One Factor ANOVA

C. Perez

Analysis of Variance (ANOVA)

The ANOVA is a great way to determine if there are any statistically significant differences between the averages of three or more independent groups. In this case I want to determine if there are any significant differences between the average number of citibike trips across four months of the year, which correspond to the four different seasons in the year. The hypothesis test underlying the ANOVA is as follows:

Null Hypothesis (**Ho**): All sample averages are equal (**mu(1) = mu(2) = ... = mu(n)**).

Alternative Hypothesis (**Ha**): At least one average is statistically different from the rest.

The analysis is a test that relies on a comparison of variation. There are three types (arguably two) of variation in the one-way (one factor) ANOVA; Treatment Variation (between groups), Random Variation (within group), and Total Variation (which is the sum of the previous two). The between group variation is divided by the associated degrees of freedom (k-1) and the value becomes the numerator. The within group variation is divided by the associated degress of freedom k(n-1) and the value becomes the denominator. The final result is a test statistic which is then compared to a critical F-value and based on that comparison, the conclusion is to either reject or fail to reject Ho. (** k = the # of groups, while n = the sample size per group)

The easiest way to identify the result is to compare the p-value produced by the test to alpha (the significance level of the test). Should the p-value be smaller than alpha, Ho is rejected. The same applies to a majority of the p-values produced by R with the exception of a few tests, such as the Kolmogorov-Smirnov test. When this test is used to compare two samples, the higher the p-value, the more likely it is the samples come from the same distribution, as this value relates to the likelihood of finding an absolute difference when comparing cumulative distributions. However, this interpretation can get confusing when p-values such as (.54) are produced as p-values this high typically supports the null hypothesis, but for the ks.test() it simply tells us that half of our data is "alike".

The ANOVA assumptions are as follows:

1. The population for each sample should be normally distributed.
2. The variances for the populations are approx. equal
3. There should be minimal to no outliers (usually of greater importance in unbalanced designs)
4. Observations should be independent across and within groups.

It is typically very easy to accept these assumptions without verification depending upon the context. For example, if the ANOVA was being used to measure the effectiveness of 3 prep courses on students, we know something like test scores is normally distributed, and therefore picking a random sample of students is enough to satisfy 1,2, and 4; and in establishing the baseline by measuring the initial scores prior to the treatments, outliers would tend to reveal themselves. This would assist in verifying assumption 3.

The data used in this analysis are samples of the population of NYC Citibike riders, and because of this I assume that the aggregate day totals, grouped by month, is very likely to be normal with roughly equivalent variances across months (groups). However this will still be checked. The observations are also independent as each days aggregate totals do not relate to the previous day; the same applies for the groups (months).

Data

Citibike data used for this analysis is publicly available and can be found here.

Methodology

1. Download all individual 2019 monthly files for four months: January, April, July, and November.
2. Find the quantity of trips per month, per day, for the first 28 days. Seeing as the sample size is quite small, it is even more important that the assumptions are checked prior to conducting the analysis.
3. Conduct the ANOVA
    - Exploratory Data Analysis (EDA)
    - Assumption Checks
    - Outlier Comparison
    - Results and Reasonability

Analysis

I created the following code to handle steps 1 and 2. Notice that a function was created to eliminate what I feel to be unrealistic values in the data. Citibike policy states that trips exceeding a 24 hour period are subject to a fine of 1,200 dollars, and because of this it seems unlikely to include trips where the trip duration was greater than 24 hours as these are most likely incidents of loss, theft, or broken equipment that needs to be serviced. In fact, the data specifies when a bike is taken to what appears to be a repair station or Citibike facility and thus makes it easy to remove those observations.

The notes that accompany the data also state that trips less than 60 seconds in duration are excluded as this is usually the result of human error in trying to return or retrieve the equipment, which makes sense. The policy further states that each type of user (customer or subscriber) is alotted a maxium time per trip for which a charge of 15 cents per minute applies to the portion of time exceeding the alottment. This charge is equivalent to 9 dollars per hour. I find it very likely that someone would pay 9 dollars per hour for a bike rental in NYC, however I find it very unlikely that someone would pay this for 24 hours.

Because of the above, I chose to exclude trips that exceed 12 hours as the total cost for a 12 hour rental exceeds 100 dollars and becomes cost prohibitive as retail stores have bikes available with a purchase price starting at 100 dollars. Further, I chose to exclude all trips that are less than 5 minutes in length, where the pick-up station is the same as the drop-off station. It is likley that someone can cross an avenue in 5 minutes via citibike, but I doubt a roundtrip in that time is feasible. Also the original data inlcudes 61 second trips which, quite frankly, are not that different from 60 second trips (**already excluded**). I think 5 minutes with the second condition is a good way to isolate and remove small trips that are likley human error and/or are unrealistic.

```
#One-Way Analysis of Variance (ANOVA) - Balanced Design

#clear global environment
rm(list = ls(all.names = TRUE))

#Libraries----
library(tidyverse)
library(readxl)
library(openxlsx)
library(lubridate)
```

```r
library(car)
library(gridExtra)

#----Functions-----------------------------------------------


#created function to clean citibike input data and eliminate trips greater than 11 hours
#and less than 5 minutes.
citi_clean <- function(input_df){

  function_df <-  input_df %>%
    filter(!(tripduration < 300 & `start station id` == `end station id`)) %>%
    filter(tripduration < 43200 & (day(starttime) %in% c(1:28))) %>%
    mutate(day = day(starttime)) %>% mutate(month = month(starttime, label = TRUE))

  return(function_df)
}



#----import and clean----------------------------------------------------


#load citibike data for 2019 by month
january <- read_csv("~/Documents/R/RProjects-Public/ANOVA-Data/201901-citibike-tripdata.csv",
                    col_names = TRUE)
april <- read_csv("~/Documents/R/RProjects-Public/ANOVA-Data/201904-citibike-tripdata.csv",
                  col_names = TRUE)
july <- read_csv("~/Documents/R/RProjects-Public/ANOVA-Data/201907-citibike-tripdata.csv",
                 col_names = TRUE)
november <- read_csv("~/Documents/R/RProjects-Public/ANOVA-Data/201911-citibike-tripdata.csv",
                     col_names = TRUE)


#clean and retructure data with citi_clean() function using lapply()
months_raw <- list(january, april, july, november)

months_cleaned <- lapply(months_raw, citi_clean)



#----Main----------------------------------------


# prepare data for visual inspection and anova
january_vi <- months_cleaned[[1]] %>%
  select(day, month) %>% group_by(month,day) %>% summarise(trips = n()) %>% arrange(month)

april_vi <- months_cleaned[[2]] %>%
  select(day, month) %>% group_by(month,day) %>% summarise(trips = n()) %>% arrange(month)

july_vi <- months_cleaned[[3]]%>%
  select(day, month) %>% group_by(month,day) %>% summarise(trips = n()) %>% arrange(month)
```

```
november_vi <- months_cleaned[[4]] %>%
  select(day, month) %>% group_by(month,day) %>% summarise(trips = n()) %>% arrange(month)
```

The following code prepares the data for visual inspection (anova format, aggregated by month and day) and produces a histogram of each sample.

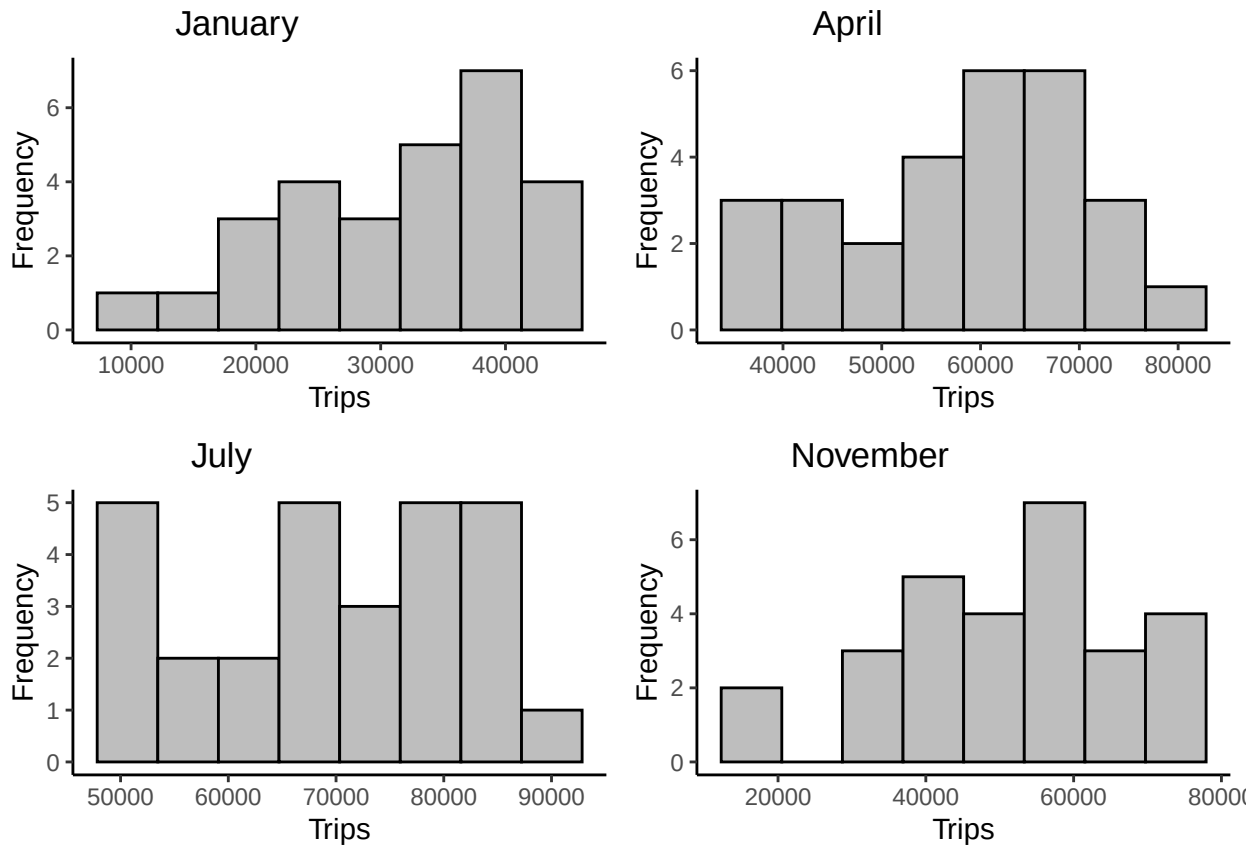
```
# graphics for assumption check via visual inspection
january_hist <- ggplot(data = january_vi, mapping = aes(x = trips))+
  geom_histogram(bins = 8, fill = "grey", color = "black")+
  theme_classic()+ labs(title = "January")+ ylab("Frequency")+ xlab("Trips")+
  theme(plot.title = element_text(hjust = .25))

april_hist <- ggplot(data = april_vi, mapping = aes(x = trips))+
  geom_histogram(bins = 8, fill = "grey", color = "black")+
  theme_classic()+ labs(title = "April")+ ylab("Frequency")+ xlab("Trips")+
  theme(plot.title = element_text(hjust = .25))

july_hist <- ggplot(data = july_vi, mapping = aes(x = trips))+
  geom_histogram(bins = 8, fill = "grey", color = "black")+
  theme_classic()+ labs(title = "July")+ ylab("Frequency")+ xlab("Trips")+
  theme(plot.title = element_text(hjust = .25))

november_hist <- ggplot(data = november_vi, mapping = aes(x = trips))+
  geom_histogram(bins = 8, fill = "grey", color = "black")+
  theme_classic()+ labs(title = "November")+ ylab("Frequency")+ xlab("Trips")+
  theme(plot.title = element_text(hjust = .25))
```

```
#arrange into a grid for easy viewing
grid.arrange(january_hist,  april_hist, july_hist, november_hist, ncol = 2)
```



The histograms show a normal like distribution for the most part, however there are some questions with the skewness of January and the high initial frequency in July. To verify the normality assumption, the Shapiro-Wilk test is implemented in the following code and a boolean test vector was created to store the result of testing whether the p-value is greater than an alpha of 5% (95% confidence level for the tests). The Shapiro-Wilk test essentially fits a gaussian curve to the data and measures both, the obervations that overlap with the curve and the observations that do not. The hypothesis underlying the test is as follows:

Ho: The data is a sample from some normal distribution.
Ha: The data is not a sample from some normal distribution.

As long as the p-values are greater than the level of significance it is safe to assume normality as this indicates a failure to reject the null hypothesis.

Levene's Test is also conducted in the following code to verify assumption #2 (roughly equivalent variances). The hypothesis underlying Levene's Test is as follows:

Ho: The sample variances are equal.
Ha: The sample variances are not equal.

Again, a p-value greater than the level of significance (alpha) shows a failure to reject the null hypothesis, which is what is needed to verify the assumption. This analysis uses alpha = .05 (5%) for each of the tests conducted.

```
#anova ready data
anova_data <- rbind(january_vi, april_vi, july_vi, november_vi) %>% group_by(month, day)
```

```
anova_data$month <- as.factor(anova_data$month)


#to apply shapiro-wilk
months_list <- list(january_vi$trips, april_vi$trips, july_vi$trips, november_vi$trips)


#Shapiro-Wilk normality tests - 1st assumption for ANOVA - population
sw_results <- lapply(months_list, shapiro.test)
sw_pvalues <- c(sw_results[[1]][2], sw_results[[2]][2], sw_results[[3]][2],
                sw_results[[4]][2])

  #results
unlist(sw_pvalues) > .05#if all true then fail to reject Ho
```

```
p.value p.value p.value p.value
   TRUE    TRUE    TRUE    TRUE
```

```
#Levene's Test for equality of variances across populations
#2nd assumption for ANOVA - population
leveneTest(trips ~ month, data = anova_data) #fails Levene's Test, confirms visuals
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   3   1.714 0.1684
      108
```

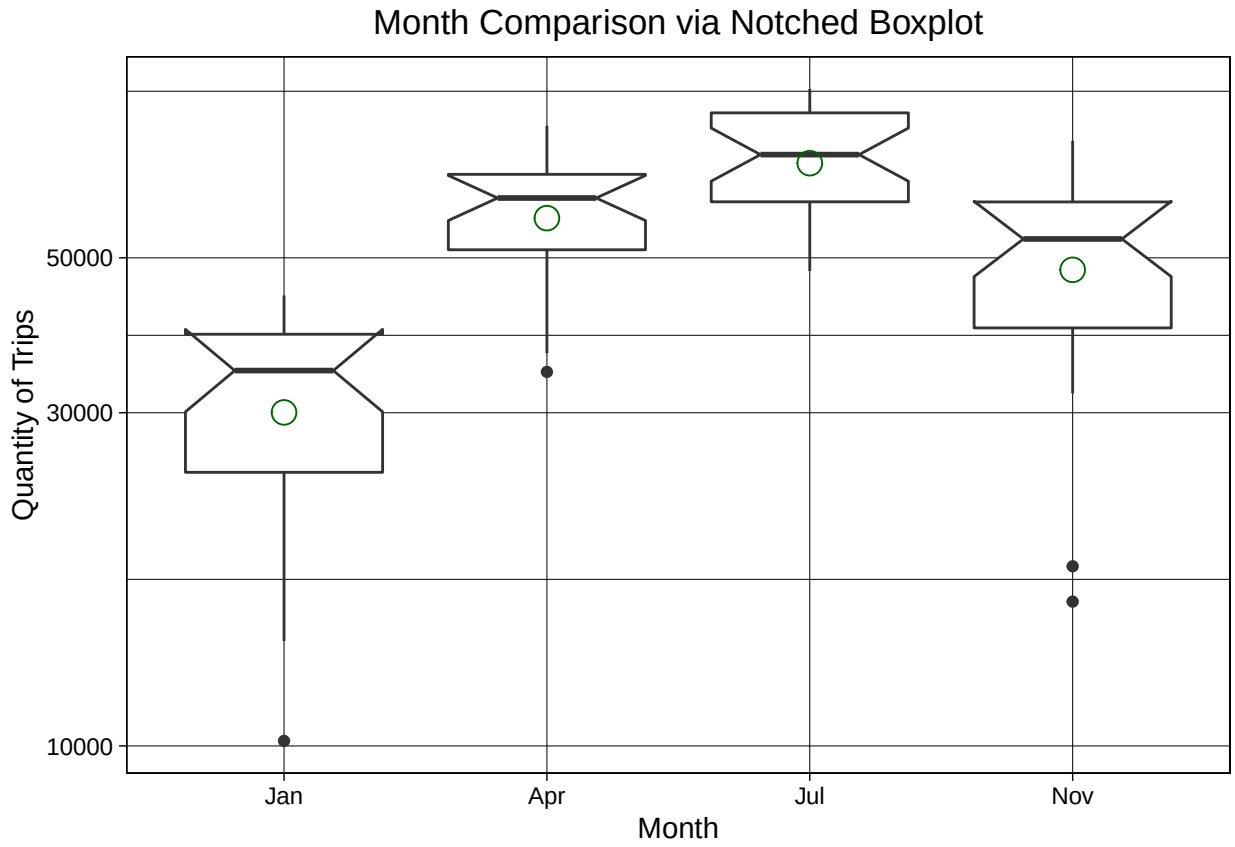
All of the assumptions required for an ANOVA analysis have been verified. The following code provides summary statistics for each month along with a notched boxplot for easy confidence interval identification. The green circles show the mean of each sample and it is clear to see the degree to which outliers affect the mean as it relates to the median. This analysis will be conducted again later after removing a few outliers to see the impact they have on the overall result.

```
#----ANOVA-Ready-Main-----------------------------------------------


#extract some summary statistics for each group (month)
sum_stats <- anova_data %>% group_by(month) %>%
  summarise(observations = n(), mean = mean(trips), st.dev = sd(trips))
sum_stats
```

```
# A tibble: 4 x 4
  month observations   mean st.dev
  <ord>        <int>  <dbl>  <dbl>
1 Jan             28 31743.  9414.
2 Apr             28 58280. 11759.
3 Jul             28 69450. 12452.
4 Nov             28 50853. 15104.
```

```r
#visualize the samples via boxplot using ggplot2 package
ggplot(anova_data, mapping = aes( x = month, y = trips))+
  geom_boxplot(notch = TRUE)+
  stat_summary(fun=mean, geom="point", shape=21, size=4, color = "darkgreen")+
  scale_y_log10()+
  ylab("Quantity of Trips")+
  xlab("Month")+
  labs(title = "Month Comparison via Notched Boxplot")+
  theme_linedraw()+
  theme(plot.title = element_text(hjust = .5))
```



The following F-test is the overall result from the ANOVA and it reveals an incredibly small p-value indicating the **Ho** (null hypothesis) should be rejected. This shows that at least one month has a statistically different average number of trips. The question now is, **which months have averages that are statistically different from each other?**

```r
#ANOVA analysis
output_anova <- aov(trips ~ month, data = anova_data)
summary.aov(output_anova)
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
month         3 2.112e+10 7.040e+09   46.15 <2e-16 ***
Residuals   108 1.647e+10 1.525e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The following test (**Tukey's Honest Significant Difference Method**) is a great way to determine which means are statistically different from each other without inflating the likelihood of a Type I error. The output shows an adjusted p-value which takes into consideration the increased error associated to family pair-wise comparisons. The results of this post hoc test indicate that there is a significant difference between each pair with the exception of **Nov-Apr**. Because the p-value is greater than 5% for this pair, it's a failure to reject the Ho which means there is not a statistical difference between the averages of these two months. The underlying hypothesis for each pair-wise test is as follows:

Ho: The means are the same.
Ha: The means are not the same.

```r
#Tukey method for comparison
TukeyHSD(output_anova)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = trips ~ month, data = anova_data)

$month
             diff        lwr       upr     p adj
Apr-Jan  26536.50  17923.332 35149.668 0.0000000
Jul-Jan  37707.14  29093.975 46320.311 0.0000000
Nov-Jan  19109.50  10496.332 27722.668 0.0000004
Jul-Apr  11170.64   2557.475 19783.811 0.0054222
Nov-Apr  -7427.00 -16040.168  1186.168 0.1165095
Nov-Jul -18597.64 -27210.811 -9984.475 0.0000008
```

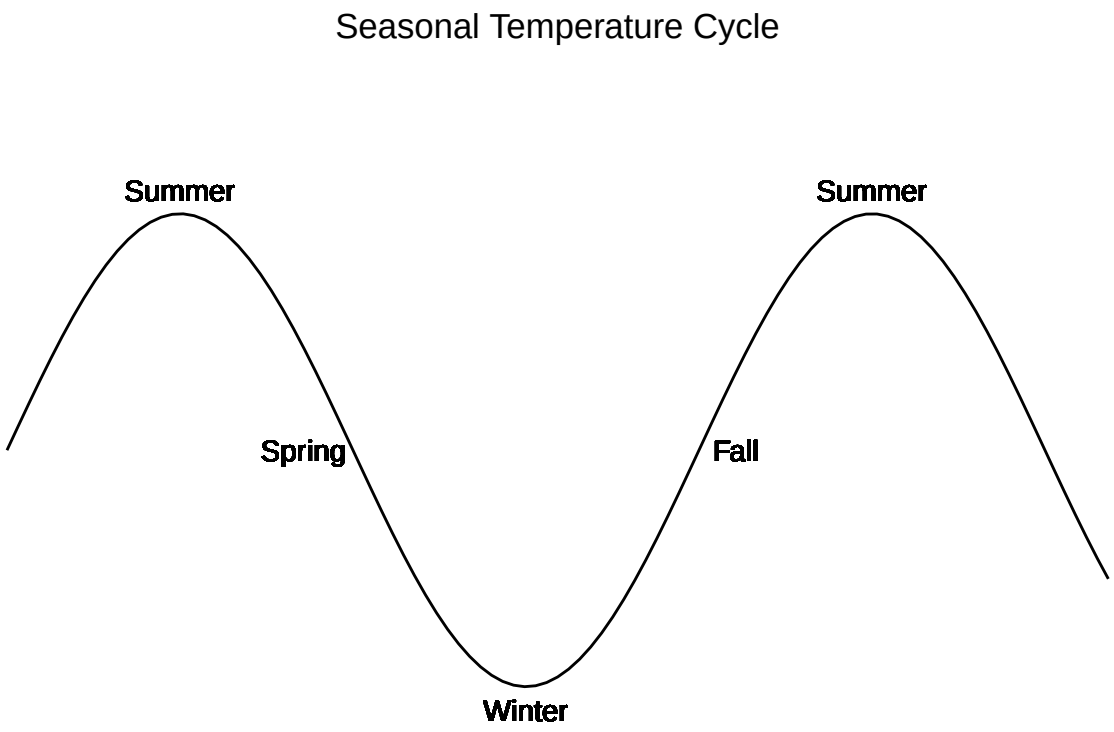
The above result corroborates an initial hunch based on the corresponding seasons. Typically, better weather results in more people riding bikes, which shows there is a correlation between weather patterns and the quantity of people riding bikes. The following graph represents my initial hunch on ridership as it relates to weather.

```r
# sinusoidal curve
x <-  seq(0,10,.1)
y <- sin(x)
sin_df <- as.data.frame(cbind(x,y))

sin_curve <- ggplot(data = sin_df, mapping = aes(x = x, y = y))+
  geom_line()+
  theme_classic()+
```

```
  geom_text( x = pi/2, y = 1.1, label = "Summer")+
  geom_text( x = 1.5*pi, y = -1.1, label = "Winter")+
  geom_text( x = pi/2+2*pi, y = 1.1, label = "Summer")+
  geom_text( x = 3*pi/3.5, y = 0, label = "Spring")+
  geom_text( x = 3*pi/3.5 + 1.25*pi, y = 0, label = "Fall")+
  ylim(c(-1.5,1.5))+
  theme(axis.text = element_blank(), axis.ticks = element_blank(),
        axis.title = element_blank(),
        axis.line = element_blank(), plot.title = element_text(hjust = .5))+
  ggtitle("Seasonal Temperature Cycle")

sin_curve
```



Because Winter and Summer represent the extremes of ridership, and Fall and Spring have similar temperature averages and probability of precipitation, it would make sense that November and April would have similar Citibike trip averages, and hence would be the only pair that do not show a statistical difference across averages.

The following code removes a few outliers from the data (while maintaining a balanced design) and shows the result of the test with the smaller samples. The overall analysis and result is the same, confirming the hint of assumption #3 and the original result of the test (ANOVA).

```
#----sample size reductions to remove outliers---
```

```r
# prepare data for visual inspection and anova
january_vir <- months_cleaned[[1]] %>%
  select(day, month) %>% group_by(month,day) %>% summarise(trips = n()) %>% arrange(trips)

april_vir <- months_cleaned[[2]] %>%
  select(day, month) %>% group_by(month,day) %>% summarise(trips = n()) %>% arrange(trips)

july_vir <- months_cleaned[[3]]%>%
  select(day, month) %>% group_by(month,day) %>% summarise(trips = n()) %>% arrange(trips)

november_vir <- months_cleaned[[4]] %>%
  select(day, month) %>% group_by(month,day) %>% summarise(trips = n()) %>% arrange(trips)


#anaova ready data
anova_data_r <- rbind(january_vir[2:28,], april_vir[2:28,], july_vir[2:28,],
                      november_vir[2:28,]) %>%
  group_by(month, day)
anova_data_r$month <- as.factor(anova_data_r$month)


#to apply shapiro-wilk
months_list_r <- list(january_vir$trips, april_vir$trips, july_vir$trips,
                      november_vir$trips)


#Shapiro-Wilk normality tests - 1st assumption for ANOVA - population
sw_results_r <- lapply(months_list_r, shapiro.test)
sw_pvalues_r <- c(sw_results[[1]][2], sw_results[[2]][2], sw_results[[3]][2],
                  sw_results[[4]][2])

#results
sw_pvalues_r > .05 #if all true then fail to reject Ho
```

```
p.value p.value p.value p.value
   TRUE    TRUE    TRUE    TRUE
```

```r
#Levene's Test for equality of variances across populations
#2nd assumption for ANOVA - population
leveneTest(trips ~ month, data = anova_data_r) #fails Levene's Test, confirms visuals
```

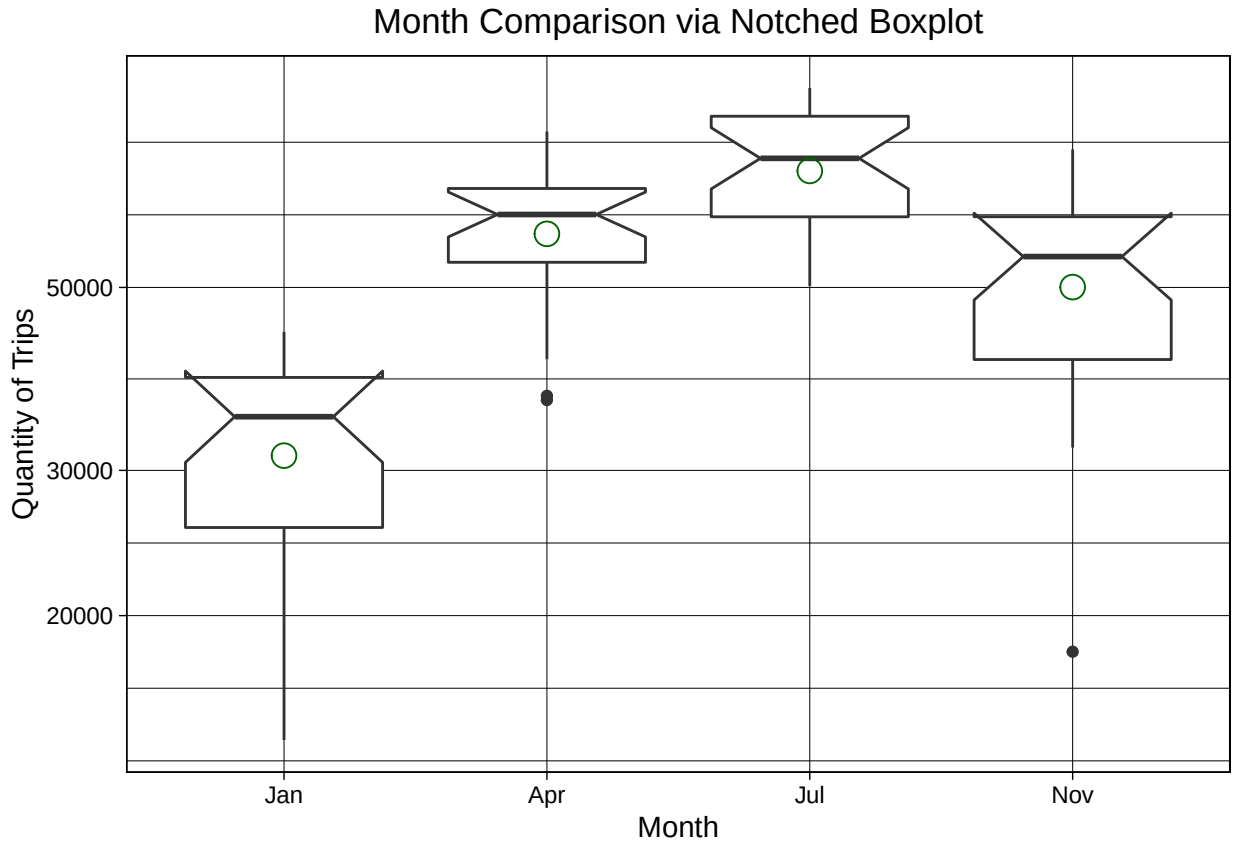
```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   3   1.734 0.1646
      104
```

```r
#----ANOVA-Ready-Main-----------------------------------------------


#extract some summary statistics for each group (month)
sum_stats <- anova_data %>% group_by(month) %>%
```

```r
  summarise(observations = n(), mean = mean(trips), st.dev = sd(trips))


#visualize the samples via boxplot using ggplot2 package
ggplot(anova_data_r, mapping = aes( x = month, y = trips))+
  geom_boxplot(notch = TRUE)+
  stat_summary(fun=mean, geom="point", shape=21, size=4, color = "darkgreen")+
  scale_y_log10()+
  ylab("Quantity of Trips")+
  xlab("Month")+
  labs(title = "Month Comparison via Notched Boxplot")+
  theme_linedraw()+
  theme(plot.title = element_text(hjust = .5))
```




```r
#ANOVA analysis
output_anova_r <- aov(trips ~ month, data = anova_data_r)
summary.aov(output_anova_r)
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
month         3 2.035e+10 6.784e+09   51.65 <2e-16 ***
Residuals   104 1.366e+10 1.313e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Tukey method for comparison
TukeyHSD(output_anova_r)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = trips ~ month, data = anova_data_r)

$month
              diff        lwr       upr     p adj
Apr-Jan  26624.519  18480.295 34768.742 0.0000000
Jul-Jan  37707.148  29562.924 45851.372 0.0000000
Nov-Jan  19597.704  11453.480 27741.927 0.0000000
Jul-Apr  11082.630   2938.406 19226.853 0.0031745
Nov-Apr  -7026.815 -15171.039  1117.409 0.1160755
Nov-Jul -18109.444 -26253.668 -9965.221 0.0000004
```