2023

PRACTICAL DATA ANALYTICS – BEGINNER TO PRO

ULOH C. ARMSTRONG

…

BEBETTER ANALYTICS |

This note was created for the "Practical data analytics – Beginner to Pro" Boot camp, and for now I have decided to leave it open and available for anyone who is interested in data analytics, data science and/or the basics of statistics.

It was designed to give you a meaningful structure, to serve as a guide along the way. For those that have enrolled in the boot camp, it'll keep you abreast of the steps, and also informed on all that is going on in the boot camp.

The ideal use case of this note is that you print this note and keep it close to you while you're learning and practicing all that you see in it.

You will find the SQL, Excel and Python codes to be italicized, note that the codes are examples, and you can switch it up and attune the codes as you learn and develop.

Data analytics opens itself up to you, if you spend time with it. Employ consistency and don't fire her.

To humble beginnings.

Uloh C. Armstrong
Email: armstronguloh@gmail.com

Table of Contents

UNDERSTANDING DATA ANALYTICS

WHO IS A DATA ANALYST?

A data analyst is a professional who collects, processes, and interprets small/large sets of data to discover insights, trends, and patterns.

Imagine a data analyst as a detective in a mystery story.

Data Analyst as a Detective:

- Gathering Clues (Data Collection): Just like a detective collects clues from a crime scene, a data analyst gathers information from various sources, like numbers, pictures, and spreadsheets.
- Analyzing Clues (Data Analysis): Similar to how a detective analyzes clues to solve a mystery, a data analyst examines data to uncover patterns, trends, and insights hidden within the data collected.
- Finding Solutions (Making Decisions): Once the detective figures out the mystery, they use their findings to solve the case. Similarly, a data analyst uses their insights to help businesses and organizations make smarter decisions, solve problems, and improve processes.
- Storytelling (Reporting Insights): Just as a detective presents their findings in a report or to a jury, a data analyst shares their discoveries using visualizations, reports, and presentations to explain what the data is saying.

In essence, a data analyst is like a detective for data. They collect, investigate, and interpret data to help organizations make informed decisions and solve problems.

So, while a detective solves mysteries, a data analyst solves data puzzles to reveal meaningful insights!

Think of a data analyst as a skilled chef in a bustling restaurant kitchen.

Data Analyst as a Skilled Chef:

- Ingredients Collection (Data Collection): Similar to how a chef gathers various ingredients from the pantry, a data analyst collects different types of data from databases, spreadsheets, and sources like a well-stocked pantry.
- Recipe Creation (Data Analysis): Just as a chef combines ingredients to create a delicious dish, a data analyst blends and examines data using analytical tools and methods to uncover meaningful insights, much like crafting a recipe.
- Taste Testing (Decision-Making): After preparing a dish, a chef tastes it to ensure it's just right before serving. Similarly, a data analyst evaluates and validates their findings to ensure the insights are accurate and reliable before presenting them to stakeholders for decision-making.
- Presenting the Dish (Reporting Insights): Like a chef presents a beautifully plated dish, a data analyst communicates their analyzed data through compelling visualizations, reports, and presentations, making it easily digestible for business leaders and stakeholders.

So, while a chef creates delectable dishes from ingredients, a data analyst crafts valuable insights from data sets, both aiming to deliver something remarkable and satisfying to their audience!

SKILLS REQUIRED OF A DATA ANALYST

It can be divided into two categories, the hard and soft skills, and with these skills do you have a complete data analyst.

Hard Skills:

- Analytical Skills: The ability to dissect and analyze data sets to identify trends, correlations, and patterns.
- Statistical Knowledge: Understanding statistical concepts and methods to interpret data accurately, perform hypothesis testing, and derive meaningful conclusions.
- Programming Skills: Proficiency in programming languages like Python, R, or SQL for data manipulation, analysis, and querying databases.

- Data Visualization: Proficiency in creating visually engaging charts, graphs, and dashboards using tools like Tableau, Power BI, matplotlib etc.
- Database Management**:** Knowledge of database systems and query languages (e.g., SQL) to efficiently retrieve, manage, and manipulate data.

Soft Skills:

- Critical Thinking: Ability to approach problems logically, think critically, and find innovative solutions using data-driven methodologies.
- Communication Skills: Effective communication to convey complex findings, reports, and insights in a clear and understandable manner to diverse audiences.
- Attention to Detail: Precision in data analysis, ensuring accuracy, identifying errors, and maintaining data integrity.
- Problem-Solving Skills: Aptitude for solving complex business problems by leveraging data insights and recommending actionable solutions.
- Business Acumen: Understanding of the industry domain and business processes to contextualize data insights for informed decision-making.
- Curiosity and Learning Mindset: Eagerness to explore new techniques, tools, and advancements in data analytics to continuously improve skills.
- Ethical Consideration: Awareness of ethical considerations in data handling, ensuring compliance with privacy regulations and maintaining data confidentiality.

RESPONSIBILITIES OF A DATA ANALYST

You must not be Kafkaesque as a data analyst!

Your duty as a data analyst in simpler terms is to answer questions. You are the go-to person when it comes to finding answers from data.

The responsibilities of a data analyst typically involve a range of tasks related to data collection, analysis, interpretation, and reporting.

Here are the key responsibilities:

- Data Collection**:** Gather data from various sources such as databases, spreadsheets, API's, and other repositories.
- Data Cleaning and Preparation: Cleanse, organize, and preprocess data to ensure accuracy, completeness, and consistency for analysis.
- Data Analysis: Apply statistical methods, algorithms, and analytical techniques to analyze large datasets, identify trends, patterns, and correlations.
- Data Modeling: Develop models and algorithms to extract insights, make predictions, or optimize processes based on the analyzed data.
- Data Visualization: Create visualizations, charts, graphs, and dashboards to present data insights in a clear and understandable format using tools like Tableau, Power BI, matplotlib etc.
- Reporting and Presentation**:** Prepare reports, summaries, and presentations to communicate findings, actionable insights, and recommendations to stakeholders.
- Collaboration: Collaborate with cross-functional teams, including business analysts, engineers, and managers, to understand business needs and provide data-driven solutions.
- Continuous Monitoring: Monitor and evaluate data for accuracy, completeness, and consistency, ensuring data quality and integrity over time.
- Problem-solving: Use data insights to identify business problems, explore opportunities, and propose solutions to improve processes or decision-making.
- Adherence to Standards: Ensure compliance with data governance, security, and privacy regulations while handling sensitive or confidential data.
- Continuous Learning**:** Stay updated with the latest trends, tools, and advancements in the field of data analytics to enhance skills and techniques.

These responsibilities may vary based on the organization's specific requirements, but they form the core duties that a data analyst typically performs in their role.

DIFFERENCE BETWEEN DATA ANALYTICS, DATA ANALYSIS, AND DATA SCIENCE

Data analytics is a science that deals with analysis of data.

Data analysis, or Analysis of data is the process of inspecting, cleansing, transforming, and interpreting data to extract meaningful insights, patterns, and trends. It involves examining raw data to uncover information that can aid in decision-making, problem-solving, or understanding various phenomena within a specific context or domain.

Data science is like an umbrella that covers data analytics, statistics, machine learning etc.

In summary, data analytics is a scientific approach to solving data problems, data analysis focuses on examining and interpreting data to derive insights and make informed decisions, and data science involves a broader range of techniques, including predictive modeling, machine learning, and advanced analytics, to solve complex problems and generate valuable insights.

TYPES OF DATA ANALYTICS

There are primarily four types of data analytics, each serving a different purpose in extracting insights from data. These types are:

Descriptive Analytics:

Involves summarizing historical data to gain insights into past trends, patterns, and events.

Focuses on answering the question, "What has happened?"

Examples include generating reports, creating dashboards, and using Key Performance Indicators (KPI's) to track performance.

Diagnostic Analytics:

Aims to identify the reasons behind past events or trends.

Explores why certain patterns or anomalies occurred in the data.

Helps in understanding causation and correlation.

Seeks to answer the question, "Why did it happen?"

Examples include root cause analysis and exploratory data analysis (EDA).

Predictive Analytics:

Involves using statistical algorithms and machine learning techniques to forecast future trends or outcomes based on historical data.

Utilizes predictive models to anticipate what might happen.

Answers the question, "What is likely to happen?"

Examples include predictive modeling, regression analysis, and time series forecasting.

Prescriptive Analytics:

Focuses on recommending actions or strategies to optimize future outcomes based on predictive analysis.

Helps in decision-making by suggesting the best course of action.

Answers the question, "What should be done?"

Examples include optimization algorithms, decision support systems, and recommendation engines.

These four types of data analytics form a progressive approach, starting from understanding historical data (descriptive), investigating reasons behind trends (diagnostic), predicting future outcomes (predictive), and finally, recommending actions (prescriptive) based on those predictions.

THE DATA ANALYTICS PROCESS

The data analytics process involves a series of steps aimed at extracting valuable insights and knowledge from data. While specific methodologies may vary across industries and organizations, a typical data analytics process follows these fundamental stages:

- **Defining the Problem:** Understand the business problem or objective. Define clear goals and outcomes expected from the data analysis.
- **Setting a clear metric:** This is deciding what will be measured, and how it will be measured.

- **Data collection:** Gather relevant data from various sources such as databases, spreadsheets, apis, sensors, or external repositories.
- **Data preparation:** Cleanse, preprocess, and transform the raw data to ensure accuracy, consistency, and completeness. Handle missing values, remove duplicates, and format data for analysis.
- **Exploratory data analysis (eda):** Explore the dataset using statistical summaries, visualization techniques, and preliminary analysis to identify patterns, trends, and anomalies.
- **Data modeling:** Apply statistical methods, machine learning algorithms, or predictive models to analyze and interpret the data. This step involves hypothesis testing, regression, clustering, or other analytical techniques.
- **Interpreting results:** Evaluate the outcomes of data analysis. Interpret the findings, identify significant trends, insights, or correlations within the data.
- **Data Visualization and Reporting:** Present the analyzed data in a visually compelling manner using charts, graphs, dashboards, or reports. Communicate the insights effectively to stakeholders. May I add that properly presenting your findings is as important as the analysis itself.
- **Decision-Making and Implementation:** Use the insights gained from data analysis to make informed decisions, solve problems, or optimize strategies. Implement actionable recommendations derived from the analysis.
- **Monitoring and Iteration:** Continuously monitor the implemented strategies or solutions. Evaluate their performance and iterate the data analytics process as needed for continuous improvement.
- **Documentation and Communication:** Document the entire data analytics process, methodologies used, findings, and conclusions. Communicate the results to relevant stakeholders.

Adjustments to these steps might occur based on the specific requirements or nature of the analysis being conducted.

DATA ANALYSIS TOOLS

There are various data analysis tools available, each serving different purposes and catering to various aspects of the data analysis process. Some major data analysis tools widely used across industries include:

- **Microsoft Excel:** A widely used spreadsheet software with powerful data analysis capabilities, including functions, pivot tables, and charting tools.
- **SQL (Structured Query Language):** A querying (programming) language used to manage and manipulate relational databases for data querying, retrieval, and analysis.
- **Python**: A versatile programming language with rich libraries (such as Pandas, numpy, Matplotlib) used for data manipulation, statistical analysis, and machine learning.
- **R**: A programming language specifically designed for statistical analysis, data visualization, and data mining.
- **Tableau**: A data visualization tool that enables users to create interactive and shareable dashboards and visualizations from various data sources.
- **Power BI**: Microsoft's business analytics tool for data visualization, interactive dashboards, and business intelligence reporting.
- **Google Analytics**: A web analytics tool used to track website traffic, user behavior, and performance metrics.
- **SAS (Statistical Analysis System):** A software suite used for advanced analytics, multivariate analysis, predictive modeling, and data management.
- **MATLAB**: A programming and numeric computing environment used for data analysis, algorithm development, and statistical modeling.
- **SPSS (Statistical Package for the Social Sciences):** A software package for statistical analysis, data management, and reporting commonly used in social sciences research.

The choice of data analysis tools depends on the specific requirements, preferences, and complexity of analysis needed for a particular project or task. Each tool has its strengths and applications, and professionals often use a combination of tools to perform comprehensive data analysis.

STATISTICS

MEANING AND SCOPE OF STATISTICS

WHAT IS STATISTICS?

Statistics is a branch of mathematics and a scientific discipline concerned with collecting, organizing, analyzing, interpreting, and presenting data. It involves methods and techniques used to gather, summarize, and draw conclusions or inferences from data.

TYPES OF STATISTICS

There are two kinds of statistics

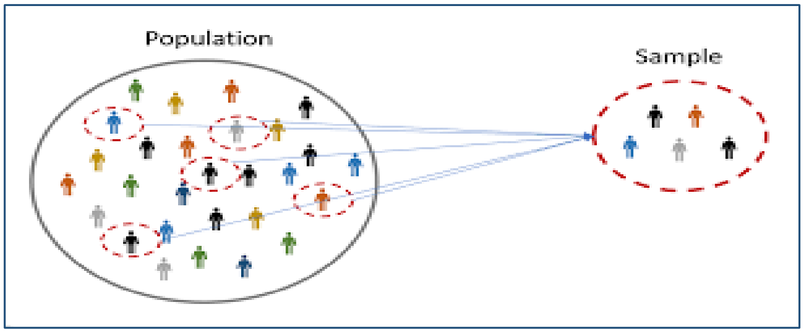
Descriptive Statistics: Describes and summarizes data through measures such as mean, median, mode, range, variance, and standard deviation. It provides insights into the characteristics of the dataset.

Inferential Statistics: Uses sample data to make inferences or predictions about a larger population. It includes hypothesis testing, confidence intervals, and regression analysis to draw conclusions beyond the available data.

Worthy to define a 'population' and 'sample'.

Population*: This is a set of all objects or units about which conclusions are to be drawn.*

Sample*: This is a part of a population. A subset of a population.*



Statistics finds applications in various fields such as science, economics, social sciences, business, engineering, and more. It helps in decision-making, identifying trends, making predictions, testing hypotheses, and providing insights based on empirical evidence derived from data analysis.

STATISTICAL DATA

Statistical data refers to a collection of information or observations gathered through systematic methods, often involving measurements, counts, or observations of various characteristics or phenomena. These data points are numerical, or qualitative representations of real-world entities, events, or attributes.

TYPES OF DATA

- **Qualitative Data**: Descriptive or categorical information that doesn't have a numerical value but rather represents qualities or characteristics. Examples include gender, color, marital status, or types of products.

  When such information is classifiable into two or more categories, they are referred to as *nominal data*. Data that can be ranked are called *ordinal data*.
- **Quantitative Data**: Numerical measurements that represent quantities or amounts. Examples include heights, weights, ages, temperatures, and income.

  Quantitative data can be further divided into two(2):

  i. ***Discrete Data****: Numerical data with specific values, often in whole numbers and countable. For instance, the number of students in a class or the number of cars in a parking lot.*

  ii. ***Continuous Data:*** *Numerical data that can take any value within a range. Examples include measurements like time, temperature, or height.*

CLASSES AND SOURCES OF DATA

Data can be classified based on various criteria such as its nature, origin, or format. Here are common classifications and sources of data:

Classes of Data:

- **Structured Data:** Data that adheres to a predefined format or structure. It's organized and easily searchable, often stored in databases. Examples include tables in relational databases or spreadsheets.
- **Unstructured Data:** Data that lacks a specific structure or organization. It doesn't fit neatly into traditional databases and includes text, images, videos, social media posts, and raw text files.
- **Semi-Structured Data:** Data that doesn't conform to a rigid structure like traditional databases but contains some organizational properties. Examples include XML files, JSON data, or NoSQL databases.

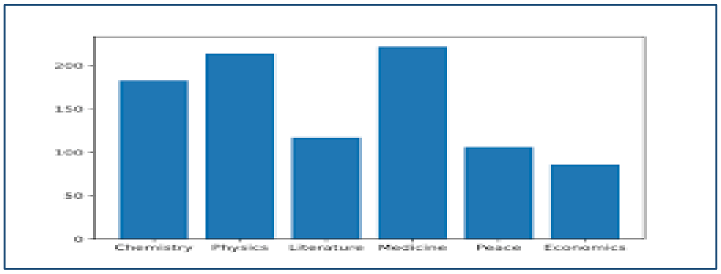
Sources of Data:

- **Primary Data:** Data collected firsthand through experiments, surveys, interviews, observations, or direct measurements specifically for the purpose of the current investigation or study.
- **Secondary Data:** Data that is obtained from existing sources such as books, journals, databases, reports, or other studies. It wasn't originally collected for the current research purpose.

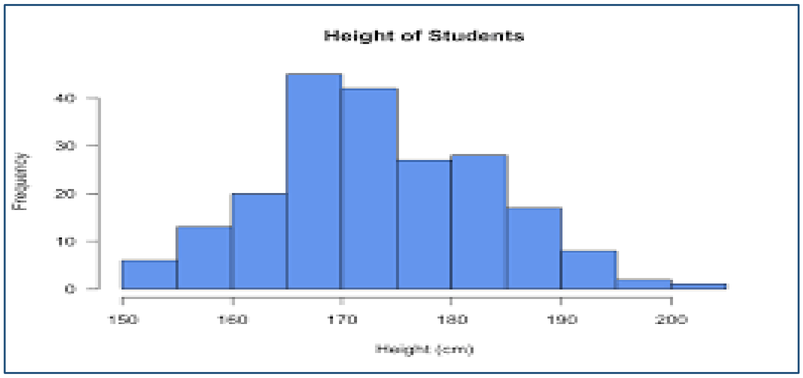
GRAPHICAL PRESENTATION OF DATA

There are several ways to graphically present data, each suitable for displaying different types of information and emphasizing specific aspects of the data. Some common methods for graphical representation of data include:

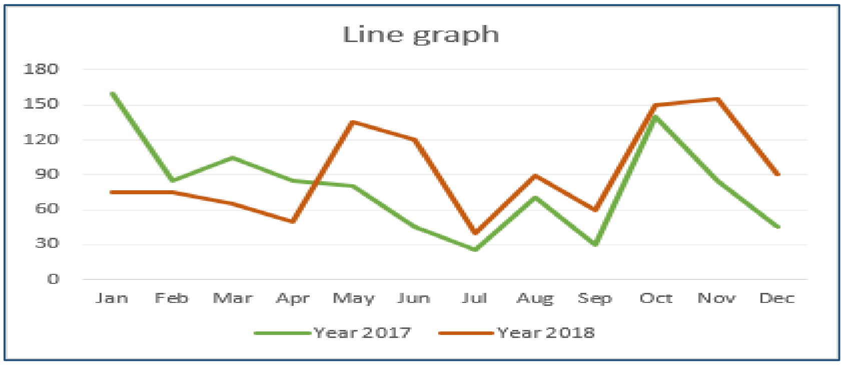
- **Bar Graphs:** Represent data using rectangular bars of varying lengths or heights to compare categories or quantities. Useful for comparing discrete categories or showing changes over time.



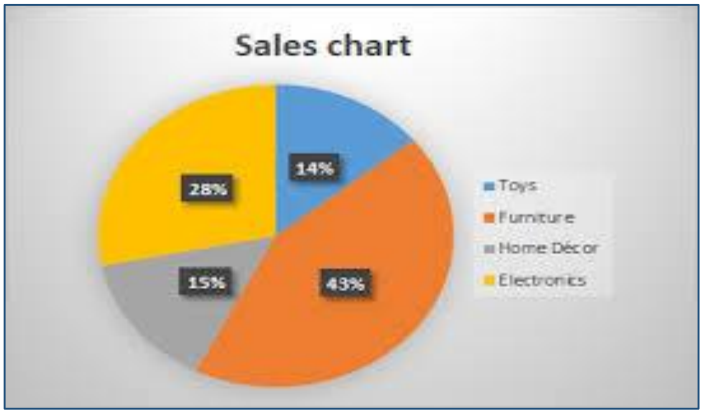
- **Histograms:** Display the distribution of continuous numerical data by dividing it into intervals (bins) and showing the frequency of values within each interval using bars. Useful for visualizing data distributions and identifying patterns.



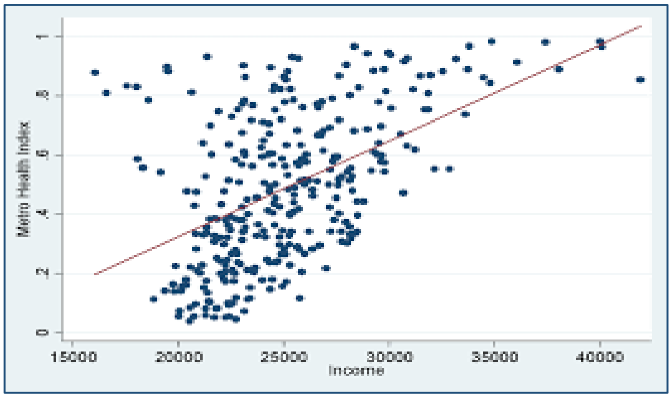
- **Line Graphs:** Use lines to connect data points, showing trends or changes over continuous intervals or time periods. Ideal for illustrating trends, relationships, or patterns in data.



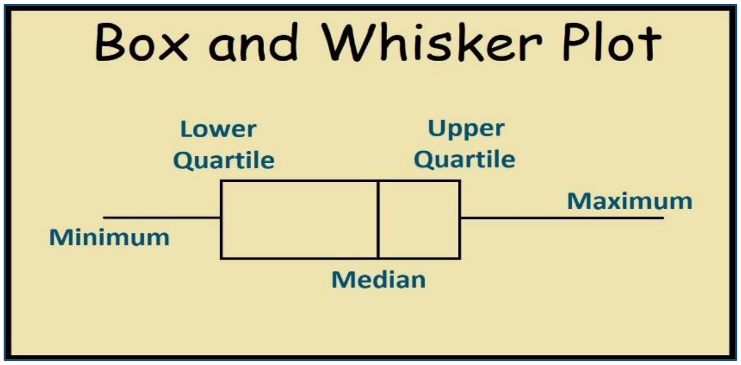
- **Pie Charts:** Present data as a circular graph divided into slices to represent the proportion or percentage of each category relative to the whole. Suitable for displaying parts of a whole and comparing percentages.



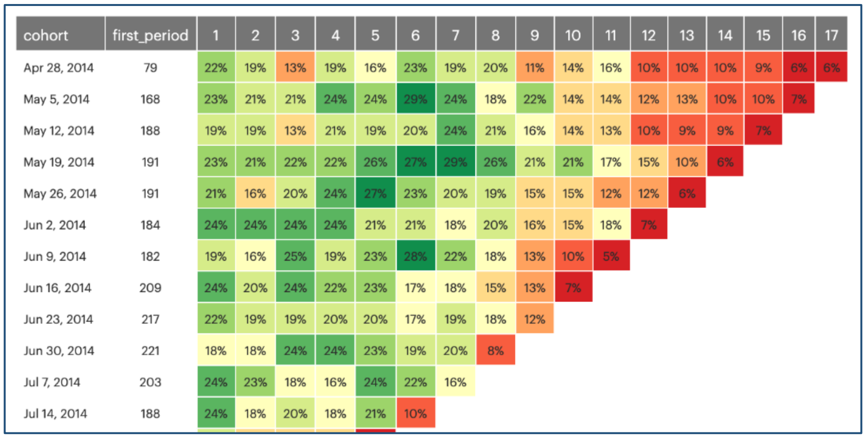
- **Scatter Plots:** Display individual data points as dots on a two-dimensional graph to visualize the relationship between two variables. Useful for identifying correlations or patterns between variables.



- **Box-and-Whisker Plots (Boxplots):** Represent the distribution of numerical data through quartiles, showing the median, quartiles, and outliers. Useful for identifying the spread and central tendency of data sets.

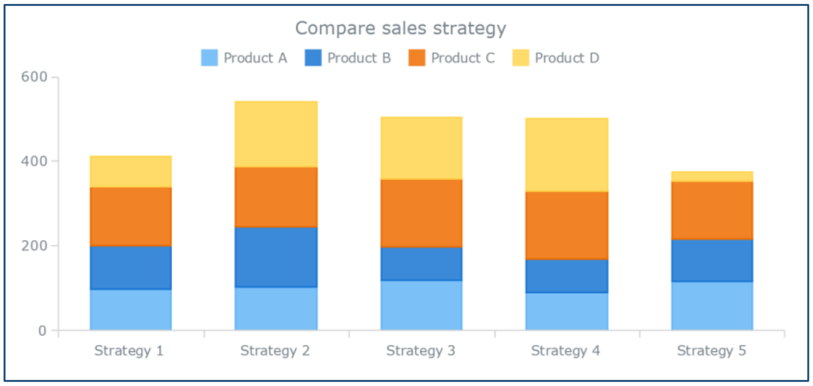


- **Heatmaps:** Use color variations to represent values in a matrix or grid. Useful for visualizing large datasets, correlations, or highlighting patterns.

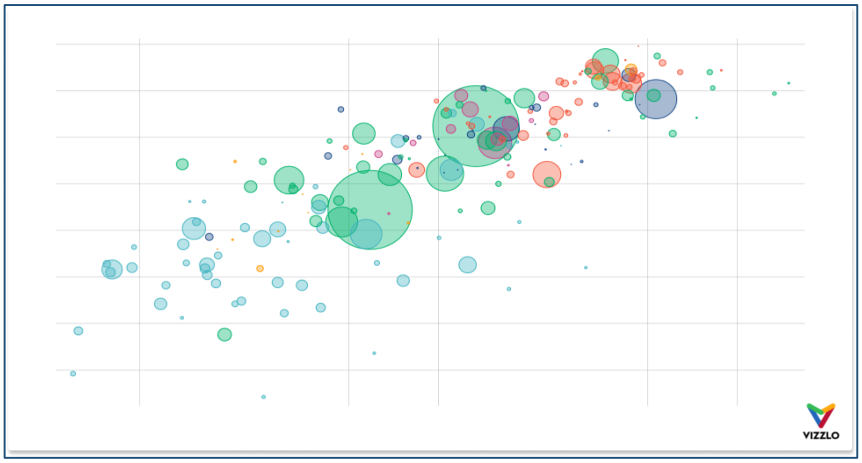


cohort	first_period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Apr 28, 2014	79	22%	19%	13%	19%	16%	23%	19%	20%	11%	14%	16%	10%	10%	10%	9%	6%	6%
May 5, 2014	168	23%	21%	21%	24%	24%	29%	24%	18%	22%	14%	14%	12%	13%	10%	10%	7%	
May 12, 2014	188	19%	19%	13%	21%	19%	20%	24%	21%	16%	14%	13%	10%	9%	9%	7%		
May 19, 2014	191	23%	21%	22%	22%	26%	27%	29%	26%	21%	21%	17%	15%	10%	6%			
May 26, 2014	191	21%	16%	20%	24%	27%	23%	20%	19%	15%	15%	12%	12%	6%				
Jun 2, 2014	184	24%	24%	24%	24%	21%	21%	18%	20%	16%	15%	18%	7%					
Jun 9, 2014	182	19%	16%	25%	19%	23%	28%	22%	18%	13%	10%	5%						
Jun 16, 2014	209	24%	20%	24%	22%	23%	17%	18%	15%	13%	7%							
Jun 23, 2014	217	22%	19%	19%	20%	20%	17%	19%	18%	12%								
Jun 30, 2014	221	18%	18%	24%	24%	23%	19%	20%	8%									
Jul 7, 2014	203	24%	23%	18%	16%	24%	22%	16%										
Jul 14, 2014	188	24%	18%	20%	18%	21%	10%											

- **Stacked Charts:** Display categories or parts of a whole as stacked segments in a single bar or area chart. Suitable for comparing totals across multiple categories.



- **Bubble Charts:** Show data in a scatter plot with varying sizes or colors of bubbles to represent three dimensions of data. Useful for visualizing relationships among three variables.



The choice of graphical representation depends on the nature of the data, the message to be conveyed, and the audience. Selecting the appropriate visualization method is crucial to effectively communicate insights and patterns inherent in the data.

MEASURES OF CENTRAL TENDENCY AND PARTITION

ARITHEMETIC MEAN

The arithmetic mean, commonly known as the average, is a measure of central tendency used to represent a set of numerical data. It is calculated by adding up all the values in a dataset and dividing the sum by the total number of values.

The formula for calculating the mean of a set of ‘n’ values is given by:

$$\text{Arithmetic Mean} = \frac{\text{(Sum of all values)}}{\text{(Number of all values)}}$$

Key points about the arithmetic mean:

- **Representative Measure:** It represents the typical value in a dataset and is sensitive to extreme values (outliers) in the dataset.
- **Application**: Widely used in various fields such as statistics, mathematics, economics, and sciences for summarizing and analyzing data.
- **Properties**: It is affected by changes in the dataset, as any addition, removal, or alteration of values affects the calculated mean.
- **Usage:** Provides a quick summary of data but might not always accurately represent the central tendency, especially if the dataset contains extreme values or is highly skewed.

The arithmetic mean is one of the most commonly used measures of central tendency, but it should be used cautiously, especially in situations where outliers significantly impact the overall average, leading to a skewed representation of the data.

MEDIAN

The median is a measure of central tendency in statistics that represents the middle value in a dataset when arranged in ascending or descending order. It divides the dataset into two equal halves, where half the values lie below and half lie above the median.

To find the median manually:

For Odd Number of Values:

Arrange the data in ascending or descending order and select the middle value as the median. For example, in the dataset {2, 5, 7, 9, 12}, the median is 7.

For Even Number of Values:

Take the average of the two middle values after arranging the data. For instance, in the dataset {3, 6, 8, 11}, the median is (6 + 8) / 2 = 7.

Key points about the median:

- **Robust Measure:** The median is less influenced by extreme values (outliers) compared to the mean. It provides a better representation of the central value in skewed datasets.
- **Application:** Widely used when there are outliers or when the distribution of data is not symmetrically distributed.
- **Properties:** The median does not consider the actual values but focuses on the position of values in the dataset.
- **Usage:** Useful when dealing with ordinal or interval data, or when a more robust measure of central tendency is needed.

The median is an essential measure of central tendency that provides a clearer understanding of the central value, especially in datasets where extreme values could significantly skew the arithmetic mean.

MODE

The mode in statistics refers to the value or values within a dataset that occur most frequently or have the highest frequency of occurrence. It is a measure of central tendency used to identify the most common or prevalent value(s) in a set of data.

Key points about the mode:

- **Multiple Modes**: A dataset can have one mode (unimodal), more than one mode (multimodal) if multiple values have the same highest frequency, or no mode if all values occur with equal frequency.
- **Application**: Useful for categorical or discrete data analysis, such as identifying the most common category in a dataset.
- **Robustness:** Unlike the mean and median, the mode does not take into account the magnitude of values; it solely focuses on identifying the value(s) with the highest frequency of occurrence.

- **Usage**: Commonly used in various fields, especially when dealing with nominal or categorical data, for instance, finding the most common color, type, or category.

The mode helps in understanding the central tendency of a dataset by highlighting the values that occur most frequently, providing insights into the most prevalent characteristics or categories within the data.

QUARTILES

Quartiles are values that divide a dataset into four equal parts, each representing 25% of the data when arranged in ascending or descending order. These values help analyze the distribution of numerical data by dividing it into quarters.

There are three quartiles:

- First Quartile (Q1): The value below which 25% of the data points lie. It separates the lowest 25% of the dataset from the rest.
- Second Quartile (Q2): Also known as the median. It divides the dataset into two halves, with 50% of the data points below and 50% above this value.
- Third Quartile (Q3): The value below which 75% of the data points lie. It separates the lowest 75% of the dataset from the highest 25%.

To manually calculate quartiles:

Arrange the dataset in ascending or descending order.

Find the median (Q2).

Calculate Q1: Find the median of the lower half of the dataset.

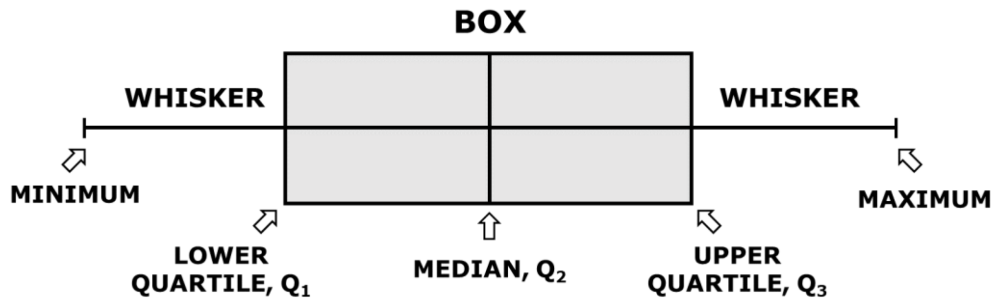
Calculate Q3: Find the median of the upper half of the dataset.

Quartiles are valuable in statistical analysis to understand the spread and distribution of data, especially in box plots, where the range between Q1 and Q3 represents the interquartile range (IQR), showing the middle 50% of the data. They help identify potential outliers, measure variability, and assess the dispersion of values within a dataset.

BOX AND WHISKERS PLOT (BOX PLOT)

A box plot, also known as a box-and-whisker plot, is a graphical representation that displays the distribution, central tendency, and variability of numerical data through five summary statistics: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.

Figure 4.5.2.1 Building a box and whisker plot



Components of a Box Plot:

- **Minimum and Maximum (Whiskers):** Lines extending from the box's ends to the minimum and maximum values within the dataset that are not outliers.
- **Box (Interquartile Range - IQR):** Represents the middle 50% of the data, stretching from Q1 (lower quartile) to Q3 (upper quartile). The length of the box denotes the range of the middle 50% of the dataset.
- **Median (Line inside the box):** Represents the central tendency or the middle value of the dataset.
- **Outliers (Data Points outside Whiskers):** Individual data points that fall significantly above or below the whiskers, considered as potential outliers.

Box plots are useful for comparing distributions of different datasets, identifying central tendency, detecting outliers, and visualizing the spread of numerical data. They offer a clear and concise summary of the data's variability and central values in a graphical format.

MEASURES OF DISPERSION

THE VARIANCE

Variance is a statistical measure that quantifies the spread or dispersion of a set of data points around the mean (average) of the dataset. It indicates how much the data points deviate from the mean value.

Key points about variance:

- **Spread of Data**: Variance measures the average squared distance of data points from the mean. Larger variance indicates greater dispersion, while smaller variance indicates less dispersion around the mean.
- **Squared Deviations:** The calculation involves squaring the differences between individual data points and the mean, which emphasizes larger deviations and disregards the sign of the deviations.
- **Units Squared:** Variance is in squared units of the original data, which might not be easily interpretable in the same units as the data.
- **Standard Deviation**: The square root of the variance is the standard deviation, a widely used measure of dispersion that shares the same unit as the original data.

  Variance is a fundamental statistical concept used to understand the variability and distribution of data points in a dataset. It provides valuable insights into the data's spread and variability around the mean value.

THE STANDARD DEVIATION

The standard deviation is a statistical measure that quantifies the amount of variation or dispersion of a set of data points relative to the mean (average) of the dataset. It measures how spread out the values are from the average value, providing insight into the data's variability.

The standard deviation (σ for a population or s for a sample) is calculated as the square root of the variance. It is a square root of the average squared differences between each data point and the mean of the dataset.

Key points about standard deviation:

- **Measure of Dispersion:** Standard deviation measures the average distance of data points from the mean. Higher standard deviation indicates greater dispersion, while lower standard deviation indicates less variability around the mean.
- **Sensitivity to Outliers**: Like variance, standard deviation is sensitive to outliers as it involves squaring the deviations from the mean.
- **Interpretability**: Standard deviation is expressed in the same units as the original data, making it more interpretable and providing a better understanding of the data's spread compared to variance.

Standard deviation is widely used in various fields of study, including statistics, science, finance, and social sciences, to analyze and interpret the variability and distribution of data points within a dataset.

THE RANGE

The range in statistics refers to the difference between the highest and lowest values within a dataset. It is a simple measure of dispersion that provides the span or spread covered by the data points, specifically the difference between the maximum and minimum values.

Key points about the range:

- **Simple Measure**: The range is easy to compute and understand, providing a quick overview of how widely dispersed the data is in terms of the highest and lowest values.
- **Sensitivity to Outliers**: The range can be greatly influenced by outliers (extreme values) in the dataset, as it solely depends on the maximum and minimum values.
- **Limited Information**: While the range gives an idea of the spread of data, it does not consider the distribution of values between the extremes or provide information about the central tendency or variability within the dataset.

The range is a basic and straightforward measure of dispersion, primarily useful for providing a quick insight into the spread or variability of a dataset, especially when dealing with small or simple datasets. However, it might not capture the full extent of variability present in larger or more complex datasets due to its sensitivity to outliers.

THE MEAN DEVIATION

The mean deviation, also known as the mean absolute deviation (MAD), is a measure of dispersion that calculates the average absolute differences between individual data points and the mean of a dataset. It quantifies how much, on average, each data point varies from the mean value of the dataset.

Key points about mean deviation:

- **Absolute Differences:** Unlike variance and standard deviation that square the differences from the mean, mean deviation uses absolute differences (that is, for each number in the group, we look at how far away it is from the average of all the numbers (the middle value).), disregarding the sign of deviations.
- **Interpretability**: Mean deviation provides a measure of dispersion in the same units as the original data, making it more interpretable compared to variance or standard deviation.
- **Sensitivity to Outliers**: Mean deviation, similar to range, is sensitive to extreme values, as it considers the absolute differences, making it potentially less robust against outliers.

Mean deviation is a measure of dispersion that describes the average absolute deviation of data points from the mean. While it offers interpretability and simplicity, it might not be as commonly used as variance or standard deviation due to its sensitivity to extreme values and less efficient handling of squared deviations.

In simpler terms, mean deviation helps us understand, on average, how much the numbers in a group differ from their average value, without considering if they're higher or lower than that average.

CORRELATION AND REGRESSION

CORRELATION ANALYSIS

Correlation analysis is a statistical technique used to measure and describe the **strength** and **direction** of the relationship between two variables. It helps in understanding how changes in one variable are associated with changes in another variable.

Correlation analysis is like looking at how two things are related. Imagine you have two friends: whenever one friend does something, the other friend reacts in a certain way.

For example:

If one friend gets excited about something, the other friend also gets excited.

If one friend feels sad, the other friend also feels sad.

Correlation analysis is like checking if there's a pattern in how your friends' moods match up. If they tend to feel the same way a lot, we'd say they have a strong "correlation" in their moods. If their feelings don't match up much, we'd say there's not much "correlation" between their moods.

In a similar way, in statistics, correlation analysis checks how two things, like hours of study and exam scores or temperature and sales, relate to each other. It helps us see if there's a strong connection, a weak one, or maybe no connection at all between different things we're studying.

Correlation Coefficient: This is a numerical measure that ranges between -1 and +1, representing the strength and direction of the relationship between two variables.

A correlation coefficient of +1 indicates a **perfect positive correlation**, meaning that as one variable increases, the other variable also increases proportionally.

A correlation coefficient of -1 represents a **perfect negative correlation**, suggesting that as one variable increases, the other variable decreases proportionally.

A correlation coefficient close to zero (0) signifies a **weak or no linear relationship** between the variables.

Types of Correlation:

Positive Correlation: When one variable increases, the other variable also tends to increase. The correlation coefficient is positive (closer to +1). Example, suppose we want to examine the relationship between the number of hours studied for an exam and the exam scores obtained by a group of students.

Scenario: As the number of hours studied increases, do the exam scores tend to increase?

Correlation Result: If we find a positive correlation close to +1 between hours studied and exam scores, it indicates that as students spend more time studying (one variable increases), their exam scores also tend to increase (the other variable increases).

Negative Correlation: When one variable increases, the other variable tends to decrease. The correlation coefficient is negative (closer to -1). Example, Consider the relationship between outdoor temperature and sales of winter clothing in a retail store.

Scenario: As the temperature outside decreases (gets colder), do the sales of winter clothing tend to increase?

Correlation Result: If there's a negative correlation closer to -1 between outdoor temperature and winter clothing sales, it suggests that as the temperature drops (one variable decreases), sales of winter clothing tend to rise (the other variable increases).

No Correlation: When there is no systematic relationship between the variables. The correlation coefficient is close to zero (0). Let's take an example of the relationship between shoe size and intelligence level.

Scenario: Is there a relationship between a person's shoe size and their intelligence level?

Correlation Result: If we find a correlation coefficient close to zero (0) between shoe size and intelligence, it suggests no meaningful relationship between the variables. There's no evidence to support that having a larger or smaller shoe size affects intelligence level.

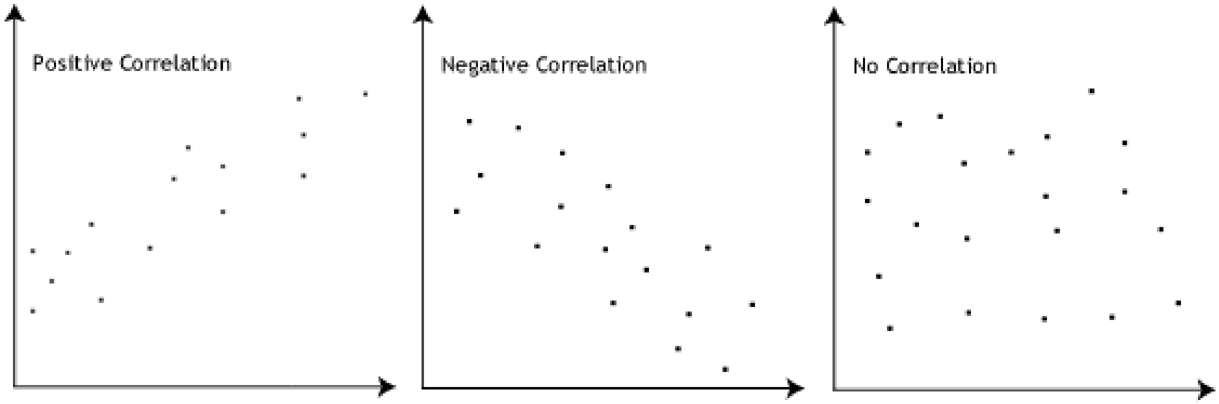


Fig. 1: Scatter plot/diagram showing the types of correlation

It is important to note that correlation does not imply causation. Even if two variables are strongly correlated, it doesn't necessarily mean that one causes the change in the other. Other factors or variables might be influencing the relationship. For example, on days when more umbrellas are sold, there tends to be more rainfall.

If you observe sales data and rainfall records, you might find a positive correlation: higher umbrella sales correlate with increased rainfall.

However, correlation does not imply causation. The relationship between umbrella sales and rainfall doesn't mean that buying more umbrellas causes it to rain or vice versa.

The correlation is due to a common factor: the weather. Rainy days prompt more people to buy umbrellas because they need them to stay dry. Therefore, the increase in umbrella sales is a response to the rainfall, not the cause of it.

In conclusion, correlation analysis helps in understanding and quantifying relationships between variables, whether positive, negative, or non-existent, aiding in drawing insights from data and making informed decisions, it does not necessarily imply causation.

Read more here

Scatter Diagram/Plot

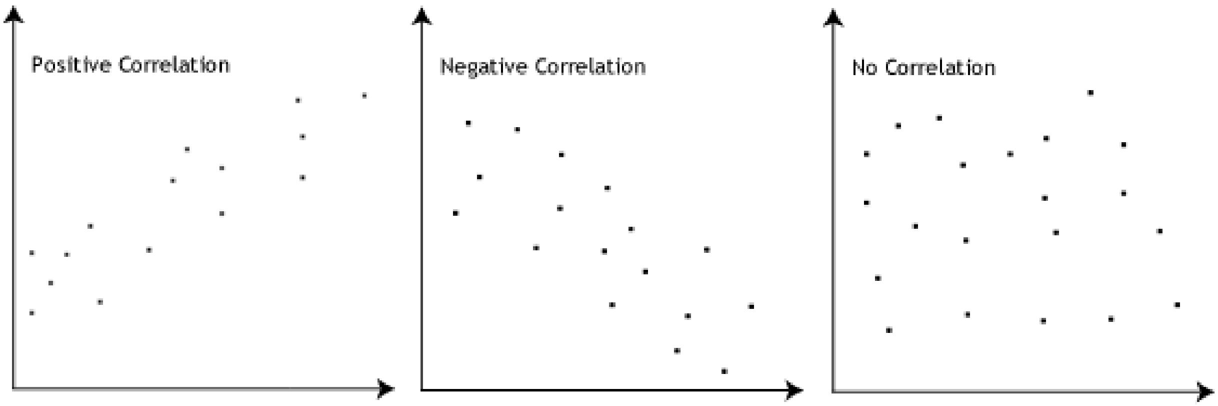
A scatter plot, also known as a scatter diagram or scatter graph, is a type of data visualization that displays the relationship or association between two numerical variables. It helps to visualize how changes in one variable are related to changes in another variable.

Explanation:

- **Horizontal Axis (X-axis):** It represents one variable, usually the independent variable, and is plotted along the horizontal axis.
- **Vertical Axis (Y-axis):** It represents the other variable, often the dependent variable, and is plotted along the vertical axis.
- **Data Points**: Each data point on the plot represents a pair of values—one for the horizontal variable (X-axis) and one for the vertical variable (Y-axis). These points are plotted at the intersection of their respective values.
- **Scatter of Points:** When you plot multiple data pairs on the graph, you create a scatter of points across the plot, and the pattern they form can reveal the relationship between the variables.

Interpretation:

- **Correlation Analysis:** By observing the pattern of the points, you can visually assess whether there's a relationship between the two variables.
- **Positive Correlation**: If the points tend to form an upward trend as you move from left to right, it indicates a positive correlation.
- **Negative Correlation**: If the points show a downward trend as you move from left to right, it suggests a negative correlation.
- **No Correlation**: If the points appear randomly scattered without any particular trend, it signifies no correlation between the variables.



Usage:

Scatter plots are widely used in various fields like statistics, data analysis, sciences, social sciences, and economics to visualize and understand relationships between variables.

They help in identifying patterns, trends, clusters, or outliers within the data.

For example, in a scatter plot of "hours studied" versus "exam scores," each point represents a student's study hours on the X-axis and their corresponding exam score on the Y-axis. Observing the pattern of points helps to understand if more study hours relate to higher exam scores (positive correlation), lower exam scores (negative correlation), or no particular relationship (no correlation).

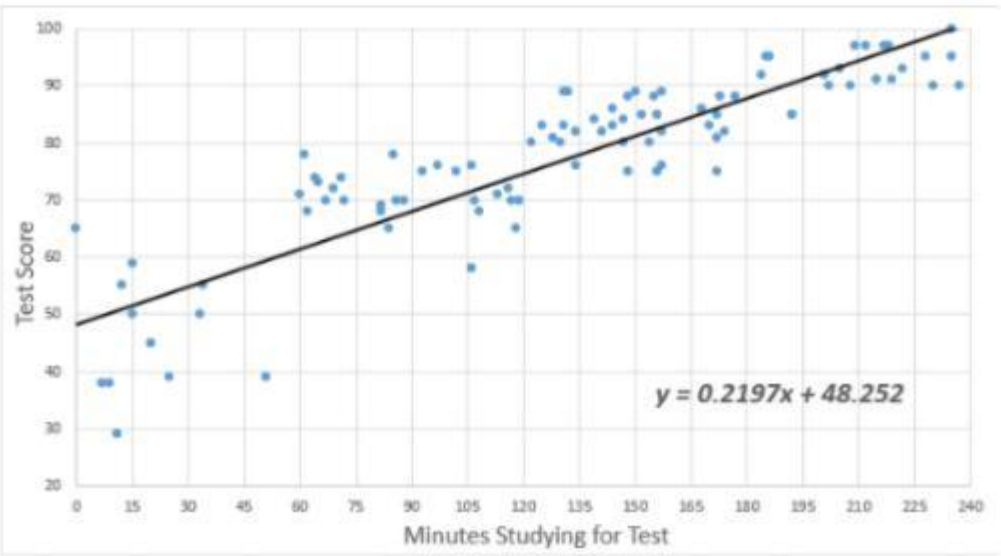


Fig 2. Scatter plot of Minutes studying for test and test score

In conclusion, a scatter plot is a graphical representation used to visualize the relationship between two variables by displaying their data points, allowing for a quick assessment of correlation or patterns within the data.

REGRESSION ANALYSIS

Regression analysis is a statistical method used to examine and quantify the relationship between a dependent variable and one or more independent variables. It helps in predicting and understanding how changes in the independent variables are associated with changes in the dependent variable.

Imagine you're trying to figure out how the number of hours you spend studying relates to the grades you get in exams, regression Analysis is like finding a line that best shows how changes in your study hours might affect your exam grades.

This line helps predict how your grades might change based on how much time you spend studying.

Explanation:

- **Dependent Variable**: It's the variable that we want to predict or understand. It's also called the response variable.
- **Independent Variable(s**): These are the variables used to predict or explain the behavior of the dependent variable. They are also known as predictor variables.
- **Regression Line**: Regression analysis calculates a line (or curve) that best fits the data points on a scatter plot, indicating the relationship between the dependent and independent variable(s).

There are various types of regression, but we will restrict our studies to just two types:

Simple Linear Regression: When there is only one independent variable.

Multiple Linear Regression: When there are multiple independent variables.

Example 1:

Let's consider an example of predicting house prices based on the size (in square feet) of the house.

Dependent Variable: House Price (the variable we want to predict).

Independent Variable: House Size (the variable used to predict the price).

Simple Linear Regression:

We collect data on house sizes and their corresponding prices.

Using regression analysis, we create a regression line that best fits the data points on a scatter plot, showing the relationship between house size and price.

The regression line helps us predict the price of a house based on its size. For instance, if a house is 2,000 square feet, the regression analysis could estimate its price.

Example 2:

If you spend more hours studying (independent variable), regression analysis helps predict if your grades will likely be higher or lower (dependent variable).

It finds the pattern in your study hours and grades to give you a good guess about your future grades based on your study time.

Usage:

People use regression analysis in lots of areas to understand how one thing (like study time) affects another (like exam grades).

It's used in things like predicting sales based on advertising spending, understanding how temperature affects ice cream sales, or even predicting house prices based on size.

Regression analysis is used in various fields like economics, finance, sciences, social sciences, and machine learning for predictive modeling, forecasting, and understanding relationships between variables.

Interpretation:

By analyzing the regression results (slope, intercept, coefficients), we can understand the strength, direction, and significance of the relationship between variables and make predictions based on this relationship.

EXCEL FOR DATA ANALYTICS

BASIC EXCEL SKILLS

UNDERSTANDING EXCEL INTERFACE

Ribbon**:** Toolbar divided into tabs like Home, Insert, Formulas, etc., containing commands.

Cells**:** Intersection of rows and columns where data is entered.

Columns & Rows: Columns are labeled with letters (A, B, C...), rows are labeled with numbers (1, 2, 3...).

Sheet Tabs**:** At the bottom, allowing multiple sheets within a workbook.

Navigation:

Moving Around: Use arrow keys or click on cells to navigate.

Sheet Navigation**:** Click on sheet tabs to switch between different sheets.

Scrolling**:** Scroll bars or mouse wheel for vertical/horizontal movement.

DATA ENTRY, FORMATTING, AND CELL REFERENCING

Data Entry:

Typing Data: Enter data directly into cells.

Copy/Paste: Use Ctrl + C (Copy) and Ctrl + V (Paste) for data duplication.

Formatting:

Cell Formatting: *Right-click cells* > *Format Cells* for customizing fonts, alignment, borders, etc.

Number Formatting: Change data type (currency, date, percentage etc) through Number Format options.

Cell Referencing:

Relative Referencing: Referring to cells by their relative position (e.g., A1).

Absolute Referencing: Fixing cell references (A1) to prevent them from changing when copied.

BASIC FORMULAE AND FUNCTIONS

Basic Formulas:

- **SUM:** Adds values in a range:
  Type the following in a cell ***=SUM(A1:A10*****).**
- **AVERAGE**: Calculates the average of values:
  Type the following in a cell ***=AVERAGE(A1:A10).***
- **COUNT**: Counts numbers in a range:
  Type the following in a cell ***=COUNT(A1:A10).***

Basic Functions:

IF Function: Makes logical comparisons: *=IF(condition, value_if_true, value_if_false).*

VLOOKUP: Searches for a value in the first column and returns a value in the same row from another column.

SORTING, FILTERING, TABLES AND NAMED RANGES

Sorting:

Sorting Data: *Select data* > *Data tab* > *Sort* to arrange data based on a particular column.

Custom Sorting: Sort by multiple columns or custom criteria.

Filtering:

Autofilter: *Data tab* > *Filter* for applying filters to columns.

Filtering Options: Filter by specific criteria or custom filters.

Tables:

Excel's tables are powerful tools that allow users to efficiently manage and analyze data.

Creating a Table:

1. Select Data Range:
    - To create a table, start by selecting the range of cells containing your data.
2. Insert Table:
    - Go to the "Insert" tab on the Excel ribbon.
    - Click on "Table."
3. Confirm Range:
    - Excel will automatically detect the range based on your selection. Confirm the range and check the box if your data has headers.
4. Table Design:
    - Excel will convert the selected range into a table, applying a default table design.

Benefits of Using Tables:

1. Dynamic Range:
    - Tables automatically expand to accommodate new data, ensuring your range is always up-to-date.
2. Readability:
    - Banded rows and frozen headers enhance the readability of large datasets.
3. Ease of Formulas:
    - Structured references simplify the creation of formulas, improving formula clarity and reducing errors.
4. Filtering and Sorting:
    - Quickly filter and sort data within the table without affecting the surrounding data.

5. Total Row Calculations:
    - Easily calculate totals, averages, or other functions using the Total Row feature.
6. Data Validation:
    - Apply data validation rules directly to a column within the table.

Tips for Working with Tables:

1. Naming Tables:
    - Consider giving your table a meaningful name using the "Table Tools Design" tab. This name can be used in formulas.
2. Ctrl + T Shortcut:
    - Highlight your data range and use the shortcut Ctrl + T to quickly convert it into a table.
3. Table Styles:
    - Experiment with different table styles to find the one that suits your preferences and enhances readability.

Named Ranges:

Named ranges in Excel provide a powerful way to organize and manage data in spreadsheets. Instead of referring to cells by their coordinates (like A1 or B2), you can assign a meaningful name to a range of cells. This enhances clarity, simplifies formula creation, and improves the overall usability of your Excel workbooks.

Creating Named Ranges:

1. Using the Name Box:
    - Select the range of cells you want to name.

- Look at the left of the formula bar; there is a box called the "Name Box." Click on it and type a name for your range.

2. Using the Name Manager:

   - Go to the "Formulas" tab and click on "Name Manager."
   - Click "New" and define the name, scope, and the range of cells. You can also add a comment for reference.

Advantages of Named Ranges:

1. **Clarity and Readability:**

   - Named ranges make formulas more readable and reduce the likelihood of errors. Instead of referencing a cell like A1, you can use a name that describes the data, like "SalesData" or "Expenses."

2. **Ease of Navigation:**

   - Navigating large spreadsheets becomes more manageable when using named ranges. Instead of scrolling through columns and rows, you can jump directly to a named range.

3. **Dynamic Formulas:**

   - Named ranges can be dynamic. If your data expands or contracts, you can adjust the named range to automatically include new data without modifying your formulas.

4. **Simplified Formula Creation:**

   - When writing formulas, typing the name of a range is often faster than selecting cells manually. This is especially true for complex formulas.

5. **Enhanced Data Analysis:**

- Named ranges are valuable for PivotTables and charts. You can easily reference named ranges when setting up dynamic data ranges for these elements.

Best Practices:

1. **Choose Descriptive Names:**
    - Pick names that clearly describe the content or purpose of the range. This enhances understanding, especially when sharing workbooks with others.
2. **Avoid Spaces and Special Characters:**
    - While Excel allows spaces and certain special characters in named ranges, it's best to use underscores or camelCase to ensure compatibility and prevent errors.

Data Visualization:

Data visualization is the graphical or pictorial representation of data. You can create visualizations in excel by:

Charts: Insert tab > Charts for creating visual representations (bar, pie, line charts, etc.).

INTERMEDIATE EXCEL SKILLS

ADVANCED FUNCTIONS: VLOOKUP, INDEX-MATCH etc

VLOOKUP Function:

Looks for a value in the first column of a table and returns a value in the same row from another column.

Syntax: =*VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup]).*

INDEX-MATCH Function (Alternative to VLOOKUP):

INDEX returns the value of a cell in a specified row and column.

MATCH returns the relative position of an item in a range.

Combining them creates a powerful lookup: =*INDEX(return_range, MATCH(lookup_value, lookup_range, 0)).*

CONDITIONAL STATEMENTS:

Conditional statements in Excel are powerful tools that allow users to create logical tests, make decisions based on those tests, and automate data analysis and processing. These statements can be implemented through functions like IF, and nested IF statements.

IF Statement:

The IF statement is fundamental and enables you to perform different actions based on a specified condition.

Syntax:

=IF(logical_test, value_if_true, value_if_false)

- ***logical_test****: The condition you want to evaluate.*
- ***value_if_true****: The result if the condition is true.*
- ***value_if_false****: The result if the condition is false.*

Example:

=IF(A1>10, "Yes", "No")

Nested IF Statements:

Nested IF statements allow for more complex decision-making by embedding one IF statement inside another.

Syntax:

=IF(logical_test1, value_if_true1, IF(logical_test2, value_if_true2, value_if_false2))

Example:

=IF(A1>10, "High", IF(A1>5, "Medium", "Low"))

DATA CLEANING TECHNIQUES

Data cleaning/cleansing or data scrubbing is simply the process of identifying and correcting anomalies or irregularities in data. These anomalies can range from missing data, inconsistent data, invalid data down to duplicate data. The data cleaning process is relatively not a linear process, and neither is it straightforward, rather, the approach you employ is dependent on the kind of error that is prevalent in your data.

However, we can employ three (3) steps in this process, and they are:

Step A. Find/identify the problem.

Step B. Solve/correct the problem.

Step C. Repeat steps A and B.

Step A. Find/identify the problem.

Look for the following:

- Are there rows/columns with empty values?
- What data is missing and why?

- How is the data distributed? (That is, is it normally distributed or not. *Read more about the distribution of data here*)
- Is there consistency in my data? Are the dates written in the same format?
- Are there columns/rows that contains invalid data?

These and more are the possible questions to ask to check if there are errors in your data or not.

Step B. Solve/correct the problem.

Depending on the kind of data problem you have, you'll need to employ varying strategies.

The possible data problems can be placed in different categories, and they are:

1. INVALID DATA
2. INCONSISTENT DATA
3. MISSING DATA
4. DUPLICATE DATA
5. CONTAMINATED DATA
6. OUTLIERS
7. DATA TYPE ERRORS

1. **INVALID DATA**: Invalid data are data that are illogical. For example, in an age column, you find someone who is aged 255, or in an exam column, you see someone scoring 500 where the exam limit is 100. Invalid data can result from data entry errors. a step towards understanding your data will aid in correcting this error, for instance, it could be that in the first scenario, the person inputting the data wanted to input 25, but pressed the 5 twice and this resulted to 255, same applies to the second case, or this error can arise from functions and transformations that were made earlier to the data. You can drop such rows in the case of uncertainty, and another way of correcting this can be to amend the functions and transformations which caused the data to be invalid.

2. **INCONSISTENT DATA:** This can arise when you have different formats of the same thing in a column. E.g., 'BOOKS', 'Books', 'Book', etc. In this case, as a human being, I understand that even though all the entries are not in the same format, they mean the same thing - 'Books', but computers don't understand so, that's the only place where they act dumb, they consider the entries with different formats to be different. It (computer) gives you what you feed it.

   Inconsistency is something common in data sets. The best way to spot inconsistent representations of the same elements in your database is to visualize them.

   Plot a bar chart of the column together with its frequency (count), when you spot the inconsistency, standardize all the entries/elements into the same format.

3. **MISSING DATA:** There are different ways of handling missing data, but the three(3) main approaches to handling it:
   - Drop rows/columns with missing (empty cells) data.
   - Recode missing data into a different format.
   - Fill in missing values with "best guesses". That is, you can estimate the most probable value of data at that point. You can fill with the succeeding or preceding values, you can also fill with the mean of the column (in the case of a numeric column).

4. **DUPLICATE DATA:** This means having the same values repeating for an observation point. This can be costly to your analysis because it can either deflate/inflate our numbers (we can record more sales than there are, or the profit made from a campaign changes because values were entered multiple times). You can handle duplicate data by:
   - Finding the same records and deleting all but one.
   - Pairwise match records, comparing them and taking the most relevant one (most recent one).

5. **CONTAMINATED DATA:** Example of a contaminated data is having your exam score record in an age column/data set, or having a sales record in an address column/data set. With a contaminated or corrupt data, there is not much that can be done except for removing such records. This requires a lot of domain knowledge to do.

6. **OUTLIERS:** Outliers are data points which are at an extreme. They are usually very high or very low values. Outliers usually signify either very interesting behavior or a broken collection process. We can deal with outliers by:

    - Remove outliers from the analysis. Having outliers can ruin your analysis, and since the mean is greatly affected by it, it can bring the mean up or down, and by so doing distort your statistics.
    - Segment data so that outliers are in a different group. Place all the normal-looking data in one group and outliers in another.
    - Keep outliers, but use a different statistical method for analysis. Weighted means (which places more weight on the "normal" part of the data) can be used for such analysis without suffering the negative consequence of outliers.

7. **DATA TYPE ERRORS:** Depending on which data type you are working with (Objects, strings, integers, DateTime, floats, decimals), problems specific to data types can be encountered. A way of solving this is by standardizing the column/data set or converting the column/data set to the correct data type.

Step C. Repeat steps A and B.

Once cleaned, repeat steps A and B, this is helpful because:

- You might have missed something. Cleaning again helps you correct the things you missed.

- Through cleaning, you learn more about your data set. Every time you sweep through your dataset and look at the distributions of values, you learn more about your data, and this gives you hunches as to what to question and analyze.

Reports have it that data scientists spend 80% of their time cleaning and organizing data because of the associated benefits. Clean data is the pillar upon which data-driven decision-making rests.

If you don’t take out time to clean and organize you data you lose more time. To gain speed in your analysis, spend more time in preparing your data.

DATA TRANSFORMATION

Data transformation in Excel is a crucial step in the data analysis process. It involves restructuring and manipulating data to make it suitable for analysis. Excel provides a variety of tools and functions to perform data transformations, allowing users to clean, organize, and derive insights from their data.

Common Data Transformation Techniques:

1. Text to Columns:
    - Use the "Text to Columns" feature to split text into multiple columns based on a delimiter. This is helpful when dealing with data that needs to be separated into distinct fields.
2. Transpose:
    - The "Transpose" function allows you to switch rows to columns and vice versa. This is useful when the orientation of your data needs to be changed.
3. Concatenate:
    - Concatenate function combines text from multiple cells into one cell. This is beneficial when merging data from different columns.
4. Find and Replace:

- The "Find and Replace" tool helps in locating specific values and replacing them with others. This is handy for cleaning data by correcting errors or standardizing formats.

5. Filter and Sort:

   - Utilize Excel's filter and sort functions to organize and arrange data based on specific criteria. This is essential for better data visibility and analysis.

6. PivotTables:

   - PivotTables are powerful tools for summarizing and analyzing data. They allow you to rearrange and aggregate data dynamically, providing a more insightful view of your dataset. We will talk more about this in the next section.

7. Formulas and Functions:

   - Excel's functions like IF, VLOOKUP, HLOOKUP, and INDEX-MATCH are powerful for data transformation. They can be used to derive new information, categorize data, or perform calculations based on specific conditions.

8. Remove Duplicates:

   - Excel provides a "Remove Duplicates" feature to eliminate duplicate values in a dataset. This is crucial for ensuring data quality and preventing redundancy.

Best Practices:

1. Backup Your Data:

   - Before starting any data transformation, create a backup of your dataset to avoid accidental loss of data.

2. Document Your Steps:

   - Keep a record of the steps you take during data transformation. This documentation aids in reproducing analyses and ensures transparency.

3. Use Consistent Formats:

   - Maintain consistent formatting throughout your dataset to avoid discrepancies and errors in analysis.

4. Automate Repetitive Tasks:

   - Leverage Excel's automation features, such as macros, to streamline repetitive data transformation tasks.

5. Test and Validate:

   - Regularly test and validate your transformations to ensure that the output aligns with your expectations. This is crucial for accurate analysis.

One may ask: **What is the difference between Data Cleaning and Data Transformation?** There is a difference between the two. **Data cleaning**, involves the process of identifying and correcting errors or inconsistencies in a dataset. It focuses on resolving issues like missing values, outliers, duplicates, and inaccuracies to ensure data accuracy and reliability. While, **Data transformation** involves converting and reformatting raw data into a structure that is suitable for analysis. It includes tasks such as reshaping data, aggregating information, creating new variables, and organizing data in a way that facilitates meaningful analysis.

The primary goal of **data cleaning** is to improve data quality by addressing issues that might hinder accurate analysis. It ensures that the dataset is free from errors and discrepancies. Whereas the primary purpose of **data transformation** is to prepare and organize data in a way that makes it more suitable for analysis. It involves shaping the data to meet specific requirements, creating derived variables, eliminating irrelevant variables or changing the structure of the dataset.

STATISTICAL ANALYSIS (CORRELATION, REGRESSION)

Correlation Analysis:

Measures the strength and direction of the relationship between two variables.

=*CORREL(array1, array2)* calculates correlation between two data sets.

Regression Analysis:

Studies the relationship between dependent and independent variables.

=*LINEST(known_y's, [known_x's], [const], [stats])* for linear regression.

DATA VISUALIZATION + Slicers

Data visualization is a powerful technique for presenting complex information in a clear and easily understandable way.

Data visualization is important for various reasons, it aids analysis, it brings clarity, and promotes communication.

Excel has different visuals/Charts, some of which are:

- **Column Charts:** Compare values across categories.
- **Line Charts:** Show trends over a period.
- **Pie Charts:** Display parts of a whole.
- **Bar Charts:** Similar to column charts but with horizontal bars.
- **Scatter Plots:** Visualize relationships between two variables.
- **Heat Maps:** Depict variations in data using colors.

To create a chart/visual in excel:

- Select Data: Highlight the data you want to visualize.
- Insert Chart: Choose a chart type from the "Insert" tab.
- Format: Customize colors, labels, and other chart elements.

Slicers:

Slicers are visual controls that allow users to filter data interactively. They are primarily associated with PivotTables and PivotCharts. Generally, they simplify data filtering and enhance dashboard interactivity.

ADVANCED EXCEL SKILLS

PIVOT TABLES

Pivot Tables aids in analyzing large data sets and in creating summaries. Pivoting helps us work with large sets of data in excel, unlike in python where large data sets have no impact, in excel, it can get daunting when working with data that has hundreds of thousands of rows or even millions of rows, Pivoting provides a means of escape. It helps quick and even deep analysis and visualization of data. You can Drag-and-drop fields to rows/columns/values to create summaries and comparisons.

DASHBOARD ARCHITECTURE

A dashboard in Excel is a visual representation of data that provides a comprehensive view of key performance indicators (KPIs) and metrics. Designing an effective dashboard requires careful consideration of architecture to ensure clarity, accessibility, and relevance of information. Some of the steps to take in achieving success in your dashboard architecture includes:

Define Objectives:

- Clearly define the objectives of the dashboard. Understand the audience and the specific insights they need. This helps in determining the types of visualizations and data to include.

Data Source and Integration:

- Identify the data sources for your dashboard. Excel allows for direct connections or importing data from various sources. Ensure data integrity and consistency for accurate analysis.

Data Cleaning and Transformation:

- Perform necessary data cleaning and transformation steps within Excel to ensure the data is ready for analysis. This may include handling missing values, filtering, and structuring data appropriately.

Choose Appropriate Visualizations:

- Select visualizations that effectively communicate the insights. Excel offers a variety of charts and graphs. Consider bar charts, line graphs, pie charts, and gauges based on the nature of your data.

Layout and Organization:

- Organize your dashboard layout logically. Group related information together and maintain a consistent flow. Use color coding and formatting to guide the user's attention and enhance readability.

Interactive Elements:

- Implement interactive elements like drop-down lists, slicers, and buttons. Excel allows you to create dynamic dashboards where users can customize views and explore data by interacting with the dashboard.

Charts and Graphs Formatting:

- Customize the appearance of charts and graphs to align with the dashboard's theme and enhance visual appeal. Adjust colors, fonts, and styles for a cohesive and professional look.

Dashboard Navigation:

- If your dashboard contains multiple sheets, consider creating a navigation system. Use hyperlinks or index pages to help users move between different sections of the dashboard seamlessly.

SQL FOR DATA ANALYTICS

INTRODUCTION TO SQL, DATABASES, AND DBMS

UNDERSTANDING DATABASES AND DBMS

Databases:

Structured collections of data organized for efficient retrieval.

Store information that can be managed, accessed, and updated.

DBMS (Database Management System):

Software managing databases, allowing users to interact with data.

Examples: mysql, Oracle, SQL Server,sqlite etc.

UNDERSTANDING SQL SYNTAX AND DATABASE CONCEPTS

SQL (Structured Query Language):

Language used to interact with databases.

Allows users to perform operations like querying, updating, and managing databases.

Database Concepts:

Tables: Organized data in rows and columns.

Relationships: Connections between tables via keys (primary, foreign).

Indexes: Enhance database performance by speeding up data retrieval.

CREATING A DATABASE

Database Creation:

Using DBMS software (e.g., mysql Workbench, pgadmin) to create a new database.

Syntax: *CREATE DATABASE database_name;.*

Table Creation:

Creating tables within a database to organize data.

Syntax: *CREATE TABLE table_name (column1 datatype, column2 datatype, ...).*

DESIGNING A DATA MODEL

Data Modeling:

Logical representation of data and its relationships within a database.

Includes entity-relationship diagrams (ERD) to visualize tables and their connections.

Normalization:

Process of organizing data to minimize redundancy and dependency.

Dividing large tables into smaller, related tables to avoid data duplication.

BASIC SQL QUERY AND COMMANDS

Basic SQL Query:

SELECT: Retrieves data from a database.

Syntax: *SELECT column1, column2 FROM table_name;.*

SQL Commands:

Data Definition Language (DDL): Commands for defining database schema and structure.

Data Manipulation Language (DML): Commands for managing data within database objects.

Data Control Language (DCL): Commands for managing access rights and permissions.

Transaction Control Language (TCL): Commands for managing transactions in the database.

Data Definition Language (DDL) Commands:

- **CREATE:**

CREATE DATABASE db_name;

CREATE TABLE table_name (column1 datatype, column2 datatype, ...);

- **ALTER:**

  ALTER TABLE table_name ADD column_name datatype;

  ALTER TABLE table_name MODIFY column_name datatype;

- **DROP:**

  DROP DATABASE db_name;

  DROP TABLE table_name;

- **TRUNCATE:**

  TRUNCATE TABLE table_name;

Data Manipulation Language (DML) Commands:

- **INSERT:**

  INSERT INTO table_name (column1, column2, ...) VALUES (value1, value2, ...);

- **SELECT:**

  SELECT column1, column2 FROM table_name WHERE condition;

- **UPDATE:**

  UPDATE table_name SET column_name = value WHERE condition;

- **DELETE:**

  DELETE FROM table_name WHERE condition;

Data Control Language (DCL) Commands:

- **GRANT:**

  GRANT permission ON object TO user;

- **REVOKE:**

  REVOKE permission ON object FROM user;

Transaction Control Language (TCL) Commands:

- **COMMIT:**

  COMMIT;
- **ROLLBACK:**

  ROLLBACK;
- **SAVEPOINT:**

  SAVEPOINT savepoint_name;
- **SET TRANSACTION:**

  SET TRANSACTION;

FILTERING DATA WITH WHERE CLAUSE

WHERE Clause:

Filters records based on specified conditions.

Syntax: *SELECT * FROM table_name WHERE condition;.*

Filtering Data:

Retrieves specific data meeting particular criteria.

Conditions include comparisons, logical operators (AND, OR), etc.

SORTING AND GROUPING DATA USING ORDER BY AND GROUP BY

ORDER BY Clause:

Sorts retrieved data in ascending or descending order.

Syntax: *SELECT * FROM table_name ORDER BY column_name ASC/DESC;.*

GROUP BY Clause:

Groups rows that have the same values into summary rows.

Syntax: *SELECT column_name, COUNT(*), SUM(column_name) FROM table_name GROUP BY column_name;.*

INTERMEDIATE SQL

AGGREGATION FUNCTIONS

COUNT: Counts the number of rows in a specified column.

Syntax: *SELECT COUNT(column_name) FROM table_name;*

SUM: Calculates the sum of values in a specified column.

Syntax: *SELECT SUM(column_name) FROM table_name;*

AVG: Computes the average of values in a specified column.

Syntax: *SELECT AVG(column_name) FROM table_name;*

MAX: Retrieves the maximum value in a specified column.

Syntax: *SELECT MAX(column_name) FROM table_name;*

MIN: Fetches the minimum value in a specified column.

Syntax: *SELECT MIN(column_name) FROM table_name;*

JOINS AND SUBQUERIES FOR COMPLEX DATA RETRIEVAL

Joins: Combines rows from two or more tables based on a related column between them.

Types: INNER JOIN, LEFT JOIN, RIGHT JOIN, FULL OUTER JOIN.

Syntax*: SELECT * FROM table1 INNER JOIN table2 ON table1.column = table2.column;*

Subqueries: Nesting queries within other queries for complex data retrieval.

Example: *SELECT column_name FROM table_name WHERE column_name IN (SELECT column_name FROM another_table);*

POWER BI FOR DATA ANALYTICS

POWER BI FUNDAMENTALS

Introduction to Power BI and its Interface

Power BI Overview:

- Business intelligence tool by Microsoft for data visualization and analysis.
- Empowers users to connect, transform, and visualize data.

Power BI Components:

- Power BI Desktop: Application for creating reports and dashboards.
- Power BI Service: Cloud-based service for sharing and collaborating on reports.
- Power BI Mobile: Mobile app for accessing reports and dashboards on the go.

Power BI Interface:

- Home Tab**:** Landing page with options to open recent files, access online services, and get help.
- Ribbon: Toolbar containing various commands for data modeling, visualization, and formatting.
- Data View & Report View: Sections to manage data and create visualizations.
- Visualization Pane: Area to drag and drop fields for creating visualizations.

DATA IMPORTING, TRANSFORMATION, AND MODELING

Data Importing:

- Get Data: Import data from various sources like Excel, SQL Server, CSV, Web, etc.
- Query Editor**:** Tool for data transformation and cleaning before loading into the data model.

Data Transformation:

- Data Cleaning: Removing duplicates, handling null values, renaming columns, etc.
- Data Shaping**:** Pivot, unpivot, merge, and transform data using the Query Editor.

Data Modeling:

- Data Model**:** Relationships between tables created in Power BI's data view.
- Manage Relationships: Define connections between tables based on related columns.

CREATING RELATIONSHIPS BETWEEN DATA TABLES

Defining Relationships:

- Primary & Foreign Keys**:** Identify common columns to establish relationships.
- Diagram View**:** Visual representation of tables and their relationships.

Types of Relationships:

- One-to-One**:** Each record in one table is related to only one record in another table.
- One-to-Many**:** Each record in one table can be related to multiple records in another table.

Enforcing Relationships:

- Ensure proper relationship cardinality (1:1, 1:M) and referential integrity for accurate data analysis.

DATA ANALYSIS AND VISUALIZATION

Building Interactive Reports and Dashboards

Interactive Reports:

- Reports designed for user interaction and exploration of data.
- Incorporate dynamic elements for users to filter, sort, and manipulate data.

Dashboards:

- Summarize information visually through multiple interactive elements like charts, graphs, and tables.
- Provide an overview of key metrics and allow drill-down for detailed insights.

DAX (DATA ANALYSIS EXPRESSIONS) FOR CALCULATED COLUMNS AND MEASURES

Calculated Columns:

- Use DAX to create new columns based on calculations using existing data.
- Syntax: ***newcolumn = <expression>***, where expression defines the calculation.

Measures:

- DAX formulas to compute aggregations or perform calculations dynamically.
- Created in Power BI for aggregations like SUM, AVERAGE, or custom calculations.

UTILIZING SLICERS, FILTERS, AND DRILL-DOWNS FOR DYNAMIC VISUALS

Slicers:

- Allow users to filter data in visualizations interactively.
- Select specific categories or values to view data subsets.

Filters:

- Apply conditions to visuals or datasets to display only desired data.
- Dynamic filtering based on user-defined criteria.

Drill-Downs:

- Capability to navigate from a summarized view to detailed information.
- Users can drill into specific sections to explore granular data.
-

BEST PRACTICES FOR INTERACTIVE REPORTING AND VISUALIZATION:

Understand User Needs:

- Design reports and dashboards based on user requirements and intended audience.

Use Consistent Design:

- Maintain a consistent layout, color scheme, and formatting across visuals for easy comprehension.

Optimize Performance:

- Limit data volume displayed initially for faster load times; use incremental loading if applicable.

Test and Iterate:

- Test reports with end-users, gather feedback, and make iterative improvements.

PYTHON FOR DATA ANALYTICS

PYTHON BASICS

INTRODUCTION TO PYTHON PROGRAMMING

Python Basics:

- Python is a high-level, interpreted programming language known for its simplicity and readability.
- It supports multiple paradigms: procedural, object-oriented, and functional programming.

Installation:

- Download and install Python from the official website (python.org).
- Choose the appropriate version (Python 3.x) for your operating system.

Integrated Development Environment (IDE):

- Select an IDE (Integrated Development Environment) like pycharm, Jupyter Notebook, or Visual Studio Code to write and execute Python code.

DATA TYPES, STRUCTURES, LOOPS, CONDITIONAL STATEMENTS

Data Types:

- Numeric (int, float), String, Boolean.
- Understand the characteristics and usage of each data type.

Data Structures:

- Lists (*list*), Tuples (*tuple*), Dictionaries (*dict*), Sets (*set*), etc.
- Used to store and organize data in different ways.

Loops (for and while):

- **For** loop: Iterate over a sequence (list, tuple, string, etc.).
- **While** loop: Execute code as long as a condition is true.

Conditional Statements:

- **If**, **elif**, **else**: Control the flow of the program based on certain conditions.

Functions, Modules, and Libraries

Functions:

- Reusable blocks of code; they take inputs, perform operations, and return outputs.
- Defined using **def** keyword: ***def function_name(parameters)*****:**.

Modules:

- Python files containing functions, classes, and variables.
- Import modules using ***import module_name***.

Libraries:

- Collections of modules that extend Python's capabilities.
- Examples: numpy (for numerical computations), Pandas (for data manipulation), Matplotlib (for data visualization).

DATA MANIPULATION AND ANALYSIS

DATA STRUCTURES: LISTS, DICTIONARIES, TUPLES, SETS

Lists:

- Ordered collection of elements, mutable (modifiable).

- Syntax: ***my_list = [1, 'apple', 3.14, True]****.*

Dictionaries:

- Unordered collection of key-value pairs, mutable.
- Syntax: ***my_dict = {'key1': 'value1', 'key2': 'value2'}****.*

Tuples:

- Ordered, immutable collection of elements.
- Syntax: ***my_tuple = (1, 'apple', 3.14, True)****.*

Sets:

- Unordered collection of unique elements.
- Syntax: ***my_set = {1, 2, 3, 4}****.*

Pandas Library for Data Manipulation

Python library for data manipulation and analysis.

Offers data structures (Series, dataframe) for handling structured data.

Dataframes:

- 2-dimensional labeled data structure with columns of potentially different data types.
- Import Pandas: ***import pandas as pd****.*
- Create dataframe: ***df = pd.dataframe(data)****.*

Series:

- 1-dimensional labeled array capable of holding data of any type.
- Extract Columns*: **df['column_name']**.*

Data Manipulation:

- Handling missing data (fillna, dropna).

- Data merging, joining, and concatenation.
- Grouping data (groupby) for aggregations.
- Filtering Data: ***df[df['column'] > value]***.
- Grouping Data: ***df.groupby('column').mean()***.

Data Import/Export:

- Read Data: ***pd.read_csv('file.csv')***.
- Write Data: ***df.to_csv('new_file.csv')***.

DATA VISUALIZATION USING MATPLOTLIB AND SEABORN

Matplotlib:

- Comprehensive library for creating static, interactive, and animated visualizations.
- Create Line Chart: ***plt.plot(x, y)***.
- Create Bar Chart: ***plt.bar(x, height)***.
- Save Plot: ***plt.savefig('plot.png')***.

Seaborn:

- Based on Matplotlib, provides a high-level interface for drawing attractive and informative statistical graphics.
- Create Scatter Plot: ***sns.scatterplot(x, y, data=df)***.
- Create Histogram: ***sns.histplot(data=df, x='column_name')***.

Styling Plots:

- Adding Titles and Labels: ***plt.title('Title')***, ***plt.xlabel('X-axis label')***.
- Adjusting Figure Size: ***plt.figure(figsize=(width, height))***.

- Customizing Colors and Styles.

Best Practices for Python Programming:

Code Readability:

- Follow PEP 8 style guidelines for clean, readable code.
- Use meaningful variable and function names.

Documentation:

- Add comments and docstrings to explain code functionality.
- Document functions and modules for better understanding.

Testing and Debugging:

- Write test cases for functions and debug code regularly.
- Use debugging tools like print statements or debuggers.

Continuous Learning:

- Explore Python's vast ecosystem of libraries and keep learning new features and updates.

AUTHORED BY

ULOH C. ARMSTRONG *from BeBetter Analytics*

Uloh C. Armstrong is a statistician, data analyst and data analytics instructor.

He has authored multiple books on data analysis and data analysis tools in his mind, and hopes to move them from being abstract to being real/tangible, just like this particular note.

BeBetter Analytics is an Analytics firm that offers data solutions to individuals and corporate organizations alike. They answer questions on your business problems and prescribe solutions based on your data, they analyze data on research works both for undergraduate and post-graduate students, and finally, they train individuals and organizations on data analytics and use of data analysis tools.

NEXT STEPS:

Enroll for free:

Jovian data analysis with python boot camp here.

FreeCodeCamp data analysis with python here.