Capstone Project

Car Insurance Price Prediction

By Matt Elmajian

Executive Summary

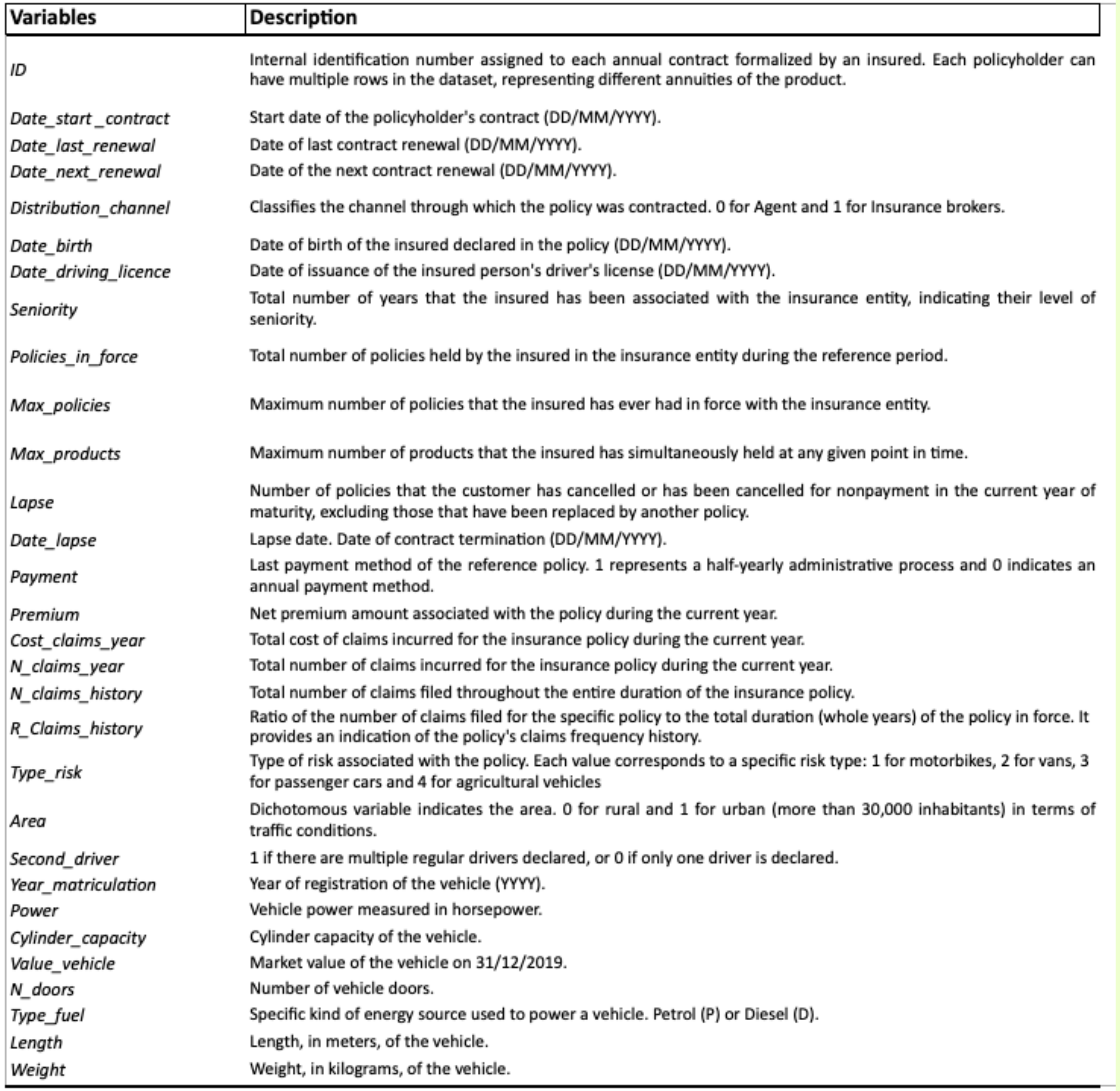
The automotive industry is a massive market that many people interact with everyday. Largely, these interactions are driven by the purchase and sale of cars. With purchasing a new car, there are lots of obvious decisions surrounding the make, model, color, and other features based on customers preferences. However, one of the less contemplated aspects is understanding the subsequent costs that come with your car. The price is one thing, but an often overlooked cost inherent with owning a car is the insurance associated. I intend to build out a model that can predict insurance costs based on a variety of features within a car. This will assist any customer looking to understand their all in costs for a car, ultimately allowing them to make a more informed decision and avoid unexpected costs.

The dataset I will be using to build this model is one from a motor vehicle insurance portfolio, accessible through Mendeley Data with over 100k entries. It includes features such as Insurance Premium, Number and Cost of claims, renewal date, value of the vehicle, power, and other car features/ policy details.

Dataset Features & Strategy

Target Variable: Premium

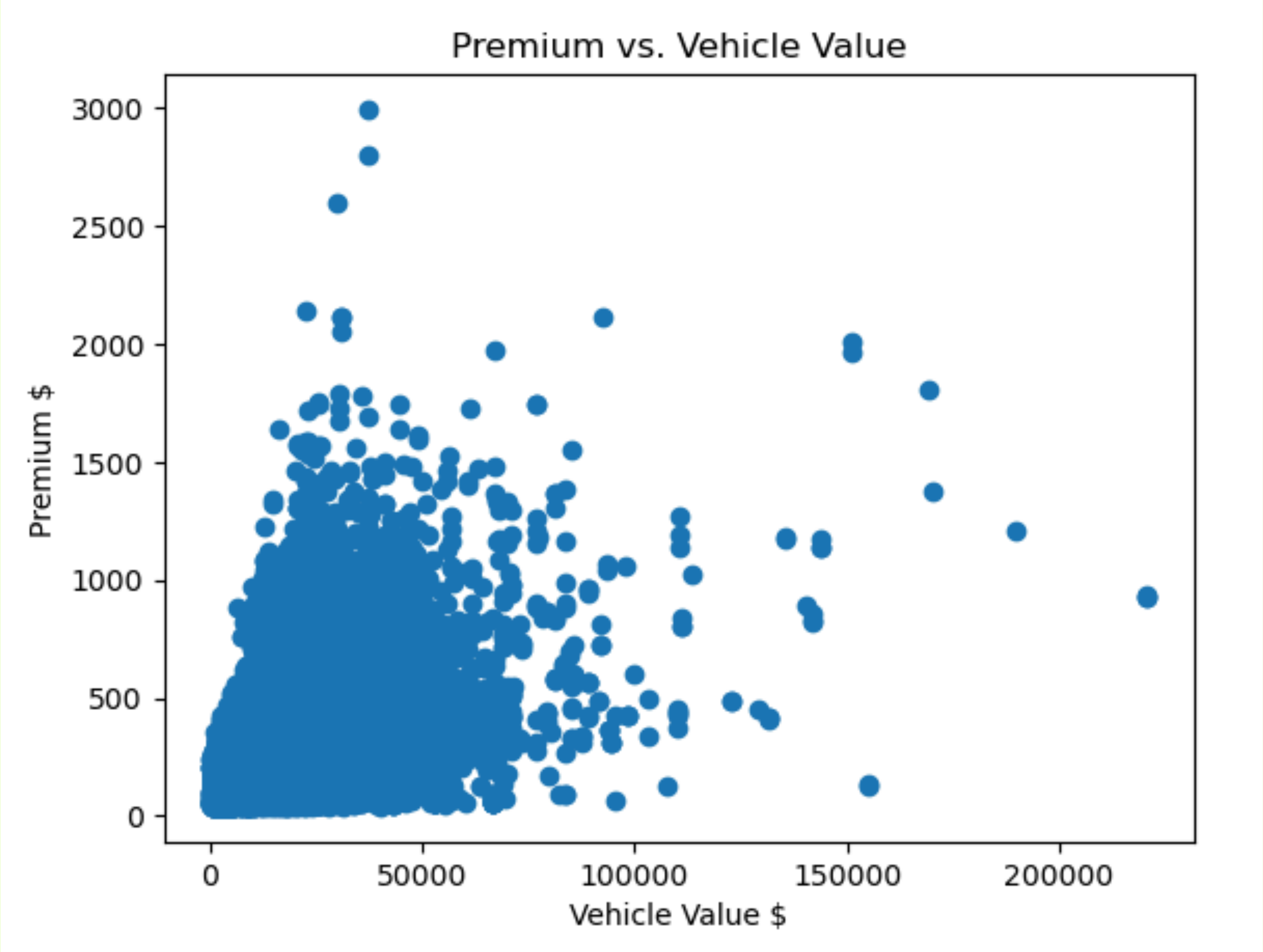
My strategy will involve building out multiple models and comparing results to try and find the best fit. First, I intend to search for correlations between features and the target variable to gain an understanding of how to best fit a linear regression model. With this I will build out a few linear regression models to test predictive power. Secondarily, I will run the dataset through a Random Forest model and compare the results.

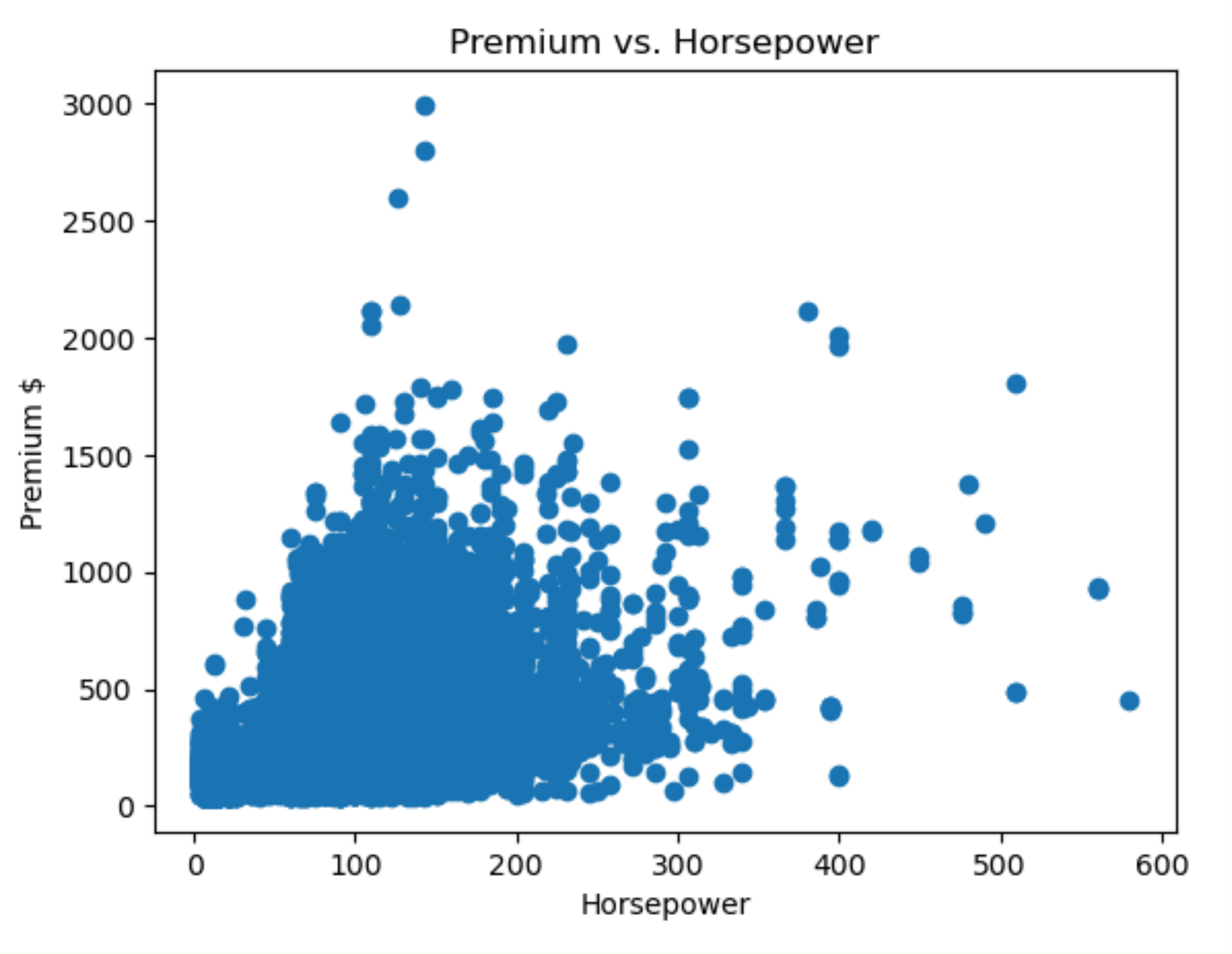


Variables	Description
ID	Internal identification number assigned to each annual contract formalized by an insured. Each policyholder can have multiple rows in the dataset, representing different annuities of the product.
Date_start_contract	Start date of the policyholder's contract (DD/MM/YYYY).
Date_last_renewal	Date of last contract renewal (DD/MM/YYYY).
Date_next_renewal	Date of the next contract renewal (DD/MM/YYYY).
Distribution_channel	Classifies the channel through which the policy was contracted. 0 for Agent and 1 for Insurance brokers.
Date_birth	Date of birth of the insured declared in the policy (DD/MM/YYYY).
Date_driving_licence	Date of issuance of the insured person's driver's license (DD/MM/YYYY).
Seniority	Total number of years that the insured has been associated with the insurance entity, indicating their level of seniority.
Policies_in_force	Total number of policies held by the insured in the insurance entity during the reference period.
Max_policies	Maximum number of policies that the insured has ever had in force with the insurance entity.
Max_products	Maximum number of products that the insured has simultaneously held at any given point in time.
Lapse	Number of policies that the customer has cancelled or has been cancelled for nonpayment in the current year of maturity, excluding those that have been replaced by another policy.
Date_lapse	Lapse date. Date of contract termination (DD/MM/YYYY).
Payment	Last payment method of the reference policy. 1 represents a half-yearly administrative process and 0 indicates an annual payment method.
Premium	Net premium amount associated with the policy during the current year.
Cost_claims_year	Total cost of claims incurred for the insurance policy during the current year.
N_claims_year	Total number of claims incurred for the insurance policy during the current year.
N_claims_history	Total number of claims filed throughout the entire duration of the insurance policy.
R_Claims_history	Ratio of the number of claims filed for the specific policy to the total duration (whole years) of the policy in force. It provides an indication of the policy's claims frequency history.
Type_risk	Type of risk associated with the policy. Each value corresponds to a specific risk type: 1 for motorbikes, 2 for vans, 3 for passenger cars and 4 for agricultural vehicles
Area	Dichotomous variable indicates the area. 0 for rural and 1 for urban (more than 30,000 inhabitants) in terms of traffic conditions.
Second_driver	1 if there are multiple regular drivers declared, or 0 if only one driver is declared.
Year_matriculation	Year of registration of the vehicle (YYYY).
Power	Vehicle power measured in horsepower.
Cylinder_capacity	Cylinder capacity of the vehicle.
Value_vehicle	Market value of the vehicle on 31/12/2019.
N_doors	Number of vehicle doors.
Type_fuel	Specific kind of energy source used to power a vehicle. Petrol (P) or Diesel (D).
Length	Length, in meters, of the vehicle.
Weight	Weight, in kilograms, of the vehicle.

Linear Regression Model

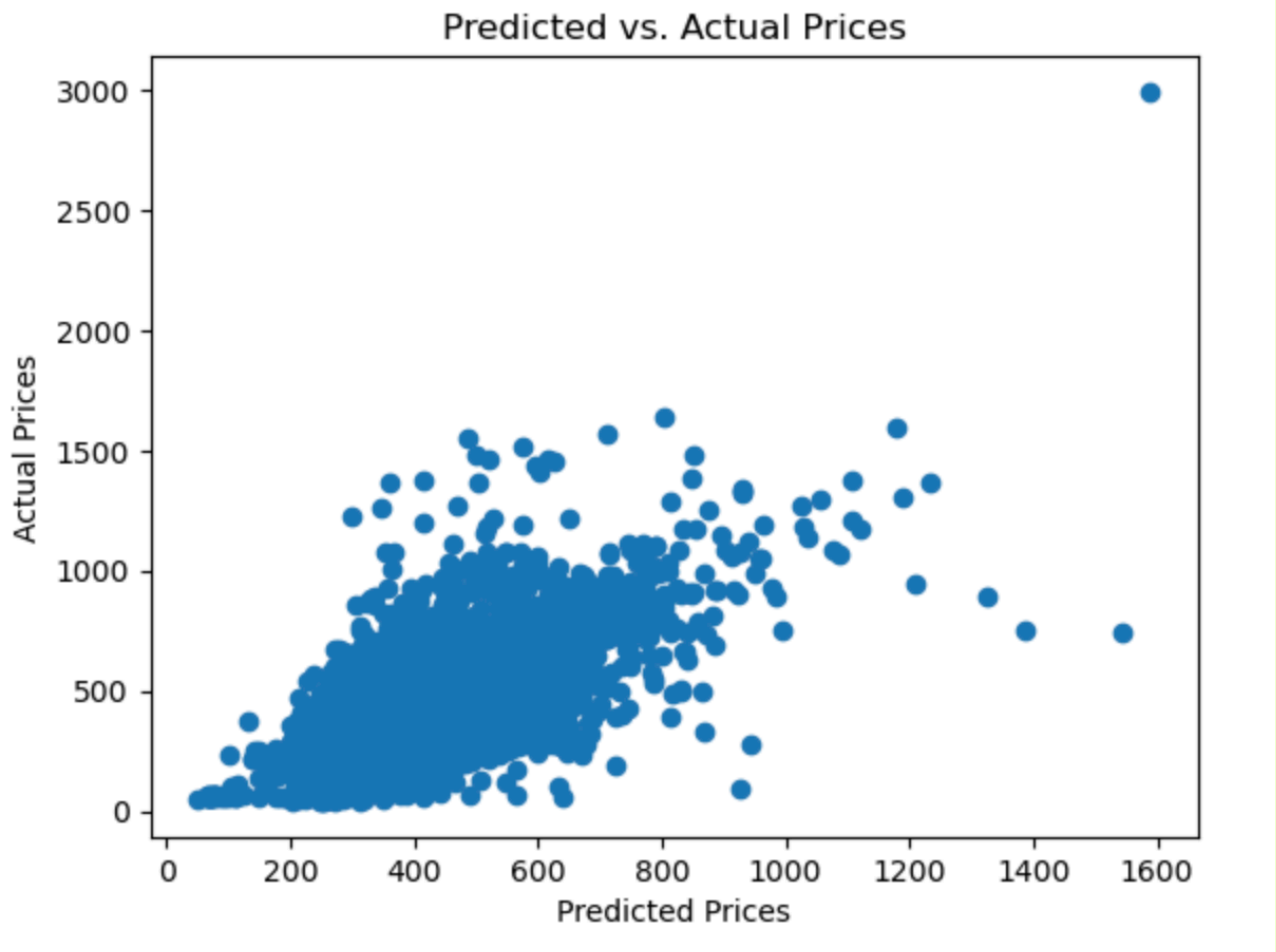
- Unfortunately, there were no major correlations between Premium and other features variables.
- The two largest correlations were with the value of the vehicle and the horsepower it produces. I still completed linear regression models, however as you can see from the scatterplots, the correlation was relatively minimal.
    - Models ran: OLS (statsmodel), LinearRegression (sklearn), Ridge Regression (sklearn), Lasso (sklearn)
    - Model predictive power of approximately 25%





Random Forest Model

- The Random Forest model showed a substantial increase in predictive capabilities. After seeing some success through initial testing with the Random Forest algorithm, I ran a few models to adjust for hyper parameters such as max depth and number of estimators.
  - Final hyper-parameters: n_estimators: 1000, max_depth = 25
  - Model r2_score = 57%
  - Model RMSE = $89.53
- Although the predictive power and error metrics are not nearly where I would like them to be, I was encouraged to see such an increase from running the Random Forest model vs Linear Regression.



Conclusion / Next Steps

- In conclusion, I was pleased to find an increase in model performance when switching from a Linear Regression algorithm to a Random Forest. However, I believe there is more work to be done that could further increase performance. Below are my recommendations:
    - Run an unsupervised clustering algorithm on the dataset to further draw insights within the data in regards to grouping. Potential ability to add another feature based on clustered groups.
    - Further subset the data and create separate models based on certain criteria (type of vehicle, number of doors, length/width combinations)
- Next steps would be to create a model/models that provide a much greater ability to predict pricing based on the recommendations above. With this information, customers can evaluate insurance pricing for cars prior to purchasing and ultimately make a more informed decision.