Great Learning

Insurance Fraud (Final Submission)

Submitted By:

Sunny Kumar

1. Abstract

India is a huge market for insurance, but the industry is bleeding losses due to increasing fraud from inside and outside of the system. Insurance frauds leads to close to INR 40,000 crores loss which is close to 8.5% revenue of the insurance sector in India. These huge losses have prompted insurance companies to set-up a new anti-fraud department whose job is to identify the risk and loss to these frauds and to find out ways to reduce fraudulent claims.

2. Introduction

Insurance fraud occurs when either the buyer or the vendor of an insurance contract commits a criminal conduct. Selling insurance from non-existent companies, failing to submit premiums, and churning policies to generate more fees are all examples of issuer fraud. Exaggerated claims fabricated medical histories, post-dated insurance, viatical fraud, and forged deaths can all be examples of buyer fraud (James Chen, 2020).

Insurers and their extortion groups are recapturing ground and learn what unused practices see like to reply to extortion. Predictive analytics is playing a more grounded part as is entity analytics, the understanding of who a person is, and on the off chance that they are who they claim to be. Analytics engines can presently run these checks and raise concerns amid the on-boarding handle (Reuters Events, 2021).

This study will be solved insights and predictive information about acceptance of fraud cases and chances to raise issues and reasons. Study solves further solution to control and keep frauds rates down.

Trade protections scope secures businesses from misfortunes because of occasions that will happen amid the typical course of commerce. There are numerous sorts of protections for businesses counting scope for property harm, lawful obligation, and employee-related risks. Companies assess their protection needs to be based on potential dangers, which can change depending on the sort of environment in which the company works (Julia Kagan, 2020).

It is critical for little trade proprietors to weigh and assess their trade protections needs since they may have more individual monetary introduction within the occasion of a misfortune. In the event that a trade proprietor does not feel he or she has the capacity to successfully survey commerce chance and the requirements for scope, they ought to work with a legitimate, experienced, and licensed insurance broker. You'll be able to get a list of allowed operators in your state through your state's division of protection or the National Affiliation of Protections Commissioners. Also known as commercial lines protections, these inclusions incorporate property and casualty protections items for businesses. Commercial lines Protections makes a difference keep the economy running easily by ensuring businesses from potential misfortunes they couldn’t afford to cover on their possess, which permits businesses to function when it might have something else be as well unsafe to do so (Julia Kagan, 2020).

Data Description:

This data file is categorial data which include place, time, unique key number, and different classes. In project -1 we discuss about all the data references. In our finding we found we are having 75200 obs. Of 32 variables. And in apex. Number 1 you can see unique value of all variables. Apex. 2 showing class of all available variables. Apex 3 we found total blank space is 3725 in Date_distribution variables. We fixed this to replace with 'null' character.

Data frame & unique value: 75200 obs. of 32 variables:

Uniquekey	Txt_Policy_Year	Boo_Endorsement	Txt_Location_RTA
75200	15	2	5598

Txt_Policy_Code	Txt_Class_Code	Txt_Zone_Code	Num_Vehicle_Age
5	10	6	28

Txt_CC_PCC_GVW_Code	Txt_Colour_Vehicle	Num_IDV	Txt_Permit_Code
19	103	33303	5

Txt_Nature_Goods_Code	Txt_Road_Type_Code	Txt_Vehicle_Driven_By_Code	Txt_Driver_Exp_Code
2	5	2	6

Txt_Claims_History_Code	Txt_Driver_Qualification_Code	Txt_Incurred_Claims_Code	Boo_TPPD_Statutory_Cover_only
6	4	9	2

Txt_Claim_Year	Date_Accident_Loss	Txt_Place_Accident	Date_Claim_Intimation
14	2076	15552	1911

Txt_TAC_NOL_Code	Date_Disbursement	Boo_OD_Total_Loss	DRV_CLAIM_AMT
27	1765	2	26127

DRV_CLAIM_STATUS	Boo_AntiTheft	Boo_NCB	Num_Net_OD_Premium
2	2	2	3554

Class of data variables:

Uniquekey	Txt_Policy_Year	Boo_Endorsement	Txt_Location_RTA
"integer"	"character"	"integer"	"character"

Txt_Policy_Code	Txt_Class_Code	Txt_Zone_Code	Num_Vehicle_Age
"integer"	"integer"	"integer"	"integer"

Txt_CC_PCC_GVW_Code Txt_Colour_Vehicle Num_IDV Txt_Permit_Code

"integer" "character" "numeric" "integer"

Txt_Nature_Goods_Code Txt_Road_Type_Code Txt_Vehicle_Driven_By_Code Txt_Driver_Exp_Code

"integer" "integer" "integer" "integer"

Txt_Claims_History_Code Txt_Driver_Qualification_Code Txt_Incurred_Claims_Code Boo_TPPD_Statutory_Cover_only

"integer" "integer" "integer" "integer"

Txt_Claim_Year Date_Accident_Loss Txt_Place_Accident Date_Claim_Intimation

"character" "character" "character" "character"

Txt_TAC_NOL_Code Date_Disbursement Boo_OD_Total_Loss DRV_CLAIM_AMT

"integer" "character" "integer" "numeric"

DRV_CLAIM_STATUS Boo_AntiTheft Boo_NCB Num_Net_OD_Premium

"character" "integer" "integer" "integer"

Fixing blank space

Uniquekey Txt_Policy_Year Boo_Endorsement Txt_Location_RTA

0 0 0 0

Txt_Policy_Code Txt_Class_Code Txt_Zone_Code Num_Vehicle_Age

0 0 0 0

Txt_CC_PCC_GVW_Code Txt_Colour_Vehicle Num_IDV Txt_Permit_Code

0 0 0 0

Txt_Nature_Goods_Code Txt_Road_Type_Code Txt_Vehicle_Driven_By_Code Txt_Driver_Exp_Code

0 0 0 0

Txt_Claims_History_Code Txt_Driver_Qualification_Code Txt_Incurred_Claims_Code Boo_TPPD_Statutory_Cover_only

0 0 0 0

Txt_Claim_Year Date_Accident_Loss Txt_Place_Accident Date_Claim_Intimation

0	0	0	0

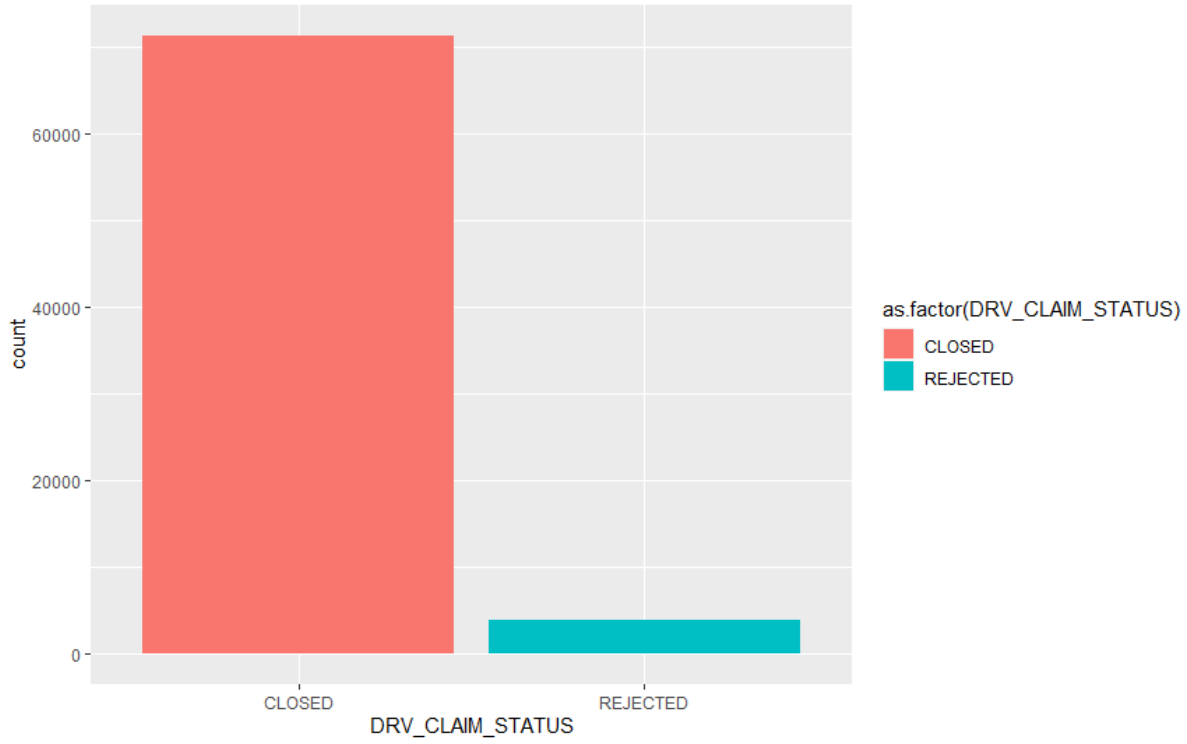
Txt_TAC_NOL_Code	Date_Disbursement	Boo_OD_Total_Loss	DRV_CLAIM_AMT
0	3735	0	0

DRV_CLAIM_STATUS	Boo_AntiTheft	Boo_NCB	Num_Net_OD_Premium
0	0	0	0

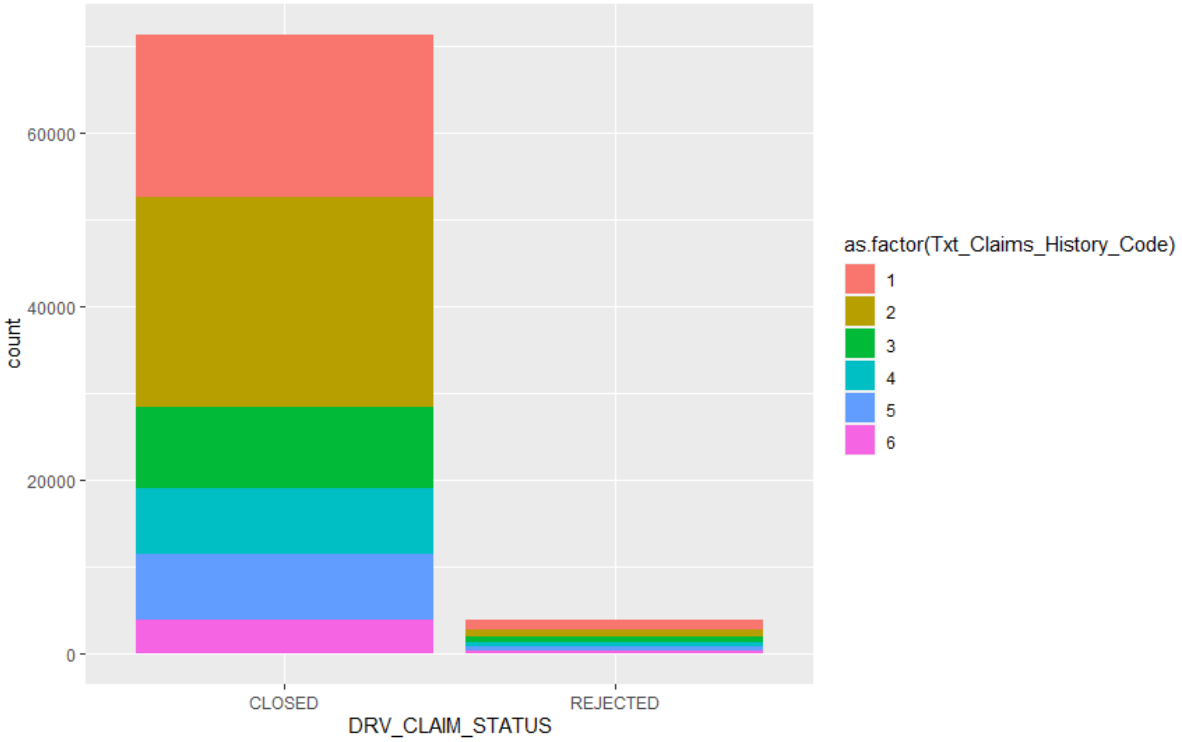
EDA:

We find the frequency of our variables before modelling are:

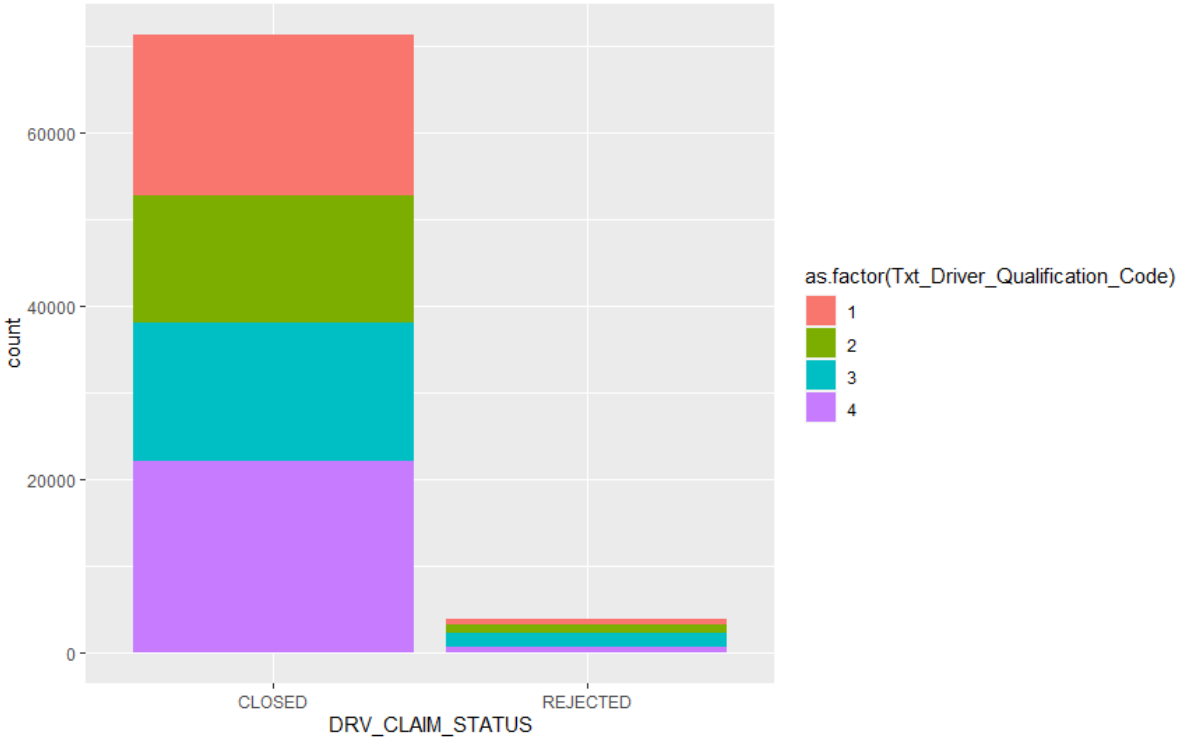
In Claim status we can see in bar plot closed status is more than 60k and Rejected status is less than 3000.



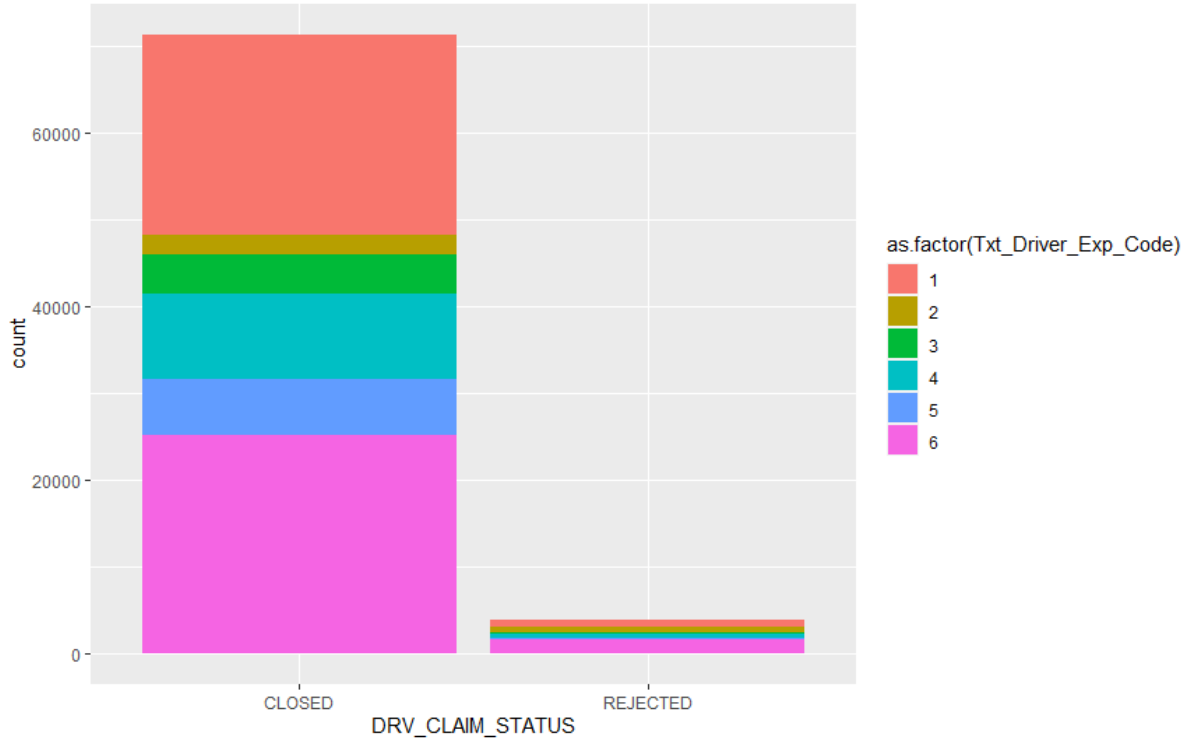
Claims history bar graph represents claims taking 1 time is 25000 and no claims is 20000. 2 times claims taking by customers 10000 times. 3- & 4-times claims taking by customers are 8000 and 5 or more than 5 taking by customers is 4000 times.



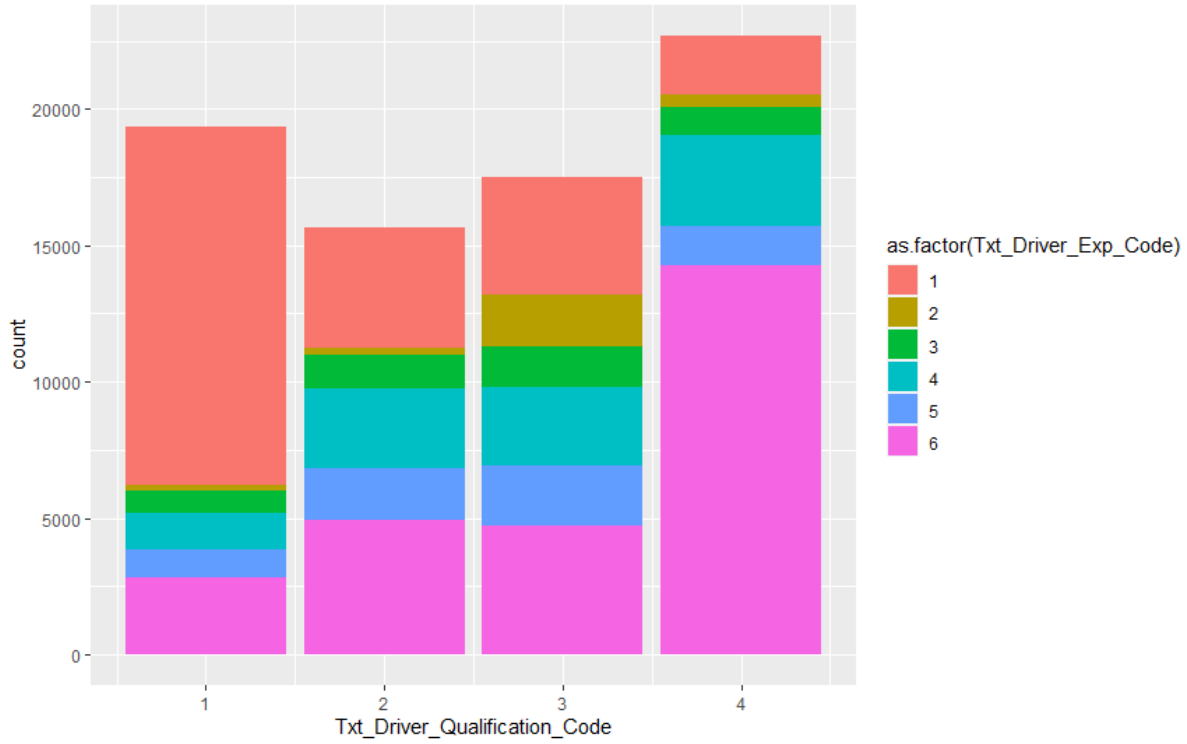
Qualification status of drivers are 22500 drivers is graduate and postgraduate according to qualification code. 19000 drivers are less than10 standard. 12th standard is 17500 drivers and more than 15000 is 10th standard qualification. So, our mostly claim drivers are having lower education which is increased the reason of more accident and frauds.



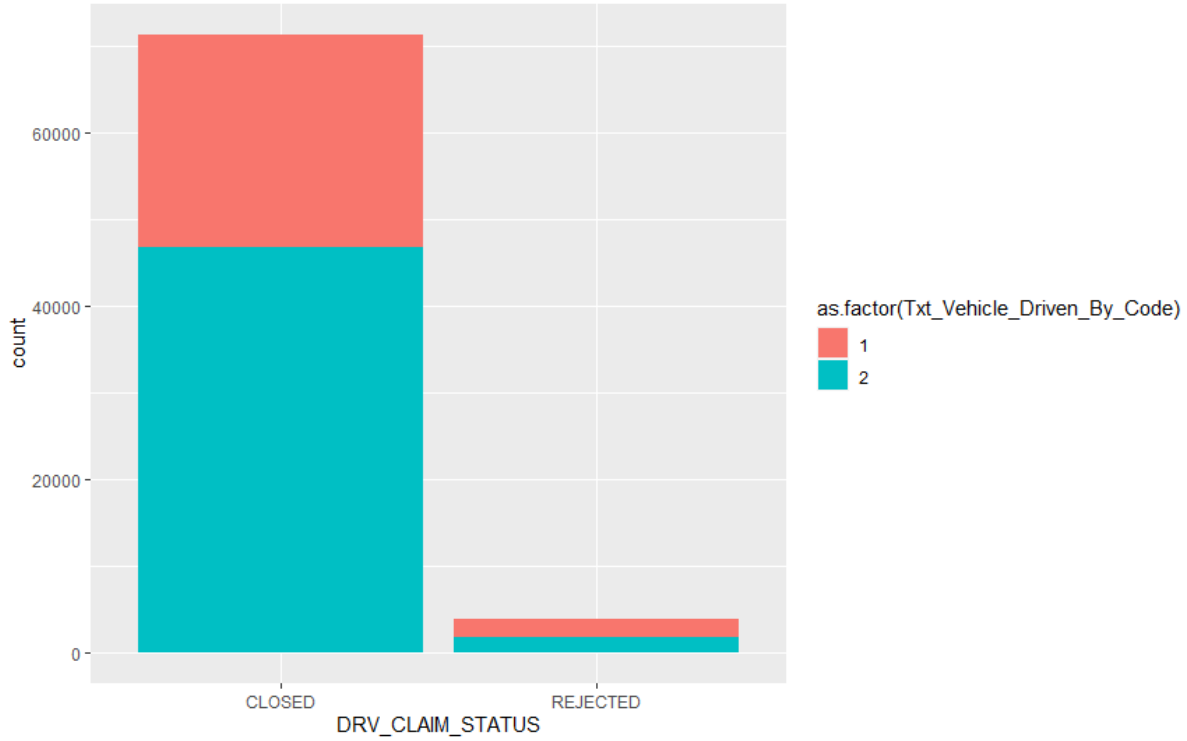
Driving experience is almost 28000 drivers having is more than 15 years' experience and 24000 drivers having less than 1 year experience which increase more accident probability.



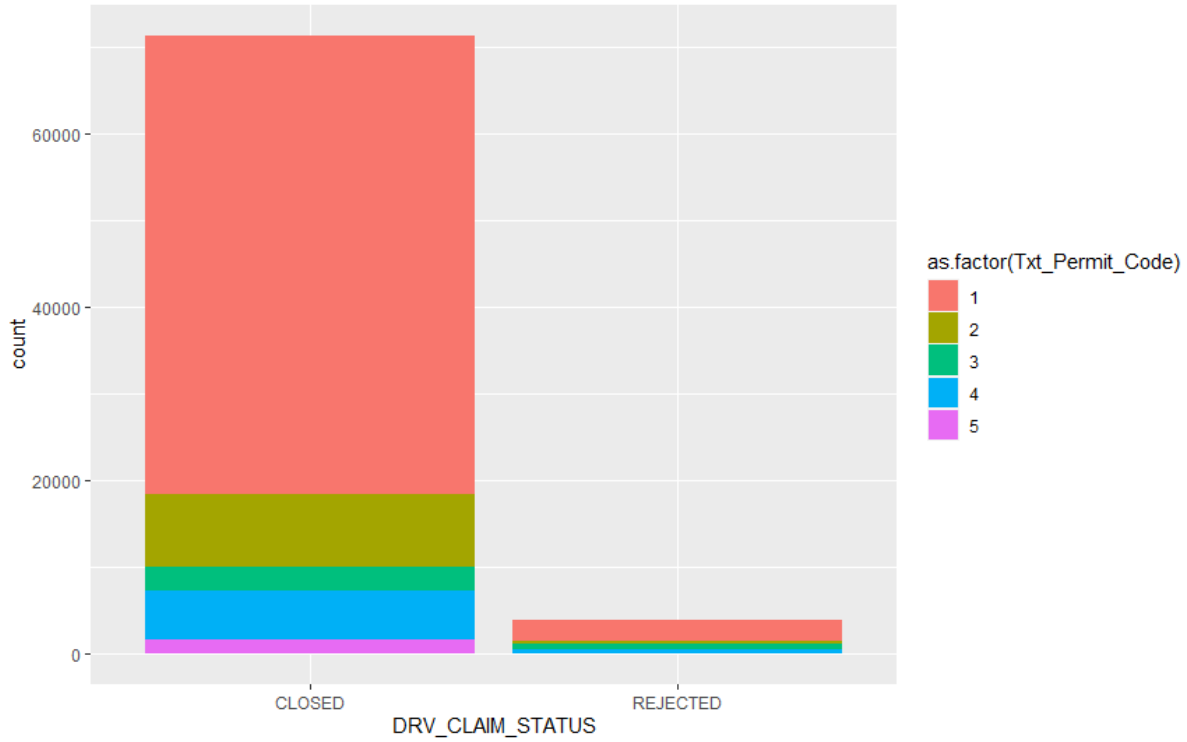
Corelation of qualification and driving experience proved here both the theory of less expireance and less than 12th qualification people are make more accidents and cause more claims.



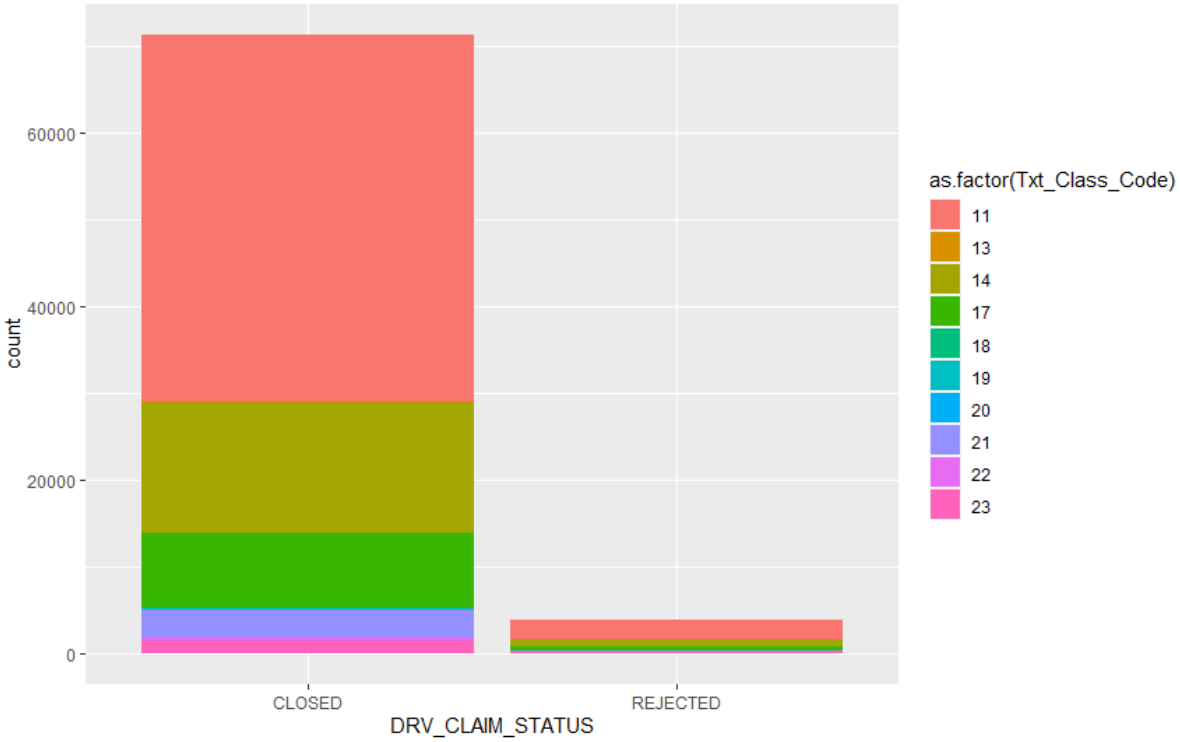
Vehicle drive by mostly other drivers 58000 that's why it's more chances of careless behaviour.



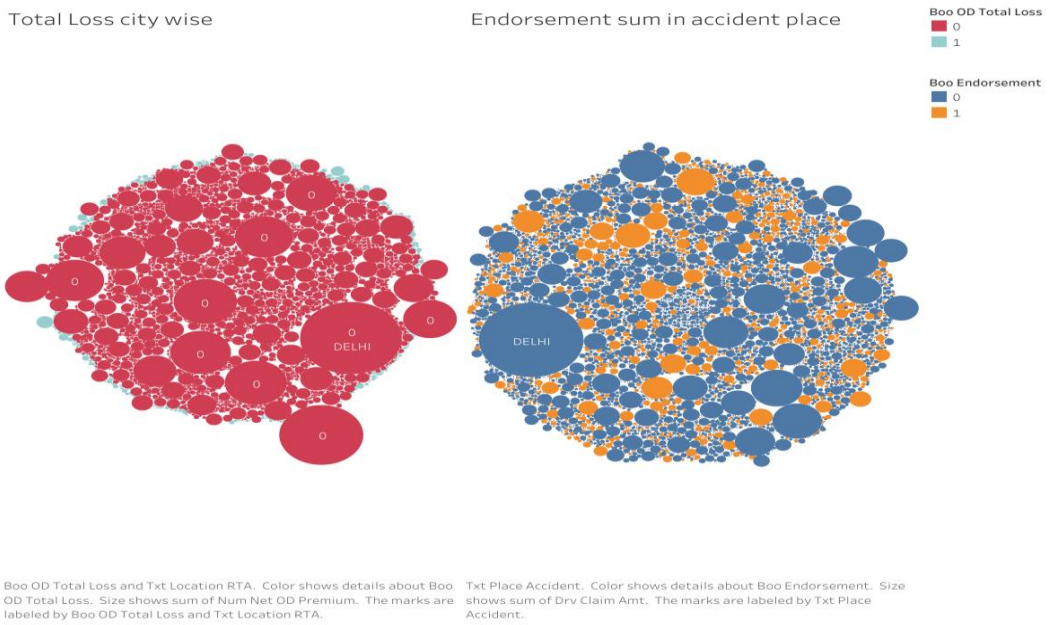
Permit of car is mostly for local areas and having more claims then others.



Class of automobile claimers are private car and 2 wheelers auto. But major area is covered by cars 45000.

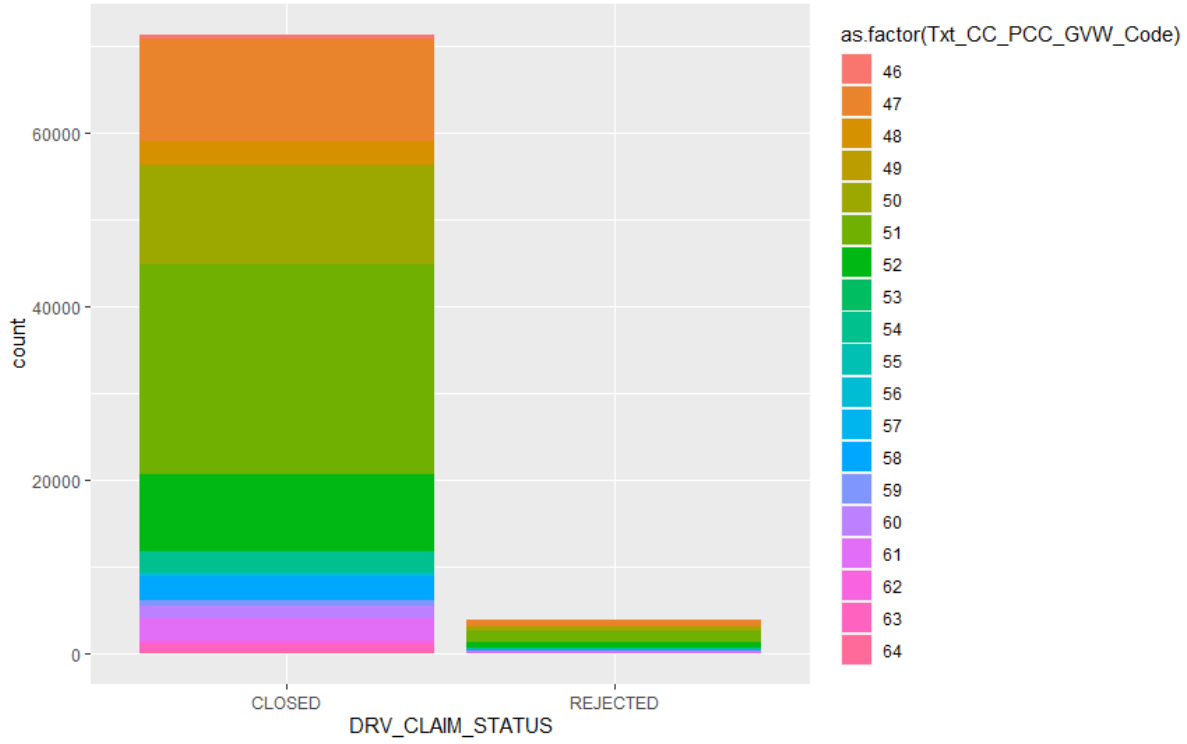


2.1.1. Correlation Between Total loss and endorsement sum in accident place



Total loss city wise and accident places are more Delhi Ncr and other big cities there belong graduate and private car in local area.

In CC and power auto is having more part covered in insurance Calame. Code 51 is more in graph which means 1000cc to 1500 cc for private cars and taxis.



Data pre-processing is a predominant step in machine learning to yield highly accurate and insightful results. Greater the quality of data, the greater is the reliability of the produced results. **Incomplete, noisy, and inconsistent data** are the inherent nature of real-world datasets. Data pre-processing helps in increasing the quality of data by filling in missing incomplete data, smoothing noise, and resolving inconsistencies.

- **Incomplete data** can occur due to many reasons. Appropriate data may not be persisted due to a misunderstanding, or because of instrument defects and malfunctions.
- **Noisy data** can occur for several reasons (having incorrect feature values). The instruments used for the data collection might be faulty. Data entry may contain human or instrument errors. Data transmission errors might occur as well.

There are many stages involved in data pre-processing.

- **Data cleaning** attempts to impute missing values, removing outliers from the dataset.
- **Data integration** integrates data from a multitude of sources into a single data warehouse.

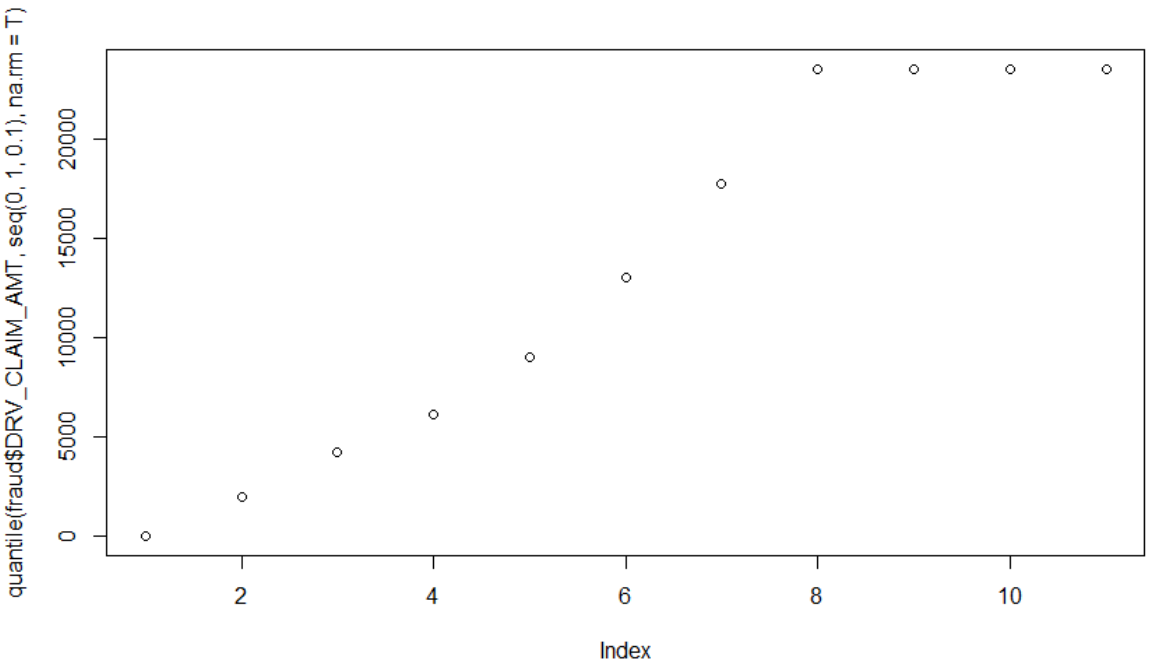
- **Data transformation** such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurement.

- **Data reduction** can reduce the data size by dropping out redundant features. Feature selection and feature extraction techniques can be used.

Outlier treatment

We found 0-23531 range is perfect for our data set and it's showing continue data flow.

```
> #Check outliers
> plot(quantile(fraud$DRV_CLAIM_AMT, seq(0,1,0.1),na.rm = T))
> quantile(fraud$DRV_CLAIM_AMT, seq(0,1,0.1),na.rm = T)
    0%    10%    20%    30%    40%    50%    60%    70%    80%    90%   100%
   0.0  2000.0  4261.0  6125.7  9000.0 13000.0 17712.8 23531.0 23531.0 23531.0 23531.0
> fraud$DRV_CLAIM_AMT[fraud$DRV_CLAIM_AMT >23531]<-23531
```



Machine Learning Models Building

Building machine learning models for building machine learning models there are a few models display interior the Sklearn module. Sklearn gives two sorts of models i.e. regression and classification. Our dataset's target variable is to foresee whether extortion is detailed or not. So for this kind of issue, we utilize classification models. But some time recently fitting our dataset to its demonstration, to begin with we got to isolate the indicator. variable and the target variable, then we pass this variable to the train_test_split strategy to form an arbitrary test and prepare subset.

What is train_test_split, it may be a work in sklearn demonstrate choice for part information clusters into two subsets for preparing information and testing information. With this function, you don't ought to partition the dataset physically. By default, sklearn train_test_split will make irregular segments for the two subsets. In any case, you'll too indicate a irregular state for the operation.

It gives four yields x_train, x_test, y_train and y_test. The x_train and x_test contains the preparing and testing indicator factors whereas y_train and y_test contains the preparing and testing target variable. After performing train_test_split we got to select the models to pass the preparing variable. We can construct as numerous models as we want to compare the precision given by these models and to choose the leading demonstrate among them.

Logistic regression

We may also wish to see measures of how well our model fits. This can be particularly useful when comparing competing models.

The output produced by '**summary(mylogit)'** included indices of fit (shown below the coefficients), including the null and deviance residuals and the AIC. One measure of model fit is the significance of the overall model. This test asks whether the model with predictors fits significantly better than a model with just an intercept (i.e., a null model). The test statistic is the difference between the residual deviance for the model with predictors and the null model. The test statistic is distributed chi-squared with degrees of freedom equal to the differences in degrees of freedom between the current and the null model (i.e., the number of predictor variables in the model). To find the difference in deviance for the two models (i.e., the test statistic) we can use the command:

```
with(mylogit, null.deviance - deviance)

## [1] 26080.57
```

The degrees of freedom for the difference between the two models is equal to the number of predictor variables in the mode, and can be obtained using:

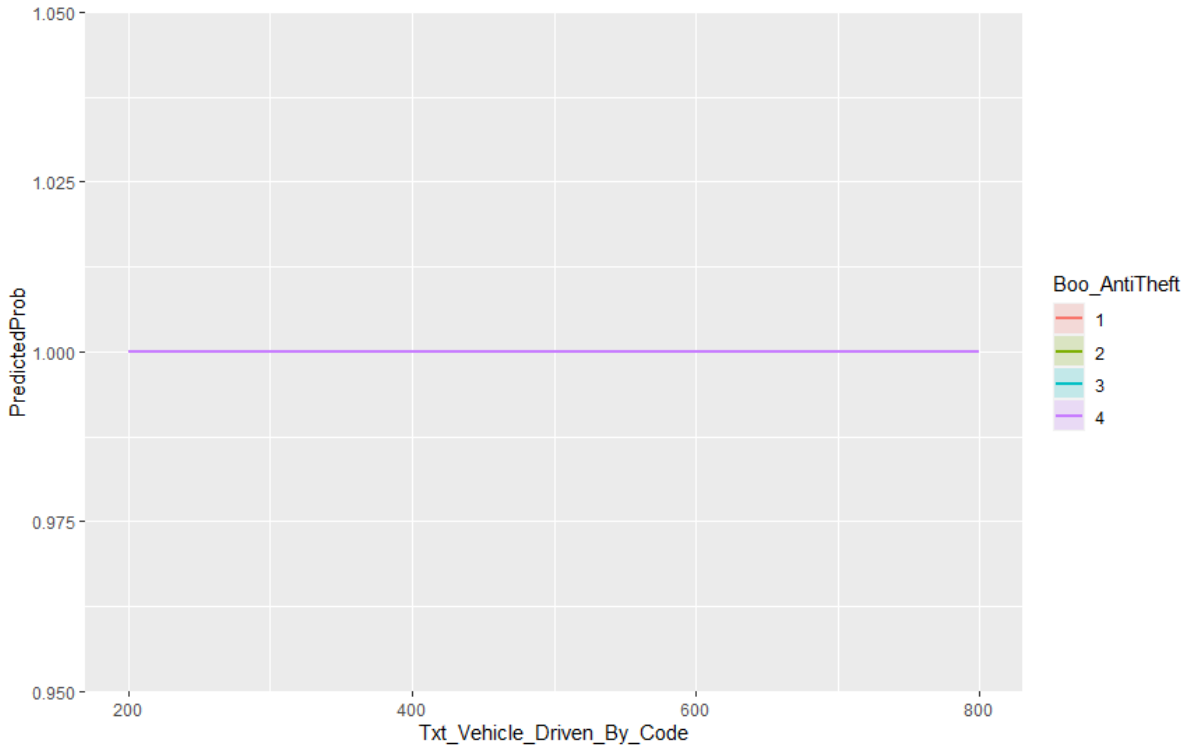
```
with(mylogit, df.null - df.residual)

## [1] 4
```

Finally, the p-value can be obtained using:

```
with(mylogit, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))

## [1] 0
```

The chi-square of 41.46 with 5 degrees of freedom and an associated p-value of less than 0.001 tells us that our model fits significantly better than an empty model. This is sometimes called a likelihood ratio test (the deviance residual is -2*log likelihood). To see the model's log likelihood, we type:

```
logLik(mylogit)

## 'log Lik.' -24371.3 (df=5)
```



So, I found logistic regression model more accurate.

```
# By using Decision Tree

dtc = DecisionTreeClassifier()
dtc.fit(x_train, y_train)
preddtc = dtc.predict(x_test)
print("Accuracy Score: {}".format(accuracy_score(y_test, preddtc)*100))
print("f1_Score: {}".format(f1_score(y_test, preddtc)*100))
print(confusion_matrix(y_test, preddtc))
print(classification_report(y_test, preddtc))
```

```
Accuracy Score: 78.53982300884957
f1_Score: 79.40552016985139
[[168  50]
 [ 47 187]]
              precision    recall  f1-score   support

           0       0.78      0.77      0.78       218
           1       0.79      0.80      0.79       234

    accuracy                           0.79       452
   macro avg       0.79      0.78      0.79       452
weighted avg       0.79      0.79      0.79       452
```

Decision trees can be built by an algorithmic approach that can part the dataset in several ways based on distinctive conditions. The two primary substances of a tree are choice hubs, where the information is part and takes off, where we get the result.

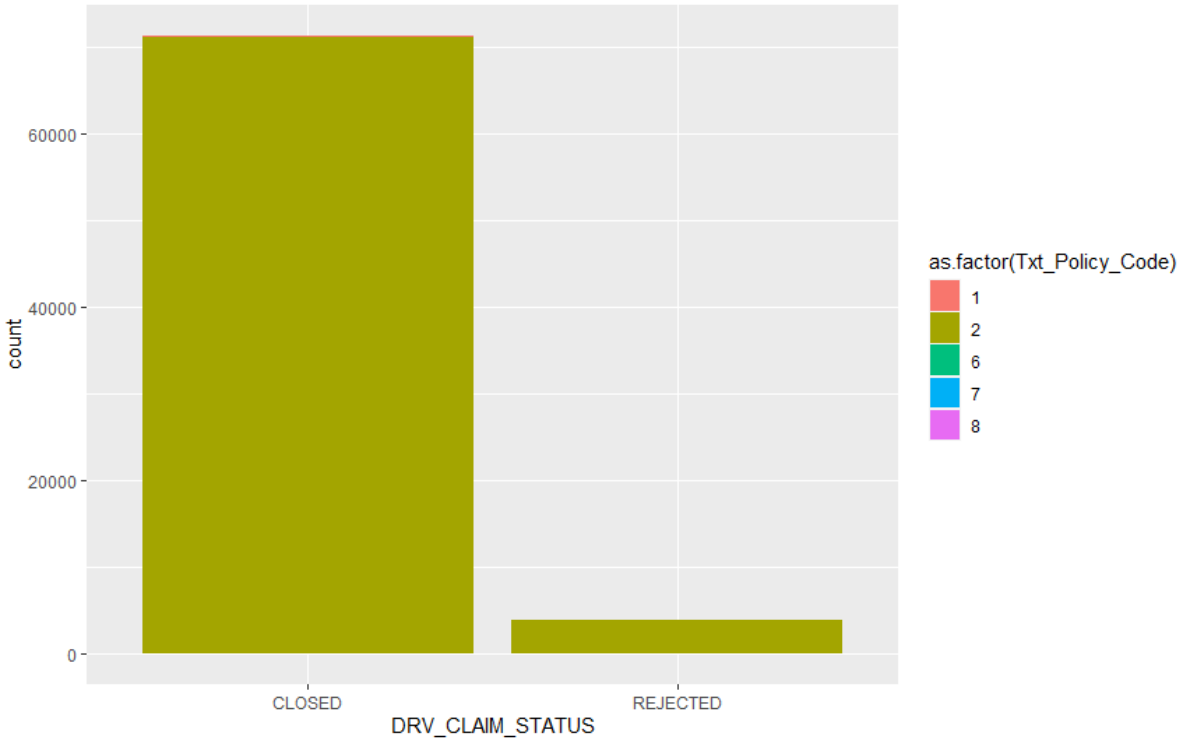
So also, a RANDOM FOREST makes decision trees on information tests and after that gets the expectation from each of them and at long last chooses the finest arrangement by implies of voting. It is an outfit strategy that is superior to a single decision tree since it diminishes the over-fitting by averaging the result.

```
# By using Random Forest

rfc = RandomForestClassifier()
rfc.fit(x_train, y_train)
predrfc = rfc.predict(x_test)
print("Accuracy Score: {}".format(accuracy_score(y_test, predrfc)*100))
print("f1_Score: {}".format(f1_score(y_test, predrfc)*100))
print(confusion_matrix(y_test, predrfc))
print(classification_report(y_test, predrfc))
```

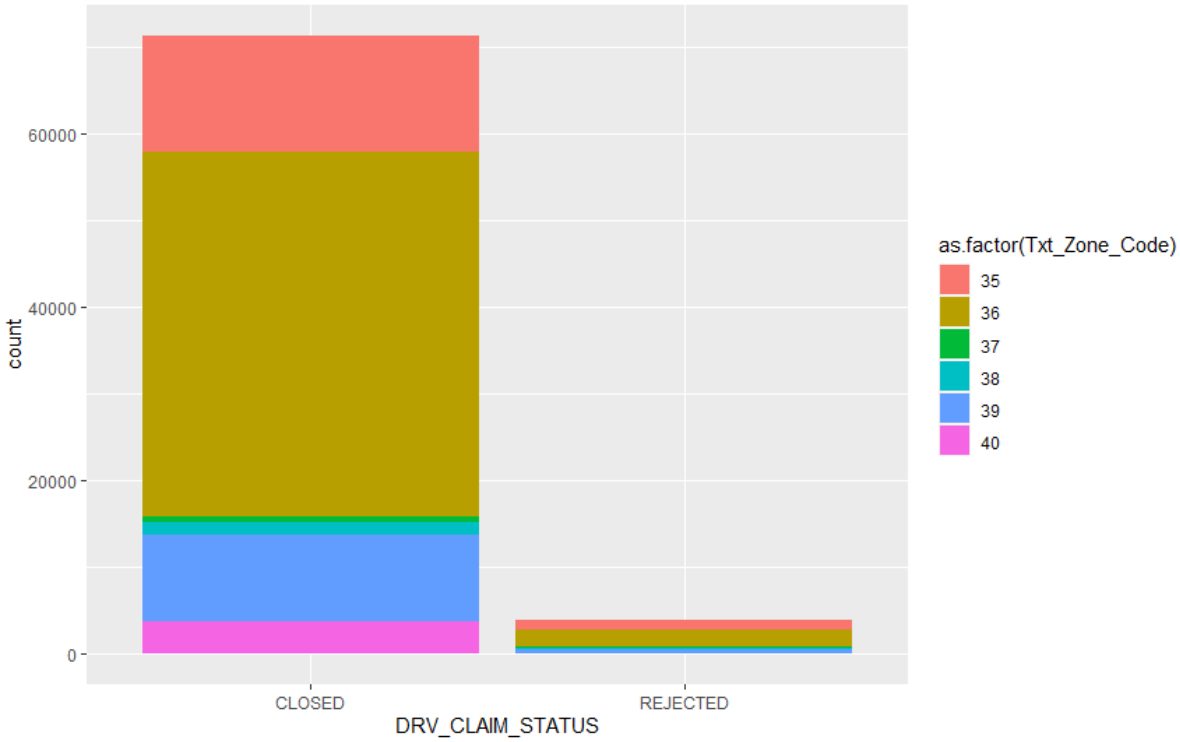
```
Accuracy Score: 85.39823008849558
f1_Score: 85.20179372197309
[[196  22]
 [ 44 190]]
              precision    recall  f1-score   support

           0       0.82      0.90      0.86       218
           1       0.90      0.81      0.85       234

    accuracy                           0.85       452
   macro avg       0.86      0.86      0.85       452
weighted avg       0.86      0.85      0.85       452
```

Focused on these points where more chances of frauds :

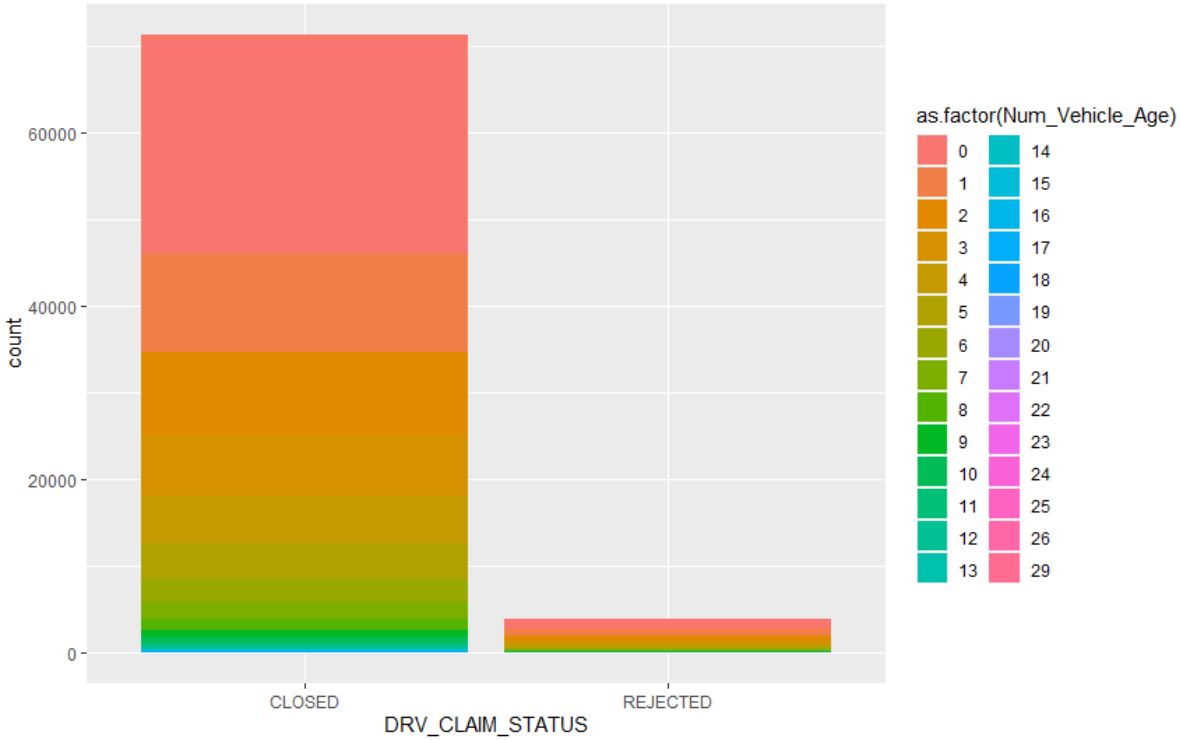
All policy code which insurance company sale and customer purchase was package policy



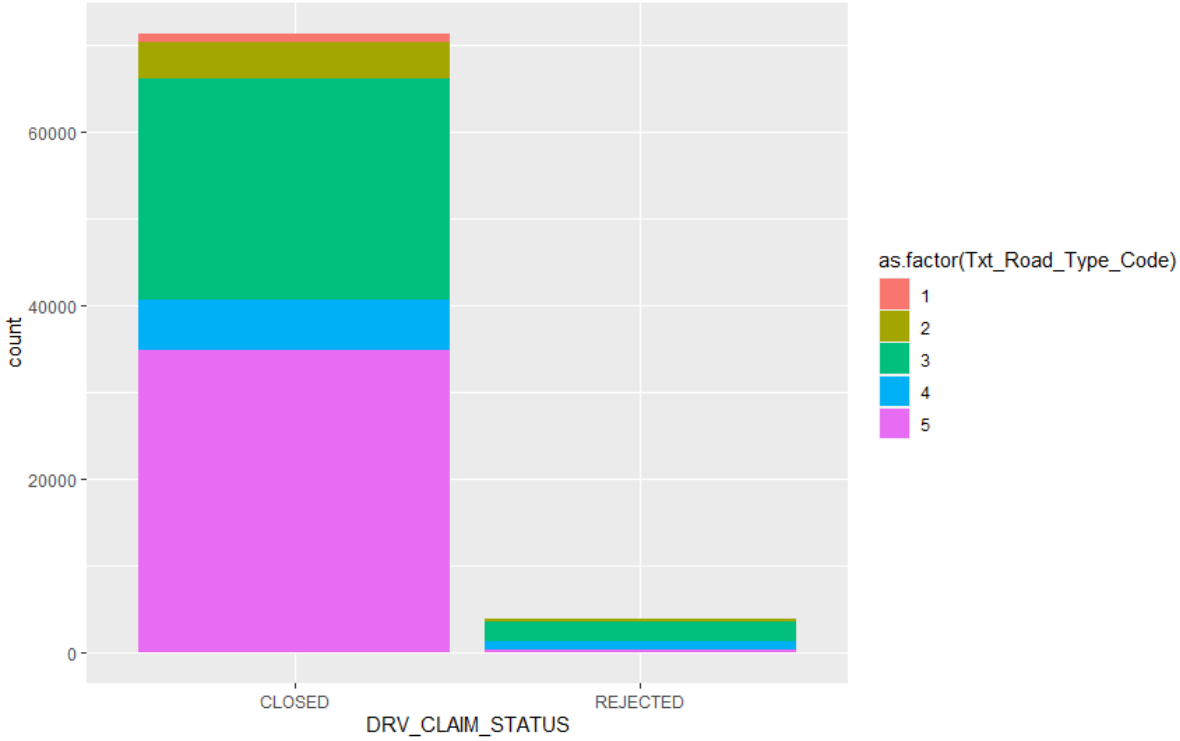
- Mostly claimed zone is zone 2



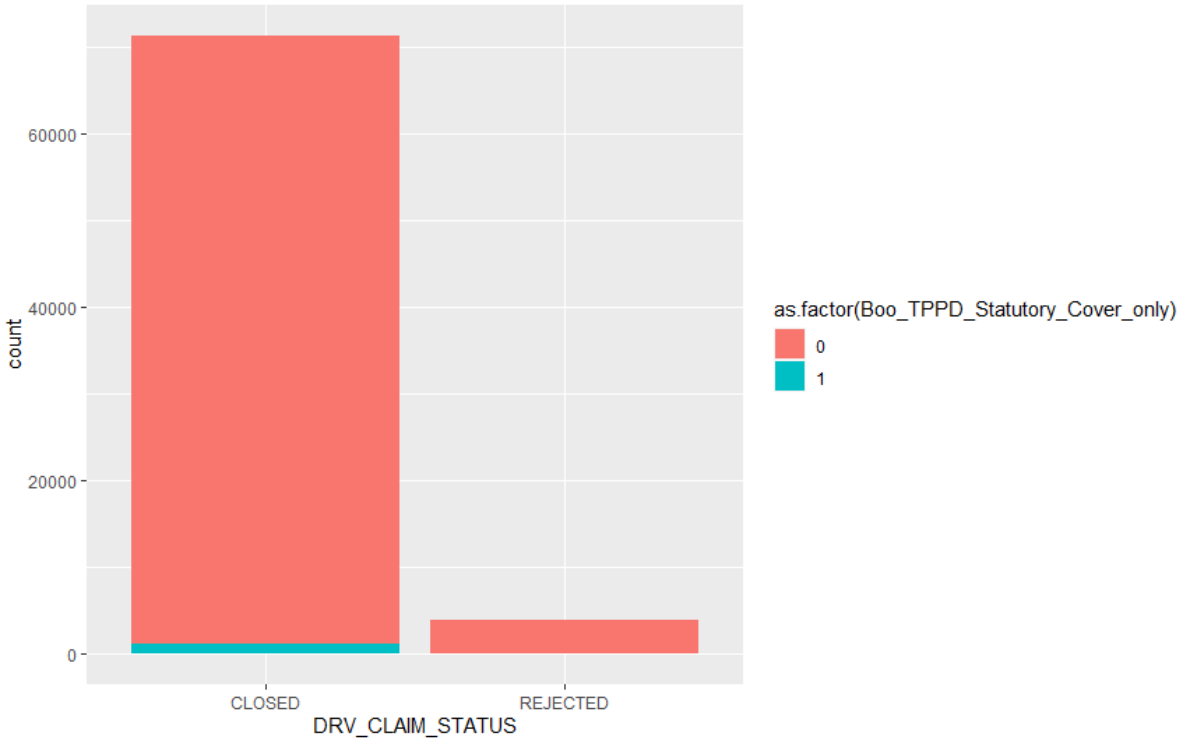
New cars age (0-5) years old cars having more claims. But after 6-19 years car we have to be more careful for claims because already they are old and need several works.



Road type others are coming more cases which showing more chances of frauds or lack of information. Number 3 code showing city/town which we found earlier more accidents and claims conduct in big cities.



Claims type is statuary in more cases



Remarks

This venture has built a demonstrate that can identify auto protections extortion. In doing so, the show can diminish misfortunes for protections companies. The challenge behind extortion location in machine learning is that fakes are distant less common as compared to legit protections claims.

The leading and last fitted show was a weighted Logistic regression that shouted a F1 score of 0.85 and a ROC AUC of 1.0. The demonstrate performed amazing. The model's F1 score and ROC AUC scores were the most elevated among the other models. In conclusion, the show was able to accurately recognize between extortion claims and legit claims with tall exactness.

Claims history bar graph represents claims taking 1 time is 25000 and no claims is 20000. 2 times claims taking by customers 10000 times. 3- & 4-times claims taking by customers are 8000 and 5 or more than 5 taking by customers is 4000 times. Qualification status of drivers are 22500 drivers is graduate and postgraduate according to qualification code. 19000 drivers are less than10 standard. 12th standard is 17500 drivers and more than 15000 is 10th standard qualification. So, our mostly claim drivers are having lower education which is increased the reason of more accident and frauds.
Driving experience is almost 28000 drivers having is more than 15 years' experience and 24000 drivers having less than 1 year experience which increase more accident probability. Corelation of qualification and driving experience proved here both the theory of less experience and less than 12th qualification people are make more accidents and cause more claims.
Road type others are coming more cases which showing more chances of frauds or lack of information. Number 3 code showing city/town which we found earlier more accidents and claims conduct in big cities.

Lack of location ratio it’s hard to making map but major big cities having more claimers and frauds. So, have to fix some policy strategies where we can make down frauds.