**Product value modelling of summer clothes in E-commerce Wish**

**Final Report**

**University of Texas at Arlington**

**BSTAT 5325 -003**

**Prof. Patricia Swafford**

**Group 11**

(*Devi Prasad Chitti Pothula, Jitesh Khemchandani, Bharath Kumar Reddy Peram, Faria Tasnim*)

Foreword

This report provides an insight on how advanced statistics helps businesses to increase their profitability / Revenue when they use their data. The data is collected from an e-commerce website called Wish for a period of one month (Aug – 2020). We have applied few statistical techniques to model an equation to determine a product is profitable or not with a quantitative value.

We have used clustering, Multiple Regression and Logistic regression analysis in this report.

Please note due to unobtainability of few data points, we have taken few assumptions to determine a product is profitable or not.

**Group 11 Students,**

**BSTAT 5325, Sec-003.**

Contents

|  |  |  |
| --- | --- | --- |
| Section | Heading | Page Number |
| Data Overview | Description of Dataset  Identification of Qualitative and Quantitative Variable | 4 |
| Cleaning and Preparing data for EDA |  |
| Exploratory Data Analysis | Statistical description of Dataset | 7 |
| Validating the correlations between variables  Finalizing the variables to run the models |  |
| Model Fitting | Cluster Analysis | 11 |
| Regression Analysis | 12 |
| Logistic Regression | 14 |
| Model Output | Equation for product value | 15 |
|  |
|  |

Executive Summary

Concentrating on top items requires something other than item postings. You additionally need to realize what sells well and what doesn't.

We have chosen e-commerce data set which is collected from website called Wish for a period of one month and we have performed few analyses to identify the key components which impacts the sales which eventually contribute to revenue.

In addition to above, we have also identified an equation which sellers can use to get the revenue potential value of product using different variables and also an equation to identify high value product or non-high value product.

**Data Overview:**

Our chosen dataset is collected from Kaggle website under the title “Sales of summer clothes in E-commerce Wish” – ([Kaggle](https://www.kaggle.com/jmmvutu/summer-products-and-sales-in-ecommerce-wish)) and describes the sales of summer clothes in E-commerce according to their user rating, product price and merchandise rating. This data set from the year 2020 contains 1574 observations in total and 43 different variables which we can divide into two groups: product level data and merchandise level data.

Here are the main variables which is going to determine our statistical analysis in below table:

|  |  |
| --- | --- |
| **Variables** | **Description** |
| title | is something with an information about the product with which one can check list through e-commerce. |
| price | price is the money the customers must pay to purchase. |
| retail\_price | is the final price depending upon the supply and demand conditions at which the product is sold to customer. |
| units\_sold | no. of products sold in a period of time. |
| uses\_ad\_boosts | Gives information whether ad booster is used or not |
| rating | Average rating of product |
| rating\_count | total number of ratings given on product |
| rating\_five\_count | # of people given rating 5 |
| rating\_four\_count | # of people given rating 4 |
| rating\_three\_count | # of people given rating 3 |
| rating\_two\_count | # of people given rating 2 |
| rating\_one\_count | # of people given rating 1 |
| badges\_count | Count of special badges added on this product to attract customers to purchase the product |
| badge\_local\_product | Products which are manufactured locally |
| badge\_product\_quality | Products with high quality |
| badge\_fast\_shipping | Products with fast shipping facility |
| tags | Tags on products which helps in searching the product |
| product\_color | visual representation of the product's main colour |
| product\_variation\_size\_id | available sizes of the product. |
| product\_variation\_inventory | is the number of variations of the product available at the seller |
| shipping\_option\_name | available different shipping options |
| shipping\_option\_price | charges for shipping the product |
| shipping\_is\_express | whether the shipping is express or not. Standard delivery services. |
| countries\_shipped\_to | no. of countries the product is shipped to |
| inventory\_total | total inventory stocked by the seller against each item |
| has\_urgency\_banner | Does the product got any urgency banner? |
| urgency\_text | "Limited quantity!"; "Discount on wholesale purchases" |
| origin\_country | country where the product was made |
| merchant\_rating\_count | # of ratings given to the customer |
| merchant\_rating | consolidated for a particular merchant |
| merchant\_has\_profile\_picture | Convenience Boolean that says whether there is a `merchant\_profile\_Picture` |

Though many columns are text values and not useful for the analysis, we have identified below columns which we are planning to consider for analysis.

|  |  |
| --- | --- |
| **Variables** | **Description** |
| price | price is the money the customers must pay to purchase. |
| retail\_price | is the final price depending upon the supply and demand conditions at which the product is sold to customers |
| units\_sold | no. of products sold in a period of time. |
| uses\_ad\_boosts | Gives information whether ad booster is used or not |
| Rating | Average rating of product |
| rating\_count | total number of ratings given on product |
| badge\_local\_product | Products which are manufactured locally |
| badge\_product\_quality | Products with high quality |
| badge\_fast\_shipping | Products with fast shipping facility |
| tags | Tags on products which helps in searching the product |
| product\_color | visual representation of the product's main color |
| product\_variation\_size\_id | available sizes of the product. |
| merchant\_rating | consolidated for a particular merchant |



In order to perform analysis on tags, product color and size columns, we have modified the data by adding columns with 0 and 1 value for the values present in those columns. Tag’s and color column are added for the each tag which is top 10 based on frequency. Size column is modified by clubbing the sizes to S,M and L.





For the Analysis purpose, we have created few extra columns using price and retail price and units sold.

**Discount­\_%** column data is created using (retail\_price – price)/retail\_price.

**Revenue** column data is created using price\*units\_sold.

In order to identify the profit on each product we have taken few assumptions around profit margin on product based on discount\_%.

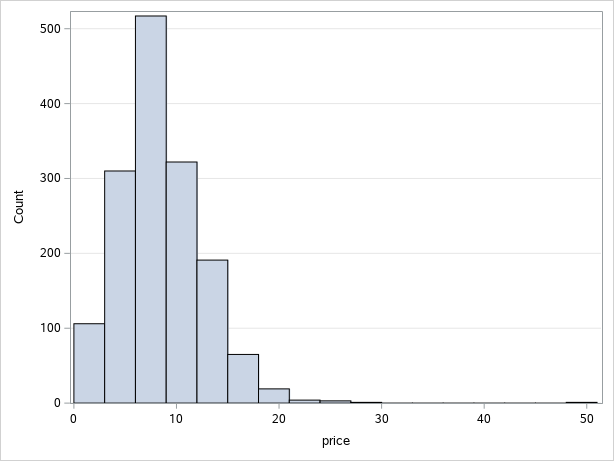
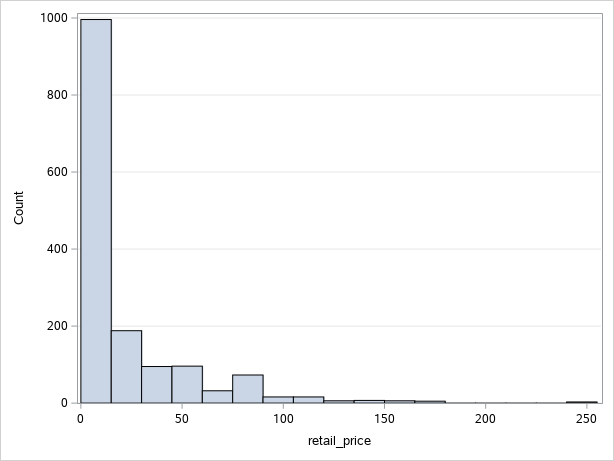
Below table provides the assumed margin % on each product based on discount %.

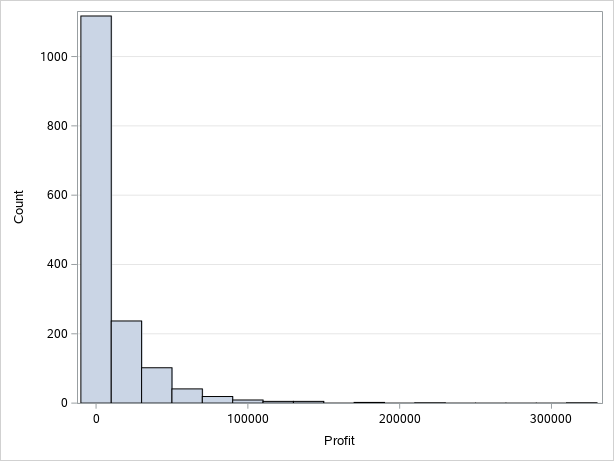
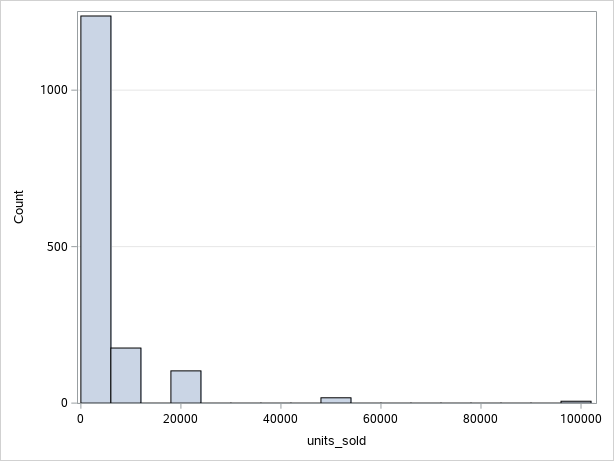


**profit** on each product in that particular month is created using Revenue\*profit margin%.

**high\_value\_product** column is created by taking the assumption that top 20 percentile products in terms of profit.

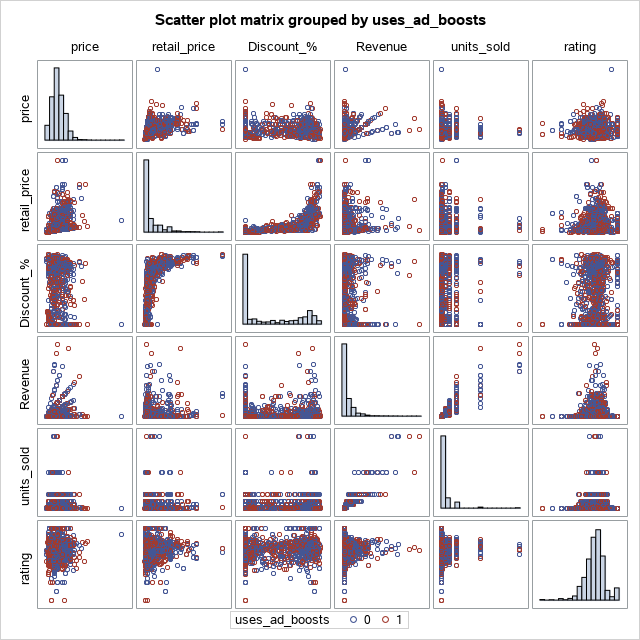
**Exploratory Data Analysis:**

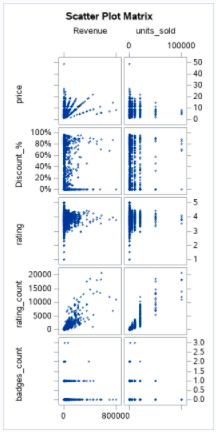
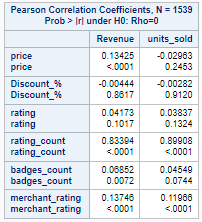
 

Distribution of ‘price’ data clearly shows the most of the products fall in the range of 0 – 20 EUR and retail-price are below 100 EUR.

Profit and units\_sold in the month of August on each product distribution is shown above.



Based on the discount % and rating plot, we can create a clusters analysis which might show a significant impact in our analysis.

We see a good correlation between Revenue and rating\_count and units\_sold, which is expected as more number of units sold leads to more customers which results in more reviews on products.

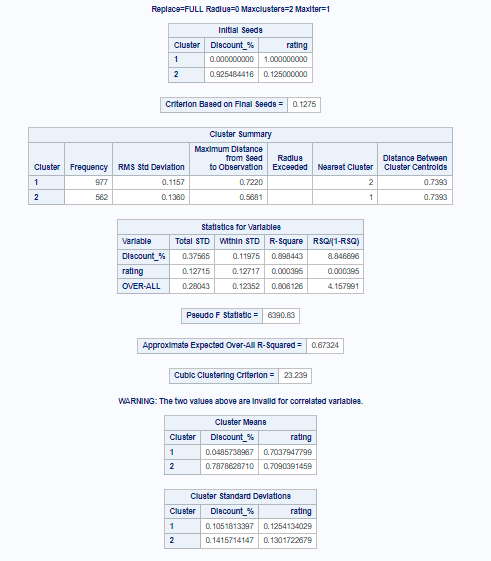
Even though we have created many variables with tag, color and size, we are eliminating all red color highlited columns from our analysis due to no significant distribution.



**Model Fitting:**

**Clustering:**

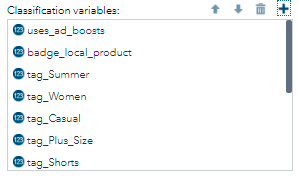
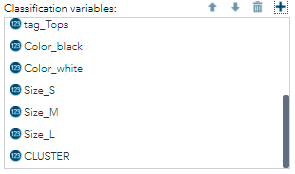
As we found in the EDA, we trying to divide data into three clusters using K-means clustering method based on Discount % and rating. Below are the output:

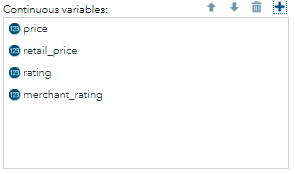


We have divided data into two clusters with frequency of 977, 562.

**Multiple Regression:**

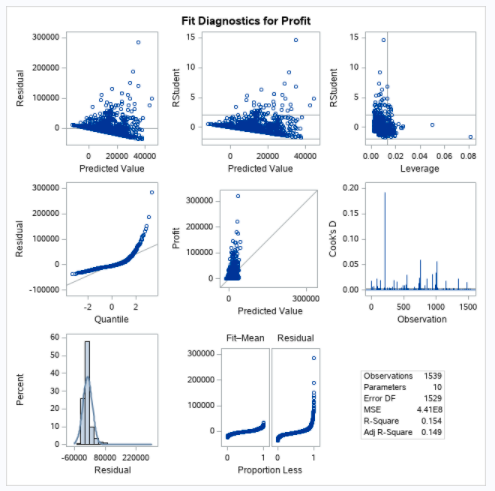
Now we will try to generate a equation for product potential value in Eur per month. In order to achive this, we have selected below Classification and Continuous varibles and ran multiple regression with 95% confidence interval.



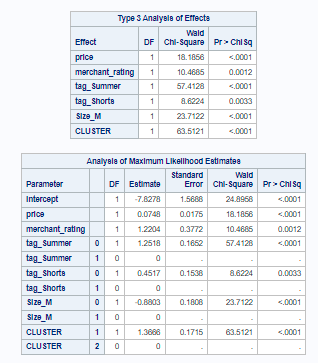
The output of multiple regression is shown below:





**Logistic Regression:**

We ran logistic regression to get an equation to identify whether a product fall under is high potential value product. Below are the results:



**Model Output:**

**Equation of Product potential value:**

**For Cluster 1:**

Product Potential Value = 624.0433\*price + 10335\*merchant\_rating + 14153\*tag\_summer + 3591.4516\*tag\_shorts - 2240.373\*tag\_tops - 4528.2934\*Size\_s – 13511\*size\_M – 8035.5587\*Size\_L + 9213.8524-24127

**For Cluster 2:**

Product Potential Value = 624.0433\*price + 10335\*merchant\_rating + 14153\*tag\_summer + 3591.4516\*tag\_shorts - 2240.373\*tag\_tops - 4528.2934\*Size\_s – 13511\*size\_M – 8035.5587\*Size\_L -24127

**Equation to Predict High potential value product:**

**For Cluster 1:**

High\_potential\_product = 0.0748\*price + 1.2204\*merchant\_rating + 1.2518\*tag\_summer + 0.4517\*tag\_shorts-0.8803\*Size\_M + 1.3666 – 7.8278

**For Cluster 2:**

High\_potential\_product = 0.0748\*price + 1.2204\*merchant\_rating + 1.2518\*tag\_summer + 0.4517\*tag\_shorts-0.8803\*Size\_M – 7.8278

Above equations can be used by sellers while deciding their inventory for selling.