Project Part 1

Data Set Description

Background Information

This data set(Reference 1) contains data on Americans' preferences and usage practices regarding the Oxford comma. I understand this is a statistics course, but I appreciated the juxtaposition of statistics regarding an English subject. Anyways, for those unfamiliar with the Oxford comma, it is the comma separating a list of three or more things before "and". For example, in the sentence: "Statistics is fun, cool, and exciting." the Oxford comma is the one after "cool" and before "and." This data set examines Americans' usage of the Oxford comma and what factors determine whether an individual uses the Oxford comma or not.

Data Organization

This data is composed of responses to survey questions. Therefore, each row in the dataset represents a different respondent to the survey. There were 1,129 total respondents to the survey and each one is represented as a separate row in the dataset. Each column header represents a different question that was asked to the respondents. The data inside the columns then represents the response of the individual. The primary question is "In your opinion, which sentence is more gramatically correct?" with response options of "It's important for a person to be honest, kind, and loyal." and "It's important for a person to be honest, kind and loyal." This then shows whether the respondent uses the Oxford comma or not. The following survey questions attempt to categorize the respondents to determine why they use or don't use the Oxford comma.

Data Charateristics

This data is a sample of the American population. This sample was collected through a survey ran by FiveThiryEight and SurveyMonkey Audience from June 3 to 5 and polled 1,129 Americans. FiveThirtyEight is a polling aggregate website that uses polling and statistical analyses to look at topics ranging from sports to politics to pop culture. SurveyMonkey Audience is a survey tool that allows surveyors to reach a specific audience, such as a representative sample of the American population. With SurveyMonkey Audience a surveyor

is able to design a survey and pick an appropriate audience. The audience is made up of "a diverse online population that voluntarily joined a program to take surveys."(Reference 2)

Potential Issues and Analysis

One potential issue with this dataset is the number of no responses for some of the questions, which creates a sort of nonresponse bias. For example, income level was left unanswered for 25.98% of the total respondents which makes this variable particularly dangerous to use. Education level was also left unanswered for 9.13% of respondents. This is a significant problem because variables like these would be very useful for easily classifying respondents into different segments but now that is impossible to do for some of the respondents. Another potential issue is whether the group of people who voluntarily take surveys online on comma usage are really a reliable sample of the general American population. 54.96% of survey respondents said they had acheived a Bachelor's degree or higher. This number rises to 60.49% of respondents said they had acheived a bachelor's degree or higher when not counting the blank/null responses to educaton. The latest Census data shows that just 33.4% of Americans have earned a bachelor's degree or higher(Reference 3). Therefore, it seems that there is some sampling bias in our data, as being more than 20 points off the actual American population education average just by chance seems unlikely.

```
## R code
blankincome <- which(comma$Income == "")
#percentage no response to income:
length(blankincome) / (nrow(comma)) * 100
```

```
## [1] 25.95217
```

```
blankedu <- which(comma$Education == "")
#percentage no response to education level:
length(blankedu) / (nrow(comma)) * 100
```

```
## [1] 9.123118
```

```
edu <- which(comma$Education == "Graduate degree" | comma$Education == "Bachelor degree"
#Bachelor degree percentage of respondents:
length(edu) / (nrow(comma) ) * 100
```

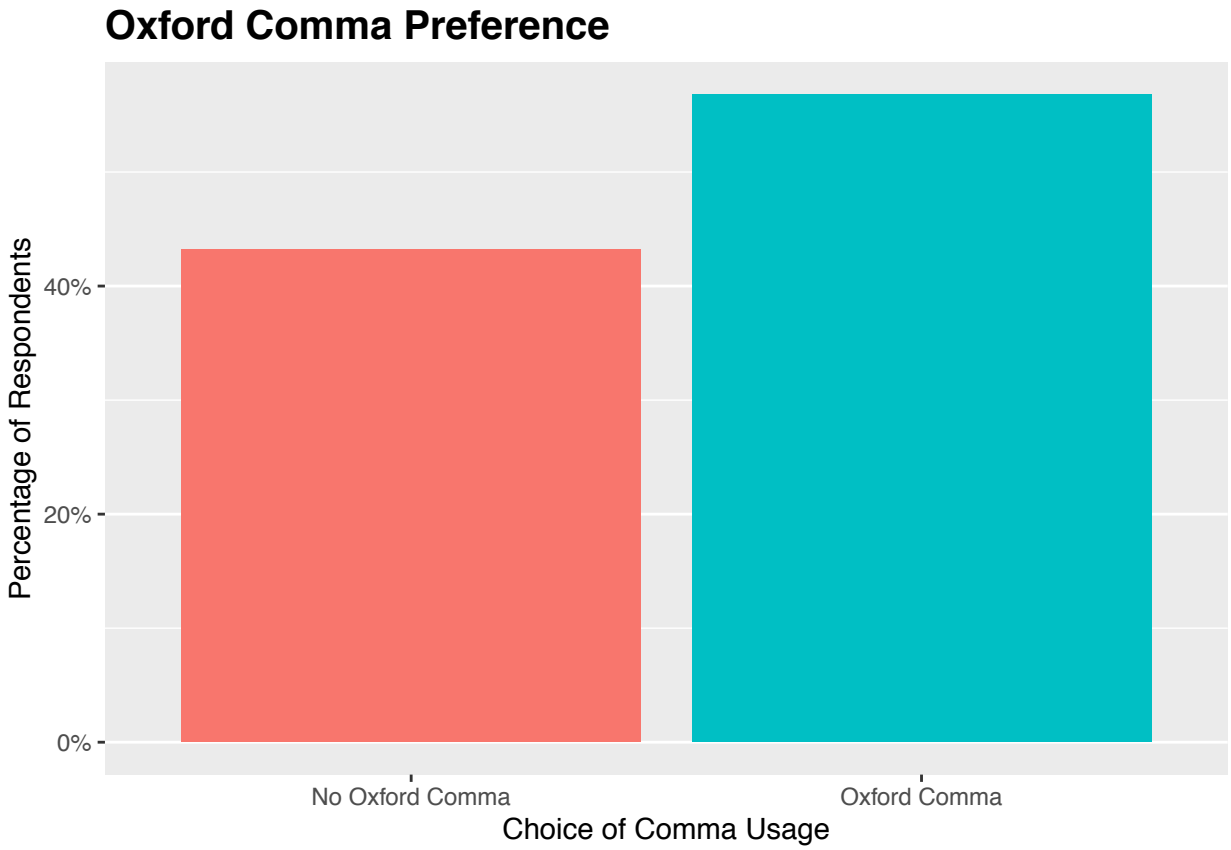
```
## [1] 54.91585

#Bachelor degree percentage of respondents w/o null responses:
length(edu) / (nrow(comma) - length(blankedu) ) *100

## [1] 60.42885
```

Summary 1

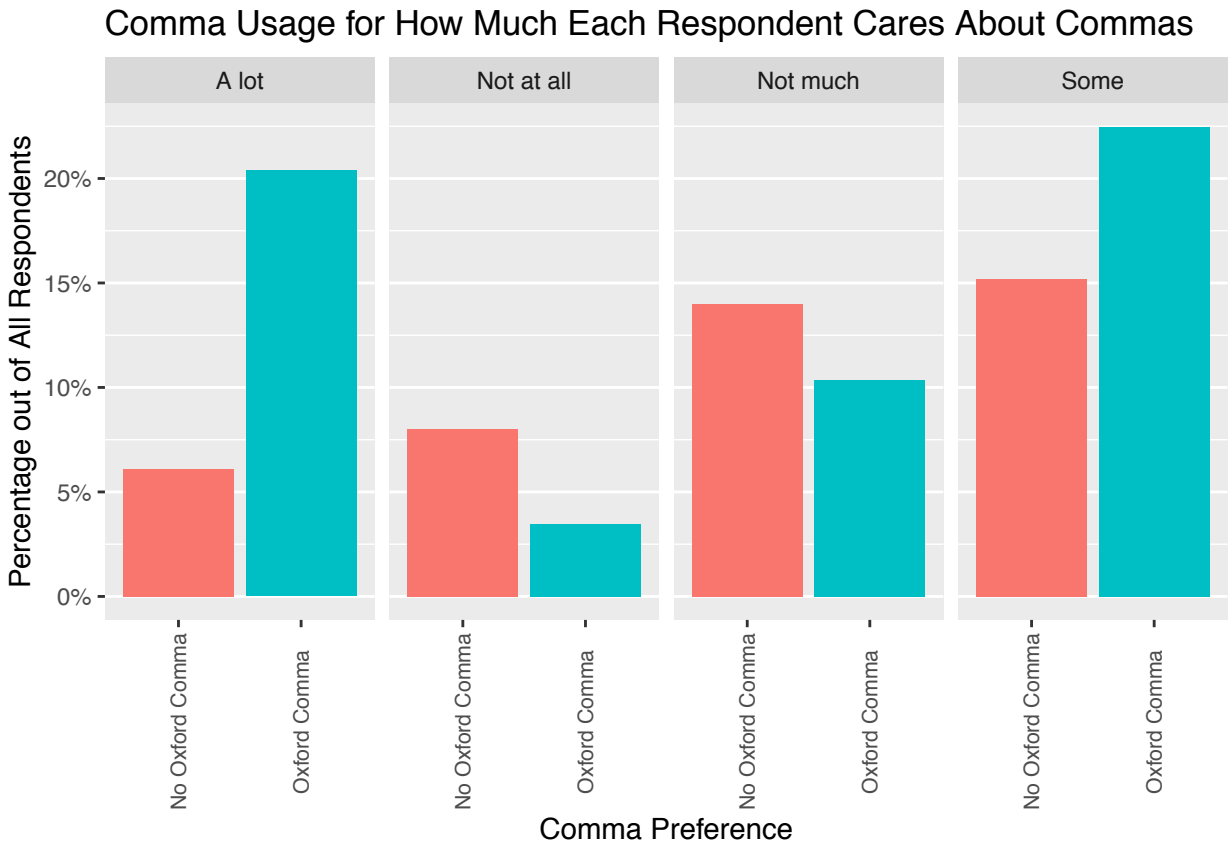
```
response1 <- comma[,2]
comma$CommaChoice = factor(comma$CommaChoice, labels=c("No Oxford Comma","Oxford Comma")
ycomma <- which(response1 == "Oxford Comma")
ycommacount <- response1[ycomma] #responses with Oxford comma
ycommapercent <- length(ycommacount) / nrow(comma) #percentage use Oxford comma
ncommacount <-response1[-ycomma] #responses without Oxford comma
ncommapercent <- length(ncommacount) / nrow(comma) #percentage don't use Oxford comma
commapref<- ggplot(comma, aes(x=CommaChoice)) +
  geom_bar(aes(y=(..count..)/sum(..count..), fill= CommaChoice)) + #Reference4
  theme(panel.grid.major.x=element_blank(), panel.grid.minor.x=element_blank()) +
  labs(title= "Oxford Comma Preference", y= "Percentage of Respondents",
       x= "Choice of Comma Usage") +
  theme(plot.title=element_text(face="bold", size=15)) +
  scale_y_continuous(labels = scales::percent_format(accuracy=1)) +
  theme(legend.position = "none")
commapref
```



This bar graph shows the comparison between the percentages of respondents who chose the sentence without the Oxford comma compared to with the Oxford comma. The exact value of those who chose the Oxford Comma is ycommapercent= 56.78%, and the exact value of those who didn't choose the Oxford Comma sentence is ncommapercent = 43.22%

Summary 2

```
noblank <- which(comma$CommaImportance != "")
commaNoBlank <- comma[noblank,] #remove blank rows for comma importance
commapref<- ggplot(commaNoBlank, aes(x=CommaChoice)) +
  geom_bar(aes(y=(..count..)/sum(..count..), fill= commaNoBlank$CommaChoice)) +
  theme(panel.grid.major.x=element_blank(), panel.grid.minor.x=element_blank()) +
  scale_y_continuous(labels = scales::percent_format(accuracy=1))
commapref + facet_grid(.~commaNoBlank$CommaImportance) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, size=8)) +
  labs(title= "Comma Usage for How Much Each Respondent Cares About Commas",
       x= "Comma Preference", y= "Percentage out of All Respondents") +
  theme(legend.position = "none")
```



This summary shows the relationship between how much a respondent cares about comma usage and whether that individual chose the sentence with or without the Oxford comma. From this we can see that those respondents who care a lot about commas were approximately 4x as likely to use the Oxford comma, whereas those respondents who don't care at all about comma usage were almost 3x as likely to not use the Oxford comma. This shows the high correlation between importance of comma usage to respondents and actual Oxford comma usage.

Summary 3

```
noBlankEdu <- comma[-blankedu,] #remove blank responses
#data set with only bachelor's degree or higher
bach <- noBlankEdu[which(noBlankEdu$Education == "Graduate degree"
                         | noBlankEdu$Education == "Bachelor degree" ), ]
bachcomma <- (length(which(bach$CommaChoice == "Oxford Comma")) / nrow(bach)) *100
bachnocomma <- 100- bachcomma
lowedu <- noBlankEdu[which(noBlankEdu$Education == "High school degree"
                           | noBlankEdu$Education == "Less than high school degree" ), ]
```

```
loweducomma <- (length(which(lowedu$CommaChoice == "Oxford Comma")) / nrow(lowedu)) *100
lowedunocomma <- 100 - loweducomma
#Percentage of Bachelor's degree or higher respondents who use the Oxford comma is
round(bachcomma, 2)
```

```
## [1] 59.84
```

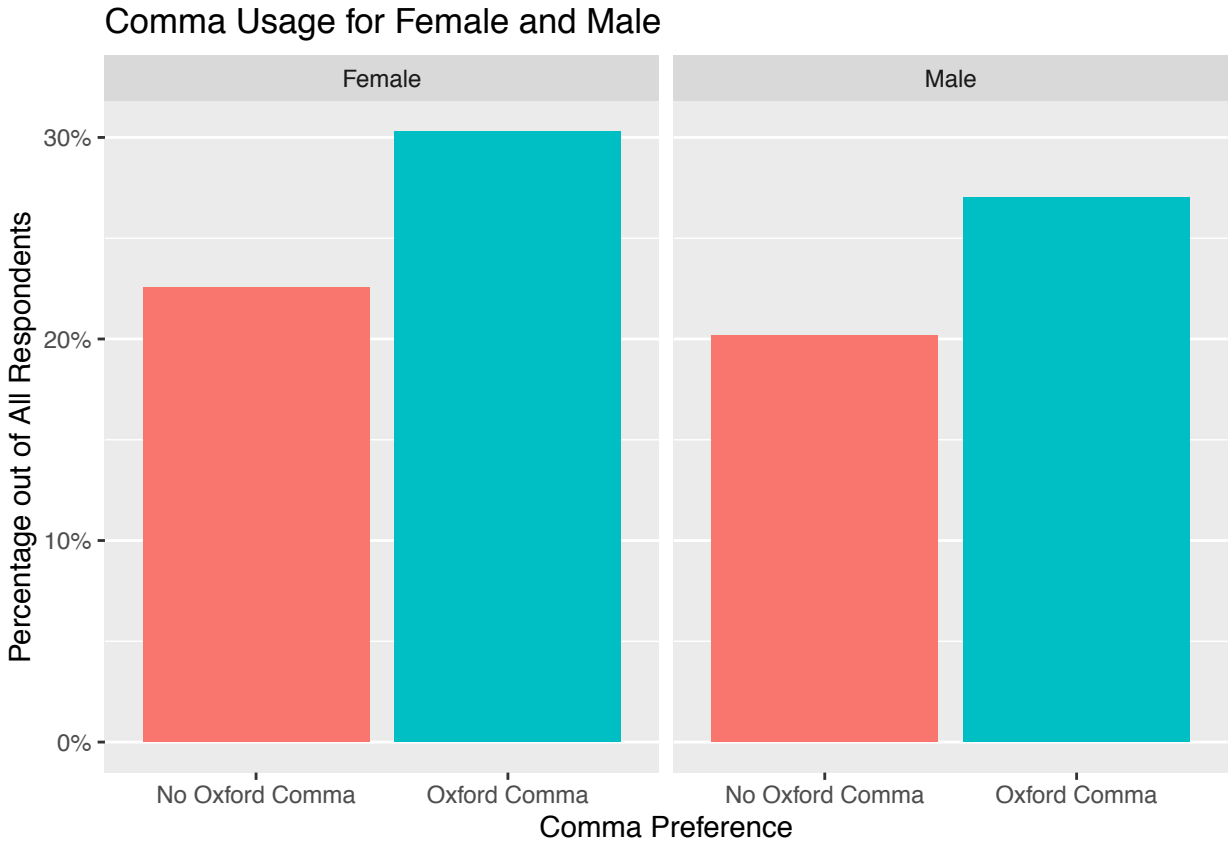
```
#Percantage of high school degree or lower respondents who use the Oxford comma is
round(loweducomma, 2)
```

```
## [1] 54.05
```

Another factor that seems like it would be important in whether a respondent uses the Oxford comma or not is the education level of the respondent. After removing no responses and filtering the data appropriately, I determined that individuals with a Bachelor's degree or higher chose the Oxford comma sentence %59.84 of the time and those with a high school degree or lower chose the Oxford comma sentence %54.05 of the time. The average for all respondents was ycommapercent = %56.78, so it seems that education level does not have that large of an effect on Oxford comma usage.

Summary 4

```
noblank <- which(comma$Gender != "")
genderNoBlank <- comma[noblank,] #remove blank rows for gender
genderpref<- ggplot(genderNoBlank, aes(x=CommaChoice)) +
  geom_bar(aes(y=(..count..)/sum(..count..), fill= genderNoBlank$CommaChoice)) +
  theme(panel.grid.major.x=element_blank(), panel.grid.minor.x=element_blank()) +
  scale_y_continuous(labels = scales::percent_format(accuracy=1))
genderpref + facet_grid(.~genderNoBlank$Gender) +
  labs(title= "Comma Usage for Female and Male",
       x= "Comma Preference", y= "Percentage out of All Respondents") +
  theme(legend.position = "none")
```



This graph looks at whether respndents' gender makes a difference on comma usage. However, the difference in percentages for Oxford Comma and No Oxford Comma is approximately the same for both Female and Male. This makes it seem that gender does not have a significant effect on Oxford comma usage. Overall, it seems that an individual's own personal opinion on importance of comma usage is the most important factor for determing comma usage.

References

1. https://github.com/fivethirtyeight/data/blob/master/comma-survey/comma-survey.csv
2. https://help.surveymonkey.com/articles/en_US/kb/SurveyMonkey-Audience
3. https://thehill.com/homenews/state-watch/326995-census-more-americans-have-college-degrees-tha
4. https://stackoverflow.com/questions/3695497/show-instead-of-counts-in-charts-of-categorical-variab