

## DATA PREPROCESSING

### DATAFRAME CREATION:

```
#creating data frame from the csv file
import pandas as pd
df=pd.read_csv(r"C:\\Users\\Nithesh\\Downloads\\shopping_trends.csv")
df
```

### OUTPUT:

Home Page - Select...

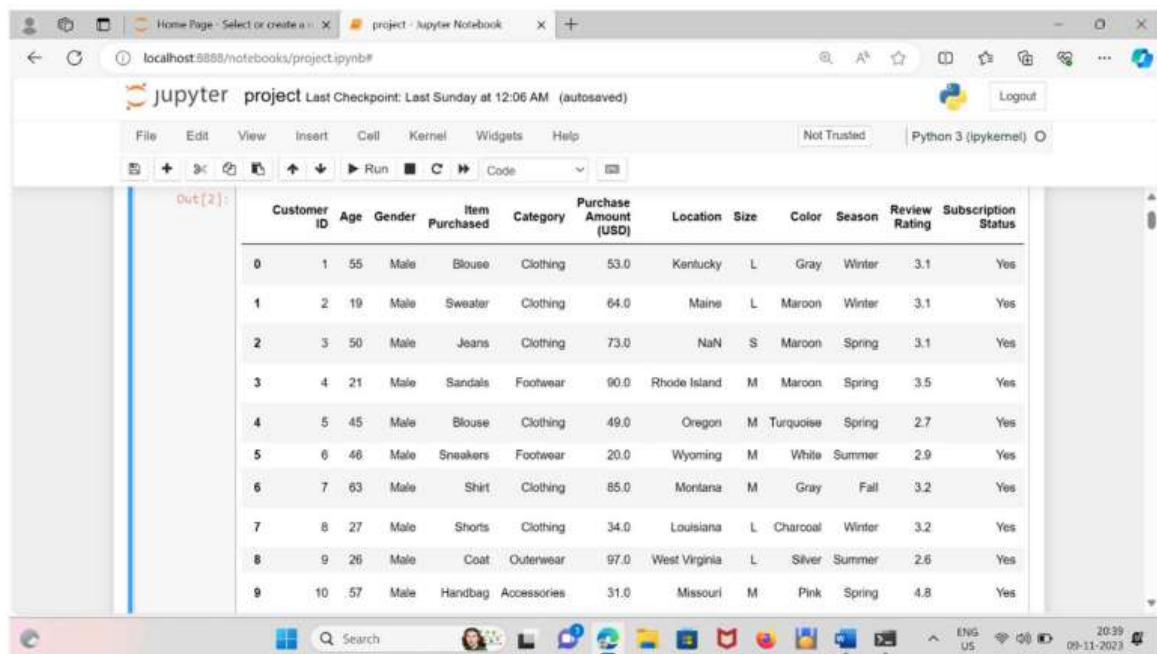
project - Jupyter Note...

Microsoft Edge | Whur...

pdf.pdf

Data Frame of Shopping trends dataset

## Data Frame of Shopping trends dataset



The screenshot displays a Jupyter Notebook window in a web browser. The browser's address bar shows 'localhost:8888/notebooks/project.ipynb#'. The Jupyter interface includes a top bar with the project name 'project', a 'Last Checkpoint' timestamp, and a 'Logout' button. Below this is a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations, running cells, and code execution. The main area shows a code cell with the output 'Out[2]:' followed by a DataFrame table. The table has 14 columns: Customer ID, Age, Gender, Item Purchased, Category, Purchase Amount (USD), Location, Size, Color, Season, Review Rating, and Subscription Status. It contains 10 rows of data, indexed 0 to 9. The Windows taskbar at the bottom shows the search bar, several application icons, and the system clock indicating 20:39 on 09-11-2023.

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status
0	1	55	Male	Blouse	Clothing	53.0	Kentucky	L	Gray	Winter	3.1	Yes
1	2	19	Male	Sweater	Clothing	64.0	Maine	L	Maroon	Winter	3.1	Yes
2	3	50	Male	Jeans	Clothing	73.0	NaN	S	Maroon	Spring	3.1	Yes
3	4	21	Male	Sandals	Footwear	90.0	Rhode Island	M	Maroon	Spring	3.5	Yes
4	5	45	Male	Blouse	Clothing	49.0	Oregon	M	Turquoise	Spring	2.7	Yes
5	6	46	Male	Sneakers	Footwear	20.0	Wyoming	M	White	Summer	2.9	Yes
6	7	63	Male	Shirt	Clothing	85.0	Montana	M	Gray	Fall	3.2	Yes
7	8	27	Male	Shorts	Clothing	34.0	Louisiana	L	Charcoal	Winter	3.2	Yes
8	9	26	Male	Coat	Outerwear	97.0	West Virginia	L	Silver	Summer	2.6	Yes
9	10	57	Male	Handbag	Accessories	31.0	Missouri	M	Pink	Spring	4.8	Yes

## DATA PREPROCESSING:

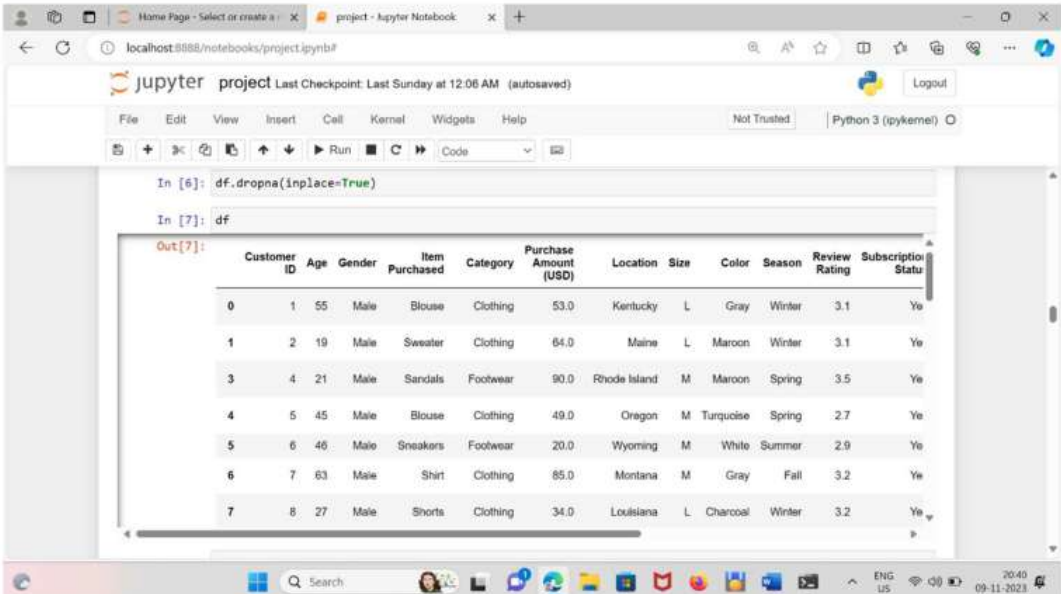
#Handling Missing Values

```
import pandas as pd
```

```
df=pd.read_csv(r"C:\\Users\\Nithesh\\Downloads\\shopping_trends.csv")
```

```
df.dropna(inplace=True)
```

**OUTPUT:**



The screenshot shows a Jupyter Notebook window with the following code and output:

```
In [6]: df.dropna(inplace=True)
```

```
In [7]: df
```

Out[7]:

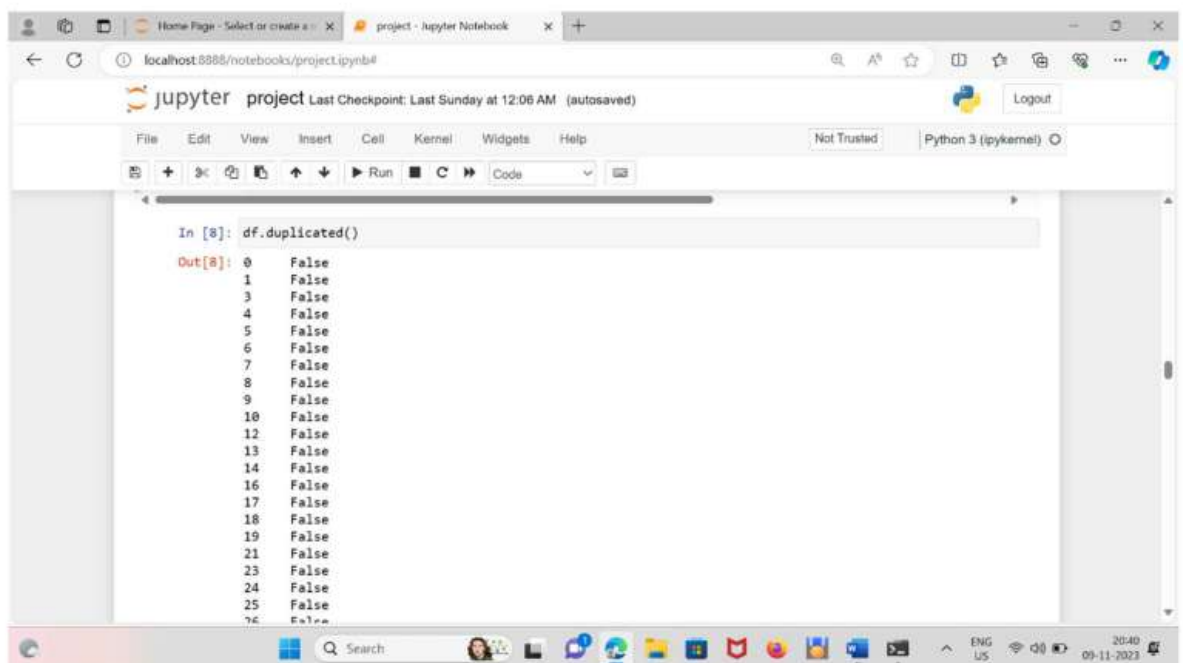
	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status
0	1	55	Male	Blouse	Clothing	53.0	Kentucky	L	Gray	Winter	3.1	Ye
1	2	19	Male	Sweater	Clothing	64.0	Maine	L	Maroon	Winter	3.1	Ye
3	4	21	Male	Sandals	Footwear	90.0	Rhode Island	M	Maroon	Spring	3.5	Ye
4	5	45	Male	Blouse	Clothing	49.0	Oregon	M	Turquoise	Spring	2.7	Ye
5	6	46	Male	Sneakers	Footwear	20.0	Wyoming	M	White	Summer	2.9	Ye
6	7	63	Male	Shirt	Clothing	85.0	Montana	M	Gray	Fall	3.2	Ye
7	8	27	Male	Shorts	Clothing	34.0	Louisiana	L	Charcoal	Winter	3.2	Ye

Handling Missing values

#Identifying the duplicates in a entire dataframe

```
df.duplicated()
```

**OUTPUT:**



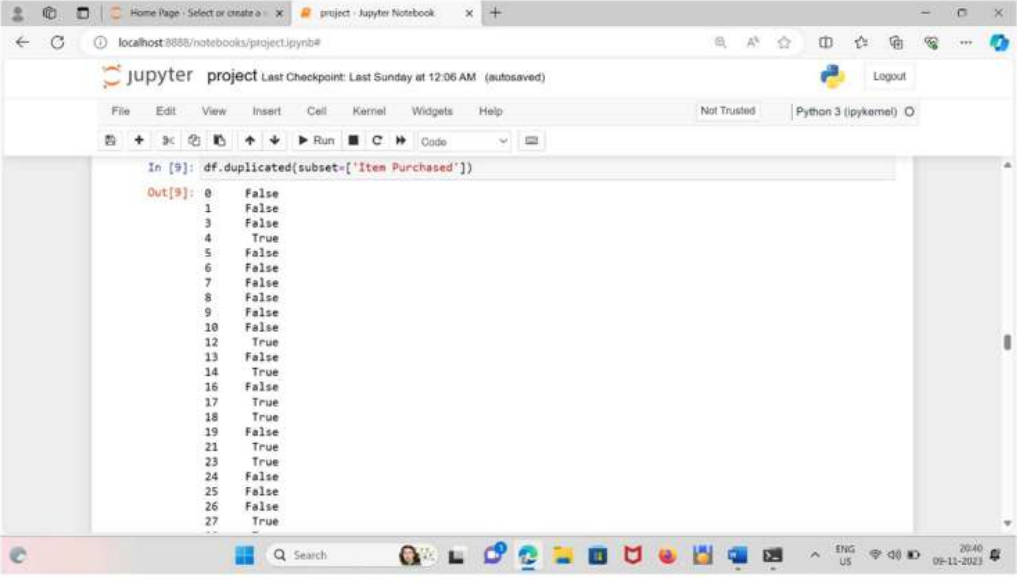
```
In [8]: df.duplicated()
Out[8]: 0    False
        1    False
        3    False
        4    False
        5    False
        6    False
        7    False
        8    False
        9    False
       10    False
       12    False
       13    False
       14    False
       16    False
       17    False
       18    False
       19    False
       21    False
       23    False
       24    False
       25    False
       26    False
```

Duplicates in a Data Frame

#Identifying the duplicates in a particular column

```
df.duplicated(subset=['Item Purchased'])
```

**OUTPUT:**

A screenshot of a Jupyter Notebook interface. The browser address bar shows 'localhost:8888/notebooks/project1jpyrn2#'. The notebook title is 'project'. The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for running, saving, and other actions. The code cell contains the command 'In [9]: df.duplicated(subset=['Item Purchased'])'. The output cell shows a series of boolean values for each row index from 0 to 27. The values are: 0: False, 1: False, 3: False, 4: True, 5: False, 6: False, 7: False, 8: False, 9: False, 10: False, 12: True, 13: False, 14: True, 16: False, 17: True, 18: True, 19: False, 21: True, 23: True, 24: False, 25: False, 26: False, 27: True. The Windows taskbar is visible at the bottom with the search bar and various application icons.

```
Out[9]: 0    False
        1    False
        3    False
        4     True
        5    False
        6    False
        7    False
        8    False
        9    False
       10    False
       12     True
       13    False
       14     True
       16    False
       17     True
       18     True
       19    False
       21     True
       23     True
       24    False
       25    False
       26    False
       27     True
```

Duplicates in a particular column

#Removing the duplicates in a particular column

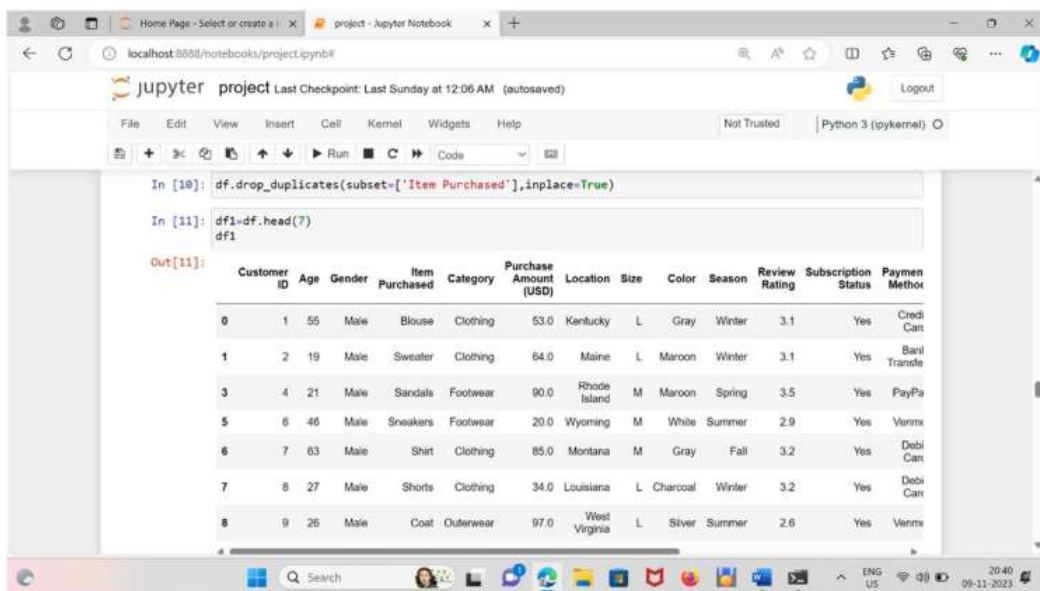
```
df.drop_duplicates(subset=['Item Purchased'],inplace=True)
```

#Accessing the top elements in a data frame using head

```
df1=df.head(7)
```

```
df1
```

## OUTPUT:



The screenshot shows a Jupyter Notebook interface with the following code and output:

```
In [10]: df.drop_duplicates(subset=['Item Purchased'],inplace=True)
```

```
In [11]: df1=df.head(7)
```

The output of the second cell is a DataFrame with 7 rows and 13 columns. The columns are: Customer ID, Age, Gender, Item Purchased, Category, Purchase Amount (USD), Location, Size, Color, Season, Review Rating, Subscription Status, and Payment Method.

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Payment Method
0	1	55	Male	Blouse	Clothing	53.0	Kentucky	L	Gray	Winter	3.1	Yes	Credit Card
1	2	19	Male	Sweater	Clothing	64.0	Maine	L	Maroon	Winter	3.1	Yes	Bank Transfer
3	4	21	Male	Sandals	Footwear	90.0	Rhode Island	M	Maroon	Spring	3.5	Yes	PayPal
5	6	46	Male	Sneakers	Footwear	20.0	Wyoming	M	White	Summer	2.9	Yes	Venmo
6	7	63	Male	Shirt	Clothing	85.0	Montana	M	Gray	Fall	3.2	Yes	Debit Card
7	8	27	Male	Shorts	Clothing	34.0	Louisiana	L	Charcoal	Winter	3.2	Yes	Debit Card
8	9	26	Male	Coat	Outerwear	97.0	West Virginia	L	Silver	Summer	2.6	Yes	Venmo

Accessing top elements using head

#Accessing the bottom elements in a data frame using tail

df.tail()

OUTPUT:

Home Page - Select or create a...

project - Jupyter Notebook

+

localhost:8888/notebooks/project.ipynb#

jupyter project Last Checkpoint: Last Sunday at 12:06 AM (autosaved)

Logout

FileEditViewInsertCellKernelWidgetsHelp

Not TrustedPython 3 (pykernel)

+

↶

↷

↺

↻

↱

↲

↴

↵

↶

↷

↺

↻

↱

↲

↴

↵

Code

In [12]: df.tail()

Out[12]:

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Payment Method
34	35	36	Male	T-shirt	Clothing	91.0	North Dakota	L	Violet	Spring	4.6	Yes	D C
37	38	35	Male	Jeans	Clothing	45.0	Indiana	S	Cyan	Summer	2.8	Yes	D C
41	42	67	Male	Scarf	Accessories	39.0	Alaska	M	Orange	Spring	4.5	Yes	C
44	45	39	Male	Hat	Accessories	53.0	Kentucky	S	Silver	Summer	4.8	Yes	B Trans
45	46	50	Male	Socks	Clothing	21.0	Tennessee	XL	Indigo	Fall	2.9	Yes	Ver

In [13]: import matplotlib.pyplot as plt  
x = df1['Item Purchased']  
y = df1['Purchase Amount (USD)']  
plt.xlabel('Item Purchased')

20:40 09-11-2023

ENG US

Accessing bottom elements using tail