Capstone Project

Insurance Claim Analysis

Project Note

*By, Akash Mishra*

**Table of Content**

1. Introduction
   1. Problem Understanding
   2. Need of Project
   3. Understanding Business Opportunity
2. Data Report
   1. Understanding Data
   2. Data descriptive details
   3. Variable Information
3. Exploratory Data Analysis (EDA)
   1. Feature Engineering
   2. Data Summary
   3. Outlier and Missing Value Treatment
   4. Univariate Analysis
   5. Bivariate Analysis
4. Insights from EDA
   1. Check data balance
   2. Key Insights
5. Clustering
   1. K-Means Clustering and Understanding the Customer Profile with highest Rejection Ratio
6. Model Building and Performance Measurement
   1. Logistic Regression
   2. Naïve Bayes
   3. Random Forest
   4. Bagging
7. Business Recommendations
8. **Introduction**
   1. **Problem Understanding**

The given problem has financial impact on the organization. The objective of the organization through this analysis is to identify the fraudulent claims.

This project will support the Claims team to effectively manage the claims through predicting whether the submitted claim is to be accepted or rejected. Any claim rejected by the model can be scrutinized by the team in detail and hence, take a final call whether to accept the claim or reject the same.

* 1. **Need for the Project**

There is a huge need for such projects in the insurance sectors, for the below reasons,

* + 1. The project has financial implication on the organization
    2. Improve the TAT for claim processing
    3. Help the claims team to increase scrutiny focus on certain cases over other
    4. Improve the customer experience of claim submission
  1. **Understanding of Business Opportunity**

Fraudulent claims in Insurance sector leads to close to INR 40,000 Cr. loss which forms to close to 8.5% of the total revenue,

Any reduction in the count of fraudulent claims has huge financial implication on the firm. Even 1% reduction in claims leads to close to saving of INR 400 Cr. for the organization.

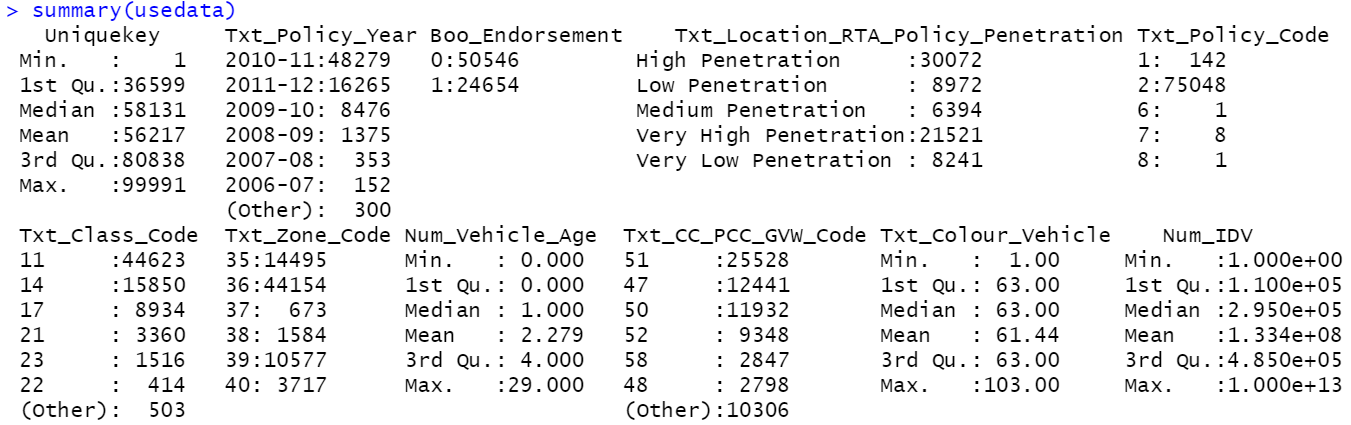
Moving ahead in the report to get a basic understanding of the data.

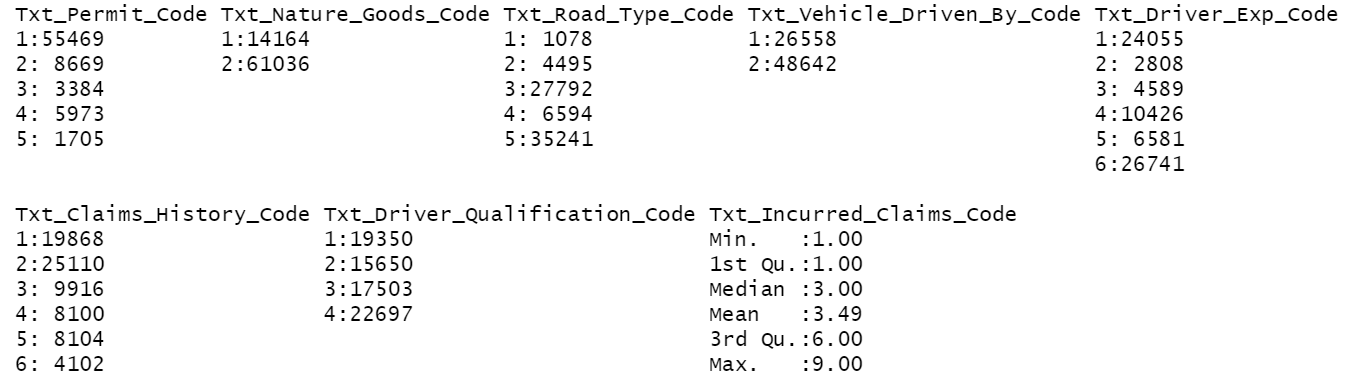
1. **Data Report**
   1. **Understanding Data**
      1. The dataset provided, has 75200 observations and 32 variables
      2. The data has been collected at a daily frequency
      3. The claim data has been collected for the period 1999-2012 i.e. 14 years
   2. **Data descriptive details**
      1. The data consists of 32 variables – 5 continuous, 3 character, 5 date and 19 discrete
      2. The output variable consists of 2 levels – Closed and Rejected
      3. 5 Continuous variables include – Uniquekey, Num\_Vehicle\_Age, Num\_IDV, DRV\_CLAIM\_AMT and Num\_Net\_OD\_Premium
      4. 3 Character variables include – Txt\_Location\_RTA, Txt\_Colour\_Vehicle and Txt\_Place\_Accident
      5. 5 Date variables include – Txt\_Policy\_Year, Txt\_Claim\_Year, Date\_Accident\_Loss, Date\_Claim\_intimation and Date\_Disbursement
      6. 20 Discrete variables include – Boo\_Endorsement, Txt\_Policy\_Code, Txt\_Class\_Code, Txt\_Zone\_Code, Txt\_CC\_PCC\_GVW\_Code, Txt\_Permit\_Code, Txt\_Nature\_Goods\_Code, Txt\_Road\_Type\_Code, Txt\_Vehicle\_Driven\_By\_Code, Txt\_Driver\_Exp\_Code, Txt\_Claims\_History\_Code, Txt\_Driver\_Qualification\_Code, Txt\_Incurred\_Claims\_Code, Boo\_TPPD\_Statutory\_Cover\_only, Txt\_TAC\_NOL\_Code, Boo\_OD\_Total\_Loss, DRV\_CLAIM\_STATUS, Boo\_AntiTheft and Boo\_NCB
   3. **Variable Information**

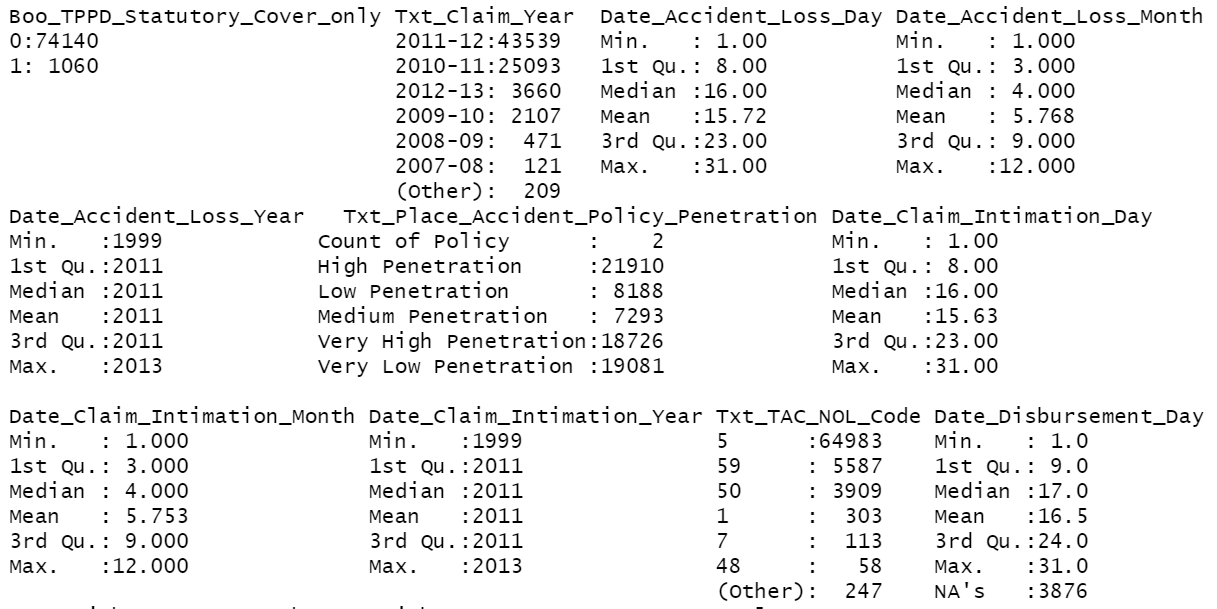
* Txt\_Policy\_Year: Range is from 1989-99 to 2012-13. No missing values
* Boo\_Endorsement: Levels – 0 and 1. No missing values
* Txt\_Location\_RTA: 5501 unique values with no missing value
* Txt\_Policy\_Code: Levels (5) – 1,2,6,7,8. No missing values
* Txt\_Class\_Code: Levels (10) – 11,13,14,17,18,19,20,21,22,23. No missing values
* Txt\_Zone\_Code: Levels (6) – 35,36,37,38,39,40. No missing values
* Num\_Vehicle\_Age: Range is from 0 to 29. No missing values
* Txt\_CC\_PCC\_GVW\_Code: Levels (19) – 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64. No missing values
* Txt\_Colour\_Vehicle: 103 unique values with no missing value
* Num\_IDV: Range is from 1 to 1e+13. No missing values
* Txt\_Permit\_Code: Levels (5) – 1,2,3,4,5. No missing values
* Txt\_Nature\_Goods\_Code: Levels – 1 and 2. No missing values
* Txt\_Road\_Type\_Code: Levels (5) – 1,2,3,4,5. No missing values
* Txt\_Vehicle\_Driven\_By\_Code: Levels – 1 and 2. No missing values
* Txt\_Driver\_Exp\_Code: Levels (6) – 1,2,3,4,5,6. No missing values
* Txt\_Claims\_History\_Code: Levels (6) – 1,2,3,4,5,6. No missing values
* Txt\_Driver\_Qualification\_Code: Levels (4) – 1,2,3,4. No missing values
* Txt\_Incurred\_Claims\_Code: Levels (9) – 1,2,3,4,5,6,7,8,9. No missing values
* Boo\_TPPD\_Statutory\_Cover\_only: Levels – 0 and 1. No missing values
* Txt\_Claim\_Year: Range is from 1999-00 to 2012-13. No missing values
* Date\_Accident\_Loss: Range is from 5th Apr’1999 to 2nd Dec’2013. No missing values
* Txt\_Place\_Accident: 15346 unique values. No missing values
* Date\_Claim\_Intimation: Range is from 17th May’1999 to 7th Dec’2013. No missing values
* Txt\_TAC\_NOL\_Code: Levels (27) – 1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13, 14, 32, 34, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 59. No missing values
* Date\_Disbursement: Range is from 26th Jun’1999 to 6th Jan’2014. There are blank values as there are no date of disbursement for Rejected cases
* Boo\_OD\_Total\_Loss: Levels – 0 and 1. No missing values
* DRV\_CLAIM\_AMT: Range is from 0 to 821600. No missing values
* DRV\_CLAIM\_STATUS: Levels – Closed and Rejected. No missing values
* Boo\_AntiTheft: Levels – 0 and 1. No missing values
* Boo\_NCB: Levels – 0 and 1. No missing values
* Num\_Net\_OD\_Premium: Range is from 13 to 96795. No missing values

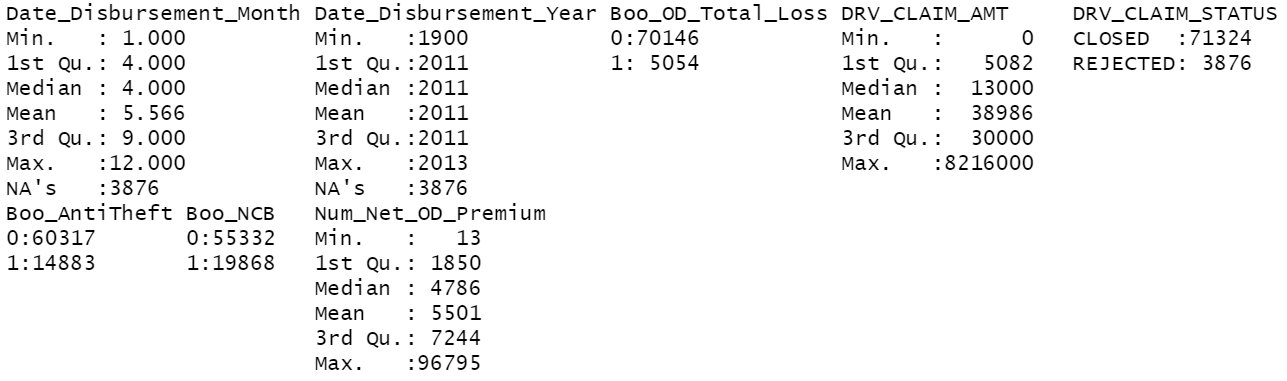
1. **Exploratory Data Analysis**
   1. **Feature Engineering**
      1. Date\_Accident\_Loss variable has been re-defined into 3 variables - *Date\_Accident\_Loss\_Day*, *Date\_Accident\_Loss\_Month*, *Date\_Accident\_Loss\_Year*
      2. Date\_Claim\_Intimation variable has been re-defined into 3 variables - *Date\_Claim\_Intimation\_Day*, *Date\_Claim\_Intimation\_Month*, *Date\_Claim\_Intimation\_Year*
      3. Date\_Disbursement variable has been re-defined in to 3 variables - *Date\_Disbursement\_Day*, *Date\_Disbursement\_Month*, *Date\_Disbursement\_Year*
      4. Additionally, we have created 2 variable *Location\_RTA\_Policy\_Penetration* and *Place\_Accident\_Policy\_Penetration*, which is used as a flag to identify if the RTA registration place and place of accident are same or different
   2. **Data Summary**

Pre-outlier Treatment









* 1. **Outlier and Missing Value Treatment**

From the above data we observe that there are outliers in 3 variables,

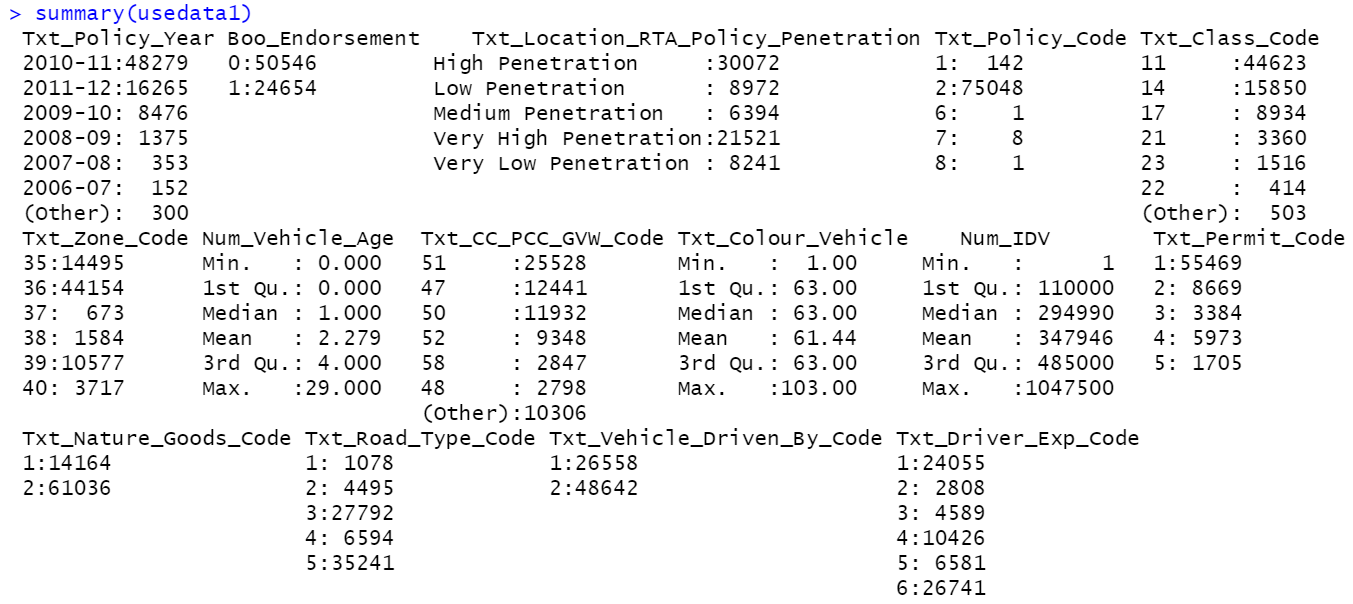
* Num\_IDV: Outlier at the higher end. No outlier on the lower end
* DRV\_CLAIM\_AMT: No outlier on the lower end
* Num\_Net\_OD\_Premium: No outlier on the lower end

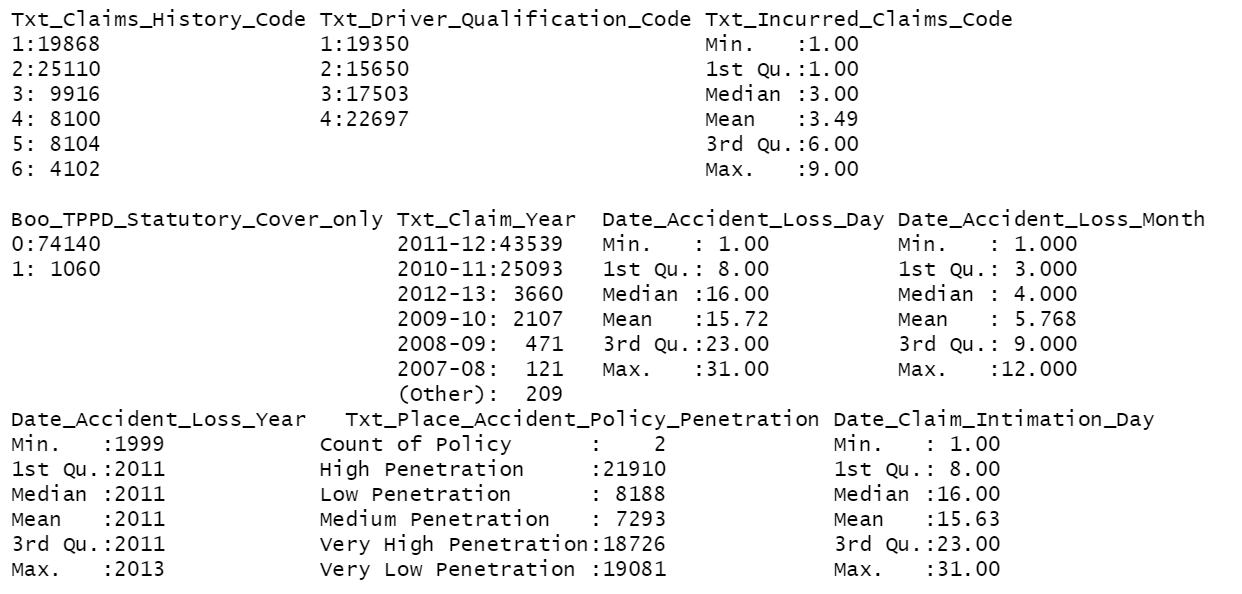
The outlier values have been treated by replacing them with the value of Q3+(1.5\*(Q3-Q1)) of each respective variable.

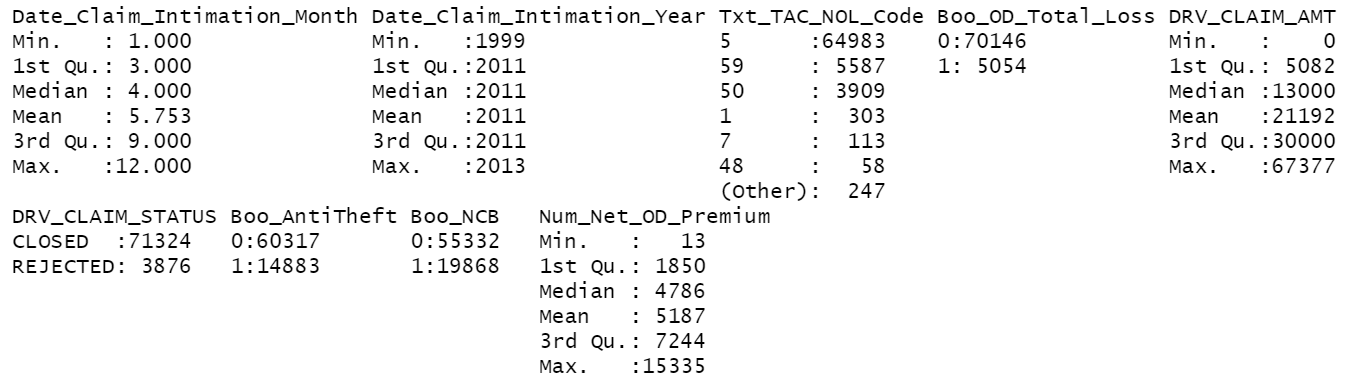
We also, find missing values in variables Date\_Disbursement\_Day, Date\_Disbursement\_Month and Date\_Disbursement\_Year. The count of missing values in each is 3876. However, we will not treat them because, as per business logic, the date disbursement date would be blank for all the rejected cases, which is the case in this scenario.

Hence, we will go ahead with further steps with these missing values and cautiously approach the modelling for claim identification keeping in mind these data points.

Summary Post-outlier Treatment



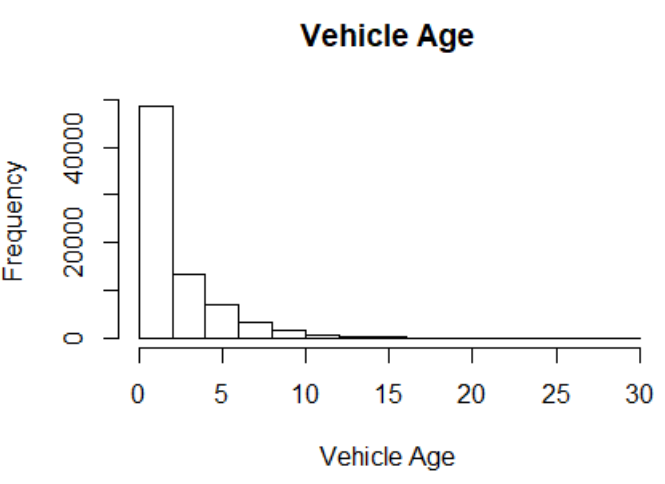
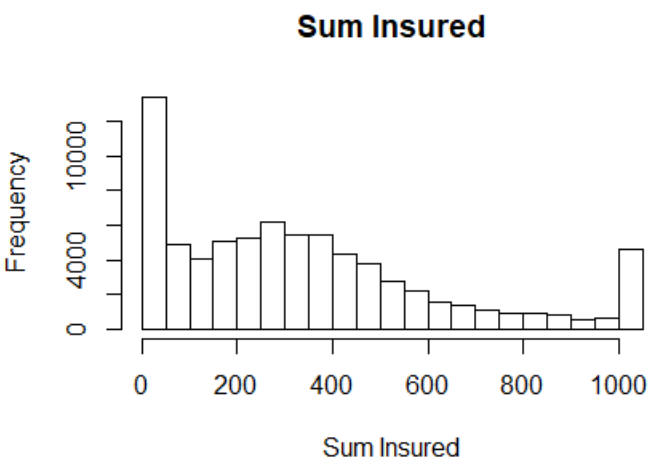




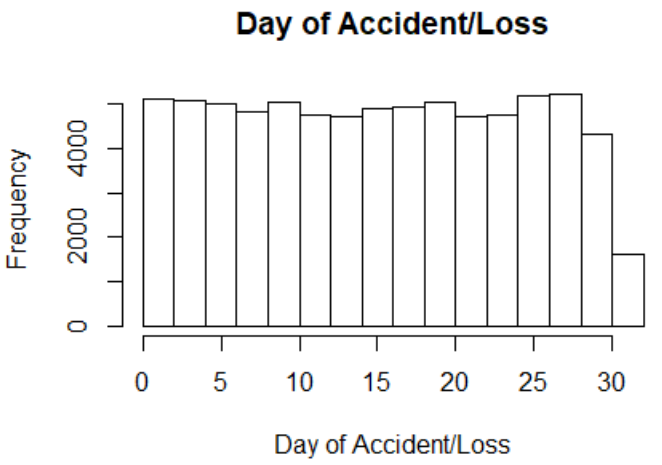
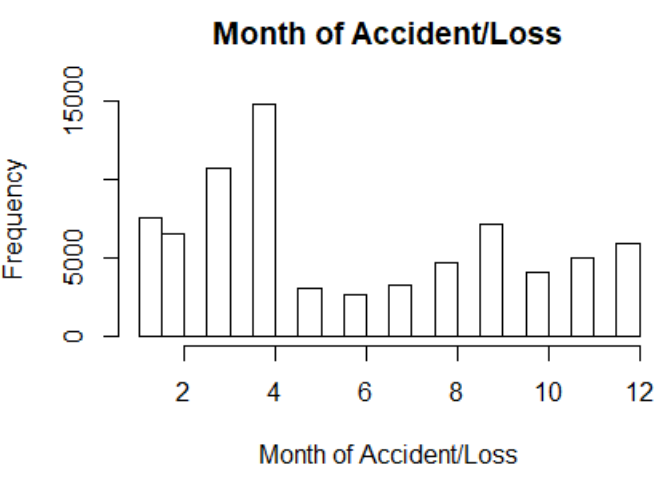
* 1. **Univariate Analysis**

Continuous Variables

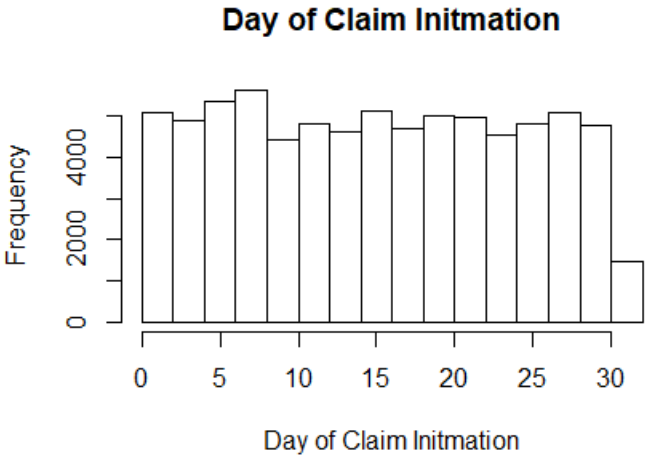
**Num\_Vehicle\_Age – Right Skewed** **Num\_IDV – Semi Right Skewed**

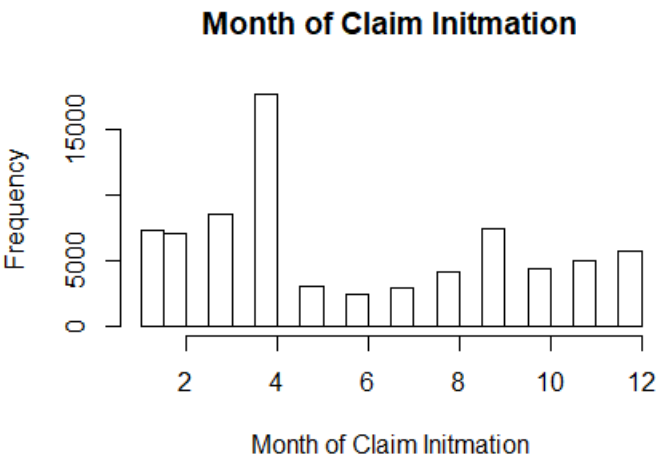
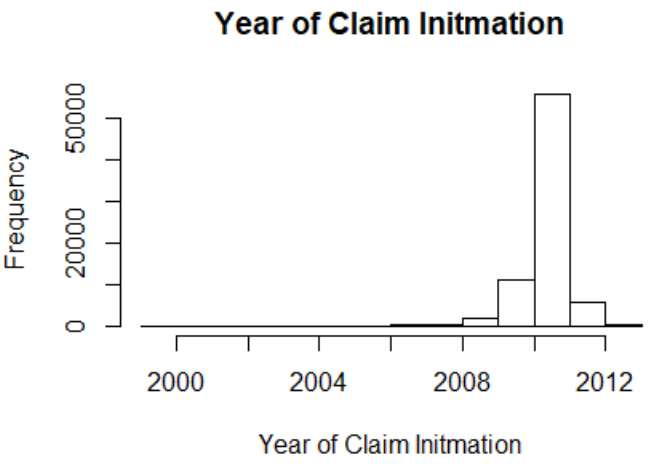
**Date\_Accident\_Loss\_Day – Normal** **Date\_Accident\_Loss\_Month–Semi Right Skew**

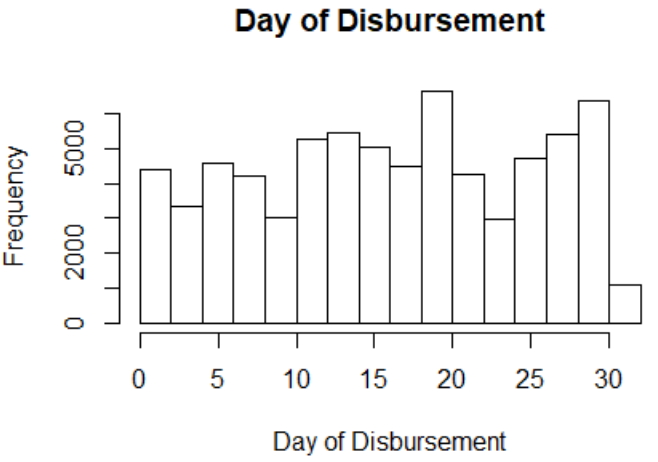
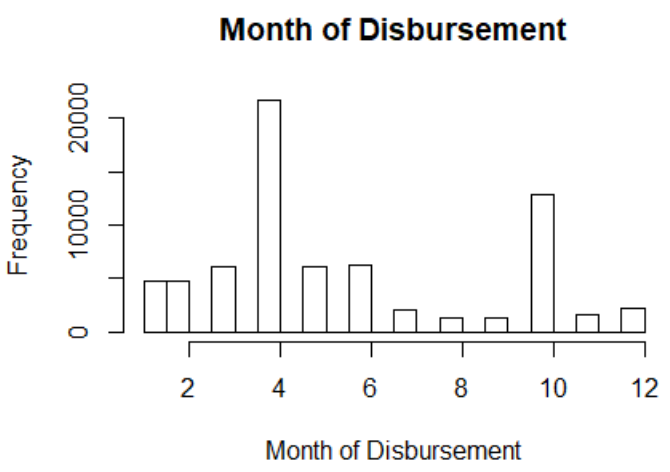
**Date\_Accident\_Loss\_Year – Left Skewed Date\_Claim\_Intimation\_Day – Normal**

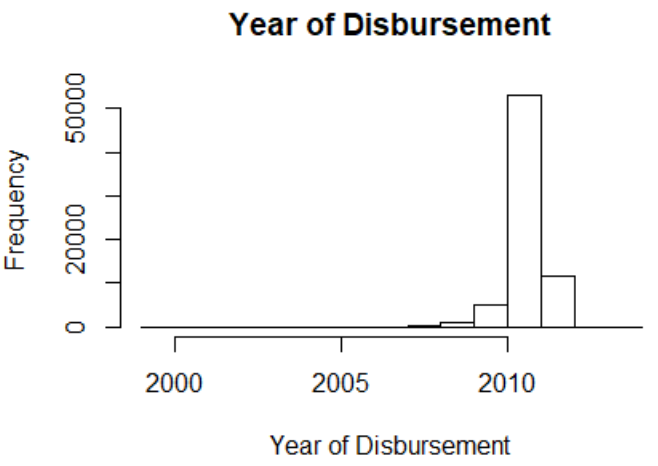
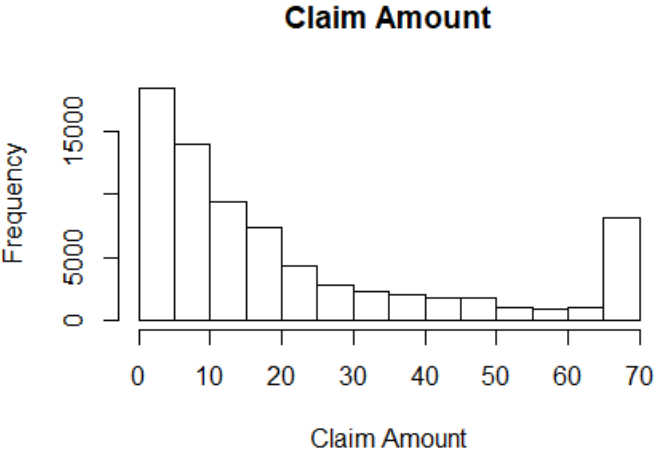
**Date\_Claim\_Intimation\_Month–Semi Rt. Skew Date\_Claim\_Intimation\_Year–Left Skew**

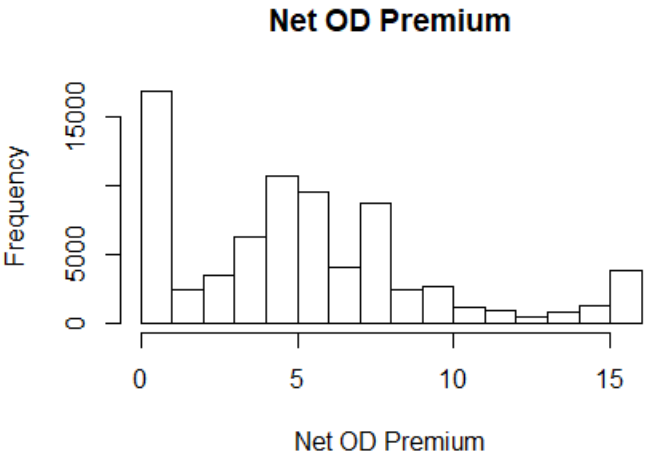
**Date\_Disbursement\_Day – Normal Date\_Disbursement\_Day – Semi Right Skew**

**Date\_Disbursement\_Year – Left Skewed DRV\_CLAIM\_AMT – Right Skewed**

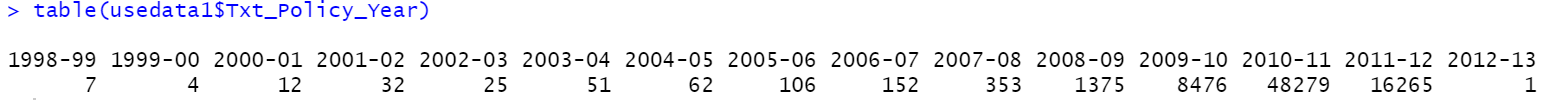
 

**Num\_Net\_OD\_Premium – Almost Normal**

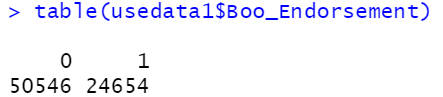


Discrete Variable

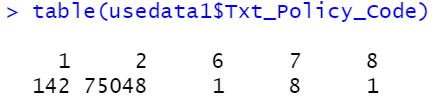
**Txt\_Policy\_Year: Maximum no. of claims were raised in policies purchased in 2010-11**



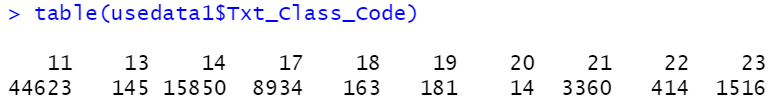
**Boo\_Endorsement: 67% of claims are not endorsed**



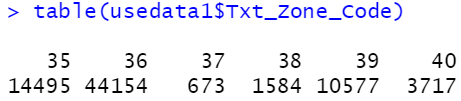
**Txt\_Policy\_Code: 99.8% of claims are from Package Policy**



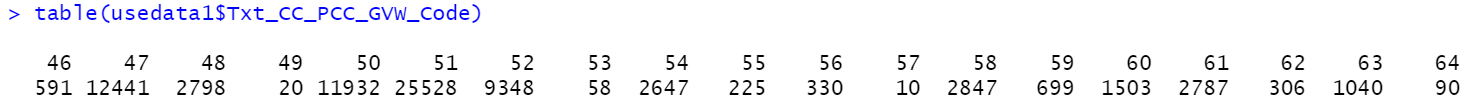
**Txt\_Class\_Code: 92% of claims are registered by Private Cars (59%), Two Wheelers (21%) & Goods Carrying vehicles other than three wheelers – Public (12%)**



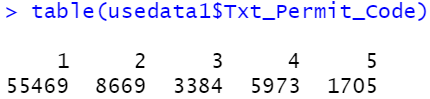
**Txt\_Zone\_Code: 92% of the claims are registered Zone A - Sections 2 ,3,4.C.1 and 4.C.4 (19%), Zone B - Sections 2 ,3,4.C.1 and 4.C.4 (59%) and Zone C - Sections 4.A, 4.C.2,4.C.3 and 4.D (14%)**



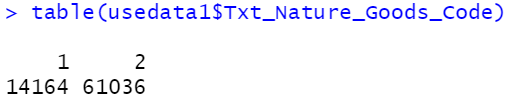
**Txt\_CC\_PCC\_GVW\_Code: 66% of the claims are made from people driving 75 to 150 cc (applicable for two wheelers) (17%), CC Not exceeding 1000CC (applicable for Private cars &Taxis) (16%) & CC Between 1000CC and 1500 CC (applicable for Private cars & Taxis) (34%)**



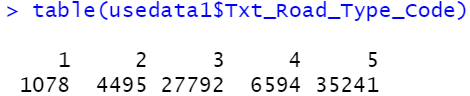
**Txt\_Permit\_Code: 74% of the claims are made by people with local permit**



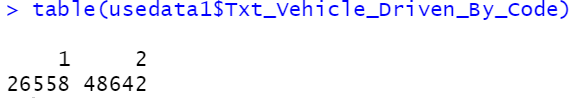
**Txt\_Nature\_Goods\_Code: 81% of the claims are made by non-hazardous goods carrier**



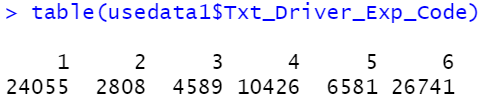
**Txt\_Road\_Type\_Code: 84% of the claims are made by vehicle running on City/Town Roads and Others**



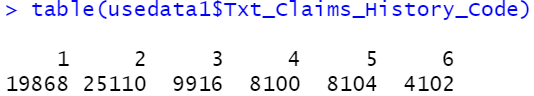
**Txt\_Vehicle\_Driven\_By\_Code: 65% of the claims are made by non-Owners**



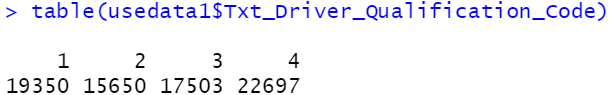
**Txt\_Driver\_Exp\_Code: 68% of the claims are made by rider with “<1 year” or “15 years and above” experience category**



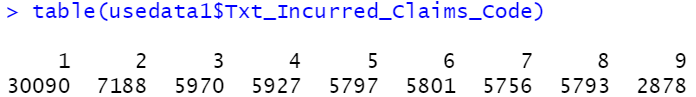
**Txt\_Claims\_History\_Code: 60% of the claims are made with no prior claims or only 1 claim history**



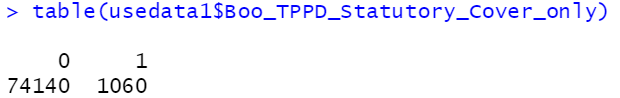
**Txt\_Driver\_Qualification\_Code: There is no clear impact of qualification on claims made**



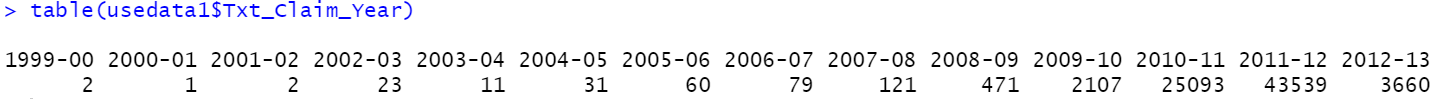
**Txt\_Incurred\_Claims\_Code: 40% of the claims made had only 1 previous claim history before the current policy**



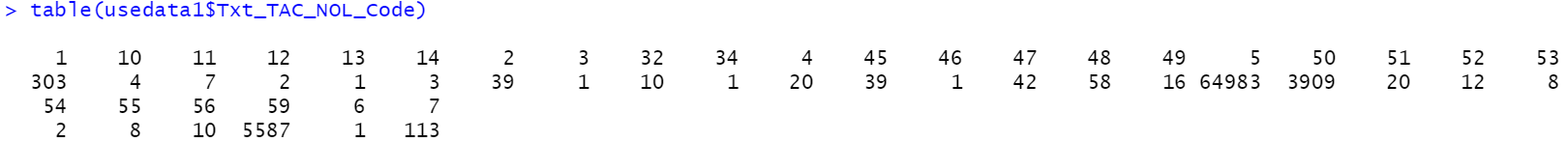
**Boo\_TPPD\_Statutory\_Cover\_only: 98.6% of people who have made claim had opted for Wider Cover**



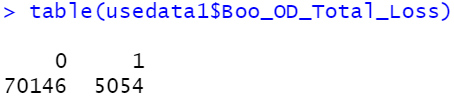
**Txt\_Claim\_Year: Maximum number of claims were made in the year 2011-12**



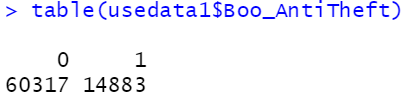
**Txt\_TAC\_NOL\_Code: 86% of the claims made were due to Accident by External Means**



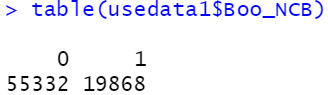
**Boo\_OD\_Total\_Loss: 93% of claim cases did not have a total loss**



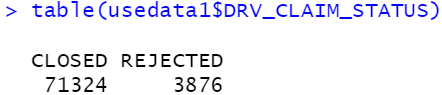
**Boo\_AntiTheft: in 80% of the claim cases no theft discount was given**



**Boo\_NCB: 74% did not receive any NCB of the total claim cases**

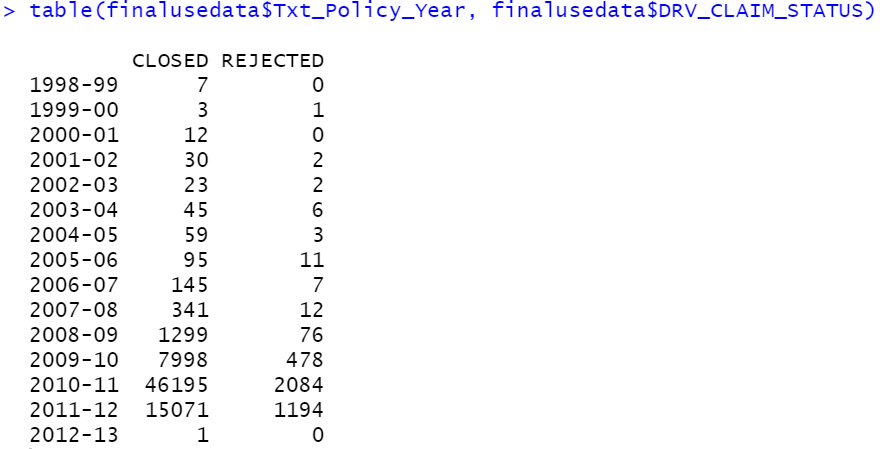


**Dependent Variable (DRV\_CLAIM\_STATUS): Only 5% of the cases are rejected of the total claim made**

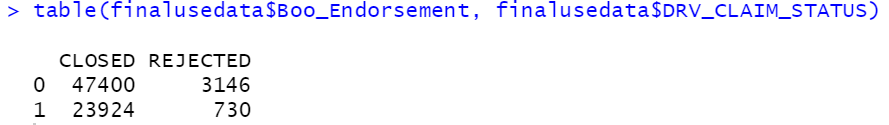


* 1. **Bivariate Analysis**

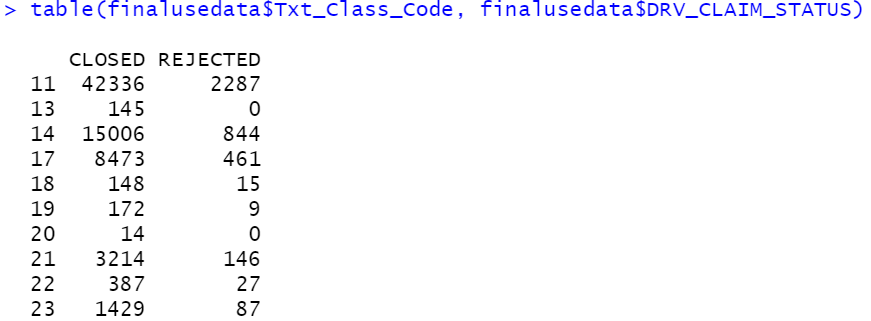
**Policy Year ~ Claim Status: Maximum Claim Rejection Ratio is for 2011-12**



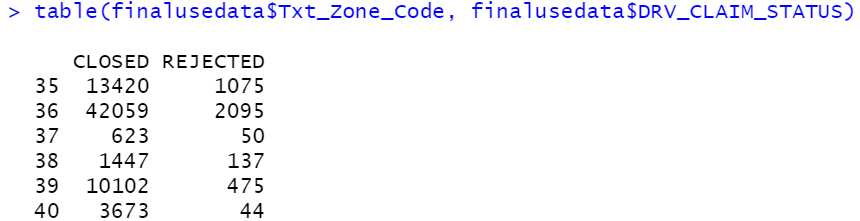
**Endorsement ~ Claim Status: Rejection Ratio is higher in non-endorsed cases**



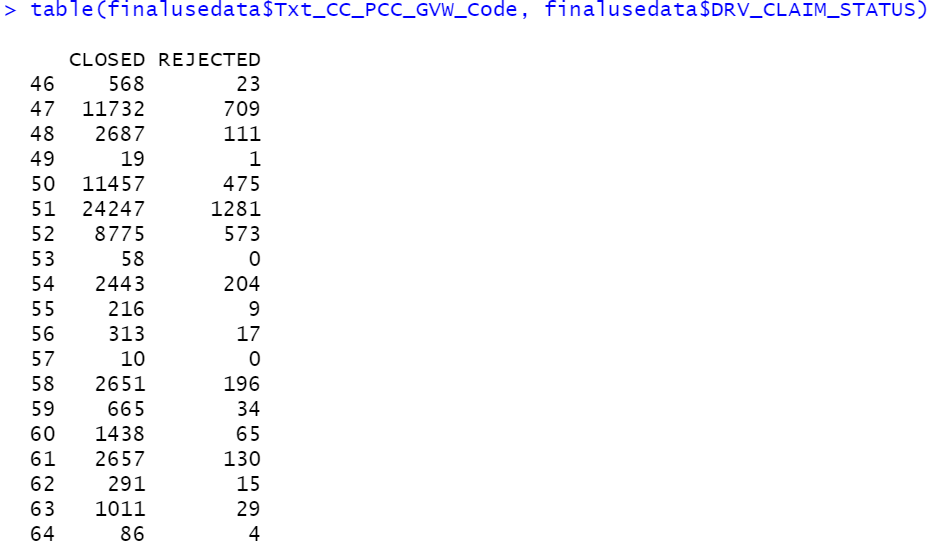
**Class Code ~ Claim Status: Class Code with more than 1000 claims, Rejection Ratio is highest for Class 23 (Special type of vehicles)**



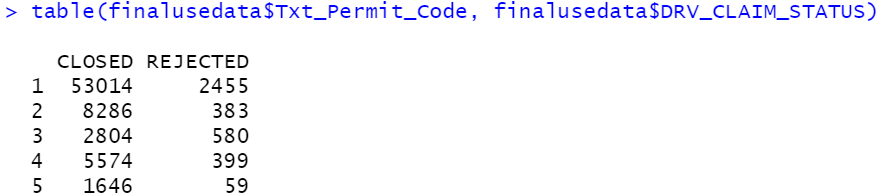
**Zone Code ~ Claim Status: Zone Code with more than 1000 claims, Rejection Ratio is highest for Zone Code 38 (Zone B - Sections 4.A, 4.C.2,4.C.3 and 4.D)**



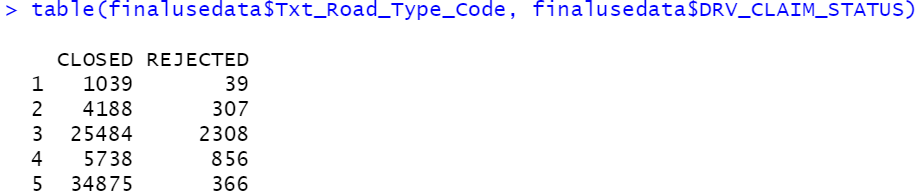
**CC\_PCC\_GVW\_Code ~ Claim Status: In this Code with more than 5000 claims, maximum Rejection Ratio is for code 52 (Exceeding 1500 CC (applicable for Private cars & Taxis))**



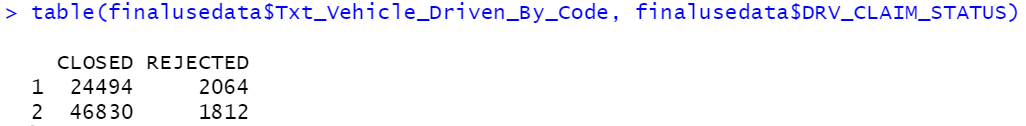
**Permit Code ~ Claim Status: Permit Code 3 (Zonal (More than one state) has the highest Rejection Ratio of 21%**



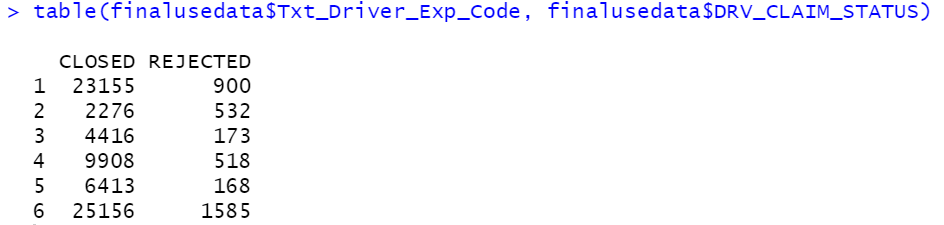
**Road Type ~ Claim Status: Road type 4 (District Roads) has the highest Rejection Ratio at 15%**



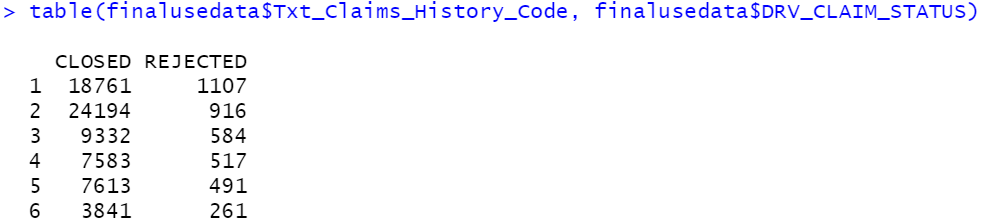
**Vehicle Driven By ~ Claim Status: Owners tend to put more fraudulent claims with Rejection Ratio of 8%**



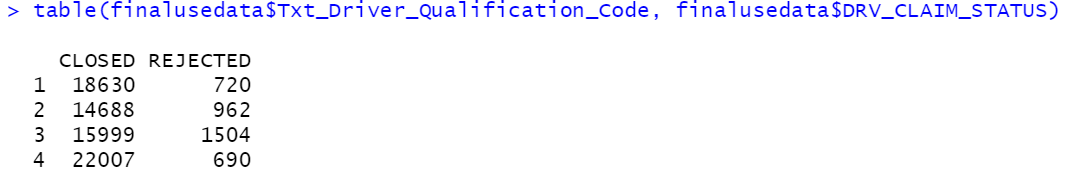
**Driver Exp ~ Claim Status: 1 year to <3 years tend to have highest Rejection Ratio at 23%**



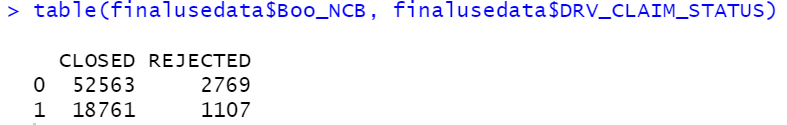
**Claim History ~ Claim Status: Claims with prior history of 3 claims tend to have higher Rejection Ratio at 6.8%**



**Driver Qualification ~ Claim Status: Maximum Rejection Ratio at 9.4% is from drivers with qualification of 12th Standard Pass**



**NCB ~ Claim Status: Rejection Ratio where NCB discount has been given are higher at 6%**



**Rejection Ratio = Count of Rejections/Count of Closed**

1. **Insights from EDA**

This dataset is a case of **imbalance nature** as the **dependent variable** in purview has only **5.2%** of “**Rejected**” cases which is of concern.

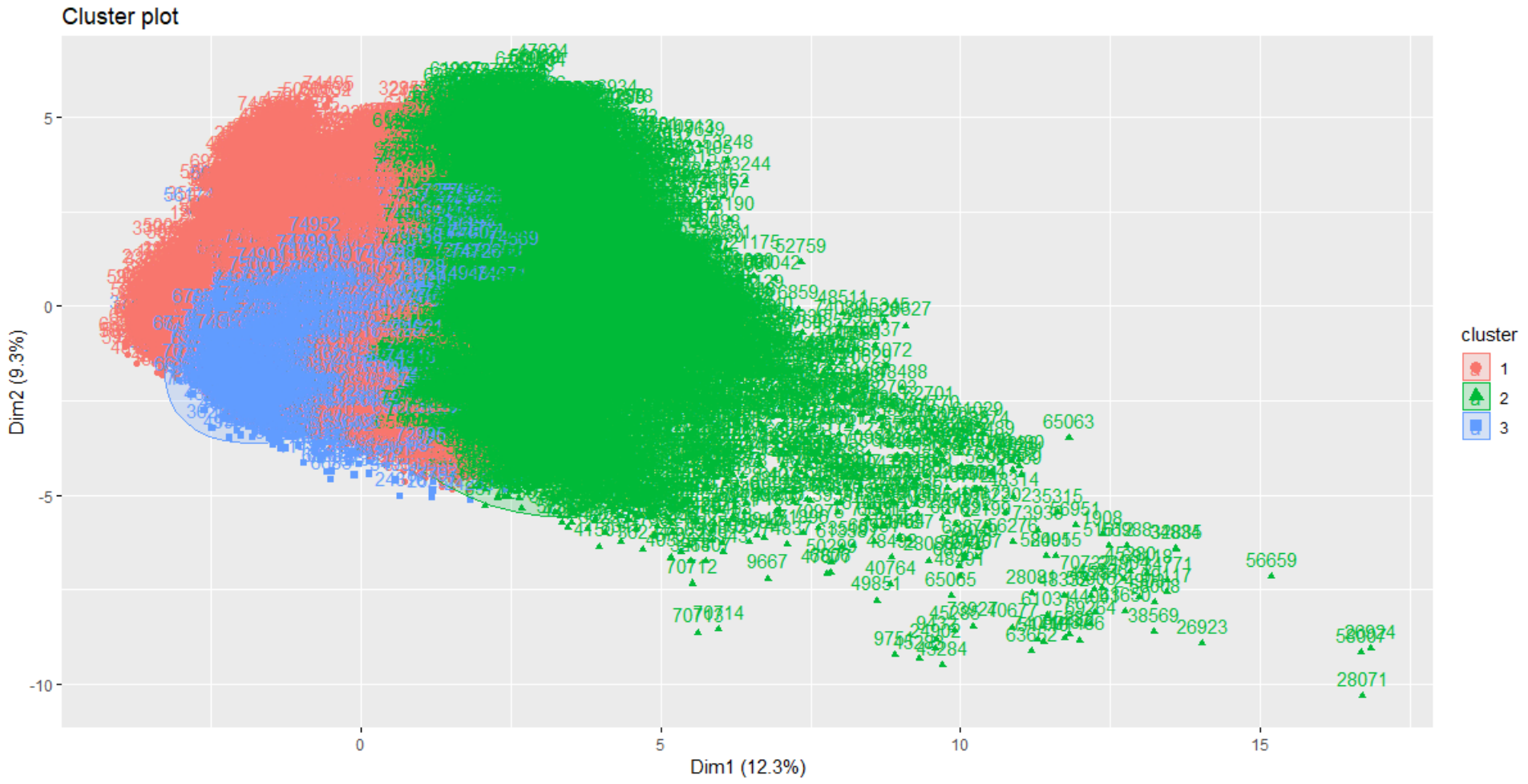
To overcome the issue during modelling, we will either **upscale minority value** or **down-scale the majority value** in dataset. We will use techniques such as **SMOTE** to create synthetic data so that our model is appropriate and given optimum outcome.

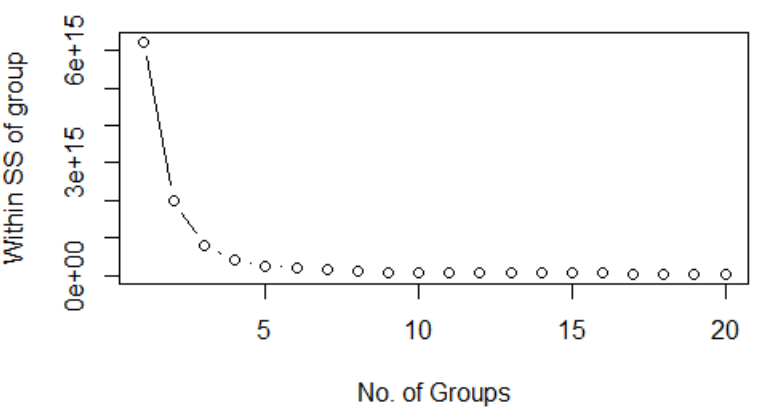
**Key Insights generated from the dataset are as follows:**

* Maximum number of claims were raised in policies purchased in 2010-11. Maximum number of claims were made in the year 2011-12. Maximum Claim Rejection Ratio is for year policy 2011-12
* 67% of claims are from not endorsed policies and claim Rejection Ratio is higher in non-endorsed cases
* 99.8% of claims belong to Package Policy
* 92% of claims are registered by Private Cars (59%), Two Wheelers (21%) & Goods Carrying vehicles other than three wheelers – Public (12%). But, in the Class Code with more than 1000 claims, Rejection Ratio is highest for Special type of vehicles
* 92% of the claims are registered Zone A - Sections 2 ,3,4.C.1 and 4.C.4 (19%), Zone B - Sections 2 ,3,4.C.1 and 4.C.4 (59%) and Zone C - Sections 4.A, 4.C.2,4.C.3 and 4.D (14%). But, among, Zone with more than 1000 claims, Rejection Ratio is highest for Zone B - Sections 4.A, 4.C.2,4.C.3 and 4.D
* 66% of the claims are made from people driving 75 to 150 cc (applicable for two wheelers) (17%), CC Not exceeding 1000CC (applicable for Private cars &Taxis) (16%) & CC Between 1000CC and 1500 CC (applicable for Private cars & Taxis) (34%). Moreover, in this Code with more than 5000 claims, maximum Rejection Ratio is for code 52 (Exceeding 1500 CC (applicable for Private cars & Taxis))
* 74% of the claims are made by people with local permit. But Zonal Permit holders (More than one state) has the highest Rejection Ratio of 21%
* 81% of the claims are made by non-hazardous goods carrier
* 84% of the claims are made by vehicle running on City/Town Roads and Others. But, District Roads category has the highest Rejection Ratio at 15%
* 65% of the claims are made by non-Owners. However, owners tend to put more fraudulent claims with Rejection Ratio of 8%
* 68% of the claims are made by rider with “<1 year” or “15 years and above” experience category but drivers with experience from 1 year to <3 years tend to have highest Rejection Ratio at 23%
* 60% of the claims are made with no prior claims or only 1 claim history. Claims with prior history of 3 claims tend to have higher Rejection Ratio at 6.8%
* There is no clear impact of qualification on claims made. However, maximum Rejection Ratio at 9.4% is from drivers with qualification of 12th Standard Pass
* 40% of the claims made had only 1 previous claim history before the current policy
* 98.6% of people who have made claim had opted for Wider Cover
* 86% of the claims made were due to Accident by External Means
* 93% of claim cases did not have a total loss
* In 80% of the claim cases no theft discount was given
* 74% did not receive any NCB of the total claim cases. However, Rejection Ratio where NCB discount has been given are higher at 6%

1. **Clustering**

We have implemented K-means clustering for analyzing the customer profile of maximum rejected customers. This will help the company identify the customer segment for whom there could be higher due diligence while giving the policy. Hence, this will act as the 1st stage filtering criteria.





There are 3 clusters chosen basis the above elbow plot.

We note that, the maximum rejections are observed in cluster 1, which will help us create the negative customer profile.

We find the least rejection cases (% of total cases) in cluster 2.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Cluster 1** | **Cluster 2** | **Cluster 3** |
| **CLOSED** | 39224 | 19024 | 13076 |
| **REJECTED** | 2970 | 493 | 413 |
| **% Rejection** | **7.03%** | **2.53%** | **3.06%** |

Thus, basis above output we can plot the Customer Profile leading to maximum rejection,

|  |  |  |
| --- | --- | --- |
| **Field Names** | **Description** | **Customer Profile** |
| Boo\_Endorsement | Y for Endorsement, otherwise N | Not endorsed |
| Txt\_Policy\_Code | Type of the Policy (Reference: Policy Code Master) | Package Policy |
| Txt\_Class\_Code | Type of vehicle (Reference: Class Code master) | Private Care |
| Txt\_Zone\_Code | Reference: Zone Code Master | Zone C - Sections 4.A, 4.C.2,4.C.3 and 4.D |
| Num\_Vehicle\_Age | Vehicle Age in completed years on inception/ renewal | <1.5 years |
| Txt\_CC\_PCC\_GVW\_Code | Reference: CC/ PCC/ GVW Code master | Exceeding 1500 CC (applicable for Private cars & Taxis) |
| Txt\_Colour\_Vehicle | Descriptive Entry | Other Colour |
| Num\_IDV | Sum insured rounded off to rupees | Higher SI policies are susceptible to fraudulent claims |
| Txt\_Permit\_Code | Reference: Permit Code Master | Local |
| Txt\_Nature\_Goods\_Code | Reference: Nature of Goods Code Master | Others |
| Txt\_Road\_Type\_Code | Reference: Road Type Code Master | Others |
| Txt\_Driver\_Exp\_Code | Reference: Driver Experience Code Master | <1 year |
| Txt\_Claims\_History\_Code | Reference: Claims History Code Master | 1 claim in last 5 years |
| Txt\_Driver\_Qualification\_Code | Reference: Driver Education Code Master | Below 10th Standard |
| Boo\_TPPD\_Statutory\_Cover\_only | If opted for restricted cover - "Y". If Wider cover applicable - "N" | Wider Cover Applicable |
| Boo\_AntiTheft | Y if Antitheft discount given, otherwise N | No antitheft discount given |

Thus, we recommend that the cases fulfilling the above profile should be duly investigated even before the policy purchase.

1. **Model Building**

We have built 4 models, Logistic Regression, Naïve Bayes, Random Forest and Bagging.

Data being imbalance in nature, we have built the models on both the as is dataset and over-sampled dataset. For over-sampling, we have used ROSE package in R. We have removed the imbalanced nature of the data and improved the minority class data proportion to 50%.

For the purpose of modelling, we also carried out PCA, reducing the total number of independent variables to 16, thus removing multicollinearity between the independent variables.

**Training Dataset**

With Balance Dataset

Model Performance Comparison and Model Selection

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameter** | **Logistic Regression** | **Naïve Bayes** | **Random Forest** | **Bagging** |
| Accuracy | 54.20% | 66.96% | 100.00% | 87.60% |
| Sensitivity | 97.90% | 38.67% | 100.00% | 86.87% |
| Specificity | 10.51% | 95.25% | 100.00% | 88.33% |
| ROC | 87.72% | 66.96% | 100.00% | 87.59% |
| KS | 62.97% | 33.91% | 100.00% | 75.20% |
| Gini | 35.90% | 9.53% | 49.32% | 16.58% |

With Imbalanced Dataset

Model Performance Comparison and Model Selection

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameter** | **Logistic Regression** | **Naïve Bayes** | **Random Forest** | **Bagging** |
| Accuracy | 94.84% | 87.89% | 100.00% | 96.29% |
| Sensitivity | 99.98% | 89.74% | 100.00% | 99.50% |
| Specificity | 0.00% | 53.81% | 100.00% | 37.33% |
| ROC | 91.51% | 71.77% | 100.00% | 68.42% |
| KS | 71.19% | 43.56% | 100.00% | 36.84% |
| Gini | 72.77% | 9.72% | 92.86% | 2.28% |

**Testing Dataset**

With Balance Dataset

Model Performance Comparison and Model Selection

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameter** | **Logistic Regression** | **Naïve Bayes** | **Random Forest** | **Bagging** |
| Accuracy | 93.15% | 41.53% | 96.84% | 87.30% |
| Sensitivity | 97.68% | 38.63% | 99.22% | 87.27% |
| Specificity | 9.97% | 95.01% | 53.14% | 87.79% |
| ROC | 87.56% | 66.82% | 98.27% | 87.53% |
| KS | 63.63% | 33.64% | 88.75% | 75.19% |
| Gini | 48.91% | 14.27% | 84.88% | 11.87% |

With Imbalanced Dataset

Model Performance Comparison and Model Selection

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameter** | **Logistic Regression** | **Naïve Bayes** | **Random Forest** | **Bagging** |
| Accuracy | 94.84% | 88.02% | 97.07% | 96.00% |
| Sensitivity | 100.00% | 89.88% | 99.27% | 99.40% |
| Specificity | 0.00% | 53.83% | 56.67% | 33.62% |
| ROC | 91.25% | 71.85% | 98.45% | 66.51% |
| KS | 71.62% | 43.71% | 89.00% | 36.84% |
| Gini | 72.94% | 9.65% | 89.99% | 2.20% |

The above comparison clearly indicates that Bagging has performed the best on the test dataset. The model presents an accuracy of over 87% with Sensitivity and Specificity being high.

We also note that the models are more stable in case of using balanced dataset. Hence, the final model consumption would be post balancing the dataset, using SMOTE or other over/under sampling techniques.

1. **Business Recommendations**

* Implementation of Bagging algorithm should be a part of the process which issuing policy or renewing an old policy
* Strict due diligence process to be set up where both the model recommends “REJECTED” output and the customer profile meets the negative profiling parameters
* Further, as a few models are slightly overfitting the data, it is recommended that the model is fine tuned every 3-6 months with addition of new database to fine tune the models.