# Analyses 4 (Lab 5)

*Anthony Hu*

*06/22/2021*

## IN THE CONTEXT OF YOUR FINAL PROJECT DATA:

### Data Import and Merging

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.2
```

```
## -- Attaching packages --------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
## Warning: package 'tibble' was built under R version 3.6.2
```

```
## Warning: package 'tidyr' was built under R version 3.6.2
```

```
## Warning: package 'readr' was built under R version 3.6.2
```

```
## Warning: package 'purrr' was built under R version 3.6.2
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
## Warning: package 'forcats' was built under R version 3.6.2
```

```
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 3.6.2
```

```
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 3.6.2
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 3.6.2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(interactions)
```

```
## Warning: package 'interactions' was built under R version 3.6.2
```

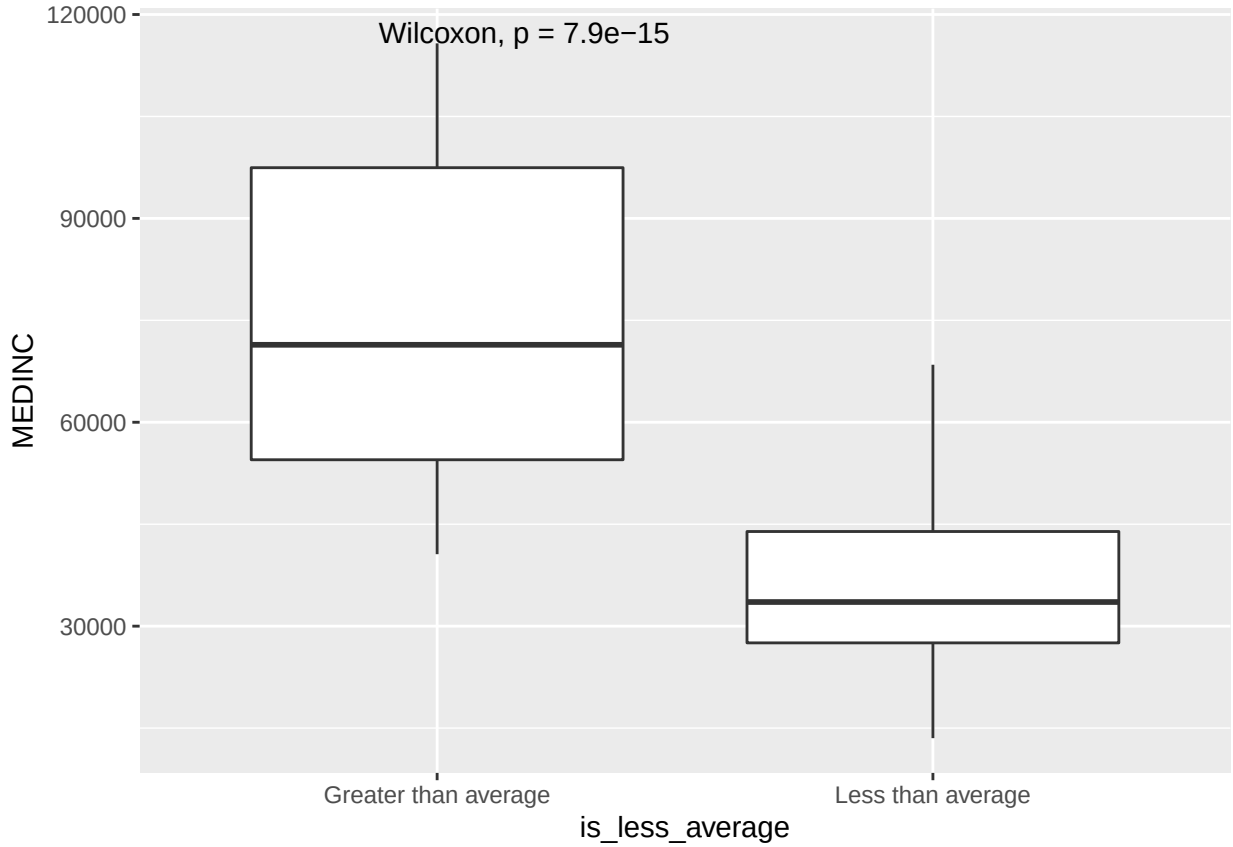
```
library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 3.6.2
```

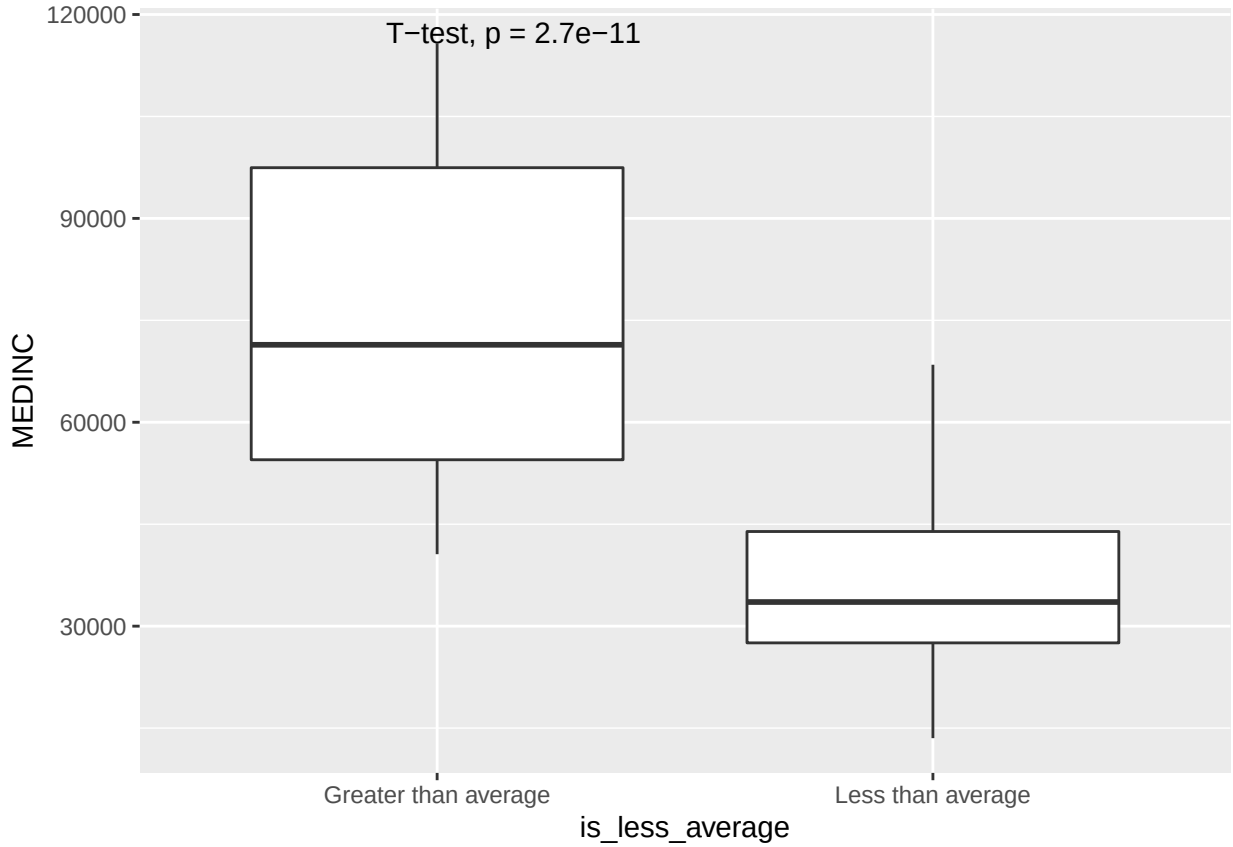
```
Chicago_census <- read.csv(url("https://datahub.cmap.illinois.gov/dataset/1d2dd970-f0a6-4736-96a1-3caeb

Chicago_socionomic <- read.csv(url("https://data.cityofchicago.org/api/views/kn9c-c2s2/rows.csv"))

Chicago_schools <- read.csv(url("https://data.cityofchicago.org/api/views/9xs2-f89t/rows.csv"))

Chicago_combined <- Chicago_socionomic %>%
  rename(GEOG = "COMMUNITY.AREA.NAME") %>%
  left_join(Chicago_census, by = "GEOG")
```

## Perform T-test

```
ttest <- Chicago_combined %>%
  filter(WHITE > 0) %>%
  filter(TOT_POP > 0) %>%
  mutate(white_pct = WHITE/TOT_POP) %>%
  mutate(is_less_average = ifelse(white_pct < mean(white_pct), "Less than average", "Greater than avera
  select(is_less_average, everything())


a <- ggplot(data = ttest, mapping = aes(x = is_less_average, y = MEDINC)) +
  geom_boxplot()
a + stat_compare_means()
```




```
a + stat_compare_means(method = "t.test")
```

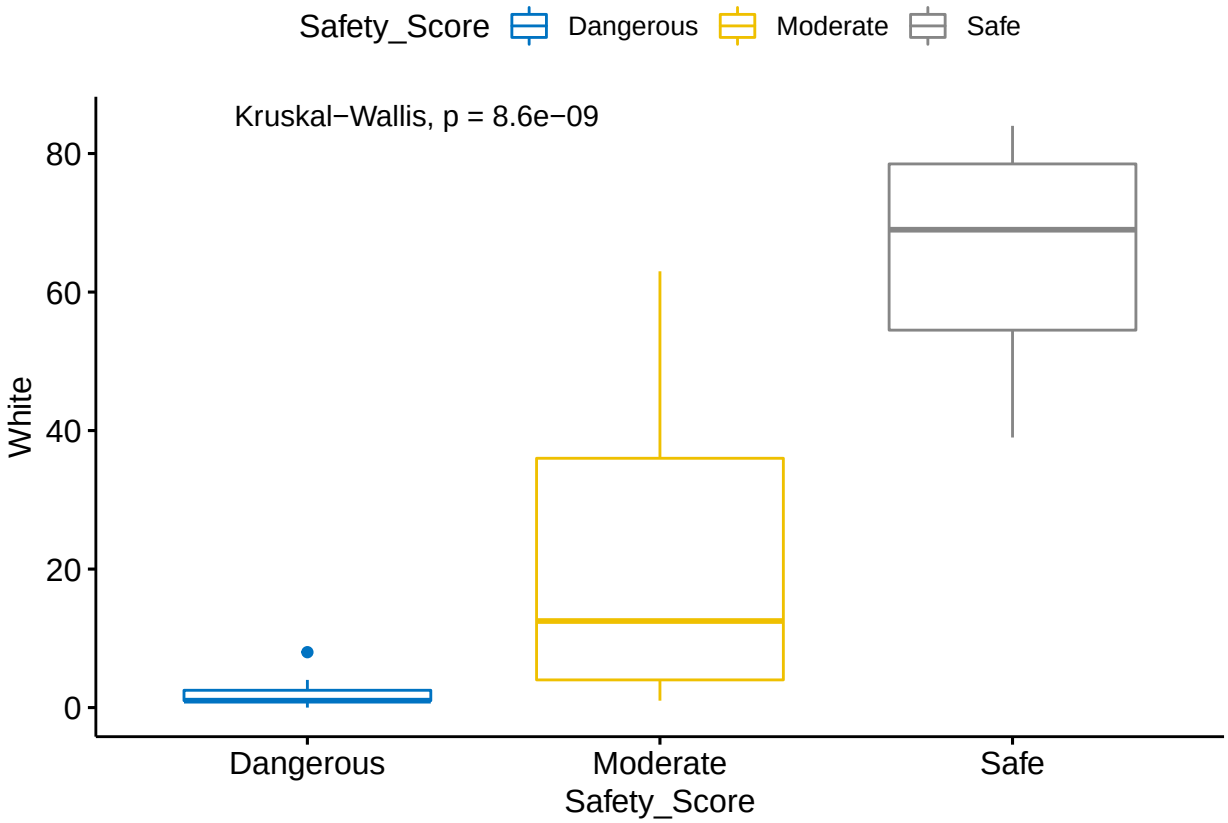



(INCLUDE VISUALIZATION) DESCRIBE OUTPUT: There is a statistically significant difference between median income of neighborhoods and the percentage of white citizens in neighborhoods.

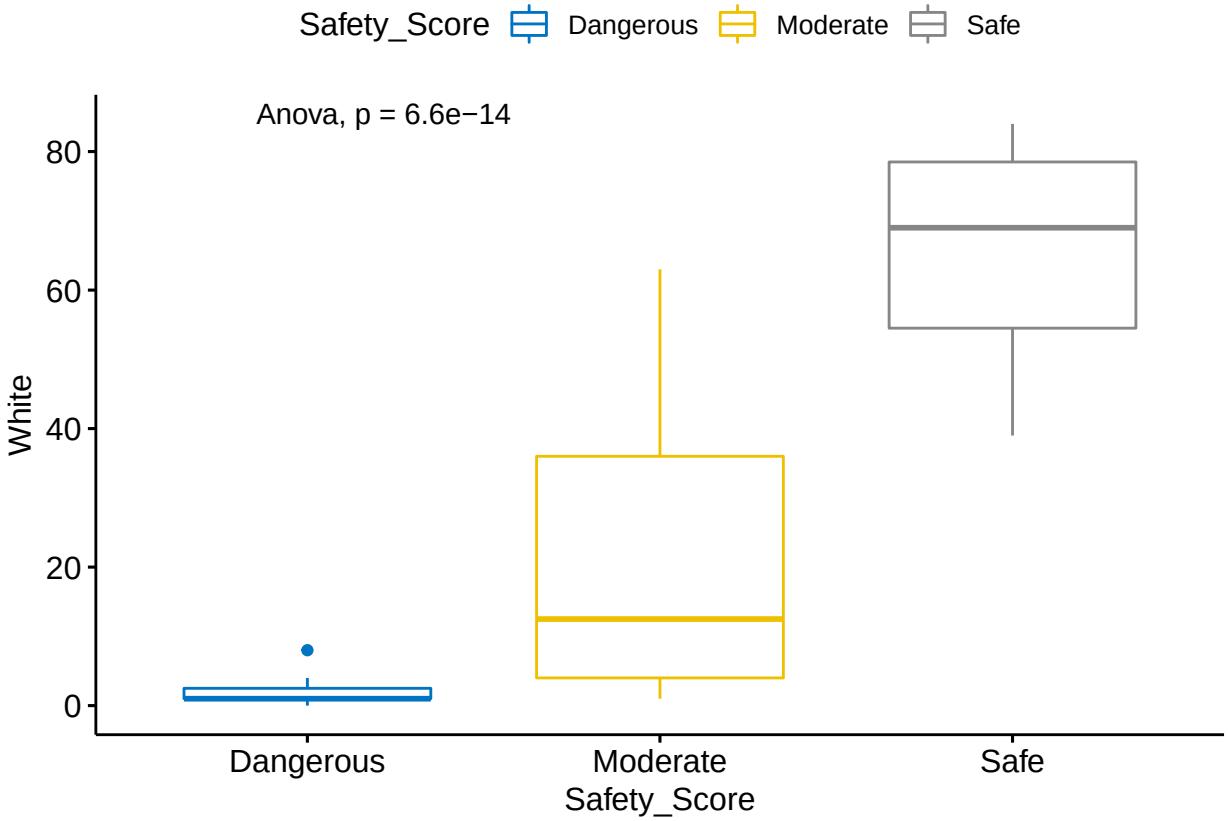
## Perform ANOVA(How does ethnicity affect safety?)

```
anova_alt_data_raw <- Chicago_schools %>%
  select(c( "Community.Area.Name","Safety.Score")) %>%
  transmute(GEOG = Community.Area.Name, Safety_Score = Safety.Score) %>%
  filter(!(Safety_Score == "NDA")) %>%
  group_by(GEOG) %>%
  summarise(Mean = mean(Safety_Score), n=n()) %>%
  select(c("GEOG", "Mean"))

anova_alt_dataa <- Chicago_combined %>%
  select(c("TOT_POP", "WHITE", "GEOG"))

anova_alt_data <- toupper(anova_alt_dataa$GEOG)

 anova_test_data1 <- cbind(anova_alt_dataa, anova_alt_data)
 anova_test_data <- anova_test_data1 %>%
   transmute(TOT_POP = TOT_POP, WHITE = WHITE, GEOG = anova_alt_data) %>%
   left_join(anova_alt_data_raw, by ="GEOG") %>%
   transmute(GEOG = GEOG, WHITE_percent = round(100*(WHITE / TOT_POP)), Safety = round(Mean, 1)) %>%
   drop_na()
```

```
anova_test_data2 <- cut(anova_test_data$Safety, breaks = c(0, 100/3, 200/3, 100),
                   labels = c("Dangerous", "Moderate", "Safe"))
anova_test_data3 <- cbind(anova_test_data, anova_test_data2)
anova_test_data4 <- anova_test_data3 %>%
  transmute(GEOG = GEOG, White = WHITE_percent, Safety_Score = anova_test_data2)


ggboxplot(anova_test_data4, x = "Safety_Score", y = "White",
          color = "Safety_Score", palette = "jco")+
  stat_compare_means()
```




```
# Change method to anova
ggboxplot(anova_test_data4, x = "Safety_Score", y = "White",
          color = "Safety_Score", palette = "jco")+
  stat_compare_means(method = "anova")
```




(INCLUDE VISUALIZATION) ] DESCRIBE OUTPUT: There is a statistically significant difference between the white percentage in neighborhoods that are considered dangerous, moderate, or safe. The anova test suggests that the neighborhoods with high white percentages are safer due to historical systems.

## Perform Multiple Regression (please try interaction terms, if possible.)

```
Q2 <- Chicago_combined %>%
  select(c("PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA",
           "GEOG", "PER.CAPITA.INCOME", "TOT_POP", "WHITE")) %>%
  mutate(Percent_Minorities = ((TOT_POP - WHITE)/TOT_POP) * 100,
         Income = PER.CAPITA.INCOME,
         Education = PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA)

lm_1 = lm(Income ~ Percent_Minorities * Education, data = Q2)
summary(lm_1)
```

```
##
## Call:
## lm(formula = Income ~ Percent_Minorities * Education, data = Q2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

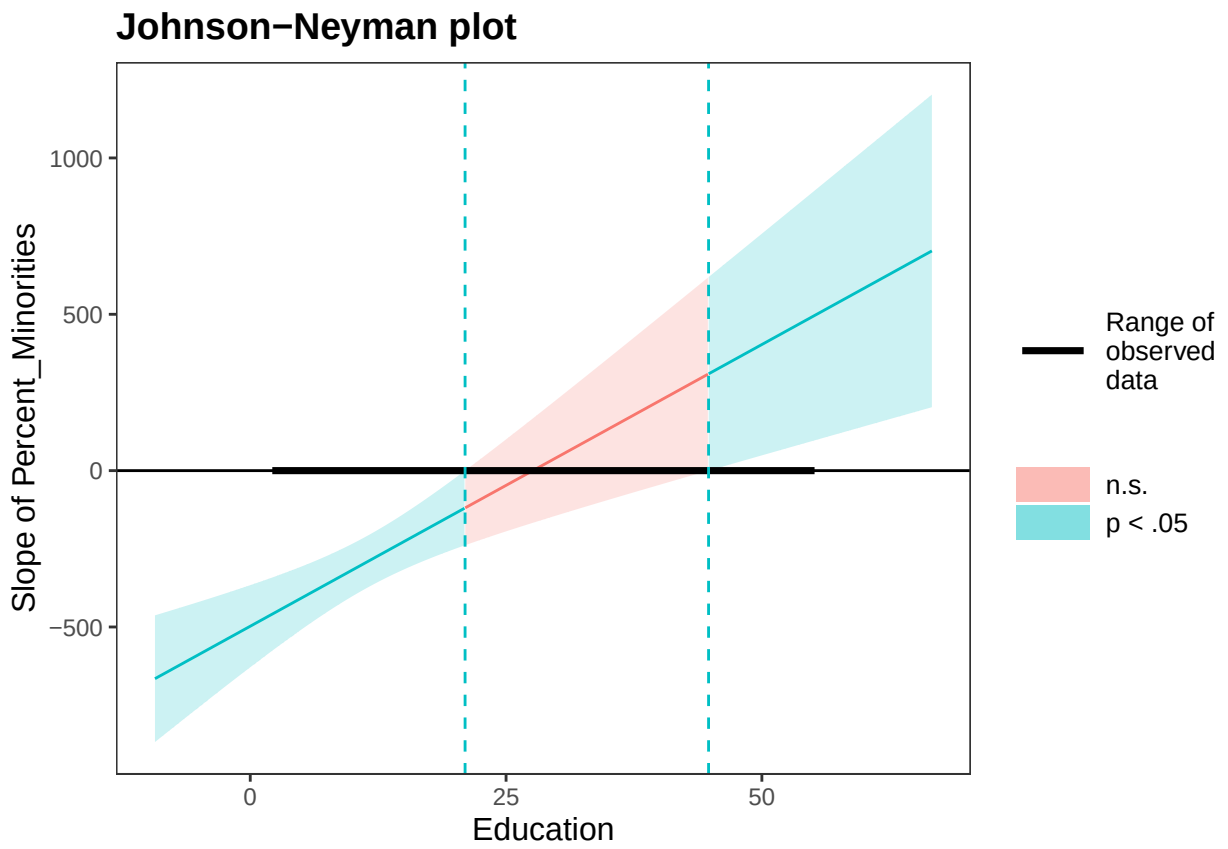
```
## -22538  -4069   -315   2764  34654
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  72137.592   4694.979  15.365  < 2e-16 ***
## Percent_Minorities            -497.555     65.790  -7.563 1.26e-10 ***
## Education                    -2001.213    387.362  -5.166 2.21e-06 ***
## Percent_Minorities:Education    18.018      4.475   4.027 0.000143 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7747 on 69 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.741,  Adjusted R-squared:  0.7297
## F-statistic: 65.79 on 3 and 69 DF,  p-value: < 2.2e-16
```

```
sim_slopes(lm_1, pred = Percent_Minorities, modx = Education, johnson_neyman =  FALSE)
```

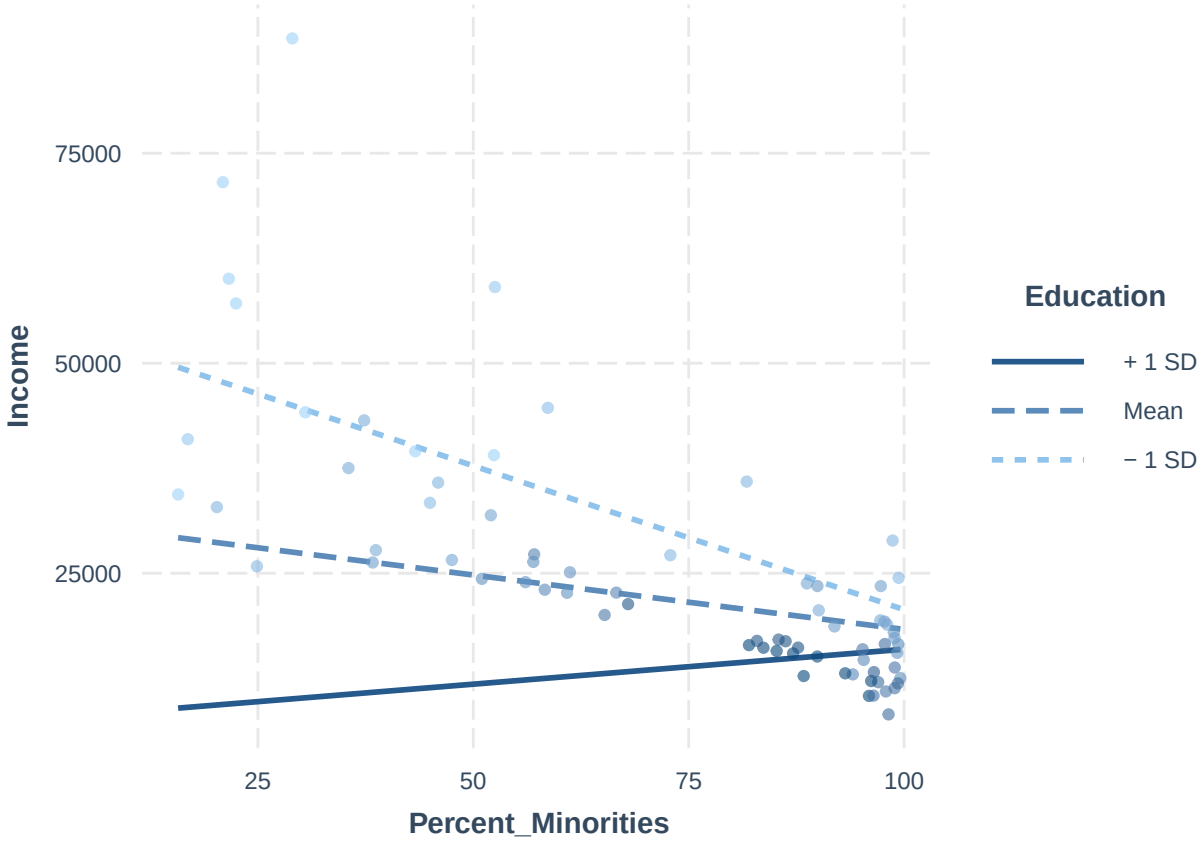
```
## SIMPLE SLOPES ANALYSIS
##
## Slope of Percent_Minorities when Education =  8.61 (- 1 SD):
##
##      Est.    S.E.   t val.      p
## --------- ------- -------- ------
##   -342.46   43.42    -7.89   0.00
##
## Slope of Percent_Minorities when Education = 20.42 (Mean):
##
##      Est.    S.E.   t val.      p
## --------- ------- -------- ------
##   -129.63   57.98    -2.24   0.03
##
## Slope of Percent_Minorities when Education = 32.23 (+ 1 SD):
##
##    Est.     S.E.   t val.      p
## ------- -------- -------- ------
##   83.20   102.11     0.81   0.42
```

```
johnson_neyman(lm_1, pred = Percent_Minorities, modx = Education, alpha = .05)
```

```
## JOHNSON-NEYMAN INTERVAL
##
## When Education is OUTSIDE the interval [20.99, 44.79], the slope of
## Percent_Minorities is p < .05.
##
## Note: The range of observed values of Education is [2.50, 54.80]
```



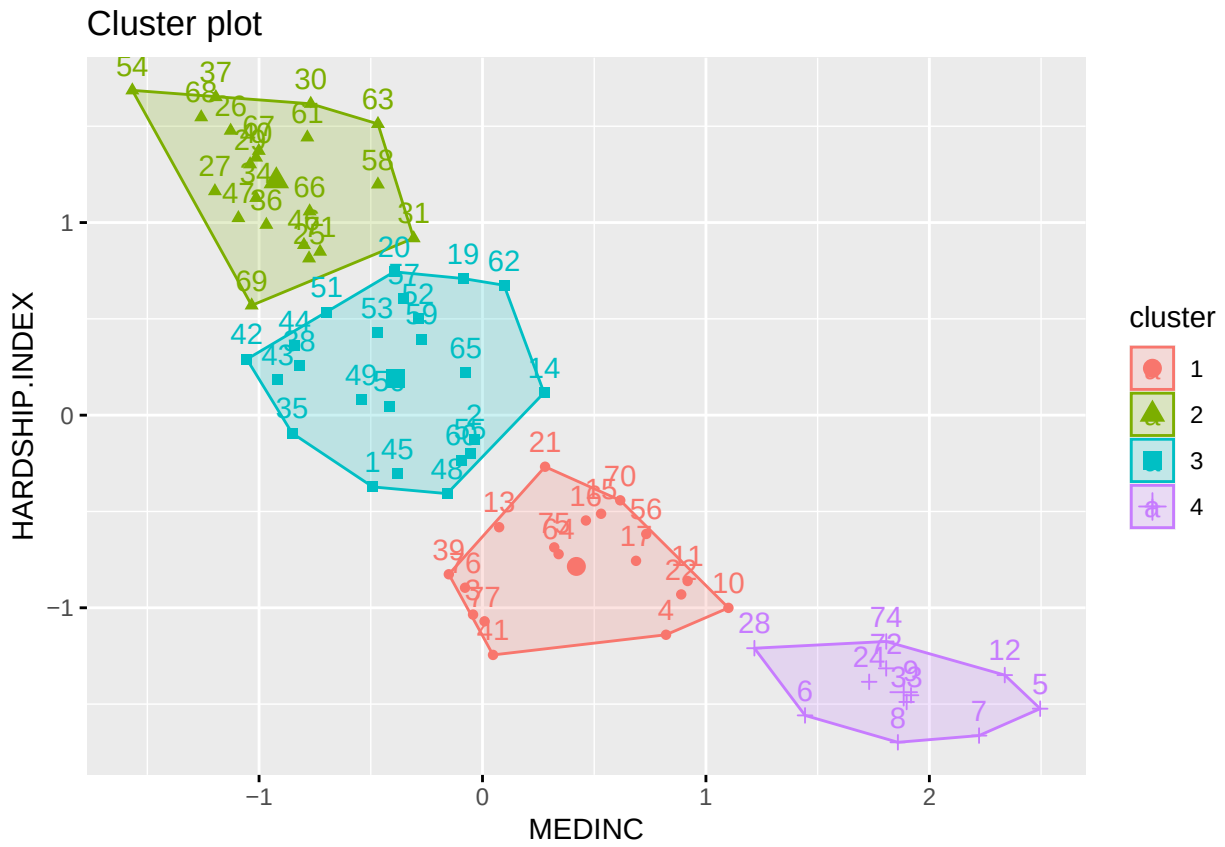

```
interact_plot(lm_1, pred = Percent_Minorities, modx = Education, plot.points = TRUE)
```




(INCLUDE VISUALIZATION) DESCRIBE OUTPUT: There is a statistically significant relationship between the Income and the Percent of Minorities in a Neighborhood + Percent of Education Level in that Neighborhood (% which completed high school). However, the Simple Slopes Analysis shows that the relationship between Education Level and Percent of Minorities in a neighborhood are not statistically significant within +- 1 standard deviation to each other. Additionally, the large range between the Slopes of Percent_Minorities and large Johnoson-Neyman interval indicates a weak/non-statistically significant relation between the education level and the number of minorities in a neighborhood.

## Perform Classification or Clustering (e.g., SVM, Decision Tree, K-Means Clustering)

```
# Number on graph is based on the order of the observation/neighborhood in Chicago-combined
D <- Chicago_combined %>%
  select(MEDINC, HARDSHIP.INDEX) %>%
  na.omit()

D1 <- scale(D)

D2 <- kmeans(D1, centers = 4, nstart = 50)


fviz_cluster(D2, data = D)
```




(INCLUDE VISUALIZATION) DESCRIBE OUTPUT: There seems to be a strong negative correlation between hardship index and median income. There also seems to be few, if any, outliers that contradict the relationship otherwise, and the observations appear to be clustered closely around 4 centers.