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 <- 889489533
```

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: 889489533
## Sample size for this study: 115
```

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 **seven variables**: They account for the **ride number**, the **fare** (in dollars and cents) of a ride offered to a driver from Lyft and Uber, as well as the **distance** of those rides (in miles). There is also a **The State the ride was taken in** from Uber and Lyft). 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)

```r
#testDF
subset(testDF, driver==5, select = (Lyft_35))
```

```
##   Lyft_35
## 5   29.55
```

```r
testDF[5,2]
```

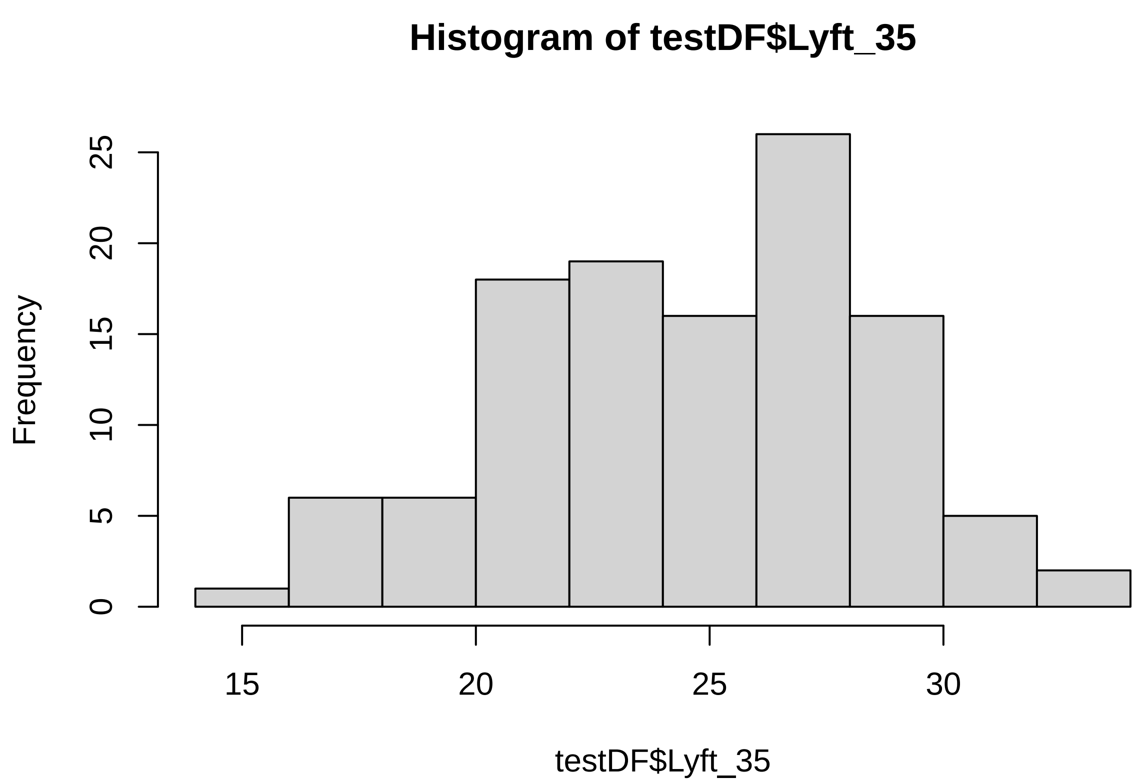
```
## [1] 29.55
```

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

```r
#summary(testDF) #Summary of the whole testDF dataframe
summary(testDF$Lyft_35)  #Summary of the Lyft_35 column of the dataframe
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.91   21.62   24.81   24.63   27.42   33.24
```

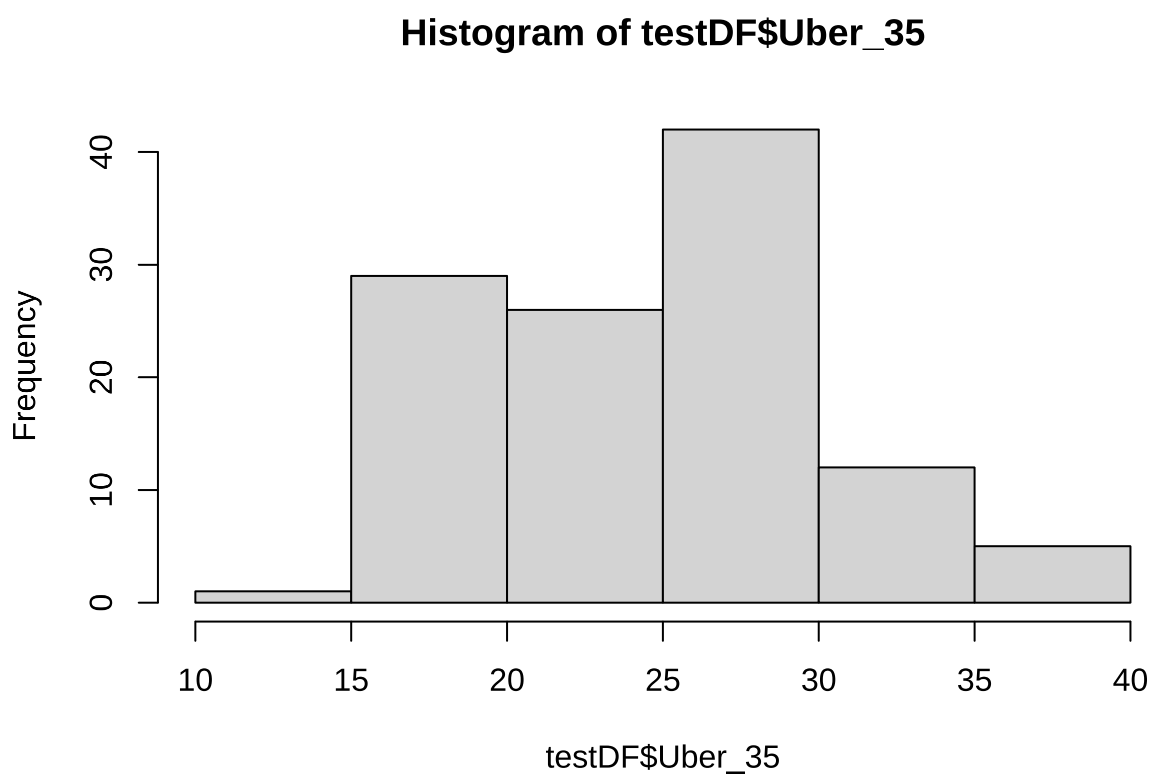
```r
hist(testDF$Lyft_35) #Histogram
```



```r
summary(testDF$Uber_35) #Summary of the Uber_35 column of the dataframe
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.51   19.32   25.17   24.75   29.16   37.63
```

```r
hist(testDF$Uber_35) #Histogram
```



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

```r
#The shape of the distribution for Lyft fares is pretty much normally distributed, with a min of 15.91,

#The shape of the distribution for Lyft fares is also normally distributed, with a min of 14.51, max of
```

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)

```r
mean_Lyft_35 <- mean(testDF$Lyft_35)
mean_Uber_35 <-mean(testDF$Uber_35)
mean_Lyft_35
```

```
## [1] 24.62817
```

```
mean_Uber_35
```

```
## [1] 24.75435
```

```
#We can observe that average Uber fares are higher than that of Lyft fares. Thus, Uber is generally mor
difference <- mean_Uber_35 - mean_Lyft_35
difference
```

```
## [1] 0.1261739
```

```
#Average Uber fares are higher than Lyft fares by $0.12617.
```

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$diff <- testDF$Uber_35 - testDF$Lyft_35
testDF
```

```
##    driver Lyft_35 Uber_35 Lyft_35_distance Uber_35_distance Lyft_35_state
## 1       1   26.46   28.66        63.770791         35.15218       Florida
## 2       2   17.86   29.88        22.701646         36.13890         Texas
## 3       3   24.45   19.08        36.321004         23.74384      New York
## 4       4   27.15   19.14        62.364721         26.71386      New York
## 5       5   29.55   31.02        63.684445         37.86171         Texas
## 6       6   20.74   25.05        32.014158         30.70382      New York
## 7       7   27.80   27.83        60.039761         35.16039         Texas
## 8       8   23.51   16.26        20.071887         21.18680       Florida
## 9       9   22.30   25.17        43.342825         32.75983      New York
## 10     10   25.74   21.67        84.490662         27.09633         Texas
## 11     11   27.24   26.87        76.967172         32.98336       Florida
## 12     12   28.04   29.14        47.006770         36.53638      New York
## 13     13   21.19   27.83        43.915779         34.80935      New York
## 14     14   24.81   21.91        29.432940         28.94104         Texas
## 15     15   26.13   26.95        60.595981         35.43072       Florida
## 16     16   28.44   33.38        19.200787         39.23079         Texas
## 17     17   23.86   17.89         4.388303         22.31568       Florida
## 18     18   22.94   29.52        53.677322         37.02667      New York
## 19     19   24.13   14.51        52.107999         21.58759         Texas
## 20     20   20.81   16.94        64.359109         24.28997         Texas
## 21     21   22.84   24.60        78.398380         31.96528         Texas
## 22     22   27.86   29.78        74.747588         35.06849         Texas
## 23     23   16.42   22.49        54.510339         26.23378      New York
## 24     24   30.79   23.15        27.782055         29.98007         Texas
## 25     25   29.15   23.06        34.290939         29.93104       Florida
## 26     26   22.18   23.13        52.217394         31.31574      New York
## 27     27   26.54   20.93        29.416690         26.56547      New York
## 28     28   31.02   25.79        84.547909         31.00649         Texas
## 29     29   33.24   37.63        38.901216         43.91328      New York
## 30     30   25.70   18.78        49.278921         23.53185      New York
## 31     31   22.36   18.33        54.060381         28.38043      New York
## 32     32   21.63   24.39        77.716822         31.42080       Florida
```

```
## 33        33  16.82  17.41        68.436793        22.90503     New York
## 34        34  21.59  19.31        58.572150        24.40327        Texas
## 35        35  23.38  25.92        37.723282        33.93535      Florida
## 36        36  26.12  31.60        50.061625        35.37445      Florida
## 37        37  28.81  17.05        69.850761        21.69321      Florida
## 38        38  29.68  35.47        46.240381        42.38979        Texas
## 39        39  25.42  26.67        54.113559        30.48443     New York
## 40        40  22.87  19.34        44.720780        25.87562        Texas
## 41        41  25.84  26.12        25.732217        35.11915      Florida
## 42        42  27.46  21.22        44.004867        26.66737      Florida
## 43        43  21.90  26.35        37.465067        33.84197        Texas
## 44        44  17.36  16.81        39.937051        22.14689        Texas
## 45        45  21.51  31.11        43.809603        37.70396      Florida
## 46        46  21.13  30.38        42.016452        38.56709        Texas
## 47        47  22.04  19.20        27.600094        28.65345        Texas
## 48        48  24.02  20.48        37.319986        26.98586        Texas
## 49        49  20.57  22.86        67.106135        27.51521     New York
## 50        50  27.57  30.39        76.550688        36.85256        Texas
## 51        51  23.12  19.18        43.966484        27.75506     New York
## 52        52  27.38  28.99        70.161971        34.92453        Texas
## 53        53  23.84  15.77        67.803279        23.12944     New York
## 54        54  29.60  33.93        75.245372        40.29426        Texas
## 55        55  22.80  28.12        55.968850        36.02117        Texas
## 56        56  20.51  27.50         8.999394        32.30784        Texas
## 57        57  27.55  33.60        43.517791        40.46107        Texas
## 58        58  26.31  28.26        51.173245        36.11737     New York
## 59        59  21.08  29.63        19.523510        35.94380     New York
## 60        60  28.41  31.54        48.113501        36.87524     New York
## 61        61  26.48  27.51        12.218052        32.89430     New York
## 62        62  19.41  26.71       101.311550        33.14916        Texas
## 63        63  21.08  18.98        39.578077        24.90000        Texas
## 64        64  19.20  16.42        74.615584        24.55165      Florida
## 65        65  23.40  21.01        59.107794        27.15312      Florida
## 66        66  26.43  25.23        50.544154        32.30993        Texas
## 67        67  22.81  22.52        29.812751        28.91299     New York
## 68        68  26.56  27.02        86.313872        33.22396     New York
## 69        69  25.16  29.26        25.038360        34.72565      Florida
## 70        70  24.34  29.68        59.490317        36.24331      Florida
## 71        71  28.81  31.54        76.298221        37.30273        Texas
## 72        72  23.38  27.50        55.325583        33.21883        Texas
## 73        73  30.50  36.73        40.495889        41.95958        Texas
## 74        74  29.64  29.39        87.101071        35.13040     New York
## 75        75  24.42  18.78        43.140847        24.73021        Texas
## 76        76  21.26  30.56        71.546722        37.36229      Florida
## 77        77  31.12  35.05        22.484211        40.47531        Texas
## 78        78  15.91  22.31        34.387315        29.97805      Florida
## 79        79  17.97  25.26        22.679629        31.53091      Florida
## 80        80  18.31  25.83        73.621583        32.06184     New York
## 81        81  26.77  29.01        51.220561        35.04548     New York
## 82        82  21.62  26.42        40.584678        33.62210      Florida
## 83        83  23.86  19.21        34.325855        25.41997     New York
## 84        84  16.82  18.63        54.937428        23.57925     New York
## 85        85  26.56  17.93        64.885562        23.81206      Florida
## 86        86  29.30  29.04        65.619232        34.51223      Florida
```

```
## 87       87   22.68   17.17         66.533183          23.95076        Florida
## 88       88   28.13   29.88         60.261951          33.01051        Florida
## 89       89   26.81   33.06         29.549442          40.51030          Texas
## 90       90   26.22   26.98         52.447600          31.60866       New York
## 91       91   25.45   20.87         53.389619          28.96279        Florida
## 92       92   27.88   30.00         52.388646          38.68978       New York
## 93       93   26.41   28.75          0.000000          33.51567          Texas
## 94       94   28.92   24.99         43.230302          29.68967       New York
## 95       95   29.94   21.23         46.790176          28.63599        Florida
## 96       96   29.50   18.36        101.306633          25.01995          Texas
## 97       97   25.43   23.10         82.002283          28.64902       New York
## 98       98   24.44   29.17         50.784485          33.16226       New York
## 99       99   20.35   20.90         52.167684          28.33590          Texas
## 100     100   18.54   20.27         47.968475          27.10845       New York
## 101     101   26.26   26.33         54.338782          34.16845          Texas
## 102     102   26.73   24.27         60.580043          28.77240       New York
## 103     103   19.49   17.79         30.458760          25.07885          Texas
## 104     104   31.38   20.81         68.847653          28.85298          Texas
## 105     105   21.30   17.80         40.691952          22.89772          Texas
## 106     106   25.17   18.90         37.787527          25.46384        Florida
## 107     107   26.78   21.47         74.264306          28.09948       New York
## 108     108   24.03   29.87         27.881436          35.82321          Texas
## 109     109   21.52   22.02         23.356135          29.04139       New York
## 110     110   18.48   16.59         84.958348          21.38529          Texas
## 111     111   32.54   35.37         72.149413          42.00545          Texas
## 112     112   23.95   17.35         95.813296          23.27680          Texas
## 113     113   26.05   29.35         40.453184          36.41615       New York
## 114     114   28.30   23.55         49.355202          28.68437          Texas
## 115     115   20.87   18.05         58.790638          25.07892          Texas
##     Uber_35_state   diff
## 1         Florida   2.20
## 2           Texas  12.02
## 3        New York  -5.37
## 4        New York  -8.01
## 5           Texas   1.47
## 6        New York   4.31
## 7           Texas   0.03
## 8         Florida  -7.25
## 9        New York   2.87
## 10          Texas  -4.07
## 11        Florida  -0.37
## 12       New York   1.10
## 13       New York   6.64
## 14          Texas  -2.90
## 15        Florida   0.82
## 16          Texas   4.94
## 17        Florida  -5.97
## 18       New York   6.58
## 19          Texas  -9.62
## 20          Texas  -3.87
## 21          Texas   1.76
## 22          Texas   1.92
## 23       New York   6.07
## 24          Texas  -7.64
```

```
## 25        Florida  -6.09
## 26       New York   0.95
## 27       New York  -5.61
## 28          Texas  -5.23
## 29       New York   4.39
## 30       New York  -6.92
## 31       New York  -4.03
## 32        Florida   2.76
## 33       New York   0.59
## 34          Texas  -2.28
## 35        Florida   2.54
## 36        Florida   5.48
## 37        Florida -11.76
## 38          Texas   5.79
## 39       New York   1.25
## 40          Texas  -3.53
## 41        Florida   0.28
## 42        Florida  -6.24
## 43          Texas   4.45
## 44          Texas  -0.55
## 45        Florida   9.60
## 46          Texas   9.25
## 47          Texas  -2.84
## 48          Texas  -3.54
## 49       New York   2.29
## 50          Texas   2.82
## 51       New York  -3.94
## 52          Texas   1.61
## 53       New York  -8.07
## 54          Texas   4.33
## 55          Texas   5.32
## 56          Texas   6.99
## 57          Texas   6.05
## 58       New York   1.95
## 59       New York   8.55
## 60       New York   3.13
## 61       New York   1.03
## 62          Texas   7.30
## 63          Texas  -2.10
## 64        Florida  -2.78
## 65        Florida  -2.39
## 66          Texas  -1.20
## 67       New York  -0.29
## 68       New York   0.46
## 69        Florida   4.10
## 70        Florida   5.34
## 71          Texas   2.73
## 72          Texas   4.12
## 73          Texas   6.23
## 74       New York  -0.25
## 75          Texas  -5.64
## 76        Florida   9.30
## 77          Texas   3.93
## 78        Florida   6.40
```

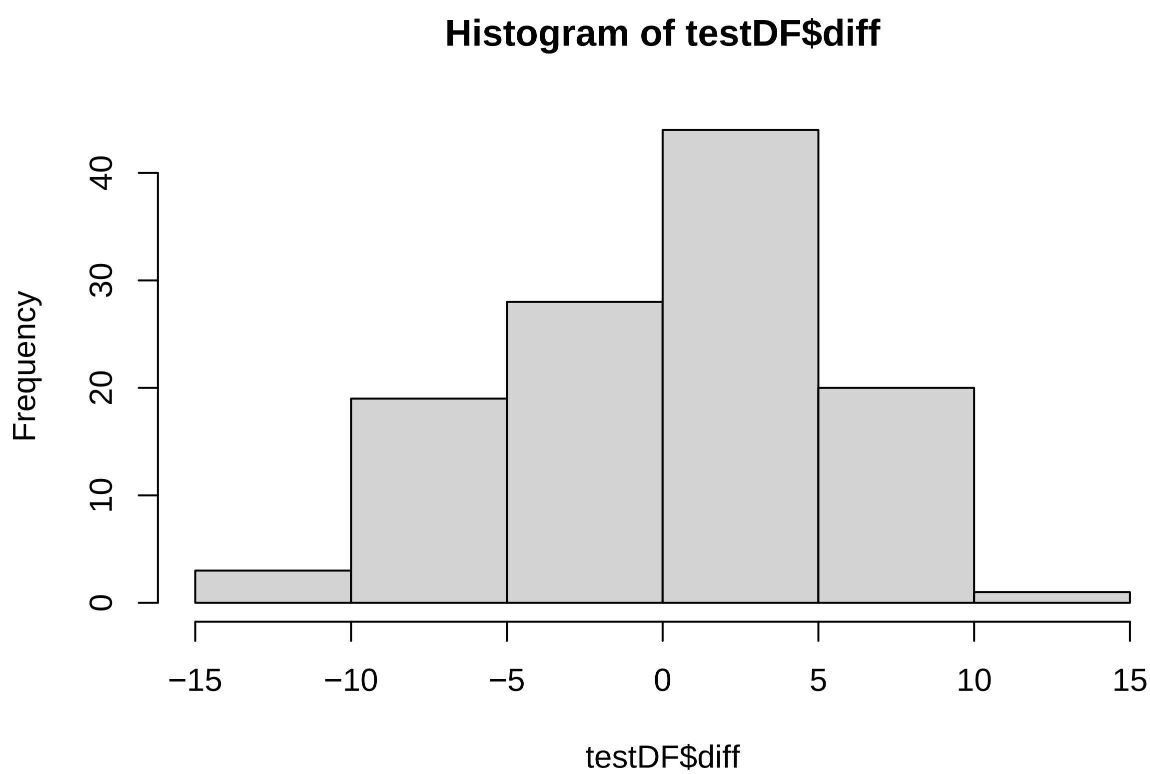
```
## 79        Florida   7.29
## 80       New York   7.52
## 81       New York   2.24
## 82        Florida   4.80
## 83       New York  -4.65
## 84       New York   1.81
## 85        Florida  -8.63
## 86        Florida  -0.26
## 87        Florida  -5.51
## 88        Florida   1.75
## 89          Texas   6.25
## 90       New York   0.76
## 91        Florida  -4.58
## 92       New York   2.12
## 93          Texas   2.34
## 94       New York  -3.93
## 95        Florida  -8.71
## 96          Texas -11.14
## 97       New York  -2.33
## 98       New York   4.73
## 99          Texas   0.55
## 100      New York   1.73
## 101         Texas   0.07
## 102      New York  -2.46
## 103         Texas  -1.70
## 104         Texas -10.57
## 105         Texas  -3.50
## 106       Florida  -6.27
## 107      New York  -5.31
## 108         Texas   5.84
## 109      New York   0.50
## 110         Texas  -1.89
## 111         Texas   2.83
## 112         Texas  -6.60
## 113      New York   3.30
## 114         Texas  -4.75
## 115         Texas  -2.82
```

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.
## -11.7600  -3.9000   0.7600   0.1262   4.1100  12.0200
```

```r
hist(testDF$diff)
```



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)

```r
sortedDF <- testDF[order(testDF$diff),]
#sortedDF
head(sortedDF, 5)
```

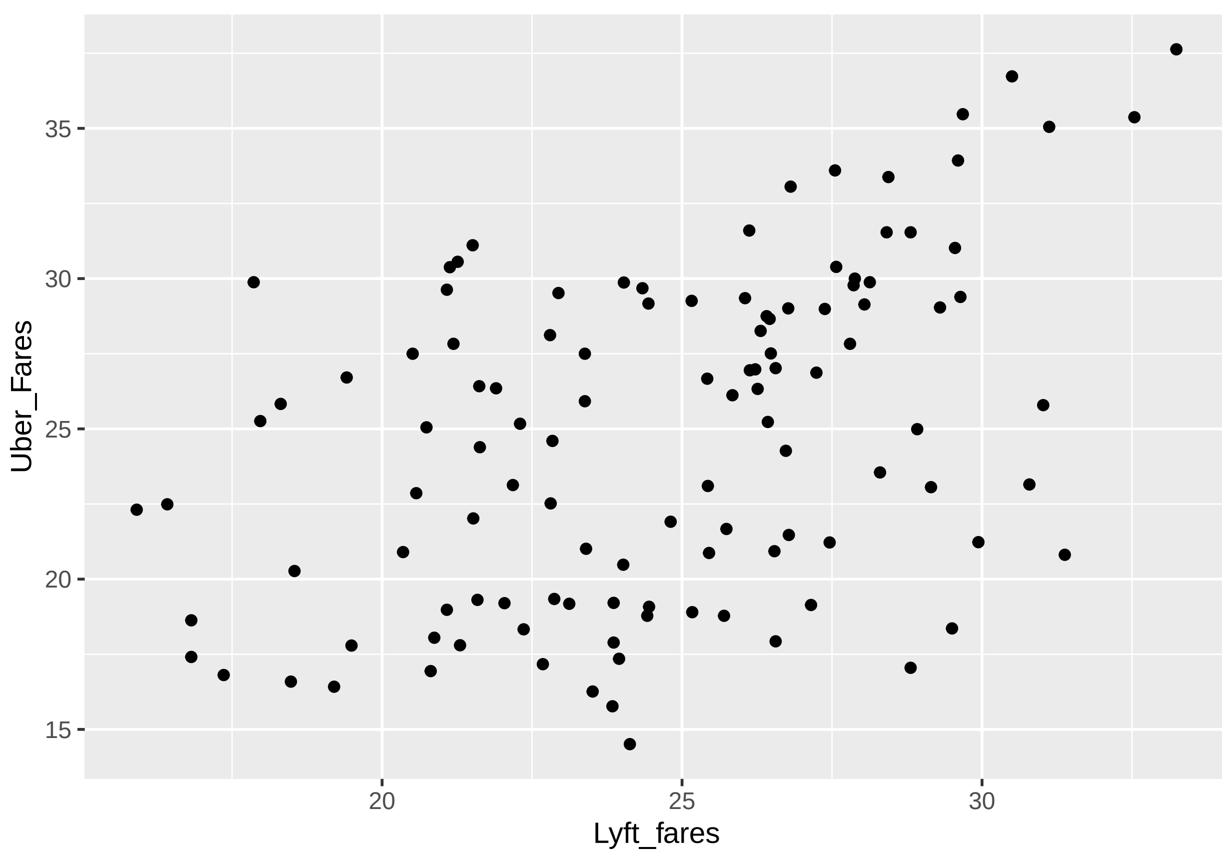
```
##     driver Lyft_35 Uber_35 Lyft_35_distance Uber_35_distance Lyft_35_state
## 37      37   28.81   17.05         69.85076         21.69321       Florida
## 96      96   29.50   18.36        101.30663         25.01995         Texas
## 104    104   31.38   20.81         68.84765         28.85298         Texas
## 19      19   24.13   14.51         52.10800         21.58759         Texas
## 95      95   29.94   21.23         46.79018         28.63599       Florida
##     Uber_35_state   diff
## 37        Florida -11.76
## 96          Texas -11.14
## 104         Texas -10.57
## 19          Texas  -9.62
## 95        Florida  -8.71
```

```r
tail(sortedDF, 5)
```

```
##    driver Lyft_35 Uber_35 Lyft_35_distance Uber_35_distance Lyft_35_state
## 59     59   21.08   29.63         19.52351         35.94380      New York
## 46     46   21.13   30.38         42.01645         38.56709         Texas
## 76     76   21.26   30.56         71.54672         37.36229       Florida
## 45     45   21.51   31.11         43.80960         37.70396       Florida
## 2       2   17.86   29.88         22.70165         36.13890         Texas
##    Uber_35_state  diff
## 59      New York  8.55
## 46         Texas  9.25
## 76       Florida  9.30
## 45       Florida  9.60
## 2          Texas 12.02
```

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

```r
library(ggplot2)
ggplot(testDF) + geom_point(aes(x=Lyft_35, y=Uber_35)) + xlab("Lyft_fares") + ylab("Uber_Fares")
```



```
#Yes, we can observe that there is indeed a relatively linear relationship of a positive slope between
```

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).

```
lmOut1 <- lm(formula = Lyft_35 ~ Lyft_35_distance, data=testDF)
summary(lmOut1)
```

```
##
## Call:
## lm(formula = Lyft_35 ~ Lyft_35_distance, data = testDF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3412 -2.6270  0.2817  2.7148  8.8876
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      23.47988    0.96200  24.407   <2e-16 ***
## Lyft_35_distance  0.02243    0.01745   1.285    0.201
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.821 on 113 degrees of freedom
## Multiple R-squared:  0.0144, Adjusted R-squared:  0.005681
## F-statistic: 1.651 on 1 and 113 DF,  p-value: 0.2014
```

```
#Lyft_35 = 0.02243 * Lyft_35_distance + 23.47988 (y=mx+b)
# Coefficient (m) is 0.03333 and the intercept (b) is 23.38682
# Lyft_13_distance is not statistically significant as the p value(0.2014) is greater than 0.05
# Adjusted R-squared is approximately 0.005681
```

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

```
lmOut2 <- lm(formula = Uber_35 ~ Uber_35_distance, data=testDF)
summary(lmOut2)
```

```
##
## Call:
## lm(formula = Uber_35 ~ Uber_35_distance, data = testDF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8059 -0.8153  0.0515  0.8842  3.3022
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5.0909     0.6730  -7.564 1.12e-11 ***
## Uber_35_distance   0.9594     0.0213  45.041  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.263 on 113 degrees of freedom
## Multiple R-squared:  0.9472, Adjusted R-squared:  0.9468
## F-statistic:  2029 on 1 and 113 DF,  p-value: < 2.2e-16
```

```
#Uber_35 = 0.9594 * Uber_35_distance - 5.0909 (y=mx+b)
# Coefficient (m) is 0.9594 and the intercept (b) is -5.0909
# Lyft_13_distance is statistically significant as the p value (2.2e-16) is lesser than 0.05
# Adjusted R-squared is approximately 0.005681
```

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

```
#The Uber model is significantly better as we can see that all the coefficients are highly significant
```

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

```
predDF <- data.frame(Lyft_35_distance=2.39)
predict(lmOut1, predDF)
```

```
##        1
## 23.53349
```

```
# it will cost $23.53349
```

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)

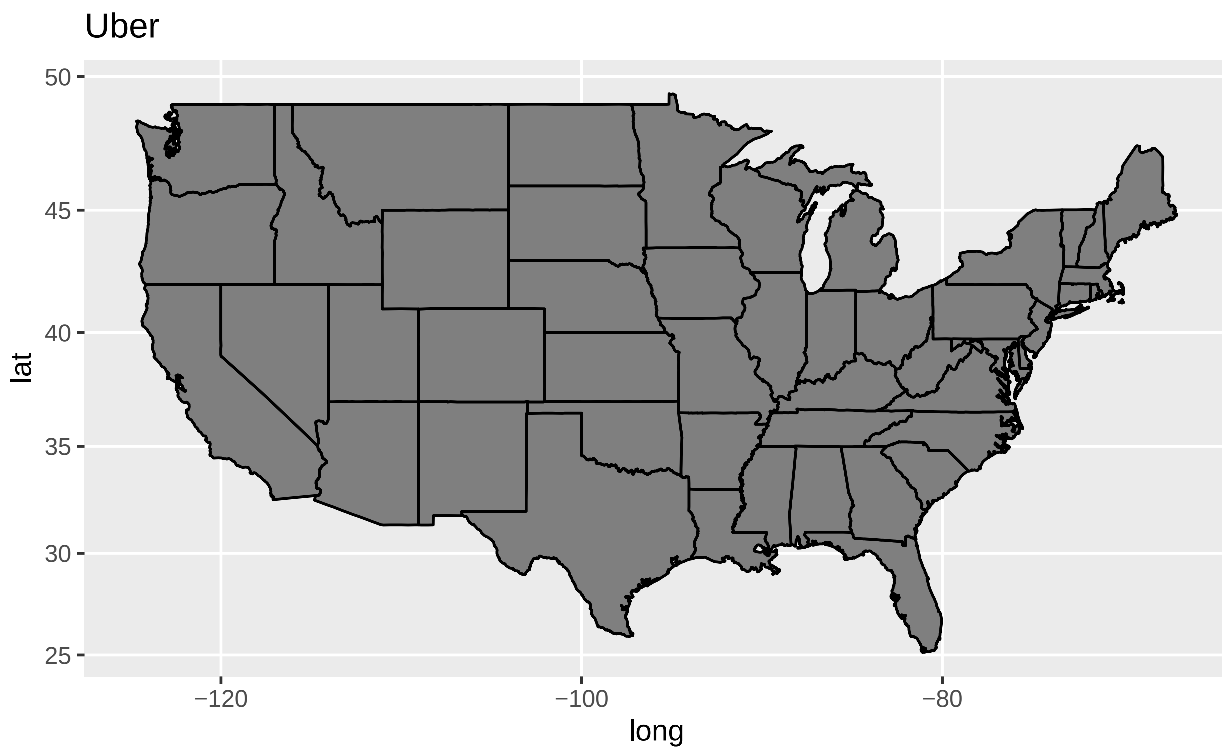
```
library(tidyverse)
```

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

## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
## v purrr   0.3.4

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

```
sortedDF <- sortedDF %>% group_by(Uber_35_state)
step111 <- aggregate(x=sortedDF$Uber_35, by= list(sortedDF$Uber_35_state), mean)
sortedDF$Uber_35_state <- tolower(sortedDF$Uber_35_state)
us <- map_data("state")
mergeUsData1000 <- merge(us, step111, by.x = "region", by.y = "Group.1", all.x = TRUE)
mergeUsData1000 <- mergeUsData1000 %>% arrange(order)
map40 <- ggplot(mergeUsData1000)
map40 <- map40 + geom_polygon(color="black", aes(x=long,y=lat, group=group, fill=x))
map40 <- map40 + coord_map() + ggtitle("Uber")
map40
```



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).

```
# According to me, the analysis of both of the fare distributions are normal and similar. When we analy
#I also observed that when we are predicting fares with respect to distance, the linear model of Uber i
```