DS 4635 - Final Project

**Joseph Fox, Ryan Gentilucci, Nicholas Latsis, Kobe Richards, Joshua Thurber**

# Questions to Solve

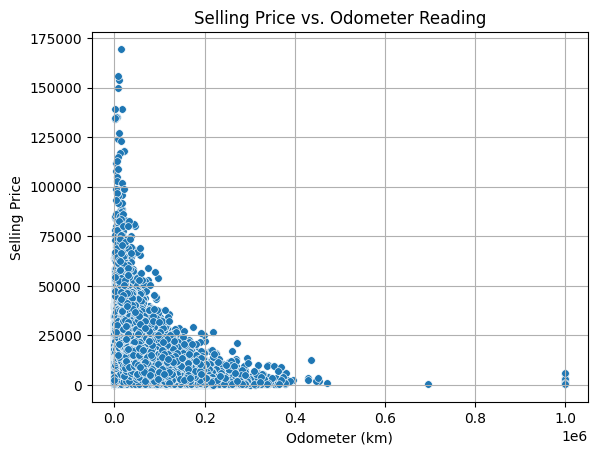
Our main goal is to develop a predictive model that accurately predicts the selling price of a car based on its various attributes. Furthermore, our model could be adapted to current market conditions by extending the time frame, enabling us to conduct a timeline analysis. We plan to use various regression analysis techniques to uncover the underlying relationships between predictors such as year, make, model, condition, odometer reading, etc., and the selling price of vehicles in the dataset. By understanding these relationships, we aimed to answer other key questions as well, such as:

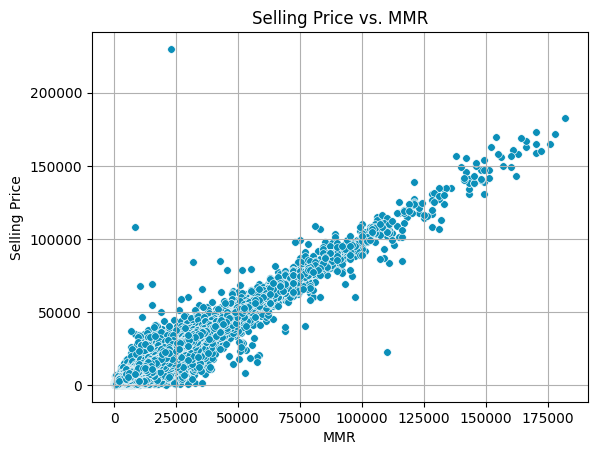
1. How do various attributes like year, make, condition, odometer reading, etc… correlate with the selling price of a car?
2. Can we identify any influential factors that significantly impact the selling price of vehicles?
3. Are there any anomalies or outliers in the data that could skew our analysis and predictions?
4. How well can we model and predict the selling price of cars based on the available attributes?

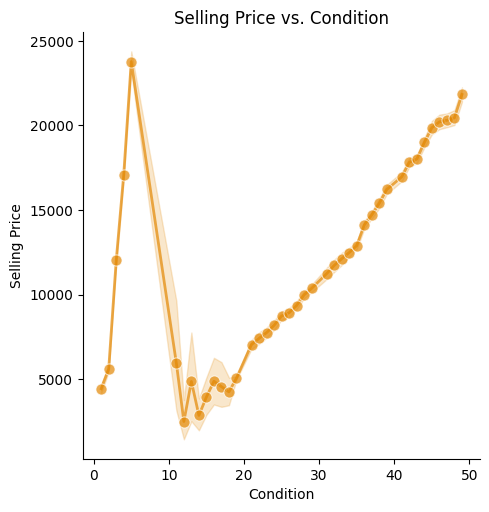
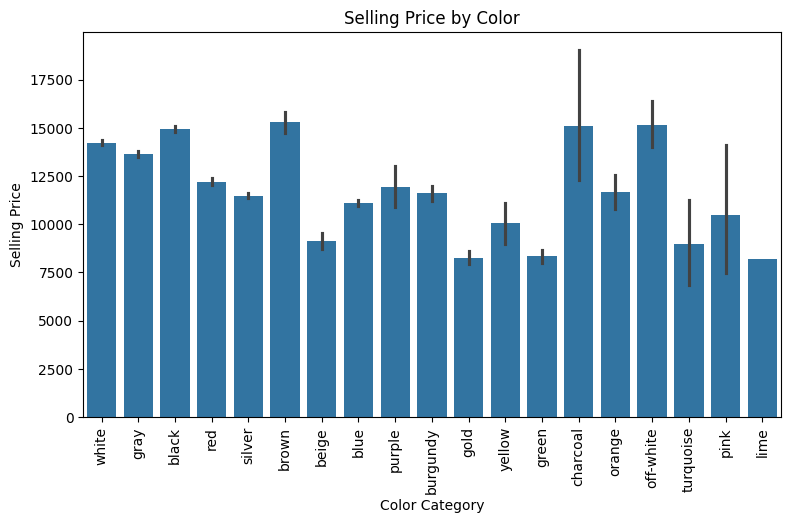
# 

# Description of Data

Our project revolves around the “Vehicle Sales data” dataset found on Kaggle. The dataset comprises approximately 560,000 rows of vehicle sales data, encompassing 16 attributes detailing various aspects of each car entry. With such an extensive dataset, there were bound to be some outliers. Some significant outliers we needed to remove from the dataset included a Ford Escape that had a selling price of $230,000, anything with a selling price below $150, and anything with a condition below 11 because when analyzing the graph there were two different scales for condition, one seemed to be from a 1-10 scale and another a 1-50. We removed the car sales below $150 and condition below 11 to accommodate for this.







## Feature Descriptions

**Year:** year the vehicle was made (1982 - 2015)

**Make:** the brand of the vehicle

**Model:** specific model of the vehicle

**Trim:** additional description of vehicle model

**Transmission:** transmission of the vehicle, either automatic or manual

**Body:** body type of the vehicle, Sedan, SUV, etc.

**Vin:** individual vehicle ID

**State:** state vehicle was sold in

**Condition:** condition of vehicle rated on an ordinal scale

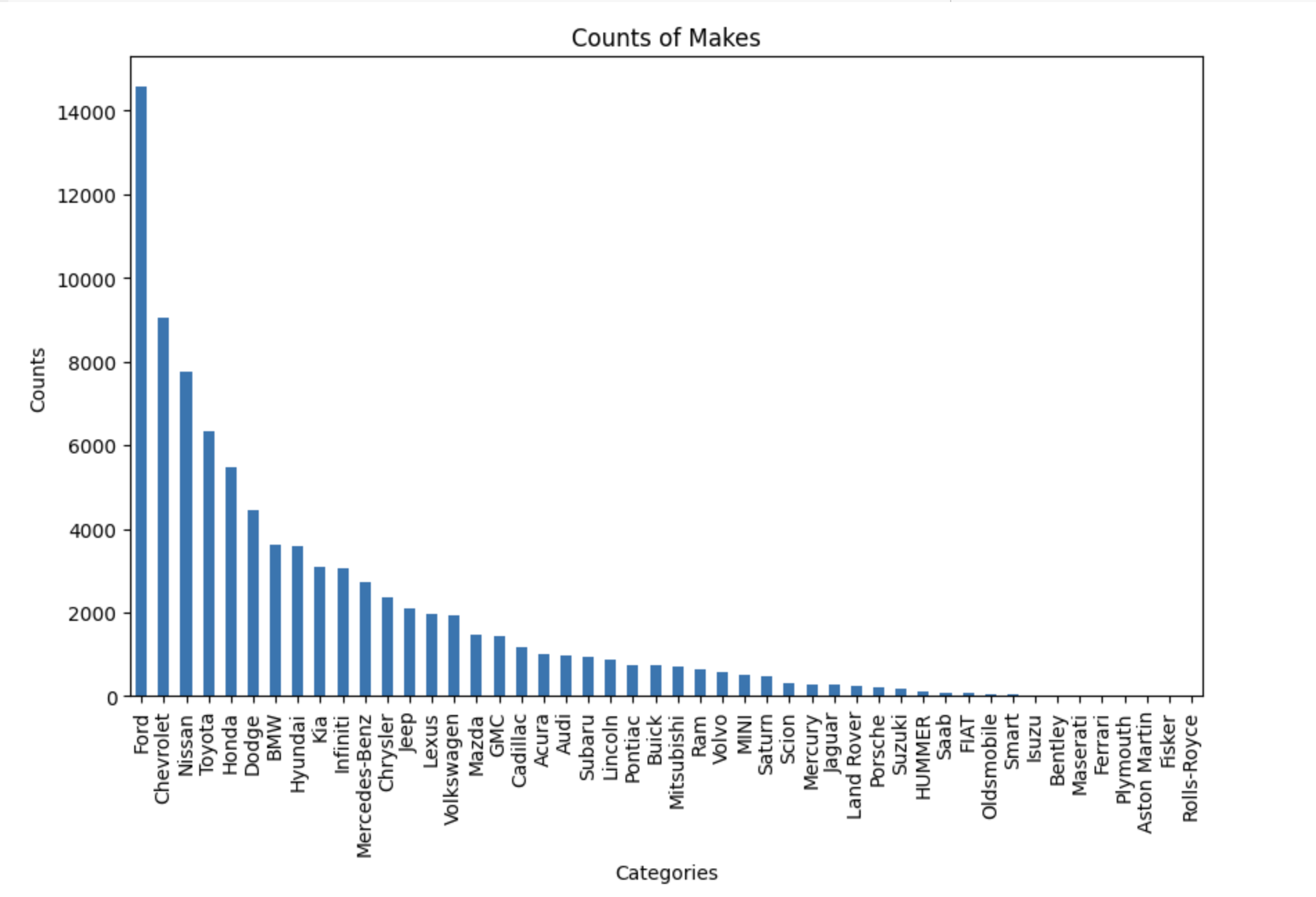
**Odometer:** miles on the vehicle when sold

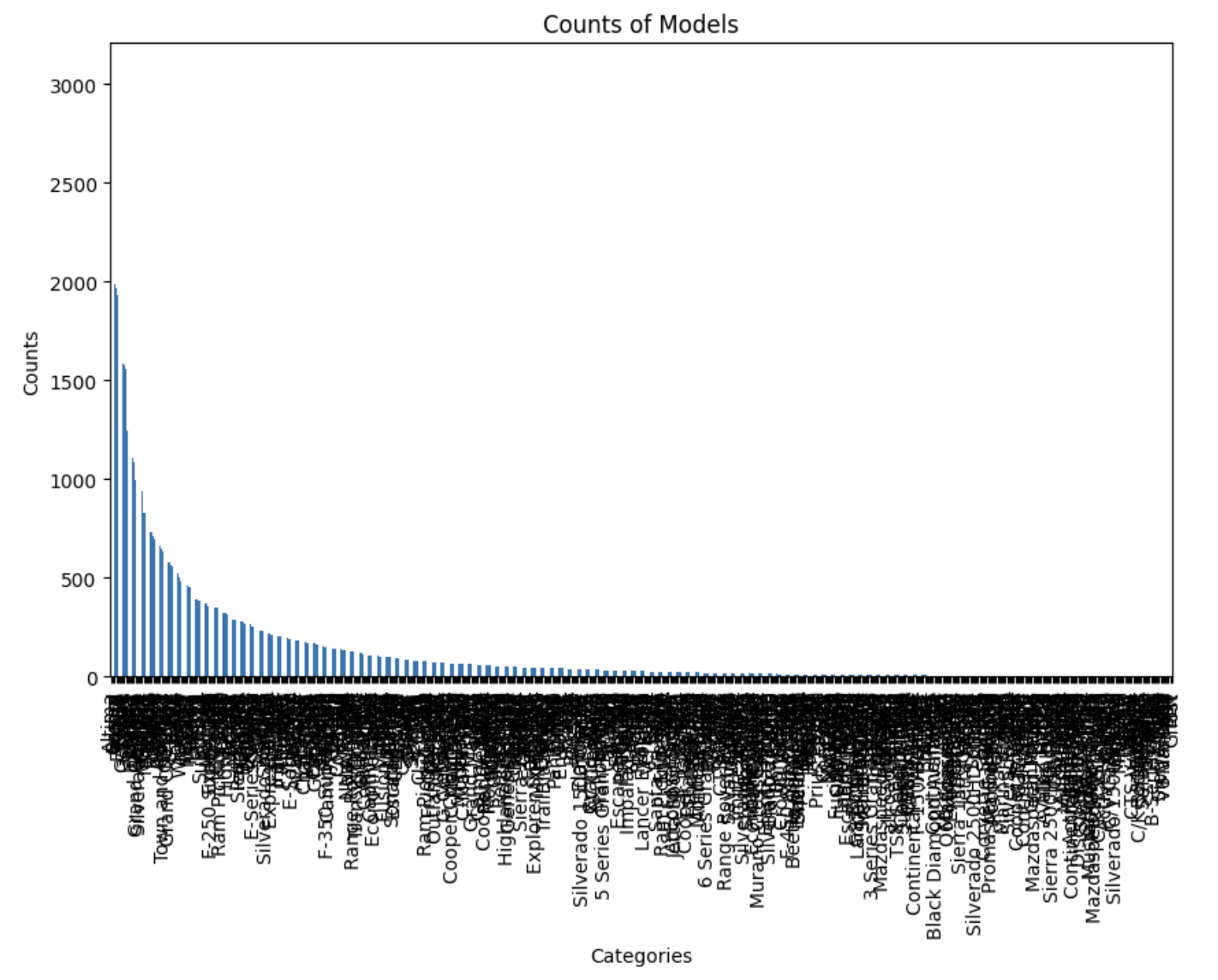
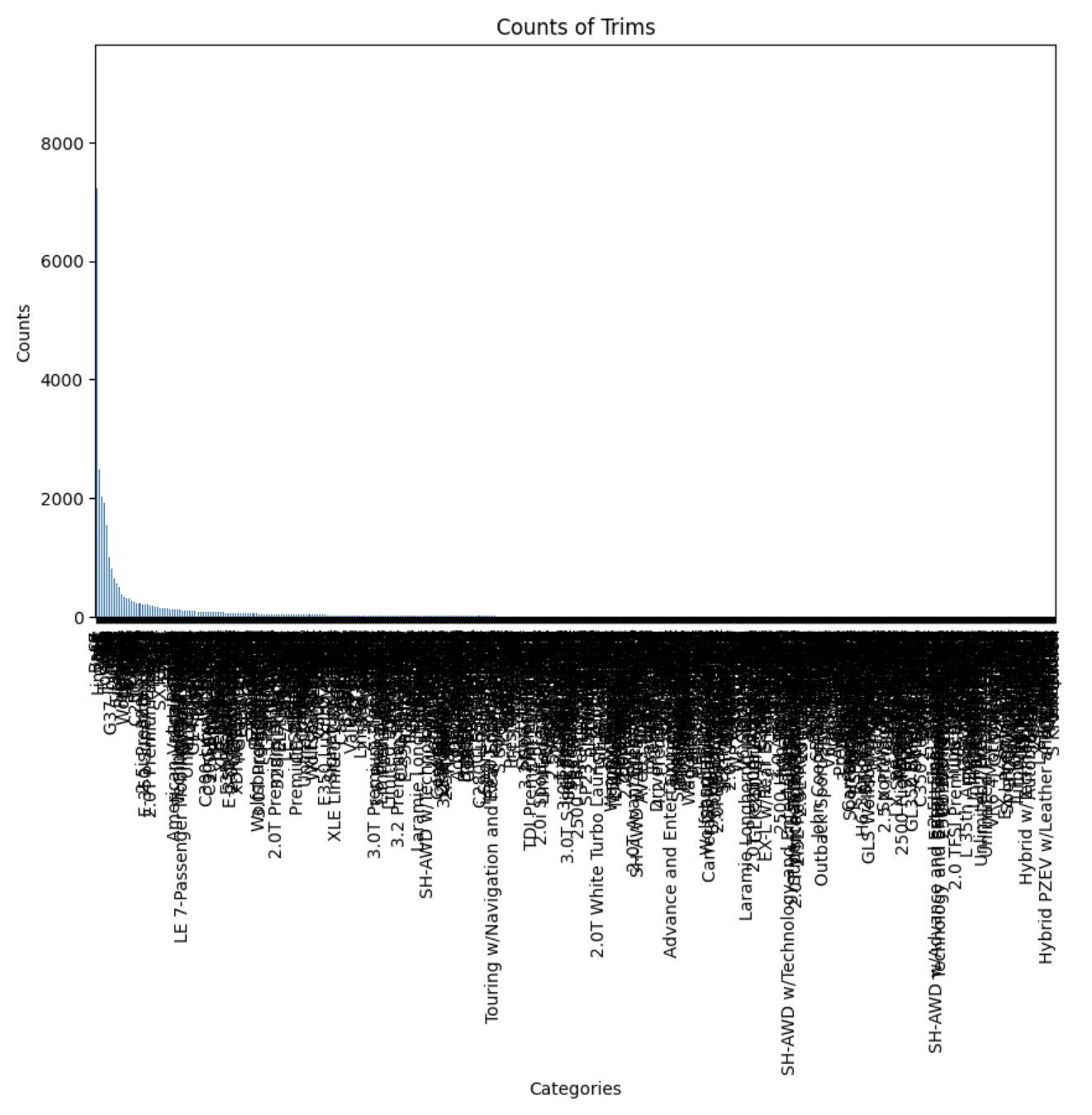
**Color:** color of the exterior of the vehicle  
**Interior:** color of the interior of the vehicle

**MMR:** The current market value of the vehicle

**Selling Price:** The target variable of the dataset, the price the car actually sold at

## Unique Values

We were interested in the distribution of the unique values in the qualitative data. We looked at the makes, models, and trims of the vehicles. Looking at the unique values gave us insight into if we felt the qualitative variable would be useful in a regression model. Below are three graphs showing the counts of the unique variables.



The make variable had a reasonable amount of categories, but the models and trim had 658 and 1154 unique values respectively. We planned to use one-hot encoding to make these variables numerical in our data cleaning process, so if we included ‘model’ and ‘trim’ we would have almost 2000 features. We decided that ‘make’ would be enough in the prediction and to cut the other two variables out.

## Feature Interactions

Before deciding which model to proceed with, we first made a scatterplot matrix containing the quantitative variables in the dataset after our data-cleaning process. Additionally, we generated a matrix of correlations between the same variables in the scatterplot matrix. From these matrices, we were able to determine some of the variables to generate models.

## 

## 

The scatterplots told us that using a linear model is out of the question, especially if we want to include ‘year’, ‘odometer’, or ‘condition’. When compared to ‘sellingprice’, the relationships of the other three appear to be non-linear, but none of them really fit a specific line. This made us look at models such as decision trees, random forest, and KNN.

It was clear that the mmr feature was highly correlated to the selling price. This concerned us in that our model may overweigh that feature. We decided to test the models with both MMR and without it to see the impact of the mmr feature.

# 

# Data Cleaning

In preparation for our data analysis, we conducted a thorough data cleaning process. Given the large size of our dataset, maintaining the dataset’s reliability and integrity was our top priority. Our approach began with addressing missing values, a common challenge when working with large datasets. Through the application of listwise deletion, we systematically removed rows containing missing values, thereby ensuring that our training and testing datasets were complete and suitable for subsequent analysis. This method, while straightforward, proved effective in handling missing data points, as it also facilitates the identification and removal of rows with other data abnormalities.

Furthermore, our data cleaning efforts extended to rectifying inconsistencies in variable names, particularly within the ‘make’ column. We encountered several instances where values were duplicates but lacked consistent naming structures, posing a potential challenge to the accuracy of our analysis. Standardizing the naming conventions allowed us to address this issue effectively, ensuring uniformity across the dataset.

Another important facet of our data cleaning process entailed dealing with outliers, which have the potential to skew analysis results and affect model performance. While we identified some outliers in our dataset, we determined that they were primarily the result of typographical errors or false entries. To address this, we employed corrective measures tailored to each scenario. For typographical errors, we aligned the corrections with the distribution of the dataset, ensuring that they remained consistent with the overall data patterns. In cases of false entries, we opted for another round of listwise deletion, removing these erroneous observations to maintain data quality and integrity.

In addition to addressing missing values, we utilized one-hot encoding to handle categorical variables such as make, transmission, body, state, color, and interior. This technique allowed us to transform these categorical variables into numerical format so they could be effectively utilized in our regression model. By creating binary columns corresponding to each category within these variables, we expanded our dataset with additional features that captured the nuances of each category. This process not only ensured compatibility with regression algorithms but also preserved the unique information contained within these categorical variables.

Overall, our thorough data-cleaning efforts were pivotal in setting the stage for our future analysis and predictive modeling initiatives. By addressing missing values, standardizing variable names, managing outliers, and using one-hot encoding for categorical variables, we created a quality dataset fit to train and test with.

# 

# Different Model Approaches

## Linear Regression Model

When deciding on a model to train with our dataset we felt that a linear model would not be the best at finding patterns in the data based on what we saw in the scatterplot matrix. We did train a linear regression model to see if it could perform at all as a baseline. We found that the linear model had a Mean Square Error (MSE) of 29394503, a Mean Absolute Error (MAE) of 3571.93, and an R2 of 0.696. This model was on average $3500 off from the true value. Not the best model for sure, but we had predicted that based on what we found with the scatter plots.

## Decision Trees

The next model we tried was a decision tree regressor. This model captured less of the training data than the linear model, with an R2 of 0.669, but it did have a better MAE of 3043.25. The MSE was also better at 26345349. Still not the greatest model, but it is predicting better than the linear model.

## Random Forest

The random forest model was essentially the decision tree model that captured more of the training data during training. The R2 value was 0.746, and the MAE was slightly better with a 2967.38. The MSE was also slightly better with a 25201333. This is the best predictive model so far, but it is still guessing about $3000 off the actual value.

## KNN

We tried a different approach using a clustering model like KNN, trying different values of n and eventually landing on n = 20. The KNN model performed the worst so far with an MSE of 48261042, an MAE of 4943.55, and an R2 of 0.39. It was worth trying but did not end up being a good model for our data.

## Cat Boosting Regression

The Cat Boosting model was the GBM model that effectively captured the most training data during training. The R2 value was 0.8522, and the MAE was slightly better with a 2173.45. The MSE was also slightly better with a 12249371. This is the best predictive model but still predicts about $2100 on average off the exact value.

## 

# Experimentation

## Models with the MMR Variable

When training the original models, we had left out the MMR variable because we felt it would impact the prediction too much. We decided to test each model with and without the MMR variable so we made two separate data frames. We tested the MMR variable with the two best-performing models without it. The results were a lot better than without it, which was what we had anticipated.

| Random Forest | Cat Boosting GBM |
| --- | --- |
| Mean Square Error: 2008710.03 | Mean Square Error: 2257492.65 |
| R Squared: 0.9747 | R Squared: 0.9728 |
| Mean Absolute Error: 924.63 | Mean Absolute Error: 890.63 |

We were hoping to be able to create a model that would not require the market value of the car to predict, but based on the MAEs of all of the models we have tested, it was not realistic, at least with good accuracy. Based on these experiments, we decided to include the MMR variable in our final model.

# Conclusions

We worked with a challenging dataset, with many missing values, mislabels for qualitative data, and over 500,000 data entries. We cleaned the data using many data-cleaning methods such as listwise deletion and standardization mapping. We also removed outliers in the ‘sellingprice’ target variable so the model was trained for a more normally priced car. When it came to one-hot encoding, we decided that the make of the car would be more important to include when encoding than both the model and trim. This prevented the model from having over 1500 input features.

We tested the cleaned data with numerous types of models: linear regression, decision tree, random forest, KNN, and Cat Boosted. After our experimentation, we determined that the best of these models is the Cat Boosted. The MAE of this model was 2173.45 and R2 of 0.8522.

We tested each model with two datasets, one that included the MMR variable, and one that removed it. We wanted to see how much the MMR variable affected the performance of the predictive models. Including the MMR variable helped the performance significantly. The Cat Boosted Regression model with MMR had an MAE of 890.63 and an R2 of 0.9728. A model that did not include it would be nice so we didn’t need the market value, just the other attributes of the car. After we looked at both, it was clear the model without MMR was not going to be good enough compared to the one with it.

# Link to dataset

<https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data?resource=download>