Final Project Step 1

2022-01-29

Introduction

Many people still extol the American dream that used to be a single income supporting a suburban home with a picket fence, two cars in the driveway and a family. However, for many there appears to be inequality in the US in who shares in this dream. There is income and wealth inequality by gender, race and even age which may seem to be caused by inequity. Personally, I am very interested in financial ideas due to my work experience and my experience in trying to be as financially successful as my older family members. In this project I hope to explore some of this income inequality to see if I can identify relationships between variables that may also help to explain some of the differences.

Research Questions

1. Is there a visible income gap between genders in the data?
2. Is there a visible income gap between races in the data?
3. Is there a relationship between age and income?
4. Is education a factor in these gaps?
5. Does career type factor in these gaps?

Approach

My hypothesis is that there are multiple variables that go into the differences in pay between genders, races, and age groups. In my approach I will be looking at the correlation and strength of correlation between variables to explore the relationships between them along with performing numerous other statistical tests. I will also explore the data visually using plots and graphs.

This approach can help identify the strength of relationships between the variables to identify if there are discernable causes for this income gap.

Datasets

1. https://www.kaggle.com/dbsimpson/us-college-graduates-wages?select=Unemployment_rate.csv
2. https://www.kaggle.com/nilimajauhari/glassdoor-analyze-gender-pay-gap
3. https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-pinc/pinc-01.html
4. https://www.kaggle.com/econdata/predicting-earnings-from-census-data

Packages Needed

Base R functions

Readxl

Ggplot2

Dplyr

Purrr May be some that I did not include that I end up using.[^1]

Plots and Tables

1. Scatterplots: To identify relationships visually
2. Box Plots: To look for outliers
3. Histograms: To check for normality

Questions I have on completing my project

I will have to figure out how I want to incorporate all my data efficiently and merge the data in a coherent way. I also will have to improve my skill in dealing with categorical variables.