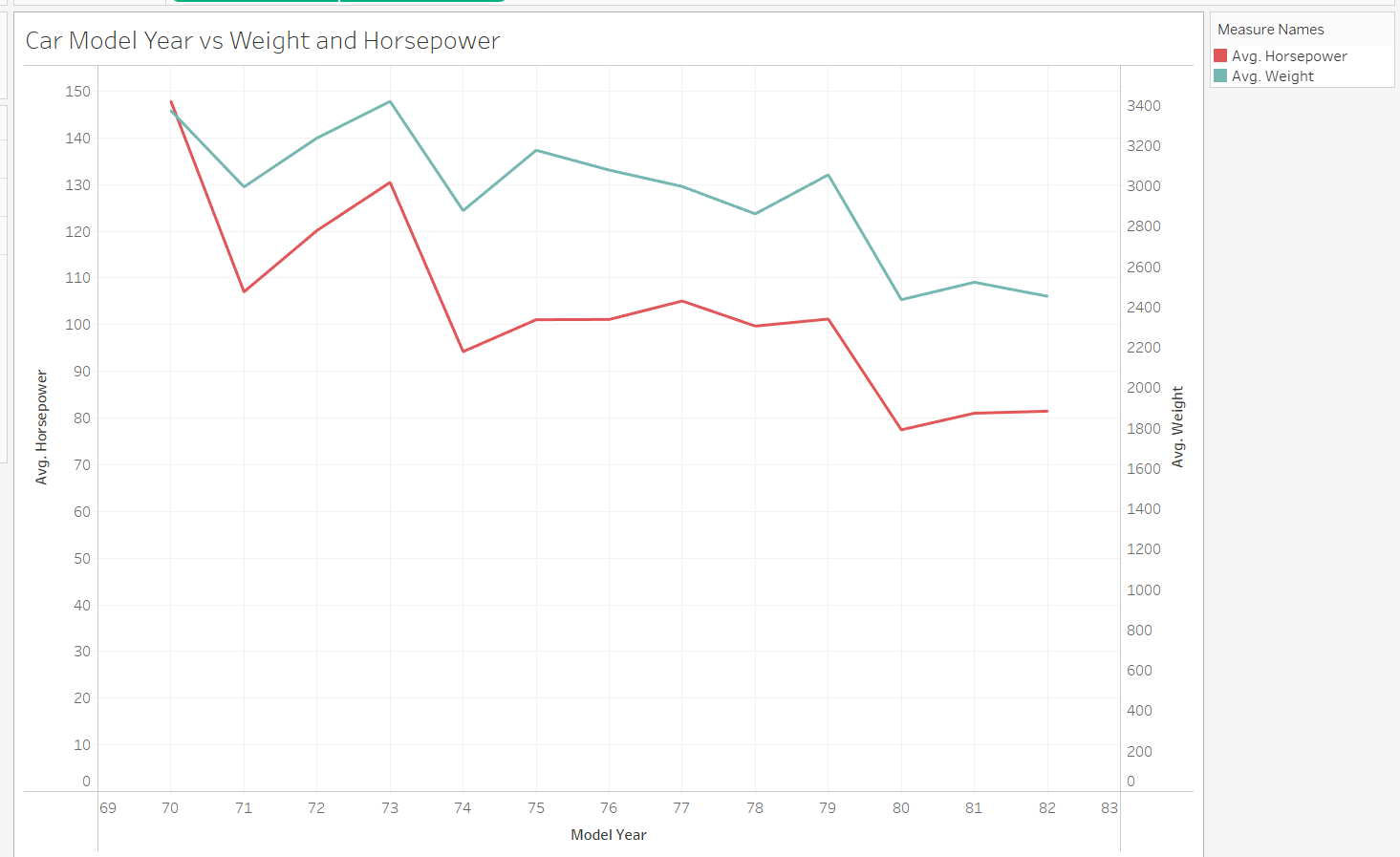
Rusi Rothschild

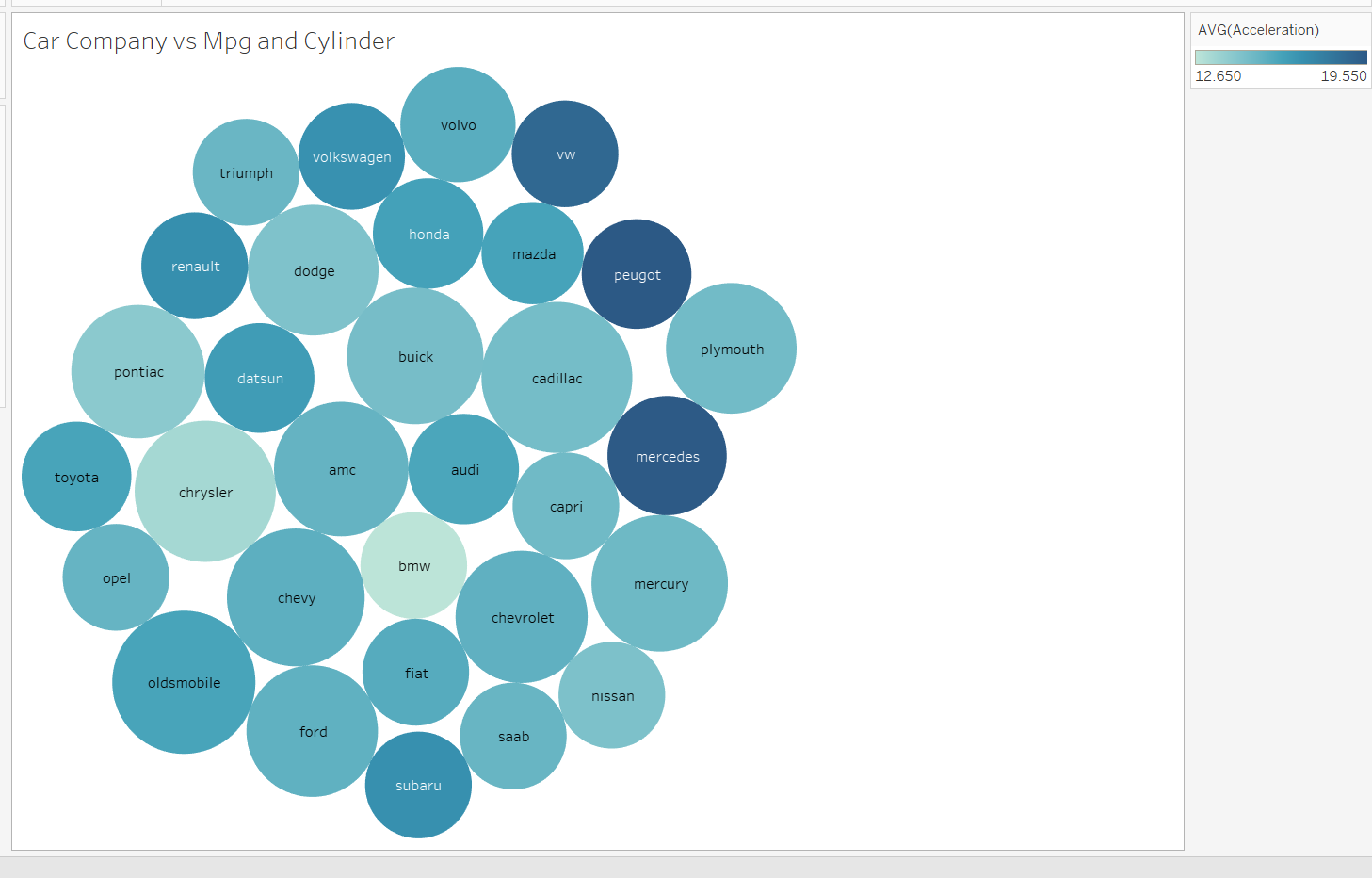
DS Final Project

**Part 1**



https://public.tableau.com/app/profile/ruth.rothschild/viz/Tableau-final\_project/Sheet1?publish=yes

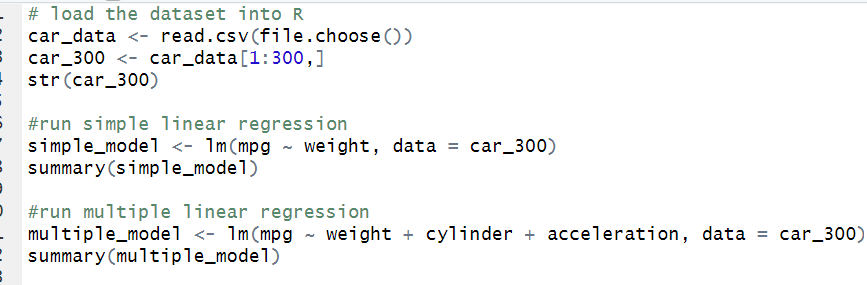
This graph compares the average Horsepower and the average Weight of cars across different model years. There seems to be a strong correlation between weight and horsepower. Heavier cars tend to have higher horsepower values. However, the relationship is not perfectly consistent across all years. Both the horsepower and the weight start off high in the 1970’s and then fluctuate over the next 10 years, with a slow decrease until the 1980’s.

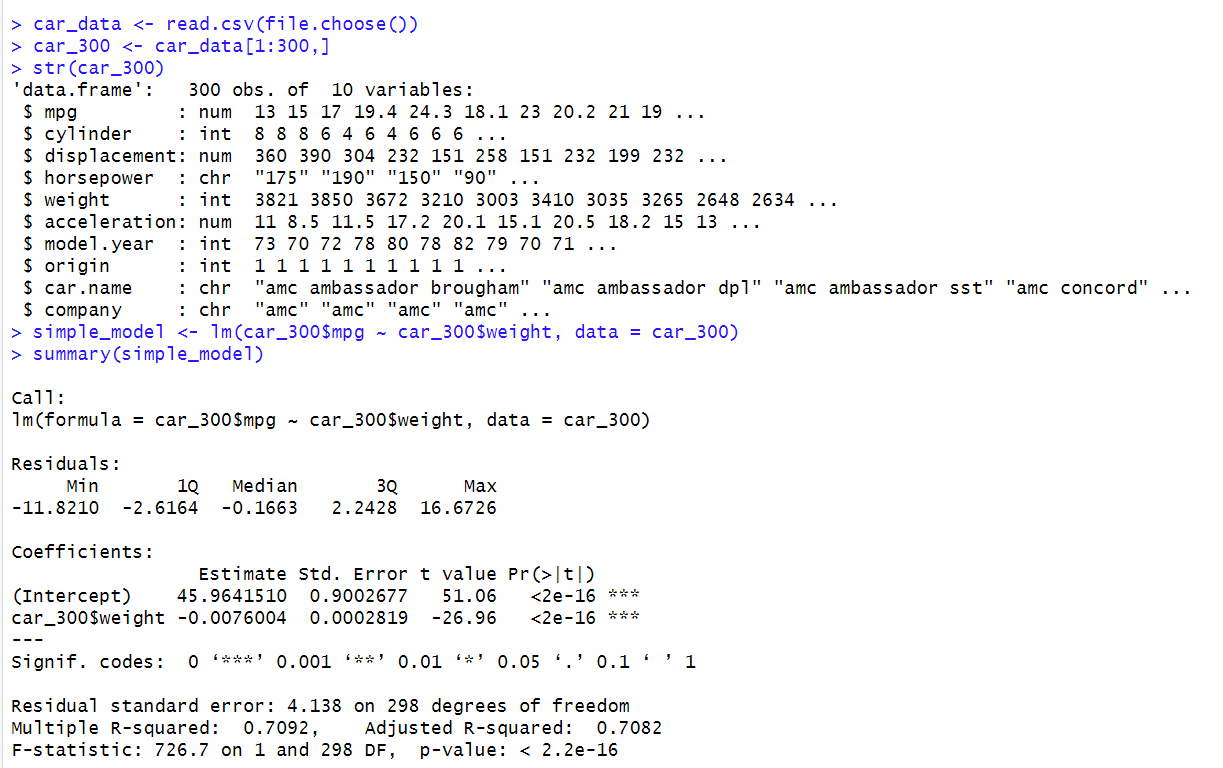


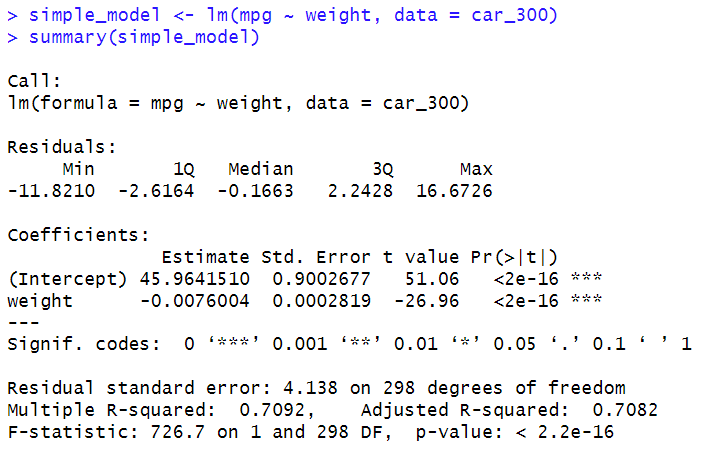
https://public.tableau.com/app/profile/ruth.rothschild/viz/Tableau-final\_project/Sheet2?publish=yes

To create this graph, I created a new attribute in the excel file called company. The new column contains the company of each car.

This bubble chart compares Car Companies by their average Mpg values and average cylinder count. The color of each bubble represents the average Mpg for cars from that company, the darker the color, the higher the mpg. The size of the bubble shows the cylinder count; the larger the bubble, the higher the cylinder count. There doesn’t seem to be a strong correlation between Mpg and cylinder. They do not rise and fall together.

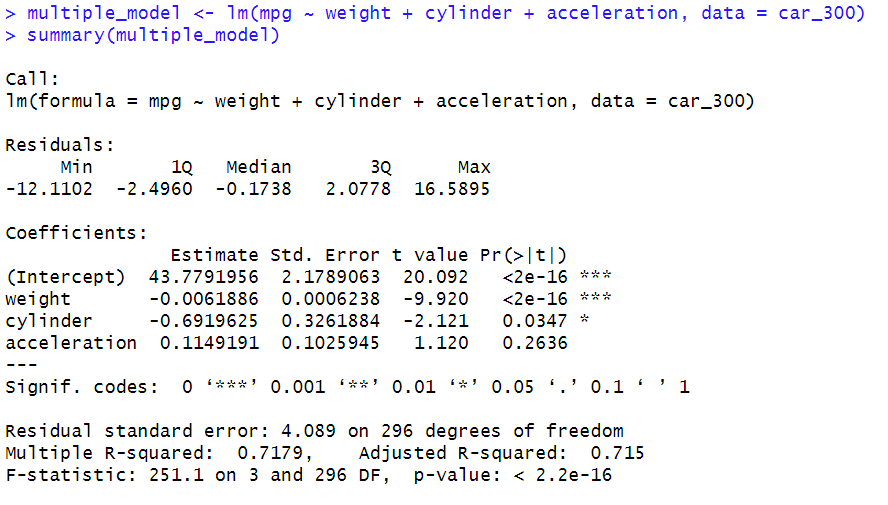






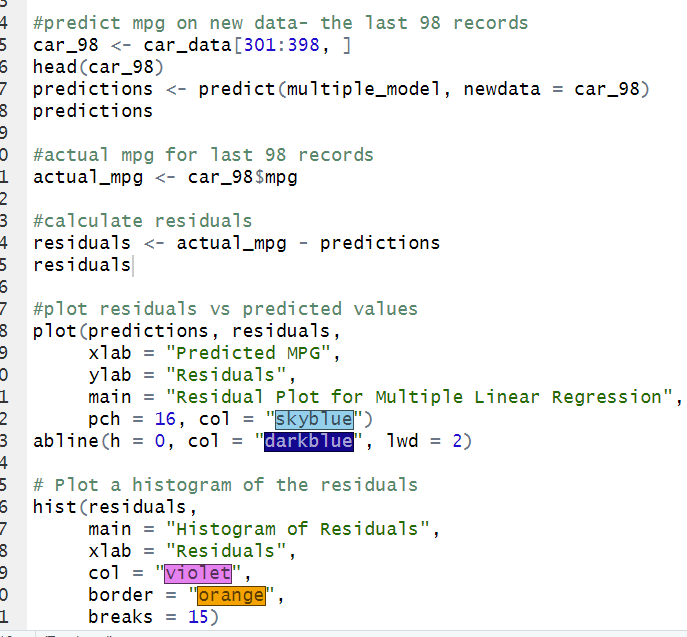
1. The multiple R squared is 0.7092.
2. The adjusted R squared is 0.7082.
3. The complete linear regression equation is:

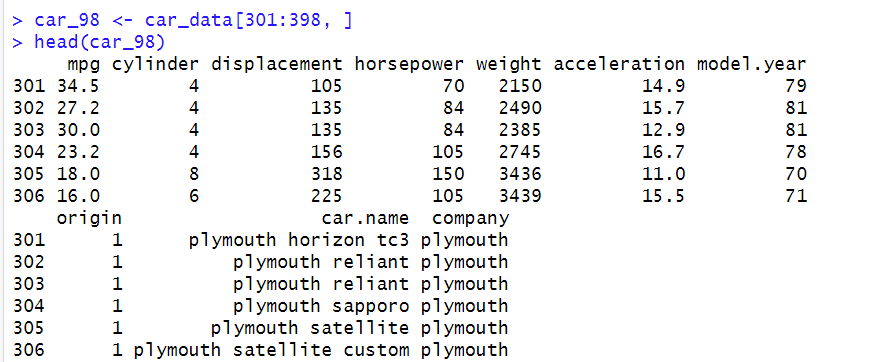
Mpg = 45.9641 – 0.0076 \* X



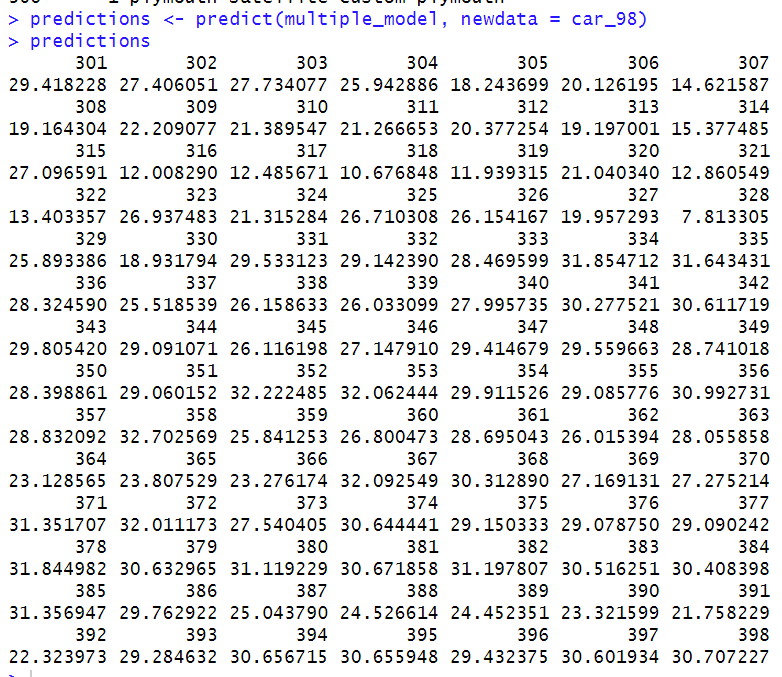
1. The multiple R squared is 0.7179.
2. The adjusted R squared is 0.715.
3. The complete linear regression equation is:

Mpg = 43.7791 – 0.0061 \* weight – 0.6919 \* cylinder + 0.1149 \* acceleration

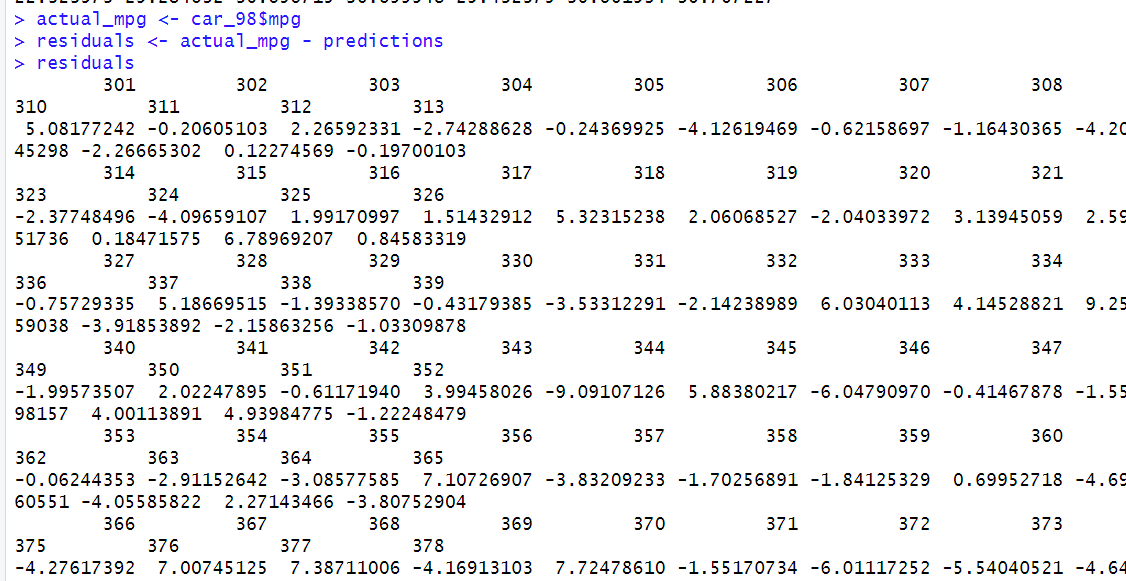




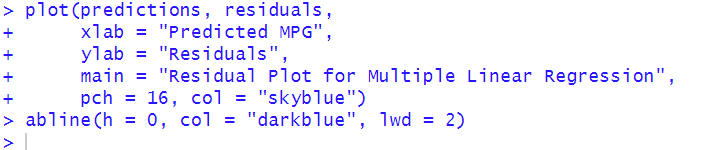
Here I stored the last 98 rows into a variable called car\_98, and displayed the first 5 rows to see.

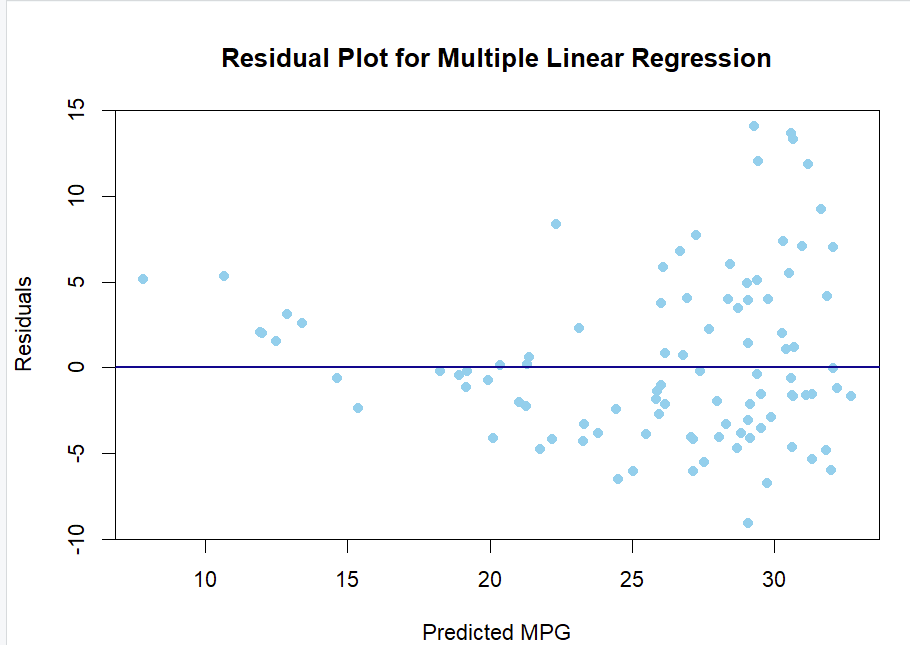


Here I did a predict; using the multiple model (the multiple linear regression model from the first 300 rows) as the “training data,” I told it to predict the mpg of the last 98 rows.

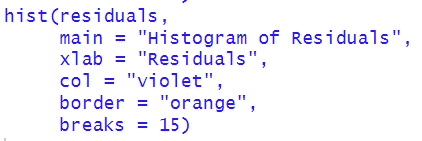


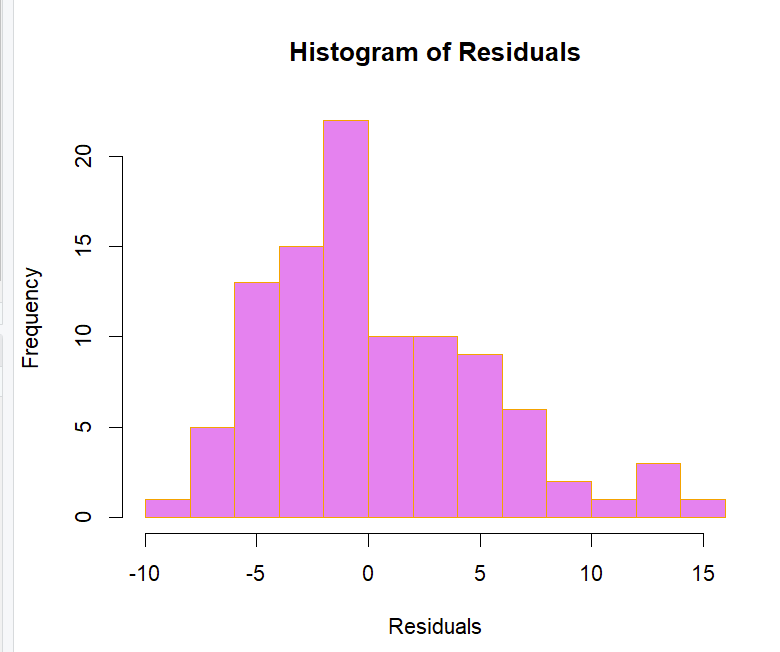
Here I found the residuals; the residuals tell me how off the predictions were from the actual.





Here I plotted the residuals so I could see them visually. Looking at it, you can see that the dots are more clustered around the 0 and sparser as you reach the top or bottom, showing that the predictions weren’t too different from the actual mpg. However, the dots are still spread out and don’t all fall on the 0 line (residuals = 0), telling us that the predictions aren’t perfect.



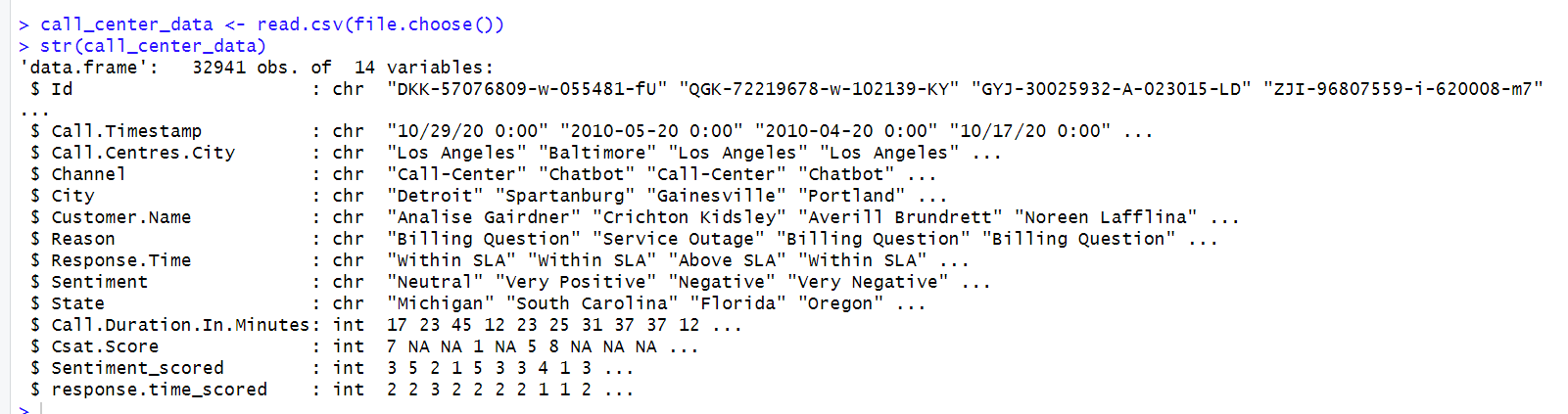


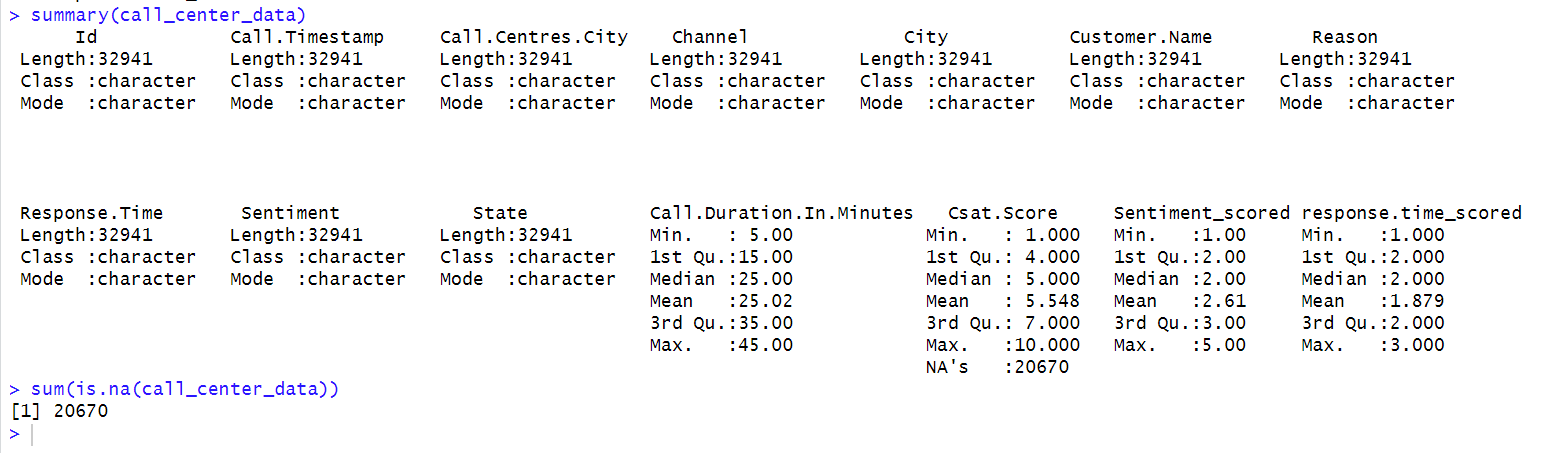
This is a histogram of the residuals.

Looking at this histogram, you can see that the highest frequency is around 0 and the frequency drops as you reach -10 or 15, showing that the predictions weren’t too different from the actual mpg. However, there is still a spread, and not everything is on the 0 line (residuals = 0), telling us that the predictions aren’t perfect.

**Part 2**

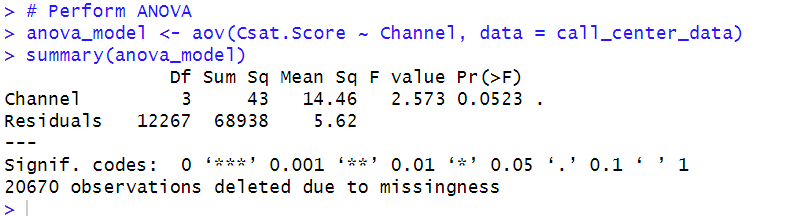
I added 2 columns, sentiment scored and response time scored, which were the same as their counterparts just continuous, so it would be easier to use for calculations.

****

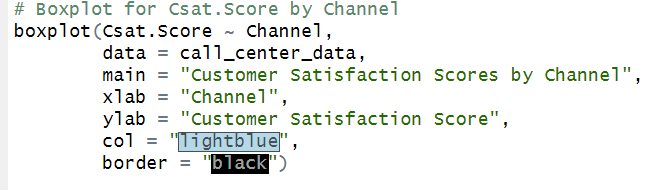
****

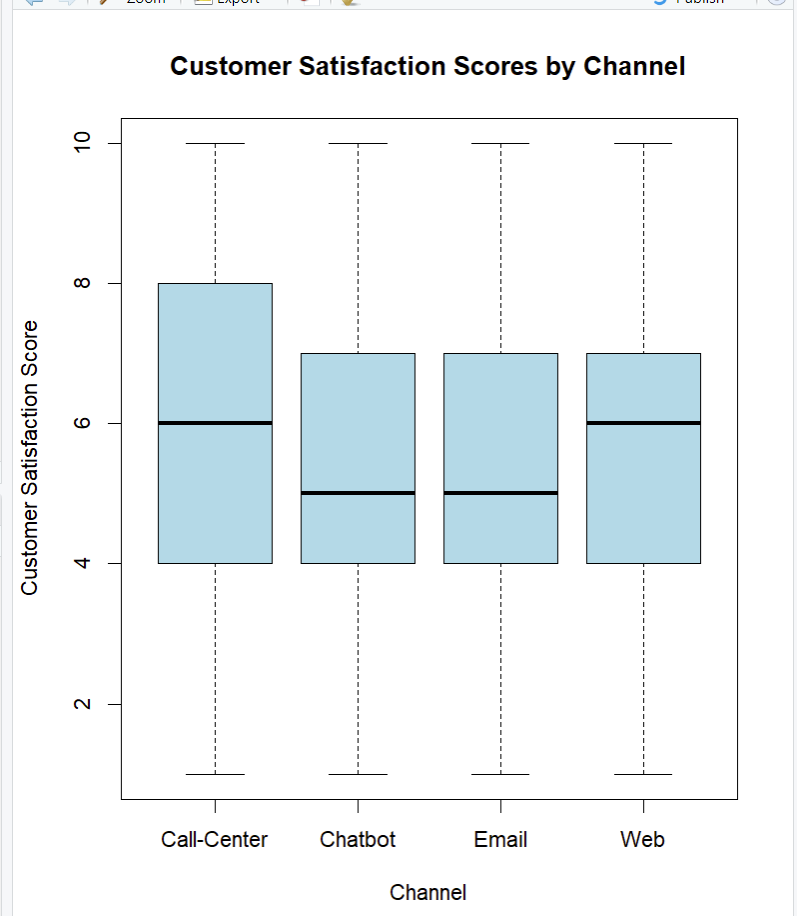
1. **Q: Is there a correlation between customer satisfaction and channel?**

**A: At the 5% significance level, there is no strong evidence to conclude that the Channel has a significant effect on Csat.Score. However, the Call-Center has the highest median satisfaction score, which suggests that customers may be more satisfied when using this channel.**

****

The p-value indicates the probability that the observed differences between the means of the Channel groups occurred due to chance. The p-value is 0.0523, which is slightly above 0.05, the conventional threshold for statistical significance.   
At the 5% significance level, there is no strong evidence to conclude that the Channel has a significant effect on Csat.Score.  
However, the p-value is very close to 0.05, suggesting there may be a weak or borderline effect.



****

Interpretation by Channel

1. Call-Center:
   * The median score is higher compared to other channels (around 6.5).
   * The spread (IQR) is large, meaning scores vary widely.
   * There are scores close to both the minimum (1) and maximum (10).
2. Chatbot:
   * The median score is lower (around 5).
   * The scores are more tightly packed (narrower IQR).
   * The whiskers extend fully, showing some lower satisfaction scores.
3. Email:
   * Similar to Chatbot in terms of median (around 5).
   * The IQR is narrow, indicating less variation in scores.
   * Whiskers extend to both ends, suggesting a broad range of responses.
4. Web:
   * The median score is similar to Email and Chatbot (around 6).
   * The IQR is wider, meaning there’s more variation in satisfaction scores compared to Chatbot and Email.

 Call**-Center** has the **highest median** satisfaction score, which suggests that customers may be more satisfied when using this channel.

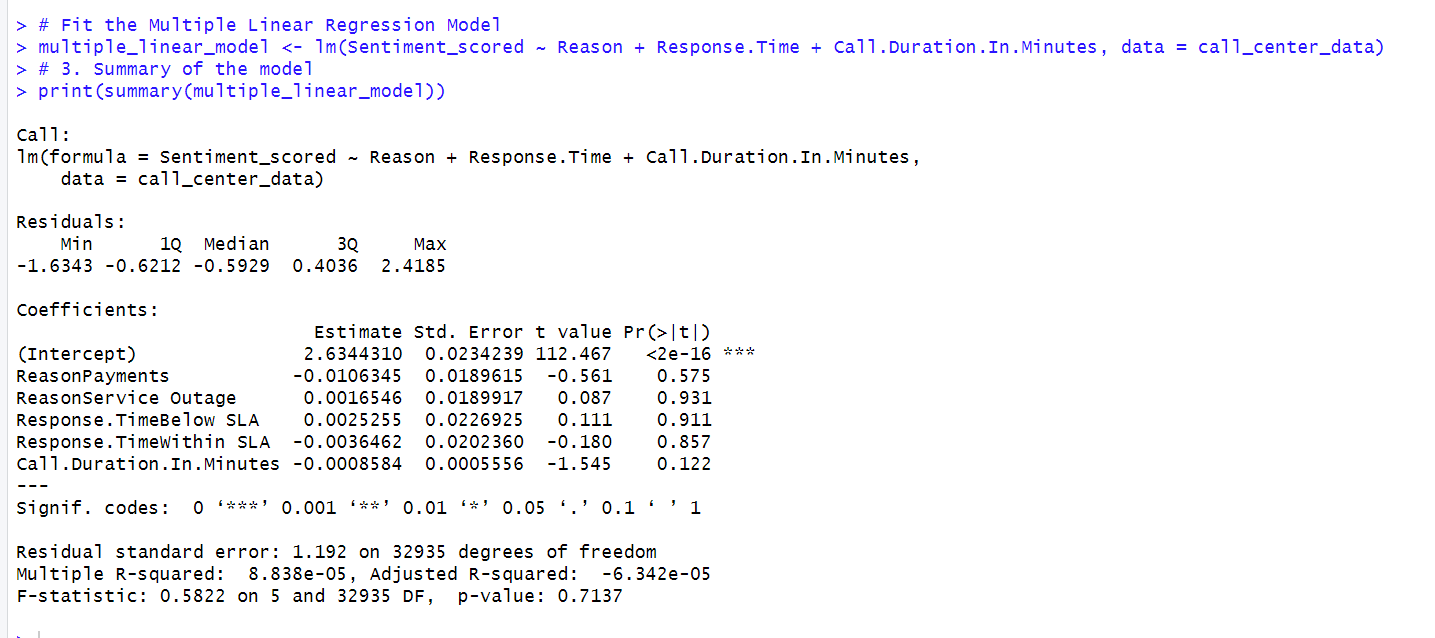
 Chatbot and **Email** channels have the **lowest medians** and narrower IQRs, indicating more consistent, but lower satisfaction.

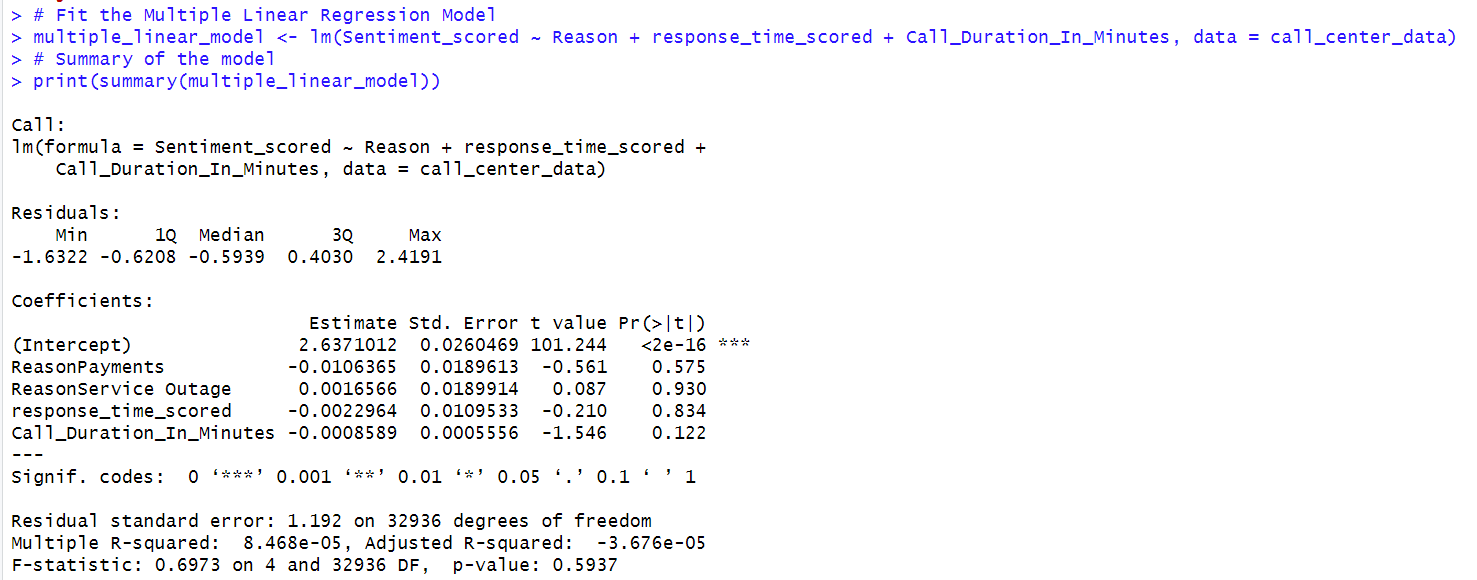
 Web has a slightly higher median than Chatbot and Email but also shows greater variability.

Based on the analysis:

1. **Investigate Call-Center performance** to identify what contributes to its higher satisfaction scores.
2. **Improve the Chatbot and Email channels** since they have the lowest satisfaction scores.
3. Further analyze the **Web channel** to understand why its scores vary widely.
4. **Q: Is there a correlation between the sentiment and the reason, response time, and call duration?**

**A: There is no correlation between sentiment and reason, response time, and call duration. None of these independent factors seem to have an influence or impact on the sentiment.**

****



**Interpretation of Coefficients:**

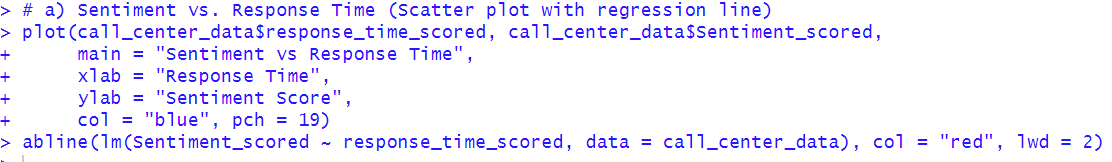
* **Intercept:**  
  The predicted Sentiment\_scored is 2.63 when all other predictors are at their baseline or 0. This value is significant (p < 2e-16), meaning the intercept is reliable.
* **Reason:**

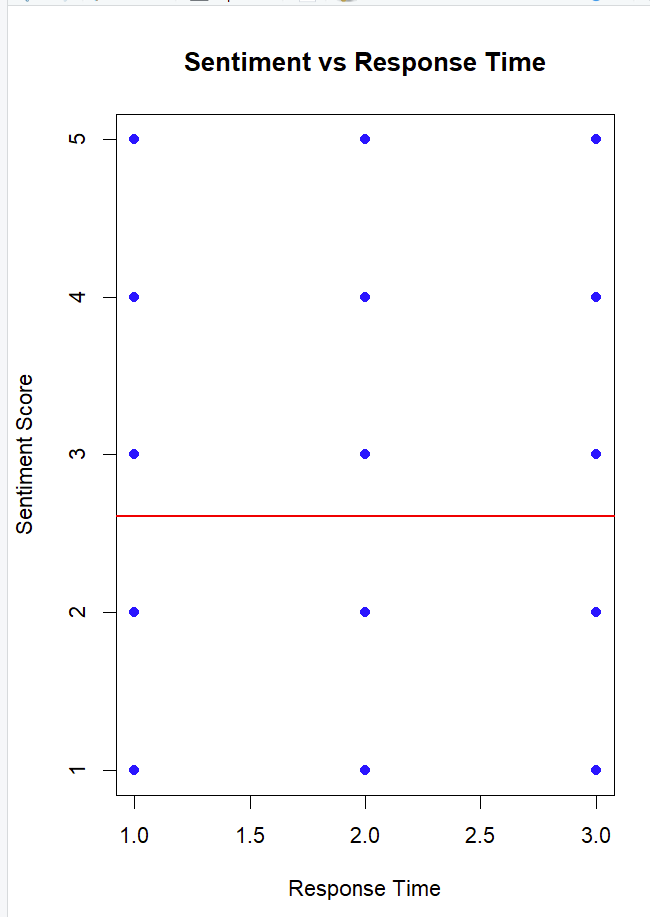
None of the levels (Payments, Service Outage, etc.) significantly affect Sentiment\_scored as all p-values > 0.05.

* **response\_time\_scored:**  
  Response time categories do not significantly predict Sentiment\_scored (p-values of 0.930 and 0.834).
* **Call\_Duration\_In\_Minutes:**  
  This variable has a negative estimate (-0.0008), suggesting that longer call durations slightly decrease sentiment, but the result is not significant (p = 0.122).
* None of the predictors are significant at the 0.05 level.

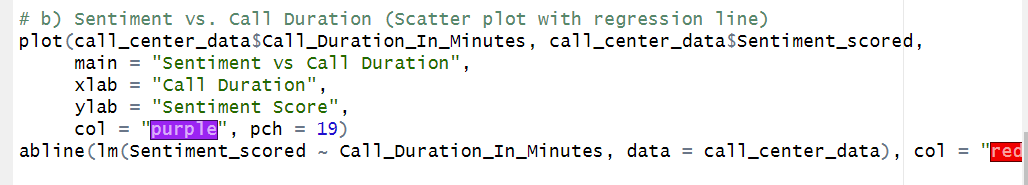
**Model Performance Metrics:**

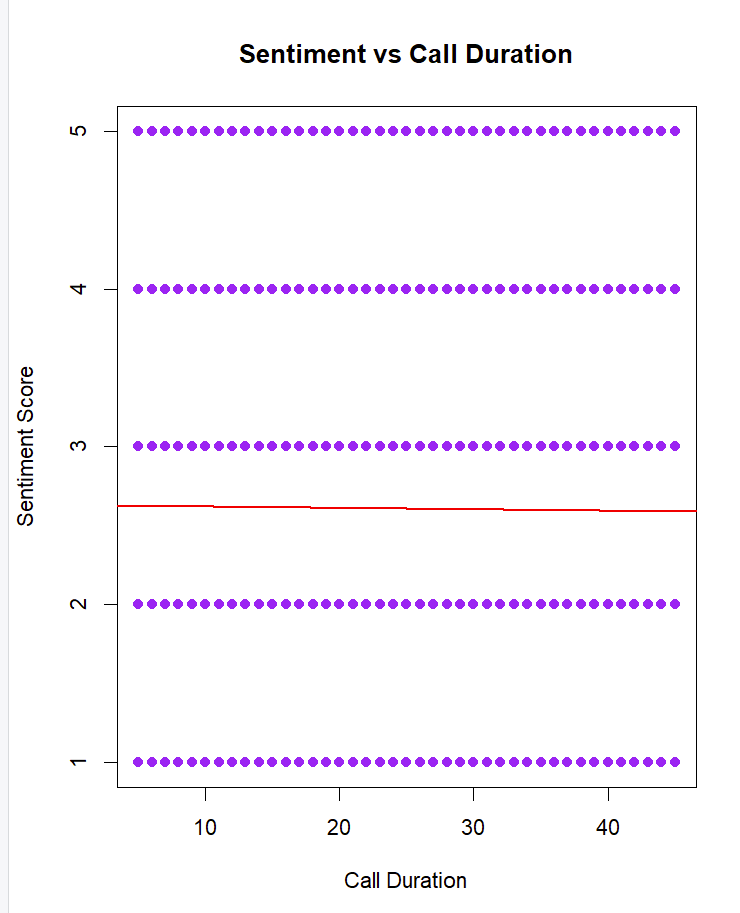
* **Residual standard error**: 1.192  
  This measures the average size of the residuals.
* **R-squared**: 8.46e-05  
  This indicates that the model explains **less than 0.01%** of the variance in Sentiment\_scored. The model has extremely poor explanatory power.
* **Adjusted R-squared**: -3.67e-05  
  Adjusted R-squared accounts for the number of predictors and confirms the poor fit of the model.
* **F-statistic**: 0.697, p-value = 0.593  
  The overall model is **not significant** (p > 0.05), meaning the predictors do not collectively explain variance in Sentiment\_scored.



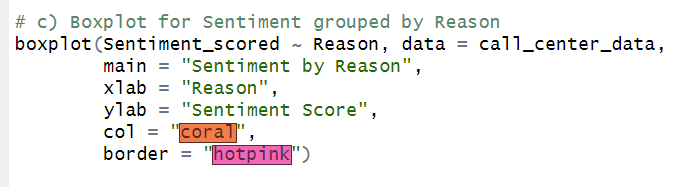


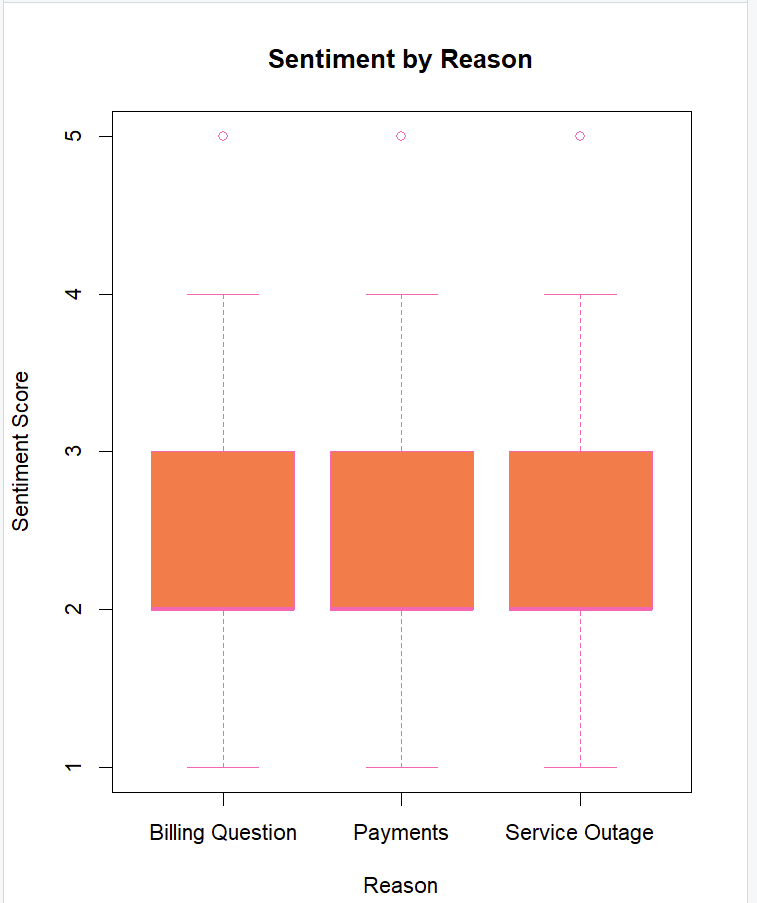
* Each blue dot corresponds to an observation in the dataset, showing the relationship between Response Time and Sentiment Score.
* The red line represents the regression line that best fits the data based on the linear model.
* The line is relatively flat, suggesting no clear relationship or correlation between Response Time and Sentiment Score.
* The Sentiment Scores (y-values) are spread across the entire range (1 to 5) for all values of Response Time (x-values: 1, 2, and 3).
* There is no visible trend (e.g., increasing or decreasing) between the two variables.
* This indicates that Response Time does not appear to influence or predict the Sentiment Score in this dataset.
* The scatter plot and regression line suggest that there is no strong linear relationship between Response Time and Sentiment Score. This is supported by the horizontal nature of the red regression line.





* Each purple dot represents an observation of the call duration and its corresponding sentiment score.
* The points are horizontally aligned for each sentiment score (1, 2, 3, 4, and 5).
* Call durations are spread evenly across the x-axis, suggesting a broad range of durations was observed.
* The regression line is nearly flat (horizontal), indicating no linear relationship between Call Duration and Sentiment Score.
* The lack of slope suggests that changes in Call Duration do not meaningfully influence Sentiment Scores.
* For any given call duration, sentiment scores appear at all levels (1 to 5). This further supports the observation that Call Duration has no clear effect on Sentiment Scores.

****



The sentiment scores range from 1 to 5, where:

* 1 represents the lowest sentiment (very negative).
* 5 represents the highest sentiment (very positive).

The boxplot groups sentiment scores by Reason:

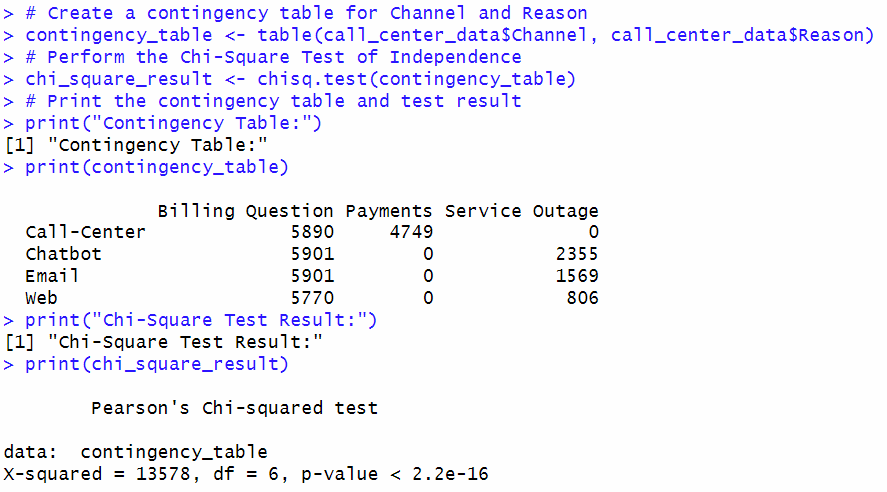
* Billing Question
* Payments
* Service Outage

For each reason:

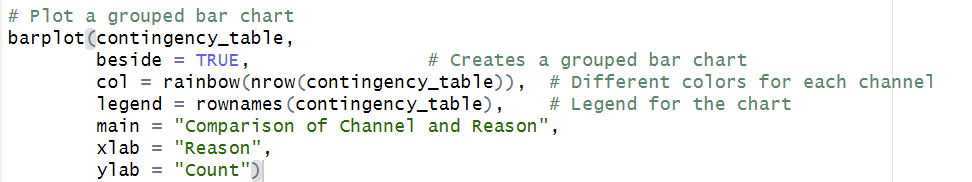
* The orange boxes represent the interquartile range (IQR), where 50% of the data lies between the lower quartile (Q1) and upper quartile (Q3).
* The horizontal line inside each box represents the median (middle value) of the sentiment scores.
* The whiskers (lines extending from the boxes) represent the range of the data, excluding outliers.
* The small circles above the whiskers represent outliers—sentiment scores that are unusually high.
* The median sentiment score for Billing Question, Payments, and Service Outage appears to be approximately the same, around 3.
* The IQR (size of the orange box) is similar for all three categories, suggesting that sentiment scores are distributed in a similar range across the reasons.
* Most sentiment scores are concentrated between 2 and 4 for all categories, suggesting neutral to slightly negative sentiment across the reasons.
* The sentiment scores are very similar across the three reasons, with no significant differences in medians or IQRs. This indicates no clear correlation between the sentiment scores and the reason.

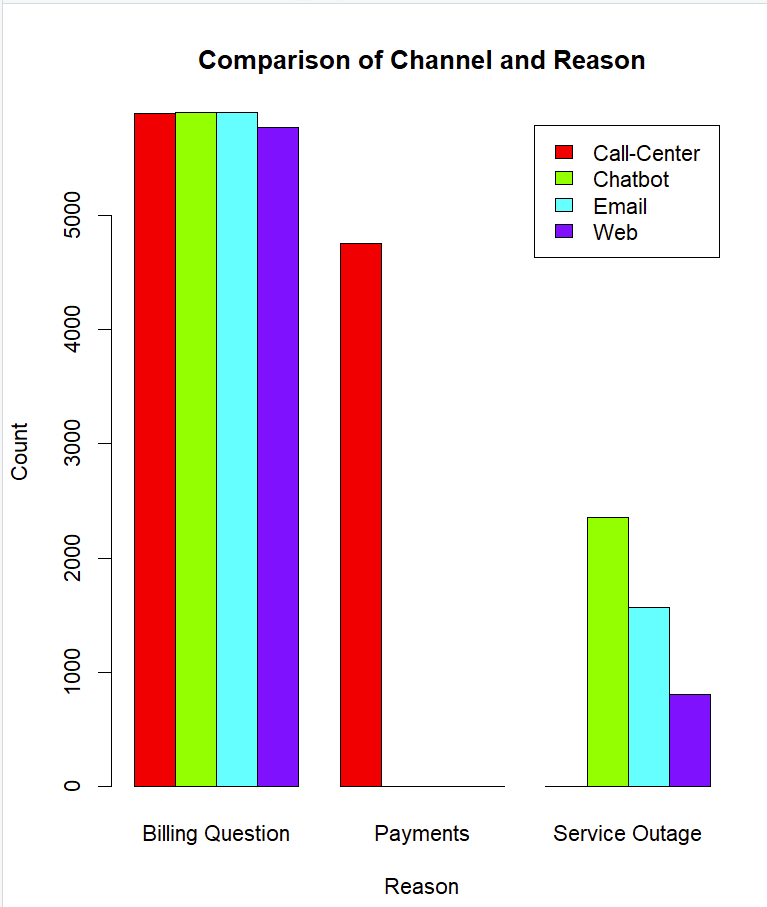
1. **Q: Is there a correlation between the channel and the reason?**

* **A: The p-value is extremely small, and this indicates that there is a statistically significant association between the communication channel (Channel) and the reason for the interaction (Reason). The type of communication channel used strongly depends on the reason for the interaction.**



* The Payments reason is exclusively handled by the Call-Center.
* Service Outage is mostly handled by Chatbot, followed by Email and Web.
* Billing Questions are distributed across all channels, but counts are slightly higher in Chatbot and Email compared to Web and Call-Center.
* The p-value is extremely small (2.2e-16), which means the null hypothesis is rejected.
* This indicates that there is a statistically significant association between the communication channel (Channel) and the reason for the interaction (Reason).
* The type of communication channel used (Call-Center, Chatbot, Email, Web) strongly depends on the reason for the interaction (Billing Question, Payments, Service Outage).





**Grouped Bar Chart**: Easier to compare counts across Reason for each Channel.

1. Billing Question:
   * All four channels (Call-Center, Chatbot, Email, and Web) handled Billing Questions.
   * The counts for Call-Center, Chatbot, Email, and Web are approximately the same, suggesting that Billing Questions are equally distributed across channels.
2. Payments:
   * Call-Center is the only channel handling Payments.
   * This indicates that Payments issues are exclusively managed through human interaction (Call-Center).
3. Service Outage:
   * Chatbot handles the highest number of Service Outage queries, followed by Email, and then Web.
   * Call-Center does not handle any Service Outage queries.
   * This shows that customers prefer self-service or automated channels like Chatbot, Email, and Web for Service Outage issues.

Recommendations:

1. Expand Support for Payments:  
   Introduce Payments handling in Chatbot, Web, or Email to reduce Call-Center workload.
2. Enhance Self-Service Options:
   * Continue to promote Chatbot, Email, and Web for Service Outage-related issues.
   * Explore if Billing Questions can be automated further.
3. Monitor Workload Balance:  
   Track how the Call-Center resources are utilized for Billing Questions and Payments, as those two reasons account for a large volume of queries.