MattYoungMedia

Sales Data Analysis Report

-Rahul Bhagat

Introduction

This report summarises the descriptive findings and insights from the Sales-Dataset from Kaggle. The Analysis was carried out on Python(Pandas,Seaborn,Matplotlib) and PowerBI. This report is a part of an assessment for the position of a Data Anlayst Intern at MattYoungMedia.

The Data

The Dataset consisted of 25 columns and 2823 rows. Here is a peek the dataset loaded in a Pandas Dataframe:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2823 entries, 0 to 2822
Data columns (total 25 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ORDERNUMBER       2823 non-null   int64
 1   QUANTITYORDERED   2823 non-null   int64
 2   PRICEEACH         2823 non-null   float64
 3   ORDERLINENUMBER   2823 non-null   int64
 4   SALES             2823 non-null   float64
 5   ORDERDATE         2823 non-null   object
 6   STATUS            2823 non-null   object
 7   QTR_ID            2823 non-null   int64
 8   MONTH_ID          2823 non-null   int64
 9   YEAR_ID           2823 non-null   int64
 10  PRODUCTLINE       2823 non-null   object
 11  MSRP              2823 non-null   int64
 12  PRODUCTCODE       2823 non-null   object
 13  CUSTOMERNAME      2823 non-null   object
 14  PHONE             2823 non-null   object
 15  ADDRESSLINE1      2823 non-null   object
 16  ADDRESSLINE2      302 non-null    object
 17  CITY              2823 non-null   object
 18  STATE             1337 non-null   object
 19  POSTALCODE        2747 non-null   object
 20  COUNTRY           2823 non-null   object
 21  TERRITORY         1749 non-null   object
 22  CONTACTLASTNAME   2823 non-null   object
 23  CONTACTFIRSTNAME  2823 non-null   object
 24  DEALSIZE          2823 non-null   object
dtypes: float64(2), int64(7), object(16)
memory usage: 551.5+ KB
```

Fig. 1

As seen in Fig.1, the columns named - `[ADDRESSLINE2 , STATE , POSTALCODE , TERRITORY]` contain null values. Luckiy for our analysis, we did not require these columns

much, hence we could either drop it or let it stay the way it was.

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	QTR_ID	MONTH_ID	YEAR_ID	MSRP
count	2823.000000	2823.000000	2823.000000	2823.000000	2823.000000	2823.000000	2823.000000	2823.00000	2823.000000
mean	10258.725115	35.092809	83.658544	6.466171	3553.889072	2.717676	7.092455	2003.81509	100.715551
std	92.085478	9.741443	20.174277	4.225841	1841.865106	1.203878	3.656633	0.69967	40.187912
min	10100.000000	6.000000	26.880000	1.000000	482.130000	1.000000	1.000000	2003.00000	33.000000
25%	10180.000000	27.000000	68.860000	3.000000	2203.430000	2.000000	4.000000	2003.00000	68.000000
50%	10262.000000	35.000000	95.700000	6.000000	3184.800000	3.000000	8.000000	2004.00000	99.000000
75%	10333.500000	43.000000	100.000000	9.000000	4508.000000	4.000000	11.000000	2004.00000	124.000000
max	10425.000000	97.000000	100.000000	18.000000	14082.800000	4.000000	12.000000	2005.00000	214.000000

Fig.2

The Fig.2 shows the statistics of the numerical columns in the Dataset. Now the columns - `['ORDERNUMBER','ORDERLINENUMBER','QTR_ID','MONTH_ID','YEAR_ID']`

have no real significance of the mean and standard deviations.

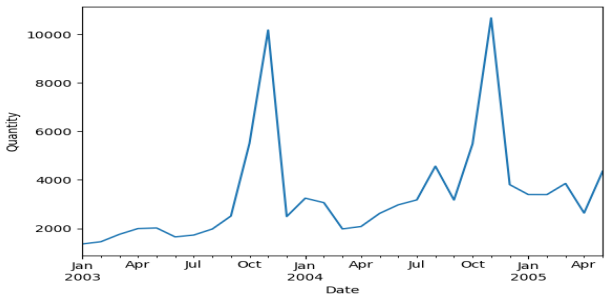
On the other hand the rest of the columns help us explore the dataset with a mathematical eye. We can see the average quantity ordered ~ 35, average sales per transaction ~ $ 3554 and more.

We also observe that the standard deviation for Price and MSRP(Retail Price) is pretty significant, indicating a good spread in them.

Let us do some Exploratory Data Analysis and some awesome Visualisations !

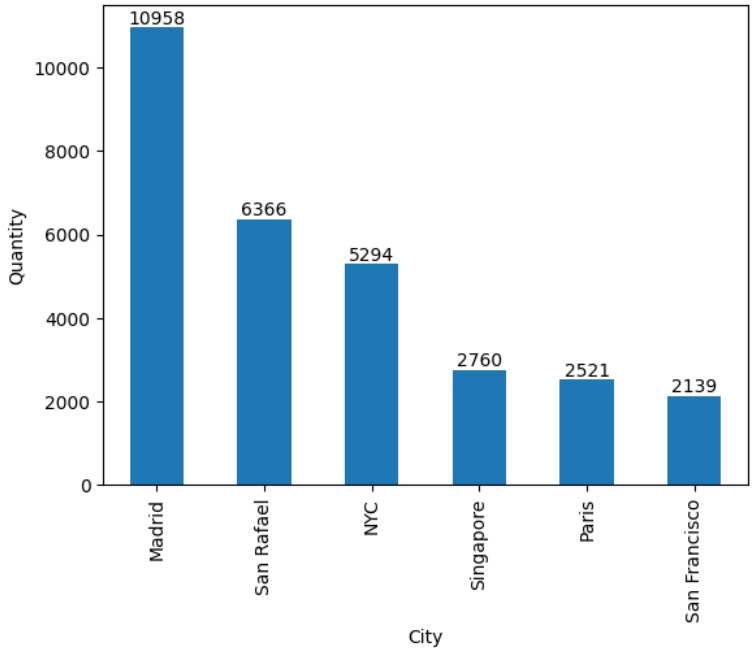
Visualisations and Analysis

We now explore the dataset even more deeply and plot visualisations for better understanding. Fig 3.



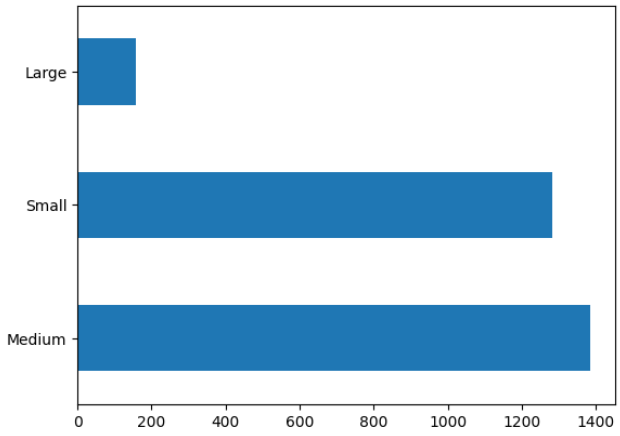
- In the figure above, we see the total sales in a month for the entire duration in the dataset. We observe an interesting trend in data, i.e. around September every year, the sale start increasing and skyrocket in October to December.

Fig. 4.



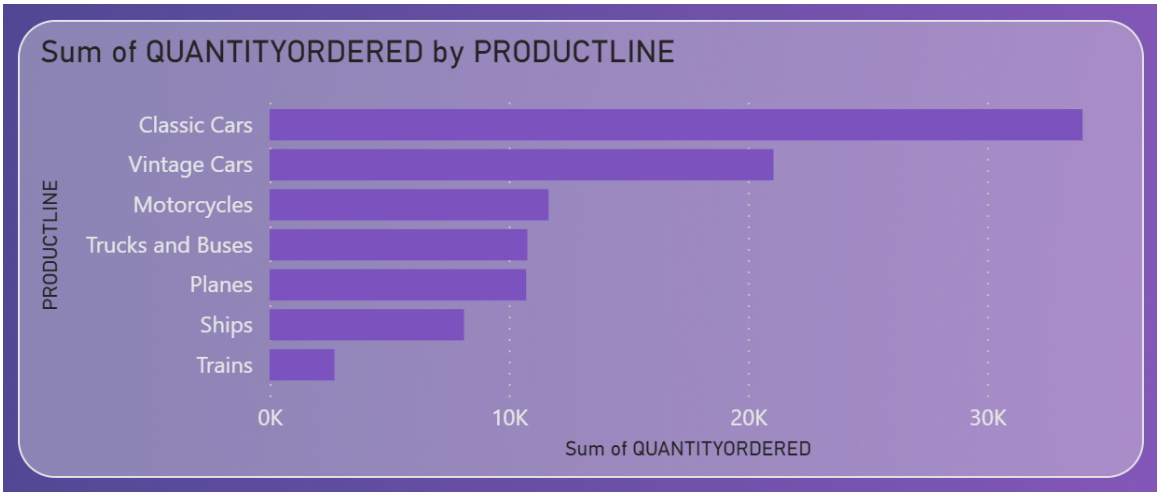
- In Figure 4, we now see the top 6 cities in terms of overall Sales/Quantity in a barplot. Madrid and San Rafael are the cities with the most sales.

  Fig. 5.



- In Figure 5, we see the countplot of the Ordersize column. We conclude that majority of the orders are either Small or Medium in volume.

  Fig.6.



- In figure 6, we can see the product-lines in the dataset. We observe that Cars and Motorcycles are the top categories in terms of order Quantity.

  The above two cards in PowerBI indicate the total quantity of order sold and the total sales generated. These are connected to two slicers in the dashboard that can help us filter in terms of the Year and top 6 Cities.

Conclusion

The dataset, as mentioned earlier had some missing values, but we did not need to tend to them because of their insignificance to the analysis. Some key insights gained were:

- The sales peak from October to December, every year. The reason for this could be multiple, but we would need more domain information and product informations ro uncover it.
- The company is significantly international in terms of sales, as evident by the top 6 cities being in multiple continents.
- Car and Motorcycle products have higher sales, which is understandable given the disparity in the distribution of the amount of vehicles on the planet.

This concludes the report on the Kaggle Sample Sales Dataset.