Team 26: Charlie Gu, Christian Robinson, Lindsey Waggoner

Shifting Priorities: The Impact of COVID-19 on US Housing Demand

Project Overview

The COVID-19 pandemic has led to significant changes in migration patterns in the United States, as remote work has given workers more freedom to choose where to live. As society became more isolated and entire industries came to a halt, changes in the crime rates and economic systems also occurred, and nearly every aspect of American life was affected in some way. This analysis aims to investigate the COVID-19 pandemic's impact on US migration patterns and housing demand and identify the key factors driving these changes.

Ultimately, the insights gained from this analysis can be valuable to government entities, real estate investors, and companies in making informed decisions about zoning laws, infrastructure development, real estate investments, marketing strategies, and operational planning. Being able to anticipate migration patterns is valuable to government entities to help create zoning laws, build necessary infrastructure, create housing supply, and attract prospective citizens. Additionally, understanding migration behavior helps companies build an efficient marketing and supply chain strategy and make informed decisions about where to invest resources and adapt to changing market conditions.

The analysis uses several time series data sets from a variety of governmental agencies and nonprofits. To simplify the data collection process, some of the data is collected on a state level instead of a city level. While the change in our original intention requires a modified focus, the value of the analysis remains unchanged.

Our plan for the analysis included first performing exploratory data analysis (EDA) to identify any relationships or trends between the independent and dependent variables. Then, we used principal component analysis and k-means clustering to identify and visualize groups of cities with similar characteristics according to our collected data. Finally, we trained a multiple linear regression model to identify the factors that have the strongest impact on housing demand. Prior to performing any analysis, we hypothesized that job potential, measured in average pay and total wages, would have the largest effect on housing demand, and that this factor may have seen the greatest change during the course of the pandemic due to changing industry patterns, work from home culture, and outside economic pressures.

Data Overview

Because our team wanted to analyze the effect of a variety of factors on housing demand, we collected data from multiple sources. While each of the data sources are government entities or non-profits and are the prime authority in their respective fields, the collection and reporting methods differed greatly between each of the sources. The labor data provided by the Bureau of Labor Statistics could be filtered by industry, county, or state, and reported wages and employment data for each combination of these filters. With hundreds of possible industries, we chose to view only the total of all industries. Additionally, county data was sometimes unavailable for all time periods, and filtering by county offered a unique challenge for metropolitan areas that are located across multiple counties. For this reason, we chose to collect data at the state level. The Federal Bureau of Investigation data was also available at the state level, which eliminated the need to filter through other reporting organizations (local police departments, state agencies, etc.) whose overlapping jurisdictions could preclude an accurate count of violent crimes in a single metropolitan area. Likewise, using only state-level data for the National Oceanic and Atmospheric Administration weather statistics prevented conflicts between multiple weather stations within the same metropolitan areas. Though our analysis may lose some granularity because of this decision, the complexity of the data reporting standards necessitated that some simplifying measures be taken prior to analysis.

From these three datasets, we collected five main independent variables: violent crime rate, minimum annual temperature, maximum annual temperature, total wages, and average annual pay. The violent crime rate is reported per 100,000 people, and is thus controlled for differences in the population of different states. Minimum and maximum annual temperature collectively represent how temperate a particular state is. Total wages, or the total amount paid to all employees in a state, was chosen because it quantifies both wages and the number of employed people, and is supplemented by the average annual pay, which is reported on a per-person basis. Together, these factors quantify the safety, weather, and job opportunities in a state.

To measure housing demand, we collected data on housing prices from the American Enterprise Institute, which reports housing prices for each of the top 100 metropolitan areas in the United States. This data was reported quarterly and divided into housing types (entry-level homes, step-up homes, and overall). To account for all types of home purchases, we chose to analyze the overall data, and used the average value over all four quarters so that the time series would align with the annual data of the other datasets. This data became our main dependent variable.

Because of the sheer amount of data available, we chose to also limit our analysis to a chosen set of states and years. To keep a balance between the pre-COVID and post-COVID data, we filtered most of the data to only the years 2018-2021. At the time of this report, the annual data from 2022 has not yet been reported, so 2021 was the most recent year available. We also chose our states based on the top 100 largest metropolitan areas, thereafter filtering out any city for which we did not have corresponding data, and selected every third city in the filtered list to narrow down our chosen data while ensuring that we were sampling evenly across the city sizes. In total, we had data from 21 states across 4 years, or 84 data points per factor. The one exception to this was the crime data, which had not yet been reported for 2021. As such, the crime data was collected from 2012 until 2020 and filtered as needed for the analysis.

Modeling Overview

To establish a baseline knowledge of how each of the independent variables has changed over the time interval of interest, we performed a simple treatment effect analysis on each variable. First, we generated a scatterplot of each factor, to gain a qualitative impression of the changes in each variable.

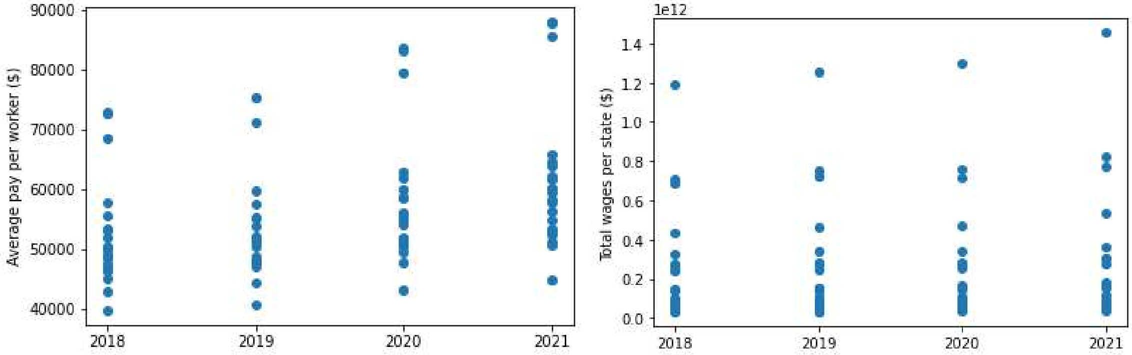

Figure 1: Labor statistics over time

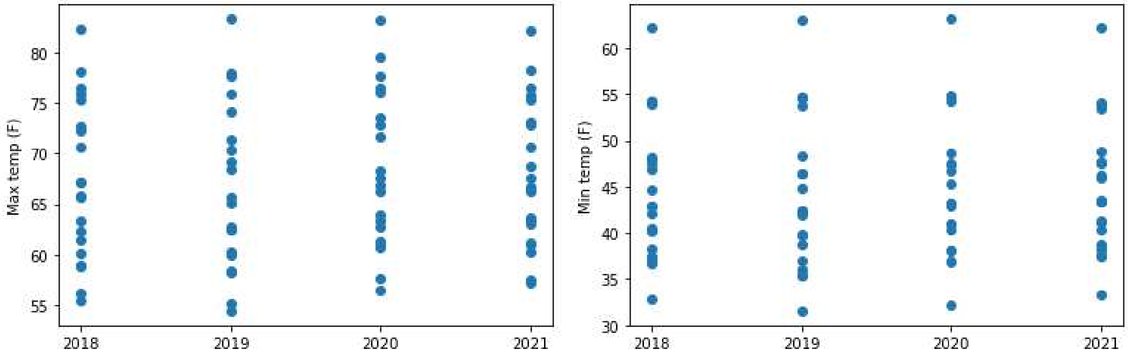

Figure 2: Weather statistics over time

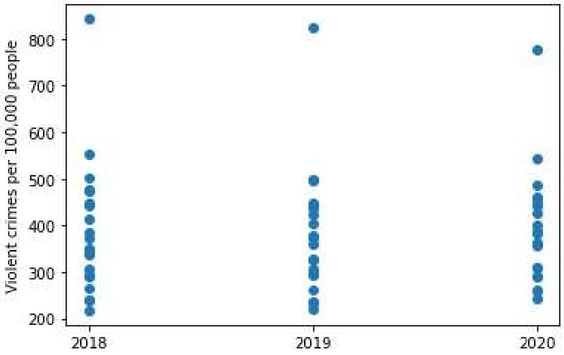

Figure 3: Violent crime statistics over time

Visually, it does not appear that there is a clear trend in these variables during the period of interest, with the exception of average pay per worker, which appears to increase consistently from year to year. To confirm this, we performed a linear regression with the value of each factor as the independent variable and a treatment factor as the dependent variable. This treatment factor was a binary variable with a value of 0 if the datapoint occurred in 2018 or 2019, and a value of 1 if the datapoint occurred in 2020 or 2021. This regression was performed for each variable, and the p-value of the treatment coefficient was used to determine whether the treatment effect was significant.

Variable	Intercept	Coefficient	P-Value
Average pay	53,344.27	6,746.10	0.0056
Total wages	2.632e11	2.456e10	0.7312
Minimum temperature	44.24	0.642	0.6999
Maximum temperature	66.98	1.200	0.4861
Violent crime rate	384.425	10.375	0.7767

Table 1: Treatment effect summary

As can be seen in Table 1, the only significant treatment effect was that of average pay, with a 99% significance level. Because average pay experienced an increase during these years but total wages did not, we can infer that those who remained employed saw an increase in wages but the overall number of employed people likely decreased, keeping the state's total wages steady. This is consistent with conventional knowledge about the effects of the pandemic. Overall, this initial analysis showed that there was not a significant change in most of our independent variables during the period of interest, and so it is unlikely that any changes in housing demand are caused solely by changes in these variables. However, shifts in priorities for home purchasers may cause the effect of these variables on housing demand to differ pre- and post-pandemic, which will be examined later in our analysis.

Next, we wanted to visualize the relationships within our collected data and group similar cities together. To do this, we used principal component analysis, balancing the number of components with the explained variances. From visual inspection, four components is the ideal number of components, with all further components only adding negligible value to the model.

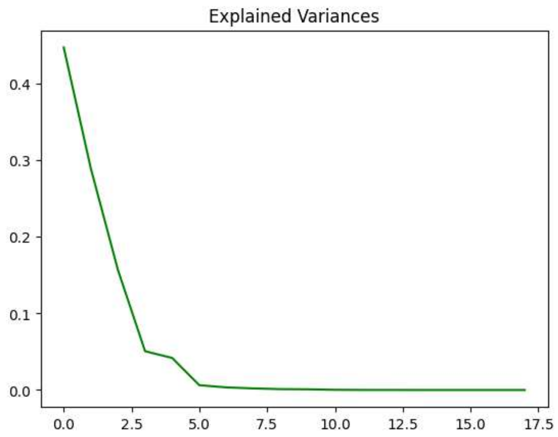

Figure 3: Total explained variance for principal components

Importantly, since principal component analysis is a variance maximizing technique, all the numerical data needs to be standardized. We used a standard scaler from sklearn to accomplish this.

Even though we determined four components is the ideal number, a 4-dimensional graph is still difficult to interpret. For accessibility, we also projected each city to its first two principal components.

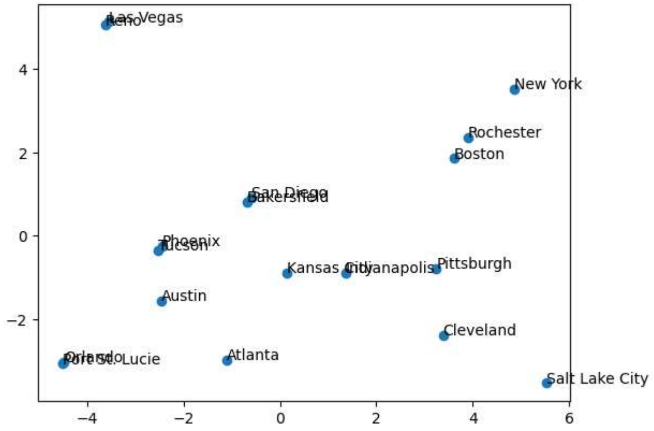

Figure 4: Cities projected to the first two principal components

A few things immediately stand out - cities that lie within the same state are projected extremely similarly to each other. This makes sense - a large portion of our data was filtered by state and not city, since we were unable to find more granular data. Additionally, geographic neighbors are projected nearby to each other on the graph, indicating location plays an important role.

The next step after principal component analysis is to use the principal components in a k-means algorithm to systemically group cities together, minimizing within-cluster sum of squared error. Again, we need to tune hyperparameters and find the ideal number of clusters to use. The sharpest bend in the total sum of squares curve is found at five clusters, so we choose to proceed with five clusters.

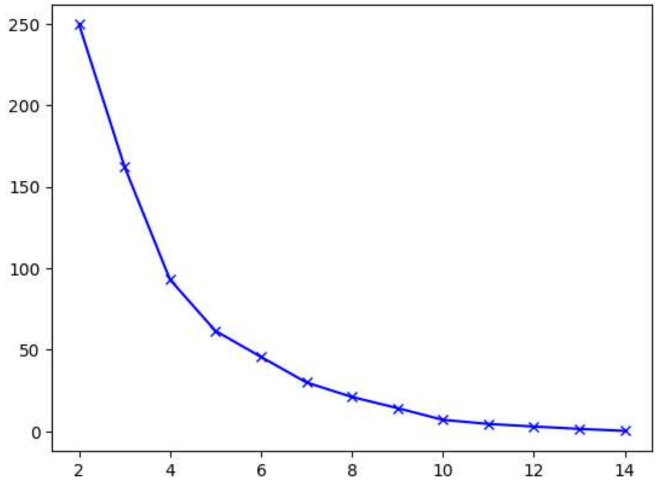

Figure 5: Total sum of squares vs number of clusters

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Indianapolis	Austin	Las Vegas	Phoenix	New York
Cleveland	Atlanta	Reno	San Diego	Boston
Pittsburgh	Orlando		Tucson	Rochester
Port St. Lucie			Bakersfield	
Salt Lake City				
Kansas City				

Table 2: Cluster assignment for each city

This clustering reveals which cities in our dataset are most similar to each other, based only on the independent variables. Though the number of cities in each cluster varies, it is interesting to note that some of the most expensive cities in the United States (namely New York, Orlando, Las Vegas, and Bakersfield) fall in different clusters, demonstrating that the factors that drive the clusters' similarities may not be what drives an increase in housing prices. This clustering also shows that our downsampling of cities maintains a breadth of characteristics, and that we did not flatten variation in our selection.

Last, we used multiple linear regression on the housing data to understand the relationship between our predictors and the response variable, and how those effects may have changed after the COVID pandemic. To begin, we evaluated each predictor, which revealed some missing data. Specifically, the violent crime data was missing all data points for 2021 and 2022 and the weather data was missing data for 2012-2017 and 2022. Therefore, before conducting analysis, we had to decide how to handle the missing data points. We chose to impute the missing data rather than remove them because the goal of the multiple linear regression model is to compare the effects before and after COVID that goal would be impossible to achieve if we threw out the post-COVID data.

Regarding the crime data, we began by looking at the crime rate for each year from 2012-2020 by city to see if there were significant differences between cities and years. There are immediately observable trends for some cities; for example, Reno had a very high violent crime rate in the mid-2010s; however, it had been trending downward in recent years. On the contrary, the violent crime rate in Atlanta, Kansas City, and Pittsburgh have been trending upward in recent years. To impute the missing data with values that held true to recent trends, we made the decision to impute the missing data from 2021-2022 with the mean violent crime rate from each city from 2018-2020.

Because the weather data included a much shorter time horizon, we did not have the luxury of ascertaining temperature trends over a long timeline. Evaluating the chart showing the minimum and maximum temperatures by city from 2018-2021, it was clear that each city had a relatively consistent average minimum and maximum temperature, so we imputed the missing data with the mean temperature for each city from 2018-2021.

Once the data was cleaned, we evaluated the linearity of each predictor on the response using a scatter plot. Here, each plot displayed a clear lack of linearity between the predictors and the response. Regarding violent crime rate, cities with both a low and a high violent crime rate tended to have lower housing prices on average, where cities with violent crime rates near the center tended to have a much larger variability in housing prices and included both the highest priced and the lowest priced housing market. For weather, cities that had higher average minimum and maximum temperatures tended to have lower-priced housing; however, the rest of the cities varied greatly, and there was no clear relationship between average temperature and housing prices. Lastly, for employment data, there was a general upward trend between average pay and housing price, though there was still a lot of variability.

After evaluating the linearity of each predictor, we added a categorical variable “PostCovid” which was assigned a 1 if the data point occurred on or after 2020, and a 0 if the data point occurred before 2020. Before performing regression, we compared the mean of each variable pre-COVID and post-COVID, which can be found in Table 3 below.

	Housing Price	Violent Crime Rate	Minimum Temperature	Maximum Temperature	Total Wage	Average Pay
Pre-COVID	$245,889	413.64	45.59	68.44	$431.4 million	$60,845
Post-Covid	$365,207	401.10	45.95	69.10	$451.5 million	$64,755

Table 3: Mean values of independent variables pre- and post-COVID

Then, we produced an initial multiple linear regression model on the full dataset, including the new categorical variable. The goal with the initial model was to create a baseline to compare subsequent models to, and to see if any additional insights could be gleaned. As we suspected from the initial scatterplots, the full linear model did not prove to explain the variability in housing price well, with an R-squared of 0.376 and an Adjusted R-squared of 0.371. A scatterplot of predicted housing prices from the initial model versus actual housing prices reveals a model that is able to predict prices better for lower-priced markets than higher-priced markets.

Because the scale varied drastically for each variable, it was necessary to normalize the data and produce another multiple linear regression model on the normalized data set. Though explainability was not improved with the new model, we were able to obtain valuable effect sizes. The coefficient for each predictor on the new regression model is outlined in Table 4 below, which revealed average pay to have the greatest effect size on housing price in our sample.

Predictor Variable	**Coefficient**
Average Pay	0.307
Maximum Temperature	0.271
Minimum Temperature	-0.243
PostCovidTRUE	0.142
Violent Crime Rate	0.050

Total Wage	0.009

Table 4: Effect size of predictor variables

Another way to compare housing prices in each city before and after the pandemic is to obtain a regression model on the pre-COVID data, then use it to predict housing prices using the post-COVID data. Comparing the predicted housing prices to actual housing prices post-pandemic, we found that the model was able to predict housing prices for cities with lower housing prices, though higher priced cities had much larger error. In fact, the pre-COVID regression model did not predict any housing prices to be above 0.4 on average (normalized), when there were actually many cities well above this threshold. The scatterplot below highlights the inability of the pre-COVID model to accurately predict post-COVID housing prices.

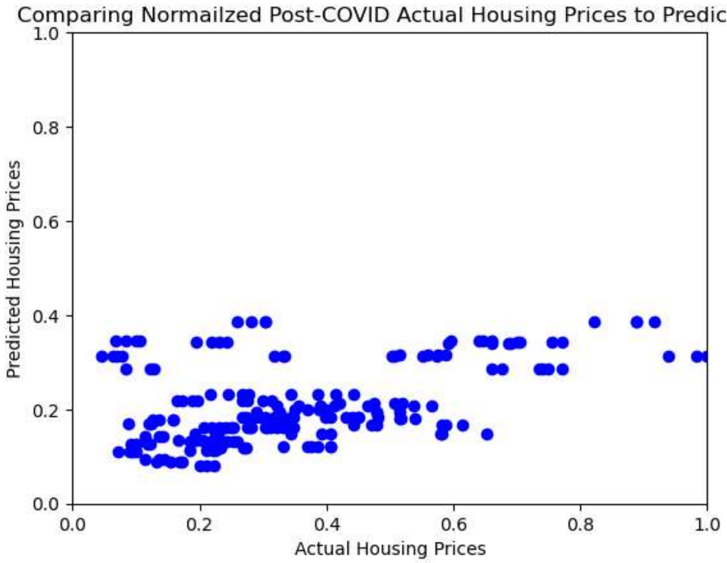

Figure 6: Predicted housing prices vs actual housing prices

Lastly, we created interaction terms with the PostCovid categorical variable with each of the predictors to measure the difference in effect size pre-COVID and post-COVID. The inclusion of interaction terms revealed a notable difference in the effect size for the average maximum temperature and average minimum temperature variables, though each had less of an effect post-COVID. A full breakdown of the coefficient of each predictor and the corresponding interaction term can be found in Table 5.

Predictor	Coefficient	Interaction Coefficient (PostCovid)
Violent Crime Rate	39.62	120.49
Minimum Temperature	-2,905.08	-1,738.46
Maximum Temperature	-1,511.57	257.18
Average Pay	4.49	4.78
Total Wage	9.66e-8	1.16e-7

Table 5: Effect size of interaction terms

Conclusions

From these models, we concluded that average pay had the greatest effect on housing prices in an area. With average pay being the only independent variable that underwent a significant change during the pandemic, we can assume that much of the increase in housing prices after 2020 was correlated with an overall increase in wages as well. However, the factors whose effect changed the most pre- and post-COVID were the factors related to climate. From these interaction coefficients, we can see that temperate climates (higher minimum and lower maximum temperatures) were correlated with higher housing prices post-pandemic, which supports our original theory that home buyers may have favored locations with better weather during and after the pandemic. Total wages had very little effect on housing prices both pre- and post-pandemic, while high average pay was correlated with high housing prices during the entirety of the time interval examined. Interestingly, high violent crime rates became increasingly correlated with high housing prices post-pandemic, which may imply the presence of an additional confounding factor not accounted for in this analysis. In general, our analysis shows that models of housing prices pre-COVID are not sufficient to predict post-COVID prices, and that we must develop new models and understandings of home buyers' priorities to understand the "new normal" of the housing market.