Little Data, Huge Expectations...

Time Series Forecasting

Achintya Gupta

Data Scientist

Bio

About Me

- Graduated from Jamia Millia in 2018, in Electrical Engineering
- Previously with Nissan as AI-Engineer.
- Currently Working as a Data Scientist, Roadzen.
- Have a sweet spot for Mathematics.

Trivia

- Love to go on Hikes!
- Although these days, just workout and read!

Agenda

4 Hours

S1

Segment 1 (60 minutes) : Basic Introduction

8:30 A.M

S1.a) The Genesis **10 min** S1.b) Introduction to Numpy **10 min** S1.c) Introduction to Pandas **15 min** S1.d) Introduction to Matplotlib **10 min**

5 min - Q/A **10 min Break**

9:30 A.M

S2

Segment 2 (60 minutes) : Time Series Foundation

9:30 A.M

S2.a) Time Series **10 min** S2.b) Time Series Work Flow **5 min** S2.c) Statistical Foundations pt-1 **30 min**

5 min - Q/A **10 min Break**

10:30 A.M

S3

Segment 3 (60 minutes) : Time Series Foundation & Modelling Concepts

10:30 A.M

S3.a) Statistical Foundations pt-2 **30 min** S3.b) Modelling Concepts **20 min**

5 min - Q/A **5 min Break**

11:30 A.M

S4

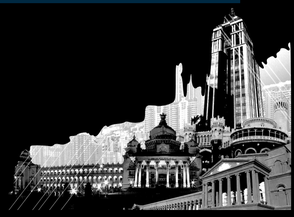
Segment 4 (60 minutes) : Modelling

11:30 A.M

S4.a) Modelling Concepts cont.. **15 min** S4.b) Modelling **40 min**

5 min - Q/A

12:30 A.M



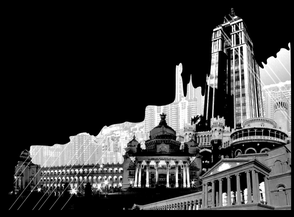
Environment Setup & Prerequisites

Prerequisites

- Understanding with Python
- Basic Mathematical & Statistical Knowledge for Machine Learning Models

1.) Install Anaconda Navigator
2.) Install Jupiter NBExtensions
3.) Install Libraries (requirements.txt) [pip install]
   3.1) ***ts***-mad
   3.2) ***statsmodels***
   3.3) ***prophet***
   3.4) ***scipy & sklearn***
   3.5) ***pandas & numpy*** *(data wrangling)*

TO UPDATE

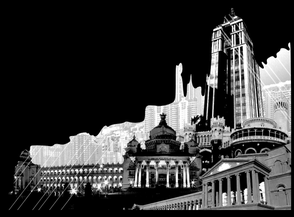


The Genesis

Definition of Time

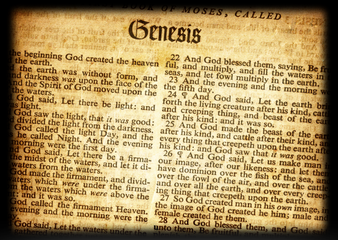
Time [1]

An indefinite continued progress of existence and events that occur in apparently irreversible succession from the past , through the present and into the future. Time is a component quantity of various measurements used to sequence events…



The Genesis

Time and its Journey

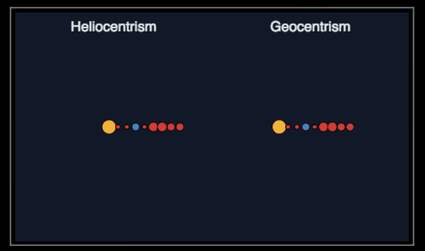


[3] Old Testament
(1200 BC) The Genesis : *Bereshit* (בְּרֵאשִׁית, literally "*In the beginning*")
First Chapter of Hebrew Bible.
Concept of *Day* and *Daytime* and *Nighttime*

Astronomical Diaries - Alexander The Great's Death (323BC)

Alexandria's Library Burnt (272 AD)
Lost most of Hellenistic Culture knowledge

Anno Domini (A.D)

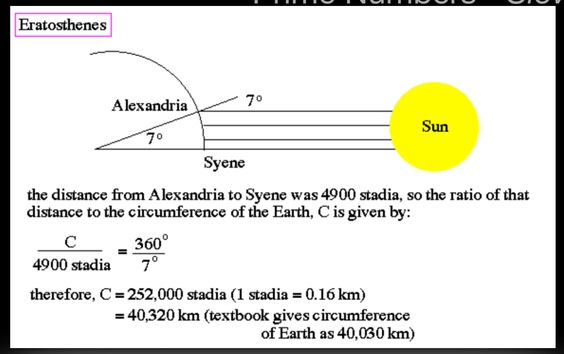


[4] Copernicus
(1500 AD) Challenges Church Doctrine which had absorbed Ptolemy's Model.

Johannes Kepler

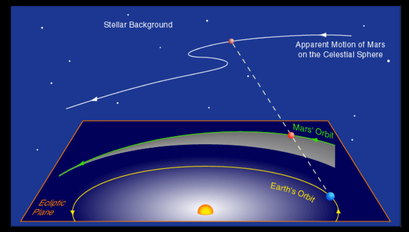
[2] Water Clock
(1500BC) Egyptian Time Measuring Devices

[4] Natural Philosophers
(500BC) Hellenistic Culture
Greeks inherited data from Babylonians … but also to think of new experiments.
Thales - was able to predict Eclipses
Eratosthenes - Father of Geography, Prime Numbers - *Sieve of Eratosthenes*)



[4] Tycho Brahe
(1580 AD)
First True Observer in Astronomy

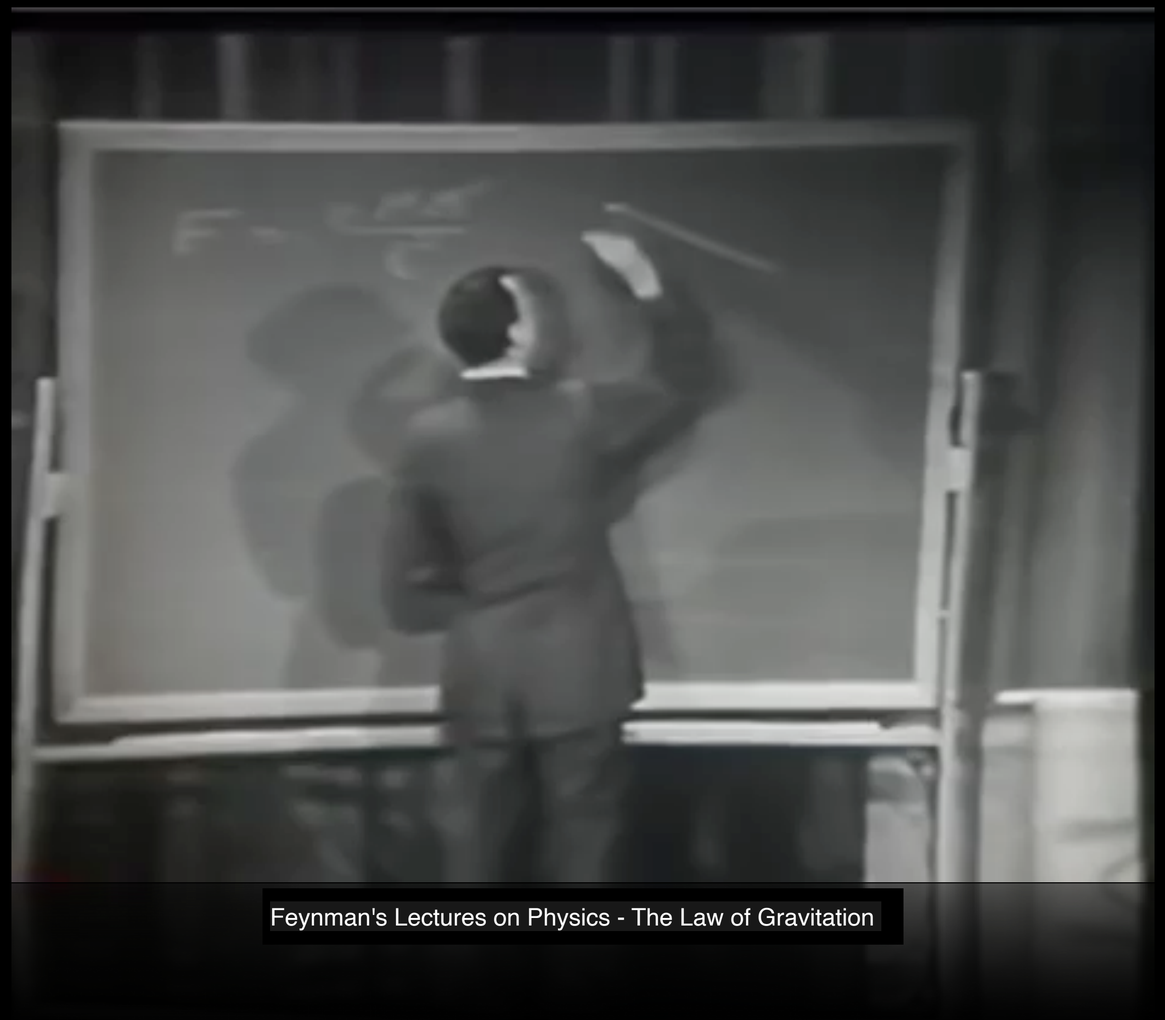
Isaac Newton



The Genesis

First Recorded Time Series Richard Feynmans Remarks

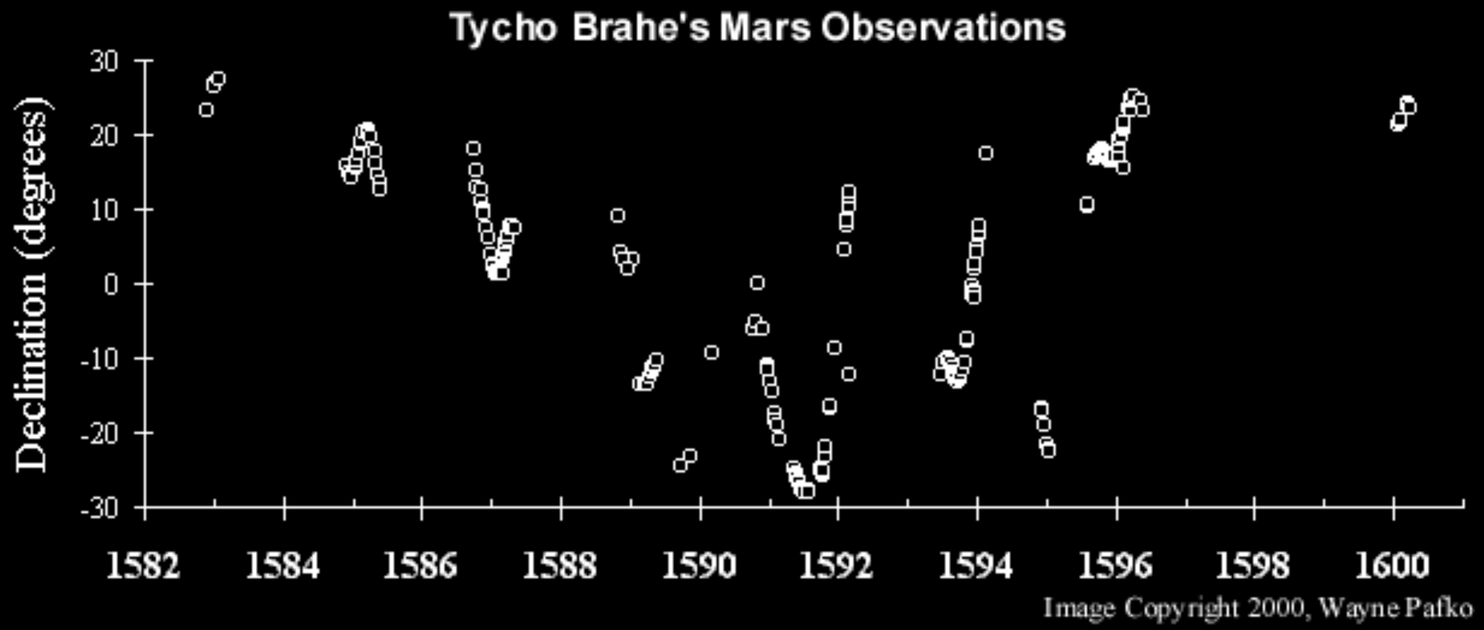
Tycho Brahe
(1580 AD)
First True Observer in Astronomy



The Genesis

One of the First Recorded Time Series [5]

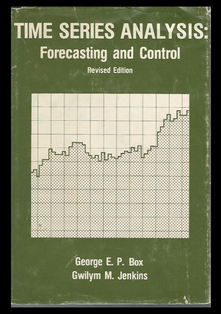
Tycho Brahe
(1580 AD)
First True Observer in Astronomy



The Genesis

Modern History

[2] G.U Yule
(1920)
Wrote Three Influential Papers
1.) On the Time-Correlation Problem…
2.) Nonsense Correlation..
3.) Wolfers Sunspots Forecasting…



[4] Box & Jenkins
(1970)
Full Time Series Modelling procedure for individual series
- Specification
- Estimation
- Diagnostics
- Forecasting

[3] Herman Wold
(1938)
1.) Wold Decomposition
2.) Theory Of Consumer Demand

The Genesis

to the Notebook…

1 The Genesis

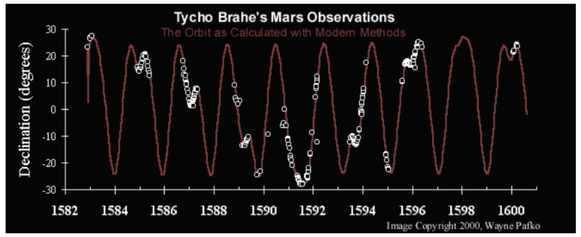
Time Series and its History and how pivotal has it been in the answering very contoversoial questions -

Time series has been there for ages, from Bybolnians of Mesopotamia, recording weather conditions, daily commodity prices and river levels against dates with some important events in their *Astronomical Diaries*.

It is through these records that we know the date of demise of Alexander - The Great's in Babylon, 323 BC

But after **The Great Library of Alexandria** was burnt in 272 AD, there is no complete records of that data left anymore sadly!

One of the oldest Time Series that we do have recorded data of is of ***Mars Declination*** angles by **Tycho Brahe**, which, to the best of my knowledge consititues as *One of the Oldest Time Series* that we have a record of.



[Tycho Brahe - Mars Declination Ang.gif] - Wayne Pafko, http://www.pafko.com/tycho/observe.html

[Astronomical Diary - Alexander the Greats Death.png] - https://www.britishmuseum.org/collection/object/W_1881-0706-403

2 Imports [...]

3 Path Declaration & Variable Initialisation [...]

4 Tycho Brahe - Mars Declination [...]

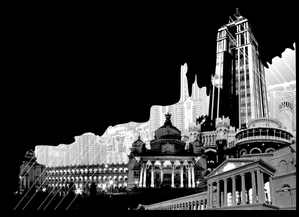
The Genesis

Our lessons

1.) Time Series has been there for ages, although the scientific and statistical development started in around early 20th Century.

2.) It has played a pivotal role in answering some of the most controversial questions in the human history, be it *“Wether the Earth is at the centre of the Universe?”* or *“Where the economy is going?”*

3.) Analyst’s and Scientists have tried extensively to tame this beast, but very few have succeeded in doing so, the sole reason, **THE FUTURE IS UNPREDICTABLE, ERRATIC AND CHAOTIC, MUCH LIKE A HUMAN BEHAVIOUR!**



BREAK

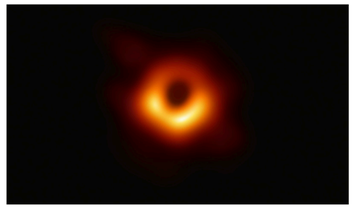
Introducing Numpy

to the Notebook...

1 Introduction

NumPy (Numerical Python) is an open source Python library that's used in almost every field of science and engineering. It's the universal standard for working with numerical data in Python, and it's at the core of the scientific Python and PyData ecosystems.

Interesting Read : The first image of a Balck Hole



2 Imports [...]

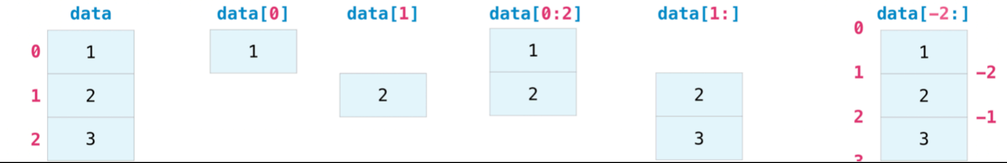
3 Python & `array` [...]

4 `numpy` to the rescue [...]

5 Basics - Array Generation [...]

6 Indicing and Reshaping Numpy Array [...]

Indicing an Numpy `ndarray` is same as indicing a Python🐍 `list`



Introducing Pandas

to the Notebook...

1 Introduction

pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

Interesting Read : mlcourse.ai : EDA with Pandas

Much of this Notebook has been adopted from `pandas` docs

2 Imports [...]

3 Paths & Variable Initialisations [...]

4 Pandas DataStructures [...]

`pandas` creates and stores data in rectangular format.

On a broad stroke there are majorly two forms of a datatype in `pandas` :-

- Series : "Series is a one-dimensional labeled array capable of holding any data type (integers, strings, floating point numbers, Python objects, etc.)."
- Dataframe : "DataFrame is a 2-dimensional labeled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dict of Series objects."

5 Pandas Datetime Operations [...]

"Pandas builds upon `dateutil`, `datetime` & `numpy.datetime64` the tools just discussed to provide a Timestamp object, which combines the ease-of-use of datetime and dateutil with the efficient storage and vectorized interface of numpy.datetime64. From a group of these Timestamp objects, Pandas can construct a DatetimeIndex that can be used to index data in a Series or DataFrame" - Jake-Python Data Science Handbook

6 Pandas TimeZones [...]

7 Pandas Resampling [...]

NOTE : Do not confuse this with Undersampling and Oversampling

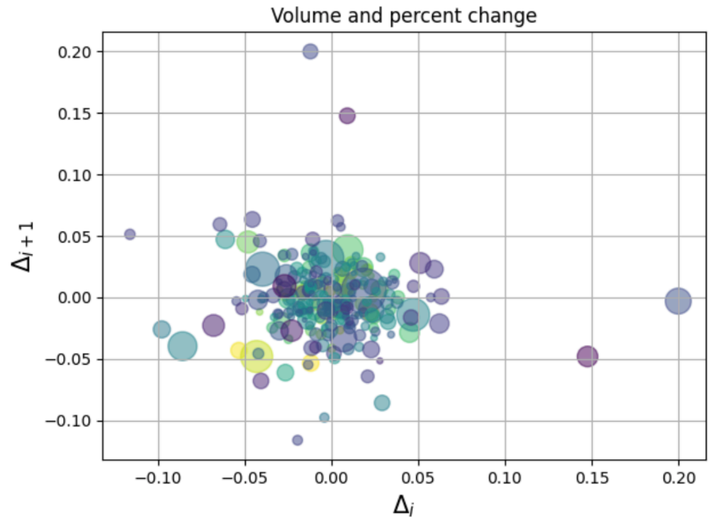
Undersampling	Oversampling

Introducing Matplotlib

to the Notebook…

1 Introduction [...]

`matplotlib` is a comprehensive library for creating static, animated, and interactive visualizations in Python.



Matplotlib Gallery

2 Imports [...]

3 Paths & Variable Initialisations [...]

4 Basics [...]

5 Simple Line Plot [...]

6 Simple Scatter Plots [...]

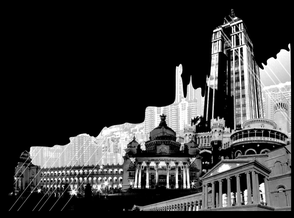
BREAK

Time Series Analysis

What is it?

An ordered sequence of values of a variable at equally spaced time intervals, since time series datapoint are adjacent in time, there is a potential for correlation among them, this is one feature that makes them stand out.

> *"It is far better to foresee even without certainty, than not to foresee at all"*
>
> *– Henri Poincare , French Mathematician 1854-1912*



Time Series Analysis

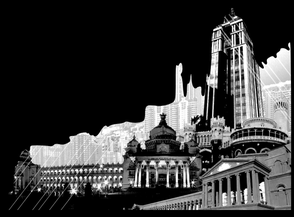
Caveat's

1.) Your analysis over the Time Series Data will only be as strong as your understanding of the **DOMAIN KNOWLEDGE**. Hence, it is advised to have either a good understanding of the Domain Knowledge yourself, or get in touch with the Domain Expert.

2.) **POINT FORECAST** are usually a bad idea, unless accompanied by **CONFIDENCE INTERVALS**.

3.) Like in any other Machine Learning problem, it is important that we define the **METRICS** keeping in mind the KPI's we are tracking.

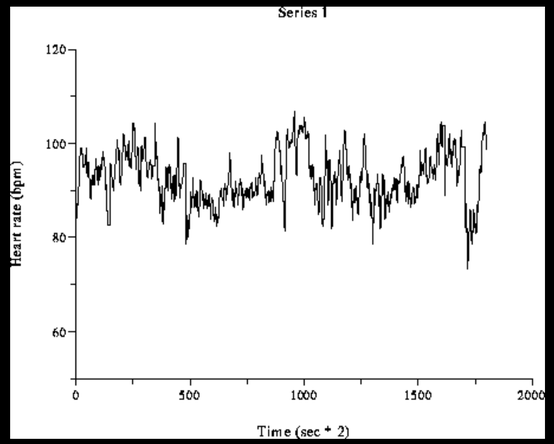
4.) **SIMPLICITY** is *elegant & effective.*



Time Series Analysis

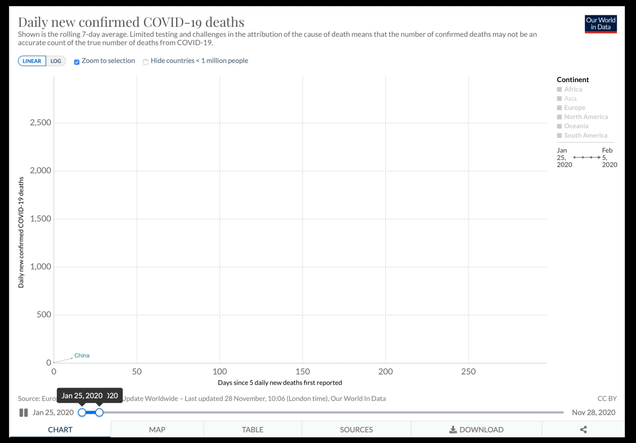
Where do they find their place? There Applications

- Obtain an understanding of the underlying forces and structure that produced the observed data
- Fit a model and proceed to forecasting, monitoring or even feedback and feedforward control.



HEALTH

FINANCE



EPEDIMIEOLOGY

Time Series Analysis

Time Series Data v/s Cross Sectional Data v/s Panel Data

Time Series Data

- Data collected against a *Single* object, can have multiple features though across an equally spaced time span.
  (Ex : HeartBeat TS,Energy Consumption TS)

- Since the readings are recorded at adjacent timestamps, there is a possibility of correlation between readings, which is also the genesis of **Auto-Correlation**.

Time Series Data

	State	Description	Rural	Urban	Combined	Date
0	Delhi	Cereals and products	106.300000	105.700000	105.800000	2013-01-01 00:00:00
1	Delhi	Cereals and products	106.900000	107.300000	107.300000	2013-02-01 00:00:00
2	Delhi	Cereals and products	107.100000	107.500000	107.500000	2013-03-01 00:00:00
3	Delhi	Cereals and products	108.200000	107.600000	107.700000	2013-04-01 00:00:00
4	Delhi	Cereals and products	108.300000	108.500000	108.500000	2013-05-01 00:00:00
5	Delhi	Cereals and products	109.500000	109.700000	109.700000	2013-06-01 00:00:00
6	Delhi	Cereals and products	111.200000	110.900000	110.900000	2013-07-01 00:00:00
7	Delhi	Cereals and products	112.600000	111.000000	111.100000	2013-08-01 00:00:00
8	Delhi	Cereals and products	112.200000	110.900000	111.000000	2013-09-01 00:00:00
9	Delhi	Cereals and products	112.800000	111.400000	111.500000	2013-10-01 00:00:00

Cross Sectional Data

- Data collected against *Multiple* objects at a single point in time.
  (Ex : Focus Group Data, Iris Dataset..)

- Although the data is not recorded in adjacent timestamps, the multiple features that are collected during the process might be correlated to each-other giving rise to **Multi-Collinearity**.

Cross Sectional Data

	State	Description	Rural	Urban	Combined	Date
0	Delhi	Cereals and products	106.300000	105.700000	105.800000	2013-01-01 00:00:00
1	Delhi	Meat and fish	106.500000	103.700000	103.900000	2013-01-01 00:00:00
2	Delhi	Egg	114.600000	116.400000	116.300000	2013-01-01 00:00:00
3	Delhi	Milk and products	104.300000	103.900000	103.900000	2013-01-01 00:00:00
4	Delhi	Oils and fats	105.700000	102.400000	102.700000	2013-01-01 00:00:00
5	Delhi	Fruits	93.400000	101.500000	101.200000	2013-01-01 00:00:00
6	Delhi	Vegetables	93.800000	91.700000	91.900000	2013-01-01 00:00:00
7	Delhi	Pulses and products	111.000000	108.000000	108.300000	2013-01-01 00:00:00
8	Delhi	Sugar and confectionery	106.900000	106.900000	106.900000	2013-01-01 00:00:00
9	Delhi	Spices	106.700000	103.200000	103.500000	2013-01-01 00:00:00

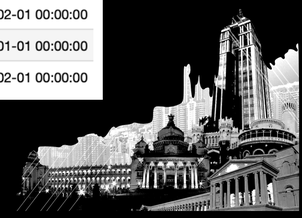
Panel Data

- Sometimes also referred to as *Longitudinal Data*, is an amalgamation Time Series data and Cross sectional data.
  (Ex : CPI Data for multiple states (M))

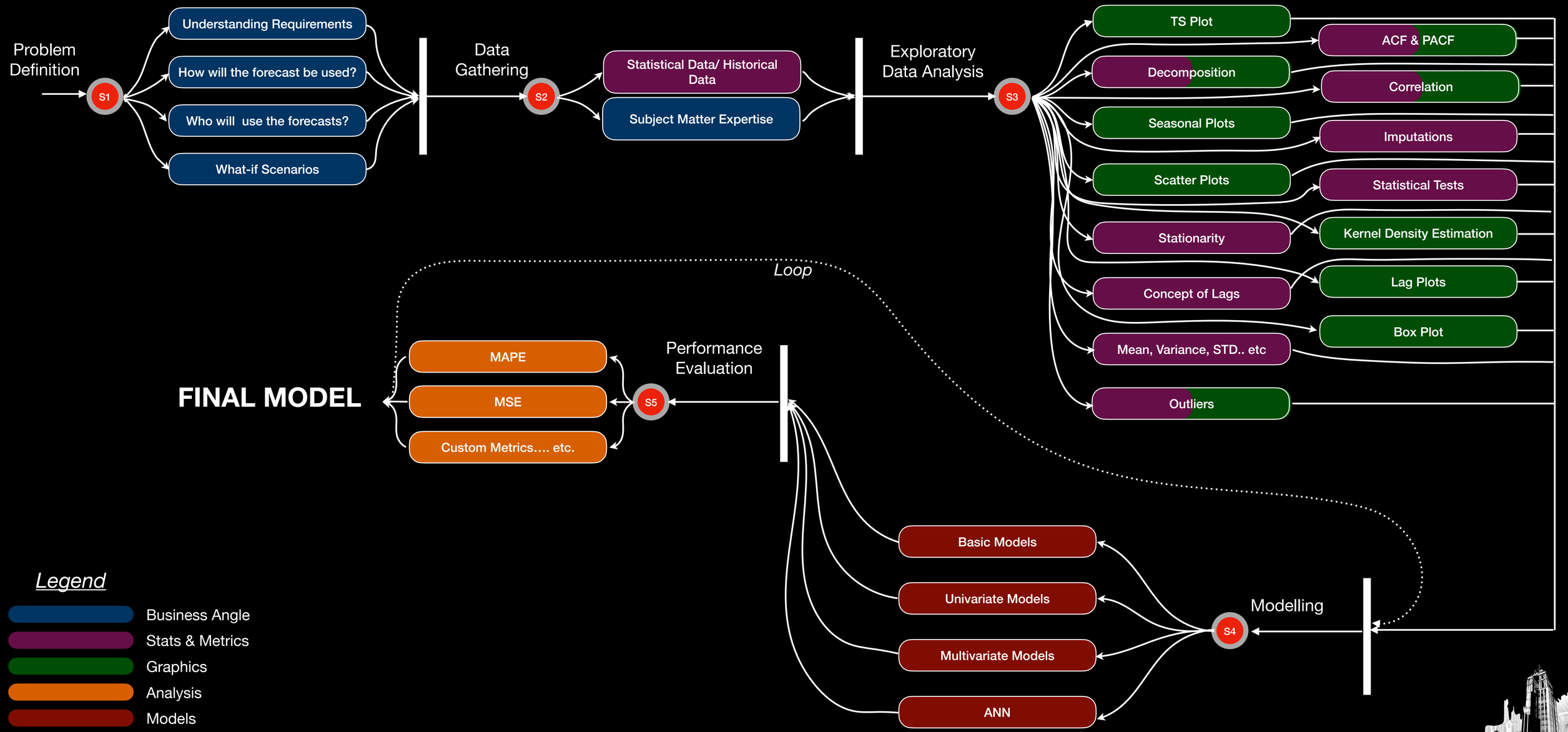
- Since the recording are having the characteristics of both Time Series & Cross sectional data, both **Auto-Correlation** & **Multi-Collinearity** can be seen

Panel Data

	State	Description	Rural	Urban	Combined	Date
0	Delhi	Cereals and products	106.300000	105.700000	105.800000	2013-01-01 00:00:00
1	Delhi	Cereals and products	106.900000	107.300000	107.300000	2013-02-01 00:00:00
2	Delhi	Meat and fish	106.500000	103.700000	103.900000	2013-01-01 00:00:00
3	Delhi	Meat and fish	112.800000	107.800000	108.100000	2013-02-01 00:00:00
4	Delhi	Egg	114.600000	116.400000	116.300000	2013-01-01 00:00:00
5	Delhi	Egg	113.700000	115.200000	115.100000	2013-02-01 00:00:00
6	Delhi	Milk and products	104.300000	103.900000	103.900000	2013-01-01 00:00:00
7	Delhi	Milk and products	104.300000	104.200000	104.200000	2013-02-01 00:00:00
8	Delhi	Oils and fats	105.700000	102.400000	102.700000	2013-01-01 00:00:00
9	Delhi	Oils and fats	106.000000	101.400000	101.800000	2013-02-01 00:00:00



Flow of Time Series Problem Statement



Statistical Foundations - Pt-1

to the Notebook...

1 Introduction

This notebook will take you through some of the foundational elements of the Time Series analysis, Including various types of dataset's and how does Time Series Data differ from them, Inferrential Statistics pertaining to the Time Series data, Various graphical visualisation involved in the time series.

Interesting Read : Engineering Statistics Handbook

R - Libraries : Some useful libraries and components for Time Series Analysis in R - by Rob.J.Hyndman

2 Imports [...]

3 Path and Variable Initialisation [...]

4 Load & Explore the Data [...]

5 Different Types of Data [...]

- Time Series Data : Data collected against a ***Single feature*** across an equally spaced time span.
- Cross-Sectional Data : Data collected against ***Multiple features*** at a single point in time.
- Panel Data : Recording data for ***Multiple features at eaqually spaced time intervals***

6 Time Series Related Operations [...]

- Windows
- Lags

BREAK

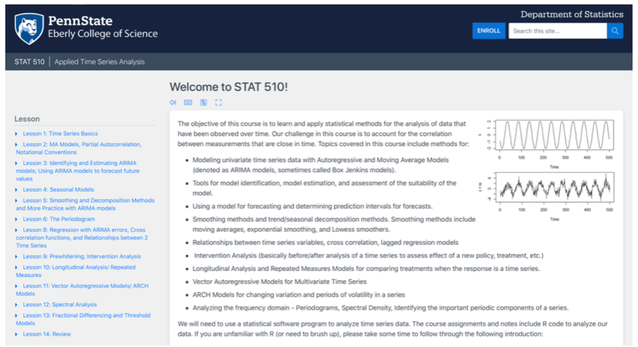
Statistical Foundations - Pt-2

to the Notebook...

1 Introduction [...]

This notebook will focus primarily statistical side of the time series modelling, including how to define Correlation's, the concept of Stationarity, Auto-Correlation & Partial Auto-Correlation Function, Auto-Regressive & Moving Average Processes in modelling and how to perform Model Diagnostics.

Interesting Read : STAT 510



2 Imports [...]

3 Path and Variable Initialisation [...]

4 Load & Explore the data [...]

5 Correlation [...]

Correlation : Two variables are said to be correlated when the value assumed by one affects the distribution of the other. It reflects the association between the two variables whose strength usually lies within the range of -1 to +1. If, as the value of X increase there is an increase in Y, then X & Y are said to be positively correlated. Also if, as the value of X decrease there is an increase in Y, then X & Y are said to be negatively correlated.

Different Types Of Correlations

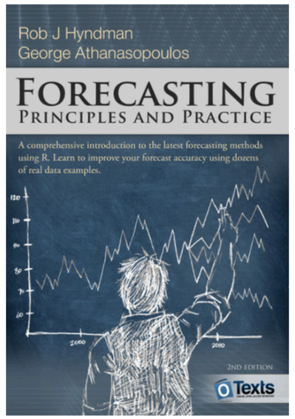
Modelling Concepts

to the Notebook...

1 Introduction [...]

This notebook will focus primarily on modelling, various types of Univariate Models.

Interesting Read : Forecasting Principles & Practices



2 Imports [...]

3 Path and Variable Initialisation [...]

4 Splitting Data [...]

There is a need to split the data such that we land with three buckets of the Data:-

- Train, CV & Test Splits
  - Training Data - Data that the Model Learns upon
  - Cross Validation Data - Data on which we determine our hyperparameters/parameters of the model on
  - Test Data - Data on which we measure our `Metrics` as to how the model is doing on a dataset which it has never seen before, neither in Training or in Cross Validation Phase.

BREAK

Sunspots Modelling

to the Notebook…

1 Introduction

Sunspots are regions on the sun's surface that are cooler than the surrounding areas and so appear darker. Sunspots have been observed continuously since 1609, and the first observation dates back a 100 years in china, If sunspots are active, more solar flares will result creating an increase in geomagnetic storm activity for Earth. Therefore during sunspot maximums, the Earth will see an increase in the Northern and Southern Lights and a possible disruption in radio transmissions and power grids.

In [1]:

```
from IPython.display import YouTubeVideo
YouTubeVideo('rx9m6H6GeLs', width=1000, height=400) # Relevant till 3:00
```

Out[1]:

2 Imports […]

3 Path and Variable Initialisations […]

4 Load and Explore the Data […]

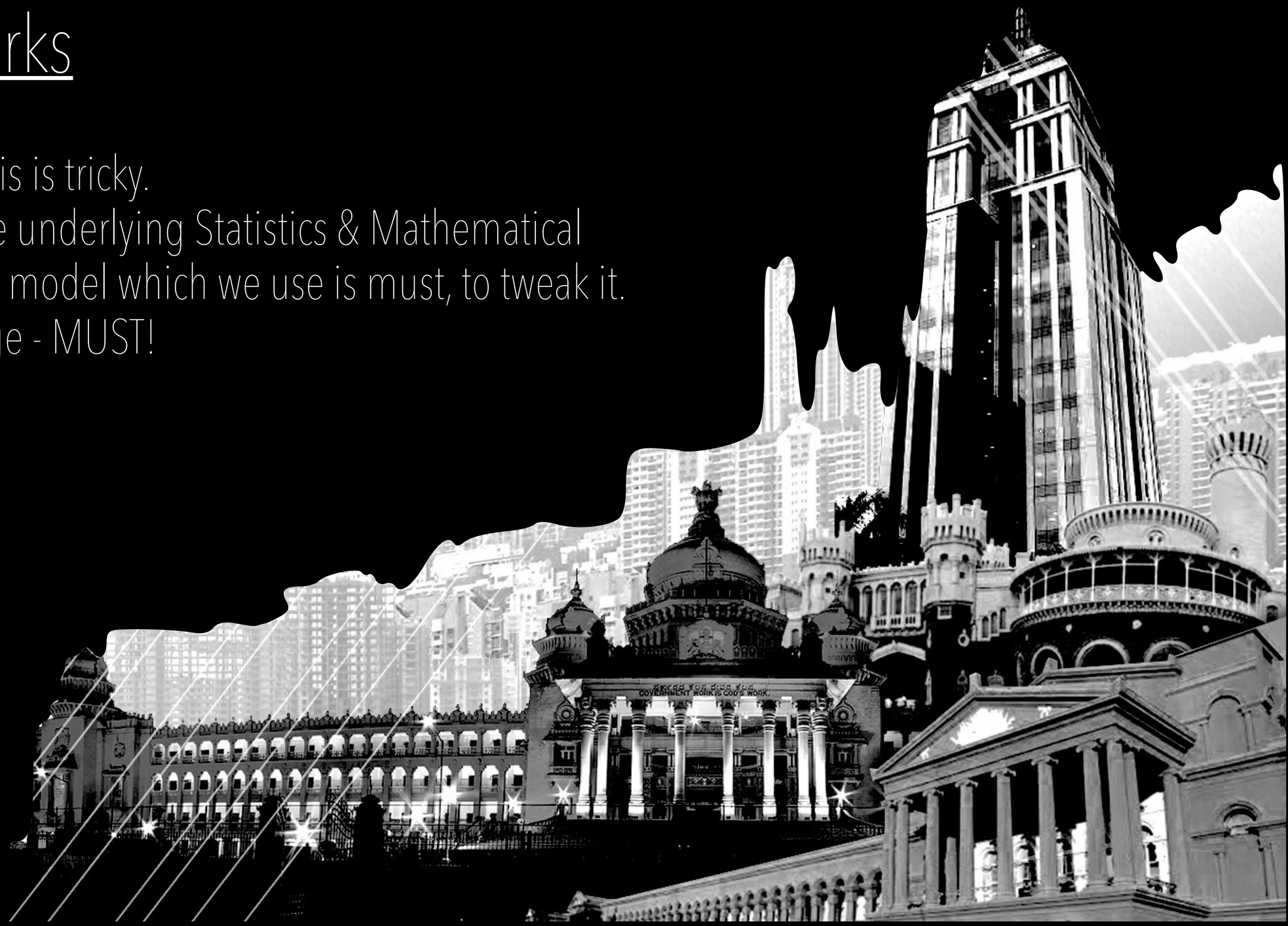
5 In-Depth Exploratory Data Analysis & Data Preperation […]

6 Gauging its Statistics […]

- Stationarity Test

Closing Remarks

- Time Series Analysis is tricky.
- Understanding the underlying Statistics & Mathematical formulation of any model which we use is must, to tweak it.
- Domain Knowledge - MUST!



Appendix

[1] https://en.wikipedia.org/wiki/Time

[2] https://en.wikipedia.org/wiki/Time#History_of_time_measurement_devices

[3] https://en.wikipedia.org/wiki/Book_of_Genesis

[4] http://abyss.uoregon.edu/~js/ast121/lectures/lec02.html

[5] http://www.pafko.com/tycho/observe.html