Brad Pawlow, Sebastian Reid, Noah Slosberg

04/24/23 - 04/28/2023

STAT 390

Progress Report 3

**Goals from Week Plan 2:**

* Task 1: Finish data cleaning to filter out the predictor variables that are significant
* Task 2: Use Lasso and Ridge regression to find which features best predict and affect the data.
* Task 3: Decide on predictor variables and overall direction of our analysis
* Task 4: Flatten dataset to only contain unique observations
* Task 5: Discard programs where the minimum age is over 25
* Task 6: Reorganize dataset with more geographic data - Noah
* Task 7: Perform PCA and factor analysis to condense dimensionality of dataset

**Progress/Accomplishments**:

* Completed Task 3 on schedule. We decided to focus our final project analysis on predicting academic opportunities across different neighborhoods based on respective programs.
  + Include a separate dataset of academic performances across different Chicago neighborhoods
  + Locate the different neighborhoods that could use more programs based on the ratio of programs dispersed across the city to maximize resources
* Resolved the issue of deciding which categories to direct our EDA analysis. For bivariate and multivariate analysis, we will depict a broad depiction of all programs across neighborhoods and income levels, but make sure to mention our focus.
* Conducted 2 meetings with the team to discuss/generate EDA plots, direction of project, and data cleaning techniques.
* All plots currently completed are attached in the *appendix* section
* Completed Task 4 on schedule.
* Completed Task 5 on schedule.
* Made significant progress on Task 1/Task 2 with about 25% of the work completed

**Problems/Challenges**:

* Encountered delays with finding and incorporating additional geographic data for Chicago neighborhoods
* Encountered delays in starting to perform PCA and factor analysis due to more focus on EDA steps

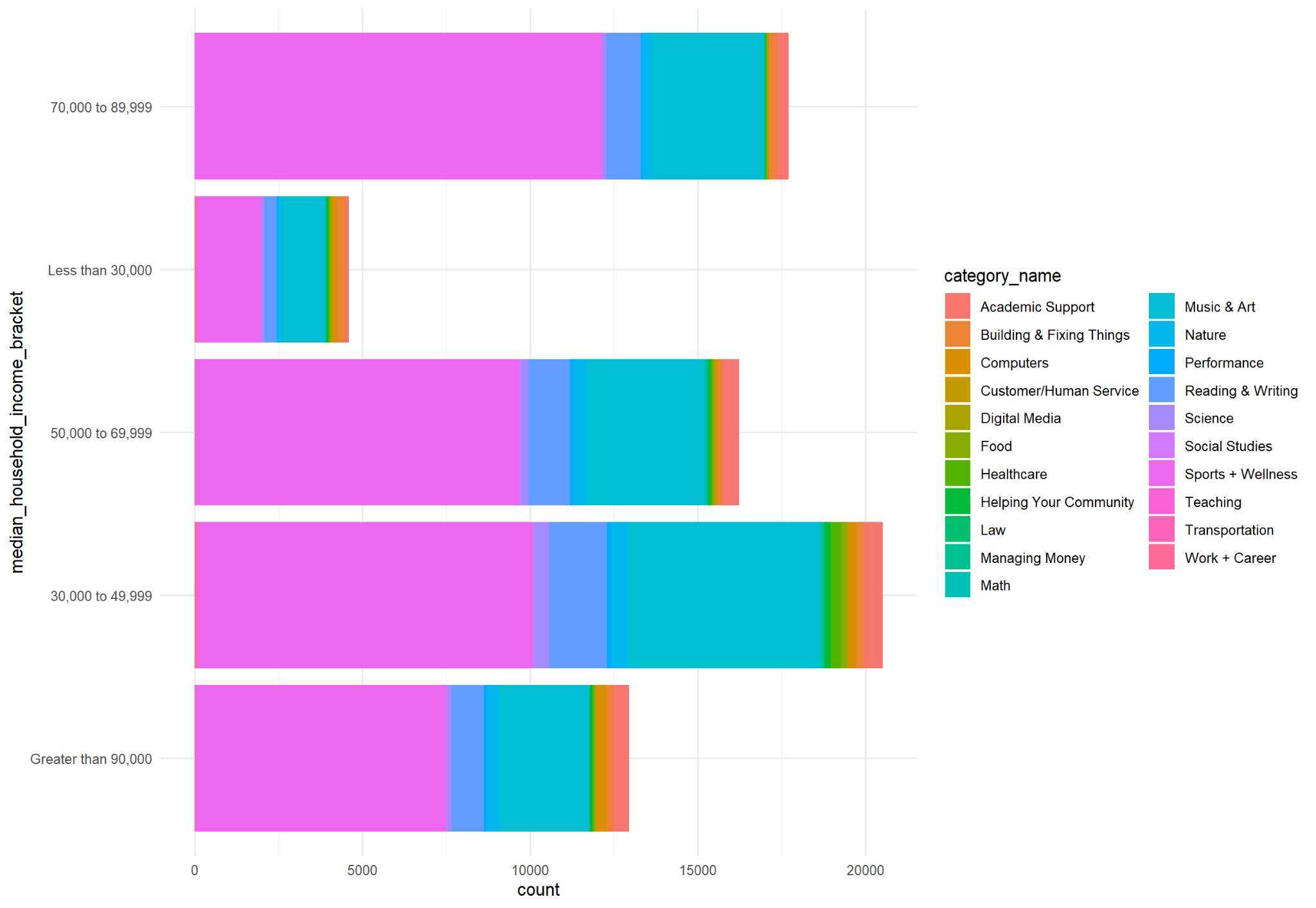
**Plan/Next Steps**:

* Begin work on identifying and merging a dataset containing educational performance of different chicago public schools within different neighborhoods
* Produce summary univariate, bivariate, and multivariate tables by our presentation time next Friday (5/5)
  + ~~(Univariate) Summary statistics of diversity indices – convert these numbers into levels that can be explained~~
  + ~~(Bivariate) Summary statistics of capacity levels for programs, capacity to neighborhood ratio~~
  + (Bivariate) Length of programs by geographic cluster and categories
* Produce a couple more descriptive plots identifying trends/patterns in the features of the dataset by next Friday (5/5)
  + ~~Pie chart/bar chart of program price range~~
  + ~~Pie charts of programs~~
  + ~~Geographical map of all programs in the Chicagoland area~~
* Discover possible correlations in different feature variables and address them by next Friday (5/5)
  + Correlation plot between features
    - Then pinpoint with scatterplot the relationship of those features
* Analyze missing or incomplete data, if any, by next Friday (5/5)
  + Plot of missing data percentages of each feature/column
* Complete full presentation slides by next Friday (5/5)

*Appendix*

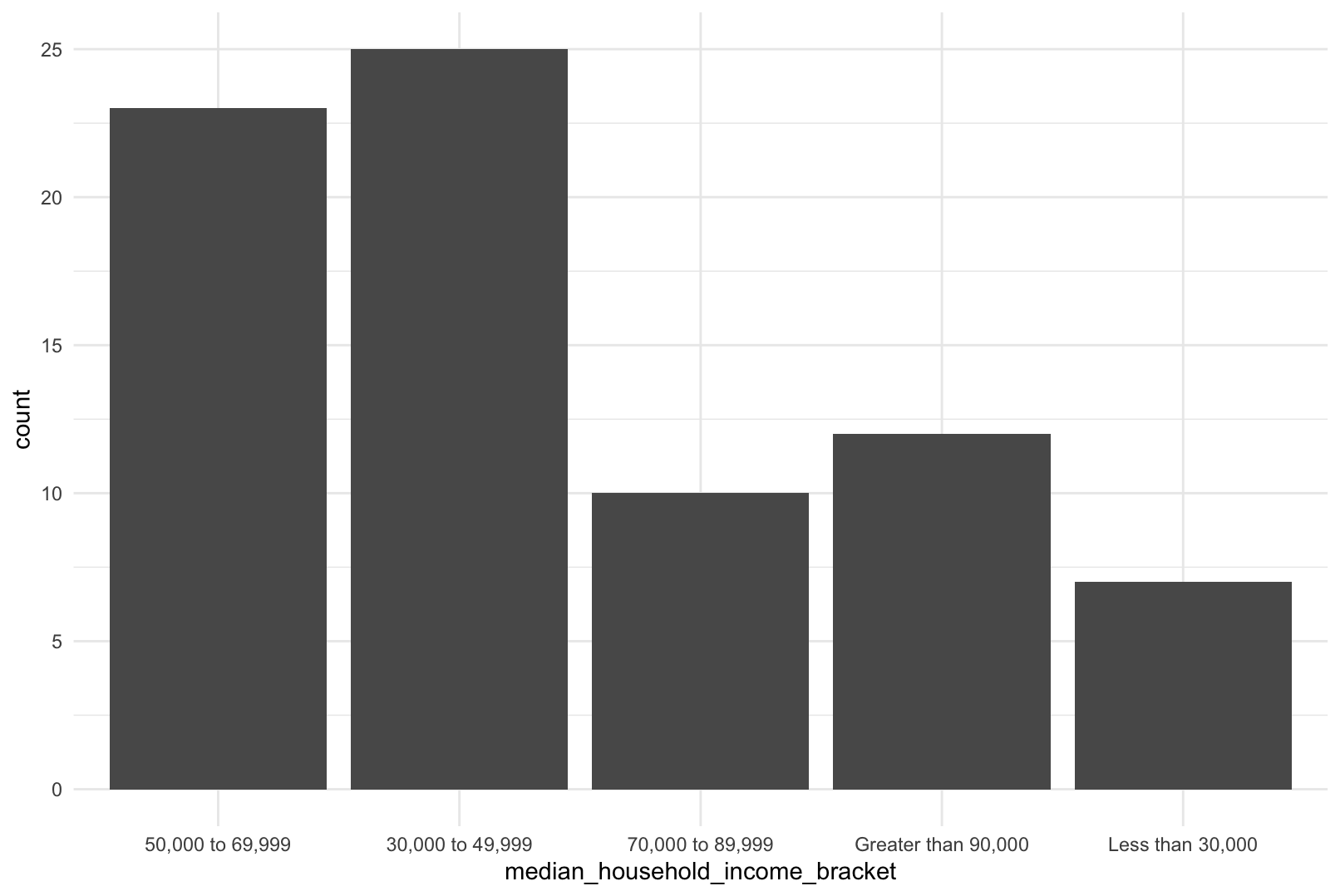
*Plot 1*

*Number of programs by income and category*

**

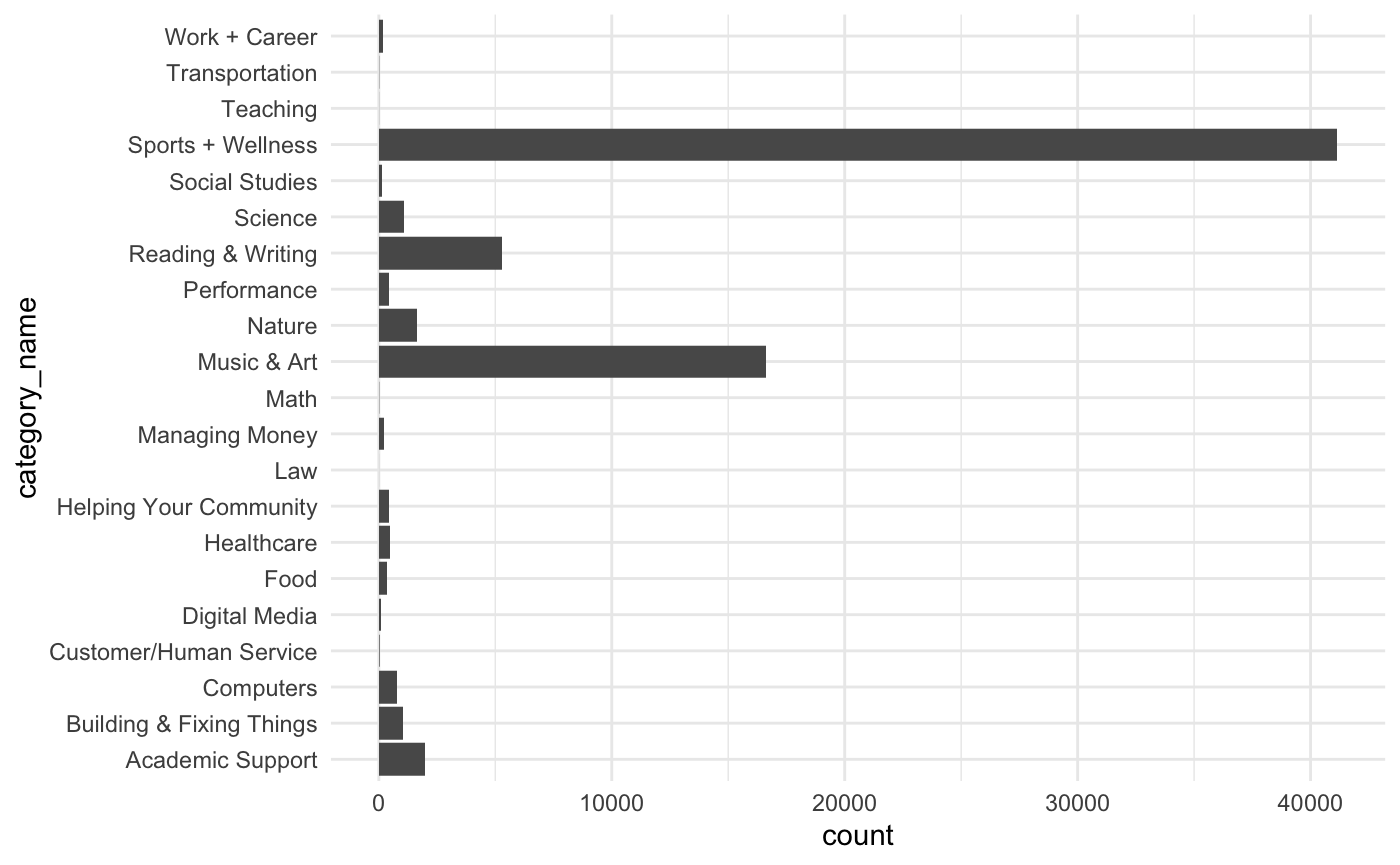
*Plot 2*

*Income Breakdown of Chicago by Community Area*

**

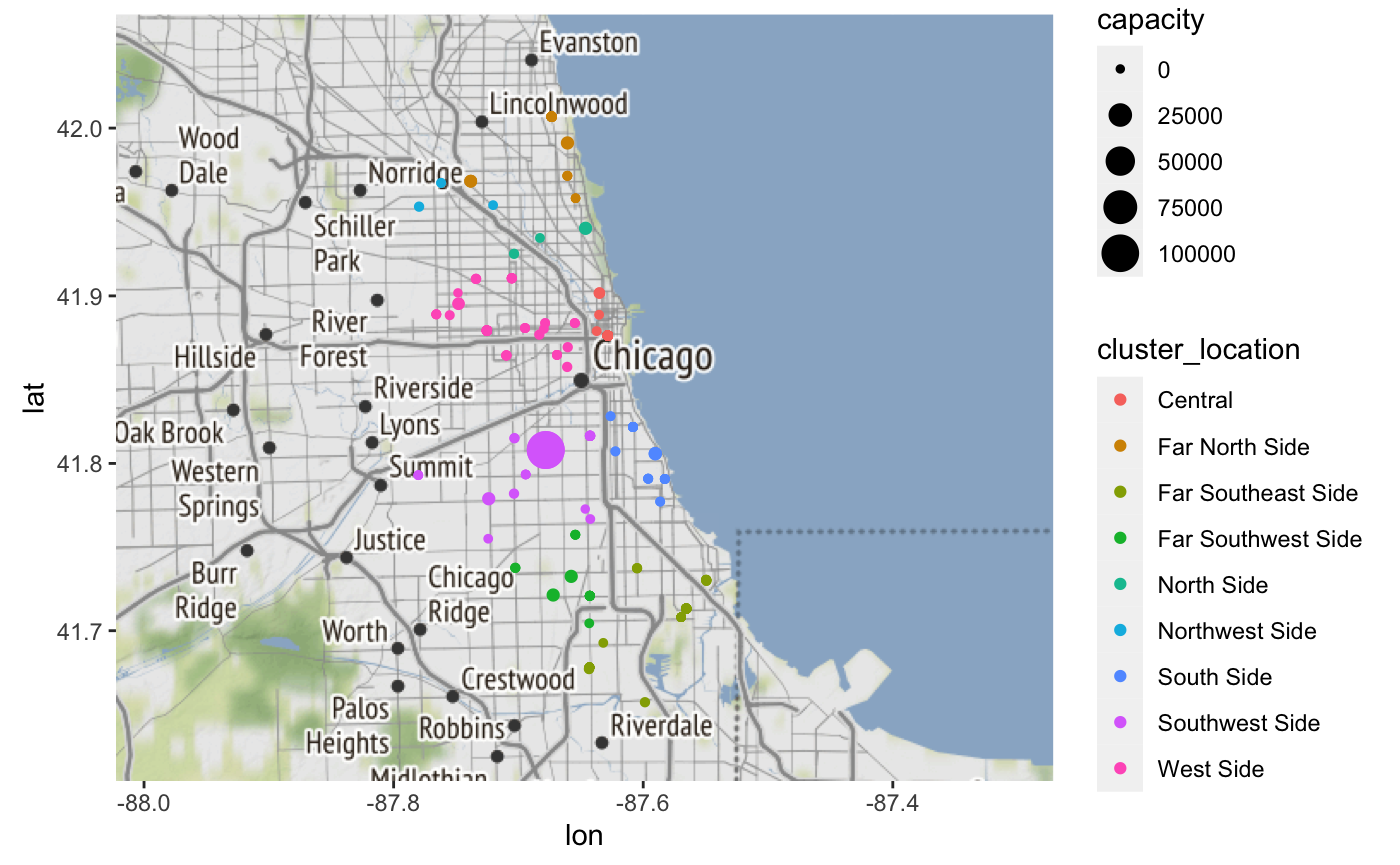
*Plot 3*

*Number of programs by category*

**

*Plot 4*

*Computer Programs Offered by Capacity and region*

**