MSDS 601 Final Project

Group Members: Richard Aw, Ashwani Rajan, Ka Ying Yam

Introduction

Consumer demand for used cars has returned to pre-COVID-19 levels [1], resulting in a recent surge in used car prices [2]. This leads to the question: what are the *stable* factors that have a nontrivial influence on used car prices? In this project, we seek to identify predictors that have a significant impact on a used car's sales price, as well as understand the *relative* impact of these predictors on a used car's sales price.

To this end, we analyzed a dataset [3] containing multiple such features for cars manufactured by the automotive manufacturer Hyundai Motor Company. Based on our analysis, the factors which have a significant impact on a used Hyundai car's price are: (i) age, (ii) mileage, (iii) mpg, (iv) engine size, (v) transmission, and (vi) fuel type.

Description of Data

The dataset has 4860 rows and 9 columns, with the following variables:

- "**model**": Hyundai model of the car.
- "**year**": Year of registration for the car.
- "**price**": Price of the car (in Euros).
- "**transmission**": Car's gearbox type.
- "**mileage**": Distance used up by the car.
- "**fuelType**": Type of engine fuel used by the car.
- "**tax**": Road tax of the car.
- "**mpg**": Miles per gallon, i.e., the distance per gallon the car can travel up to.
- "**engineSize**": Engine size (in litres) of the car.

Overview of Methodology

Our regression analysis was conducted via the following broad sequence of steps:

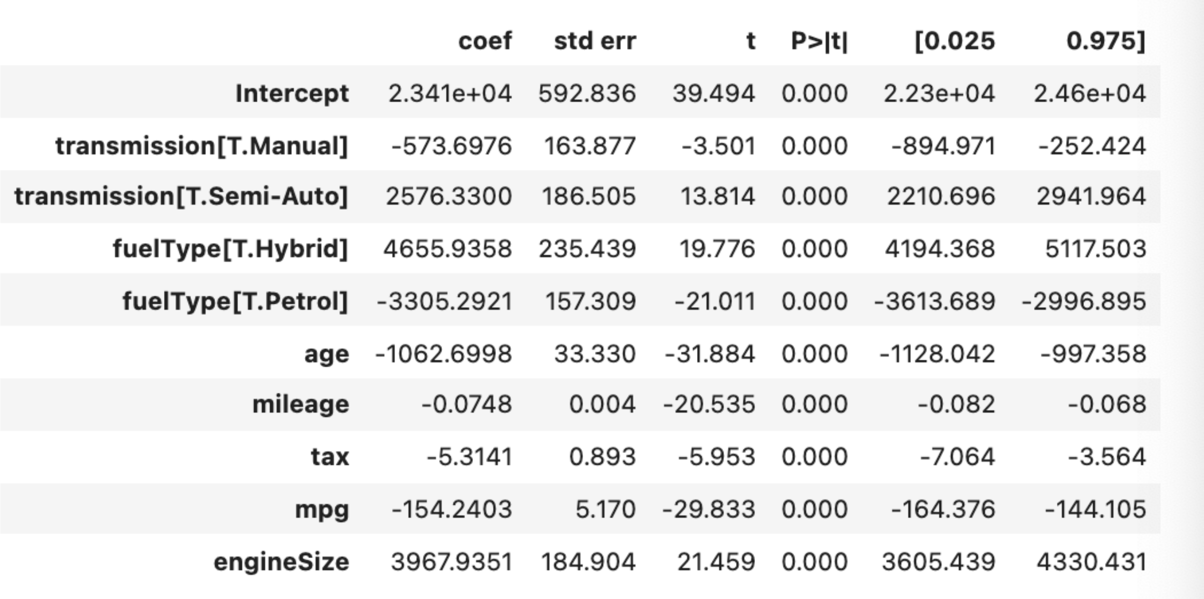
1. Check for multicollinearity in the data using the variance inflation factor (VIF) score of each predictor.
2. Fit a multiple linear regression model on the *full* data.

3. Check for influential points in the data using the Cook’s distance and externally studentized residual for every observation.
4. Check for (possible) heteroscedasticity using the Breusch-Pagan test.
5. Check for (possible) violation of normality using the Jarque-Bera test.
6. Select our most preferred model via the Best Subsets approach using, primarily, Akaike’s information criterion (AIC) and Bayesian information criterion (BIC).

Exploratory Data Analysis

We began by identifying the variables that may assist in answering the research questions. From the aforementioned 9 feature columns, we determined that all except “**model**” would be good predictors to explore further. We excluded “**model**” because it was a categorical variable with too many categories (16 categories). Including it might lead to overfitting and would add excessive computational strain on our analysis. We also calculated the cars’ ages from their year of registration and used that (in place of year of registration itself) as a predictor in our model. The reasons for this are that a car’s age (i) seems more *directly* relevant than its year of registration as a predictor of its price, and (ii) is more suited for linear regression analysis (of the two; in the sense that year of registration would be more suited for time series analysis).

The individual t test results from regressing price against age, mileage, tax, mpg, engine size, transmission, and fuel type (*'price ~ age + mileage + tax + mpg + engineSize + transmission + fuelType'*) suggested that all of the predictors included in this initial model were significant predictors at alpha = 0.05 (i.e., with a confidence level of at least 95%).

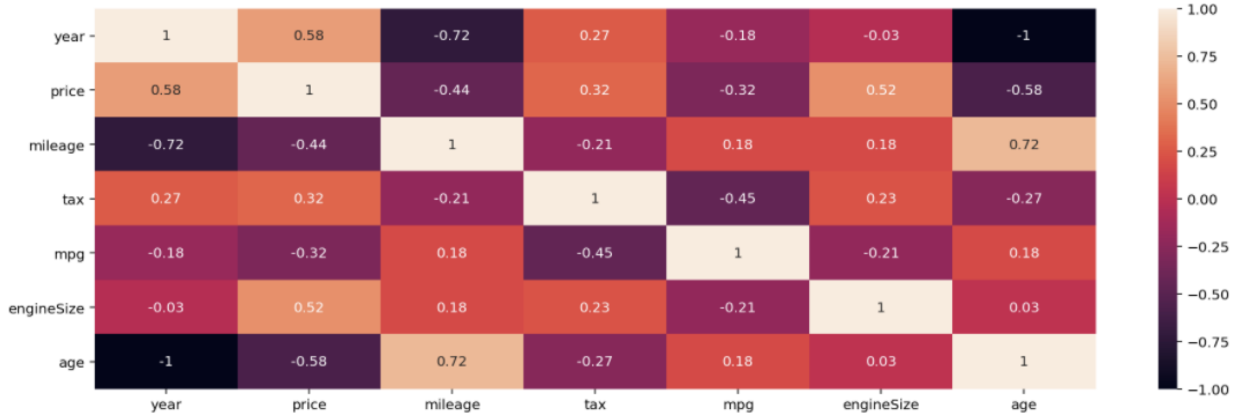


	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.341e+04	592.836	39.494	0.000	2.23e+04	2.46e+04
transmission[T.Manual]	-573.6976	163.877	-3.501	0.000	-894.971	-252.424
transmission[T.Semi-Auto]	2576.3300	186.505	13.814	0.000	2210.696	2941.964
fuelType[T.Hybrid]	4655.9358	235.439	19.776	0.000	4194.368	5117.503
fuelType[T.Petrol]	-3305.2921	157.309	-21.011	0.000	-3613.689	-2996.895
age	-1062.6998	33.330	-31.884	0.000	-1128.042	-997.358
mileage	-0.0748	0.004	-20.535	0.000	-0.082	-0.068
tax	-5.3141	0.893	-5.953	0.000	-7.064	-3.564
mpg	-154.2403	5.170	-29.833	0.000	-164.376	-144.105
engineSize	3967.9351	184.904	21.459	0.000	3605.439	4330.431

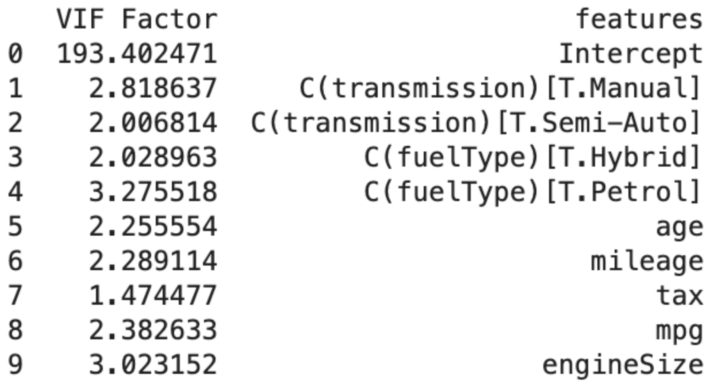
Regression Analysis

Model Diagnosis: Checking for Multicollinearity

The matrix of pairwise correlation between predictors showed that mileage and age were correlated, which made intuitive sense as an older car might be used more and therefore had higher mileage. However, since the correlation plot below only showed pairwise relationships, we needed to calculate the VIF to evaluate all features together.



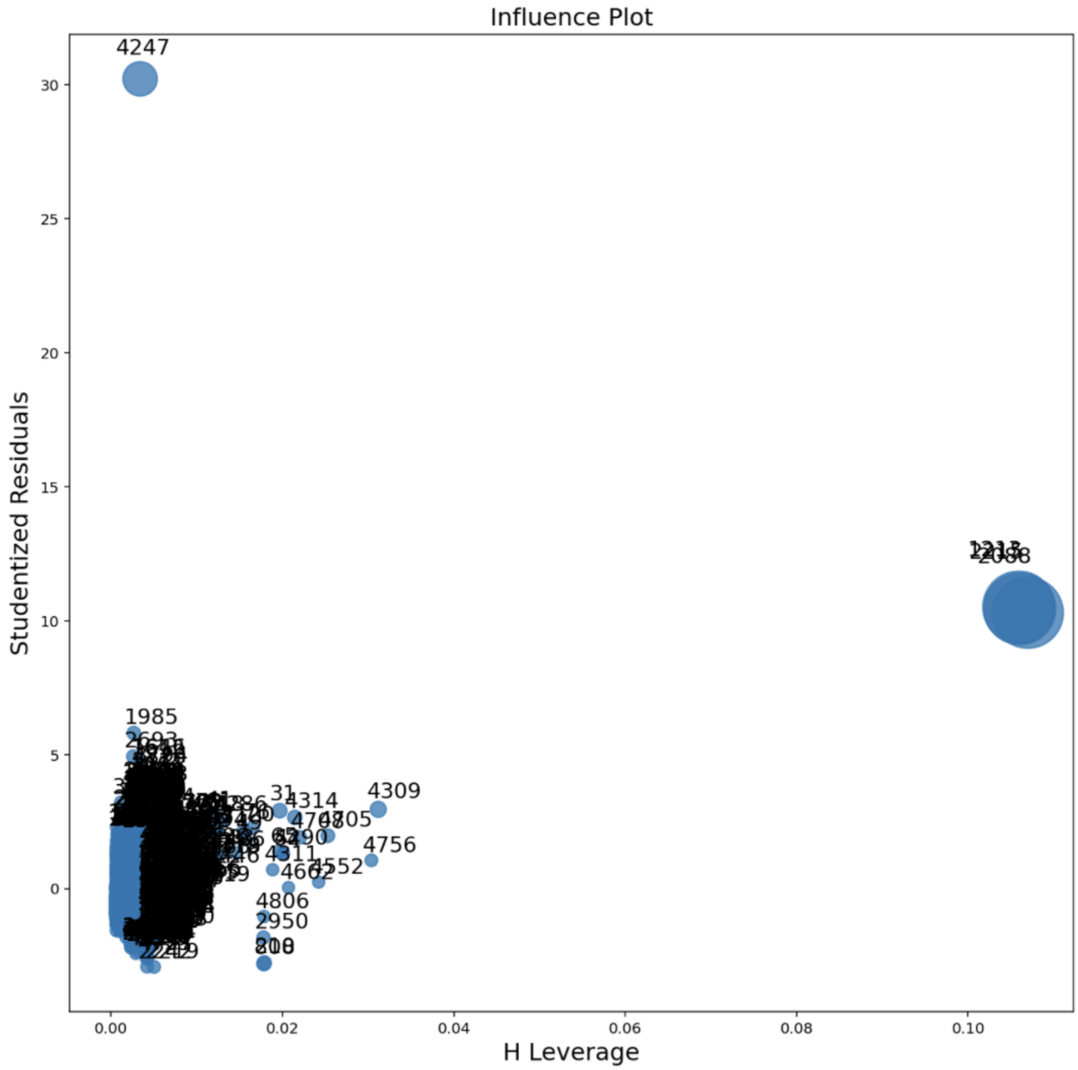
As shown below, the VIF scores were not indicative of significant multicollinearity among the predictor variables. They are all less than 4, a common upper bound for *light* multicollinearity.



```
   VIF Factor                        features
0  193.402471                       Intercept
1    2.818637     C(transmission)[T.Manual]
2    2.006814  C(transmission)[T.Semi-Auto]
3    2.028963         C(fuelType)[T.Hybrid]
4    3.275518         C(fuelType)[T.Petrol]
5    2.255554                             age
6    2.289114                         mileage
7    1.474477                             tax
8    2.382633                             mpg
9    3.023152                      engineSize
```

Model Diagnosis: Checking for Influential Points

We then proceeded to check for the presence of influential points by using the residual plot and influence plot. The points with large sizes on the influence plot indicated high Cook's distance values and were therefore flagged.

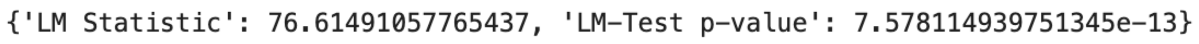


We then continued with the regression analysis in a parallel manner, considering the data with *and* without the influential points.

With Influential Points:

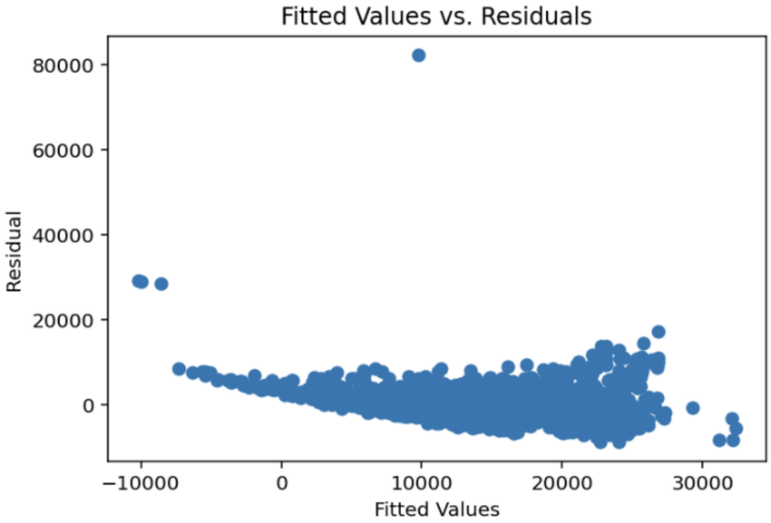
Model Diagnosis: Checking for Heteroscedasticity

From the residual plot and the Breusch-Pagan (BP) test, our initial model *did* present heteroscedasticity. As seen below, the BP test yielded a p-value (much) smaller than 0.05, suggesting that the data (with influential points) contained heteroscedasticity.



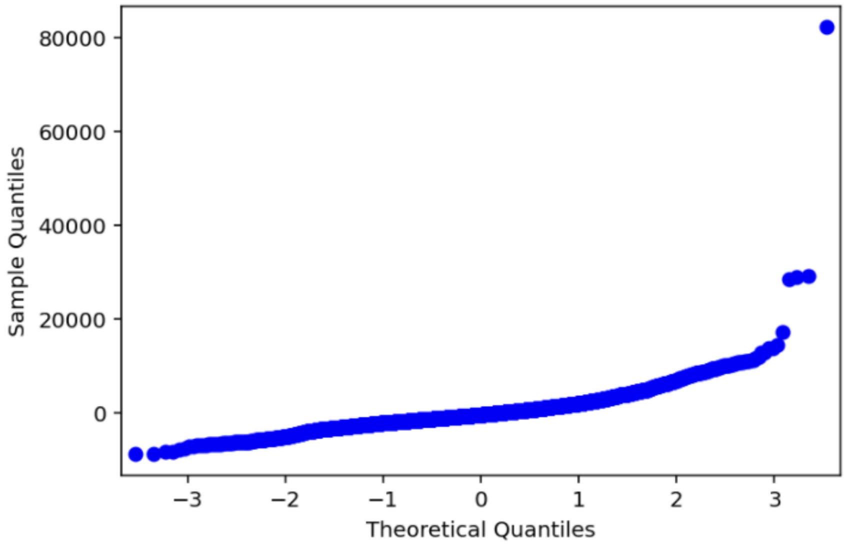
```
{'LM Statistic': 76.61491057765437, 'LM-Test p-value': 7.578114939751345e-13}
```

From the residual plot below, we can also see that the vertical spread of the residuals changes nontrivially as we move along the x-axis.



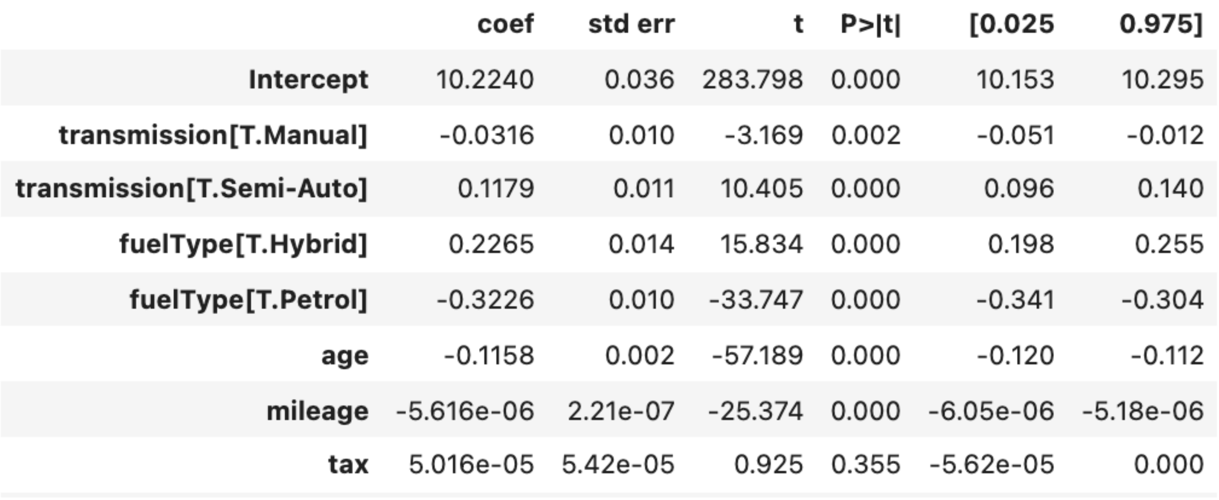
Model Diagnosis: Checking for Normality

The **QQ plot** below also seems *nonlinear*, which suggests that the data (with influential points) deviated from normality.



To mitigate the impact of heteroscedasticity and non-normality, we decided to perform a natural-log transformation on the values for the response variable (i.e., “price”) and fit the modified model - *‘log_price ~ age + mileage + tax + mpg + engineSize + transmission + fuelType’* - on **both** the data with influential points **and** the data without the influential points.

The summary table below for the modified model and data without influential points indicated that “tax” was no longer a significant predictor, per the t-test:

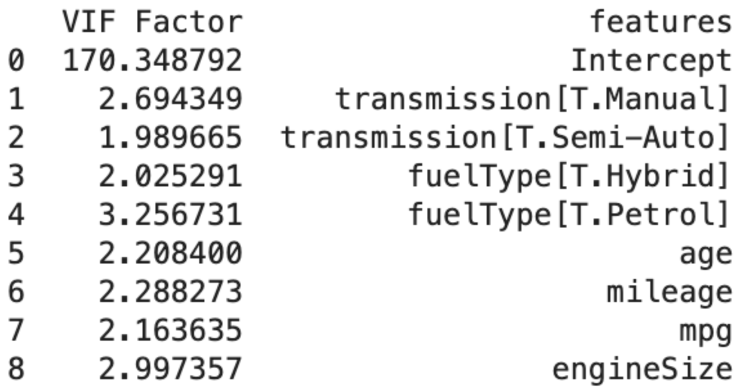


	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	10.2240	0.036	283.798	0.000	10.153	10.295
transmission[T.Manual]	-0.0316	0.010	-3.169	0.002	-0.051	-0.012
transmission[T.Semi-Auto]	0.1179	0.011	10.405	0.000	0.096	0.140
fuelType[T.Hybrid]	0.2265	0.014	15.834	0.000	0.198	0.255
fuelType[T.Petrol]	-0.3226	0.010	-33.747	0.000	-0.341	-0.304
age	-0.1158	0.002	-57.189	0.000	-0.120	-0.112
mileage	-5.616e-06	2.21e-07	-25.374	0.000	-6.05e-06	-5.18e-06
tax	5.016e-05	5.42e-05	0.925	0.355	-5.62e-05	0.000

Hence, we decided to exclude “tax” and proceed with the further modified model on the data without influential points:

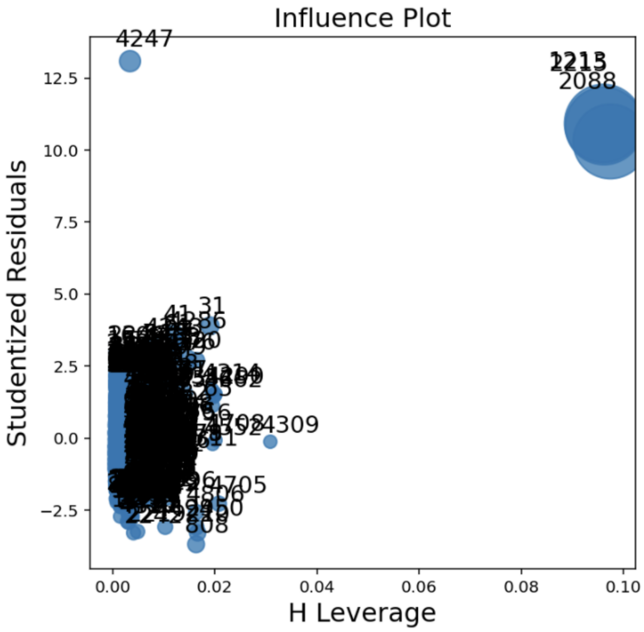
log_price ~ age + mileage + mpg + engineSize + transmission + fuelType

For the model with influential points, multicollinearity seemed *light* , with all VIF scores falling below 4.

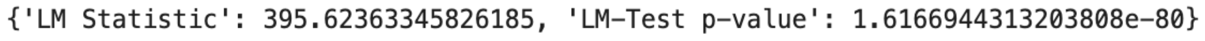


```
   VIF Factor                  features
0  170.348792                 Intercept
1    2.694349    transmission[T.Manual]
2    1.989665  transmission[T.Semi-Auto]
3    2.025291         fuelType[T.Hybrid]
4    3.256731         fuelType[T.Petrol]
5    2.208400                       age
6    2.288273                   mileage
7    2.163635                       mpg
8    2.997357                engineSize
```

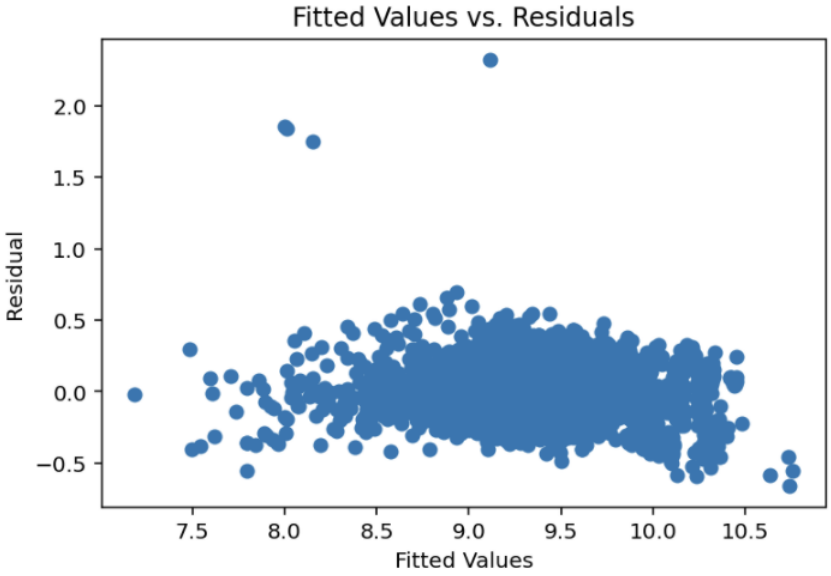
From the influence plot below, we observe that (new) influential points were still detected despite our performing the natural-log transformation.



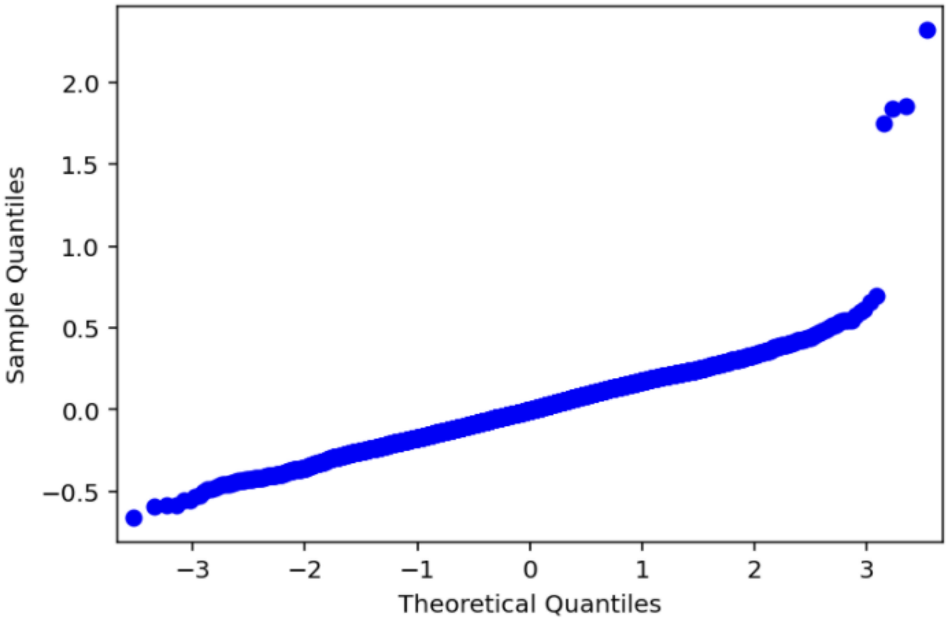
From the residual plot shown below, heteroscedasticity still seemed to present a problem. We could, however, see a slight improvement. (We also present the BP test results below, but avoid reading too much into it *in relation to previous BP test results* since the BP test is known to be sensitive to outliers.)



```
{'LM Statistic': 395.62363345826185, 'LM-Test p-value': 1.6166944313203808e-80}
```

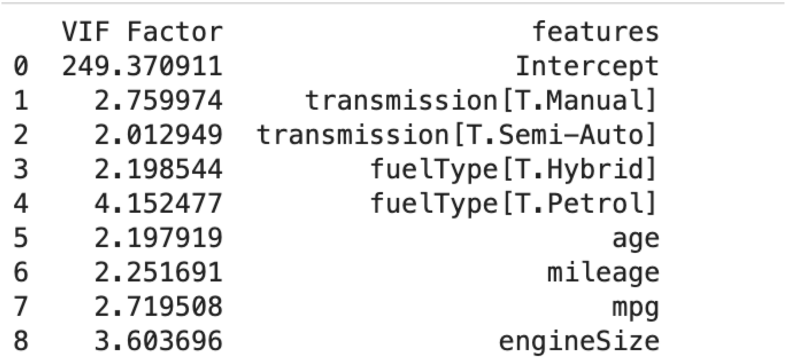


Based on the **QQ plot**, we observed a slight improvement with normality as well, although the normality assumption appeared to still be violated based on the JB test (p-value <= 0).



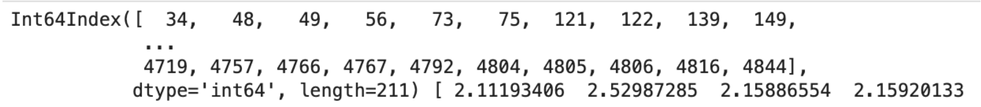
Without Influential Points:

For the transformed model without influential points, multicollinearity did not seem to be a problem. All the VIF scores were less than 4, which indicated only *light* multicollinearity.



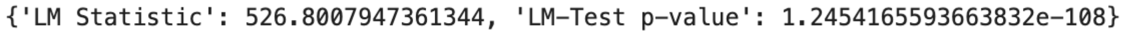
```
   VIF Factor                   features
0  249.370911                  Intercept
1    2.759974     transmission[T.Manual]
2    2.012949  transmission[T.Semi-Auto]
3    2.198544         fuelType[T.Hybrid]
4    4.152477         fuelType[T.Petrol]
5    2.197919                        age
6    2.251691                    mileage
7    2.719508                        mpg
8    3.603696                 engineSize
```

However, there were still 211 influential points identified by the externally studentized residuals.

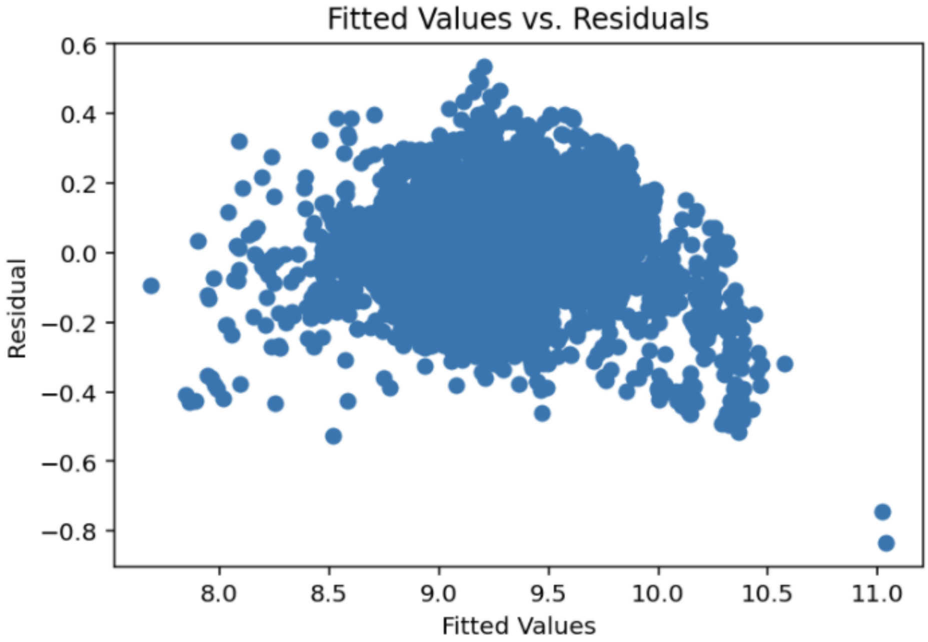


```
Int64Index([  34,   48,   49,   56,   73,   75,  121,  122,  139,  149,
            ...
            4719, 4757, 4766, 4767, 4792, 4804, 4805, 4806, 4816, 4844],
           dtype='int64', length=211) [ 2.11193406  2.52987285  2.15886554  2.15920133
```

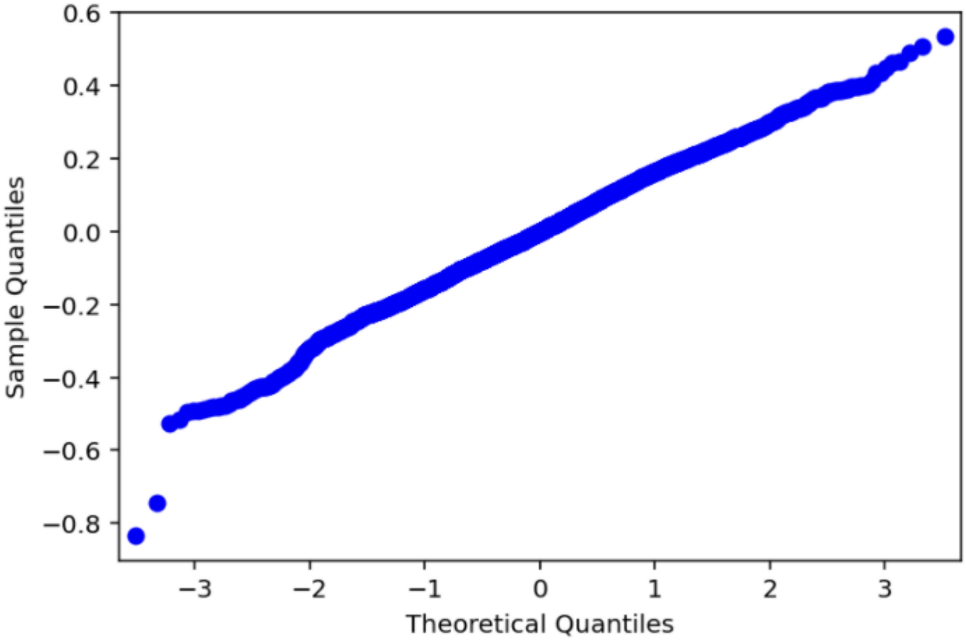
Heteroscedasticity also still seemed to present a problem, as shown by the residual plot below. (Again, we avoided reading too much into the BP test results shown below *in relation to previous BP test results* since the BP test is known to be sensitive to outliers.)



```
{'LM Statistic': 526.8007947361344, 'LM-Test p-value': 1.2454165593663832e-108}
```



Further, normality was still an issue here as well (p-value of JB test < 0). However, we could still see a significant improvement in the QQ plot.



Potential Problems with Data and Analysis

After performing natural log-transformation on "price", our model *log_price ~ age + mileage + mpg + engineSize + transmission + fuelType* (with and without influential points) still registered (new) influential points, heteroscedasticity, and non-normality. These can present significant issues in the following ways:

1. Influential points: fitted line pulled towards the uncommon points, resulting in inaccurate predictions.
2. Heteroscedasticity: the estimators β_j stay unbiased, but $Var(\beta_j)$ would be inconsistent and biased. This could cause $se(\beta_j)$ to be unreliable, resulting in biased individual t test results and confidence intervals.
3. Non-normality: T distribution, chi-square distribution, F distribution in tests will be "shifted." This is typically described as less important than other assumptions since β_j has an approximate normal distribution when sample size is large.

Model Selection

After dealing with model diagnostic problems i.e., dealing with heteroscedasticity, non-normality, and selecting influential points from our dataset, we are now ready to choose the final model. There are two main approaches for Model selection: The best subset approach and the stepwise selection approach. In the best subset approach, we compare all the models which are possible with our set of predictors. It requires high computation, but it is considered foolproof as it compares all the possible models. Since we have narrowed down our predictor candidates to **six,** we will deal with 63 possible subsets (excluding the Null model). This won't need high computation so we can accommodate the best subset approach here.

For the comparison, we iteratively regress all the 63 models and calculate the Adjusted R^2 , Mallows' Cp and the AIC and BIC values. Since Mallow's Cp is calculated with reference to the full model, we cannot use it to compare the full model against other models. Mallow's Cp also becomes unreliable when the data is very complicated or when we don't have the information on all the variables that could be related to the response variable.

In our case, we see very high Mallow's Cp values. The reason for high Cp values could be the high number of observations and low mean squared error for the full model. Due to the high Cp values, overfitting is less of an issue as compared to underfitting. Hence we don't have to worry about the full model overfitting the data. However, it would be difficult to conclude anything by looking at the Cp value, so we can exclude it from our analysis. Also, AIC is equivalent to Mallows' Cp for Gaussian errors and linear models, so we are not losing much by excluding Cp. Adjusted R^2 is also a good criteria to shortlist the candidate models, since it represents how much the model explains variance in y and it is sensitive to change in number of predictors.

Therefore, we can use Adjusted R^2, AIC and BIC to compare the 63 different models and select a pool of candidate models.

From our analysis, we observe that for 25 model candidates, the Adjusted R^2 values are higher than 0.7 and for 11 models the Adjusted R^2 is higher than 0.8. Since we are studying the dependence of car prices on predictors which are not only intuitively relevant (from contextual knowledge) but also physical in nature, we expect a relationship between the response and predictor variables. Adjusted R^2 values are generally expected to be high in such cases.

Shortlisted Models based on Adjusted R^2:

Model	Adj. R sq.	AIC	BIC
log_price~age+mpg+fuelType	0.83023	-2912.9775	-2880.7565
log_price~age+mileage+mpg+fuelType	0.84744	-3408.9004	-3370.2352
log_price~age+mileage+engineSize+transmission	0.80009	-2152.5451	-2113.8800
log_price~age+mpg+engineSize+fuelType	0.83836	-3140.1772	-3101.5121
log_price~age+mpg+transmission+fuelType	0.83790	-3126.0278	-3080.9185
log_price~age+mileage+mpg+engineSize+transmission	0.80336	-2228.0726	-2182.9632
log_price~age+mileage+mpg+engineSize+fuelType	0.85746	-3723.8334	-3678.7240
log_price~age+mileage+mpg+transmission+fuelType	0.85413	-3615.1827	-3563.6292
log_price~age+mileage+engineSize+transmission+fuelType	0.81367	-2477.3873	-2425.8338
log_price~age+mpg+engineSize+transmission+fuelType	0.84586	-3358.8698	-3307.3163
log_price~age+mileage+mpg+engineSize+transmission+fuelType	0.86402	-3940.7665	-3882.7687

It is also noteworthy that these are also the models with the least AIC and BIC values. Out of the above listed models, we can select our best model by looking at the AIC and BIC values. The **Full Model,** i.e. *'log_price~age+mileage+mpg+engineSize+transmission+fuelType'* has the least AIC and BIC values. It also has the highest Adjusted R^2 value. We also note that the model '*log_price ~ age+mileage+mpg+engineSize+fuelType'* also has very good Adjusted R^2 and ABC/BIC values. It is our second best model.

Though the second best model excludes the 'transmission' predictor, the drop in Adjusted R^2 is negligible. This can be explained by the individual t-statistics. 'Transmission' is a categorical variable and is mainly populated by the 'manual' category (76%). The t-statistic for this category suggests that it is not significant. However, we keep the "transmission" predictor in our best model as at least one of the categories for "transmission" is significant from the individual t-test, which suggests that the entire categorical variable needs to be included.

Therefore, we propose the following model as our best model:

"log_price~age+mileage+mpg+engineSize+transmission+fuelType"

Conclusion

The Hyundai Used Car dataset has 4860 entries of used cars along with their prices and other attributes like engine size, fuel type, mileage, mpg, transmission type, model type and year of manufacture. We studied the relationship between all these attributes and the Price of the used car using Multiple Linear Regression. We also attempted to find the best linear model to predict the car prices based on the given attributes.

Initially, we performed data cleaning and some initial exploratory analysis. As a result, we excluded the 'Model' predictor and excluded the 'Others' category in 'engineSize' and 'transmission' section. We then performed model diagnostics tests to deal with model problems like multicollinearity, heteroskedasticity, influential points and non-normality of residuals. Based on the Variance Inflation Factors, we reject the possibility of multicollinearity. Looking at the Influence plot, we can see the presence of influential points. We use both Cook's distance and external studentized residuals to detect the influential points and we exclude the ones which have high Cook's distance and high externalized studentized residuals. Heteroscedasticity is evident from the results of the Breusch–Pagan test and the residuals vs fitted values plot. We perform log transformation to deal with heteroskedasticity. Excluding influential points and log transformation of the response variable helps us deal with non-normality as well.

We look at the individual t-test results to further exclude the 'tax' predictor as it is not a significant addition to our model given all other predictors. Finally, we perform model selection using the best subset approach. This approach can be used as we have a manageable number of predictors. Using Adjusted R^2 values and AIC and BIC values, we decide on our final model, i.e., *"log_price~age+mileage+mpg+engineSize+transmission+fuelType"*.

Limitations

Based on our analysis, the factors which have a significant impact on a used Hyundai car's price are: (i) age, (ii) mileage, (iii) mpg, (iv) engine size, (v) transmission, and (vi) fuel type. We see a few limitations as a result of model diagnostics which are listed below:

- High Adjusted R^2 suggests that overfitting might be an issue. However, it would be difficult to conclude as we are not doing prediction and the relationship we are trying to study is physical in nature, so we can expect high R^2 values.
- Though the influential points which had high Cook's distance and high externalized studentized distribution were excluded, some of the influential points still remain. It is not ideal to remove those influential points without consulting car manufacturers and other subject matter experts to understand the significance of these influential points, if any.

- After doing the log-transformation, the fitted vs residual plot in case of heteroscedasticity seems to improve but the BP test result worsens. This could be because of the presence of the outliers. Since the BP test is sensitive to outliers, we cannot conclude whether Heteroscedasticity improves or not after log transformation just by looking at the BP-test results.
- The Jarque–Bera test results suggest non-normality. Consequently, this may affect the sensitivity of our results from t-test and ANOVA, which may lead to missing out on the best model.

References

1. Furcher, T., Holland-Letz, D., Rupalla, F., and Tschiesner, A. (2021, Aug 27). Car buying is on again, and mobility is picking up. Retrieved from: https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/how-consumers-behavior-in-car-buying-and-mobility-changes-amid-covid-19.
2. Yang, Y. (2021, Sep 17). Used Car Prices on the Rise, Raising Inflationary Specter. Retrieved from: https://www.bloomberg.com/news/articles/2021-09-17/used-car-prices-in-u-s-rise-again-in-first-half-of-september.
3. Dataset: https://www.kaggle.com/mysarahmadbhat/hyundai-used-car-listing