Google: Data Analytics

Course: 1

Foundations: Data, Data, Everywhere

Analytical skills

* Qualities and characteristics accosiated with problem solving using facts

Types of Analytical skills

1. Curiousity: will to learn and accept new challanges
2. Understanding context: The condition in which something exist.
3. Technical mindset: break things in smaller steps
4. Data Design: The way of organizing information
5. Data Strategy: Management of things.

Analytical thinking

* Defining a problem and solve it in a step by step way.

5 Key aspects of Analytical thinking

1. Visualization: Graphs, maps, design elements, plots
2. Strategic mindset: A step by step manner
3. problem oriented mindset: Always think about problem
4. Correlation finding: Always identify, find correlations
5. Big picture thinking: look from a different, large view.

Data Analysis Process phases

* Ask → * Prepare → * Process → * Analyzes → * Share → * Act.

Life Cycle of data

* Plan → * Capture → * Manage → * Analyze → * Archive → * destroy

Ask

* Ask questions about problem from stakeholders

Prepare

* Collect and store data

Process

* Clean the data (data preprocessing technique)

Analyze

* Do analyses on the clean and prepared data
  → Use tools (we will learn spreadsheet and SQL)

Share

* share the analyzes results with stakeholder

Act

* stakeholder take some action based on your analysis

Spread Sheets

- Microsoft Excel
- Google sheets

Query language

- SQL

Visualizations

- Tableau
- Looker

Spread Sheets

* Columns are ordered alphabetically
* Rows are ordered numerically
* you can access cells by it the column and row.
  e.g. first cell = A1
  second " = A2

* Text wrapping options in spreadsheets
* Colums labels (names) are called attributes
* Sort data in spreadsheets
* Formulas = e.g: $\boxed{=A1+A2+A3}$

SQL

* Use it for larger data
  - you can use SQL with following databases:
    - → Oracle
    - → MySQL
    - → PostgreSQL
    - → Microsoft SQL Server
    - → BigQuery
* A basic query is = Select from where

Visualization

* we can use spreadsheets for visualization

Issue

* A topic to investigate

Question

* used to discover information

Problem

* An obstacle to be taken care of

C #1 ⇒ Complete

Course #2: Ask

Ask: define problem properly by asking Question

Don't ask leading or closed ended questions
" " vague questions.

* **Quantitative Data** => Measurable, The what, How many, How often
* **Qualitative Data** => Unmeasurable, helps to answer Why, e.g = reviews

* **Reports** => use preset data
* **Dashboard** => use live data

* To meet a specific goal we use specific metrics

* At the beginning of every projects, Ask these 3 questions
  - Who are primary & Secondary stakeholders?
  - Who is managing data?
  - Who can help you with queries?

* To communicate, Ask these 4 questions
  - Who is your audience?
  - What they already know?
  - What they want to know?
  - What is the best way to communicate?

C-2 / Complete

Course #3 — Prepare data

Types of data

- **Qualitative**
  - **Ordinal** (e.g = good/ok/best)
  - **Nominal** (e.g = Yes/No/not sure)
- **Quantitative**
  - **Discrete** (e.g = no. of students in a class)
  - **Continuous** (e.g = height of a person)

- Columns are also called as Fields
- Rows " " " " Records

* Randomization can be used to deal with bias

Types of bias

- Sampling bias
- Observer bias
- Interpretation bias
- Confirmation bias

* Metadata is data about actual data

Databases

* Relational database = It consists of tables which are connected by a relationship.
  → To build a relationship b/w two tables there should one common attribute.

Key in database

* A Key is actually an columns (also called as attribute or field)

Type of Keys

* **Primary key**: A column in which every record is a unique value.
  → It cannot be null or blank
* **Foreign key**: A column in a table, that is a primary key in another table

• A table can has only 1 primary key but many foreign keys

• → Primary key is used access a unique record from table
• → Foreign key " " to create connection b/w two tables

A SQL query

SELECT * FROM customers WHERE city = "Jhang"

customers → Table of database
city → column of table customers

Organizing data

* **Naming Conventions**: Naming should describe content, date or version
* **Archive**: A collection of old files (Used to cut down clutter)
* **Foldering**: Putting files into a parent folder

*Relational database helps to avoid data duplication.
• **International Date Convention**: year - month - day

Spreadsheet Security

* **Access control features**: Password protection & User permission
  * Hidden tabs can be unhidden by anyone

Online Presence

★ LinkedIn ★ Github

END (course 3)

Course #4: Process

- **Sample**: A Representative of whole population.
- **Confidence level**: Probability that your sample size accurately reflects the whole population.
- **Margin of error**: It tells you how close sample results are to the results (if you use whole population).

★ As the confidence level decrease the sample size also decrease

★ Sample size $\propto \frac{1}{\text{Margin of error}}$

- To calculate sample size we need: • Population size, • Confidence level • Margin of error
- To calculate Margin of error we need: • Sample size, • population size • Confidence interval (level)

Data cleaning Tools

- **Conditional formatting**: > It changes the appearence of the data cell, based on certain condition
- **Duplicate removing**: > Dataset shouldn't include duplicates
- **Date Formatting**: > Dataset should have one single date formatting.
- **Split the text**: > It splits the text based on a character and creates new cells for each part of string

Functions for data cleaning

- Count if(): it returns no. of cells that match a value
  → basically tells no. of occurence

★ LEN() , ★ LEFT() , ★ RIGHT() , ★ MID()
★ CONCATENATE() , ★ TRIM()

Pivot Table

★ It is use to summarize and analyze data
  ★ It group values together.

e.g. you have data of employees and no. of items they have sold.
So columns are (Emp_name , Item, Quantity sold)

→ So now with pivot table, you can filter it through Item name and see which employee has sold which item in what quantity.

VLook UP()

★ It search for value based on certain condition.
e.g. = What is salary , where name is "ubaid"

Working with data from two or more databases

★ Always do data mapping (compare field names of both datasets)
★ Compatibility of two datasets means if they are able to worked out together

SQL Queries

★ SELECT , ★ INSERT
- To remove duplicates : ★ SELECT DISTINCT

★ LENGTH , ★ SUBSTR , ★ TRIM
★ CAST() , ★ CONCAT() , ★ COALESCE()

* **Verification**: It tells that data is accurate and reliable after got cleaned.
* **Change log**: list of all modifications of project

* CASE () => sql

Resume Building

* Add PAR Statements in your Resume.
  ↳ P = What was the Problem.
  A = What was your Action.
  R = What Results you achieved.

Course # 4 Ends

Course # 5 => Analyze Data

Analysis

* It consists of four phases:
  → Organize
  → Format and adjust = sort, filter etc.
  → Get input from others
  → Transform = Identify relationships and patterns

* Sort in SQL = Order By
* filter in SQL = Where

Sort in Spreadsheet

* Sort Sheet = Sort all colums
* Sort Range = Sort a specific column only by isolating it from others

* You can do sorting by multiple colums also

Data Validation

* You can preset the options for a colomns, so that not any other kind of data can be entered in that colomns.

* CONCAT() , * LIMIT , * GROUP BY

sql

A Usecase of substring

* let's say you got data and time in your data.
"2022-09-21 02:05:11"

→ first of all convert this into a text (string).
then split it using delimeter as space " ".
to seperate the date and time.

Data Aggregation

* Collecting many seperate pieces into a whole

• Value() = in spreadsheet, it is used to convert text to int.

Join in Database

* It is used to combine rows from two or more tables based on a related column.

* Types of join : ① Inner Join ⇒ returns matching values from both table
② Left Join
③ Right Join
④ Outer Join

• We use foreign key in joins

* SUB Queries
* Count and count distinct

~~Course #~~

* Count if() in spreadsheets
* Sum if() " "
* Sumproduct() " "

- Not equal to symbol in sql ⇒ != , <>
  ↳ Both can be used
- Group By in sql
- Order By in sql
* Temporary tables in sql

Course # 6 (Share your findings)

Data Visualizations

* Bar Graph
* Line Graph
* Pie Chart
* Maps
  * Your Data visuals shouldn't look confusing
* Histogram
* Time Series Chart
  - Correlation and Causation are two different things.

Types of Visualizations

① Static
↳ Tools = Spreadsheet

② Dynamic
↳ Tools - Tableu

Elements of Visualization

* Line
* Shape
* Color
* Space
* Movement.

* The goal of data analysis is to find out story of data and to tell this story to stakeholders
and
the best way to tell this story is a good presentation

[A Good Presentation]

* It should have a clear start, middle and end.
* There must be Titles and subtitles in every slide
* There shouldn't be more than 1 or 2 graphs in a slide
* There shouldn't be long text statements
* It must be easy and clear to understand, even to people who don't have much knowledge about analysis.
* There should be a consistent them followed in every slide of your presentation

Course #7 ⇒ R - Programming

* Introduction
* Tidyverse
* Tibbles
* Pipes
* ggplot2

Course #8 ⇒ Case Study

* Do a case study or a project.

THE END