Capstone Project 1: Milestone Report

Customer/Client: General Public or Real Estate agents

Problem Statement:
Determine when the housing price is likely to go up or down based on demand and supply information from 45 days to 6 month in advance.

How would this help the clients/customers?

For the general public, it would be so they can determine when to buy or sell. Without this information, the general public would have to rely on hearsay from their acquaintances and real estate agents without a hard data they can rely on.

For real estate agents, they can use the information as marketing material and persuade potential clients that it is a good time to buy or sell with them. Having data will make it easier to sell the strategy to the clients.

It generally takes about 30-60 days to complete the closing process. The amount of time to find and buy a house is different for everyone but 6 months is generally the amount of time you are in contract with a real estate agent. Using these factors, 45 days as lower boundary and 6 months as upper boundary for forecasting will serve most non-commercial real estate dealings.

Data sets used:

Data set	URL	Level	Time range	Time interval / frequency	Format
Zillow Economics data - Sales Price	https://www.kaggle.com/zillow/zecon#County_time_series.csv	County	1996 - 2017	Monthly (month end)	CSV
Zillow Economics data - # of days on Zillow	https://www.kaggle.com/zillow/zecon#County_time_series.csv	County	1996 - 2017	Monthly (month end)	CSV

Zillow Economics data - # of Monthly listings	https://www.kaggle.com/zillow/zecon#County_time_series.csv	County	1996 - 2017	Monthly (month end)	CSV
Unemployment data	https://fred.stlouisfed.org/tags/series?t=bls%3Bcounty&ob=pv&od=desc	County	1990 - 2019	Monthly	Excel / CSV
Mortgage Rate	http://www.freddiemac.com/pmms/docs/30yr_pmmsmnth.xls	National	1970 - 2019	Monthly	Excel
Historical Housing Affordability Index	https://www.car.org/marketdata/data/haitraditional	County - CA only	2013 - 2018	Quarterly	Excel
Population data (Estimate)	https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html#par_textimage_739801612	County	2010 - 2018	Annual	CSV
Crime data - Violent and Property crimes	https://oag.ca.gov/crime	State level - CA only	2009 - 2018	Annual	Excel

Data wrangling approach:

1. 3 representative counties in California were selected.
2. To clean the csv and excel files before consuming it from python code:
    a. Filled missing data to NA.
    b. Set date format to yyyy-mm-dd.
    c. Removed unnecessary columns and rows
    d. Formatted column names to be consistent - start with capital letters
3. Read in each csv files into dataframes.
4. For Quarterly and Annual data, resampled them to monthly to match the rest of the dataset.
5. Melted data to set date as a value rather than column name.

6. Combined all dataframes into one dataframe using RegionName and Date as key.
   Because affordability data has only years 2013 - 2018 data, even though other dataset have more data, dataframe is restricted to 2013-2018 to join all variable data into one dataframe.
7. Annual data (population, violent crimes and property crimes) was included in the dataframe but ultimately not used because the number of data was too small.
8. Seasonal data was smoothed out by averaging in rolling window of 12 months - # of properties listed in Zillow.
9. % change value was calculated for Sales price and mortgage rate.

Findings from Initial Exploratory Analysis:

Housing sales prices years 2013-2018 in 3 representative markets (Alameda, Sacramento, L.A.) in CA have been going up.

Few variables are used to correlate with sale prices.

1. Annual Mortgage Rate:
2. Monthly Unemployment Rate
3. # of days properties were listed on Zillow
4. # of new properties listed each month on Zillow
5. Quarterly Affordability index

EDA and inferential statistic method used:

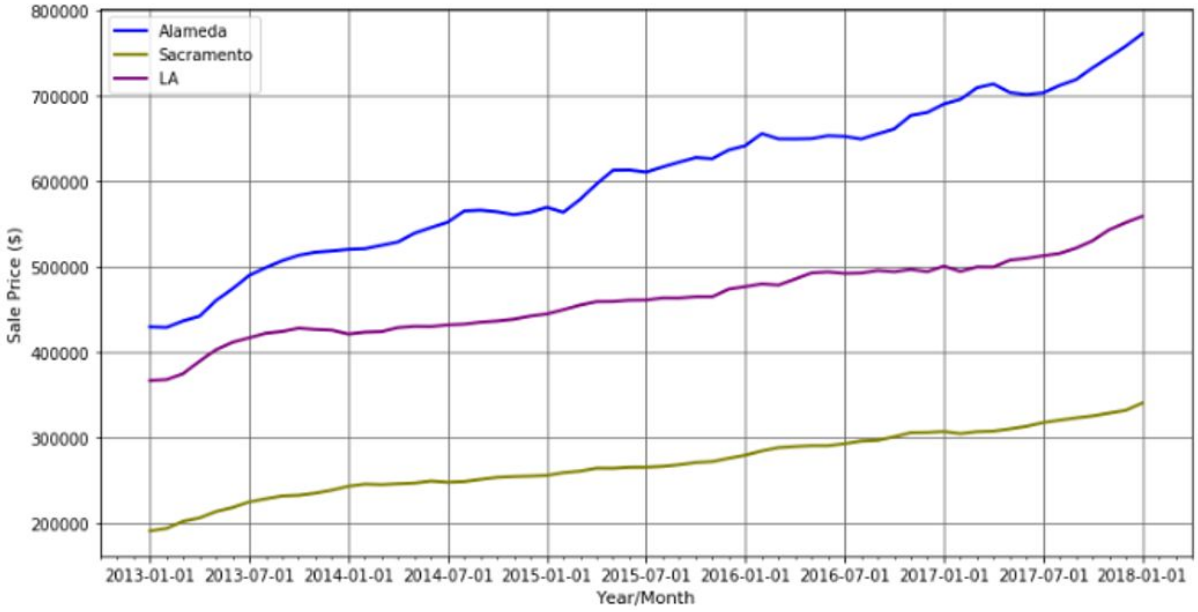
1. Scatter plot between sales price and variables.
   Mortgage rate lagged by 1-3 month are also used for comparison to account for any delay in effect on sales price.
2. Linear correlation plot
3. Correlation heatmap
4. Pearson R coefficient

It seems to all comes down to how many people are employed and feel secure about their jobs is what correlates the most with housing price trend in CA.
Other economic factors such as affordability index have strong correlation as well compared to factors related to supply and demand for the three representative California counties in 2013 to 2018 Jan.

Heatmap and Pearson R coefficient results:

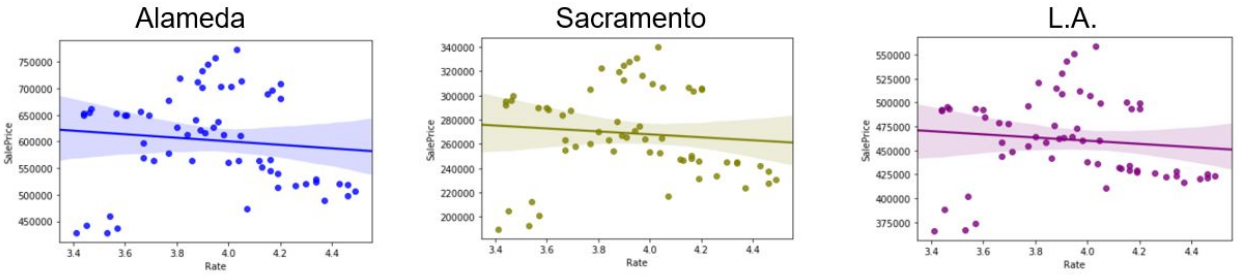
1) Alameda: Unemployment Rate is -0.96 at strongest negative correlation. Mortgage rate has weakest correlation at -0.11. Only number of monthly listing in Zillow is over -0.5.
2) Sacramento: Unemployment Rate is -0.95 at strongest negative correlation followed by Affordability index at -0.89. Mortgage rate has weakest correlation at -0.09. Surprisingly, for this market, number of days on Zillow has correlation of -0.55 and number of monthly listing in Zillow is at 0.16. Just looking at Pearson coffecient, Sacramento market has different characteristics than Alameda.
3) L.A.: Unemployment Rate is -0.93 at strongest negative correlation followed by Affordability index and Monthly listing at -0.67. Mortgage rate has weakest correlation at -0.11. L.A. is yet again different from Sacramento and Alameda in that affordability index and number of monthly listing is over -0.5.

Housing prices trend

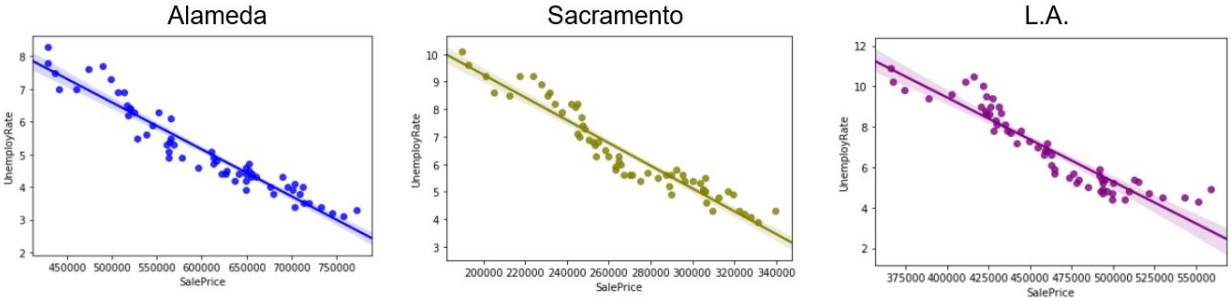


Linear Regression:

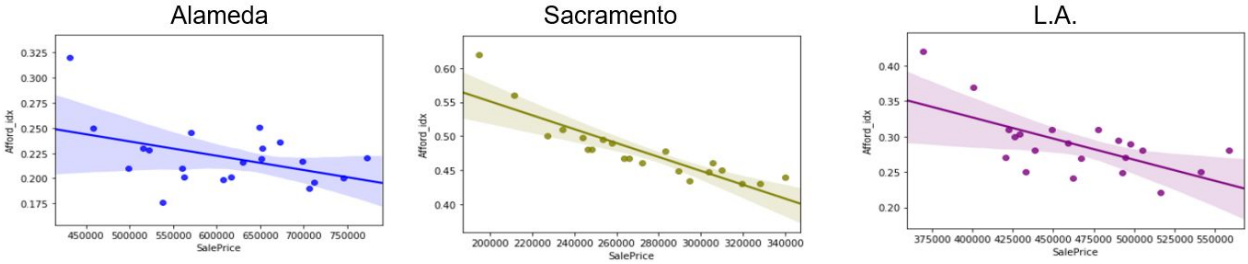
1. Mortgage Rate: Linear correlation is fairly weak, almost none. Pearson R coefficient for Alameda is -0.11. Other counties have similar numbers.



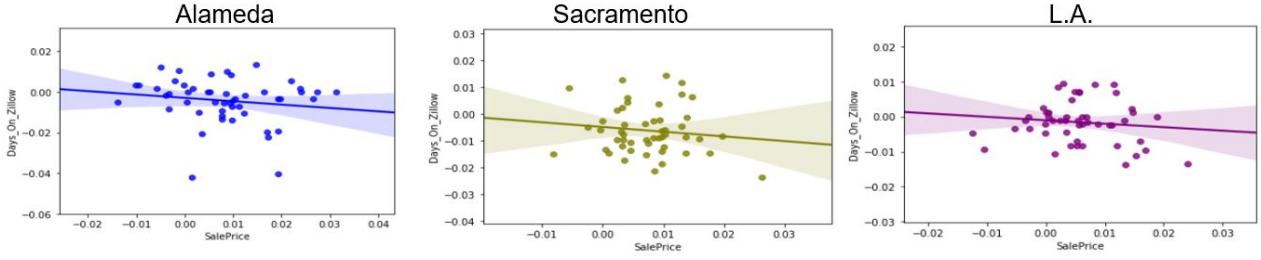
2. Unemployment Rate: Unemployment rate definitely seem to strongly correlate negatively to the sales price. Pearson R coefficient is -0.96, -0.95 and -0.93 respectively for each county.



3. Affordability Index: Affordability is in negative correlation with sales price. For Sacramento and L.A., Pearson R Coefficients are -0.89 and -0.67 respectively but for Alameda, it did not go over -0.5. You can see the relationship in linear regression graph below. Sacramento and L.A. have steeper angle compared to Alameda.

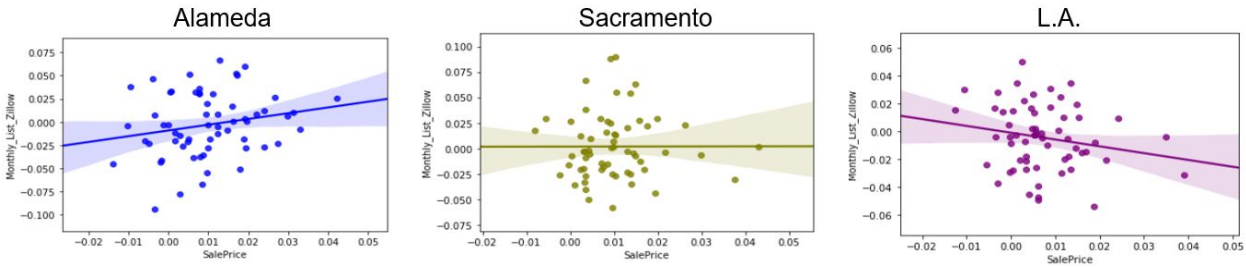


4. Number of days property is listed on Zillow: For all counties, Pearson R Coefficient are below -0.5. Correlation with sale price is very weak.



5. Number of properties listed each month on Zillow: Pearson R Coefficient was slightly over -0.5 for Alameda and L.A. but for Sacramento, it was only at 0.16. Linear regression

graph below shows the relationship. Overall, correlation with sales price is there but not as strong as unemployment rate and affordability index.



Presentation deck:
https://docs.google.com/presentation/d/1dge2W830n8LZkilj7NK2rZiQ0nmB0Y8ULvQU3NYq3AY/edit?usp=sharing