MGT 6203 Group Project Final Report

US Key Regional Housing Price Analysis

Team 1:

Kevin Hsu

Karthick Mani

Venkataraghavan Punnapakkam Krishnan

Julia Lee

Yoonseo Lee

Contents

Introduction

This project investigates the correlation between housing prices and income levels in five selected cities, as well as other factors such as structural features of the houses. In the last decade, the US has seen a constant increase in housing prices across the country after recovering from the burst of the housing bubble in 2008. During the recovery and growth of the economy, some cities have thrived and grown tremendously in terms of GDP per capita, mainly those that have a high concentration of high-tech corporations. However, to study this relationship, we recognize that there are many factors that can affect housing prices, such as the country's economy, mortgage rates, location, and others, which makes forecasting housing prices extremely difficult. On the other hand, overall salary levels are easier to predict, as they are often tied to company performance, industry growth, etc. By investigating changes in salary level with changes in housing prices in a certain area, we may be able to predict how its housing prices change, since expectations are that as income levels rise, people are willing to pay more to purchase a house, and vice versa.

As such, this project will compare the change in housing prices with the change in salaries in three cities where many big-technology corporations are located, San Francisco, Austin, and Seattle, and two cities whose economies rely mostly on sectors outside of technology, Cleveland and Nashville. The inclusion of Cleveland and Nashville will allow us to study housing prices patterns in areas where we do not expect income levels to have risen as rapidly in the past 20 years.

The analysis performed for this project serves two purposes:

1. Regression analysis - Establish a regression model across different cities using house properties to predict housing prices
2. Time series forecasting - Investigate how housing prices vary with changes in the income level of the residents in selected cities

Initial Hypothesis

We expect changes in income to be closely related to changes in housing prices across all cities in the US, which will allow our models to accurately predict housing prices.

Data collection, cleaning, transformation

This section summarizes the datasets collected and the data cleaning process that took place before we began building our models.

- Datasets for time series analysis
    - We extracted historical housing data for our cities of interest from Zillow[i], an American marketplace website. The Zillow Home Value Index (ZHVI) All Homes (SFR, Condo/Co-op) Time-series, Smoothed, Seasonally Adjusted dataset[i] consists of monthly home values in the 35th to 65th percentile range from 2000 to 2022 for all metropolitan cities in the United States. As part of the data cleaning process, a subset was created for the 5 cities selected for this project. This process included removing cities that have the same names but in different states, as well as columns in the dataset that were unnecessary for the analysis, such as the Region ID and Size Rank.

	RegionName	State	X2000.01.31	X2000.02.29	X2000.03.31	X2000.04.30	X2000.05.31	X2000.06.30	X2000.07.31	X2000.08.31	···	X2021.12.31	X2022.01.31
	<chr>	<chr>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	···	<int>	<int>
1	San Francisco	CA	451990	454719	458279	465835	474124	483099	492214	501766	···	1637944	1645882
2	Austin	TX	199517	200598	201305	202289	202752	202596	202694	203839	···	639959	653688
3	Cleveland	OH	83742	83984	84178	84619	85228	85710	86086	86308	···	99595	100921
4	Nashville	TN	141650	141992	142231	142864	143343	143584	143877	144130	···	403359	413655
5	Seattle	WA	270916	272728	274382	277607	280946	283647	286658	288575	···	961135	977007

    - The Federal Reserve Economic Data (FRED)[ii] provides the annual per capita personal income from 1969 to 2020. Data from all 5 cities was collected, and all income data prior to 2000 was removed afterward.

	1969-01-01	1970-01-01	1971-01-01	1972-01-01	1973-01-01	1974-01-01	1975-01-01	1976-01-01	1977-01-01	1978-01-01	...	2011-01-01	2012-01-01	2013-01-01	2014-01-01	2015-01-01	2016-01-01	2017-01-01	2018-01-01	2019-01-01
SEAT653PCPI	4660	4758	4873	5219	5816	6450	7190	7891	8641	9829	...	50833	55280	56117	60226	62952	65485	68680	72685	7597(
SANF806PCPI	5323	5713	6066	6555	7011	7710	8510	9241	10072	11216	...	64682	69579	70336	75094	81229	85648	91236	97681	10240(
NASH947PCPI	3456	3629	3895	4253	4787	5227	5526	6157	6788	7607	...	44236	46653	46895	49400	52149	53903	55903	59105	6151(
CLEV439PCPI	4484	4631	4869	5322	5878	6502	6908	7553	8394	9302	...	42522	44632	45065	47211	48874	50102	51860	53881	5529-
AUST448PCPI	3363	3672	3990	4312	4676	5141	5779	6366	6874	7905	...	44698	47710	48565	51355	52657	53419	57183	60764	6246(

- Dataset for Regression analysis
    - Zillow[iii] has millions of for-sale and rental listings on its website. To collect data on all current listings on Zillow for our 5 cities, we utilized Bardeen, a workflow automation tool, to scrape the listings on the search pages for each city. This allows us to store each property's attributes, such as the price for sale, number of bedrooms, size, etc. to a CSV file for each city. A column was added in each CSV file to label which city the listings were for before all CSV files were compiled into one. Listings that were missing certain attributes or contained N/A cells were removed. The column that specifies which year the house was built was transformed into a numeric number to indicate how old the house is as of 2022.

	Price	Bed	Bath	House.sqft	old	Heat.system	Cooling.system	Garage	Total.sqft	City
	<dbl>	<int>	<int>	<dbl>	<dbl>	<chr>	<chr>	<dbl>	<dbl>	<chr>
1	18500000	5	7	8040	0	Electric	Central air, electric	4	34412.4	Austin
2	18000000	3	8	14025	151	Fireplace(s), multiple heating units, natural gas	No data	1	3920.0	Austin
3	14999000	5	7	8549	20	Central, fireplace(s)	Ceiling fan(s), central air	5	574992.0	Austin
4	13000000	5	7	6528	83	Natural gas	Multi units	3	17424.0	Austin
5	11250800	5	8	7947	0	Central	Central air, electric, see remarks	4	50529.6	Austin
6	9950000	4	5	4346	67	Central	Ceiling fan(s), central air, multi units	2	26571.6	Austin

Initial Discoveries (Exploratory Data Analysis)

Before building our models, we wanted to confirm that the five cities we selected were appropriate for our analysis and in line with our expectations. We expected San Francisco, Seattle, and Austin, which have some of the world's largest technology companies, to experience an obvious upward trend in housing prices during the past 20 years, while Cleveland and Nashville would see only small changes in housing prices. Figure A below shows all five cities' median (35th to 65th percentile) housing price trend using the Zillow dataset for our time-series analysis.

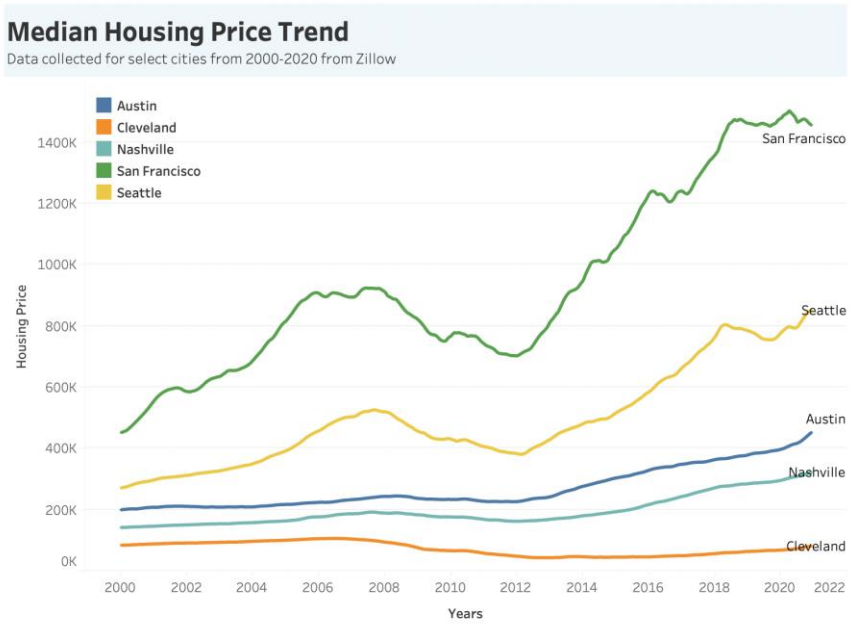


Figure A – Housing Price Trend in Austin, Cleveland, Nashville, San Francisco, Seattle

An unexpected observation is that Nashville's housing price appears to increase at a similar rate as Austin's. We decided to proceed using Nashville as a selection, as it may allow a mini comparison between Nashville and Austin when investigating how both cities' income levels have changed, given how vastly different each city's economy is structured.

In Figure B below, when comparing how housing prices and income levels have changed in each city year over year, our observations were mostly in line with our expectations. Across all cities, we noticed income levels do not fluctuate as much as housing prices. The percentage increase in per capita income has maintained between 0% to 10% in most years except in 2009 due to the Great Recession. Housing prices have fluctuated more, as the graph shows double-digit percent increases in all five cities and a negative growth in housing prices even in years when income level has increased. Overall, the graphs for Cleveland and Nashville have flatter lines, except during the Great Recession and a short period after that, as we had expected. In San Francisco and Seattle, which have long been tech hubs, housing prices have grown over 10% multiple times, while salaries in numerous years have also almost reached 10% increases throughout the past 20 years. Austin's lines are flatter than expected, which is also understandable as it is only starting to become a tech hub with companies like Amazon and Google having only just opened an office there in recent years. The impact of big-tech corporations raising the income level in the area due to high salaries and increases in the cost of living may need to be observed a few more years before an obvious trend is apparent.

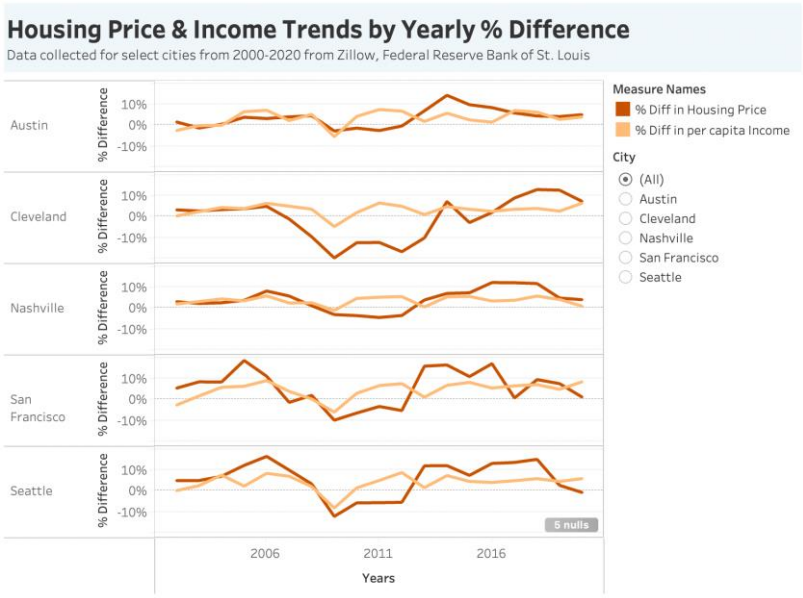


Figure B – Housing Price and Income Trends by Yearly Percentage Difference

Models

As our analysis consists of two components, regression analysis and time-series forecasting, the following models are created:

- Regression analysis- Linear regression, random forest
- Time-series forecasting – Multivariate (Vector Autoregression)

Linear Regression

Before the linear regression models were created, we created a correlation plot in Figure C to identify the relationship of the key variables amongst each other. We noticed that the number of bedrooms and bathrooms were highly correlated, and both were highly correlated with the size of the house as well, meaning some sort of transformation would have to be done to these variables for our regression models.

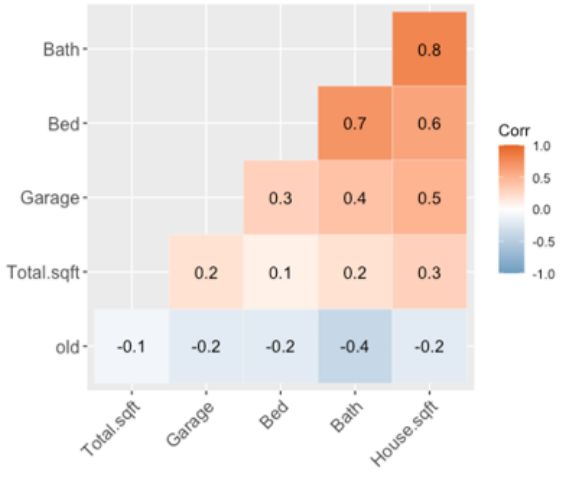


Figure C – Correlation Plot

For our regression analysis, the following models were created using the Zillow current listing dataset.

Model	Model Formula	R^2	Adjusted R^2	P>0.05
1	Price ~ Bed + Bath + City	0.1077	0.1047	CityNashville
2	log(Price)~ Bed+Bath+log(House.sqft)+old+Garage+log(Total.sqft)+City	0.8555	0.8547	Garage
3	log(Price)~ Bed+Bath+log(House.sqft)+old+log(Total.sqft)+City	0.8555	0.8548	-
4	log(Price)~ Sum+log(House.sqft)+old+log(Total.sqft)+City	0.8265	0.8257	old
5	**log(Price)~ Sum+log(House.sqft)+log(Total.sqft)+City**	**0.8263**	**0.8256**	**-**

We experimented incorporating and removing different variables based on the Adjusted R^2 values of each model and the P-values of each variable. The "City" variable is a dummy variable with Austin as the base case and the "old" variable is the age of the house as of 2022. In the first model, we used the number of bedrooms, number of bathrooms, and the "City" variables to see how well the model can explain housing prices but resulted in a low Adjusted R^2. In the second model, we performed a log transformation on price, and included more variables. We noticed the number of garages in the house was not a significant factor, thus it was removed in the third model. Next, we noticed there was a negative coefficient for the "Bed" variable, which is counterintuitive, as one would expect the house of a price to increase the more bedrooms it has. The correlation plot in Figure C also shows the "Bed" and "Bath" variables are highly correlated, so we replaced the two variables with a new variable, "Sum", which is the total number of bedrooms and bathrooms. Our fourth model indicated that the age of the house was not a significant factor, which led us to our fifth and final model, shown below.

```
Call:
lm(formula = log(Price) ~ Sum + log(House.sqft) + log(Total.sqft) +
    City, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2954 -0.2941 -0.0249  0.2435  3.5102

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.035437   0.308256  22.823  < 2e-16 ***
Sum             0.080832   0.008962   9.019  < 2e-16 ***
log(House.sqft) 0.749267   0.046968  15.953  < 2e-16 ***
log(Total.sqft) 0.066997   0.014167   4.729 2.43e-06 ***
CityCleveland  -1.707338   0.040325 -42.340  < 2e-16 ***
CityNashville  -0.321566   0.033199  -9.686  < 2e-16 ***
CitySF          0.719643   0.039826  18.070  < 2e-16 ***
CitySeattle     0.215239   0.037933   5.674 1.62e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4925 on 1783 degrees of freedom
Multiple R-squared:  0.8263,    Adjusted R-squared:  0.8256
F-statistic:  1212 on 7 and 1783 DF,  p-value: < 2.2e-16
```

Figure D – Final Regression Model Summary

This model achieved an acceptable Adjusted R^2 of 0.8256. The negative coefficients of the "Cleveland" and "Nashville" variable indicate that log housing prices are lower in those cities compared to Austin, while keeping other variables

constant. The remaining variables have positive coefficients, so an increase in any of these variables, while keeping the others constant, increases the log price of a house, which we expected. Additionally, to confirm that income level is a significant factor, we removed the "City" variable and the Adjusted R^2 fell to 0.4918. Clearly, the location of a house is a significant factor in determining its price, and we believe the main reason is due to the area's overall income level. Although this conclusion was in line with our expectations, we noticed some non-linearity even after the log transformations, shown in the Q-Q plot and Residuals vs Fitted plot below.

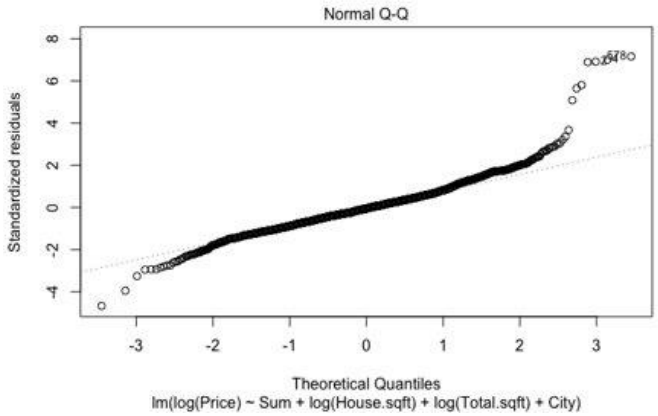


Figure E – Q-Q Plot

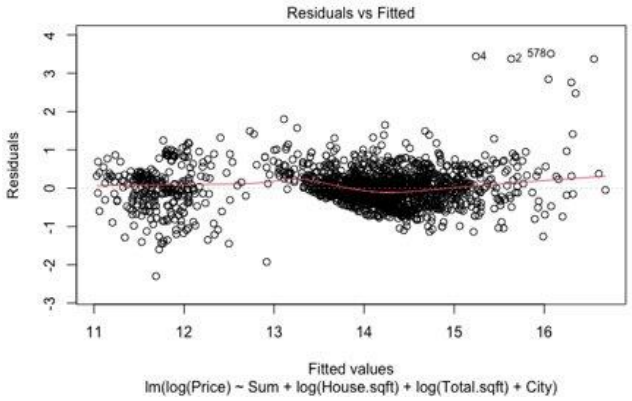


Figure F – Residuals vs Fitted Plot

Random Forest

The next step of our analysis consisted of using random forest to further analyze the independent variables we identified. In our first random forest model, we included all variables prior to any modification. Although the model achieved a high level of percentage variation explained, we observed the Mean Decrease Accuracy (%IncMSE) was the lowest for the "Bed" variable, which is counterintuitive again since this suggests that the variable does not provide much accuracy to the model. We then created a second model in Figure G and replaced the "Bed" and "Bath" variable with "Sum", similar to the linear regression model earlier.

```
Call:
 randomForest(formula = log(Price) ~ . - Bed - Bath, data = train,
               Type of random forest: regression
                     Number of trees: 1000
No. of variables tried at each split: 2

          Mean of squared residuals: 0.1836393
                    % Var explained: 86.74
```

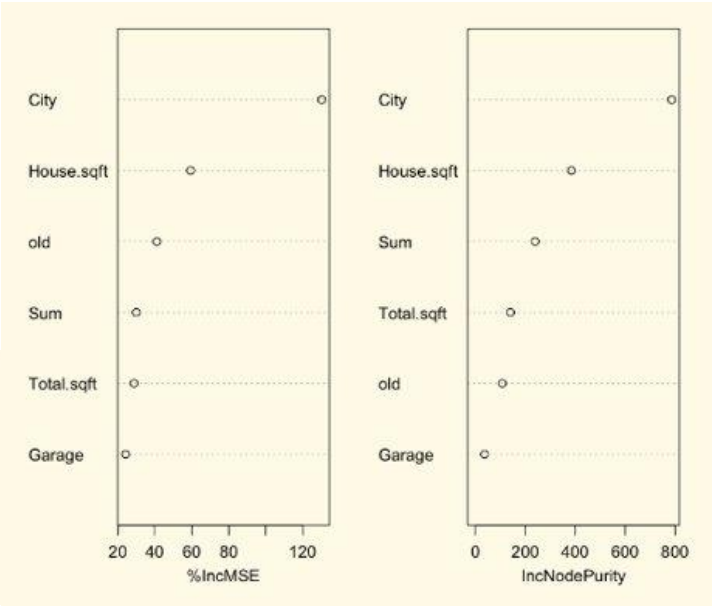


Figure G – Final Random Forest

By looking at the Mean Decrease Gini (IncNodePurity) we observed that the "City", "House.sqft", "Sum", "Total.sqft" are the most important variables, in descending order. This result is consistent with the regression model created earlier,

but the R^2 value is relatively higher under the random forest than regression. The linear regression model earlier assumes a linear relationship, but the higher R^2 in the random forest model suggests the independent variables explain housing prices better when we do not assume a linear relationship. This is in line with our interpretations of the Q-Q plot and Residuals vs Fitted plots earlier.

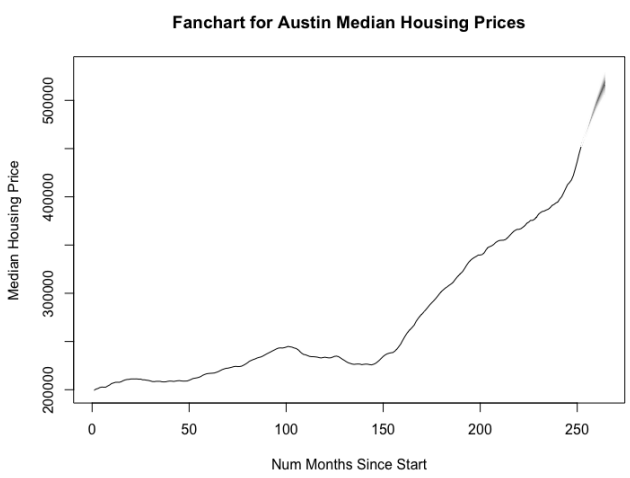
Multivariate Time Series Model (Vector Autoregression)

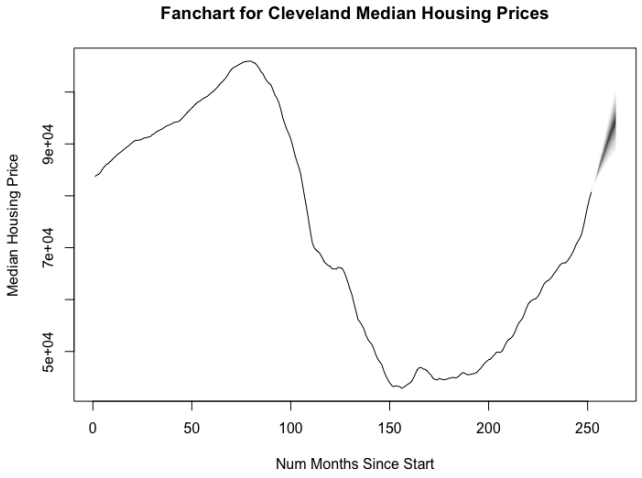
There were some limitations we faced in our previous models, such that they were unidirectional in their relationship, meaning one dependent variable had many independent variables. So, the next model we tried was a Vector Autoregression (VAR) model, which is a multivariate time series linear model that takes endogenous variables in the function as lagged values of the other endogenous variables. These variables act similar to dependent variables, as they are changed by their relationship with other variables in a model[viii]. Using a VAR model would allow our prediction of a city's median housing price to be based on each respective city's income (and vice versa). Since our analysis is to only predict housing prices, we will not be showcasing any predicted incomes based on housing prices here in the report but will show San Francisco's income prediction within the R code as reference.

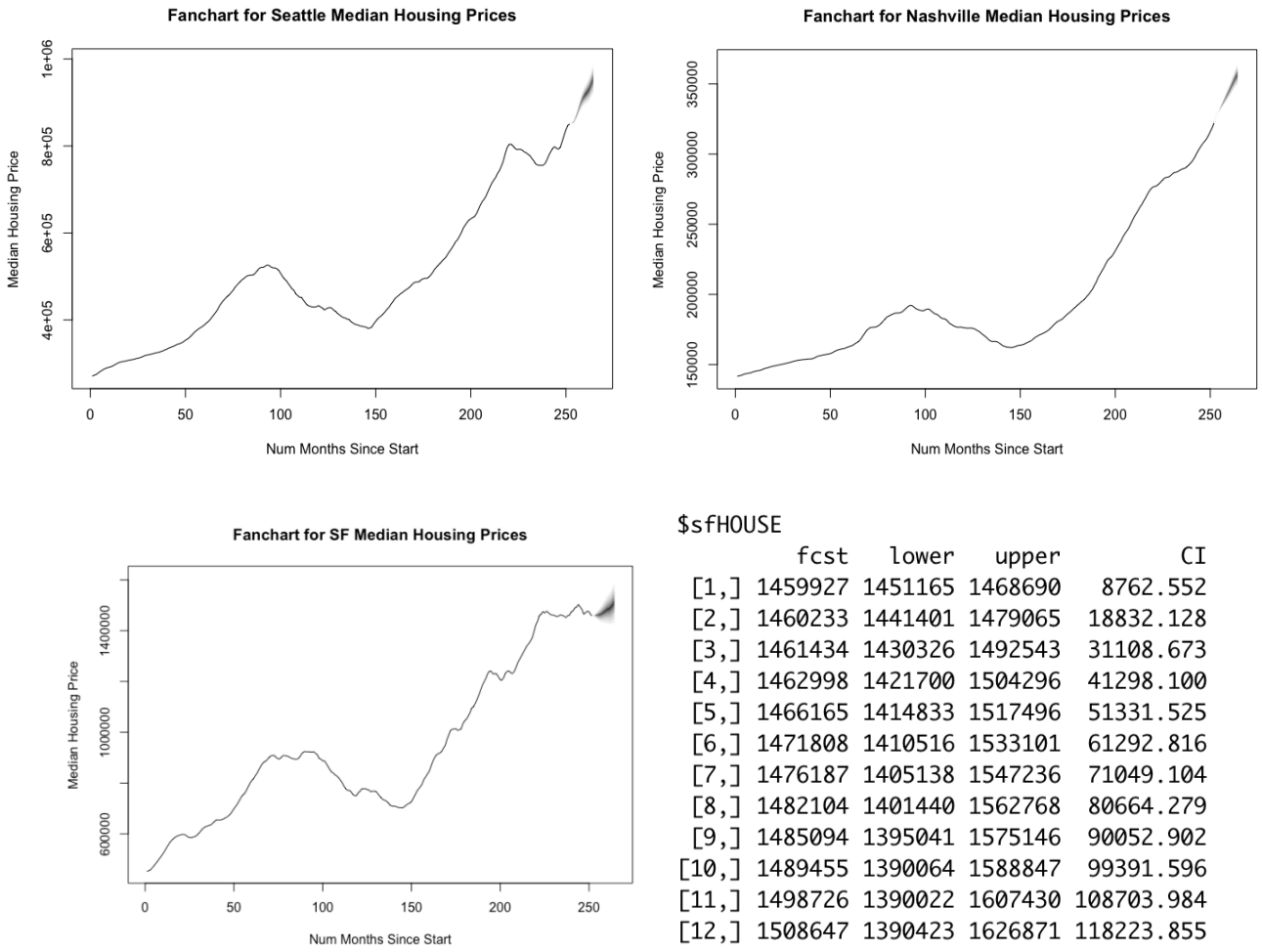
Since the FRED income dataset is only up to the end of 2020, and the Zillow median housing price dataset starts from 2000, we filtered the data according to those data ranges. Another problem we faced is that the income dataset was on an annual basis, whereas the housing price data was monthly, so to have our forecast on a smaller time scale, the annual income for each year was applied to each month of that year in order to match the housing price per month scaling. We recognize this would affect our model on accuracy as each data point may not be as close to their true value as it could be with a monthly income per capita data point.

Using the vars package in R, we have our 5 models, one for each city of research. Here we have the model for San Francisco: *sf_model1 <- VAR(sf_1, p = 14, type = "const", season = NULL, exog = NULL)* with *sf_1* being the bound VAR variables (income + median housing price) for VAR estimation like so: *sf_1 <- cbind(sfIN, sfHOUSE)*, with "sfIN" and "sfHOUSE" representing income and housing price as time series data, respectively. Note that we chose 14 as the integer lag order based on our understanding from literature papers[x], but using VARselect() in the vars package would choose your lag integer based on your system.

Below we have each forecasted median housing price for Austin, San Francisco, Seattle, Nashville and Cleveland. We represented these forecasts as fan charts which are time series VAR forecasts with shaded confidence regions of each prediction. An example of San Francisco's prediction code in R: *sf_forecast <- predict(sf_model1, n.ahead = 12, ci = 0.95)*, which takes each respective city's model, a forecast of 12 months ahead, with a 95% confidence interval. Note that the x-axis for each chart is the number of months since 2000-01-31, so the tick '100' would represent 100 months after 2000-01-31, which is 2008-05-31.







$sfHOUSE

	fcst	lower	upper	CI
[1,]	1459927	1451165	1468690	8762.552
[2,]	1460233	1441401	1479065	18832.128
[3,]	1461434	1430326	1492543	31108.673
[4,]	1462998	1421700	1504296	41298.100
[5,]	1466165	1414833	1517496	51331.525
[6,]	1471808	1410516	1533101	61292.816
[7,]	1476187	1405138	1547236	71049.104
[8,]	1482104	1401440	1562768	80664.279
[9,]	1485094	1395041	1575146	90052.902
[10,]	1489455	1390064	1588847	99391.596
[11,]	1498726	1390022	1607430	108703.984
[12,]	1508647	1390423	1626871	118223.855

Figure I: Forecasted 12-month Median Housing Prices for San Francisco

A summary for San Francisco's median housing prices can be seen above with each forecasted price, lower and upper bound and each confidence interval for 12 months out. Note that this is forecasting 12 months out from 2020-12-31, which would be the year 2021. Since our median housing price dataset is up to 2022, we compared how our model prediction performed versus real housing prices in 2021.

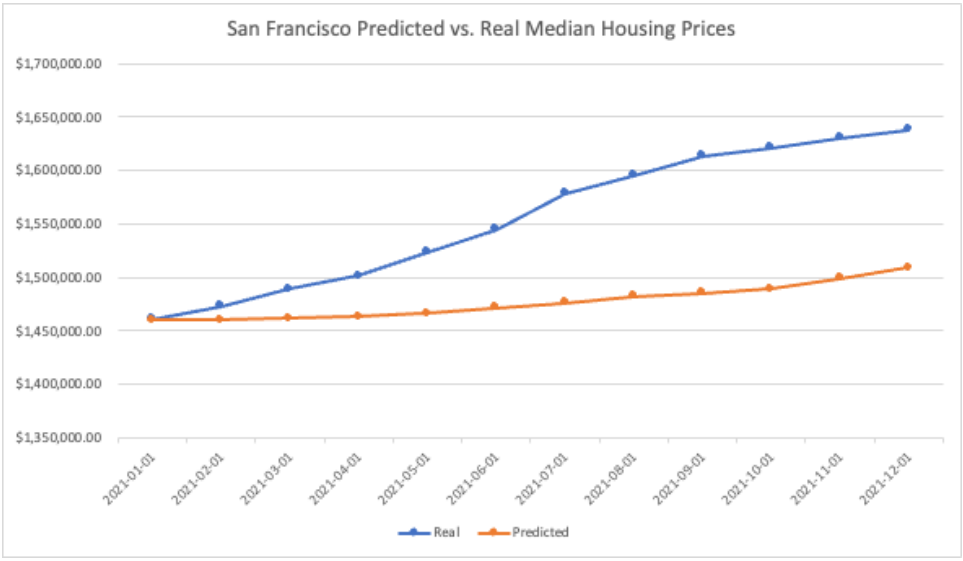


Figure J: Comparing Predicted and Real Median Housing Prices in San Francisco

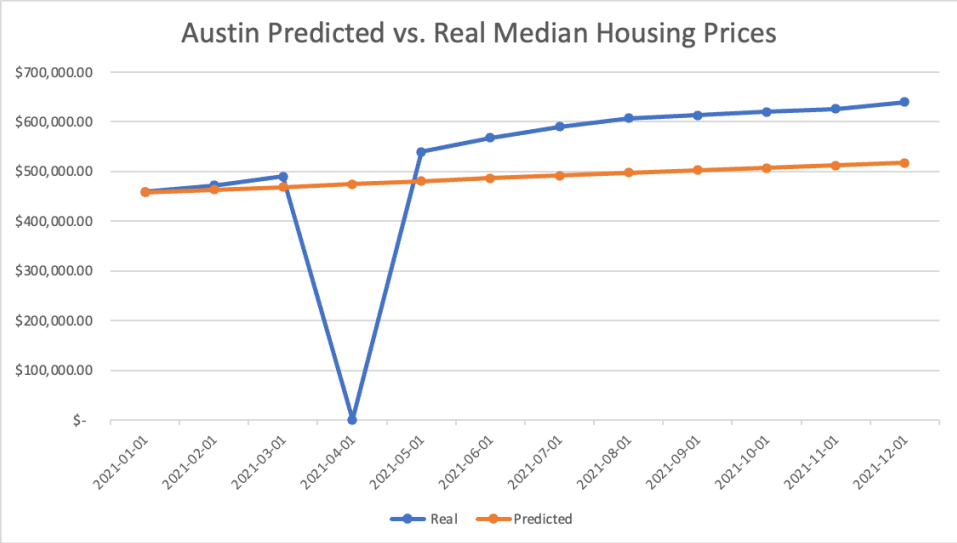


Figure K: Comparing Predicted and Real Median Housing Prices in Austin
(Note: Austin's median housing price data point on 2021-04 was unavailable)

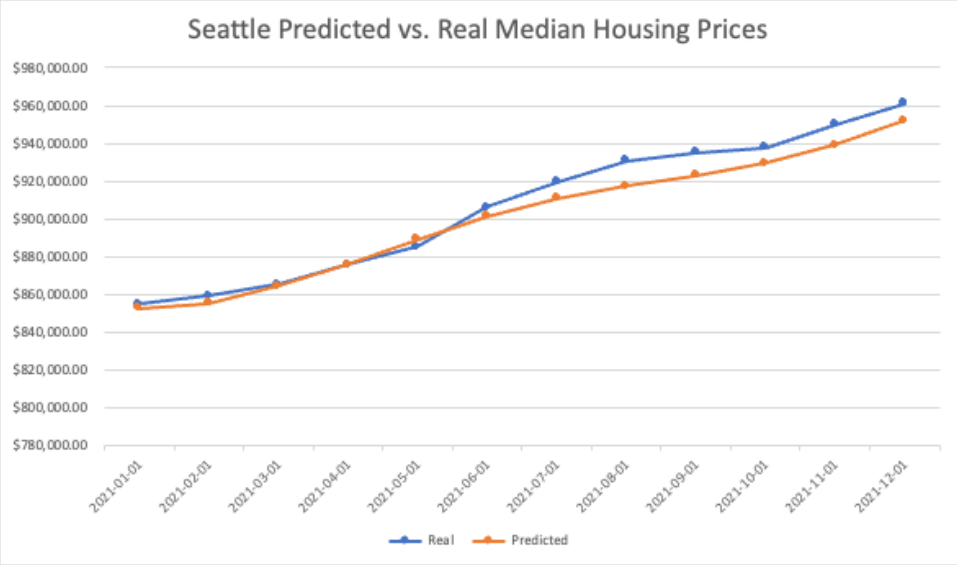


Figure L: Comparing Predicted and Real Median Housing Prices in Seattle

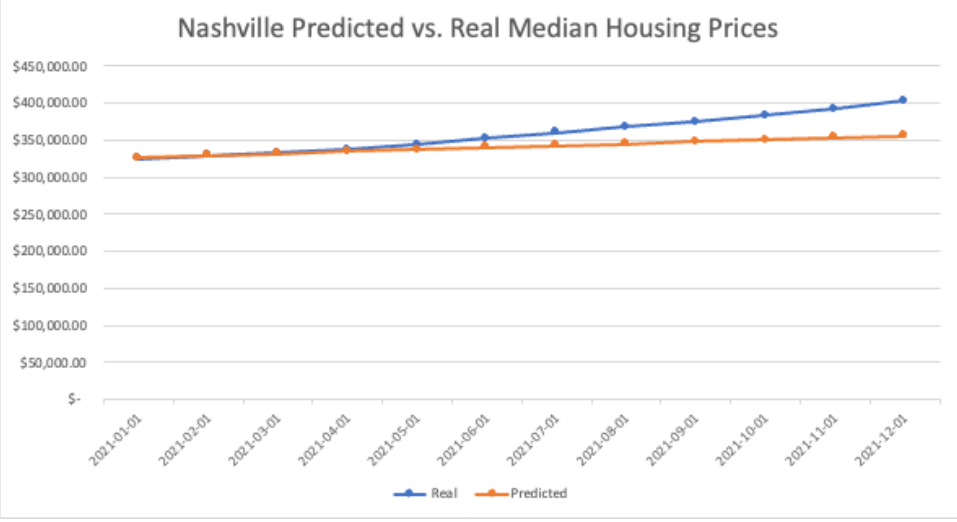


Figure M: Comparing Predicted and Real Median Housing Prices in Nashville

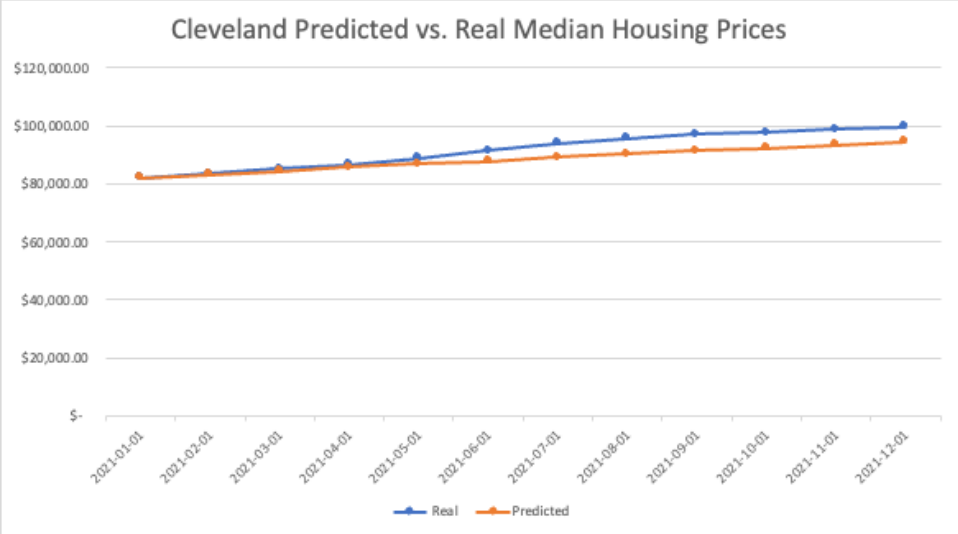


Figure N: Comparing Predicted and Real Median Housing Prices in Cleveland

We can see from Figure J above that the real median housing prices in San Francisco have increased at a much higher rate than what was predicted by our model. Our VAR model for Seattle, Nashville and Cleveland have predicted median housing price for these cities fairly well compared to the actual observed prices for 2021. Generally, our model for each city is underpredicting, which is another sign that our data is autocorrelated, a typical characteristic of time series data. Having only two variables, it is not a surprise to see that all 5 city models were autocorrelated, but one could experiment with the lag orders to combat that issue and hopefully show signs of no autocorrelation- which would be the next step in improving our models. Similar to autocorrelation, we saw heteroskedasticity in all models except for Cleveland's. Cleveland's model shows no degree of heteroskedasticity, most likely from the drastic rise and falls in median housing price throughout the 20 years, an extreme we do not see in the other cities, as shown in Figure B.

From these results, it is clear that more factors should be considered for a city's housing prices. The overall market, state taxes, development in the city and many other variables need to be considered when deciding appropriate housing prices. Like our regression analysis above, it would be a natural next step to try VAR using more than one variable and see how our model performs. Having more variables would reduce correlation between the variables along with reducing heteroskedasticity of the model.

Conclusion

Our linear regression model produced an Adjusted R^2 of 0.8256, which indicates that using the total number of bedroom and bathrooms, log of the house size in square-feet, log of the total area in square-feet, and the "City" variables was able to reasonably explain the log of housing prices in these cities. However, there appears to be a non-linear relationship between the independent and dependent variables, despite the log transformations. Our random forest model confirmed this observation, as it resulted in a higher R^2 than the regression model. As such, although our regression model appears to explain log of housing price well, a linear relationship between the variables should not be assumed.

Our multivariate time-series model was able to predict housing prices in each city for 2021 by using just historical income and housing price datasets. We compared the results to actual 2021 housing prices and noticed that our model was able to predict housing prices in Seattle, Nashville, and Cleveland relatively well. San Francisco's prediction had the largest difference, which suggests that income level is clearly not the only factor that can affect housing prices although the trends for both variables move in the same direction.

Overall, income level and housing prices often have similar trends, but our models from the regression analysis and time-series forecasting show that multiple variables work together to ultimately determine the price of a house. To further improve our analysis, we will consider other macro-economic factors in both our regression and time-series models. In addition, we can consider different types of houses and extending the model to include more cities. A

difference in difference analysis can also be conducted to see if housing price trends change after a certain event takes place.

Works cited

[i]. "Zillow Home Value Index (ZHVI) All Homes (SFR, Condo/Co-Op) Time-Series, Smoothed, Seasonally Adjusted." *Zillow Research*, 10 Oct. 2022, https://www.zillow.com/research/data/.

[ii]. "Per Capita Personal Income in Seattle-Tacoma-Bellevue, WA (MSA)." *FRED*, 16 Nov. 2021, https://fred.stlouisfed.org/series/SEAT653PCPI.

"Per Capita Personal Income in Austin-Round Rock, TX (MSA)." *FRED*, 16 Nov. 2021, https://fred.stlouisfed.org/series/AUST448PCPI.

"Per Capita Personal Income in Nashville-Davidson--Murfreesboro--Franklin, TN (MSA)." *FRED*, 16 Nov. 2021, https://fred.stlouisfed.org/series/NASH947PCPI.

"Per Capita Personal Income in Cleveland-Elyria, OH (MSA)." *FRED*, 16 Nov. 2021, https://fred.stlouisfed.org/series/CLEV439PCPI.

"Per Capita Personal Income in San Francisco-Oakland-Hayward, CA (MSA)." *FRED*, 16 Nov. 2021, https://fred.stlouisfed.org/series/SANF806PCPI.

[iii]. "Real Estate, Apartments, Mortgages & Home Values." *Zillow*, https://www.zillow.com/.

[iv]. Clever Real Estate, Eylul Tekin. "How Home Prices and Household Incomes Changed since 1960." *Clever Real Estate*, Clever Real Estate, 7 Aug. 2022, https://listwithclever.com/research/home-price-v-income-historical-study/.

[v]. Hermann, Alexander. "Price-to-Income Ratios Are Nearing Historic Highs." *Price-to-Income Ratios Are Nearing Historic Highs | Joint Center for Housing Studies*, 13 Sept. 2018, https://www.jchs.harvard.edu/blog/price-to-income-ratios-are-nearing-historic-highs.

[vi]. Lee, Don. "Is Austin, Texas, Becoming the next Silicon Valley?" *Governing*, Governing, 9 Feb. 2022, https://www.governing.com/now/is-austin-texas-becoming-the-next-silicon-valley.

[vii]. Trubetskoy, Gregory. "Gregory Trubetskoy." *Holt-Winters Forecasting for Dummies - Part III - Gregory Trubetskoy*, https://grisha.org/blog/2016/02/17/triple-exponential-smoothing-forecasting-part-iii/.

[viii]. Kenton, Will. "Endogenous Variable: Definition, Meaning, and Examples." *Investopedia*, Investopedia, 8 Nov. 2022, https://www.investopedia.com/terms/e/endogenous-variable.asp.

[ix]. "Forecasting: Principles and Practice (2nd Ed)." *11.2 Vector Autoregressions*, https://otexts.com/fpp2/VAR.html.

[x]. Christiano, Lawrence J. "Christopher A. Sims and Vector Autoregressions - New York University." *Christopher A. Sims and Vector Autoregressions*, https://pages.stern.nyu.edu/~dbackus/Identification/Christiano_on_Sims_SJE_12.pdf.