Techfest

TATA COMMUNICATIONS

TATA Comm Datathon

Speakers:
Vedant Kaushik
Srikar Verma
Utkarsh Pandey

THE STRATEGY

1

Understanding the problem statement

2

Exploring and understanding the data

3

Developing a forecasting model

UNDERSTANDING THE PROBLEM STATEMENT

What is the Problem Statement?

- Given: the business potential of 3,915 regions over 6 years,
- Forecast: the potential over the next 15 months for each region.

What is business potential?

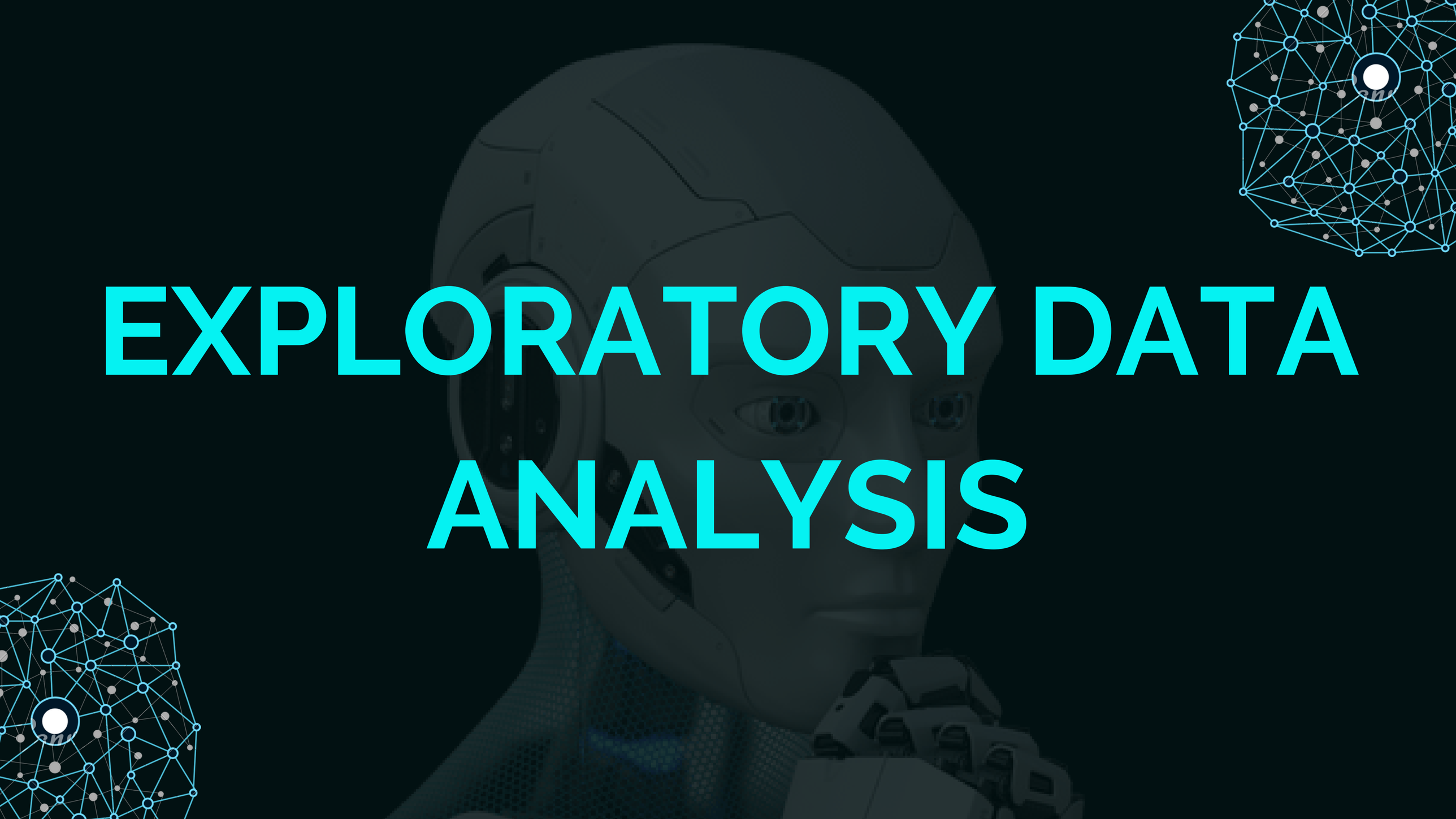
A means of quantifying the extent to which a business can grow in the future.

How to tackle the data?

Find and remove

- the potential outliers
- trends
- seasonality.

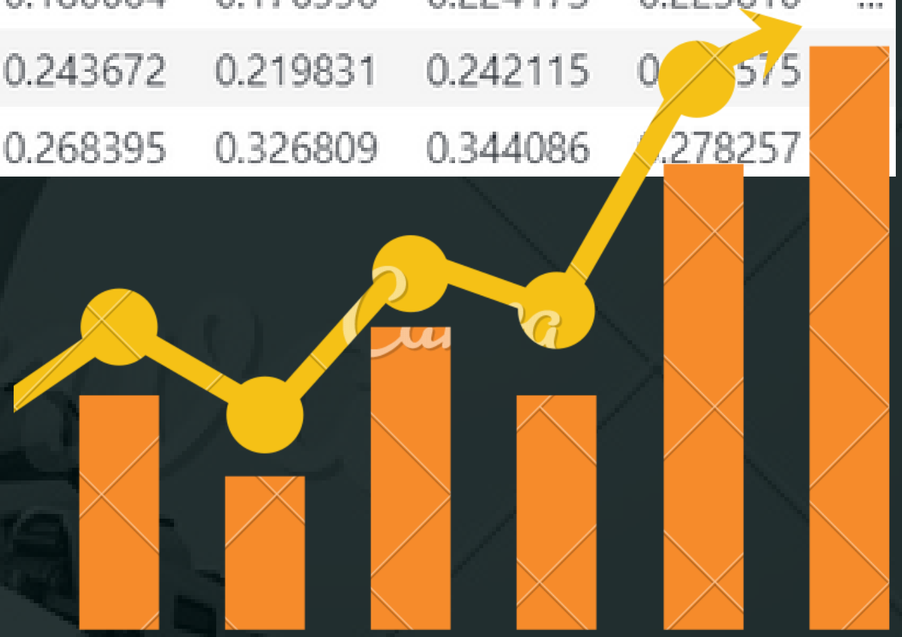
We find correlated regions with highly similar business potential curves.



EXPLORATORY DATA ANALYSIS

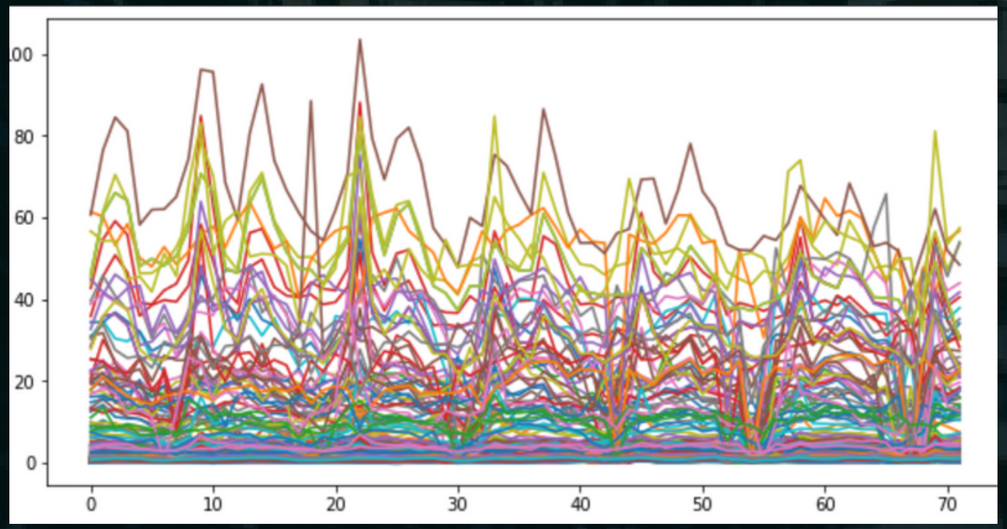
BUSINESS POTENTIAL DATA

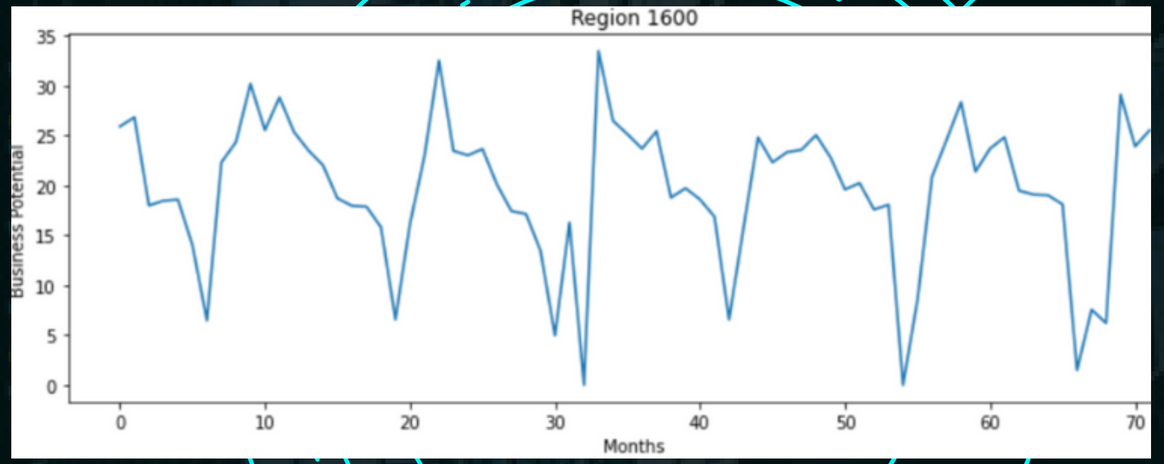
Region_Name	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10	...
Region 1	0.154325	0.122901	0.127583	0.501557	0.103551	0.051983	0.140272	0.157586	0.182546	0.181801	...
Region 2	0.119708	0.119831	0.112970	0.326049	0.098852	0.041502	0.116759	0.136964	0.178476	0.162621	...
Region 3	0.178236	0.165497	0.190754	0.321733	0.203126	0.048345	0.180064	0.176990	0.224173	0.223810	...
Region 4	0.225016	0.220602	0.224240	0.330916	0.268934	0.171322	0.243672	0.219831	0.242115	0.[illegible]575	
Region 5	0.318976	0.316835	0.328486	0.483001	0.380898	0.292074	0.268395	0.326809	0.344086	[illegible].278257	

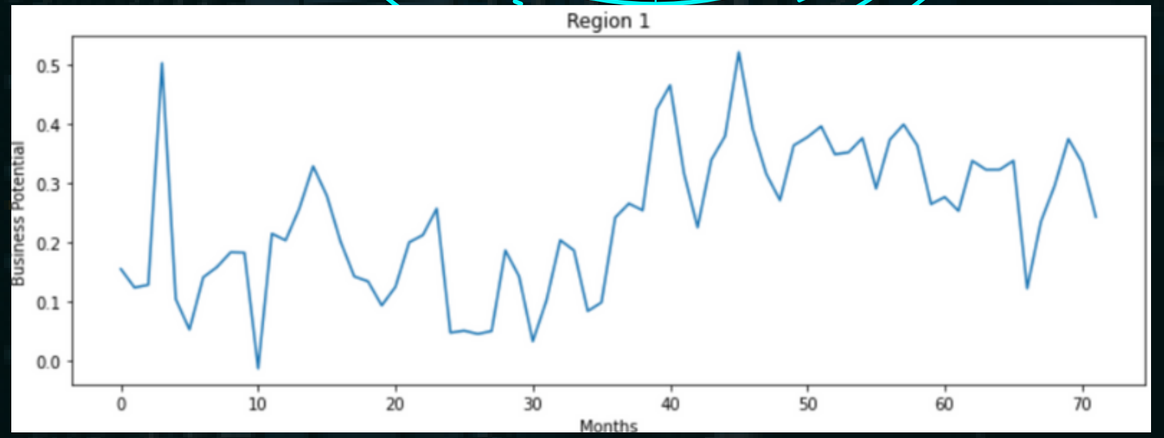


PLOTS AND INFERENCE

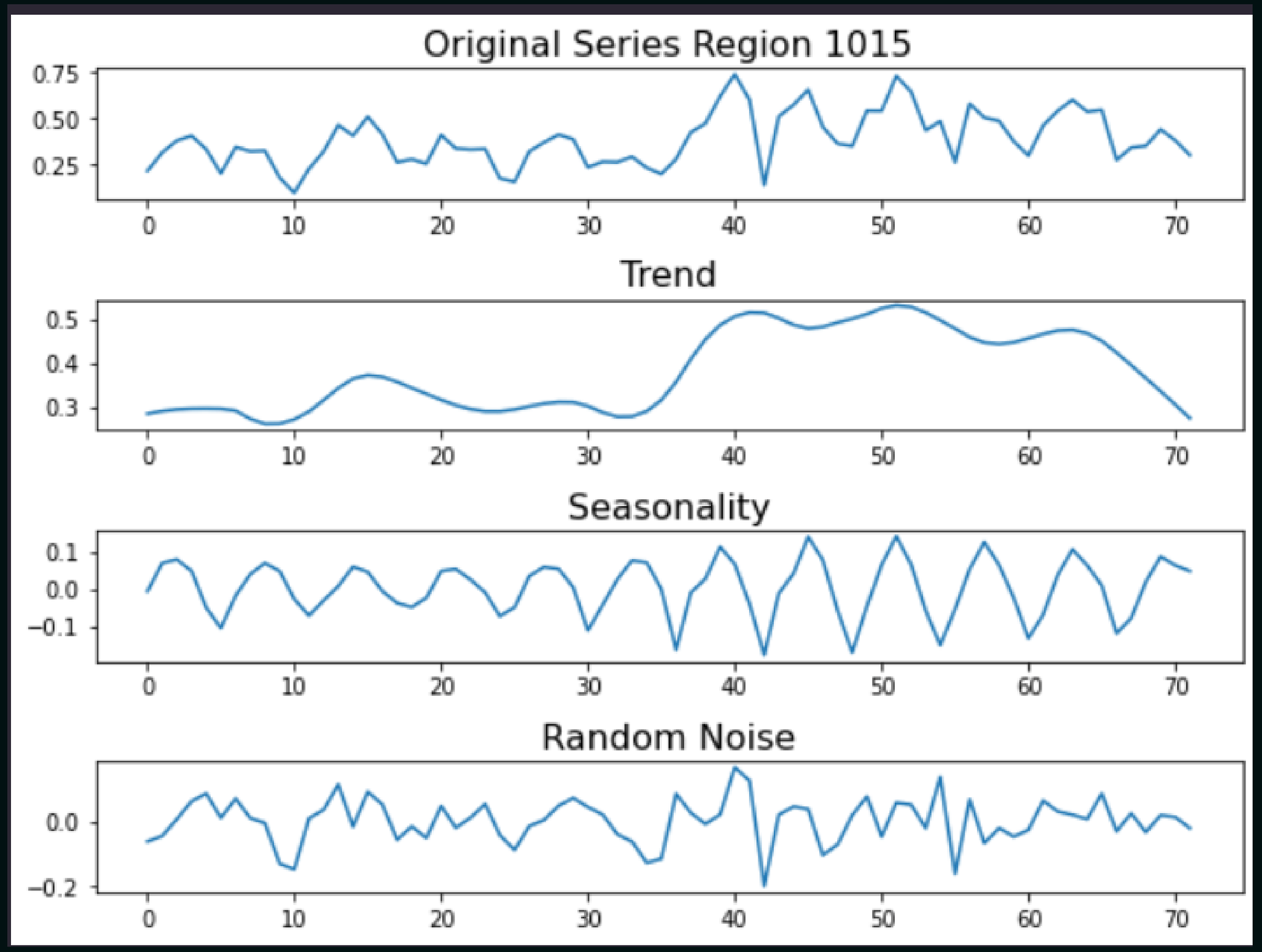
- Seasonality
- Outliers
- High regional variations
- Trends





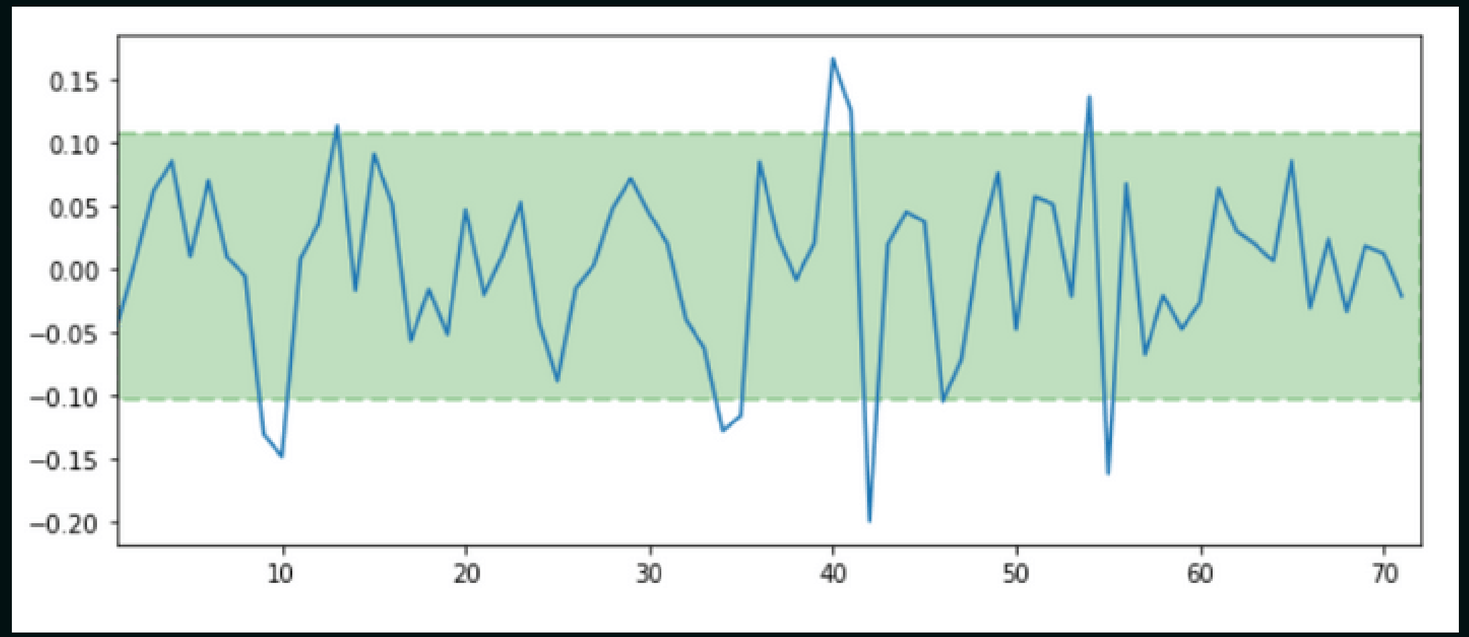


ANOMALY AND OUTLIER DETECTION

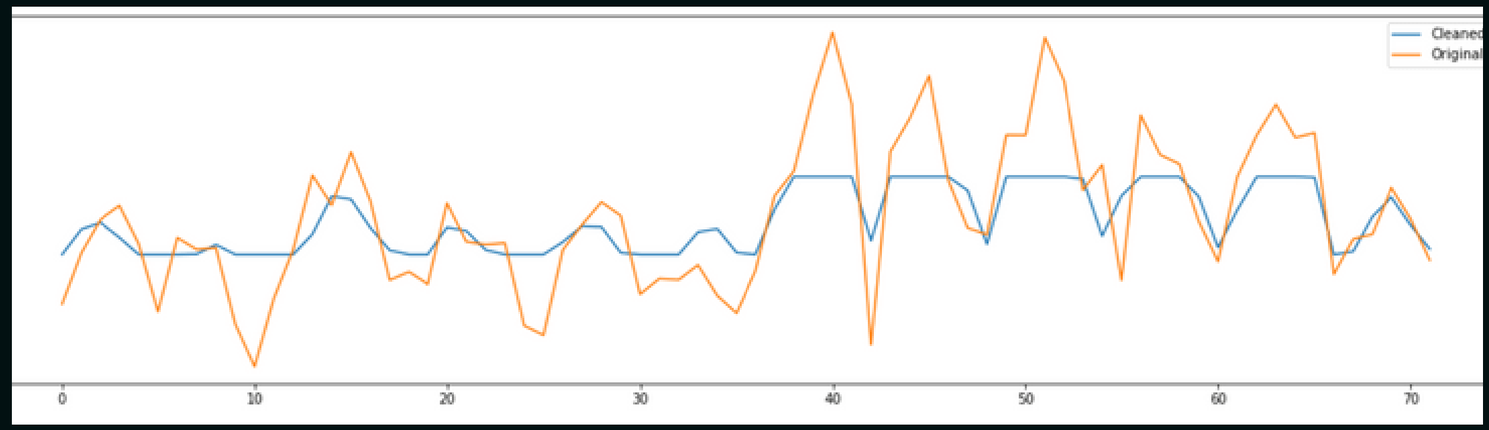


DETECTING OUTLIERS USING THE QUARTILE METHOD

- Green region shows values within 25-75% of median.
- All points outside the green region are outliers.
- Outliers are replaced with the previous value assuming Markov chain hypothesis.
- The figure below shows the plot with outliers removed(in blue).

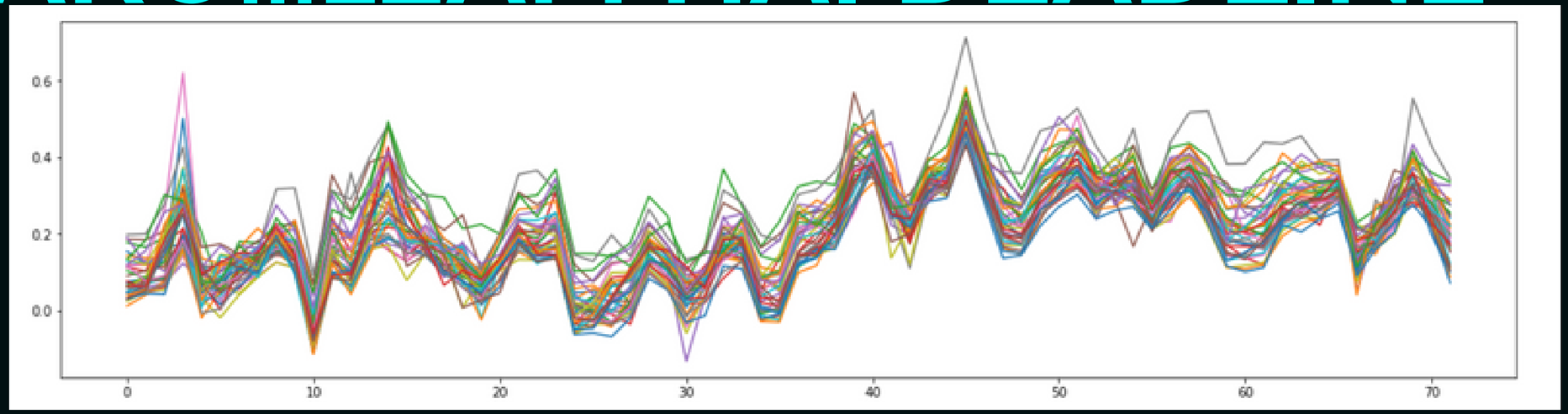


Region 1015

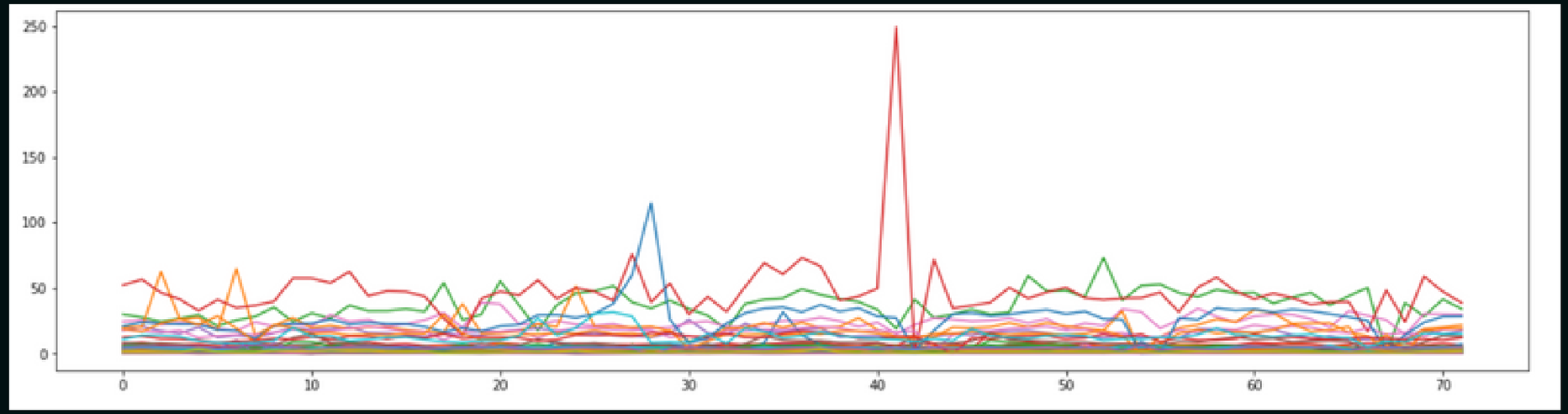


JALDI KARO...12AM HAI DEADLINE

- Regions with "HIGH" Correlation



- Regions with "LOW" correlation



SELECTING REGIONS TO TRAIN

- We found Pearson's Correlation
- Divided the highly correlated data into different regions.
- Took random Samples from the regions different regions with high correlation to train
- Then we took regions whose correlation was very low and used them to train too
- In total, we selected 1313 regions.

```
cor = dfr.iloc[:,1:].corr(method = "pearson")
corr = np.array(cor)
cor.shape
```

```
def find_corr_cols(l_thres, u_thresh):
    related = []
    unrelated = []
    for i in range(0, len(cor.iloc[0, :])):
        if ind[i] == False:
            list = []
            list.append(i+1)
            for j in range(i+1, len(cor.iloc[0, :])):
                if l_thres < corr[i, j] and corr[i, j] <= u_thresh and ind[j] == False:
                    list.append(j+1)
                    ind[i] = True
                    ind[j] = True
            if len(list) == 1:
                unrelated.append(i+1)
            else:
                related.append(list)

    return related, unrelated
```

```
related9, unrelated9 = find_corr_cols(0.9, 1)
related8, unrelated8 = find_corr_cols(0.8, 0.9)
related7, unrelated7 = find_corr_cols(0.7, 0.8)
```

PREPARING TRAINING DATA

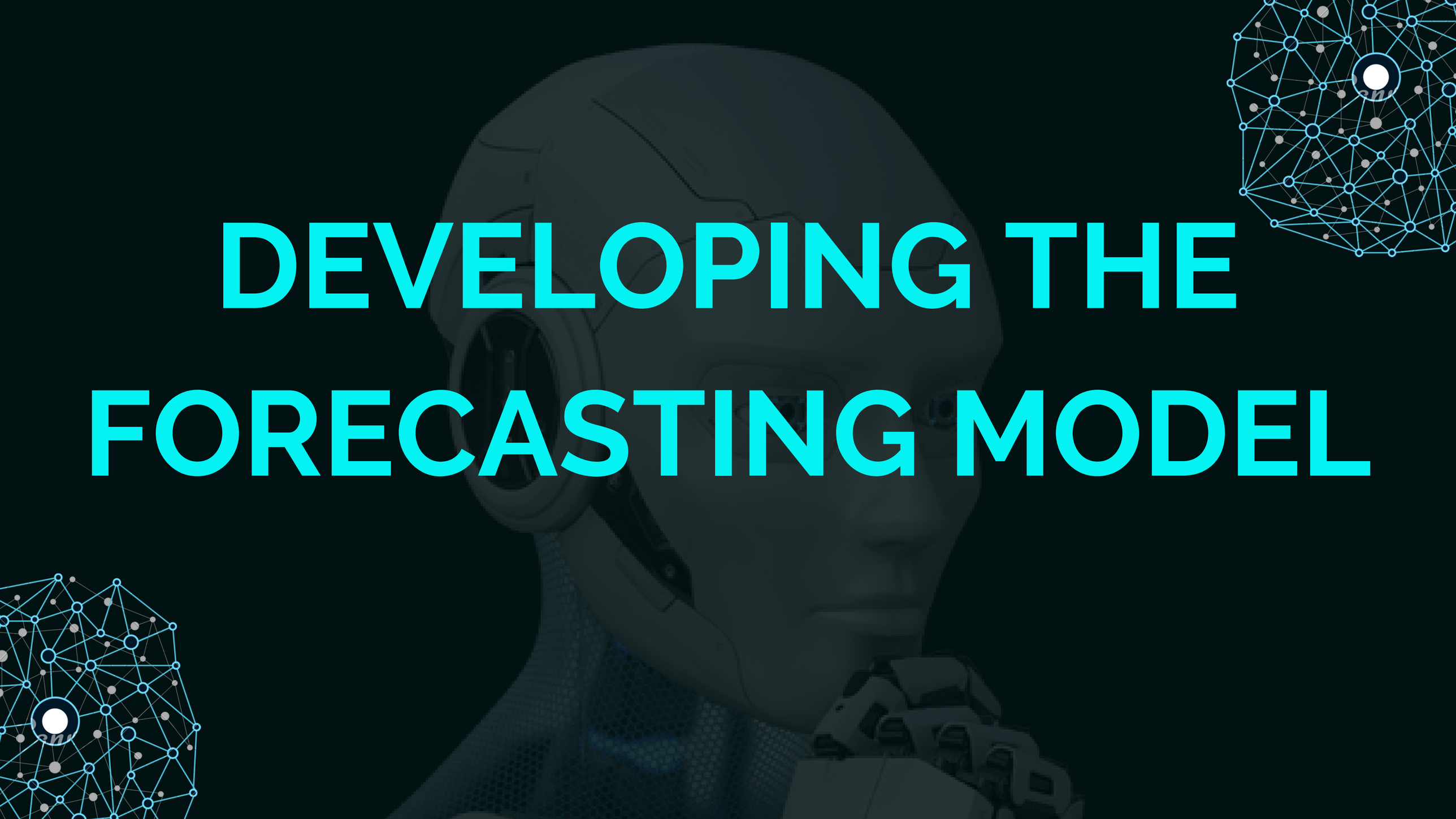
```
def prepare_data(timeseries_data, n_features):
    X, y = [],[]
    for i in range(len(timeseries_data)):
        end_ix = i + n_features
        if end_ix > len(timeseries_data) - 1:
            break
        seq_x, seq_y = timeseries_data[i:end_ix], timeseries_data[end_ix]
        X.append(seq_x)
        y.append(seq_y)
    return np.array(X), np.array(y)
```

For a given region:

- Let a_n denote the business potential for the nth month.
- Then we define a sequence A, with A_n containing {a_n, a_n+1, a_n+2, a_n+3, a_n+4, a_n+5}.
- The training data consists of the sequences A_1, A_2, A_3.....A_65, A_66, which can be written as a 2D NumPy array of shape (66, 6).
- Let B denote a sequence, with B_n defined as a_n+6. Thus, the shape of B will be (66, 1).
- B_n is the label for the sequence A_n.

- We prepare the training data of multiple regions by running the prepare_data function in a loop.
- So that the final shape of X_train is (1313, 66, 6), where N is the number of regions.
- The final shape of the prediction label is (1313, 66, 1).

```
# ls_correg = [[2, 12]]
np.random.shuffle(allrelated)
trainx= []
trainy = []
for ind,pee in enumerate(allrelated):

#     if(ind > 200):break

    region = "Region " + str(pee[0])

    X, y = prepare_data(dfr[region], 6)

    X_train = X[:]

    y_train = y[:]

    # X_train = X_train.reshape((X_train.shape[0], 6))

    trainx.append(X_train)
    trainy.append(y_train)
    mont = 15
    n_features = 1
print(len(trainx))
trainx = np.array(trainx)
trainy = np.array(trainy)
```

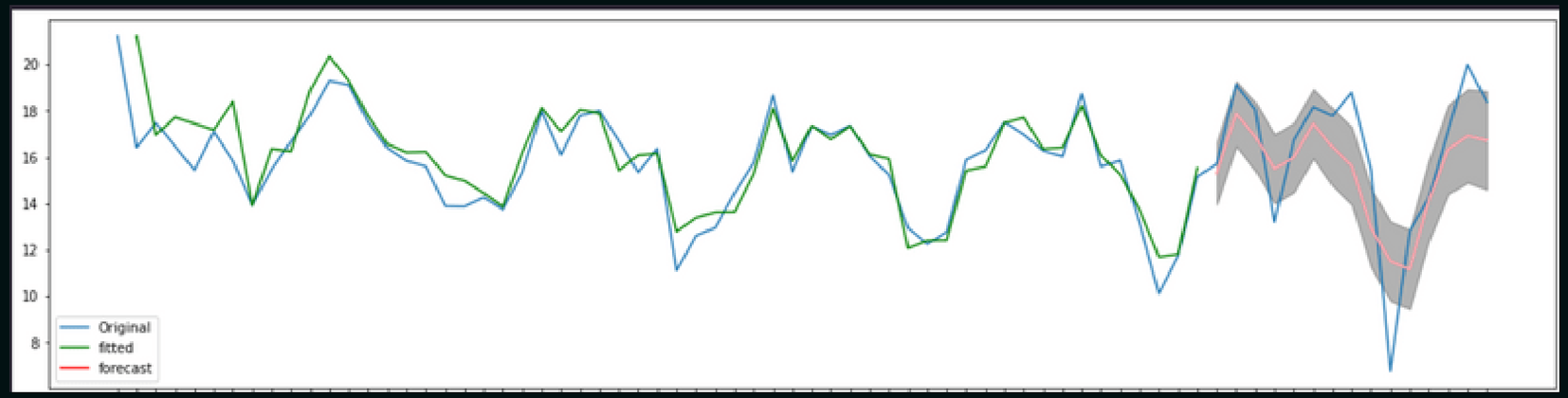

DEVELOPING THE
FORECASTING MODEL

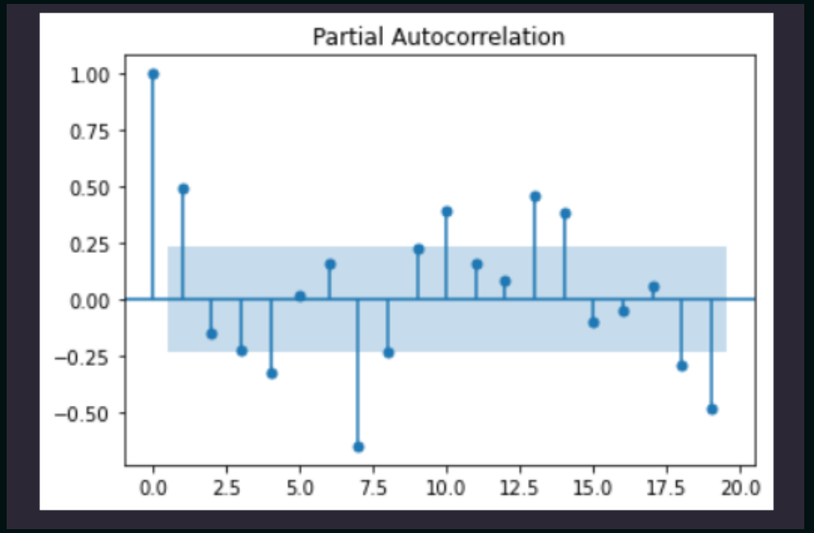
MODELS WE TRIED

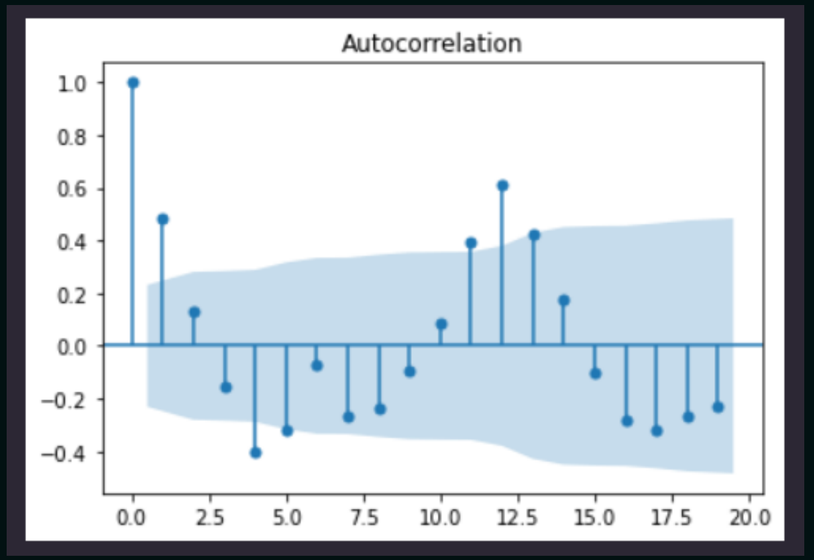
1. ARIMA
2. AUTOARIMA
3. PROPHET
4. LSTM

ARIMA

- Number of Autoregressive terms(p) using Partial Autocorrelation
- Number of non-seasonal differences(d) needed for stationarity taken by calculating p-values.
- Number of Lagged forecasts(q) using Autocorrelation







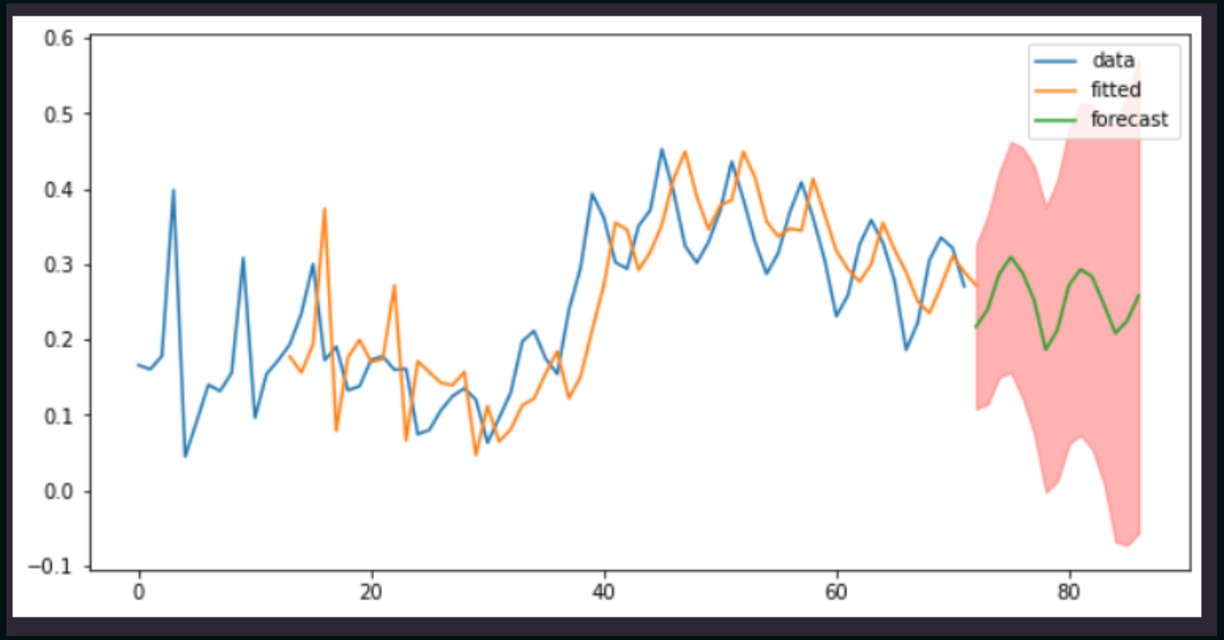
AUTOARIMA

AutoArima using pmdarima

Best values for

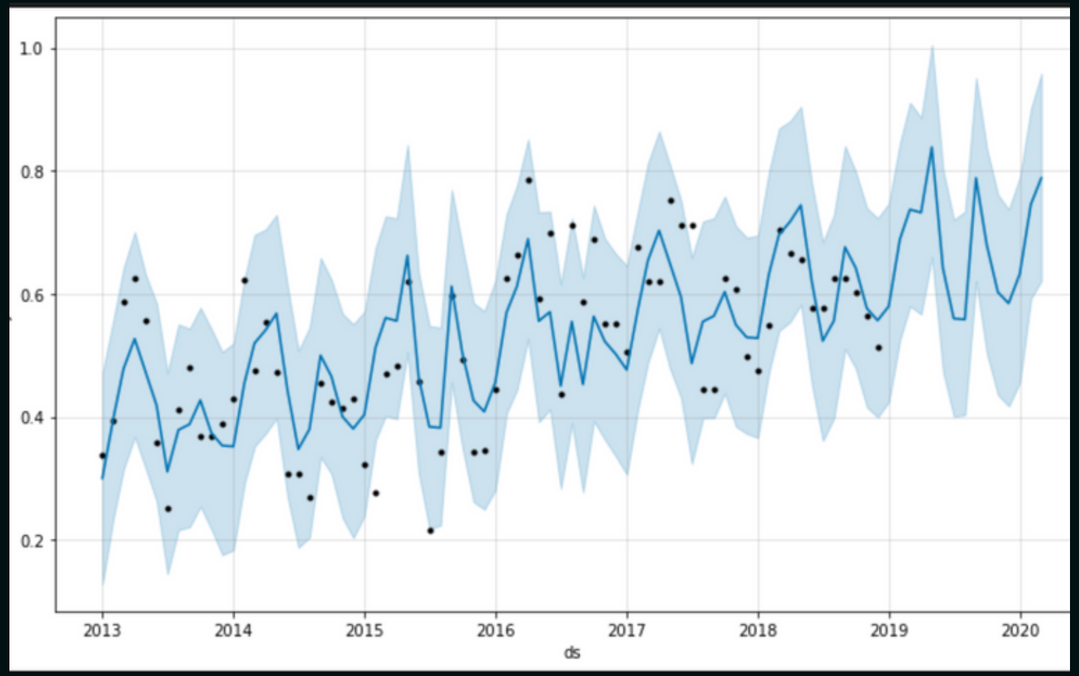
1. Autoregressive terms(p)
2. Non-Seasonal Difference(d)
3. Lagged Forecasts(q)

selected by trying different values in a matrix.



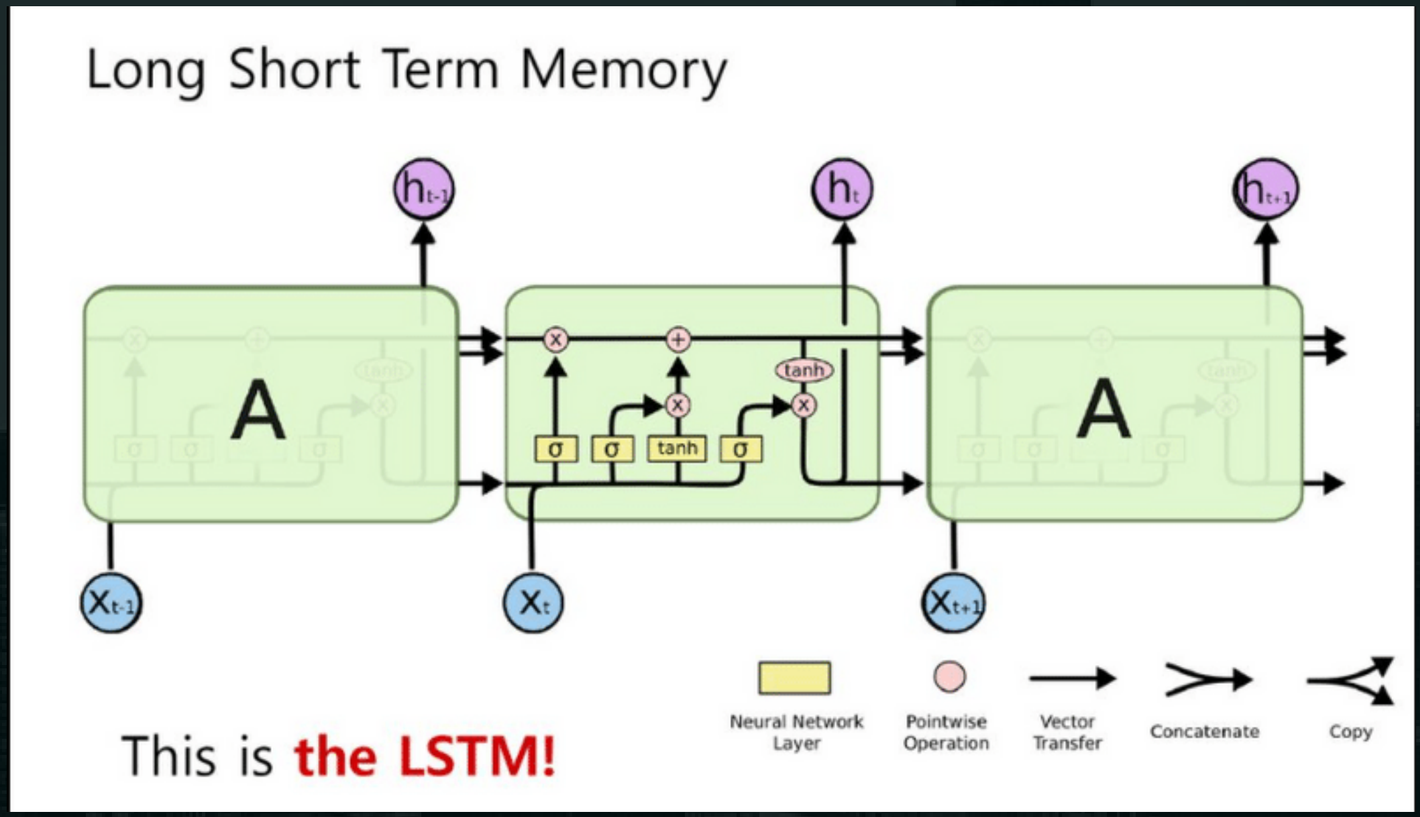
PROPHET

- Black dots ⟶ data points.
- The blue curve is the fitting and prediction made using Prophet, for the region 1067
- Blue shaded region is uncertainty band; it depicts the max uncertainty from the prediction.
- Interval width taken : 0.95
- Frequency of predicting future dataframe set to monthly.



LSTM

Long-Short Term Memory, as the name says, makes future prediction taking into account the immediate past features and also the old features.



Different LSTM Variations

- **Bidirectional LSTM** - It may be beneficial to make the LSTM model to learn the input sequence both forward and backwards and concatenate both interpretations.

- **Conv LSTM** - The ConvLSTM was developed for reading two-dimensional spatial-temporal data, but can be adapted for use with univariate time series forecasting.

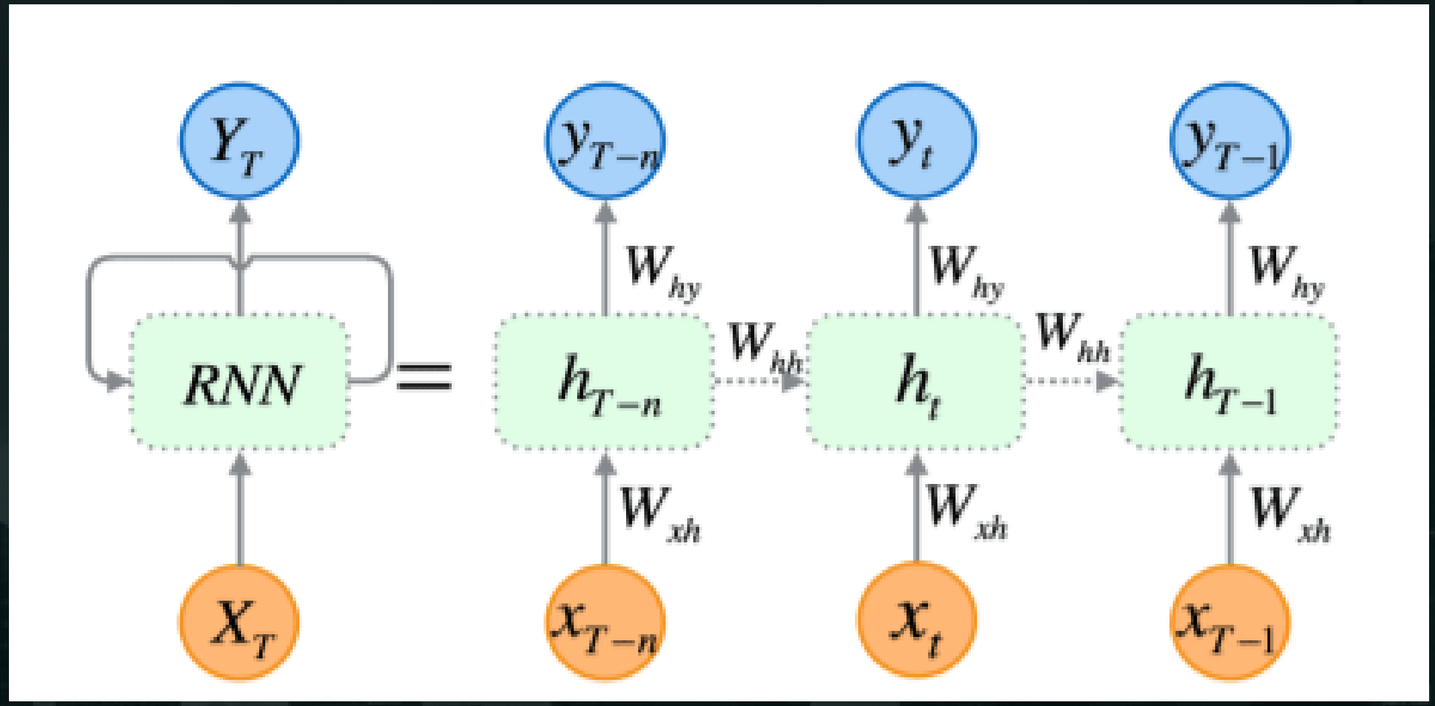
- **Stacked LSTM**

- Changing the **sample size**(time interval), the LSTM model uses to do forecasting
- On the basis of the **seasonality-period**

We can use the same output layer or layers to make each one-step prediction in the output sequence. This can be achieved by wrapping the output part of the model in a **TimeDistributed wrapper**.

STACKING LSTM

- We stacked layers of LSTM so that we can customize our model more.
- Also, since we make the next prediction on the basis of the previous 6 months' data, a deeper network is needed.



MODEL SUMMARY

We used normalized the data before feeding it into the LSTM, because LSTM models dont train good with big values

Layer (type)	Output Shape	Param #
lstm_10 (LSTM)	(None, None, 1024)	4222976
lstm_11 (LSTM)	(None, None, 1024)	8392704
lstm_12 (LSTM)	(None, None, 512)	3147776
lstm_13 (LSTM)	(None, None, 256)	787456
lstm_14 (LSTM)	(None, None, 256)	525312
lstm_15 (LSTM)	(None, None, 256)	525312
lstm_16 (LSTM)	(None, 128)	197120
dense (Dense)	(None, 128)	16512
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 64)	8256
dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 32)	2080
dropout_2 (Dropout)	(None, 32)	0
dense_3 (Dense)	(None, 32)	1056
dense_4 (Dense)	(None, 1)	33

Total params: 17,826,593
Trainable params: 17,826,593

HYPER-PARAMETER TUNING

Choosing the optimizer

Tried - Adam, SGD, Adagrad, Nadam
Used - Adam

Setting the learning rate

Selected by Random Sampling.
Used - 0.001

Adding dropout and regularizer

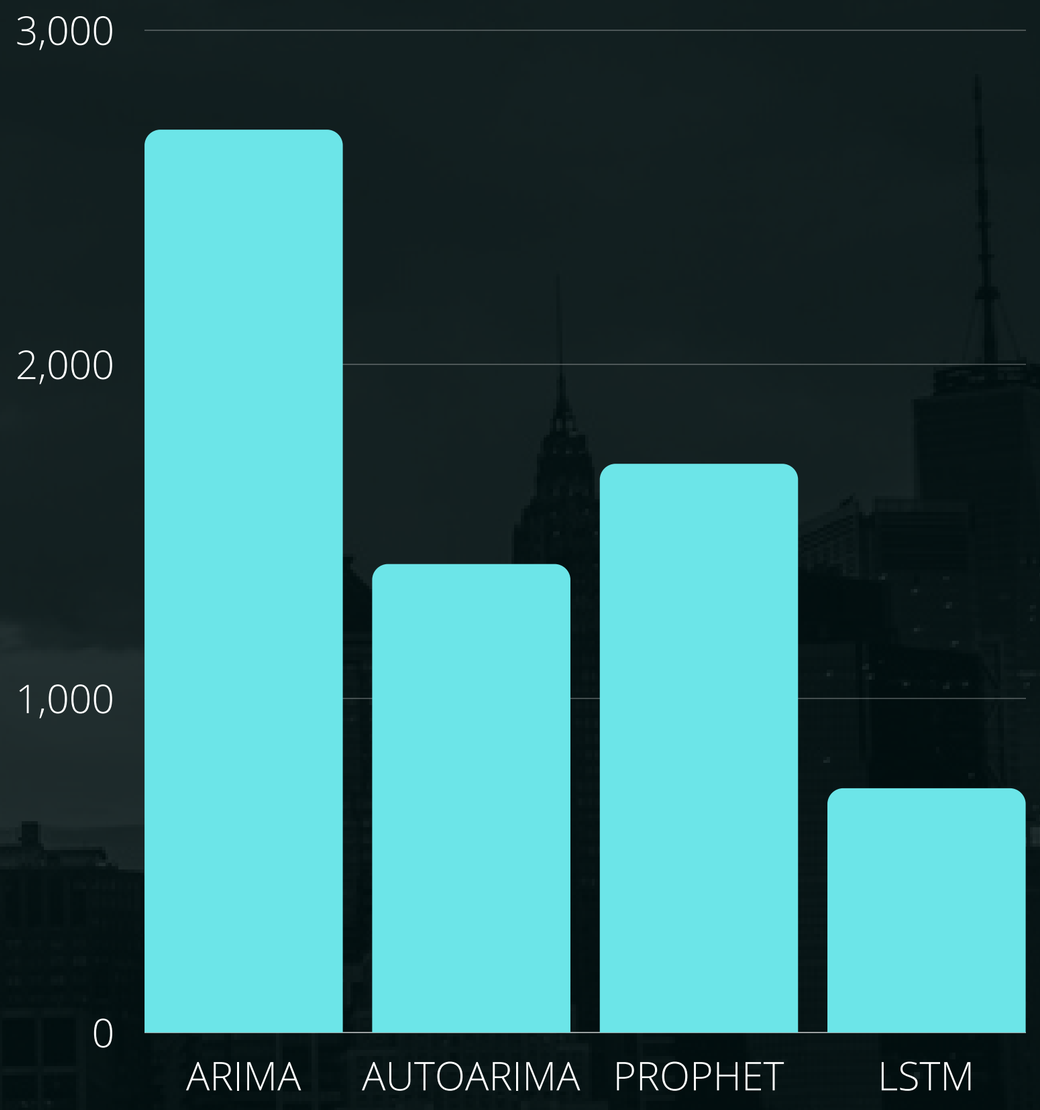
To avoid overfitting and vanishing and exploding gradients

Number of epochs

Selecting the one that gave best score

FINALIZING THE MODEL

After trying all the models - ARIMA, AutoARIMA, Prophet, and LSTMs, we obtained the following scores on submission. The lowest (thus, the best) score was given by the LSTM model, which encouraged us to finalize it.



Techfest

TATA COMM DATATHON

Thank You

Utkarsh Pandey

Vedant Kaushik

Srikar Verma