Institute of Business Administration Karachi

Leadership and Ideas for Tomorrow

INTRO TO DATA MINING

(Class:5704, 5705, 5706, 5758 & 5837) - Fall 2021

Pakistan's Largest E-Commerce Dataset

(Class Project)

Instructor: Dr. Sajjad Haider

Group Members:

Agha Muhammad Usman – 19750

Muhammad Saad Karim – 18565

Sarmed Ahmed Usmani – 19673

Syed Nameer Ali – 18606

BS – CS

Advice taken from Instructor about the dataset:

The dataset is great! Please go ahead. I would suggest that instead of solving all the problems, you do a quick analysis of the available solutions (against different questions) posted in the form of code and then pick one of them to analyze further.

Data Set:

Pakistan's Largest E-Commerce Dataset

Project Description:

E-commerce is the buying and selling of goods and services, or the transmitting of funds or data, over an electronic network, primarily the internet. It means business transactions through the internet, telephone, credit card, etc. without the help of a cheque or physical payment of money on the part of the buyer. The money is paid by the bank or company. It is the most modern method of transaction and is in practice in the developed countries of the world. E-commerce is in turn driven by the technological advances of the semiconductor industry and is the largest sector of the electronics industry.

The dataset we chose contains detailed information of half a million e-commerce orders in Pakistan from March 2016 to August 2018. It contains item details, shipping method, payment method like credit card, Easy-Paisa, Jazz-Cash, cash-on-delivery, product categories like fashion, mobile, electronics, appliance etc., date of order, SKU, price, quantity, total and customer ID. This is the most detailed dataset about e-commerce in Pakistan that you can find in the public domain.

Problem Statement:

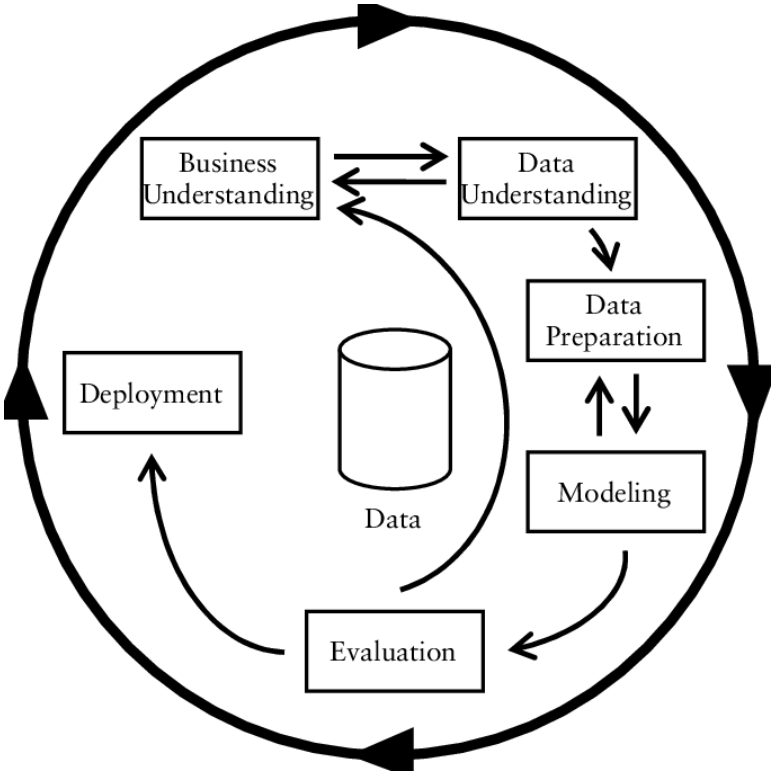
- What is the best-selling category?
    - Bestselling category by year(count).
    - Bestselling category by revenue.
    - Bestselling category by quantity ordered.
    - Bestselling categories by no. of orders
    - Bestselling categories after checking if order was completed.
    - Bestselling categories by payment methods.

Since the dataset is collected from multiple resources as a source for research study there is no set problem statement, however we will be doing a detailed analysis of the above points mentioned.

Objective:

By using python libraries and our understanding of CRISP-DM model during the course Intro to data mining. We will show data with different types of visualizations for the bestselling category analysis to be properly understood.

CRISP – DM Model:



We will follow the Crisp-DM Model to understand the flow of our entire Project. The first thing which follows in Crisp-DM is Business Understanding.

Business Understanding

As already stated above that we were required to analyze the bestselling category with respect to many different measures.

Data Understanding:

```
data.dtypes
✓ 0.1s
item_id                    float64
status                      object
created_at          datetime64[ns]
sku                         object
price                      float64
qty_ordered                float64
grand_total                float64
increment_id                object
category_name_1             object
sales_commission_code       object
discount_amount            float64
payment_method              object
working_date                object
bi_status                   object
_mv_                        object
year                       float64
month                      float64
customer_since              object
m-y                         object
fy                          object
customer_id                float64
dtype: object
```

There are 26 columns in the dataset. Names of the column are mentioned below:

- Item_id : Primary key for item
- Status: complete/ canceled/ order refunded/ received/ refund/closed/ fraud/ holded/ exchange/ pending PayPal/ paid/ cod/ pending/ processing/ payment_review
- Created_at : Date when item was enlisted
- Sku : name / description of product
- Price: price of product
- Qty_order : no. of these items ordered
- Grand_total : price * quantity
- Increment_id : ID
- Category_name_1: Women's Fashion / Beauty & Grooming / Soghaat / Mobiles & Tablets / Appliances / Home & Living / Men's Fashion / Kids & Baby / Others / Entertainment / Computing / Superstore / Health & Sports /Books / School & Education
- Sales_commission_code : Referral code

- Discount_amount: amount discounted
- Payment_method: method of Payment
- Working Date: same as created date
- BI Status: unknown
- MV: unknown
- Year: Year of order
- Month: Month of Order
- Customer Since: First order of customer
- M-Y: Month - Year
- FY: unknown
- Customer ID: ID of customer that placed the order
- Unnamed: null
- Unnamed: null
- Unnamed: null
- Unnamed: null
- Unnamed: null

Data Preparation:

Now we have seen what our Data is like and what steps needed to be taken while Data Preparation. Let us move ahead with our Crisp-DM Cycle and prepare the data based on the insights we have gathered.

Initially there are total of 26 columns with 1048575 entries.

```
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 26 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   item_id                584524 non-null  float64
 1   status                 584509 non-null  object
 2   created_at             584524 non-null  object
 3   sku                    584504 non-null  object
 4   price                  584524 non-null  float64
 5   qty_ordered            584524 non-null  float64
 6   grand_total            584524 non-null  float64
 7   increment_id           584524 non-null  object
 8   category_name_1        584360 non-null  object
 9   sales_commission_code  447349 non-null  object
 10  discount_amount        584524 non-null  float64
 11  payment_method         584524 non-null  object
 12  Working Date           584524 non-null  object
 13  BI Status              584524 non-null  object
 14   MV                    584524 non-null  object
 15  Year                   584524 non-null  float64
 16  Month                  584524 non-null  float64
 17  Customer Since         584513 non-null  object
 18  M-Y                    584524 non-null  object
 19  FY                     584524 non-null  object
 20  Customer ID            584513 non-null  float64
 21  Unnamed: 21            0 non-null       float64
 22  Unnamed: 22            0 non-null       float64
 23  Unnamed: 23            0 non-null       float64
 24  Unnamed: 24            0 non-null       float64
 25  Unnamed: 25            0 non-null       float64
dtypes: float64(13), object(13)
```

After analyzing the data, it was found that the values of the last five columns were null therefore they are required to be dropped.

```
data = data.iloc[:, :-5]
```
✓ 0.2s

Also, any null entries in the data set are also dropped.

```
data=data.dropna(how='all')
```
✓ 0.9s

'MV' columns contains space in its name and should be removed to avoid any error.

```
data.rename(columns={' MV ':'MV'},inplace=True)
```

Data type of some columns are also needed to be changed.

```
data['item_id']=data['item_id'].astype(int)
data['Customer ID']=data['Customer ID'].astype(str)
data['qty_ordered']=data['qty_ordered'].astype(int)
data['Year']=data['Year'].astype(int)
data['Month']=data['Month'].astype(int)
```

✓ 0.5s

After the Data Cleaning process our data set is reduced to 21 columns with 584524 entries.

```
Int64Index: 584524 entries, 0 to 584523
Data columns (total 21 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   item_id               584524 non-null  int32
 1   status                584509 non-null  object
 2   created_at            584524 non-null  object
 3   sku                   584504 non-null  object
 4   price                 584524 non-null  float64
 5   qty_ordered           584524 non-null  int32
 6   grand_total           584524 non-null  float64
 7   increment_id          584524 non-null  object
 8   Category              584360 non-null  object
 9   sales_commission_code 447349 non-null  object
 10  discount_amount       584524 non-null  float64
 11  payment_method        584524 non-null  object
 12  Working Date          584524 non-null  object
 13  BI Status             584524 non-null  object
 14  MV                    584524 non-null  object
 15  Year                  584524 non-null  int32
 16  Month                 584524 non-null  int32
 17  Customer Since        584513 non-null  object
 18  M-Y                   584524 non-null  object
 19  FY                    584524 non-null  object
 20  Customer ID           584524 non-null  object
dtypes: float64(3), int32(4), object(14)
```

Now we are all set to pass our processed data for modeling and evaluation.

Model & Evaluation:

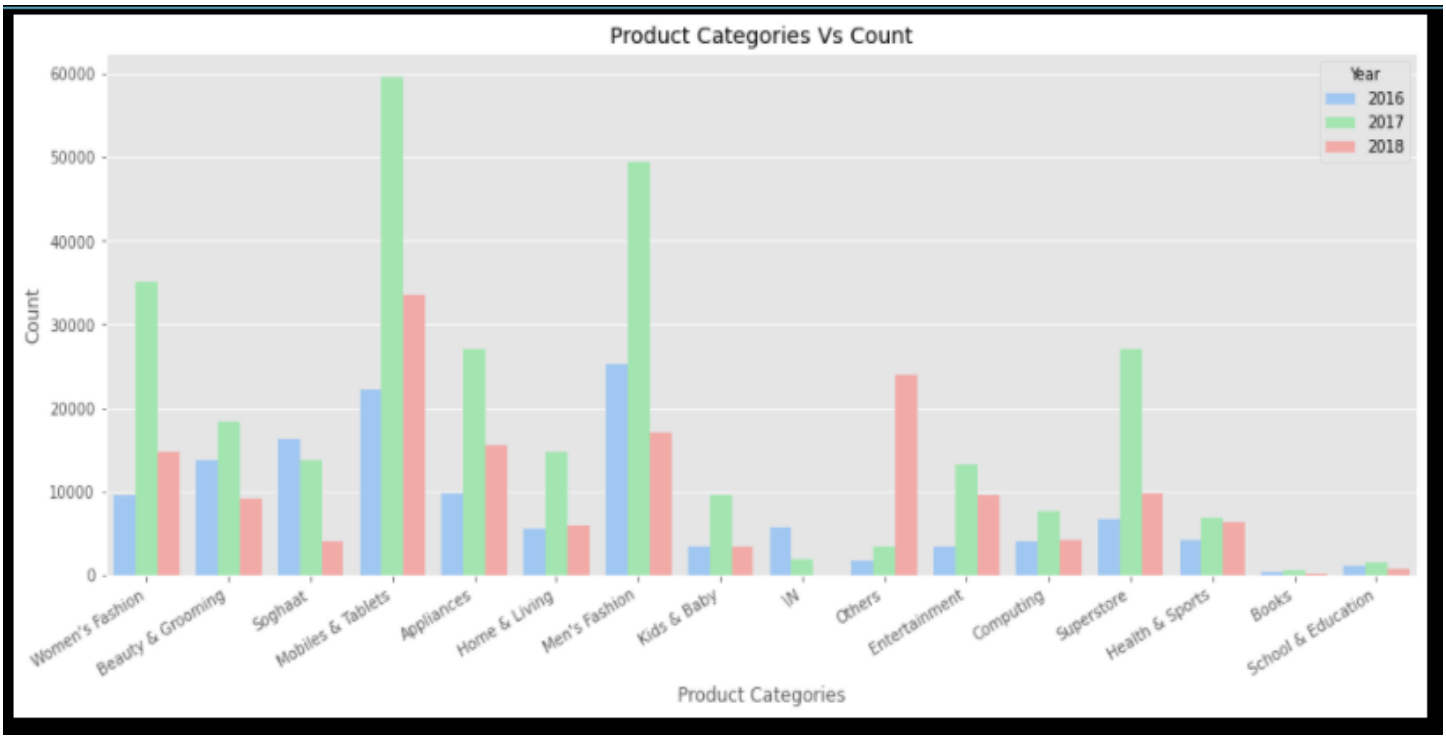
As we are predicting the bestselling category, we will be inspecting the target Variables (Classifier).

Bestselling category by yearly(count)

```
fig, ax = plt.subplots(figsize=(16, 6))

# add the plot
sns.countplot(x = 'category_name_1', data = df, hue = 'Year').set(title = 'Product Categories Vs Count')

# add Labels
ax.set(xlabel = 'Product Categories')
ax.set(ylabel = 'Count')
ax.set_xticklabels(ax.get_xticklabels(), rotation = 30, horizontalalignment = 'right')

plt.show()
```

[79] ✓ 0.8s

On inspecting Product categories vs count for the whole dataset, it was found that in 2016 Men's Fashion was the bestselling category while in 2017 and 2018 the bestselling category was Mobiles & Tablets. Count ranges from 0-60k. So, overall Best-Selling Category from 2016 to 2018 is: Mobiles & Tablets.

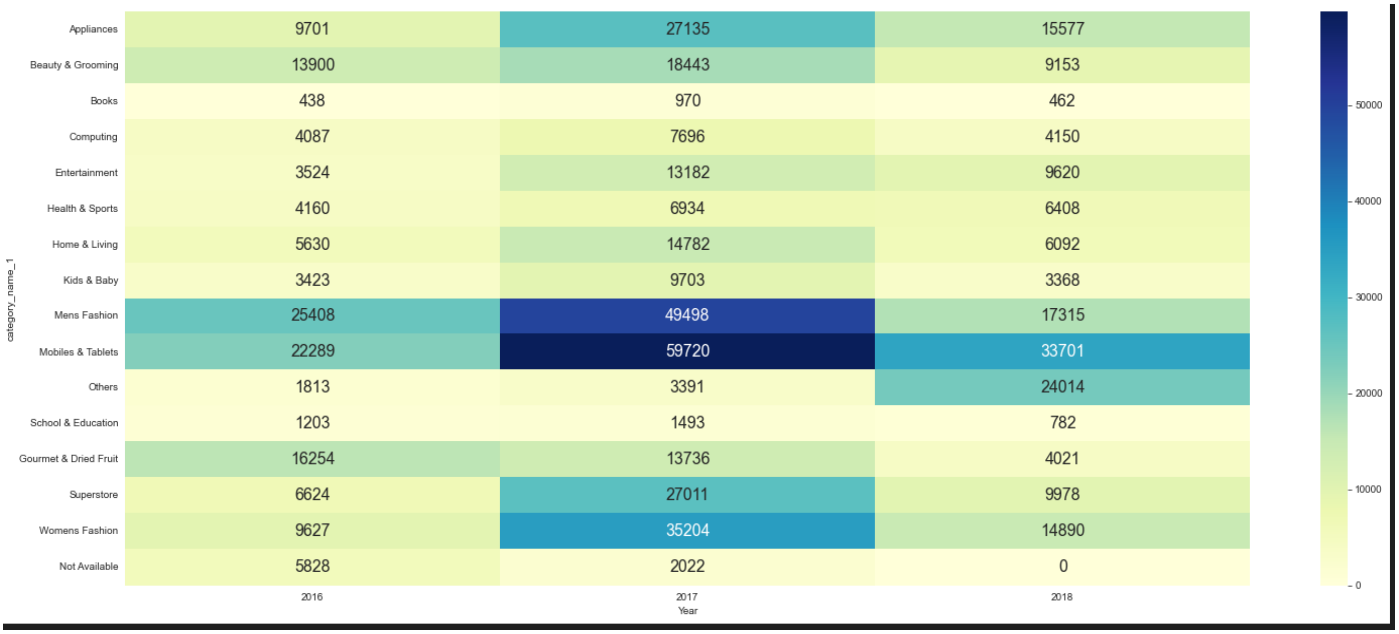


Heatmap of Categories group by years

```
heatmapData = (df
 .groupby('Year')
 .category_name_1
 .value_counts()
 .unstack()
 .fillna(0)
)

labels_x = ['2016','2017','2018']
labels_y = ['Appliances','Beauty & Grooming','Books','Computing','Entertainment','Health & Sports','Home & Living','Kids & Baby',
            'Mens Fashion','Mobiles & Tablets','Others','School & Education','Gourmet & Dried Fruit','Superstore','Womens Fashion','Not Available']

plt.figure(figsize=(25,10))

sns.heatmap(heatmapData.T,
            cmap = 'YlGnBu',
            annot = True,
            fmt = '.0f',
            center = 30000,
            linecolor = 'black',
            xticklabels = labels_x,
            yticklabels = labels_y,
           annot_kws ={'fontsize': 16,}
           )
```

✓ 0.5s

```
<AxesSubplot:xlabel='Year', ylabel='category_name_1'>
```



According to the diagrams above we can say:

Mobiles & Tablets are Best Selling category in Ecommerce.

Top 3 categories in 2016:

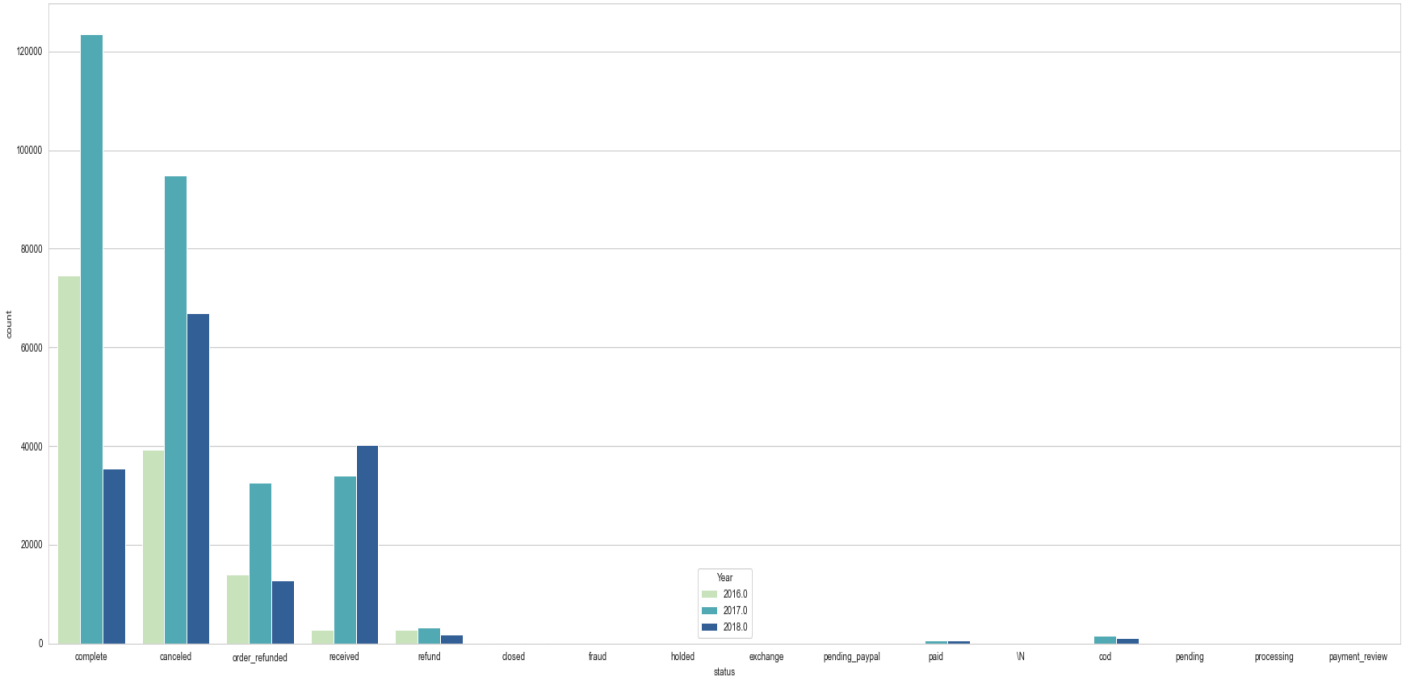
1. Men's Fashion
2. Mobiles & Tablets
3. Gourmet & Dried Fruit

Top 3 categories in 2017:

1. Mobiles & Tablets
2. Men's Fashion
3. Women's Fashion

Top 3 categories in 2018:

1. Mobiles & Tablets
2. Others
3. Men's Fashion

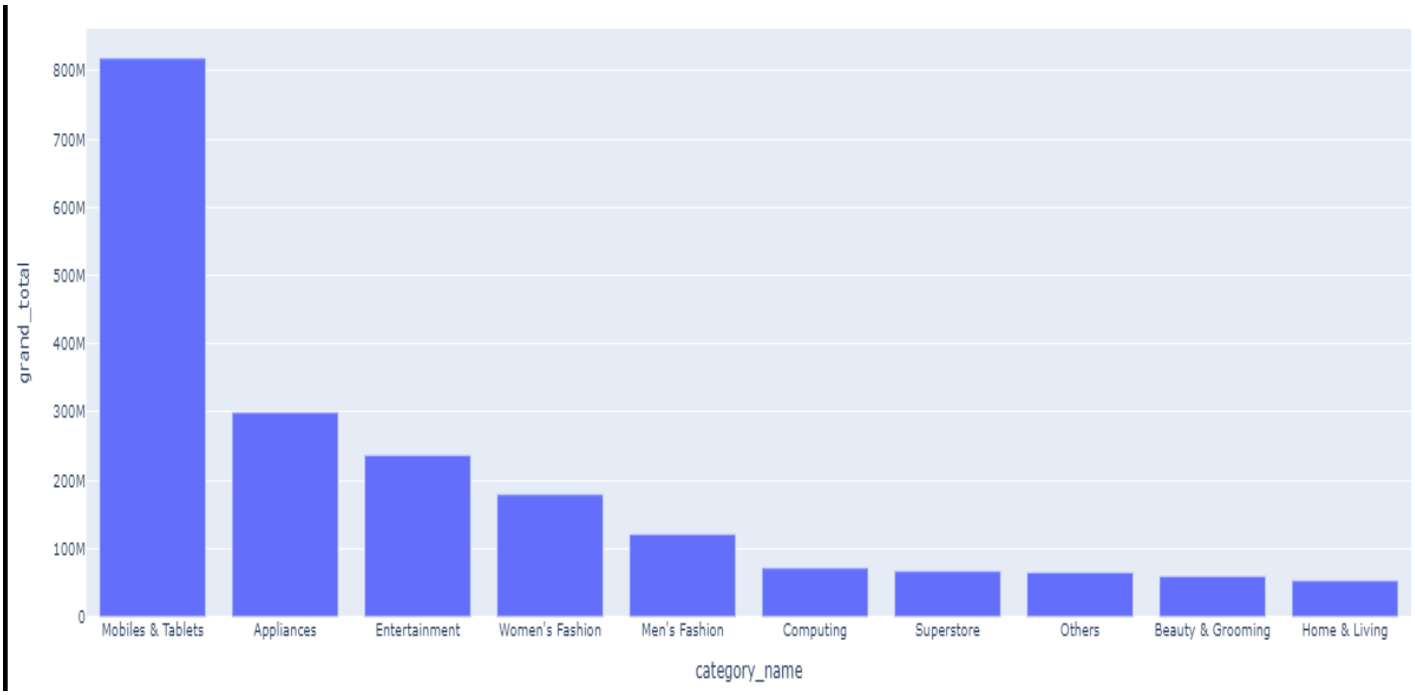


Best Selling category by revenue

```
# Best category by Grand Total
Bar_chart1 =py.bar(df3,x='category_name',y='grand_total')
Bar_chart1.show()
```

✓ 0.8s

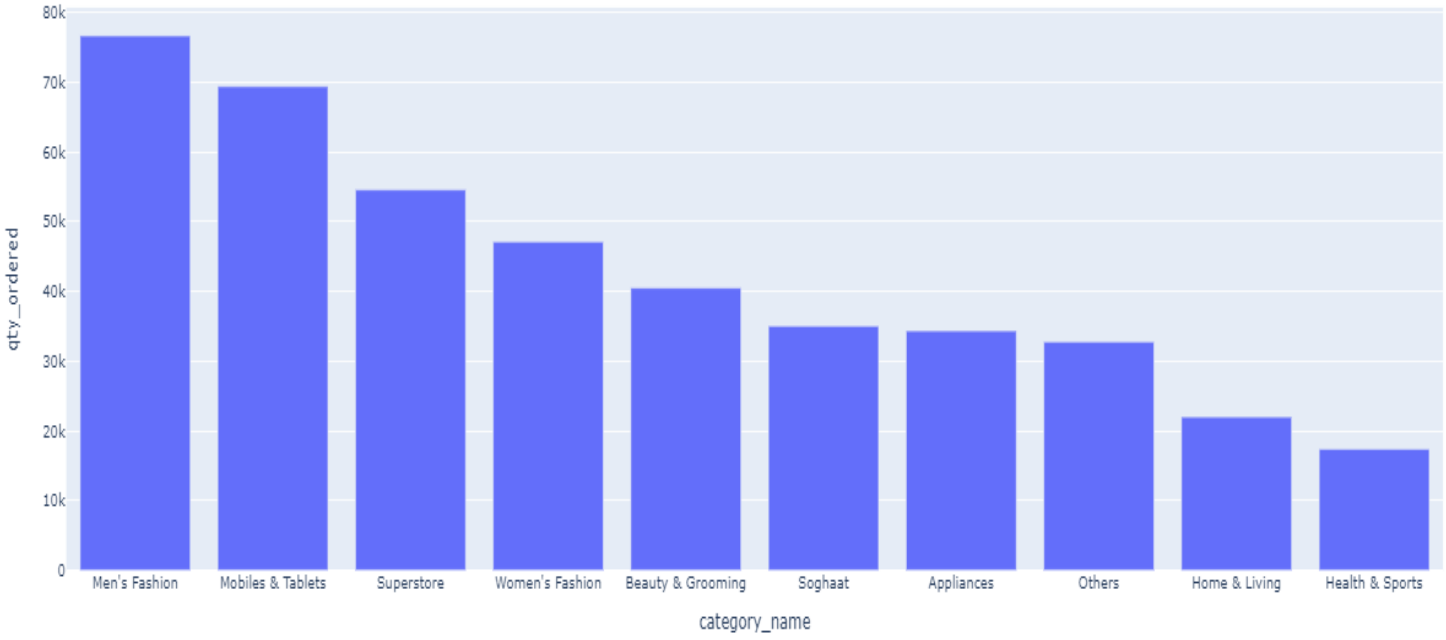
On inspecting Product categories vs Revenue (Grand Total) for the whole dataset, it was found that the bestselling category was Mobiles & Tablets by far. Count ranges from 0-800 million.



Best Selling category by quantity ordered

```
Bar_chart2=py.bar(df3,x='category_name',y='qty_ordered')
Bar_chart2.show()
✓ 0.1s
```

On inspecting Product categories vs Quantity ordered for the whole dataset, it was found that the bestselling category was Men's Fashion. Count ranges from 0-80k.

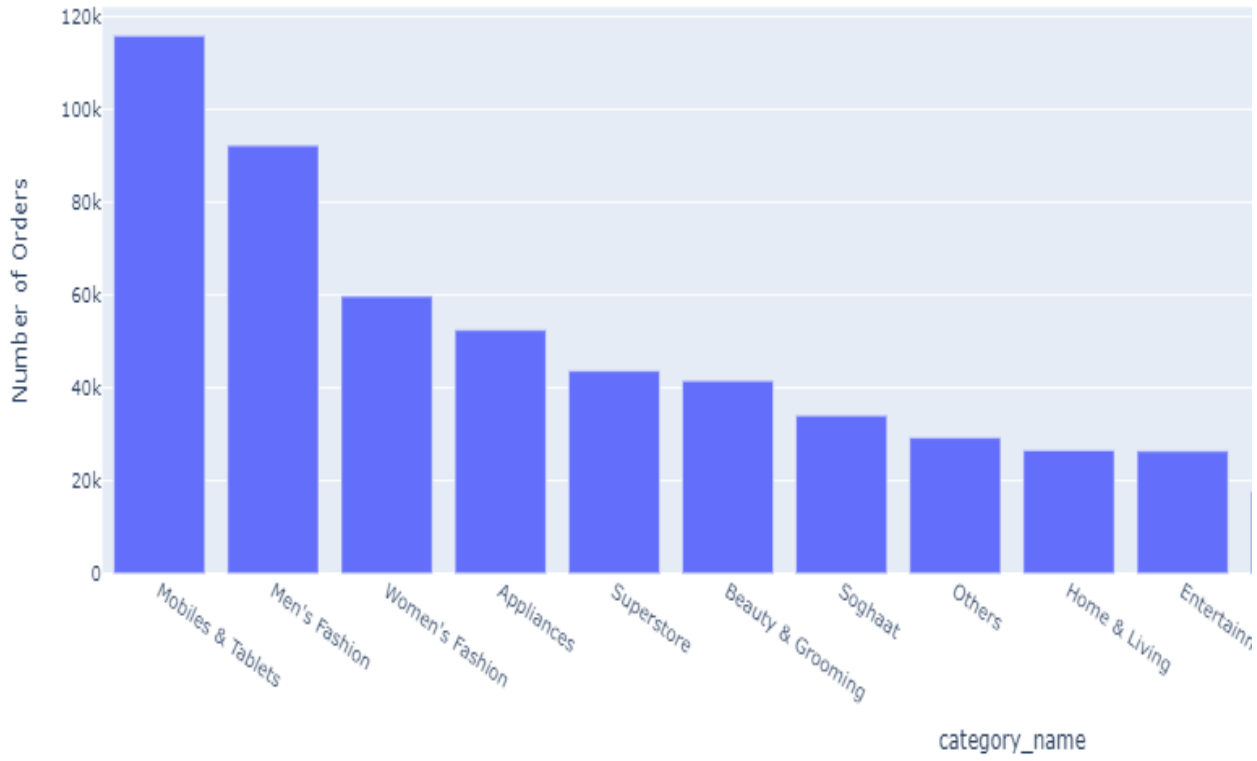


Bestselling categories by no. of orders

```
categories_by_order_count = data_sub.groupby('category_name_1').size().reset_index(name='Number of Orders').sort_values('Number of Orders',ascending=False)

bar = px.bar(categories_by_order_count, y='Number of Orders', x='category_name_1',
             title='Number of orders by Category',
             hover_data=['category_name_1'], labels={'category_name_1':'category_name'})
bar.show()
```

✓ 0.6s

On inspecting Product categories vs number of orders for the whole dataset, it was found that the bestselling category was Mobiles & Tablets. Count ranges from 0-120k.



There following are the types of status that order can have:

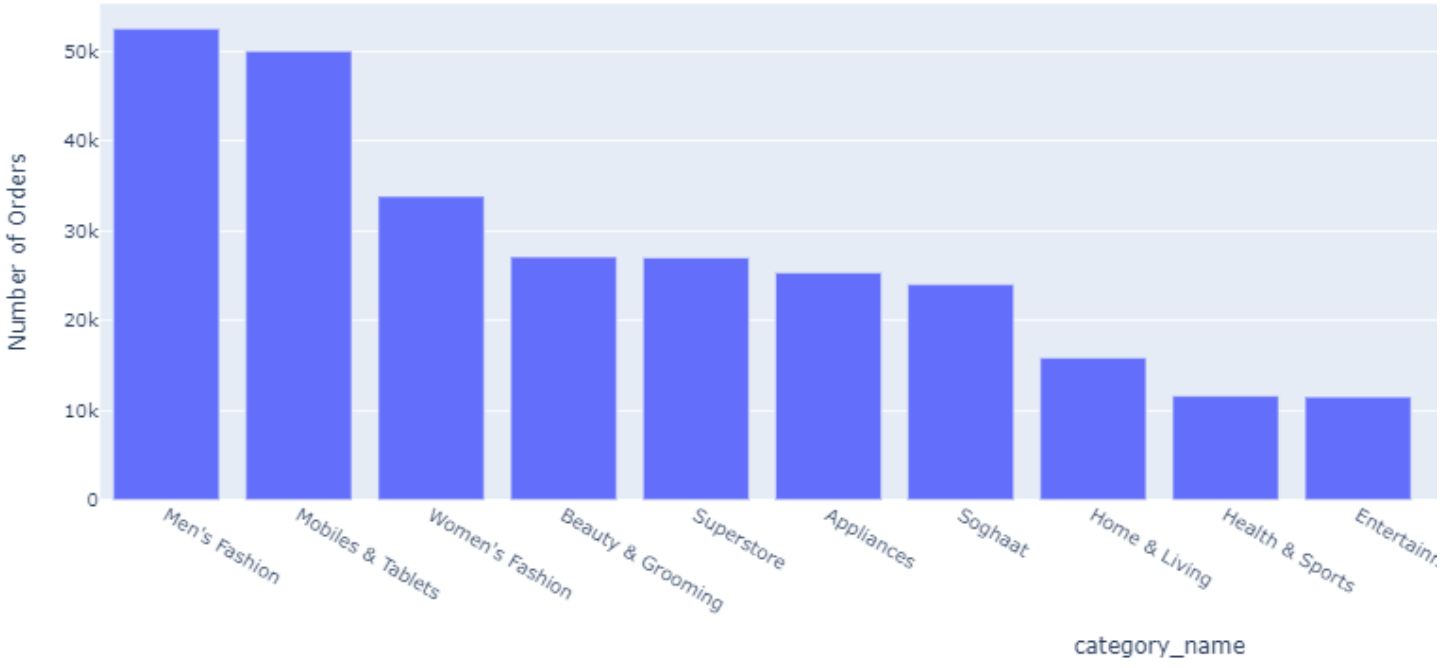
```
array(['complete', 'canceled', 'order_refunded', 'received', 'refund',
       'closed', 'fraud', 'holded', 'exchange', 'pending_paypal', 'paid',
       '\\N', 'cod', 'pending', 'processing', 'payment_review'],
      dtype=object)
```

Best Category by completion of order

Considering order statuses 'complete', 'paid', 'received' as completed. The following code is:

```
completed_orders = data_sub[(data_sub['status'] == 'complete') | (data_sub['status'] == 'paid') | (data_sub['status'] == 'received')]
completed_orders_by_category_count = completed_orders.groupby('category_name_1').size().reset_index(name='Number of Orders').sort_values('Number of Orders', ascending=False)

pie = px.bar(completed_orders_by_category_count, y='Number of Orders', x='category_name_1',
             title='Number of completed orders by Category',
             hover_data=['category_name_1'], labels={'category_name_1':'Category name'})
pie.show()
✓ 0.4s
```

On inspecting Product categories vs number of completed orders for the whole dataset, it was found that the bestselling category was Men’s Fashion. Count ranges from 0-50k.



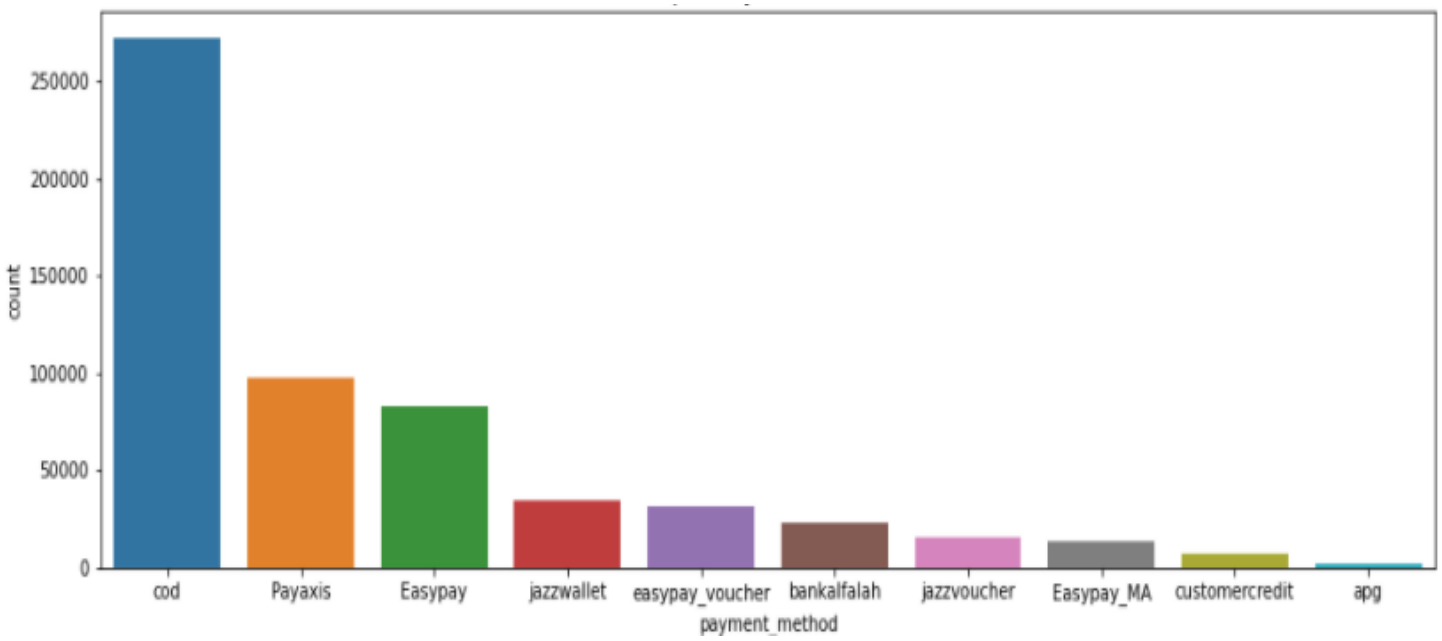
Some more findings

Following are the top 10 payment methods:

payment_method	count
cod	271960
Payaxis	97641
Easypay	82900
jazzwallet	35145
easypay_voucher	31176
bankalfalah	23065
jazzvoucher	15633
Easypay_MA	14028
customercredit	7555
apg	1758

```
plt.figure(figsize=(15,5))
graph=sns.barplot(x='payment_method',y='count',data=paymethod)
graph.set_title('Top 10 Payment Method')
plt.show()
```
✓ 0.2s

On inspecting payment methods vs count for the whole dataset, it was found that the best payment method was Cash on delivery (COD). Count ranges from 0-250000.



Limitation:

As we preferred on using cleaned data to our finalized model, due to nature of the data some of the anomalies are missed which can in turn affect the result of the above modelling.

Advice to the upcoming ventures:

More details about the customers can be recorded like their feedback etc.

Python links:

Heatmap: https://www.kaggle.com/hussainaliarif/ecommerce-best-selling-category-analysis#Heatmap-of-Categories-Group-by-Years

Yearly Best Seller Category: https://www.kaggle.com/mazhar01/task-1-best-seller-category#Question-1:-Best-Selling-Category

Yearly comparison: https://www.kaggle.com/mohsinmahmood83/best-category-pakistan-largest-ecommerce-dataset#Would-be-interesting-to-see-if-this-data-set-and-this-hierarchy-or-Top-N-categories-for-grand-total-and-Qty-Ordered-holds-true-on-an-year-to-year-basis.-We-have-three-unique-years-beint-2016,17-and-18

Order completed/not completed: https://www.kaggle.com/fazalerabbi/best-selling-category-pak-largest-e-com-dataset#Worst-category-by-not-completed-order

https://www.kaggle.com/mfaisalqureshi/pakistan-e-commerce-data-analysis