# analysis

Phani Sai Kamal Lingam

11/21/2019

# Data Preparation

## 1. Introduction

Question of Interest: How the Payment Preferences for tickets in Buses impacts Boarding Time of the Passengers?

Quantitative Variable (Continous): Boarding Duration in Seconds (The time from first step in the bus till the ticket is issued)

Categorical Variable (Nominal): Mode of Payment, Cash (Coins and Notes) and Pass (Card, Transfer Pass, Application)

Analysis Method: Difference in Means

Statistical Test: Welch Two Sample T-test

Additional Varibles:

Ethnicity: Four Level Factor (Categorical; Nominal)

W - Whites

HL - Hispanic/Latinos

AA - African Americans

A - Asians and Others

Gender: Two Level Factor (Categorical; Nominal)

M - Male

F - Female

AgeGroup: Four Level Facotr (Categorical; Ordinal)

Y - Youth (0 - 17)

A - Adults (18 - 34)

M - Middle Aged (35 - 54)

S - Seniors (55+)

Employment: Two Level Factor (Categorical; Nominal)

E - Employed

U - Unemployed

Population Data Points: 150 (Cash - 42, Pass - 108 )

Sample Data Points: 80 (Cash - 40, Pass - 40; n > 30)

Collection Methodology: Random Buses and Random Routes are chosen on random days to minimize bias. Data Points are collected at few random stops instead of all stops throughout the routes or of a particular bus to prevent Sampling Bias and Selection Bias.

Some nuances araised due to the missing data points in crowded buses and high density stops.

Inference Population: Bus Commuters in the City of Seattle

Inference Parameter: Mean of Boarding Time of Passenger

Explanatory Variable: Mode of Payment

Respone Variable: Boarding Duration

Population Variance: Unknown

```
# intalling packages
install.packages("ggplot2")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

```
install.packages("dplyr")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

```
# importing libraries
library("ggplot2")
library("dplyr")
```

```
##
## 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
```

# 2. Sampling Data

```
# importing data from CSV to a Dataframe
myData <- read.csv(file = "dataset.csv")
head(myData)
```

```
##   Ethnicity Gender AgeGroup Employment ModeOfPayment BoardingDuration
## 1        AA      M        M          E             P             4.45
## 2         W      M        S          E             C             7.36
## 3         W      M        A          E             C             3.90
## 4         W      F        M          E             C             3.66
## 5         A      F        A          E             P             9.42
## 6         A      F        M          E             P             4.35
```

```
# Entire collected data
myData
```

```
##     Ethnicity Gender AgeGroup Employment ModeOfPayment BoardingDuration
## 1          AA      M        M          E             P             4.45
## 2           W      M        S          E             C             7.36
## 3           W      M        A          E             C             3.90
## 4           W      F        M          E             C             3.66
## 5           A      F        A          E             P             9.42
## 6           A      F        M          E             P             4.35
## 7           W      F        M          E             C             4.22
## 8          HL      M        A          E             P             8.11
## 9           W      M        M          E             P             2.36
## 10          W      M        M          E             P             6.49
## 11         HL      M        A          U             P             8.49
## 12          W      M        S          E             P             8.60
## 13          A      F        S          E             P             2.60
## 14          W      M        A          U             P             9.99
## 15          W      M        A          E             P             2.08
## 16         HL      M        S          U             C             2.62
## 17         HL      M        M          E             P             7.97
## 18         AA      M        A          E             P             9.12
## 19         AA      F        A          E             P             2.95
## 20         HL      F        S          U             P             6.85
## 21          W      M        A          E             P             1.86
## 22         HL      M        A          E             P             8.10
## 23         AA      F        S          U             P             5.33
## 24          W      M        M          E             P             2.46
## 25         AA      M        M          E             C             9.96
## 26         HL      M        A          U             C             6.23
## 27          W      F        M          E             P             2.81
## 28          A      F        M          E             P             4.34
## 29         HL      F        A          E             C             9.02
## 30         HL      F        M          U             P             8.87
## 31          W      F        M          E             C             3.40
## 32          W      M        S          E             P             5.67
## 33         HL      F        M          E             C             4.97
## 34         HL      M        M          E             P             2.21
## 35          W      F        S          E             P             8.06
## 36         AA      M        M          E             C             4.86
## 37          W      M        S          E             C             4.75
## 38          W      F        M          E             C             6.88
## 39          W      M        A          E             C             3.90
## 40          W      F        A          E             P             6.00
## 41          W      M        A          E             P             3.61
## 42          W      M        S          E             P             1.90
## 43         HL      F        M          E             P             7.37
## 44          W      F        M          E             C             5.80
## 45          A      M        M          U             C             8.27
## 46          W      M        M          E             P             8.08
## 47          W      M        A          E             P             5.50
## 48          W      F        M          E             P            10.27
## 49          W      M        A          U             P             8.29
## 50         AA      M        M          E             P             3.78
## 51          W      F        S          E             P             6.96
## 52          W      F        M          E             P             2.79
```

| | | | | | | |
|---|---|---|---|---|---|---|
| ## 53 | AA | M | M | E | P | 8.56 |
| ## 54 | AA | F | M | E | C | 8.08 |
| ## 55 | W | M | Y | E | C | 9.99 |
| ## 56 | HL | M | S | U | P | 10.01 |
| ## 57 | HL | M | A | E | P | 7.33 |
| ## 58 | A | F | S | E | C | 10.00 |
| ## 59 | AA | M | A | E | P | 3.72 |
| ## 60 | A | M | A | E | P | 9.59 |
| ## 61 | HL | M | M | E | P | 10.04 |
| ## 62 | W | M | S | E | P | 6.68 |
| ## 63 | W | M | A | U | P | 2.07 |
| ## 64 | W | M | A | E | P | 9.30 |
| ## 65 | W | M | A | E | P | 4.75 |
| ## 66 | AA | M | S | E | P | 7.37 |
| ## 67 | W | F | A | E | P | 4.05 |
| ## 68 | W | M | Y | E | P | 4.49 |
| ## 69 | A | M | A | E | P | 3.27 |
| ## 70 | AA | F | M | U | P | 8.07 |
| ## 71 | AA | F | M | E | P | 3.75 |
| ## 72 | W | M | A | E | P | 6.32 |
| ## 73 | W | M | S | E | P | 5.43 |
| ## 74 | AA | M | S | E | C | 5.38 |
| ## 75 | W | F | M | E | P | 5.00 |
| ## 76 | W | M | A | E | P | 9.14 |
| ## 77 | AA | F | M | E | P | 3.30 |
| ## 78 | HL | F | S | E | P | 7.54 |
| ## 79 | A | M | A | E | C | 3.19 |
| ## 80 | AA | F | A | E | P | 7.72 |
| ## 81 | W | F | A | E | C | 9.75 |
| ## 82 | HL | M | M | E | C | 5.85 |
| ## 83 | W | M | A | E | C | 9.54 |
| ## 84 | W | M | M | U | C | 8.37 |
| ## 85 | HL | M | M | E | P | 3.81 |
| ## 86 | HL | M | M | E | P | 8.43 |
| ## 87 | HL | M | M | U | P | 4.54 |
| ## 88 | W | F | M | E | P | 8.95 |
| ## 89 | HL | M | A | E | P | 2.68 |
| ## 90 | HL | F | M | E | P | 5.61 |
| ## 91 | HL | F | M | E | P | 5.00 |
| ## 92 | W | F | M | E | P | 5.61 |
| ## 93 | W | M | M | U | P | 3.57 |
| ## 94 | W | F | A | E | C | 9.34 |
| ## 95 | W | M | A | E | P | 6.23 |
| ## 96 | W | F | S | E | P | 3.17 |
| ## 97 | W | M | M | E | C | 6.32 |
| ## 98 | W | F | A | E | P | 5.66 |
| ## 99 | AA | F | M | E | C | 5.09 |
| ## 100 | W | M | S | E | P | 4.08 |
| ## 101 | HL | M | A | E | P | 4.25 |
| ## 102 | W | F | Y | E | P | 1.97 |
| ## 103 | W | M | A | E | C | 2.26 |
| ## 104 | AA | M | A | E | C | 9.57 |
| ## 105 | HL | M | Y | E | P | 5.73 |
| ## 106 | AA | M | S | E | C | 8.51 |

```
## 107        W      F        M        U        P         7.19
## 108        W      M        M        E        P         9.60
## 109       HL      M        A        E        P         2.63
## 110        W      M        S        E        P         7.21
## 111        W      F        A        E        P         4.71
## 112        W      M        M        E        P         2.21
## 113        A      F        M        E        P         4.05
## 114       HL      M        M        E        P         4.02
## 115        W      F        S        E        P        10.26
## 116        W      M        S        E        P         5.53
## 117        W      M        M        E        C         5.13
## 118        W      M        A        E        P         2.16
## 119        W      F        A        U        C         6.02
## 120        W      M        M        E        P         5.67
## 121        W      M        M        E        P         7.13
## 122        W      M        S        U        C         6.06
## 123        W      M        A        E        P         1.98
## 124        A      M        M        E        C         6.34
## 125       HL      F        S        E        P         8.43
## 126        W      F        A        E        C         6.12
## 127        W      M        M        U        C         6.89
## 128        W      F        M        E        P         5.54
## 129       AA      F        S        E        P         7.77
## 130        A      M        A        E        P         6.88
## 131        W      M        A        E        P         9.65
## 132        W      M        S        E        P         2.22
## 133       HL      F        M        E        C         8.87
## 134        W      M        S        U        P         5.42
## 135        W      F        A        E        P         2.05
## 136        W      M        M        E        P         3.55
## 137       HL      M        M        U        C         8.54
## 138       HL      F        A        E        P         9.08
## 139       HL      M        A        E        C         3.50
## 140        W      F        M        E        P         5.54
## 141        A      M        A        U        P         6.11
## 142        W      M        A        U        C         8.70
## 143        W      M        M        E        P         3.31
## 144        W      F        M        E        P         1.94
## 145        W      F        S        U        P         2.63
## 146       HL      M        A        E        C         9.44
## 147        W      M        A        E        P         7.38
## 148        W      M        S        E        P         4.26
## 149        W      M        M        E        P         5.45
## 150       AA      M        M        E        P         8.52
```

```
# Dataset Summary
summary(myData)
```

```
##  Ethnicity Gender AgeGroup Employment ModeOfPayment BoardingDuration
##  A :13     F:56   A:53     E:126      C: 42         Min.   : 1.860
##  AA:21     M:94   M:62     U: 24      P:108         1st Qu.: 3.833
##  HL:33            S:31                             Median : 5.700
##  W :83            Y: 4                             Mean   : 5.944
##                                                    3rd Qu.: 8.107
##                                                    Max.   :10.270
```

In order to do statistical analysis we chose 40 data point from each category to come up with a reliable conclusion. In the Population data the “Pass Category” is dominating the “Cash Category” as it is the most preferred method of payment for tickets in real life because of it’s tap and go approach instead of inserting the exact amount of cash into the machine and wait for the ticket. In our Sample data we took 40 data points of each category to reduce the impact of dominance of category on the other and thereby preventing the possible Sample Bias.

```
# Choosing a random seed number to introduce randomization in sampling process.
set.seed(7384)
```

```
# Sample data of 40 passengers using Cash for ticket payment (40 out of 42 data points).
cashData <- filter(myData, myData$ModeOfPayment == 'C')
cashDataSample <- cashData[sample(nrow(cashData), 40), ]
cashDataSample
```

```
##    Ethnicity Gender AgeGroup Employment ModeOfPayment BoardingDuration
## 24         W      M        A          E             C             9.54
## 18         W      M        Y          E             C             9.99
## 33         W      F        A          U             C             6.02
## 5         HL      M        S          U             C             2.62
## 19         A      F        S          E             C            10.00
## 39        HL      M        M          U             C             8.54
## 31        AA      M        S          E             C             8.51
## 9          W      F        M          E             C             3.40
## 3          W      F        M          E             C             3.66
## 21         A      M        A          E             C             3.19
## 11        AA      M        M          E             C             4.86
## 20        AA      M        S          E             C             5.38
## 41         W      M        A          U             C             8.70
## 40        HL      M        A          E             C             3.50
## 26         W      F        A          E             C             9.34
## 35         A      M        M          E             C             6.34
## 28        AA      F        M          E             C             5.09
## 30        AA      M        A          E             C             9.57
## 34         W      M        S          U             C             6.06
## 15         W      F        M          E             C             5.80
## 6         AA      M        M          E             C             9.96
## 25         W      M        M          U             C             8.37
## 1          W      M        S          E             C             7.36
## 37         W      M        M          U             C             6.89
## 17        AA      F        M          E             C             8.08
## 27         W      M        M          E             C             6.32
## 36         W      F        A          E             C             6.12
## 10        HL      F        M          E             C             4.97
## 14         W      M        A          E             C             3.90
## 38        HL      F        M          E             C             8.87
## 12         W      M        S          E             C             4.75
## 13         W      F        M          E             C             6.88
## 23        HL      M        M          E             C             5.85
## 32         W      M        M          E             C             5.13
## 4          W      F        M          E             C             4.22
## 7         HL      M        A          U             C             6.23
## 22         W      F        A          E             C             9.75
## 8         HL      F        A          E             C             9.02
## 42        HL      M        A          E             C             9.44
## 16         A      M        M          U             C             8.27
```

```
# Cash Data Statistics
summary(cashDataSample)
```

```
##  Ethnicity Gender AgeGroup Employment ModeOfPayment BoardingDuration
##  A : 4     F:15   A:13     E:31       C:40          Min.   : 2.620
##  AA: 7     M:25   M:19     U: 9       P: 0          1st Qu.: 5.060
##  HL: 9            S: 7                             Median : 6.330
##  W :20            Y: 1                             Mean   : 6.762
##                                                    3rd Qu.: 8.742
##                                                    Max.   :10.000
```

```
# Choosing a random seed number to introduce randomization in sampling process.
set.seed(4250)
```

```
# Sample data of 40 passengers using Pass for ticket payment (40 out of 108 data points).
passData <- filter(myData, myData$ModeOfPayment == 'P')
passDataSample <- passData[sample(nrow(passData), 40), ]
passDataSample
```

```
##     Ethnicity Gender AgeGroup Employment ModeOfPayment BoardingDuration
## 89          W      M        A          E             P             1.98
## 12         HL      M        M          E             P             7.97
## 78         HL      M        A          E             P             2.63
## 44          W      M        A          U             P             2.07
## 92         AA      F        S          E             P             7.77
## 5           W      M        M          E             P             2.36
## 52         AA      F        M          E             P             3.75
## 71          W      F        A          E             P             5.66
## 10          W      M        A          U             P             9.99
## 30          W      M        M          E             P             8.08
## 4          HL      M        A          E             P             8.11
## 74          W      F        Y          E             P             1.97
## 96          W      M        S          U             P             5.42
## 58         HL      F        S          E             P             7.54
## 50          A      M        A          E             P             3.27
## 95          W      M        S          E             P             2.22
## 38         HL      M        S          U             P            10.01
## 31          W      M        A          E             P             5.50
## 39         HL      M        A          E             P             7.33
## 45          W      M        A          E             P             9.30
## 91          W      F        M          E             P             5.54
## 46          W      M        A          E             P             4.75
## 19          W      M        M          E             P             2.46
## 81          W      M        M          E             P             2.21
## 99         HL      F        A          E             P             9.08
## 105         W      M        A          E             P             7.38
## 24         HL      M        M          E             P             2.21
## 86          W      M        A          E             P             2.16
## 57         AA      F        M          E             P             3.30
## 108        AA      M        M          E             P             8.52
## 13         AA      M        A          E             P             9.12
## 48          W      F        A          E             P             4.05
## 87          W      M        M          E             P             5.67
## 75         HL      M        Y          E             P             5.73
## 11          W      M        A          E             P             2.08
## 98          W      M        M          E             P             3.55
## 106         W      M        S          E             P             4.26
## 90         HL      F        S          E             P             8.43
## 80          W      F        A          E             P             4.71
## 33          W      M        A          U             P             8.29
```

```
# Pass Data Statistics
summary(passDataSample)
```

```
##  Ethnicity Gender AgeGroup Employment ModeOfPayment BoardingDuration
##  A : 1     F:11   A:19     E:35       C: 0          Min.   : 1.970
##  AA: 5     M:29   M:12     U: 5       P:40          1st Qu.: 2.587
##  HL:10            S: 7                             Median : 5.460
##  W :24            Y: 2                             Mean   : 5.411
##                                                    3rd Qu.: 7.997
##                                                    Max.   :10.010
```

```
# Sampled Data chosen for Statistical Analysis.
sampleData <- rbind(cashDataSample, passDataSample)
head(sampleData)
```

```
##    Ethnicity Gender AgeGroup Employment ModeOfPayment BoardingDuration
## 24         W      M        A          E             C             9.54
## 18         W      M        Y          E             C             9.99
## 33         W      F        A          U             C             6.02
## 5         HL      M        S          U             C             2.62
## 19         A      F        S          E             C            10.00
## 39        HL      M        M          U             C             8.54
```

```
# Sampled Data
sampleData
```

```
##     Ethnicity Gender AgeGroup Employment ModeOfPayment BoardingDuration
## 24          W      M        A          E             C             9.54
## 18          W      M        Y          E             C             9.99
## 33          W      F        A          U             C             6.02
## 5          HL      M        S          U             C             2.62
## 19          A      F        S          E             C            10.00
## 39         HL      M        M          U             C             8.54
## 31         AA      M        S          E             C             8.51
## 9           W      F        M          E             C             3.40
## 3           W      F        M          E             C             3.66
## 21          A      M        A          E             C             3.19
## 11         AA      M        M          E             C             4.86
## 20         AA      M        S          E             C             5.38
## 41          W      M        A          U             C             8.70
## 40         HL      M        A          E             C             3.50
## 26          W      F        A          E             C             9.34
## 35          A      M        M          E             C             6.34
## 28         AA      F        M          E             C             5.09
## 30         AA      M        A          E             C             9.57
## 34          W      M        S          U             C             6.06
## 15          W      F        M          E             C             5.80
## 6          AA      M        M          E             C             9.96
## 25          W      M        M          U             C             8.37
## 1           W      M        S          E             C             7.36
## 37          W      M        M          U             C             6.89
## 17         AA      F        M          E             C             8.08
## 27          W      M        M          E             C             6.32
## 36          W      F        A          E             C             6.12
## 10         HL      F        M          E             C             4.97
## 14          W      M        A          E             C             3.90
## 38         HL      F        M          E             C             8.87
## 12          W      M        S          E             C             4.75
## 13          W      F        M          E             C             6.88
## 23         HL      M        M          E             C             5.85
## 32          W      M        M          E             C             5.13
## 4           W      F        M          E             C             4.22
## 7          HL      M        A          U             C             6.23
## 22          W      F        A          E             C             9.75
## 8          HL      F        A          E             C             9.02
## 42         HL      M        A          E             C             9.44
## 16          A      M        M          U             C             8.27
## 89          W      M        A          E             P             1.98
## 121        HL      M        M          E             P             7.97
## 78         HL      M        A          E             P             2.63
## 44          W      M        A          U             P             2.07
## 92         AA      F        S          E             P             7.77
## 51          W      M        M          E             P             2.36
## 52         AA      F        M          E             P             3.75
## 71          W      F        A          E             P             5.66
## 101         W      M        A          U             P             9.99
## 301         W      M        M          E             P             8.08
## 43         HL      M        A          E             P             8.11
## 74          W      F        Y          E             P             1.97
```

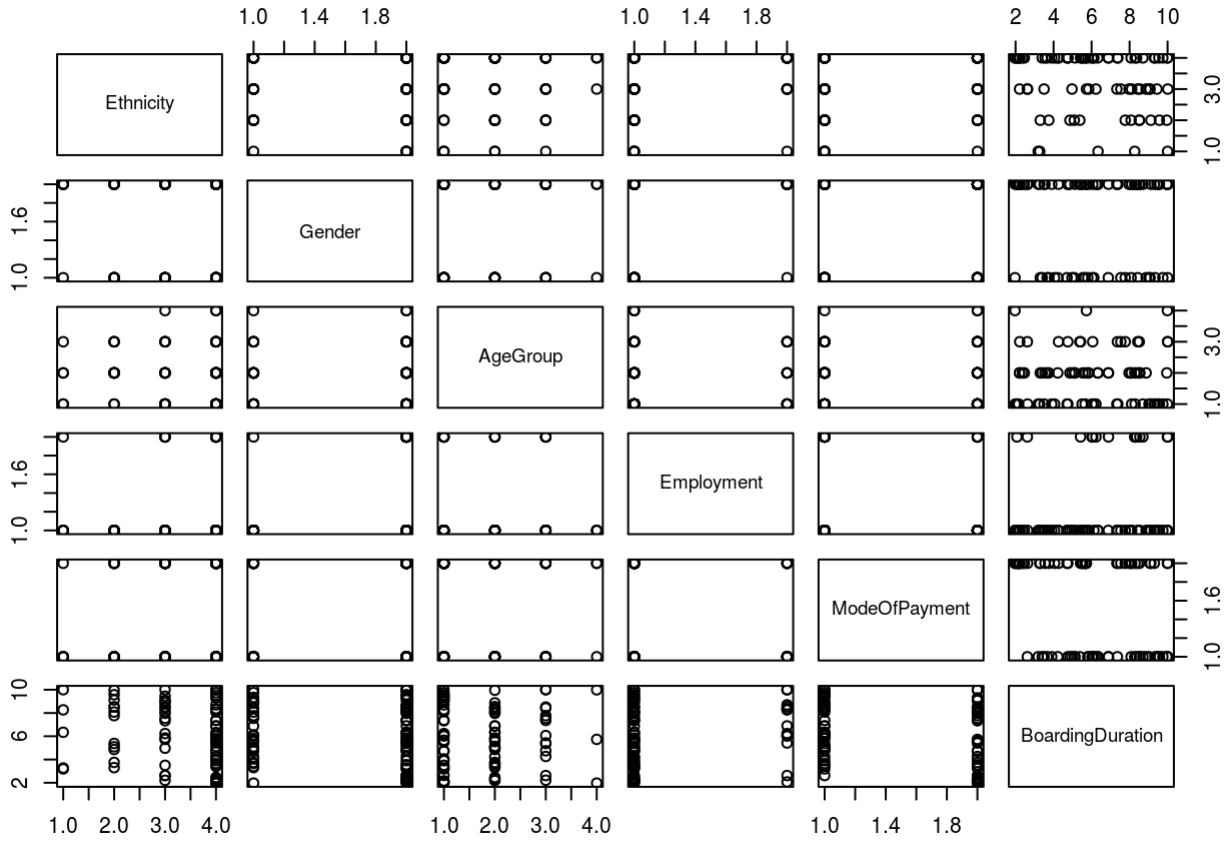
```
## 96          W      M        S          U              P             5.42
## 58         HL      F        S          E              P             7.54
## 50          A      M        A          E              P             3.27
## 95          W      M        S          E              P             2.22
## 381        HL      M        S          U              P            10.01
## 311         W      M        A          E              P             5.50
## 391        HL      M        A          E              P             7.33
## 45          W      M        A          E              P             9.30
## 91          W      F        M          E              P             5.54
## 46          W      M        A          E              P             4.75
## 191         W      M        M          E              P             2.46
## 81          W      M        M          E              P             2.21
## 99         HL      F        A          E              P             9.08
## 105         W      M        A          E              P             7.38
## 241        HL      M        M          E              P             2.21
## 86          W      M        A          E              P             2.16
## 57         AA      F        M          E              P             3.30
## 108        AA      M        M          E              P             8.52
## 131        AA      M        A          E              P             9.12
## 48          W      F        A          E              P             4.05
## 87          W      M        M          E              P             5.67
## 75         HL      M        Y          E              P             5.73
## 111         W      M        A          E              P             2.08
## 98          W      M        M          E              P             3.55
## 106         W      M        S          E              P             4.26
## 90         HL      F        S          E              P             8.43
## 80          W      F        A          E              P             4.71
## 331         W      M        A          U              P             8.29
```

```
# Sample Data Summary
summary(sampleData)
```

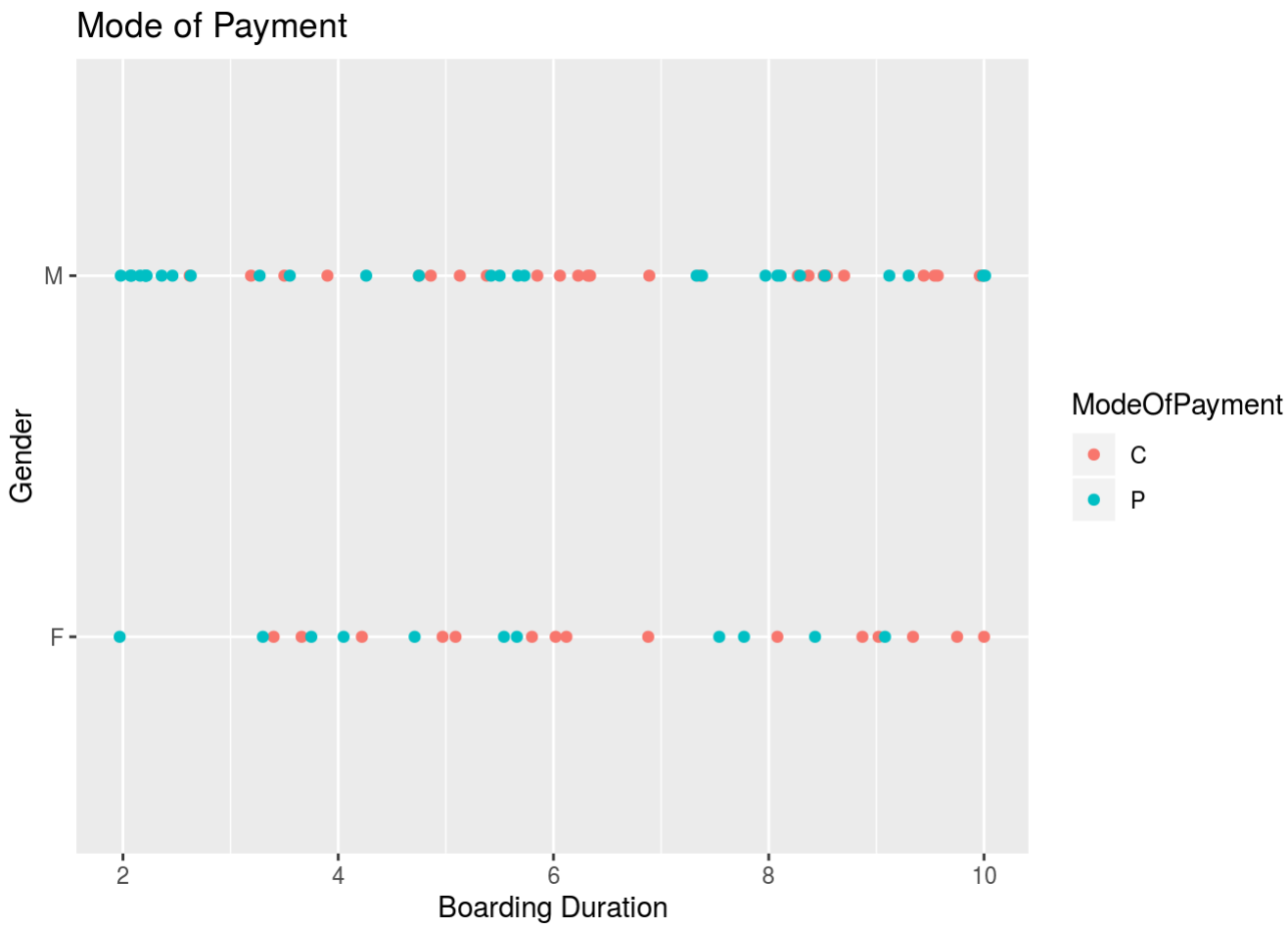
```
##  Ethnicity Gender AgeGroup Employment ModeOfPayment BoardingDuration
##  A : 5     F:26   A:32     E:66       C:40          Min.   : 1.970
##  AA:12     M:54   M:31     U:14       P:40          1st Qu.: 3.862
##  HL:19            S:14                             Median : 5.935
##  W :44            Y: 3                             Mean   : 6.087
##                                                    3rd Qu.: 8.385
##                                                    Max.   :10.010
```

# Exploratory Analysis

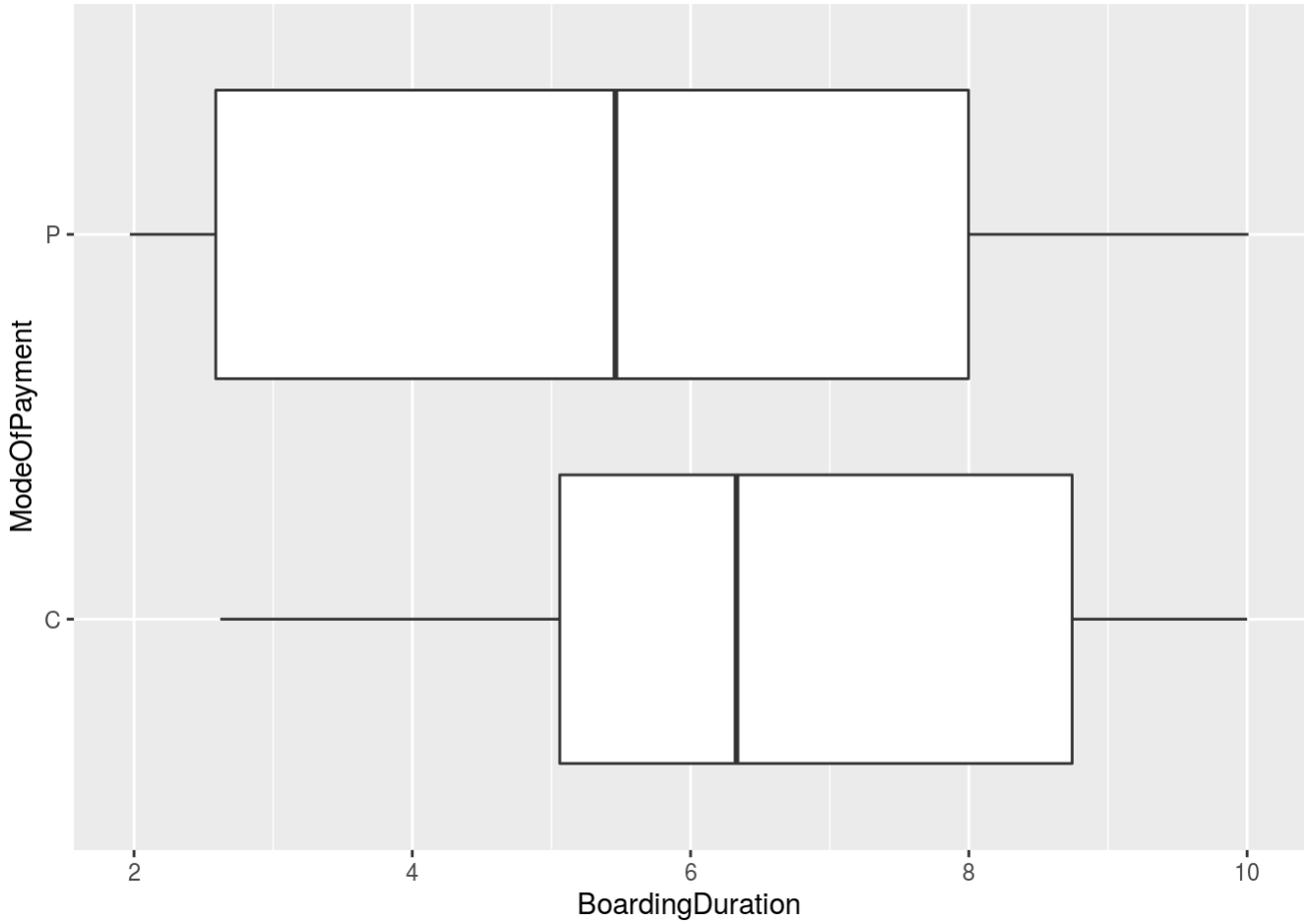
```
# graphs of all pairs of variables
pairs(sampleData)
```




```
ggplot(sampleData,aes(x=BoardingDuration,y=Gender,color=ModeOfPayment))+geom_point()+ylab('Gende
r')+xlab('Boarding Duration')+ ggtitle('Mode of Payment')
```



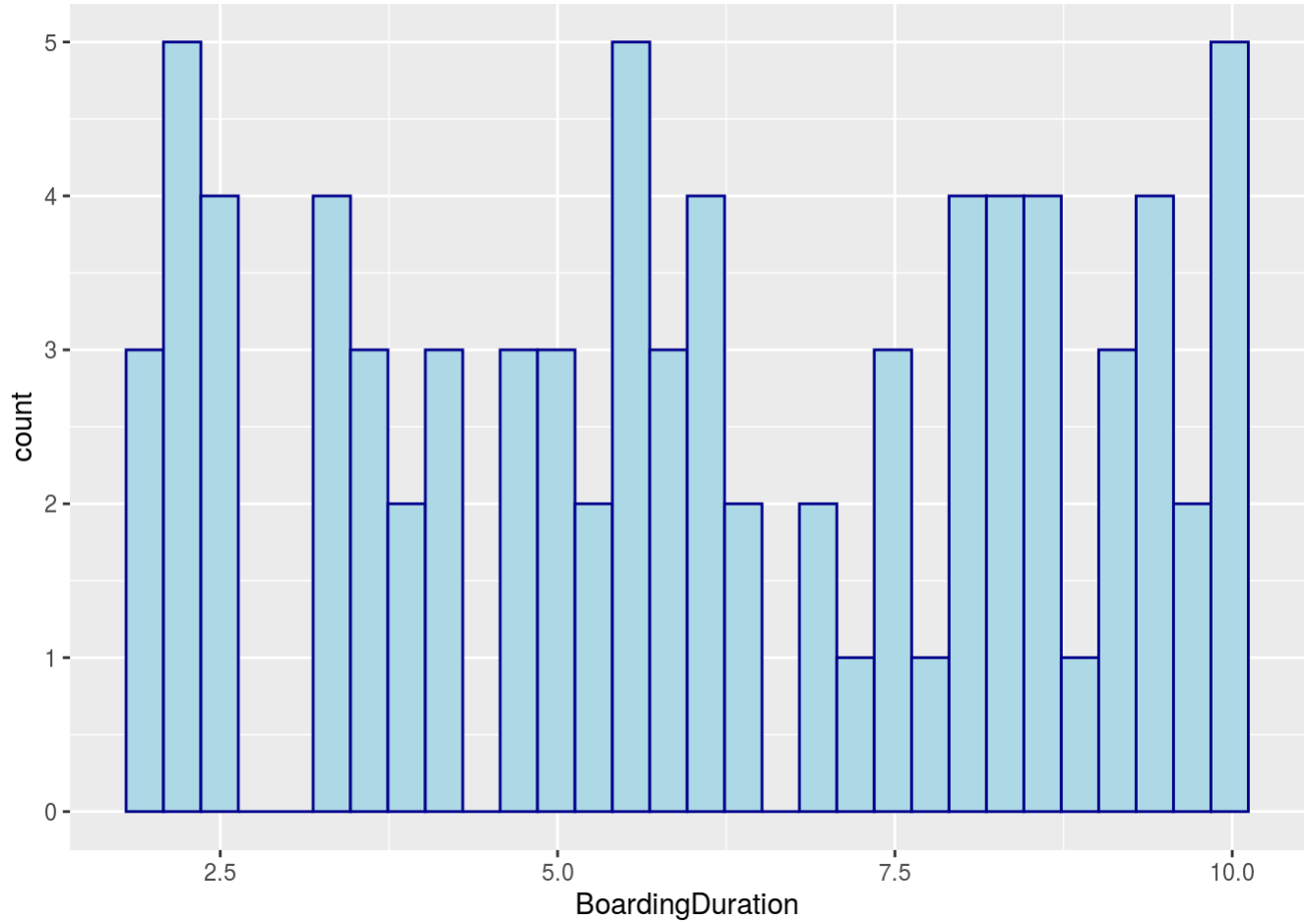

```
# Box plot for Mode of Payment vs Boarding Duration
ggplot(sampleData, aes(x = ModeOfPayment, y = BoardingDuration)) + geom_boxplot() + coord_flip()
```




Boxplot depicting the effect of mode of payment preference on Boarding Duration

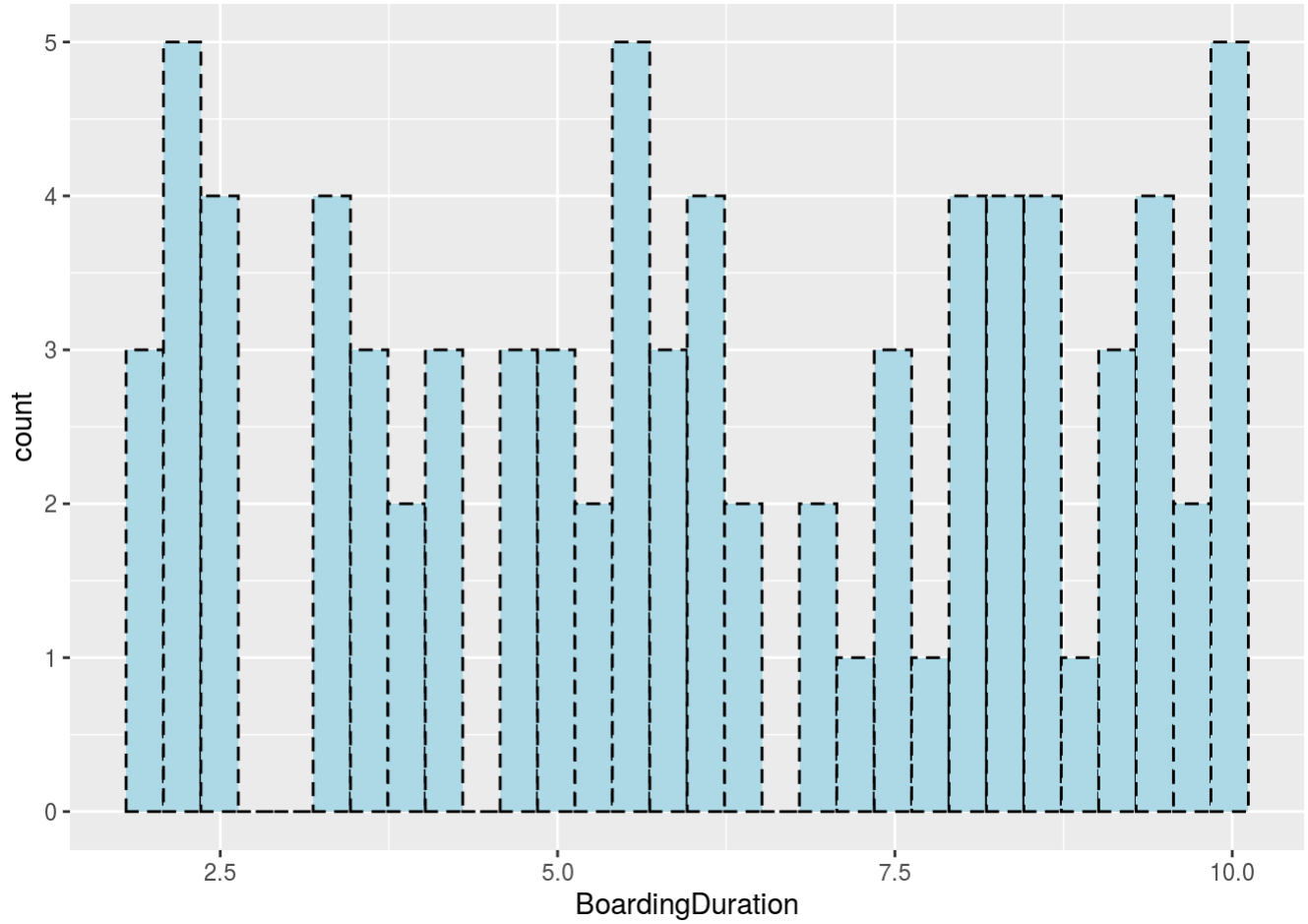
```
# histogram with fill color
ggplot(sampleData, aes(x=BoardingDuration))+
  geom_histogram(color="darkblue", fill="lightblue")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



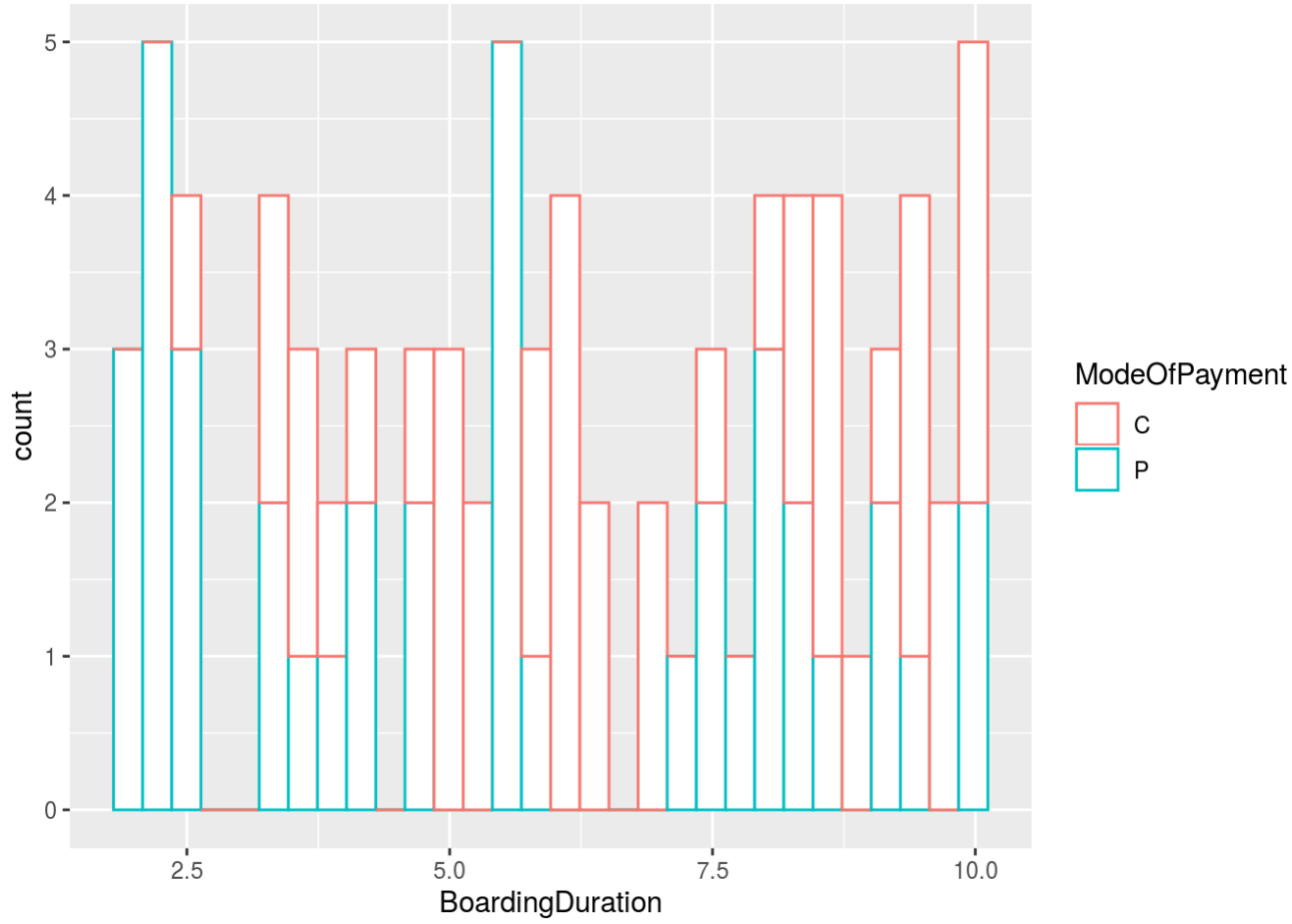

```
# dashed line
ggplot(sampleData, aes(x=BoardingDuration))+
  geom_histogram(color="black", fill="lightblue",
                 linetype="dashed")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



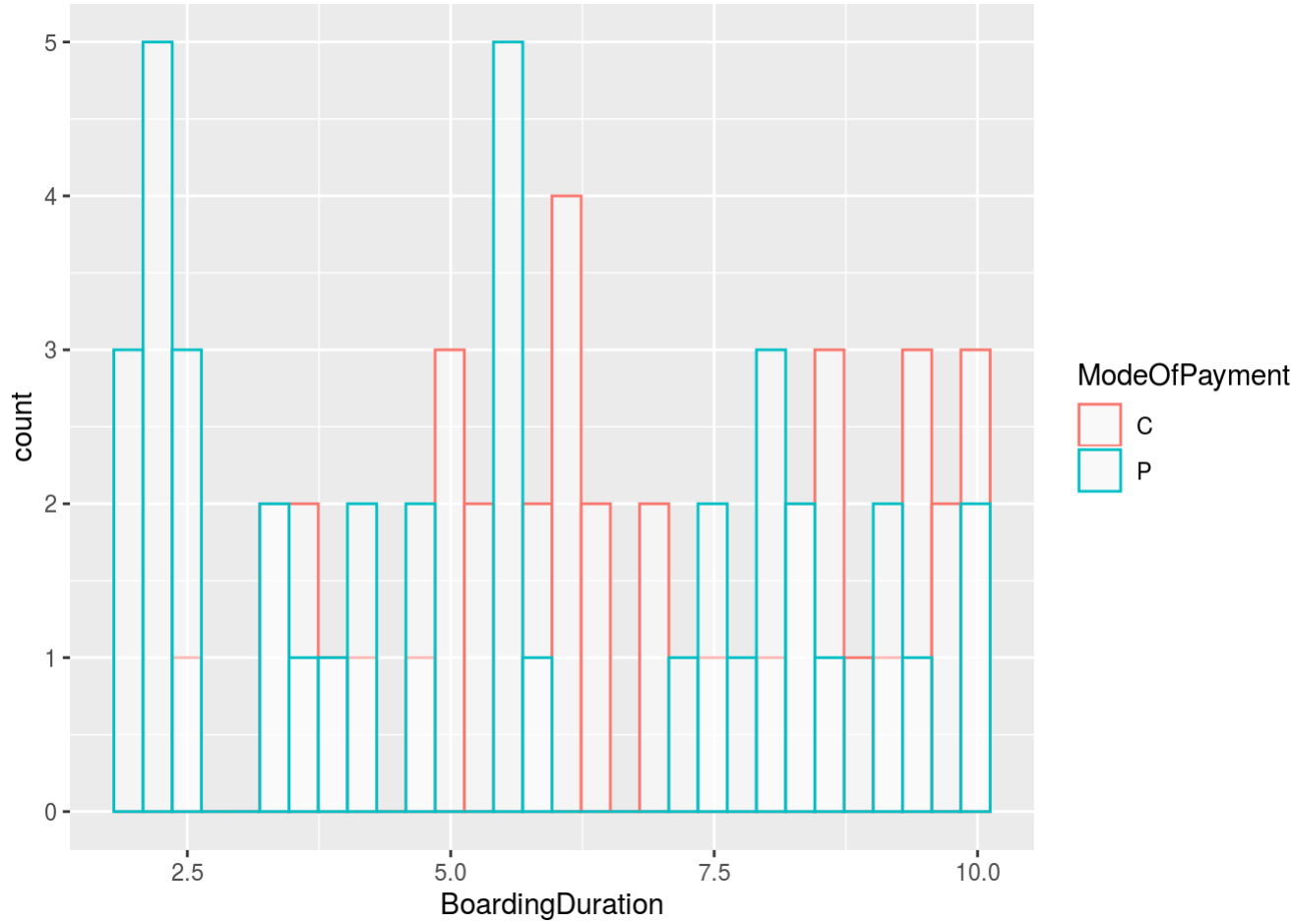

```
# histogram plot by groups
ggplot(sampleData, aes(x=BoardingDuration, color=ModeOfPayment)) +
  geom_histogram(fill="White")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Overlaid histograms
ggplot(sampleData, aes(x=BoardingDuration, color=ModeOfPayment)) +
  geom_histogram(fill="White", alpha=0.5, position="identity")
```

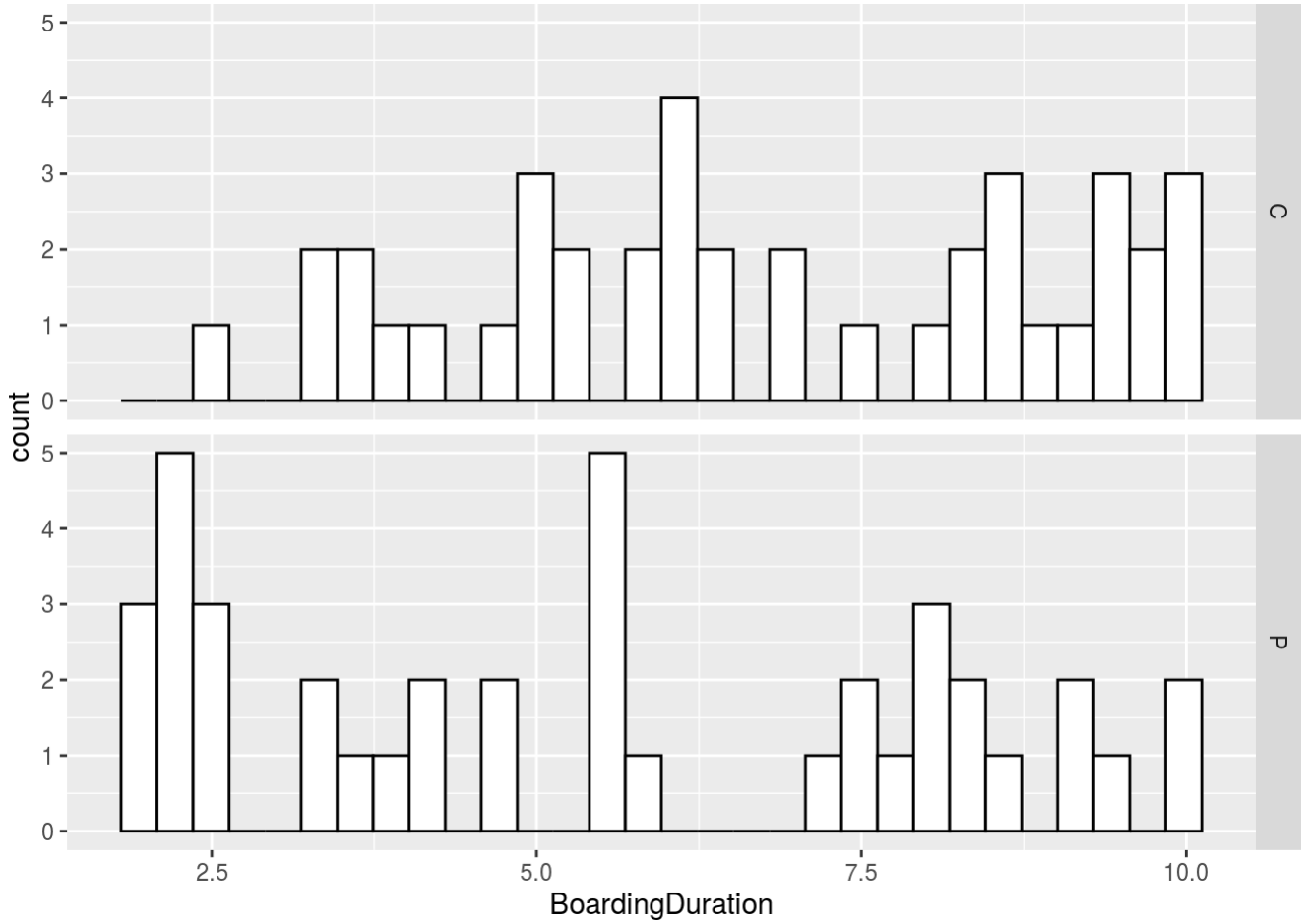
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




Histogram depicting the distribution of Data points based upon each category type.

```
# Boarding Duration vs Mode of Payment
ggplot(sampleData, aes(x=BoardingDuration))+
  geom_histogram(color="black", fill="white")+
  facet_grid(ModeOfPayment ~ .)
```

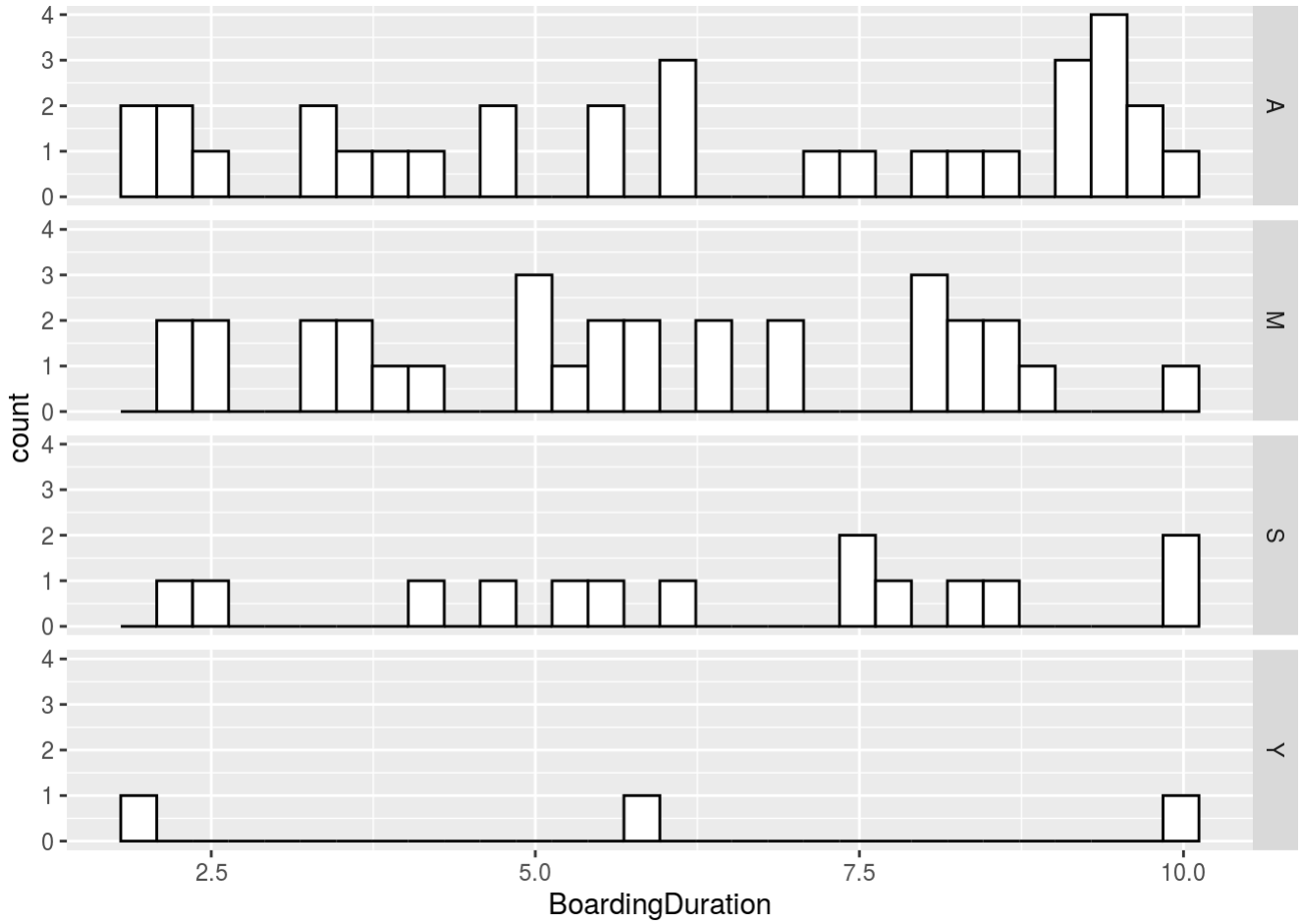
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




Histogram depicting the effect of Mode of Payment on Boarding Duration

```
# Boarding Duration vs Age Groups
ggplot(sampleData, aes(x=BoardingDuration))+
  geom_histogram(color="black", fill="white")+
  facet_grid(AgeGroup ~ .)
```

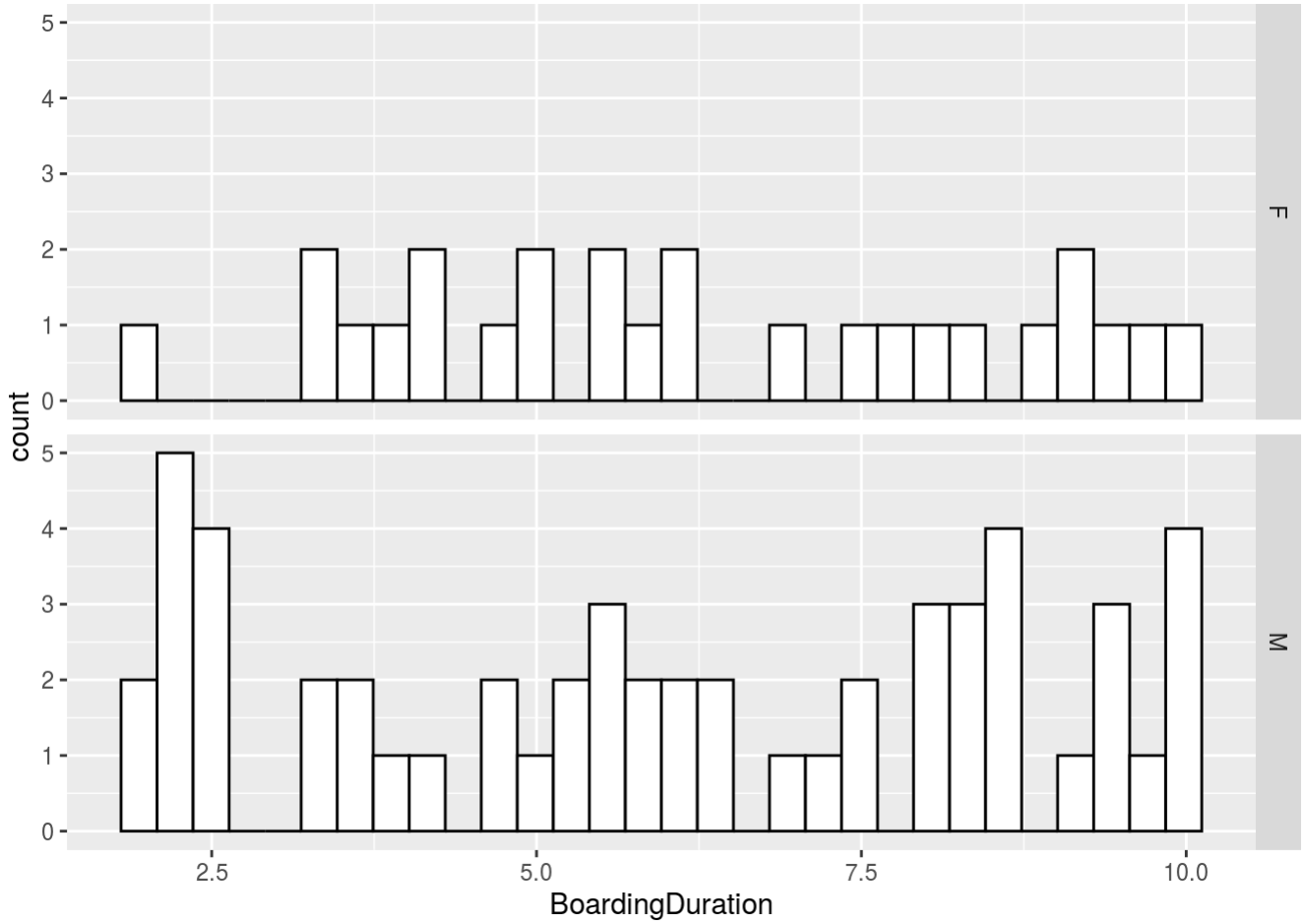
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




Histogram depicting the effect of Age Group on Boarding Duration

```
# Boarding Duration vs Gender
ggplot(sampleData, aes(x=BoardingDuration))+
  geom_histogram(color="black", fill="white")+
  facet_grid(Gender ~ .)
```

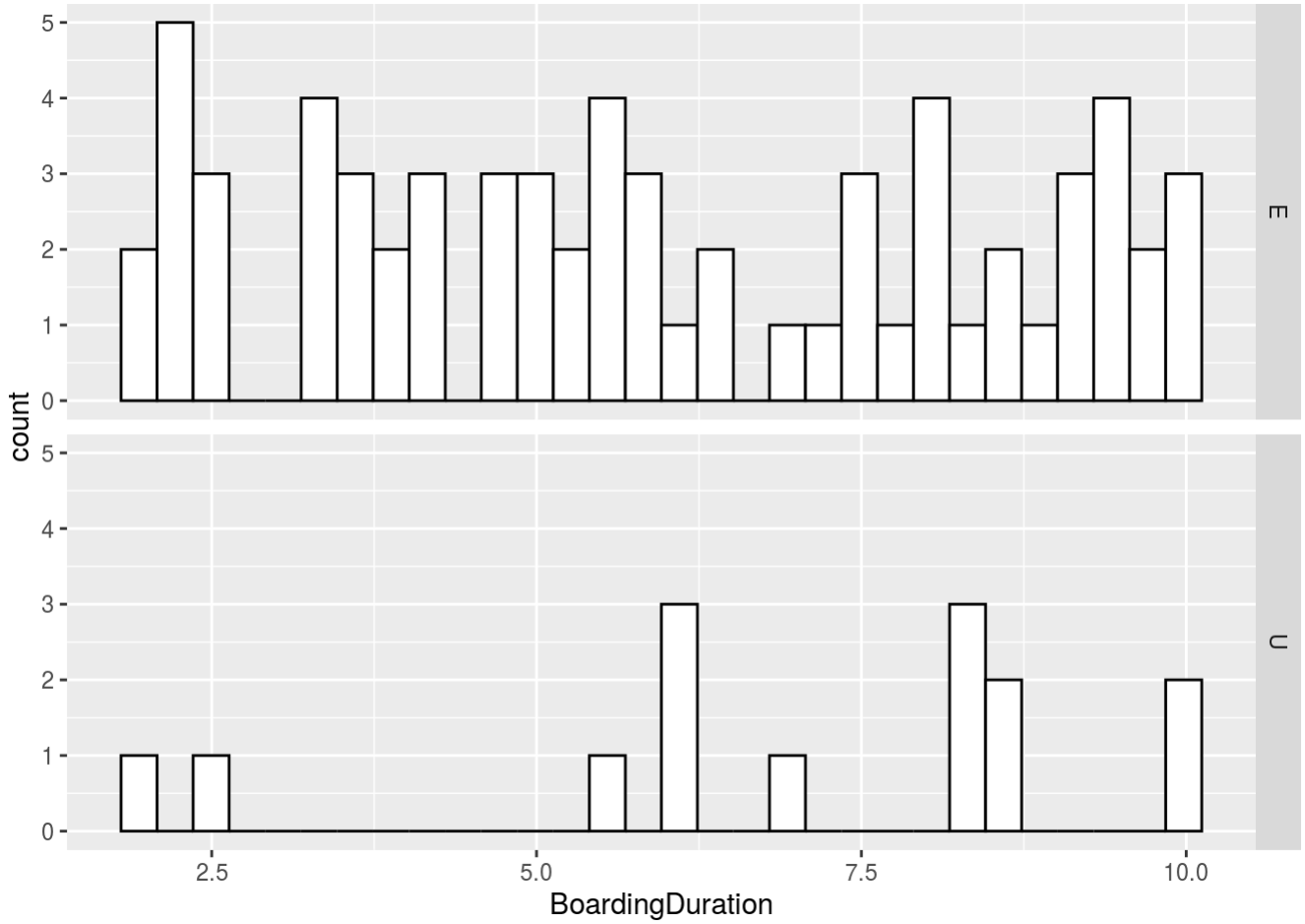
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




Histogram depicting the effect of Gender on Boarding Duration

```
# Boarding Duration vs Employment
ggplot(sampleData, aes(x=BoardingDuration))+
  geom_histogram(color="black", fill="white")+
  facet_grid(Employment ~ .)
```

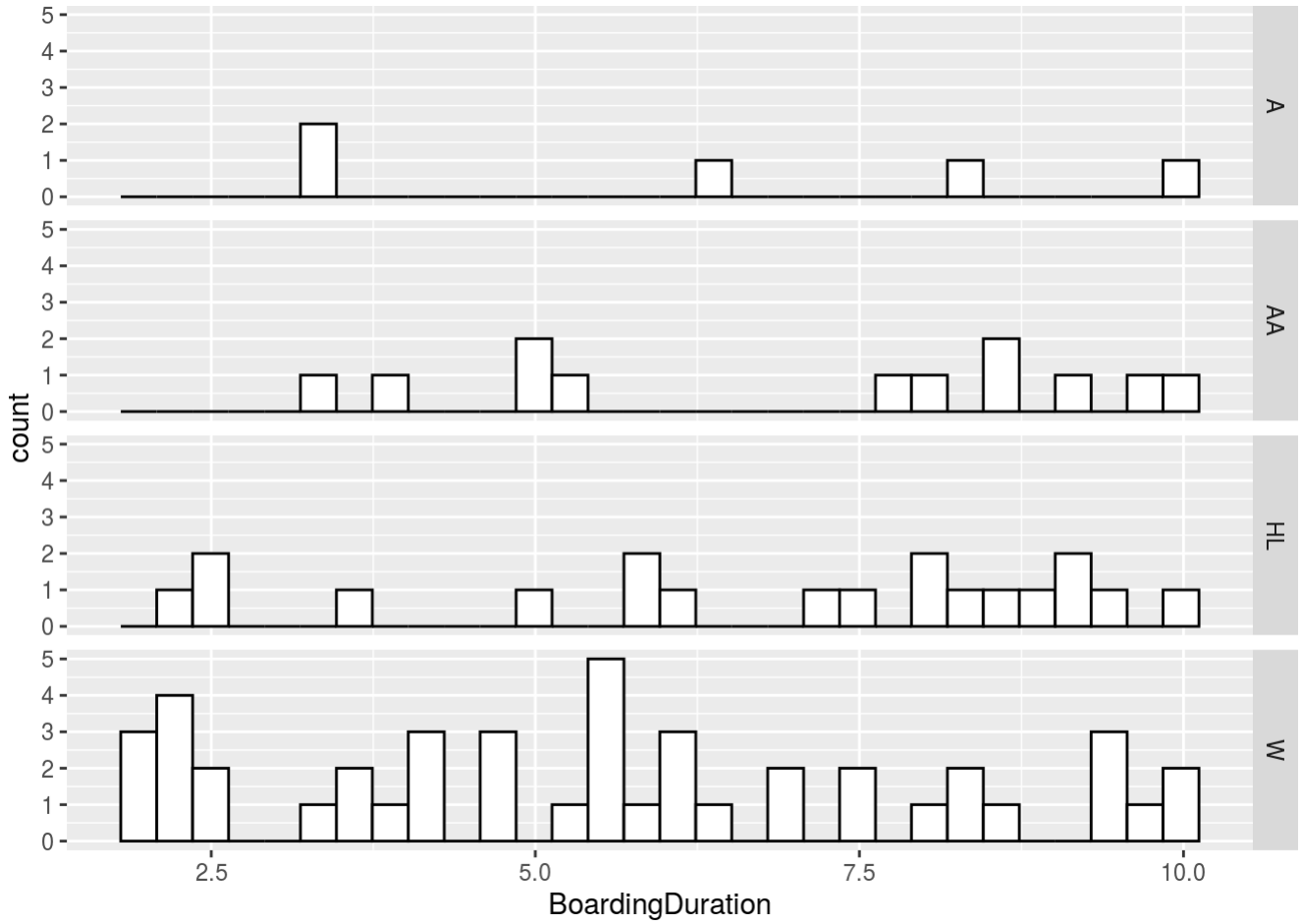
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



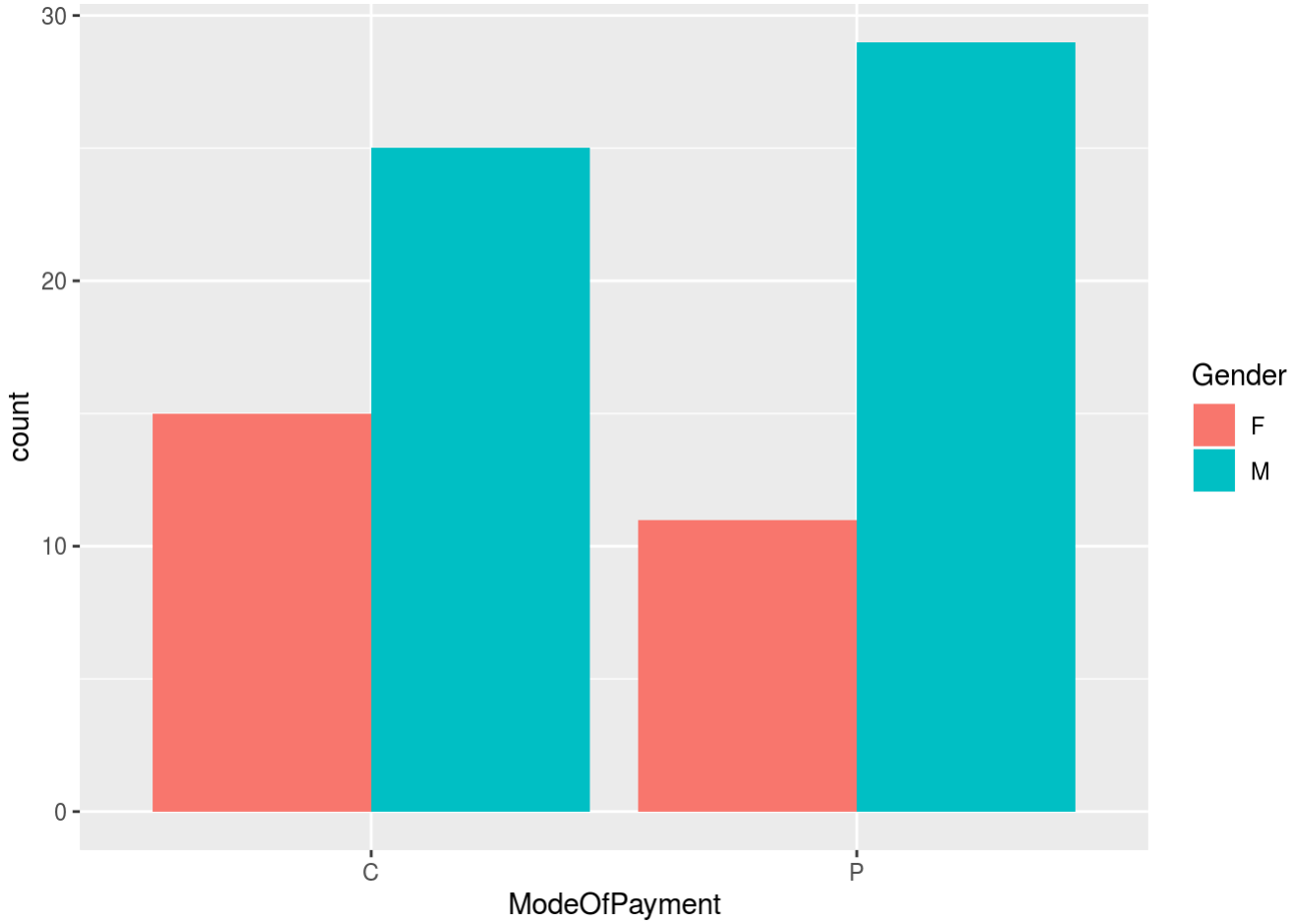

Histogram depicting the effect of Employment status on Boarding Duration

```
# Boarding Duration vs Ethnicity
ggplot(sampleData, aes(x=BoardingDuration))+
  geom_histogram(color="black", fill="white")+
  facet_grid(Ethnicity ~ .)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

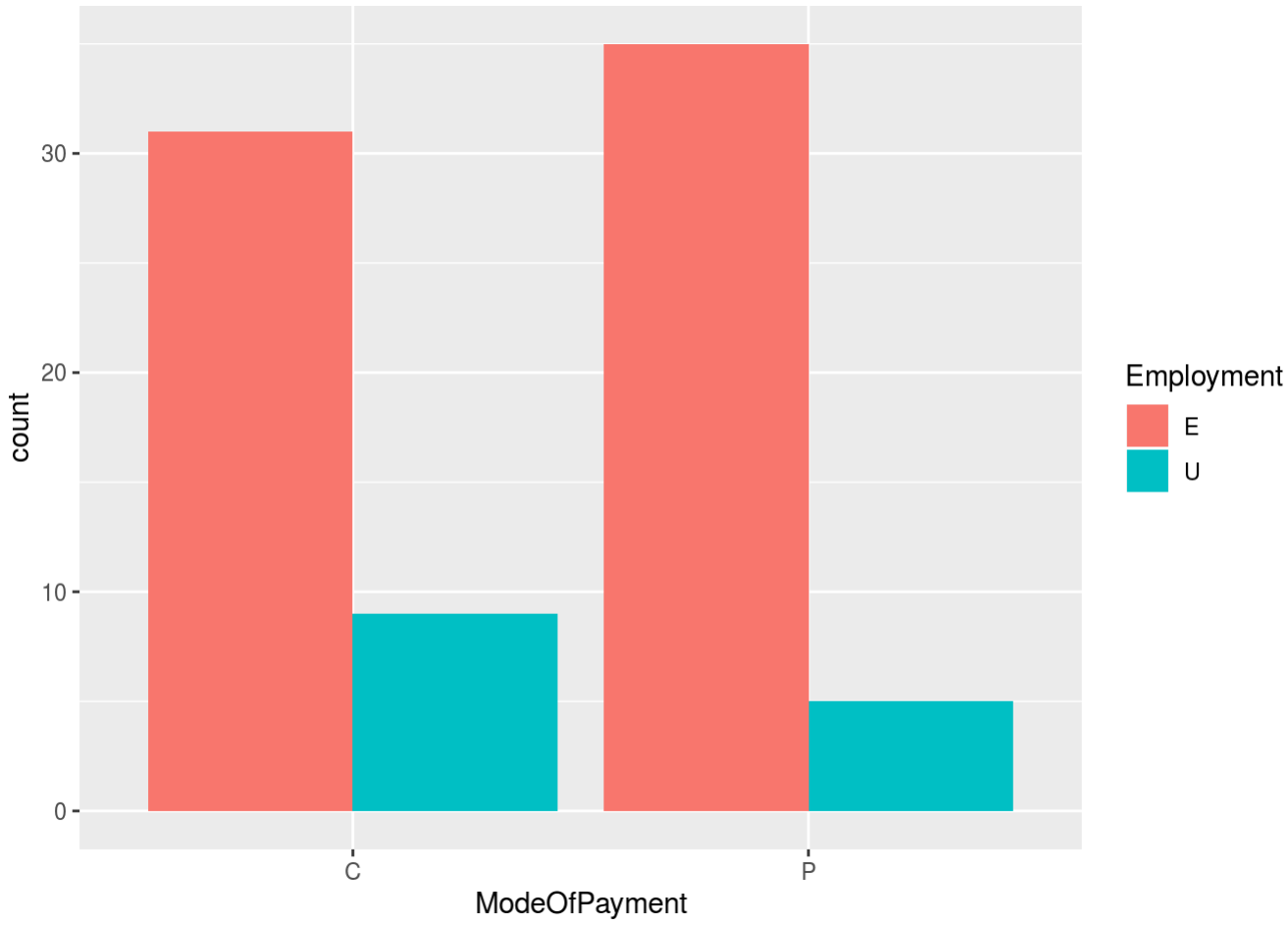



```
# Mode of Payment vs Gender
ggplot(sampleData, aes(ModeOfPayment, ..count..)) + geom_bar(aes(fill = Gender), position = "dod
ge")
```



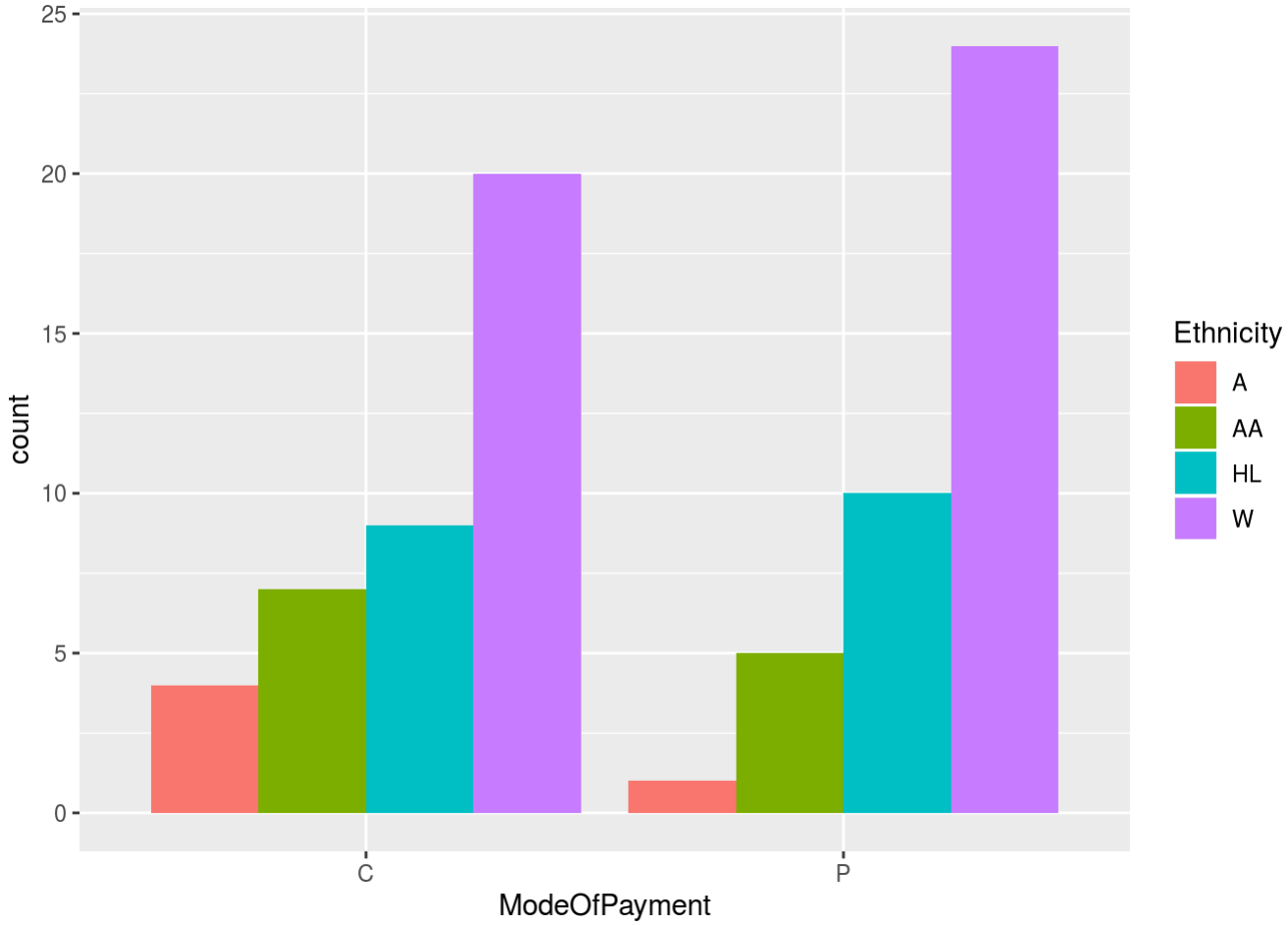

Bar Graph depicting the preference of Both the gender for mode of payment

```
# Mode of Payment vs Employment
ggplot(sampleData, aes(ModeOfPayment, ..count..)) + geom_bar(aes(fill = Employment), position =
"dodge")
```



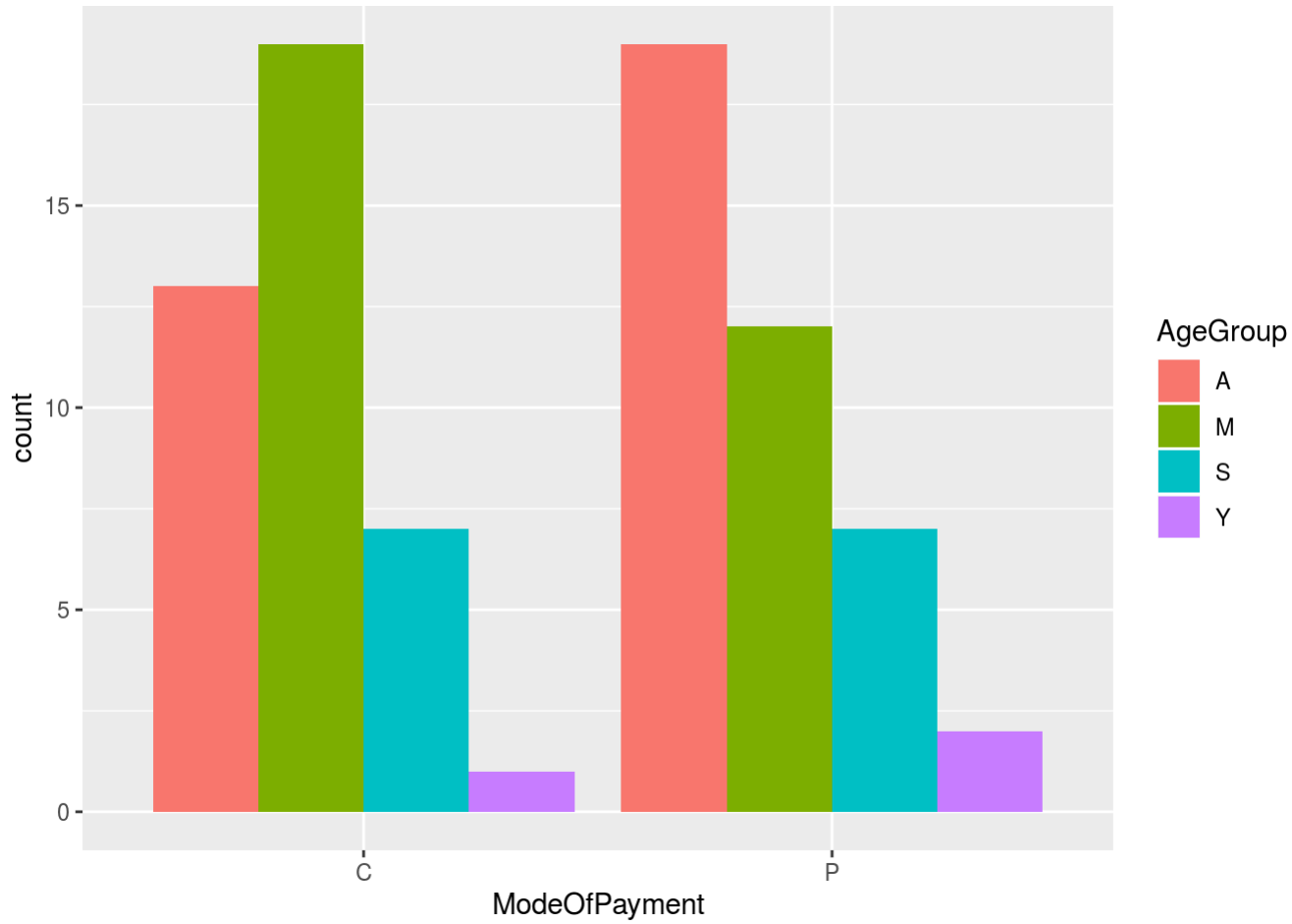

Bar Graph depicting the preference of sectors for mode of payment

```
# Mode of Payment vs Ethnicity
ggplot(sampleData, aes(ModeOfPayment, ..count..)) + geom_bar(aes(fill = Ethnicity), position =
"dodge")
```




Bar Graph depicting the preference of individuals from different ethnical background for mode of payment

```
# Mode of Payment vs Age Group
ggplot(sampleData, aes(ModeOfPayment, ..count..)) + geom_bar(aes(fill = AgeGroup), position = "d
odge")
```



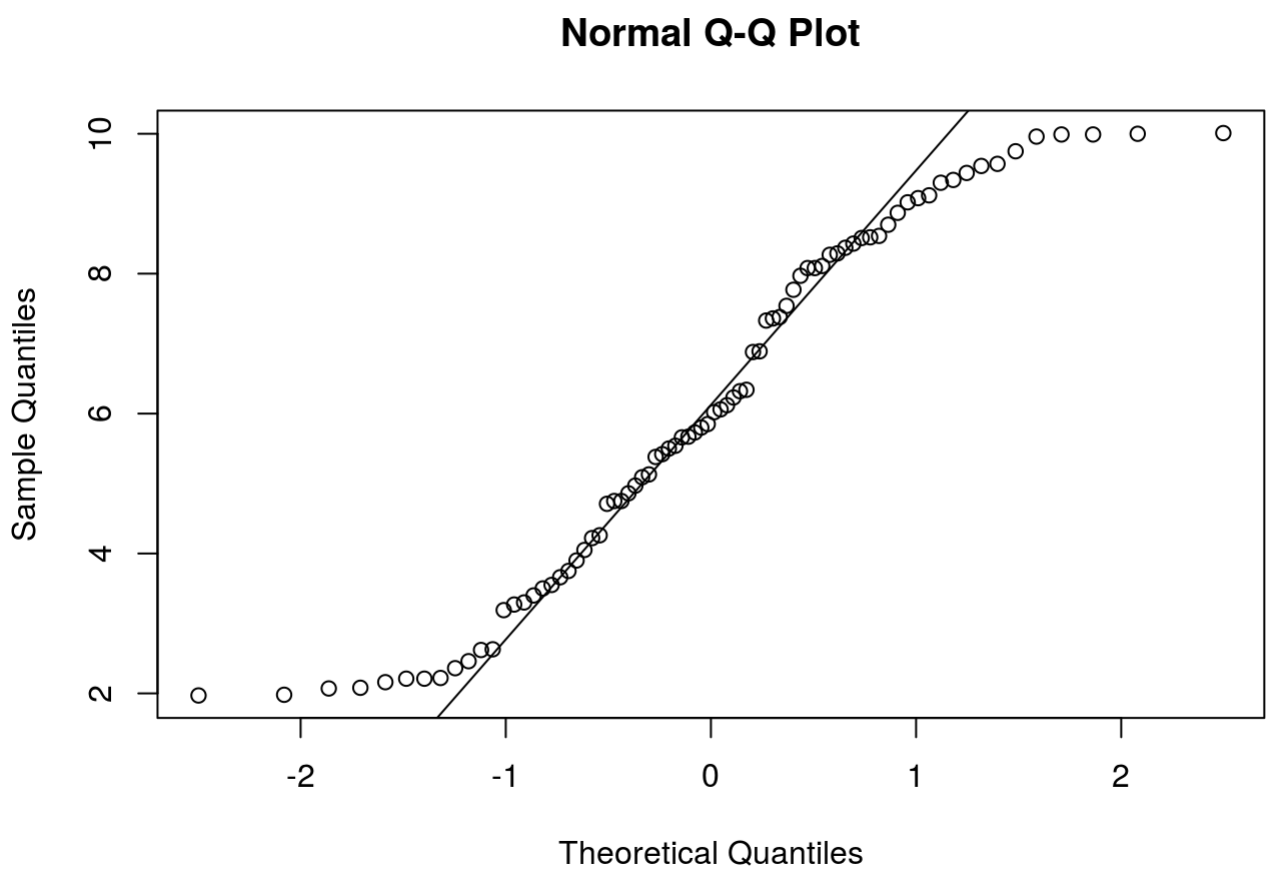

Bar Graph depicting the preference of age groups for mode of payment

# Statistical Analysis

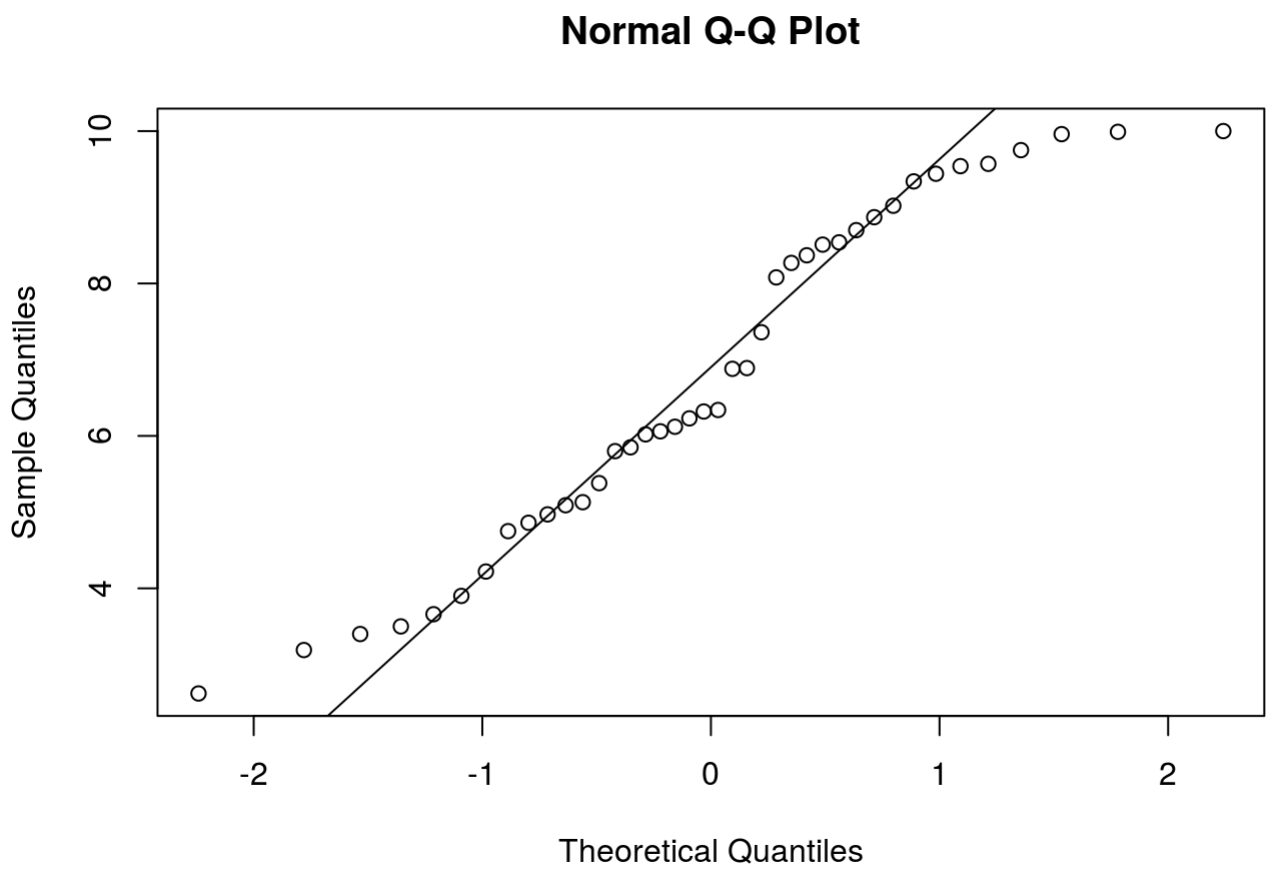
## 1. Visualizing Sample Data

A Normal Q-Q Plot helps us to visualize whether our sample data is normally distributed or not. In real word scenarios no dataset is perfectly normally distribution. In a perfect Normal Distribution the Q-Q Plot just represent the $Y = X$ Line. In real-life the line would be slightly curved over some data points.

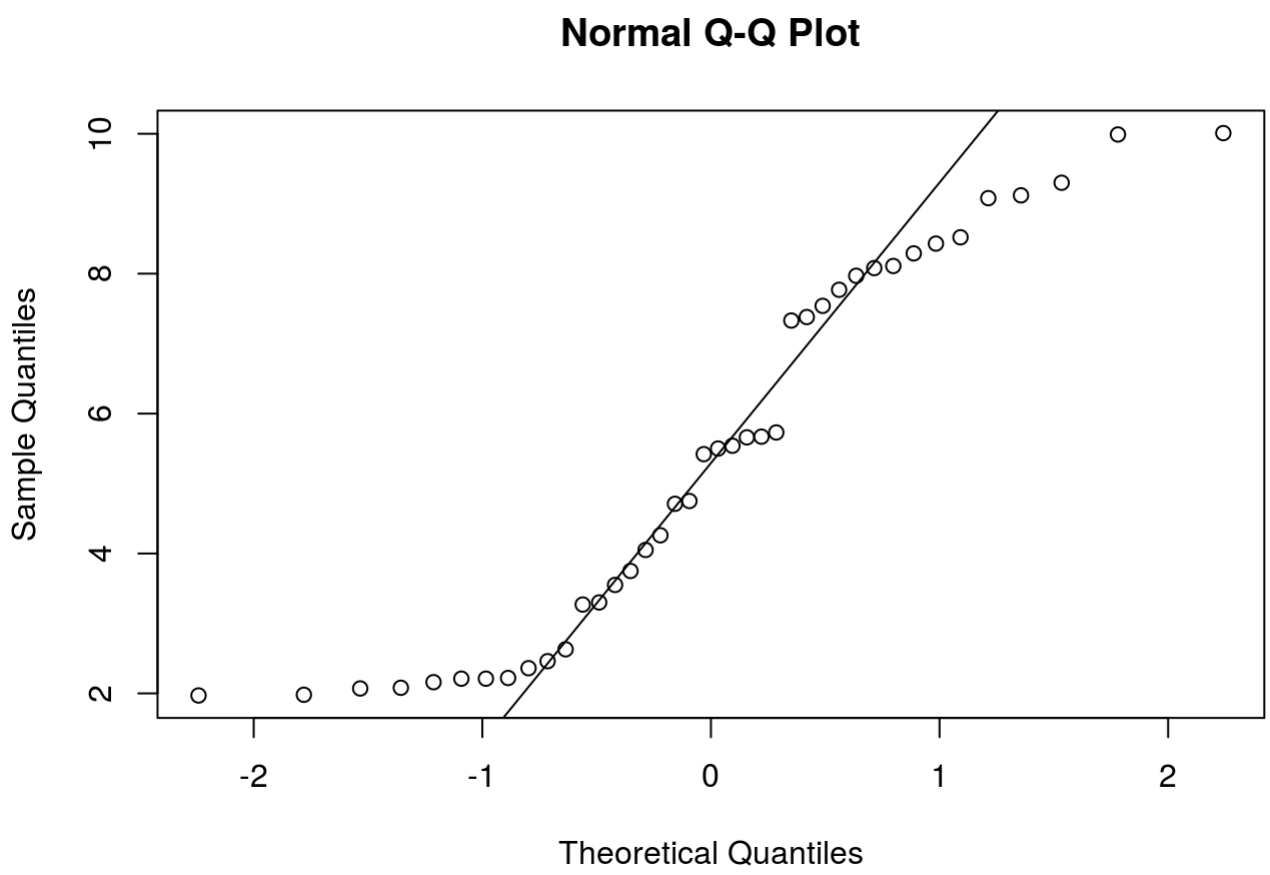
```
# Q-Q Plot of Sample Data
qqnorm(sampleData$BoardingDuration)
qqline(sampleData$BoardingDuration)
```




```
# Q-Q Plot of Passengers using Cash in the Sample Data
qqnorm(sampleData$BoardingDuration[sampleData$ModeOfPayment == "C"])
qqline(sampleData$BoardingDuration[sampleData$ModeOfPayment == "C"])
```




```
# Q-Q Plot of Passengers using Pass in the Sample Data
qqnorm(sampleData$BoardingDuration[sampleData$ModeOfPayment == "P"])
qqline(sampleData$BoardingDuration[sampleData$ModeOfPayment == "P"])
```




The above graphs may not be perfectly normally distributed but it can be said that they are close to that of a normal distribution with slight curves.

# 2. Statistical Testing

## Hypothesis:

Null Hypothesis: The true population mean Boarding Duration of Commuters using Cash is equal to the true population mean Boarding Duartion for those who use Pass.

$$H_0 : \mu_c - \mu_p = 0 \; or \; \mu_c = \mu_p$$

Alternate Hypothesis:

A1. The true population mean Boarding Duration of Commuters using Cash is not equal to (or different than that of) the true population mean Boarding Duartion for those who use Pass.

$$H_{A1} : \mu_c - \mu_p \neq 0 \; or \; \mu_c \neq \mu_p$$

A2. The true population mean Boarding Duration of Commuters using Cash is greater than that of the true population mean Boarding Duartion for those who use Pass.

$$H_{A2} : \mu_c - \mu_p > 0 \; or \; \mu_c > \mu_p$$

## Sample Statistic: Difference in Means

$$\overline{x}_c - \overline{x}_p$$

## Test-statistic

$$t = \frac{(\overline{x}_c - \overline{x}_p) - (\mu_c - \mu_p)}{\sqrt{\frac{\sigma_c^2}{n_c} + \frac{\sigma_p^2}{n_p}}}$$

$$\mu_0 = \mu_c - \mu_p = 0$$

## Two Sample t-test

Calculating p-value using R in-built function t.test()

```
# Two sided t-test
t.test(sampleData$BoardingDuration~sampleData$ModeOfPayment)
```

```
##
##  Welch Two Sample t-test
##
## data:  sampleData$BoardingDuration by sampleData$ModeOfPayment
## t = 2.4445, df = 75.408, p-value = 0.01684
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2501977 2.4528023
## sample estimates:
## mean in group C mean in group P
##         6.76225         5.41075
```

Calculating p-value manually using given formulas.

```
# Mean Boarding Time of Passengers using Cash
mu_c <- mean(sampleData$BoardingDuration[sampleData$ModeOfPayment == 'C'])
mu_c
```

```
## [1] 6.76225
```

```
# Mean Boarding Time of Passengers using Pass
mu_p <- mean(sampleData$BoardingDuration[sampleData$ModeOfPayment == 'P'])
mu_p
```

```
## [1] 5.41075
```

```
# Null Hypothesis
mu_0 <- 0
```

```
# Variance of Boarding Time of Passengers using Cash
var_c <- var(sampleData$BoardingDuration[sampleData$ModeOfPayment == 'C'])
var_c
```

```
## [1] 4.980228
```

```
# Variance of Boarding Time of Passengers using Pass
var_p <- var(sampleData$BoardingDuration[sampleData$ModeOfPayment == 'P'])
var_p
```

```
## [1] 7.246971
```

```
# Sample Size of Passengers using Cash
n_c <- length(sampleData$BoardingDuration[sampleData$ModeOfPayment == 'C'])
```

```
# Sample Size of Passengers using Pass
n_p <- length(sampleData$BoardingDuration[sampleData$ModeOfPayment == 'P'])
```

```
# t-value (test statistic)
t <- (mu_c - mu_p - mu_0)/sqrt(var_c/n_c + var_p/n_p)
t
```

```
## [1] 2.444458
```

```
# p-value for 2 sided t-test
p_value <- pt(q = t, df = min(n_c, n_p) - 1, lower.tail = FALSE)*2
p_value
```

```
## [1] 0.01912974
```

```
# Lower Boundary of Confidence Interval
lowerBound <- mu_c - mu_p + qt(0.05, min(n_c, n_p) - 1)*sqrt(var_c/n_c + var_p/n_p)
lowerBound
```

```
## [1] 0.4199606
```

```
# Upper Boundary of Confidence Interval
upperBound <- mu_c - mu_p + qt(0.95, min(n_c, n_p) - 1)*sqrt(var_c/n_c + var_p/n_p)
upperBound
```

```
## [1] 2.283039
```

```
# Sample Statistic
ss <- mu_c - mu_p
ss
```

```
## [1] 1.3515
```

## P-value for Alternate Hypothesis 2

### One-sided t-test (Upper t-test)

```
# Upper t-test
t.test(sampleData$BoardingDuration[sampleData$ModeOfPayment == 'C'], sampleData$BoardingDuration
[sampleData$ModeOfPayment == 'P'], alternative = "greater")
```

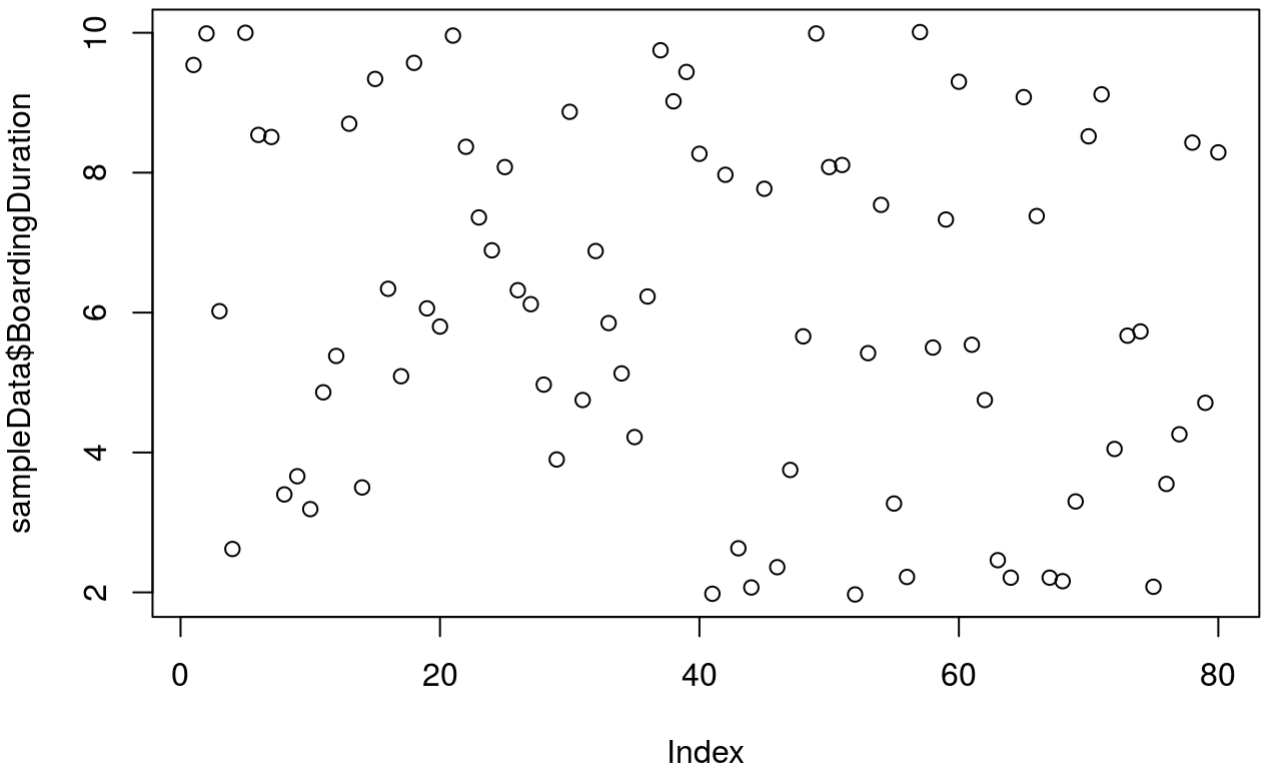
```
##
##  Welch Two Sample t-test
##
## data:  sampleData$BoardingDuration[sampleData$ModeOfPayment == "C"] and sampleData$BoardingDu
ration[sampleData$ModeOfPayment == "P"]
## t = 2.4445, df = 75.408, p-value = 0.008421
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4307764       Inf
## sample estimates:
## mean of x mean of y
##   6.76225   5.41075
```

# 3. Results

### Interpretation

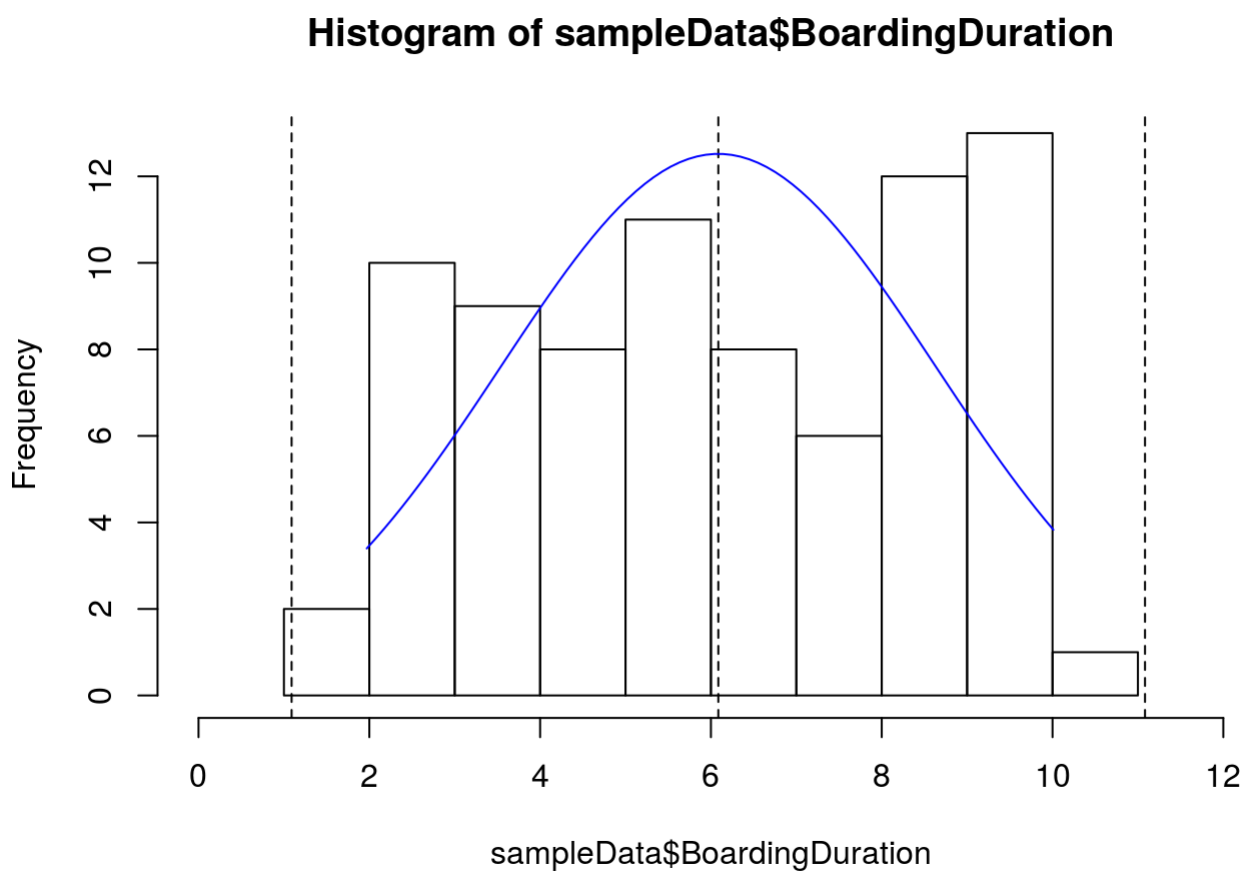
Since the p-value is 0.01912974 which is less than 0.05, there is strong evidence for us to suggest that true population mean Boarding Duration of Commuters using Cash is different from the true population mean Boarding Duration for those who use Pass (i.e, true difference in means is not equal to 0). We reject the Null Hypothesis that the mean Boarding Duration of Commuters using Cash and Pass is the same at the alpha = 0.05 level. We can say that the true difference in mean Boarding Duration between those using Cash and those using Pass is between 0.4199606 and 2.283039 with a confidence of 95%. The null hypothesized difference between the mean Boarding Duration is zero is not in the 95% confidence interval which is consistent with the rejection of the null hypothesis and the values of the confidence interval suggest that on average those using Cash for purchasing tickets in Bus have a greater Boarding Duration than those of using Pass for purchasing tickets in Bus.

```
# Scatterplot of Data points representing Random Sampling and Randomness in data.
plot(sampleData$BoardingDuration)
```

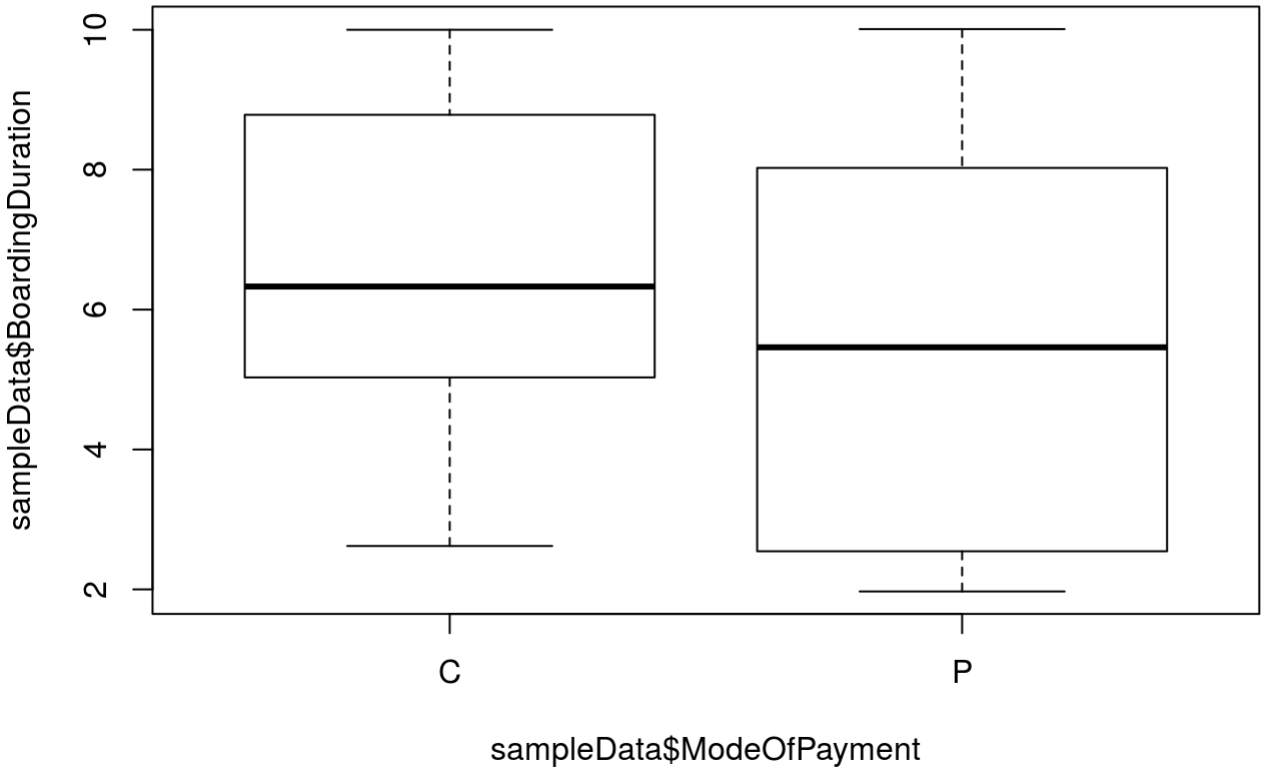



This graph shows the randomness in the sample data and proves that methodology to be Simple Random Sampling.

```
# Histogram of Sampling Distribution
mu <- mean(sampleData$BoardingDuration)
sd <- sd(sampleData$BoardingDuration)
h <- hist(sampleData$BoardingDuration, xlim = c(0,12))
lb <- mu - 1.96*sd
ub <- mu + 1.96*sd
abline(v = c(mu, lb, ub), lty = 2)
xx <- seq(min(sampleData$BoardingDuration),max(sampleData$BoardingDuration),length=80)
yy <- dnorm(xx, mu, sd)*length(xx)
lines(xx, yy, col = "blue")
```

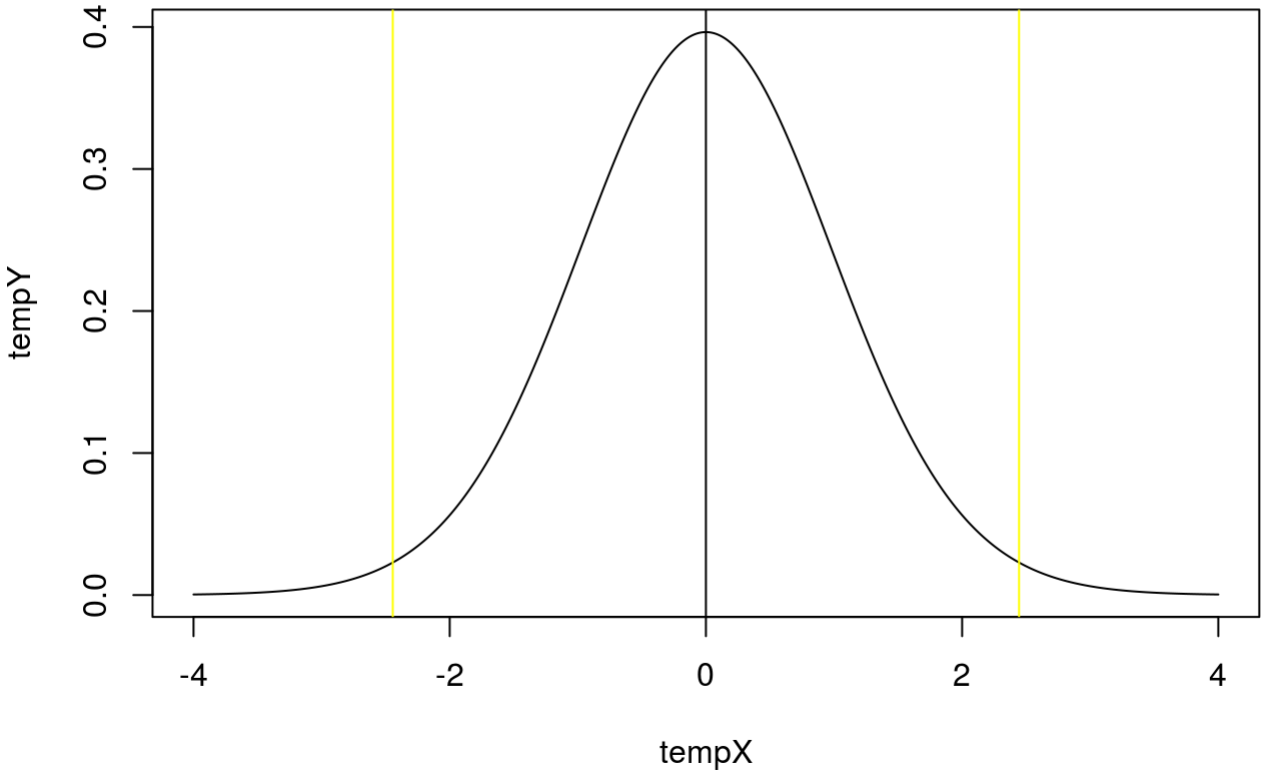



```
# Boarding Time Intervals for Both Categories
plot(sampleData$BoardingDuration~sampleData$ModeOfPayment)
```

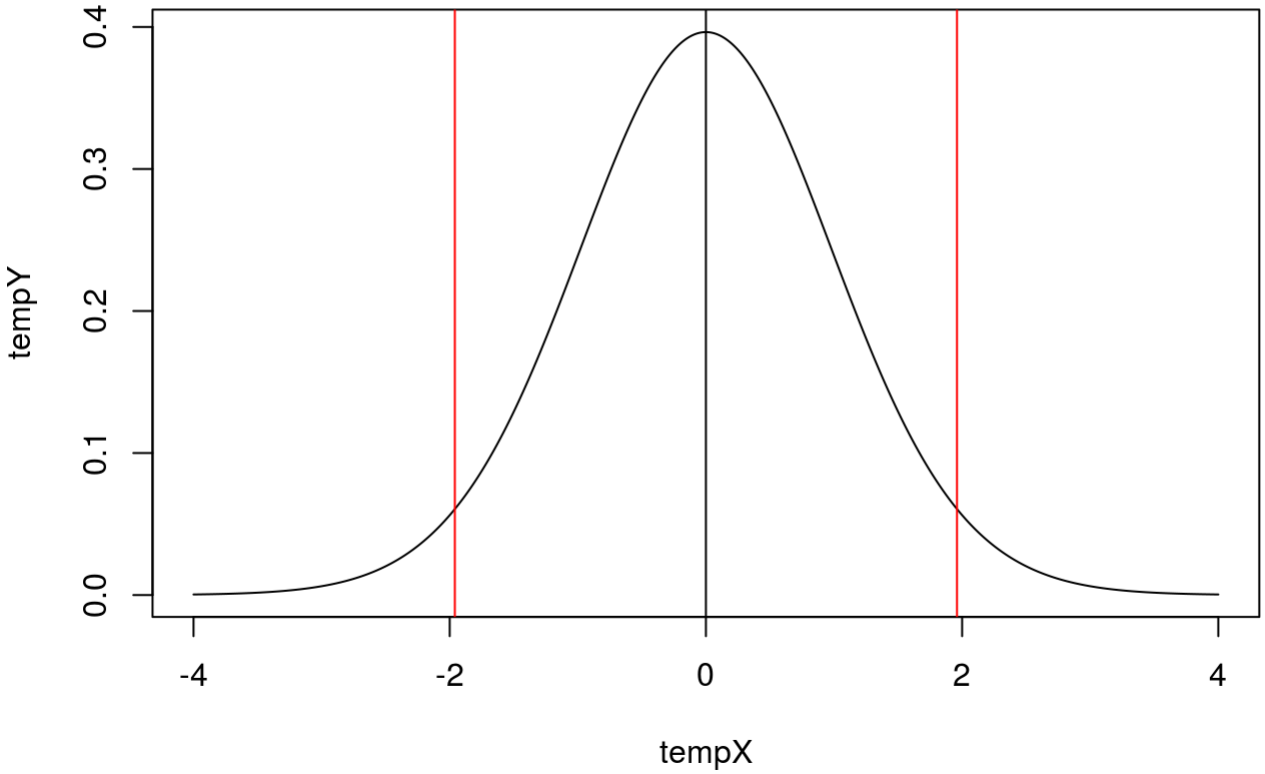



This shows that the mean of Boarding Time of those using Cash is greater than (not equal to) the Boarding Time of those using Pass

```
# test statistic graph
n <- min(n_c, n_p)
tempX <- seq(-4, 4, .01)
tempY <- dt(tempX, n-1)
plot(tempX, tempY, type = 'l')
abline(v = c(t, -t),  col = "yellow")
abline(v = 0, col = "black")
```

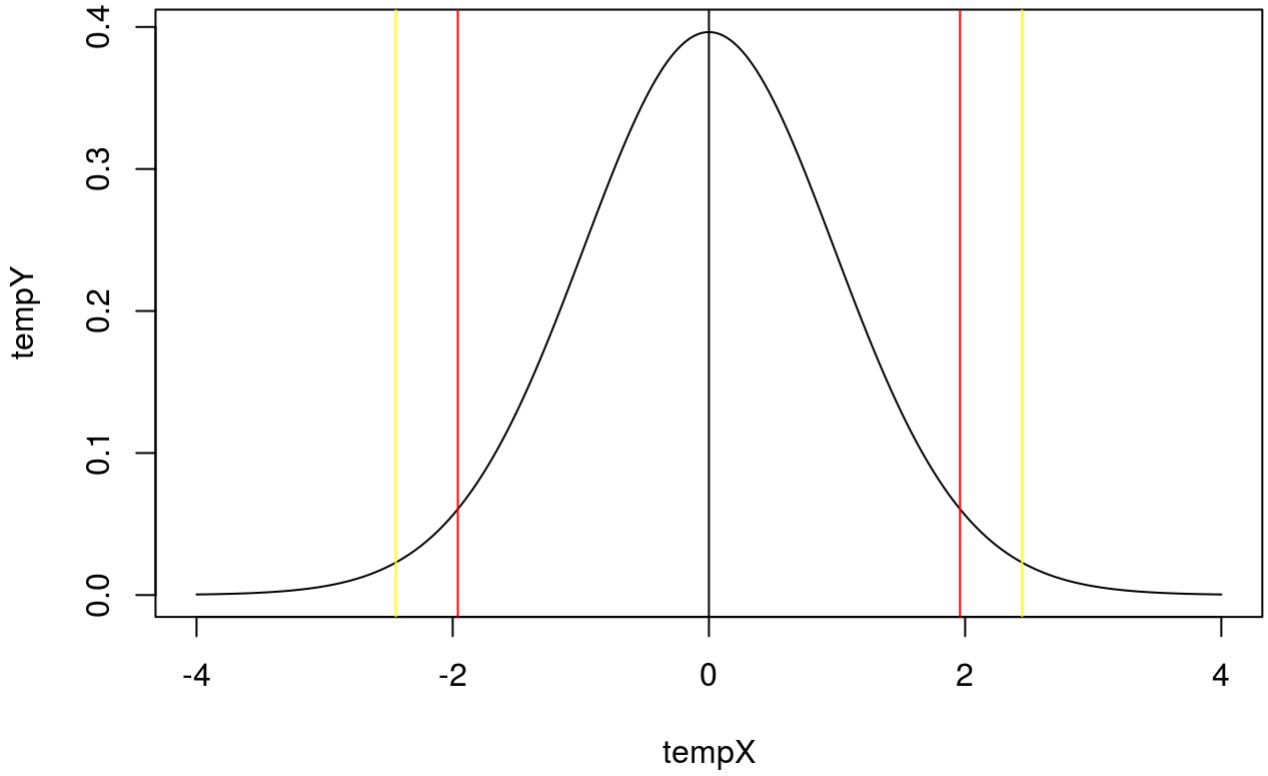



```
# Confidence Intervals graph
plot(tempX, tempY, type = 'l')
abline(v = qnorm(0.975), col = "red")
abline(v = qnorm(0.025), col = "red")
abline(v = 0, col = "black")
```



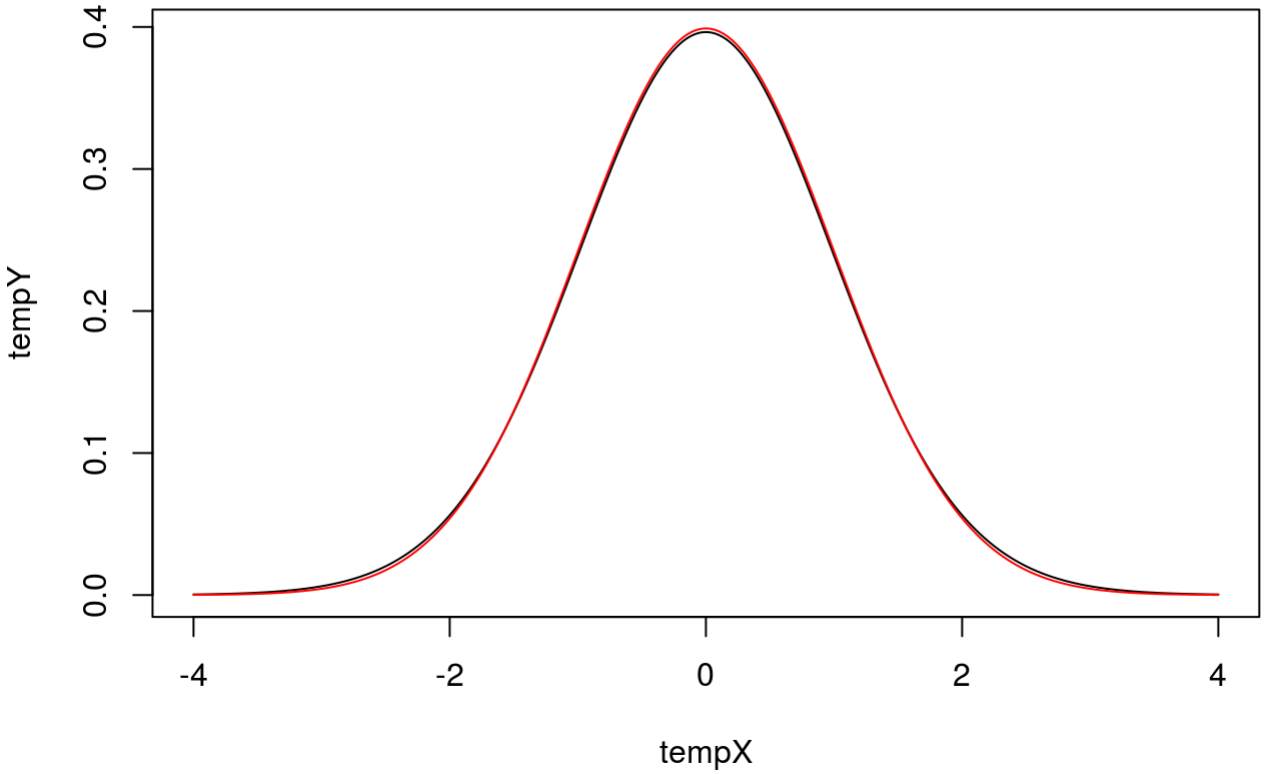

This graph interprets the confidence interval. The area under the curve between the red lines is our 95% confidence interval. These intervals are symmetric over the line X = 0.

```
plot(tempX, tempY, type = 'l')
abline(v = qnorm(0.975), col = "red")
abline(v = qnorm(0.025), col = "red")
abline(v = 0, col = "black")
abline(v = c(t, -t),  col = "yellow")
```




```
# confidence interval shading
# plot(tempX, tempY, type = 'l')
# polygon(c(-1.96, tempX, 1.96), c(0, tempY, 0), col="red")
```

```
# T-distribution vs Normal distribution
plot(tempX, tempY, type = 'l')
lines(tempX, dnorm(tempX), col = "red")
```




This graphs shows the deviation of our T-Distribution data from that of a Normal Distribution (which is the red line).

# 4. Appendix

```
knitr::opts_chunk$set(echo = TRUE)
# intalling packages
install.packages("ggplot2")
install.packages("dplyr")
# importing libraries
library("ggplot2")
library("dplyr")
# importing data from CSV to a Dataframe
myData <- read.csv(file = "dataset.csv")
head(myData)
# Entire collected data
myData
# Dataset Summary
summary(myData)
# Choosing a random seed number to introduce randomization in sampling process.
set.seed(7384)
# Sample data of 40 passengers using Cash for ticket payment (40 out of 42 data points).
cashData <- filter(myData, myData$ModeOfPayment == 'C')
cashDataSample <- cashData[sample(nrow(cashData), 40), ]
cashDataSample
# Cash Data Statistics
summary(cashDataSample)
# Choosing a random seed number to introduce randomization in sampling process.
set.seed(4250)
# Sample data of 40 passengers using Pass for ticket payment (40 out of 108 data points).
passData <- filter(myData, myData$ModeOfPayment == 'P')
passDataSample <- passData[sample(nrow(passData), 40), ]
passDataSample
# Pass Data Statistics
summary(passDataSample)
# Sampled Data chosen for Statistical Analysis.
sampleData <- rbind(cashDataSample, passDataSample)
head(sampleData)
# Sampled Data
sampleData
# Sample Data Summary
summary(sampleData)
# graphs of all pairs of variables
pairs(sampleData)
ggplot(sampleData,aes(x=BoardingDuration,y=Gender,color=ModeOfPayment))+geom_point()+ylab('Gende
r')+xlab('Boarding Duration')+ ggtitle('Mode of Payment')
# Box plot for Mode of Payment vs Boarding Duration
ggplot(sampleData, aes(x = ModeOfPayment, y = BoardingDuration)) + geom_boxplot() + coord_flip()

# histogram with fill color
ggplot(sampleData, aes(x=BoardingDuration))+
  geom_histogram(color="darkblue", fill="lightblue")
# dashed line
ggplot(sampleData, aes(x=BoardingDuration))+
  geom_histogram(color="black", fill="lightblue",
                 linetype="dashed")
# histogram plot by groups
ggplot(sampleData, aes(x=BoardingDuration, color=ModeOfPayment)) +
```

```
  geom_histogram(fill="White")
# Overlaid histograms
ggplot(sampleData, aes(x=BoardingDuration, color=ModeOfPayment)) +
  geom_histogram(fill="White", alpha=0.5, position="identity")
# Boarding Duration vs Mode of Payment
ggplot(sampleData, aes(x=BoardingDuration))+
  geom_histogram(color="black", fill="white")+
  facet_grid(ModeOfPayment ~ .)
# Boarding Duration vs Age Groups
ggplot(sampleData, aes(x=BoardingDuration))+
  geom_histogram(color="black", fill="white")+
  facet_grid(AgeGroup ~ .)
# Boarding Duration vs Gender
ggplot(sampleData, aes(x=BoardingDuration))+
  geom_histogram(color="black", fill="white")+
  facet_grid(Gender ~ .)
# Boarding Duration vs Employment
ggplot(sampleData, aes(x=BoardingDuration))+
  geom_histogram(color="black", fill="white")+
  facet_grid(Employment ~ .)
# Boarding Duration vs Ethnicity
ggplot(sampleData, aes(x=BoardingDuration))+
  geom_histogram(color="black", fill="white")+
  facet_grid(Ethnicity ~ .)
# Mode of Payment vs Gender
ggplot(sampleData, aes(ModeOfPayment, ..count..)) + geom_bar(aes(fill = Gender), position = "dod
ge")
# Mode of Payment vs Employment
ggplot(sampleData, aes(ModeOfPayment, ..count..)) + geom_bar(aes(fill = Employment), position =
"dodge")
# Mode of Payment vs Ethnicity
ggplot(sampleData, aes(ModeOfPayment, ..count..)) + geom_bar(aes(fill = Ethnicity), position =
"dodge")
# Mode of Payment vs Age Group
ggplot(sampleData, aes(ModeOfPayment, ..count..)) + geom_bar(aes(fill = AgeGroup), position = "d
odge")
# Q-Q Plot of Sample Data
qqnorm(sampleData$BoardingDuration)
qqline(sampleData$BoardingDuration)
# Q-Q Plot of Passengers using Cash in the Sample Data
qqnorm(sampleData$BoardingDuration[sampleData$ModeOfPayment == "C"])
qqline(sampleData$BoardingDuration[sampleData$ModeOfPayment == "C"])
# Q-Q Plot of Passengers using Pass in the Sample Data
qqnorm(sampleData$BoardingDuration[sampleData$ModeOfPayment == "P"])
qqline(sampleData$BoardingDuration[sampleData$ModeOfPayment == "P"])
# Two sided t-test
t.test(sampleData$BoardingDuration~sampleData$ModeOfPayment)
# Mean Boarding Time of Passengers using Cash
mu_c <- mean(sampleData$BoardingDuration[sampleData$ModeOfPayment == 'C'])
mu_c
# Mean Boarding Time of Passengers using Pass
mu_p <- mean(sampleData$BoardingDuration[sampleData$ModeOfPayment == 'P'])
mu_p
# Null Hypothesis
```

```
mu_0 <- 0
# Variance of Boarding Time of Passengers using Cash
var_c <- var(sampleData$BoardingDuration[sampleData$ModeOfPayment == 'C'])
var_c
# Variance of Boarding Time of Passengers using Pass
var_p <- var(sampleData$BoardingDuration[sampleData$ModeOfPayment == 'P'])
var_p
# Sample Size of Passengers using Cash
n_c <- length(sampleData$BoardingDuration[sampleData$ModeOfPayment == 'C'])
# Sample Size of Passengers using Pass
n_p <- length(sampleData$BoardingDuration[sampleData$ModeOfPayment == 'P'])
# t-value (test statistic)
t <- (mu_c - mu_p - mu_0)/sqrt(var_c/n_c + var_p/n_p)
t
# p-value for 2 sided t-test
p_value <- pt(q = t, df = min(n_c, n_p) - 1, lower.tail = FALSE)*2
p_value
# Lower Boundary of Confidence Interval
lowerBound <- mu_c - mu_p + qt(0.05, min(n_c, n_p) - 1)*sqrt(var_c/n_c + var_p/n_p)
lowerBound
# Upper Boundary of Confidence Interval
upperBound <- mu_c - mu_p + qt(0.95, min(n_c, n_p) - 1)*sqrt(var_c/n_c + var_p/n_p)
upperBound
# Sample Statistic
ss <- mu_c - mu_p
ss
# Upper t-test
t.test(sampleData$BoardingDuration[sampleData$ModeOfPayment == 'C'], sampleData$BoardingDuration
[sampleData$ModeOfPayment == 'P'], alternative = "greater")
# Scatterplot of Data points representing Random Sampling and Randomness in data.
plot(sampleData$BoardingDuration)
# Histogram of Sampling Distribution
mu <- mean(sampleData$BoardingDuration)
sd <- sd(sampleData$BoardingDuration)
h <- hist(sampleData$BoardingDuration, xlim = c(0,12))
lb <- mu - 1.96*sd
ub <- mu + 1.96*sd
abline(v = c(mu, lb, ub), lty = 2)
xx <- seq(min(sampleData$BoardingDuration),max(sampleData$BoardingDuration),length=80)
yy <- dnorm(xx, mu, sd)*length(xx)
lines(xx, yy, col = "blue")
# Boarding Time Intervals for Both Categories
plot(sampleData$BoardingDuration~sampleData$ModeOfPayment)
# test statistic graph
n <- min(n_c, n_p)
tempX <- seq(-4, 4, .01)
tempY <- dt(tempX, n-1)
plot(tempX, tempY, type = 'l')
abline(v = c(t, -t),  col = "yellow")
abline(v = 0, col = "black")
# Confidence Intervals graph
plot(tempX, tempY, type = 'l')
abline(v = qnorm(0.975), col = "red")
abline(v = qnorm(0.025), col = "red")
```

```
abline(v = 0, col = "black")
plot(tempX, tempY, type = 'l')
abline(v = qnorm(0.975), col = "red")
abline(v = qnorm(0.025), col = "red")
abline(v = 0, col = "black")
abline(v = c(t, -t),  col = "yellow")
# confidence interval shading
# plot(tempX, tempY, type = 'l')
# polygon(c(-1.96, tempX, 1.96), c(0, tempY, 0), col="red")
# T-distribution vs Normal distribution
plot(tempX, tempY, type = 'l')
lines(tempX, dnorm(tempX), col = "red")
sampleData
```

```
sampleData
```

```
##     Ethnicity Gender AgeGroup Employment ModeOfPayment BoardingDuration
## 24          W      M        A          E             C             9.54
## 18          W      M        Y          E             C             9.99
## 33          W      F        A          U             C             6.02
## 5          HL      M        S          U             C             2.62
## 19          A      F        S          E             C            10.00
## 39         HL      M        M          U             C             8.54
## 31         AA      M        S          E             C             8.51
## 9           W      F        M          E             C             3.40
## 3           W      F        M          E             C             3.66
## 21          A      M        A          E             C             3.19
## 11         AA      M        M          E             C             4.86
## 20         AA      M        S          E             C             5.38
## 41          W      M        A          U             C             8.70
## 40         HL      M        A          E             C             3.50
## 26          W      F        A          E             C             9.34
## 35          A      M        M          E             C             6.34
## 28         AA      F        M          E             C             5.09
## 30         AA      M        A          E             C             9.57
## 34          W      M        S          U             C             6.06
## 15          W      F        M          E             C             5.80
## 6          AA      M        M          E             C             9.96
## 25          W      M        M          U             C             8.37
## 1           W      M        S          E             C             7.36
## 37          W      M        M          U             C             6.89
## 17         AA      F        M          E             C             8.08
## 27          W      M        M          E             C             6.32
## 36          W      F        A          E             C             6.12
## 10         HL      F        M          E             C             4.97
## 14          W      M        A          E             C             3.90
## 38         HL      F        M          E             C             8.87
## 12          W      M        S          E             C             4.75
## 13          W      F        M          E             C             6.88
## 23         HL      M        M          E             C             5.85
## 32          W      M        M          E             C             5.13
## 4           W      F        M          E             C             4.22
## 7          HL      M        A          U             C             6.23
## 22          W      F        A          E             C             9.75
## 8          HL      F        A          E             C             9.02
## 42         HL      M        A          E             C             9.44
## 16          A      M        M          U             C             8.27
## 89          W      M        A          E             P             1.98
## 121        HL      M        M          E             P             7.97
## 78         HL      M        A          E             P             2.63
## 44          W      M        A          U             P             2.07
## 92         AA      F        S          E             P             7.77
## 51          W      M        M          E             P             2.36
## 52         AA      F        M          E             P             3.75
## 71          W      F        A          E             P             5.66
## 101         W      M        A          U             P             9.99
## 301         W      M        M          E             P             8.08
## 43         HL      M        A          E             P             8.11
## 74          W      F        Y          E             P             1.97
```

```
## 96          W      M      S          U          P          5.42
## 58         HL      F      S          E          P          7.54
## 50          A      M      A          E          P          3.27
## 95          W      M      S          E          P          2.22
## 381        HL      M      S          U          P         10.01
## 311         W      M      A          E          P          5.50
## 391        HL      M      A          E          P          7.33
## 45          W      M      A          E          P          9.30
## 91          W      F      M          E          P          5.54
## 46          W      M      A          E          P          4.75
## 191         W      M      M          E          P          2.46
## 81          W      M      M          E          P          2.21
## 99         HL      F      A          E          P          9.08
## 105         W      M      A          E          P          7.38
## 241        HL      M      M          E          P          2.21
## 86          W      M      A          E          P          2.16
## 57         AA      F      M          E          P          3.30
## 108        AA      M      M          E          P          8.52
## 131        AA      M      A          E          P          9.12
## 48          W      F      A          E          P          4.05
## 87          W      M      M          E          P          5.67
## 75         HL      M      Y          E          P          5.73
## 111         W      M      A          E          P          2.08
## 98          W      M      M          E          P          3.55
## 106         W      M      S          E          P          4.26
## 90         HL      F      S          E          P          8.43
## 80          W      F      A          E          P          4.71
## 331         W      M      A          U          P          8.29
```