IMPACT OF COVID – 19 CASES ON HOUSING PRICES AND PREDICTION

Credit: Getty Images/iStockphoto

When the Covid –19 pandemic hit the USA, it took a toll not only on people's life, but also on the economy. The International Monetary Fund estimated the median global GDP dropped by 3.9% from 2019 to 2020, making it the next bad economic downturn after the great depression. Tens of millions of people lost their job. Employment began to rebound within a few months, but unemployment remained high throughout 2020. The great depression between 2007 – 2009 observed plummeting of the U.S housing bubble. The reason for this study is to see if there was any impact of COVID- 19 cases on the housing market. Did the housing prices go up or down during COVID? Does/ did the number of covid cases affect the prices? Also predicting housing prices helps people to plan to buy a house so they can know the price range in the future and plan their finances well. Housing price prediction also allows the investors to learn about the trend of the prices in a certain location.

1) DATA SOURCE

The dataset considered were:

- COVID -19 data from John Hopkins University : This is a daily updating version of COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). The data updates every day at 6am UTC, which updates just after the raw JHU data typically updates. The data set can be downloaded as .csv file

from KAGGLE website following the link given above or at my GitHub repository following the link: COVID-19 data . I have used **RAW_us_confirmed_cases** dataset for the number of covid- 19 cases across the USA.

- The housing price data from Zillow. I have used **Zillow Home Value Index (ZHVI)** file for metro and U.S geography. ZHVI is A smoothed, seasonally adjusted measure of the typical home value and market changes across a given region and housing type. It reflects the typical value for homes in the 35th to 65th percentile range. The raw version of that mid-tier ZHVI time series is also available. This dataset file can be downloaded from: Metro zhvi.
- Since COVID data is aggregated by Counties and the Housing data is aggregated by Metro area, city and county mapping data is required. This is collected from https://simplemaps.com/data/us-cities .
- I have also included interest rates data from https://fred.stlouisfed.org/series/MORTGAGE30US to see its impact on housing prices along with covid data.

All the raw data can be downloaded from my GitHub repository .

2) DATA WRANGLING

- The housing dataset needed to include the following features: Metro/ county region name, State name, Date (monthly – last date of the month) and house price in $. Since there were housing price data for all states of the USA, I wanted to limit the scope of the analysis to the major housing market states. I filtered the data for states including **Texas (TX), Washington (WA), California (CA), Colorado (CO) and Virginia (VA).** The region name had the state id attached to it at the end. Hence this was cropped from the region name. The date feature in the raw dataset file was in a wide format, having individual dates as the feature. So, I converted it to a long format, to include date as a feature. All the features were converted to the appropriate format and NaN values dropped.

- The covid dataset contained every day data cumulated. Hence, the month end date was only considered, since we have monthly housing data. Again, the states were filtered to include only the major 5 states of the housing market. State, county, number of covid case (monthly cumulative) and dates features were only kept and all other columns were dropped. The data types were converted accordingly.

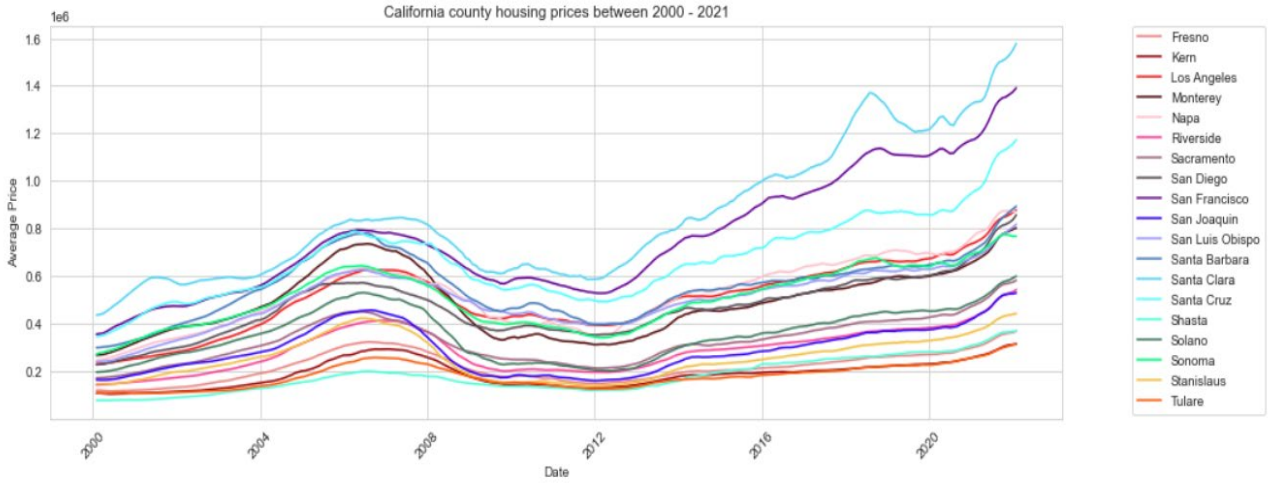
- The city-county mapping dataset and mortgage dataset was taken again for only the five states. Unwanted columns were dropped. The mortgage rates were not monthly, but annual rates across different state. Hence, this data was copied for all the months in a

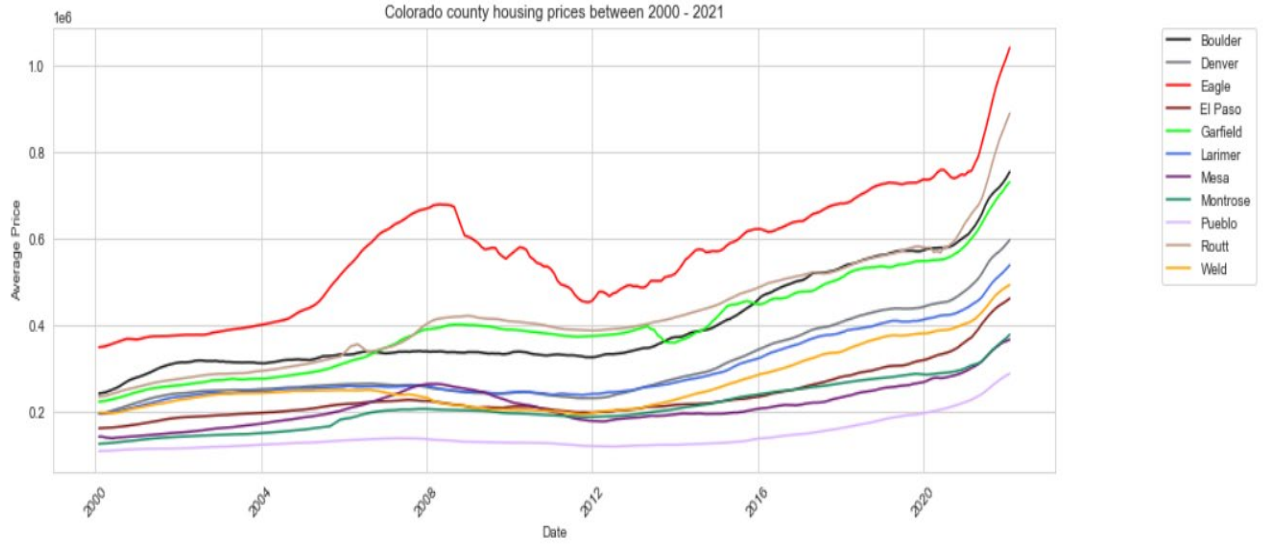
particular year. This is to match with the processed housing price and covid cases dataset. I merged all the datasets together to include the covid case number, mortgage rate and housing price data for each month from 2019 to mid-2021 (June month data). I set aside a validation set, which includes the data from July 2021 to 2022 data, to check my model prediction.

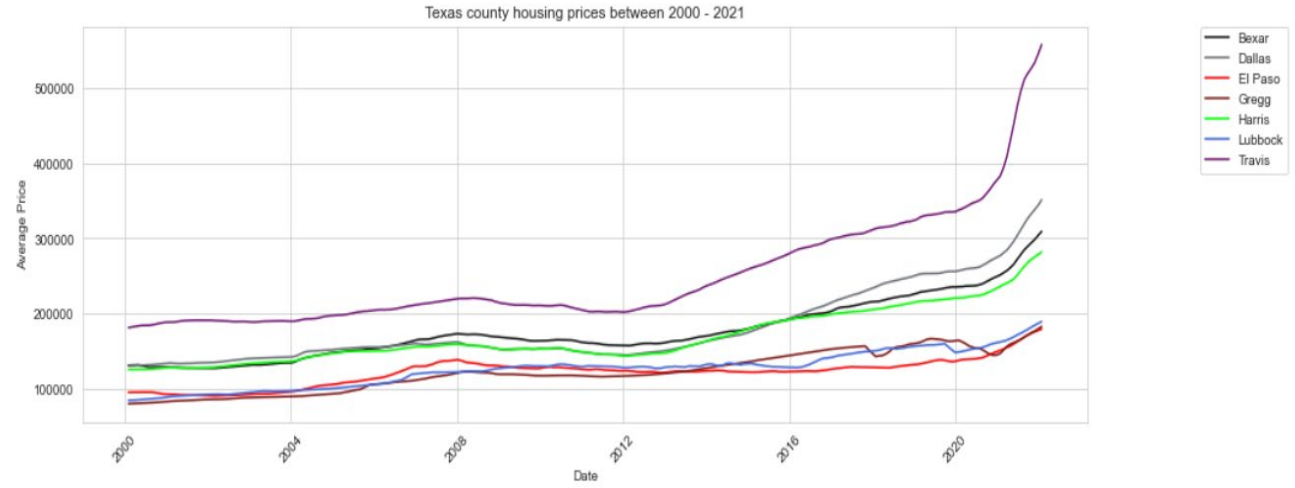
- The processed files are saved in the processed data folder of my GitHub repository.

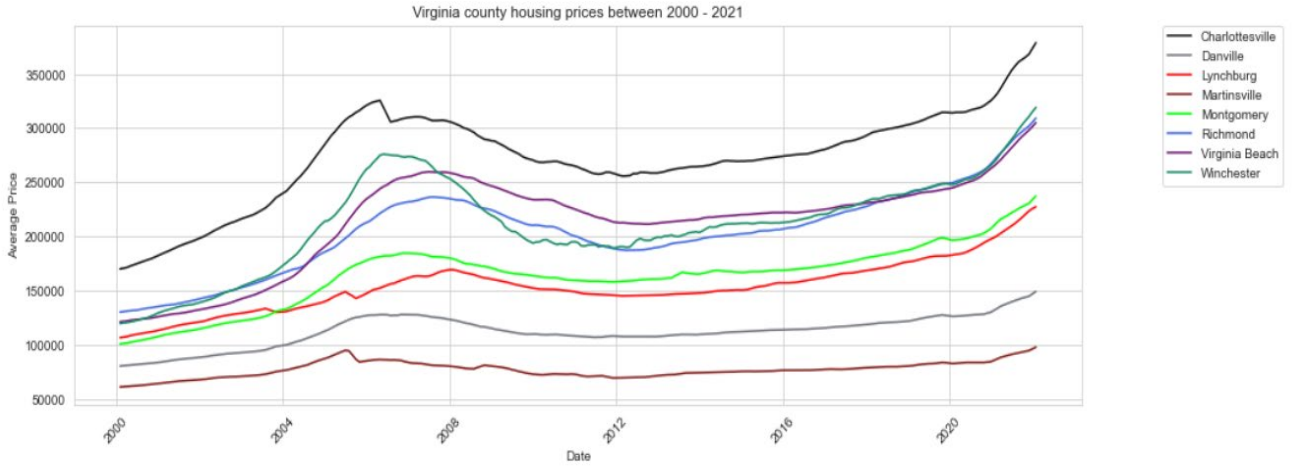
3) EXPLORATORY DATA ANALYSIS

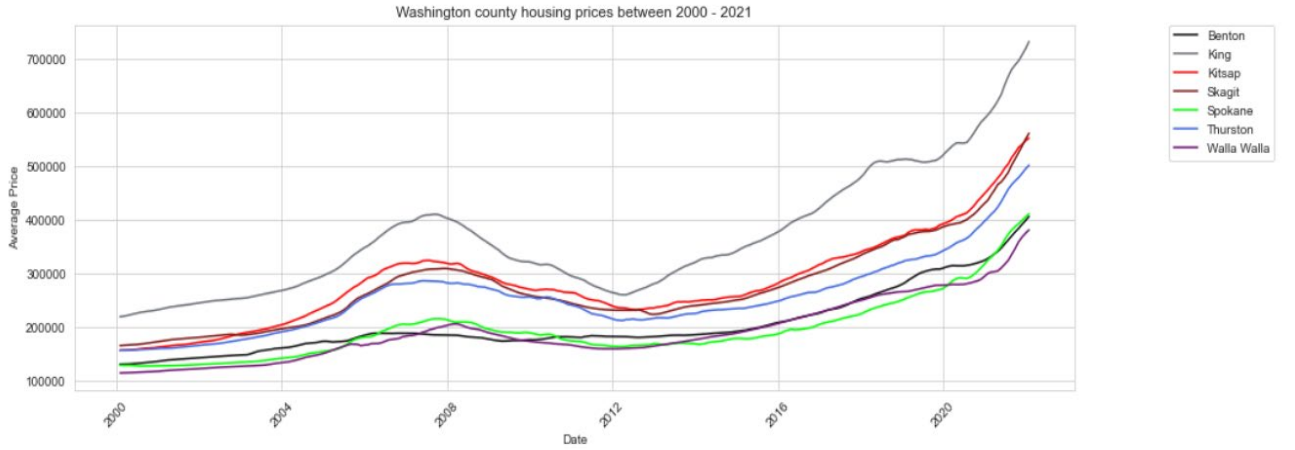
- **STATE- COUNTY HOSING PRICES FROM YEAR 2000 TO 2021**





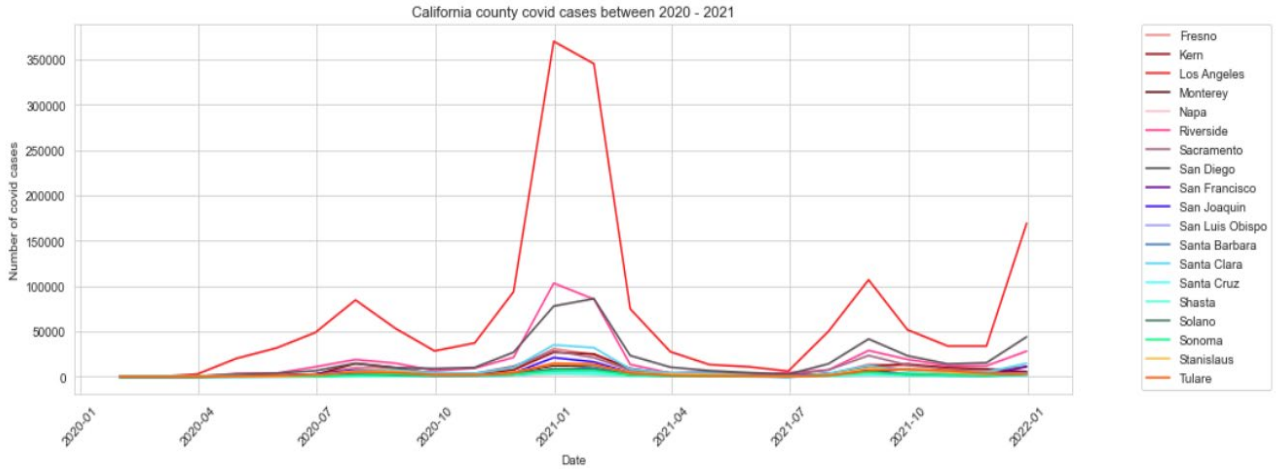


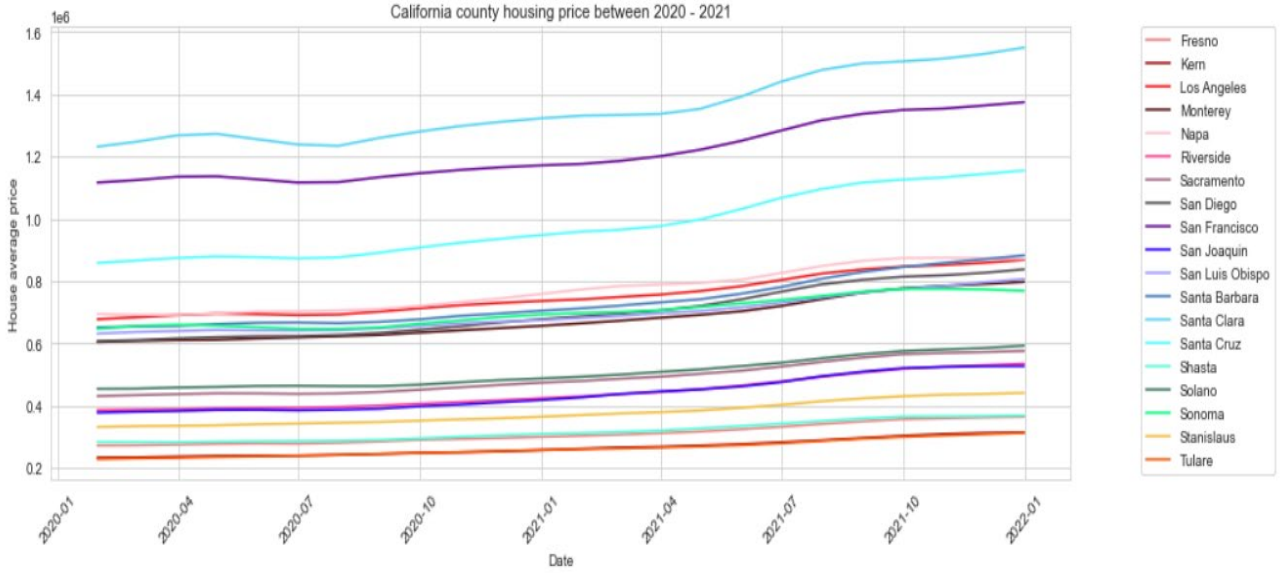


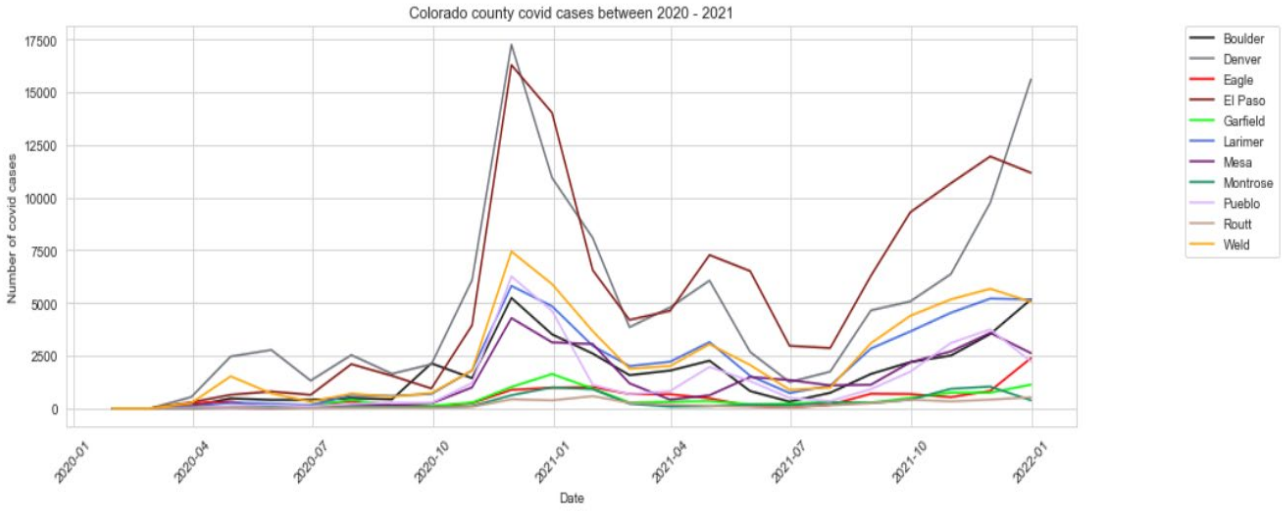


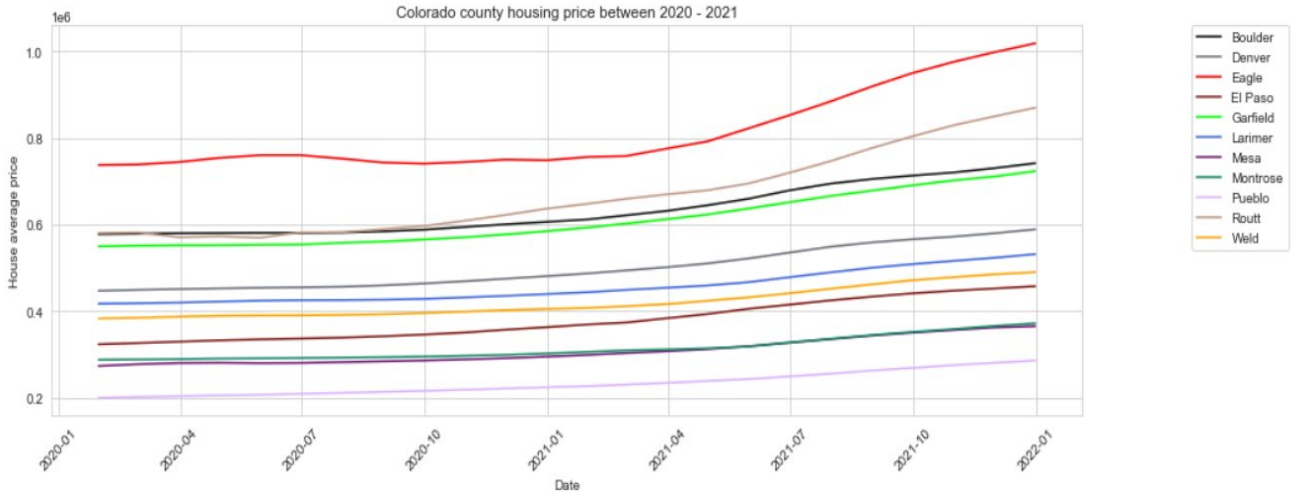
The housing prices between 2000 to 2021 for most of the states, shows a dip around late 2008 to early 2012. This is because of the Great Recession that began in December 2007 and ended in June 2009. After the Great Recession, Housing market was affected across the USA. Also, there is an increase in the prices after late 2020 during COVID-19 pandemic. This maybe because commuting reduced and people wanted to buy houses on the outskirts, instead of paying for house rents in the metro areas.

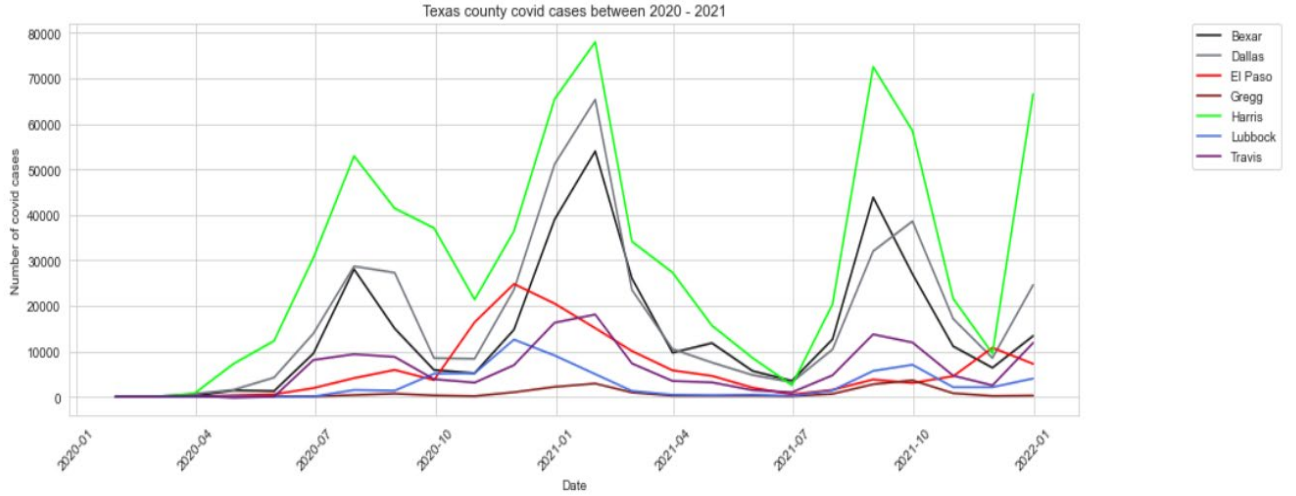
- **STATE-COUNTY COVID CASES FOR 2000 AND 2021 COMPARED TO STATE -COUNTY HOUSING PRICES DURING COVID-19 PANDEMIC**



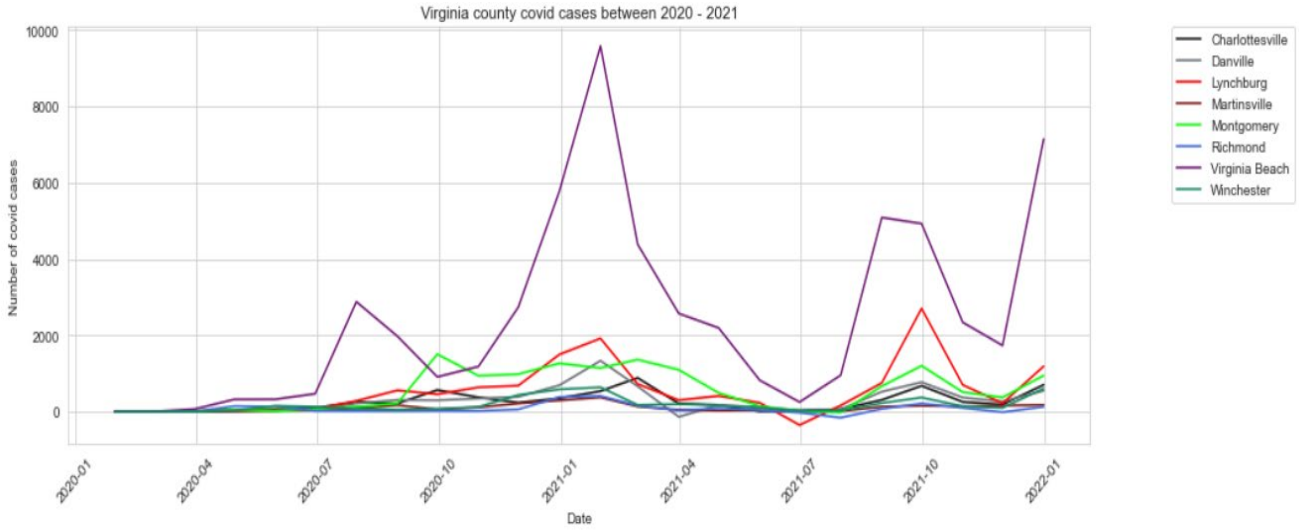


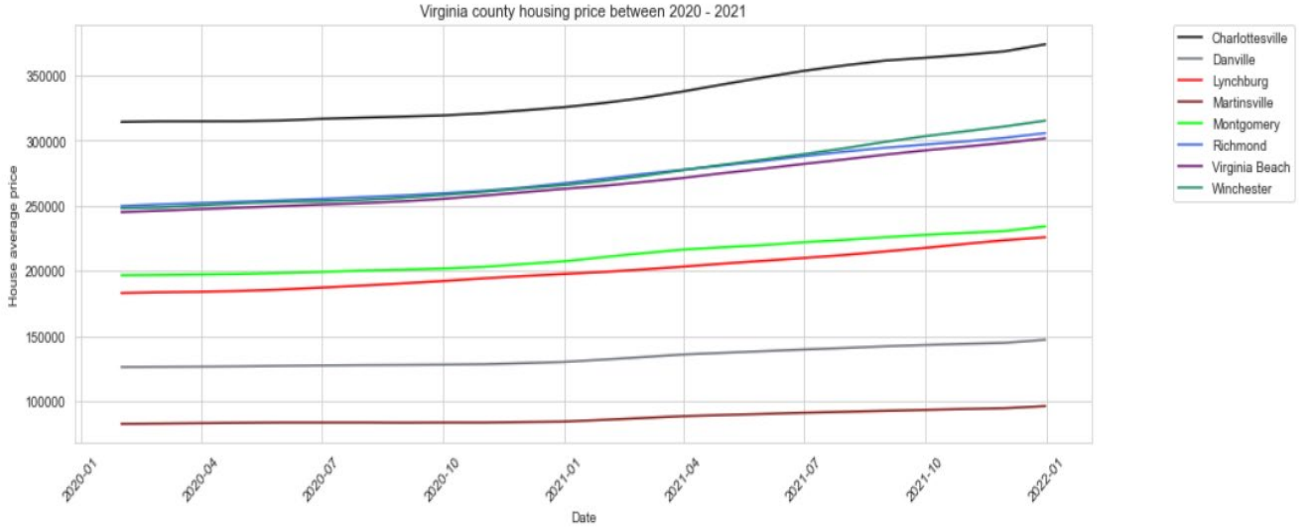


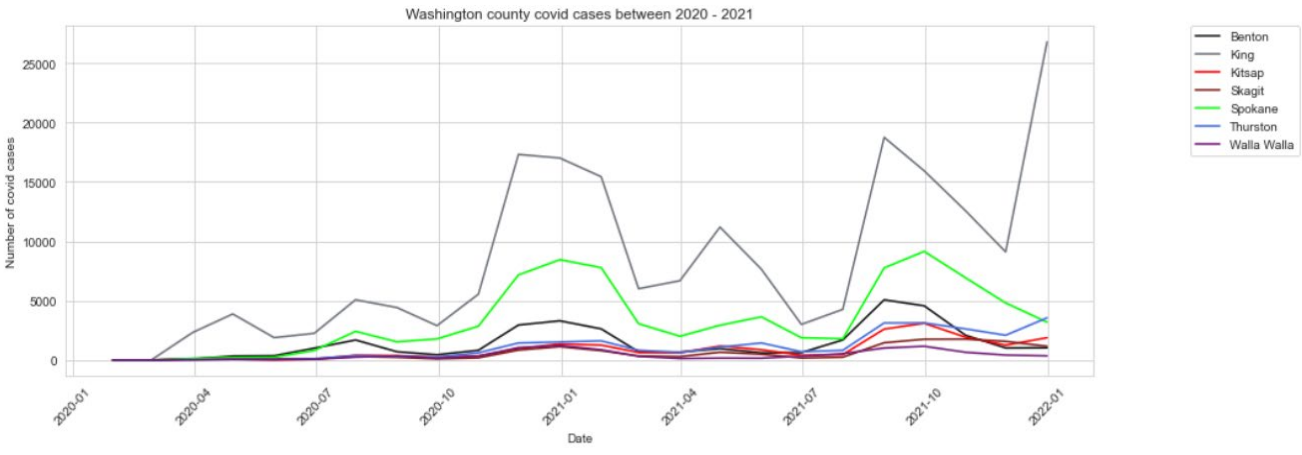

Colorado county housing price between 2020 - 2021
1e6
1.0
0.8
0.6
0.4
0.2
House average price
2020-01
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
Date
Boulder
Denver
Eagle
El Paso
Garfield
Larimer
Mesa
Montrose
Pueblo
Routt
Weld

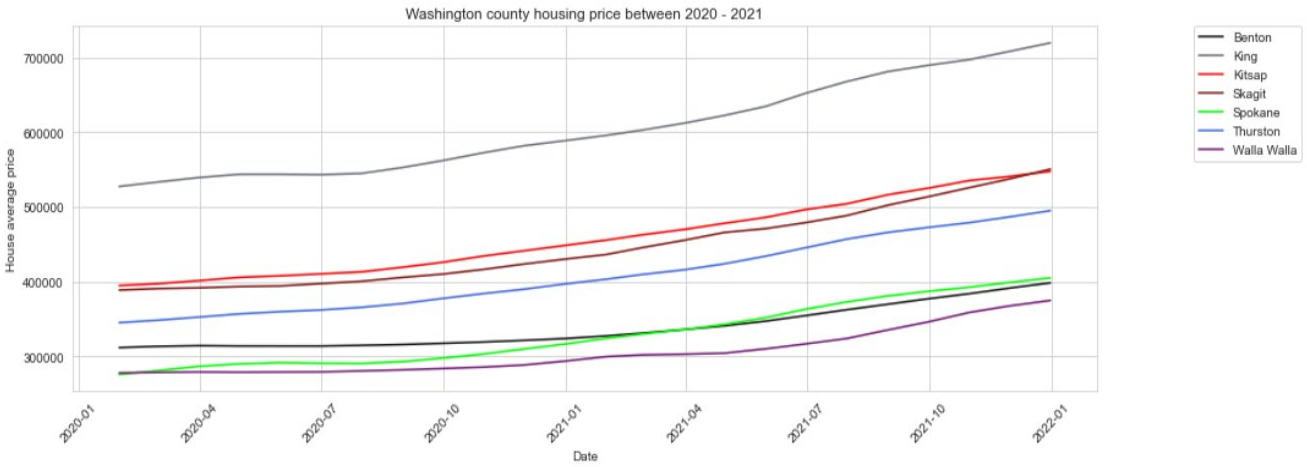

Texas county covid cases between 2020 - 2021
80000
70000
60000
50000
40000
30000
20000
10000
0
Number of covid cases
2020-01
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
Date
Bexar
Dallas
El Paso
Gregg
Harris
Lubbock
Travis


Texas county housing price between 2020 - 2021
550000
500000
450000
400000
350000
300000
250000
200000
150000
House average price
2020-01
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
Date
Bexar
Dallas
El Paso
Gregg
Harris
Lubbock
Travis


Virginia county covid cases between 2020 - 2021
Number of covid cases
Date
Charlottesville
Danville
Lynchburg
Martinsville
Montgomery
Richmond
Virginia Beach
Winchester

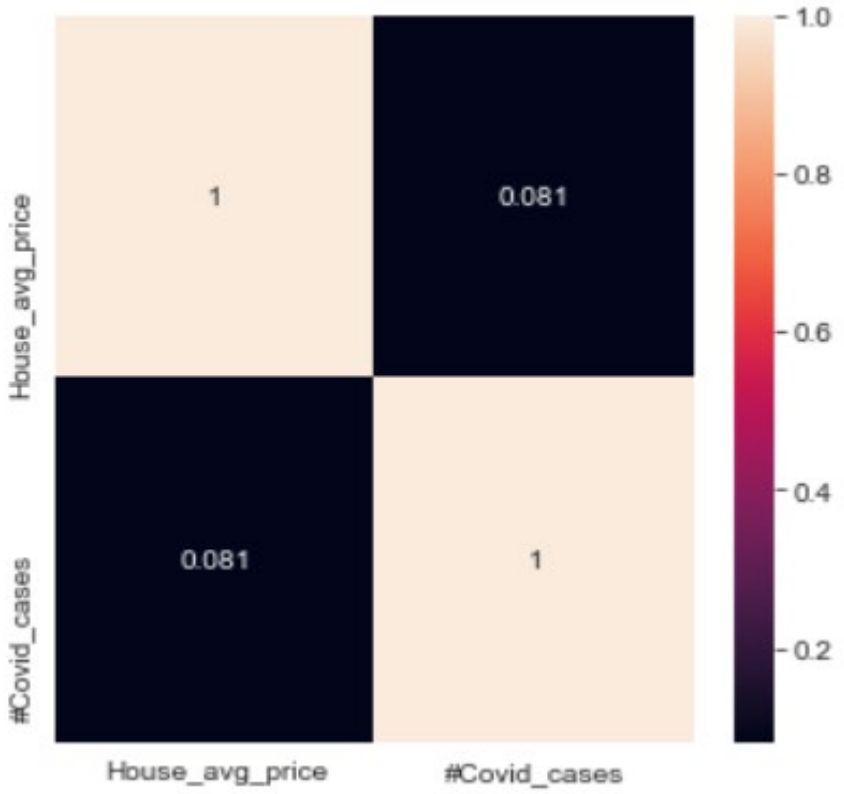

Virginia county housing price between 2020 - 2021
House average price
Date
Charlottesville
Danville
Lynchburg
Martinsville
Montgomery
Richmond
Virginia Beach
Winchester


Washington county covid cases between 2020 - 2021
Number of covid cases
Date
Benton
King
Kitsap
Skagit
Spokane
Thurston
Walla Walla



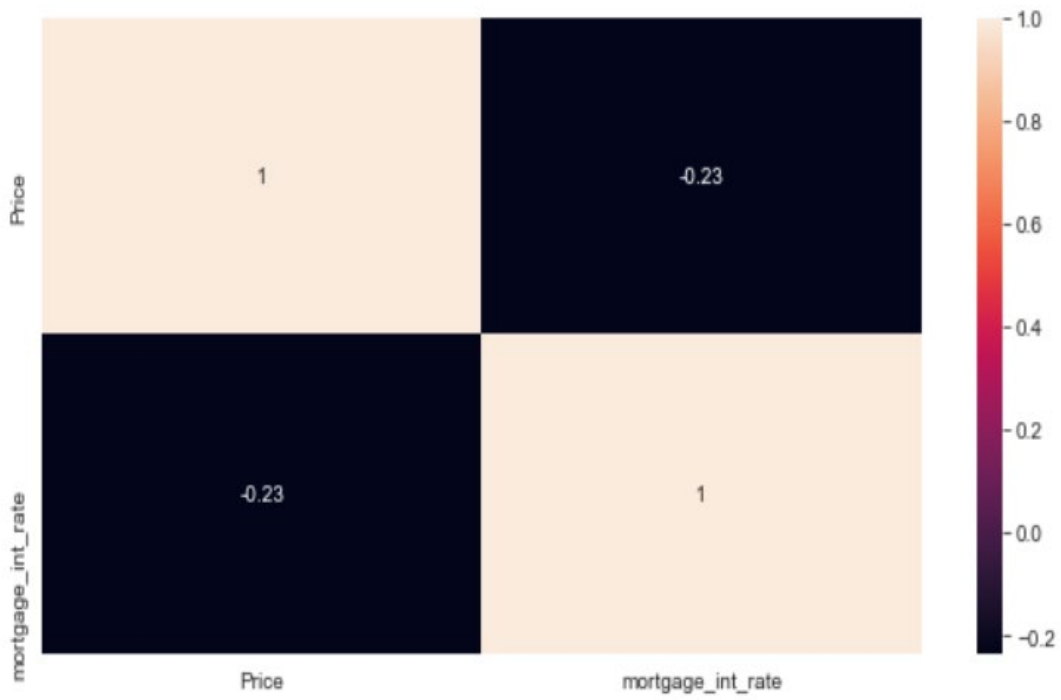
Visualizing the State-county house prices and covid cases during 2020-2021, there seems to be no much correlation between them. The number of recorded covid cases did not have any kind of impact on buying houses. In fact, around this time, the rate of increase in the housing prices has increased steadily. Any kind of a dip or spike in the number of covid cases has not affected directly. The onset of the pandemic in general increased the housing prices, but not quantitatively.

We can better confirm that there is no correlation between the covid case number and the housing price with a Pearson's correlation map below. The color map clearly shows the same.



Pearson's correlation for # covid cases and housing prices

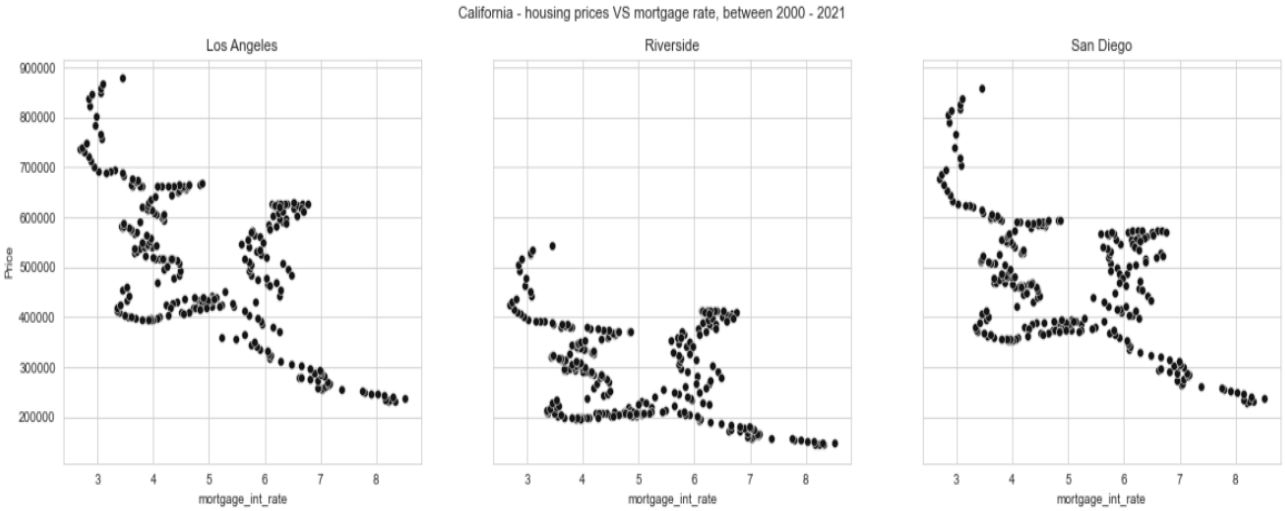
- The correlation between interest rates and housing price is negative, but there is a small correlation, which can be seen in the below correlation map.



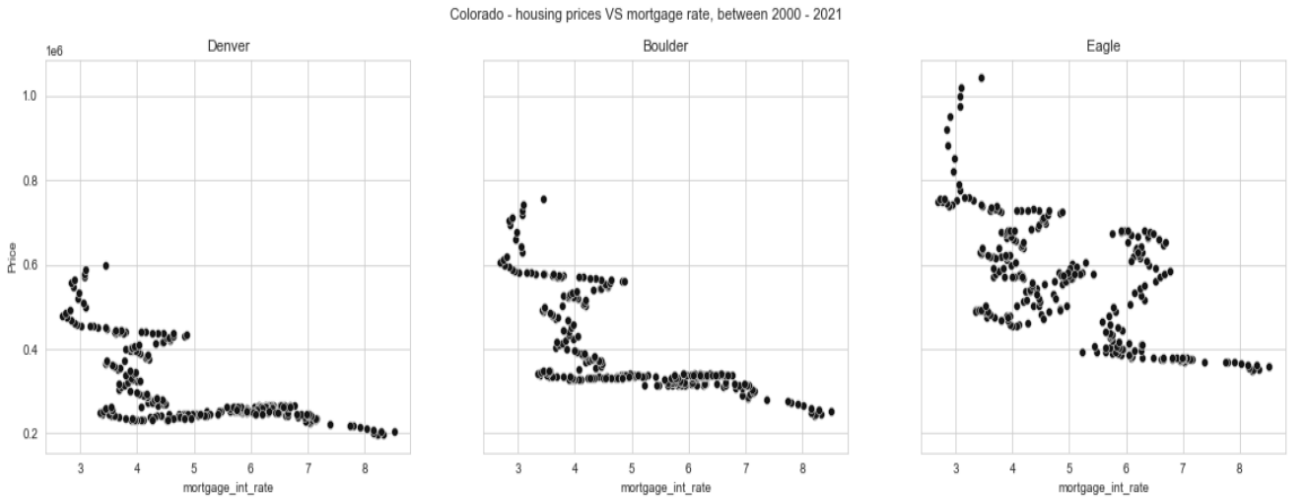
Pearson's correlation for interest rates and housing prices

- Visualizing relationship between mortgage interest rates and average housing prices between 2000- 2021, for random 3 counties in each state below, we can conclude that there is not so strong but a small negative correlation. All the plots also have a similar pattern. We can use this as a feature for linear regression model to predict housing prices.

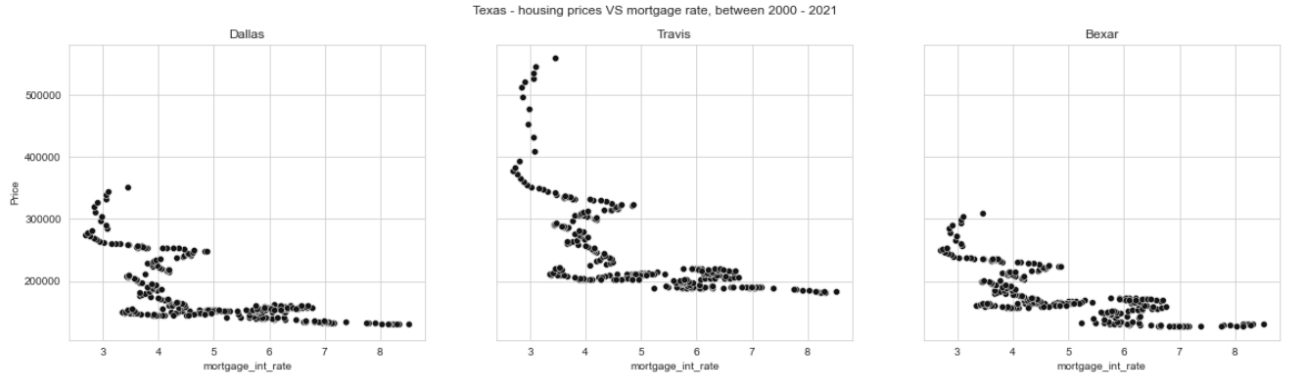
CALIFORNIA



COLORADO



TEXAS



VIRGINIA

WASHINGTON

4) MACHINE LEARNING MODEL TRAINING AND PREDICTION

- I have used Linear regression and time series model training and compared the evaluation metrics to select a better model for housing price prediction. Also, the great depression caused sudden dip and after the depression caused sudden increase in the housing price rates. So I have taken data after 2012 to train the model.

- **LINEAR REGRESSION**: The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric. The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (β). One additional coefficient is also added, giving the line an additional degree of freedom (e.g., moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient. For example, in a simple regression problem (a single x and a single y), the form of the model would be:

  $y = \beta 0 + \beta 1*x$

  In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g., β 0 and β 1 in the above example).
- I used **sklearn** package for linear regression model, training and testing split, standard scaler and metrics. The model is applied to each county- state combination. Since there are a lot of counties, I just considered the top county in each state. They are as follows:
  Los Angeles, CA
  Denver, CO

Dallas, TX
Virginia Beach, VA
King, WA

- The date value is considered as one of the feature columns for the linear regression. But we will have to convert it to a either monotonically increasing number to represent dates or convert to ordinal values. Ordinal date is a simple method used to manipulate the objects of DateTime class. It returns proleptic Gregorian ordinal of the date, where January 1 of year 1 has ordinal 1. The function returns the ordinal value for the given DateTime object.
- I have considered single linear regression (with date as a feature) and multiple linear regression (date and interest rates) with a combination of scaled and unscaled data to train and test the data. Below is the model metrics.

county, state	multiple LR	scaled	R2	R2_pred	MAE	MSE	RMSE
Los Angeles, California	Yes	No	0.96	-33.09	10027.79	241756906.7	15548.53
Denver, Colorado	Yes	No	0.97	-15.13	8251.81	112505449.4	10606.86
Dallas, Texas	Yes	No	0.99	-10.09	2896.51	20162314.39	4490.25
Virginia Beach, Virginia	Yes	No	0.87	-27.84	4179.71	32690749.57	5717.58
King, Washington	Yes	No	0.96	-17.65	12755.23	275919923.7	16610.84
Los Angeles, California	Yes	Yes	0.96	-33.09	10027.79	241756906.7	15548.53
Denver, Colorado	Yes	Yes	0.97	-15.13	8251.81	112505449.4	10606.86
Dallas, Texas	Yes	Yes	0.99	-10.09	2896.51	20162314.39	4490.25
Virginia Beach, Virginia	Yes	Yes	0.87	-27.84	4179.71	32690749.57	5717.58
King, Washington	Yes	Yes	0.96	-17.65	12755.23	275919923.7	16610.84
Los Angeles, California	No	No	0.96	-30.08	9929.13	226597526.6	15053.16
Denver, Colorado	No	No	0.97	-13.43	8680.02	108224797.1	10403.11
Dallas, Texas	No	No	0.98	-9.13	3458.25	21221382.78	4606.67
Virginia Beach, Virginia	No	No	0.83	-32.66	4986.37	43692919.06	6610.06
King, Washington	No	No	0.96	-16.12	13533.84	278507955.3	16688.56
Los Angeles, California	No	Yes	0.96	-30.08	9929.13	226597526.6	15053.16
Denver, Colorado	No	Yes	0.97	-13.43	8680.02	108224797.1	10403.11
Dallas, Texas	No	Yes	0.98	-9.13	3458.25	21221382.78	4606.67
Virginia Beach, Virginia	No	Yes	0.83	-32.66	4986.37	43692919.06	6610.06
King, Washington	No	Yes	0.96	-16.12	13533.84	278507955.3	16688.56

- The R2 score for all case scenario of Linear regression looks great, but the model performs very bad with prediction, as is seen with the predicted R2 score.

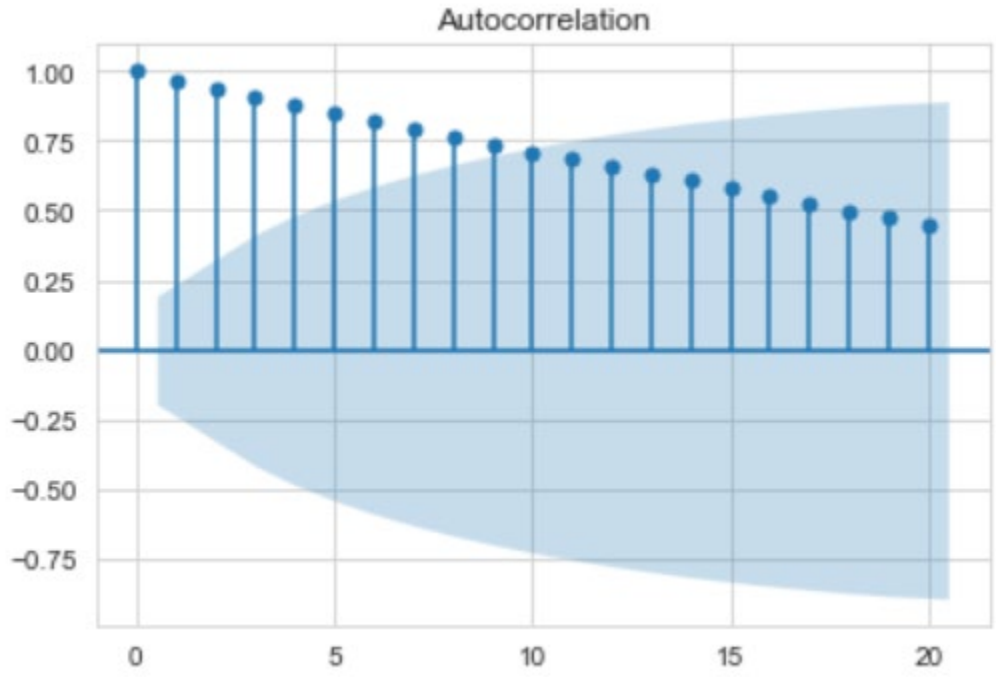
- **TIME SERIES MODEL:** Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record

data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. However, this type of analysis is not merely the act of collecting data over time.

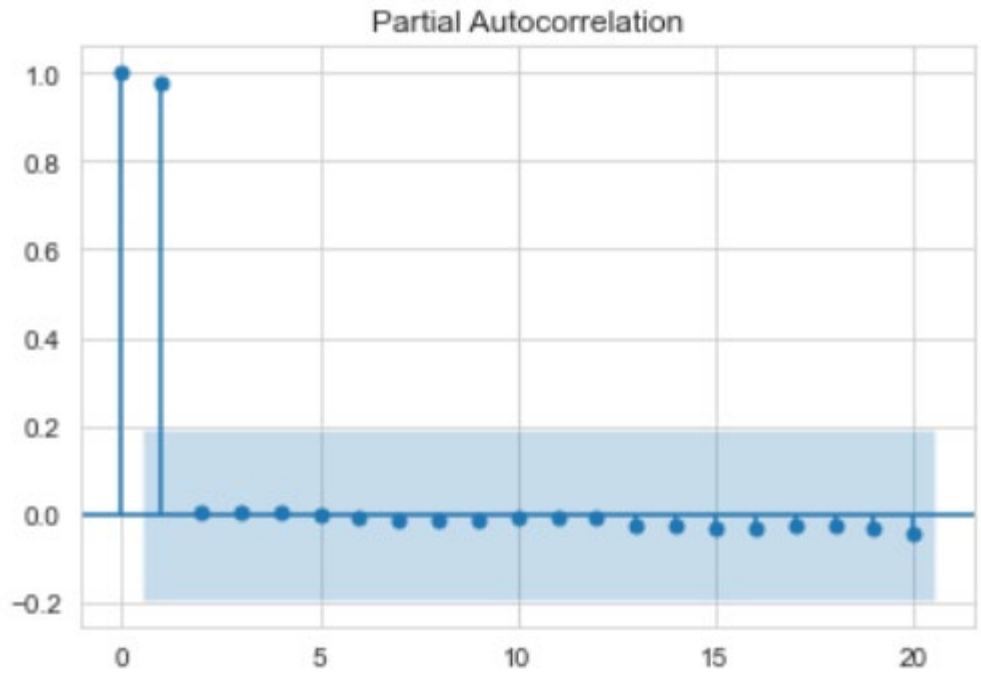
- I used date feature as index and trained a time series ARIMA model. The housing price data shows some seasonality and trend. I tested for stationarity of the data using **Dickey fuller** test and **kpss** test for each of the major county-state data. The series was not stationary and required a log transform and 1st order differencing to become stationary.

- PLOTTING ACF and PACF:

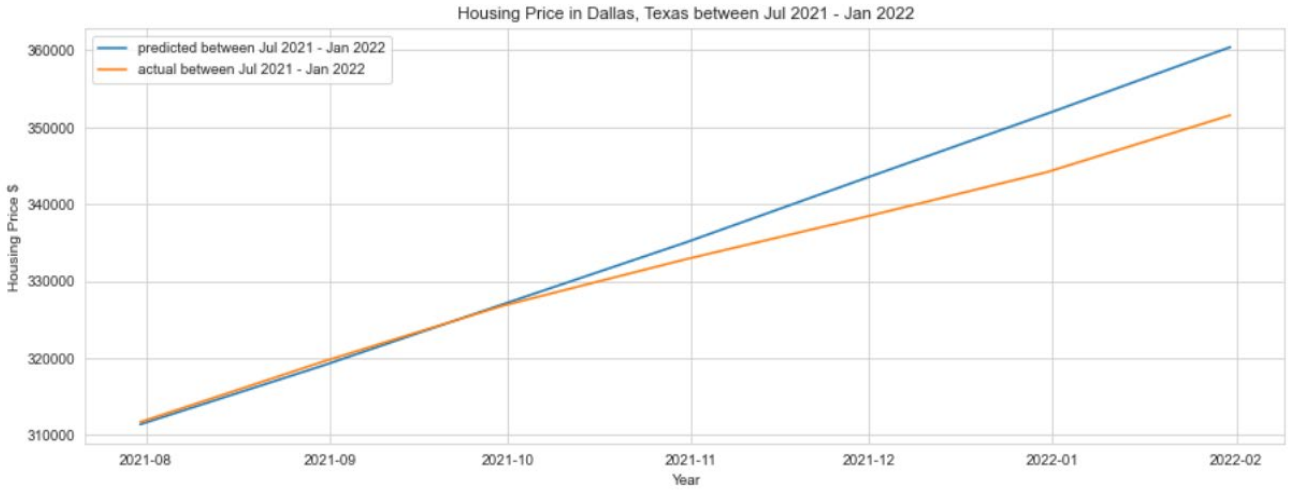
  ACF plot for Dallas, Texas data



  PACF plot for Dallas, Texas data



- The above plots have ACF that is tailing off and PACF cutting off at lag p = 1. Hence, this is an AR (1) model.
- The prediction R2 score is 0.93 for Dallas, Texas. Time series model predicts pretty well for Dallas, Texas data.



5) LEARNINGS AND FUTURE WORK

- The time series model seems to work okay for Dallas, Texas. But to train this model for every county/ region and generalize this time series model across all county-state data would not give better performance for every county-state data.
- We could try different non-linear model and see if its performance is better than the time series model. One such model to consider is Random Forest.

- In addition, we have only considered interest rates and date as feature in linear regression model. There may be multiple factors affecting housing price rates. In future, we could study and consider additional features for housing price prediction. Employment numbers, other government benefits during covid like stimulus check, location of houses etc. can be some features.