Analysis of Purchase Data

An Internship Report

Submitted by

Rutvik Bharat Pradhan

210500107039

In partial fulfilment for the award of the degree of

BACHELOR OF ENGINEERING

in

Computer Engineering

Sigma Institute of Engineering

Gujarat Technological University, Ahmedabad

April, 2024-25

Sigma Institute of Engineering

Ajwa-Nimeta Road, Bakrol, Vadodara-390019

CERTIFICATE

This is to certify that the internship report submitted titled "**Analysis Of Purchase Data**" has been carried out by **Rutvik Bharat Pradhan** under my guidance in partial fulfilment for the degree of Bachelor of Engineering in Computer Engineering, 7th Semester of Gujarat Technological University, Ahmedabad during the academic year 2024-25.

Dr. Sheshang Degadwala	**Dr. Sheshang Degadwala**
Internal Guide	**Head of the Department**

TO WHOM IT MAY CONCERN

This is to certify that Rutvik Bharat Pradhan, a student of Sigma Institute of Engineering has successfully completed his internship in the field of Data Analysis from 28th June 2024 to 12th July 2024 (Total number of weeks:2) under the guidance of Aparna Mahajan (Human Resource Manager). His internship activities include learning of various tools for Data Analysis. During this period of her internship program with us, He was exposed to different processes and was found diligent, hardworking, and inquisitive.

We wish his every success in life and career.
For Sigma Institute of Engineering

Authorized Signature

COMPANY JOINING LETTER

DYNAMISITY PRIVATE LIMITED

Be immersed with the Dynamics world and witness the SUCCESS.....!

Date: 29.06.2024

Subject: Offer Letter for Internship

Dear **Rutvik Pradhan,**

We are pleased to offer you an opportunity to join our organization as an Intern at **Dynamisity Private Limited** focusing on AI technology for the position of **Python Developer.** We are thoroughly impressed with your qualifications and believe that your skills and enthusiasm would make a valuable contribution to our team. Your internship will span two weeks, starting from **29th June 2024.**

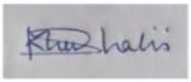
As a prerequisite, we kindly request you to submit a No Objection Certificate (NOC) from your college, affirming your commitment to completing the entire three-month internship period. It is noteworthy that the internship is provided without any charges, and no fees are applicable.

You are required to submit the following documents (self-attested):

- Certificates in proof of age and proof of residence – Aadhar Card.
- Marksheets of previous semester/Degree(s)/or post-Graduation.
- University / Institutes Id.
- 1 Passport size photo

We are excited to have you join our team for this internship program and look forward to working with you. We believe it will be a valuable experience that will contribute to your professional growth.

For Dynamisity



Authorized Signatory

Khushali Trivedi (HR)

Address: B2-505, Westgate Business Bay, Opp. Andaj Party Plot, S.G. Highway Ahmedabad, Gujarat, India – 380015

Email - info@dynamisity.com
Web - https://Dynamisity.com
Phone: (+91) 79 4800 1825

COMPANY CERTIFICATE

DYNAMISITY PRIVATE LIMITED

Be immersed with the Dynamics world and witness the SUCCESS.....!

Date: 13.07.2024

Subject: Internship Experience Letter

We are glad to inform you that **Mr. Rutvik Pradhan** has completed his internship at **Dynamisity Pvt. Ltd.** From **29th June 2024 to 13th July 2024.** He has worked as a "**Python Developer Intern".**

During his internship, we found him extremely inquisitive and hard-working. He was very much interested in learning the functions of our core division and also willing to put in his best efforts and get into the depth of the subject to understand it better.

His association with us was very fruitful and we wish him all the best in his future endeavors.

For Dynamisity

Authorized Signatory

Khushali Trivedi

HR & Admin Head

Address: B2-505, Westgate Business Bay, Opp. Andaj Party Plot, S.G. Highway Ahmedabad, Gujarat, India – 380015

Email - info@dynamisity.com
Web - https://Dynamisity.com
Phone: (+91) 79 4800 1825

Sigma Institute of Engineering

Ajwa-Nimeta Road, Bakrol, Vadodara-390019

DECLARATION

I hereby declare that the Internship report submitted along with the Internship entitled as Data Analysis Internship in partial fulfillment for the degree of Bachelor of Engineering in Computer Engineering to Gujarat Technological University, Ahmedabad is a bonafide record of original project work carried out by me at Sigma Institute of Engineering under the supervision of Dr. Sheshang Degadwala and that no part of this report has been directly copied from any student's reports or taken from any other source, without providing due reference.

Name of Student
Rutvik Bharat Pradhan

Sign of Student

ACKNOWLEDGEMENT

First, I would like to thank Cognispark solutions Pvt. Ltd. for giving me the opportunity to do an internship within the organization. I also would like to thank all the people who worked along with me at Cognispark solutions Pvt. Ltd. It is indeed a great sense of pleasure and immense sense of gratitude that I acknowledge the help of these individuals. I am highly indebted to Dr. Sheshang Degadwala (Internal Guide Head of Department), for the facilities provided to accomplish this internship.

Your sincerely,

Rutvik Bharat Pradhan (210500107039)

ABSTRACT

This report is a detailed overview of my internship journey at Cognispark Solutions Pvt. Ltd. During my internship I have learned a lot about how the industry of Retail and B2C organizations actually works, what are the parameters, and how to work on an actual working and project. I have known about the Data Analysis's roles and responsibilities. I have learned to work in a corporate space which not only enriched me professionally but also helped me grow personally as well. My contribution was appreciated by my supervisor and other members of the department. The career path I would be selecting for myself is quite influenced by my internship as I have had a great opportunity to practically see how a Data Analysis Department is working and evolving in the entire Globe. I have summarized my overall experience, with my learning and challenges faced as an intern.

Table of Contents

Declaration i

Acknowledgment ii

Abstract iii

List of Figures……. ..... ……..…iv

List of Tables v

Chapter 1 Overview of The Company…….. 01

1.1 About the company…….. 01

1.2 Scope of Work……..……..…….. 01

1.3 Company Vision……..…..……..… 02

Chapter 2 Introduction To Internship……. 04

2.1 Project Name……..…….……..... 04

2.2 Introduction To Internship…..……..…… 04

2.3 Internship Goal……...…….. 04

2.4 Internship Scheduling…….……...…… 05

Chapter 3 Installing PyCharm professional …… 06

3.1 Register on jetbrains ide…… 07

3.2 Download pycharm professional from web……...…. 07

3.1.2 Downloading Libraries on python…….……. 07

Chapter 4 Numpy Libraries and its commands…….. 08

4.1 import numpy…….. 09

4.1.1 array()…… ……..…...…….... 9

4.1.2 ones()…… …..…… 9

4.1.3 arange()……... ……..…….... 9

4.1.4 linspace()……..…… 10

4.1.5 Full()…… 10

4.1.6 Random()……..……. 10

4.1.7 Eye()…… ……. 11

4.1.8 Diag()……..……... 11

Chapter 5 Pandas Library and its commands ……..…….. 12

5.1 Import pandas…….…… 12

5.2 Dataframe()…..…….. 13

5.3 rename()…… 13

5.4 Nan()……..…14

5.5 Dupicated()…….14

Chapter 6 Matplotlib and Seaborn library and it's commands …….…….. 15

6.1 import matplotlib()…….15

6.2 import seaborn()..……15

6.3 catplot()……..…….16

6.4 countplot()……..…….17

6.5 distplotplot()……..…….18

Chapter 7 Project details……..19

Chapter 8 Conclusion……. 26

References…….. 27

LIST OF TABLES

CHAPTER 1: OVERVIEW OF THE COMPANY

1.1 ABOUT THE COMPANY

To architecture & implement practical and proven business specific solutions &capabilities on the Dynamics 365 Platform for organizations & Businesses worldwide that are aligned with the organizational values, initiatives andprinciples that help achieve the strategic initiatives We intend to listen,understand and provide customers with cost-effective Dynamics 365implementation experience from beginning to end, with skilled, experienced teams, easy-to-follow project steps, clear and secure solution architectures, and fast, quality delivery.

1.2 Scope of work:

Python programmers have an intensive scope in the field of networking and AI. By learning the advanced concepts from Python programming classes, you can explor job opportunities in this field and work as a network engineer or an AI analyst. Becoming a python expert may seem inaccessible. However, it is a learning process that requires knowing advanced concepts and developing a skillset. Here are the top skills you need to acquire to become a successful Python Developer.

1.3 Company vision:

Our vision is to ignite a passion for learning by creating highly engaging and interactive eLearning courses. We strive to revolutionize the educational experience by leveraging the best authoring tools available, ensuring that each course is impactful and resonates with learners. Through innovative design approaches and dynamic interactions, we aim to set a new standard in eLearning, providing immersive and effective learning solutions that inspire and empower individuals to reach their full potential. Join us in exploring a new era of education where learning is not just a task, but an exciting journey.

CHAPTER 2 : INTODUCTION TO INTERNSHIP

2.1 PROJECT NAME

Analysis Of Purchase Data

2.2 INTRODUCTION TO INTERNSHIP

In today's data-driven world, the ability to analyze and interpret vast amounts of information is an invaluable skill. A data analysis internship offers aspiring analysts the opportunity to dive deep into the world of data, gaining hands-on experience and practical knowledge that bridges the gap between academic learning and real-world application.

A data analysis internship is designed to provide interns with a comprehensive understanding of how data can be used to drive decision-making processes within an organization. Interns learn to collect, process, and analyze data, using statistical tools and software to identify trends, patterns, and insights. These insights are crucial for businesses and organizations, as they inform strategies, optimize operations, and contribute to achieving overall goals.

During the internship, interns are typically exposed to a variety of tasks and responsibilities. They might work on cleaning and preparing data, creating data visualizations, building predictive models, and generating reports. This hands-on experience with real datasets allows interns to apply theoretical knowledge in practical scenarios, enhancing their problem-solving skills and proficiency with data analysis tools such as Excel, SQL, Python, R, and various data visualization platforms like Tableau or Power BI.

2.3 INTERNSHIP GOAL:

Your internship in Data analysis will focus on helping you obtain real-world experience and build the skills you need to thrive in a career in Computer Engineering. You will specifically try to gain technical expertise: You'll have practical experience using various IT systems and applications during your internship.

Your objective will be to improve your technical knowledge by learning how to analyze , visualize and manipulate data according to needs. Learn about industry best practices: As an intern, you will have the chance to become familiar with the guidelines and norms followed by Data Analysis sector. Your objective will be to develop a thorough understanding of these procedures and how they are applied in practice.

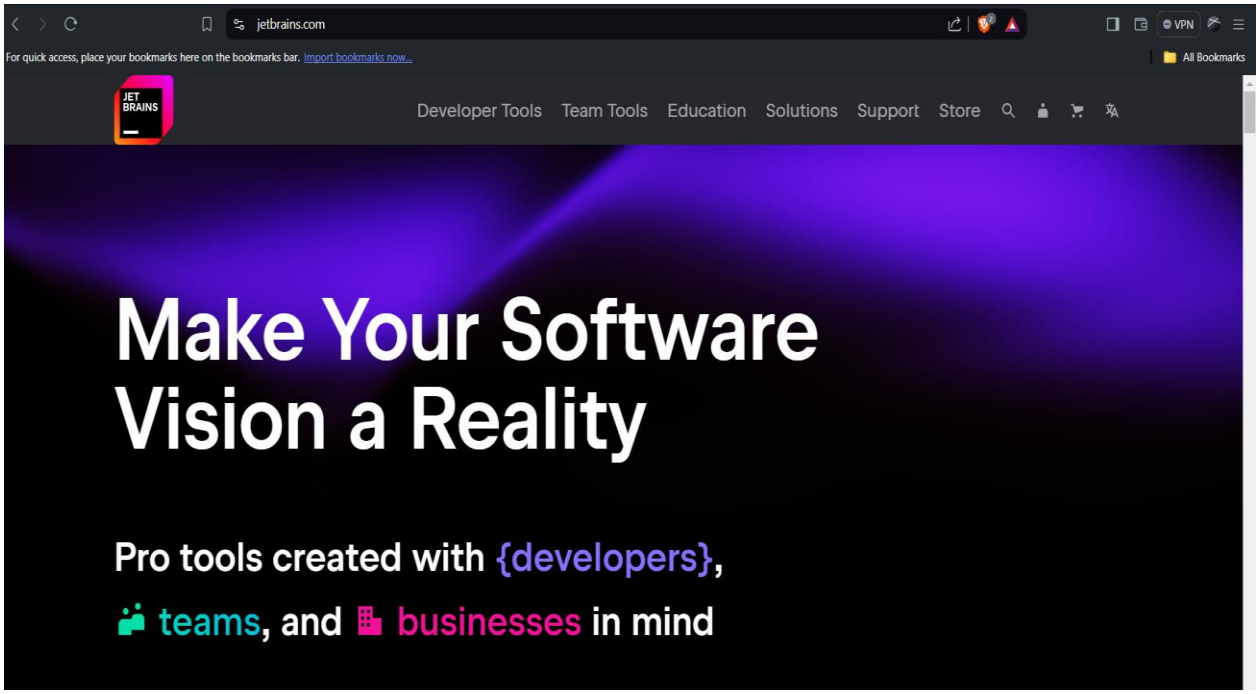
2.4 INTERNSHIP SCHEDULING:

Week wise	**Schedule**
Week 1	Learning about the libraries and installing ipynb notebook .
Week 2	Visualizing and completing the project

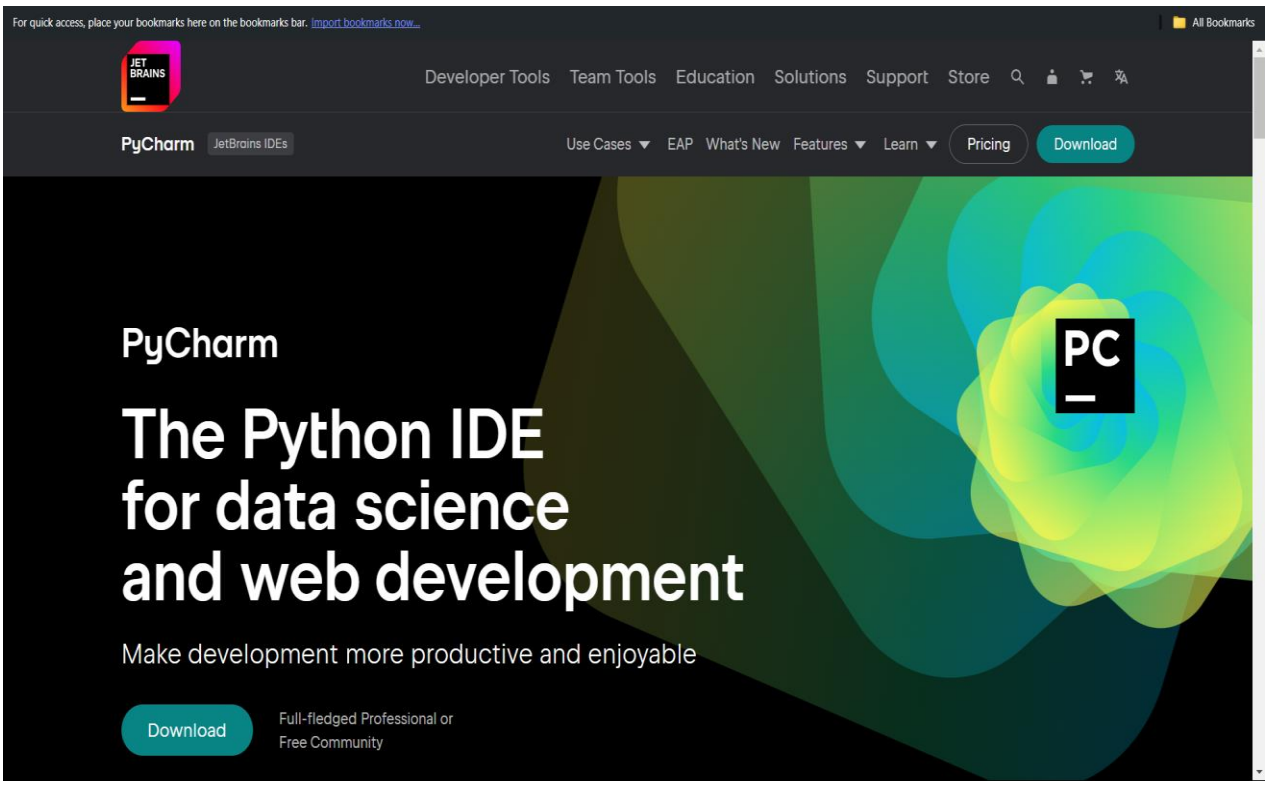
Table 2.5.1 Internship Scheduling

CHAPTER 3:Installing PyCharm Professional

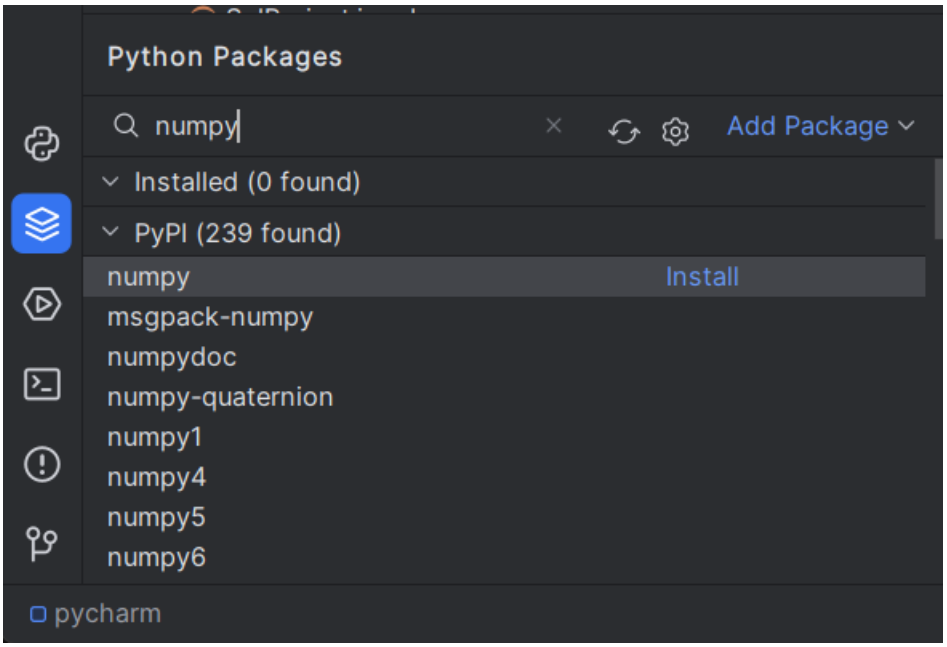
Step 1: Register on jetbrains ide.



Step 2 : Download pycharm professional from web.



Step 3 : Downloading libraries on pycharm.



1.User Interface:

- Jupyter Notebook: Jupyter Notebook offers an interactive web-based interface where you can create and run code cells in a document-like format. It's excellent for creating interactive and shareable documents.

2.Code Exceution:

- Jupyter Notebook: Code is written and executed in cells, which allows for step-by-step execution. This is great for experimenting and documenting your work.

3.Documentation and Visualization:

- Jupyter Notebook: Jupyter is well-suited for creating rich documents that mix code, visualizations, and explanations. It supports various programming languages through kernels.

4.Extension and Integration:

- Jupyter Notebook: Jupyter supports multiple programming languages and has a rich ecosystem of extensions and widgets.

5.Collaboration:

- Jupyter Notebook: Jupyter notebooks can be shared easily, allowing for collaboration and reproducibility. They can also be converted to various formats like HTML or PDF.

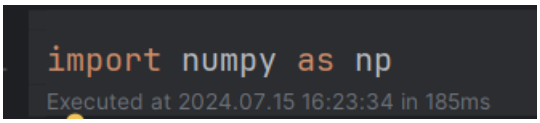
6.Learning Curve:

- Jupyter Notebook: Jupyter's cell-based approach may have a steeper learning curve for those new to it, but it's worth it for its interactive and documentation capabilities.

Chapter 4: NumPy Library and it's commands

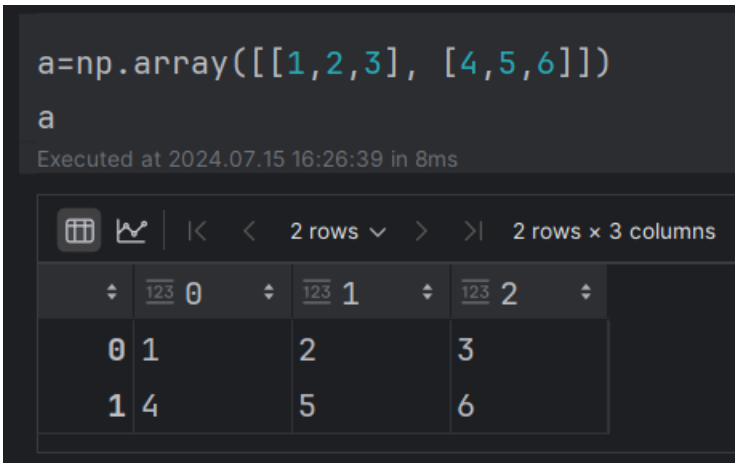
1. Import numpy as np()

   This command will help us to import the library in jupyter notebook. np is a literal which will be used instead of numpy .



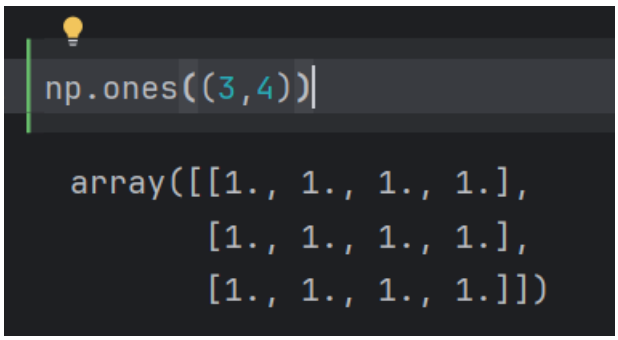
2. array()

   The np,array() function will create the array of the particular size which a user will provide.



3. ones()

   np.ones() will create the array with user defined size. This array will have all the values in 1 particularly.



4. arange()

   np.arange() is used to create an array from starting point to end point. We can also create jump steps in between using this function.

```
np.arange(1,10,2)

Executed at 2024.07.15 16:31:41 in 7ms

array([1, 3, 5, 7, 9])
```

5. linspace()

   np.linspace() has the similar function like np.arange() in which the user can create an array by giving the starting and ending values. The main unique feature of linspace function is that the steps you give in this function will create the array with equal distance according to the value provide by the user.

```
np.linspace(1,10,5)

array([ 1.  ,  3.25,  5.5 ,  7.75, 10.  ])
```

6. full()

   np.full() will create an matrix with desired rows and columns and will fill each of the value of an amatrix with the desired value provided by user.

```
data=np.full((3,4), 8)
data

Executed at 2024.07.15 16:35:30 in 9ms
```

3 rows | 3 rows × 4 columns | Change page size

	0	1	2	3
0	8	8	8	8
1	8	8	8	8
2	8	8	8	8

7. random()

np.random() is used to generate random values between 0 and 1 which can be used for experiments.

```
np.random.randint(3,7,16)

array([4, 3, 5, 3, 6, 5, 4, 6, 5, 4, 6, 4, 6, 3, 6, 3])
```

8. eye()

np.eye() function will create a matrix with provided number of diagonals and each value of diagonal starting from left to right will have a value as one. It is basically used to make an diagonal matrix .

```
1  np.eye(3)

   array([[1., 0., 0.],
          [0., 1., 0.],
          [0., 0., 1.]])
```

9. diag()

np.diag() function is used to create a matrix which will have the diagonals with user's desires and rest of the values will be 0. This is also a diagonal matrix with different diagonal values.

```
a=np.diag([1,2,3,4])
a
Executed at 2024.07.15 16:45:11 in 8ms
```

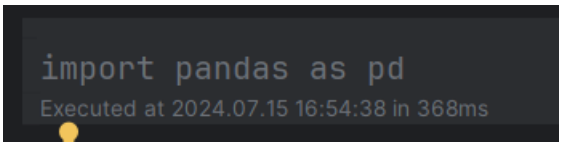
4 rows × 4 columns

	0	1	2	3
0	1	0	0	0
1	0	2	0	0
2	0	0	3	0
3	0	0	0	4

CHAPTER 5: Pandas Library and it's commands.

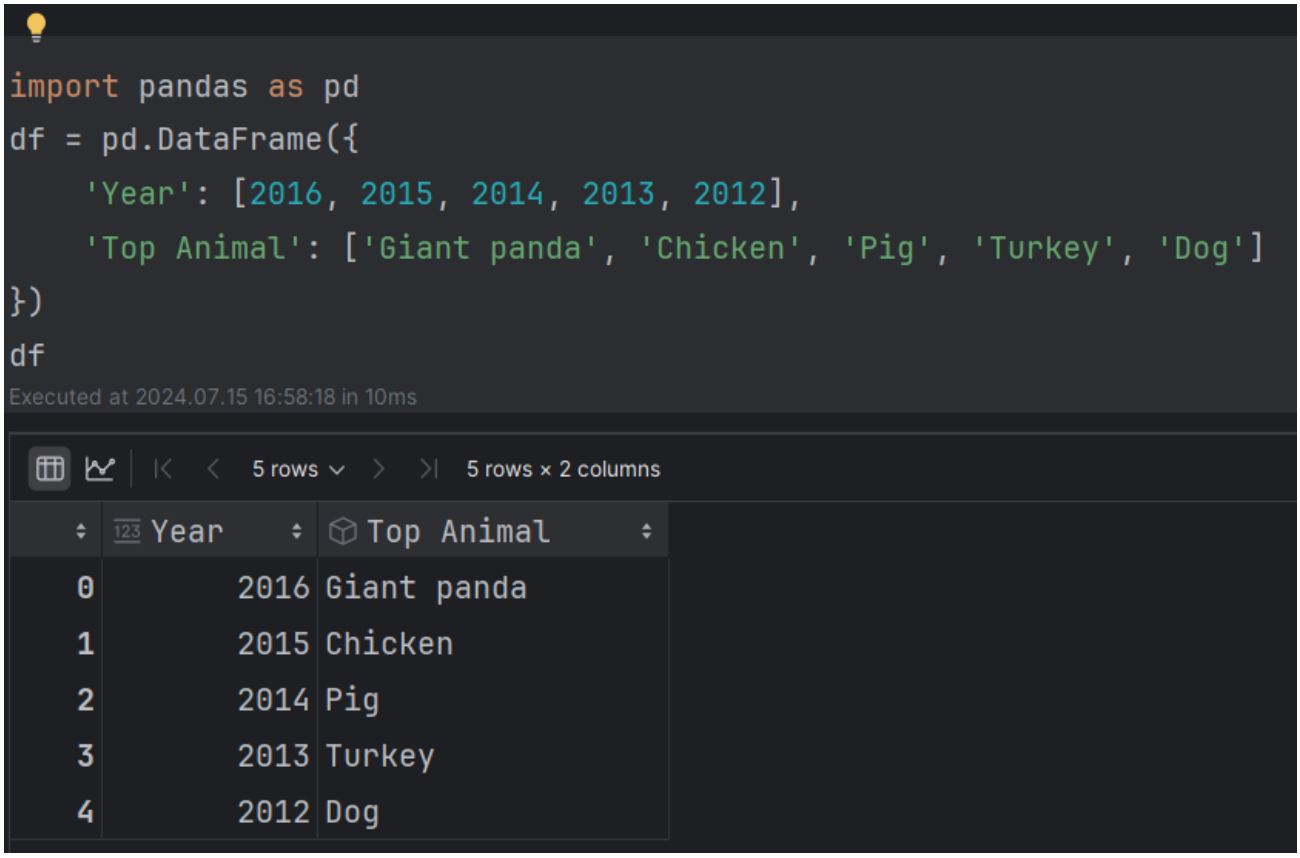
1. Import pandas as pd()

   This is used to import the pandas library. Pd is used as a literal instead of pandas.



2. DataFrame()

   This function is used to make the dataframes in pandas librabry.



3. rename()

df.rename() function is used to rename the columns made earlier in the dataframe.

```
df.rename(columns={
    'Year': 'Calendar Year',
    'Top Animal': 'Favorite Animal',
}, inplace=True)
df
```

Executed at 2024.07.15 16:58:08 in 7ms

5 rows | 5 rows × 2 columns

	Calendar Year	Favorite Animal
0	2016	Giant panda
1	2015	Chicken
2	2014	Pig
3	2013	Turkey
4	2012	Dog

4. Nan function

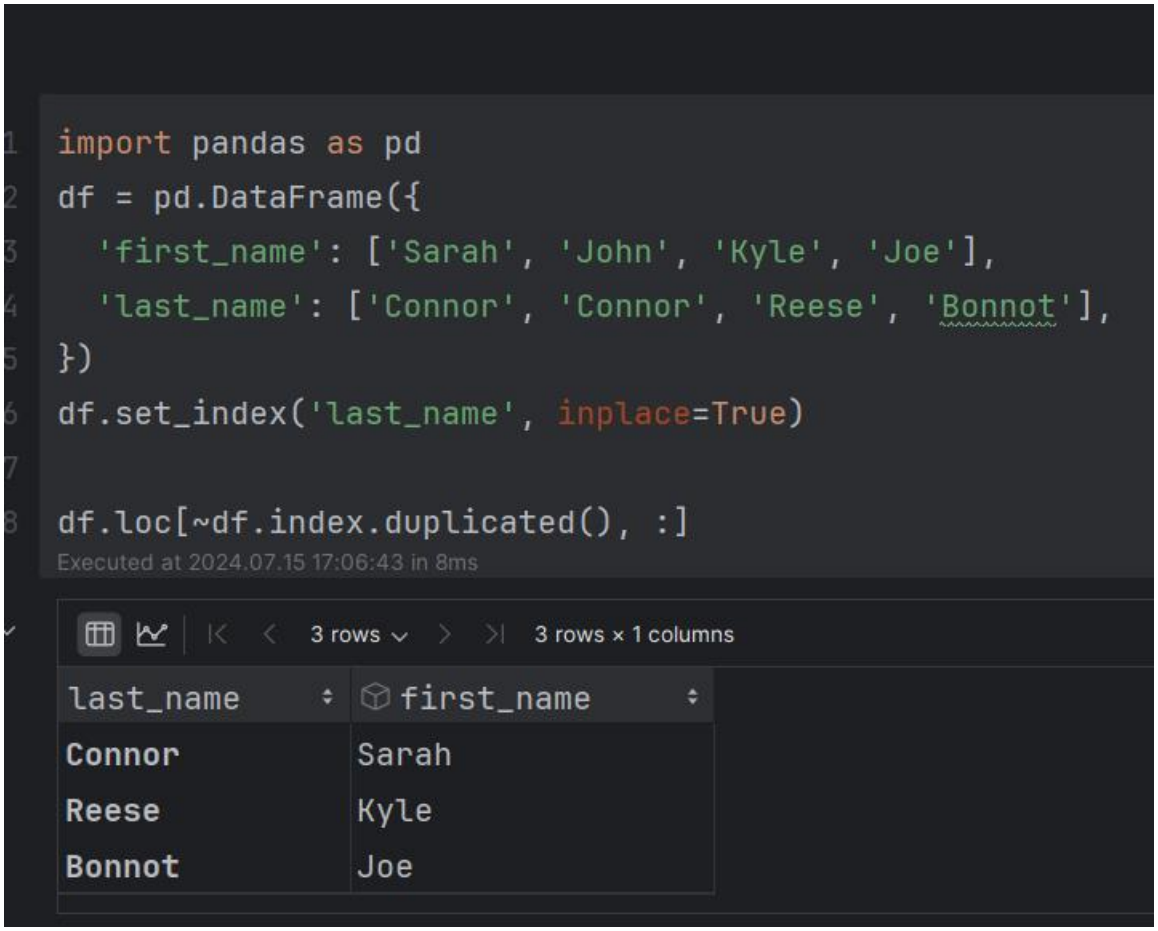
This function is used to replace null values present in a column.

```
import numpy as np
import pandas as pd
df = pd.DataFrame({
  'dogs': [5, 10, np.nan, 7],
})

df['dogs'].replace(np.nan, 0, regex=True)
```

5. Duplicated()

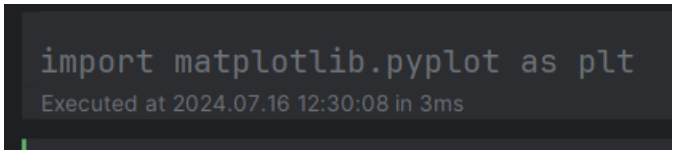
Usually it may happen that the rows are duplicated in the dataset so we have to remove it using duplicated() function.



Chapter 6: Matplotlib and Seaborn Library and it's commands.

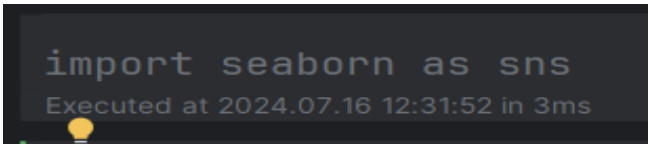
1. Import matplotlib.pyplot as plt

   This will import the matplotlib library in your ipynb file. Plt is used as a literal so it is used to call the library .



2. Import seaborn as sns

   This will import the seaborn library in your ipynb file. Sns is used as a literal so it is used to call the library.

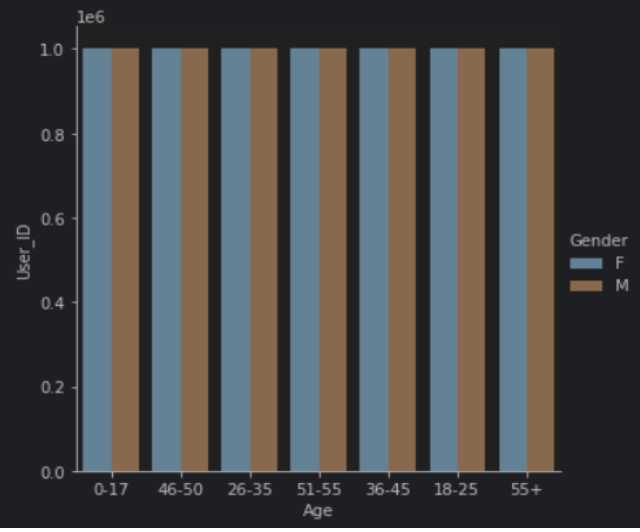


3. Catplot()

   Catplot is used to plot the categorical plots with different values. It is used for visualization of categorical data.

```
sns.catplot(data=data, x="Age", y="User_ID", kind="bar", hue='Gender')
```

```
<seaborn.axisgrid.FacetGrid at 0x7fc95e669b80>
```

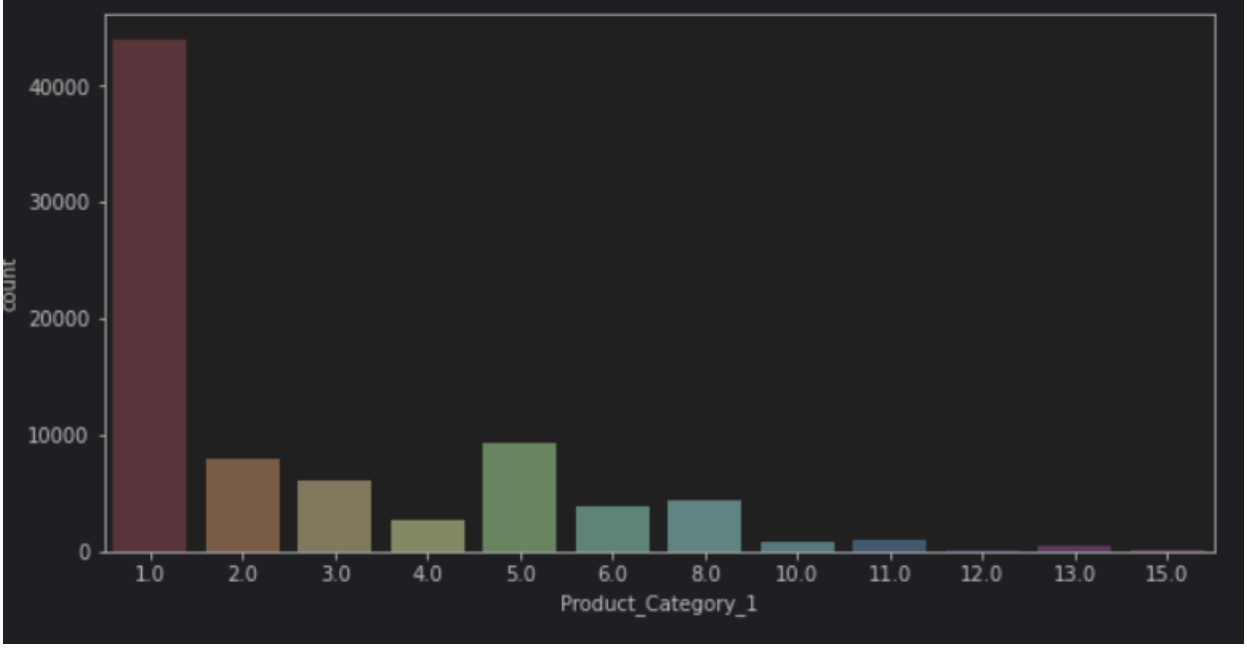


4. Countplot()

   Countplot is used to show the counts of observations in each categorical bin using bars. Widely used visualization plot.

```
plt.figure(figsize=(10,5))
sns.countplot(data['Product_Category_1'])
```

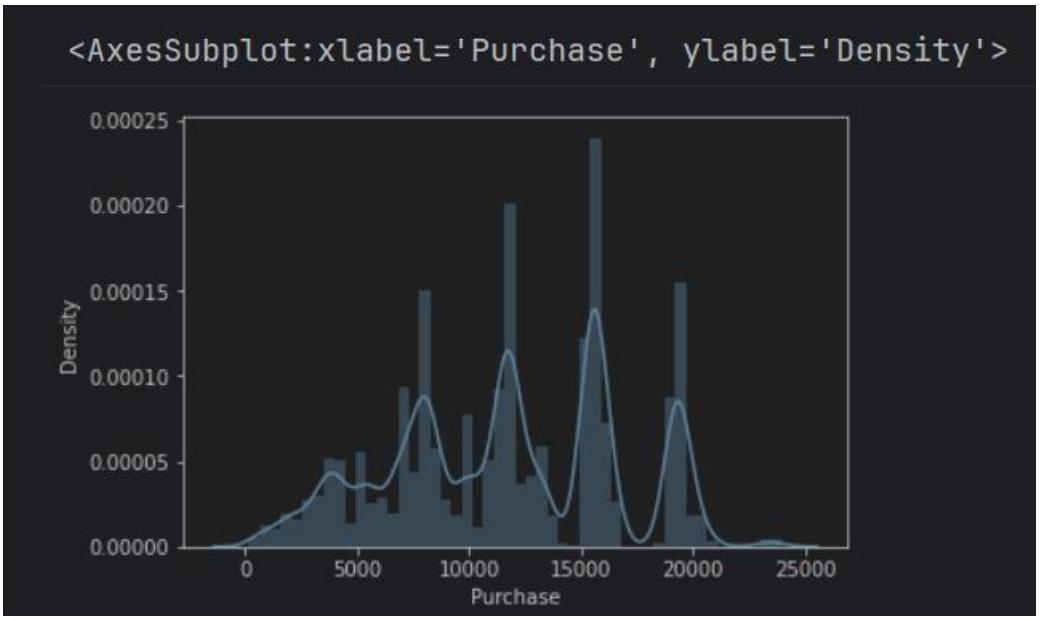
```
<AxesSubplot:xlabel='Product_Category_1', ylabel='count'>
```



5. Distplot()

   Distplot is used to visualize the distribution amongst the dataset present in you project. It is used to visualize the distributions of the data.

```
sns.distplot(data['Purchase'])
```


<AxesSubplot:xlabel='Purchase', ylabel='Density'>
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
Density
0
5000
10000
15000
20000
25000
Purchase

CHAPTER 7: Project Details

1. Problem Statement:

   You have been provided purchase data for various customers across a vertical. You need to apply your learnings from Data Manipulation, Data Visualization, and statistical analysis to come up with actionable insights about the data.

2. Tasks to be performed:

   I. Perform a Detailed EDA for the Data with inferences from each of the actions.
   II. Using Statistical Analysis, find out statistical evidence for the following:
   a. It was observed that the average purchase made by the Men of the age 18-25 was 10000. Is it still the same?
   b. It was observed that the percentage of women of the age that spend more than 10000 was 35%. Is it still the same?
   c. Is the average purchase made by men and women of the age 18-25 same?
   d. Is the percentage of men who have spent more than 10000 the same for the ages 18-25 and 26-35?

3. EDA

```
#importing the required libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```
Executed at 2024.07.16 12:26:19 in 4ms

```
#Data loading
data = pd.read_csv("PURCHASE.CSV")
#data = pd.read_csv(r"C:\Users\Dell\Downloads\train (6).csv")
```
Executed at 2024.07.16 12:26:30 in 197ms

```
data
```

	User_ID	Product_ID	Gender	Age	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
0	1000001	P00069042	F	0-17	A	2	0.0	3.0	NaN	NaN	8370.0
1	1000001	P00248942	F	0-17	A	2	0.0	1.0	6.0	14.0	15200.0
2	1000001	P00087842	F	0-17	A	2	0.0	12.0	NaN	NaN	1422.0
3	1000001	P00085442	F	0-17	A	2	0.0	12.0	14.0	NaN	1057.0
4	1000002	P00285442	M	55+	C	4+	0.0	8.0	NaN	NaN	7969.0
...	...	...	...	...	...	...	...	...	...	...	...
263010	1004473	P00041942	M	36-45	B	3	0.0	5.0	18.0	NaN	3722.0
263011	1004473	P00115142	M	36-45	B	3	0.0	1.0	8.0	17.0	19253.0
263012	1004473	P00188442	M	36-45	B	3	0.0	5.0	7.0	NaN	3608.0

Add Code Cell | Add Markdown Cell

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 263015 entries, 0 to 263014
Data columns (total 11 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   User_ID                     263015 non-null  int64
 1   Product_ID                  263014 non-null  object
 2   Gender                      263014 non-null  object
 3   Age                         263014 non-null  object
```

```
#print top 5 rows of the dataset
data.head()
```

	User_ID	Product_ID	Gender	Age	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
0	1000001	P00069042	F	0-17	A	2	0.0	3.0	NaN	NaN	8370.0
1	1000001	P00248942	F	0-17	A	2	0.0	1.0	6.0	14.0	15200.0
2	1000001	P00087842	F	0-17	A	2	0.0	12.0	NaN	NaN	1422.0
3	1000001	P00085442	F	0-17	A	2	0.0	12.0	14.0	NaN	1057.0
4	1000002	P00285442	M	55+	C	4+	0.0	8.0	NaN	NaN	7969.0

```
#column names
data.columns
```

```
data['Stay_In_Current_City_Years'].unique()
```

```
array(['2', '4+', '3', '1', '0', nan], dtype=object)
```

```
#Changing all values from 4+ to 4
data['Stay_In_Current_City_Years'] = data['Stay_In_Current_City_Years'].replace('4+', '4')
```

```
#to check the null values
data.isnull().sum()
```

```
User_ID                         0
Product_ID                      1
Gender                          1
Age                             1
City_Category                   1
Stay_In_Current_City_Years      1
Marital_Status                  1
Product_Category_1              1
Product_Category_2          81514
```

```
#changing the datatype to integer
data.dropna(inplace=True)
data['Stay_In_Current_City_Years'] = data['Stay_In_Current_City_Years'].astype(int)
```

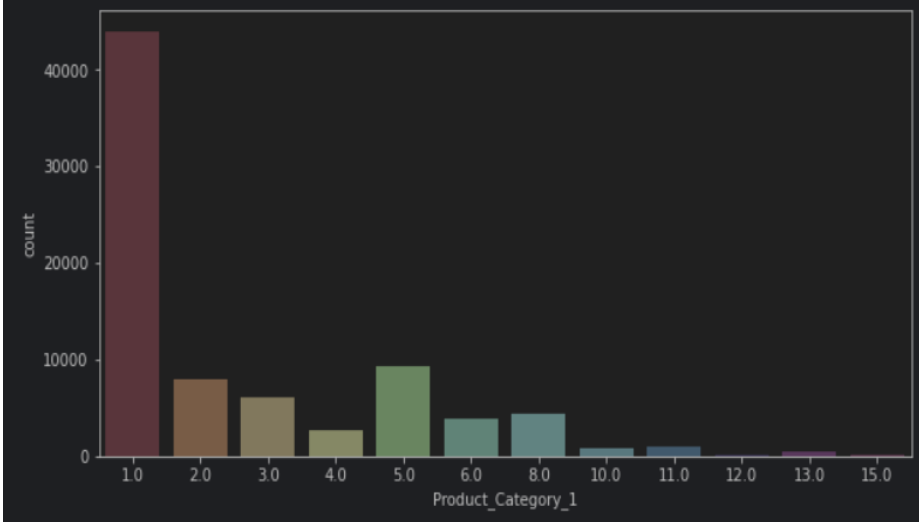
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

4. Observations:

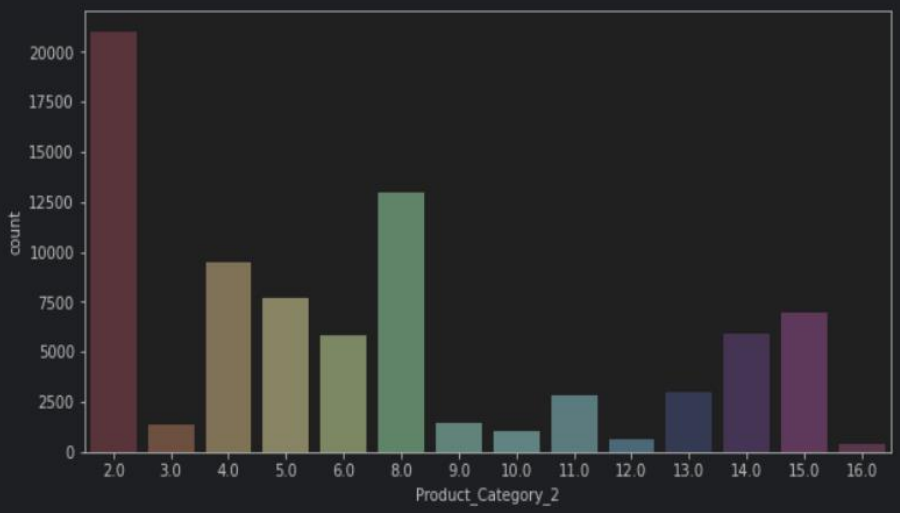
```
Warnings.warn(
```

```
<AxesSubplot:xlabel='Product_Category_1', ylabel='count'>
```

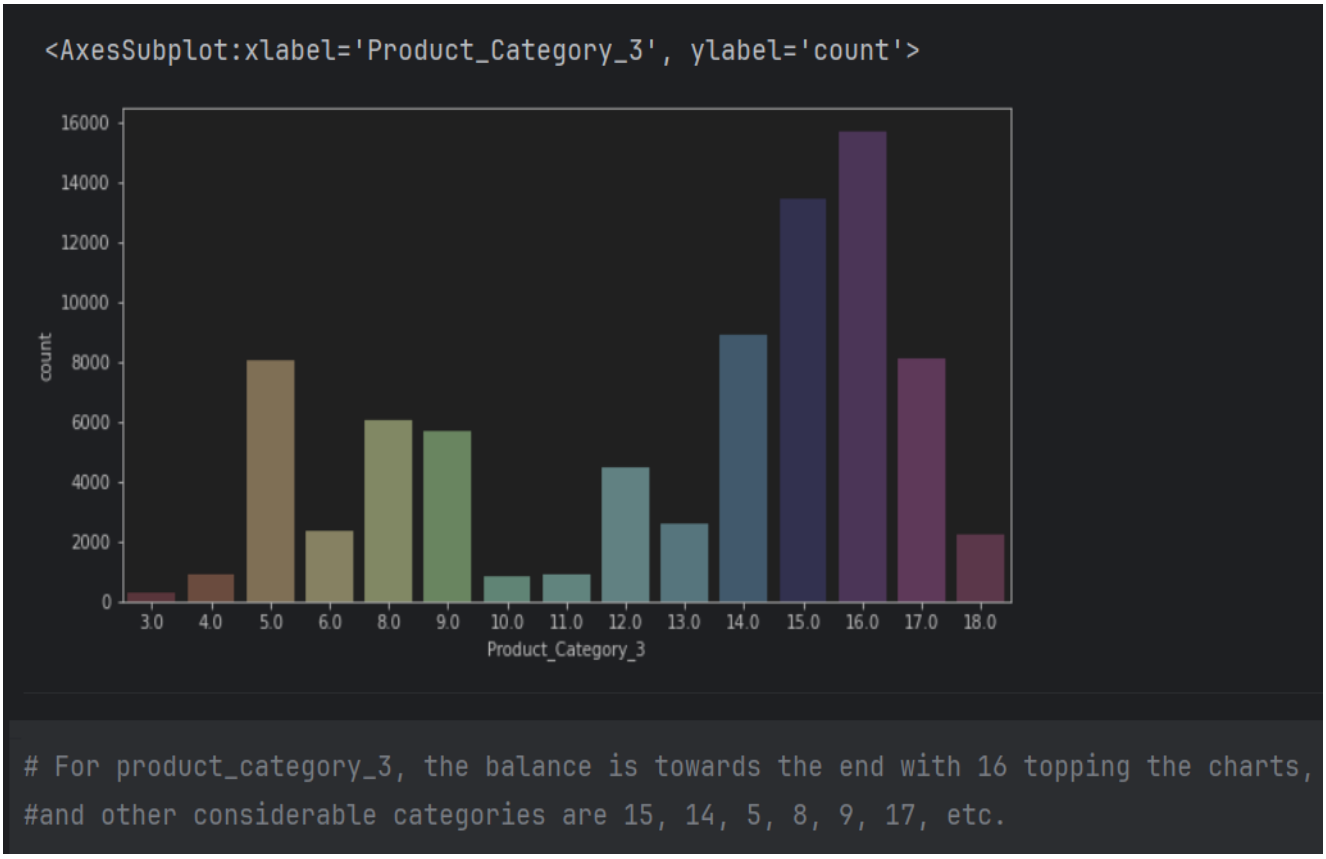



```
### The product_category_1 sees a great rise of product category 1 and lowest in 12th category
```

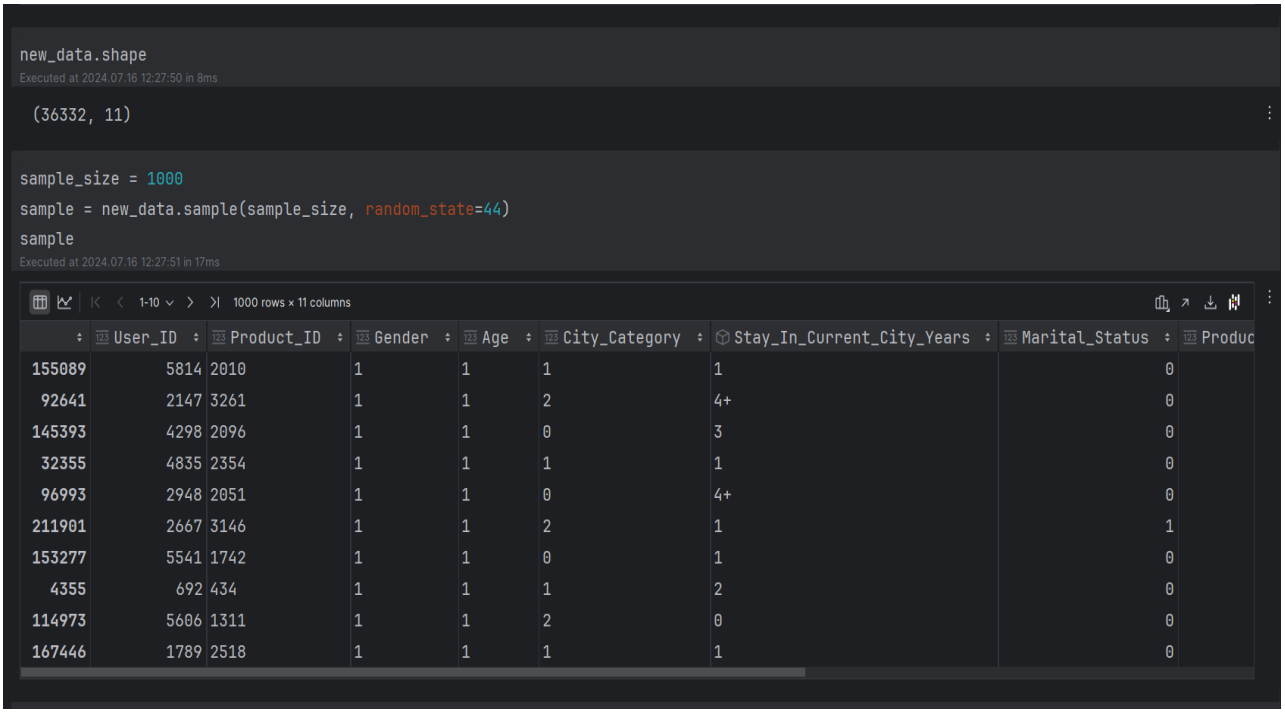
```
<AxesSubplot:xlabel='Product_Category_2', ylabel='count'>
```




```
# Product_Category_2 sees a considerate balance among categories.
#With category 2 topping the charts, and other considerable categories are 8,,4,5,6,14,15,etc.
```



- Evidence 1: It was observed that the average purchase made by the Men of the age 18-25 was 10000. Is it still the same?



	User_ID	Product_ID	Gender	Age	City_Category	Stay_In_Current_City_Years	Marital_Status	Produc
155089	5814	2010	1	1	1	1	0	
92641	2147	3261	1	1	2	4+	0	
145393	4298	2096	1	1	0	3	0	
32355	4835	2354	1	1	1	1	0	
96993	2948	2051	1	1	0	4+	0	
211901	2667	3146	1	1	2	1	1	
153277	5541	1742	1	1	0	1	0	
4355	692	434	1	1	1	2	0	
114973	5606	1311	1	1	2	0	0	
167446	1789	2518	1	1	1	1	0	

```
p_mean = 10000
```
Executed at 2024.07.16 12:27:52 in 5ms

```
sample_mean = sample["Purchase"].mean()
print(sample_mean)
```
Executed at 2024.07.16 12:27:54 in 5ms

```
9393.114
```

```
#one sample t test
from scipy.stats import ttest_1samp
```
Executed at 2024.07.16 12:27:55 in 6ms

```
t_stat, p_value = ttest_1samp(sample['Purchase'], p_mean)
print(t_stat, p_value)
```
Executed at 2024.07.16 12:27:56 in 9ms

```
-3.7532811608846286 0.00018459576198200494
```

```
#p value is less than 0.05 , reject the null hypothesis.
#Therefore, the mean purchase for men ages 18-25 is not 10000.
```

- Evidence 2: It was observed that the percentage of women of the age that spend more than 10000 was 35%. Is it still the same?

```
#NUll hypothesis  - proportion is 35%
#Alternate hypothesis - proportion is not 35%
```

```
data_new = data.loc[(data['Purchase']>10000)]
```
Executed at 2024.07.16 12:29:25 in 11ms

```
data_new.shape
```
Executed at 2024.07.16 12:29:26 in 5ms

```
(91057, 11)
```

```
#no of women in the sample
count = data_new["Gender"].value_counts()[0]

#no. of obs
nobs = len(data_new["Gender"])

#hypothesise value
p0 = 0.35
```
Executed at 2024.07.16 12:29:26 in 4ms

```
count
```
Executed at 2024.07.16 12:29:26 in 4ms

```
18685
```

```
data_new["Gender"].value_counts()/nobs
```

Executed at 2024.07.16 12:29:27 in 11ms

```
Gender
1    0.794799
0    0.205201
Name: count, dtype: float64
```

```
#Ztest - used to determine whether two population means are different when the variances are known.
```

Executed at 2024.07.16 12:29:27 in 2ms

```
from statsmodels.stats.proportion import proportions_ztest
```

Executed at 2024.07.16 12:29:27 in 5ms

```
z_stat, p_value = proportions_ztest(count=count,nobs=nobs, value=p0)
print(z_stat, p_value)
```

Executed at 2024.07.16 12:29:28 in 13ms

```
-108.19403500300525 0.0
```

```
#p-value is less than 0.05, reject the null hypothesis i.e.., proportion is not 35%
```

Executed at 2024.07.16 12:29:28 in 5ms

- Evidence 3 : Is the average purchase made by men and women of the age 18-25 same?

```
data_men = data.loc[(data['Gender'] == 1)& (data['Age'] == 1)]
data_women = data.loc[(data['Gender'] == 0) & (data['Age'] == 1)]
```
Executed at 2024.07.16 12:29:29 in 14ms

```
#creating samples
data_men_sample = data_men.sample(500, random_state=0)
data_women_sample = data_women.sample(500, random_state=0)
```
Executed at 2024.07.16 12:29:29 in 7ms

```
#checking variances of the two samples
print(data_men_sample.Purchase.var())
print(data_women_sample.Purchase.var())
```
Executed at 2024.07.16 12:29:29 in 3ms

```
27003556.191919837
21090163.437611222
```

```
#sample means
print(data_men_sample.Purchase.mean())
print(data_women_sample.Purchase.mean())
```
Executed at 2024.07.16 12:29:30 in 4ms

```
9919.908
8553.708
```

```
#compute f statistic
from scipy.stats import f #f-test is used to compare the variances
F = data_men_sample.Purchase.mean()/data_women_sample.Purchase.mean()
F
```

Executed at 2024.07.16 12:29:30 in 6ms

```
1.15972020555296
```

```
#calculating the degrees of freedom
#Degrees of freedom is the number of independent pieces of
#information used to calculate a statistic.
df1 = len(data_men_sample) - 1
df2 = len(data_women_sample) - 1
print(df1, df2)
```

Executed at 2024.07.16 12:29:30 in 4ms

```
499 499
```

```
#p-value
#cdf - The cumulative distribution function is used
#to describe the probability distribution of random variables
import scipy
scipy.stats.f.cdf(F, df1, df2)
```

Executed at 2024.07.16 12:29:32 in 7ms

```
0.9508800574313439
```

```
#the p-value is greater than 0.05, do not reject the null hypothesis.
#the null hypothesis is true. The average purchases are same.
```

Executed at 2024.07.16 12:29:32 in 3ms

- Evidence 4 : Is the percentage of men who have spent more than 10000 the same for the ages 18-25 and 26-35?

```
data_age1 = data.loc[(data['Age'] == 1) & (data['Purchase'] > 10000)]
data_age2 = data.loc[(data['Age'] == 2) & (data['Purchase'] > 10000)]
```
Executed at 2024.07.16 12:29:37 in 15ms

```
data_age1_sample = data_age1.sample(1000, random_state=0)
data_age2_sample = data_age2.sample(1000, random_state=0)
```
Executed at 2024.07.16 12:29:37 in 8ms

```
count = [(data_age1_sample['Gender'] == 1).sum(), (data_age2_sample['Gender'] == 1).sum()]

nobs = [(len(data_age1_sample)), len(data_age2_sample)]
```
Executed at 2024.07.16 12:29:37 in 5ms

```
count
```
Executed at 2024.07.16 12:29:38 in 5ms

```
[790, 811]
```

```
nobs
```
Executed at 2024.07.16 12:29:38 in 5ms

```
[1000, 1000]
```

```
from statsmodels.stats.proportion import proportions_ztest
stat_2sample, p_value_2sample = proportions_ztest(count=count, nobs=nobs)
```
Executed at 2024.07.16 12:29:38 in 4ms

```
p_value_2sample
```
Executed at 2024.07.16 12:29:38 in 4ms

```
0.2399792241715063
```

```
#p value is more than 0.05, accept the null hypthesis.
#therefore, Percentage of the men in the age groups is same
```
Executed at 2024.07.16 12:29:38 in 3ms

CHAPTER 8: CONCLUSION

During my internship as a Data Analyst, I had the invaluable opportunity to apply my theoretical knowledge in a real-world setting, significantly enhancing my skills in data manipulation, analysis, and visualization. Over the course of the internship, I worked on multiple projects that required a deep dive into data sets, utilizing various tools and techniques to extract meaningful insights and support decision-making processes.

Key accomplishments during this internship include:

1. **Data Cleaning and Preparation:**
    - Successfully cleaned and prepared multiple large data sets, ensuring data integrity and accuracy for analysis. This involved handling missing values, correcting data entry errors, and standardizing formats.
2. **Data Analysis:**
    - Conducted thorough exploratory data analysis (EDA) to uncover patterns, trends, and anomalies. Applied statistical methods to derive insights that informed strategic business decisions.
3. **Data Visualization:**
    - Created clear and compelling data visualizations using tools like Power BI, Tableau, and Excel. These visualizations helped communicate complex data findings to non-technical stakeholders effectively.
4. **Collaboration and Communication:**
    - Worked closely with cross-functional teams, including marketing, finance, and product development, to understand their data needs and deliver actionable insights. Improved my ability to translate technical findings into business recommendations.
5. **Tool Proficiency:**
    - Gained hands-on experience with industry-standard tools such as SQL, Python, Power BI, and Excel. Developed automated scripts and dashboards that streamlined data processing and reporting tasks.
6. **Project Impact:**
    - Contributed to projects that had a tangible impact on the company, including optimizing marketing campaigns, improving customer segmentation, and enhancing financial forecasting models.

This internship has solidified my passion for data analysis and has equipped me with practical skills and experiences that I will carry forward in my career. I am grateful for the mentorship and support from my colleagues, which has been instrumental in my professional growth. I look forward to continuing to apply my skills in data analysis and contributing to data-driven decision-making in future role

REFERENCES

1. https://github.com/Rutvik2803/Purchase-Data-Analysis
2. https://www.w3schools.com/datascience/ds_analyze_data.asp