Exploratory Data Analysis Project

Connor Lydon

03/02/2024

Introduction

For this project, we will be analyzing the High School and Beyond data set which was established to track and study the development of young people through the lens of education, vocation, and personal metrics. From this data set of 11 variables and 200 observations, we will be looking at program type and Science test scores.I am interested to see if there is a correlation between the type of program students chose and their science test scores, there might be some interesting relationships we can find in the data.

```
library(ggplot2)
library(forcats)
library(RColorBrewer)
```

```
hsb2 <- read.table("C:/Users/Admin/OneDrive/Documents/math130/data/HighSchoolAndBeyond.txt", header=TRU
```

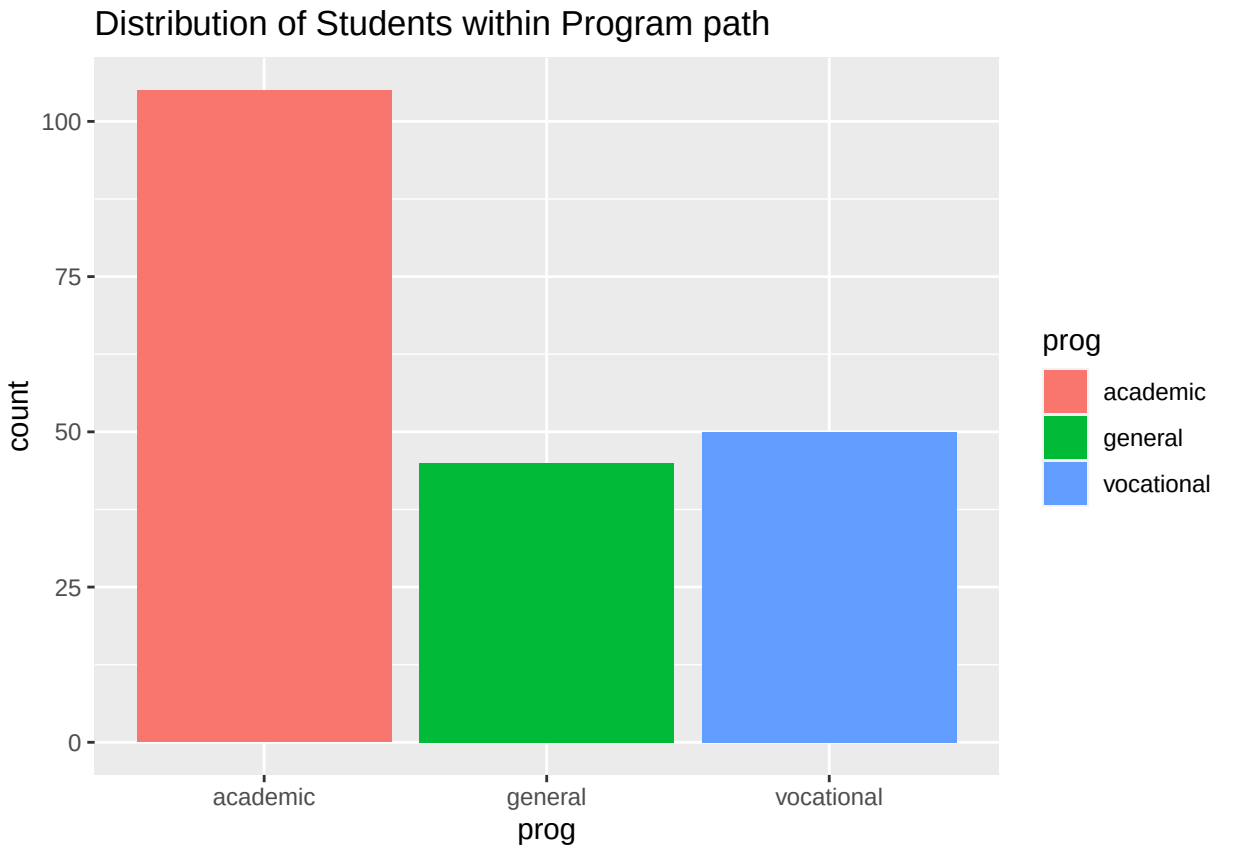
Univariate Exploration

Our first variable will be the distribution of programs that students chose.

```
table(hsb2$prog)
```

```
##
##   academic    general vocational
##        105         45         50
```

```
ggplot(hsb2, aes(x=prog, fill=prog)) + geom_bar() + ggtitle("Distribution of Students within Program pa
```



This bar chart shows that academic programs are more than twice as popular than general programs or vocational programs within the study. Within this bar chart we can see that there is a major bias towards academic programs within the students of our data set.

The next variable is Science test scores.

```
table(hsb2$science)
```

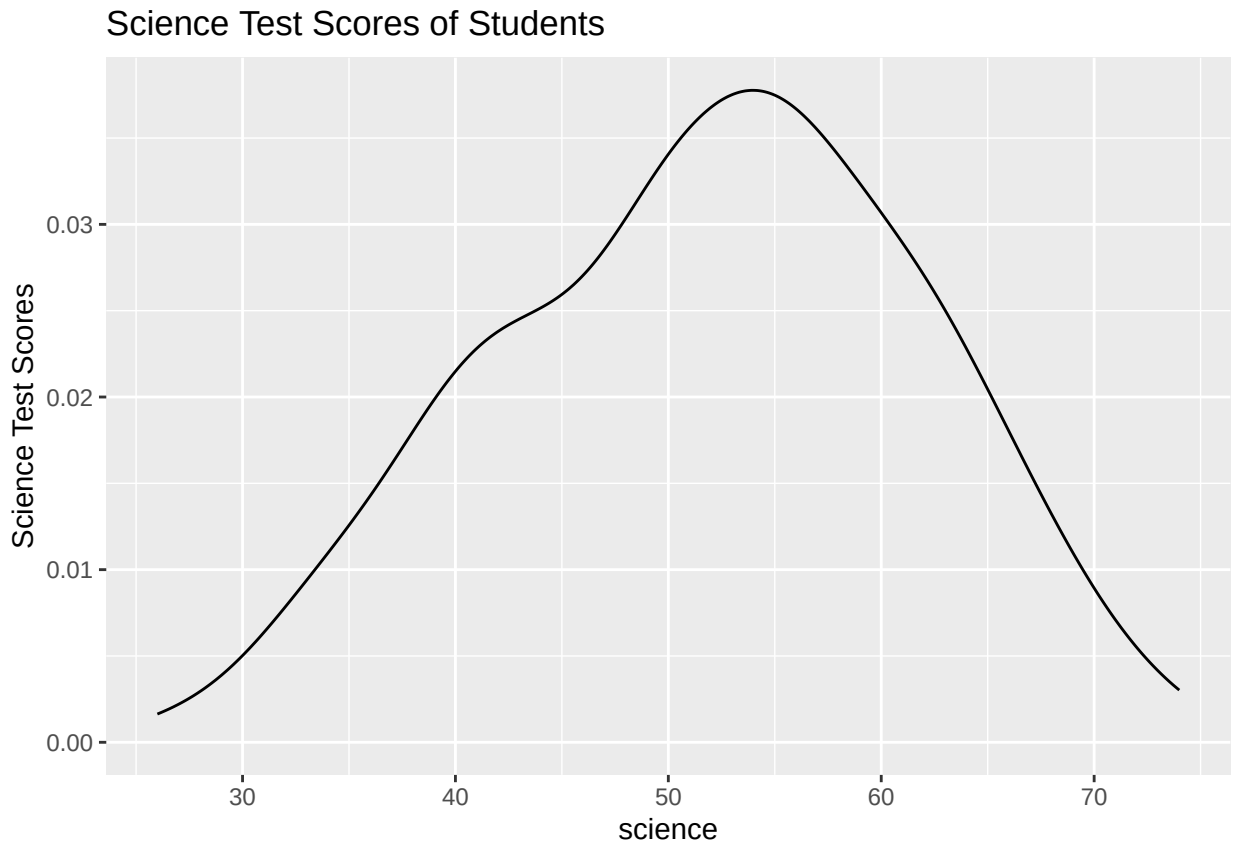
```
##
## 26 29 31 33 34 35 36 39 40 42 44 45 46 47 48 49 50 51 53 54 55 56 57 58 59 61
##  1  1  2  2  5  1  4 13  2 12 11  1  1 11  2  2 21  2 15  3 18  2  1 19  1 14
## 63 64 65 66 67 69 72 74
## 12  1  1  9  1  6  2  1
```

```
summary(hsb2$science)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.00   44.00   53.00   51.85   58.00   74.00
```

As can be seen from the summary statistics, the highest score on the science exam was a 74 and the lowest score was 26 with an mean test score of 51.85. This shows that there is a large spread of exam scores with an average test score right in the middle of the low and high test score.

```
ggplot(hsb2, aes(x=science)) + geom_density() + ggtitle("Science Test Scores of Students") +
  ylab("Science Test Scores")
```



Using all the data we now have at our disposal shows that there are more data points on the higher end of the range of scores. This can be seen as the median score is skewed slightly higher than the mean test score showing that there are slightly more higher scores and that some of the lower scores are dragging the mean down.

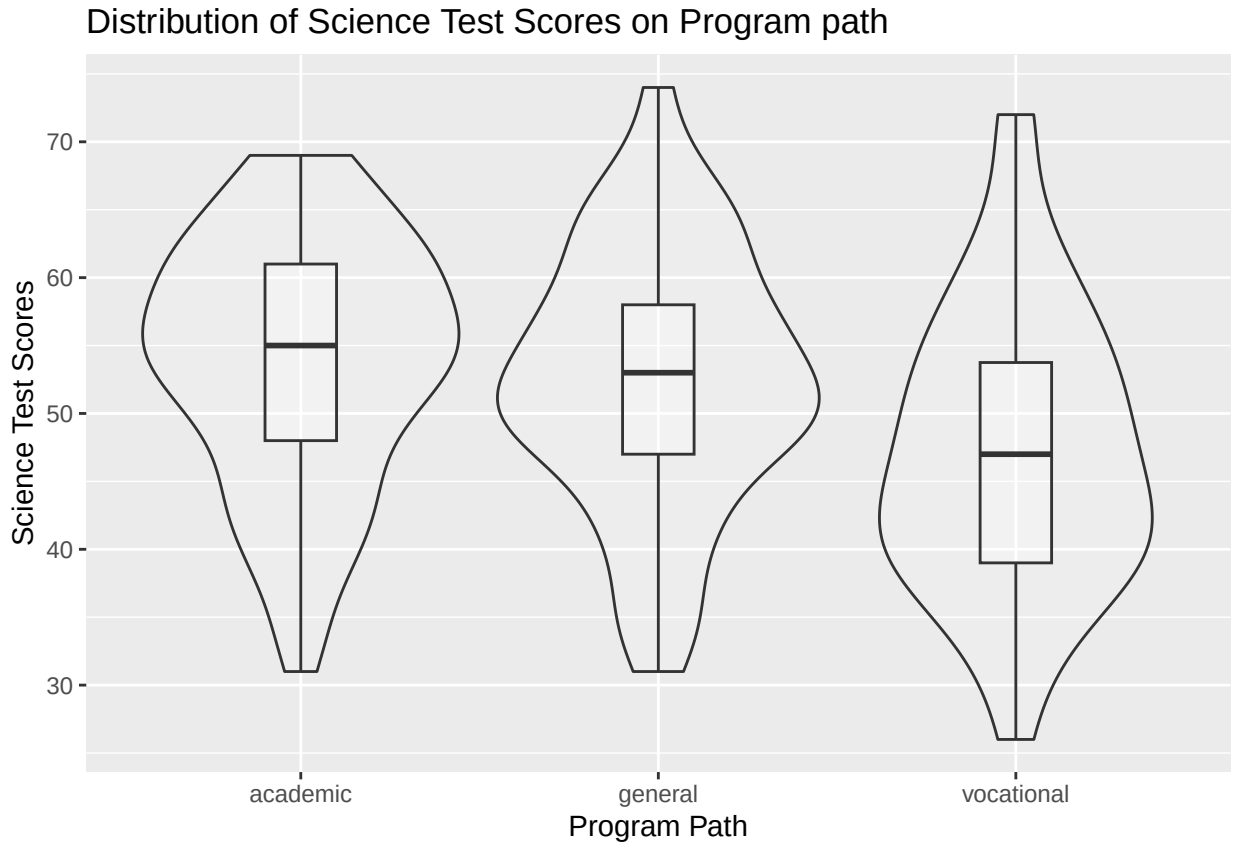
Bivariate Exploration

```
table(hsb2$prog, hsb2$science) %>% prop.table(margin=1) %>% round(3)
```

```
##
##                 26    29    31    33    34    35    36    39    40    42    44
##   academic   0.000 0.000 0.010 0.010 0.019 0.000 0.010 0.048 0.010 0.038 0.057
##   general    0.000 0.000 0.022 0.000 0.022 0.022 0.022 0.022 0.000 0.067 0.022
##   vocational 0.020 0.020 0.000 0.020 0.040 0.000 0.040 0.140 0.020 0.100 0.080
##
##                 45    46    47    48    49    50    51    53    54    55    56
##   academic   0.010 0.010 0.029 0.010 0.010 0.086 0.010 0.076 0.019 0.124 0.000
##   general    0.000 0.000 0.089 0.000 0.000 0.200 0.000 0.089 0.000 0.067 0.022
##   vocational 0.000 0.000 0.080 0.020 0.020 0.060 0.020 0.060 0.020 0.040 0.020
##
##                 57    58    59    61    63    64    65    66    67    69    72
##   academic   0.010 0.105 0.010 0.105 0.067 0.010 0.010 0.048 0.010 0.048 0.000
##   general    0.000 0.111 0.000 0.022 0.089 0.000 0.000 0.067 0.000 0.022 0.000
##   vocational 0.000 0.060 0.000 0.040 0.020 0.000 0.000 0.020 0.000 0.000 0.040
```

```
##
##                 74
##   academic   0.000
##   general    0.022
##   vocational 0.000
```

```
ggplot(hsb2, aes(x=prog, y=science)) +
  geom_violin(alpha=.1) +
  geom_boxplot(alpha=.3, width=.2) +
  ggtitle("Distribution of Science Test Scores on Program path") +
  ylab("Science Test Scores") +
  xlab("Program Path")
```



The boxplot is helping us to see some differences between the three different groups of students, those that took the vocational school route had a lower average science test scores. Overall, the academic school students had a higher average test scores than the general route, with the vocational route students having the lowest science test average. Another interesting thing to look at is that the vocational school had the highest range from the lowest to the highest test score.

Conclusion

My expectation that students who pursue the academic route of schooling had higher science test scores was correct. Something that I didn't expect was that the general route was very close to the average test scores and overall distribution of science test scores. Overall my main takeaway is that while there are differences

between the students path in which they take, I am sure there is more to take into factor to look at their future success than some science test scores.