Price Prediction in the Used Cars Industry

Group Project Using R Programming and Data Science Techniques

MOKRANE Naïm

OHEIX Romain

Business Statistics, Analytics, and Data Science
Serge Nyawa

December 3, 2024

Table of Contents

1. Discovery
   1.1 Business Domain
   1.2 Project Objective
   1.3 Available resources
   1.4 Framing the business problem
   1.5 Initial Hypotheses
2. ELT Process
   2.1 Extracting and cleaning the data
   2.2 Transforming the data for analysis
   2.3 Final data preparation
3. Model Planning
   3.1 Variable selection
   3.2 Correlation analysis
   3.3 Outlier detection
   3.4 Validation strategy
4. Model Building
   4.1 Fitting a linear regression model
   4.2 Coefficient interpretation and model performance
   4.3 Residual Analysis
   4.4 Testing the model on another set
5. Recommendations

1. Discovery

Business Domain

The used car market is rapidly growing, especially with online platforms enabling buyers to compare prices and vehicle features. The factors influencing car prices include the vehicle's condition, mileage, year of manufacture, brand, model, as well as specifics like fuel type, transmission or colors. For sellers and platforms, accurately estimating vehicle prices is essential to attract buyers while remaining competitive.

Project Objective

The goal of this project is to predict the price of used cars based on their available features. We aim to answer the following question:

> *What are the key factors influencing the price of used cars, and how can we use this information to create a reliable predictive model?*

Available Resources

We have a dataset that includes the following information for each vehicle:

- **General Features**:
  - brand model.
  - model_year (year of manufacture).
- **Technical Features**:
  - milage (mileage).
  - fuel_type and transmission (automatic or manual gearbox, electric or gasoline vehicles).
  - engine (engine description).
- **History and Legal Status**:
  - accident (presence of past accidents).
  - clean_title (legal status of the vehicle).
- **Target Variable**:
  - price (price of the car).

This dataset provides a solid foundation for analysis, although some variables will need cleaning or transformation before use.

Framing the business problem

The business problem is reframed as an analytical challenge:

> *Develop a regression model to predict the price of a used car based on its features.*

Initial Hypotheses

Before diving into the data, we propose the following hypotheses:

1. Vehicles with higher mileage have lower prices.
2. Newer vehicles (recent manufacturing years) are priced higher.
3. Cars with a history of accidents or non-clean titles have reduced prices.
4. Fuel type and transmission play a role: electric or automatic cars may be priced higher.
5. Some colors are rare and could affect the price as well.

These hypotheses will guide the analysis and modeling phases.

2. ETL Process

In this project, we worked through several steps to clean, transform, and prepare the dataset for analysis and modeling following the hypothesis above and always keep in mind that we want to do a linear regression.

1. Extracting and Cleaning the Data

The dataset contains multiple columns, including the vehicle price, mileage, and other important attributes. The first step was to ensure that the data was clean and free of inconsistencies, such as special characters or missing values.

- **Price Cleaning**: The **price** column initially included dollar signs and commas, which are not suitable for numerical analysis. We removed these characters to create a new **price_clean** column that holds the numeric values for vehicle prices.

- **Mileage Cleaning**: The **milage** column contained non-numeric characters like commas and text. We extracted the numeric values (mileage) and created a new column, **milage_clean**.

- **Vehicle Age Calculation**: We created a new column called **vehicle_age**, which represents the age of the vehicle. The formula used was simply subtracting the **model_year** from the current year (2024).

- **Handling Missing Titles**: In the **clean_title** column, missing values (including empty spaces) were replaced with "No" to ensure uniformity, because if a car has no information about its legal cleanliness then we can consider that it’s not a clean title. This column was then encoded into a binary format, creating the **clean_title_encoded** column where "Yes" was represented as 1, and "No" as 0.

- **Handling Missing Fuel Types**: In the **fuel_type** column, we replaced missing values (blank space) with "Electric" to maintain consistency because we saw that the rows with a blank space were related to electrical cars and we deleted the “--” rows due to a problem in the importation of the CSV file.

- **Accident Data Cleaning**: The **accident** column had missing values and inconsistent entries. We filtered out rows where the **accident** information was empty, and then encoded the remaining values as binary: 1 for vehicles with at least one reported accident, and 0 for others. The original **accident** column was then dropped.

- **Factor Encoding**: Several categorical columns, such as **fuel_type**, **transmission**, **brand**, and **model**, were converted into factors. This was done to ensure that these categorical variables were handled properly during modeling.

2. Transforming the Data for Analysis

After cleaning the dataset, we focused on transforming certain columns to ensure that they could be used effectively for regression modeling.

- **Fuel Type Encoding**: We created a **fuel_dict** to assign numeric values to the various fuel types. Each fuel type was given a unique numeric value based on its order (from the most polluting fuel type to the cleaner) in the dictionary ("Diesel" = 1 , "Gasoline" = 2, "E85 Flex Fuel" = 3, "Hybrid" = 4, "Plug-In Hybrid" = 5, "Electric" = 6, "not supported (hydrogen" = 7). This allowed us to encode the **fuel_type** column into a numeric form using this mapping.

- **Engine Power (HP) Extraction**: The **engine** column contained textual descriptions of engine specifications, including horsepower (HP). We used a regular expression to extract the horsepower values (e.g., "300.0HP") and created a new column, **hp**, which holds the extracted numeric horsepower data. Any rows with missing HP values were removed from the dataset to ensure completeness. Unfortunately, not all the textual descriptions contained the HP so we had to remove around 700 rows.

- **Color Encoding**: The **ext_col** (external color) column, which contains the color of the vehicle, was encoded similarly to the **fuel_type** column. We created a **color_dict** mapping each color to a numeric value (from the most common to the rarest ("Black" = 1 , "White" = 2, "Gray" = 3, "Silver" = 4, "Blue" = 5, "Red" = 6, "Green" = 7, "Brown" = 8, "Gold" = 9 , "Beige"= 10 , "Orange" = 11 , "Yellow" = 12, "Purple" = 13, "Pink" = 14). This transformation allows the color to be used as a predictor in the regression model.

3. Final Data Preparation

After all the necessary transformations, we ensured that the dataset was ready for analysis by removing any remaining irrelevant or redundant columns. For instance, the original **price** and **milage** columns were dropped as they were replaced by the clean versions. We also ensured that no rows with missing values remained for important features like **hp**, **fuel_type**, and **color**.

The **ETL** process for this dataset involved extracting and cleaning data from various columns, transforming categorical variables into numeric formats suitable for regression analysis, and calculating new variables like **vehicle_age** and **hp**. With the cleaned and transformed data, we were able to create a dataset ready for regression modeling, where we could analyze factors such as the vehicle's color, fuel type, and engine power, all of which play a crucial role in determining the price of used cars.

This cleaned and prepared dataset will be used in the next stages of the analysis, including building and evaluating a regression model to predict vehicle prices based on these factors.

3. Model Planning

In this phase, we plan which model will be best suited to predict the target variable, in this case, the price of used cars. Given that the target variable is numeric, linear regression is an appropriate starting choice. However, before applying the regression model, we must ensure proper variable selection and address any issues that might impact model accuracy.

Key Considerations

1. **Variable Selection**
   We will identify the most relevant variables for predicting car prices based on our previous data cleaning and transformation:
   - **vehicle_age**: The age of the car.
   - **hp**: The horsepower of the vehicle.
   - **milage_clean**: The total mileage of the car, indicating how much it has been driven.
   - **fuel_type_clean**: The type of fuel used by the car.
   - **accident_binary**: Indicates whether the car has been in an accident.
   - **color**: The external color of the car, which may influence buyer preference.

clean_title_encoded was removed because it only contained the value 1 since we deleted a lot of rows in our previous cleaning step.

2. **Correlation Analysis**

A correlation analysis will be conducted to check the relationships between these variables and the target variable. Additionally, we will ensure no variables are highly correlated with each other to avoid multicollinearity issues.

	price_clean	vehicle_age	hp	milage_clean	fuel_type_clean	accident_binary	color
price_clean	1.000000000	-0.2035685	0.34808725	-0.290014229	0.04074396	-0.09118848	0.003730158
vehicle_age	-0.203568521	1.0000000	-0.30868676	0.598463746	-0.24425626	0.17663654	0.138569341
hp	0.348087254	-0.3086868	1.00000000	-0.345798977	0.16110077	-0.13003692	-0.044289706
milage_clean	-0.290014229	0.5984637	-0.34579898	1.000000000	-0.19085279	0.28199249	-0.004383727
fuel_type_clean	0.040743956	-0.2442563	0.16110077	-0.190852793	1.00000000	-0.10435249	0.002289010
accident_binary	-0.091188483	0.1766365	-0.13003692	0.281992490	-0.10435249	1.00000000	-0.019962674
color	0.003730158	0.1385693	-0.04428971	-0.004383727	0.00228901	-0.01996267	1.000000000

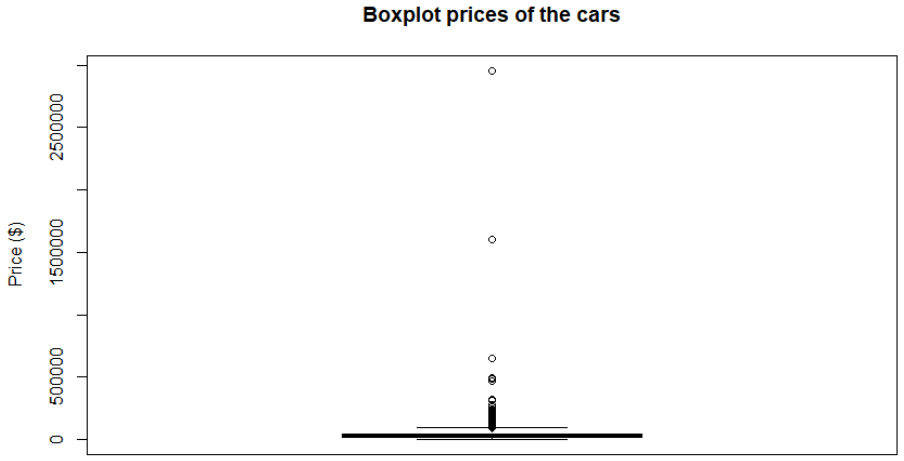
From the analysis, we observed the following:

- **hp** shows a moderate positive correlation with the target variable **price_clean**, indicating that cars with higher horsepower tend to have higher prices.
- **vehicle_age** has a weak negative correlation, suggesting that older vehicles are slightly less expensive.
- **milage_clean** is moderately negatively correlated with **price_clean**, meaning that cars with higher mileage typically have lower prices.
- **fuel_type_clean** and **color** show very weak or negligible correlations, indicating that these factors have little direct impact on car price.
- **accident_binary** has a weak negative correlation, showing a minor price reduction for cars with a history of accidents.

Regarding relationships between explanatory variables:

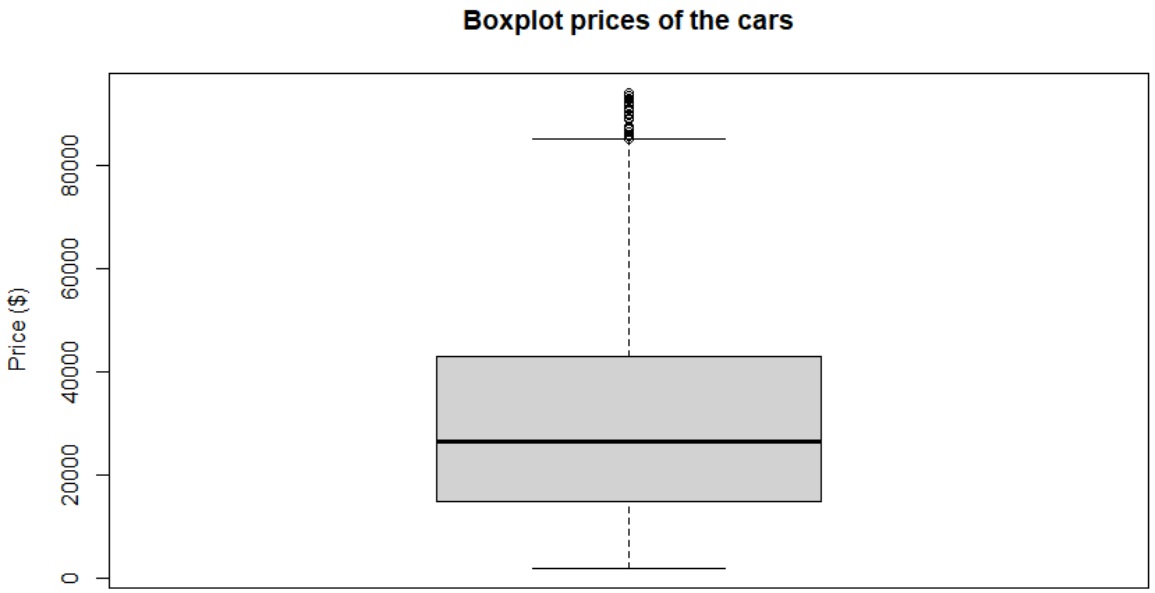
- **vehicle_age** and **milage_clean** have a strong positive correlation, as older cars generally have higher mileage. This strong correlation could potentially lead to multicollinearity issues, so we have to be careful about this variable.

3. **Outlier Detection**
   Outliers in the target variable (price) will be identified and potentially removed, as they can distort regression results.



In the initial boxplot, we observed that some prices were extremely high, which could distort the results of the regression model. To address this, we set a threshold using the Interquartile Range (IQR) method. Prices outside 1.5 times the IQR above the third quartile or below the first quartile were identified as outliers. These extreme outliers were removed from the dataset.

After removing the outliers, we re-plotted the data, resulting in a more reasonable price distribution, which is less likely to bias the regression model. This step ensures that the linear regression model can provide more accurate and reliable predictions by reducing the influence of extreme values.



4. **Validation Strategy**
   To evaluate the model's performance, the dataset will be split into two parts:
   - **Training Set**: 70% of the data will be used to train the model.
   - **Testing Set**: The remaining 30% will be used to test the model on unseen data.

So at the end of our cleaning we had 2927 observations, 2051 (70%) will be used to train the data and 876 (30%) will be used to test the model after.

4. Model Building

After finalizing the model plan, we proceed with building and evaluating the linear regression model.

1. Fitting a Linear Regression Model

Using the **lm()** function in R, we will fit a linear regression model. The model will predict the target variable **price_clean** based on the selected features:

- hp
- vehicle_age
- milage_clean
- fuel_type_clean
- accident_binary
- color

```
Call:
lm(formula = price_clean ~ hp + vehicle_age + fuel_type_clean +
    color + milage_clean + accident_binary, data = train_data)

Residuals:
   Min     1Q Median     3Q    Max
-34506  -7201  -2055   4886  56820

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.548e+04  1.281e+03  19.883  < 2e-16 ***
hp               8.935e+01  2.399e+00  37.244  < 2e-16 ***
vehicle_age     -1.105e+03  5.428e+01 -20.351  < 2e-16 ***
fuel_type_clean -9.790e+02  2.710e+02  -3.612 0.000311 ***
color           -7.584e+01  1.047e+02  -0.724 0.468890
milage_clean    -1.228e-01  5.980e-03 -20.538  < 2e-16 ***
accident_binary -2.398e+03  5.726e+02  -4.188 2.93e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11290 on 2044 degrees of freedom
Multiple R-squared:  0.7017,    Adjusted R-squared:  0.7009
F-statistic: 801.5 on 6 and 2044 DF,  p-value: < 2.2e-16
```

2. Coefficient interpretation and model performance

hp : Each additional unit of horsepower increases the car's price by 89.35 on average, holding other variables constant.

vehicle_age : Each additional year reduces the price by 1105 units on average.

fuel_type_clean : Cars with a different fuel type (e.g., diesel instead of petrol) reduce the price by 979 units on average.

color : The variable **color** does not significantly influence the price (**p-value = 0.469**)

milage_clean : For every 1000 km increase in mileage, the price drops by 123 units on average.

accident_binary : Cars with a history of accidents decrease in value by 2398 units on average.

All the pvalue are significant except for color, so we will remove it and try to see if it can improve the model.

```
Call:
lm(formula = price_clean ~ hp + vehicle_age + fuel_type_clean + 
    milage_clean + accident_binary, data = train_data)

Residuals:
   Min     1Q Median     3Q    Max 
-34752  -7177  -2096   4866  56960 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      2.524e+04  1.239e+03  20.366  < 2e-16 ***
hp               8.941e+01  2.397e+00  37.296  < 2e-16 ***
vehicle_age     -1.111e+03  5.352e+01 -20.763  < 2e-16 ***
fuel_type_clean -9.833e+02  2.709e+02  -3.629 0.000291 ***
milage_clean    -1.223e-01  5.945e-03 -20.581  < 2e-16 ***
accident_binary -2.388e+03  5.724e+02  -4.172 3.14e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11280 on 2045 degrees of freedom
Multiple R-squared:  0.7017,    Adjusted R-squared:  0.7009 
F-statistic: 961.9 on 5 and 2045 DF,  p-value: < 2.2e-16
```

Now all of our variable are significant enough to stay in the model (pvalue below 5%)

Model Performance

- **Multiple R-squared :** About 70% of the variance in car prices is explained by the model.
- **Adjusted R-squared :** Indicates a robust fit, slightly better than the previous model due to the removal of the non-significant variable.
- **F-statistic :** The model is highly significant overall, confirming that at least one predictor has a substantial effect on the price.
- The removal of color which was insignificant has not harmed the model

Interpretation

- **Strong predictors : hp**, **vehicle_age**, and **milage_clean** are the strongest predictors, all with highly significant p-values and reasonable effect sizes.

- **Moderate predictors : fuel_type_clean** and **accident_binary** are still giving meaningful information.

3. Residual Analysis

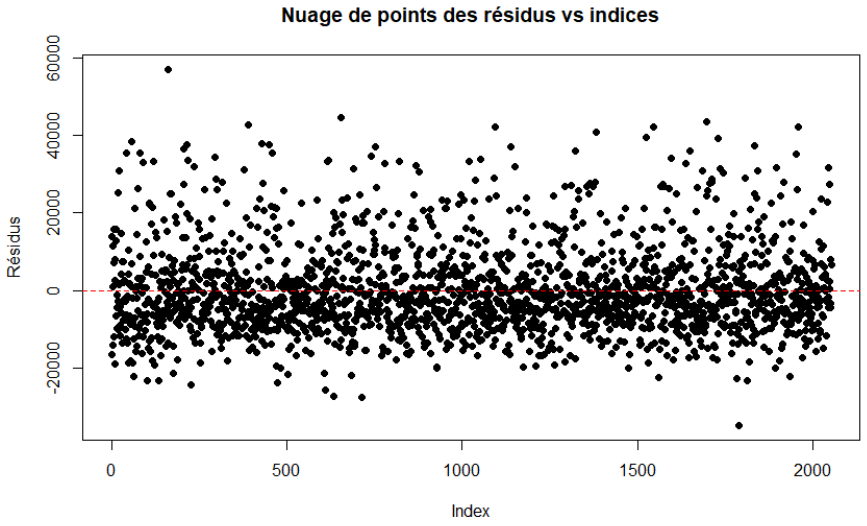
Residual analysis will confirm that the linear regression assumptions are met. Specifically, we will ensure:

- Residuals have a mean close to 0

```
> mean_residuals
[1] -4.269592e-13
```

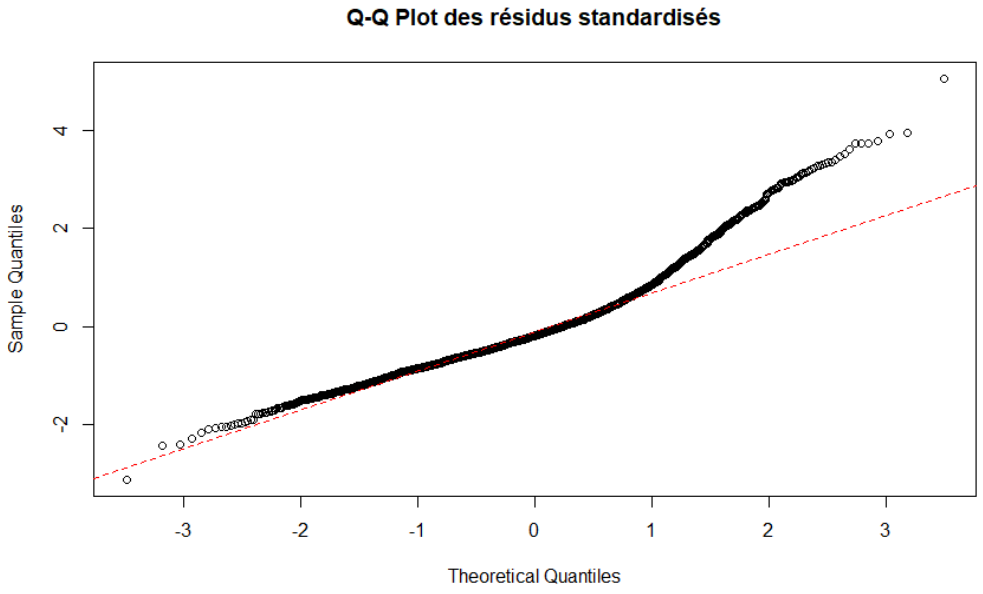
the mean is extremely close to 0, that means our hypothesis is accepted

- Residuals are not autocorrelated



The scatter plot is showing us that the residuals are randomly distributed around 0 meaning that they are not autocorrelated

- Residuals are following a normal distribution



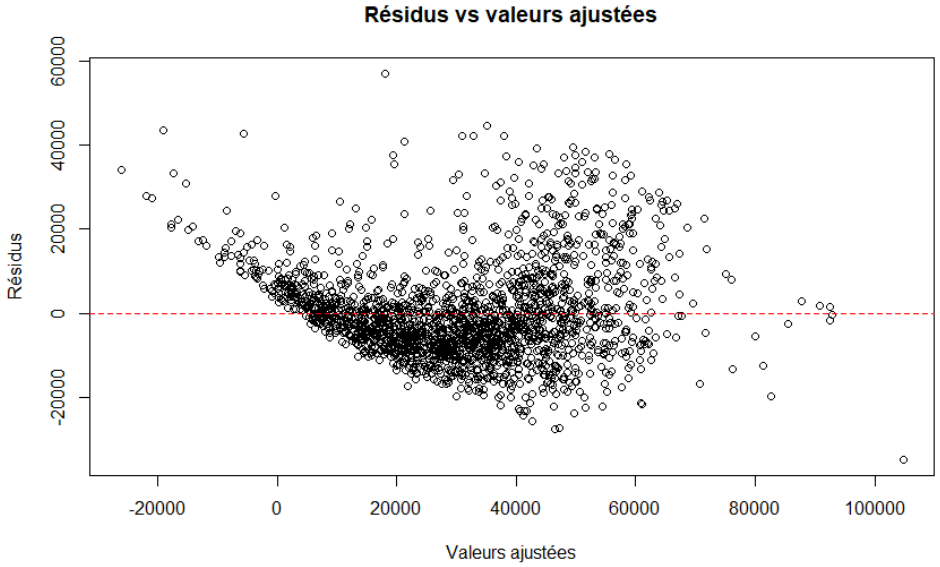
Residuals are globally following the normal distribution, even if we see some exceptions that are not following the line, we will perform a shapiro-wilk test to ensure the normality of the residuals

```
        Shapiro-Wilk normality test

data:  model2$residuals
W = 0.93479, p-value < 2.2e-16
```

pvalue is below 5% so we reject the null hypothesis and we can't confirm that our residuals follow a normal distribution

- Residuals have the same variance



```
        studentized Breusch-Pagan test

data:  model2
BP = 100.82, df = 5, p-value < 2.2e-16
```

Same here we cannot conclude that our residuals have the same variance, we can see on the plot that the variance goes up as the adjusted values goes up too, same for the test with a pvalue below 5% meaning that we have to reject the null hypothesis.

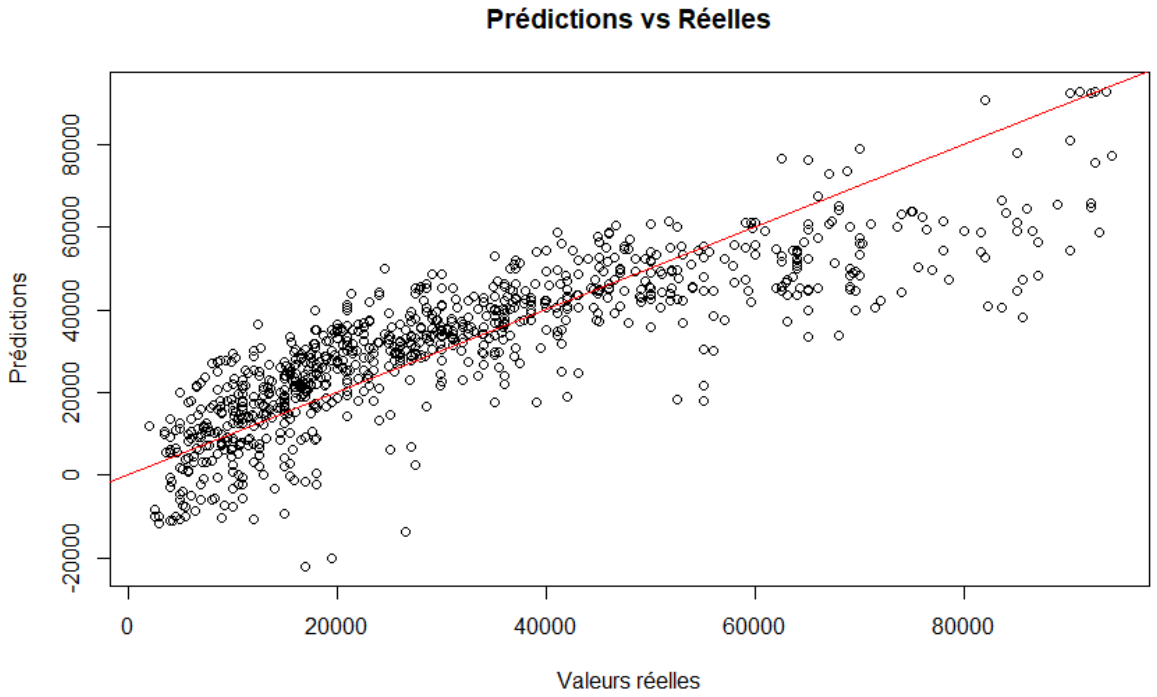
To solve this problem of non normality and variance we tried to apply the GLS method and the log transformation on the target variable (price) but we still had the same problem on our residuals. However in our project it will not create a big issue since the model is pretty robust and that we are trying to predict the price and not doing statistical inference or hypothesis testing. The primary goal of our project is to develop a predictive model for car prices, and as long as the model's predictions are accurate, the non-normality and heteroscedasticity of the residuals should not significantly impact the performance of the model.

The use of generalized least squares (GLS) and log transformation, even though they didn't fully resolve the issues with residuals, helped mitigate potential biases in coefficient estimation, ensuring that the model can still provide reliable predictions.

4. Testing the model on another set

```
    Real Predicted
1  10300  28262.94
3  31000  31326.42
5  41927  44879.86
8  62000  54589.67
9  29990  34816.48
11 11000  16545.89
```

```
> print(paste("RMSE:", rmse))
[1] "RMSE: 11519.3202911932"
> print(paste("R²:", r_squared))
[1] "R²: 0.703763470259386"
> print(paste("MAE:", mae))
[1] "MAE: 8774.66936121709"
```



As we can see on the plot and the previous performance analysis, the model performed reasonably well despite some issues with non-normality and heteroscedasticity in the residuals, with an R^2 of 0.70 and an RMSE of 11,519.32. These results suggest that the model is effective at predicting used car prices, offering valuable insights into the relationship between features such as horsepower, vehicle age, mileage, and accident history. While further refinements, such as addressing the heteroscedasticity, could improve the model, the results are sufficiently robust for our purposes. Therefore, the model can be considered a useful tool for predicting used car prices in a business context.

5. Communicate results

At the end of this project, the main goal was to create a model capable of predicting the price of used cars based on several key characteristics. We used linear regression to analyze the impact of factors like engine power, vehicle age, mileage, fuel type, and accident history.

The results of our model show that several variables play a significant role in the car price. For instance, engine horsepower (HP) and vehicle age are very important indicators. Each increase of one unit in horsepower increases the price by an average of 89.35 units, while each additional year of vehicle age decreases the price by 1105 units. Similarly, mileage is inversely proportional to the price.

Regarding other factors, fuel type also has a notable effect. Vehicles with more eco-friendly fuel types (like electric cars) tend to sell for a higher price, while cars with an accident history have their price reduced by an average of 2398 units. However, the car's color didn't seem to have a significant impact on the price, which led us to remove it from the model.

Despite a few challenges related to the normality of residuals and heteroscedasticity, the model remains robust and reliable for making predictions. These issues didn't significantly affect the model's performance, since our primary goal was to create reliable price predictions rather than perform in-depth statistical analysis. In fact, with an R^2 of 0.70, the model explains a large portion of the variance in car prices. The root mean square error (RMSE) of 11,519.32 is also reasonable, indicating that our predictions are quite accurate.

In conclusion, while further adjustments (particularly to address heteroscedasticity) could slightly improve the model's precision, it is already robust enough to be used in a business context. This model provides valuable insights that can help sellers better assess the price of their used cars, taking into account the most influential factors.