Team FitBit - Data Exploration

Rikka, Shiva Ram Reddy (00759744)
srikk1@unh.newhaven.edu

Chilukuri, Gayathri (00755590)
gchil4@unh.newhaven.edu

Potlapelli, Adithya (00761532)
apotl1@unh.newhaven.edu

October 16, 2022

GitHub Repository

1 Introduction

Poverty is one of the topics studied by many economists and data scientists. It is one of the economic issues that most governments wish to address. As one of the world's most powerful economies, the United States faces the challenge of domestic poverty. This project will concentrate on the poverty rate in New York City because it is one of the most representative cities in the United States, as well as having the country's highest income inequality. The project will make use of decision tree model analysis to predict poverty in New York and identify the features that can generate the most accurate predictions.

2 Dataset

The data utilized in this project is derived from the NYCgov poverty measure data, which is generated annually by poverty research unit of the Mayor's Office of Economic Opportunity. This dataset has 68,644 observations and 79 distinct variables. The dataset includes various characteristics of New York City households (with unique identities), such as education levels, work status, annual income, and so on. Based on these characteristics, a decision is reached on whether or not a certain family is poor.

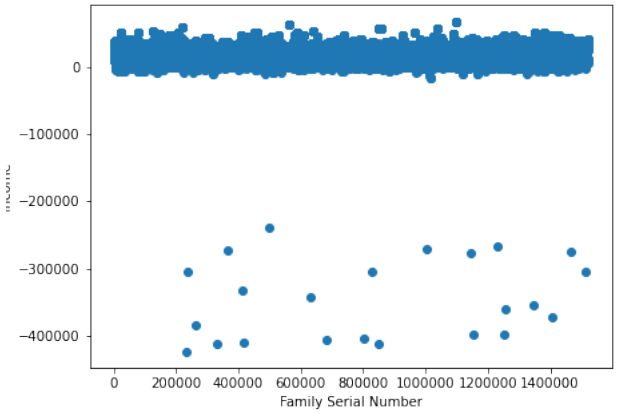
3 Exploration Techniques

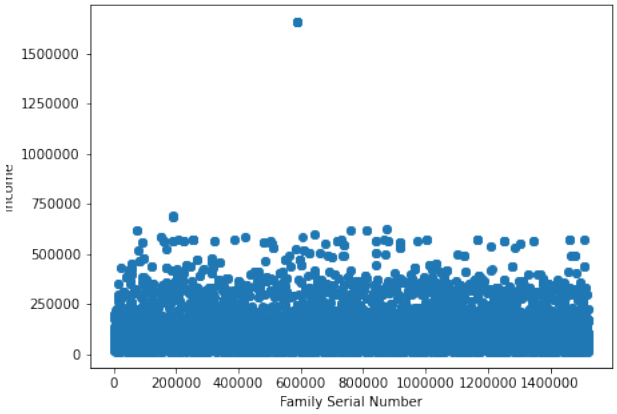
A summary statistics and exploratory data analysis was performed to gain insights of the data that included varied visualization techniques such as scatter plot, bar graph and histogram.

4 Data Exploration

Scatter Plot

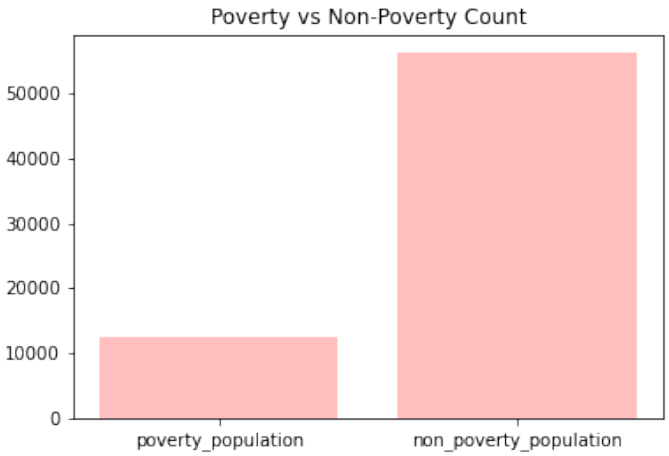
To observed the data distribution and a relationship (if any), Scatter plot of income of families in Poverty and Scatter plot of income of families in Non-Poverty.





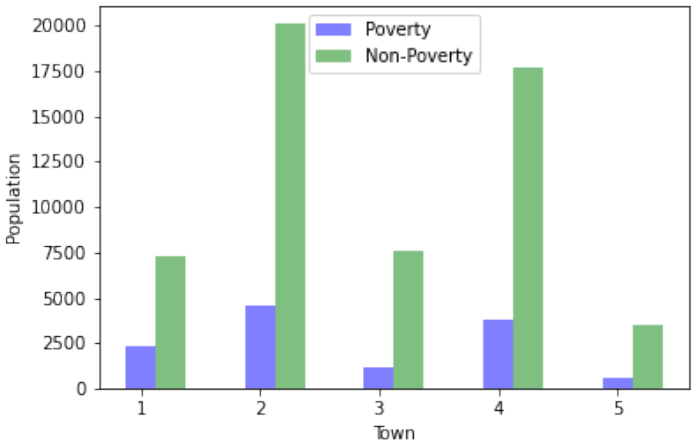
Bar Graph

Bar graph to illustrate and compare the population count in poverty and non-poverty.



Histogram

To understand the poverty and non-poverty population in different towns of New York.



5 Conclusion

The data analysis provided a detailed understanding of the data distribution. A significant finding is that the number of families in poverty is lower than the number of families in non-poverty (the above bar graph bar to illustrate this imbalance).