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

1. 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.

2. 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-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-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

	State Household Income | Department of Numbers (deptofnumbers.com)
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 | Department of Numbers (deptofnumbers.com)
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: California Consumer Price Index 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/ in a CSV format
Population	The total number of persons inhabiting a country, city, or any district or area. Sourced from Search Results in a CSV format

3. Explanatory Data and Variable Analysis

a. Data description and basic statistical analysis

- Population

	Total Population t-1
count	1.581000e+03
mean	3.617541e+07
std	1.043379e+07
min	0.000000e+00
25%	3.824870e+07
50%	3.907132e+07
75%	3.950093e+07
max	3.987306e+07

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

	confirmed_cases
count	2.619000e+04
mean	2.741497e+04
std	1.028101e+05
min	0.000000e+00
25%	2.280000e+02
50%	2.549000e+03
75%	1.555400e+04
max	1.244917e+06

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

	Value
count	5.200000e+01
mean	3.092162e+06
std	3.975898e+06
min	1.010350e+05
25%	1.011564e+06
50%	1.014525e+06
75%	3.219552e+06
max	9.994392e+06

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

	Median Family Income(California) t-1
count	5930.000000
mean	81.402659
std	5.017310
min	73.866000
25%	76.946000
50%	82.066000
75%	85.530000
max	91.377000

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

	Median Household Income(California) t-1
count	5930.000000
mean	71.464236
std	4.448384
min	65.068000
25%	67.812000
50%	72.001000
75%	74.888000
max	80.440000

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

	Total Crime t-1
count	1581.000000
mean	485711.218564
std	146542.128466
min	0.000000
25%	478795.300000
50%	502558.500000
75%	522485.800000
max	600021.870000

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

	CASTHPI
count	17.000000
mean	648.835294
std	37.389765
min	581.120000
25%	624.580000
50%	648.450000
75%	671.540000
max	718.340000

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.

c. ANOVA and Tukey HSD

- ANOVA CASTHPI

Pingouin Test

	Source	SS	DF	MS	F	p-unc	np2
0	CASTHPI	6.783618e+06	46	147469.953435	50.216024	1.092791e-67	0.942471
1	Within	4.140763e+05	141	2936.711080	NaN	NaN	NaN

OLS Method

	sum_sq	df	F	PR(>F)
CASTHPI	6.783618e+06	46.0	50.216024	1.092791e-67
Residual	4.140763e+05	141.0	NaN	NaN

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

Multiple Comparison of Means - Tukey HSD, FWER=0.05

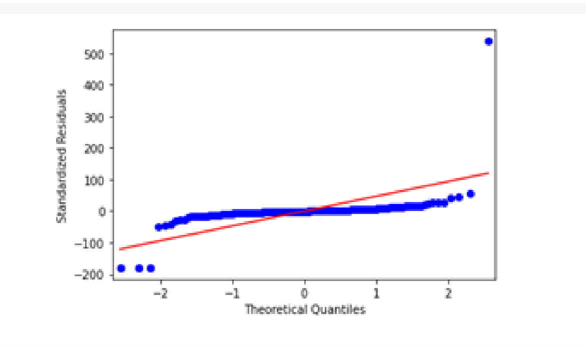
group1	group2	meandiff	p-adj	lower	upper	reject
tahun_1975	tahun_1976	8.175	0.9	-147.2379	163.5879	False
tahun_1975	tahun_1977	21.2725	0.9	-134.1404	176.6854	False
tahun_1975	tahun_1978	33.765	0.9	-121.6479	189.1779	False
tahun_1975	tahun_1979	47.1075	0.9	-108.3054	202.5204	False
tahun_1975	tahun_1980	60.8325	0.9	-94.5804	216.2454	False
tahun_1975	tahun_1981	70.6925	0.9	-84.7204	226.1054	False
tahun_1975	tahun_1982	64.6825	0.9	-90.7304	220.0954	False
tahun_1975	tahun_1983	73.725	0.9	-81.6879	229.1379	False
tahun_1975	tahun_1984	77.45	0.9	-77.9629	232.8629	False
tahun_1975	tahun_1985	84.0825	0.9	-71.3304	239.4954	False
tahun_1975	tahun_1986	92.9275	0.9	-62.4854	248.3404	False
tahun_1975	tahun_1987	107.27	0.7118	-48.1429	262.6829	False
tahun_1975	tahun_1988	131.035	0.2843	-24.3779	286.4479	False
tahun_1975	tahun_1989	167.76	0.0168	12.3471	323.1729	True
tahun_1975	tahun_1990	185.1825	0.0029	29.7696	340.5954	True
tahun_1975	tahun_1991	184.41	0.0032	28.9971	339.8229	True
tahun_1975	tahun_1992	181.5775	0.0043	26.1646	336.9904	True
tahun_1975	tahun_1993	173.0775	0.0101	17.6646	328.4904	True
tahun_1975	tahun_1994	162.4975	0.0271	7.0846	317.9104	True
tahun_1975	tahun_1995	158.685	0.038	3.2721	314.0979	True
tahun_1975	tahun_1996	158.185	0.0397	2.7721	313.5979	True
tahun_1975	tahun_1997	163.5525	0.0246	8.1396	318.9654	True
tahun_1975	tahun_1998	181.0625	0.0045	25.6496	336.4754	True
tahun_1975	tahun_1999	199.04	0.001	43.6271	354.4529	True
tahun_1975	tahun_2000	229.4975	0.001	74.0846	384.9104	True
tahun_1975	tahun_2001	264.3325	0.001	108.9196	419.7454	True
tahun_1975	tahun_2002	298.73	0.001	143.3171	454.1429	True
tahun_1975	tahun_2003	341.45	0.001	186.0371	496.8629	True
tahun_1975	tahun_2004	429.015	0.001	273.6021	584.4279	True
tahun_1975	tahun_2005	540.01	0.001	384.5971	695.4229	True
tahun_1975	tahun_2006	599.585	0.001	444.1721	754.9979	True
tahun_1975	tahun_2007	564.785	0.001	409.3721	720.1979	True
tahun_1975	tahun_2008	443.5	0.001	288.0871	598.9129	True
tahun_1975	tahun_2009	378.0225	0.001	222.6096	533.4354	True
tahun_1975	tahun_2010	365.41	0.001	209.9971	520.8229	True
tahun_1975	tahun_2011	344.13	0.001	188.7171	499.5429	True
tahun_1975	tahun_2012	345.055	0.001	189.6421	500.4679	True
tahun_1975	tahun_2013	393.065	0.001	237.6521	548.4779	True
tahun_1975	tahun_2014	445.5575	0.001	290.1446	600.9704	True
tahun_1975	tahun_2015	480.235	0.001	324.8221	635.6479	True
tahun_1975	tahun_2016	517.1975	0.001	361.7846	672.6104	True
tahun_1975	tahun_2017	554.75	0.001	399.3371	710.1629	True
tahun_1975	tahun_2018	593.94	0.001	438.5271	749.3529	True
tahun_1975	tahun_2019	614.0425	0.001	458.6296	769.4554	True
tahun_1975	tahun_2020	640.7025	0.001	485.2896	796.1154	True
tahun_1975	tahun_2021	135.9525	0.211	-19.4604	291.3654	False
tahun_1976	tahun_1977	13.0975	0.9	-142.3154	168.5104	False
tahun_1976	tahun_1978	25.59	0.9	-129.8229	181.0029	False
tahun_1976	tahun_1979	38.9325	0.9	-116.4804	194.3454	False
tahun_1976	tahun_1980	52.6575	0.9	-102.7554	208.0704	False
tahun_1976	tahun_1981	62.5175	0.9	-92.8954	217.9304	False
tahun_1976	tahun_1982	56.5075	0.9	-98.9054	211.9204	False
tahun_1976	tahun_1983	65.55	0.9	-89.8629	220.9629	False
tahun_1976	tahun_1984	69.275	0.9	-86.1379	224.6879	False

tahun_2012	tahun_2015	135.18	0.2217	-20.2329	290.5929	False
tahun_2012	tahun_2016	172.1425	0.011	16.7296	327.5554	True
tahun_2012	tahun_2017	209.695	0.001	54.2821	365.1079	True
tahun_2012	tahun_2018	248.885	0.001	93.4721	404.2979	True
tahun_2012	tahun_2019	268.9875	0.001	113.5746	424.4004	True
tahun_2012	tahun_2020	295.6475	0.001	140.2346	451.0604	True
tahun_2012	tahun_2021	-209.1025	0.001	-364.5154	-53.6896	True
tahun_2013	tahun_2014	52.4925	0.9	-102.9204	207.9054	False
tahun_2013	tahun_2015	87.17	0.9	-68.2429	242.5829	False
tahun_2013	tahun_2016	124.1325	0.4149	-31.2804	279.5454	False
tahun_2013	tahun_2017	161.685	0.0292	6.2721	317.0979	True
tahun_2013	tahun_2018	200.875	0.001	45.4621	356.2879	True
tahun_2013	tahun_2019	220.9775	0.001	65.5646	376.3904	True
tahun_2013	tahun_2020	247.6375	0.001	92.2246	403.0504	True
tahun_2013	tahun_2021	-257.1125	0.001	-412.5254	-101.6996	True
tahun_2014	tahun_2015	34.6775	0.9	-120.7354	190.0904	False
tahun_2014	tahun_2016	71.64	0.9	-83.7729	227.0529	False
tahun_2014	tahun_2017	109.1925	0.6789	-46.2204	264.6054	False
tahun_2014	tahun_2018	148.3825	0.0875	-7.0304	303.7954	False
tahun_2014	tahun_2019	168.485	0.0157	13.0721	323.8979	True
tahun_2014	tahun_2020	195.145	0.001	39.7321	350.5579	True
tahun_2014	tahun_2021	-309.605	0.001	-465.0179	-154.1921	True
tahun_2015	tahun_2016	36.9625	0.9	-118.4504	192.3754	False
tahun_2015	tahun_2017	74.515	0.9	-80.8979	229.9279	False
tahun_2015	tahun_2018	113.705	0.6017	-41.7079	269.1179	False
tahun_2015	tahun_2019	133.8075	0.2406	-21.6054	289.2204	False
tahun_2015	tahun_2020	160.4675	0.0325	5.0546	315.8804	True
tahun_2015	tahun_2021	-344.2825	0.001	-499.6954	-188.8696	True
tahun_2016	tahun_2017	37.5525	0.9	-117.8604	192.9654	False
tahun_2016	tahun_2018	76.7425	0.9	-78.6704	232.1554	False
tahun_2016	tahun_2019	96.845	0.8902	-58.5679	252.2579	False
tahun_2016	tahun_2020	123.505	0.4276	-31.9079	278.9179	False
tahun_2016	tahun_2021	-381.245	0.001	-536.6579	-225.8321	True
tahun_2017	tahun_2018	39.19	0.9	-116.2229	194.6029	False
tahun_2017	tahun_2019	59.2925	0.9	-96.1204	214.7054	False
tahun_2017	tahun_2020	85.9525	0.9	-69.4604	241.3654	False
tahun_2017	tahun_2021	-418.7975	0.001	-574.2104	-263.3846	True
tahun_2018	tahun_2019	20.1025	0.9	-135.3104	175.5154	False
tahun_2018	tahun_2020	46.7625	0.9	-108.6504	202.1754	False
tahun_2018	tahun_2021	-457.9875	0.001	-613.4004	-302.5746	True
tahun_2019	tahun_2020	26.66	0.9	-128.7529	182.0729	False
tahun_2019	tahun_2021	-478.09	0.001	-633.5029	-322.6771	True
tahun_2020	tahun_2021	-504.75	0.001	-660.1629	-349.3371	True

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.

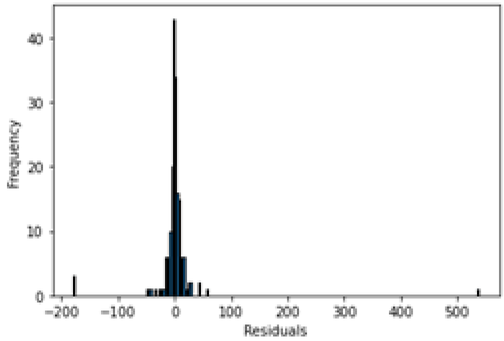
QQ Plot



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

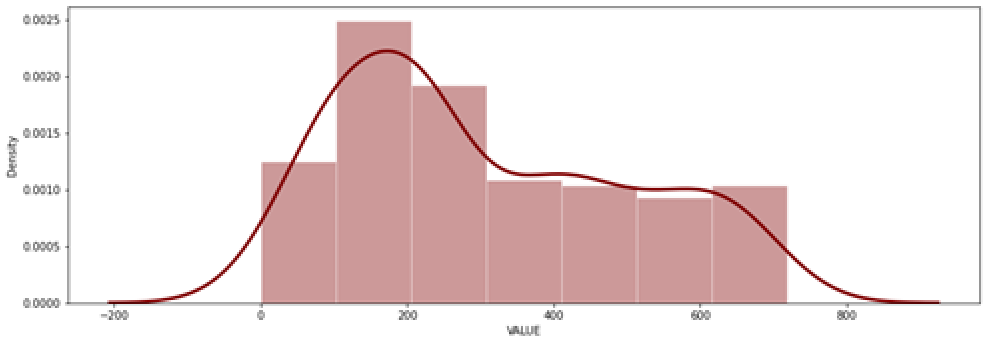


Interpretation:

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

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2557: FutureWarning: `distplot` is a deprecated function and
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7fb6554dc290>
```



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

```
Statistics=0.984, p=0.000
Data Tidak Berdistribusi Normal (Tolak H0)
```

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

```
test=1.004, p=0.478
Kesimpulan:
Gagal tolak H0. Sampel dari populasi homogen
```

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

```
           Source            SS   DF  ...            F          p-unc       np2
0  populationyear  6.947079e+16    4  ...  9849.262316  4.671731e-304  0.992818
1          Within  5.025548e+14  285  ...          NaN            NaN       NaN

[2 rows x 7 columns]
```

OLS method

```
                      sum_sq     df            F         PR(>F)
populationyear  6.947079e+16    4.0  9849.262316  4.671731e-304
Residual        5.025548e+14  285.0          NaN            NaN
```

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.

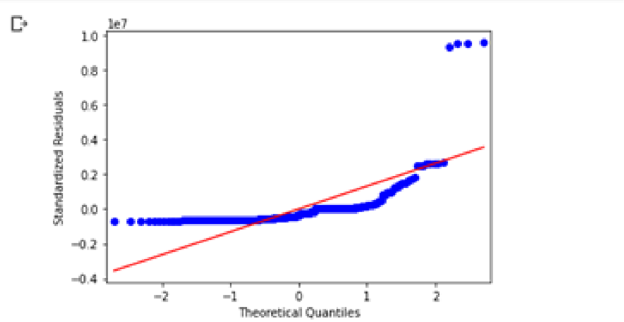
Tukey test

```
          Multiple Comparison of Means - Tukey HSD, FWER=0.05
======================================================================
 group1   group2     meandiff   p-adj     lower          upper     reject
----------------------------------------------------------------------
pop 2017 pop 2018      7407.069    0.9  -669564.3805   684378.5185  False
pop 2017 pop 2019     9068.9829    0.9  -667902.4666   686040.4324  False
pop 2017 pop 2020   -14183.0862    0.9  -691154.5357   662788.3633  False
pop 2017 pop 2021    38694434.5 0.001 38017463.0505 39371405.9495   True
pop 2018 pop 2019      1661.914    0.9  -675309.5355   678633.3635  False
pop 2018 pop 2020   -21590.1552    0.9  -698561.6047   655381.2943  False
pop 2018 pop 2021  38687027.431 0.001 38010055.9815 39363998.8805   True
pop 2019 pop 2020   -23252.0691    0.9  -700223.5186   653719.3803  False
pop 2019 pop 2021 38685365.5171 0.001 38008394.0676 39362336.9666   True
pop 2020 pop 2021 38708617.5862 0.001 38031646.1367 39385589.0357   True
----------------------------------------------------------------------
```

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.

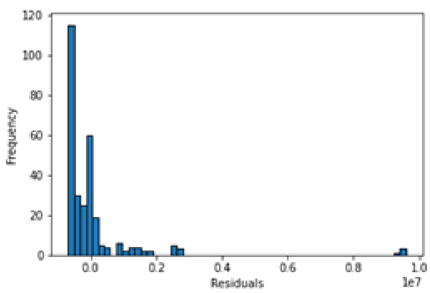
QQ Plot



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

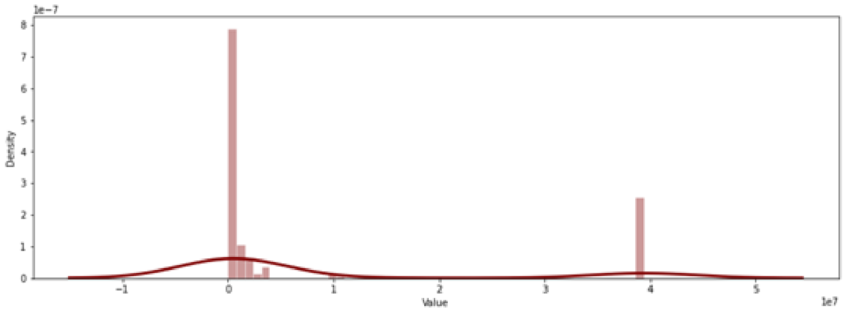


Interpretation:

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

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2557: FutureWarning: `distplot` is a deprecated function an
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f0bd8736dd0>
```



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

```
Statistics=1.000, p=0.000
Data Tidak Berdistribusi Normal (Tolak H0)
```

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).

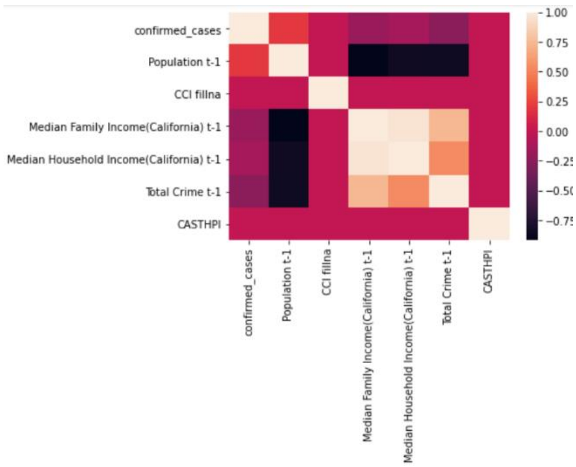
Levene Test

```
test=2.667, p=0.033
Kesimpulan:
Tolak H0. Sampel dari populasi heterogen
```

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.

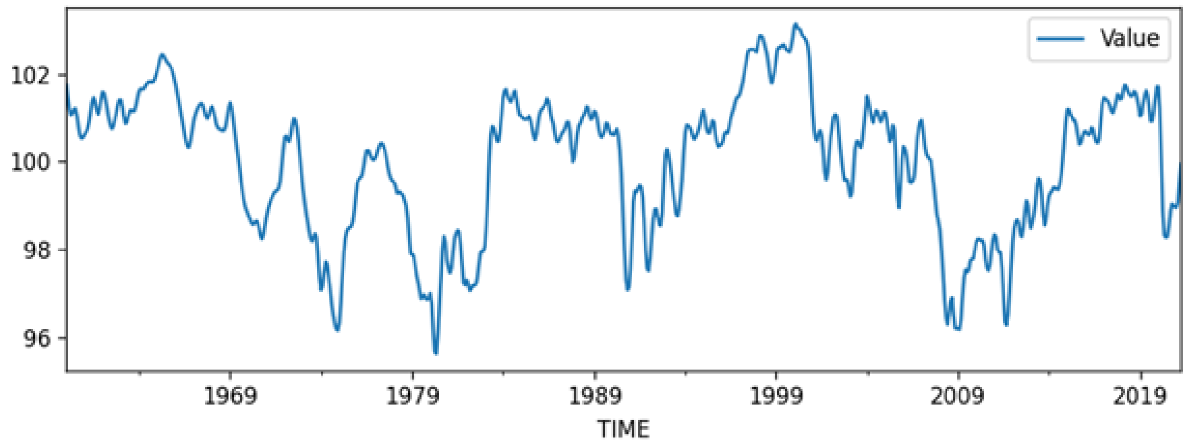
d. Matrix correlation

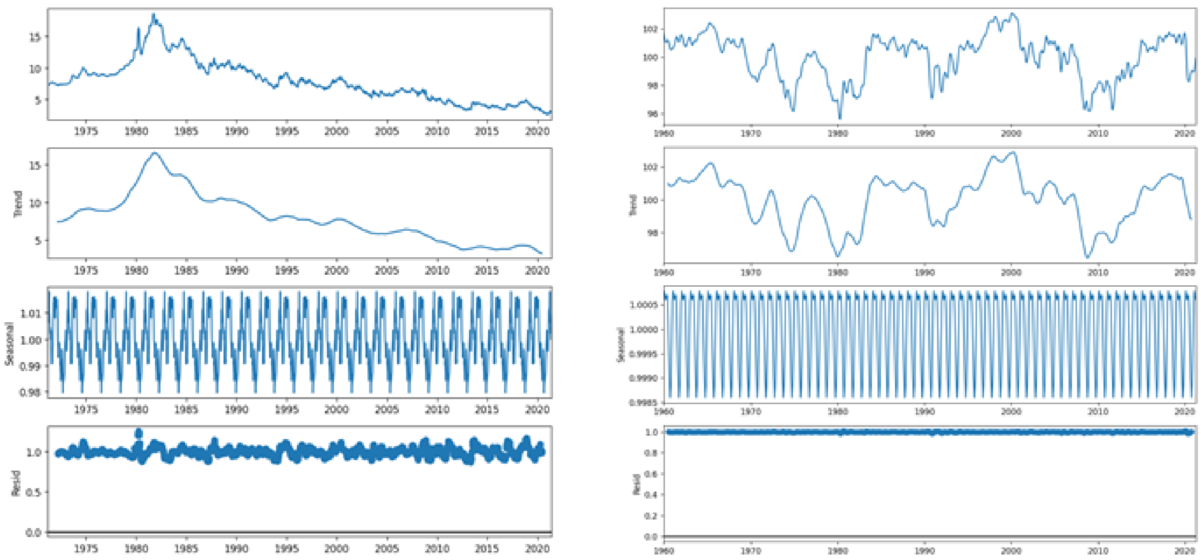


4. Time Series Decomposition

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.

5. Time Series Analysis

a. ARIMA

An ARIMA model 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 Y_t depends only on its own lags. That is, Y_t is a function of the 'lags of Y_t'.

$$Y_t = a + b_1 Y_{t-1} + b_2 Y_{t-2} + \cdots + b_p Y_{t-p} + e$$

- **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. Y_t depends only on the lagged forecast errors.

$$Y_t = a + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}$$

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

$$Y_t = a + b_1 Y_{t-1} + b_2 Y_{t-2} + \cdots + b_p Y_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}$$

or

$$Y_t = elements\ of\ p + elements\ of\ q$$

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.

Y_t = 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.

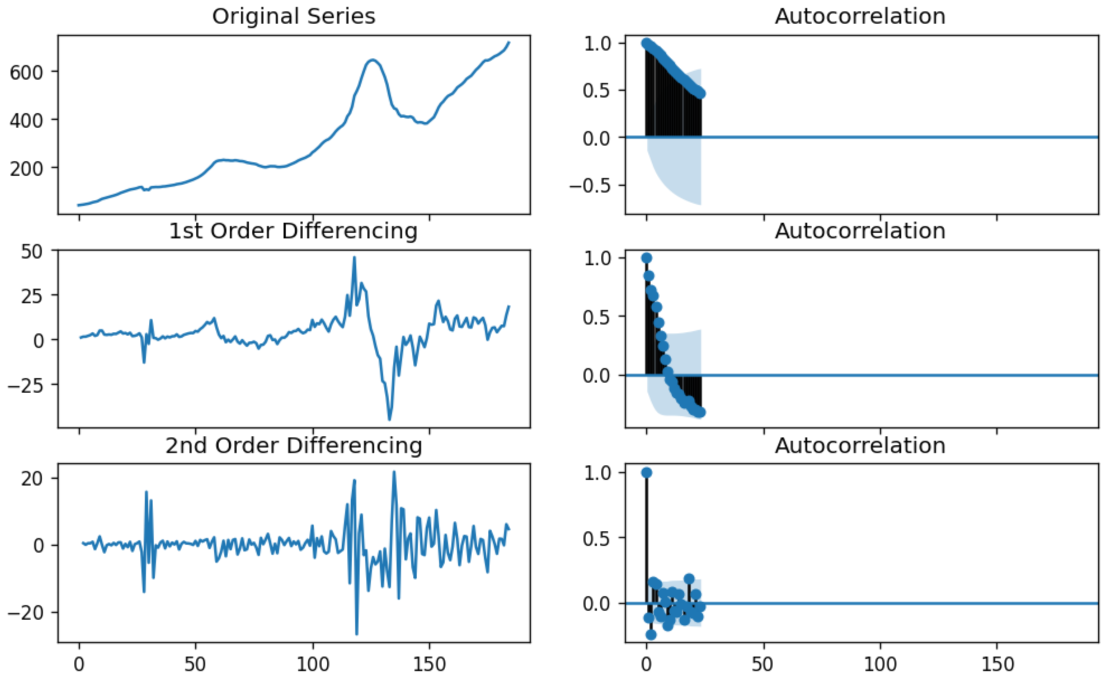
If the autocorrelations are positive for many number of lags (10 or more), then the series needs further differencing. On the other hand, if the lag 1 autocorrelation itself is too negative, then the series is probably over-differenced. In the event, you can't really decide between two orders of differencing, 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

```
ADF Statistic: -0.028599
p-value: 0.956102
```

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

b. Finding (d)



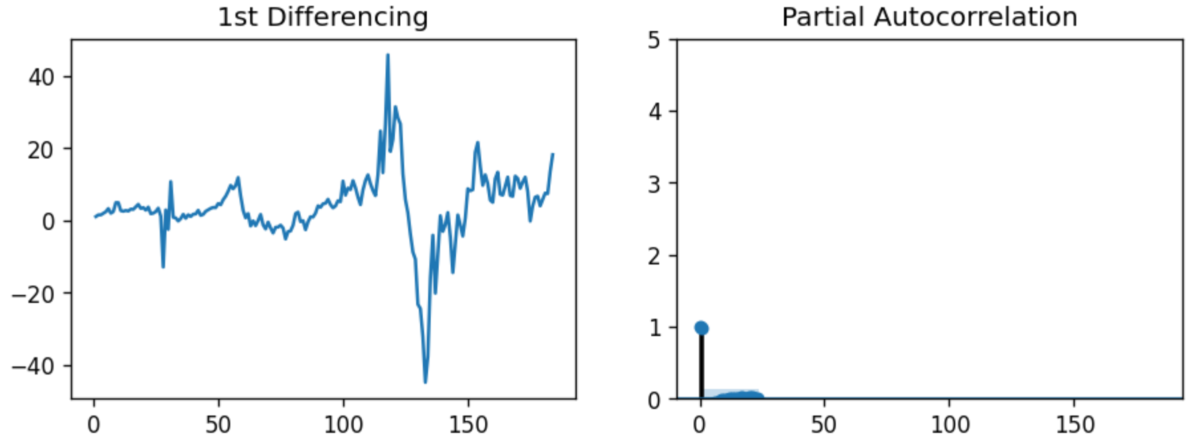
By visualization above, authors concluded that second order differencing is the best (d) parameter value as the autocorrelation becomes stable. Otherwise, it could be over differencing if we go further

c. 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

$$Y_t = a_0 + a_1 Y_{t-1} + a_2 Y_{t-2} + a_3 Y_{t-3} \dots$$

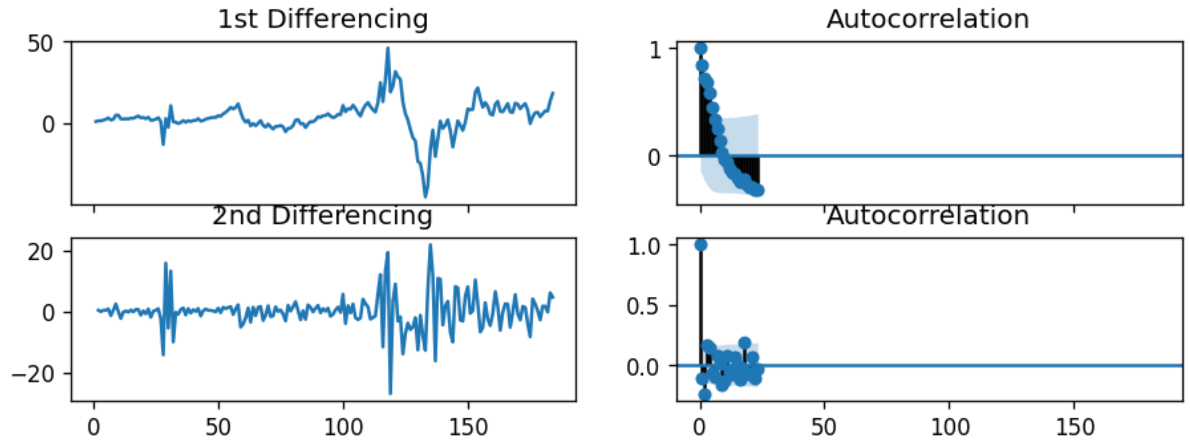
Partial autocorrelation of lag (k) of a series is the coefficient of that lag in the autoregression equation of Y. Hence, suppose, if Y_t is the current series and Y_{t-1} is the lag 1 of Y, then the partial autocorrelation of lag 3 (Y_{t-3}) is the coefficient alpha of Y_{t-3} in the equation above. 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.

d. 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.

e. 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.

6. Machine Learning Models

a. 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:                           6
Covariance Type:            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 X_1 - 1.187e-17 X_2 + 1.213e-08 X_3 + 4.123e-12 X_4 - 4.912e-12 X_5 - 4.055e-18 X_6

$$Y = 648.7978 + 2.491 * 10^{-16}\ X_1 - 1.187 * 10^{-17}\ X_2 + 1.213 * 10^{-8}\ X_3 + 4.123 * 10^{-12}\ X_4 - 4.912 * 10^{-12}\ X_5 - 4.055 * 10^{-18}\ X_6$$

X_1 = confirmed_cases
X_2 = Population t-1
X_3 = CCI
X_4 = Median Family Income (California) t-1
X_5 = Median Household Income (California) t-1
X_6 = Total Crime t-1

Interpretation:

Parameter	Description
b0 = 648.7978	CASTHPI 648.7978 we can conclude that if the variables (X_1), (X_2), (X_3), (X_4), (X_5), and (X_6) are zero (0). So, if CASTHPI is not influenced by other variables, it will be worth 648.7978.
b1 = 2.491e-16.	confirmed_cases (X_1) is positive, then the confirmed_cases (X_1) variable has a positive effect / proportional to CASTHPI, if X_1 increases then CASTHPI also increases. The parameter value is 2.491e-16, so when X_1 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 (X_3) is positive, then the CCI variable (X_3) has a positive effect / proportional to CASTHPI, if X_3 increases then CASTHPI also increases. Parameter value 1.213e-08.

	then when X_3 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 (X_4) is positive, then the variable Median Family Income (California) t-1 (X_4) has a positive effect / proportional to CASTHPI, if X_4 increases then CASTHPI also increases. The parameter value is 4.123e-12 then when X_4 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 (X_5) is negative. if the Median Household Income (California) t-1 (X_5) goes up, then CASTHPI actually goes down. The value is -4.912e-12 when X_5 increases then CASTHPI decreases by 4.912e-12
b6 = -4.055e-18.	Total Crime t-1 (X_6) is negative. if Total Crime t-1 (X_6) goes up, then CASTHPI actually goes down. The value of 4.055e-18 when X_6 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

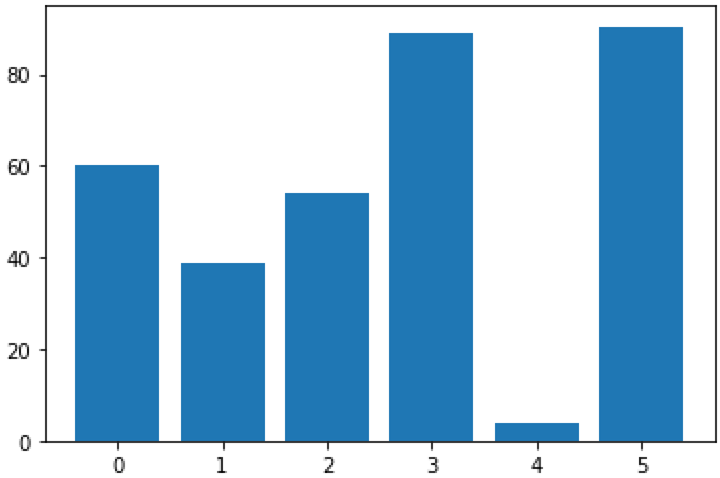
```
R2 =  5.302363424330991e-06

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.

b. 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.

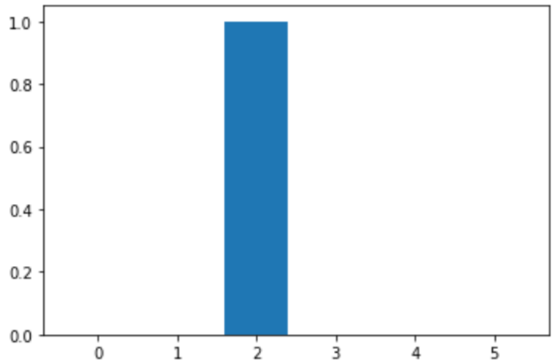
For boosting pharameter, we use tree based models. The reason is that it is a simple model and is not sensitive to scale differences

From the model we obtain the following evaluation.

```
R2    : -24.94%
RMSE  : 1.79
```

```
MAE    : 0.49
```

We assume that the negative value of R^2 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.

7. 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-forecasting-python/

Jason Brownlee. (August 17, 2016). A Gentle Introduction to XGBoost for Applied 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/

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-coronavirus-data