Final Report

Milestone I: A Comprehensive Study of California Housing Price Behaviors

SIADS 591 & 592 (December 2020 - January 2021)

Mel Nguyen & Do Young Kim

[Motivation](#)

[Data Sources](#)

[Data Manipulation Methods](#)

[Analysis and Visualization](#)

[Statement of Work](#)

[Bibliography](#)

Motivation

Buying a house is one of the biggest decisions a person would make in his or her course of life. Beyond financial reasons, a house is where a person, or a family spends the majority of their time. Our project will explore the housing scene in California. We are interested in looking at the data primarily from a buyer's point of view; for instance, as someone who is relocating to the area and would like to understand the real estate scene.

Through this project, we hope to achieve two objectives. We want to help buyers understand which cities would be the most suitable for their budgets through an analysis of house prices in cities, how these prices have changed over time, and for a buyer who is considering moving to and settling in California in the next 6 months to 1 year, what can they expect from the price movements.

Secondly, we will investigate indicators that could help buyers gauge the home prices in a certain city. For example, our analysis looks into whether crime rate could be an indicator to a city being more expensive. While we will not be attempting to determine a causal relationship in this project, we hope that by examining the

statistical relationships between house prices and demographic variables, buyers will be able to draw useful insights that will aid their decision making.
Overall, we hope that our analysis will serve not only as a guidebook to existing and future Californian dwellers to be better informed, but also lay the groundwork for future, more in-depth analyses that could unravel underlying forces that drive the valuation of houses in this state.

Data Sources

Zillow Datasets

Headquartered in Seattle, Zillow is a real estate and rental marketplace, serving the full spectrum of living and owning a home. The company possesses a database of more than 110 million U.S homes, including homes in California.

- Name: There are three Zillow datasets obtained, which are Zillow Data, Zillow Indicators, and Zillow Regions.
- Location: https://www.quandl.com/databases/ZILLOW/data
- Access Method: API
- Format: CSV
- Size: 4.51 GB
- Total Records: Zillow Data: 123,301,858, Zillow Indicators: 78,532, Zillow Regions: 56
- Time Period Used: Monthly data, between Jan 2017 and Jan 2021
- Important Variables: date, value, ZIP code, county, city.

US Zip Code Latitude and Longitude Dataset

The dataset provides information about each ZIP code's latitude and longitude coordinates.

- Name: US Zip Code Latitude and Longitude
- Location: https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/
- Access Method: Web download
- Format: JSON
- Size: 17 MB
- Total Records: 43,191
- Important Variables: city, zip code, latitude and longitude. Because the original file is in JSON, we use Spark to unpack the fields 'fields' and 'geopoint' to get to the variables.

FBI UCR California Offenses Known to Law Enforcement Dataset

Sitting under the Federal Bureau of Investigation (FBI), Uniform Crime Reporting (UCR) provides statistics for use in law enforcement, as well as criminal justice research and the public in general.

- Name: 2019 California Offenses Known to Law Enforcement
- Location: https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/tables/table-8/table-8-state-cuts/california.xls
- Access Method: Web download
- Format: XLS
- Size: 99 KB
- Total Records: 456
- Time Period Used: 2019
- Important Variables: city, population, number of crimes in different categories.

NASDAQ Historical Prices Dataset

Data on NASDAQ historical values. Historical data is inflation-adjusted using the headline CPI and each data point represents the month-end closing value.

- Name: NASDAQ Historical Prices
- Location: https://www.macrotrends.net/1320/nasdaq-historical-chart
- Access Method: Web download
- Format: csv
- Size: 17 KB
- Total Records: 48
- Time Period Used: We extract data from the same period as the Zillow datasets as much as possible, between Jan 2017 and Dec 2020.
- Important Variables: closing value.

30 Year Fixed Mortgage Rate Dataset

Historical data of the 30 year fixed rate mortgage average in the United States since 1971.

- Name: 30 Year Fixed Mortgage Rate
- Location: https://www.macrotrends.net/2604/30-year-fixed-mortgage-rate-chart
- Access Method: Web download
- Format: csv
- Size: 10 KB
- Total Records: 598
- Time Period Used: We extract data from the same period as the Zillow datasets as much as possible, between Jan 2017 and Dec 2020.

- Important Variables: 30 Year Fixed Mortgage Rate.

FRED Economic datasets

Time series economic data of California including New Private Housing Units, Real Trade-Weighted Value of the dollar, Housing Inventory, and Average Hourly Earnings of All Employees.

- Name: FRED Economic datasets
- Location: https://fred.stlouisfed.org/
- Access Method: Web download
- Format: csv
- Size: 24 KB
- Total Records:
    - Housing Inventory 54
    - Average Hourly Earnings of All Employees 167
    - New Private Housing Units 395
    - Real Trade-Weighted Value of the dollar 395
- Time Period Used: We extract data from the same period as the Zillow datasets as much as possible, between Jan 2017 and Nov 2020.
- Important Variables: Housing Inventory, Average Hourly Earnings of All Employees, New Private Housing Units, Real Trade-Weighted Value of the dollar

California State GDP data

Quarterly Real GDP in chained dollars of California.

- Name: California State GDP data
- Location: https://apps.bea.gov/itable/iTable.cfm?ReqID=70&step=1&acrdn=1
- Access Method: Web download
- Format: csv
- Size: 18 KB
- Total Records: 64
- Time Period Used: We extract data from the same period as the Zillow datasets as much as possible, between Jan 2017 and Dec 2020.
- Important Variables: GDP of all industry total.

2019 ACS data

2019 American Community Survey(ACS) data that contains demographics of California such as population, median age, median income, etc.

- Name: 2019 ACS data
- Location: https://data.census.gov/cedsci/
- Access Method: Web download

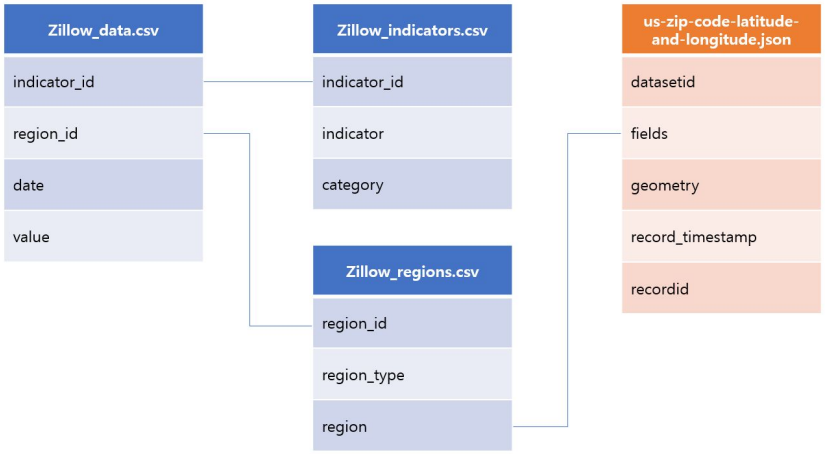
- Format: csv
- Size: 20 MB
- Total Records:
    - ACS Demographic And Housing Estimates 1,766
    - Geographic Mobility By Selected Characteristics In The US 1,766
    - Means Of Transportation To Work By Selected Characteristics 1,766
    - Households And Families 1,766
- Time Period Used: 2019
- Important Variables: population, median age, median income, educational attainment.

Data Manipulation Methods

Zillow Datasets

Since the size of the dataset was relatively large, we decided to utilize PySpark to efficiently process the data. First, we loaded the dataset into Spark Dataframe and created a temporary view of each table so we could process them using SQL queries. The next step is to merge and filtering the Spark Dataframes, which was done with the following:

- We started with merging Zillow datasets using SQL queries. We decided to break the list into the zip code level because that was as micro as we could get.
- The region column's data was separated by semicolons. Here's how it was formatted: "*zip code; state; city; county; community*". We utilized the UDF function to split them into separate columns. Then we extracted California data using a WHERE statement.
- We imported the us-zip-code-latitude-and-longitude.json file and merged them with the Zillow dataset. Here's how they are connected:



Lastly, we converted the Spark Dataframe to Pandas DataFrame so we could utilize pandas functions to proceed with the further manipulations, and filtered out rental values and converted the data column into Pandas Datetime object

City Crime/Population

We computed the crime rate out of given data using broadcasting. The crime rate was defined as the ratio between the total number of crimes in a city, and that city's population

NASDAQ Historical Prices Dataset / 30 Year Fixed Mortgage Rate Dataset / FRED Economic datasets

The original datasets contained only spot values of indexes. we computed the monthly rate of changes using the pct_change method in Pandas.

ACS Demographics data

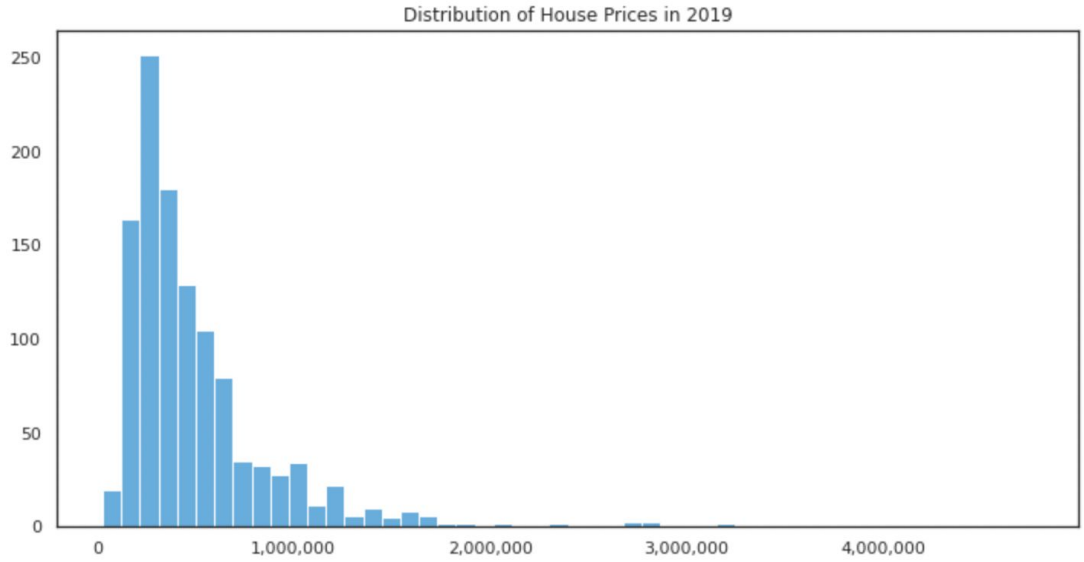
We processed the Geographic Area Name column using the apply method in Pandas to match their format to a 5-digit zip code. As the dataset contains the standard deviation information of features as well as estimates, we filtered them out using str.contains methods in Pandas.

Analysis and Visualization

House Prices Distribution And Trend Analysis

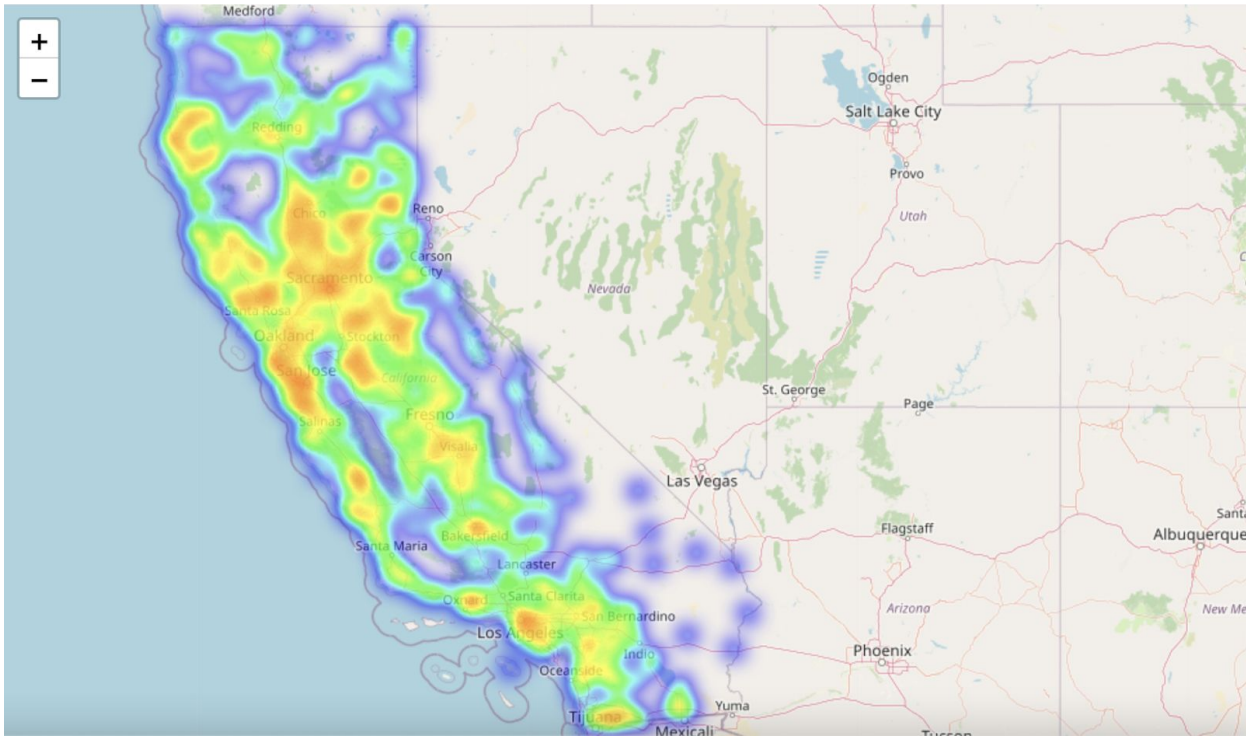
House Prices Distribution

After data cleaning and filtering out rental records, as the analysis only covers house buying/selling, we start by looking at the distribution of house prices in California. We have selected 2019 data instead of 2020 for various reasons, but the two most important reasons are to set a baseline expectation in terms of housing prices, given that we don't fully understand the impact of COVID-19 on the real estate market, and to consistently match subsequent analyses with other variables, as some of the social-economic datasets do not have data for 2020 yet.



In 2019, the mean price for a house in California was $516,013.7, making California one of the most expensive states to purchase a house in, ranking only after Hawaii and District of Columbia (Experian, 2019). The histogram above shows a long right tail, suggesting that there are a number of outliers. These are houses that could cost up to $4M and more. Indeed, there is a large standard deviation for house prices in California ($443,052.18), and consequently, a significant difference between the mean and median house prices ($381,236.67).

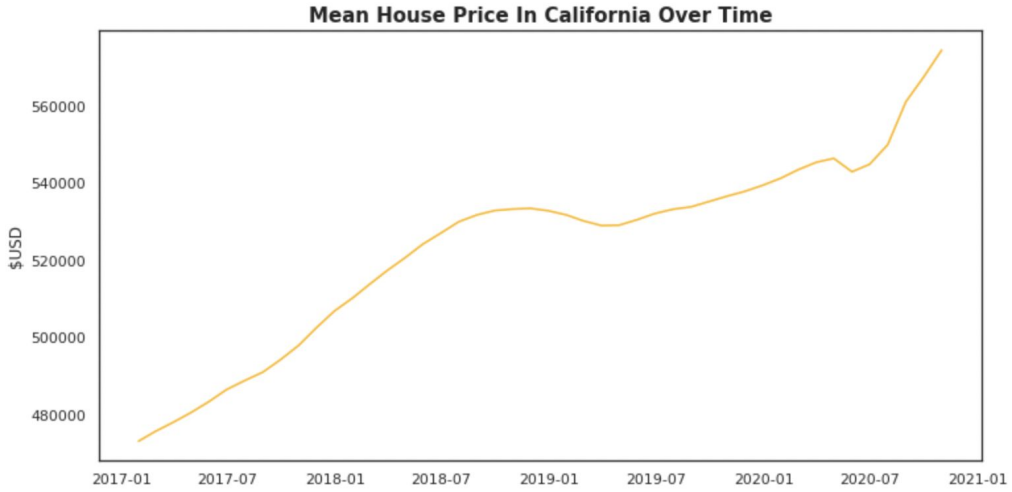
Keeping this in mind, let us take a closer look at the state of California, and see how the average house prices differ between cities/counties:



We see a red cluster in the northern region of California, and another, albeit smaller, red cluster in Southern California. These clusters correspond to the largest metropolitans in California. In the south, we have Los Angeles, whereas in the north, cities like San Jose, Sacramento, and San Francisco emerge as cities with the highest cost of purchasing a home.
In short, consumers should bear in mind the state-level average cost to purchase a house in California, but this figure is likely to be contributed significantly by the big cities in California. The fact that areas surrounding these cities have a much lower average house price (represented by the heatmap color being closer to blue) means that consumers with a lower budget can still reasonably afford a house in California by staying slightly further away from the city centers.

House Prices Trend Analysis

Now that we know how much it would take on average to buy a house in California, and where the more expensive areas are, let's proceed to look into how these prices have changed over time.

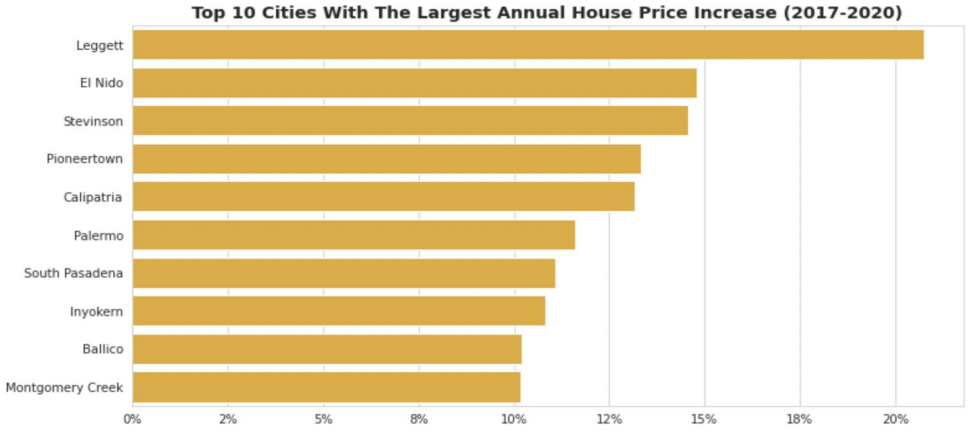


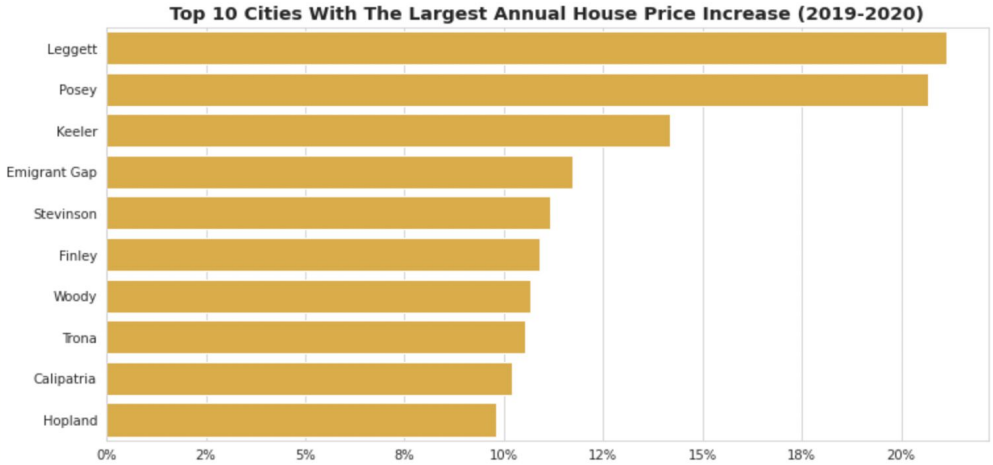
The average house price in California has been on a steady increase since 2017. There are slight fluctuations along the way, but in general, it has been an uptrend, with the mean house value increasing from $480,000 USD in 2017 to around $570,000 USD in 2020, an average growth of 5% per annum. The average house price has stayed on an upward trend even with the pandemic going on, and in fact, even increased more sharply since July 2020.
Contrary to our initial assumption that the COVID-19 pandemic will dampen the real estate market in California, the growth trajectory has not changed. One of the possible reasons is due to the lower mortgage rates that happened off the back of the Federal Reserve's interest rate cuts in order to counter negative impacts in the financial markets. This leads to lower borrowing costs and more affordable housing in the near future, thus driving the demand in the real estate market (California

Association Of Realtors, 2020). Another factor contributing to the increased average house price is because of the lower supply of newly-constructed homes in California. COVID-19 has disrupted the global supply chain, which in turn leads to rising material costs to build houses and consequently less new houses being built. Combined with the rising demand covered in the point above, this creates an upward pressure on the existing inventory of houses in California, and as a result, keep the price increase steady throughout the pandemic (California Association Of Realtors, 2020).

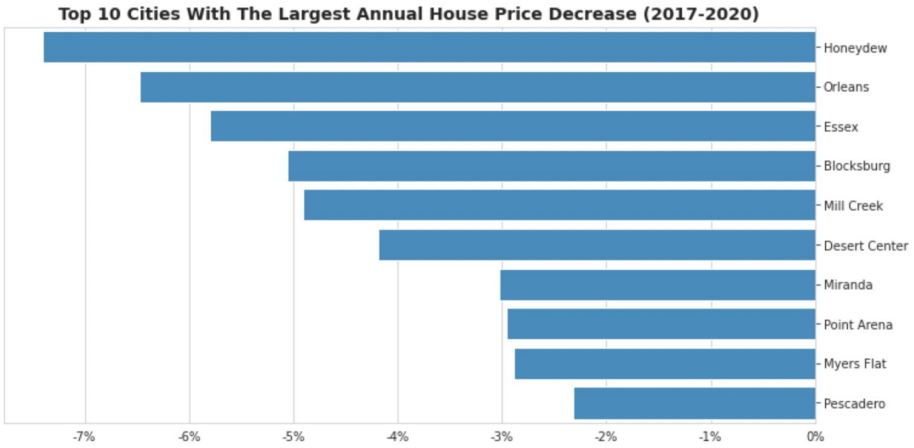
However, knowing that the average house price is very different between cities in our previous analysis, it is likely that house price has been growing differently between cities as well.

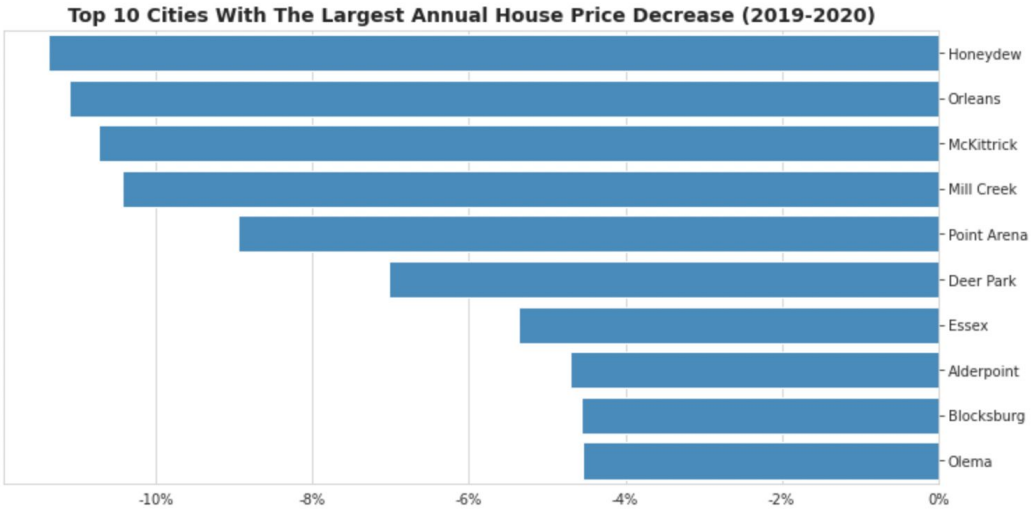
We first look into the top ten cities with the largest increase in house prices, over the entire 4 years worth of data we have collected, and over the period of 2019-2020 in order to zoom into the COVID-19 year.





Note that none of the big metropolitan cities are in the top 10 list. Let's now also take a look at the bottom 10 cities:





House price increase over the years has been starkly different among cities in California. On the one hand, there are cities that have seen double-digit house growth over the years, while on the other hand, some cities have gone against the current and seen the values of their homes depreciating over the years. Leggett tops the list both year-on-year and between 2019-2020, while some other rising cities recently are Posey, Keeler, and Emigrant Gap.

The bottom 10 is relatively consistent over the years, however, as we see very few changes in the cities whose house prices have been declining substantially. Interestingly, none of the metropolitan cities (cities in the 'red' clusters seen above) is in either the top 10 or the bottom 10, suggesting that even though these are the most expensive cities in California, house price has not grown as much as other cities.

	city	avg-annual-change	2019-vs-2020-change
476	Los Angeles	0.05663	0.03049
737	San Francisco	0.04484	-0.00179
743	San Jose	0.04202	0.01989

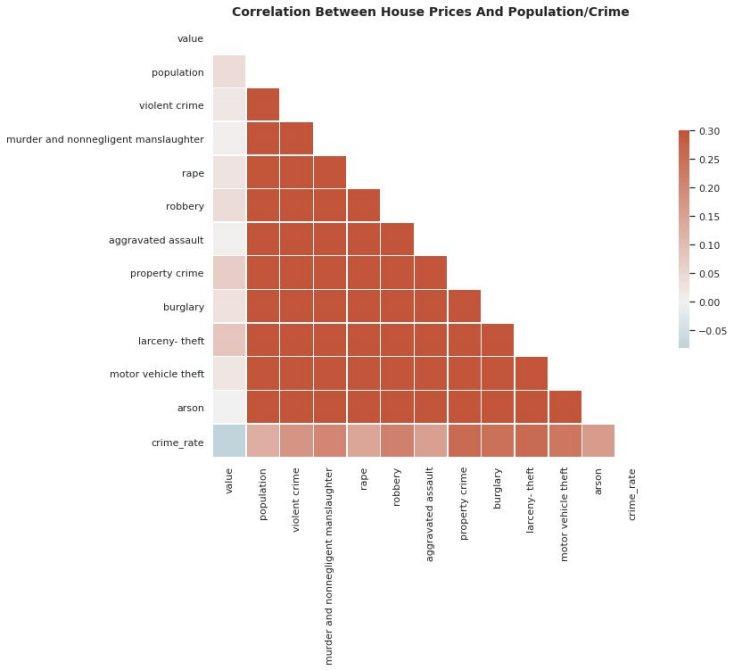
The growth rate has been relatively low for metropolitan areas. Between 2019 and 2020, house prices in San Francisco have in fact been declining. COVID-19 could be a driving force behind this trend. Between March and September 2020, more than 6000 residents requested for outbound moving assistance, according to data from website moveBuddha.com (Madden, 2020).

As remote work becomes increasingly the norm, there is no longer a compelling reason to stay in more expensive cities when people could move to cheaper cities while retaining the same earning power. Notably, big tech companies such as Facebook and Google have been supporting remote work as well, allowing employees to work remotely in the long run (Hadden et al., 2020). Some companies are even going as far as offering a one-time bonus for employees who agree to move to less expensive cities, in exchange for a salary reduction (Melin, 2020). Similarly, we can observe slowing house price growth for other cities like Los Angeles and San Jose. Albeit less tech-concentrated as compared to San Francisco, the trend is likely to be driven by the same reason.

House Prices versus Population and Crime Levels

In this section, we will investigate the statistical relationship between house prices and crime-related variables. Safety is one of the most important considerations to home buyers. Logically, buyers would prefer, and are more likely to pay more for a house in a neighborhood with better security.

To understand the relationship between house prices and crime levels, we use both the Zillow datasets and the FBI UCR California Offenses dataset. As the latter contains data in 2019 only, we extract 2019 data from the Zillow dataset. The FBI UCR dataset contains data on several types of crimes, as well as the population, at the city level. We also compute the crime_rate, which is computed as the ratio between the total number of crimes in a city, and that city's population.



The visualization is a correlation matrix between house price (column “value”), the population, and various types of crimes. We can observe the correlation between house price and the rest of the variables by looking at the vertical edge of the triangle.

Unsurprisingly, the correlation matrix shows a medium positive correlation between the crimes. In other words, a city with higher levels of theft is likely to also have high levels of robbery, murder, and other types of assault. Contrary to our initial assumption, however, there is almost no, or weak positive correlation (less than 0.15) between the average house price and the population, or between the price and other crime variables.

Nevertheless, correlation alone is not a sufficient measure to conclude that there is no relationship between house price and crimes. For one, the correlation is computed using figures at a city level, which may have obscured the impact at higher levels of details (eg. neighborhoods). Secondly, we do not have sufficient data to investigate the impact of crimes on short and medium-term house prices. These could be possible directions for future, more in-depth analyses on the relationship between these variables.

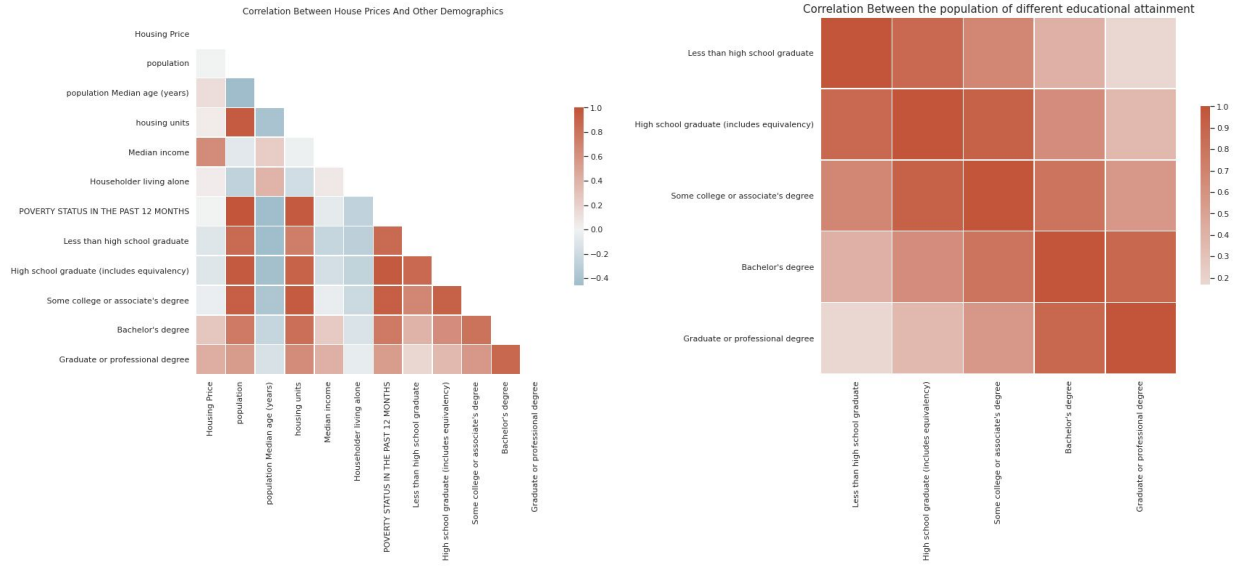
House Prices versus 2019 ACS Demographics

In this section, we will utilize 2019 American Community Survey(ACS) data to see if any demographic variables correlate with house prices. Since there are many features in this survey, we chose some of them we thought would be relevant to be analyzed.

- Population / Housing Unit

    - Although we already talked about the population in the previous section, we still decided to include it since the ACS data provides the information at the zip code level, allowing us a more detailed approach.
- Median Age
    - We wanted to see if the generation gap applies to house prices too.
- Median Income / Poverty Status In The Past 12 Months
    - We wanted to see if the dwellers' financial status has a relationship with the house prices of that area.
- Householder Living Alone
    - Typically, the areas that have many single households are the ones that are located near the CBD.
- Educational Attainment
    - People who have received a high level of education are more likely to be in stable, well-paying jobs. That means that they would have enough purchasing power to afford expensive housings.

Just like how we did in the previous section, we calculated the correlations between each variable and created a heatmap based on the variables' correlation matrix.



As shown from the chart, the median income and the number of the population who has a graduate or professional degree are the variables that show a medium positive correlation with house prices. Intuitively, this seems obvious because median income and high education level are related to higher purchasing power either directly or indirectly. Of course, correlation alone does not tell any causal relationship.

Another interesting thing we have found is that each educational attainment population has a stronger correlation with those with the adjacent level of education, and the correlation gets weaker when the educational level moves

further away. The population of graduate or professional degree holders has the strongest correlation with bachelor's degree holders and has the weakest correlation with less than high school graduates. That is also valid in vice versa.

House Prices vs Other Economic Indicators

We wanted to see if other economic variables affect California's housing price. We chose the following economic indexes for investigation;

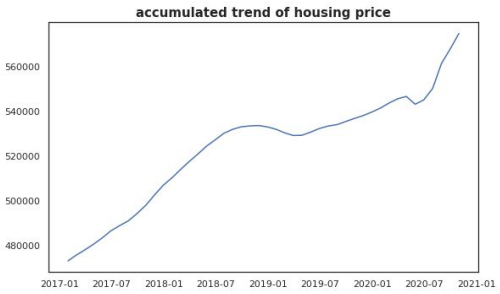
- NASDAQ value
    - The stock market would be an alternative investment for the housing market. Theoretically, a surge in the price of one market would shift the other market's demand curve towards the right, resulting in a surge in both markets' prices.
- Mortgage rate
    - Normally, when investing in housings, investors would establish mortgage loans for various reasons. Thus, the increase in the mortgage rate would make investing more expensive, lowering demand for housing properties.
- Real Trade-Weighted Value of the dollar for California
    - Inflation is an important factor when investing in an asset. People would invest in real estate properties to hedge the effect of inflation.
- New Private Housing Units Authorized by Building Permits for California / Housing Inventory: Active Listing Count in California
    - These indexes are the indicators for the supply for California's housing market.
- Average Hourly Earnings of All Employees in California
    - Purchasing power is also one factor that shifts the demand curve to the right. Also, well-paying business establishments would lure people into moving in.

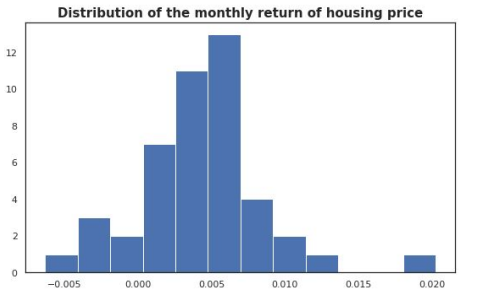
House prices' trend and distribution

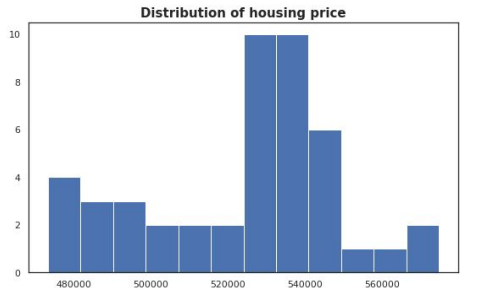
Before jumping into the main analysis, we have first explored the trend and return rate of California's housing market. We created line plots and histograms to see the trend and distribution of the monthly return and accumulated trend of California's housing price. From the visualization, we could point out a couple of things;

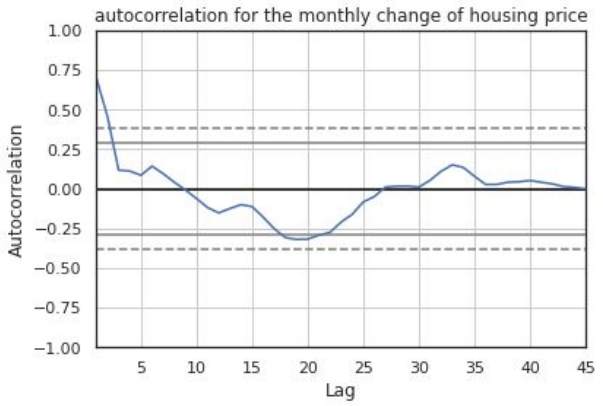
- Housing price rarely drops
    - Looking at the line plot of the Monthly return of housing price, the major portion of the return rates are positive, meaning that there were very few losses during the period.

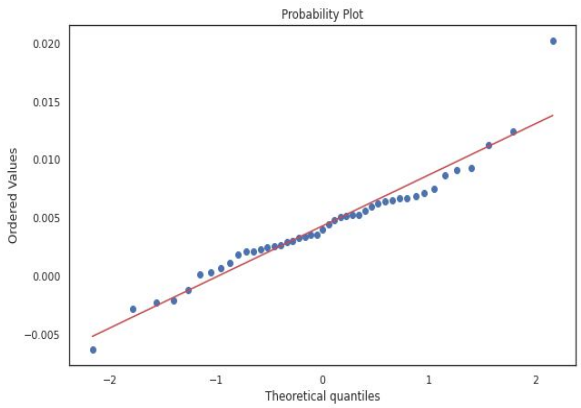
- The distribution of the monthly return of housing prices looks similar to the normal distribution
    - As shown in the QQ plot, it seems that the ordered values are relatively well aligned with the theoretical quantiles.









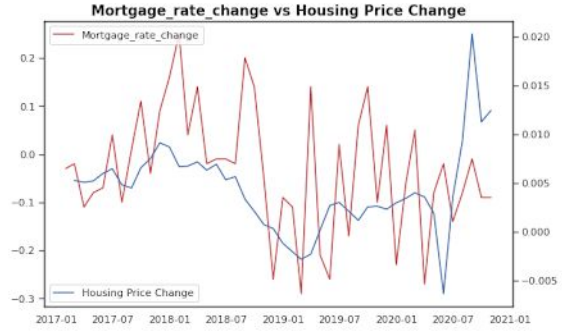


The relationship between house prices and other indexes

We created a series of line plots of housing prices vs. other economic indexes for the next step. We also created a bar chart to compare the standard deviation of each index. NASDAQ value and average hourly earnings of all employees seem to align with the housing price trend. The mortgage rate, on the other hand, shows a somewhat negative relationship with the housing price. Other variables don't seem to have any sort of relationship with the housing price. Also, looking at the bar chart, it seems that the housing price is pretty stable compared to other indexes.


NASDAQ_change vs Housing Price Change
NASDAQ_change
Housing Price Change

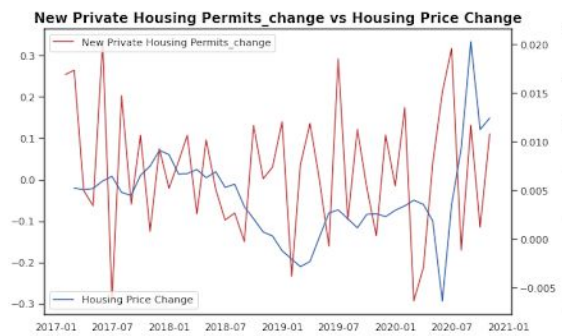

NASDAQ_value vs Housing Price
NASDAQ_value
Housing Price


Mortgage_rate_change vs Housing Price Change
Mortgage_rate_change
Housing Price Change

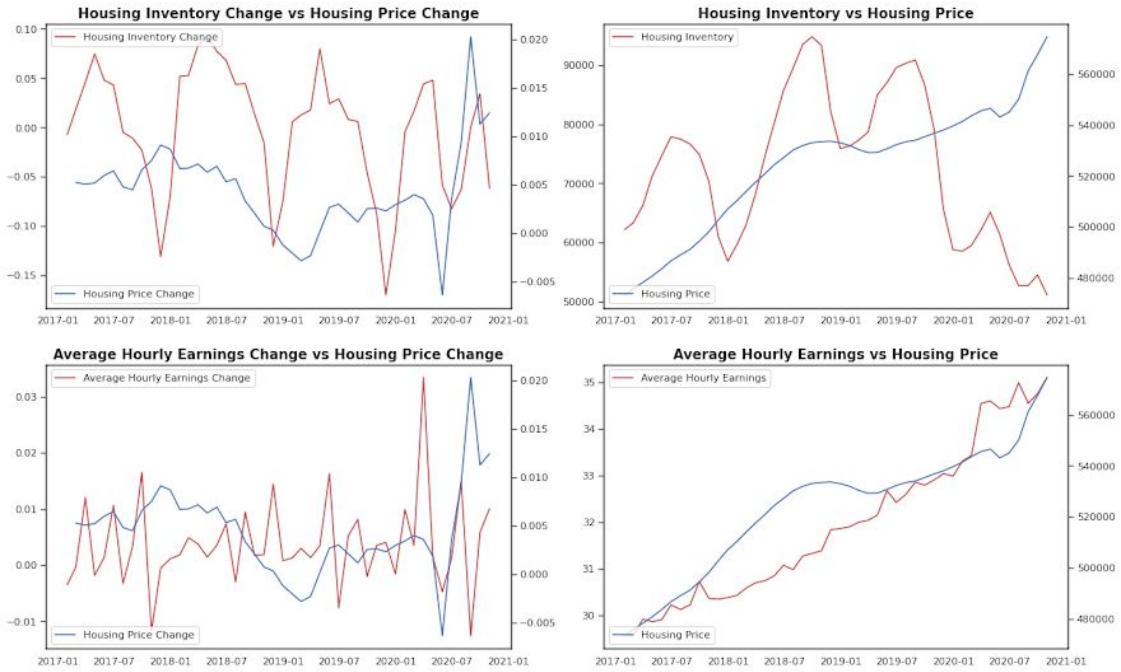

Mortgage_rate_value vs Housing Price
Mortgage_rate_value
Housing Price

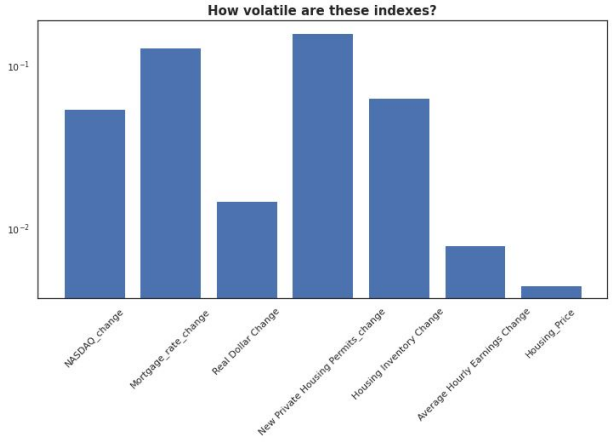

Real Dollar Change vs Housing Price Change
Real Dollar Change
Housing Price Change


Real Dollar vs Housing Price
Real Dollar
Housing Price


New Private Housing Permits_change vs Housing Price Change
New Private Housing Permits_change
Housing Price Change


New Private Housing Permits vs Housing Price
New Private Housing Permits
Housing Price



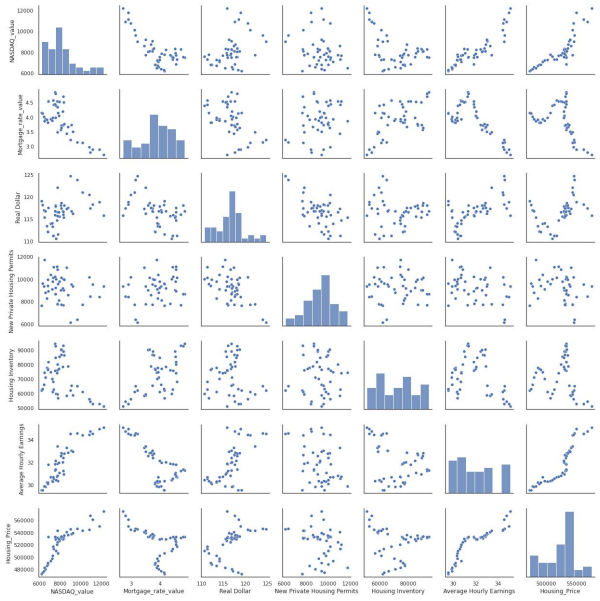


After checking on the trend and monthly change of each index visually, We decided to calculate the Pearson correlation coefficient between the variables to mathematically measure the statistical relationships. We also created a scatter plot matrix to compare them visually. We first started with the accumulated trends of the indexes.

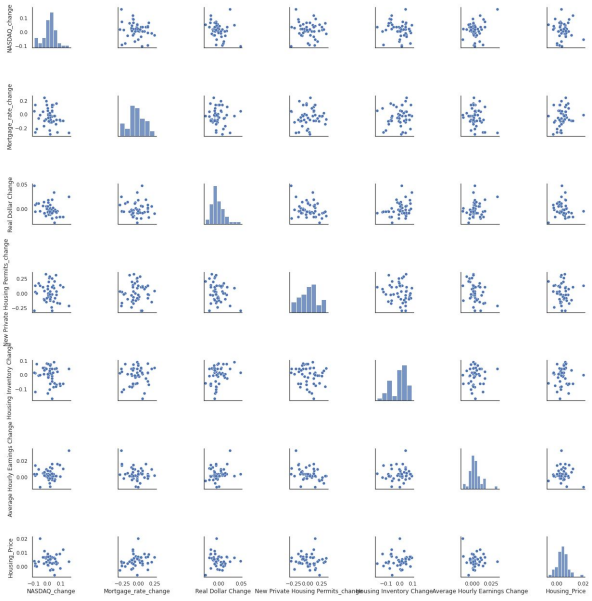
The chart shows that the trend of housing price is correlated with the trend of NASDAQ value and Average Hourly Earnings of Employees in California. Also, there seems to be a somewhat negative correlation between the trend of housing price and the mortgage rate trend. Intuitively, this makes sense because the more earning power in the state makes the housing more attractive, and a higher mortgage rate

means expensive interests for a mortgage loan, meaning there would be less demand. **But there is also a huge chance that this might be just a spurious correlation** because each index follows a certain trend. Just because they follow the same trend does not mean any causal relationship between them. For example, the US spending on science, space, and technology has a 99.79% correlation with Suicides by hanging, strangulation, and suffocation(Tyler Vigen, 2015). This certainly doesn't mean that they are causally related.

Next, we conducted the same process to the change rate of the variables. Unlike accumulated trends, there is no strong correlation between each index's rate of change.



[The trend of house prices vs. other economic indexes]



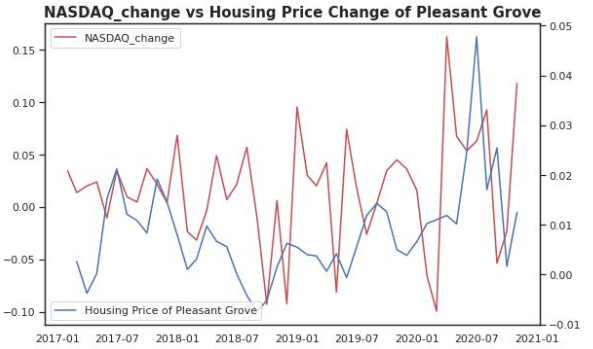
[The rate of change of house prices vs. other economic indexes]

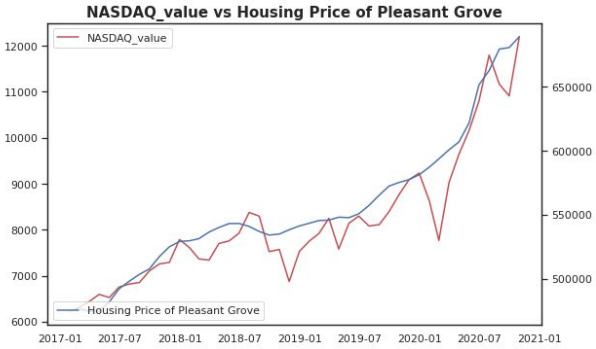
We were able to point out several things from the procedures;

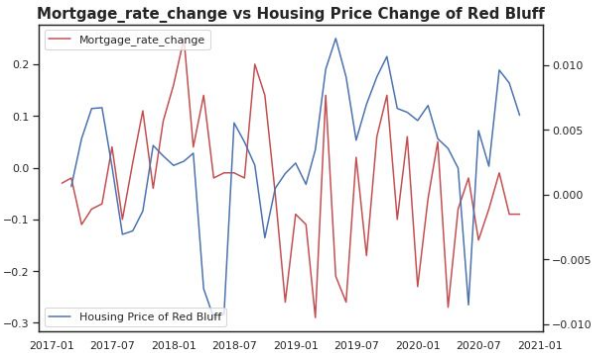
- The housing price of California has very low volatility compared to other economic indexes. This means that a house is a relatively stable asset.
- The trend of housing price had a strong correlation with the trend of other indexes, which wasn't the case between its monthly return rate and other indexes. The housing price shares the same trend with some indexes, but that might be just a spurious correlation.

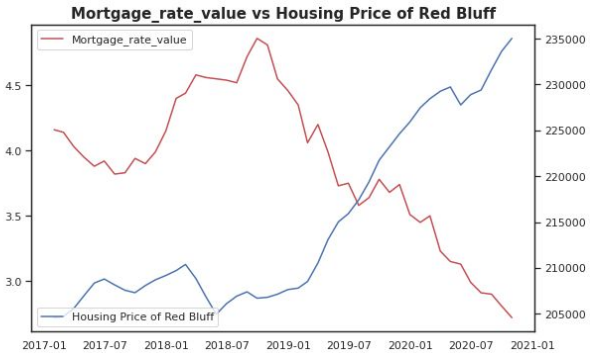
After checking on California's housing price at the state level, we decided to break the list down to the city level to figure out which city has the strongest correlation with the indexes. We specifically chose NASDAQ and mortgage rate for comparison. Then, we plotted them to visually confirm the resemblance. We chose 4 cities; Pleasant Grove, Red Bluff, Creston, and Corcoran. Each of them scored the largest

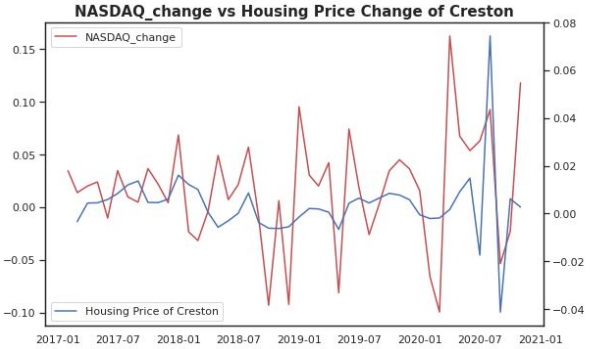
points in positive correlation with NASDAQ value, negative correlation with the mortgage rate, positive correlation with NASDAQ return rate, and negative correlation with the mortgage rate's change rate, respectively.

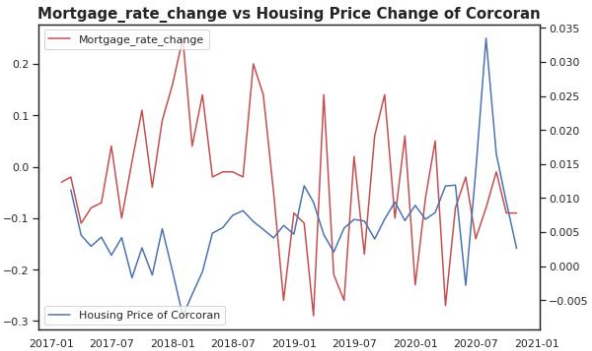


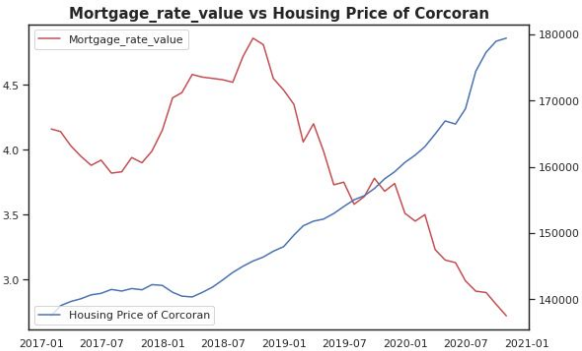












Looking at the charts, they all look like they have some sort of statistical relationship with the indexes. Again, however, there is a chance that this might be just spurious correlations.

Statement of Work

There is a collaborative effort with equal participation between Mel and Doyoung in the project. Our primary mode of communication is via Slack, and we ensure that each person has access to the entire project artefacts at any point in time. We use Google Colaboratory to host our Python notebook, which allows us to contribute simultaneously, while the source data is stored in Google Drive and seamlessly connected to the notebook.

Doyoung contributed significantly to the data collection and manipulation stages, especially with the Zillow, city crime, and ACS datasets. Mel assisted with additional cleaning/manipulation depending on the analysis.

Mel contributed to the analysis of how house prices change over time, and the statistical relationship between house prices and demographic variables such as crime and population, while Doyoung looked into the statistical relationship between house prices and 2019 ACS data, and the relationship between changes in house prices and other economic indexes.

From there, we both discussed and agreed upon the conclusions and recommendations presented in this report.

Bibliography

California Association Of Realtors. (2020). *Coronavirus Impacts on California's Housing Market*. California Association Of Realtors. https://www.car.org/knowledge/pubs/newsletters/Newsline/Coronavirus

Experian. (2019, November 18). *Median Home Values By State*. Experian. https://www.experian.com/blogs/ask-experian/research/median-home-values-by-state/

Hadden, J., Casado, L., Sonnemaker, T., & Borden, T. (2020, Dec 14). *21 major companies that have announced employees can work remotely long-term*. Business Insider. https://www.businessinsider.com/companies-asking-employees-to-work-from-home-due-to-coronavirus-2020

Madden, A. (2020, Dec 1). *Californians are still leaving the Bay Area during the pandemic*. MoveBuddha. https://www.movebuddha.com/blog/californians-leave-sf-pandemic/#sources

Melin, A. (2020, Sep 16). *Stripe Workers Who Relocate Get $20,000 Bonus and a Pay Cut*. Bloomberg. https://www.bloomberg.com/news/articles/2020-09-15/stripe-employees-who-relocate-to-get-20-000-bonus-and-a-pay-cut

Tyler Vigen (2015). *Spurious Correlations*. Hachette Books. https://www.tylervigen.com/spurious-correlations