ALLEGHENY COUNTY (PITTSBURGH) – NEIGHBORHOOD SEGMENTATION

Kelly Rhoton
April 13, 2019

I. INTRODUCTION

Background

Pittsburgh remains one of the most affordable cities in the United States. When families are looking for an area to purchase a home, they use many factors such as school district rating, affordability, walkability, and the proximity to venues like restaurants, parks or libraries. I will use various datasets to segment Pittsburgh. I will first segment Pittsburgh by North Hills, South Hills, East, West Hills and the City of Pittsburgh. Are there areas in each group that are comparable? (Similar school district ranking, average lot/home size, average home price per sq. foot, average age of homes, types of homes (single family, apartments, etc.), etc.)

Problem

The desired result of this project is to cluster/segment the four suburbs and the city of Pittsburgh to find neighborhoods or census tracts that are most comparable.

Interest

When a family moves to the Pittsburgh area who is unfamiliar with all the suburbs/neighborhoods, there are many websites that provide details, but I was not able to find one that compared neighborhoods from different suburbs

II. DATA ACQUISITION AND CLEANING

Data Sources

i. **Allegheny County Property/Real Estate Sale Information (AC_PROPERTY)**

- Allegheny County has parcel information here : https://catalog.data.gov/dataset/allegheny-county-property-sale-transactions This includes every property in Allegheny County with address, property type, sale information, etc. as well as details about the property like whether it is residential or commercial. This also includes building information like how many bedrooms, bathrooms, stories, etc.

	PARID	PROPERTYHOUSENUM	PROPERTYFRACTION	PROPERTYADDRESS	PROPERTYCITY	PROPERTYSTATE	PROPERTYUNIT	PROPERTYZIP
0	0001M00009000000	247.0		FORT PITT BLVD	PITTSBURGH	PA		15222.0
1	0001M00010000000	7.0		WOOD ST	PITTSBURGH	PA		15222.0
2	0001M00013000000	14.0		WOOD ST	PITTSBURGH	PA		15222.0
3	0001M00018000000	1.0		SMITHFIELD ST	PITTSBURGH	PA		15222.0
4	0001H00327110200	300.0		4TH AVE	PITTSBURGH	PA	UNIT 1102	15222.0

5 rows × 86 columns

- Because I needed the longitude and latitude information, I found parcel centroid information to get the LAT and LNG and the census tract information here https://data.wprdc.org/dataset/parcel-centroids-in-allegheny-county-with-geographic-identifiers/resource/4b68a6dd-b7ea-4385-b88e-e7d77ff0b294. This gives the census tract for each parcel as well as the X,Y coordinates and other identifying information like districts for elections.

	PIN	MAPBLOCKLO	x	y	STATEFP10	COUNTYFP10	TRACTCE10	BLOCKCE10	GEOID10	NAME10	...	level_SenatePA
0	0460E00021000002	460-E-21-0-2	-79.831190	40.361378	42	3	551900	1000	420036000000000	Block 1000	...	PA Senate
1	0455D00224000001	455-D-224-0-1	-79.818869	40.413092	42	3	509400	1000	420035000000000	Block 1000	...	PA Senate
2	0455D00224000002	455-D-224-0-2	-79.818869	40.413092	42	3	509400	1000	420035000000000	Block 1000	...	PA Senate
3	0013J00077000000	13-J-77	-79.982418	40.417202	42	3	481000	1000	420035000000000	Block 1000	...	PA Senate
4	0100L00375000000	100-L-375	-80.072533	40.389166	42	3	470300	1000	420035000000000	Block 1000	...	PA Senate

5 rows × 47 columns

ii. Demographic Information (CENSUS)

- The census bureau allows you to query very detailed datasets here: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_17_5YR_DP03&prodType=table
  - I pulled data by census tract for employment status, household income and average family income groupings

	GEO_KEY	GEO_ID	CENSUS_TRACT	EMPLOYMENT STATUS	IN_LABOR_FORCE	IN_LABOR_FORCE_PCT	NOT_IN_LABOR_FORCE	NOT_IN_LABOR_F
0	1400000US42003486700	42003486700	4867.0	1606	996	62.0	610	
1	1400000US42003101600	42003101600	1016.0	684	345	50.4	339	
2	1400000US42003982200	42003982200	9822.0	4644	1482	31.9	3162	
3	1400000US42003422000	42003422000	4220.0	3975	2401	60.4	1574	
4	1400000US42003563300	42003563300	5633.0	1120	736	65.7	384	

5 rows × 59 columns

  - I also pulled the racial profiles of each census track to determine how diverse each area is.

	GEO_KEY	GEO_ID	CENSUS_TRACT	TOTAL_POPULATION	WHITE_ONLY	BLACK_ONLY	AMERICAN_INDIAN_ALASKA_NATIVE_ONLY	ASIAN_TC
0	1400000US42003010300	42003010300	103.0	6546	4299	1738	27	
1	1400000US42003020100	42003020100	201.0	4823	3845	313	30	
2	1400000US42003020300	42003020300	203.0	943	816	59	0	
3	1400000US42003030500	42003030500	305.0	2177	187	1704	22	
4	1400000US42003040200	42003040200	402.0	1742	720	751	0	

- I was able to find by census tract, the number of renters vs. home owners here https://data.wprdc.org/dataset/allegheny-county-homewonership-and-rentals/resource/85ec763a-457e-4025-8263-112baeba0fdb
- I also found walk scores by census tract which grades each area by how walker friendly the area is. It can be found here: https://catalog.data.gov/dataset/allegheny-county-walk-scores

iii. School District Ranking (SCHOOL_RANK)

- I used the school digger API (www.schooldigger.com)to find the top school districts in Allegheny County. I created custom tiers of each district. Some of the parcels did not have a matching school district so I lumped them into school tier 5.

```
#Create school tier classification
def school_class(row):
    if row["Rank"] <= 7 :
        return "1"
    elif row["Rank"] >=8 and (row["Rank"] < 15):
        return "2"
    elif row["Rank"] >=15 and (row["Rank"] < 25):
        return "3"
    elif row["Rank"] >=26 and (row["Rank"] < 30):
        return "4"
    else:
        return "5"
```

	Rank	DistrictName	City	Zip	County	SCHOOL_TIER
47	1	Upper St. Clair Area	Pittsburgh	15241	Allegheny	1
27	2	Mt. Lebanon	Pittsburgh	15228	Allegheny	1
18	3	Fox Chapel Area	Pittsburgh	15238	Allegheny	1
44	4	South Fayete Township	South Fayete Township	15057	Allegheny	1
33	5	Pine-Richland	Gibsonia	15044	Allegheny	1

Data Cleansing

For the most part, the data was pretty clean because most of my sources shared a common key the Census Tract ID (GEO_ID). I did have to make some adjustments.
School District names did not match up between the county website and the school digger ranking so I changed the school digger data so it matched the county data. Census Tract data was missing for some of the parcels in the real estate file. I dropped them because there was just a small amount.

Feature Selection

Data Set	Kept	Dropped	Reason
AC_PROPERTY	LOTAREA, FINISHEDLIVINGAREA, SALEPRICE,USECODE,USED ESC,PARID	Everything Else	I needed to calculate the price per square foot by property type
AC_CENTROID	geo_id_tract, Group_Neighborhoods		I just needed the GEO_ID and Neighborhood to join to the property assessment details. Later I pulled the coordinates for the GEO_ID
SCHOOL_RANK	Rank, DistrictName	Everything else	I only needed the school rank
CENSUS_INCOME		Everything else	I only wanted high level information about income and didn't want to look at industry or any other demographic information
CENSUS_RACE	Rent_PCT, Own_PCT	'TotalPopulation', 'OwnedMortgage','OwnedFree', 'TotalOwned', 'RenterOccupied'	I calculated the PCT and dropped the details
CENSUS_WALKSCORE	GEO_ID,CENSUS_TRACT,WALK_SCORE	N/A	N/A

III. METHODOLOGY

Clustering property information

I used K-Means clustering to segment the census tracts. I first used Foursquare and determined that the county file was actually more accurate and up to date. Foursquare had missing and outdated information. Ultimately, I wanted to look at the mix of property types rather than the types of restaurants and other businesses in the area. Areas with a high concentration of retail/business, were classified as commercial. When I looked at the break down of the types of properties in each cluster, I named them Multi-Family, Commercial, Single Family and Mixed. Then, I charted each area of Pittsburgh to see the distribution.

Table 1 - Cluster Labels

Cluster	Description
0 - Multi-Family	Row houses, condominiums, and some single family homes
1 - Commercial	Office buildings, government building and some homes
2 – Single Family	Mostly single family homes
3 – Mixed	A mix between multi-family, single family

Table 2 - Cluster Distribution

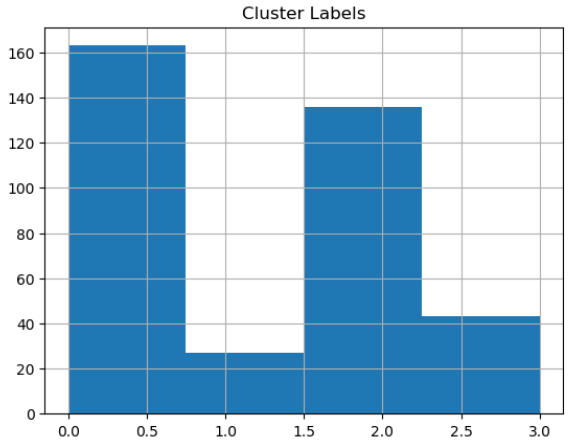
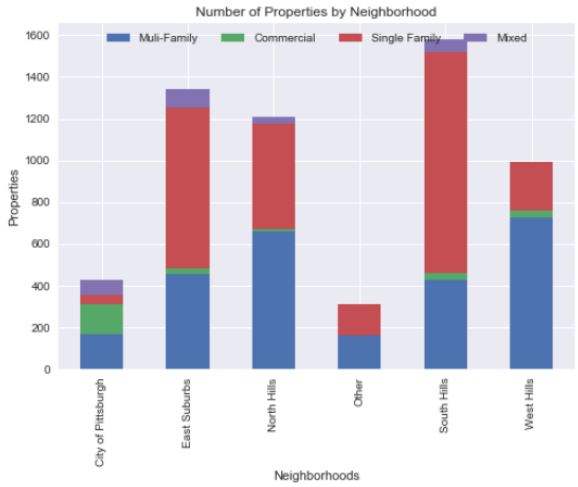
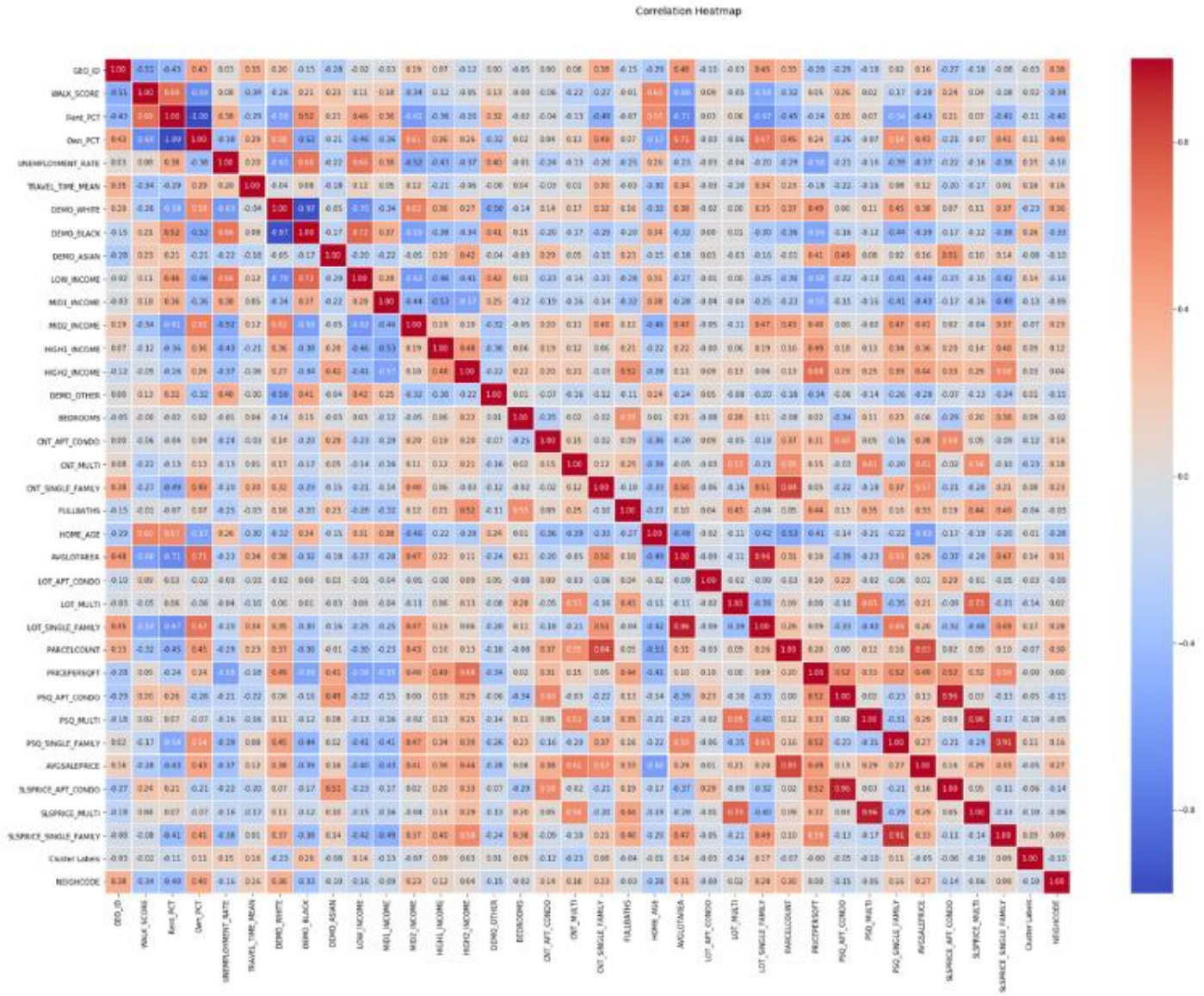


Table 3 - Neighborhood Cluster Distribution



I created this heat map to see if there was any strong correlation between the price per square foot of a home and the other features. There is some correlation, but it was not high enough for me to explore further. (ee Table 4)

Table 4 - Correlation Heat Map

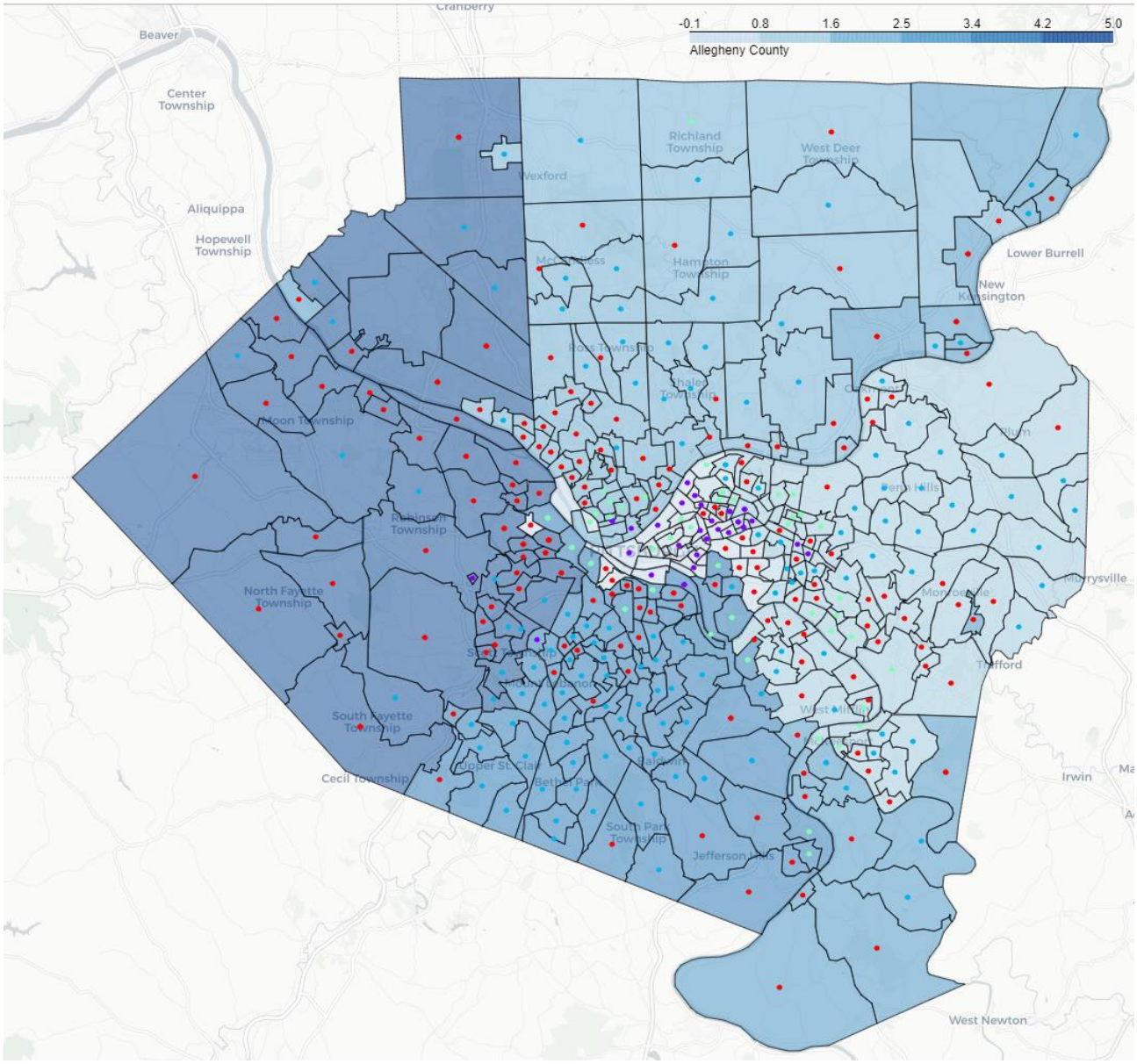


Results

Mapping the Pittsburgh Area – Clustered

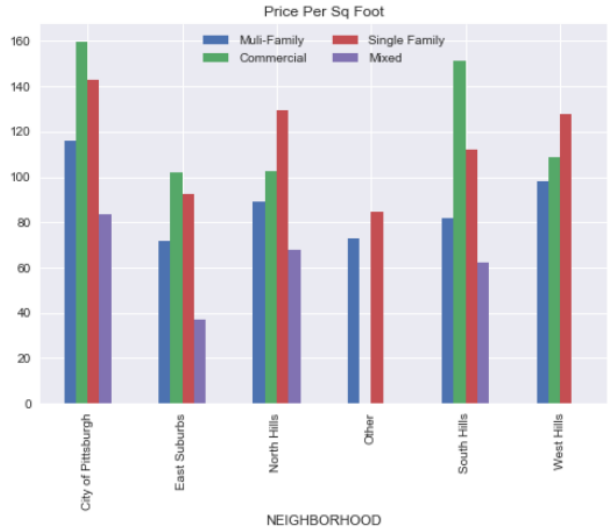
I used Folium to create a map of Allegheny County. It is broken out by the City of Pittsburgh, West Hills, North Hills, East Suburbs and South Hills. Each of those sections is segmented by census tracts outlined in black. When you hover over the cluster dots, it will give the cluster number and description.

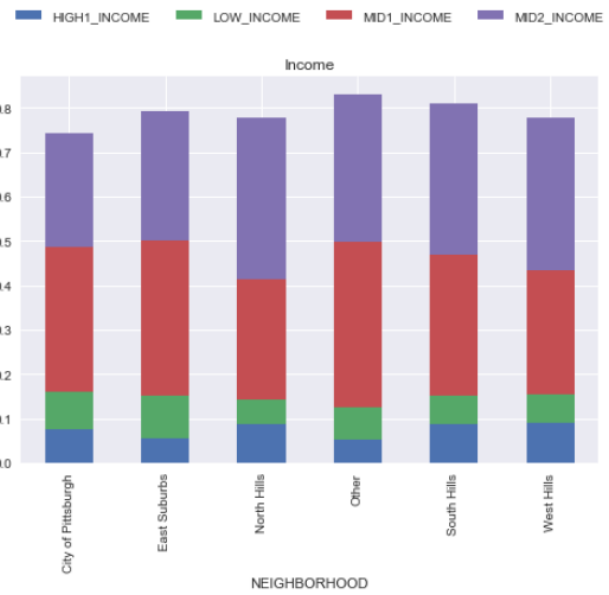
Table 5 - Map of Allegheny County

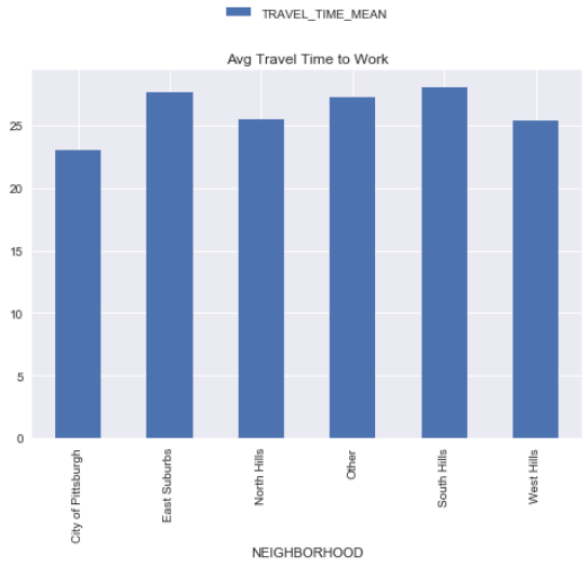


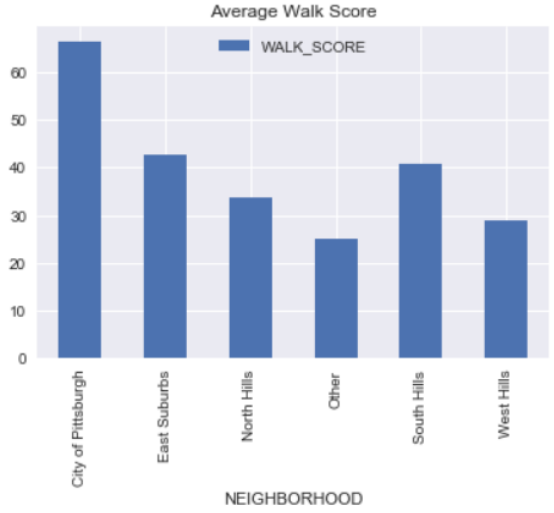
Coursera - IBM Applied Data Science Capstone

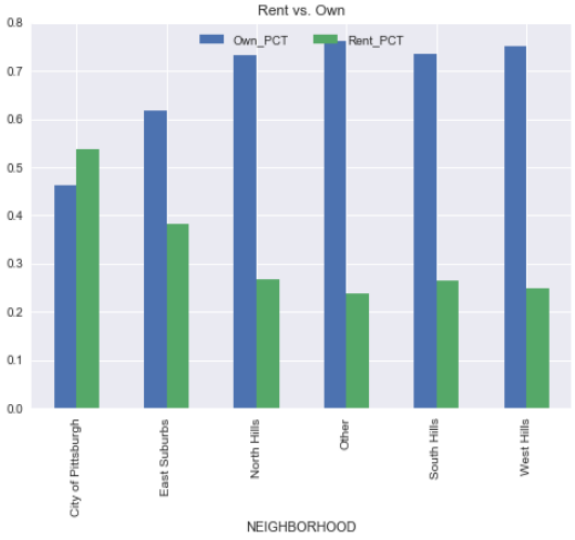
Kelly Rhoton

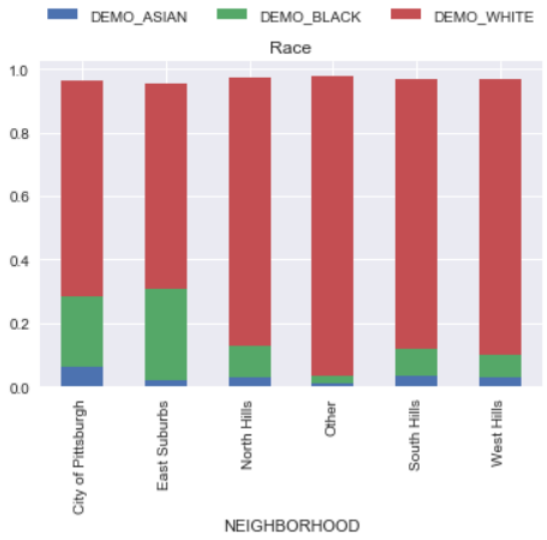












IV. DISCUSSION

Pittsburgh is an affordable city with many neighborhoods to choose from. Even though there is diversity in the type of home, neighborhood and distance from the city, there is not very much racial diversity in any part of town.

The City of Pittsburgh – The city has some of the highest price per square foot which might explain why it also has the highest number of renters. I was surprised to see that the average commute time to work was the same as other neighborhoods. Obviously, the city has the highest racial diversity, walk score and the most commercial clusters. There are neighborhoods within Pittsburgh with really high price per square foot and some with low prices.

West Hills – The west hills' price per square foot is slightly higher than other areas (for single family homes). The school districts are mostly average with the exception of Quaker Valley in Sewickley.

The East Suburbs – Single family home prices are the lowest in the east and is slightly more diverse (racially and income) than the other suburbs. The schools, however, are not ranked the highest.

South Hills – The south hills have the second highest walk score. It has a low percentage of renters and schools that are ranked high (Upper St. Clair & Mt. Lebanon)

North Hills – The north hills has options for highly ranked schools. (Fox Chapel, North Allegheny, etc) It also has a slightly lower commute time.

V. CONCLUSION/NEXT STEPS

If I were to continue this analysis, I would look more closely at the census tract level and add in other features like political party. There seems to be a lot of information available online about each community and I found it hard to stop looking for the "best" data.

VI. GITHUB REPOSITORY/PRESENTATION