ANALYSIS OF AUTO-INSURANCE CUSTOMERS USING K-MEANS CLUSTERING

BY

RISIKAT HAMEED

VIVIAN OGWUIHE

TABLE OF CONTENTS

Introduction

1.0 Background

According to Statista.com, the share of net premium written in Canada for vehicle insurance between 1990 and 2020 is 26.96 billion CAD, a figure that is always rising due to current inflation. Furthermore, the net claim incurred by vehicle insurance for the same period is 17.86 billion CAD, which is more than 65 percent of the premium received during the same period. This does not appear to be a profitable venture. Many vehicle insurance customers complain of excessive rates owing to the usage of postal code information and other irrelevant driving attributes to compute premiums in the middle of these high premiums and claims.

Although there seem to be a whole lot of intricacies, conflicting opinions and legalities involved in other technological savvy and accurate ways of calculating car insurance premiums, such as Telematics which involves the use of diagnostics and GPS to record movements on a computerized map, there is a need for a fairer, more robust and more transparent model of premium calculation in the car insurance industry to increase profitability in the auto insurance industry as well as increase consumer satisfaction.

By completing this project, insurance companies will have a better model that will help in underwriting auto insurance premiums other than relying on postal codes which lumps very different customers together. Since this project aims at grouping only similar customers together, their unique characteristics will help determine which type of customer to classify as high risk, low risk and medium risk, thereby reducing the unfairness of determining premiums through postal codes and other irrelevant driving information.

2.0 Problem Statement

Customer service as an essential component of insurance administration involves the use of fair, transparent and less complex methods to determine car premiums. Also, Auto Insurance rate evasion is a mounting problem in different countries. Rate evasion is providing wrong information in order to obtain a cheaper insurance rate.

In Canada, Aviva Canada, an insurance company reported that their data shows 9.1% address misrepresentation claims and almost 7% national address misrepresentation claims. Misrepresentation is when the wrong home address is provided to get cheaper rates.

As a result, Insurance companies must determine the critical success criteria in offering auto-insurance policies for their client's happiness by using only relevant driving and personal information in calculating premiums. Customers should not be charged high auto insurance just because of their choice of where to live or their postal codes.

3.0 Objectives & Measurement

This project will segment insurance customers based on variables that are deemed important in determining the risk level of auto insurance customers. Since auto insurance premium calculations have remained competitive amongst insurance companies as customers continue to grow more sensitive to premiums, customer retention and acquisition are critical for maintaining a competitive advantage in the industry. This project will help the auto insurance firms to build a strong model for underwriting based on only important and relevant information for high customer satisfaction.

Furthermore, this project will assist in determining the coverage that should be offered based on specific traits or commonalities discovered in distinct consumer groups or segments.

The variables used as relevant or important for fair, transparent and cost effective auto premium calculations will be determined using Decision Tree and its accuracy should be more than 80%. Also, the final model produced from this project will be measured based on the similarities and dissimilarities of individual characteristics found in each segment or cluster produced.

4.0 Assumptions and Limitations

Segmentation of customers, in particular, clustering analysis using KMeans that involves continuous, discrete or interval variables arguably produces a more interpretable result than the one with more categorical features, else, a more in-depth statistical model may be used. Our dataset had lots of categorical variables that are grouped together with little or no natural ordinal relationship between them.

Also, although our dataset was obtained from a public site (kaggle.com) with a few parent model projects, none of those projects did a cluster analysis using the data, hence, our knowledge on the efficacy of our model choice using this dataset is limited.

Furthermore, we could not get the year our dataset was created and as a result, our knowledge on what effects a 'data shift' might cause due to the age of our dataset is somewhat limited. We have made the following assumptions in the course of this project:

- The datasets we used for this project is correct and meets all quality standards including method of data collection
- The property and casualty insurance sector is willing to adjust premium calculation methods as a result of the outcome of this project

- Our model accurately captures all categories of insurance customers.
- Available dataset is similar to what is obtainable in the Canadian Auto Insurance industry and as a result, the model is reproducible in the aforementioned industry.

Data Sources

5.0 Dataset Introduction

The dataset used in this project is a car insurance claims dataset and was sourced from a public website kaggle.com (https://www.kaggle.com/sagnik1511/car-insurance-data).

The dataset contains 10,000 observations and 19 features. It is a multivariate dataset with more categorical features than numerical features. We found that past projects done using this dataset were mostly on car claims prediction and claims classification. Also, while going through past works done using this dataset, we found that some parts of this dataset were generated by a Kaggle user by the moniker 'sagnik1511', aside this, the dataset is real data from an auto insurance company.

6.0 Exclusions

We excluded three variables from our datasets. We removed the id column from our dataset because it does not add any relevance to our model. The Postal code feature was also removed because of the granularity of the information it holds and finally, the race column was also removed in order to promote the idea of global citizenship.

7.0 Data Dictionary

Since we excluded three variables from the onset of our analysis, we would only be providing explanations for the remaining 16 variables.

Variables	**Description**
Age	Age of the customers (16-25, 26-39, 40-64, 65 and above)
Gender	Customer's gender (Male and Female)
Driving_experience	No of years a customer has in driving (0-9, 10-19, 20-29, 30 and above)

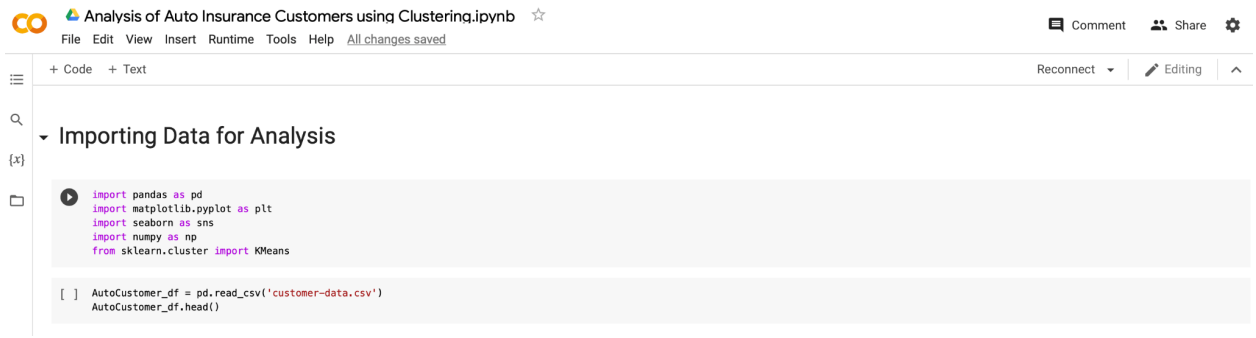
Education	Level of customer's education (none, high school and university)
Income	Income level of the customer (poverty, working class, middle class and upper class)
Credit_score	Credit score of the customer (Between 0 and 1)
Vehicle_ownership	Does the customer own a vehicle or not?(True or False)
Vehicle_year	The year of the vehicle (before 2015 and after 2015)
Married	The marital status of the customer (True if Married and False if not)
Children	Does the customer have children? (True if he/she has children and False if he/she does not)
Annual_Mileage	Accumulated mileage of the vehicle owned
Vehicle_type	The type of vehicle (sedan or sports car)
Speeding_violations	Number of speeding tickets the customer has
DUIs	Number of tickets gotten from driving under the influence
Past_accidents	Number of accidents the customer has had
Outcome	Has the customer filed a claim in the past year or not? (1 if he/she has and 0 if he/she has not)

Data Exploration

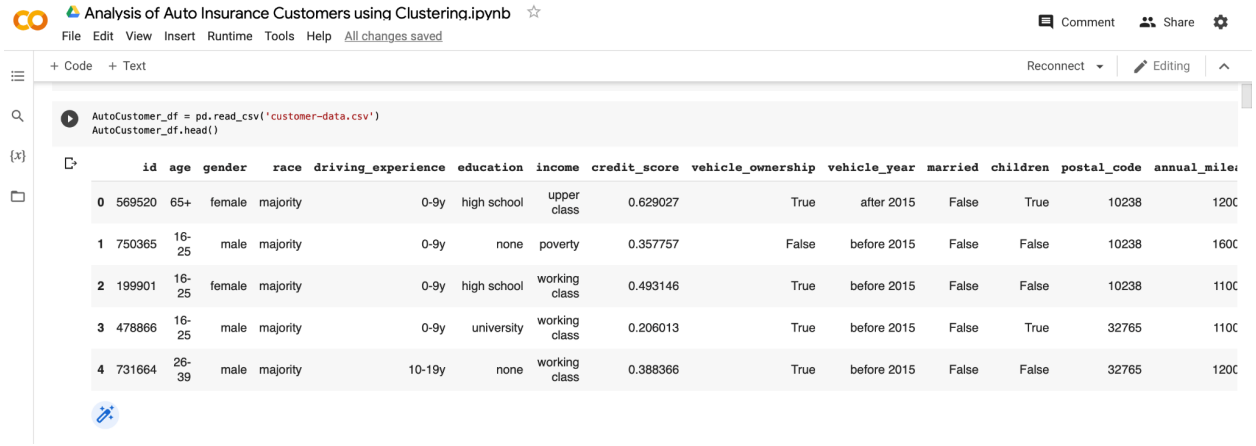
8.0 Data Exploration Techniques

Data Exploration was completed using different libraries in Python. ‘Pandas’, ‘MatPlotlib’ and ‘Seaborn’ libraries were utilized in this project to load the dataset, group variables, examine correlation and create plots/charts as required.

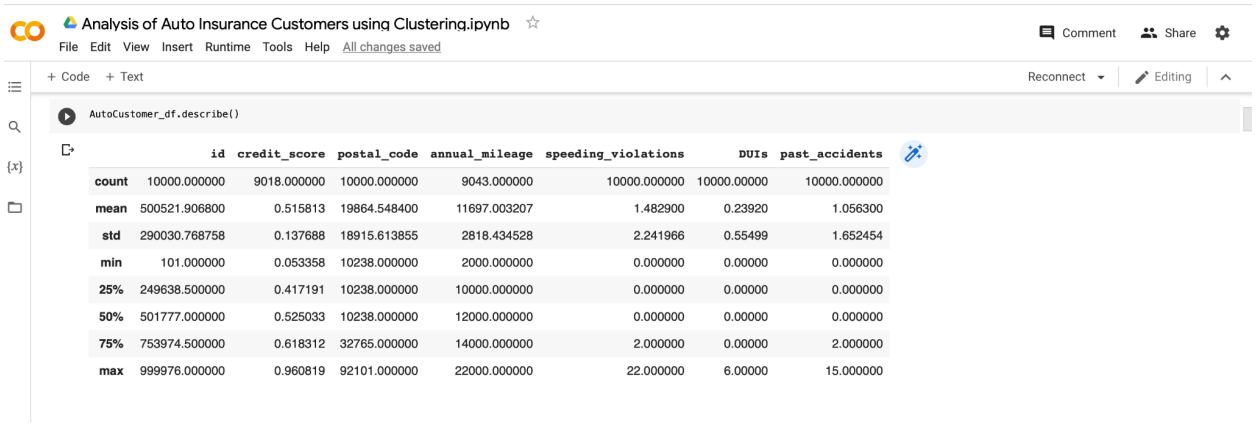
Importing the libraries and dataset



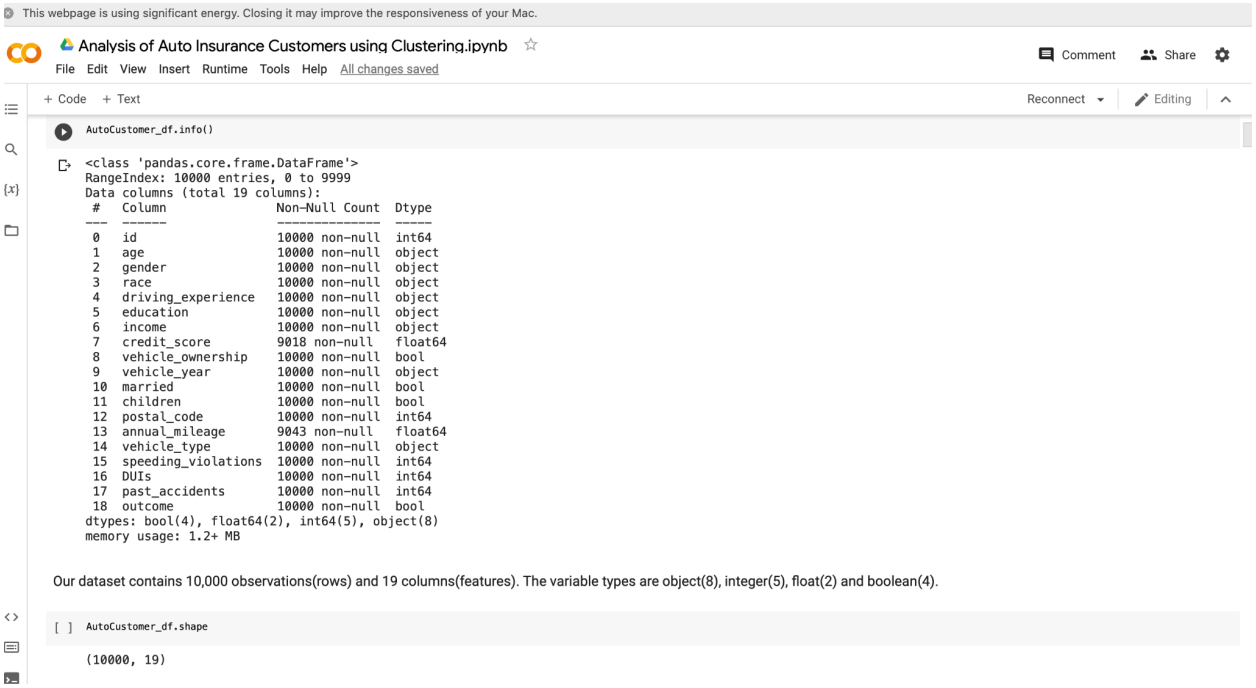
Preliminary data analysis



	id	age	gender	race	driving_experience	education	income	credit_score	vehicle_ownership	vehicle_year	married	children	postal_code	annual_mile
0	569520	65+	female	majority	0-9y	high school	upper class	0.629027	True	after 2015	False	True	10238	120
1	750365	16-25	male	majority	0-9y	none	poverty	0.357757	False	before 2015	False	False	10238	160
2	199901	16-25	female	majority	0-9y	high school	working class	0.493146	True	before 2015	False	False	10238	110
3	478866	16-25	male	majority	0-9y	university	working class	0.206013	True	before 2015	False	True	32765	110
4	731664	26-39	male	majority	10-19y	none	working class	0.388366	True	before 2015	False	False	32765	120

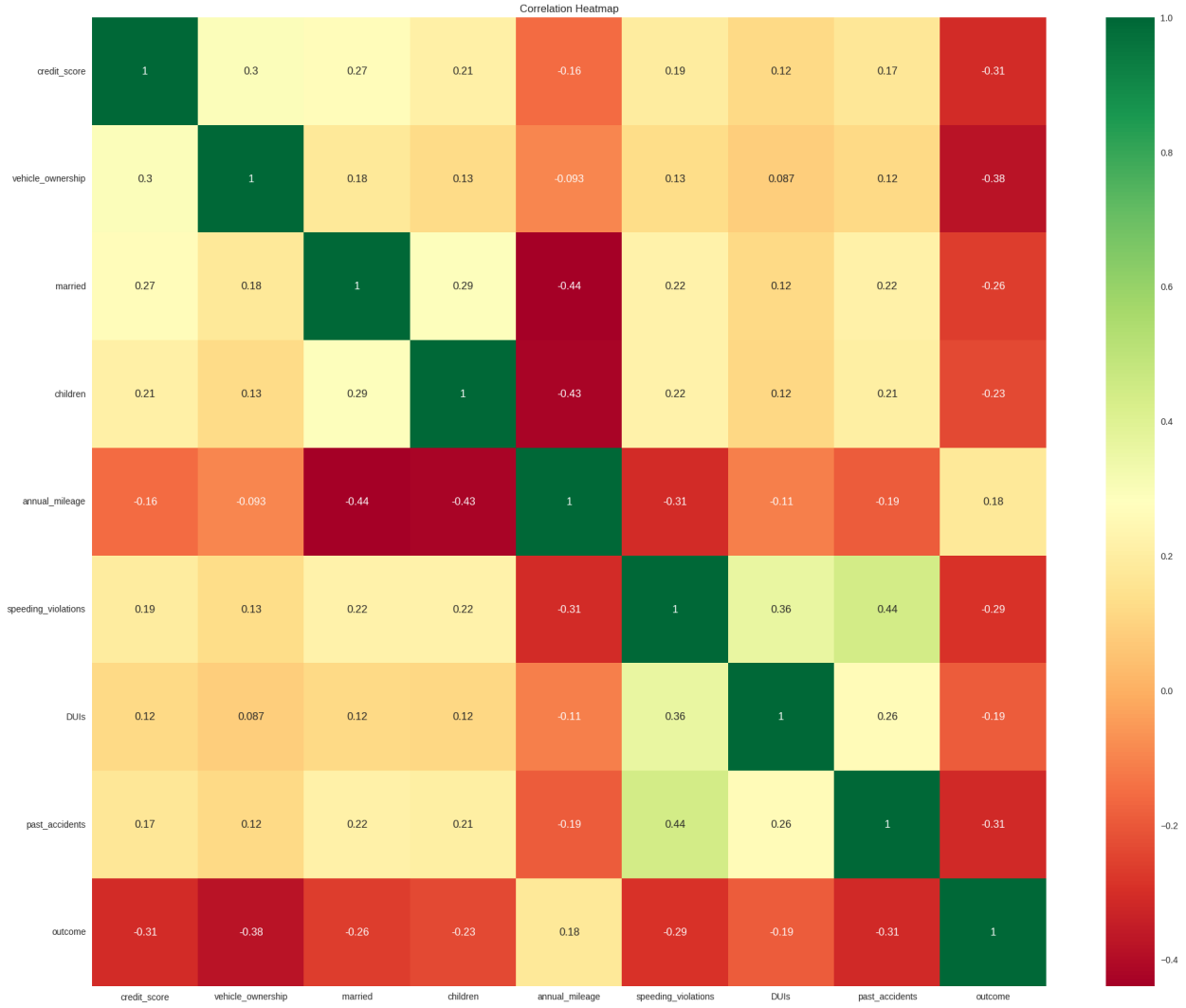


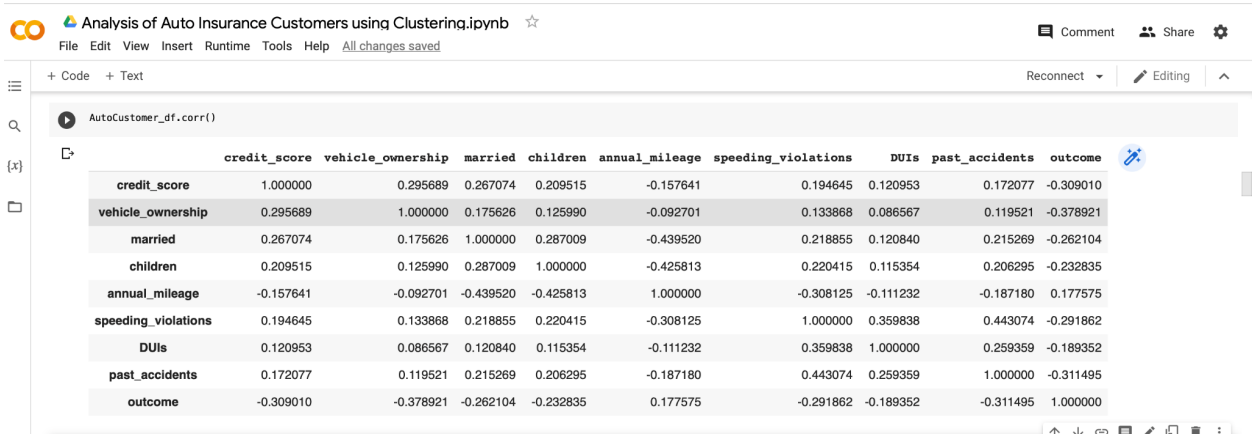
	id	credit_score	postal_code	annual_mileage	speeding_violations	DUIs	past_accidents
count	10000.000000	9018.000000	10000.000000	9043.000000	10000.000000	10000.00000	10000.000000
mean	500521.906800	0.515813	19864.548400	11697.003207	1.482900	0.23920	1.056300
std	290030.768758	0.137688	18915.613855	2818.434528	2.241966	0.55499	1.652454
min	101.000000	0.053358	10238.000000	2000.000000	0.000000	0.00000	0.000000
25%	249638.500000	0.417191	10238.000000	10000.000000	0.000000	0.00000	0.000000
50%	501777.000000	0.525033	10238.000000	12000.000000	0.000000	0.00000	0.000000
75%	753974.500000	0.618312	32765.000000	14000.000000	2.000000	0.00000	2.000000
max	999976.000000	0.960819	92101.000000	22000.000000	22.000000	6.00000	15.000000



In the preliminary data analysis, we get the summary of the data, the data type of each column, the size and shape of the dataset.

Correlation Analysis (Heatmap)





	credit_score	vehicle_ownership	married	children	annual_mileage	speeding_violations	DUIs	past_accidents	outcome
credit_score	1.000000	0.295689	0.267074	0.209515	-0.157641	0.194645	0.120953	0.172077	-0.309010
vehicle_ownership	0.295689	1.000000	0.175626	0.125990	-0.092701	0.133868	0.086567	0.119521	-0.378921
married	0.267074	0.175626	1.000000	0.287009	-0.439520	0.218855	0.120840	0.215269	-0.262104
children	0.209515	0.125990	0.287009	1.000000	-0.425813	0.220415	0.115354	0.206295	-0.232835
annual_mileage	-0.157641	-0.092701	-0.439520	-0.425813	1.000000	-0.308125	-0.111232	-0.187180	0.177575
speeding_violations	0.194645	0.133868	0.218855	0.220415	-0.308125	1.000000	0.359838	0.443074	-0.291862
DUIs	0.120953	0.086567	0.120840	0.115354	-0.111232	0.359838	1.000000	0.259359	-0.189352
past_accidents	0.172077	0.119521	0.215269	0.206295	-0.187180	0.443074	0.259359	1.000000	-0.311495
outcome	-0.309010	-0.378921	-0.262104	-0.232835	0.177575	-0.291862	-0.189352	-0.311495	1.000000

The correlation map shows the correlation coefficients for the variables in the dataset excluding categorical variables.

There is a significant positive correlation between these pairs :

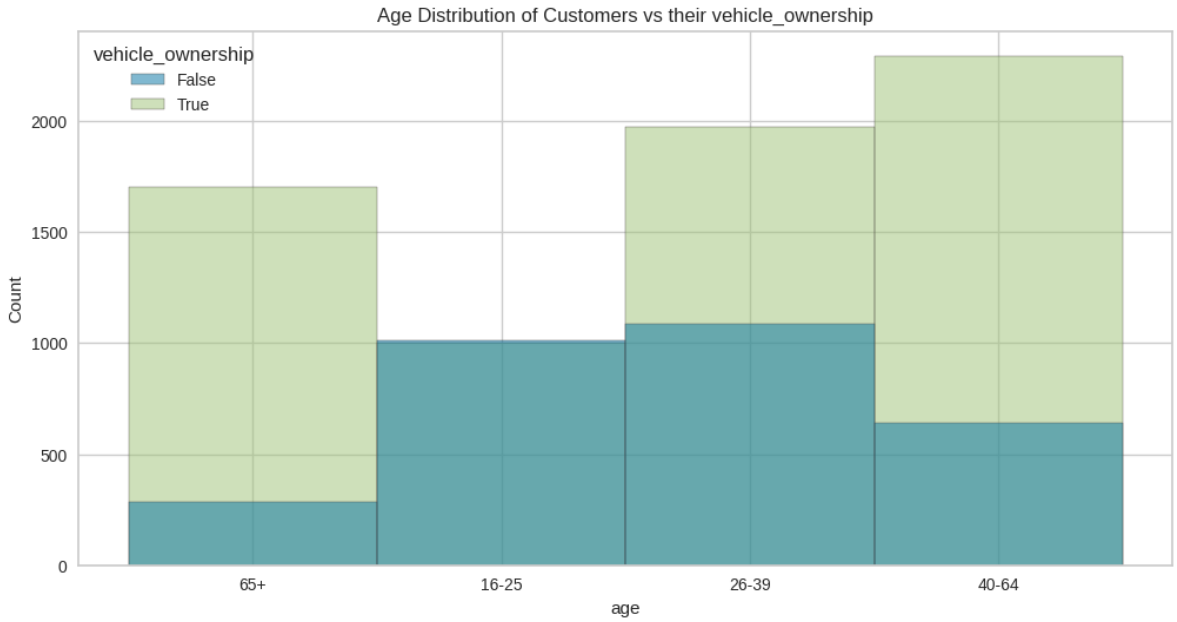
- **Speeding violations and past accidents.**
- **DUIs and speeding violations.**
- **Credit score and vehicle ownership.**
- **Married customers and credit score.**
- **Having children and credit score.**
- **Vehicle ownership and being married/ having children.**
- **Speeding violations and Customers with children. The strongest positive correlation is between Speeding violations and past accidents(0.44).**

Outcome is negatively correlated with all variables except annual mileage.

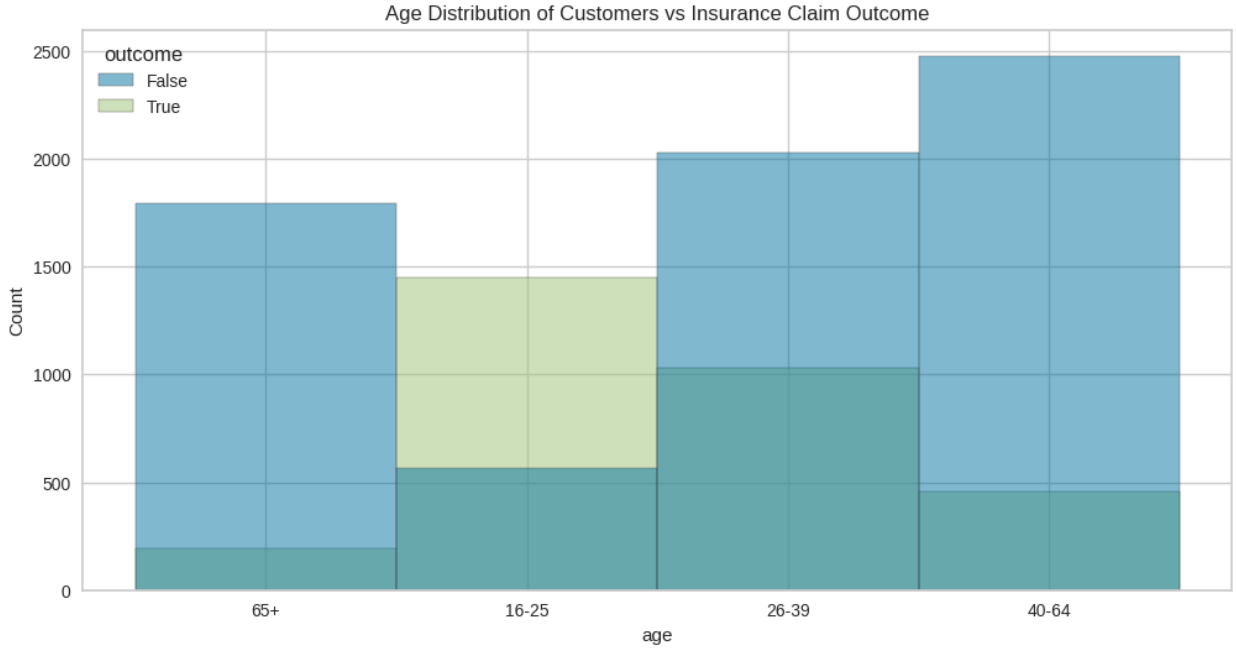
Annual mileage and Married customer are highly negatively correlated (-0.44)

Charts

The chart shows the age distribution of the customers in the dataset. We have more customers within the age range 26-39 years.

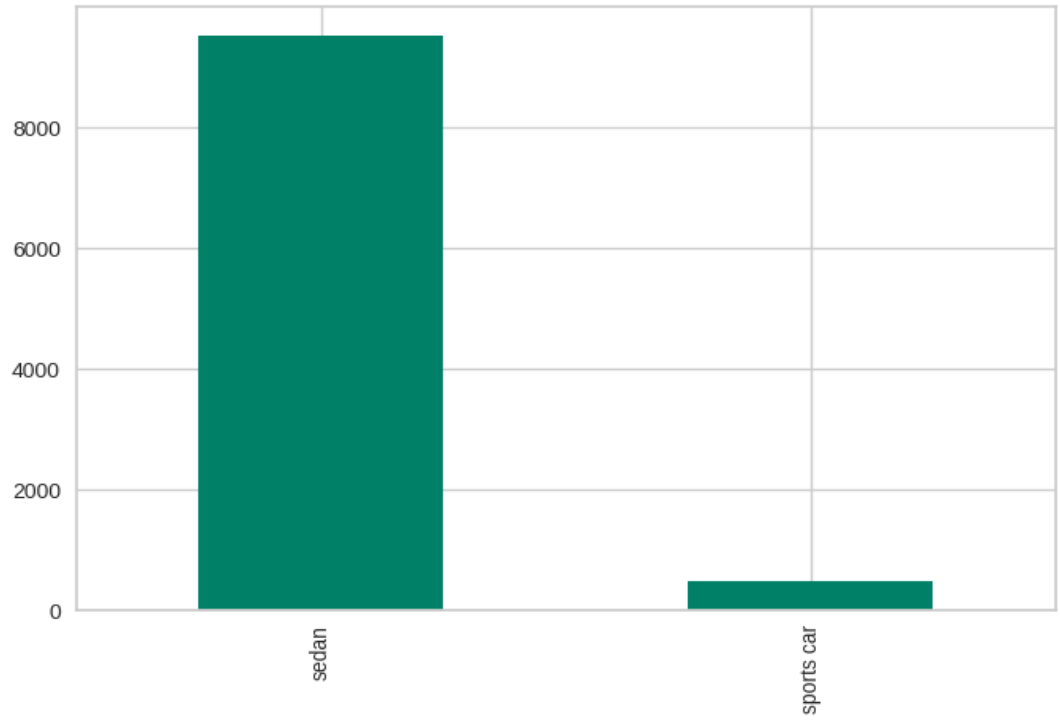


The age distribution vs vehicle ownership shows that Customers between age "40-64" own more vehicles than other age groups.

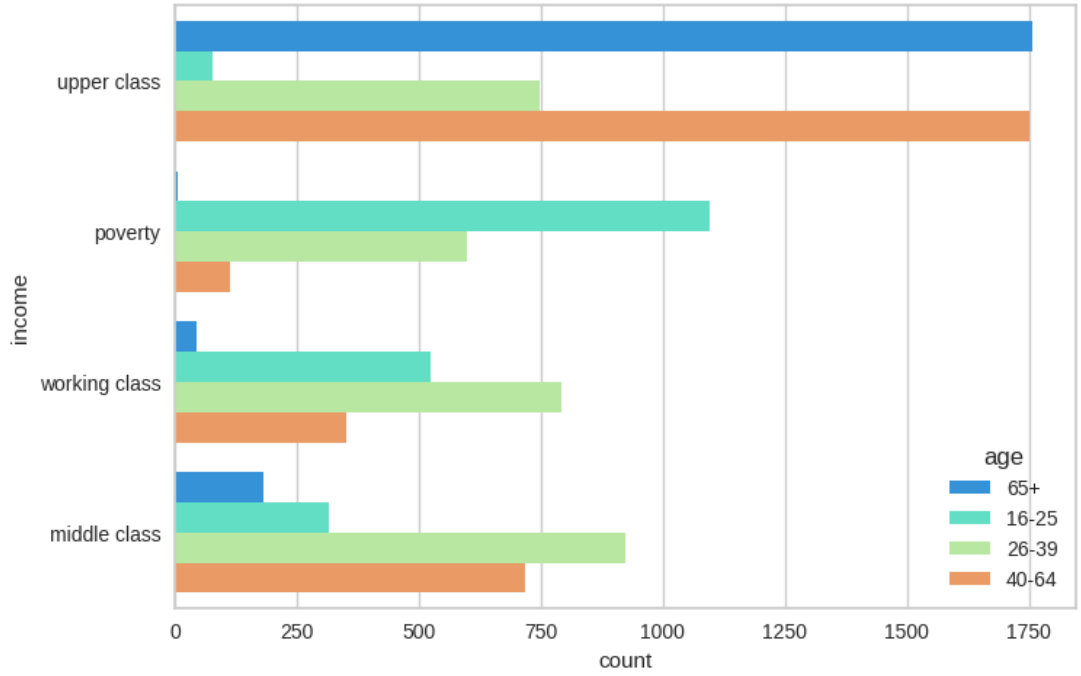


The outcome variable indicates 1 if a customer has filed an insurance claim in the last year when they encounter an accident else 0.

The chart shows that customers between age 16-25 have filed claims.



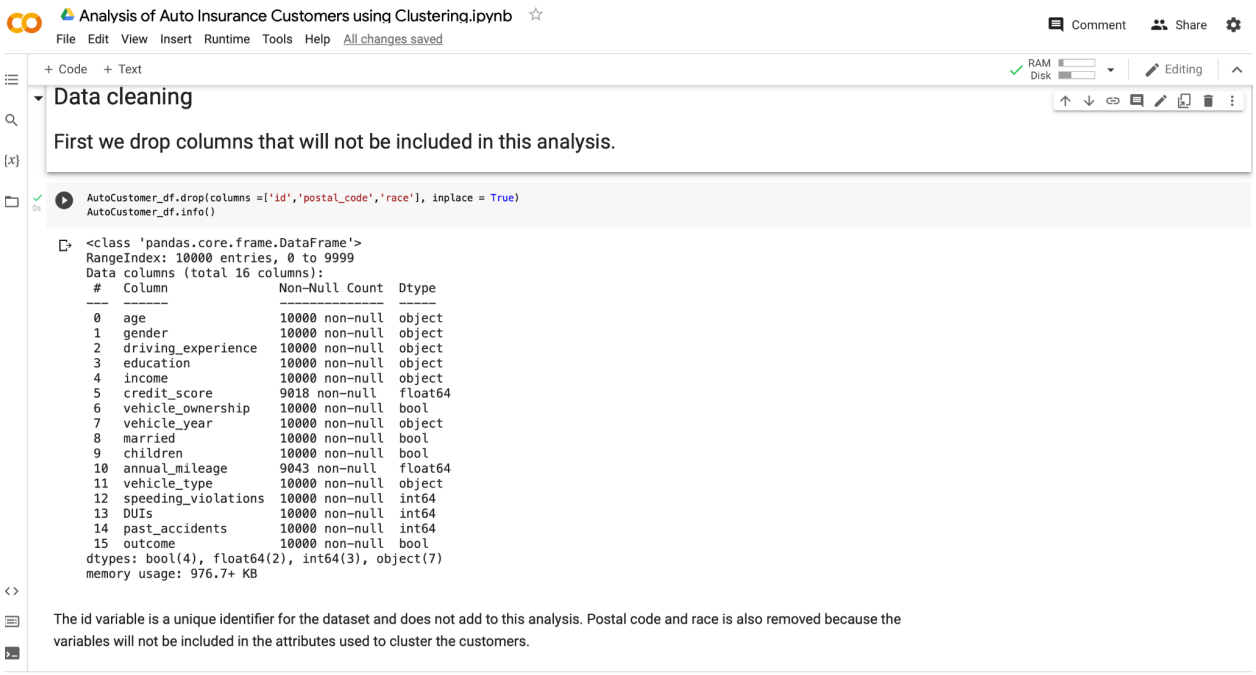
The bar graph shows the distribution of vehicle type, we have more sedan car owners in the dataset.



The chart shows the income distribution in our dataset. Age 40-64 and 65+ have higher counts in "upper class" income level, age 16-25 has more counts in "poverty" income level.

9.0. Data Cleansing

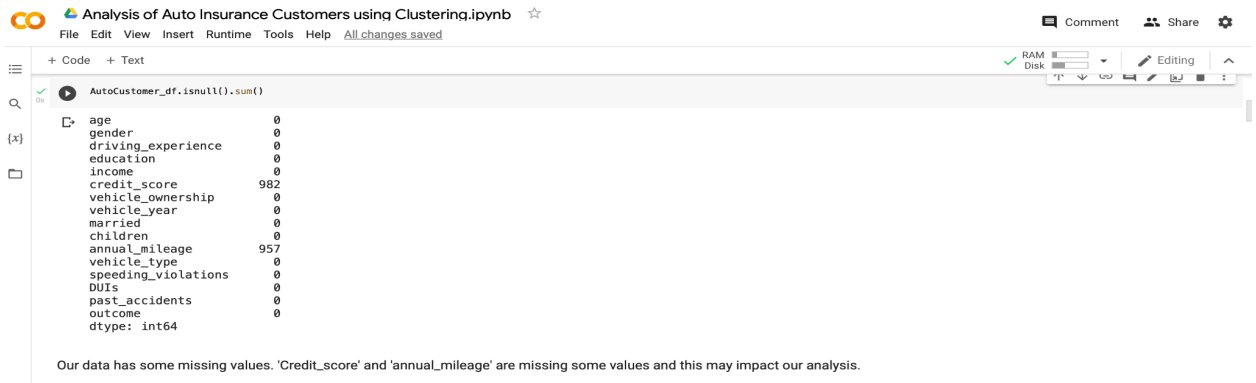
The approach to cleaning the data was to first drop columns that will not be included in the analysis.



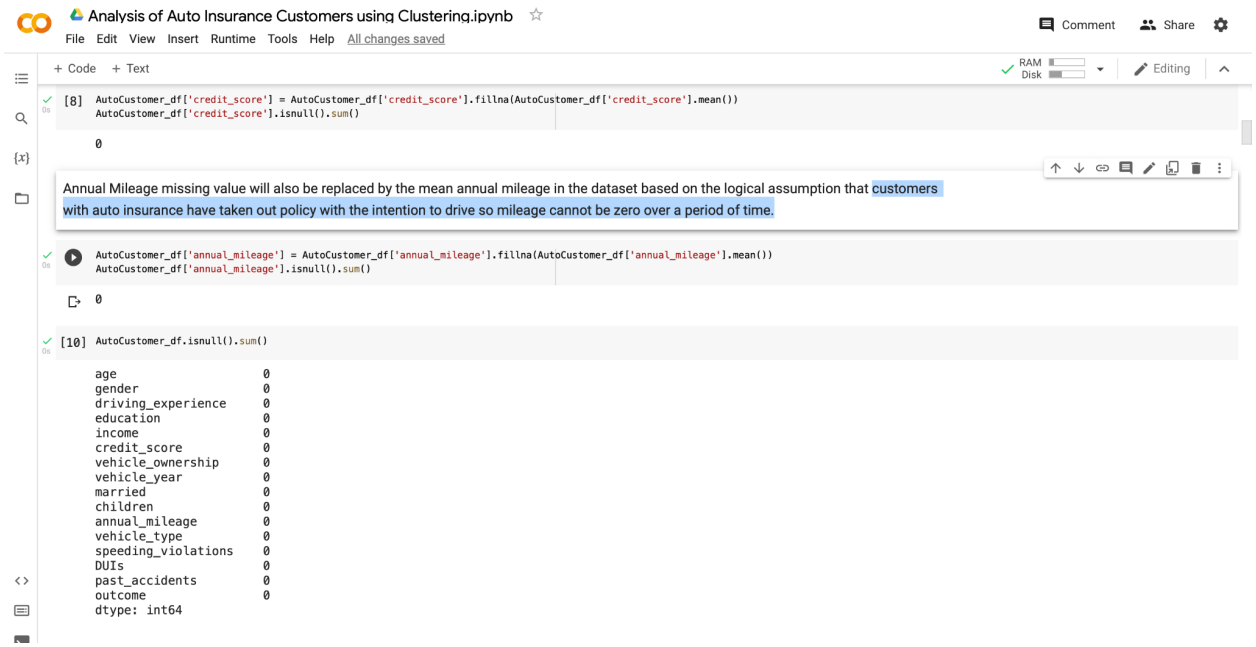
'Id' column dropped, the unique identifier does not add to this analysis. The analysis is being done to present the insurance company with a viable clustering option that does not include postal code, so the Postal code column is also dropped and race is also removed because it will not be included in the attributes used to cluster the customers.

Missing values

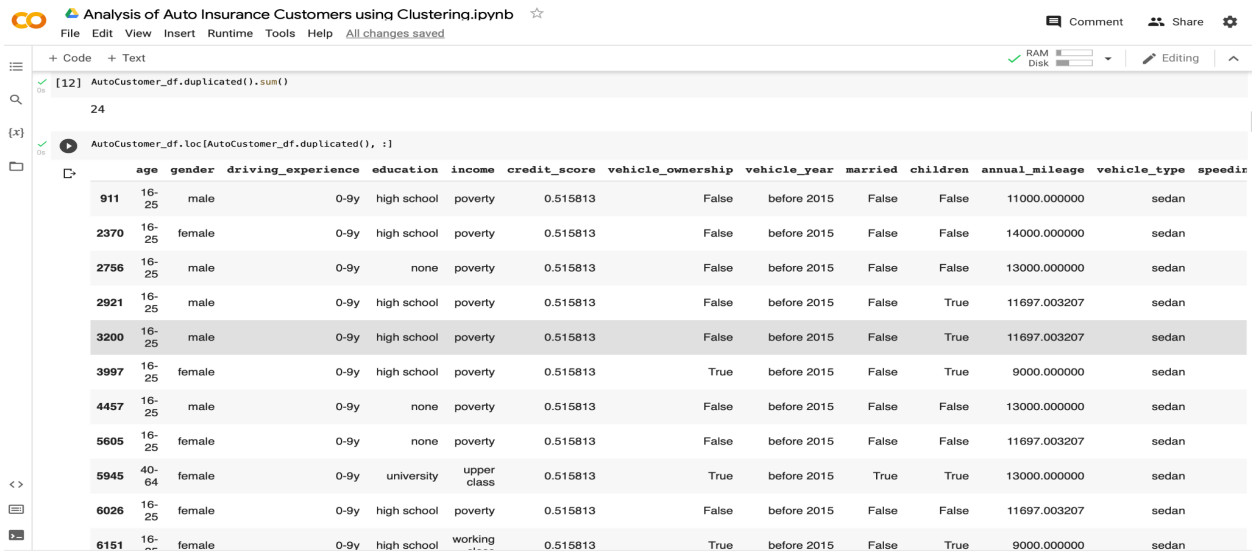
We explore our dataset for missing values and handle missing values using any of the techniques in data analysis as it is applicable to the variable based on logical reasoning.



‘Credit_score’ and ‘annual_mileage’ have some counts of missing values, 982 and 957 respectively.
Both columns will be replaced by mean values. For credit score, no individual has “zero” credit score so we will calculate the average of the values provided and use that to replace the missing values.
For annual mileage, customers with auto insurance have taken out a policy with the intention to drive, so mileage cannot be zero over a period of time.



Duplicate Values



	age	gender	driving_experience	education	income	credit_score	vehicle_ownership	vehicle_year	married	children	annual_mileage	vehicle_type	speedir
911	16-25	male	0-9y	high school	poverty	0.515813	False	before 2015	False	False	11000.000000	sedan	
2370	16-25	female	0-9y	high school	poverty	0.515813	False	before 2015	False	False	14000.000000	sedan	
2756	16-25	male	0-9y	none	poverty	0.515813	False	before 2015	False	False	13000.000000	sedan	
2921	16-25	male	0-9y	high school	poverty	0.515813	False	before 2015	False	True	11697.003207	sedan	
3200	16-25	male	0-9y	high school	poverty	0.515813	False	before 2015	False	True	11697.003207	sedan	
3997	16-25	female	0-9y	high school	poverty	0.515813	True	before 2015	False	True	9000.000000	sedan	
4457	16-25	male	0-9y	none	poverty	0.515813	False	before 2015	False	False	13000.000000	sedan	
5605	16-25	female	0-9y	none	poverty	0.515813	False	before 2015	False	False	11697.003207	sedan	
5945	40-64	female	0-9y	university	upper class	0.515813	True	before 2015	True	True	13000.000000	sedan	
6026	16-25	female	0-9y	high school	poverty	0.515813	False	before 2015	False	False	11697.003207	sedan	
6151	16-25	female	0-9y	high school	working class	0.515813	True	before 2015	False	True	9000.000000	sedan	

The dataset has 24 duplicate rows but this will be ignored because no one observation has the same information for duplicated rows. At least, one column has different information from the previous row.

Model Exploration

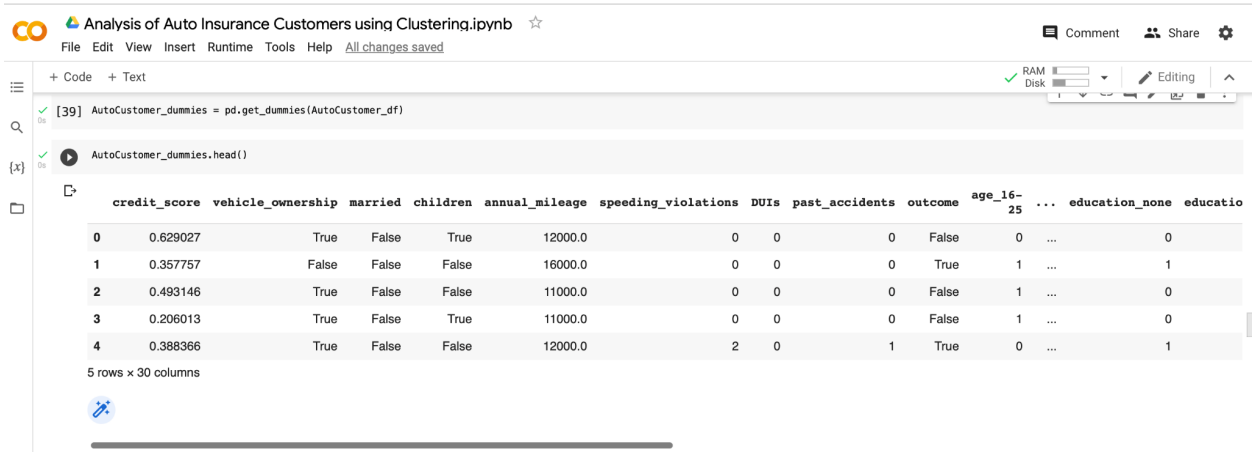
11.0. Modeling Approach/Introduction

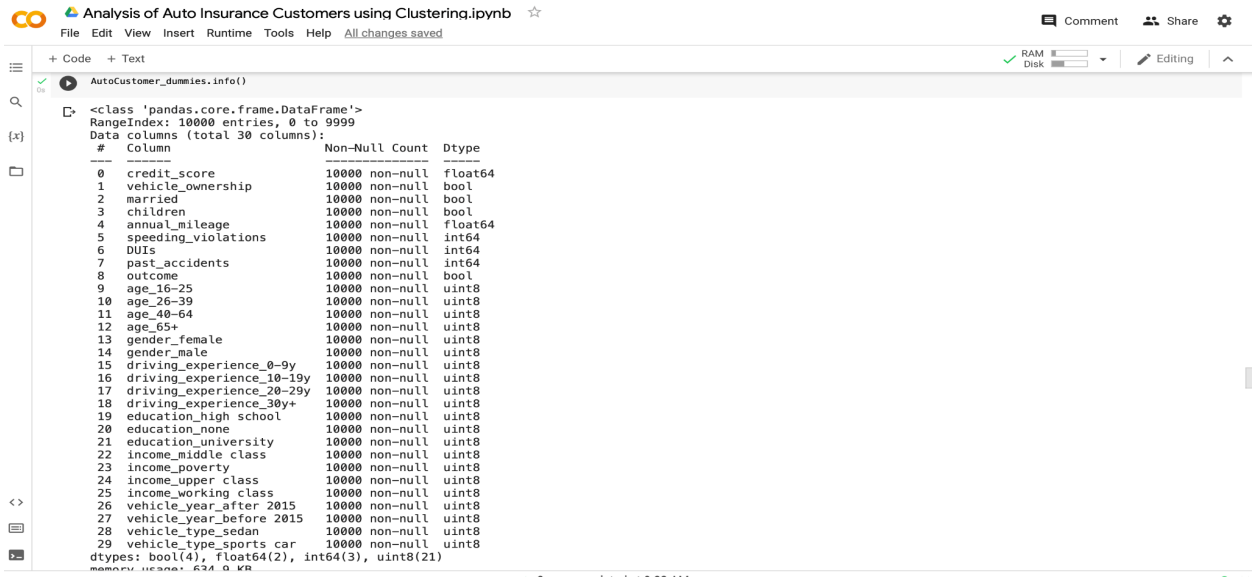
K-means Clustering

For this project, we will be adopting the K-means clustering approach to cluster the Auto-Insurance customers. K-means is an unsupervised learning algorithm that groups a dataset into k numbers of clusters. “K” is the number of clusters and it varies (JavaTpoint, n.d.).

```
from sklearn.cluster import KMeans
```

To apply K means clustering, all features must be numeric. Some of the data are not in numeric format, we solve this by applying the dummy or one hot encoding method.

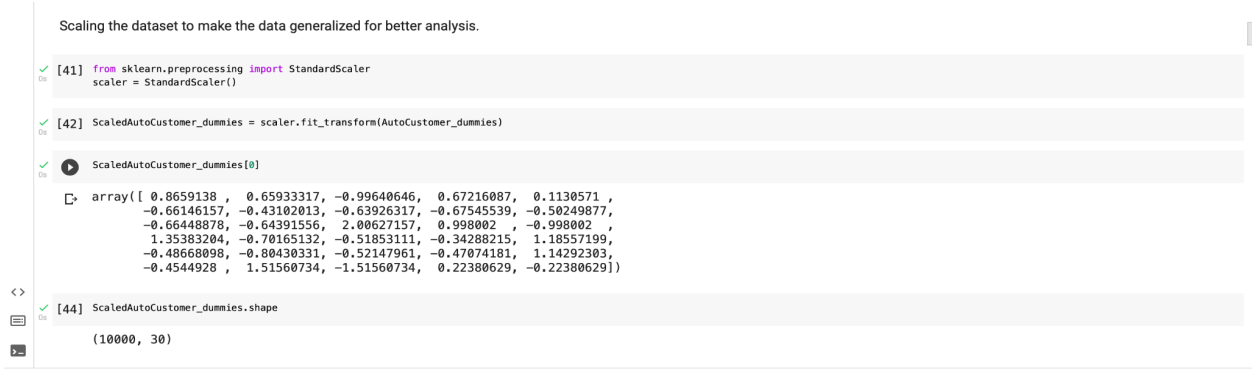




Scaling dataset

Scaling the dataset to make the data generalized for better analysis. We used the standard scaler for this project.

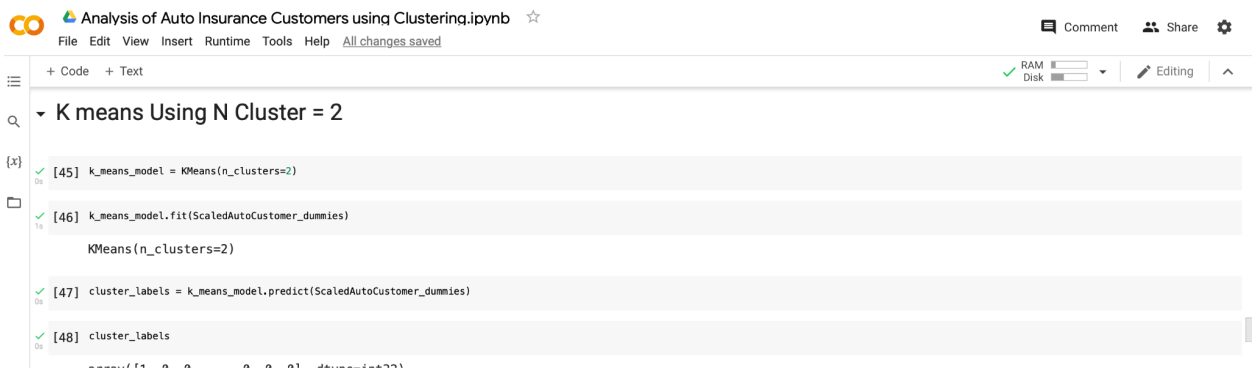
This scaling technique standardizes features by removing the mean and scaling to unit variance.



12.0. Model Technique #1

Assigning random number of clusters

Choosing k = 2.



The model will predict the clusters assigning each row randomly to either of the 2 clusters. Cluster 0 and Cluster 1.

Merge cluster to dataset

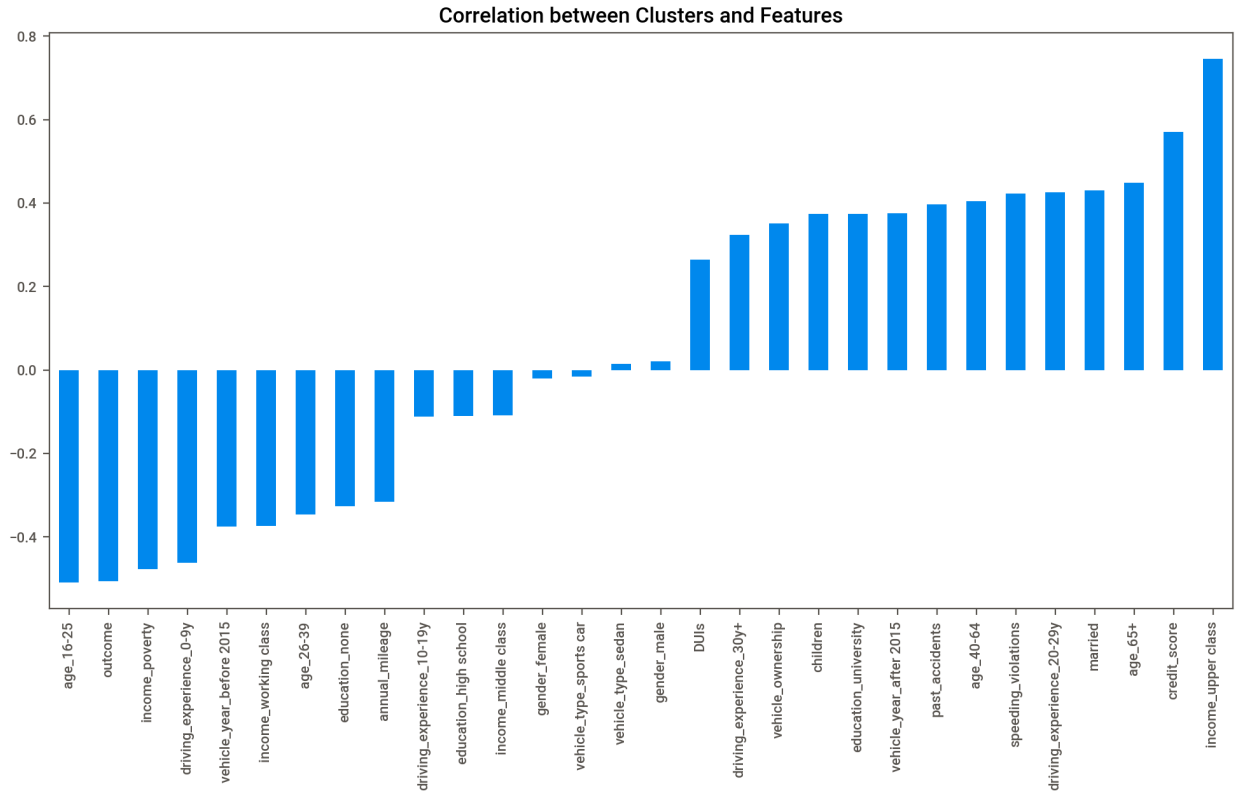
```
[51] AutoCustomer_dummies2["Cluster"] = cluster_labels
```

```
[52] AutoCustomer_dummies2.head()
```

iversity	income_middle class	income_poverty	income_upper class	income_working class	vehicle_year_after 2015	vehicle_year_before 2015	vehicle_type_sedan	vehicle_type_sports car	Cluster
0	0	0	1	0	1	0	1	0	1
0	0	1	0	0	0	1	1	0	0
0	0	0	0	1	0	1	1	0	0
1	0	0	0	1	0	1	1	0	0
0	0	0	0	1	0	1	1	0	0

The cluster label contains the clustering distribution. To visualize our clusters, we get:

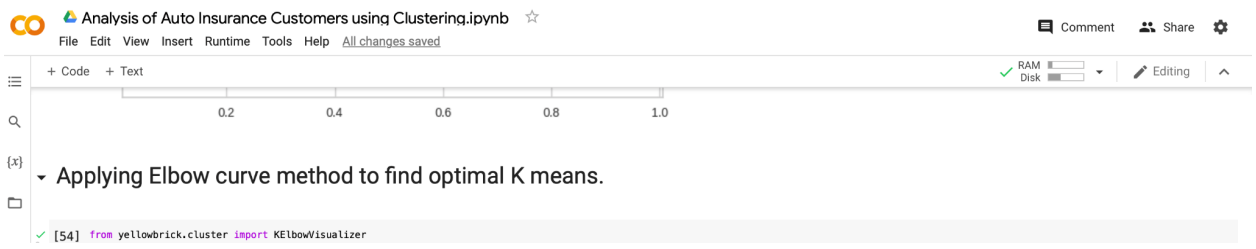
Exploring correlation between clusters and features

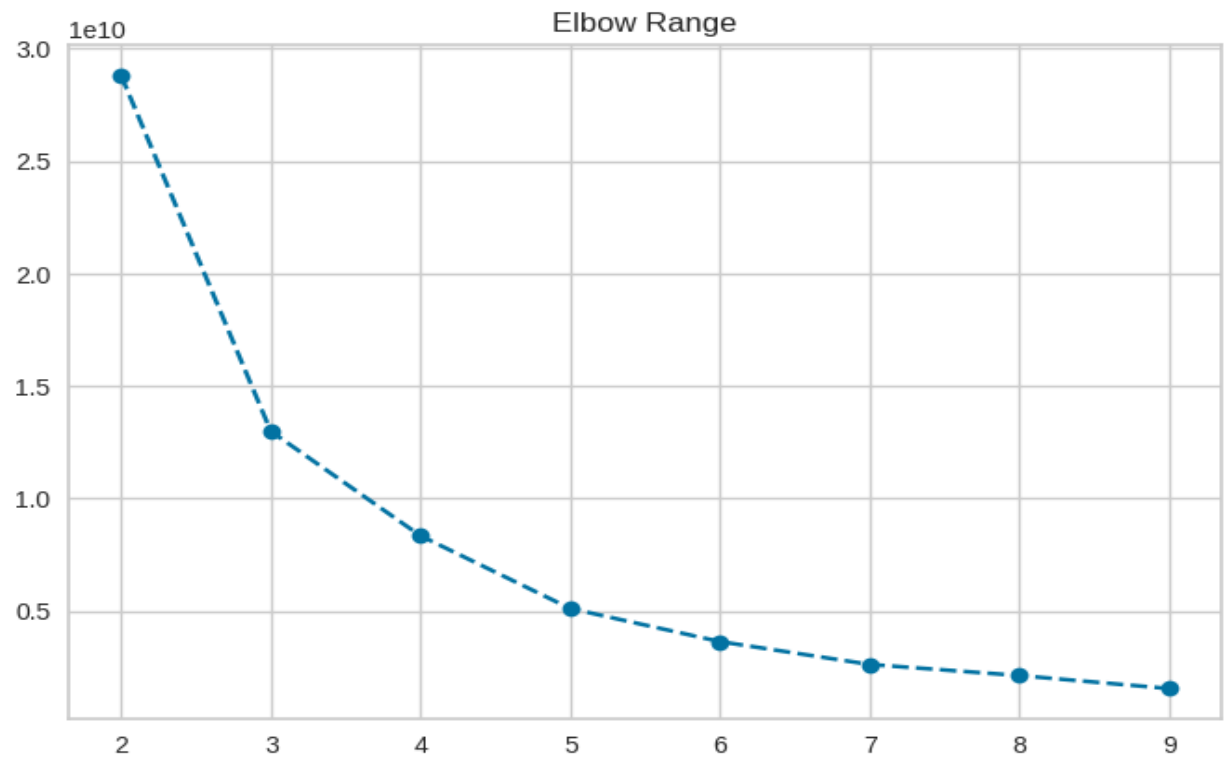


13.0. Model Technique #2

Elbow Curve Method and K means clustering

The Elbow curve method is used to determine the optimal number of clusters. We will apply this to get value for k.

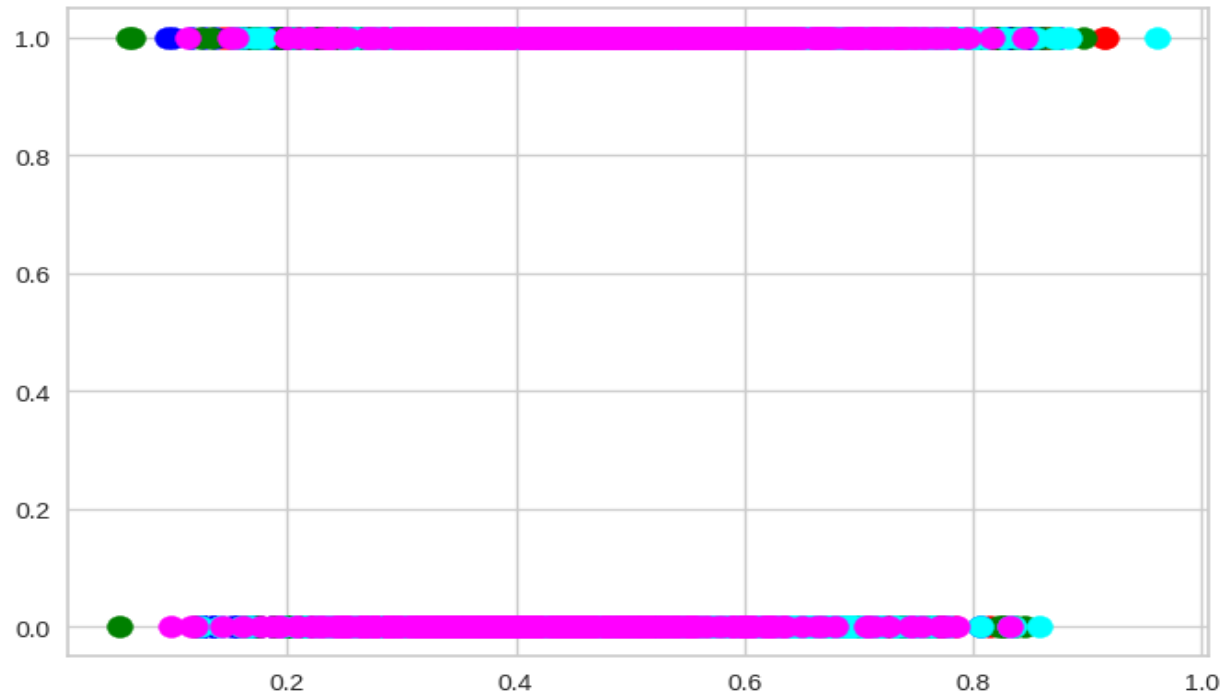




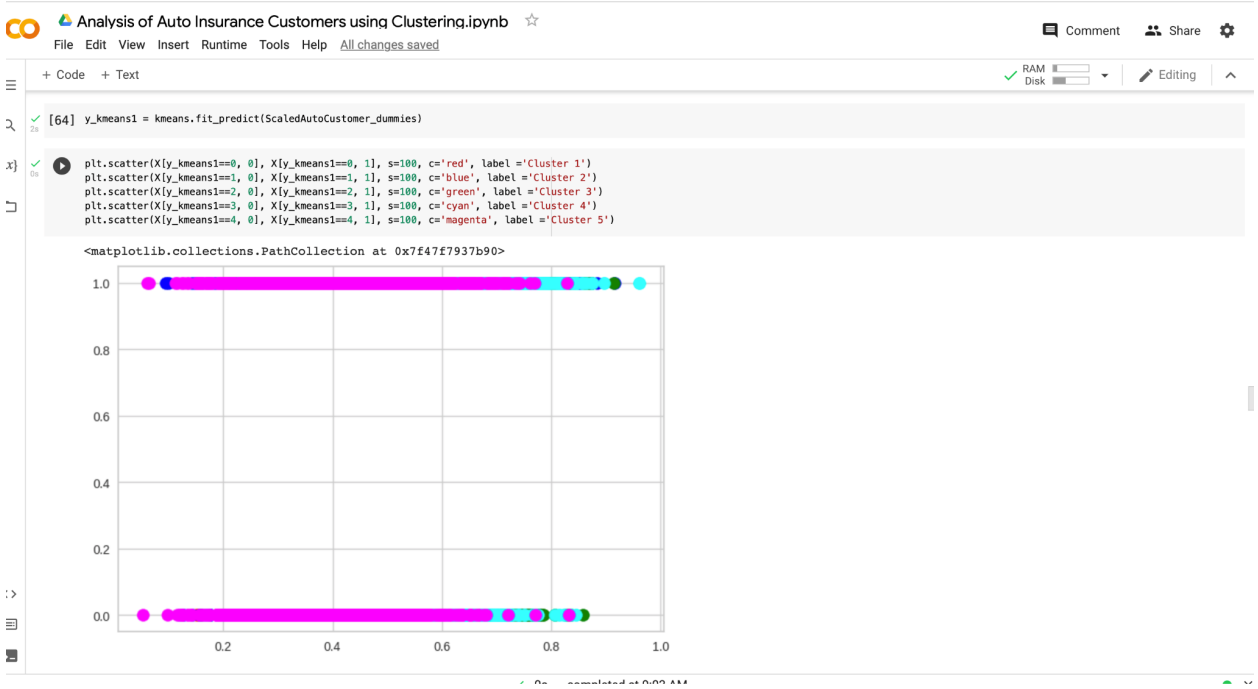
There is an elbow at 5. K value/ optimal value of clusters is 5.

Applying K value = 5 to K-means on our dataset.

The visualization:



Applying K value = 5 to K-means on our scaled dataset.



There's not much variation in the data points of the clusters from the chart above. For this project, we need distinct attributes for the customers where there is not too much overlapping.

14.0. Model Technique #3

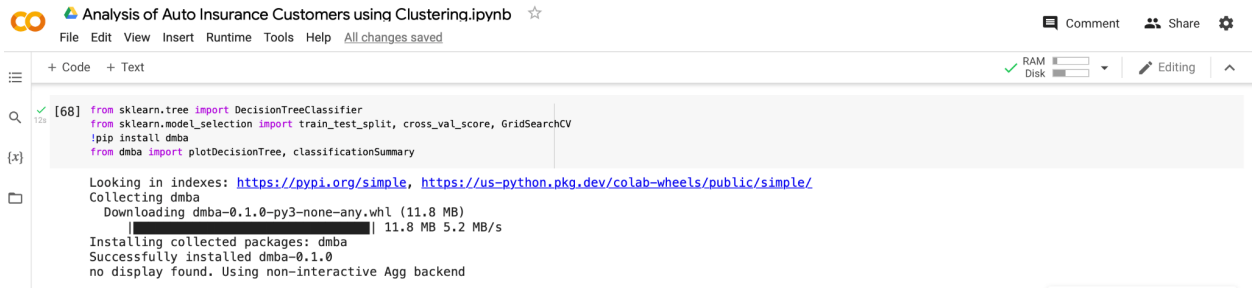
Dimensionality reduction, Elbow curve method and K means

First, we will use Decision Tree to let our dataset decide on the appropriate classification and the most important variables.

Assumption

We are making the assumption that the 'outcome' variable in our dataset which tells us whether a customer has filed a claim or not in the past year is a direct indicator of the customer's risk level. This is due to the fact that a customer who has filed a claim once is likely to file yet another claim.

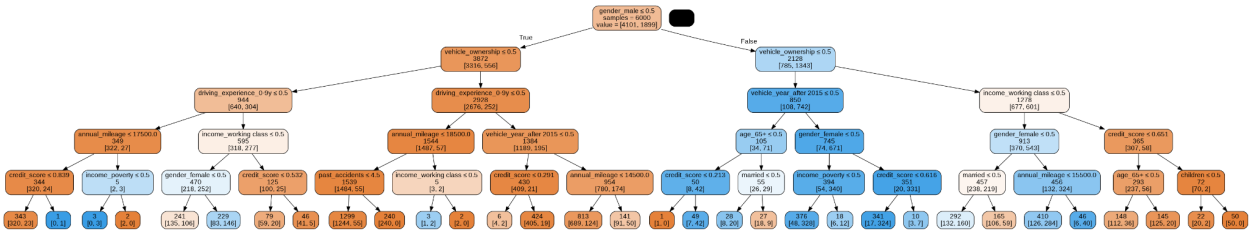
Decision Tree



We create a Decision Tree with all our datasets and set aside some of our data for validation of our model by using the train test split.

Test size used is 0.4

The resulting decision tree:

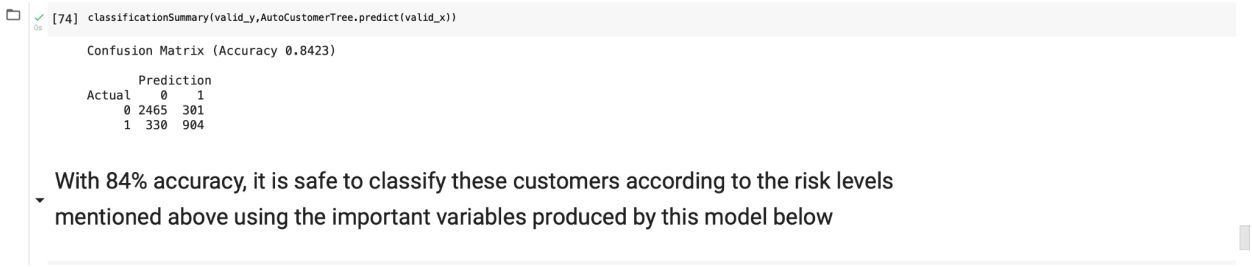


Confusion Matrix

We use the confusion matrix to assess the performance of our algorithm.

Assumption

Accuracy of at least 80% is sufficient to proceed with this model technique.

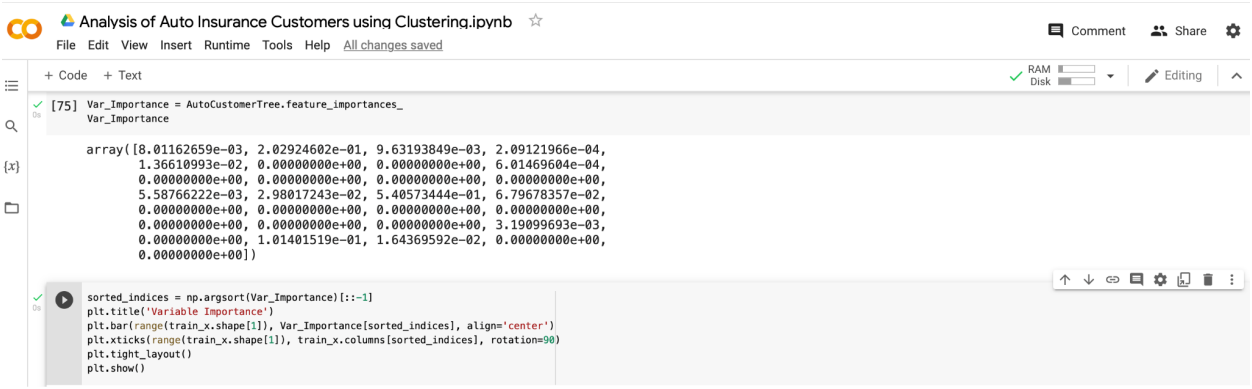


```
[74] classificationSummary(valid_y,AutoCustomerTree.predict(valid_x))

     Confusion Matrix (Accuracy 0.8423)

           Prediction
     Actual    0    1
          0 2465  301
          1  330  904
```

With 84% accuracy, it is safe to classify these customers according to the risk levels mentioned above using the important variables produced by this model below

With 84% accuracy, it is safe to classify these customers according to the risk levels mentioned above using the important variables produced by the Decision Tree.

Important Variables



Analysis of Auto Insurance Customers using Clustering.ipynb

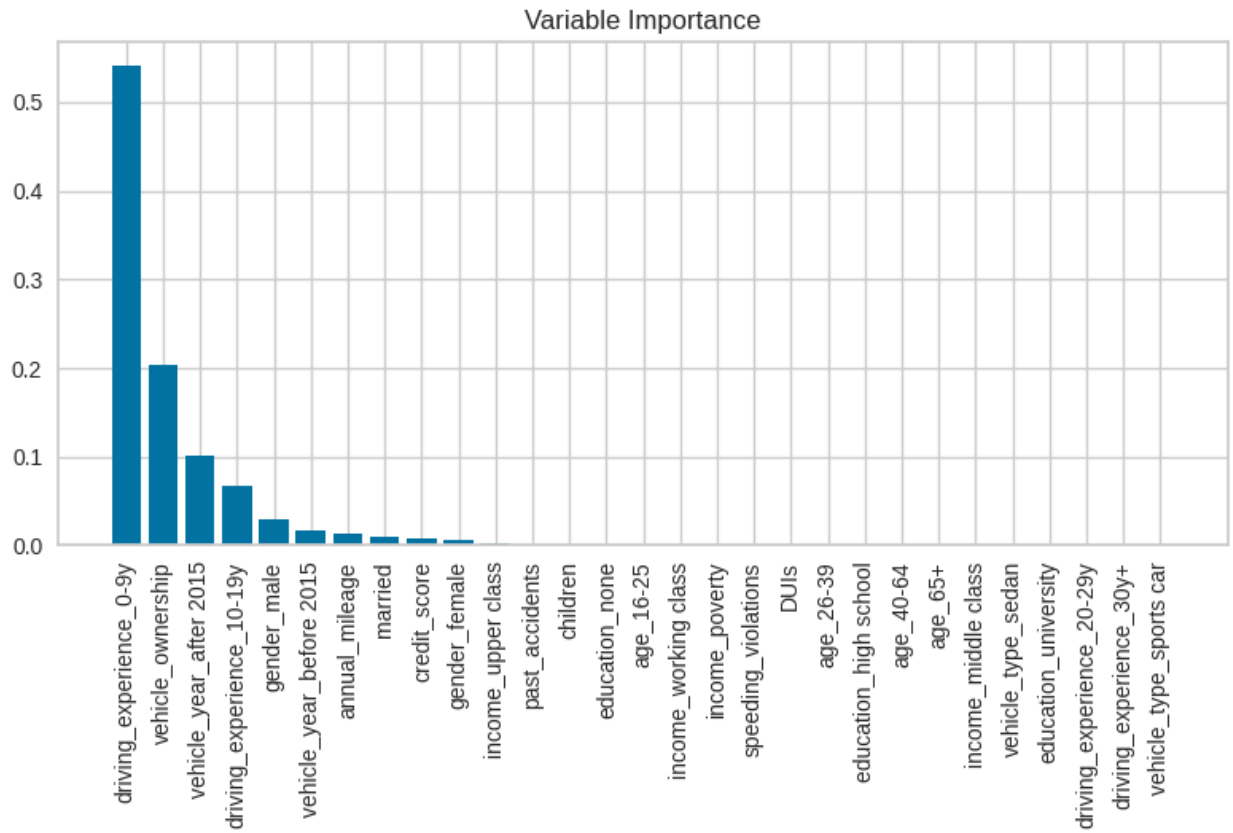
File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

```
[75] Var_Importance = AutoCustomerTree.feature_importances_
     Var_Importance

     array([8.01162659e-03, 2.02924602e-01, 9.63193849e-03, 2.09121966e-04,
            1.36610993e-02, 0.00000000e+00, 0.00000000e+00, 6.01469604e-04,
            0.00000000e+00, 0.00000000e+00, 0.00000000e+00, 0.00000000e+00,
            5.58766222e-03, 2.98017243e-02, 5.40573444e-01, 6.79678357e-02,
            0.00000000e+00, 0.00000000e+00, 0.00000000e+00, 0.00000000e+00,
            0.00000000e+00, 0.00000000e+00, 0.00000000e+00, 3.19099693e-03,
            0.00000000e+00, 1.01401519e-01, 1.64369592e-02, 0.00000000e+00,
            0.00000000e+00])

sorted_indices = np.argsort(Var_Importance)[::-1]
plt.title('Variable Importance')
plt.bar(range(train_x.shape[1]), Var_Importance[sorted_indices], align='center')
plt.xticks(range(train_x.shape[1]), train_x.columns[sorted_indices], rotation=90)
plt.tight_layout()
plt.show()
```

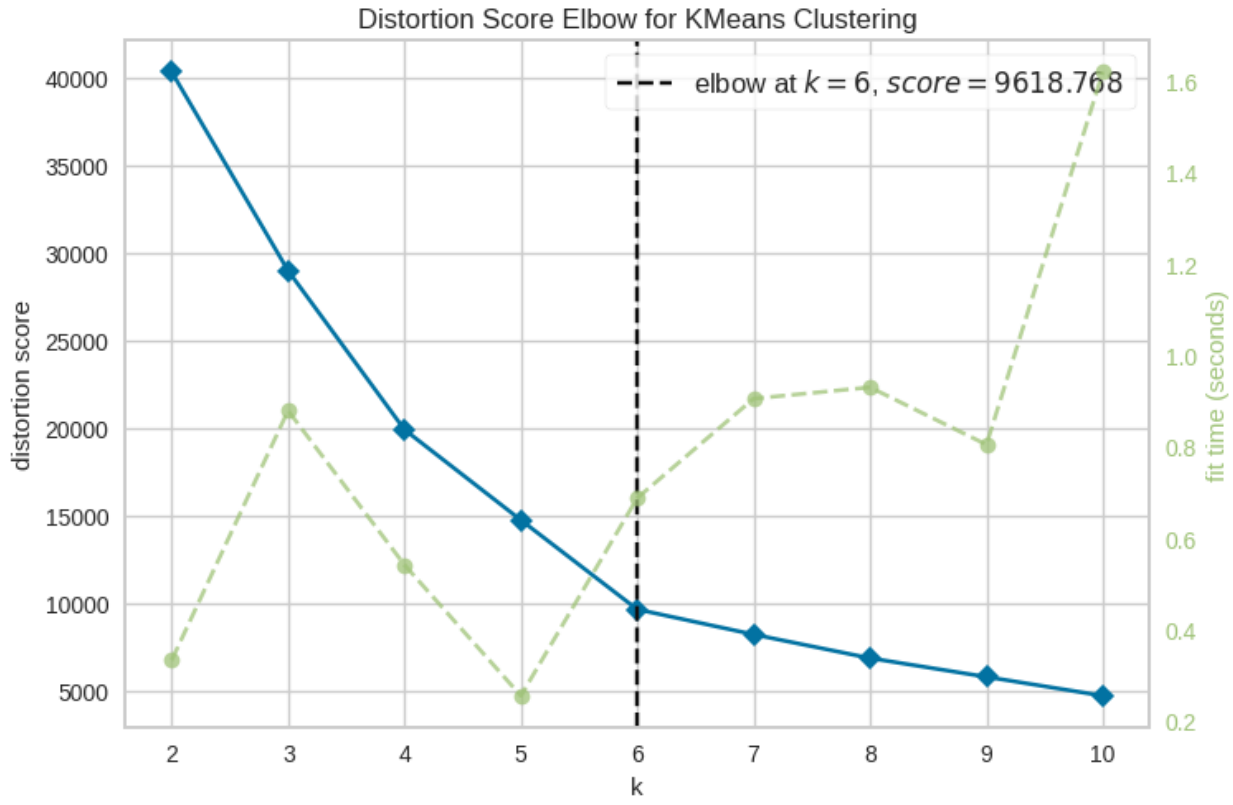


The important variables are:
Driving experience (0-9years),
Vehicle ownership,
Vehicle year (after 2015),
Driving experience(10-19years),
Gender(male),
Vehicle year(before 2015),
Annual mileage,
Married,
Credit score,
Gender(female) in descending order.

Scaling Important Variables and applying PCA

PCA is applied to reduce the size of our data while creating new and important features to remove the curse of dimensionality.

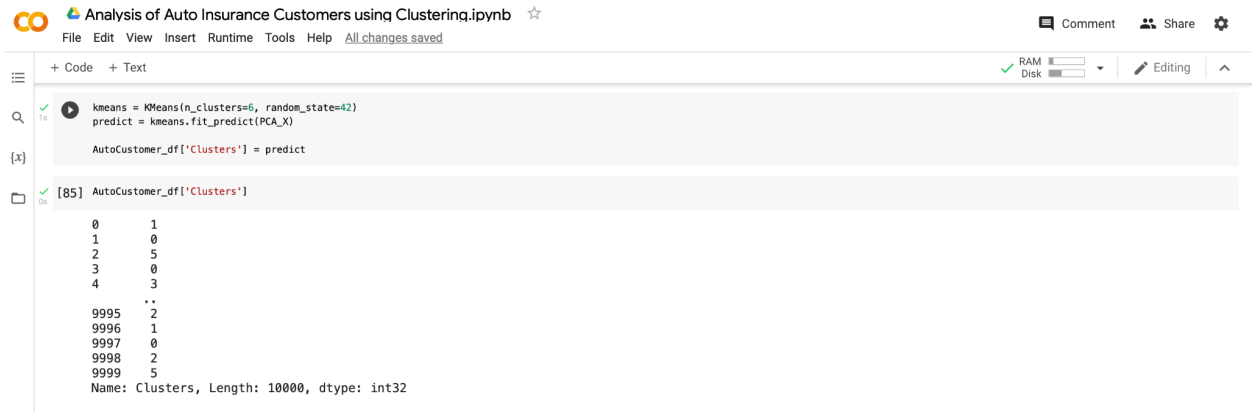
Elbow Curve result



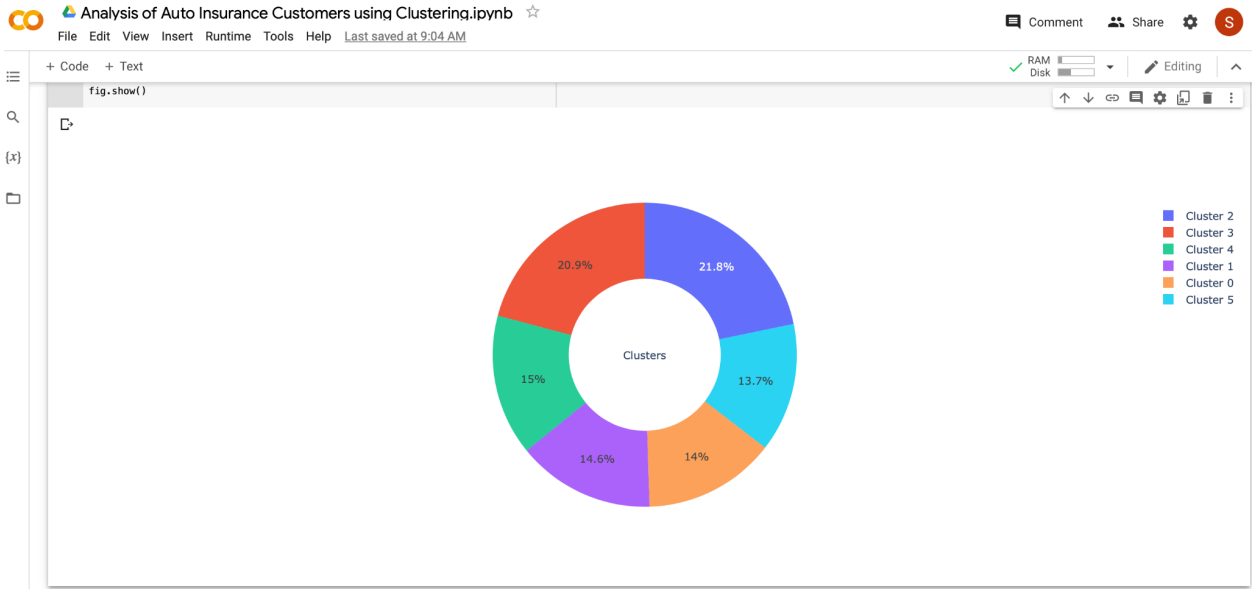
K =6

The result of applying the Elbow curve is k =6, which means that the optimal number of clusters is 6.

Cluster Analysis and Visualization



Distribution of Clusters.



The above chart shows the percentage distribution of our clusters. There are 6 clusters using the K Means clustering after Dimension reduction.

Cluster 0 - 14%

Cluster 1 - 14.6%

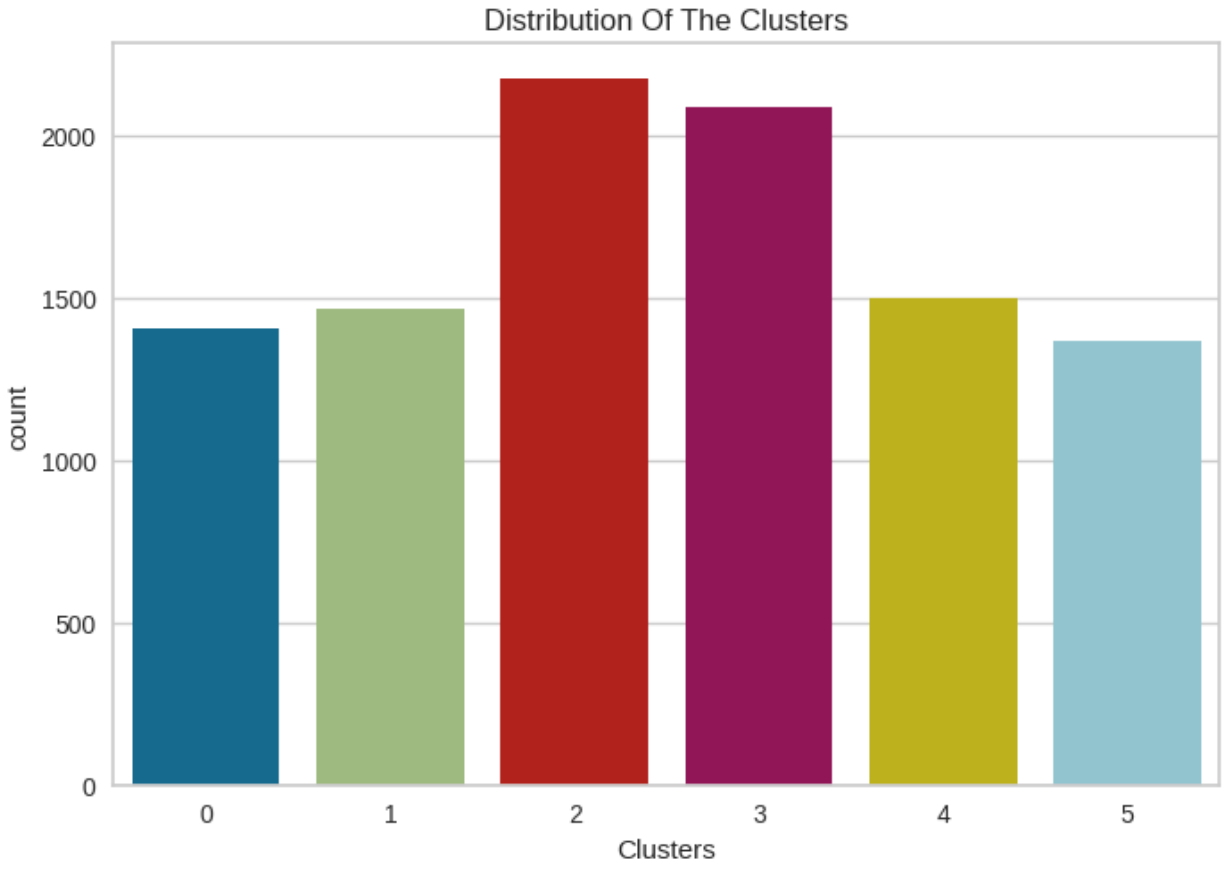
Cluster 2 - 21.8%

Cluster 3 - 20.9%

Cluster 4 - 15%

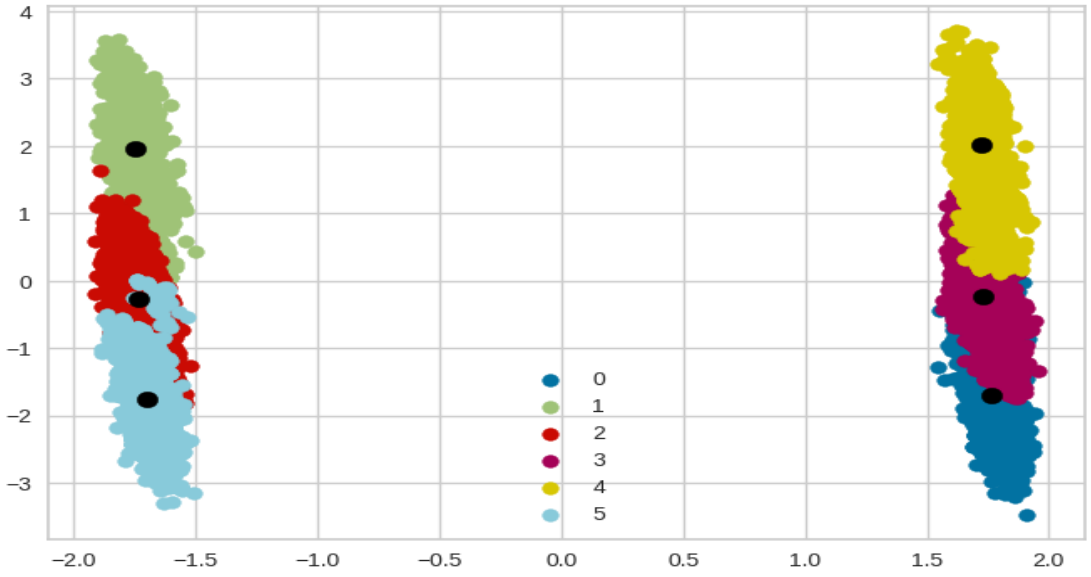
Cluster 5 - 13.7%

More customers fall in Cluster 2 and 3. Cluster 5 has the least number of customers.



Cluster Centroids

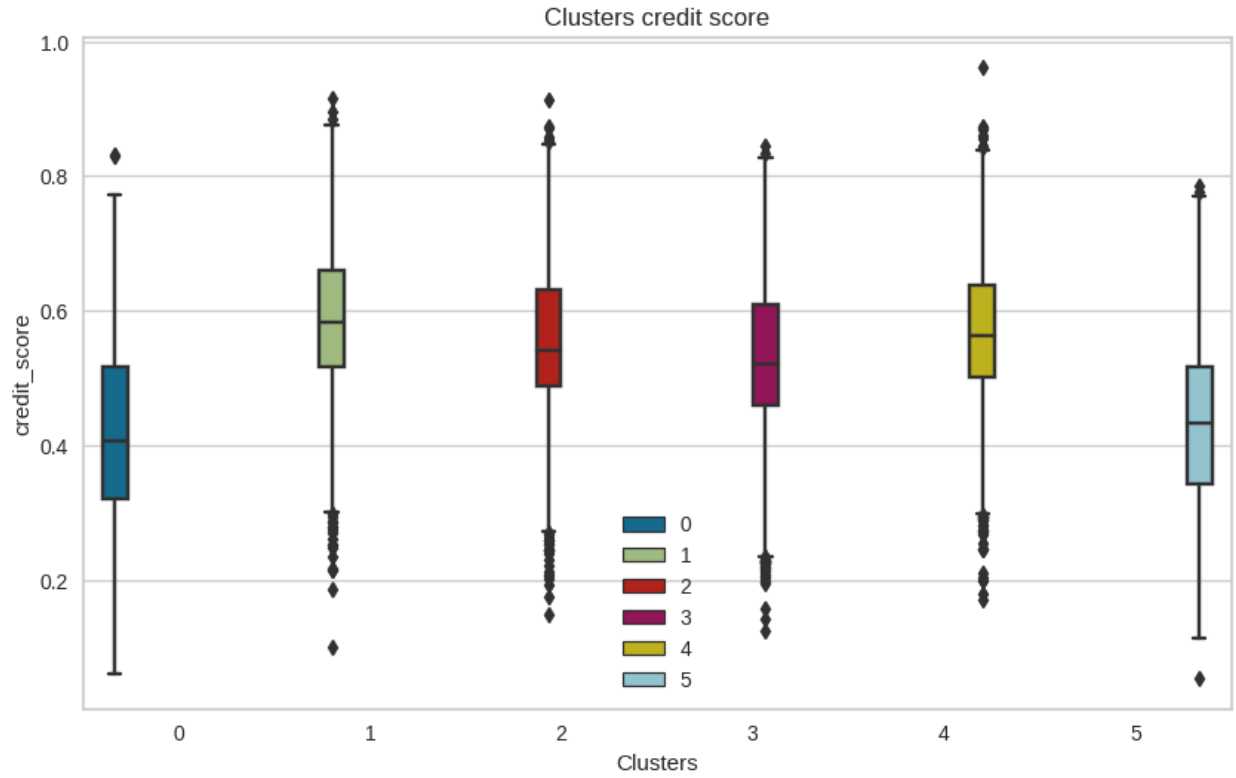
We plot the clusters against the center of the clusters.



Clustering Customers

Clusters Credit score distribution.

Credit score was provided in the dataset as values between 0 and 1.



Cluster 0 - Customers with relatively high credit score

Cluster 1- Customers with high credit score

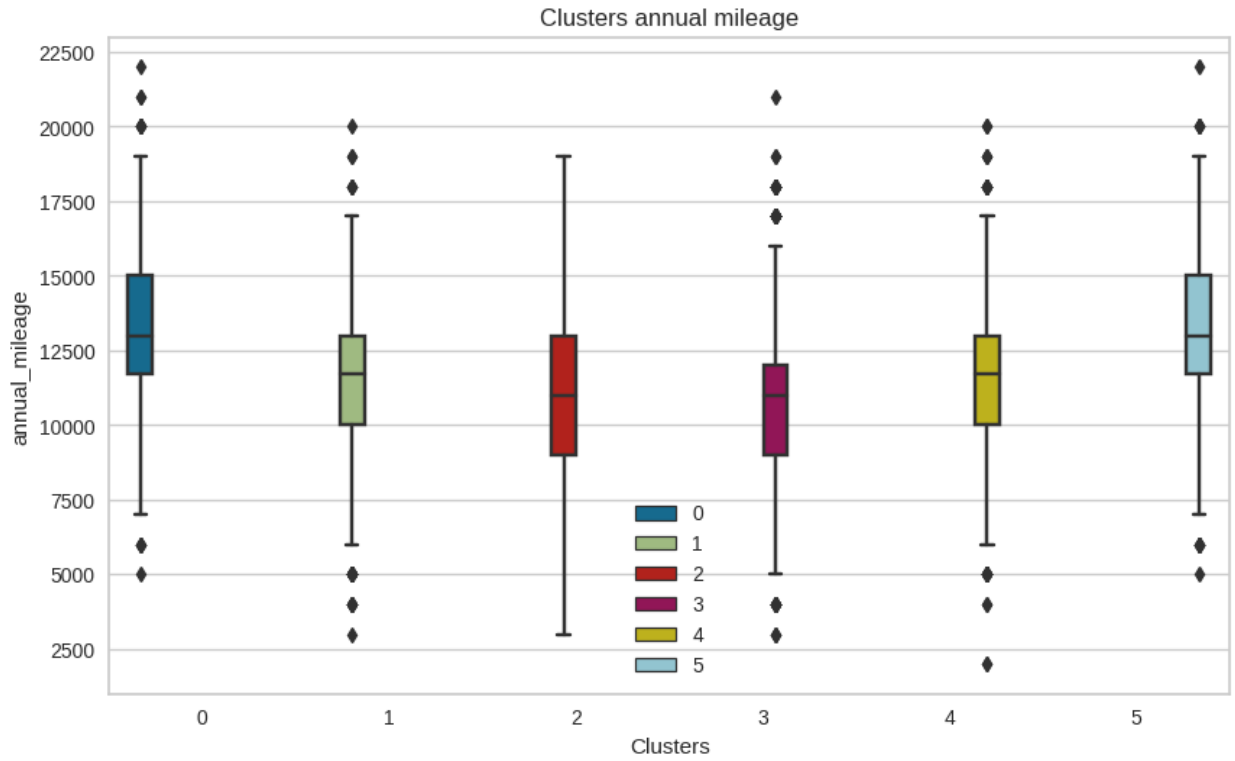
Cluster 2- Customers with high credit score

Cluster 3- Customers with relatively high credit score

Cluster 4- Customers with high credit score

Cluster 5- Customers with average credit score

Clusters Annual Mileage distribution.



Cluster 0- Customers with higher vehicle annual mileage

Cluster 1- Customers with relatively high annual mileage

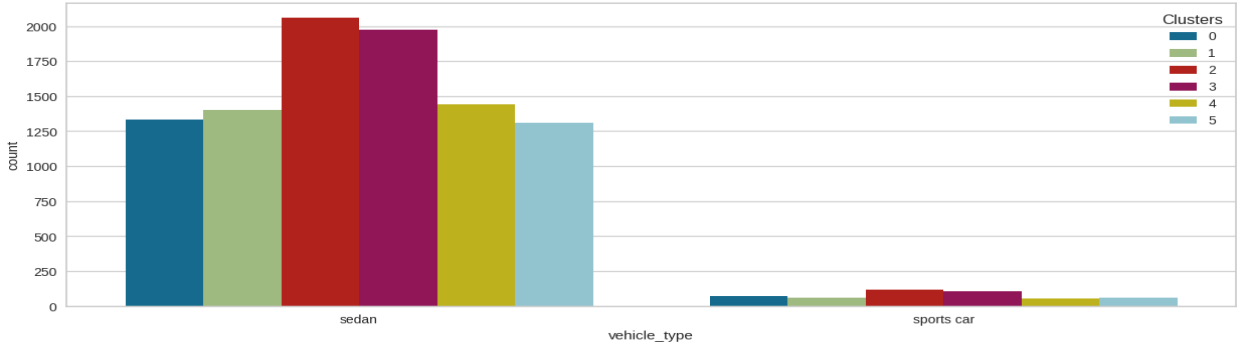
Cluster 2- Customers with relatively high annual mileage

Cluster 3- Customers with average annual mileage

Cluster 4- Customers with relatively high annual mileage

Cluster 5- Customers with higher vehicle annual mileage

Vehicle Type distribution



Cluster 0- More Sedan car owners than sports cars.

Cluster 1- More Sedan car owners than sports cars.

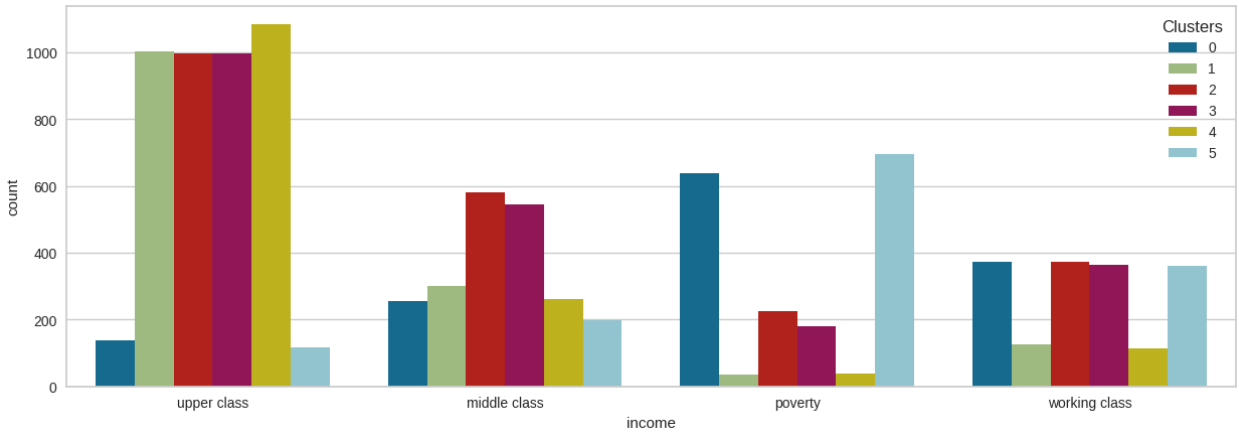
Cluster 2- More Sedan car owners than sports cars. Has the highest number of sports car users.

Cluster 3- More Sedan car owners than sports cars.

Cluster 4- More Sedan car owners than sports cars.

Cluster 5- More Sedan car owners than sports cars.

Clusters Income Distribution



Cluster 0- Low number of customers in upper class, relatively low number of customers in middle class, high number of customers in poverty class and working class.

Cluster 1- High number of customers in upper class, relatively high number of customers in middle class, very low number of customers in poverty class and low number of working class customers.

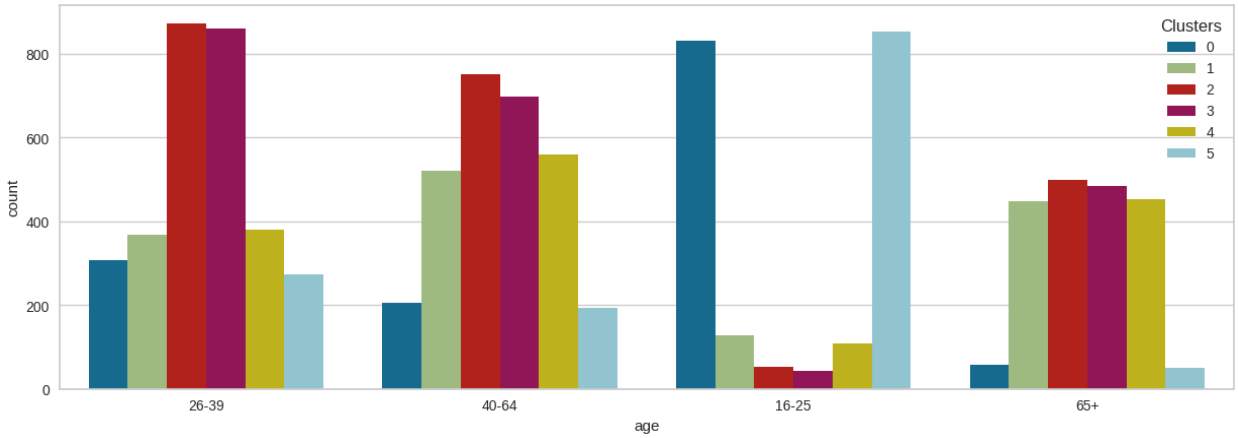
Cluster 2- High number of customers in upper class, high number of customers in middle class,relatively low number of customers in poverty class and average number of working class customers.

Cluster 3- High number of customers in upper class, high number of customers in middle class,relatively low number of customers in poverty class and average number of working class customers.

Cluster 4- Very high number of customers in upper class, low number of customers in middle class,very low number of customers in poverty class and low number of working class customers.

Cluster 5- Very low number of customers in upper class,low number of customers in middle class, very high number of customers in poverty class and high number of working class.

Clusters Age distribution



Cluster 0- More customers in age bracket 16-25, some customers in 26-39 and 40-64 age brackets, least customers in 65+

Cluster 1- More customers in age bracket 40-54, 65+ and relatively high numbers in 26-39 age bracket, least customers in 16-25.

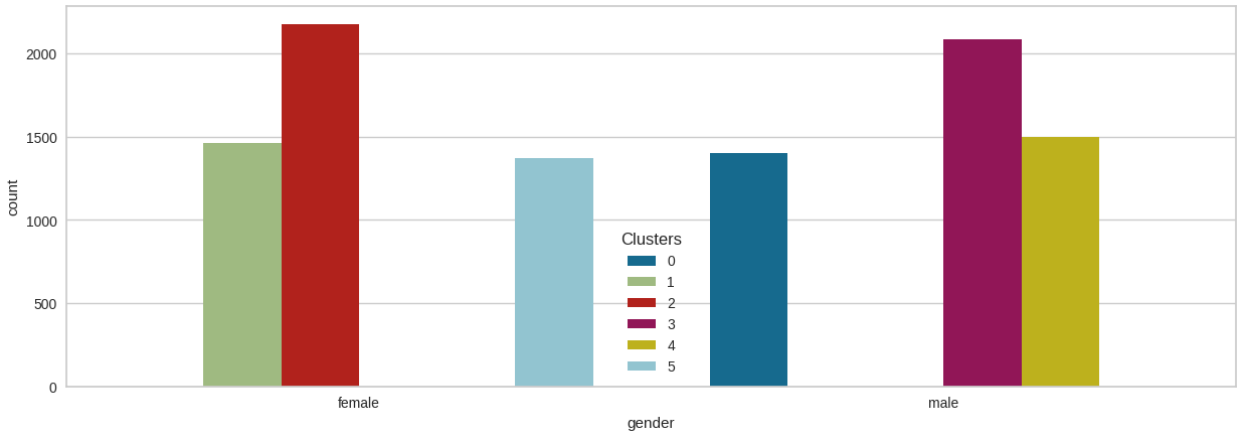
Cluster 2- More customers in age bracket 26-39, 40-64 and relatively high numbers in 65+, least customers in 16-25.

Cluster 3- More customers in age bracket 26-39, 40-64 and relatively high numbers in 65+, least customers in 16-25.

Cluster 4- More customers in age bracket 40-64, 65+ and some customers in 26-39 age bracket, least customers in 65+

Cluster 5- More customers in age bracket 16-25, some customers in 26-39 and 40-64 age brackets, least customers in 65+

Gender Distribution



Cluster 0- Male customers.

Cluster 1- Female customers.

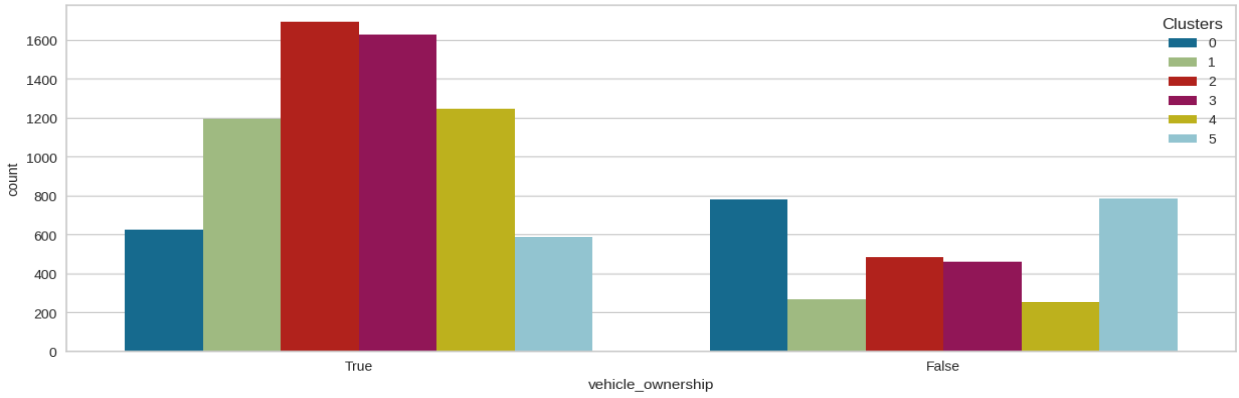
Cluster 2- Female customers.

Cluster 3- Male customers.

Cluster 4- Male customers.

Cluster 5- Female customers.

Clusters Vehicle Ownership



Cluster 0- Most of the customers do not own cars. Some own cars

Cluster 1- More car owners than otherwise.

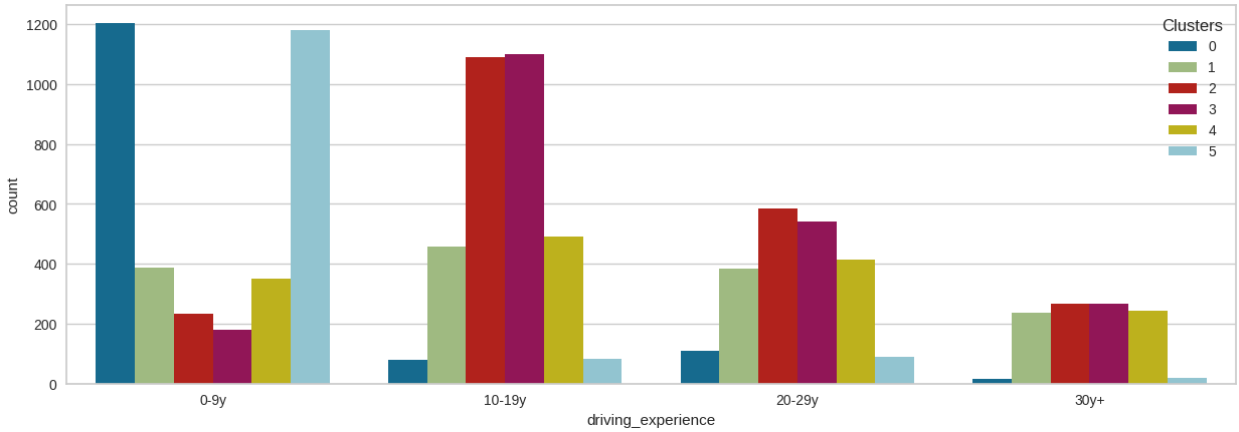
Cluster 2- More car owners than otherwise.

Cluster 3- More car owners than otherwise.

Cluster 4- More car owners than otherwise.

Cluster 5- Most of the customers do not own cars. Some own cars.

Clusters Driving Experience



Cluster 0- Most of the customers have driving experience between 0-9years. Cluster contains the least customers with driving experience of 30y+.

Cluster 1- Most of the customers have driving experience between 10-19years and some customers with 0-9 years, 20-29 years and 30y+ driving experience.

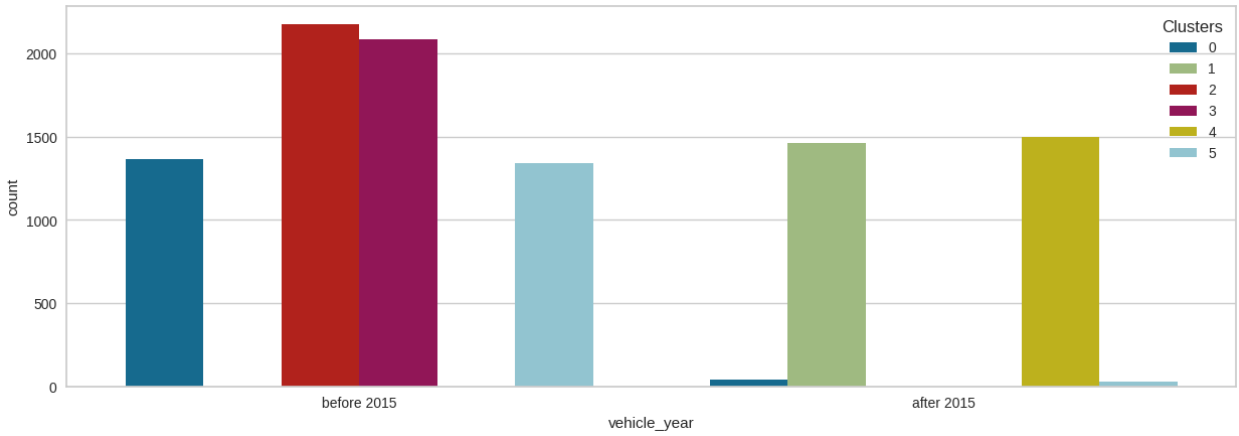
Cluster 2- Most of the customers have driving experience between 10-19years and some customers with 20-29 years driving experience. Relatively low count of customers with 0-9 years and 30y+ driving experience.

Cluster 3- Most of the customers have driving experience between 10-19years and some customers with 20-29 years driving experience. Relatively low count of customers with 0-9 years and 30y+ experience.

Cluster 4- Most of the customers have driving experience between 10-19years and some customers with 0-9 years, 20-29 years and 30y+ driving experience.

Cluster 5- Most of the customers have driving experience between 0-9years. Relatively low count of customers with 10-19 years, 20-29years and 30y+ driving experience.

Clusters Vehicle year



Cluster 0- Own more vehicles with model year before 2015 and very few 'after 2015'.

Cluster 1- Own only vehicles with model year after 2015.

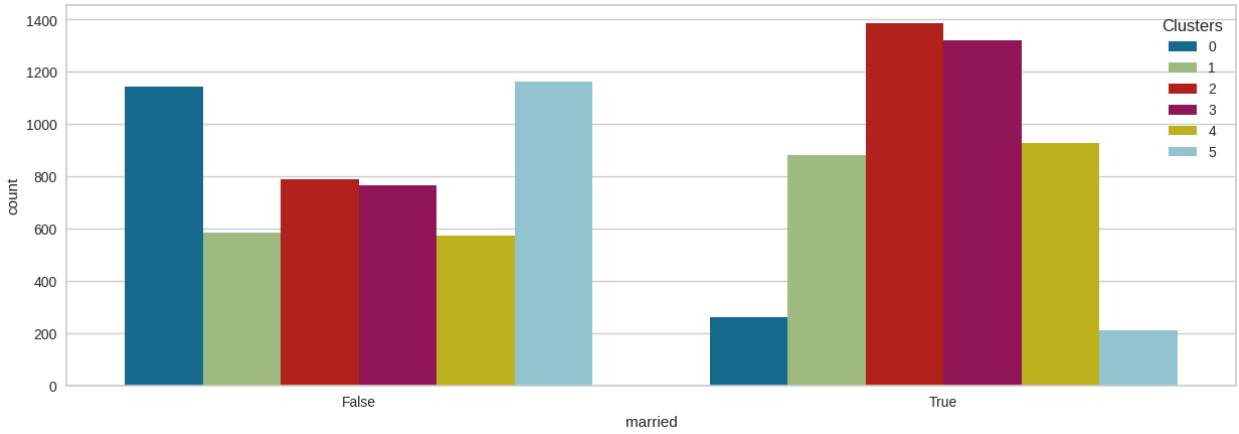
Cluster 2- Own only vehicles with model year before 2015.

Cluster 3- Own only vehicles with model year before 2015.

Cluster 4- Own only vehicles with model year after 2015.

Cluster 5- Own more vehicles with model year before 2015 and very few 'after 2015'.

Clusters “Married status” distribution



Cluster 0 - More unmarried customers than married.

Cluster 1- More married customers than unmarried.

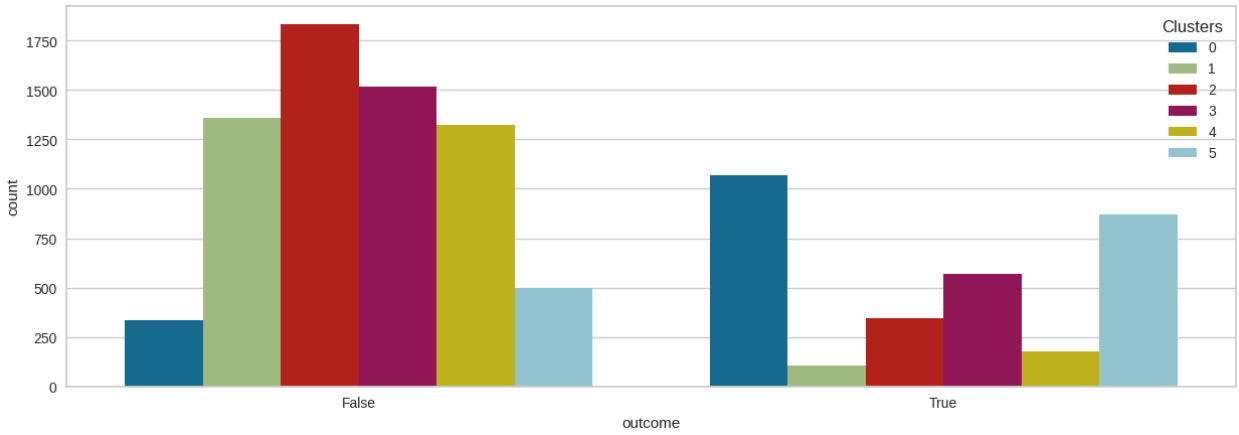
Cluster 2- More married customers than unmarried.

Cluster 3- More married customers than unmarried.

Cluster 4- More married customers than unmarried.

Cluster 5- More unmarried customers than married.

Clusters “Past Outcome” distribution



Cluster 0- More claims filed in the past year than not. Filed the most claims between the clusters.

Cluster 1- Less claims filed in the past year. Filed the least claims between the clusters.

Cluster 2- Less claims filed in the past year.

Cluster 3- Less claims filed in the past year.

Cluster 4- Relatively less claims filed in the past year.

Cluster 5- More claims filed in the past year than not.

Building Customers Profile based on Clustering of Important Variables.

Cluster 0 (Cobalt)

- Most of the customers have driving experience between 0-9years. Cluster contains the least customers with driving experience of 30y+.
- Most of the customers do not own cars. Some own cars.
- Own more vehicles with model year before 2015 and very few 'after 2015'.
- Male customers.
- Customers with higher vehicle annual mileage.
- Customers with relatively high credit scores.
- More unmarried customers than married.
- More claims filed in the past year than not. Filed the most claims between the clusters.

Cluster 1 (Green)

- Most of the customers have driving experience between 10-19years and some customers with 0-9 years, 20-29 years and 30y+ driving experience.
- More car owners than otherwise.
- Own only vehicles with model year after 2015.
- Female customers.
- Customers with relatively high annual mileage.
- Customers with high credit scores.
- More married customers than unmarried.
- Less claims filed in the past year. Filed the least claims between the clusters.

Cluster 2 (Red)

- Most of the customers have driving experience between 10-19years and some customers with 20-29 years driving experience. Relatively low count of customers with 0-9 years and 30y+ driving experience.
- More car owners than otherwise.
- Own only vehicles with model year before 2015.
- Female customers.
- Customers with relatively high annual mileage.
- Customers with high credit scores.
- More married customers than unmarried.
- Less claims filed in the past year.

Cluster 3 (Magenta)

- Most of the customers have driving experience between 10-19years and some customers with 20-29 years driving experience. Relatively low count of customers with 0-9 years and 30y+ experience.
- More car owners than otherwise.
- Own only vehicles with model year before 2015.
- Male customers.
- Customers with average annual mileage.
- Customers with relatively high credit scores.
- More married customers than unmarried.
- Less claims filed in the past year.

Cluster 4 (Lime)

- Most of the customers have driving experience between 10-19years and some customers with 0-9 years, 20-29 years and 30y+ driving experience.
- More car owners than otherwise.
- Own only vehicles with model year after 2015.
- Male customers.
- Customers with relatively high annual mileage.
- Customers with high credit scores.
- More married customers than unmarried.
- Relatively less claims filed in the past year.

Cluster 5 (Blue)

- Most of the customers have driving experience between 0-9years. Relatively low count of customers with 10-19 years, 20-29years and 30y+ driving experience.
- Most of the customers do not own cars. Some own cars.
- Own more vehicles with model year before 2015 and very few 'after 2015'.
- Female customers.
- Customers with higher vehicle annual mileage.
- Customers with average credit scores.
- More unmarried customers than married.
- More claims filed in the past year than not.

16.0 Model Recommendation

For this project, we will select Model 3 because it returns a better cluster that can be used for customer clustering.

The Limitations:

The identified limitations during the course of work are:

- It is very difficult to integrate with Categorical data, to navigate this- dummy variables were used for all categorical data.
- It does not work well with high dimensional data so PCA was applied for dimensionality reduction.

17.0 Validation and Governance

In this part of the project, we would objectively measure how well our model performs generally against fundamental errors, drifts and incorrect use. The model we have built is an unsupervised learning technique that segments auto insurance customers according to the level of risk they pose. We have made lots of assumptions with a few limitations, with this insight in mind; risks abound in forms of inaccurate data, incorrect assumptions, inappropriate conceptual framework, human error and inappropriate interpretations. We have used Decision Tree (a supervised learning technique) to select important variables to build our cluster analysis and we would be validating our model based on these important variables.

Major Sources of Risks in this Project

a. Risks emanating from choice of model
b. Risks emanating from data used
c. Risks emanating from the business and human process errors
d. Risks emanating from technology
e. Risks emanating from conceptual change and timeframe
f. Risks emanating from legal assessments and processes

Variable Level and Drift Monitoring

As mentioned above, the dataset used for this model was filtered from the original dataset using 'Variable Importance' metrics from the Decision Tree algorithm. These variables are driving experience (0-9years), Vehicle ownership, Vehicle year (after 2015), driving experience(10-19years), gender(male), vehicle year(before 2015), annual mileage, married, credit score, gender(female) in descending order. The acceptable ranges as well as actions for abnormalities for these variables are drafted in a tabular format below;

Variable	**Current Distribution**		**Acceptable Range(s)**	**Handling Missing variables**	**Tolerance for Drifting Variables**	**Handling Statistical Significance (Confusion Matrix as a statistical metric)**
	min	max				
driving experience (0-9years)	0	1	0/1	Impute Mode	More than 2% drift	Accuracy must not be below 80%
Vehicle ownership	0	1	0/1	Impute Mode	More than 2% drift	Accuracy must not be below 80%
Vehicle year (after 2015)	0	1	0/1	Impute Mode	More than 2% drift	Accuracy must not be below 80%
driving experience(10-19years)	0	1	0/1	Impute Mode	More than 2% drift	Accuracy must not be below 80%
gender(male)	0	1	0/1	Impute Mode	Not applicable	Accuracy must not be below 80%
vehicle year(before 2015)	0	1	0/1	Impute Mode	More than 2% drift	Accuracy must not be below 80%
annual mileage	2000	22000	13000	Impute Mean	More than 2% drift	Accuracy must not be below 80%
credit score	0.0534	0.9608	0.6076	Impute Mean	More than 2% drift	Accuracy must not be below 80%
gender(female)	0	1	0/1	Impute Mode	Not applicable	Accuracy must not be below 80%

18.0 Model Health and Stability

In this project, we used the Silhouette Score to determine the goodness of our clustering technique. The Silhouette Score or coefficient value ranges from -1 to 1 and the formula is shown below;

$$Silhouette\ Score = \frac{y - x}{\max(x, y)}$$

Where :

'x' is the average intra-cluster distance, that is the average distance between each point within a cluster.

'y' is the average inter-cluster distance, that is the average distance between all clusters.

For the values of the Silhouette Score, a value of 1 means that the clusters are well apart from each other and are clearly specified.

A value of 0 means that the clusters are indifferent, that is, the distance between clusters is not significant.

A value of -1 means that the clusters are assigned in the wrong way.

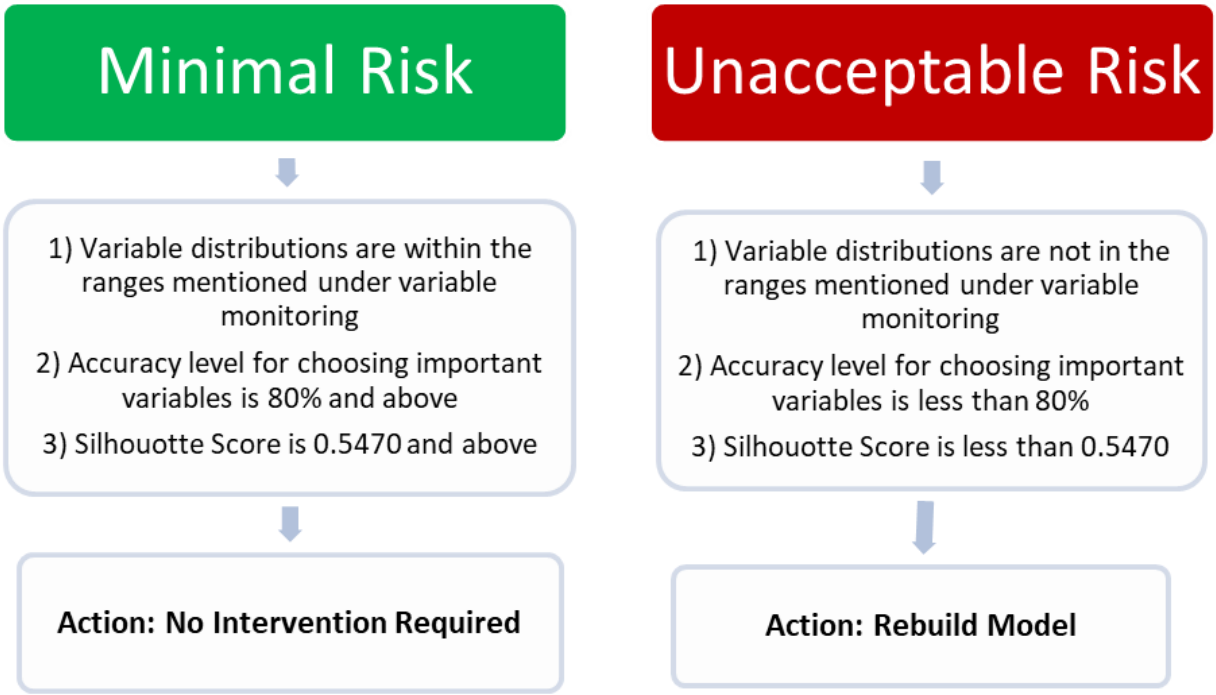
We have established that the optimum number of clusters for our dataset is 6 (k=6) and with this number of clusters, we generated a Silhouette score of 0.5470.

This score falls between 0 and 1 but closer to 1.

Due to the many assumptions and limitations on this project, as well as the nature of our data, a Silhouette score that is less than 0.5470 is unacceptable.

19.0 Model Risk Tiering

Having assessed the tolerance level of our variables, how to handle uncertainties that may occur in the future due to conceptual and or data change, as well as the stability of our model due to these mentioned changes, we would classify these potential harms according to their risk levels and actions to be taken.



Conclusion and Recommendations

We have profiled the customers in the dataset according to the risk level they have in all the features and have narrowed down their general risk levels according to the clusters in the table below.

Features	Clusters						Ratings or Scaling
	Cobalt	Green	Red	Magenta	Lime	Blue	
Driving Experience	1	2	3	4	5	6	Scale of 1 to 6
Risk Level	High	Low	Ave	Ave	Low	Low	
Vehicle Ownership	1	2	2	2	2	3	Scale of 1 to 3
Risk Level	High	High	High	High	High	Low	
Vehicle Year	Hybrid	After 2015	Before 2015	Before 2015	After 2015	Hybrid	
Risk Level	Ave	Low	High	High	Low	Ave	
Gender	Male	Female	Female	Male	Male	Female	
Risk Level	High	Low	Low	High	High	Low	
Annual Mileage	3	2	2	1	2	3	Scale of 1 to 3
Risk Level	High	High	High	Low	Low	Low	
Credit Score	2	3	3	2	3	1	Scale of 1 to 3
Risk Level	Low	Low		Low	Low	High	
Marital Status	Unmarried	Married	Married	Married	Married	Unmarried	Married/Unmarried
Risk Level	High	Low	Low	Low	Low	High	
Claims History	5	1	3	3	2	4	Scale of 1 to 5
Risk Level	High	Low	Ave	Ave	Low	High	
Risk Profiling	**High Risk Customers**	**Low Risk Customers**	**Medium Risk Customers**	**Medium Risk Customers**	**Low Risk Customers**	**High Risk Customers**	

From the table, auto insurance customers who belong to the Cobalt and Blue Clusters are considered high-risk customers. They have low driving experience compared to other clusters but have filed the most claims. High vehicle mileage impacts the overall condition of the car and performance. These customers should be well scrutinized and their insurance policies must be given strict underwriting.

The customers in the Red and Magenta clusters are the average or medium-risk customers. These customers are considered to have an average of all the most important features and as such, while they do not require strict scrutiny in terms of their policies, they need to be constantly monitored. They have more driving experience which means more experience behind the wheel and can mitigate road risks, hence, they are considered “safe”.

Lastly, the customers who are classified as low-risk customers are in the Green and Lime clusters. These customers are the ‘friends’ of insurance companies as their policies are considered a good and profitable business for the insurance companies. They have the best scores of features to be considered when underwriting a car insurance policy. Hence, their policies should be given utmost consideration and minimal premiums so as to encourage their good behaviors and records. The green cluster filed the least claims among the clusters and Lime clusters have relatively fewer claims filed. The clusters include customers with various years of driving experience.

Recommendations

During the course of this project, we ran into a lot of difficulties due to a lot of things but mostly, the nature of the datasets. Our dataset had lots of categorical variables with grouped information. Important variables such as age, income level, and driving experience were all grouped thereby poking holes in our goal of not lumping different customers together.

We would suggest running this analysis on a dataset with more numerical and continuous variables so as to properly segment these customers according to their risk levels.

We also implore the insurance industries to make their insurance datasets publicly available after removing the sensitive information. This will help to properly granulate these customers and help both the insurance companies and their customers stay happy and satisfied.

Appendix

Code sheet

Copy of Analysis_of_Auto_Insurance_Customers_using_Clustering-Final doc.ipynb - Colaboratory (google.com)

References

Bhardwaj, Ashutosh. “Silhouette Coefficient : Validating Clustering Techniques.” Medium, Towards Data Science, 27 May 2020, https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c.

georsara1. “Data Shift in Machine Learning: What Is It and How to Detect It.” Georgios Sarantitis, 24 June 2021, https://gsarantitis.wordpress.com/2020/04/16/data-shift-in-machine-learning-what-is-it-and-how-to-detect-it/.

McCormack, Caitlin. “Telematics and Car Insurance: What Is It, and What Are the Benefits?” LowestRates, LowestRates.ca, 22 Jan. 2021, https://www.lowestrates.ca/blog/auto/telematics-car-insurance-what-are-the-benefits.

JavaTPoint. (n.d.). *K-Means Clustering Algorithm*. Https://Www.Javatpoint.Com/k-Means-Clustering-Algorithm-in-Machine-Learning. Retrieved August 19, 2022, from https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning