Group Module 1

Duc Le, Evan Yuan, Yassaman Davilu

9/30/2020

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

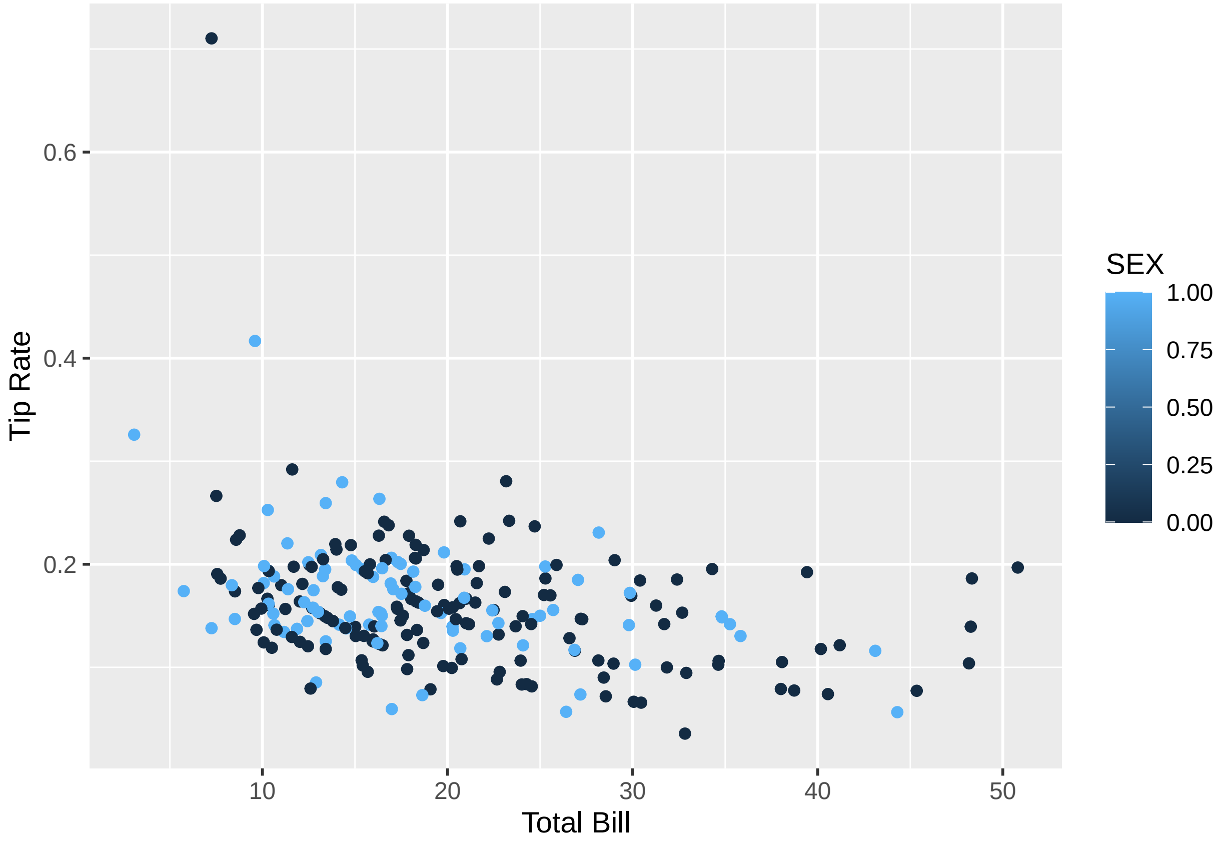
```
library(ggplot2)

tips = read.csv("tips.csv")
attach(tips)
```

Question 1

Examine the association between the total bill and tip. How does this relationship change based on sex, smoker status, and the combination of the two (i.e. male smokers, female non-smokers, etc.)? We want to analyze the effect the variable "sex" has on the correlation between Total Bill vs. Tip Rate (Tip Rate = Tips/Total Bill)

```
tip.rate = TIP/TOTBILL
qplot(TOTBILL, tip.rate, xlab = "Total Bill", ylab = "Tip Rate", col = SEX)
```



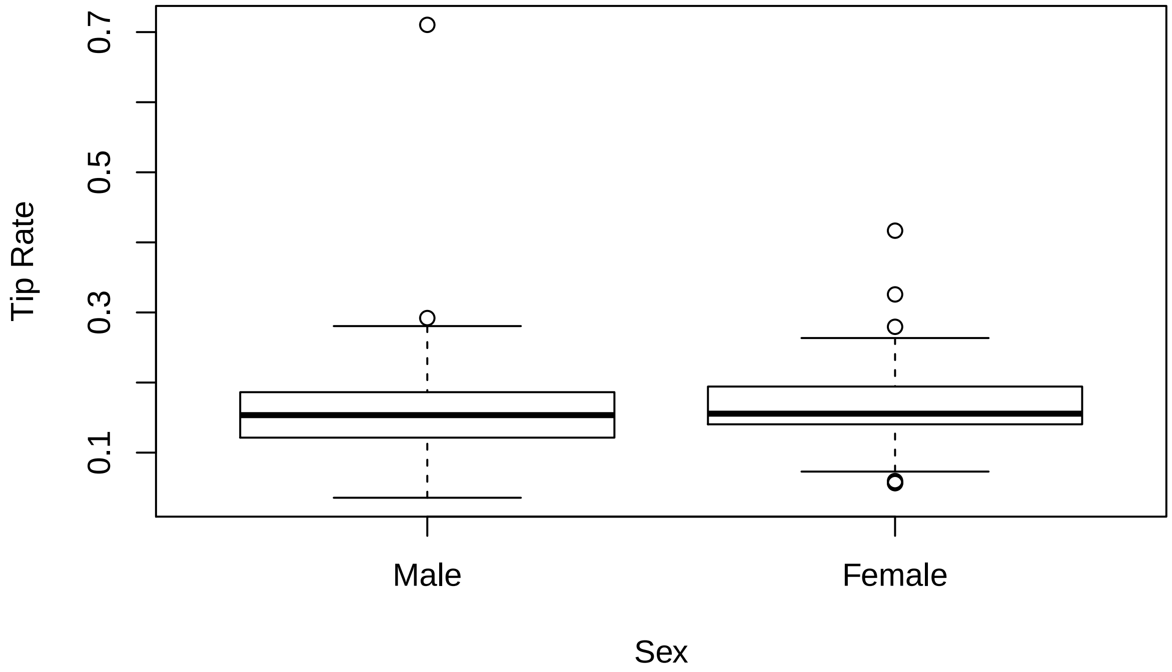
As far as we can see, there doesn't seem to be any disparity determined by sex when it comes to analyzing Total Bill vs. Tip Rate. Next, we ran an ANOVA test to see if there's any significant correlation between Tip Rate & Sex.

```
anova(lm(tip.rate~SEX, data = tips))
```

```
## Analysis of Variance Table
##
## Response: tip.rate
##            Df  Sum Sq   Mean Sq F value Pr(>F)
## SEX         1 0.00437 0.0043747  1.1737 0.2797
## Residuals 242 0.90197 0.0037271
```

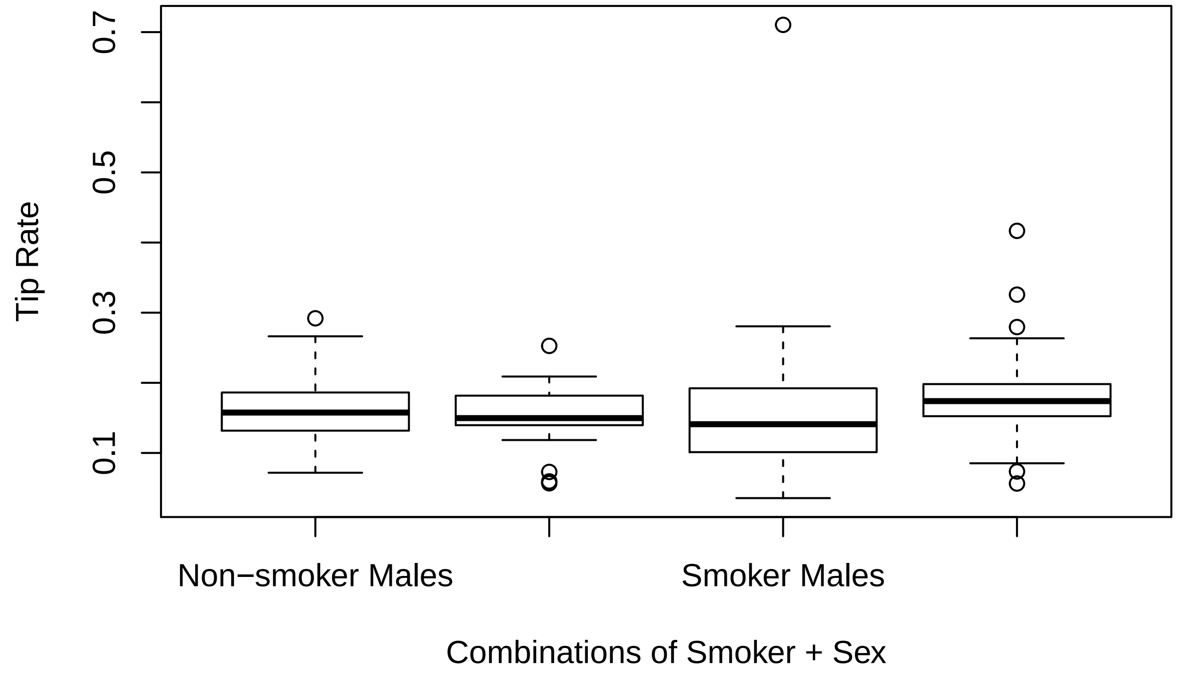
An ANOVA statistic test provide the results of insignificant F-statistics. These results show that "SEX" & "SMOKER" (status) don't really have an effect on the Tip Rate. Next, we construct a boxplot to see if there's any significant difference of variance between the 2 sexes vs. Tip Rate.

```
boxplot(tip.rate~SEX, xlab = "Sex", ylab = "Tip Rate",
        names = c("Male", "Female"))
```



The plot shows a slightly higher spread in the Male Tip Rate vs. Female but the medians are the same. Then, we analyze the relationship between the Tip Rate vs. different combinations of Sex & Smoker status (Male Smoker, Female Non-smoker, etc) with a boxplot.

```
smoker.sex = paste(SMOKER, SEX)
boxplot(tip.rate~smoker.sex, xlab = "Combinations of Smoker + Sex",
        ylab = "Tip Rate",
        names = c("Non-smoker Males", "Non-smoker Females",
                  "Smoker Males", "Smoker Females"))
```



Slight variances between the groups but no significant differences in terms of the variances + medians.

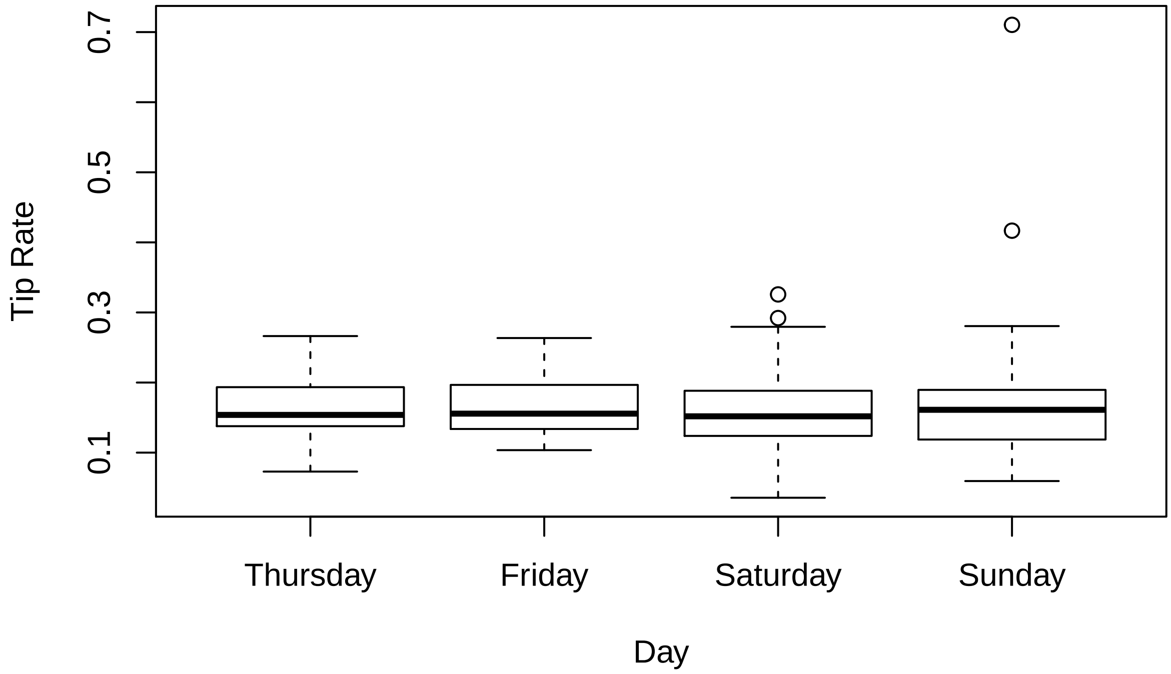
Lastly, we ran ANOVA tests for Tip Rate vs. Sex + Smoker

```
anova(lm(tip.rate~SEX+SMOKER, data = tips))
```

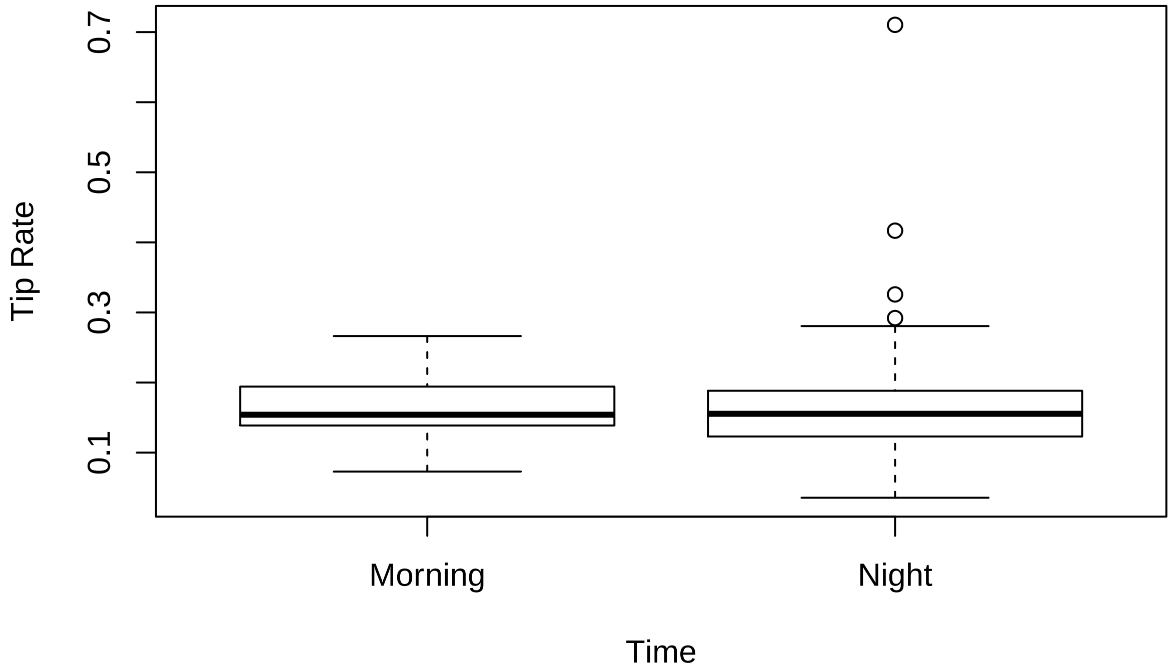
```
## Analysis of Variance Table
##
## Response: tip.rate
##            Df  Sum Sq   Mean Sq F value Pr(>F)
## SEX         1 0.00437 0.0043747  1.1700 0.2805
## SMOKER      1 0.00087 0.0008719  0.2332 0.6296
## Residuals 241 0.90110 0.0037390
```

Once again, the F-Statistics are too small for us to admit to any significance variables like Smoker Status or Sex might have on the Tip Rate. #### Conclusion for Q1 After various statistical tests & visualizations, we don't find any significant correlation/effect variables such as Sex + Smoker have on the Tip Rate (or Total Bill vs. Tip). We think the data set is limited and biased since these records were saved by one person. ### Question 2 #### At what time & day is the tip rate the highest/lowest? First, we use boxplots to see the relationship between Tip Rate vs. Day & Tip Rate vs. Time.

```
boxplot(tip.rate~DAY, xlab = "Day", ylab = "Tip Rate",
        names = c("Thursday", "Friday", "Saturday", "Sunday"))
```

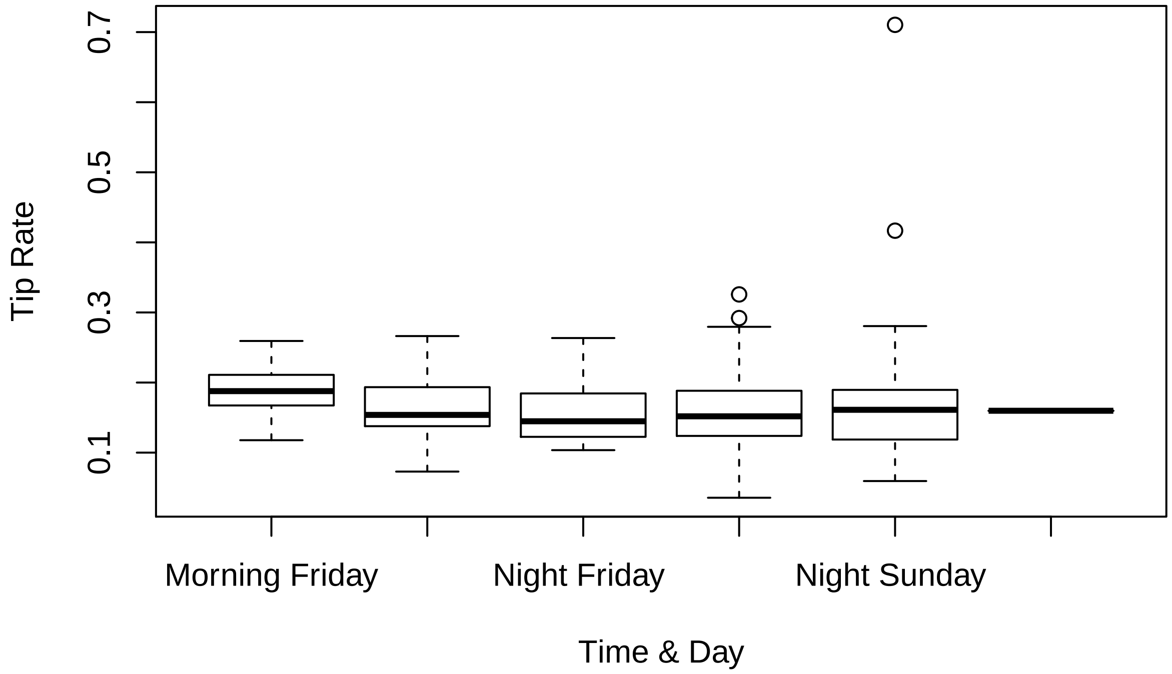


```r
boxplot(tip.rate~TIME, xlab = "Time", ylab = "Tip Rate",
        names = c("Morning", "Night"))
```



Outside of notable outliers, no significant differences can be drawn from the box plots. Now we look at the relationship between the Tip Rate vs. the combinations of the Days & Times.

```
time.day = paste(TIME, DAY)
time.day2 = mapvalues(time.day, c("0 3", "0 4", "1 3", "1 4", "1 5", "1 6"),
          c("Morning Thursday", "Morning Friday", "Night Thursday",
            "Night Friday ", "Night Saturday ", "Night Sunday "))
boxplot(tip.rate~time.day2, xlab = "Time & Day", ylab = "Tip Rate")
```



Few takeaways from the boxplot:

1. There are more night observations than day. The waitress's schedule isn't consistent as she obviously has recorded more night shifts than day shifts.
2. Although the variances aren't exactly identical between each variable, there isn't a big disparity between their medians.
3. We cannot do much with only 1 table recorded for Thursday Night. ANOVA test

```
anova(lm(tip.rate~TIME+DAY, data = tips))
```

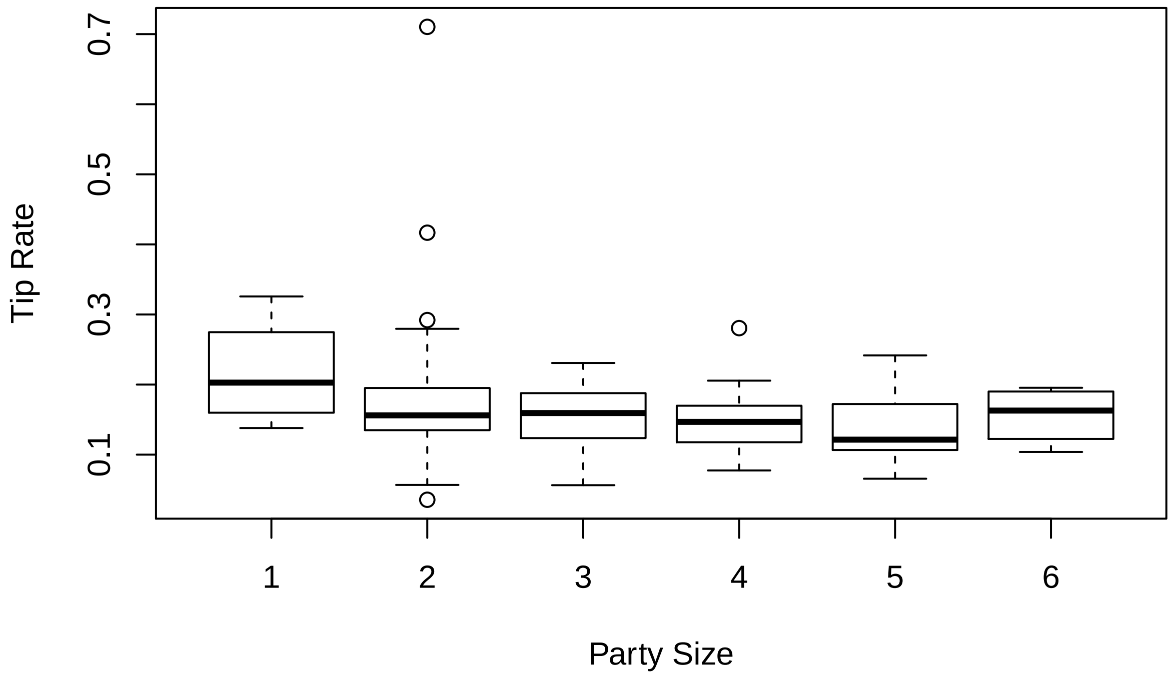
```
## Analysis of Variance Table
##
## Response: tip.rate
##            Df  Sum Sq   Mean Sq F value Pr(>F)
## TIME        1 0.00104 0.0010425  0.2797 0.5974
## DAY         1 0.00717 0.0071652  1.9227 0.1668
## Residuals 241 0.89814 0.0037267
```

The results from the ANOVA test suggest we cannot reject the null hypothesis (P-values > 0.05).

Conclusion for Q2 The given data recorded by the waitress is not sufficient enough for us to find any conclusive evidence that may suggest serving tables at day time results in a better tip rate than vs. night time.

Bonus Analysis We want to lastly investigate to see if the tip rate is at all affected by the party size.

```
boxplot(tip.rate ~ tips$SIZE, data = tips,xlab = "Party Size", ylab ="Tip Rate")
```



The plot interestingly shows that the median tip rate from parties of 1 is higher than the rest. Parties of 2 have the highest variance in contrast to parties of 6, which have the lowest range for tip rate.

```
anova(lm(tip.rate~SIZE, data = tips))
```

```
## Analysis of Variance Table
##
## Response: tip.rate
##            Df  Sum Sq   Mean Sq F value  Pr(>F)
## SIZE        1 0.01850 0.0184975  5.0418 0.02565 *
## Residuals 242 0.88785 0.0036688
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, an ANOVA test that produces a p-value < 0.05, which means we reject the null hypothesis. The means of the different groups are not equal. #### Conclusion We think the reason behind a higher usual tip rate (~20%) from a single party is due to a probably a smaller overall bill, therefore the customer may be more comfortable tipping a slightly higher percentage. In comparison to the tip rate from party size of 6, their bills will mostly be larger therefore the range of tips should be smaller.