**Best Pittsburgh Neighborhoods for New Residential Buildings**

**Chad Dominick**

**March 9, 2020**

**1. Introduction**

**1.1 Background**

The city of Pittsburgh has transitioned from being a blue-collar, industrial town to a vibrant center of education and innovation. The town went from being centered around steel mills to being centered by technology and medicine. The growth in all these areas has made Pittsburgh one of the fastest growing cities in the country. The increase in population has caused many new venues and residential buildings to be built in many neighborhoods. There is a need and opportunity for even more residential buildings throughout the city’s neighborhoods. These apartments buildings can also be utilized for tourists and visitors, which is a specific need for Pittsburgh to host more major events such as the Super Bowl.

**1.2 Problem**

There are numerous factors and data points that can be used to classify the overall nature of each individual neighborhood. Population size, median income, available land, and the venue types in close proximity are all factors that are used to make decisions on residential property development. This project uses these factors along with population demographic information to put Pittsburgh neighborhoods into different clusters and classify them in order to find the best neighborhoods to build new apartment buildings.

**1.3 Interest**

Apartment building developers would be very interested in this project, especially developers form out of town who are not familiar with the dynamics of Pittsburgh neighborhood and are interested in entering and developing the Pittsburgh market. Other groups that would have interest are city officials, citizens, and businesses and venues looking to develop in Pittsburgh as well.

**2. Data**

**2.1 Data sources**

Neighborhood demographics and information data was gathered from multiple datasets from the 2010 Census for Pittsburgh available at <https://catalog.data.gov/dataset/pgh-snap> . I also manually utilized the Google Search Engine to find and input the Latitude and Longitude of each Pittsburgh neighborhood into the dataset. Lastly, venue location and information were pulled in utilizing the Foursquare location data and its Developer API.

**2.2 Data cleaning**

The data downloaded from the Census datasets was relatively clean and usable. There were five neighborhoods that missing data for a majority of all of the fields. In all of these cases the neighborhoods had a very small population and overall area which means they could easily be dropped for this analysis. Venue information was pulled into a separate dataset and then grouped by venue category and neighborhood. The data set then contained the frequency for the top five most popular categories in each individual neighborhood. Both datasets were then joined using the corresponding Neighborhood of both datasets.

**2.3 Feature selection**

Following minimal data cleansing, there were 85 neighborhoods left with each having 43 features in the data. There were eight separate datasets available from 2010 Census data so there many of the features were redundant or exactly the same. There were also many features that referenced data from the 2000 Census. There were also some features that would not pertain directly or be of interest to apartment building development. After removal of these features there were a remaining 20 features in the dataset. Many of these that were numerical were in percentages. I used the total population and area to convert the percentages to totals. Then these features were combined with the top five venue category features for a total of 25 features selected to be used for the analysis. This also excludes the Neighborhood name, Sector, Latitude, and Longitude features that were removed for the analysis and then added back in for the map creation and cluster visualization.