Initial Analysis

Kyle Ligon

Read the data in and library calls

```
library(tidyverse)
library(zoo)
library(lubridate)

crime <- read_csv("BPD_Part_1_Victim_Based_Crime_Data.csv", progress = FALSE)
```

Looking at the data

```
head(crime)
```

```
## # A tibble: 6 x 15
##   CrimeDate CrimeTime CrimeCode Location    Description   `Inside/Outside`
##   <chr>     <time>    <chr>     <chr>       <chr>         <chr>
## 1 9/2/2017  23:30     3JK       4200 AUDRE~ ROBBERY - RE~ I
## 2 9/2/2017  23:00     7A        800 NEWING~ AUTO THEFT    O
## 3 9/2/2017  22:53     9S        600 RADNOR~ SHOOTING      Outside
## 4 9/2/2017  22:50     4C        1800 RAMSA~ AGG. ASSAULT  I
## 5 9/2/2017  22:31     4E        100 LIGHT ~ COMMON ASSAU~ O
## 6 9/2/2017  22:00     5A        CHERRYCRES~ BURGLARY      I
## # ... with 9 more variables: Weapon <chr>, Post <int>, District <chr>,
## #   Neighborhood <chr>, Longitude <dbl>, Latitude <dbl>, `Location
## #   1` <chr>, Premise <chr>, `Total Incidents` <int>
```

```
names(crime)
```

```
##  [1] "CrimeDate"       "CrimeTime"       "CrimeCode"
##  [4] "Location"        "Description"     "Inside/Outside"
##  [7] "Weapon"          "Post"            "District"
## [10] "Neighborhood"    "Longitude"       "Latitude"
## [13] "Location 1"      "Premise"         "Total Incidents"
```

Looks like we have information about the crime, where it happened, when it happened, what happened in the form of Description, and the responding Post.

Counting up the Number of Crimes

```
descCounts <- crime %>%
                group_by(Description) %>%
                tally() %>%
                arrange(desc(n))
descCounts
```

```
## # A tibble: 15 x 2
##    Description             n
##    <chr>               <int>
##  1 LARCENY             60528
```

```
##  2 COMMON ASSAULT       45518
##  3 BURGLARY             42538
##  4 LARCENY FROM AUTO    36295
##  5 AGG. ASSAULT         27513
##  6 AUTO THEFT           26838
##  7 ROBBERY - STREET     17691
##  8 ROBBERY - COMMERCIAL  4141
##  9 ASSAULT BY THREAT     3503
## 10 SHOOTING              2910
## 11 ROBBERY - RESIDENCE   2866
## 12 RAPE                  1637
## 13 HOMICIDE              1559
## 14 ROBBERY - CARJACKING  1528
## 15 ARSON                 1464
```

Counting up Where the Crimes Occurred

```
hoodCounts <- crime %>%
                group_by(Neighborhood) %>%
                tally() %>%
                arrange(desc(n))
head(hoodCounts)
```

```
## # A tibble: 6 x 2
##   Neighborhood            n
##   <chr>               <int>
## 1 Downtown             9048
## 2 Frankford            6642
## 3 Belair-Edison        5977
## 4 Brooklyn             4516
## 5 Cherry Hill          4086
## 6 Sandtown-Winchester  4026
```

Let's focus on Arsons. Particurlarly, let's see if the number of Arsons committed in one month are more varied in the winter months than in the other nine months of the year. For this I will:

- Summarize the number of arsons by month
- Run an F test on number of arsons between the two groups
- Write a conclusion for the test

1) Are the distributions of arsons in the winter months less varied that other 9 months?

```
arsons <- crime %>%
            filter(Description == "ARSON")
head(arsons)
```

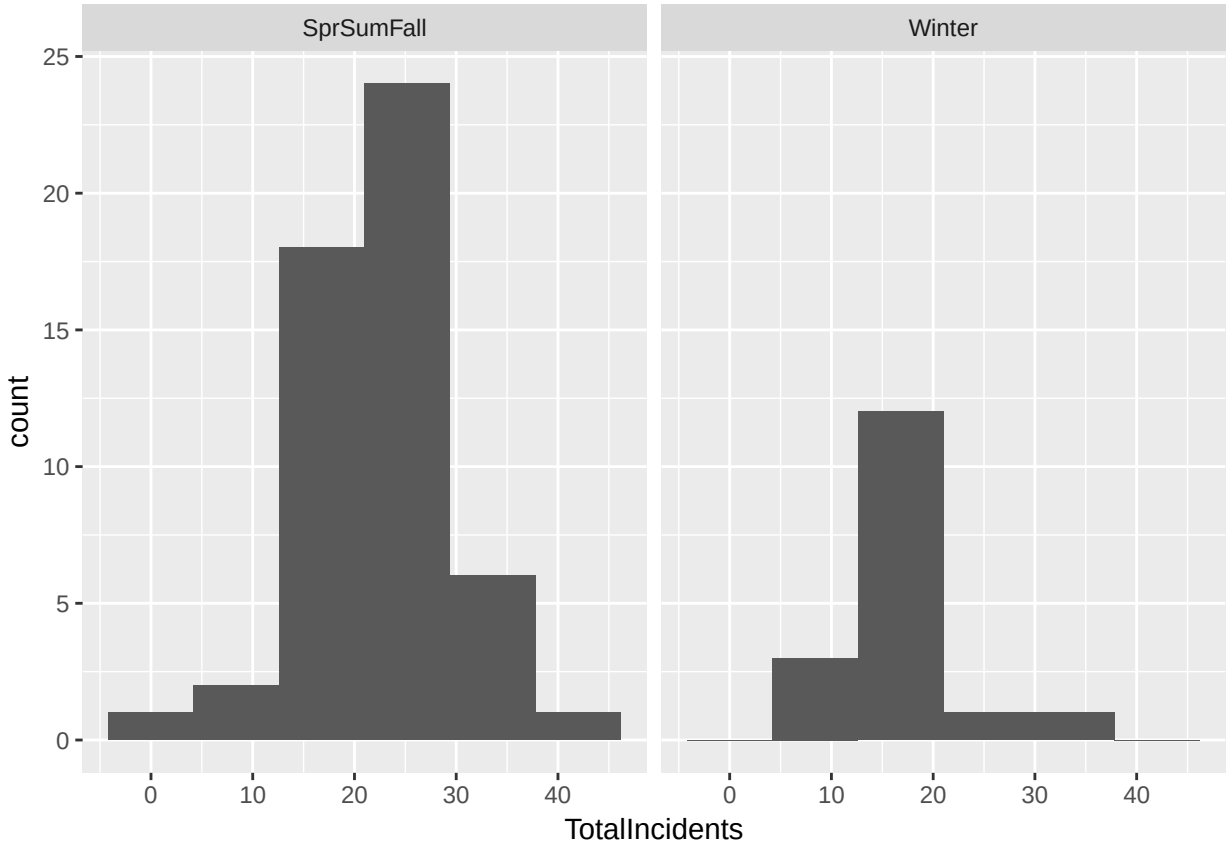
```
## # A tibble: 6 x 15
##   CrimeDate CrimeTime CrimeCode Location      Description `Inside/Outside`
##   <chr>     <time>    <chr>     <chr>         <chr>       <chr>
## 1 9/1/2017  22:00     8AO       300 N FREMON~ ARSON       I
## 2 8/30/2017 22:00     8H        2600 FLORA ST ARSON       <NA>
## 3 8/30/2017 19:30     8H        3700 CLIFTMO~ ARSON       O
```

```
## 4 8/30/2017 15:26     8AV       4600 PARK HE~ ARSON       I
## 5 8/29/2017 03:30     8H        800 RAPPOLLA~ ARSON       <NA>
## 6 8/28/2017 06:50     8H        3300 TIVOLY ~ ARSON       O
## # ... with 9 more variables: Weapon <chr>, Post <int>, District <chr>,
## #   Neighborhood <chr>, Longitude <dbl>, Latitude <dbl>, `Location
## #   1` <chr>, Premise <chr>, `Total Incidents` <int>
```

Checking the dimensions of the new frame

```
dim(arsons)
```

```
## [1] 1464   15
```



With "normal" distributions, we will proceed with the hypotheses.

Hypotheses: H_0: var_Winter = var_SprSumFall H_1: var_Winter < var_SprSumFall

Variance of the Winter Months

```
#degrees of freedom for W
arsons_grouped %>% filter(WinterBin == "Winter") %>% tally() - 1
```

```
##    n
## 1 16
```

```
var_W
```

```
## # A tibble: 1 x 1
##   Variance
##      <dbl>
## 1     27.3
```

Variance of the Other Months

```
arsons_grouped %>% filter(WinterBin != "Winter") %>% tally() - 1
```

```
##    n
## 1 51
```

```
var_SSF
```

```
## # A tibble: 1 x 1
##   Variance
##      <dbl>
## 1     45.9
```

Test Statistic for F Test

```
f <- as.numeric(round(var_W/var_SSF, 4))
f
```

```
## [1] 0.5945
```

Rejection Region

```
qf(0.975, 16, 51)
```

```
## [1] 2.075301
```

```
f < qf(0.025, 16, 51)
```

```
## [1] FALSE
```

```
f > qf(0.975, 16, 51)
```

```
## [1] FALSE
```

Since our F Test Statistic is not larger and not smaller than the F Stat for alpha, we do not have enough evidence to reject the null hypothesis that the variances are equal. It does not appear that the the Winter months experience a less varied number of arsons than the other 9 months.

Are there more Auto Thefts on Weekends over Weekdays?

```
auto_thefts <- crime %>%
                filter(Description %in% c("ROBBERY - CARJACKING", "AUTO THEFT"))
head(auto_thefts)
```

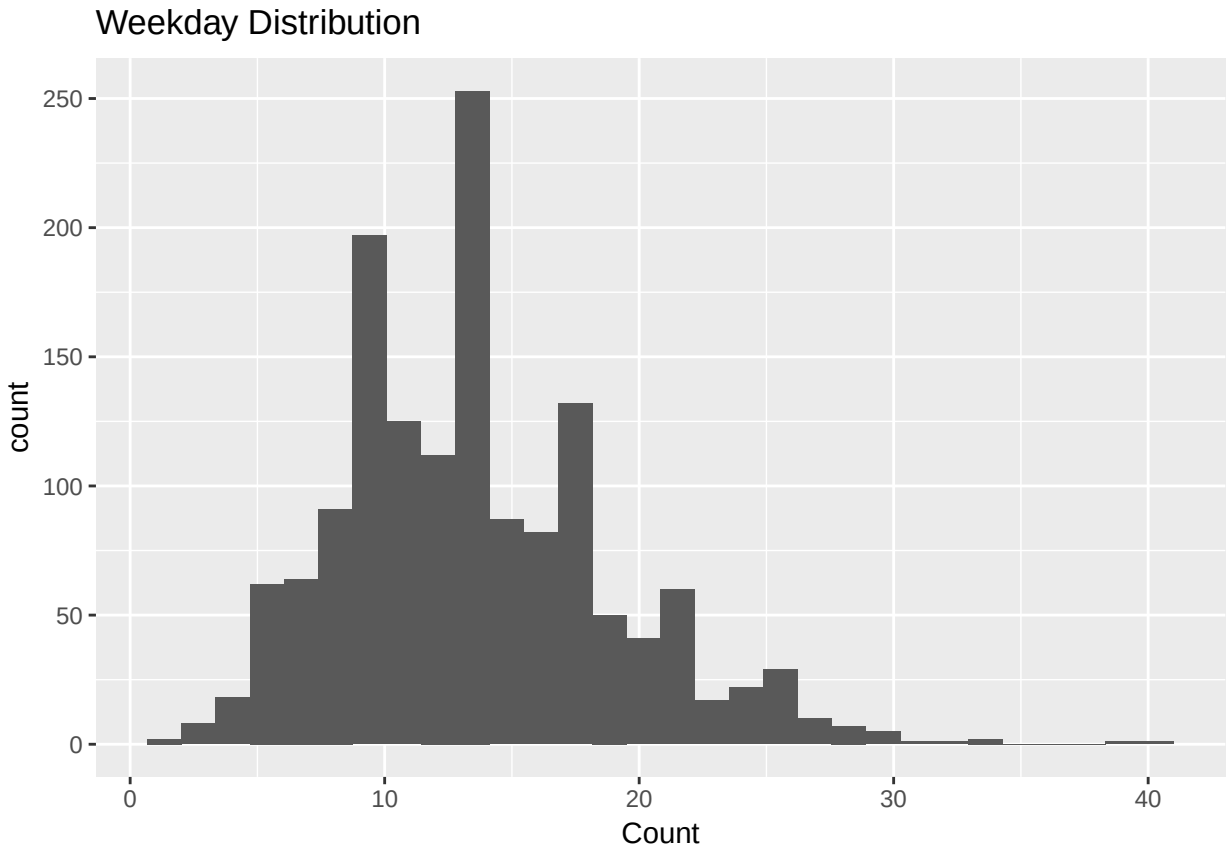
```
## # A tibble: 6 x 15
##   CrimeDate CrimeTime CrimeCode Location      Description `Inside/Outside`
##   <chr>     <time>    <chr>     <chr>         <chr>       <chr>
## 1 9/2/2017  23:00     7A        800 NEWINGTO~ AUTO THEFT  O
## 2 9/2/2017  08:00     7A        4700 HOMESDA~ AUTO THEFT  I
## 3 9/2/2017  02:00     7C        1500 RUSSELL~ AUTO THEFT  O
## 4 9/1/2017  22:30     7A        300 E LORRAI~ AUTO THEFT  O
## 5 9/1/2017  21:30     7A        3500 CHESTER~ AUTO THEFT  O
## 6 9/1/2017  20:45     7A        OSTEND ST & ~ AUTO THEFT  O
## # ... with 9 more variables: Weapon <chr>, Post <int>, District <chr>,
## #   Neighborhood <chr>, Longitude <dbl>, Latitude <dbl>, `Location
## #   1` <chr>, Premise <chr>, `Total Incidents` <int>
```

```
at_form <- auto_thefts %>%
                  mutate(CrimeDate= as.Date(CrimeDate, format = "%m/%d/%Y"),
```

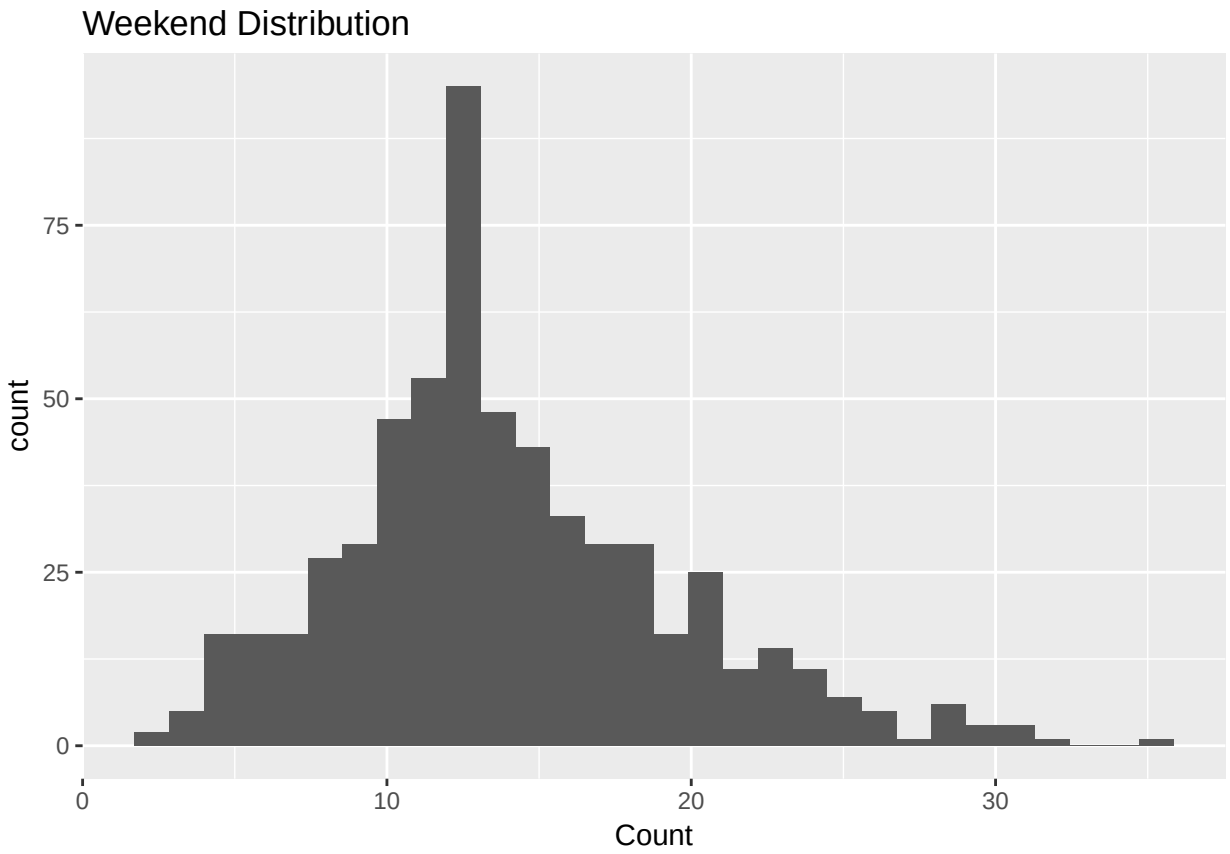
```
                          DateName = wday(CrimeDate, label = TRUE))
weekend <- at_form %>%
              filter(DateName %in% c("Sat", "Sun")) %>%
              group_by(CrimeDate, DateName) %>%
              summarize(Count = sum(`Total Incidents`))

weekday <- at_form %>%
              filter(!(DateName %in% c("Sat", "Sun"))) %>%
              group_by(CrimeDate, DateName) %>%
              summarize(Count = sum(`Total Incidents`))

ggplot(data = weekday, aes(x = Count)) + geom_histogram() + ggtitle("Weekday Distribution")
```




```
ggplot(data = weekend, aes(x = Count)) + geom_histogram() + ggtitle("Weekend Distribution")
```



With our data cleaned, we can now go about testing to see if the mean of the weekend set is larger than the mean of the weekday set. But first... equal variance check. Which, is just an F test:

H_0: Var_Weekday = var_Weekend H_1: var_Weekday < var_weekend

```
#Weekday Variance
var(weekday$Count)
```

```
## [1] 28.16756
```

```
#df of Weekday
nrow(weekday)-1
```

```
## [1] 1479
```

```
#Weekend Variance
var(weekend$Count)
```

```
## [1] 30.29062
```

```
#df of Weekends
nrow(weekend)-1
```

```
## [1] 591
```

Test Statistic

```
f_w <- var(weekday$Count)/var(weekend$Count)
f_w
```

```
## [1] 0.9299104
```

Rejection Region:

```
up <- qf(0.975, 1479, 591)
up
```

```
## [1] 1.146795
```

```
dwn <- qf(0.025, 1479, 591)
dwn
```

```
## [1] 0.8754564
```

```
f_w > up
```

```
## [1] FALSE
```

```
f_w < dwn
```

```
## [1] FALSE
```

Conclusion: Since our F Stat was less than the upper rejection region and more than the lower rejections region, we do not have enough evidence to reject the hypothesis that the variances are equal. It does not appear that the variances are different. Now we can test the means.

Hypotheses: H_0: mean_weekday = mean_weekend H_1: mean_weekday < mean_weekend

Test Statistic:

```
t.test(x = weekday$Count, y = weekend$Count)
```

```
##
##  Welch Two Sample t-test
##
## data:  weekday$Count and weekend$Count
## t = -1.5301, df = 1054.2, p-value = 0.1263
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9252956  0.1144848
## sample estimates:
## mean of x mean of y
##  13.57432  13.97973
```

Rejection Region

```
lower <- qt(0.05, 2070)
lower
```

```
## [1] -1.64559
```

```
-1.5301 < lower
```

```
## [1] FALSE
```

Conclusion: Since -1.5301 is greater than our rejection region, we do not have enough proof to reject our null hypothesis that the means are the same. It does not appear that the there are more Auto thefts on weekends in comparison to weekdays.

3) On a given night are there more shootings inside a residence or outside?

```
shootings <- crime %>%
#              mutate(CrimeTime = strptime(CrimeTime, format = "%H:%M")) %>%
```

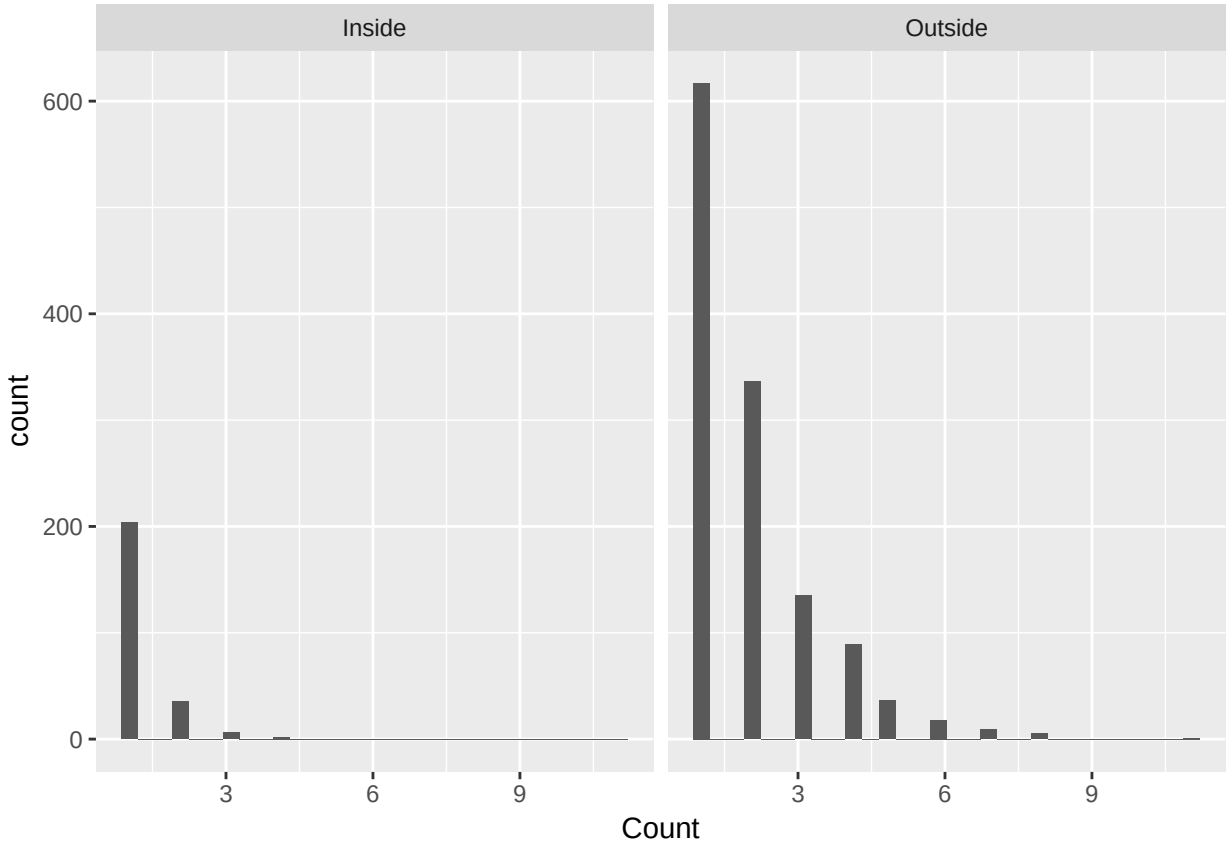
```
              filter(Description == "SHOOTING", CrimeTime > "17:00") %>%
              group_by(CrimeDate, `Inside/Outside`) %>%
              summarize(Count = sum(`Total Incidents`))
head(shootings)
```

```
## # A tibble: 6 x 3
## # Groups:   CrimeDate [5]
##   CrimeDate `Inside/Outside` Count
##   <chr>     <chr>            <int>
## 1 1/1/2012  Outside              1
## 2 1/1/2013  Inside               1
## 3 1/1/2013  Outside              1
## 4 1/1/2014  Outside              2
## 5 1/1/2015  Outside              1
## 6 1/1/2016  Outside              3
```

```
dim(shootings)
```

```
## [1] 1493    3
```

```
ggplot(shootings, aes(x = Count)) + geom_histogram() + facet_grid(~ `Inside/Outside`)
```



With non-normal distributions, our route that we should run down is to check to see if the medians are difference between the two groups. Using the Wilcoxon Rank Sum test, we'll see what if the outisde median is larger than the inside median.

```
wilcox.test(Count ~ `Inside/Outside`, data = shootings)
```

```
##
```

```
##  Wilcoxon rank sum test with continuity correction
##
## data:  Count by Inside/Outside
## W = 98797, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

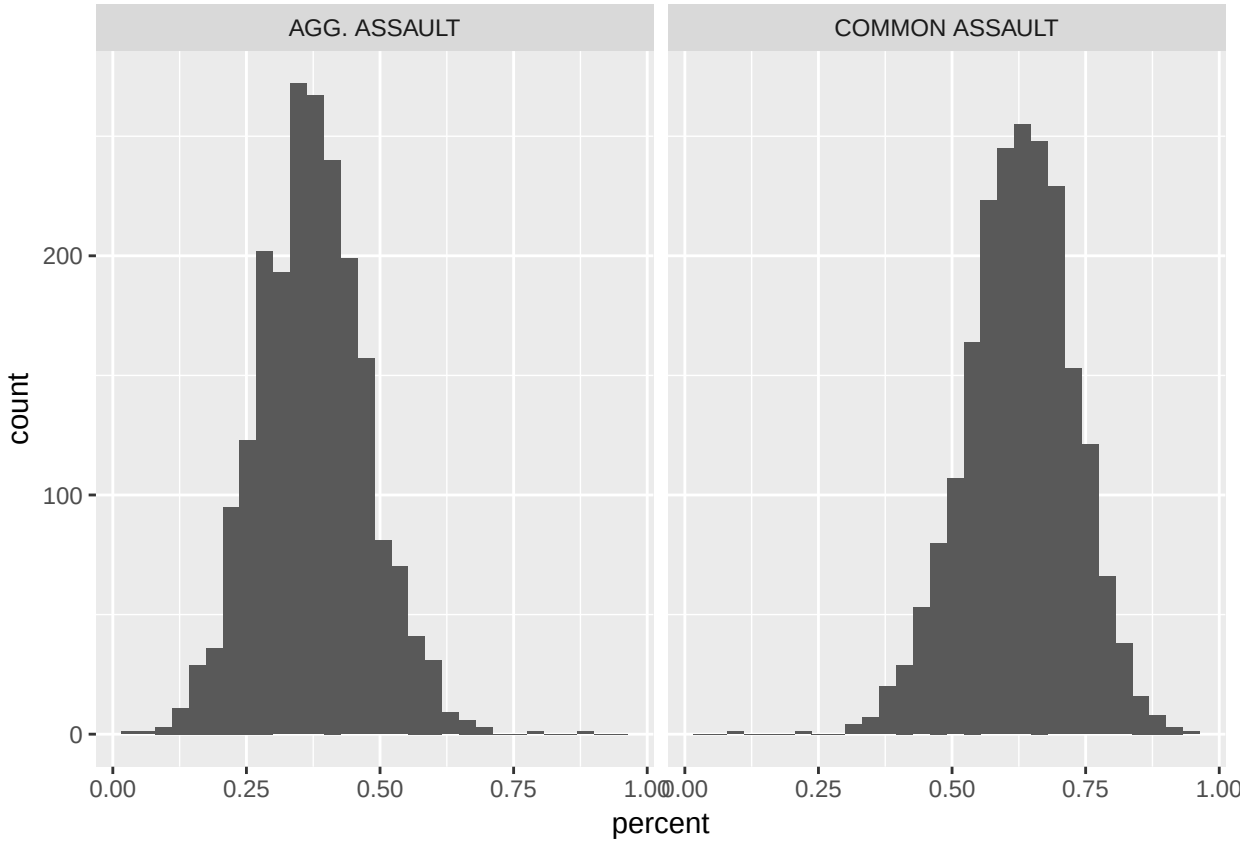
4) On a given day if you are assaulted Downtown s it more likely to be a common assault or aggravated assault?

```
assaults <- crime %>%
              filter(Description %in% c("AGG. ASSAULT", "COMMON ASSAULT")) %>%
              select(CrimeDate, Description, `Total Incidents`) %>%
              group_by(CrimeDate, Description) %>%
              summarize(Total = sum(`Total Incidents`)) %>%
              ungroup()

day_merge <- assaults %>%
              group_by(CrimeDate) %>%
              summarize(Count = sum(Total))

assaults <- assaults %>%
              group_by(CrimeDate) %>%
              left_join(day_merge) %>%
              mutate(percent = Total/Count) %>%
              ungroup() %>%
              mutate(CrimeDate = as.Date(CrimeDate, format = "%m/%d/%Y")) %>%
              arrange(desc(CrimeDate))

ggplot(data = assaults, aes(x =percent)) + geom_histogram() + facet_grid(~Description)
```


AGG. ASSAULT
COMMON ASSAULT
count
200
100
0
0.00
0.25
0.50
0.75
1.00
percent