STAT 311 Lab 2

Group 17

Steven Tran & Owen Cheung

Tuesday, July 26, 2022 @ 11:59pm

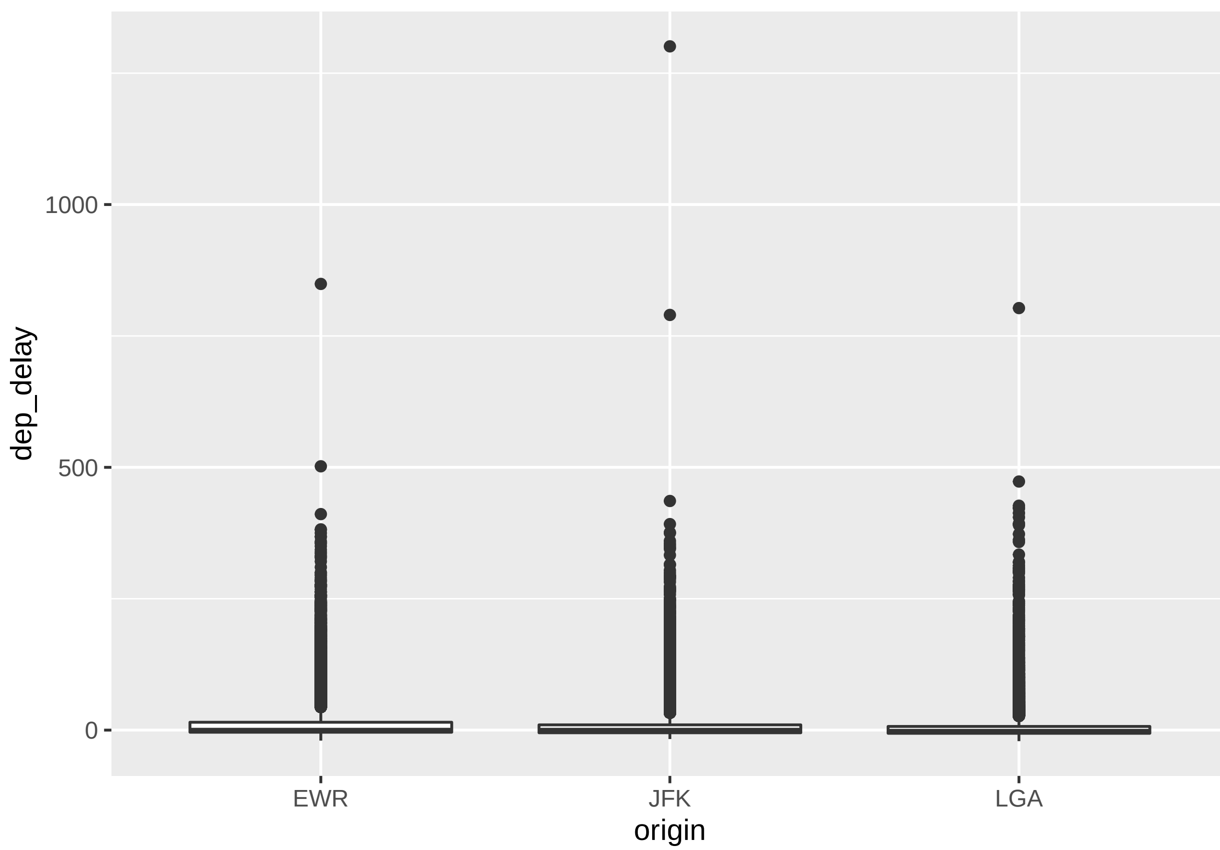
Exercise 1: R Practice

Part a)

```
data = nycflights

bplots <- ggplot(data, aes(x=origin, y=dep_delay)) +
  geom_boxplot()

bplots
```



I think that one reason this might not be the most useful visual as the boxplots are all extremely hard to read due to the scale of the plots and because the data points are all clustered together. You really can't the difference between all 3 nor the distrubutions.

Part b)

```
data = nycflights

mean_and_med <- data %>%
  group_by(origin) %>%
  summarize(mean=mean(dep_delay), median=median(dep_delay))

mean_and_med
```

```
## # A tibble: 3 x 3
##   origin  mean median
##   <chr>  <dbl>  <int>
## 1 EWR     15.3     -1
## 2 JFK     12.3     -1
## 3 LGA     10.1     -3
```

Part c)

One explanation for the large difference between the means and the medians would be that there are a lot of negative departure delays as well as positive departure delays in the hundreds so if you took a mean/average of that group you would be more likely to get a positive number leaning more towards the positive departure delays while if you took a median of the group its reasonable to get a negative number since there are a large quantity of negative departure delays in the group. I think that I would want to use mean since it might be more representative of the spread of the data

Another explanation could be that there are a lot of outliers in the data. The big difference between the means and the medians could probably be explained by these outliers. Even in the boxplot, there seems to be a large portion of the data clustered around the 0 departure delay time and some outliers as big as 1000. Thus, with this in mind, it's better to use the median as it is less susceptible to large outliers.

Part d)

```
data = nycflights

newset <- data %>%
  mutate(delayed = dep_delay > 5) #this should work

head(newset)
```

```
##   year month day dep_time dep_delay arr_time arr_delay carrier tailnum flight
## 1 2013     6  30      940        15     1216        -4      VX  N626VA    407
## 2 2013     5   7     1657        -3     2104        10      DL  N3760C    329
## 3 2013    12   8      859        -1     1238        11      DL  N712TW    422
## 4 2013     5  14     1841        -4     2122       -34      DL  N914DL   2391
```

```
## 5 2013     7  21     1102        -3     1230         -8       9E  N823AY   3652
## 6 2013     1   1     1817        -3     2008          3       AA  N3AXAA    353
##   origin dest air_time distance hour minute delayed
## 1    JFK  LAX      313     2475    9     40    TRUE
## 2    JFK  SJU      216     1598   16     57   FALSE
## 3    JFK  LAX      376     2475    8     59   FALSE
## 4    JFK  TPA      135     1005   18     41   FALSE
## 5    LGA  ORF       50      296   11      2   FALSE
## 6    LGA  ORD      138      733   18     17   FALSE
```

Part e)

```
data = nycflights

newset2 <- newset %>%
  group_by(origin) %>%
  #add_count() %>%
  summarize(total=n(), num_delayed=sum(delayed))

newset2
```

```
## # A tibble: 3 x 3
##   origin total num_delayed
##   <chr>  <int>       <int>
## 1 EWR    11771        4093
## 2 JFK    10897        3212
## 3 LGA    10067        2625
```

Part f)

EWR had the highest percentage of delayed flights with 35% compared to JFK with 29% and LGA with 26%

Exercise 2: Exercise and General Health

Part a)

```
contingency <- table(cdc$exerany, cdc$genhlth)
print.table(contingency)
```

```
##
##     poor fair good very good excellent
##   n  384  857 1731      1352       762
##   y  293 1162 3944      5620      3895
```

```r
contingency_margins <- addmargins(contingency)
print.table(contingency_margins)
```

```
##
##         poor  fair  good very good excellent   Sum
##   n      384   857  1731      1352       762  5086
##   y      293  1162  3944      5620      3895 14914
##   Sum    677  2019  5675      6972      4657 20000
```

Part b)

```r
contingency <- table(cdc$exerany, cdc$genhlth)
print.table(contingency)
```

```
##
##     poor fair good very good excellent
##   n  384  857 1731      1352       762
##   y  293 1162 3944      5620      3895
```

```r
contingency_margins <- addmargins(contingency)
print.table(contingency_margins)
```

```
##
##         poor  fair  good very good excellent   Sum
##   n      384   857  1731      1352       762  5086
##   y      293  1162  3944      5620      3895 14914
##   Sum    677  2019  5675      6972      4657 20000
```

```r
contingency_prop <- prop.table(contingency)
print(contingency_prop)
```

```
##
##         poor    fair    good very good excellent
##   n 0.01920 0.04285 0.08655   0.06760   0.03810
##   y 0.01465 0.05810 0.19720   0.28100   0.19475
```

The proportion of those who have exercised in the past month which is the yes to the variable exerany is 14914/20000, which is 0.7457. The proportion of the sample reporting excellent health is 4657/20000 or 0.23285. These numbers are supported by both `contingency_prop` and `contingency_margins`.

Part c)

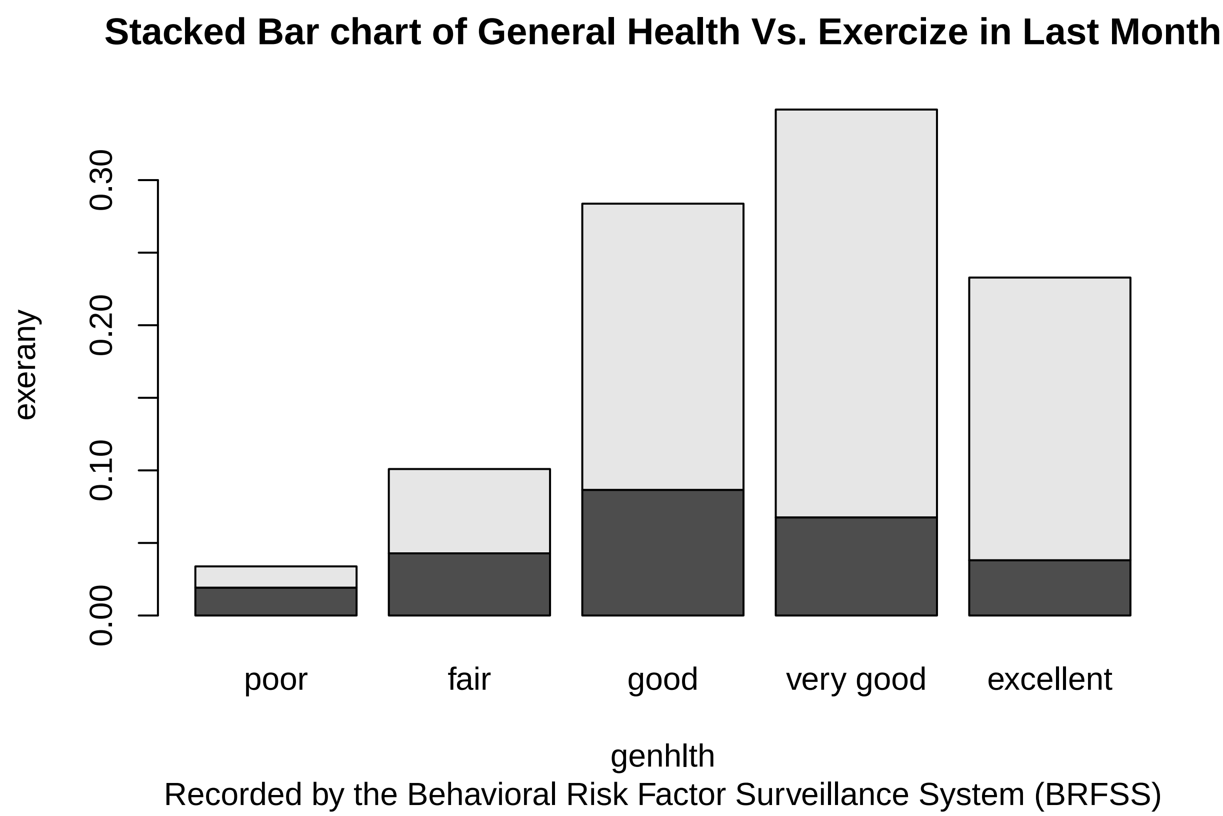
```r
contingency_prop <- prop.table(contingency)
print(contingency_prop)
```

```
##
##         poor    fair    good very good excellent
##   n 0.01920 0.04285 0.08655   0.06760   0.03810
##   y 0.01465 0.05810 0.19720   0.28100   0.19475
```

Among the people who exersised in the past month, a proportion of about 0.19475 of them reported excellent health. Among the people who didn't exersize in the past month, only a proportion of 0.03810 of them reported excellent health.

Part d)

```
bar_chart <- barplot(contingency_prop,
                     main = "Stacked Bar chart of General Health Vs. Exercize in Last Month",
                     sub = "Recorded by the Behavioral Risk Factor Surveillance System (BRFSS)",
                     xlab = "genhlth",
                     ylab = "exerany",
                     axes = TRUE)
```

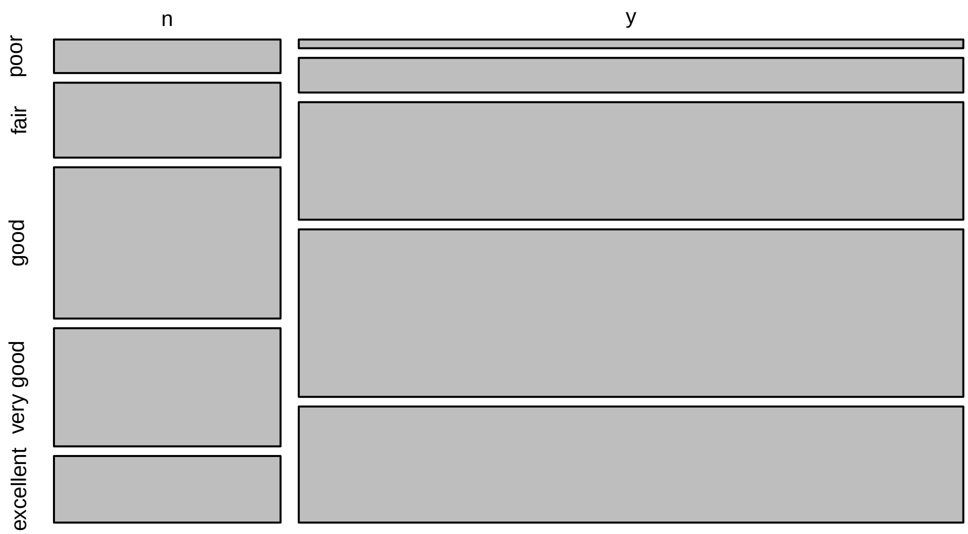


It seems as though most people would rate their health as very good, though those who exercize more make up more of the excellent, very good, good, and fair categories. I also see some response bias here where those who responded tended to have better health, which may perhaps mean that they had strong feelings for this topic. There doesn't seem to be a large 'no exercize group'.

Part e)

```
mosaic_plot <- mosaicplot(contingency_prop,
                          main = "Mosaic Plot of General Health Vs. Exercize in Last Month",
                     sub = "Recorded by the Behavioral Risk Factor Surveillance System (BRFSS)")
```

Mosaic Plot of General Health Vs. Exercize in Last Month



Recorded by the Behavioral Risk Factor Surveillance System (BRFSS)

Based on the plots, both the stacked bar chart and the mosaic plot, it seems as though those who exercized in the past month rated their general health higher than the population that didn't. It also seems like the people who didn't exercize may have been subjected to wording bias, maybe they felt bad for having not exercized. There is also response and non-response bias where there is a lot more responses from the people who exercized, and those who didn't exercize perhaps wouldn't have worse health than shown here on the graph.

Part f)

No, it doesn't seem like the two variables exerany and genhealth are independent. The people who said they exercized in the past month tend to indicate a higher level of general health. However, all we can say is that there is a correlation, not a causation, because there many be many other confounding variables that may impact a person's general health.

Exercise 3: More Research Questions

Research Question 1

Proposed Question

What is the relationship between mother's age and lengths of pregnacy in weeks?

Proposed Statistical Method Since there are 2 numerical variables, we can use a histogram to analyze the data. We chose this method because we feel as though it works the best with 2 numerical variables. Also, we wanted to know if the mother's age had an influence in the development of the child, or if older mothers would produce more premie babies.

We could also use a scatter plot with x being mother's age and y being lengths of pregnancy. It just depends on which method shows the correlation between the 2 variables best.

Research Question 2

Proposed Question

Is there a relationship between the maturity status of the mother and the premie status of the baby?

Proposed Statistical Method Since there are 2 categorical variables, in the maturity status of the mother and the premie status of the baby, we can use a chi-squared test. Using a chi-square test, we can test if there is a correlation between the staus of the mother and premie status of the baby, or if it is due to random chance.

Research Question 3

Proposed Question

Is there a trend between the weight gained by the baby measured in pounds and the smoking status of the mother?

Proposed Statistical Method In this question, there is 1 categorical and 1 numerical variable. As such, we can run a t-test to see if the average weight gained by the baby when the mother smokes is greater or less than when the mother doesn't smoke.