LAB 03

Objective: This lab introduces you to three useful ideas

- Import data from webpages.
- Pivot Tables
- VLOOKUP function.

1. IMPORT DATA FROM WEBPAGE

Suppose we want to extract information about Alumni of IEOR into spreadsheet where details for alumni is given in the page: https://www.ieor.iitb.ac.in/people/alumni

Follow below steps to import data from a webpage:

- Open a new Google Sheet file. In the cell A1, use the following formula to extract specific data from the above webpage:

```
=IMPORTHTML("https://www.ieor.iitb.ac.in/people/alumni","table", 2)
```

- Function Help: **`=IMPORTHTML("XXXX", "TY",#)`**

  where, "XXXX" is the link for webpage, followed by "TY" representing type of data to be imported (use 'table' for table and 'list' for list) and # represent identity or sequence number. If there are multiple tables in webpage the "#" helps identify the table to import.

2. PIVOT TABLE

- A pivot table is a powerful tool that can be used to summarise and analyse data in a spreadsheet. It allows us to quickly and easily create interactive reports that can help you make better decisions.
- *How does a pivot table work?* A pivot table works by taking your data and rearranging it into a new table that is more meaningful to you. You can choose which columns to include in the pivot table and how to summarise the data. For e.g., you could create a pivot table showing total sales for each product category or the average monthly sales.
- *How to create a pivot table in Google Sheets?*

1. Open the spreadsheet that contains your data.
2. Select the cells that contain the data that you want to summarise.
3. Click on the Insert menu and select Pivot Table.
4. Pivot table needs to be created in a new sheet using the Pivot Table Editor. You can customize the pivot table by adding or removing columns, changing the summary functions, and filtering the data.

Here are some basic ideas for pivot tables:

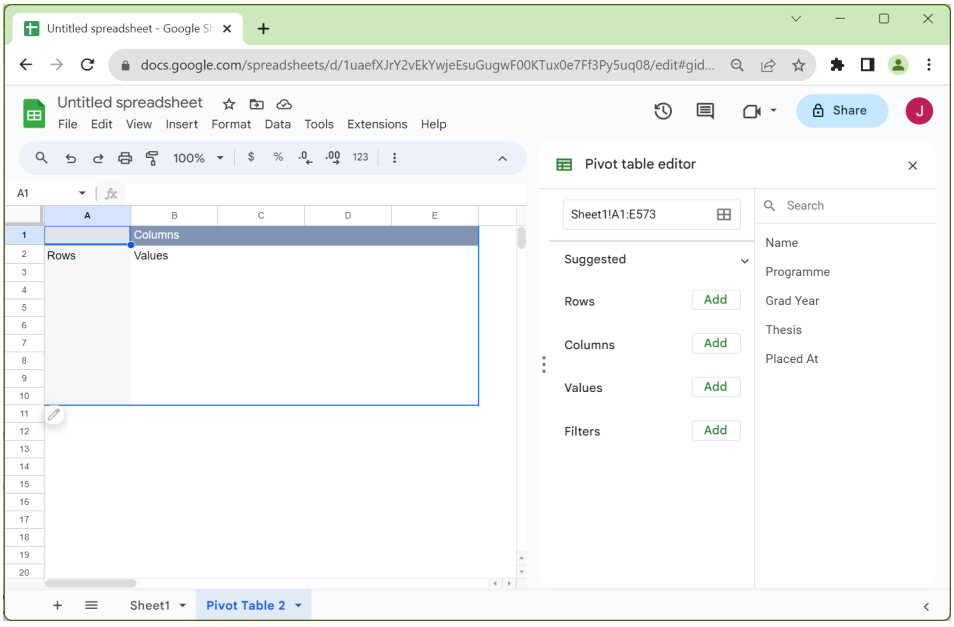
- You can use pivot tables to summarise data in a variety of ways, including counting, averaging, summing, and calculating percentages.
- You can use pivot tables to sort and filter data.
- You can use pivot tables to create calculated fields.
- You can use pivot tables to create interactive reports.

Here are some tips for using pivot tables:

- Make sure that your data is well-formatted before you create a pivot table. The columns should have headers, and the data should be consistent.
- Use the Pivot Table wizard to help you create your pivot table. The wizard will guide you through the process of selecting the data and choosing the summary functions.
- Experiment with different settings to see how they affect the pivot table. You can always change the settings later.
- Use pivot tables to create interactive reports. You can add filters and slicers to your pivot table to let users explore the data on their own.

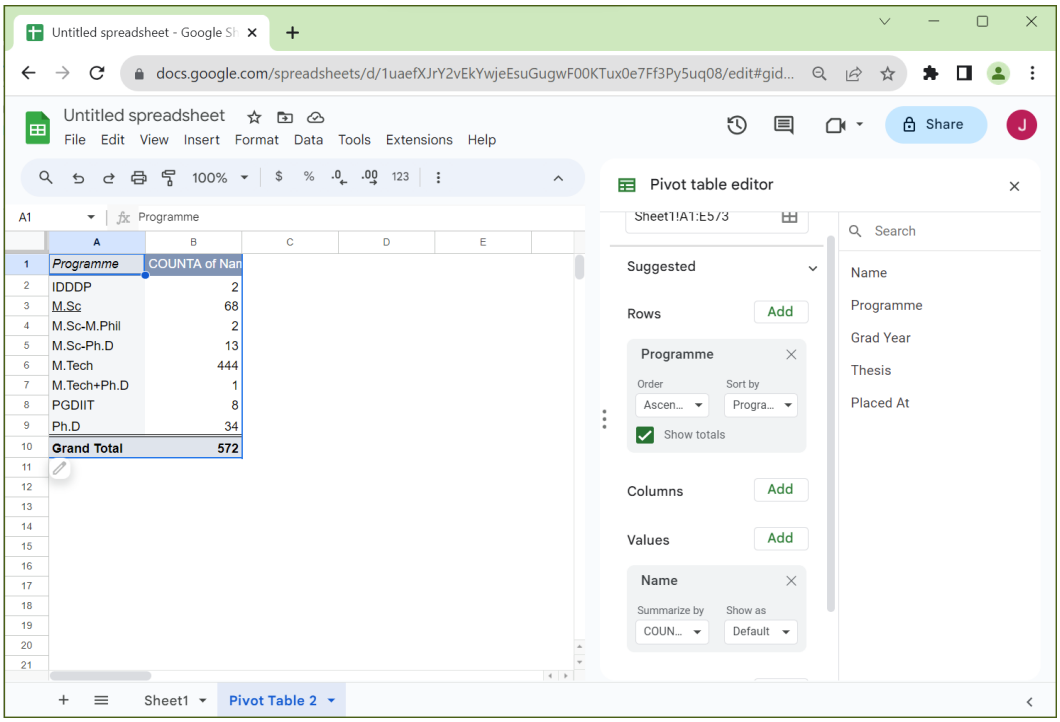
Follow below steps to create a pivot table:

- Use the same spreadsheet where you have imported the alumni data from IEOR website.
- Select cells A1 to E573
- Click Insert>>Pivot Table In Pop-up, select ‘New Sheet’. Click Create.
- In the new sheet, you should see a ‘Pivot Table Editor’ with two columns as shown in figure below. We need to use the Editor to create the summary tables we need.



Task A: Suppose we want to make a summary of the number of students who graduated with various degrees till now.

- Click the Add button next to ROWS and select ‘Programme’ (Alternately, you can drag the ‘Programme’ from the right-most panel to under ROWS.)
- Add ‘Name’ under VALUES
- The Pivot Table on the left side (spreadsheet should now be populated with the data).
- Observe that under VALUES, the Name is being Summarized using COUNTA function. Click it to see the various functions possible.



Task B1: Suppose we want to make a summary of year-wise number of students who graduated with various degrees till now.

- Add ‘Grad Year’ under ROWS (remove Programme)
- Add ‘Programme’ under COLUMNS
- Add ‘Name’ under VALUES [This should already be there. So, nothing to do]

This should give you a large table with year in the ROWS and Programme in the COLUMNS

Task B2: Suppose we want to display the graduation count from say, **year 2010 onwards**.

- Add ‘Programme’ under FILTER
- Click the drop-down below ‘GradYear’ under FILTER
    - Click ‘Filter by Condition’
        - Click ‘Greater Than or equal to’
        - Enter “2010” (without quotes)

- Click OK

This should give you a smaller table from year 2010 onwards

Task C (to do on your own): Keep 'Grad Year' and 'Programme' both under ROWS. Observe the Pivot Table

Task D (to do on your own):

Create a pivot table showing the number of years since 2005, the different degrees had graduating students.

For example, from 2005 to 2022 every year we had MTechs graduating. Hence, Table should show 18. However, PhDs were in only 15 years. The output you should get is shown.

Programme	Num Years
IDDDP	2
M.Sc	11
M.Sc-M.Phil	1
M.Sc-Ph.D	7
M.Tech	18
M.Tech+Ph.D	1
PGDIIT	4
Ph.D	15
Grand Total	**18**

3. VLOOKUP

The VLOOKUP function in a spreadsheet is a function that allows you to search for a value in a table and return the corresponding value from another column in the same row. Some practical example includes suppose you want to know customer information by name, find price of product using product ID, create a product catalog, etc. The VLOOKUP function has four arguments:

- Search_key: The value that you want to search for.
- Range: The range of cells that contains the data that you want to search.
- Index: The column number that contains the value that you want to return.
- Is_sorted: Indicates whether the range is sorted in ascending or descending order.

How VLOOKUP works?

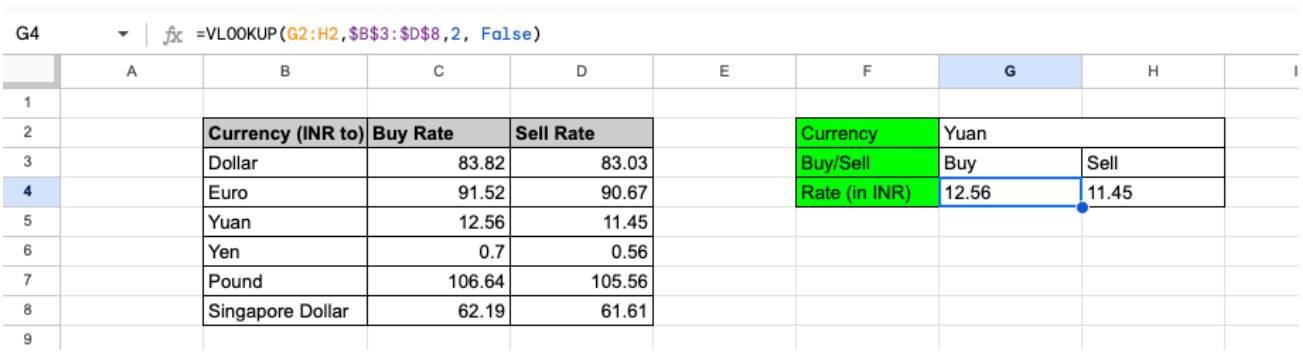
The VLOOKUP function works by first searching for the search_key value in the first column of the range. If the search_key value is found, the VLOOKUP function returns the value from the column specified by the index argument. If the search_key value is not found, the VLOOKUP function returns the #N/A error.

Follow below steps to use VLOOKUP function in spreadsheet

- Consider the table for exchange rate given below for various buy rate and sell rate in Indian Rupee (INR) from B2:D8. Use following sheet CLICK HERE
- Lets say, we want to know buy rate of Chinease Yuan using INR. Figure below describes cell formula to be used for buy rate.

```
=VLOOKUP(G2:H2, $B$3:$D$8, 2, False)
```

- First argument in VLOOKUP() is G2:H2 representing *search_key* used to as specific value you are looking for in dataset. $B$3:$D$8 is *range* used search pre-defined search_key. Third argument is *index* used which return value corresponding to *search_key* in *range.* Fourth argument is an optional used to specify sorted nature of *range.* True as default assume range is sorted and vice-versa.
- G4 is buy rate of Yuan in INR. Similarly, you can drage and drop G4 to H4 for sell rate as the range is already prefixed using $ sign in formula. But now, third argument need to be changed from 2 to 3 as index 3 represent sell rate.

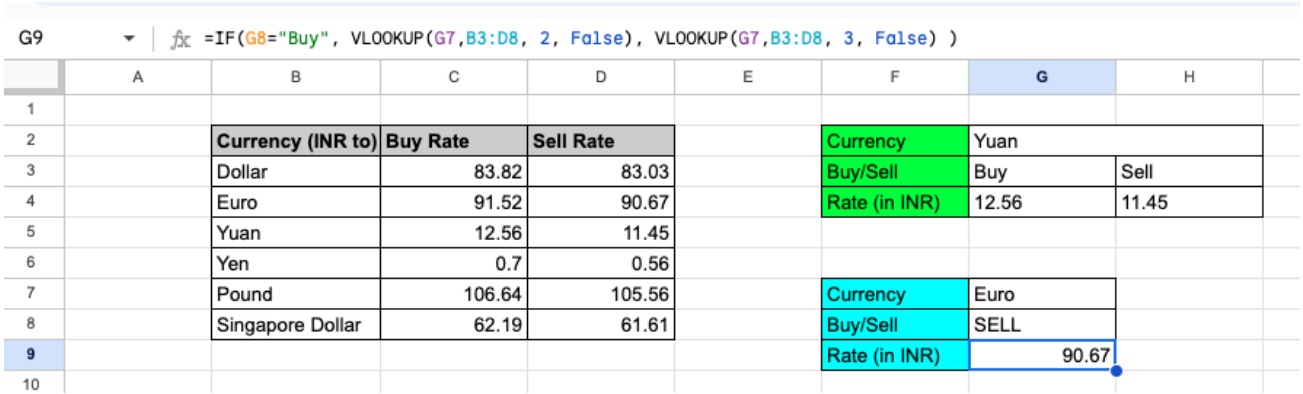


Currency (INR to)	Buy Rate	Sell Rate
Dollar	83.82	83.03
Euro	91.52	90.67
Yuan	12.56	11.45
Yen	0.7	0.56
Pound	106.64	105.56
Singapore Dollar	62.19	61.61

Currency	Yuan	
Buy/Sell	Buy	Sell
Rate (in INR)	12.56	11.45

VLOOKUP with conditional statements

Try out the following where VLOOKUP is defined itself in conditional statement, which takes two arguments in G7 (Currency) and G8 (To be bought or sold) cell.



Currency (INR to)	Buy Rate	Sell Rate
Dollar	83.82	83.03
Euro	91.52	90.67
Yuan	12.56	11.45
Yen	0.7	0.56
Pound	106.64	105.56
Singapore Dollar	62.19	61.61

Currency	Yuan	
Buy/Sell	Buy	Sell
Rate (in INR)	12.56	11.45

Currency	Euro
Buy/Sell	SELL
Rate (in INR)	90.67

Exercises: The exercises below are designed to give you practice in using spreadsheets for basic data summary, and data visualization.

NETFLIX. The wikipedia page has details about the 2023 Feature films, those released and those that are yet to release in 2023. https://en.wikipedia.org/wiki/List_of_Netflix_original_films_(since_2023) Do the following.

1. Import the list of Feature Films into Google spreadsheet (ignore Documentaries and Specials)

Make the following Summary Tables/ Charts. Use as many different tabs as needed, add new columns, insert charts etc. But ensure your spreadsheet is neat, with properly labeled tabs (sheets), with headings and comments, where needed. ***Hint:*** *Make pivot tables and use them to make charts, as needed.*

2. Pivot chart of the language-wise and genre-wise number of films released or awaiting release in 2023.
3. Bar chart of the number of films released or awaiting release in each month. *Hint: You may want to extract the MONTH in a separate column first., and the create a pivot table*.
4. Stacked bar chart of language-wise number of films released or awaiting release in each month.
5. Histogram of the run-time of the feature films. ***HINT****, you may need to compute the runtime in a separate column first.*
6. Which are the top 3 most popular genres?

SOLAR PROJECT DATA: In 2017-2020, we coordinated a national level solar lamp distribution project. Now, some (modified) data based on that actual project is shared in the file below (you have view access; feel free to make a copy of the same for lab use)
https://docs.google.com/spreadsheets/d/13o8q6VlenkbFuWxUVx7na8lkcZ525GHM_syn9KdoDUA/edit?usp=sharing
The file has 73000+ rows showing the employee-wise and location-wise number of lamps sold. The data set has the following columns

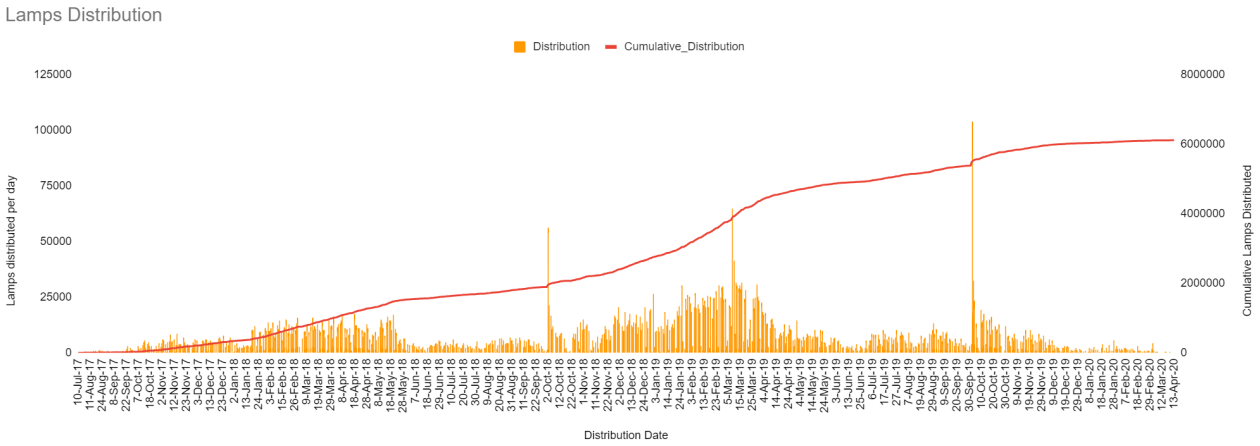
- *Sno* and *Sno_T2* columns: Are just serial numbers.
- *DistributionDate* shows the exact date the lamps were distributed. In YYY-MM-DD format.
- *EmployeeCode* is the code of the employee
- *BlockCode* refers the demand location

- *LampIssued* refers to the quantity given to the employee on a day (usually sold the same day on in future). *UnsoldLamps* refers to the quantity the employee doesnt sell same day. Could be returned and reissued later, or kept with the employee of sale in future. These two columns is quite 'noisy'
- *NoofBeneficiaries* indicated the quantity actually sold
- *ModelNo* indicate the lamp model. There were two models, 1 and 2.
- *ADCode* indicates the local distributor from where the lamps are sold to the market. AD stands for Assembly and Distribution Center.
- *Phase* indicates the different phases of the project, T1, T2 and T3.

Now, using the data, do the following:

1. How many unique Employees were employed? Unique Blocks? Unique AD centers?
2. Make a summary table of AD center wise distribution quantities.
3. Make a summary table of the number of markets (blocks) associated with each AD center.
4. Create a day-wise distribution bar chart, showing the quantity sold per day, from the first day to the last day of sales. Ensure that the days are properly arranged.
5. Create a cumulative distribution line chart, showing the cumulative quantity sold until that day, from the first day to the last day of sales. Ensure that the days are properly arranged.

If 4 and 5 are correctly done, then you should get a chart like this:



6. For each block, compute the total number of distribution days. A particular date is counted as a distribution day iff on that date at least 1 unit was sold.
7. For each block, list the first date of distribution and last date of distribution; and compute the difference between these two dates. Compare this with the results you get in 6.