INDE546_HW3

Michael Shieh

1/18/2020

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────────────────────
──────────────────────────────────────────────── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.2.1     ✓ purrr   0.3.3
## ✓ tibble  2.1.3     ✓ dplyr   0.8.3
## ✓ tidyr   1.0.0     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.4.0
```

```
## ── Conflicts ────────────────────────────────────────────────────────────
──────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

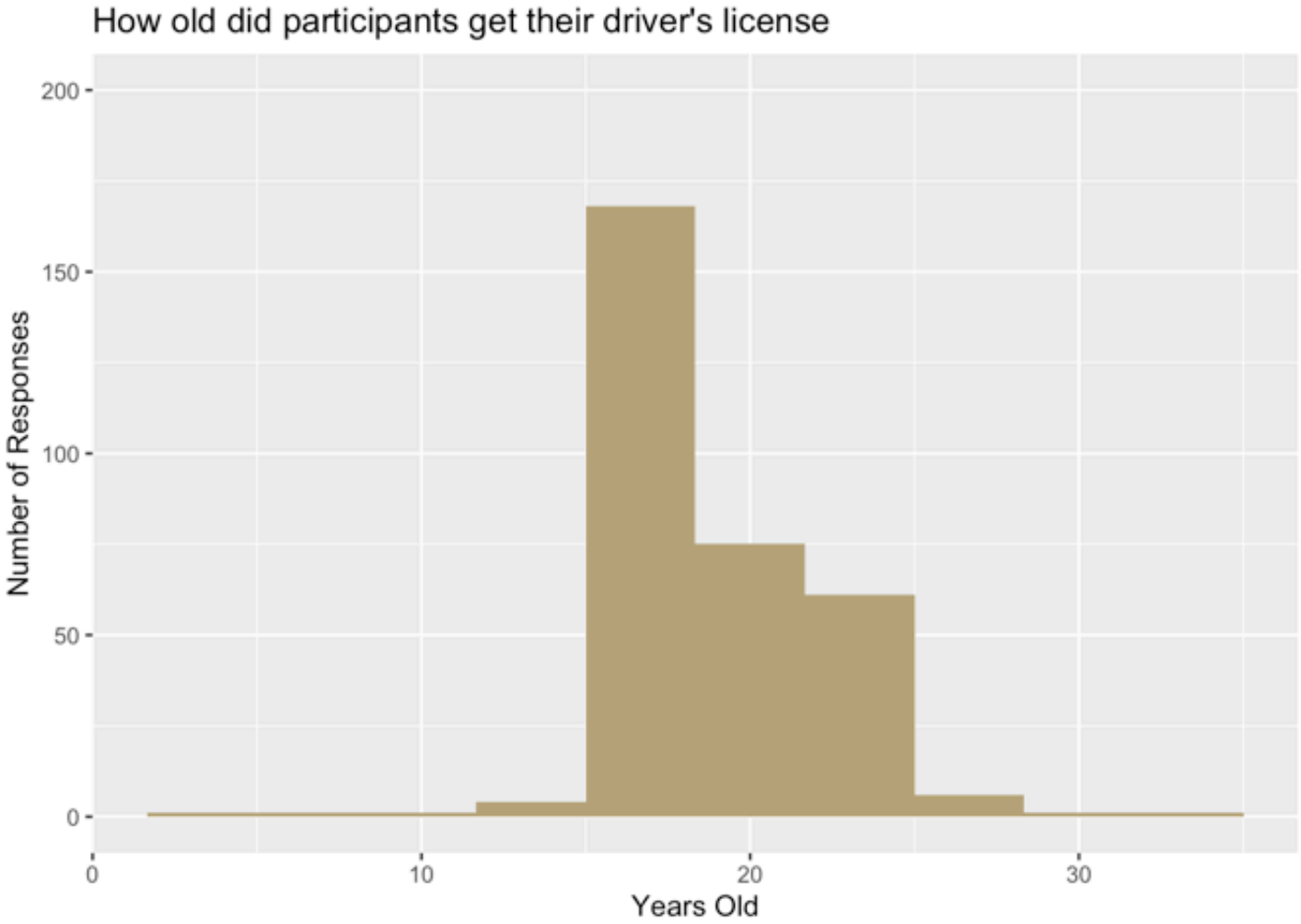
```
setwd("~/Documents/UW/2019Winter/INDE546_InferentialDataAnalysis/HW/HW3")
```

```
data <- read.csv("Class_Survey_W20.csv", header = TRUE)
data <- rename(data, DL_Age = How.old.were.you.when.you.got.your.driver.s.license., Y
earsOld = How.old.are.you., CommuteTime = On.average..how.many.minutes.does.it.take.y
ou.to.get.to.the.U..Washington.from.your.home., MeansofTranport = What.is.your.primar
y.means.of.transportation.to.and.from.the.U..Washington., Maps = What.navigation.app.
do.you.use.most.often., Who = Are.you..1, USD = Do.you.have.a.US.driver.s.license.)
```

Problem1

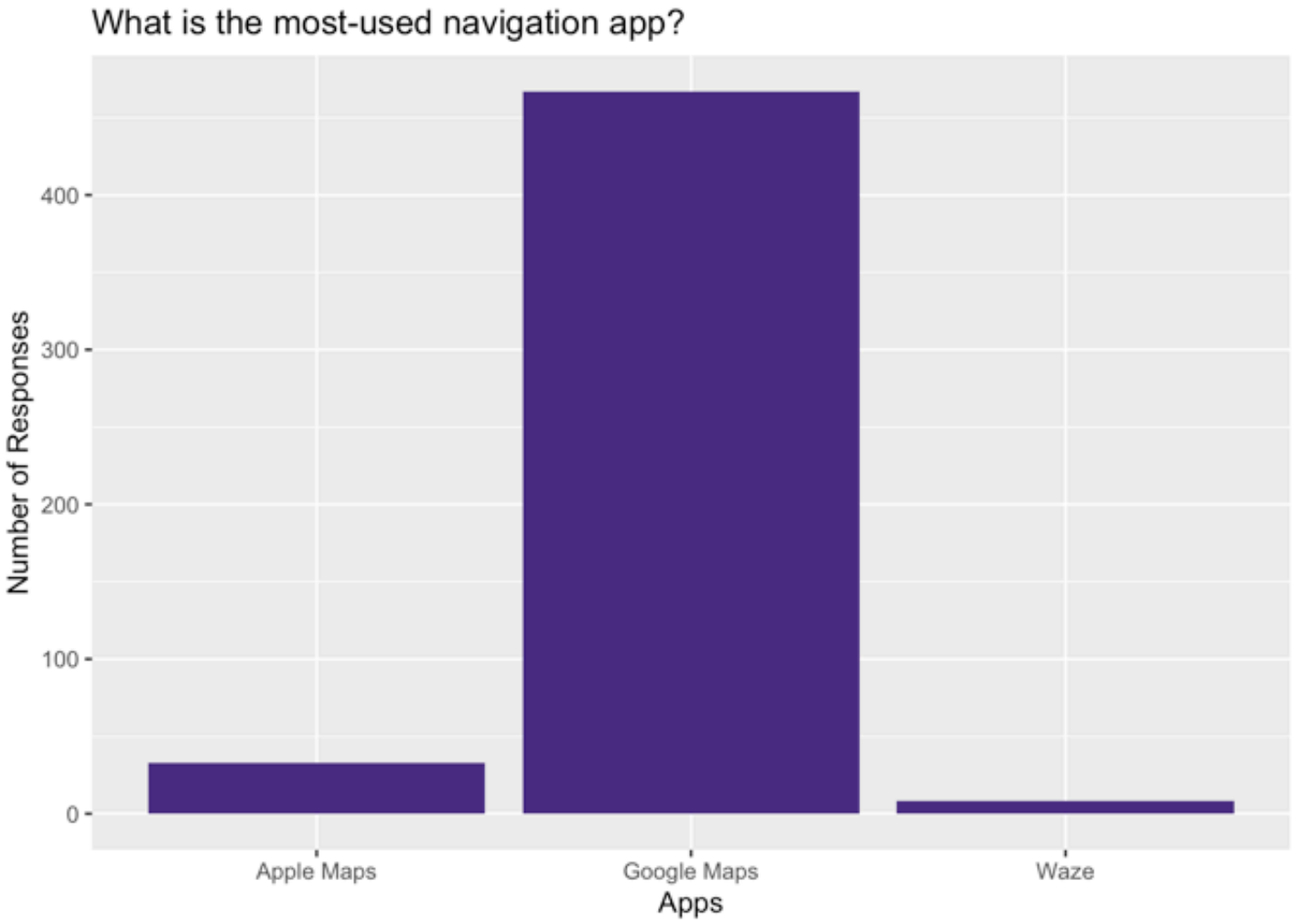
Histogram

```
data.dl_age <- data %>% filter(data$DL_Age >= 0)
ggplot(data.dl_age, aes(x = data.dl_age$DL_Age)) + geom_histogram(bins = 10, fill = "
#b7a57a") + labs(title = "How old did participants get their driver's license", x = "
Years Old", y = "Number of Responses") + ylim(c(0, 200))
```



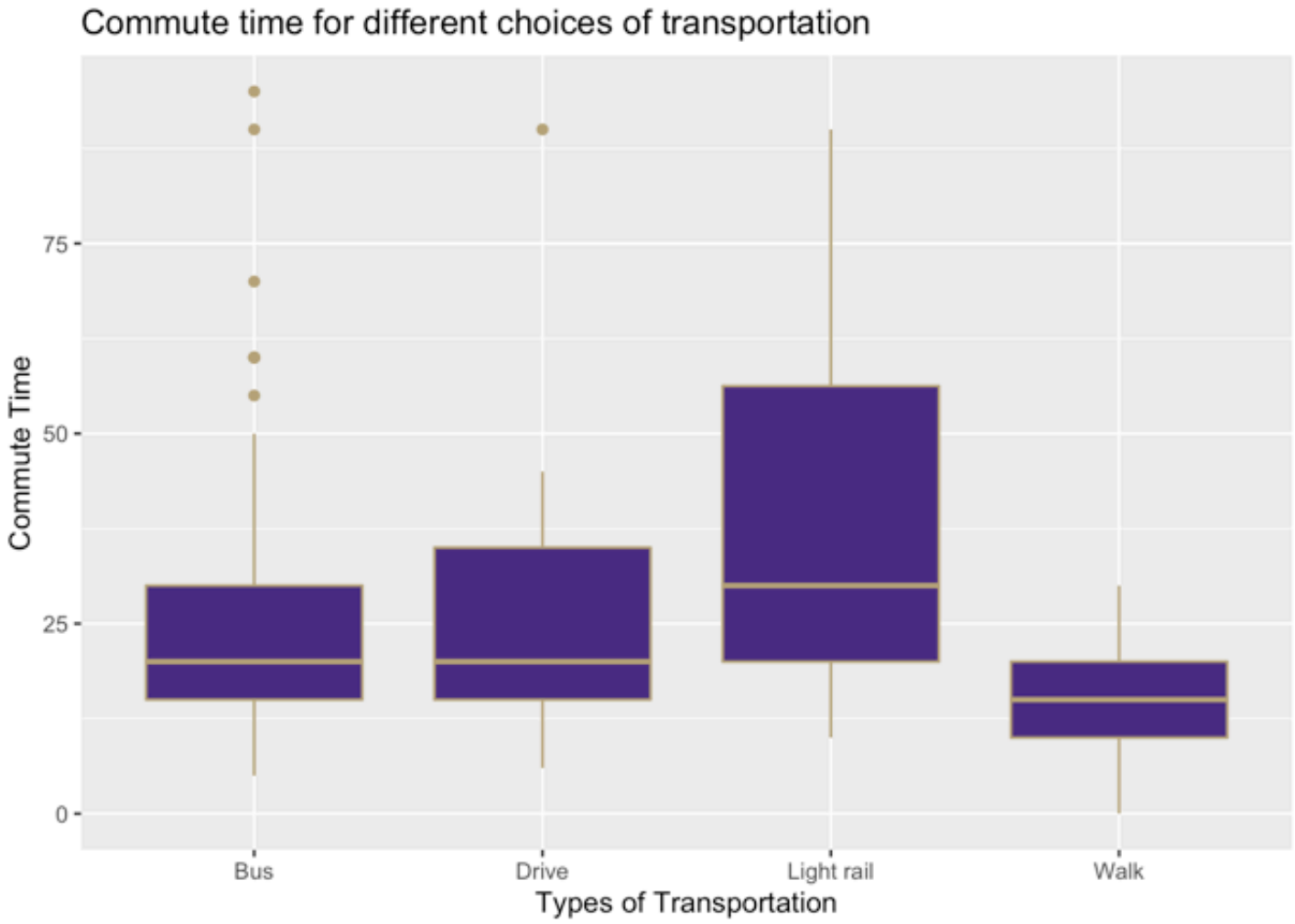
Barchart

```
data.Maps <- filter(data, Maps %in% c('Apple Maps', 'Waze', 'Google Maps'))
ggplot(data.Maps, aes(x = data.Maps$Maps)) + geom_bar(fill = "#4B2E83") + labs(title
= "What is the most-used navigation app?", x = "Apps", y = "Number of Responses")
```



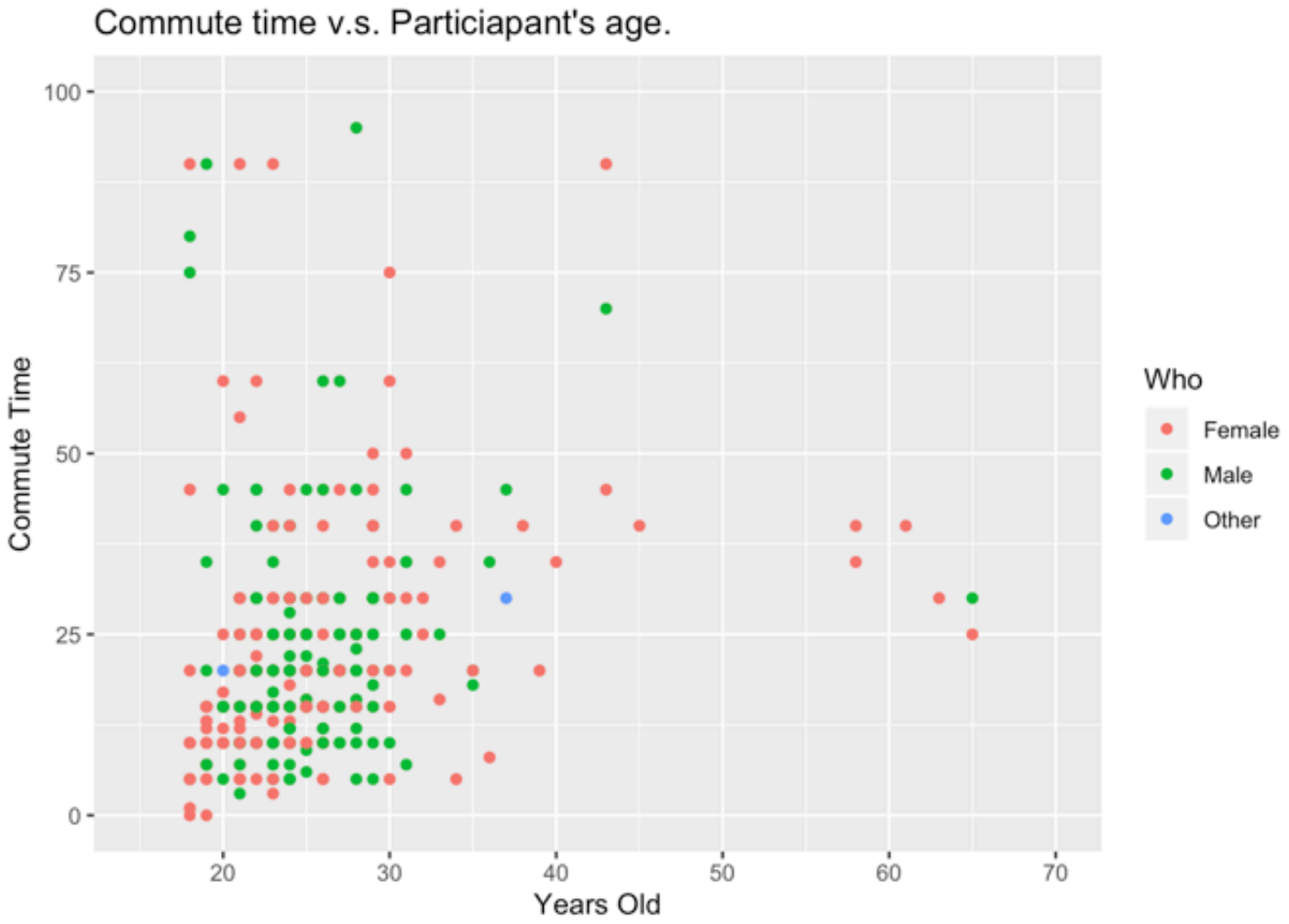
Boxplot

```
data.CT <- data %>% filter(CommuteTime >= 0, MeansofTranport %in% c('Drive', 'Bus', '
Light rail', 'Walk'))
ggplot(data.CT, aes(x = data.CT$MeansofTranport, y = data.CT$CommuteTime)) + geom_box
plot(fill = "#4B2E83", color = "#b7a57a") + labs(x = "Types of Transportation", y = "
Commute Time", title = "Commute time for different choices of transportation")
```



Scatterplot

```
data.scatter <- data %>% filter(YearsOld >= 0, CommuteTime >= 0, Who %in% c('Male', '
Female', 'Other'))
ggplot(data.scatter, aes(x = data.scatter$YearsOld, y = data.scatter$CommuteTime, col
or = Who)) + geom_point() + labs(x = "Years Old", y = "Commute Time", title = "Commut
e time v.s. Particiapant's age.") + xlim(c(15, 70)) + ylim(0, 100)
```



Problem 2

A histogram presents quantitative data and indicates the distribution of non-discrete variables. A barchart presents categorical data and indicates comparisons of discrete variables. The bars for histograms can not be reordered while touching each other, for barcharts, they can be reordered while not touching each other. A pareto chart combines bars and a line graph where the line always rises from left to right. The bars in Pareto charts are usually arranged in a descending order with the tallest bar on the very left. This often occurs when demonstrating the pareto philosophy(the 80/20 rule).

Problem 3

Perform chi-square test on whether the gender of participants are independent to whether they have a U.S. Driver licences. I choose chi-square test because this test is commonly used for testing relationshipsbetween categorical variables, which is the case for my question. The following is my hypotheses:

The Null Hypothesis(H0) is gender and having driver's license are independent.

The Alternative Hypothesis(H1) is that they are not independent.

```
data.P3 <- data %>% filter(Who %in% c('Male', 'Female'))
test <- table(data.P3$USD, data.P3$Who)
test <- test[, 2:3]
chisq <- chisq.test(test)
chisq
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  test
## X-squared = 5.7578, df = 1, p-value = 0.01642
```

As shown above, since I get the p-value less than the significant level of 0.05, the Null Hypothesis is rejected. Therefore, I conclude that the two variables are in fact dependent.