**Report Analysis**

**Executive Summary:**

**Topic:** Used car analysis and modeling on Price

**Raw Data Set:** [kaggle-usedcars-data](https://www.kaggle.com/austinreese/craigslist-carstrucks-data)

N = 423,868

Number of Columns: 24

Description of the original data set:

Summary: There are total 25 columns, here is the list of the columns:

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Names** | **Column description** | **Remove or Keep** | **Reason** |
| ID | Entry ID | Remove | Irrelevant to analysis |
| url | listing URL | Remove | Irrelevant to analysis |
| Region | Craigslist Region | Remove | Irrelevant to analysis |
| Region URL | Region URL | Remove | Irrelevant to analysis |
| Price | Entry Price | Keep |  |
| Year | Entry Year | Keep |  |
| Manufacturer | Manufacturer of vehicle | Keep |  |
| Model | model of vehicle | Remove | Low data quality |
| Condition | Condition of Vehicle | Keep |  |
| Cylinders | Number of Cylinders | Keep |  |
| Fuel | Fuel type | Keep |  |
| Odometer | Miles traveled by vehicle | Keep |  |
| Title\_Status | Title Status of Vehicle | Keep |  |
| Transmission | Transmission of Vehicle | Keep |  |
| Vin | Vehicle Identification Number | Remove | Irrelevant to analysis |
| Drive | Type of Drive | Keep |  |
| Size | Size of Vehicle | Keep |  |
| Type | Generic type of Vehicle | Keep |  |
| Paint\_Color | Color of Vehicle | keep |  |
| Image\_URL | Image URL | Remove | Irrelevant to analysis |
| Description | Listed Description of Vehicle | Remove | Irrelevant to analysis |
| County | Useless column left in by mistake | Remove | Irrelevant to analysis |
| State | State of listing | Keep | For geographic graph |
| lat | Latitude of listing | Keep | For geographic graph |
| long | Longitude of listing | Keep | For geographic graph |

**Data Set Included in this analysis:**

N = 146,495

Number of the Features = 13

**Methods in used:**

Data Visualizations: usmap, ggplot

Machine Learning: Random Forest, Lasso/Ridge regression, multivariate regression

Lasso/Ridge regression: Empirical Risk presents the prediction power of model f(x) to all samples in training data set; Expected Risk presents the prediction power of model f(x) to all samples in universe (including training and testing data set); Structural Risk is adding the regularization (penalty) after the Empirical Risk. As Empirical Risk smaller, the more complex the model f(x) is, which indicates more parameters in f(x). When Empirical Risk is small enough, there will be overfitting issue. In order to avoid overfitting, we need to lower the complexity of the model. It means we want to minimize penalty. We want to minimize both empirical risk and model complexity (penalty) at the same time. Lasso penalty and Ridge penalty will help to reduce the overfitting risk and to make sure the model can be very well generalized. Ridge penalty (makes all theta(s) close but not equal to 0 to avoid overfitting. Thus, Ridge penalty can lead to lower RMSE and make better prediction. Lasso penalty () penalize theta(s) to 0 easier than Ridge. It imports sparsity; thus, it can be used for features selection.

Random Forest:

A method of combining multiple random trees into one big classifier using even more randomization (under the concept of bagging – Bootstrap Aggregation).

Step 1: Randomly select number of observations with replacement (Bootstrap).

Step 2: Randomly pick number of features without replacement

One of method is information gain:

Entropy - H(Y), when it is most certain, it is 100% or 0%; when it is most uncertain, it is 50%.

Information Gain: H(Y) – H(Y|X)

Step3: Repeat step 2 until cannot be further split.

Step4: Repeat step1 ~ step 3 to form random forest

**Findings:**

The range of the average used car odometer in most states is 90k to 100k. In most states, the most popular used car brand in transaction is Ford. The favorite color is white. East and west coast states favor sedan, while middle states prefer SUV.

And the best price prediction model is Random Forest.

**Introduction:**

To build a model to predict the used car’s **price** based on Car Year, Car Manufacturer, Car Condition, Number of Cylinder, Type of Fuel, Car Odometer, Type of Title, Type of Transmission, Type of Drive, Car Size, Car Type, and Paint Color.

To build a model to identify the used car’s **condition** based on Car Year, Car Manufacturer, Car Price, Number of Cylinder, Type of Fuel, Car Odometer, Type of Title, Type of Transmission, Type of Drive, Car Size, Car Type, and Paint Color.

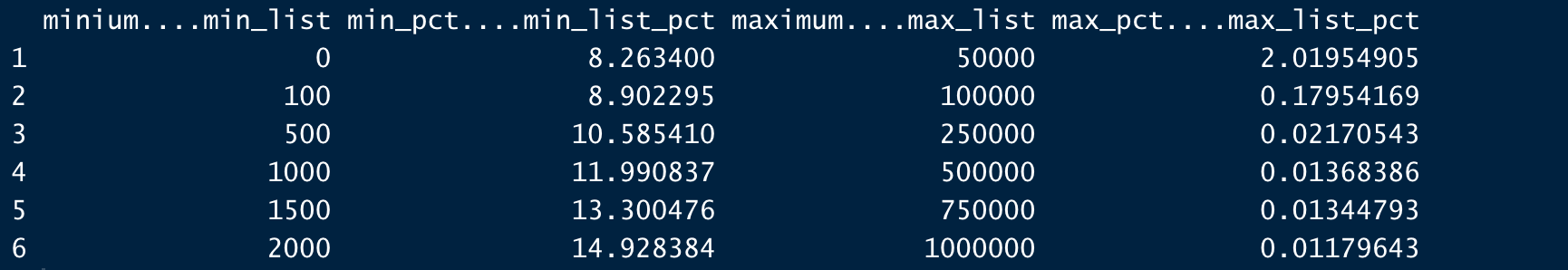
Organizations: Car dealer, Car owner, Potential used car buyer

Challenges: During cleansing data set process, it is hard to find a best way to deal with Null values and the categorical variables with too many levels.

**Data Cleansing:**

Total the number of observations is around 420,000:

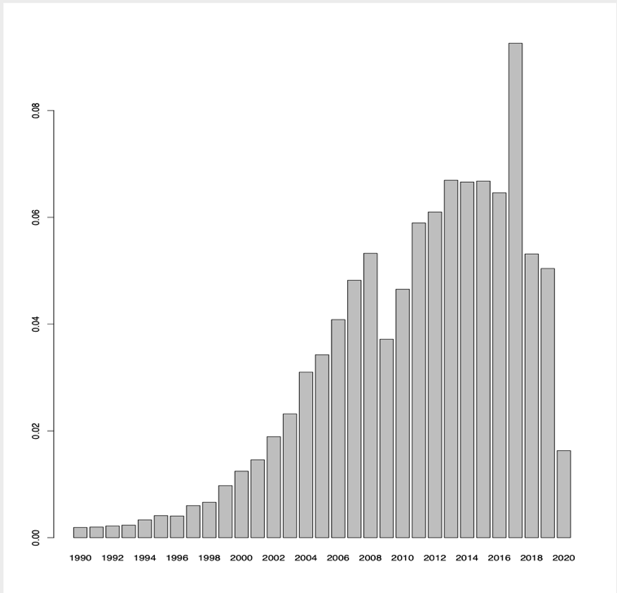
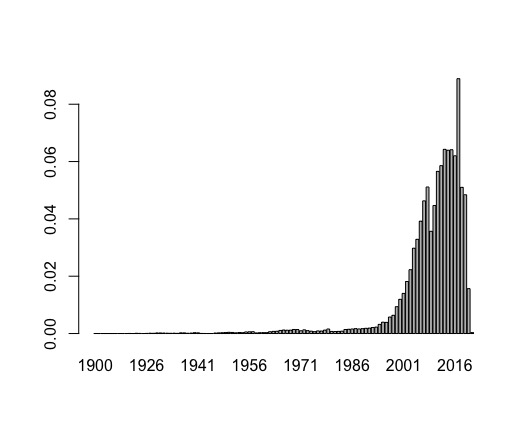
Price Column: Remove NA and redefine the price range:



Conclusion: We only include the car price in the range of [$500, $250,000], because we think it is a reasonable range of the price for a used car, which covers 90% of data points.

After refining price range, total the number of observations is around 380,000

Year Column: Remove NA and redefine the Year range:



Conclusion: The original data set’s manufacturer year is from 1900 to 2021 and we find most of used cars’ manufacturer year is clustered in a range from 1990 to 2020. Moreover, data from 2021 is not completed due to 2021 hasn’t come yet and we believe data before year 1990 are untrustful.

After refining year range, total the number of observations is around 280,000

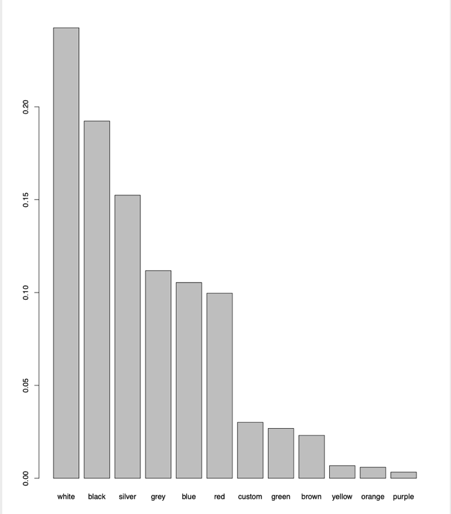
Odometer Columns: Remove NA

After removing NA observations from Odometer, total the number of observations is around 240,000.

Manufacturer, Condition, and Size Columns: These are important parameters about a used car, thus if more than one of three columns contain NA, it is removed.

After removing NA observations from Manufacturer, Condition, and Size Columns, total the number of observations is around 150,000.

Color Column: regrouping the levels as only keeping 6 commonly used color and setting rest colors as other.



After regrouping the levels of color column, total number of observations is around **146,000**.

State, Long and Lat Columns: Removed from further analysis.

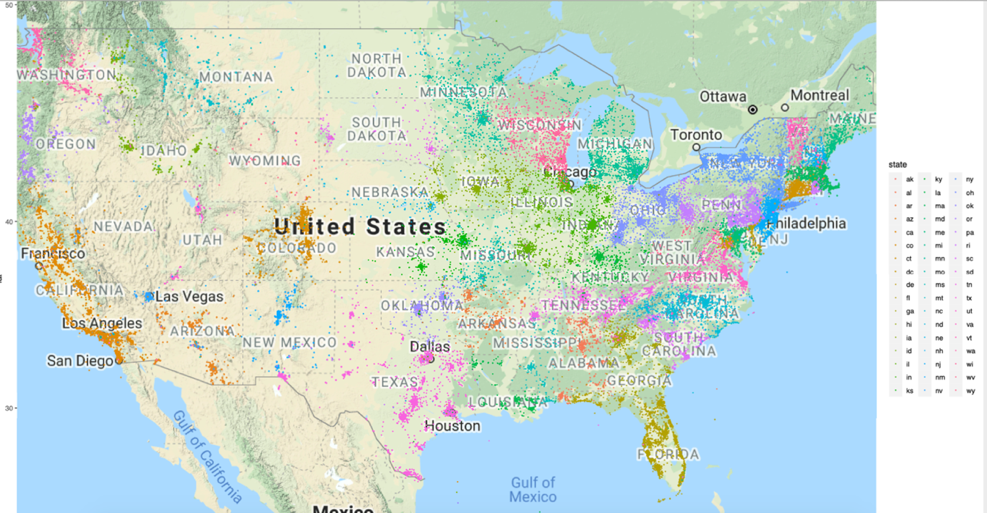
**Data Transformation:**

Transform categorical variables to binary variables:

|  |  |  |
| --- | --- | --- |
| **Data transformation summary** | | |
| **Column Name** | **Levels** | **Transformed Columns** |
| manufacturer | 41 | gmc, hyundai, toyota, ford, chevrolet… |
| condition | 7 | excellent, good, fair, like new, NA, new, salvage |
| cylinders | 8 | 10 cylinders, 12 cylinders, 3 cylinders, 4 cylinders, na… |
| fuel | 5 | diesel, electric, gas, hybrid, na |
| title\_status | 6 | clean, lien, rebuilt, salvage, parts only, NA |
| transmission | 3 | automatic, manual, NA |
| drive | 4 | 4wd, fwd, rwd, NA |
| size | 5 | NA , full-size, mid-size, compact, sub-compact |
| type | 13 | truck, SUV, sedan… |
| paint | 7 | red, grey, blue, white, Other, silver, black |

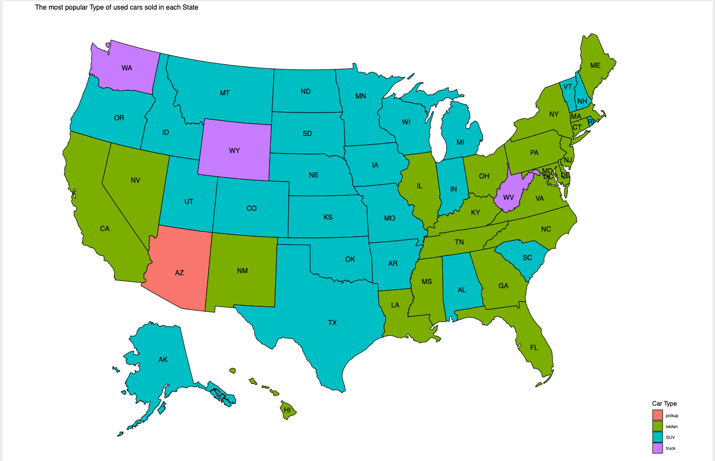
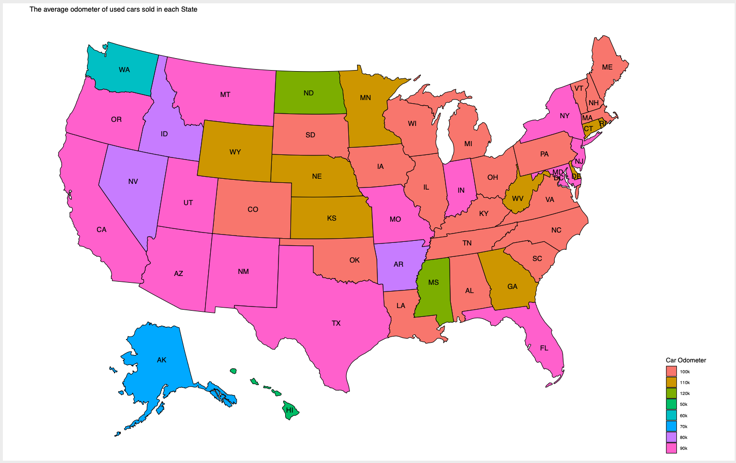
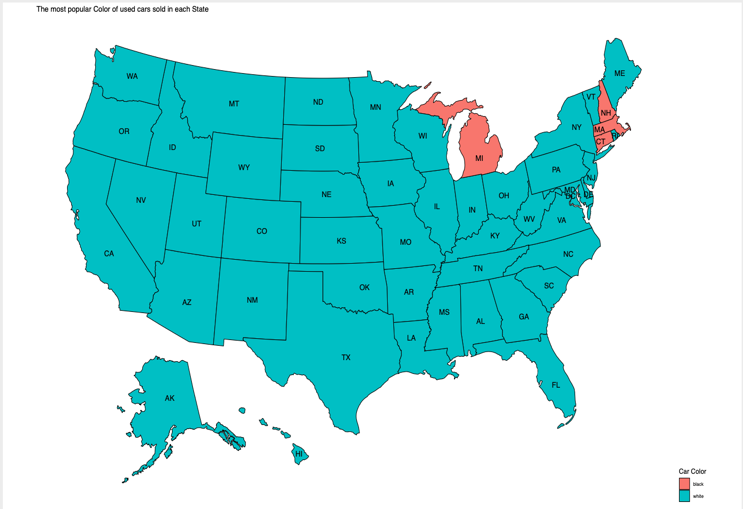
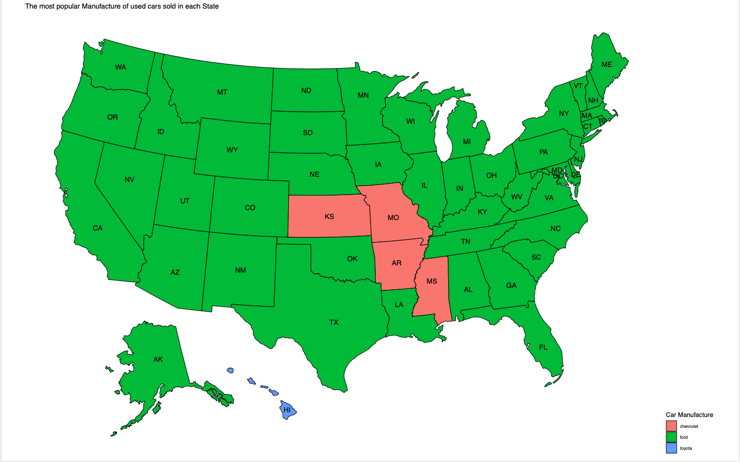
Data analysis:

Geographic analysis



Conclusion: Most of the used cars are sold from east coast to mid and west coast.

Consumer Insights:



Insights: East and west coast states favor sedan, while middle states prefer SUV. The favorite color is white. In most states, the most popular used car brand in transaction is Ford. The range of the average used car odometer in most states is 90k to 100k.

Model Analysis:

Executive Summary:

N = 14.2k

Reason: the number of observations included in my analysis below is around 14.2k after removing 2.5% of observations with very low and extremely high odometers since we don’t want to include new car and non-sense data.

Categorical Variables are included:

|  |
| --- |
| manufacturer |
| condition |
| cylinders |
| fuel |
| title\_status |
| transmission |
| drive |
| size |
| type |
| paint |

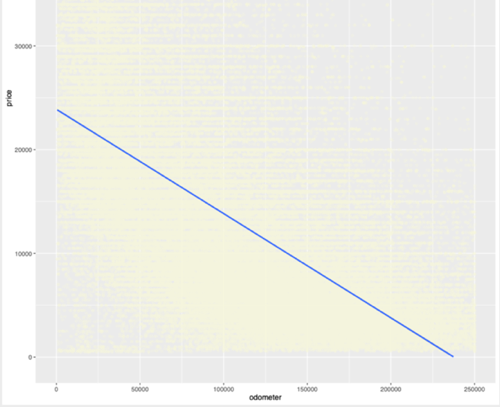
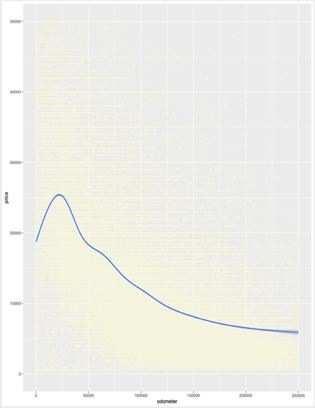
Numerical Variables are included:

Price(Y), Year (set initial year back to 0), Odometer

Partition:

I divide the dataset as 50% training data, 30% validation data, and 20% testing data for final model RMSE reporting.

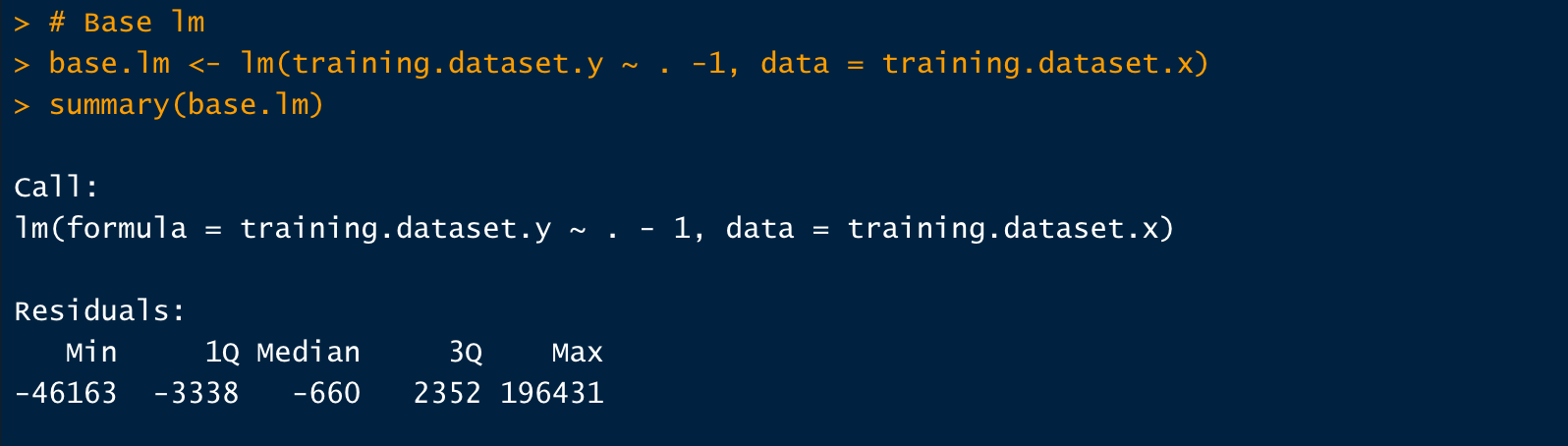
In order to run a regression model, I have checked the assumptions about linear regression and explored the relationship among the features and Y(Price):

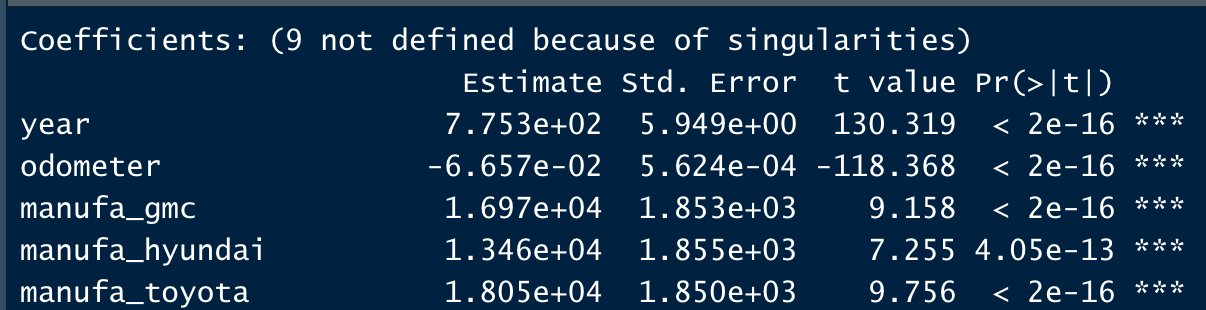


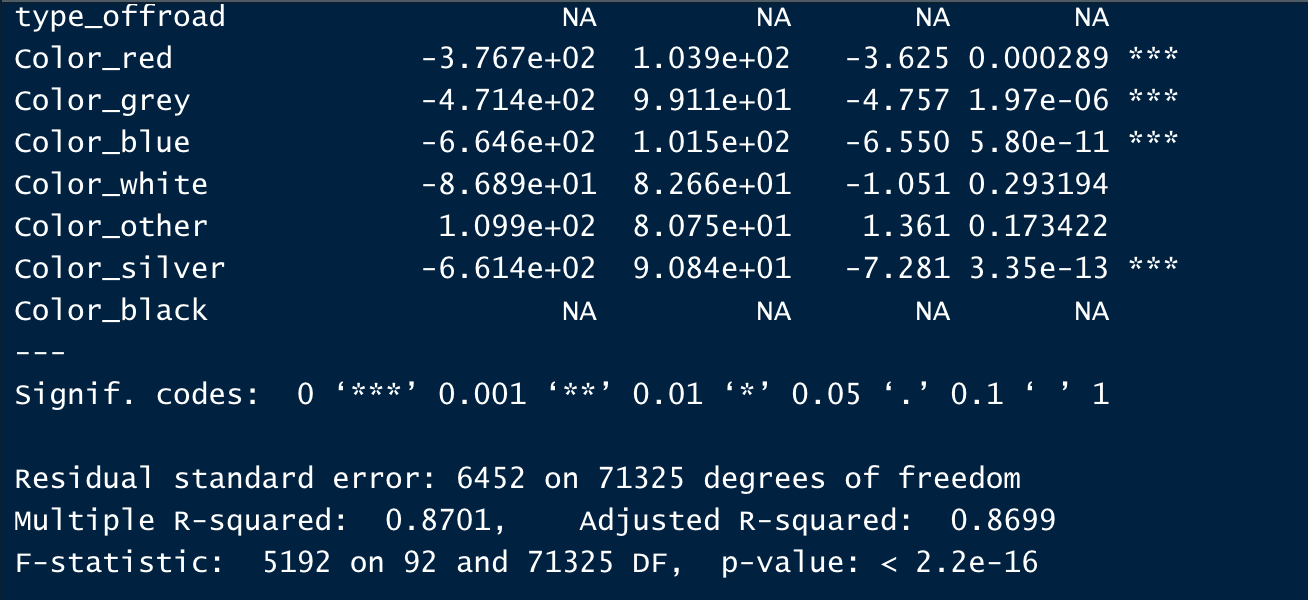
Conclusion: Odometer = 0 has a lot of noises and I assumed that odometer < 500 is almost likely new cars. After keeping the range between 500 and 300000 (removed 2.5% of total 14.6k observations), it shows the linear relationship among Odometer and Price.

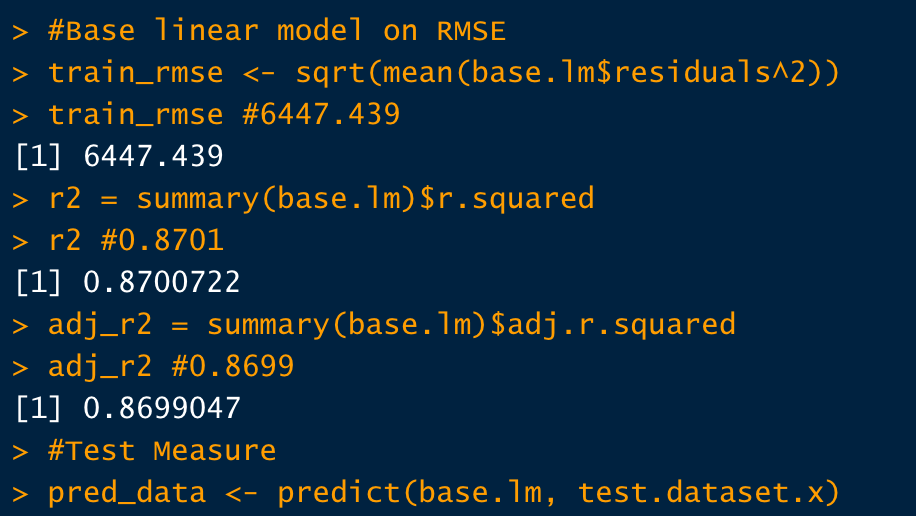
Year Column: setting the initial year back to 0, in order to mitigate the impact on Beta from larger feature.

Base linear model result:

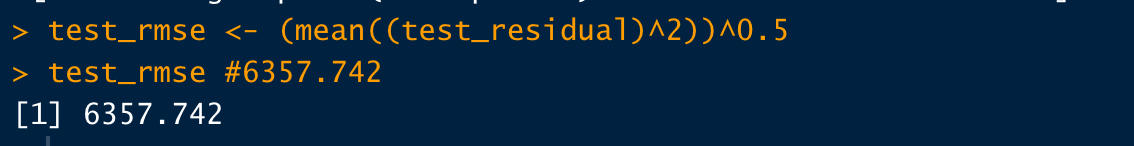




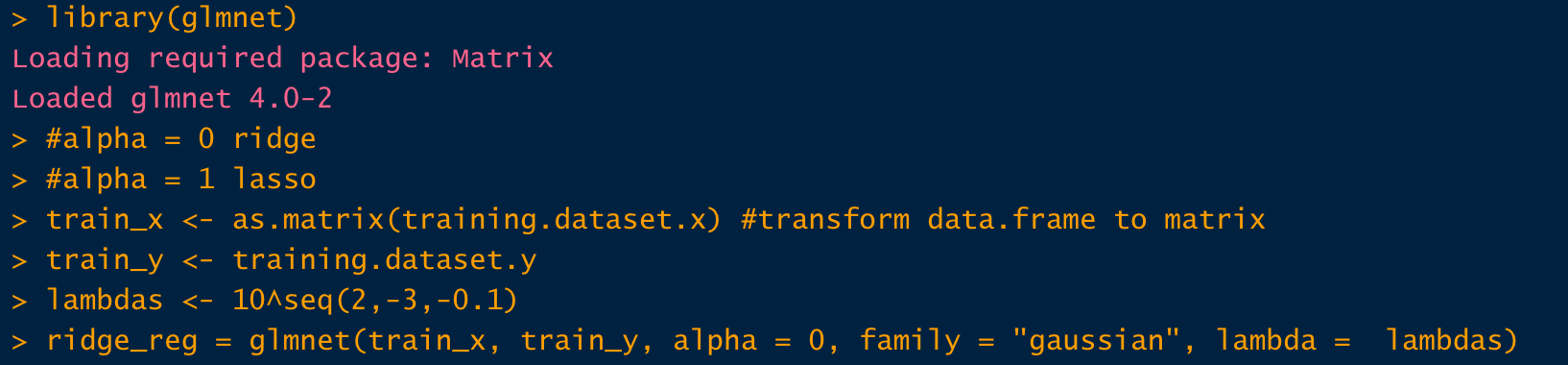


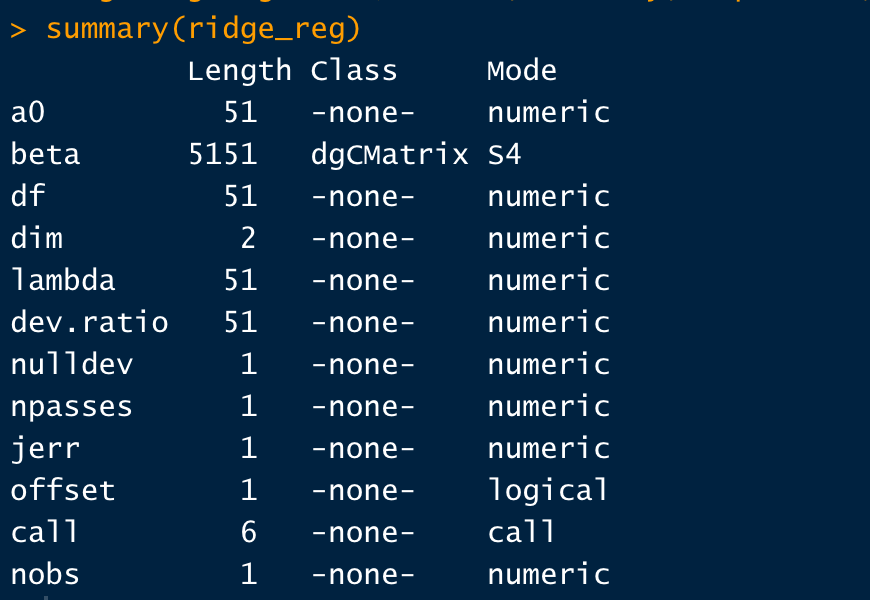


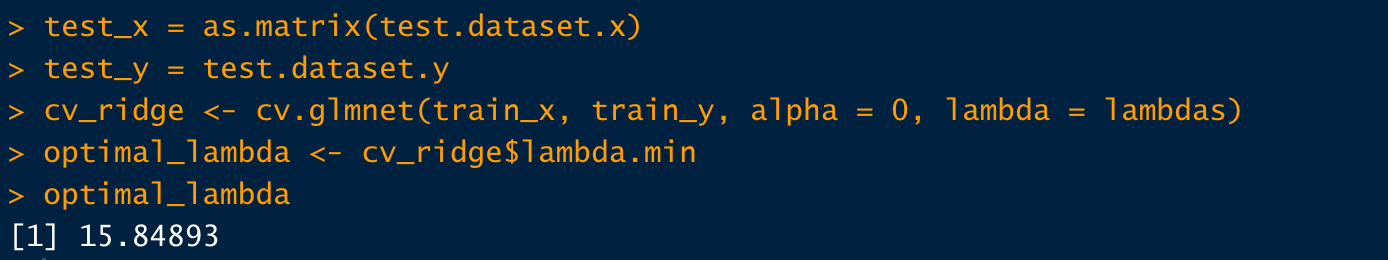


Since the data set contains binary variables, in order to prevent from perfect multicollinearity, I do not include intercept in my linear model.

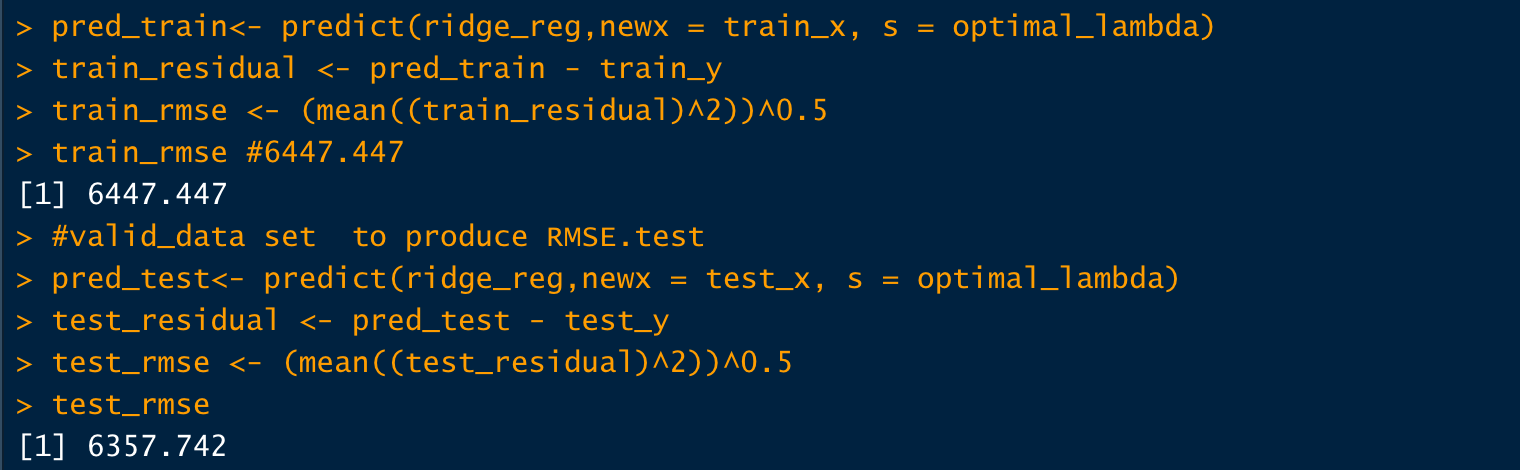
Ridge and Lasso summary:

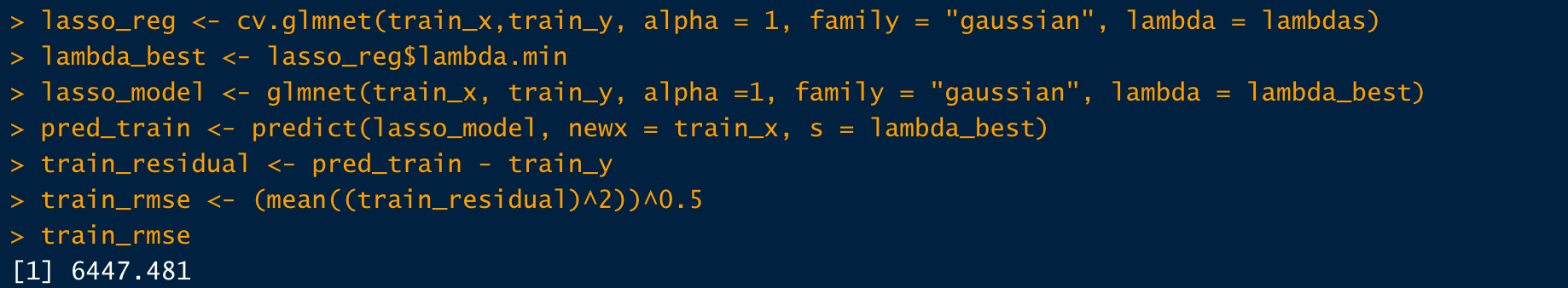




I picked 51 lambdas in a range from 0.001 to 100, and use cross validation to train the optimal lambda. 

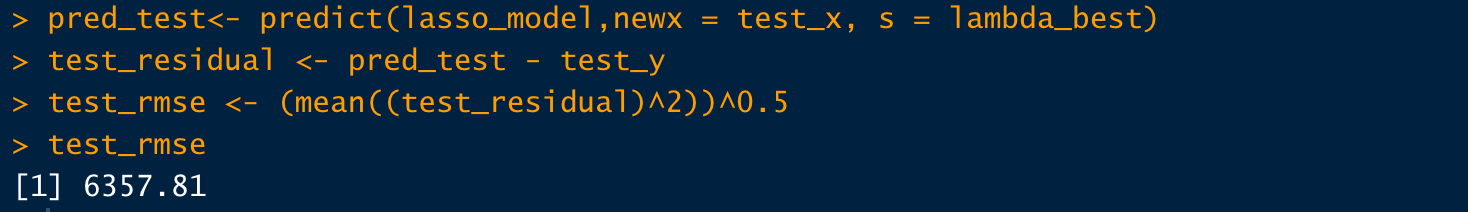
Optimal\_lambda (Ridge) = 15.8489





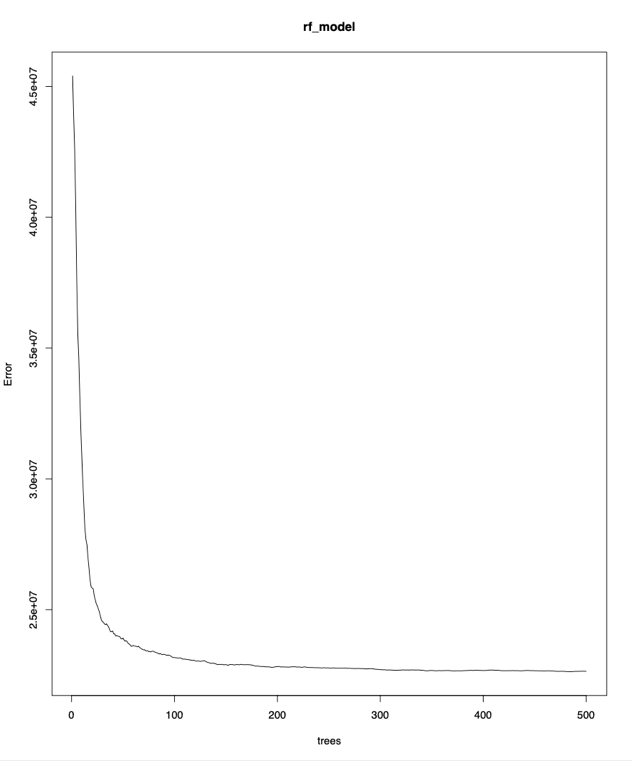
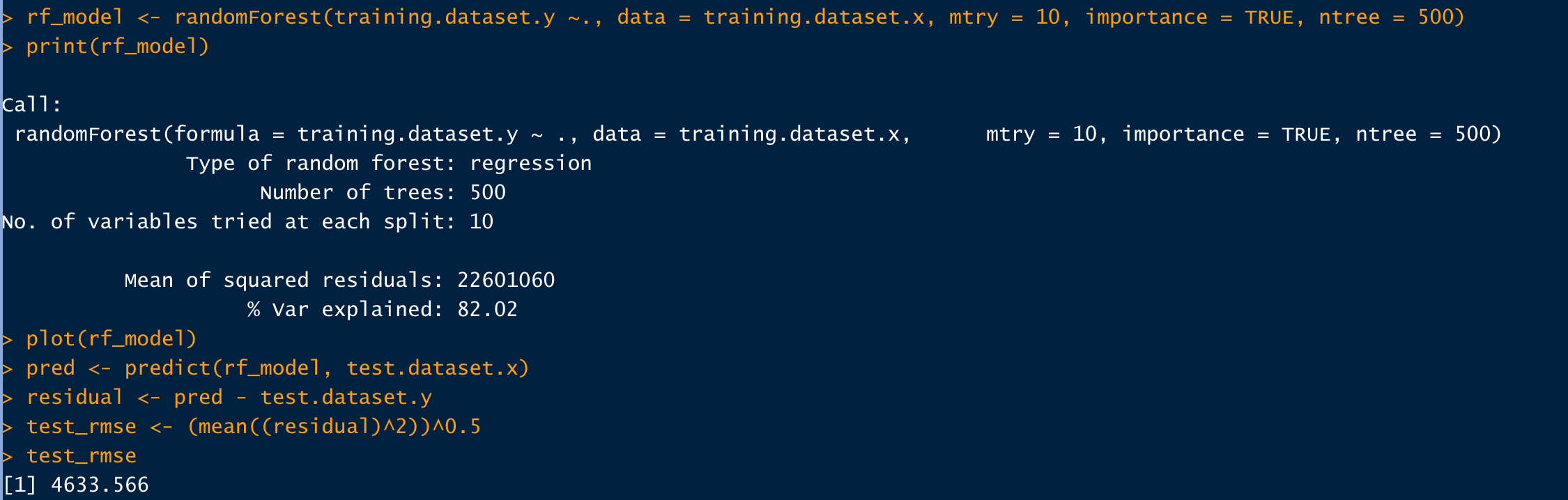


Lasso Optimal Lambda = 2.51



Overall, my RMSE based on Lasso and Ridge regression doesn’t outperform basic linear regression significantly, one of the reasons is that Lasso and Ridge are trying to resolve overfit issue but the base linear model doesn’t show any overfit.

Random Forest Result:



Conclusion: based on this plot, I set my ntree equal 500, and mtry equal 10 because it is the reasonable size of tree based on square root of total number of features around 100. Train.mse is around 4754 and test.mse around 4633, it shows the model isn’t overfit and 82 % variance is explained.

**Discussion and Conclusion:**

The range of the average used car odometer in most states is 90k to 100k, thus a recommendation for the car owners to re-sell their cars in this range. In most states, the most popular used car brand in transaction is Ford, thus a recommendation for the car dealer to promote Ford. The favorite color is white, thus a recommendation for the car dealer to display more diversity white color cars. East and west coast states favor sedan, while middle states prefer SUV, thus a recommendation for the car dealer to focus on different sale strategy in different regions.

And the best price prediction model is Random Forest which is a method of combining multiple random trees into one big classifier using even more randomization (under the concept of bagging).

**Reference**

@drsimonj .(2017). How and When: Ridge Regression with Glmnet, Retrieved from <https://drsimonj.svbtle.com/ridge-regression-with-glmnet>

kumar, U. (2020). Random Forest Approach for Regression in R Programming, Retrieved from <https://www.geeksforgeeks.org/random-forest-approach-for-regression-in-r-programming/>

Eremenko, K. (2020). R Programing A-Z: R For Data Science With Real Exercises!, Retrieved from <https://www.udemy.com/course/r-programming/learn/lecture/4679128#questions>

Eremenko, K. (2020). R Programing: Advanced Analytics In R For Data Science, Retrieved from <https://www.udemy.com/course/ranalytics/learn/lecture/5192738?start=105#overview>

Prabhakaran, S. (2016). The Complete ggplot2 Tutorial - Part 2 | How To Customize ggplot2 (Full R code), Retrieved from <http://r-statistics.co/Complete-Ggplot2-Tutorial-Part2-Customizing-Theme-With-R-Code.html#Legend%20Title>

Lorenzo, P. (2020). Mapping the US, Retrieved from <https://cran.r-project.org/web/packages/usmap/vignettes/mapping.html>

Lorenzo, P. (2020). Advanced Mapping, Retrieved from <https://cran.r-project.org/web/packages/usmap/vignettes/advanced-mapping.html>

Zhou, Z. (2019). Machine learning, Beijing: Qinghua University Publisher, pp 171 – 196.