Final Exam - Battle of the Neighborhoods – Week 1 Submission part 2

Data to be Used

For this analysis, the following data is needed, in the form of CSV files that will be downloaded and generated as part of the reproducible source input data for the project:

- File containing text names of Seattle neighborhoods and associated (parent) Districts.
- File containing zip codes of the Seattle neighborhoods.
- Files containing recent Real Estate Housing sales and rents of Seattle neighborhoods.
- Four-Square data of current restaurant venues, including number of direct competitors, in the various Seattle neighborhoods.
- Four-Square data of containing counts of current general venues in a given neighborhood to ascertain additional "magnets" and points of interest that would help increase foot traffic and potential "draw in" business in the area.

In an ideal world, a breakdown of overall retails sales or sales tax revenues by district, would be most valuable, in order to target the highest revenue generating areas, which would be most favorable for a high end upscale restaurant. While the city of Seattle and King County have very good data portals, they currently do not provide this level of detail. Instead, relevant sales revenue and sales taxes are currently only available as aggregate summaries by city, not by district or neighborhood. So as an approximation, housing prices and rent data (by neighborhood) will be used to select those districts and neighborhoods with the highest median values. These will be used to derive the best potential for an upscale restaurant.

The collected data will be used to determine which neighborhoods exist in the Seattle area, what are the income characteristics of each neighborhood, what competitors exist in each neighborhood, which neighborhoods might be being "under-served" (very few offerings or competitors in the area), and what "magnets" does the respective neighborhood have to draw people into the area. This information will be used to rank and select the top ranked 20 % of the neighborhoods. From there, additional analysis will be performed to whittle do the recommended locations to a "top 10" list.

Data Sources to be Used

The following data sources will be used by the project:

- Text Names of Seattle neighborhoods and districts – screen scraped via Python BeautifulSoup package from the following URL:
  https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Seattle.html
  The results will be saved in a CSV file, which will later be loaded into a Pandas "Neighborhoods" DataFrame.

- Zip Codes of Seattle neighborhoods – the Nominatim geocode() API will be used to obtain the Zip code for each Seattle neighborhood, using the Neighborhood names contained in the Pandas "Neighborhoods" DataFrame.

Nominatim was selected because of Google's recently increased pricing for obtaining geo-coded data. Because Nominatim has a policy of no more than 1 request per second (which out getting throttled or rejected), a 2 second timer will be used between invocations. The results will be saved in a CSV file, that will later be merged into the Pandas "Neighborhoods" DataFrame.

- Real Estate Housing sales and rents in the Seattle neighborhoods – will be extracted from Zillow's median Housing sales price and Rents Index data by neighborhood, which is downloaded from the following "Zillow Data" URL:
  https://www.zillow.com/research/data/

  The results will be saved in a CSV file, that will later be merged into the Pandas "Neighborhoods" DataFrame.

- Four-Square data of current restaurant venues – invoked via FourSquare API and counted, and merged into the "Neighborhoods" DataFrame.

- Four-Square data with counts of current general venues in a given neighborhood – invoked via FourSquare API and counted, and merged into the "Neighborhoods" DataFrame.