Name: Poornima Muthukumar
Date: Nov 10th, 2021

Motivation:

In this project, I plan to analyze the impact of covid on the housing market. One specific area that specifically experienced the brunt of this unprecedented crisis is the housing market. As a home buyer during the pandemic, my husband and I were going through the process of searching for a home. But the housing market at the start of 2021 gave us immense anxiety and worry as the demand for the houses was very high, prices of the houses kept escalating at a very fast rate, the inventory of houses kept declining and the bidding war to top it off made things worse. We spent countless hours every day trying to read the news, watching youtube videos, and talking to other buyers to understand the housing market to help with the confusion.

A lot of the home-buyers across the country are frustrated with this increasing trend in housing prices. This topic is particularly human-centered as it affects a vast majority of Americans who are looking to purchase homes. If the prices keep increasing at this rate, a lot of the people from lower-income brackets will suffer as they will not be able to afford to house, or will be pushed to outskirts or suburbs thus causing a widening of the wealth gap between the rich and the poor. Because homeownership is an important tool for building long-term wealth and children of homeowners are likely to become homeowners, this trend can further exacerbate wealth inequality for future generations.

By performing this analysis I hope to learn if there was any correlation between the number of covid cases and housing prices and hope to see how the market fluctuated during the pandemic. I also want to see if the housing prices have gone down recently or if they are still on the increasing trend. This analysis is of particular scientific interest as it will help others see trends between housing prices and a fluctuating economy during a pandemic.

Research Question / Hypothesis:

As the pandemic progressed, the cases started increasing nationwide. The state and national government issued several lockdowns to control the spread of the virus, a lot of the businesses were shut and there was widespread fear of another recession. Tracking changes in the housing market will provide us insights into the economy of North Carolina state, help us contextualize growth and decline in this county, and give us insight into the market. As part of this project, I plan to explore the following questions to help understand the trend in the housing market since the start of the pandemic.

1. During the COVID-19 pandemic did the housing prices go up or down from January 2020 through August 2021?
2. Did the number of COVID-19 cases have an impact on the housing prices from January 2020 through August 2021?
3. Where there any other trends in the data related to covid cases and the housing market?

The entire analysis will be focused on the state of North Carolina and County Mecklenburg.

Data Used:

To perform the analysis I will use the following different datasets.

1. The RAW_us_confirmed_cases.csv file from the Kaggle repository of John Hopkins University COVID-19 data.
2. The weekly housing market data from Redfin - https://redfin-public-data.s3-us-west-2.amazonaws.com/redfin_covid19/weekly_housing_market_data_most_recent.tsv

The Redfin weekly housing market data has data for each county on a weekly basis. The data is broken down by property type (All Residential, Single Family, Condo, Multi-Family, Townhouse, etc). Redfin has published this page to define each column in the dataset and how to interpret the column. This data set is licensed under Redfin's Terms of Use. The guidelines for using the data states to cite the data source appropriately and provide a link to Redfin.

The following columns from the Redfin data set are of interest that I will use in my analysis.

COLUMN NAME	**DESCRIPTION**
region	County Name
property_type	Type of the property (All Residential, Single Family, Condo, Multi-Family, Townhouse etc)
median_sale_price	Median sale price of homes in that county
median_list_price	Median listed prices of homes in that county
inventory	Number of homes in the inventory for that month.
total_homes_sold	Number of homes sold in that month
start_period	30 day start time
end_period	30 day end time
total_new_listing	Total number of new listings.

Joining the housing market data with the RAW_us_confirmed_cases.csv dataset will help in my analysis of how housing prices changed during the pandemic from January 2020 till August 2021.

I feel there are no strong ethical considerations of using the housing market dataset as housing price is publicly available information and does not intrude on the privacy of any individual. However, there is a possibility that this data set released by Redfin is not accurate and could lead to misleading results.

Unknowns and Dependencies:

Although plotting covid cases and housing prices can help us find correlations between the two, there could be other factors that might have contributed to the fluctuation in the housing market such as low-interest rates, remote work during the pandemic, people purchasing houses because of an increase in the rental price, etc. Due to constraints of time, I will primarily focus my analysis to see patterns and trends between housing price and covid cases and deaths, leaving other factors out.

Methodology:

Step1: Data Processing and Cleaning

In this step, I will merge the two data set RAW_us_confirmed_cases.csv and weekly_housing_market_data_most_recent.tsv based on date and county. After the data set is merged I plan to perform data cleaning to remove any NA values.

Step2: Correlation between covid cases and housing prices

Here I will perform exploratory data analysis by creating visualizations to see the correlation between housing market data and covid cases and deaths week over week.

Here I will plot the following visualization from the redfin data set.

1. Weekly confirmed covid cases and the number of new listings.
2. Weekly confirmed covid cases and the total number of homes sold.
3. Weekly confirmed covid cases and median list price.
4. Weekly confirmed covid cases and median sale price.

Step3: Linear Regression

I will also perform linear regression to predict housing prices for 2020 and 2021 and compare it with actual housing prices to see if there is a difference between predicted and actual housing prices for 2020 and 2021.

Linear regression suits best to find the relationship between a dependent continuous variable (Median Sale Price) and one or more explanatory independent variables (Month/Year). Linear regression suits best here because we can see a linear trend in the dataset for housing prices and housing prices are normally distributed.

I will fit a univariate linear regression model using historical data from (2010- 2019) where the feature is the weekly dates and the target is the median housing price. We are all aware of the housing market crash in 2007-2008 and it took the market some time to stabilize after the crash, hence I have decided to train the model with data post-2010. I will split the data into (80-20 train test split) and fit a linear regression model using python scikit learn. I will also compute the RMSE (root mean square error) as a measure of model performance and use the model to predict housing prices for 2020 and 2021 and compare if the prediction is higher or similar to actual prices.

Timeline to completion:

To approach the above problem I envision it would take me the following tasks to complete -

1. 2-3 days for downloading the datasets, cleaning, and merging the data sets. The housing data is already aggregated on a weekly basis. I will also aggregate the covid cases data to produce weekly aggregates and join the two datasets by county and the month. (20th November 2021)
2. I plan to spend another 2-3 days to produce initial exploratory data analysis visualizations to see the correlation between housing prices and covid cases and deaths. (25th November 2021)
3. I plan to spend 1 week doing the linear regression analysis model to predict housing prices. (1st December 2021)
4. I plan to spend another 2-3 days documenting the entire steps on Github for easy reproducibility along with all the different steps. (3rd December 2021)
5. I plan to spend 2-3 days doing the project presentation highlighting my findings. (7th December 2021)
6. I plan to spend 1 week finishing the report and organizing everything in GitHub for submission. (14th Dec 2021)