Capstone Project

Car Insurance Price Prediction Model

By Matt Elmajian

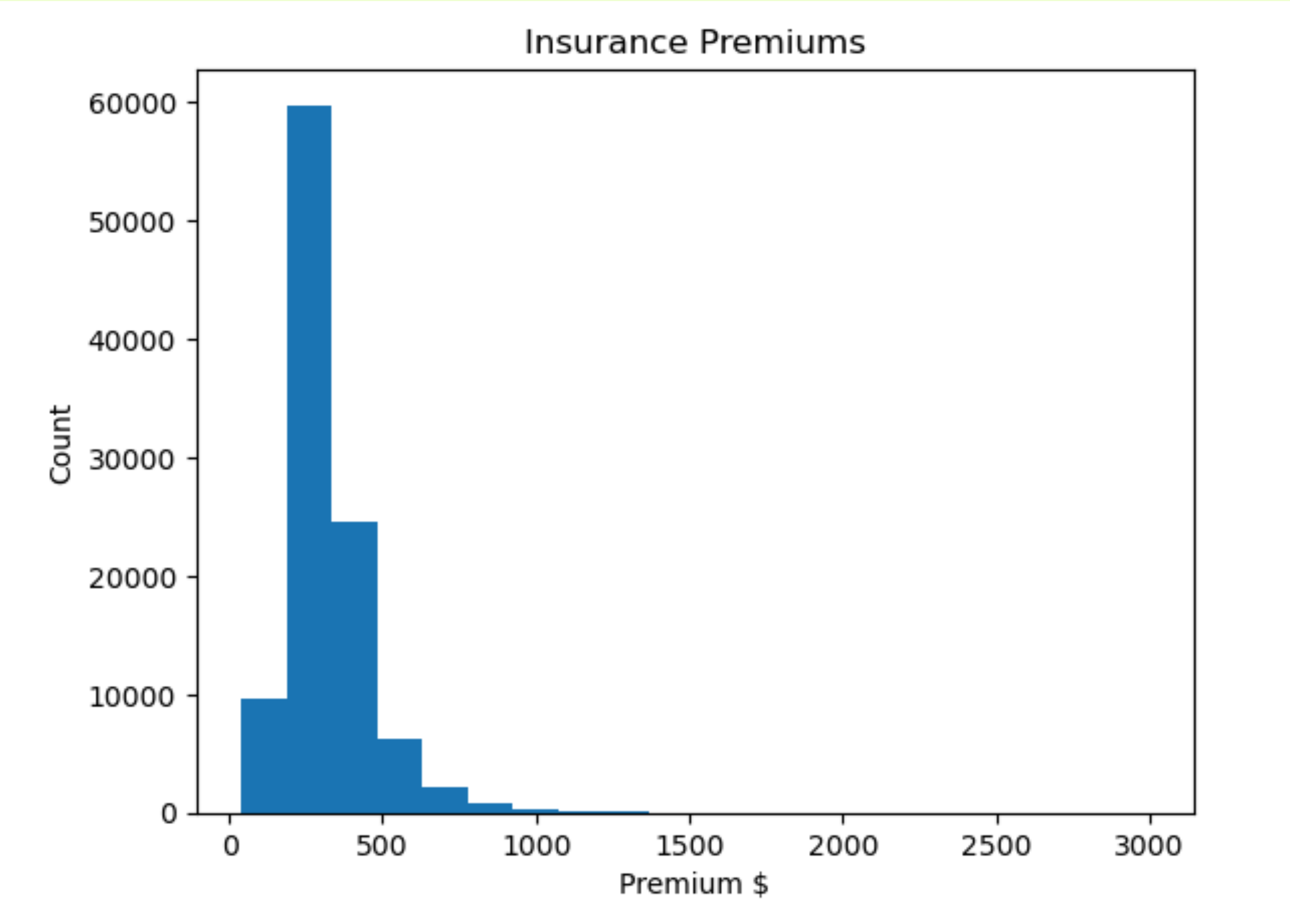
Dataset Features & Strategy

- Target Variable: Premium
- Features Added:
  - Age of Customer (Date Last Renewal - DoB)
- Key Variables:
  - Value of Vehicle
  - Year_matriculation
  - Age
- Dropped during modeling:
  - Date ranges

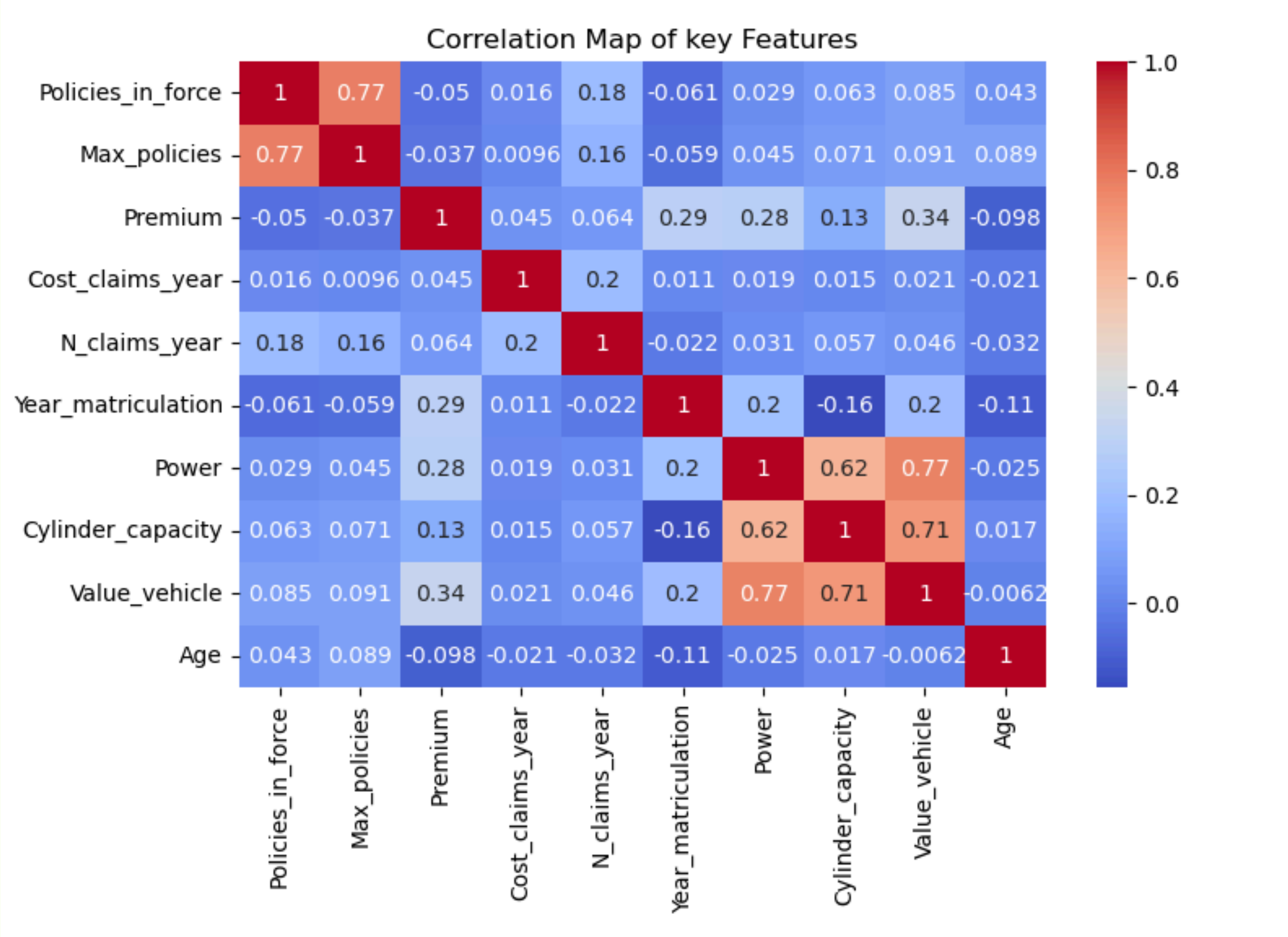
Variables	Description
ID	Internal identification number assigned to each annual contract formalized by an insured. Each policyholder can have multiple rows in the dataset, representing different annuities of the product.
Date_start_contract	Start date of the policyholder's contract (DD/MM/YYYY).
Date_last_renewal	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customer has cancelled or has been cancelled for nonpayment in the current year of maturity, excluding those that have been replaced by another policy.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.

Target Variable & Correlations

- Target Variable: Premium
- Skewed Right Distribution
  - Mean: $320
  - Median: $300

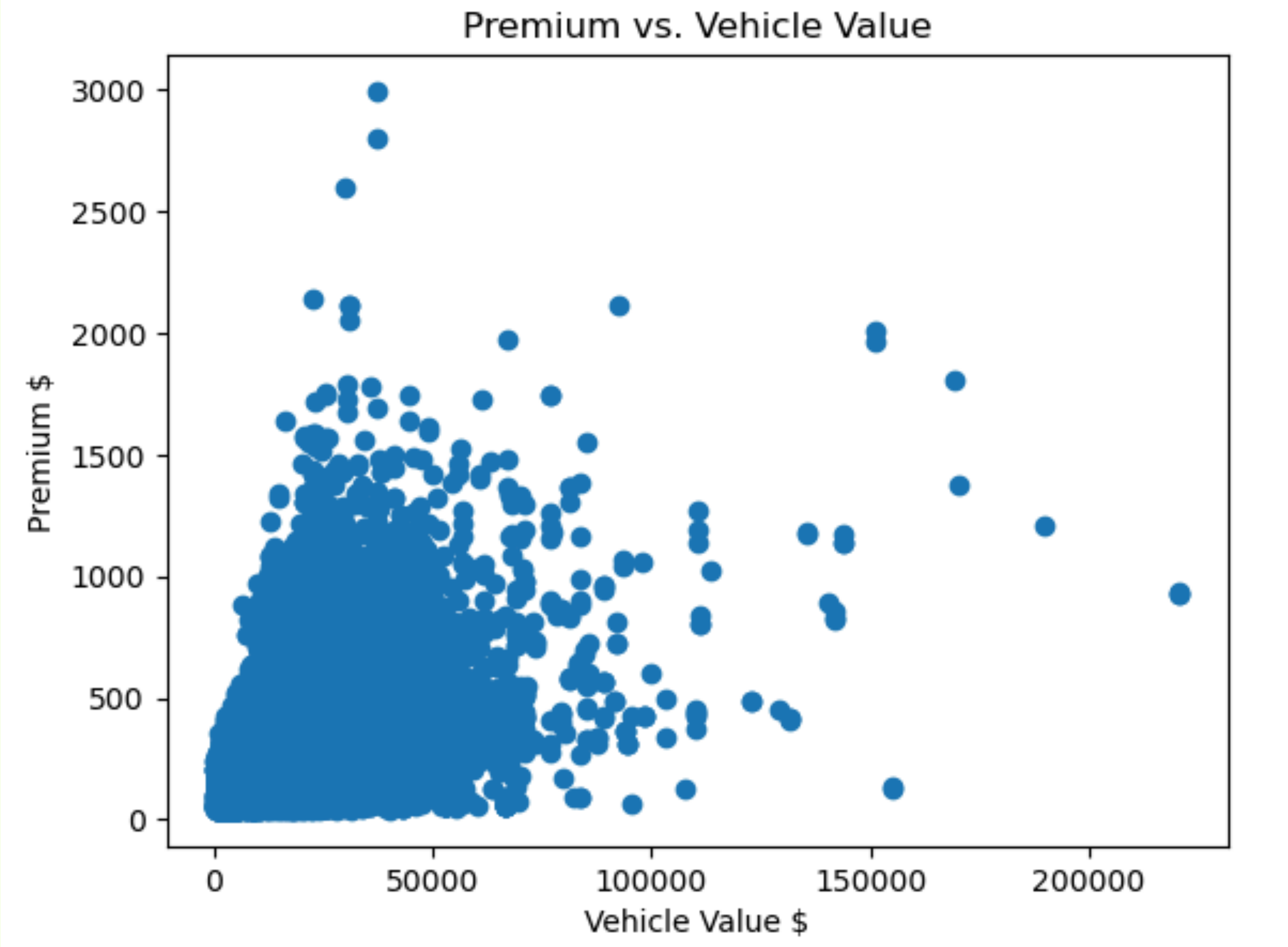


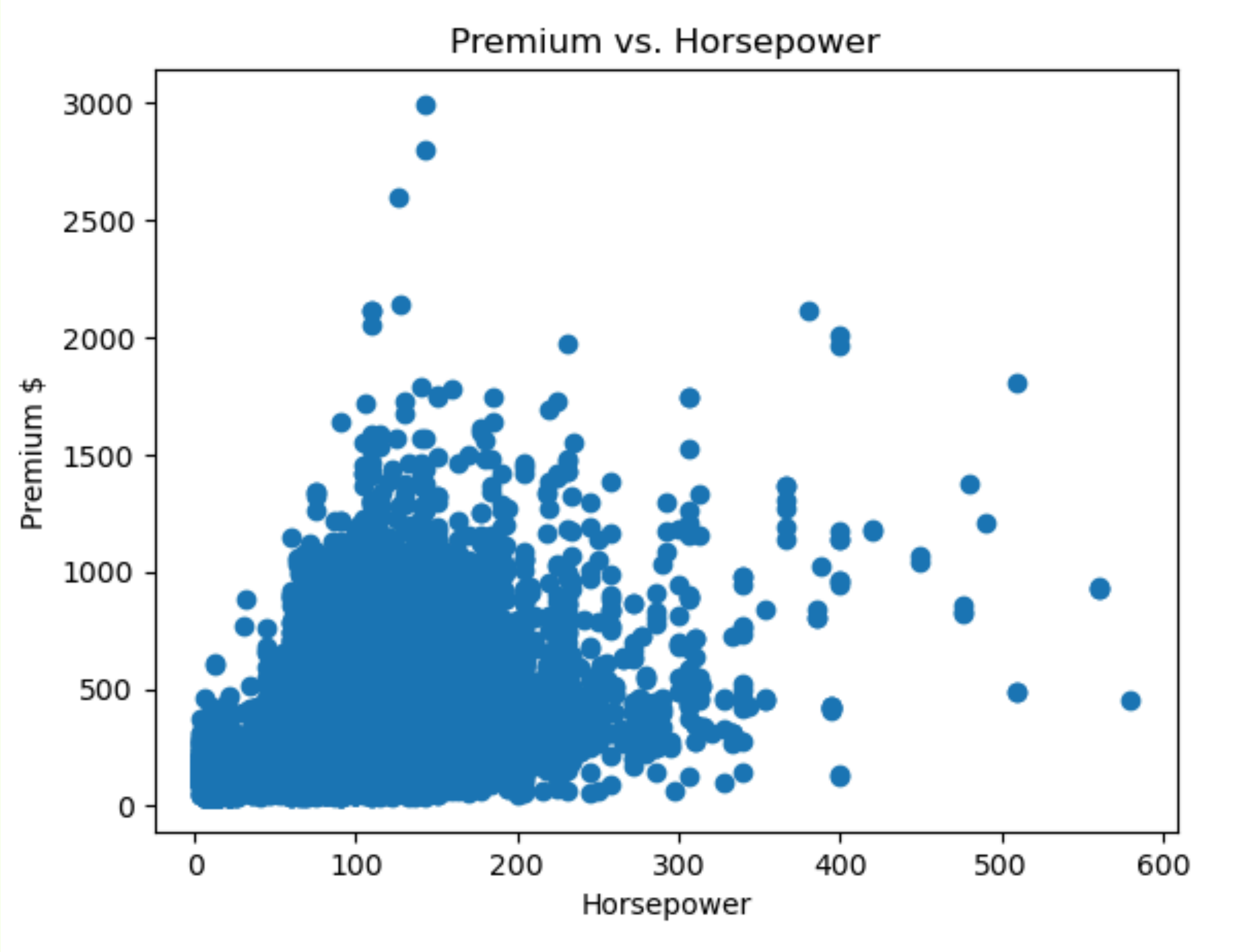
- Key Correlations: Vehicle Value, Year Matriculation, Power



EDA

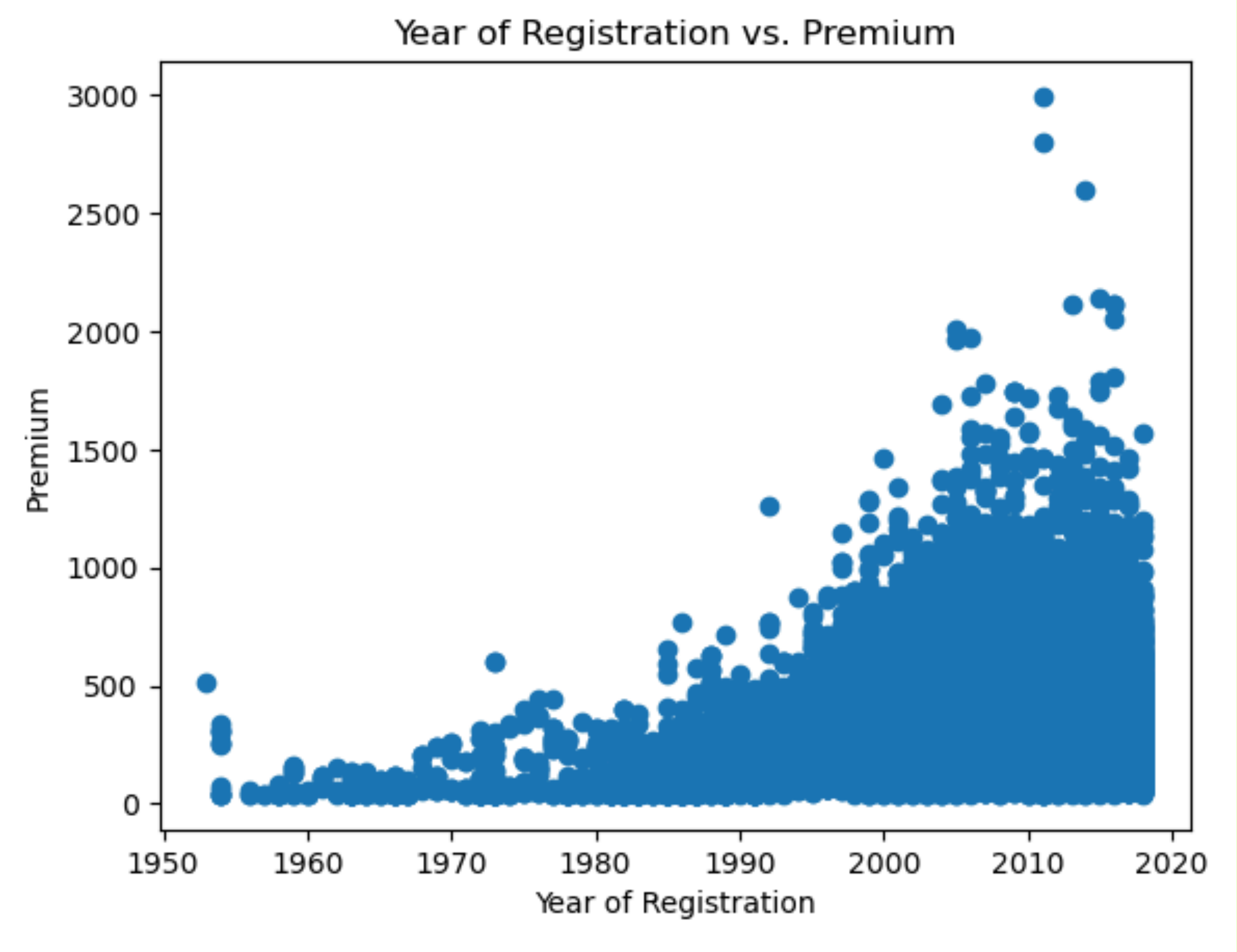
- Correlated features with Premium:
  - Vehicle Value vs. Premium - .34
  - Horsepower vs. Premium - .28
- Relatively weak correlations
- Single variable regression analysis returned weak R2 scores as well
  - Vehicle Value - .163
  - Horsepower - .166

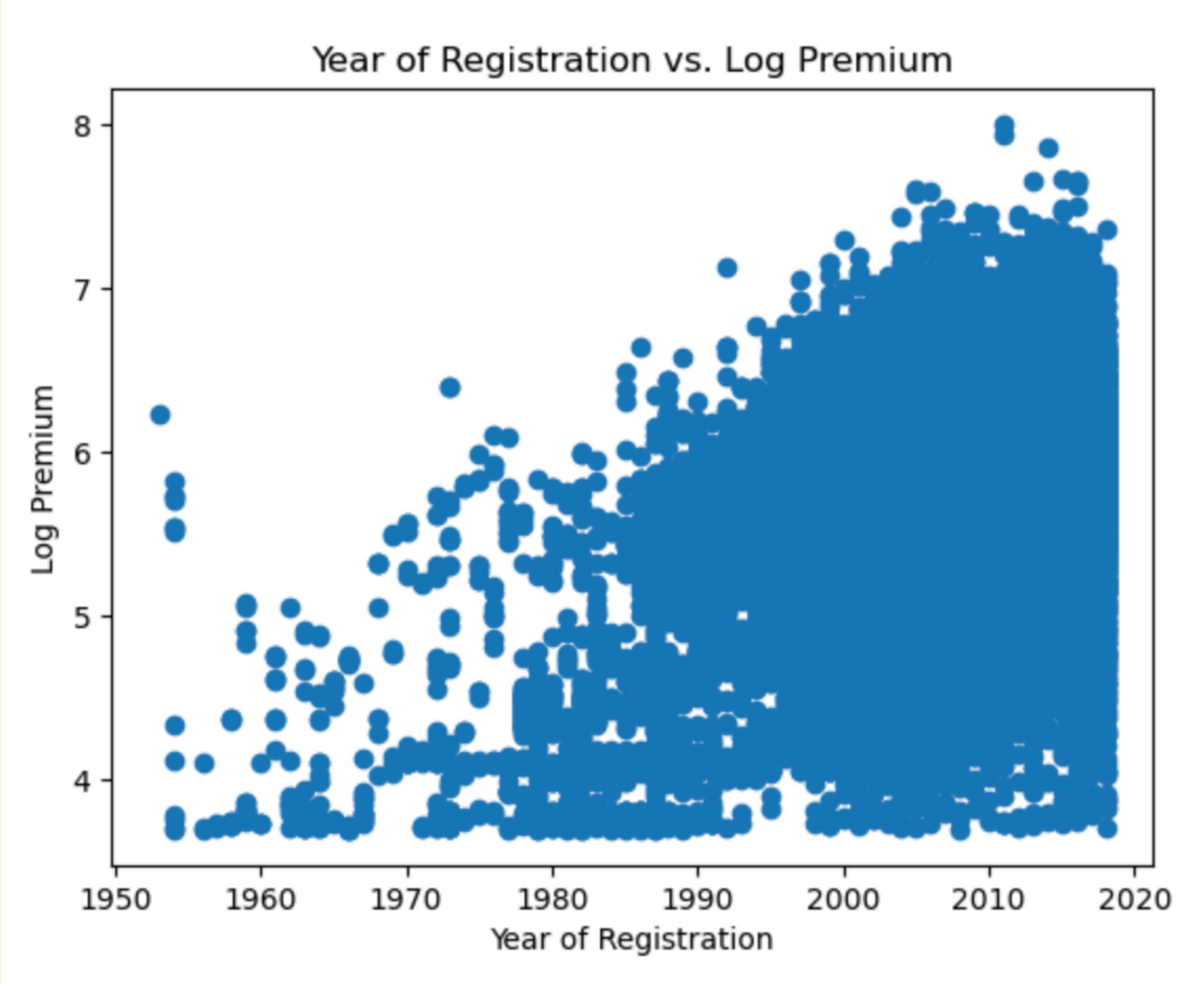




EDA - continued

- Year of Registration also showed relatively high correlation with Premium
  - Initial relationship appeared to show more exponentially
- Scatterplot below showed no large correlation increase when taking the log of premium and comparing
- Due to the large amount of data, all correlations tested had p-values under .05 indicating statistical significance





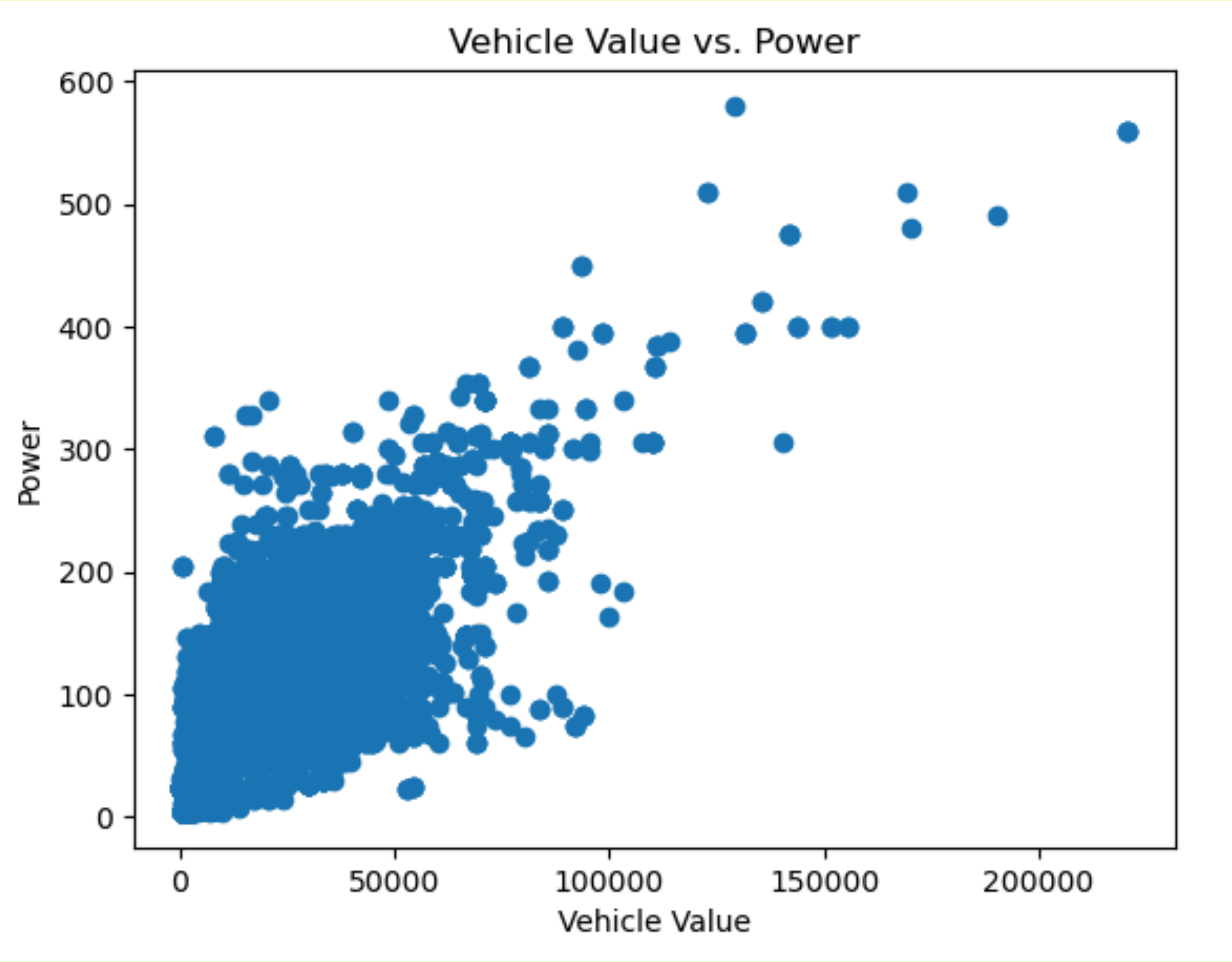
Feature Importance

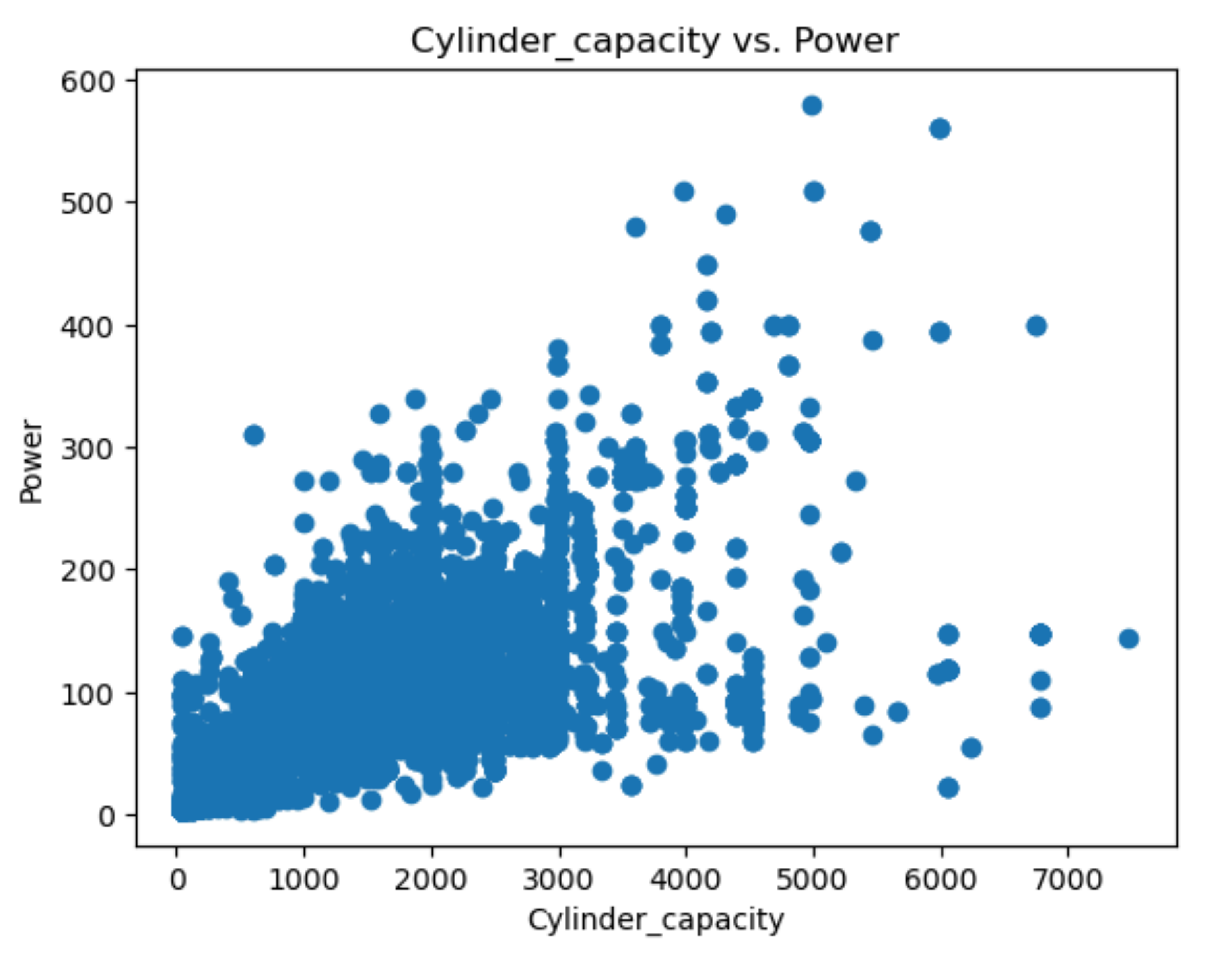
- Features showing highest correlation with Premium (correlation coefficient)
  - Vehicle Value - .34
  - Power - .28
  - Year of Matriculation - .29
- Feature importance based on Random Forest Model
  - Vehicle Value
  - Year of Matriculation
  - Age
- Similar features important with exception of Age

	Feature	Importance
17	Value_vehicle	0.154670
14	Year_matriculation	0.128611
22	Age	0.121584
0	ID	0.085314
11	R_Claims_history	0.061122
21	Weight	0.054609
20	Length	0.053521
15	Power	0.048825
16	Cylinder_capacity	0.047924
2	Seniority	0.042417
10	N_claims_history	0.038449
7	Payment	0.035794
8	Cost_claims_year	0.021196
3	Policies_in_force	0.020461
13	Second_driver	0.017971
4	Max_policies	0.015217
6	Lapse	0.012341
1	Distribution_channel	0.010572
12	Area	0.010332
9	N_claims_year	0.006840
19	Type_fuel	0.004965
18	N_doors	0.003895
5	Max_products	0.003372

Linear Regression Models

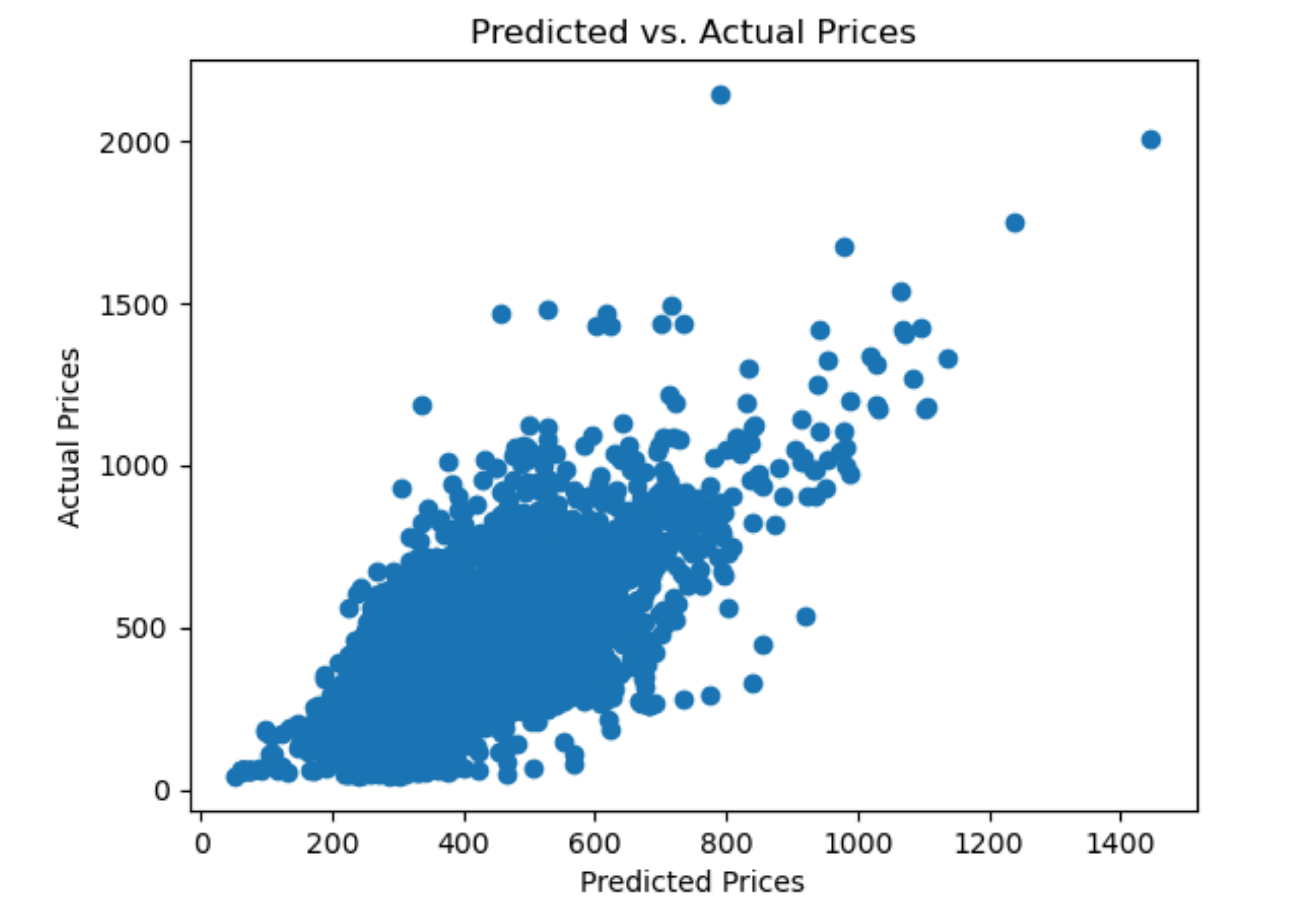
- Higher correlated features were still relatively weak
- Models ran with all features:
  - OLS - R-squared: .255
  - Lasso - R-squared: .2552
  - Ridge - R-squared: .2553
- Models ran with reduced features to account for collinearity & overfitting
  - OLS (dropping collinear features) - R-squared: .252
  - OLS (using top three features of importance) - R-squared: .174
  - Lasso (dropping collinear features - R-squared: .252





Tree Based Models

- Due to the lack of correlation between features and the target variable, Tree Based Models performed better
- Models tested:
    - Random Forest Regressor
        - R-squared score: .5957
        - RMSE: $85.00
    - Gradient Boost Regressor
        - R-squared score: .5421
        - RMSE: $90.46
- Final hyper-parameters: n_estimators: 1000, max_depth = 25



Results

- Linear Regression models did not perform very well
  - None were able to surpass .25 R-squared scores regardless of features used
- Tree based models showed much more promise
  - Results did not change very much with hyper parameter tuning
  - Strongest model - Random Forest Regressor

	Model	Hyperparameters	R2_score	RMSE
1	OLS - All Features	Default Hyperparameters	0.255	N/A
2	OLS - All Features (Scaled)	Default Hyperparameters	0.255	N/A
3	OLS - Top 3 Features	Default Hyperparameters	0.174	N/A
4	Lasso Regression - All Features	aplha = .205	0.255	119.29
5	Random Forest Regressor - All Features	n_estimators: 1000, max_depth: 25	0.595	85.0
6	Gradient Boost Regressor - All Features	n_estimators: 189, learning_rate: .078, max_depth: 9	0.542	90.46

Conclusion / Next Steps

- Model not ready for deployment yet
- Target R2 Score: > .85
- Target RMSE: < $50
- Potential next steps:
    - Run clustering algorithm on data for further feature engineering
    - Consider running SVM model
    - Further sub-set data into groups and run separate models