Exploring the Implications of COVID-19 on the Real Estate Market and Trying to Predict Rental/Selling Prices During the Pandemic

Sindhuja Rao, Tomas Ortega

Professor Dr. Anasse Bari

Big Data Science

Table of Contents

Abstract

It is clear that the COVID-19 pandemic has affected almost every aspect of our lives, so we ask, How has the pandemic affected the real estate industry? In this paper, we examine the relationship between the residential real estate market and the increase in cases of COVID-19 in the New York metropolitan area. To see how the prices in New York's real estate market have been affected by COVID-19, we collected data from zillow.com and transformed it into a time series so that we could perform time series analysis.

Our time series analysis involves Granger's Causality analysis followed by ARIMA and VAR modelling and conclusively prove that major shifts in the real estate market can be attributed to the rise in cases of COVID-19. The impact of COVID-19 could be better explained by additional factors such as vaccination rates as and when new data is collected which provides more leeway for future work. A different but equally important perspective would be to explore the relationship between our two key areas at a more granular level which can lead to more actionable insights.

Introduction

With an increase in testing and the release of vaccines, the pandemic appears to be under control, but on a closer look, the disruption caused over the past year (2020) intends to cause long-term effects in the migration patterns of citizens all over the country. The reasons for this migration vary widely, from people having to move out of their homes due to the loss of income during the pandemic, to moving for safety measures, or even in some cases, in need of ease of restrictions. In either case, this signifies a considerable change in population density for neighborhoods, counties, and even states by extension. Understanding these consequences and how they affect the real estate market, would allow firms to revise their property portfolios and make data-driven decisions on where and when to invest.

To make predictions about the state of the housing market, we must understand the factors that drive the current shifts in the market. First, the COVID-19 pandemic is causing a dramatic reduction in consumption, a drop in prices, and a decrease in workers' per capita income. Second, the acceptance of remote work is growing rapidly, and it seems that it might stay relevant for the foreseeable future. This means that the variability in the market will increase because people don't need to live near their workplace and are free to move away to a different state. Finally, remote learning also has a substantial effect on the rental prices in cities where a large population are students from out of state or international students.

Studying the Real estate market through changes in prices according to the impact of COVID-19 regionally might help us make decisions reliant upon where the consumer demand actually is. Migration patterns also help companies and governments understand the social and economic impact that can be attributed to the pandemic. This can in turn help organizations to direct

resources for development wherever necessary and anticipate demand appropriately. Our project thus intends to analyze this correlation and analyze causality (if it is present) between one of the basic needs of people and a deadly pandemic in one of the most turbulent times of our modern world.

Literature Review

As part of the process we reviewed four published papers in order to get deeper insights on the kind of work done with similar kinds of data beforehand albeit under different circumstances. In this section, we will go over them and explain how they inspired the decisions made in this project.

In the first paper "*COVID-19 Infects Real Estate Markets: Short and Mid-Run Effects on Housing Prices in Campania Region*", researchers tried to estimate real estate price changes resulting from the COVID-19 pandemic. To understand the impact of COVID-19, the researchers first explored more conventional models used to evaluate and predict the Average Housing Prices in a specific region. These models used for real estate usually use inputs such as the GDP, unemployment rate, residential real estate prices, and other real estate transactions.

In this paper, researchers determined that social and regional factors are the real estate market drivers that are most altered by the pandemic. They used regional data as the unit of analysis because it allowed them to make community-specific and region-specific assessments of the macroeconomic drivers of real estate prices. They found that the pandemic affected prices in several ways. Including permanent or temporary closing of neighborhoods or cities. Using this information, they were able to find a correlation between COVID-19 and real estate, and they predicted that the average housing price will have a decrease of -4.16% in the short term and -6.49% in the mid-run (from late 2020 to early 2021)

To get a better idea of what has been done in this field, we examined the paper titled "*Financialization, Real Estate and COVID-19 in the UK*". The focus of this paper was on

understanding how COVID-19 impacted the Real estate market combined with its recent changes in housing policies. In the UK, financialization has transformed the housing market. The deregulation of financial markets that took place from the 1980s onwards, combined with the privatization of social housing, has transformed UK real estate from an ordinary good, insulated to some extent from consumer and financial markets, into a valuable financial asset.

The financialization of real estate has had a largely negative impact on the UK's housing market, the wider economy and individual communities; wealth inequality, financial instability, gentrification and homelessness have all increased as the role of the financial sector in UK property has increased. The COVID-19 pandemic further worsened this situation and accelerated the downfall of real estate prices. Their conclusion was that the UK is sleepwalking into a potential eviction crisis, and ongoing loose monetary policy and a pandemic therefore is likely to prevent a significant and necessary correction in house prices over the long term.

The paper "*A First Look at the Impact of Covid-19 on Commercial Real Estate Prices*" focuses on the movements in stock prices of REITS (Real Estate Investment trusts) who own CREs (Commercial Real estate) assets. Movements in a firm's stock price are largely driven by the perceived current and future productivity of the firm's underlying assets; therefore, it is important to understand how the COVID-19 shock transmits to the equity markets from a firm's asset base. To examine how the growth rates of COVID-19 cases affect firms differently through their asset holdings, the research involved constructing a novel firm-level measure of geographically weighted COVID-19 growth that varies daily during sample periods.

Findings highlight the importance of the asset-level attributes of a firm's portfolio to stock price reactions to the pandemic. Specifically, the key drivers are the property type (business) focus of

the firm, the geographic allocation of assets, and the interaction between these two attributes.

Their conclusion was whether the shock of COVID-19 on stock prices remains significant in the long run crucially depends on the resilience of the overall economy and, perhaps more importantly, how perceptions of risk change after the pandemic.

Lastly the paper "*Time Series Analysis of COVID-19 Data to Study the Effect of Lockdown and Unlock in India*" talks about the ongoing COVID-19 pandemic that has caused worldwide socio-economic unrest, forcing governments to introduce extreme measures to reduce its spread. Being able to accurately forecast the effect of unlocking in India would allow governments to alter their policies accordingly and plan ahead. The study investigated prediction forecasts using the ARIMA model on the COVID-19 data on the lockdown period and the unlock period. In this work, they considered not only the number of positive COVID-19 cases but also considered the number of tests carried out. The time series data sample was collected till June 2020, and the prediction and analysis are done for August 2020. The model developed and the forecasted results align very closely with the actual number of cases, and some important inferences have been drawn through the experimentation.

Business Understanding

Since the advent of the pandemic there have been major shifts in the lifestyle of an average individual concerning all aspects of financial expenses. Our problem focuses on if there has been a significant impact of the number of COVID-19 cases on the rental price market of New York for the year 2020 and part of 2021. Our hypothesis is that an increase in the number of cases should have led to a decrease in the average rental and selling price in New York City with its five major boroughs - Manhattan, Queens, Brooklyn, Bronx and Long Island under consideration.

We approached this problem as a Time series analysis as our data is distributed evenly and can be explored as a record corresponding to the mean count of cases on that day and the mean rental price on the same day. While we have more attributes such as case rates, testing counts and rates, death counts and rates on COVID-19 and number of listings etc. on Rent to explore upon we chose case count and mean rent prices as a starting point. A time series analysis also provided us with the room necessary to account for lags in reflection of prices i.e. changes in case count could not possibly reflect immediate effects in rental prices on the same day which seems a reasonable assumption to make.

While our hypothesis revolved around finding a causal relationship, we decided to set up the conditions for predicting rental prices in case a considerable relationship did emerge. On the other hand in the absence of a relationship we planned to continue exploring on the other attributes mentioned above and/or look for relationships on a more granular level i.e. borough

wise or county wise. The involvement of heavy statistical analysis meant we had to account for experimentation with various hyperparameter values every step of the way. Interpretation of these results were also subject to understanding what value was added on by each of these tests. Granger causality was a key point we needed to establish for definite causal relationships and ARIMA and VAR modelling provided us a way to predict multivariate time series. A number of tests such as Augmented Dickey Fuller test (ADF), KPSS test, Johansen cointegration test, etc. were significant in confirming our hypothesis along the way.

Data Understanding

For this project, we decided to collect information about real estate properties from the popular listing site zillow.com. In order to get the features from the properties listed on the site, we had to implement a web scraper. The decision to implement our own scraper rather than using a third party scraper was made because of the various mechanisms that Zillow implements to discourage scraping such as captcha pages and IP and user-agent detection. For this scraper, we used python and several of its libraries to collect the data and the steps we took to implement it are specified below.

First, we had to inspect how Zillow shows the information to the user in order to identify the pages with the information that we want to scrape. Zillow only shows the first 500 properties for any search that we make, so to get as many properties as possible in a zip code, we had to make the searches more specific. To solve this issue, we just searched for properties in a small range of prices, for example, properties in the price range of 100,000 to 200,000 so that we get fewer than 500 entries. Then, we coded tools to create the query URLs containing that information. These URLs had to have information such as the area from which we want the properties, the zip codes, map information to display the properties, the range of prices, etc. Once we were able to create a URL to get the properties in a specific area, we used BeautifulSoup to parse the web pages and get the features from each property.

To avoid spending too much time collecting the data, we used multi-threading. Since we didn't want to take up too many resources from Zillow's servers; we attempted to achieve a balance between speed and the resources used, so we only used a total of 10 threads to collect all our data. We used object-oriented programming to build the architecture of the scraper in a modular

way so that in future iterations, the system could be easily extendable to scrape other listing websites such as realtor.com or trulia.com.

We were able to collect data from 2007 until April 25th 2021. The number of properties that we collected from the zip codes that we specifically wanted is 209,492. From this two hundred thousand properties, 48,321 are properties for rent and 161,171 are properties either for sale or sold. From these properties, we selected only the ones that are currently for sale or for rent, or that were sold or rented from January 2020 until April 2021.

We collected COVID-19 data from tracktherecovery.org which is a website that aims to track the recovery from the COVID-19 pandemic in the United States. It contains data from several sources such as the CDC and Google's mobility data. The COVID-19 data includes the daily count and rate per 100,000 people of confirmed COVID-19 cases, deaths or tests performed. The data is already in a time series format and has 0 missing since it only contains data from when the pandemic started until the day it was collected by us.

Data Preparation

To prepare the real estate data for time series analysis, we first removed outliers from our dataset. We defined outliers in the usual way, we first calculated the interquartile range (IQR) and then removed the properties with values higher than 1.5*IQR above the third quartile and below the first quartile. After removing outliers, we aggregated the data by date in order to create a time series. The functions that we used to aggregate the data were the mean and the number of entries for each date. Since there are days for which we don't have any data available, we had to fill in missing values. We tried several ways to fill missing values such as linear, quadratic, cubic, and slinear interpolation, and replacing them with zeros, but we decided to use a spline function with order equal to 3. Spline interpolation uses low-degree polynomials in each of the intervals, and chooses the polynomial pieces such that they fit smoothly together. The resulting function is called a spline.

We then decided to smooth the data using a running average of 14 days. We attribute a higher window size due to the amount of noise possibly present in the raw data. Several smoothing techniques were tested, and we decided to go with a running average because it was the one that gave us better results. Furthermore, we also experimented with several ways of calculating the running average. The window operations that we compared were a trailing window and a centered window. We decided to use the trailing window because we are trying to make a forecast and if the model is deployed, we are not going to be able to get the values necessary for the centered window operation.

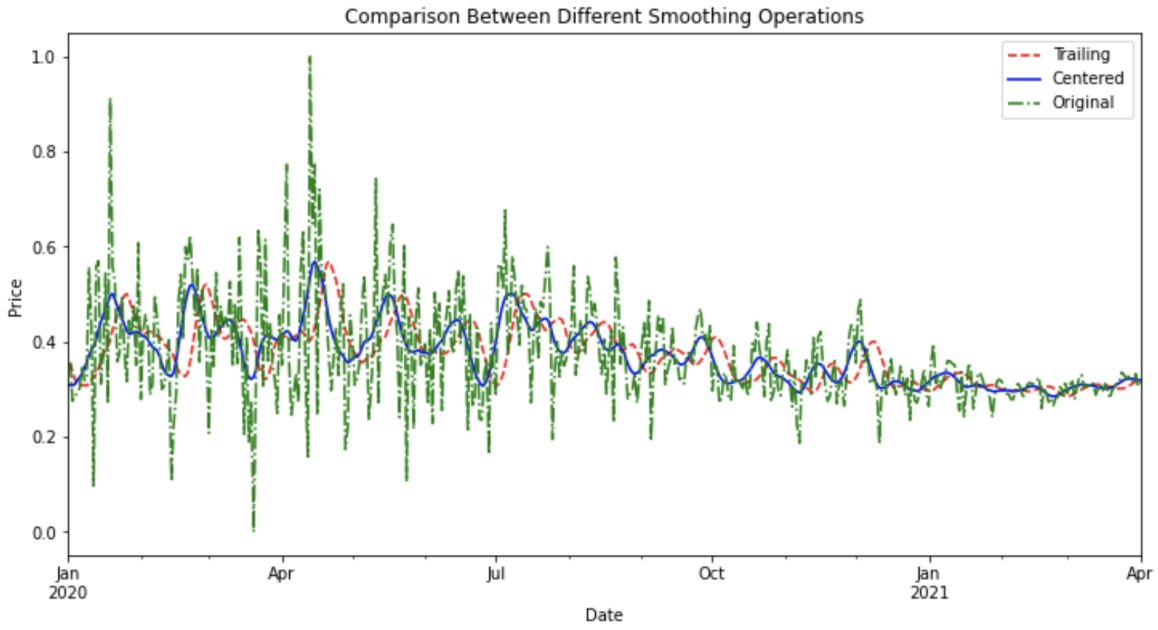


Figure 1: Comparison between different smoothing operations

Additionally, we also tried using the log transformation on the COVID-19 time series to remove any noise that might be present in the data. We ended up not using the log transformation because we achieved better results in the correlations and the forecasts without applying the logarithm.

Another preprocessing step that we took was scaling the data using min-max scaling. We applied the same scaling algorithm to every feature so that they are all in the range from 0 to 1. In the figure below, we can see the real estate data after it has been scaled and smoothed using the previously mentioned operations.

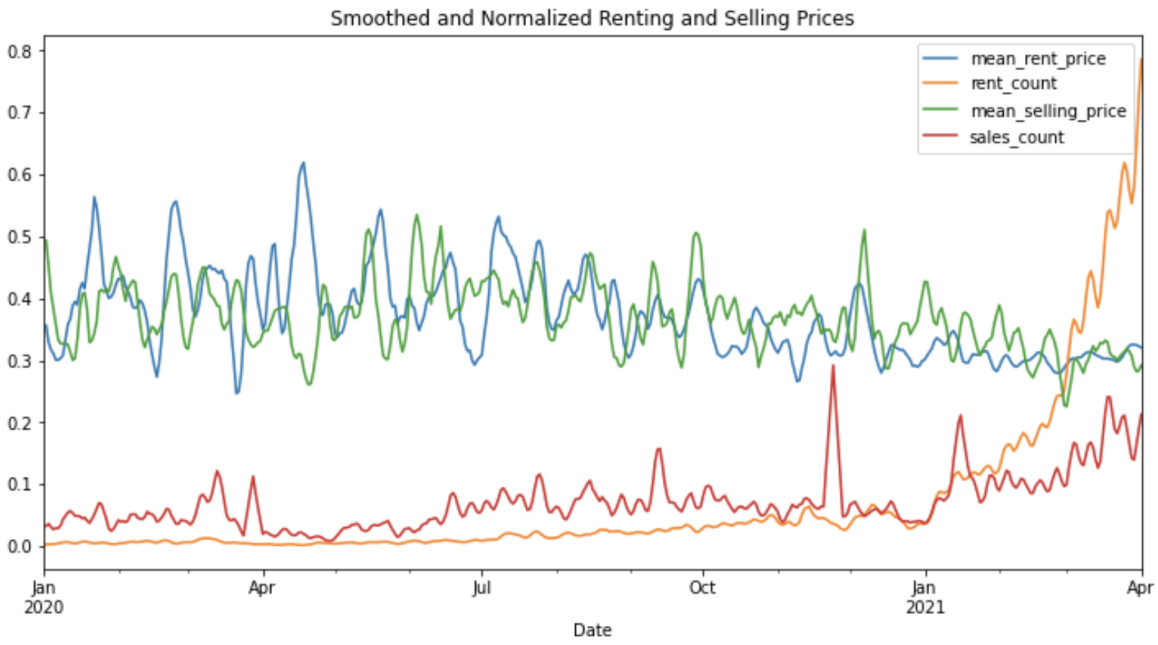


Figure 2: Smoothed and Normalized Renting and Selling Prices

We also applied min-max scaling to the COVID-19 data as shown in the figure below.

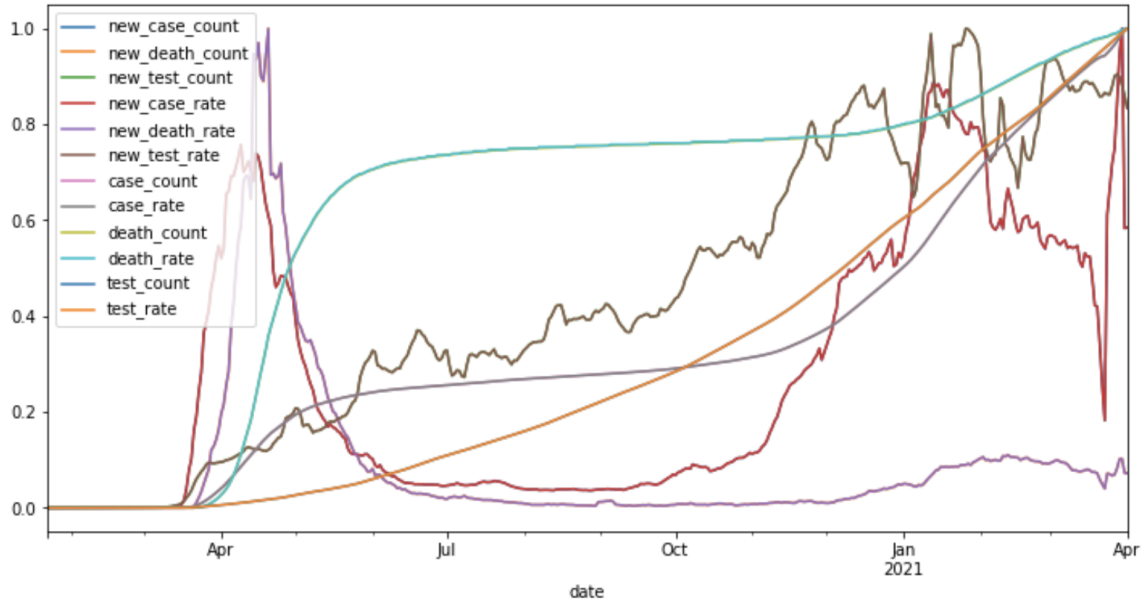


Figure 3: Min-max scaling of Covid-19 data

Time Series Decomposition

Time series decomposition is extremely useful when it comes to studying time series data, and exploring changes over time, but it can also be used for forecasting and for removing the trend and seasonality of a time series. This technique involves thinking of a time series as a combination of level, trend, seasonality, and noise components. Decomposition provides a useful abstract model for thinking about the different aspects of a time series, and for better understanding problems during time series analysis and forecasting, which will be our goal in the following few sections.

Time series can usually be broken down into two distinct components, these are called, systematic and unsystematic components. Any component of a time series that have consistency or recurrence and can be described and modeled we call a systematic component. On the other hand, components that cannot be directly modeled are called non-systematic components. Generally, a given time series is thought to consist of three systematic components; level, trend, and seasonality, but in addition to those components, we also have one non-systematic component called noise. The level of a time series can be defined as the average value in the series (we did not really focus on this component during this analysis). The trend is the increasing or decreasing component in the series. The seasonality components contains the short-term cycle in the series. And finally, the noise component contains the random (or not so random) variations in the series. This final component is often also called the residual since it is the part of the series that remains after removing the other components.

The two ways of decomposing a time series are: multiplicative decomposition and additive decomposition. The names are pretty self-explanatory, since in the first one, the systematic

components are added together to form the original time series, while in the second, the systematic components are multiplied together. We decided to use additive decomposition because it was the one that seemed most appropriate for our data.

For this project, we used the python module statsmodels and its seasonal_decompose function. This function allowed us to easily plot the components individually in order to examine them. It also made it extremely easy to create a seasonally adjusted version of our time series and to remove the trend and seasonality to create a stationary version of our original data. In the following two figures, we can see the results of doing additive decomposition on the average rent prices, the average selling prices, and the case rate of COVID-19.

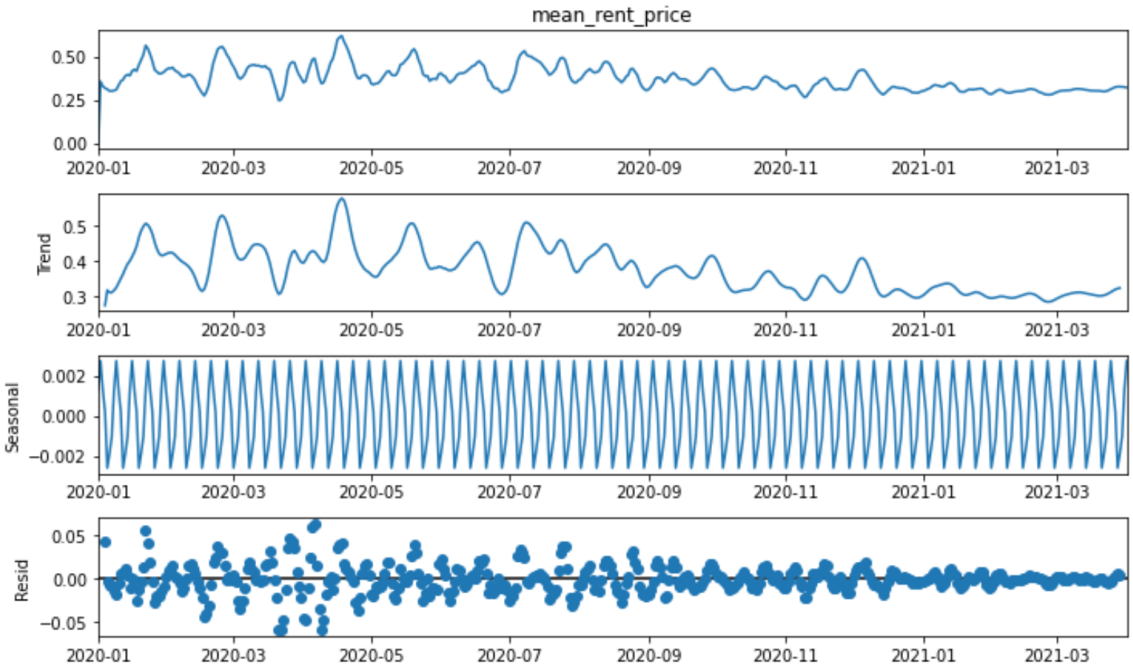


Figure 4: Time series Decomposition of rental prices

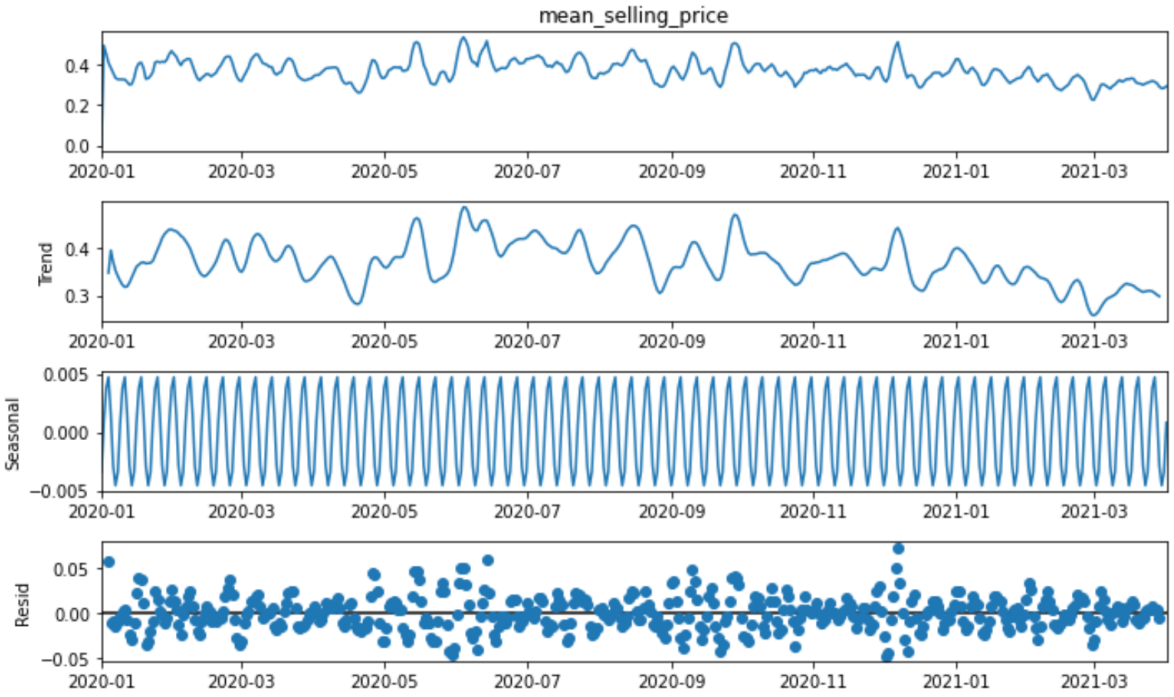


Figure 5: Time series Decomposition of selling prices

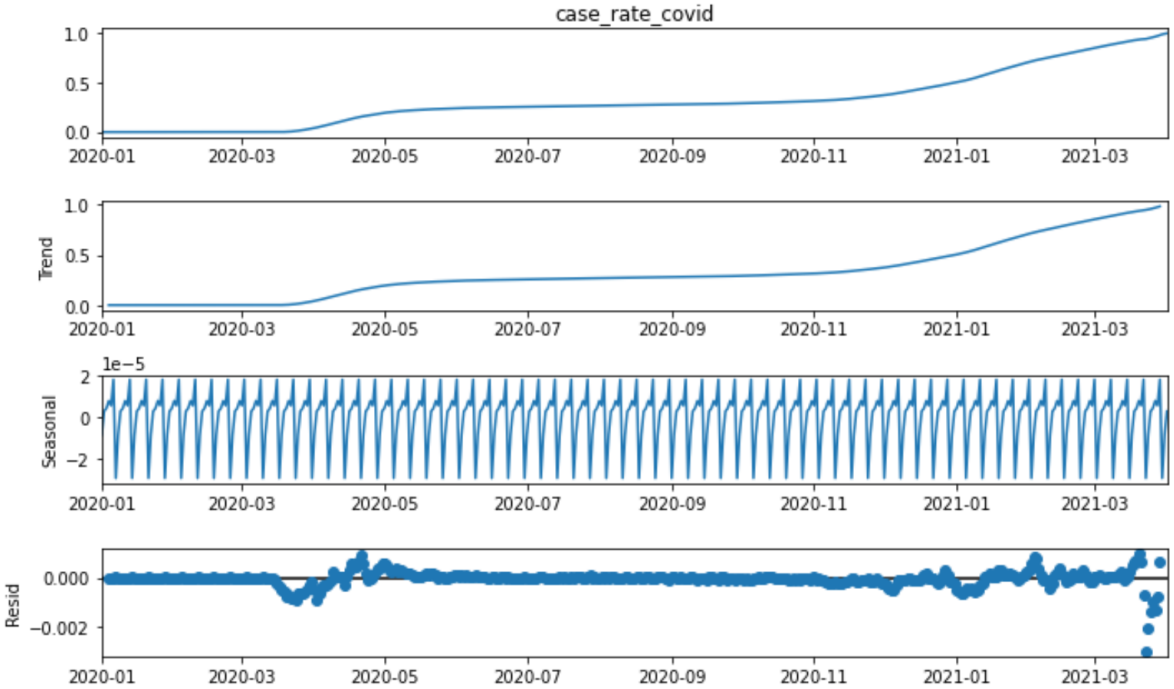


Figure 6: Time series Decomposition of COVID-19 case rates

Autocorrelation (ACF)

The coefficient of correlation between two values in a time series is called the autocorrelation function (ACF). In other words, autocorrelation represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals. It is often important to look at autocorrelation plots to decide which model to use during forecasting, and to identify hidden patterns such as seasonality in our data. Furthermore, it is also used to calculate the parameters of autoregressive models (normally used to determine p and q values), and to create lagged features because it measures the relationship between a variable's current value and its past values. The most important thing to keep in mind when looking at an autocorrelation plot, is that an autocorrelation of +1 represents a perfect positive correlation, while an autocorrelation of negative 1 represents a perfect negative correlation.

To plot the autocorrelation of each feature, we used the autocorrelation_plot function from the Pandas module. According to the documentation, the horizontal lines in the plot correspond to 95% and 99% confidence bands, where the dashed line is 99% confidence band. To get a closer look and to compare the autocorrelation plots to the partial autocorrelation plots, we also used the plot_acf function from the statsmodels module. These two modules use slightly different algorithms to compute the correlation plot, but the results in our case are only marginally different. In the figure below, we can see the autocorrelation plots of the real estate data and some features from the COVID-19 time series.

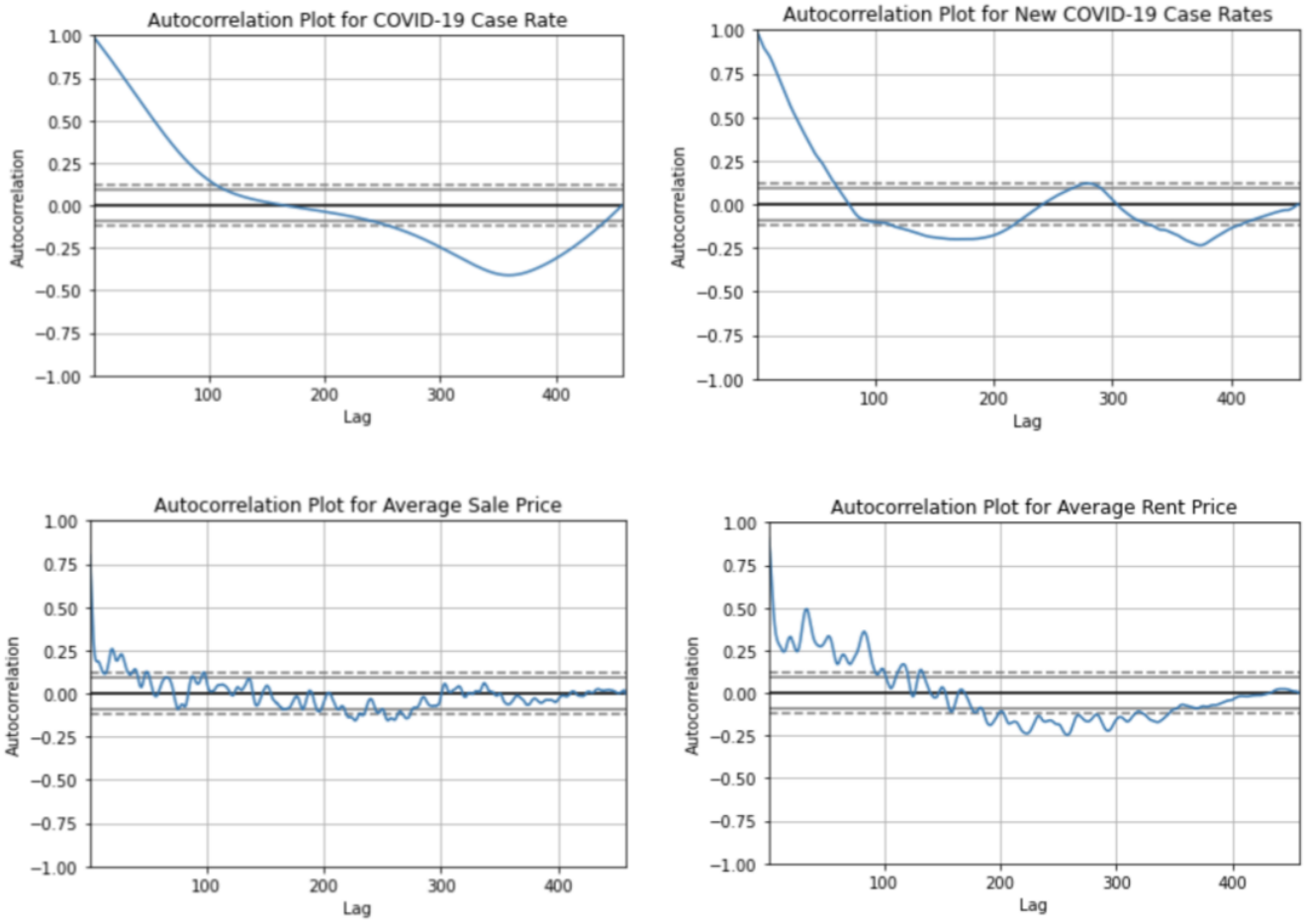
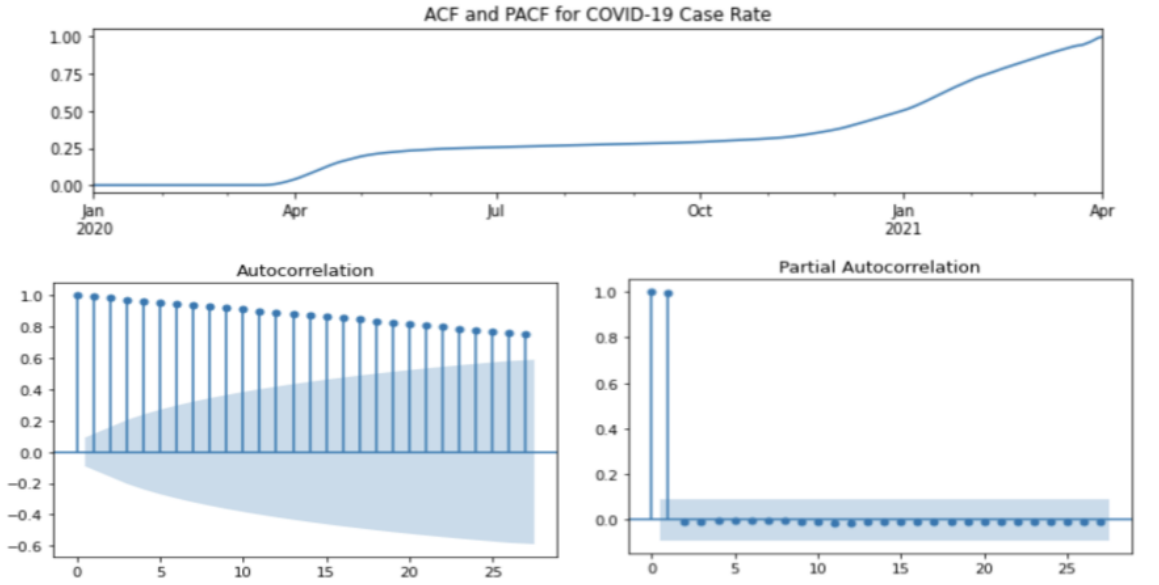


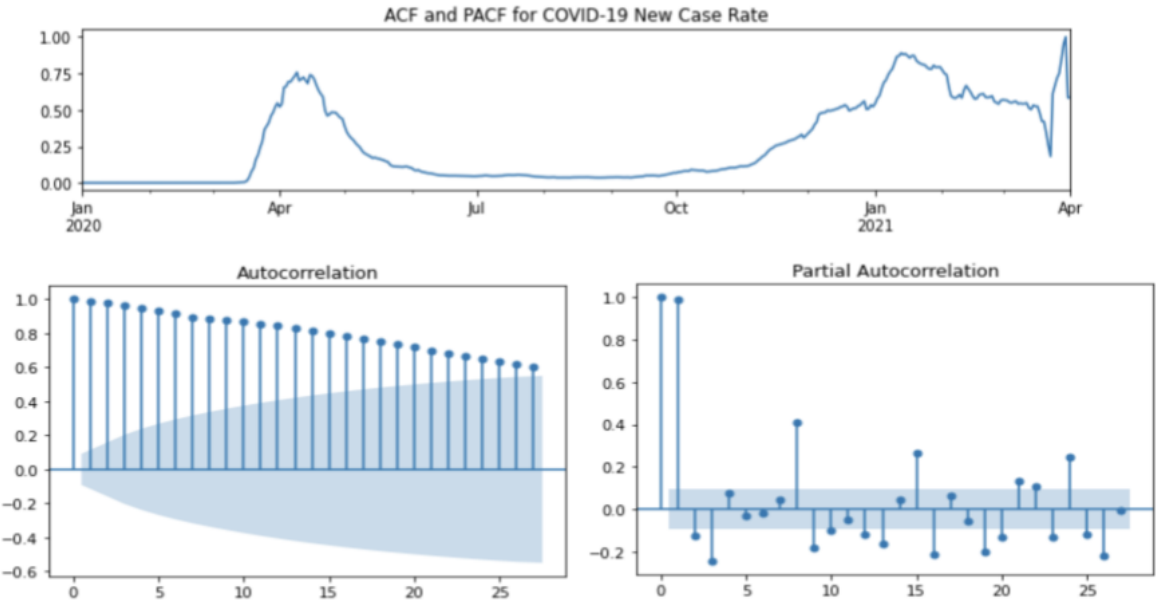
Figure 7: Autocorrelation plots for rental prices and COVID-19 case rates

ACF and The Partial Autocorrelation Function (PACF)

Analyzing the autocorrelation function (ACF) and partial autocorrelation function (PACF) in conjunction is necessary for selecting the appropriate ARIMA model for any time series prediction. A partial autocorrelation is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed. In the book “Introductory Time Series with R”, the authors explain that

the partial autocorrelation at lag k is the correlation that results after removing the effect of any correlations due to the terms at short lags. This means that we can use PACF to determine the value of k for which no other value beyond k has a correlation, or in other words, a cut-off point for the value of k in the ACF plot.





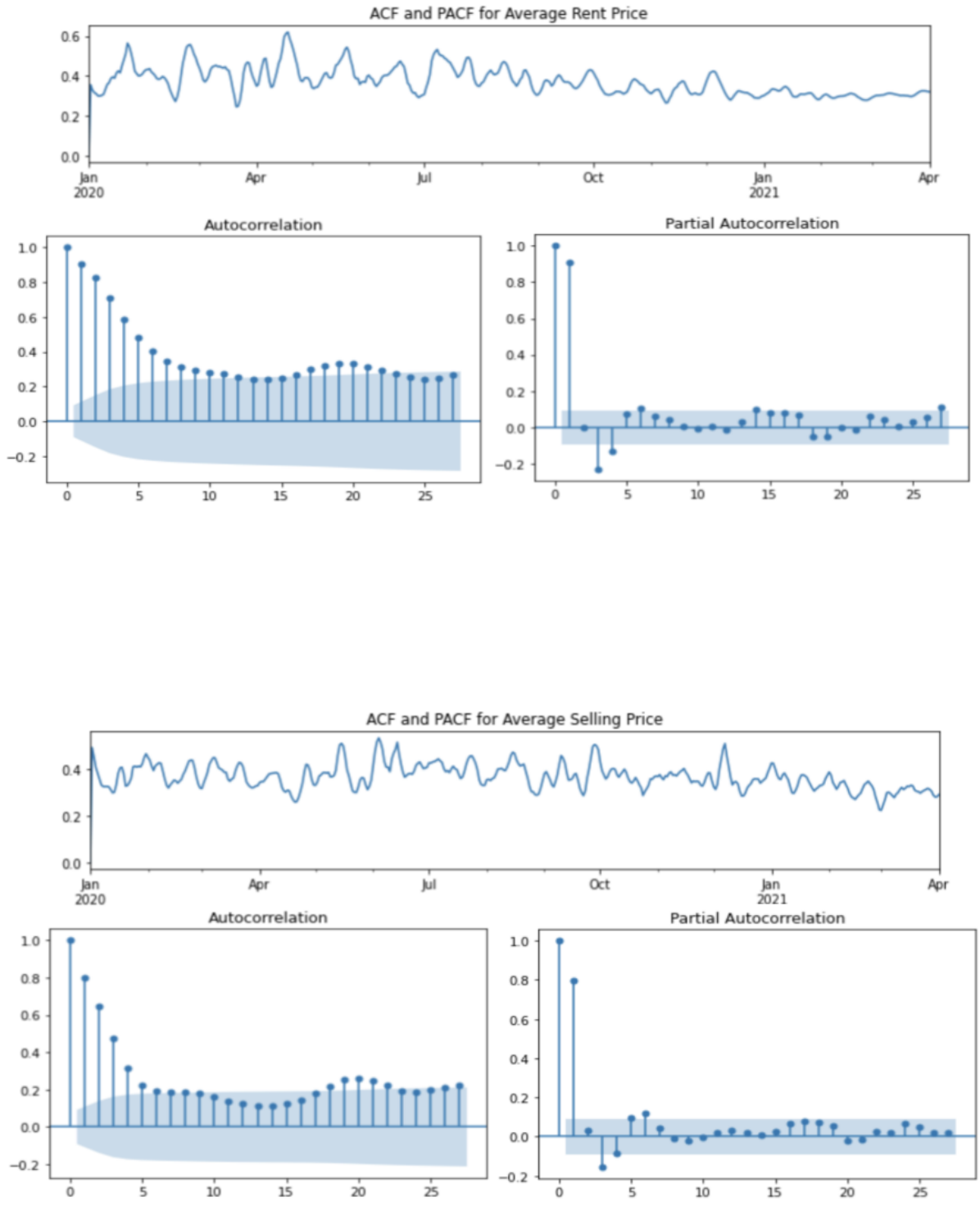


Figure 8: Partial Autocorrelation plots for rental and selling prices as well as COVID-19 case rates

The ACF and PACF plots are often used to select the values of p and q to use in ARIMA models like auto-regressive (AR) and moving average (MA) models. These plots show that there is in fact a seasonal component in the real estate data, but this seasonal component is not present in the COVID-19 time series. We can also see that we should probably use a lag value of 2 for the real estate data and a period of 4 (or close to 4).

Cross Correlation

Cross correlation is similar to autocorrelation, but it uses two time series variables. In this context, we can define cross-correlation as a measure of similarity of between two time series as a function of the displacement of one relative to the other. In Figure 9 and 10, we can see that for values of k larger than 0, the correlation is just slightly higher than when k is 0 so in this case, having a lagged variable is not necessarily needed.

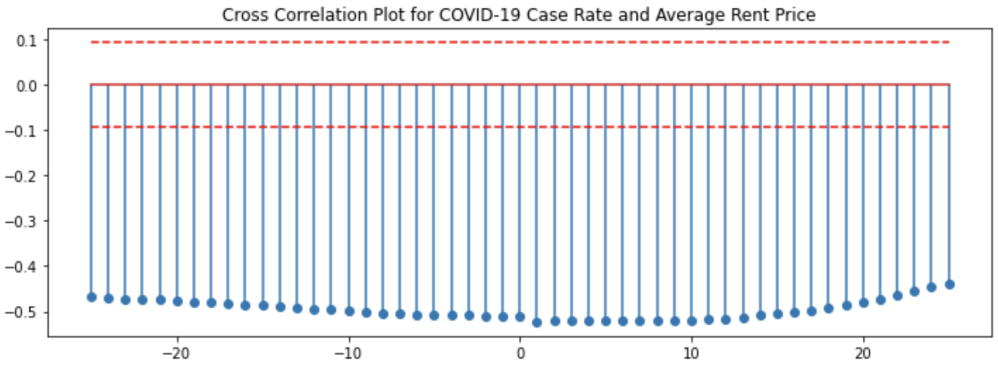


Figure 9: Cross Correlation plots for rental prices and COVID-19 case rates

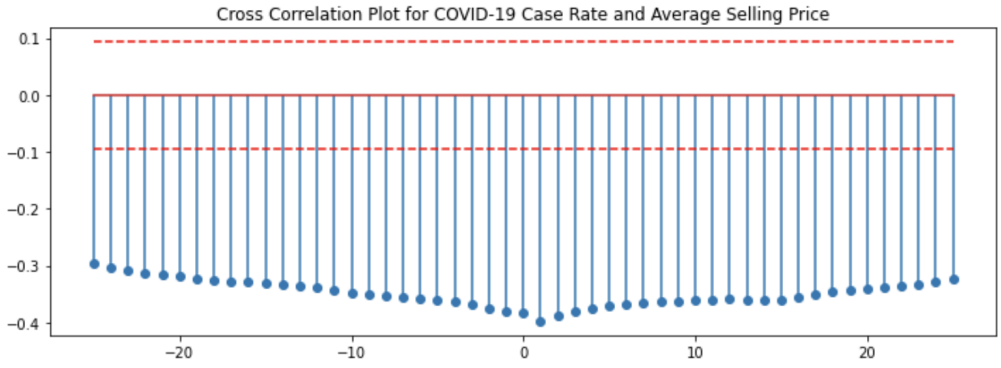


Figure 10: Cross Correlation plots for selling prices and COVID-19 case rates

Lag Plots

Creating a lag plot enables you to check for randomness. Random data will spread fairly evenly both horizontally and vertically. If you cannot see a pattern in the graph, your data is most probably random. On the other hand a shape like in our case or trend to the graph (like a linear pattern) indicates the data is not random.

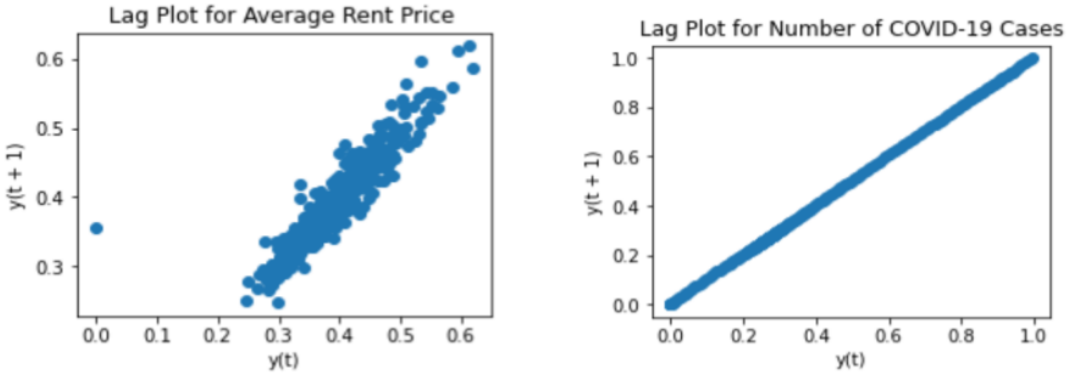


Figure 20: Lag order for rental prices and COVID-19 cases

KDE Plots

Kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample.

The vertical or y-axis of a KDE plot represents the Kernel Density Estimate of the Probability Density Function of a random variable, which is interpreted as a probability differential. The probability of a value being between the points x1 and x2 is the total shaded area of the curve under the two points.

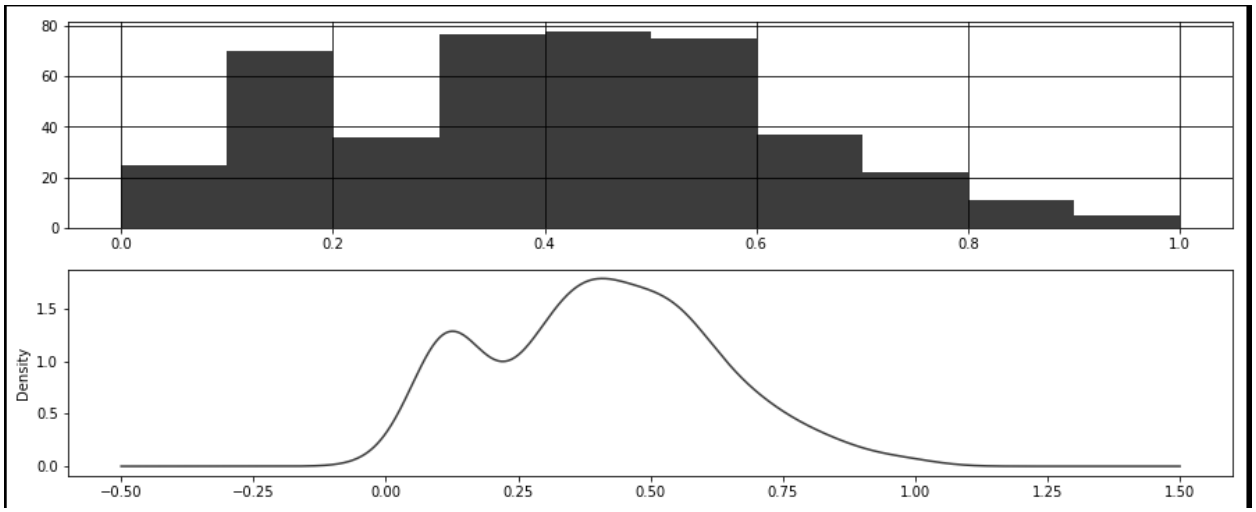


Figure 11: KDE plot for rental prices

In the above graph, we can observe that the peak is just below 0.5. In a histogram, we have found that the concentration of values lies somewhere in the bin of 0.25–0.65. But this density plot gives us a more precise location. It also gives a continuous distribution view. This proves our data is normally distributed around a definite mean and hence we can move forward with the time series analysis and forecasting.

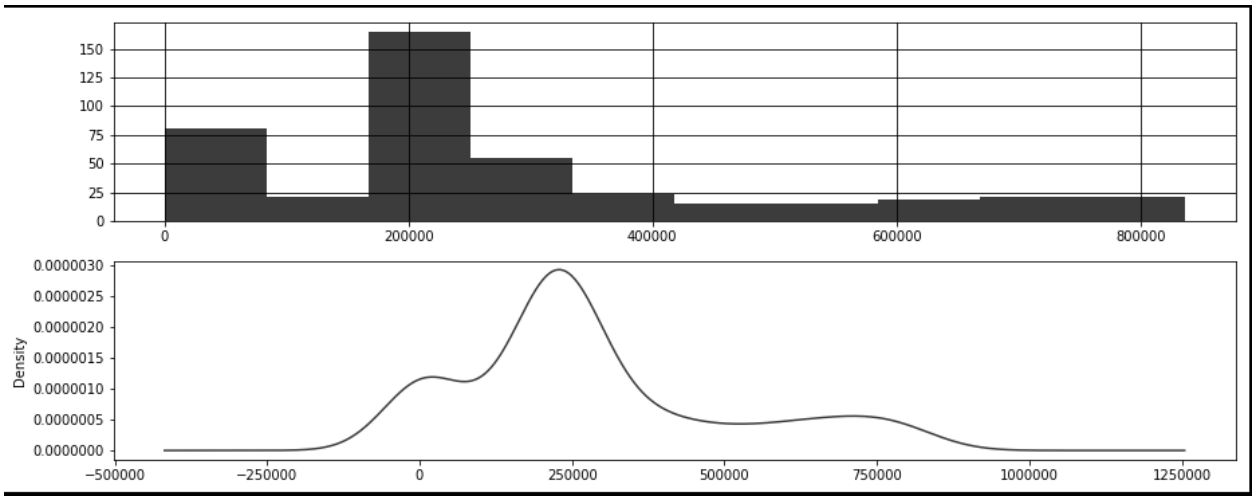


Figure 12: KDE plot for COVID-19 cases

Here we have a peak at the 250000 level, and we have also found that the concentration of values lies somewhere in the bin of 180000–250000 values.

Stationarity Checks

Establishing if a time series is stationary or not is an essential part of preparing the data for the modeling phase. The models used for forecasting a time series often require that the data be stationary before fitting the model. In order to identify which features in our data set are stationary, we applied some test and interpreted the results. First, we will explain how we dealt with series that were not stationary and then we will show the results of the tests before and after we attempt to make the series stationary.

Differencing

One way to make a non-stationary time series stationary — compute the differences between consecutive observations. This is known as differencing.

Transformations such as logarithms can help to stabilize the variance of a time series. Differencing can help stabilize the mean of a time series by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality. In our case the following second order differencing was necessary to achieve stationarity.

$$\begin{aligned} y''_t &= y'_t - y'_{t-1} \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2}. \end{aligned}$$

Y" will have T−2 values. Then, we would model the "change in the changes" of the original data. In practice, it is almost never necessary to go beyond second-order differences.

ADF (Augmented dickey fuller tests)

Time series is different from more traditional classification and regression predictive modeling problems. It has several characteristics like trend, seasonal, residual etc. and highly dependent over time. In simple words, stationary time series data do not depend on time. Time series are stationary if they do not have trend or seasonal effects. Summary statistics calculated on the time series are consistent over time like the mean or the variance of the observation. When a time series is stationary, it can be easier to model.

We can check stationary by two ways. One is manual checks of mean and variance of time series and another way is by using ADF test function. We opted for the latter as the results are more self-evident.

The ADF test is fundamentally a statistical significance test. That means, There is a hypothesis testing involved with a null and alternate hypothesis and as a result a test statistic is computed

and p-values get reported. From the statistic test and the p-values, we can make an inference as to whether a given time series is stationary or not.

A key point to remember here is: Since the null hypothesis assumes the presence of unit root, that is α=1, the p-value obtained should be less than the significance level (say 0.05 or 0.01) in order to reject the null hypothesis. Thereby, inferring that the series is stationary.

The hypotheses for the test are:

- The null hypothesis for this test is that there is a unit root.
- The alternate hypothesis is that the time series is stationary (or trend-stationary).

Like a sanity check we can apply ADF to our data once after preparation and check for unit roots. The figures below show ADF results before and after differencing.

```
Test Statistic                  -1.614051
p-value                          0.475849
No Lags Used                    18.000000
Number of Observations Used    417.000000
Critical Value (1%)             -3.446129
Critical Value (5%)             -2.868496
Critical Value (10%)            -2.570475
dtype: float64
Conclusion:====>
Fail to reject the null hypothesis
```

Figure 13: ADF results for rental prices

Here, we noticed that the statistical test value is greater than the critical value and the p-value is also greater than the significant value(0.05). So we can say the time series is non-stationary.

```
Test Statistic                  1.627808
p-value                         0.997939
No Lags Used                   16.000000
Number of Observations Used    419.000000
Critical Value (1%)             -3.446054
Critical Value (5%)             -2.868463
Critical Value (10%)            -2.570458
dtype: float64
Conclusion:====>
Fail to reject the null hypothesis
```

Figure 14: ADF results for COVID-19 cases

Similarly, for case count too we noticed that the statistical test value is greater than critical value and p-value is also greater than significant value(0.05). So the time series is again non-stationary.

```
Test Statistic                 -3.856129
p-value                         0.002383
No Lags Used                   14.000000
Number of Observations Used    419.000000
Critical Value (1%)             -3.446054
Critical Value (5%)             -2.868463
Critical Value (10%)            -2.570458
dtype: float64
Conclusion:====>
Reject the null hypothesis
Data is stationary
```

Figure 15: ADF results for rental prices (post differencing)

Now after the second difference, we notice that the statistical test and p-value are lower than the critical value and significant value (0.05) respectively. So this is stationary. And we also found that given time series would become stationary on second difference.

```
Test Statistic                  -3.856129
p-value                          0.002383
No Lags Used                    14.000000
Number of Observations Used    419.000000
Critical Value (1%)             -3.446054
Critical Value (5%)             -2.868463
Critical Value (10%)            -2.570458
dtype: float64
Conclusion:====>
Reject the null hypothesis
Data is stationary
```

Figure 16: ADF results for COVID-19 cases (post differencing)

Similarly, we notice that the test statistic and p-value are lower than the critical value and significant value (0.05) respectively. Hence, data is now stationary.

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test

In our analysis, we use the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Kwiatkowski, Phillips, Schmidt, & Shin, 1992). In this test, the null hypothesis is that the data are stationary, and we look for evidence that the null hypothesis is false. Consequently, small p-values (e.g., less than 0.05) suggest that differencing is required.

```
KPSS Statistic: 1.219741529479865
p-value: 0.01
num lags: 18
Critial Values:
   10% : 0.347
   5% : 0.463
   2.5% : 0.574
   1% : 0.739
Result: The series is not stationary
KPSS Statistic: 2.025603182610288
p-value: 0.01
num lags: 18
Critial Values:
   10% : 0.347
   5% : 0.463
   2.5% : 0.574
   1% : 0.739
Result: The series is not stationary
```

Figure 17: KPSS results for rental prices and COVID-19 cases

The test statistic is much bigger than the 1% critical value, indicating that the null hypothesis is rejected. That is, the data is not stationary, reinforcing the need for differencing. This is in line with our ADF test results.

```
KPSS Statistic: 0.02376221234137097
p-value: 0.1
num lags: 18
Critial Values:
   10% : 0.347
   5% : 0.463
   2.5% : 0.574
   1% : 0.739
Result: The series is stationary
KPSS Statistic: 0.06633185645150017
p-value: 0.1
num lags: 18
Critial Values:
   10% : 0.347
   5% : 0.463
   2.5% : 0.574
   1% : 0.739
Result: The series is stationary
```

Figure 18: KPSS results for rental prices and COVID-19 cases (post differencing)

This time, the test statistic is less than 0.05, and well within the range we would expect for stationary data. So we can conclude that the differenced data are stationary.

Pearson Correlation

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

Metrics for evaluation

Strength: The correlation coefficient can range in value from −1 to +1. The larger the absolute value of the coefficient, the stronger the relationship between the variables.

For the Pearson correlation, an absolute value of 1 indicates a perfect linear relationship. A correlation close to 0 indicates no linear relationship between the variables.

Direction: The sign of the coefficient indicates the direction of the relationship. If both variables tend to increase or decrease together, the coefficient is positive, and the line that represents the correlation slopes upward. If one variable tends to increase as the other decreases, the coefficient is negative, and the line that represents the correlation slopes downward.

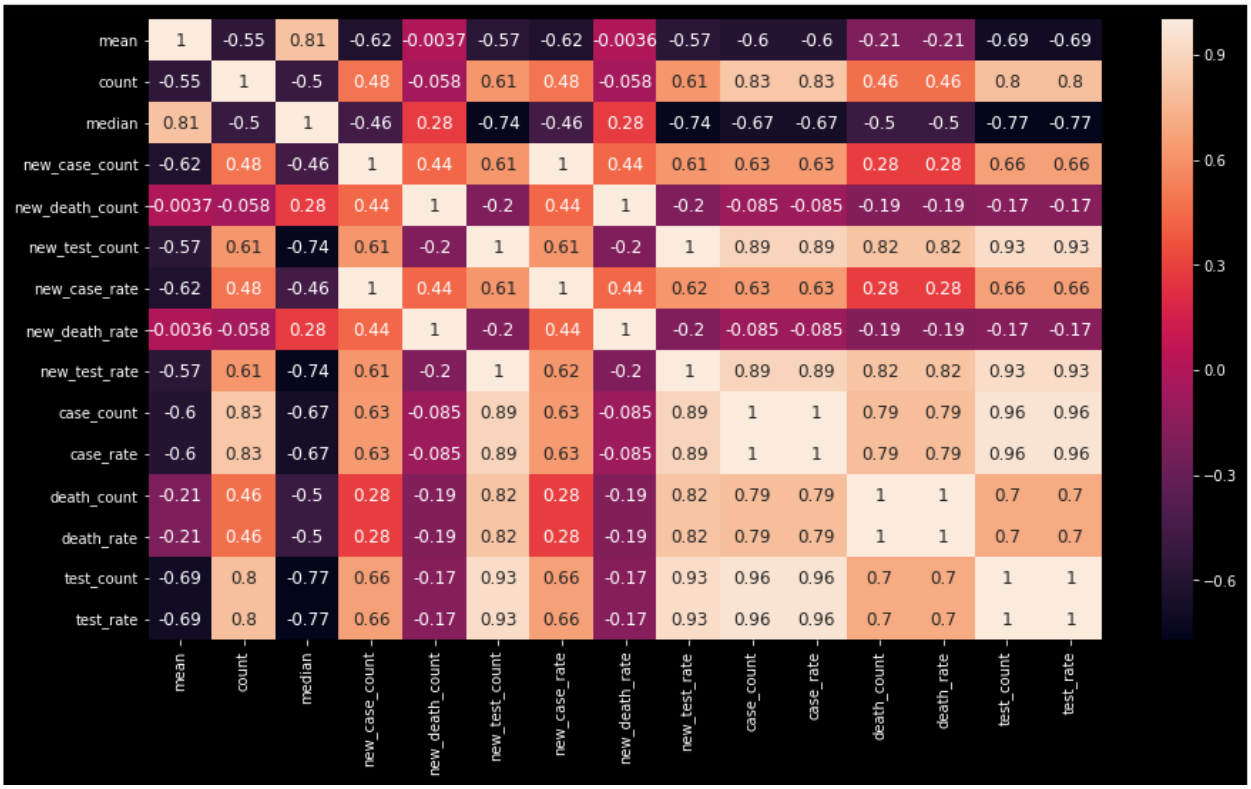


Figure 19: Pearson Correlation for all attributes

From the calculated plot we can make a few key observations. Our variables of interest “case_count” and “mean” (depicts mean rent price) have a correlation coefficient of -0.6 indicating there is a negative correlation between them as expected. The value significance is moderate in comparison but still significant enough to check for further analysis.

Data Modelling

ARIMA

In the modeling phase of our project, we decided to begin with a univariate analysis of the features that we have collected in order to learn more about forecasting and to get a better understanding of our data. Any ‘non-seasonal’ time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models. ARIMA, short for ‘Auto Regressive Integrated Moving Average’ is one of the most wildly used approaches for time series forecasting. Instead of just being a model, ARIMA is actually a class of models that ‘explains’ a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that an equation can be used to forecast future values. An ARIMA model is characterized by 3 terms: p, d, q where, p is the order of the AR term, q is the order of the MA term and d is the number of differencing required to make the time series stationary.

To model the rate of COVID-19 cases, we used the auto_arima model from the library pmdarima. This function allows us to optimize the parameters of ARIMA to minimize the AIC. In the case of selling price, the optimum parameters for the ARIMA model seems to be ARIMA(1,1,0)(10,1,0)[4] which results in the following forecast:

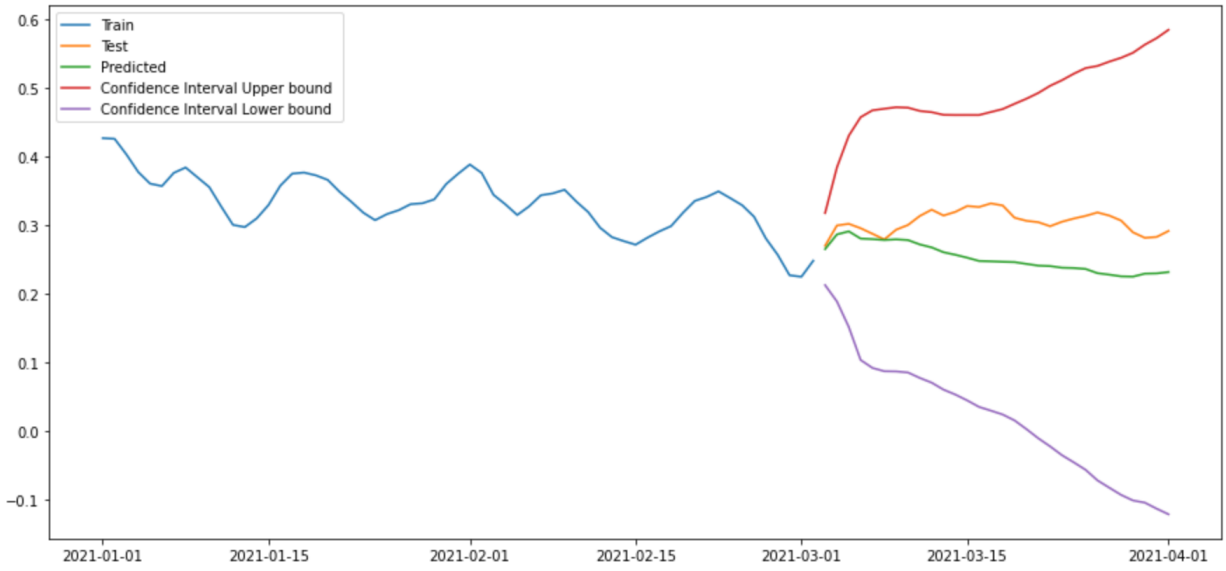


Figure 21: ARIMA plots for COVID-19 cases

In Figure 21 we can see that the horizon for the forecast is 30 days, but it seems that the forecast is only accurate for a shorter period of time.

To model the movement of the mean rental prices in NYC, we used the auto_arima model again which in this case, the optimum parameters seems to ARIMA(1,1,0)(10,1,0)[4] which results in the forecast shown in Figure 22. As in the previous case, the forecast seems to only work in the short term. It is also important to note that the confidence interval in both cases is extremely large which means that the model might not be completely reliable.

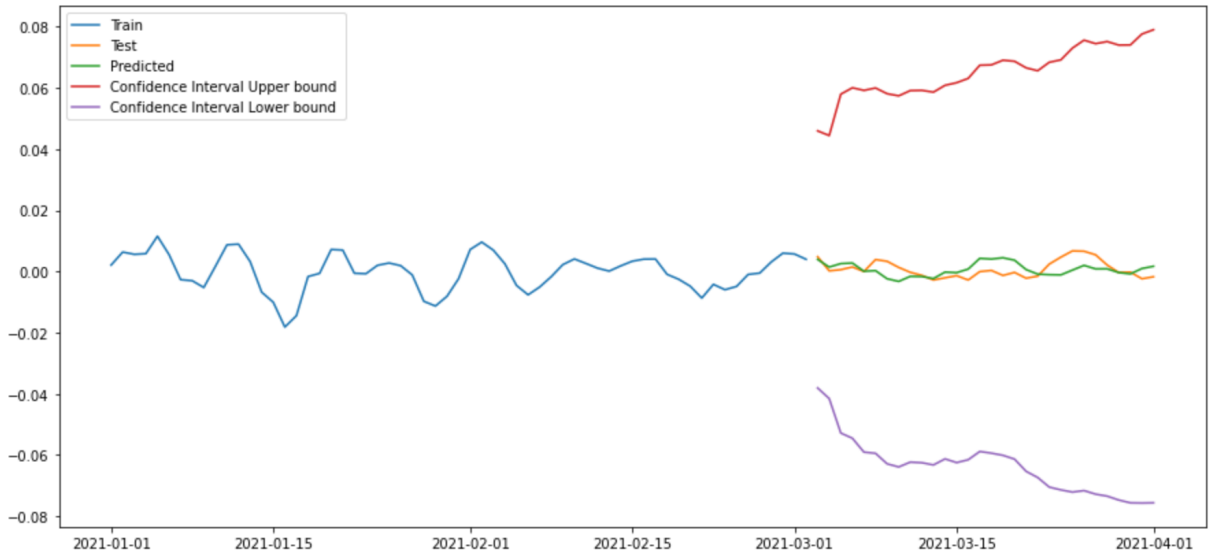


Figure 22: ARIMA plots for rental prices

Granger's Causality

Granger causality is a statistical test for determining whether one time series can forecast another.

According to Granger causality, if a X1 "Granger-causes" (or "G-causes") a X2, then past values of X1 should contain information that helps predict X2 above and beyond the information contained in past values of X2 alone.

	mean_x	case_count_x
mean_y	1.0000	0.0065
case_count_y	0.0017	1.0000

Figure 23: Granger's causality results for rental prices (mean) and COVID-19 cases (case_count)

The p-values of our statistic in Granger causation is significantly lesser than the standard p value

of 0.05. This shows that these variables can be “causally correlated”. This is a key indicator of causation and thus with correlation and causation proved we now try to model the data to predict values both in a multivariate fashion using VAR.

Durbin Watson Statistic

We can check for Serial Correlation of Residuals (Errors) using Durbin Watson Statistic. The value of this statistic can vary between 0 and 4. The closer it is to the value 2, then there is no significant serial correlation. The closer to 0, there is a positive serial correlation, and the closer it is to 4 implies negative serial correlation.

According to our calculations, Durbin Watson Statistic for mean rent is 1.8 and for case_count is 2.12.

Our data is closer to that of the expected value of 2, and we can now be sure that the residual errors are not being carried forward by our model during predictions.

Johansen Cointegration test

A cointegration test is the co-movement among underlying variables over the long run. This long-run estimation feature distinguishes it from correlation. Two or more variables are cointegrated if and only if they share common trends.

But COVID-19 is a recent phenomenon and the data available is limited only for the past year (2020)

```
Column Name > Test Stat > C(95%) => Signif
 ----------------------------------------
mean    >  113.73    > 12.3212   =>  True
case_count >  46.45     > 4.1296    =>  True
```

Figure 24: Johansen Cointegration results for rental prices (mean) and COVID-19 cases (case_count)

Our results above indicate that mean rent can be significantly predicted whereas case counts cannot.

VAR

VAR models (vector autoregressive models) are used for multivariate time series. The structure is that each variable is a linear function of past lags of itself and past lags of the other variables.

VAR(p) model:

$$Y_t = a + A_1 Y_{t-1} + A_2 Y_{t-2} + \ldots + A_p Y_{t-p} + \varepsilon_t$$

where:

- $Y_t = (y_{1t}, y_{2t}, \ldots, y_{nt})'$: an (nx1) vector of time series variables
- a: an (nx1) vector of intercepts
- A_i (i=1, 2, …, p): (nxn) coefficient matrices
- ε_t: an (nx1) vector of unobservable i.i.d. zero mean error term (white noise)

Our data is split such that 70% of data goes into the train dataset and 30% goes into test dataset. Also in a time series we preserve the order and not randomize the dataset.

Hyperparameter optimization: Lag order

The most common approach for lag order selection is to inspect among different information criteria and choose the model that minimizes these indicators. There are several Information Criterion alternatives such as AIC (Akaike information criterion), SIC (Schwarz information criterion, aka Bayesian information criterion BIC), HQIC (Hannan-Quinn information criterion), Akaike's Final Prediction Error etc. and they vary on the weight they put on prediction error and parameters.

In our case we opted to choose the lag order using the lowest AIC values.

Why AIC?

The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

A lag order of 9 provided us with the lowest AIC values for our data.

At Lag Order = 9 the following values were observed:

AIC : -27.0121

BIC : -26.6300

FPE : 1.857185273229239e-12

HQIC: -26.8607

Modelling results and interpretation

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Tue, 04, May, 2021
Time:                      13:39:19
----------------------------------------------------------------------
No. of Equations:         2.00000    BIC:                   -26.6301
Nobs:                     396.000    HQIC:                  -26.8607
Log likelihood:           4262.60    FPE:                1.85719e-12
AIC:                     -27.0121    Det(Omega_mle):     1.69102e-12
----------------------------------------------------------------------

Forecast accuracy of Rent
RMSE:  0.01
MAE:  0.11
```

Figure 25: VAR results for rental prices

The forecast accuracy of mean rent shows an RMSE of 0.01 which is considerably low and shows that rental price prediction is a definite possibility with both past and present values of rental prices along with past case count history of COVID-19. Our data shows that the impact of COVID-19 is far significant and has indeed adversely affected the rental prices.

```
Forecast accuracy of CASE_COUNT
RMSE:  0.85
MAE:  0.92
```

Figure 26: VAR results for COVID-19 cases

The forecast accuracy of cases is worse in comparison which lines up with the idea that with the intervention of tests and vaccinations it is no longer a predictable variable and thus is in turn dependent on other factors.

Correlation matrix of residuals

	mean	case_count
mean	1.000000	-0.042632
case_count	-0.042632	1.000000

Figure 27: Correlation matrix of residual errors

The correlation matrix of residuals shows that almost no past errors/low error is being carried forward by our model which shows our model is fairly reasonable in its results.

Evaluation

Key findings in Granger's causality helps us evaluate that increase in COVID-19 cases leads to a fall in rental prices and vice versa. But stating COVID-19 cases increased due to fall in rental prices is an overstretch which is simply not true. Additional evidence indicates that the housing rental prices were already in turmoil and was decreasing even before COVID-19 started. The situation was only worsened due to COVID-19.

It is also important to note that while these conclusions hold true for NYC as a whole, a further investigation into borough wide data would help us explore which of these neighborhoods within NYC were more adversely affected. Our ARIMA model provides a univariate analysis of rental prices and successfully predicts Rental prices in the short term, but it cannot seem to hold the same accuracy for long term forecasting. On the other hand, our VAR model with its multivariate analysis of both Rental prices and COVID-19 cases has even lesser RMSE and thus more explainability on predicting data.

Conclusion

We can safely predict the average rent prices in New York in tandem with the number of COVID-19 cases in the city. Although our ARIMA models gave us a large confidence interval and the predictions are not accurate in the long term, it seems to give a very good forecast in the short term. However, going forward we have new variables such as vaccination rates which can interfere with the model. VAR allows us to even evaluate the relationship between all 3 when enough data is available for vaccination. This could be the goal of future iterations of the project, such as evaluating the impact of the vaccine, on the ease of restrictions due to COVID-19 etc.

Also, other models such as VECM or bleeding edge techniques like LSTMs could also be the way to go, and thus more variables can be brought into the picture.

Alternatively we are also expanding upon the predictions borough wise for New York which would give us more granular insights into the situation in each borough. We started to work on a more modular way of analyzing our data in order to easily compare the results from different boroughs, but we decided to focus on NYC as a whole in this paper due to the time restrictions of this course. This will help governments or related entities to identify key problematic areas and analyze the reasons behind the same. A more targeted approach would definitely help us combat the COVID-19 crisis better which sheds new light on Real estate situation in NYC.

References

Del Giudice, V., De Paola, P., & Del Giudice, F. P. (2020). "*COVID-19 infects real Estate markets: Short And Mid-Run effects on housing prices in Campania Region (Italy)*". *Social Sciences, 9*(7), 114.

Ling, Wang, Zhou., (2020). "*A First Look at the Impact of COVID-19 on Commercial Real Estate Prices: Asset-Level*". *The Review of Asset Pricing Studies*, Volume 10, Issue 4, December 2020, Pages 669–704.

Blakeley, Grace., (January 2021). Financialization, Real Estate and COVID-19 in the UK. Community Development Journal Vol. 56, Iss. 1, 79-99.

Brownlee, J. (2020, December 9). *How to Decompose Time Series Data into Trend and Seasonality*. Machine Learning Mastery. https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/.

Elementary Statistics for the rest of us! Statistics How To. (2020, December 18). https://www.statisticshowto.com/.

Appendix

Detailed equations for VAR modelling analysis consists of components such as coefficient, standard error, t-statistics and probabilities for each iteration of 9 lags of the model.

Results for equation mean

	coefficient	std. error	t-stat	prob
const	-0.000045	0.000803	-0.056	0.955
L1.mean	0.967966	0.043443	22.281	0.000
L1.case_count	-5.181653	4.956486	-1.045	0.296
L2.mean	-0.070850	0.057263	-1.237	0.216
L2.case_count	4.021303	8.199537	0.490	0.624
L3.mean	0.003125	0.057288	0.055	0.956
L3.case_count	-0.071264	7.898112	-0.009	0.993
L4.mean	-0.021027	0.056851	-0.370	0.711
L4.case_count	11.571922	7.863471	1.472	0.141
L5.mean	-0.014146	0.056935	-0.248	0.804
L5.case_count	-1.298983	7.826636	-0.166	0.868
L6.mean	0.028097	0.056899	0.494	0.621
L6.case_count	-18.417837	7.905096	-2.330	0.020
L7.mean	0.003356	0.056866	0.059	0.953
L7.case_count	13.266141	8.081841	1.641	0.101
L8.mean	-0.632071	0.056334	-11.220	0.000
L8.case_count	-14.390559	8.425666	-1.708	0.088
L9.mean	0.533365	0.042568	12.530	0.000
L9.case_count	10.526386	5.066653	2.078	0.038

Results for equation case_count

	coefficient	std. error	t-stat	prob
const	0.000015	0.000008	1.904	0.057
L1.mean	-0.000180	0.000428	-0.420	0.674
L1.case_count	1.461899	0.048850	29.926	0.000
L2.mean	0.000656	0.000564	1.162	0.245
L2.case_count	-0.522169	0.080813	-6.461	0.000
L3.mean	-0.001408	0.000565	-2.494	0.013
L3.case_count	0.254461	0.077842	3.269	0.001
L4.mean	0.000687	0.000560	1.226	0.220
L4.case_count	-0.084218	0.077501	-1.087	0.277
L5.mean	-0.000215	0.000561	-0.384	0.701
L5.case_count	-0.213434	0.077138	-2.767	0.006
L6.mean	0.000554	0.000561	0.988	0.323
L6.case_count	0.346831	0.077911	4.452	0.000
L7.mean	-0.000266	0.000560	-0.475	0.635
L7.case_count	-0.702033	0.079653	-8.814	0.000
L8.mean	0.000458	0.000555	0.825	0.409
L8.case_count	0.783270	0.083042	9.432	0.000
L9.mean	-0.000944	0.000420	-2.250	0.024
L9.case_count	-0.330135	0.049936	-6.611	0.000