
Insurance

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Communication
- Conclusion

EXECUTIVE SUMMARY

- In this project, we will predict who would be interested in buying a caravan insurance policy using several machine learning classification algorithms.
- The main steps in this project include:
  - Data collection, wrangling, and formatting
  - Exploratory data analysis
  - Interactive data visualization
  - Machine learning prediction
- Our graphs show that some multivariate features of the rocket launches have a correlation with the outcome of the decision made by the client.
  - It is also concluded that a random forest classifier may be the best machine learning algorithm to predict if the a client will purchase a caravan insurance policy.

INTRODUCTION

- This data used in the project contains information on customers of an insurance company. The data consists of 86 features and includes product usage data and socio-demographic data. Information about customers consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes.
- The training set contains over 5000 descriptions of customers, including the information of whether or not they have a caravan insurance policy. A test set contains 4000 customers of whom only the organisers know if they have a caravan insurance policy. Note that sociodemographic data (attribute 1-43) and product ownership (attributes 44-86).

METHODOLOGY

- The overall methodology includes:

1. Data collection, wrangling, and formatting by reading the files into dataframes and assigning appropriate feature names.

2. Exploratory data analysis (EDA), using:

- Pandas and NumPy

3. Data visualization, using:

- Matplotlib and Seaborn

4. Machine learning prediction, using

- Logistic regression
- Random Forest Classifier
- K-means clustering and PCA for data dimensionality reduction of features.

METHODOLOGY
Data collection, wrangling, and formatting

```
df type: <class 'pandas.core.frame.DataFrame'>
df shape: (5822, 86)
```

[3]:

	MOSTYPE	MAANTHUI	MGEMOMV	MGEMLEEF	MOSHOOFD	MGODRK	MGODPR	MGODOV	MGODGE	MRELGE	...	APERSONG	AGEZONG	AWAOREG	ABRA
0	33	1	3	2	8	0	5	1	3	7	...	0	0	0	
1	37	1	2	2	8	1	4	1	4	6	...	0	0	0	
2	37	1	2	2	8	0	4	2	4	3	...	0	0	0	
3	9	1	3	3	3	2	3	2	4	5	...	0	0	0	
4	40	1	4	2	10	1	4	1	4	7	...	0	0	0	

5 rows × 86 columns

METHODOLOGY
Data collection, wrangling, and formatting

- The data contains 5,822 observations and 86 features, inclusive of the target variable "CARAVAN" and the dataset has no missing values.

- All features are given as integers but have respective categories since they are multivariate features i.e. MOSTYPE (Customer Subtype – socioeconomic class), MGEMLEEF(Age Category), PWAPART(Contribution to private third party insurance) and MOSHOOFD(Segments by Lifestyle).

Exploratory Data Analysis

```
df['CARAVAN'].value_counts(normalize=True)
```

```
CARAVAN
0    0.940227
1    0.059773
Name: proportion, dtype: float64
```

Exploratory Data Analysis …

- There is a class imbalance in the data as observed that 94% of the total clients whose information was obtained did not buy caravan insurance policies whereas about 6% had purchased caravan insurance.

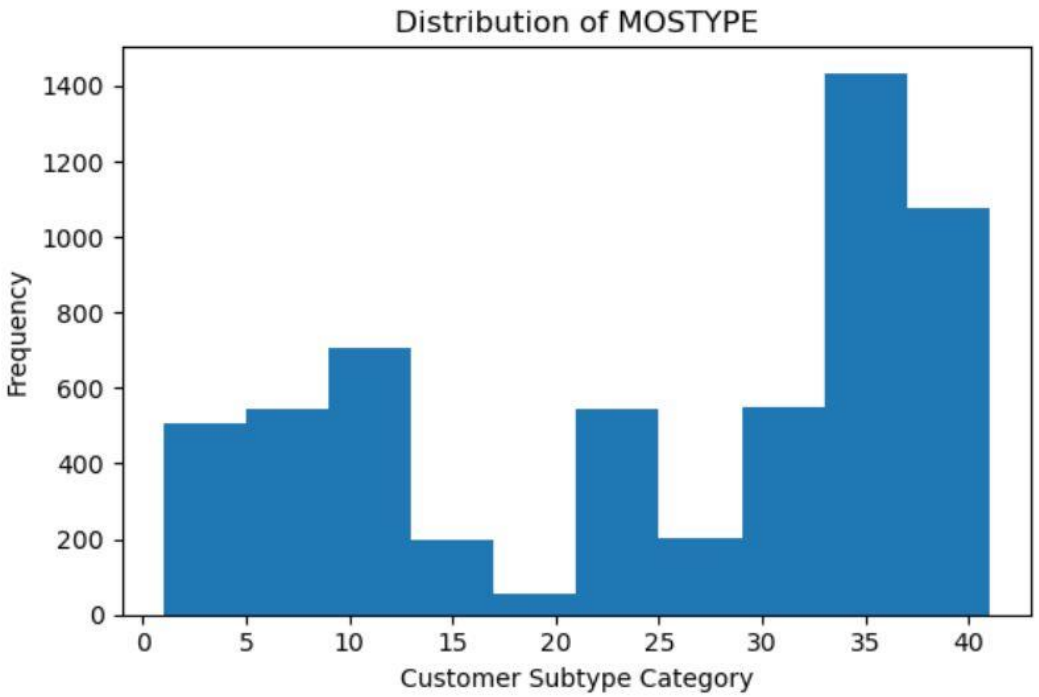
Descriptive Statistics

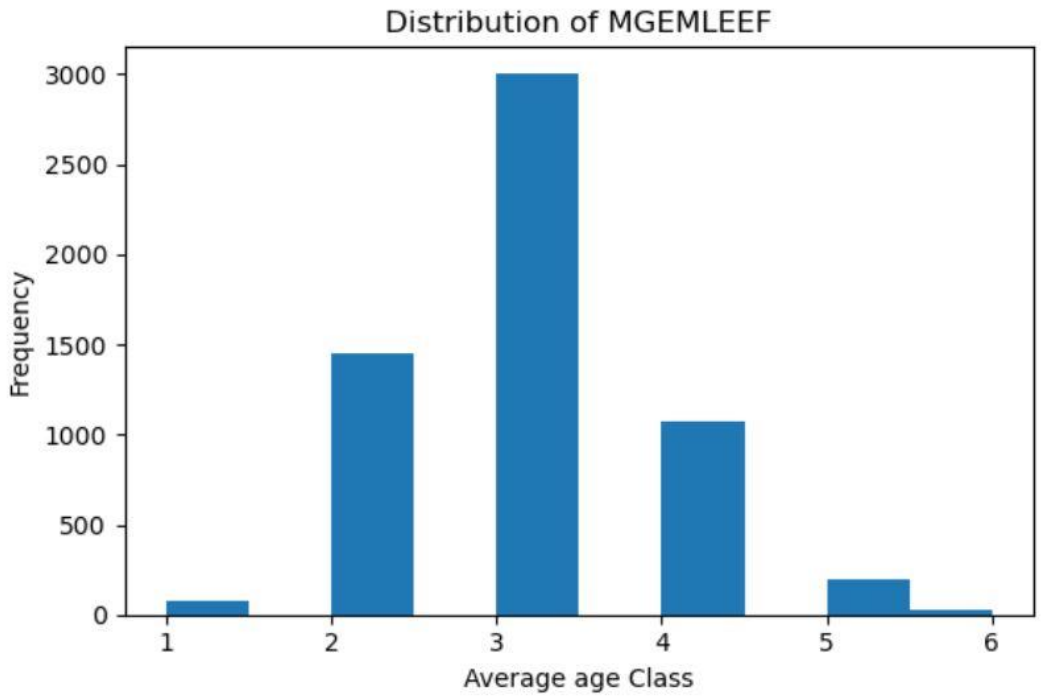
	MOSTYPE	MGEMLEEF	MOSHOOFD	MGODRK	PWAPART
count	5822.000000	5822.000000	5822.000000	5822.000000	5822.000000
mean	24.253349	2.991240	5.773617	0.696496	0.771213
std	12.846706	0.814589	2.856760	1.003234	0.958623
min	1.000000	1.000000	1.000000	0.000000	0.000000
25%	10.000000	2.000000	3.000000	0.000000	0.000000
50%	30.000000	3.000000	7.000000	0.000000	0.000000
75%	35.000000	3.000000	8.000000	1.000000	2.000000
max	41.000000	6.000000	10.000000	9.000000	3.000000

Exploratory Data Analysis …

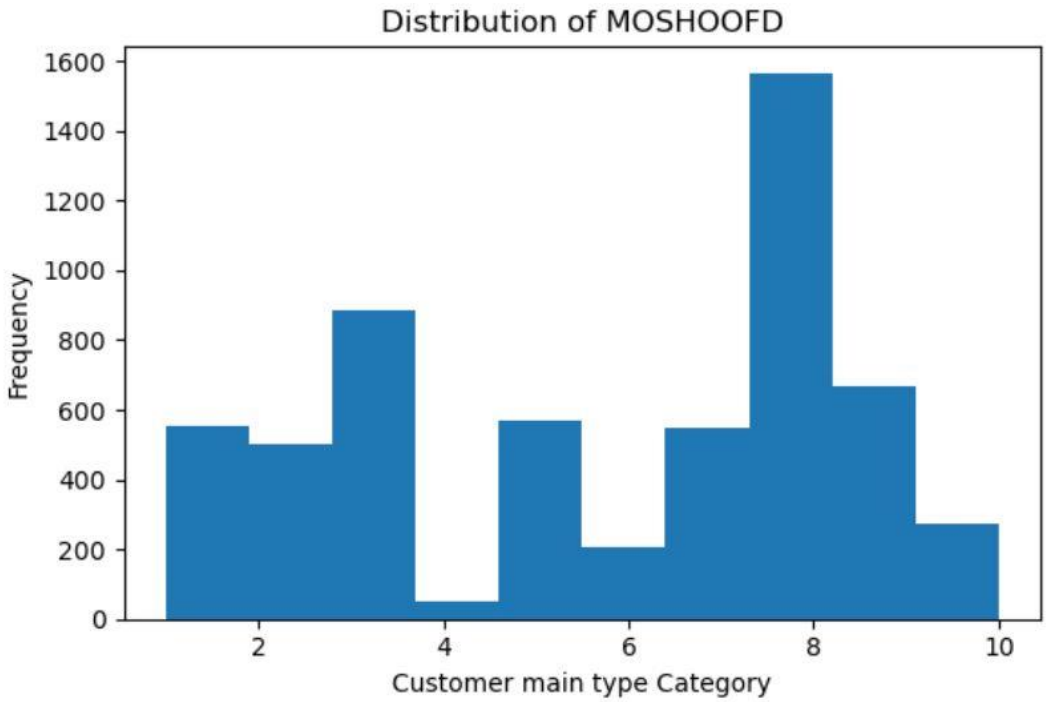
- Considering the multivariate features in the figure shown above, MOSTYPE has the highest average with clients/ customers under “Religious elderly singles” being the median class. i.e. half of the clients in the mentioned group could be considered to respond towards purchasing caravan insurance policy and 50% of them negatively responded to purchasing the named type of insurance.

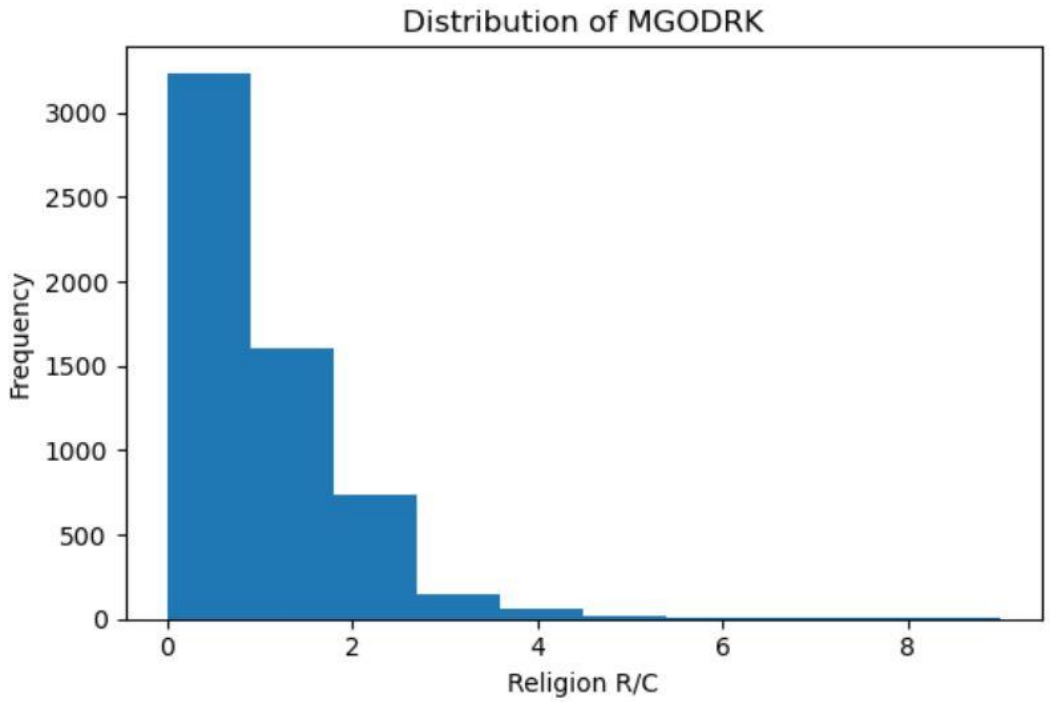
Distributions of multivariate data features



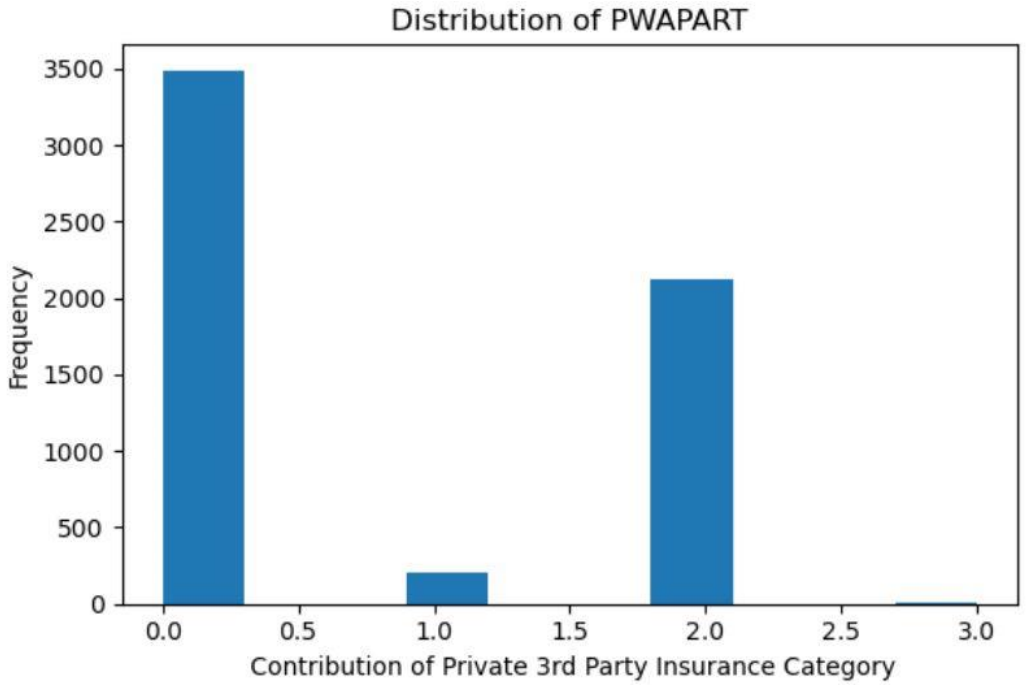


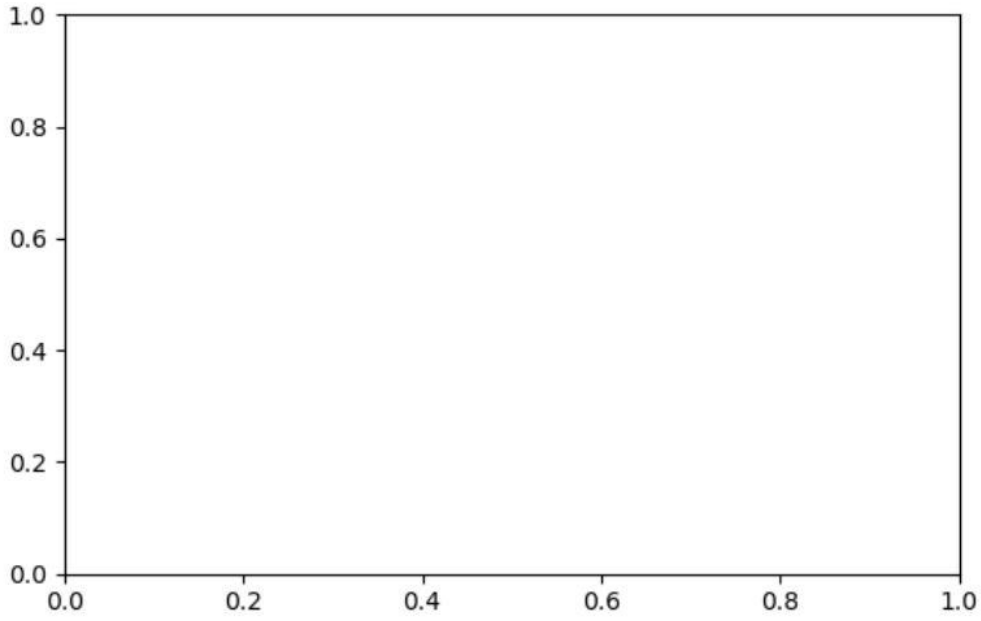
Distributions of multivariate data features





Distributions of multivariate data features





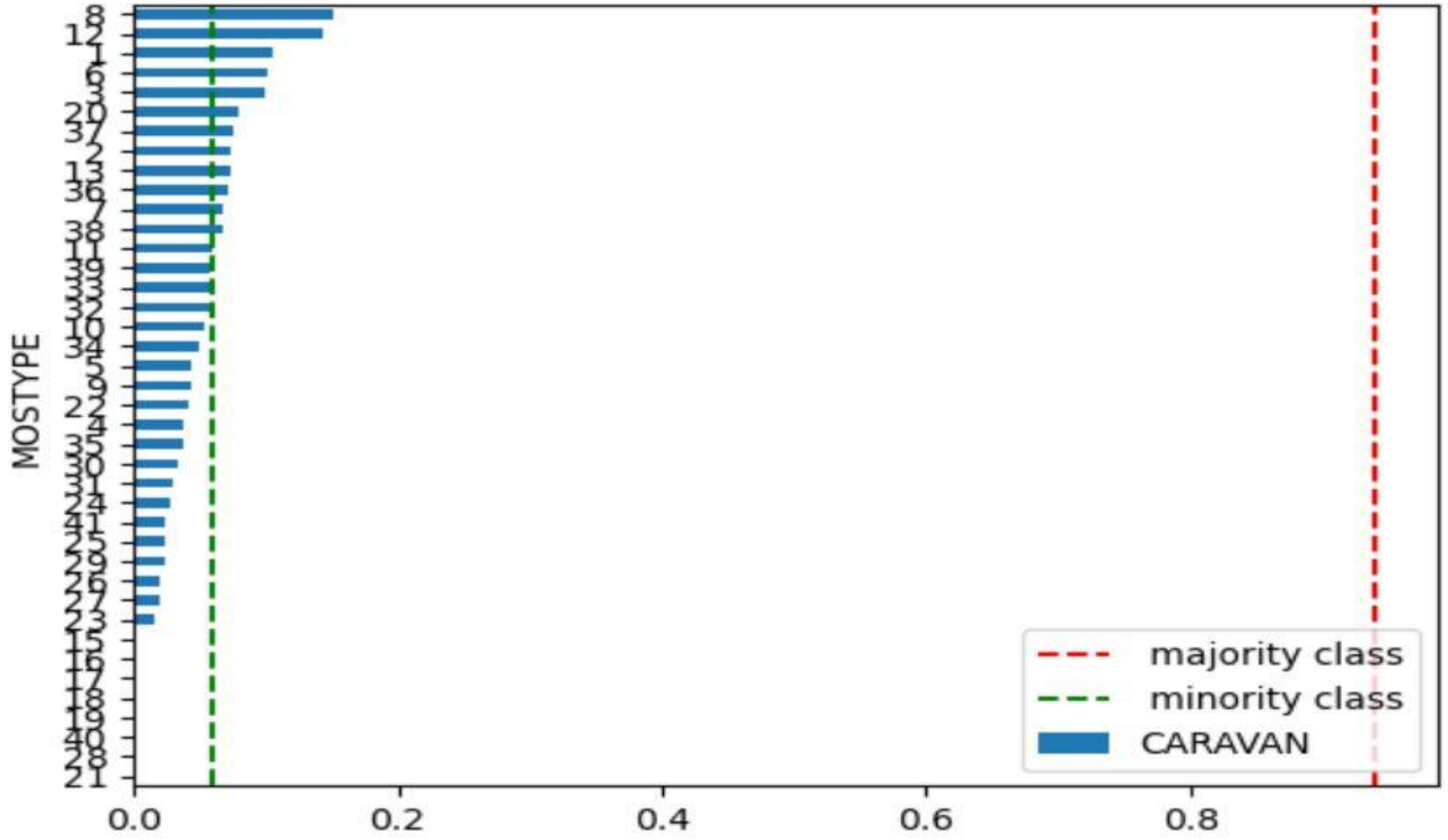
Distributions of multivariate data features explained

- **MOSTYPE**: Groups/ Categories ranging from 33 to 38 are observed to have a high likelihood towards purchasing caravan insurance.

- **MGEMLEEF**: Clients ranging from age 40-50 years are observed to have the highest likelihood to purchase caravan insurance, followed by those ranging from 30-40years. Clients whose age ranges from 20-30years have a the least likelihood of taking on this insurance arrangement.

- **MOSHOOFD:** Clients with families with grown ups tend to have a high likelihood towards purchasing caravan insurance, followed by those with an average household number. Career loners have a low response towards purchasing insurance.

Exploratory Data Analysis …



Exploratory Data Analysis …

- Clients captured in segments:-

- 8 (middle class families),
- 12 (affluent young families),
- 1 (high income, expensive child),
- 6 (career and childcare),
- 3 (high status seniors),
- 20 (ethically diverse), and 37 (mixed small town dwellers) are slightly above the minority class(positive class).

RESULTS

- The model baseline accuracy is 94%, therefore, our predictive model should be able to learn patterns in the data (also data that the model has never seen before) with a higher accuracy.

```
[9]: acc_baseline = y_train.value_counts(normalize =True).max()
     print("Baseline Accuracy:", round(acc_baseline, 2))

     Baseline Accuracy: 0.94
```

RESULTS

- Used a **StandardScaler** to preprocess the data by improving the performance of the model and interpretability of the data. The model used is a Random Forest Classifier with balanced class weights so as to harmonize the data imbalance and returned a training accuracy of 99% and test accuracy of 93%.
- The model memorized training data but did not generalize effectively on new data.

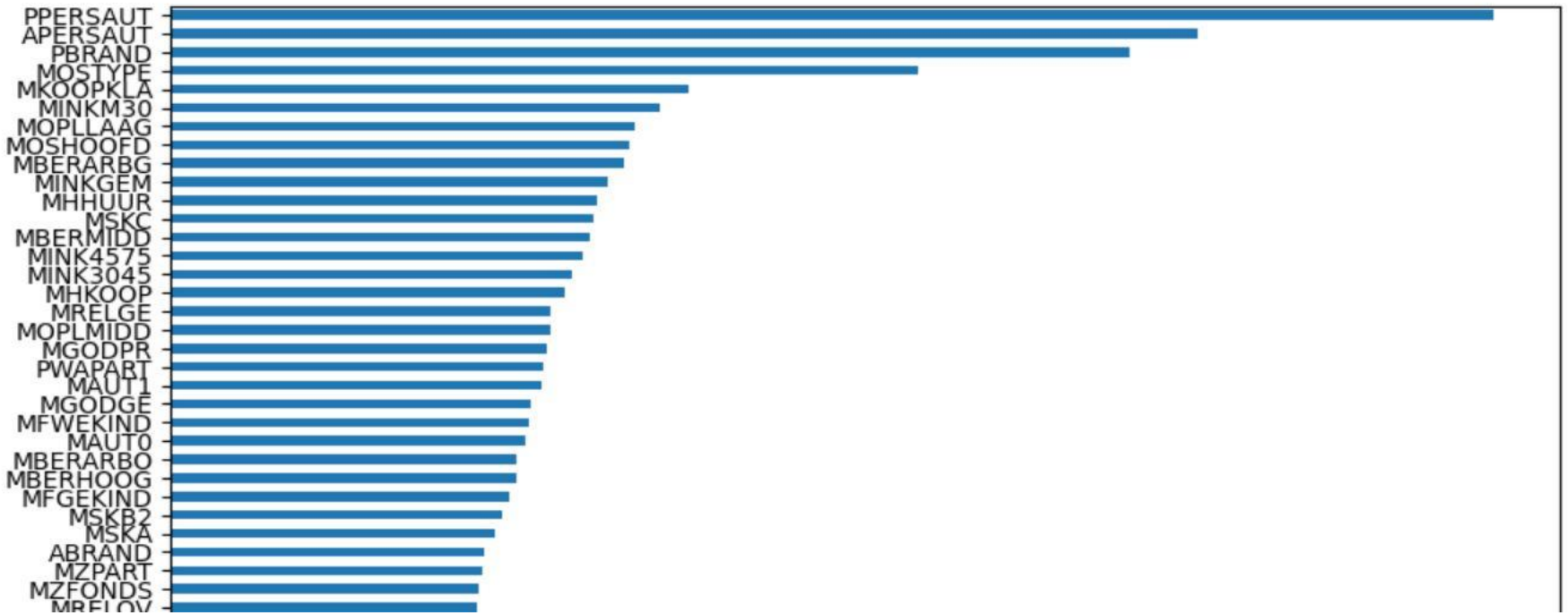
[11]:
```
acc_train = accuracy_score(y_train, model.predict(X_train))
acc_test = model.score(X_test, y_test)

print("Training Accuracy:", round(acc_train, 2))
print("Test Accuracy:", round(acc_test, 2))
```

```
Training Accuracy: 0.99
Test Accuracy: 0.93
```

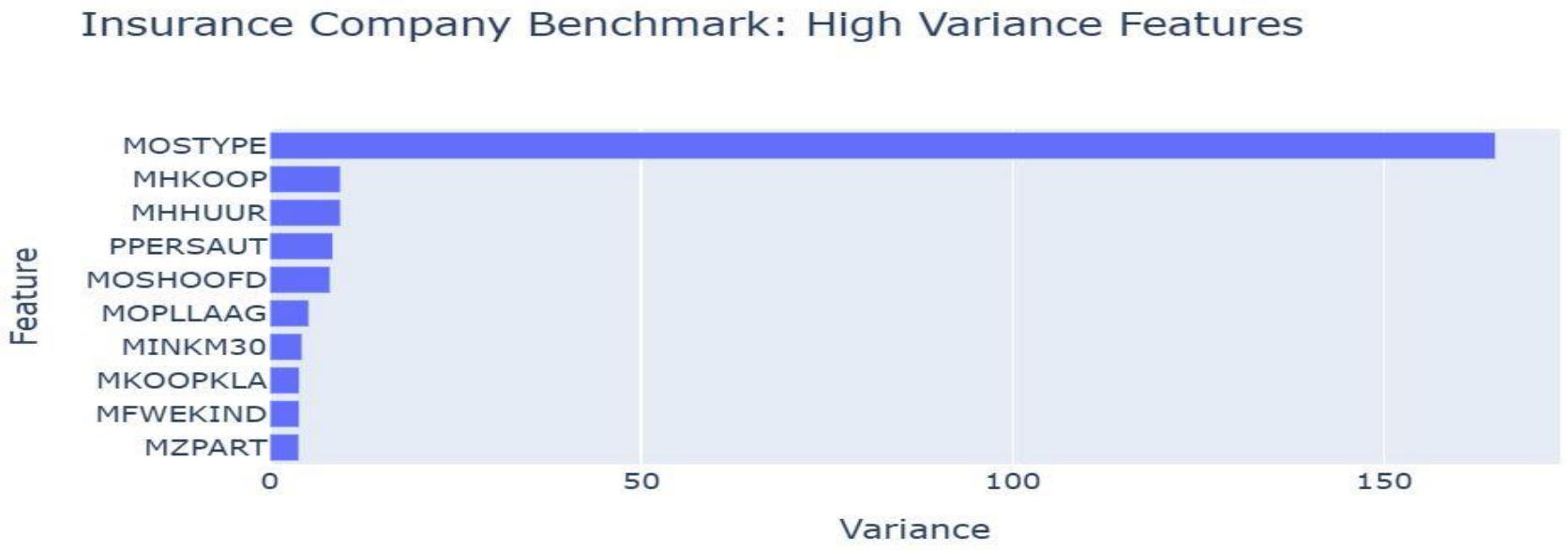
RESULTS

- The model predicted that:- PPERSAUT(6.501%), APERSAUT(5.042%), PBRAND(4.714%), MOSTYPE(3.674%), MKOOPKLA(2.546%), MINKM30(2.4%), etc. had the highest impact on the model.



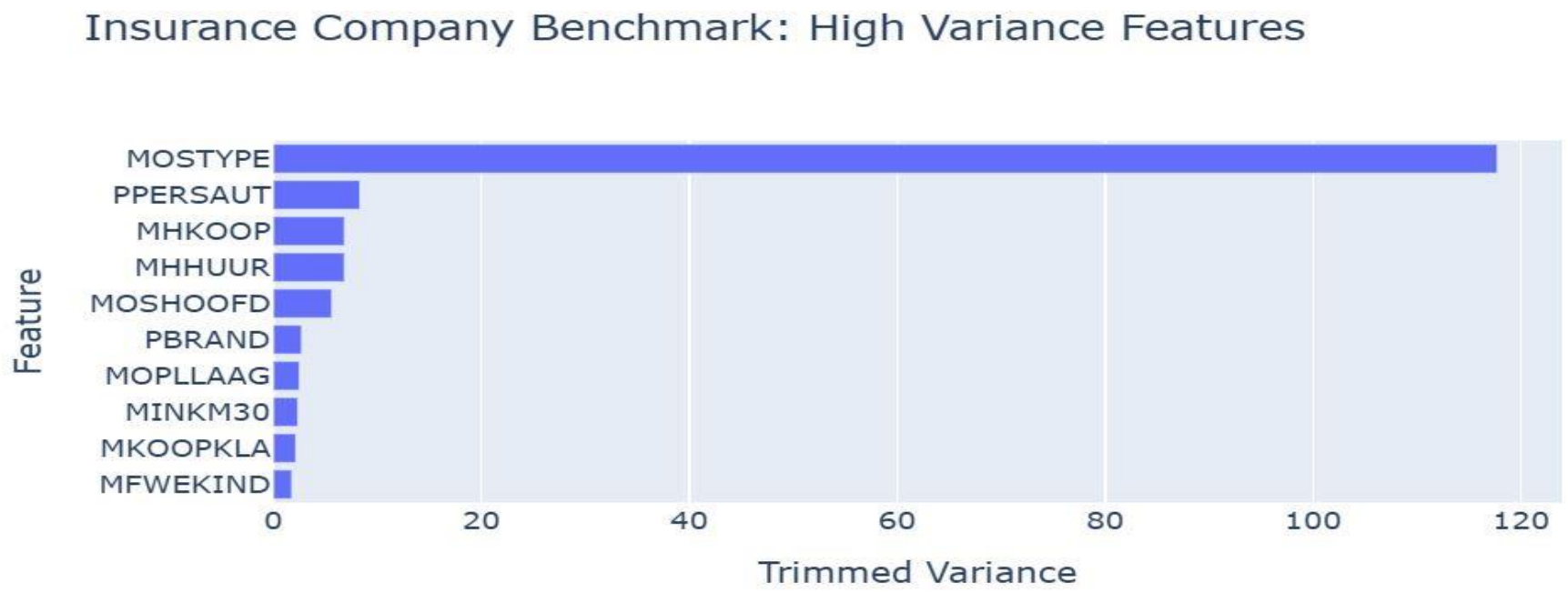
RESULTS

- MOSTYPE, MHKOOP, PPERSAUT, MHHUUR and MOSHOOFD have a high variance. This enables insurers allocate the appropriate deductibles and excess amounts on householders/homeowners and motor comprehensive insurance policies.



RESULTS

- Trimmed variance enables us to work with accurate data without having the effect of outliers as data point are widely spread.



RESULTS

- Inertia: - Is a key concept applied in clustering algorithms(K-Means) . It measures how well data points are assigned to their respective clusters by quantifying the sum of squared distances between each data point and its assigned cluster centroid.

- Low inertia implies that data points are close to their centroids,

- High inertia implies that clusters are loosely grouped, indicating poor data separation.

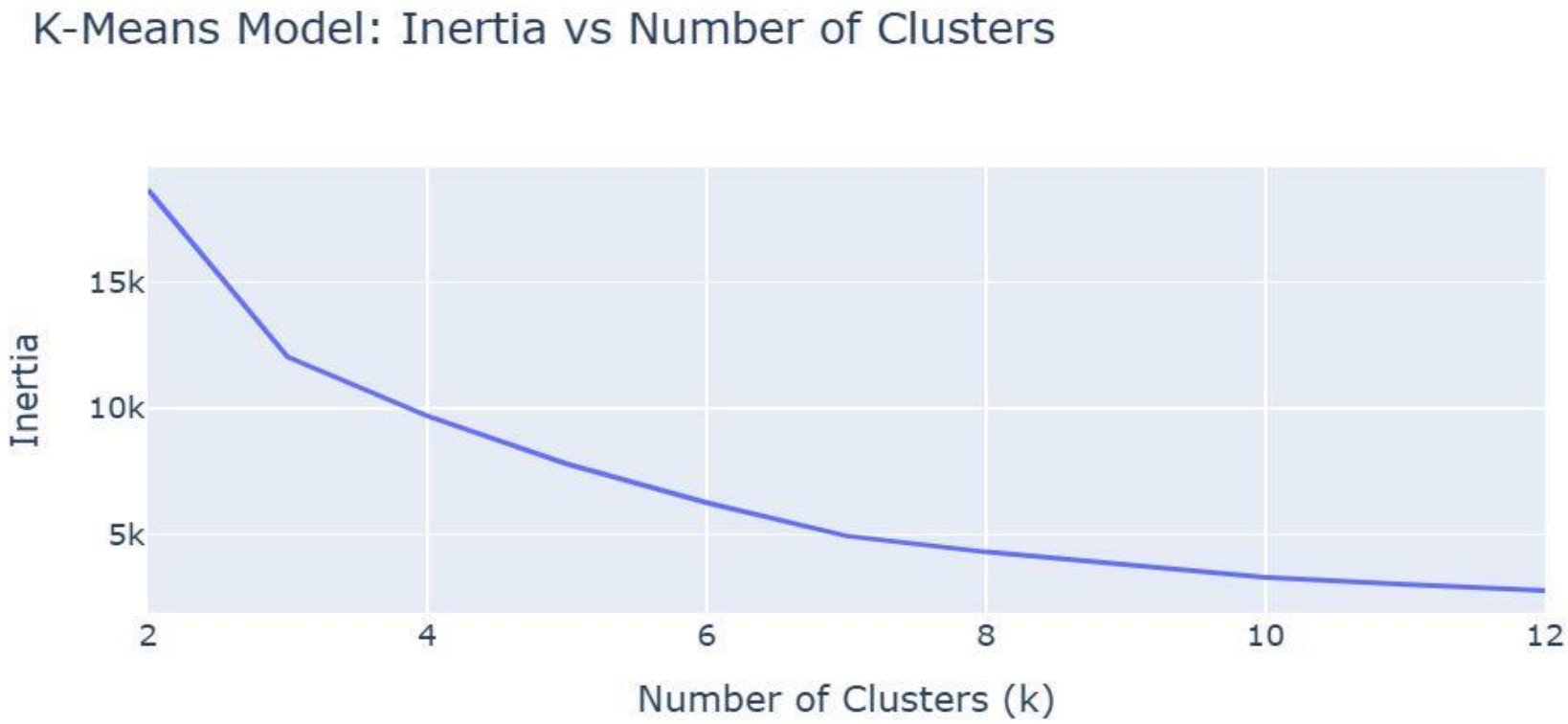
RESULTS

- Silhouette score:- Used to evaluate the quality of clusters in a clustering algorithm. It measures how similar an object is to its own cluster compared to other clusters. *(ranges from -1 to +1)*

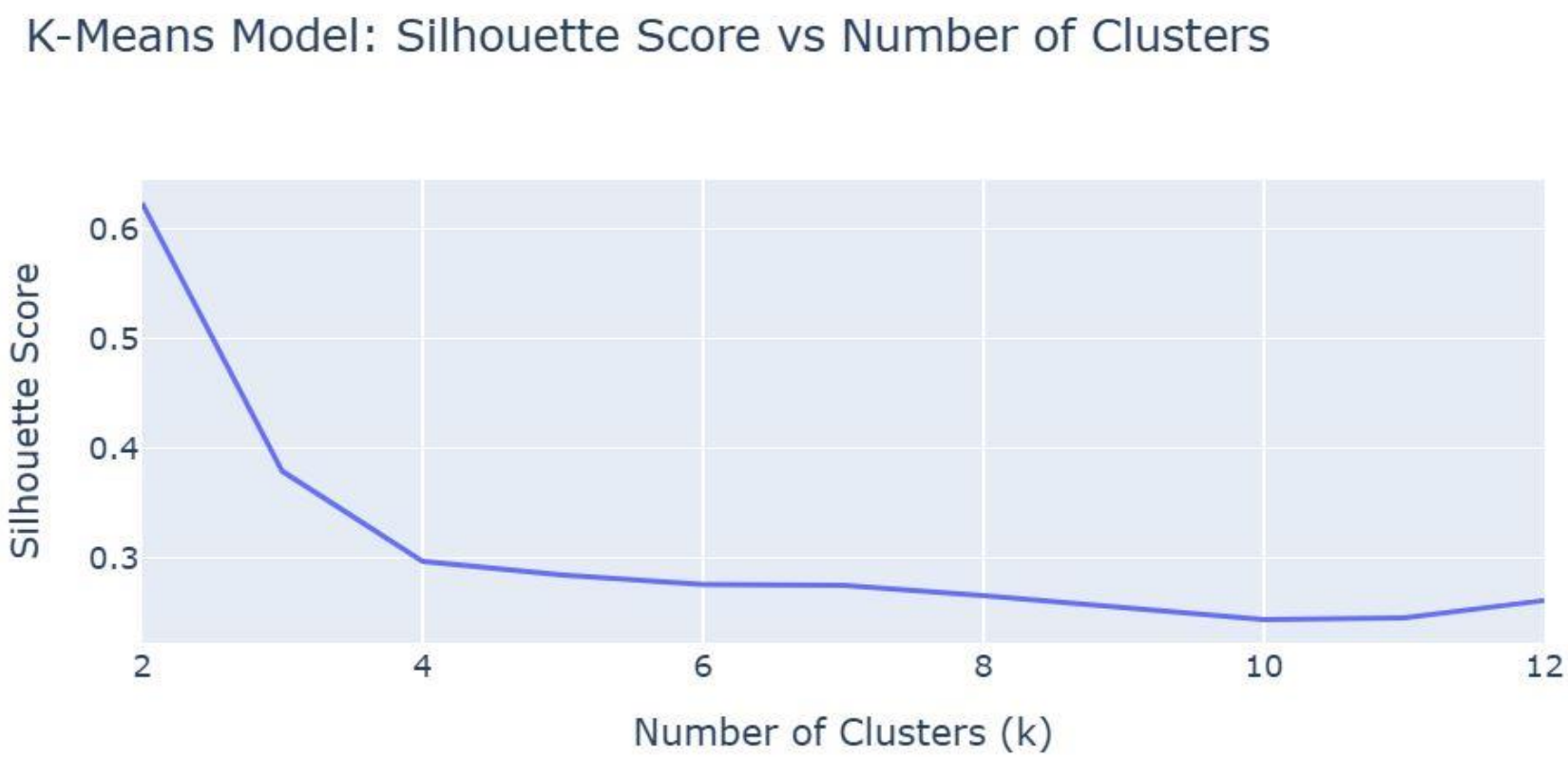
```
inertia_errors type: <class 'list'>
inertia_errors len: 11
Inertia: [18649.70966578131, 12021.204643397441, 9693.16799950018, 7784.120718723589, 6247.736151689862, 4929.558901300045, 4293.671370530936, 3753.4
347154578318, 3284.7296894290257, 3017.375839124871, 2765.2078787950422]

silhouette_scores type: <class 'list'>
silhouette_scores len: 11
Silhouette Scores: [0.623539354976519, 0.3788773987927405, 0.29662078166798217, 0.28392320737133936, 0.27520479041170476, 0.2745879326430291, 0.26515
186987062844, 0.25502695151241517, 0.24328348002753897, 0.2447081806046869, 0.2607075783524319]
```

RESULTS


K-Means Model: Inertia vs Number of Clusters
Inertia
15k
10k
5k
2
4
6
8
10
12
Number of Clusters (k)

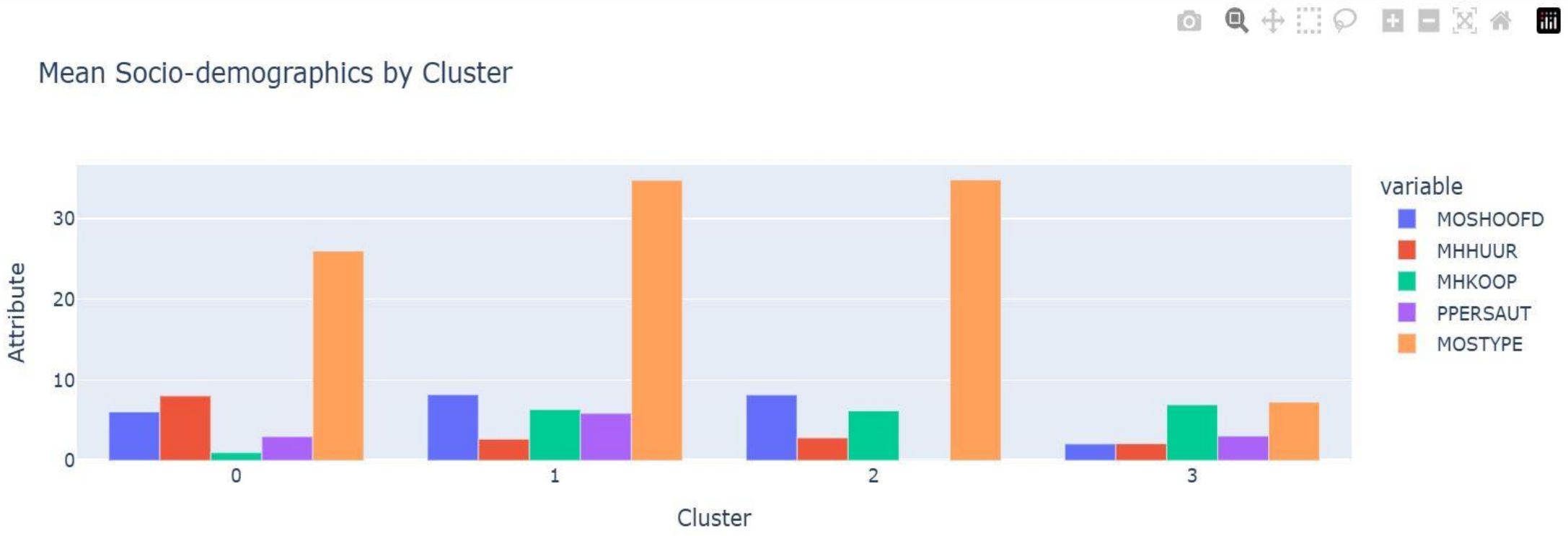
RESULTS


K-Means Model: Silhouette Score vs Number of Clusters
Silhouette Score
0.6
0.5
0.4
0.3
2
4
6
8
10
12
Number of Clusters (k)

COMMUNICATION

- In cluster 0, about 4% of the socioeconomic category are houseowners and 11% have made a contribution to car insurance policies.
- In cluster 1, about 18% of the socioeconomic category are houseowners and 16% have made a contribution to car insurance policies.
- In cluster 2, about 17% of the socioeconomic category are houseowners and none of them have acquired any car insurance policy.
- In cluster 3, 96% of the clients in the socioeconomic category are houseowners and 42% of these clients in the same cluster purchased car insurance policies.

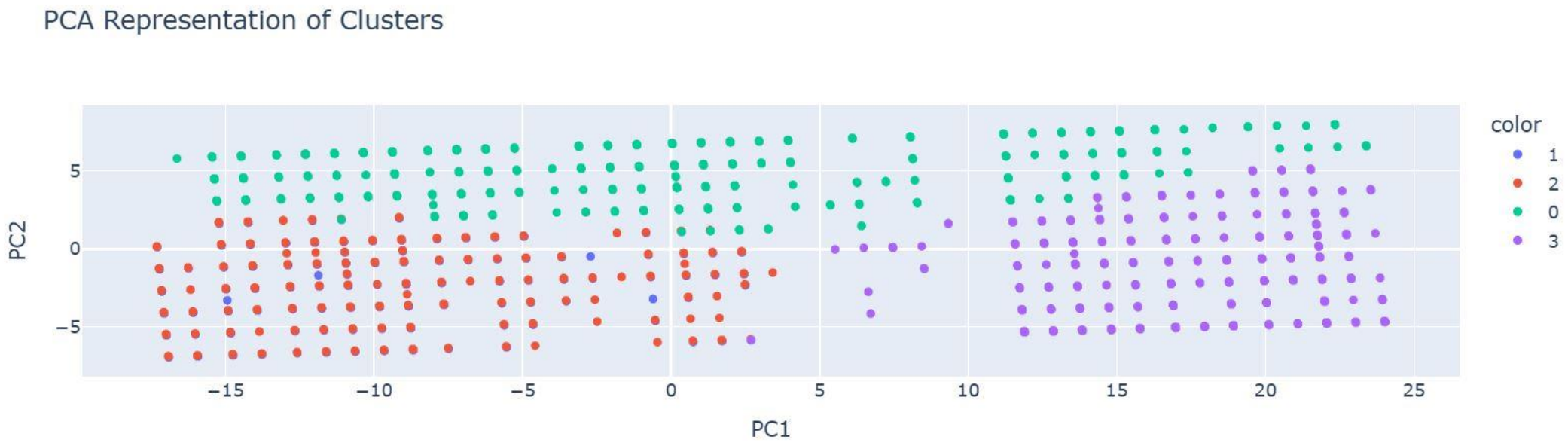
COMMUNICATION



COMMUNICATION

- MHHUUR and MOSHOOFD have the least representation in the socioeconomic class. Slightly a big number of homeowners are in the socioeconomic class (MOSTYPE) and a bigger proportion of clients with car insurance policies are highly expected to drive the cause to purchase caravan insurance since the policy covers against fire, theft(burglary insurance), accidental damage (third party liabilities) and enhances purchase of householders/houseowners insurance policies.

COMMUNICATION



Conclusion / Way forward

- Under MOSTYPE, middle class families and affluent young families can be properly sensitized to enhance caravan insurance purchasing dur to their viable financial capacity to take on these policies since it also financially secures them against uncertainties.
- Under MOSHOOFD, encourage successful hedonists to take up insurance policies that best suit their needs and lifestyle. Insurance brokers should reach out this target class and procure cover for them through prudent underwriters / insurers so as to obtain appropriate insurance cover on competitive premiums.
- Risk officers should technically advise on the risk strategies underwriters should undertake during risk assessment as this saves the company from paying out excess amount of claims which adversely affects their financial capacity.

Conclusion / Way forward

- The insurance company as a service provider should tailor appropriate cover for clients in average families i.e. advise them to take up education policies that benefit their children whereby premiums are paid on a monthly basis and a lumpsum with interest is paid out at maturity. This safeguards the financial capacity of these clients(parents) and insurance pays out in events of death and children cannot sustain themselves.
- Discounts can be offered to clients with installed CCTV surveillance system in their homes, installed car trackers in their motor vehicles and safety and fire equipment used incase of a fire, for example fire extinguishers, horse reels in apartment/rented homes, etc.