Python Foundations - Project Support

Agenda

○ Python Programming & EDA Quiz

○ Getting Started with Data Analysis

○ Common Statistical Measures

○ Significance of Data Visualization

○ Choosing plots for Univariate/Bivariate Analysis

○ FoodHub - Business Context & objective

○ FoodHub - Full-Code/Low-Code Version

○ FoodHub - Q/A

Let’s begin the discussion by answering a few questions on Python programming and Exploratory Data Analysis

Python Programming & EDA Quiz

Which of the following represents the correct sequence of steps to begin data analysis?

- **A** Import Libraries => EDA => Load dataset
- **B** Load dataset => Import Libraries => EDA
- **D** Import Libraries => Load dataset => EDA

Python Programming & EDA Quiz

Which of the following represents the correct sequence of steps to begin data analysis?

A Import Libraries => EDA => Load dataset

B Load dataset => Import Libraries => EDA

D Import Libraries => Load dataset => EDA

Getting Started with Data Analysis

1 Importing Packages

In this step, we import all the necessary packages such as **numpy**, **pandas**, **matplotlib**, **seaborn** etc.

2 Loading the Dataset

Using **pandas** functions, we load the dataset in a dataframe. For csv files, '**pd.read_csv( )**' is used. For excel files, 'pd.read_excel( )' is used.

3 Exploratory Data Analysis

In this step, we look for the **shape** of the dataset, the different **data types,** check for **anomalous and missing values,** and analyse the attributes individually as well as relationships between them through visualizations to identify key business insights

Python Programming & EDA Quiz

Consider the file foodhub_order.csv stored in the following folder hierarchy
Python => Project => Dataset
Which of the following code snippets is the correct way to load the file into a pandas dataframe in Google Colab?

A

```
df = pd.read_csv("Python/Project/Dataset/foodhub_order.csv")
```

B

```
df = pd.read_csv("Python\Project\Dataset\foodhub_order.csv")
```

C

```
df = pd.read_csv("Python//Project//Dataset//foodhub_order.csv")
```

D

```
df = pd.read_csv("Python\\Project\\Dataset\\foodhub_order.csv")
```

Python Programming & EDA Quiz

Consider the file foodhub_order.csv stored in the following folder hierarchy
Python => Project => Dataset
Which of the following code snippets is the correct way to load the file into a pandas dataframe in Google Colab?

A `df = pd.read_csv("Python/Project/Dataset/foodhub_order.csv")`

B `df = pd.read_csv("Python\Project\Dataset\foodhub_order.csv")`

C `df = pd.read_csv("Python//Project//Dataset//foodhub_order.csv")`

D `df = pd.read_csv("Python\\Project\\Dataset\\foodhub_order.csv")`

Loading Datasets in Pandas

read_csv - pandas function used to load datasets in CSV format into a pandas dataframe

Syntax: `df = pd.read_csv(“file_path/file_name.csv”)`

Pandas has to be imported with alias pd - `import pandas as pd`

The file name has to be enclosed in quotation marks (single or double)

Above syntax works when the file (dataset) is in the same working directory as the Python notebook

When the file (dataset) and the Python notebook are not in the same working directory, the path to the file has to be specified

Python Programming & EDA Quiz

Which of the following measures condense the dataset down to one representative central value?

A Mean, Median, Mode

B Standard Deviation, Variance, Range

C Correlation Coefficient

D Maximum, Median, Minimum

Python Programming & EDA Quiz

Great Learning
POWER AHEAD

Which of the following measures condense the dataset down to one representative central value?

A [✓ Mean, Median, Mode]

B Standard Deviation, Variance, Range

C Correlation Coefficient

D Maximum, Median, Minimum

Common Statistical Measures

Central tendency measures condense the dataset down to one representative central value

Allows us to compare one dataset to another

Mean	Median	Mode
The **mean** is the arithmetic average of a set of given numbers. df['column_name'].mean( )	The **median** is the middle score in a set of given numbers. df['column_name'].median( )	The **mode** is the most frequent score in a set of given numbers. df['column_name'].mode( )[0]
The **mean** can be **used to represent the typical value** and therefore serves as a yardstick for all observations.	Since the mean is highly affected by the outliers, the **median** is a better choice for a dataset with extreme values	Mode is the preferred measure when data is categorical.

Python Programming & EDA Quiz

Consider a dataframe df with two attributes "height" and "weight". Which of the following methods can be used to check the correlation between these two variables?

A `df.corr()`

B `sns.heatmap(df)`

C `sns.histplot(data=df, x='height')`

D `plt.scatter(df['height'], df['weight'])`

Python Programming & EDA Quiz

Great Learning
POWER AHEAD

Consider a dataframe df with two attributes "height" and "weight". Which of the following methods can be used to check the correlation between these two variables?

A [✓ `df.corr()`]

B [✓ `sns.heatmap(df)`]

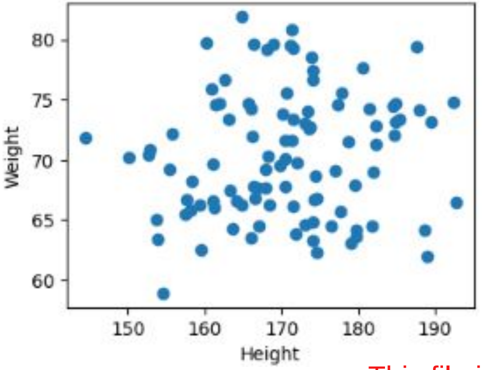
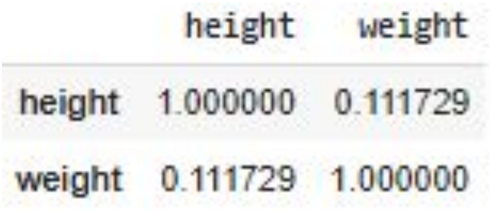
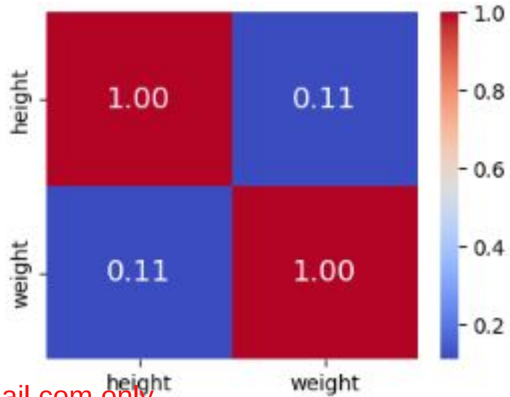
C `sns.histplot(data=df, x='height')`

D [✓ `plt.scatter(df['height'], df['weight'])`]

Common Statistical Measures

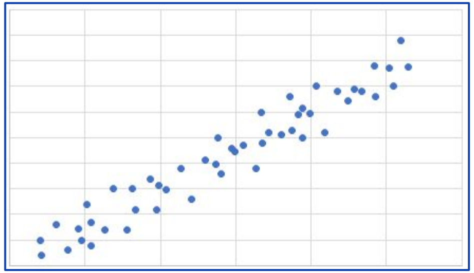
Correlation is a **measure of association** between two variables

Correlation coefficient is a statistical measure of the **strength of the linear relationship** between two variables.

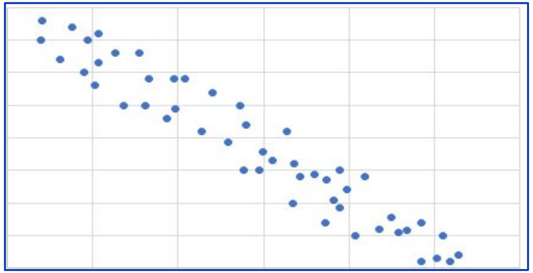
`plt.scatter(df['height'], df['weight'])`	`df.corr()`	`sns.heatmap(df)`
	 height weight height 1.000000 0.111729 weight 0.111729 1.000000	

Common Statistical Measures

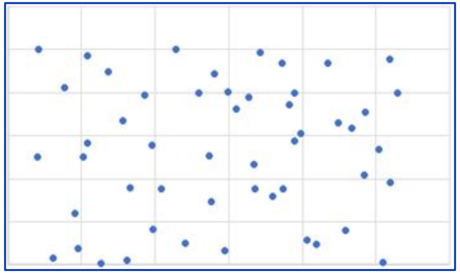
Based on direction of change in the value of one variable as the value of the other changes, the two variables are said to have a positive relationship, negative relationship, or no relationship at all.



+ve correlation



-ve correlation



No correlation

Python Programming & EDA Quiz

Consider a dataframe df containing information about CustomerID, Region, Purchase Amount. which of the following statements is true?

A `df.info()` provides information about data types of columns

B `df.value_counts('CustomerID')` returns a single number representing the total count of the values in the 'CustomerID' column

C `df.describe()` returns the counts, mean, standard deviation, min, max, and quartiles of numeric columns

D `df.groupby('Region')[‘Purchase Amount’].sum()` provides the total sum of amount of purchase by different regions

Python Programming & EDA Quiz

Consider a dataframe df containing information about CustomerID, Region, Purchase Amount. which of the following statements is true?

A. `df.info()` provides information about data types of columns

B. `df.value_counts('CustomerID')` returns a single number representing the total count of the values in the 'CustomerID' column

C. `df.describe()` returns the counts, mean, standard deviation, min, max, and quartiles of numeric columns

D. `df.groupby('Region')[‘Purchase Amount’].sum()` provides the total sum of amount of purchase by different regions

Pandas

One of the most commonly used Python libraries for data manipulation and analysis

df.head( )

The **df.head( )** function returns the **first 5 rows** of the dataframe

df.shape

The **df.shape** returns the number of **rows** and **columns** of the dataframe

df.astype( )

The **df.astype( )** function **convert the data type** of an existing column in a dataframe

df.info( )

The **df.info()** function returns information about the dataframe including the **data types** of each column and **memory usage**

df.describe( )

The **df.describe()** function returns the statistical info like percentile, mean, standard deviation, etc. of the dataframe

df.unique( )

The **df.unique()** function returns the unique values present in a dataframe

df.groupby( )

The **df.groupby( )** function function is used to **split** the data **into groups**

df.value_counts( )

The **df.value_counts( )** returns a Series containing the **counts of unique values**.

Python Programming & EDA Quiz

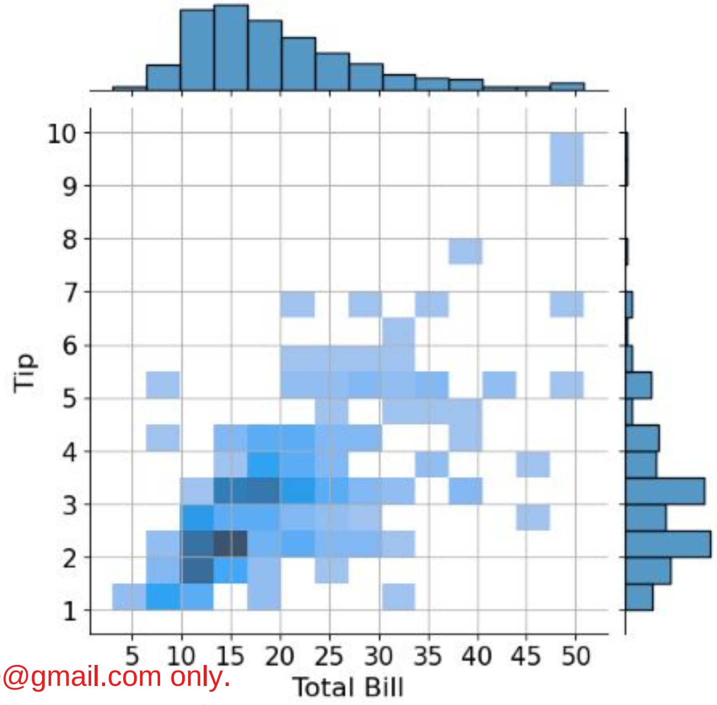
According to the jointplot below, where is the highest density of data points?

A Total bill ~(15 – 20) & Tip ~(3 - 4)

B Total bill ~(5 - 10) & Tip ~(1.5 - 2.5)

C Total bill ~(10 - 20) & Tip ~(1.5 - 2.5)

D Total bill ~(25 - 35) & Tip ~(3 - 4)



Python Programming & EDA Quiz

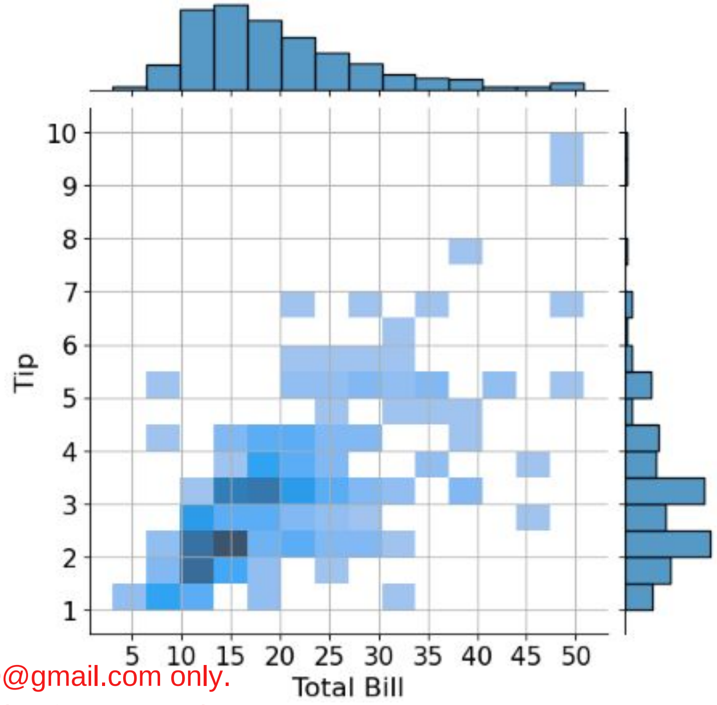
According to the jointplot below, where is the highest density of data points?

A Total bill ~(15 – 20) & Tip ~(3 - 4)

B Total bill ~(5 - 10) & Tip ~(1.5 - 2.5)

C Total bill ~(10 - 20) & Tip ~(1.5 - 2.5)

D Total bill ~(25 - 35) & Tip ~(3 - 4)



Significance of Data Visualization

Gives us a better idea of the information stored in data by giving it visual context through various plots

Allows us to **visualize large volumes of data** in an understandable and coherent way

Also enables us to **identify relationships and patterns within data**

Helps us comprehend the information and **draw conclusions and insights**

Enables data storytelling to easily create a narrative through graphics and diagrams

Python Programming & EDA Quiz

Which of the following the combination of plot and type of data is generally used for univariate analysis?

A Boxplot - Numerical Data

B Histogram - Numerical data

C Lineplot - Categorical Data

D Countplot - Categorical Data

Python Programming & EDA Quiz

Which of the following the combination of plot and type of data is generally used for univariate analysis?

A Boxplot - Numerical Data

B Histogram - Numerical data

C Lineplot - Categorical Data

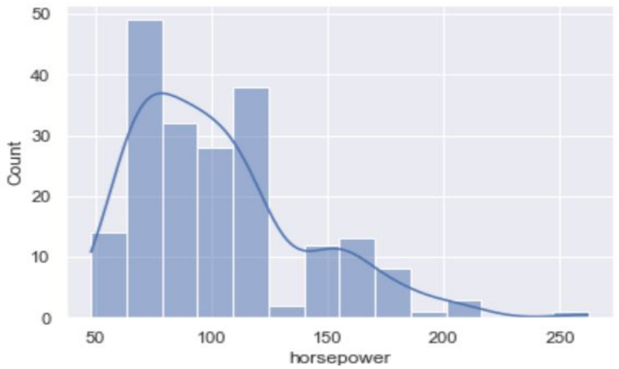
D Countplot - Categorical Data

Choosing plots for Univariate Analysis

When to use a Histogram

When **the data is numeric** and **you want to see the shape of the data distribution**, determine whether the data is distributed approximately normally (bell shaped) or not

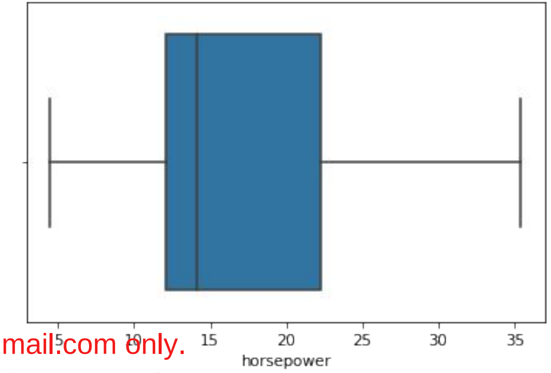
```
sns.histplot( data =  , x = ‘ ’, kde = True )
```



When to use Boxplot

When **the data is numeric** and **you want to understand the centre, spread, and presence of outliers**

```
sns.boxplot( data =  , x = ‘ ’)
```

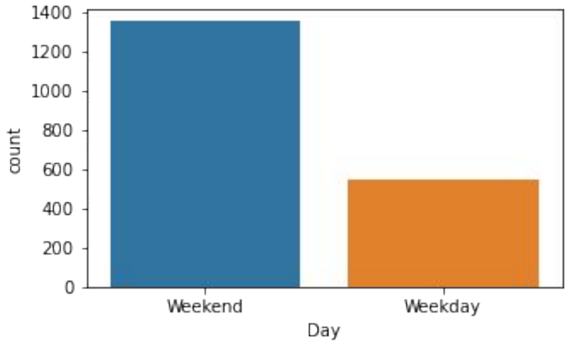


Choosing plots for Univariate Analysis

When to use a Count plot

When **the data is categorical** and you want to show the **counts of observations in each categorical** bin

```
sns.countplot( data =  , x = ‘ ’)
```

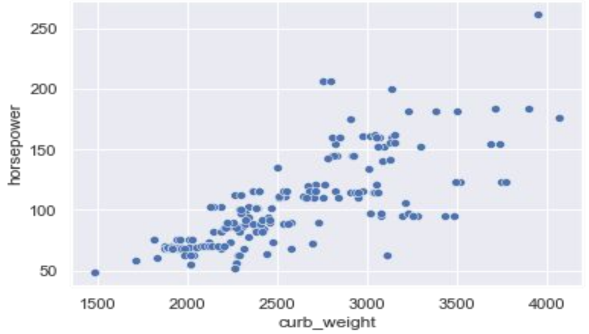


Choosing plots for Bivariate Analysis

When to use a scatter plot

When **the data is numeric** and **you want to** determine whether the two variables are related, and see if it's a positive or negative correlation.

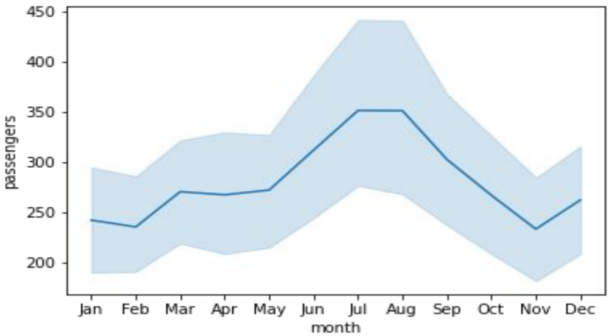
```
sns.scatterplot( data =  , x = ‘ ’, y = ‘ ’ )
```



When to use a line chart

When **the data is continuous** and **you want to see the** how the value of something changes over short and long periods of time.

```
sns.lineplot( data =  , x = ‘ ’, y = ‘ ’ )
```



FoodHub - Business Context and Objective

The number of restaurants in New York is increasing day by day. Lots of students and busy professionals rely on those restaurants due to their hectic lifestyles. Online food delivery service is a great option for them. It provides them with good food from their favorite restaurants. A food aggregator company FoodHub offers access to multiple restaurants through a single smartphone app.

The app allows the restaurants to receive a direct online order from a customer. The app assigns a delivery person from the company to pick up the order after it is confirmed by the restaurant. The delivery person then uses the map to reach the restaurant and waits for the food package.

FoodHub - Business Context and Objective

Once the food package is handed over to the delivery person, he/she confirms the pick-up in the app and travels to the customer's location to deliver the food. The delivery person confirms the drop-off in the app after delivering the food package to the customer. The customer can rate the order in the app. The food aggregator earns money by collecting a fixed margin of the delivery order from the restaurants.

The food aggregator company has stored the data of the different orders made by the registered customers in their online portal. They want to analyze the data to get a fair idea about the demand of different restaurants which will help them in enhancing their customer experience. Suppose you are hired as a Data Scientist in this company and the Data Science team has shared some of the key questions that need to be answered. Perform the data analysis to find answers to these questions that will help the company to improve the business.

FoodHub - Data Dictionary

Data	Data Description
order_id	Unique ID of the order
customer_id	ID of the customer who ordered the food
restaurant_name	Name of the restaurant
cuisine_type	Cuisine ordered by the customer
cost	Cost of the order
day_of_the_week	Indicates whether the order is placed on a weekday or weekend (The weekday is from Monday to Friday and the weekend is Saturday and Sunday)

FoodHub - Data Dictionary

Data	Data Description
rating	Rating given by the customer out of 5
food_preparation_time	Time (in minutes) taken by the restaurant to prepare the food. This is calculated by taking the difference between the timestamps of the restaurant's order confirmation and the delivery person's pick-up confirmation.
delivery_time	Time (in minutes) taken by the delivery person to deliver the food package. This is calculated by taking the difference between the timestamps of the delivery person's pick-up confirmation and drop-off information

How to Load Dataset in Google Colab

Step 1: Upload the csv file in the Google Drive

Step 2: Create a new notebook / open an existing notebook

Step 3: Import pandas library into the notebook. The following code can be used for the same

```
import pandas as pd
```

Step 4: Mount Google Drive in the notebook. This can be done via two approaches:

How to Load Dataset in Google Colab

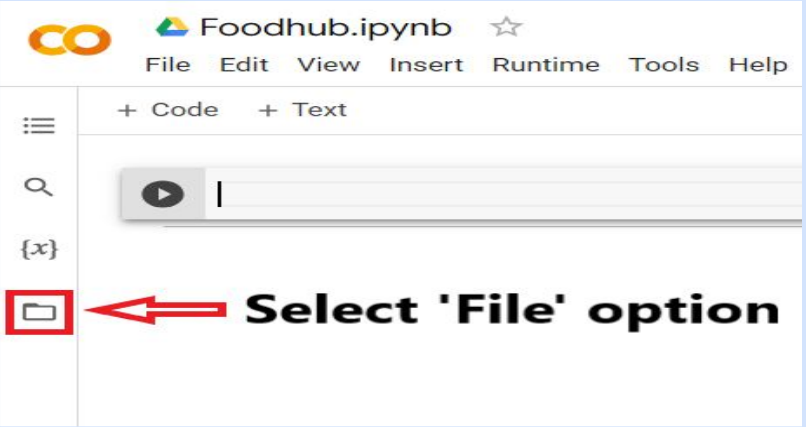
Approach 1

Step i: Click on the *Files* option on the left

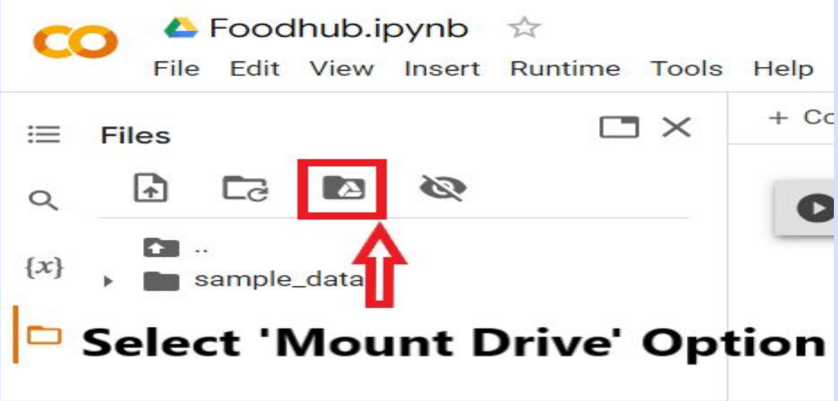
Step ii: Select the *Mount Drive* option

Step iii: In the pop-up that appears, select *Connect to Google Drive* option

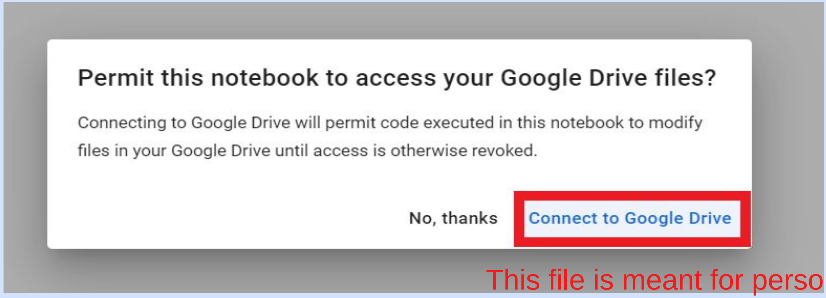
How to Load Dataset in Google Colab



Select 'File' option



Select 'Mount Drive' Option



Permit this notebook to access your Google Drive files?

Connecting to Google Drive will permit code executed in this notebook to modify files in your Google Drive until access is otherwise revoked.

No, thanks | Connect to Google Drive

How to Load Dataset in Google Colab

Approach 2

Step i: Run the following command in the notebook

```
from google.colab import drive
drive.mount('/content/drive')
```

Step ii: In the pop-up that appears, select *Connect to Google Drive* option

How to Load Dataset in Google Colab

Step 5: Expand the *Drive* option, and browse to your working directory

Step 6: Right-click on the file and select *Copy path*

For example, if we want to load the file foodhub.csv, which is present in the *Colab Notebooks* folder in *MyDrive*, we would navigate to the folder and right-click on the file to get the file path

How to Load Dataset in Google Colab

Step 7: Create a variable *path* and set the copied file path as the value of the variable (you can simply paste the copied file path for this)

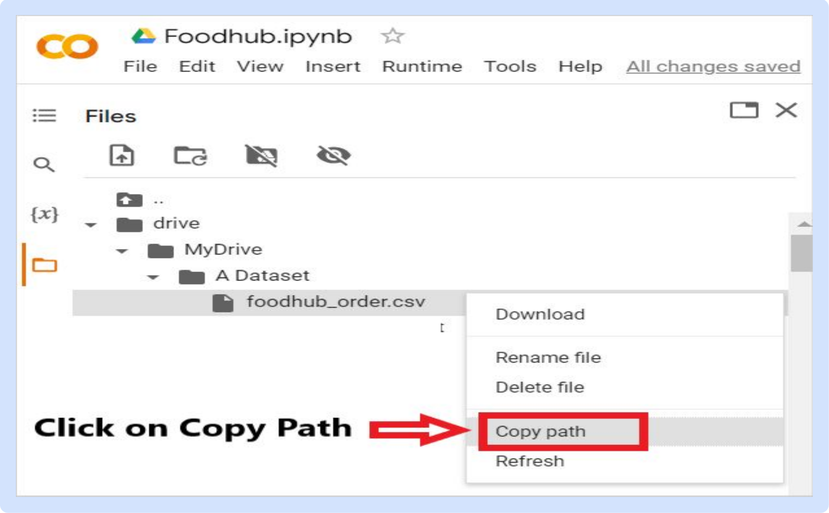
Step 8: Pass the *path* variable as an argument of the pandas *read_csv()* function to load the file into a pandas dataframe and store it in a variable

For example: `df = pd.read_csv(path)`

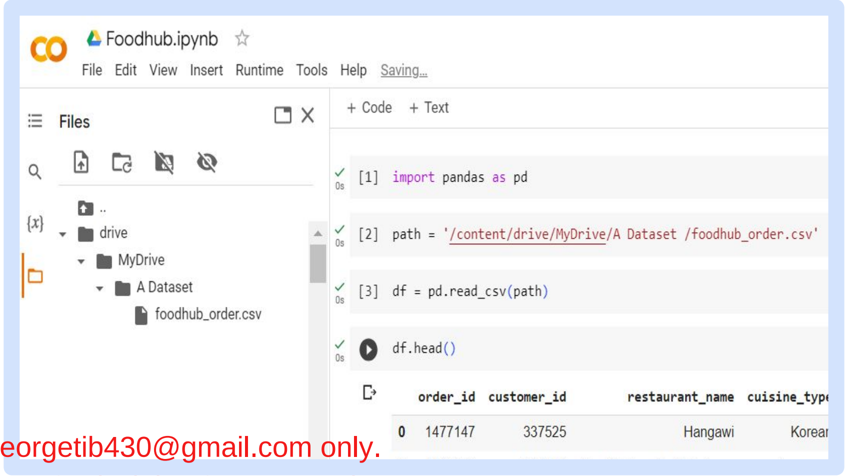
Step 9: Call the head() function of the dataframe to check if the data is imported correctly

For example: `df.head()`

How to Load Dataset in Google Colab



Click on Copy Path



```
[1] import pandas as pd

[2] path = '/content/drive/MyDrive/A Dataset /foodhub_order.csv'

[3] df = pd.read_csv(path)

    df.head()
```

```
   order_id  customer_id      restaurant_name  cuisine_typ
0  1477147   337525           Hangawi          Korear
```

How to Load Dataset in Jupyter Notebook

Step 1: Download the CSV file you want to work with

Step 2: Locate the file in the Local Drive

Step 3: Right-click on the file and click on Properties and copy the file location

Step 4: Import numpy and pandas

How to Load Dataset in Jupyter Notebook

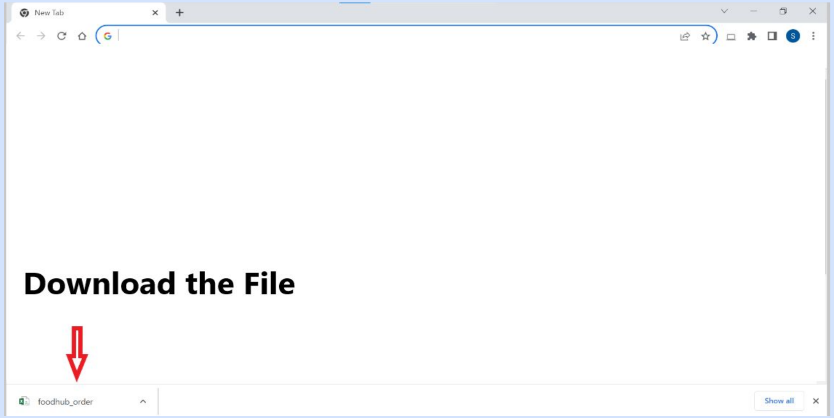
Step 5: Paste the path in the variable path and add the filename at the end, as shown below

It is important to replace the single backslash (i.e., \) in the file path with a double backslash (i.e., \\), a single forward slash (i.e., /), or a double forward slash (i.e., //).

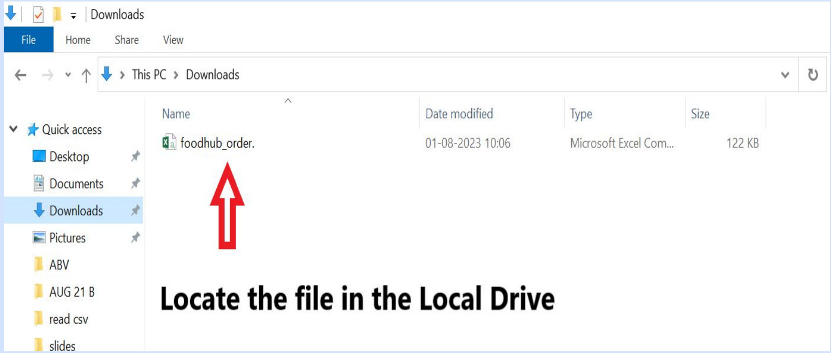
For example: if the filename is Foodhub.csv and the file path is `C:\Users\User\Downloads` , then the path variable should be defined as one of the following:

```
path = 'C:\\Users\\User\\Downloads\\Foodhub.csv'

path = 'C:/Users/User/Downloads/Foodhub.csv'

path = 'C://Users//User//Downloads//Foodhub.csv'
```

How to Load Dataset in Jupyter Notebook



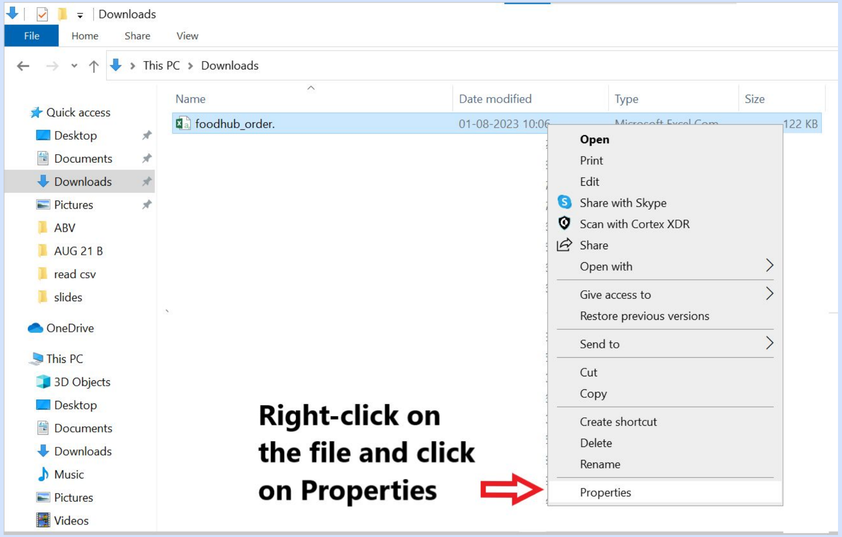
Download the File



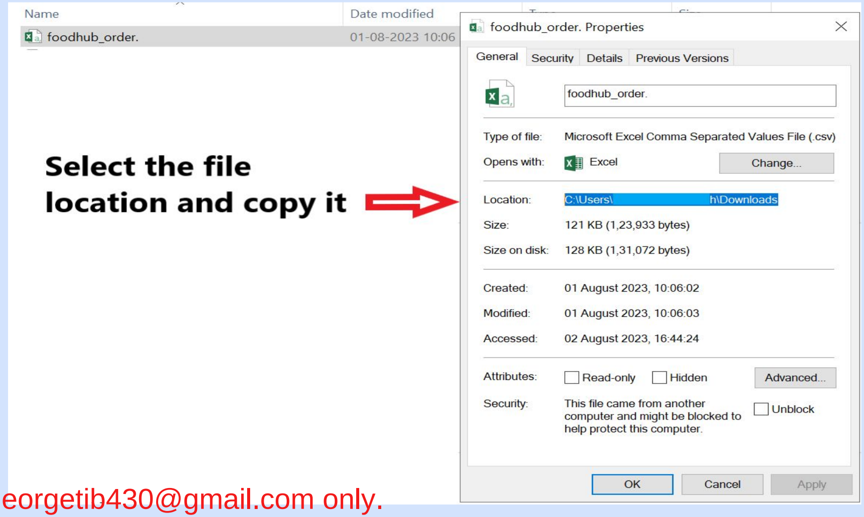
Name	Date modified	Type	Size
foodhub_order.	01-08-2023 10:06	Microsoft Excel Com...	122 KB

Locate the file in the Local Drive

How to Load Dataset in Jupyter Notebook



Right-click on the file and click on Properties



foodhub_order.

Type of file: Microsoft Excel Comma Separated Values File (.csv)

Opens with: Excel

Location: C:\Users\[redacted]h\Downloads

Size: 121 KB (1,23,933 bytes)

Size on disk: 128 KB (1,31,072 bytes)

Created: 01 August 2023, 10:06:02

Modified: 01 August 2023, 10:06:03

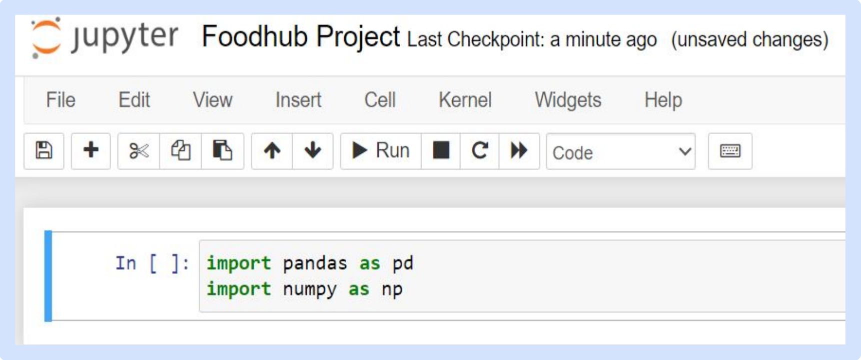
Accessed: 02 August 2023, 16:44:24

Attributes: □ Read-only □ Hidden

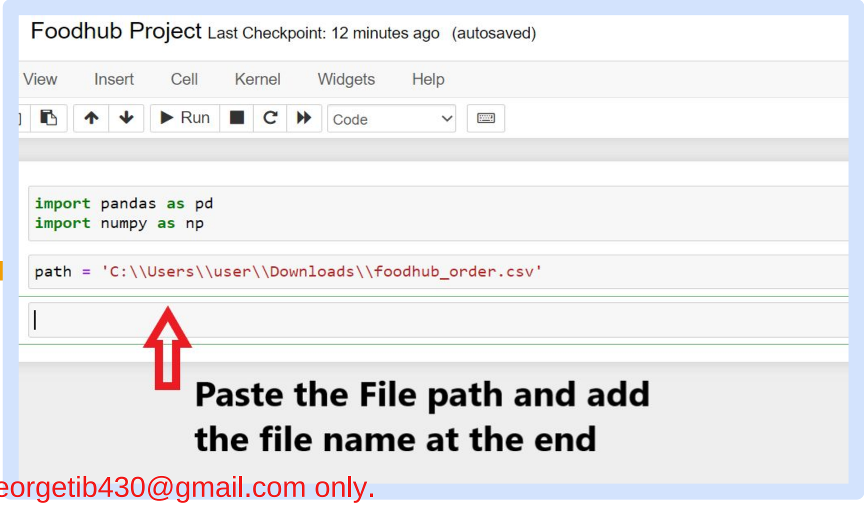
Security: This file came from another computer and might be blocked to help protect this computer. □ Unblock

Select the file location and copy it

How to Load Dataset in Jupyter Notebook


jupyter
Foodhub Project Last Checkpoint: a minute ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Run
Code

```
In [ ]: import pandas as pd
        import numpy as np
```


Foodhub Project Last Checkpoint: 12 minutes ago (autosaved)
View Insert Cell Kernel Widgets Help
Run
Code

```
import pandas as pd
import numpy as np
```

```
path = 'C:\\Users\\user\\Downloads\\foodhub_order.csv'
```

Paste the File path and add the file name at the end

How to Load Dataset in Jupyter Notebook

Step 6: Call the path variable in the read_csv() function of pandas to load the file into a pandas dataframe, and store it in a variable

For example: **`df = pd.read_csv(path)`**

Step 7: Call the head() function of the dataframe to check if the data is imported correctly

For example: `df.head()`

FoodHub - Low-code Version

For learners who aspire to be in managerial roles in the future, focusing on solution review, interpretation, recommendations, and communication with business stakeholders

Download the dataset and the *Learner Notebook - Low Code* (this is a template notebook)

Fill in the blanks in the notebook to complete and execute the code to solve the questions and perform all the tasks as per the grading rubric

Once the notebook is completely executed and necessary outputs obtained, a business presentation (using Microsoft PowerPoint, Google Slides, etc.) has to be created

FoodHub - Low-code Version

The presentation should contain observations, insights, and recommendations for the business problem

The presentation template provided can be referred to as a sample

Once the presentation is complete, convert the presentation to .pdf format

The presentation should be submitted as a PDF file (.pdf) and NOT as a .pptx file

Please make sure that all the sections mentioned in the grading rubric have been covered in the submission

FoodHub - Full-code Version

For learners who aspire to be in hands-on coding roles in the future, focusing on building solution codes from scratch

Download the dataset and the *Learner Notebook - Full Code* (this is a template notebook containing high-level steps to perform and insight-based questions)

Write necessary code to solve the questions and perform all the tasks as per the grading rubric

Clearly write down observations, insights, and recommendations for the business problem based on the analysis performed

Once the notebook is complete, download it as a **.ipynb** file and convert it to a **.html** file

FoodHub - Full-code Version

The notebook should be submitted as an HTML file (.html) and NOT as a notebook file (.ipynb)

The conversion can be done via one of the following ways:

Jupyter Notebook:
Google Colab: Use free online tools

Please make sure that all the sections mentioned in the grading rubric have been covered in the submission

FoodHub - Q/A

FoodHub - Project FAQs

How to approach Question 13?

The company wants to provide a promotional offer in the advertisement of the restaurants. The condition to get the offer is that the restaurants must have a rating count of more than 50 and the average rating should be greater than 4. Find the restaurants fulfilling the criteria to get the promotional offer.

Step 1: Filter the restaurant's column for those restaurants that do not have a rating as 'Not given'

Step 2: Convert the rating column created above from object to integer datatype

FoodHub - Project FAQs

Step 3: Create a dataframe that contains the restaurant names with their rating counts

Step 4: Get all the restaurant names that have a rating count of more than 50

Step 5: find the mean rating of the restaurants by using the group by function

FoodHub - Project FAQs

How to approach Question 14?

The company charges the restaurant 25% on the orders having cost greater than 20 dollars and 15% on the orders having cost greater than 5 dollars. Find the net revenue generated by the company across all orders.

Step 1: Create a function with conditional statements (for each category, 25%, 15% and 0%(else condition) and mention the revenue for each condition.

Step 2: Apply these conditions on the cost_of_the_order column to calculate the revenue, same the value in a revenue column.

Step 3: Taking summation of the revenue column will give the total revenue.

FoodHub - Project FAQs

Is there a way to transfer the graphs from Google Colab to the presentation without it looking blurry?

There are multiple ways to transfer the graphs :

1. Use the following line of code just after the visualization code:

```
plt.savefig("output.jpg", bbox_inches='tight')
```

For example:

```
sns.histplot(data=data, x='column')

plt.savefig("output.jpg", bbox_inches='tight')
```

FoodHub - Project FAQs

2. Use the snipping tool to snip the visual plot from the Jupyter notebook and paste the snip in ppts.

3. Right-click on the image and click on copy and paste the copied plot in the ppt or document.

Happy Learning !