THE SIMPLE FEATURES PACKAGE

Kamal Abdelrahman[1]

[1] City University of New York - Brooklyn College

Author Note

Kamal Abdelrahman is an undergraduate at the City University of New York - Brooklyn College in Brooklyn, NY majoring in psychology with a focus in statistical programming.

Correspondence concerning this article should be addressed to Kamal Abdelrahman, Brooklyn, NY. E-mail: kamalabdel97@gmail.com

Abstract

R is a statistical programming software that empowers its users with the ability to process, analyze, and present data. One of the key tools allow it to acheive this is through packages. Each package is comprised of various tools that serve a particular purpose. That could include data processing, visualization, and statistical analysis. This study will explore the sf (simple features) package in R to make a choropleth map. Given its user friendly interface, and integration with ggplot2, the grammer of graphics for data visualization, it should allow users to fell confident to develop this skill.

Keywords: R, simple features, geospatial, data visualization

Word count: X

THE SIMPLE FEATURES PACKAGE

Introduction

Data visualization is a key component in being able to communicate information to an audience. (Keim, 2002) Common forms of data visualization include bar charts, scatter plots, and among others courtesy of the ggplot2 package (Wickham, 2016) to help analysts understand the circumstances of their studies. Spatial data visualization has also been a key component in building off of that same aspect, especially with the sf package. (Pebesma, 2018) The various capabilities of being able to explore not just what is going on in a scenario, but also the ability to provide that with spatial context. That advantage with spatial data could allow analysts and users to visually connect areas of need within a given study area.

Methods

The sample in this study was pre-determined by the standardized number of community districts across New York City. The income levels across New York City contains data on both the borough and community district level as well as data across four household types spanning from 2005 to 2017. For the purpose, of simplifying the demenstration, the data was subsetted to the community district for only “All Households” types during the year 2017.

Participants

The Median Income dataset contains 3381 rows of data and six columns of data. The six columns are Location, Household Type, TimeFrame, DataFormat, Income, and Fips. The Location column identifies the both the borough and community district level locations throughout New York City, but only the community district values were used. The

Household Type indicates the type of household that was analyzed; All households, families, families with children, families without children. Only the "All households level was used in this demonstration. The TimeFrame column indicates the year the of the record, spanning from 2005 to 2017. This demonstration only used 2017. DataFormat indicates the unit, which in this dataset would be dollars to represent money for income. Income indicates the median income value throughout a particular area. Fips is the identifying value for each area. Only Fips that have 3 numbers were used in this demonstration were used to as a unique identifier of the community district level.

Material

For this demonstation, R Studio, the sf package, the ggplot2 package, the community districts shapefile, and the income levels for each community district are needed. The R Studio Integrated Development Environment (IDE) provides the platform to run the analysis. The sf package is used to process the spatial data. The ggplot2 package will be used to access data visualization capabilities. The community district shapefile will be used to identify the areas of interest for New York City. The income levels will be used to shade in the geographical boundaries in proportion to the values.

Procedure

To conduct the geospatial analysis of creating a choropleth, the following steps were taken. First, the sf and ggplot2 packages were loaded into R. Next, the shapefile of New York City's community districts was loaded into R with the read_sf function from New York City Open Data Portal. The shapefile could be loaded into R from user's computer filepath as well as straight from the website. If the user is loading the shapefile from their computer's filepath, they must download the shapefile. The downloaded folder must be unzipped. The

shapefile must be kept in the same folder of the files it is located in, regardless if user keeps it in its original folder or moves it to another folder. If the user decides to load the shapefile from straight from the website, they must copy the link to the GeoJson version into the read_sf function. The dataset containing the income levels was downloaded from the Citzens' Committee For Children of New York website. To reproduce this analysis, the user must unzip the folder containing the dataset and load the csv containing the data into R with the read.csv function.

The dataset for the Median Income levels contains data spanning from 2005 to 2017, as well as data on both the borough and community district level, and four different household types (All Households, Families, Families With Children, and Families Without Children). The data was subsetted to only view areas of a Fips code with 3 characters, only the "All Households" Household type, and the year of 2017. Fips codes with only 3 characters indicates the community districts, as opposed to the boroughs that have 5 characters. The Income dataset was merged with the shapefile to combine attribute data with the spatial data. Since the income dataset and the shapefile have different column names for the Fips codes, they were merged together by their respective coloumn names ("boro_cd" from the shapefile, "Fips" from the income dataset). There were non-matching identifiers given that the Income dataset did not have data for (number) of the total Fips across New York City. They were retained in the merged.

To visualize the choropleth, the `ggplot` function was called to activate the ggplot2 syntax. The `geom_sf` function was called to load in the shapefile dataset containing both the shapefile and fill the community district boundaries by the value of their correpsonding income levels. The choropleth was further editted to change colors to accurately represent the analysis being presented.

Results

The choropleth map indicates the income levels across New York City. Thus, individuals are able to spatially reference the income levels across New York City. Though, this possible could have been acheived with a bar chart or even a pie chart, but then they take away that dynamic spatial reference. The choropleth is shaped to New York City, and is shaded in relating to the density of the data. So areas of higher income are darker green, areas of lower incomes are light gray. The darker gray values indicate values that do not have data for them. By analyzing the map, analysts can observe there is a greater cluster of higher income community districts in lower Manhattan. In the lower Bronx, there is a greater clustering of lower incomes.

Discussion

The sf package in R provides a multitude of ways for analysts to process geospatial data. This opens upon different avenues in the field of data science, especially designing predictive models of how income levels may continue to disperage across New York City. This would allow policymakers to truly have a more proactive approach to understand where exactly these disparities are occuring. For the user, there great deal of tools at their disposal with the sf package. They include converting non-spatial data frames to spatial data frames. For example, a dataset containing the Longitude and Latitude location of universities across New York City, but it is not a feature. That could be accomplished with the sf package.

References

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 1–8.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*. Retrieved from https://journal.r-project.org/archive/2018/RJ-2018-009/index.html

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

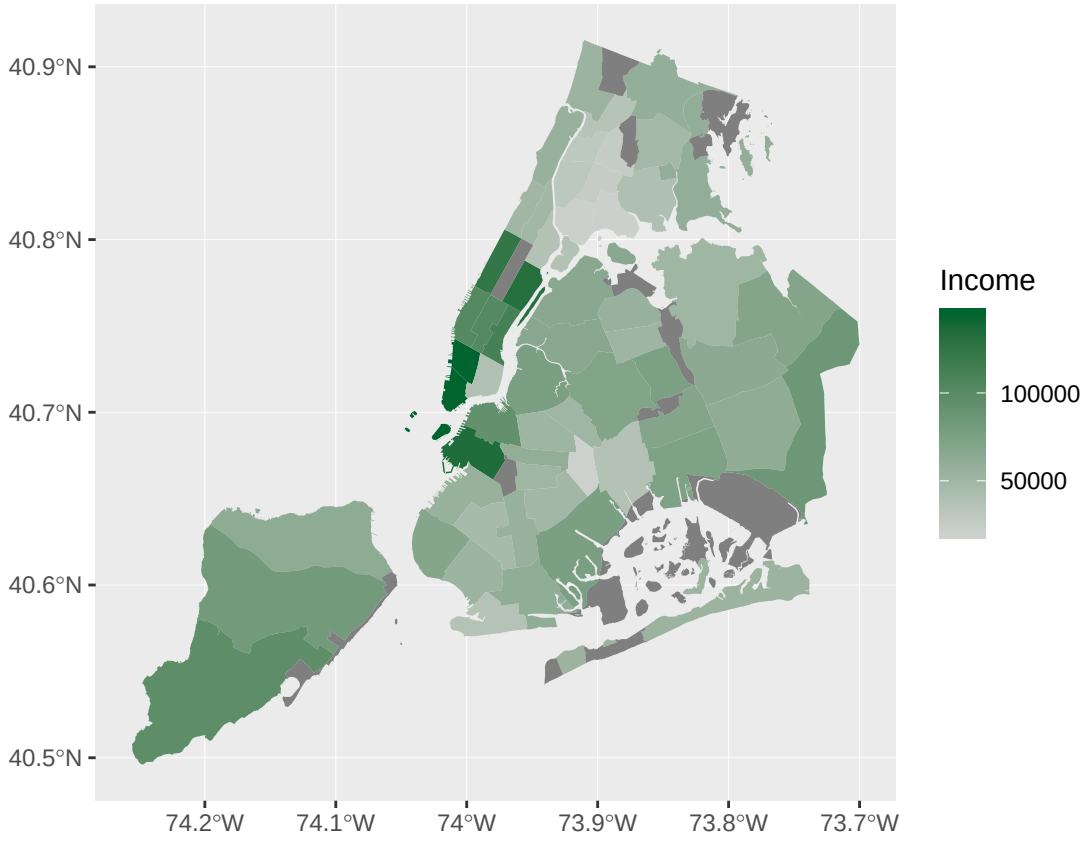

Figure 1. A choropleth map that indicates the income levels across New York City at the community district level