

Capstone Project

Pakistan retail/ecommerce data analysis & future sales prediction

- *Vivek Dumbre*

Content

- Problem Statement
- Approach
- Data Summary
- Exploratory Data Analysis
- RFM analysis
- Model building:
- Model Performance and Evaluation
- Conclusion and Recommendations
- Reference slide

Problem Statement

The largest retail e-commerce orders dataset from Pakistan. It contains half a million transaction records from March 2016 to August 2018. The data was collected from various e-commerce merchants as part of a research study. The major problems to solve or to analyze are as mentioned below -

- **What is the best-selling category?**
- **Visualize payment method and order status frequency**
- **Find a correlation between payment method and order status**
- **Find any hidden patterns that are counter - intuitive for a layman**
- **Can we predict number of orders, or item category or number of customers/amount in advance?**

Approach

The following approach was followed in the completion of the project:

- **Data Collection & Load**
 - Data Collection and Preprocessing
 - Data Cleaning
 - Missing Data Handling
 - Merging the Datasets
- **Exploratory Data Analysis**
 - Hypotheses
 - Categorical Features
 - Continuous Features
 - EDA Conclusion and Validating Hypotheses
- **Data Modelling and Manipulation**
 - Feature Engineering
 - Outlier Detection and Treatment
- **RFM Analysis by visualizing data**
 - Elbow curve and Silhouette analysis to determine No. segments
 - Segmentation
- **Modelling**
 - Train -Test split
 - Comparing the model fitment
 - Feature Scaling
 - Categorical Data Encoding
- **Model Performance Evaluation and Selection**
 - Visualizing Models
 - Testing Model performance
- **Finding Hidden Trends**
 - Visualizing Data for finding various hidden patterns and trends based on seasonal behaviours
- **Conclusive Insight Recommendations.**

Data Collection & Load.

The dataset contains variables as follows with the following summary details.

- item_id	- category_name_1	- Month	- Unnamed: 24
- status	- sales_commission_code	- Customer Since	- Unnamed: 25
- created_at	- discount_amount	- M-Y	
- sku	- payment_method	- FY	
- price	- Working Date	- Customer ID	
- qty_ordered	- BI Status	- Unnamed: 21	
- grand_total	- MV	- Unnamed: 22	
- increment_id	- Year	- Unnamed: 23	

Dataset Dimensions : **(1048575, 26)**

The Data set is a huge collection of orders placed by customers from various ecommerce businesses describing various demographic and natural behaviours in the orders of pakistan.

the data collected includes orders from 2016 to 2018.

Data Cleaning & Preprocessing

Although the data acquired is ample in amount and describes a lot of information but after cleansing the dataset shrinks down to only 50% of valid records post handling missing values.

EDA is probably the most important step in any kind of Data analytics case study as it lays down the basis and direction of the analysis towards proper modelling.

Thus we must avoid completely getting rid of any record and should emphasize on mean, median or mode replacement in the missing value cells.

Moving to our data to achieve the objectives mentioned earlier in the Problem statement we will be carefully handling the data and make necessary adjustments.

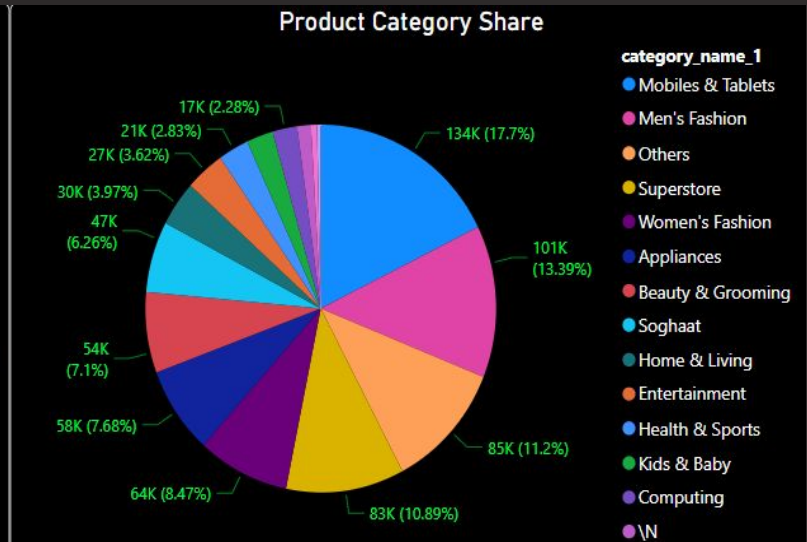
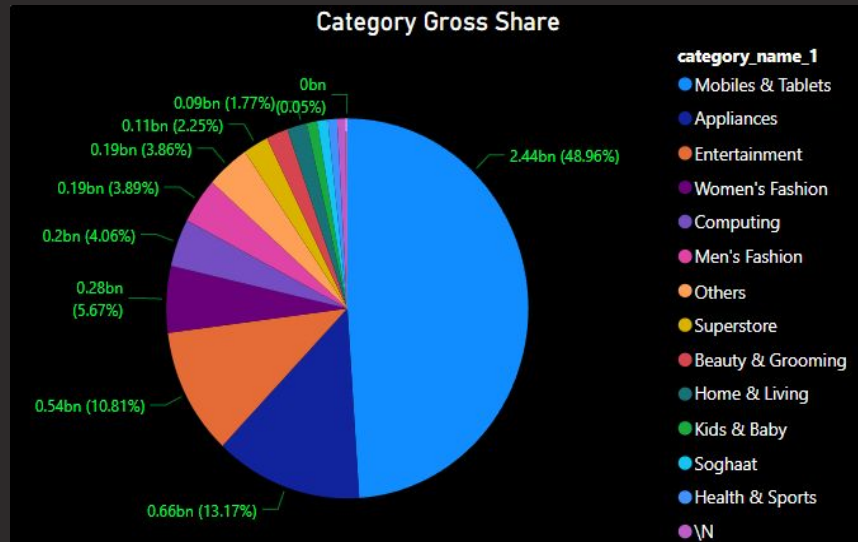
Let's move forward to the technical analysis and the code of the study.



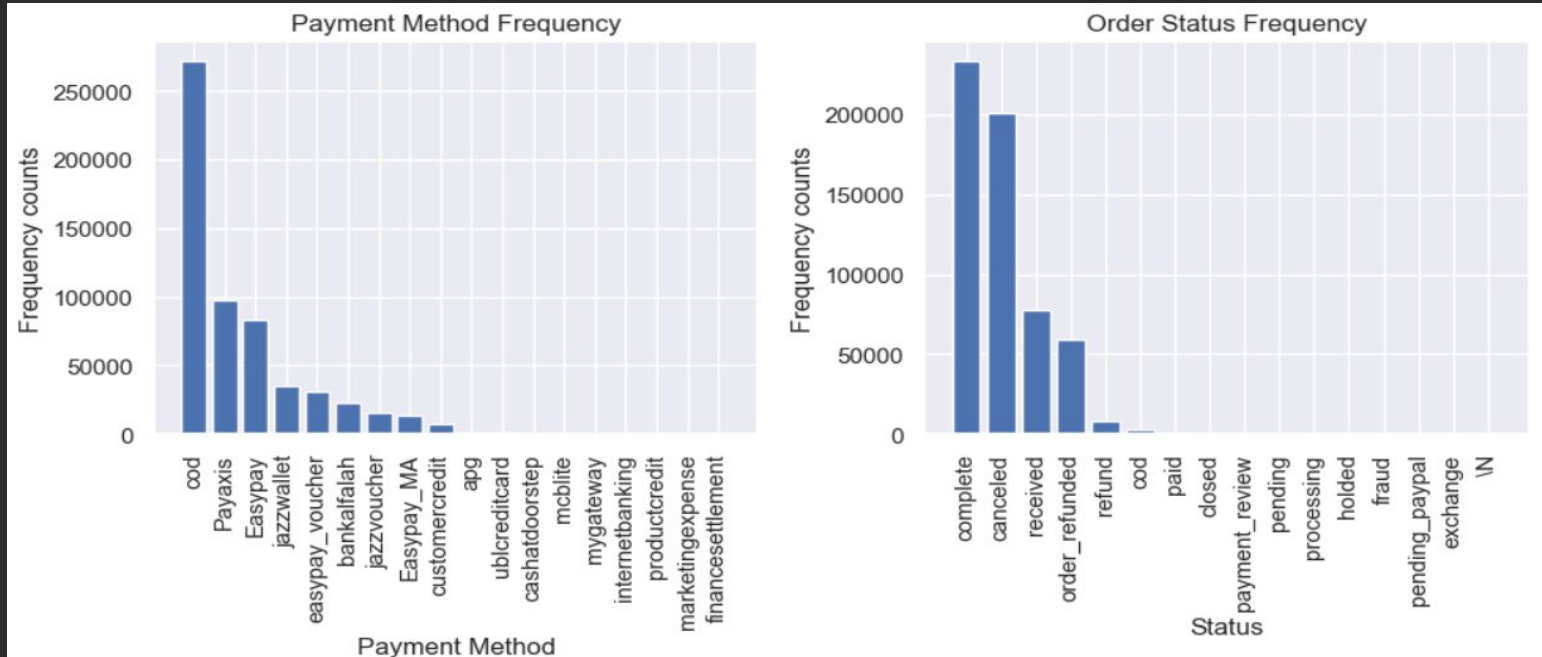
EDA

What is the best-selling category?

The categorical sales is clearly dominated by “**Mobiles \$ Tablets**” in both segments i.e. in Gross sales and number of products sold per category as shown below in the designed pie plot designated for all categories sales.



Visualize payment method and order status frequency



Find a correlation between payment method and order status




As we already know payment method and Offer Status are both categorical variables to find the correlation between them we need the Chi- Square Test to map their inter-dependence as the variables are non numeric.

The chi-square : **208092.8816361259**, p-value : **0.0**

The chi-square statistic indicates the strength of association between the two categorical variables. A higher chi-square value suggests a stronger association.

The p-value is a measure of the statistical significance of the association. In this case, the p-value is very close to zero (0.0), which suggests that the association between the variables is statistically significant. In other words, it is highly unlikely to observe such an association by chance alone.

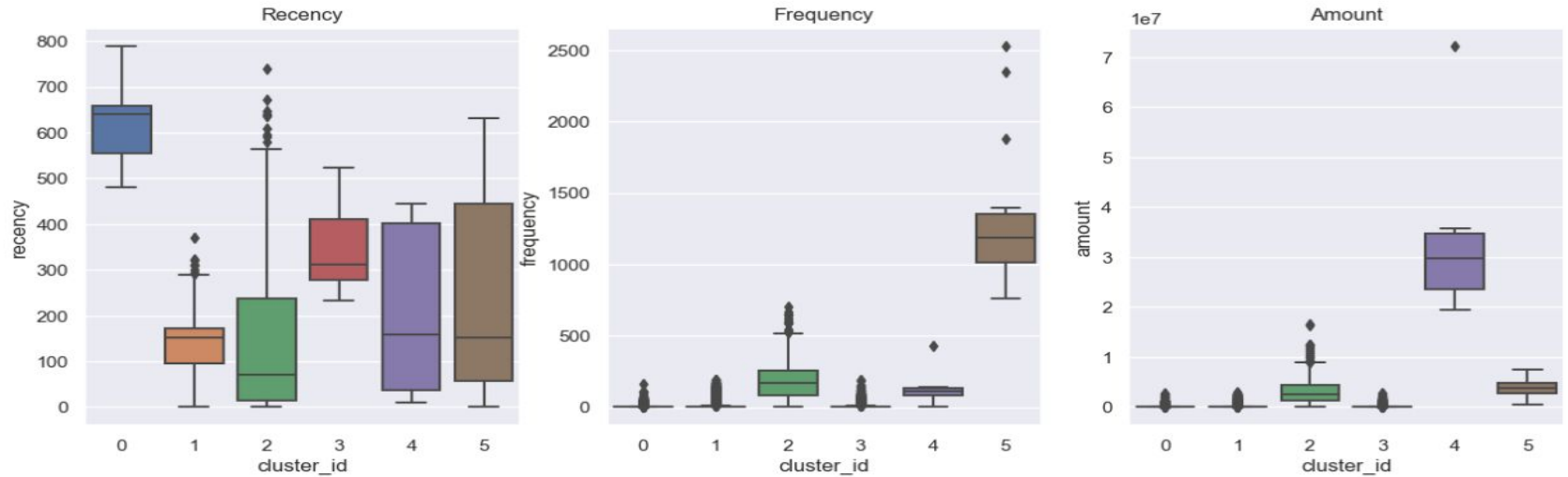
These results indicate that there is a significant association or dependency between the two categorical variables you analyzed.



RFM Analysis & Segmentation

RFM analysis is a marketing technique used to quantitatively rank and group customers based on the recency, frequency and monetary total of their recent transactions to identify the best customers and perform targeted marketing campaigns.

RFM results

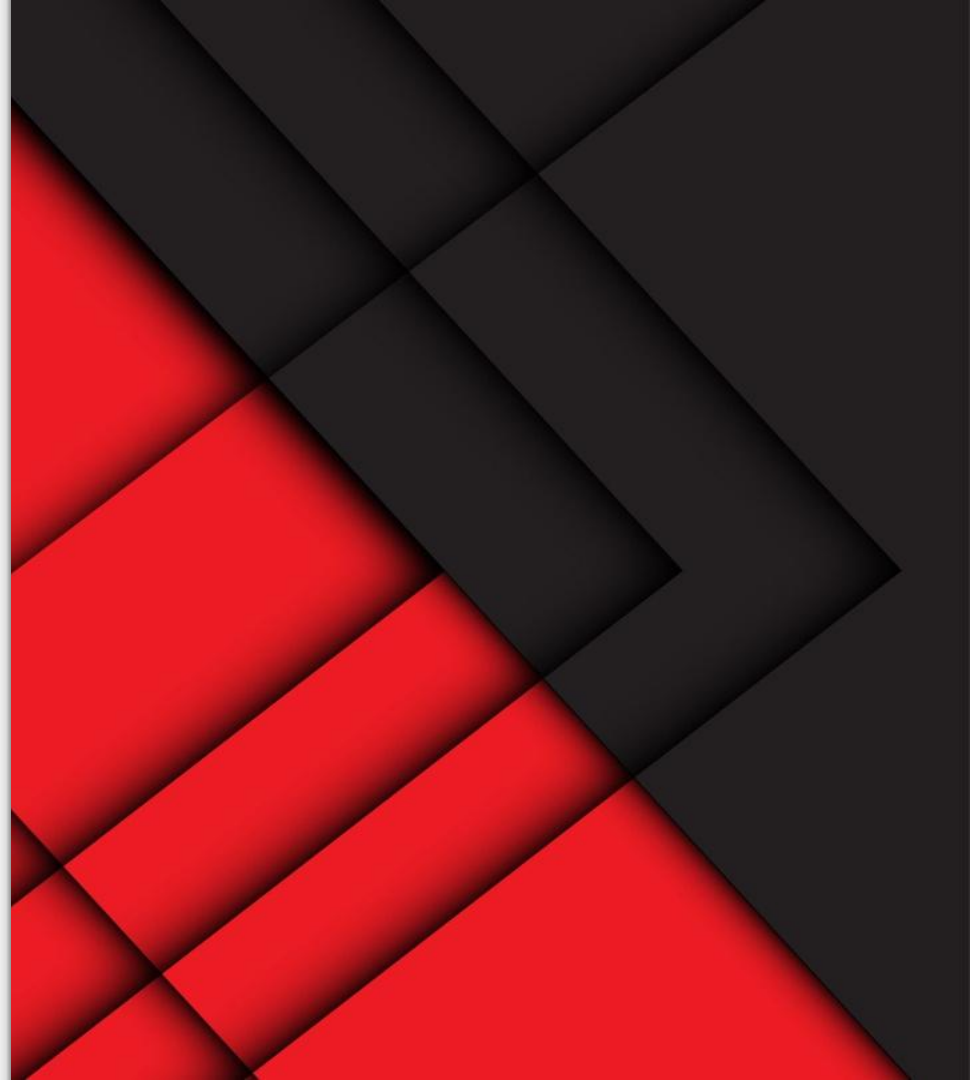


Model Building

In our problem we will be using the ARIMA and SARIMAX models for testing our hypothesis and build our models

The ARIMA model is great, but to include seasonality and exogenous variables in the model can be extremely powerful. Since the ARIMA model assumes that the time series is stationary, we need to use a different model.

Above is the the of the SARIMAX model. This model takes into account exogenous variables, or in other words, use external data in our forecast.



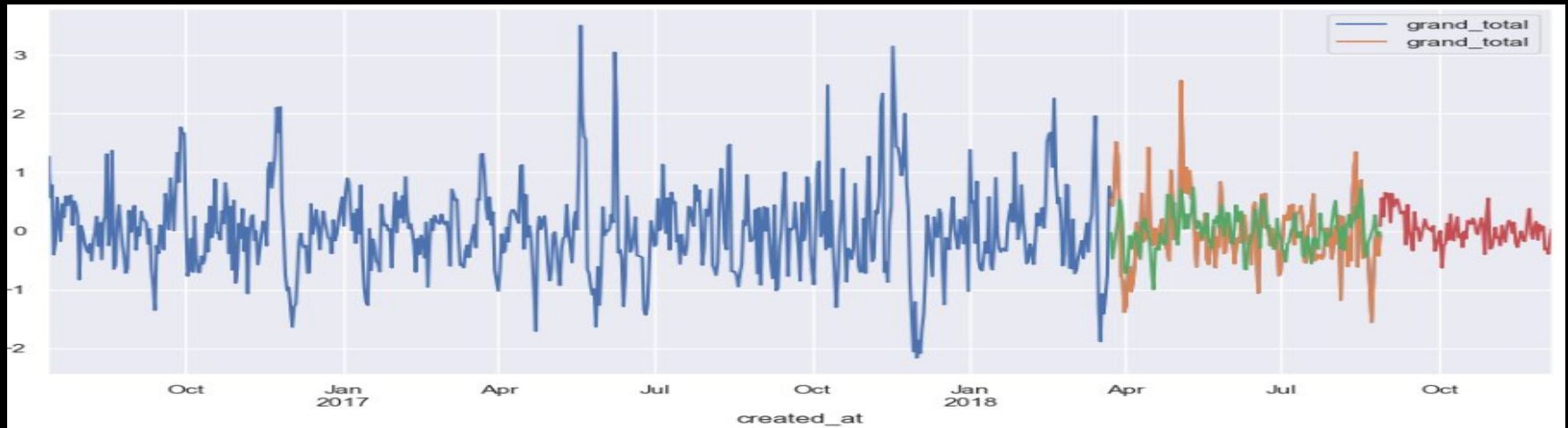
ARIMA model results

Red line shows predictions

Green Line shows training part of Dataset

Orange Line shows Test part of Dataset

Mean_absolute_error 0.4340195628655157
0.57113410026802



SARIMAX model results

Naturally the SARIMAX model performs slightly better than the ARIMA model as it is more advanced in capturing features and making predictions due to its ability to take account of the seasonality and other external factors that might affect the analysis.



Insights and Hidden trends found in data based on above analysis

We observe overall trend from july-2016 to aug-2018 and concluded following points

- Purchases increases in last Three months (October,November and December) of 2016 and 2017
- Most e-commerce buyers purchases products in November
- November is most suitable month for sellers to market their products
- Overall volume of ecommerce sales is increasing year by year
For example,you can compare first six months of 2017 with 2018

Thanks!

References:

[Time-series forecasting](#)

<https://neptune.ai/blog/arma-sarima-real-world-time-series-forecasting-guide>

