REPORT

1. What is your business proposition?

We are a medium-sized car dealer company that wants to improve our estimated price for a newly manufactured car.

Our business proposition is “We utilize the Manheim market report alongside car features including model, body type, and condition to forecast car prices. This approach sets a benchmark for companies to attract customers while optimizing profitability.“

2. Why this topic is interesting or important to you? (Motivations)

As college students, we recognize that car prices fluctuate over time. With our financial limitations, our goal is to find cars that fit within our budget constraints while offering desirable features. Additionally, we aim to pinpoint the most favorable months for purchasing.

Through the lens of an intermediary company positioned between manufacturers and customers, we delve into past data to uncover trends in car prices over time and identify popular selling features. By leveraging this analysis, we aim to forecast prices for newly manufactured cars. Moreover, adjusting pricing strategies to maximize profit and gain a competitive edge in the market.

3. How did you analyse the data? What conjectures you made? Which conjecture you used as the basis of developing your model? Why?

We performed dataset analysis by grouping features together with their respective selling prices to explore correlations. Our conjectures involve examining up to two features in conjunction with the selling price to clearly identify their relationships. Subsequently, these relationships are visualized using stacked bar graphs, pie charts, and correlation tables.

More specifically, our conjectures revolve around selling prices with numerical factors, categorial factors, and all. Furthermore, we also analyze monthly and yearly sales, as well as the popularity of exterior and interior colors and car body types with models.

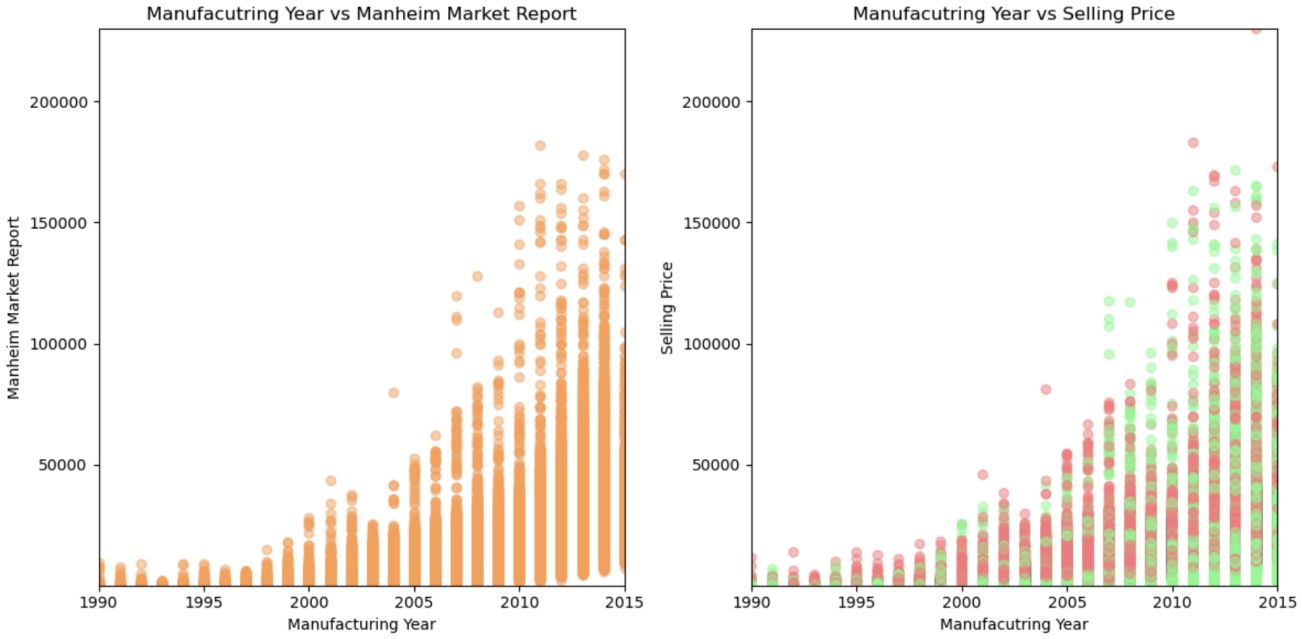
We utilized conjectures that all factors could influence the selling prices as the foundation for developing our model as our goal is to predict car prices based on all car features and market reports.

4. How does your analysis support your business proposition? (please include figures or tables in the report, but no source code)

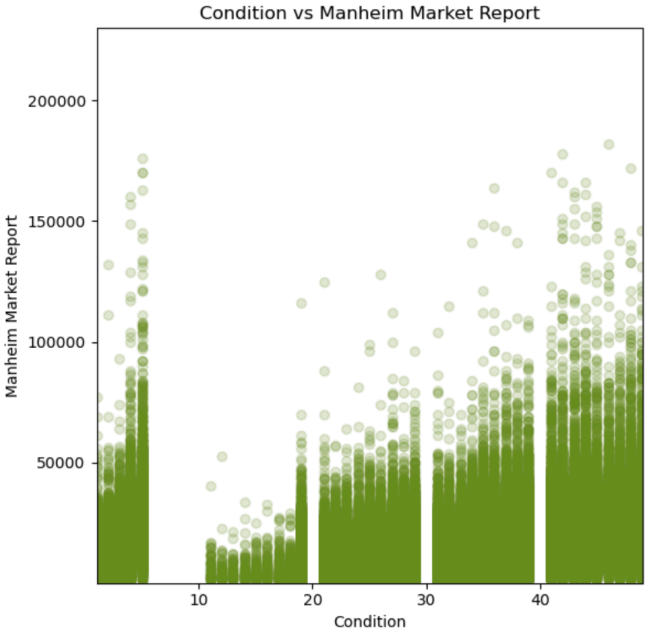
    a. Conjecture 1: All numerical factors influence the pricing

        i. Manheim Market Report vs Selling Price

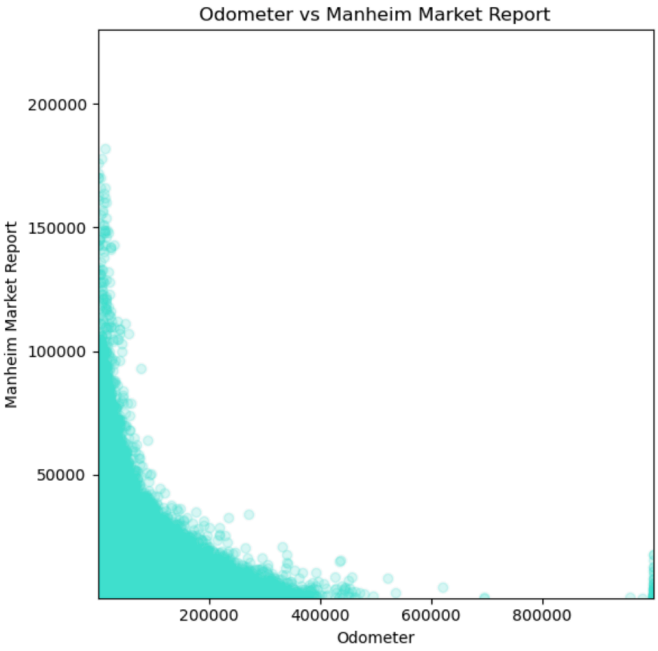
ii. Manufacturing Year vs Price



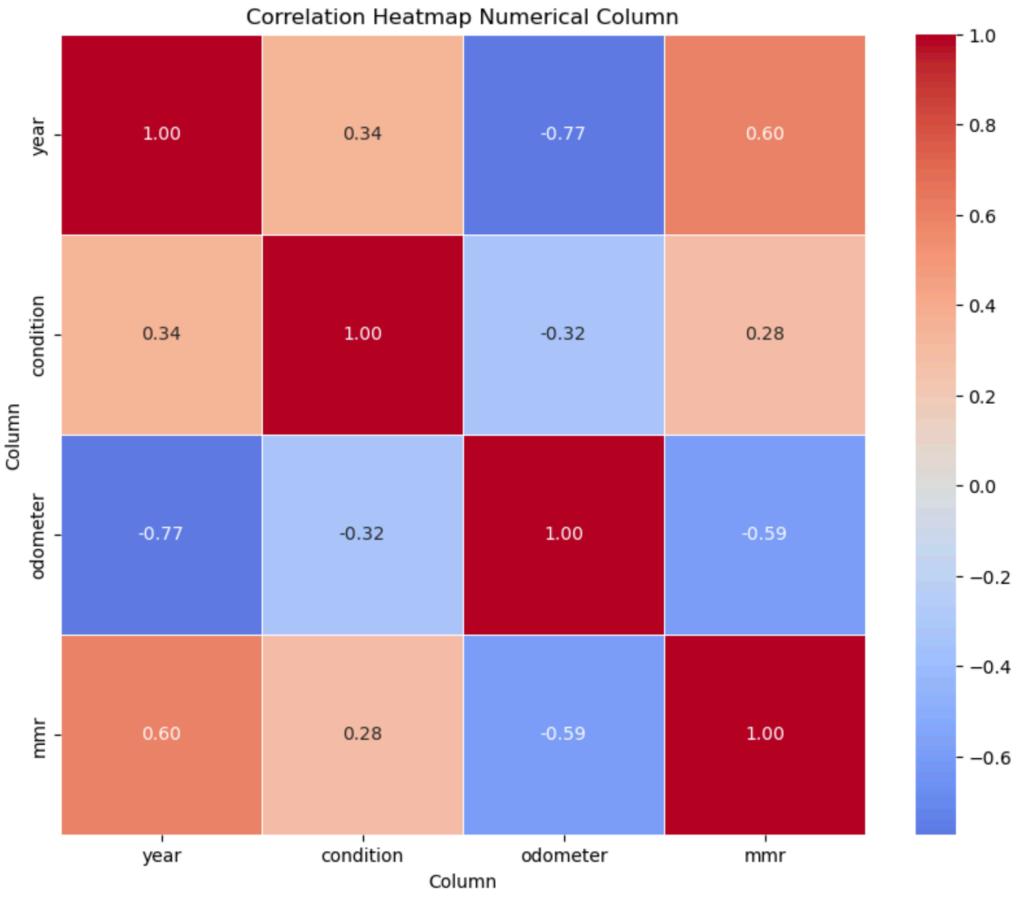
iii. Condition vs Price



iv. Odometer vs Price



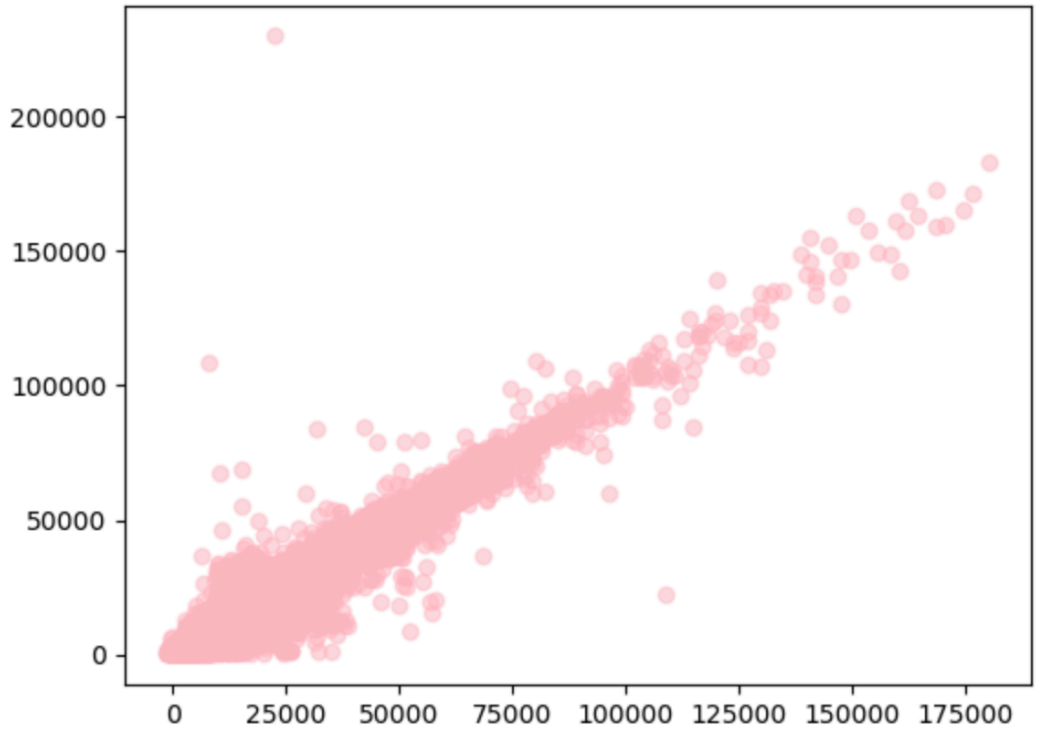
v. Correlation between all factors



vi. Train model of all numerical factors with the selling price

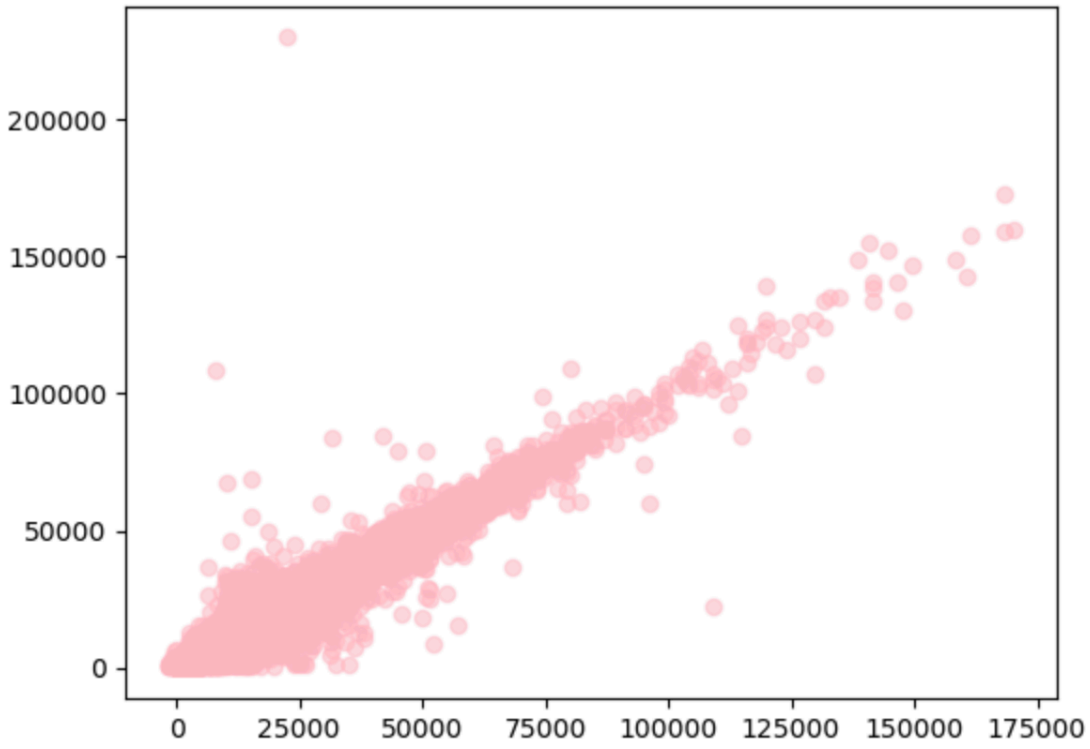
1. Train with 30% dataset

```
Score: 0.9657714914365501
```



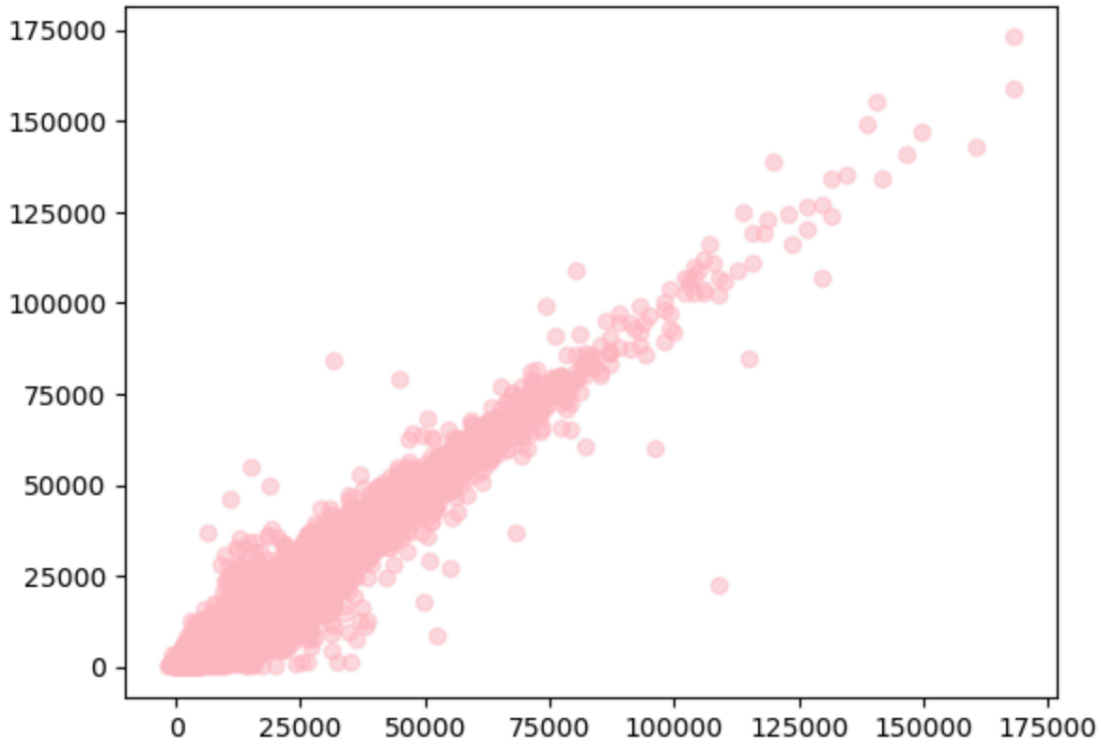
2. Train with 50% dataset

```
Score: 0.9641786342193437
```

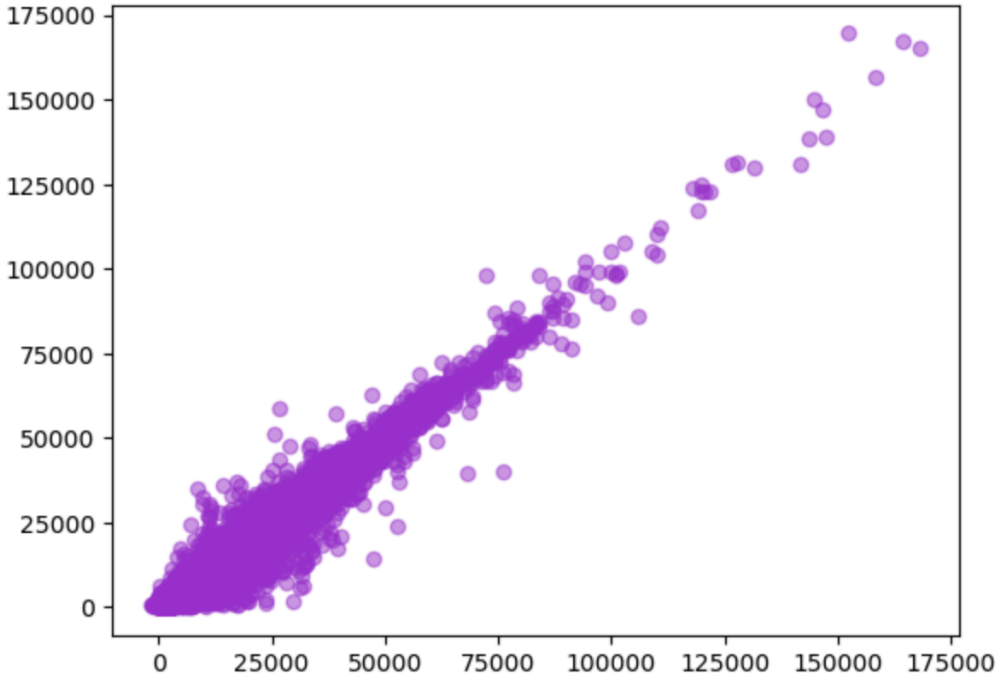


3. Train with 70% dataset

```
Score: 0.9676966401642308
```



4. Test the model

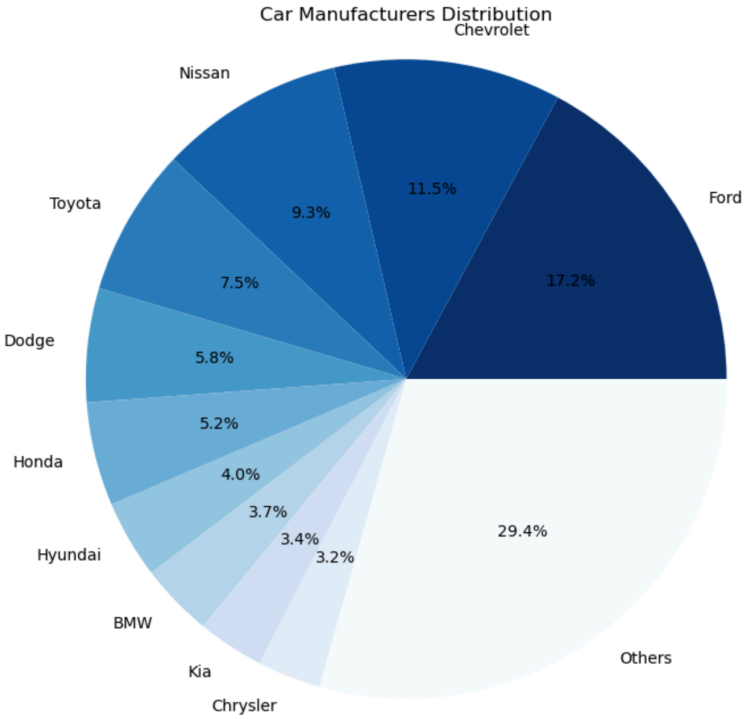



```
Linear Regression score: 0.9686613928409918
```

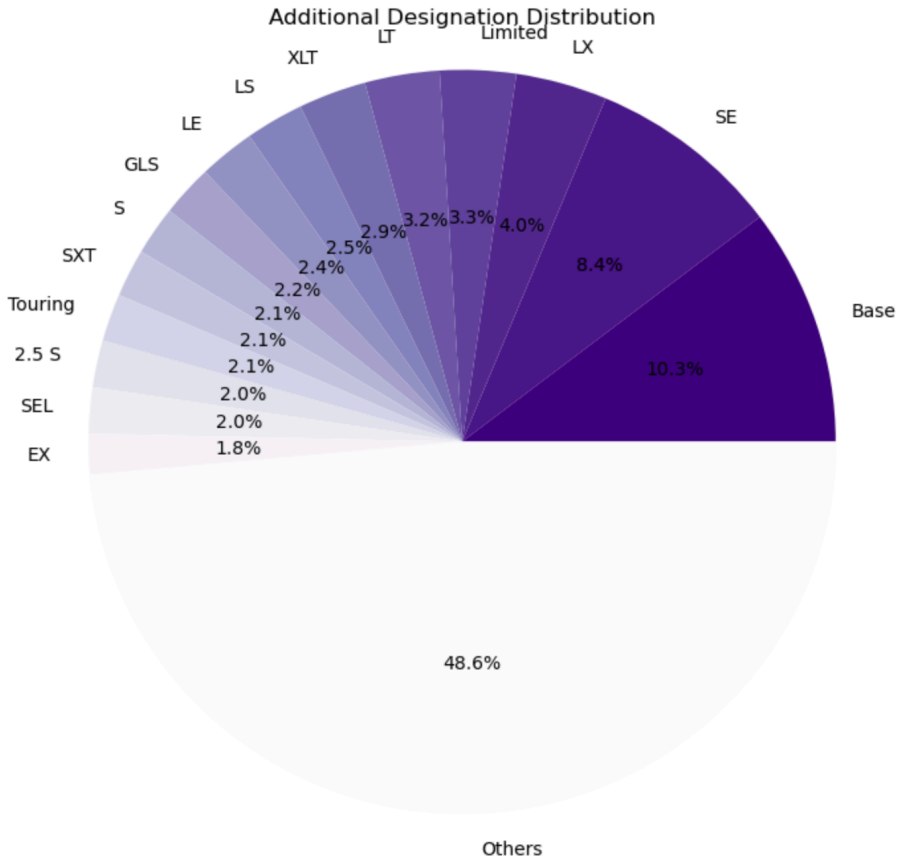
- Based on the visualization of conjecture 1, we saw a linear relation between the numerical factors with the selling point. When we analyze the model, the test result is pretty high (0.96866), meaning these are important factors that influence the pricing.

b. Conjecture 2: All categorical factors influence the pricing

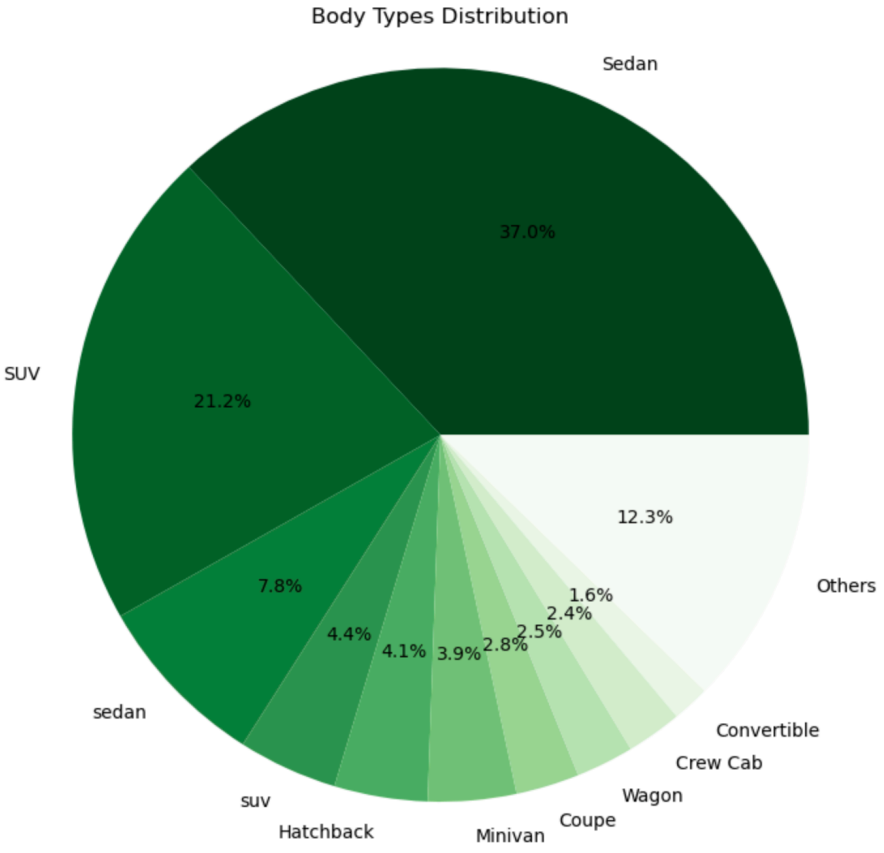
i. Car Manufacturers Distribution



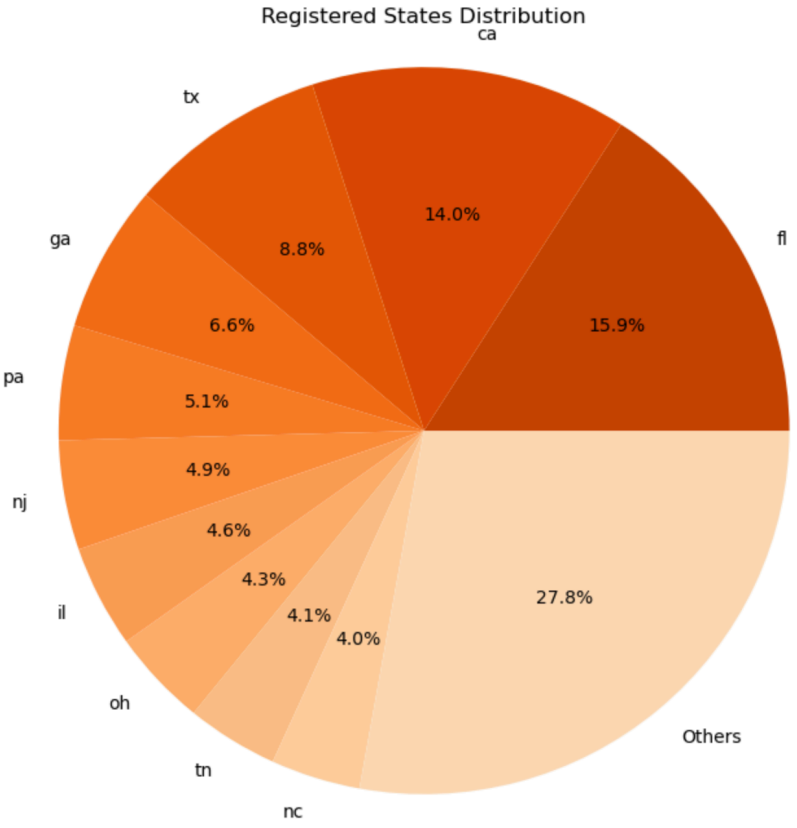
ii. Additional Designation Distribution



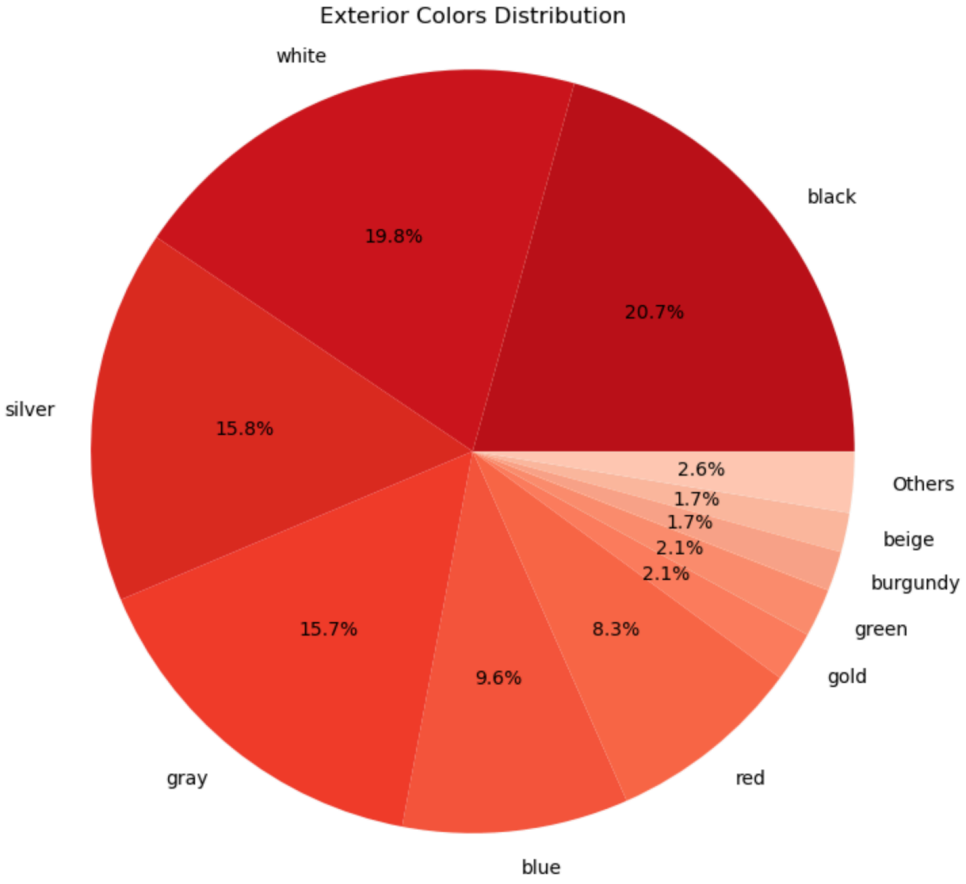
iii. Body Types Distribution



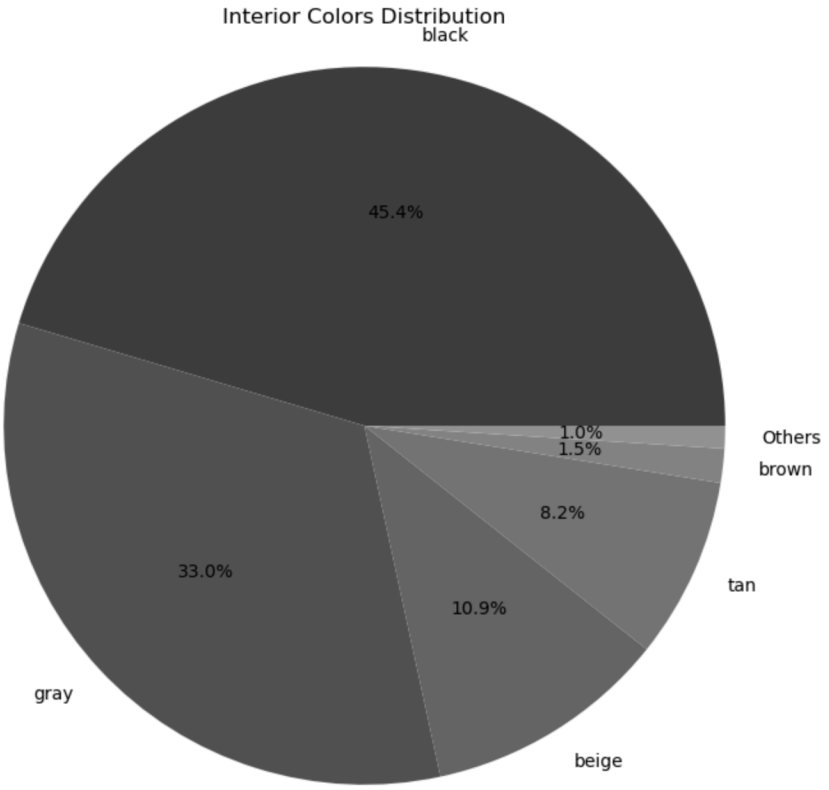
iv. Registered States Distribution



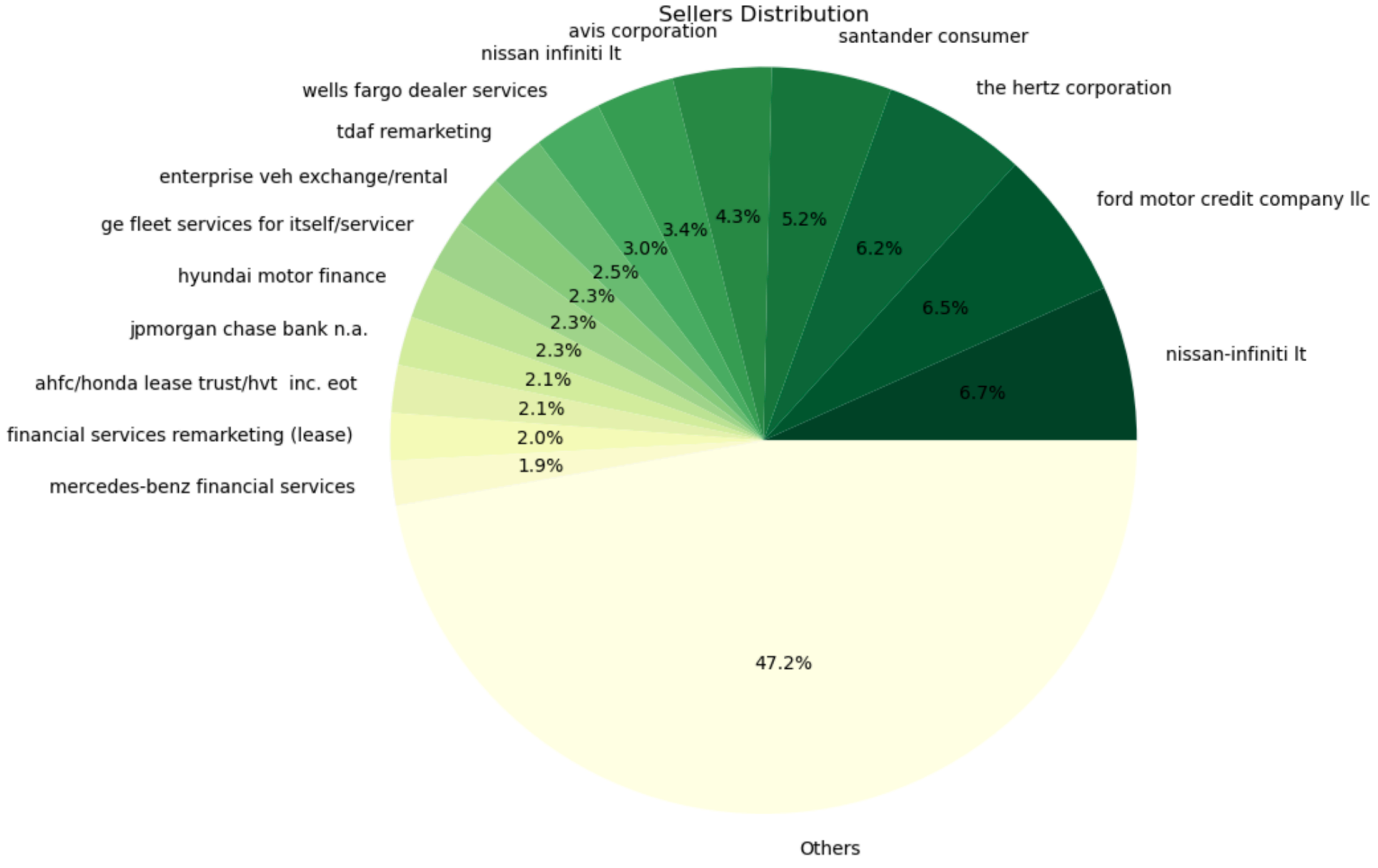
v. Exterior Colors Distribution



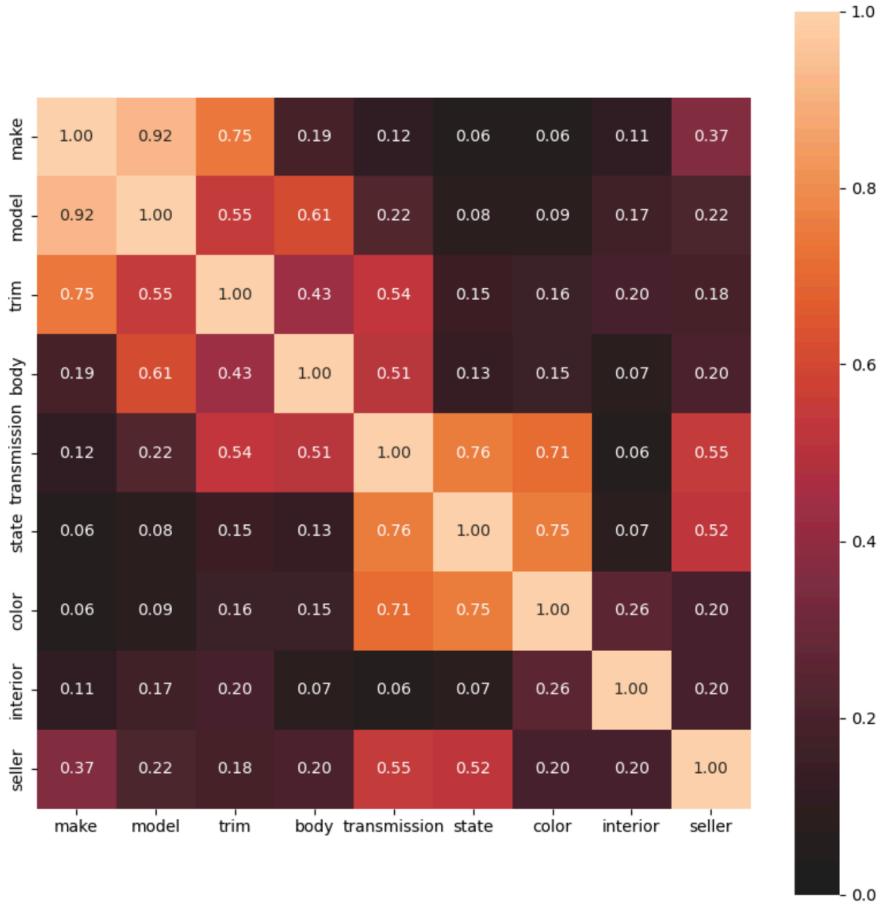
vi. Interior Colors Distribution



vii. Sellers Distribution

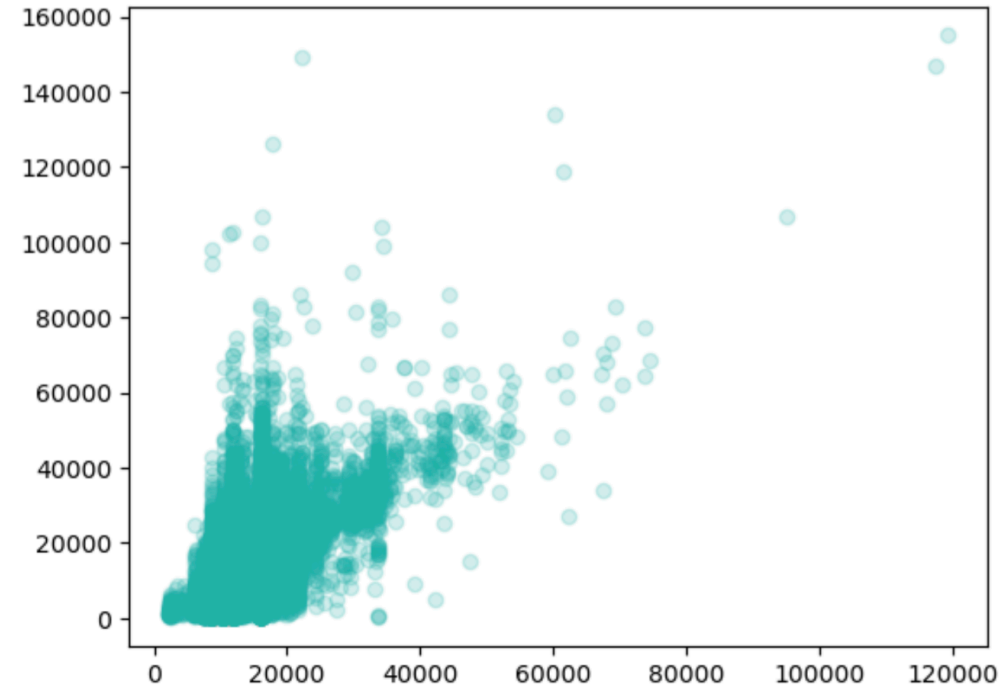


viii. Correlation between all factors



ix. Train model of all categorical factors with the selling price

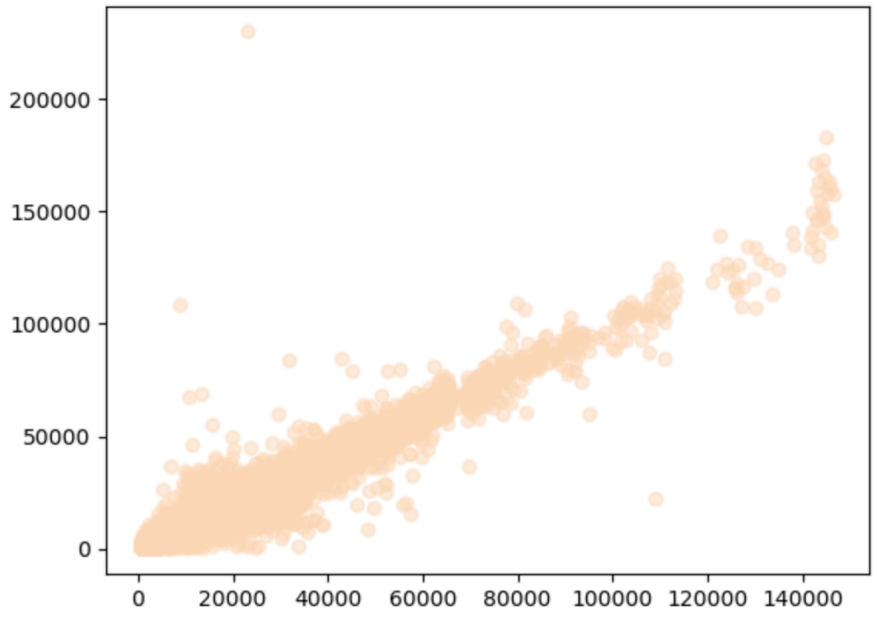
```
Random Forest R^2 score: 0.2984820448858492
```



- Based on the visualization of conjecture 2, we couldn't identify a relation between the categorical factors with the selling point. When we analyze the model, the test result is pretty low (0.29848), indicating some of these factors are not important factors that influence the pricing.

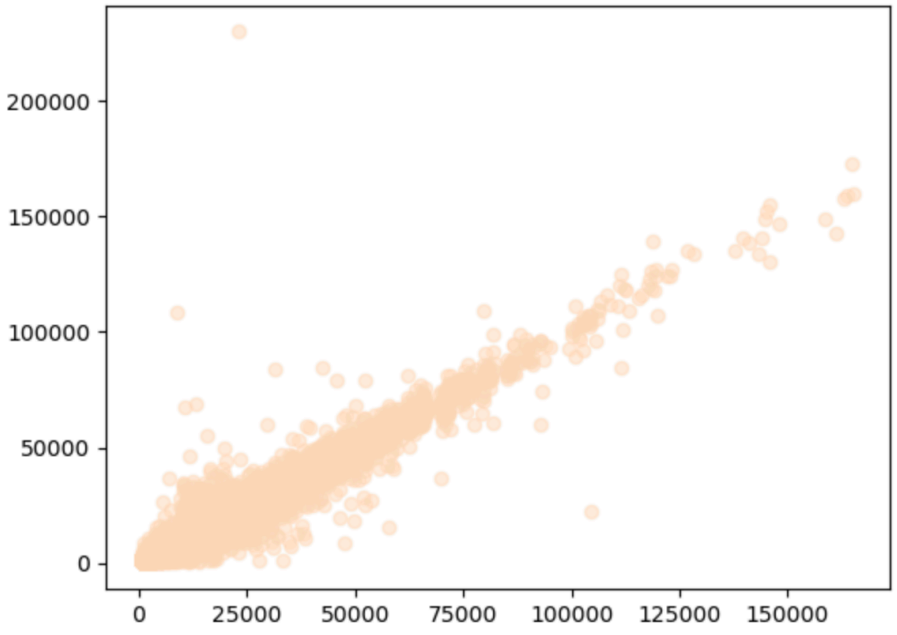
c. Conjecture 3: All factors influence the pricing
   - Based on the above conjectures, we want to estimate the price when using both numerical and category factors.
   - Train model of all factors with selling price
     - Train with 30% dataset

```
Random Forest R^2 score: 0.972061345806958
```



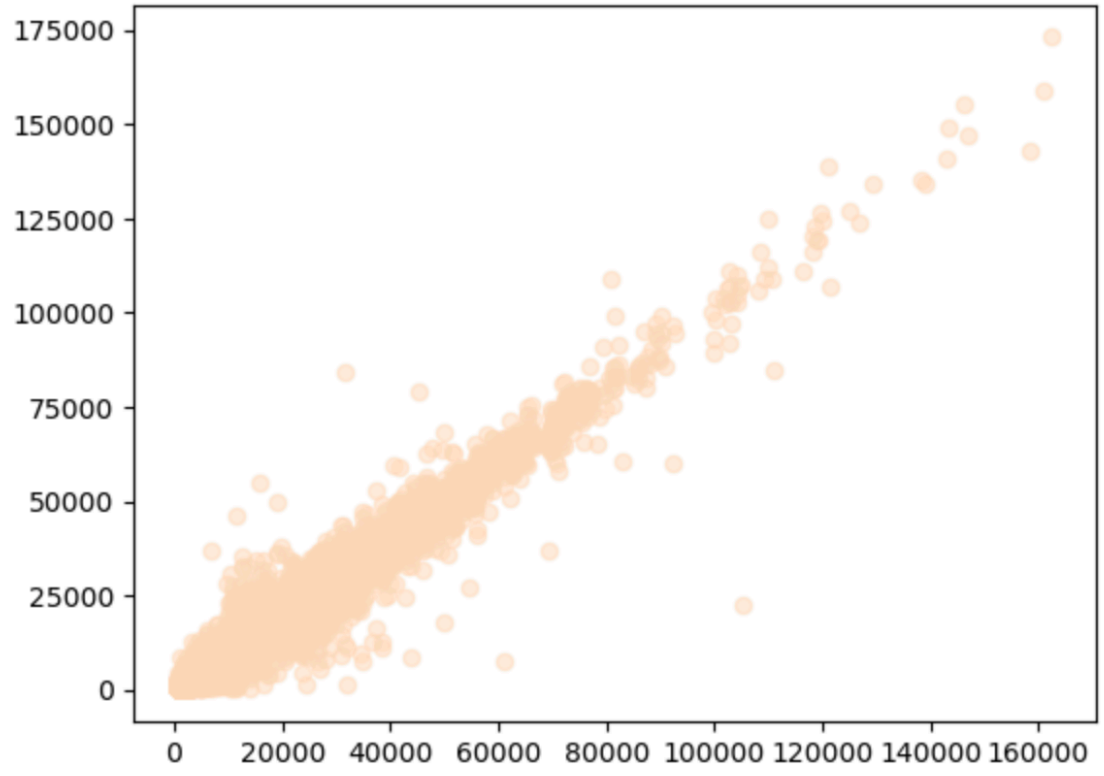
     - Train with 50% dataset

```
Random Forest R^2 score: 0.9710564398946824
```

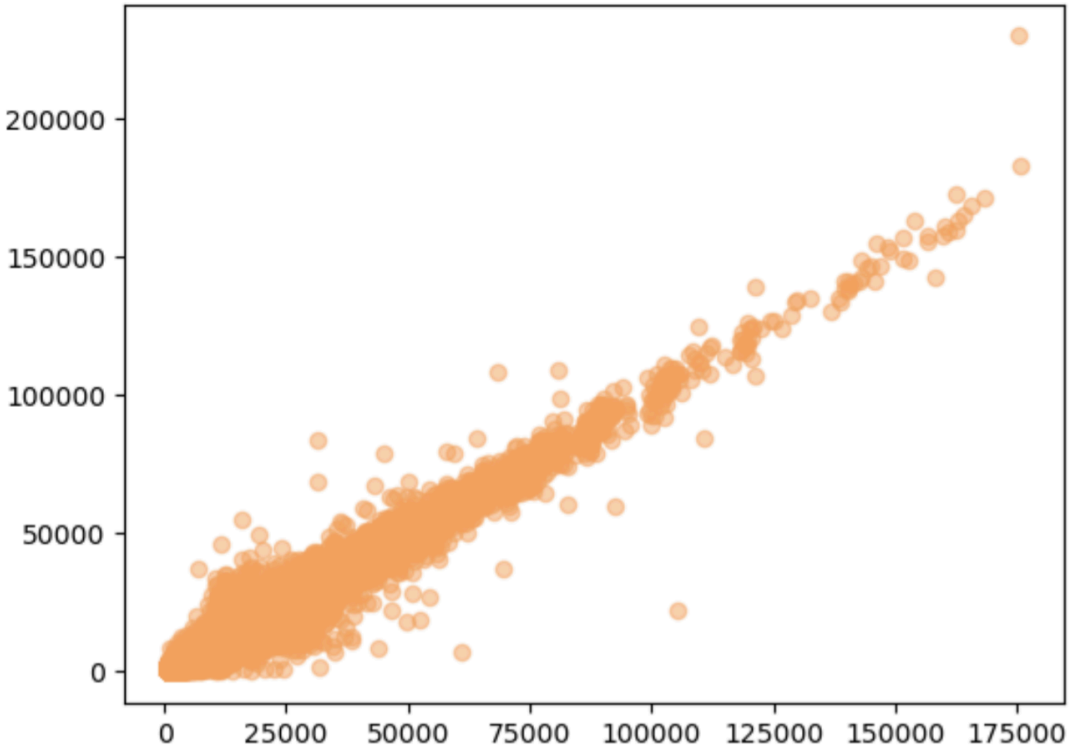


- Train with 70% dataset

```
Random Forest R^2 score: 0.9745367078667203
```



- Test the model




```
Random Forest R^2 score: 0.9772568458402271
```

- The test result is even higher than that of the numerical factors (0.97726 > 0.96866), meaning some of these categorical factors are important factors that could influence the pricing. Also, this model seems like a good prediction model for newly manufactured cars.

(5) (4 points) How does the model tie in with the business proposition?

The model ties directly to the business proposition, as in we train based on all of the features except the selling date. This allows us to leverage the model effectively when a newly released car enters the market, enabling us to predict its selling price in line with current market conditions. By integrating our predictive model into our pricing strategy, we can better cater to customer demands and maintain competitiveness in the market.

(6) (bonus 5 points) How did your team work together as a group from ideation to implementation? Write in one page.

Our team's idea was to create a model with the input from the car features to predict the selling price and compare it with the market report to see whether the predicted price can generate profit or not.

Duyen commenced by analyzing the dataset, focusing on the relationship between the selling price and MMR (Market Report) to discern any correlations. Additionally, she scrutinized the distributions of the categorical dataset to gain further insights.

Dan delved into exploring the relationships between pairs of features and the selling price, generating correlation tables to facilitate visualization and understanding of these connections.

Drawing upon the insights gleaned from the initial analyses, Phong developed several predictive models tailored to our dataset's nuances. These models were designed to effectively leverage the correlations identified between numerical and categorical features, ensuring a comprehensive approach to price prediction.

Upon making conjecture 1, we can see that most of the numerical features have clear relations so we decided to use a regression model to predict the price based on numeric data only.

When making conjecture 2, we couldn’t see a clear relation between the categorical factors and the selling price. After examining the model, the result shows that our conjecture was wrong.

With conjecture 3, since the dataset has both categorical data and numerical data, we decided to use Random Forest Regressor for the correlation between categorical and numerical ones. These turned out to be a good fit for the prediction.