MidTerm

Midterm Exam Part 2: Hands-on Coding Assignment (12 points; 14 questions)

Instructions

Type in your SUID in place of the zeros below and run the cell (click Ctrl + Enter):

```
suid <- 326926205
```

The block of code below creates a custom data set for you to analyze. Your dataset is different from every other student's dataset. The goal of this part of the exam is to write code and comments that address the research questions described below. The quality of your comments is critical to your success on this exam! You will only be submitting this file and there are several important results that require an explanation in plain language. Pay close attention to the research questions described below when writing your code and comments.

Do not modify any of the code, just run it as is:

```
if (suid == 0) {cat("Please update your SUID (above) before running this code.")} else {cat(paste("Lyft/
```

```
## Lyft/Uber Fare Comparison Study Number: 326926205
## Sample size for this study: 117
```

Your Assignment: rYdZ Analysis

The code you just ran generates a unique dataframe called **testDF**.

You can explore it by running, e.g. head(testDF).

There is an upstart in the ride-sharing market: The new start-up firm **rYdZ** (pronounced rides) is driver-owned and operated. In addition to providing safe rides at competitive prices, the?mission of **rYdZ** is to provide a working wage to **rYdZ** drivers. But the leadership team at **rYdZ** believes there is a problem: the two giants in the industry, **Lyft** and **Uber**, are coordinating to set prices for rides that are artificially low? The team at **rYdZ** has produced a data set of more than 100 fares offered to drivers from **Lyft** and **Uber**. Your job is to analyze this data set and infer whether there is some sort of coordination between **Lyft** and **Uber** to set prices, as well as understand if either is pricing based on miles driven, or perhaps, based on geography.

Data Set Description:

Your data set contains **five variables**: The **ride number**, the **fare** (in dollars and cents) of a ride offered to a driver from Lyft, and the **distance** of that ride (in miles). There is also a **fare for a different ride** offered to a driver through Uber (and the **distance of that trip**). There are at least 100 observations (rows) in your dataset, and possibly more. Each observation was done at roughly the same time for Uber and Lyft (the data for the ride in a row was collected at roughly the same time).

Research Questions (tasks to do):

1. Output the 5th Lift fare (0.5 pts)

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

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

```
testDF %>% slice(5) %>% select(Lyft_13)
```

```
##   Lyft_13
## 1   26.04
```

2. Describe the fares provided by Lyft and Uber (separately) using descriptive statistics that you calculate in R (1 pts):

```
summary(testDF$Lyft_13)
```

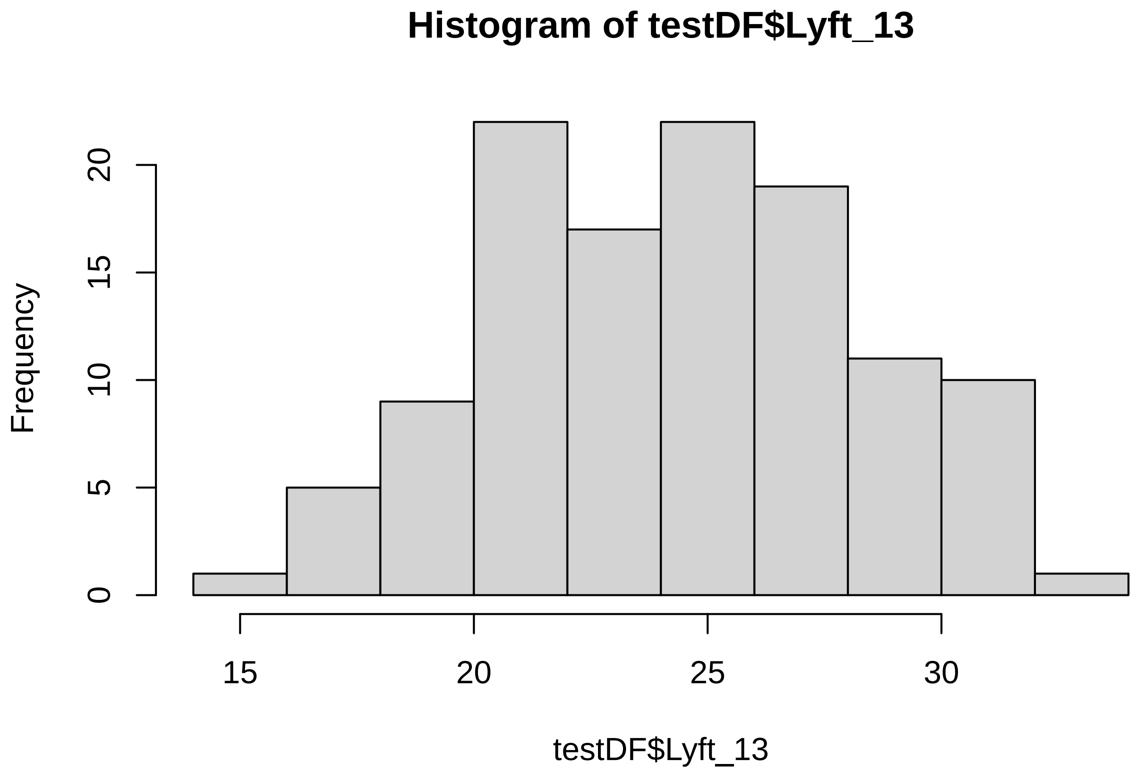
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.57   21.67   24.37   24.36   26.93   32.61
```

```
summary(testDF$Uber_13)
```

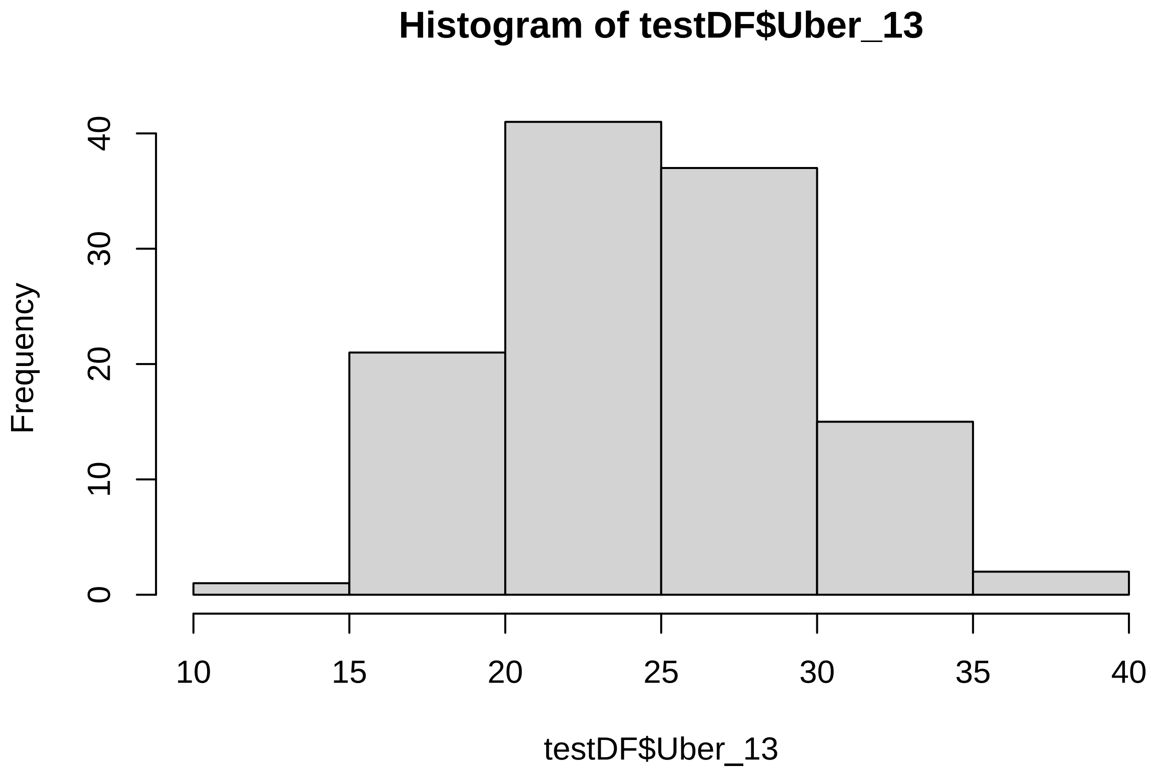
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.58   20.72   24.13   24.44   28.01   35.75
```

3. Describe the shape of the distribution for Lyft fares. Do the same for Uber fares (1 pt)

```
# fares are normally distributed. histogram plot looks close to bell shaped curve.
# it is also evident from
# the statistics above : mean and median of fares are almost equal
hist(testDF$Lyft_13)
```



```
# fares are normally distributed. histogram plot looks close to bell shaped curve.
# it is also evident from
# the statistics above : mean and median of fares are almost equal
hist(testDF$Uber_13)
```



4. Based on the fares offered by both companies, on average, which company is more expensive, Lyft or Uber? By how much? (0.5 pts)

```
testDF$Lyft_13 %>% mean()
```

```
## [1] 24.36214
```

```
testDF$Uber_13 %>% mean()
```

```
## [1] 24.44436
```

```
mean(testDF$Lyft_13) - mean(testDF$Uber_13)
```

```
## [1] -0.08222222
```

```
# Uber is little expensive by approx 0.08$
```

5. Create a new attribute, called 'diff' in testDF, that represents the difference in fares between Uber and Lyft for each observation - in other words, the difference for each row(0.5 pts):

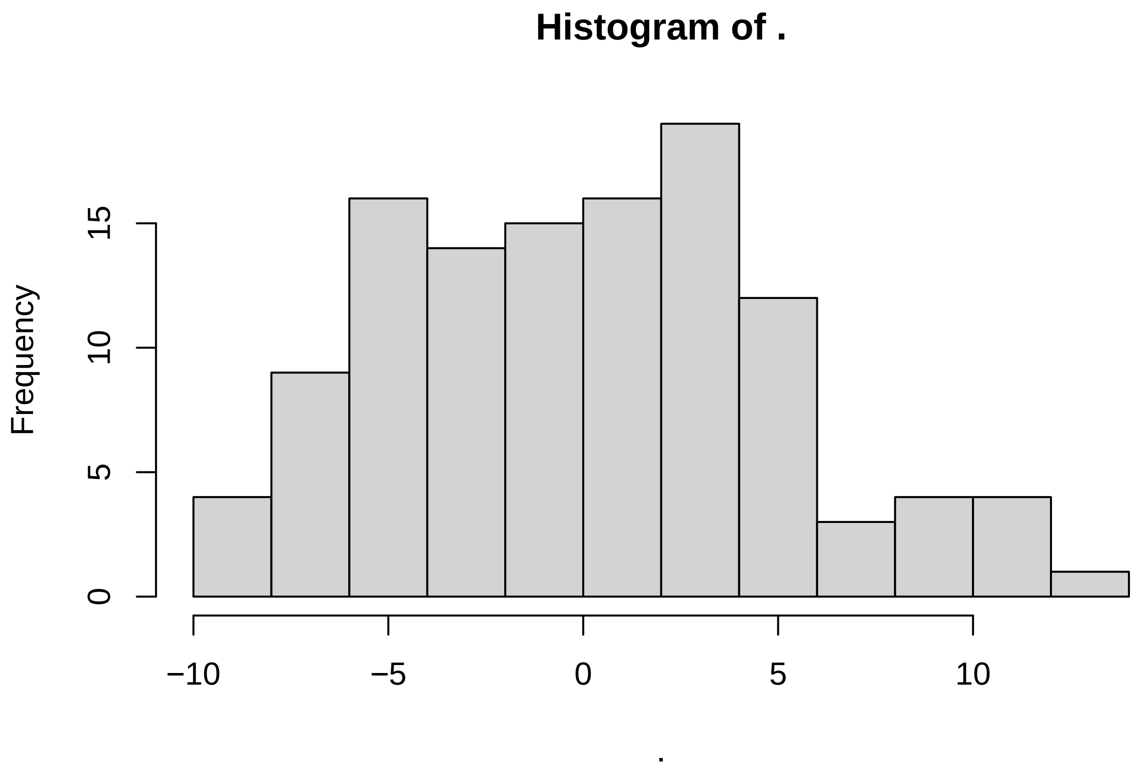
```
testDF <- testDF %>% mutate(diff=Uber_13-Lyft_13)
```

6. Describe the shape of the distribution for this new variable(0.5 pts)

```
summary(testDF$diff)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -9.97000 -3.25000  0.11000  0.08222  3.09000 12.11000
```

```
testDF$diff %>% hist()
```



```
# diff is not normally distributed. mean and median seems to be far apart
```

7. Sort testDF, based on the new attribute (*diff*), and store the sorted dataframe in 'sortedDF'. Show the first and last row in the sortedDF dataframe (1 pt)

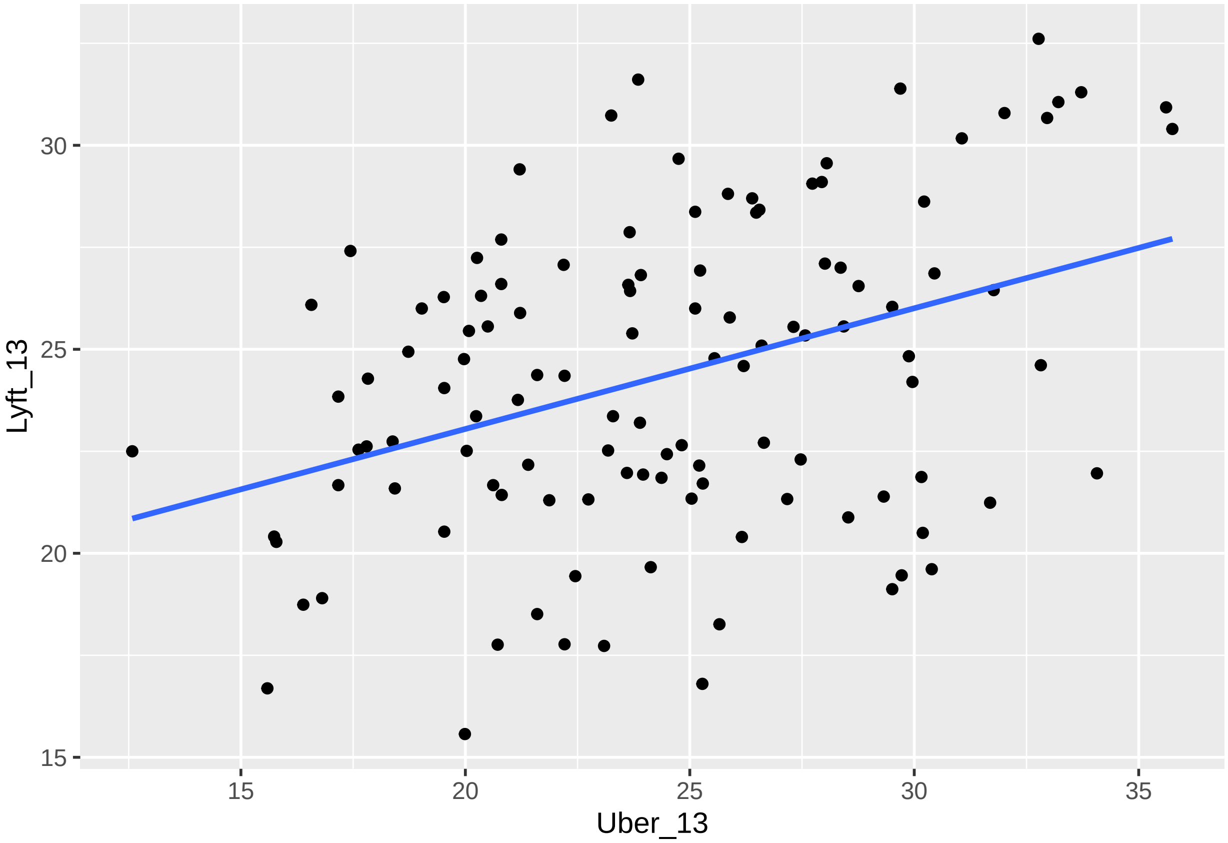
```
sortedDF <- testDF %>% arrange(diff)
sortedDF %>% slice(c(1,n()))
```

```
##   driver Lyft_13 Uber_13 Lyft_13_distance Uber_13_distance Lyft_13_state
## 1    107   27.41   17.44         16.01565         22.75163       Florida
## 2    110   21.96   34.07         62.91889         41.04287         Texas
##   Uber_13_state  diff
## 1       Florida -9.97
## 2         Texas 12.11
```

8. Create an X-Y scatterplot of the Lyft and Uber fares (make sure to provide informative labels for each axis). Does the scatterplot show an obvious pattern/relationship? (1 pt total)

```
library(ggplot2)
faresPlot <- testDF %>% ggplot() + aes(x=Uber_13, y= Lyft_13) +
  geom_point() + geom_smooth(method="lm", se=FALSE )
faresPlot
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
# from the plot, it appears to be there is a very weak positive correlation
# between uber and lyft fares
```

9. Generate a linear model trying to predict Lyft fares based on the distance of the trip. Interpret the coefficients of the statistically significant predictors in the model (1 pt).

```
lyft_reg <- lm(Lyft_13~Lyft_13_distance,data=testDF) # simple linear regression
summary(lyft_reg)
```

```
##
## Call:
## lm(formula = Lyft_13 ~ Lyft_13_distance, data = testDF)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4112 -2.8754  0.0923  2.3908  9.2128
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      22.39715    0.78593  28.498  < 2e-16 ***
## Lyft_13_distance  0.03891    0.01399   2.781  0.00633 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.725 on 115 degrees of freedom
## Multiple R-squared:  0.06303,    Adjusted R-squared:  0.05488
## F-statistic: 7.736 on 1 and 115 DF,  p-value: 0.006328
```

```
# looking at the asterisk, Lyft_13_distance is statistically significant
# co-efficient of Lyft_13_distance indicates with every additional
# increase in miles, lyft fare is predicted to
# increase by 0.03891$
# intercept is also significant: even when there is no distance covered,
# minimum fare is 22.39715$
```

10. Generate a similar model for the Uber trips. Interpret the coefficients of the statistically significant predictors in the model (0.5 pts)

```
uber_reg <- lm(Uber_13~Uber_13_distance,data=testDF) # simple linear regression
summary(uber_reg)
```

```
##
## Call:
## lm(formula = Uber_13 ~ Uber_13_distance, data = testDF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3317 -0.8426 -0.0519  0.8154  3.3701
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -4.70592    0.64131  -7.338 3.32e-11 ***
## Uber_13_distance  0.95169    0.02066  46.066  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.127 on 115 degrees of freedom
## Multiple R-squared:  0.9486, Adjusted R-squared:  0.9481
## F-statistic:  2122 on 1 and 115 DF,  p-value: < 2.2e-16
```

```
# looking at the asterisk, Uber_13_distance is statistically significant
# co-efficient of Uber_13_distance indicates with every additional
# increase in miles, lyft fare is predicted to
# increase by 0.95169$
# intercept is also significant: even when there is no distance covered,
# minimum fare is -4.70592$
```

11. Which model is better? Please explain your answer (0.5 pts)

```
#uber regression model is better since adjusted R square is more when
# compared to lyft regression model
```

12. What would be your model's prediction of the Lyft fare for a 2.39 mile trip? (1 pt).

```
predDF <- data.frame(Lyft_13_distance=2.39)
predict(lyft_reg, predDF)
```

```
##        1
## 22.49015
```

13. Generate a map where each state is shaded according to the average fare for Uber. Make sure even states with no data are visible on your map (2 pts)

```
library(maps)
```

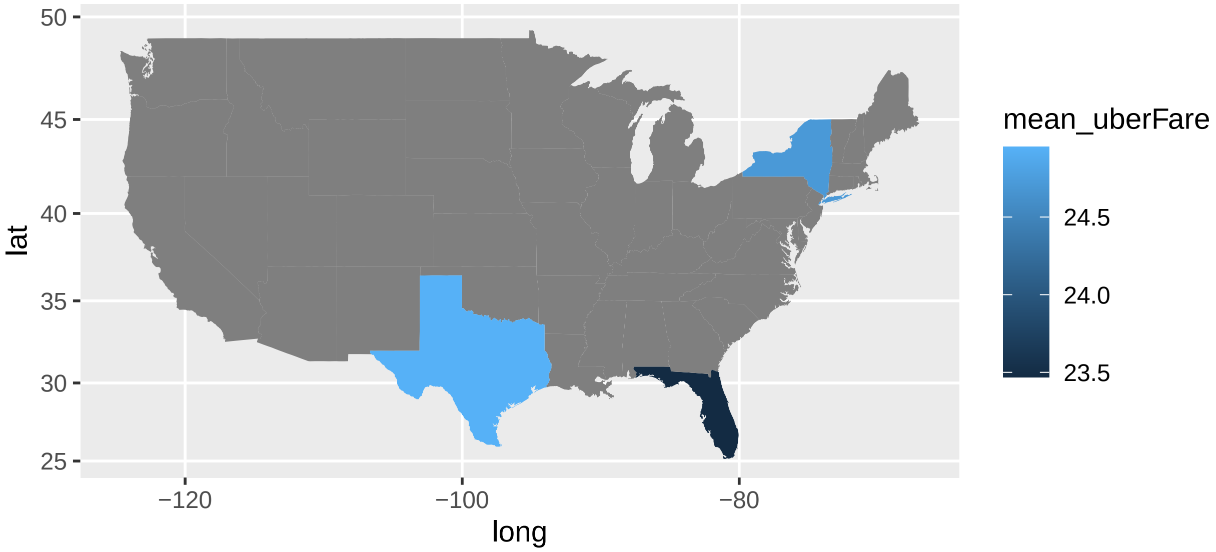
```
##
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':
##
##     map
```

```
library(mapproj)
US_States <- map_data("state")

# ensuring state name to be consistent across data frames
testDF$Uber_13_state <- testDF$Uber_13_state %>% tolower()
mergedUberDF <- merge(US_States, testDF,
                      all.x = TRUE, by.x="region", by.y="Uber_13_state")

# ensure order is sorted in ascending
mergedUberDF <- mergedUberDF %>% arrange(order)
dfSimple <- mergedUberDF %>% group_by(region)%>%
  summarize(mean_uberFare = mean(Uber_13))
finalmerge <- merge(mergedUberDF, dfSimple, all.x = TRUE,
                    by.x="region", by.y="region")
finalmerge %>% ggplot() + aes(x=long, y=lat, group=group,
                              fill=mean_uberFare) +
  geom_polygon() + coord_map()
```



14. Include a comment indicating whether or not you think Lyft and Uber fares are related based only on your data analysis. If the distributions of Lyft fares and Uber fares look similar and the distribution of the differences variable is normal and the X-Y scatterplot shows a clear pattern or relationship, then they may be related, i.e. they may be coordinating prices (1 pt).

```
# uber and lyft are not related
# distributions of Lyft fares and Uber fares look similar but,
# distribution of the differences variable does not seem to be normal
# X-Y scatterplot shows a weak relationship
```