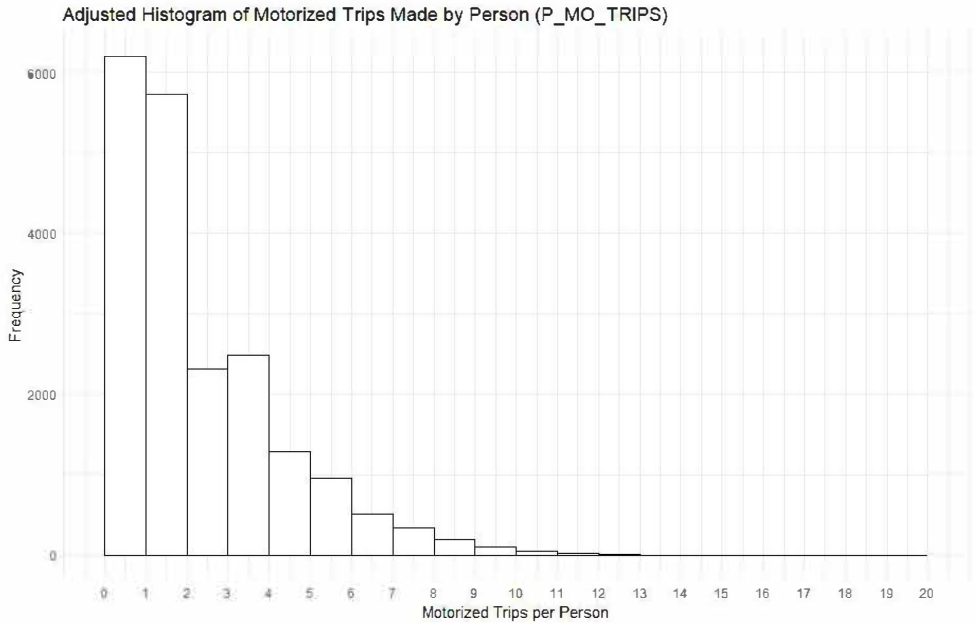
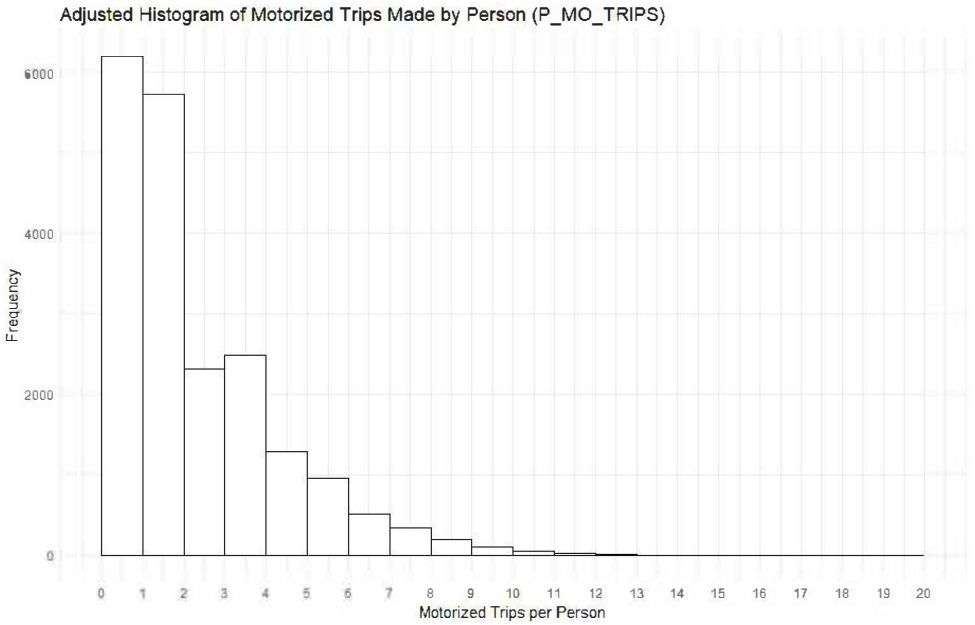
1) Using the 2012 Philadelphia household travel survey, plot a histogram of the total number of trips people made (P_TOT_TRIPS). Describe the distribution of trip-making.





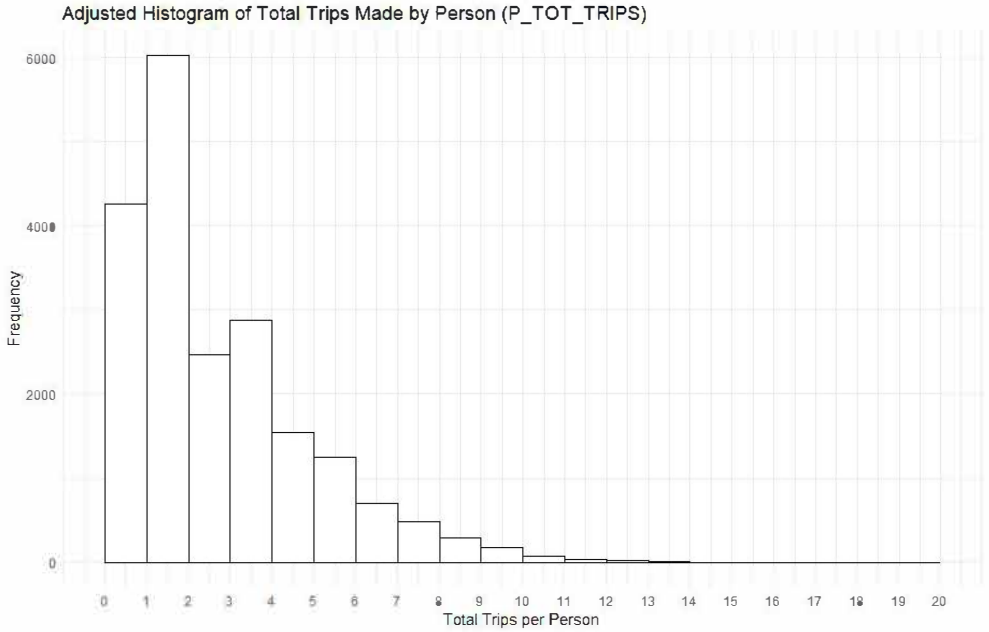
We derived the following three charts from the 2012 Philadelphia Household Travel Survey.

- From the 'Adjusted Histogram of Motorized Trips Made by Person,' it can be observed that the most frequent number of motorized trips per person falls within the 0-2 range, with a frequency of around 6000. Subsequently, the frequency of motorized trips per person in the 2-4 range decreases by more than half, to approximately 2500. As the number of motorized trips per person increases, the frequency of travelers decreases significantly, displaying an inverse relationship.

- The 'Adjusted Histogram of Non-Motorized Trips Made by Person' reveals that the frequency of non-motorized trips per person is significantly higher within the 0-1 range, around 17500. The number of non-motorized trips per person exceeding 1 is much lower, all below 1300, and their quantity is inversely proportional to frequency.

Philadelphia Household Travel Survey Transportation Modeling

1) Using the 2012 Philadelphia household travel survey, plot a histogram of the total number of trips people made (P_TOT_TRIPS). Describe the distribution of trip-making.



- The 'Adjusted Histogram of Total Trips Made by Person' indicates that the highest frequency of total trips per person is within the 1-2 range, with approximately 6000. Subsequently, it is followed by the 0-1, 3-4, and 2-3 ranges, with a gradual decrease in frequency beyond 4

From the comparison of these three charts, it can be observed that the 0-1 range in the 'Adjusted Histogram of Non-Motorized Trips Made by Person' has a high frequency break, with the most frequent occurrences at 0. Few people travel by non-motorized means, which suggests that more people opt for motorized travel due to its convenience and ability to reach more distant locations. Additionally, from the analysis of Chart 3, it can be seen that the majority of people travel 1-2 times or 0-1 times, with a relatively small number of individuals traveling more than 4 times. The frequency in Chart 1 aligns with the trends in Chart 3, reinforcing the idea that most people prefer motorized means of travel.

Philadelphia Household Travel Survey Transportation Modeling

2) Create a new variable that equals 1 if a person did not take any trips. How many people in the sample took no trips? Summarize the race, age, and income of those who took trips on the survey day vs. those that did not.

First, here is a brief overview of the general steps we will take to address this question:

- Step 1: Data Loading and Cleaning
  Data Loading: Read individual data from the CSV file (2_Person_Public.csv) and household data from the CSV file (1_Household_Public.csv).
  Data Cleaning: Handle missing values and error codes, such as marking "98 Don't know" and "99 Refused" for income as NA.

- Step 2: Data Preprocessing and Feature Engineering
  Merge Datasets: Combine individual data and household data based on HH_ID to calculate the average income for each household.
  Create New Variable: Create a new variable "No_Trip" in the individual dataset to indicate whether a person has not taken any trips (trip count is 0).

- Step 3: Statistical Analysis
  Number of Non-Travelers: Calculate the number of individuals who have not taken any trips.
  Group Statistics: For each "No_Trip" group, compute the average age category (Avg_Age_Cat), race distribution (Race_Distribution), and average income (Avg_Income).

- Step 4: Visual Analysis
  Race and Age Distribution: Use bar charts to display the travel status of different races and age groups.
  Income Distribution: Use box plots to show the income distribution for individuals with and without trips.

- Step 5: Interpretation of Results
  Race and Age: Through the bar charts, we can observe whether certain races or age groups are more inclined to travel or stay at home.
  Income: The box plots can reveal the median, quartiles, and outliers of income for individuals with and without trips, helping to understand how income levels influence travel behavior.

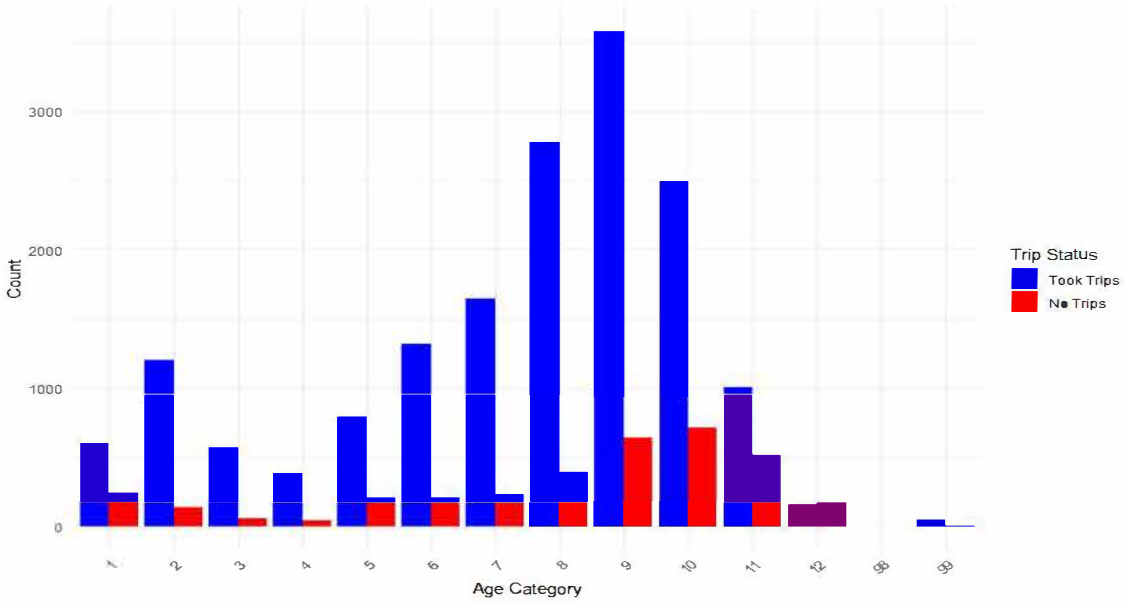
2) Create a new variable that equals 1 if a person did not take any trips. How many people in the sample took no trips? Summarize the race, age, and income of those who took trips on the survey day vs. those that did not.

So here is the answer:

The number of non-travelers can be obtained using sum(person_data$No_Trip, na.rm = TRUE).

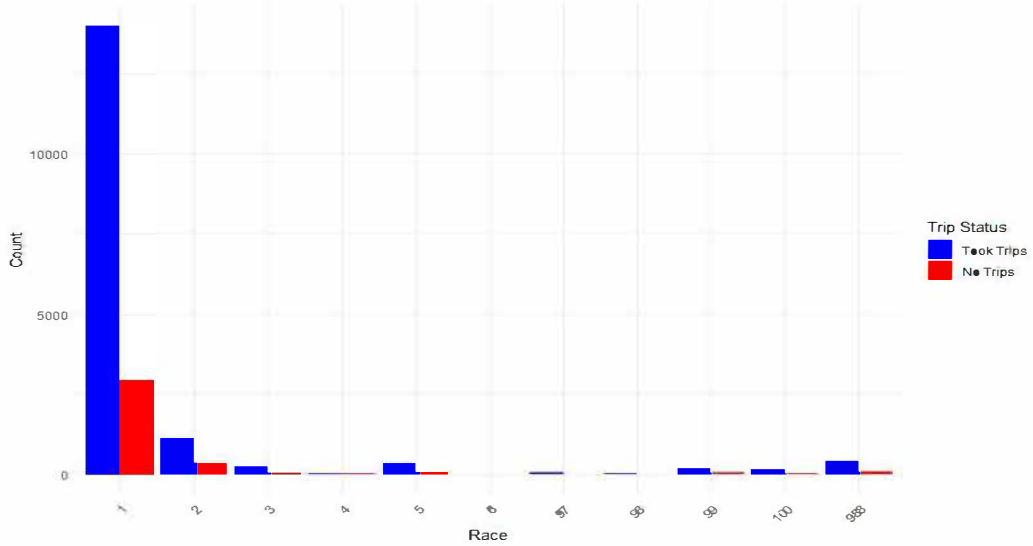
The number of people took no trips is: 3625.

Avg_Age_Cat represents the average age category grouped by travel status.



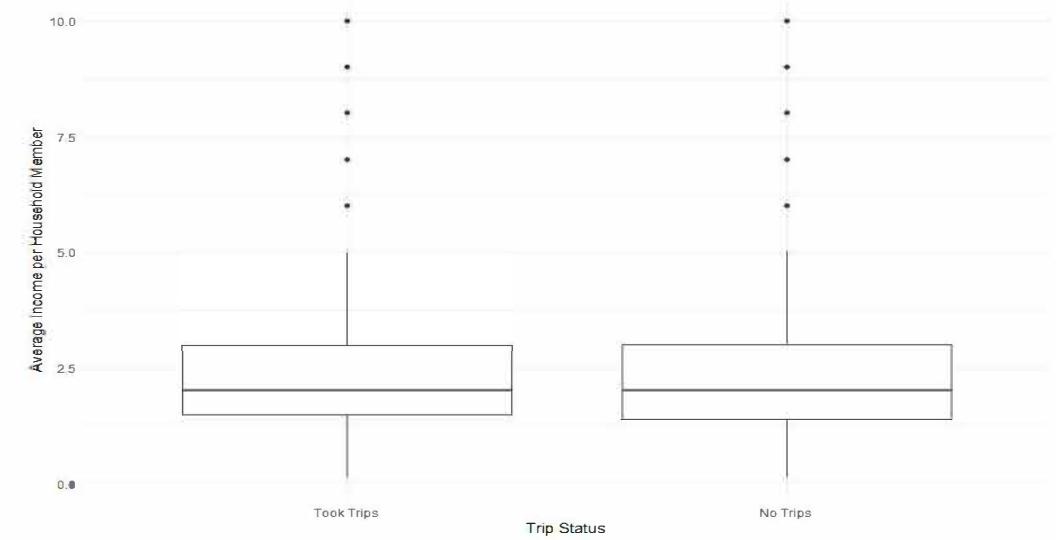
From this table, it is evident that individuals who had "no trips" outnumbered those who "took trips" only in the age group 12 (86 years and above). Among the surveyed populations, those in the age group 9 (55-64 years) had the highest number of individuals who "took trips," and the difference in numbers between this group and those with "no trips" was the most significant. We hypothesize that this is due to the fact that many individuals in this age group have recently retired, possessing both the financial means and the time to travel. Other travel patterns are also correlated with the life situations and economic conditions corresponding to different age groups. During childhood, individuals often travel with their families, resulting in a higher frequency of "took trips." In adolescence and young adulthood, the number of "took trips" decreases, likely due to academic commitments, while the number of "no trips" remains relatively stable. As individuals enter the workforce and gain economic independence, the number of "took trips" increases until aging sets in.

Race_Distribution displays the distribution of different races within the groups of individuals with and without travel.



The information in the table indicates a correlation between travel patterns and racial/ethnic backgrounds. While the basic quantities surveyed for each racial/ethnic group differ, the disparity in proportions between "took trips" and "no trips" reveals distinctions based on racial/ethnic identity. For example, in the case of racial/ethnic group 1 (White), there is a higher number of individuals who took trips.

Avg_Income showcases the average income levels for individuals with and without travel.



This table highlights the correlation between travel patterns and economic conditions. We converted the average income per household member into the midpoint average value. The average income per household member for those who "took trips" is higher than for those who had "no trips." This suggests that economic circumstances directly impact the likelihood of traveling; individuals with better economic conditions are more likely to embark on journeys.

2) Create a new variable that equals 1 if a person did not take any trips. How many people in the sample took no trips? Summarize the race, age, and income of those who took trips on the survey day vs. those that did not.

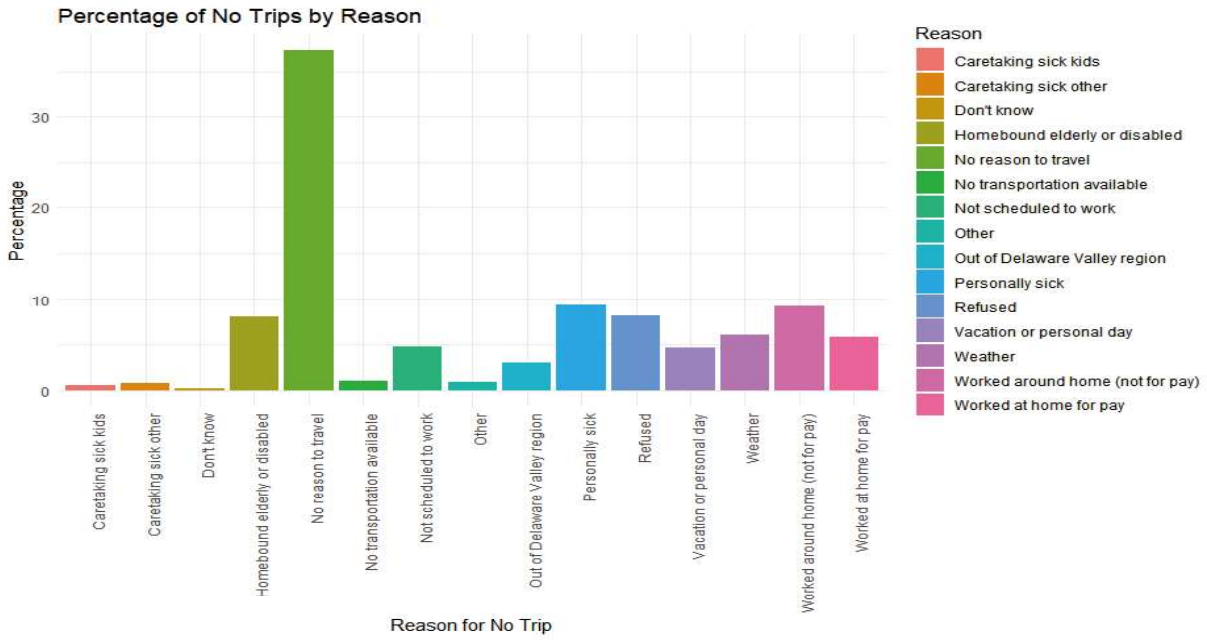
Appendix: Avg_Age_Cat, in the given context, is a statistical metric we plan to calculate, representing the average age category grouped by travel status (No_Trip). It is derived by taking the average of AGECAT values for all individuals within their respective No_Trip groups.

In the provided R code, AGECAT is a variable that indicates the age group to which individuals belong. This variable has multiple categories, such as:
a: 5 years or younger
b: 6 to 12 years
c: 13 to 15 years
and so on.

The mean(AGECAT, na.rm = TRUE) within the summarise function calculates the average age category for individuals in the two groups: those with and without travel. This can provide an intuitive understanding of the age distribution within these two groups. For instance, a higher Avg_Age_Cat value might indicate that a group is generally older, while a lower value suggests that the group is generally younger. Here, na.rm = TRUE is a parameter that instructs R to ignore NA (missing) values when calculating the mean.

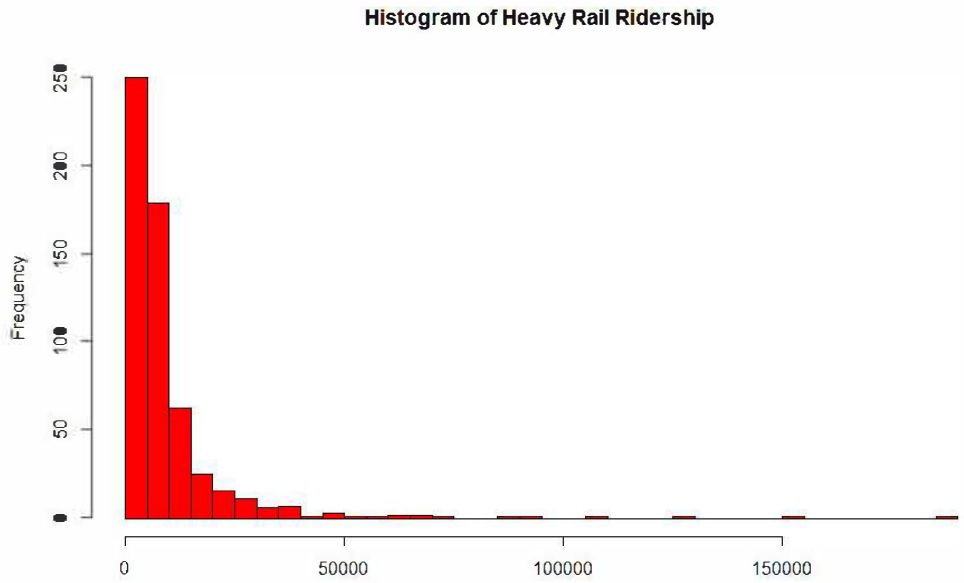
3) Create a table showing the percent of people who did not take a trip by the reason they did not take a trip.

Reason_Code	Percentage	Reason
1	9.40689655	Personally sick
2	4.63448276	Vacation or personal day
3	0.55172414	Caretaking sick kids
4	0.8	Caretaking sick other
5	8.05517241	Homebound elderly or disabled
6	5.82068966	Worked at home for pay
7	4.74482759	Not scheduled to work
8	9.29655172	Worked around home (not for pay)
9	0.99310345	No transportation available
10	3.06206897	Out of Delaware Valley region
11	6.09655172	Weather
12	37.2689655	No reason to travel
97	0.91034483	Other
98	0.19310345	Don't know
99	8.16551724	Refused

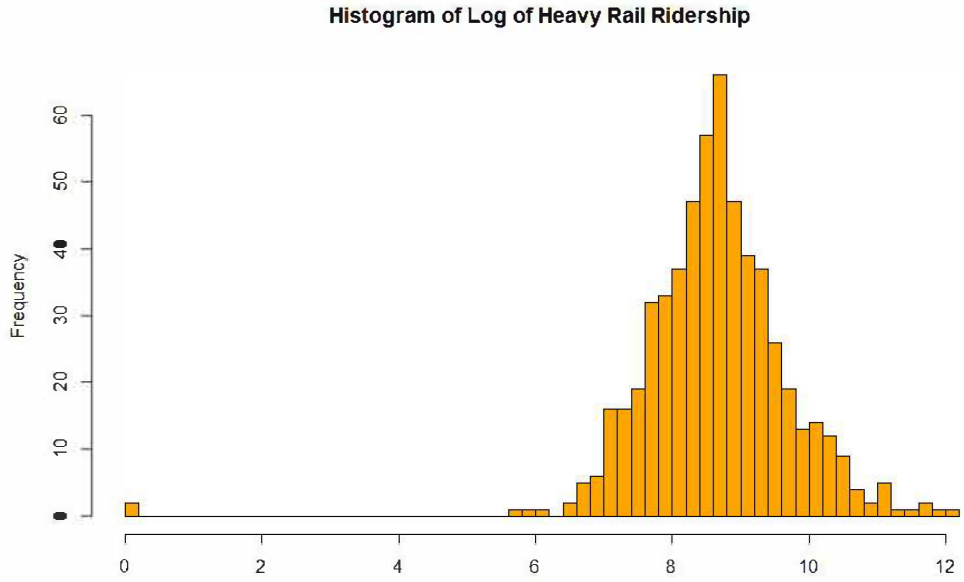


On the left is a table showing the percentage of non-travelers corresponding to the reasons for not traveling. On the right is the generated histogram. It can be clearly observed that the most dominant reason for not traveling is "no reason to travel," while the rest of the reasons for non-traveling have relatively lower percentages, and they are distributed fairly evenly. These reasons include "Personally sick," "Weather," "Worked at home," and others.

4) Plot a histogram of heavy rail ridership and a histogram of the natural log of heavy rail ridership. Describe the two plots.

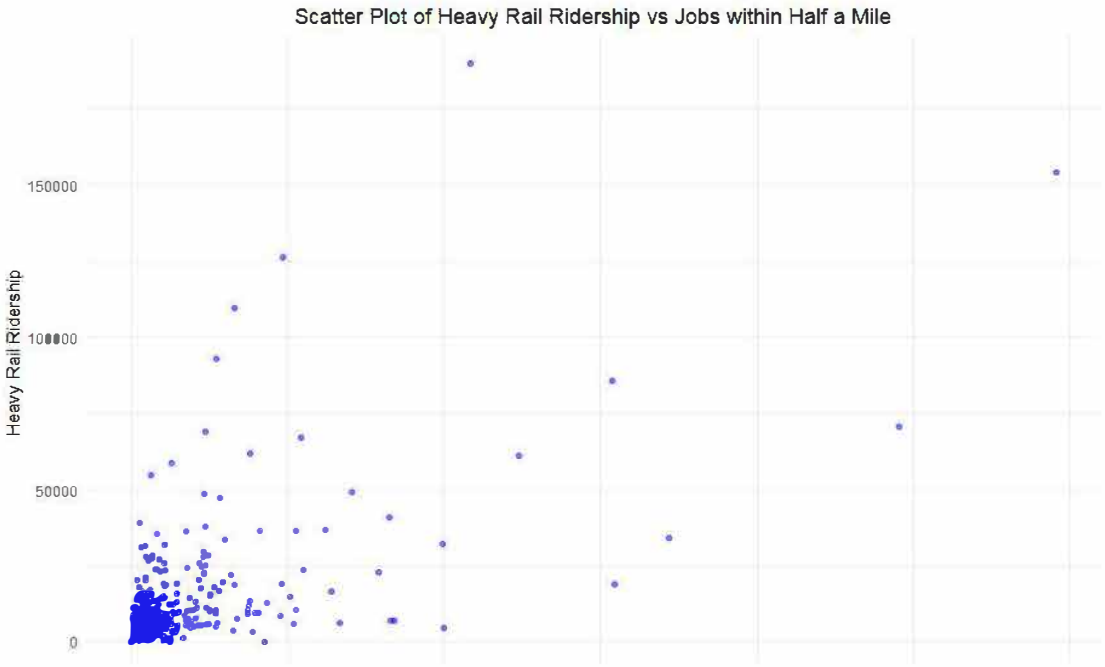


Heavy Rail Ridership Counts Histogram: This histogram illustrates the frequency distribution of passenger counts at different levels. From the histogram, it can be discerned that the data may be skewed, as the heights of certain bars significantly exceed those of others, indicating that passenger counts at some stations are much higher than the average level. Within the chart, ridership within the 0-5000 range exhibits a considerably higher frequency compared to other ridership ranges, and the frequency gradually decreases as ridership increases.



Natural Log Histogram of Heavy Rail Ridership Counts: By taking the logarithm, the distribution of passenger count data may appear to be closer to a normal distribution. This is because taking the logarithm often makes the data more concentrated. In this histogram, the data distribution appears to be more symmetric compared to the distribution of the original data. Here, the width of the histogram represents the data's level of dispersion. A wider histogram in this context indicates more dispersed data. Additionally, in the chart, the highest frequency of frequency occurs at a Log(ridership) value of 8.8, which is consistent with the results from the previous table.

5) Plot a scatter plot of heavy rail ridership (y axis) against the jobs within a half mile of stations. Describe the relationship.



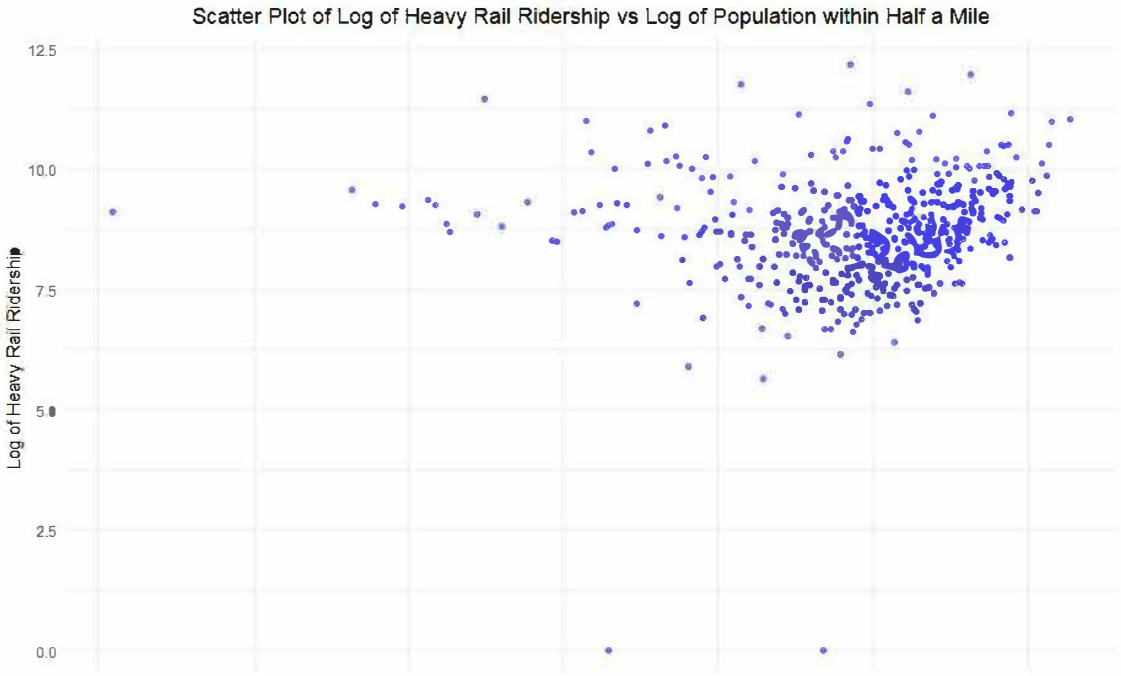
According to the scatter plot, we can observe the relationship between the number of job positions within a half-mile radius of the station (x-axis) and the average weekday transit passenger counts (y-axis). Based on the distribution of points observed in the plot, we can make the following descriptions and inferences:

Scatter Plot Distribution:
The distribution of points may display a certain degree of positive relationship, suggesting that areas with a higher number of job positions may also have higher passenger counts.
If the points show an upward trend, it indicates a potential positive correlation between the two variables.

Correlation Inference:
If there is a positive correlation, it could be because a higher number of job positions attract more people to use heavy rail transportation, especially if these job locations are close to heavy rail stations.
Another possibility is that heavy rail stations with higher passenger counts in their vicinity are more likely to develop more job positions due to the convenience of transportation.
It's important to note that correlation does not imply causation. Other factors such as the station's geographical location, nearby transportation facilities, or the overall economic activity of the community may also influence both variables.

From the plot, we can see a positive correlation between the number of job positions within a half-mile radius of the station (x-axis) and the average weekday transit passenger counts (y-axis). Therefore, it aligns with the inference mentioned above and suggests a reasonable mutual influence and correlation between these variables.

6) Plot a scatter plot of the natural log of heavy rail ridership (y axis) against the natural log of people within a half mile of stations. Describe the relationship.



From the graph, we can observe a positive correlation. If the natural log of people within a half mile of stations increases, the natural log of heavy rail ridership (y-axis) also increases. It is evident that as the population grows, the demand for travel and traffic density increases, leading to an increase in passenger counts.

Based on the plotted scatter chart, we can observe the relationship between the natural logarithm of Heavy Rail Ridership (ln(Heavy Rail Ridership)) and the population within a half-mile radius (ln(Population within a half mile)). From the chart, it is evident that as the population increases, there appears to be an upward trend in passenger counts. This suggests a potential positive correlation between the two variables. However, for a more precise description of the strength and form of this relationship, further in-depth statistical analysis, such as calculating correlation coefficients or conducting regression analysis, is necessary.

7) Predict station level ridership (linear) as a function of jobs within a half mile, population within a half mile, whether the station is a terminal, whether it connects to an airport. Provide the output of the regression.

We created a new data frame, "heavy_rail_data," specifically to contain data for Heavy Rail Transit (HRT) stations.

We calculated the logarithms of passenger counts, job positions, and population numbers.

We fitted a linear model using the lm() function, with the logarithmically transformed passenger counts as the dependent variable and logarithmically transformed job positions, logarithmically transformed population numbers, whether the station is a terminus, and whether the station is connected to the airport as independent variables.

We used the summary() function to generate a detailed summary of the model, providing information such as coefficients, R-squared values, and other statistical details.

```
Call:
lm(formula = log_rider ~ log_jobs_halfmile + log_pop_halfmile +
    terminal_d + airport_d, data = heavy_rail_data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5015 -0.4311  0.0194  0.5240  2.5136

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.83156    0.45077   8.500  < 2e-16 ***
log_jobs_halfmile 0.47331   0.03696  12.806  < 2e-16 ***
log_pop_halfmile  0.11179   0.03710   3.013   0.0027 **
terminal_d        0.54800   0.12620   4.342 1.67e-05 ***
airport_d         0.54767   0.43708   1.253   0.2107
---
Signif. codes:  0 ¡®***¡¯ 0.001 ¡®**¡¯ 0.01 ¡®*¡¯ 0.05 ¡®.¡¯ 0.1 ¡® ¡¯ 1

Residual standard error: 0.9432 on 569 degrees of freedom
Multiple R-squared:  0.2432,  Adjusted R-squared:  0.2379
F-statistic: 45.71 on 4 and 569 DF,  p-value: < 2.2e-16
```

8) Describe the statistical relationship between the variables and transit ridership.

When modeling to predict passenger counts at Heavy Rail Transit (HRT) stations, several factors should be considered:

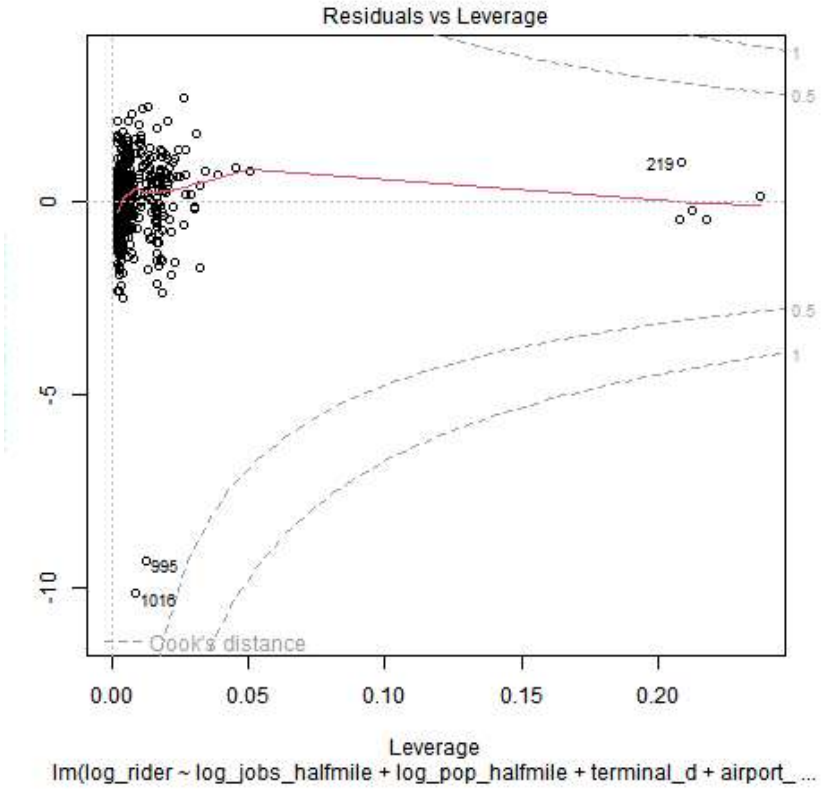
- Variable Selection: Choose variables closely related to passenger counts, such as job positions and population within a half-mile radius, whether the station is a terminus, and whether the station is connected to the airport.
- Data Transformation: For non-normally distributed data, like passenger counts and job positions, log transformation can help improve the model's performance and predictive accuracy.
- Model Type: Utilize a linear regression model to predict passenger counts, as it can provide the interpretability and predictive capability we require.
- Model Evaluation: Assess the model's explanatory power and predictive accuracy using metrics such as the coefficient of determination (R-squared) and other statistical tests.

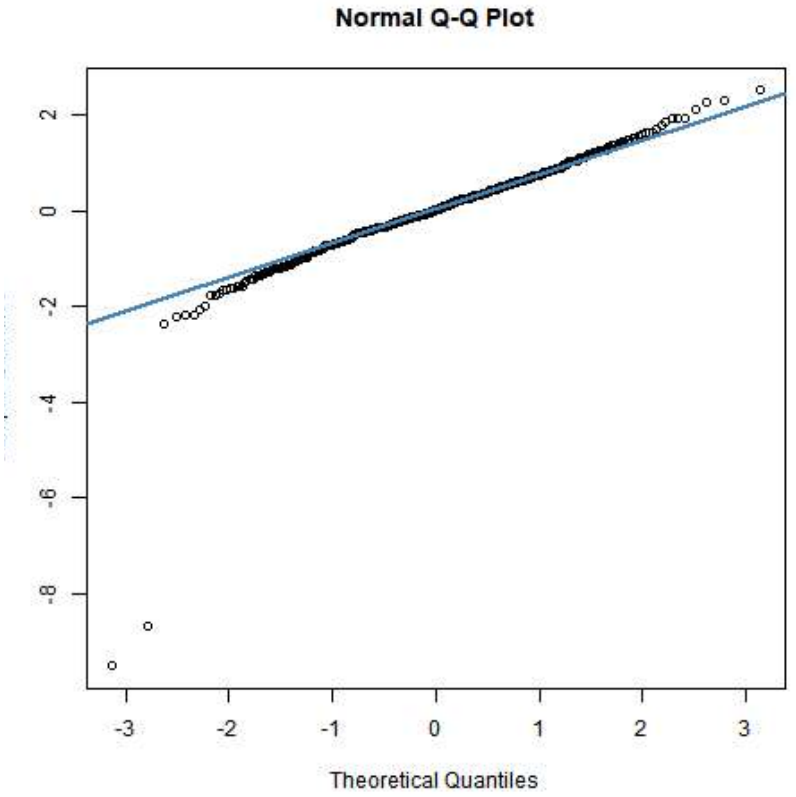
Here is the rationale and logic behind the model, along with some key considerations:

- Selecting job positions and population counts as predictor variables is based on their direct influence on the transportation demand in an area. More job positions imply that more people are likely to use public transportation to commute to work, and higher population density is often positively correlated with transportation usage.
- Logarithmic transformation is applied to help stabilize the variance of variables, reduce the impact of outliers, and better align the model with the normality assumption of linear regression.
- For model evaluation, we can examine the coefficient of determination (R-squared) to assess the model's ability to explain the variability in the target variable. An R-squared value close to 1 indicates that the model can effectively explain the variability in the target variable. Additionally, we consider p-values to test the statistical significance of model coefficients.
- Terminus stations often have higher passenger counts because they serve as starting or ending points for passengers. Similarly, stations connected to airports may experience higher passenger counts due to travelers.
- Lastly, logarithmic transformation helps mitigate the impact of outliers. However, if outliers persist, further diagnostic and treatment steps may be required, such as removal or robust regression analysis.

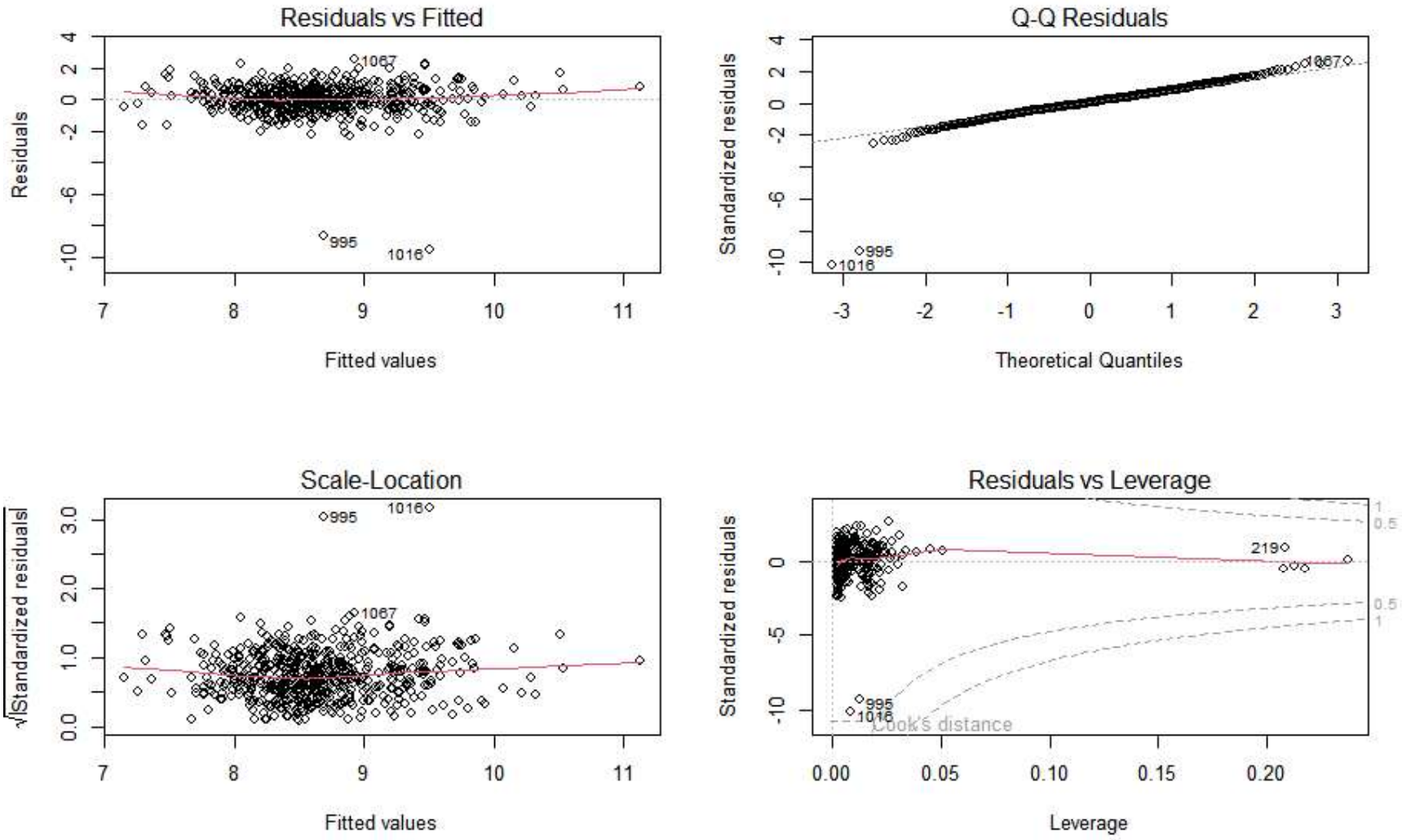
Through these steps and logic, we have developed a model capable of predicting passenger counts at HRT stations and providing insights into the factors that most influence passenger counts.

8) Describe the statistical relationship between the variables and transit ridership.

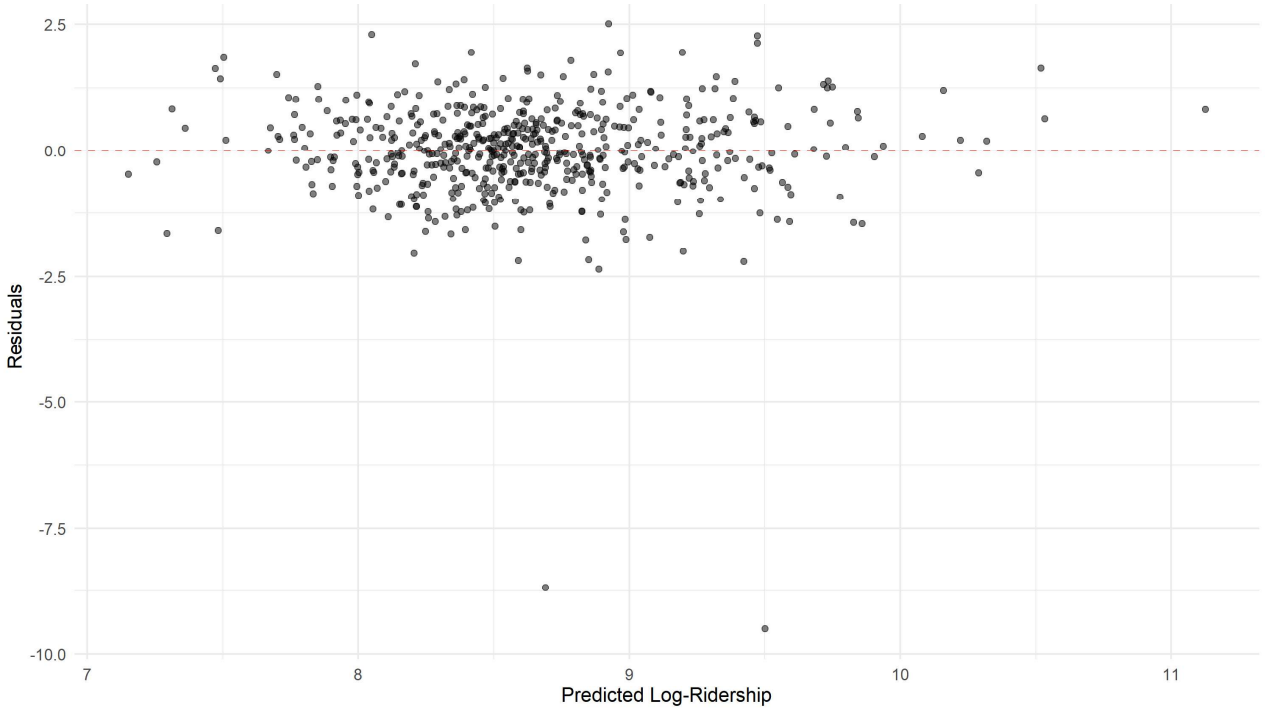




8) Describe the statistical relationship between the variables and transit ridership.



9) Plot the predicted ridership against the error terms and provide a graphic in your homework. Describe this residual plot.



In linear regression analysis, residuals (also known as errors) are the differences between actual observed values and model-predicted values. A residual plot is a commonly used tool to assess the goodness of fit of a linear regression model. An ideal residual plot shows residuals that are randomly distributed with no evident patterns, indicating that the model's errors are random and lack systematic bias.

Residual plots help us identify several issues:

- Heteroscedasticity (variance of residuals increasing or decreasing with increasing predicted values).
- Nonlinear relationships (residuals exhibiting some systematic pattern).
- Outliers or leverage points (residual values far from other data points).

10) Add the dummy variable for whether the station is a heavy rail station. Does this improve the model? Explain your answer.

Original:

```
Call:
lm(formula = log_rider ~ log_jobs_halfmile + log_pop_halfmile +
    terminal_d + airport_d, data = heavy_rail_data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5015 -0.4311  0.0194  0.5240  2.5136

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.83156    0.45077   8.500  < 2e-16 ***
log_jobs_halfmile 0.47331    0.03696  12.806  < 2e-16 ***
log_pop_halfmile  0.11179    0.03710   3.013   0.0027 **
terminal_d        0.54800    0.12620   4.342 1.67e-05 ***
airport_d         0.54767    0.43708   1.253   0.2107
---
Signif. codes:  0 ¡®***¡¯ 0.001 ¡®**¡¯ 0.01 ¡®*¡¯ 0.05 ¡®.¡¯ 0.1 ¡® ¡¯ 1

Residual standard error: 0.9432 on 569 degrees of freedom
Multiple R-squared:  0.2432,  Adjusted R-squared:  0.2379
F-statistic: 45.71 on 4 and 569 DF,  p-value: < 2.2e-16
```

Modified:

```
Call:
lm(formula = log_rider ~ log_jobs_halfmile + log_pop_halfmile +
    terminal_d + airport_d + hrt_d, data = heavy_rail_data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5015 -0.4311  0.0194  0.5240  2.5136

Coefficients: (1 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.83156    0.45077   8.500  < 2e-16 ***
log_jobs_halfmile 0.47331    0.03696  12.806  < 2e-16 ***
log_pop_halfmile  0.11179    0.03710   3.013   0.0027 **
terminal_d        0.54800    0.12620   4.342 1.67e-05 ***
airport_d         0.54767    0.43708   1.253   0.2107
hrt_d                NA         NA      NA       NA
---
Signif. codes:  0 ¡®***¡¯ 0.001 ¡®**¡¯ 0.01 ¡®*¡¯ 0.05 ¡®.¡¯ 0.1 ¡® ¡¯ 1

Residual standard error: 0.9432 on 569 degrees of freedom
Multiple R-squared:  0.2432,  Adjusted R-squared:  0.2379
F-statistic: 45.71 on 4 and 569 DF,  p-value: < 2.2e-16
```

10) Add the dummy variable for whether the station is a heavy rail station. Does this improve the model? Explain your answer.

Original Model Summary Key Points:

Four explanatory variables: log_jobs_halfmile, log_pop_halfmile, terminal_d, airport_d.
All these variables have defined coefficient estimates.
Statistical metrics of the model (e.g., R2, adjusted R2, F-statistic) indicate the model's explanatory power.
Modified Model Summary Key Points:

Addition of one explanatory variable: hrt_d.
The coefficient for hrt_d is undefined, indicating a potential issue of multicollinearity.
Statistical metrics of the model remain unchanged, suggesting that adding hrt_d did not enhance the model's explanatory power.

Comparative Analysis:

Variable Significance: In the original model, all variables have defined coefficient estimates, and most are statistically significant. In the modified model, the coefficient for hrt_d is undefined, often attributed to perfect multicollinearity among variables.

Model Explanatory Power: R2 and adjusted R2 show no change in both model summaries, suggesting that despite the addition of hrt_d, the model's explanatory power has not increased.

Multicollinearity Issue: The undefined coefficient for hrt_d suggests a high correlation with one or more variables in the model. This may be because hrt_d is a variable derived from others or is highly correlated with existing variables, making it challenging for the model to distinguish its independent effects.

10) Add the dummy variable for whether the station is a heavy rail station. Does this improve the model? Explain your answer.

Conclusion:

In the modified model, despite attempting to capture the impact of being a heavy rail station by adding the hrt_d dummy variable, the model did not show any improvement due to multicollinearity issues. In practical terms, this indicates a need to reexamine how hrt_d is included or structured in the data to reduce collinearity and reassess the model.

To enhance the model's quality and explanatory power, we may need to take the following steps:

Conduct variable selection, identifying and excluding highly correlated variables.
Use regularization regression methods to address multicollinearity.
Consider potential interactions or nonlinear relationships to more comprehensively capture the impact of explanatory variables on the response variable.

To enhance the model's quality and explanatory power, we may need to take the following steps:

Conduct variable selection, identifying and excluding highly correlated variables.
Use regularization regression methods to address multicollinearity.
Consider potential interactions or nonlinear relationships to more comprehensively capture the impact of explanatory variables on the response variable.