The Battle of Neighborhoods – Capstone Project

Introduction and Business Problem

Location, location, location! You have heard it before. It’s a common mantra in real estate. Many factors can influence the value of a home, but the surrounding area is one of the most influential. The value of the home rises and falls with the value of the properties around it and the commercial or recreational activities that develop nearby. It would therefore be beneficial to use location data when predicting house prices.

Traditionally historical sales data is used to predict prices in individual neighborhoods. Can we combine historical sales data with location data to make even better predictions? An accurate prediction on the house price is important to a lot of different stakeholders as prospective homeowners, real estate agencies, developers, investors, mortgage lenders and insurers etc.

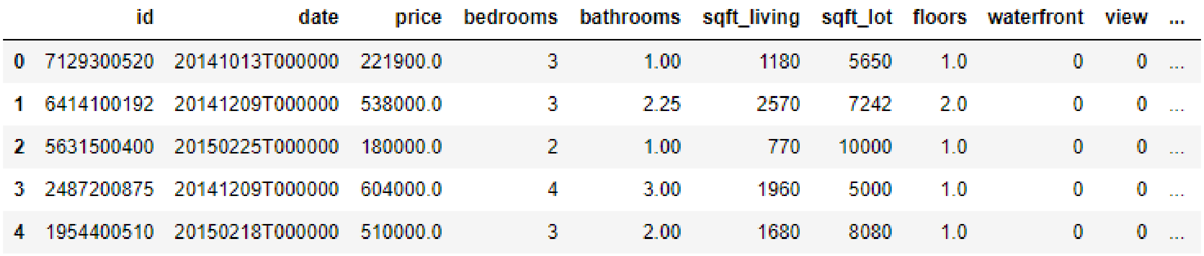
The goal is to build a data-driven decision support tool predicting house prices in different locations combining historical sales data with location data.

Data

House Sales in King County, USA

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. The dataset includes attributes and features such as square footage, number of bedrooms, building year etc. It also contains the latitude and longitude coordinates for the location of the house.

Example of the data:

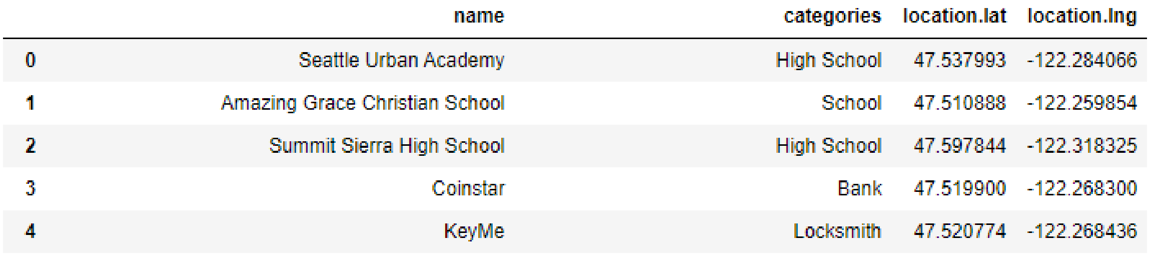


	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0	...
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0	...
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0	...
3	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0	...
4	1954400510	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0	...

Foursquare Location Data

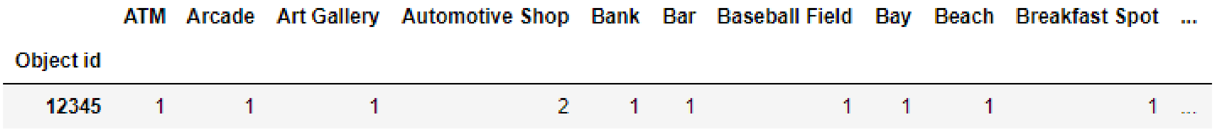
Since we know the location of the houses sold in King County, we can use the Foursquare API to find out which services and entertainment or recreation opportunities like movie theaters, parks, and golf courses can be found in the immediate area of each house and include that information when building our prediction model.

Example of data we collect for one house:



	name	categories	location.lat	location.lng
0	Seattle Urban Academy	High School	47.537993	-122.284066
1	Amazing Grace Christian School	School	47.510888	-122.259854
2	Summit Sierra High School	High School	47.597844	-122.318325
3	Coinstar	Bank	47.519900	-122.268300
4	KeyMe	Locksmith	47.520774	-122.268436

For each house we will summarize the data to something like this:



	ATM	Arcade	Art Gallery	Automotive Shop	Bank	Bar	Baseball Field	Bay	Beach	Breakfast Spot	...
Object id											
12345	1	1	1	2	1	1	1	1	1	1	...