PSTAT 126 Final Project Report

Aliza Samad, Michelle Brataatmadja, Madeleine Bulman, Isaiah Singer

2024-11-29

Introduction

The used car market plays a vital role in the economy, providing an affordable alternative to new vehicles and offering flexibility for buyers across various income levels. For students and young professionals, purchasing a used car is often the most practical choice, balancing budget constraints with the need for reliable transportation. As economic factors and consumer priorities evolve, understanding the dynamics of resale value becomes increasingly important.

This study explores the intricate relationships between factors that influence used car pricing, with a focus on variables such as car age, mileage, engine capacity, and transmission type. These factors are closely tied to maintenance and operational costs, which are crucial considerations for cost-conscious buyers like students. By examining these variables and their interactions, we aim to shed light on the predictors of resale value and uncover patterns that can help consumers make informed purchasing decisions.

Specifically, this research addresses the following questions:

1. How does car age influence the resale price, considering mileage, engine capacity, and transmission type as indicators of maintenance?
2. How does power correlate with the price of the car, and does this relationship differ across car brands?

In order to investigate these research questions, we came up with two sets of testable hypotheses:

1) Hypothesis Set 1

H_0 : The age of a car or any of its interactions with mileage, engine, and transmission do not have an effect on resale price

H_A : The age of a car or any of its interactions with mileage, engine, and transmission do have an effect on the resale price

2) Hypothesis Set 2

H_0 : The relationship between power and price does not vary by car brand

H_A : The relationship between power and price does varies by car brand

The Second Hand Car Price Data from Kaggle includes the price and specifications of 100 cars. It includes the following variables which have been separated into the predictors and response variables:

- Predictor Variables: Car ID, Brand, Model, Year, Kilometers Driven, Fuel Type, Transmission, Owner Type, Mileage, Engine, Power, Seats
- Response Variable: Price

Note: We separated brands into premium brands and economy brands in a new variable called ***`Segment`****. Premium brands have been defined as Audi, BMW and Mercedes, which we selected based on outside research from CARFAX. Economy brands are all other brands in our data set.*

1.1 Regression Model Application

1.1.1 Simple and Multiple Linear Regression

```
# simple linear regression
model1.1 <- lm(Price ~ Year, data=data)

model1.2 <- lm(Price ~ Power, data = data)

# multiple regression
model2.1 <- lm(Price ~ Year + Mileage + Transmission + Engine, data=data)

model2.2 <- lm(Price ~ Power + Segment, data=data)

# post-stepwise selection
## referred to as model 1
model1_official <- lm(Price ~ Engine + Transmission + Mileage + Year +
    Transmission:Mileage + Engine:Mileage + Engine:Year + Mileage:Year +
    Engine:Transmission + Engine:Transmission:Mileage + Engine:Mileage:Year,
    data = data)

## referred to as model 2
model2_official <- lm(log(Price) ~ I(Power^-0.95)*Segment, data = data)
```

Models 2.1 and 2.2 correspond to research question 1 and 2 respectively. For RQ 1, we wanted to focus on key variables that would represent what we considered to be "maintenance" factors. These predictors variables originally included `Mileage`, `Transmission`, `Engine`, `Fuel_Type`, `Owner_Type`, and `Kilometers_Driven`. In the effort to prevent over-fitting, we decided to limit these predictors to just three, in addition to `Year`. Our choice in variables was made by examining our preliminary EDA (see appendix), specifically Figures 2 and 6. For RQ 2, selecting the variables was simple since we wanted to observe the relationship between `Power` and `Brand`. Moving forward, we will be conducting stepwise selection of the ideal model, including interaction terms.

1.1.2 Model Selection and Evaluation

Model 1:

Comparing AIC/BIC

```
## [1] "original model:"
```

```
## [1] 2002.523
```

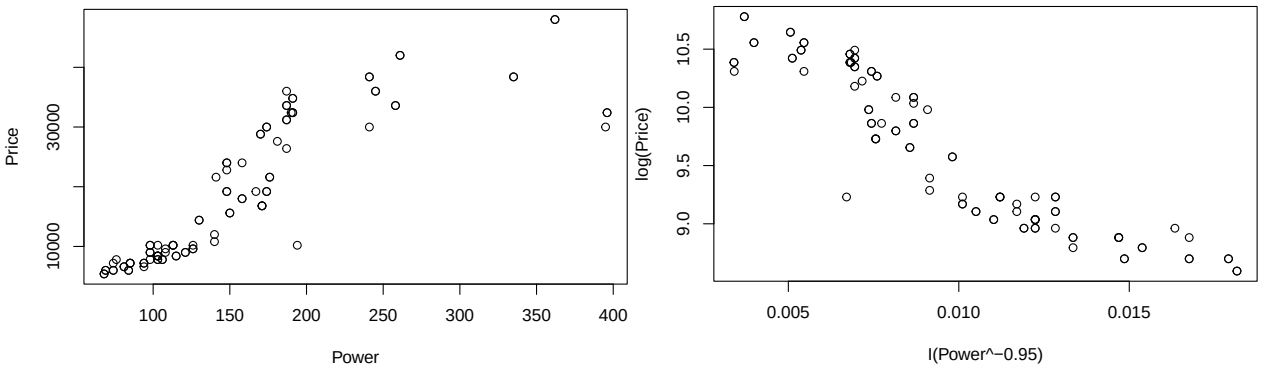
```
## [1] 2046.81
```

```
## [1] "stepwise model:"
```

```
## [1] 1998.454
```

```
## [1] 2032.321
```

Model 2:



Comparing AIC/BIC (same for original & official):

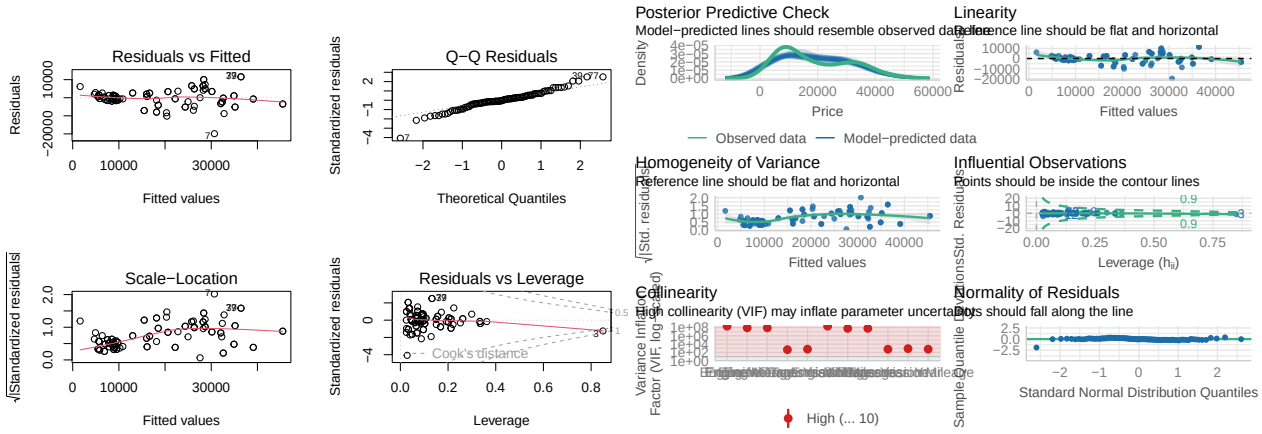
```
## [1] 1931.07
```

```
## [1] 1944.096
```

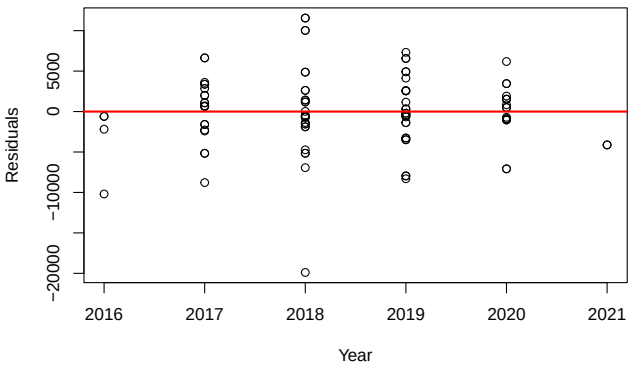
```
##
##  Shapiro-Wilk normality test
##
## data:  model2_official$residuals
## W = 0.97437, p-value = 0.04815
```

1.2 Diagnostic Checking

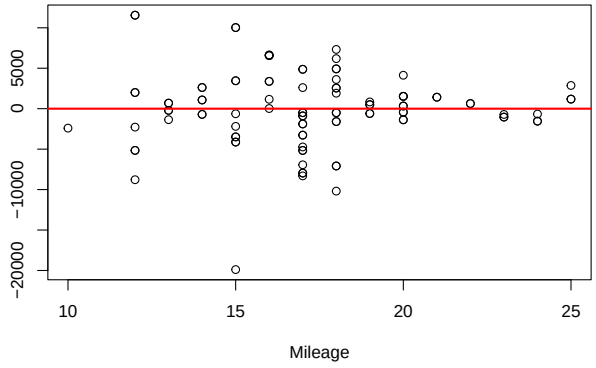
Model 1:

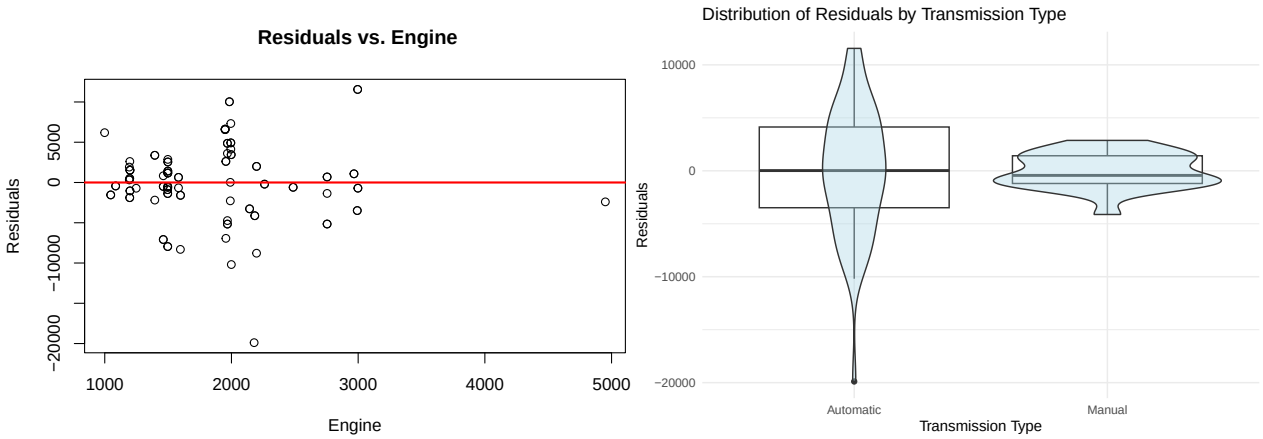


Residuals vs. Year



Residuals vs. Mileage





For models 1 and 2, we can assume all observations are independent of each another.

To check the normality assumption we can look at the Normal Q-Q plot and the Shapiro-Wilks Test. Looking at the Q-Q plot we can see that the majority of the points are on the line, however the points do move away from the line on both ends. The Shapiro-Wilks Test resulted in a p-value $= 0.0008075 < 0.05$, suggesting that the errors do not follow normal distribution. As a result, the normality of residuals plot shows a residuals straying away from the line at either end even though they are centered around the line for the bulk of the plot.

There could be several reasons this is the case. The first reason could be that the predictors have a non-linear relationship with `Price`. In order to test this, we plotted all the predictors against the residuals of our model. Based on the analysis of all these graphs, there do not seem to be any detectable patterns and the data seems to be relatively random. Thus, no transformations have to be conducted. The only anomaly is the outlier in the Residuals vs. Engine plot, with a residual plotted at the `Engine` value 5000. This brings us to the second reason: outliers. When looking at outliers for Model 1 there is at least one significant outlier (point 7) and several other flagged observations in the influencePlot for model 1 (see section 1.2.4).

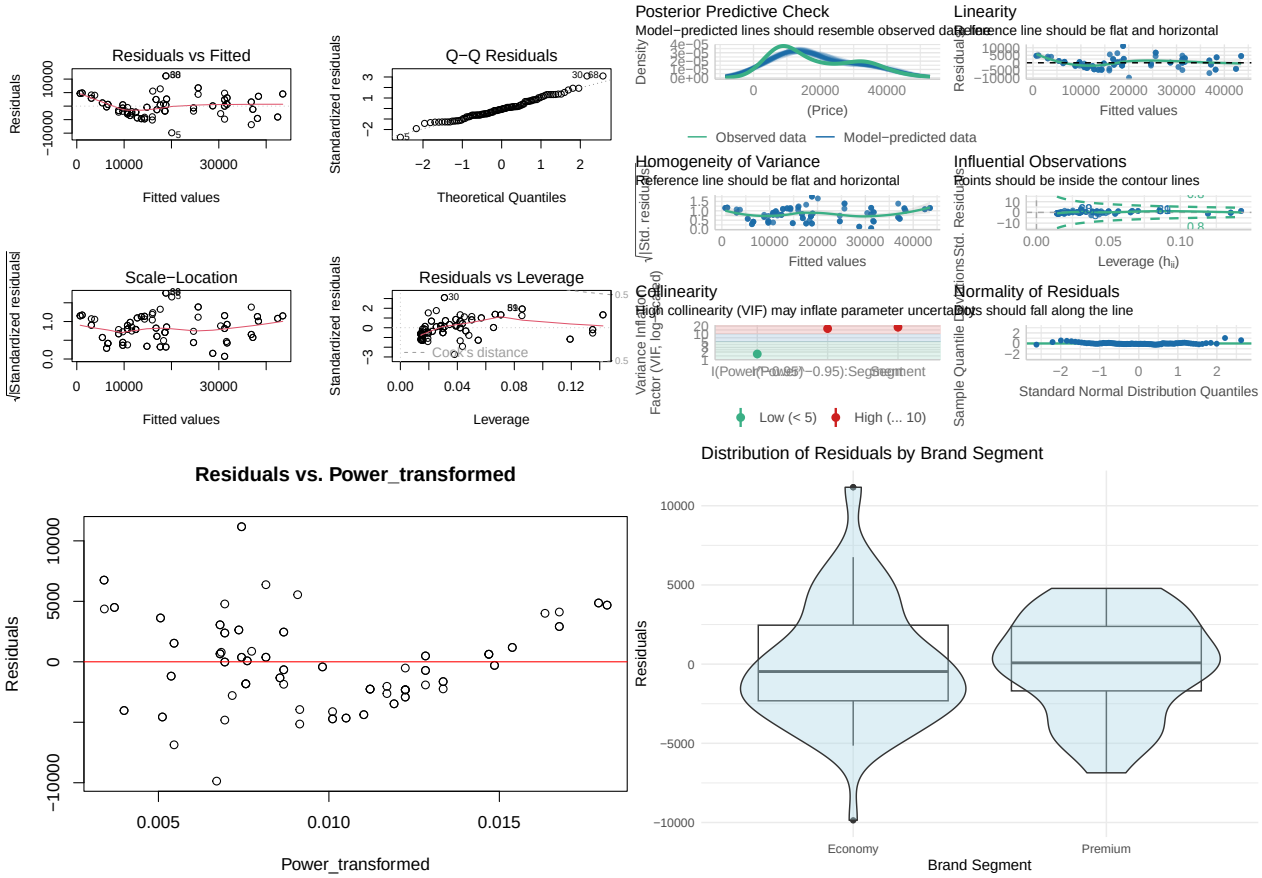
To check linearity we can look at the linearity plot under 1.2.2. The reference line does have slight curvature, however it does closely fit the horizontal line. Thus, this suggests that the linearity assumption is met.

The homogeneity of variance graph shows low heteroscedasticity; however, looking at the fitted vs residuals graph we see that there is a slight cone shape as fitted values increase, which suggests that variance might not be constant. Again, the presence of any unequal variance is most likely due to the influence of outliers.

Though the collinearity appears to be very high in this plot, this is due to the presence of interaction terms. The VIF test conducted below under section 1.2.3 reveals that the main effects are not collinear.

Thus, the assumptions for model 1 have been met.

Model 2:



Looking at the linearity plot which compares fitted values and residuals, the reference line is mostly flat and horizontal, but it does curve slightly at the tips. Nevertheless, the reference line does suggest that the linearity assumption is met.

Looking at the fitted vs residuals graph, we see that there is a slight cone shape in the points which suggests that there is not a constant variance. This is most likely due to the presence of outliers or other influential points. However, we can also observe a slight pattern in the Residuals vs. Power_transformed graph. Despite this pattern, this is the best fit transformation for linearity and normality after having attempted several other transformations on `Mileage` including a log, square root, and inverse.

The Q-Q plot for model 2 suggests that the majority of residuals fall along the line with the exception of a few points at each end. The Shapiro-Wilks test resulted in a high W, but a p-value of $0.04815 < 0.05$ which suggests that the normality assumption is not met. This is the largest p-value we could accomplish after transforming our variables, suggesting that there may be some external factors at play (such as outliers). The residual plots of both predictor variables show the bulk of the residuals centered around 0, as we would expect.

Finally, the VIF analysis under section 1.2.3 shows that despite the high multicollinearity in the graph above, there is no collinearity between the main effects `Power` and `Segment` (collinearity is due to interaction terms).

1.2.3 Multicollinearity

```
## [1] "model 1:"
```

```
##       Engine Transmission      Mileage         Year
##     2.179279     1.228469     1.876988     1.153077
```

```
## [1] "model 2:"
```

```
## I(Power^-0.95)        Segment
##       1.569709       1.569709
```

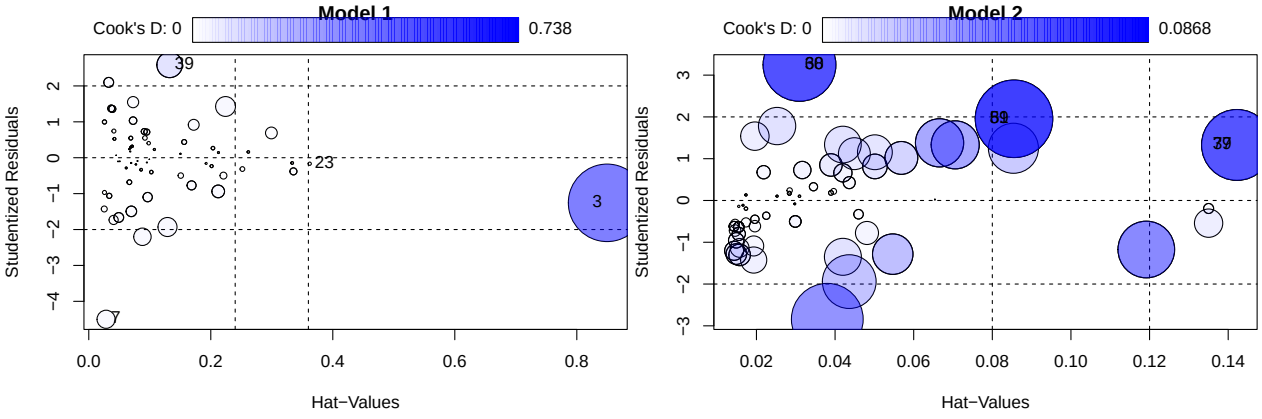
1.2.4 Outliers and Influential Points

```
##       StudRes        Hat       CookD
## 3  -1.2563705 0.84965036 0.738493365
## 7  -4.4997514 0.02836377 0.040415735
## 23 -0.1668405 0.36218014 0.001331903
## 39  2.5860343 0.13254536 0.079984483
```

```
##    rstudent unadjusted p-value Bonferroni p
## 7 -4.499751         2.0951e-05    0.0020951
```

```
##     StudRes        Hat      CookD
## 30 3.245133 0.03095406 0.07650149
## 39 1.329149 0.14224246 0.07266033
## 51 1.955028 0.08551881 0.08680599
## 68 3.245133 0.03095406 0.07650149
## 77 1.329149 0.14224246 0.07266033
## 89 1.955028 0.08551881 0.08680599
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 30 3.245133          0.0016216      0.16216
```



1.3 Interpretation of Results

1.3.1 Interpret Coefficients

Model 1:
Research Question 1:

The intercept shows that when all predictors are at zero and Transmission is automatic, the baseline price for a used car is $7,311,000. This does not make sense in context since you cannot have a `Year` 0, and you also cannot have 0 engine capacity.

The coefficient for `Year` is approximately -3581.75. In other words, for every year increase in age, the resale price decreases by $3581.75 units per year.

The coeffcient for `TransmissionManual`is priced on average 86,622.76 less than automatic transmissions. This shows that used cars with manual transmission tend to have a much lower resale price.

The coefficient for mileage is 47,470,000. In other words, for every unit increase in mileage, the price of a used car increases by $47,470,000. Since higher mileage typically does not increase the value of a used car, this result likely reflects other factors at play, such as newer vehicles with high mileage, which may obscure the usual relationship between mileage and resale price. Additionally this unreasonably high rate of price increase does not make sense in context.

The coefficient for engine is approximately 5,574.09. For every unit increase in engine capacity, the price of a used car increases by $5,574.09. This implies that a stronger engine has a positive correlation with the price of a used car.

The coefficient for Mileage:Year is 284.90 units. This implies that for each added year of age, the relationship between mileage and price decreases by $284.90. In other words, as a car gets older, the increase in price associated with the coefficient of mileage becomes less significant

The coefficient for Engine:Year is -2.77, which shows that for each added year of age, the relationship between engine and price decreases by $2.77. While it does show a decreasing trend, the p-value is not significant enough to draw any conclusions from.

The coefficient of Engine:Mileage:Year is 39.48. This shows that for each added year of age, the engine power and mileage increase price by $39.48 Once again, since the number is not significant, conclusions can't be drawn from it.

Model 2
Research Question 2:

Intercept:
The price for an Economy-brand car with a transformed horsepower of 0 is $31,400. In context, this doesn't make sense since no car can have a power of 0, so it acts as more of a baseline than a true interpretable intercept.

`I(Power^-0.95)`:
For every one unit increase in `I(Power^-0.95)`, `Price` decreases by $1,689,727. Since the transformed power variable has an inverse relationship to the raw `Power` variable, we can generally conclude that as `Power` increases, so does `Price` (assuming brand segment is held constant).

`SegmentPremium`:
In comparison to the base line `SegmentEconomy`, `SegmentPremium` results in the `Price` being $26,182 more expensive than the baseline, assuming `I(Power^-0.95)` is held constant.

`I(Power^-0.95):SegmentPremium`:
For Premium cars, the slope of the relationship between `I(Power^-0.95)` and `Price` decreases by an additional $2,105,594 compared to Economy-brand cars. In context, this means that a one-unit increase in `Power` results in a much steeper `Price` increase compared to Economy-brand cars.

1.3.2 Model Fit

Model 1:

Model 1 has an R^2 value of .8488 which shows that the model can explain 84.88% of variance in price. Since this value of R^2 is relatively high *and* we have accounted for multicollinearity, it implies that this model may be a good fit for the data. The adjusted R^2 value is .8299 which shows that after accounting for model complexity, the model still explains 82.99% of variance in price. Since the adjusted-R^2 value is close to R^2, overfitting does not seem to be an issue in the model.

Additionally, in an earlier step we compared the AIC and BIC values of the model before performing stepwise selection to our model (after stepwise selection). Our model has both a lower AIC and BIC value.

The value of the RSS is 4951, which relative to the scale of the dataset, is a small value. With that being said, the inclusion of the predictors on the response variable help explain the variance in price. Therefore, it can be said that this regression model is a good fit for the dataset.

Model 2:
Similar to Model 1, Model 2 has a high R^2 value (0.9096) and adjusted-R^2 value (0.9067). This shows that the model can explain 90.96% of the variance in price, which is a very high amount. Again, we have accounted for multicollinearity, so the high R^2 value does not ring too many alarm bells. Additionally, the earlier AIC/BIC analysis post-stepwise selection yielded the lowest AIC/BIC values possible considering the chosen predictor variables. com This information combined with the relatively low RSS of 3666 suggests that this regression model is also a good fit for the dataset.

1.3.3 Statistical Significance

Model 1
From the model summary, we can see that `Year` has the largest p-value of 0.734241. At the $\alpha = 0.05$ significance level, the p-value for `Year` is not statistically significant. Now, let us take a look at the significance of interactions between `Year` and `Engine`, `Transmission`, and `Mileage` on `Price`. The p-value for `Year:Engine` is 0.624141 and the p-value for `Year:Mileage` is 0.628107. Both will be rejected at the $\alpha = 0.05$ significance level. `TransmissionManual:Year` was removed from the final model during step-wise selection.

Given that none of the interactions between `Year` and the other predictor variables is significant *and* the direct impact of `Year` on `Price` is not significant, we can conclude that we **do not** have sufficient evidence to reject the null hypothesis that used-car age does not significantly influence resale price, either directly or through its relationship with other maintenance factors.

Model 2
From the model summary, we can see that the p-value for `I(Power^-0.95)` is $2 * 10^{-16} \approx 0$. At the $\alpha = 0.05$ significance level, the p-value for `I(Power^-0.95)` is statistically significant, meaning that `I(Power^-0.95)` plays a statistically significant role in determining `Price`. The p-value for `SegmentPremium` is also $2 * 10^{-16} \approx 0$. Hence, the p-value for `SegmentPremium` is also statistically significant, meaning that being in the Premium vs. Economy segment is significant in determining price. Finally, the p-value for the interaction between `I(Power^-0.95):SegmentPremium` is 0.000049. This p-value is also statistically significant, meaning that the effect of power on price differs by brand segment.

Given that both main effects and their interaction are statistically significant, we have sufficient evidence to reject the null hypothesis and conclude that there is a relationship between power and price, and the

relationship between power and price differs across brands.

1.3.4 Conclusion

The Second-Hand Car Prices dataset focuses on the various factors that influence a vehicle's resale price. Through variable selection and various other selection methods, our group determined and transformed which predictors were significant to our model. Initially, we used our intuition and previous knowledge on cars to discover which variables would be relevant when considering a second-hand car's price. In step 2, we completed our data visualizations to view the predictors' correlation to the response variable. Then, we used stepwise regression to affirm many of our past findings.

When comparing the two models' R-squared values, we discovered that model 2 exhibited a slightly better fit and explains more variability in car prices. It's important to note that some transformations and adjustments were made to improve model performance and meet linearity assumptions.

In terms of our research findings, they are as follows:

Model 1 did not find `Year` to be statistically significant in predicting the response variable, which implies that the age of the car does not affect its price. This model fails to reject the null hypothesis for RQ 1, thus we conclude that the age of a car or any of its interactions with mileage, engine, and transmission do not have an effect on resale price.

Model 2 showed that `Power` significantly influences `Price`, and this relationship differs between economy and premium car brands (`Segment`). This model rejects the null hypothesis for RQ2, and concludes that `power` and `brand` are key determinants of car `price.` Premium cars generally have a higher price, and the effect of power on price is more pronounced in premium brands compared to economy brands.

Of course, we did face some limitations in our study. Although we found that `Year` does not significantly affect `Price`, the graph of median salary over time suggests otherwise. We noticed a pattern that car price experienced a steady decline from 2016 to 2021, with a sharp decline in 2020 indicative of how Covid affected the used-car market. Most likely, the reason that `Year` was not statistically significant was because we had too narrow a range of car age (only 5 years). If we were to conduct this research again, we would like to investigate how a broader scope of car ages may affect maintanence factors and thus price.

To all those out there looking to purchase a used-car, there are a couple factors you will have to consider. If you are specifically looking for cars with higher horsepower, you will have to spend that extra money to purchase a premium brand. However, if you are simply looking for a car to, for example, commute to work in, we suggest getting a cheaper Economy-brand car (not an Audi, Mercedes, or BMW). Additionally, it is important to note that though age does not play a significant role in car price, transmission surely does. So if you are looking for a cheap option and are comfortable driving a manual car, that is our best suggestion.

Appendix

EDA (Step 2)

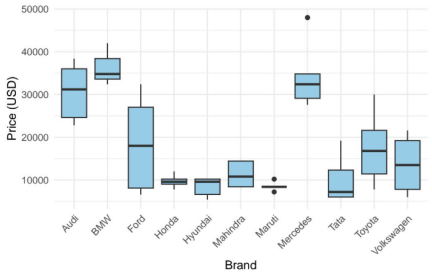


Figure 2: Price (USD) vs. Car Brand

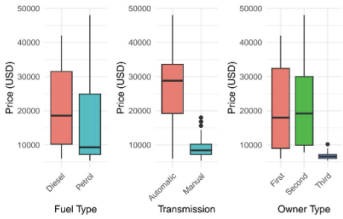


Figure 3: Boxplots of Price (USD) vs. Fuel Type, Transmission, and Owner Type

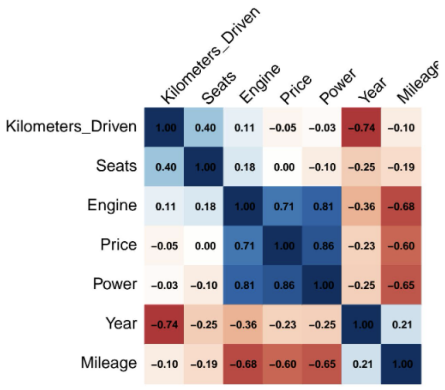


Figure 6: Heat Correlation Matrix

Citations

CARFAX. (2022, October 25). Luxury car brands. CARFAX. Retrieved December 4, 2024, from https://www.carfax.com/blog/luxury-car-brands

Mandala, S. (2020). Second hand car price prediction [Data set]. Kaggle. https://www.kaggle.com/datasets/sujithmandala/second-hand-car-price-prediction