Student: Arthur Grossmann—Le Mauguen H00494101

F79MB: Statistical Models B

Assessment Project 2 2025 – Edinburgh; Academic Year: 2024-2025

Task 1

Question A:

The data represents the characteristics of 200 vehicles from various brands in 2022, including variables such as brand, age, fuel type, price, and vehicle type. Below are the numerical and graphical summaries with key findings:

1) Vehicle Price (in USD):

Minimum (in USD)	1st Quartile (in USD)	Median (in USD)	Mean (in USD)	3rd Quartile (in USD)	Maximum (in USD)	Standard deviation (in USD)	Interquartile range (IQR) (in USD)	Mode (in USD)
281	6,925	13,992	17,882	25,925	69,999	14,335.23	19,000	7,995

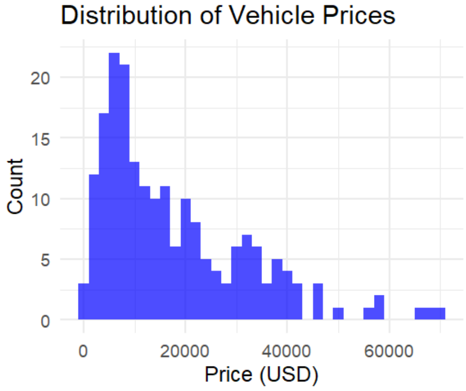


Figure 1: *Summary statistics of vehicle price (in USD)*

Figure 2: *Histogram of vehicle price distribution (in USD)*

The histogram of vehicle prices shows a **right-skewed** distribution, where most vehicles are priced below $30,000 (75% of car sold are 25,925$ or less), while a few high-priced models (up to $69,999) inflate the mean. This is confirmed by the mean ($17,882) being higher than the median ($13,992) and a relatively high maximum. Outliers are present, particularly in the upper tail, contributing to price variability.

2) Vehicle Age (as of 2022):

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Standard deviation	Interquartile range (IQR)	Mode
2	5	9	10.18	13	85	7.780557	8	5

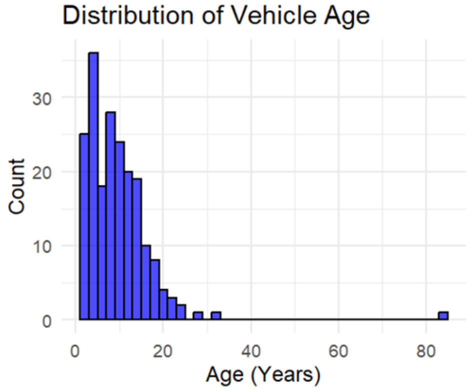


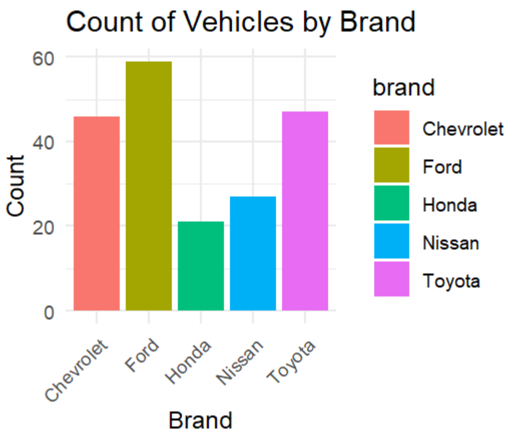
Figure 3: *Summary statistics of vehicle age* (*in years*)

Figure 4: *Histogram of vehicle age* (*in years*)

The histogram of vehicle age shows a **right-skewed** distribution, which is supported by the summary statistics: the median age being lower than the mean. The fleet appears relatively recent, with three quarters of the vehicles being less than 13 years old. However, there is a notable outlier, a vehicle that is 85 years old, which significantly inflates the mean. This is likely a vintage or classic car included in the dataset.

3) Vehicle brand:

Figure 5: Bar plot of the vehicle distribution by manufacturer



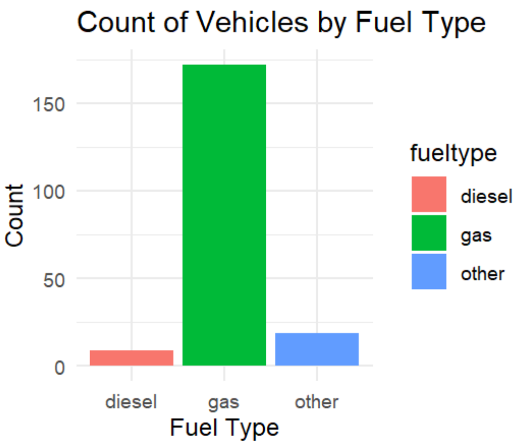
Figures 5 and 6 illustrate the distribution of vehicle brands within the dataset, which is notably **unbalanced**. Ford is the most represented brand, accounting for nearly a third of the vehicles (29.5%). Toyota and Chevrolet follow closely, representing 23.5% and 23% of the sample, respectively. Nissan appears less frequently, making up only 13.5% of the vehicles, while Honda is the least represented brand, with just 10.5%.

Brand	Count	Frequency (%)
Ford	59	29.5
Toyota	47	23.5
Chevrolet	46	23.0
Nissan	27	13.5
Honda	21	10.5

Figure 6: Frequency table of vehicle manufacturers

4) Vehicle fuel type:

Figure 7: *Bar plot of the fuel type distribution*



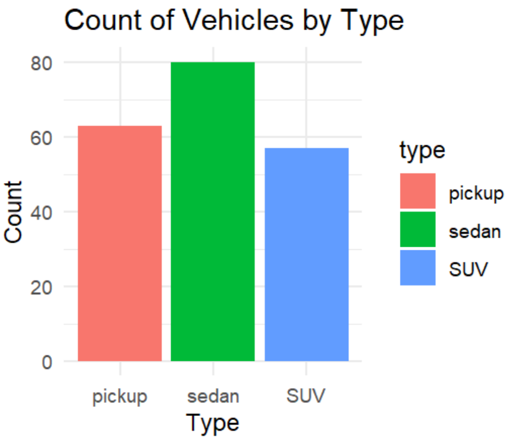
The distribution of vehicle fuel types is highly **unbalanced**. The majority of vehicles (86%) run on gasoline, while diesel-powered vehicles are rare, making up only 4.5% of the sample. The remaining 9.5% fall into the "other" category, which may include GPL, hybrid, or electric vehicles, but they still represent a minority.

Figure 8: Frequency table of fuel types

Fuel Type	Count	Frequency (%)
Gas	172	86.0
Other	19	9.5
Diesel	9	4.5

5) Vehicle Type:

Figure 9: *Bar plot of the vehicle counts by type*



The distribution is relatively even. Sedans are the most common, representing 40% of the vehicles. Pickups follow with 31.5%, while SUVs are the least common, making up 28.5% of the total.

Figure 10: Frequency table of vehicle types

Vehicle Type	Count	Frequency (%)
Sedan	80	40.0
Pickup	63	31.5
SUV	57	28.5

Observations:

Vehicle price and age are both right skewed with notable outliers, suggesting a few old or expensive vehicles disproportionately affect the averages.
The dataset is dominated by gas-powered vehicles, which may reflect the general market in 2022. There is brand imbalance too, with Ford, Toyota, and Chevrolet together making up more than 75% of the dataset.

Question B:

Note: for visualization purposes, outliers will not be displayed in the following scatter plot.

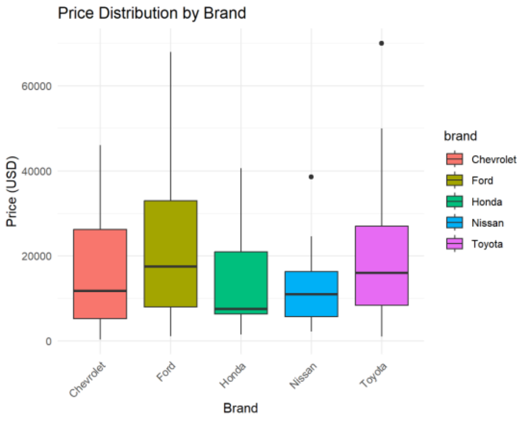


Figure 11: Boxplot of the vehicle price for each brand

Figure 11 depicts the link between vehicle prices and their respective brands. Although median prices are generally close across brands, Ford clearly has the highest median around $17,500 and the largest price spread. Nissan, by contrast, has the tightest price range and the most consistent distribution, reflecting minimal variation in its vehicle prices. All other brands exhibit positively skewed price distributions, with Honda standing out: its median price is positioned near the first quartile, meaning that half of Honda's vehicles are priced below $7,500, despite a wider price range than that of Nissan.

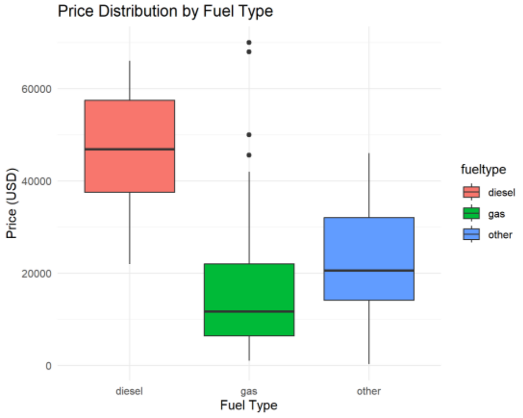


Figure 12: Boxplot of the vehicle price for each fuel type

Figure 12 explores how vehicle prices vary by fuel type. The distribution is highly **imbalanced**, with gas-powered vehicles making up 86% of the dataset. Diesel vehicles are markedly more expensive than the rest, with a median price of $46,900, and even the least expensive ones cost more than 75% of gas vehicles. In contrast, gas vehicles are the most economical option, with half priced below $11,700. Vehicles using other fuel types fall in the middle: their median price is $20,600, and their price range closely resembles that of gas vehicles. However, due to the small sample size for non-gas vehicles, these observations should be interpreted with caution.

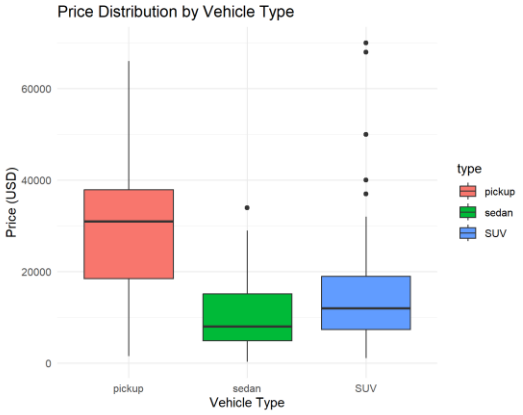


Figure 13: Boxplot of the vehicle price for each vehicle type

Figure 13 compares vehicle prices across different vehicle types. Sedans and SUVs have similarly right-skewed price ranges, though SUVs command a slightly higher median price at $12,000 compared to the $8,000 for sedans. A subset of SUVs reaches premium prices, with some at nearly $70,000, the highest in the dataset. Pickups, however, follow a distinct trend, with a **left-skewed** distribution and significantly higher prices. Half of all pickups cost $31,000 or more, placing 75% above the majority of sedans. Their price range is also the widest among the vehicle types analysed.

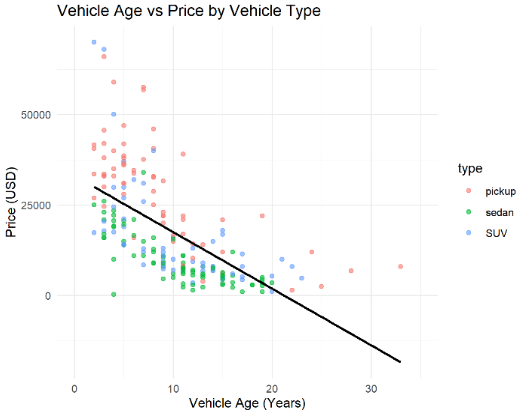


Figure 14: Scatter Plot of vehicle price by age and type of vehicle

Figure 14 shows the relationship between vehicle age and price. An outlier (an 85-years old car) was excluded to maintain scale and readability. The plot demonstrates a clear negative correlation: as vehicles age, their prices generally decrease. This trend is most noticeable between 2 and 14 years, where prices decline almost linearly. New vehicles are the most expensive, with none older than 10 years priced above $40,000. On the other end, vehicles under 6 years old rarely cost less than $10,000, except for one priced at $281 the lowest in the dataset so it might be a mistake. Most entries are concentrated around newer, lower-priced vehicles, indicating a dataset dominated by recent models under $30,000. Some older vehicles still command high prices, likely due to rarity or collector interest.

Question C:

The objective is to estimate a 99% confidence interval for the Pearson correlation between vehicle age and price in the population, using Fisher's transformation.

The Pearson correlation is defined as: $r = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$.

In this case, the correlation between price and age computed in R is r = -0.414, indicating a moderate negative or inverse relationship in the sample. A Pearson correlation near ±1 suggests a strong linear relationship, and the negative sign here implies that as the age of the vehicle increases, its price tends to decrease.

To apply Fisher's transformation, we compute: $W = \frac{1}{2} log\left(\frac{1+r}{1-r}\right)$. Using R, we get W = -0.440.

For a 99% confidence interval, we use the critical value from the standard normal distribution: $Z_{0.005}$ = 2.56. The standard error for the Fisher transformation is SE=$\frac{1}{\sqrt{\mathrm{N}-3}}$. Given a sample size of N = 200, we get: SE=$\frac{1}{\sqrt{197}}$. So, the confidence interval for W is: $\mathrm{W} \pm \mathrm{Z}_{0.005} * \frac{1}{\sqrt{\mathrm{N}-3}} = \mathrm{W} \pm 2.56\frac{1}{\sqrt{197}}$.

Finally, we apply the inverse Fisher transformation to convert the interval back to the correlation scale. Using R, we find the 99% confidence interval for the true correlation to be: (-0.554; -0.251). This means we can be 99% confident that the true correlation between vehicle age and price in the population lies between a weak to moderate negative correlation. However, these results should be interpreted with caution. While 200 observations is a reasonably large sample, it may still be sensitive to outliers. For instance, the dataset includes an 85-year-old vehicle priced at $30,000 and a 4-year-old vehicle sold for just $281, both of which could significantly influence the correlation and the resulting interval.

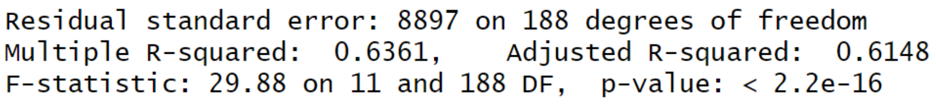
Question C bis:

Instead of using Fisher's transformation, we can compute the confidence interval for Pearson's correlation using bootstrap resampling. Resampling is a statistical technique that involves repeatedly drawing samples from the observed data, with replacement, to estimate the sampling distribution of a statistic. In the case of the bootstrap, we generate a large number of "resampled" datasets and compute the correlation coefficient for each one. This method does not rely on the assumption of normality and is more robust to outliers and non-linearity.

Question D:

A **multiple linear regression model** was used to investigate the impact of various features on vehicle prices. In this model, the dependent or **response variable** is the vehicle price, while the **explanatory (independent) variables** include age, type, fuel type, and brand. An interaction term between age and type was incorporated, but no such terms were added for fuel type or brand. The model was fitted in R using the lm() function with the formula: ***price ~ age * type + fueltype + brand***

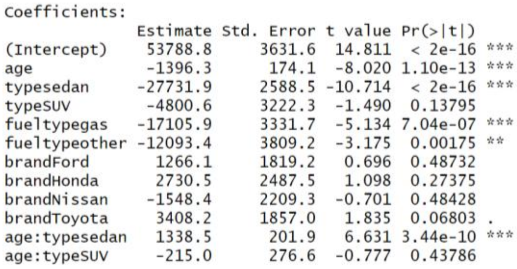
Figure 15: Regression model summary



```
Residual standard error: 8897 on 188 degrees of freedom
Multiple R-squared:  0.6361,    Adjusted R-squared:  0.6148
F-statistic: 29.88 on 11 and 188 DF,  p-value: < 2.2e-16
```

The output of the model indicates that the F-statistic strongly favors the fitted model over the null, with a p-value below 2.2×10^{-16}. This suggests that the model captures a substantial portion of the variation in vehicle prices. The R^2 value is 0.6361, meaning approximately 63.61% of the variability in price is explained by the model.

Figure 16: Each term p-value coefficient

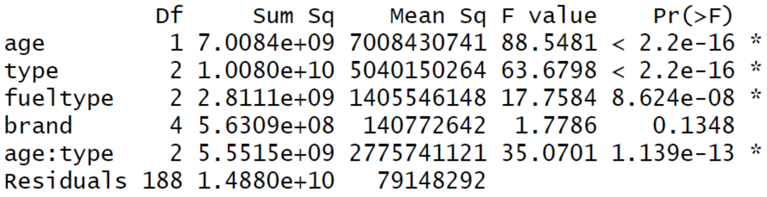


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    53788.8     3631.6  14.811  < 2e-16 ***
age            -1396.3      174.1  -8.020 1.10e-13 ***
typesedan     -27731.9     2588.5 -10.714  < 2e-16 ***
typeSUV        -4800.6     3222.3  -1.490  0.13795    
fueltypegas   -17105.9     3331.7  -5.134 7.04e-07 ***
fueltypeother -12093.4     3809.2  -3.175  0.00175 ** 
brandFord       1266.1     1819.2   0.696  0.48732    
brandHonda      2730.5     2487.5   1.098  0.27375    
brandNissan    -1548.4     2209.3  -0.701  0.48428    
brandToyota     3408.2     1857.0   1.835  0.06803 .  
age:typesedan   1338.5      201.9   6.631 3.44e-10 ***
age:typeSUV     -215.0      276.6  -0.777  0.43786    
```

Each coefficient in the model was tested for statistical significance using p-values obtained from individual t-tests. For each predictor, the null hypothesis (H_0) states that its coefficient is equal to zero, meaning the variable has no effect on the response (vehicle price). The alternative hypothesis (H_1) posits that the coefficient is different from zero, indicating a significant impact on price. A variable is considered statistically insignificant if its p-value exceeds the conventional 0.05 threshold, suggesting we fail to reject H_0. In this case, the coefficients for brandFord (p = 0.487), brandHonda (p = 0.274), brandNissan (p = 0.484), and brandToyota (p = 0.068) all have p-values above 0.05. This means we do not have sufficient evidence to reject H_0 for any of these brand categories. Consequently, the *brand* variable was deemed statistically insignificant and excluded from the model to improve interpretability and reduce unnecessary complexity. Although the p-value for typeSUV is relatively high, the type variable was retained due to the significant contribution of its other levels. The most statistically significant variables include typesedan ($p < 2.2\times10^{-16}$), age ($p = 1.10\times10^{-13}$), fueltypegas ($p = 7.04\times10^{-7}$), and the interaction age:typesedan ($p = 3.44\times10^{-10}$).

To further assess the model, an analysis of variance was performed using the anova() function in R. The ANOVA confirmed that age ($p < 2.2\times10^{-16}$), type ($p < 2.2\times10^{-16}$), fueltype ($p = 8.62\times10^{-8}$), and the interaction between age and type ($p = 1.14\times10^{-13}$) were all highly significant, whereas brand (p = 0.135) was not. This justified the removal of the brand variable.

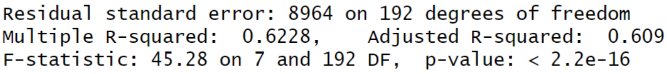
Figure 17: Model's variance analysis



```
           Df     Sum Sq    Mean Sq F value    Pr(>F)  
age         1 7.0084e+09 7008430741 88.5481 < 2.2e-16 *
type        2 1.0080e+10 5040150264 63.6798 < 2.2e-16 *
fueltype    2 2.8111e+09 1405546148 17.7584 8.624e-08 *
brand       4 5.6309e+08  140772642  1.7786    0.1348  
age:type    2 5.5515e+09 2775741121 35.0701 1.139e-13 *
Residuals 188 1.4880e+10   79148292                    
```

A second model, referred to as Model 2, was fitted with the formula: ***price ~ age * type + fueltype***

Figure 18: New model's summary

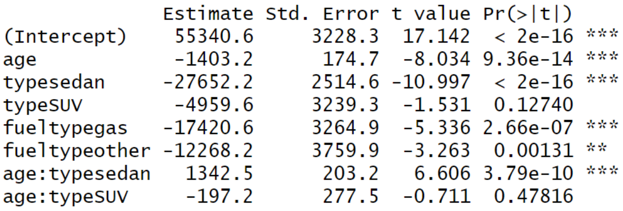


```
Residual standard error: 8964 on 192 degrees of freedom
Multiple R-squared:  0.6228,	Adjusted R-squared:  0.609 
F-statistic: 45.28 on 7 and 192 DF,  p-value: < 2.2e-16
```

After removing the brand variable, the R^2 slightly decreased from 63.61% to 62.28%, a minimal change that supports this simplification.

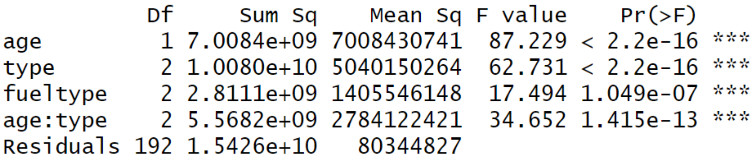
In Model 2, age (p = 9.36e-14), typesedan ($p < 2\times10^{-16}$), fueltypegas (p = 2.66e-07), and age:typesedan (p = 3.79e-10) remained highly significant. Although fueltypeother had a higher p-value (p = 0.00131), it was still considered significant enough to keep. While typeSUV (p = 0.12740) and age:typeSUV (p = 0.47816) appeared insignificant individually, removing the SUV category was deemed inappropriate as it might capture trends that other variables miss.

Figure 19: Each term p-value coefficient for then new model



```
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    55340.6     3228.3  17.142  < 2e-16 ***
age            -1403.2      174.7  -8.034 9.36e-14 ***
typesedan     -27652.2     2514.6 -10.997  < 2e-16 ***
typeSUV        -4959.6     3239.3  -1.531  0.12740    
fueltypegas   -17420.6     3264.9  -5.336 2.66e-07 ***
fueltypeother -12268.2     3759.9  -3.263  0.00131 ** 
age:typesedan   1342.5      203.2   6.606 3.79e-10 ***
age:typeSUV     -197.2      277.5  -0.711  0.47816    
```

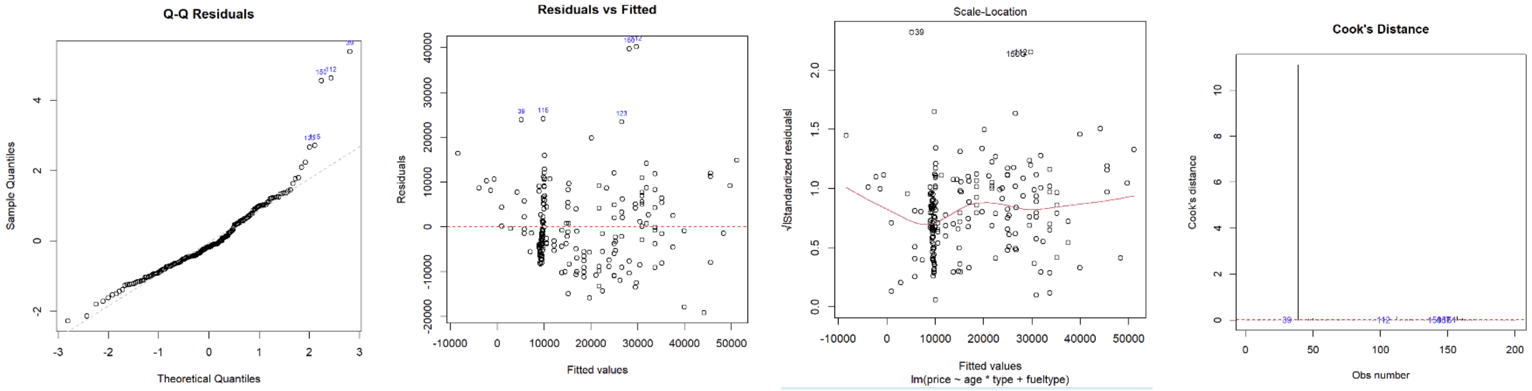
Figure 20: New model's variance analysis



```
           Df     Sum Sq    Mean Sq F value    Pr(>F)    
age         1 7.0084e+09 7008430741  87.229 < 2.2e-16 ***
type        2 1.0080e+10 5040150264  62.731 < 2.2e-16 ***
fueltype    2 2.8111e+09 1405546148  17.494 1.049e-07 ***
age:type    2 5.5682e+09 2784122421  34.652 1.415e-13 ***
Residuals 192 1.5426e+10   80344827                      
```

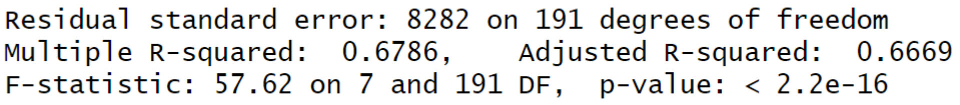
The ANOVA results table for Model 2 reaffirmed that age, type, fueltype, and their interaction were significant contributors to the model.

Figure 21: New model's residuals plot



The residual diagnostic plots for Model 2 were examined next. The Q-Q plot indicated **approximate normality in residuals**, except for some deviation at the tails. The Residuals vs Fitted plot showed **no distinct patterns**, with residuals centered around zero, a positive sign. The Scale-Location plot suggested **consistent variance across fitted values**. However, the Cook's Distance plot identified observation 39 as highly influential, with a distance above 10. This point corresponds to a vehicle aged 85 years, previously flagged as an outlier (see Question 1(b)).

Model 3 was then built by excluding observation 39 and using the formula: ***lm(price ~ age * type + fueltype, data = vehicle_clean)***

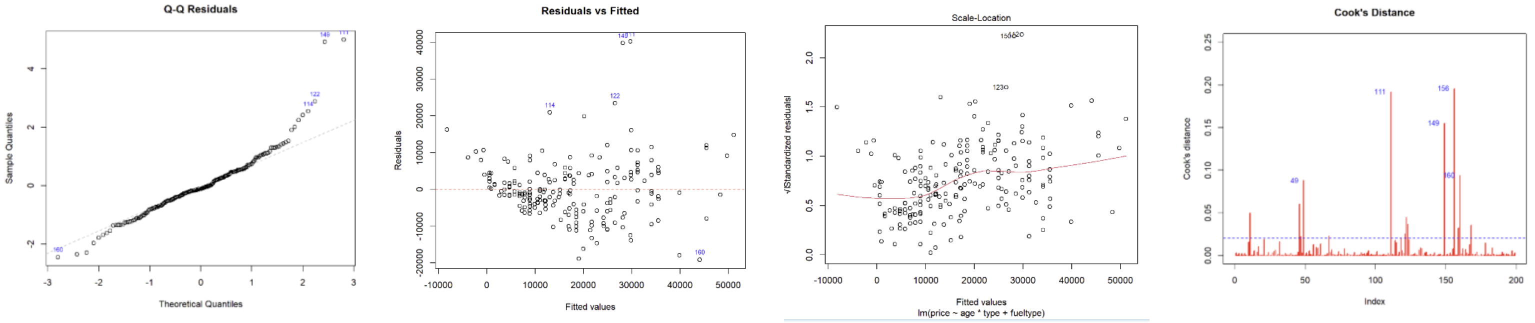


```
Residual standard error: 8282 on 191 degrees of freedom
Multiple R-squared:  0.6786,    Adjusted R-squared:  0.6669
F-statistic: 57.62 on 7 and 191 DF,  p-value: < 2.2e-16
```

Figure 22: Model 3's variance summary

This adjustment led to a noticeable improvement in model performance. Model 3's coefficient of determination R² increased from 62.28% to 67.86%, showing that removing this influential point had a major positive effect.

Figure 23: 3rd model's residuals plot



Model 3's diagnostic plots showed improved residual behavior. The Q-Q plot followed the expected distribution more closely, with minor deviations at the upper end, indicating slight **right skewness**. The Residuals vs Fitted plot remained centered around zero with less dispersion. The Scale-Location plot revealed a small increase in variance for higher predicted values, suggesting mild heteroscedasticity. The Cook's Distance plot no longer showed any extreme outliers, with all values under 1.

The 3rd and final is the best-performing model,

$\text{Price}_i = 55{,}401.0 - 1{,}411.8 \times \text{Age}_i - 17{,}952.2 \times I\{\text{Type}_i = \text{Sedan}\} - 5{,}374.9 \times I\{\text{Type}_i = \text{SUV}\} - 17{,}065.7 \times I\{\text{Fuel}_i = \text{Gas}\} - 14{,}192.4 \times I\{\text{Fuel}_i = \text{Other}\} + 368.4 \times \text{Age}_i \times I\{\text{Type}_i = \text{Sedan}\} - 188.5 \times \text{Age}_i \times I\{\text{Type}_i = \text{SUV}\}$

Here, $I(\,)$ represents the indicator function, with "Pickup" and "Diesel" serving as the reference categories for type and fuel, respectively.

For a diesel pickup, which is the baseline, the price declines by $1,411.8 for each additional year of age. Sedans depreciate faster, losing an extra $368.4 per year compared to pickups, while SUVs lose an additional $188.5 per year. The estimated starting price for a new diesel pickup is $55,401.0. The initial price is also adjusted based on type and fuel: sedans and SUVs are cheaper than pickups by $17,952.2 and $5,374.9 respectively, while gas-powered and other-fuel vehicles are priced lower by $17,065.7 and $14,192.4 compared to diesel models.

The model clearly demonstrates that vehicle price declines with age, with the rate of depreciation varying by vehicle type. However, it struggles with extreme values, particularly at the distribution tails, as shown in the residuals Q-Q plot. This is likely because very expensive or very old vehicles, such as classic or collectible cars, do not follow typical pricing trends and were not well represented in the dataset. The removal of the 85-year-old vehicle outlier significantly improved the model's performance for exemple. Even with Model 3, about 32.14% of price variation remains unexplained (R^2 = 67.86%). This suggests that other factors, such as mileage, could have a significant impact on vehicle pricing but are not accounted for in this model. Finally, removing the brand variable improved the model's efficiency. This makes sense, as car manufacturers often produce a wide range of models at various price points, making brand a less reliable predictor of price.

Task 2

Question A:

Figure 24: Bar Plot of proportion of satisfied travelers by class

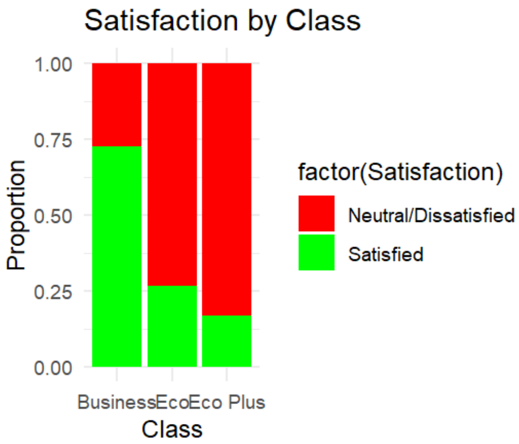


Figure 24 highlights differences in passenger satisfaction by travel class. Business class passengers report the highest satisfaction levels, with nearly 73% expressing positive feedback. In comparison, only 27% of Economy passengers and just 17% of those in Economy Plus report being satisfied. This lower satisfaction in Economy Plus is unexpected, as it typically offers enhanced comfort over Economy.

Figure 25: Boxplot of the satisaftion status by distance traveled

Figure 25 illustrates the relationship between travel distance and passenger satisfaction. Both satisfied and dissatisfied groups display right-skewed distributions, though the skew is minimal among dissatisfied passengers. The boxplot reveals that no dissatisfied individuals took flights exceeding 2000 km, suggesting that longer journeys are associated with higher satisfaction. Additionally, the range of travel distances is much narrower for dissatisfied passengers than for those who are satisfied. Despite this, the median travel distances remain relatively close approximately 1000 km for satisfied passengers and 600 km for those dissatisfied.

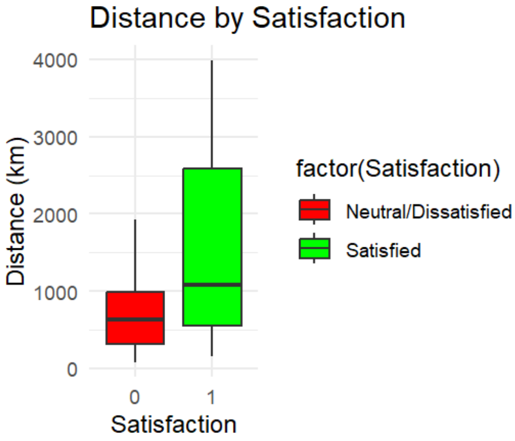
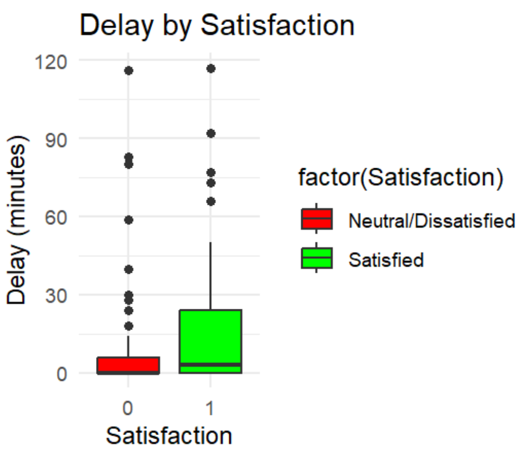


Figure 26: Boxplot of the travel delay by satisfaction status



This boxplot shows the relationship between flight delay (in minutes) and passenger satisfaction. Satisfied passengers tend to experience more variable delays. Their delays range widely, the median delay for this group is also higher at around 25 minutes. Neutral or dissatisfied passengers have shorter and more consistent delays, with most delays under 15 minutes. Their median delay is noticeably lower—approximately 5 minutes. Despite these differences, it's important to note that some satisfied passengers experienced significant delays (with several outliers above 90 minutes), while most dissatisfied passengers had minimal delays. This result is somewhat counterintuitive—it suggests that satisfied passengers may tolerate longer delays, or that their satisfaction is influenced more by other factors (like service quality or comfort) rather than just punctuality. On the other hand, dissatisfied passengers report shorter delays, which might indicate that even minor delays can contribute to a negative overall experience, or again, that delay isn't the main factor driving their dissatisfaction.

Question B:

Gender and Satisfaction:

Figure 27: Contingency table of satisfaction status for the flight and gender and travel status:

Satisfaction status	Gender of the passenger		Row total
	Female	Male	
Not satisfied (0)	30	20	50
Satisfied (1)	28	17	45
Column total	58	37	95

We use the χ^2 test of independence to test:

- H_0: Satisfaction status and Gender are independent
- H_1: Satisfaction status and Gender are not independent

To calculate the expected value under the null hypothesis H_0 for a chi-squared test of independence, we use the following formula for each cell in the contingency table:

$$E_{i\,j} = \frac{(Row\ total_i) \times (Column\ total_j)}{Grand\ total}$$

Where:

- $E_{i\,j}$ is the expected count for the cell in row i and column j
- $(Row\ total_i)$ is the total number of observations in row i
- $(Column\ total_j)$ is the total number of observations in column j
- $Grand\ total$ is the total number of observations in the entire table.

Figure 28: Table of the expected frequencies under H_0:

Satisfaction status	Gender of the passenger	
	Female	Male
Not satisfied (0)	30.53	19.47
Satisfied (1)	27.47	17.53

$$\chi 2 = \sum \frac{(O_{i\,j} - E_{i\,j})^2}{E_{i\,j}} \qquad \text{df = (r−1) × (c−1)}$$

Where $O_{i\,j}$ are the observed counts and $E_{i\,j}$ are the expected counts. We then compare the calculated chi-squared statistic to the critical value from the chi-squared distribution with the appropriate degrees of freedom (df=1) to determine if the null hypothesis can be rejected. The degrees of freedom (df) for a Chi-squared test of independence are determined by the formula: with r the number of rows in the contingency table and c the number of columns.

Since the expected frequency in each cell is greater than 5, the Chi-squared test statistic χ^2 = 0.00012297 with 1 degree of freedom, and the associated p-value is $P(\chi^2_1 > 0.00012297) = 0.9912$ is much greater than 0.05, we fail to reject the null hypothesis (H_0: Satisfaction status and Gender are independent). This indicates that there is no significant evidence to suggest that gender and satisfaction are dependent. In other words, gender does not have a significant effect on the satisfaction status of passengers for this flight. Therefore, we conclude that satisfaction status and gender are independent in this sample.

Travel Class and Satisfaction:

Figure 29: Contingency table of satisfaction status for the flight and the travel class of the passenger:

Satisfaction status	Travel class of the passenger			Row total
	Business	Eco	Eco Plus	
Not satisfied (0)	12	33	5	50
Satisfied (1)	32	12	1	45
Column total	44	45	6	95

We use the χ^2 test of independence to test:

- H_0: Satisfaction status and Travel class are independent
- H_1: Satisfaction status and Travel class are not independent

Figure 30: Table of the expected frequencies under H_0:

Satisfaction status	Travel class of the passenger		
	Business	Eco	Eco Plus
Not satisfied (0)	23.16	23.68	3.16
Satisfied (1)	20.84	21.32	2.84

Since the expected frequency in each cell is greater than 5, the Chi-squared test statistic χ^2 = 21.354 with 2 degrees of freedom (2 satisfaction statuses and 3 travel classes), and the associated p-value is $P(\chi^2_2 > 21.354)$ = 2.307e-05, which is much smaller than 0.05. Therefore, we reject the null hypothesis. This suggests that there is a significant association between travel class and satisfaction. In other words, travel class has a significant effect on the satisfaction status of passengers for this flight. Therefore, we conclude that satisfaction status and travel class are dependent in this case.

Justification for the Tests

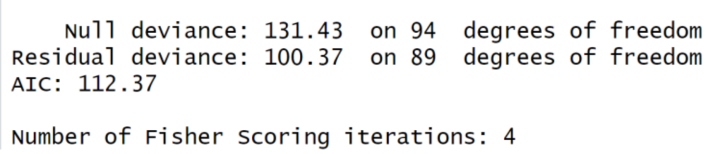
The Chi-squared test of independence is appropriate here because we are testing the relationship between two categorical variables (Satisfaction and Gender or Class). The assumptions of the Chi-squared test are: the data are categorical, the expected frequencies in each cell should be at least 5, which is met in both tests and the observations are independent. Both tests were performed using the appropriate contingency tables and passed the assumptions.

Question C:

To assess the factors influencing passenger satisfaction in the airline dataset, a generalized linear model was employed. The binary response variable was Satisfaction, indicating whether a passenger was satisfied or not. The initial ("full") model included four explanatory variables: Gender, Class, Distance, and Delay. The model was fitted using the glm() function in R with a binomial family and a logit link:

Satisfaction ~ Gender + Class + Distance + Delay, data = airline_data

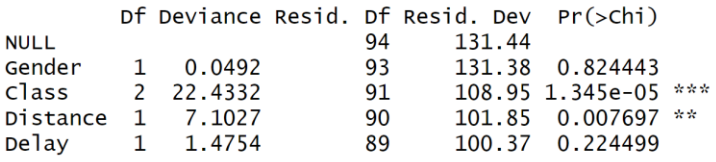
Figure 31: Model summary



```
    Null deviance: 131.43  on 94  degrees of freedom
Residual deviance: 100.37  on 89  degrees of freedom
AIC: 112.37

Number of Fisher Scoring iterations: 4
```

The first model's output indicated a residual deviance of 100.37 (on 89 degrees of freedom), which represents a substantial improvement from the null deviance of 131.43 (on 94 degrees of freedom). The associated AIC (Akaike Information Criterion) value is 112.37, a useful metric for comparing models. A lower AIC value generally indicates a better-fitting model, as it penalizes models with more parameters to prevent overfitting. The Fisher scoring iterations are 4. It is an iterative process used in maximum likelihood estimation (MLE) to estimate the model parameters. This value is considered good, as a small number of iterations indicates that the model converges quickly.

Figure 32: deviance and Chi-squared tests

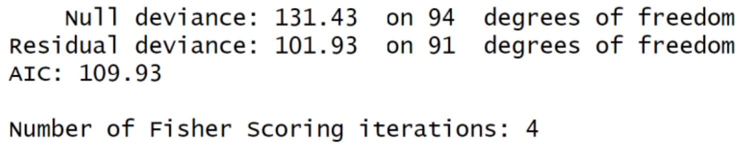


```
         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                        94     131.44              
Gender    1   0.0492        93     131.38  0.824443    
Class     2  22.4332        91     108.95 1.345e-05 ***
Distance  1   7.1027        90     101.85  0.007697 ** 
Delay     1   1.4754        89     100.37  0.224499    
```

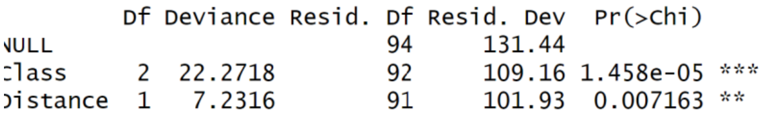
The table shows the results of a model comparison based on deviance and Chi-squared tests. The NULL model has a deviance of 131.44, and the addition of "Gender" does not significantly improve the model, as its p-value is 0.8244. However, both "Class" (p-value < 0.0001) and "Distance" (p-value = 0.0077) significantly improve the model, indicating that these predictors are important. On the other hand, "Delay" does not have a significant effect, with a p-value of 0.2245. Overall, the results suggest that "Class" and "Distance" are key predictors, while "Gender" and "Delay" are not.

To enhance interpretability and simplify the model, only statistically significant variables were retained. The reduced model was fitted with: ***Satisfaction ~ Class + Distance, data = airline_data***

Figure 32: New model summary



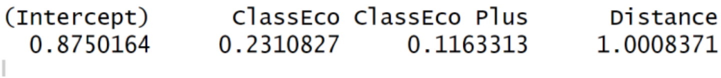
```
    Null deviance: 131.43  on 94  degrees of freedom
Residual deviance: 101.93  on 91  degrees of freedom
AIC: 109.93

Number of Fisher Scoring iterations: 4
```



```
         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                        94     131.44              
Class     2  22.2718        92     109.16 1.458e-05 ***
Distance  1   7.2316        91     101.93  0.007163 ** 
```

Figure 33: New model deviance and Chi-squared tests

This led to a minor increase in residual deviance but removed unnecessary complexity. A likelihood ratio test (anova(reduced_model, model, test = "Chisq")) confirmed that the simpler model was not significantly worse, validating the exclusion of non-significant terms. In Model 2, "Class" (p = 1.46e-05) and "Distance" (p = 0.0072) remained highly significant, indicating that these predictors are crucial for the model.



```
(Intercept)      ClassEco ClassEco Plus      Distance
  0.8750164     0.2310827     0.1163313     1.0008371
```

Figure 34: New model exponential coefficients

To aid interpretability, exponentiated coefficients were computed using exp(coef(reduced_model)), providing odds ratios. For instance, if you are in the "ClassEco" category, you are approximately 0.23 times as likely to experience the outcome compared to the reference category, meaning you're less likely to experience the event. Similarly, being in the "ClassEco Plus" category makes you 0.12 times as likely to experience the outcome, again indicating a reduced likelihood. For "Distance," the odds are nearly unchanged (1.001), meaning that for each unit increase in distance, the likelihood of the outcome slightly increases.

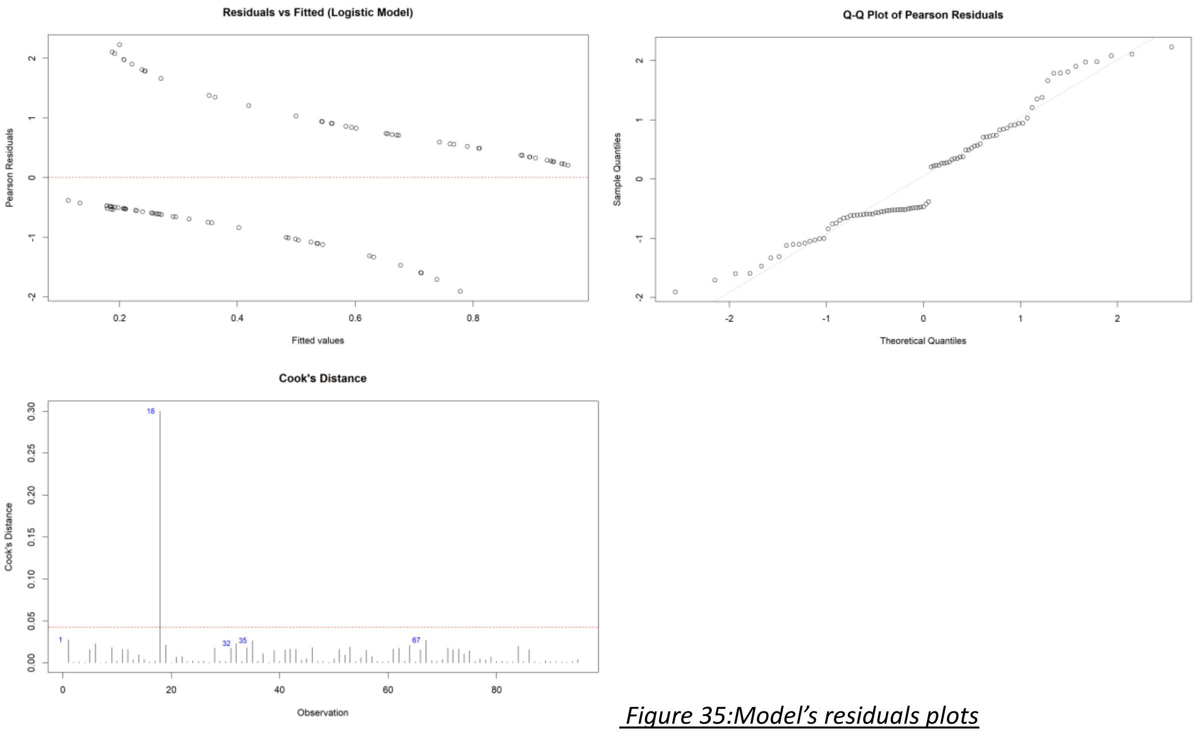
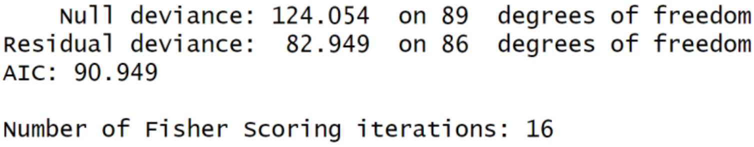


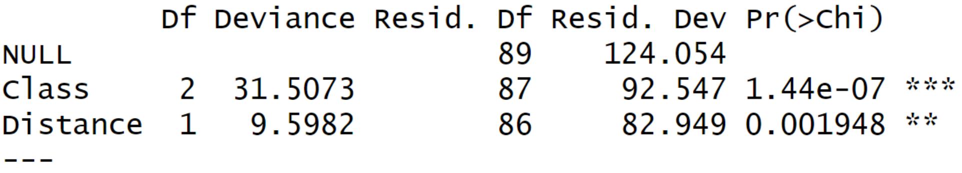
Figure 35:Model's residuals plots

The diagnostic plots for the model provide several insights into its performance and potential issues. The Cook's Distance plot highlights that observation 18 is highly influential, suggesting it has a substantial impact on the model's coefficients and should be examined more closely. The Residuals vs Fitted plot indicates that residuals are generally centered around zero without distinct patterns, which is a positive sign, although some outliers are present. The Q-Q plot of Pearson Residuals shows that the residuals are approximately normally distributed, with some deviations at the tails, indicating slight non-normality at the extremes.



```
    Null deviance: 124.054  on 89  degrees of freedom
Residual deviance:  82.949  on 86  degrees of freedom
AIC: 90.949

Number of Fisher Scoring iterations: 16
```

Figure 36:3rd Model's summary

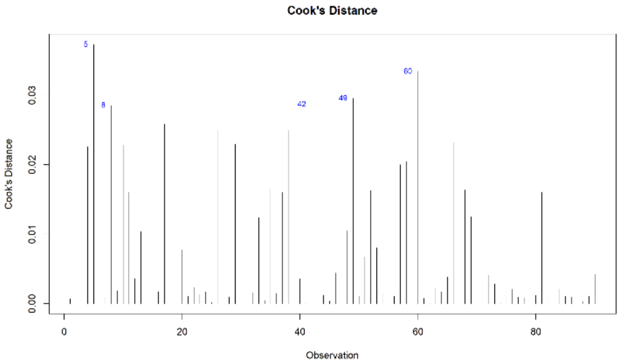


```
         Df Deviance Resid. Df Resid. Dev Pr(>Chi)    
NULL                        89    124.054             
Class     2  31.5073        87     92.547 1.44e-07 ***
Distance  1   9.5982        86     82.949 0.001948 ** 
---
```

Figure 37:3rd Model's deviance and Chi-squared tests

The comparison between the 2nd model and the 3rd one after removing influential observations reveals several insights, particularly concerning the Number of Fisher Scoring iterations, which increased from 4 to 16. Initially, the model 2, demonstrated significant predictors in Class and Distance, with p-values of 0.00431 and 0.01683, respectively. The model had a residual deviance of 101.93 on 91 degrees of freedom and an AIC of 109.93, converging quickly in just 4 iterations. After cleaning the data by removing influential observations, the new model, showed enhanced performance with a residual deviance of 82.949 on 86 degrees of freedom and a lower AIC of 90.949. The significance of Class and Distance remained robust, with p-values of 1.44e-07 and 0.001948, respectively. However, the model required 16 Fisher Scoring iterations to converge, indicating a more complex fitting process due to the data adjustments. This increase in iterations suggests that while the cleaned data improved overall model fit and reduced complexity, it also necessitated more computational effort to achieve convergence. Overall, the optimized model's improved performance metrics underscore the benefits of data cleaning in enhancing model reliability and fit, despite the increased computational demands reflected in the higher number of iterations.

Figure 38: 3rd Model's cook's distance



The Cook's Distance plot shown here represents a significant improvement over the previous model that included outliers. In this updated plot, while there are still some observations with higher Cook's Distance values, such as observations 5, 8, 39, 42, 49 and 60, the overall influence of individual data points appears to be more balanced compared to the previous model. This suggests that the removal of the most influential outliers has led to a more robust and stable model.

Although some observations still exhibit significant Cook's Distance values, these points are important as they may represent valid variations in the data that should not be disregarded. Their presence indicates that while the model has been refined, it remains sensitive to certain data points, which is crucial for maintaining the integrity and accuracy of the analysis.

Question D:

Figure 39: Predicted Satisfaction table for the 10 first observations

Observation	Predicted Probability	Actual Satisfaction	Correct Prediction? (0.5 threshold)
1	0.188	1	NO
2	0.906	1	YES
3	0.192	0	YES
4	0.937	1	YES
5	0.352	1	NO
6	0.208	1	NO
7	0.961	1	YES
8	0.881	1	YES
9	0.244	1	NO
10	0.790	1	YES

The predicted probabilities are scattered across a wide range of values (from 0.2 to 0.9), which corresponds to varying degrees of confidence in predicting passenger satisfaction. When comparing the predicted probabilities with the actual satisfaction values, the model appears to correctly predict satisfaction for some of the passengers but not for others. This is reflected in the fact that the model has an accuracy of 60% (i.e. 6 out of 10 predictions are correct).

By experimenting with different thresholds, we observe that a threshold of 0.55 results in an accuracy of approximately 76% (on the full dataset), a notable improvement compared to the 71% accuracy at a threshold of 0.5. This indicates that adjusting the decision threshold could enhance model performance.

However, this improvement comes at a cost, as modifying the threshold can influence other aspects of the model's performance, particularly precision and recall. Precision, which refers to the proportion of correctly predicted positive outcomes (satisfied passengers) among all the predicted positive outcomes, can be impacted when the threshold is increased. As the model becomes more conservative by classifying fewer passengers as satisfied, this may lead to a reduction in false positives but also decrease the number of true positives. On the other hand, recall, which is the proportion of correctly predicted positive outcomes among all actual positive outcomes, may suffer when the threshold is raised. The model might miss more true positives, thus lowering recall, but this adjustment could simultaneously increase precision by reducing false positives.

In addition to generalized linear models (GLM), other machine learning models could be considered for improving the prediction of passenger satisfaction. One particularly interesting alternative is Random Forests. Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting.

Appendices

```
getwd() # Print the current working directory


###
#1#
###
# a)
# Read the data from the csv file into R
vehicle_data <- read.csv("vehicle.csv", header = TRUE)
# Numerical summaries
head(vehicle_data) # Display the first few rows of the dataset
str(vehicle_data)  # Check structure of the dataset
summary(vehicle_data) # Summary statistics
nrow(vehicle_data)  # Number of rows = total number of vehicles
# Categorical summaries
#table(vehicle_data$brand) # Count of vehicles by brand
# Frequency table with counts and percentages for vehicle brands
brand_counts <- table(vehicle_data$brand)
brand_percent <- prop.table(brand_counts) * 100
data.frame(Brand = names(brand_counts), Count = as.vector(brand_counts), Percentage = round(as.vector(brand_percent), 2))
#table(vehicle_data$fueltype) # Count of vehicles by fuel type
# Frequency table with counts and percentages
fueltype_counts <- table(vehicle_data$fueltype)
fueltype_percent <- prop.table(fueltype_counts) * 100
data.frame(Fueltype = names(fueltype_counts), Count = as.vector(fueltype_counts), Percentage = round(as.vector(fueltype_percent), 2))
#table(vehicle_data$type) # Count of vehicles by type
# Frequency table with counts and percentages for vehicle types
```

```
type_counts <- table(vehicle_data$type)
type_percent <- prop.table(type_counts) * 100
data.frame(Type = names(type_counts), Count = as.vector(type_counts), Percentage = round(as.vector(type_percent), 2))
# Summary for numeric variables
summary(vehicle_data$age)     # Summary statistics for age
summary(vehicle_data$price)   # Summary statistics for price
# Standard Deviation
sd(vehicle_data$age, na.rm = TRUE)    # Standard deviation for age
sd(vehicle_data$price, na.rm = TRUE)  # Standard deviation for price
# Interquartile Range (IQR)
IQR(vehicle_data$age, na.rm = TRUE)   # IQR for age
IQR(vehicle_data$price, na.rm = TRUE) # IQR for price
# Mode function since R doesn't have a built-in one for numerical data
get_mode <- function(v) {
  uniq_vals <- unique(v)
  uniq_vals[which.max(tabulate(match(v, uniq_vals)))]}
get_mode(vehicle_data$age)    # Mode for age
get_mode(vehicle_data$price)  # Mode for price


# Graphical summaries
library(ggplot2)
# Histogram of vehicle prices
ggplot(vehicle_data, aes(x = price)) +  geom_histogram(binwidth = 2000, fill = "blue", alpha = 0.7) +
  labs(title = "Distribution of Vehicle Prices", x = "Price (USD)", y = "Count") + theme_minimal()
# Boxplot of vehicle number by fuel type
ggplot(vehicle_data, aes(x = brand, fill = brand)) + geom_bar() +
 labs(title = "Count of Vehicles by Brand", x = "Brand", y = "Count") + theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Distribution of vehicle age

ggplot(vehicle_data, aes(x = age)) + geom_histogram(binwidth = 2, fill = "blue", alpha = 0.7, color = "black") +
 labs(title = "Distribution of Vehicle Age", x = "Age (Years)", y = "Count") + theme_minimal()

# Boxplot of vehicle number by fuel type

ggplot(vehicle_data, aes(x = fueltype, fill = fueltype)) + geom_bar() +
 labs(title = "Count of Vehicles by Fuel Type", x = "Fuel Type", y = "Count") + theme_minimal()

# Boxplot of vehicle number by type

ggplot(vehicle_data, aes(x = type, fill = type)) + geom_bar() +
 labs(title = "Count of Vehicles by Type", x = "Type", y = "Count") + theme_minimal()




# b)

# Boxplot: Price distribution by brand

ggplot(vehicle_data, aes(x = brand, y = price, fill = brand)) +
 geom_boxplot() + labs(title = "Price Distribution by Brand", x = "Brand", y = "Price (USD)") +
 theme_minimal() + theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Scatter plot: Age vs Price

ggplot(vehicle_data, aes(x = age, y = price, color = brand)) +
 geom_point(alpha = 0.6) + geom_smooth(method = "lm", se = FALSE, color = "black") +
 labs(title = "Vehicle Age vs Price", x = "Vehicle Age (Years)", y = "Price (USD)") + theme_minimal()

# Boxplot: Price distribution by fuel type

ggplot(vehicle_data, aes(x = fueltype, y = price, fill = fueltype)) + geom_boxplot() +
  labs(title = "Price Distribution by Fuel Type", x = "Fuel Type", y = "Price (USD)") + theme_minimal()

# Boxplot: Price distribution by vehicle type

ggplot(vehicle_data, aes(x = type, y = price, fill = type)) + geom_boxplot() +

  labs(title = "Price Distribution by Vehicle Type", x = "Vehicle Type", y = "Price (USD)") + theme_minimal()

# Scatter plot of price vs age vs fuel type

ggplot(vehicle_data, aes(x = age, y = price, color = fueltype)) + geom_point(alpha = 0.6) +

  geom_smooth(method = "lm", se = FALSE, color = "black") + labs(title = "Vehicle Age vs Price by Fuel Type", x = "Vehicle Age (Years)", y = "Price (USD)") +

  theme_minimal() + xlim(0, 35)

# Scatter plot of price vs age vs type

ggplot(vehicle_data, aes(x = age, y = price, color = type)) + geom_point(alpha = 0.6) +

  geom_smooth(method = "lm", se = FALSE, color = "black") + labs(title = "Vehicle Age vs Price by Vehicle Type", x = "Vehicle Age (Years)", y = "Price (USD)") +

  theme_minimal() + xlim(0, 35)


# c)

# Compute Pearson correlation between price and age

correlation <- cor(vehicle_data$price, vehicle_data$age, method = "pearson")

# Sample size (n)

n <- nrow(vehicle_data)

# Fisher’s Z-transformation

z <- 0.5 * log((1 + correlation) / (1 - correlation))

# Standard error of Z

se_z <- 1 / sqrt(n - 3)

# 99% Confidence Interval for Z

z_critical <- qnorm(0.995)  # For 99% CI, use 0.995 (two-tailed)

z_lower <- z - z_critical * se_z
```

```
z_upper <- z + z_critical * se_z

# Convert back to correlation scale

r_lower <- (exp(2 * z_lower) - 1) / (exp(2 * z_lower) + 1)

r_upper <- (exp(2 * z_upper) - 1) / (exp(2 * z_upper) + 1)

# Display results

cat("Pearson Correlation between Price and Age:", round(correlation, 3), "\n")

cat("99% Confidence Interval for Correlation: [", round(r_lower, 3), ",", round(r_upper, 3),
"]\n")


# confidence interval for Pearson’s correlation using bootstrap resampling.

# Load necessary libraries

library(boot)

# Define function to compute Pearson correlation

correlation_fn <- function(data, indices) {

  sample_data <- data[indices, ]  # Resample data

  return(cor(sample_data$price, sample_data$age, method = "pearson"))}

# Perform bootstrap with 10,000 resamples

set.seed(123)  # For reproducibility

bootstrap_results <- boot(data = vehicle_data, statistic = correlation_fn, R = 10000)

# Compute 99% Confidence Interval (Percentile Method)

ci_boot <- boot.ci(bootstrap_results, type = "perc", conf = 0.99)

# Display results

cat("Bootstrap 99% Confidence Interval for Pearson Correlation:
[",round(ci_boot$percent[4], 3), ",", round(ci_boot$percent[5], 3), "]\n")




# d)

# Categorical values
```

```
vehicle_data$type <- as.factor(vehicle_data$type)
vehicle_data$brand <- as.factor(vehicle_data$brand)
vehicle_data$fueltype <- as.factor(vehicle_data$fueltype)
# First model
model1 = lm(price ~ age * type + fueltype + brand, data = vehicle_data)
summary(model1)
anova(model1)

#second model
model2 = lm(price ~ age  * type + fueltype, data = vehicle_data)
summary(model2)
anova(model2)
#plotting
std_res <- rstandard(model2)
# 5 largest residuals
top5_residuals <- order(abs(std_res), decreasing = TRUE)[1:5]
# Residuals vs Fitted
plot(model2$fitted.values, model2$residuals,
   xlab = "Fitted values", ylab = "Residuals",
   main = "Residuals vs Fitted", pch = 1)
abline(h = 0, col = "red", lty = 2)
# Adding the number of the largest point
text(model2$fitted.values[top5_residuals],
   model2$residuals[top5_residuals],
   labels = top_resid,
   cex = 0.6,
   pos = 3, col = "blue")
# Q-Q Residuals
```

```
qqnorm(std_res, main = "Q-Q Residuals", pch = 1)
qqline(std_res, col = "gray", lty = 2)
# Adding the number of the largest point
text(x = qqnorm(std_res, plot.it = FALSE)$x[top5_residuals],
   y = std_res[top_resid],
   labels = top_resid,
   cex = 0.6,
   pos = 3, col = "blue")
# Scale location
plot(model2, which = 3)
# Cook's distances
cooks_distance <- cooks.distance(model2)
# 3 most influential points
top3_cooks <- order(cooks_distance, decreasing = TRUE)[1:5]
# Plot Cook's Distance
plot(cooks_distance, type = "h", main = "Cook's Distance", ylab = "Cook's distance", xlab = "Obs number")
text(x = top3_cooks, y = cooks_distance[top3_cooks], labels = top3_cooks, pos = 2, col = "blue", cex = 0.8)
abline(h = 4 / nrow(vehicle_data), col = "red", lty = 2)


# Remove the influential point
print(vehicle_data[39, ])
vehicle_clean <- vehicle_data[c(-39), ]
nrow(vehicle_clean)
# Rebuild the model
model2_clean <- lm(price ~ age * type + fueltype, data = vehicle_clean)
# Summary and anova
summary(model2_clean)
```

```
anova(model2_clean)

# Plotting

std_res <- rstandard(model2_clean)

top5_residuals <- order(abs(std_res), decreasing = TRUE)[1:5]

plot(model2_clean$fitted.values, model2_clean$residuals,

   xlab = "Fitted values", ylab = "Residuals",

   main = "Residuals vs Fitted", pch = 1)

abline(h = 0, col = "red", lty = 2)

text(model2_clean$fitted.values[top5_residuals],

   model2_clean$residuals[top5_residuals],

   labels = top_resid, cex = 0.7, pos = 3, col = "blue")

qqnorm(std_res, main = "Q-Q Residuals", pch = 1)

qqline(std_res, col = "gray", lty = 2)

text(x = qqnorm(std_res, plot.it = FALSE)$x[top5_residuals],

   y = std_res[top5_residuals],

   labels = top_resid,

   cex = 0.7, pos = 3, col = "blue")

plot(model2_clean, which = 3)

cooks_distance <- cooks.distance(model2_clean)

top5_cooks <- order(cooks_distance, decreasing = TRUE)[1:5]

plot(cooks_distance, type = "h",

   main = "Cook's Distance",

   ylab = "Cook's distance",

   col = "red", lwd = 2,

   ylim = c(0, 0.25),

   xlim = c(0, nrow(vehicle_clean)))

abline(h = 4 / nrow(vehicle_clean), col = "blue", lty = 2)

text(x = top5_cooks, y = cooks_distance[top5_cooks],
    labels = top5_cooks, pos = 2, col = "blue", cex = 0.8)


###
#2#
###
# a)
# Read the data from the csv file into R
airline_data <- read.csv("Airline.csv", header = TRUE, sep=",")

# Numerical summaries
head (airline_data) # Display the first few rows of the dataset
str(airline_data) # Check structure of the dataset
summary(airline_data) # Summary statistics
nrow(airline_data) # Number of rows in the dataset = total number of passengers
table(airline_data$Gender) # Count the number of male and female passengers
table(airline_data$Class) # Count the number of passengers by class
table(airline_data$Satisfaction) # Count the number of satisfied and unsatisfied passengers
summary(airline_data$Distance) # Summary statistics for distance
summary(airline_data$Delay) # Summary statistics for delay
summary(airline_data$Age) # Summary statistics for age

# Graphical summaries
library(ggplot2)
# Bar plot of satisfaction by class
ggplot(airline_data, aes(x = Class, fill = factor(Satisfaction))) +
  geom_bar(position = "fill") + labs(title = "Satisfaction by Class", x = "Class", y = "Proportion") +
  scale_fill_manual(values = c("red", "green"), labels = c("Neutral/Dissatisfied", "Satisfied")) + theme_minimal()
# Boxplot of satisfaction by distance
ggplot(airline_data, aes(x = factor(Satisfaction), y = Distance, fill = factor(Satisfaction))) +
  geom_boxplot() + labs(title = "Distance by Satisfaction", x = "Satisfaction", y = "Distance (km)") +
  scale_fill_manual(values = c("red", "green"), labels = c("Neutral/Dissatisfied", "Satisfied")) + theme_minimal()
# Scatter plot of distance vs delay colored by satisfaction
ggplot(airline_data, aes(x = Distance, y = Delay, color = factor(Satisfaction))) +
  geom_point(alpha = 0.6) + labs(title = "Distance vs Delay by Satisfaction", x = "Distance (km)", y = "Delay (minutes)") +
  scale_color_manual(values = c("red", "green"), labels = c("Neutral/Dissatisfied", "Satisfied")) + theme_minimal()
# Box plot of delay by satisfaction
ggplot(airline_data, aes(x = factor(Satisfaction), y = Delay, fill = factor(Satisfaction))) +
  geom_boxplot() + labs(title = "Delay by Satisfaction", x = "Satisfaction", y = "Delay (minutes)") +
  scale_fill_manual(values = c("red", "green"), labels = c("Neutral/Dissatisfied", "Satisfied")) + theme_minimal()

# b)
gender_table <- table(airline_data$Satisfaction, airline_data$Gender)
print(gender_table)
chi_gender <- chisq.test(gender_table)
chi_gender
chi_gender$expected # Expected counts
```

```
class_table <- table(airline_data$Satisfaction, airline_data$Class)
print(class_table)
chi_class <- chisq.test(class_table)
chi_class
chi_class$expected # Expected counts


# c)
# Fit the full model
model <- glm(Satisfaction ~ Gender + Class + Distance + Delay, data = airline_data, family = binomial)
summary(model)
anova(model, test = "Chisq")
# I choose to retain only the significant variables
reduced_model <- glm(Satisfaction ~ Class + Distance, data = airline_data, family = binomial)
summary(reduced_model)
anova(reduced_model, test = "Chisq")
# Compare the full and reduced models
anova(model, reduced_model, test = "Chisq")
# Odds ratios are great for interpreting effects (e.g., flying Business Class increases the odds of being satisfied by X times)
exp(coef(reduced_model))


# Standardized Pearson residuals
pearson_resid <- rstandard(reduced_model, type = "pearson")
fitted_vals <- fitted(reduced_model)
# Residuals vs Fitted
plot(fitted_vals, pearson_resid,
   xlab = "Fitted values", ylab = "Pearson Residuals",
    main = "Residuals vs Fitted (Logistic Model)", pch = 1)
abline(h = 0, col = "red", lty = 2)
# Q-Q Plot of residuals
qqnorm(pearson_resid, main = "Q-Q Plot of Pearson Residuals")
qqline(pearson_resid, col = "gray", lty = 2)
# Cook's Distance
cooks_d <- cooks.distance(reduced_model)
plot(cooks_d, type = "h", main = "Cook's Distance", ylab = "Cook's Distance", xlab = "Observation")
abline(h = 4 / nrow(airline_data), col = "red", lty = 2)
top_cooks <- order(cooks_d, decreasing = TRUE)[1:5]
text(x = top_cooks, y = cooks_d[top_cooks], labels = top_cooks, pos = 2, col = "blue", cex = 0.8)


# Remove the influential points
airline_data_clean <- airline_data[-top_cooks, ]
# Rebuild the model
opti_model <- glm(Satisfaction ~ Class + Distance, data = airline_data_clean, family = binomial)
summary(opti_model)
anova(opti_model, test = "Chisq")


# Standardized Pearson residuals
pearson_resid <- rstandard(opti_model, type = "pearson")
fitted_vals <- fitted(opti_model)
# Residuals vs Fitted
plot(fitted_vals, pearson_resid,
    xlab = "Fitted values", ylab = "Pearson Residuals",
    main = "Residuals vs Fitted (Logistic Model)", pch = 1)
```

```
abline(h = 0, col = "red", lty = 2)
# Q-Q Plot of residuals
qqnorm(pearson_resid, main = "Q-Q Plot of Pearson Residuals")
qqline(pearson_resid, col = "gray", lty = 2)
# Cook's Distance
cooks_d <- cooks.distance(opti_model)
plot(cooks_d, type = "h", main = "Cook's Distance", ylab = "Cook's Distance", xlab = "Observation")
abline(h = 4 / nrow(airline_data_clean), col = "red", lty = 2)
top_cooks <- order(cooks_d, decreasing = TRUE)[1:5]
text(x = top_cooks, y = cooks_d[top_cooks], labels = top_cooks, pos = 2, col = "blue", cex = 0.8)



# d)
# Prediction of probabilities using the reduced model
predicted_probs <- predict(reduced_model, type = "response")
# Show the first 10 predicted probabilities
pred <- data.frame(
  Predicted_Probability = round(predicted_probs[1:10], 3),
  Actual_Satisfaction = airline_data$Satisfaction[1:10])
print(pred)
# Show the percentage of satisfied passengers with a threshold of 0.5
pred$Predicted_Label <- ifelse(pred$Predicted_Probability >= 0.5, 1, 0)
pred$Correct <- pred$Predicted_Label == pred$Actual_Satisfaction
mean(pred$Correct)
# We can try to improve the model by modifying the threshold for classification
# It will change depending on the priority of the model (increasing sensitivity/recall or precision)
```

```
thresholds <- seq(0.4, 0.6, by = 0.05)

accuracies <- sapply(thresholds, function(t) {

  pred_label <- ifelse(predicted_probs >= t, 1, 0)

  mean(pred_label == airline_data$Satisfaction)})

data.frame(thresholds, accuracies) # The best one seems to be 0.55 with 0.758 accuracy
```