Capstone Project - Portland, Oregon’s Bicycle Traffic

Lisa Myers
November 14, 2020

2. Data

2.1 Data sources

These sources will be used to extract/generate the required information:

- *Coordinates* of **Portland** center will be obtained using **GeoPy API geocoding** for Portland, Or location.

- *Bicycle parking* will be extracted from: PortlandMaps - Open Data, updated on 11/4/2020 ... Downloaded *Bicycle_Parking.km*l file and used the pykml parser to read the data with 7618 records.

- *Bicycle routes* will be extracted from: PortlandMaps - Open Data</a> (Bicycle Routes), updated in 11/5/2020 ... Downloaded *Recommended_Bicycle_Route_Points.kml* file and used the pykml parser to read the data with 554 records.

- *Median House Sale Price* will be scraped from a web table: Portland Neighborhoods by the Numbers 2019: The City... web table consists of 96 records of Portland neighborhood data.

- Number of *venues* and *location* in every *neighborhood* will be obtained using **Foursquare API.**

2.2 Data cleaning

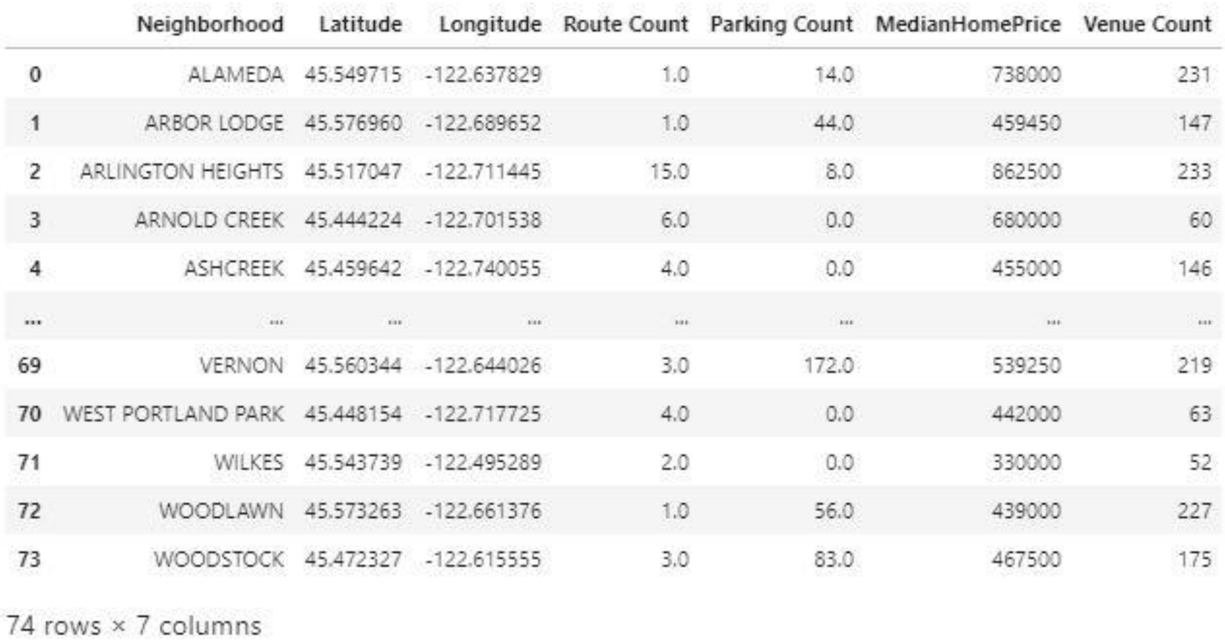
Data downloaded or scraped from multiple sources will be combined into one table using outer joins. There is the potential for missing values, so a neighborhood will be eliminated if the datasets don’t provide coordinates or Median House Sale Price. I will only use the most current bicycle route and parking data, and will use housing data from 2019, as to get the median from a completed year.

There are several problems with the housing dataset. First, the neighborhood names are all in upper case, so I have decided to make all datasets contain Neighborhoods with uppercase letters. Then there are several neighborhood names that used abbreviations, so I will have to update the neighborhood names to match the other datasets. Also, the Median House Sales Price is stored as a string with commas. The column will be modified by removing the commas and converting the sale price to type float.

Both the Bicycle Parking and Route datasets will be extracted with coordinates, and will have to be split into the Latitude and Longitude values. Also, these datasets didn't contain address information. I used the reverse API call from **Google Maps API geocoding** to retrieve the Neighborhood from the dataset coordinates.

2.3 Feature selection

After all data transformations and cleansing have occurred, the combined dataset will be reduced to having seven features and 74 rows. Features that don't relate to housing prices or bicycle ridership will be candidates for elimination. An example, all neighborhood divorce information was removed from the housing dataset.



	Neighborhood	Latitude	Longitude	Route Count	Parking Count	MedianHomePrice	Venue Count
0	ALAMEDA	45.549715	-122.637829	1.0	14.0	738000	231
1	ARBOR LODGE	45.576960	-122.689652	1.0	44.0	459450	147
2	ARLINGTON HEIGHTS	45.517047	-122.711445	15.0	8.0	862500	233
3	ARNOLD CREEK	45.444224	-122.701538	6.0	0.0	680000	60
4	ASHCREEK	45.459642	-122.740055	4.0	0.0	455000	146
...	...	...	...	...	...	...	...
69	VERNON	45.560344	-122.644026	3.0	172.0	539250	219
70	WEST PORTLAND PARK	45.448154	-122.717725	4.0	0.0	442000	63
71	WILKES	45.543739	-122.495289	2.0	0.0	330000	52
72	WOODLAWN	45.573263	-122.661376	1.0	56.0	439000	227
73	WOODSTOCK	45.472327	-122.615555	3.0	83.0	467500	175

74 rows × 7 columns