Google Data Analytics Notes

Overview -

1. Foundations :
    - Responsibilities of a data analyst.
    - Spreadsheet ,database, visualization basics
    - Showing trends and patterns with data visualization
    - Ensuring your data analysis is fair.
2. Ask :
    - Use of analytics for data driven decisions.
    - Spreadsheet formulas and functions.
    - Dashboard basics and introduction to tableau.
    - Data reporting basics.
3. Prepare :
    - Different data types, fields and values.
    - Accessing Database and importing data
    - Bias and credibility
    - SQL functions
    - Metadata
    - Organizing and protecting data
4. Process :
    - Data integrity
    - Clean and dirty data
    - Cleaning small datasets using spreadsheet
    - Cleaning large datasets by writing SQL queries.
    - Documenting data cleaning processes.
5. Analyze :
    - Data organization
    - Spreadsheet calculation and pivot tables.
    - SQL queries
    - Temporary tables
    - Data validation - Converting and formatting data
6. Share :
    - Creating and visualizations and dashboards in tableau
    - Data driven storytelling
    - Strategies for creating an effective data presentation
7. Act :
    - Function in R

- Accessing data
- Cleaning data
- R visualization tools
- R markdown for documentation , creating structure and emphasis.

8. Capstone project

1. **Course - Foundations Data Data Everywhere -**
   Data analysis is the collection, transformation, organization of data in order to draw conclusions, make predictions and drive informed decision making.

   6 steps of Data analysis.
   a. Ask
   b. Prepare
   c. Process
   d. Analyze
   e. Share
   f. act

   Data Ecosystems - Various elements that interact with one another in order to produce , manage, store, organize and share data
   These elements include hardware and software tools and people who use them.
   Data analyst use data driven decision making and follow a step by step process -
   1. **Ask** questions and define the problem.
   2. **Prepare** data by collecting and storing information.
   3. **Process** data by cleaning and checking information.
   4. **Analyze** data to find patterns, relationships and trends.
   5. **Share** data with your audience.
   6. **Act** on the data and use the analysis results.

   Same process
   1. Ask :- business challenges /objective/questions
   2. Prepare :- data generation, collection, storage and data management.
   3. Process :- data cleaning /data integrity
   4. Analyze :- data exploration, data visualization and analysis

5. Share:- communicating and interpreting results.
6. Act:- putting your insights to work to solve the problem.

Analytical Skills

Qualities and characteristics associated with solving problems using facts.

a. Curiosity :- new challenges and experiences
b. Understanding context :- condition in which something exists or something happens.
c. Having a technical mindset :- ability to break down things in smaller steps and and work with them in an orderly and logical way.
d. Data design :- how you organize information.
e. Data strategy : - management of people, process and tools used in data analysis.

Data Life Cycle -

1. Plan : decide what kind of data is needed, how it will be managed and who will be responsible for it.
2. Capture: collect data from a variety of different sources.
3. Manage : care for and maintain data. This includes determining how and where it is stored and the tools used to do so.
4. Analyze: use data to solve problems, make decisions and support business goals.
5. Archive: keep relevant data stored for long-term and future use.
6. Destroy : remove data from storage and delete any shared copies of data.
7.

2. Course - Ask Questions to make Data Driven Decisions

Structured Thinking - The process of recognizing current problems or situations, organizing available information, revealing gaps and opportunities and identifying the options.

Quantitative and qualitative data -

1. Quantitative data : related to numbers
   E.g : how many positive reviews and negative reviews?
2. Qualitative data : related to context
   E.g. : - Why are the most frustrating reviews ?

Data formats:

1. **Primary vs secondary :**

   Primary - first hand source

   Secondary - gathered by other people

2. **Internal vs external :**

   Internal - data lives inside a company's own system.

   External - data lives outside of the company.

3. **Continuous vs discrete :**

   continuous : data that is measured and can have almost any numeric value.

   E.g:- height of kids in 3rd grade class (52.5 inches,64.2 inches), Runtime marker in video,
   Temperature

   Discrete : data that is counted and has a limited number of values.

   E.g :- number of people who visit the hospital on a daily basis(10,20,200), rooms maximum capacity allowed, tickets sold in the current month.

4. **Qualitative vs quantitative :**

   Qualitative : - measure of qualities and characteristics

   E.g: - exercise activity most enjoyed, favorite brand, fashion preference of young adults.

   Quantitative : -measure of numerical facts.

   E.g:- percentage of board certified women doctors, population of elephants in africa, distance from earth to mars.

5. **Nominal vs ordinal : -**

   Nominal :- a type of qualitative data that isn't categorized with a set order.

   E.g:-

   1. first time customer, returning customer, regular customer.
   2. New job applicant, existing applicant, internal applicant.
   3. New listing, reduced price listing, foreclosure

   Ordinal : - a type of qualitative data with a set order or scale.

   E.g:-

   1. Movie ratings(number of stars : 1 star, 2 star, 3 star)

2. Income level (low income, middle income, high income)

1. **Structured vs Unstructured :-**

   Structured: - data organized in certain format like rows and columns
   E.g : expense report, tax returns, store inventory
   Unstructured data :- data isn't organized in an easily identifiable manner.
   E.g:- social media posts, emails, videos

Data Anonymization -

Data anonymization is one of the ways that we can keep data private and secure.

- Here is the list of data that can be anonymized
  1. Telephone number
  2. License plate
  3. Name
  4. Social security number
  5. Email id
  6. etc

Metadata - Metadata is data about data , that is deep, it tells you from where data come, when and how it was created and what its all about.

3 common types of the metadata

1. Descriptive
2. Structural
3. Administrative

- Metadata repository :- is a database created specially to store metadata.

3. **Prepare Data for Exploration** -

- **Statistical power** : - The probability of getting meaningful results from tests.
- **Hypothesis testing** : -a way to see if a survey or experiment has meaningful results.
- **Statistically significant** :- if the test is statistically significant , it means the results of the test are real and not an error caused by random chance.

- **Proxy data** : - sometimes data isn't readily available .this is when proxy data is useful.
- **Sample size** :-a part of the population that is representative of the population.
- **Confidence level** : the probability that your sample size accurately reflets the greater population.

  E.g: if company want confidence level 90%

  Margin of error = 10% or margin of error =3%

  It is not like confidence level and margin of error should add 100%

Data Cleaning:

- **Dirty data** : data that is incomplete , incorrect and irrelevant to the problem you are trying to solve.

  Types of Dirty Data : -
- Duplicate data
- Outdated data
- Incomplete data
- Incorrect/inaccurate data
- Inconsistent data

- **Clean data** : data that is complete , correct and relevant to the problem you are trying to solve.

SQL -

- Read question of query carefully if they ask for what is customer_name on row 12 then you should use

  (limit by 12)in your query
- In sql to avoid duplicates use distinct.
- For string cleaning in sql use trim() and distinct() functions.
- Length():- length function gives the total number of characters in the string.

  Syntax: - length(name)

- Substring(string, starting point, how many character :-

  E.g : substring(name,1,5)

  E.g:2:-

```
SELECT
customer_id,
```

```
substr(state,1,2) as new_state
FROM
customer
ORDER BY
state DESC
limit 9
```

- **Trim()-** a function that removes leading , trailing and repeated spaces in data.
- **Cast() :-** cast() can be used to convert anything from one data type to another.

  E.g:

```
SELECT cast(purchase_price as float64)
 FROM customer_data.customer_purchase
 order by cast(purchase_price as float64)
 LIMIT 1000
```

```
E.g:- date conversion

select cast(date as date) as date_only, purchase_price
from customer_data.customer_purchase
where
cast(date as date) between '2020-12-01' and '2020-12-31'
```

- **Typecasting** : - converting data from one type to another.
- **Concat() :-** adds string together to create new text strings that can be used as unique keys.

  E.g:-

```
select concat(product_code, product_color) as
new_product
from
`logical-factor-357715.customer_data.customer_purchas
e`
where product ='couch'
```

- COALESCE() :- can be used to return non null values in a list.

- E.g:-here logical factor is the project name and customer_data is dataset and customer_purchase is the table name

```
select coalesce(product,product_code) as product_info
from
`logical-factor-357715.customer_data.customer_purchase
`
```

This query will return product but if product is null then it will return product_code.

How to correct the most common problems?

Make sure you identified the most common problems and corrected them, including:

- **Sources of errors**: Did you use the right tools and functions to find the source of the errors in your dataset?
- **Null data**: Did you search for NULLs using conditional formatting and filters?
- **Misspelled words**: Did you locate all misspellings?
- **Mistyped numbers**: Did you double-check that your numeric data has been entered correctly?
- **Extra spaces and characters**: Did you remove any extra spaces or characters using the **TRIM** function?
- **Duplicates**: Did you remove duplicates in spreadsheets using the **Remove Duplicates** function or **DISTINCT** in SQL?
- **Mismatched data types**: Did you check that numeric, date, and string data are typecast correctly?
- **Messy (inconsistent) strings**: Did you make sure that all of your strings are consistent and meaningful?

- **Messy (inconsistent) date formats**: Did you format the dates consistently throughout your dataset?

- **Misleading variable labels (columns)**: Did you name your columns meaningfully?

- **Truncated data:** Did you check for truncated or missing data that needs correction?

- **Business Logic**: Did you check that the data makes sense given your knowledge of the business?

Changelog - A changelog is a document used to record the notable changes made to a project over its lifetime across all of its tasks.