Dastgyr Technologies

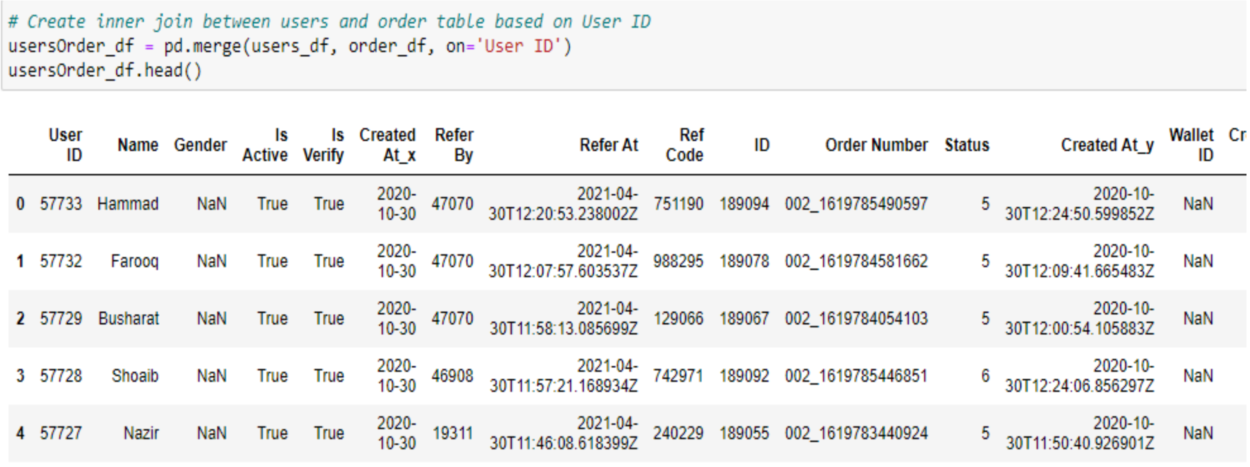
Data Analysis Case Study

Q1.

Ans. Create inner join Orders placed by customers through referral code of the agent

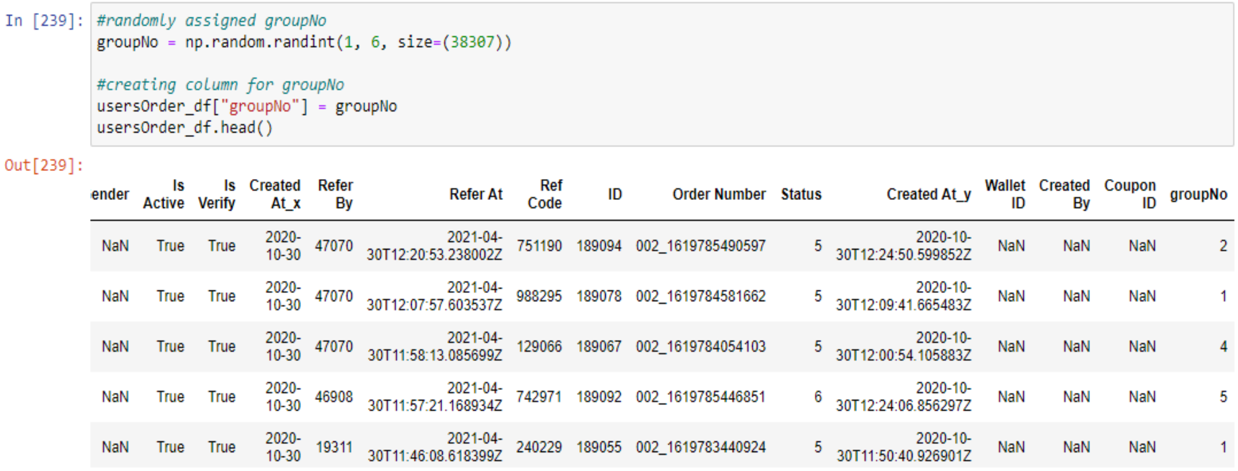
- Make 5 groups
- Identify Leader of the group

1. Create inner join between users and order table based on user id. Randomly assigned group number from 1 to 5.



```
# Create inner join between users and order table based on User ID
usersOrder_df = pd.merge(users_df, order_df, on='User ID')
usersOrder_df.head()
```

	User ID	Name	Gender	Is Active	Is Verify	Created At_x	Refer By	Refer At	Ref Code	ID	Order Number	Status	Created At_y	Wallet ID	Cr
0	57733	Hammad	NaN	True	True	2020-10-30	47070	2021-04-30T12:20:53.238002Z	751190	189094	002_1619785490597	5	2020-10-30T12:24:50.599852Z	NaN	
1	57732	Farooq	NaN	True	True	2020-10-30	47070	2021-04-30T12:07:57.603537Z	988295	189078	002_1619784581662	5	2020-10-30T12:09:41.665483Z	NaN	
2	57729	Busharat	NaN	True	True	2020-10-30	47070	2021-04-30T11:58:13.085699Z	129066	189067	002_1619784054103	5	2020-10-30T12:00:54.105883Z	NaN	
3	57728	Shoaib	NaN	True	True	2020-10-30	46908	2021-04-30T11:57:21.168934Z	742971	189092	002_1619785446851	6	2020-10-30T12:24:06.856297Z	NaN	
4	57727	Nazir	NaN	True	True	2020-10-30	19311	2021-04-30T11:46:08.618399Z	240229	189055	002_1619783440924	5	2020-10-30T11:50:40.926901Z	NaN	

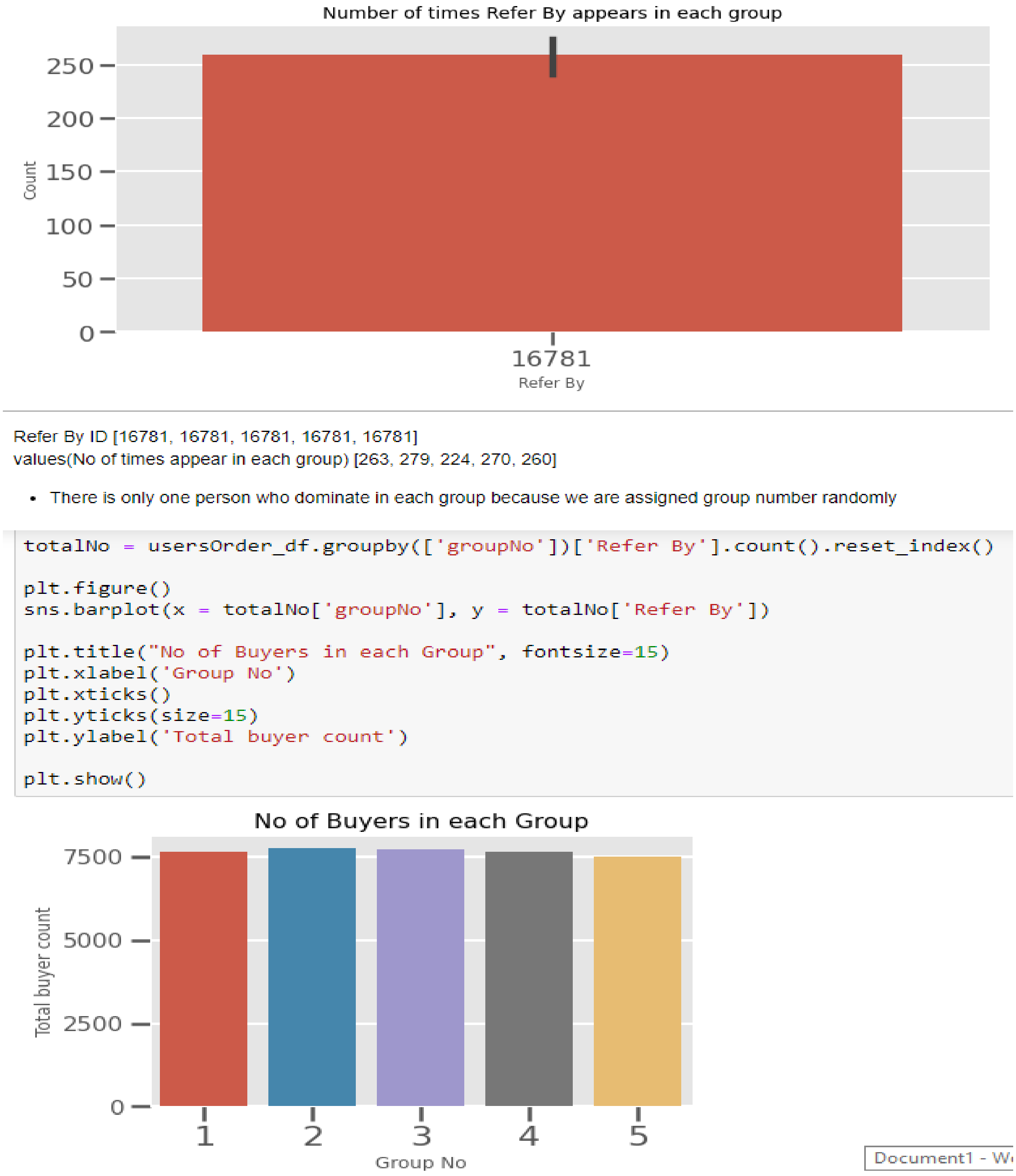


```
In [239]: #randomly assigned groupNo
          groupNo = np.random.randint(1, 6, size=(38307))

          #creating column for groupNo
          usersOrder_df["groupNo"] = groupNo
          usersOrder_df.head()
```

Out[239]:

ender	Is Active	Is Verify	Created At_x	Refer By	Refer At	Ref Code	ID	Order Number	Status	Created At_y	Wallet ID	Created By	Coupon ID	groupNo
NaN	True	True	2020-10-30	47070	2021-04-30T12:20:53.238002Z	751190	189094	002_1619785490597	5	2020-10-30T12:24:50.599852Z	NaN	NaN	NaN	2
NaN	True	True	2020-10-30	47070	2021-04-30T12:07:57.603537Z	988295	189078	002_1619784581662	5	2020-10-30T12:09:41.665483Z	NaN	NaN	NaN	1
NaN	True	True	2020-10-30	47070	2021-04-30T11:58:13.085699Z	129066	189067	002_1619784054103	5	2020-10-30T12:00:54.105883Z	NaN	NaN	NaN	4
NaN	True	True	2020-10-30	46908	2021-04-30T11:57:21.168934Z	742971	189092	002_1619785446851	6	2020-10-30T12:24:06.856297Z	NaN	NaN	NaN	5
NaN	True	True	2020-10-30	19311	2021-04-30T11:46:08.618399Z	240229	189055	002_1619783440924	5	2020-10-30T11:50:40.926901Z	NaN	NaN	NaN	1

2.Please see the picture description

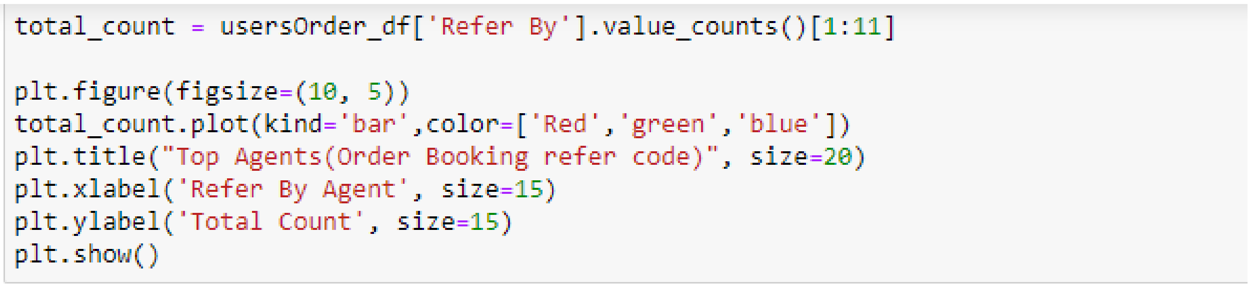


Refer By ID [16781, 16781, 16781, 16781, 16781]
values(No of times appear in each group) [263, 279, 224, 270, 260]

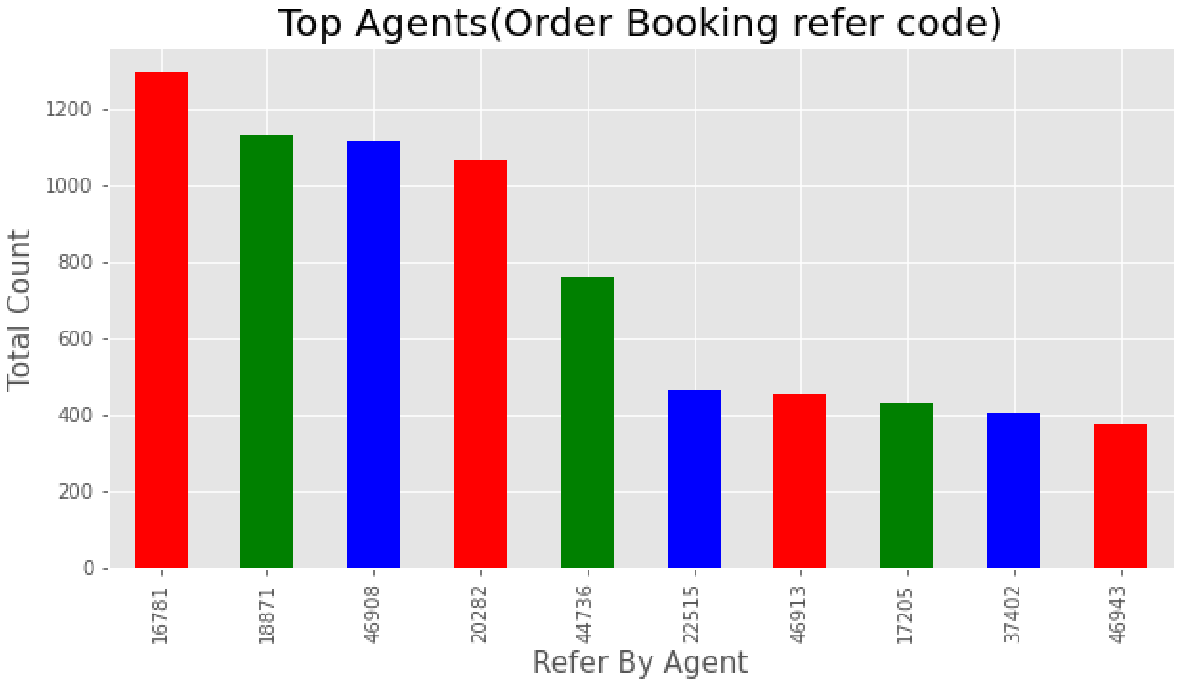
- There is only one person who dominate in each group because we are assigned group number randomly

```
totalNo = usersOrder_df.groupby(['groupNo'])['Refer By'].count().reset_index()

plt.figure()
sns.barplot(x = totalNo['groupNo'], y = totalNo['Refer By'])

plt.title("No of Buyers in each Group", fontsize=15)
plt.xlabel('Group No')
plt.xticks()
plt.yticks(size=15)
plt.ylabel('Total buyer count')

plt.show()
```

TOP 10 Agents (Whose refer code is mostly used in order booking)

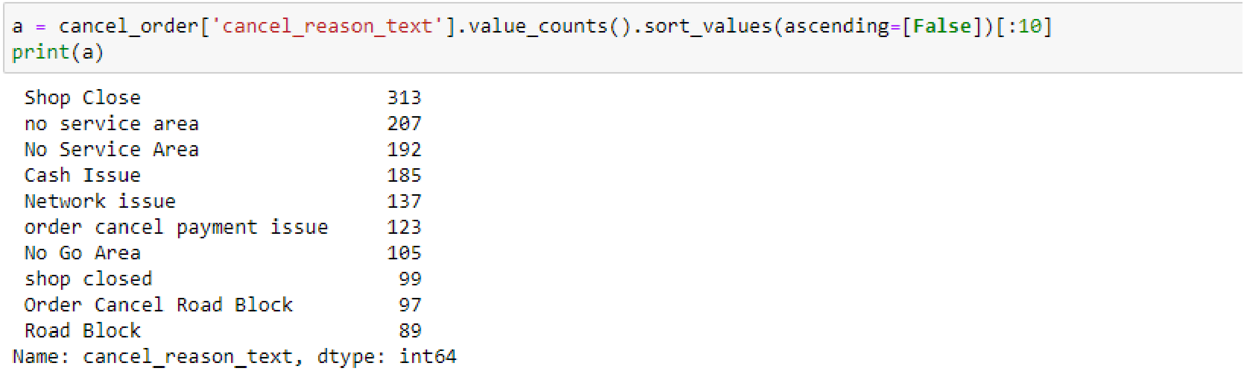


```
total_count = usersOrder_df['Refer By'].value_counts()[1:11]

plt.figure(figsize=(10, 5))
total_count.plot(kind='bar',color=['Red','green','blue'])
plt.title("Top Agents(Order Booking refer code)", size=20)
plt.xlabel('Refer By Agent', size=15)
plt.ylabel('Total Count', size=15)
plt.show()
```



Q2.

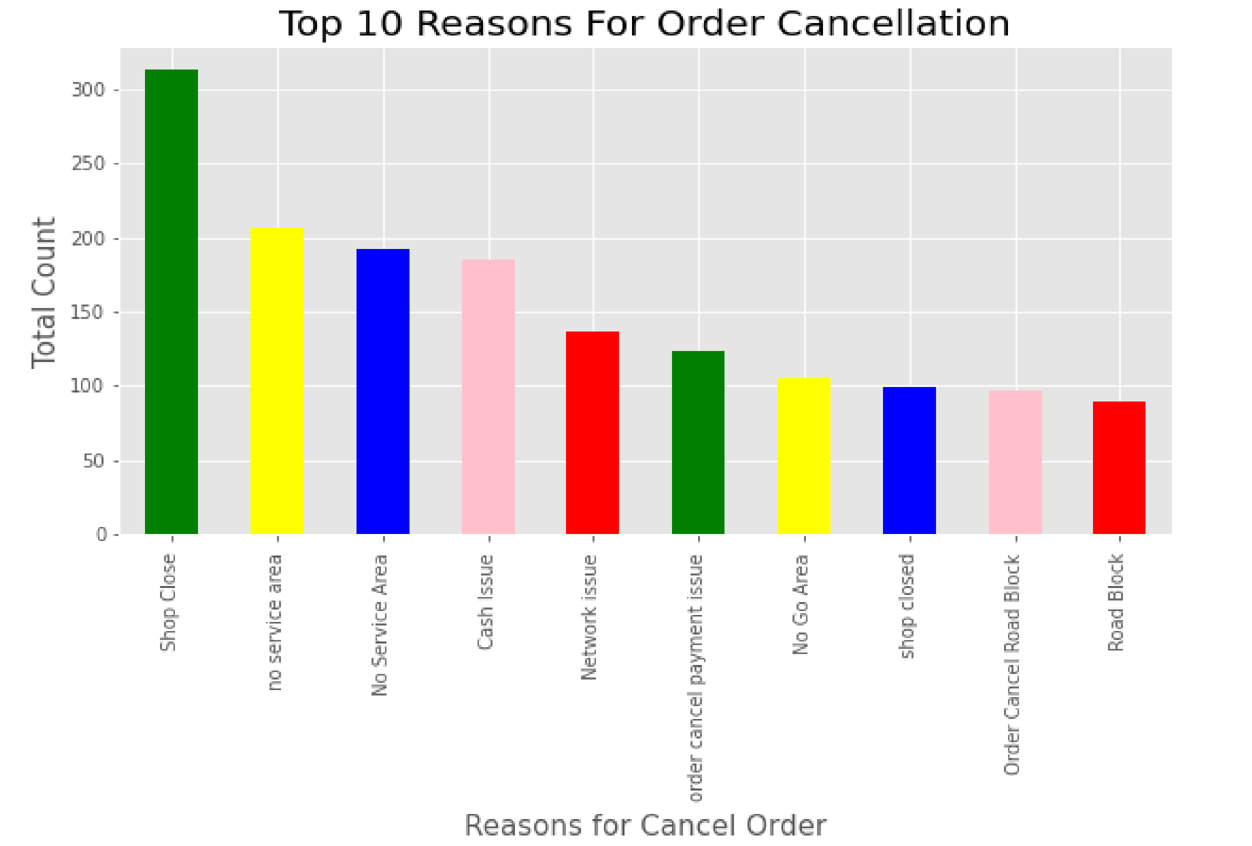
Ans. Frist we will get the reasons why orders are getting cancelled and then calculate how many order have been cancelled based on cancellation reasons.



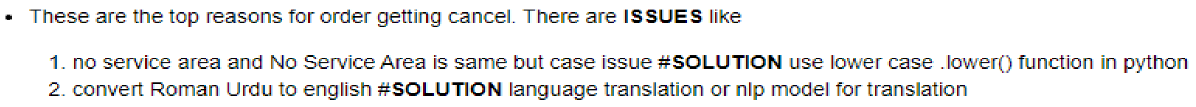
```
a = cancel_order['cancel_reason_text'].value_counts().sort_values(ascending=[False])[:10]
print(a)
```

```
 Shop Close                     313
 no service area                207
 No Service Area                192
 Cash Issue                     185
 Network issue                  137
 order cancel payment issue     123
 No Go Area                     105
 shop closed                     99
 Order Cancel Road Block         97
 Road Block                      89
Name: cancel_reason_text, dtype: int64
```

1.These are the top 10 reasons why orders are cancelled



Issues



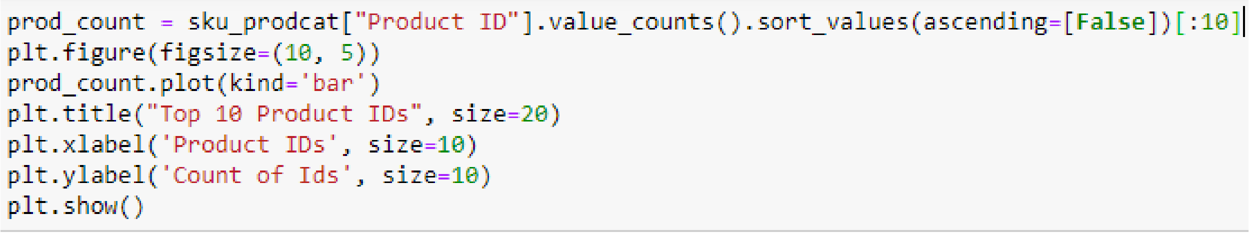
- These are the top reasons for order getting cancel. There are **ISSUES** like

  1. no service area and No Service Area is same but case issue **#SOLUTION** use lower case .lower() function in python
  2. convert Roman Urdu to english **#SOLUTION** language translation or nlp model for translation

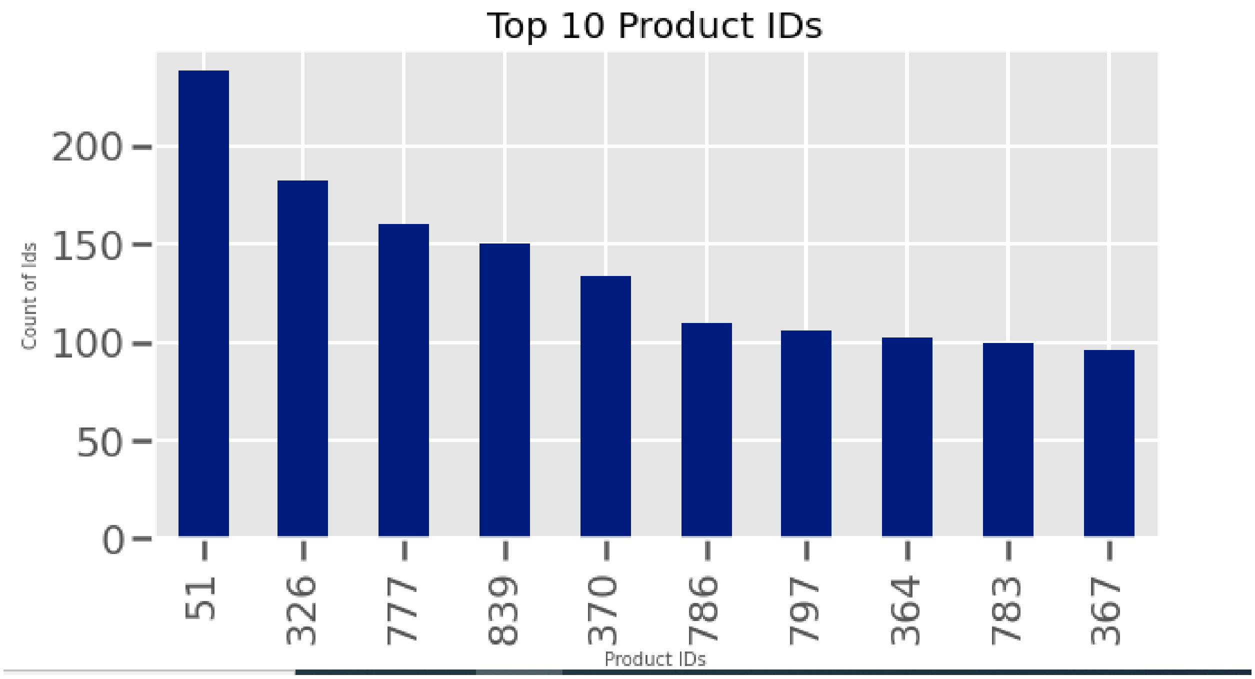
Q3.

Ans.

First we calculate and visualize the top 10 most buying products.

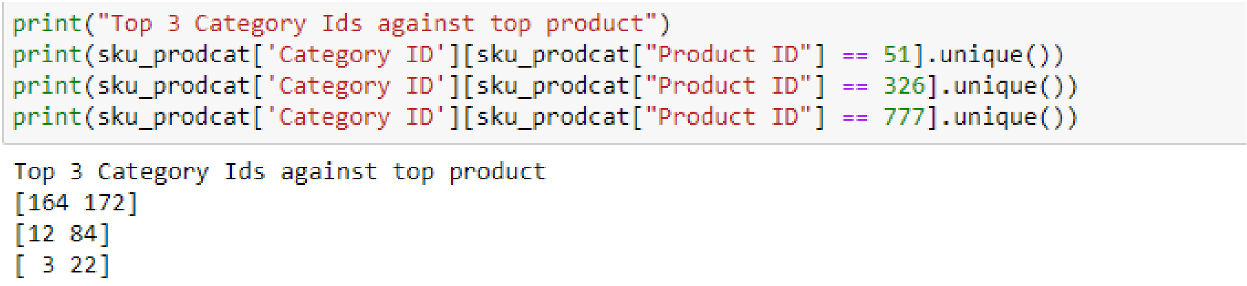


```
prod_count = sku_prodcat["Product ID"].value_counts().sort_values(ascending=[False])[:10]
plt.figure(figsize=(10, 5))
prod_count.plot(kind='bar')
plt.title("Top 10 Product IDs", size=20)
plt.xlabel('Product IDs', size=10)
plt.ylabel('Count of Ids', size=10)
plt.show()
```



2. **Get the most buying Product Categories (Top 3)**

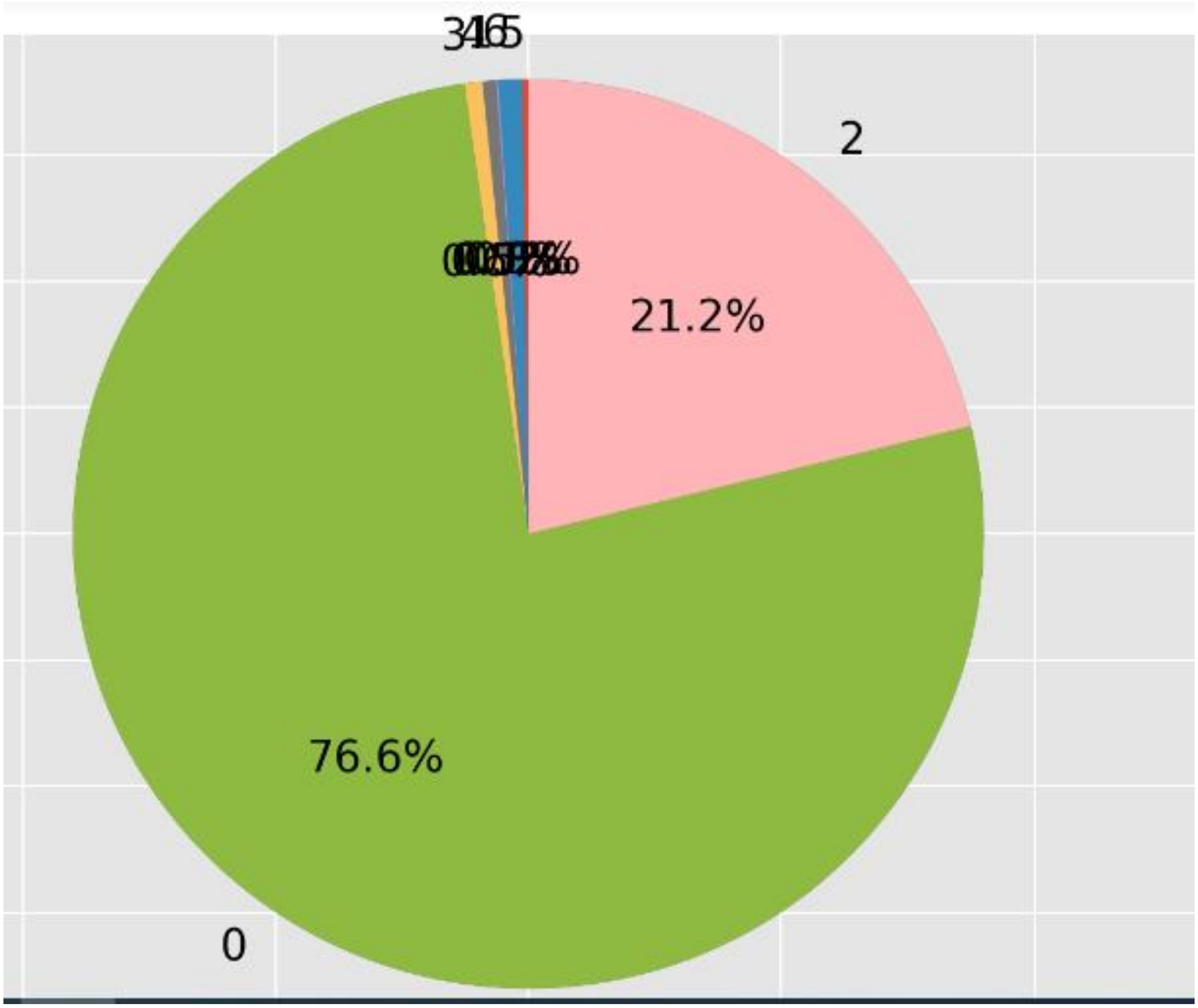
- Every product has two categories

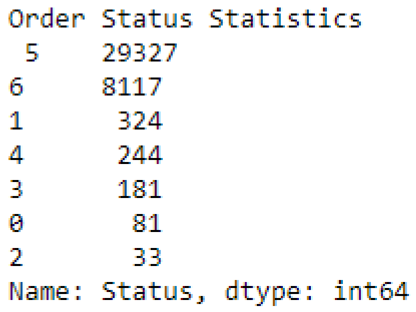


```
print("Top 3 Category Ids against top product")
print(sku_prodcat['Category ID'][sku_prodcat["Product ID"] == 51].unique())
print(sku_prodcat['Category ID'][sku_prodcat["Product ID"] == 326].unique())
print(sku_prodcat['Category ID'][sku_prodcat["Product ID"] == 777].unique())
```

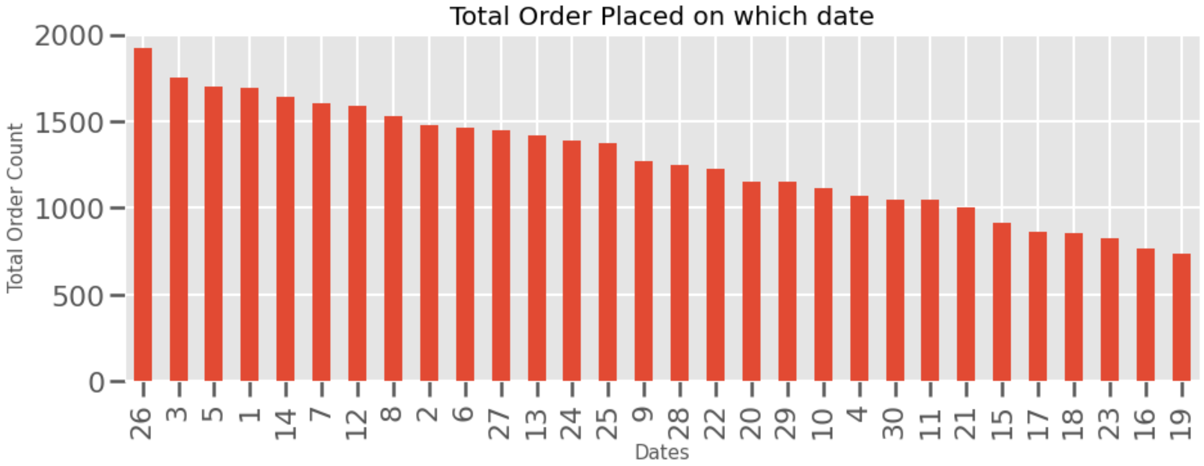
```
Top 3 Category Ids against top product
[164 172]
[12 84]
[ 3 22]
```

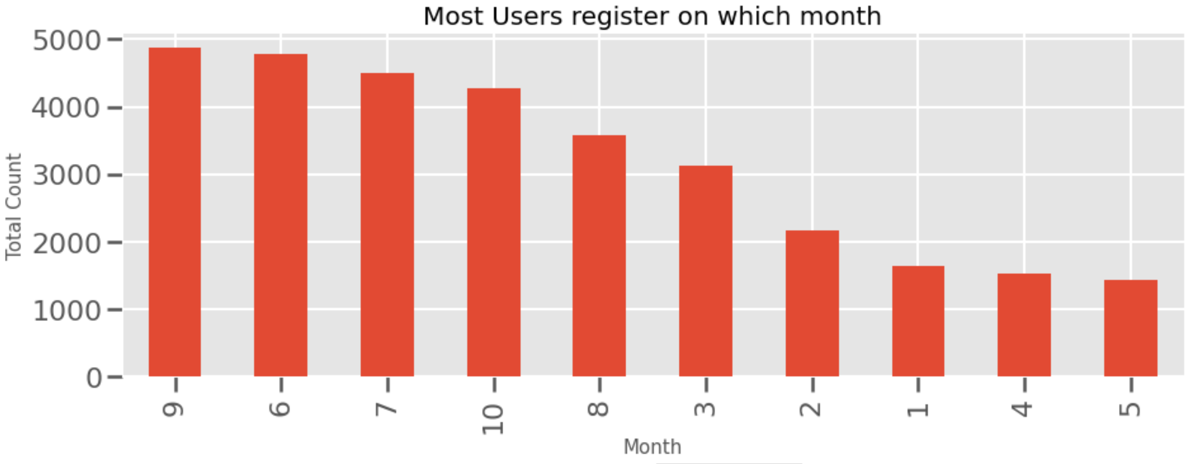
Extra work





```
Order Status Statistics
 5    29327
6     8117
1      324
4      244
3      181
0       81
2       33
Name: Status, dtype: int64
```





# You can find more extra EDA Here

Colab Notebook Link:

https://colab.research.google.com/drive/1MYcbJdQ-Wejnt48gd2dmgXnoB_nq6Auc?usp=sharing