EDA Project

R Markdown

INTRODUCTION: I am going to explore if there a discernable difference in the grades between scholars in private and public schools using the High School and Beyond data set. This data was collected by the National Longitudinal Studies Program and was meant to study people's development from children to adults. This survey includes factors concerning their education, vacational, and personal development in their lives. I think that private schools' grades will be higher than public schools because private schools have more funding and individual attention to offer to their students versus public schools.

```
HS2B <- read.delim ("/Users/jessicamoody/Desktop/math130/data/hsb2.txt", header=TRUE, sep="\t")
head(HS2B)
```

```
##    id gender  race    ses schtyp       prog read write math science socst
## 1  70   male white    low public    general   57    52   41      47    57
## 2 121 female white middle public vocational   68    59   53      63    61
## 3  86   male white   high public    general   44    33   54      58    31
## 4 141   male white   high public vocational   63    44   47      53    56
## 5 172   male white middle public   academic   47    52   57      53    61
## 6 113   male white middle public   academic   44    52   51      63    61
```

UNIVERIATE EXPLORATION: There were a variety of observations to choose from, but I decided to first investigate the school type variable.

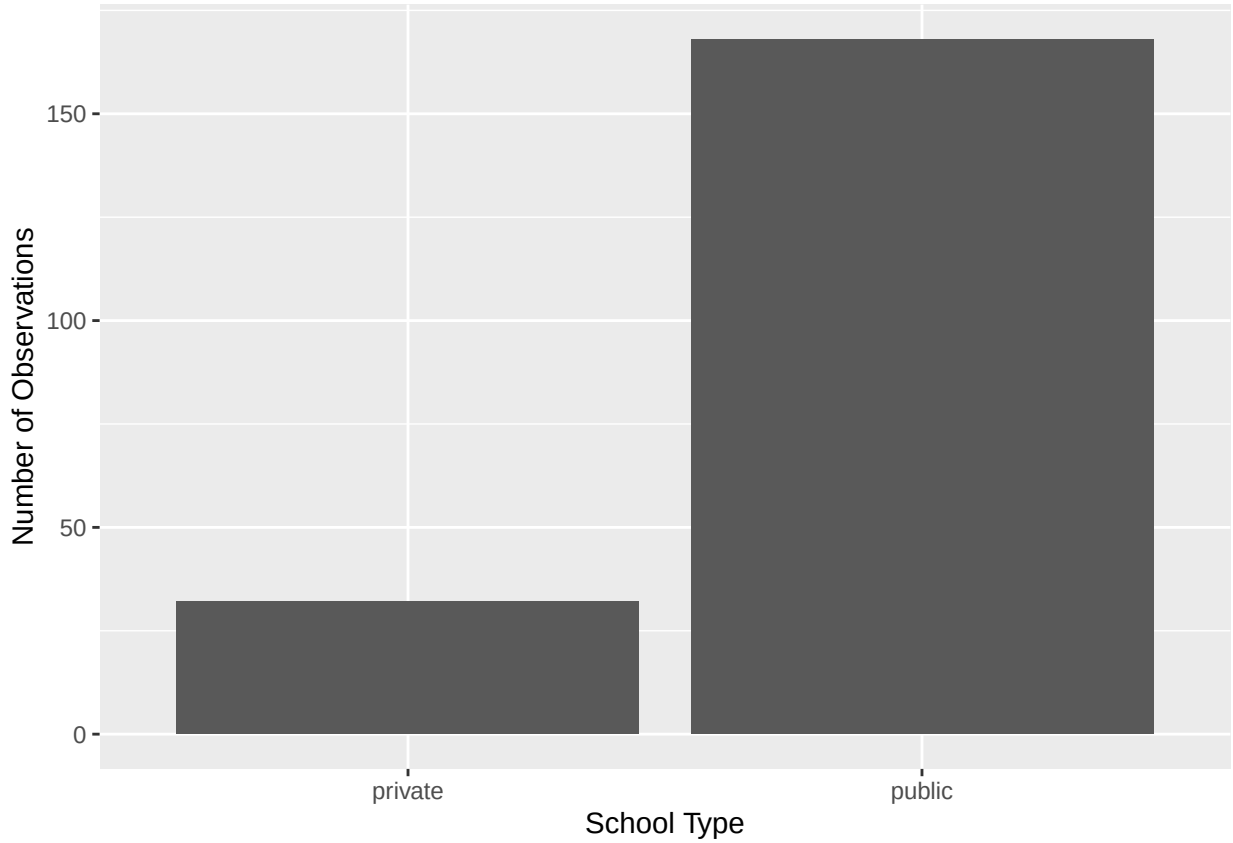
```
table(HS2B$schtyp)
```

```
##
## private  public
##      32     168
```

```
library(ggplot2)
```

```
## Warning in register(): Can't find generic 'scale_type' in package ggplot2 to
## register S3 method.
```

```
ggplot(HS2B, aes(x=schtyp))+geom_bar(position="dodge")+xlab("School Type")+ylab("Number of Observations
```



This data is mainly public school students. I then decided to make a total_grade variable in order to compare how well students performed in all aspects of academics.

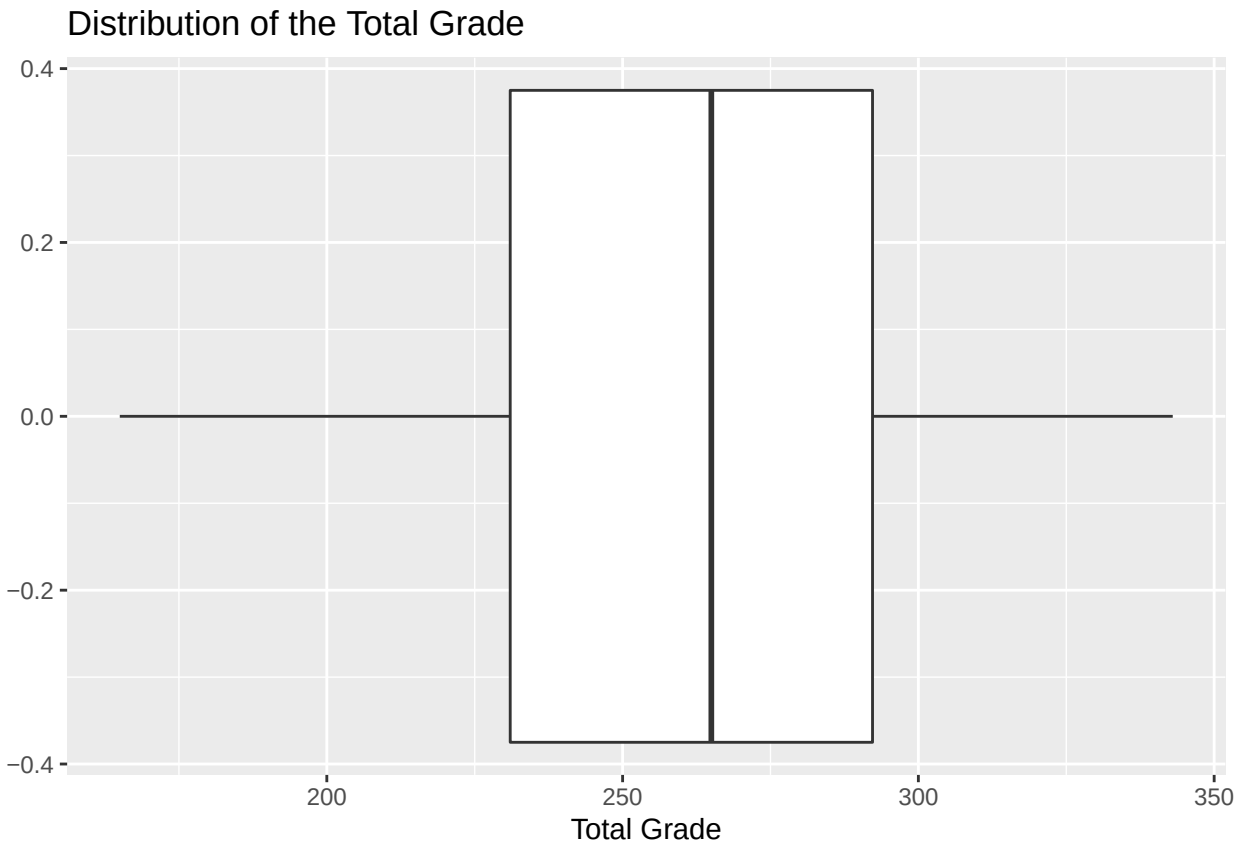
```
library(dplyr)
```

```
##
## 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
```

```
HS2B= mutate(HS2B, total_grade= read+write+math+science+socst)
head(HS2B$total_grade)
```

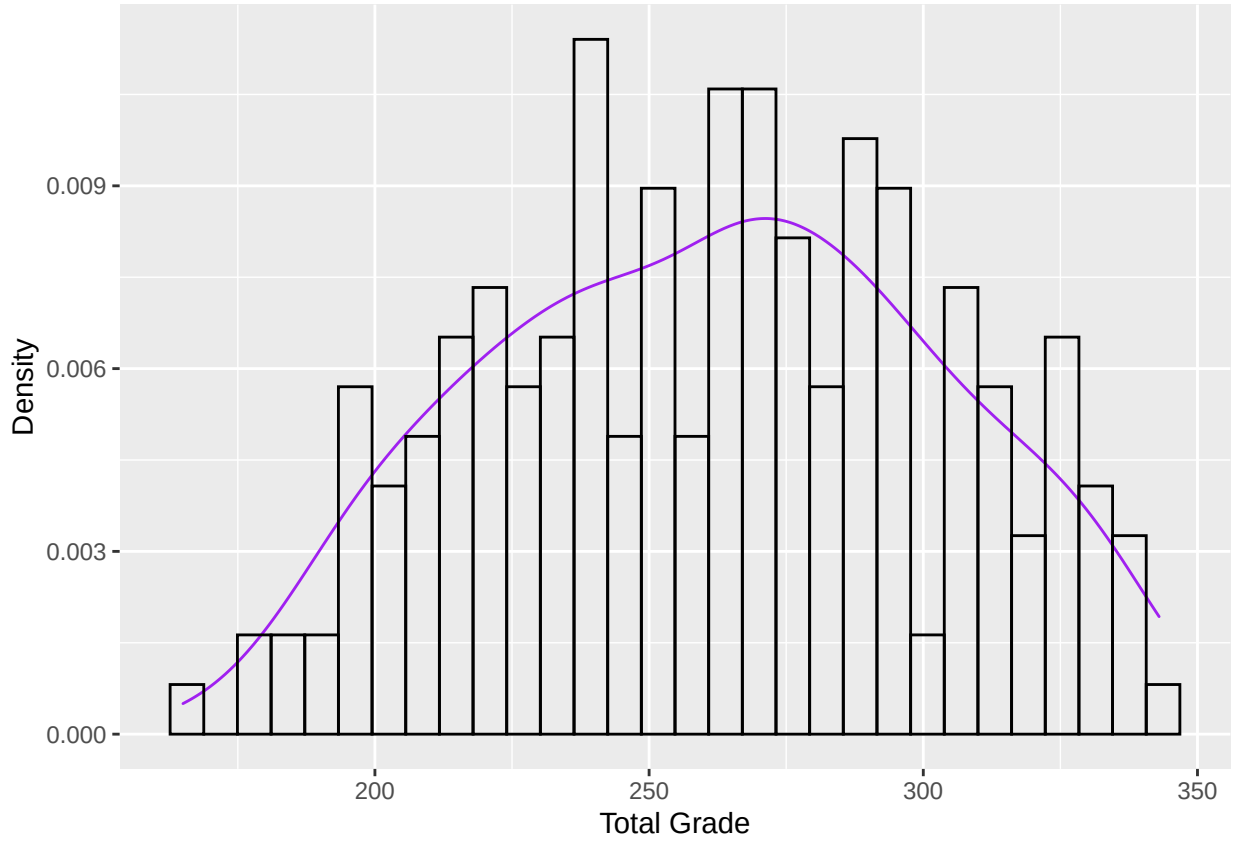
```
## [1] 254 304 220 263 270 271
```

```
ggplot(HS2B, aes(x=total_grade))+geom_boxplot()+ggtitle("Distribution of the Total Grade")+xlab("Total
```




```
ggplot(HS2B, aes(x=total_grade))+geom_density(col="purple")+geom_histogram(aes(y=..density..), col="bla
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The distribution of the combined, total grade is pretty regular. There are no extreme outliers and most of the data sits in the center of the distribution. The mean of the total grade is about 265. I decided to use a boxplot and density overlaid histogram to showcase the distribution because it is easy to see where the majority of the data lies and will make it easier to compare to other distribution graphs.

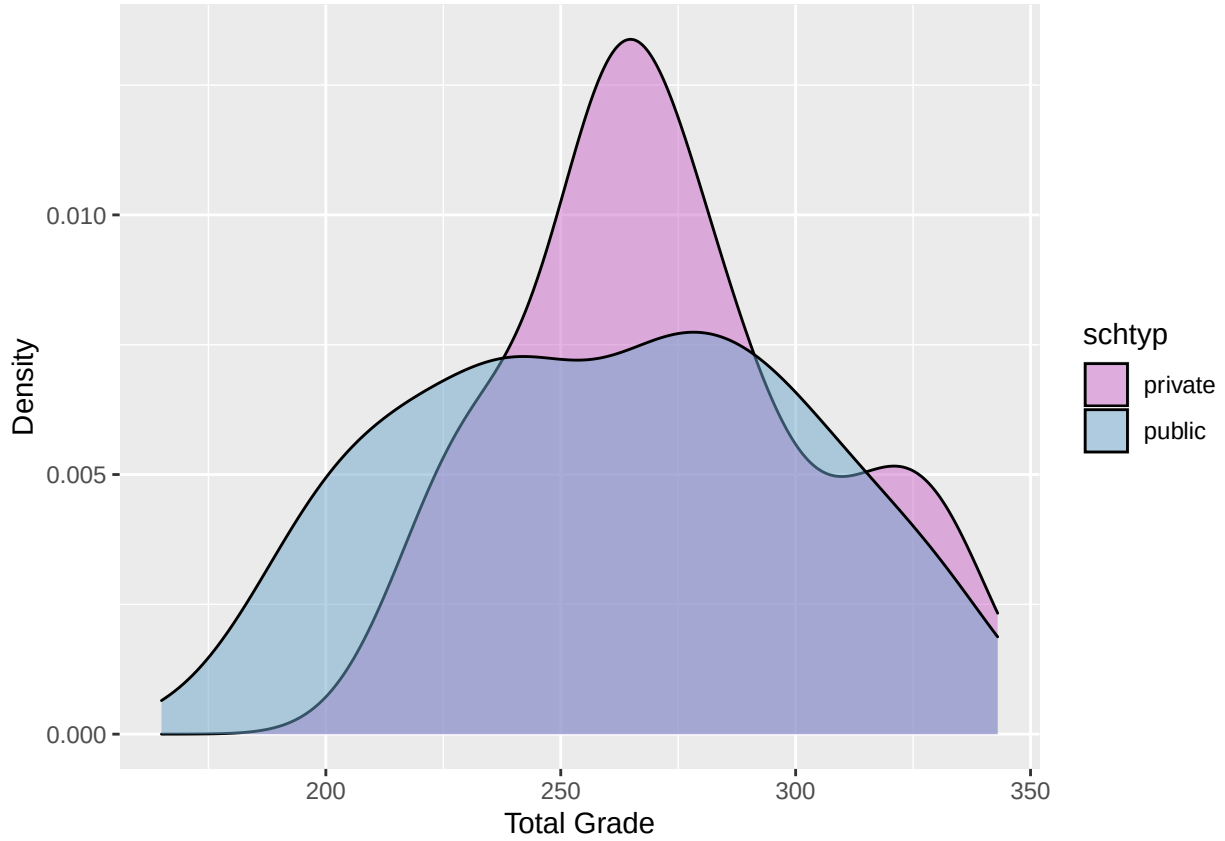
BIVARIATE EXPLORATION: The data I observed above made me wonder if there would be a difference between the grades for each school type. I decided to investigate this by creating a mean total grade for each school type and compare them.

```
HS2B%>%group_by(schtyp)%>%select(schtyp, total_grade)%>%summarize(mean_total_grade=mean(total_grade, na
```

```
## # A tibble: 2 x 2
##   schtyp  mean_total_grade
##   <chr>              <dbl>
## 1 private             273.
## 2 public              260.
```

The mean for private schools is noticiably larger than public schools and the overall, combined total grade mean. The mean for private schools would sit in the upper quadrant of the distribution of total grade graph, meaning that overall their students are performing better than those in public schools. I then use different graphs to visualize the distribution of total grades in private versus public schools to understand how students are performing compared to the average.

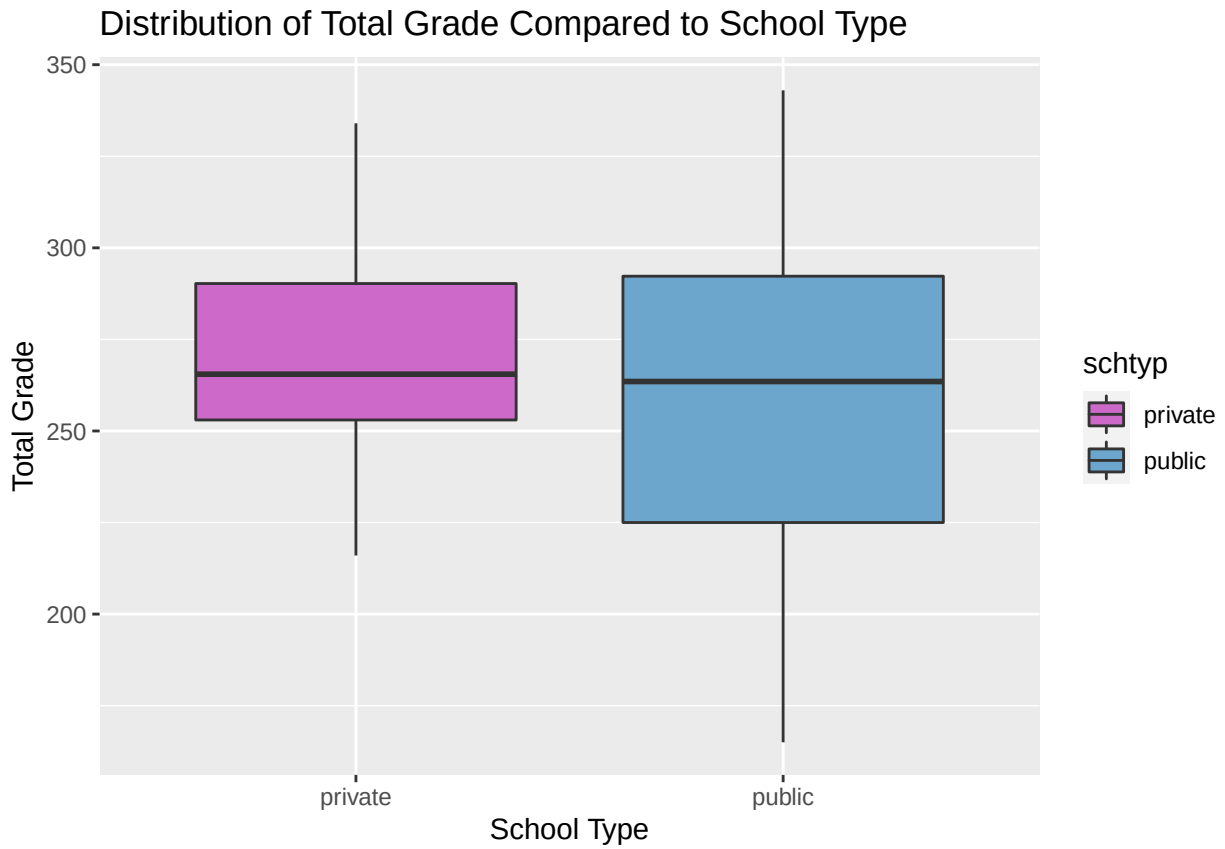
```
ggplot(HS2B, aes(x=total_grade, fill=schtyp))+geom_density(alpha=.5)+xlab("Total Grade")+ylab("Density")
```



This graph shows how the distribution of grades in public schools are mostly even, as the rise in the graph occurs right around the mean. This graph is also similar to the total grade density curve displayed above, meaning it is following the overal data trends. There is no large group of people who are scoring high total grades. I think that means that most students are probably learning at the same level and receiving the same amount of attention.

In private schools, the graph shows how there is a large group of people who score very close to the mean and a small group that scores above. The group of people who score below the mean is drastically smaller compared to those below the mean in public schools. I think this means that the more advanced students are given more attention and benefits to succeed while those below average must adjust to the others' learning styles.

```
ggplot(HS2B, aes(x=schtyp, y=total_grade, fill=schtyp))+geom_boxplot()+ggtitle("Distribution of Total G
```



This boxplot shows the concentration of the distribution of grades in each school type. The public school box is much more elongated and spans over almost the entire graph. The upper and lower quadrant are fairly large, with the lower quadrant being larger than the upper.

The private school box is more condensed on the upper half of the graph. Its lower tail is also much shorter than the public schools'. Its lower quadrant is also smaller than its upper quadrant.

CONCLUSION: I think all these aspects of this graph conicide with my statements about the previous graph and my initial hypothesis. Even though each school type's mean is very similar, private schools' grades are higher than public schools, the majority of their students score above average, and they have much less students scoring below average. I think this is due to how each type of school is structured and the amount of attention each student receives in the group and individually.