Project Proposal

staRstistions - Will Lieber, Wania Iftikhar Khan, AJ Tenser, Kami Akala

```
library(tidyverse)
library(tidymodels)
library(patchwork)
library(ggplot2)
#install.packages("psych")
library(psych)
#install.packages("readxl")
library(readxl)
library(dplyr)
```

Introduction

Data description

```
ca_incarceration <- read_excel("data/California incarceration rate data.xlsx")
ca_lead <- read_excel("data/ZIPCodeData2022 (1).xlsx")

ca_demographic <- read_csv("data/demographic.csv", skip = 1)
ca_income <- read_csv("data/income.csv", skip = 1)
```

```
#transforming the data

#ca_demographic <- read_csv("data/demographic.csv", skip = 1)

ca_demographic <- ca_demographic |>
  select(!((contains("Estimate") | contains("Margin")) & !37)) |>
  select(c(1:6, 8:17, 40, 41, 42, 47, 55, 60)) |>
  rename_with(~ str_replace(., ".*!!", ""))
```

```r
ca_demographic <- ca_demographic |>
  mutate(`Geographic Area Name` = substr(`Geographic Area Name`, 7, 11)) |>
  mutate(across(!1, as.numeric)) |>
  mutate(`0_to_19` = `Under 5 years` + `5 to 9 years` + `10 to 14 years` + `15 to 19 years`,
         `20_to_44` = `20 to 24 years` + `25 to 34 years` + `35 to 44 years`,
         `45_to_64` = `45 to 54 years` + `55 to 59 years` + `60 to 64 years`) |>
  select(!7:16) |>
  rename(zip_code = `Geographic Area Name`)

#ca_income <- read_csv("data/income.csv", skip = 1)

ca_income <- ca_income |>
  select(c(1, 2, 25, 27)) |>
  rename_with(~ str_replace(., ".*!!", "")) |>
  mutate(`Geographic Area Name` = substr(`Geographic Area Name`, 7, 11)) |>
  mutate(across(!1, as.numeric)) |>
  rename(zip_code = `Geographic Area Name`)
```

```r
#renaming the columns
ca_incarceration <- ca_incarceration |>
  rename(zip_code = `California ZIP codes    `) |>
  mutate(zip_code = as.numeric(zip_code))


ca_lead <- ca_lead |>
  rename(zip_code = `ZIP Code`) |>
  mutate(zip_code = as.numeric(zip_code))
```

```r
#left_join to pare down to least observation dataset

joined_data1 <- left_join(ca_lead, ca_incarceration, by = "zip_code")
joined_data2 <- left_join(joined_data1, ca_demographic, by = "zip_code")

joined_data <- left_join(joined_data2, ca_income, by = "zip_code")

joined_data_clean <- joined_data |>
  janitor::clean_names() |>
  select(!c(geography_y, city, census_population_2020)) |>
  rename(city = postal_district_name) |>
  rename(geography = geography_x)
```

```r
joined_data_clean %>%
  nrow()
```

```
[1] 1777
```

Exploratory data analysis

...

Analysis approach

...

Data dictionary

The data dictionary can be found here [Update the link and remove this note!]