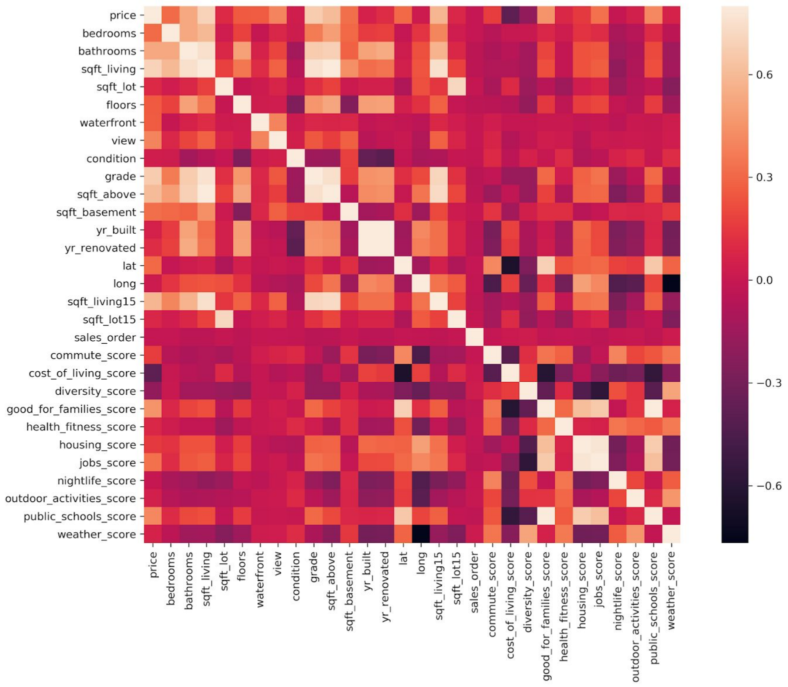
Capstone Project 1 Data Storytelling Report

Hye Joo Han

The dataset contained 21,597 observations (house sales) each with 44 features. I did univariate, bivariate and multivariate analyses. The followings are the summary of what I found the most interesting through the analyses. For full analysis, see the Jupyter notebook https://github.com/math470/Springboard_Capstone_Project_1

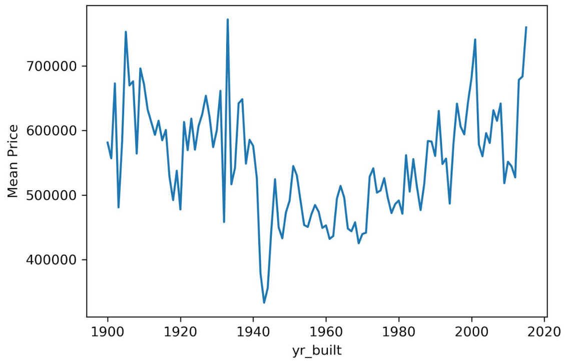


1. What are the features correlated with house prices?

- The features related to house living space including sqft_living, grade, and bathrooms are highly or moderately correlated with house prices. The square footage of home is the most correlated with a house price. The number of bathrooms is much more correlated with a house price than the number of bedrooms.

- The house location features such as good_for_families_score, public_schools_score, and job_score correlate moderately with a house price. The latitude is also correlated with price.

2. Why are the built years not correlated with house prices (corr=0.05)?



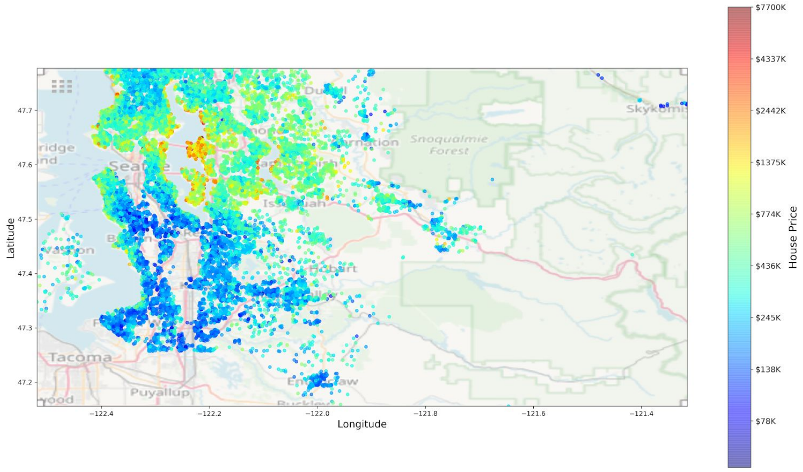
I found an interesting pattern in the plot of house price vs. year built. The mean or median house prices tend to be lower for older houses if they were built between 1960 to 2015 (as expected). However, prices tend to be higher for older houses if they were built between 1900 and 1960. Through further analysis, I found 2 possible reasons that can add up to the pattern.

- Plots of 'bedrooms', 'bathrooms', 'sqft_living', 'floors', 'grade', 'sqft_above', 'good_for_families_score', 'public_schools_score', and 'jobs_score' as a function of year built have the similar pattern. They have lower values for older houses from built year 2015 to 1960, but tend to get better values for older houses from 1960 to 1900.

- Latitude tends to decrease and longitude tends to increase as year built increases. The pattern seems to show how the Seattle metropolitan area has expanded from Seattle to the east and south suburb cities since 1900 (within King county). The 'commute_score' and 'nightlife_score' tend to decrease over time since 1900, but increase after 2000. This pattern is possibly due to the dispersion of Seattle just like the pattern found in latitude and longitude.

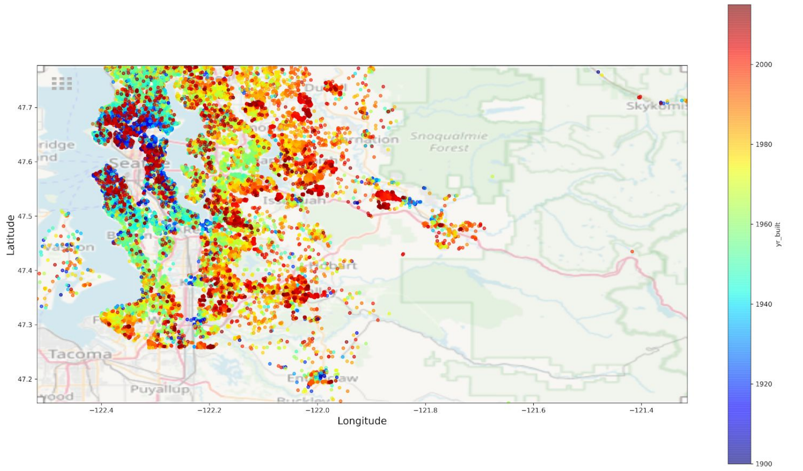
3. Since the dataset contains exact house locations (longitude and latitude features), I was able to draw houses on the King county map using colors for another feature. There were three interesting findings, which are very consistent with findings from the plots of year built (summarized in 2).

- Houses near Lake Washington and east and west sides of the lake tend to be more expensive and the extremely expensive houses are on the waterfront. This is possibly why latitude of a house is positively correlated with a price. Are waterfront houses significantly more expensive? I cannot answer this until I do some hypothesis testing, but the boxplot showed there is a big difference in prices between waterfront and not waterfront houses.



- The house maps colored with good_for_families_score and public_schools_score show similar color patterns, more expensive houses on the west and east of Lake Washington and cheaper houses below the lake.

- The last house map colored with year built showed very old houses are mostly concentrated in Seattle and relatively new houses built around 1970 to 2000 are dispersed to the east and south of Seattle.


2000
1980
1960
1940
1920
1900
yr_built
Latitude
47.7
47.6
47.5
47.4
47.3
47.2
Longitude
−122.4
−122.2
−122.0
−121.8
−121.6
−121.4
Snoqualmie Forest
Tacoma
Puyallup
Skykomi