Data Analysis Report

Fall' 2019

Date: 11/25/2019

Phani Sai Kamal Lingam
Soumyajeet Patra

Effect of Different Fare Collection System on Boarding Times

How the payment preferences for Tickets in Buses impacts the boarding time of the passengers?

ABSTRACT

Understanding the Transit Boarding time has potential benefits for both users and operators. In this article, multiple statistical methods are used to analyze the influence of different payment methods and confounding variables like gender ethnicity, Age Group, Employment on explaining observed variation in boarding times. Using the real time boarding average times, performance comparisons are simulated assuming a service with payment inside buses, a prepaid pass validated inside buses and cash transactions. Results show the specific effect of all the variables involved in determining the bus boarding times in particular, that substantial time savings are accruable if payment methods are upgraded from slow techniques, such as cash transactions to the fastest one, bus pass validated inside buses (with or without contact) or left as is.

INTRODUCTION

An important part of the total running time for buses is spent at stops and stations in the process of boarding of passengers. Understanding the nature of this process has potential benefits for both users and operators, if after a detailed characterization of the time that a bus is stopped for boarding passengers, some recommendations can be made to reduce it. A possible reduction in this time can be translated into cost savings for the operator, if the total running time is reduced by a noticeable margin, and benefits for users as well, perceived as a reduction in their overall travel time, a benefit that can be monetized using the users' value of travel time savings. In this project we estimate the marginal contribution of all factors involved in the process of serving passengers at stops, that is to say, the number of passengers boarding , the payment method and any relevant dead time that takes place in between this time (time in which a bus is stopped not serving passengers). Multiple statistical methods are estimated to this end, ranging from the most basic model in which only the total number of passengers boarding is considered, to more detailed models where passengers are disaggregated in categories that are observed to have different boarding times: age groups, ethnicity, gender, employment, using data collected from dwell time surveys in Seattle, King

County Metro with different fare collection systems, such as prepaid card, cash transactions inside buses . Once the nature of the boarding time has been captured using statistical models, we then compare the total boarding time with several fare collection methods for different demand levels. The results of this study can be used to evaluate the costs and benefits of alternative fare payment technologies.

DATA COLLECTION STRATEGY

Boarding time surveys were conducted 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.

These are bus services in a high population area in central Seattle run by a public operator (King County Metro).

In these services, passengers either pay directly with cash to a machine which vends out ticket or Use the Monthly Bus Pass (Also Known as ORCA Card). In every bus, two devices are set next to the front door (one at the right and one at the left, close to the driver) to pay by cash to the ticket vending machine and for passengers to validate them card, which has to be introduced near the machine in order to be read.

Surveys were conducted on Monday on the 18th November 2019 by two observers on board the buses, equipped with a stopwatch. For every bus stop observation, the following items are recorded:

Time in which doors are open, plus door opening and closing times. Any extra time in which the bus is stopped but no one boarded is not recorded.

Number of passengers boarding, distinguishing:

- Age Groups
  - Youth (0-17)
  - Adults (18-34)
  - Middle Age (35-54)
  - Seniors (55+)
- Separating by a payment method:
  - Cash
  - Bus Pass (ORCA)
- Ethnicity –
  - Whites,
  - African American,

    - Asian and others,
    - Hispanic/Latino
- Gender –
    - Male
    - Female
- Employed (Based on Observation)
    - Employed
    - Unemployed

Payment Methods Considered:

- Bus Pass (ORCA)
- Cash

Here Boarding Time refers to the time from when a passenger steps first step in the bus to the time, he is vended a ticket in case of cash transactions. Or in the case of Bus pass holders to the time the beep from the validation machine to let him go to his seat.

Observations with an extraordinary long Boarding time due to exceptional events have been disregarded.

Population Size		
	Population Data Points	**Sample Data Points**
Cash	42	40
Bus Pass	108	40
Total	150	80

In order to do statistical analysis, we chose 40 data points 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 its taps 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.

We chose a random seed to introduce randomization in the sample selection methodology from the population data

DATA VARIABLE DEFINITION

Variable Name	Variable /Scale	Category	Use
Payment Methods	Categorical / Nominal	• Cash • Bus Pass	This Refers to the Payment method used to pay for the ticket. It is the primary variable in our study that affects the boarding time. The payment method acts as the explanatory variable with the boarding time as the Response Variable. In King County Metro Bus Service Cash payment system, Cash is Directly paid to a Ticket Vending machine on the right of the driver with no change return policy and once cash is verified ticket is vended out from the same machine The Bus Pass method refers to the ORCA Monthly card or prepaid card which is tapped against another machine on the right of the driver which is a device with a digital display indicating the success of the tap either, pass or failure. We will ignore scenarios where ORCA Pass is rejected.

Boarding Duration	Quantitative / Continuous	Time	The Boarding time is the primary response variable in our project, We consider Boarding Time refers to the time from when a passenger steps first step in the bus to the time, he is vended a ticket in case of cash transactions. Or in the case of Bus pass holders to the time the beep from the validation machine to let him go to his seat. Boarding times were recorded in real time using the stopwatch feature of phone by two individuals in real time at random bus stops in Seattle
Gender	Categorical / Nominal	• Male • Female	Gender can be considered as a confounding variable as demonstrated in the exploratory analysis in the histogram. We considered the factor that the female population in the survey might have to take the orca card out of their clutch to validate whereas It will be quicker for men who put the bus pass in their pockets thereby reducing overall boarding times.
Age	Categorical / Ordinal	• Youth (0-17) • Adults (18-34) • Middle Age (35-54) • Seniors (55+)	Age can be considered as a confounding variable as demonstrated in the exploratory analysis below. We have considered the scenario that younger people will be faster to board the bus than seniors.
Ethnicity	Categorical /	• African American	People come from different

	Nominal	• Asian • Whites • Hispanic\Latino	backgrounds and different cultures. They may be willing to try new things but their routine activities are dependent on their culture. This also helps us identify how the diversity is affecting the scenario.
Employment	Categorical / Nominal	• Employed • Unemployed	Most corporate companies give their employees a transit pass to promote public transportation making more Employed population to use a pass rather than cash.

SHORTCOMINGS

Some nuances might have arisen while collecting data points in the scenario a crowd has boarded the bus and it was difficult getting all the data points correctly as there might have been some observation bias in such a scenario as it was difficult to keep track of the sound of bus pass tap or verify when a person has collected the ticket in case of cash transactions. There might also have been some observation bias on mis judgement of employment status and ethnicity of a person just by observation.

We are also limiting our study to Seattle city, though proper randomization was used to achieve the most unbiased data, we might have missed out on specific locations or people that might represent the population sample more effectively

EXPLORATORY ANALYSIS

In this section we compare the different effects of our confounding variables and payment method on the Boarding time of the bus, whether or not age, gender, ethnicity and employment status.

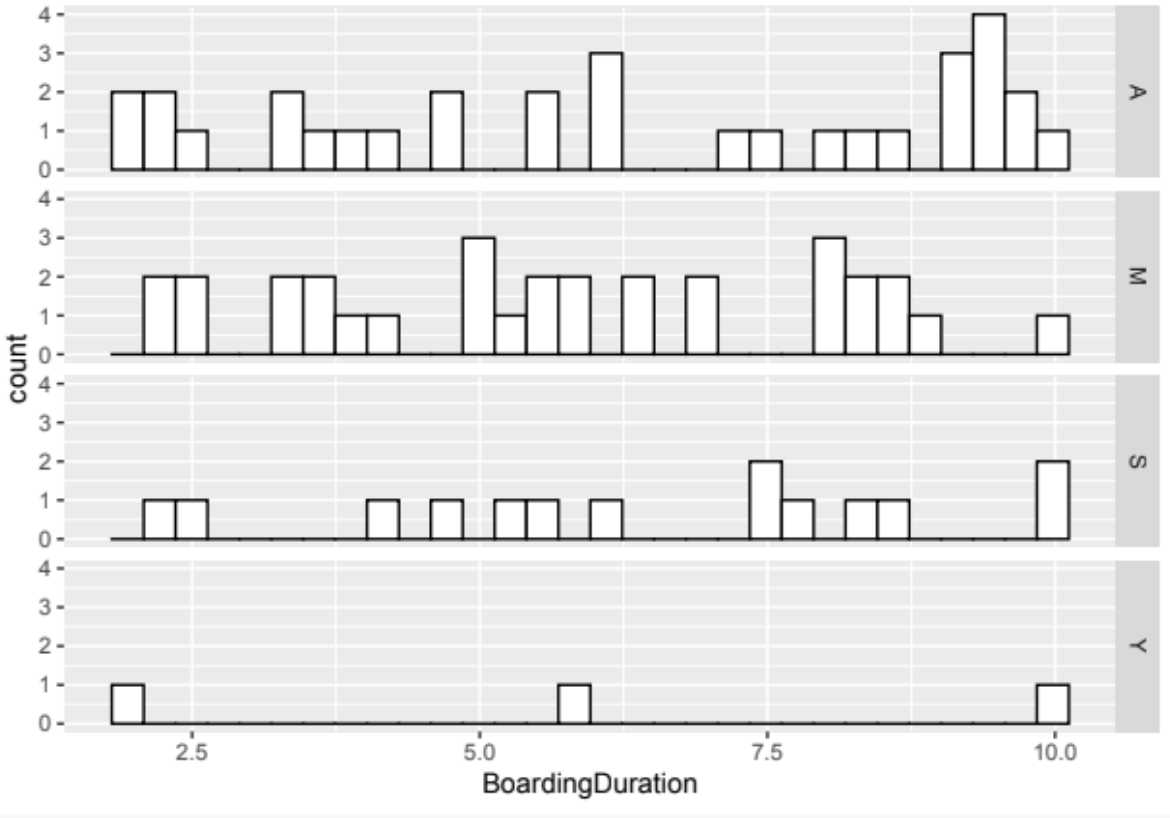


Fig 1: Histogram depicting the effect of Age Group on Boarding Duration

The above graph depicts the count and effect of Age Group on the boarding time of bus.

Here the Age groups are classified as below

- Youth (0-17) Depicted by Y
- Adults (18-34) Depicted by A
- Middle Age (35-54) Depicted by M
- Seniors (55+) Depicted by S

From the above exploratory analysis, we can infer that Age group 18-34 have the highest time of boarding with many taking above 7.5 seconds as well as has the

maximum people with boarding times in the range of 2.5 to 5 seconds.

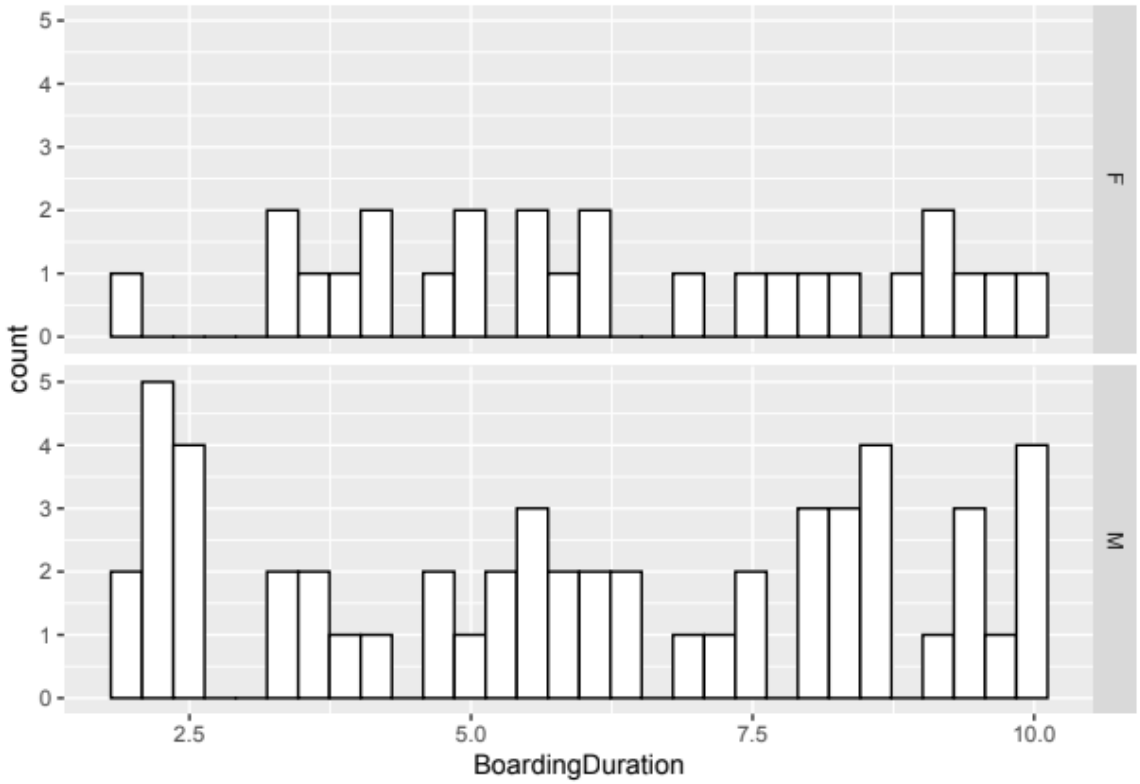


Fig 2: Histogram depicting the effect of Gender on Boarding Duration

The above graph depicts the count and effect of Gender on the Boarding Duration. Gender here is classified as below

- Male: Depicted by M
- Female: Depicted by F

The graph shows the distribution of both genders and the respective Boarding times. From the Graph we can infer that we most female individuals have a higher boarding time then male individual.

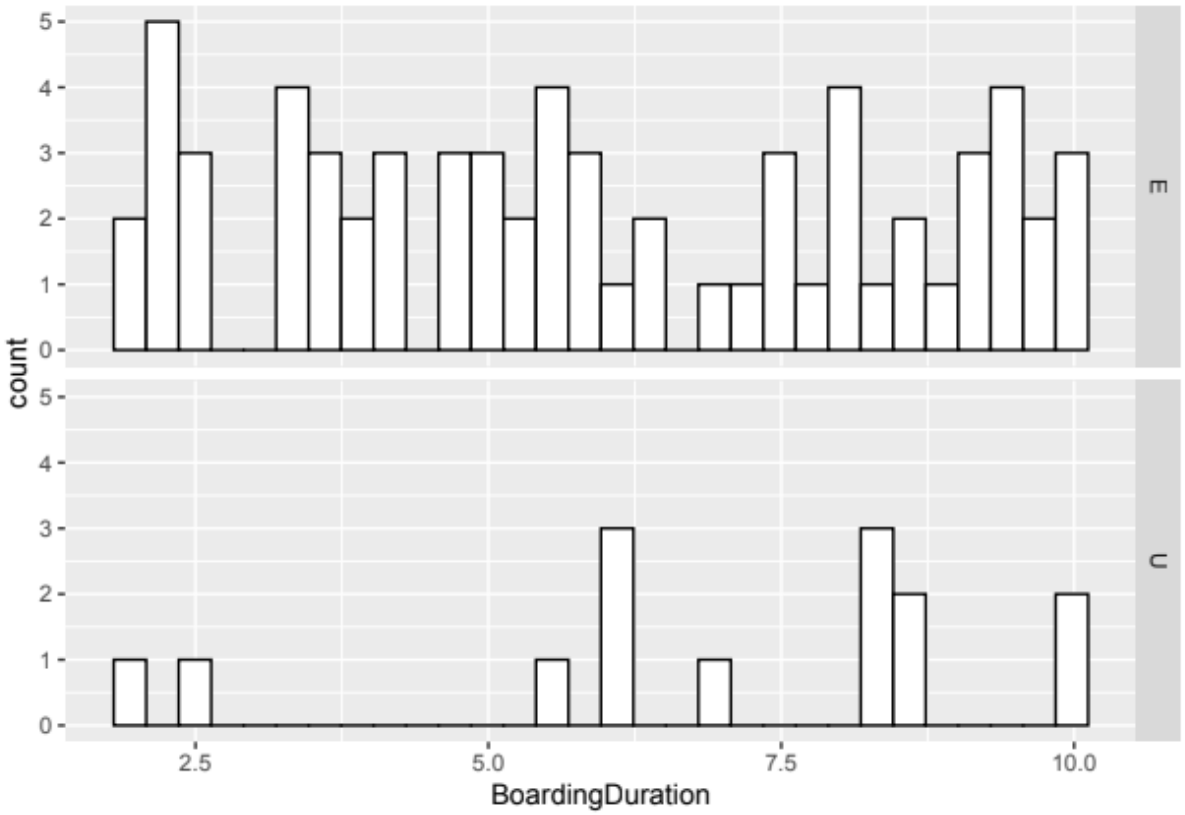


Fig 3: Histogram depicting the effect of employment status on Boarding Duration

The above graph depicts the count and effect of Employment Status on the Boarding Duration. Status here is classified as below

- Employed Depicted by E
- Unemployed depicted by U

The graph shows the distribution of both the categories and the respective Boarding times. From the Graph we can infer that we employed individuals have a lower boarding time of less than 5 seconds then unemployed individuals. This could be due to the rush of office hours and also a variety of other factors and office times.

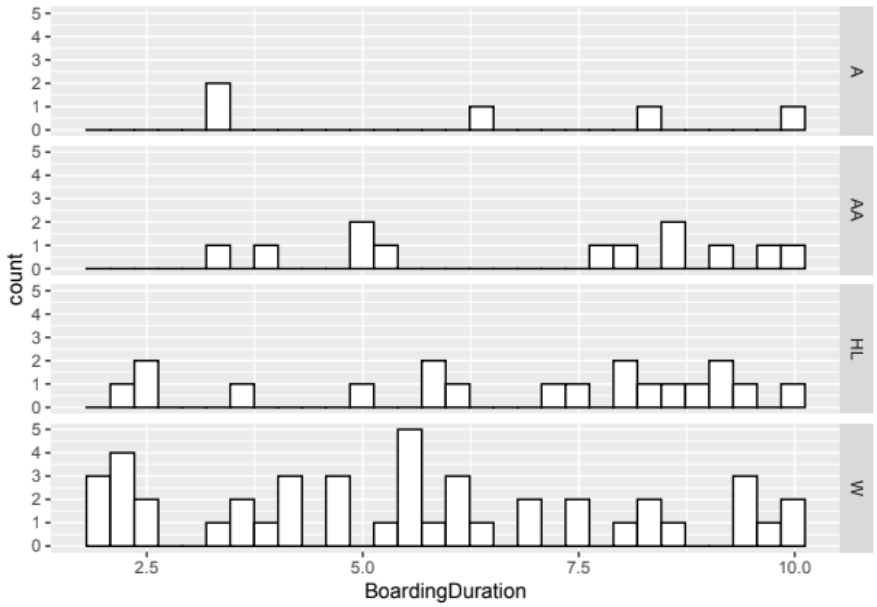


Fig 4: Histogram depicting the effect of ethnicity on Boarding Duration

The above graph depicts the count and effect of ethnicity on the boarding time of bus. Here the ethnic groups are classified as below

- Asian Depicted by A
- African American Depicted by AA
- Hispanic\ Latino Depicted by HL
- Whites Depicted by W

From the above exploratory analysis, we can infer that Whites have the greatest number of people with low boarding times below 2.5 seconds while African American and Hispanic Latino's have the highest boarding times with majority of the population above 5 seconds

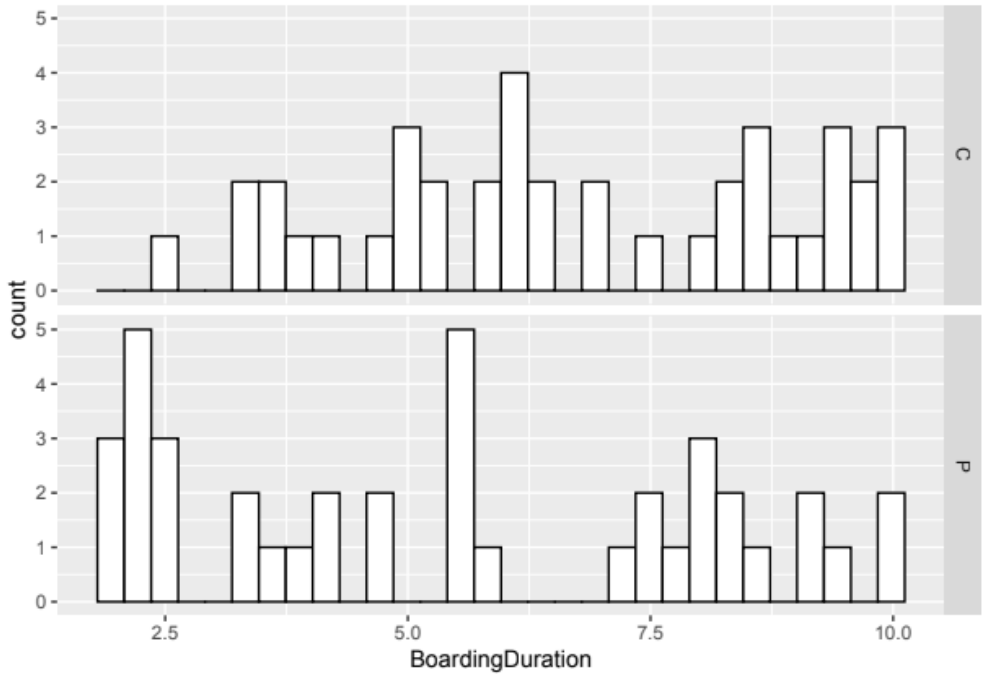


Fig 5: Histogram depicting the effect of mode of payment preference on Boarding Duration

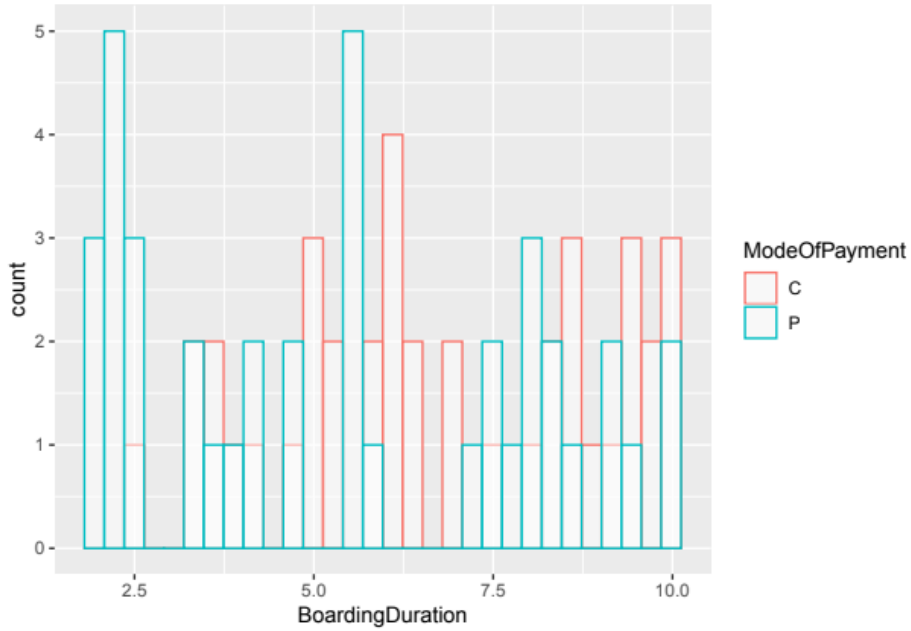


Fig 6: Histogram depicting the effect of mode of payment preference on Boarding Duration

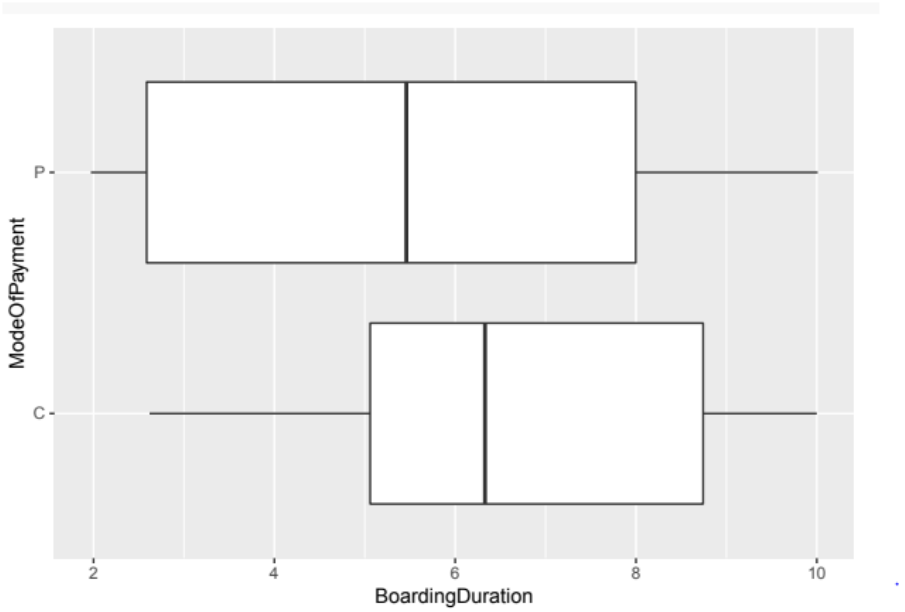


Fig 7: Boxplot depicting the effect of mode of payment preference on Boarding Duration

The above graphs and box plots depict the count and effect of mode of payment preference on the Boarding Duration. Status here is classified as below

- Cash denoted by C
- Bus Pass denoted by P

The graph shows the distribution of both the categories and the respective Boarding times. From the Graph we can infer that most people who prefer Bus pass have low boarding times as compared to people who prefer cash, this is in accordance with our topic of interest and provide us an exploratory conclusion on how Payment method mode affects Boarding Duration.

When we plot the Boxplot, we can conclude that the mean of the boarding duration for Bus Pass is less than the mean of the boarding duration for individuals who prefer Cash payments. This result is in accordance with our earlier graphs and thereby provides exploratory results on our question of interest

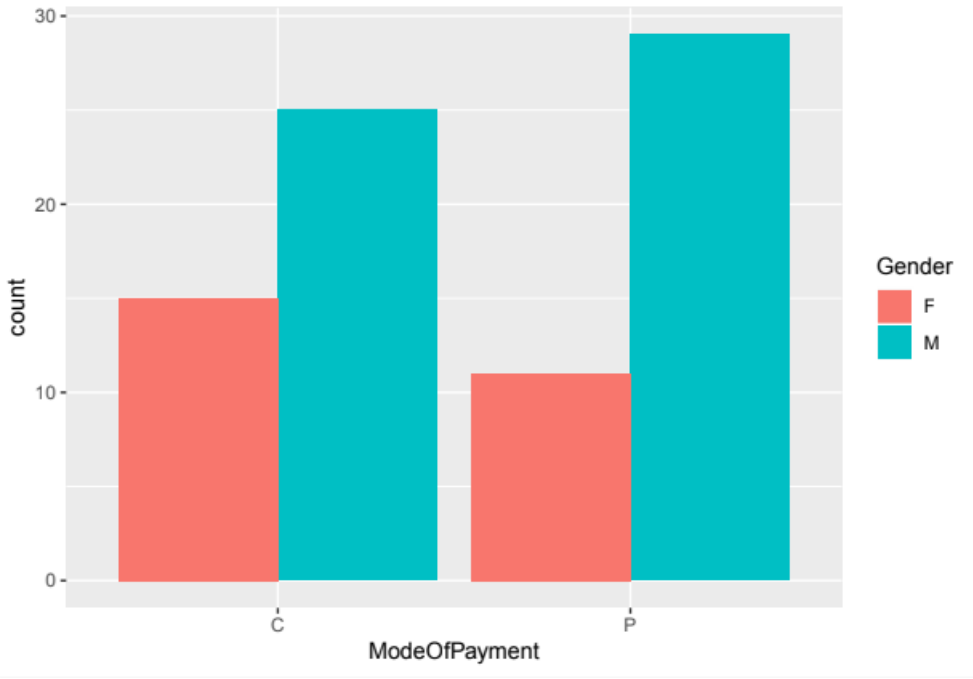


Fig 8: Bar Graph depicting the preference of Both the gender for mode of payment

From the above bar graph we can conclude that a greater number of male prefer Bus Pass while a greater number of females prefer Cash Transactions, this is probably in accordance with our previous exploratory analysis in Fig 2 of boarding times on the basis of gender where we saw that female individuals had a greater boarding time than male individuals which is in accordance with Fig 5 and 6 that cash payment transactions have a greater boarding time than transactions made with Bus pass.

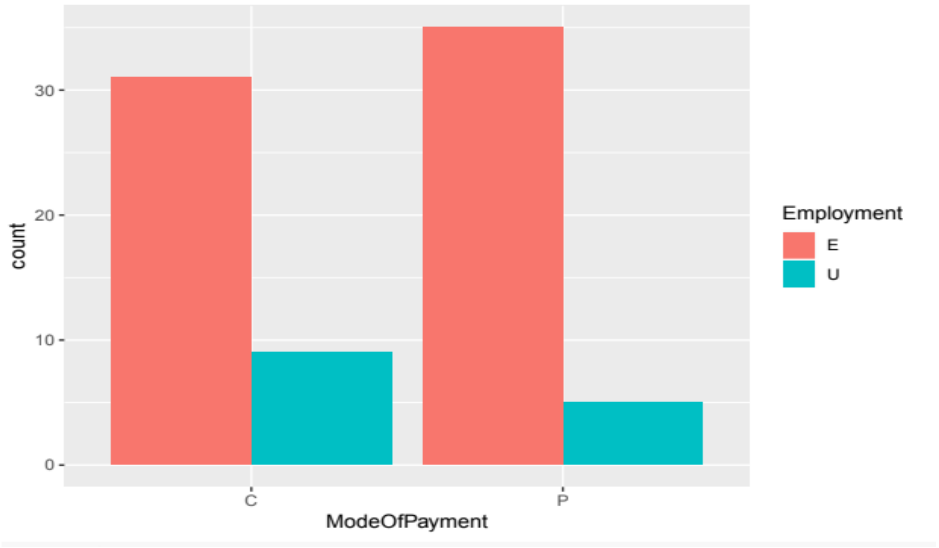


Fig 9: Bar Graph depicting the preference of sectors for mode of payment

From the above bar graph we can conclude that a greater number of employed individuals prefer Bus Pass while a greater number of Unemployed prefer Cash Transactions, this is probably in accordance with our previous exploratory analysis in Fig 3 of boarding times on the basis of employment status where we saw that unemployed individuals had a greater boarding time than employed individuals which is in accordance with Fig 5 and 6 that cash payment transactions have a greater boarding time than transactions made with Bus pass.

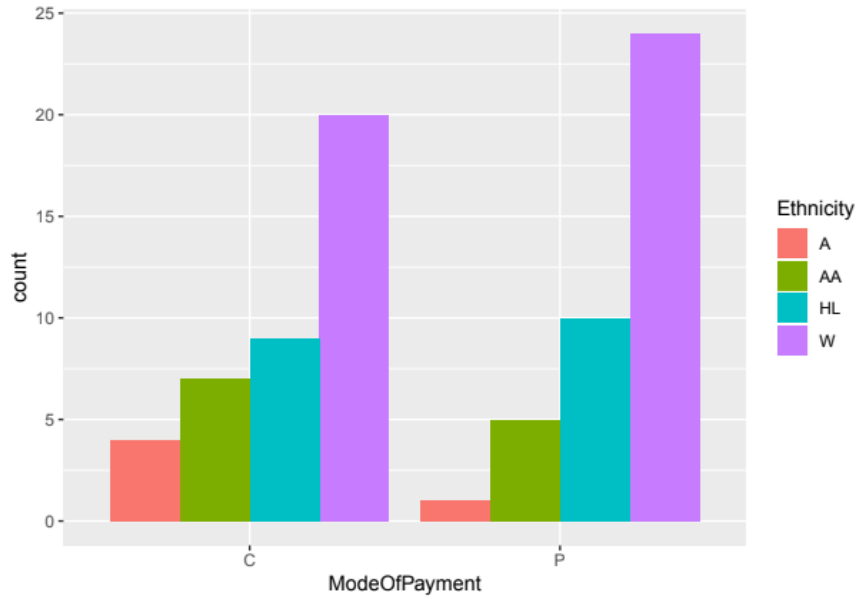


Fig 10: Bar Graph depicting the preference of individuals from different ethnical background for mode of payment

From the above bar graph we can conclude that a greater number of Whites prefer Bus Pass while a greater number of African American prefer Cash Transactions, this is probably in accordance with our previous exploratory analysis in Fig 4 of boarding times on the basis of ethnicity where we saw that White individuals had a lesser boarding time than African Americans which is in accordance with Fig 5 and 6 that cash payment transactions have a greater boarding time than transactions made with Bus pass.

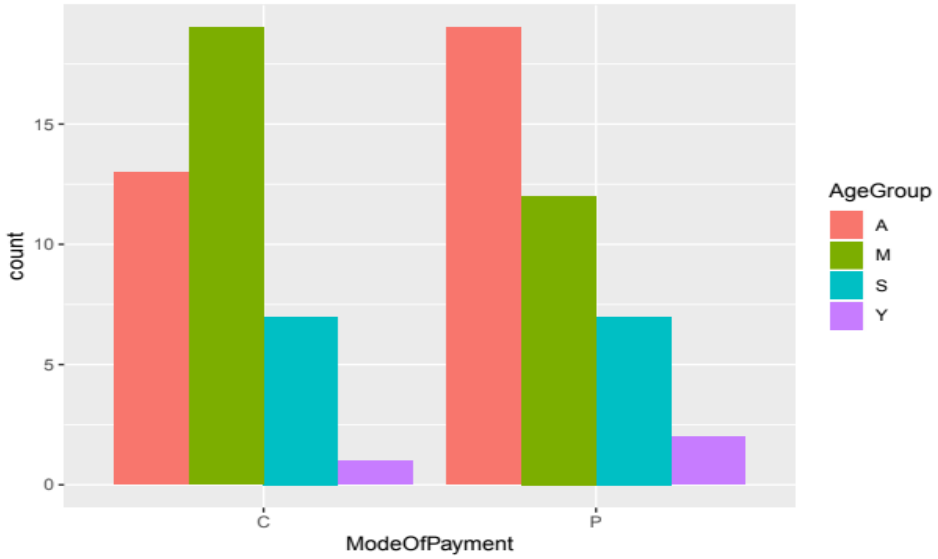


Fig 11: Bar Graph depicting the preference of age groups for mode of payment

From the above bar graph we can conclude that a greater number of Adults prefer Bus Pass, this is probably in accordance with our previous exploratory analysis in Fig 1 of boarding times on the basis of Age where we saw that Adult individuals had a lesser boarding time which is in accordance with Fig 5 and 6 that cash payment transactions have a greater boarding time than transactions made with Bus pass.

From our exploratory analysis we can see that Individuals using Bus Pass have a shorter boarding time than individuals relying on cash.

STATISTICAL ANALYSIS

Population: Our population consisted of random travelers using King County Metro Bus for a daily commute. These travelers provided us our data points on their boarding time.

Sample: In order to do statistical analysis, we chose 40 data points from each category of mode of payment (Cash and Bus Pass) to come up with a reliable conclusion. This will prevent the dominance of one category in the survey and thereby providing us with an equal weightage of boarding times in the study from both the categories

Population parameter: The mean boarding time from both the categories of payment is our population parameter.

In our case the population parameter is the mean of the sample using Cash payment for transaction and the mean of the sample using Bus Pass for transaction

u_c = mean of the sample of population using cash payments for transactions

u_p = mean of the sample of population using pass as a mode of transaction

Choice of Test:

A two sample T-test is used because data was categorized into two different population (payment method) and this will aide in calculating the difference in mean for the two populations.

It will help us in measuring the difference of means of two population samples (Cash or Bus pass) and thereby determine if there is any significance difference in the mean of the boarding times or the difference is just by chance, thereby allowing us to draw inference from the same..

For a two sample T-test our population data satisfies all the conditions:

The data are continuous (not discrete).(Boarding time is quantitative, Continuous)

The data follow the normal probability distribution

(N>30) in our case for each population

The two samples are independent. There is no relationship between the individuals in one sample cash transaction as compared to those preferring Bus pass other (as there is in the paired t-test).

Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample.

Normal Q-Q Plot

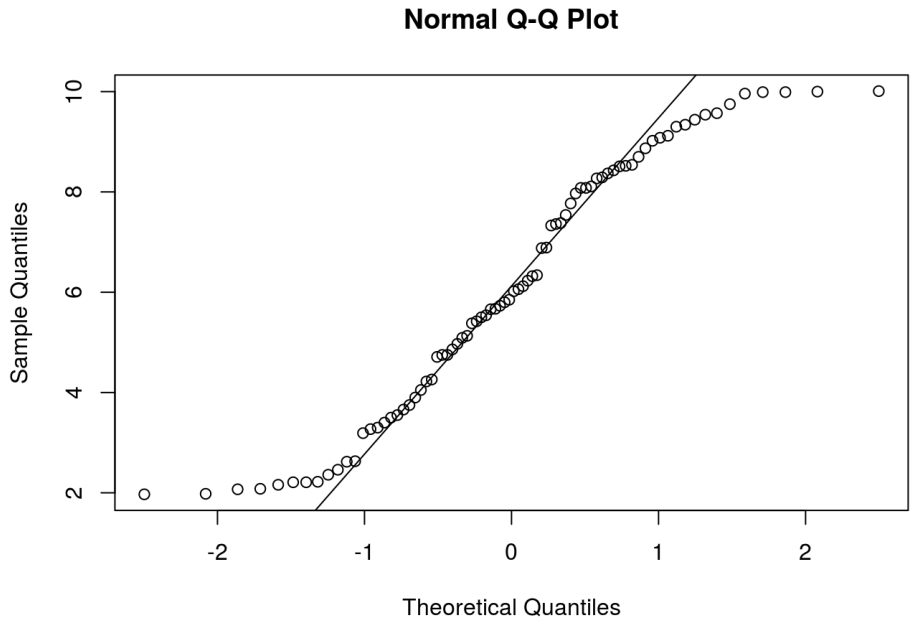


Fig 12. Q-Q Plot of Entire Sample Data

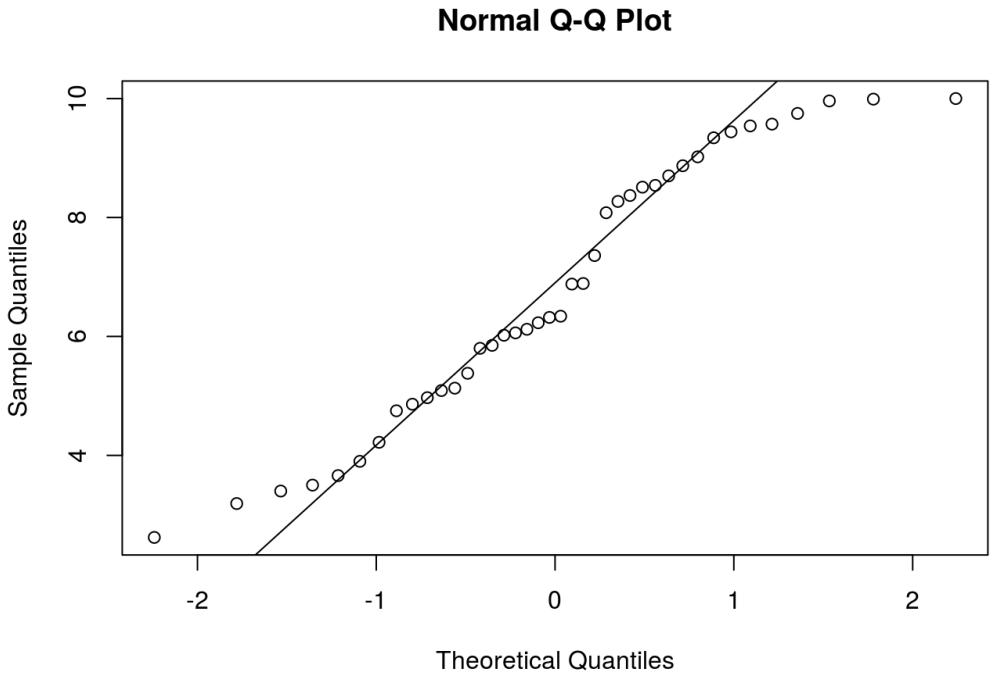


Fig 13. Q-Q Plot of Population using Cash

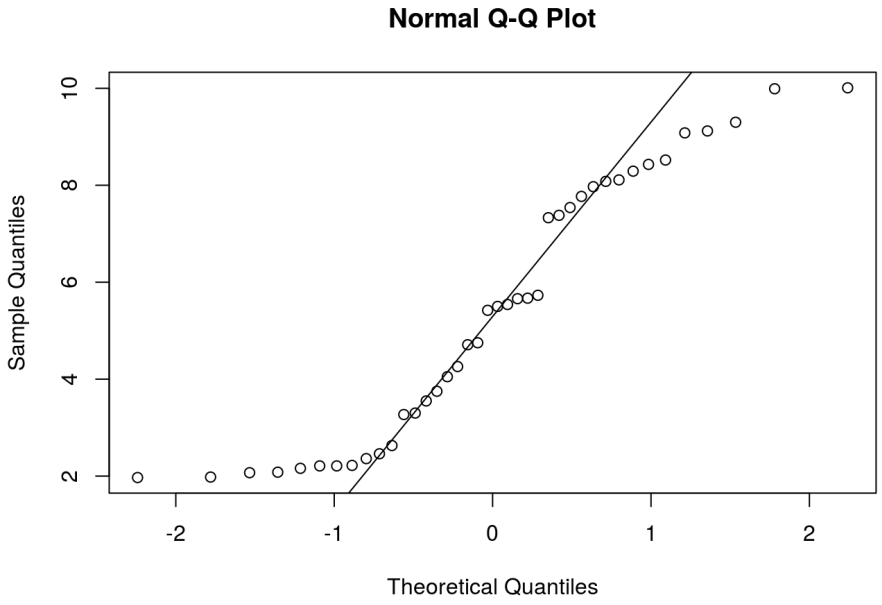
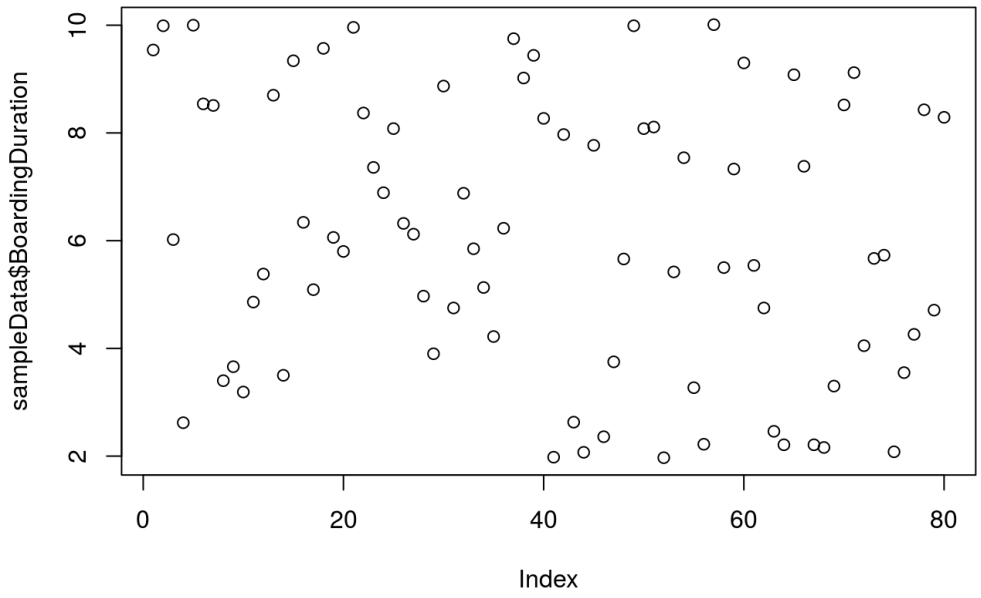


Fig 14. Q-Q Plot of Population using Pass

Fig 15. Scatterplot depicting the randomness in data sampling



Hypothesis

Null Hypothesis: The true population mean of Boarding Duration of Commuters using Cash is equal to

the true population mean of Boarding Duration 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 Duration 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 Duration 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 :

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

p-value for null hypothesis

The computer computed p value is as follows:

```
##
##  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
```

The Manually calculate p value is **0.019129754.**

P-value for Alternate Hypothesis 2

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

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

```
##
##  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
```

Confidence Interval

Using the below formula to calculate the confidence Interval in for our two-sample t-distribution i

$$P(\bar{x} - t_{(\frac{\alpha}{2},n-1)}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{(\frac{\alpha}{2},n-1)}\frac{s}{\sqrt{n}}) = (1-\alpha)$$

The lower bound of our confidence interval is 0.4199606

The upper bound of our confidence interval is 2.283039

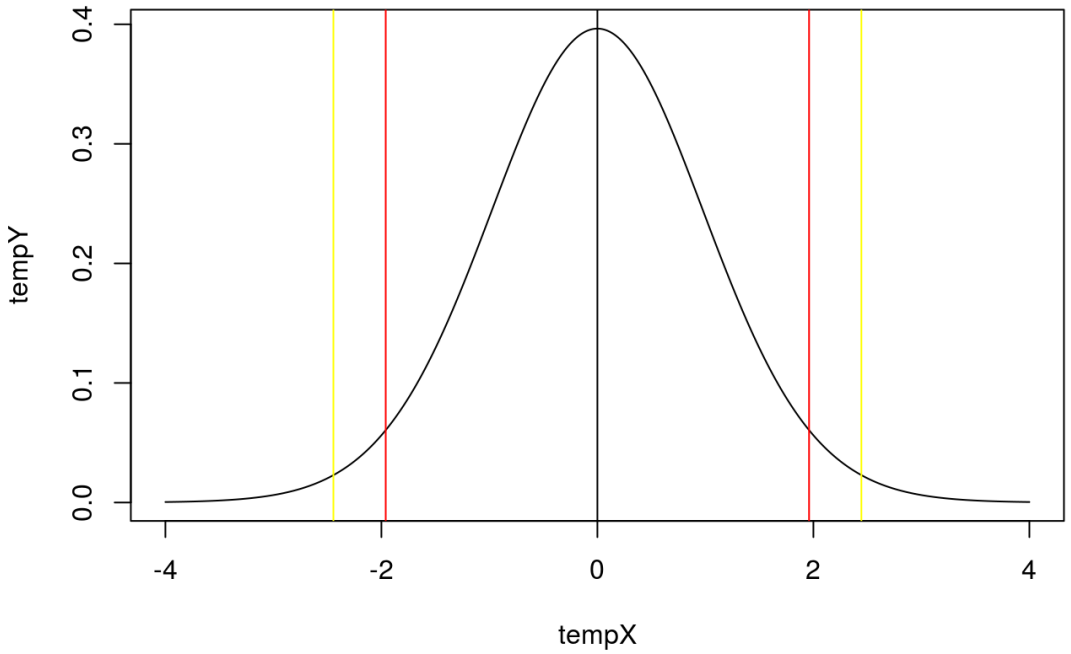


Fig 16. Test Statistic and Confidence Interval Line on Distribution graph

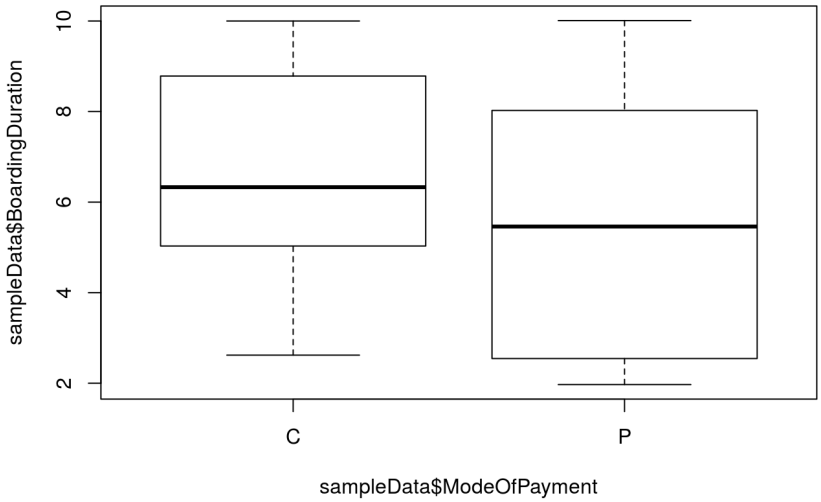


Fig 17. Box plot depicting the difference in means.

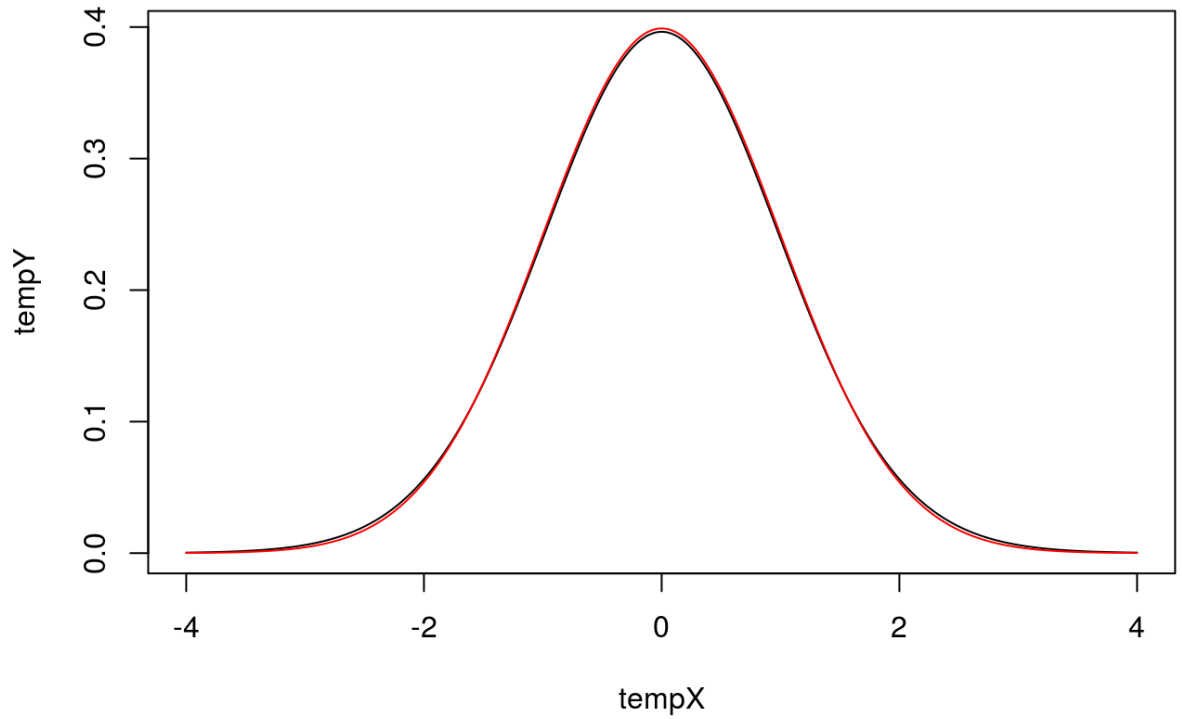


Fig. 18 Deviation of T-Distribution from Normal Distribution (Red)

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.

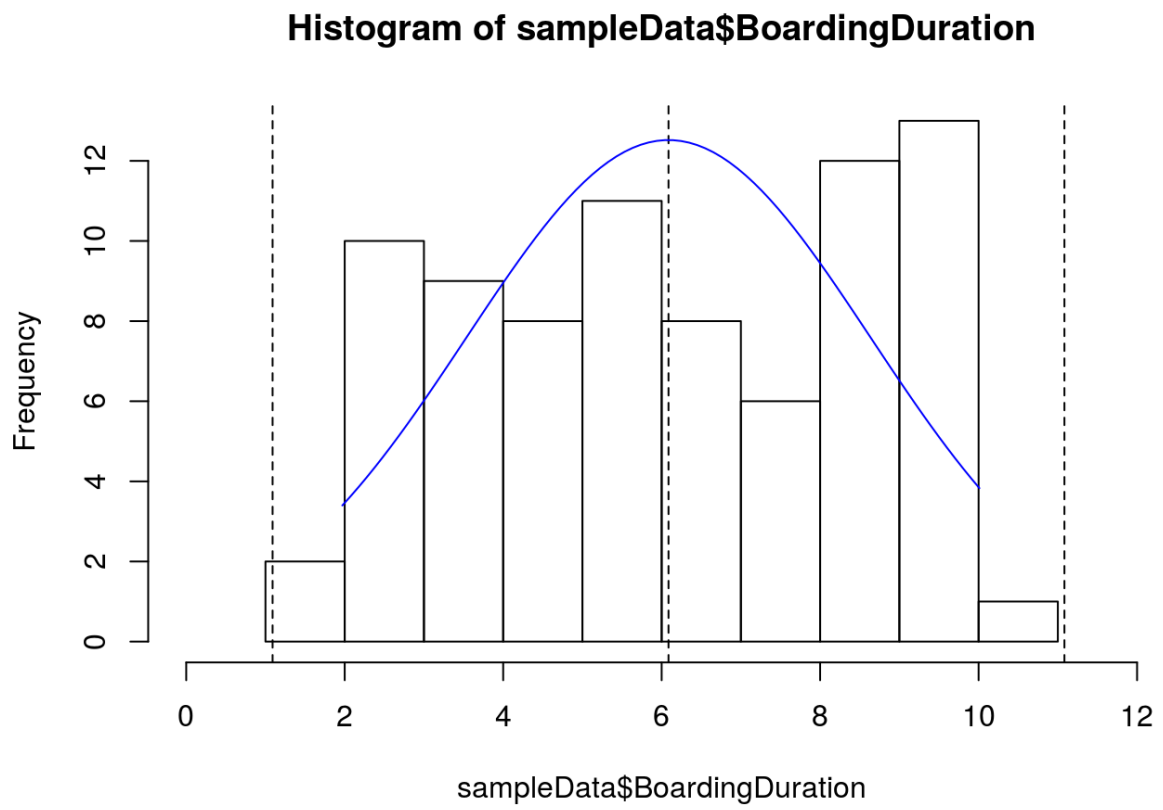


Fig 18. Histogram of Sampling Distribution

The above Histogram depicts Mean, bounds of 95% Confidence Interval of Mean of our sampling distribution.

DISCUSSION

Based on the findings and conclusions in this study, the following recommendations are made:

1. Instead of having two modes of payment for ticketing system King County Metro System should implement a single mode so that it will save them the cost of infrastructure and bus travel time.
2. There is a significant difference of 1.3 seconds between those who use cash and those use pass for each passenger hopping on the bus, which can be mitigated to make the system more time efficient.

IMPLICATIONS

As discussed earlier our understanding the nature of this process can provide potential benefits for both users and operators, after a detailed analysis and inference of how boarding time can be reduced , It can have a significant impact on the time that a bus is stopped for boarding passengers,

New form of payment method even faster than bus pass can be introduced and thus a possible reduction in this time can be translated into cost savings for the operator, if the total running time is reduced by a noticeable margin, and benefits for users as well, perceived as a reduction in their overall travel time, a benefit that can be monetized using the users' value of travel time savings.

LIMITATIONS

While the outcome of this research has lead to favourable results in favour of bus pass. The direct implication to unify the mode of payment into a single method might result in undesirable circumstances for people who might have lost their bus pass furthermore while we have had the chance to observe gather data for individuals whose bus pass tapping has been successful in one go itself, there might be scenarios where there might be device failure scenarios which might provide a lot of outliers to our boarding time and thereby affecting the entire inference itself and show cash transactions as a better form of payment than Bus pass

NEXT STEP

While our research provided decisive conclusion, we can further our research to consider a larger dataset.

We can also consider other forms of payment which may be faster than the existing prevalent forms of payment use the data to apply multiple regression models that can estimate the influence of other factors like, the existence of steps at doors, the age of passengers and the possible friction between users boarding, alighting and standing, on explaining observed variation in total dwelling times in buses for each individual.

APPENDIX

R - Code:

```
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('Ge
nder')+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 = "dodge")
# 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 = "dodge")
# 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")
```

Raw Data:

```
##     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
```