Project Proposal

Matthew Roland

2023-10-28

Data Preparation

I will be using 2022 National Health Interview Survey (NHIS) data collected by the CDC (https://www.cdc.gov/nchs/nhis/2022nhis.htm). The dataset in question hosts observations related adult physical and mental health, as well as healthcare utilization.

```
# load data

link <- "https://media.githubusercontent.com/media/Mattr5541/DATA_606_Final_Project/main/adult22.csv?to
nhis <- read.csv(url(link), stringsAsFactors=FALSE)

colnames(nhis)

##I will subset my variables of interest, including participant IDs, race, region (metropolitan vs. non

nhis_vars <- nhis %>% subset(select = c(HHX, RACEALLP_A, URBRRL, SEX_A, AGEP_A, MHRX_A, MHTHRPY_A, MHTH
```

Research question

My research question is as follows:

What is the impact of living in metropolitan regions (large or small) versus a rural region on levels of depression? Furthermore, does level of education affect levels of depression?

H0: There are no significant differences in depression levels among regions of residence

H0: There are no significant differences in depression levels among education levels

There are no significant differences in depression levels due to the interaction between region and education level

Cases

There are 27,651 observations and 637 variables. For my analysis, I will be using a subset of those variables

Data collection

The data were collected through in-person or phone-based interviews and surveys

Type of study

This study is observational in nature

Data Source

The data were collected from the cdc's 2022 NHIS adult survey: https://www.cdc.gov/nchs/nhis/2022nhis.htm

Dependent Variable

The dependent variable that I will be focusing on is quantitative in nature

Independent Variable(s)

The independent variables I will be using are qualitative in nature

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

Ages

```
unique(nhis_vars$AGEP_A)
```

```
##  [1] 85 64 37 72 84 74 31 80 42 81 68 69 24 57 78 59 27 58 38 83 46 43 19 32 25
## [26] 45 47 62 61 35 44 55 63 60 23 65 36 53 50 34 76 20 26 66 39 52 41 33 73 71
## [51] 40 21 67 29 18 30 77 49 82 51 54 48 56 22 75 79 70 28 97 99
```

```
range(nhis_vars$AGEP_A)
```

```
## [1] 18 99
```

```
age_groups <- nhis_vars %>% mutate(group = case_when(
  AGEP_A >= 18 & AGEP_A <= 25 ~ "18-25",
  AGEP_A >= 26 & AGEP_A <= 35 ~ "26-35",
  AGEP_A > 35 & AGEP_A <= 45 ~ "36-45",
    AGEP_A > 45 & AGEP_A <= 55 ~ "46-55",
    AGEP_A > 55 ~ "56+",
  TRUE ~ "Unknown"
))

age_groups %>% count(group) %>% mutate(prop = n / sum(n))
```

```
##   group     n       prop
## 1 18-25  2073 0.07497016
## 2 26-35  4097 0.14816824
## 3 36-45  4257 0.15395465
## 4 46-55  3855 0.13941630
## 5   56+ 13369 0.48349065
```

As we can see, the largest group in this study is composed of individuals 56 years of age or older. For the purposes of this analysis, however, I will limit my focus to the 26-35 age group, because I am interested in observing the trends among that group (it helps that I am within that age group)

```
nhis_vars <- nhis_vars %>% filter(AGEP_A >= 26 & AGEP_A < 36)

nhis_vars %>% count(AGEP_A) %>% mutate(prop = n / sum(n))
```

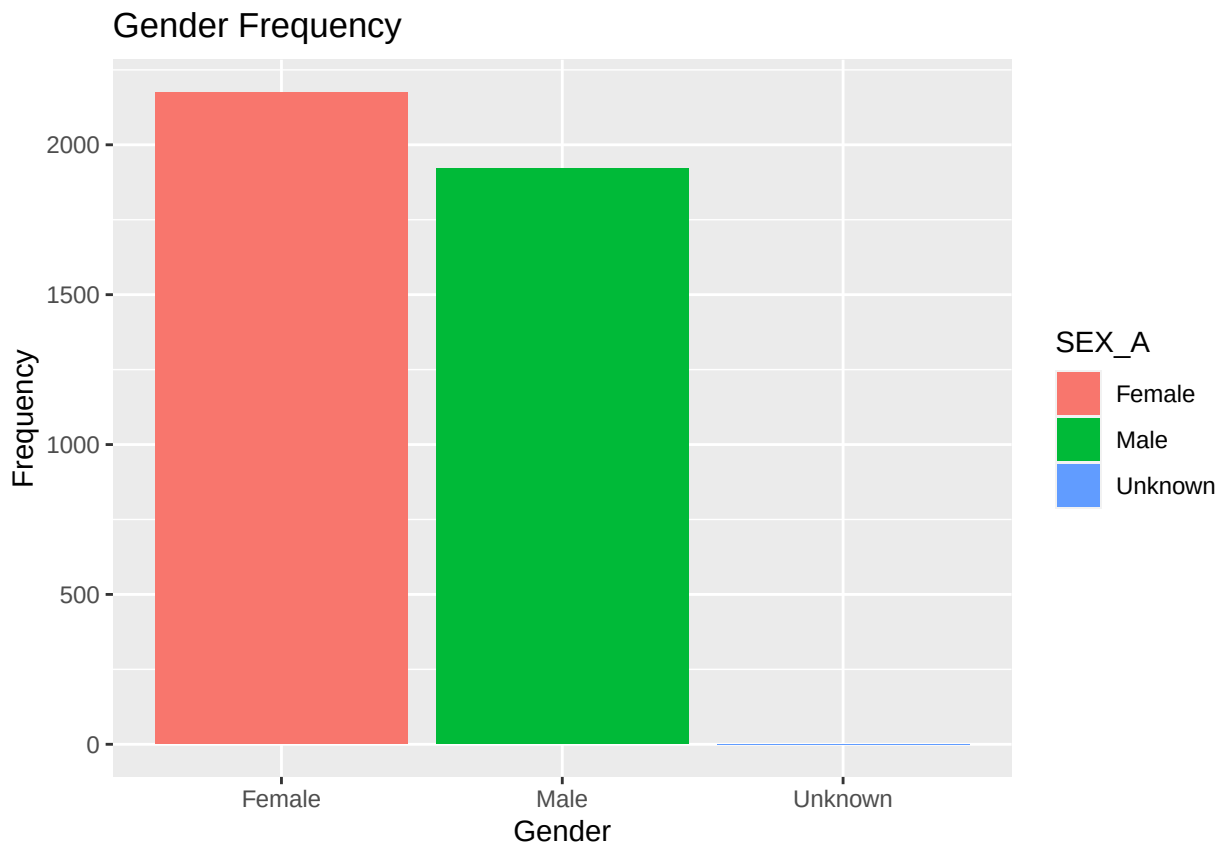
```
##    AGEP_A   n       prop
## 1      26 351 0.08567244
## 2      27 369 0.09006590
## 3      28 359 0.08762509
## 4      29 401 0.09787649
## 5      30 439 0.10715157
## 6      31 451 0.11008055
## 7      32 423 0.10324628
## 8      33 420 0.10251403
## 9      34 461 0.11252136
## 10     35 423 0.10324628
```

Gender Ratio

```
#Note, 1 = Male; 2 = Female; 9 = Don't Kn ow
nhis_gender <- nhis_vars %>% count(SEX_A) %>% mutate(prop = n / sum(n))

kable(nhis_gender)
```

SEX_A	n	prop
1	1921	0.4688797
2	2175	0.5308763
9	1	0.0002441

```
nhis_gender <- nhis_gender %>% mutate(SEX_A = recode(SEX_A, "1" = "Male", "2" = "Female", "9" = "Unknow

ggplot(nhis_gender, aes(SEX_A, n, fill = SEX_A)) + geom_bar(stat = "identity")  +  labs(title = "Gender
```



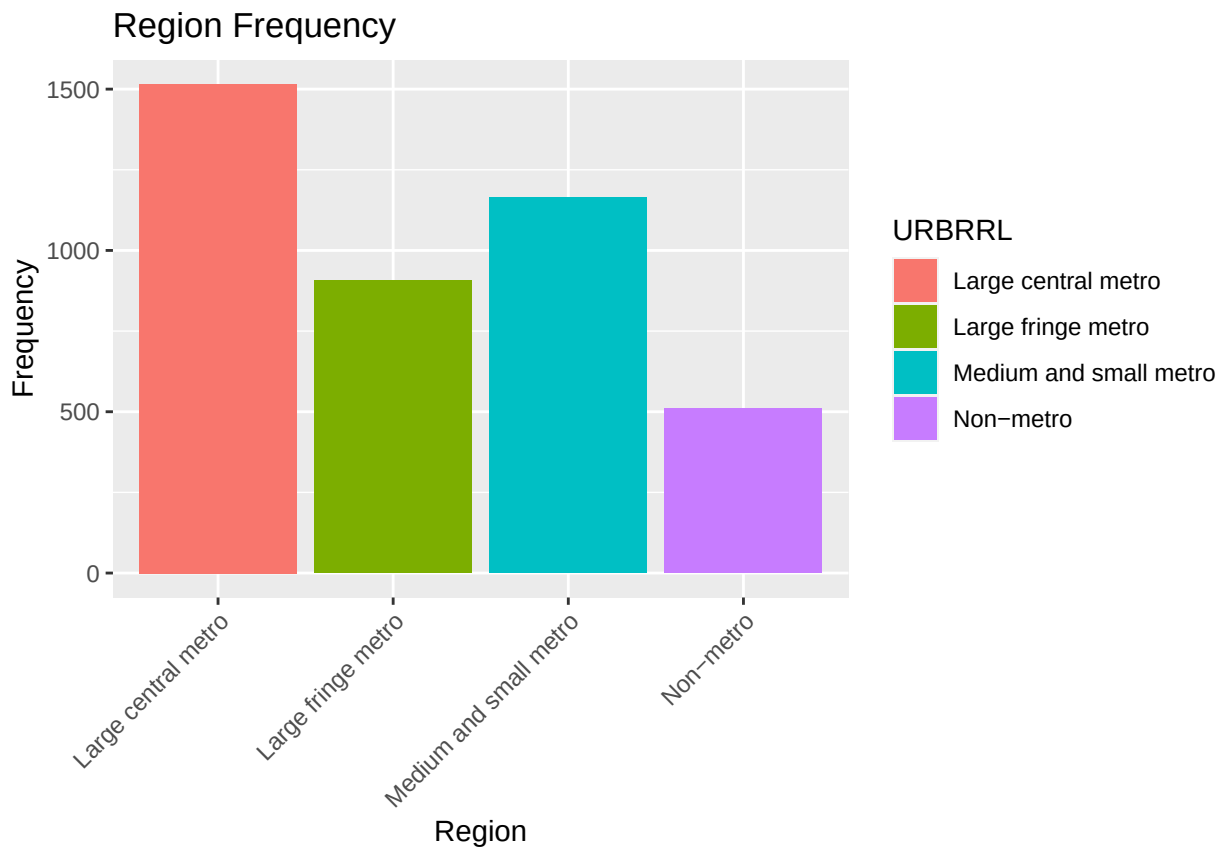
As we can see, the proportion of men to women is fairly similar, with slightly more women participants

Region Data

```
# Note: 1 = large central metro, 2 = large fringe metro, 3 = medium and small metro, and 4 = non-metrop
nhis_region <- nhis_vars %>% count(URBRRL) %>% mutate(prop = n / sum(n))

nhis_region <- nhis_region %>%
  mutate(URBRRL =
           recode(URBRRL,
                  "1" = "Large central metro",
                  "2" = "Large fringe metro",
                  "3" = "Medium and small metro",
                  "4" = "Non-metro"))

kable(nhis_region)
```

URBRRL	n	prop
Large central metro	1516	0.3700268
Large fringe metro	908	0.2216256
Medium and small metro	1163	0.2838662
Non-metro	510	0.1244813

```
ggplot(nhis_region, aes(URBRRL, n, fill = URBRRL)) + geom_bar(stat = "identity")  +  labs(title = "Regi
```



Most participants within this age range were drawn from metro areas, as opposed to non-metro areas

Education Levels

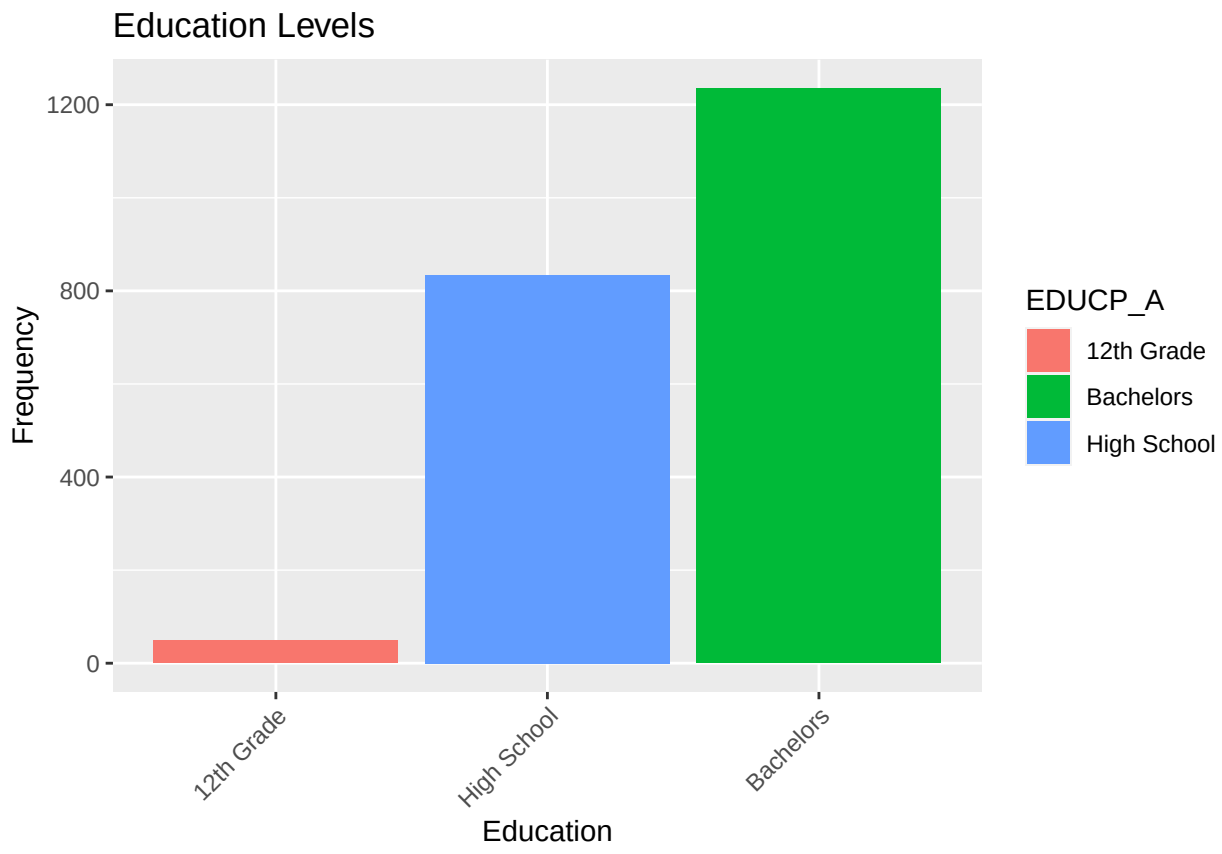
```
unique(nhis_vars$EDUCP_A)
```

```
##  [1]  8  4 10  9  2  5  1 97  6  7  3 99
```

```
#Note, I will only be using the following codes: 02 = 12th grade/no diploma, 04 = High School Graduate,
nhis_edu <- nhis_vars %>% count(EDUCP_A) %>% mutate(prop = n / sum(n))

nhis_edu <- nhis_edu %>% filter(EDUCP_A %in% c(2, 4, 8))

nhis_edu <- nhis_edu %>% mutate(EDUCP_A = recode(EDUCP_A, "2" = "12th Grade", "4" = "High School", "8" =

kable(nhis_edu)
```

EDUCP_A	n	prop
12th Grade	49	0.0119600
High School	834	0.2035636
Bachelors	1235	0.3014401

```
ggplot(nhis_edu, aes(reorder(EDUCP_A, n), n, fill = EDUCP_A)) + geom_bar(stat = "identity")  +  labs(ti
```



This shows us that most participants have at least a Bachelors

Depression Summary

It should be noted that depression was rated on a likert scale. However, it is not uncommon for such data to be treated as continuous/quantitative for the purposes of analysis

```
unique(nhis_vars$PHQCAT_A)
```

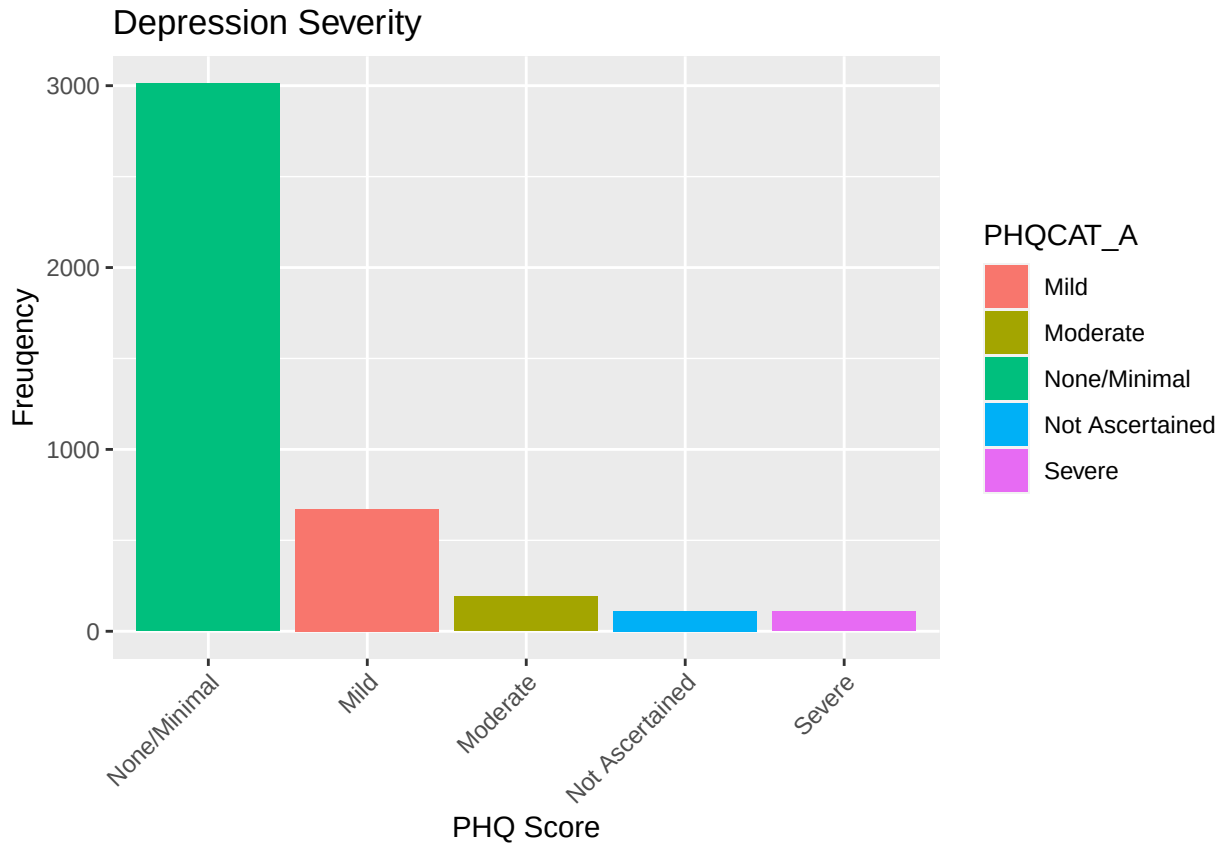
```
## [1] 1 2 4 8 3
```

```
#The PHQ is a measure of depression. The PHQCAT_A aggregates depression scores on each subscale of the
#Note: 1 = None/Minimal, 2 = Mild, 3 = Moderate, 4 = Severe, 8 = Not Ascertained
nhis_dep <- nhis_vars %>% count(PHQCAT_A) %>% mutate(prop = n / sum(n))
```

```r
nhis_dep <- nhis_dep %>%
  mutate(PHQCAT_A =
           recode(PHQCAT_A,
                  "1" = "None/Minimal",
                  "2" = "Mild",
                  "3" = "Moderate",
                  "4" = "Severe",
                  "8" = "Not Ascertained"))

kable(nhis_dep)
```

PHQCAT_A	n	prop
None/Minimal	3011	0.7349280
Mild	673	0.1642665
Moderate	192	0.0468636
Severe	109	0.0266048
Not Ascertained	112	0.0273371

```r
ggplot(nhis_dep, aes(reorder(PHQCAT_A, -n), n, fill = PHQCAT_A)) + geom_bar(stat = "identity")  +  labs
```



As one would expect of a random sample reflexive of the population, most individuals sampled are considered to not have abnormally high depression scores. Finding the mean and median may also be helpful in understanding levels of depression in this sample:

```r
nhis_vars %>% summarize(mean = mean(PHQCAT_A))
```

```
##       mean
## 1 1.529168
```

```r
nhis_vars %>% summarize(median = median(PHQCAT_A))
```

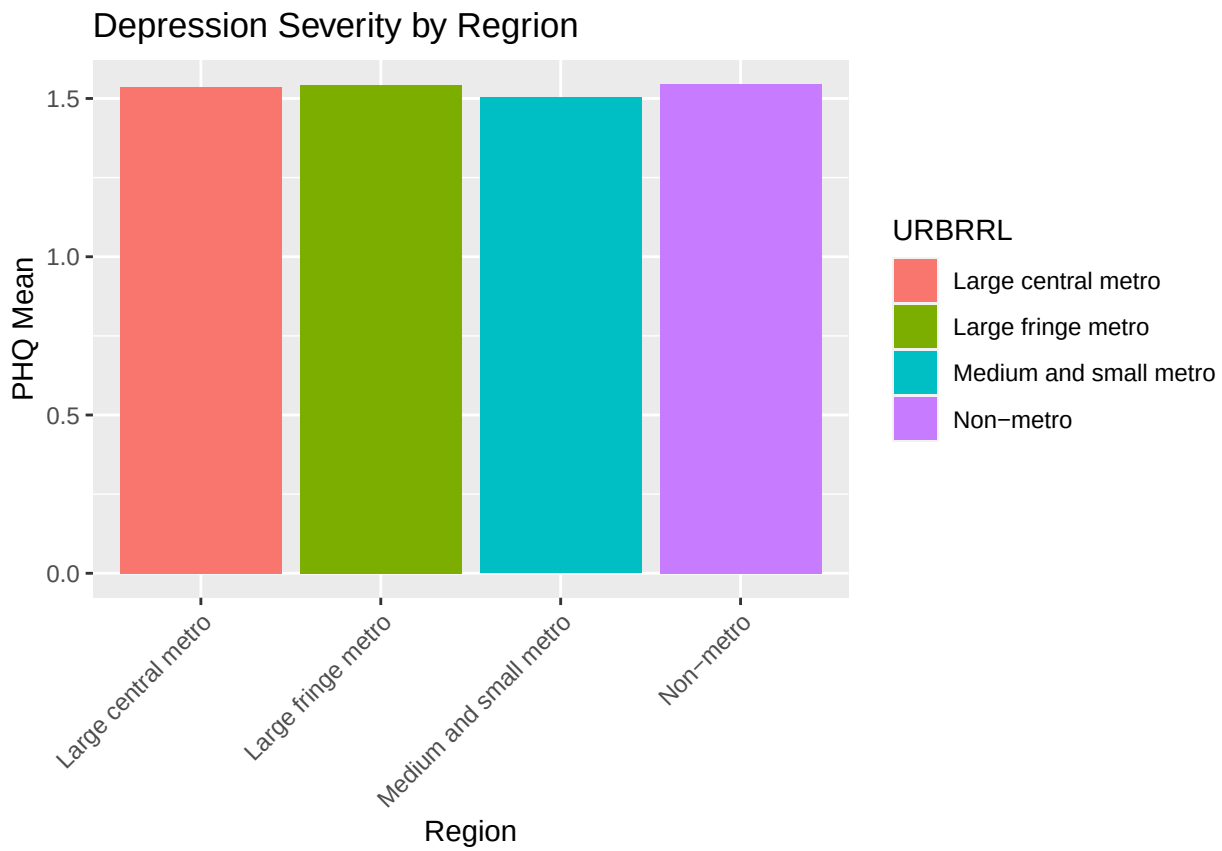
```
##   median
## 1      1
```

Depression by Region

```r
nhis_dep_regrion <- nhis_vars %>% group_by(URBRRL) %>% summarize(mean = mean(PHQCAT_A))

nhis_dep_regrion <- nhis_dep_regrion %>%
  mutate(URBRRL =
           recode(URBRRL,
                  "1" = "Large central metro",
                  "2" = "Large fringe metro",
                  "3" = "Medium and small metro",
                  "4" = "Non-metro"))

kable(nhis_dep_regrion)
```

URBRRL	mean
Large central metro	1.536280
Large fringe metro	1.541850
Medium and small metro	1.503010
Non-metro	1.545098

```r
ggplot(nhis_dep_regrion, aes(URBRRL, mean, fill = URBRRL)) + geom_bar(stat = "identity")  +  labs(title
```



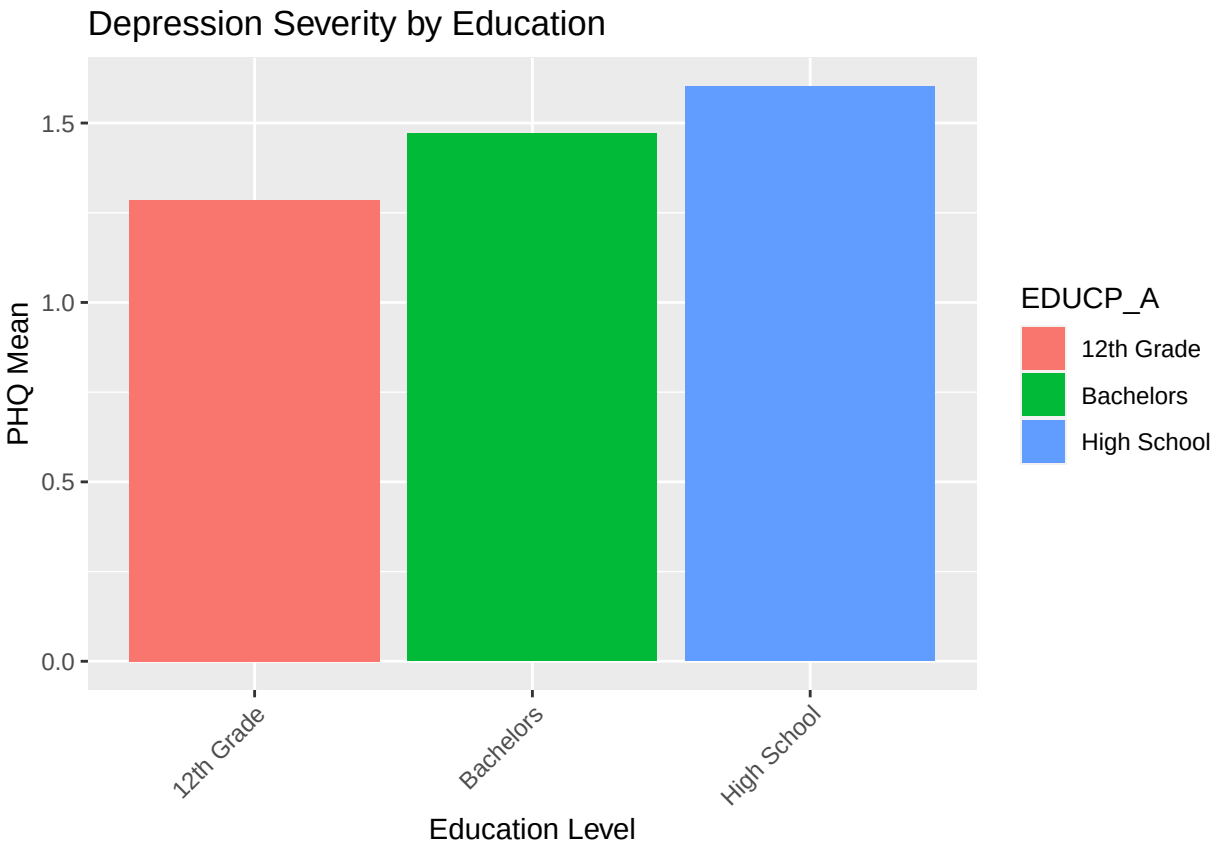
It would appear that the distribution representing depression levels by region is rather uniform, with slightly higher mean levels in Non-metro areas, and lower levels in medium and small metro areas

Depression by Education Level

```
nhis_edu_region <- nhis_vars %>% group_by(EDUCP_A) %>% summarize(mean = mean(PHQCAT_A))

nhis_edu_region <- nhis_edu_region %>% filter(EDUCP_A %in% c(2, 4, 8))

nhis_edu_region <- nhis_edu_region %>% mutate(EDUCP_A = recode(EDUCP_A, "2" = "12th Grade", "4" = "High

kable(nhis_edu_region)
```

EDUCP_A	mean
12th Grade	1.285714
High School	1.601919
Bachelors	1.470445

```
ggplot(nhis_edu_region, aes(EDUCP_A, mean, fill = EDUCP_A)) + geom_bar(stat = "identity")  +  labs(titl
```



These findings show that there are some mean differences in depression depending on one's level of education. Specifically, those with a high school education are more likely to have higher depression scores, followed by those with a Bacehlors, and surprisingly, followed by a 12th grade education.