Project Proposal

STOR 320.(01 OR 02) Group 1

Topic

- **Chicago Neighborhoods**

Dataset

Chicago Crime and Public Schools (https://www.kaggle.com/ortizmacleod/ibm-sql-course-chicago-crime-and-public-schools)

Chicago Community Area CDS data (https://datahub.cmap.illinois.gov/dataset/1d2dd970-f0a6-4736-96a1-3caeb431f5e4/resource/8c4e096e-c90c-4bef-9cf1-9028d094296e/download/ReferenceCCAProfiles20142018.csv)

Initial Questions

- **Question 1:** Relationship between income and college enrollment, per capita, in each neighborhood
- **Question 2:** Relationship between ethnicity and college enrollment
- **Question 3:** Relationship between transit to work, and hardship index
- **Question 4:** Relationship between unemployment and percent of households below poverty
- **Question 5:** Relationship between population growth and diversity for each neighborhood, from 2010 to 2020
- **Question 6:** Relationship between language Level and college enrollment per capita

Data Wrangling and Analysis Plan

Data wrangling

- Merge the two datasets together by neighborhood name
- Missing data in the excel spreadsheet- will have to delete the missing data
- Plotting relationships using various ggplot graphs

Question Analyses Plan:

- Question 1
  - Scatterplot
    - X axis is median income (Independent variable)
    - Y axis is college enrollment per capita (Dependent variable)
    - Each neighborhood is a single dot– 77 data points total
- Question 2
  - Scatterplot (regression analysis)
    - Each neighborhood is a single dot

        - X axis is ethnicity (Independent variable)
            1. We are given total # number of ppl in an ethnicity, but we believe a percentage would be better
        - Y axis is college enrollment (Dependent variable)
- Question 3
    - Stratified scatterplots (Regression Analysis for each group)
        - Stratified based on the method of commute
        - X axis is percentage of the commute method used in each neighborhood
        - Y axis is hardship index percentile
            1. Hardship index is ranked from 99 to 1 for each school. The percentile is based on where the school is ranked amongst other schools in the dataset.
        - Each datapoint is a school
- Question 4
    - Two different Heat map (2010 to 2020)
        - Different shades of colors on Chicago city map based on average inflation adjusted income
        - Statistical test also displaying whether the changes in incomes between the neighborhoods are significant between the 10-year period
- Question 5
    - Stratified map plot
        - Stratified by ethnic group
        - Each ethnic group will have 3 maps:
            1. Map with percentage of ethnic group, in each neighborhood based on 2010 figures
            2. Map with percentage of ethnic group, in each neighborhood based on 2020 figures
            3. Map with percentage change for ethnicity, in each neighborhood, from 2010 to 2020
- Question 6
    - Bar chart
        - Percentage of college enrollment based on language spoken at home
        - X axis: language
        - Y axis: college enrollment