Exploratory Data Analysis

Patrick Roehling

2/25/2021

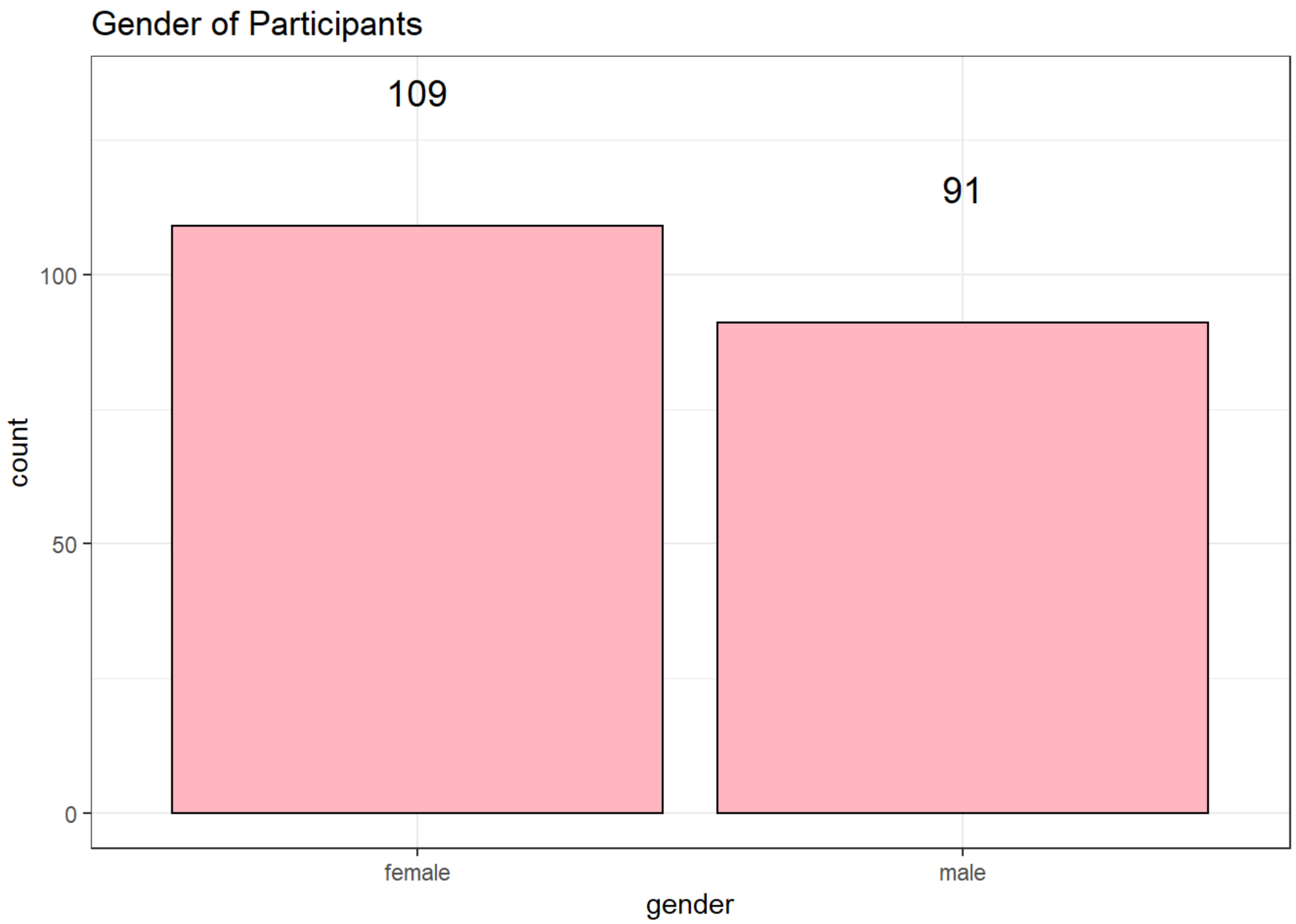
Introduction

The data set that I chose for my Exploratory Project is the `High School and Beyond` data set. The set encompasses 11 variables with 200 observations. It is a study by NCES' National Longitudinal Studies Program that tracks the development of students from youth to adulthood. This started for each participant while they were in their elementary years or while in high school. The data followed each participant either to postsecondary education or to the workforce. The data was not only collected from the students themselves, but also collected from teachers, partents, school administrators, and others. The data collected gives insight into educational, vocational, and personal development of each participant. The variables that I will be analyzing include: gender, socioeconomic status (SES), and race to get an idea of the demographic that was studied.

Univariate Descriptions of the Variables

1. Gender

```
ggplot(HSAB, aes(x=gender)) + theme_bw() + geom_bar(aes(y=..count..), color="black", fill="light
pink") + ggtitle("Gender of Participants") + geom_text(aes(y=..count.. + 25, label=..count..), s
tat='count', size = 5)
```



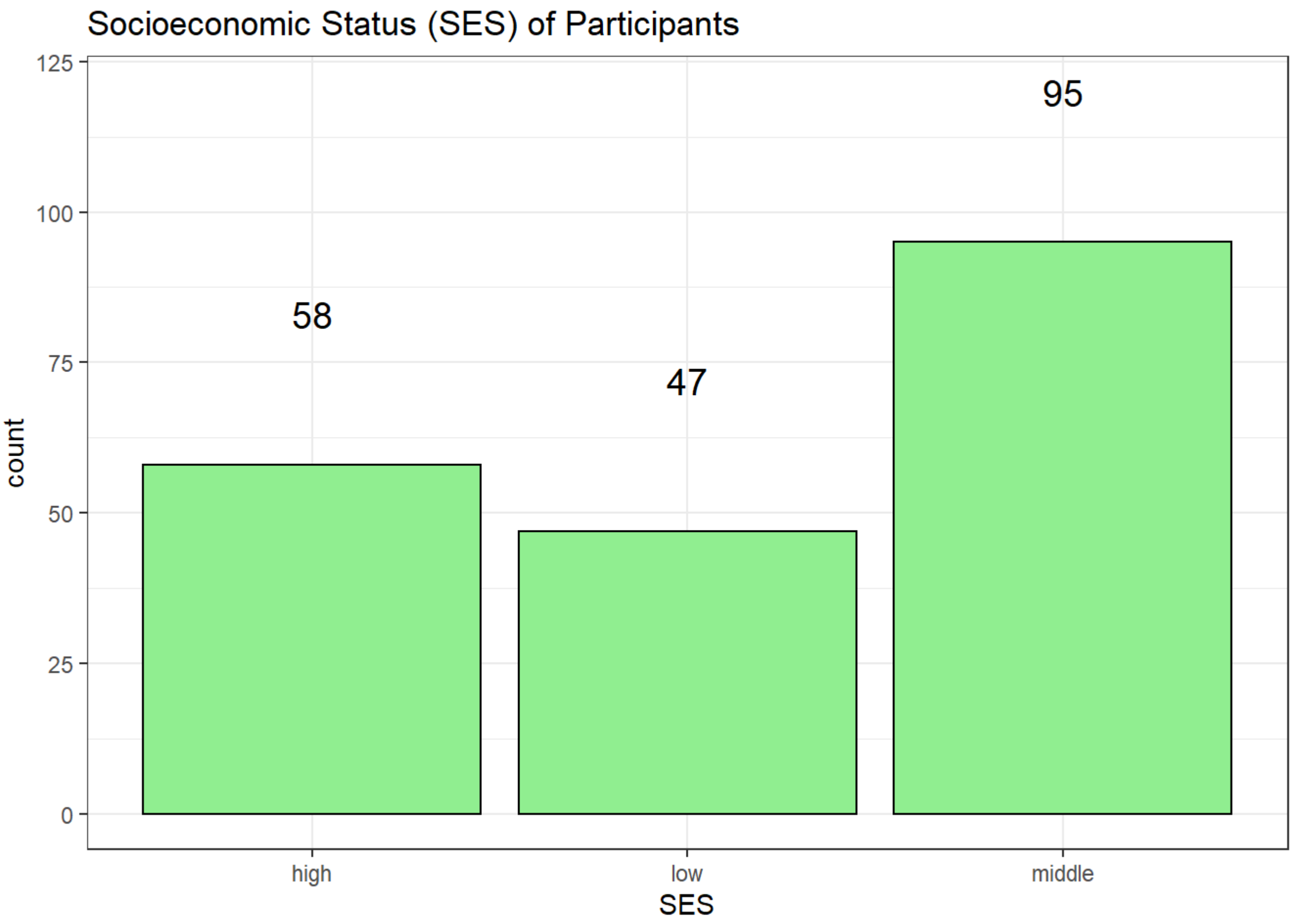
```
table(HSAB$gender) %>% prop.table * 100
```

```
##
## female   male
##   54.5   45.5
```

The barchart and table above show that the study follows 109 females (54.5%) and 91 males (45.5%).

2. Socioeconmic Status

```
ggplot(HSAB, aes(x=ses)) + theme_bw() + geom_bar(aes(y=..count..), color="black", fill="light green") + ggtitle("Socioeconomic Status (SES) of Participants") + xlab("SES") + geom_text(aes(y=..count.. + 25, label=..count..), stat='count', size = 5)
```



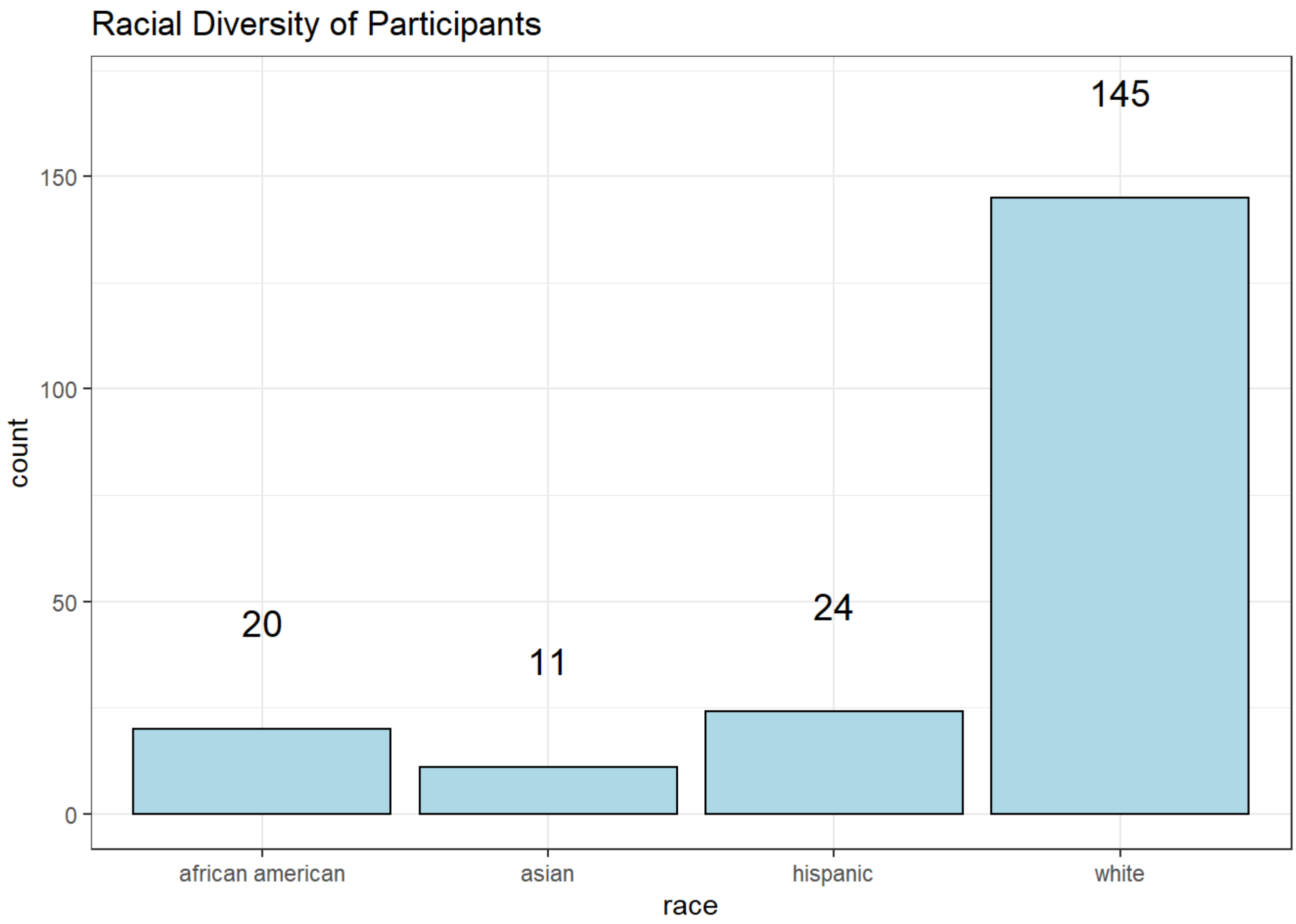
```
table(HSAB$ses) %>% prop.table * 100
```

```
## 
##   high    low middle 
##   29.0   23.5   47.5
```

Looking at the barchart and table above reveals that the study had a majority of middle class participants at 47.5%, while having relatively equal amounts of lower and higher class participants at a respective 23.5% and 29.0%.

3. Race

```
ggplot(HSAB, aes(x=race)) + theme_bw() + geom_bar(aes(y=..count..), color="black", fill="light blue") + ggtitle("Racial Diversity of Participants") + geom_text(aes(y=..count.. + 25, label=..count..), stat='count', size = 5)
```



```
table(HSAB$race) %>% prop.table * 100
```

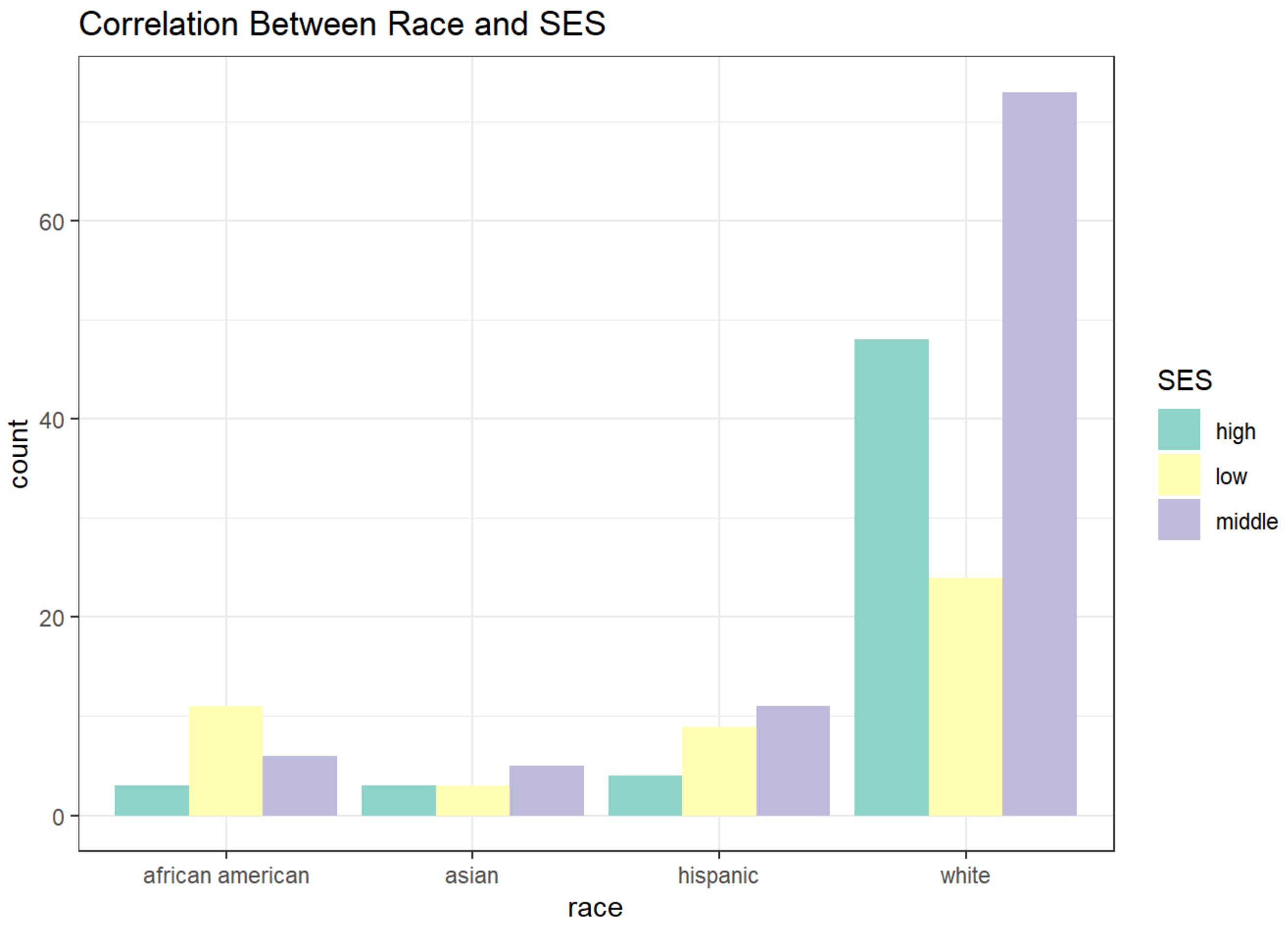
```
##
## african american            asian         hispanic            white
##             10.0              5.5             12.0             72.5
```

In terms of race, this study mainly focuses on a white demographic, which makes up 72.5% of the participants as shown by the table above. The others are all very low with Asians being represented the least at 5.5%.

Bivariate Comparison Between Two Variables of Interest

Race and Socioeconomic Status (SES)

```
ggplot(HSAB, aes(x=race, fill=ses)) + theme_bw() + geom_bar(position = "dodge") + ggtitle("Corre
lation Between Race and SES") + scale_fill_brewer(palette="Set3", name="SES")
```



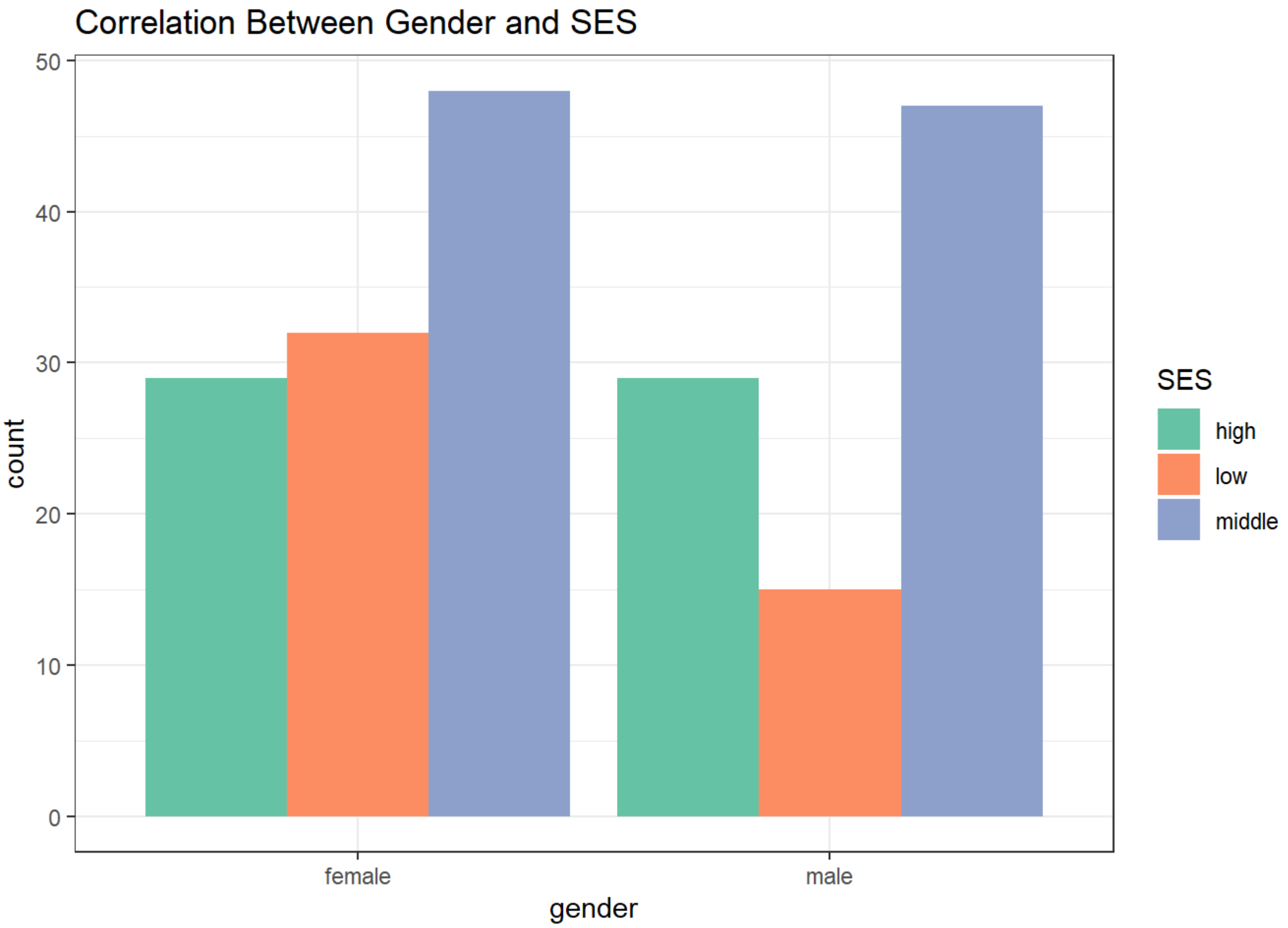
```
table(HSAB$race, HSAB$ses) %>% prop.table(margin=1) %>% round (3) * 100
```

```
##
##                    high  low middle
##   african american 15.0 55.0   30.0
##   asian            27.3 27.3   45.5
##   hispanic         16.7 37.5   45.8
##   white            33.1 16.6   50.3
```

The barchart above gives a visual representation of the socioeconomic status of each racial demographic. For example the barchart clearly shows that not only are there more white participants, but of those participants the majority of them are of high or middle class. This is confirmed by the two-way frequency table, which shows that 33.1% of the white participants are high class while 50.3% are of middle class. Alternatively, not only are the other three demographics: Asian, African American, and Hispanic represented less, but they are on average less likely to be high class. The only exception to this is Asians inwhich their proportions of high and lower class are both equal at 27.3%. However, like African Americans and whites, Asians too have a higher middle class percentage compared to the other classes. Another outlier is that African Americans, instead of having a majority of middle class like all the other demographics represented, they instead have a majority of lower class representation at 55%. This is shocking considering that this 55% is the greatest percentage given by the two-way frequency table.

Gender and Socioeconomic Status (SES)

```
ggplot(HSAB, aes(x=gender, fill=ses)) + theme_bw() + geom_bar(position = "dodge") + ggtitle("Cor
relation Between Gender and SES") + scale_fill_brewer(palette="Set2", name="SES")
```



```
table(HSAB$gender, HSAB$ses) %>% prop.table(margin=2) %>% round (3) * 100
```

```
##
##          high  low middle
##   female 50.0 68.1   50.5
##   male   50.0 31.9   49.5
```

Looking at both the barchart and two-way frequency table above it shows that both the female and male particpants have relatively equal representation of of both middle and high class. Females make up 50.5% of the middle can participants while males make up the other 49.5%. Similarly, in terms of high class it is equal with each having 50.0% representation. Where males and females differ in SES is lower class. In this class, 68.1% are females and 31.9% are males.