Introduction

The aim of the project is to identify a model for predicting whether an insurance claim will be made or not by training several models to find the optimized one. As any acceptable project, I have started with an exploratory data analysis to get a general view of the dataset, as well as I could without access to metadata.

Our initial exploration reveals that the following 3 columns contain null values and require handling:

1. Claim: The target of our model contains a null value. Since the data in this column is binary 1, 0 signifying whether a claim was made or not, replacing the column with average would not work. Instead, we use the median to replace the null value.
2. Category_anomaly: A column with binary value whose nulls will be replaced by the median of the column.
3. Repair_date: Since the column is supposed to be a date, first we convert the datatype to datetime, then we replace the missing value with the median.

Now, our dataset does not contain any null values but this does not indicate that the data is clean.

Using a df.describe() returns valuable information regarding the numeric values. We can compare the min, max with mean in order see if the data is skewed or compare max and 75% to find whether there are glaring outliers in our data.

We can see clearly that our data has several issues that need addressing. First, the engine_size column shows up as a string with a float followed by 'L' . We can address this issue by striping the 'L' and casting the values as float.

The Runned_miles column has a unique problem in that the minimum value it contains is a negative column but there is no possible way to perceive that a car has runned negative miles, thus we convert the values of the column to their absolute values assuming the negative was placed by mistake. The issue deepens further once we create a boxplot which shows the large number of outliers. We have a decision to make as to whether these outliers should be replaced or kept because they can be informative. I have decided to create a new column that replaces the outliers with the average of the column and test it in my models. Same treatment was applied to repair_hours.

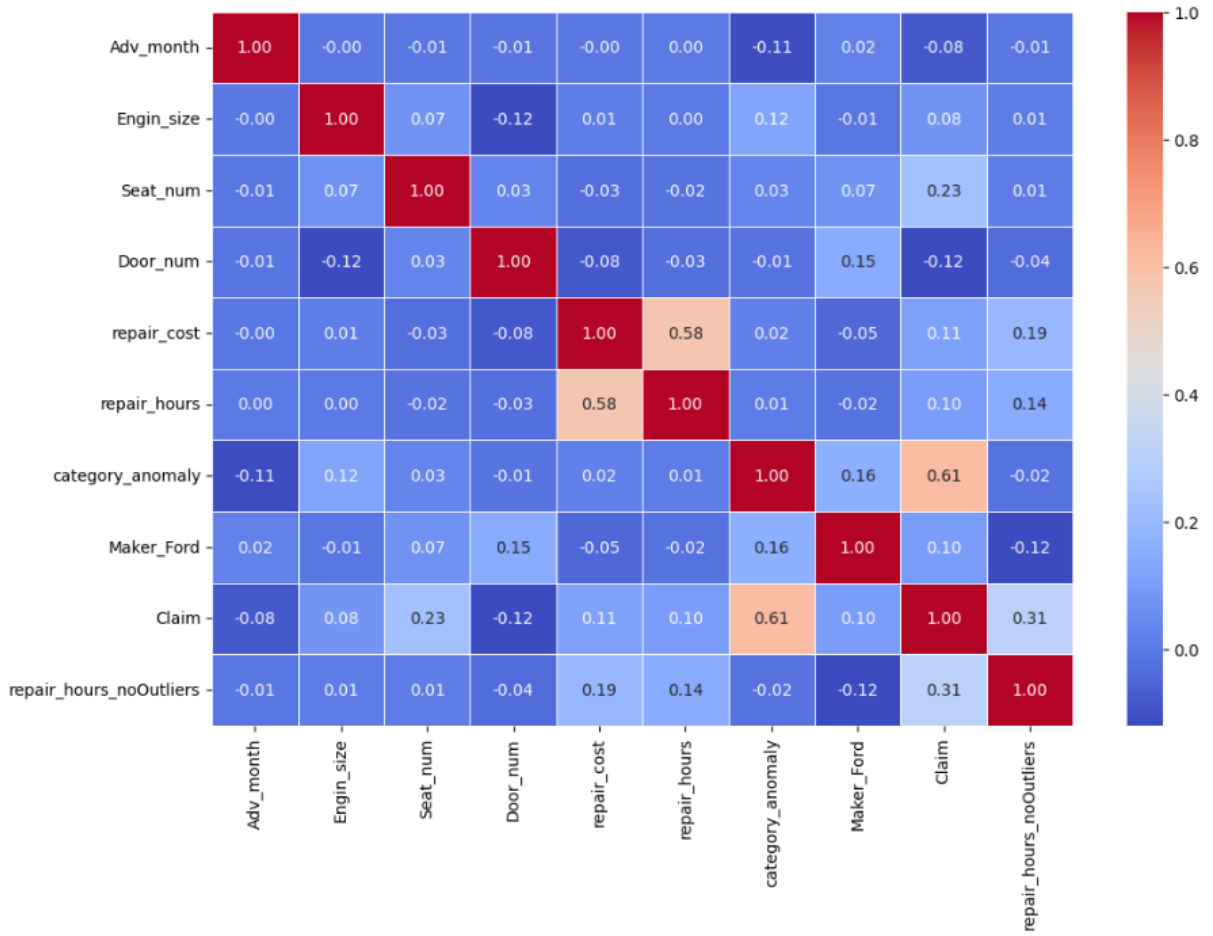
I have dropped the repair_date, breakdown_date and the issue_id from my cleaned data before exporting it to a csv file. The reason is that the dates are usable only if we turn them into numeric values but that would cause a problem for our model since the regression model would treat later dates as higher values. In addition, the Adv_month column can already account for part of the possible significance of these columns and issue_id is non-informative.

Finally, the categorical features are turned into dummy variables in order for the logistic regression to be able to use them.

I have moved on to use forward feature selection. In separate notebooks, I have used Chi-square and backward elimination, but the metrics suggested by forward feature selection have shown more promise.

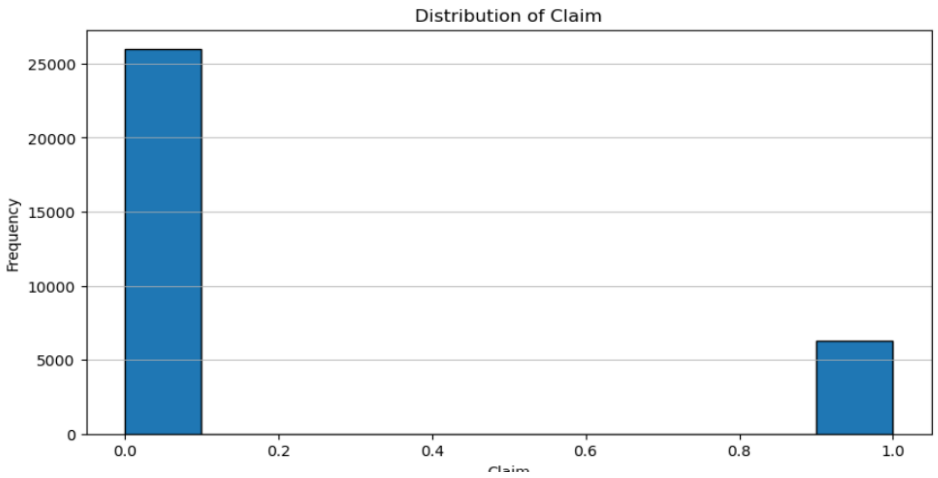
The result of ffs is the following features: 'Adv_month', 'Engin_size', 'Seat_num', 'Door_num', 'repair_cost', 'repair_hours', 'category_anomaly', 'Maker_Ford', 'Model_B-Max', 'Model_Focus', 'Color_Black', 'Color_Blue', 'Color_Gelb', 'Color_Silver', 'Bodytype_Wood'

Here, I will follow by a heatmap to show the correlation between the features that were selected as most significant to one another and to the Claim:

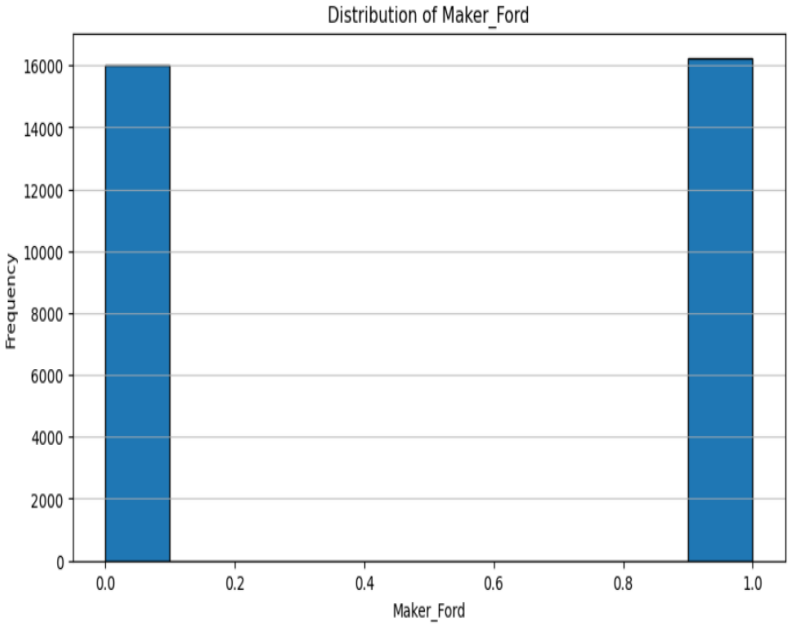


The heatmap shows that Claim is highly correlated with category_anomaly much more than any other feature and removing the anomalies from the repair_hour strengthens the correlation with Claim. A sign of health of our correlation matrix is the strong positive correlation between repair_hours and repair_cost.

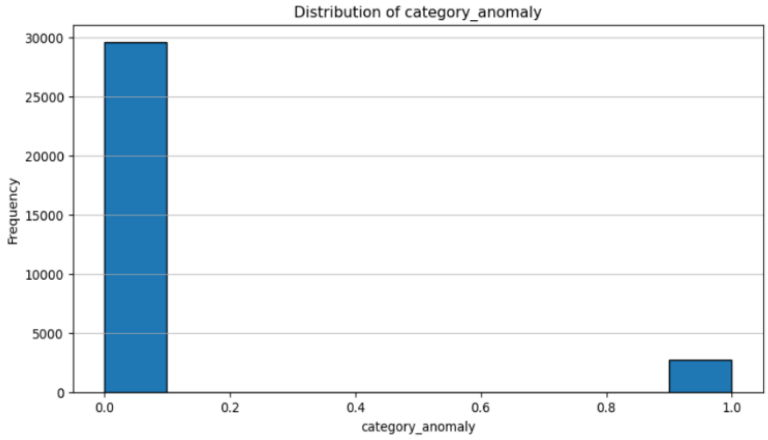
We can show the distribution of all features to have a better understanding:



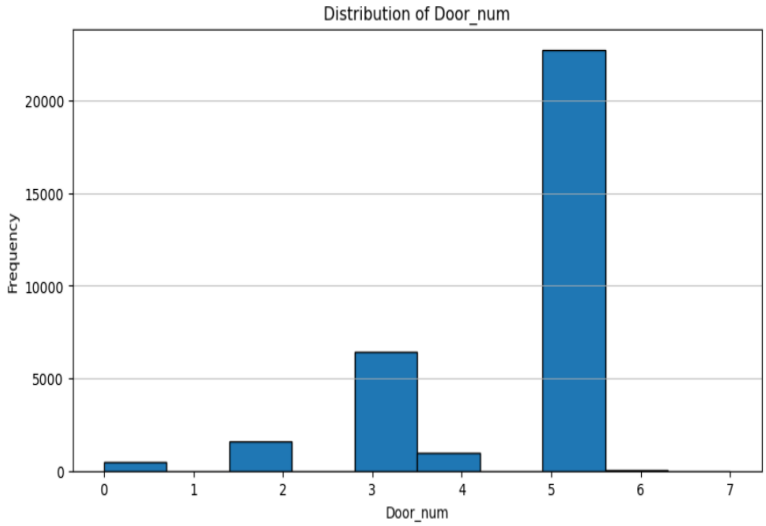
Majority of the values in the Claim are 0 possibly meaning either a claim was not made or was not accepted.



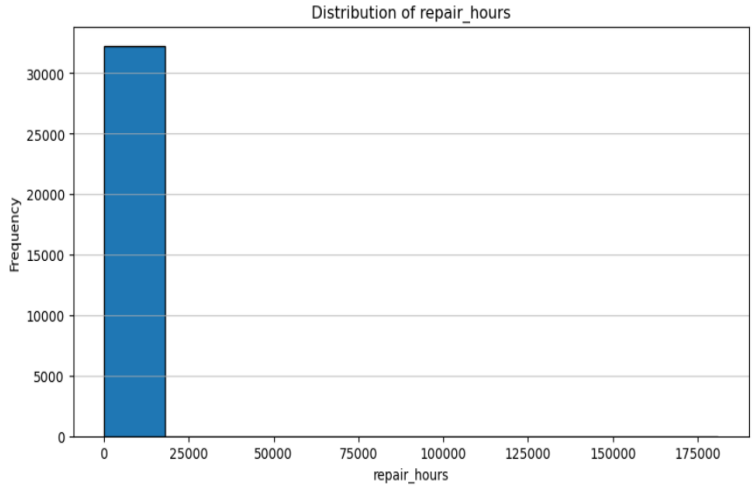
It seems like almost half the cars in the dataset were made by Ford so no wonder this feature is significant.



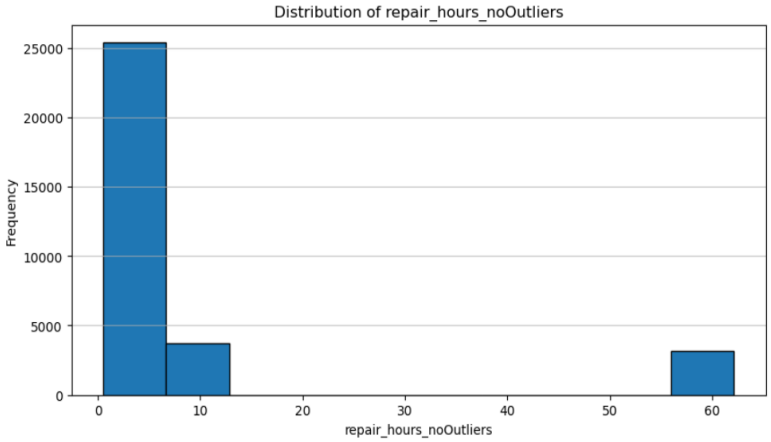
Significant number of the claims registered are of the category anomaly 0.



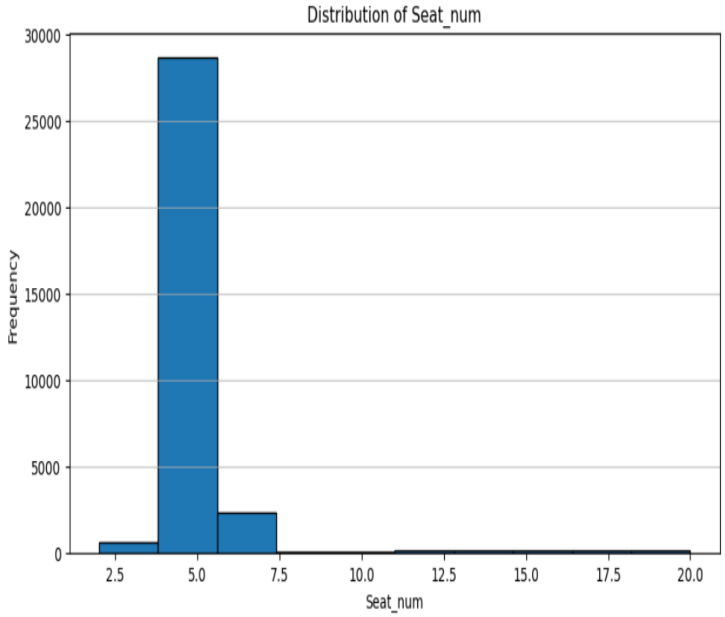
The figure shows that cars with higher number of doors have been registered in the dataset.



Repair_hours with the outliers is not really informative as the X axis is stretched.

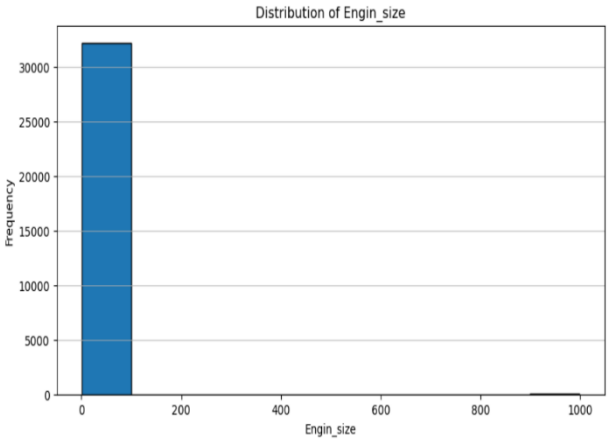


Eliminating the Outliers helps get a better understanding of repair hours.

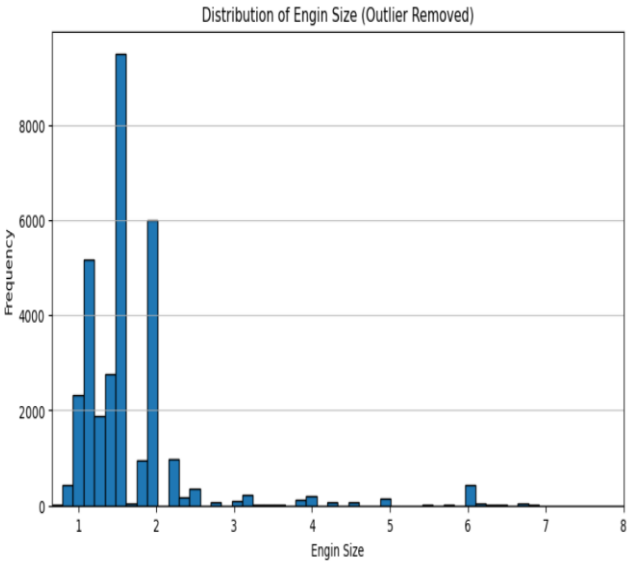


Seat numbers don't have a large variance but seems like few vehicles with unusually high number of doors are registered.

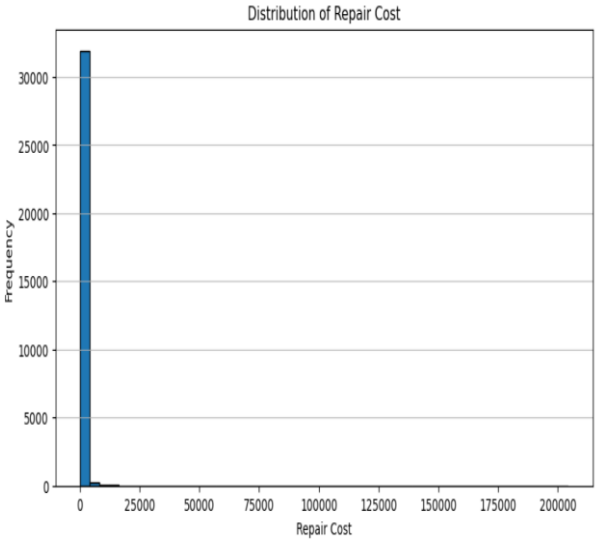
Removing the outliers in Engine size



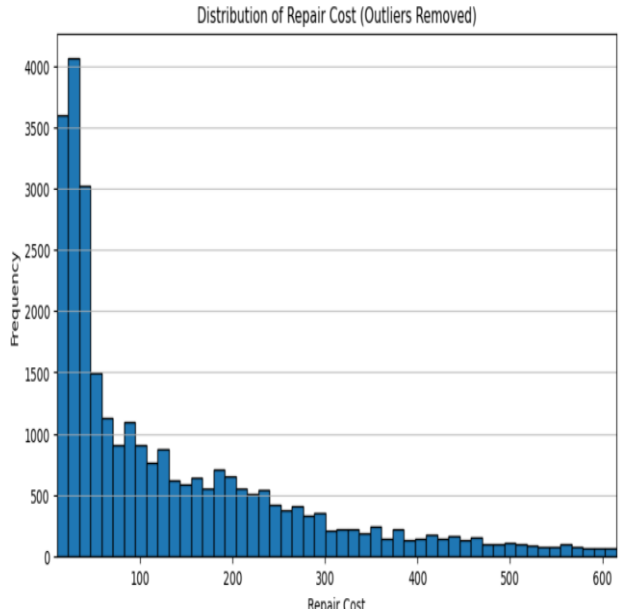
Without removing outliers, the distribution is not informative.



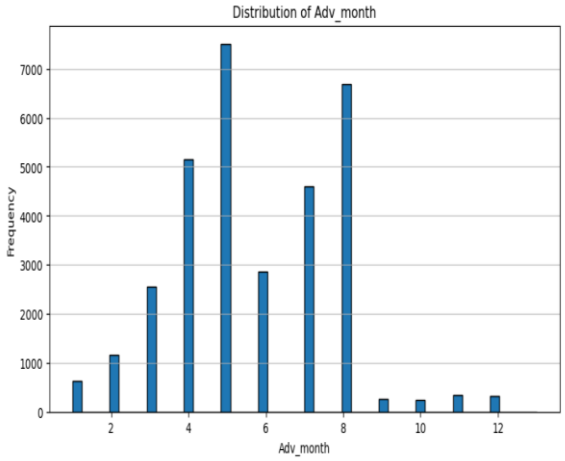
Once the outliers are removed, we can see that Engines with lower capacity form vast majority of



The repair cost has outliers that make the distribution not informative



Removing the outliers shows data is skewed towards lower costs making claims



The first 8 months show higher number of claims being made.

The features selected through application of ffs were used to build models to find the best performing one. The table belows offers the features and metrics received from those models:

Features	Avg Accurac	Std Accurac	Avg Precisio	Std Precisio	Avg Reca	Std Reca	Avg F1-scor	Std F1-scor
Adv_month, Engin_size, Seat_num, Door_num, repair_cost, repair_hours, category_anomaly, Maker_Ford, Model_B-Max, Model_Focus, Color_Black, Color_Blue, Color_Gelb, Color_Silver, Bodytype_Wood	0.945671471	0.002010432	0.988664987	0.005091095	0.7288183	0.01209938	0.839024681	0.008371258
Adv_month, Engin_size, Seat_num, Door_num, repair_cost, repair_hours, category_anomaly, Maker_Ford, Model_B-Max, Model_Focus, Color_Black, Color_Blue	0.945710223	0.002002159	0.988669728	0.00508594	0.72902449	0.01196594	0.839163605	0.008252607
Adv_month, Engin_size, Seat_num, Door_num, repair_cost, repair_hours, category_anomaly, Maker_Ford, Model_B-Max, Model_Focus	0.945748975	0.002034117	0.988922107	0.004764355	0.72901672	0.01228907	0.839248952	0.00852897
Adv_month, Engin_size, Seat_num, Door_num, repair_cost, repair_hours, category_anomaly, Maker_Ford	0.945787735	0.001983048	0.988929484	0.004751467	0.72923224	0.01186197	0.839396663	0.008169878
Adv_month, Engin_size, Seat_num, Door_num, repair_cost, repair_hours, category_anomaly, Maker_Ford (With Robust Scaling)	0.945826487	0.002036067	0.988680617	0.005076306	0.72962482	0.01214412	0.839562832	0.008328399

The highlighted row shows the best performing model which was exported to be used for offering predictions.