EDA_username

Mianna Taylor

9/21/24`

##Introduction

For my data analysis project I am going to explore the HIV data set. The data set contains information on 252 children who have parents with HIV. The variables I choose for my project are the mother's job status, the mother's education level, and how old the children smoked.

Questions I wish to explore are if a lower job status is associated with more younger smoking age, and if education level also changes the odds of smoking younger.

Univariate Exploration

```
knitr::opts_chunk$set(fig.width=6, fig.height=4) # This sets all figure sizes in the document unless otherwise specified.
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)
```

Now I will load the data set into the global environment.

```
parHIV <- read.table("../data/parHIV.txt", header=TRUE, sep="\t")
```

Now I will use the use factor() to label categorical variables

```
parHIV$JOBMO_fac <- factor(parHIV$JOBMO, labels=c("Employed", "Unemployed","Disabled/Retired"))
table(parHIV$JOBMO_fac)
```

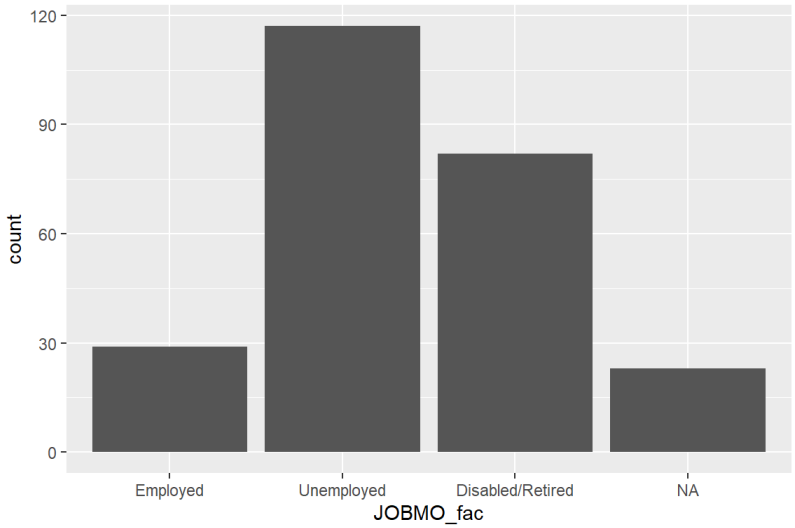
```
##
##         Employed       Unemployed Disabled/Retired
##               29              117               82
```

```
parHIV$EDUMO_fac <- factor(parHIV$EDUMO, labels=c("Did Not Complete High school","High school diploma/GED","More than High s
chool"))
table(parHIV$EDUMO_fac)
```

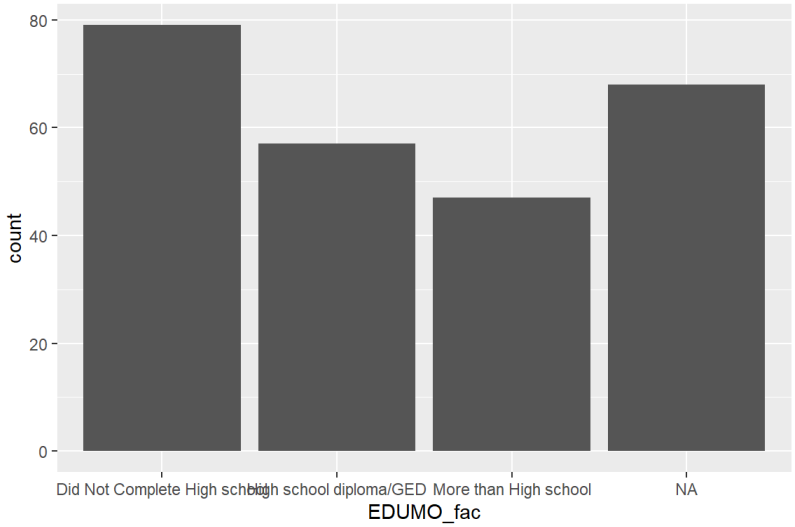
```
##
## Did Not Complete High school      High school diploma/GED
##                           79                           57
##        More than High school
##                           47
```

A barchart of the categorical variables:

```
ggplot(parHIV, aes(x=JOBMO_fac)) + geom_bar()
```




```
ggplot(parHIV, aes(x=EDUMO_fac)) + geom_bar()
```



Now I will do a Histogram for the continuous variable:

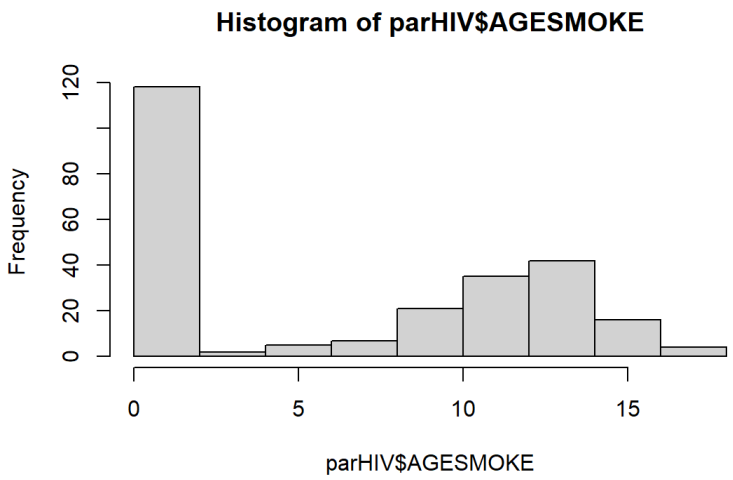
```
summary (parHIV$AGESMOKE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.000   6.500   6.292  12.000  17.000       1
```

```
sd(parHIV$AGESMOKE)
```

```
## [1] NA
```

```
hist(parHIV$AGESMOKE)
```



##Bivariate Exploration

Now I will try to answer my first question: Is a lower job status associated with a younger smoking age.

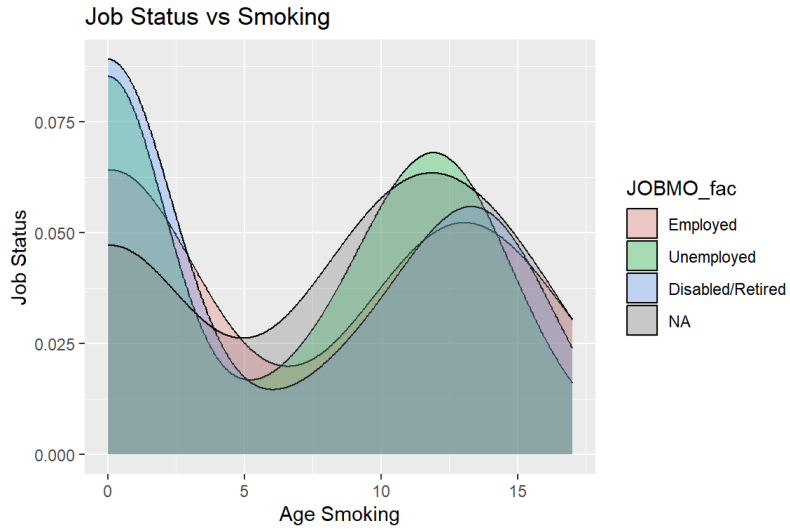
```
table(parHIV$JOBMO_fac,parHIV$AGESMOKE)
```

```
##
##                    0  4  5  6  7  8  9 10 11 12 13 14 15 16 17
##   Employed         14  1  0  0  0  0  0  1  1  5  1  2  2  1  1
##   Unemployed       52  1  1  2  2  3  3  8  9 11 12  6  3  0  3
##   Disabled/Retired 44  0  0  2  0  1  3  2  4  1  9  9  6  1  0
```

Now I will plot a density graph and show their association

```
ggplot(parHIV, aes(x=AGESMOKE, fill=JOBMO_fac)) + geom_density(alpha=.3)+ggtitle("Job Status vs Smoking") +
  ylab("Job Status") + xlab("Age Smoking")
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_density()`).
```



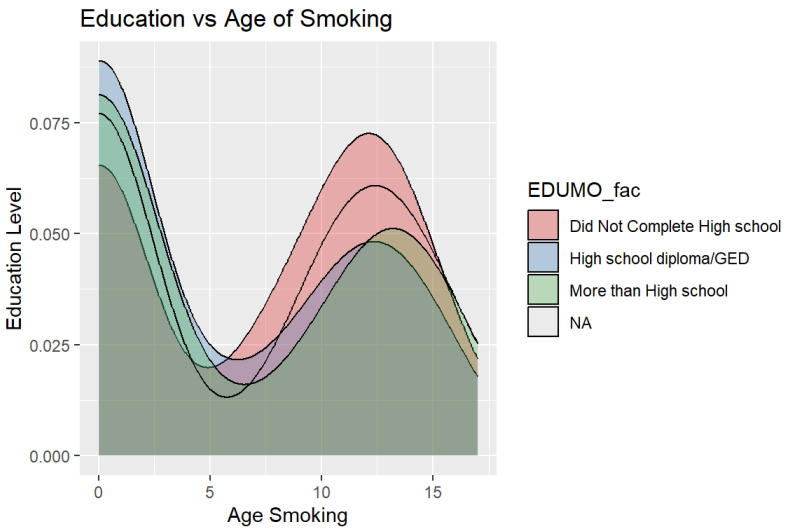
Now, I am going to answer my second question: Is education level a good indicator of how young someone smokes among children of parents with HIV?

```
table(parHIV$EDUMO_fac,parHIV$AGESMOKE)
```

```
##
##                               0  4  5  6  7  8  9 10 11 12 13 14 15 16 17
##   Did Not Complete High school 29  1  0  1  3  3  3  5  4 11  7  7  2  1  2
##   High school diploma/GED      30  1  1  2  0  0  2  3  3  3  4  3  3  1  0
##   More than High school        26  0  0  1  0  1  0  1  3  1  6  3  4  1  0
```

```
library(RColorBrewer)
ggplot(parHIV, aes(x=AGESMOKE, fill=EDUMO_fac)) + geom_density(alpha=.3)+ggtitle("Education vs Age of Smoking") +
  ylab("Education Level") + xlab("Age Smoking")+  scale_fill_brewer(palette="Set1")
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_density()`).
```



Conclusion

From my univariate exploration, I learned that most of the moms with HIV were unemployed and did not complete high school. From the histogram, it is clear that in early teens is a popular time to smoke. I was shocked by the mean value of children who smoke being 6.2, showing there is a skew towards the younger side.Second, in my bivariate exploration it is more interesting to see the continuous variable and categorical variable interact. The first density plot shows that those parents who were employed actually had children who smoked earlier. The other density plot shows that parents who did not complete high school had the highest level of children who smoked. These conclusions could be due to many outside factors along with these variables presented. It would be unfair to say all parents afflicted with HIV who have a lower job status and lower education level will have kids who smoke inherently. But it is safe to say that a child with parents who have lower education and lower job status are at a higher risk of smoking.