**Additional notes related to project**

[**ARIMA** 1](#_Toc176343951)

[**Outlier vs Anomaly** 1](#_Toc176343952)

[**SARIMA** 1](#_Toc176343953)

[**Training vs Test data** 1](#_Toc176343954)

[**Deep learning – LSTM** 1](#_Toc176343955)

[**Prophet** 1](#_Toc176343956)

1. When time series forecasting is required – when data change significantly as time changes; time should impact target variable. e.g., weather prediction, stock market prediction, temperature change etc. in time series you always try to predict the future.
2. GDP - depends on time interval, by year it is time series, but by week not, change should be significant
3. Coal production will not change drastically in minute or day but in year it will change drastically
4. you can create time series data using non-time series data by feature engineering - date time data to day or month or year
5. if datetime data is given as object, change it to datetime using pd.to\_datetime, it is important to work with few libraries

In time series data analysis, you specify frequency to python and here it is month start

df.index.freq = 'MS' ## monthly time series data specifying to python (month start)

you can define it by below as well –

df.index.freq = 'Y'

df.index.freq = 'DD’

df.index.freq = 'HH’

df.index.freq = 'S’

## **ARIMA**

Autoregression means there should be time variation.

It is statistical inference approach - auto regressive integrated moving average

Regression - there is a variable which impacts another variable. Means data here should be time dependent. Then only you can use ARIMA.

Moving average thing is also there. Models calculates moving average behind the scene for the prediction.

Stride of moving average is 1 so moving average should be calculated by 1 jump. If it is 3, it should be calculated by 3 jumps.

Moving average – Average over a window, window is like 5 months, 7months etc. stride term is also there, it is similar to how many months jump

p - lag order which is also called stride

q - size of moving average,

d = order of diff - difference between normal numbers as it is or diff between moving average

Residuals – there will be time when data will not follow the seasonality then what will be the difference shown here.

## **Outlier vs Anomaly**

Outlier - An outlier is a data point that significantly deviates from the other observations in a dataset. Outliers can occur due to variability in the data or experimental errors. They are often identified through statistical methods and can impact the results of data analysis if not handled appropriately.

Characteristics of Outliers -

Extreme Values: Outliers are extreme values that lie far from the central tendency of the data.

Statistical Methods: They are often detected using statistical techniques such as the interquartile range (IQR), Z-scores, and standard deviations.

Influence on Analysis: Outliers can significantly influence the results of statistical analyses, potentially leading to misleading conclusions if not properly addressed.

Anomaly - An anomaly, on the other hand, refers to a pattern in the data that does not conform to expected behaviour. Anomalies can be indicative of novel, rare, or unexpected events. In many cases, detecting anomalies is crucial for identifying issues such as fraud, network intrusions, or system failures.

Characteristics of Anomalies -

Unexpected Patterns: Anomalies represent unexpected patterns that deviate from the norm.

Context-Dependent: The detection of anomalies is often context-dependent and requires a deep understanding of the data and the domain.

Significance: Anomalies are significant in various applications, including fraud detection, health monitoring, and network security.

outlier - 1000 age not ok, but 105 is ok 105 is anomaly, anomalies very rare some says you can remove some not, it is matter of opinion.

age 105 should have Aadhar card or not, they should be approached for that, here you can't remove anomaly.

## **SARIMA**

Similar to ARIMA, it just adds few seasonal parameters in ARIMA

additive model - used when change is linear

multiplicative model – used when change is exponential

What would be strategy to split data into train and test. You can't select randomly, there should be some strategy. Random indexing can't work here.

Which model is best?

depends on the dataset and what you are trying to predict

for 6 months prediction ARIMA would have been very good. if we compare RMSE with mean, it looks like a good model

8/136\*100 = 6% very good this loss is very low, we will be able to predict value less than or more than 6% of actual value

for cost of house +-5 is very good but not for probability of cancer.

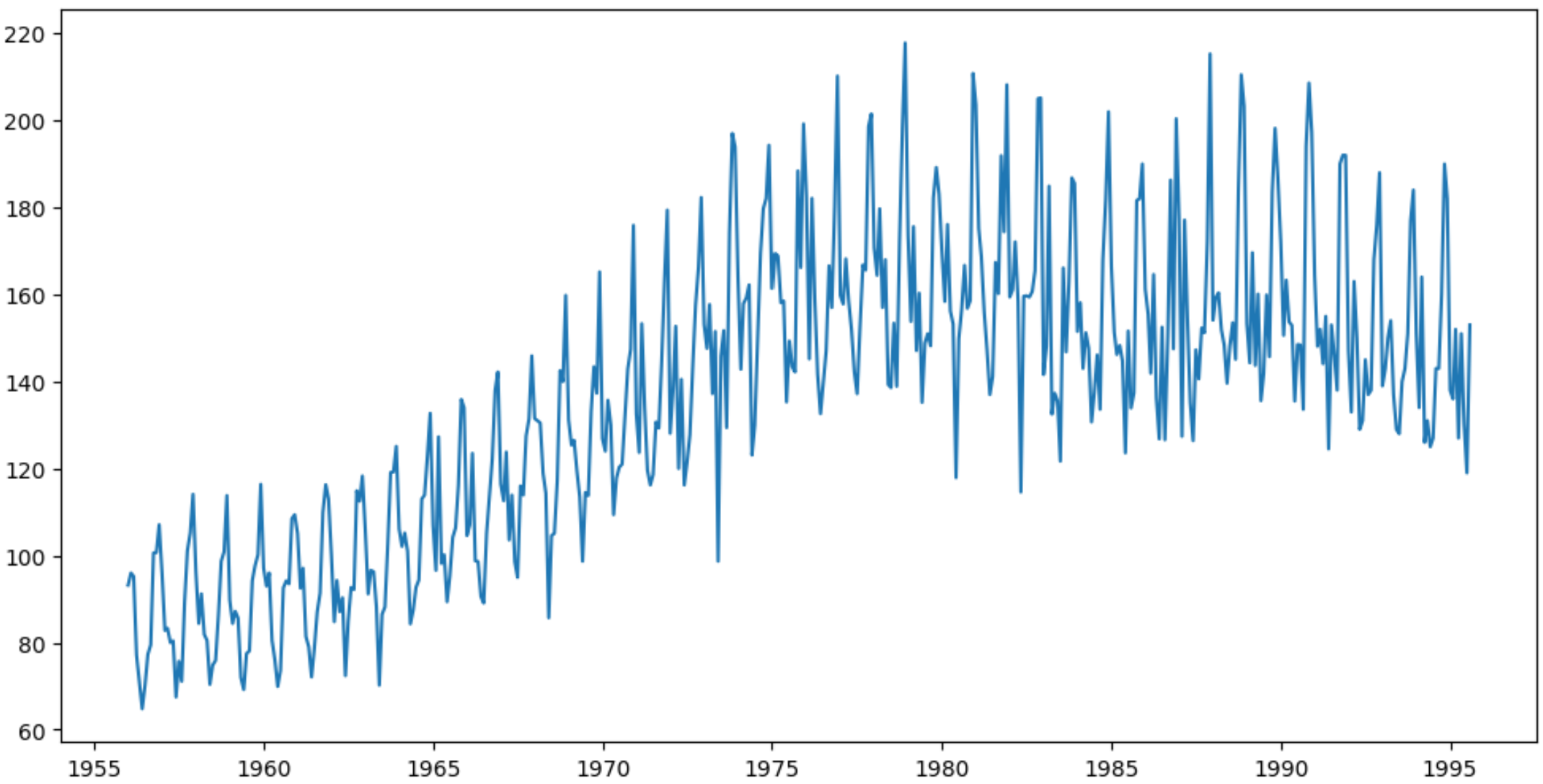
SARIMA is using Oct data to predict Nov not Sep and Oct so it just takes previous month data to predict future value not the whole historical data. There is also moving average component to this.

That is why initial values it is predicting are very well but as you move forward it deviates a lot.

## **Training vs Test data**

How do we go about training vs test data in time series modelling?

K fold will not make sense here, suppose we have 1955-1995 data, you give training data from 1975-1995, why would it predict past data? It will doesn’t work here.

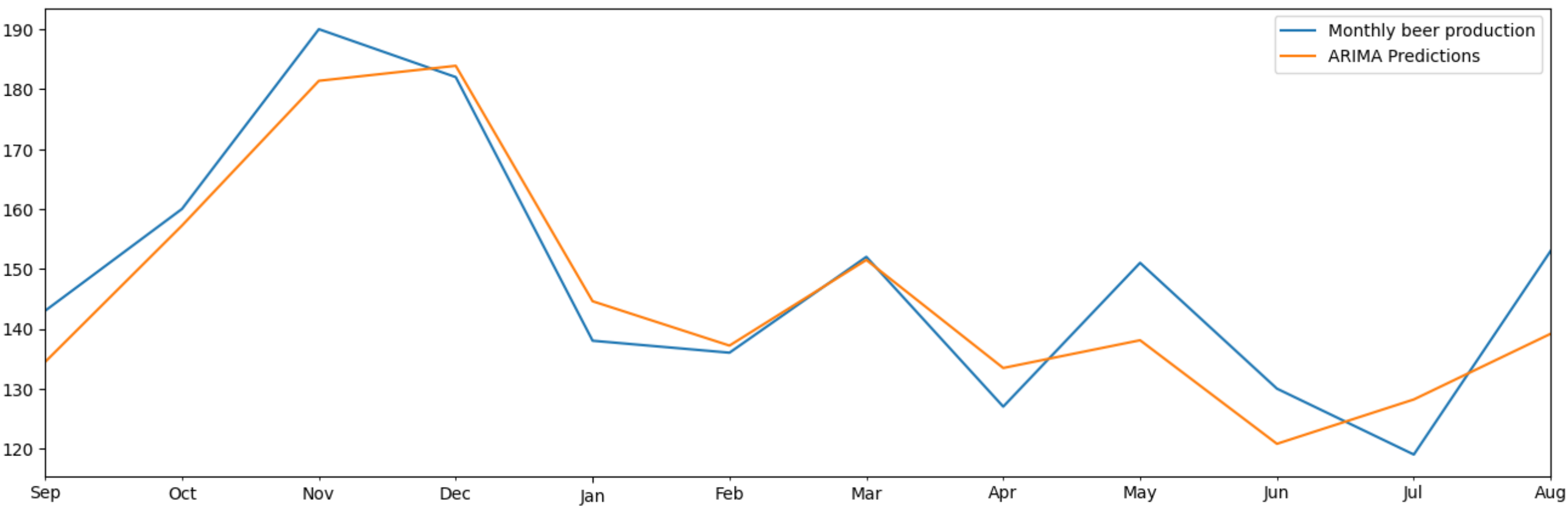


We draw an imaginary line like above and take data pre imaginary line as train, post imaginary line as validation and test.

Seasonality is repeating here in 12 months so test data also should be multiple of that.

sarima\_model = SARIMAX(train\_data['Monthly beer production'], order = (2,1,1), seasonal\_order = (4,0,3,12)) -- these seasonal orders are initial mention only. Model can correct them.

12 is overall period, 4 month is seasonality



## **Deep learning – LSTM**

LSTM – long short term memory, it takes short and long term memory both in context.

Scaling doesn’t mean only increasing, it could be reducing as well. Change your subconscious on this.

n\_input = 12 means seasonality

your training dataset has to be much much larger than test dataset because capturing the seasonality, trend etc is important.

One very good thing to do before you do any deep learning model is scaling your data

With generator, you can't do fit, you have to do fit\_generator

## **Prophet**

It is one the best non deep learning library used for time series forecasting.

Very old time series modelling libraries. It is an additive model. It is so powerful that now there are premium versions of it as well

in time series, loss is generally on the higher side

training dataset has to be very large than test bcoz capturing the seasonality etc is ...trying to keep a split of 95:5

there is no need to create validation dataset

it captures seasonality pattern more but it is off as compare to ARIMA model

Out-of-time testing, also known as out-of-sample testing, is a method used in the evaluation of predictive models, particularly in the context of time series analysis and forecasting. The idea behind out-of-time testing is to assess the performance of a model on data that was not used during the training phase. This helps to gauge how well the model generalizes to new, unseen data, and it provides a more realistic measure of its predictive capabilities.