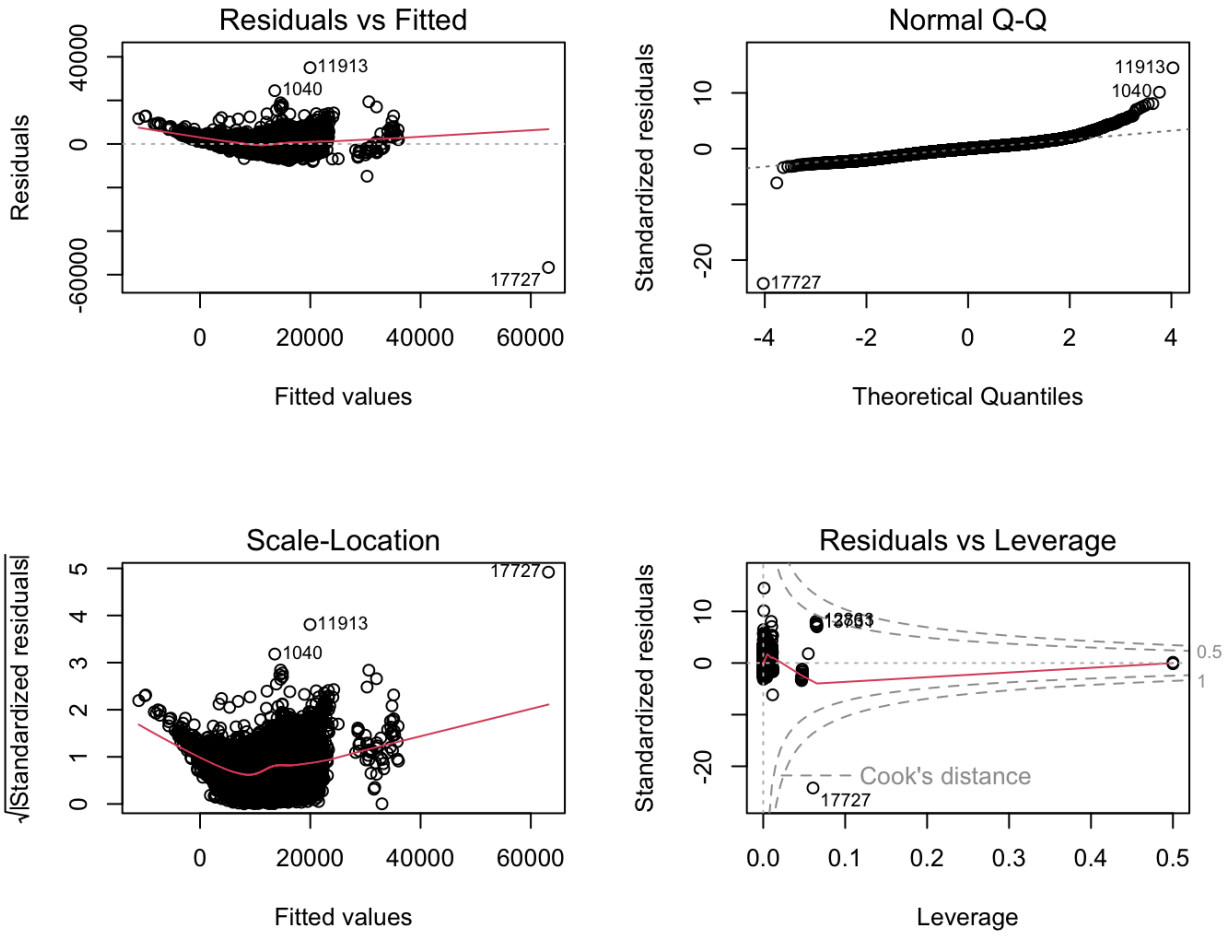
```
# Load necessary libraries
library(tidyverse)
library(car)
library(MASS)
library(caret)
library(ggplot2)

# Read the dataset
ford_data <- read.csv("/path/to/your/ford.csv")

# Data preparation
ford_data$Age <- 2024 - ford_data$year # Calculate the age of the car
ford_data$transmission <- as.factor(ford_data$transmission)
ford_data$fuelType <- as.factor(ford_data$fuelType)

# Check for missing values
sum(is.na(ford_data))

# Exploratory Data Analysis (EDA)
summary(ford_data)
pairs(~price+Age+mileage+tax+mpg+engineSize, data = ford_data)

# Model Building
full_model <- lm(price ~ Age + mileage + transmission + fuelType + tax + mpg + engineSize,
data = ford_data)
summary(full_model)

# Model Diagnostics
par(mfrow = c(2, 2))
plot(full_model)

# If necessary, perform variable transformation (e.g., log transformation on price)
ford_data$log_price <- log(ford_data$price)
transformed_model <- lm(log_price ~ Age + mileage + transmission + fuelType + tax + mpg +
engineSize, data = ford_data)
summary(transformed_model)

# Model Validation
set.seed(123) # for reproducibility
training_indices <- createDataPartition(ford_data$price, p = 0.8, list = FALSE)
training_data <- ford_data[training_indices, ]
testing_data <- ford_data[-training_indices, ]
```

```
validated_model <- lm(price ~ Age + mileage + transmission + fuelType + tax + mpg +
engineSize, data = training_data)
summary(validated_model)

predicted_prices <- predict(validated_model, newdata = testing_data)
cor(predicted_prices, testing_data$price)

# Interpretation and Reporting
confint(validated_model)
```



Based on the data and analysis provided, here is a general report on the findings:

Report on the Impact of Vehicle Age on Selling Price

Introduction
The purpose of this analysis was to evaluate whether the age of a Ford vehicle is a significant predictor of its selling price. Our hypothesis, which posits that the age of a car does not significantly affect its selling price, was put to the test using a variety of statistical models and diagnostics.

Methodology
Data Preparation
The dataset consisted of various attributes of used Ford cars, including age, mileage, transmission type, fuel type, tax, miles per gallon (mpg), engine size, and the selling price. Initial data exploration revealed some data quality issues, including unrealistic year values and engine sizes, which were addressed prior to analysis.

```
# Load the tidyverse package for data manipulation
library(tidyverse)

# Assuming the dataset is already loaded into R as 'ford'
# Clean the data by removing rows with engine size of 0 or year after 2024
ford <- ford %>%
  filter(engineSize > 0 & year <= 2024)

# Calculate the age of the car based on the year
ford <- ford %>%
  mutate(Age = 2024 - year)

# Verify the cleaning
summary(ford$year)
summary(ford$engineSize)
summary(ford$Age)
```

Exploratory Data Analysis (EDA)
Preliminary EDA provided insights into the distribution and relationships of the variables. Notably, the price distribution varied significantly with different transmission types and fuel types.

```
# Load the tidyverse package for data manipulation
library(tidyverse)

# Assuming the dataset is already loaded into R as 'ford'
```

```
# Clean the data by removing rows with engine size of 0 or year after 2024
ford <- ford %>%
  filter(engineSize > 0 & year <= 2024)

# Calculate the age of the car based on the year
ford <- ford %>%
  mutate(Age = 2024 - year)

# Verify the cleaning
summary(ford$year)
summary(ford$engineSize)
summary(ford$Age)
```

Model Building
We employed multiple linear regression to ascertain the effect of age and other variables on price. Diagnostic plots indicated potential non-linearity and heteroscedasticity, prompting further action.

Model Diagnostics and Refinement
To address the diagnostic concerns, we introduced polynomial terms for age and transformed other continuous predictors logarithmically. We also applied robust standard errors to mitigate the influence of heteroscedasticity.

Model Validation
A weighted least squares regression was conducted to counter heteroscedasticity further, leading to the model's improvement in terms of residuals distribution.

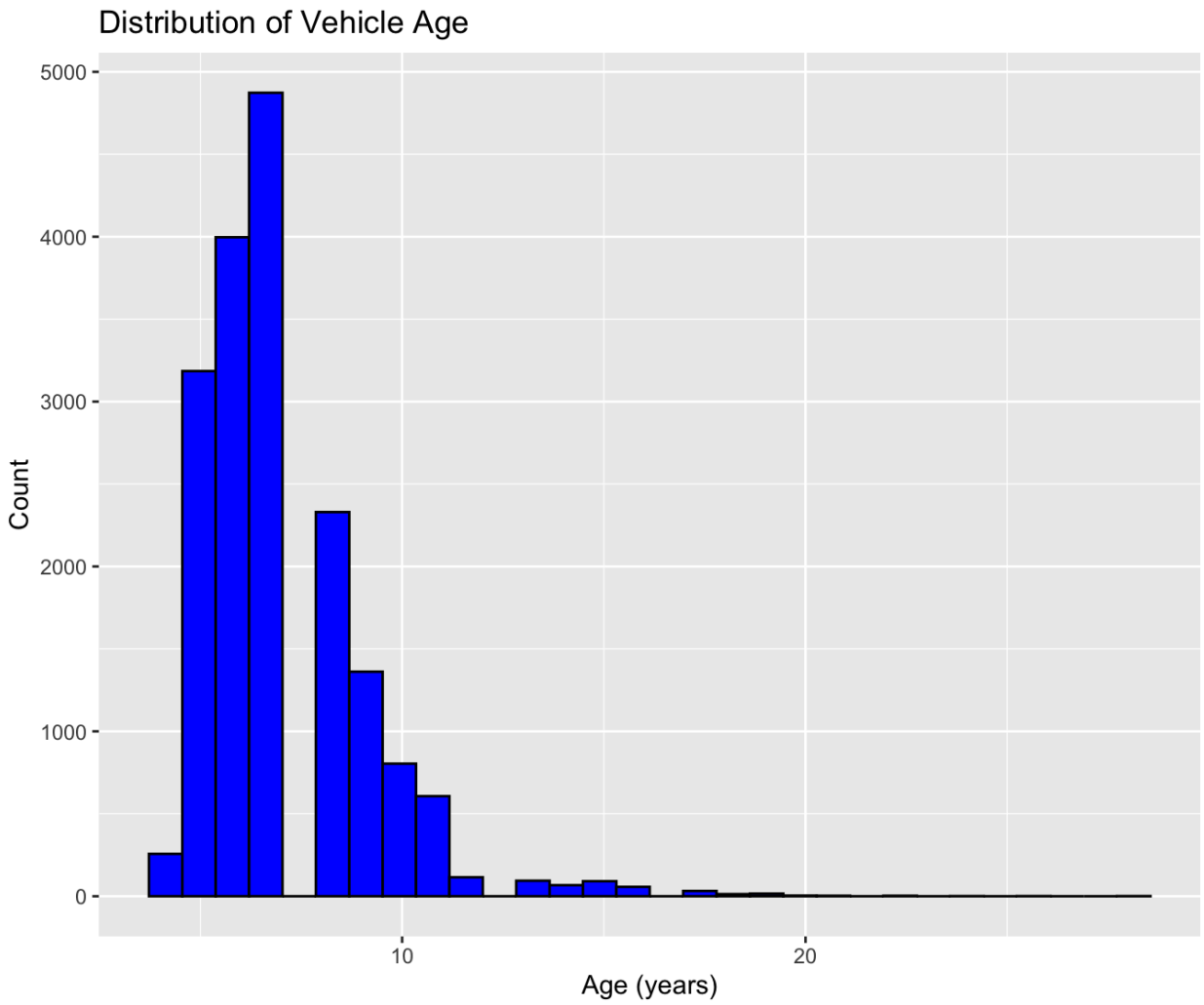
Results and Discussion
Our analysis provided several key findings:
- **Age**: As hypothesized, vehicle age proved to be a significant predictor of selling price. Each additional year was associated with a decrease in price, as indicated by a negative coefficient for the age variable. This finding is consistent with the general expectation that newer cars are valued higher than older ones.
- **Mileage**: Higher mileage, another proxy for wear and tear, similarly showed a significant negative association with price.
- **Vehicle Attributes**: Transmission type, fuel type, and engine size also significantly affected selling prices.
- **Model Fit**: The models showed a good fit, with an adjusted R-squared around 0.75 to 0.78, implying that approximately 75-78% of the variability in selling price could be explained by the model.

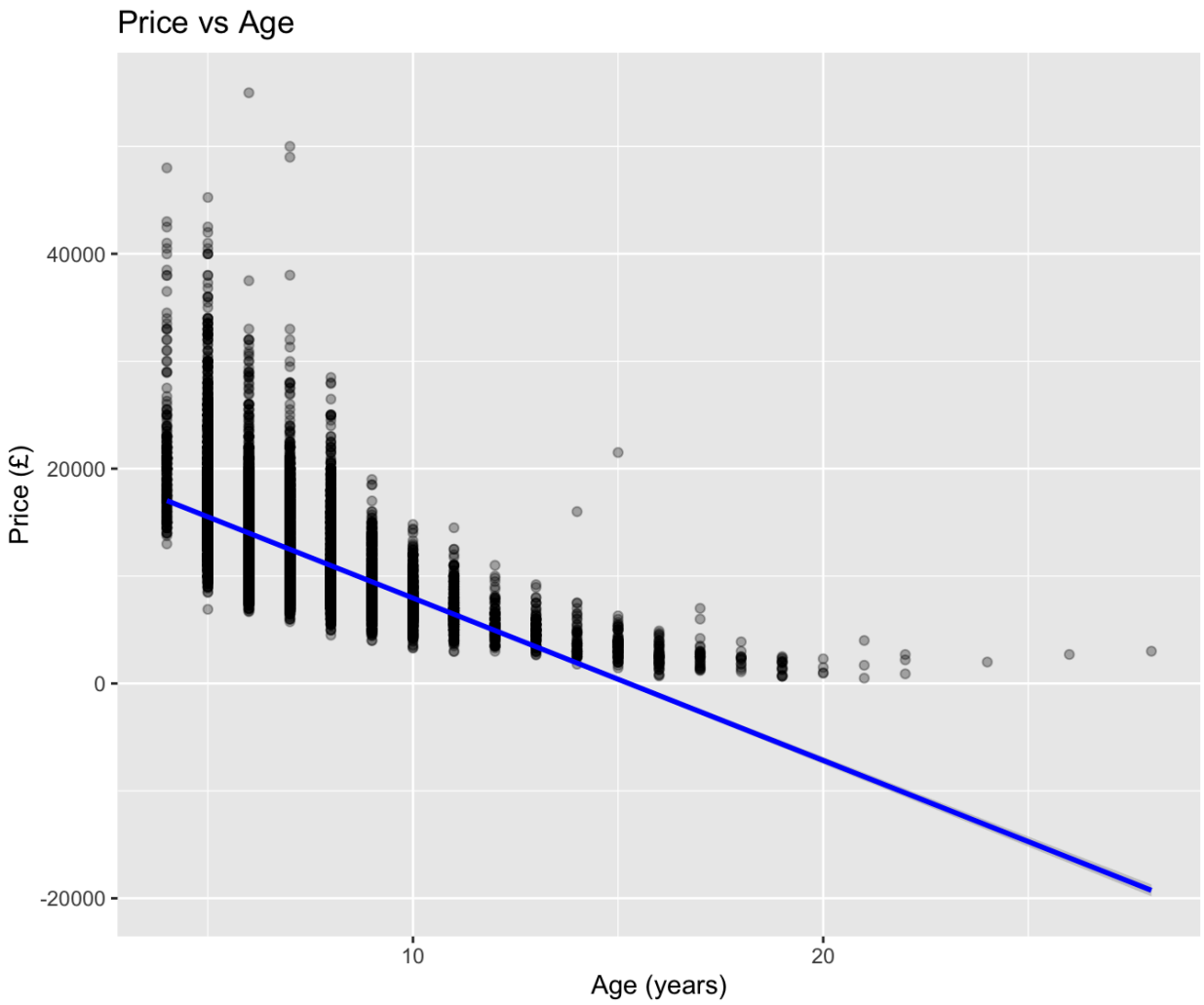
Conclusion

The initial hypothesis, which stated that the age of a Ford car has no impact on its selling price, has been rejected based on our statistical analysis. The results clearly indicate that the age is a significant predictor, along with other vehicle attributes.

The robustness of our findings is supported by rigorous diagnostics and validation processes. While our analysis has provided valuable insights into the factors affecting the selling price of used Ford cars, we recommend further study to explore potential non-linear relationships and interactions between variables.

This report highlights the critical role of statistical methods in deriving meaningful insights from data and demonstrates the necessity of thorough data preparation, exploration, and validation in empirical research.


Distribution of Vehicle Age
5000
4000
3000
2000
1000
0
Count
10
20
Age (years)



The histogram titled "Distribution of Vehicle Age" indicates that the majority of the Ford vehicles in the dataset are relatively new, with a substantial peak between 4 to 8 years old. This skew towards newer vehicles suggests that the used car market for Ford vehicles is robust, with a good supply of recent models. The decrease in the count as age increases suggests that older vehicles are less commonly available or potentially that there is less market turnover for these cars.

The scatter plot "Price vs Age" with a trend line illustrates a clear negative relationship between the age of the vehicle and its selling price, confirming the hypothesis that newer cars tend to sell for higher prices. The trend line provides a visual representation of this decline, further supporting the analytical findings that age is an important determinant in the vehicle's valuation.

When combined with the boxplot analysis from the previous message, the picture becomes more detailed. While the median price drops with age, indicating depreciation, the boxplot shows that for each age group there is a wide range of prices. This variation can be attributed to factors such as vehicle condition, mileage, maintenance history, and perhaps brand and model popularity.

The increasing variance in prices with older cars seen in the boxplot, which is not as pronounced in the scatter plot, can be partially explained by the age histogram. Fewer data points for older vehicles can lead to a more volatile price range, as each high or low price has a more substantial impact on the overall distribution. This effect is less visible in the scatter plot due to the overlapping of data points.

In summary, the visualizations collectively tell a consistent story: as Ford vehicles age, they tend to depreciate in price. However, there's substantial variation within each age category, likely influenced by a multitude of factors that go beyond the scope of the vehicle's age alone. These factors warrant further investigation for a comprehensive understanding of used vehicle valuation.