Data Analytics Assisgnment 1

Sabbir Dewan

2024-10-09

Contents

Let's load the appropriate library first

```
library(bootstrap)
```

Let's load the CNC router sample data

```
cnc = c(138.13, 138.14, 138.15, 138.12)
```

Fit the jackknife sampling on CNC dataset . I am setting $theta = sd$ as we are looking for the standard deviation of the data.

```
cnc.jk.samples <- jackknife( cnc , theta = sd , na.rm = TRUE)
```

Average the all sample standard deviation

```r
#mean(average) of all sample standard deviation
mean(cnc.jk.samples$jack.values)
```

```
## [1] 0.01263763
```

Let's check the true(all data) standard deviation

```r
sd(cnc)
```

```
## [1] 0.01290994
```

Let's check the Difference between original and jackknife

```r
mean(cnc.jk.samples$jack.values) - sd(cnc)
```

```
## [1] -0.0002723183
```

There' slight difference (-0.0002723) between sample standard deviation and jackknife standard deviation.

Let's look at jackknife standard error

```r
cnc.jk.samples$jack.se
```

```
## [1] 0.004568503
```

Jackknife standard error is 0.00456 Let's look at the bias of jackknife

```r
cnc.jk.samples$jack.bias
```

```
## [1] -0.000816955
```

Bias is -0.0008169

Is the Jackknife estimator Unbiased?

If the jackknife estimator is close to true standard deviation, then we can say it's unbiased or nearly unbiased. Also if the bias is small relative to standard error, the estimator is nearly unbiased. For our case, true standard deviation is 0.01291cm and jackknife standard deviation is 0.01263cm. Then we got bias -0.0002723 which is smaller than of standard error 0.004568.

So, considering minimal bias compared to the standard error, the jackknife estimator of CNC router is almost unbiased but with a minor bias. The minor bias indicates that the estimate is near to the true value but is not perfectly unbiased.

0.1 Question 2

Consider the variable AGE in the "Dating Profiles" dataset, which records the age of the OkCupid user.

(a) Use appropriate graphical displays and measures of centrality and dispersion to summaries the AGE variable. Provide a reasonable explanation for why the AGE data might have the distribution you observe.

Let's read the Dating Profile data first

```
dating_profile <- read.csv("C:\\University Study [MU]\\Semester 2\\ICT 513 - Data Analytics\\Datasets-2
```

To measure of centrality we can check the Mean, Median and Mode of Age data.

Let's check the data first to check if there's any null value or not. Null value (i.e - NA) can destroy the function, so we have to handle null value carefully.

```
sum(is.null(dating_profile$AGE))
```

```
## [1] 0
```

Great. We don't have any null value in AGE column. Let's find the mean first.

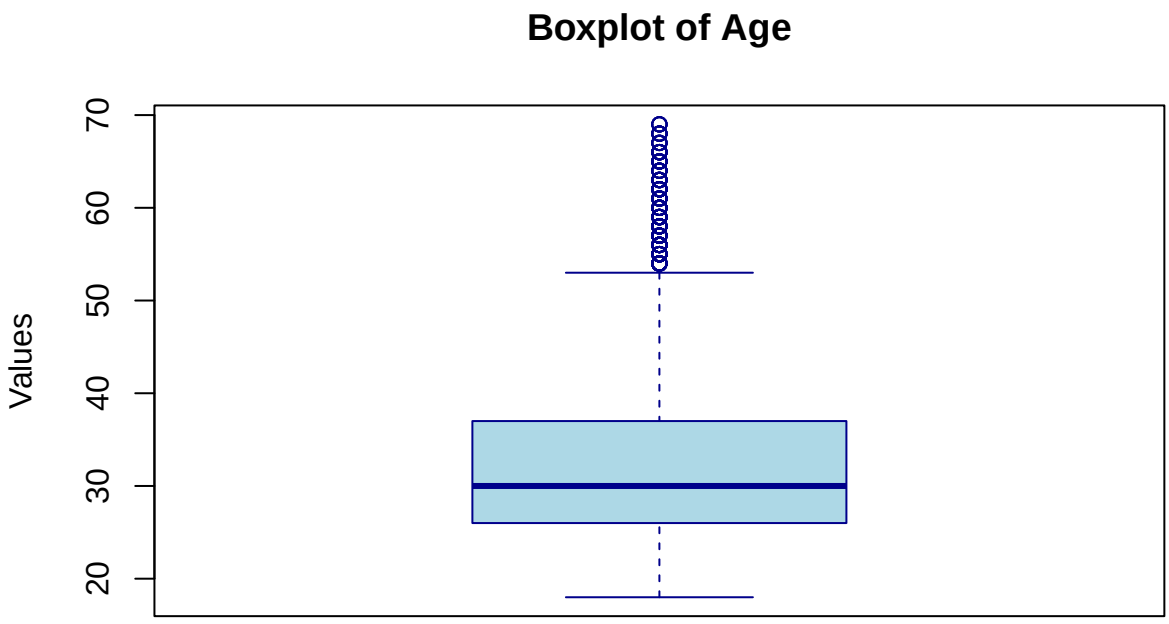
Mean:

```
mean_age = mean(dating_profile$AGE)
mean_age
```

```
## [1] 32.4444
```

The average Age is 32.45years. But we are not sure this average is misleading or not as Mean is highly sensitive by Outliers. Let's plot a box plot to find any outliers

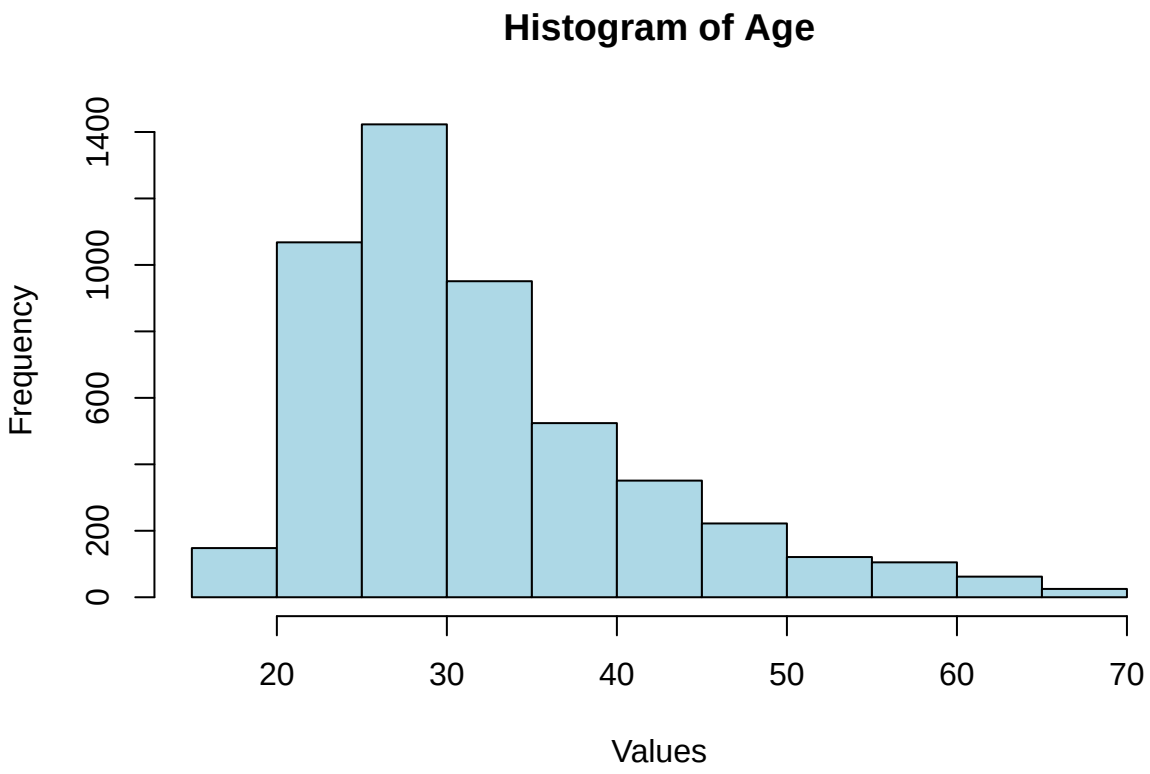
```
# Boxplot for Age Column
boxplot(dating_profile$AGE,
        main = "Boxplot of Age",
        ylab = "Values",
        col = "lightblue",
        border = "darkblue")
```



Box plot suggest there's some data points are extremely high. We can consider as outliers. So, Mean age are not supports a central point of age.

Let's try the histogram to understand the distribution of Age column.

```
# Basic histogram
hist(dating_profile$AGE,
     main = "Histogram of Age",
     xlab = "Values",
     col = "lightblue",
     border = "black")
```



The histogram represents a right-skewed distribution, indicating that most of the individuals are younger, with a small number of older individuals extending the distribution.This right-skewed distribution making the mean of Age less representative of where most of the data points lie.

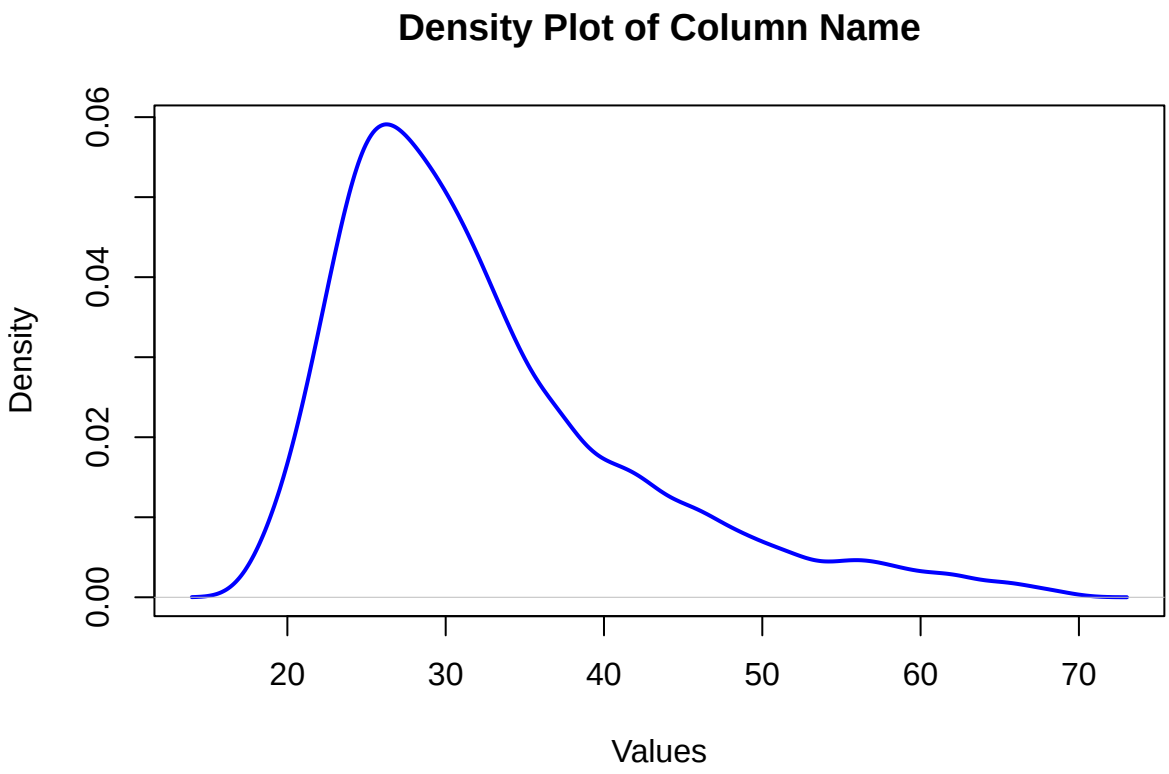
##Let's find the Median age

```
median_age = median(dating_profile$AGE)
median_age
```

```
## [1] 30
```

Median Age is 30 which is much more central as per our data. As Age has outliers and right-skewed distribution , median more accurate measure of central tendency in these situations.

```
# Basic density plot
plot(density(dating_profile$AGE),
     main = "Density Plot of Column Name",
     xlab = "Values",
     ylab = "Density",
     col = "blue",
     lwd = 2)
```



The density plot suggest that data has right tail. Meaning more larger values are on the right. Positively skewed.

```
mean_median_table <- data.frame(measures = c("Mean_age","Median_Age"), Value = c(mean_age, median_age ))
mean_median_table
```

```
##     measures   Value
## 1   Mean_age 32.4444
## 2 Median_Age 30.0000
```

##Let's check now the range and Inter Quartile range of Age

```
age_range <- range(dating_profile$AGE)
age_range
```

```
## [1] 18 69
```

```
range = max(dating_profile$AGE - min(dating_profile$AGE))
range
```

```
## [1] 51
```

The youngest person on this dataset is 18 and the oldest person age is 69.

0.2 Let's check the IQR

```
# Calculate quantiles
quantiles <- quantile(dating_profile$AGE, probs = c(0.25, 0.50, 0.75))

# Calculate IQR
iqr <- IQR(dating_profile$AGE)

quantiles
```

```
## 25% 50% 75%
##  26  30  37
```

```
iqr
```

```
## [1] 11
```

The age distribution analysis reveals the 25th percentile (Q1) is 26 years, the median is 30 years, and the 75th percentile (Q3) is 37 years, with an interquartile range(IQR) of 11 years. Ages range from 18 to 69 years, indicating a broad age range (i.e, 51) within the dataset.

#Let's calculate mean absolute deviation

```
# Calculate the absolute deviations from the mean
absolute_deviations <- abs(dating_profile$AGE - mean_age)

# Calculate the Mean Absolute Deviation (MAD)
mad <- mean(absolute_deviations)
mad
```

```
## [1] 7.364612
```

#Let's check the Variance and Standard Deviation

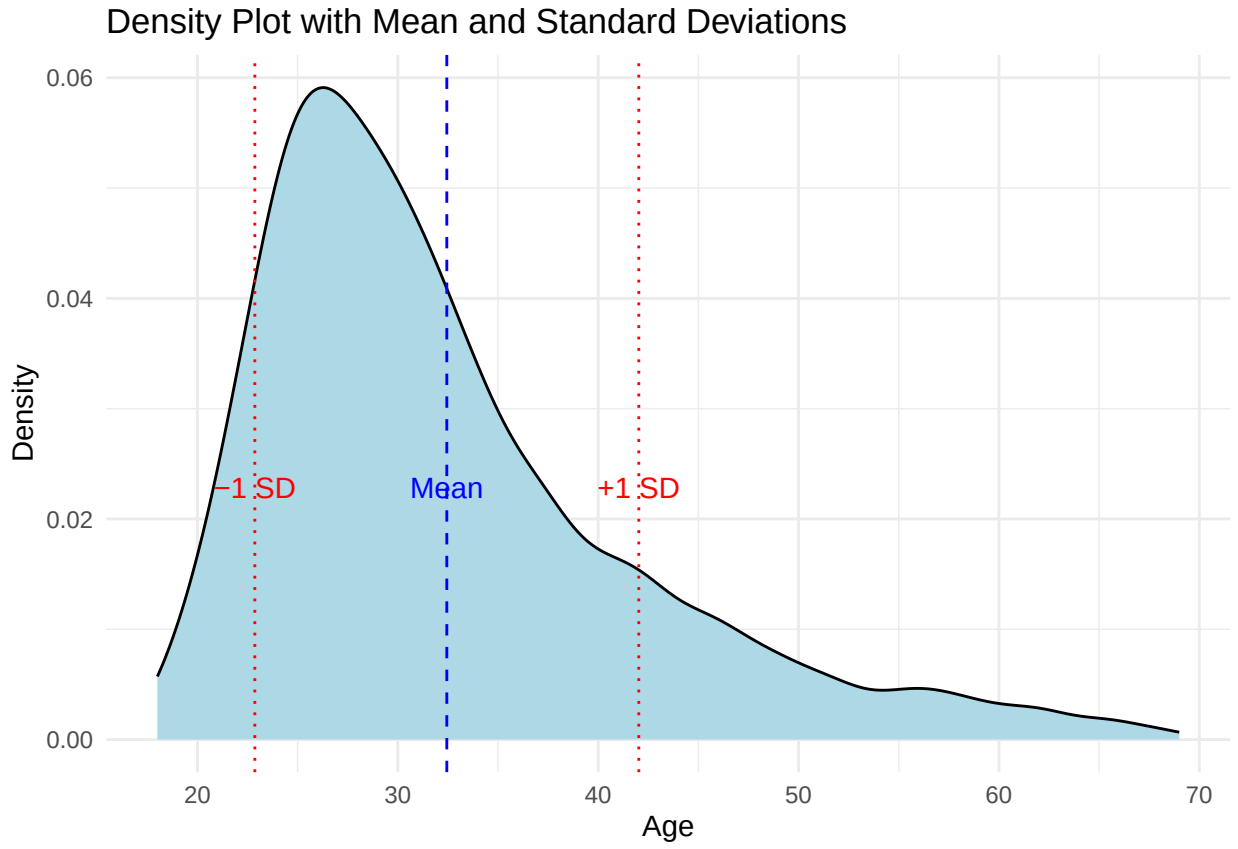
```
# Calculate variance
variance_age <- var(dating_profile$AGE)

# Calculate standard deviation
std_dev_age <- sd(dating_profile$AGE)

variance_age
```

```
## [1] 91.79687
```

```
std_dev_age
```

```
## [1] 9.581068
```

```r
# Load ggplot2 package
library(ggplot2)
# Create the density plot
ggplot(dating_profile, aes(x = AGE)) +
  geom_density(fill = "lightblue", color = "black") +
  geom_vline(aes(xintercept = mean_age), color = "blue", linetype = "dashed") +
  geom_vline(aes(xintercept = mean_age + std_dev_age), color = "red", linetype = "dotted") +
  geom_vline(aes(xintercept = mean_age - std_dev_age), color = "red", linetype = "dotted") +
  annotate("text", x = mean_age, y = 0.02, label = "Mean", color = "blue", vjust = -1) +
  annotate("text", x = mean_age + std_dev_age, y = 0.02, label = "+1 SD", color = "red", vjust = -1) +
  annotate("text", x = mean_age - std_dev_age, y = 0.02, label = "-1 SD", color = "red", vjust = -1) +
  labs(title = "Density Plot with Mean and Standard Deviations",
       x = "Age",
       y = "Density") +
  theme_minimal()
```



A mean age of 32.44 years with a standard deviation of 9.58 years, indicating a moderate spread around the mean. The Mean absolute deviation, at 7.36 years, supports this by showing the average deviation from the mean in a more robust manner as it's less effected by outliers. Overall, these measures suggest that while the average age is around 32 years, there is considerable variability in the ages, with most ages falling within roughly one standard deviation (i.e+/- 9.58 years) of the mean

0.3 QSN 2(b)

For the most appropriate measure of centrality and measure of dispersion you have selected for AGE, produce a table of the form shown below that presents:

1 the particular measures (i.e., statistics) you have chosen,

2 those measures (i.e., statistics) as calculated for the variable AGE,

3 the jackknife and bootstrap estimators for those statistics,

4 the jackknife and bootstrap standard errors for those statistics, and

5 the jackknife and bootstrap estimates of bias for those statistics.

#Do these measures of centrality and dispersion appear to be biased or unbiased estimators? (8 marks)

I am choosing Median and IQR. Median age is 30 and IQR is 11 years.

Let's calculate the jackknife and bootstrap estimators of Median Age.

```
library(bootstrap)
set.seed(0)
nbootstaps<-10000

#find jacknife estimator
dating.jk.samples <- jackknife( dating_profile$AGE , theta = median , na.rm = TRUE)

#mean(average) of all jackknife sample median
mean(dating.jk.samples$jack.values)
```

```
## [1] 30
```

```
#find bootstrap estimator
dating.bs.samples <- bootstrap(dating_profile$AGE,
                        nboot = nbootstaps,
                        theta = median,
                        na.rm = T)
```

```
#mean(average) of all bootstrap sample median
mean(dating.bs.samples$thetastar)
```

```
## [1] 29.9999
```

Standard Error and Bias of Jackknife

```
dating.jk.samples$jack.se
```

```
## [1] 0
```

```
dating.jk.samples$jack.bias
```

```
## [1] 0
```

Standard Error and Bias of Bootstrap

```
## bootstrap estimate of the standard error
sd(dating.bs.samples$thetastar)
```

```
## [1] 0.01
```

```
## bias (estimate - bootstrapped estimator)
median(dating_profile$AGE, na.rm = TRUE) - mean(dating.bs.samples$thetastar)
```

```
## [1] 1e-04
```

```
#find jackknife estimator for dispersion (IQR)
dating.jk.samples.iqr <- jackknife( dating_profile$AGE , theta = IQR , na.rm = TRUE)

#mean(average) of all jackknife sample IQR
mean(dating.jk.samples.iqr$jack.values)
```

```
## [1] 11
```

```
dating.jk.samples.iqr$jack.se
```

```
## [1] 0
```

```
dating.jk.samples.iqr$jack.bias
```

```
## [1] 0
```

```
#find bootstrap estimator for IQR
dating.jk.samples.iqr <- bootstrap(dating_profile$AGE,
                         nboot = nbootstaps,
                         theta = IQR,
                         na.rm = T)
```

```
#mean(average) of all bootstrap sample IQR
mean(dating.jk.samples.iqr$thetastar)
```

```
## [1] 11.00745
```

```
## bootstrap estimate of the standard error of IQR
IQR(dating.jk.samples.iqr$thetastar)
```

```
## [1] 0
```

```r
## bootstrap bias (estimate - bootstrapped estimator) of IQR
IQR(dating_profile$AGE, na.rm = TRUE) - mean(dating.jk.samples.iqr$thetastar)
```

```
## [1] -0.00745
```

Below is the table summary

```r
knitr::include_graphics("C:\\University Study [MU]\\Semester 2\\ICT 513 - Data Analytics\\Assignment 1\
```

Measure of Centrality

Measure of Centrality (Median ~=30)	**Jackknife**	**Bootstrap**
Estimator	30	≈ 30
Standard error	0	0.01
Bias	0	≈ 0

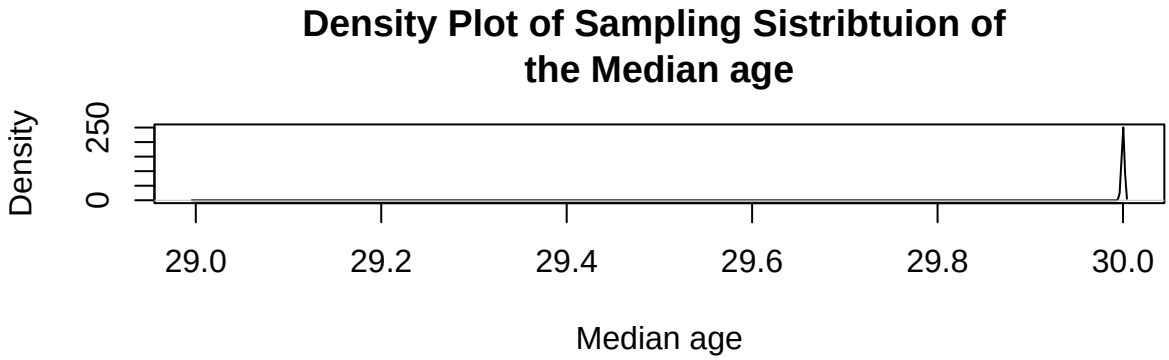
Measure of Dispersion

Measure of Centrality (IQR =11)	**Jackknife**	**Bootstrap**
Estimator	11	11.011
Standard error	0	0.425
Bias	0	-0.011

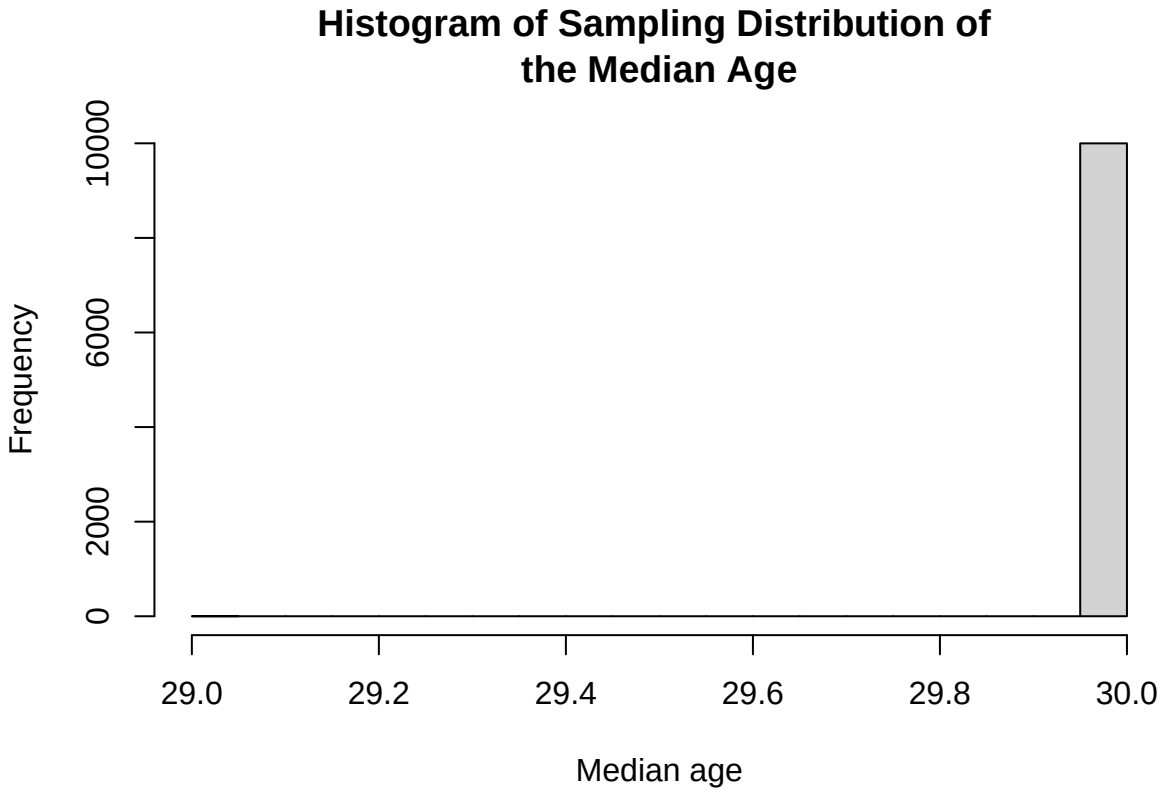
Biases for both the median and IQR are extremely small, age column essentially unbiased. The observed biases are so minor that they are unlikely to affect the validity of conclusions. Therefore, we can consider the age column as being unbiased based on the jackknife and bootstrap results.

Qsn 2(c) - Produce graphical displays of the sampling distributions of the measure of centrality and measure of dispersion you have selected for AGE. Comment on the shapes of these distributions. Additionally, produce a 95% bootstrap percentile confidence interval for both your measure of centrality and measure of dispersion and interpret them. If there is anything unusual about 95% bootstrap percentile confidence intervals, comment on that.

```r
## visual representation of the bootstrap samples
par(mfrow = c(2,1))
plot(density(dating.bs.samples$thetastar),
     xlab = 'Median age',
     main = 'Density Plot of Sampling Sistribtuion of \nthe Median age')
```




```
hist(dating.bs.samples$thetastar,
     freq = TRUE,
     xlab = 'Median age',
     main = 'Histogram of Sampling Distribution of \nthe Median Age')
```




```
## 95% confidence interval
quantile(dating.bs.samples$thetastar, probs=c(0.025, 0.975))
```

```
##  2.5% 97.5%
##    30    30
```

For the median, the distribution shows almost no variation (only 1 out of 1,000 bootstrap simulations gave a median different from the sample median). This resulted in a 95% bootstrap percentile confidence interval of (30, 30). However, this interval doesn't seem reasonable because both ends are the same, which suggests that we are 100% sure the true population median is 30.

```
## 95% confidence interval
quantile(dating.jk.samples.iqr$thetastar, probs=c(0.025, 0.975))
```

```
##  2.5% 97.5%
##    10    12
```

The distribution of the interquartile range is symmetric but clearly discrete, having only five unique values. It is also multi-modal and deviates from a normal distribution. Due to the symmetry in its sampling distribution, a bootstrap percentile confidence interval is suitable. A 95% bootstrap percentile confidence interval for the interquartile range is (10, 12), indicating that we are 95% confident the true interquartile range falls between 10 and 12 years.

6 Qsn 3 Now consider the relationship between average airfare (FARE) and average weekly number of passengers (PASSENGERS).

```
airfare = read.csv("C:\\University Study [MU]\\Semester 2\\ICT 513 - Data Analytics\\Datasets-20230713\\
```

Qsn 3(a) For a regression model based on these two variables, why would we most naturally consider (FARE) as the response variable?

Ans - FARE is a response variable or dependent variable, its value is anticipated or its variation is explained by the explanatory variable. Changes of fare can determine the number of Passengers. In practical, airline can not identify total number of passengers rather they can control fare which we are trying to predict in next solutions.

3 (b) Clearly and accurately state the linearity, independence, normality, and equal variances (i.e., homoscedasticity) assumptions of linear regression as they pertain to these data, and assess them for a linear model of FARE on PASSENGERS: FAREi = b0 + b1PASSENGERSi + i

This assessment should include reference to appropriate graphical displays.

Ans -

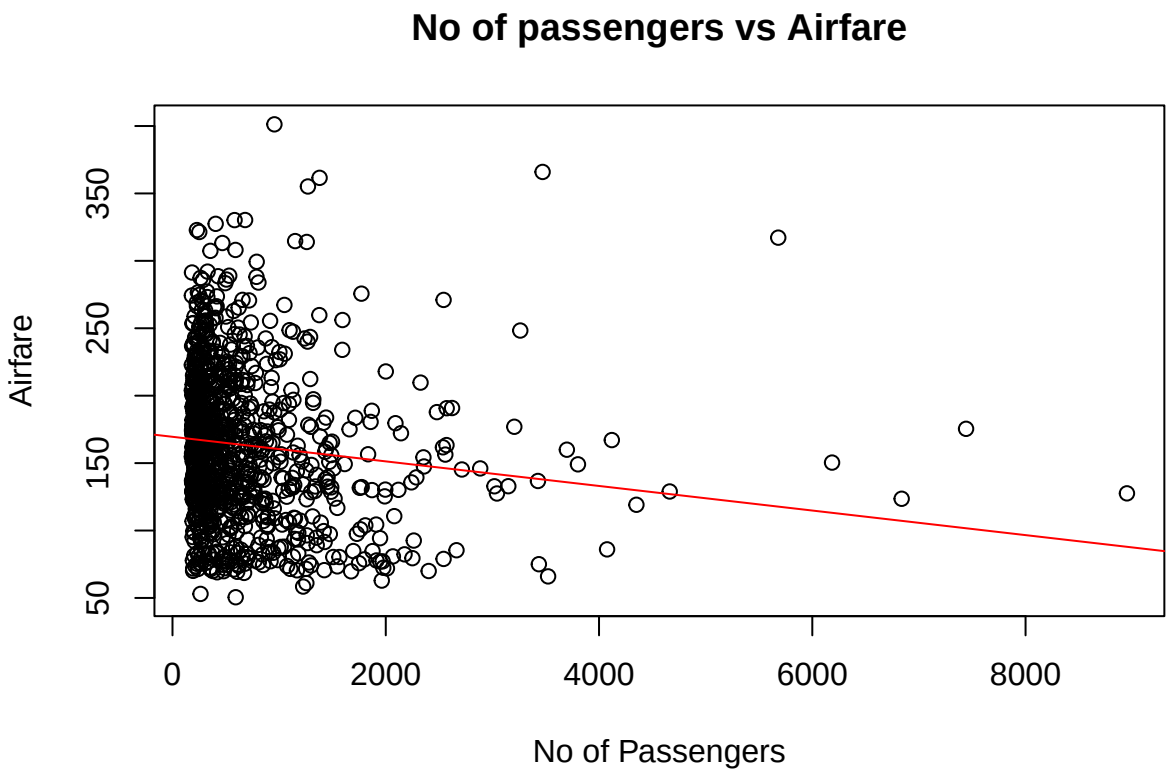
Let's fit the model first

```
#Let's fit the model first
airfare.model <- lm(FARE~PASSENGERS, data = airfare)
names(airfare.model)
```

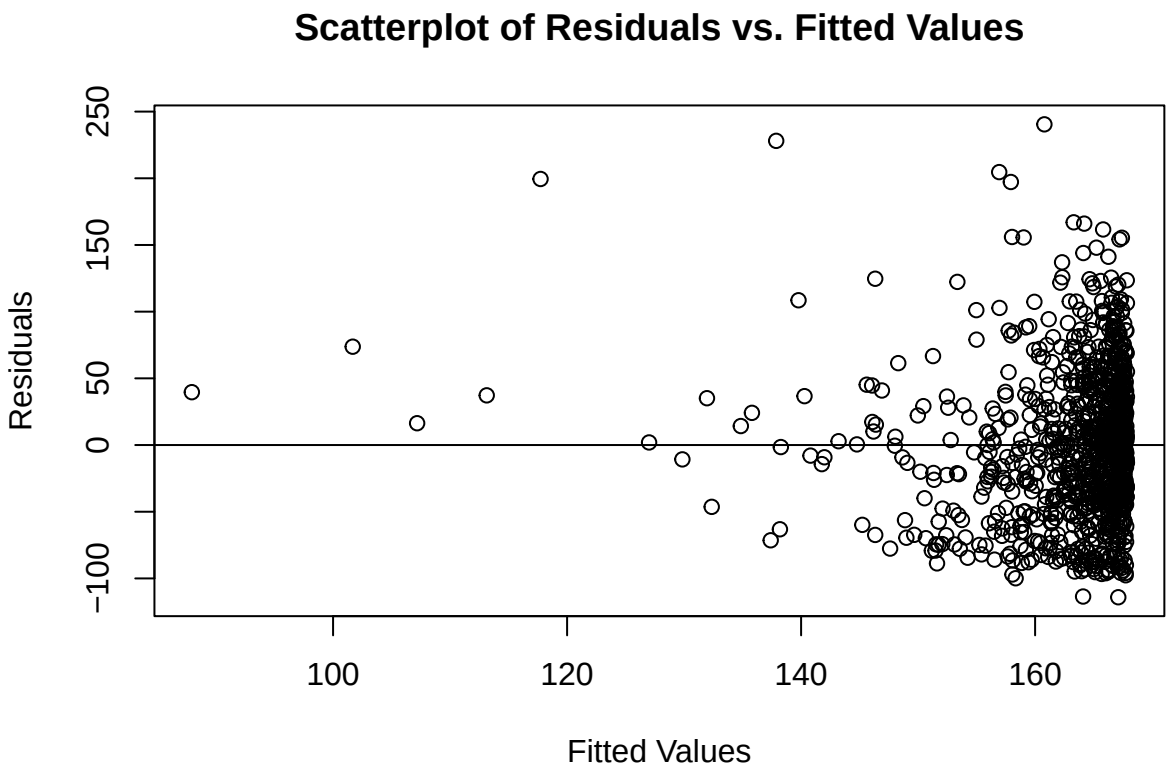
```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

Let's check the linearity first

```
# Let's check the linearity first
plot(airfare$PASSENGERS, airfare$FARE, xlab = "No of Passengers", ylab = "Airfare", main = "No of passe

abline(lm(FARE ~ PASSENGERS,  data = airfare), col = "red")
```




```
# Scatter plot of residuals vs. fitted values.
plot(airfare.model$fitted.values, airfare.model$residuals, xlab = "Fitted Values", ylab = "Residuals", m
abline(h = 0)
```

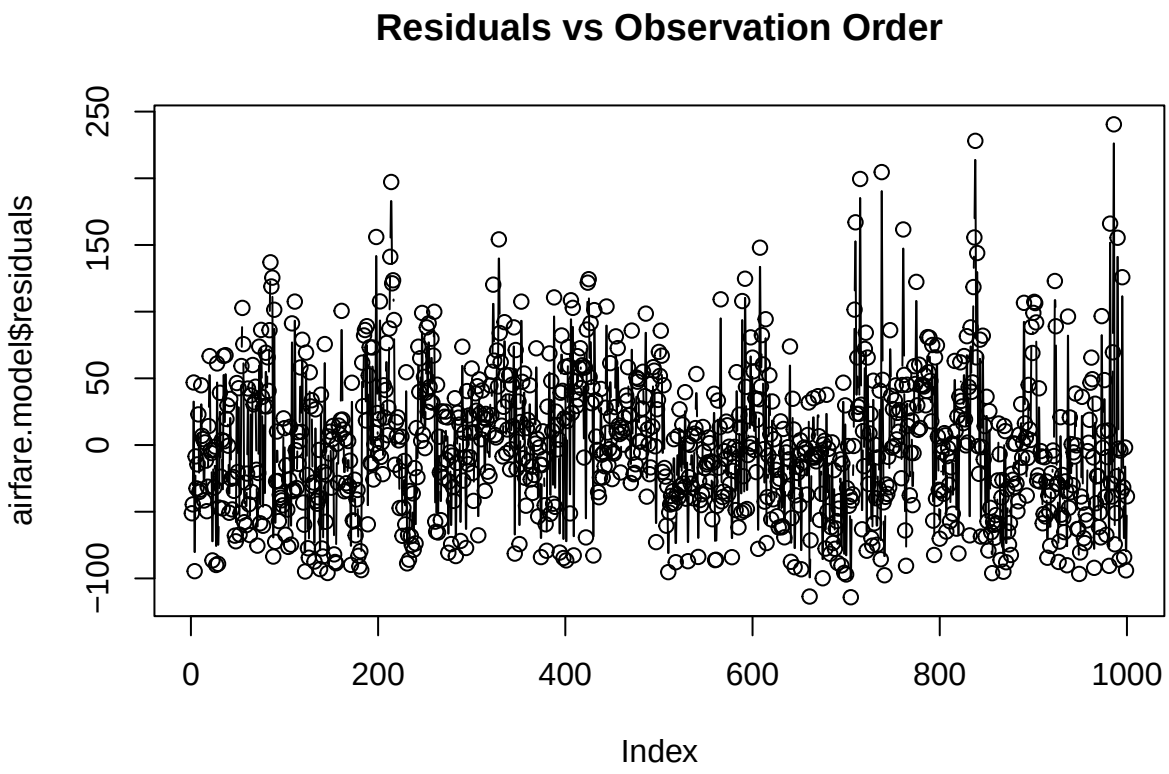


The residuals show a pattern that does not fully satisfy the linearity assumption. For the linearity assumption to hold, the residuals should be randomly dispersed about the horizontal line with a constant spread.

The plot clearly reveals a funnel-like shape, with the spread of residuals increasing as the fitted values grow.This shows that the connection between the variables is not strictly linear, or that there are concerns with nonlinearity or heteroscedasticity (unequal variance in residuals)

Let's check the Independence

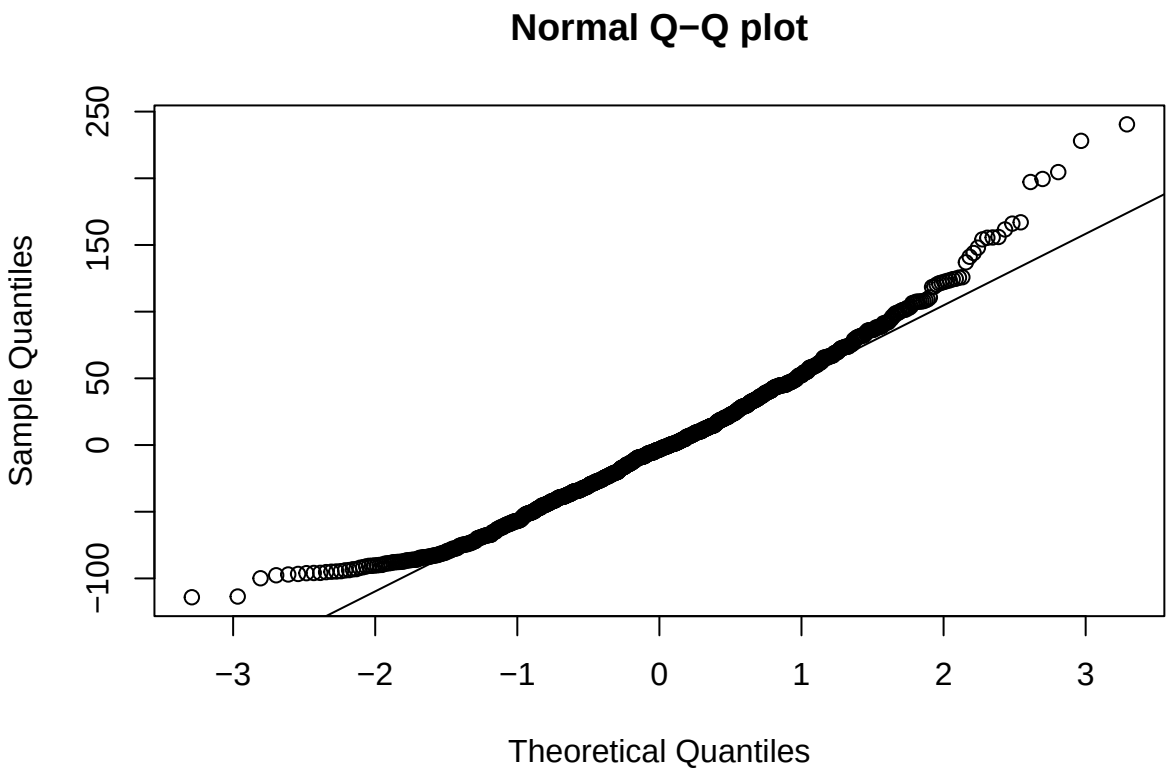
```
#Let's check the Independence
plot(airfare.model$residuals, type = "b", main = "Residuals vs Observation Order")
```



Plot has no strong evidence to support a significant violation of the independence of error term. The residuals appear to be randomly distributed with no visible patterns. Some slight grouping may occur, however this could simply be due to random noise.

Let's check the normality

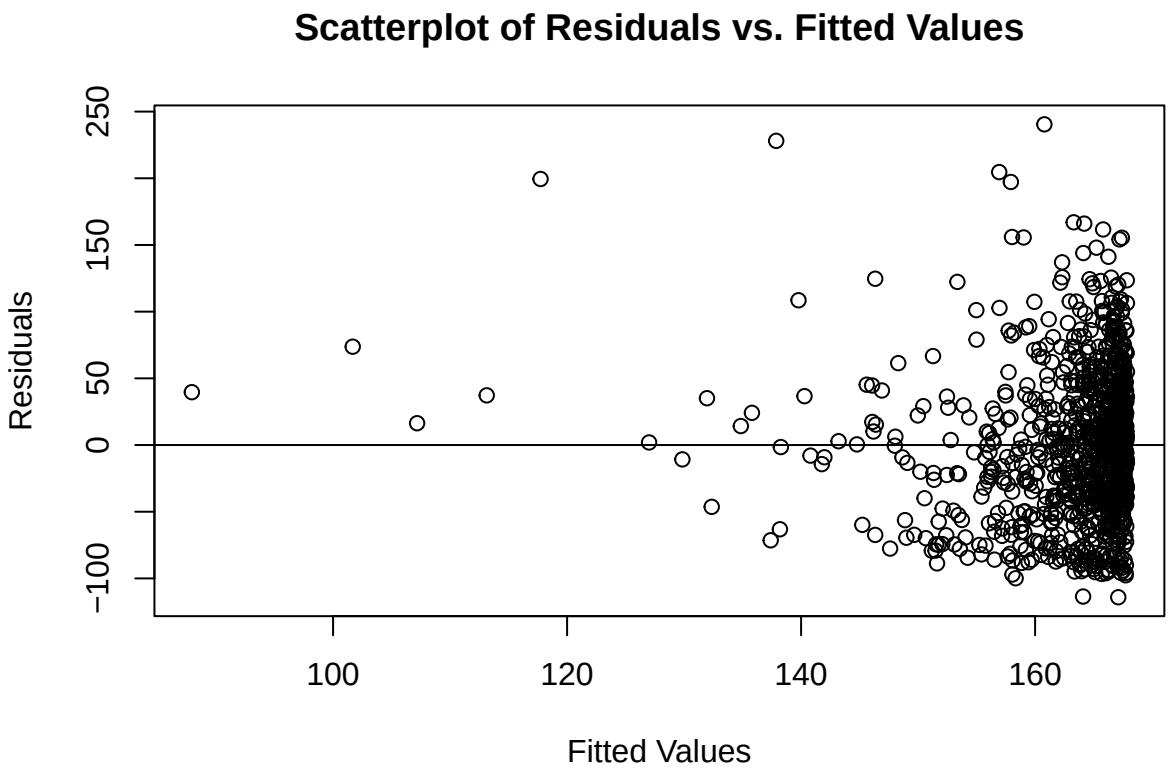
```
#Let's check the normality
#Normal Q-Q Plot for residuals
qqnorm(airfare.model$residuals, main = "Normal Q-Q plot")
qqline(airfare.model$residuals)
```



The Q-Q plot indicates a violation of the normality assumption. The residuals do not precisely follow the normal distribution, particularly at the extremes, indicating that the residual distribution may have larger tails(both lower and upper) or be skewed.

Let's check the equal variance (Homoscedasticity)

```
#Let's check the equal variance (Homoscedasticity)
# Scatter plot of residuals vs. fitted values.
plot(airfare.model$fitted.values, airfare.model$residuals, xlab = "Fitted Values", ylab = "Residuals", m
abline(h = 0)
```

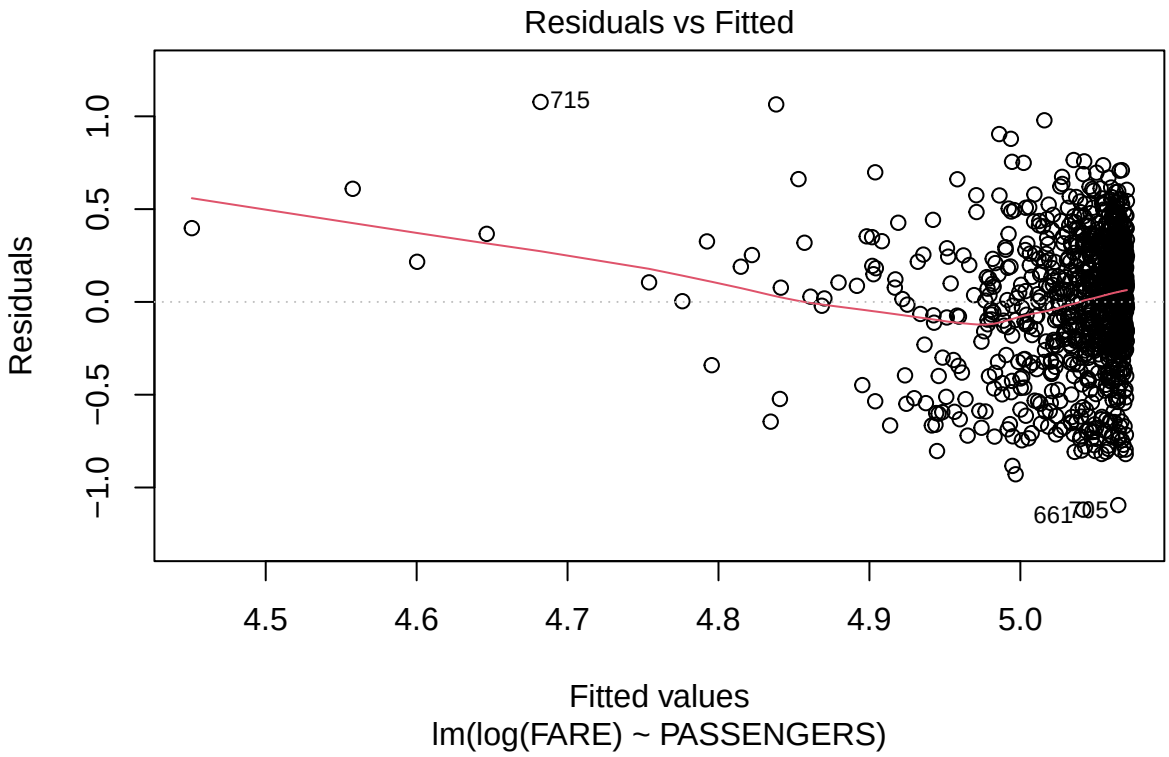


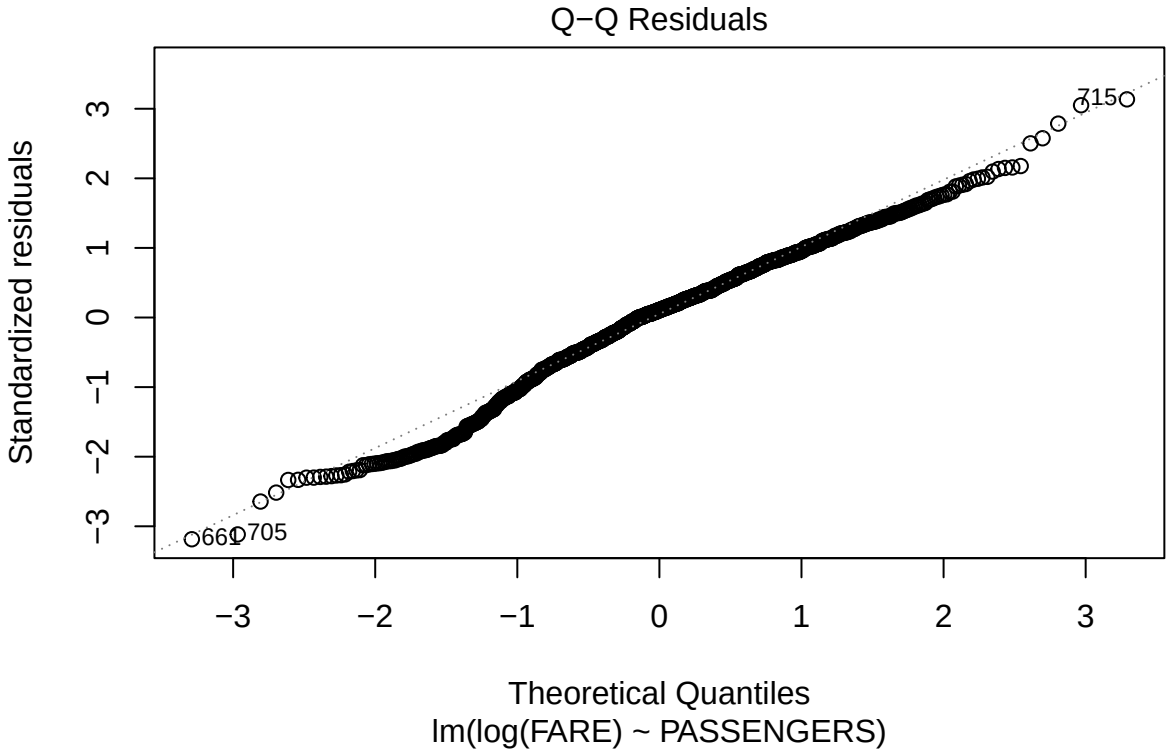
This plot shows a funnel-shaped pattern in residuals vs. fitted values plot which violates the homoscedasticity assumption. It strongly suggests that the condition where the variance of the residuals is not constant.

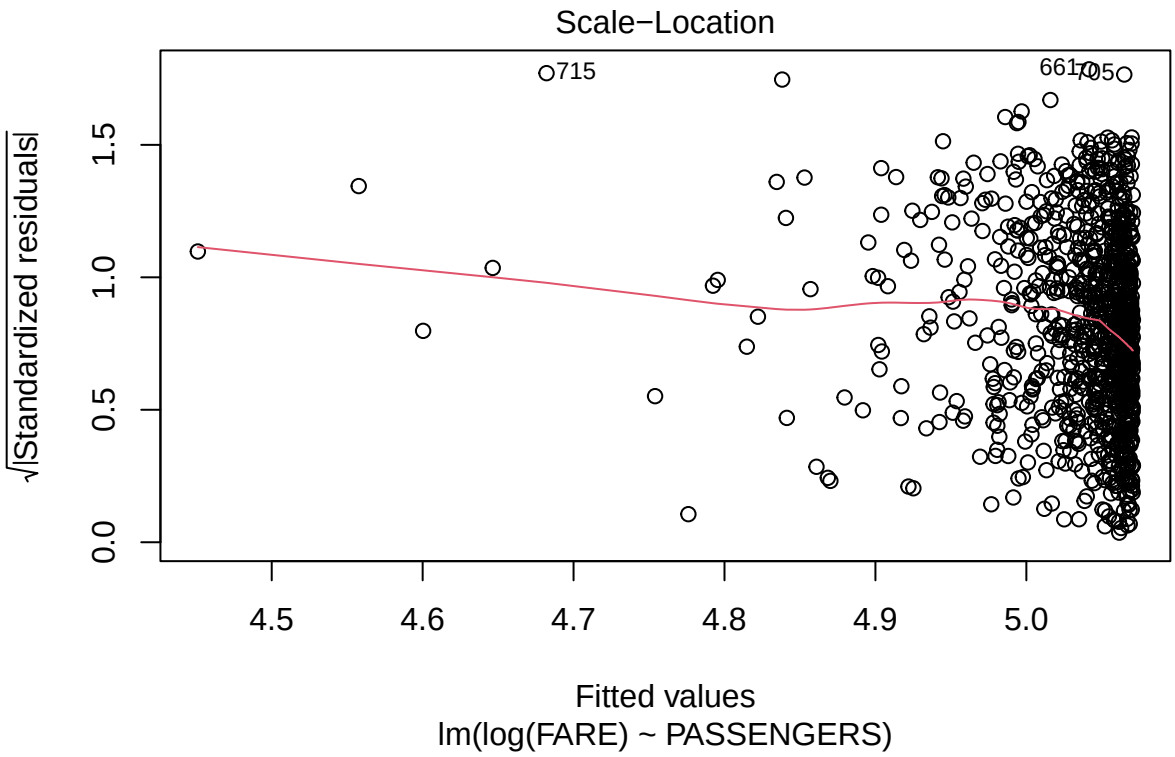
Qsn 3 (c) - Consider common transformations of the data and present the form of the linear model which you believe would be best when attempting to assess the relationship between FARE and PASSENGERS. Present and discuss relevant diagnostic plots for assessing the assumptions of linear regression for this model, clearly noting any violations of assumptions that may still exist

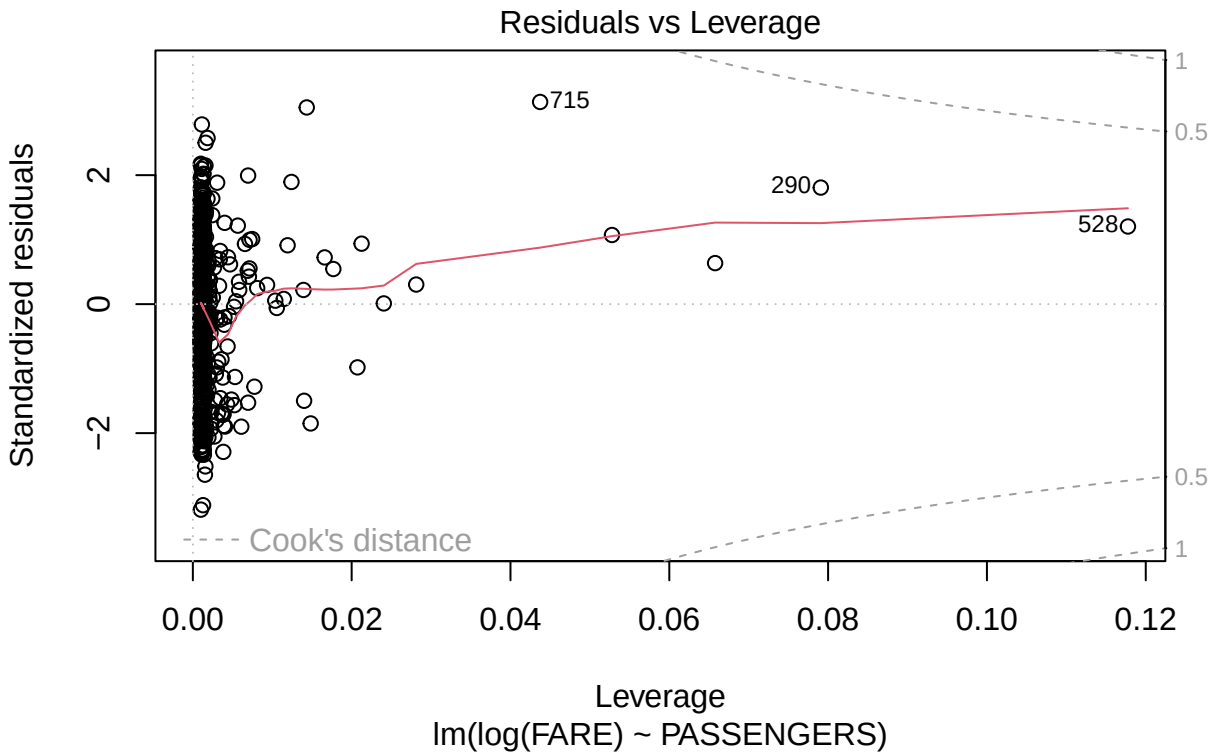
Let's apply the log transformation on the FARE variable to satisfy the normality in data

```
# Apply the logarithmic transformation to the Fare variable
log_fare_model <- lm(log(FARE) ~ PASSENGERS, data = airfare)

# Check the new residual plots to assess model assumptions
plot(log_fare_model)
```


Residuals vs Fitted
715
661
705
Residuals
1.0
0.5
0.0
−0.5
−1.0
4.5
4.6
4.7
4.8
4.9
5.0
Fitted values
lm(log(FARE) ~ PASSENGERS)


Q−Q Residuals
Standardized residuals
−3
−2
−1
0
1
2
3
715
661
705
Theoretical Quantiles
lm(log(FARE) ~ PASSENGERS)


Scale−Location
715
661
705
√|Standardized residuals|
0.0
0.5
1.0
1.5
4.5
4.6
4.7
4.8
4.9
5.0
Fitted values
lm(log(FARE) ~ PASSENGERS)

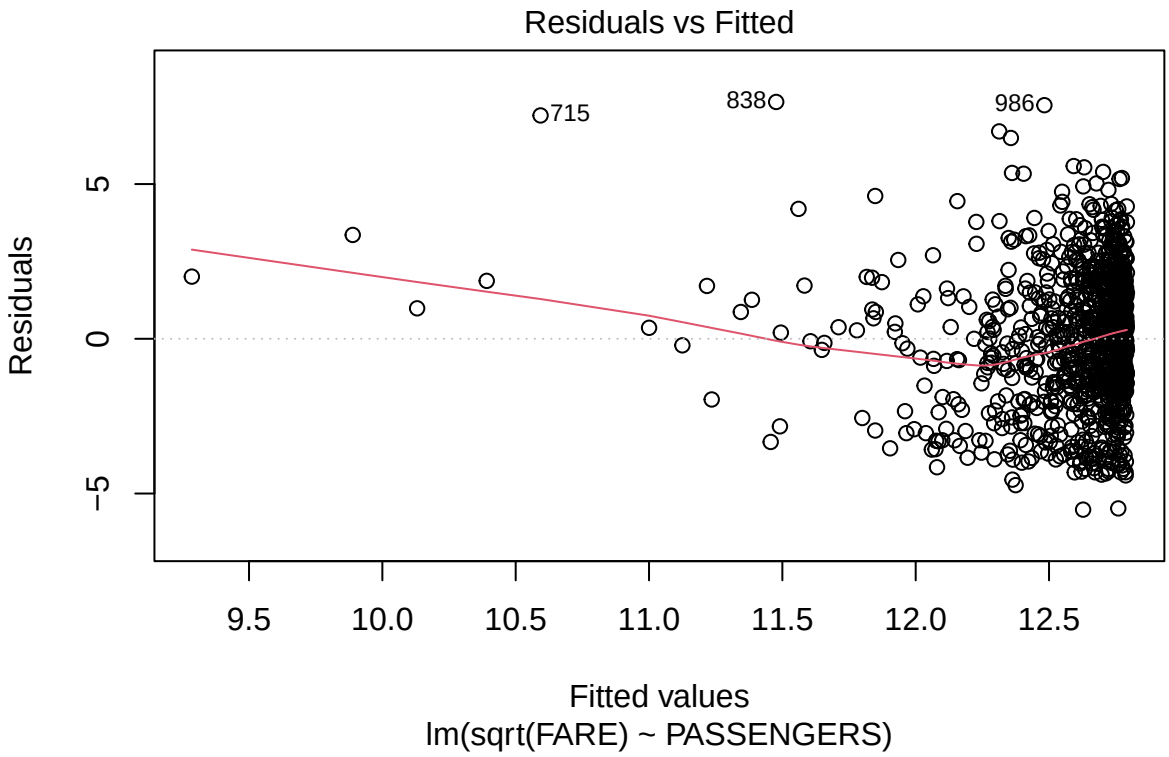


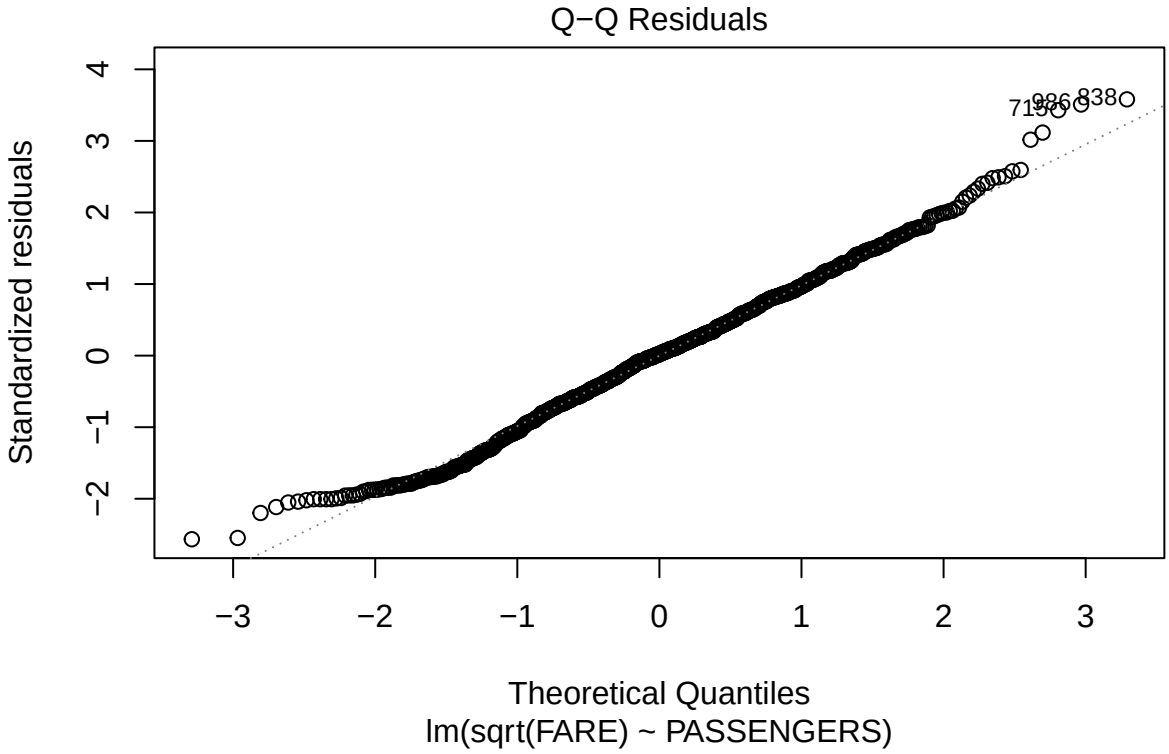
After log transform on FARE variable, no significant change on linearity assumption(1st image) and normality assumption(2nd picture) and doesn't satisfy the assumption. Still The plot clearly reveals a funnel-like shape also. But Q-Q plot showing slight improvement on the top value but still remain the fluctuation on the tail. It's not satisfying the normal distribution 100% but sightly better from our sample original dataset.

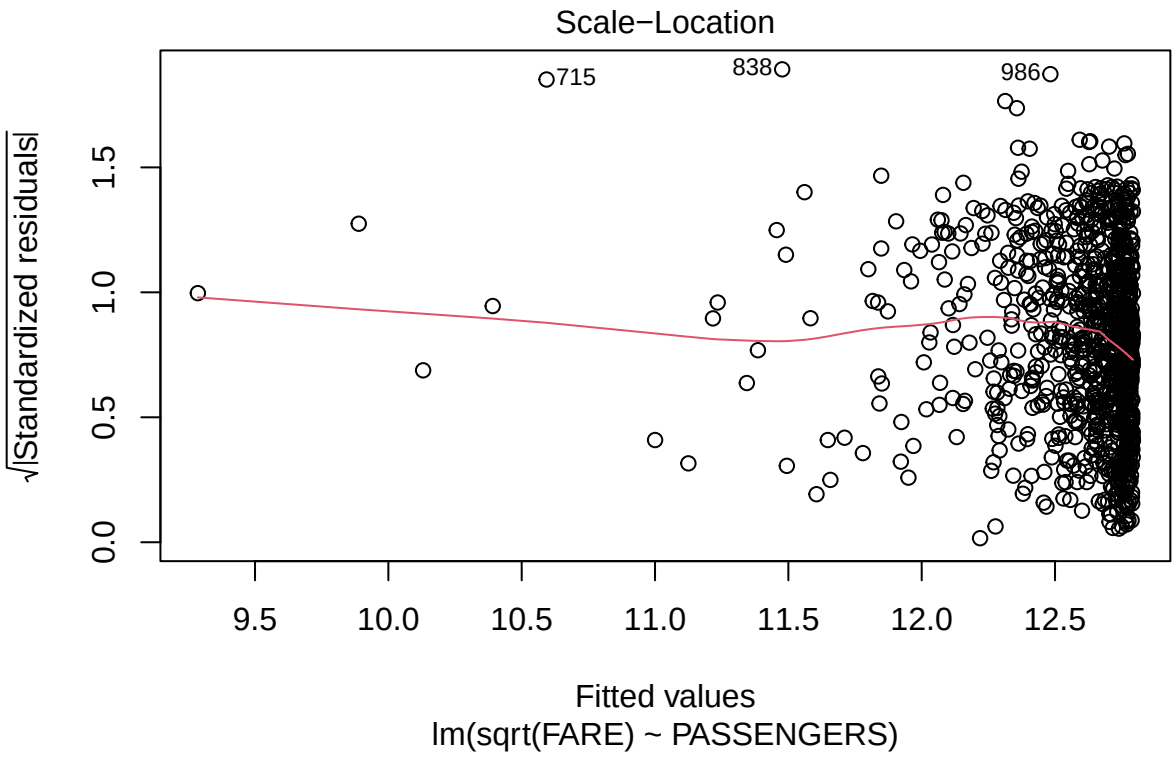
The assumption of Homoscedasticity(picture 3) and Independence (Picture 4) violate as still have the funnel shape on the data.

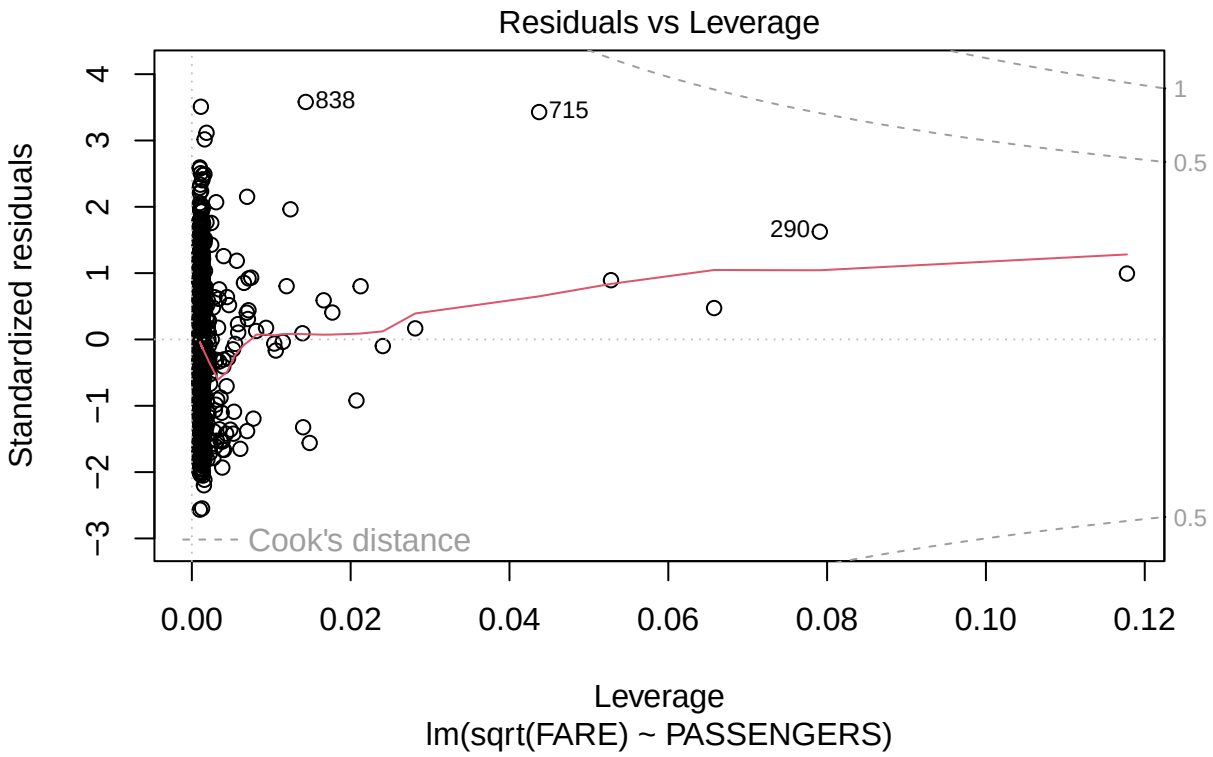
7 Let's try the square-root transformation on Fare variable

```
# Apply the Square root transformation to the Fare variable
sqrt_fare_model <- lm(sqrt(FARE) ~ PASSENGERS, data = airfare)

# Check the new residual plots to assess model assumptions
plot(sqrt_fare_model)
```


Residuals vs Fitted
715
838
986
Residuals
5
0
−5
9.5
10.0
10.5
11.0
11.5
12.0
12.5
Fitted values
lm(sqrt(FARE) ~ PASSENGERS)


Q−Q Residuals
Standardized residuals
−2
−1
0
1
2
3
4
838
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(sqrt(FARE) ~ PASSENGERS)


Scale−Location
715
838
986
√|Standardized residuals|
0.0
0.5
1.0
1.5
9.5
10.0
10.5
11.0
11.5
12.0
12.5
Fitted values
lm(sqrt(FARE) ~ PASSENGERS)



After square root transform on FARE variable, no significant change on linearity assumption(1st image) and normality assumption(2nd picture) and doesn't satisfy the assumption. Still The plot clearly reveals a funnel-like shape also. But Q-Q plot showing slight improvement comparing to log transformation but still remain the fluctuation on the tail. It's not satisfying the normal distribution 100% but sightly better from log transformation.

The assumption of Homoscedasticity(picture 3) and Independence (Picture 4) violate as still have the funnel shape on the data.

#Qsn 3 (d) Assuming that the model presented in Part (b) is wholly appropriate (i.e., there are no #violations of the assumptions of linear regression), provide a table of relevant R output #for that model and comment on whether there is a significant "effect" of average weekly #number of passengers on average airfare. If so, interpret this "effect

```
#model fit summary
summary(airfare.model)
```

```
##
## Call:
## lm(formula = FARE ~ PASSENGERS, data = airfare)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -114.118  -38.649   -3.235   33.680  240.437
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 169.502290   2.311839  73.319  < 2e-16 ***
## PASSENGERS   -0.009114   0.002268  -4.018 6.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.95 on 998 degrees of freedom
## Multiple R-squared:  0.01592,    Adjusted R-squared:  0.01493
## F-statistic: 16.15 on 1 and 998 DF,  p-value: 6.307e-05
```

The equation of the linear regression model is:

$$FARE = 169.502290 - 0.009114 \times PASSENGERS$$

Passenger is a an important feature as p value suggests $p < 0.05$).We can conclude, A one-unit increase in the average weekly number of passengers causes a decrease of approximately 0.0091 units in the average airfare.

Let's check the 95% confidence interval

```
confint(airfare.model, level = 0.95)
```

```
##                    2.5 %       97.5 %
## (Intercept) 164.96566712 174.038913638
## PASSENGERS   -0.01356441  -0.004662752
```

With 95% confidence,we can say the mean airfare decreases by between 0.0047 and 0.0136 units for each additional passenger.