Initial Analysis

Kyle Ligon

Read the data in and library calls

```r
library(tidyverse)
library(zoo)
library(lubridate)
library(gridExtra)

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

Looking at the data

```r
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>
```

```r
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

```r
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. Particularly, 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?

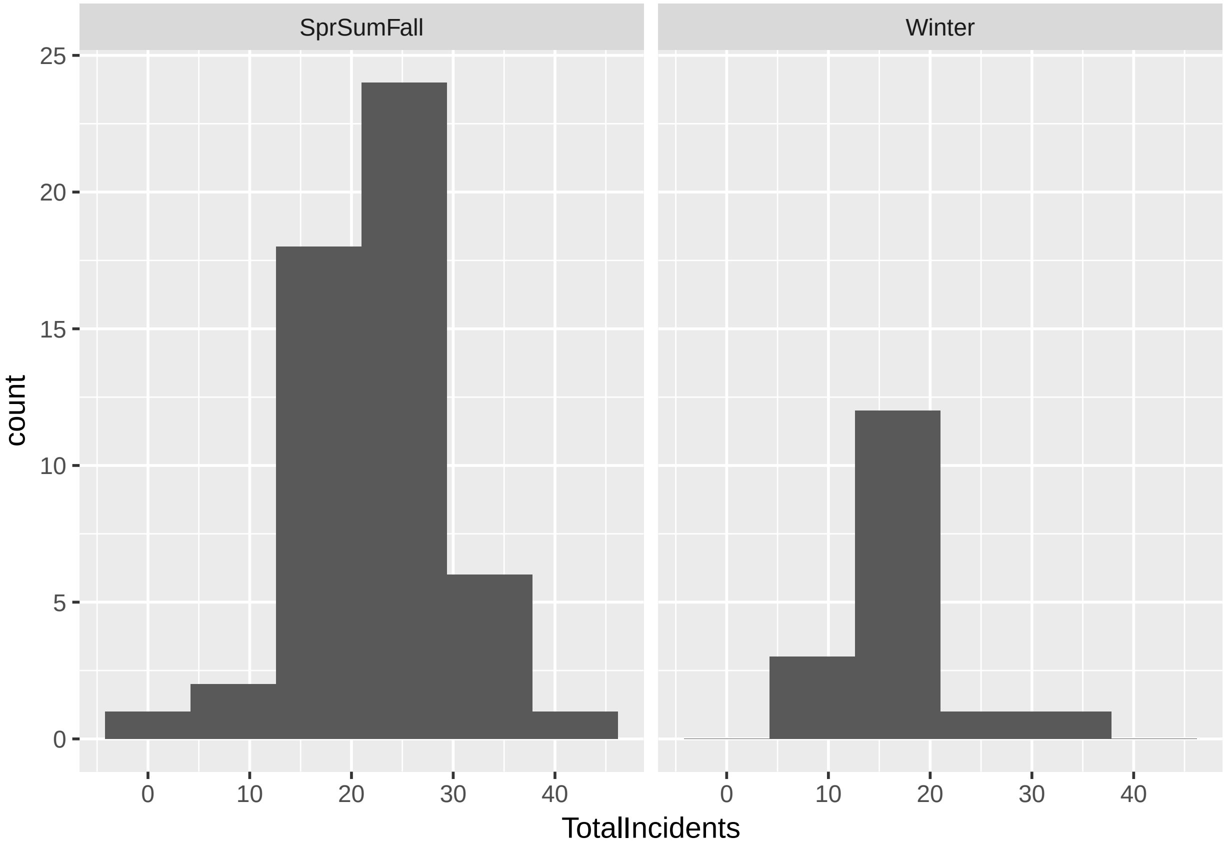
```
arsons <- crime %>%
            filter(Description == "ARSON")
head(arsons)
```

```
## # 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
```

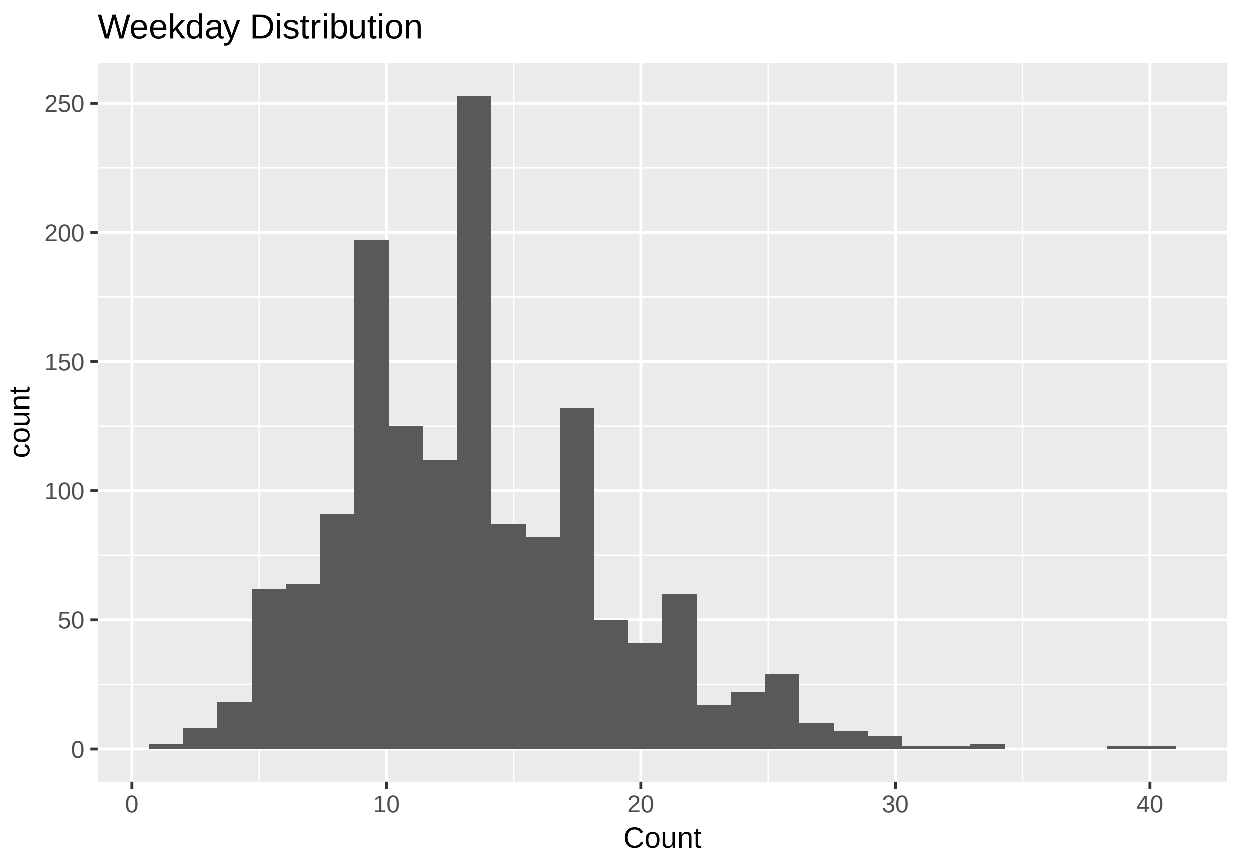
Since our F Test Statistic is not larger and not smaller than the F Stat for alpha, we do not have enough evidence to reject the null hypothesis that the variances are equal. It does not appear that the the Winter months experience a less varied number of arsons than the other 9 months.

2) Are there more Auto Thefts on Weekends over Weekdays?

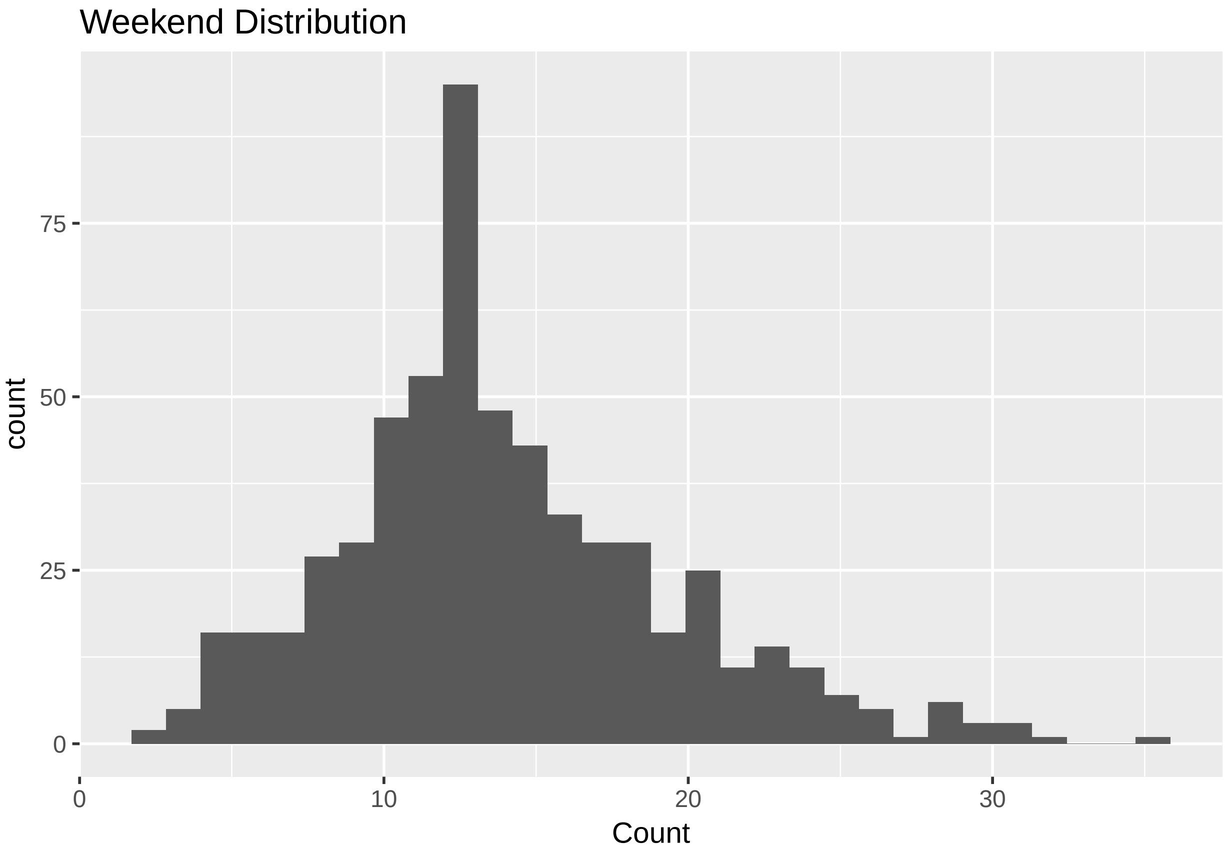
```
auto_thefts <- crime %>%
                filter(Description %in% c("ROBBERY - CARJACKING", "AUTO THEFT"))
head(auto_thefts)
```

```
## # 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>
```

```r
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")
```



```r
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 than outside?

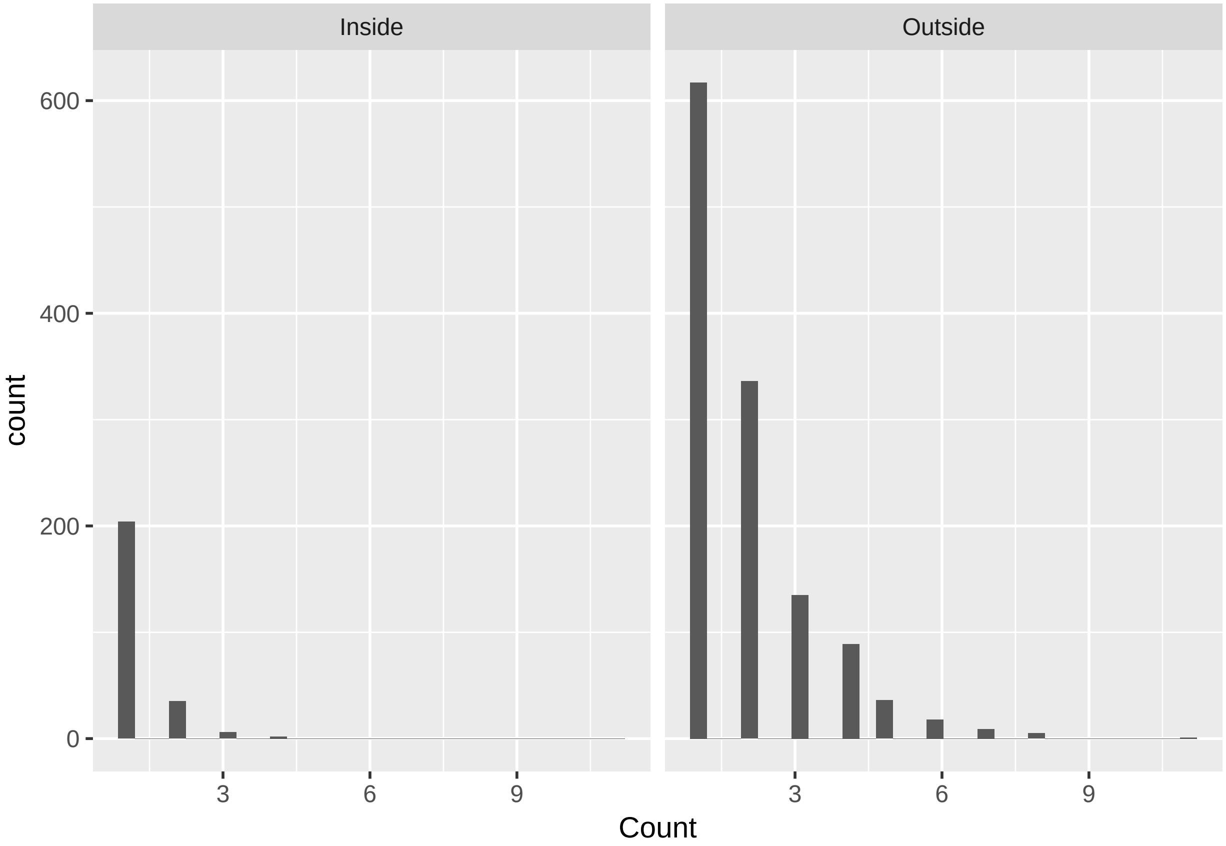
```
shootings <- crime %>%
#              mutate(CrimeTime = strptime(CrimeTime, format = "%H:%M")) %>%
              filter(Description == "SHOOTING", CrimeTime > "17:00") %>%
              group_by(CrimeDate, `Inside/Outside`) %>%
              summarize(Count = sum(`Total Incidents`))
head(shootings)
```

```
## # 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 different between the two groups. Using the Wilcoxon Rank Sum test, we'll see what if the outisde median is larger than the inside median.

Hypothesis: H_0: M_Inside = M_Outside H_1: M_Inside > M_Outside

```
w <- wilcox.test(Count ~ `Inside/Outside`, data = shootings, alternative = "greater")
```

Test Statistic

```
w$statistic
```

```
##     W
## 98797
```

Rejection Region

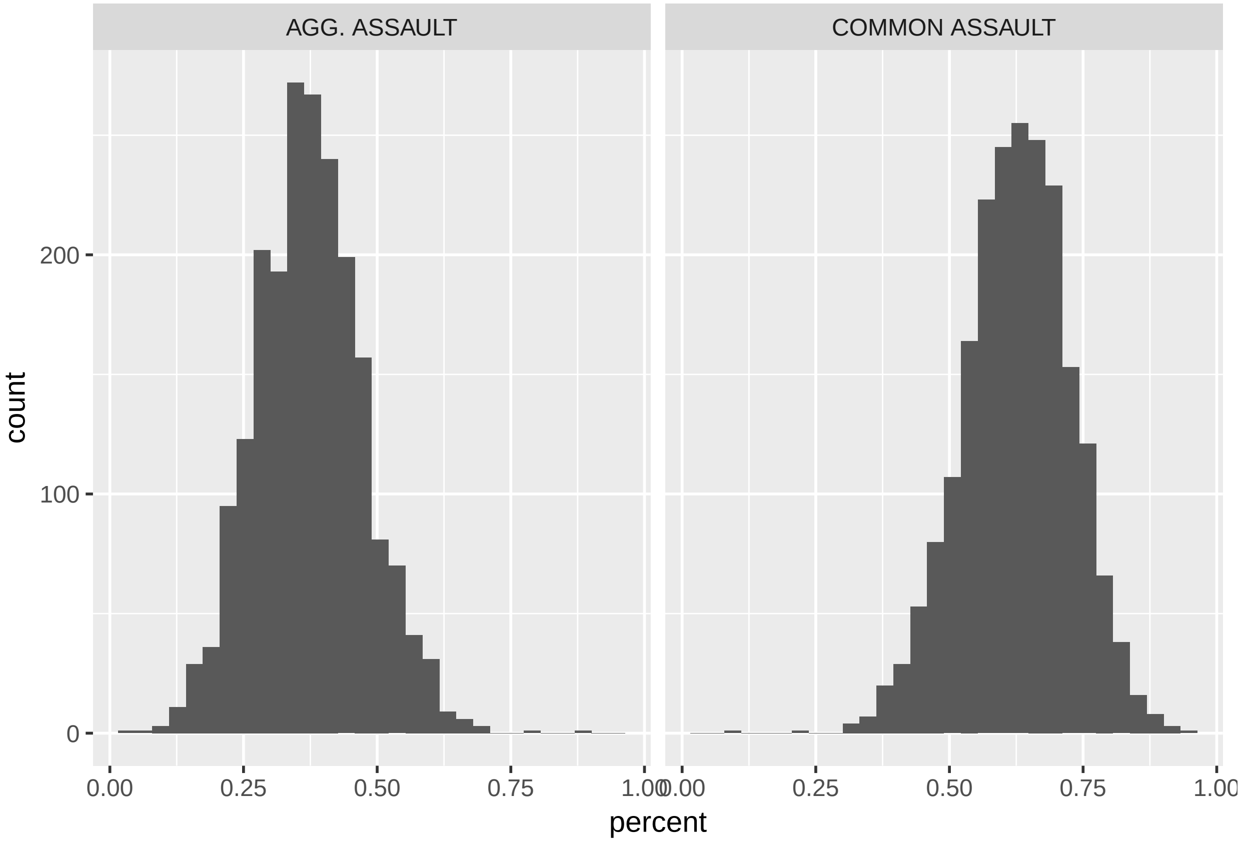
```
## [1] FALSE
```

Conclusion/Interpretation: Since our p-value is greater than 0.05, we do not have enough evidence to reject the null hypothesis that the Medians are the same. It does not appear the there are more shootings inside a residence than outside.

4) On a given day if you are assaulted, is it more likely to be a common assault?

Hypotheses: H_0: Mean_aggAssault = Mean_CommonAssault H_1: Mean_aggAssault < Mean_CommonAssault

```
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)
```



```
assaults_split <- t.test(formula = percent ~ Description, data = assaults, alternative = "less")
```

Test Statistic/p-value

```
assaults_split$statistic
```

```
##         t
## -80.19273
```

```
assaults_split$p.value
```

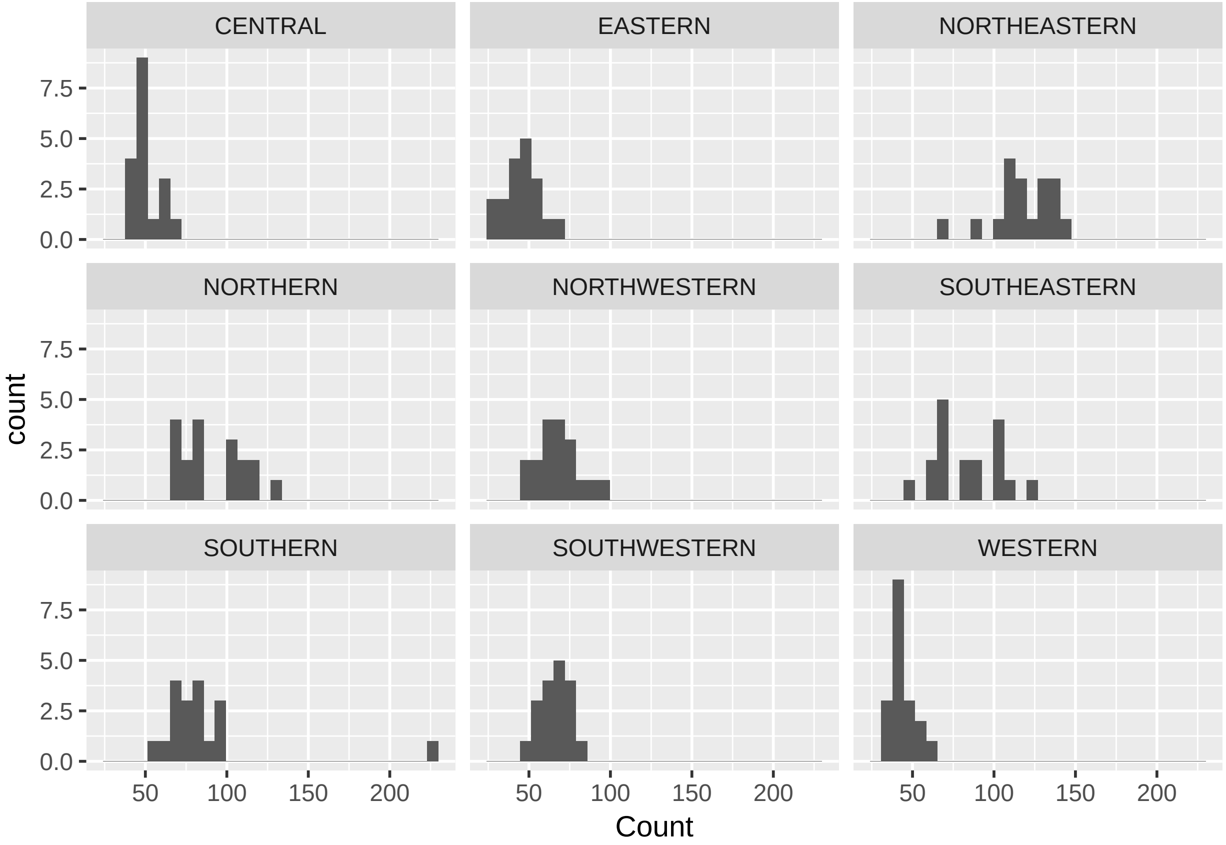
```
## [1] 0
```

Conclusion: Since our p-value is less than 0.05, we have sufficient evidence to reject the null hypothesis that the means are the same. It appears that the Mean percent of the Aggravated Assault group is less than the Mean percent of the Common Assault.

5) Is there a District with a distinctly higher number of Burglary's in the Summer Months? If so, which are significantly different?

H_0: means are all the same H_1: at least one mean is different

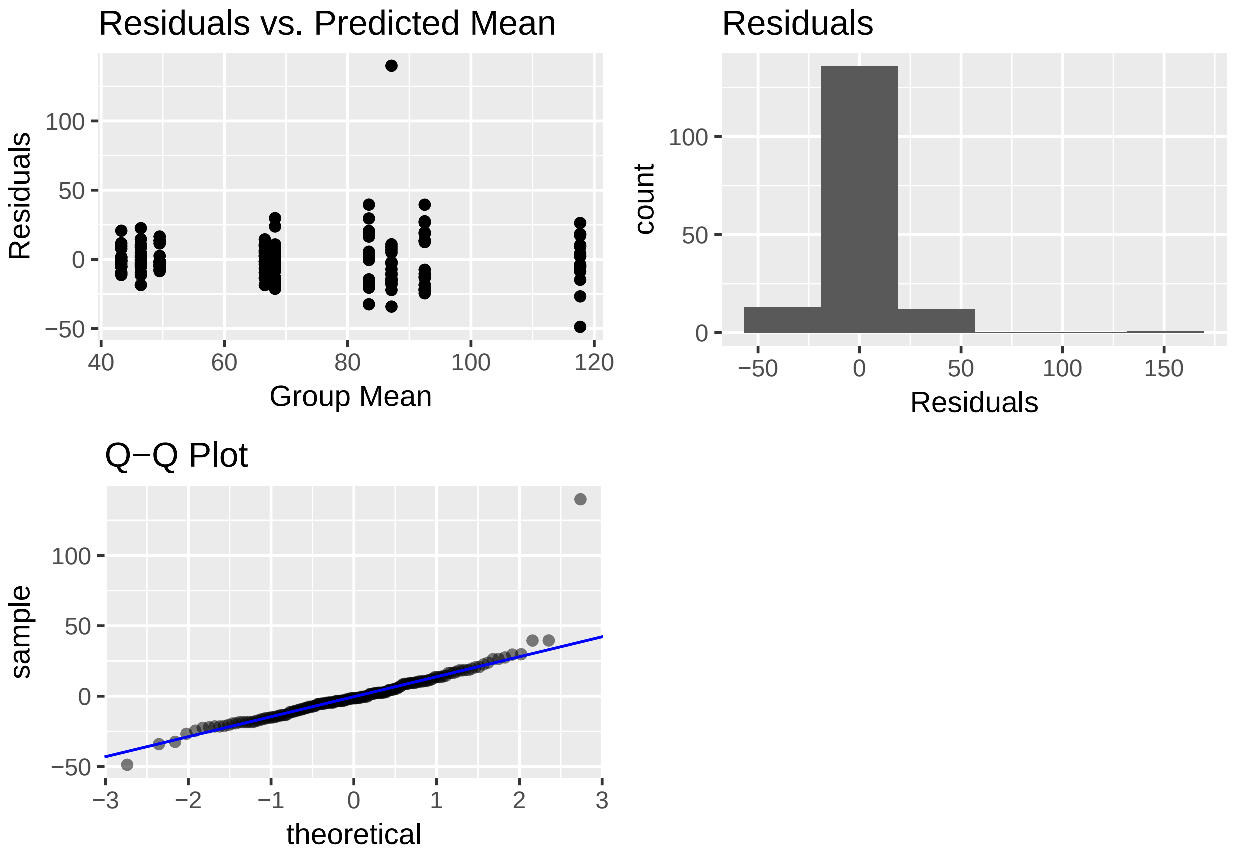
```
summer_burglaries <- crime %>%
                      filter(substr(CrimeDate, 1,1) %in% c('6','7','8'), Description == "BURGLARY") %>%
                      mutate(CrimeDate <- as.Date(CrimeDate, format = "%m/%d/%Y")) %>%
                      mutate(YearMon = as.yearmon(CrimeDate,  "%m/%d/%Y")) %>%
                      group_by(District, YearMon) %>%
                      summarize(Count = sum(`Total Incidents`)) %>%
                      filter(District != "NA") %>%
                      ungroup()

ggplot(data= summer_burglaries, aes(x = Count)) + geom_histogram() + facet_wrap(~ District)
```



Test Choice: Since there is a mess of distributions that appear to fit an ANOVA model, I'll use the ANOVA test to see if there's one mean that's different.

```
summer_burglaries <- summer_burglaries %>%
                            mutate(District = as.factor(District))


sum_burg_anova_check <- aov(Count ~ District, data = summer_burglaries)

y <- quantile(sum_burg_anova_check$residuals[!is.na(sum_burg_anova_check$residuals)], c(0.25, 0.75))
x <- qnorm(c(0.25, 0.75))
slope = diff(y)/diff(x)
int <- y[1L] - slope*x[1L]


norm_res <- ggplot(data = sum_burg_anova_check, aes(x = sum_burg_anova_check$residuals)) + geom_histogra


resid_jitter <- ggplot(data = sum_burg_anova_check, aes(x = sum_burg_anova_check$fitted.values, y = sum_
```

```r
qq <- ggplot(data = sum_burg_anova_check) + stat_qq(aes(sample = sum_burg_anova_check$residuals), alpha
    geom_abline(slope = slope, intercept = int, color = "blue") + ggtitle("Q-Q Plot")


res_sum <- as.table(summary(sum_burg_anova_check$residuals), nrow= 2, ncol = 6)
grid.arrange(resid_jitter, norm_res, qq, nrow = 2, ncol = 2)
```



Checking ANOVA/Test Statistic/p-value

```r
summary(sum_burg_anova_check)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)    
## District      8  88078   11010   32.54 <2e-16 ***
## Residuals   153  51772     338                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since our p-value is considerably less than 0.05, we can conclude that it appears that one mean is different than the other. We will proceed with Tukey's HSD to show which district is different from another.

```r
thsd_sum_burg <- TukeyHSD(sum_burg_anova_check)
thsd_sum_burg
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Count ~ District, data = summer_burglaries)
##
```

```
## $District
##                                 diff         lwr        upr     p adj
## EASTERN-CENTRAL            -3.055556 -22.3503538  16.239243 0.9998975
## NORTHEASTERN-CENTRAL       68.222222  48.9274240  87.517020 0.0000000
## NORTHERN-CENTRAL           43.000000  23.7052018  62.294798 0.0000000
## NORTHWESTERN-CENTRAL       18.722222  -0.5725760  38.017020 0.0647509
## SOUTHEASTERN-CENTRAL       33.944444  14.6496462  53.239243 0.0000046
## SOUTHERN-CENTRAL           37.611111  18.3163129  56.905909 0.0000003
## SOUTHWESTERN-CENTRAL       17.055556  -2.2392427  36.350354 0.1295154
## WESTERN-CENTRAL            -6.222222 -25.5170205  13.072576 0.9839710
## NORTHEASTERN-EASTERN       71.277778  51.9829795  90.572576 0.0000000
## NORTHERN-EASTERN           46.055556  26.7607573  65.350354 0.0000000
## NORTHWESTERN-EASTERN       21.777778   2.4829795  41.072576 0.0145920
## SOUTHEASTERN-EASTERN       37.000000  17.7052018  56.294798 0.0000004
## SOUTHERN-EASTERN           40.666667  21.3718684  59.961465 0.0000000
## SOUTHWESTERN-EASTERN       20.111111   0.8163129  39.405909 0.0340026
## WESTERN-EASTERN            -3.166667 -22.4614649  16.128132 0.9998657
## NORTHERN-NORTHEASTERN     -25.222222 -44.5170205  -5.927424 0.0020286
## NORTHWESTERN-NORTHEASTERN -49.500000 -68.7947982 -30.205202 0.0000000
## SOUTHEASTERN-NORTHEASTERN -34.277778 -53.5725760 -14.982980 0.0000036
## SOUTHERN-NORTHEASTERN     -30.611111 -49.9059094 -11.316313 0.0000556
## SOUTHWESTERN-NORTHEASTERN -51.166667 -70.4614649 -31.871868 0.0000000
## WESTERN-NORTHEASTERN      -74.444444 -93.7392427 -55.149646 0.0000000
## NORTHWESTERN-NORTHERN     -24.277778 -43.5725760  -4.982980 0.0035837
## SOUTHEASTERN-NORTHERN      -9.055556 -28.3503538  10.239243 0.8645865
## SOUTHERN-NORTHERN          -5.388889 -24.6836871  13.905909 0.9937698
## SOUTHWESTERN-NORTHERN     -25.944444 -45.2392427  -6.649646 0.0012955
## WESTERN-NORTHERN          -49.222222 -68.5170205 -29.927424 0.0000000
## SOUTHEASTERN-NORTHWESTERN  15.222222  -4.0725760  34.517020 0.2488070
## SOUTHERN-NORTHWESTERN      18.888889  -0.4059094  38.183687 0.0601195
## SOUTHWESTERN-NORTHWESTERN  -1.666667 -20.9614649  17.628132 0.9999991
## WESTERN-NORTHWESTERN      -24.944444 -44.2392427  -5.649646 0.0024032
## SOUTHERN-SOUTHEASTERN       3.666667 -15.6281316  22.961465 0.9995979
## SOUTHWESTERN-SOUTHEASTERN -16.888889 -36.1836871   2.405909 0.1381106
## WESTERN-SOUTHEASTERN      -40.166667 -59.4614649 -20.871868 0.0000000
## SOUTHWESTERN-SOUTHERN     -20.555556 -39.8503538  -1.260757 0.0273407
## WESTERN-SOUTHERN          -43.833333 -63.1281316 -24.538535 0.0000000
## WESTERN-SOUTHWESTERN      -23.277778 -42.5725760  -3.982980 0.0064007
```

- The Northeastern District has distinctly larger number of burglaries than all other groups.
- The Northwestern District has a distinctly larger number of burglaries than the Western, Northern, and Eastern.
- The Southern District has a distinctly larger number of burglaries in the Summer months over the Central, Eastern, and the Southwest.
- The Southwest District has a distinctly larger number of burglaries in the Summer months over the Eastern, Northern, and the Western districts.
- The Western District has a distinctly larger number of burglaries in the Summer months over the Northern and southern Districts.
- The Nothern District has a distincly larger number of burglaries in the Summer months over the Central and Eastern Districts.
- The Southeastern District has a distinctly larger number of burglaries in the Summer months over the

Central and Eastern Districts.