Question 1

Test if there is a significant association between the passenger's rating and the passenger's gender for the rides taken on weekends. What does being associated mean in this context? Interpret your findings.

At first we read the CSV file and store it in a variable named "rides".

```
rides <- read.csv("Rides.csv")
head(rides)
```

```
##   TripId DriverAge DriverGender PassengerGender DriverRating PassengerRating
## 1      1        18         Male          Female            4               4
## 2      2        56   Non-binary          Female            5               4
## 3      3        29       Female            Male            3               3
## 4      4        51         Male          Female            3               4
## 5      5        47       Female            Male            4               4
## 6      6        63       Female      Non-binary            4               3
##      PickupLoc    DropoffLoc  Fare TripDist Duration Weather VehicleType
## 1        Manly Circular Quay  59.2     20.5     34.7   Clear         SUV
## 2        Manly       Central  82.3     21.1     32.7   Rainy         SUV
## 3      Newtown    Paddington  64.8     20.0     26.0   Rainy       Sedan
## 4   Parramatta   Kings Cross 102.4     32.6     39.1   Sunny         Van
## 5 Darlinghurst      Cronulla  83.6     29.9     47.5   Clear       Sedan
## 6   Parramatta Circular Quay  61.7     22.0     25.4   Rainy       Sedan
##      PickupTime Tip DayofWeek
## 1        Midday   2   Tuesday
## 2 Early Morning   3   Tuesday
## 3        Midday   6  Saturday
## 4       Evening   7   Tuesday
## 5     Afternoon   2    Friday
## 6 Early Morning   4    Sunday
```

Then we create a new data frame named "weekend_rides" where we filter the rows where the day of the week is either Saturday or Sunday and select the passenger gender and passenger rating column only for analysis.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
weekend_rides <- rides %>%
  filter(DayofWeek == "Saturday" | DayofWeek == "Sunday") %>%
  select(PassengerGender, PassengerRating)
head(weekend_rides)
```

```
##   PassengerGender PassengerRating
## 1            Male               3
```

```
## 2      Non-binary               3
## 3          Female               3
## 4            Male               4
## 5          Female               3
## 6      Non-binary               4
```

Now we check the size of the "weekend_rides" data set.

```
nrow(weekend_rides)
```

```
## [1] 80
```

For a better understanding we create a table where we can understand the number of ratings from each gender.

```
table(weekend_rides$PassengerGender)
```

```
##
##     Female       Male Non-binary
##         27         49          4
```

Hypothesis:

H0: There is no significant association between the passenger's rating and gender for rides taken on weekends.

HA: There is a significant association between the passenger's rating and gender for rides taken on weekends.

Now we perform a chi-squared test on our observed data set.

```
obs_chisq <- chisq.test(table(
  weekend_rides$PassengerGender,
  weekend_rides$PassengerRating
))
obs_chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(weekend_rides$PassengerGender, weekend_rides$PassengerRating)
## X-squared = 7.8572, df = 6, p-value = 0.2488
```

We get a p-value of 0.2488 which is very high.

Since the data set is small and we cannot confirm if the data is normally distributed, we have to do simulation.

```
perm_chisq <- replicate(1000, {
  perm_rating <- sample(weekend_rides$PassengerRating)
  chisq.test(table(weekend_rides$PassengerGender,
                   perm_rating))$statistic
})
```

```
p_val <- sum(perm_chisq >= obs_chisq$statistic) / 1000
p_val
```

```
## [1] 0.265
```

Conclusion

After the simulation we get a p-value of 0.265 which is higher than our threshold (0.05). We can conclude that we do not have enough evidence to reject the null hypothesis i.e. we do not have enough evidence which shows that there is a significant association between the passenger's rating and gender for rides taken on weekends.

In this context associated means a potential relationship between the passenger's gender and their rating for the rides.

Question 2

Test whether the mean of tip that were held on Thursday's for male drivers are greater than female drivers.

Hypothesis:

H0: Mean of tip that were held on Thursday's for male drivers are equal to female drivers.

HA: Mean of tip that were held on Thursday's for male drivers are greater than female drivers.

At first we create a data frame named "thursday_rides" where we filter the rows where the day of the week is Thursday and the gender is either male or female and select the gender of the driver and tip only.

```r
library(dplyr)
thursday_rides <- rides %>%
  filter(DayofWeek == "Thursday") %>%
  filter(DriverGender == "Male" | DriverGender == "Female") %>%
  select(DriverGender, Tip)
head(thursday_rides)
```

```
##   DriverGender Tip
## 1         Male   3
## 2       Female   5
## 3       Female  10
## 4       Female   8
## 5         Male   3
## 6         Male  10
```

Now we check the size of the "thursday_rides" data set.

```r
nrow(thursday_rides)
```

```
## [1] 35
```

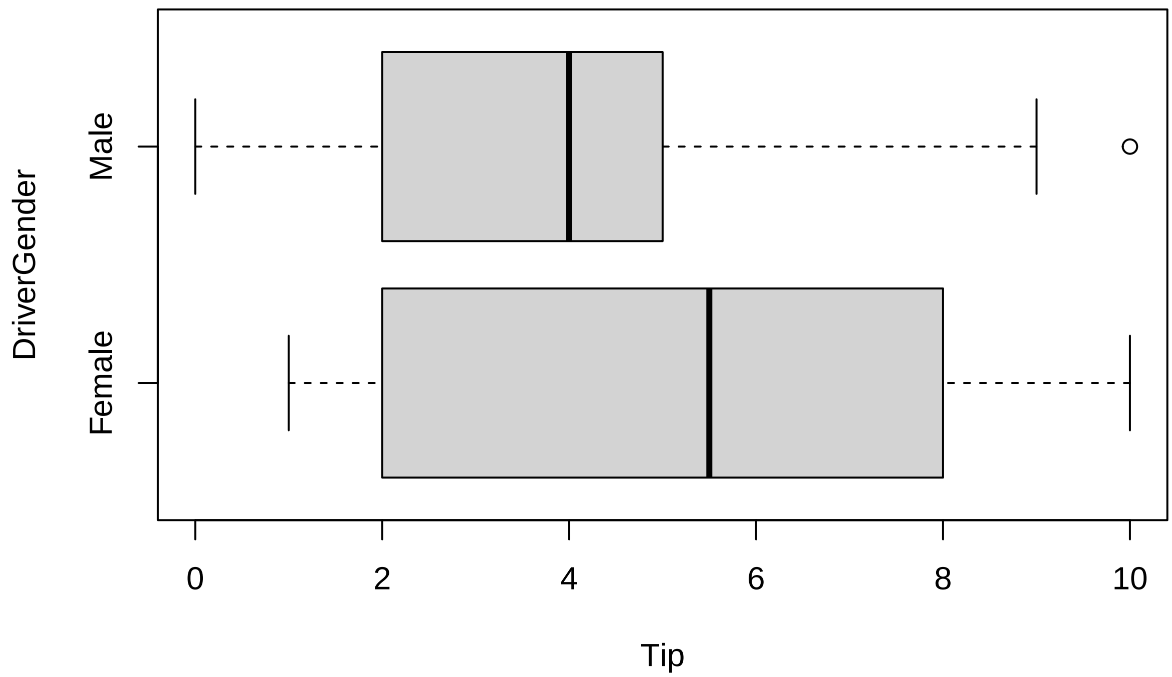
For a better understanding of the data frame we create a table summarizing the total number of male and female.

```r
table(thursday_rides$DriverGender)
```

```
##
## Female   Male
##     14     21
```

We now demonstrate the data in a box plot to understand the mean, the spread and the outliers of male and female.

```r
boxplot(Tip~DriverGender, thursday_rides, horizontal = TRUE)
```



From the box plot above we can observe that the mean tip of male is lower than that of female. To further confirm our findings, we can simulate our data and find the p-value.

At first we find the difference of mean between the genders.

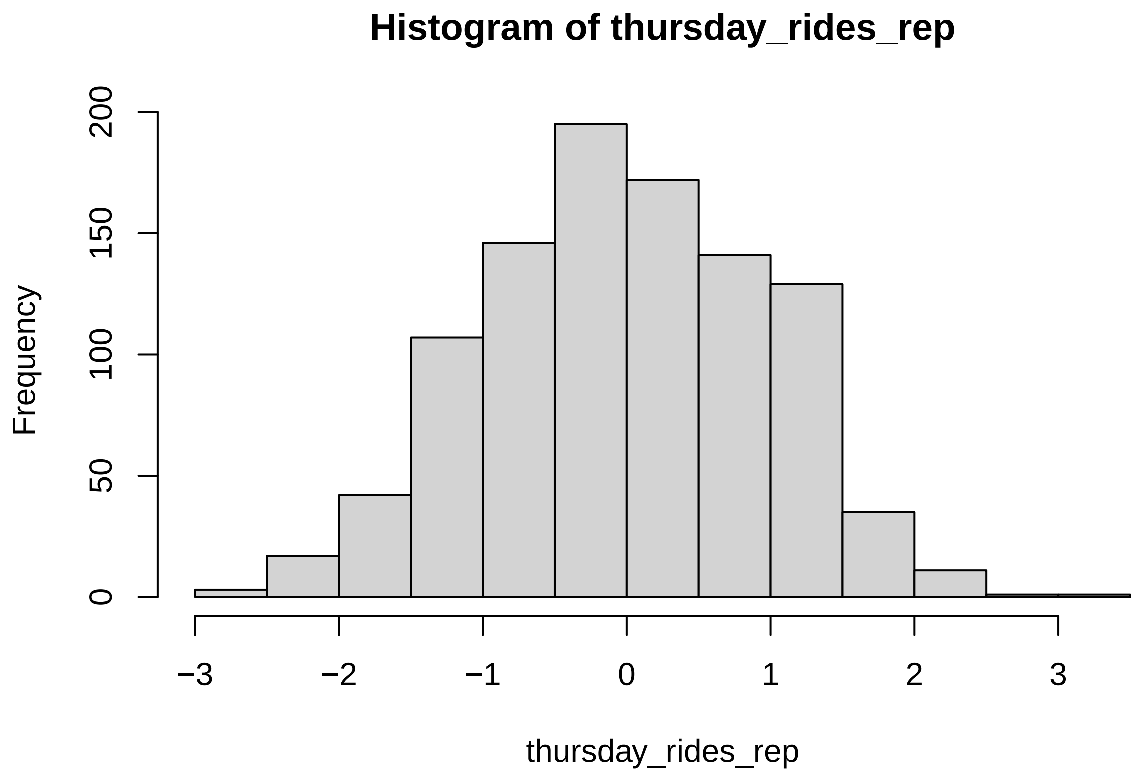
```
m <- mean(thursday_rides$Tip[thursday_rides$DriverGender == "Male"])
f <- mean(thursday_rides$Tip[thursday_rides$DriverGender == "Female"])
dif <- m - f
dif
```

```
## [1] -1.261905
```

The difference of mean is -1.2619048.

Now we simulate our data set 1000 times.

```
thursday_rides_rep <- replicate(1000, {
  DriverGender.sim <- sample(thursday_rides$DriverGender)
  - diff(aggregate(Tip ~ DriverGender.sim, thursday_rides, mean)$Tip)
})
hist(thursday_rides_rep)
```




```
pVal <- mean(thursday_rides_rep > dif)
pVal
```

```
## [1] 0.918
```

Conclusion

After the simulation we get a p-value of 0.918 which is higher than our threshold (0.05). So we can conclude that we do not have enough evidence to reject the null hypothesis i.e. we do not have enough evidence to say that the mean of tip that were held on Thursday's for male drivers are greater than female drivers.

Question 3

Compute the 98% confidence interval for the difference in the mean fare charged for rides starting from Olympic Park versus those starting from Circular Quay.

- First, use bootstrapping to compute the confidence interval.
- Then approximate the confidence interval based on a t-distribution.
- How do the results compare? Justify your answer.

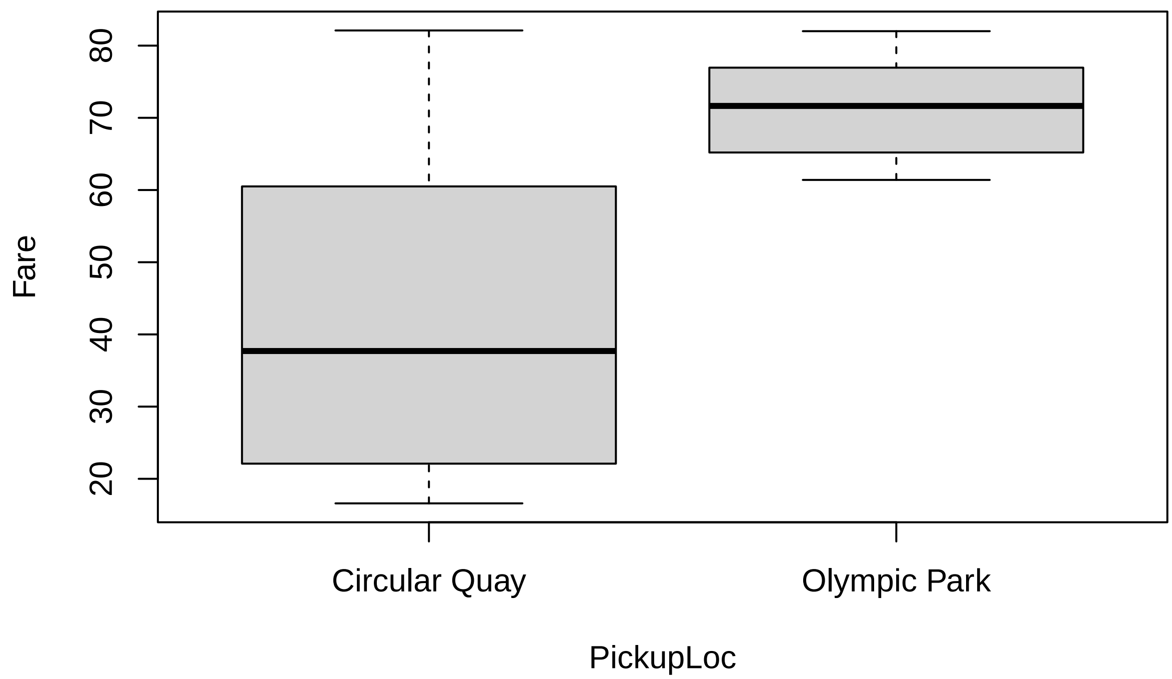
To compute the 98% confidence interval for the difference in the mean fare charged for rides starting from Olympic Park versus those starting from Circular Quay at first we create a data frame. In the data frame we filter the rows where the rides start from Olympic Park or Circular Quay and then we select the pickup location and fare for the rides.

```
olympic_circular_rides <- rides %>%
  filter(PickupLoc == "Olympic Park" |
```

```
           PickupLoc == "Circular Quay") %>%
  select(PickupLoc, Fare)
head(olympic_circular_rides)
```

```
##       PickupLoc Fare
## 1 Circular Quay 77.5
## 2 Circular Quay 37.7
## 3 Circular Quay 22.1
## 4 Circular Quay 44.3
## 5 Circular Quay 31.9
## 6 Circular Quay 25.9
```

```
boxplot(Fare~PickupLoc, olympic_circular_rides)
```



We can observe that the variances are not equal.

Now we check the size of the "olympic_circular_rides" data set.

```
nrow(olympic_circular_rides)
```

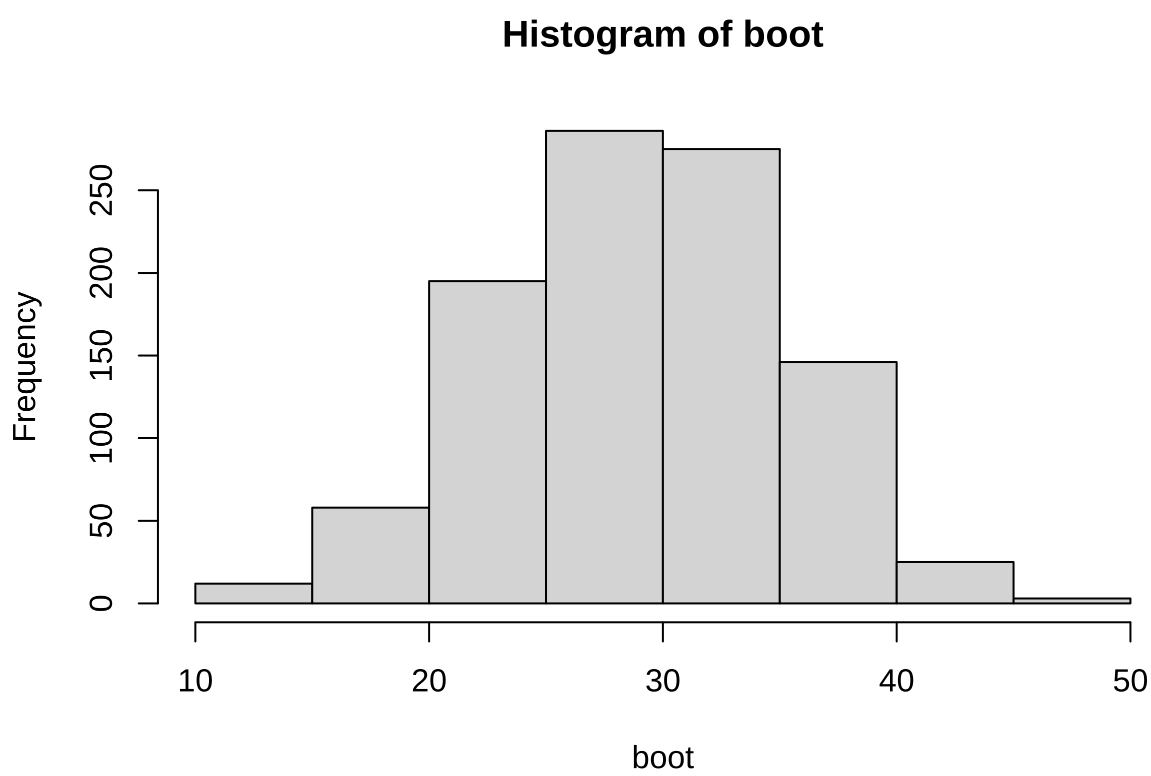
```
## [1] 25
```

Since the "olympic_circular_rides" data set size is small (less than 30), we calculate the confidence interval from bootstrapping.

```
olympic <-
  olympic_circular_rides$Fare[olympic_circular_rides$PickupLoc == "Olympic Park"]
circular <-
  olympic_circular_rides$Fare[olympic_circular_rides$PickupLoc == "Circular Quay"]
```

```r
boot <- replicate(1000, {
  sc <- sample(olympic, replace = TRUE)
  sw <- sample(circular, replace = TRUE)
  mean(sc) - mean(sw)
})

hist(boot)
```



From the histogram above we can observe that the average difference in the mean fare is close to 30 where the spread is between 10 and 45. To further investigate, we will calculate the 98% confidence interval for the difference.

```r
CI <- quantile(boot, c(0.01, 0.99))
CI
```

```
##       1%      99%
## 14.38882 42.11051
```

```r
t.test(olympic, circular,
       alternative = "two.sided",
       paired = FALSE,
       var.equal = FALSE,
       conf.level = 0.98)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  olympic and circular
## t = 4.7056, df = 21.343, p-value = 0.0001158
## alternative hypothesis: true difference in means is not equal to 0
## 98 percent confidence interval:
##  13.56940 44.71001
## sample estimates:
## mean of x mean of y
##  71.37500  42.23529
```

Conclusion

We are 98% confident that the true population difference is between 14.3888162 and 42.1105074.

Question 4

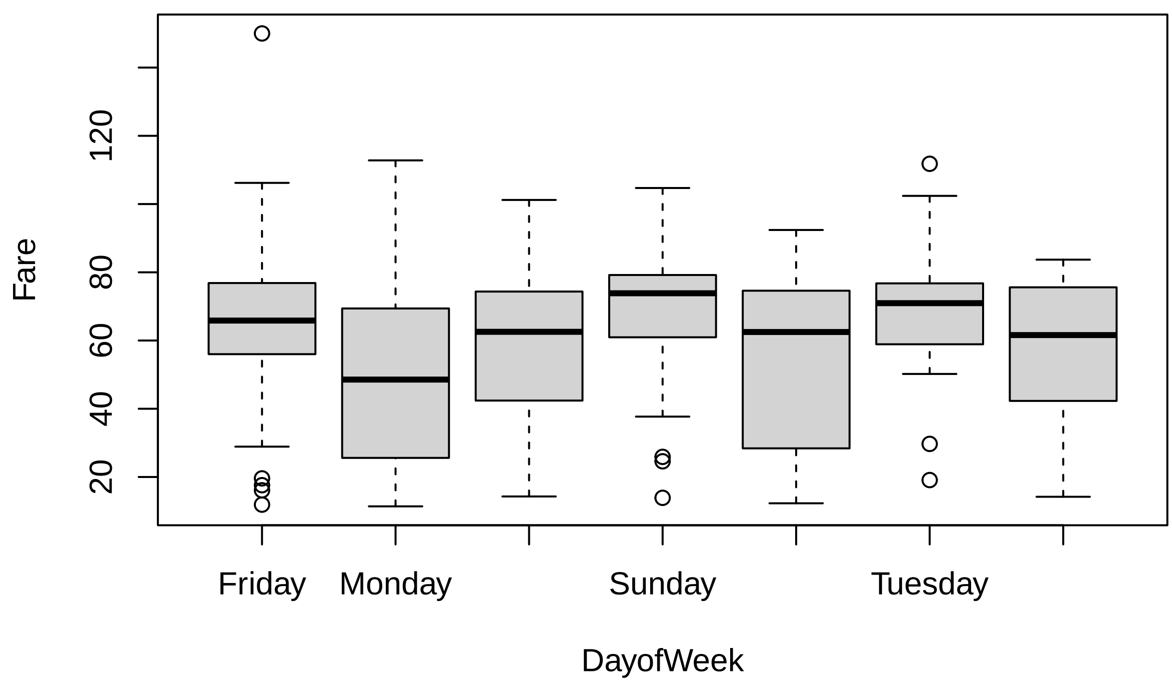
Test if the mean fare charged for rides is different for the days of the week? If so, find which day has the highest fare charged.

Hypothesis:

H0: Mean fare charged for rides is not different for the days of the week.

HA: Mean fare charged for rides is different for the days of the week.

```
boxplot(Fare ~ DayofWeek, rides)
```



From the box plot above we can observe that the variances are not equal.

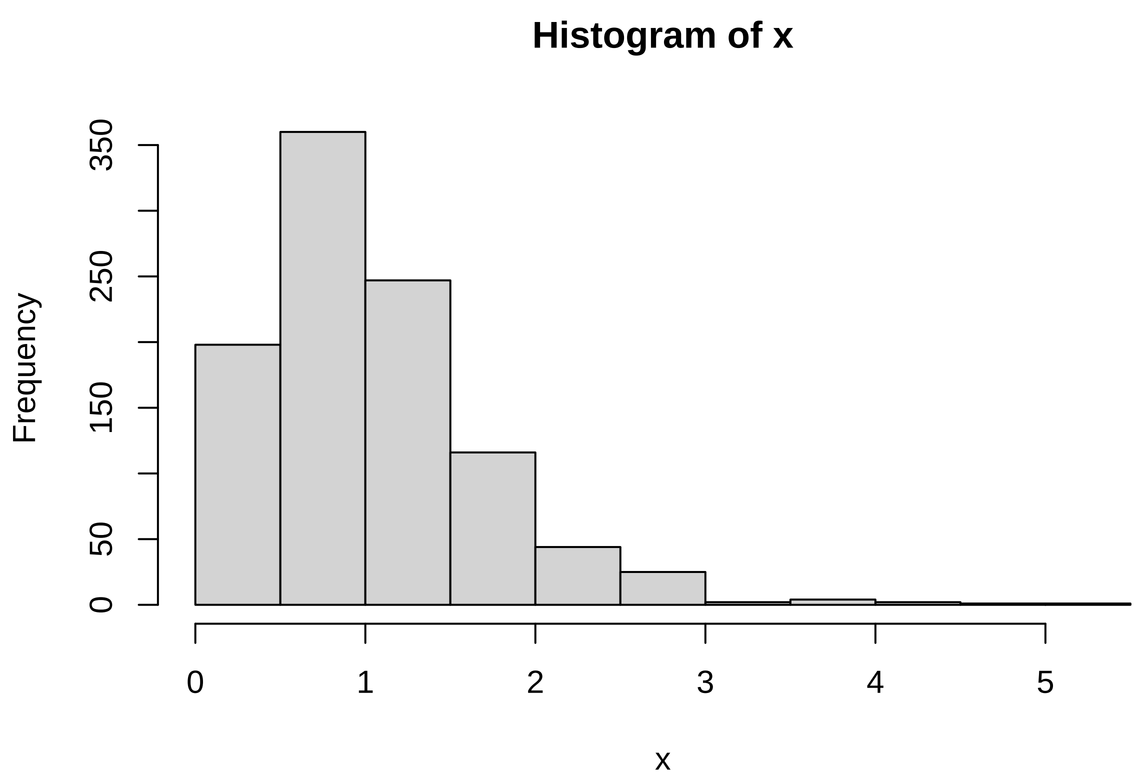
```
table(rides$DayofWeek)
```

```
##
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##        40        38        24        56        37        32        29
```

From the box plot above we can observe the mean fare charged on different days of the week. The box plot suggests that the mean values are different for the days of the week. To further investigate this, we have to calculate the f-statistic and p-value.

At first we replicate the data set 1000 to find the simulated result.

```
x <- replicate(1000, {
  DayofWeek.perm <- sample(rides$DayofWeek)
  oneway.test(Fare ~ DayofWeek.perm, data = rides, var.equal = FALSE)$statistic
})
hist(x)
```



From the histogram above we can suggest that the replication resulted in a right-skewed histogram. This means that the data is positively skewed, meaning that the majority of the observations are on the lower end of the distribution, while a few observations have higher values.

Now we calculate the f-statistic for the original data set.

```
Fstat <- oneway.test(Fare ~ DayofWeek, data = rides, var.equal = FALSE)$statistic
Fstat
```

```
##        F
## 4.405524
```

P-value is calculated by comparing the simulated f-statistics to the f-statistic of the original data set:

```
pVal <- mean(x > Fstat)
pVal
```

```
## [1] 0.002
```

Conclusion

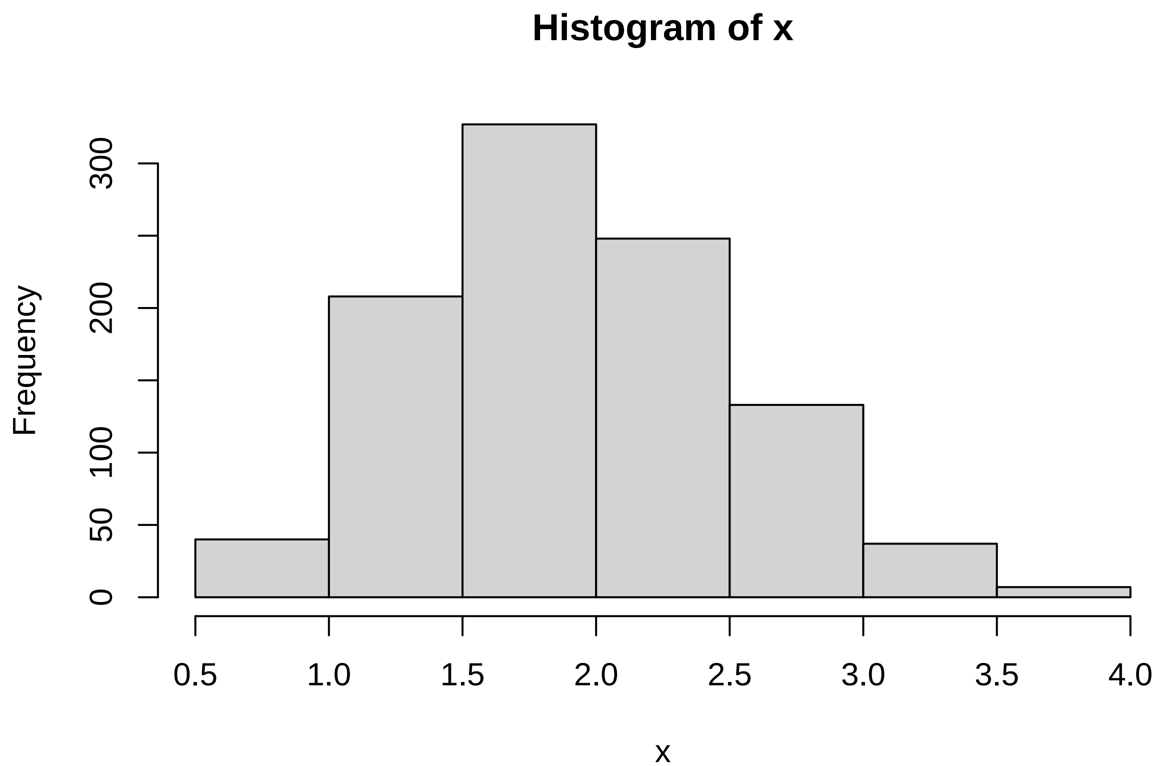
Since p-value is 0.002 which is lower than our threshold (0.05), we can say that there is evidence of a difference of the mean fare charged for rides and we have enough evidence to reject the null hypothesis.

```
ns <- table(rides$DayofWeek) # obtain sample size of each category
ns
```

```
##
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##        40        38        24        56        37        32        29
```

To find the day when the highest fare was charged, first we have to do the post-hoc pairwise comparison.

```
x <- replicate(1000, {
  DayofWeek.perm <- sample(rides$DayofWeek) # shuffle the categories
  fit0 <- aov(Fare ~ DayofWeek.perm, data = rides) # compute ANOVA to obtain MSE
  MSE <- summary(fit0)[[1]][2, 3] # Extract the MSE
  means <- aggregate(Fare ~ DayofWeek.perm, data = rides, mean)[, 2] # compute means of categories
  Ts <- outer(means, means, "-") / sqrt(outer(1 / ns, 1 / ns, "+")) # t-statistics
  Ts = Ts / sqrt(MSE) # Scale by pooled standard deviation
  max(abs(Ts)) # keep largest t statistic
})
hist(x) # examine distribution of maximum t statistics
```



Now we compute the t statistic for each pair of categories from the original data.

```
fit = aov(Fare ~ DayofWeek, data = rides)
MSE = summary(fit)[[1]][2, 3]
means = aggregate(Fare ~ DayofWeek, data = rides, mean)[, 2]
Ts = outer(means, means, "-") / sqrt(outer(1 / ns, 1 / ns, "+"))
Ts = Ts / sqrt(MSE)
Ts
```

```
##
##                Friday   Monday   Saturday      Sunday  Thursday      Tuesday
##   Friday     0.0000000 3.171549  0.9559270 -0.91557607  1.789470 -0.81773354
##   Monday    -3.1715491 0.000000 -1.8088571 -4.32020489 -1.343444 -3.80276778
##   Saturday  -0.9559270 1.808857  0.0000000 -1.78855010  0.615612 -1.63226464
##   Sunday     0.9155761 4.320205  1.7885501  0.00000000  2.821267 -0.01985676
##   Thursday  -1.7894702 1.343444 -0.6156120 -2.82126652  0.000000 -2.49417856
##   Tuesday    0.8177335 3.802768  1.6322646  0.01985676  2.494179  0.00000000
##   Wednesday -1.1389493 1.787180 -0.1121937 -2.04267558  0.525736 -1.83990368
##
##              Wednesday
##   Friday     1.1389493
##   Monday    -1.7871795
##   Saturday   0.1121937
##   Sunday     2.0426756
##   Thursday  -0.5257360
##   Tuesday    1.8399037
##   Wednesday  0.0000000
```

```r
pVal <- mean(x > Ts[4,2]) # Sunday vs Monday
pVal
```

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

The p-value is 0.

```r
TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Fare ~ DayofWeek, data = rides)
##
## $DayofWeek
##                           diff        lwr        upr     p adj
## Monday-Friday      -16.327368 -31.630597 -1.0241398 0.0280202
## Saturday-Friday     -5.609167 -23.051795 11.8334617 0.9627124
## Sunday-Friday        4.307500  -9.677716 18.2927156 0.9698429
## Thursday-Friday     -9.275946 -24.684861  6.1329692 0.5564333
## Tuesday-Friday       4.407500 -11.614577 20.4295772 0.9829902
## Wednesday-Friday    -6.312759 -22.788809 10.1632917 0.9154222
## Saturday-Monday     10.718202  -6.895717 28.3321203 0.5433385
## Sunday-Monday       20.634868   6.436591 34.8331463 0.0004464
## Thursday-Monday      7.051422  -8.551126 22.6539705 0.8306931
## Tuesday-Monday      20.734868   4.526482 36.9432552 0.0033712
## Wednesday-Monday    10.014610  -6.642673 26.6718927 0.5579811
## Sunday-Saturday      9.916667  -6.565068 26.3984013 0.5570550
## Thursday-Saturday   -3.666779 -21.372597 14.0390386 0.9962803
## Tuesday-Saturday    10.016667  -8.225271 28.2586041 0.6615564
## Wednesday-Saturday  -0.703592 -19.345522 17.9383380 0.9999998
## Thursday-Sunday    -13.583446 -27.895572  0.7286802 0.0753253
## Tuesday-Sunday       0.100000 -14.870279 15.0702789 1.0000000
## Wednesday-Sunday   -10.620259 -26.075437  4.8349193 0.3905632
## Tuesday-Thursday    13.683446  -2.624762 29.9916543 0.1658658
## Wednesday-Thursday   2.963187 -13.791243 19.7176176 0.9984587
## Wednesday-Tuesday  -10.720259 -28.040283  6.5997653 0.5224131
```

The p-value for Sunday versus Monday after performing TukeyHSD is 0.0004464 which is the lowest. So we can confirm they have the highest difference of mean i.e. the highest fare is charged on Sunday.

Conclusion

We can conclude that Sunday has the highest fare charged.

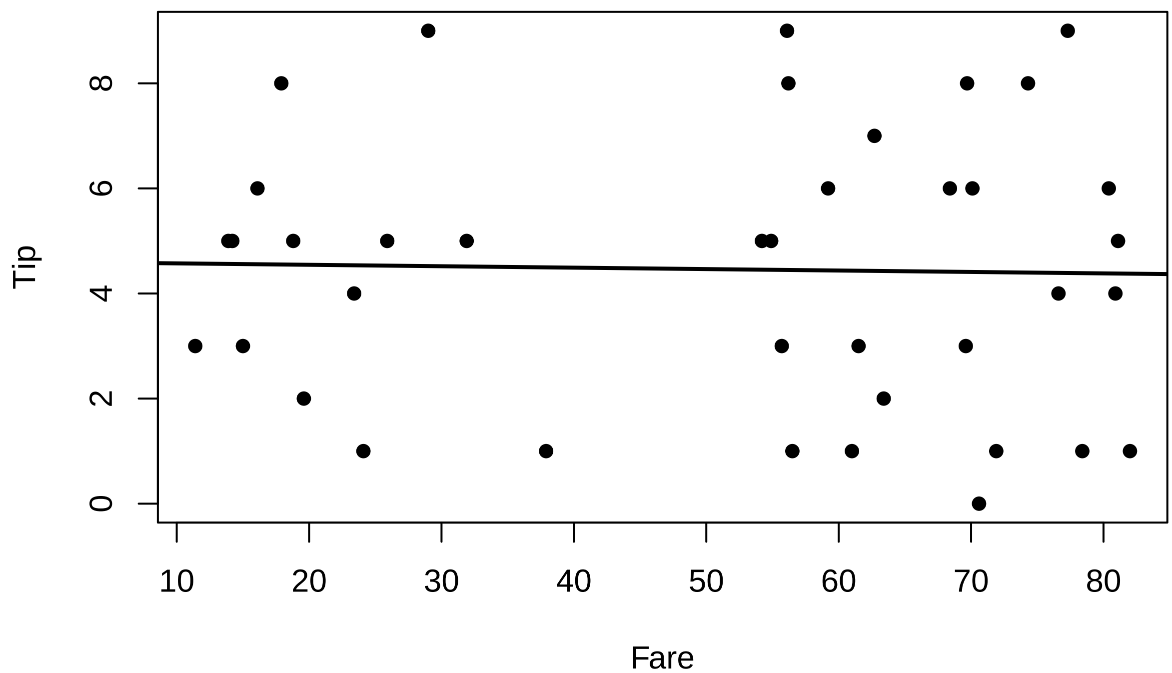
Question 5

- Draw an appropriate plot to show the relationship between the tip provided by the passenger and the fare charged for the ride taken in the mornings. Interpret your plot.
- Test if there is a linear relationship between the tip provided by the passenger and the fare charged for the ride taken in the mornings.
- Can we predict the tip provided by the passenger based on the fare charged for the ride taken in the mornings?
- If so predict the tip provided by the passenger when fare charged for the ride is 83.4 AUD.
- How good is your estimate? Discuss the suitability and/or strength of your model.

```r
library(dplyr)
morning_rides <- rides %>%
  filter(PickupTime == "Morning") %>%
  select(Fare, Tip, PickupTime)
head(morning_rides)
```

```
##   Fare Tip PickupTime
## 1 80.4   6    Morning
## 2 69.7   8    Morning
## 3 69.6   3    Morning
## 4 62.7   7    Morning
## 5 24.1   1    Morning
## 6 29.0   9    Morning
```

```r
plot(Tip ~ Fare, data = morning_rides, pch = 16)
fit = lm(Tip ~ Fare, data = morning_rides)
abline(fit, lwd = 2)
```



Hypothesis:

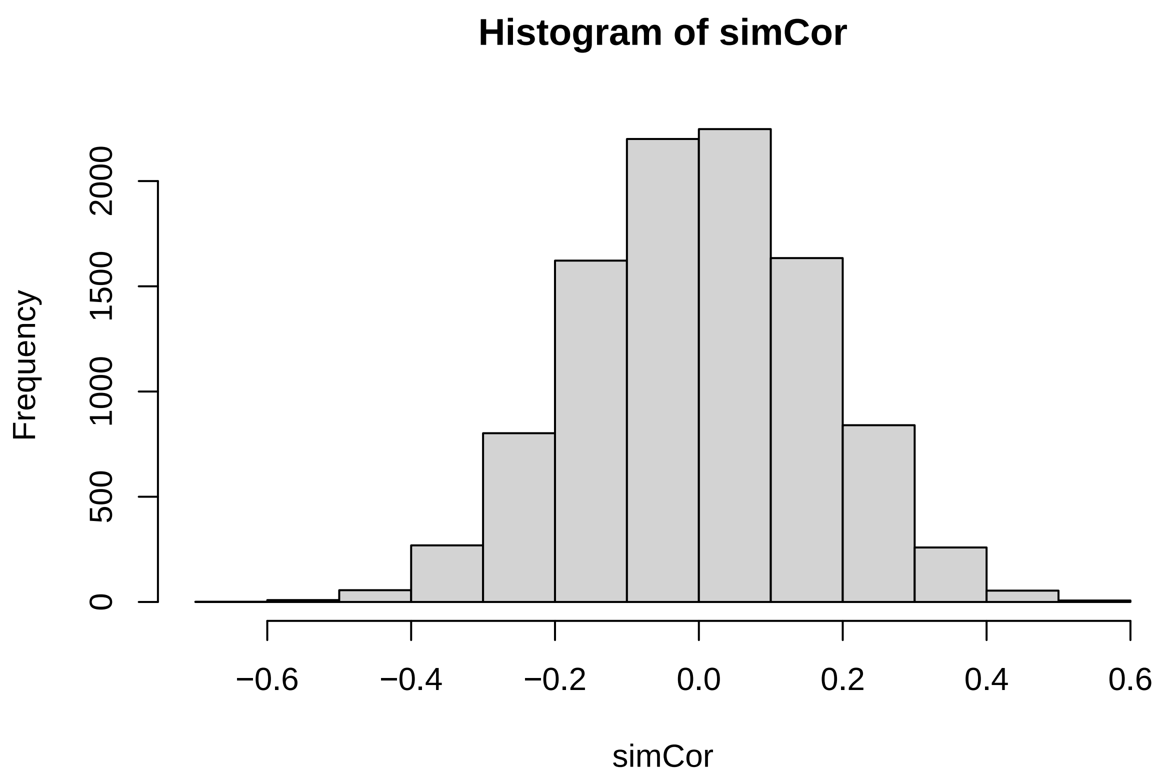
H0: There is no linear relationship between the tip provided by the passenger and the fare charged for the ride taken in the mornings.

HA: There is a linear relationship between the tip provided by the passenger and the fare charged for the ride taken in the mornings.

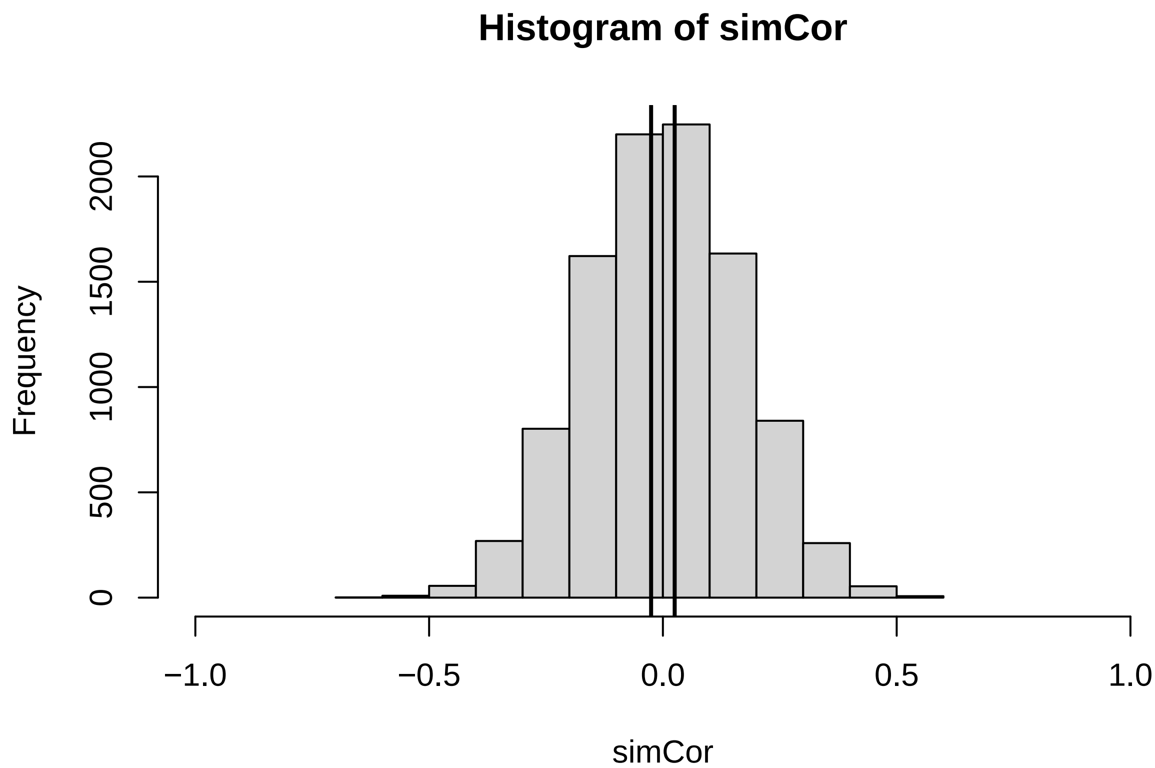
```r
cor0 <- cor(morning_rides$Fare, morning_rides$Tip)
cor0
```

```
## [1] -0.02518471
```

```
simCor = replicate(10000, {
  postShuffle = sample(morning_rides$Tip)
  cor(morning_rides$Fare, postShuffle)
})
hist(simCor)
```




```
hist(simCor, xlim = c(-1, 1))
abline(v = c(-cor0, cor0), lwd = 2)
```




```
pVal = mean(simCor > cor0) + mean(simCor < (-cor0))
pVal
```

```
## [1] 1.1157
```

Conclusion

P-value is 1.1157 which is higher than our threshold (0.05), so we can say that we do not have enough evidence to reject the null hypothesis i.e. we do not have enough evidence to say that there is a linear relationship between the tip provided by the passenger and the fare charged for the ride taken in the mornings

Since the correlation value is -0.0251847 which is very low, we can suggest that we cannot predict the tip provided by the passenger based on the fare charged for the ride taken in the mornings.

Although the model is very weak, we can still calculate the tip using the following method:

```
print(fit)
```

```
##
## Call:
## lm(formula = Tip ~ Fare, data = morning_rides)
##
## Coefficients:
## (Intercept)         Fare
##    4.600602    -0.002723
```

Prediction for the tip provided by the passenger when fare charged for the ride is 83.4 AUD:

```r
y = 4.600602 + (-0.002723) * 83.4
y
```

```
## [1] 4.373504
```

Predicted Tip: 4.373504

```r
summary(fit)
```

```
##
## Call:
## lm(formula = Tip ~ Fare, data = morning_rides)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4084 -1.9988  0.4381  1.5880  4.6099
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.600602   1.001687   4.593 4.92e-05 ***
## Fare        -0.002723   0.017769  -0.153    0.879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.648 on 37 degrees of freedom
## Multiple R-squared:  0.0006343,  Adjusted R-squared:  -0.02638
## F-statistic: 0.02348 on 1 and 37 DF,  p-value: 0.879
```

From the Multiple R-squared: 0.0006343, which is close to 0, we can suggest that it's bad regression model which may lead to a bad estimate.