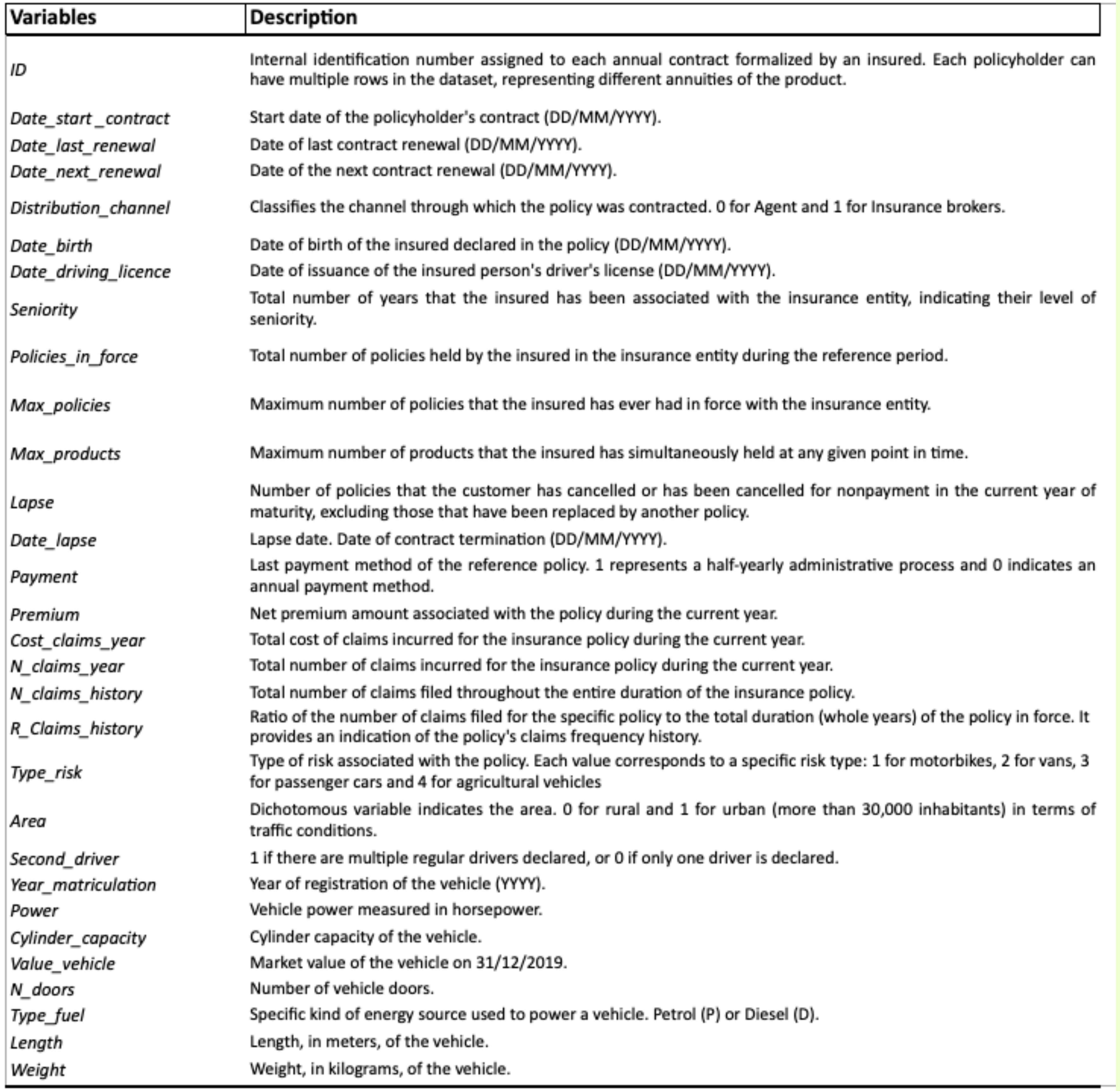
Capstone Project

Car Insurance Price Prediction

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

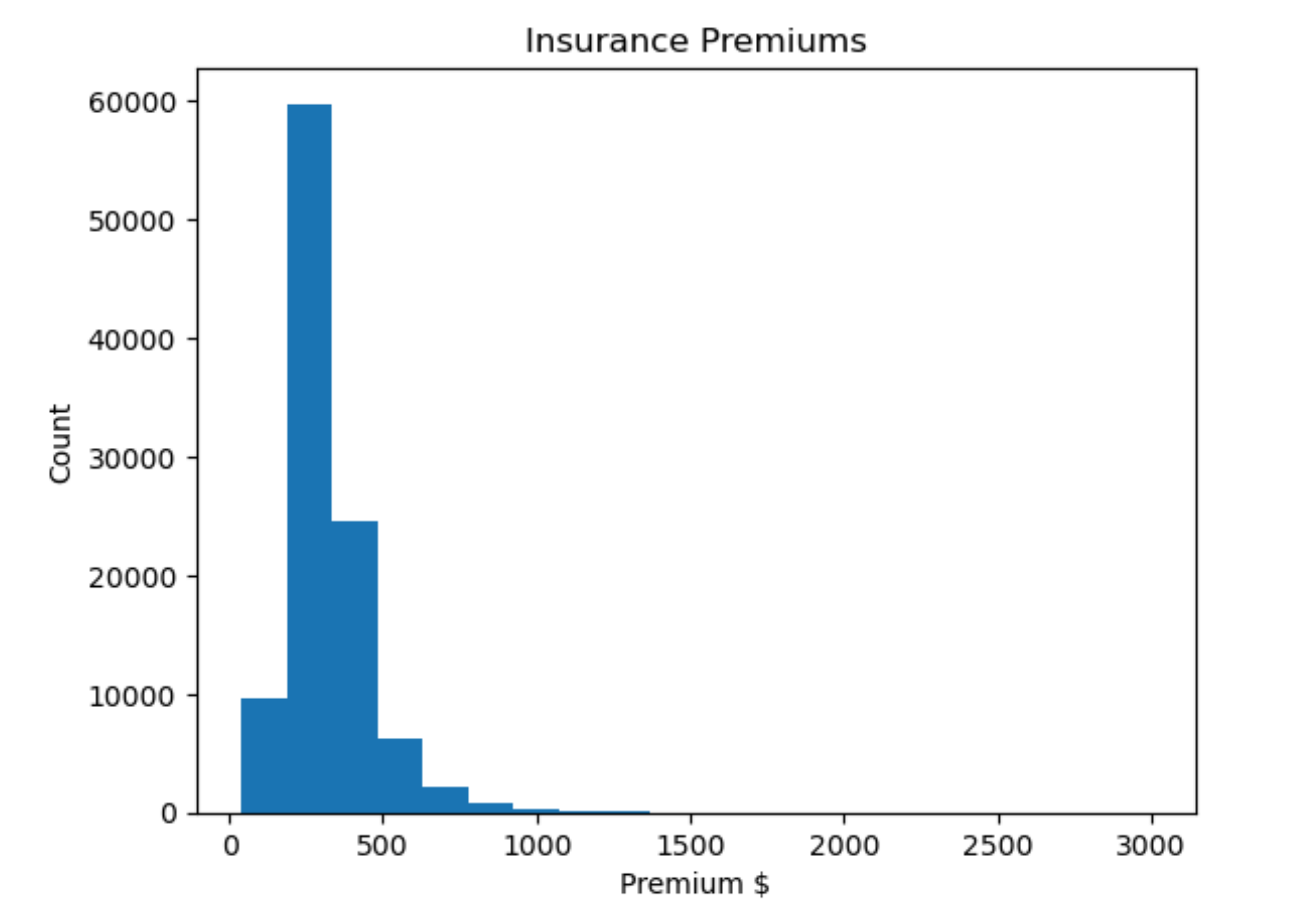


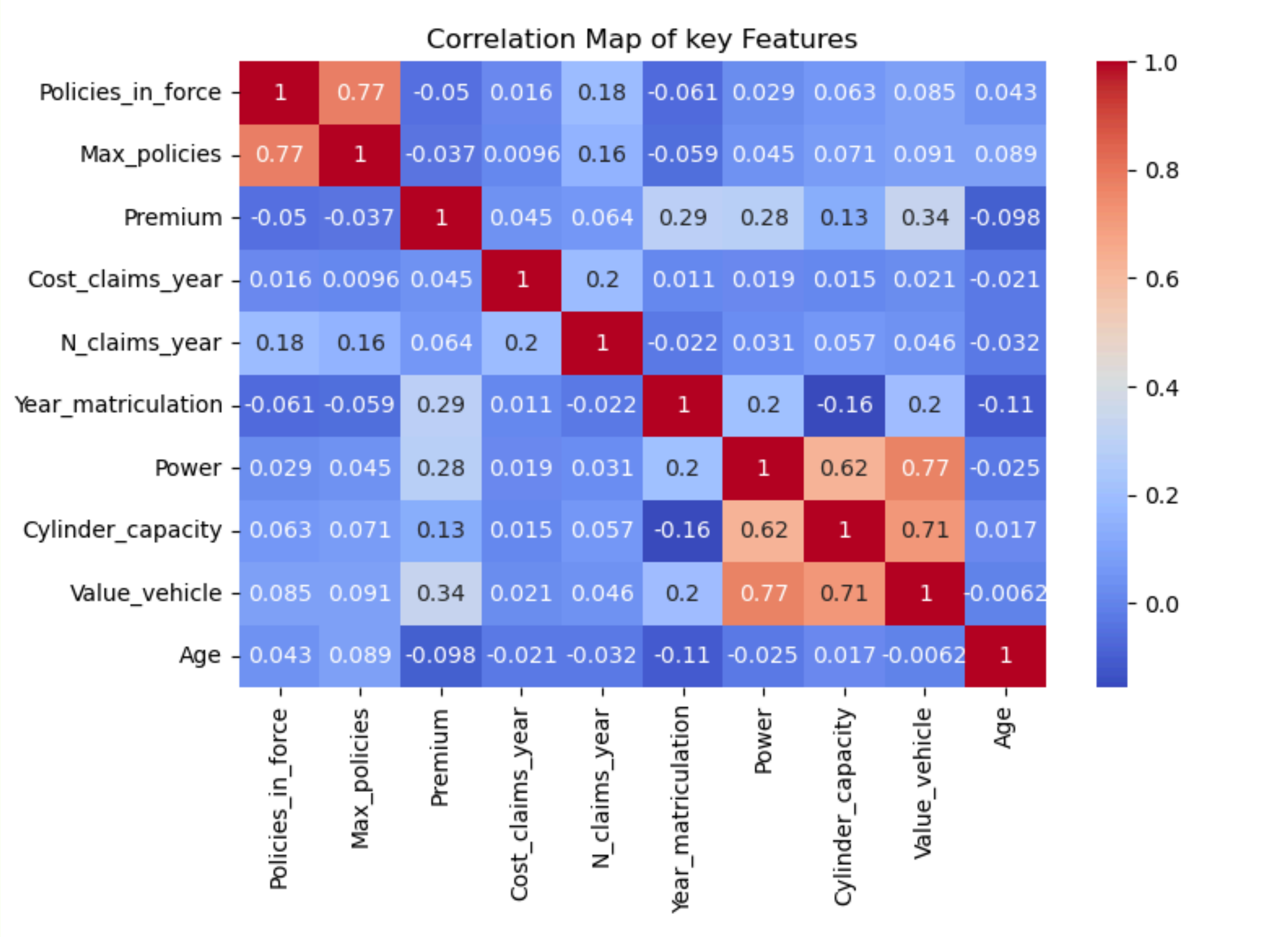
Variables	Description
ID	Internal identification number assigned to each annual contract formalized by an insured. Each policyholder can have multiple rows in the dataset, representing different annuities of the product.
Date_start_contract	Start date of the policyholder's contract (DD/MM/YYYY).
Date_last_renewal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customer has cancelled or has been cancelled for nonpayment in the current year of maturity, excluding those that have been replaced by another policy.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



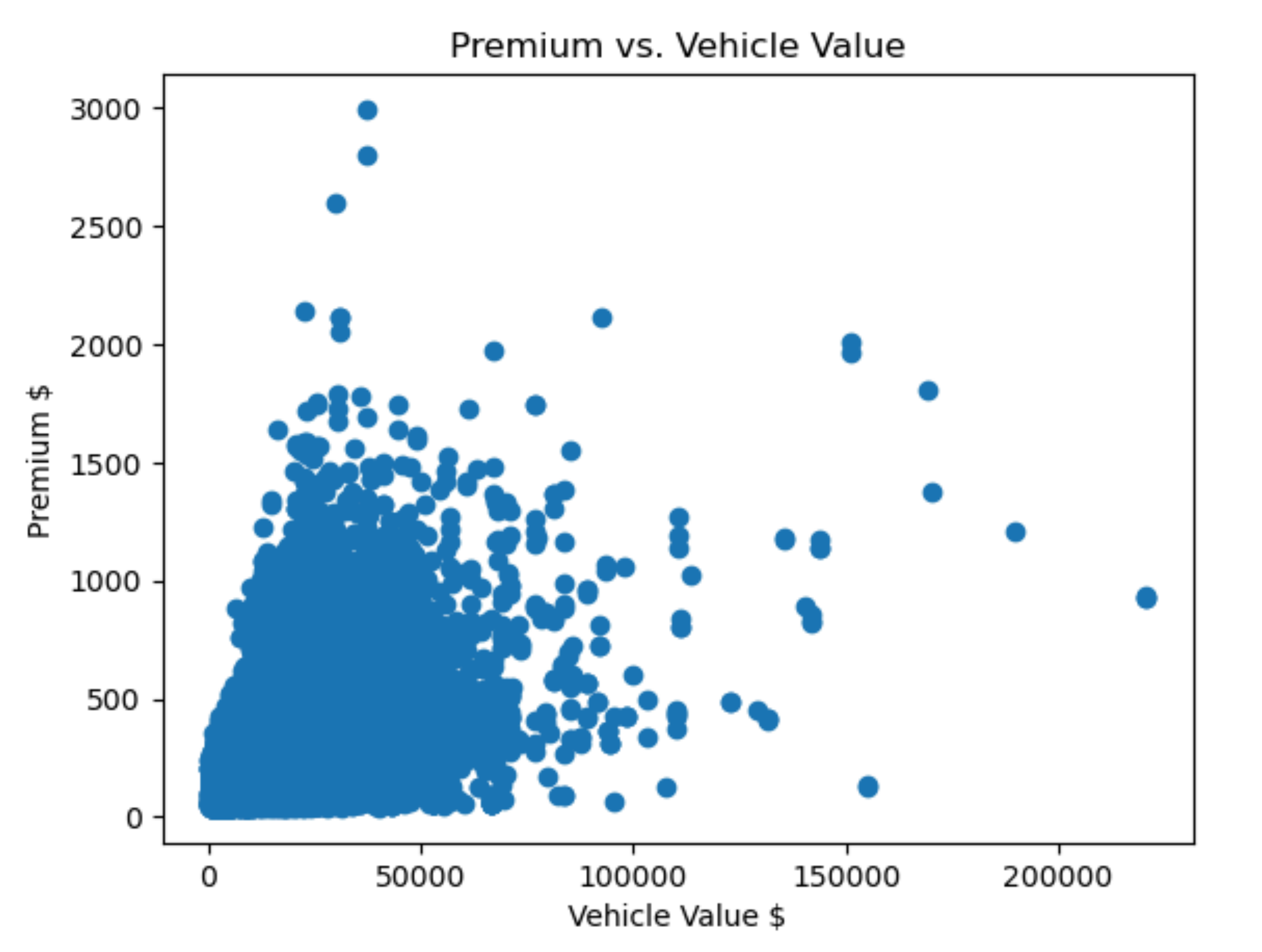

Correlation Map of key Features

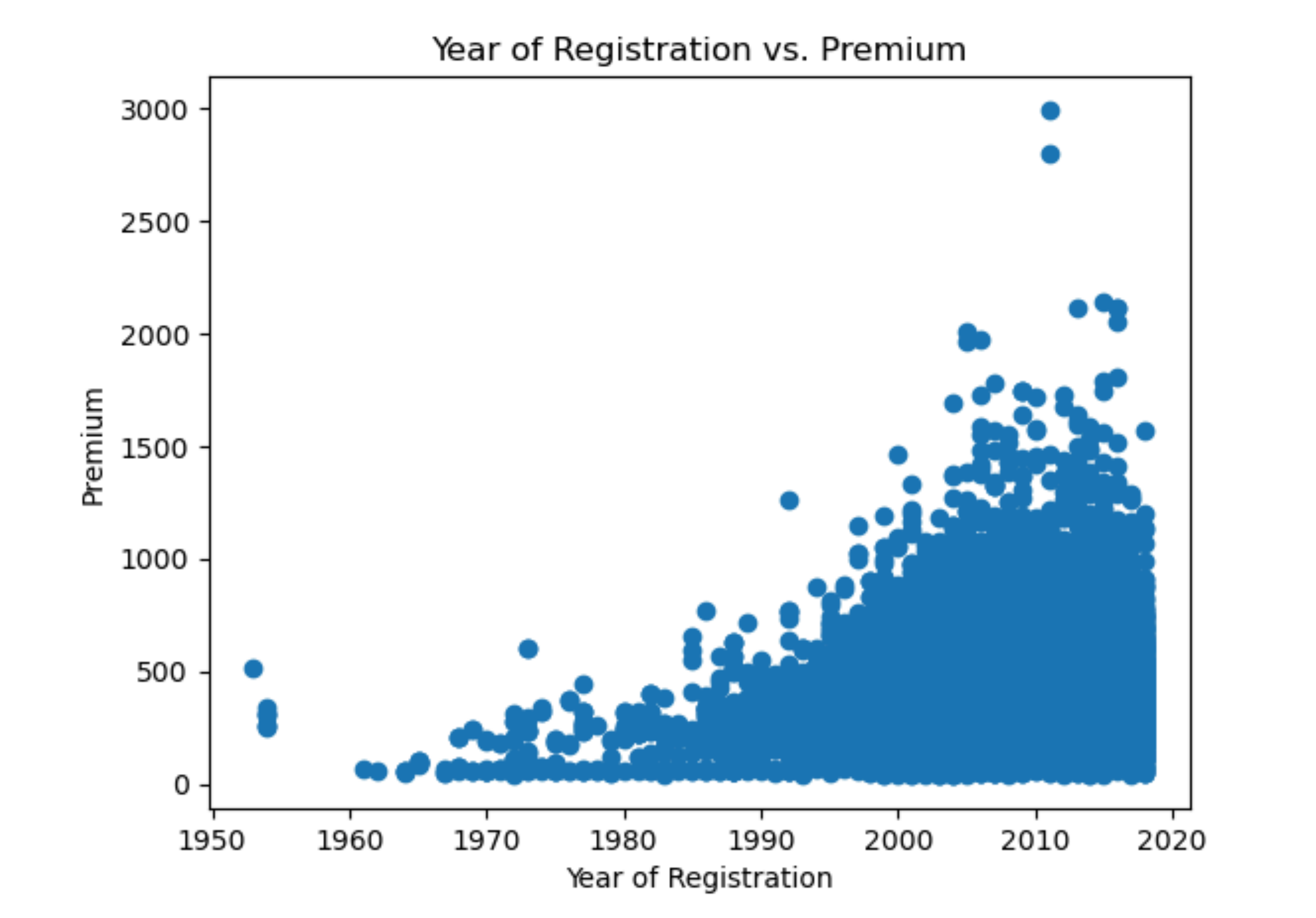
	Policies_in_force	Max_policies	Premium	Cost_claims_year	N_claims_year	Year_matriculation	Power	Cylinder_capacity	Value_vehicle	Age
Policies_in_force	1	0.77	-0.05	0.016	0.18	-0.061	0.029	0.063	0.085	0.043
Max_policies	0.77	1	-0.037	0.0096	0.16	-0.059	0.045	0.071	0.091	0.089
Premium	-0.05	-0.037	1	0.045	0.064	0.29	0.28	0.13	0.34	-0.098
Cost_claims_year	0.016	0.0096	0.045	1	0.2	0.011	0.019	0.015	0.021	-0.021
N_claims_year	0.18	0.16	0.064	0.2	1	-0.022	0.031	0.057	0.046	-0.032
Year_matriculation	-0.061	-0.059	0.29	0.011	-0.022	1	0.2	-0.16	0.2	-0.11
Power	0.029	0.045	0.28	0.019	0.031	0.2	1	0.62	0.77	-0.025
Cylinder_capacity	0.063	0.071	0.13	0.015	0.057	-0.16	0.62	1	0.71	0.017
Value_vehicle	0.085	0.091	0.34	0.021	0.046	0.2	0.77	0.71	1	-0.0062
Age	0.043	0.089	-0.098	-0.021	-0.032	-0.11	-0.025	0.017	-0.0062	1

Linear Regression Model

- No major correlations in dataset
- Models ran:
  - OLS - R-squared: .255
  - Lasso - R-squared: .2552
  - Ridge - R-squared: .2553

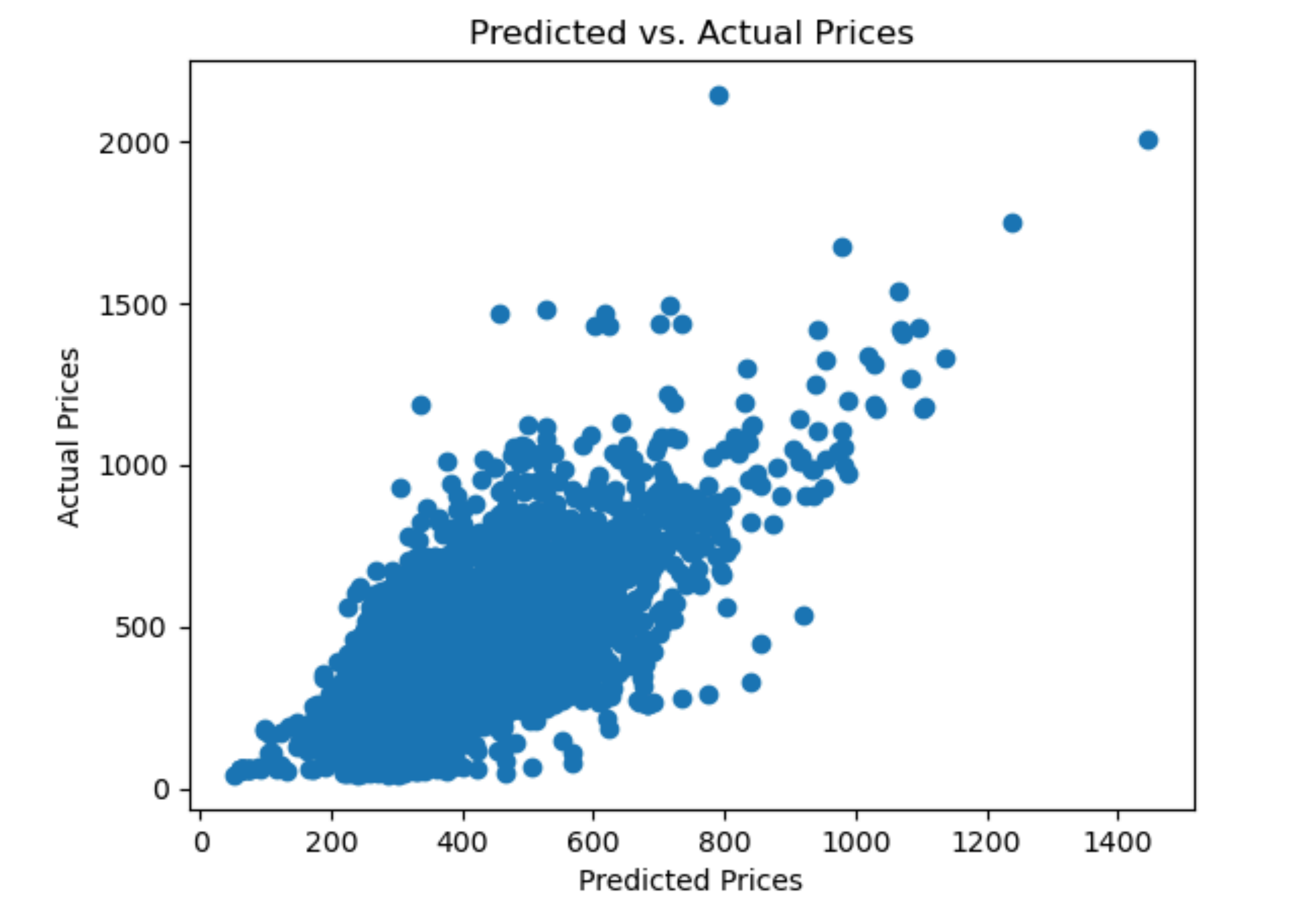
- Pearson Correlation Coefficients:
  - Value of the Vehicle: .3359
  - Year_matriculation: .2920
  - Power: .2795



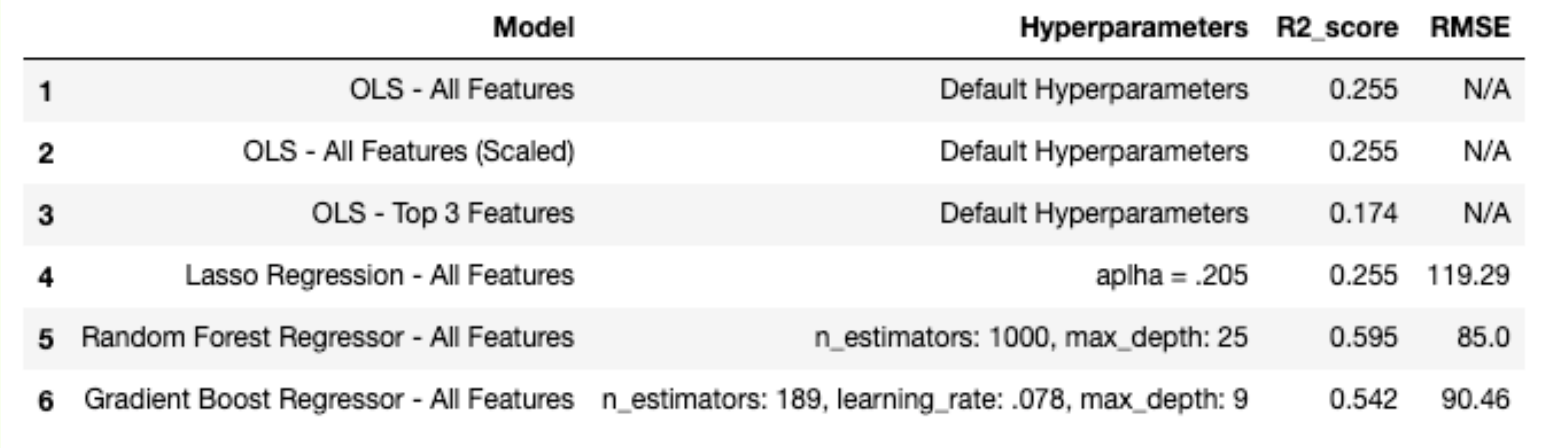


Tree Based Models

- 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



Conclusion / Next Steps



	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

- 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