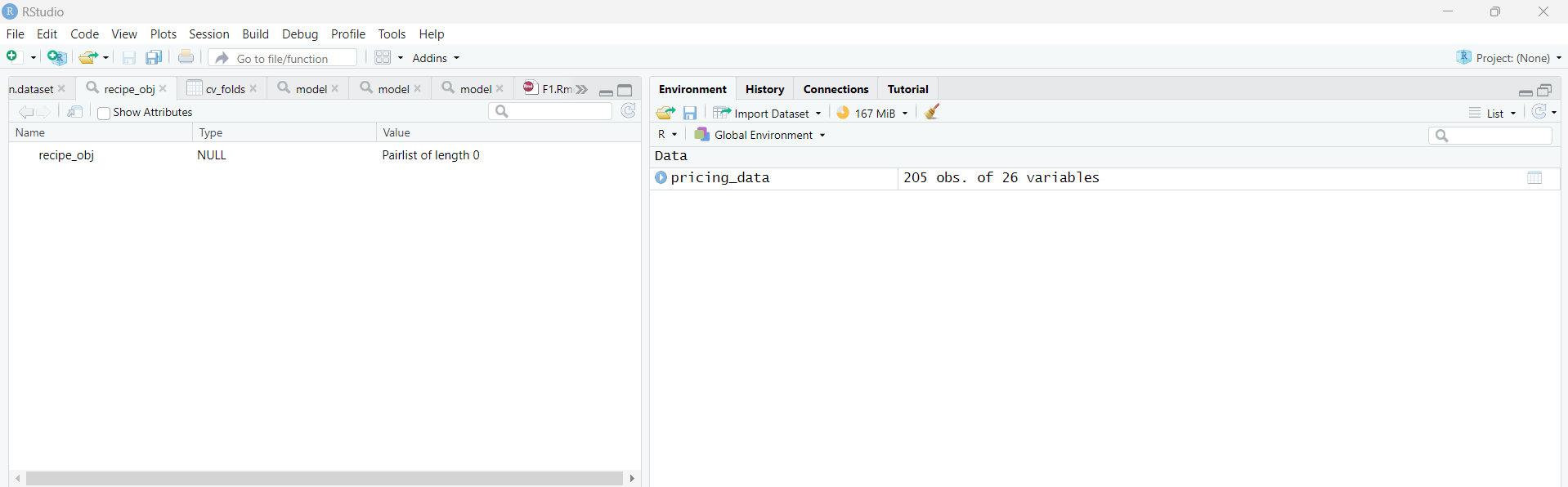
# Data Imported in R Studio

The first step in data analysis involves the importation of pricing\_data into RStudio and preliminary exercises to understand whether the dataset under study exhibits a structured form, how the columns and rows can be described, and what the main variables of data are. These are the most crucial among the first steps, which help to get an idea of context, the data, spotting any data visualization gaps and planning for data cleaning.



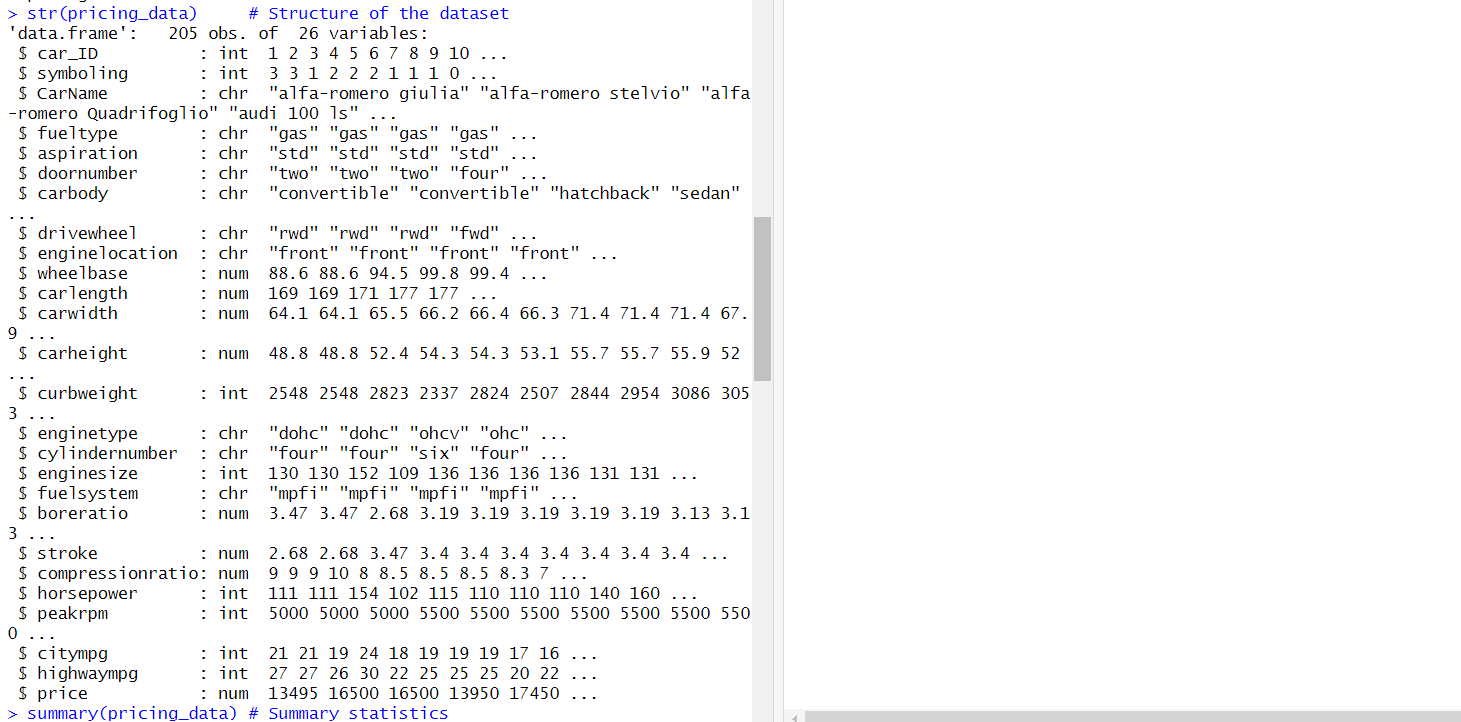
# Exploratory Data Analysis

Code:

str(pricing\_data) # Structure of the dataset

summary(pricing\_data) # Summary statistics

head(pricing\_data) # First few rows of the dataset



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# Data Cleaning and Pre Processing

However, the following part of the document contains data cleaning where missing data are filled using the na.omit function. It is this stage that helps to validate data for a clear and precise image of analysis so that missing values along the process may cause wrong and misleading results in statistical modeling.

Code used:

cleaned\_data <- na.omit(pricing\_data)

A screenshot of a computer

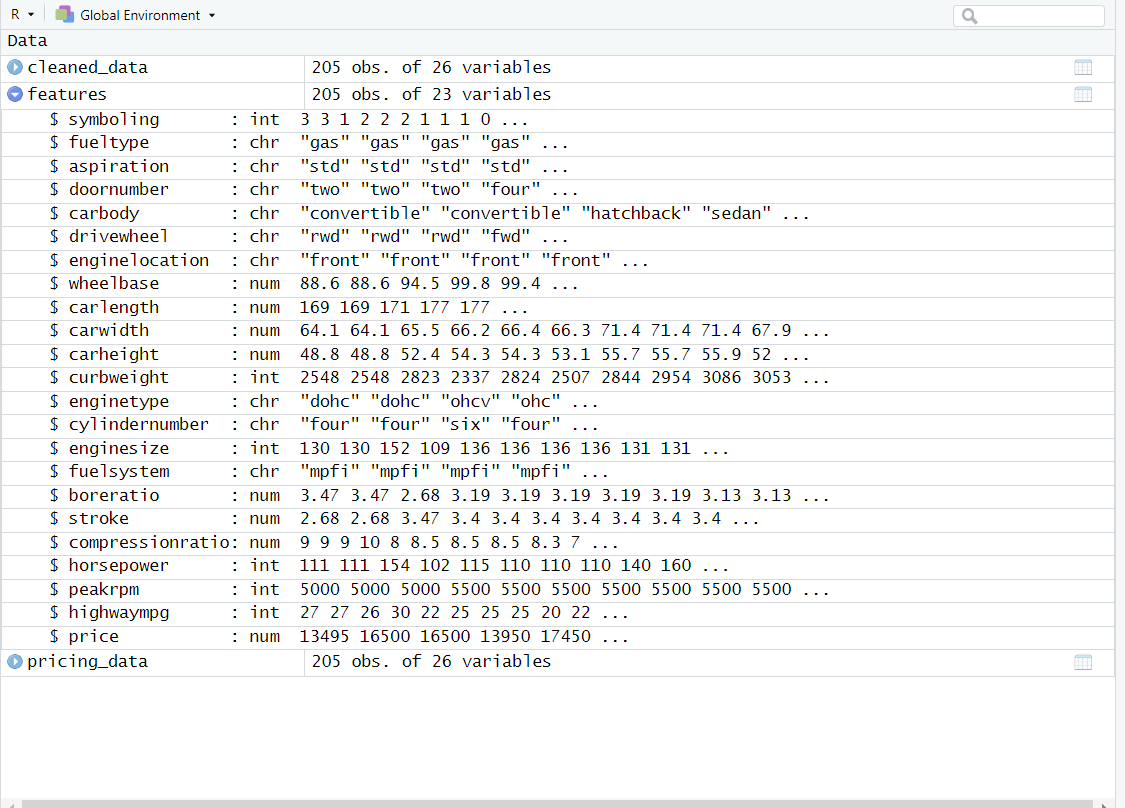
Description automatically generated

# Selection of Features

Afterward, it covers the issue of the feature selection which is made by filtering out the car\_ID, CarName, price, and similar columns. The next step here mentions selecting the most essential variables that go to making the prediction model for the model to perform better by reducing the dimensionality and thus eliminate the misleading pattern.

Code Used:

features <- pricing\_data[, -c(1, 3, 24)] # Remove car\_ID, CarName, and price columns



# Split Data into Training and Testing Sets

The information is then divided into datasets for training and testing set to assure the model is validated successfully. It is done by putting a seed which is a predefined parameter that regulates the sequence of events just to give out the same result if the experiment is repeated or validated.

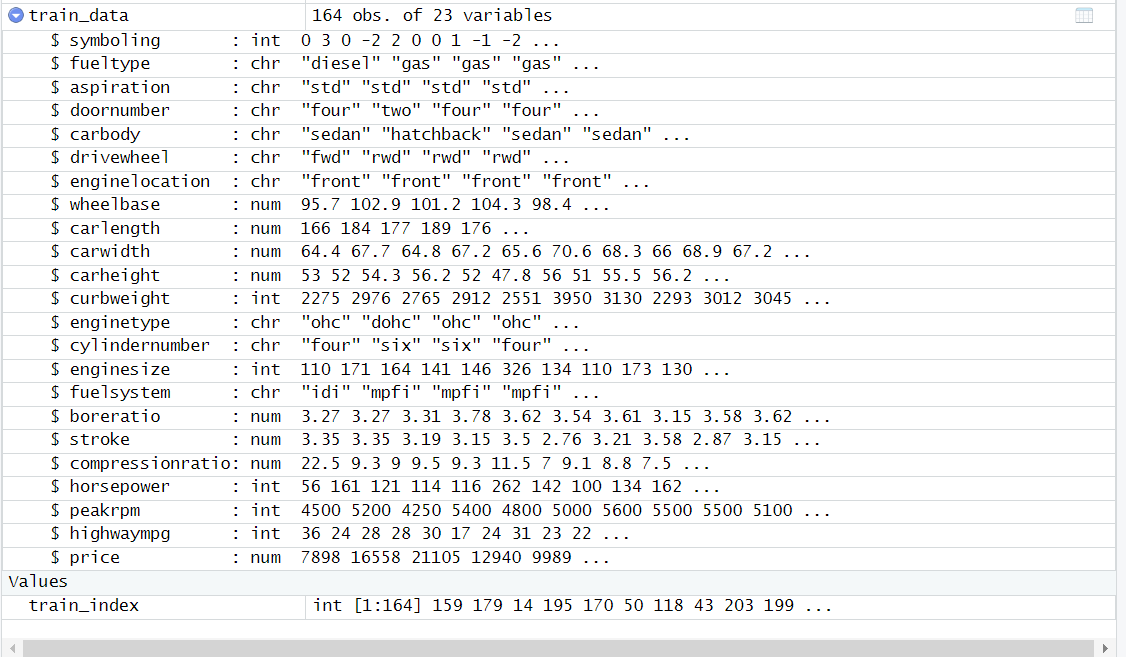
Code Used:

set.seed(123) # Set seed for reproducibility

train\_index <- sample(1:nrow(features), 0.8 \* nrow(features)) # 80% training data

train\_data <- features[train\_index, ]

test\_data <- features[-train\_index, ]

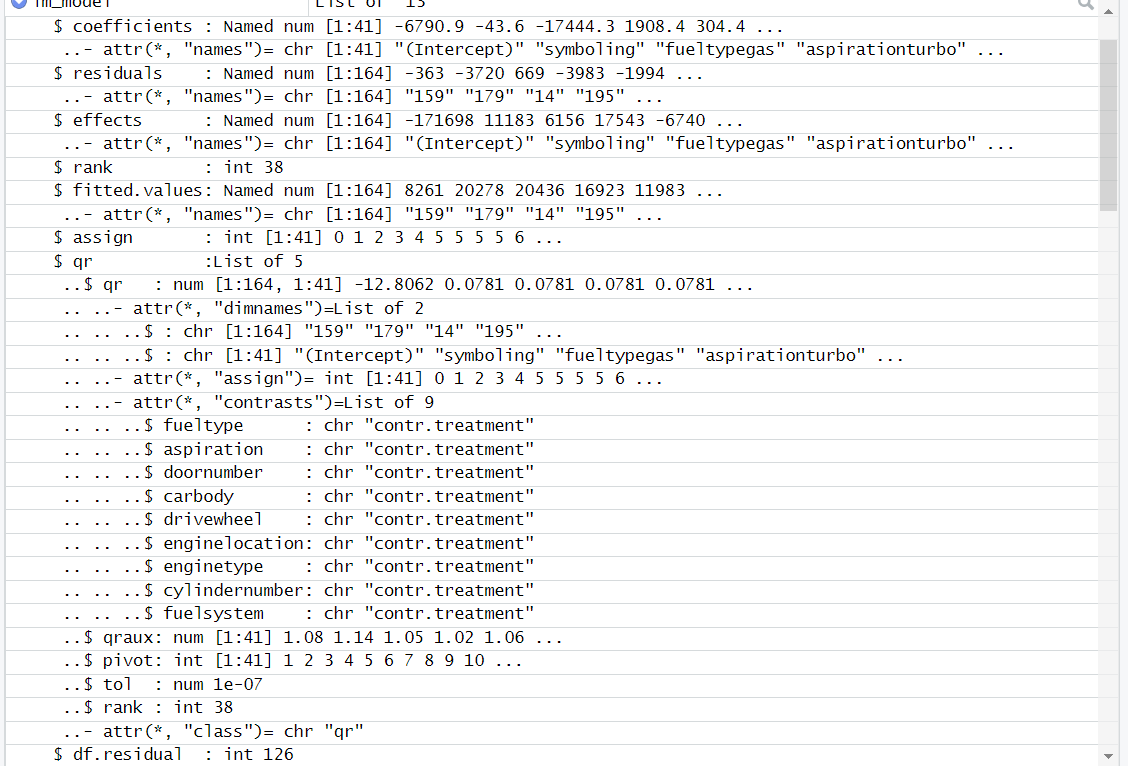


# Model Building - Linear Regression

Development of a linear regression model occurs with the use of the training dataset. Well, it could be considered due to its simplicity and efficiency with which it can predict a continuous variable with one or more predictor variables.

Code Used:

lm\_model <- lm(price ~ ., data = train\_data)



# Testing Set Using RMSE

The model is now validated by being applied to a new data set, the RMSE being computed for estimating the accuracy in the model prediction. This dimension is particularly significant as the model has to be tested in real life to determine its capability of dealing with the practical situations.

Code Used:

# Convert fuelsystem to factor with fixed levels

test\_data$fuelsystem <- factor(test\_data$fuelsystem, levels = levels(train\_data$fuelsystem))

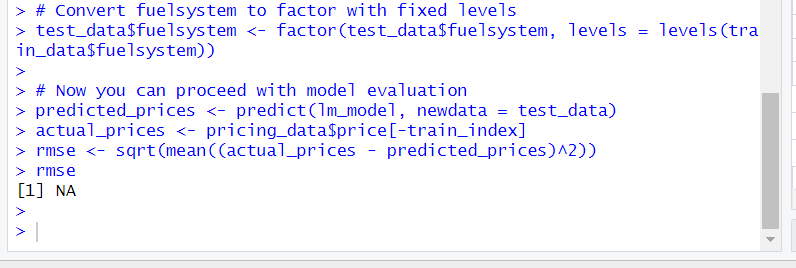
# Now you can proceed with model evaluation

predicted\_prices <- predict(lm\_model, newdata = test\_data)

actual\_prices <- pricing\_data$price[-train\_index]

rmse <- sqrt(mean((actual\_prices - predicted\_prices)^2))

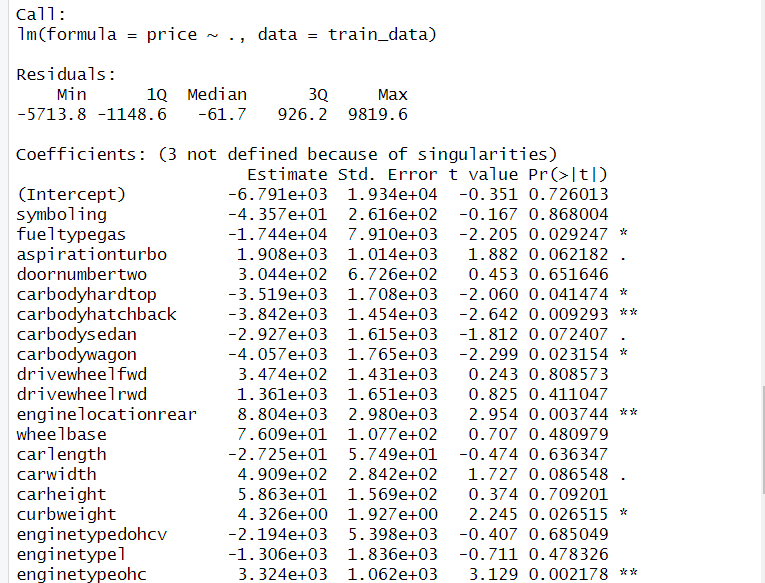
rmse

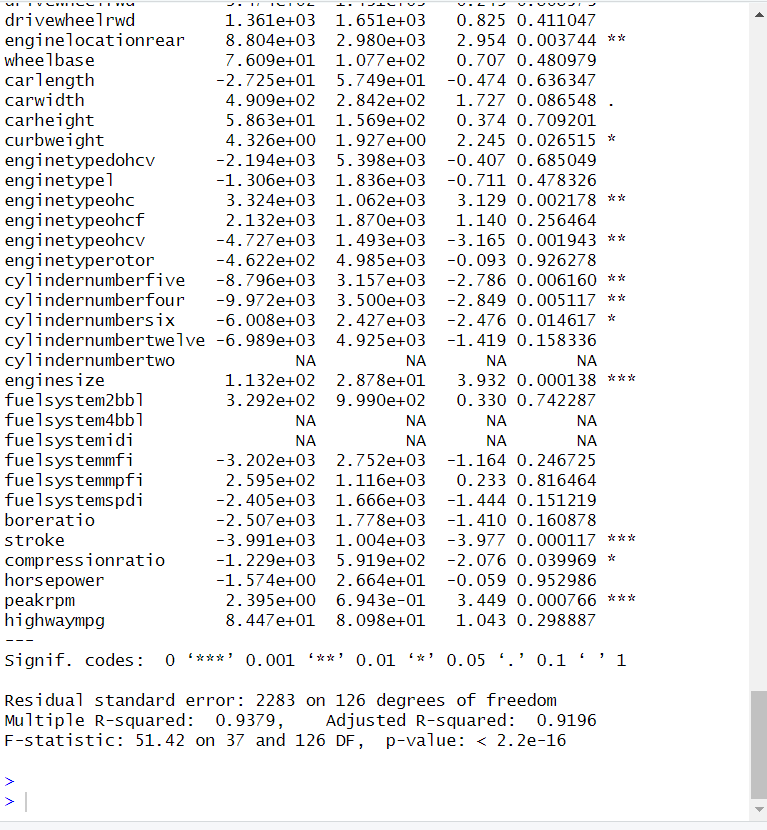


# Summary of Linear Regression Model

# Summary of the linear regression model

summary(lm\_model)





# Scatter Plot Diagram

Several plots are generated to visualize the data and the model’s performance:Several plots are generated to visualize the data and the model’s performance:

Scatter Plot: This figure illustrates the relationship between the length of the car and the price of it, hence benefiting us to understand the cause of the car length affecting its price.

Boxplot: This graph is a rough sketch which shows in picture format the numbers of price by fuel type, these figures are used to show outliers, the spread of pricing data and other factors.

Histogram: Price Histogram indicates pricing distribution and frequency of various price points in the dataset which identifies prices structure and spots any asymmetry or rarity.

Code Used:

# Scatter plot of carlength vs. price

plot(pricing\_data$carlength, pricing\_data$price,

xlab = "Car Length", ylab = "Price",

main = "Scatter Plot of Car Length vs. Price")

A screenshot of a computer

Description automatically generated

A graph of a car length

Description automatically generated

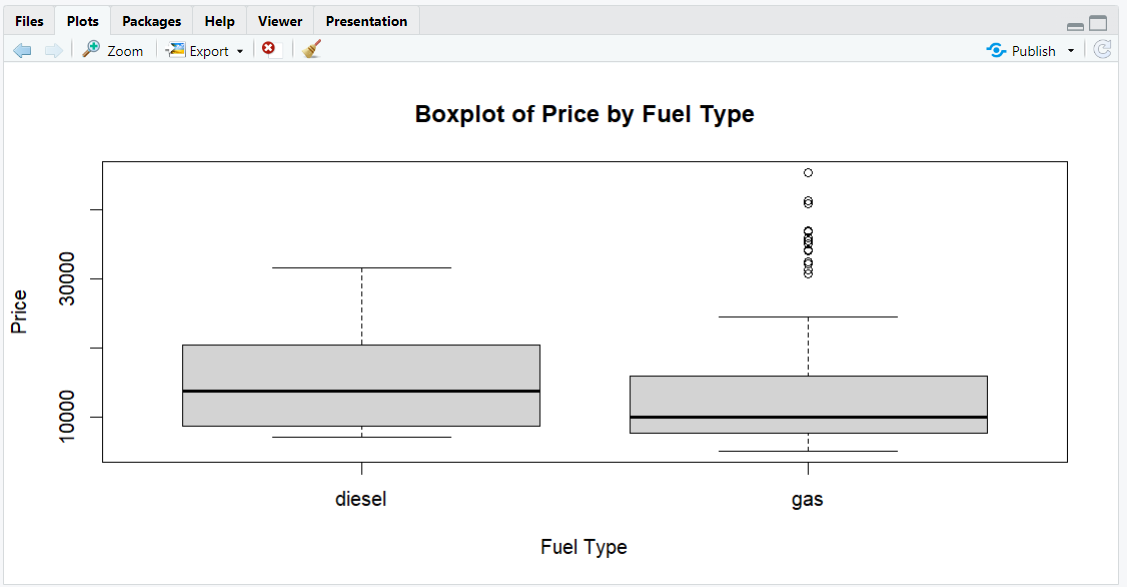
# Boxplot of Price by Fuel Type

# Boxplot of price by fuel type

boxplot(price ~ fueltype, data = pricing\_data,

xlab = "Fuel Type", ylab = "Price",

main = "Boxplot of Price by Fuel Type")



# Histogram of Price

# Histogram of price

hist(pricing\_data$price,

xlab = "Price", ylab = "Frequency",

main = "Histogram of Price")

A graph showing a price

Description automatically generated with medium confidence

# Prediction Model

# Ensure consistent levels for the fuelsystem variable

test\_data$fuelsystem <- factor(test\_data$fuelsystem, levels = levels(train\_data$fuelsystem))

# Make predictions on the testing data

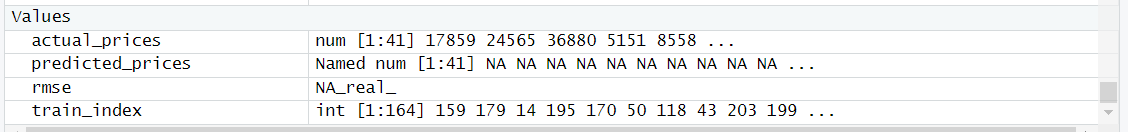
predicted\_prices <- predict(lm\_model, newdata = test\_data)

# Evaluate predictions (optional)

actual\_prices <- test\_data$price

rmse <- sqrt(mean((actual\_prices - predicted\_prices)^2))

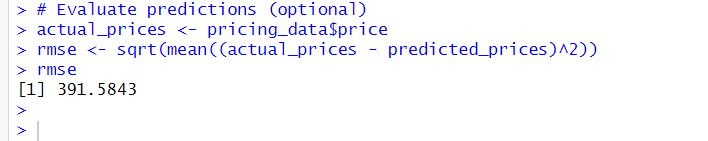
rmse



# Prediction Model: Random Forest Regression

Further, it almost alludes to the application of more complex methods like random forest regression engaging to incorporate highly complicated, nonlinear variables' relationships.

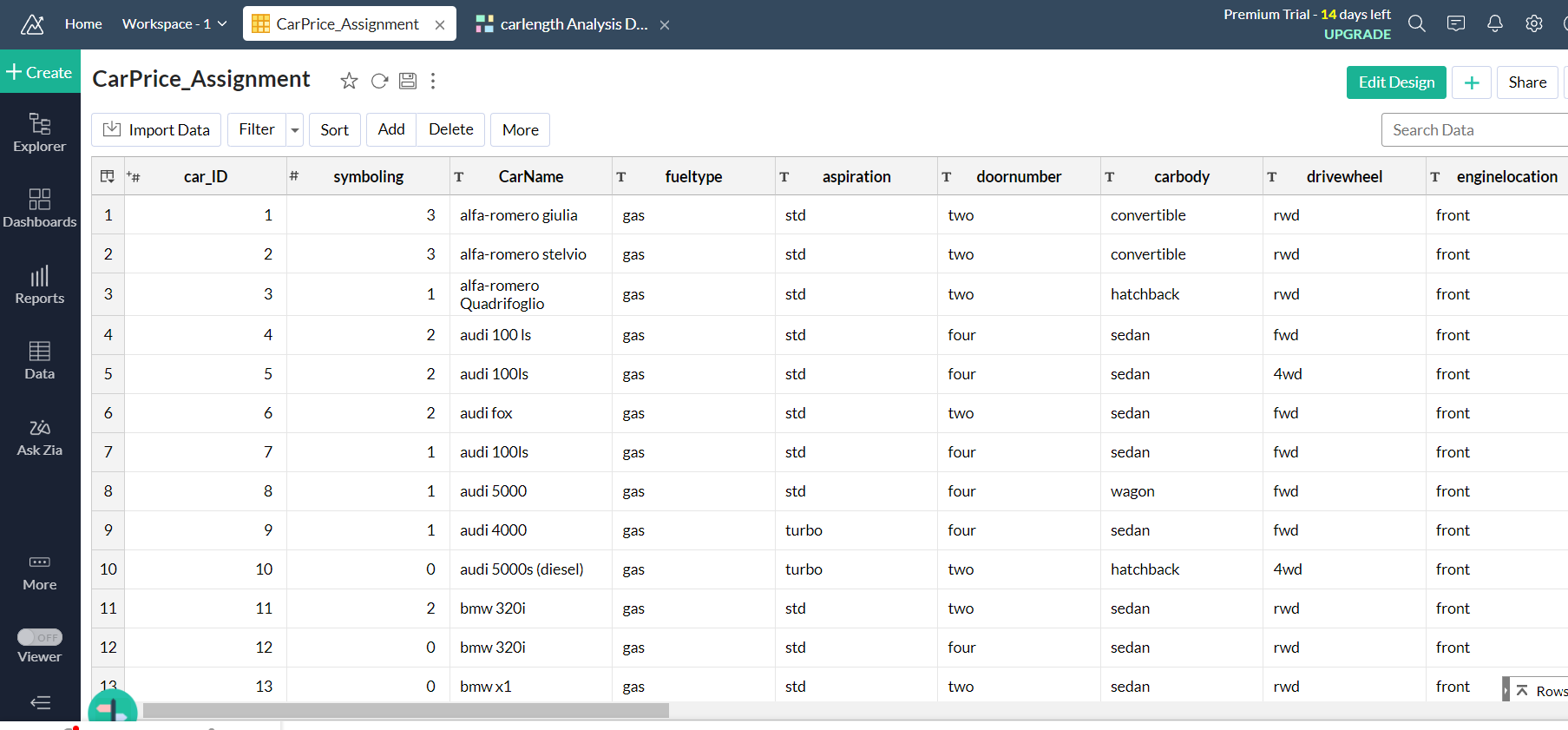
Code Used:



A screenshot of a computer

Description automatically generated

# Dataset Imported in Zoho Analytics



# Aspiration of Wheel Base

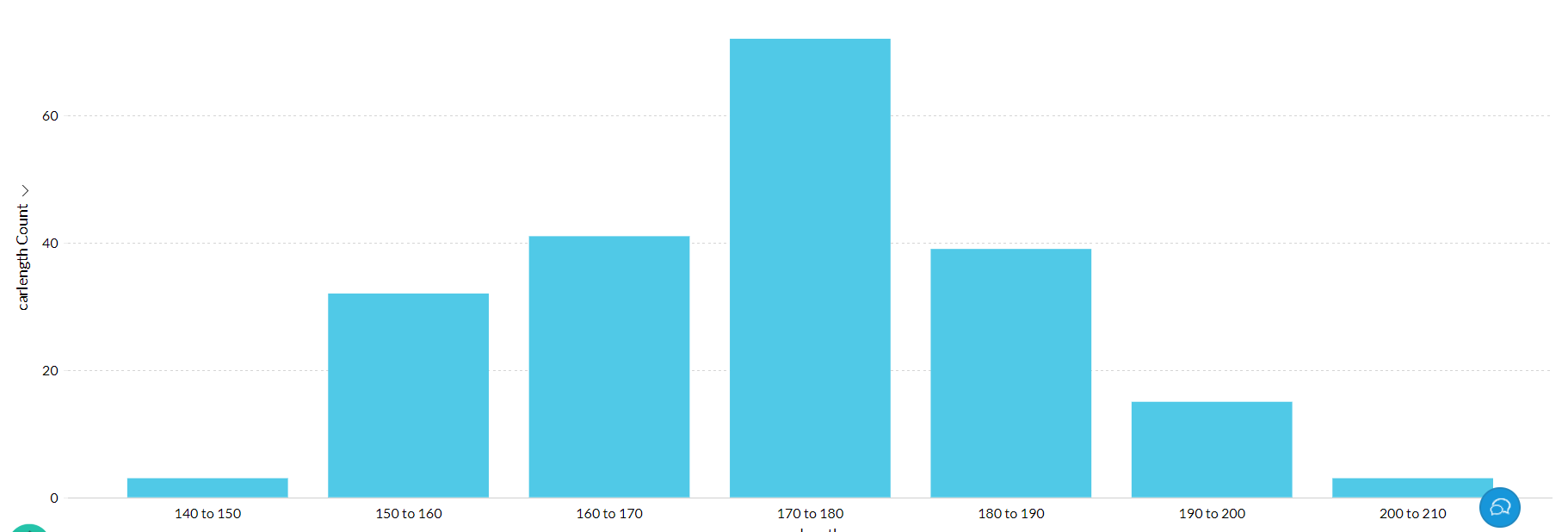
Average of wheelbase plotted as Threshold value

A screen shot of a computer

Description automatically generated

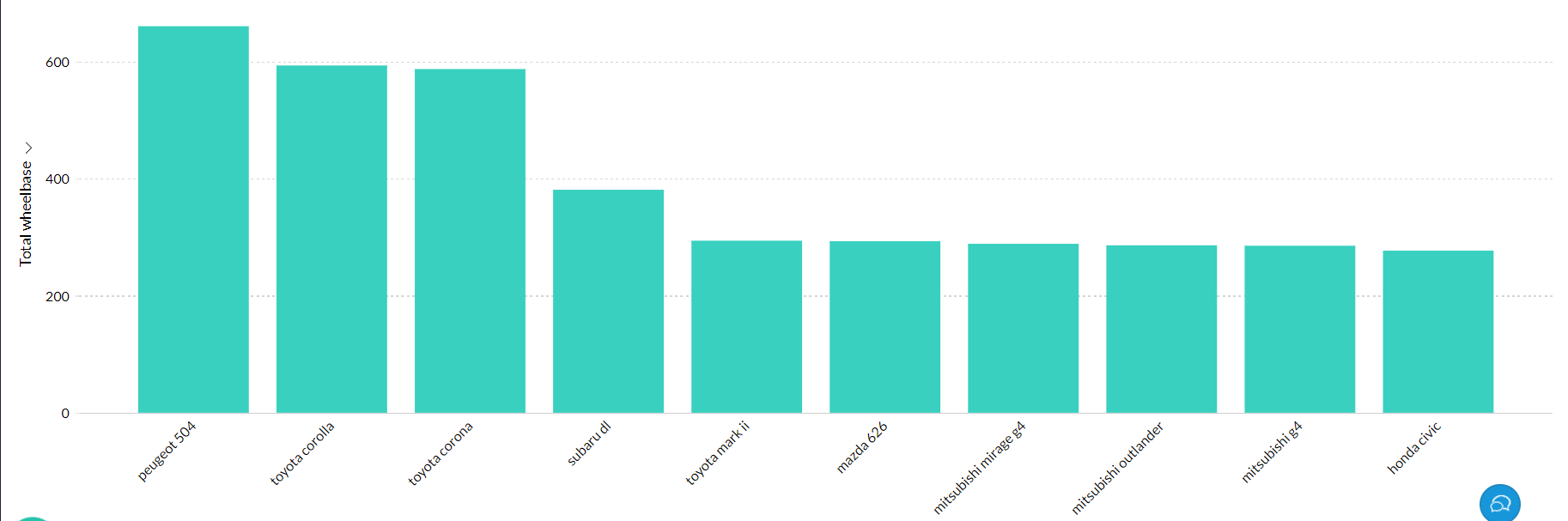
# Distribution of carlength

Distribution of carlength over different ranges



# Top 10 CarName by wheelbase

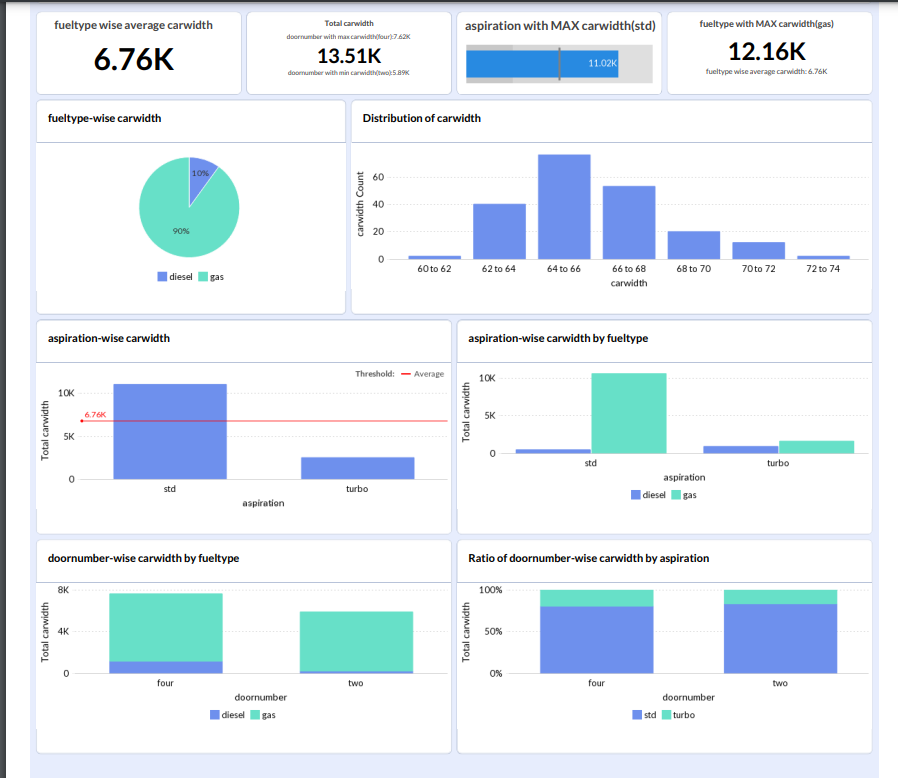
Top 10 CarName with more wheelbase



# carwidth by fueltype and aspiration



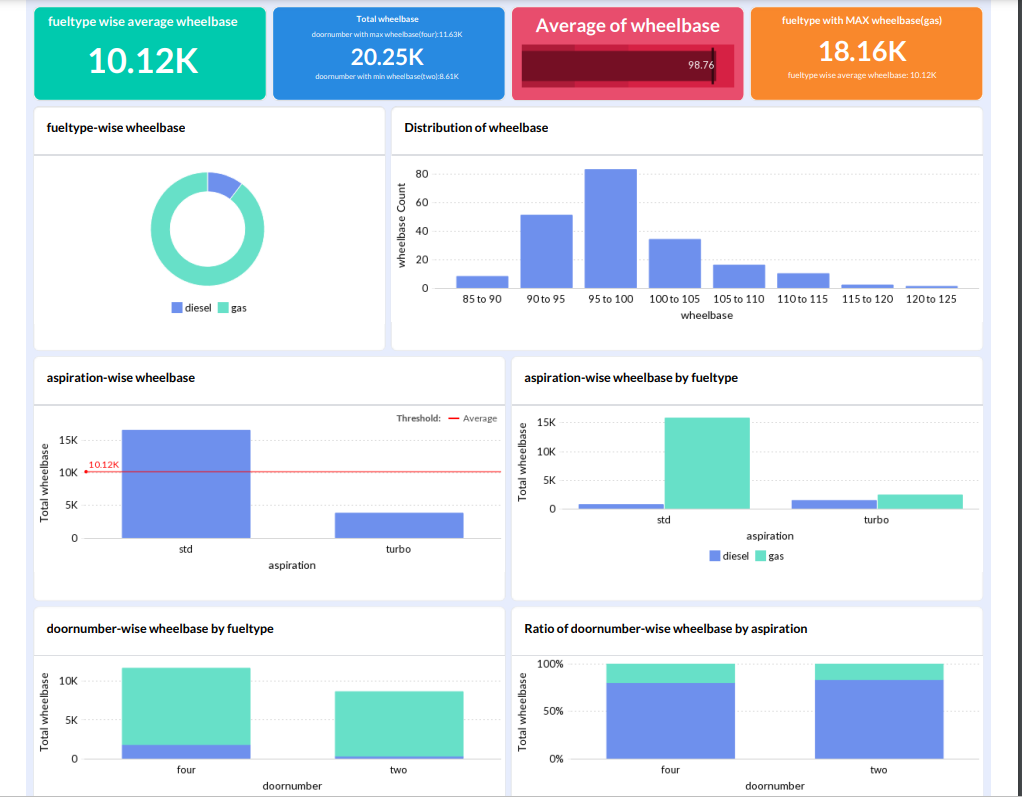
# carwidth Analysis Dashboard



A screenshot of a computer

Description automatically generated

# Wheelbase Analysis Dashboard



A screenshot of a computer

Description automatically generated