Project – Jogesh Pugazhendhi

USA Real Estate listings pre/during COVID and additional analysis

Concept: The concept is to analyze trends of real estate listings, price, and metrics in all states of USA for period Jan 2020 to June 2020. I choose this concept to understand how COVID have influenced home sales listings, trends starting from starting of the year to till now

The data is presented at year month and zip code level of granularity

The analysis is further expanded to add covid cases by month/zip name, mortgage rates by month, car sales by month to know if there is any co-relation between home listings and other metrics from the other datasets

I expanded the analysis with covid cases vs mortgage rate trend by month and also the car sales trend against the covid cases per month overall in USA

Dataset(s):

The data that is presented here has information related to home sales listings in all of USA for the month of June 2020. The data has info like active listings, new listings, median price listings, month on month and year on year change of all metrics listed below

Multiple files will be brought in to build history of listings data dating back to Jan 2020

Real Estate Listings (Source):
Source for this data: https://www.realtor.com/research/data/

Current Month (View Zip Data):
https://econdata.s3-us-west-2.amazonaws.com/Reports/Core/RDC_Inventory_Core_Metrics_Zip.csv

State Abbreviation reference info data:
Also, additional datasets have been brought in to enhance the data with region information and present with various dimensions (Region, State and Sub region)

state_region.csv

Contains USA state abbreviation, full state name, Region and sub region

COVID Cases Data:

COVID Cases data will be brought in to tie up to the Home Sales data to present any co-relation between COVID cases by states and real estate listing metrics

Source for the data: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

The data is presented by month and at zip code level. I plan to join this data at a Zip code level to the home sales listing data

Mortgage Interest Rate Data:
https://ycharts.com/indicators/30_year_mortgage_rate

I plan to scrape data from this website to get average mortgage rates for each month and plot it against other metrics to know if the mortgage rates had any influence on the home listings as well

Car Sales data by Brand by Month 2020:

https://www.goodcarbadcar.net/2020-us-vehicle-sales-figures-by-brand/

I plan to scrape data from the above website to get the car brand and sales for each month to see if there is any relation to car buying behavior based on other data metrics collected earlier to this. This data will be joined with the other datasets on month level aggregate to understand if metrics in this data has any co-relation with the metrics in the other data set

Dimensions and Metrics to be analyzed and measured (for Year 2020):

Dimensions for all metrics mentioned below: Month, City/County, State, Region
Measures (real estate listings): Average listing price, active listing count, avg days in the market, avg square feet, YoY change of all metrics mentioned earlier
Measures (COVID): Total covid cases by month
Measures (Mortgage rates): avg mortgage rate by month

Measures (Car Sales data): cars sold by month, cars sold by brand
Note: Car Sales data is only at month level, hence cannot be broken by state or region in the final report

What top Questions I am trying to answer?

1. Compare USA region's average listing price and their corresponding listing count
2. Overall, how the states fare against each other on median price
3. What's the trend of median price and price reduced count for the top 10 and bottom 10 states from Jan to Aug 2020
4. How the top 10 and bottom 10 states (based on the active listing count) fare against each other with comparing price vs days in the market vs active listing count

5. Is there any correlation between increase in covid cases vs mortgage interest rate from Jan to Aug 2020
6. Is there any correlation between increase in covid cases vs car sales from Jan to Aug 2020
7. Is there any correlation between increase in covid cases vs home listings and covid cases vs home median price from Jan to Aug 2020

USA Real Estate sales, listings, and YoY change on various metrics as of June 2020

Dataset: The data that is presented here has information related to home sales listings in all of USA for the month of June 2020. The data has info like active listings, new listings, median price listings, month on month and year on year change of all metrics listed below

Also, additional datasets have been brought in to enhance the data with region information and present with various dimensions (Region, State and Sub region)

Source:
Source for this data: https://www.realtor.com/research/data/

Current Month (View Zip Data):
https://econdata.s3-us-west-2.amazonaws.com/Reports/Core/RDC_Inventory_Core_Metrics_Zip.csv

State Abbreviation reference info data:


state_region.csv

Contains USA state abbreviation, full state name, Region and sub region

Data Exploration:

The dataset has home sale listings as of June 2020 with various metrics like active listings, median price, median days in the market, median sq ft of the house listed , month on month change of all metrics (compared to May 2020) and year on year change of all metrics (compared to June 2019)

The dataset has information of zip codes and the zip name for which the metrics are calculated

Extracting state abbreviation: Used the zip_name column to extract the last 2 characters as state abbreviation and created new column called Abb (state abbreviation)

Join reference info for state name and region: Imported new public dataset with the fields mentioned above and joined this dataset to the original dataset on state abbreviation

Report Dataframes:

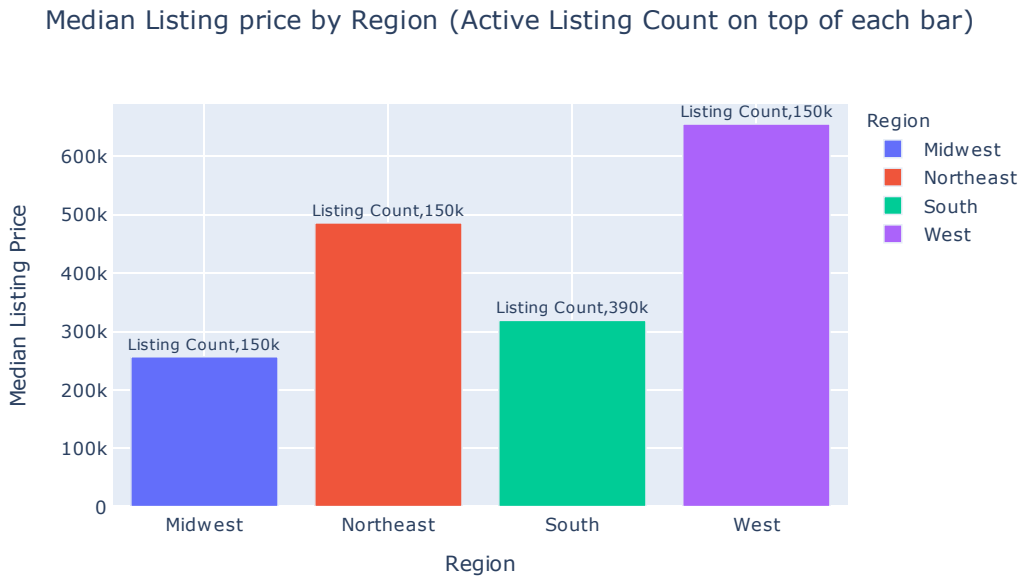
Report 1 (region_stats): From the original dataset, created a report on region level with metrics median_listing_price, active_listing_count and median_days_in_market

region_stats.csv

Output File Interpretation: As per the output file, West region has high median_listing price compared to other regions, south region having the most listings as of June 2020 and west region homes stay less time in the market

How the dataframe was created:
Used the average function for median_listing_price and median_days_in_market
Used the sum function to get total number of active_listing_count
report dataframe was created using the sqldf function in python – used group by region

Created a plot with information on median listing price and Listing count info by region



Report 2 (state_stats): From the original dataset, created a report on state level with metrics median_listing_price, active_listing_count and median_days_in_market

state_stats.csv

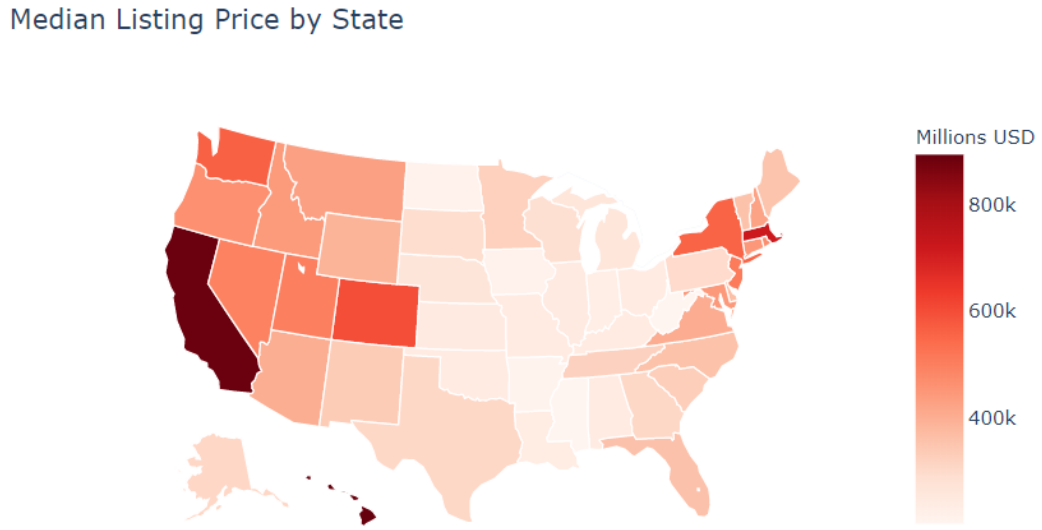
Output File Interpretation: California, Seattle, New york, Colorado, new jersey and Massachusetts are the states with high median listing price of homes compared to the other states

Mid-west states has the low prices for buying a home

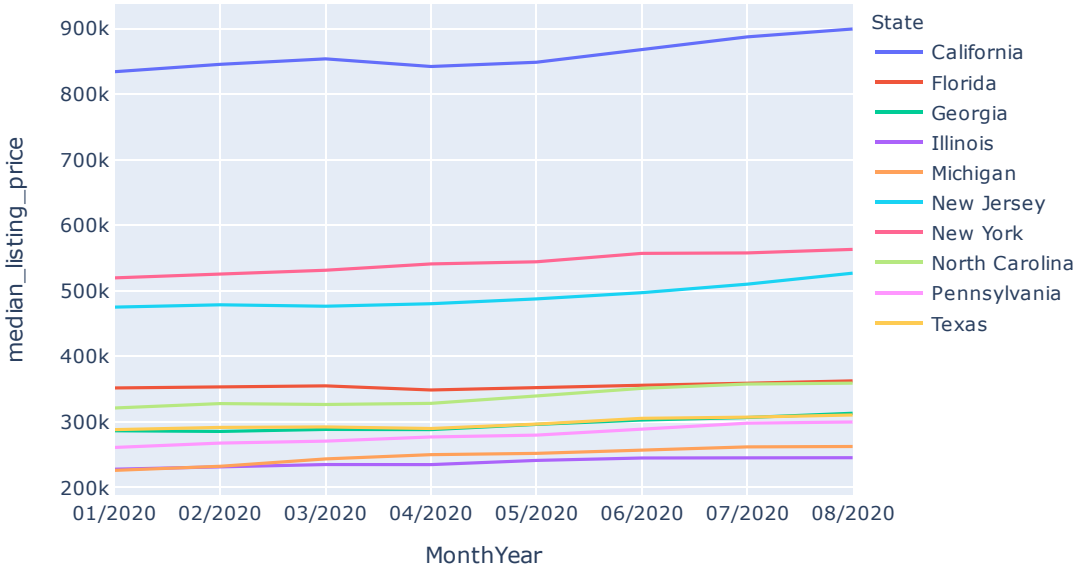
How the dataframe was created:
Used the average function for median_listing_price and median_days_in_market
Used the sum function to get total number of active_listing_count

Report dataframe was created using the sqldf function in python – used group by state

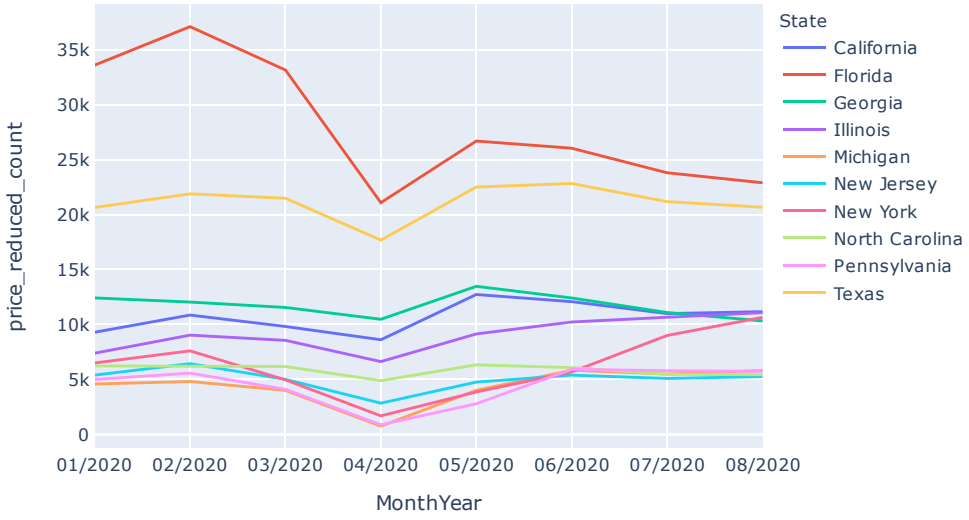
Created a plot with information on median listing price by state



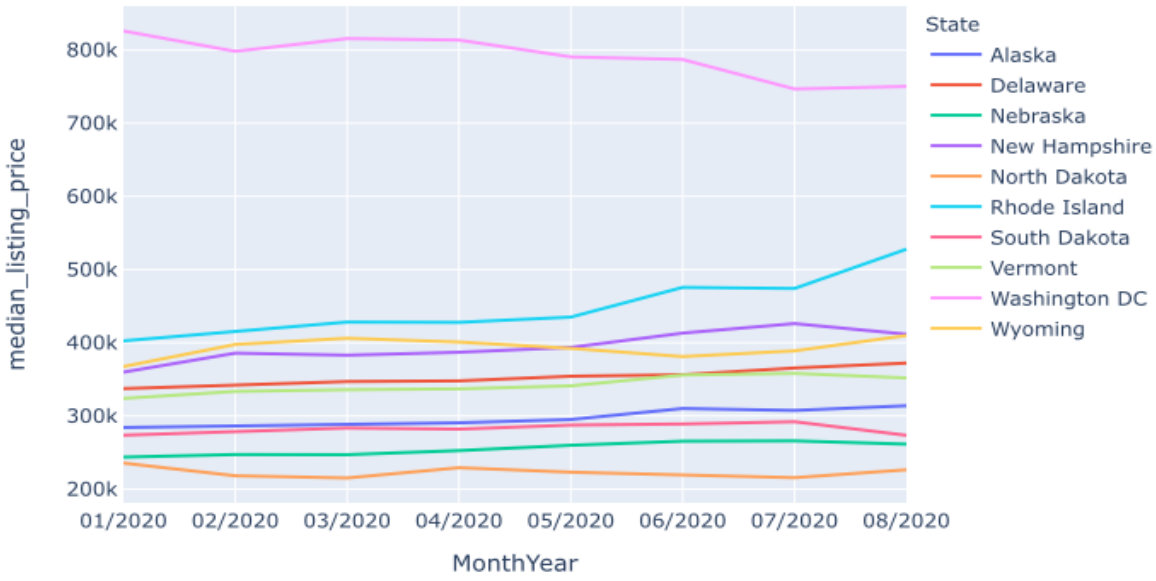
Top 10 States (Based on Active Listing Count) – Median Listing Price trend in 2020



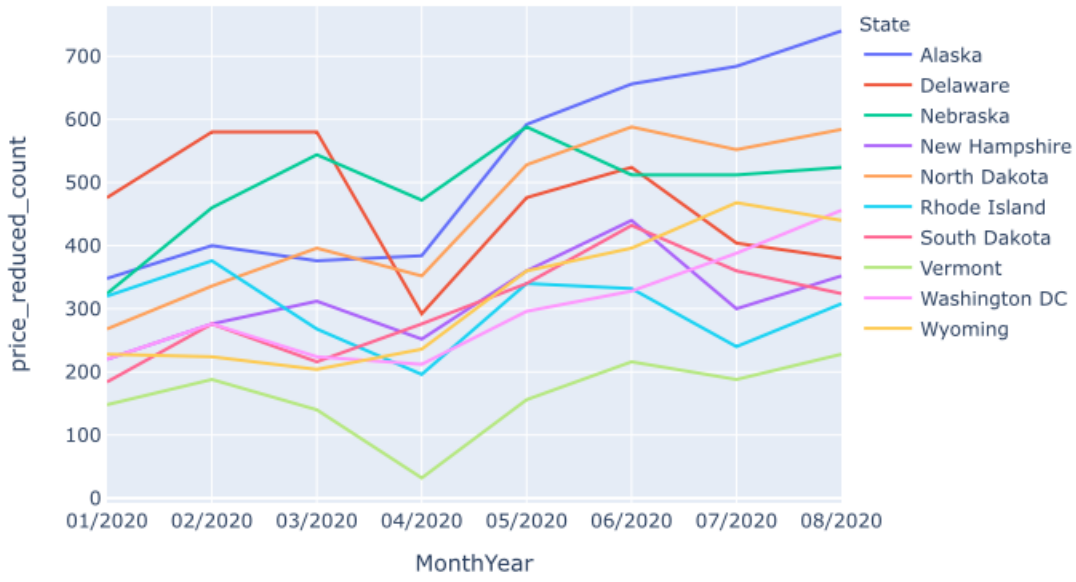
Top 10 States(Based on Active Listing Count) – Price Reduced Count trend in 2020



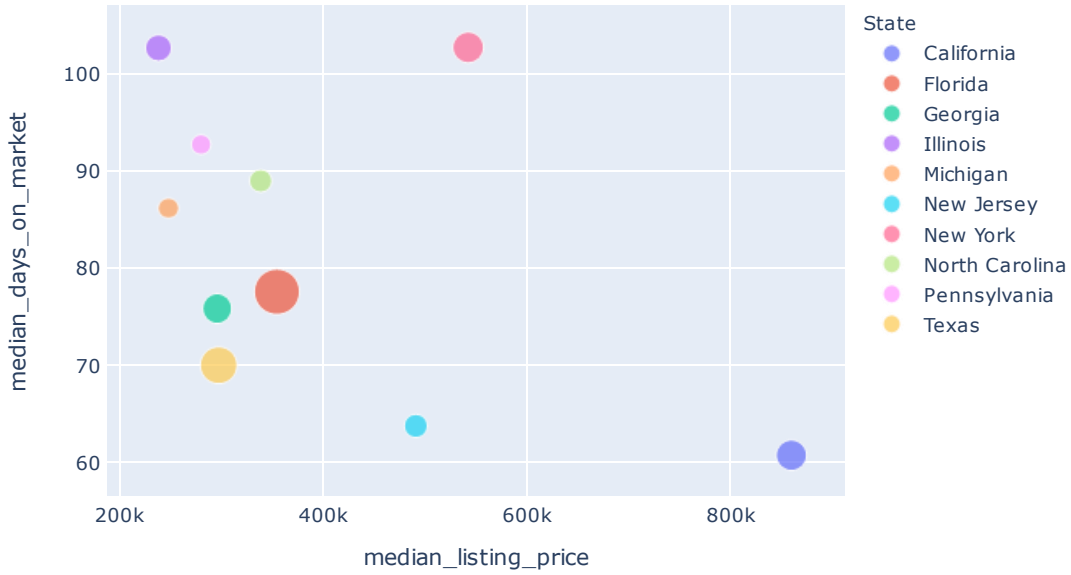
Bottom 10 States (Based on Active Listing Count) – Median Listing Price Trend in 2020



Bottom 10 States (Based on Active Listing Count) – Price Reduced Count trend in 2020



Top 10 States – Median Listing Price Vs Median days in the market (Bubble size Active Listing Count)



Bottom 10 States – Median Listing Price Vs Median days in the market (Bubble size Active Listing Count)

USA Covid Cases, Mortgage Rate and Car Sales from Jan 2020 to Jul 2020

Dataset (Covid Cases): The data that is presented here has information related to COVID cases that was pulled in from a website in CSV format . The data has info like zip code, state, cumulative covid cases over time on daily basis

Source for the data: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

Mortgage Rate: Additional datasets were scraped from the below websites to bring in information weekly mortgage rates from Jan 2020 to Jul 2020

Data has been scraped from below website link

https://ycharts.com/indicators/30_year_mortgage_rate

Car Sales data by Brand by Month 2020:
Monthly car sales data has been scraped from the below website to get car sales data from Jan 2020 to Jul 2020

https://www.goodcarbadcar.net/2020-us-vehicle-sales-figures-by-brand/

Contains car brand name, car sales

Data Exploration:

Aggregating Covid cases on MonthYear level: created aggregation dataset by summing up the covid cases by month

Mortgage Rate: Mortgage rates were scraped from the website using pandas html library. Read each column from the website and concatenated the monthYear and mortgage rates

Car Sales 2020: Scraped car sales data from the website, aggregated the data on MonthYear level to get the total sales by month

Final Merged Dataset: Combined all aggregated data from covid, mortgage rate and car sales on MonthYear level and created a final dataset for the report

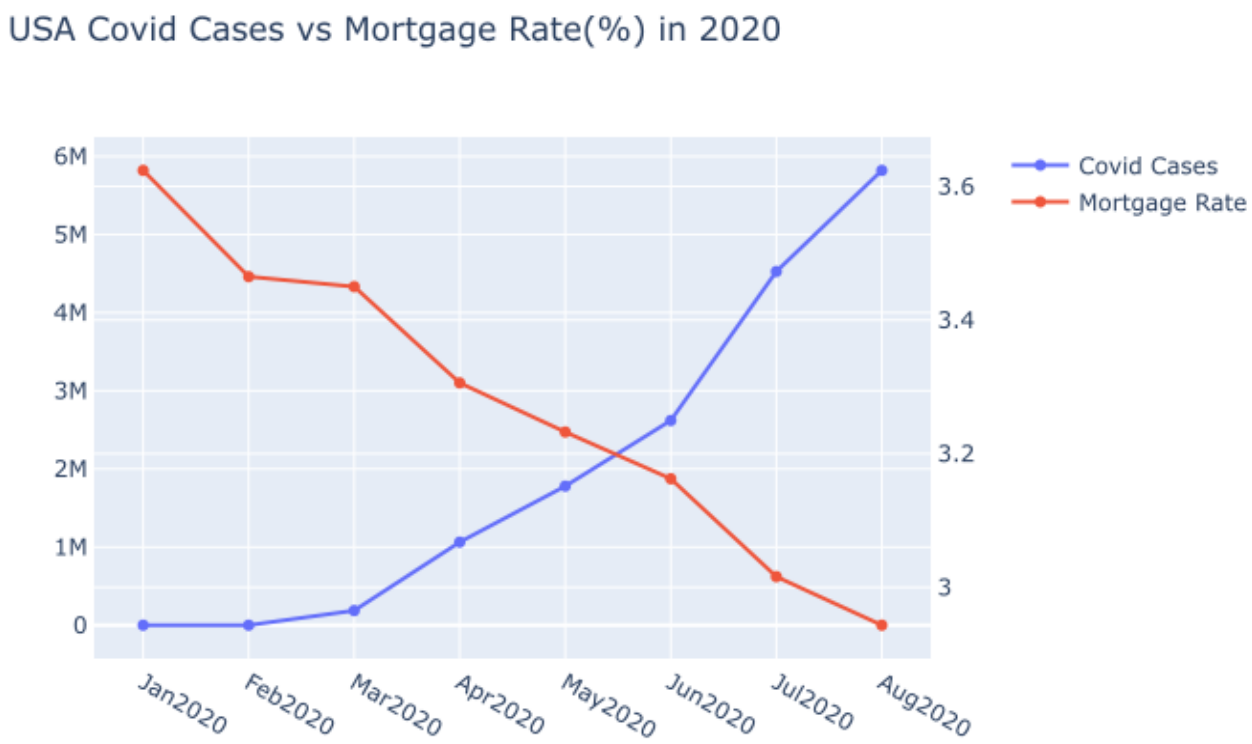
Report Dataframe:

Report : Created a final report with all metrics on monthYear Level from all the datasets combined.
Metrics – Covid Cases, Mortgage rate, USA car sales

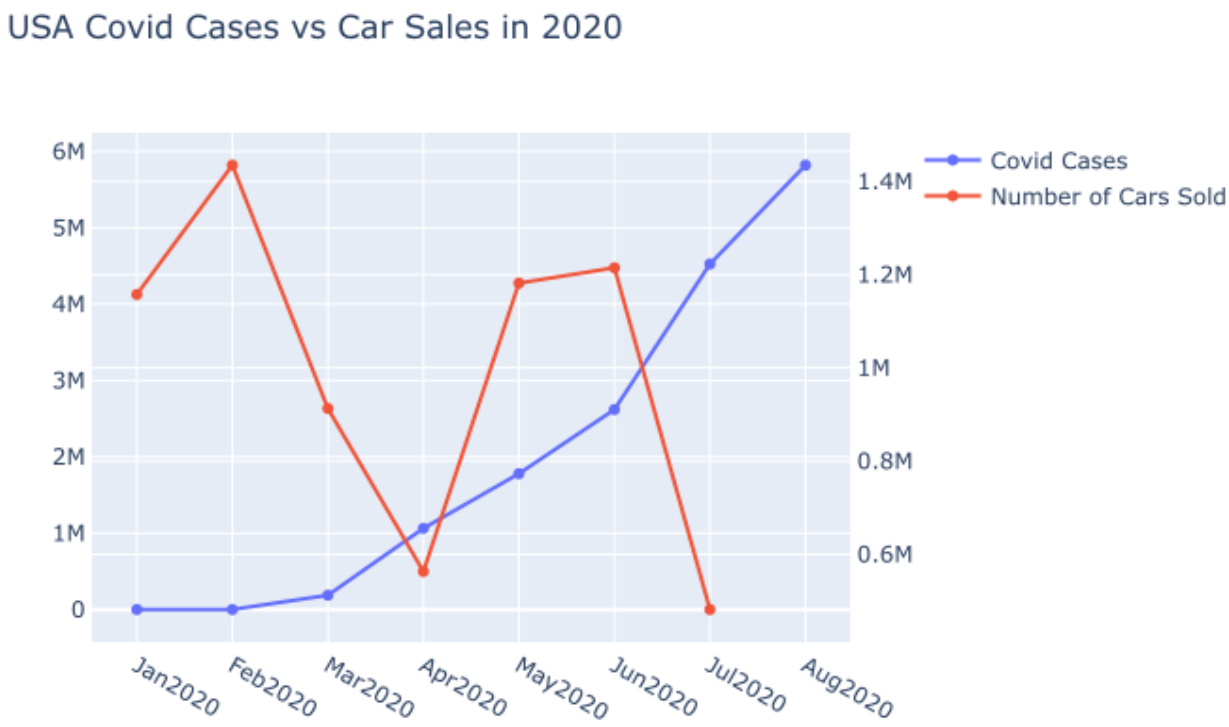
	MonthYear	covid_cases	rate_percen	car_sales
0	Jan-20	8	3.624	1,157,415
1	Feb-20	34	3.465	1,434,716
2	Mar-20	186750	3.45	912,800
3	Apr-20	1063314	3.306	563,122
4	May-20	1779689	3.2325	1,181,756
5	Jun-20	2619623	3.1625	1,214,706
6	Jul-20	4526434	3.016	481,587
7	Aug-20	5820085	2.9433333	

How the dataframe was created:
Used the average function for weekly mortgage rate on month level
Used the sum function to sum up all covid cases for all zip/states combined
report dataframe was created using the sqldf function in python – used group MontYear

Plot 1:

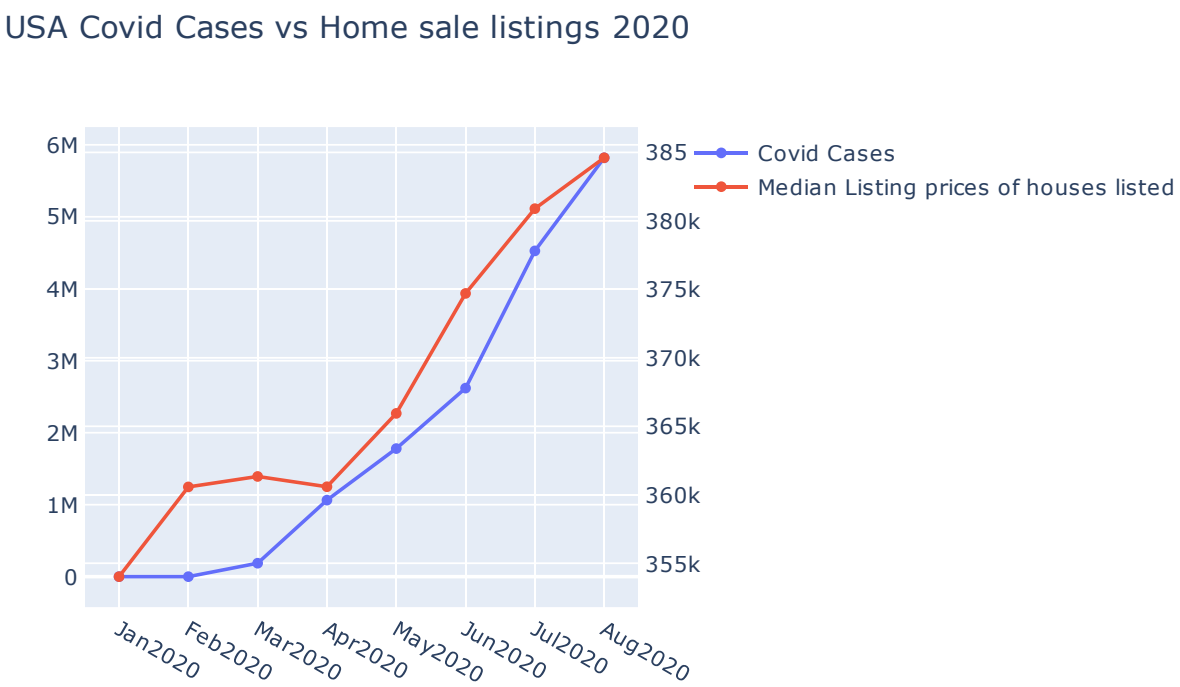


Plot 2:



Bringing in Home Sales Data to Covid final Dataset:

	MonthYear	covid_cases	rate_percent	car_sales	active_listing_count	median_listing_price
0	Jan-20	8	3.624	1157415	970991	354018.9308
1	Feb-20	34	3.465	1434716	948516	360572.0756
2	Mar-20	186750	3.45	912800	953363	361335.8697
3	Apr-20	1063314	3.306	563122	975109	360584.8586
4	May-20	1779689	3.2325	1181756	954991	365941.9481
5	Jun-20	2619623	3.1625	1214706	892559	374703.0173
6	Jul-20	4526434	3.016	481587	838158	380893.6395
7	Aug-20	5820085	2.935		795404	384607.8977



Conclusion:

Conclusion is as the covid cases increased, there was a decrease in the mortgage rates for the same period. The mortgage rate decrease could be due to rates policy or economic impact or other factors

Even though there was a lockdown in most parts of USA in May and June, there was steep increase in the number of cars sold. However, when lockdown was lifted for most parts of the country in Jun/Jul 2020 – there was a drastic decrease in the car sales during that time. Not very sure what factors influenced the increase in car sales from May to Jun 2020. Could be lot of cars sold at premium discounts, cheap prices and better credit terms offered by vendors during that time ?

There was also not a huge increase or decrease in the median prices in the top 10 and bottom 10 states during 2020. However, there were some biggest increase and decreases in the price reduction count of houses listed over period

California and Washington DC had the highest median price listings with listings sold out very soon based on listings of days in the market

One interesting trend is, as the COVID cases increased over time in 2020, the median price increased in all of US. This could be attributed to the lower mortgage interest rates