MATH 130 Exploratory Data Analysis Project

by Angel De Trinidad

Selected Data set: Parental HIV

For this exploratory data analysis project, I decided to utilize the **Parental HIV** data set. According to the site where the data was obtained, this study investigated the "behavioral interventions" of families with a parent of HIV. This study focused on the children and their environment. It looked at various variables such as ethnicity, neighboorhood environments, education, and more. This study was conducted by Dr. Mary Jane Rotheram-Borus at the University of California, Los Angeles. The variables I will be examining in this study are **ethnicity** and **neighborhood violence and crime**.

```
parhiv <- read.table("/cloud/project/PARHIV_081217.txt", header=TRUE, sep="\t")
```

These are the packages used in order to investigate and create this data analysis:

```
library(forcats)
library(ggplot2)
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(knitr)
library(markdown)
```

Univariate Descriptions

According to the Centers for Disease Control and Prevention, there are specefic ethnicities/races that are more at risk for contracting HIV. This includes **African Americans, Hispanic/Latino, Native American/Alaskan Native, Asians,** and **Native Hawaiians/other Pacific Islanders**. Becuase of this, I was curious in examining the variable **ethnicity (ETHN)** within the Parental HIV data. Although there are only three groupings within the ethnicity variable—Black, Hispanic/Latino(a), and Other—I was still interested to see the data.

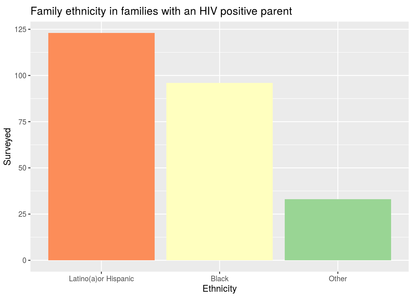
Due to the labeling of ethnicity categories (Black, Hispanic, etc.,) by numerical values, I renamed each category for ease of understanding (This was also done with the other variable examined).

```
parhiv$ETHNrename <- factor(parhiv$ETHN, labels=c( "Latino(a)or Hispanic", "Black","Other"))
```

```
summary(parhiv$ETHNrename)
```

```
## Latino(a)or Hispanic                Black                Other
##                  123                   96                   33
```

```
ggplot(parhiv, aes(x=ETHNrename, fill=ETHNrename)) + geom_bar() + xlab("Ethnicity") + scale_fill_brewer(palette=
"Spectral", guide=FALSE) + ylab("Surveyed") +ggtitle("Family ethnicity in families with an HIV positive parent")
```



Here we can see that **Latino/Hispanic** and **Black** ethnicities surveyed have a higher rate of HIV compared to other ethnicity categories. **"Latino(a)/Hispanic"** had most HIV positive parents (123 individuals), then **"Black"** (96 individuals), and lastly **"Other"** (33 individuals). This somewhat corresponds with the CDC's claims regarding HIV suceptibility based on ethnicity.

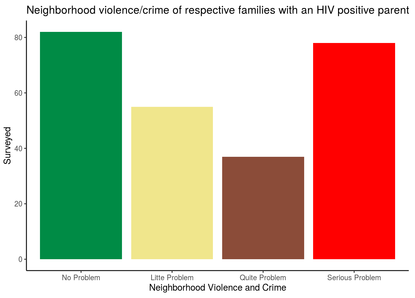
Next, I want to examine **neighborhood violence and crime**. Based on past courses, I know that cases of HIV are higher in areas of poverty and crime, thus I wanted to examine this variable to see how it compares to my previous knowledge.

```
parhiv$nbhrename <- factor(parhiv$NGHB6, labels=c( "No Problem", "Litte Problem","Quite Problem", "Serious Proble
m"))

summary(parhiv$nbhrename)
```

```
##       No Problem    Litte Problem   Quite Problem Serious Problem
##               82               55               37               78
```

```
ggplot(parhiv, aes(x=nbhrename, fill=nbhrename)) + geom_bar() + xlab("Neighborhood Violence and Crime") + scale_f
ill_manual(values=c("springgreen4", "khaki", "salmon4","red"), guide=FALSE) + ylab("Surveyed") +ggtitle("Neighbor
hood violence/crime of respective families with an HIV positive parent") + theme_classic()
```



This was very interesting, contrary to my previous belief, the data shows that the highest surveyed responses were **"No Problem"** and **"Serious Problem"**. So perhaps HIV does not affect individuals according to neighborhhood crime diffrerently like I previously assumed.

Bivariate Descriptions

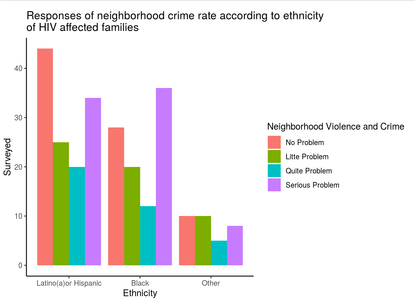
Now I would like to examine how ethnicity and neighborhood crime relate according to this data. Once again, due to previous knowledge, I know that People of Color (POC) ethnicities tend to be centered in areas of poverty and crime, thus I would like to examine these two variables in regards to eachother.

```
table(parhiv$ETHNrename, parhiv$nbhrename) %>% prop.table(margin=1) %>% round(3)
```

```
##
##                      No Problem Litte Problem Quite Problem Serious Problem
##   Latino(a)or Hispanic      0.358         0.203         0.163           0.276
##   Black                     0.292         0.208         0.125           0.375
##   Other                     0.303         0.303         0.152           0.242
```

Here we can see each ethnicity's response to how their neighborhood crime is. It will be easier visualize the responses in a grouped barchart.

```
ggplot(parhiv, aes(x=ETHNrename, fill=nbhrename)) + geom_bar(position = "dodge") + xlab("Ethnicity") +ylab("Surve
yed") + ggtitle("Responses of neighborhood crime rate according to ethnicity
of HIV affected families") + scale_fill_discrete(name="Neighborhood Violence and Crime")+ theme_classic()
```



This is a better visual of the two categorical variables. We can see that both **Latino(a)/Hispanic** and **Black ethnicities** either have significant crime and violence problems in their neighborhoods, or none at all. What is very apparrent is that **Other** ethnicities has lower data overall. This could be due to the study's sampling size, or it is a potential testiment to the difference in HIV susceptibility among ethnicity groups.

Conclusion

Overall, there are a wide variety of different variables that impact different group susceptibility to HIV. What was interesting about the data was the great scope of information it provided on family dynamics. There are so many variables I did not get to truly investigate in this data, but I really believe it really just speaks to the complexity of this disease.