Leveraging Public Data to Enhance Your Analysis

There's more to than meets the eye

Nikhil Bhandari
Rock Creek Analytics, LLC
www.RockCreekAnalytics.com
nikhil@rockcreekanalytics.com

October 26, 2020

Abstract

This article provides an introduction to the various public data sources that exist at the Federal, State, County and local levels that can help enhance the typical data analysis assignment.

Contents

1 Introduction

The various agencies of the US government at different levels (Federal, State, County and local) collect vast quantities of data and make these data available to public. The challenge is to merge the data together to generate meaningful information. In this brief note, we discuss how publicly available data can be used to enhance typical data analytics project.

For the analysis shown below, we use the R software platform though the analysis could be done via a number of software platforms and/or programming language options. We recognize that there are a number of commercial and open-source tools that are much more powerful for specific analyses (especially when it comes to processing, manipulating, analyzing and displaying geographical data) but for the purposes of this article, we will only use one software platform.

We will review both the commonly used data and the less commonly used data. We start with a general map and see what can be added to the map to create something that is, hopefully, more than the sum of its parts. We will use the State of Maryland and Montgomery County, MD, data for all the analysis discussed in this article.

2 Commonly Used Data

2.1 State and County Maps

The US Census Bureau provides cartographic boundary files at different geographic levels (national, county, congressional districts, divisions, metropolitan areas, urban areas, zip code tabulation areas, etc.) in shapefile and KML formats [1]. These files are available for different years and for some cases, different levels of accuracy, and are part of the Census Bureau's MAF/TIGER geographic database. An example of a map created by using one of the census shapefile is shown in Figure 1 that shows the county boundaries for Maryland.

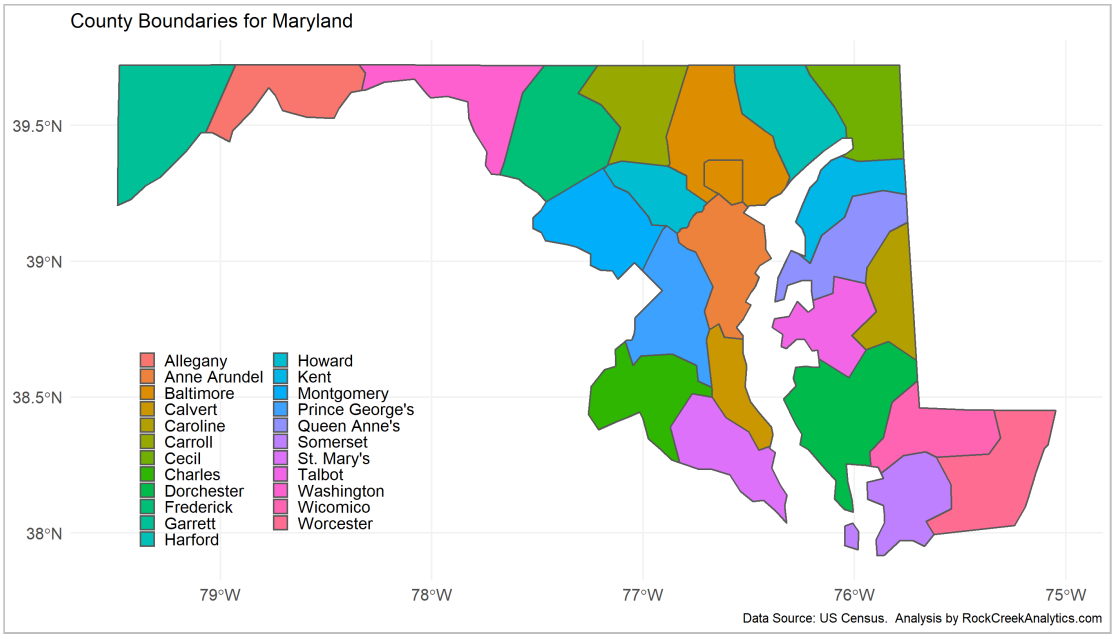

Figure 1: County Boundary Map for Maryland using US Census Shapefile

2.2 Demographic Data

Demographic data generally convers population, employment, income, housing units, etc. and their distribution across a variety of variables such as age, sex, education, family size, etc. The US Census Bureau provides very detailed demographic data to the public via their excellent APIs [4]; this data forms the basis for some of the most common data analyses projects. For example, Figure 2 shows the population and housing characteristics for Maryland counties.

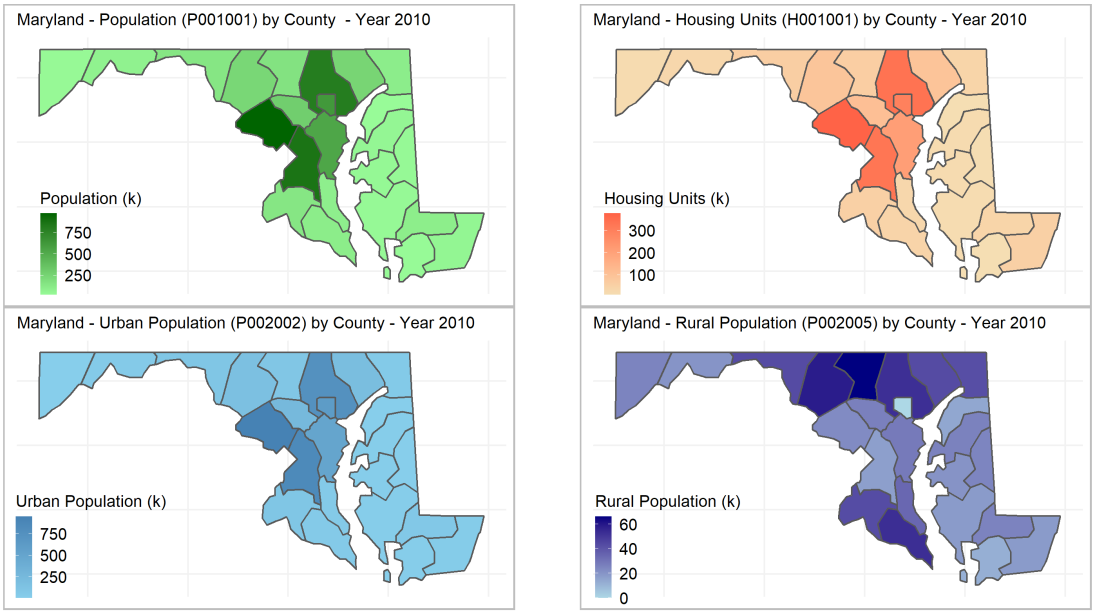

Figure 2: Population and Housing Units in Maryland by County - 2010

3 Less Commonly Used Data

3.1 Parcel Data

The parcel data typically provides information on the parcel boundaries for properties within the county. These data sets are generally accompanied by property tax data that can be quite useful in several different analyses.

The data is managed at a county level and therefore there is some level of inconsistency in how the data is formatted, field names used and the update cycle across the counties in the US. Even with this shortcoming, this is a very useful data source since most counties provide the data in shapefile formats and the data attributes are quite similar, and it is relatively easy to merge data from different counties. Note that these are quite large datasets so the data processing times may vary depending on the computing equipment.

An example of the parcel data is shown in Figure 3 for Montgomery County, Maryland. The data used is downloaded from the County's planning department [2] and the specific dataset is contained in the *property.zip* file [3]. The associated data that comes with this dataset includes: owner's name, address, property tax details such as assessment date, assessment values, property details such as the address, zoning category, type of dwelling, etc., sales details such as transfer date, sale price, etc.

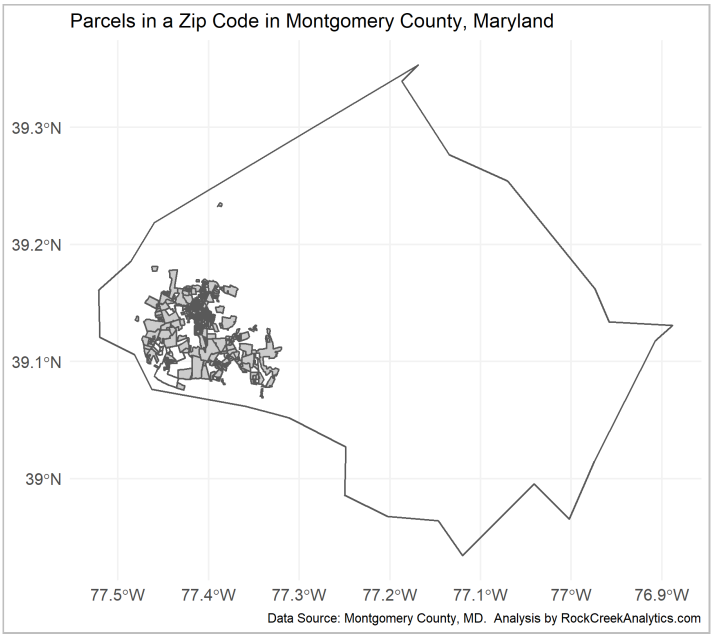

Figure 3: Parcel Boundaries for a Zip Code in Montgomery County, Maryland

To see the usefulness of this data, consider Figure 4 that shows the parcel boundaries on one street in Montgomery County, Maryland; the total assessed value (the sum of land assessment, improvement assessment and preferential land assessment values) and the land use type of the property are displayed in the two panels of the figure.

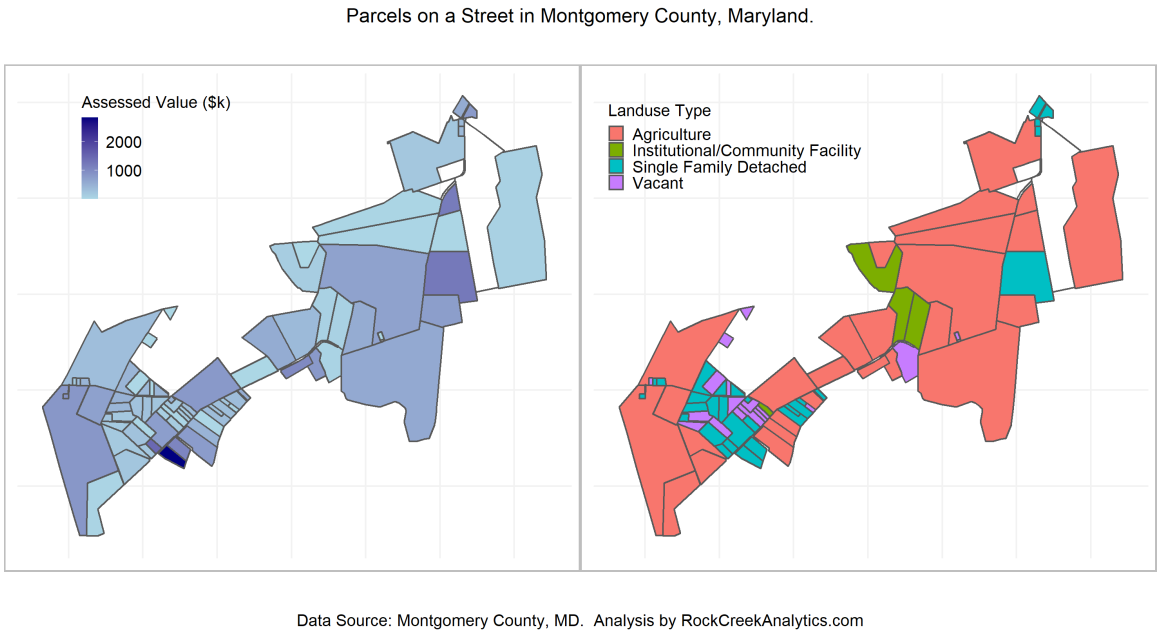

Figure 4: Parcel Boundaries for Properties on a Street in Montgomery County, MD

3.2 Police Data

The police departments across the country provide valuable crime incident data which can be used for a variety of purposes. Data elements typically include information on the incident such as address, date and time of the dispatch, type of incident, police station information, etc. An example of such data from Montgomery County, Maryland [5] is shown in Figure 5 that shows the incidents for the year 2020 by dispatch time.

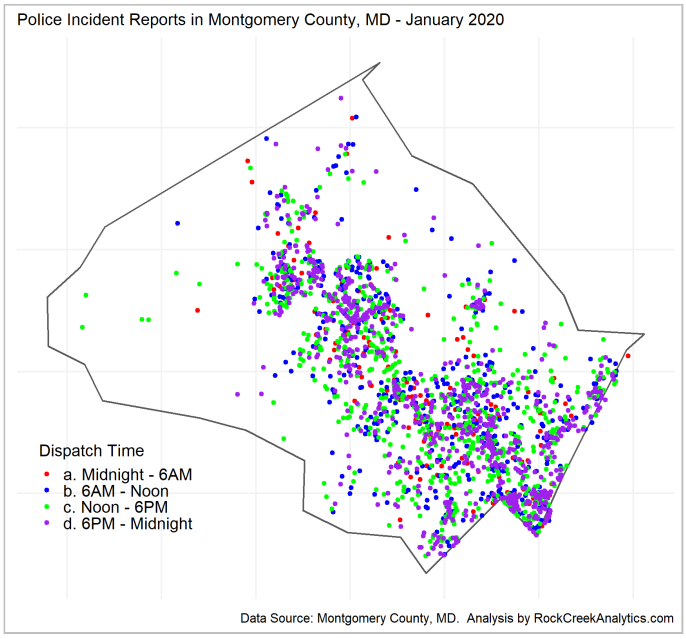

Figure 5: Police Incident Reports in Montgomery County, Maryland - January 2020

3.3 Construction Data

There is significant amount of construction related data available from the counties. Figure 6 shows the locations for the commercial building permits issued in the year 2020 in Montgomery County, MD [6]. Additional data elements include street address of work location, permit dates, building area, declared valuation, description of work, work type, type of structure, etc.

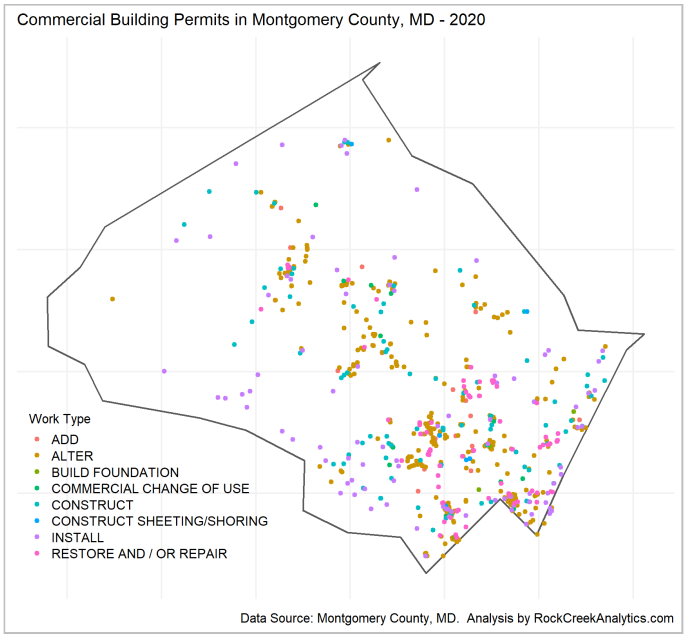

Figure 6: Commercial Building Permits Issued in Montgomery County, Maryland - 2020

3.4 Other Data

There are several other data elements that we did not explore in this article. These include structure data, meteorological and weather data, transportation data, financial and insurance data, etc. Depending on the objectives, one can add this data to their approach and create a more robust and complete analyses.

4 Closure

This article has provided an introduction to the various public data sources that exist at the Federal, State, County and local levels that can help enhance your data analysis. The data is often in somewhat different formats, and additional steps have to be performed so that the data is ready to be combined; the analyst needs to be especially careful about the geospatial data to make sure that the same reference system is used. At times, some geocoding is necessary to convert address data to geographical coordinates.

References

[1] US Census Bureau. *Cartographic Boundary Files - Shapefile.* https://www.census.gov/geographies/mapping-files/time-series/geo/carto-

boundary-file.html.

[2] Montgomery County Planning Department. *GIS and Mapping - Data Downloads*. https://montgomeryplanning.org/tools/gis-and-mapping/data-downloads/.

[3] Montgomery County Planning Department. *Montgomery County Property Data File*. https://mcatlas.org/tiles/00_Shapefiles/property.zip. Data downloaded on 24 October 2020.

[4] US Census Bureau. *Decennial Census*. https://www.census.gov/data/developers/data-sets/decennial-census.html.

[5] Montgomery County Government. *Crime Incident Map*. https://data.montgomerycountymd.gov/Public-Safety/Crime-Incident-Map/df95-9nn9?referrer=embed. Data downloaded on 25 October 2020.

[6] Montgomery County Government. Data for all *Commercial Building Permits* issued since 2000, including status and work performed. https://data.montgomerycountymd.gov/Property/CommercialPermits-API/98z8-bqz4. Data downloaded on 25 October 2020.