

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group # | Following Instructions (11 pts) | Writing Quality (11 pts) | Abstract (12 pts) | Data Collection / Cleaning / Exploration (11 pts) | Data Exploration Insights (11 pts) | Methodology (11 pts) | Predictions (11 pts) | Inference (11 pts) | Conclusion (11 pts) | Additional Discretionary  Points | Numeric Grade | Letter Grade | Notes |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 12 | 11 | 11 | 12 | 11 | 11 | 11 | 11 | 9 | 11 | 0 | 98 | A |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**IST 718**

**Project Report**

*Used Car Listings Price Prediction*

**Group 12**

Pranali Shenvi

Srushti Samant

Mikhail Pinto

**Total Pages:**

13

Table of Contents

[Abstract 1](#_Toc121618211)

[*Overview of the project* 1](#_Toc121618212)

[*Description of dataset* 1](#_Toc121618213)

[*Predictors* 1](#_Toc121618214)

[*Link to the dataset* 2](#_Toc121618215)

[*Interesting insights* 2](#_Toc121618216)

[*Inferences* 3](#_Toc121618217)

[*Challenges* 3](#_Toc121618218)

[*Conclusion Summary* 3](#_Toc121618219)

[Data Collection and Cleaning 4](#_Toc121618220)

[Data Exploration 4](#_Toc121618221)

[*Insights* 4](#_Toc121618222)

[*Visualizations* 5](#_Toc121618223)

[*Data Facts* 7](#_Toc121618224)

[*Statistical Summary* 7](#_Toc121618225)

[Methodology 7](#_Toc121618226)

[*Method Description* 7](#_Toc121618227)

[*Feature Engineering* 8](#_Toc121618228)

[*Scoring Metrics* 8](#_Toc121618229)

[*Workflow diagram* 8](#_Toc121618230)

[Model Prediction (Regression models) 9](#_Toc121618231)

[*Linear Regression* 9](#_Toc121618232)

[*Random Forrest Regressor* 9](#_Toc121618233)

[*Gradient Boosted Tree Regressor* 10](#_Toc121618234)

[Model Inference (Regression Models) 10](#_Toc121618235)

[*Linear Regression (Inference)* 10](#_Toc121618236)

[*Random Forrest Regressor (Inference)* 11](#_Toc121618237)

[*Gradient Boosted Tree Regressor (Inference)* 11](#_Toc121618238)

[Conclusion 12](#_Toc121618239)

[*Model Comparison* 12](#_Toc121618240)

[*Analysis* 12](#_Toc121618241)

[Appendix 13](#_Toc121618242)

[*References* 13](#_Toc121618243)

[*Tables* 14](#_Toc121618244)

[*Figures* 14](#_Toc121618245)

## *Tables*

|  |  |  |
| --- | --- | --- |
| Table number | Table name | Page number |
| 1.1 | Data attributes and their description | 1 |
| 1.2 | Predictor Columns | 1 |
| 1.3 | Statistical summary of Price and Mileage | 2 |
| 5.1 | Scoring metrics and values for linear regression | 9 |
| 5.2 | Scoring metrics and values for RF Regressor | 9 |
| 5.3 | Scoring metrics and values for GBT Regressor | 10 |

## *Figures*

|  |  |  |
| --- | --- | --- |
| Figure number | Figure name | Page Number |
| 1.1 | Car Prices Density Distribution | 2 |
| 2.1 | Schema (Data types of all attributes) | 4 |
| 3.1 | Correlation Matrix | 4 |
| 3.2 | Model Manufacturing Year by Price | 5 |
| 3.3 | Map of Average price per US States | 5 |
| 3.4 | Car makers by count of cars available | 6 |
| 3.5 | Average price of used cars per top 10 manufacturers | 6 |
| 3.6 | Statistical summary of the data | 7 |
| 4.1 | Workflow diagram of our methodology | 8 |
| 6.1 | Feature importance table for Linear Regression | 10 |
| 6.2 | Feature importance table for RF Regressor | 11 |
| 6.3 | Feature importance table for GBT Regressor | 11 |
| 7.1 | Model Comparison | 12 |

# 

# Abstract

## *Overview of the project*

We know that the prices of brand-new cars are increasing day by day. This has led to an increase in the number of customers who prefer buying used cars. We aim to examine the used car pricing data, gain insights, and predict the prices of the used cars. Using these insights, people who need to sell their car can decide an estimated price and anyone who would be interested in purchasing the car will be able to determine the appropriate price.

## *Description of dataset*

The attributes in dataset are either integer or string. Dimensions: 728,567 rows \* 10 columns. There are 2 index columns \_c0 and Id present in the data as well.

|  |  |  |  |
| --- | --- | --- | --- |
| Sr. No. | Attributes | Description | Data Type |
| 1 | Price | Listed price of the car | integer |
| 2 | Year | Model year of the car | integer |
| 3 | Mileage | No. of miles on the odometer of the listed car | integer |
| 4 | City | City that the car is listed in | string |
| 5 | State | State that the car is listed in | string |
| 6 | Vin | Vehicle Identification Number | string |
| 7 | Make | Car manufacturer | string |
| 8 | Model | Car model | string |

Table 1.1: Data attributes and their description

## *Predictors*

We are using the following columns as our predictor columns:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor columns | | | | |
| Year | Mileage | State | Make | Model |

Table 1.2: Predictor columns

## *Link to the dataset*

The following link can take us to the site to download the dataset:

<https://www.kaggle.com/datasets/brentpafford/true-car-listings-2017-project>

## *Interesting insights*

|  |  |  |
| --- | --- | --- |
| Summary | Price | Mileage |
| count | 728,567 | 728,567 |
| mean | 20,940.54 | 48,286.19 |
| stddev | 11,367.64 | 96,878.96 |
| min | 1,790 | 5 |
| max | 429,900 | 77,587,763 |

Table 1.3: Statistical summary of Price and Mileage

* According to the summary statistics for Price and Mileage, the standard deviation values are extremely large and the maximum and minimum values are far apart
* It is surprising that some cars having extremely low mileage have been put up for sale

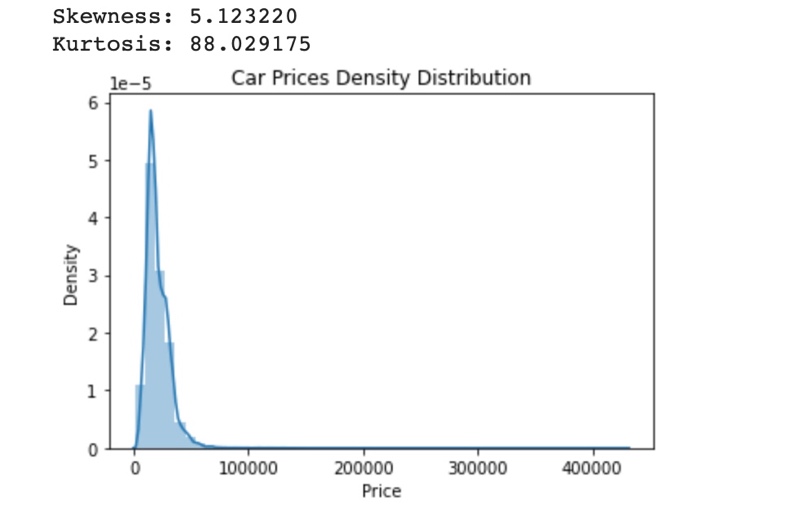


Figure 1.1: Car Prices Density Distribution

* The density plot shown above for car prices is right skewed
* The Kurtosis value is around 88 which is high
* This suggests that there are many outliers in the data which need to be removed
* If we see the states, Wyoming and Texas have the highest average price of used cars. On the other hand, lowest prices of used cars can be seen in the state of Ohio and Hawaii

## *Inferences*

We have used a user defined function to check which are the best columns and in what order should the variables be added to the regression equation to generate the least MSE. Using that, the features that impact the price prediction of used cars are:

* Year
* Make

## *Challenges*

The following were the challenges faced:

* The data had a lot of outliers, and even after cleaning, there were still a lot of them. As a result, the prediction model did not perform as predicted, leading to poor scores.
* About 2500 unique values were included in the City and Model column (high cardinality), which affected the performance of Random Forrest and GBT since a large number for the "Max Bins" parameter was required. They had to be taken out of the prediction model as a result.
* There were far more rows in the data than there were columns. The effectiveness of the model was also impacted as a result.

## *Conclusion Summary*

None of the models performed as predicted, with fairly large Mean squared errors and low R squared values, due to the data's high degree of skewness and small number of columns. In comparison to linear regression and random forest regression models, the GBT regression model generated the best results for us. More variables would be needed if we want better results.

# Data Collection and Cleaning

Text

Description automatically generated

Figure 2.1: Schema (Data types of all attributes)

There are no null or duplicate values in the dataset. We don’t need to make any changes to the data type as the column attributes' data types are integer and string which are correct. As we saw above, there are many outliers in our dataset, we tried to remove the numeric data points that lie outside the Interquartile range of the 20th to 80th percentile. This would help eliminate any bias and inaccuracies caused by very high or very low values. We also converted the column names State, Make and City into upper case as there were different groups of these column values being formed.

# Data Exploration

## *Insights*

Chart

Description automatically generated

Figure 3.1: Correlation Matrix

From the correlation matrix of numeric variables in the dataset we can see that:

* Year and Mileage are negatively correlated to each other
* Price and Mileage are also negatively correlated to each other
* Year and Price are positively correlated to each other

Chart, bar chart

Description automatically generated

Figure 3.2: Model Manufacturing Year by Price

From the above plot, the newer the model based on the model year, the selling price is higher. This trend can be seen in all model years except for years 1997 and 2016.

## *Visualizations*

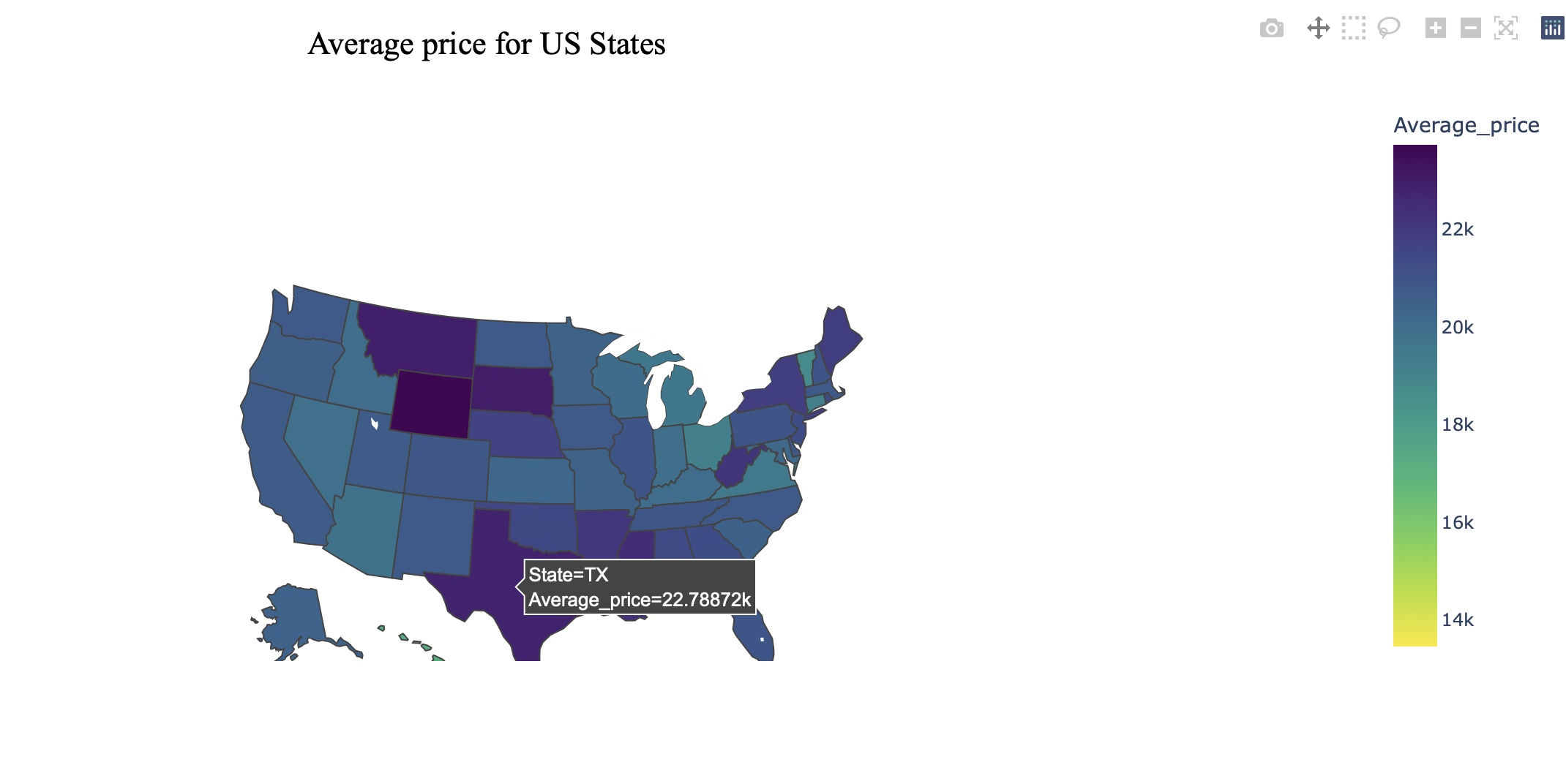


Figure 3.3: Map of Average price per US States

From the map above, we can visualize the average price of used cars for different states in the US. For example, we can see that Wyoming has the highest average price for a used car and Hawaii has the least.

Chart, bar chart, histogram

Description automatically generated

Figure 3.4: Car makers by count of cars available

From the plot above, we can visualize the count of cars at dealerships in the US broken down by make. We can also see that Ford and Chevrolet have the greatest number of used cars available.

Chart, histogram

Description automatically generated

Figure 3.5: Average price of used cars per top 10 manufacturers

From the plot above, we can visualize the Average price of cars at dealerships in the US broken down by make. We can also see that McLaren and Ferrari have the average prices.

## *Data Facts*

We see the following interesting facts from the plots above:

* From figure 3.3, we can see that instead of California or New York, Wyoming has the highest average price. This could be because of greater distance in places of importance in the state, lack of public transport and higher demand for cars.
* Ford and Chevrolet have the highest number of cars available at dealerships as seen in figure 3.4. This could be because of the two being US manufacturers making quality utility vehicles that are used the most in the US
* We can also see from figure 3.5 that McLaren and Ferrari have the highest resale price. This could be because of these brands manufacturing limited number of vehicles and most of these vehicles are not used for daily use but for leisure.

## *Statistical Summary*

Table

Description automatically generated

Figure 3.6: Statistical summary of the data

The figure above shows the descriptive statistical summary.

# Methodology

## *Method Description*

To improve the performance of our models we have dropped some columns, used String Indexing, and Scaled a column. Features like VIN, Model, and City were dropped. VIN was dropped as it has all unique values which would not help in predicting price. Model and City were dropped as they have too many unique values which was creating a problem when tuning models.

## *Feature Engineering*

String indexer was used for categorical variables and a Standard scaler was used to scale just Mileage which has a lot of variances. Finally using vector assembler, we have combined all the features to create a vector of features i.e Scaled\_Mileage, Year, state\_in, make\_in.

## *Scoring Metrics*

In order to predict the price of the used cars at dealerships, we have used regression models. In order to evaluate the regression models, we have used the following metrics

* MSE
* RMSE
* R squared value

## *Workflow diagram*

Diagram

Description automatically generated

Figure 4.1: Workflow diagram of our methodology

We first take the used cars data and perform data cleansing. Next, we take our categorical column and put it through a string indexer. Furthermore, we scale down the Mileage using Standard scaler. We use vector assembler to combine the predictors into a single vector column. Finally, we feed the data to our regression models for price prediction.

# Model Prediction (Regression models)

We have used 3 regression models and have compared their scores to find the model and features that best predict the price.

## *Linear Regression*

This is a regressor type of model. We have used ‘Price’ as the dependent variable (column to be predicted). The scoring metrics that we have used are Mean squared error (MSE), Root mean squared error (RMSE) and R squared value to show how our model is performing. Since the model's performance is not significantly impacted by the hyperparameters, grid search was not conducted for this model. The scores of the model are shown in Table 5.1.

|  |  |
| --- | --- |
| MSE | 110779471.7 |
| RMSE | 10525.18 |
| R squared | 0.14 (14%) |

Table 5.1: Scoring metrics and values for linear regression

## *Random Forrest Regressor*

This model is of the regressor type. Price is the dependent variable that we have employed (column to be predicted). Mean squared error (MSE), Root mean squared error (RMSE), and R squared value are the scoring metrics we've utilized to demonstrate how well our model is working. Grid search was used for this model in order to tune the hyperparameters. Table 5.2 displays the model's scores.

|  |  |
| --- | --- |
| MSE | 50172498.5 |
| RMSE | 7083.2 |
| R squared | 0.60 (60%) |

Table 5.2: Scoring metrics and values for Random Forrest Regressor

## *Gradient Boosted Tree Regressor*

This model is of the regressor type. Price is the dependent variable that we have employed (column to be predicted). Mean squared error (MSE), Root mean squared error (RMSE), and R squared value are the scoring metrics we've utilized to demonstrate how well our model is working. Grid search was used for this model in order to tune the hyperparameters. Table 5.3 displays the model's scores.

|  |  |
| --- | --- |
| MSE | 44753112.1 |
| RMSE | 6689.77 |
| R squared | 0.64 (64%) |

Table 5.3: Scoring metrics and values for Gradient Boosted Tree Regressor

# Model Inference (Regression Models)

From the regression models used, we can infer the following things:

## *Linear Regression (Inference)*

The inference's objective is to evaluate the influence of special characteristics on the predicted variable. In the end, the least significant elements from the predictor are removed in order to fine-tune the best model. In order to determine the most significant positive and negative features, scores have been ranked based on absolute value. More positive scores have a greater positive influence, while more negative scores have a greater negative impact.

Table

Description automatically generated with medium confidence

Figure 6.1: Feature importance table for Linear Regression

## *Random Forrest Regressor (Inference)*

By obtaining the feature importance from the model, the inference seeks to evaluate the influence of various features on the predicting variable. Finally, the least significant elements from the predictor are removed in order to fine-tune the best model. The most important feature by a great margin is make\_in (which is the string indexed Make column) with the score of 0.66. Year is the second most important variable with the score of 0.18

Graphical user interface, table

Description automatically generated with medium confidence

Figure 6.2: Feature importance table for Random Forrest Regressor

## *Gradient Boosted Tree Regressor (Inference)*

By obtaining the feature importance from the model, the inference seeks to evaluate the influence of various features on the predicting variable. Finally, the least significant elements from the predictor are removed in order to fine-tune the best model. The most important feature by a great margin is make\_in (which is the string indexed Make column) with the score of 0.495. Year is the second most important variable with the score of 0.22

Table

Description automatically generated

Figure 6.3: Feature importance table for Gradient Boosted Tree Regressor

# Conclusion

The 3 regression models can be compared in order to find the best one suited for the data.

## *Model Comparison*

Table

Description automatically generated

Figure 7.1: Model Comparison

We may conclude that the Gradient Boosted Tree Regression algorithm yields the best results in comparison based on the model comparison shown in the above figure. The feature-engineered predictors account for around 64% of the price change. This is followed by the Random Forrest Regressor and the Linear Regression algorithm. The Linear Regression does not provide a particularly good R squared value which means that only 15% of the change in price is explained by the feature engineered predictors.

## *Analysis*

Although we get a decent accuracy using the GBT regressor for predicting the price, more variables are required such as:

* Condition
* Title category
* Number of Owners

The strongest predictor of price prediction is the vehicle's condition.

Whether the car has a clean title, or a salvage title may also affect how much it costs.

The state has a significant impact on the used automobile sale price.

There are plenty of used automobiles available in most American cities on the West Coast, but not as many on the East. So, whether you're looking to purchase or sell a car, a West Coast city is where you're most likely to get the finest offers and discounts.

# Appendix

## *References*

1. Link to the dataset: <https://www.kaggle.com/datasets/brentpafford/true-car-listings-2017-project>
2. Linear Regression: <https://spark.apache.org/docs/latest/ml-classification-regression.html>
3. Random Forrest Regression: <https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.mllib.tree.RandomForest.html>
4. Gradient Boosted Tree Regression: <https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.regression.GBTRegressor.html>
5. Hyperparameter tuning: <https://spark.apache.org/docs/latest/ml-tuning.html>
6. Feature selection: <https://stackoverflow.com/questions/53528481/feature-selection-in-pyspark>
7. Stepwise regression: <https://towardsdatascience.com/stepwise-regression-tutorial-in-python-ebf7c782c922>
8. Plotly Maps: <https://plotly.com/python/choropleth-maps/>
9. Plotly Bars: <https://plotly.com/python/bar-charts/>
10. Seaborn Bars: <https://seaborn.pydata.org/generated/seaborn.barplot.html>