**RMDS Q2 2021 Data Science Competition**

California House Price Prediction in Post- Covid Period Analysis

**Muhammad Syamil Iklil Abdul Barr Yogi Anggara**

**Ergidya Liviana Riyanna Shabrina**

**Team Zero**

# Introduction

Covid-19 is an ongoing pandemic of coronavirus disease started since late 2019. The disease is a severe acute respiratory syndrome (thus named SARS-CoV 2). The outbreak first identified in Wuhan Province, China in December 2019. This pandemic halted many aspects of social and economic activity since there are restrictions on physical activity. After a quarter, the pandemic began to show its severity of effects. The disruption on economic even caused global recession with most country have negative economic growth.

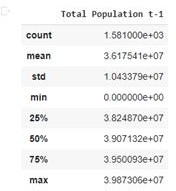
As the economic activity reduced, price level also affected, including in property sector. In most cases, deflation happened in the first and/or second quarter of 2020 due to sharp decrease in demand of goods and services. Despite that, with the positive or increasing in trend of property prices, especially in dense and high-demanded places like California, the pandemic effect could be eliminated or at least not as powerful compared to other aspects. Regarding this fact, authors did a research on how pandemic affects house prices in California using house price index as proxy target variable.

# Variables and Dataset Sources

|  |  |
| --- | --- |
| **Variable** | **Data Source and Description** |
| House Price Index | Web-based presentations of California criminal justice statistical data. Sourced from [https://openjustice.doj.ca.gov/exploration/crime-](https://openjustice.doj.ca.gov/exploration/crime-statistics) [statistics](https://openjustice.doj.ca.gov/exploration/crime-statistics) in a CSV format |
| US Mortgage Rate | The average interest rate on a 30-year fixed-rate mortgage in the United States. Sourced from <https://fred.stlouisfed.org/series/MORTGAGE30US> in a CSV format |
| Covid-19 Infection Rate | Covid-19 infection rate data gathered by LA Times. The data is real-time data stored in [https://github.com/datadesk/california-coronavirus-](https://github.com/datadesk/california-coronavirus-data) [data](https://github.com/datadesk/california-coronavirus-data) in a CSV format |
| Median Household Income | Median household income refers to the income level earned by a given household where half of the households in the geographic area of interest earn more and half earn less. Sourced from [California](https://fred.stlouisfed.org/series/MORTGAGE30US) |

|  |  |
| --- | --- |
|  | [State Household Income | Department of Numbers](https://www.deptofnumbers.com/income/california/) [(deptofnumbers.com)](https://www.deptofnumbers.com/income/california/) |
| Median Family Income | The median family income is a measure of family's ability to meet food, clothing, housing, health costs care, transportation, child care, and higher education Sourced from [California State Household Income |](https://fred.stlouisfed.org/series/MORTGAGE30US) [Department of Numbers (deptofnumbers.com)](https://www.deptofnumbers.com/income/california/) |
| Consumer Confidence Index | This consumer confidence indicator provides an indication of future developments of households’ consumption and saving, based upon answers regarding their expected financial situation, their sentiment about the general economic situation, unemployment and capability of savings. An indicator above 100 signals a boost in the consumers’ confidence towards the future economic situation, as a consequence of which they are less prone to save, and more inclined to spend money on major purchases in the next 12 months. Values below 100 indicate a pessimistic attitude towards future developments in the economy, possibly resulting in a tendency to save more and consume less. Sourced from OECD and in a CSV format |
| Consumer Price Index | This variable shows CPI data from 2017 until 2021. Sourced from [Office of the Director - Research Unit:](https://www.dir.ca.gov/OPRL/capriceindex.htm) [California Consumer Price Index](https://www.dir.ca.gov/OPRL/capriceindex.htm) in CSV format |
| Crimes | Crime dataset in California cities. The data is aggregated to create crime in California. Sourced from [https://ucr.fbi.gov/crime-in-the- u.s/](https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/tables/table-8/table-8-state-cuts/california.xls) in a CSV format |
| Population | The total number of persons inhabiting a country, city, or any district or area. Sourced from [Search](https://www.census.gov/content/census/en/search-results.html?stateGeo=none&q=california%20population%20ca&searchtype=web&page=1) [Results](https://www.census.gov/content/census/en/search-results.html?stateGeo=none&q=california%20population%20ca&searchtype=web&page=1) in a CSV format |

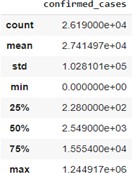
1. Explanatory Data and Variable Analysis
   1. Data description and basic statistical analysis
      * Population



Interpretation:

By using .describe() function, we get the value of mean (3.617541e+07), standard deviation (1.043379e+07), minimum (0.00), and maximum (3.987306e+07) of Population data from 2017 to 2021 in April.

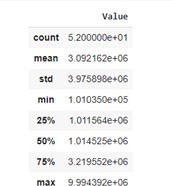
* + - Covid-19 Infection Rate



Interpretation:

Based on the output, we can conclude that the value of mean is 2.741497e+04, standard deviation is 1.028101e+05, the minimum is 0.00 and the maximum is 1.244917e+06.

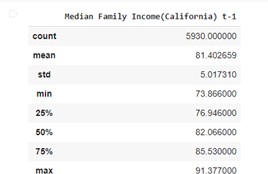
* + - CCI (Consumer Confidence Index) Data



Interpretation:

Based on the output after using .describe() function, we can conclude that the value of mean is 3.092162e+06, standard deviation is 3.975898e+06, the minimum is 1.010350e+05 and the maximum is 9.994392e+06.

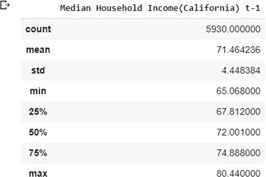
* + - Median Family Income



Interpretation:

By using .describe() function, we get the value of mean (81.402659), standard deviation (5.017310), minimum (73.866000), and maximum (91.377000) of the Median Family Income data.

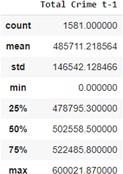
* + - Median Household Income



Interpretation:

Based on the output after using .describe() function, we get the value of mean (71.464236), standard deviation (4.448384), minimum (65.068000), and maximum (80.440000) of the Median Household Income data.

* + - Crimes



Interpretation:

By using .describe() function, in the output we get the value of mean (485711.218564), standard deviation (146542.128466), minimum (0.00), and

maximum (600021.870000).

* + - House Price Index



Interpretation:

Based on the output, we can conclude that the value of the mean is 648.835294, the standard deviation is 37.389765, the minimum is 581.120000, and the maximum is 718.340000.

1. ANOVA and Tukey HSD
   * ANOVA CASTHPI Pingouin Test

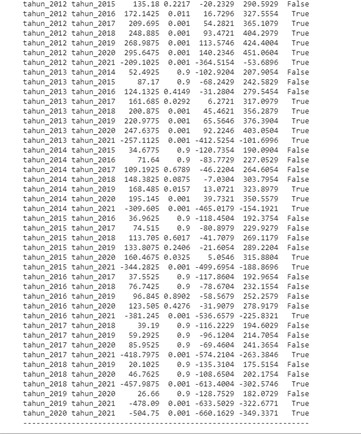
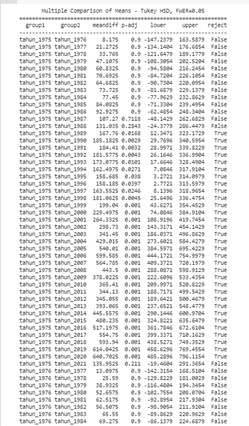
OLS Method



Interpretation:

Because the p-value is less than alpha, the conclusion is to reject H0 so that it can be concluded that there is a difference for the average impact of the several years of CASTHPI.

Tukey Test

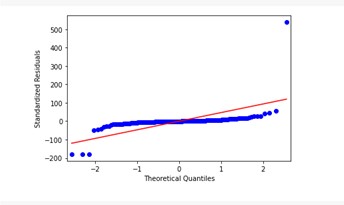


Interpretation:

The use of Tukey HSD to examine differences between groups showed that there was a statistically significant difference. We can see that for the most part, 2015-2021, 2016-2021, 2017-2021, 2018-2021, 2019-2021, and 2021-2021 indicate that the

average year for CASTHPI will always be different in 2021.

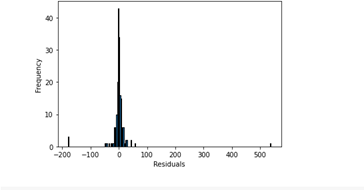
QQ Plot



Interpretation:

Because these points are located around the linear line or follow the diagonal line, it can be concluded that the residual value is normally distributed

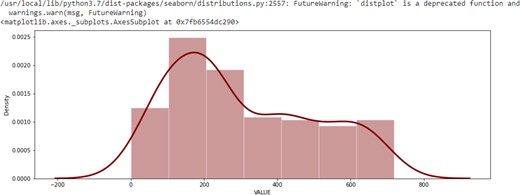
Histogram



Interpretation:

In the histogram graph, it can be seen that it does not form a perfect bell-shape, so it can be concluded that the data is not normally distributed

Check the normality of test data



Interpretation:

From the density plot above, it can be seen that the data does not form a bell-shaped perfectly, so it is possible that the assumption is that the test data here are not normally distributed. To be sure it should be further tested using the Kolmogorov Smirnov Uji Test

Kolmogorov Smirnov



Interpretation:

From the KS test, it can be concluded that the residual data is not normally distributed or rejects H0 because the p-value (0.000) is smaller than alpha (0.05).

Levene Test



Interpretation:

To check whether the sample from the population is homogeneous or not, the Levene test is used because the data is not normally distributed, and the conclusion is that it fails to reject H0 or the sample from the population is homogeneous.

* + - ANOVA Population Pingouin



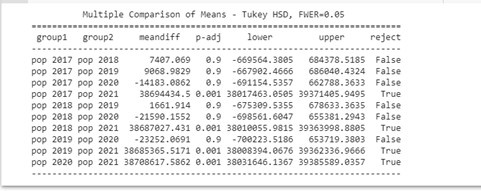
OLS method



Interpretation:

Because the p-value is less than alpha (0,05), the conclusion is that H0 is rejected so that it can be concluded that there is a difference for the population average per year.

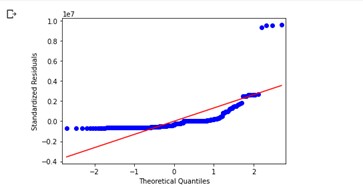
Tukey test



Interpretation:

The use of Tukey HSD to examine differences between groups showed that there was a statistically significant difference. We can see for the most part that for the population in 2017-2021, the population in 2018-2021, the population in 2019-2021, the population in 2020-2021 indicates that the population average is always different in 2021.

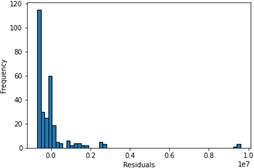
QQ Plot



Interpretation:

Because most of these points are not located around the linear line, it can be concluded that the residual value is not normally distributed

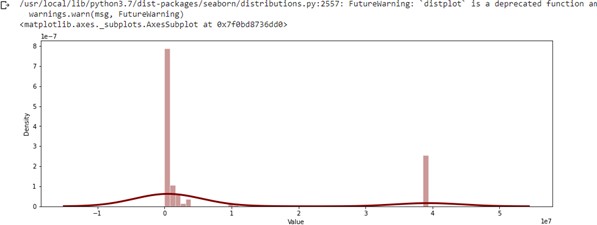
Histogram



Interpretation:

In the histogram graph, it can be seen that it does not form a perfect bell-shape, so it can be concluded that the data is not normally distributed

Check the normality of test data



Interpretation:

From the density plot above, it can be seen that the data does not form a bell-shaped perfectly, so it is possible that the assumption is that the test data here are not normally distributed. To be sure it should be further tested using the Kolmogorov Smirnov Uji Test

Kolmogorov Smirnov



Interpretation:

From the KS test, it can be concluded that the residual data is not normally distributed or rejects H0 because the p-value (0.000) is smaller than alpha (0.05).

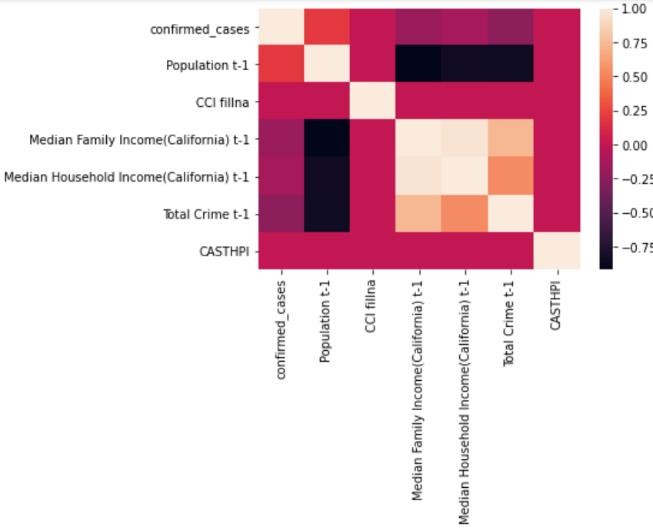
Levene Test



Interpretation:

To check whether the sample from the population is homogeneous or not, the Levene test is used because the data is not normally distributed, and it is concluded to reject H0 or the sample from the population is heterogeneous.

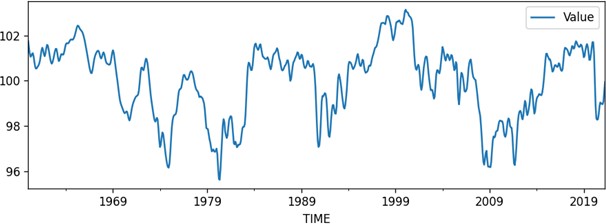
1. Matrix correlation

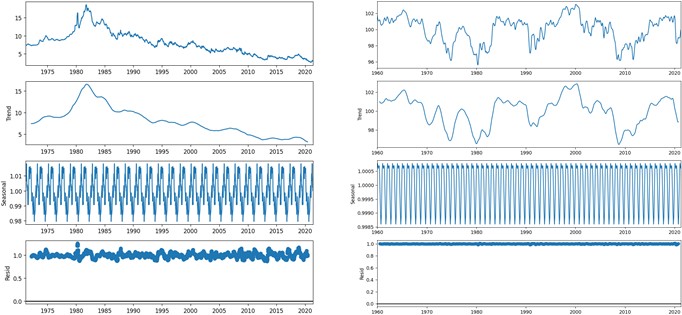


# Time Series Decomposition

Before we go to the time-series forecast, we must understand that there are variables-based on literature review- that significantly affect house price, here we use HPI as the proxy of house price. Those variables mentioned often in prior literature are consumer confidence index as the proxy of consumer expectation and mortgage rate as factor that affect directly to houses’ feasibility to be bought.

Here, authors decompose those two variables into time series decomposition to give insight and intuition how two move and may affect house price index





Based on graph above, we can clearly see that there is season and trend in the data that can be decomposed with low residual level. However, as both season and trend differ, we need further analysis including other variables including house price index and Covid-19 spread itself as our main topic in this study.

# Time Series Analysis

* 1. ARIMA

An [ARIMA model](https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average) is a class of statistical models for analyzing and forecasting time series data. In statistics and econometrics, and in particular in time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.

ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a generalization of the simpler AutoRegressive Moving Average and adds the notion of integration.

This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are:

* **AR**: *Autoregression*. A model that uses the dependent relationship between an observation and some number of lagged observations 𝑌𝑡 depends only on its own lags. That is, 𝑌𝑡 is a function of the ‘lags of 𝑌𝑡 ’.

𝑌𝑡 = 𝑎 + 𝑏1𝑌𝑡− 1 + 𝑏2𝑌𝑡 −2 + ⋯ + 𝑏𝑝 𝑌𝑡− 𝑝 + 𝑒

* **I**: *Integrated*. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.
* **MA**: *Moving Average*. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations. 𝑌𝑡 depends only on the lagged forecast errors.

𝑌𝑡 = 𝑎 + 𝑒𝑡 + 𝜃1 𝑒𝑡−1 + 𝜃2 𝑒𝑡−2 + ⋯ + 𝜃𝑞 𝑒𝑡−𝑞

An ARIMA model is one where the time series was differenced at least once to make it stationary and you combine the AR and the MA terms. Hence, the equation becomes

𝑌𝑡 = 𝑎 + 𝑏1𝑌𝑡− 1 + 𝑏2𝑌𝑡− 2 + ⋯ + 𝑏𝑝 𝑌𝑡− 𝑝 + 𝑒𝑡 + 𝜃1𝑒𝑡−1 + 𝜃2𝑒𝑡−2 + ⋯ + 𝜃𝑞 𝑒𝑡−𝑞

or

𝑌𝑡 = 𝑒𝑙𝑒𝑚𝑒𝑛𝑡𝑠 𝑜𝑓 𝑝 + 𝑒𝑙𝑒𝑚𝑒𝑛𝑡𝑠 𝑜𝑓 𝑞

Thus, the model called ARIMA with parameter p, d, q. Moreover, each of these components are explicitly specified in the model as a parameter. A standard notation is used of ARIMA (p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.

The parameters of the ARIMA model are defined as follows:

* **p**: The number of lag observations included in the model, also called the lag order.
* **d**: The number of times that the raw observations are differenced, also called the degree of differencing.
* **q**: The size of the moving average window, also called the order of moving average.

In other words ARIMA is a linear regression model is constructed including the specified number and type of terms, and the data is prepared by a degree of differencing in order to make it stationary, i.e. to remove trend and seasonal structures that negatively affect the regression model.

𝑌𝑡 = Constant + Linear combination Lags of 𝑌 (up to 𝑝 lags) + Linear Combination of Lagged forecast errors (up to 𝑞 lags)

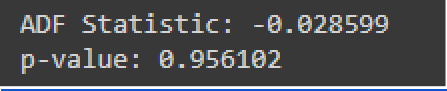
A value of 0 can be used for a parameter, which indicates to not use that element of the model. This way, the ARIMA model can be configured to perform the function of an ARMA model, and even a simple AR, I, or MA model. Model ARIMA (p,d,q) merupakan model umum dari regresi deret waktu sebab ARIMA (p,0,0) sama dengan AR (p), ARIMA (0,0,q) sama dengan MA (p) dan ARIMA (p,0,q) sama dengan ARMA (k,p)

Here, we create a model by adjusting the parameters from differencing the data to get a proper parameter in ARIMA model. So how to determine the right order of differencing?

The right order of differencing is the minimum differencing required to get a near-stationary series which roams around a defined mean and the ACF plot reaches to zero fairly quick.

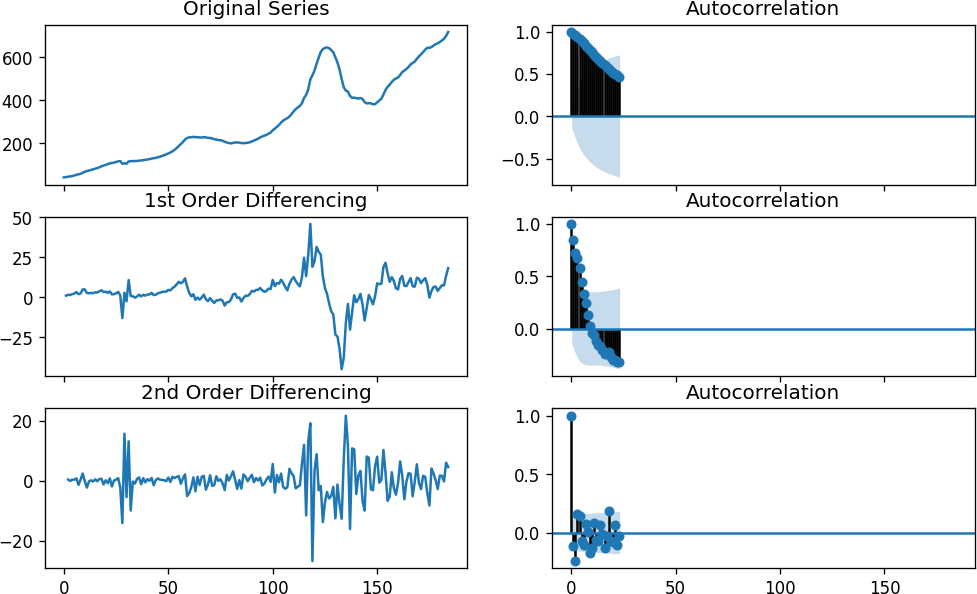
If the autocorrelations are positive for many number of lags (10 or more), then the series needs further differencing. On the other hand, if the lag 1 autocorrelation itself is too negative, then the series is probably over-differenced. In the event, you can’t realy decide between two orders of differencing, then go with the order that gives the least standard deviation in the differenced series.

First, we need to check if the series is stationary using the Augmented Dickey-Fuller test



Since P-value is greater than the significance level, let’s difference the series and see how the autocorrelation plot looks like in finding (d) section

* 1. Finding (d)



By visualization above, authors concluded that second order differencing is the best (d) parameter value as the autocorrelation becomes stable. Otherwise, it could be over differencing if we go further

* 1. Finding (p)

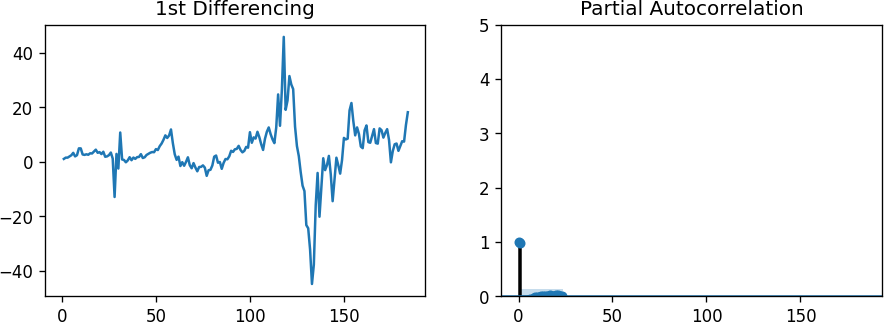
The next step is to identify if the model needs any AR terms. You can find out the required number of AR terms by inspecting the Partial Autocorrelation (PACF) plot. Partial autocorrelation can be imagined as the correlation between the series and its lag, after excluding the contributions from the intermediate lags. So, PACF sort of conveys the pure correlation

between a lag and the series. That way, you will know if that lag is needed in the AR term or not

𝑌𝑡 = 𝑎0 + 𝑎1𝑌𝑡− 1 + 𝑎2𝑌𝑡− 2 + 𝑎3𝑌𝑡− 3 …

Partial autocorrelation of lag (k) of a series is the coefficient of that lag in the autoregression equation of 𝑌. Hence, suppose, if 𝑌𝑡 is the current series and 𝑌𝑡−1 is the lag 1 of 𝑌, then the partial autocorrelation of lag 3 (𝑌𝑡−3 ) is the coefficient alpha of 𝑌𝑡−3 in the equation above. Any autocorrelation in a stationarized series can be rectified by adding enough AR terms. So, we initially take the order of AR term to be equal to as many lags that crosses the significance limit

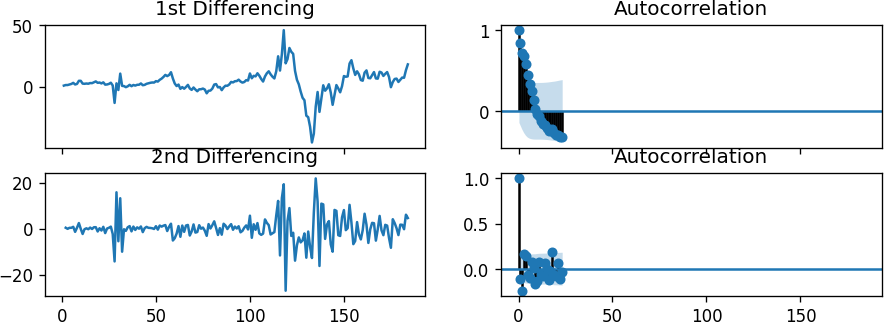
in the PACF plot



It is observed that the PACF lag 1 is quite significant since it is well above the significance line. However, it is still tentative as the parameter have not been tested yet in the model.

* 1. Finding (q)

Just like how we looked at the PACF plot for the number of AR terms, you can look at the ACF plot for the number of MA terms. An MA term is technically, the error of the lagged forecast. The ACF tells how many MA terms are required to remove any autocorrelation in the stationarized series



Here, we found that the parameter of (q) is 2. However, it is just like (p) that we need to adjust later if it cannot provide a proper time series forecast.

* 1. Testing the parameters

- ARIMA (1, 1, 2)

ARIMA Model Results

==============================================================================

|  |  |  |  |
| --- | --- | --- | --- |
| Dep. Variable: | D.CASTHPI | No. Observations: | 184 |
| Model: | ARIMA(1, 1, 2) | Log Likelihood | -566.567 |
| Method: | css-mle | S.D. of innovations | 5.242 |
| Date: | Sat, 12 Jun 2021 | AIC | 1143.134 |
| Time: | 13:48:50 | BIC | 1159.208 |
| Sample: | 1 | HQIC | 1149.649 |

===================================================================================

coef std err z P>|z| [0.025 0.975]

-- - - - - - - - - - - - - - - - - - const 4.0853 2.837 1.440 0.152 -1.476 9.646

ar.L1.D.CASTHPI 0.8927 0.040 22.456 0.000 0.815 0.971

ma.L1.D.CASTHPI -0.0263 0.084 -0.314 0.754 -0.191 0.138

ma.L2.D.CASTHPI -0.1534 0.072 -2.119 0.035 -0.295 -0.012

Roots

=============================================================================

Real Imaginary Modulus Frequency

-- - - - - - - - - - - - - - - - AR.1 1.1202 +0.0000j 1.1202 0.0000

MA.1 2.4685 +0.0000j 2.4685 0.0000

MA.2 -2.6400 +0.0000j 2.6400 0.5000

-- - - - - - - - - - - - - - - -

- ARIMA (4, 2, 2)

ARIMA Model Results

==============================================================================

|  |  |  |  |
| --- | --- | --- | --- |
| Dep. Variable: | D2.CASTHPI | No. Observations: | 148 |
| Model: | ARIMA(4, 2, 2) | Log Likelihood | -457.544 |
| Method: | css-mle | S.D. of innovations | 5.150 |
| Date: | Sat, 12 Jun 2021 | AIC | 931.087 |
| Time: | 14:10:14 | BIC | 955.065 |
| Sample: | 2 | HQIC | 940.829 |

====================================================================================

coef std err z P>|z| [0.025 0.975]

-- - - - - - - - - - - - - - - - - const 0.0166 0.022 0.763 0.447 -0.026 0.059

ar.L1.D2.CASTHPI 1.6882 0.082 20.575 0.000 1.527 1.849

ar.L2.D2.CASTHPI -0.7686 0.158 -4.852 0.000 -1.079 -0.458

ar.L3.D2.CASTHPI 0.3160 0.158 2.006 0.047 0.007 0.625

ar.L4.D2.CASTHPI -0.2612 0.080 -3.260 0.001 -0.418 -0.104

ma.L1.D2.CASTHPI -2.0000 0.039 -51.491 0.000 -2.076 -1.924

ma.L2.D2.CASTHPI 1.0000 0.039 25.775 0.000 0.924 1.076

Roots

=============================================================================

Real Imaginary Modulus Frequency

-- - - - - - - - - - - - - - - - AR.1 1.0239 -0.1320j 1.0323 -0.0204

AR.2 1.0239 +0.1320j 1.0323 0.0204

AR.3 -0.4189 -1.8485j 1.8954 -0.2855

AR.4 -0.4189 +1.8485j 1.8954 0.2855

MA.1 1.0000 -0.0000j 1.0000 -0.0000

MA.2 1.0000 +0.0000j 1.0000 0.0000

-- - - - - - - - - - - - - - - -

After adjusting the parameters we found that (4, 2, 2) have more accuracy in predicting house price index over time. This accuracy can be concluded regarding the facts that the most coefficient of the model is significant. Furthermore, most metrics say that the model is good enough to predict the future.

|  |  |
| --- | --- |
| Metrics | Value |
| MAPE | 0.092 |
| MAE | 57.93 |
| RMSE | 78.59 |

From the metrics value, we can see high value of mean absolute error and root mean squared error. This mean that the model have high error value. However, even though ARIMA (4,4,2) have both MAE and RMSE at high level, the model still be accurate, checked using mean absolute percentage error. By scaling the error to a percentage, we get error just about 9% or in other word we can say that ARIMA (4,4,2) is roughly 91% accurate thus we can use the model.

# Machine Learning Models

* 1. Multiple Linear Regression

OLS Regression Results

==============================================================================

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dep. Variable: | CASTHPI | R -squared: | | 0.000 | | | |
| Model: | OLS | Adj. R-squared: | | -0.000 | | | |
| Method: | Least Squares | F -statistic: | | 0.02228 | | | |
| Date: | Sun, 13 Jun 2021 | Prob (F- statistic): | | 1.00 | | | |
| Time: | 11:13:01 | Log-Likelihood: | | -34274. | | | |
| No. Observations: | 25222 | AIC: | | 6 .856e+04 | | | |
| Df Residuals: | 25215 | BIC: | | 6 .862e+04 | | | |
| Df Model:  Covariance Type: | 6  nonrobust |  | |  | | | |
| ==============================================================  coef std err | | | | ================  t | =======  P>|t| | ============ [0.025 | ==========  0.975] |
| const | | 648.7978 | 2.690 | 241.201 | 0.000 | 643.525 | 654.070 |
| confirmed\_cases | | 2.491e- 16 | 6. 65e-08 | 3.74e- 09 | 1.000 | -1.3e- 07 | 1.3e-07 |
| Population t-1 | | - 1.187e- 17 | 4. 98e-08 | -2.39e- 10 | 1.000 | -9.76e- 08 | 9 .76e-08 |
| CCI | | 1.213e- 08 | 3. 32e-08 | 0.366 | 0.715 | -5.29e- 08 | 7 .71e-08 |
| Median Family Income(California) t-1 | | 4.123e- 12 | 0.052 | 7.97e- 11 | 1.000 | -0.101 | 0.101 |
| Median Household Income(California) t-1 | | - 4.912e- 12 | 0.064 | -7.7e- 11 | 1.000 | -0.125 | 0.125 |
| Total Crime t-1 | | - 4.055e- 18 | 2 .2e-07 | -1.85e- 11 | 1.000 | -4.31e- 07 | 4 .31e-07 |
| ====================  Omnibus: | ==========================================================  23623 .934 Durbin-Watson: 0.282 | | | | | | |
| Prob(Omnibus): | 0 .000 | Jarque-Bera (JB): | | 12868679704.675 | | | |
| Skew: | -2 .408 | Prob(JB): | | 0.00 | | | |

Kurtosis: 3502 .310 Cond. No. 1.76e+10

==============================================================================

Y = 648.7978 + 2.491e-16 X1 - 1.187e-17 X2 + 1.213e-08 X3 + 4.123e-12 X4 - 4.912e-12 X5 - 4.055e-18 X6

𝑌 = 648.7978 + 2.491 ∗ 10−16 𝑋1 − 1.187 ∗ 10−17 𝑋2 + 1.213 ∗ 10−8 𝑋3 + 4.123

∗ 10−12 𝑋4 − 4.912 ∗ 10−12 𝑋5 − 4.055 ∗ 10−18 𝑋6

X1 = confirmed\_cases X2 = Population t-1 X3 = CCI

X4 = Median Family Income (California) t-1

X5 = Median Household Income (California) t-1 X6 = Total Crime t-1

Interpretation:

|  |  |
| --- | --- |
| Parameter | Description |
| b0 = 648.7978 | CASTHPI 648.7978 we can conclude that if the variables (X1), (X2), (X3), (X4), (X5), and (X6) are zero (0). So, if CASTHPI is not influenced by other variables, it will be worth 648.7978. |
| b1 = 2.491e-16. | confirmed\_cases (X1) is positive, then the confirmed\_cases (X1) variable has a positive effect / proportional to CASTHPI, if X1 increases then CASTHPI also increases. The parameter value is 2.491e-16, so when X1 increases by 1 unit, CASTHPI increases by 2.491e-16. assuming the other variables are constant. |
| b2 = - 1.187e-17. | Population t-1 (X2) is negative. If Population t-1 (X2) increases then CASTPHI actually decreases. The value is 0.1084 when X2 increases then the house price decreases by 1.187e-17 |
| b3 = 1.213e-08. | CCI (X3) is positive, then the CCI variable (X3) has a positive effect / proportional to CASTHPI, if X3 increases then CASTHPI also increases. Parameter value 1.213e-08. |

|  |  |
| --- | --- |
|  | then when X3 increases by 1 unit then CASTHPI increases by 1.213e-08 assuming other variables are constant |
| b4 = 4.123e-12. | Median Family Income (California) t-1 (X4) is positive, then the variable Median Family Income (California) t-1 (X4) has a positive effect / proportional to CASTHPI, if X4 increases then CASTHPI also increases. The parameter value is 4.123e-12 then when X4 goes up a thousand dollars then CASTHPI goes up by 4.123e-12 assuming other variables are constant |
| b5 = -4.912e-12. | Median Household Income(California) t-1 (X5) is negative. if the Median Household Income (California) t-1 (X5) goes up, then CASTHPI actually goes down. The value is  -4.912e-12 when X5 increases then CASTHPI decreases by 4.912e-12 |
| b6 = -4.055e-18. | Total Crime t-1 (X6) is negative. if Total Crime t-1 (X6) goes up, then CASTHPI actually goes down. The value of 4.055e-18 when X6 increases then CASTHPI decreases by 4.055e-18 |

Concurrent Test Result

F-statistic = 234.72334155427433 P-value = 2.7031835586322296e-150

Interpretation:

Because the P-value (2.7031835586322296e-150) is smaller than alpha (0.05), the decision is to reject H0, then there is a linear relationship between X and Y variables

Coefficient of Determination Value (R-Square) and Adjusted R square

R2 = 5.302363424330991e-06

adjusted R2 = -0.00023264997390737285

Interpretation:

The values of R2 and adjusted R2 obtained indicate that the model is not accurate because the Y variable cannot be explained by the X variable, and the value of the r- square value is not close to 1.

Linear Regression Feature Importance

Feature: 0, Score: 60.32347

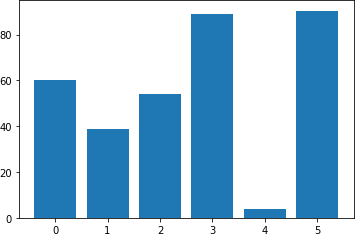
Feature: 1, Score: 38.67115

Feature: 2, Score: 54.03354

Feature: 3, Score: 88.82734

Feature: 4, Score: 4.03574

Feature: 5, Score: 90.44347



Interpretation:

From the result of linear regression feature importance, we can assume that the score indicates the model finds six important features.

* 1. Gradient Boosting

XGBoost is an algorithm that has recently dominated the application of machine learning for structured or tabular data. XGBoost is an implementation of a gradient boosting decision tree designed for speed and performance.

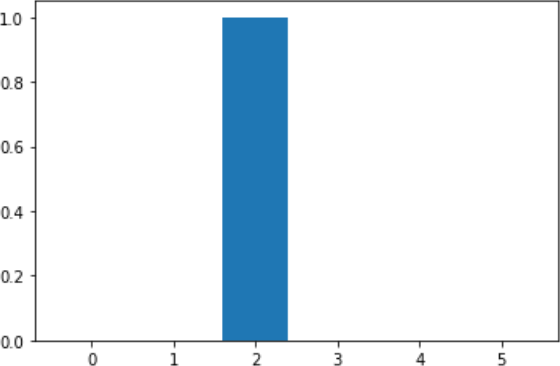
For boosting pharameter, we use tree based models. The reason is that it is a simple model and is not sensitive to scale differences

From the model we obtain the following evaluation.

R2 : -24.94%

RMSE : 1.79

MAE : 0.49

We assume that the negative value of R2 is because the target variable is not so varied. As a result, it is difficult for us to determine how well the feature describes the target variation. RMSE and MAE scored 1.79 and 1.49, respectively. This values indicates that the error in the model is quite small. Furthermore, we present what features are important in the model.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature: | 0, | Score: | 0.00000 |
| Feature: | 1, | Score: | 0.00000 |
| Feature: | 2, | Score: | 1.00000 |
| Feature: | 3, | Score: | 0.00000 |
| Feature: | 4, | Score: | 0.00000 |
| Feature: | 5, | Score: | 0.00000 |

From this it can be seen that the only feature that has an effect on the model is only feature 2, namely CCI.

1. References

Prabhakaran, S. (n.a.). ARIMA Model – Complete Guide to Time Series Forecasting in Python. [https://www.machinelearningplus.com/time-series/arima-model-time-series-](https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/)

[forecasting-python/](https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/)

Jason Brownlee. (August 17, 2016). A Gentle Introduction to XGBoost for Applied Machine Learning. [https://machinelearningmastery.com/gentle-introduction-xgboost-applied-](https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/)

[machine- learning/](https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/)

FBI. (n.a.). Crime Rate in the US. [https://ucr.fbi.gov/crime- in-the-u.s/](https://ucr.fbi.gov/crime-in-the-u.s/)

American Community Survey. (n.a.). California Household Income. <https://www.deptofnumbers.com/income/california/>

US Census Bureau. (n.a.). Population. <https://www.census.gov/content/census/en/> DOJ. (n.a.). Crime Statistics. <https://openjustice.doj.ca.gov/exploration/crime-statistics>

St. Louis Fed. (n.a.). 30-Year Fixed Rate Mortgage Average in the United States. <https://fred.stlouisfed.org/series/MORTGAGE30US>

LA Times. (n.a.). california-coronavirus-data. [https://github.com/datadesk/california-](https://github.com/datadesk/california-coronavirus-data)

[coronavirus-data](https://github.com/datadesk/california-coronavirus-data)