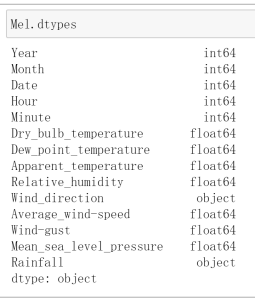
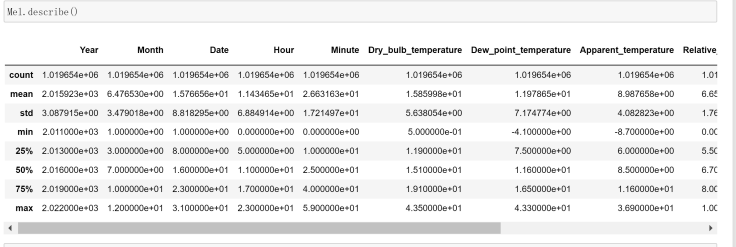
Siming Wu

**Individual Report Of ADS2001 Predict Melbourne Climate**

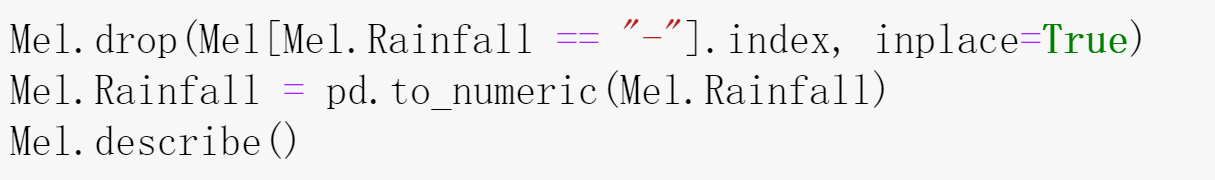
In each of individual report, is divided by the types of the model that are applied to the dataset. In this section of report is aimed to discovery the effect of using SARIMA model and use this model to predict future climate features based on testing of accuracy and measurement of errors.

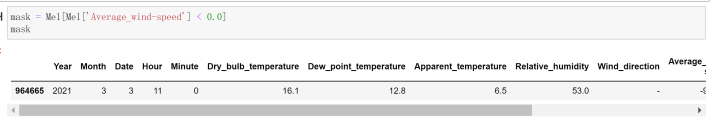
**PART1**

**The first process of data preparation before setting up model is about data cleaning and arrangement, and this is important to generate a good visualization and testify resample dataset.** We use the function df.describe() to look over the count of each features of the raw dataset and Identify the type of each feature.

****

Base on this information, are aimed to find any invalid value and remove it. For example, rainfall shouldn’t exist negative value. So mask the value which exist non-logical value and NaN like below. We can observe that the rainfall is not numeric, is object, so we need to convert it to numeric.

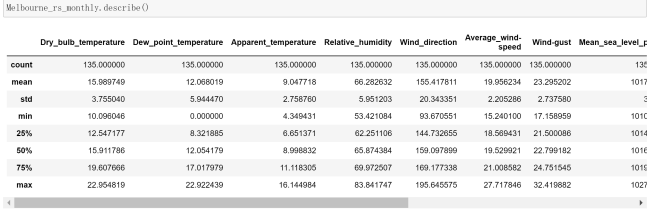


****

****

**PART 2 and PART 3**

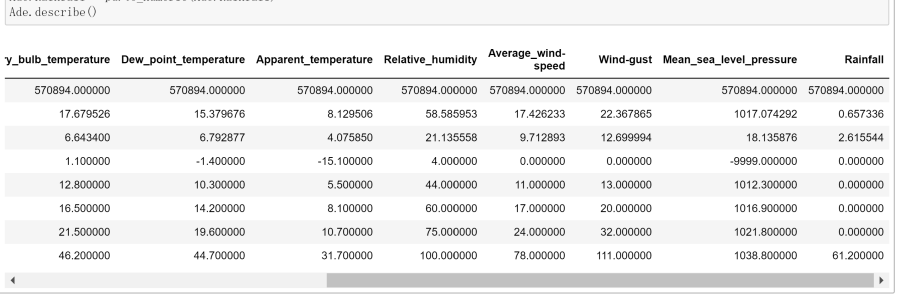
In order to make the dataset with clear timestamp and reduce the frequency of the dataset, convert feature of Year, Month, Day , Hour, Minute to timestamp by applying pd.to\_datetime(). As the visualization of the dataset suggests, the raw data contains up to 500000 rows of data, it is not rational to cover all the rows, so it is better to do resample and convert it to frequency to monthly over 10 years observation, also check and remove for its Nan and invalid value if it exists.

****

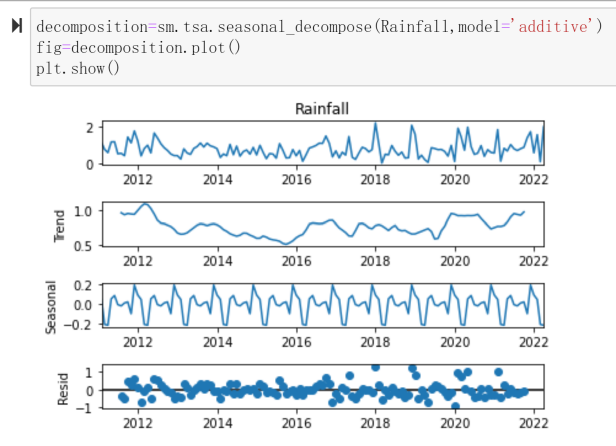
****

****

In general, data cleaning is manual, but resampling, rearrangement and visualization of data is automatic. At the beginning I had doubt on the feature rainfall because the measurement of rainfall using annual precipitation with hundreds statistics, but most of the data are 0, later I figured out this is not cumulative precipitation, it only records the precipitation during the interval. By comparing dataset of Melbourne and Adelaide, this convinces the rainfall records a non-cumulative precipitation. In fact, it is rational that most of the time, it doesn’t rain so mode should be 0.

****

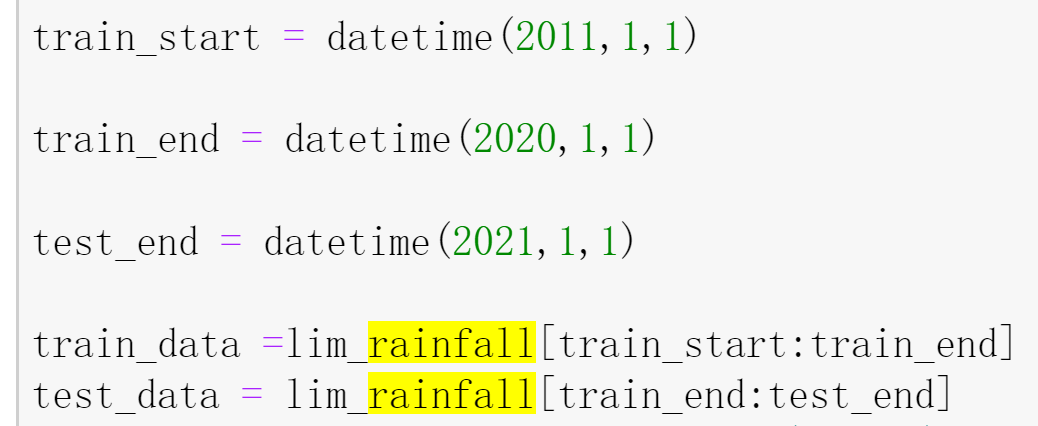
By plotting, we can observe how the dataset look like overall, and we can use criteria such as residual to measure whether it is a good graph.

****

We can see that the plot show seasonality of rainfall over 10 years, also that the scatter of residual evenly distributed both side of 0, which demonstrate the resample dataset well-performed.

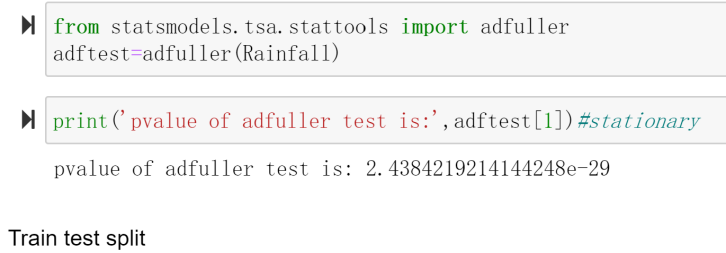
The second step is to train the model and testify its accuracy with different measurement.

The process of splitting train and test by date, the testing dataset is based on previous 9 years from 2011 to 2020.



**Before applied SARIMA model, we must testify that the dataset Rainfall is stationary.**

**The input of SARIMA model is the feature rainfall from monthly resample dataset, and we expect that the output demonstrate the output of prediction and its accuracy based on RMSE and MAPE.**

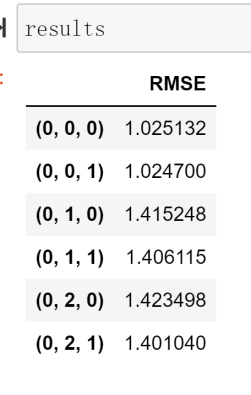
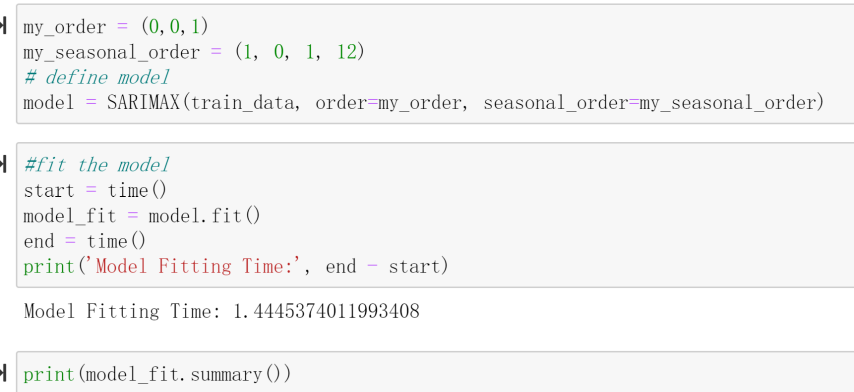
****

**The pvalue of adfuller test is quite low and less than 0.05 which demonstrates it is a stationary function. After this we can apply dataset to model SARIMA.**

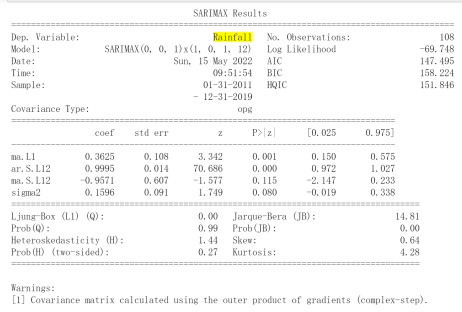
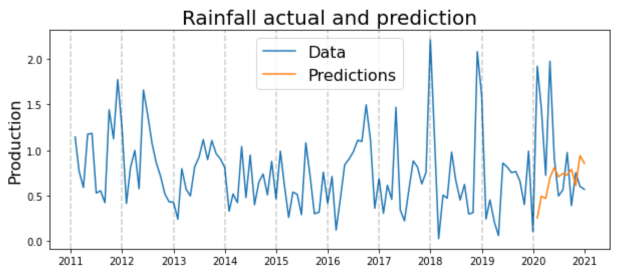
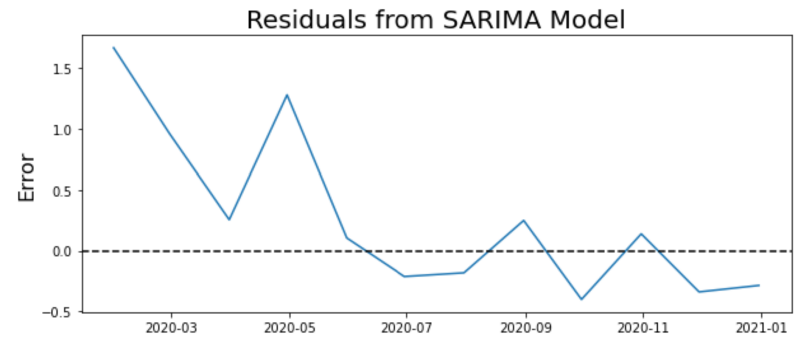
ARIMA(Autoregressive Integrated Moving Average) is models are fitted to [time series](https://en.wikipedia.org/wiki/Time_series" \o "Time series) data either to better understand the data or to predict future points in the series ([forecasting](https://en.wikipedia.org/wiki/Forecasting" \o "Forecasting)). The parameters P,D,Q, as one of the input of the model, the choice of these parameter will interfere the result of prediction, so that determination of P, D, Q is a very important step of SARIMA. **Also, series Rainfall is also an input. The next step is to identify the parameter P,Q,D which related to AR(autoregressive), I(intergrated) and MA(moving average) respectively.**

**ACF is related to p and PACF is related to q. the ACF should show a sharp drop after a certain q number of lags while PACF should show a geometric or gradual decreasing trend**.For example, ARIMA(1, 0, 12) means that if you are describing some response variable (Y) by combining a 1st order Auto-Regressive model and a 12th order Moving Average model.

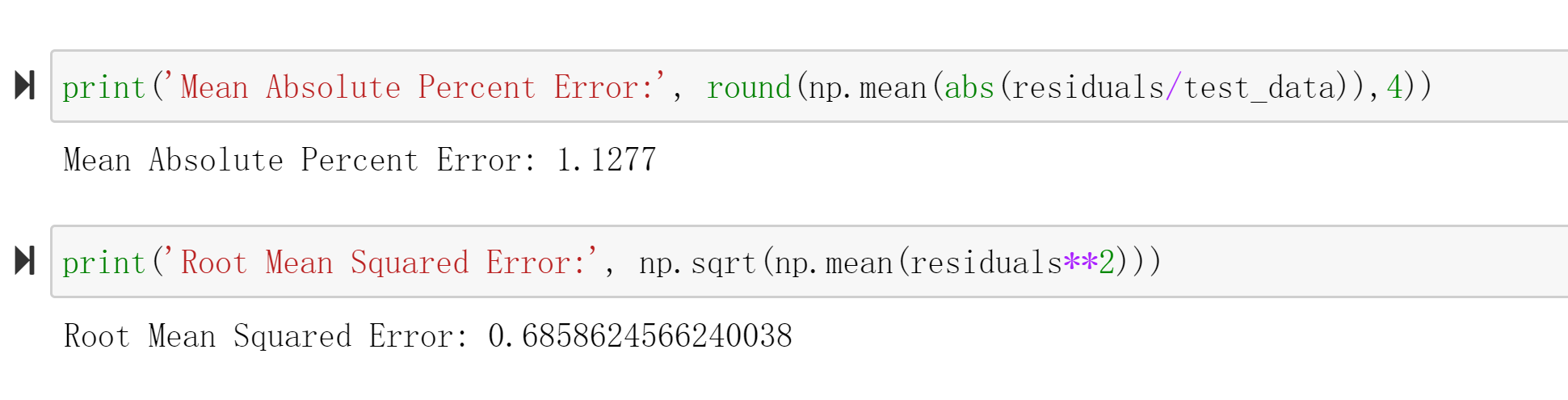
**There are 3 ways to determine the parameters, but iteration is the best approach to view all RMSE one by one. (0,0,1) provides the lowest RMSE and we apply order=(0,0,1) to the model.**

**** ****

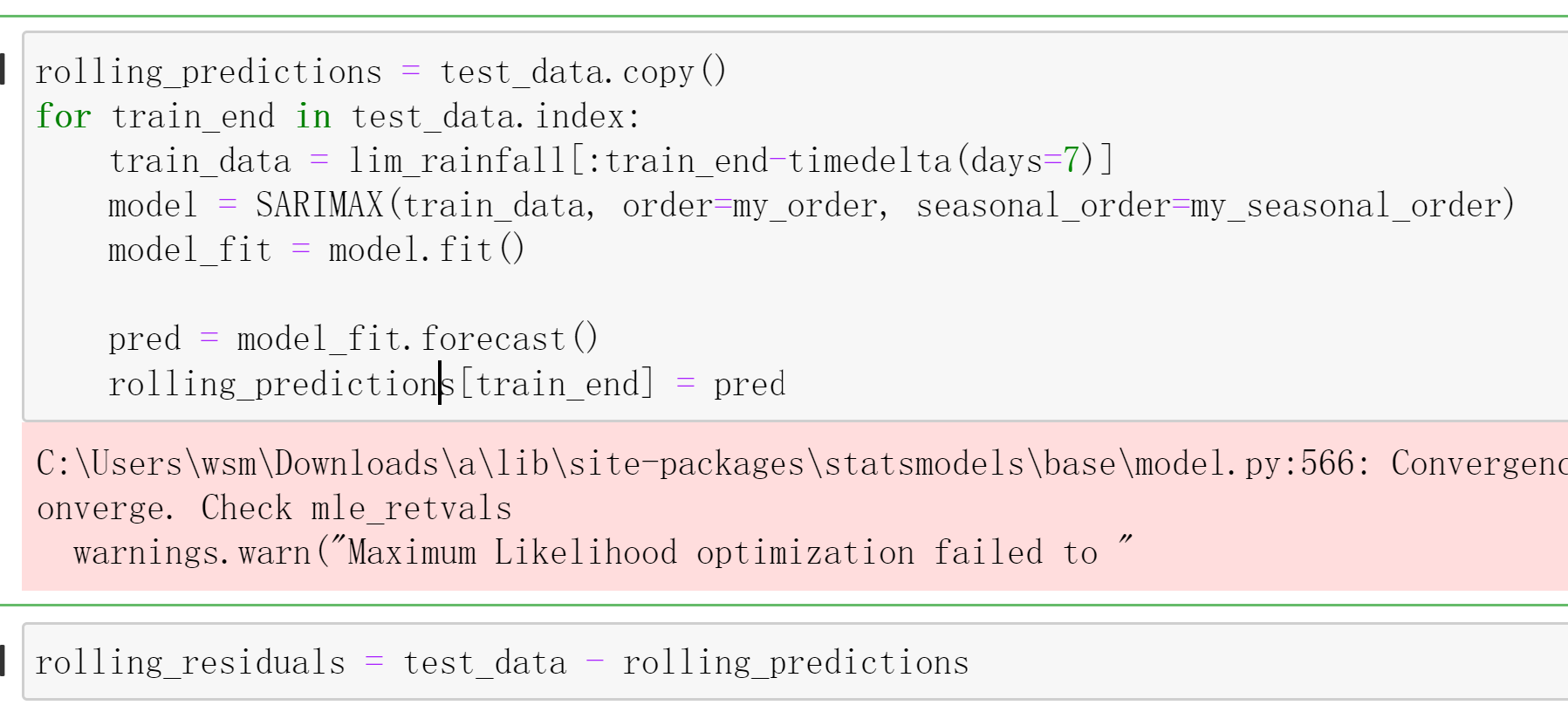
**We observe the fit model by summary and plot of residual and discovery the model fit is not very ideal. Also the residual of the first 5 months are above the 0.**

**** ****

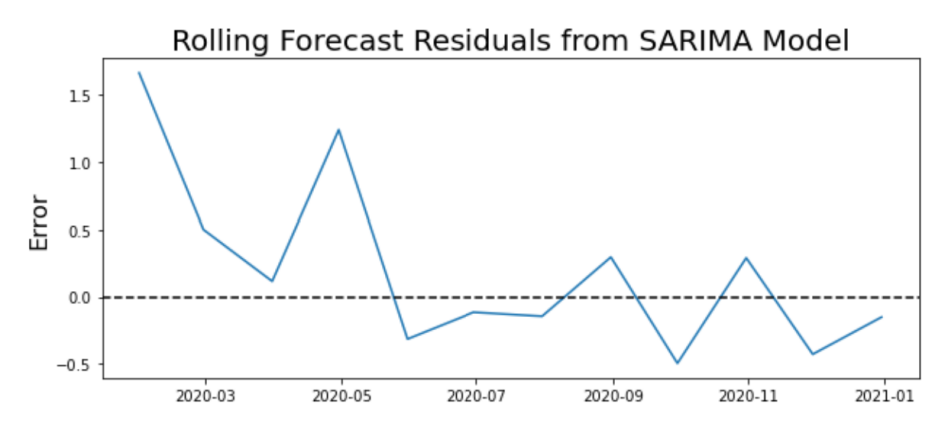
Mean Absolute Percent Error and Root Mean Square Error is quite high as well.

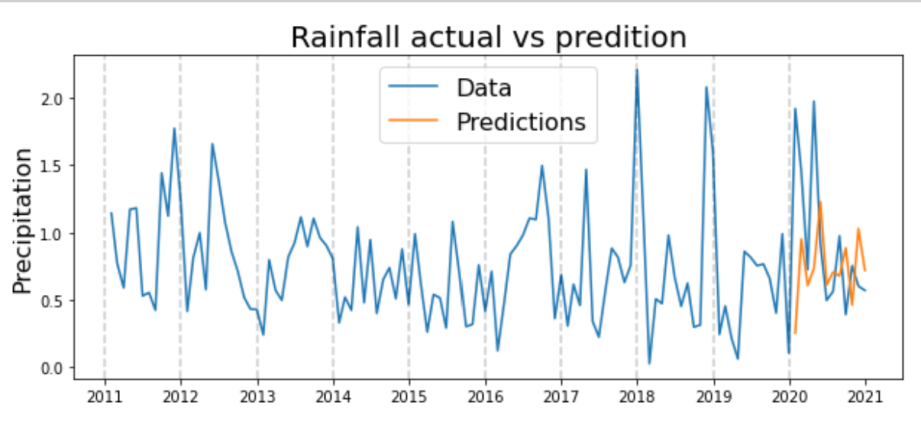


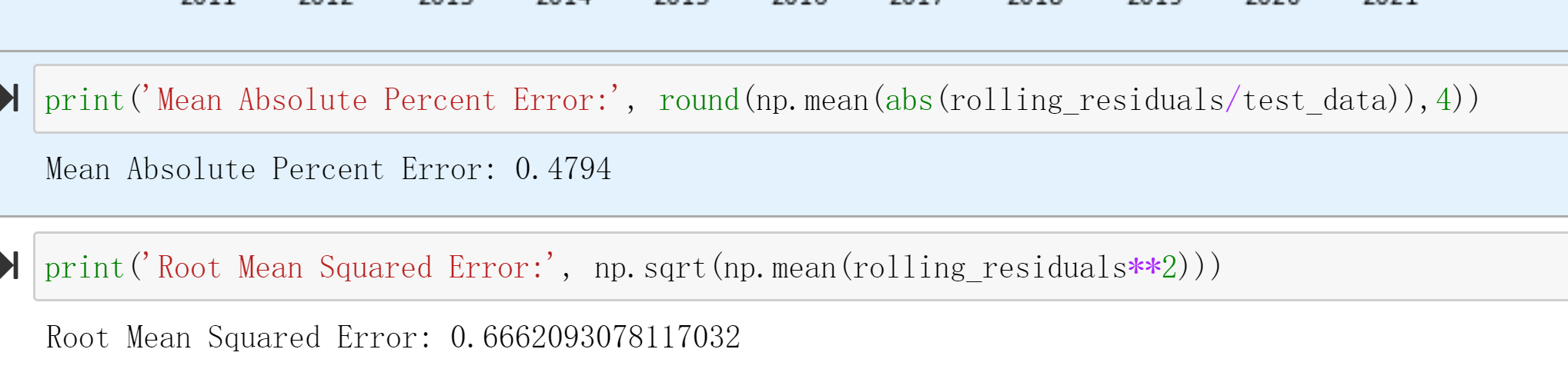
To enhance this, rolling forecast is a good way to enhance its accuracy. Rolling forecasts allow for continuous planning with a constant number of periods. For example, if forecast period lasts for 12 months as each month ends another month will be added, then it means the model always forecasting 12 months in the future.



About accuracy, rolling forecasts allow make quick tweaks along the way rather than letting mistakes count up and only giving one shot to make those changes. The residual, Prediction are shown as below. The rolling forecasts has better accuracy than previous one since its MAPE and RMSE is less than the previous one.







In SARIMA model, the aim is to predict the future trend rather than classification, so overfit is not the correct word to measure a SARIMA model, it focus on the accuracy more than whether it is overfitted.

Identify why rolling forecast doesn’t vary like actual data does. One of the reason is the training data is covered from 2011 to 2020, and we can observe that the first 5 year the seasonality is not very obvious and trends without regularity until 2017, to balance the variance so the fluctuation of prediction didn’t vary that much as actual data does, but anyone it still grasps the trend of up-side-down of how actual rainfall does with the same amount of lags.