Sales Data Analysis

Business Understanding

Given 12 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 or 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 180000 records with 6 features.

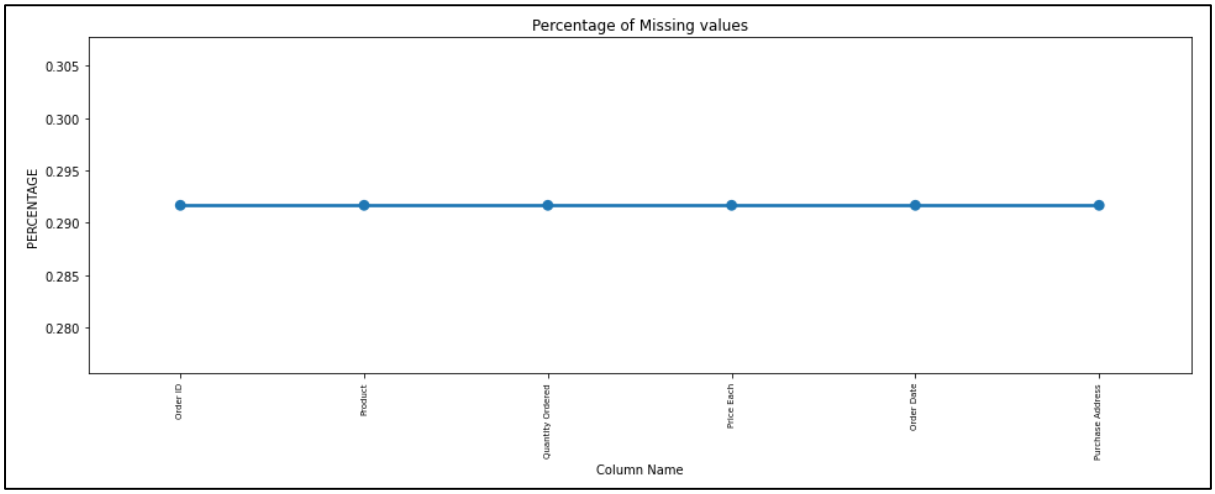
```
df.shape
```

```
(186850, 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            545
Product             545
Quantity Ordered    545
Price Each          545
Order Date          545
Purchase Address    545
dtype: int64
```



```
missing_percentage
```

```
Order ID            0.291678
Product             0.291678
Quantity Ordered    0.291678
Price Each          0.291678
Order Date          0.291678
Purchase Address    0.291678
dtype: float64
```

```
df[df.isnull().any(axis=1)]
```

	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
1299	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...
22945	NaN	NaN	NaN	NaN	NaN	NaN
22962	NaN	NaN	NaN	NaN	NaN	NaN
23309	NaN	NaN	NaN	NaN	NaN	NaN
23996	NaN	NaN	NaN	NaN	NaN	NaN
24730	NaN	NaN	NaN	NaN	NaN	NaN

545 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 decided to drop such records.

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

```
df.shape
```

```
(185950, 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
```

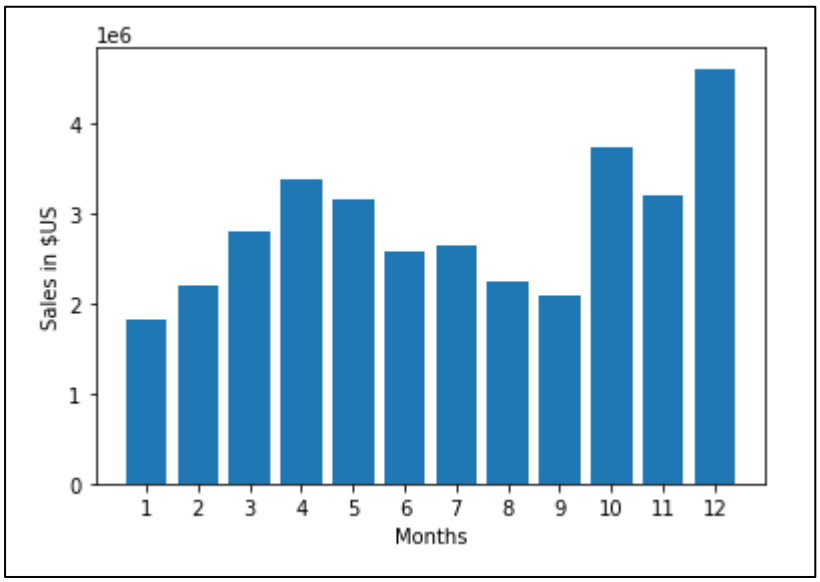
Unearthing Insights

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

```
result=df.groupby(['Month'])['Sales'].sum()
result
```

```
Month
1     1822256.73
2     2202022.42
3     2807100.38
4     3390670.24
5     3152606.75
6     2577802.26
7     2647775.76
8     2244467.88
9     2097560.13
10    3736726.88
11    3199603.20
12    4613443.34
Name: Sales, dtype: float64
```

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

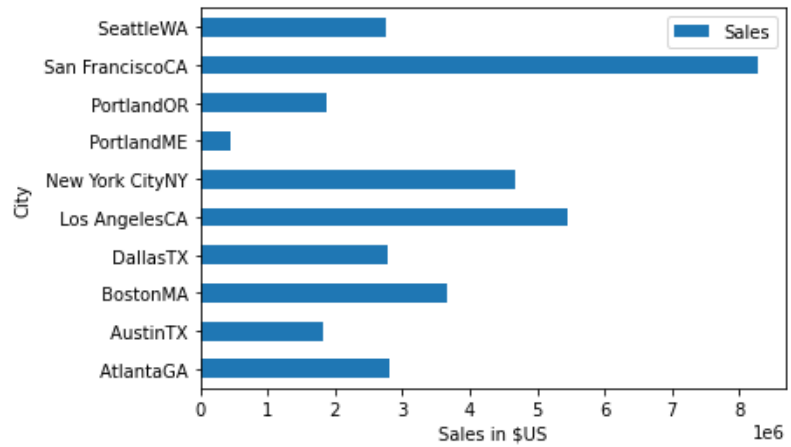


Insight: San Francisco region observed the highest sales and the Portland, Maine region the least.

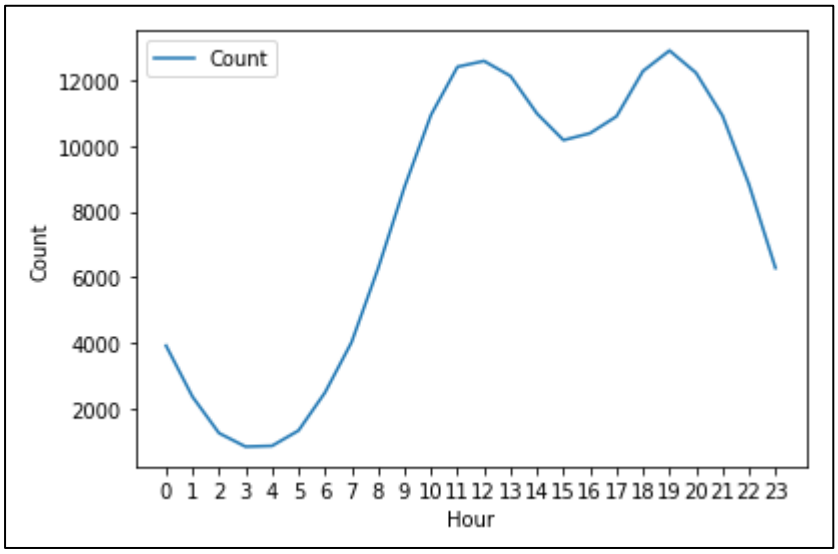
City	
AtlantaGA	2795498.58
AustinTX	1819581.75
BostonMA	3661642.01
DallasTX	2767975.40
Los AngelesCA	5452570.80
New York CityNY	4664317.43
PortlandME	449758.27
PortlandOR	1870732.34
San FranciscoCA	8262203.91
SeattleWA	2747755.48

```
city_df.plot(kind='barh',x='City',y='Sales')
plt.xlabel('Sales in $US')
```

```
Text(0.5, 0, 'Sales in $US')
```



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

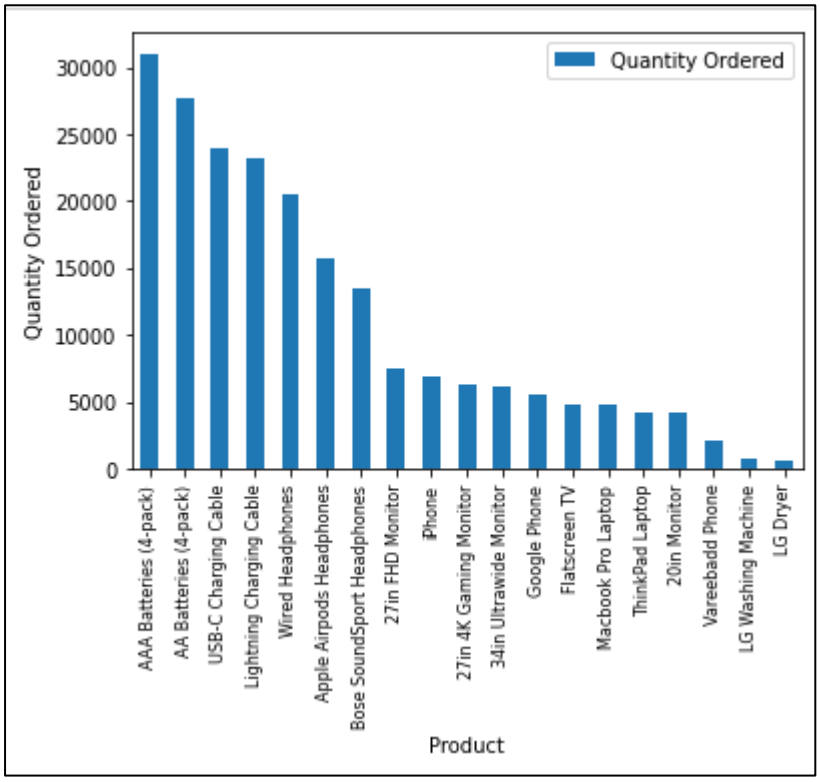


Insight: A combination of an iPhone and Lightning charging cables were purchased together most often.

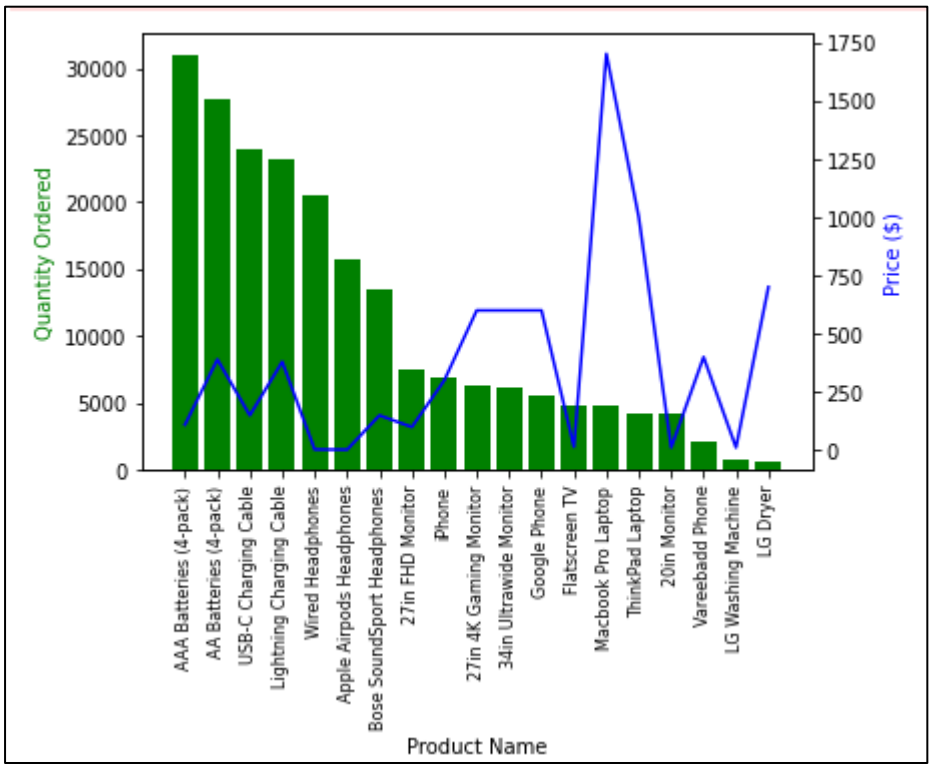
```
[(('iPhone', 'Lightning Charging Cable'), 1005),
 (('Google Phone', 'USB-C Charging Cable'), 987),
 (('iPhone', 'Wired Headphones'), 447),
 (('Google Phone', 'Wired Headphones'), 414),
 (('Vareebadd Phone', 'USB-C Charging Cable'), 361),
 (('iPhone', 'Apple Airpods Headphones'), 360),
 (('Google Phone', 'Bose SoundSport Headphones'), 220),
 (('USB-C Charging Cable', 'Wired Headphones'), 160),
 (('Vareebadd Phone', 'Wired Headphones'), 143),
 (('Lightning Charging Cable', 'Wired Headphones'), 92)]
```

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 that 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 dataset and the EDA process has unearthed several insights. Some of them are as follows:

- The highest number of sales has happened in the month of December. This could mainly be attributed to the holiday season (Christmas) which is famous for its gifting tradition.
- San Francisco has observed the highest number of sales, while the Portland, Maine region has 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 to post ads. 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.
- iPhones along with lightning charging cables seem to be the most shopped pair of items. This could be attributed to the fact that charging cables are no longer complimentary with iPhones. Insights such as these could help the business tailoring their recommendation engines, give out suggestions to customers when purchasing certain items, and perhaps helping the business in handing out discounts on certain products when purchased as a pair.
- AAA batteries are sold the most, whereas LG dryers are sold the least. On first inspection, this could be attributed to inexpensive everyday items attracting more customers, while more expensive items tend be slower moving. The hypothesis has been tested out and to a large extent, the hypothesis has been verified with a few anomalies.