Math 130 Final Project by Sidd Sharma

Introduction

The data set I chose to use for my final project is named ParentalHIV and was located under the following link: https://norcalbiostat.netlify.app/data/. This data set was part of a clinical data trial in which an intervention was implemented for families who have a parent with HIV. There are 252 observations (children with parents whom have HIV) and 123 variables.

I will limit the large number of available variables and focus of the following two for the remainder of this project; whether the parent bonding scale (the likely-hood that parents were over-protective of their children) had an impact on whether or not children played hookey and skipped school and for how long they did so. The goal of this report is to explore the data set and find a potential link between parental control and a child's instinct to "break free", act more rebellious and potential skip school.

NHOOKEY -> Number of days absent from school without a reason PB23 - Is over-protective of you

```
hiv <- read.table("C:/Users/sidd0/math130version2/data/PARHIV_081217.txt", header=TRUE, sep="\t")
```

```
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
```

```
library(ggplot2)
```

Univariate Analysis of Variables

1. Living with Parents

```
table(hiv$PB23)
```

```
##
##  1  2  3  4
## 53 42 77 78
```

```
hiv$PB23_fac <- factor(hiv$PB23, labels = c("Very Unlike", "Somewhat Unlike", "Somewhat Like", "Very Li

table(hiv$PB23_fac)
```

```
##
##     Very Unlike Somewhat Unlike   Somewhat Like       Very Like
##              53              42              77              78
```

The numbers (1, 2, 3, and 4) from the PB23 subset relate to how the individuals tested felt about the following question: How likely is it that your parents are over-protective of you?

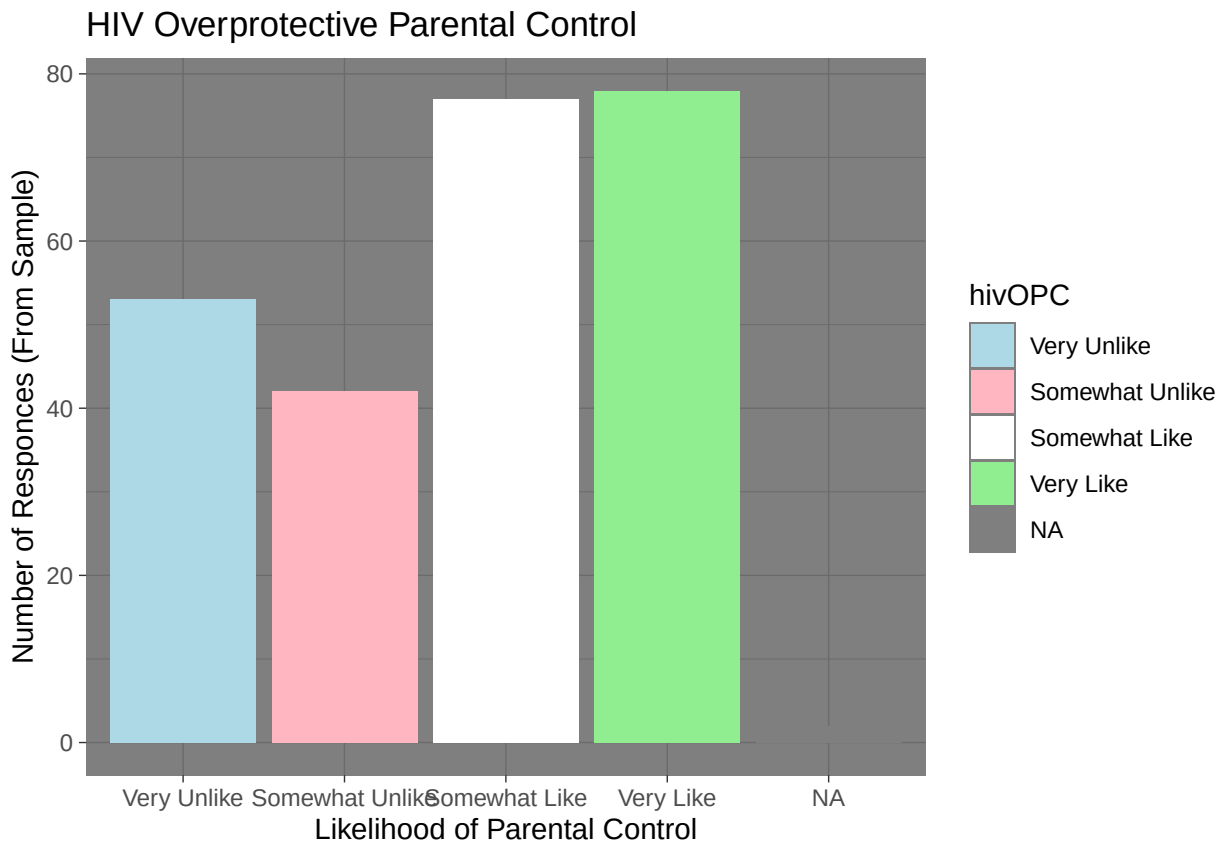
The numerical scale is not useful for charts, so I changed them to text-based likelihood statements.

```
hivOPC <- hiv$PB23_fac

table(hivOPC)
```

```
## hivOPC
##     Very Unlike Somewhat Unlike   Somewhat Like       Very Like
##              53              42              77              78
```

```
ggplot(hiv, aes(x=hivOPC, fill=hivOPC)) + geom_bar() + xlab("Likelihood of Parental Control") + ylab("Nu
```



I then took the data from the PB23 table and created the bar graph seen above. From the graph, we can tell that, of the children whom came from parents with HIV, the majority of them felt as though they were

being raised by overprotective parents. Of the 250 responses, 155 (over half) responded by saying that it felt "somewhat" and "very" overprotective parents, which can be seen illustrated via the white and green bars.

```
summary(hiv$NHOOKEY)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    1.00    2.27    4.00   15.00
```

```
table(hiv$NHOOKEY)
```

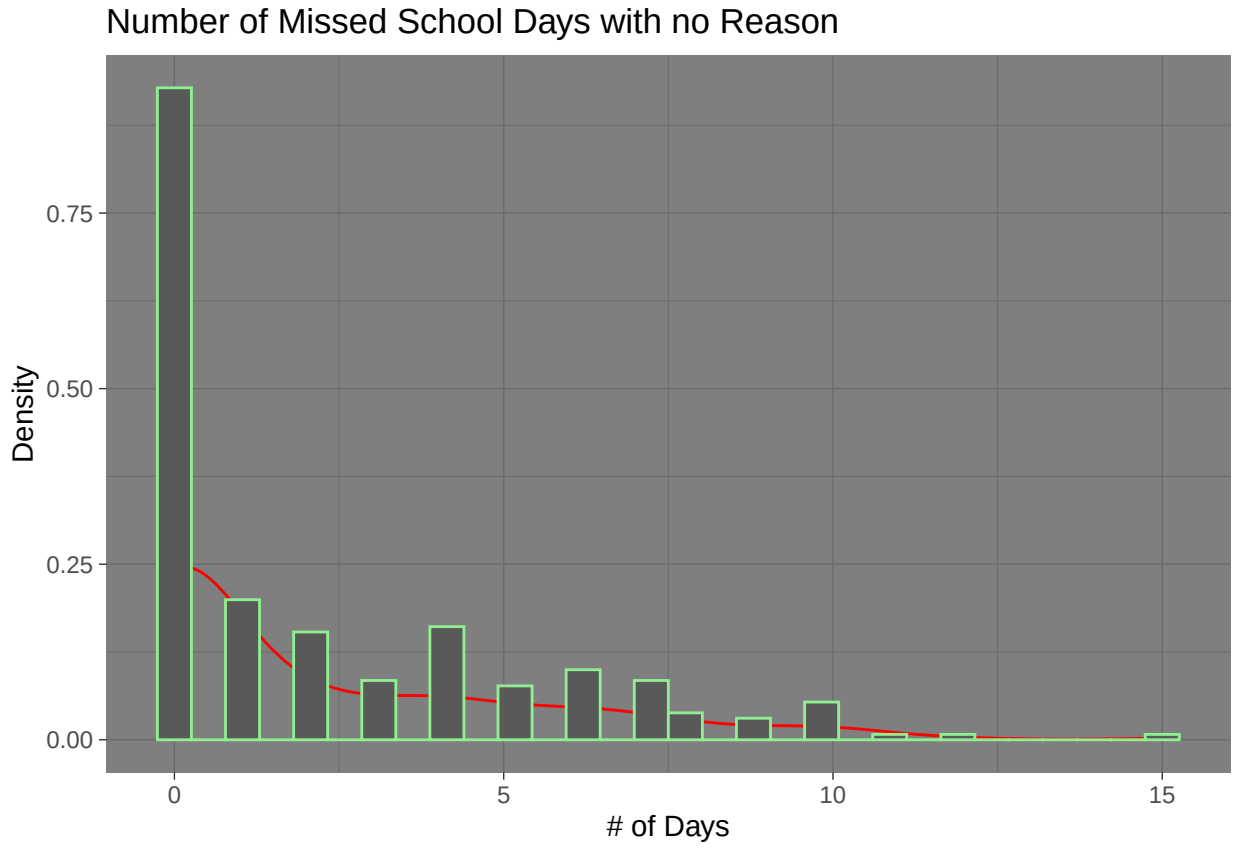
```
##
##   0   1   2   3   4   5   6   7   8   9  10  11  12  15
## 121  26  20  11  21  10  13  11   5   4   7   1   1   1
```

```
mean(hiv$NHOOKEY)
```

```
## [1] 2.269841
```

```
ggplot(hiv, aes(x=NHOOKEY)) + geom_density(col="red") + geom_histogram(aes(y=..density..), colour="ligh
```

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



The second variable of interest was the NHOOKEY variable, one that looked at the number of days kids, how had parents with HIV, skipped school without a reason. From the density plot that was constructed above, we can see that the most common number of days skipped is roughly 2.5 days. This can be confirmed by plugging this data subset into the mean function, which tells us the more accurate answer; 2.269 days.

Bivariate Analysis

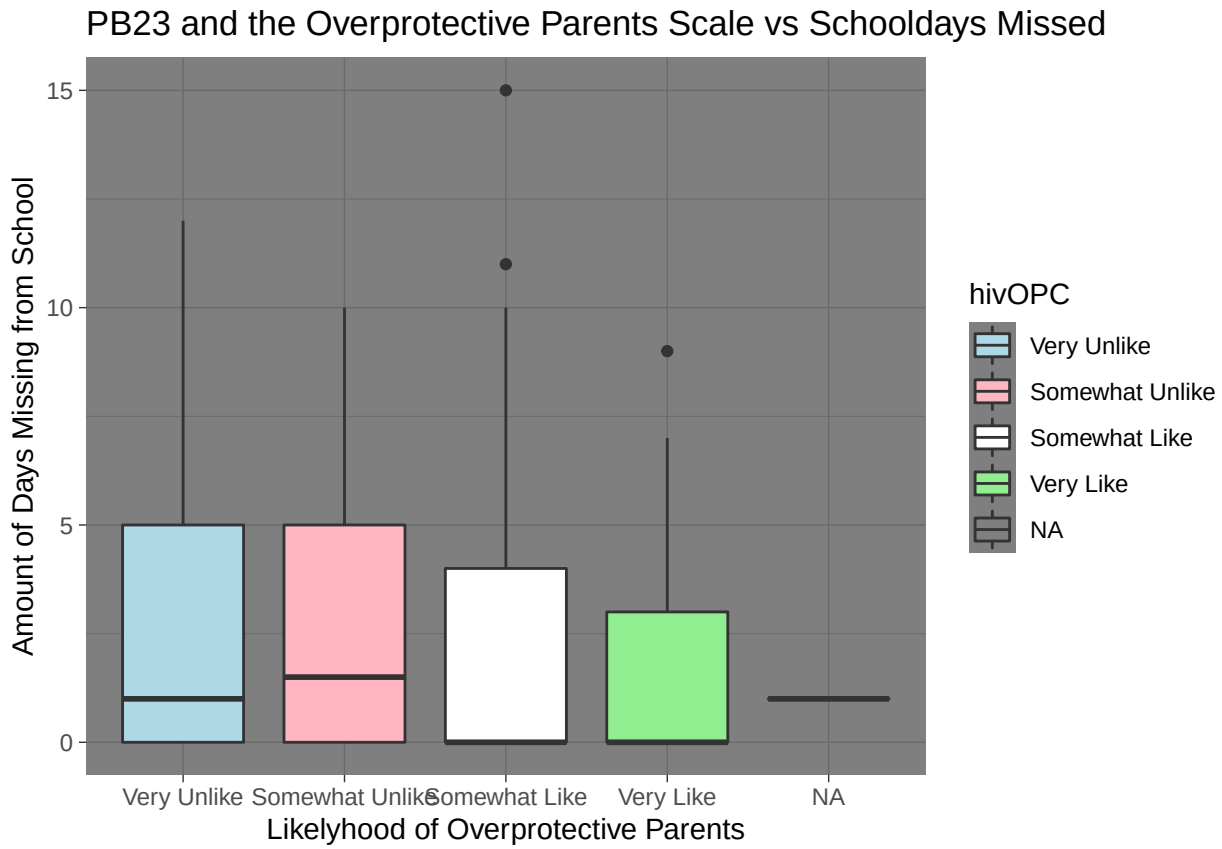
PB23 Vs NHOOKEY

The Likelyhood that a Child Skips School vs the Likelyhood that a Child has Overprotective Parents

```
table(hiv$NHOOKEY, hivOPC)
```

```
##     hivOPC
##      Very Unlike Somewhat Unlike Somewhat Like Very Like
##   0           23              16            40        42
##   1            4               5             7         8
##   2            5               7             4         4
##   3            1               1             4         5
##   4            5               1             3        12
##   5            2               3             4         1
##   6            3               2             4         4
##   7            6               2             2         1
##   8            1               1             3         0
##   9            1               1             1         1
##   10           1               3             3         0
##   11           0               0             1         0
##   12           1               0             0         0
##   15           0               0             1         0
```

```
ggplot(hiv, aes(x=hivOPC,y=NHOOKEY, fill=hivOPC))+theme_dark() +
geom_boxplot() + xlab("Likelyhood of Overprotective Parents")+ylab("Amount of Days Missing from School")
```



When comparing both subsets of data against each other, we can use a table and a box plot to better understand their relationship. From the colored box plot graph, we can see the bars (Overprotective Parents scale) are roughly the same size. On top of that, and over the bars themselves, we see very few outlines. Due to this, there is an indication that there is no relationship/correlation between the likely-hood of an overprotective parent and the number of days children skip school.

The outliers that are present, however, represent the scale level on which, on average, more days of school are skipped, which happen to be the "somewhat" overprotective and "very" overprotective levels.

Conclusion

From the report, and observation from the data tested, no relationship seemed to be found. While it seemed logical to think that children may want to "break free" from controlling parents, the hiv data set did not prove this to be true; for the children tested at least.