Stat515-005- Final Project - Team 2

Buying Caravan Insurance Policies

Preethal Reddy Yellareddygari

Nivedita J

Grace Manasseh

Agenda

- Introduction & Descriptive Statistics
- Can we predict if a new customer will buy a caravan insurance policy based on their profile?
- Can we group caravan insurance policy holders into clusters?
- Is there a relationship between buying a caravan insurance and having other insurance policy types?
- Wrap up

Dataset

- **Dependent Variable**: Having mobile home policy 0-1
- **Other Variables (85):**
  - Socio-Demographic Attributes (1-43)

CustomerSubtype, NbHouses, AvgSizeHousehold, AvgAge, CustomerMainType, RomanCatholic, Protestant, OtherReligion, NoReligion, Married, LivingTogether, OtherRelation, Singles, HouseholdNochildren, HouseholdWithchildren, HighLevelEducation, MediumLevelEducation, LowerLevelEducation, HighStatus, Entrepreneur, Farmer, MiddleManagement, SkilledLabourers, UnskilledLabourers, SocialclassA, SocialclassB1, SocialclassB2, SocialclassC, SocialclassD, RentedHouse, HomeOwners, 1car, 2cars, Nocar, NationalHealthService, PrivateHealthInsurance, Income<30, Income30- 45.000, Income45-75.000, Income75-122.000, Income>123.000, AverageIncome, PurchasingPowerClass,

  - Product Purchase Attributes (44-85)

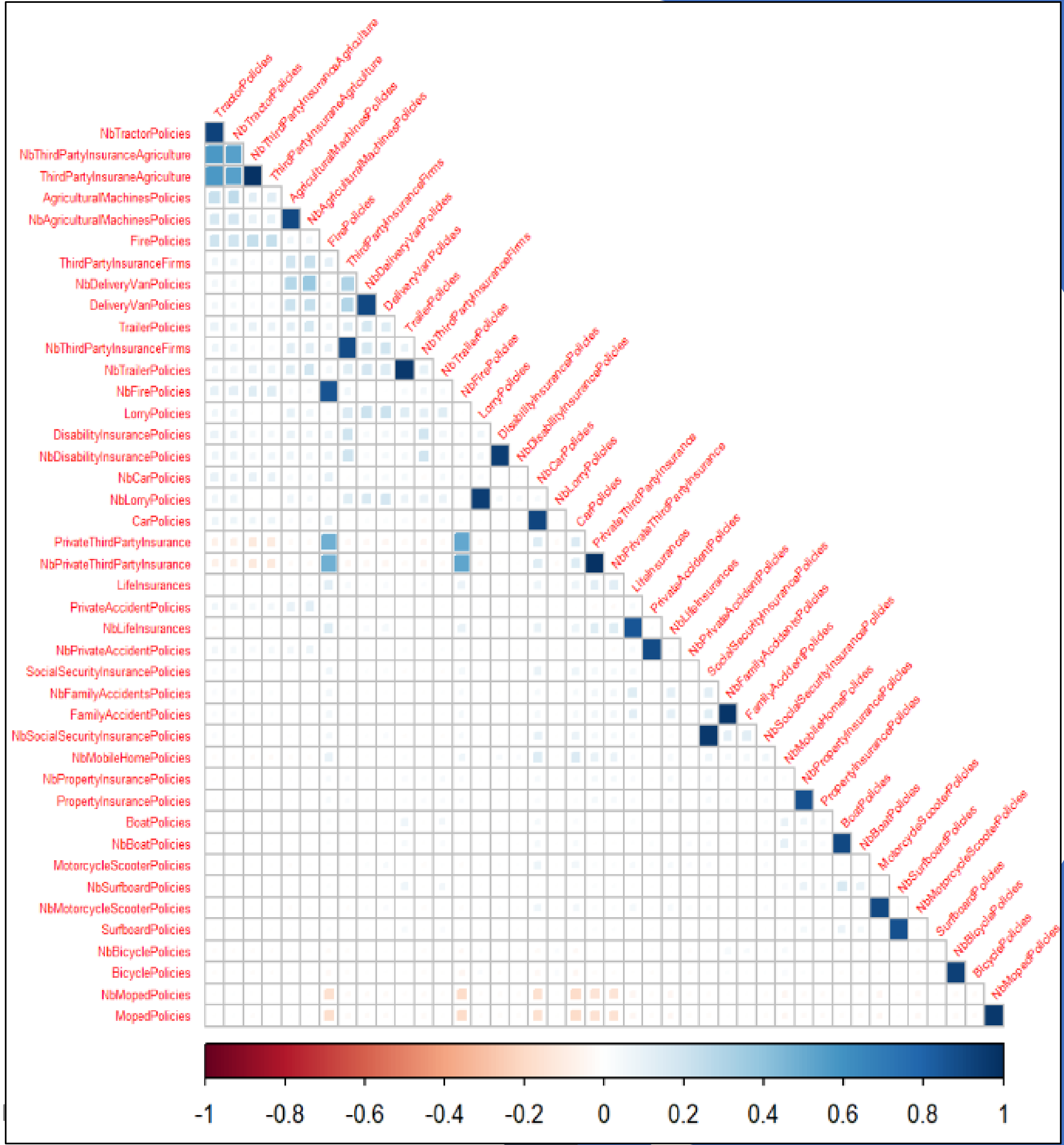
PrivateThirdPartyInsurance, ThirdPartyInsuranceFirms, ThirdPartyInsuraneAgriculture, CarPolicies, DeliveryVanPolicies, MotorcycleScooterPolicies, LorryPolicies, TrailerPolicies, TractorPolicies, AgriculturalMachinesPolicies, MopedPolicies, LifeInsurances, PrivateAccidentPolicies, FamilyAccidentPolicies, DisabilityInsurancePolicies, FirePolicies, SurfboardPolicies, BoatPolicies, BicyclePolicies, PropertyInsurancePolicies, SocialSecurityInsurancePolicies, NbPrivateThirdPartyInsurance, NbThirdPartyInsuranceFirms, NbThirdPartyInsuranceAgriculture, NbCarPolicies, NbDeliveryVanPolicies, NbMotorcycleScooterPolicies, NbLorryPolicies, NbTrailerPolicies, NbTractorPolicies, NbAgriculturalMachinesPolicies, NbMopedPolicies, NbLifeInsurances, NbPrivateAccidentPolicies, NbFamilyAccidentsPolicies, NbDisabilityInsurancePolicies, NbFirePolicies, NbSurfboardPolicies, NbBoatPolicies, NbBicyclePolicies, NbPropertyInsurancePolicies, NbSocialSecurityInsurancePolicies

Dataset

- **Training Dataset** = 5822 Observations having 348 classified with Target = 1 and 5474 with Target = 0. (~ 6.3% having 1)
- **Test Dataset** = 4000 Observations having 238 classified with Target = 1 and 3762 with Target = 0. (~6.3% having 1)

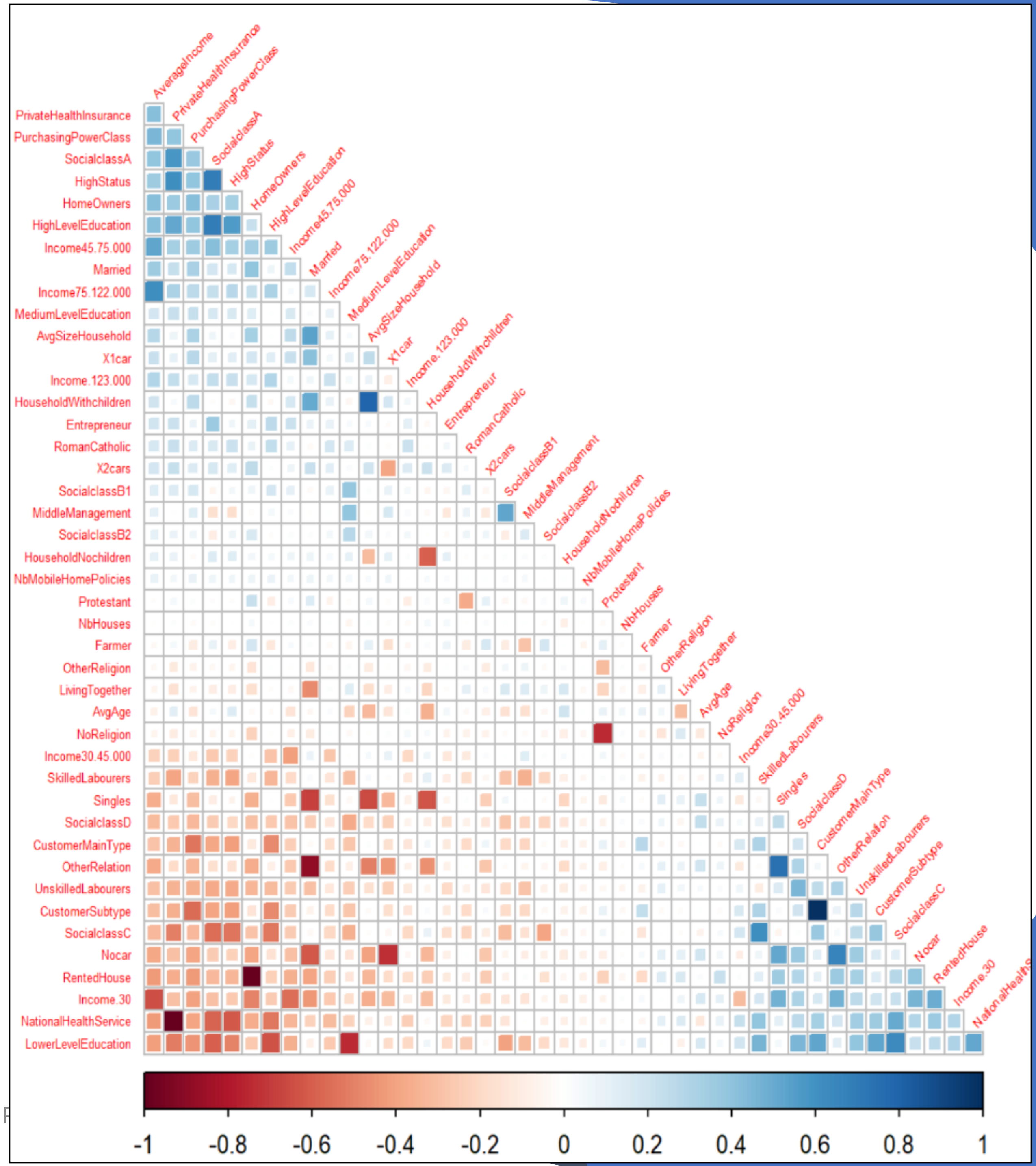
Correlations

- Set 1: Insurance product purchase attributes (44-85)
- A pattern in the very highly correlated variables: If the customer has insurance (yes/no) v.s. the number of policies of each insurance product
- During the analysis, we created 2 tests considering all v.s. keeping only 1 from each



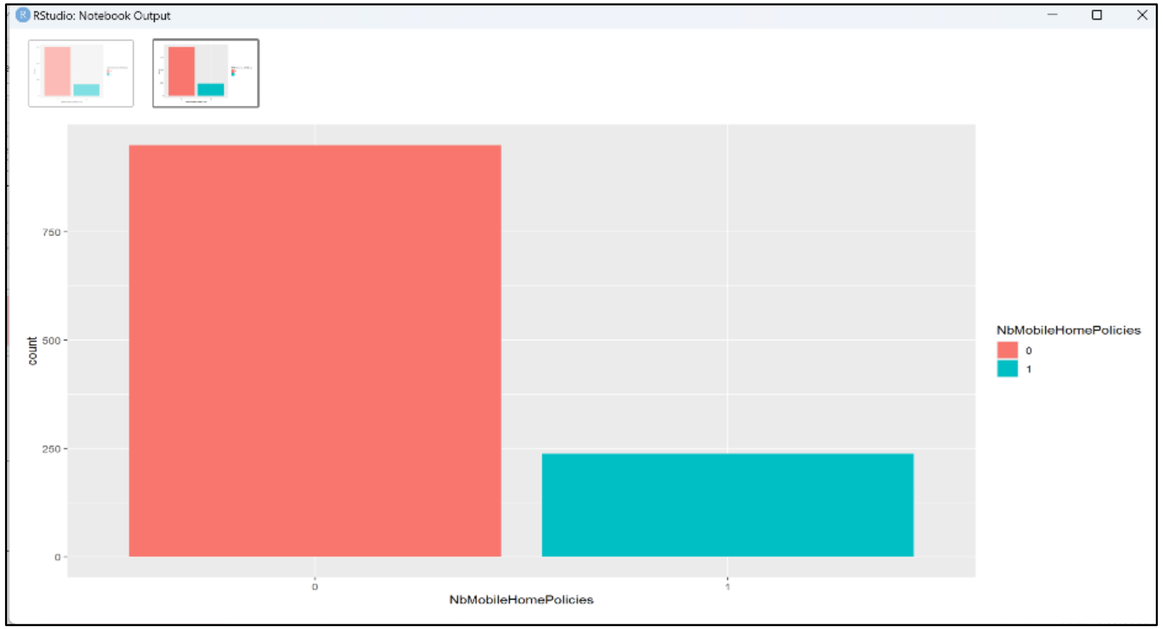
Correlations

- Set 2: Attributes (1:43) related to socio-demographics
- Highly Correlated:
  - ➢ OtherRelation v.s. Married
  - ➢ Singles v.s. Married
  - ➢ RentedHouse v.s HomeOwners
  - ➢ HouseholdNoChildren v.s. HouseholdWithChildren
  - ➢ CustomerMainType v.s. CustomerSubType



Imbalanced Data

- Treated Imbalanced data by removing a portion of the 0’s from both Training and Test datasets
- During the modeling, we compared results before and after this treatment



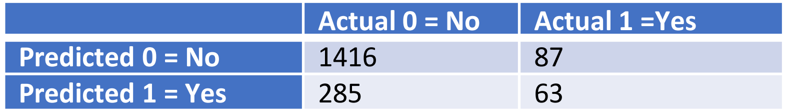
Question 1:

- Predicting Customers Purchasing Mobile Home Insurance
- 2 methods were used:
  - ❖Random Forest (3 models were compared)
    - ❖Full Dataset
    - ❖Having Imbalanced data treated
    - ❖Removing one from each of the highly correlated attributes
  - ❖Logistic Regression
    - ❖Using the dataset that had imbalanced data and highly correlated attributes treated

Question 1: Random Forest Models

- The best of the 3 was when using the dataset **having imbalanced data treated**

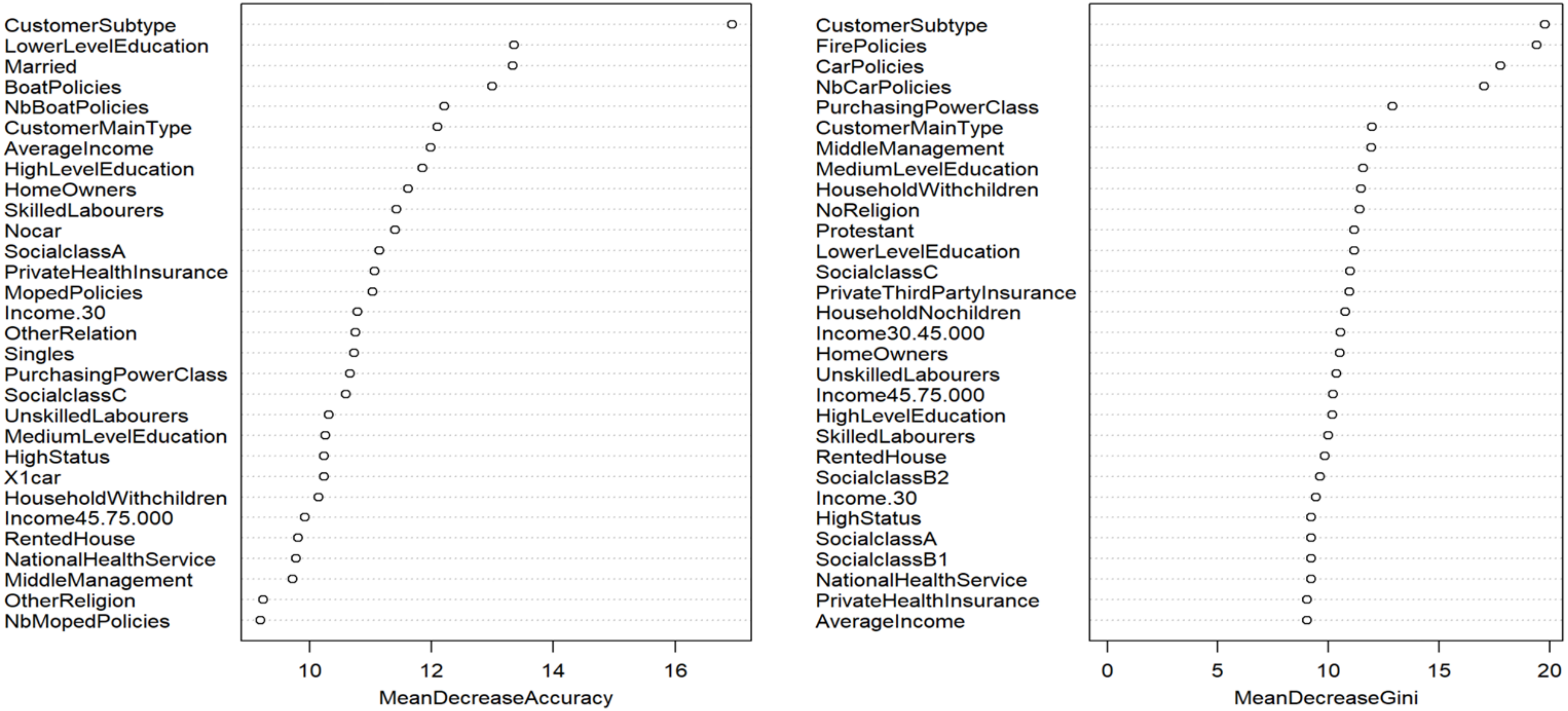
  Confusion Matrix =>



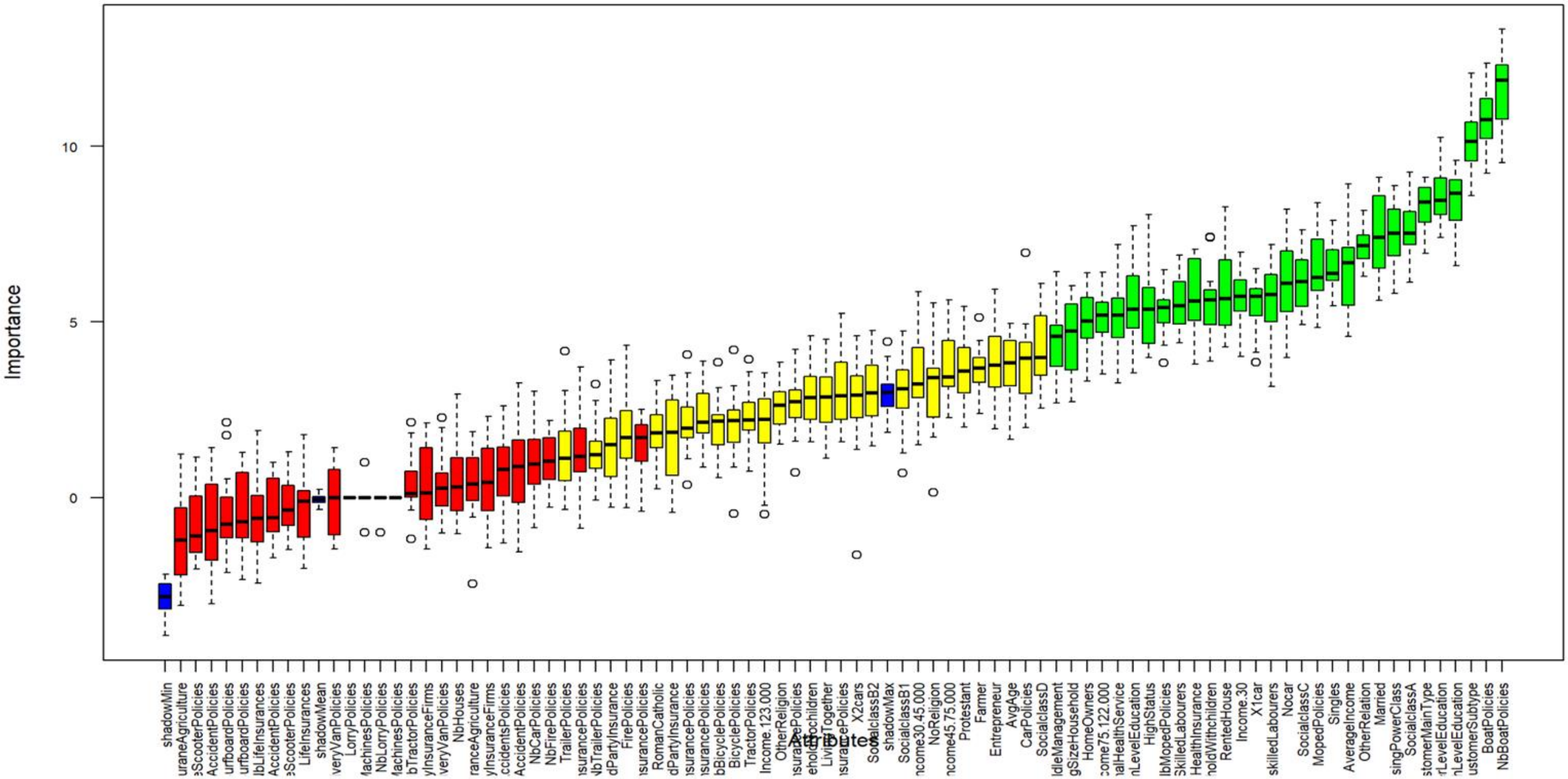
	Actual 0 = No	Actual 1 =Yes
Predicted 0 = No	1416	87
Predicted 1 = Yes	285	63

- We noticed the OOB (Out-of-bag) error rate increased to 20.1% v.s. the other 2 models (~6.8% and 5.5%).
- However, if we look at the True Positive matches, we notice a much higher % of true positives (42% instead of 11% in both other models)

Question 1: RF - Feature Importance



Question 1: RF - Feature Importance (Boruta)



Question 1: Logistic Regression

- Having Imbalanced and highly correlated attributes treated
- When using 50% as classifying probability, only 25 of the test observations are predicted to purchase insurance. The true positive rate is **~11% (25/238).**

Confusion Matrix =>

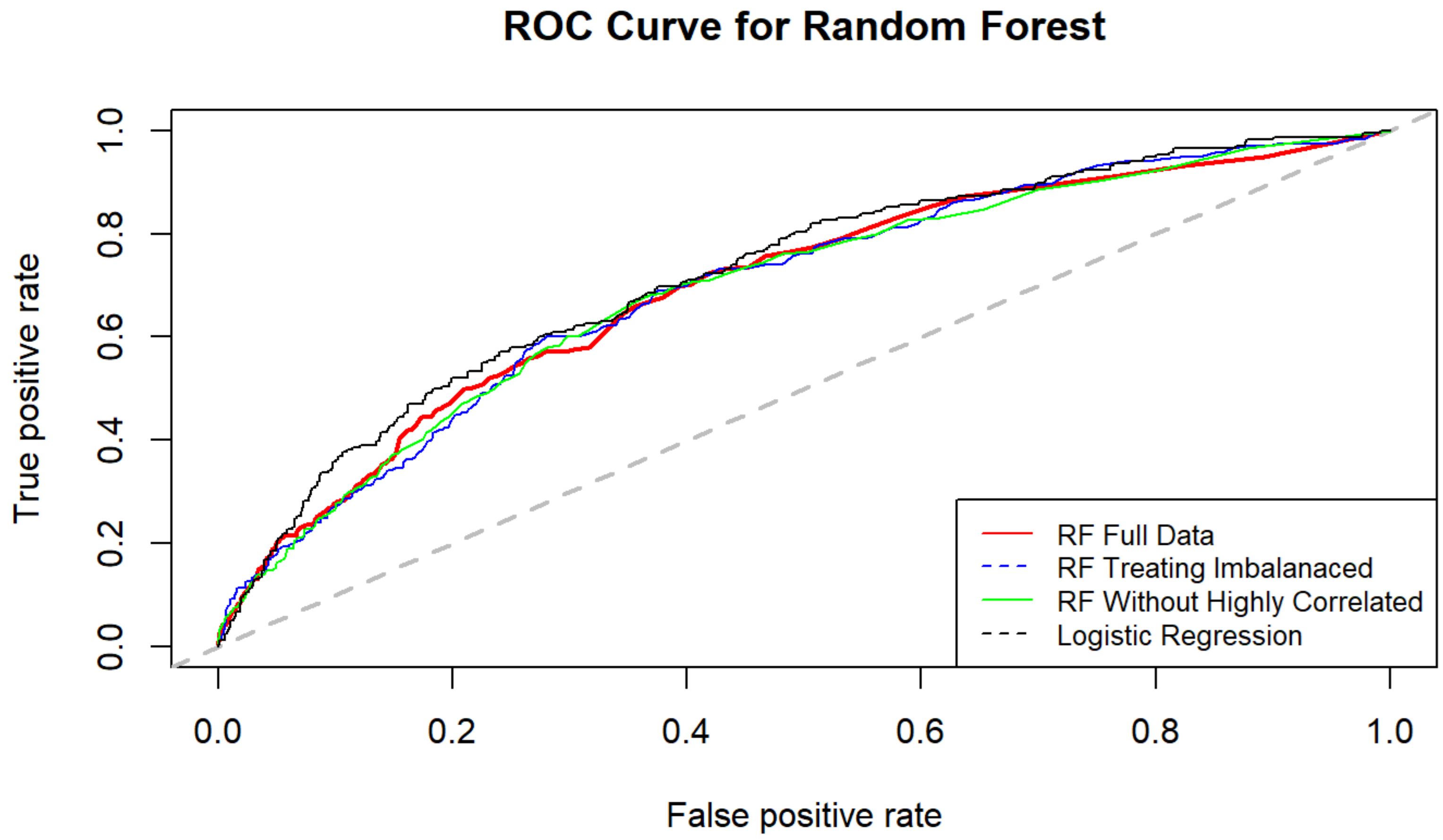
	Actual 0 = No	Actual 1 =Yes
Predicted 0 = No	928	213
Predicted 1 = Yes	22	25

- When using 25% as classifier, we get better results, we have 124 predicted to purchase insurance with a true positive rate of **~52% (124 / 238)**.

Confusion Matrix =>

	Actual 0 = No	Actual 1 =Yes
Predicted 0 = No	757	114
Predicted 1 = Yes	193	124

Question 1: ROC Curve Comparison

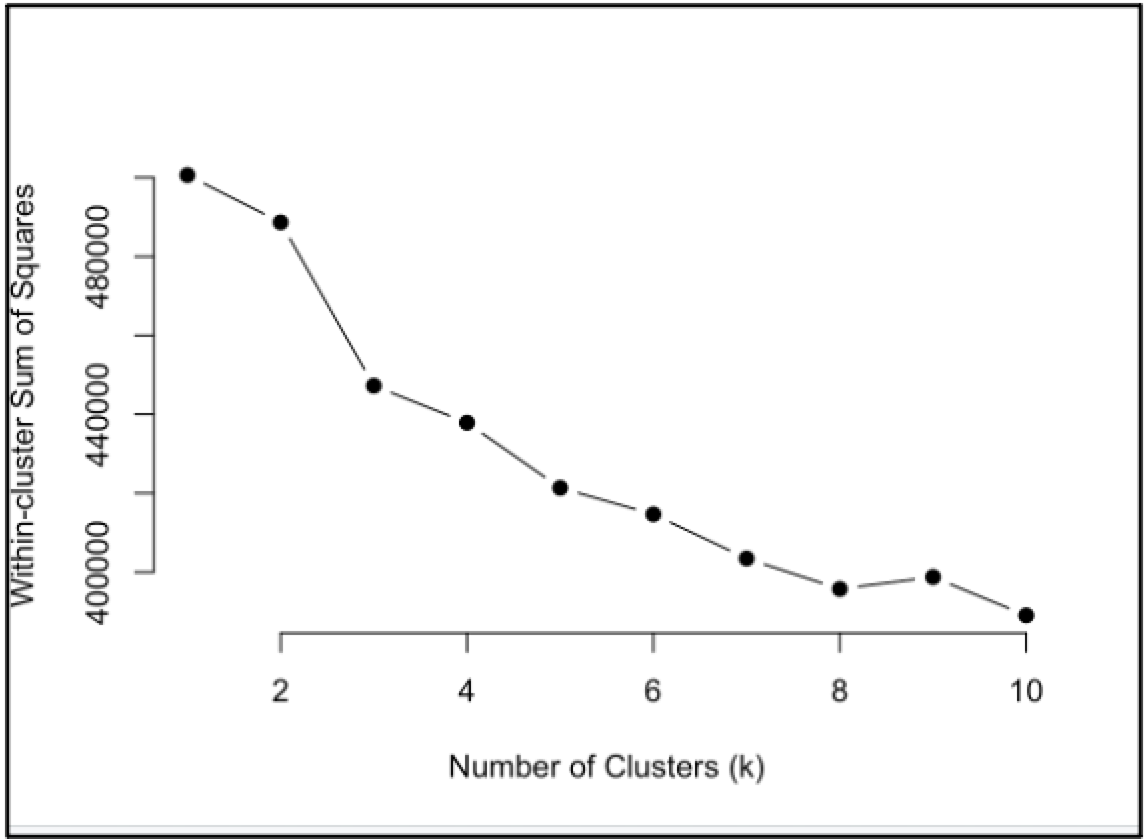


Question 2

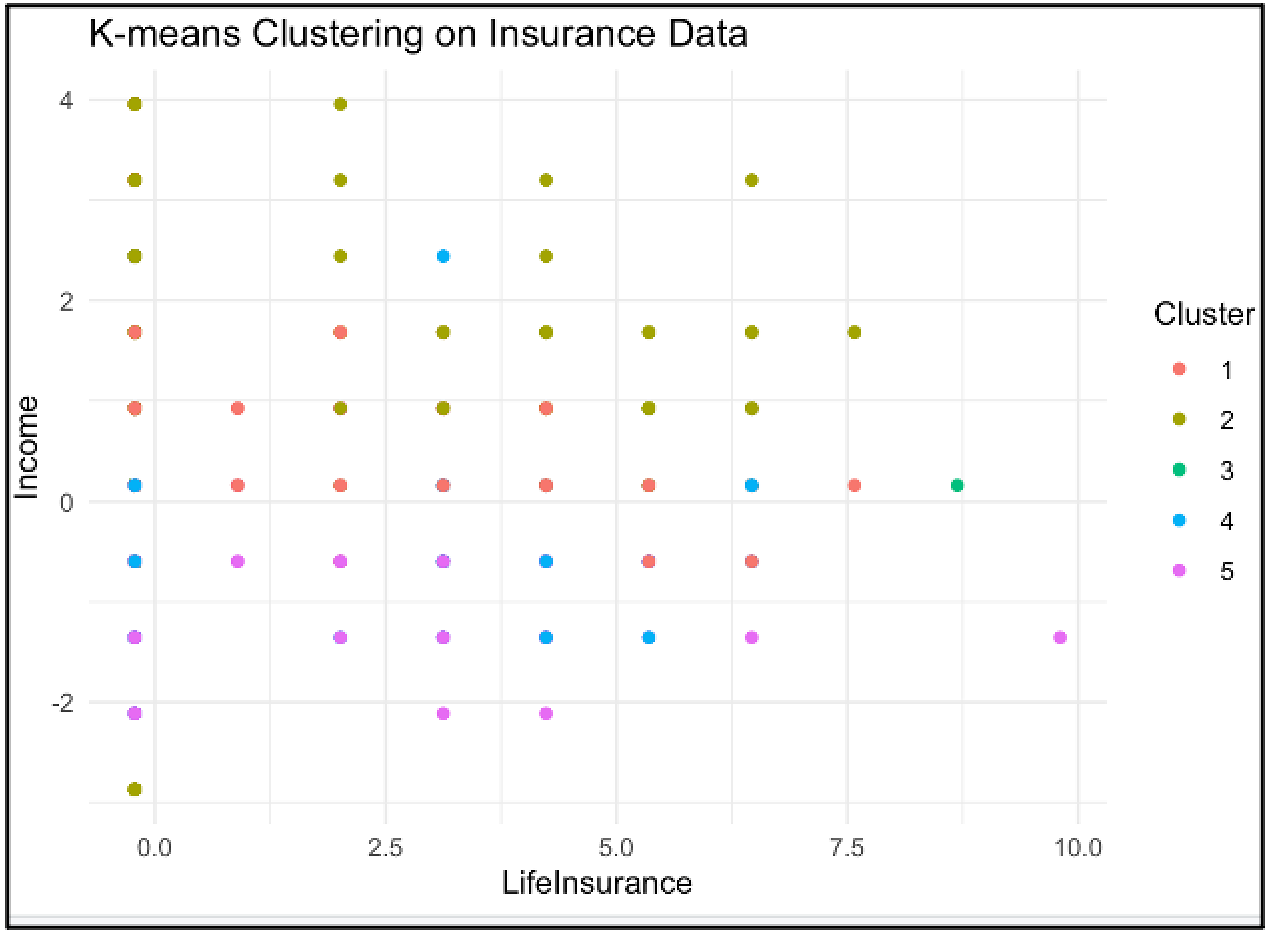
Grouping caravan insurance policy holders into clusters based on their income, social class, family size and other demographic characteristics?

K-Means Clustering

Kmeans clustering model on Insurance data to find customer segmentation. Trained a clustering model with different K ranging from [1,10] and plotted the elbow curve. This plot is an objective function(Sum of Squared errors) on the y-axis and the number of clusters k on the x-axis. Considering this, I have chosen an optimal k value of 5.



To understand different segmentation, I have plotted the Income vs Number of Life Insurance Policies different customers have brought with their income range.No clear separation between the clusters.



Potential reason why this could have happened?even after removing outliers from the data using Inter Quartile Range Method and normalising values.

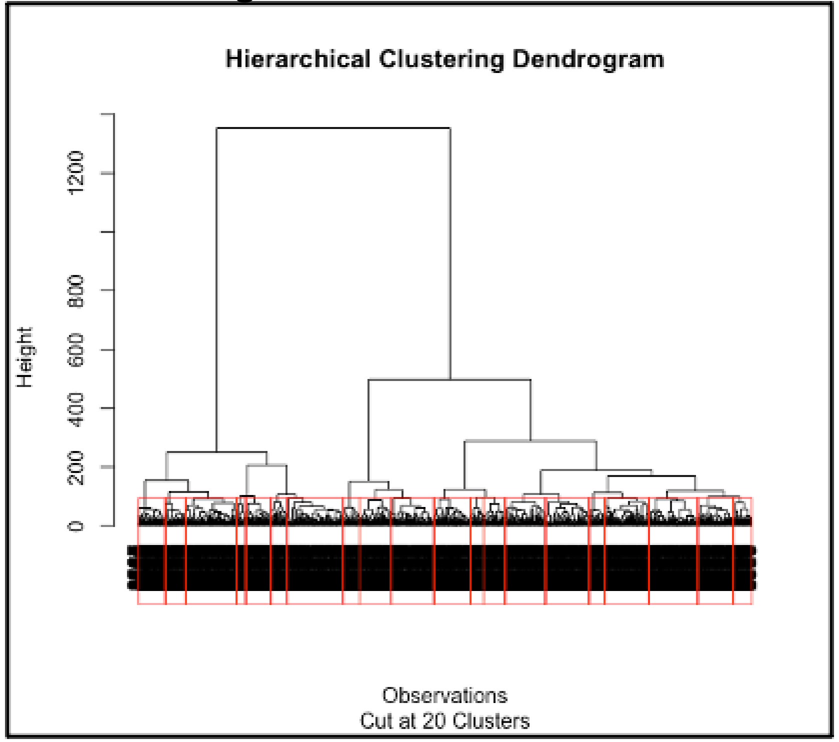
1. While k means using Euclidean distance to calculate similarities between different instances in the dataset, making the clusters look spherical and equal in size. While my data has distinct scales, as some values are categorical and some are numerical and Euclidean distance does not work well with this kind of data type, or the clusters probably have different shapes and densities.
2. The centroids in the k-means clustering algorithm are initialised by random selection, which could have produced less than ideal-outcomes.
3. This dataset could have overlapping clusters or lack a distinct separation, and k-means clustering would not work well for this data.

With this limitation in mind, we have considered using a hierarchical clustering algorithm to perform clustering.

Hierarchical Clustering

Ward's linkage, a method to minimise the rise in within-cluster variance, has been used as the distance measure. With this technique, we can capture intricate relationships in the data.

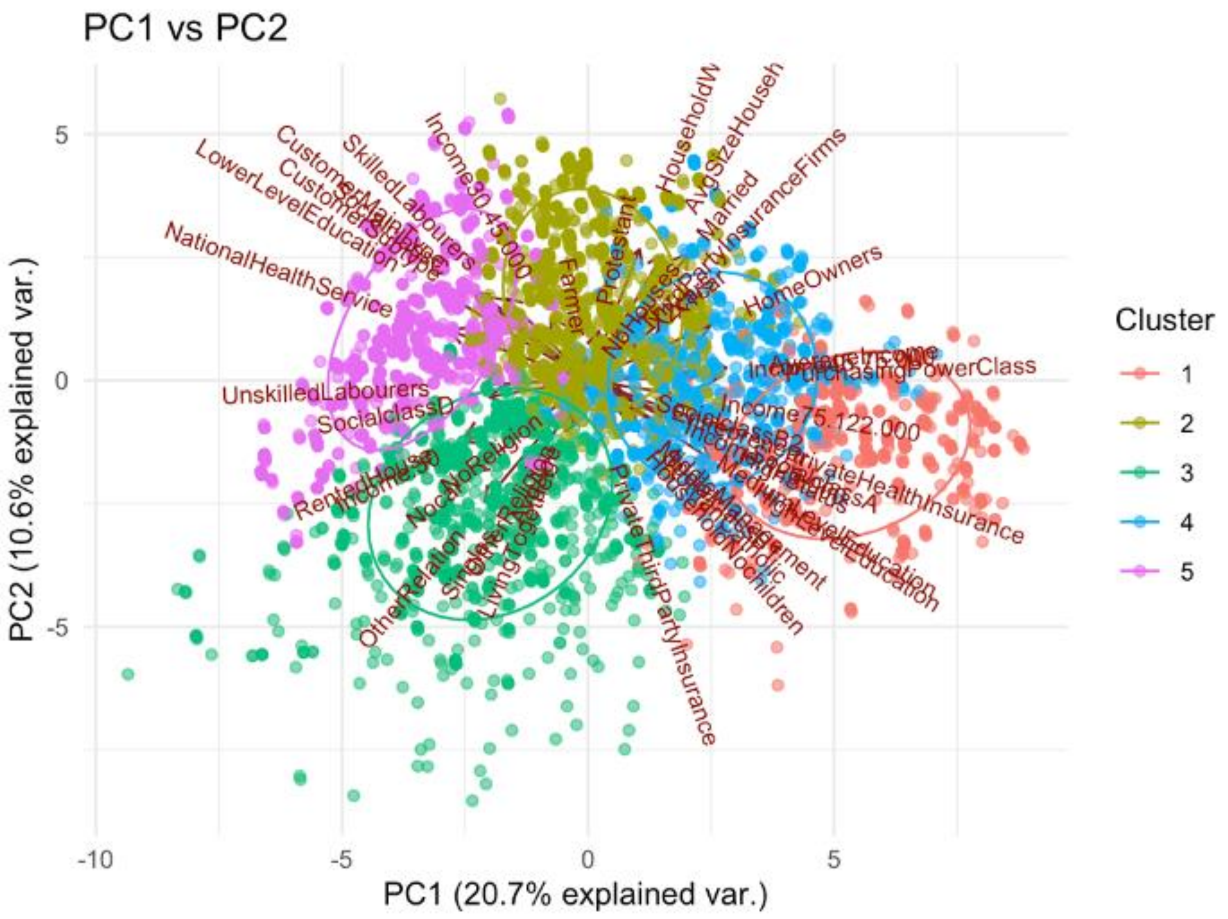
Below is a dendrogram with 20 clusters cut in observations,



The drawback of this model for our dataset is that we have many observations and variables; it is tough to interpret this large dendrogram visualisation, making it difficult to identify meaningful clusters and patterns.

PCA + K-Means Clustering

Tested on Dimensionality reduction techniques to have meaningful clusters. Used PCA to reduce the dimensionality of the data as it works by capturing linear combinations of variables.Here we can clearly see which Demographic Variables have influenced formation of a cluster.



Question 3

- Relationship between Caravan Insurance and Other Insurance Policies
- Methods used:
    - ❖Hypothesis test : Chi Squared test
        - ❖Data from columns 44-64
        - ❖Target Variable: Column 86
    - ❖Comparison of results using interactive plot

Question 3

Results of Chi Squared Test depending on p value

- **Significant associations:**
  - Buying caravan insurance is significantly associated with:
    - Private third-party insurance
    - Car policies
    - Motorcycle/scooter policies
    - Bicycle policies
    - Fire policies
    - Social security insurance policies

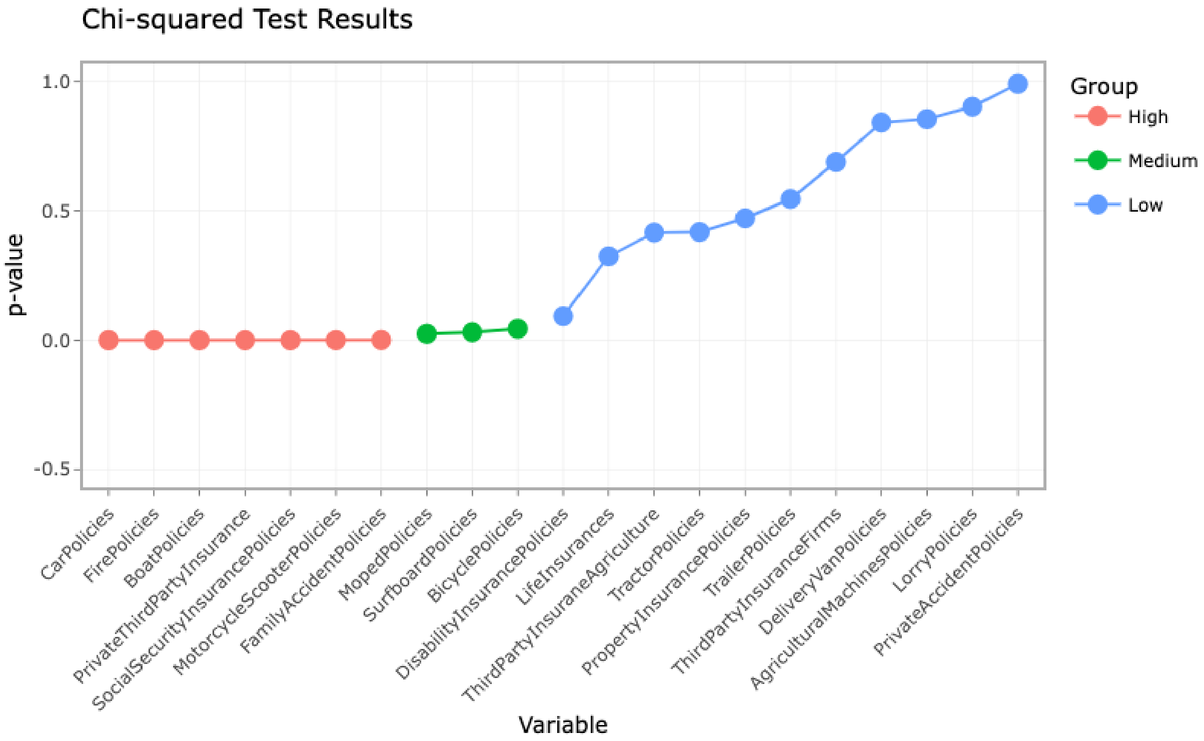
- **Non-significant associations:**

  No significant relationship between buying caravan insurance and:

  - Insurance from third-party firms
  - Agriculture-related third-party
  - Delivery van policies
  - Lorry policies
  - Trailer policies
  - Tractor policies
  - Agricultural machines policies
  - Life insurances
  - Family accident policies
  - Disability insurance policies
  - Fire policies
  - Surfboard policies
  - Boat policies
  - Property insurance policies

Question 3

- Visualization depending on Chi Squared Test Results
  - ❖Grouped the variables based on p-values into high, medium, and low significance categories
  - ❖Sorted the values accordingly
  - ❖Used ggplot and plotly packages for visualization
  - ❖Lines and points colored based on significance level groups



Conclusion / Future Work

- We notice similarity in results when doing the Chi-Squared in Q3 and what we got found from Random Forest - feature selection in Q1. Common influential attributes where noted: customer who bought caravan insurance most likely have a Car Policy, Boat Policy or Fire Policy.
- We can study in the future the predictions based on the clusters detected in Q2

Thank you