Case Study 2

CAB 330

Jack Farrel – N8640866

Anton Polison – N9907637

Dennis Tran – N9937072

2017

Table of Contents

[Task 1: Data Selection and Distribution. 2](#_Toc492835509)

[Task 2: Predictive Modelling Using Decision Trees 5](#_Toc492835510)

[Task 3: Predictive Modelling Using Regression 8](#_Toc492835511)

[Task 4: Predictive Modelling Using Neural Networks 11](#_Toc492835512)

[Task 5: Comparing Predictive Models 13](#_Toc492835513)

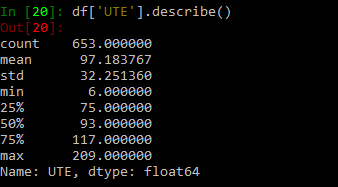
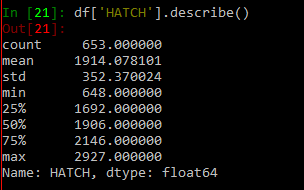
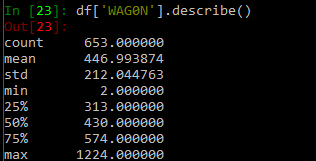
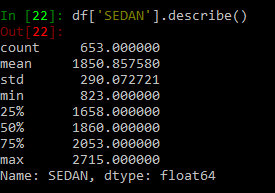
[Appendix A 16](#_Toc492835514)

# Task 1: Descriptive Data Mining - Clustering.

1. There were some changes made to the original dataset to ensure that the information from the analysis was accurate. There are several quality problems with “model\_car\_sales.csv” data set such as missing values and redundant or irrelevant variables
   1. ***Missing Values:*** The dataset contains 22 rows which have empty values. These values were deleted as they are irrelevant for our research.
   2. ***Dropped Variables:*** In total 4 variables were dropped due to various reasons. The below table shows the dropped variables and their corresponding reasons.

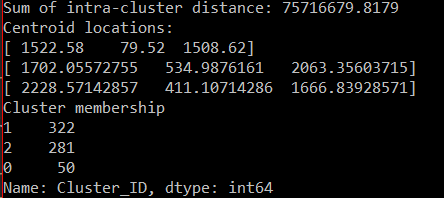
|  |  |
| --- | --- |
| *Variable* | *Reason For Drop* |
| *REPORT\_DATE* | Irrelevant. The variable has the date when the report was generated and doesn’t provide any useful information. |
| *DEALER\_CODE* | Redundancy. “LOCATION” variable provides more organised information which is at the same time easier to work with as it is numerical variable. |
| *K\_SALES\_TOT* | Irrelevant. The company wants to find the minimum number of product sale segments and it is not interested in identifying what is a number of total car sales. |

1. After preprocessing and analyzing the dataset it was identified that “UTE” is clearly underperforming in sales in comparison with “HATCH”, “SEDAN” and “WAGON”.

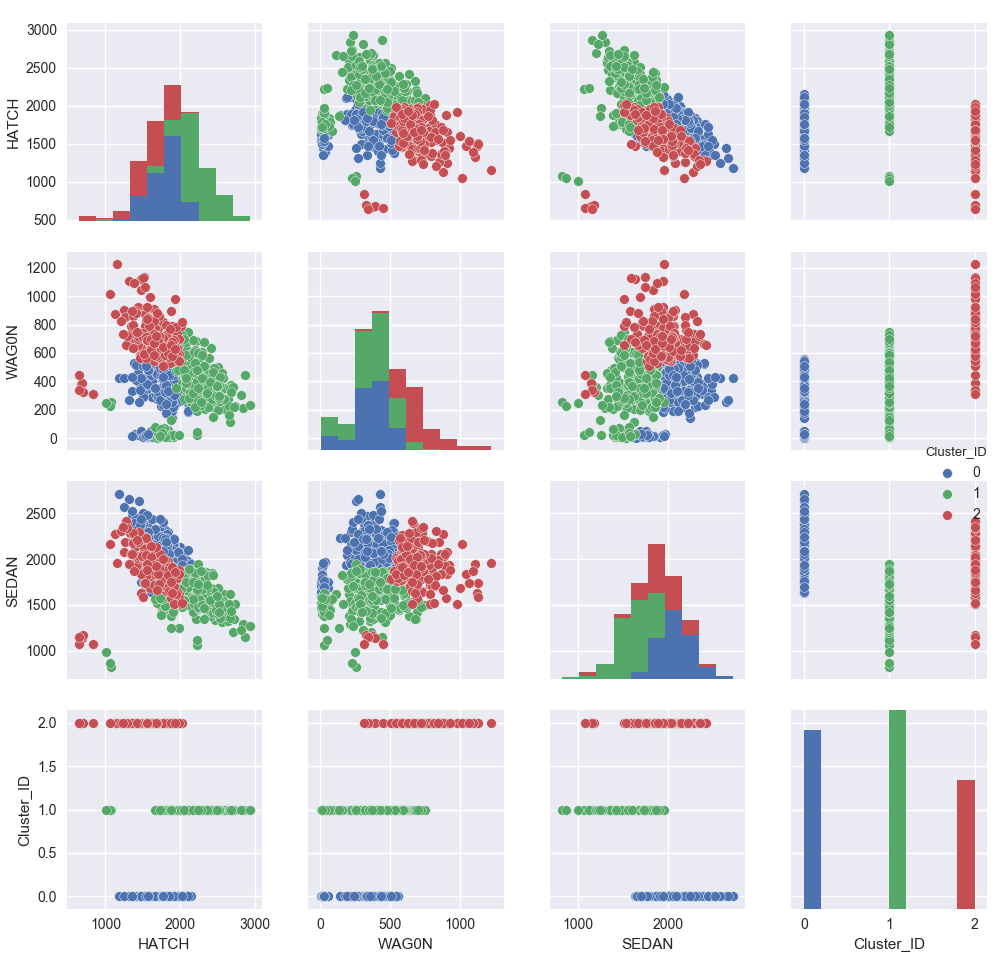
Form the tables above we can determine that “mean” of the variables “HATCH”, “WAGON”, “SEDAN” is significantly higher than of “UTE” variable.

1. “HATCH”, “SEDAN”, “WAGON” and “LOCATION” variables were included in the analyses.
2. a)



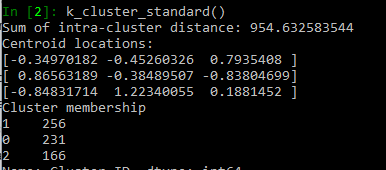
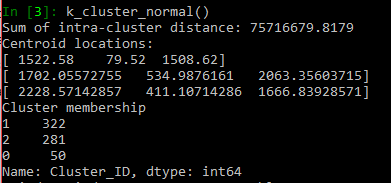
From cluster membership table we can identify that 231 records were assigned to **cluster 0**, 256 records to **cluster 1** and 166 records to cluster 2.

b)



The pairplot shows the sales distribution on different variables.

* If we look at the first row first column we can notice that **cluster 0** has average to high sales. cluster 2 has low to average sales and finally, **cluster 1** has high sales.
* Wagon variable distribution is the following: **cluster 0** has low to average sales, cluster 2 has average to high sales and **cluster 1** has low to average sales with a slight win over **cluster 0.**
* SEDAN variable has the following distribution: **cluster 0** has high sales, **cluster 1** has low to average sales and cluster 2 average to high sales.

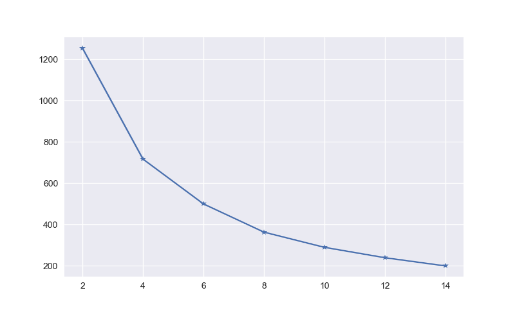
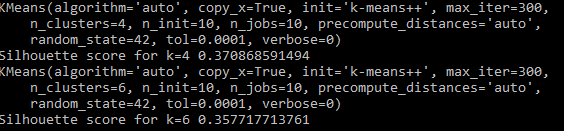
 

Normal distribution Standardization

The standardization made dramatic changes to our model. First and most obvious change is that intra-cluster distance was significantly reduced by standardization which is one of the main purpose of clustering. Furthermore, cluster membership changed as well “cluster 0” 50 =>231, “cluster 1” 322=>256, “cluster 2” 281 => 166.

If we look at the “pairplots” we can see 2 absolutely different results due to increased “cluster 0” membership and changed intra-cluster distance.

As we can see, all the above-mentioned facts proving the fact that standardization brings better clustering solution and better visualization outcome.

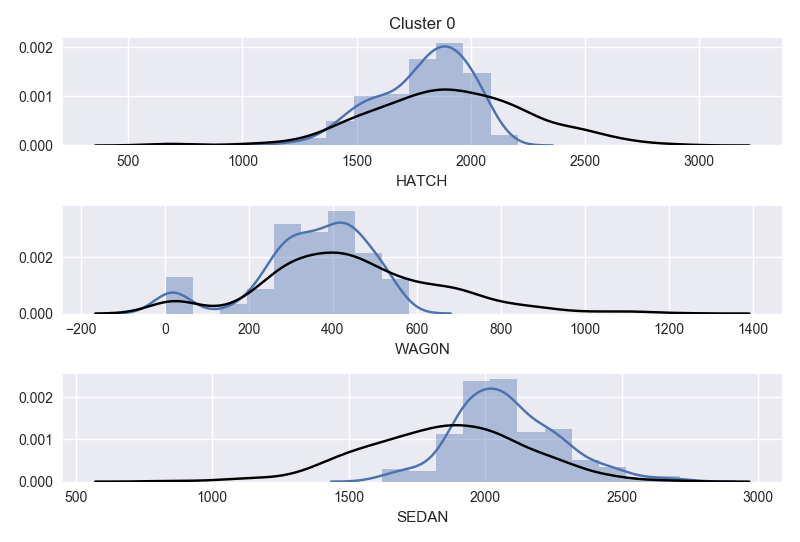
  “Elbow Method” “Silhouette Method”

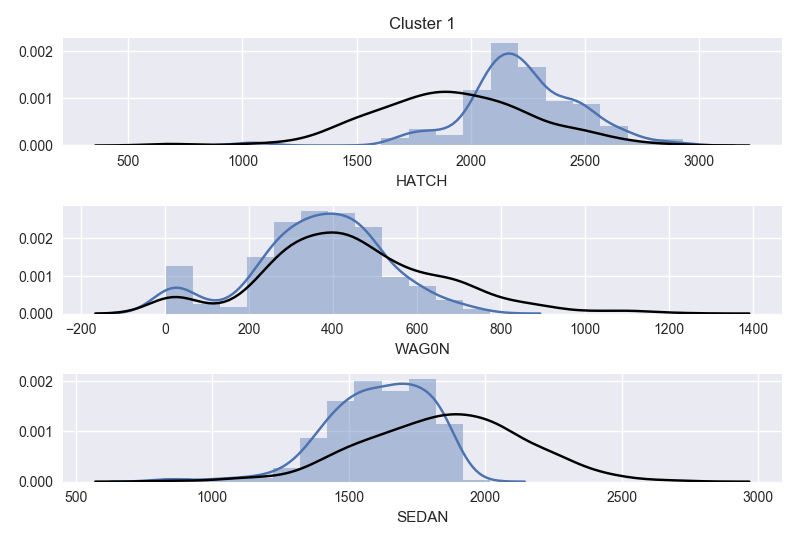
First, we look at the elbow method to determining better K. The graph shows that our best K is between 4 and 6. Though, this method doesn’t always provide with the best K.

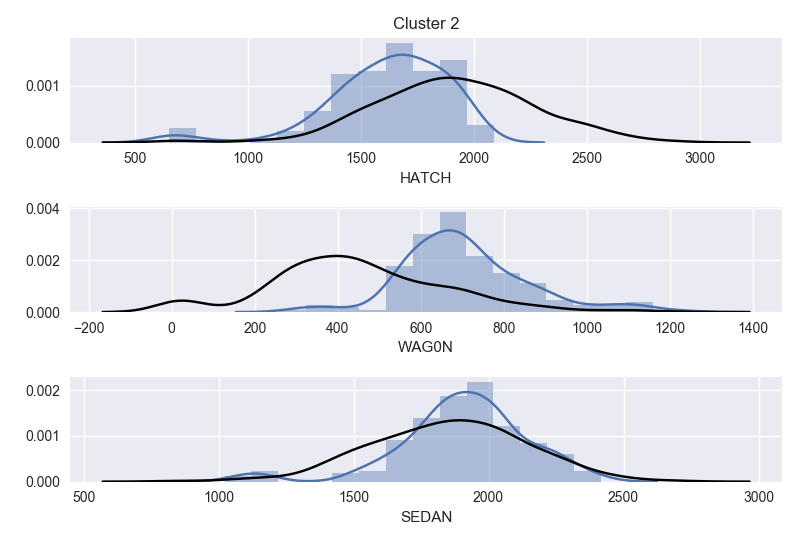
“Silhouette Method” is a more precise technique in identifying best K. It measures how close each point in a cluster to the point in the neighborhoodlike cluster. Silhouette value lies in the range [-1, 1]. One of the main goal of clustering is to minimise inter-cluster similarity. Therefore, higher the value better the cluster configuration. In our case, k=4 has a better inter-cluster similarity (~0.3709) in comparison with k=6 which has (~0.3577). Based on this method we can state that k=4 is the best model.

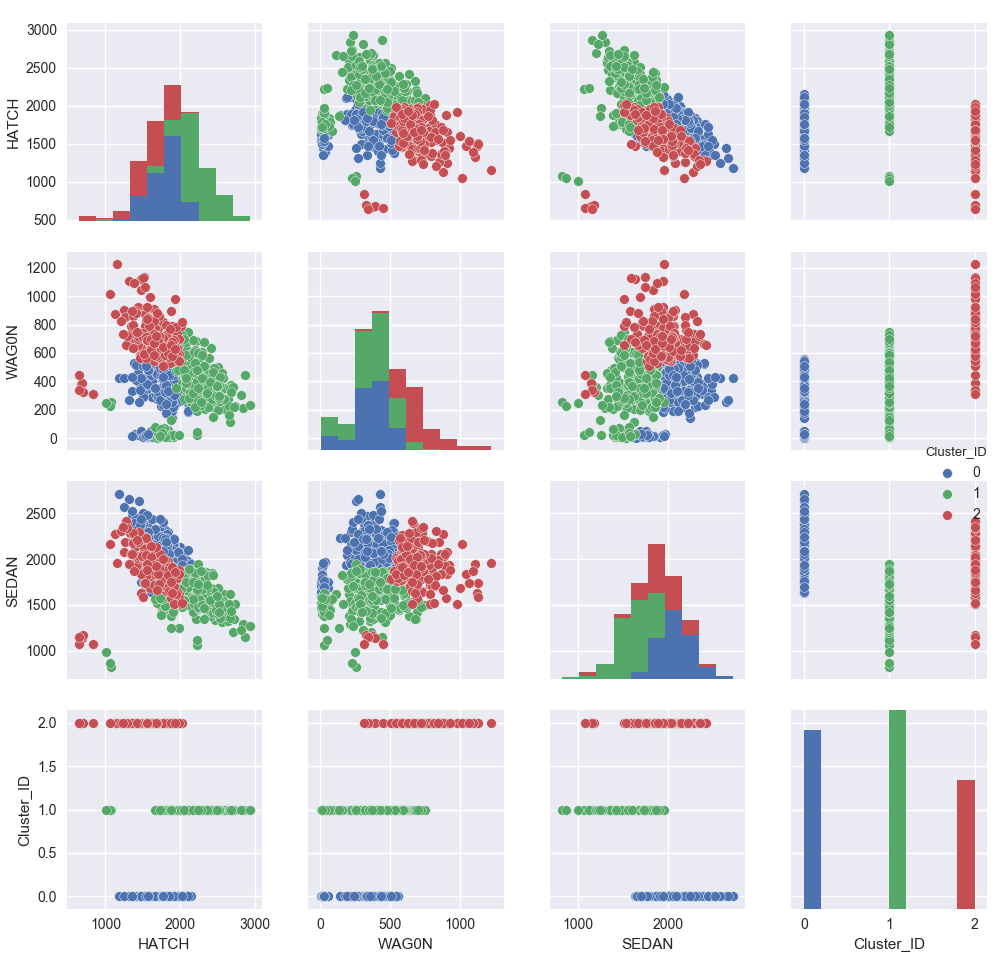
However, even though K=4 is the best model for this task we will use K=3 in our research as it’s easier to interpret and provide more distribution understandably.











After careful examination of all distplots and pairplot with K=3 we can sum up our findings.

**Cluster 0**: has mostly average sales for HATCH. At the same time WAGON has low to medium sales. On the other hand, SEDAN has average to high sales.

**Cluster 1**: showing that sales for HATCH is medium to high, WAGON has low to medium sales and SEDAN is sold on an average level.

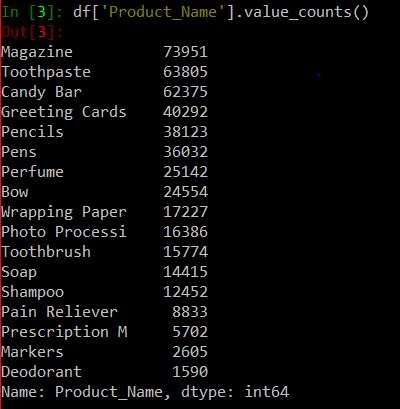
**Cluster 2**: sales on HATCH are from low to medium, WAGON has medium to high sales and SEDAN has medium to high sales.

# Task 2: Descriptive Mining – Association

|  |  |
| --- | --- |
| *Variable* | *Importance* |
| *Location* | Irrelevant. The variable has no use to our research |
| *Transactin\_Date* | Irrelevant. The transaction date’s values are similar and of no use. |
| *Quantity* | Irrelevant. Each row of Quantity variable has a value of “1”. |

1. “pos\_transaction.csv” doesn’t have obvious erroneous or missing values. Some of the variables were dropped due to irrelevancy in our research.

1. “Transaction\_Id” and “Product\_Name” were 2 most valuable variable which we were using to perform association mining. “Transaction\_Id” variable include the transaction ids which group the products under the particular transaction. At the same time “Product\_Name” attribute has the following values:



3.



* 1. The table is illustrating the association rules between different items. It’s noticed that 2 resulting rules have the highest lift of 3.601370. They are “**Perfume => Toothbrush**” and “**Toothbrush =>Bow**” with lift level of 3.081236.
  2. Furthermore, it is identified that the highest confidence value has “**Toothbrush=> Perfume**” rule with the value 0.323979.
  3. If we plot together with the confidence, lift and support we get the following scenario. The customers who buy **Perfume** are likely to purchase **Toothbrush** with it. The lift of this rule is quite high which denotes strong association rules between these two products. On the other hand, the rule “**Toothbrush => Perfume**” has the same lift level as “**Toothbrush => Perfume**” but higher confidence level implying that if there is **Toothbrush** in the transaction it’s likely that **Perfume** will be in the transaction as well. Further, we can observe interesting rules is **“Toothbrush => Bow”** with a quite a high lift level of 3.081236 but if we take a closer look at the confidence level it appears to be significantly lower than confidence level of other association rules with the same lift level. However, from the business prospective this rule is “interesting” as it has a positive correlation (positive lift value). Furthermore, there are a couple of other association rules which can attract business’s attention such as **“Candy Bar, Magazine => Greeting Cards”**, **“Pencils, Toothpaste => Candy Bar”**, **“Magazine, Toothpaste => Greeting Cards”**. Lift level is approximately the same around 2.7 for these rules. Also, it can’t go without saying that confidence level for these rules is almost twice higher than for the rules with higher lift.

1. a. From the table above, we can notice 4 association rules for items which are bought together with “Candy Bar”. They are **“Pencils, Toothpaste => Candy Bar”**, **“Magazine**, **Greeting Cards => Candy Bar”**, **“Magazine, Toothpaste => Candy Bar”**, **“Toothpaste, Greeting Cards => Candy Bar”**.

b. If we look closer to the above rule, we can state that the association rule **“Pencils, Toothpaste => Candy Bar”** has the highest Lift and confidence level. It means that the customers who bought pencils and toothpaste will likely to buy candy bar as well. It would be a wise business decision for the store to place these items close to each other to increase their sales. From the question 4.a. we notice that **Magazine** and **Greeting Cards** are also frequently bought with Candy Bars.

# 