Background

A top performing barista at Glen Edith Coffee shop located in Rochester, NY decided to keep track of all of his tips acquired for the span of three Thursday-Sunday shifts. He recorded a total of 244 descriptive observations regarding each customer's total bill, tip amount, gender, week day, party size, and whether it was during a lunch or dinner shift. You are considering opening a similar type coffee shop in the Rochester, NY area and want to better understand the local customer base, specifically, factors that lead to higher or lower tips.

Tip Amount vs. Total Bill

In Figure 1 below, it is clear that there is a positive correlation with tip amount and total bill. In this case, we can rightfully assume that the higher the bill, the higher the tip. There are a few outliers, where one customer tipped $5 on a $7 total bill, and another that tipped $1 on a $33 total bill. Without knowing more detail regarding the customer's individual experience during their visit to Glen Edith, we cannot really form a more specific opinion about the Rochester coffee shop customer with this information.

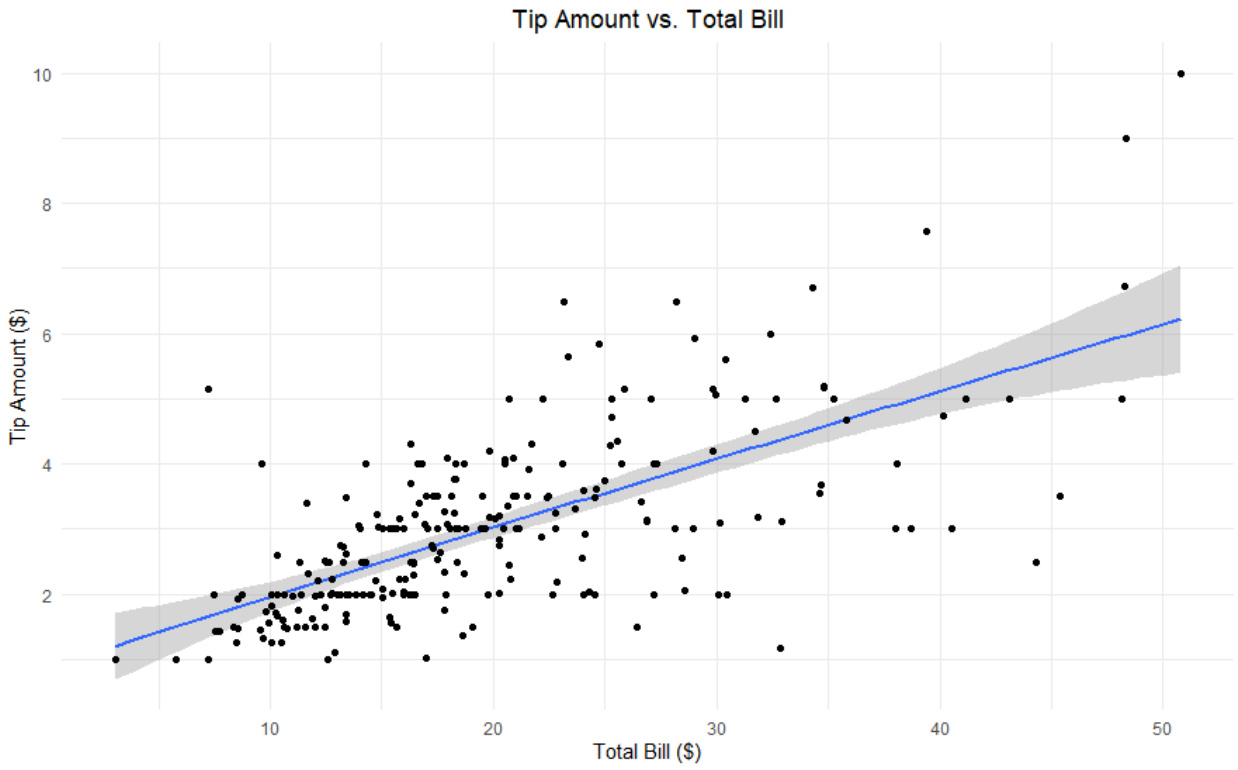

Figure 1

Tip Percentage of Total Bill Frequency

We can take the total bill verses tip amount a step farther by analyzing the percentage of total bill that resulted in the tip amount. The highest number of customer's tipped around 14-20% of their total bill amount, shown in Figure 2. The plot is skewed right, which means that only a smaller number of customers tipped higher than the average.

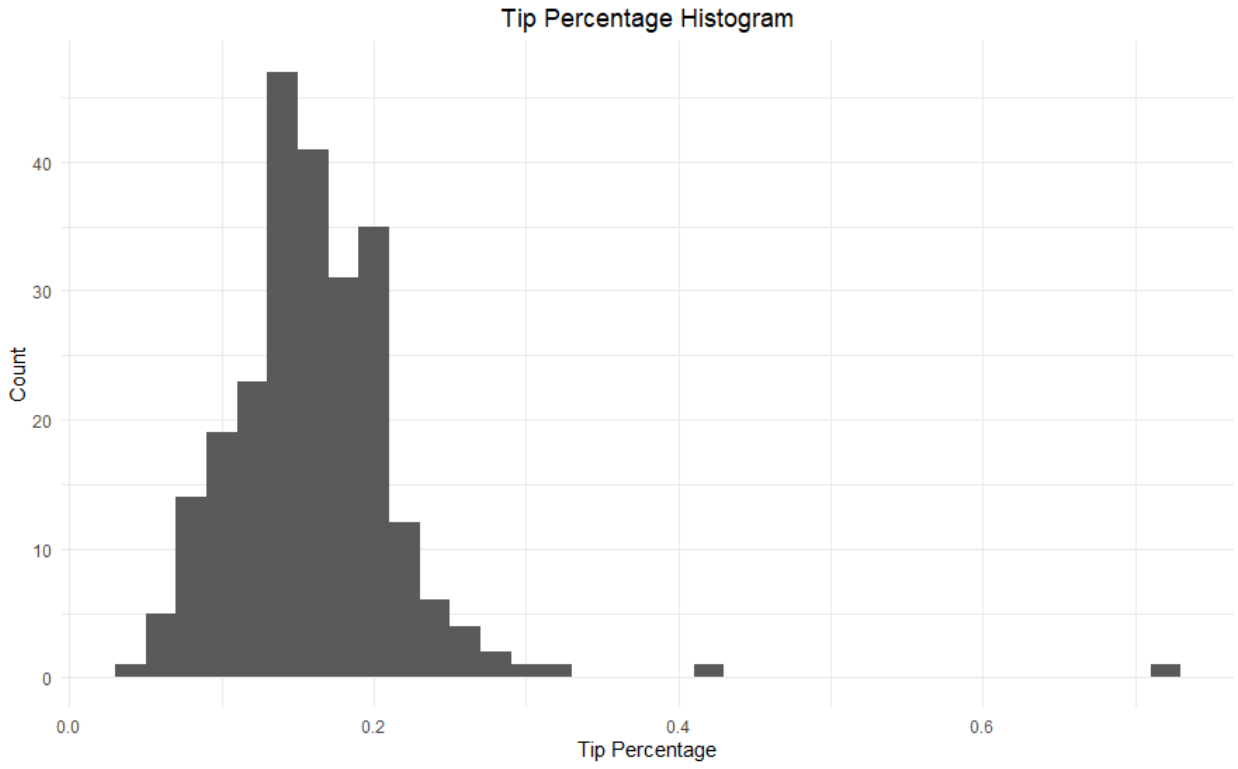

Figure 2

Tip Percentage of Total Bill by Day of Week

Figure 3 below separates out the tip percentage of the total bill by each day of the week. This graph gives us a better idea of any differences that we may see day by day. Firstly, we see very quickly that our barista received much fewer tips on Friday than Thursday, Saturday, or Sunday. Surprisingly, the Saturday and Thursday plots are very similar, but Saturday appears to have a higher variance in tip percentage, while Thursday is very closely centered around the average percentage. Sunday tips appear to have the highest variance, and is skewed very right, meaning that there are a few customers that tipped a much higher percentage than the average. In Figure 4 below, the differences in percentage of total bill by day of the week don't appear as significant, but this scatter plot allows us to see exactly where each individual tip percentage lands among the rest.

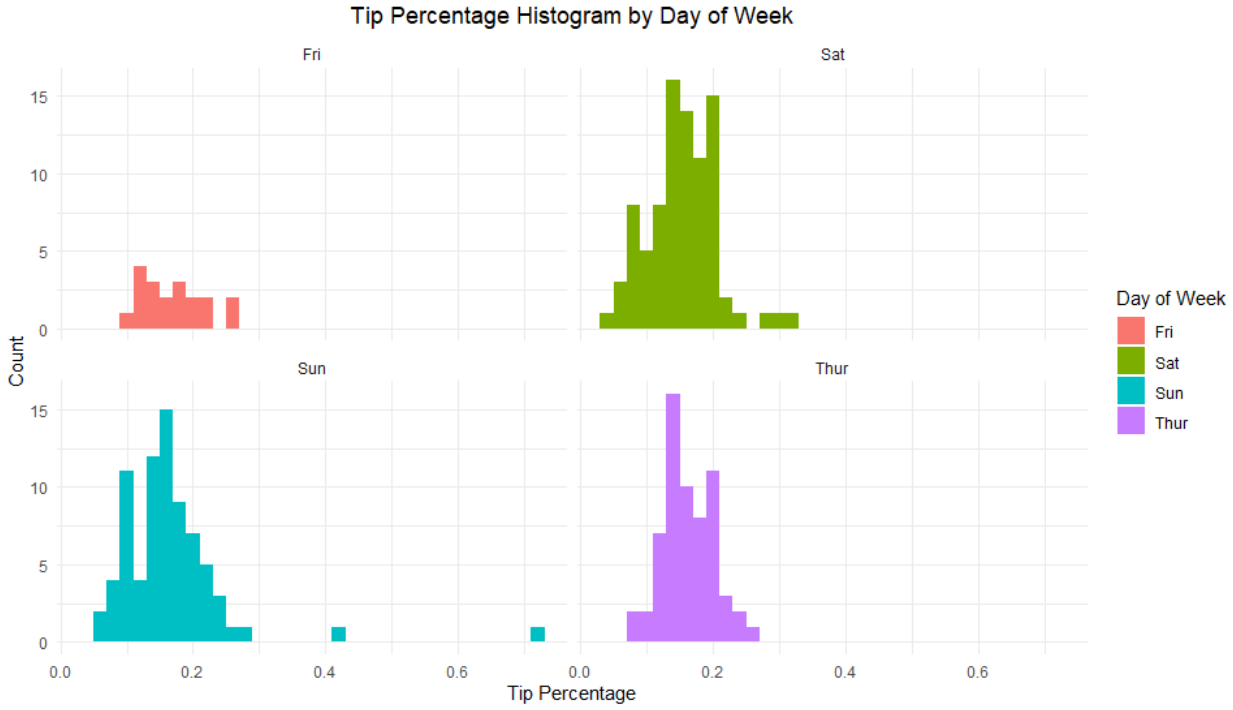

Figure 3

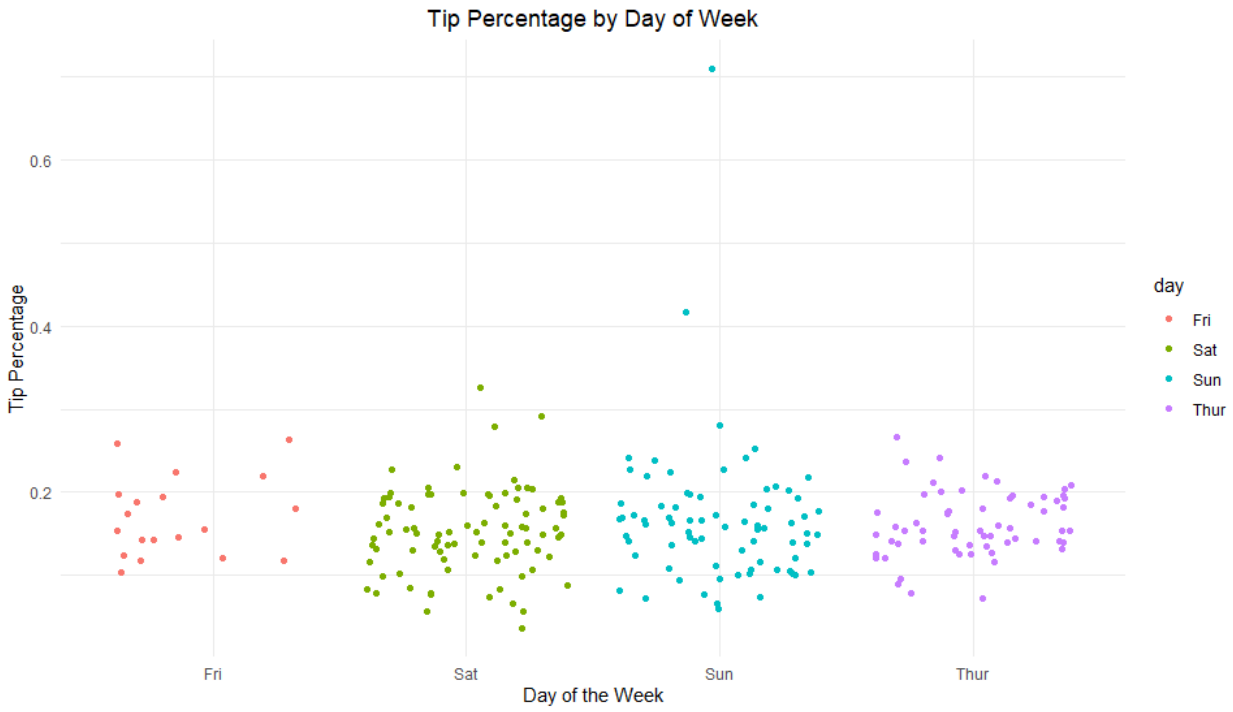
Tip Percentage by Day of Week
Tip Percentage
0.6
0.4
0.2
Fri
Sat
Sun
Thur
Day of the Week
day
Fri
Sat
Sun
Thur

Figure 4

Tip Percentage by Gender

When we break out the tip percentage of total bill by gender and display as a box plot, we are able to see any slight differences between the two groups. The female red box being slightly higher than the blue indicates that some females did tip slightly higher than males. The thick horizontal median line is essentially equal around 15% for both gender groups, meaning that the median tip percentage across male and females is basically the same. The main difference is that males tip percentage varies significantly, while females typically cluster around the average. The location of the median line for each group will also indicate if the tip percentage is skewed or not, if the horizonal median line is not exactly centered around the correspondingly colored box. It appears that females who tipped our barista were right skewed, meaning that a few females tipped higher than average, while males are very centered around the average (see Figure 6 for more detail regarding skew). Finally, the dots much higher or lower than the colored boxes represent specific outliers, or people who tipped far outside the groups average. Females also have a couple more outliers than males did in this data set.

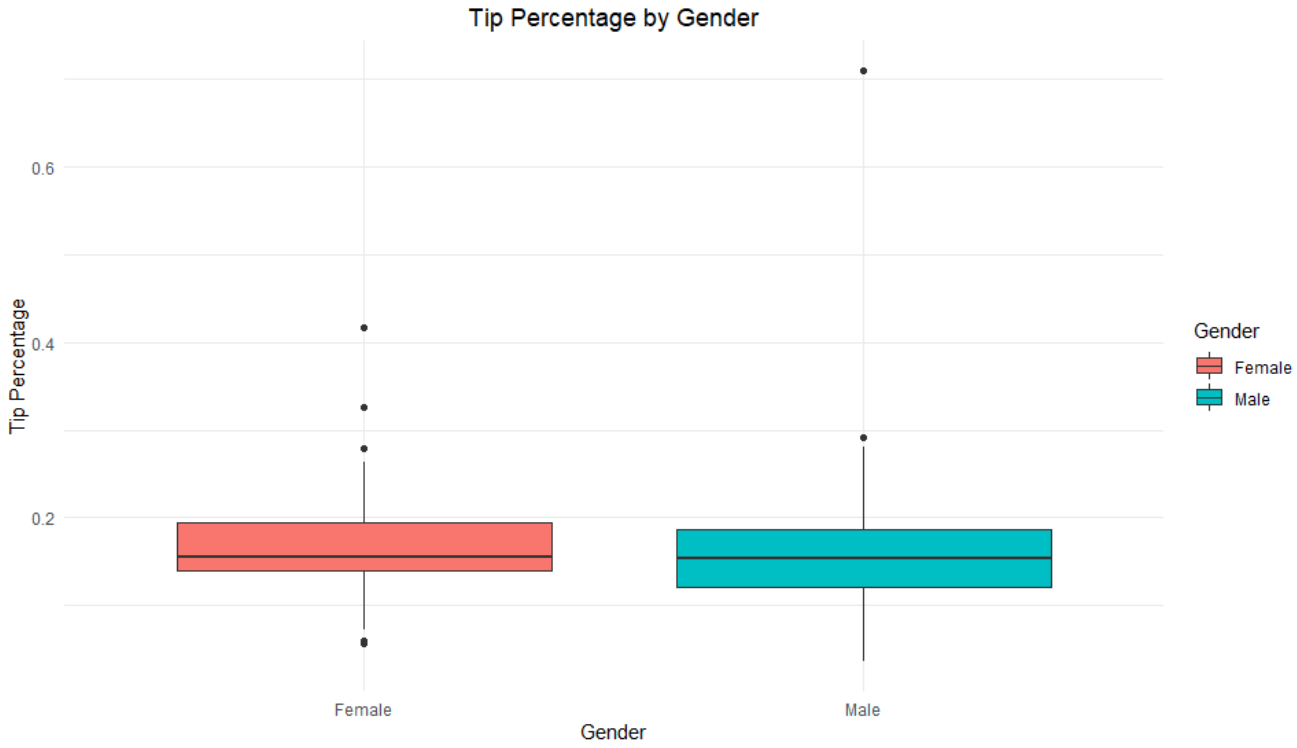

Figure 5

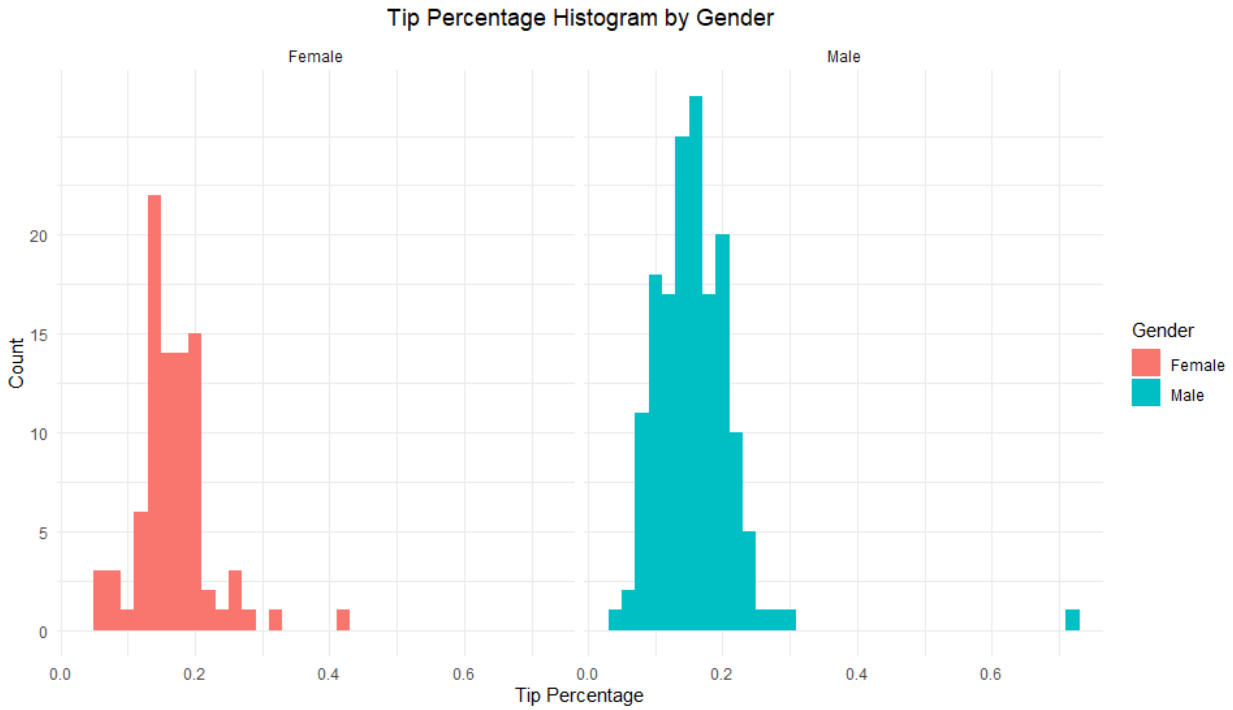
Tip Percentage Histogram by Gender
Female
Male
Count
0
5
10
15
20
0.0
0.2
0.4
0.6
Tip Percentage
Gender
Female
Male

Figure 6

Tip Percentage by Party Size

Figure 7 below displays tip percentage of total bill by number of people in the party. Most of the customers tipping our barista arrive in a party size of 2. They also appear to cluster around a higher tip percentage than other party sizes. Because the sampling does not appear to be random party sizes, it is difficult to fully understand differences among the groups. Although, we understand that incomplete data might mislead us, Figure 8 adds a trend line that resembles our assumption. It appears that party sizes of 1 tip the highest percentage of their total bill, while party sizes of 3 or 4 tip the lowest percentage of their bill.

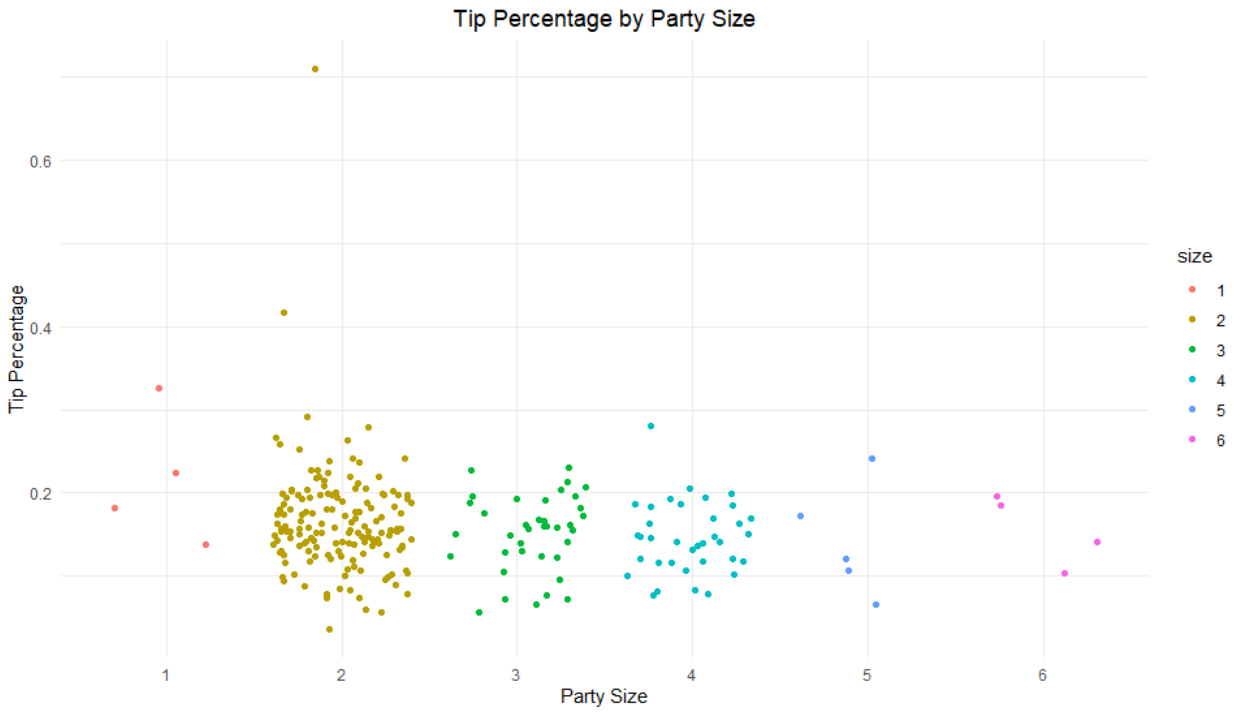

Figure 7

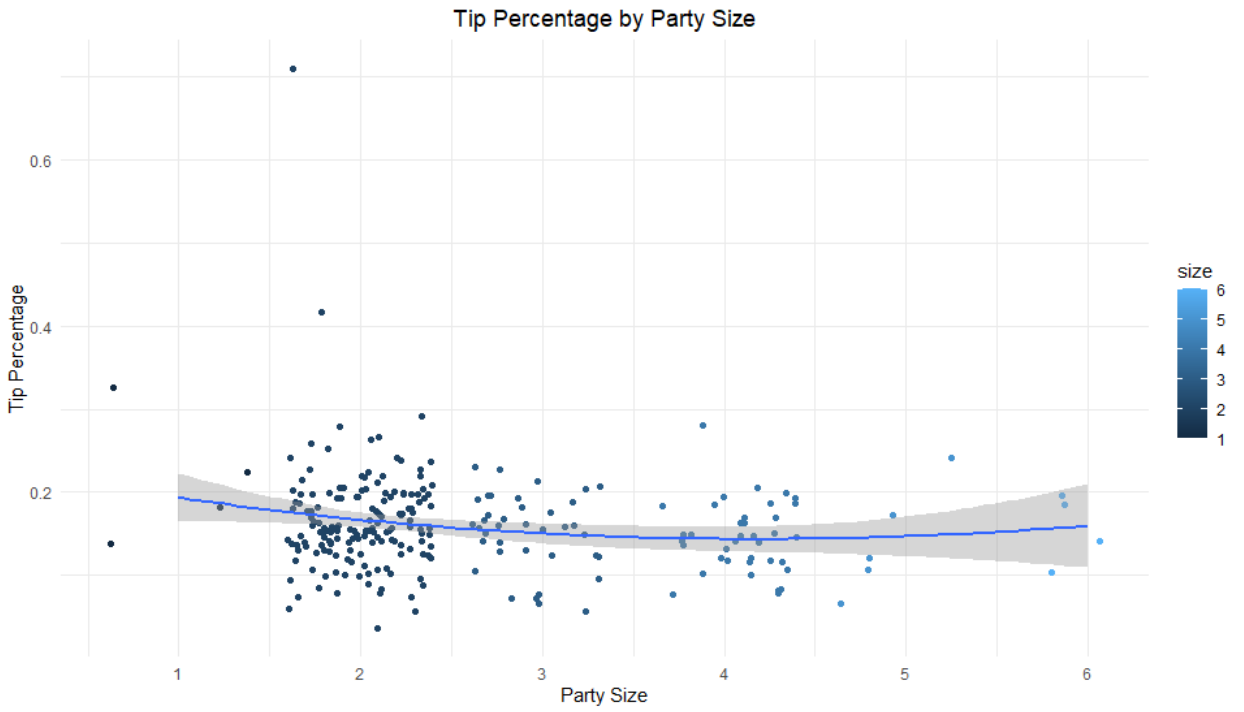

Figure 8

Summary

Not surprisingly, we discovered very quickly that the higher the total bill, the higher you can expect to be tipped. After calculating tip percentage of total bill, we were able to normalize the data so that total bill amount did not skew the remaining analysis. Glen Edith's customers mostly tipped around the 15% mark, with a small amount of people tipping over that benchmark, verses patrons who tipped under the average. By expanding our analysis to days of the week, we were not really able to discover many insights. Just that Sunday patrons were more varied and random in their tip percentage than Saturday or Thursday, and Thursday customers seemed to cluster more around the average than Saturday customers. After grouping our data by gender, we noticed that both sexes did tend to tip around the same average. Females tended to be right skewed, meaning that a small number tipped slightly higher than the average, but not enough to say they did indeed tip higher than males. Finally, we split up our data into groups by party size and discovered that our data tended to be couples. When analyzing party size by tip percentage, there was a surprising decrease in tip percentage clustering with groups of 3 or 4. Given this sample, it would be interesting to apply predictive analytics to see if based on a customer's characteristics, are they more or less likely to tip a higher percentage of their total bill. Through this analysis we did discover a significant error with sampling because the data does not appear to be random and the sample size is quite small. If we decide to proceed with forming business decisions, it would be a good idea to partner with Glen Edith and see if they are willing to provide additional data.

Appendix: R Code for Visualizations

```
##link to data set: https://www.kaggle.com/jsphyg/tipping


#install.packages("pacman")
pacman::p_load(tidyverse, ggplot2)


tips <- read.csv(file.choose())


str(tips)
summary(tips$tip)



#tip vs total scatter (FIGURE 1)
ggplot(data=tips, mapping=aes(x=total_bill, y=tip))+
  geom_smooth(span=10) +
  geom_point()+
  theme_minimal()+
  theme(plot.title=element_text(hjust=0.5))+
  ggtitle("Tip Amount vs. Total Bill")+
  xlab("Total Bill ($)") +
  ylab("Tip Amount ($)") +
  scale_y_continuous(breaks=seq(0,12,2))



#tip percentage analysis (FIGURE 2)
tips$percentage <- tips$tip / tips$total_bill
ggplot(data=tips)+
  geom_histogram(mapping=aes(x=percentage), binwidth = .02)+
  theme_minimal()+
  theme(plot.title=element_text(hjust=0.5))+
  ggtitle("Tip Percentage Histogram")+
```

```
  xlab("Tip Percentage")+
  ylab("Count")+
  scale_y_continuous(breaks=seq(0,50,10))

#frequency plot with days of the week like sum or count (FIGURE 3)
#see which days got the most tips
#can also break down by lunch vs. dinner, but missing thursday dinner data
ggplot(data=tips)+
  geom_histogram(mapping=aes(x=percentage, fill=day), binwidth = .02)+
  theme_minimal()+
  theme(plot.title=element_text(hjust=0.5))+
  ggtitle("Tip Percentage Histogram by Day of Week")+
  xlab("Tip Percentage")+
  ylab("Count")+
  scale_fill_discrete(name="Day of Week")+
  scale_y_continuous(breaks=seq(0,20,5))+
  facet_wrap(~day)

#(FIGURE 4)
ggplot(tips) +
  geom_point(aes(x=day,y=percentage, color=day), position="jitter")+
  theme_minimal()+
  theme(plot.title=element_text(hjust=0.5))+
  ggtitle("Tip Percentage by Day of Week")+
  xlab("Day of the Week")+
  ylab("Tip Percentage")

#Parse out by gender now (FIGURE 5)
#boxplot tip percentage by gender
ggplot(data=tips)+
  geom_boxplot(mapping=aes(x=sex, y=percentage, fill=sex))+
```

```
  theme_minimal()+
  theme(plot.title=element_text(hjust=0.5))+
  ggtitle("Tip Percentage by Gender")+
  xlab("Gender")+
  ylab("Tip Percentage")+
  scale_fill_discrete(name="Gender")

#histogram tip percentage by gender (FIGURE 6)
#boxplot tip percentage by gender
ggplot(data=tips)+
  geom_histogram(mapping=aes(x=percentage, fill=sex), binwidth = .02)+
  theme_minimal()+
  theme(plot.title=element_text(hjust=0.5))+
  ggtitle("Tip Percentage Histogram by Gender")+
  xlab("Tip Percentage")+
  ylab("Count")+
  scale_fill_discrete(name="Gender")+
  scale_y_continuous(breaks=seq(0,20,5))+
  facet_wrap(~sex)

#do something with party size vs tip/percentage (FIGURE 8)
ggplot(tips) +
  geom_point(mapping=aes(x=size,y=percentage, color=size), position="jitter")+
  geom_smooth(mapping=aes(x=size,y=percentage),span=10)+
  theme_minimal()+
  theme(plot.title=element_text(hjust=0.5))+
  ggtitle("Tip Percentage by Party Size")+
  xlab("Party Size")+
  ylab("Tip Percentage")+
  scale_x_continuous(breaks=seq(1,6,1))
```

```
#make size a factor for the pretty graph (FIGURE 7)
tips$size <- as.factor(tips$size)
ggplot(tips) +
  geom_point(mapping=aes(x=size,y=percentage, color=size), position="jitter")+
  theme_minimal()+
  theme(plot.title=element_text(hjust=0.5))+
  ggtitle("Tip Percentage by Party Size")+
  xlab("Party Size")+
  ylab("Tip Percentage")+
  scale_fill_discrete(name="Party Size")

##ideas to possibly expand farther but didn't make the cut
##grouping by shift, discovered significant gaps in time data
#ggplot(tips) +
#  geom_col(aes(x=day,y=tip, fill=time))

##barchart by gender
#ggplot(tips) +
#  geom_col(aes(x=day,y=tip, fill=sex))

##heatmap count by gender
#tips %>%
#  count(day,sex) %>%
#  ggplot(aes(x=day,y=sex)) +
#  geom_tile(mapping=aes(fill=n))
```