Exploring Dataset

Using Pandas

Presented by

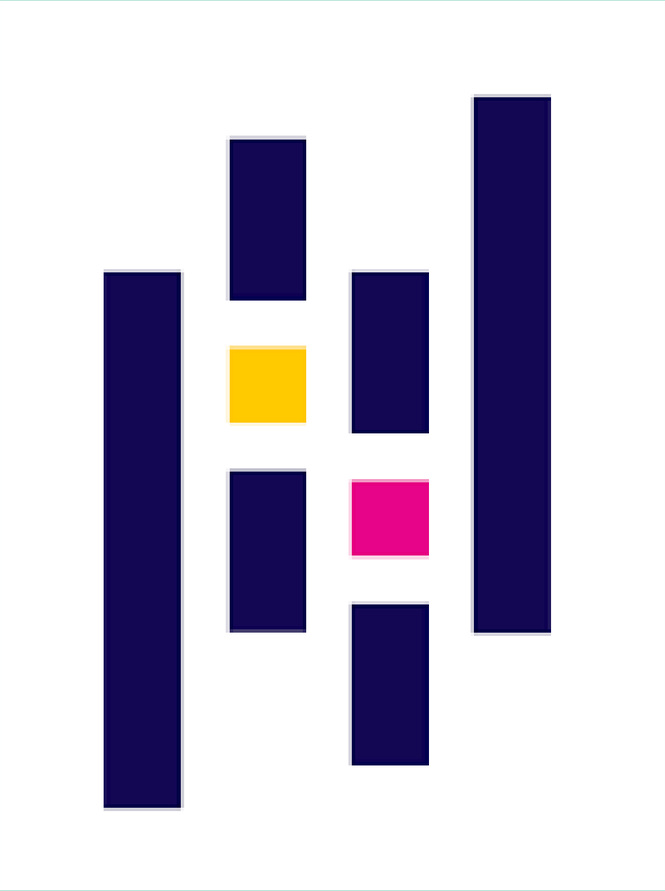
Rajesh Kanna R.
99220041074

What is Pandas?

Pandas is a Python library renowned for its versatility in data manipulation and analysis.

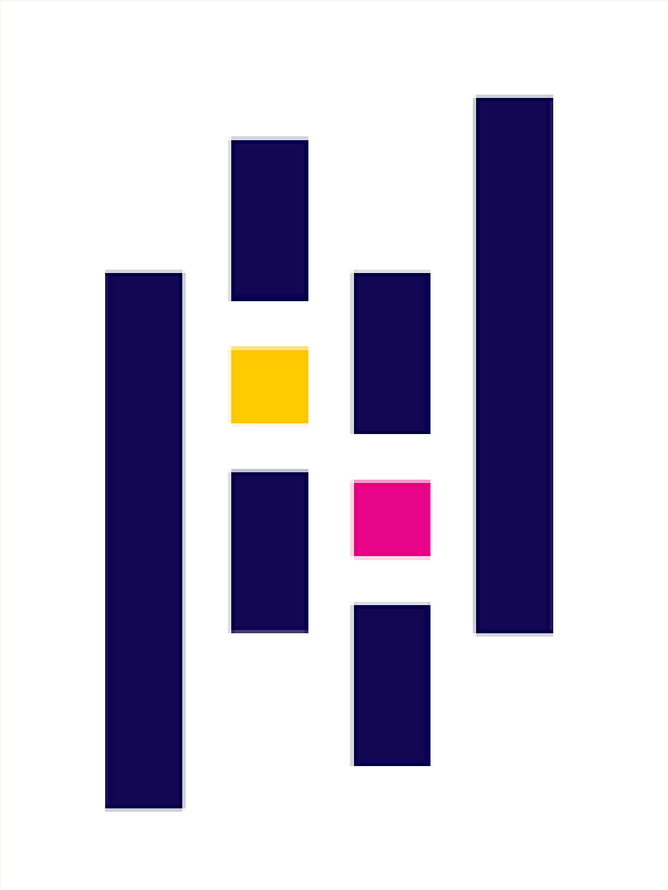
With intuitive data structures like DataFrame and Series, Pandas simplifies tasks such as cleaning, transforming, and analyzing structured data.

Its seamless integration with other libraries enhances its capabilities, making it indispensable for efficient data processing and exploration.



Objective

The objective of this project is to conduct an exploratory data analysis (EDA) of the 'E-Commerce Purchase' dataset using Pandas, with a focus cleaning and preparing the data, visualizing the data, visualizing key findings, and identifying patterns and trends in customer demographics, purchasing behavior, and preferences.



Packages installed

```
[2]: import pandas as pd
     import matplotlib.pyplot as plt
```

Implementation of the project

Basic things

- Reading the first 10 column in the Dataset

```
# Basic
# 1.Display Top 10 rows in the Dataset

data.head(10)
```

	Address	Lot	AM or PM	Browser Info	Company	Credit Card	CC Exp Date	CC Security Code	CC Provider	Email	Job	
0	16629 Pace Camp Apt. 448\nAlexisborough, NE 77...	46 in	PM	Opera/9.56.(X11; Linux x86_64; sl-SI) Presto/2...	Martinez-Herman	6011929061123406	02/20	900	JCB 16 digit	pounlap@yahoo.com	Scientist, product/process development	14
1	9374 Jasmine Spurs Suite 508\nSouth John, TN 8...	28 en	PM	Opera/8.93.(Windows 98; Win 9x 4.90; en-US) Pr...	Fletcher, Richards and Whitaker	333775816664335 6	11/18	561	Mastercard	anthony41@reed.com	Drilling engineer	
2	Unit 0065 Box 5052\nDPO AP 27450	94 vE	PM	Mozilla/5.0 (compatible; MSIE 9.0; Windows NT ...	Simpson, Williams and Pham	675957666125	08/19	699	JCB 16 digit	amymiller@morales-harrison.com	Customer service manager	1
3	7780 Julia Fords\nNew Stacy, WA 45798	36 vm	PM	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_8_0 ...	Williams, Marshall and Buchanan	6011578504430710	02/24	384	Discover	brent16@olson-robinson.info	Drilling engineer	
4	23012 Munoz Drive Suite 337\nNew Cynthia, TX 5...	20 IE	AM	Opera/9.58.(X11; Linux x86_64; it-IT) Presto/2...	Brown, Watson and Andrews	6011456623207998	10/25	678	Diners Club / Carte Blanche	christopherwright@gmail.com	Fine artist	
5	7502 Powell Mission Apt. 768\nTravisland, VA 3...	21 XT	PM	Mozilla/5.0 (Macintosh; U; PPC Mac OS X 10_8_5...	Silva-Anderson	3024618519628 7	07/25	7169	Discover	yngoyen@gmail.com	Fish farm manager	
6	93971 Conway Causeway\nAndersonburgh, AZ 75107	96 XI	AM	Mozilla/5.0 (compatible; MSIE 7.0; Windows NT ...	Gibson and Sons	6011398782655569	07/24	714	VISA 16 digit	olivia04@yahoo.com	Dancer	1
7	260 Rachel Plains Suite 360\nCastroberg, WV 24...	96 pG	PM	Mozilla/5.0 (X11; Linux i686) AppleWebKit/5350...	Marshall-Collins	561252141909	06/25	256	VISA 13 digit	phillip48@parks.info	Event organizer	2
8	2129 Dylan Burg\nNew Michelle, ME 28650	45 JN	PM	Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_7...	Galloway and Sons	180041795790001	04/24	899	JCB 16 digit	kdavis@rasmussen.com	Financial manager	14
9	3795 Dawson Extensions\nLake Tinafort, ID 88739	15 Ug	AM	Mozilla/5.0 (X11; Linux i686; rv:1.9.7.20) Gec...	Rivera, Buchanan and Ramirez	4396283918371	01/17	931	American Express	sjcoleman@hunt-huerta.com	Forensic scientist	

```
# Basic
# 1.Display Last 10 rows in the Dataset
```

Data Cleaning

- Handling missing values in the dataset

```
[15]: # basic
      # 2.Check Null values in the dataset

      data.isnull().sum()
```

```
[15]: Address             0
      Lot                 0
      AM or PM            0
      Browser Info        0
      Company             0
      Credit Card         0
      CC Exp Date         0
      CC Security Code    0
      CC Provider         0
      Email               0
      Job                 0
      IP Address          0
      Language            0
      Purchase Price      0
      dtype: int64
```

Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is an essential step in the data analysis process, as it helps us understand the structure of the dataset, identify patterns, and generate hypotheses for further analysis.
- In this project, we conducted a thorough EDA of the 'E-Commerce Purchase' dataset, using Pandas to perform basic and intermediate analyses.

Basic Analysis

- We counted the number of people with French as their language and job titles containing "Engineer."

```
[26]: # basic
      # 9.How many people have French 'fr' as their language ?

      len(data[data['Language']=='fr'])

[26]: 1097

[35]: # basic
      # 10.How many people's job title contains Engineer ?

      len(data[data['Job'].str.contains('engineer',case=False)])

[35]: 984
```

Intermediate Analysis

- We identified people with Mastercard as their credit card provider who made purchases above a certain threshold.
- We counted the number of people with a credit card that expires in 2020.
- We identified the top 5 most popular email providers.

```
[22]: # Intermediate
      # 1.How many people have Mastercard as their credit card provider and made a purchase above 50 ?

      len(data[(data['CC Provider']=="Mastercard") & (data['Purchase Price']>50)])

[22]: 405

[34]: # Intermediate
      # 2.How many people have a credit card that expires in 2020 ?

      len(data[data['CC Exp Date'].apply(lambda x:x[3:]=='20')])

[34]: 988

[35]: # Intermediate
      # 3.Tp 5 most populare Email providers (e.g. gamil.com , yahoo.com etc...)

      list1=[]
      for email in data['Email']:
          list1.append(email.split('@')[1])

[36]: data['temp']=list1  # creating a new column in dataset

[38]: data['temp'].value_counts().head()

[38]: temp
      hotmail.com     1638
      yahoo.com       1616
      gmail.com       1605
      smith.com         42
      williams.com      37
      Name: count, dtype: int64
```

Data Visualization

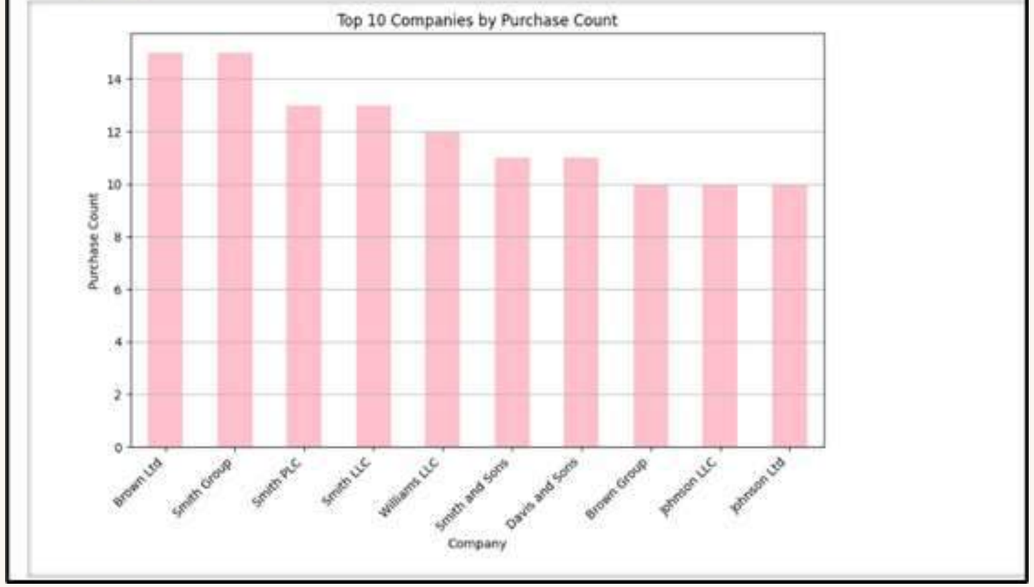
- Data visualization plays a crucial role in understanding complex datasets, as it allows us to visually explore patterns, trends, and relationships that may not be apparent from the raw data alone.
- In this project, we used various data visualization techniques to enhance our understanding of the 'E-Commerce Purchase' dataset and communicate our findings effectively.

Bar Plot of Purchase Counts by Company

- This visualization allows us to compare the purchase counts across different companies.
- It contributes to the analysis by highlighting the popularity of certain companies among customers.

```
# DATA VISUALIZATION using pandas
# 3. Bar Plot of Purchase Counts by Company

plt.figure(figsize=(10, 6))
data['Company'].value_counts().head(10).plot(kind='bar', color='pink')
plt.title('Top 10 Companies by Purchase Count')
plt.xlabel('Company')
plt.ylabel('Purchase Count')
plt.xticks(rotation=45, ha='right')
plt.grid(axis='y')
plt.show()
```



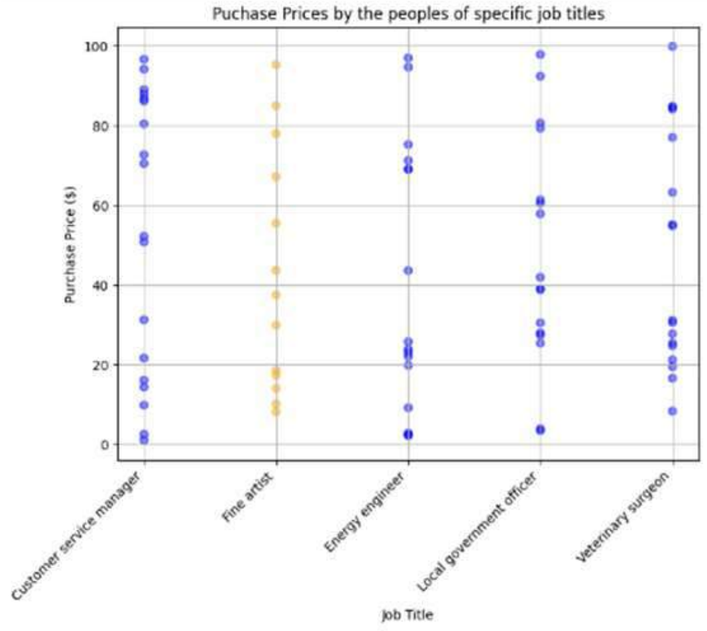
Scatter Plot of Purchase Prices by Specific Job Titles

- The scatter plot helps us visualize the relationship between purchase prices and specific job titles.
- It contributes to the analysis by identifying any patterns or trends in purchasing behavior based on occupation.

```
# DATA VISUALIZATION using pandas
# 4.Puchase Prices by the peoples of specific job titles

job_title1 = "Fine artist"
job_title2 = "Energy engineer"
job_title3 = "Local government officer"
job_title4 = "Veterinary surgeon"
job_title5 = "Customer service manager"


data_job_specific = data[data['Job'].isin([job_title1, job_title2,job_title3,job_title4,job_title5])]
plt.figure(figsize=(8, 6))
colors = ['orange' if job == job_title1 else 'blue' for job in data_job_specific['Job']]
plt.scatter(data_job_specific['Job'], data_job_specific['Purchase Price'], alpha=0.5, color=colors)
plt.title('Puchase Prices by the peoples of specific job titles')
plt.xlabel('Job Title')
plt.ylabel('Purchase Price ($)')
plt.xticks(rotation=45, ha='right')
plt.grid(True)
plt.show()
```



Conclusion

- The analysis of the 'E-Commerce Purchase' dataset using Pandas has provided valuable insights into customer behavior and trends within then e-commerce platform. Through data cleaning, exploratory data analysis (EDA), and data visualization, we were able to uncover key patterns and trends that can inform business decisions and strategies.

- In conclusion, the analysis of the 'E-Commerce Purchase' dataset has demonstrated the power of data analysis and visualization in extracting meaningful insights from complex datasets. By leveraging these insights, businesses can make informed decisions that drive growth and enhance the customer experience.

Thank you very much!

pandas