TATA STEEL

SEASON #2

Advance Analytics – Predictive Model

Presented By:
Kirty Goyal
Branch: Computer Science
College: College of Engineering and Technology ,Bhubaneswar

TATA STEEL

INTRODUCTION

women of mettle THE SCHOLARSHIP PROGRAMME SEASON #2

- Tata Steel buys nearly 13 million tons of coking coal annually and a major fraction of this is FOB Australia coal.

- The aim of the project is creating **monthly forecasts** – for a maximum of **twelve months** into the future.
- **Time series forecasting** is the use of a model to predict future values based on previously observed values.

TATA STEEL

COKING COAL DATASET

women of mettle THE SCHOLARSHIP PROGRAMME SEASON #2

This graph represents the dataset for the “Coking Coal Price Prediction” from years “1971-2018”.

- Total number of values : **12355**
- The number of missing dates : **4942**
- **Range()** and **interpolate()** function is used to fill the missing values.

After filling missing dates

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 17297 entries, 1971-01-01 to 2018-05-10
Freq: D
Data columns (total 1 columns):
DEXUSUK    17297 non-null float64
dtypes: float64(1)
memory usage: 910.3 KB
```

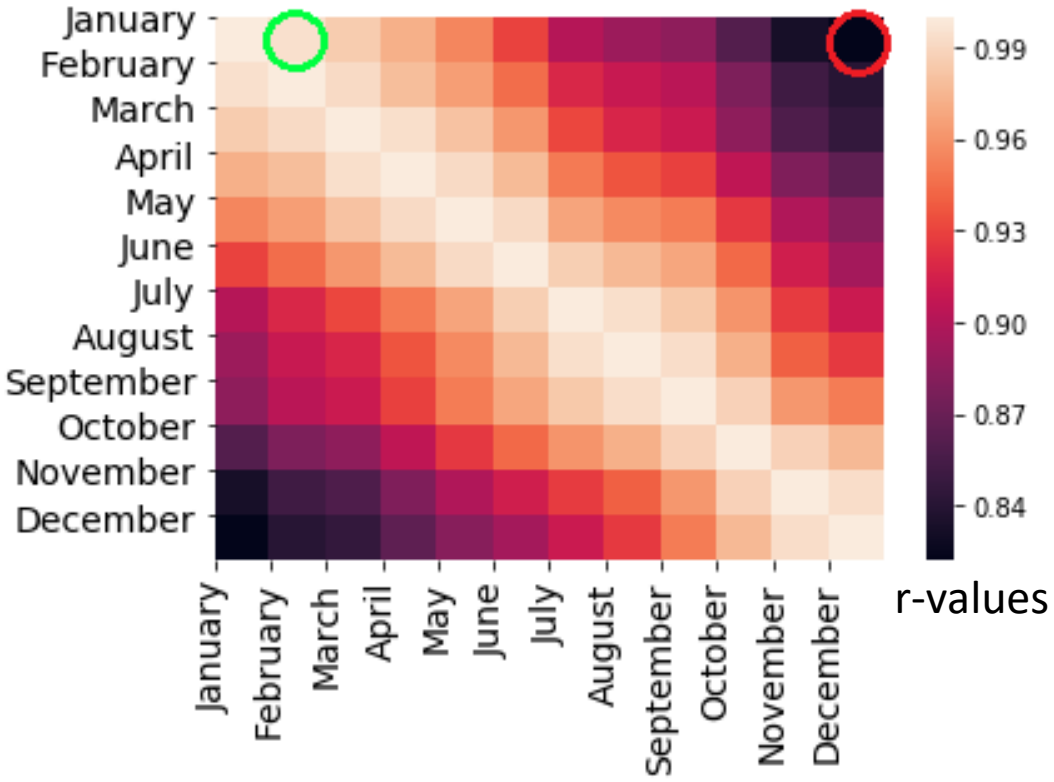
TATA STEEL

DATA VISUALIZATION

women of mettle THE SCHOLARSHIP PROGRAMME SEASON #2

month Year	1	2	3	4	5	6	7	8	9	10	11	12
1971	2.405306	2.417670	2.418794	2.417697	2.418616	2.418887	2.418500	2.434111	2.468760	2.489406	2.493385	2.526697
1972	2.570852	2.603583	2.617787	2.610347	2.612390	2.572123	2.443723	2.450148	2.441392	2.394849	2.351167	2.345355
1973	2.355758	2.424311	2.471742	2.483657	2.530929	2.576414	2.539668	2.475855	2.421043	2.428635	2.388440	2.317806
1974	2.225565	2.274170	2.339452	2.389333	2.413645	2.389700	2.389748	2.346444	2.316741	2.332990	2.325456	2.329837
1975	2.361471	2.393898	2.417558	2.369400	2.320169	2.278792	2.184587	2.114274	2.084183	2.056500	2.048818	2.021911
1976	2.028460	2.025891	1.939694	1.846057	1.804787	1.763817	1.784994	1.782635	1.730987	1.638448	1.637547	1.677768
1977	1.711874	1.709714	1.717390	1.718980	1.718345	1.719120	1.722623	1.739629	1.742900	1.771035	1.816815	1.858117

Visualizing the dataset month-wise.



CROSS CORRELATION

- This **Heatmap** represents the correlation between months.
- We conclude that for a given dataset, a particular month is most correlated to the following month.
- For example: '**January**' is least correlated to **'December'** and most correlated to **'February'**.

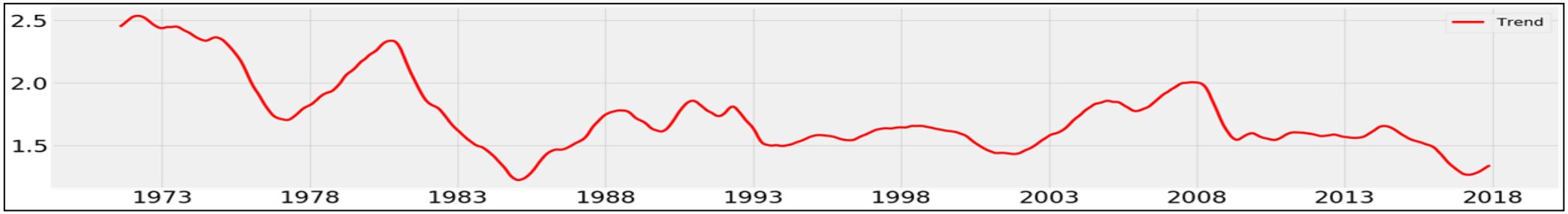
TATA STEEL

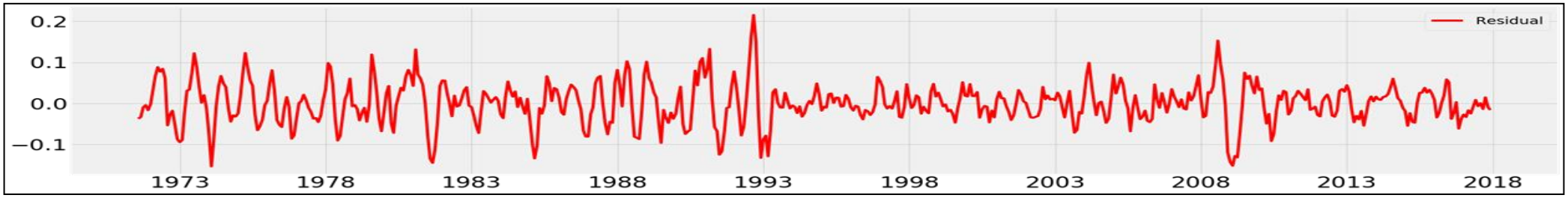
women of mettle
THE SCHOLARSHIP PROGRAMME
SEASON #2

MANN-KENDALL TEST

After performing “Mann-Kendall Test” to predict the trend in the dataset, result was: “There is a trend and it is **decreasing** from left to right”.

DECOMPOSITION

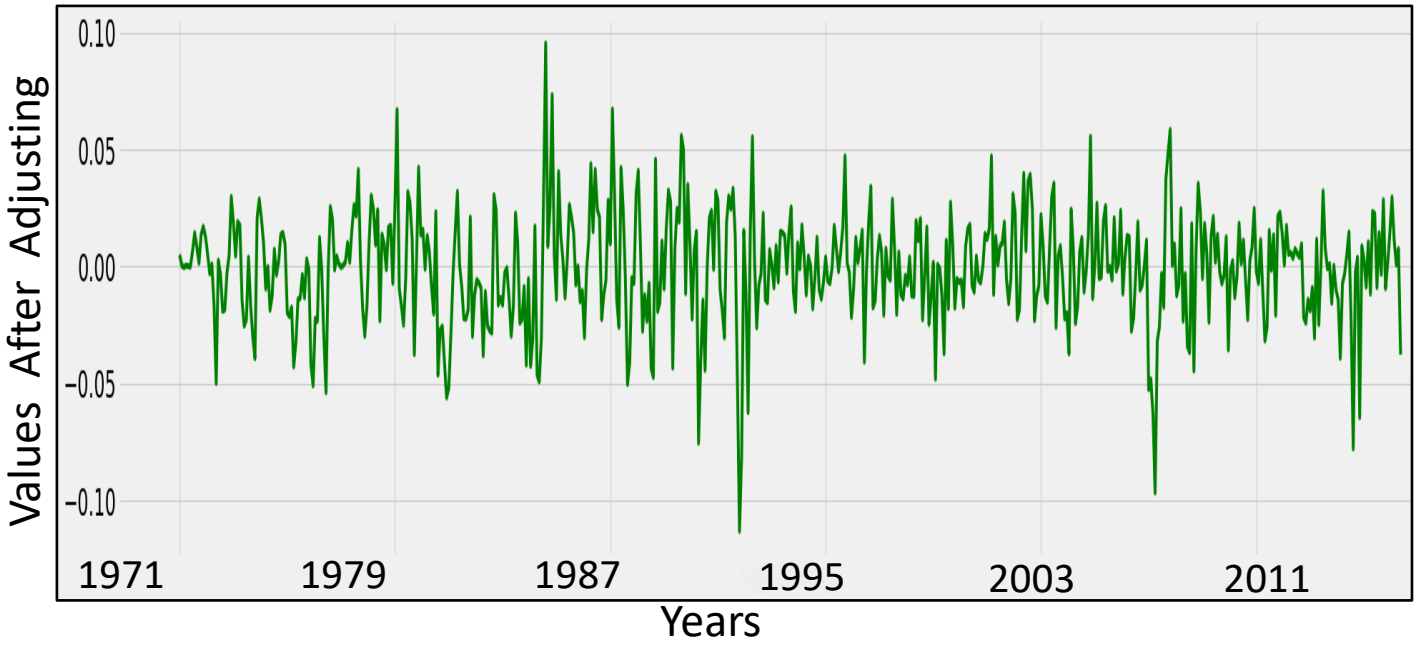




TATA STEEL

Differencing of the dataset

- After performing logarithm and first differencing, the series looked like this.
- It can be seen that it has constant mean and variance.



Augmented Dicky Fuller Test

```
ADF Statistic: -11.335895
p-value: 0.01
Critical Values:
        1%: -3.442
        5%: -2.867
        10%: -2.570
```

- This is the test which decide the **stationarity** of the time series.
- The **ADF Statistic** is much less than the 1% critical value.
- The series is now stationary with more than **99%** confidence.

TATA STEEL

UNIVARIATE TIME SERIES MODELS

women of mettle THE SCHOLARSHIP PROGRAMME SEASON #2

❑ **Autoregressive Model (AR):**

An AR model is one in which Y_t depends only on its own past values Y_{t-1}, Y_{t-2},Y_{t-3},etc

Thus,

$Y_t=f(Y_{t-1}, Y_{t-2},Y_{t-3},.....,\mathcal{E}_t)$

Where $\mathcal{E}_t$ is the error term at time t.

❑ **Moving Average Model (MA):**

A MA model is one when Y_t depends only on the error terms which follow a random process.

$Y_t=f(\mathcal{E}_{t-1}, \mathcal{E}_{t-2}, \mathcal{E}_{t-3},.....)$

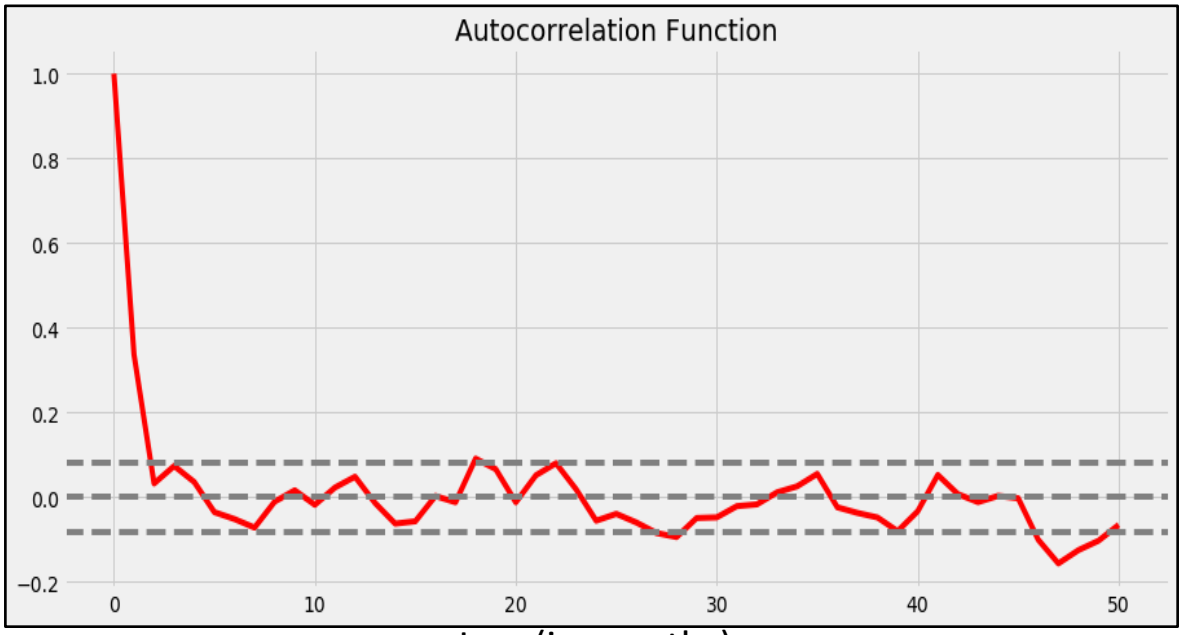
❑ **AutoRegressive Moving Average Model (ARMA):**

There are situations where the time-series may be represented as a mix of both AR and MA models referred to as ARMA(p,q).

❑ **AutoRegressive Integrated Moving Average Model (ARIMA):**

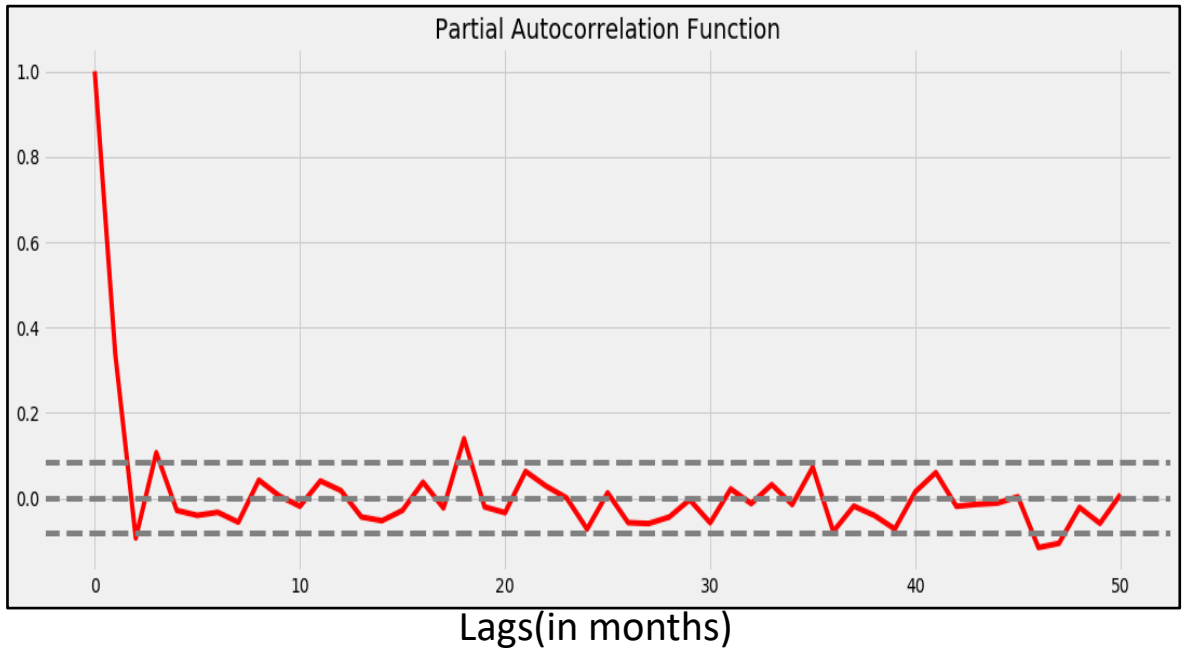
It is a generalization of an ARMA model where an initial differencing step can be applied one or more times to eliminate the non-stationarity.

ACF and PACF PLOTS



By looking at the autocorrelation function(**ACF**) and partial autocorrelation (**PACF**) plots of the differenced series, one can identify the numbers of AutoRegressive (AR) and Moving Average (MA) terms that are needed.

Here, AR =2, MA=2.

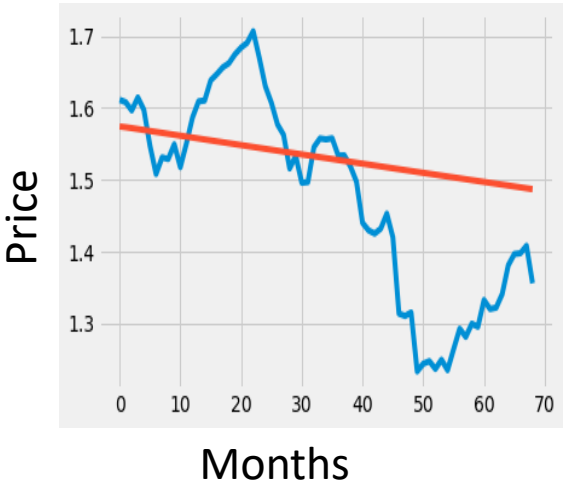
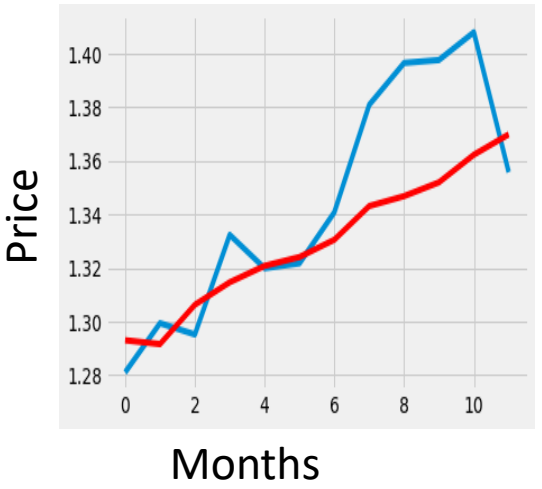
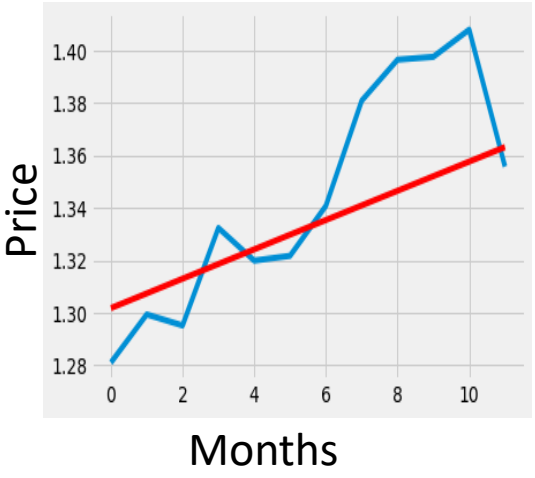


MODEL	ACF	PACF
AR(p)	Spikes decay towards zero	Spikes cutoff to zero
MA(q)	Spikes cutoff to zero	Spikes decay towards zero
ARMA(p,q)	Spikes decay towards zero	Spikes decay towards zero

TATA STEEL

Observations of Different Models

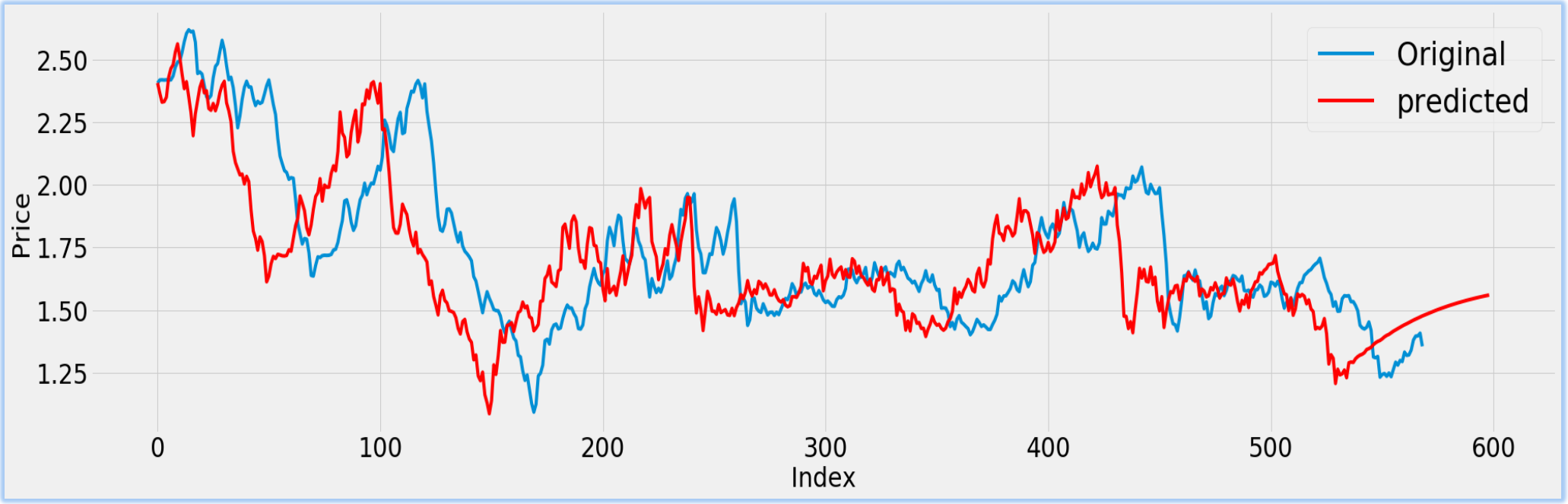
women of mettle THE SCHOLARSHIP PROGRAMME SEASON #2

MODEL	ARIMA(2,1,2)	AR(2)	ARMA(2,2)
RMSE= $\sqrt{\frac{\sum_{n=1}^{n=N}(P_n - O_n)^2}{N-1}}$	0.13	0.027	0.0287
AIC= $-2\log L_i + 2p_i$	-2643.1	-2648.4	-2644.9
Graphs: Predicted Vs Original			

TATA STEEL

AR Model's Out of Sample Forecast

women of mettle THE SCHOLARSHIP PROGRAMME SEASON #2



Predictions of next Twelve Months

Months	June	July	August	September	October	November
Prediction	1.375488	1.379287	1.385578	1.393574	1.399366	1.403430

Months	December	January	February	March	April	May
Prediction	1.409722	1.416087	1.421297	1.426215	1.431317	1.436865

TATA STEEL

CRITICAL CASE QUESTIONS (PART I)

women of mettle THE SCHOLARSHIP PROGRAMME SEASON #2

1. The aim of the project is creating monthly forecasts – for a maximum of twelve months into the future. What are the critical process, methods, Model type to be implemented to get the solution in place?

According to the analysis, the critical process involved are:

- Pre-processing (includes the filling of missing data)
- Cross-Correlation between different months.
- Performing Mann-Kendall Test.
- Taking logarithm of the series and differencing it.
- Performing Augmented Dicky Fuller Test.
- Training different models (AR,ARMA,ARIMA).

The best model is the AR(2) model.

2. Software solutions, platform needed for the seamless execution of the model?

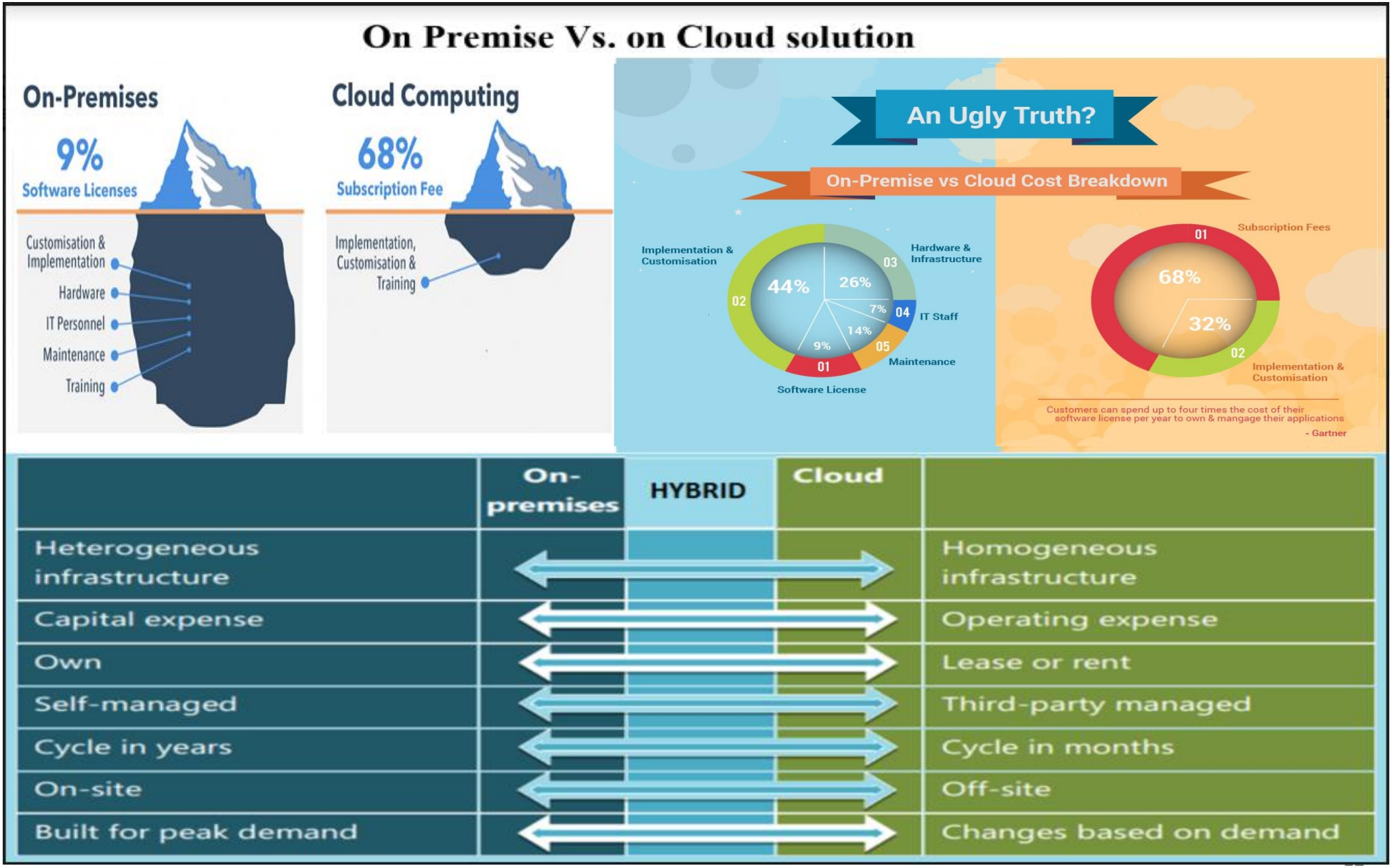
- Code language- ”**Python**”
- Libraries**: Pandas**, **Statsmodels**, **Numpy**, **Matplotlib**
- Web Application: The ”**Jupyter Notebook**”

TATA STEEL

CRITICAL CASE QUESTIONS (PART II)

women of mettle THE SCHOLARSHIP PROGRAMME SEASON #2

3. Best methods to be compared (Cloud Vs Premise solution)?



TATA STEEL

CONCLUSION

women of mettle THE SCHOLARSHIP PROGRAMME SEASON #2

- In a typical time series forecasting, it is important to identify if the time series has any trend, seasonality and cyclicality.
- Once identified, these factors are taken out of the time series to produce the stationary part of the time series.
- There are various ways for modelling the stationary part.
- The best solution differs (depends on the dataset).
- We use Time Series Analysis and Forecasting for many applications such as:
  - **Budget Analysis, Financial Market Analysis, Economic Forecasting, etc.**

TATA STEEL

women of mettle
THE SCHOLARSHIP PROGRAMME
SEASON #2

THANK YOU