Capstone Project 1 Final Report

The objective of this project is to analyze trends within the San Francisco Bay Area housing market using the Zillow API. It looks at how housing prices differ between the nine Bay Area counties and with counties throughout the United States. Many state that the Bay Area has become an unattractive place to live due to the rising and expensive cost of living. This project could be useful to a vast variety of people.

Home-seekers looking for affordable housing in the Bay Area could find this helpful by identifying which areas are within their budget. Realtors, home-sellers, and developers could get an idea of how to price their homes in a given market. Both scenarios have been personally observed as I've had 4 different families move in next door only to move out within a year and houses that have been up for sale for months without interest. This project could also be useful for politicians and tech companies as the housing market is deeply impacted by local landmarks and businesses. Employees at big companies like Facebook, Apple, and Google (which are headquartered in the Bay Area) are likely to buy houses close to work or public transit to keep commute times to a minimum. There are countless of other groups that housing prices affect.

My proposal for this project is to explore how housing has grown, and continues to grow, for the San Francisco Bay Area. I'll be checking if there are any statistical price differences among counties for my data analysis phase. It's quite possible that houses are more expensive because their localized to big businesses, public transportation, or urban areas. When it comes to modeling the project, I'll be using a method called autoregressive integrated moving averages, or ARIMA, which will be explained later.

Data was pulled from Zillow's Research Data hub which allowed me to get the data I needed for this project quickly and easily instead of making thousands of API calls. Setting the

data type to Zillow Home Value Index (ZHVI) on all single-family homes, I was allowed to download a csv file of the median estimated home value across each zip code.

After looking at the returned csv file, it seemed to have included all zip codes for the entire United States. To get zip codes that corresponded to those in the Bay Area, I wanted to start large and work my way down. I systematically started from state to counties to zip codes. My first task to narrow the dataset was to filter out every state except California. By doing this, it reduces the chance of matching a county name outside of California with similar names. Then I had to create a list of the county names that belong to the Bay Area.

I decided to create this list by pulling it from a webpage, the California Metropolitan Transportation Commision. By doing a little web scraping and making a list comprehension, I was able to create the list of the 9 counties of the Silicon Valley. Next was to filter out my dataset to only contain counties that matched with this newly created list and delete any unnecessary columns that I wouldn't need for this analysis like 'State' since the Bay Area is constrained to be in the California.

The dataset I was left with was one that indexed by zip code with the city name, county name, and the median home value for each month from 1996-2018 with each month corresponding to a different column. I wanted a dataset that only contained the median home values indexed by zip code. I decided to split the dataset into two, one with the zip code information (city and county names) in case it was needed later and another with just the home values which will be used to the majority of my analysis.

After briefly looking over the dataset and values, I noticed that there were some zip codes that started off as 'NaN' but suddenly had values as time progressed. This probably meant that the area was being developed and had no homes until a specific date. I decided to leave these values as 'NaN' for the time being because filling them with '0' will create huge

outliers and might skew the analysis at the time homes were being bought/sold in those developing areas.

The first step before modeling the data is to explore and analyze it. The way I approached this problem was to look at the median house price by zip code throughout the United States and how they've changed since the Great Recession ended. I used the median housing price in January 2010 to the most recent prices as of April 2018. I used those numbers to do my exploratory analysis on because the Great Recession heavily impacted house prices across the United States and hit a metaphorical 'reset' button on the housing market. The first method I looked at was to compare most recent median house price per zip code and compare the Bay Area to the rest of the United States. Secondly, I calculated how much much house prices changed by zip code as a percentage and did the same comparison. Lastly, I noticed that there seemed to be a division between Bay Area counties, mainly the 3 counties that inhabit most major businesses, therefore I briefly looked at any patterns within the Bay Area. In all my hypothesis testing, I used an alpha or significance level of 0.01.

2018 House Prices: Bay Area vs. United States

I started by testing the hypothesis that the mean price of a house in the Bay Area is the same for the United States. If statistically that isn't significant or reasonable, I'll reject that hypothesis and conclude that they aren't the same.

- $H_0 : \mu_{ba2018} = \mu_{us2018}$
- $H_1 : \mu_{ba2018} \neq \mu_{us2018}$

Given that I have the data for the entire population of United States homes and my sample sizes are above 30, I decided to use a z-test to compare the means. I manually calculated the z-score to be 17.65 which resulted in a p-value of 5.22×10^{-70}, well below my alpha. This means that

there's statistically significant difference in the mean housing price in the Bay Area and the United States.

2010-2018 House Price Percent Changes: Bay Area vs. United States

Next was to compare how, percentage-wise, how zip codes across the U.S. have changed between January 2010 and April 2018. The following formula was used to calculate the percent change.

$$percent\ change = \frac{(2018{-}04\ data) - (2010{-}01\ data)}{(2010{-}01\ data)}$$

Given that equation, if data wasn't present, or 0, in 2010, the percent change rises to *infinity*. That distorts the data and created problems during my analysis but since those instances only accounted for 3% and 4% of my Bay Area and U.S. data respectively, I decided to ignore those data points and analyze the remaining zip codes.

Once again, since we have the population statistics and the data is sufficiently large, the z-test was used again to calculate the p-value that the mean percent change is the same for the Bay Area and the United States. With this test, I got a z-score of 29.39 and a p-value of 3.84×10^{-190} which again is significant enough to reject the hypothesis that the percent changes are the same for the Bay Area and United States.

2018 House Prices: Bay Area Counties

Like I mentioned earlier, I noticed a few Bay Area counties (San Francisco, San Mateo, Santa Clara, and Marin) had a mean house price that seemed to be clustered together, separate from the other 5 counties. I wanted to investigate if their prices had a statistically significant difference. I ran t-test between each county (36 total comparisons) and calculated the

p-value of their significance assuming the null hypothesis that their means were the same. The following conclusions could be made from this analysis.

1. Napa county is only statistically different from San Francisco and Solano county. That may be due to Napa county's small sample size (7 zip codes). The confidence intervals were calculated and it's highly improbable that Napa county could have a higher mean than San Francisco. The confidence interval for the Napa county mean is between ($430,669 and $1,224,073) and San Francisco is between ($1,431,117 and $2,908,924).
2. Solano county, the cheapest county in the Bay Area on average, has statistically significant differences in mean price with every other Bay Area county. That means Solano county could be considered separate from other "lower" tier counties.
3. The "upper" and "lower" tier county house prices seem to be statistically different from each other. When comparing an "upper" tier county to a "lower" tier county, I was able to reject the null hypothesis on all instances except a few comparisons with Napa county. Again, that could be due to the small sample size of Napa county. When grouping "upper" tier counties and "lower" tier counties and running another t-test, the resulting p-value was 6.48×10^{-15}. That justifies rejecting the null hypothesis that the "upper" tier mean is the same as "lower" tier mean.

2010-2018 house Price Percent Changes: Bay Area Counties

As when comparing percent change between 2010 and 2018 Bay Area prices with the United States, I did the same when comparing Bay Area counties. T-tests were ran again between the 9 Bay Area counties due to small sample sizes. The following conclusions could be drawn from the t-tests.

1. Again Napa county suffers from a small sample size which makes drawing conclusive analysis difficult.
2. The county with the highest percent change since 2010, Santa Clara county, was statistically different from all other Bay Area counties.
3. Three North Bay counties (Marin, Napa, and Sonoma) have all seem the three smallest percent change since 2010 and are each statistically different from the other six counties.
4. The only other significant test was between San Mateo county and San Francisco county. Of the remaining 5 counties (Alameda, Contra Costa, San Francisco, San Mateo, and Solano), those two had seen the highest and lowest percent change respectively with San Mateo seeing a 105% percent change and San Francisco seeing a 87% percent change.

As discussed earlier, the housing market in the Bay Area is more aggressive when compared to other parts of the United States meaning the prices are higher and still seem be growing. The next step was to model the data to forecast future prices.

Since the data is a time series, methods like linear regression weren't appropriate for this project. Instead, an ARIMA model was used to forecast the data. ARIMA stand for autoregressive integrated moving average, which means it uses a number of different parameters to model data to control a certain aspect of the ARIMA model. One parameter represents an autoregressive lag (p), another the differencing order (d), and lastly the moving average order (q). The ARIMA model is most effective if the data being modeled is considered stationary, that means that over time, the mean, variance, and autocorrelation are constant over

time. If your data isn't stationary, there are a number of methods to induce to become stationary.

There are two methods to test whether your data is stationary. The first is to visually plot your data and visually inspect if there's any pattern to the mean and variance over time. If the plots look relatively flat and without form, then you could say your data is stationary. For a numerical/statistical approach, the Dickey-Fuller test could be used. This test returns a statistic and p-value which could be used with a predetermined confidence level to conclude is your data is stationary.

The most effective way to induce a stationary dataset is to difference out trends such as the moving average, weighted moving average, shifted data, and seasonality. The the number of times the data is differenced will determine you parameter *d*. Parameters *p* and *q* are determined by plotting the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the stationary data. Where these plots reach a predetermined significance level will determine your parameters *p* and *q*.

After a model is fitted to the data and given a RSS/RSME score, it is used to forecast the housing price for the next period in time with a 95% confidence interval. Functions were created in Python to help automate the process.

For purposes of this project, the data was aggregated by county and the average of the prices were taken. That left a group of the nine Bay Area counties with housing price data dating back to January 2012. January of 2012 was taken instead of 2010 as in the previous reports because it was proving difficult to induce the data to become stationary. The data was then split 70:30 into training and testing sets. The first 70% of the data was taken chronologically to be the training set and the last 30% to be the testing set.

The logarithm was taken of each of the 9 data sets to reduce the scale and then each was passed to the four differencing techniques mentioned above to determine which produced the most stationary dataset.

Alameda County

For the Alameda training set, differencing provided a significant enough p-value (0.023) to consider it stationary. Fitting the ARIMA model to the Alameda training set with parameters according to the first order differenced stationary data, it produced a reasonable model with roughly normalized residuals and a RMSE of $17.44. When forecasting the next 24 months and comparing it to the test data, the RMSE inflates to $32,834, however the overall trend of the data seems to be captured as the actual data lies within the 95% confidence interval predicted by the model. Another thing to look for going forward is the overall trend of the forecast and 95% confidence intervals provided by the model. For Alameda, both have an upward trend which alludes to the fact that the cost of living in Alameda county will keep growing.

Contra Costa County

The data for Contra Costa county proved to be difficult to get to become stationary with the weighted moving average technique providing the smallest p-value (0.064) and most stationary result. The fitted ARIMA model produced a roughly normalized residuals and a RMSE of $165.71 on the actual data. When forecasting data and comparing it to the test set, the values tend to diverge from each other with the actual values outside the 95% predicted confidence interval after about 12 months. Regardless, both the test data and predictions seem to be rising hinting at its growth but at different rates. More fine tuning of the model will need to be considered to capture the trend better or perhaps more differencing to produce a more stationary training set.

Marin County

Unlike Contra Costa county, Marin county didn't pose as many problems to induce it into becoming stationary. The decomposition technique produced a residuals that gave a p-value much smaller than 0.01. Since the decomposition technique removes the overall trend and seasonality trend, the differencing order becomes 2. However, the resulting model didn't produce normalized residuals of the fitted values and stationary values; the plot seems to be right skewed. The model still performs well producing a RMSE of $59.55 on the training set and $15,022 on the test set capturing the trend of the data. One thing to note though is the exponentially increasing confidence interval of the model which signifies the models short term value to forecast data reliably. The long term predictions indicate that the cost of living in Marin county could explode or crash at any moment. A combination of fine tuning parameters or more data could help improve the model.

Napa County

Napa county training data produced the most stationary data set when differencing the exponentially weighted moving average (p-value $\ll$ 0.01). The Napa county model predicted the training set with a RMSE of $219.45 and test set with $59,141. The model was also able to predict the test set within the 95% confidence interval. The continual increase of the confidence interval also indicates that Napa county housing prices are steadily rising.

San Francisco County

The only method that produced stationary values of the San Francisco training set was the decomposition technique but even then, the residual plot doesn't appear to be normal and skewed to the right. The following ARIMA model predicted the training set with a RMSE of $30.21 but miscalculated the test set forecast by giving a $275,505 RMSE. By looking at the forecast plot, even though the test data was captured by the models 95% confidence interval,

the model predicted a negative trend in the forecast and exponentially increasing confidence intervals signifying either poor fit or unpredictability in the housing price in San Francisco county. Like the case of Marin county, the model indicates that prices with explode or crash drastically which doesn't provide confidence. Fine tuning the parameters or obtaining more data might be necessary to provide more insight.

San Mateo County

The model for San Mateo county was used to fit the residuals of the decomposed San Mateo data. It gave an output that predicted the training set with a RMSE of $507.42 and while it scored a RMSE of $63,676 on the test set, it correctly predicted the data set with 95% confidence given the volatility of the data in recent months. The model seems to have captured the trend of the data and indicates that San Mateo county will continue to grow according to the predicted values and confidence interval.

Santa Clara County

Double differencing the Santa Clara county training set produced the most stationary data I was able to work with for this instance. The model performed well on the training set scoring a RMSE of $56.32 but when it came to the test set, like San Francisco county, it scored a high RMSE of $186,182 while correctly predicting the test data within the 95% confidence interval of the model. That's partly because the model has an exponentially increasing confidence interval which alludes to the short term value of this model for this data set. This example is similar to Marin and San Francisco counties where more fine tuning and/or data is necessary to create a better model.

Solano County

For Solano county, double differencing also produced the most stationary data set resulting in an ARIMA model that scored a RMSE of $33.63 on the training set. For Solano

county's test set, the model was able to capture the data within it's 95% confidence interval and score a RMSE of $12,061. When looking closely at the confidence interval, it indicates that the Solano house prices will either stabilize or continue to grow but not decrease (barring any political or economical setback).

Sonoma County

Lastly, for Sonoma county, decomposing it produced stationary residuals that the model was able to use to predict the training set with a RMSE of $113.03. Forecasting it and comparing it to the test set scored a RMSE of $5,106 which is exceptionally well when compared to the other Bay Area counties. Like most of the other Bay Area counties, the overall trend of the test and predictions values indicate that Sonoma county will continue to grow as well.

The objective behind this project was to analyze how the cost of living within the San Francisco Bay Area has changed/grown over the years and to look at any developing patterns to forecast future prices. We saw that the Bay Area has grown faster than most counties around the United States and are still growing. Models were developed for each of the nine Bay Area counties and according to the predicted values, most of them will continue to grow in the future. Of course, the models could be improved upon by including more data, higher degrees of differencing to induce more stationary data, or fine tuning parameters.

In the future, those methods could be used to produce better models but modeling strictly the time-series data of living costs is highly restrictive. Housing prices depend on a number of factors like politics, economics, region, size, and amenities to name a few. More complex models could be developed to incorporate these factors to get more precise predictions.