Google Data Analytics

1. Foundations: Data, Data, Everywhere

WEEK 1

- **Data** is basically a collection of facts or information, and through analysis we can use the data to draw conclusions, and make predictions, and decisions.

- **Data analytics** in the simplest terms is the science of data. It's a very broad concept that encompasses everything from the job of managing and using data to the tools and methods that data workers use each and every day.

- **Data analysis** is the collection, transformation, and organization of data in order to draw conclusions, make predictions, and drive informed decision-making. Data evolves over time which means this analysis or analytics, as we call it, can give us new information throughout data's entire life cycle. In simple terms, turning data into insights.

- **Data analyst** is someone who collects, transforms, and organizes data in order to help make informed decisions.

- **Cycle of Data Analysis -**
  a) **Ask -** We define the problem to be solved and we make sure that we fully understand stakeholder expectations.
  b) **Prepare -** Building a timeline and collecting data.
  c) **Process -** cleaning data to make sure it was complete, correct, relevant, and free of errors and outliers.
  d) **Analyze -** Analyzing the data you've collected involves using tools to transform and organize that information so that you can draw useful conclusions, make predictions, and drive informed decision-making.
  e) **Share -** shared findings and recommendations with team leaders
  f) **Act -** leadership acted on the results and focused on improving key areas

- **Data Ecosystem** is a group of elements that interact with one another in order to produce, manage, store, organize, analyze, and share data.

- **Cloud** is a place to keep data online, rather than on a computer hard drive. It is a term we use to describe the virtual location.

- **Data science** is defined as creating new ways of modeling and understanding the unknown by using raw data.

- **Data-driven decision-making** is defined as using facts to guide business strategy like - Recommendation Systems

WEEK 2

- **Analytical skills** are qualities and characteristics associated with solving problems using facts. There are a lot of aspects to analytical skills, but five are most essential. They are -
  a) **Curiosity -** Wanting to learn something.
  b) **Understanding context -** Try to understand the context and eliminate which is not.
  c) **Having a technical mindset -** Divide the whole problem into smaller ones.
  d) **Data design -** Arrangement of data
  e) **Data strategy -** Data strategy is the management of the people, processes, and tools used in data analysis.

- **Analytical thinking** involves identifying and defining a problem and then solving it by using data in an organized, step-by-step manner. The five key aspects to analytical thinking are -
  a) **Visualization -** Visualization is the graphical representation of information. Some examples include graphs, maps, or other design elements.
  b) **Strategy -** Strategizing helps data analysts see what they want to achieve with the data and how they can get there. Strategy also helps improve the quality and usefulness of the data we collect.
  c) **Problem-orientation -** Data analysts use a problem- oriented approach in order to identify, describe, and solve problems. It's all about keeping the problem top of mind throughout the entire project.
  d) **Correlation -** It is like a relationship between 2 or more aspects.
  e) **Big-picture and Detail-oriented thinking -** Able to see the big picture of any aspect as well as have clarity on details.

- **Root cause** is the reason why a problem occurs. If we can identify and get rid of a root cause, we can prevent that problem from happening again. **Ask 5 WHYs.**

- **Gap analysis** lets you examine and evaluate how a process works currently in order to get where you want to be in the future.

WEEK 3

- **Data Life Cycle -**
  a) **Plan -** During planning, a business decides what kind of data it needs, how it will be managed throughout its life cycle, who will be responsible for it, and the optimal outcomes.
  b) **Capture -** This is where data is collected from a variety of different sources and brought into the organization.
  c) **Manage -** Storing & arranging data somewhere like a database.
  d) **Analyze -** Data is used to solve problems, make great decisions, and support business goals.
  e) **Archive -** Archiving means storing data in a place where it's still available, but may not be used again.
  f) **Destroy -** Destroy data.

- **Tools for Data Analysis -**
  a) Spreadsheet
  b) Query Language
  c) Visualization Tools - **Tableau, R with RStudio**

- It's really important that our data should be **fair and unbiased. Sampling should be done properly** including equivalent numbers of variations.

2. Ask Questions to Make Data-Driven Decisions

WEEK 1

- **Types of Problems -**
  a) **Making predictions -** This problem type involves using data to make an informed decision about how things may be in the future.
  b) **Categorizing things -** This means assigning information to different groups or clusters based on common features.
  c) Spotting something unusual - Data analysts identify data that is different from the norm.
  d) Identifying themes - Identifying themes takes categorization as a step further by grouping information into broader concepts.
  e) Discovering connections - Enables data analysts to find similar challenges faced by different entities, and then combine data and insights to address them.
  f) Finding patterns - Data analysts use data to find patterns by using historical data to understand what happened in the past and is therefore likely to happen again.

- **SMART Question -**
  a) 'SMART' stands for Specific, Measurable, Attainable, Relevant, and Timely.
  b) Avoid closed-ended questions like Yes/No type questions
  c) Be specific
  d) Try to be measurable
  e) Action-oriented questions encourage change
  f) Relevant questions are important and have significance to the problem we're trying to solve.
  g) Try to limit within a time frame.
  h) Questions should not create biases like **“Is this good, ist it?”**.

- **Data-inspired decision-making** explores different data sources to find out what they have in common.

- **Algorithm** is a process or set of rules to be followed for a specific task.

WEEK 2

- **Two kinds of data -**
  a) **Quantitative data** is all about the specific and objective measures of numerical facts. This can often be the what, how many, and how often about a problem.
  b) **Qualitative data** describes subjective or explanatory measures of qualities and characteristics or things that can't be measured with numerical data, like your hair color.

- **Reports and dashboards** are both useful for data visualization. But there are pros and cons for each of them.
  a) A **report** is a static collection of data given to stakeholders periodically.
  b) A **dashboard** on the other hand, monitors live, incoming data.

- **Pivot table** is a data summarization tool that is used in data processing. Pivot tables are used to summarize, sort, re-organize, group, count, total, or average data stored in a database.
- **Metric** is a single, quantifiable type of data that can be used for measurement. It can also be combined into formulas that you can plug your numerical data into.

- **Metric Goal** is a measurable goal set by a company and evaluated using metrics.

- **Dashboard** is a tool that organizes information, typically from multiple data sets, into one central location for tracking, analysis, and simple visualization through charts, graphs, and maps. And just like filters and spreadsheets and queries.

- **Types of Dashboards -**
  a) **Strategic -** Focuses on long term goals and strategies at the highest level of metrics. Ex : Key Performance Indicators (KPIs) over a year.
  b) **Operational -** Short-term performance tracking and intermediate goals. Ex : Customer Service team dashboard
  c) **Analytical -** Consists of the datasets and the mathematics used in these sets. Ex : Financial Performance.

- **Cell Reference in Excel** is a single cell or range of cells in a worksheet that can be used in a formula. Ex - E2, H7

- **IFERROR function -** IFERROR(<formula>, <error message>)

- **Two kinds of data -**
  a) **Small Data -**
    1. Describes a data set made up of specific metrics over a short, well-defined time period.
    2. Usually organized and analyzed in spreadsheets.
    3. Likely to be used by small and midsize businesses.
    4. Simple to collect, store, manage, sort, and visually represent.
    5. Usually already a manageable size for analysis.

  b) **Big Data -**
    1. Describes large, less-specific data sets that cover a long time period.
    2. Usually kept in a database and queried.
    3. Likely to be used by large organizations.
    4. Takes a lot of effort to collect, store, manage, sort, and visually represent.
    5. Usually needs to be broken into smaller pieces in order to be organized and analyzed effectively for decision-making.

- **4 V's of Big Data**
  a) **Volume -** The amount of data.
  b) **Variety -** The different kinds of data.
  c) **Velocity -** How fast the data can be processed.
  d) **Veracity -** The quality and reliability of the data.

- **Types of Errors in spreadsheet -**
  a) **#DIV/0! -** When divided by 0.
  b) **#ERROR -** Tell us the formula can't be interpreted as it is input. This is also known as a parsing error. Ex - =COUNT(B1:D1 C1:C10) is invalid because the cell ranges aren't separated by a comma.
  c) **#NA -** Tells you that the data in your formula can't be found by the spreadsheet. Ex - The cell being referenced can't be found.
  d) **#NAME? -** The name of a formula or function used isn't recognized. Ex - The name of a function is misspelled.
  e) **#NUM -** Tells us that a formula's calculation can't be performed as specified by the data. Ex - =DATEDIF(A4, B4, "M") is unable to calculate the number of months between two dates because the date in cell A4 falls after the date in cell B4.
  f) **#VALUE -** can indicate a problem with a formula or referenced cells. It's often not clear right away what the problem is, so this error might take a little more effort to fix. Ex - There could be problems with spaces or text, or with referenced cells in a formula; you may have additional work to find the source of the problem.
  g) **#REF! -** A formula is referencing a cell that isn't valid. Ex - A cell used in a formula was in a column that was deleted.

- **Problem Domain** is the specific area of analysis that encompasses every activity affecting or affected by the problem. Before we can do anything else, we need to understand the problem domain and all of its parts and relationships so that we can discover the whole story.

- **Structured thinking** is the process of recognizing the current problem or situation, organizing available information, revealing gaps and opportunities, and identifying the options.

- **Types Of Data based on what insight we get?**

- **How do we do Structured thinking?**

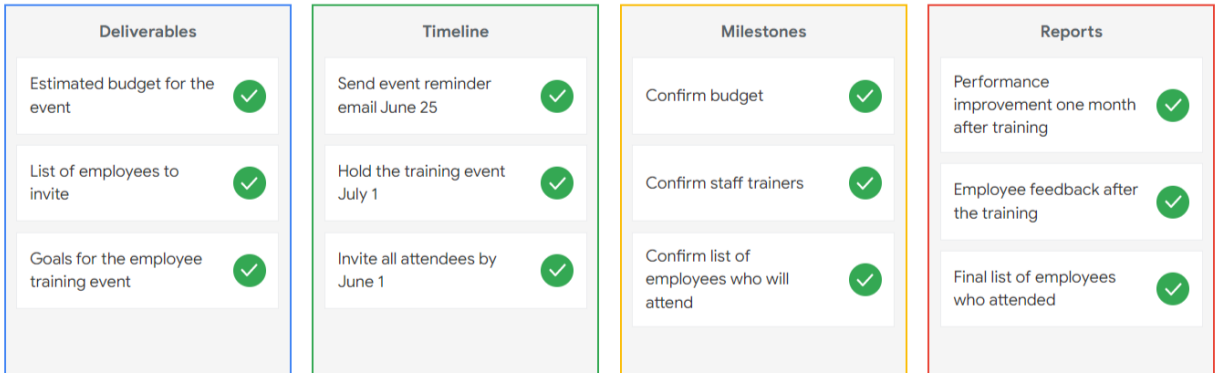
  Here's what might be in scope of work: **deliverables, timeline, milestones, and reports**.

  a) **Deliverables** - Preplan, example - the wedding planner and couple will need to decide on the invitation, make a list of people to invite, collect their addresses, print the invitations, address the envelopes, stamp them, and mail them out. Now let's check out the timelines.

  b) **Milestones** - What needs to be done which keep us on track.

c) **Timeline -** Upto when it should be done.

d) **Reports -** reports, which give our couple some peace of mind by telling them when each step is complete.



WEEK 4

- **Stakeholders** are people that have invested time, interest, and resources into the projects that you'll be working on as a data analyst.

- **Initial communication initiatives -**
  a) Who is the audience?
  b) What they already know.
  c) What do they need to know?
  d) How can you communicate efficiently?

- **Limitations of Data -**
  a) Can be incomplete
  b) Don’t miss misaligned data (understand the context)
  c) Deal with dirty data
  d) Tell a clear story
  e) Be the judge

- **Do(s) in Meetings -**
  a) Come Prepared
  b) Be on time
  c) Pay attention
  d) Ask Questions

- **Don’t(s) in Meetings -**
  a) Come Un-prepared
  b) Arrive Late
  c) Be distracted
  d) Dominate the conversation
  e) Talk over others

3. Foundations: Data, Data, Everywhere

WEEK 1

- **Common ways to collect data -**
  a) Interview
  b) Observations
  c) Forms
  d) Questionnaires
  e) Surveys
  f) Cookies - Most effective way to track people's online activities and interests.

- **Data collection consideration -**
  a) How data will be collected?
  b) Choose data sources.
     i) **First-party data -** This is data collected by an individual or group using their own resources.
     ii) **Second-party data -** Data collected by a group directly from its audience and then sold.
     iii) **Third-party data -** Data collected from outside sources who did not collect it directly. This data might have come from a number of different sources before you investigated it. It might not be as reliable, but that doesn't mean it can't be useful.
  c) Check data for accuracy, bias, and credibility.
  d) Decide which data to choose.
  e) Decide what time to choose.
  f) Analyze the size of the sample.
  g) Also store them correctly and in the right data type.

- **Population** refers to all possible data values in a certain data set. It's hard to collect data from the whole population.

- **Sample** is a part of a population that is representative of the population. If you figure out 3 attributes - Population size, Confidence level, Error of margin then you can calculate sample size online.

- **Types of data according to values -**
  a) **Discrete Value -** 1 to 10
  b) **Continuous Value -** temperature
  c) **Nominal data -** Qualitative data that's categorized without a set order.They could respond "Yes," "No," or "Not sure."
  d) **Ordinal data -** Type of qualitative data with a set order or scale. Example - Ratings.

- **Types of data according to arrangements -**
  a) **Structured Data -** Structured data is data that's organized in a certain format, such as rows and columns. Spreadsheets and relational databases are two examples of software that can store data in a structured way.
  b) **Unstructured Data -** This is data that is not organized in any easily identifiable manner. Audio and video files are examples of unstructured data because there's no clear way to identify or organize their content. Unstructured data might have internal structure, but the data doesn't fit neatly in rows and columns like structured data.

- **Data Model -** Structured data works nicely within a data model, which is a model that is used for organizing data elements and how they relate to one another. Data models help to keep data consistent and provide a map of how data is organized.This makes it easier for analysts and other stakeholders to make sense of their data and use it for business purposes.

- **Data modeling** is the process of creating diagrams that visually represent how data is organized and structured. These visual representations are called data models. You can think of data modeling as a blueprint of a house.

  There are a lot of approaches when it comes to developing data models, but two common methods are the Entity Relationship Diagram (ERD) and the Unified Modeling Language (UML) diagram.

  ERDs are a visual way to understand the relationship between entities in the data model.

  UML diagrams are very detailed diagrams that describe the structure of a system by showing the system's entities, attributes, operations, and their relationships.

- **Levels / Types of data modeling**
  a) **Conceptual data modeling** gives a high-level view of the data structure, such as how data interacts across an organization. For example, a conceptual data model may be used to define the business requirements for a new database. A conceptual data model doesn't contain technical details.
  b) **Logical data modeling** focuses on the technical details of a database such as relationships, attributes, and entities. For example, a logical data model defines how individual records are uniquely identified in a database. But it doesn't spell out actual names of database tables. That's the job of a physical data model.
  c) **Physical data modeling** depicts how a database operates. A physical data model defines all entities and attributes used; for example, it includes table names, column names, and data types for the database.

- **Wide data**, every data subject has a single row with multiple columns to hold the values of various attributes of the subject. **Long data** is data in which each row is one time point per subject, so each subject will have data in multiple rows.

- **Data transformation** is the process of changing the data's format, structure, or values. As a data analyst, there is a good chance you will need to transform data at some point to make it easier for you to analyze it.

  Data transformation usually involves:

  - Adding, copying, or replicating data
  - Deleting fields or records
  - Standardizing the names of variables
  - Renaming, moving, or combining columns in a database
  - Joining one set of data with another
  - Saving a file in a different format. For example, saving a spreadsheet as a comma separated values (CSV) file.

  Goals for data transformation might be:

  - Data organization: better organized data is easier to use
  - Data compatibility: different applications or systems can then use the same data
  - Data migration: data with matching formats can be moved from one system to another
  - Data merging: data with the same organization can be merged together
  - Data enhancement: data can be displayed with more detailed fields
  - Data comparison: apples-to-apples comparisons of the data can then be made

WEEK 2

- **Data bias** is a type of error that systematically skews results in a certain direction. Maybe the questions on a survey had a particular slant to influence answers, or maybe the sample group wasn't truly representative of the population being studied.

  Types of bias -

  a) **Sampling bias** is when a sample isn't representative of the population as a whole.
  b) **Observer bias/ Experimenter bias/ Research bias** - It's the tendency for different people to observe things differently. Example - Rounding of time or blood pressure.
  c) **Interpretation bias** - The tendency to always interpret ambiguous situations in a positive, or negative way.

- **Good Data Sources -** Reliable, Original, Comprehensive, Current, Cited

- **Data Ethics -** It is related to transparency and privacy are part of the process. Data ethics tries to get to the root of the accountability companies have in protecting and responsibly using the data they collect.

  There are lots of different aspects of data ethics but we'll cover six: ownership, transaction transparency, consent, currency, privacy, and openness.

- **Data privacy** means preserving a data subject's information and activity any time a data transaction occurs. This is sometimes called information privacy or data protection. It's all about access, use, and collection of data. It also covers a person's legal right to their data.

- **Data anonymization** - A data analyst removes personally identifying information from a dataset.

- **Data Interoperability** is the ability of data systems and services to openly connect and share data. For example, data interoperability is important for health care information systems where multiple organizations such as hospitals, clinics, pharmacies, and laboratories need to access and share data to ensure patients get the care that they need. This is why your doctor is able to send your prescription directly to your pharmacy to fill.

- **Sites and resources for open data**
  Luckily for data analysts, there are lots of trustworthy sites and resources available for open data. It is important to remember that even reputable data needs to be constantly evaluated, but these websites are a useful starting point:

  1. U.S. government data site: Data.gov is one of the most comprehensive data sources in the US. This resource gives users the data and tools that they need to do research, and even helps them develop web and mobile applications and design data visualizations.
  2. U.S. Census Bureau: This open data source offers demographic information from federal, state, and local governments, and commercial entities in the U.S. too.
  3. Open Data Network: This data source has a really powerful search engine and advanced filters. Here, you can find data on topics like finance, public safety, infrastructure, and housing and development.
  4. Google Cloud Public Datasets: There are a selection of public datasets available through the Google Cloud Public Dataset Program that you can find already loaded into BigQuery.
  5. Dataset Search: The Dataset Search is a search engine designed specifically for data sets; you can use this to search for specific data sets.

WEEK 3

- **Metadata -** Data about data. In data analytics, metadata helps data analysts interpret the contents of the data within a database. Metadata creates a single source of truth by keeping things consistent and uniform. We data analysts love consistency. Metadata also makes data more reliable by making sure it's accurate, precise, relevant, and timely.

  Types of Metadata -

  a) **Descriptive metadata** is metadata that describes a piece of data and can be used to identify it at a later point in time. For instance, the descriptive metadata of a book in a library would include the code you see on its spine, known as a unique International Standard Book Number, also called the ISBN.
  b) **Structural metadata**, which is metadata that indicates how a piece of data is organized and whether it's part of one or more than one data collection. Let's head back to the library. An example of structural data would be how the pages of a book are put together to create different chapters.
  c) **Administrative metadata** is metadata that indicates the technical source of a digital asset. When we looked at the metadata inside the photo, that was administrative metadata. It shows you the type of file it was, the date and time it was taken, and much more. Here's one final thought to help you understand.

- **Metadata repository** is a database specifically created to store metadata. Metadata repositories can be stored in a physical location, or they can be virtual, like data that exists in the cloud. These repositories describe where metadata came from, keep it in an accessible form so it can be used quickly and easily, and keep it in a common structure for everyone who may need to use it. Metadata repositories make it easier and faster to bring together multiple sources for data analysis. They do this by describing the state and location of the metadata, the structure of the tables inside, and how data flows through the repository. They even keep track of who accesses the metadata and when.

WEEK 4

- **Encryption** uses a unique algorithm to alter data and make it unusable by users and applications that don't know the algorithm. This algorithm is saved as a "key" which can be used to reverse the encryption; so if you have the key, you can still use the data in its original form.

- **Tokenization** replaces the data elements you want to protect with randomly generated data referred to as a "token." The original data is stored in a separate location and mapped to the tokens. To access the complete original data, the user or application needs to have permission to use the tokenized data and the token mapping. This means that even if the tokenized data is hacked, the original data is still safe and secure in a separate location.

4. Process Data from Dirty to Clean

WEEK 1

- **Data integrity** is the accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle. There's a chance data can be compromised every time it's **replicated, transferred, or manipulated** in any way. Data replication is the process of storing data in multiple locations. If you're replicating data at different times in different places, there's a chance your data will be out of sync.

- **Random sampling** is a way of selecting a sample from a population so that every possible type of the sample has an equal chance of being chosen.

- **Statistically significant** means the results of the test are real and not an error caused by random chance. Usually, you need a statistical power of at least 0.8 or 80% to consider your results statistically significant.

- **Confidence level** is the probability that your sample accurately reflects the greater population. But most industries hope for at least a 90 or 95 percent confidence level. Industries like pharmaceuticals usually want a confidence level that's as high as possible when they are using a sample size.

WEEK 2

- **Data validation** is a tool for checking the accuracy and quality of data before adding or importing it.

- **Data merging** is the process of combining two or more datasets into a single dataset.

- In data analytics, **compatibility** describes how well two or more datasets are able to work together.

- **Data mapping** is the process of matching fields from one database to another. This is very important to the success of data migration, data integration, and lots of other data management activities.

WEEK 3

SQL Concepts :)

WEEK 4

- **Documentation** which is the process of tracking changes, additions, deletions and errors involved in your data cleaning effort. It is important to maintain it.

- **Changelog** is a file containing a chronologically ordered list of modifications made to a project. You can use and view a changelog in spreadsheets and SQL to achieve similar results.

WEEK 5 & WEEK 6

- **Things to involve in data analytics resume -**
    a) Effective Communication
    b) **PAR** Pattern to follow - Problem, Action, Result
    c) Spreadsheet
    d) SQL
    e) R
    f) Tableau
    g) **Soft skills** like - Presentation skills, Collaboration, Communication, Research, Problem-solving skills, Adaptability, Attention to detail.

5. Analyze Data to Answer Questions

WEEK 1

- **Analysis** is the process used to make sense of the data collected. It means taking the right steps to proceed and think about your data in different ways. The goal of analysis is to identify trends and relationships within the data so that you can accurately answer the question you're asking.

- **Phases of analysis -**
    a) **Organize data -** Make it available in sheet or database
    b) **Format and adjust data -** Sort or filter data
    c) **Get input from others -** Ask for opinions
    d) **Transform data -** By observing relationships between data points and making calculations.

WEEK 2

- **Convert Celcius to farenheit (Spreadsheet) -** `=CONVERT(A1,"C","F")`

- **Data validation** - It allows you to control what can and can't be entered in your worksheet.

- **CONCATENATE** is a function that joins together two or more text strings in a spreadsheet.

- **Common functions for string** - LEN, LEFT, RIGHT and FIND in a spreadsheet.

WEEK 3

- **Data aggregation** is the process of gathering data from multiple sources in order to combine it into a single summarized collection. In data analytics, a summarized collection, or summary, describes identifying the data you need and gathering it all together in one place.

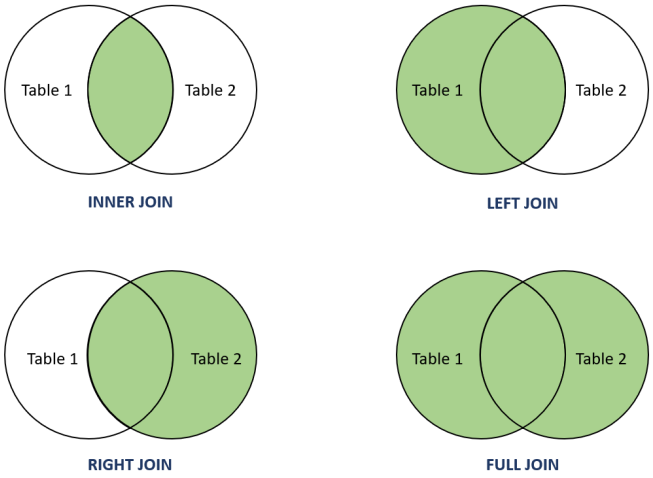
- **VLOOKUP -** if you enter FALSE as the last input parameter in a VLOOKUP function, VLOOKUP will search for exact match, else similar match for TRUE. When using VLOOKUP, there are some common limitations that data analysts should be aware of. One of these limitations is that VLOOKUP only returns the first match it finds, even if there are many possible matches within the column.

- A data analyst creates an absolute reference around a function array. The purpose of the absolute reference is to lock the function array so rows and columns don't change if the function is copied.

- **More SQL things -**
  a) SQL HAVING: This is an overview of the HAVING clause, including what it is and a tutorial on how and when it works.
  b) SQL CASE: Explore the usage of the CASE statement and examples of how it works.
  c) SQL IF: This is a tutorial of the IF function and offers examples that you can practice with.

- **Subquery** is a SQL query that is nested inside of a larger query.

- **JOIN** is a SQL clause that's used to combine rows from two or more tables based on a related column. Basically, you can think of a JOIN as a SQL version of VLOOKUP which we just covered. There are four common JOINs data analysts use, **inner, left, right, and outer (or full)**.



WEEK 4

- **SUMPRODUCT -** Multiplies the components of a given array and then returns the sum of the products in excel

- **GROUP BY** in DB is a command that groups rows that have the same values from a table into summary rows. The GROUP BY command is used with SELECT statements. In a basic SELECT FROM or SELECT-FROM-WHERE query, GROUP BY comes at the end of the query.

```
SELECT column_name FROM table_name WHERE condition GROUP BY
      column_name(s) ORDER BY column_name(s);
```

- The purpose of the **EXTRACT** command in a query is to extract a part from a given date. The EXTRACT command can extract any part from a date/time value.

```
SELECT EXTRACT(<part like year, month> FROM "2017-06-15");
```

- The data validation process is a form of data cleaning.

- **Types of data validation -**
  a) Data type check
  b) Data Range
  c) Data Constraints
  d) Data Consistency
  e) Data Structure
  f) Code Validation

- **Temporary table** is a database table that is created and exists temporarily on a database server. Temp tables as we call them store subsets of data from standard data tables for a certain period of time. Then they're automatically deleted when you end your SQL database session.

Syntax 1 :
```
With <Temp_table_name> AS ( SELECT * FROM student where
section = 'A');
```

Syntax 2 :
```
SELECT * INTO <Temp_table_name> FROM student where
section = 'A';
```

Syntax 3 :
```
Create Table <Temp_table_name> AS ( SELECT * FROM
student where section = 'A'); -----> Others can also excess.
```

6. Share Data Through the Art of Visualization

WEEK 1

- **Data visualization** is the graphic representation and presentation of data. In reality, it's just putting information into an image to make it easier for other people to understand. The graphic should include four key elements: t**he information or data, the story, the goal and the visual form.**

- **Tableau** is a business intelligence and analytics platform that helps people see, understand, and make decisions with data. Visualizations in Tableau are automatically interactive.

- **Types of graphs / charts -**
    a) Bar graphs
    b) Line Graphs
    c) Pie Charts
    d) Maps
    e) Histogram
    f) Correlation charts
    g) Scatter plots
    h) Distribution graph

- **Correlation** in statistics is the measure of the degree to which two variables move in relationship to each other. An example of correlation is the idea that “As the temperature goes up, ice cream sales also go up.” It is important to remember that correlation doesn’t mean that one event causes another. But, it does indicate that they have a pattern with or a relationship to each other. If one variable goes up and the other variable also goes up, it is a positive correlation. If one variable goes up and the other variable goes down, it is a negative or inverse correlation. If one variable goes up and the other variable stays about the same, there is no correlation.

- **Causation** refers to the idea that an event leads to a specific outcome. For example, when lightning strikes, we hear the thunder (sound wave) caused by the air heating and cooling from the lightning strike. Lightning causes thunder.

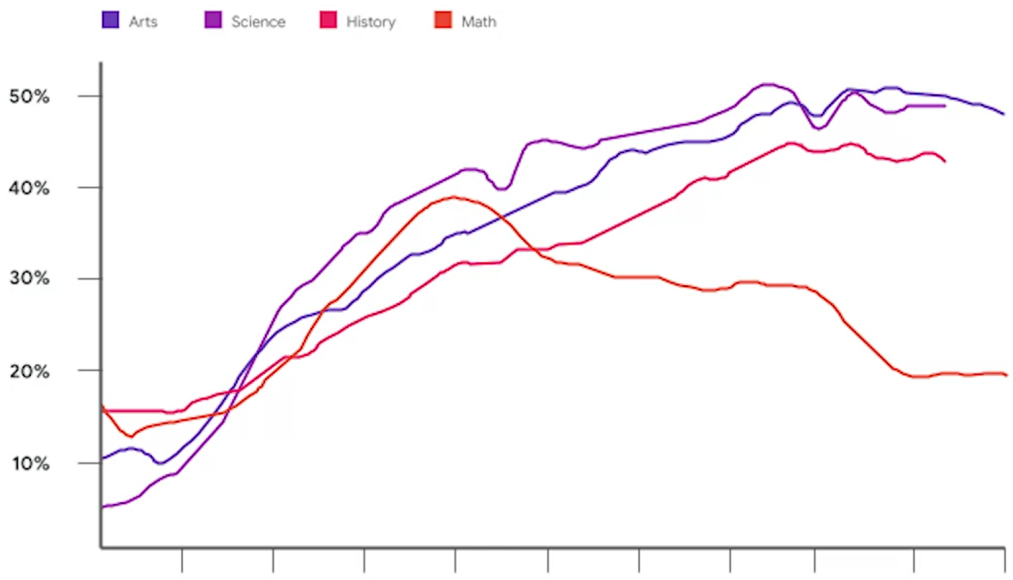
- **Decision tree** is a decision-making tool that allows you, the data analyst, to make decisions based on key questions that you can ask yourself. Each question in the visualization decision tree will help you make a decision about critical features for your visualization.

- **Elements of Charts -**
  a) **Line -** Line helps to show trends in data.
  b) **Shape -** Shapes with symmetry are usually more familiar to people, so there's less work for the audience to do when viewing symmetrical data viz. But the asymmetrical shapes in this map are still instantly recognizable as countries.
  c) **Color -** The value is how light or dark the colors are in a visualization. In more scientific terms value indicates how much light is being reflected. Dark values with some black added are called shades of color, like these shades of green. Light values with white added are called tints, like these tints of blue.
  d) **Space -** Space is the area between, around and in the objects. There should always be space in data visualizations, just not too much or too little.
  e) **Movement -** Movement is used to create a sense of flow or action in a visualization.

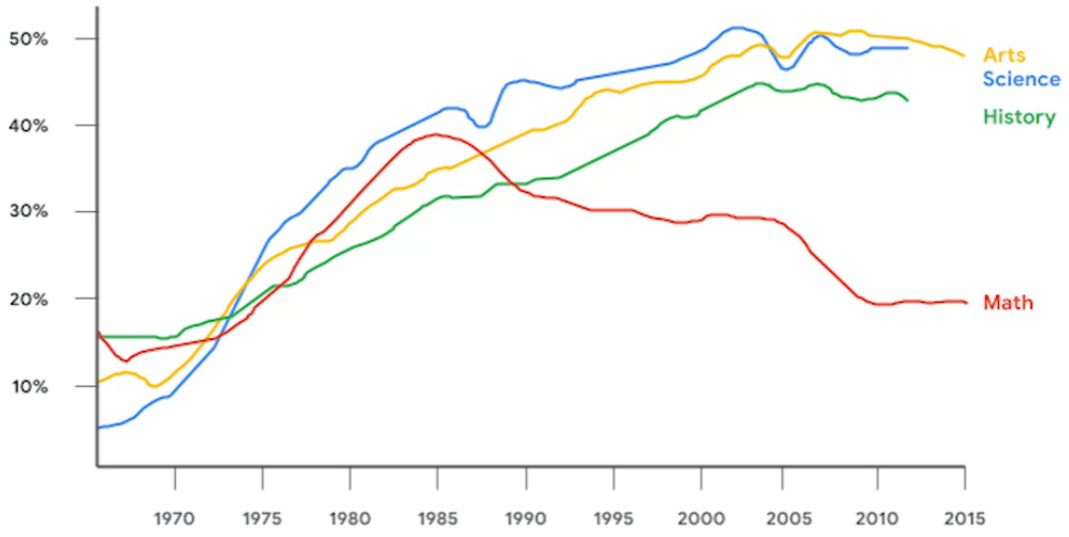
- **Five phases that you can use when creating data visualizations -**
  a) **Empathize -** In the empathize phase you think about the emotions and needs of the target audience of your data viz, whether it's stakeholders, team members or the general public. Here you should avoid areas where people might face obstacles interacting with your visualizations. For example, Maybe you're thinking of using a color scheme that you like, but you realize that these colors might be a challenge to some people. Or the colors might not have enough contrast for people who have color vision deficiencies.
  b) **Define -** The define phase helps you to find your audiences needs, their problems, and your insights.
  c) **Ideate -** This might involve creating drafts of your visualization with different color combinations or maybe experimenting with different shapes. Creating as many examples as possible will help you refine your ideas. The key here is to always remember your audience when coming up with ideas and strategies. You want to think about how you can position your visualizations to meet the needs and expectations of your audience.
  d) **Prototype -** Here you'll start putting your charts, dashboards or other visualizations together.
  e) **Test -** Test different options and find the best one.

- **Accessible visualizations -**

a) **Labeling is important**



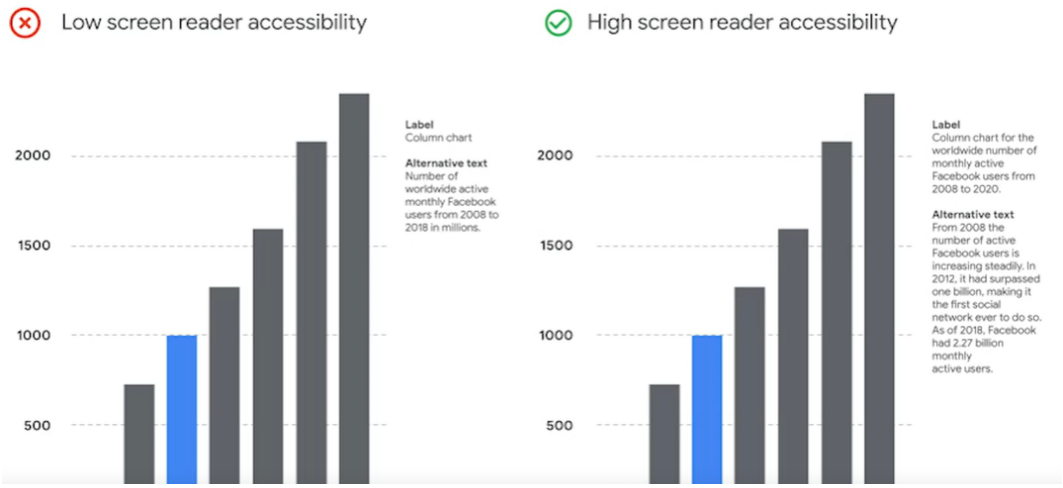
VS



b) **Use contrast background with charts, so that the audience can see a clear picture.**

c) **Use similar patterns so that the audience can distinguish them easily.**

d) **Text alternatives**



- **Design thinking** is a process used to solve complex problems in a user-centric way.

WEEK 2

Understanding Tableau, will understand later in more depth :)

WEEK 3

- **Static data** involves providing screenshots or snapshots in presentations or building dashboards using snapshots of data. There are pros and cons to static data.

- **Live data** means that you can build dashboards, reports, and views connected to automatically updated data.

- **Things need to take care while presenting the analysis -**
  a) **Characters -** Analyze audiences and their interests.
  b) **Settings -** Which describes what's going on, how often it's happening, what tasks are involved, and other background information about the data project that describes the current situation.
  c) **Plan -** This could be a challenge from a competitor, an inefficient process that needs to be fixed, or a new opportunity that the company just can't pass up. This complication of the current situation should reveal the problem your analysis is solving and compel the characters to act.

d) **Big Reveal -** how the data has shown that you can solve the problem the characters are facing by becoming more competitive, improving a process, inventing a new system, or whatever the ultimate goal of your data project may be.
e) **Aha moment -** When you share your recommendations and explain why you think they'll help your company be successful.

WEEK 4

- It’s better to share objectives and ppt, post ppt presentations. So that audience has time to go through it.
- During Q&A - Try to listen to the whole question first. Then go into details, use the Appendix slide to include information that is not very important.

7. Data Analysis with R Programming

- **R** is a programming language frequently used for statistical analysis, visualization and other data analysis. R is accessible, data-centric, open-source and has an active community of users. Here's three scenarios: reproducing your analysis, processing lots of data, and creating data visualizations where we can use R. R can handle large amounts of data much more quickly and efficiently. Finally R can create powerful visuals and has state-of-the-art graphic capabilities.

- **RStudio's** an IDE or integrated development environment. This means that RStudio brings together all the tools you might want to use in a single place. It also includes an editor for writing code, and tools for managing your data and creating visuals. In other words, it takes the output of one statement and makes it the input of the next statement.

- **Pipe**, which you'll use to make a sequence of code easier to work with and read. It's a tool in R for expressing a sequence of multiple operations.

- **Tidyverse -** Collection of packages containing reusable functions and more.

- **facet_wrap()** - Use to create a different plot for each type of cut of diamond.

- **CRAN** is a commonly used online archive with R packages and other R resources. CRAN makes sure that the R resources it shares follow the required quality standards and are authentic and valid. Packages installed in RStudio are called from CRAN. CRAN is an online archive with R packages and other R-related resources.

- **Dataframe -** A data frame is a collection of columns.

- Basically the **bias function** finds the average amount that the actual outcome is greater than the predicted outcome. It's included in the sim design package. If the model is unbiased, the outcome should be pretty close to zero.

- **Core packages of tidyverse :**
  a) **ggplot2 -** ggplot2 is used for data visualization, specifically plots. With ggplot2, you can create a variety of data viz by applying different visual properties to the data variables.
  b) **tidyr -** Tidyr is a package used for data cleaning to make tidy data.
  c) **readr -** Used for importing data.
  d) **dplyr -** dplyr offers a consistent set of functions that help you complete some common data manipulation tasks.
  e) **tibble -** tibble works with data frames.
  f) **purrr -** Works with functions and vectors helping make your code easier to write and more expressive.
  g) **stringr -** Includes functions that make it easier to work with strings.
  h) **forcats -** Provides tools that solve common problems with factors.

- In ggplot2, an **aesthetic** is a visual property of an object in your plot. In R, an argument is the information needed by a function. For example, in a scatter plot, aesthetics include the size, shape or color of your data points.

```
ggplot(data, aes(x=distance, y= dep_delay, color= carrier, size=
air_time, shape = carrier)) + geom_point()
```

- In ggplot2, a **geom** is the geometrical object used to represent your data. Geoms include points, bars, lines, and more.

- Sometimes it can be hard to understand trends in your data from scatter plots alone. **Smoothing** enables the detection of a data trend even when you can't easily notice a trend from the plotted data points. Ggplot2's smoothing functionality is helpful because it adds a smoothing line as another layer to a plot.

WEEK 4

- **Annotations** are a useful way to add notes to your plot. They help you explain the plot's purpose, highlight important data points, or comment on any data trends or findings the plot illustrates.

- ggsave() used to save plots in R.

WEEK 5

- **R Markdown** is a file format for making dynamic documents with R. You can use an R Markdown file as a code notebook to save, organize, and document your analysis using code chunks, comments, and other features. When you finish your data cleaning and exploration, you can create a report in R Markdown to summarize your findings for stakeholders. R Markdown documents are **written in Markdown**. Markdown is a syntax for formatting plain text files.