Pricing Products

for profit maximization

Objective

Provide new business owners a starting point to price products based on the existing market

Framework

Data Cleaning › Exploratory Data Analysis (EDA) › Clustering › Multivariate Regression › Conclusion & Recommendation

Framework

- Dataset
- Process

| Data Cleaning | Exploratory Data Analysis (EDA) | Clustering | Multivariate Regression | Conclusion & Recommendation |

Framework

- Conclusion
- Further development
- Limitations and challenges

| Data Cleaning | Exploratory Data Analysis (EDA) | Clustering | Multivariate Regression | Conclusion & Recommendation |

Data Cleaning

Dataset

- Global superstore data
- Compilation of transactions from retailers worldwide
- 2012 to 2015
- >50,000 rows of data
- 51 features
- Product categories : Technology, Furniture and Office Supplies

Cleaning process

- Removed features with substantial (>70%) missing values
- Removed those that do not provide added value eg. Product ID is unique to every product but would not help with clustering.

Exploratory Data Analysis (EDA)
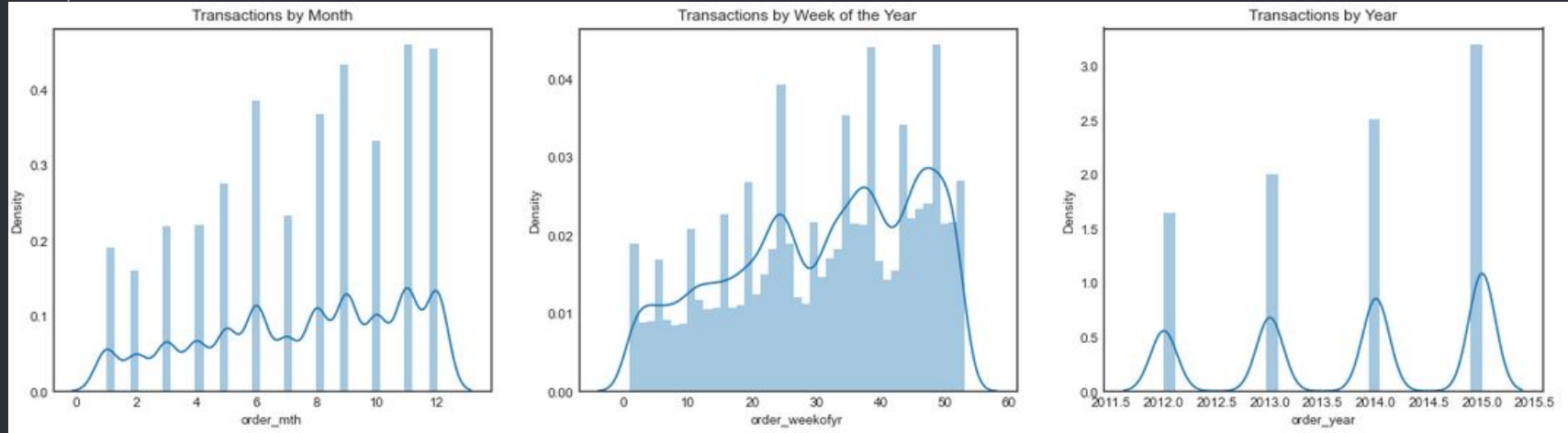
Transaction count on a monthly, weekly and yearly basis

- Growing affluence over the years

- Seasonality in sales (more sales towards the end of the year - Christmas/ bonus payouts)

Exploratory Data Analysis (EDA)
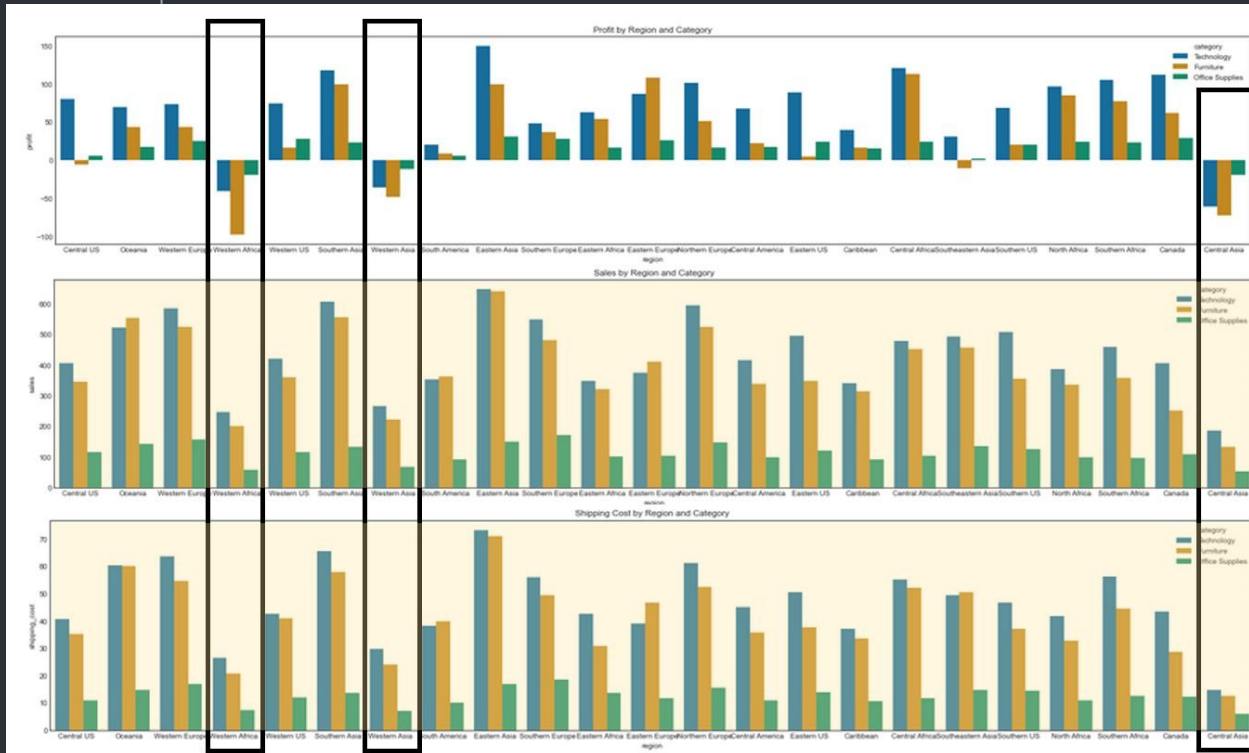
Profit, sales and shipping cost of transactions by region and category

◦ Unprofitable regions : Western Africa, Western Asia and Central Asia

◦ Generally have lower sales and shipping cost

◦ Shipping cost across regions follow the same trend as sales

Exploratory Data Analysis (EDA)
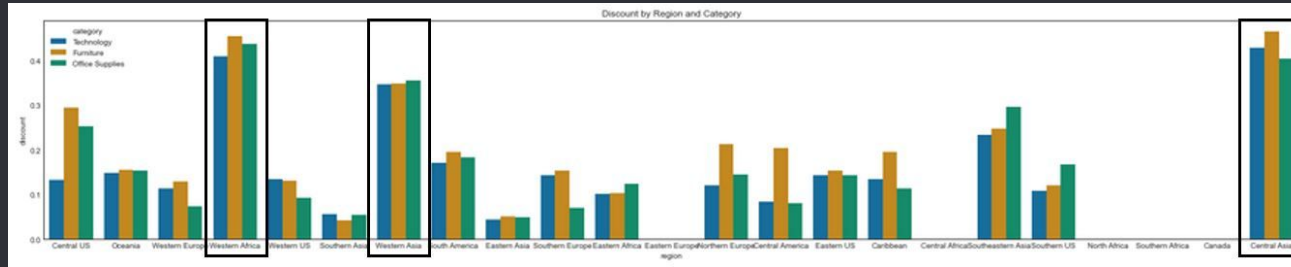
Transactions discount by region and category

- Unprofitable regions : Western Africa, Western Asia and Central Asia
- Generally have much higher discounts than other regions

Feature Engineer

- Heavily discounted (>0.3%) regions
- Non-profitable regions

Clustering - KMeans

Distortion Score Elbow for KMeans Clustering
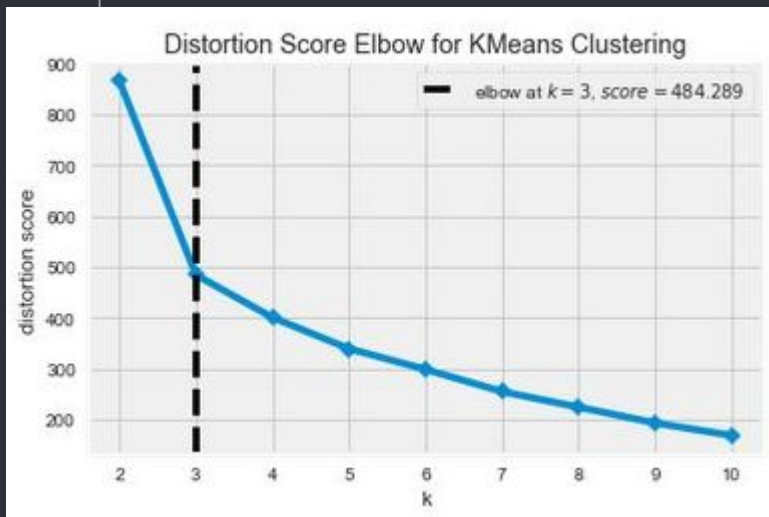
Distortion score :
Sum of squared errors from each point to its assigned centre

- **Phones** subcategory

- Determine optimal number of clusters using the elbow method

- Range of 2 to 10 clusters

- Identify a point as number of clusters increase, where the distortion score start to flatten, forming an elbow.

- Optimal number of clusters : **3**

Clustering - KMeans
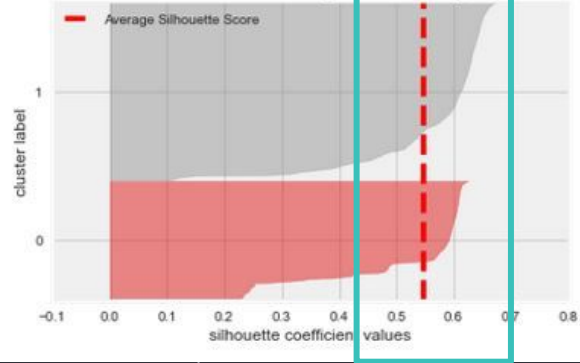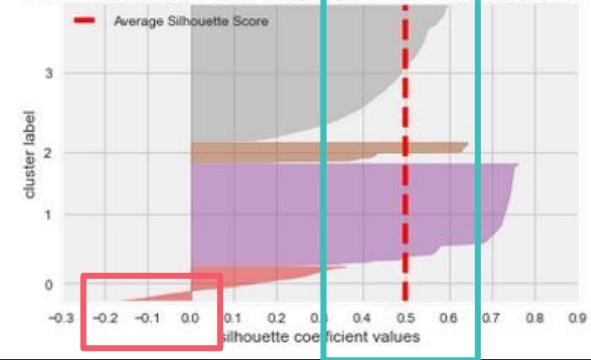
Silhouette score :

Considers both the average intra-cluster distance and average inter-cluster distance. A score close to 1 means that clusters are well apart from each other and clearly distinguished.

- Positive silhouette coefficient values
- Each individual cluster having silhouette scores above the average
- **3** centers/clusters have the highest silhouette score of **~0.61**

Clustering - DBSCAN

○ Well-suited for discovering data with arbitrary shapes

```
+-----------------------------------------------------------------------+
|             Silhouette Score on parameters and clusters               |
+-----------------------------------------------------------------------+
| Number Of Clusters | Param - Minimum Samples | Silhouette Score |
+-----------------------------------------------------------------------+
|         35         |            3            |      0.282       |
|         13         |           10            |      0.271       |
|          6         |          100            |      0.121       |
+-----------------------------------------------------------------------+
```

○ Silhouette score is significantly lower than KMeans
○ Suggests that dataset comprise data points with varying density, resulting in very high number of clusters, with very low silhouette score.

Clustering - Agglomerative Clustering

◦ Works in a "bottom-up" manner whereby each object is initially considered as a single element cluster (leaf)
◦ At each step of the algorithm, two clusters that are most similar are combined into a bigger cluster (nodes)
◦ Process is repeated until all points are a member of a single big cluster

```
+--------------------------------------+
|        Silhouette Score on Clusters  |
+--------------------+-----------------+
| Number Of Clusters | Silhouette Score |
+--------------------+-----------------+
|         2          |      0.5429      |
|         3          |      0.61        |
|         4          |      0.4698      |
|         5          |      0.4746      |
|         6          |      0.481       |
|         7          |      0.4279      |
|         8          |      0.4399      |
|         9          |      0.4418      |
|        10          |      0.4637      |
+--------------------+-----------------+
```

◦ **3** clusters are optimal with silhouette score of **0.61**

◦ Results are similar to KMeans

Clustering - Conclusion

KMeans
- 3 clusters with 0.61 silhouette score
- Neighbouring clusters (2 clusters and 4 clusters) have higher silhouette scores than Agglomerative

DBSCAN
- 35 clusters with 0.282 silhouette score

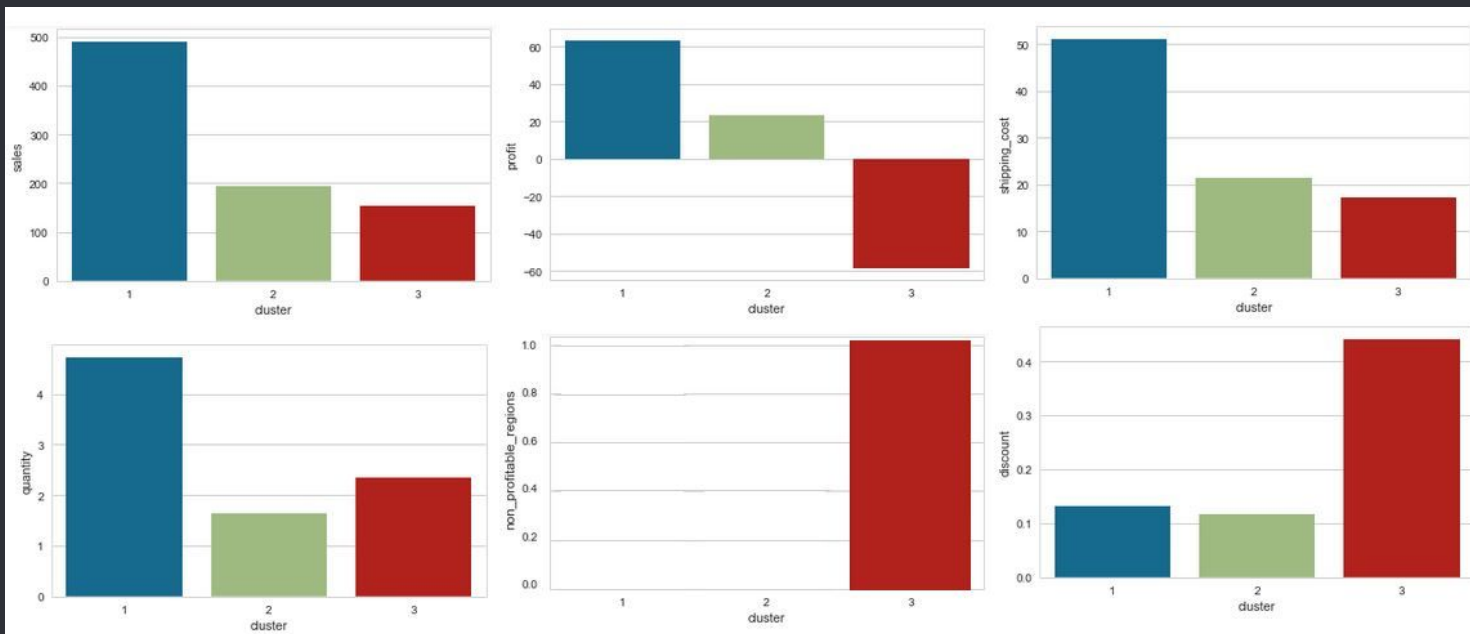### Agglomerative
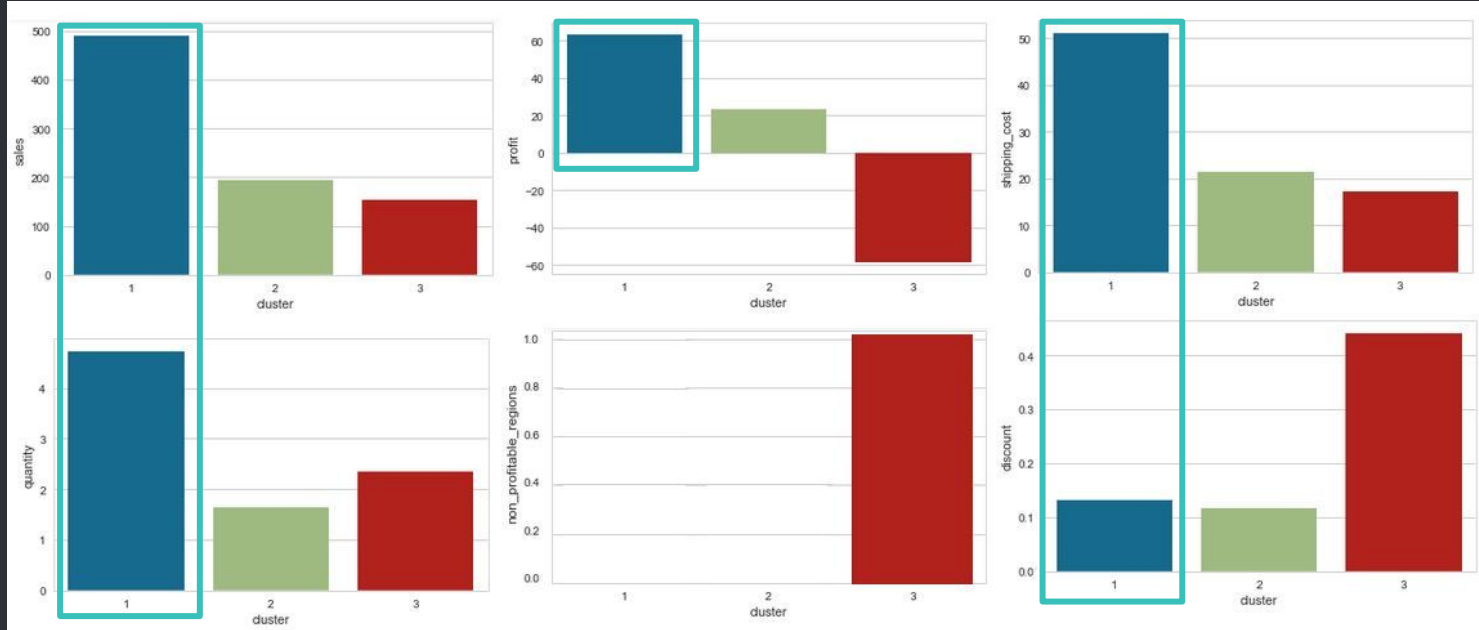- 3 clusters with 0.61 silhouette score

**Selected Model :
KMeans**

Clustering - Characteristics

- Apply cluster labels from KMeans on dataset
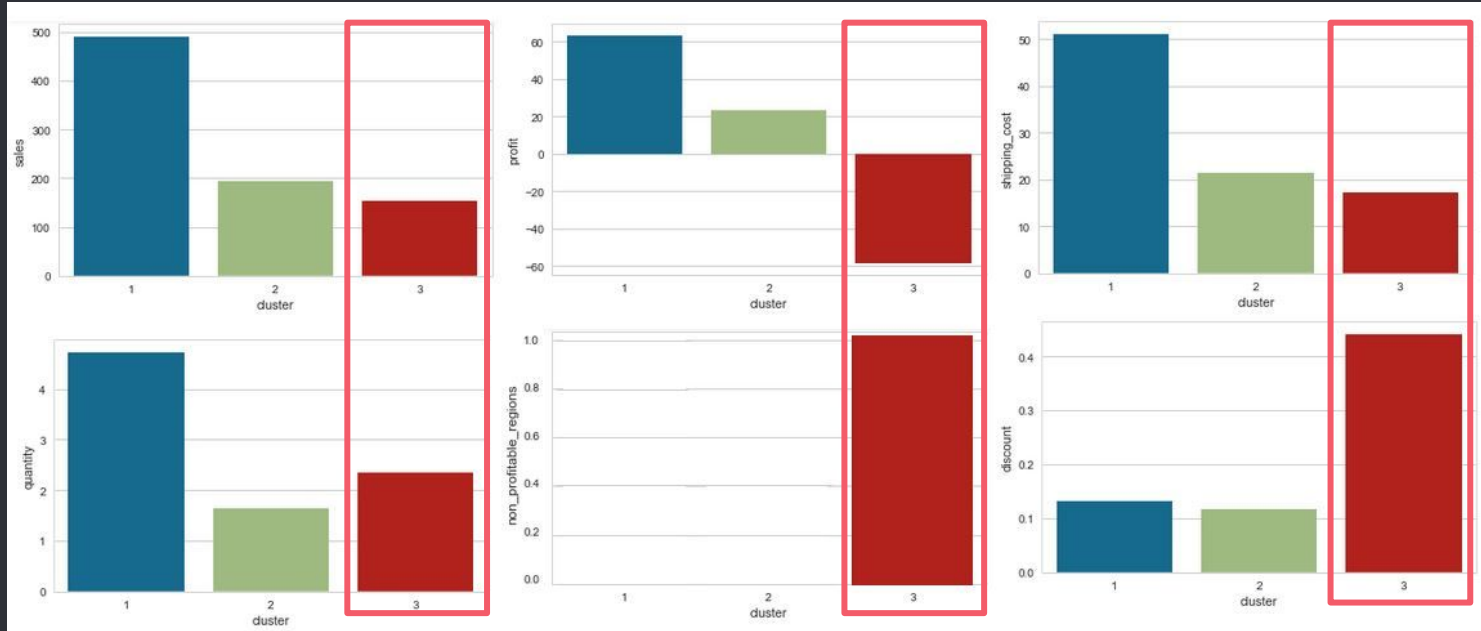- Identify characteristics

Clustering - Characteristics

Cluster 1

- Highest sales and most profitable
- Most quantity purchased (>3)
- Relatively low discounts (~0.1%)
- Excludes transactions from unprofitable regions

Clustering - Characteristics

Cluster 1	**Cluster 2**	**Cluster 3**
◦ Highest sales and most profitable	◦ Moderate sales and profit	◦ Unprofitable, lowest sales
◦ Most quantity purchased (>3)	◦ Lower quantities (<3) purchased	◦ Lower quantity (<3)
◦ Relatively low discounts (~0.1%)	◦ Relatively low discounts (~0.1%)	◦ High discounts (>0.3%)
◦ Excludes transactions from unprofitable regions	◦ Excludes transactions from unprofitable regions	◦ Non-profitable regions : Western Africa, Western Asia and Central Asia

Multivariate Regression

◦ Perform multivariate regression on each cluster
◦ Nested cross-validated R2 score to select best algorithm

◦ Generally tree-based algorithms and support vector regressor work better

Multivariate Regression

- Perform multivariate regression on each cluster
- Nested cross-validated R2 score to select best algorithm

Cluster 1

```
+-----------------------------------------------------------+
|                 Cross-validated R2 score                  |
+-----------------------------------------------------------+
|  Algorithms            | R2 Score | Standard Deviation |
+-----------------------------------------------------------+
|  Elastic Net CV        |  88.3%   |      +/- 2.1%      |
|  Lasso                 |  88.4%   |      +/- 2.1%      |
|  Random Forest Regressor |  96.0% |      +/- 1.2%      |
|  Support Vector Regressor |  88.2% |      +/- 2.4%      |
|  XGBoost Regressor     |  99.0%   |      +/- 0.4%      |
+-----------------------------------------------------------+
```

Cluster 2

```
+-----------------------------------------------------------+
|                 Cross-validated R2 score                  |
+-----------------------------------------------------------+
|  Algorithms            | R2 Score | Standard Deviation |
+-----------------------------------------------------------+
|  Elastic Net CV        |  92.4%   |      +/- 2.7%      |
|  Lasso                 |  92.5%   |      +/- 2.7%      |
|  Random Forest Regressor |  94.8% |      +/- 0.6%      |
|  Support Vector Regressor |  92.4% |      +/- 2.4%      |
|  XGBoost Regressor     | 97.399999999999999% |  +/- 1.2% |
+-----------------------------------------------------------+
```

Cluster 3

```
+-----------------------------------------------------------+
|                 Cross-validated R2 score                  |
+-----------------------------------------------------------+
|  Algorithms            | R2 Score | Standard Deviation |
+-----------------------------------------------------------+
|  Elastic Net CV        |  71.6%   |      +/- 10.5%     |
|  Lasso                 | 70.899999999999999% |  +/- 11.1% |
|  Random Forest Regressor |  90.5% |      +/- 2.5%      |
|  Support Vector Regressor |  72.7% |      +/- 9.6%     |
|  XGBoost Regressor     | 92.800000000000001% |  +/- 1.3% |
+-----------------------------------------------------------+
```

- Generally tree-based algorithms and support vector regressor work better

- For interpretive value, I select the better of regularised regression model for each of the clusters
    - Cluster 1 : Lasso
    - Cluster 2 : Lasso
    - Cluster 3 : Elastic Net

- Obtain the coefficients and y-intercepts, to plot the regression line

Conclusion and Recommendation

◦ How profit per unit changes with unit price

Regression Line of Clusters
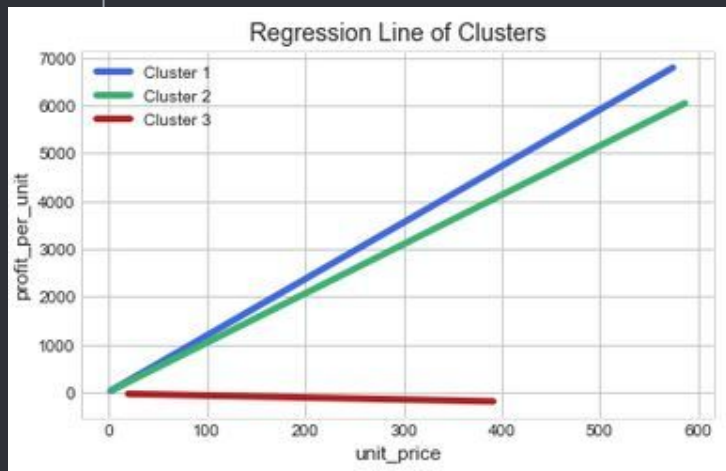
◦ Avoid selling phones at Western Africa, Western Asia, or Central Asia as all phone sales here are not profitable at any unit price

◦ Focus on selling phones in other regions

Conclusion and Recommendation

◦ How profit per unit changes with unit price

Regression Line of Clusters
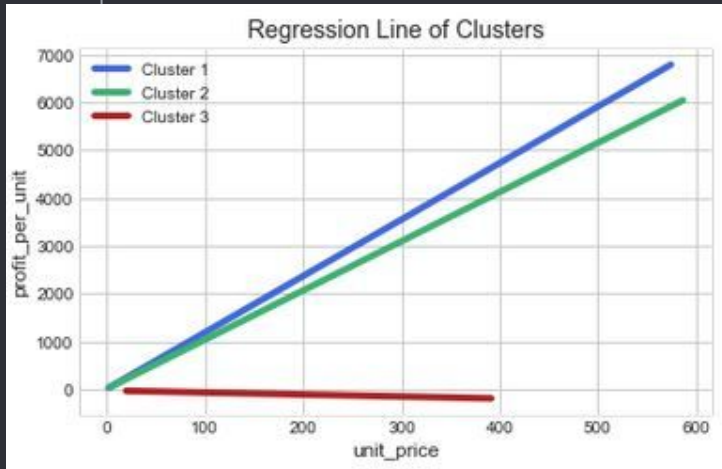
◦ Between Cluster 1 and Cluster 2, Cluster 1 comprise transactions with more quantities purchased than Cluster 2. These are probably corporate consumers.

◦ Target corporate consumers and consider providing bulk discounts or include free phone accessories if more phones are purchased.

Further Development

Customer Segmentation

- With customer details (eg. propensity to spend, age, income), can consider grouping by customers to provide more insights into consumer purchase behaviour
- Targeted marketing to each customer cluster

Breadth and/or Depth

- Analyse deeper within the product subcategory (eg. into specific phone brands, or specific phone models)
- Analyse other categories outside of phones

Limitations and Challenges

Nature of setting prices

◦ Attempted to price a product based purely on historical transactional data but pricing is as much art as it is science
◦ Several other external factors that influence prices eg. branding, marketing and advertising.

Scaling and Monitoring

◦ Consumer purchase behaviour changes over time and so there has to be consistent monitoring
◦ Scaling may be an issue as more consumers are added to the database, much more processing power is needed to handle huge amounts of data.

Thank You

Q&A