**Introduction**

**Background**

Baltimore is the 30th largest city in the United States with a population of approximately 600,000 people. Baltimore, together with Washington, D.C. make up the fourth largest combined statistical area in the country and have a combined population of about 10 million people.

**Problem**

Baltimore has been dubbed the "city of neighborhoods." Baltimore's neighborhoods have rich histories, but along with rich tradition, they also reflect a legacy of segregation and a lack of resources for communities of color. As a resident of this city, I have decided to dive into what type of services are available in which locations and juxtapose these findings with demographics, income, and available services.

**Data**

**Sources of Information**

I began by collecting the most recent available census data, collected in 2010, via the **Open Baltimore Portal** (Baltimore, n.d.). The data is organized in such a way to highlight demographics, including percentage of the population by sex, race, age, household attributes, and income. I have also incorporated a separate dataset, the Maryland Baltimore City Neighborhood (geodata.md.gov, n.d.) dataset, available on the **Maryland Open Data** platform.

I have also incorporated the **Foursquare API** (Foursquare Developers, n.d.) to gather what types of services are available across the city, such as grocery stores, cafés, and farmers markets.

**Cleaning & Normalization**

My next step was to iterate through the neighborhood dataset using Google Geocoder/ArcGIS API to fetch the latitude and longitude for each neighborhood. The neighborhood data was then cleaned by hand picking locations on the map that didn’t represent an actual neighborhood, such as the *Druid Hill Park* and *CARE* locations.

The Foursquare API was then utilized to create another data frame to store the nearest venues to the neighborhood coordinates. I then used one hot encoding to normalize the data. I then fetched the ten most common venues for each neighborhood and stored that in a separate data frame.

*Note: The census dataset required no further cleaning or normalization. In future updates to this analysis, however, I think that I will split the neighborhoods because some of the neighborhoods appear to have been grouped together.*

**Methodology**

**Data Analysis**

The target variable would be venues and services. What services are available to communities, and whether income and race inform/indicate what services are available in a given neighborhood is the basis of this study.

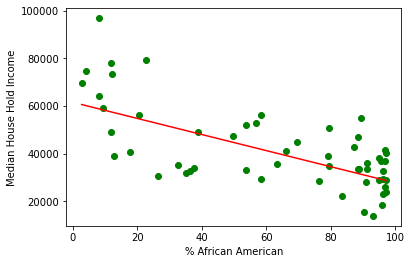
As you will soon see, racial and economic boundaries are self-evident in the data, but to what degree can they be used to predict available services in a neighborhood?

**Predictive Modeling**

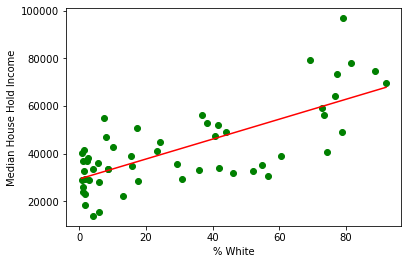
I chose to explore linear regression in analyzing income and race in these neighborhoods and k-means clustering to examine the relationship between available venues.

**Results**

I looked at percentage by race as an indicator of median household income, for African Americans and Whites. I used simple linear regression to analyze the data and displayed it on a scatterplot. I also calculated the r-squared values and annotated them below:

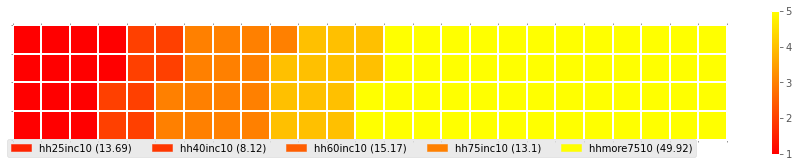
****

R2-score: 0.33

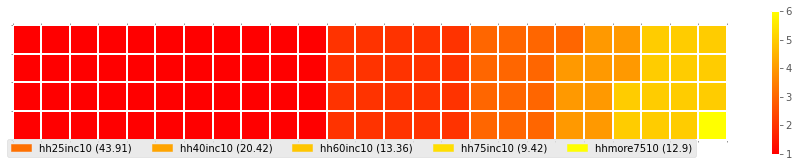


R2-score: 0.57

Next, I looked at the predominantly white neighborhood, Canton, versus the predominantly African American neighborhood, Reservoir Hill, and broke down the following income categories which can be visualized in the following charts. Most Canton respondents reported earning $75,000 or more annually, about 50%, while most Reservoir Hill respondents reported earning $25,000 or less annually, about 44%.

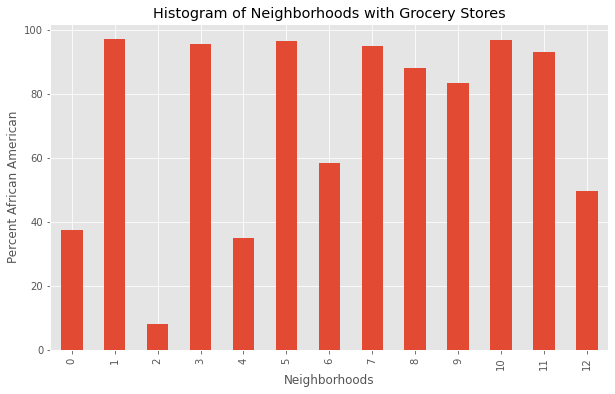


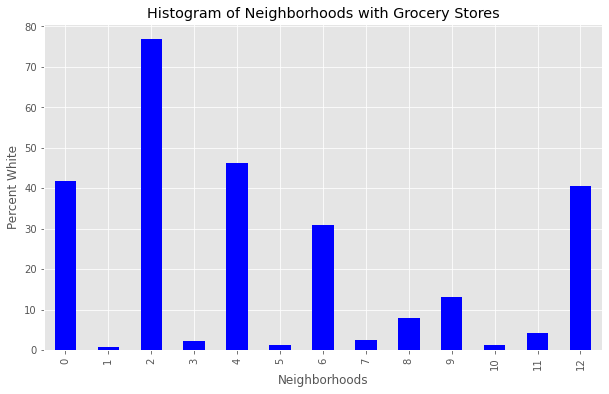
Canton (80% White)



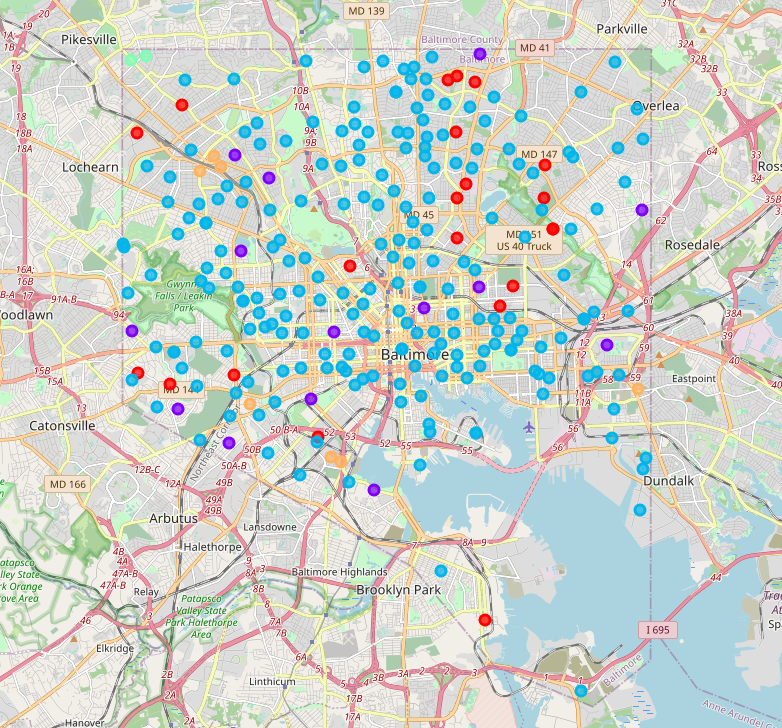
Reservoir Hill (90% African American)

I then created a list of neighborhoods that contained grocery stores from the venue data frame. I matched the neighborhood data in the census data frame and found information on thirteen neighborhoods that matched. I then created the following two bar graphs to reflect the percent of African American respondents and the percent of White respondents for each neighborhood that matched to have a grocery store.

****

****

I also created a map to represent the k-means clustering for the venue data.



**Discussion**

The first set of plots indicate that neighborhoods with a high percentage of African Americas generally have a lower median household income. Conversely, neighborhoods with a greater concentration of Whites, generally have a higher median household income. However, the regression shows that this isn’t a strong indicator in either situation.

Surprisingly, it appears that neighborhoods that contain a grocery store are more likely to have had African American respondents. In the future, I would like to continue to run through these steps with additional venue categories such as farmers markets, drug stores, schools, liquor stores, etc. and continue to compare the data.

The k-means clustering didn’t appear to suggest anything substantial except perhaps to highlight areas of higher activity.

**Conclusion**

The data appears to bear out that household income and race do not necessarily have a direct relationship with the types of services that are available in each neighborhood in Baltimore City.

This is preliminary and I plan to return to this and add additional analysis over the coming weeks and months as I continue to find more insights and sources of data. However, I am publishing this now to satisfy a requirement for the IBM Data Science professional certificate, available on Coursera. Please reach out to me at [m.melend@gmail.com](mailto:m.melend@gmail.com) with any suggestion or comments.