BMW Used Car Price Predictive Analyses

1. Case Background:

In the US, cars have always been one of the most important means of transportation. Most people in the US tend to have their vehicles from high school or college. There is an abundance of pre-owned cars in the US, which trickle down to the second-hand market. As MSBA student who is going to graduate next summer, there is a need to buy my own car after starting my career. I aim to analyze the "BMW used car" dataset and come up with a realistic and efficient prediction model to forecast the price of the used BMW cars, and choose the best deal available as per requirements.

2. Interested Questions:

- *Can I accurately predict a used BMW car's price?*

  Predicting future car prices can help me make better purchasing decisions based on budget, car condition and model preference. Additionally, the predictive model can also be used to measure the value of my own car, which is conducive to selling used cars at reasonable prices.

- *What are the important factors?*

  In this dataset, I have data of car's different attributes, such as car model, car mileage, car release year, mile per gallon, and so on. I expect to identify those factors that have significant effects on car prices, using different statistical models.

- *What are the effects of those important factors?*

  After identifying the important factors, I also want to identify what kind of effects do those important factors have on the car price. With these insights, I would be able to identify potential good deals in the BMW used-car market.

3. Data Definition:

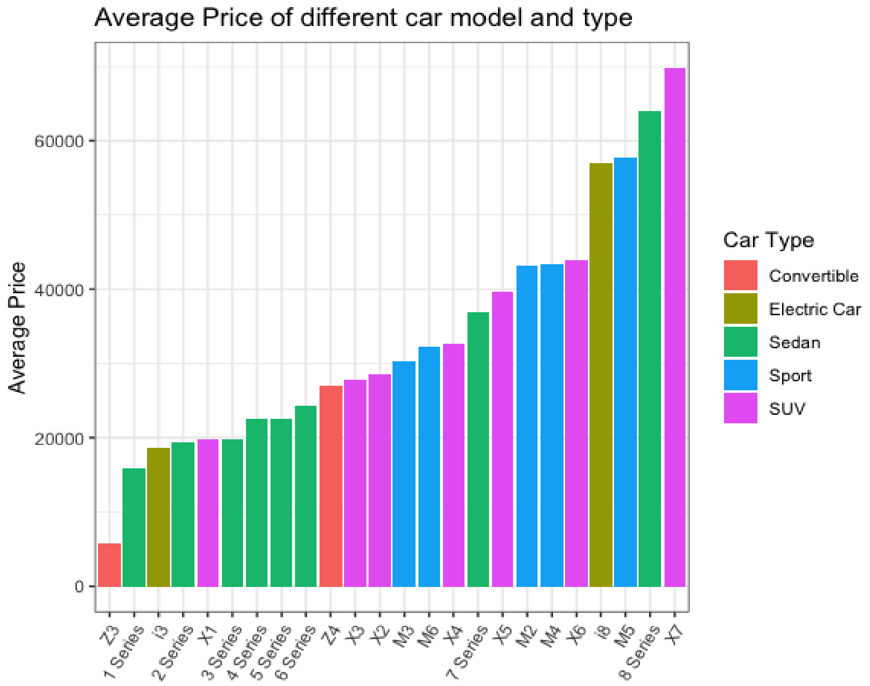
Within the dataset, I have a total of 9 variables. 3 of them are categorical, and 6 of them are numeric.

Categorical Variable		
1.	***model***	Model of the BMW used car (ex: 3 Series/ 5 Series/ X3/ …)
2.	***transmission***	Gearbox type (Manual/Auto/Semi-Auto)
3.	***fuelType***	Type of fuel used (Petrol/Diesel/Electric/Hybrid/Other)

Numeric Variable		
1.	***year***	Registration year of the car
2.	***mileage***	Total miles used of the car
3.	***price***	The price of the car in € when sold at the end of 2020
4.	***tax***	The road tax of the car in €
5.	***mpg***	Miles per gallon of the car
6.	***engineSize***	Engine size of the car in liter

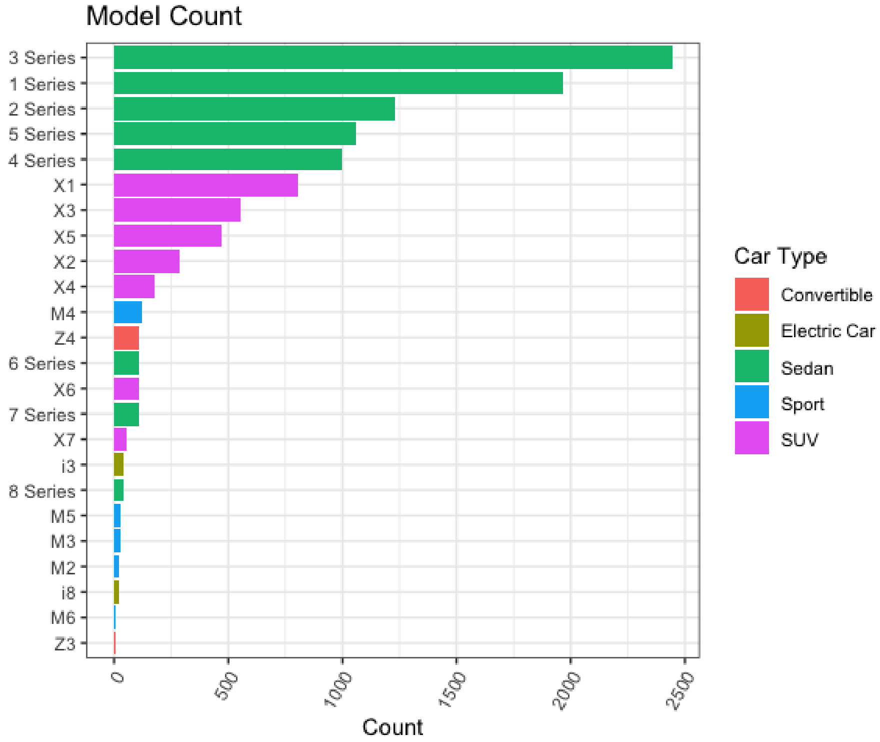
4. Data Exploration:

- *Average price of different car model:*



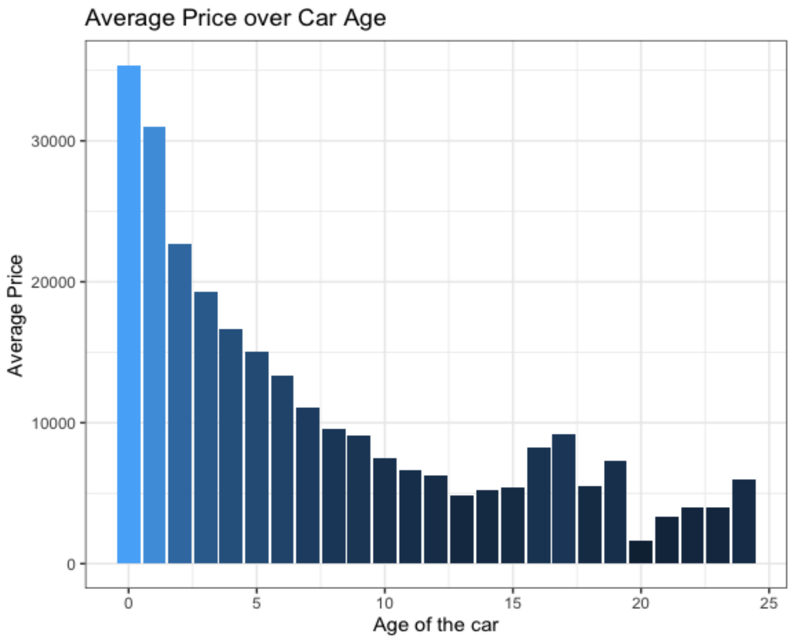
From the bar plot above, I notice that the price of different car models varies a lot. In addition, I could also tell that for each car type BMW has introduced multiple products that have very different prices so that BMW could cover more markets.

- *Popularity of different car models:*



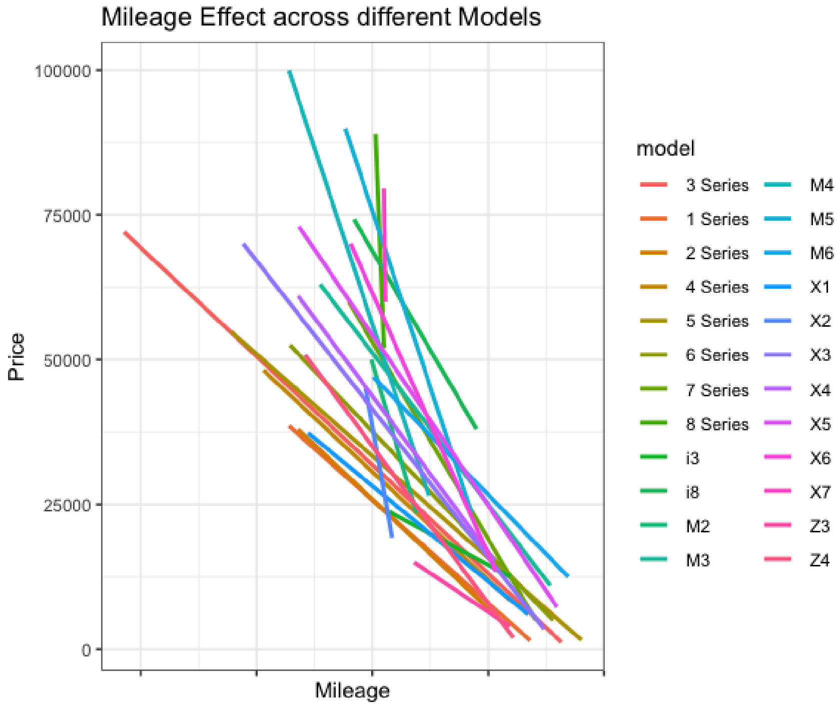
From the bar plot above, I can find that the sedans, such as the 3 Series and the 5 Series, are the most popular car type, and the SUVs are the second most popular car type.

- *Average price of different car’s age:*



From the time series plot above, I notice that the price drops dramatically in the first three years. Then, the depreciation rate starts to slow down gradually. Eventually, when the car is more than 10 years old, the price becomes stable.

- *Mileage effect across different car models:*



Each line represents a car model, and the slope implies the effect of mileage on price for that specific car model. Given that the slopes of the lines differ a lot, I can tell that the effect of mileage depends on the car model. For instance, because the line for M5 is steeper than the line for the 3 Series. I can conclude that compared to the 3 Series, M5’s price is more sensitive to mileage.

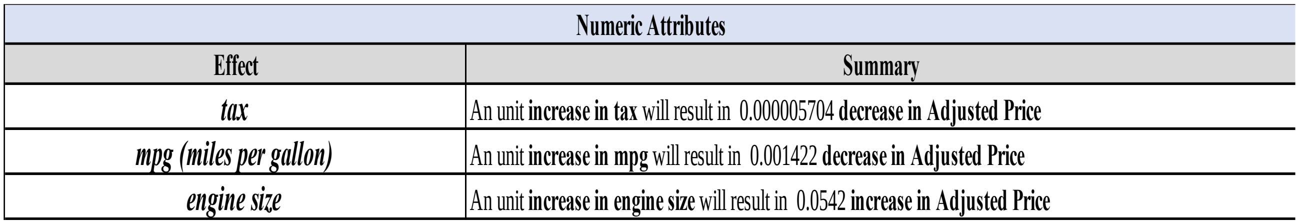
5. Analyses & Results:

- *Model Description:*

I fitted a ***multiple linear regression model***, using ***price*** as the ***outcome*** and ***mileage, tax, miles per gallon, engine size, fuel type, transmission type, car model*** as the predictors. Eventually, I found that a used BMW car's ***mileage, tax, miles per gallon, engine size, fuel type, transmission type, car model*** all have significant effects on ***log_{10} (Price)***, which will be called as ***Adjusted Price*** in the following section**.** In addition, I also notice that ***the effect of mileage*** on car price actually ***varies among different car models***. In other words, different car models have very different value preservation rates. Lastly, with these factors I could accurately predict a used BMW car's price. To be more specific, on average our ***prediction error*** is only around ***3,900 Euros***.

- *Insights:*

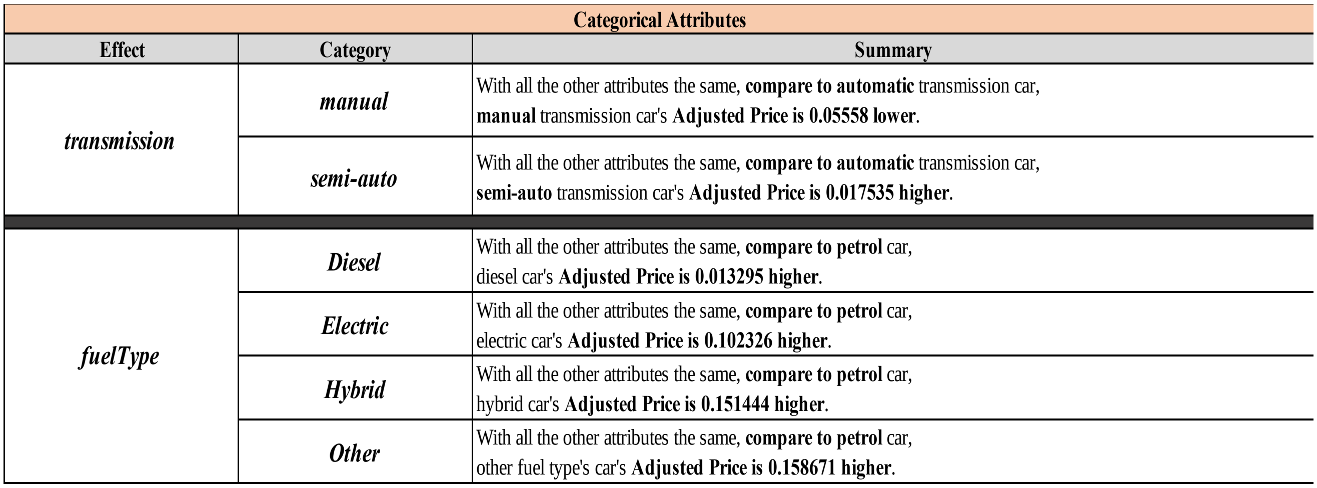
I. Numeric Attributes' Effects →



Numeric Attributes	
Effect	**Summary**
tax	An unit **increase in tax** will result in 0.000005704 **decrease in Adjusted Price**
mpg (miles per gallon)	An unit **increase in mpg** will result in 0.001422 **decrease in Adjusted Price**
engine size	An unit **increase in engine size** will result in 0.0542 **increase in Adjusted Price**

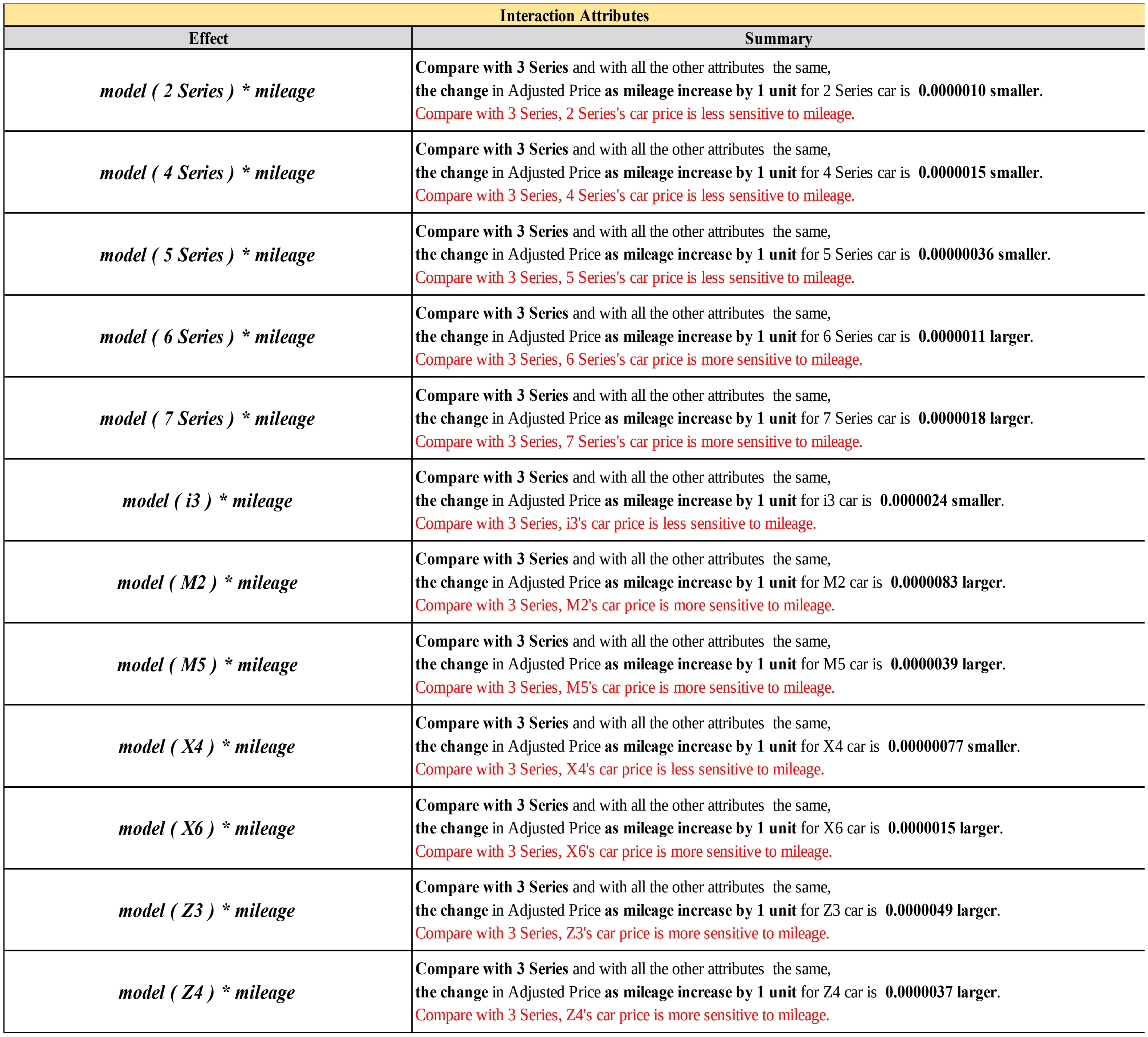
II. Categorical Attributes' Effects →

In this dataset, given that the 3 Series is the most popular and sold model. I set the **3 Series** as the base model and will compare the effects of all the other models with the 3 Series. Next, I also set **automatic transmission**, which is the most common transmission type, as the base transmission type and will compare the effects of all the other types of transmission with the automatic transmission. Lastly, I set **petrol**, which is the most common fuel type, as the base fuel type and will compare the effects of all the other types of fuel with petrol.



Categorical Attributes		
Effect	**Category**	**Summary**
transmission	***manual***	With all the other attributes the same, **compare to automatic** transmission car, **manual** transmission car's **Adjusted Price is 0.05558 lower**.
	semi-auto	With all the other attributes the same, **compare to automatic** transmission car, **semi-auto** transmission car's **Adjusted Price is 0.017535 higher**.
fuelType	***Diesel***	With all the other attributes the same, **compare to petrol** car, diesel car's **Adjusted Price is 0.013295 higher**.
	Electric	With all the other attributes the same, **compare to petrol** car, electric car's **Adjusted Price is 0.102326 higher**.
	Hybrid	With all the other attributes the same, **compare to petrol** car, hybrid car's **Adjusted Price is 0.151444 higher**.
	Other	With all the other attributes the same, **compare to petrol** car, other fuel type's car's **Adjusted Price is 0.158671 higher**.

III. Interaction Attributes’ Effects →



Interaction Attributes	
Effect	**Summary**
model (2 Series) * mileage	**Compare with 3 Series** and with all the other attributes the same, **the change** in Adjusted Price **as mileage increase by 1 unit** for 2 Series car is **0.0000010 smaller**. Compare with 3 Series, 2 Series's car price is less sensitive to mileage.
model (4 Series) * mileage	**Compare with 3 Series** and with all the other attributes the same, **the change** in Adjusted Price **as mileage increase by 1 unit** for 4 Series car is **0.0000015 smaller**. Compare with 3 Series, 4 Series's car price is less sensitive to mileage.
model (5 Series) * mileage	**Compare with 3 Series** and with all the other attributes the same, **the change** in Adjusted Price **as mileage increase by 1 unit** for 5 Series car is **0.00000036 smaller**. Compare with 3 Series, 5 Series's car price is less sensitive to mileage.
model (6 Series) * mileage	**Compare with 3 Series** and with all the other attributes the same, **the change** in Adjusted Price **as mileage increase by 1 unit** for 6 Series car is **0.0000011 larger**. Compare with 3 Series, 6 Series's car price is more sensitive to mileage.
model (7 Series) * mileage	**Compare with 3 Series** and with all the other attributes the same, **the change** in Adjusted Price **as mileage increase by 1 unit** for 7 Series car is **0.0000018 larger**. Compare with 3 Series, 7 Series's car price is more sensitive to mileage.
model (i3) * mileage	**Compare with 3 Series** and with all the other attributes the same, **the change** in Adjusted Price **as mileage increase by 1 unit** for i3 car is **0.0000024 smaller**. Compare with 3 Series, i3's car price is less sensitive to mileage.
model (M2) * mileage	**Compare with 3 Series** and with all the other attributes the same, **the change** in Adjusted Price **as mileage increase by 1 unit** for M2 car is **0.0000083 larger**. Compare with 3 Series, M2's car price is more sensitive to mileage.
model (M5) * mileage	**Compare with 3 Series** and with all the other attributes the same, **the change** in Adjusted Price **as mileage increase by 1 unit** for M5 car is **0.0000039 larger**. Compare with 3 Series, M5's car price is more sensitive to mileage.
model (X4) * mileage	**Compare with 3 Series** and with all the other attributes the same, **the change** in Adjusted Price **as mileage increase by 1 unit** for X4 car is **0.00000077 smaller**. Compare with 3 Series, X4's car price is less sensitive to mileage.
model (X6) * mileage	**Compare with 3 Series** and with all the other attributes the same, **the change** in Adjusted Price **as mileage increase by 1 unit** for X6 car is **0.0000015 larger**. Compare with 3 Series, X6's car price is more sensitive to mileage.
model (Z3) * mileage	**Compare with 3 Series** and with all the other attributes the same, **the change** in Adjusted Price **as mileage increase by 1 unit** for Z3 car is **0.0000049 larger**. Compare with 3 Series, Z3's car price is more sensitive to mileage.
model (Z4) * mileage	**Compare with 3 Series** and with all the other attributes the same, **the change** in Adjusted Price **as mileage increase by 1 unit** for Z4 car is **0.0000037 larger**. Compare with 3 Series, Z4's car price is more sensitive to mileage.

6. Recommendations:

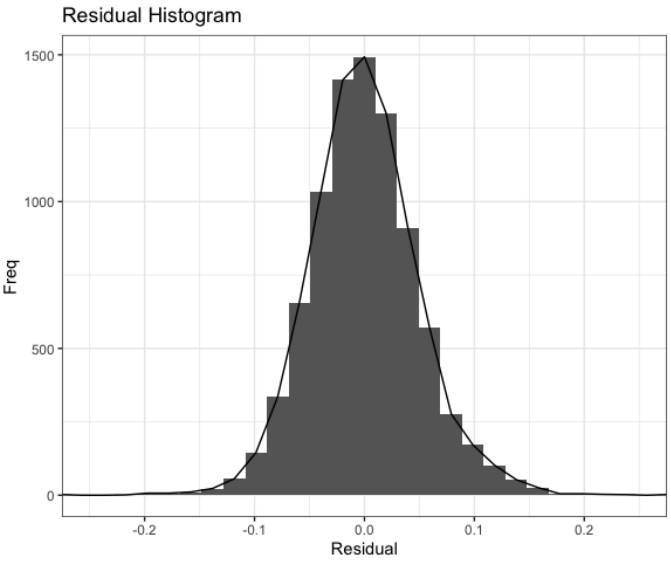
- The **manual and petrol** BMW used cars are the most economic choice.

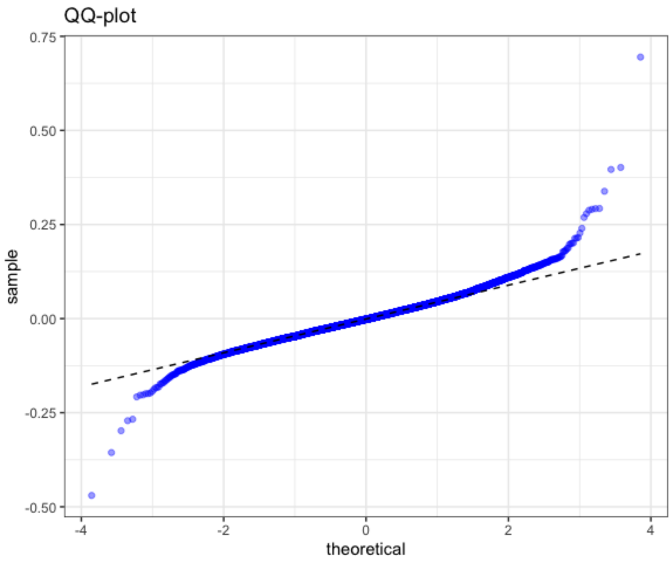
- The effects of **mileage on car price** actually depend on the **car model**. Hence, while purchasing BMW used cars, **inspect mileage and car model together**.

- With data of the attributes of a used BMW car, our model could predict the sales accurately. To be more specific, our **prediction error** is only around **3,900 Euros.**

7. Appendix:

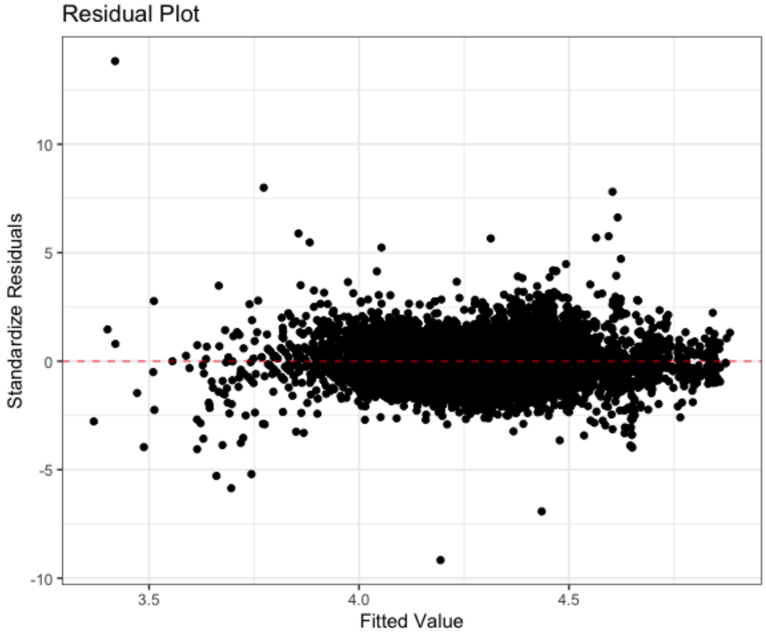
- *Assumption Check: Residual Normality*





According to the residual histogram, I can tell that the residuals are approximately bell-shaped and center at 0. Besides, according to the QQ-plot, the residuals approximately align the theoretical line. Hence, the residuals are approximately **normally distributed**.

- *Assumption Check: Constant Variance*



According to the residual plot above, I can tell that the residuals randomly scatter around 0, and there's no apparent pattern or trend. Hence, I conclude that the data has **constant variance**.