Immigration Dataset Exploratory Data Analysis Summary

David D Berberena

Bellevue University

DSC 530 Data Exploration and Analysis

Amirfarrokh Iranitalab

March 2, 2024

Immigration Dataset Exploratory Data Analysis Summary

For this project, I decided to perform EDA on an immigration dataset originating from the Department of Homeland Security, which shared immigration, refugee, and noncitizen statistics over several years. I was interested to see if an increase in the number of refugees over time increased the number of immigrants who were granted permanent residency. This question became my EDA hypothesis as I began to break down the data. The results of my analysis showed that the data did not support a positive correlation between an influx of refugees and an increased number of green card holders.

Looking back at my EDA, I do not feel that anything was missed as my hypothesis was directly answered by statistics like Pearson’s correlation coefficient and Spearman’s correlation, covariance between the two variables, a correlation permutation test, and fitting a linear regression model to the data. If any variables could have helped in the analysis, they were not in the dataset that I worked with. I would most likely have to dive deeper into the concept and pull data directly from the Department of Homeland Security, yet my lack of access would prevent that. I did assume that with the increase of refugees came the increase of green card holders, yet as the data showed, this assumption was incorrect. Instinctively this logic makes sense, however, the reality is held in the data and the results stemming from the tests and analysis.

Stumbling blocks did crop up during the EDA process for me, and while I initially had trouble overcoming them, outside research allowed me to discover what needed to be fixed to move along with the analysis. Some of these dataset roadblocks involved changing the variable names slightly so they were easier to call, finding the correct number of bins for the histograms needed, and removing the commas in the data points to convert them into numerical data points. Once they were solved, I was able to fully understand the rest of the EDA process and realize the needed results.