Sales Data Analysis

Business Understanding

Given 4 months of sales data of a leading US based electronics store, a thorough analysis has been conducted to identify various insights pertaining to the trends and patterns in the sale of various electronic products. This is of paramount importance as this would help the business concentrate on aspects which would previously not have been possible.

Understanding the data

The dataset is an open-source dataset taken from Kaggle. The dataset has a good part of over 55000 records with 6 features.

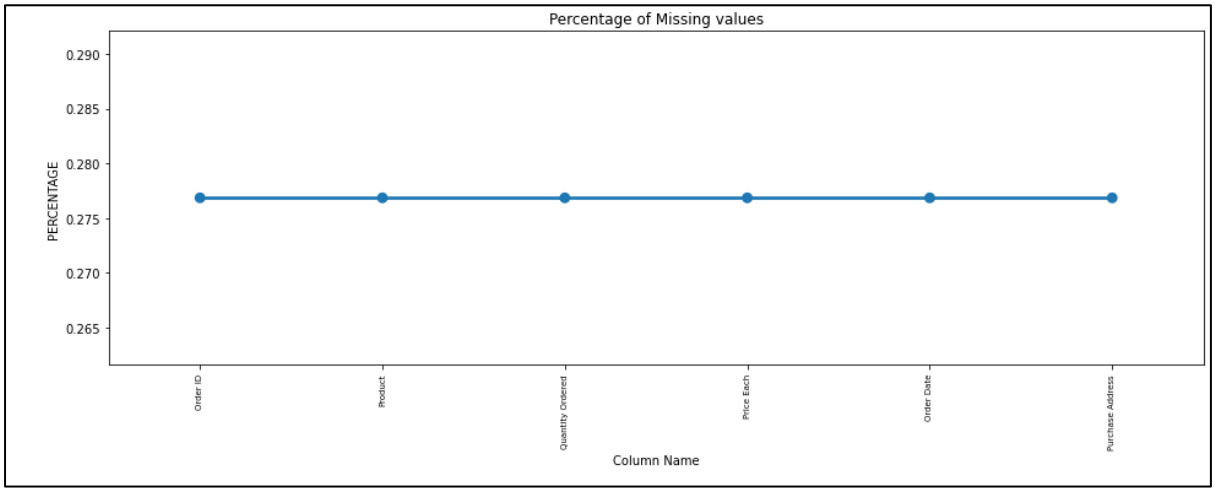
```
df.shape
```

```
(55264, 6)
```

On further inspection, it is observed that a certain sub-section of the dataset is missing and to everyone's surprise, the missing data is equally distributed across the features. PFB the screenshot depicting the same.

```
df.isnull().sum()
```

```
Order ID            153
Product             153
Quantity Ordered    153
Price Each          153
Order Date          153
Purchase Address    153
dtype: int64
```



```
Order ID            0.276853
Product             0.276853
Quantity Ordered    0.276853
Price Each          0.276853
Order Date          0.276853
Purchase Address    0.276853
dtype: float64
```

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
664	NaN	NaN	NaN	NaN	NaN	NaN
678	NaN	NaN	NaN	NaN	NaN	NaN
797	NaN	NaN	NaN	NaN	NaN	NaN
876	NaN	NaN	NaN	NaN	NaN	NaN
1283	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...
53644	NaN	NaN	NaN	NaN	NaN	NaN
53793	NaN	NaN	NaN	NaN	NaN	NaN
53941	NaN	NaN	NaN	NaN	NaN	NaN
54327	NaN	NaN	NaN	NaN	NaN	NaN
54599	NaN	NaN	NaN	NaN	NaN	NaN

153 rows × 6 columns

As the proportion of missing values is miniscule and the fact that these missing values occur as a block, it is best to drop such records.

Post dropping missing records, the data dimensions are as follows:

```
df.shape
```

```
(55111, 6)
```

Initial analysis reveals that the datatype of all attributes has been marked as ‘Object’ which is a misrepresentation of what the attributes mean. PFB the screenshot of the same.

```
df.dtypes
```

```
Order ID            object
Product             object
Quantity Ordered    object
Price Each          object
Order Date          object
Purchase Address    object
dtype: object
```

Thus, necessary changes need to be made to convert the datatypes to a suitable format. PFB the screenshot of the same after converting the datatypes to a suitable format.

```
df['Order ID']=df['Order ID'].astype(int)
df['Quantity Ordered']=df['Quantity Ordered'].astype(int)
df['Price Each']=df['Price Each'].astype(float)
df.dtypes
```

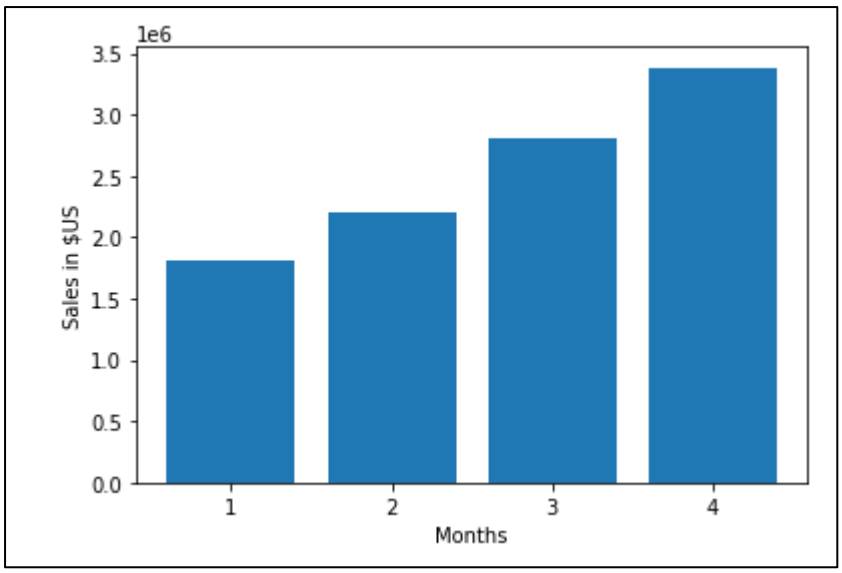
```
Order ID              int32
Product              object
Quantity Ordered      int32
Price Each          float64
Order Date           object
Purchase Address     object
dtype: object
```

Unearthing Insights

Insight: One of the key questions that is often asked is 'Which month attracts the highest sales?'. The answer is as follows:

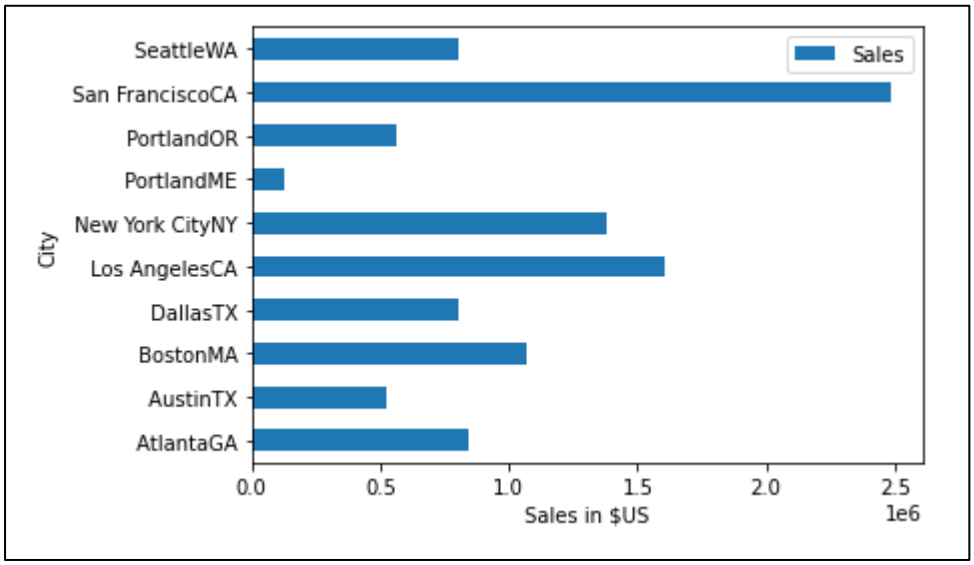
```
Month
1    1810225.54
2    2200200.25
3    2800767.91
4    3384178.56
Name: Sales, dtype: float64
```

It is seen that the highest sales have happened in the month of April. Below is a pictorial representation of the same.

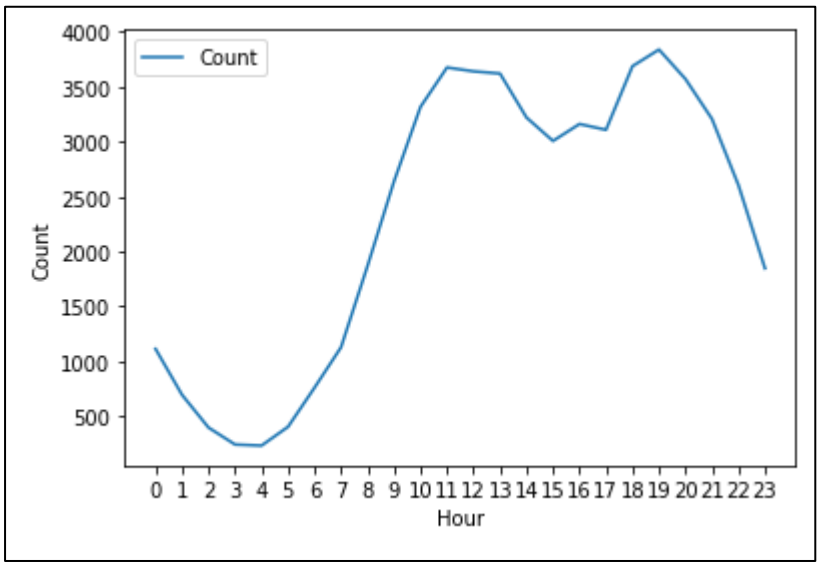


Insight: San Francisco region observed the highest sales and the Portland, Maine region, the least.

```
City
 AtlantaGA            841932.86
 AustinTX             523441.51
 BostonMA            1066863.81
 DallasTX             801042.79
 Los AngelesCA       1606134.83
 New York CityNY     1379786.34
 PortlandME           124907.07
 PortlandOR           563277.56
 San FranciscoCA     2481771.28
 SeattleWA            806214.21
Name: Sales, dtype: float64
```



Insight: Sales seem to peak just before 11AM and before 7PM. It is, therefore, best to display advertisements just before these timeslots.

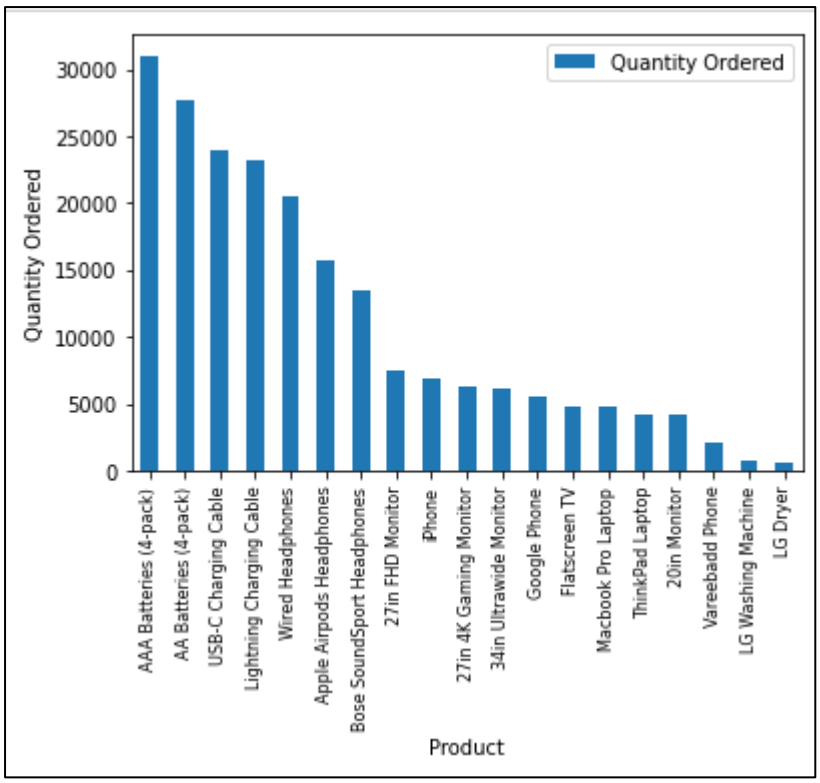


Insight: A combination of an iPhone and Lightning charging cables were purchased together most often.

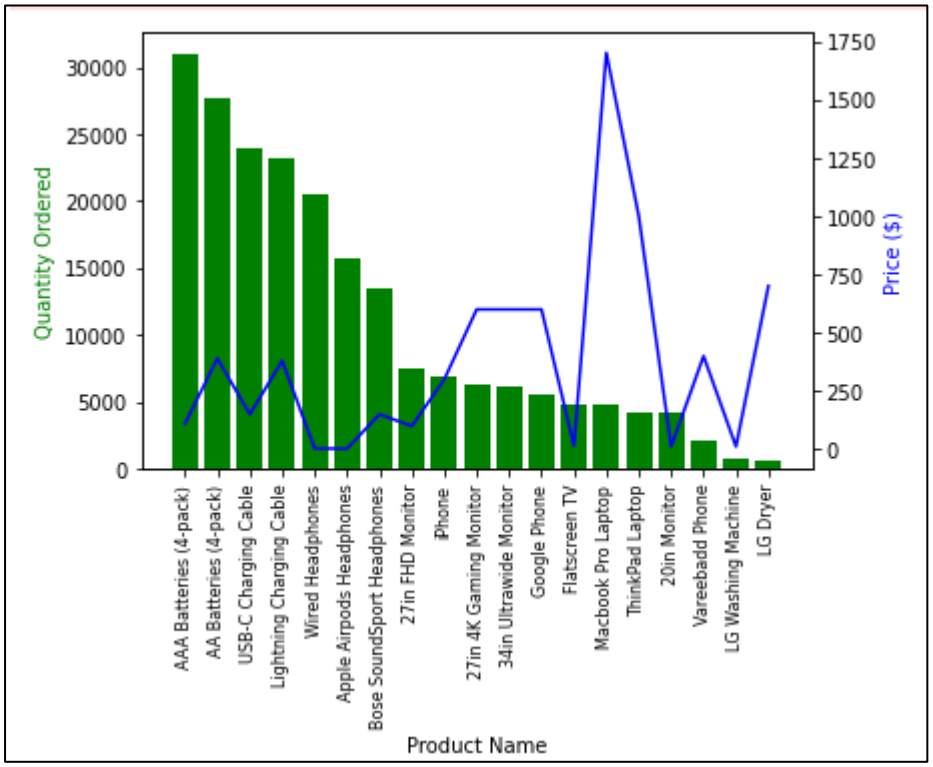
```
[(('Google Phone', 'USB-C Charging Cable'), 327),
 (('iPhone', 'Lightning Charging Cable'), 304),
 (('iPhone', 'Wired Headphones'), 136),
 (('Google Phone', 'Wired Headphones'), 125),
 (('Vareebadd Phone', 'USB-C Charging Cable'), 104),
 (('iPhone', 'Apple Airpods Headphones'), 99),
 (('Google Phone', 'Bose SoundSport Headphones'), 72),
 (('Vareebadd Phone', 'Wired Headphones'), 50),
 (('USB-C Charging Cable', 'Wired Headphones'), 48),
 (('USB-C Charging Cable', 'Bose SoundSport Headphones'), 27)]
```

Insight: AAA batteries sold the most and LG dryers sold the least.

The above observation could be attributed to AAA batteries being extremely cheap and an everyday consumable, while LG dryers being an expensive item.



We assume a hypothesis, which states "**higher the prices, lower the sales"**. We test the hypothesis by overlaying the item-price graph on top of the item-quantity graph.



Final Thoughts

The EDA process has unearthed several insights. Some of them are as follows:

- The highest number of sales happened in the month of April.
- San Francisco observed the highest sales, while the Portland, Maine region observed the least number of sales. This could very well be attributed to the tech savvy people of California, and at the same time, the people of California in general having a higher income.
- Sales seem to peak at 11AM and 7PM. This sort of information is incredibly useful for marketing teams. As per this data, it is best advised to display advertisements 15-30 minutes prior to the above-mentioned time slots. A reason for peak sales at these time slots could be attributed to office goers generally taking coffee breaks at around 11AM and most office goers returning home by 7PM.
- Google Phones along with USB-C charging cables seem to be the most shopped pair of items. This could be attributed to the fact that buyers often like keeping spare charging cables which they can use for charging their mobile devices in their vehicles. Insights such as these could help businesses tailor their recommendations to give out suggestions to customers when purchasing certain items and perhaps handing out discounts on certain products when they are purchased as a pair.
- AAA batteries sold the most. On the other hand, LG dryers sold the least. On first inspection, this could be attributed to inexpensive everyday items attracting more customers, while more expensive items tend to sell a lot lesser. This hypothesis has been tested out, and the hypothesis has been largely verified with a few anomalies.