Final Project Milestone 1: Proposal

Adam Aleksic

Data Set

Data set: https://catalog.data.gov/dataset/racial-and-social-equity-composite-index-a44fc

Understanding the variables: https://data-seattlecitygis.opendata.arcgis.com/datasets/225a4c2c50e94f2cb548a046217f49f7_0?geometry=-122.509%2C47.574%2C-122.164%2C47.655

This data set examines linguistic, racial, ethnic, income, education, and health statistics for census tracts in Seattle.

I want to examine how the percent of English language learners in census tracts correlates with factors like obesity, poverty, education level, asthma, diabetes, and mental health.

New Research Question

Will census tracts with higher percentages of English language learners have health, income, and education disadvantages relative to census tracts with lower percentages of English language learners? In this study, I plan to use public information released by the city of Seattle to visualize the correlations between ELL status and certain health, education, and income variables. I hypothesize that, because ELL speakers tend to live in institutionally disadvantaged areas, census tracts with more ELL speakers will average higher incidences of asthma, obesity, diabetes, and mental health issues, and lower education levels and income statuses.

Analyses

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

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

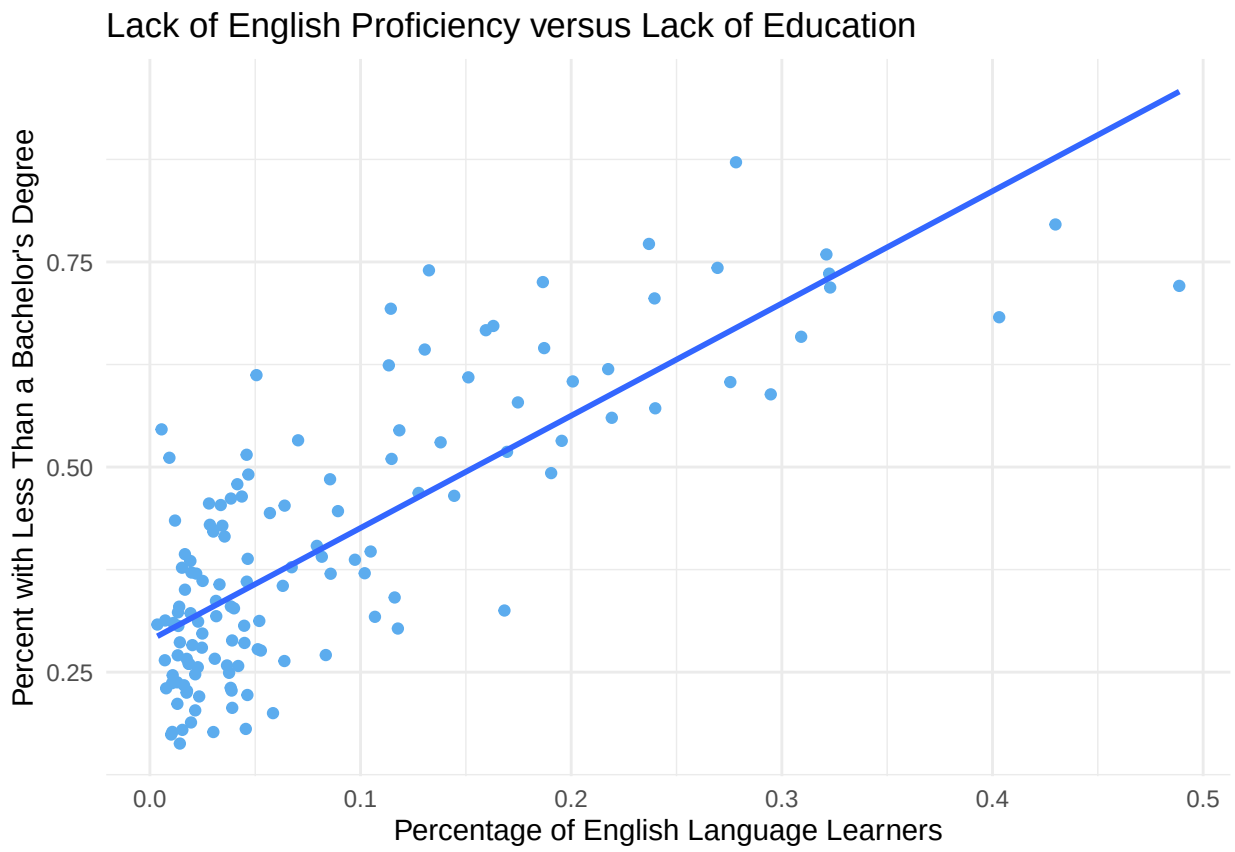
```
library(ggplot2)

seattle <- read.csv("~/Downloads/Gov/Racial_and_Social_Equity_Composite_Index.csv") %>%
 na.omit()

# setting up my data
```

```
education <- seattle %>%
ggplot(aes(x = PCT_ENGLISH_LESSTHAN_VERY_WELL, y = PCT_LESS_BACHELOR_DEGREE)) +
 geom_point(color = "steelblue2") +
 labs(title = "Lack of English Proficiency versus Lack of Education",
      x = "Percentage of English Language Learners",
      y = "Percent with Less Than a Bachelor's Degree") +
 theme_minimal() +
 geom_smooth(method = lm, se = FALSE)
education
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
education_fit <- lm(PCT_ENGLISH_LESSTHAN_VERY_WELL ~ PCT_LESS_BACHELOR_DEGREE, data = seattle)
education_fit_sum <- summary(education_fit)
education_fit_sum
```

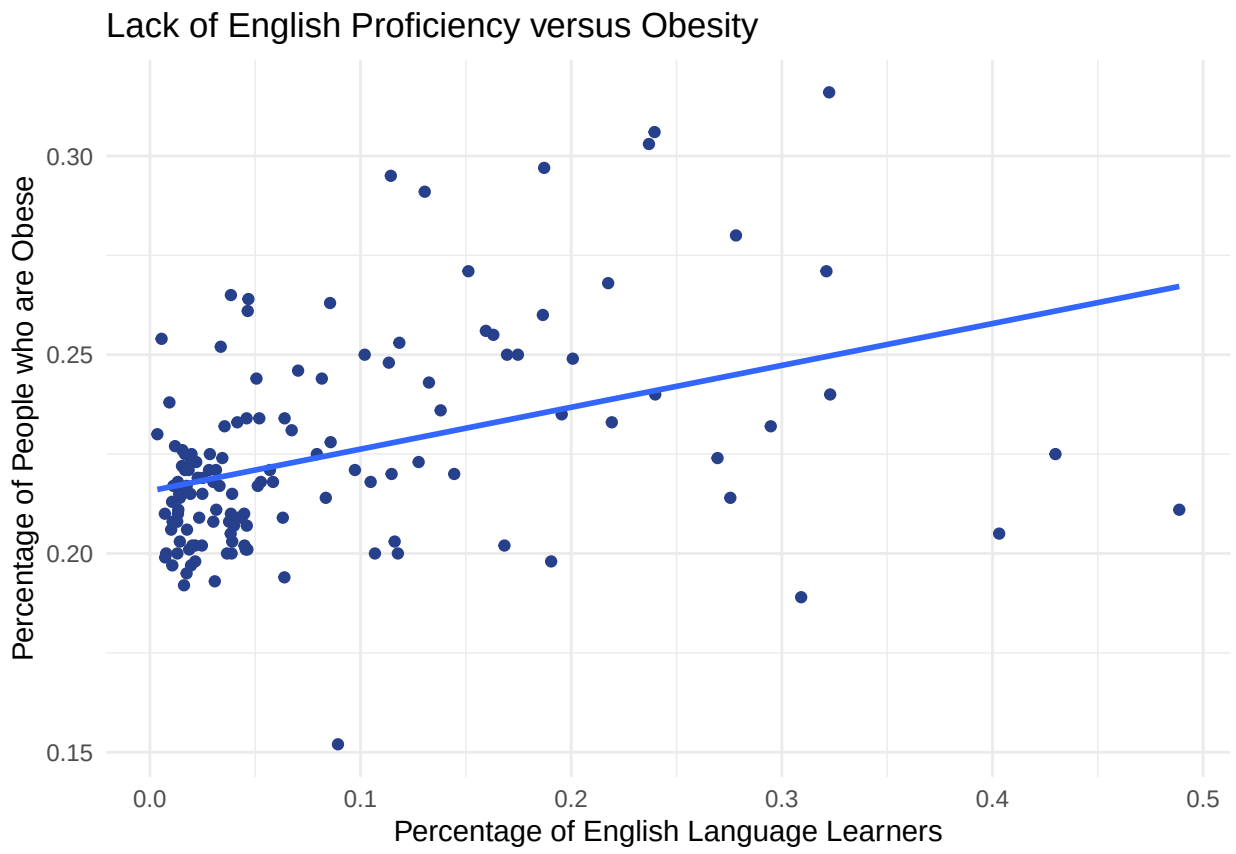
```
##
## Call:
## lm(formula = PCT_ENGLISH_LESSTHAN_VERY_WELL ~ PCT_LESS_BACHELOR_DEGREE,
##     data = seattle)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.144262 -0.037186 -0.001032  0.029943  0.259182
```

```
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -0.09899    0.01359  -7.284 2.62e-11 ***
## PCT_LESS_BACHELOR_DEGREE  0.45567    0.03081  14.789  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06005 on 132 degrees of freedom
## Multiple R-squared:  0.6236, Adjusted R-squared:  0.6208
## F-statistic: 218.7 on 1 and 132 DF,  p-value: < 2.2e-16
```

```
education_stats <- seattle %>%
  group_by(PCT_ENGLISH_LESSTHAN_VERY_WELL, PCT_LESS_BACHELOR_DEGREE) %>%
  summarise(bach_mean = mean(PCT_LESS_BACHELOR_DEGREE),
            bach_se = sd(PCT_LESS_BACHELOR_DEGREE) / sqrt(n()), .groups = "drop") %>%
  pivot_wider(names_from = PCT_ENGLISH_LESSTHAN_VERY_WELL, values_from = c(bach_mean, bach_se))
```

```
obese <- seattle %>%
ggplot(aes(x = PCT_ENGLISH_LESSTHAN_VERY_WELL, y = PCT_ADULT_OBESE)) +
 geom_point(color = "royalblue4") +
 labs(title = "Lack of English Proficiency versus Obesity",
      x = "Percentage of English Language Learners",
      y = "Percentage of People who are Obese") +
 theme_minimal() +
 geom_smooth(method = lm, se = FALSE)
obese
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# pretty standard scatterplot. I'm doing the same thing for all the other variables so I won't comment
```

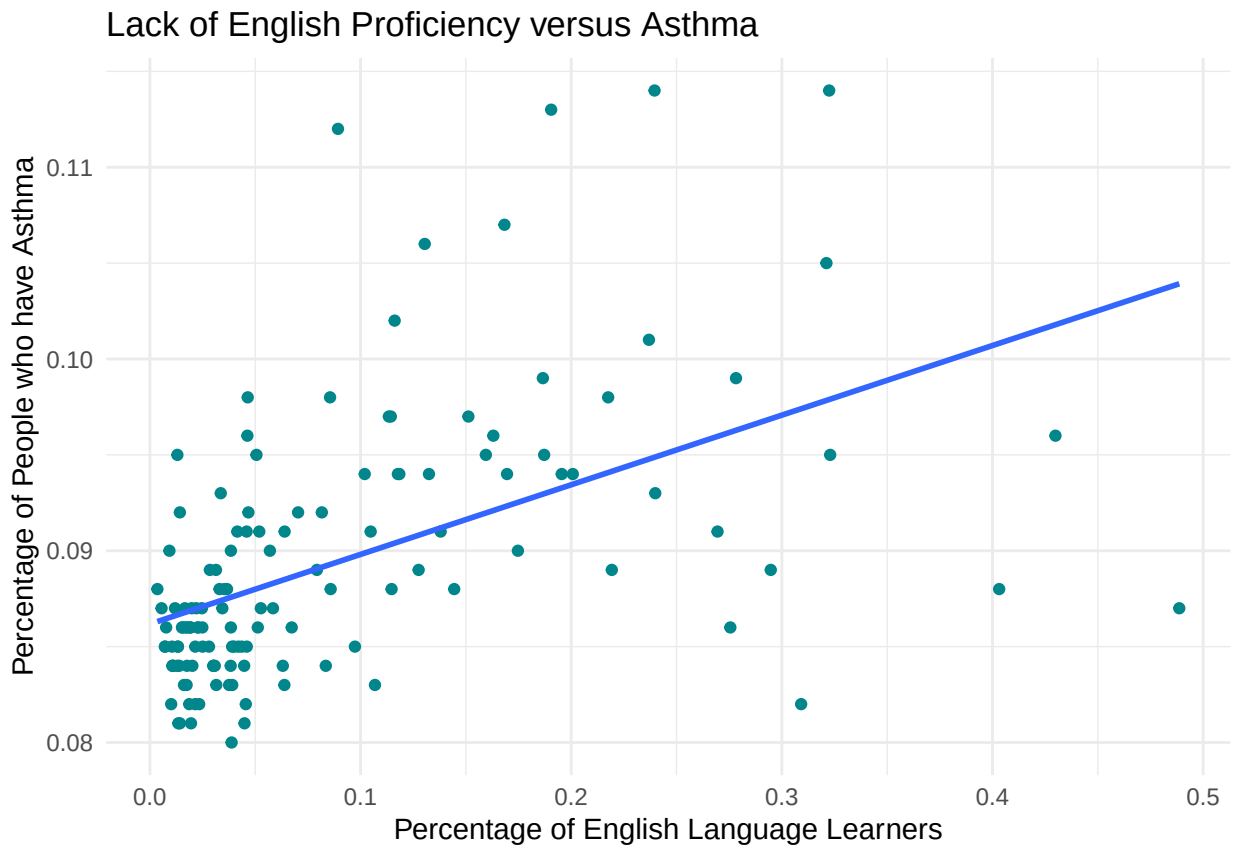
```
obese_fit <- lm(PCT_ADULT_OBESE ~ PCT_LESS_BACHELOR_DEGREE, data = seattle)
obese_fit_sum <- summary(obese_fit)
obese_fit_sum
```

```
##
## Call:
## lm(formula = PCT_ADULT_OBESE ~ PCT_LESS_BACHELOR_DEGREE, data = seattle)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.077061 -0.008218  0.000378  0.006703  0.055445
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.180514   0.004319   41.80   <2e-16 ***
## PCT_LESS_BACHELOR_DEGREE 0.108782   0.009793   11.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01908 on 132 degrees of freedom
## Multiple R-squared:  0.4832, Adjusted R-squared:  0.4792
## F-statistic: 123.4 on 1 and 132 DF,  p-value: < 2.2e-16
```

```r
# finding regression info
```

```r
asthma <- seattle %>%
ggplot(aes(x = PCT_ENGLISH_LESSTHAN_VERY_WELL, y = PCT_ADULT_WITH_ASTHMA)) +
 geom_point(color = "turquoise4") +
 labs(title = "Lack of English Proficiency versus Asthma",
      x = "Percentage of English Language Learners",
      y = "Percentage of People who have Asthma") +
 theme_minimal() +
 geom_smooth(method = lm, se = FALSE)
asthma
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```r
asthma_fit <- lm(PCT_ADULT_WITH_ASTHMA ~ PCT_LESS_BACHELOR_DEGREE, data = seattle)
asthma_fit_sum <- summary(asthma_fit)
asthma_fit_sum
```

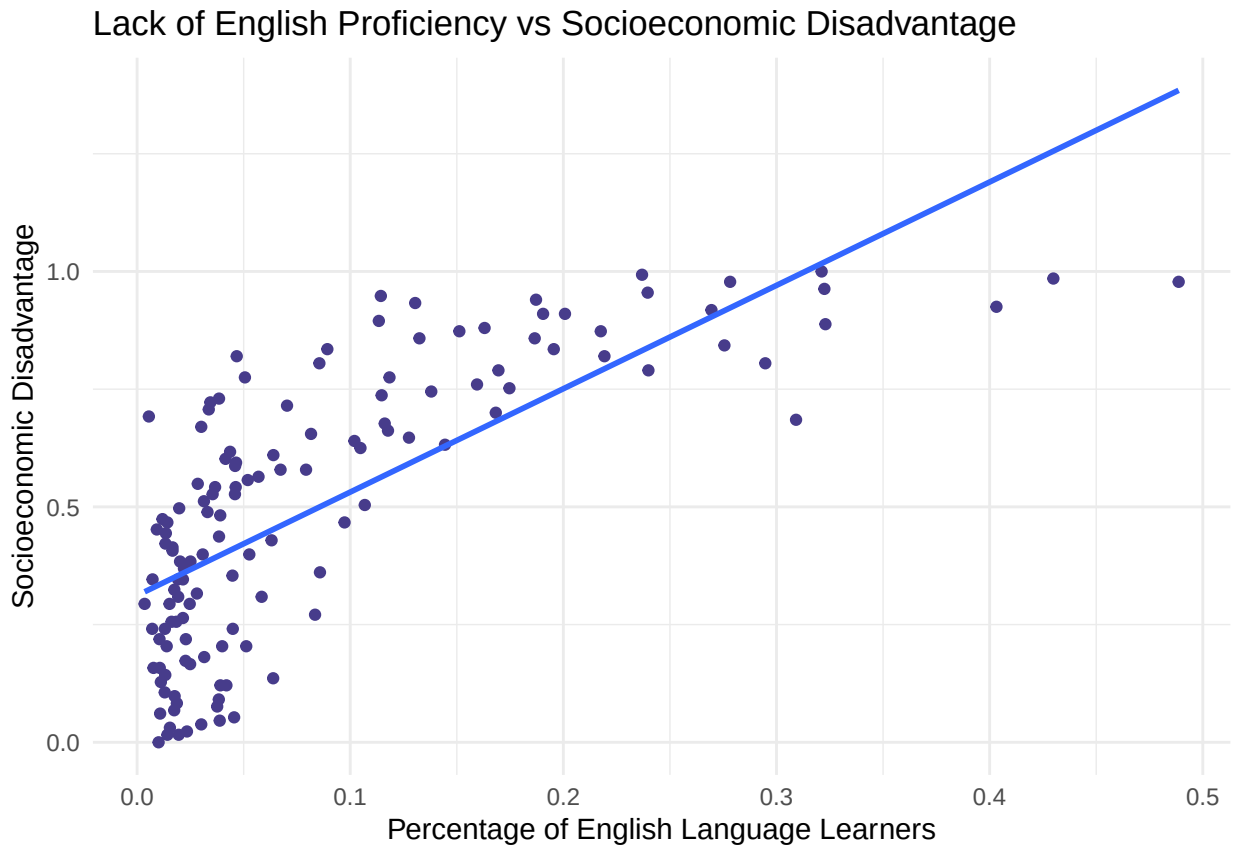
```
##
## Call:
## lm(formula = PCT_ADULT_WITH_ASTHMA ~ PCT_LESS_BACHELOR_DEGREE,
##     data = seattle)
##
## Residuals:
```

```
##        Min         1Q     Median         3Q        Max
## -0.0137371 -0.0027185 -0.0009653  0.0009999  0.0216858
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.078927   0.001234  63.943  < 2e-16 ***
## PCT_LESS_BACHELOR_DEGREE 0.025516   0.002799   9.117  1.1e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005454 on 132 degrees of freedom
## Multiple R-squared:  0.3864, Adjusted R-squared:  0.3817
## F-statistic: 83.12 on 1 and 132 DF,  p-value: 1.104e-15
```

```
# this one has the smallest r^2 out of the factors I've looked at. Clearly not going to be statistically
```

```
socioeconomic <- seattle %>%
ggplot(aes(x = PCT_ENGLISH_LESSTHAN_VERY_WELL, y = SOCIOECONOMIC_PERCENTILE)) +
 geom_point(color = "slateblue4") +
 labs(title = "Lack of English Proficiency vs Socioeconomic Disadvantage",
      x = "Percentage of English Language Learners",
      y = "Socioeconomic Disadvantage") +
 theme_minimal() +
 geom_smooth(method = lm, se = FALSE)
socioeconomic
```

```
## `geom_smooth()` using formula 'y ~ x'
```

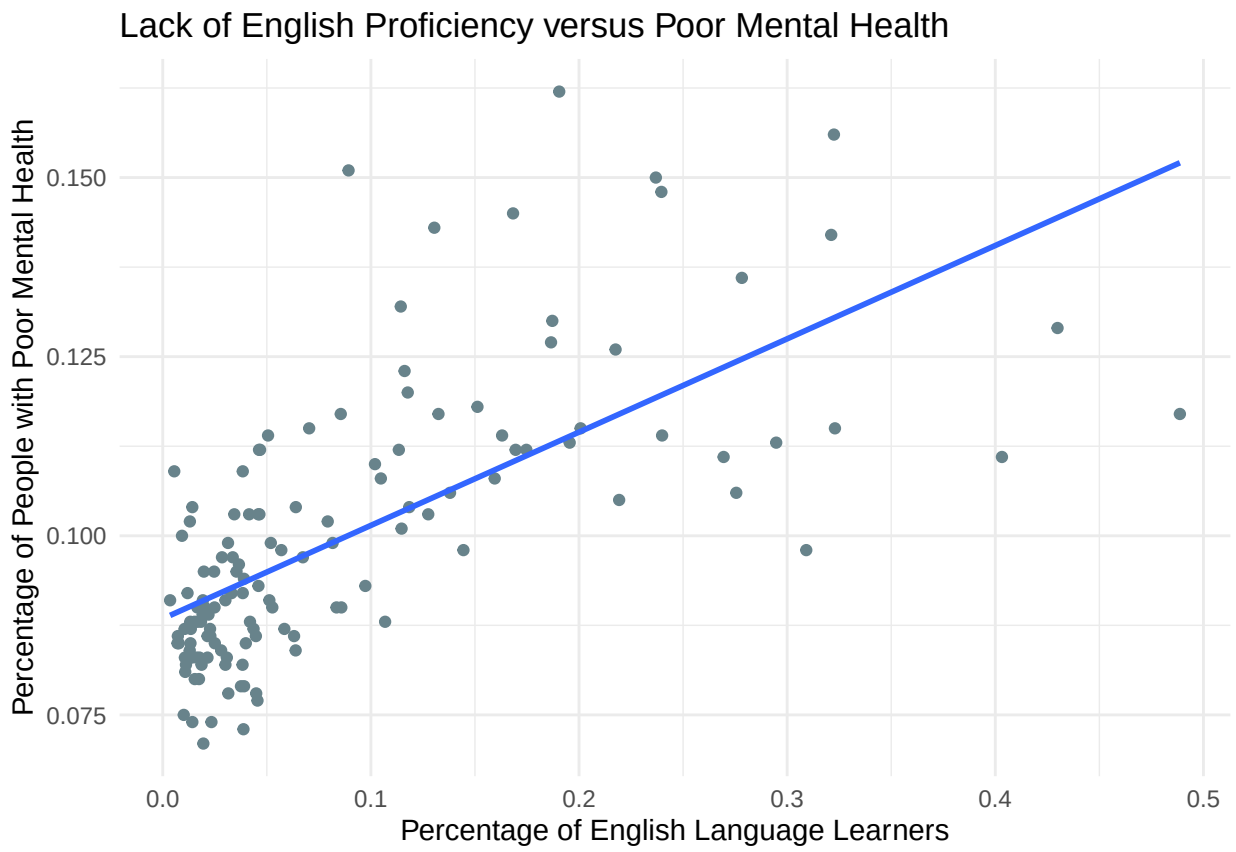



```
socioeconomic_fit <- lm(SOCIOECONOMIC_PERCENTILE ~ PCT_LESS_BACHELOR_DEGREE, data = seattle)
socioeconomic_fit_sum <- summary(socioeconomic_fit)
socioeconomic_fit_sum
```

```
##
## Call:
## lm(formula = SOCIOECONOMIC_PERCENTILE ~ PCT_LESS_BACHELOR_DEGREE,
##     data = seattle)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26026 -0.10697 -0.00230  0.08104  0.32540
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -0.12559    0.02981  -4.213 4.65e-05 ***
## PCT_LESS_BACHELOR_DEGREE  1.54050    0.06760  22.788  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1317 on 132 degrees of freedom
## Multiple R-squared:  0.7973, Adjusted R-squared:  0.7958
## F-statistic: 519.3 on 1 and 132 DF,  p-value: < 2.2e-16
```

```
mental <- seattle %>%
ggplot(aes(x = PCT_ENGLISH_LESSTHAN_VERY_WELL, y = PCT_ADULTMENTALHEALTHNOTGOOD)) +
 geom_point(color = "lightblue4") +
 labs(title = "Lack of English Proficiency versus Poor Mental Health",
      x = "Percentage of English Language Learners",
      y = "Percentage of People with Poor Mental Health") +
 theme_minimal() +
 geom_smooth(method = lm, se = FALSE)
mental
```

```
## `geom_smooth()` using formula 'y ~ x'
```



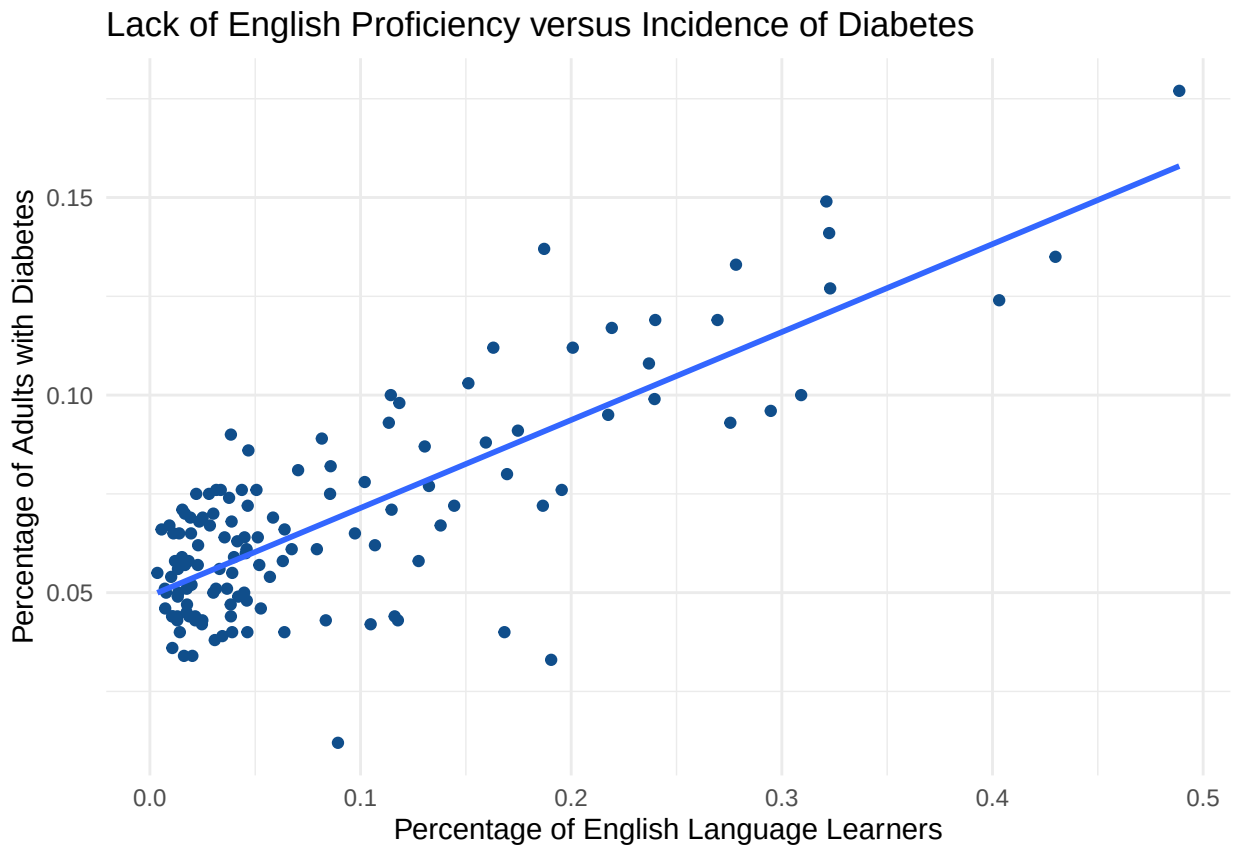

```
mental_fit <- lm(PCT_ADULTMENTALHEALTHNOTGOOD ~ PCT_LESS_BACHELOR_DEGREE, data = seattle)
mental_fit_sum <- summary(mental_fit)
mental_fit_sum
```

```
##
## Call:
## lm(formula = PCT_ADULTMENTALHEALTHNOTGOOD ~ PCT_LESS_BACHELOR_DEGREE,
##     data = seattle)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.023107 -0.006807 -0.002519  0.002525  0.055040
```

```
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.065019   0.002761   23.55   <2e-16 ***
## PCT_LESS_BACHELOR_DEGREE 0.085136   0.006260   13.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0122 on 132 degrees of freedom
## Multiple R-squared:  0.5835, Adjusted R-squared:  0.5804
## F-statistic:   185 on 1 and 132 DF,  p-value: < 2.2e-16
```

```
diabetes <- seattle %>%
ggplot(aes(x = PCT_ENGLISH_LESSTHAN_VERY_WELL, y = PCT_ADULT_WITH_DIABETES)) +
 geom_point(color = "dodgerblue4") +
 labs(title = "Lack of English Proficiency versus Incidence of Diabetes",
      x = "Percentage of English Language Learners",
      y = "Percentage of Adults with Diabetes") +
 theme_minimal() +
 geom_smooth(method = lm, se = FALSE)
diabetes
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```r
diabetes_fit <- lm(PCT_ADULTMENTALHEALTHNOTGOOD ~ PCT_ADULT_WITH_DIABETES, data = seattle)
diabetes_fit_sum <- summary(diabetes_fit)
diabetes_fit_sum
```

```
##
## Call:
## lm(formula = PCT_ADULTMENTALHEALTHNOTGOOD ~ PCT_ADULT_WITH_DIABETES,
##     data = seattle)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.027562 -0.009570 -0.003401  0.005603  0.074087
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.07693    0.00387  19.878  < 2e-16 ***
## PCT_ADULT_WITH_DIABETES  0.33277    0.05251   6.338  3.4e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01655 on 132 degrees of freedom
## Multiple R-squared:  0.2333, Adjusted R-squared:  0.2275
## F-statistic: 40.17 on 1 and 132 DF,  p-value: 3.402e-09
```

Stuff from Milestone 1

Data Set

Data set: https://catalog.data.gov/dataset/racial-and-social-equity-composite-index-a44fc

Understanding the variables: https://data-seattlecitygis.opendata.arcgis.com/datasets/225a4c2c50e94f2cb548a046217f49f7_0?geometry=-122.509%2C47.574%2C-122.164%2C47.655

This data set examines linguistic, racial, ethnic, income, education, and health statistics for census tracts in Seattle.

I want to examine how the percent of English language learners in census tracts correlates with factors like obesity, poverty, education level, asthma, and mental health.

From what I've seen of it, this data set is excellent because it has a lot of information I can draw from and is "clean" so I don't need to recode too much. It is also well under the recommended cap of 50 MB.

Proposal

Research question: How much of a socioeconomic advantage do English speakers have over non-English speakers in Seattle? In this study, I plan to use public information released by the city of Seattle to visualize the connection between ELL status and certain health, education, and income variables. I hypothesize that, because of the institutional advantages of knowing English in America, English speakers will tend to be healthier, more educated, and have higher-paying jobs. The explanatory variable in this experiment is the amount of ELL people living in census tracts (represented in the data with a percentage, but I might break this up into a few bins) and the dependent variable is the variation in education level, income, asthma rates, mental health, and diabetes rates (also measured with percentages). I will test my hypothesis by trying to make all other factors equal and then plotting ELL percentages against the other variables I'm studying. If,

after controlling for all those other factors, the lower-income census tracts have statistically significant worse health, education, and income, the pattern would support my hypothesis.

I plan to do this project by myself.