Data+ Exam Foundations

- **Data+**
    - *CompTIA Data+ is an early-career data analytics certification for professionals tasked with developing and promoting data-driven business decision-making. (CompTIA.org)*

- **Exam Description**
    - The CompTIA Data+ exam will certify the successful candidate has the knowledge and skills required to transform business requirements in support of data-driven decisions through mining and manipulating data, applying basic statistical methods, and analyzing complex datasets while adhering to governance and quality standards throughout the entire data life cycle.

- **Five Domains**
    - 15% Data Concepts and Environments
    - 25% Data Mining
    - 23% Data Analysis
    - 23% Visualization
    - 14% Data Governance, Quality, and Controls

- **Exam Details**
    - Up to 90 questions in 90 minutes
        - Multiple-choice
        - Performance-based/Simulations
            - Requires a 675 out of 900
            - Recommended Experience:
            - 18–24 months of experience in a report/business analyst job role
            - Exposure to databases and analytical tools
            - Basic understanding of statistics
            - Data visualization experience
    - Released: February 28, 2022

- **Are You Ready?**
    - Take practice exams
    - Did you score at least 85% or higher?

- If you need more practice, take additional practice exams to hone your skills before attempting the exam

- **What kind of jobs can I get?**

Data Schemas

Objective 1.1

- **OBJ 1.1:** Identify basic concepts of data schemas and dimensions.

- **Data Schemas**
    - **Data Schema**
        - Used to describe both the organization of data and the relationships between tables in a given data
    - **Database Engineers**
        - Plans the database schema before they begin to create the system

- **Relational Databases**
    - **Relational Database**
        - Uses tables to store the data that's being captured
    - **Spreadsheet**
        - Has multiple tables linked together with different relationships
    - **Tabular Schema**
        - Use rows and columns in this table format to store all of their data
    - **RDBMS** (Relational Database Management System)
        - Used with lots of different database software, such as my SQL Maria DB, and even Amazon serverless database system, Aurora.
            - Store the information about my customers:
                - First name
                - Last name
                - Phone number
                - Customer ID
        - Designing your relational database can determine what fields are going to be in each table and how they're going to be link together

- **SQL**
  - A programming language for data, and it works across all relational databases

- **REMEMBER**
  - Relational databases are going to be built using tables and those tables are going to be linked using a common field

- **Non-relational Databases**

  - **Now non-relational databases**
    - Databases that are not based on relationships
    - They're not going to use SQL or the structured query language
    - Non-relational databases are going to be able to handle extremely large amounts of traffic and data

      - Four main types of non-relational databases
        - **Document oriented databases**
          - Stores the data inside of XML documents or using Jaison format, which is the JavaScript object notation
        - **Key value stores**
          - Stores each value within a key value
        - **Column oriented databases**
          - Store data in columns instead of rows like we do in a traditional relational database
        - **Graph stores**
          - Stores are used to store individual elements as nodes inside of this database

    - Its quicker to grab information in a non-relational database than using something like a relational database

  - **REMEMBER:**
    - SQL = Relational databases
    - No SQL or Graph QL = Non-relational databases

- **Comparing Database Types**
    - **Relational databases**
        - Tables are used to store fields inside of the different columns
        - Each row is going to have the record that holds the data for that relational database
        - Tool to optimally design the tables and be able to have the least amount of information possible in those tables
        - Uses SQL, which is the structured query language to write data into the database
        - The amount of data that can be held in each field and the type of data that can be held in each field is limited
        - Each particular field you're going to do in your database is going to have a specific purpose, a specific field type
    - **Non-relational databases**
        - An alternative to relational databases
        - Stores as much information as you want in a key value pair, as opposed to having to be limited to 255 characters.
        - Easier to scale and build out for web-based or cloud-based applications
        - Uses any programming language you want
        - Stores both structured and non-structured data within those databases, because you do have more flexibility than you do in a traditional relational database
- **Data Normalization**
    - **Data Normalization**
        - Optimizes the storage and use of data within a given database
    - **First normal form (1NF)**
        - Eliminates any redundant information in individual tables
    - **Second normal form (2NF)**
        - It has all the related information applicable to multiple tables using a foreign key

- **Third normal form (3NF)**
  - Eliminates fields that do not depend on a given key

- **Fourth normal form (4NF)**
  - The data has to have a relationship that is in BC, NF or Boyce Code normal form, and has no multi valued dependencies in them

  - Ensures there are no joint dependencies inside the database

- The goal of data normalization is to establish the relationships between the different data in forms to be able to have the data that we need when we run our reports

- **Database Relationships**

  - **Primary key**
    - A unique identifier for a record that cannot contain duplicates
    - Every table needs to have a primary key

  - **Foreign key**
    - A primary key that was referenced by a different table

  - **Database relationships**
    - **One-to-one relationship**
      - One record in the table will be associated with only one record in another table
    - **One to many relationship**
      - One record in the table with a primary key is associated with multiple records in other tables
    - **Many to many relationship**
      - Many different records are associated with many other records in different table

- **Referential Integrity**

  - **Referential Integrity**

- Used to establish and maintain that records are not orphaned by ensuring the proper table has the key field established inside of it
-
- Ensures and guarantee that the primary key is being used as a foreign key in a table, and making sure that it actually exists in the table before creating records in the second table
- Prevents the occurrence of bad or missing data in any of the tables

- Referential integrity come under attack is when people start modifying or deleting parts of the database

- Update or delete data in a cascade manner

    - **Cascade Delete/Update**
        - Make sure that no data is being left orphaned as you go and make updates or deletions to the primary key and the records they control

- **Remember**
    - Make sure that all of the changes we're doing to any part of our database are cascading throughout that database, either through an update or a deletion

- **Data Denormalization**

    - **Data Denormalization**
        - Occurs whenever you take data and you make it unstructured into tables using normalization
        - There will be a lot of redundancies and repetitive data

        - Go back through all the different tables, and pull the data back out to make it redundant again

        - When you start dealing with big data and you move into data, warehousing data, mining, data analysis, and data visualization

Data Systems

Objective 1.1

- **OBJ 1.1:** Identify basic concepts of data schemas and dimensions.

- **Data Systems**

    - **Data System**
        - Any information technology system that captures processes, stores, queries, or reports on the data contained within them

- **Data Processing Types**

    - **Database Transaction**
        - An insertion, deletion, or simply query of the database

            - **Online Transactional Processing (OLTP)**
                - Systems designed to handle very large scales of transactions
                - OLTP is built for large numbers of real-time database transactions
            - **Online Analytical Processing (OLAP)**
                - Systems designed to handle longer running queries for more complex data processing
                - OLAP is built for longer running and more complex database queries

    - It is important to determine whether you should use OLTP or OLAP to prevent performance issues

- **Data Warehouse**

    - **Source System**
        - The system of record for any particular kind of data
            - **Clickstream Data**
                - Data where individual users are clicking on different pages of the website

- **Purchase Data**
  - Involves people who handles credit card to buy something that we then had to fulfill and ship out to them.

- **Catalog Information**
  - Another source system that contains the categorization and the descriptions and everything else about individual items

- **Data Warehouses**
  - Ingests data from the source systems record and combine them together
  - Considered a single source of truth
  - It combines that data into one place where we can get at it in a very efficient manner
  - Data warehouse is an example of an OLAP system where it's built to handle really big, heavy queries

- **OLAP cubes**
  - A three-dimensional structure that provides data grouped together by different dimensions
  - Optimizes the expected queries from each data marts customer

- **Data Marts**
  - A subset of data that is designed for specific groups or departments without impacting the performance of other departments or groups

- **Data Warehouse Schemas**

  - **Fact Table**
    - Contains all of the main keys associated with the queries

  - **Dimension Table**
    - Information associated with that fact table that is tied together using the course ID

  - **Star Schema**

        - Individual dimension tables branching off from that fact table that looks like a star
        - One layer dimension associated with that fact table

    - **Snowflake Schema**
        - Multiple dimensions associated with each dimension branching out like a snowflake

- **Data Lakes**
    - **Data Lake**
        - Holds both structured and unstructured data
        - It dumps raw server logs that has no structure
        - Does not required to be structured

    - **Data lake house**
        - Queries data in place on the data lake
        - In order to query, there should be some structured data, some schema data off on the side that's been built up

        - you'll get flexibility and the cost-effectiveness of a data lake, but you can still conduct queries across the entire data set

        - Either a data lake and lake house or a true data warehouse can be used to analyze large amounts of data

- **Changing Dimensional Data**

    - **Dimension table**
        - Contains metadata about the stuff in your fact table

        - How do we manage changes to that dimensional data?
        - How do we retain a history of what it used to be?

    - **Type one slowly changing dimension**
        - Where the new information is simply overriding the old information

- You can no longer query for those previous names that existed in the past

- **Type two slowly changing dimension**
  - It has the complete history of the information
  - It is retaining a history of all the previous changes to the data

- **Three slowly changing dimensions**
  - This approach is to maintain the current and previous data

Data Types

Objectives 1.2 and 1.3

- **OBJ 1.2:** Compare and contrast different data types.
- **OBJ 1.3**: Compare and contrast common data structures and file formats.

- **Quantitative & Qualitative**
    - **Quantitative data**
        - Data that is defined through different numbers
            - **Discrete data**
                - Data that can be counted with a certain number of values
            - **Continuous data**
                - Data that can be counted but with changing values
    - **Qualitative data**
        - Data is arranged into groups or categories based on some kind of quality
            - **Nominal data**
                - Information that has no natural order
            - **Ordinal data**
                - Information that has natural order
- **Data Field Types**
    - Data fields
        - Contains the different pieces of information in different databases
        - Data types are the thing that's we can control
    - **Texts and alphanumeric field data types**
        - The most common type, and you'll hear this called either a character, a text or a string
    - **Character type**
        - Means a single character, either a letter or a number that's being stored in a field
    - **Text or string type**

        - This means a grouping of characters that contains contain letters or numbers in this alpha numeric data field type
        - They can be uppercase or lowercase, depending on what we're going to be storing
        - If you store a number inside of an alphanumeric field type, such as a string, you're not going to be able to do mathematical operations on it

    - **Date data type**
        - Store exactly a calendar date in a month, day, year, or day, month year format
        - Can also store the time
        - Are you going to store the date with the time or just the date?
        - Are you going to use two digits for the month, two digits for the day and four digits for the year?
        - Are you going to go with the older convention of using two digits for the month, two digits for the day and two digits for the year?

    - **Number data type**
        - The most basic level is any type of numeric data
        - Means you're not going to be able to use text in there
        - mathematical operations can be performed
        - When you're dealing with numbers, you need to be specific

    - **Currency data type**
        - A special type of number that represents money
        - A number of data type that does allow for two decimal places

    - **Boolean data type**
        - Used for things that have only two values, either a yes or a no, a true or a false an on or an off a one or a zero
        - Used when you're doing logical operations

- **Converting Data**

    - If I give you something that's not stored as a number field, then you're not going to be able to actually do any calculations on that

- Just because somebody has already stored that data in a text field doesn't mean it has to stay as a text data type

- If we can convert it to something that's an integer or a decimal or a floating-point number, and then perform mathematical operations on it

- Data stored as a certain type doesn't mean it always has to stay in that type

- Data is not a static, it is dynamic

- Data can go back and forth into different data types based on your needs and based on the calculations you're going to do

- **Data Structures**
  - **Structured data**
    - Follow an existing convention
    - a lot of data we're going to get doesn't come as structured data
  - **Unstructured data**
    - Any data that is not organized in a predefined manner that meets standards for some kind of structured data
      - **Blob**- Microsoft Azure
      - **Bucket**- Amazon Web Services
    - How do we get this unstructured data into more of a structured manner so that we can then understand it better and make sense of all of this data?
  - **Semi-structured data**
    - A mixture of both structured and non-structured data
    - A webpage, this is a great example of semi-structured data
  - **Structured data** is a specific format in a specific data field type for each particular data you're going to get when you're dealing with unstructured data.
  - **Unstructured data** can be text, images, or video
  - **Semi-structured data** are things like XML files and webpages and zip files and emails and Jaison files

- **Data File Formats**
  - **Delimited file**
    - Files in which some form of character is going to be used to separate each field of data from the other data fields
    - The most common type of this is known as a comma separated value file or CSV
  - **Tab delimited file**
    - This uses a tab, which has five spaces to be able to separate each of the fields in that file
    - Uses a different character, like a pipe, instead of that comma
      - .CSV
      - .TAV
      - .TSV
      - .TXT
        - TXT means a text file and that can use any delimiter you want
  - **Flat file**
    - Any delimited file that is exported out of a database system and can then be sent to somebody else
    - Data has been exported from the database in real time

- **Data Languages**
  - **SQL**
    - The structured query language that is most commonly used when you're interacting with a database and working with the data
    - Uses a series of statements to provide information to the database
      - **Select statement**
        - The way you're going to query information from a database that selects the fields you want to select
      - **Where keyword**
        - Allows you to be able to select something where a certain condition happens

- **HTML**
  - The hypertext markup language
  - The language that we use to write web pages and show them to the world
  - A semi-structured environment and we use tags to be able to dictate what parts of information is being displayed at any given time

- **XML**
  - The extensible markup language
  - Another text-based market language, much like HTML, but its purpose is different
  - It interacts really well with JavaScript
  - The goal here is to transfer data, not to display to the screen

Data Acquisition

Objective 2.1

- **OBJ 2.1:** Explain data acquisition concepts.

- **Data Acquisition**
  - **Data System**
    - Any information technology system that captures processes, stores, queries, or reports on the data contained within it
  -
  - **Extract Transform Load (ETL)**
    - The process that occurs when moving data from a source system to a data warehouse by extracting data from the source, transforming the data and then loading it to the data warehouse
  -
  - **Extract Load and Transform (ELT)**
    - A modern method used when preparing data for data lakes, by holding data in preparation for future transformation

- **Extracting Data**

1. We extract data from its source, wherever that might be
2. We transform it to fit into the scheme of that we want in our database
3. load it into the data warehouse in an already transformed

  - Use ELT if you're using a data lake
  - When data is extracted from the source it is loaded directly into the data

  - **Extracting Data**
    - The process of extracting the source data and importing it into the system
    - The objective of extracting data is to connect in the data source
    - SQL, Power BI, and Power Query are tools for extracting data from external databases
  - **Comma-separated Values (CSV)**
    - The text file that uses commas to separate values

- **Transforming Data**
    - **Transforming Data**
        - The process of transforming data to another table format
        - Timestamps are very useful for end users
- **Loading Data**
    - **Loading Data**
        - The process of loading data from a source system in a data warehouse
            - **Full Load**
                - Loads all the data from the data system
            - **Delta Load**
                - Loads only the new or changed data
- **Application Programming Interface (API)**
    - **Application Programming Interface (API)**
        - Connection between the computers or other programs
        - APIs are designed to present a set of questions and define answers in the system
            - **Pull Model**
                - Continuously pull data into the system
            - **Push Model**
                - Only sends notifications when data changes
        - It's important to understand the frequency of the changes in the data
    - **Web Service**
        - Communication between or among electronic devices
        - JSON and XML are both ways of encoding structured data
            - **Synchronous**
                - Request from the web service and wait for the response
            - **Asynchronous**

        - Allows you to do other tasks while waiting for the response

    - Web service has a specific function to provide different kinds of information

- **Web Scraping**

    - **Web/Data/Screen Scraping**
        - The act of extracting data from a website
            - Selenium
            - BeautifulSoup
            - Scrapy

                - Fragile
                - Legality

        - Before you start scraping someone else's website, make sure you have permission

- **Machine Data**

    - **Machine Data**
        - The data generated by the web servers
        - The machine's data can be used as a predictive maintenance tool

- **Public Databases**
    - Do not include personal information in public data

    - **Aggregated Data**
        - Provides information that has been summarized or compiled

    - **Data.Gov**
        - Hosts various federal agencies under the OPEN Government Data Act

    - **Data.Commerce.Gov**

        - Provides information about the economy, population data, and environmental data

    - **Pew Research Center**
        - Non-government source of information on public issues

    - **Kaggle.com**
        - Has a large repository of publicly available machine learning data sets

    - **GeoPostcodes.com**
        - Provides information on postcodes or ZIP codes

- **Survey Data**

        - Bias is the enemy of a good survey

        - The types of answers are not broad enough to cover the range of options

        - When there is a shade of grey, do not force people to answer yes or no questions

        - When there are nuanced opinions, a single choice of response might not be the right type of survey

        - Make sure that the choices will cover the entire range f possible responses

    - **Likert Scale**
        - Used to scale responses

    - **Text-Based Response**
        - Allows people to share reactions in the form of writing

- **Sampling and Observation**

    - **Observation**
        - The act of collecting data by observing and then analyzing afterwards

- **Sampling**
  - Creating a smaller data set from a larger data set
    - Random sampling
    - Systematic sampling
    - Stratified sampling

Cleansing and Profiling Data

Objective 2.2

- **OBJ 2.2:** Identify common reasons for cleansing and profiling datasets.

- **Cleansing and Profiling Data**
    - **Data Profiling**
        - The process of working with data to begin to discern information and trends present in that data
            - Steps for Data Profiling
            - Identify redundant and duplicate data for consolidation
            - Data Specifications

- **Data Profiling Steps**
    - Remember to convert the data to another format
    - **Steps of Data Profiling**
        1. Identify and document the source of data
        2. Identify the field names and data types
        3. Determine the fields to be identified for reporting
        4. Check for the primary, natural, or foreign keys
        5. Recognize all the data in the dataset

- **Data Profiling Tools**
    - **Data Profiling Tools**
        - Allow both manual techniques or advanced software to start data profiling
            - Power Query in Excel
            - Power BI
            - Tableau

- **Redundant and Duplicated Data**
    - **Redundant Data**

        - Identical data stored in multiple places
        - Determine the redundant data and work on how to minimize it

    - **Duplicated Data**
        - Data repeated within the same dataset
        - To find duplicated data, use the built-in tools in a data analytics software

- **Unnecessary Data**
    - It's important to understand what data you need and what data you can ignore
    - Extra data slows down your system

        - Tools:
            - Excel
            - Power BI
            - Tableau

- **Missing Values**
    - Missing data is referred as **NO**

    - **NO**
        - Represented as blanl fiels, NO or N/A

    - When the value is not applicable to the field
    - When the dataset doesn't have the information
    - When the datasets do not match the expected information
    - When the survey data is incomplete

    - **What can you do about it?**
        - Filter out “NO” values
        - Replace missing values

- **Invalid Data**
    - Invalid Data = Incorrect Data

    - **Different reasons that data could be considered invalid**
        - Hard-coding data
        - Invalid data questions

- Extreme values
- Incorrect data
- Invisible characters
- Look for lead and trailing spaces
- Remove/Replace invalid data

- **ASCII** is a data code inside of a computer system that is invisible or non-printable characters

- **Meeting Specifications**

  - **Specifications**
    - Certain types or quality set by database engineers when designing systems

    - The most common reason that data doesn't meet specifications is wrong data type
    - Another reason is improper storage of numeric characters

- **Data Outliers**

  - **Data Outlier**
    - Any data or piece of data that is outside the normal distance from the other values in a sample

  - **Nonparametric Statistics**
    - Identifies data not assumed to come from a prescribed model that are predetermined by a small number of parameters

      - Parametric
        - Normal baseline

      - Nonparametric
        - Distribution independent

Data Manipulation

Objective 2.3

- **OBJ 2.3**: Given a scenario, execute data manipulation techniques.

- **Data Manipulation**

    - **Data Manipulation**
        - The process of recoding data so that it can be more useful during our processing, correlation, analysis, and reporting

- **Recording Data**

    - Power BI does this in a graphical manner using its conditional column
    - Power Query does this without writing any code using replace values

    - Should you replace the original data or create an additional column with the recoded data**?**

    - **Recoding Data**
        - Transforming data from one form to another

- **Derived Variables**

    - **Derived Variable**
        - A new variable or data point derived or created from existing data

            - **Recoded Data**
                - Transforming data
            - **Derived Variable**
                - Creating new data point

            - **Optimize for Speed**
                - Store derived variables
            - **Optimized for Space**
                - Store formulas

- **Value Imputation**
    - **Data Imputation**
        - Substitutes missing data with estimated values
- **Aggregation and Reduction**
    - Be careful not to introduce biases in sampling data sets
    - **Sampling**
        - Reduces data size by randomly selecting samples from the original larger dataset
            - Simple random
            - Stratified
- **Data Masking**
    - **Index Field**
        - A unique, non-personally identifiable number that cab be uses as a unique identifier
- **Transposing Data**
    - Transposing data can also be called unpivoting data
- **Appending Data**
    - **Appending Data**
        - Combines data from one data set to another data set
    - **Inline Append**
        - Combines data sets together
    - **Intermediate Append**
        - Retains individual data sets, but also creates a new data set with the combined data
            - **Inline**
                - Discards original data set
            - **Intermediate**
                - Keeps original data set

Performing Data Manipulation

Objectives 2.3 and 2.4

- **OBJ 2.3:** Given a scenario, execute data manipulation techniques.
- **OBJ 2.4**: Explain common techniques for data manipulation and query optimization.

- **Performing Data Manipulation**
  -
- **Data Blending**

  - **Data Blending**
    - Takes data and uses different text-based functions to determine how it will be displayed or stored inside a data environment
    - In a directory traversal, an attacker tries to navigate upwards and out of the web document root directory

      - Remote
      - Local

- **Parsing Strings**
  - In general, most text is going to be more unstructured

  - **Delimiters**
    - Characters that can be used to split data, like spaces, commas, periods, pipes, or tabs
    - The power of delimiters will be limited to the source data you have in your data set

- **Date Manipulation**

  - **NOW()**
    - Returns the current date and time of a calculation

  - **TODAY()**
    - Returns the current date of a calculation

- **DATEDIFF()**
  - Calculates the amount of time between two given dates

- **NETWORKDAYS()**
  - Calculates how many business days exist between a start date and an end date
  - This function does not consider holidays or days off that fall on a weekday

- **WEEKDAY([StartDate])**
  - Returns a number 1 through 7 to designate the day of the week

- **WEEKNUM([StartDate])**
  - Returns the number of the week of that year as 1 through 52

- **MONTH([StartDate])**
  - Returns a 1-12 to designate the month of the year

- **Conditional Logic**

  - **Conditional Logic**
    - Any kind of function that checks if there is a logical condition that's being met

      - **IF**
        - A logical function that uses a logical test to validate whether a condition is true or false

      - **ISNULL**
        - Returns a specified value if the expression is null

      - **AND**
        - A logical join function that tests two conditions

      - **OR**
        - A logical function that tests if either one of two conditions is true

        - AND
            - Both conditions must be true
        - OR
            - Either of the conditions must be true

- **Aggregation Functions**

    - **SUM**
        - Adds up all the records to produce a total

    - **COUNT**
        - Counts all the records as individual lines to produce a record count

    - **DISTINCT COUNT**
        - Counts all the records in that column, but will only count the field one time, even if it appears multiple times

    - **AVERAGE**
        - Adds up all the records in a column and divides them by the total record count

    - **MAX**
        - Gives the largest value in each column

    - **MIN**
        - Gives the smallest value in each column

- **System Functions**

    - **System Functions**
        - Any functions that are packaged with your reporting tool or analysis tool to perform certain functions inside of that software
        - Learn how to use different system functions and how to automate them to save you time

Querying & Filtering Data

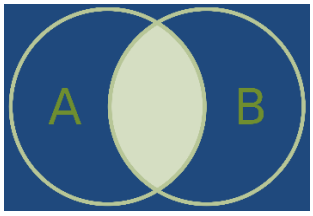
Objective 2.4

- **OBJ 2.4:** Explain common techniques for data manipulation and query optimization.

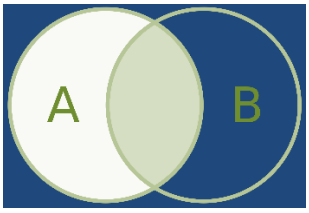
- **Querying & Filtering Data**
  - **Query**
    - A request for data or information from a database table or combination of tables

- **Querying Data**
  - **SQL Server Management Studio**
  - **Tableau**
  - **Microsoft Access**
  - The goal is to have an output that is human readable

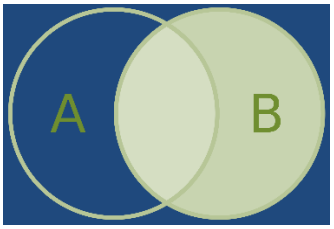
- **Join Types**
  - **Inner Join**



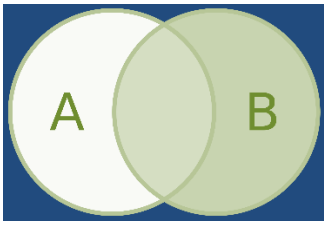
  - **Left Outer Join**



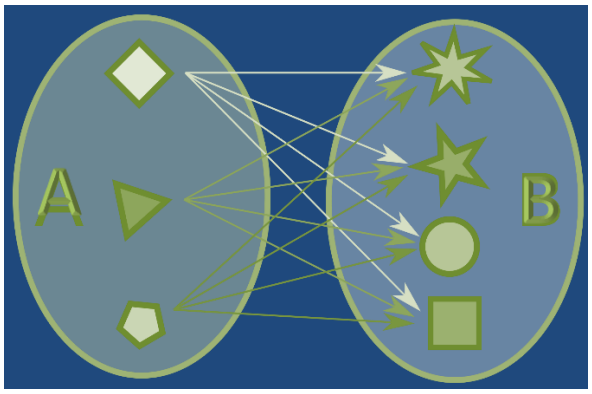
- **Right Outer Join**



- **Full Outer Join**



- **Cross Outer Join**



- **Filtering Data**
  - **Filtering Data**
    - Restricts query to a certain subset of the values in a source data
    - Another reason to filter data is to optimize query performance
- **Parameterization**
  - The concept of replacing values within the query with parameters
- **Indexing Data**

- **Index**
  - Helps speed up queries on a given column within a table
    - Takes up space
    - Takes time to produce
    - Takes time to maintain

- **Temporary Tables**
  - **Temporary Table**
    - A table that just resides in memory on the database
      - Executes much faster
      - Useful for organization purposes

- **Subsets of Records**
  - **Subquery (Nested Query)**
    - A query nested inside another query statement

- **Query Execution Plan**
  - **Query Execution Plan/ Explain Plan**
    - A visual representation that provides details about how the query executes
    - You will find things aren't really necessary by studying the query execution plan

Types of Analysis

Objective 3.3

- **OBJ 3.3:** Summarize types of analysis and key analysis techniques.

- **Determining the Analysis Type**
    - **Hypothesis Statement**
        - A statement that introduces a research question and will be tested by research

            1. What is the source of the data?
            2. Does this data come from a system where it's easy to collect and observe that data?
            3. Are you going to collect this data through a polling system, survey questions, or asking people individually for their responses?

        - An observation uses a smaller sample size instead of collecting every single piece of evidence

- **Exploratory Analysis**
    - The goal is to figure out what type of cleaning, profiling and transformation the data needs
    - Exploratory analysis is all about the initial look at a given data set

- **Performance Analysis**
    - **Performance Analysis**
        - A type of analysis that measures the performance of a particular product, outcome, or scenario against the defined objective
        - Key Performance Indicators (KPIs)
        - Measurements and goals that help identify whether a business is achieving its objectives
            - Qualitative
            - Quantitative

        - Establish KPIs that are realistic which you can meet and achieve
- **Gap Analysis**

    - **Gap Analysis**
        - Analyzes the difference between the present state and a desired or future state
    - **Delta**
        - The change between where you are and where you want to be

- **Trend Analysis**

    - **Trend Analysis**
        - Measures the trend on historical data to predict a future outcome

    - **DISCLAIMER**
        - Past performance is not a guarantee of future results
            - Market research
            - Strategic initiatives
            - Company expenses
            - Revenue

        - Trend can be observed:
            - Short term
            - Long term

- **Link Analysis**

    - **Link Analysis**
        - Determines how a single data point links to other data points

        - Figure out which links will help in achieving your desired outcome

            - 3 main components of Link Analysis
                - **Network**
                    - The set of nodes and all the links
                - **Node**

        - A single point, such as a person, an account or a product
    - **Link**
        - The relationship between different nodes

Descriptive Statistical Methods

Objectives 3.1 and 3.2

- **OBJ 3.1:** Given a scenario, apply the appropriate descriptive statistical methods.
- **OBJ 3.3**: Summarize types of analysis and key analysis techniques.

- **Descriptive Statistical Methods**
    - **Descriptive Statistics**
        - Numbers that summarize data, such as the mean, standard deviation, percentages, rates, counts, and range

- **Central Tendency**
    - **Mean (Average)**
        - The sum of a set of samples divided by the number of samples
        - Mean is susceptible to outliers
    - **Median**
        - Less susceptible to outliers than the mean
    - **Mode**
        - The most common value in a data set

- **Dispersion**
    - **Maximum**
        - The largest value in your dataset

- **Minimum**
  - The smallest value in your dataset

- **Range**
  - The maximum value minus the minimum value
  - Range can be a clue that we have outliers in the data set

- **Standard Deviation**

  - **Variance (σ2)**
    - The average of the squared differences from the mean

  - **Standard deviation**
    - Used to identify outlier

  - **Data points** that lie more than one standard deviation from the mean can be considered unusual

    - Sample Standard Deviation
      - $S=\sqrt{((\sum(x-\bar{x})^2)/(n-1))}$

- **Z-score**

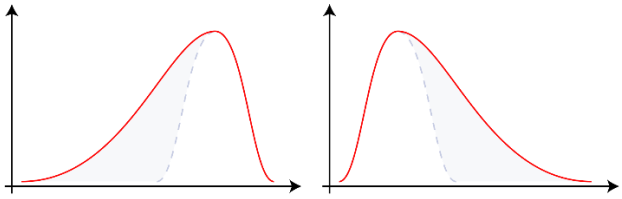
  - **Z-Score**
    - Measures how many standard deviations away from the mean each value is computing
      - $z=(x-\bar{x})/S$

- **Distribution**

  - **Normal Distribution**
    - Occurs when at least 99.74% of the data exists within three standard deviations of the mean

  - **Empirical Rule**
    - Tendency of most data points to fall within three standard distributions of the mean

- **Parametric Data**
  - A data set with at least 99.74% of data falls within three standard deviations of the mean
- **Nonparametric Data**
  - Does not fit a normal distribution
- **Skew**



- **Frequency**
  - **Frequency**
    - Number of times that the given data value appears in the dataset
  - **Frequency Percent**
    - Percentage that that value occurs relative to the entire dataset
  - **Histogram**
    - Shows plotted data points in a bar chart
- **Percentages**
  - **Percentage Change**
    - Measures the change from one data point to another
  - **Percentage Difference**
    - Overall difference relative of the two data points
    - Percentage change could be positive or negative
- **Confidence Interval**
  - Range of values within the confidence level

Inferential Statistical Methods

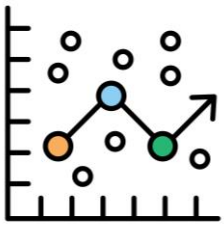
Objective 3.2

- **OBJ 3.2:** Explain the purpose of inferential statistical methods.

- **Inferential Statistical Methods**
  - **Inferential Statistics**
    - The use of statistical methods to try and reach conclusions that extend beyond the immediate data alone
  - Chi Squared

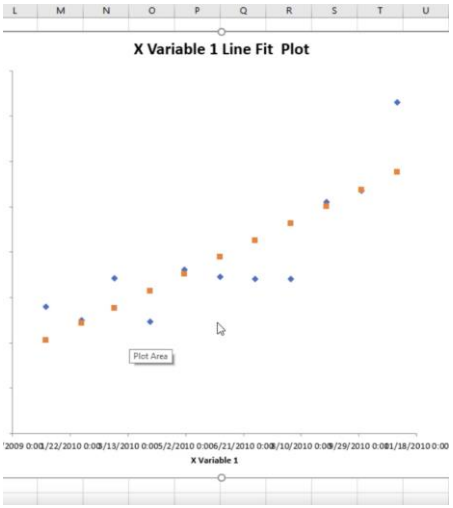
$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

  - Regression Analysis



  - High Degree of Positive Correlation

- X Variable 1 Line Fit Plot



- **T-Tests and P-Values**
  - **T-Test**
    - Compares two groups to determine if there's a significant difference between their means
      - Dependent variable
      - Independent variable
  - **Statistical Significance**
    - Helps determine if the result happened by chance or because of the independent variable that was being tested
  - **P Value**
    - Shows the probability that an observed difference occurred by chance
    - P value should be less than 0.05 or 5% to show significance
- **Hypothesis Testing**
  - **Null Hypothesis**
    - Assumes that there is no relationship between the two variables being tested

- **Alternative Hypothesis**
  - Assumes that there is a relationship between the two variables being tested

- **Type I Error**
  - Occurs when correct hypothesis gets rejected and incorrect hypothesis gets accepted

- **Type II Error**
  - Occurs when incorrect hypothesis gets accepted and correct hypothesis gets rejected
- **Null**
  - No relationship between variables

- **Alternative**
  - Relationship exists between variables

- **Chi-Square**
  - **Chi-Square Statistic**
    - Compares the size of the difference between the expected result and the actual result

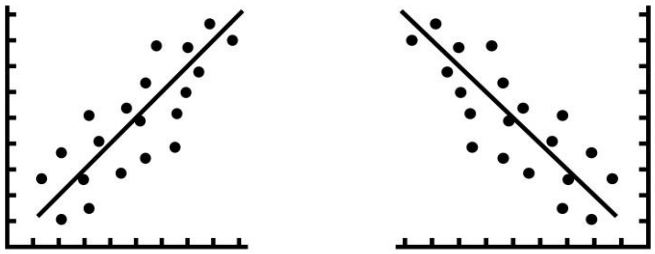
  - **Chi-Square Test**
    - Produces the chi-square statistic and determines if a difference exists between two groups

      1. Compare the actual results to what was expected to see if they're far off
      2. Rule out any observations that might've happened by chance instead of a causation effect

         - Identify confidence in the results
         - Analyze data from a random sample

  - **Test of Independence**
    - Tests against multiple variables

- **Goodness of Fit**
  - Tests against a baseline



- **Regression Analysis**
  - A statistical method that's used to estimate relationships between a dependent variable and one or more independent variables

- **Correlation**

  - **Correlation**
    - The relationship between two or more equal variables

  - **Correlation does not equal causation**
    - In causation, one variable directly changes the other variable

  - **Pearson's Correlation Coefficient (R-Value)**
    - Used to measure the linear relationship between data points

  - **R2 Value**
    - Squared R value which is then converted into a percentile
    - A perfectly positive correlation of 1 means a 100% correlation

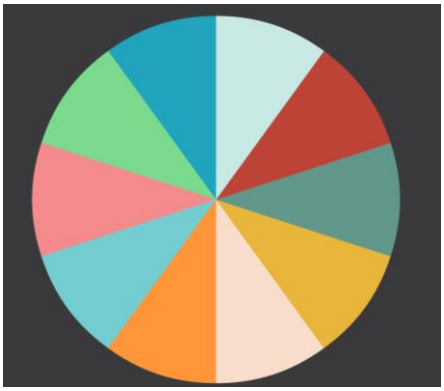
Visualization Types

Objective 4.4

- **OBJ 4.4:** Given a scenario, apply the appropriate type of visualization.

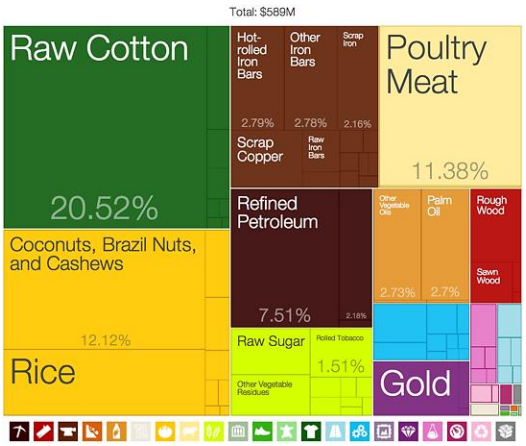
- **Visualization Types**
    - **Data Visualization**
        - The graphical representation of information and data

- **Pie Chart**
    - **Pie Chart**
        - Used to represent the percentages of information



        - Group things together for easy digestion of information

- **Tree Map**
    - **Tree Map**
        - Made for representing hierarchical data

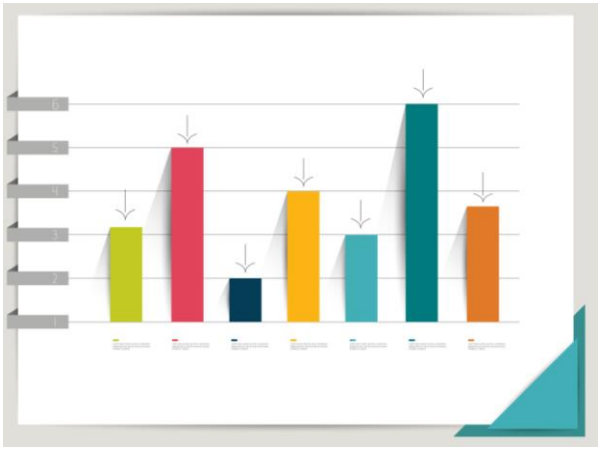


- **For easy reading, only large categories will be labeled**
- **Hierarchical data divides categories into subcategories**

- **Pie Chart**
  - Used for broad categories

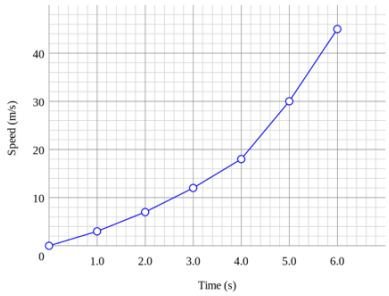
- **Tree Map**
  - Used for hierarchical subcategories

- **Column and Bar Charts**

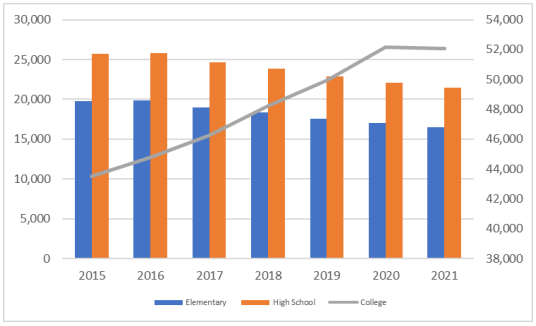


- Column charts are also used in time-based data

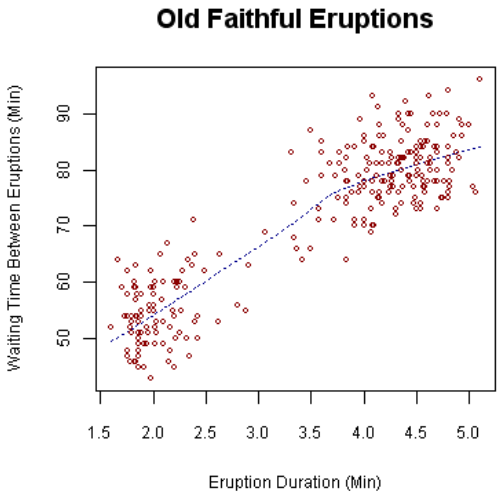
- **Line Chart**
  - **Line Chart**
    - Most used for time-based data



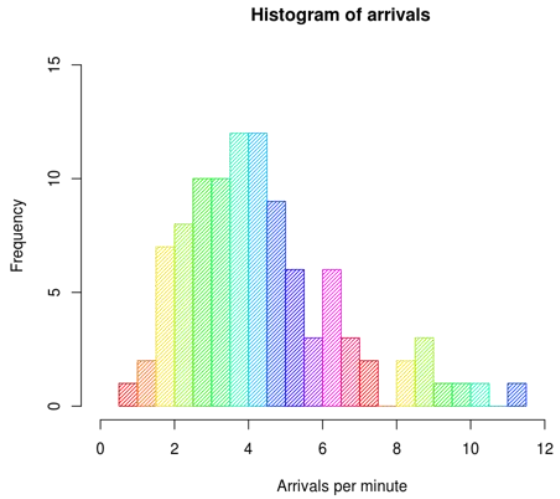
- **Combining Charts**



- **Scatter Plot and Bubble Chart**



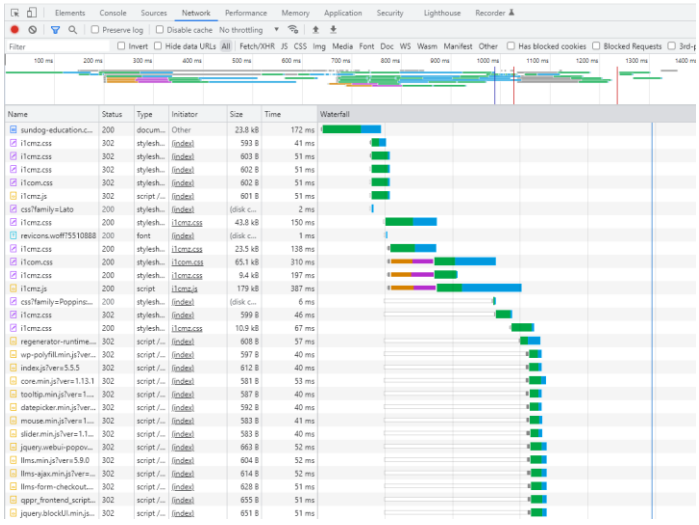
- **Histogram**



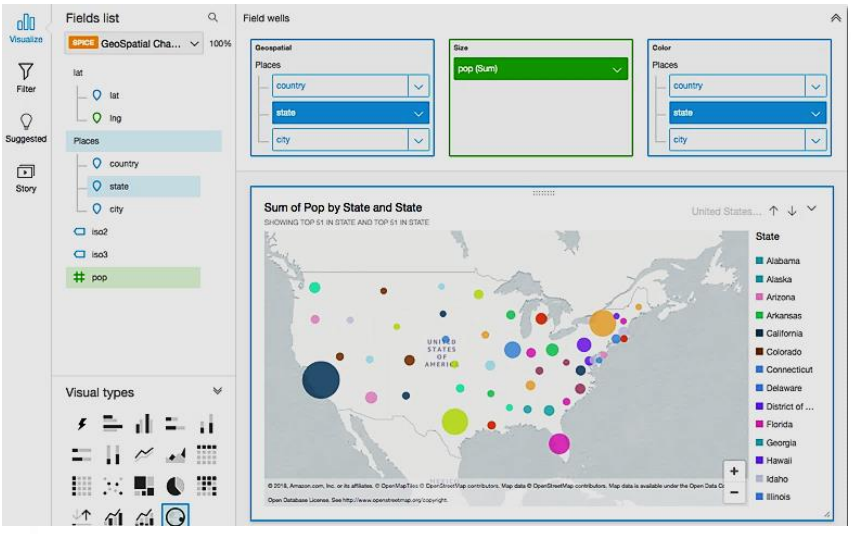
  - Make sure to include any outliers in the bins
  - Make sure not to have too many bins

- **Waterfall**

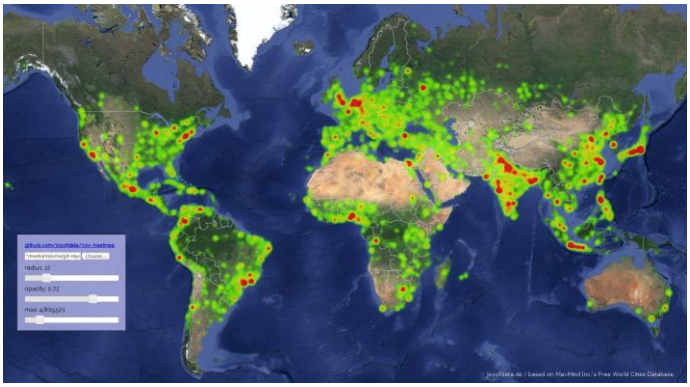
  - **Waterfall Chart**
    - Shows performance over time



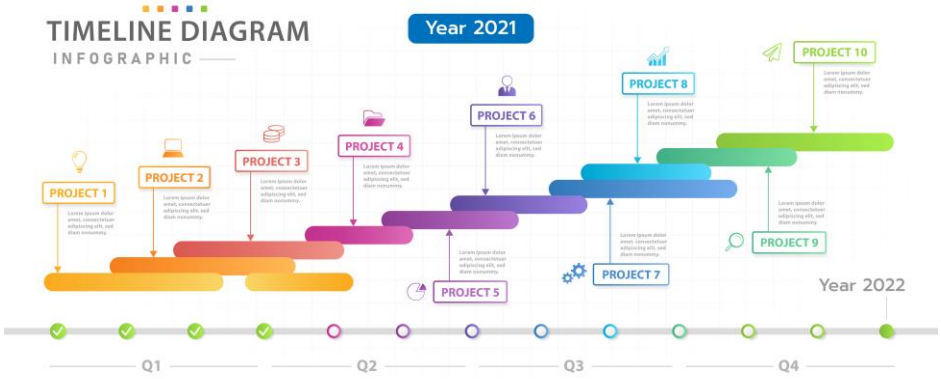
- **Geographic Maps**
  - **Geographic Maps**
    - Building maps with geographical data and visualization tools
  - **ArcGIS**
    - Program used in plotting geographical information
  - **Dot Map**
    - Uses markers to represent specific spots on the map
  - **Filled Map**
    - Uses shading to fill the borders of a geographic area
  - **Layered Map**
    - Used to convey two different dimensions of information at the same time



- **Heat Maps**
  - **Heat Map**
    - **Uses colors to highlight important data**



- **Word Clouds and Infographics**
    - **Word Cloud**
        - Word representations that are displayed in various sizes
    - **Infographic**
        - Graphics that are mostly used for marketing purposes
            - Adobe Illustrator/InDesign
            - Microsoft Word
            - Canva



Creating Reports

Objectives 4.1, 4.3 and 4.5

- **OBJ 4.1:** Given a scenario, translate business requirements to form a report.
- **OBJ 4.3**: Given a scenario, use appropriate methods for dashboard development.
- **OBJ 4.5:** Compare and contrast types of reports.

- **Creating Reports**
    - **Reporting**
        - The process of collecting and submitting data for analysis, and then taking the resulting analysis and creating a final product to inform decision-making in an organization

- **The Audience**
    - **Distribution List**
        - The list of people who will receive the report or dashboard being prepared
            - C-level executives
            - Management
            - External vendors/stakeholders
            - General public
            - Technical experts
    - C-level executives always look at the big picture
    - Know what data can be shared and not be shared with external vendors or stakeholders
    - Technical experts also want to know the methods used to get the final values

- **Data Sources**
    - **As you create a project, always document your data sources**

1. Account for any changes in roles that may occur
2. Help in troubleshooting any kind of issues that may arise
3. Have a place to pick up from when going back to a project
4. Answer questions about where the data comes from

- **Data Models**
    - **Data Model**
        - A model that organizes data and the relationships of data elements
        - The goal is to narrow down the amount of information and data that people need

- **Data Fields**
    - **Field Definition**
        - Used to clarify what information each field contains
            - Define data points
            - Use keys or legends
                - **Dimension Attribute**
                    - Categorical data
                - **Measurement Attribute**
                    - Numerical data
        - Data fields can be used as part of the final report based on the attributes they contain

- **Data Delivery**
    - **Different reporting tools**
        - **Power BI**
        - **Crystal Reports**
        - **Tableau**

- **Reporting Frequency**

- **Reports** are only as valid and up-to-date if the data is refreshed at frequent intervals

- **Recurring Report**
  - Any report that is repeatedly run on a specific date and time

    1. When will the report be created?
    2. What will be the output of the report?
    3. Who will receive the report?

- **Report Types**

  - **Static Report**
    - Report that is not automatically updated

  - **Point-in-Time**
    - Results of a report at a given time

  - **Dynamic Report**
    - Report that can be accessed on a regular time or on-demand basis

  - **Real-time Reporting**
    - Occurs when receiving up-to-date data

  - **Operational Reports**
    - Used to provide information about the status of projects, products, or organizations

  - **Compliance Reports**
    - Used for regulatory requirements

  - **Tactical Dashboards**
    - Type of dynamic report that provides operational details of a process

  - **Research-driven Reports**
    - Reports that are used to inform and change business practices

- **Ad Hoc Reports**
  - Reports that are generated for a one-time request

- **Self-service Reports**
  - Allows the end-users to build their own report

Creating Dashboards

Objectives 4.1, 4.2 and 4.3

- **OBJ 4.1:** Given a scenario, translate business requirements to form a report.
- **OBJ 4.2**: Given a scenario, use appropriate design components for reports and dashboards.
- **OBJ 4.3**: Given a scenario, use appropriate methods for dashboard development.

- **Dashboard Development**
    - **Dashboards**
        - A tool used to track, analyze, and display data to gain deeper insights into the overall wellbeing of an organization
- **Data Filtering**
    - **Natural order**
        - The order the data was input into the database
    - **Multi-sorting**
        - **Sorting within a sort**
            - **Top-end and bottom-end sorts**
                - Top-end and bottom-end concepts are both sorting and filtering operations
            - **Custom sorts**
        - **Interactive filter**
            - Adjust a slicer or a filter operation in your dashboard to narrow things down
- **Data Tables**
    - **Data Table**
        - Shows different variations of how data is represented
- **Dashboard Design**

- Don’t pick colors at random
- Make sure the chosen color scheme remains consistent

- **Documenting Dashboards**

  - Make sure the dashboards you're producing are very easy to understand
    - Change the query
    - Use transformation step
    - Show labels and legends

- **Documentation Elements**

  - **Report Header**
    - Appears only on the first page at the very top of the report that would include the title and the version number

  - **Page Header**
    - Appears at the top of every page of the report that may include field headings

  - **Page Footer**
    - Appears at the bottom of each page that may include page numbers or references

  - **Report Footer**
    - Appears at the very end of the report on the last page that may include the summary information or credits to the report writers

  - **Watermarks**
    - A good way to show end user some restrictions on how they're expected to use the report

  - **Refresh Date**
    - Report data was last updated

  - **Print Date**

- Report was printed out

- **Report Elements**
  - **The report should have:**
    - Title
    - Summary
    - Key findings
    - Significance
  - Also include key **definitions** and **calculations** on the report
  - **When presenting data, you need:**
    - Cover page
    - Instructions
    - FAQs
    - Appendix

- **Dashboard Optimization**
  - **Dashboard Optimization**
    - The art of keeping the dashboard simple and concise
      - **Visual filter**
        - Expand and collapse information on the dashboard
      - **Drillthrough**
      - **Tooltips**
        - Power BI and Tableau can create complex tooltips
  - How often is the dashboard being updated?
  - Performance of the dashboard being created
  - The impact of the security features
  - The speed the report is generated

- **Deploying Dashboards**
  - Instructions and documentation are prepared
  - Labels are clear and correct
  - The dashboard is tested
  - Approvals and permissions are checked
  - Software tools used are licensed

Data Governance

Objective 5.1

- **OBJ 5.1:** Summarize important data governance concepts.

- **Data Governance**
    - **Data Governance**
        - A term used to describe the capability of an organization to ensure that high quality data exists through the complete lifecycle of the data and that data controls are implemented to support its business objectives

- **Data Lifecycle**
    - **Data Lifecycle**
        - Period of time the data exists in a system

            ↓ **Creation**
            ↓ **Storage**
            ↓ **Use**
            ↓ **Archive**
            ↓ **Destruction**

        - **Data Creation**
            - Occurs when existing data is produced outside and imported automatically to the system

        - **Data Entry**
            - Occurs when information is manually typed into the system

        - **Data Capture**
            - Occurs when data is generated by a device into the organization

        - **Data Storage**
            - Occurs when data is not being actively used
            - When data is stored, it has to be protected by the permissions of its users

- **Data Use**
  - Used in viewing, processing, modifying, manipulating, or saving the data

- **Data Archival**
  - Copying and storing of data that can be used when needed
  - Data stored in data archives can be accessed, when necessary, but it takes longer than usual

- **Data Destruction**
  - When the data is no longer valuable or has reached its useful life and needs to be destroyed

- **When writing data governance policies, it's important to consider the data lifecycle**

- **Data Roles**
  - **The stakeholders can provide the proper classification for the data belonging to an organization**

    - **Data Ownership**
      - The process of identifying the person responsible for the confidentiality, integrity, availability, and privacy of information assets

    - **Data Owner**
      - A senior executive role who is responsible for maintaining the confidentiality, integrity, and availability of information assets

    - **Data Steward**
      - The person responsible for ensuring data is properly labeled, identified, collected, and stored

    - **Data Custodian**
      - A role that's responsible for handling the management of the system on which the data assets are going to be stored

- **Privacy Officer**
  - A role that's responsible for the oversight of any kind of privacy-related data
    - Data minimization
    - Data sovereignty
    - Data retention
    - Data destruction

- **Regulations and Compliance**

  - **Regulation**
    - Any rule implemented by an authority backed by law

  - **Compliance**
    - Shows that an organization meets the requirements for a given regulation rule or law

  - **Data Sovereignty**
    - Jurisdictional control or legal authority on collected or stored data that may be imposed by countries and states
    - Rules around data sovereignty are changing rapidly all the time

- **Data Classification**

  - **Data Classification**
    - A way to categorize or classify data based on value and sensitivity

      - **Data Classification (Commercial Sector)**
        - **Public**
          - Front-facing website
        - **Sensitive**
          - Financial data
        - **Private**
          - Personnel records
        - **Confidential**
          - Intellectual property

- **Data Classification (Government Sector)**
  - **Unclassified**
    - Front-facing website
  - **Controlled Unclassified Information (formerly SBU)**
    - Medical records
  - **Confidential**
    - Trade secrets
  - **Secret**
    - Military plans
  - **Top Secret**
    - Weapon blueprints

- **Data Type**
  - A tag or a label that's used to identify a piece of data under a subcategory of a classification

    - Personally identifiable information
    - Protected health information
    - Personally identifiable financial information
    - Intellectual property

- **Access Requirements**

  - **Non-Disclosure Agreement (NDA)**
    - Defines what data is considered confidential and cannot be shared outside of the relationship

  - **Acceptable Use Agreement**
    - Describes how data can be used and for what purposes that data can be used

  - **Memorandum of Understanding (MOU)**
    - A nonbinding agreement between two or more organizations to detail the rules, roles, and expectations for both parties
    - An MOU is often referred to as a letter of intent

- **Data Retention and Destruction**
    - **Data Retention**
        - Maintains the existence and control of data to comply with business policies and/or applicable laws and regulations
        - Consult with a legal counsel when developing these data retention policies
    - **Data Preservation**
        - Keeps information for a specific purpose outside of the data retention policy
    - **Data Removal**
        - Any process that deletes data or makes it inaccessible
        - Data removal should only be used with the least sensitive types of data
    - **Data Destruction**
        - Deletes data and tries to destroy the underlying data
    - **Data Sanitization**
        - Performs a verification function to ensure data has been wiped and no longer accessible
- **Data Processing**
    - **Data Processing**
        - **Occurs when data is collected and translated into usable information**
            - **Transaction processing**
                - Used for large volumes of information to be processed synchronously
            - **Distributed processing**
                - Takes large volumes of datasets and distributes them to multiple servers

- **Real-time processing**
  - Provides real-time output that can change things in the middle of the process

- **Batch processing**
  - Used to process large amount of data at one time

- **Multiprocessing**
  - Uses multiple processors to work on a single dataset

- **Data Security**

  - **Data Encryption**
    - Process that uses algorithms that will scramble the data into another form called ciphertext
    - Encryption key is used to unencrypt the ciphertext

  - **Data at Rest**
    - The data is being stored in a ciphertext format
    - Data at rest needs to be encrypted before it is stored again
  - **Data in Transit**
    - Data that is transferred or stored in or from another system
  - **Data In Use**
    - Data that is currently being processed
    - Data can be protected when at rest, in transit, and in use using encryption

  - **De-Identification**
    - Process of removing fields that could be used to identify an individual from a dataset
  - **Data Masking**
    - Used to minimize the amount of data shared by limiting the information shown on the system
  - **Data Breach**
    - Occurs when the information is read, modified, or deleted without proper authorization

- **Data Access**
    - **Data Access**
        - Focused on the permissions to access data based on data governance policies
            - **Read/write permissions**
                - Access to read data or to read and write the data
            - **Role-based permissions**
                - Access assigned to a person based on their job function
                - Different employees have different levels of access based on job function
            - **User group permissions**
                - Access given to people based on their user groups
                    - **Role-Based**
                        - Based on the job function
                    - **User Group**
                        - Based on the group function
- **Data Storage**
    - **Shared Drives**
        - Network hard drive that can be accessed within the organization
    - **Cloud Drives**
        - Externally shareable drive that are stored on a cloud server
    - **Local Storage**
        - Refers to the machine's local hard drive, external hard drive, and USB thumb drive
        - When storing data, think about who is going to need access to it
        - When storing on a local drive, make sure to create backups
- **Entity Relationships**
    - **Entity Relationship**
        - Provides an overview of the different systems that have data

- **Conceptual Data Model**
  - Shows where the data exists and how that data is related to other pieces of data in the system

- **Logical Data Model**
  - Shows a more detailed view that Includes data fields and the relationships between them

- **Physical Data Model**
  - Shows the actual data systems with their tables, relationships, fields, and attributes

- **Record Linkages**
  - Identify records corresponding to matching or merging records that go between different data sets, but are all linked together

- **Record Link Restriction**
  - Prevents protected datasets from combining with other data sets

- **Data Constraint**
  - Integrity rules that limit the types of data that can go into a column or table within a database system

- **Data Integrity**
  - The existence of accurate and consistent data in a database

    - **Data Integrity Rules**
      - **Domain Integrity**
        - **Accepted field values**
      - **Entity Integrity**
        - **Unique record identifier**
      - **Referential Integrity**
        - **Data integrity between tables**
      - **User-Defined Integrity**

- **Based on own business rules**

Data Quality

Objectives 5.2 and 5.3

- **OBJ 5.2:** Given a scenario, apply data quality control concepts.
- **OBJ 5.3:** Explain master data management (MDM) concepts.

- **Quality Checks**
    - **Data Validation**
        - Format and structure of the data
    - **Data Verification**
        - Accuracy of the data
    - **Sources of Error**
        - Data Transfers
        - Human Error
        - Incorrect Data Joining
        - Data Transformations
        - Data Manipulations
        - Mergers and Acquisitions
- **Quality Dimensions**
    - **Data completeness**
        - No data is missing
    - **Data accuracy**
        - Not sure if data is accurate
    - **Data consistency**
        - Data is entered consistently
- **Quality Rules and Metrics**
    - Set a concrete KPI and check if you are meeting them
    - Set the rules

- **Data Validation**
    - **Data profiling**
        - Checking results against reasonable expectations of what the results should be
    - **Cross validation**
        - Comparing your results across multiple data sets
    - **Peer review**
        - somebody else will double check your data
    - **Data audits**
        - Somebody will dive deep on your data and make sure that you have all the data that you needed to conduct the analysis
- **Automated Validation**
    - Automated Validation makes sure that the data is being validated upon entry
    - Transferring data between two systems is an opportunity for automated validate
        - Failing to import
        - Duplicated records
    - Enforcing uniqueness on the table that is being imported is another example of automated validation
    - **Automated Validation**
        - Using software to validate your data at some point, as opposed to trying to detect it
- **Data Verification**
    - **Data Validation**
        - Format and structure of the data
    - **Data Verification**
        - Accuracy of the data

- **Field Level Data**
  - Make sure that the actual data is correct, true, and accurate

- **Record Count**
  - Compare the record counts with the data to ensure they match

- **Calculations**
  - Double check the tool used to have the right calculations
    - Output of the calculation
    - Formula and equation are correct

- **Report Visuals**
  - Double check the charts, graphs, captions, and columns used in the data
  - Avoid visual inconsistencies by making sure that colors used in charts or graphs are consistent

- **Master Data Management (MDM)**

  - **Master Data**
    - Golden record or the master file for dimensional data types

  - **Master Data Management**
    - Refers to how we maintain the master data

  - **Reltio or Informatica**
    - Software packages that do master data management

    - **Master data management and data governance go hand in hand**
      - Compliance with regulations and mandates
      - Enforcing data integrity and data quality
      - Streamlining access to data
      - Automatically populating datasets consistently

- **Streamlining Data Access**

  - **How do we create these golden records?**
    - Consolidation of multiple data fields

        - Fill in the gaps of missing data with data from other systems
    - Field standardization
    - Data dictionary
        - Provides quick access to available data and how it relates to other data within the organization

Data Analytic Tools

Objective 3.4

- **OBJ 3.4:** Summarize types of analysis and key analysis techniques.

- **Data Languages**
  - **Structured Query Language (SQL)**
    - A programming and query language used to manage relational database management systems
  - **Python**
    - A high-level, interpreted, general-purpose programming language
  - **R**
    - A programming language for statistical computing and graphics that is used by data miners, statisticians, and data analysts

- **Data Transformation Tools**
  - **Microsoft Excel**
    - A spreadsheet software that has calculation capabilities, graphing tools, pivot tables, and an embedded macro programming language
  - **Tableau**
    - A leading data visualization tool used for data analysis and business intelligence
  - **Power BI**
    - An interactive data visualization software product developed by Microsoft with primary focus on business intelligence
  - **Rapid Miner**
    - A data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics

- **Data Visualization Tools**
    - **Tableau**
        - A leading data visualization tool used for data analysis and business intelligence
    - **Power BI**
        - An interactive data visualization software product developed by Microsoft with primary focus on business intelligence
    - **Qlik**
        - A business analytics platform used for big data analytics, combining data sources, and visualizing report data
    - **ArcGIS**
        - A cloud-based mapping and analysis solution used to make maps, analyze data, and to share and collaborate data
    - **AWS Quick Sight**
        - A cloud-scale business intelligence service that is used to deliver easy-to-understand insights to an organization's users in a single data dashboard that combines multiple datasets
- **Statistical Tools**
    - **Statistical Analysis Software (SAS)**
        - A statistical analysis software used for data management, advanced analytics, multivariate analysis, business intelligence, criminal investigation, nd predictive analytics
    - **IBM SPSS**
        - A statistical software suite developed by IBM for data management, advanced analytics, multivariate analysis, business intelligence, criminal investigation
    - **Stata**

        - A general-purpose statistical software package for data manipulation, visualization, statistics, and automated reporting

    - **Minitab**
        - A statistics package used to centralize and organize data, plan and visual data, and report on that data

- **Reporting Tools**

    - **SQL Server Reporting Services (SSRS)**
        - A set of on-premise tools and services used to create, deploy, and manage mobile and paginated reports
    - **Crystal Reports**
        - A business intelligence application used to create a pixel-perfect, powerful, richly formatted, and dynamic reports from virtually any data source
    - **Power BI Report Builder**
        - An interactive data visualization software product developed by Microsoft with primary focus on business intelligence

- **Platform Tools**

    - **Platform Tools**
        - Products made by software vendors that combine lots of different things in one centralized tool

    - **Business Objects**
        - An enterprise software platform for business intelligence that specializes in providing reporting and analytics tools

    - **Micro Strategy**
        - A platform tool that is a business intelligence application
        - Used to create interactive dashboards, scorecards, highly formatted reports, ad hoc queries, thresholds, and alerts, and automated report distribution

    - **Oracle Apex**

- A low code platform that enables developers to build apps in a single extensible platform and interact with an Oracle database data store

    - Low code
        - Let data analysts create their own applications

- **Data Roma**
    - A marketing cloud intelligence platform that's used to connect, analyze, and take action on our marketing data in one dashboard.

- **IBM Cognos**
    - A web-based integrated business intelligence suite that provides a tool set for reporting analytics, scorecarding and monitoring of events and metrics

- **Rapid Miner**
    - A data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics

- **Oracle Analytics**
    - A cloud native service that provides the capabilities required to address the entire analytics process from data ingestion and modeling to data preparation and enrichment to visualization and collaboration
        - Oracle
            - One of the biggest players in the database environment

- **Domo**
    - A cloud-based platform designed to provide direct simplified real-time access to business data for decision makers across the company with minimal it involvement

- **Microsoft Power Platform**
    - A platform that's used to conduct business intelligence, app development, and app connectivity using low code programming languages for expressing logic across the power platform
        - Power BI

- Power apps
- Power automation