Covid-19's Impact on the Value of Single Family Residences

Background & Motivation

The global pandemic, Covid-19, has infected much of the world; changing our spatial relationships. While many may qualitatively agree with such an assertion, we looked to quantify the impact the pandemic and the resulting work from home culture shift impacted real estate markets in California and New York. Specifically, we researched how the pandemic impacted real estate transaction trends for purchased properties: **(1)** *is there any evidence of a value shift in the residential housing market;* **(2)** *if Covid-19 did have an impact on the housing market, are there proximal socio-economic insights about Covid-19's distributive impact* (e.g If housing trends did shift, did they universal shift or only for particular housing tiers?).

Data Acquisition and Manipulation

Dataset / Information	Source
Covid-19 Cumulative Data	https://github.com/nytimes/covid-19-data/blob/master/us-counties.csv
Covid-19 Rolling Data	https://github.com/nytimes/covid-19-data/tree/master/rolling-averages
Real Estate Transaction Data	https://api.gateway.attomdata.com/propertyapi/v1.0.0/sale/snapshot
Area Median Income Data*	https://www2.census.gov/programs-surveys/saipe/datasets/2019/2019-state-and-county/est19all.xls
Mortgage Rates*	http://www.freddiemac.com/pmms/docs/30yr_pmmsmnth.xls
Home Price Index*	https://fred.stlouisfed.org/series/CSUSHPISA
CA Population Data*	*https://worldpopulationreview.com/us-counties/states/ca*
NY Population Data*	*https://worldpopulationreview.com/us-counties/states/ny*

* Informational resource for descriptive purposes or referenced for calculations.

Our primary real estate transaction dataset was from ATTOM's real estate data AP. We used the *json* and *request* python libraries for the following procedure: **(1)** query state id list; **(2)** use state id code to query county geoids for 74 California counties and 90 New York counties; **(3)** converting the geoids into a list that we fed as parameters to a "sales_scrapper_by_county" function we built to extract the nested information and export the data into a local json file that was labelled with the geoid and transaction period.

Through trial and error we learned to circumvent a "hidden" 10,000 records per query limit enforced on our api keys by slicing smaller and smaller time frames until a query returned under 9,000 records. We were only afforded 250 API calls in our API development trail so we had to sign up for several accounts to grab the data. The incremental time slicing method left us with 322 files for California, 166 New York files and over 1.5 million transaction records with 43 features for transactions between ***January 1, 2019 to March 4, 2021***. We then took to parallelizing the Python processing of files using a distributed cluster in the DASK framework. Reducing the features down to 12 of interest: ***'lot_size', 'zipcode', 'latitude', 'longitude',***

'sub_type', 'type', 'year_built', 'size', 'bathrooms', 'bedrooms', 'transaction_date', 'sale_price'. Because our analysis was time dependent we dropped any record that did not have a transaction date. We also noticed that for several records one or more features such as size, bathrooms, bedrooms had questionable values– such as fourteen bathrooms and zero bedrooms in 80 sq ft. Noticeably New York City's records were far more compromised than California's and this is why we focused our analysis on notable California cities. This also regrettably impeded our ability to glean proximal socio-economic insights from our analysis because it undermined the credibility of our property classification based on residential features and pricing.

lot_size	zipcode	latitude	longitude	sub_type	type	year_built	size	bathrooms	bedrooms	transaction_date	sale_price	Price Per Size Unit	Price Per Bed Unit	Price Per Lot Size Unit	Tier	Monthly Mortgage	Affordability	County
0.2002	93552	34.575447	-118.057502	MISCELLANEOUS INDUSTRIAL	SFR	1989	1384	2.0	3	2020-05-05	338000	244.22	112666.67	1688311.69	Bottom	1051180.0	-57719.886965	Los Angeles
0.7083	91042	34.253355	-118.288772	MISCELLANEOUS INDUSTRIAL	CONDOMINIUM	1984	756	2.0	2	2019-12-24	380000	502.65	190000.00	536495.84	Middle	1497200.0	-82253.102955	Los Angeles
0.1577	90703	33.883652	-118.042877	MISCELLANEOUS INDUSTRIAL	CONDOMINIUM	1971	853	2.0	2	2019-04-18	435000	509.96	217500.00	2758402.03	Middle	1713900.0	-94172.631014	Los Angeles

Our primary Covid-19 dataset were the New York Times csv datasets. We confirmed the integrity of these datasets against Johns Hopkins, California and New York's published datasets. We merged the features of these datasets on the 'date', 'county', and 'state' columns into a single dataframe that had the following features: ***'Date', 'County', 'State', 'Cumulative Cases', 'Cumulative Deaths', 'Day's New Cases', 'Rolling 7 Day Cases Average', 'Rolling 7 Day Cases Average Per 100k', 'Day's Deaths', 'Rolling 7 Day Deaths Average', 'Rolling 7 Day Deaths Average Per 100k'***. The date range of this dataset is ***January 25, 2020 to May 17, 2021***.

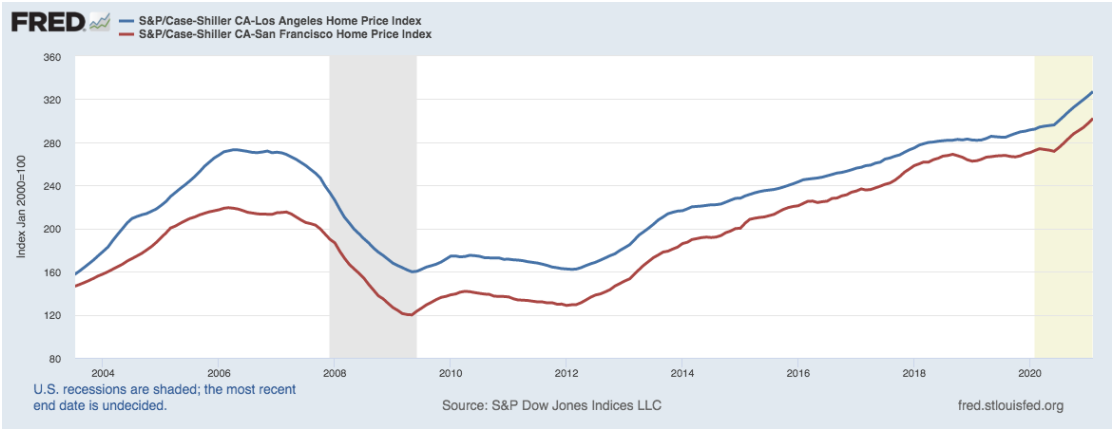
Date	County	State	CUM Cases	CUM Deaths	New Cases	R7 Cases AVG	R7 Cases AVG Per 100k	Days Deaths	R7 Deaths AVG	R7 Deaths AVG Per 100k
2020-11-19	Los Angeles	California	353316.0	7363.0	4896	3255.43	32.43	28	20.29	0.20
2020-10-09	San Francisco	California	11611.0	121.0	57	35.43	4.02	6	2.00	0.23
2020-03-14	Los Angeles	California	53.0	1.0	13	5.57	0.06	0	0.14	0.00

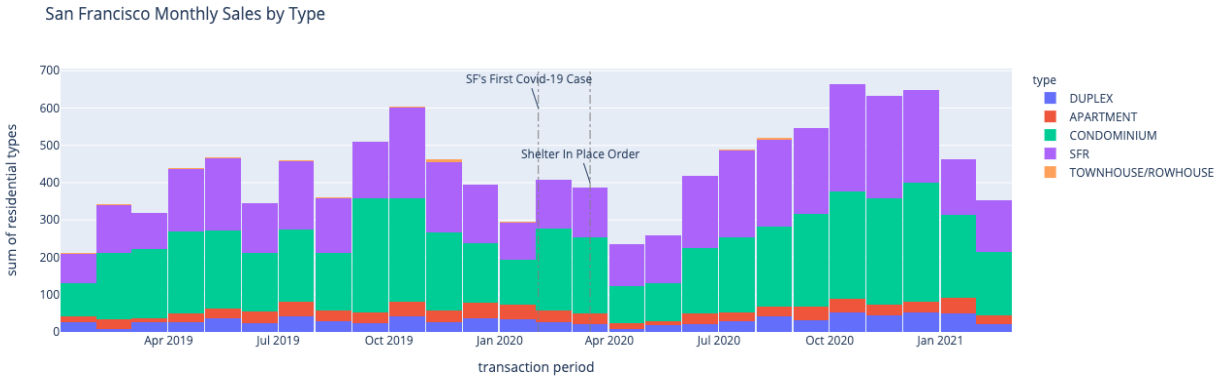
We then built a function to generate metrics, statistics and classifications. Because the analysis grounds each market within county lines we ran this function for each county. To calculate affordability we used a ratio of housing cost to income because it is the "standard", it is known as the 30 percent rule, and it is rooted in the 1969 public housing regulations. The calculations are listed in the table below.

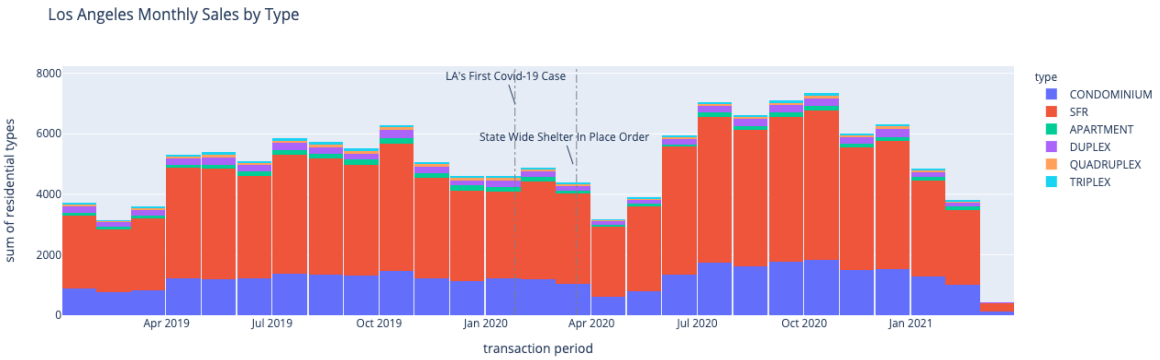
Data Type	Manipulation
Monthly Mortgage	P [i(1 + i)^n] / [(1 + i)^n – 1]
Affordability Index Value	100 * ((M_30 - Monthly Mortgage) / (M_30))
	M = monthly mortgage payment; P = the principal amount ; i = monthly interest rate; n = the number of payments over the life of the loan; M_30 = (30% of Area Median Income /12)
Rolling 30 Day Median of Daily Average Transaction Price	Group by day and report mean then create a 30 day rolling median of those averages
Property Tier	pandas qcut based on 'Price Per Size Unit' ; label ['Bottom','Middle','Top']
Price Per Size Unit	sale_price' / 'size'
Price Per Bed Unit	sale_price' /'bedrooms'
Price Per Lot Size Unit	sale_price' /'lot_size'

Analysis and Visualization

During our exploratory data analysis we decided to focus our correlational analysis and causal inference on Los Angeles (LA) and San Francisco (SF). We chose these county markets because the Difference in Difference (DiD) method we chose requires the important assumption that both treatment and control groups have similar time trends in the outcome variable prior to treatment. A particular strength of the DiD approach for drawing causal inferences is that the absolute values of outcome variable and unobservable confounders don't matter between the treatment and control groups allowing us to design DiD experiments with counties of different demographic and economic characteristics as long as the parallel trends assumption is valid. Over the last 15 years, LA and SF home prices have trended together, as verified from the St Louis Federal Reserve chart shown below. Parallel trends persisted even during economic downturns such as the 2009 recession. Thus by choosing these market's we satisfy the DiD assumptive trend requirements.

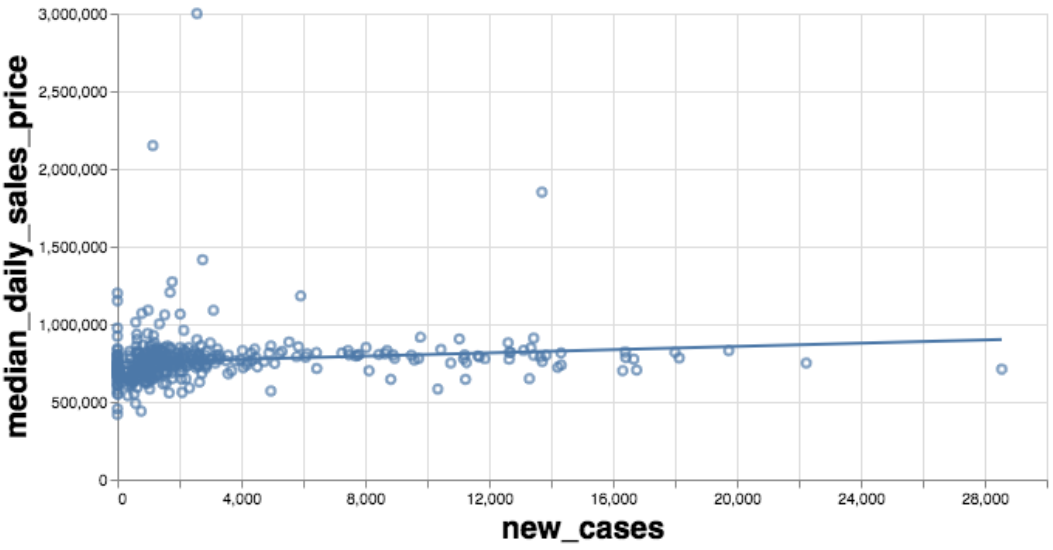






Ordinary Least Squares Regression Analysis

We began our analysis of the ***relationship between daily new Covid-19 cases and median home prices*** by merging the Covid and home price datasets by date and fitting a regression line, as seen in the chart below.

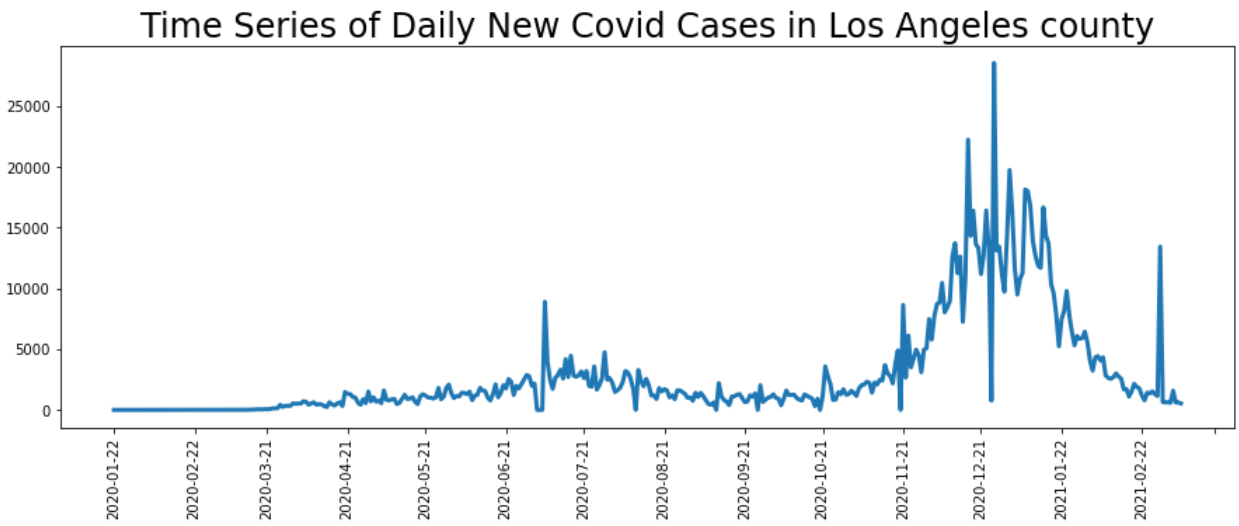


The slope indicates that as the number of new cases rises, the median sales price of Single Family Residences (SFR) in Los Angeles County rises too, albeit gradually. The fitted regression line indicates that for 1 additional daily reported case, the home price median increases by $5.20. This is to say that there is a positive linear relationship between daily cases count and median prices of homes

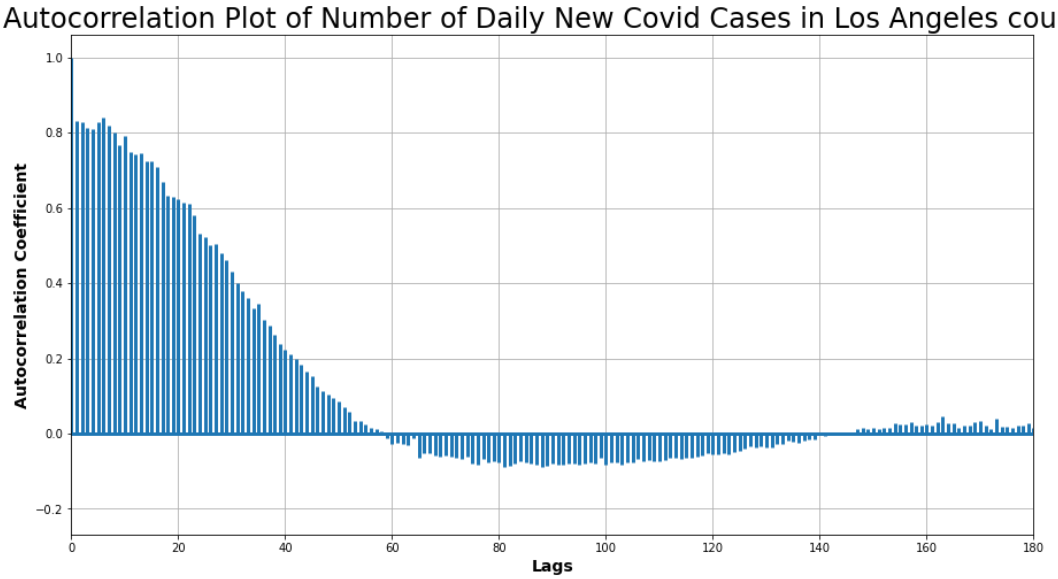
It is worth noting that this model is a generalization of the market. And that there are significant degrees of freedom not expressed in this simplification. Thus the predictive application of this model should be questioned with regard to hyperlocal factors. Nevertheless, generalizing is meaningful so we verified the significance of our results in a significance test. The 95% Confidence interval for slope is 1.2 to 9.3. The value ‘zero’ does not fall in this range. This means that the slope is statistically different from ‘zero’ and hence significant. And our p-value is 1.2% which indicates our slope is significant at 5% alpha level.

Correlation Analysis

Since our results showed significance, we then explored the volatility in daily case counts. Based on the number of cases we see on a given day, how well can we forecast the number of new cases after 1 week, 10 days, 20 days, 30 days, 60 days or 90 days? Answering such a question helps us understand the predictive power of our regression line as it is tied to case trends. We explore this question by analyzing Covid time series in time and frequency domains.



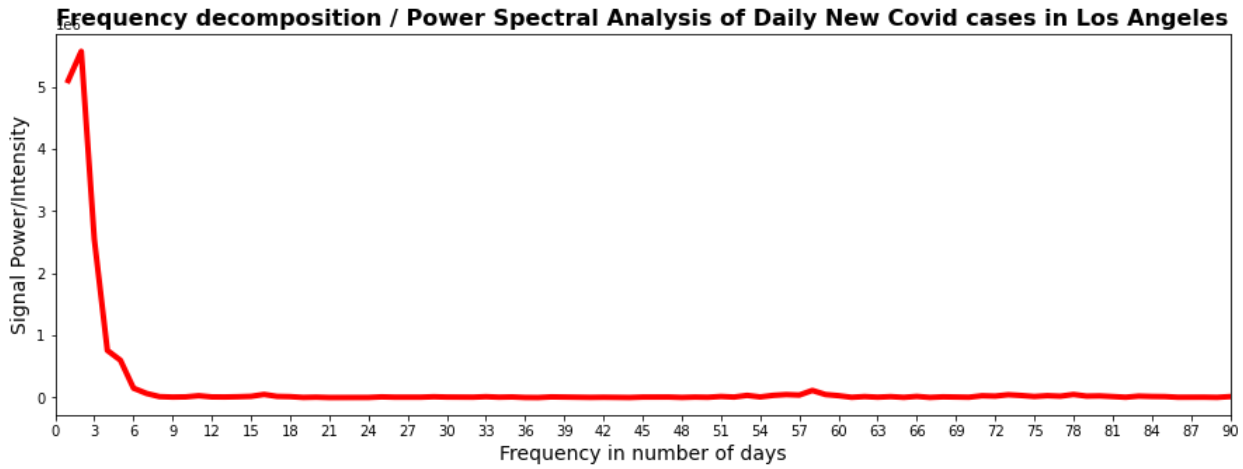
As seen above, there are a few big peaks and several small peaks which come and go. However, this line plot does not effectively express any kind of pattern or periodicity in the time series. For that, we need to plot time decomposition and autocorrelation. The autocorrelation plot shifts the time series by 1 lag, where each lag is 1 day and then we convolve it with the original (unshifted) time series. Convolving simply means finding the overlapping area in terms of percentage (%).



Interpreting the chart above, **for the first 7-8 lags, correlation remains very high, above 0.8**. This means that new Covid cases reported on a given day are very similar in number to those that will be reported in subsequent 7-8 days! Thereafter correlation steadily keeps dropping becoming progressively weaker upto 30 days. Noticeably, the number of new cases reported on a given day is completely unrelated to the number of cases that was reported 60 days (2 months) ago, as seen by correlation dropping to zero. Also, it is quite interesting to note that after 90 days, the cases are quite dissimilar. In summary, there is no evidence of a bi-monthly or quarterly pattern in time series, However there is evidence of **weekly periodicity patterns**.

Frequency Decomposition: Power Spectral Analysis

Frequency decomposition is a complementary approach to the autocorrelation method for identifying periodicity in any time series. Any time varying signal can be decomposed into a sum of periodic signals and white noise. We decompose the daily Covid cases time series into its frequency components using Fourier transform to investigate periodicity in daily Covid cases and the relative contribution of various frequencies (in number of days) to the overall signal.



The above power spectral analysis graph suggests that the most dominant frequencies which make up our Covid time series signal are less than **6 days**. Beyond 6 days, the contribution of higher frequency components is low and flat indicating randomness/white noise. This aligns with our time series autocorrelation plot insight of weekly trends.

Causal Inference

Hitherto, we calculated correlations and their statistical significance. But correlations can be spurious. Any two quantitative variables can be correlated, yet correlations can mask causal relationships or portray causality where none exists. This happens due to omitted variable bias and unobserved confounders. Thus, we needed a different analysis to explore **the causal effect of Coronavirus on home prices in Los Angeles County**.

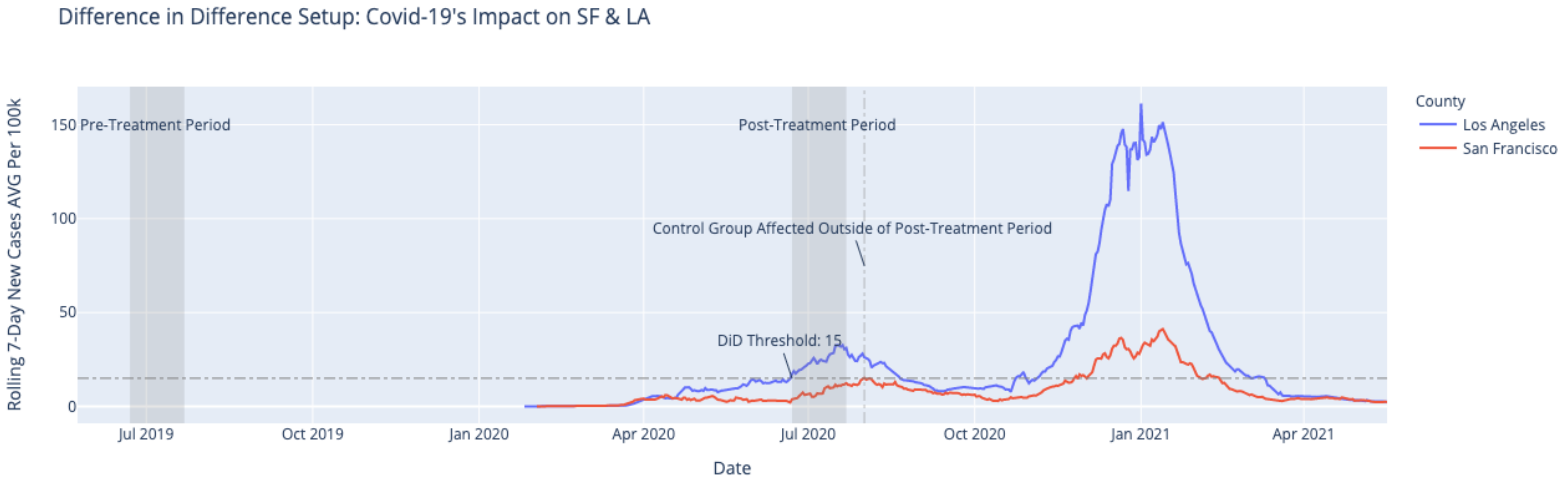
We took a null hypothesis approach and projected that home prices would fall due to Covid. In essence, this was an assumptions that Covid may have led to weakened demand for houses, due to some people losing jobs hence moving to rental units, some people falling sick hence pausing their plans of buying new house (or upgrading to a bigger house), or changes in attitudes towards perceived risk of real estate investing among other factors.

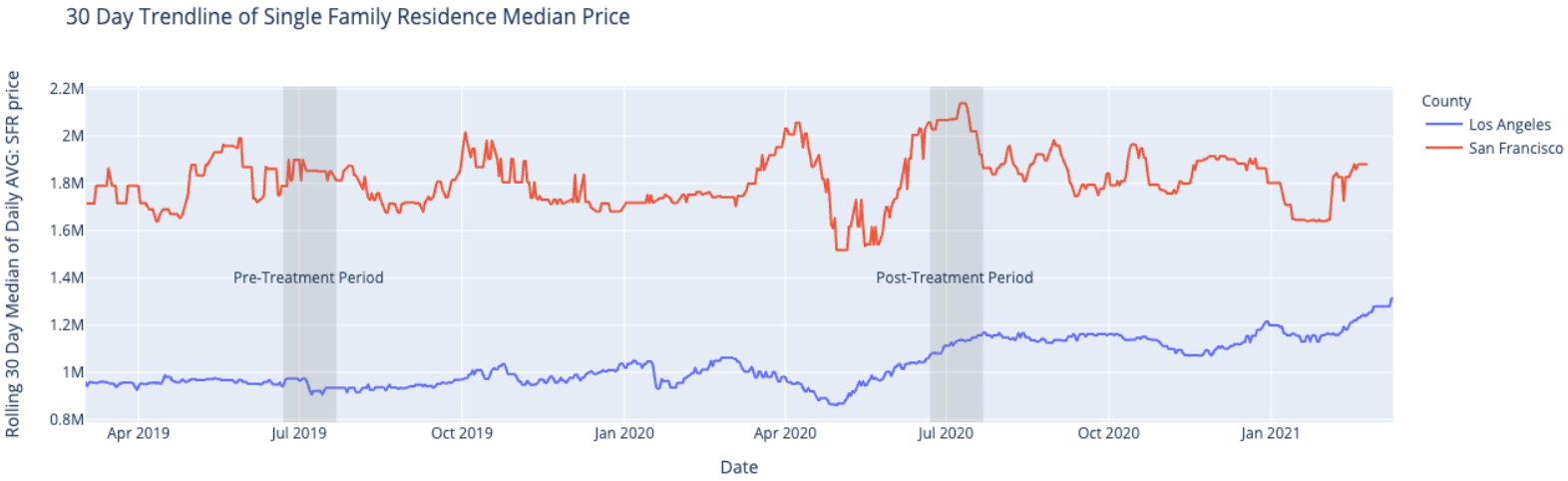
The difference in difference (DiD) method is a common statistical technique used to draw causal inferences from observational data. It calculates the ***causal effect of a treatment***, ***a sudden change like Covid hitting the world***, on an outcome variable, median home price of single family homes, by comparing the average change over time in the outcome variable for the treatment group, compared to the average change over time for the control group.

Essentially, in order *to truly know how home prices in Los Angeles county were causally impacted*, we need to consider how those home prices would have been had the county never experienced Covid-19, the counterfactual. However, the Covid-19 did hit the county and we never get to observe how those home prices would have fared without Covid-19.

DiD uses the outcome of the control group, ***in this case San Francisco was defined as the control group***, as a counterfactual proxy for what would have occurred in the treatment group had there been no treatment. *The difference in the average post-treatment outcomes between the treatment and control groups is a measure of the causal effect.*

In setting up our DiD experiment, we chose the "***7-day Rolling Average Per 100k***" of new Covid-19 cases as our ***treatment variable***. Choosing the relative indicator of per 100k was crucial to quantitatively comparing counties with dramatically different populations. Moreover, our time decomposition analysis showed that the daily case count reflects a 6 day signal so the rolling average is a strong indicator of Covid-19's impact within the treatment and control counties. In defining DiD's post-treatment period, the treatment group is affected, but the control group is not. In defining DiD's pre-treatment, neither treatment nor control group are affected. ***Again, in this case "treatment" is considered Covid-19's "sudden" impact***. To do this we had to choose a meaningful threshold for our *7-day Rolling Average* of New Covid Cases *Per 100k.* We chose ***15 as the value threshold,*** *the sudden impact moment,* in each county because it ***captured important aspects of both county's first waves***: it was the peak of SF's first wave**;** and the post treatment period captured the peak of LA's first wave. Thus we assume that this threshold allowed us to define a post treatment period which captured Covid-19's grip on LA county and its resultant economic impact. Again, there is an important parallel trend assumption that both treatment and control groups have similar time trends in the outcome variable, home price index, prior to treatment, which the Federal Reserve data supported.





Treatment Variable: 7-day Rolling Average Per 100k
Outcome Variable: Median home price of Single Family Homes (in USD)
Treatment group: Los Angeles County
Control group: San Francisco County

Post-Treatment Period:
June 22, 2020 to July 22, 2020

Pre-Treatment Period:
June 22, 2019 to July 22, 2019

Please note that in the pre-treatment period, neither groups were affected. In the post-treatment period, only the treatment group was affected, the control group was not.

Difference-in-Differences (DiD) Calculation

	Pre-treatment Period	Post-Treatment Period	Differences
Control group (San Francisco County)	$1,602,000	$1,672,500	4.5% Increase
Treatment Group (Los Angeles County)	$679,000	$760,000	12% Increase
		Difference in Differences	**7.5% Increase**

Our analysis shows that the Median Single Family Home price increased by 12% between the two time periods. However, absent the treatment, the counterfactual would have been a 4.5% increase. Hence, the average causal effect of treatment on the treatment group was a 7.5% increase in outcome variable! **In essence, Covid-19 caused single family home prices in Los Angeles to increase 7.5%**.

That said, this 4-point DiD method we used does not provide an uncertainty statistic for the DiD value. Therefore, the ***next steps*** to extending this causal analysis is to employ the more complex DiD method of 'DiD Regression with Fixed Effects', which provides a p-value and standard error for the DiD value in addition to allowing us to pick multiple counties as treatment group and control group. We could also make our results more precise by combining the 'DiD Regression with Fixed Effects' method with the 'Nearest Neighbor Matching' algorithm. This would involve 'matching' known 'treatment' units with simulated counterfactual 'control' units: characteristically equivalent units which did not receive treatment.

Statement of Work

Professor Alain Cohn reviewed our Difference in Difference causal inference and advised on the assumptive requirements needed.

Aditya Patel, collected and cleaned California's real estate data; created a dask dataframe setup for processing the data; conducted the correlation and time decomposition analysis; and led the DiD analysis; and helped build the report. Chris Westendorf, sourced the datasets; collected and cleaned New York's real estate data; conducted the exploratory data analysis; added the calculated features to the datasets; charted the DiD analysis and helped establish the assumptions that establish the DiD analysis; and helped build the report.

We collaborated through github: **https://github.com/adityahpatel/SIADS691-Covid-Real-Estate**