AUTO INSURANCE CUSTOMER SEGMENTATION

MANJU PENUMARTHI

Problem Statement

- The Vehicle Insurance company is charging the same annual premium for all the policy holders (Customer's), irrespective of their vehicle, claim history, credit score, etc.

- The company wants to move towards a more dynamic pricing approach where risky drivers pay more, and less risky drivers pay less.

Objective

Segment Customer’s into multiple groups based on their risk levels for:

- Auto insurance annual premium analysis (Dynamic Pricing approach)
- Targeted marketing campaigns

Glossary for the dataset

Name	Description
pol_number	policy number for the insurance policy
pol_eff_dt	auto insurance policy effective date
gender	gender of driver: F, M
agecat	driver's age category: 1 (youngest), 2, 3, 4, 5, 6
date_of_birth	driver's date of birth
credit_score	driver’s credit score
area	driver's area of residence: A, B, C, D, E, F
traffic_index	traffic index of driver’s area of residence
veh_age	age of vehicle(categorical): 1 (youngest), 2, 3, 4

Glossary for the dataset

Name	Description
veh_body	vehicle body type
veh_value	vehicle value, in $10,000s
months_insured	number of months vehicle insurance is bought(integer)
claim_office	office location of claim handling agent: A, B, C, D
numclaims	number of claims(integer): 0 if no claim
claimcst0	claim amount: 0 if no claim
annual_premium	total charged premium i.e. the cost of insurance

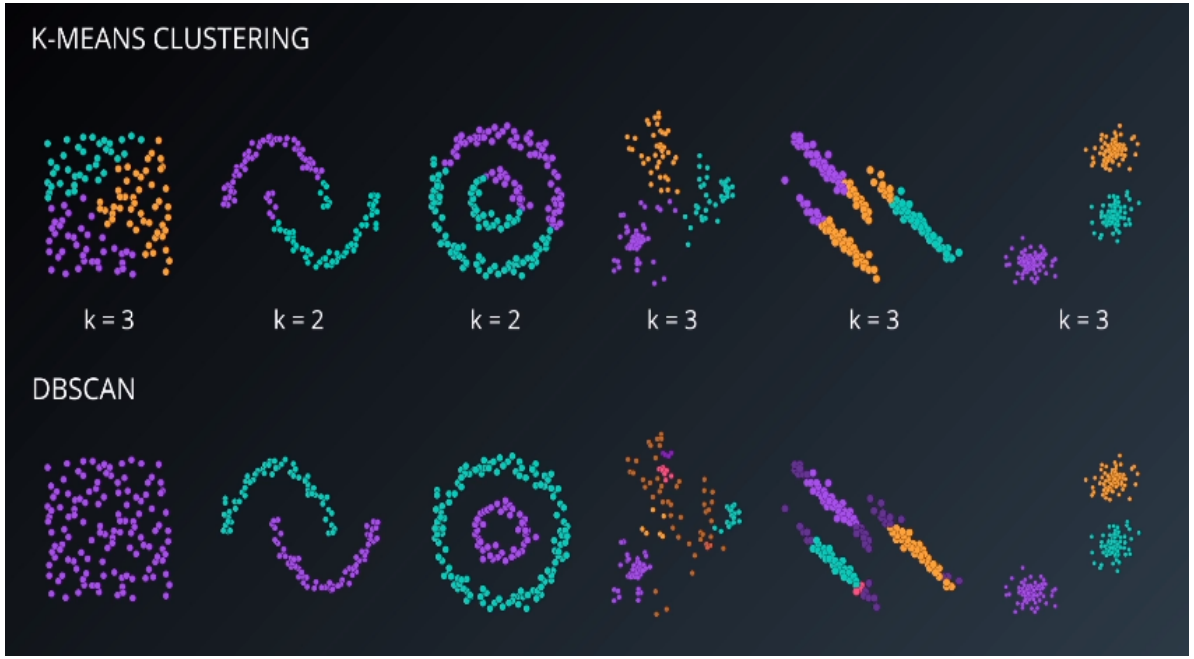
TOOLS

1. Microsoft Excel
   - Initial evaluation of data
2. IDE
   - Jupyter Notebook
3. Python
   - Data Cleaning
   - Clustering/Grouping

Approach

The Clustering algorithms I used were:

- **K-Means:** In kmeans, you initialize cluster centers and then find distance between each point and each of the cluster and then you cluster points to their nearest centers. Here the optimization problem we solve is to find the no of clusters such that sum of distances from each point and its nearest cluster is minimized.
- **DBSCAN:** DBSCAN stands for Density Based Spatial Clustering of Applications with Noise. It groups ‘densely grouped’ data points into a single cluster. It can identify clusters in large spatial datasets by looking at the local density of the data points.
- I chose K-means as the appropriate algorithm for this dataset as it presented more detailed groups that were easier to segment

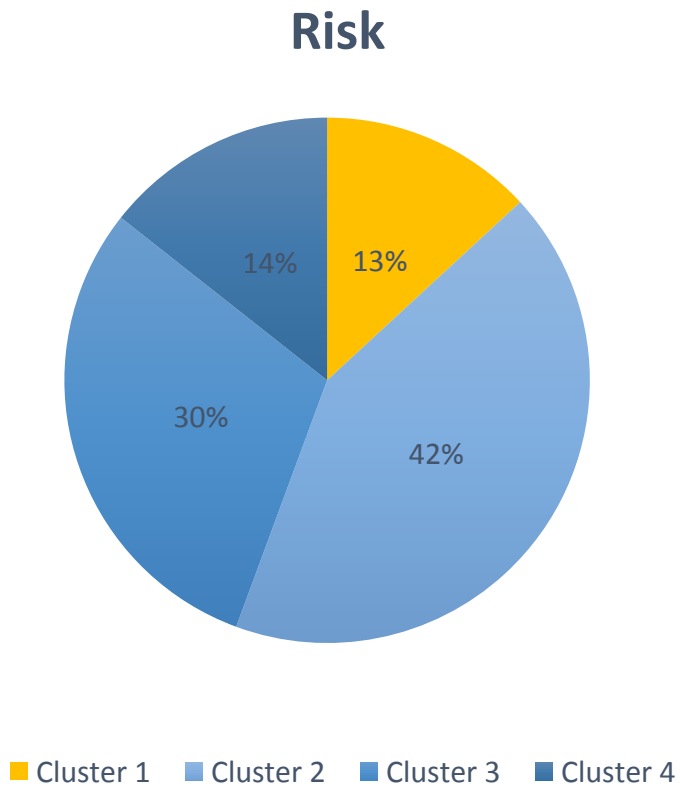


CLUSTERING RESULTS

Cluster 1 – Older Men

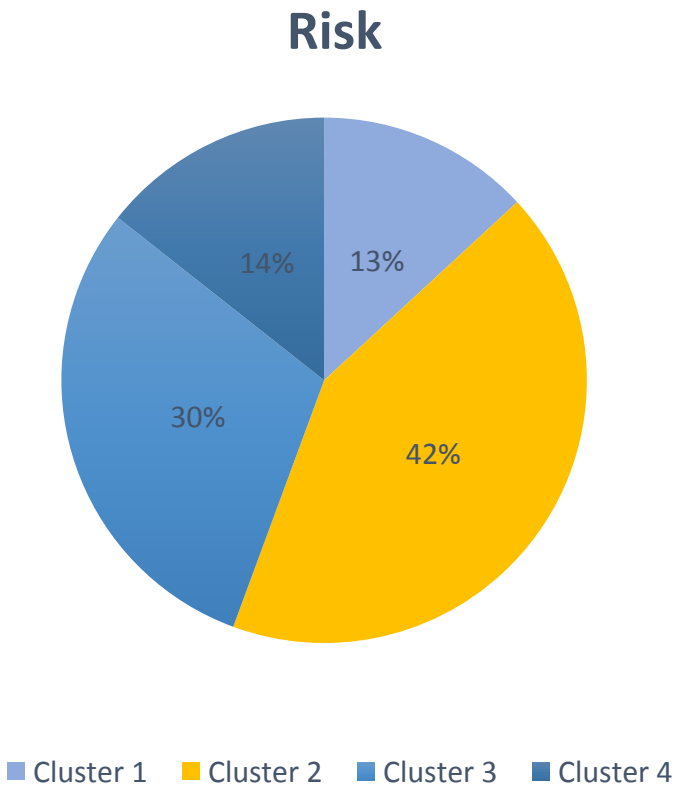
- This group consists of males who born before 1960.
- They use low range vehicles.
- Their average cost of claims is high.

Conclusion: “Above average risky drivers”



Cluster 2 – Younger Women

- This group consists of females who born on or after 1960.
- They use high range vehicles.
- Their average cost of claims is high

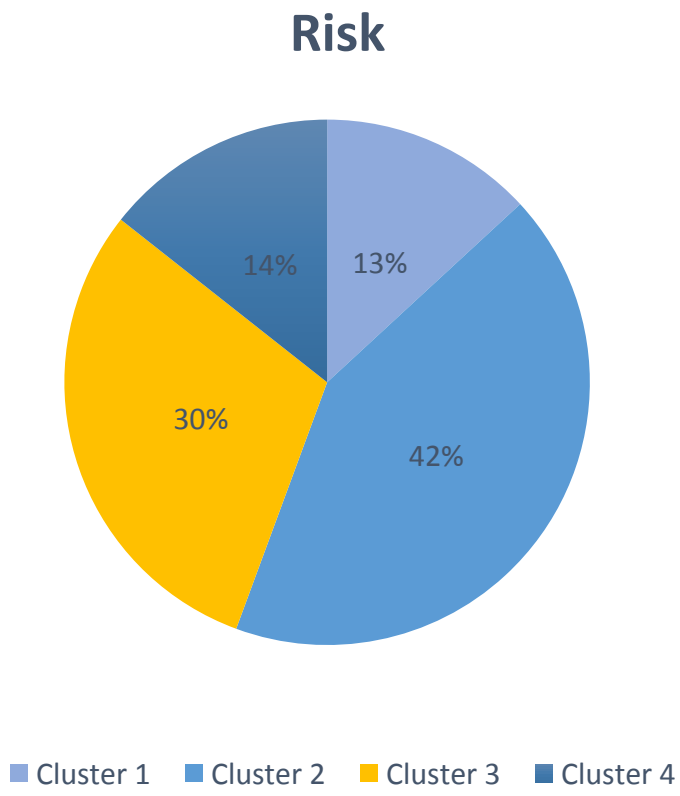


Conclusion: “High risky drivers”

Cluster 3 – Younger Men

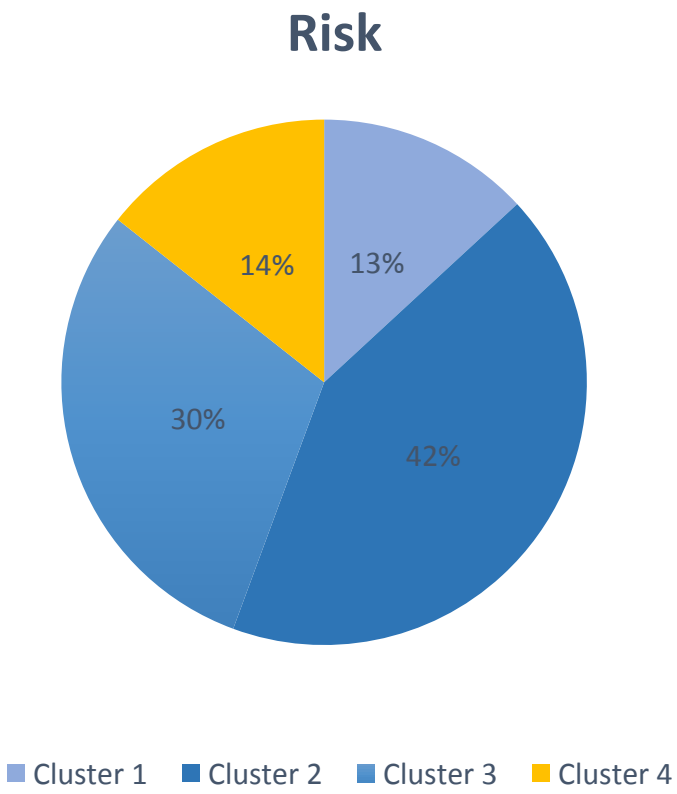
- This group consists of males who born on or after 1960.
- They use mid range vehicles.
- Their average cost of claims is low.

Conclusion: “Less risky drivers”



Cluster 4 – Older Women

- This group consists of females who born before 1960.
- They use low range vehicles.
- Their cost of claims is average.



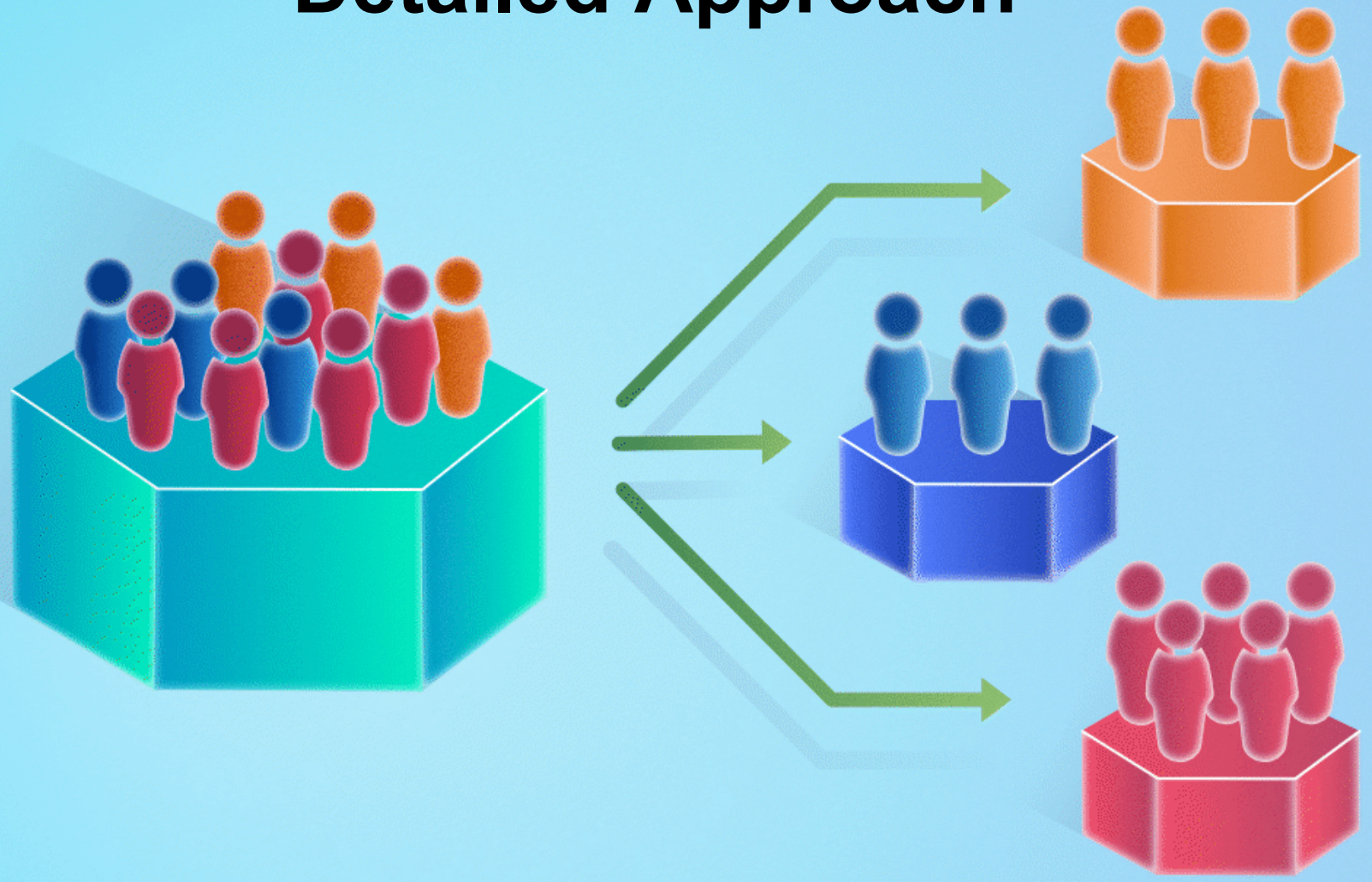
Conclusion: “Below average risky drivers”

Recommendations

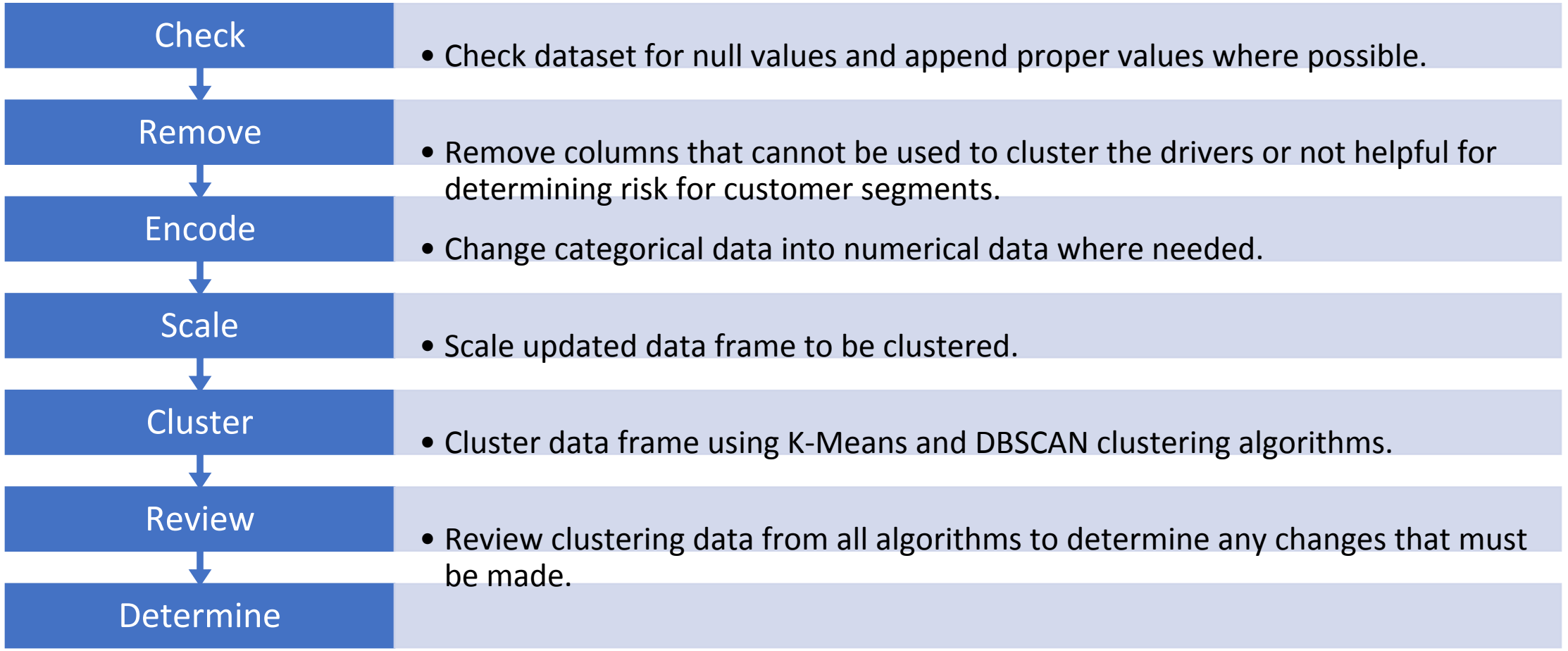
Gender	Age	Risk Level	Annual Premium
Younger Women	<= 40 years Born on or after 1960	High risky drivers	Charge higher for their annual premiums than any other group
Older Men	> 40 years Born before 1960	Above average (more than average level of risk)	Charge more than average for their annual premiums
Older Women	> 40 years Born before 1960	Below average (less than average level of risk)	Charge less than average for their annual premiums
Younger Men	<= 40 years Born on or after 1960	Less risky drivers	Charge lower for their annual premiums than any other group

Questions?

Auto Insurance Customer Segmentation
Detailed Approach



Steps


Check
• Check dataset for null values and append proper values where possible.
Remove
• Remove columns that cannot be used to cluster the drivers or not helpful for determining risk for customer segments.
Encode
• Change categorical data into numerical data where needed.
Scale
• Scale updated data frame to be clustered.
Cluster
• Cluster data frame using K-Means and DBSCAN clustering algorithms.
Review
• Review clustering data from all algorithms to determine any changes that must be made.
Determine

Cleaning Data

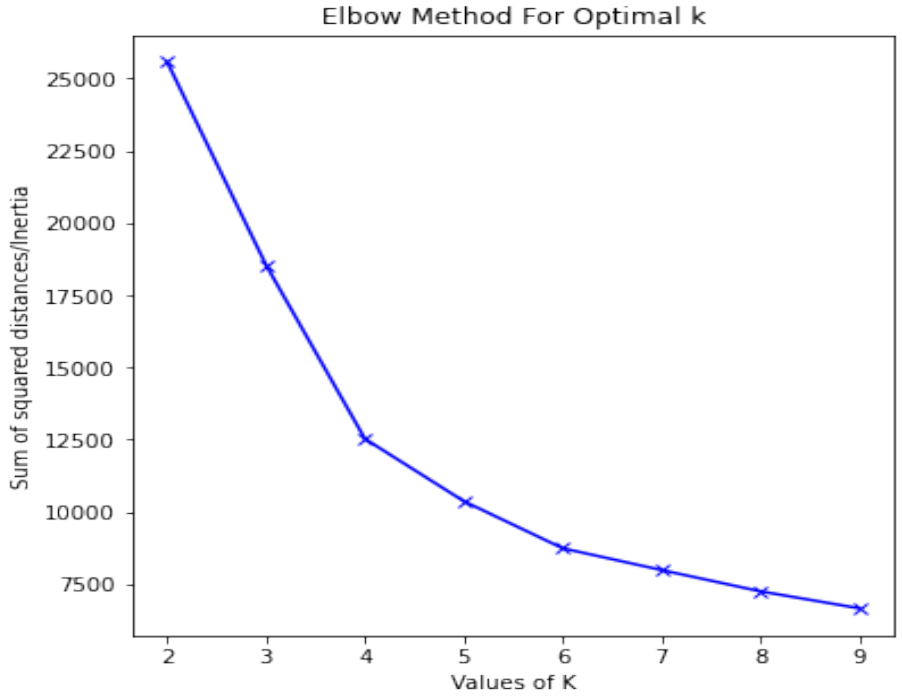
- Some columns were removed from the data frame because they did not contribute to the driver's risk:
  - Dropping "pol_number" and "pol_eff_dt" columns as they don't have any effect on customer risk level segmentation.
  - Dropping "claim_office" column as it consists of more than 80 percent of null values and has no effect on customer risk levels determination.
  - The annual premium column was removed from the data frame because the it was the same for all customers.

- Some columns were removed because they were too general to be useful:
  - Dropping "area" and "traffic_index" columns, as they don't contribute to our objective which is to apply dynamic pricing based on the customers risk level.
  - Dropping "veh_body" column as we already have "veh_age" and "veh_value" columns which helps us in further segmentation.

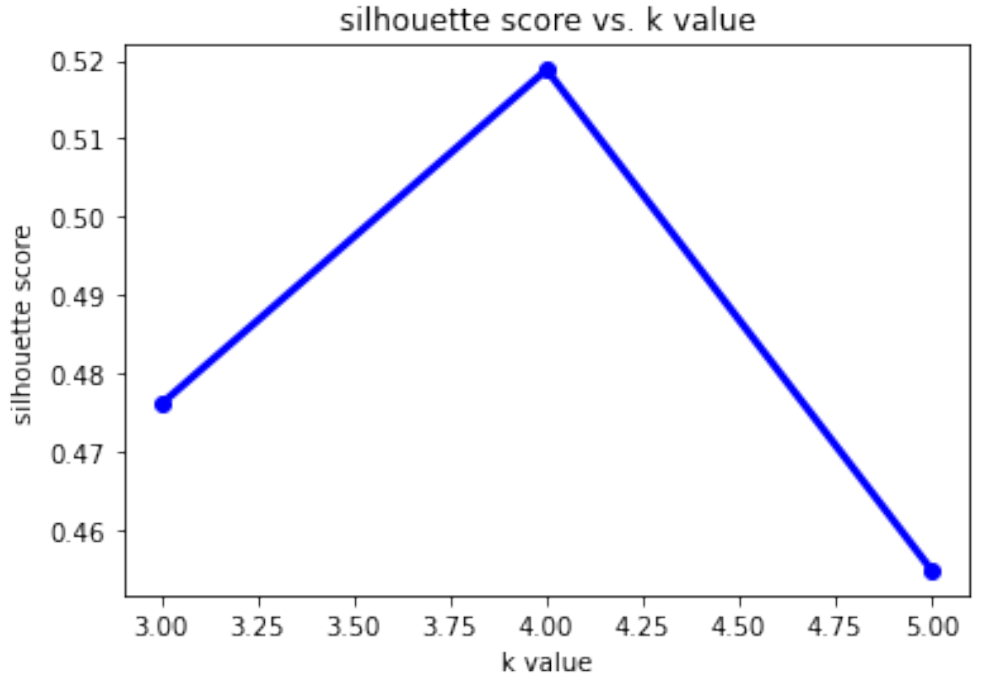
- Some columns were changed in order to cluster them in python:
  - I have used "date_of_birth" column to divide people into two categories: people born after 1960 and people born before 1960, and labelled this new column as "age_division".
  - Removed the "date_of_birth" column as the "age_division" column will give us the required information.

K-Means Clustering Algorithm

- I tested potential K values between 2 and 10 and used the elbow method to determine the best value for K.



- I also calculated the silhouette score for few K values to determine the best value for K.



In both the cases of elbow method and silhouette score I got 4 as the best value for K.

K-Means EDA

- These were the clustering results displayed using a violin plot with a k-value of 4. Based on this information, K-Means clustered the drivers based on age and gender. Their associated risk metrics are shown using claim cost and vehicle Price level in the bottom two violin plots.

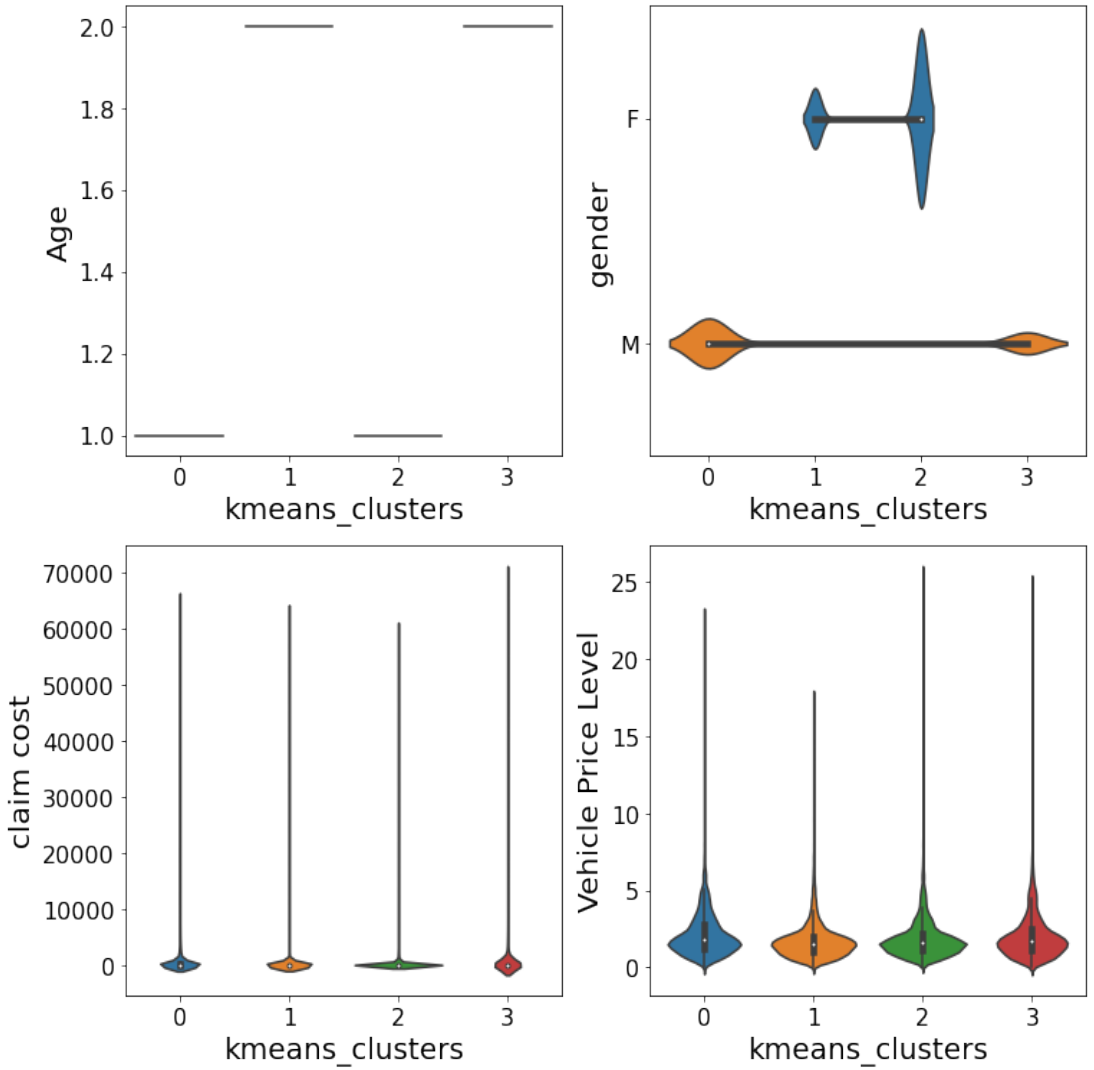
- I interpreted the results from k -means using the best k. Appended the cluster assignments to the original, unscaled dataset and performed summary statistics/visualizations on it.

    - Cluster 0 has older males (whose age is above 40 years), high claim cost history and low vehicle value.

    - Cluster 1 has younger females (whose age is below 40 years), high claim cost history and high vehicle value.

    - Cluster 2 has younger males (whose age is below 40 years), low claim cost history and median vehicle value.

    - Cluster 3 has older females (whose age is above 40 years), average claim cost history and low vehicle value.



DBSCAN Clustering Algorithm

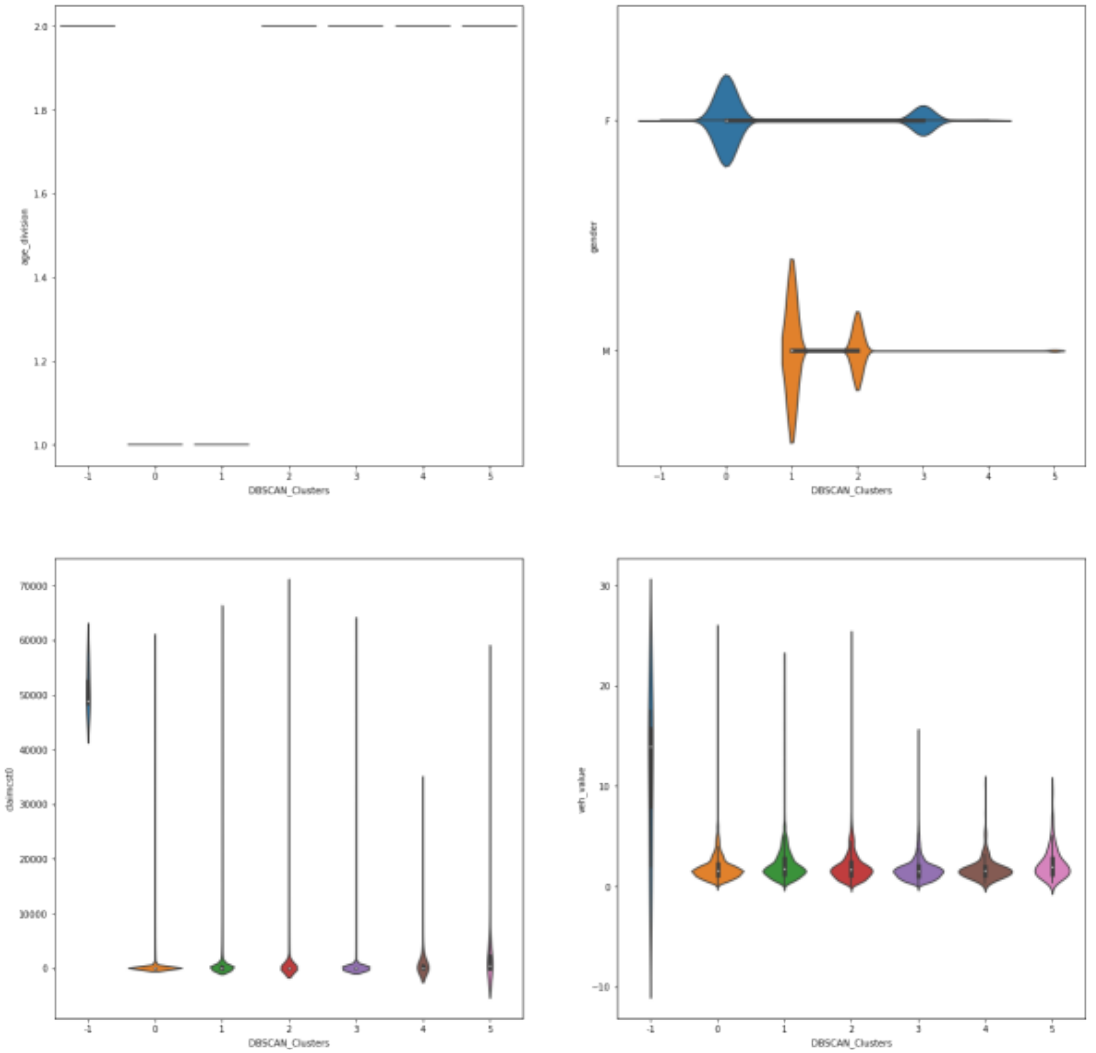
- For DBSCAN I tested all combinations of an eps of 0.3, 0.5, and 0.7 and a minpts of 8, 16, and 24.
- These are the silhouette scores received for each trial. According to this the best possible eps is 0.5 and minpts is 8.

	0	1	silhouette_score	num1_clusters
3	0.5	8	0.480339	7
5	0.5	24	0.480182	7
4	0.5	16	0.480182	7
6	0.7	8	0.478224	7
7	0.7	16	0.478224	7
8	0.7	24	0.478224	7
2	0.3	24	0.281592	24
1	0.3	16	0.281459	24
0	0.3	8	0.281409	25

DBSCAN EDA

- These are the violin plots associated with the clusters received from DBSCAN.

- My interpretation is that these clusters are informative, but not very clear. With an eps of 0.5 and minpts of 8, I got a better and highest silhouette score of 0.480339 with a total number of 7 clusters.



Conclusion

- K-Means gave clusters that were more precise for customer segmentation in determining their riskiness level and for marketing campaigns as well.

- There are many different clustering algorithms that could be used to cluster the drivers that are insured, but from the knowledge and tools available K-Means was found to be the best algorithm for this application.

Questions?