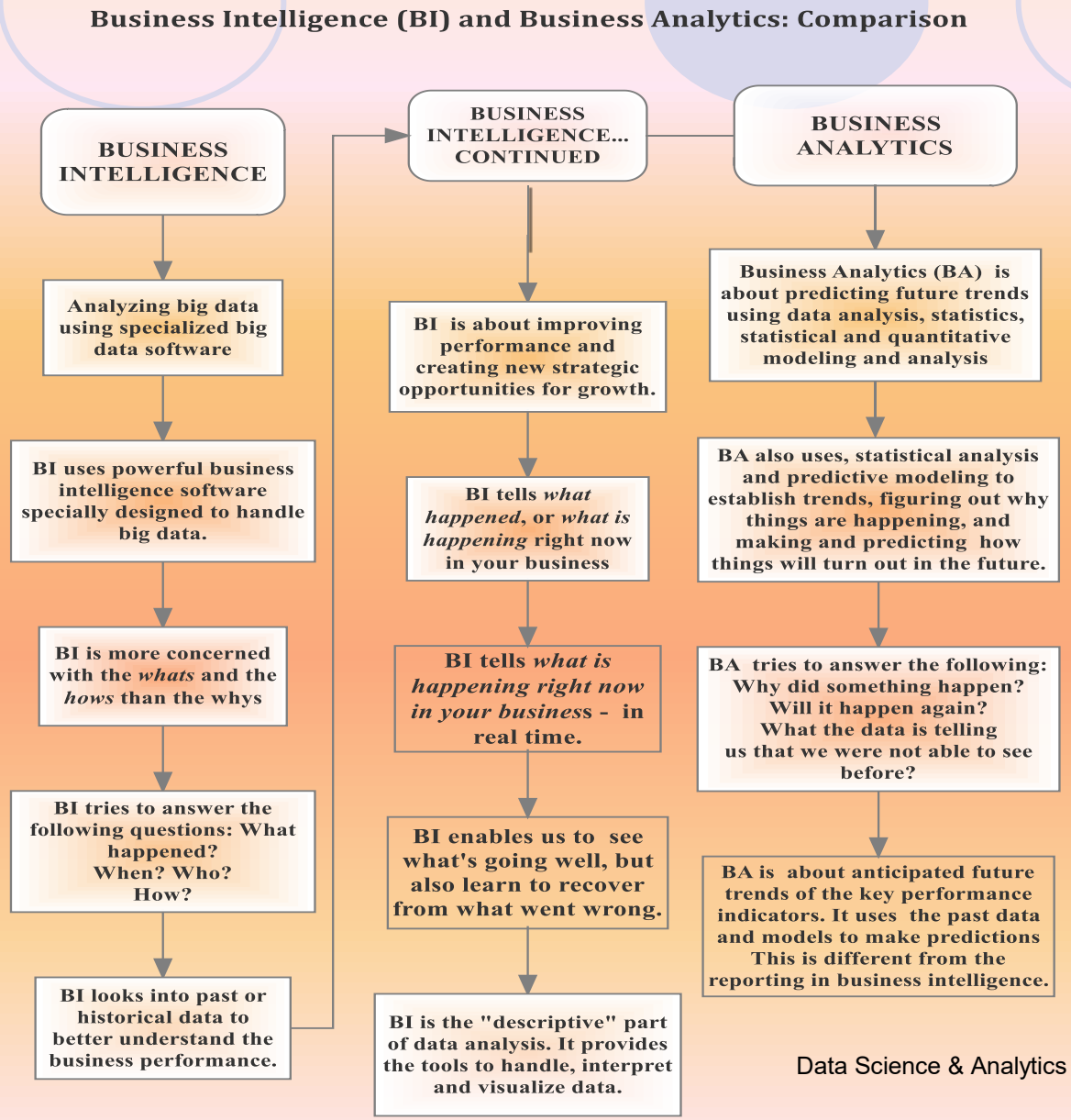
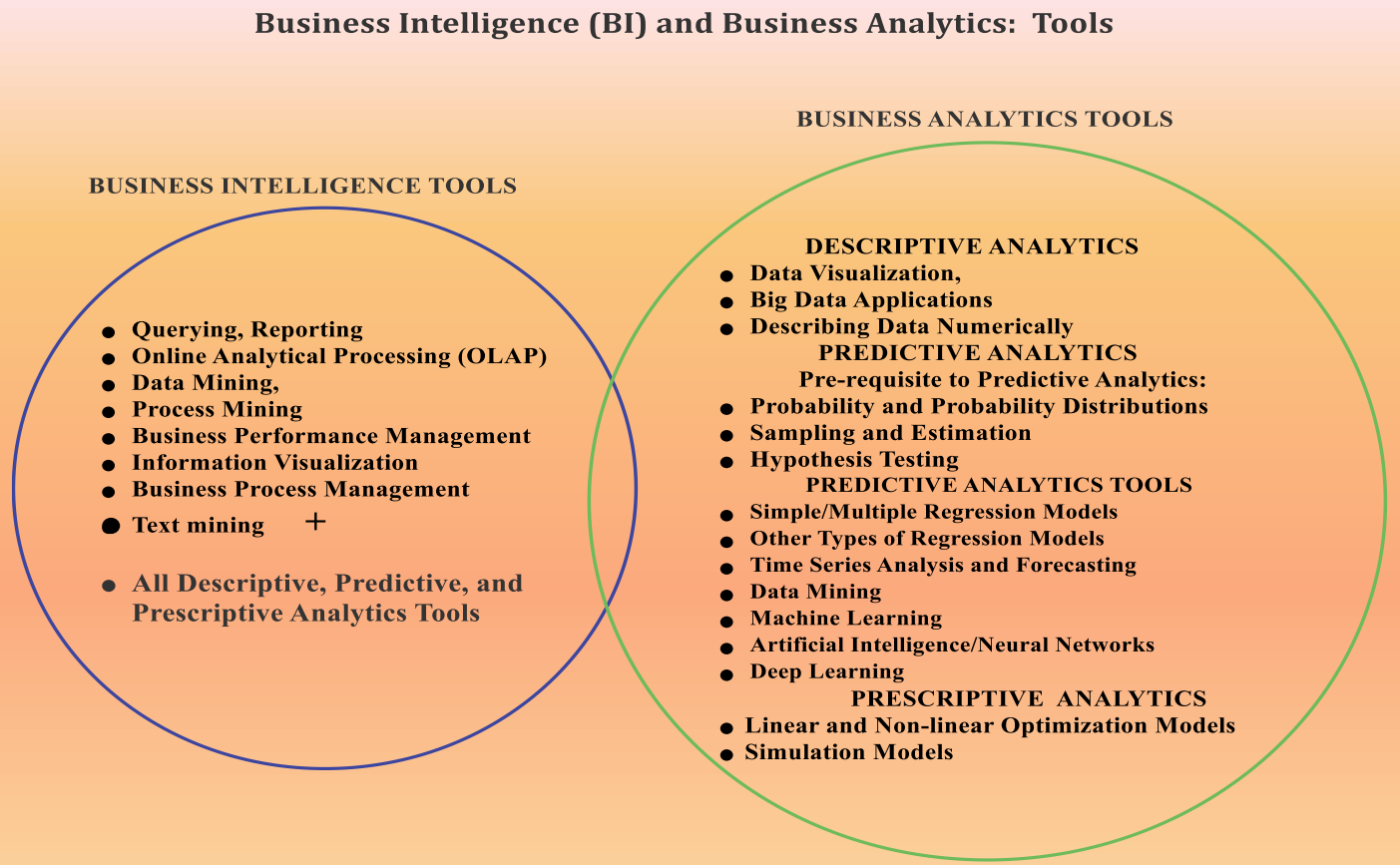


Figure 3.5: Business Intelligence (BI) and Business Analytics (BA) Tools

Figure shows the tools of business intelligence and business analytics. Note that the tools overlap in the two areas. Some of these tools are common to both.



Brief Overview of Data &Data Science Tools

Data Related Terms

- **Big Data:** Big Data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing application [Wikipedia].

As per **O'Reilly media**:

- **Big data** is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it. **O'Reilly Media** made Big Data popular.

Big Data

Gartner who was credited with the 3 ‘V’s of Big Data classified the big data as:

High-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

Figure 4.1: Classification of Data

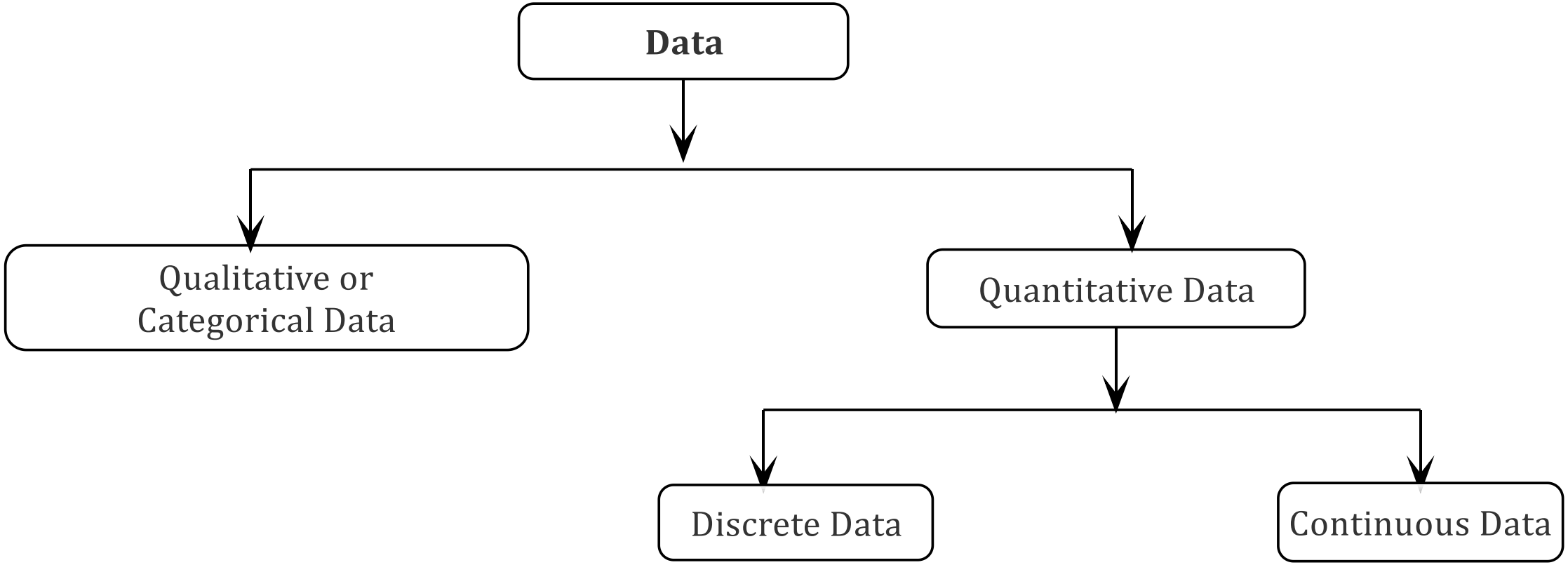
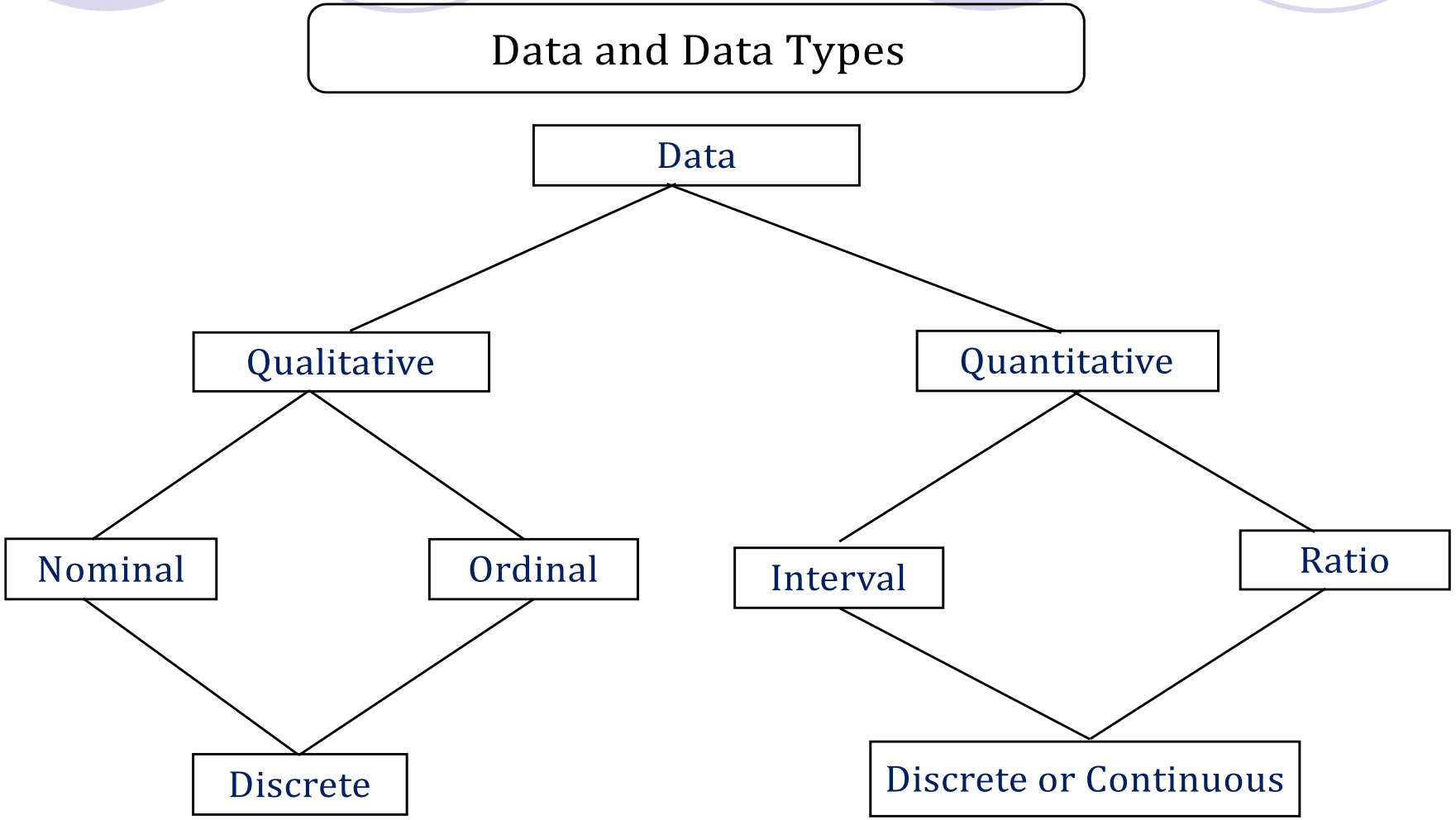


Figure 4.2 Classifications of Data



Big Data

Gartner who was credited with the 3 ‘V’s of Big Data classified the big data as:

High-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

Other Data Related Terms

- **Big Data:** Big Data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing application [Wikipedia].

As per **O'Reilly media**:

- **Big data** is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it. **O'Reilly Media** made Big Data popular.

Data Related Terms

- **Data mining:** Data mining is about finding meaningful patterns and deriving insights in large sets of data using sophisticated pattern recognition techniques. It is closely related to Analytics that we discussed earlier. To derive meaningful patterns, data miners use statistics and statistical modeling, machine learning algorithms, and artificial intelligence.
- **Data warehouse:** A **data warehouse** (**DW** or **DWH**), or **enterprise data warehouse** (**EDW**), is a system for storing, reporting, and analysis of huge amounts of data. The purpose of DW is creating reports and performing analytics which are core component of ***Business Intelligence.***

3 V's in Data Types

- **Structured vs. Unstructured Data: Structured vs. Unstructured Data:** are the 'Volume' and 'Variety' – the 'V's of Big Data in Structured data is the data that can be stored in the relational databases. This type of data can be analyzed and organized in such a way that can be related to other data via tables. Unstructured data cannot be directly put in the data bases or analyzed or organized directly. Some examples are email/text messages, social media posts and recorded human speech etc.
- **Data quality:** Data quality is affected by the way data is collected, entered in the system, stored and managed. Efficient and accurate storage (data warehouse), cleansing, and data transformation are critical for assuring data quality.

Data quality

- The process of verifying the reliability and effectiveness of data is sometimes referred to as Data quality assurance (DQA). The effectiveness, reliability, and success of business analytics (BA) and business intelligence (BI) depend on the acceptable data quality.

The following are important considerations in assuring data quality. Aspects of data quality include: [http://searchdatamanagement.techtarget.com/definition/data-quality]

- **Accuracy, Completeness,Update status, Relevance, Consistency across data sources, Reliability, Appropriate presentation, Accessibility**

- Within an organization, acceptable data quality is crucial to operational and transactional processes.

Data Visualization: Some Visuals

Figure 5.27: Cross tabulation and Bar Chat of Categorical Data

Table 2A.22 shows the pivot table. The bar chart is displayed in Figure 2A.18.

Table 2A.22

Count of Major	Column Labels					
Row Labels	**1**	**2**	**3**	**4**	**5**	**Grand Total**
Female	6	11	34	32	19	102
Male	12	28	11	31	16	98
Grand Total	**18**	**39**	**45**	**63**	**35**	**200**

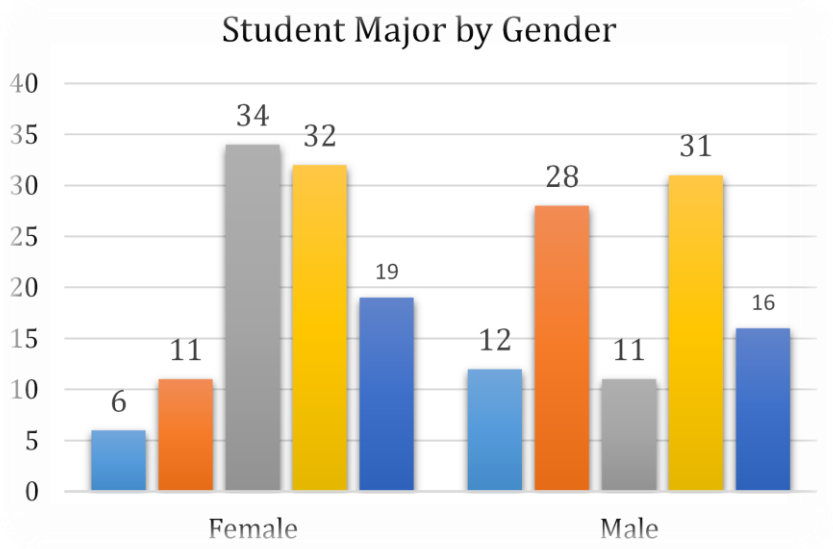


Figure 2A.18: Student Major by Gender

Figure 5.31: Histogram of 1000 Random Numbers from a Normal Distribution

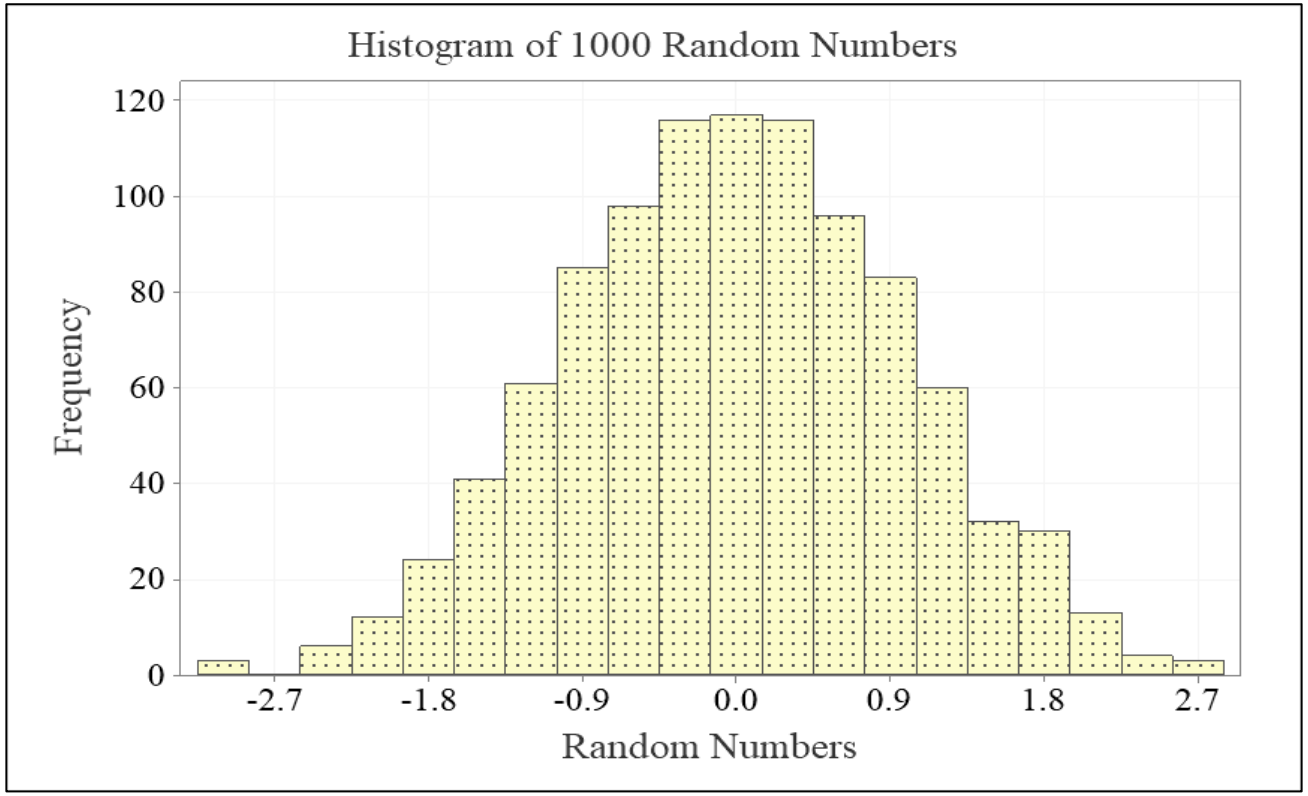


Figure 5.37: A Histogram of Home Sales data) approximating the Shape

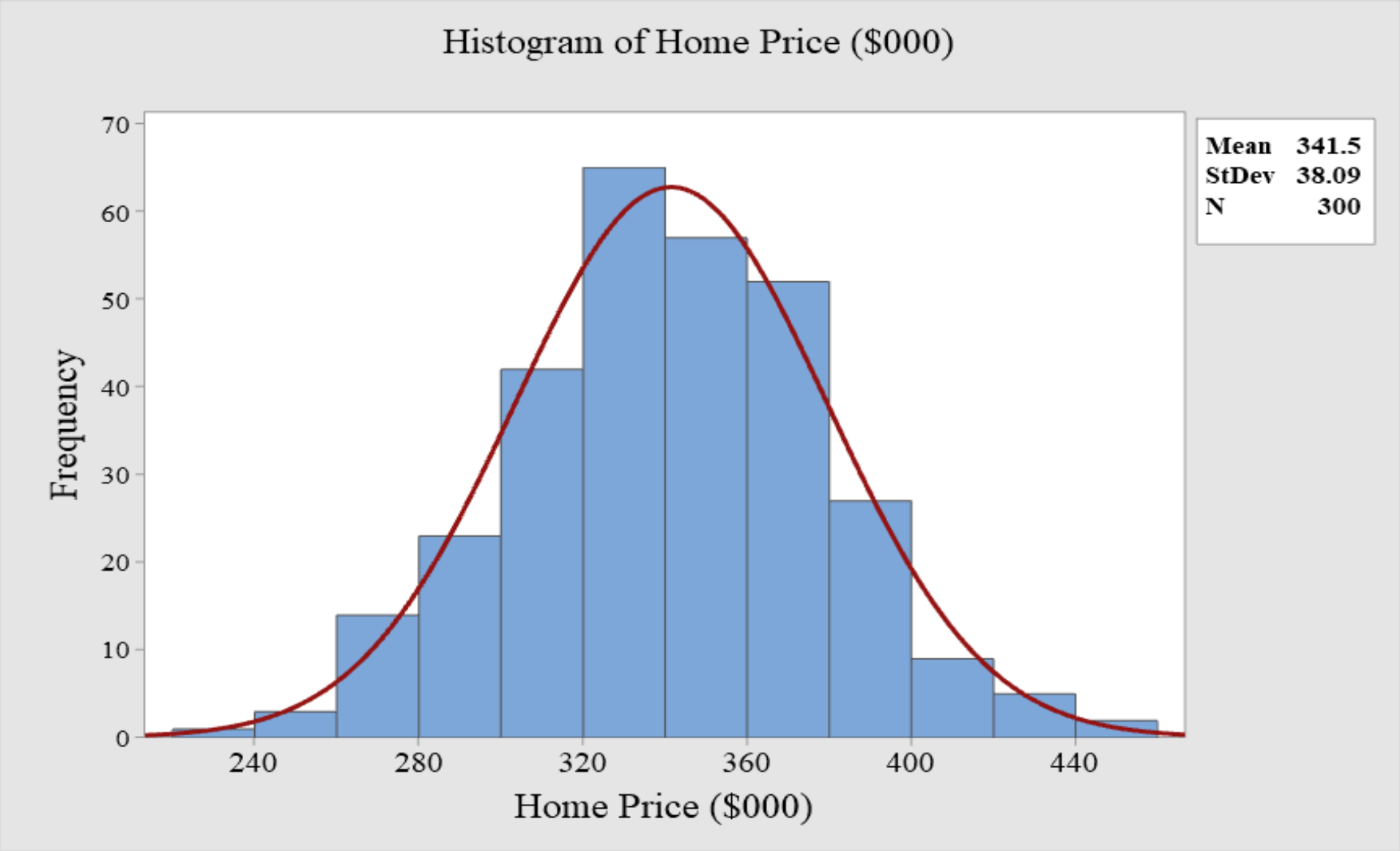


Figure 5.44: Bar Chart – largest internet companies (April 2020)

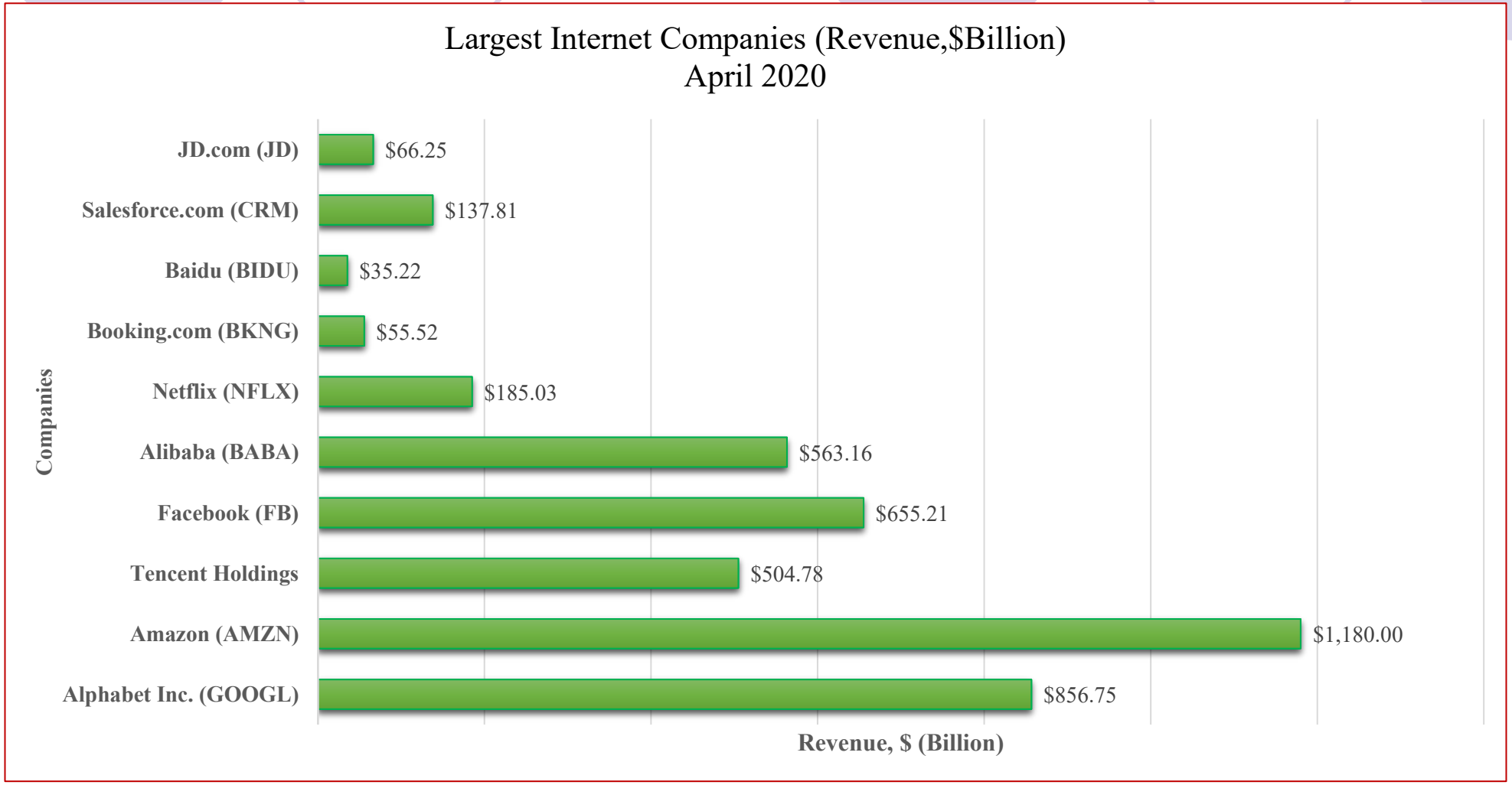


Figure 5.45: A Vertical Bar Chart – largest internet companies (April 2020)

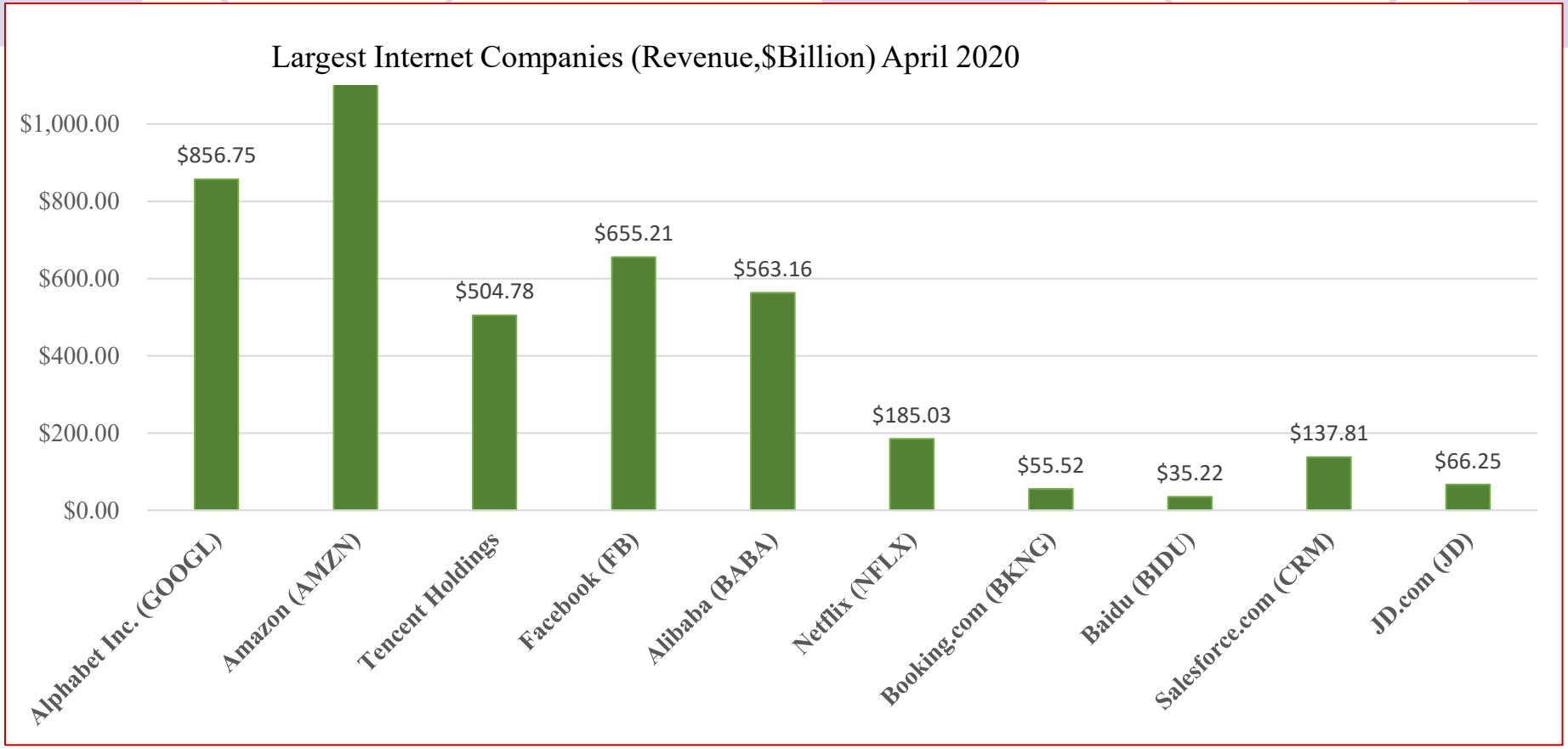


Figure 5.46: Pareto Chart – largest internet companies (April 2020)

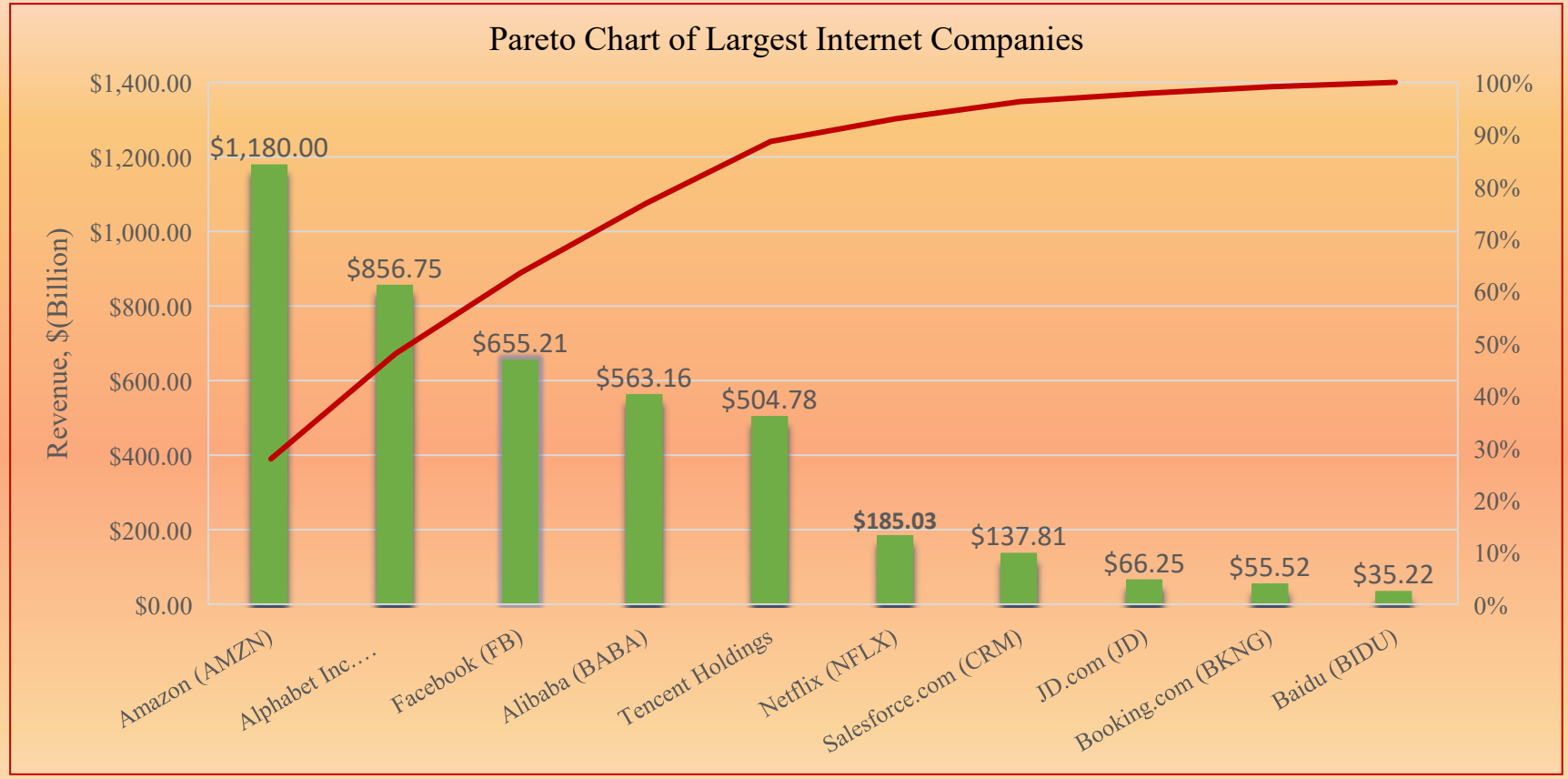


Figure 5.47: A Clustered column chart of quarterly sales

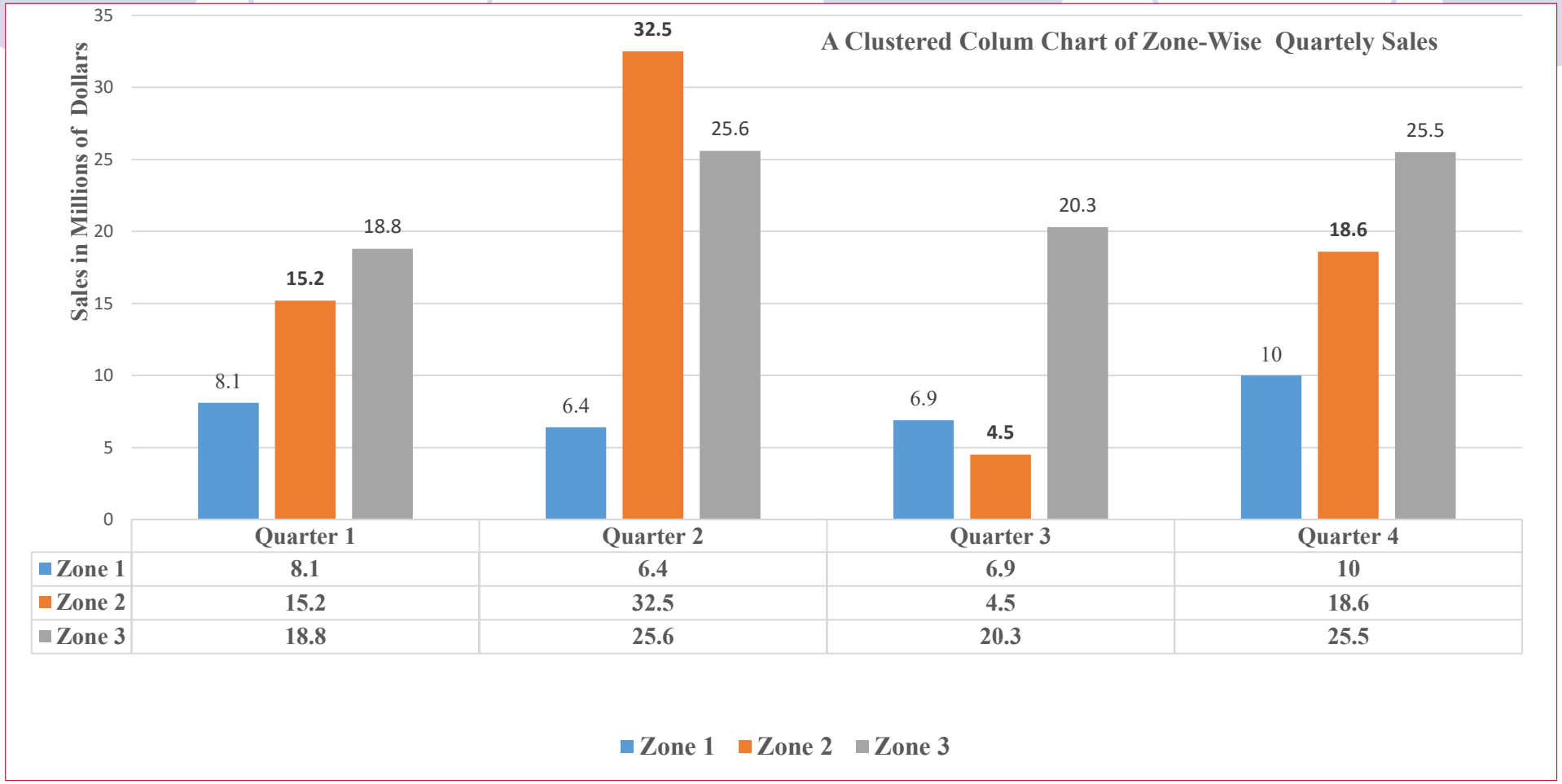


Figure 5.48: A Stacked column chart of quarterly sales

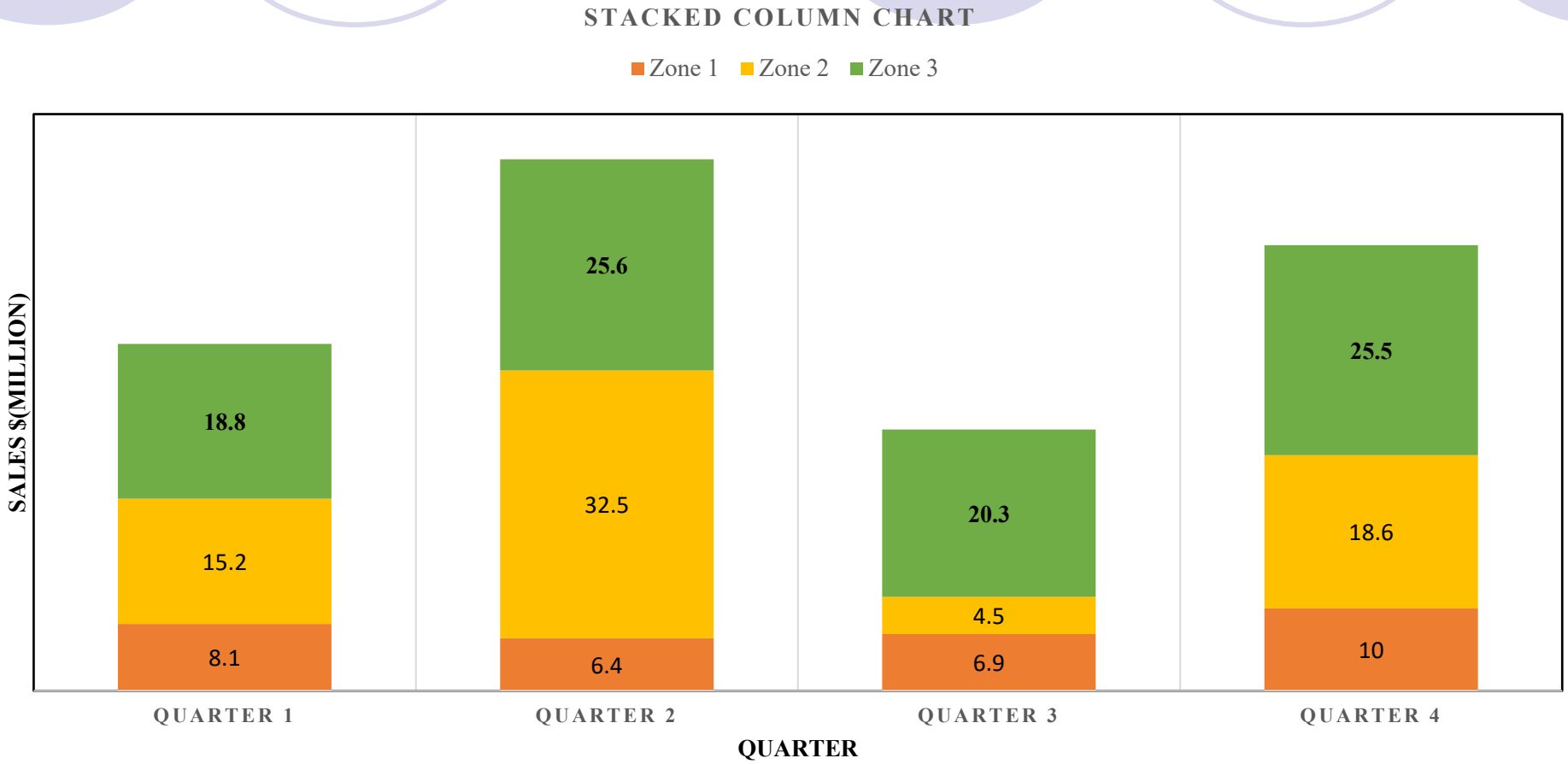


Figure 5.49: A variation of Clustered column chart of quarterly sales

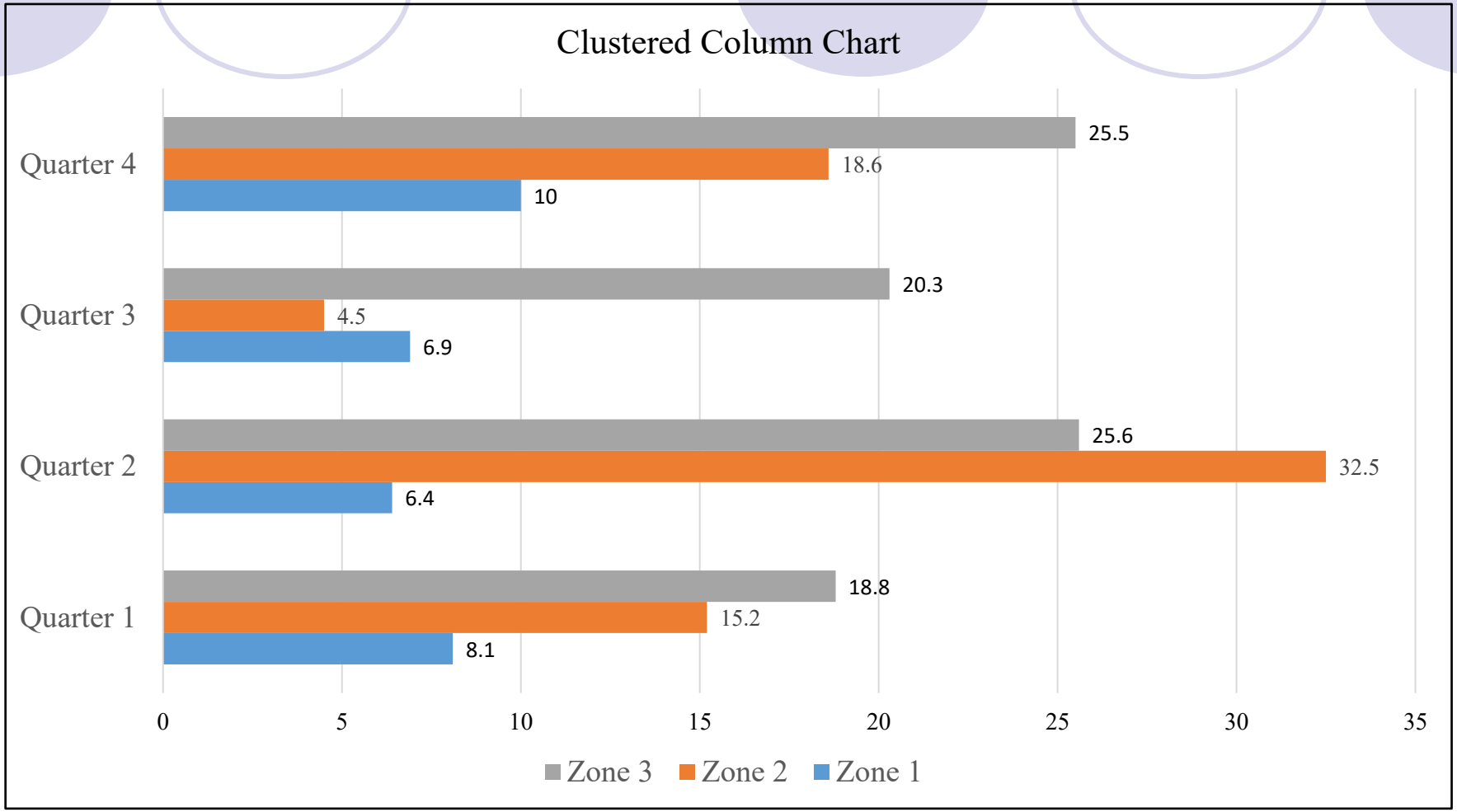


Figure 5.50: A Stacked Area chart of quarterly sales

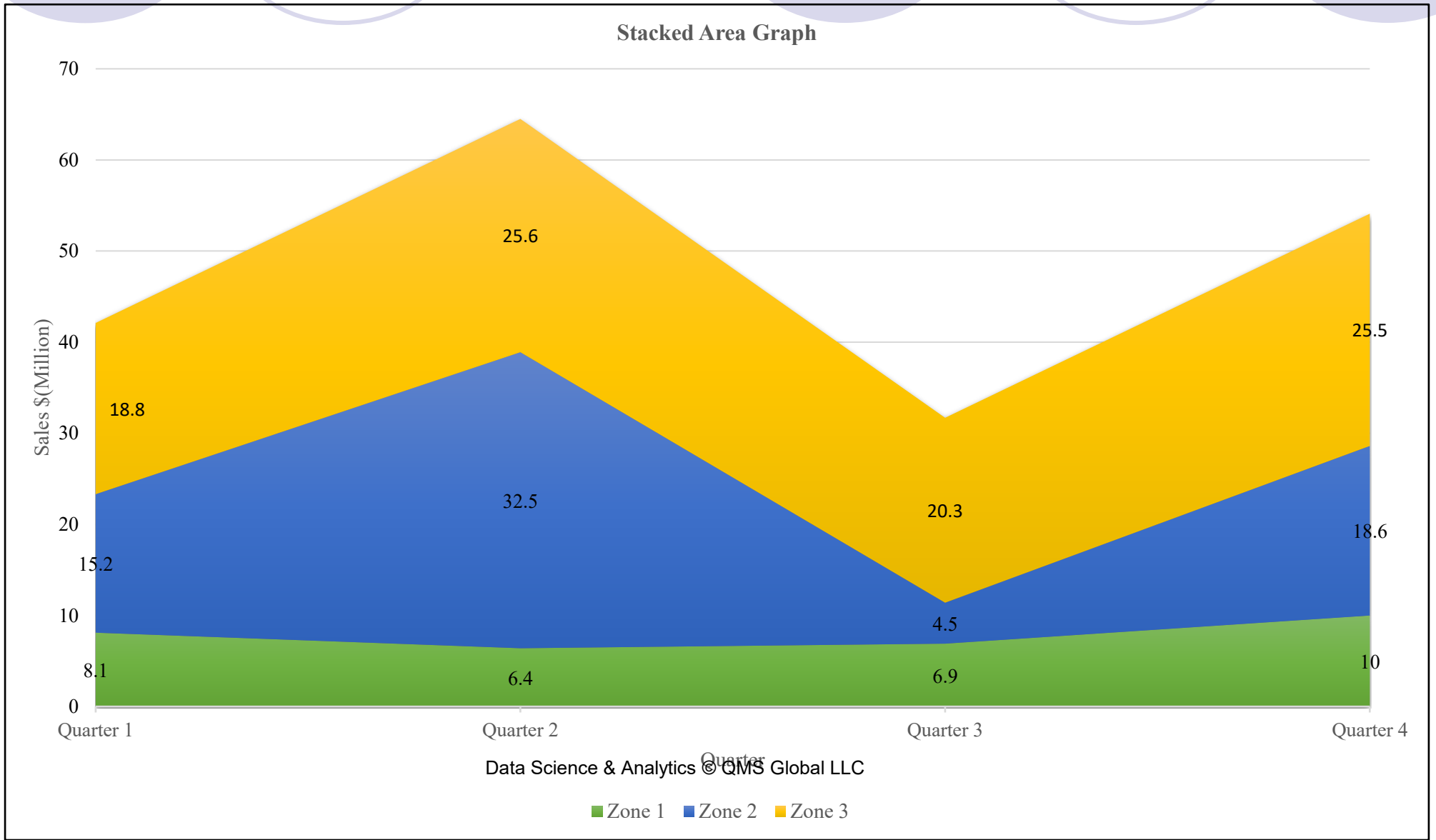


Figure 5.51: A Stacked Area Graph of Energy Consumption

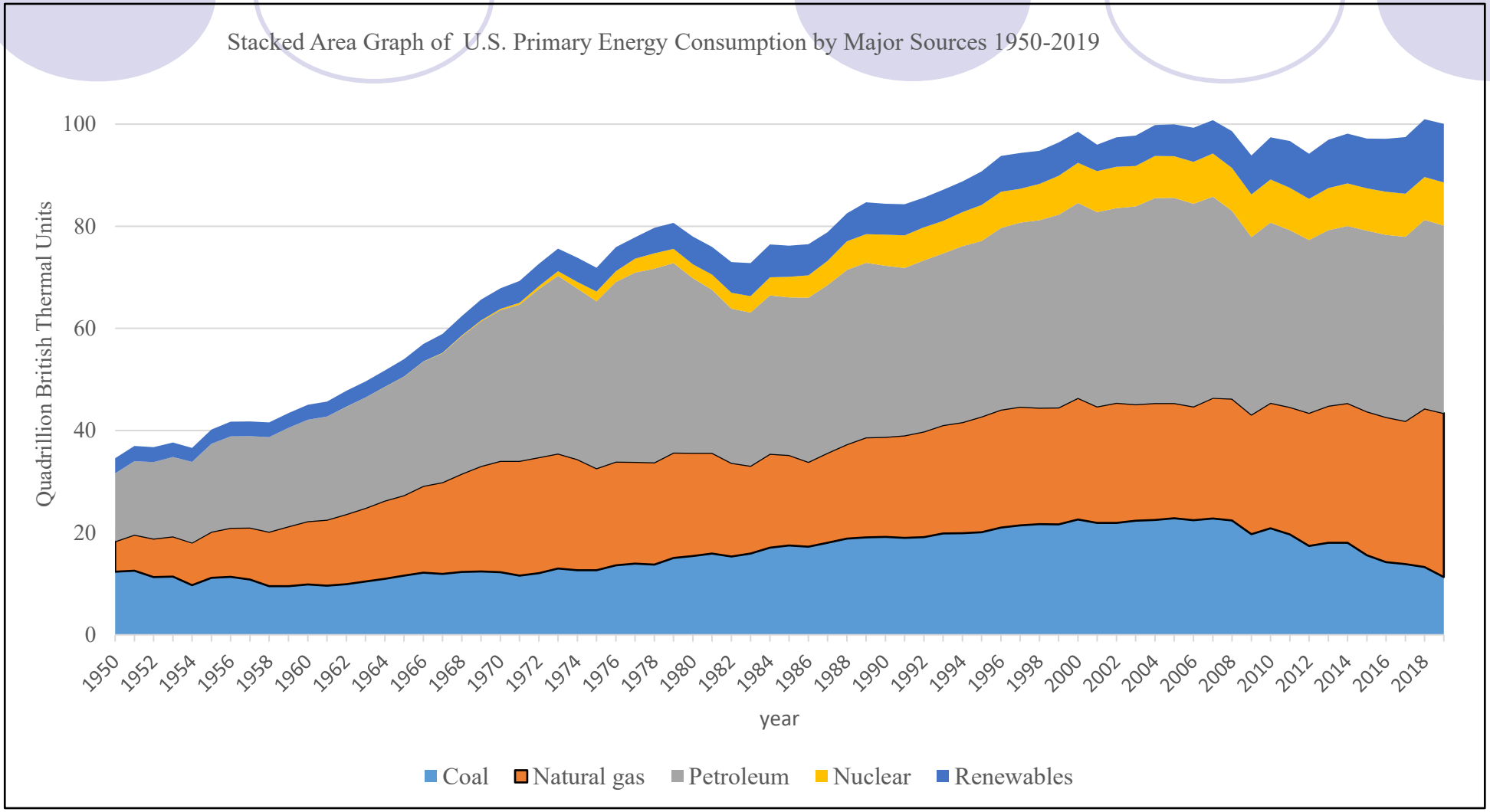


Figure 5.52: A Line Chart of Energy Consumption

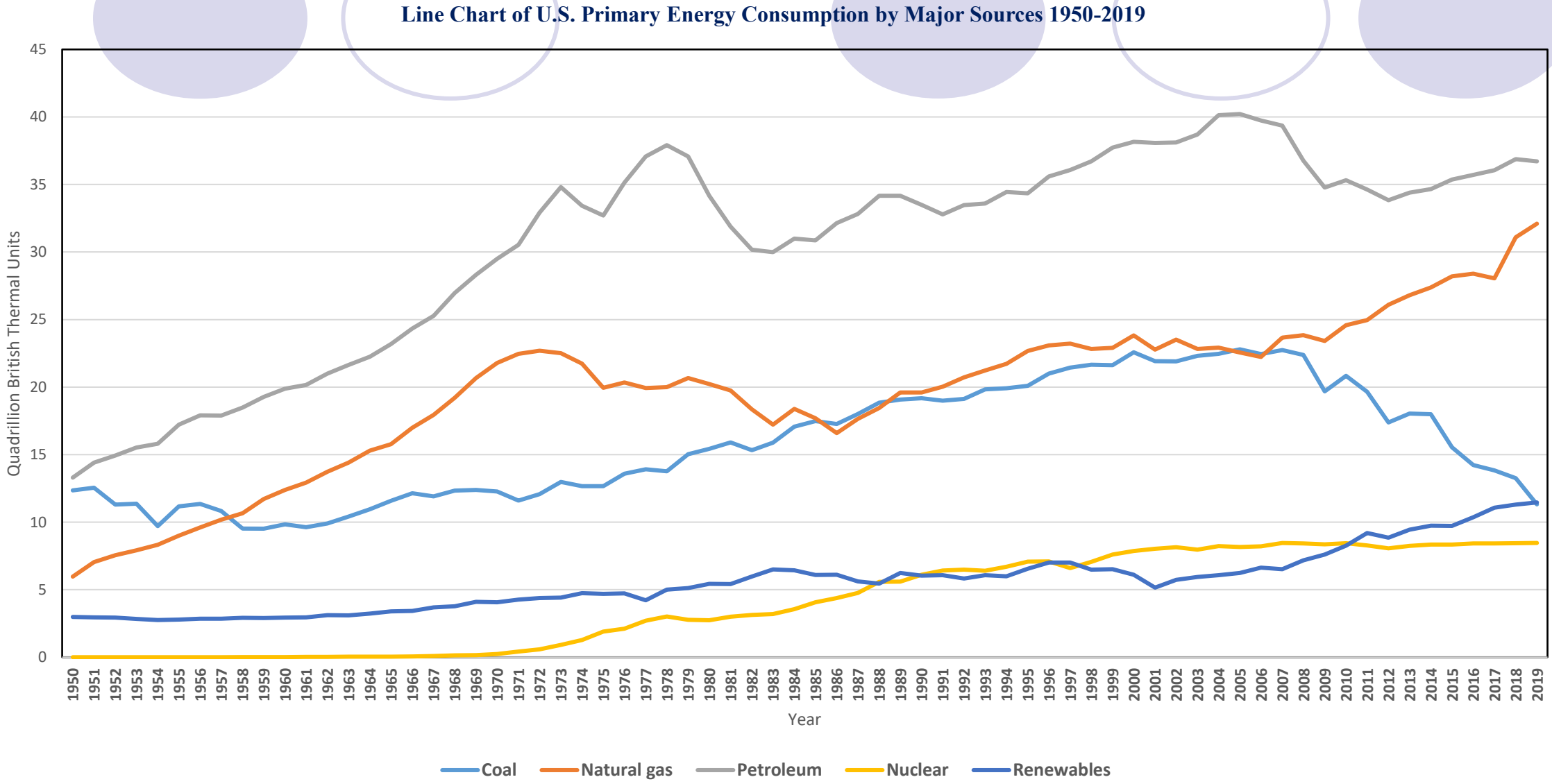


Figure 5.53: A Bar Chart of Categorical Variable showing Revenue of Amazon

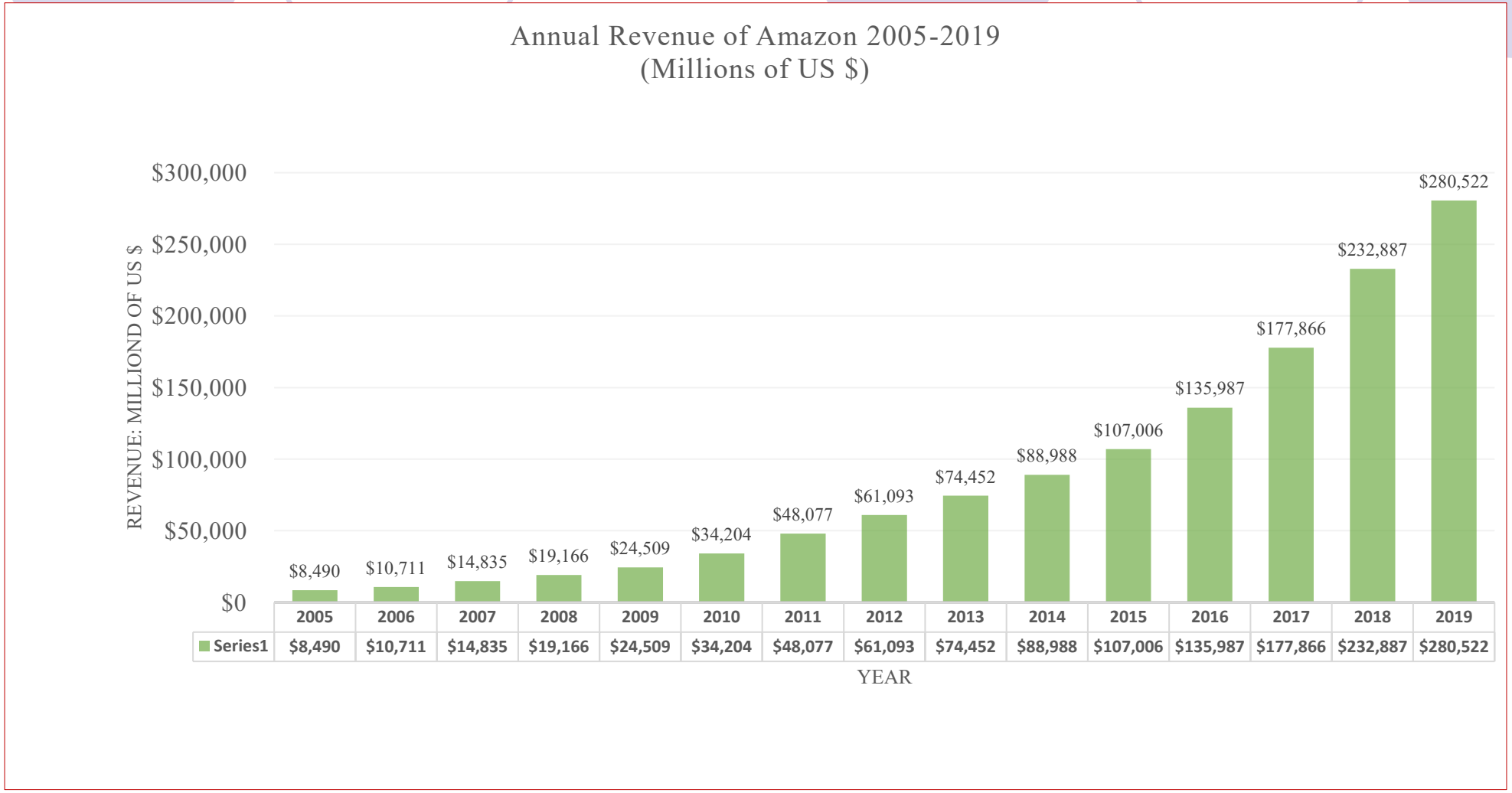


Figure 5.54: A Pie Chart of U.S. Market Share for 2018

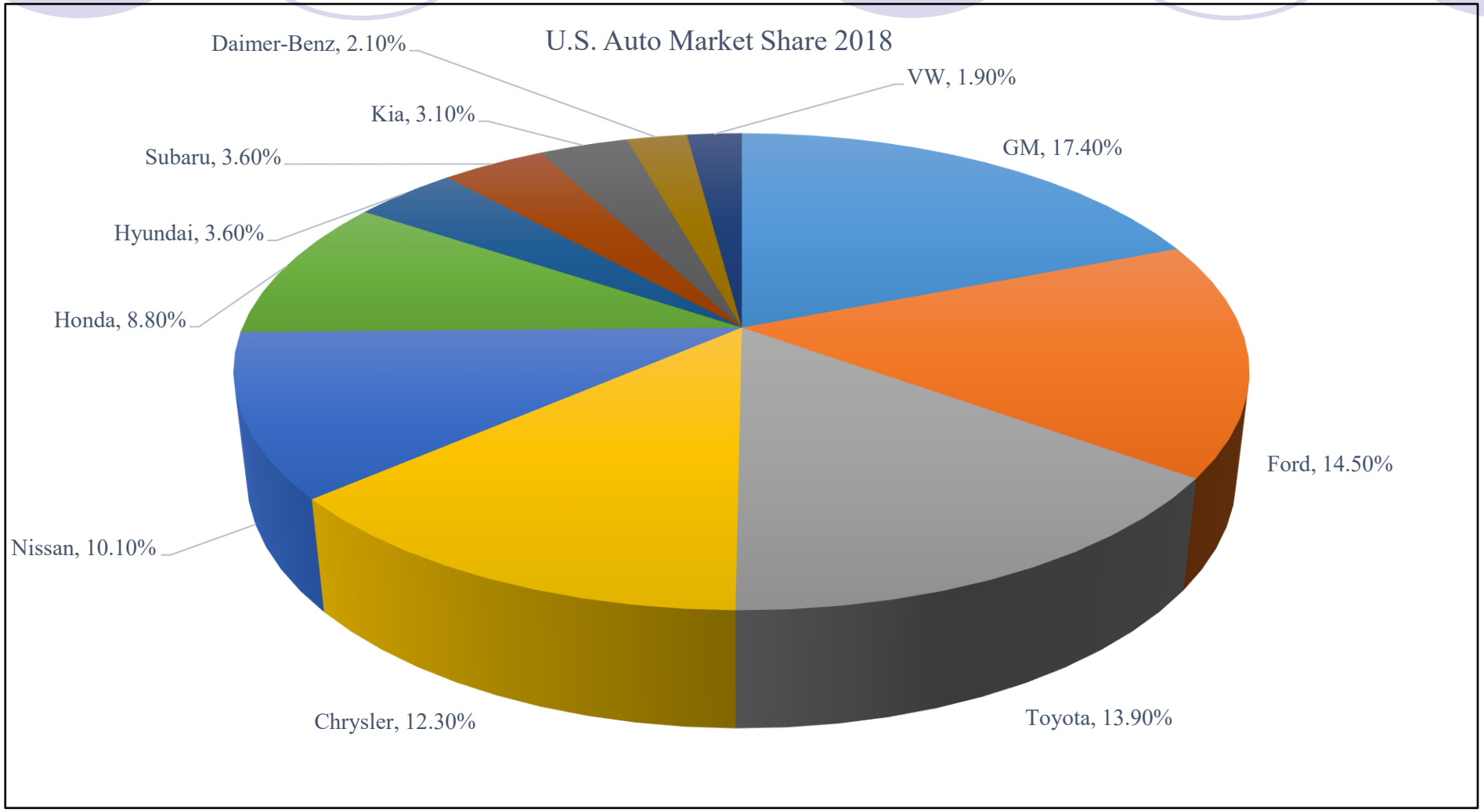
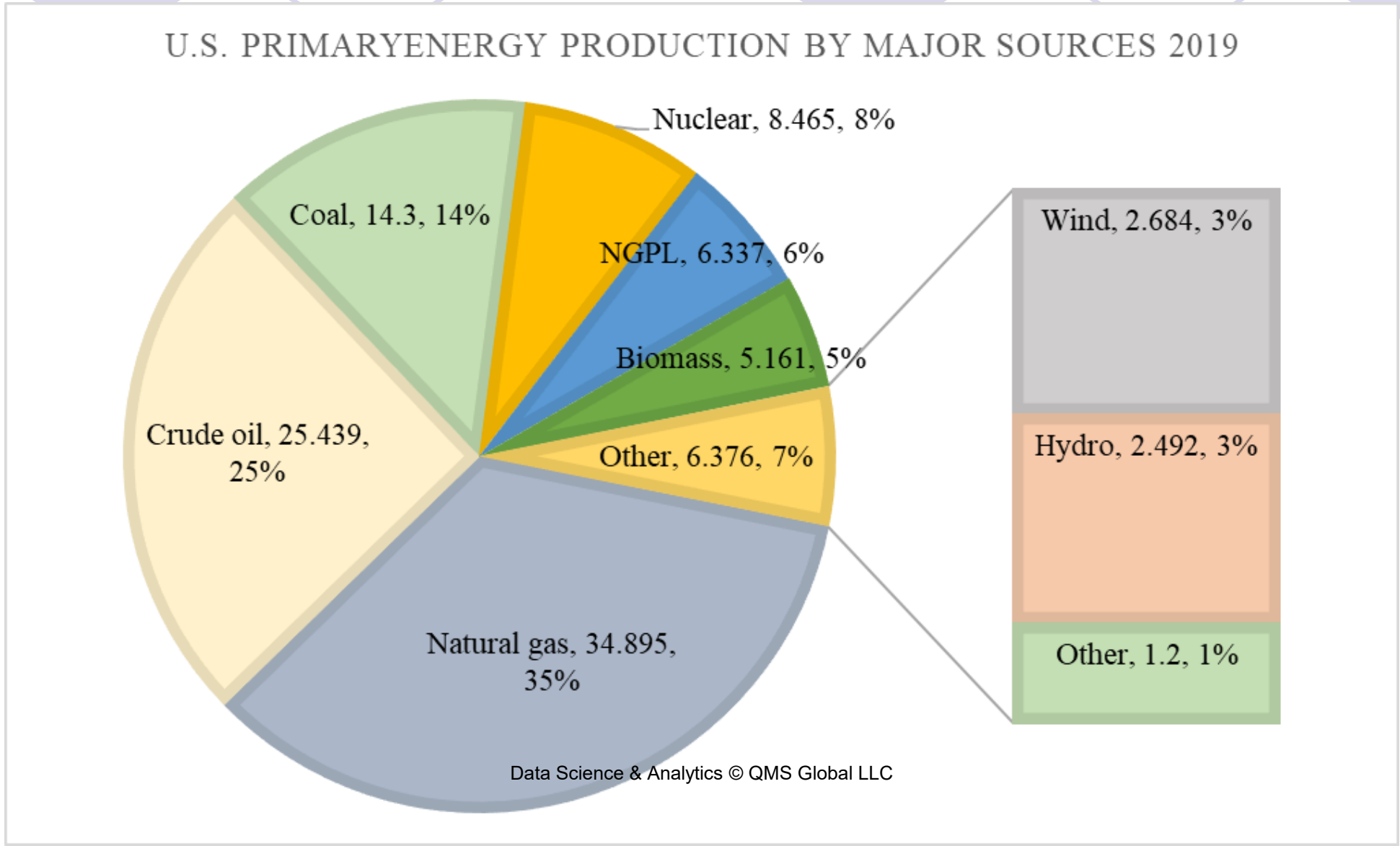


Figure 5.55: A Bar of a Pie Chart of U.S. Market Share for 2018



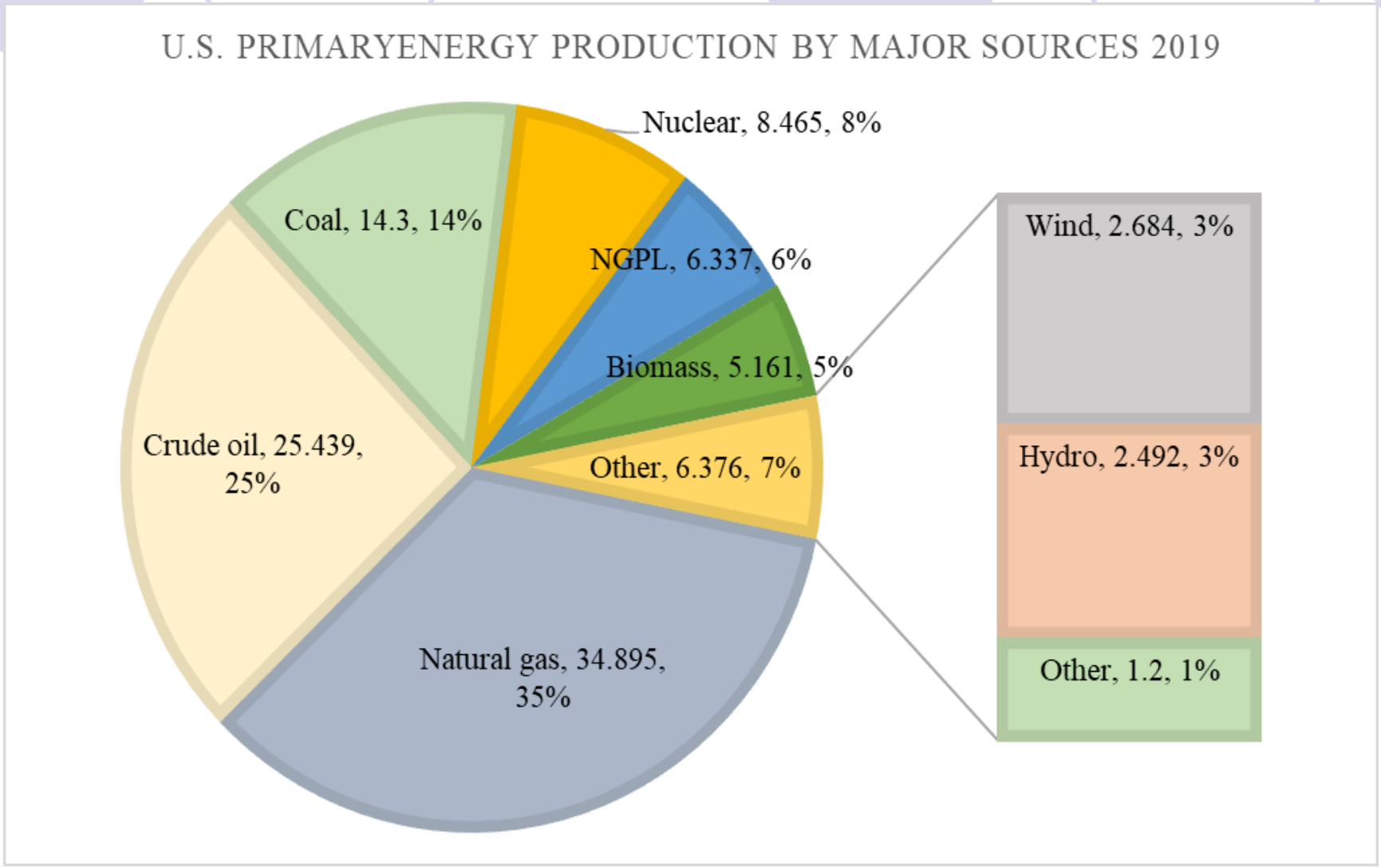

U.S. PRIMARYENERGY PRODUCTION BY MAJOR SOURCES 2019
Nuclear, 8.465, 8%
Coal, 14.3, 14%
NGPL, 6.337, 6%
Biomass, 5.161, 5%
Crude oil, 25.439, 25%
Other, 6.376, 7%
Natural gas, 34.895, 35%
Wind, 2.684, 3%
Hydro, 2.492, 3%
Other, 1.2, 1%

Figure 5.56: A Pie of pie Chart

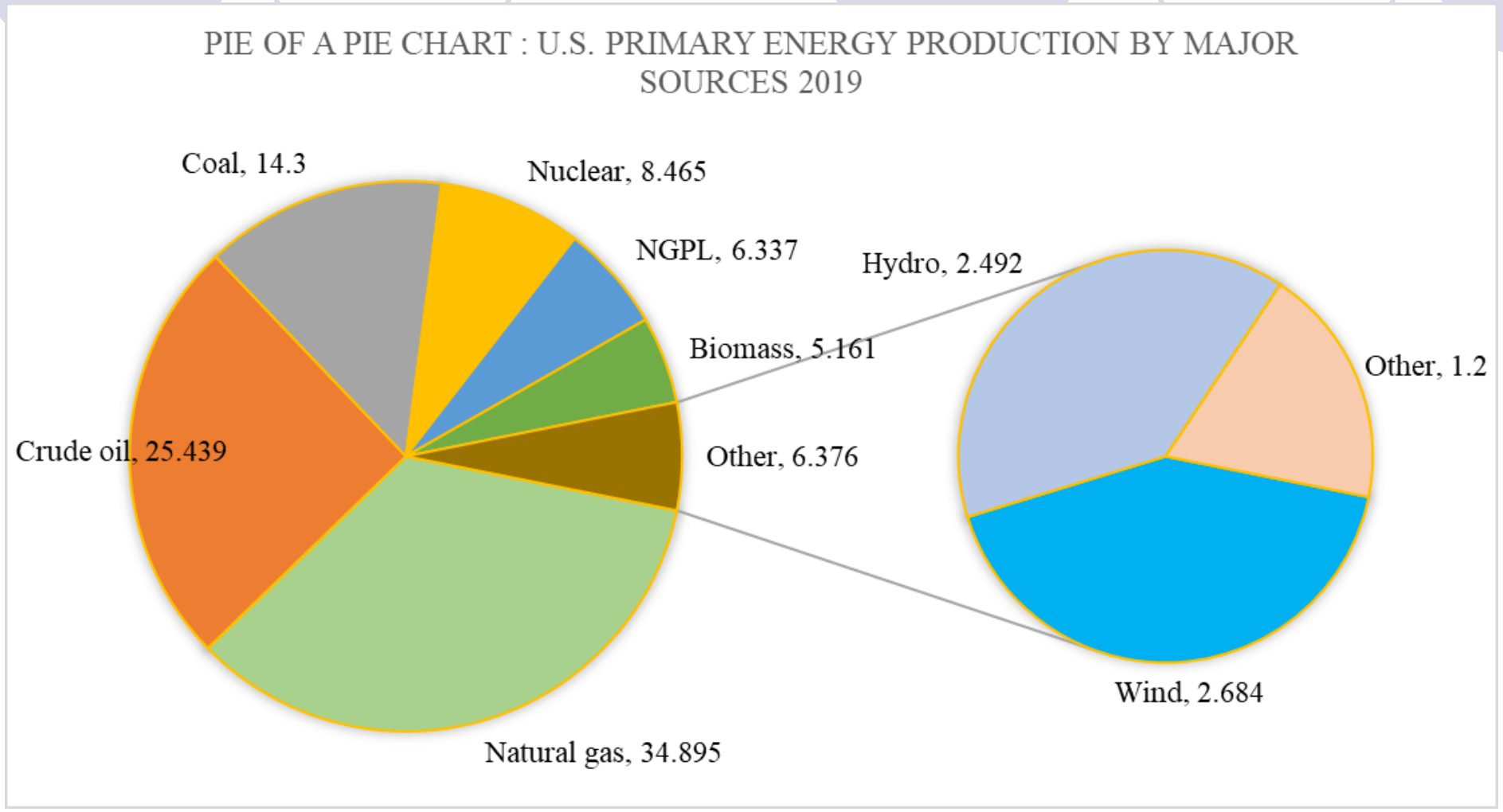


Figure 5.57: Scatter plot of Sales and Profit

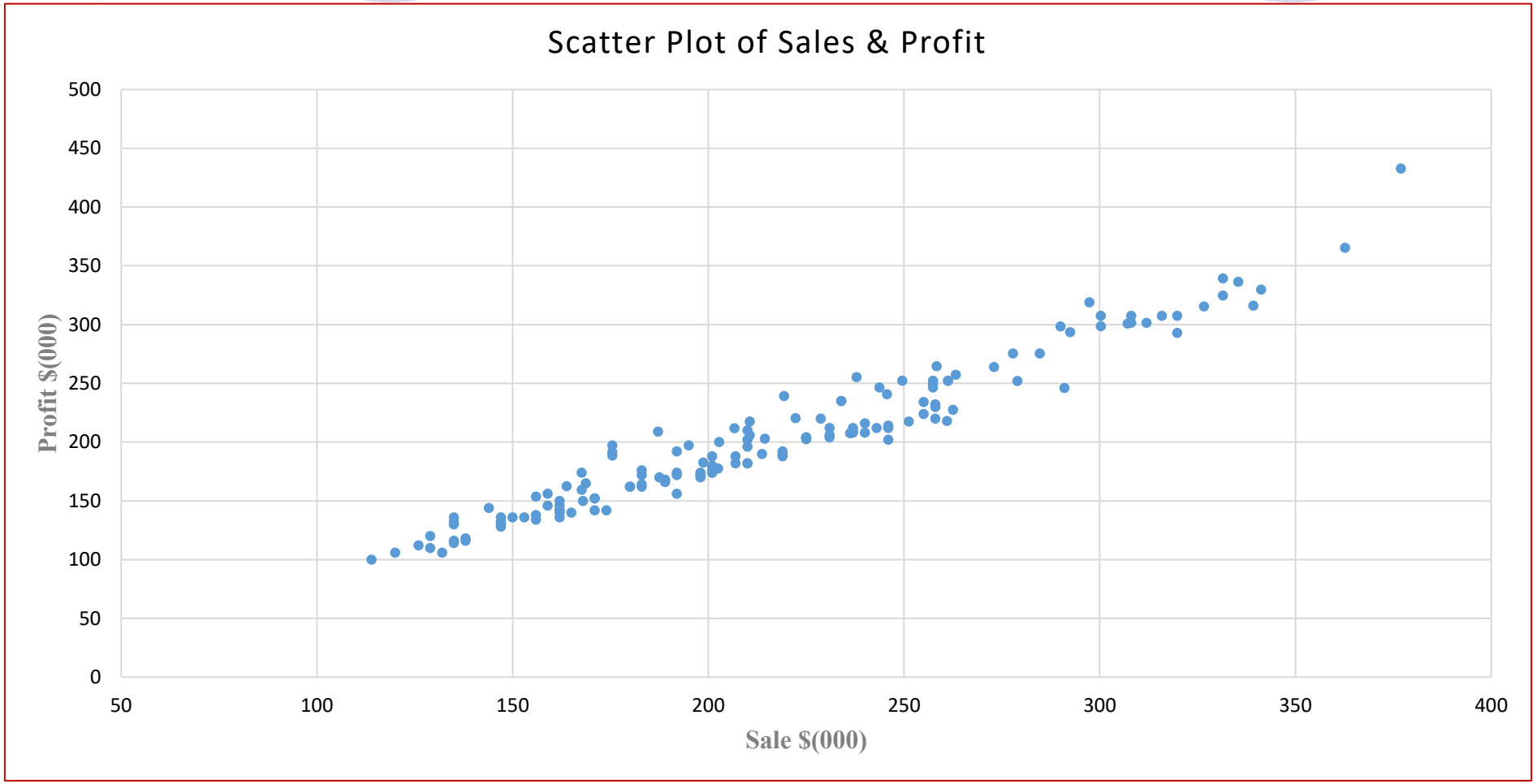


Figure 5.58: Fitted line plot of Sales and Profit

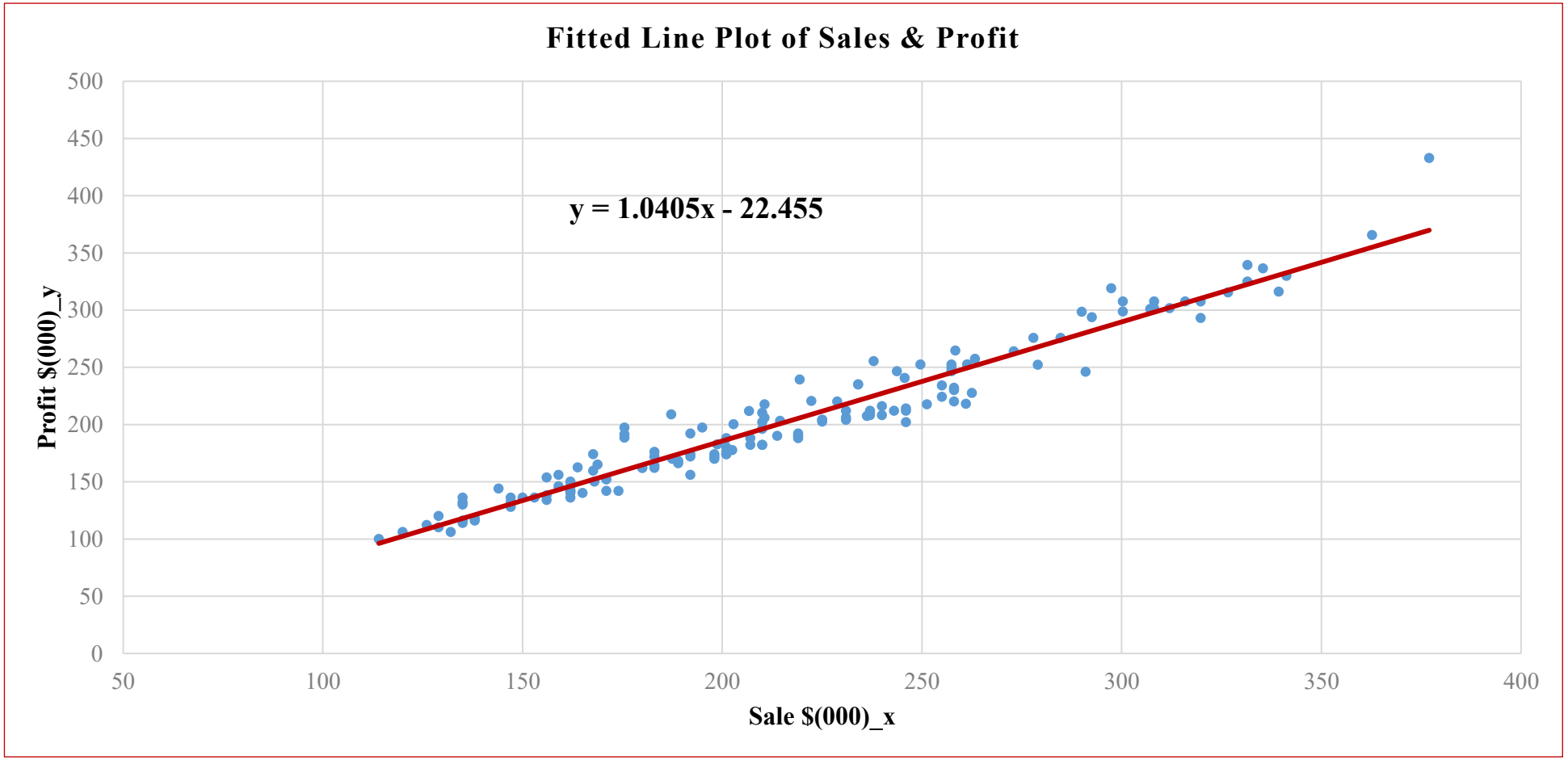


Figure 5.59: Bubble Chart of Advertisement, Sales, and Store Size

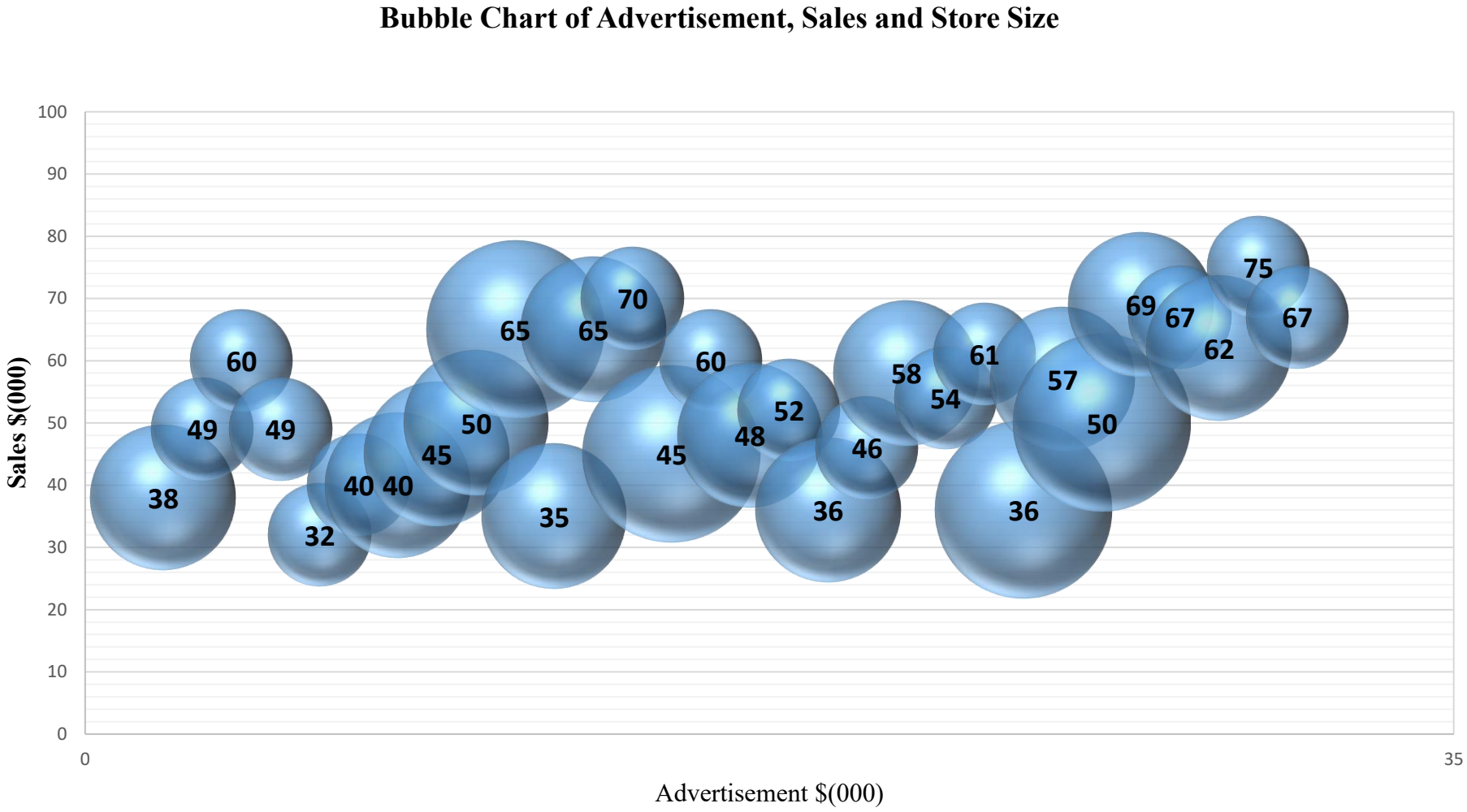


Figure 5.60: A Variation of Bubble Chart of Advertisement, Sales, and Store Size

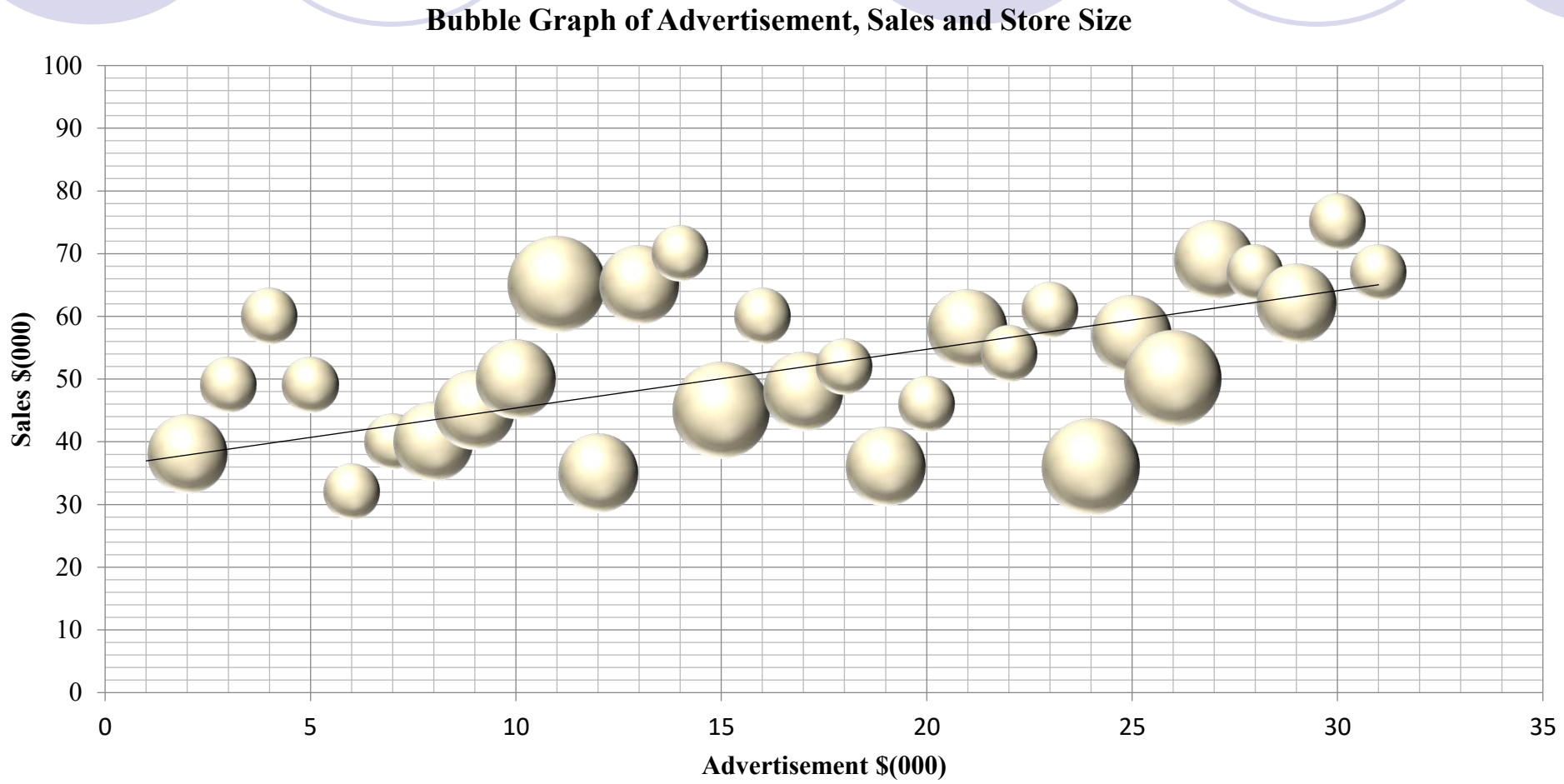


Figure 5.62: Time Series Plot of Demand Data

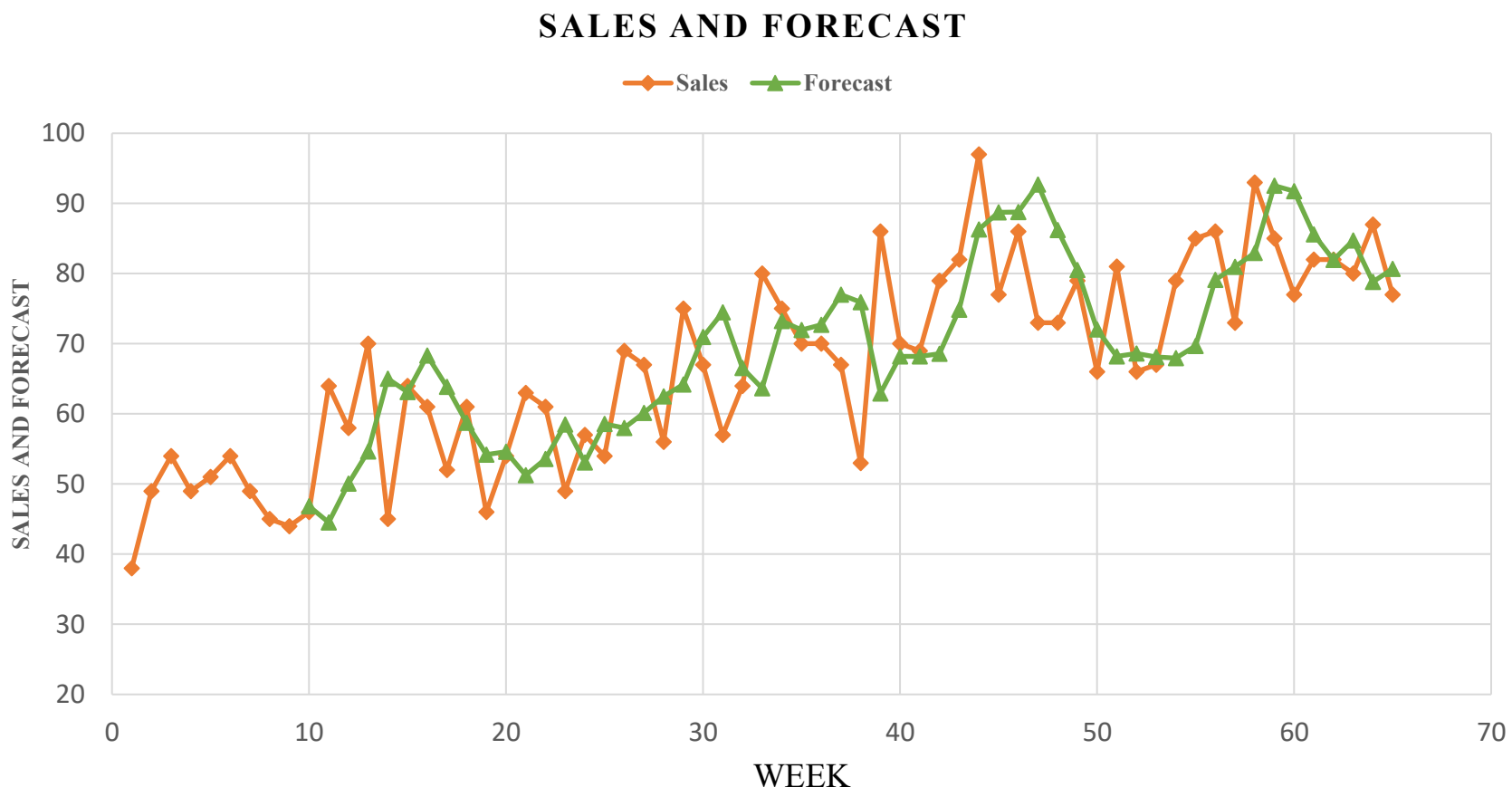
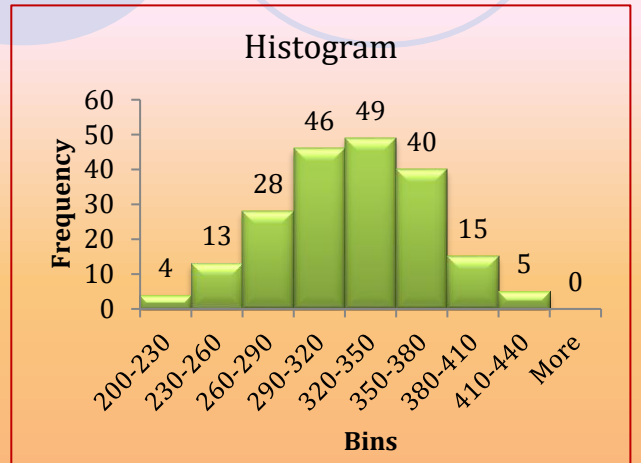
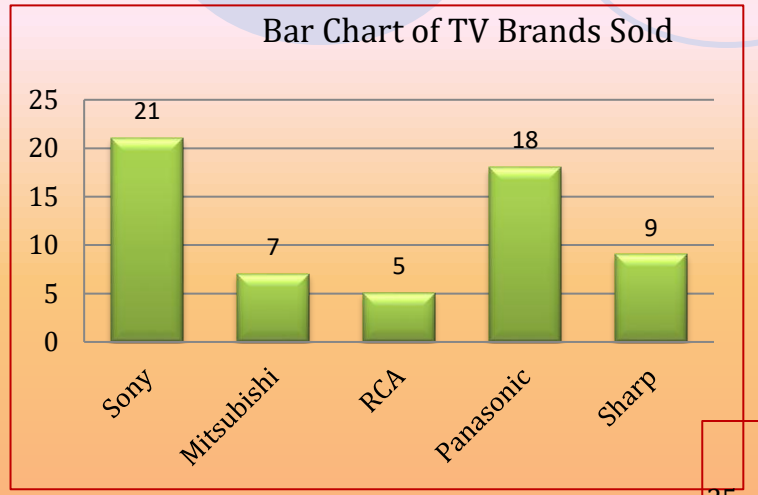
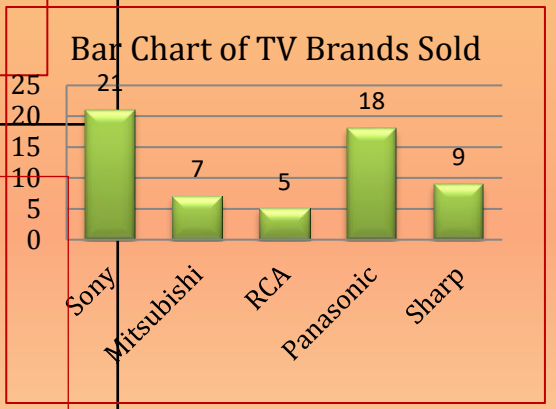
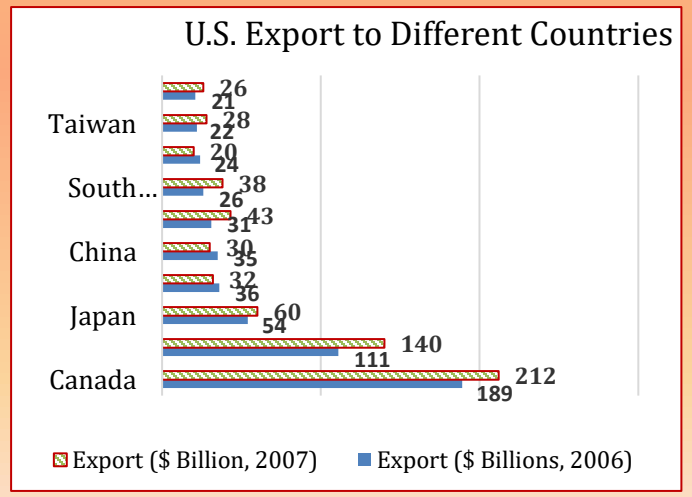


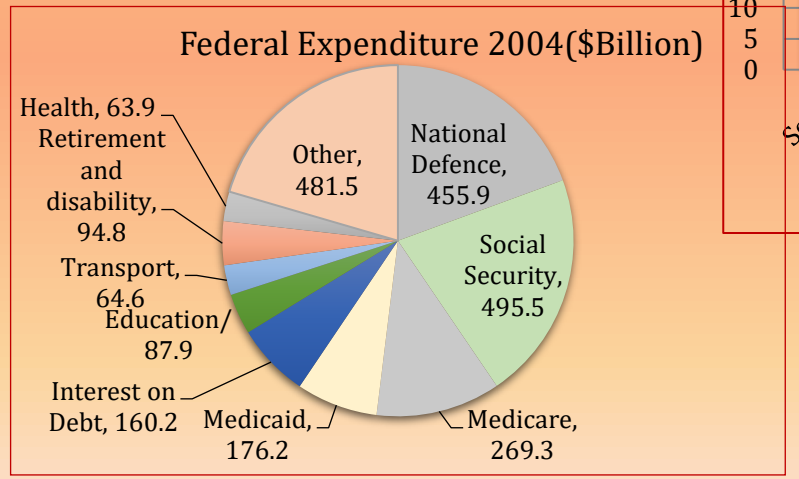
Figure 7.1: Examples of Commonly used Charts and Graphs





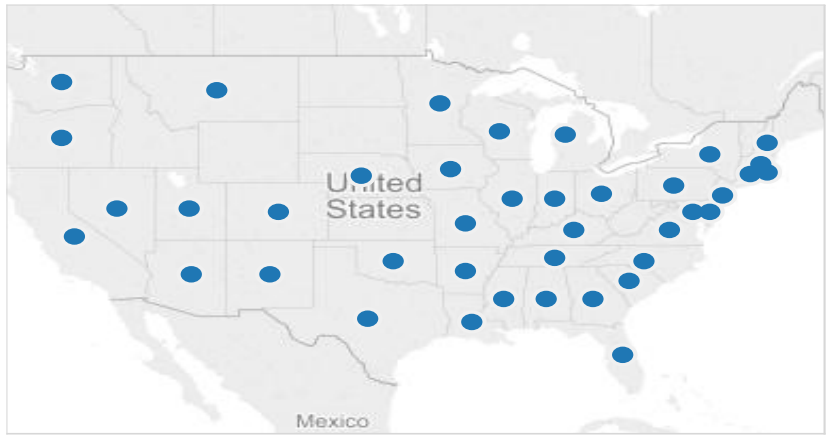




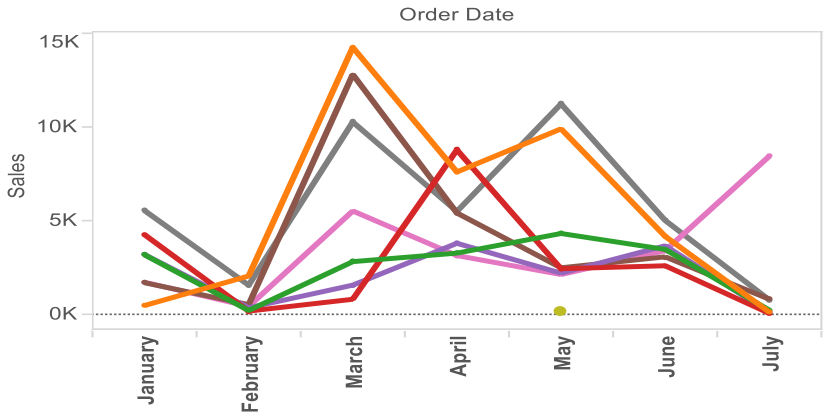


A dashboard of Online Orders Data

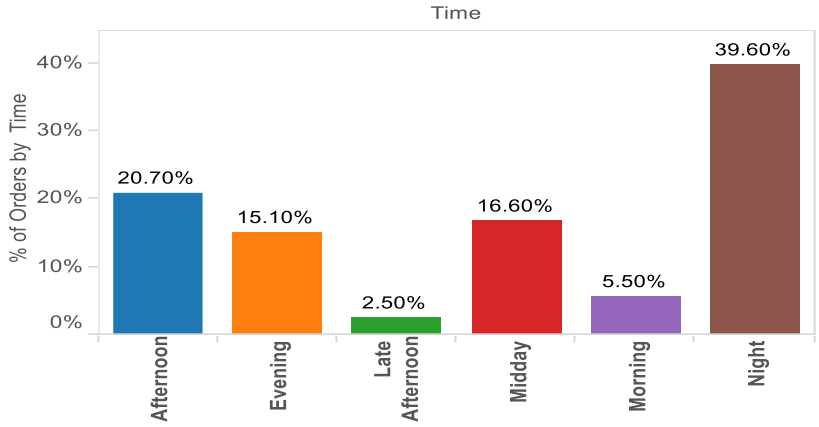
Order Map



Sales by Month



% of Orders by Time



Total Orders by Region

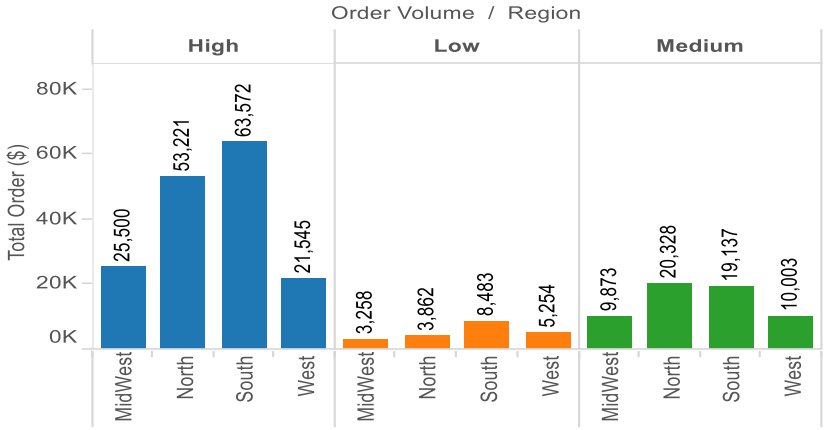


Table 7.3: Frequency distribution of 100 drivers with 60 miles per hour (mph) speed limit

Class-interval (mph)	Frequency (f)
45- 48	1
48 - 51	3
51 - 54	7
54 - 57	15
57 - 60	21
60 - 63	26
63 - 66	14
66 - 69	3
69 - 72	9
72 - 75	1
Total	$\sum f_i = 100$

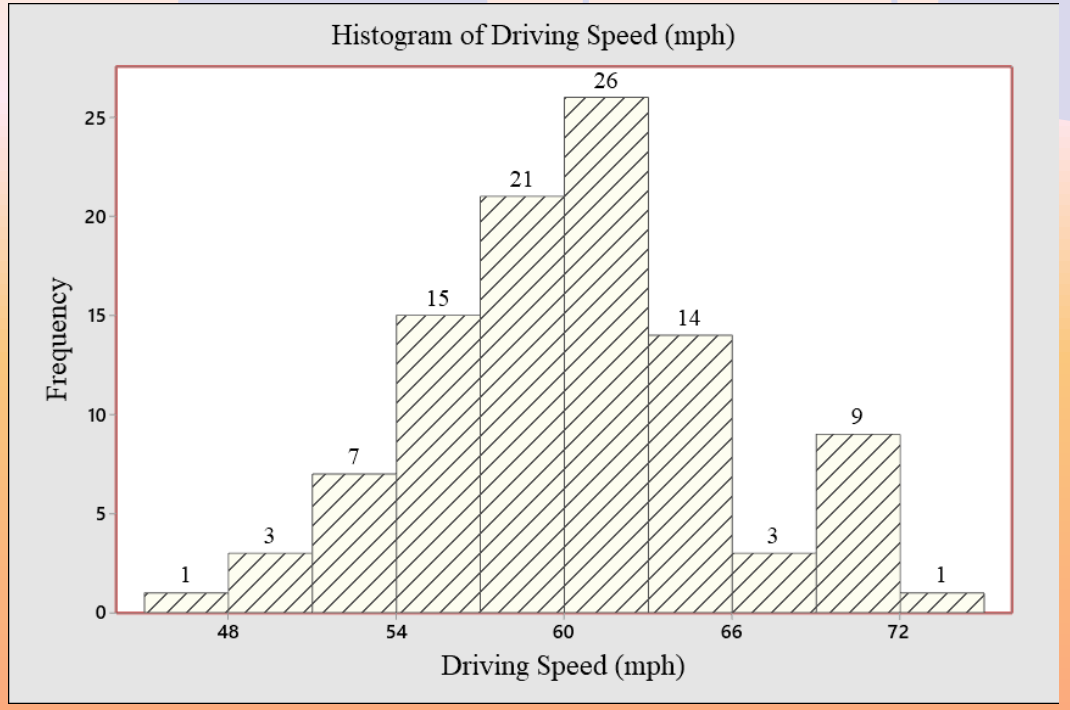


Figure 7.1(a): Histogram of Driving Speed (mph) (10 class-intervals)

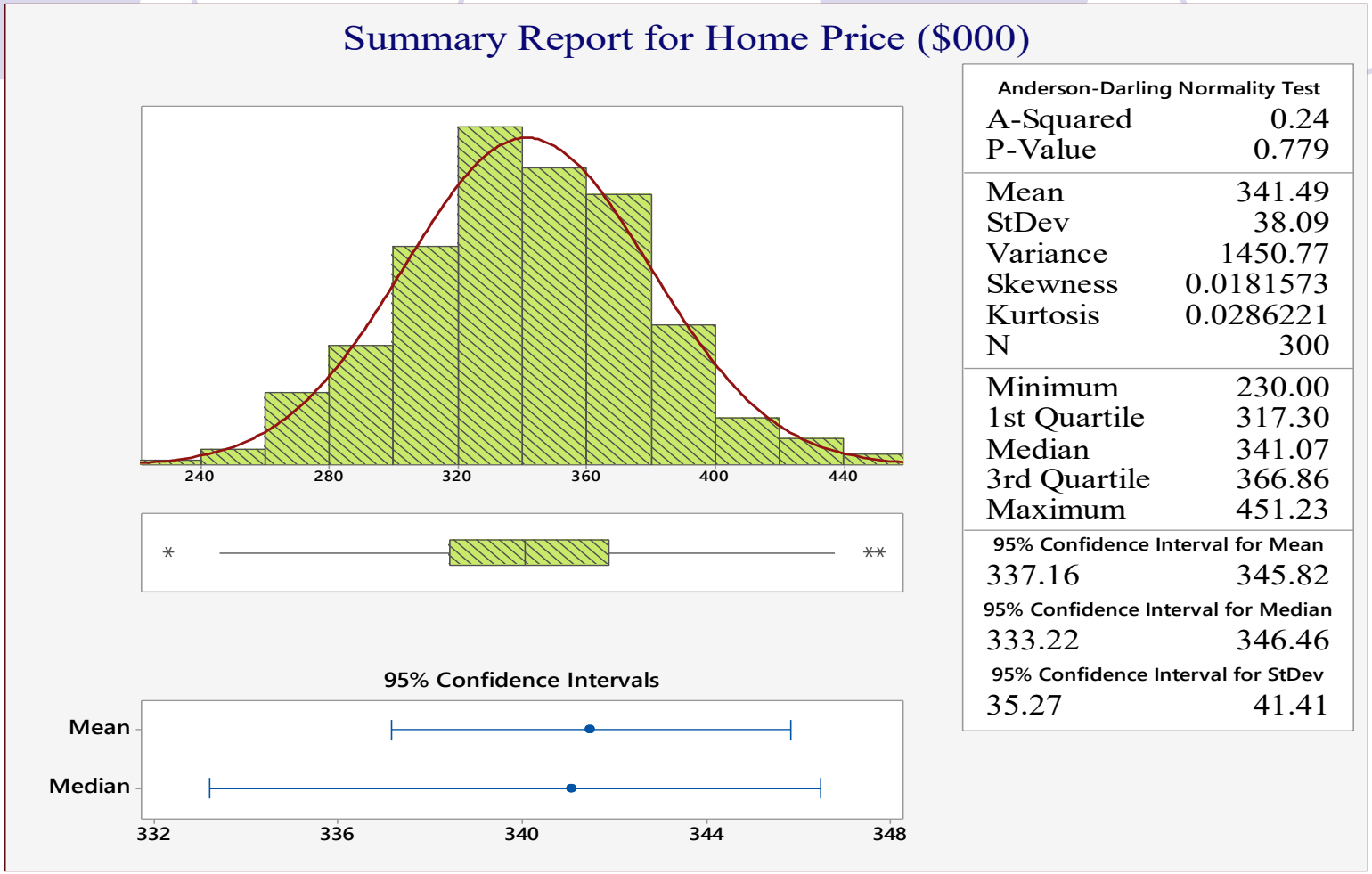


Figure 7.5: Summary Report of Home Price ($000)

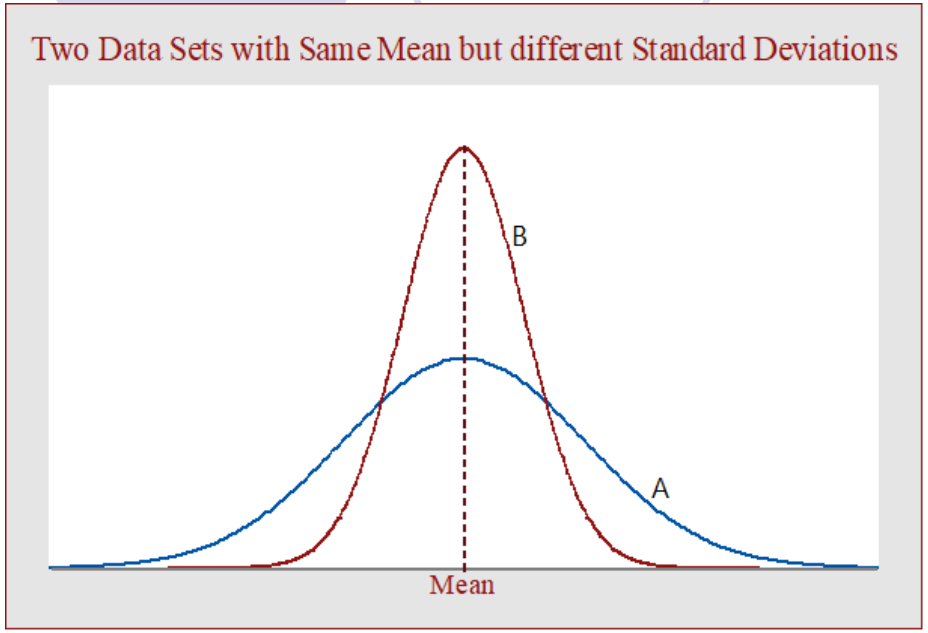


Figure 7.6: Data Sets A and B with Same Mean but Different Variations

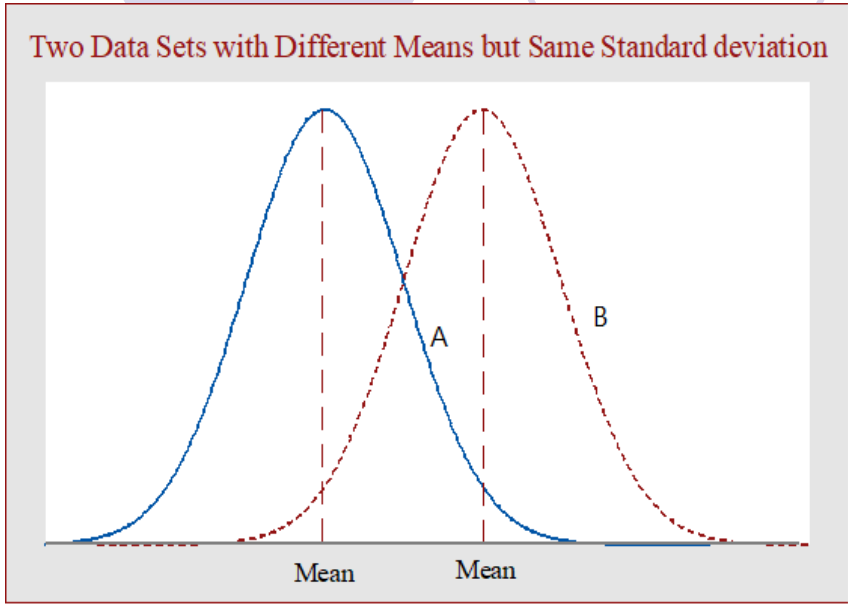


Figure 7.7: Data Sets A and B with Same Variation but Different Means

1	2	3
1	9	2
2	10	3
5	11	245
7	12	78
8	13	2
11	14	137
15	15	1229
22	16	2266778
27	17	01599
(11)	18	00013346799
17	19	03346
12	20	4679
8	21	0177
4	22	45
2	23	18

Figure 7.8: Stem-and-Leaf of Orders Received

[a] How many days were studied? **55** *(obtained by adding the numbers above and below the row median row that is, 27+11+17)*

[b] How many observations are in the fourth class? ***2***

[c] What are the smallest and largest orders? ***92, 238***

[d] List the actual values in the sixth class? ***141, 143, 147***

[e] How many days did the firm receive less than 140 orders? ***8***

[f] How many days did the firm receive 200 or more orders? ***12***

[g] How many days did the firm receive 180 orders? ***3***

[h] What is the middle value? ***180***

[i] What can you say about the shape of the data? ***Left or negatively skewed***

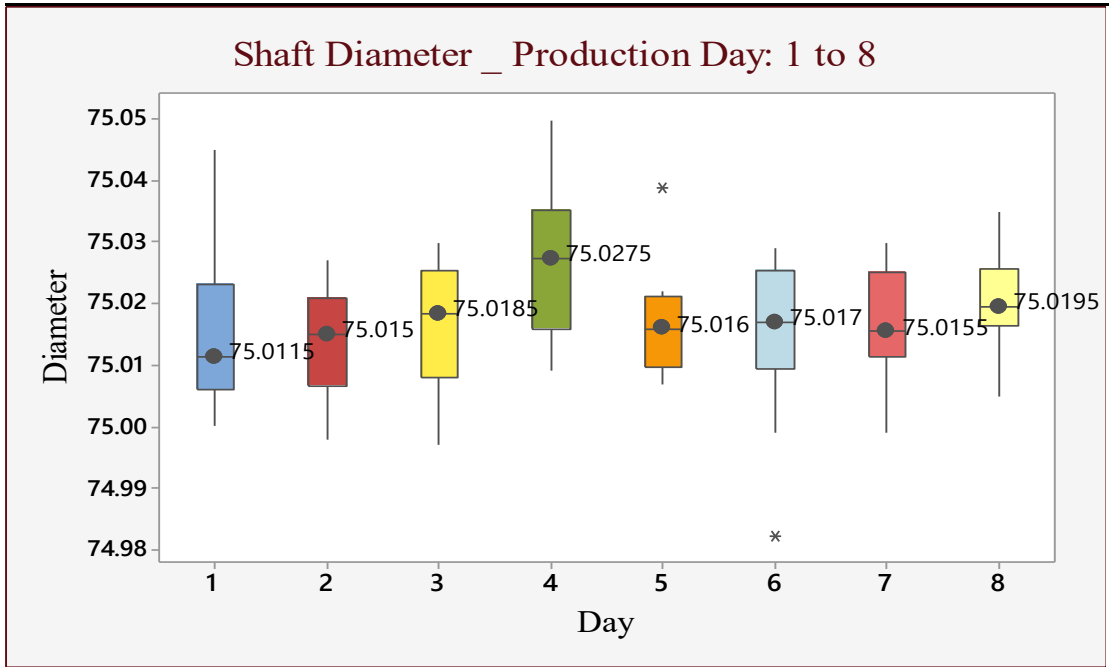


Figure 7.10: Box Plots of Shaft Diameter Over a Period of 8 Days

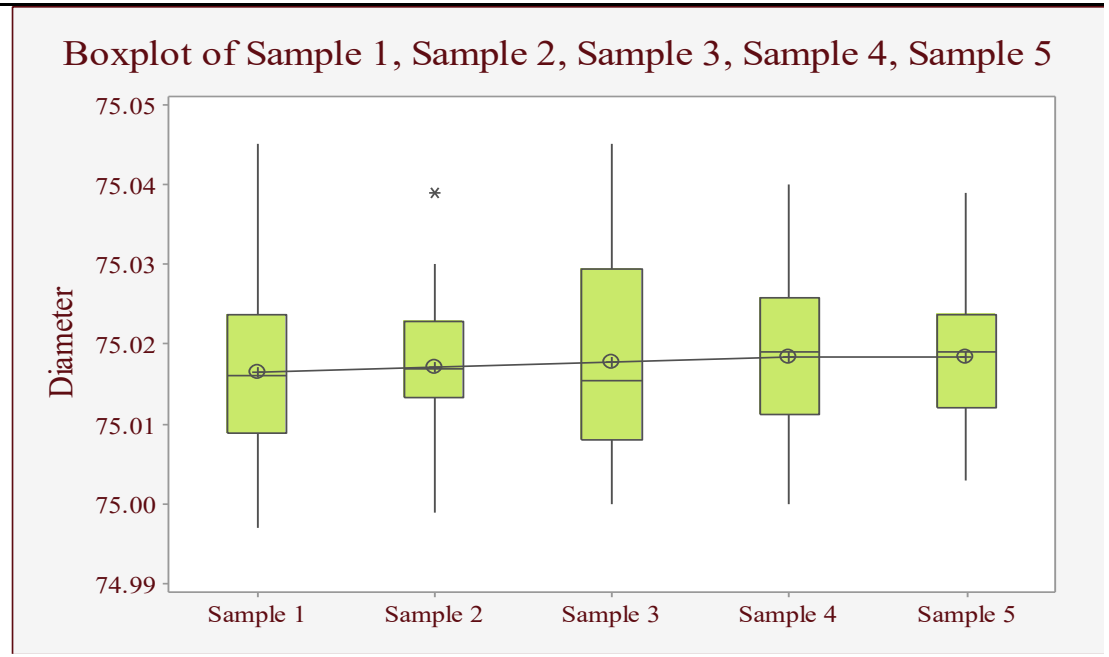


Figure 7.11: Box Plots for 5 Samples of Same Product

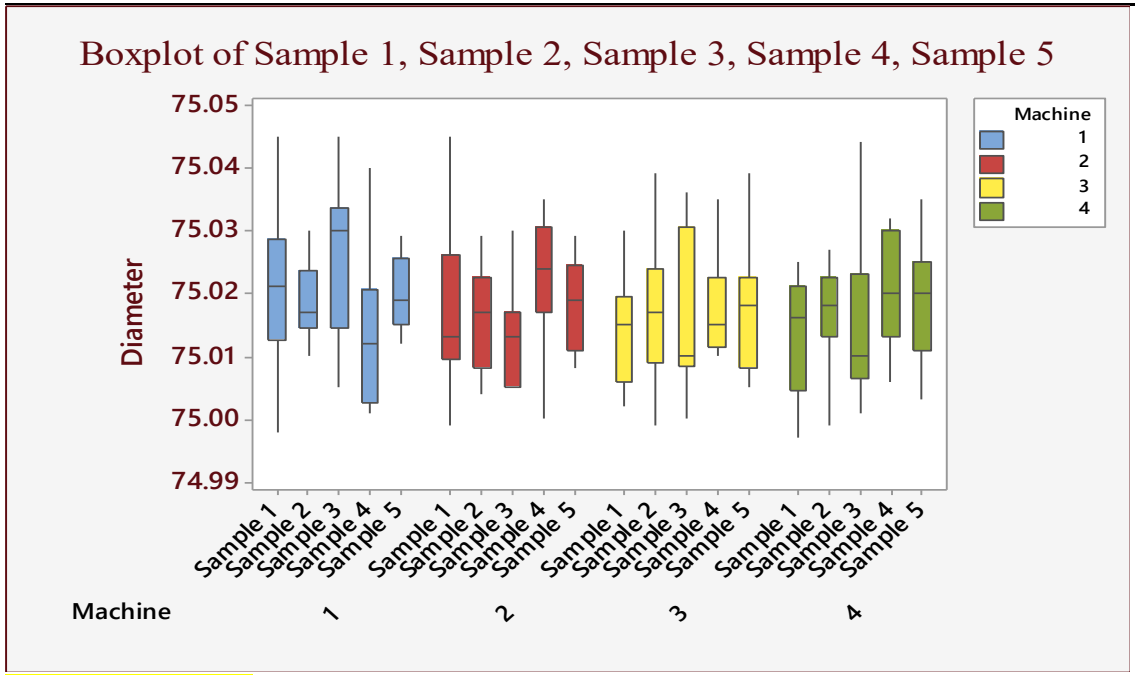


Figure 7.12: Box plots of Samples vs. Machines

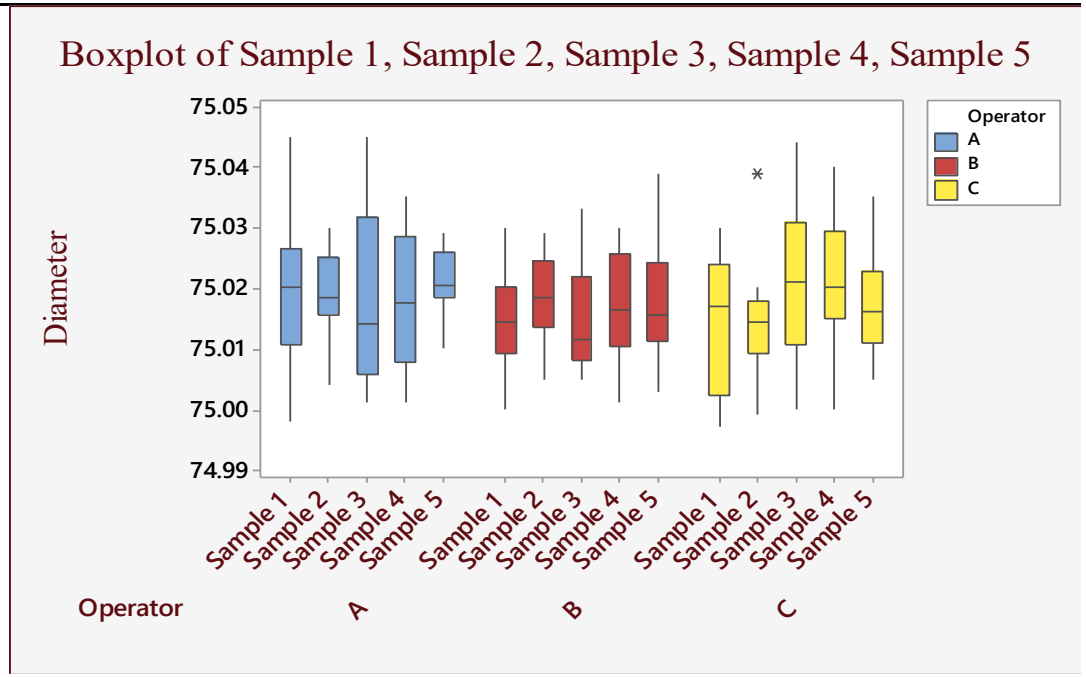


Figure 7.13: Box plots of Samples vs. Operators

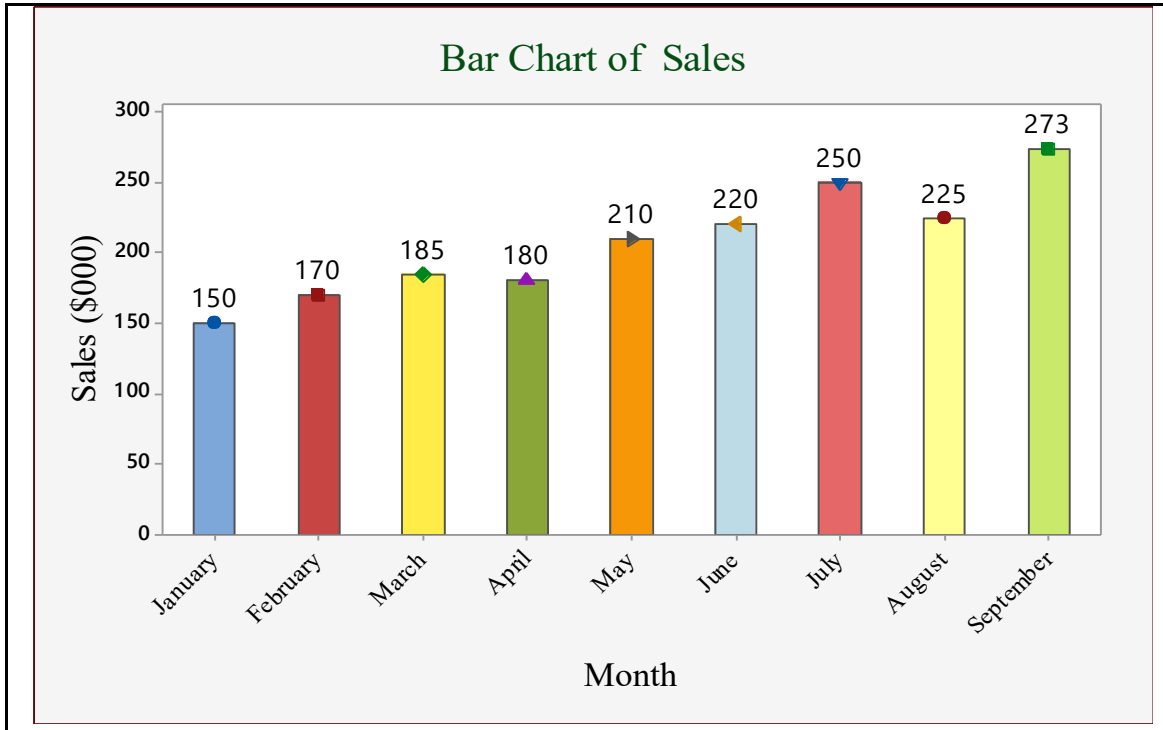


Figure 7.14: A Bar chart of Monthly Sales

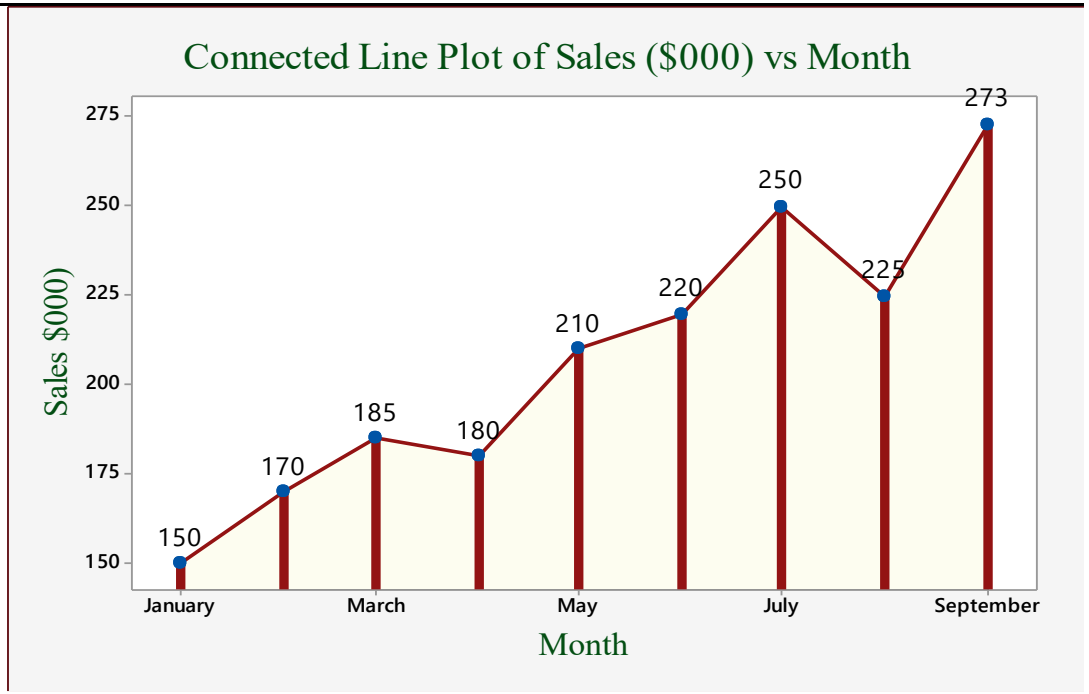


Figure 7.15: Connected Line over the Bars

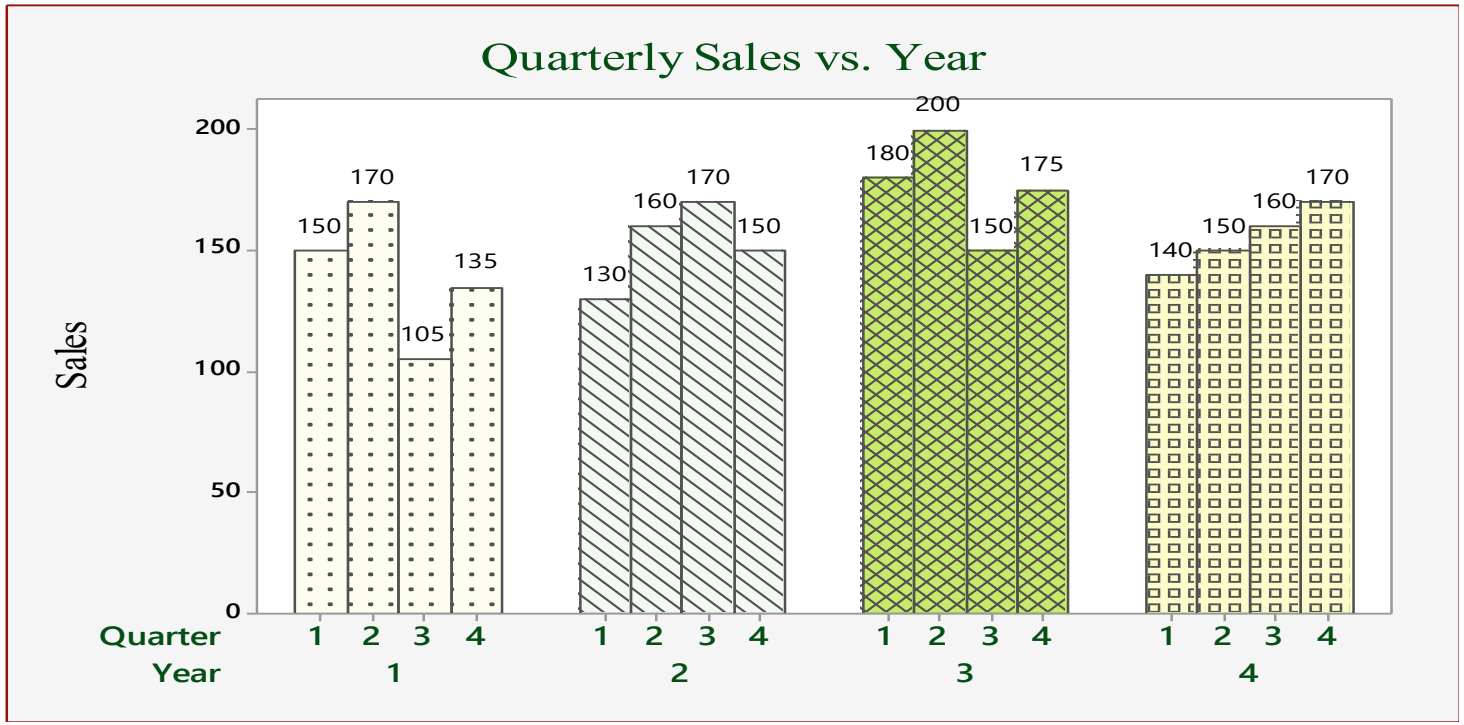


Figure 7.17: Quarterly Sales for Four Years

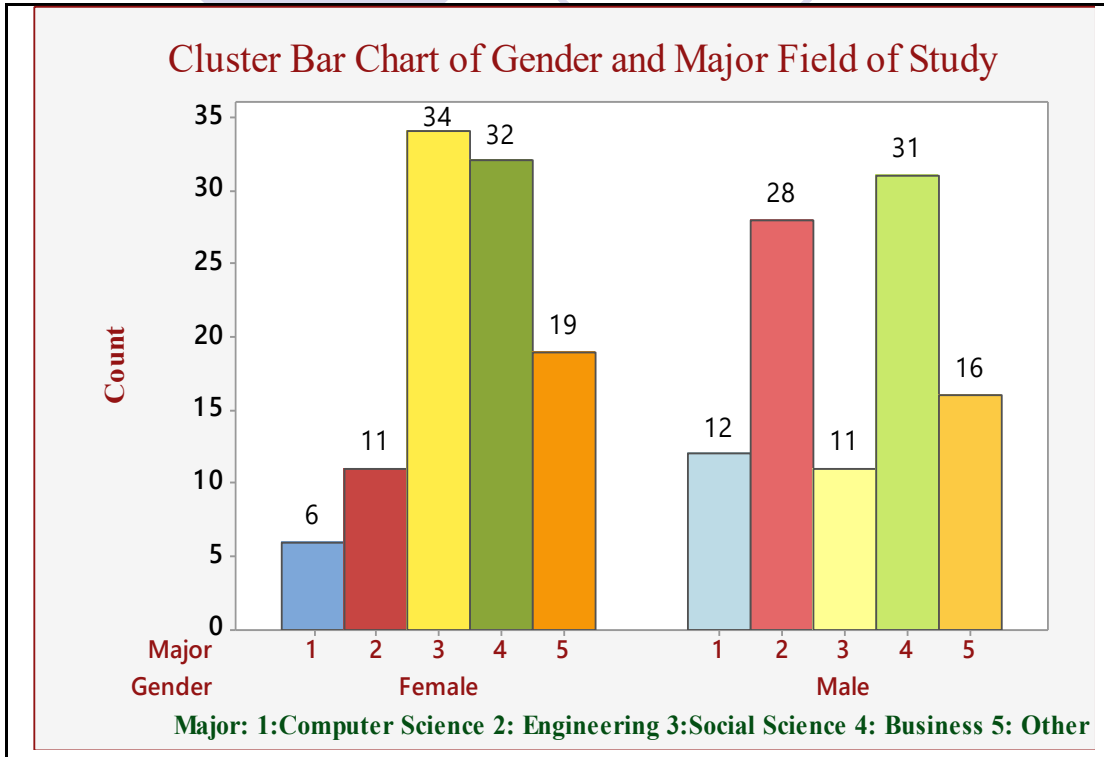


Figure 7.21: A Bar Chart of Gender and Major

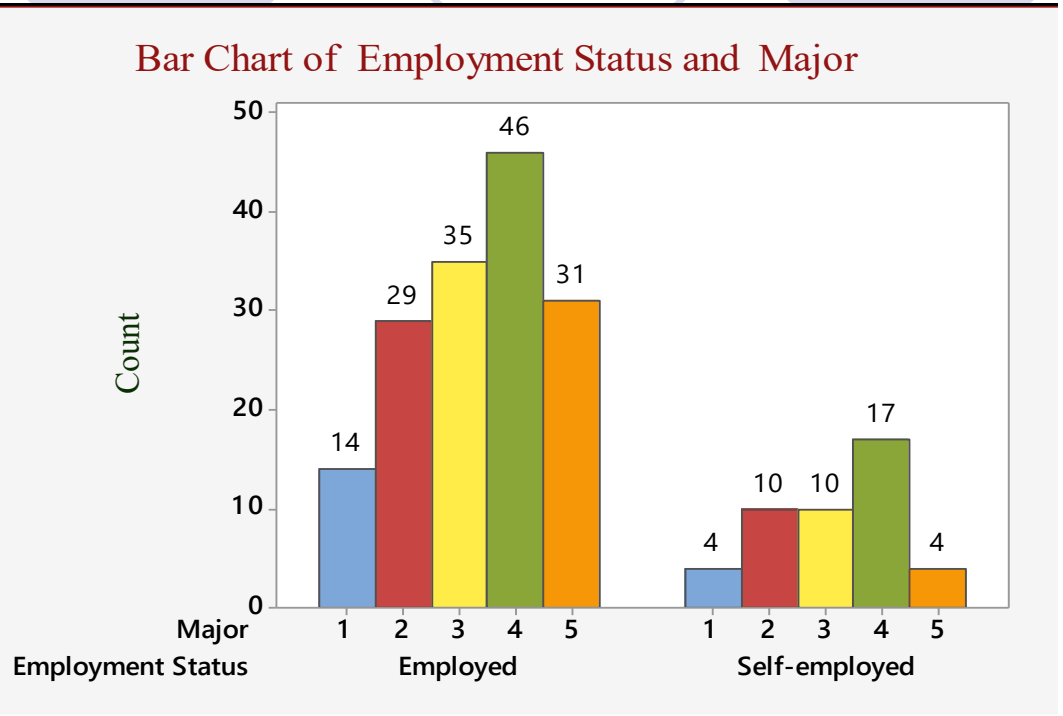


Figure 7.22: A Bar Chart of Employment Sta and Major

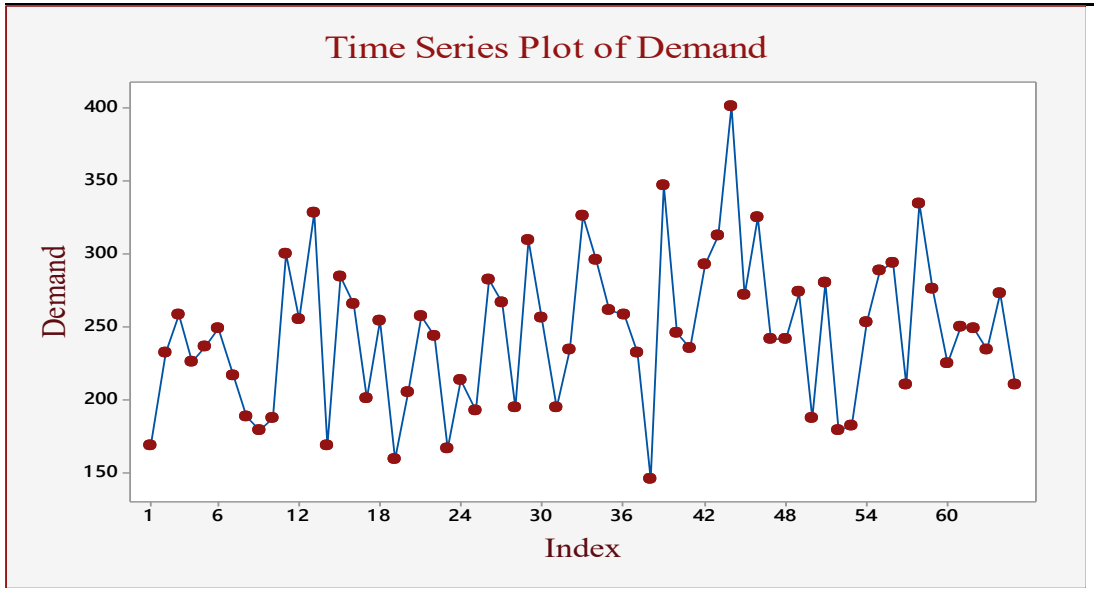


Figure 7.28: A Simple Time Series Plot of Demand

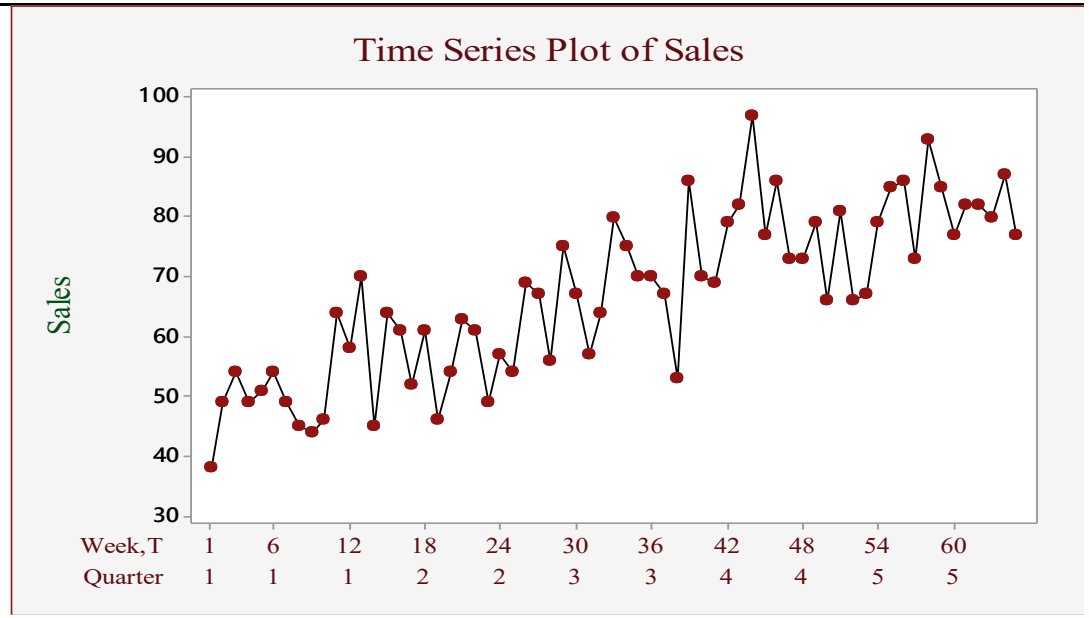


Figure 7.29: A Simple Time Series Plot of Sales

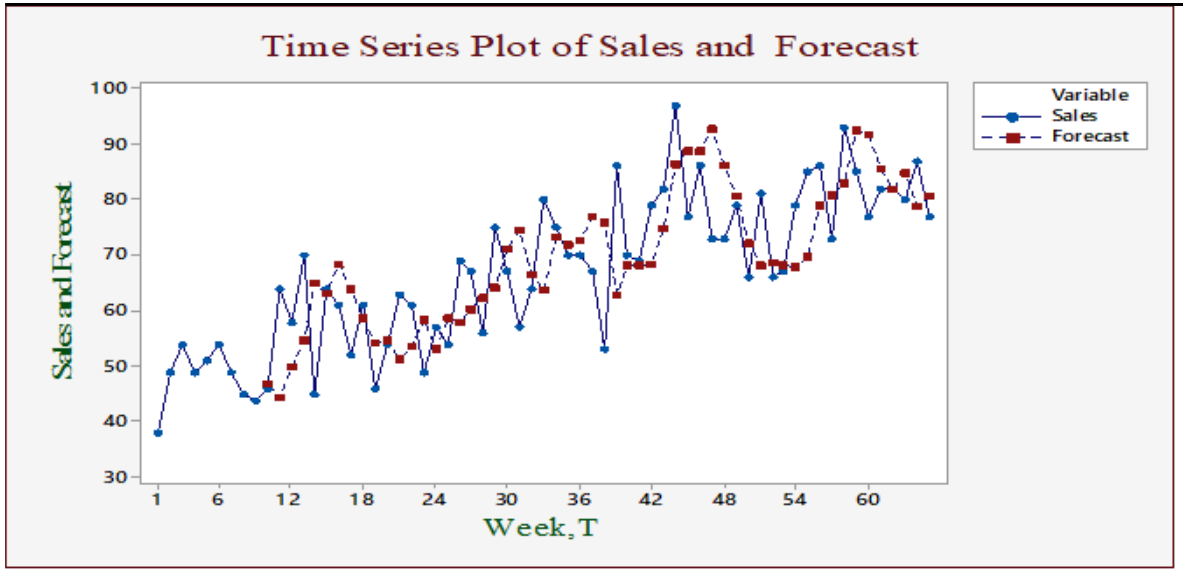


Figure 7.30: A Multiple Time Series Plot Showing Sales and Forecast

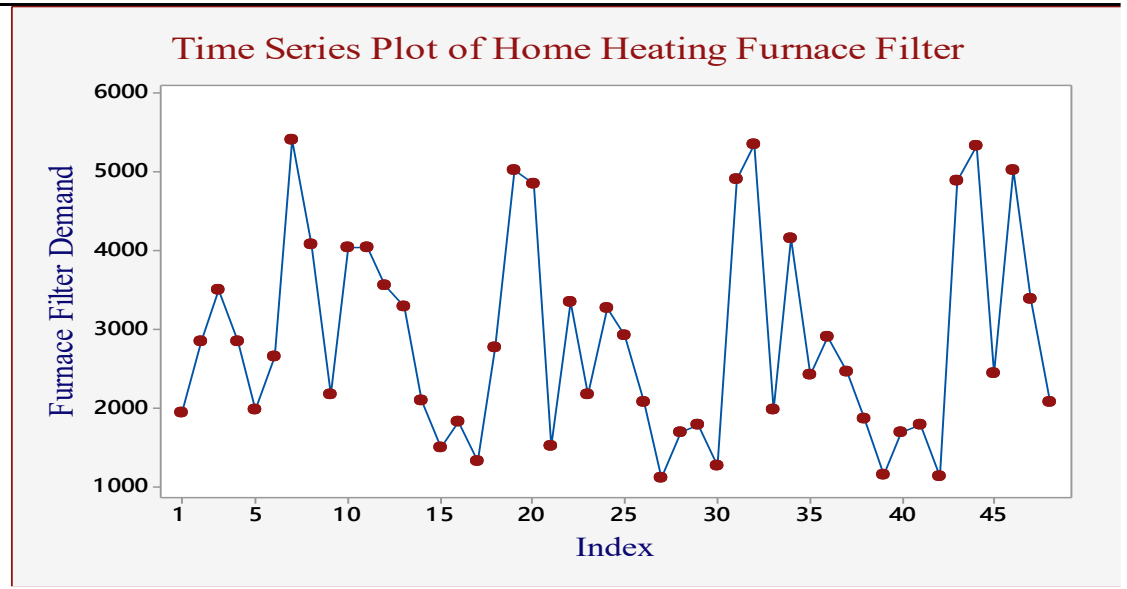


Figure 7.31: A Time Series Plot Showing Seasonal Pattern

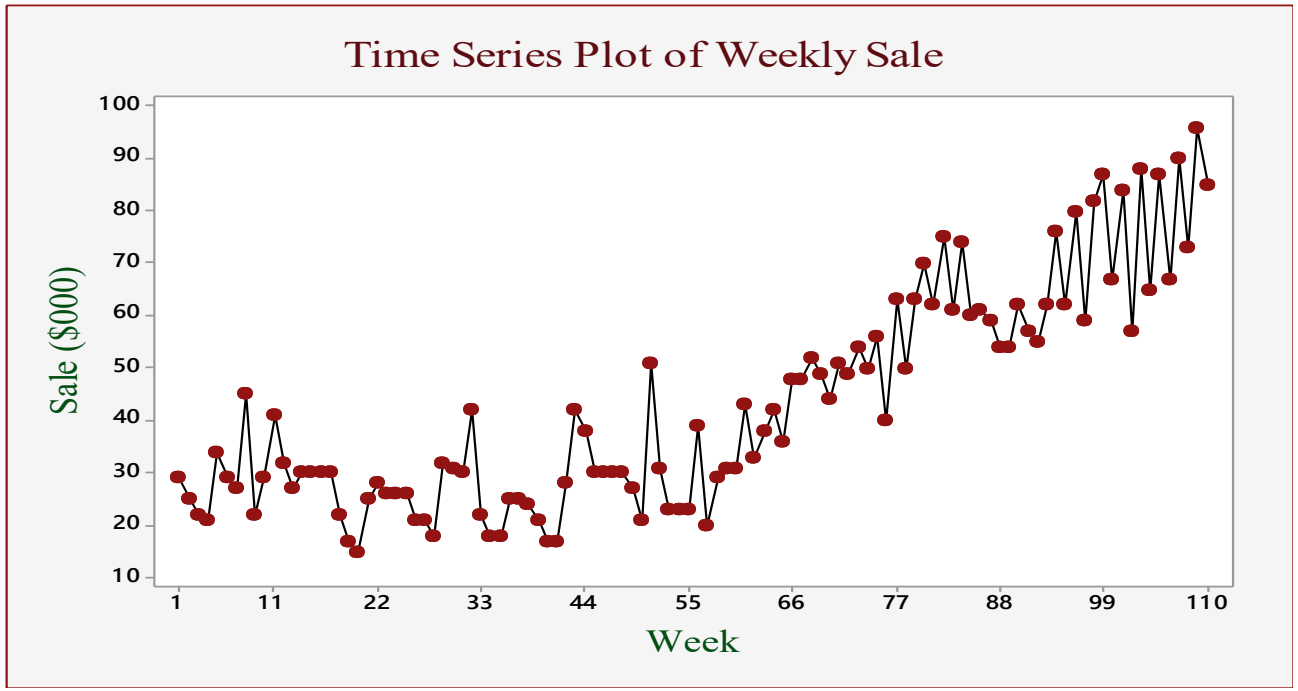


Figure 7.32: A Time Series Plot Showing a Trend

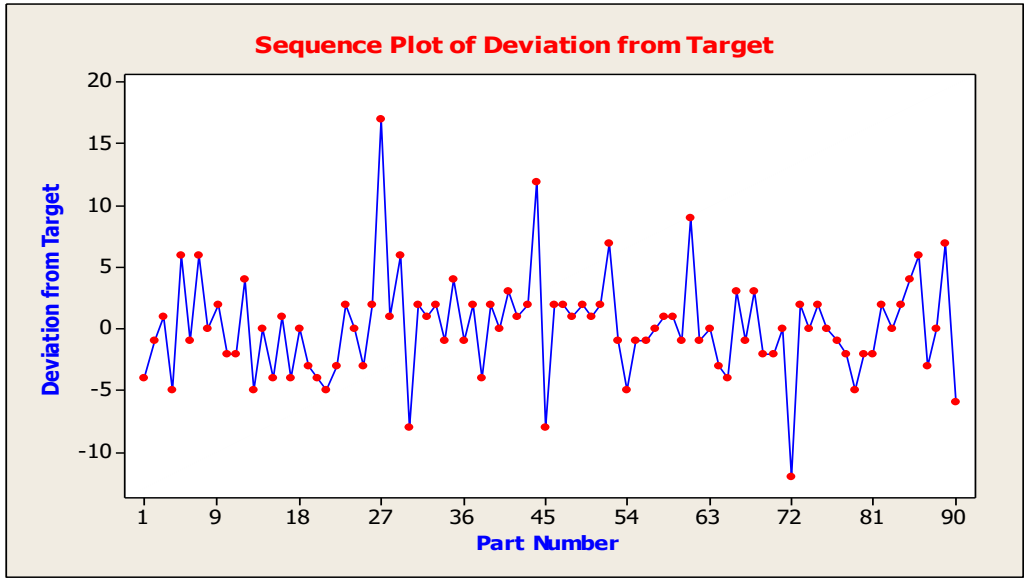


Figure 7.33: Sequence Plot of the Measurements on Machined Parts

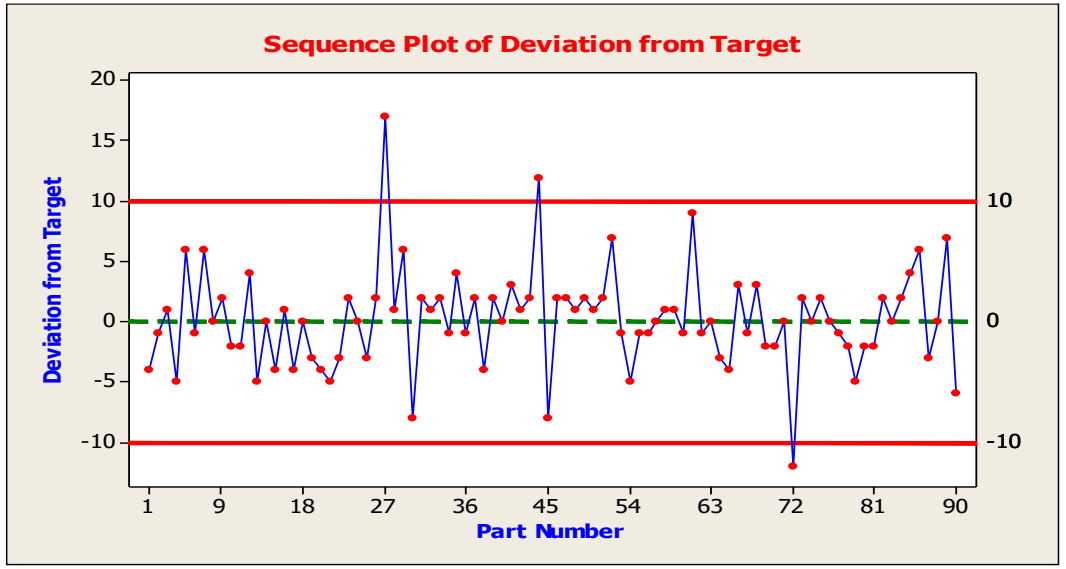


Figure 7.34: Sequence Plot with Specification Limits

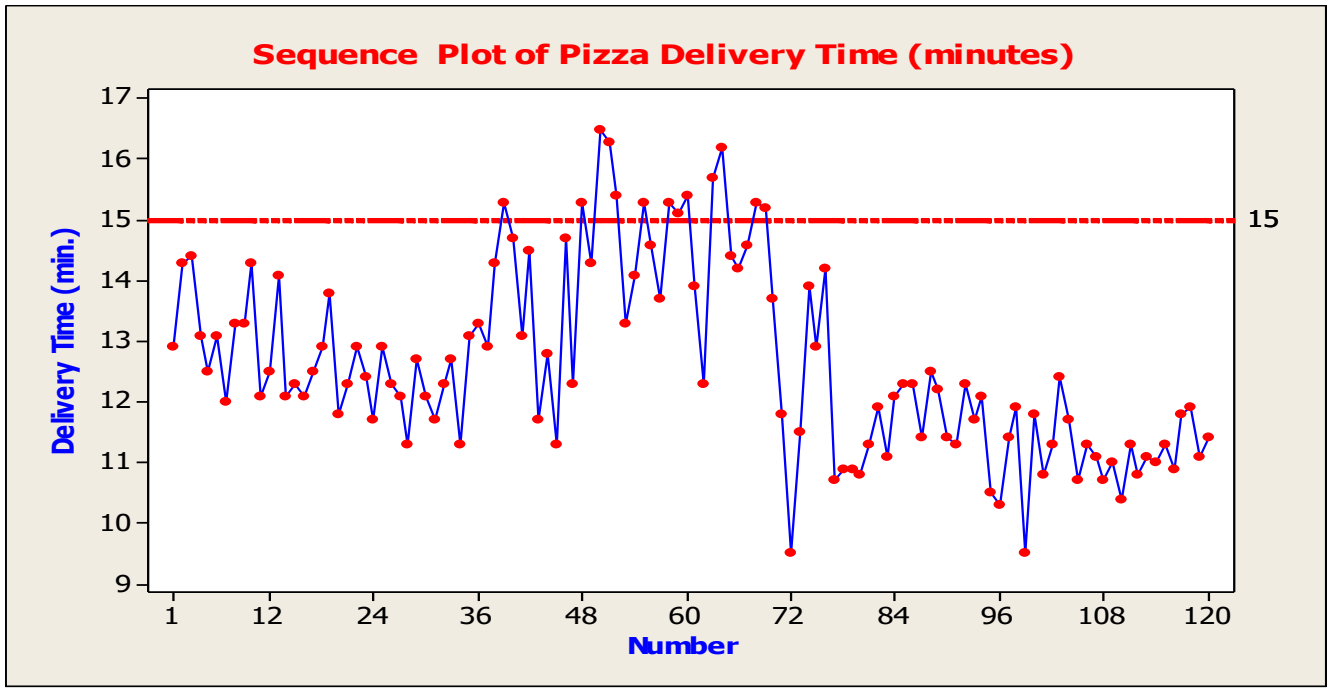


Figure 7.35: Sequence Plot of Pizza Delivery Time

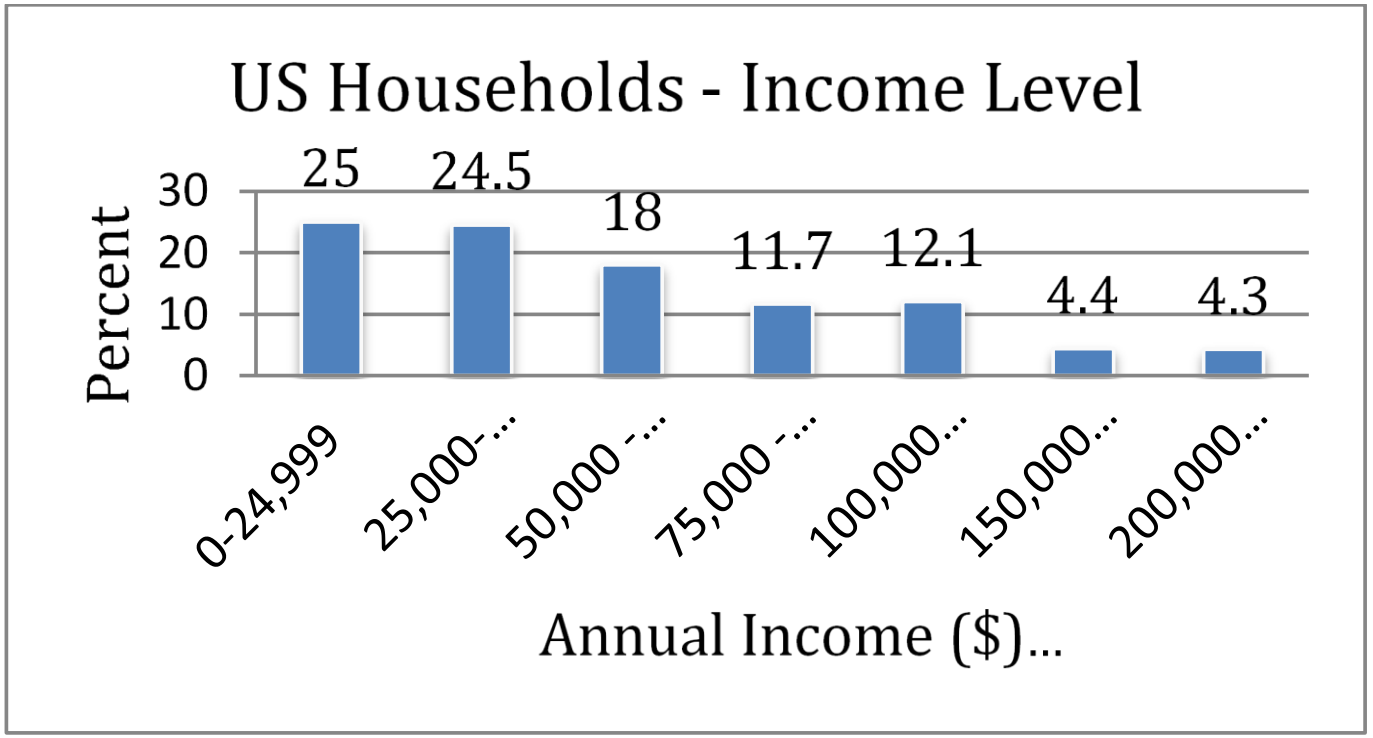


Figure 8.1: Annual Household Income of the United States for 2012

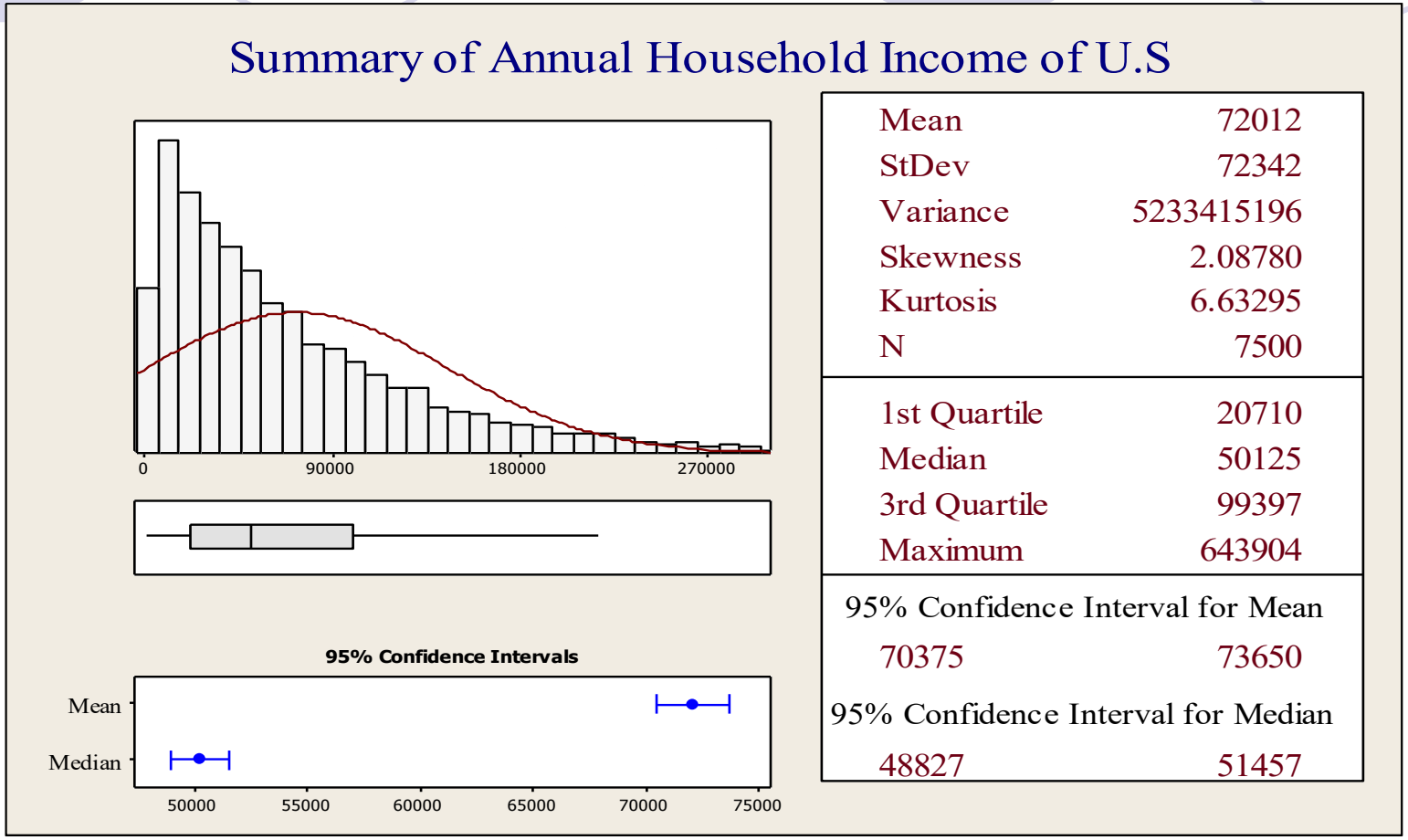


Figure 8.2: Annual Household Income and Related Statistics

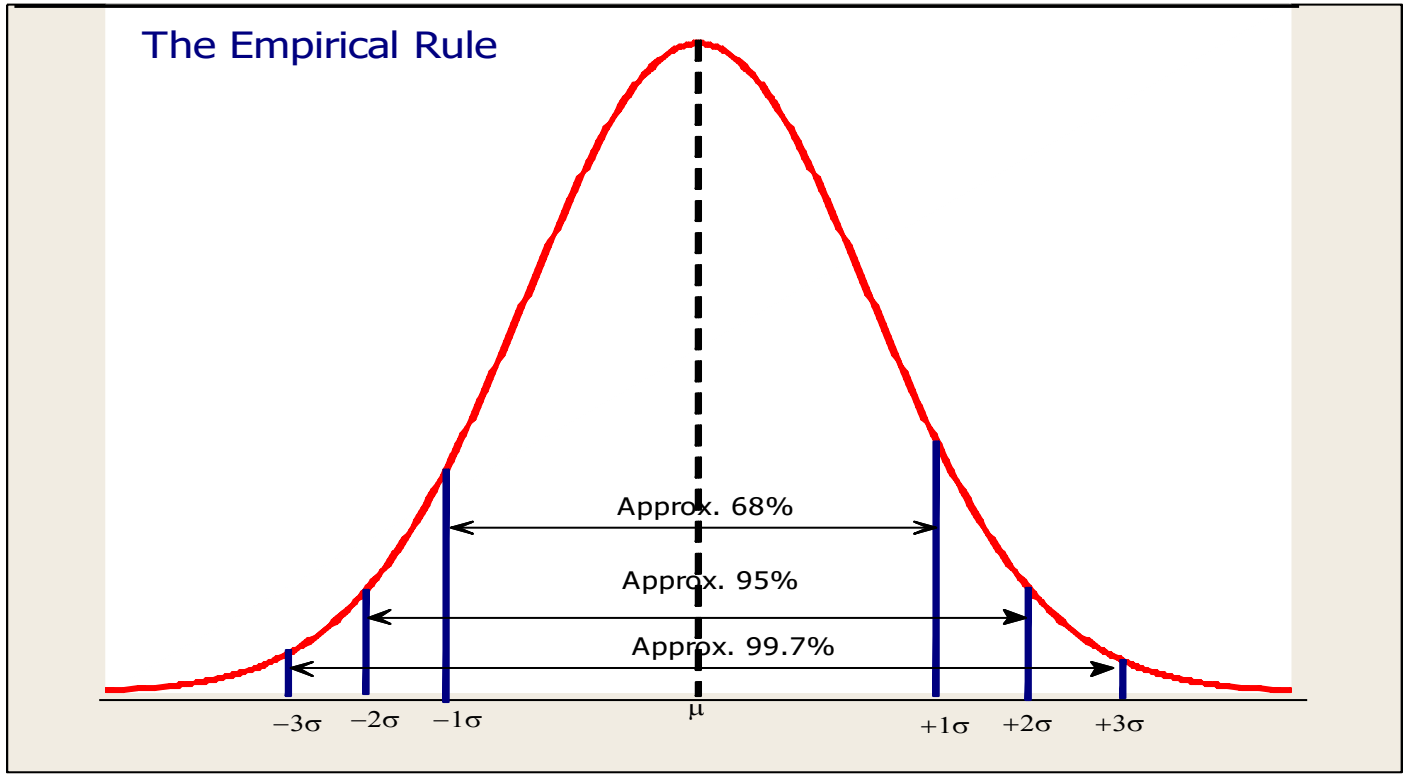


Figure 8.8: Areas under the Normal Curve

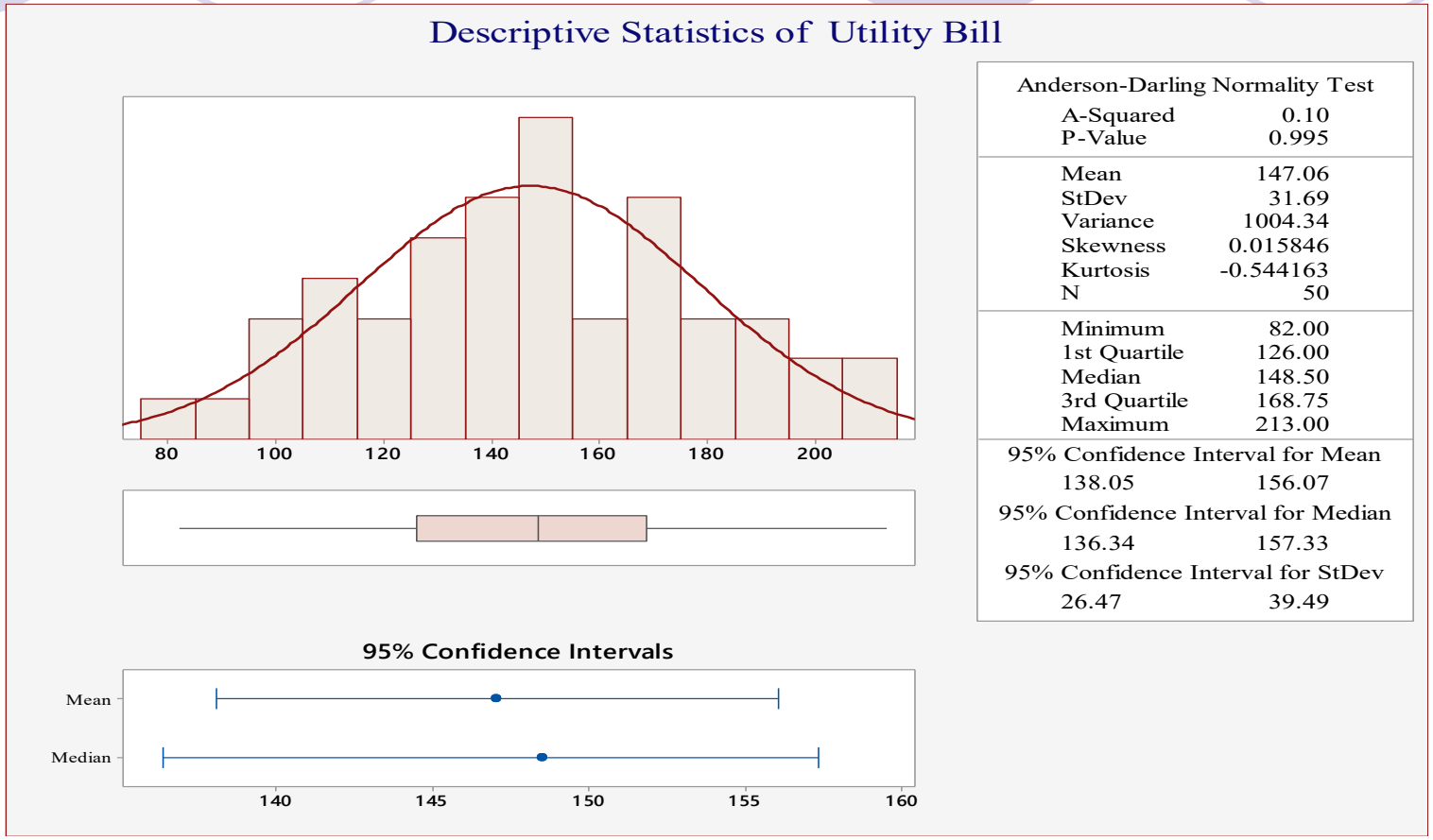


Figure 8.11: Graphical and Descriptive Summary of the Gas Bill Data

negative correlation while (d) shows a weak correlation.

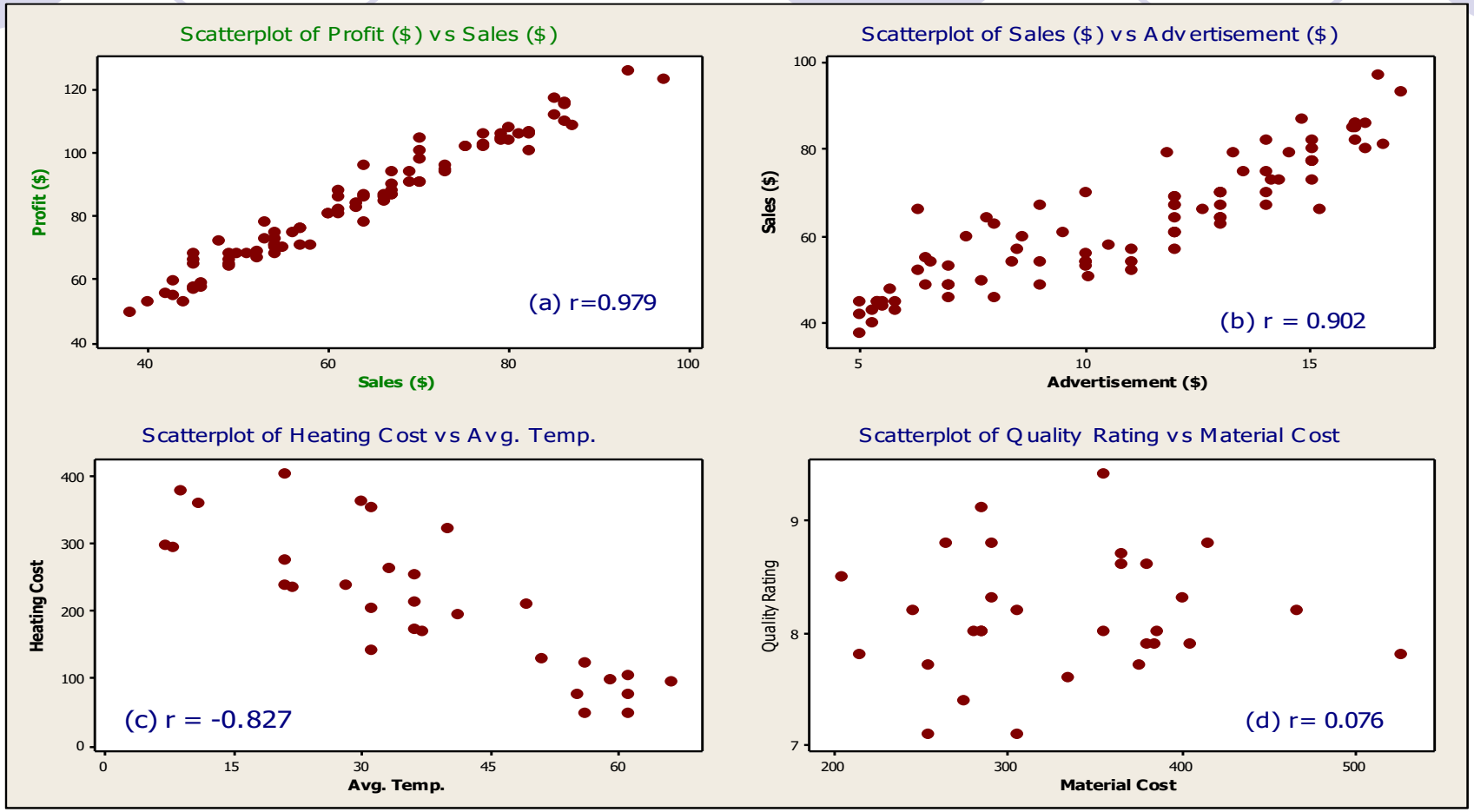
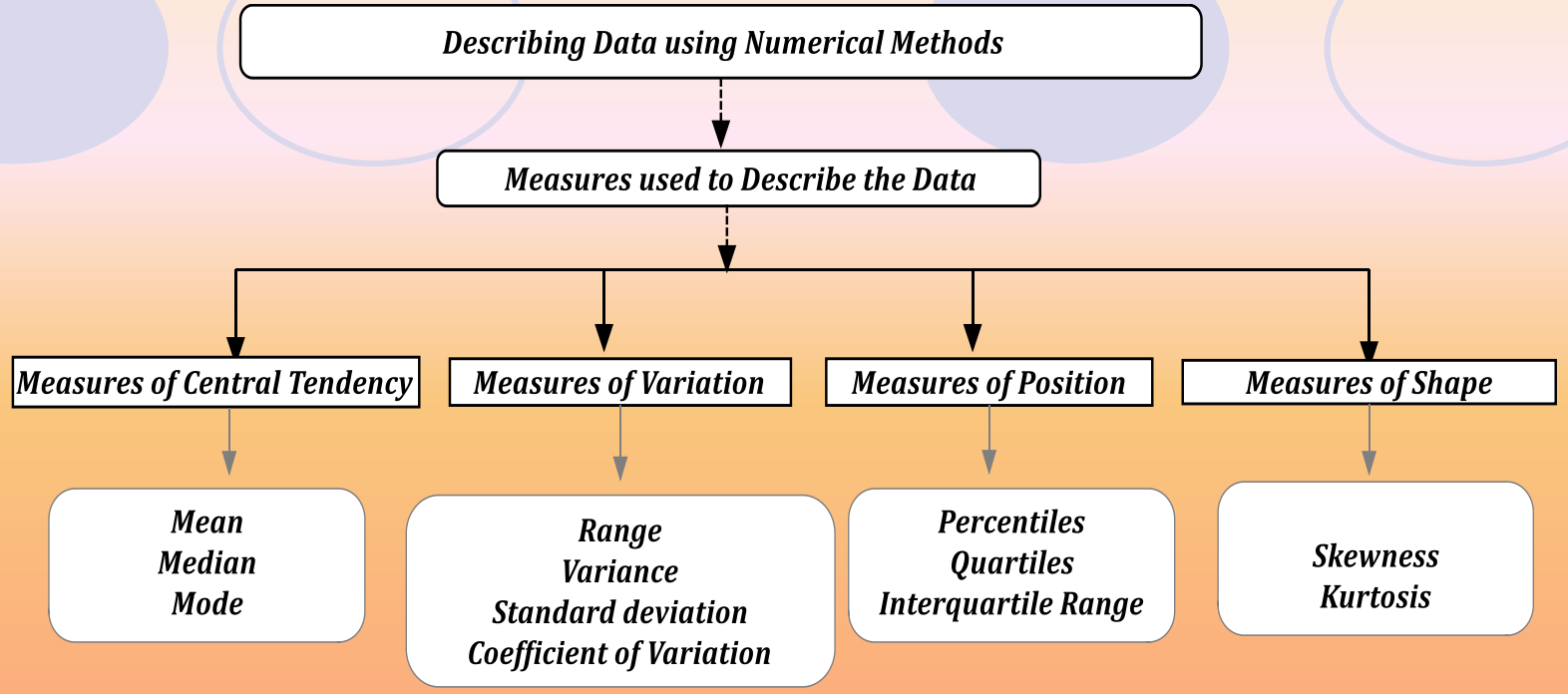


Figure 8.19: Scatterplots with Correlation (r)

Numerical Methods for Data Science



Chapter 3 : Different Measures of Describing Data - Flow Chart (1)

Descriptive Statistics : Numerical Methods

Measures of Central Tendency

Sample mean : $\bar{x} = \frac{\sum x_i}{n}$ *Population Mean:* $\mu = \frac{\sum x_i}{N}$

Median:

- Median is the middle value after the values have been arranged in ascending (or descending) order of magnitude.
- There is a distinct median when the number of observations is odd.
- Median is not affected by extreme values

When the number of observations is even

- there are two middle values and the median is obtained by taking the arithmetic mean of the middle terms

Mode:

Mode is the value that occurs most frequently in a set of observations.

Measures of Position

Location of Any Percentile

- Arrange the data in increasing order
- Find the location of the percentile using the following formula:

$$L_p = (n + 1)\frac{P}{100}$$

L_p = location of the percentile

n= total number of observations

P= desired percentile

QUARTILES: The ***quartiles*** divide the data into four parts. For a large data set, it is often desirable to divide the data into four parts. This can be done by calculating the quartiles. The quartiles are defined as

Q_1 = 1st quartile or the 25th percentile

Q_2 = 2nd quartile or the 50th percentile (median)

Q3 = 3rd quartile or the 75th percentile

Chapter 3: Measures of Central Tendency and Measures of Location- Flow Chart (2)

Descriptive Statistics: Numerical Methods...continued

Measures of Variation/Dispersion

Measures of variation or dispersion

(1) Range (2) Interquartile range (3) Variance
(4) Standard deviation (5) Coefficient of variation

Formulas for sample data

Range = *(largest value - smallest value)*

Interquartile range: $IQR = Q_3 - Q_1$

Sample variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad \text{or,} \quad s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

Sample standard deviation: Coefficient of Variation

$$s = \sqrt{s^2} \qquad c.v = \frac{s}{\bar{x}} * 100$$

Formulas for population data

Population variance:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Population standard deviation:

$$\sigma = \sqrt{\sigma^2}$$

Coefficient of variation (population):

$$CV = \frac{\sigma}{\mu} * 100$$

Measures of Shape

Measures of Shape: Skewness and Kurtosis

Skewness: Skewness is lack in symmetry. It is a measure of departure from symmetry. If the skewness a_3 is zero, the data are symmetrical; if greater than zero (positive), the data are positively or right skewed, and if the skewness is less than zero (negative), the data are negatively or left skewed.

$$s_k = \frac{n}{(n-1)(n-2)} \sum \left(\frac{(x_i - \bar{x})}{s} \right)^3$$

where:

x_i is the i[th] observation, $\bar{x}$ is mean of the observations

n is the number of non-missing observations, s is the standard deviation

Kurtosis :

Kurtosis is peakedness of the data. It is used to measure the height of the peak in the distribution. A laptokurtic distribution is more peaked than platykurtic (flatter distribution). Between laptokurtic and platykurtic is mesokurtic, which is the normal distribution. The kurtosis a_4 does not provide any information by itself; it must be compared to other distribution.

$$k_r = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{(x_i - \bar{x})}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

where: k_r is the kurtosis, x_i is the i[th] observation, $\bar{x}$ is mean of the observations, n is the number of non-missing observations, s is the standard deviation

Measures of Association Between Two Variables

Correlation coefficient : $$r_{xy} = \frac{\sum x_i y_i - (\sum x_i \sum y_i)/n}{\sqrt{\sum x_i^2 - (\sum x_i)^2/n} * \sqrt{\sum y_i^2 - (\sum y_i)^2/n}}$$

Sample covariance: $$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Chapter 3- Measures of Variation, Measures of Shape- Flow Chart (3)

Descriptive Statistics: Numerical Methods...continued

Relationship between Mean and Standard Deviation

Chebyshev's Theorem

This theorem states that ***no matter what the shape of the distribution*** (symmetrical or skewed),

at least 75% of the observations will fall within ± 2 standard deviation of the mean
at least 89% of the observation will fall within ± 3 standard deviation of the mean
at least 94% of the observations will fall within ±4 standard deviation of the mean

Within k standard deviation of the mean at least $(1-\frac{1}{k^2})$ ***percent of the values occur. Where, k is given by***

$$k=\frac{x-\bar{x}}{s} or, k=\frac{x-\mu}{\sigma}$$

Empirical Rule

The empirical rule applies to ***symmetrical or bell shaped data***. Unlike the Chebyshev's theorem, that applies to any shape (skewed or symmetrical), the empirical rule applies to symmetrical shapes. This rule states that if the data are symmetrical:

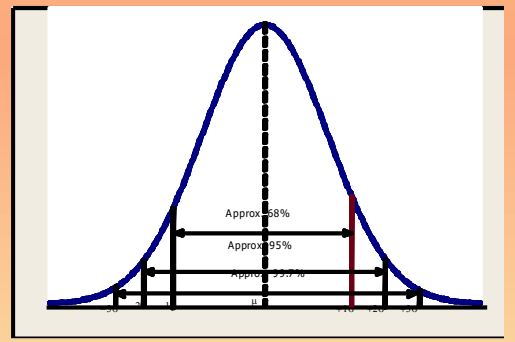
- Approximately 68% of the observations will lie within the mean and ± one standard deviation
- Approximately 95% of the observations will lie within the mean and ± two standard deviation
- Approximately 99.7% of the observations will lie within the mean and ± three standard deviation

That is,

$\mu \pm 1\sigma$ will contain approximately 68% of the observations

$\mu \pm 2\sigma$ will contain approximately 95% of the observations

$\mu \pm 3\sigma$ will contain approximately 99.7% of the observations (See the figure below)



Chapter 3: Chebyshev's and Empirical Rule- Flow Chart (4)

Descriptive Statistics: Numerical Methods...continued

Summary Measures for Grouped Data

Measures of Central Tendency for Grouped Data

Mean for the Grouped Data

$$\bar{x} = \frac{\sum f_i M_i}{n}$$

where, f_i = the frequency of class i (i=1 is class interval 1, i=2 is class interval 2 and so on)

M_i = midpoint of class i

n= number of observations which is same as $\sum f$

The midpoint is calculated by : (Lower class limit + Upper class limit)/2

Median for the Grouped Data

The median is the middle value of the data. The sample median for the grouped data is calculated using the following formula

$$M_d = L + \left[\frac{(n+1)/2 - F}{f_m}\right] w$$

Where, M_d = median
L = lower limit of the ***median class***
n=number of observations
F=sum of the frequencies up to but not including the *median class*
f_m = frequency of the *median class*
w = width of the class

Mode for the Grouped Data: In a grouped data, the mode is the average of the ***modal class***. The modal class is the class with maximum frequency.

Measures of Variation for Grouped Data

Range

The range for the above grouped data is calculated using the following formula

Range = Upper limit of the last class - lower limit of the first class

Sample variance

The sample variance of the grouped data is calculated using the following formula

$$s^2 = \frac{\sum fM^2 - n\bar{x}^2}{n-1}$$

where, f = frequency
M = midpoint
n= no. of observations
$\bar{x}$=mean

Sample Standard Deviation

The standard deviation, s is calculated by taking the square root of the variance. Therefore,

$$s = \sqrt{s^2}$$

Coefficient of Variation (CV)

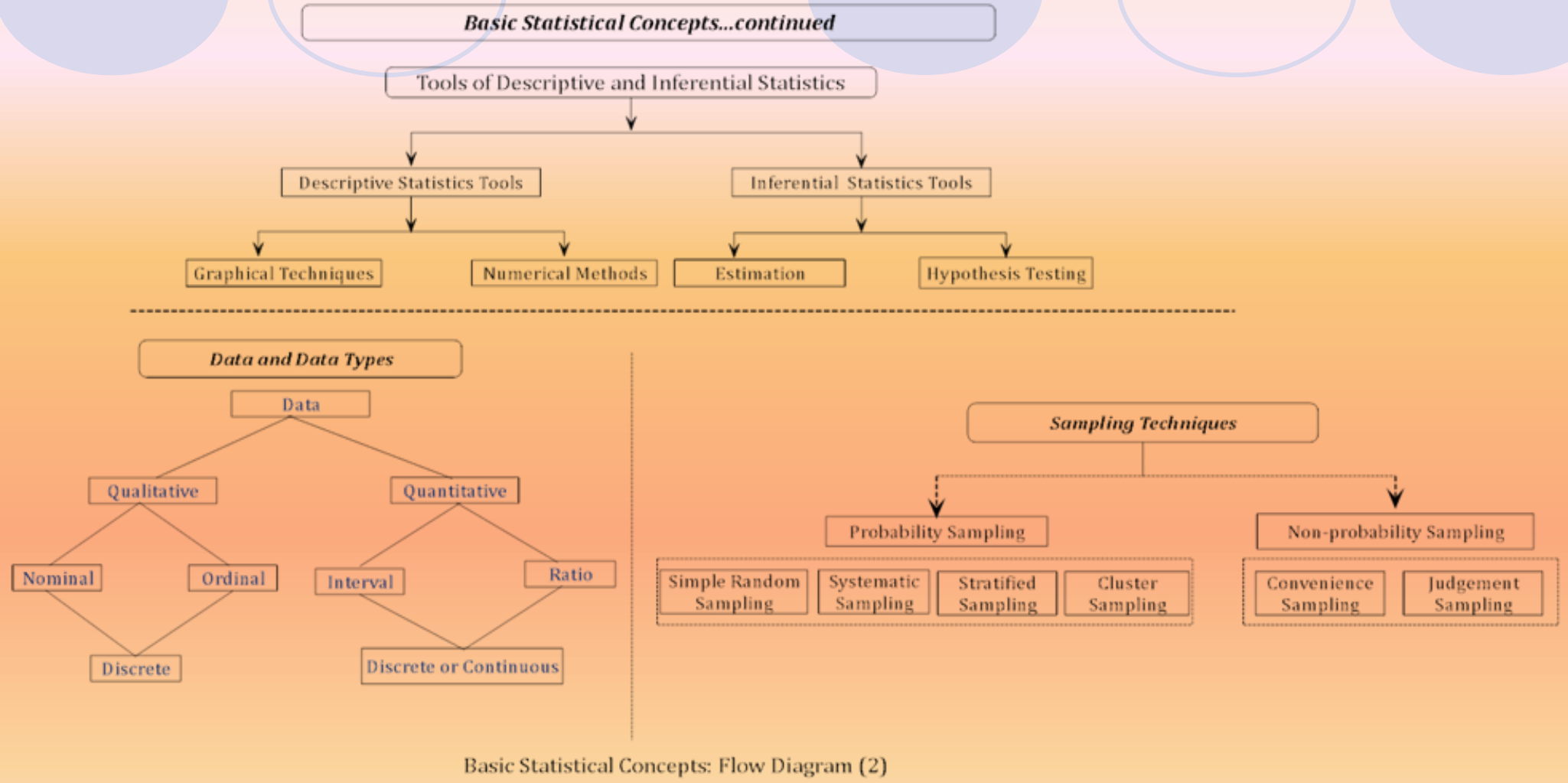
The coefficient of variation is calculated by

$$CV = \frac{s}{\bar{x}} * 100\%$$

Summarizing Data

Statistics Based on Ordered Values	Minimum, First Quartile, Median, Third Quartile, Maximum, Interquartile Range
Statistics Based on Averages	Mean, Standard Deviation, Variance, Skewness, Kurtosis
Describe a symmetrical (bell-shaped) distribution	Mean and Standard Deviation
Relating Continuous Variables	Scatterplots and Correlation

Chapter 3: Measures of Central Tendency and Variation for Grouped Data - Flow Chart (5)



Basic Statistical Concepts: Flow Diagram (2)

Probability Concepts for Data Science

Probability Theory

Probability Terms

Some Important Terms in Probability

Probability: Probability is the chance that a particular event will occur when an experiment is performed. The probability of an event A is denoted as P (A), which means "the probability that event A occurs" is between 0 and 1. That is,

$$0 \le P(A) \le 1$$

Event: An event is one or more possible outcomes of an experiment.

Experiment: An experiment is any process that produces an outcome or observation.

Sample Space: The set of all possible outcomes of an experiment is called thesample space and is denoted by S.

Mutually Exclusive Events: When the occurrence of one event excludes the possibility of another event occurring, then we say the events are mutually exclusive. In other words, only one event can take place at a time.

Exhaustive Events: The total number of possible outcomes in any trial is known as exhaustive events.

Equally Likely Events: A situation where all the events have an equal chance of occurrence or when there is no reason to expect one in preference to the other.

Counting Rules in Probability

(1) Multiple-Step Experiment or Filling Slots

Suppose an experiment can be described as a sequence of k steps in which

n1 = the number of possible outcomes on the first step

n2 = the number of possible outcomes on the second step

:

n_k = the number of possible outcomes on the k^{th} step, then the total number of possible outcomes is given by

$$(n_1)(n_2)(n_3).....(n_k)$$

(2) Permutations

The number of ways of selecting n distinct objects from a group of N objects—where the order of selection is important—is known as the number of permutations on N objects, using n at a time and is written as

$$P_n^N = \frac{N!}{(N-n)!} = (n)(n-1).....(n-k+1)$$

(3) Combinations

Combination is selecting n objects from a total of N objects. The order of selection is not important in combination. This disregard of arrangement makes the combination different from the permutation. In general, an experiment will have more permutations than combinations.

The number of combinations of N objects taken n at a time is given by

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!}$$

Note $0! = 1$

Chapter 4: Probability Theory - Flow Diagram (1)

Probability Theory...continued

Ways of Assigning Probabilities

Assigning Probabilities: There are two basic rules for probability assignment:

The probability of an event A is written as P(A) and it must be between 0 and 1. That is,

$$0 \le P(A) \le 1.0$$

If an experiment results in n number of outcomes $A_1, A_2 \ldots A_n$; then the sum of the probabilities for all the experimental outcomes must equal 1. That is,

$$P(A_1) + P(A_2) + P(A_3) + \ldots\ldots + P(A_n) = 1$$

Methods of Calculating Probabilities: There are three methods for assigning probabilities

1. **Classical Method**
2. **Relative Frequency Approach**
3. **Subjective Approach**

Different Ways of Calculating Probability

1. Classical Method

The classical approach of probability is defined as the favorable number of outcomes divided by the total number of possible outcomes. Suppose an experiment has n number of possible outcomes and the event A occurs in m of the n outcomes, then the probability that event A will occur is

$$P(A) = \frac{m}{n}$$

Note that P(A) denotes the probability of occurrence for event A. The probability that the event A will not occur is given by $P(\overline{A})$, which is read as P (not A) or 'A complement.' Thus,

$$P(A) + P(\overline{A}) = 1$$

which means that the probability that event A will occur, plus the probability that event A will not occur, must be equal to 1.

---

2. Relative Frequency Approach

Probabilities are also calculated using the relative frequency. In many problems, we define probability by relative frequency.

---

3. Subjective Probability

Subjective probability is used when the events occur only once or very few times and when little or no relevant data are available. In assigning subjective probability, we may use any information available, such as our experience, intuition, or expert opinion.

Probabilities for Mutually and Non-mutually Exclusive Events

Addition Law for Mutually Exclusive Events

If we have two events A and B, that are mutually exclusive, then the probability that A or B will occur is given by

$$P(A \cup B) = P(A)$$

Note that the "union" sign is used for "or" probability; that is, $P(A \cup B)$. This is same as P (A or B). This rule can be extended to three or more mutually exclusive events. If three events A, B, and C are mutually exclusive then the probability that A or B or C will occur

$$P(A \cup B) = P(A) + P(B) + P(C)$$

Addition Law for Non-Mutually Exclusive Events

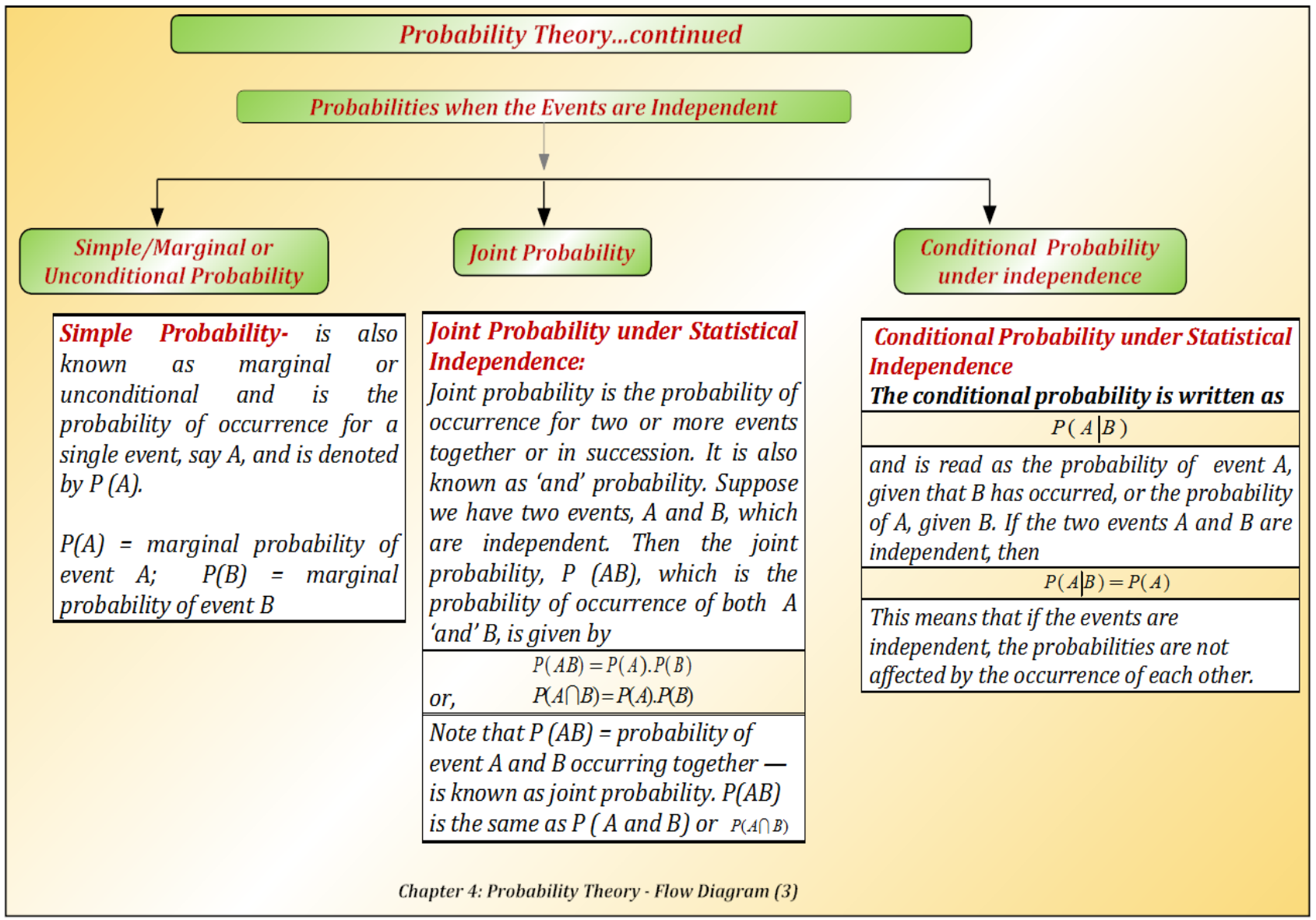
The occurrence of two events that are **non-mutually exclusive** means that they can occur together. If the events A and B are non-mutually exclusive, the probability that A or B will occur is given by

$$P(A \cup B) = P(A) + P(B) - P(A \text{ and } B)$$
$$\text{or, } P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If events A, B, and C are non-mutually exclusive, then the probability that A or B or C will occur:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \text{ and } B) - P(A \text{ and } C) - P(B \text{ and } C) + P(A \text{ and } B \text{ and } C)$$
or,
$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Chapter 4: Probability Theory - Flow Diagram (2)


Probability Theory...continued
Probabilities when the Events are Independent
Simple/Marginal or Unconditional Probability
Joint Probability
Conditional Probability under independence
Simple Probability- is also known as marginal or unconditional and is the probability of occurrence for a single event, say A, and is denoted by P (A).
P(A) = marginal probability of event A; P(B) = marginal probability of event B
Joint Probability under Statistical Independence:
Joint probability is the probability of occurrence for two or more events together or in succession. It is also known as 'and' probability. Suppose we have two events, A and B, which are independent. Then the joint probability, P (AB), which is the probability of occurrence of both A 'and' B, is given by
P(AB) = P(A).P(B)
or, P(A∩B)=P(A).P(B)
Note that P (AB) = probability of event A and B occurring together — is known as joint probability. P(AB) is the same as P (A and B) or P(A∩B)
Conditional Probability under Statistical Independence
The conditional probability is written as
P(A|B)
and is read as the probability of event A, given that B has occurred, or the probability of A, given B. If the two events A and B are independent, then
P(A|B) = P(A)
This means that if the events are independent, the probabilities are not affected by the occurrence of each other.
Chapter 4: Probability Theory - Flow Diagram (3)

Probability Theory...continued

Probabilities when the Events are Dependent

- Conditional Probability
- Joint Probability
- Marginal Probability

Conditional probability under Statistical Dependence

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A \text{ and } B)}{P(B)}$$

or,

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(B \text{ and } A)}{P(A)}$$

Joint probability under Statistical Dependence

$$P(A \cap B) = P(A|B)P(B)$$

or

$$P(A \text{ and } B) = P(A|B)P(B)$$

or, $P(B \cap A) = P(B|A)P(A)$

or

$$P(B \text{ and } B) = P(B|A)P(A)$$

Marginal probability under Statistical Dependence

The marginal probability under statistical dependence can explained using the joint probability table below. Consider the four events: D, S, R, and G and the probabilities below.

	(D)	(S)	Total
(R)	0.30	0.10	0.40
(G)	0.20	0.40	0.60
Total	0.50	0.50	1.00

P(R) = P(D and R) + P(S and R)

$$P(R) = P(D|R)P(R) + P(S|R)P(R)$$

Bay's Theorem

$$P(A_i|D) = \frac{P(A_i)P(D|A_i)}{P(A_1)P(D|A_1) + P(A_2)P(D|A_2) + \ldots + P(A_n)P(D|A_n)}$$

This equation can be used to compute any ***posterior probability*** $P(A_i|D)$ *when prior probabilities* $P(A_1), P(A_2) \ldots, P(A_n)$ *and conditional probabilities* $P(D|A_1), P(D|A_2) \ldots, P(D|A_n)$ *are known.*

Chapter 4: Probability Theory - Flow Diagram (4)

Probability Distributions Concepts for Data Science

Random Variables

A random variable is a variable that takes on different values as a result of the outcomes of a random experiment. It can also be a variable that assumes numerical values governed by chance so that a particular value cannot be predicted in advance. There are two types of random variables.

- **Discrete (Countable)** → Random variable X is either finite or countably infinite
- **Continuous(Uncountable)** → The random variable takes any value within a given range

Probability Distribution and Frequency Distribution

The probability distribution is a model that relates the value of a variable with the probability of occurrence of that value. The probability distribution describes the frequencies that occur theoretically; whereas, the relative frequency distribution describes the frequencies that have actually occurred.

Expected Value, Variance, and Standard Deviation of a Discrete Distribution

Expected Value $\mu_x = E(X) = \sum X_i P(X_i)$

Variance $\sigma^2 = \Sigma (x_i - \mu)^2 P(x_i)$

Standard Deviation $\sigma = \sqrt{\sigma^2}$

Chapter 5: Discrete Probability Distributions - Flow Chart (1)

Discrete Probability Distributions

Binomial Distribution

Binomial Distribution

The experiment or the process under study consists of ***n*** number of trials. Each trial has only two possible outcomes; success (S) and failure (F). The following properties hold:

- **there are *n* number of trials**
- **each trial has only two possible outcomes; success (S) and failure (F)**
- **the probability of success and the probability of failure remains constant across trials**
- **the outcomes are independent of each other**

A random variable X that denotes x number of successes in n Bernoulli trials is said to have a Binomial distribution.

The Binomial distribution calculates the probability of x successes in n trials using the following expression:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{where, } x = 0,1,\ldots\ldots,n$$

p(x) = probability of x number of successes, n = number of trials, p = probability of success, (1-p) = q is the probability of failure

Poisson Distribution

The Poisson Distribution

A random variable X is said to follow a Poisson distribution if it assumes only nonnegative values and its probability density function is given by:

$$p(x) = \frac{e^{-\mu}\mu^x}{x!} \quad \text{where, } x = 0,1,2,\ldots\ldots,n$$

where μ represents the mean and variance of the distribution.

Note that $\mu > 0$

The Poisson distribution occurs when there are events which do not occur as outcomes for a fixed number of trials of an experiment (unlike that of the Binomial distribution), but which occur at random points of time and space. The Poisson distribution is the correct distribution to apply when n is very large (that is, the area of opportunity is very large) and an event has a constant and very small probability of occurrence. The Poisson distribution calculates the probability of X number of occurrences.

Mean, Variance, and Standard Deviation of Binomial Distribution

The mean or expected value of the Binomial distribution is given by

$$E(x) = \mu = np$$

where, n= number of trials, and p = probability of success

Variance of a Binomial Distribution:

$$\sigma^2 = np(1-p)$$

Standard Deviation of a Binomial distribution:

$$\sigma = \sqrt{np(1-p)}$$

Chapter 5: Discrete Probability Distributions - Flow Chart (2)

Random Variables and Continuous Probability Distribution

Random Variables and Probability Distribution

Continuous Probability Distributions

Random Variables and Probability Distribution

A random variable is a numerical quantity whose value is determined by chance. A random variable must be a numerical quantity. The relationship between the values of a random variable and their probabilities is summarized by a probability distribution. Probability distributions are characterized by:

The probability density function: the probability density function, f(x), describes the behavior of a random variable and may be viewed as the shape of the distribution. The probability density function represents the entire sample space; therefore, the area under the probability density function must equal one.

$$\int_{-\infty}^{\infty} f(x) = 1$$

The cumulative distribution function: the cumulative distribution function, F(x), denotes the area beneath the probability density function to the left of x.

$$F(x) = \int_{-\infty}^{x} f(r)dr$$

Normal Distribution

To calculate the normal probability, $p(x_1 \le X \le x_2)$ where X is normal with parameters μ and σ, we need to evaluate:

$$\int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx$$

The normal distribution with μ=0 and σ=1 is called a standard normal distribution. Also, a random variable with standard normal distribution is called a standard normal random variable and is usually denoted by Z. If x is normally distributed with mean μ and standard deviation σ, then

$$Z = \frac{x-\mu}{\sigma}$$

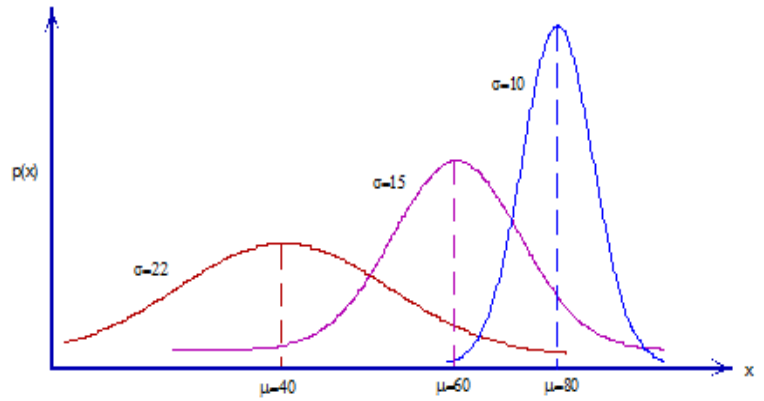
is a standard normal random variable where,
Z = distance from the mean to the point of interest (x) in terms of standard deviation units
x = point of interest
μ = the mean of the distribution, and
σ = the standard deviation of the distribution

Chapter 6: Continuous Probability Distributions - Flow Chart (1)

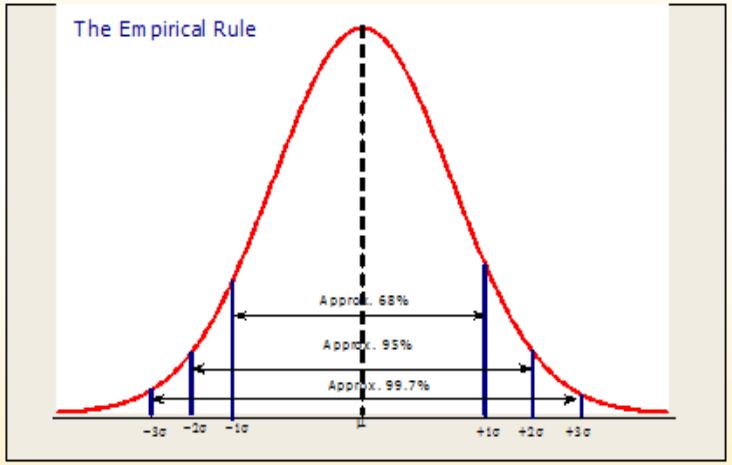
Continuous Probability Distributions...continued

Parameters of Normal Distribution

The shape of the normal curve depends upon the mean (μ) and standard deviation (σ). The mean μ and the standard deviation (σ) are the parameters of the normal distribution. The mean μ determines the location of the distribution whereas; the standard deviation σ determines the spread of the distribution.



Area Property of Normal distribution



Chapter 6: Continuous Probability Distributions - Flow Chart (2)

Exponential Distribution

Probability and Cumulative Density Function of Exponential Distribution

If the random variable X follows an exponential distribution then the probability density function is given by:

$$f(x)=\frac{1}{\mu}e^{-x/\mu} \qquad \text{where, } x>0 \text{ and } \mu>0$$

Cumulative Probabilities for exponential distribution is given by:

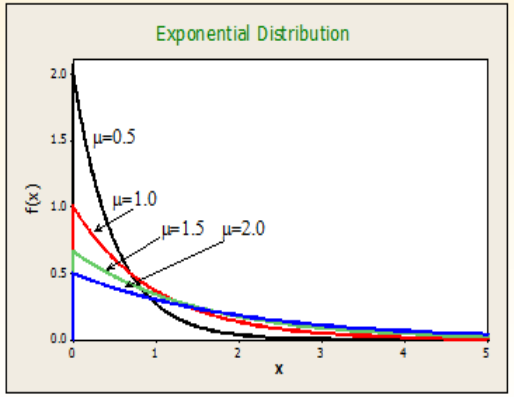
$$P(x\le x_0)=1-e^{-x/\mu} \qquad \text{for } x>0$$

The mean and standard deviation of the exponential distribution are equal and given by:

Mean = μ

Standard deviation, σ = μ

Graph of Exponential Distribution for different values of μ



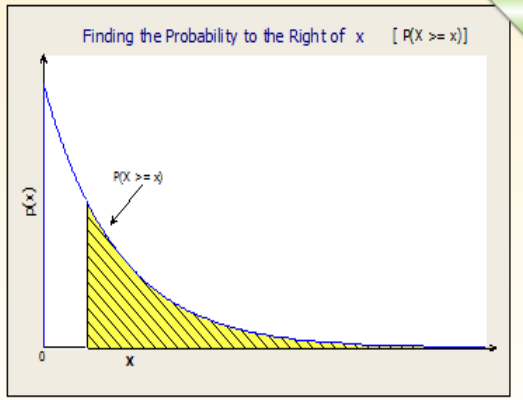
Finding Exponential Probabilities

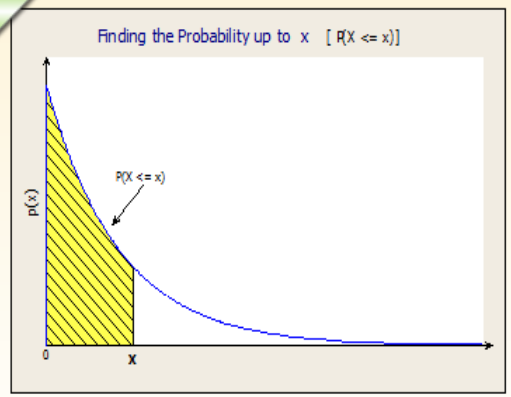
The probabilities for exponentially distributed random variables are found by evaluating the areas between the points of interest of the exponential curve described in Figure 6.39. Suppose X is an exponentially distributed random variable with parameter μ, then

$$P(X\ge x)=e^{-x/\mu} \qquad \text{for } x=0$$

$$P(X\le x)=1-e^{-x/\mu} \qquad \text{for } x>0$$

$$P(x_1\le X\le x_2)=e^{-x_1/\mu}-e^{-x_2/\mu} \qquad \text{for } x_1, x_2>0$$



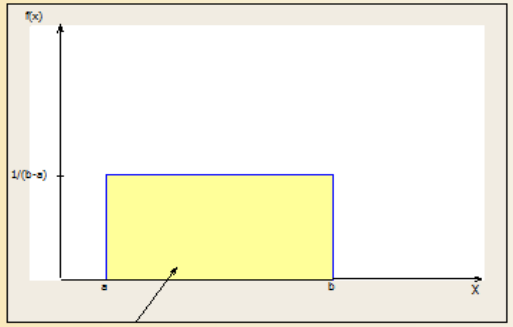


Uniform Distribution

A random variable for which all outcomes between some minimum and maximum values have equal probability of occurrence may be described by a uniform distribution. The density function of the uniform distribution is given by

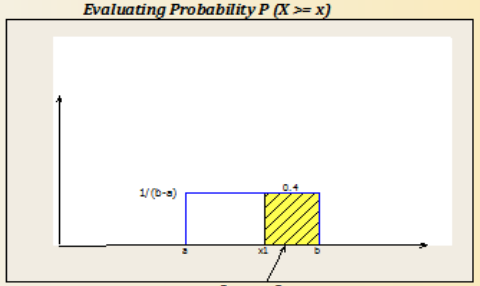
$$f(x) = \frac{1}{b-a} \qquad a = x = b$$
$$= 0 \qquad \text{otherwise}$$

- Note that *a* and *b* are two parameters of the uniform distribution where a < b. Parameter *a* is the location parameter and it controls the location of the distribution along the x-axis. The scale parameter is the difference (b-a). An increase in the difference (b-a) will elongate the distribution, whereas a decrease in the difference (b-a) will compress it.



Total shaded area= $(b-a)\left[\frac{1}{(b-a)}\right]$

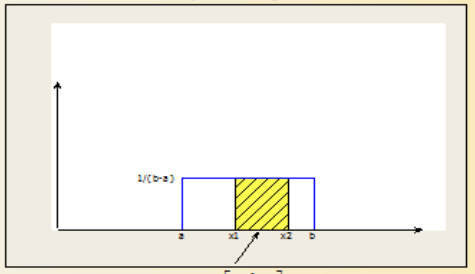
Evaluating Probability P (X >= x)



Shaded Area= $(b-x_1)\left[\frac{1}{(b-a)}\right]$

Evaluating the Probability

$P(x_1 \le X \le x_2)$



Area=$(x_2-x_1)\left[\frac{1}{(b-a)}\right]$

Mean or the Expected Value E(x) and the Variance V(x) of the Uniform Distribution

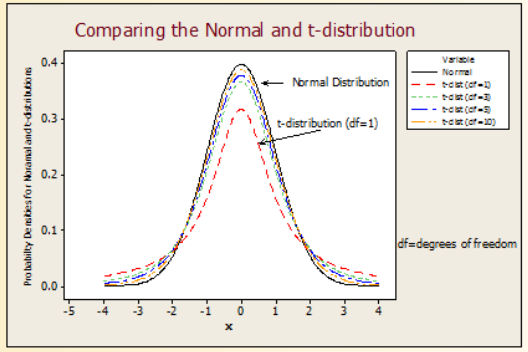
$$E(x) = \mu = \frac{(a+b)}{2} \qquad V(x) = \sigma^2 = \frac{(b-a)^2}{12}$$

Some Important Distributions Related to Normal Distribution

Distributions Related to Normal Distribution — extensively used in Conducting Several Statistical tests

t-distribution

- It was shown by Gosset that the random variable $t = [(\bar{x} - \mu)/(s/\sqrt{n})]$ follows the distribution known as the t-distribution if σ is not known and the sample size ***n*** is small,
- The statistic t has a mean=0 and a variance > 1(unlike the normal distribution whose mean=0 and variance=1).
- Since the variance is greater than 1, this distribution is less peaked at the center compared to the normal distribution and is also higher in the tails compared to the normal distribution.
- As the sample size, n becomes larger; the t-distribution comes closer and closer to the normal distribution.



Chi-square distribution

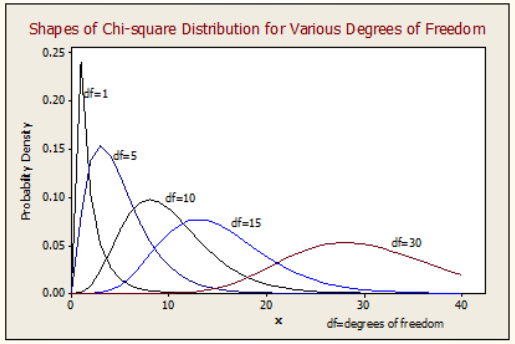
- The chi-square distribution is also defined in terms of the normal distribution. This distribution is always skewed to the right, but as the degrees of freedom increase the distribution tends toward a normal distribution.

- If a set of independent random variables $z_1, z_2, z_3, \ldots, z_k$ are normally distributed with mean zero and variance one then the sum of squares of $z_1, z_2, z_3, \ldots, z_k$ denoted by χ^2 (Chi-square) is also a random variable, and the quantity

$$\chi^2_n = z_1^2 + z_2^2 + z_3^2 + \ldots\ldots + z_k^2$$

is distributed as a chi-square distribution with n degrees of freedom. The values of χ^2 are between zero and +8 because χ^2 is the sum of squares.

- The sampling distribution of the sample variance, s^2 ; $(n-1)\, s^2 / \sigma^2$ follows a χ^2 distribution with (n-1) degrees of freedom.

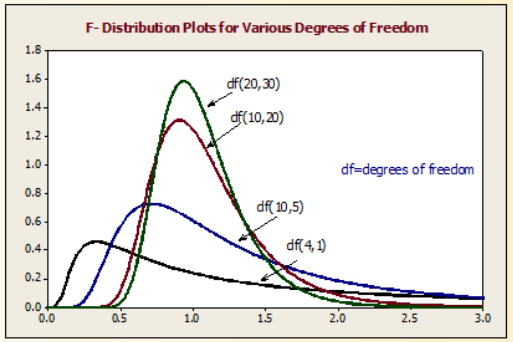


F- distribution

The F-distribution is also related to the normal distribution. Suppose, we have two independent normal variables x_1 and x_2 with the following mean and variances $x_1 \sim N(\mu_1, \sigma_1^2)$ and $x_2 \sim N(\mu_2, \sigma_2^2)$ and we draw samples of size n_1 and n_2 from the first and second normal processes. If the sample variances are s_1^2 and s_2^2, then the ratio

$$\frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

follows an F-distribution with (n_1-1) and (n_2 -1) degrees of freedom. The shape of the F-distribution depends upon the numerator and denominator degrees of freedom. As the degrees of freedom increase, the distribution approaches the normal distribution.



Chapter 6: Continuous Probability Distributions - Flow Chart (5)

Discrete Probability Distributions...continued

Hypergeometric Distribution

Hypergeometric Distribution

- In hypergeometric distribution, the trials are not independent and the probability of success changes from trial to trial.
- If the probability of success is not constant from trial to trial and the sampling is done from a population without replacement, the appropriate distribution is the hypergeometric distribution.
- The hypergeometric distribution calculates the probability of x, the specified number of successes.
- The conditions for the hypergeometric distribution are: (a) the population size is finite, (b) the sampling is done from the finite population without replacement, and (c) the sample size n is greater than 5% of the population size, N.

The hypergeometric probability function, p(x) that is used to determine the probability of x successes in a sample size of n selected without replacement and is given by:

$$p(x) = \frac{(C_x^D)(C_{n-x}^{N-D})}{C_n^N} = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$$

p(x) = probability of x number of successes in a sample of size n

D = the number of successes in the population or the number of interest

N = population size, x = the number of successes of interest, x = 0, 1, 2, ...

n = sample size

C_n^N = the combinations of all units

C_x^D = the combinations of x successes from D successes

C_{n-x}^{N-D} = the combinations of (n–x) failures from (N–D) failures

Negative Binomial or Pascal Distribution

Negative Binomial or Pascal Distribution

The basic difference between the binomial and the negative binomial distribution is that the binomial distribution is used to calculate the probability of x number of successes out of n trials where the number of trials is fixed, whereas, in a negative binomial distribution the trials are repeated until a fixed number of successes occur. We are interested in finding the probability that the r^{th} success occurs on the x^{th} trial.

Suppose we have a succession of n Bernoulli trials. Assume that the (i) trials are independent, (ii) the probability of success p in the trial remains constant from trial to trial, and (iii) the probability of failure is q = 1-p. Then the probability distribution of the random variable X, the number of trial on which the r^{th} success occurs, is given by

$$b^*(x;r,p) = \binom{x-1}{r-1} p^r q^{x-r}; x = r, r+1, r+2, \ldots\ldots$$

The probability of r^{th} success occurring on the x^{th} trial is also written as

$$p(x) = P(X = x) = \binom{x-1}{r-1} p^r q^{x-r}; x = r, r+1, r+2, \ldots\ldots$$

Note: For the Poisson distribution; the mean and the variance are equal. The equality of mean and the variance is an important characteristic of the Poisson distribution. For the binomial distribution, the mean is always greater than the variance. In some cases, the observable phenomenon gives rise to empirical distributions in which the variance is larger than the mean. In cases where the variance is larger than the mean, the negative binomial distribution provides a good model.

Chapter 5: Discrete Probability Distributions - Flow Chart (3)

Discrete Probability Distributions...continued

Geometric Distribution

The Geometric Distribution

- The geometric distribution is related to a sequence of Bernoulli trials in which the random variable X takes two values 0 and 1 with the probability q and p respectively, that is, p(X=1) = p, p(X=0)=q, and q = 1-p.
- In the geometric distribution, the number of trials is not fixed, and the random variable of interest X, is defined as the number of trials required to achieve the first success.
- The geometric distribution can be derived as a special case of negative binomial distribution. In the probability density function of negative binomial distribution above, if r = 1, we get the probability distribution for the number of trials required to achieve the first success.
- If we have a series of independent trials that can result in a success with probability p and a failure with probability q where, q=1-p, then the random variable X that denotes the number of trials on which the first success occurs, is given by

$$p(x;p) = pq^{x-1} \quad \text{where, } x = 1,2,\ldots.$$
$$= 0 \quad \text{otherwise.}$$

The distribution is called geometric because the probabilities for x=0, 1, 2,....., are the various terms of geometric progression.

All the assumptions of Binomial distribution apply to negative binomial distribution. Unlike the binomial distribution, the geometric distribution does not have a fixed sample size because the sampling continues until the first success is observed. Also, the variable x cannot be 0 because at least one trial is needed for the first success to occur.

Discrete Uniform Distribution

Discrete Uniform Distribution

In a discrete uniform distribution, the probability for a discrete variable is equal for all values. For example, in a toss of a six-sided fair dice, the probability of occurrence of each of the possible outcome 1 through 6 is 1/6. The probability density of a discrete uniform distribution is given by

$$f(x) = \frac{1}{(b-a)+1}$$

for x =a, a+1, ...,b-1, b (and zero elsewhere)

The mean and variance of the discrete uniform random variable is given by

$$\mu = \frac{a+b}{2}$$
$$\sigma^2 = \frac{(b-a)(b-a+2)}{12}$$

Multinomial Distribution

Multinomial Distribution

The multinomial distribution is a generalization of the binomial distribution. If a trial results in more than two mutually exclusive outcomes, then this leads to a multinomial distribution. Suppose $E_1, E_2, E_3, \ldots\ldots, E_k$ are k mutually exclusive outcomes of a trial with probabilities $p_1, p_2, p_3, \ldots\ldots, p_k$ then the probability that E_1 occurs x_1 times, E_2 occurs x_2 times..., and E_k occurs x_k times in n independent observations can be given by

$$p(x_1,x_2,x_3,\ldots\ldots) = \frac{n!}{x_1!x_2!\ldots.x_k!} p_1^{x_1} p_2^{x_2} \ldots\ldots p_k^{x_k}; 0 \le x \le n$$

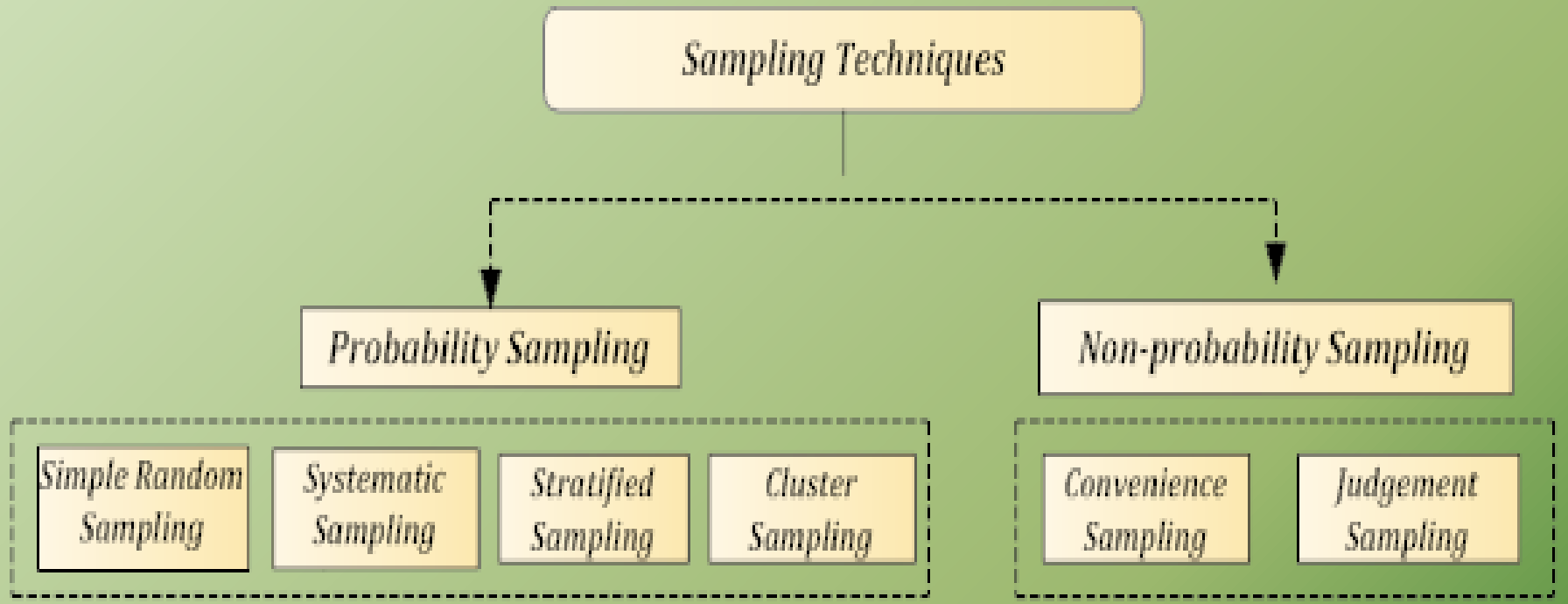
Note that

$$x_1 + x_2 + \ldots\ldots + x_k = n \quad \text{and} \quad p_1 + p_2 + \ldots\ldots + p_k = 1$$

Chapter 5: Discrete Probability Distributions - Flow Chart (4)

Sampling and Estimation Concepts for Data Science

Sampling Techniques



Sampling and Sampling Distribution

Standard Error &Sampling Distribution of the Sample Mean

Expected Value of $\bar{x}$

$$E(\bar{x}) = \mu$$

Standard deviation of $\bar{x}$, $\sigma_{\bar{x}}$ or the Standard error

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$ for an infinite population

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$ for a finite population

---

$\sigma_{\bar{x}} =$ where, n= sample size, N= population size, σ is the population standard deviation

The factor,

$$\sqrt{\frac{N-n}{N-1}}$$

is known as the finite population correction factor

if $\frac{n}{N} < 0.05$ (the ratio of the sample to population size is <0.05), do not use the finite population correction factor as it does not help reduce the standard error)

Sampling Distribution of $\bar{x}$ follows a normal distribution for $n \geq 30$

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$ For an infinite population

---

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$ For a finite population

Sampling Distribution of the Sample Proportion

Expected value of $\bar{p}$: $E(\bar{p}) = p$

$\bar{p}$ = sample proportion, p = population proportion

Standard deviation of $\bar{p}$:

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$ for an infinite population

---

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$ for a finite population

Sampling Distribution of a proportion ($\bar{p}$)

Sampling distribution of sample proportion follows a normal distribution for $n \geq 30$

$$z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

$$z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}}$$

The first formula is for infinite population; the second is for a finite population.

Estimating Population Values: Confidence Intervals

Confidence interval formulas to estimate the population mean, μ and Population Proportion (p)

Confidence interval formulas to estimate the population mean, μ

Case (1): Large sample (n), σ known: use normal distribution

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

or,

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Margin of Error for estimating μ when σ is known

$$E = \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Case (2): Large sample (n), σ unknown: use normal distribution

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

or,

$$\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Case (3): Small sample (n), σ unknown: use t- distribution

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$ *or,*

$$\bar{x} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

Some commonly used confidence intervals and corresponding Z-values

90% Confidence Interval	*Z=1.645*
95% Confidence Interval	*Z=1.96*
99% Confidence Interval	*Z=2.58*

Confidence interval formulas to estimate the population proportion, p

Assumption: sample size n is large so that normal approximation can be used

$$\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

or,

$$\bar{p} - z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \le p \le \bar{p} + z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

In the above formula,

p = population proportion

$\bar{p} = \frac{x}{n}$ *(Sample proportion)*

Determine the sample size (n)

Determine the sample size (n) to estimate μ

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$$

E=error or accuracy

n= sample size

Determine the sample size (n) to estimate p

$$n = \frac{(z_{\alpha/2})^2 p(1-p)}{E^2}$$

p= population proportion (if p is not known or given, use p=0.5)

Some less commonly used confidence intervals and their Z-values

80% Confidence Interval	*Z=1.28*
94% Confidence Interval	*Z= 1.88*
96% Confidence Interval	*Z=2.05*

Testing a Single Mean and a Single Proportion: Selecting the Right Test and Right Procedure

Hypothesis Tests for a Single Mean

$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$

Two -tailed or Two-sided Test

$H_0 : \mu \geq \mu_0$
$H_1 : \mu < \mu_0$

Left-tailed or Left-sided Test

$H_0 : \mu \leq \mu_0$
$H_1 : \mu > \mu_0$

Right-tailed or Right-sided Test

Case (1) : Sample size (n) large, s known: the sample mean follows a Normal distribution and the test statistics is given by

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Case (2) : Sample size (n) large, s unknown: the sample mean follows a Normal distribution and the test statistics is given by

$$z = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Case (3) : Sample size (n) small, s unknown: the sample mean follows a t- distribution and the test statistics is given by

$$t_{n-1} = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Test Methods

***(1) Z-value approach**: calculate the z-value using the test statistics formula and compare it to $Z_{critical}$.*

Decision rules:

Two-sided test	*Right-sided test*	*Left-sided test*
Reject H_0 if $Z > Z_{critical}$ or, if $Z < z_{critical}$	*Reject H_0 if $Z > Z_{critical}$*	*Reject H_0 if $Z < Z_{critical}$*

Note: If you are using a t-distribution then $z_{critical}$ is replaced by $t_{critical}$ and z is replaced by t_{n-1}.

***(2) p-value approach**: determine the p-value and compare it to the level of significance a. The decision rule is given by:*

Decision rule:

If $p = \alpha$; do not reject H_0
If $p < \alpha$; reject H_0

***(3) Critical value approach:** calculate the critical value of $\bar{x}$ (known as $\bar{x}_c$) and compare it to the $\bar{x}$ of the sample data. Decision rule*

Two-sided test	*Right-sided test*	*Left-sided test*
Reject H_0 if $\bar{x} > \bar{x}_c$ or, if $\bar{x} < \bar{x}_c$	*Reject H_0 if $\bar{x} > \bar{x}_c$*	*Reject H_0 if $\bar{x} < \bar{x}_c$*

$\bar{x}_c$ is calculated using the following formulas:

$\overline{x_c} = \mu \pm z \frac{\sigma}{\sqrt{n}}$ *For a two-sided test*

$\overline{x_c} = \mu + z \frac{\sigma}{\sqrt{n}}$ *For a right-sided test*

$\overline{x_c} = \mu - z \frac{\sigma}{\sqrt{n}}$ *For a left-sided test*

If you are using t-distribution then z in the above formula, Z is replaced by t_{n-1} and appropriate α , and the σ is replaced by s.

***(4) Confidence interval approach**: Calculate the confidence intervals using appropriate formulas.*
Decision rule: Reject H_0 if the hypothesized value (m) is outside the calculated confidence interval.

Confidence interval formulas:

$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$ *When n is large and σ is known*

$\bar{x} \pm z \frac{s}{\sqrt{n}}$ *When n is large and σ is unknown*

$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$ *When n is small and σ is unknown*

Calculation of type II error β and the power of the test (1-β) See the examples.Relationship between type I error α and β [See the examples]
IV. Sample size for one-sided test

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_a)^2}$$

Chapter 9 : Introduction to Hypothesis Testing - Flow Chart (1)

Testing a Single Mean and a Single Proportion: Selecting the Right Test and Right Procedure ...continued

Testing a Single Proportion

Hypothesis:

Two-tailed or Two-sided Test	Left-tailed or Left-sided Test	Right-tailed or Right-sided Test
$H_0 : p = p_0$ $H_1 : p \neq p_0$	$H_0 : p \geq p_0$ $H_1 : p < p_0$	$H_0 : p \leq p_0$ $H_1 : p > p_0$

Test Statistics:

If the sample size n is large, the sample proportion $\overline{p}$ follows a Normal distribution and the test statistics is given by

$$z = \frac{\overline{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

$\overline{p}$ = sample proportion = x /n p_0 = population proportion (or, the hypothesized value)

Test Methods

*(1) **Z-value approach**: calculate the z-value using the test statistics formula and compare it to $Z_{critical}$.*

Decision rule:

Two-sided test	Right-sided test	Left-sided test
Reject H_0 if $Z > Z_{critical}$ or, if $Z < z_{critical}$	Reject H_0 if $Z > Z_{critical}$	Reject H_0 if $Z < Z_{critical}$

(2) p-value approach: determine the p-value and compare it to the level of significance a. The decision rule is given by:

Decision rule:

If $p \geq \alpha$; do not reject H_0
If $p < \alpha$; reject H_0

*(3) **Critical value approach**: calculate the critical value of $\overline{p}$ ($\overline{p_c}$) and compare it to the $\overline{p}$ of the sample data.*

Decision rule:

Two-sided test	**Right-sided test**	**Left-sided test**
Reject H_0 if $\overline{p} > \overline{p_c}$ or, if $\overline{p} < \overline{p_c}$	Reject H_0 if $\overline{p} > \overline{p_c}$	Reject H_0 if $\overline{p} < \overline{p_c}$

(4) Confidence interval approach: Calculate the confidence intervals using appropriate formulas.

Decision rule:
Reject H_0 if the hypothesized value (p_0) is outside of the calculated confidence interval.

Confidence interval formula:

$$\overline{p} \pm z\sqrt{\frac{\overline{p}(1 - \overline{p})}{n}}$$

$\overline{p_c}$ is calculated using the following formulas:

$$\overline{p_c} = p_0 \pm z\sqrt{\frac{p_0(1 - p_0)}{n}}$$ *(For a two-sided test)*

Note: for a right sided test, the ± sign in the above formula is replaced by a (+) sign, and for a left-sided test the ± sign is replaced by a negative (-). Also, note that in a two-sided test, $\alpha = \alpha/2$ and for a one sided test, $\alpha = \alpha$.

Chapter 9: Hypothesis Testing - Flow Chart (2)

Machine Learning and Machine Learning Models

What is Machine Learning (ML)?

- Machine learning applications are a way of analyzing and creating models from huge amount of data commonly referred to as ***big data***. Machine learning is closely related to artificial intelligence (AI). In fact, it is an application of artificial intelligence (AI).

- Machine learning is a method of designing systems that can learn, adjust, and improve based on the data fed to them without being explicitly programmed.

Machine Learning Models

- Machine learning methods are used to develop complex models and algorithms that help to understand and make predictions from the data.

- The analytical models in machine learning allow the analysts to make predictions by learning from the trends, patterns and relationships in the historical data.

- Machine learning automates model building. The algorithms in machine learning are designed to learn iteratively from data without being programmed.

Machine Learning (ML) _ Some Applications.

- According to Arthur Samuel, machine learning gives "computers the ability to learn without being explicitly programmed."[2][3].
- Samuel, an American pioneer in the field of *computer gaming* and artificial intelligence, coined the term "machine learning" in 1959 while at IBM.
- ***Machine learning algorithms are extensively used for data-driven predictions and in decision making***.
- Some applications where machine learning has been used are email filtering, detection of network intruders or detecting a data breach, optical character recognition (OCR), learning to rank, computer vision, and a wide range of engineering and business applications.

Machine Learning Methods and Tasks

- At the fundamental level, Machine learning tasks are typically classified into following broad categories depending on the nature of the learning "signal" or "feedback" available to a learning system. These are[20]
- (1) Supervised learning
- (2) Unsupervised learning

- The **supervised learning** problems can be divided into

(1) Classification problems and

(2) Regression problems

Machine Learning Methods and Tasks

(2) Regression Regression is a supervised problem where the outputs are continuous rather than discrete.

- Regression problems are used to predict continuous labels. There are several regression models used in machine learning. These are linear and non-linear regression, logistic regression and others. Besides these the commonly used algorithms are support vector machines (SVM). These are the most widely used regression algorithms:
- Ordinary Least Squares Regression (OLSR)
- Linear Regression/ Multiple Regression
- Logistic Regression Stepwise Regression Multivariate Adaptive Regression Splines (MARS) Locally Estimated Scatterplot Smoothing (LOESS)

Unsupervised Learning

- **Clustering:** In **clustering**, data is assigned to some number of discrete groups. In this type of problem, a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.

- Clustering technique is used to find natural groupings or clusters in a set of data without pre specifying a set of categories.

Types of Clustering

- **K-means clustering**,
- **Spectral clustering**, and
- **Gaussian mixture models**.

More Applications of Machine Learning

- Another application of machine learning is in the area of deep learning [25] which is based on artificial neural networks.[6] In his application the learning tasks may contain more than one hidden layer) or tasks with a single hidden layer known as shallow learning.
- **Artificial neural networks [** *Artificial neural network* **]** artificial neural network [6,7] (ANN) learning algorithm, usually called "neural network" (NN), is a learning algorithm that is inspired by the structure and functional aspects of biological neural networks. **Deep learning** [8]

Note: Neural networks use machine learning algorithms extensively; whereas machine leaning is an application of *artificial intelligence* [9] *that automates analytical model building by using* ***algorithms*** *that iteratively learn from data without being explicitly programmed.*

Some Technologies used in Machine Learning

- Python is a programming language with simple syntax that is commonly used for data science.[35] There are a number of python libraries that are used in data science and machine learning applications including **NumPy, pandas, Matplot, Scikit Learn, and others**.
- R statistical analysis, a programming language that was designed for statistics and data mining[36] applications and is one of the popular application package used by data scientists and analysts.
- TensorFlow is a framework for creating machine learning models developed by Google used machine learning models and applications.
- Pytorch is another framework for machine learning developed by Facebook.
- Jupyter Notebook is an interactive web interface for Python that allows faster experimentation and is used in machine learning applications of data science.
- Tableau makes a variety of software that is used for data visualization[37]. It is a widely used software for big data applications and is used for descriptive analytics, and data visualization.
- Apache Hadoop is a software framework that is used to process data over large distributed systems.

Summary

- Data science is a multidisciplinary field focused on finding actionable insights from large sets of raw, structured, and unstructured data.
- It uses **several different areas** from data and statistical analysis and modeling, programming from varied areas of computer science, analytics, predictive analytics, information systems, and machine learning to parse through massive datasets in an effort to find solutions to problems that haven't been thought of yet.
- Data science is a data-driven decision-making approach to solve varied problems in businesses, science and engineering.

Summary

- The field of data science provides a unifying approach by combining varied areas ranging from statistics, mathematics, analytics, business intelligence, computer science, programming and information systems.
- Data science continues to evolve as one of the most sought-after areas by companies.

My books in Data Science and Analytics