New York's 2019 Volunteer Analysis

Presented By
Padraig Reilly & Sydney Tran

01/26/2022

Project Outline

Dataset: 544 Rows, 32 Columns

PART 1	Problem Statement
PART 2	Background
PART 3	Exploratory Data Analysis
PART 4	Preprocessing & Modeling
PART 5	Conclusion & Next Steps

Problem Statement

Going into 2022, the city of New York is trying to identify ways to improve engagement and ensure more volunteers for the upcoming year.

Utilizing New York's 2019 Volunteers Count Report Boroughs, this project aims to explore the volunteer count compared to area, organization type, interest areas, and boroughs served, to see if there is a relationship between the amount of volunteers and these features.

THE GLOBAL GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS

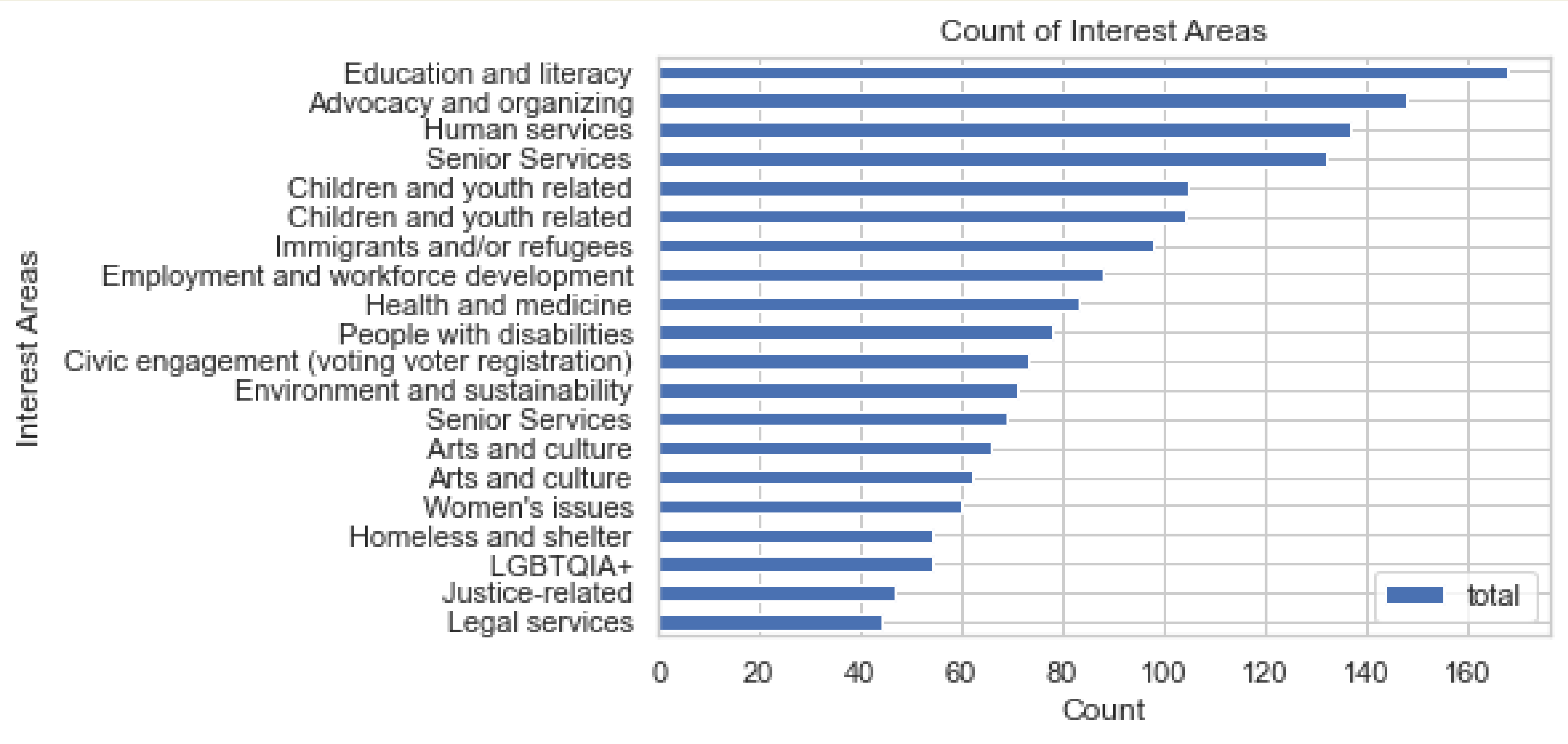
Background

- The annual NYC Volunteers Count report is the City's largest scan of residents volunteering at organizations across New York City

- Organizations are surveyed to understand how residents volunteer within the city's infrastructure to strengthen communities at the neighborhood level

- Each year, survey as many organizations that engage volunteers in service as possible to include the diversity of services provided and the number of residents civically engaged as volunteers

Interest Areas



168

Education & Literacy

148

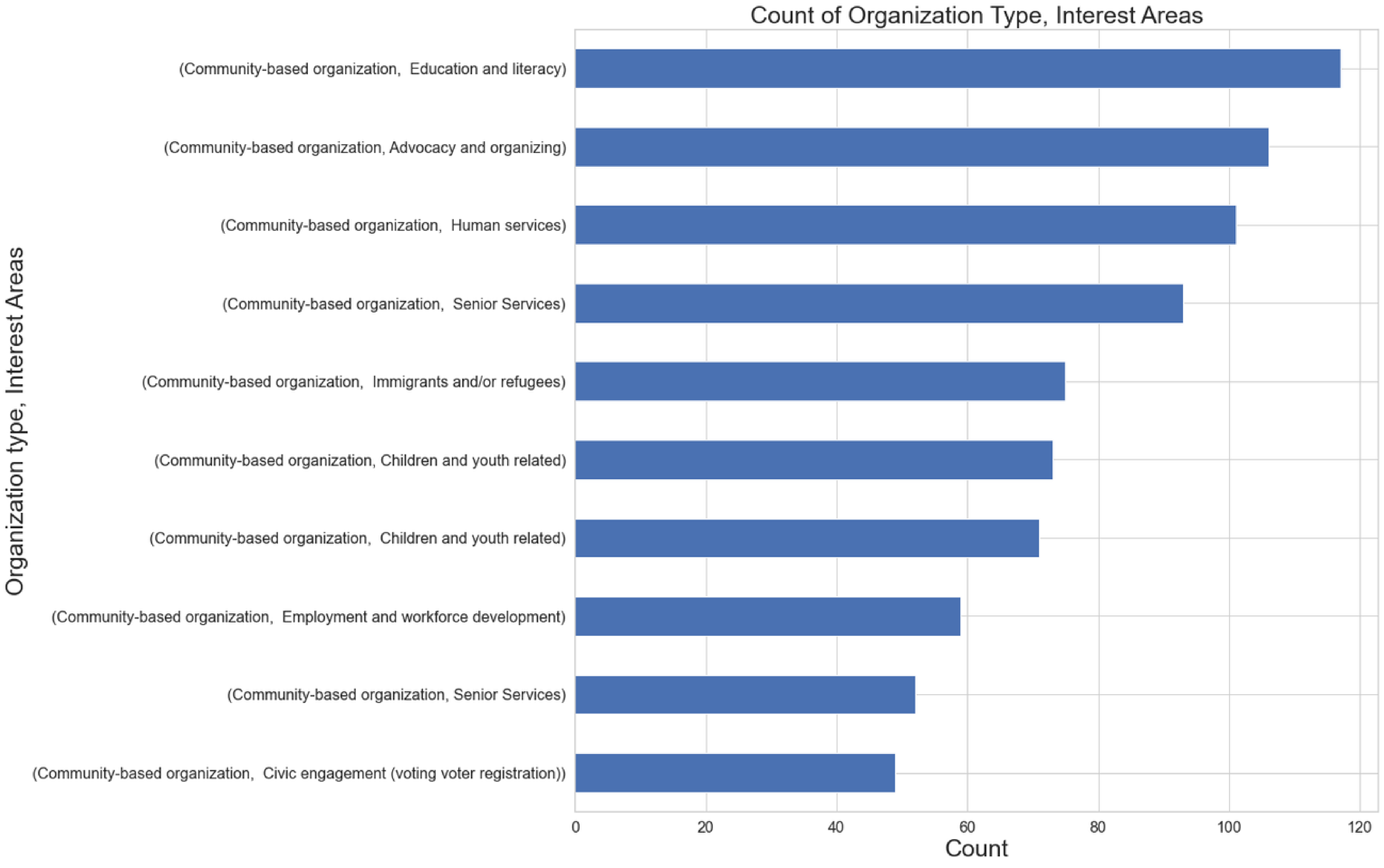
Advocacy & Organizing

137

Human Services

Organization Type & Interest Areas

The top 3 organizations were community-based organizations that had interest areas in education and literacy, advocacy and organizing, and human services



117

Community-based, Education & Literacy

106

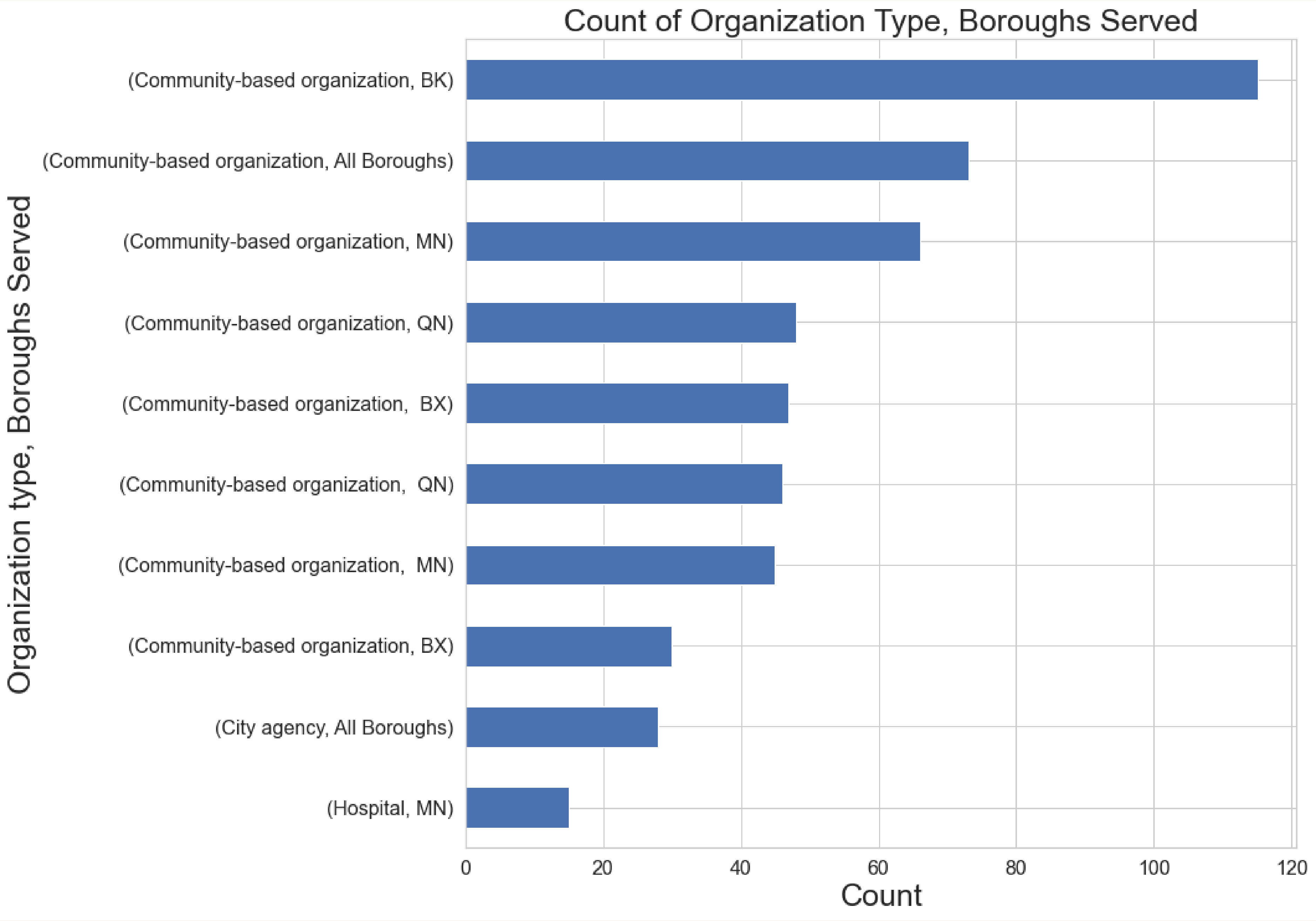
Community-based, Advocacy & Organizing

101

Community-based, Human Services

Organization Type & Borough's Served

The top 3 organizations were community-based organizations that were in Brooklyn, all boroughs, or Manhattan



115

Community-based, Brooklyn

73

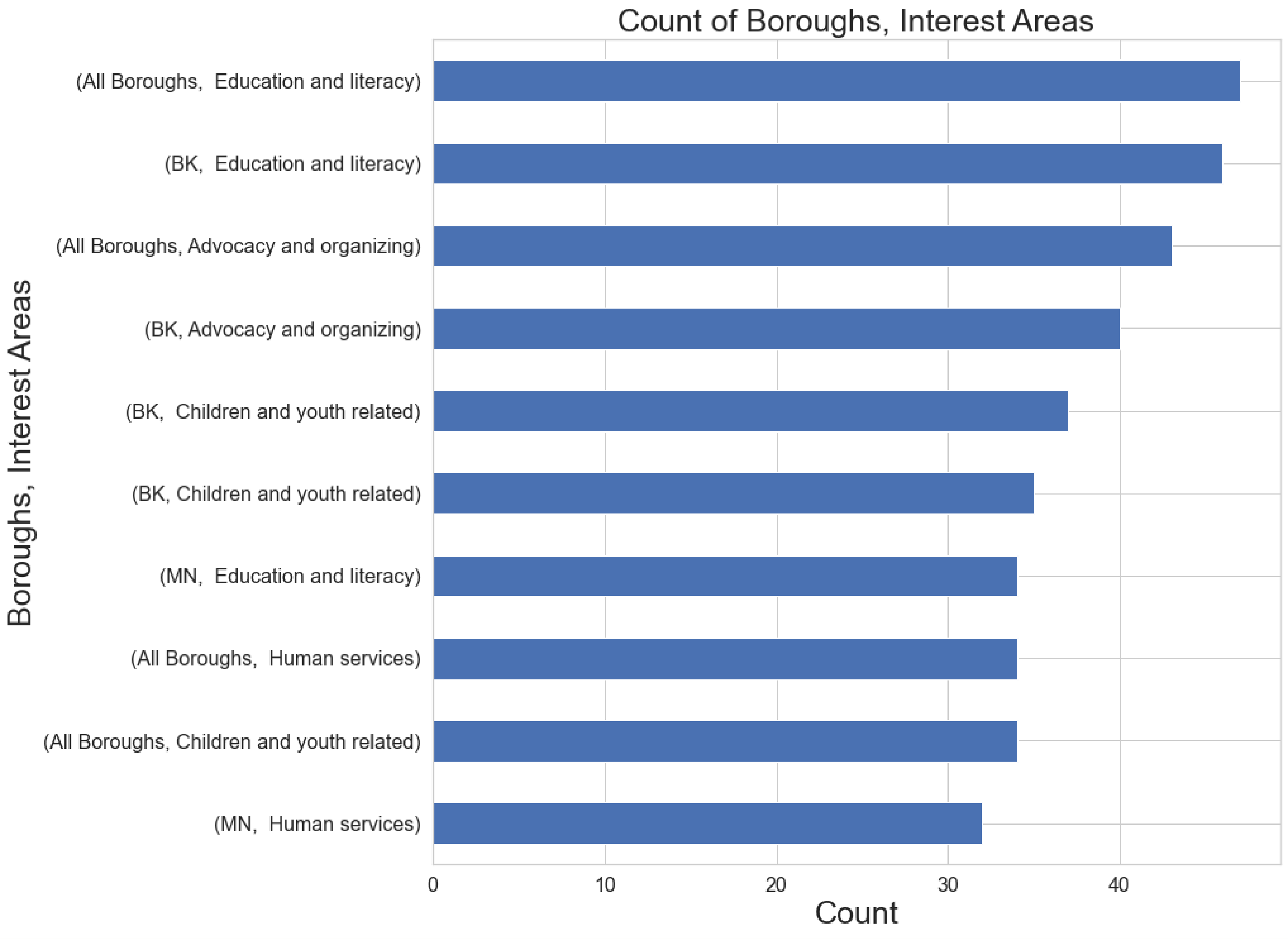
Community-based, All Boroughs

66

Community-based, Manhattan

Borough's Served & Interest Areas

The top 3 boroughs served and their interest areas were all boroughs combined for education and literacy and advocacy in organizing, and Brooklyn for education and literacy



47

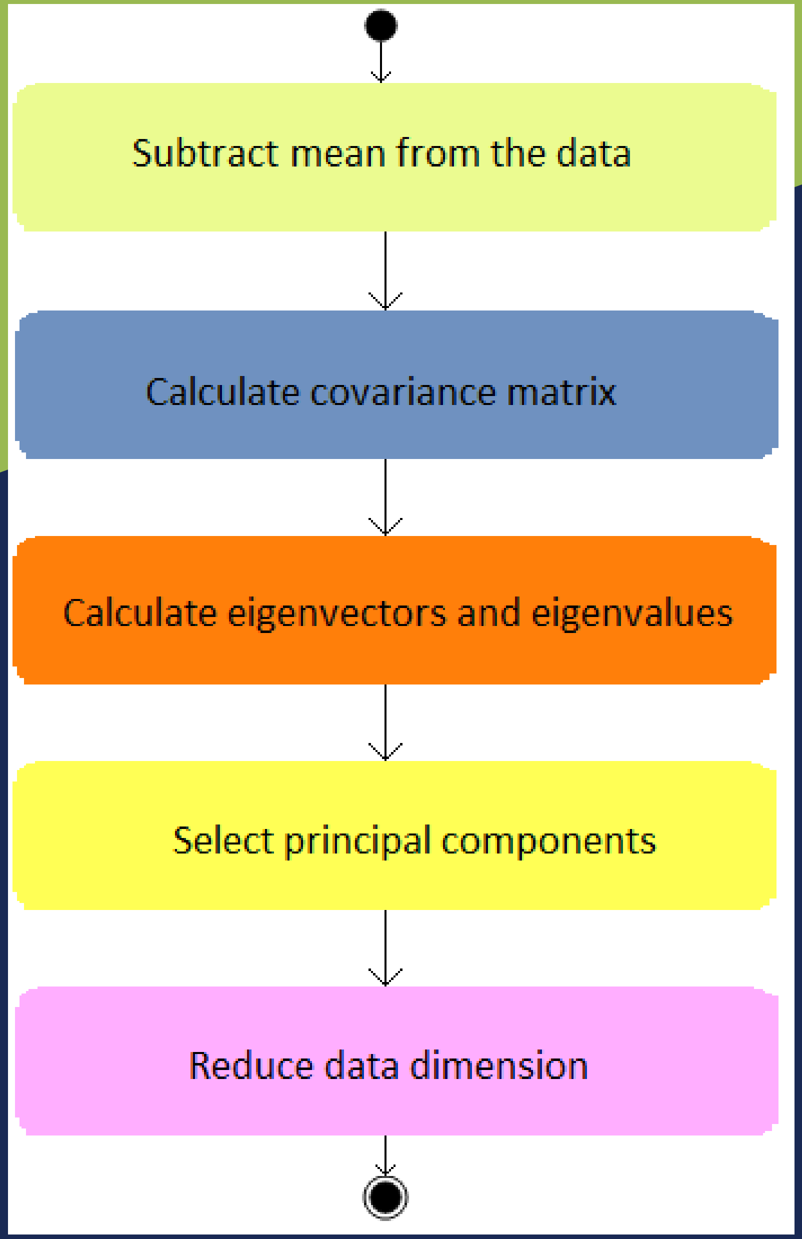
All Boroughs,
Education & Literacy

46

Brooklyn,
Education & Literacy

43

All Boroughs,
Advocacy & Organizing



Subtract mean from the data

Calculate covariance matrix

Calculate eigenvectors and eigenvalues

Select principal components

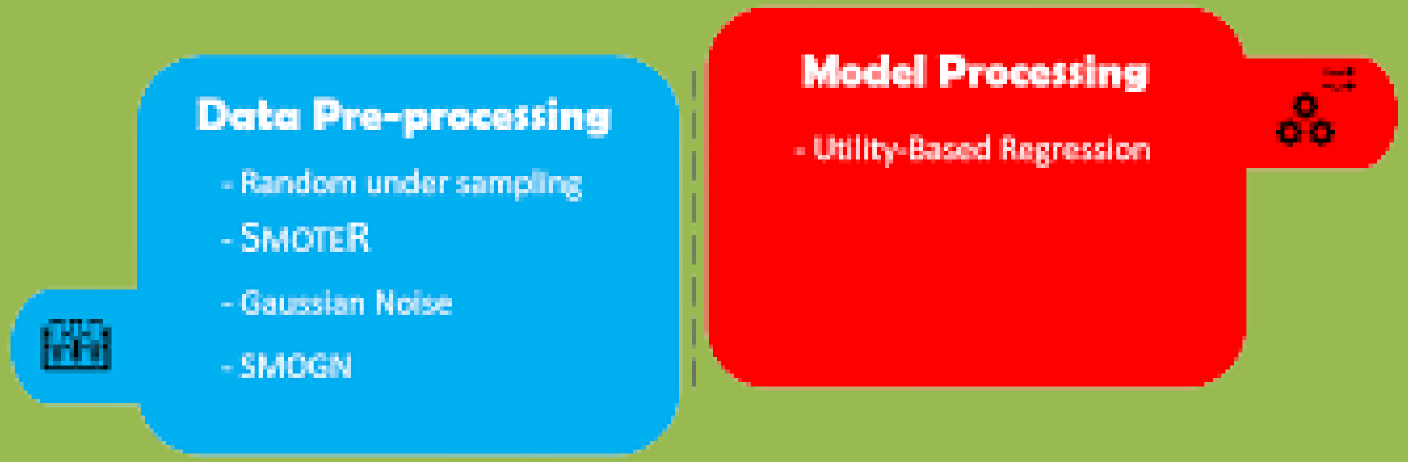
Reduce data dimension

PCA: Principal Component Analysis

- Used to reduce dimensionality
- Helps identify important relationships in our data
- Transforms the data then quantifies the importance of these relationships

SMOGN:

Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise



Data Pre-processing

- Random under sampling
- SMOTER
- Gaussian Noise
- SMOGN

Model Processing

- Utility-Based Regression

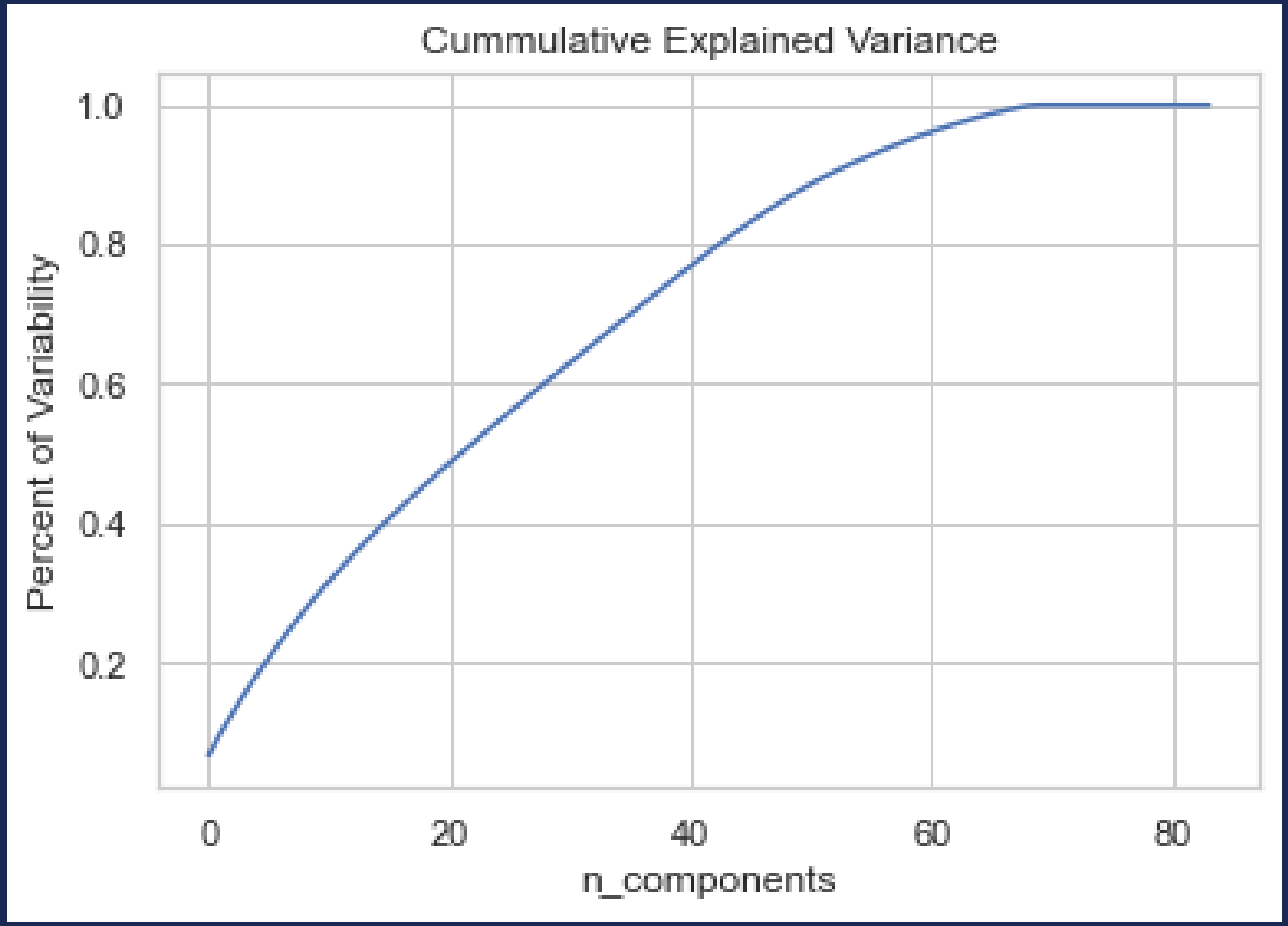
- Preprocessing step
- Resampling the rare cases for regression problems
- Conducts SMOTER and SMOTER-GN
  - Selects between the two techniques by the KNN distances
  - If distance is close, SMOTER is used
    - Otherwise, SMOTER-GN

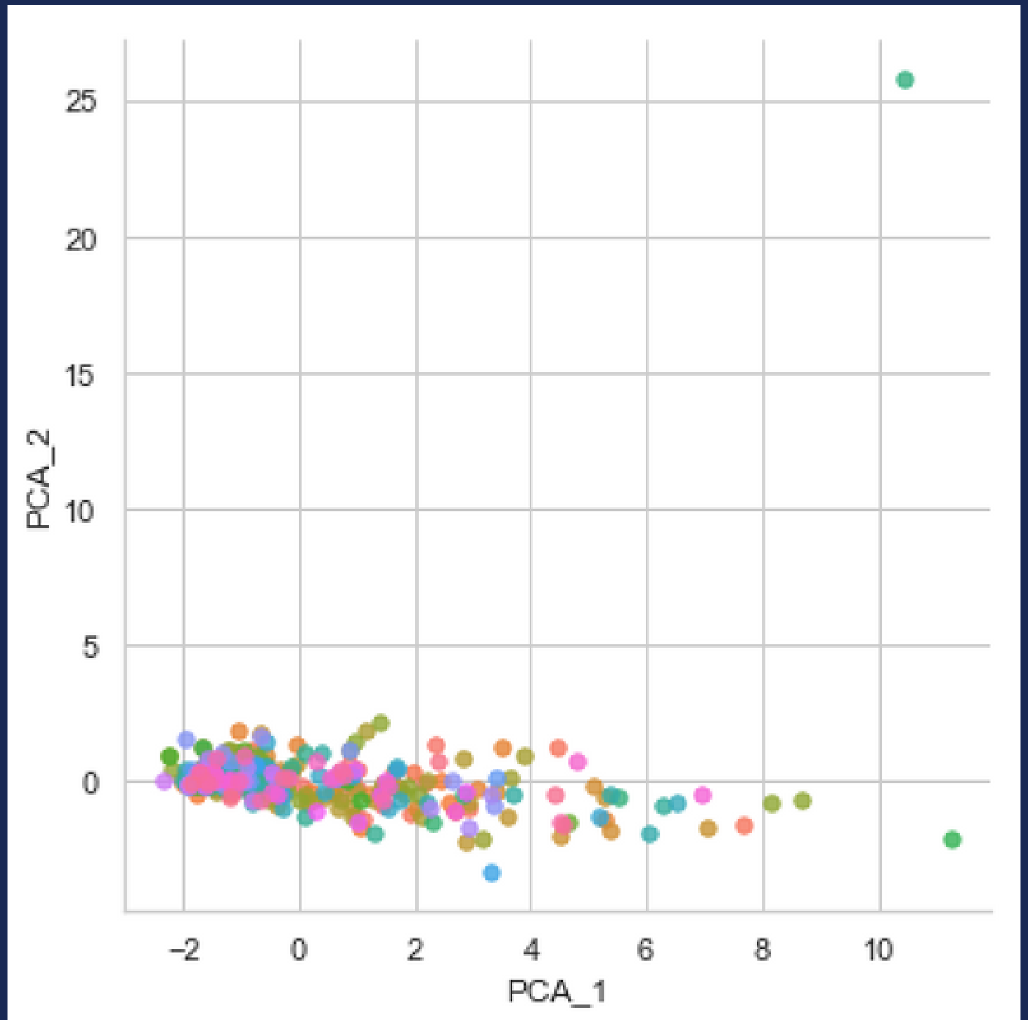
Baseline Model

Linear Regression

Train Score: 0.1813
Test Score: -1.079

Preprocessing: PCA



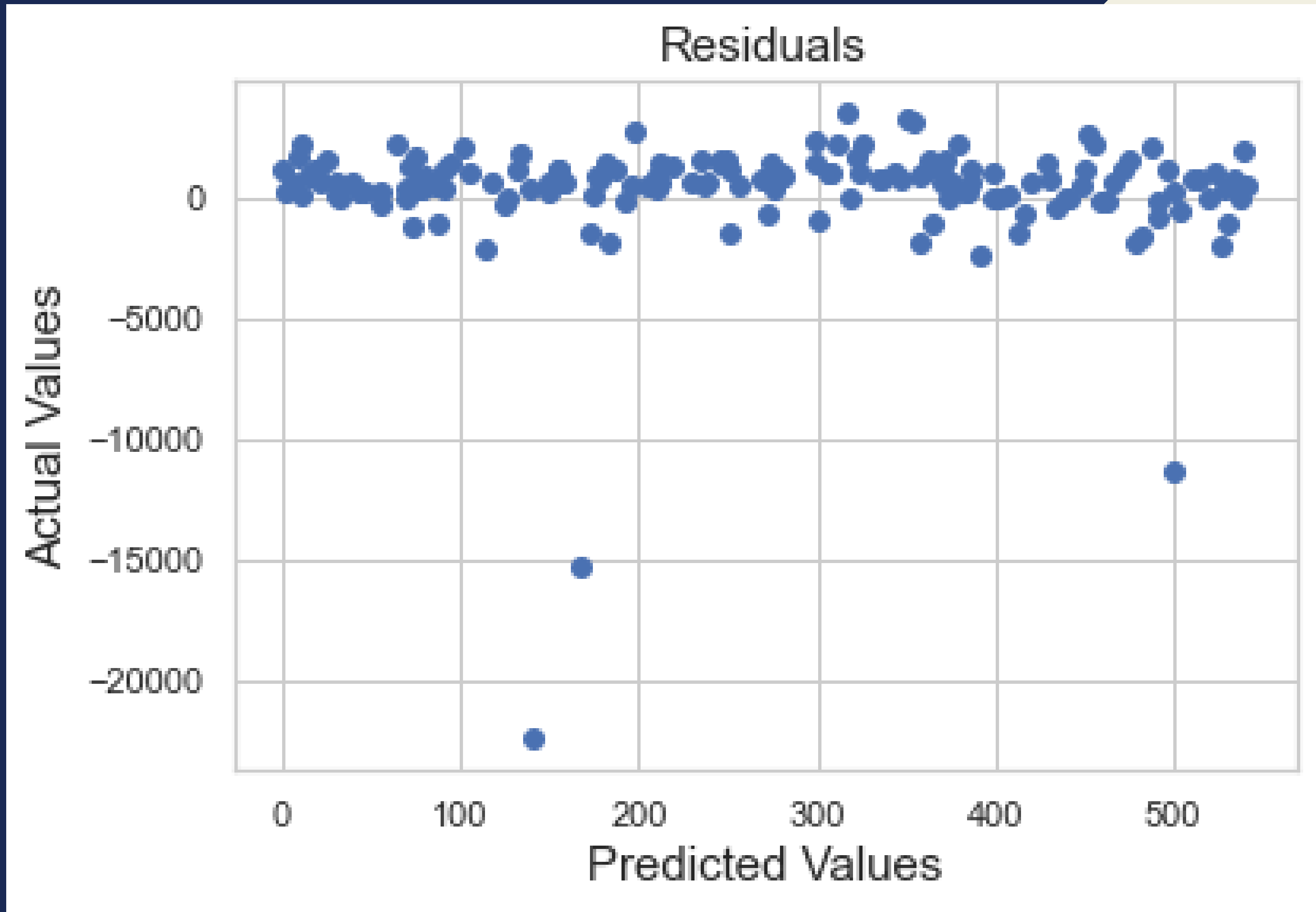


80% of variability ~ 40 n_components

Train Score: 0.0836

Test Score: 0.019

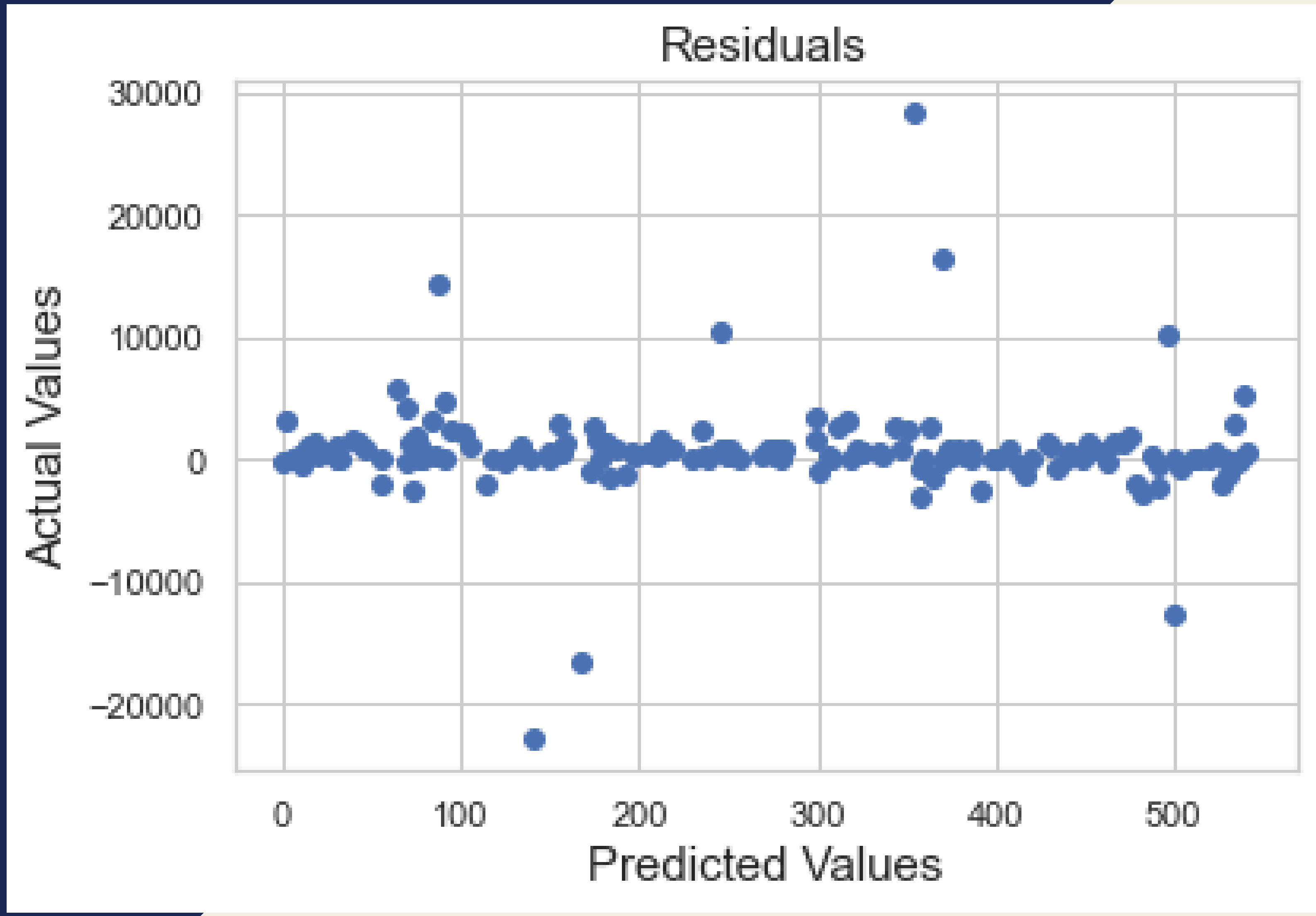
Model 1



Linear Regression with PCA

Train Score: 0.0497
Test Score: 0.0316

Train RMSE: 4487.48
Test RMSE: 2600.04



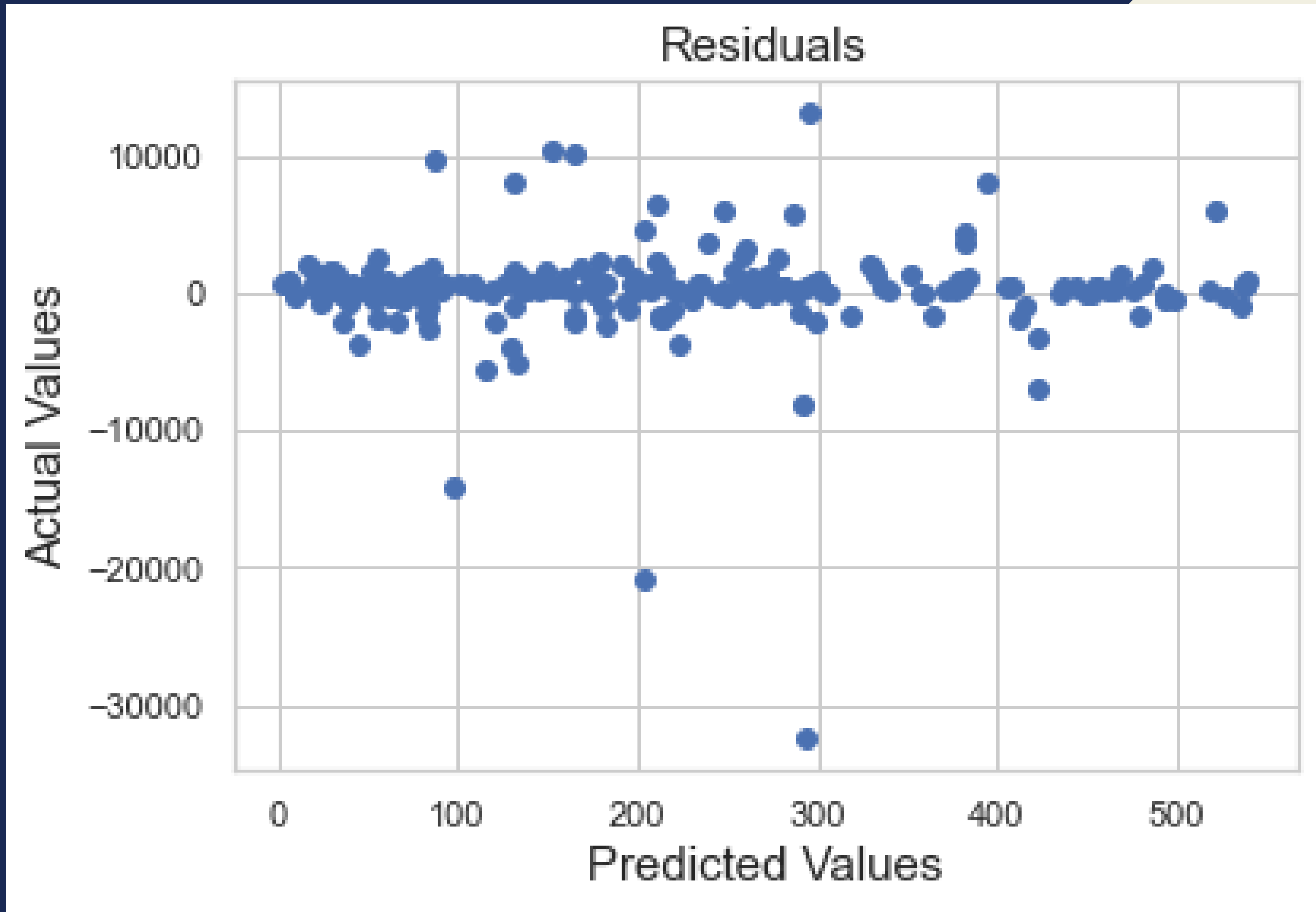
Model 2

Random Forest with PCA

Train Score: 0.774
Test Score: -1.466

RMSE Train: 2189.48
RMSE Test: 4148.92

Model 3



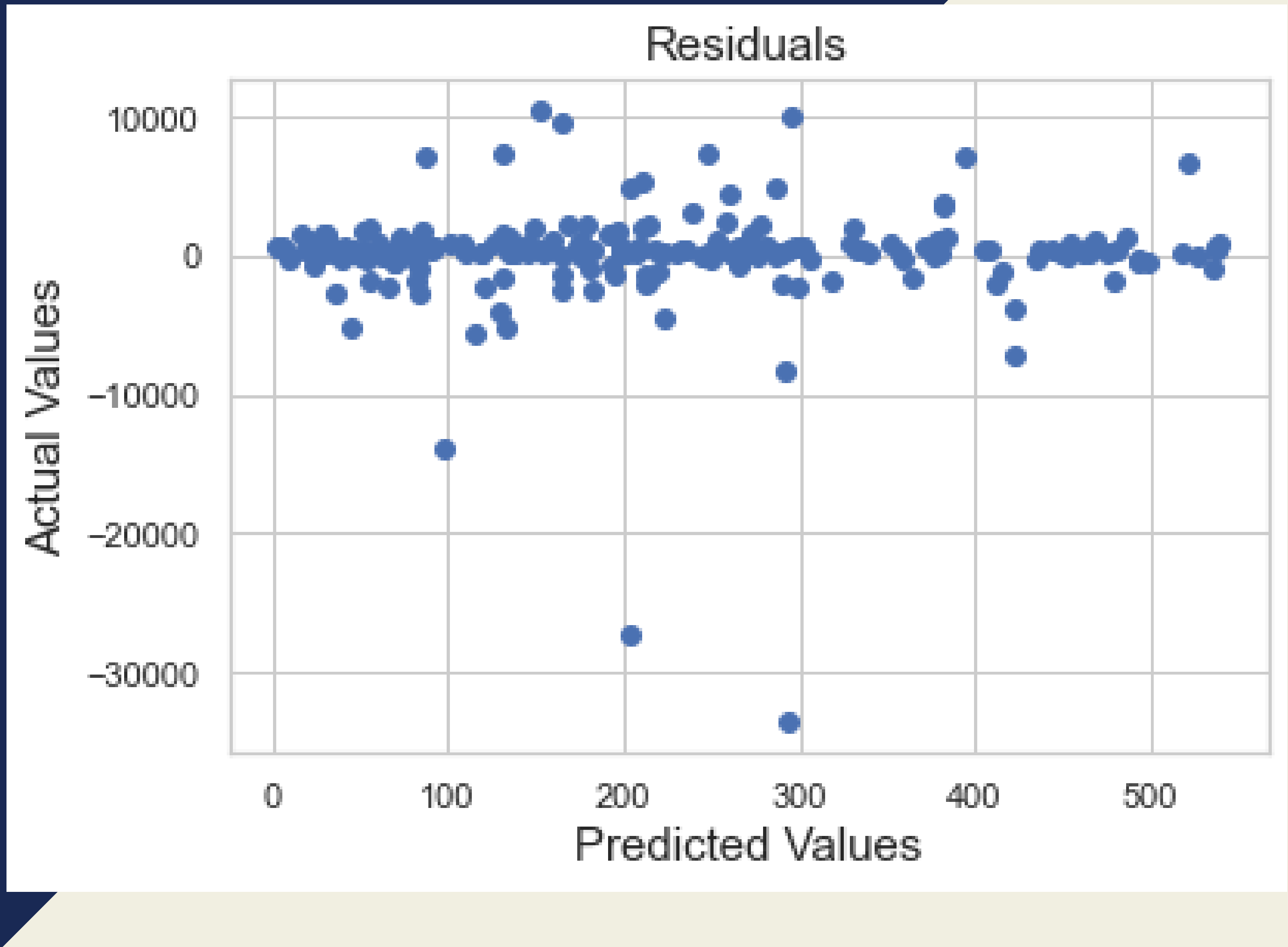
Random Forest with SMOGN

Train Score: 0.713
Test Score: 0.529

RMSE Train: 2670.93
RMSE Test: 3645.51

Model 4

Random Forest with SMOGN: Gridsearch



Train Score: 0.633
Test Score: 0.485

RMSE Train: 3018.84
RMSE Test: 3811.52

Conclusion

- Overall, all of our models were overfit
- Adjusted problem statement due to lack of data/missing data
- Used PCA and SMOGN to help build a Random Forest model

NEXT STEPS 1

Special Populations Served

- Clean, explode, dummify feature
- Add to model

NEXT STEPS 2

Further clean dummified columns

- Combine rows < 10

NEXT STEPS 3

- Bring in outside data
- Census data

Sources:

https://www.nycservice.org/pages/pages/151

https://www.neuraldesigner.com/blog/principal-components-analysis

https://towardsdatascience.com/regression-for-imbalanced-data-with-application-edf93517247c

https://pypi.org/project/smogn/