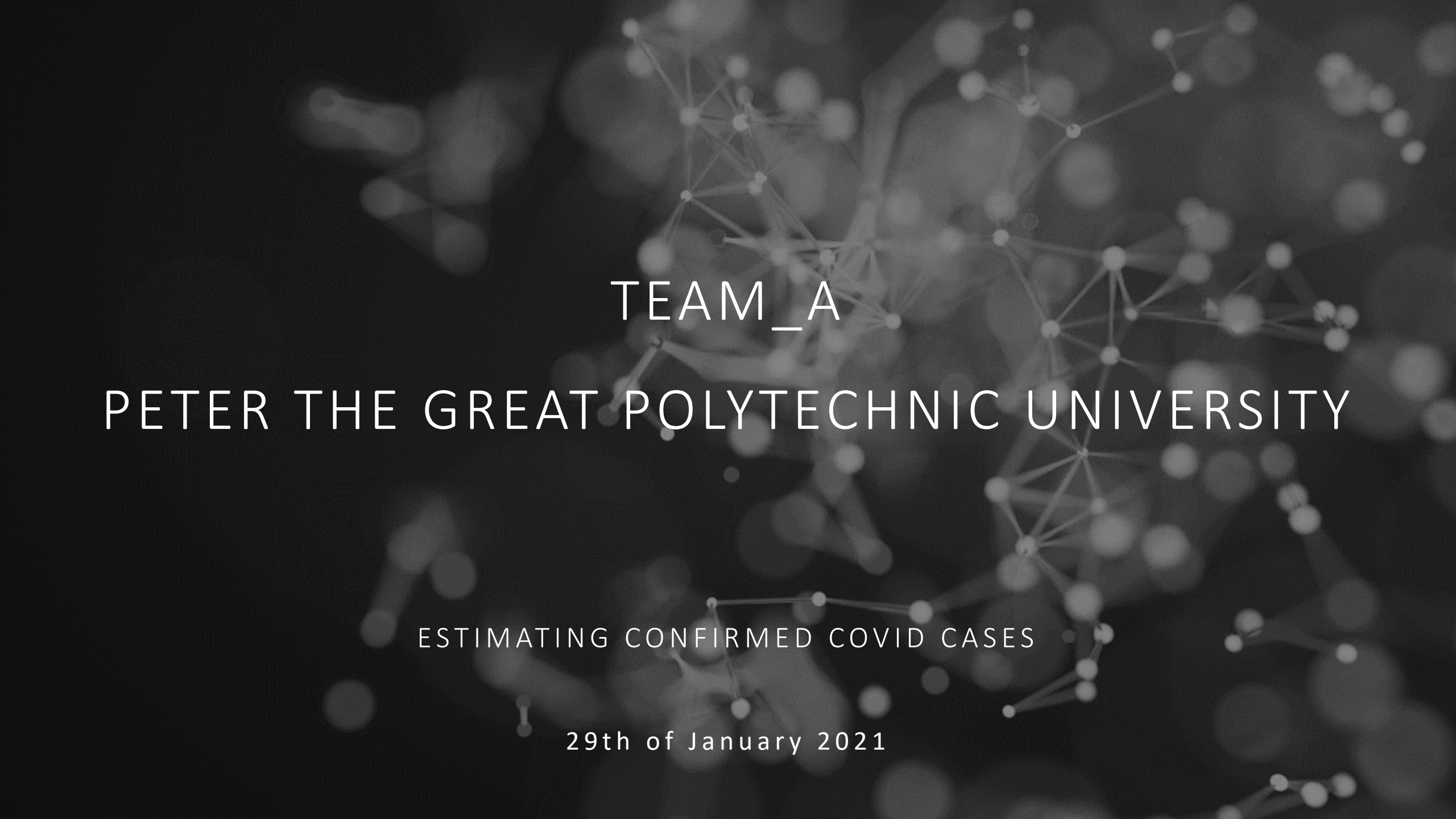


TEAM_A

PETER THE GREAT POLYTECHNIC UNIVERSITY

ESTIMATING CONFIRMED COVID CASES

29th of January 2021

AGENDA

INTRODUCTION

MATHEMATICAL METHODS & ALGORITHMS

RESULTS AND ERRORS

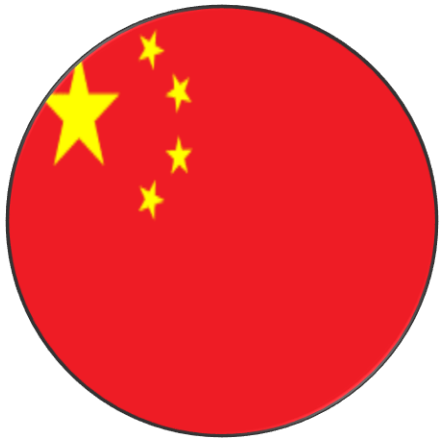
INSIGHTS AND CONCLUSION

FUTURE APPROACHES

INTRODUCTION

Shaima Almeer

Liu Zhiyuan



Daniel Denk

TABLE OF RESPONSIBLES

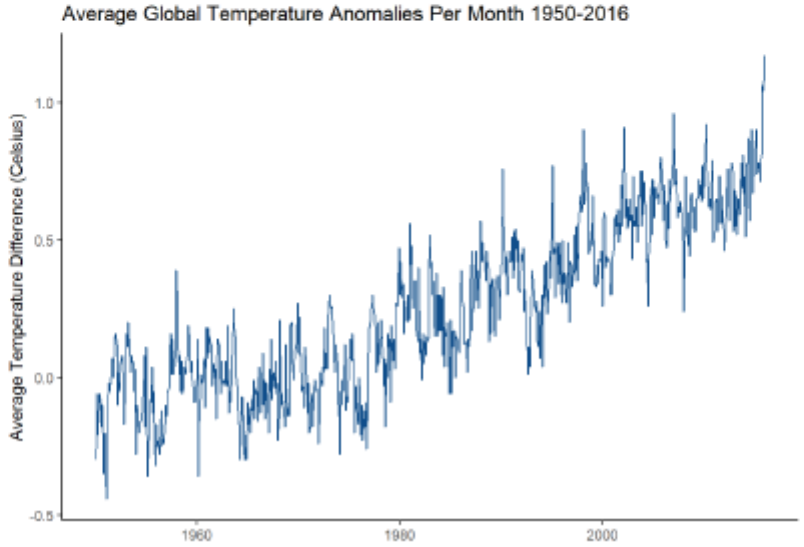
Task	Persons_Involved
Data Analysis	Whole Team
Data Preparation	Whole Team
Building Data Model	Whole Team
Performing Predictions	Whole Team
Data Visualization	Whole Team
Error Estimation	Whole Team
Database Schema	Whole Team
Data Transfer into a Database	Whole Team
Description of Mathematical Models	Whole Team
Conclusions of the Data	Whole Team
Definition of the next Stages of Development	Whole Team
Definition of Data Visualization Tools for the Database	Whole Team

MATHEMATICAL METHODS & ALGORITHMS

TIME SERIES DATA

Time series data

A time series is a set of observations on the values that a variable takes at different times. Such data may be collected at regular time intervals such as monthly, weekly, quarterly or annually. In describing time series data, the patterns are classified into **trend, seasonal, cyclic, random components**



TIME SERIES DATA

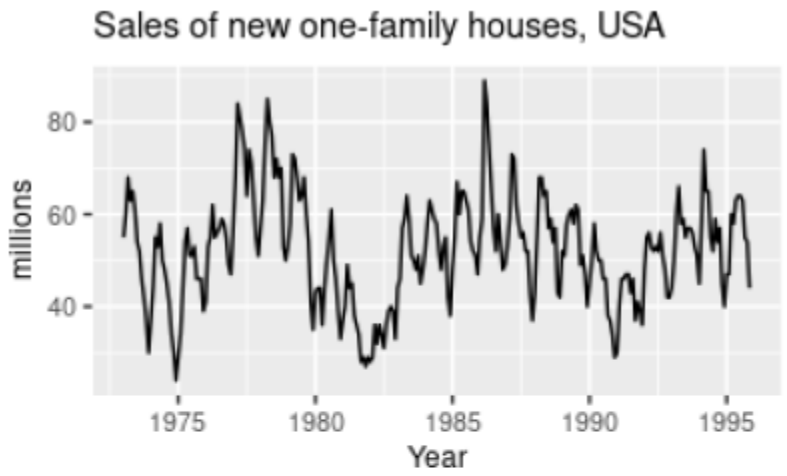
- TREND:
  A TREND EXISTS WHEN THERE IS A LONG-TERM INCREASE OR DECREASE IN THE DATA. IT DOES NOT HAVE TO BE LINEAR. SOMETIMES WE WILL REFER TO A TREND AS “CHANGING DIRECTION", WHEN IT MIGHT GO FROM AN INCREASING TREND TO A DECREASING TREND.

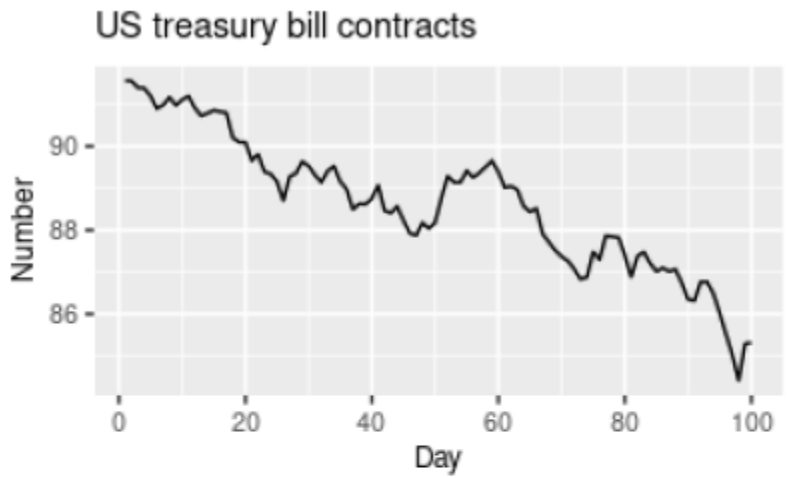
- SEASONAL:
  A SEASONAL PATTERN OCCURS WHEN A TIME SERIES IS AFFECTED BY SEASONAL FACTORS SUCH AS THE TIME OF THE YEAR OR THE DAY OF THE WEEK. SEASONALITY IS ALWAYS OF A FIXED AND KNOWN FREQUENCY.

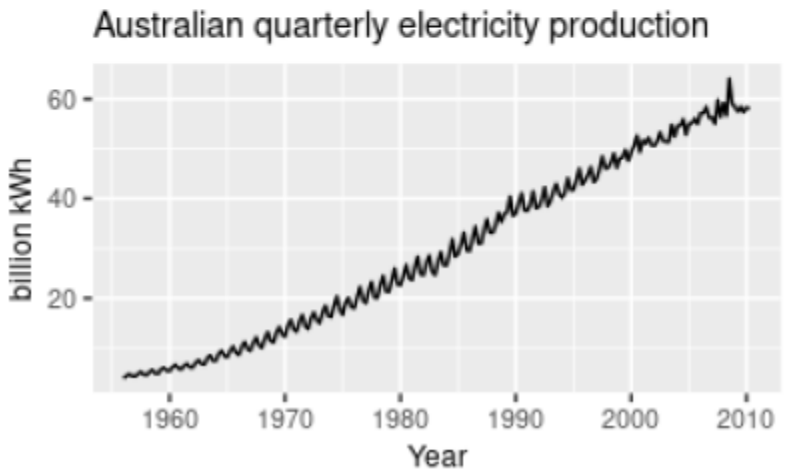
- CYCLIC:
  A CYCLE OCCURS WHEN THE DATA EXHIBIT RISES AND FALLS THAT ARE NOT OF A FIXED FREQUENCY. THESE FLUCTUATIONS ARE USUALLY DUE TO ECONOMIC CONDITIONS AND ARE OFTEN RELATED TO THE \BUSINESS CYCLE". THE DURATION OF THESE FLUCTUATIONS IS USUALLY AT LEAST 2 YEARS.

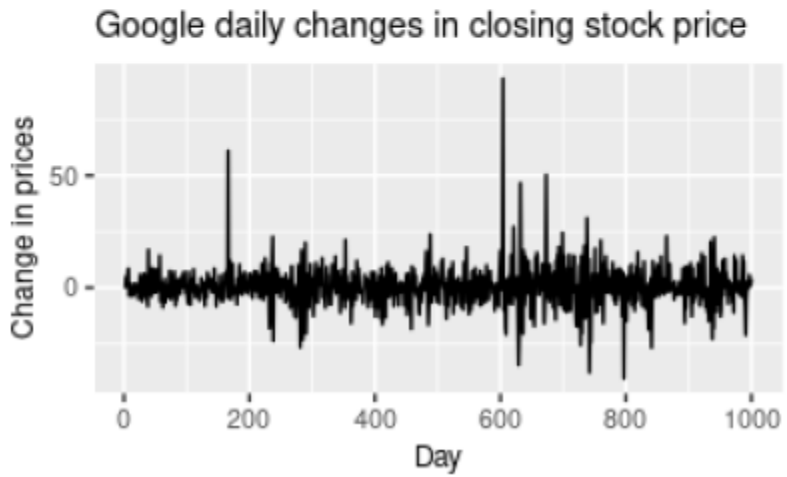
- RANDOM COMPONENTS:
  THE COMPONENT OF A TIME SERIES DATA THAT IS OBTAINED AFTER THESE THREE PATTERNS HAVE BEEN EXTRACTED OUT OF THE SERIES IS THE RANDOM COMPONENT.

TIME SERIES DATA









STATIONARY

Strictly stationary

A time series $\{r_t\}$ is said to be **strictly stationary** if the joint distribution of $(r_{t_1} \cdots r_{t_k})$ is identical to that of $(r_{t_{1+t}} \cdots r_{t_{k+t}})$ for all t, where k is an arbitrary positive integer and $(t_1 \cdots t_k)$ is a collection of k positive integers.

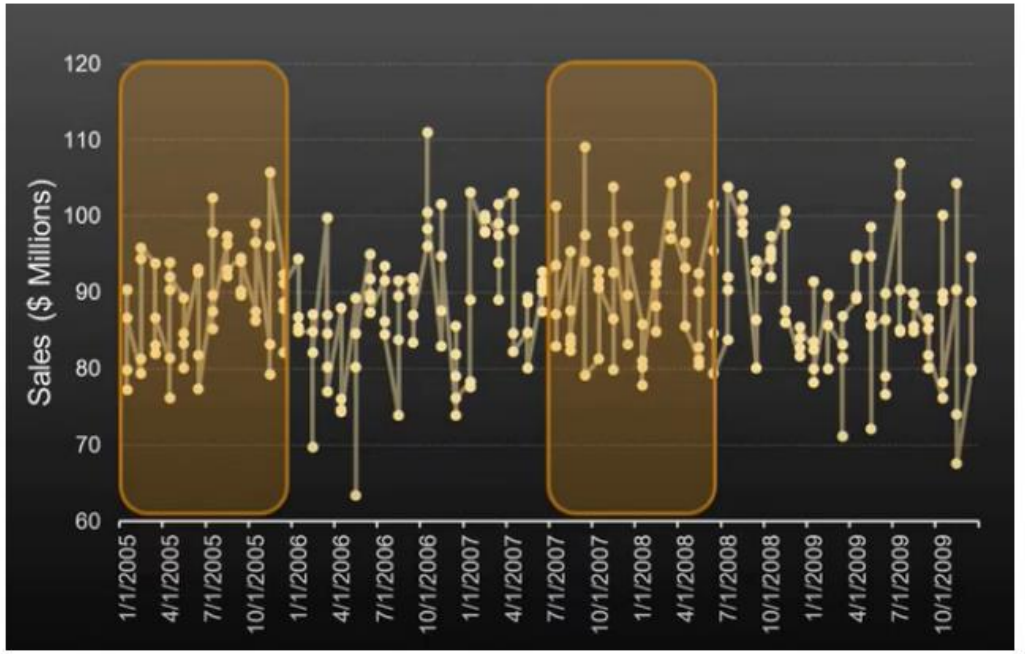
Weakly stationary

A time series $\{r_t\}$ is **weakly stationary** if both the mean of r_t and the covariance between r_t and r_{tl} are time invariant, where l is an arbitrary integer.

STATIONARY

Why we need stationarity?

To make good prediction. Most time series models assume that each point is independent of one another. The best indication of this is when the dataset of past instances is stationary. For data to be stationary, the statistical properties of a system do not change over time. If a time series has trend or seasonality, then it is not stationary



NOTICE: SAME WIDTH, SAME DISTRIBUTION

AUTOREGRESSIVE MODEL

In an **autoregressive model**, we forecast the variable of interest using a linear combination of past values of the variable. The term autoregressive indicates that it is a regression of the variable against itself. The past values in the series are called **lags**.

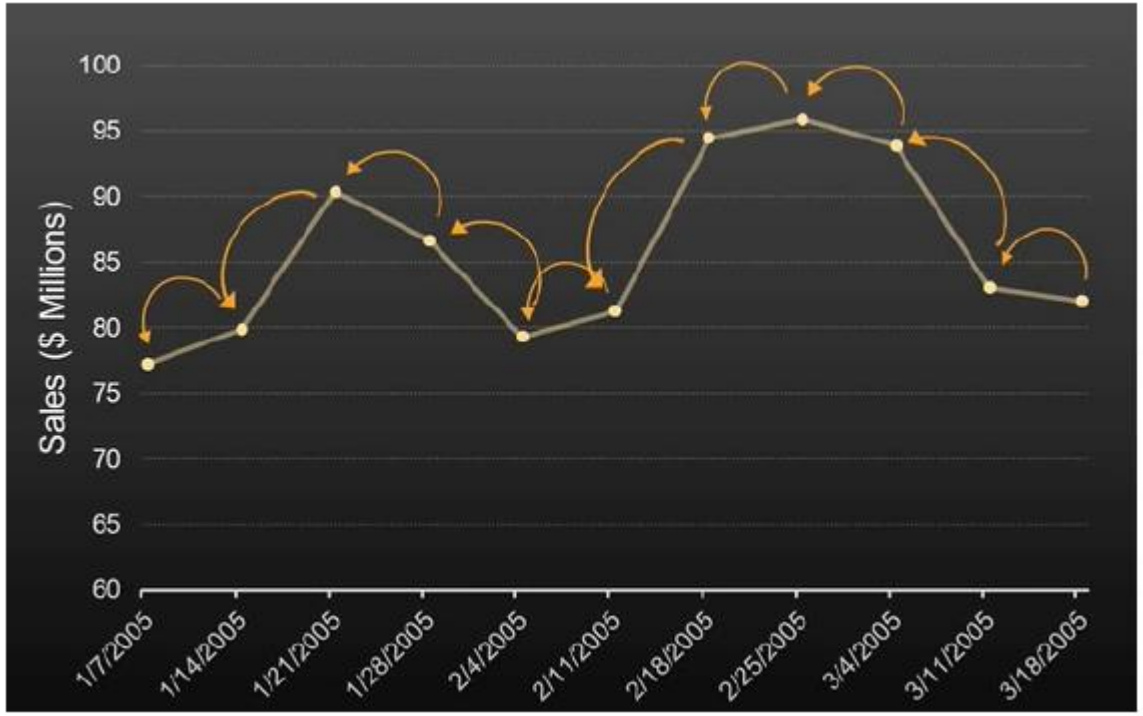
An example of **an autoregressive model of order one (one lag)** is

$$Y_t = c + \phi Y_{t-1} + e_t$$

c - constant

e_t - random components (white noise)

Y_{t-1} - lagged target.



The first observation has a small effect on what's going on today

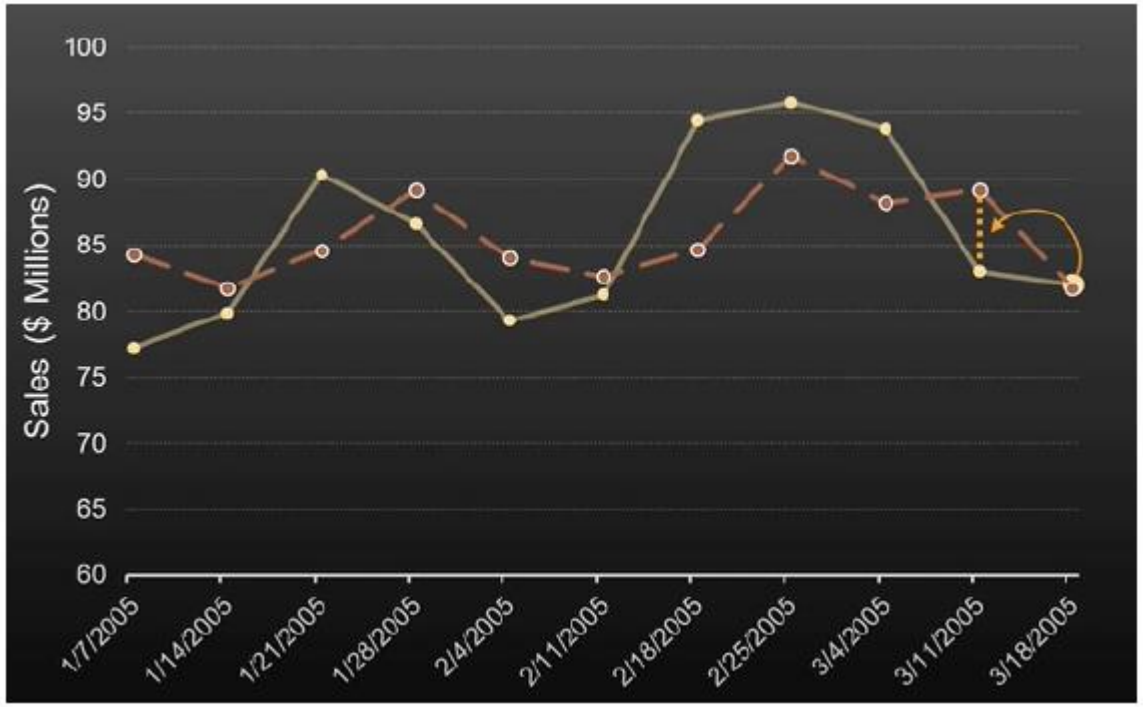
MOVING AVERAGE MODEL

Rather than using past values of the forecast variable in a regression, a moving average model uses **past forecast errors** in a regression-like model. It is not really a regression in the usual sense because we do not observe the values of past errors directly. The past forecast errors are called **error lags**. An example of a moving average model of order one (one error lag) is:

$$Y_t = c + \theta e_{t-1} + e_t$$

c - constant

e_t, e_{t-1} lagged error



As long as you go far enough in the future, the effect of present shocks will have no effect

ARIMA MODEL

Its model is a combination of autoregressive model with moving average model **with one more differencing step**, so this model is called Auto Regressive Integrated Moving Average model. In this context, “integration" is the reverse of differencing.

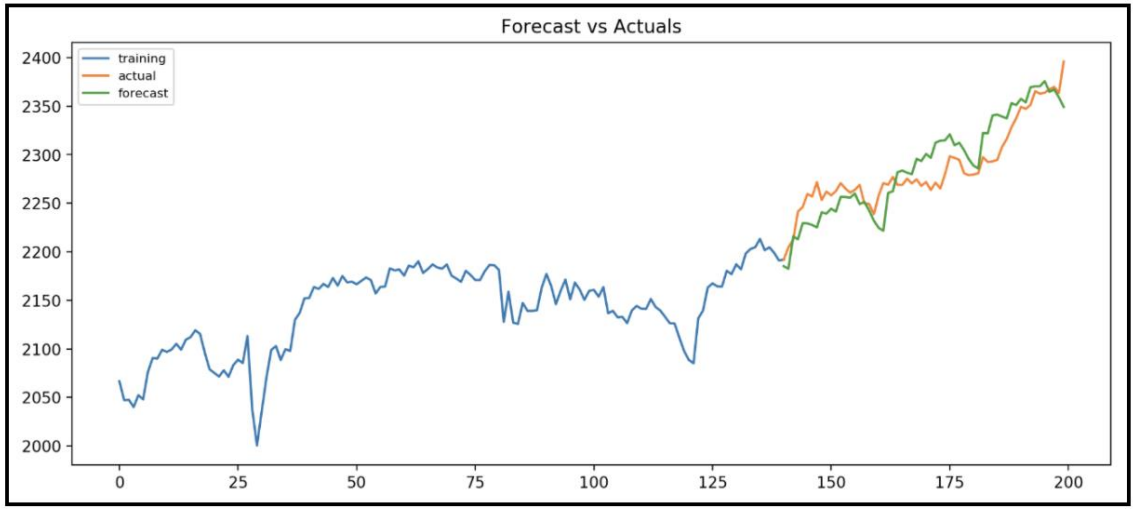
$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t$$

p - the order of the autoregressive part

d - times to perfom lag-1 differencing

q - the order of the moving average part

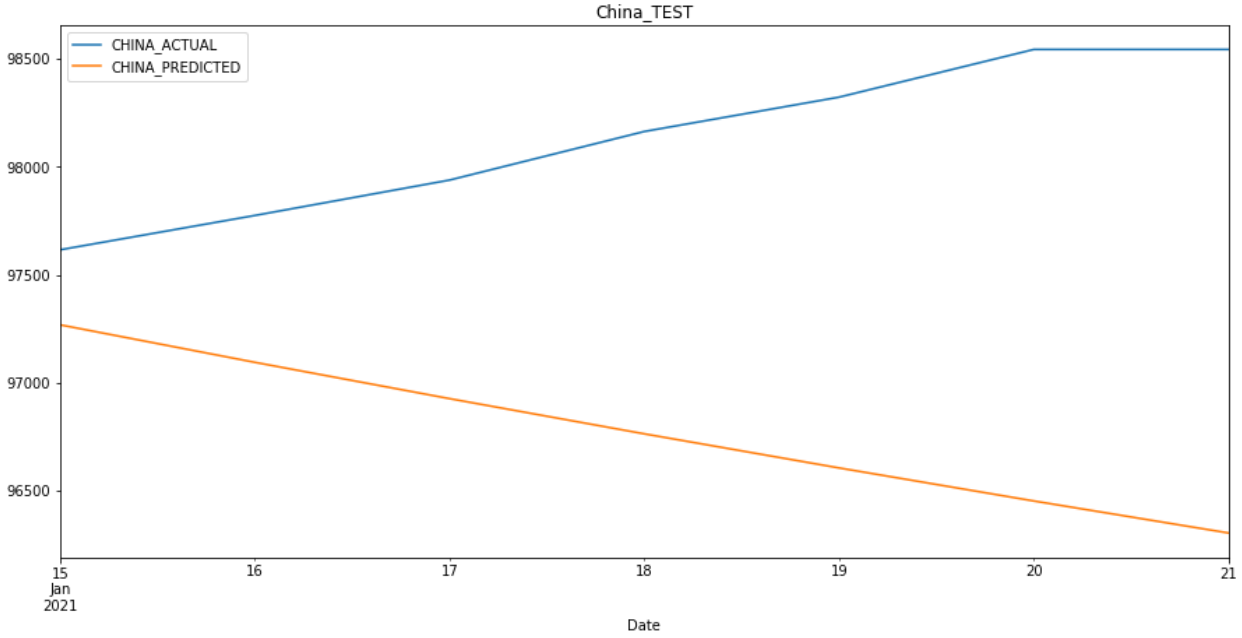
SPECIAL CASE SARIMAX (SEASONAL)



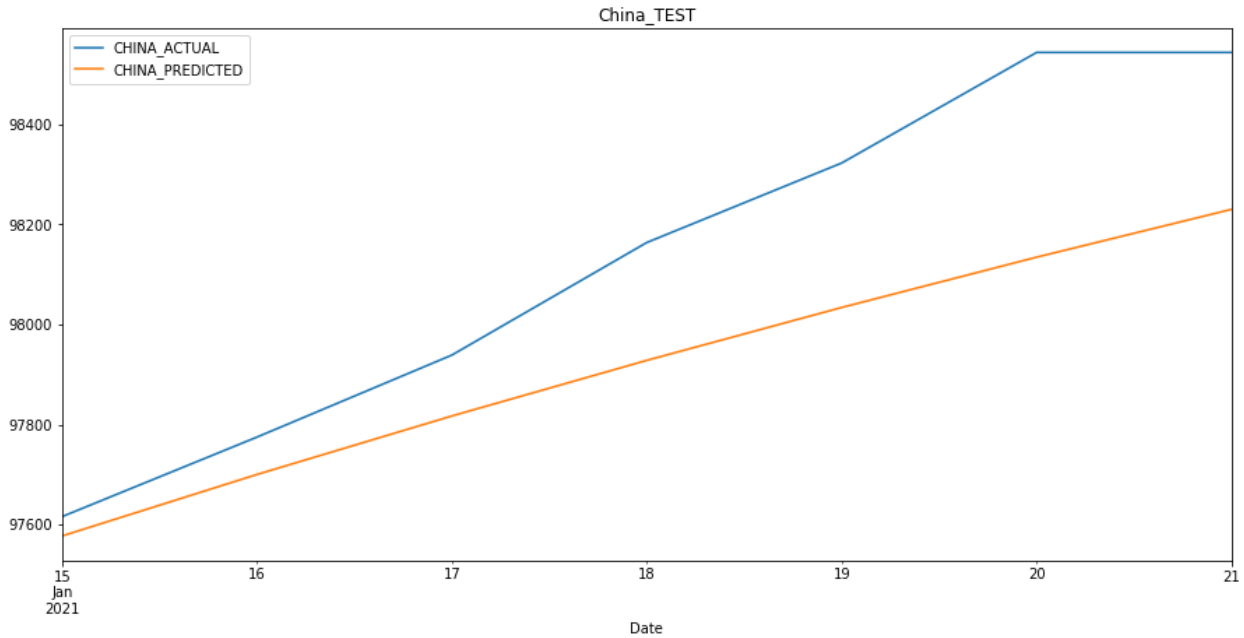

```
Test RMSE: 23.580
Test Percentage Error: 0.010%
```

How to choose p,q,d is beyond our knowledge now, but some methods like MINIC,SCAN,ESACF can be considered. Typical choices like (1,1,1) but beware of overdifferencing

COMPARISON OF PREDICTIONS



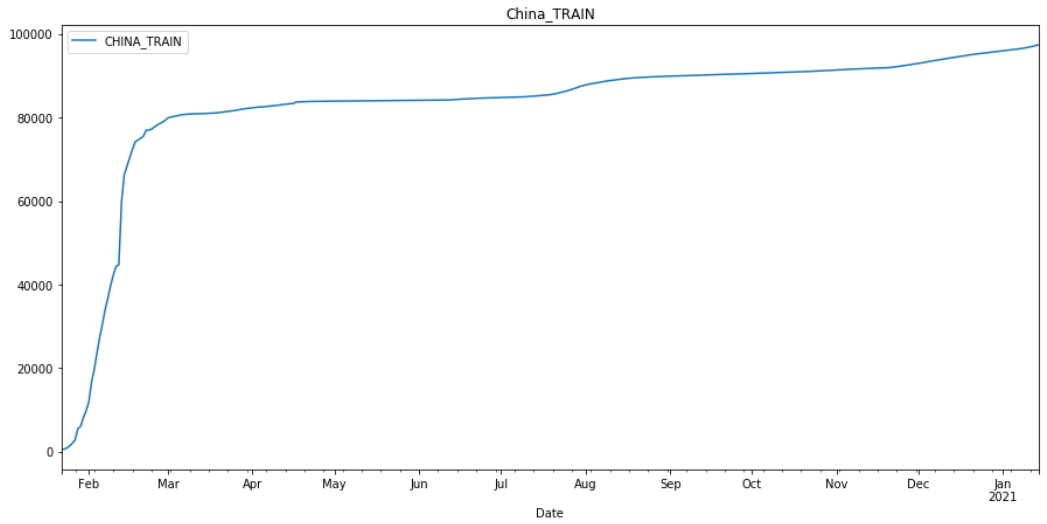
AUTOREGRESSION

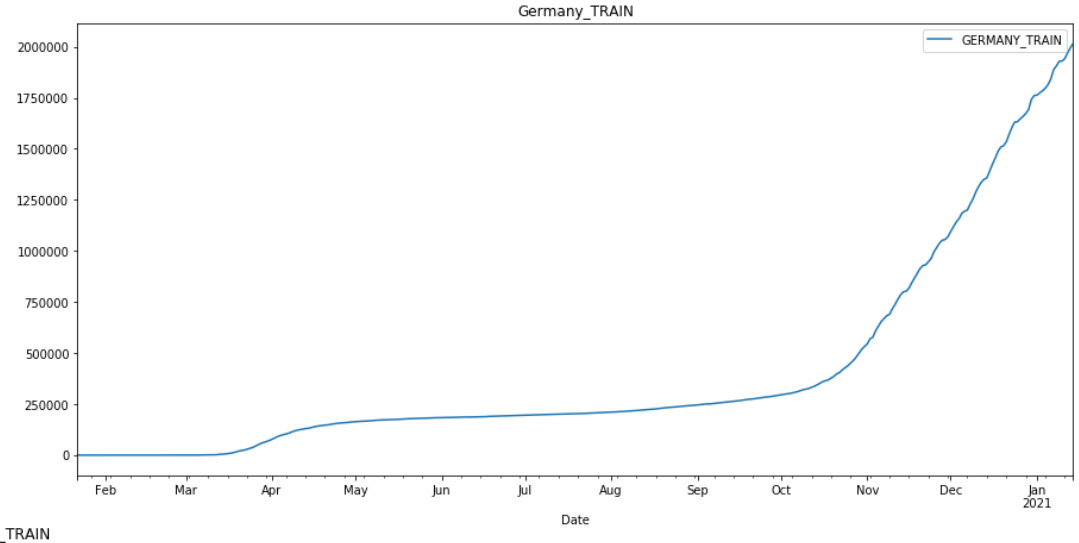


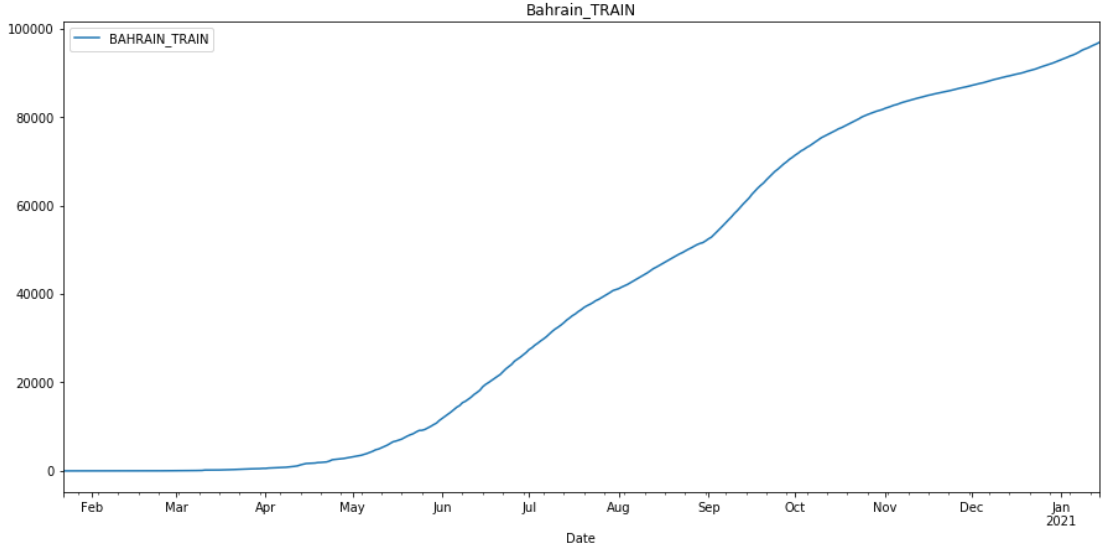
ARIMA

WE CHOSE ARIMA

TRAINING SAMPLE







RESULTS AND ERRORS

OUR ERROR FUNCTIONS

- r2_score
- mean_squared_error
- n_absolute_percentage_error

How well the regression line fits the data
HIGH is good (lower deviation)

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Average squared difference between the estimated values and the actual value
LOW is good

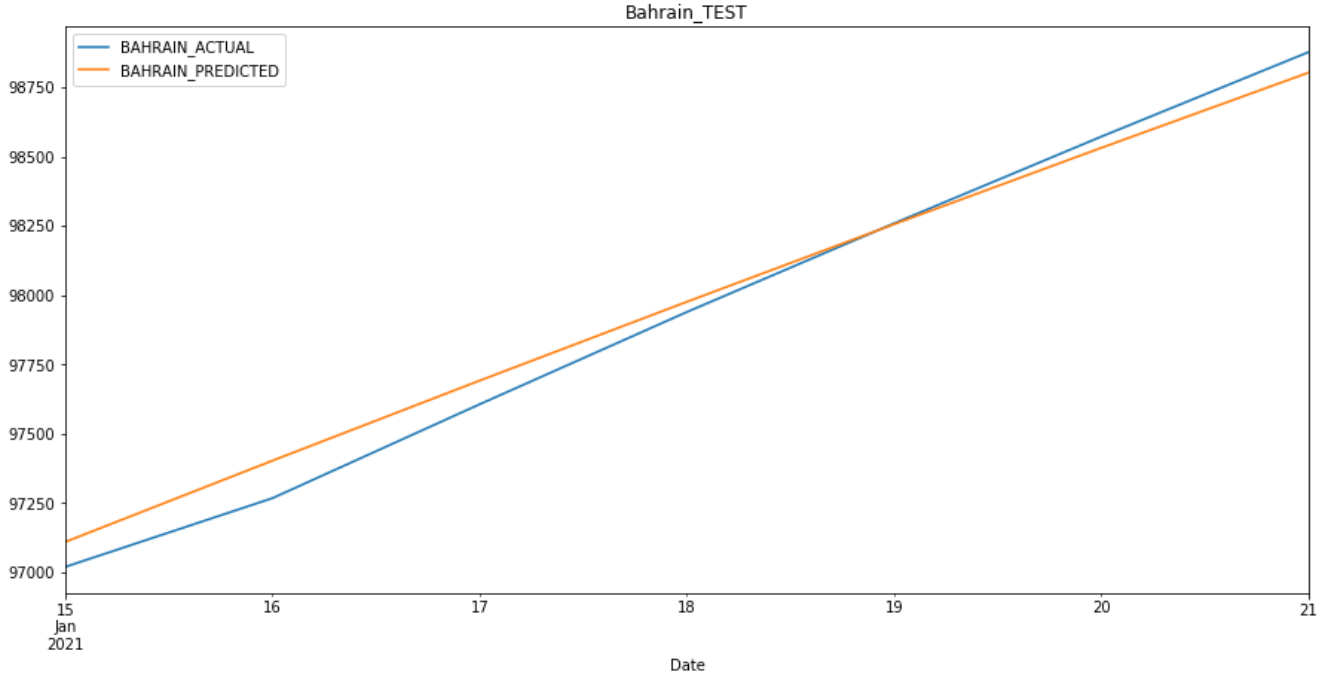
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (e_i)^2 = \frac{1}{n}\mathbf{e}^\mathsf{T}\mathbf{e}$$

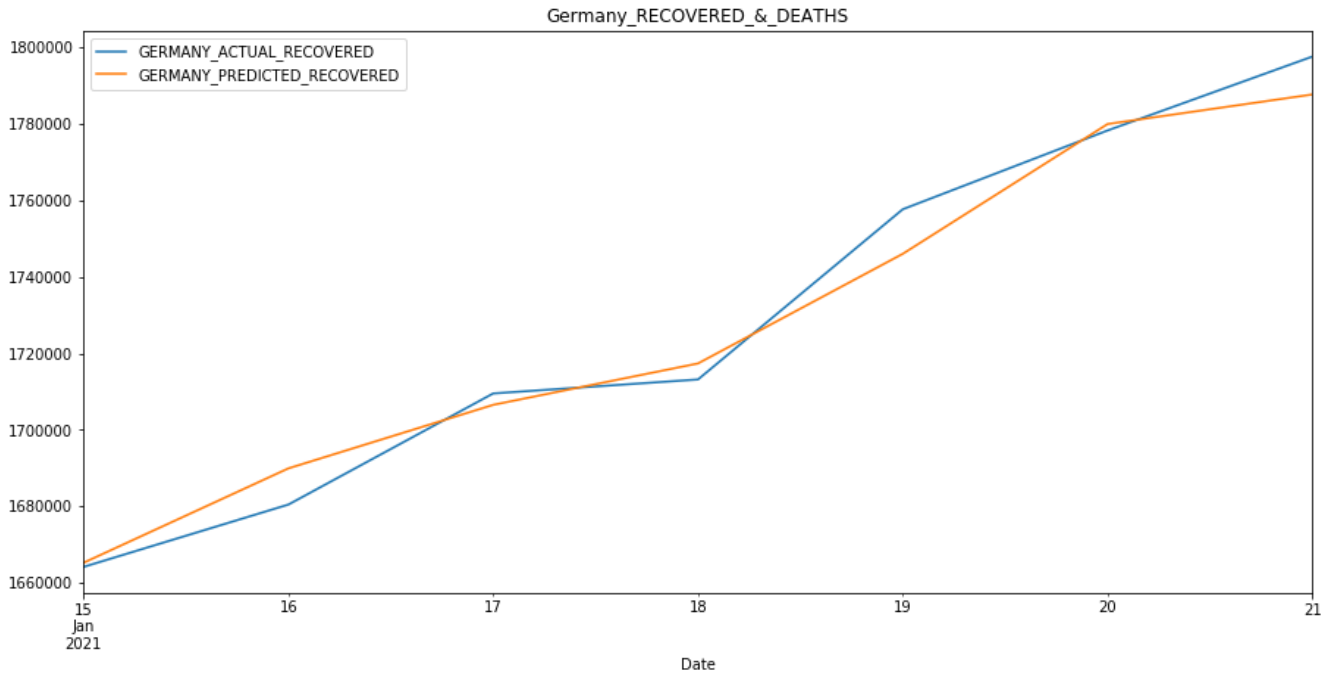
Great Intuitive Interpretion
LOW is good

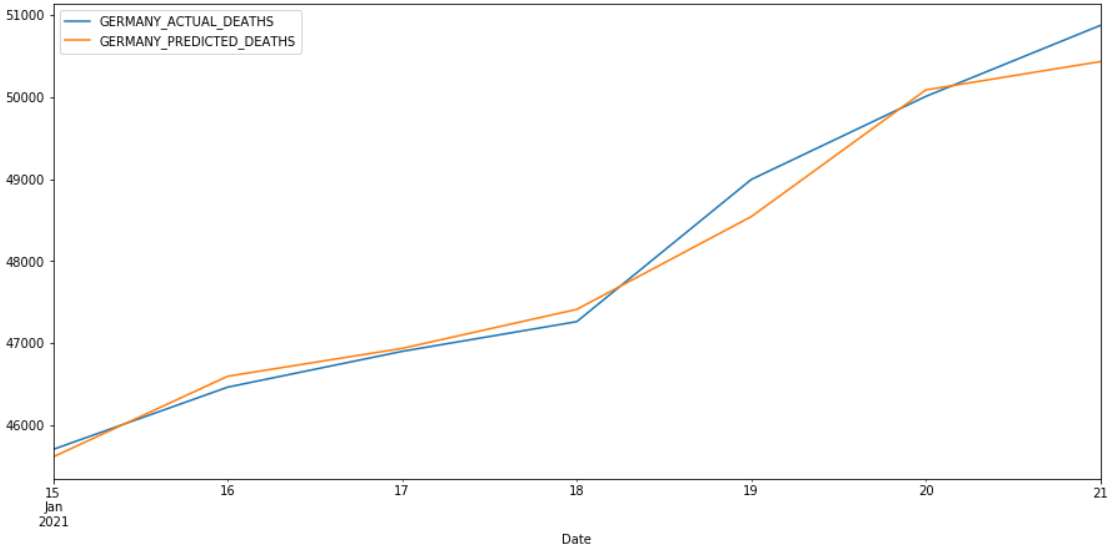
$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{A_t - F_t}{A_t}\right|,$$

=> All: Take in actual value and compare it to the predicted one

PERFORMANCE ESTIMATE







PERFORMANCE IN NUMBERS


```
"Country":"China",
"Prediction":{
   "2021-01-15 00:00:00":97574,
   "2021-01-16 00:00:00":97697,
   "2021-01-17 00:00:00":97815,
   "2021-01-18 00:00:00":97930,
   "2021-01-19 00:00:00":98041,
   "2021-01-20 00:00:00":98148,
   "2021-01-21 00:00:00":98251
},
"Actual_Values":{
   "2021-01-15 00:00:00":97616,
   "2021-01-16 00:00:00":97775,
   "2021-01-17 00:00:00":97939,
   "2021-01-18 00:00:00":98164,
   "2021-01-19 00:00:00":98323,
   "2021-01-20 00:00:00":98544,
   "2021-01-21 00:00:00":98544
}
```

```
"Country":"Germany",
"Prediction":{
   "2021-01-15 00:00:00":2033814,
   "2021-01-16 00:00:00":2052388,
   "2021-01-17 00:00:00":2070958,
   "2021-01-18 00:00:00":2089524,
   "2021-01-19 00:00:00":2108085,
   "2021-01-20 00:00:00":2126641,
   "2021-01-21 00:00:00":2145193
},
"Actual_Values":{
   "2021-01-15 00:00:00":2023828,
   "2021-01-16 00:00:00":2038645,
   "2021-01-17 00:00:00":2050129,
   "2021-01-18 00:00:00":2059382,
   "2021-01-19 00:00:00":2071615,
   "2021-01-20 00:00:00":2100618,
   "2021-01-21 00:00:00":2108895
}
```

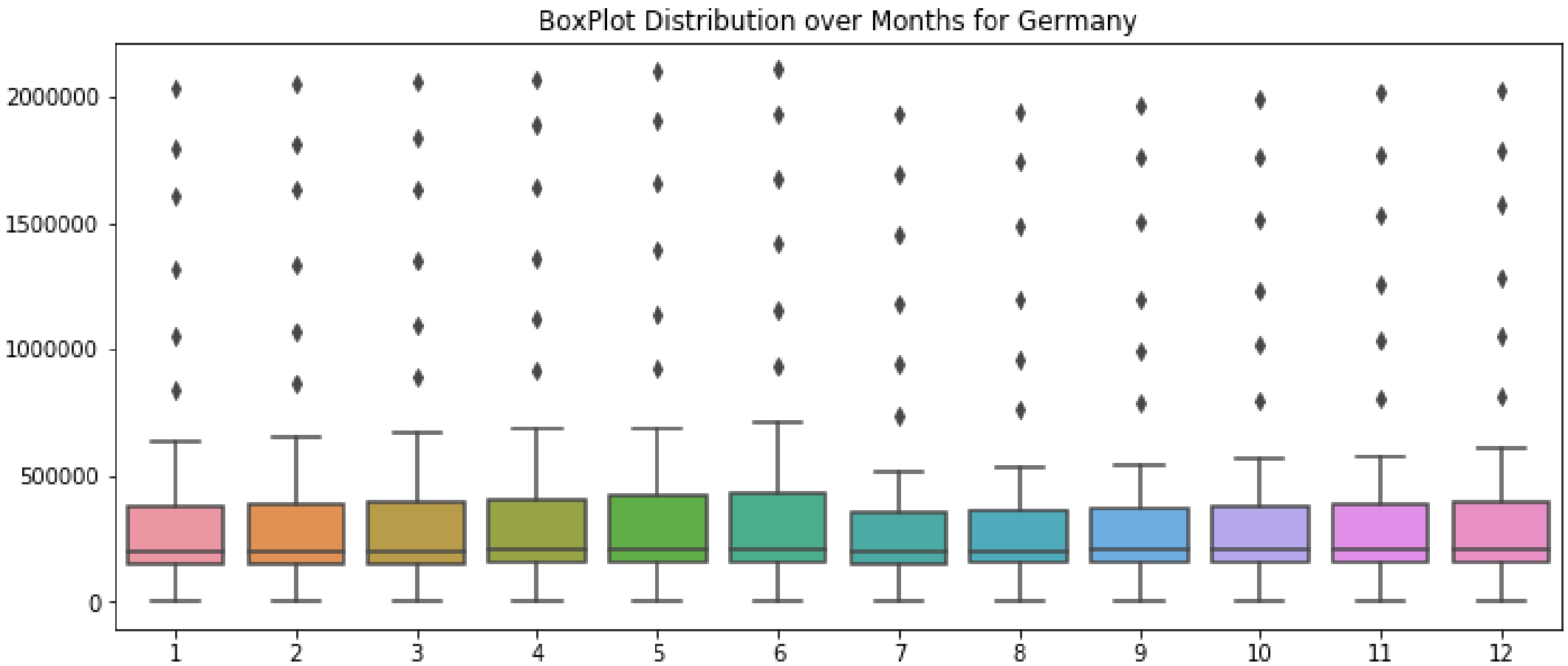
```
"Country":"Bahrain",
"Prediction":{
   "2021-01-15 00:00:00":97110,
   "2021-01-16 00:00:00":97403,
   "2021-01-17 00:00:00":97692,
   "2021-01-18 00:00:00":97976,
   "2021-01-19 00:00:00":98256,
   "2021-01-20 00:00:00":98532,
   "2021-01-21 00:00:00":98803
},
"Actual_Values":{
   "2021-01-15 00:00:00":97020,
   "2021-01-16 00:00:00":97268,
   "2021-01-17 00:00:00":97607,
   "2021-01-18 00:00:00":97940,
   "2021-01-19 00:00:00":98260,
   "2021-01-20 00:00:00":98573,
   "2021-01-21 00:00:00":98878
}
```

DATA STORAGE STRUCTURE

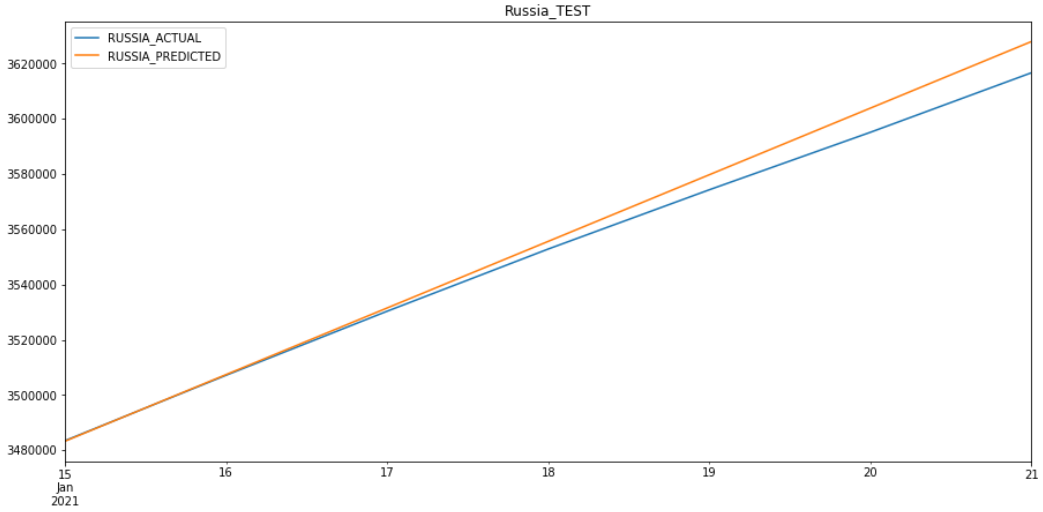
We choose **MongoDB** to store our data. MongoDB is an object-oriented, simple, dynamic, and scalable NoSQL database. One of the biggest advantage of mongodb is that it is dynamic and there is **no rigid schema** for the documentation to store inside. An example of our data storage structure is as followed:

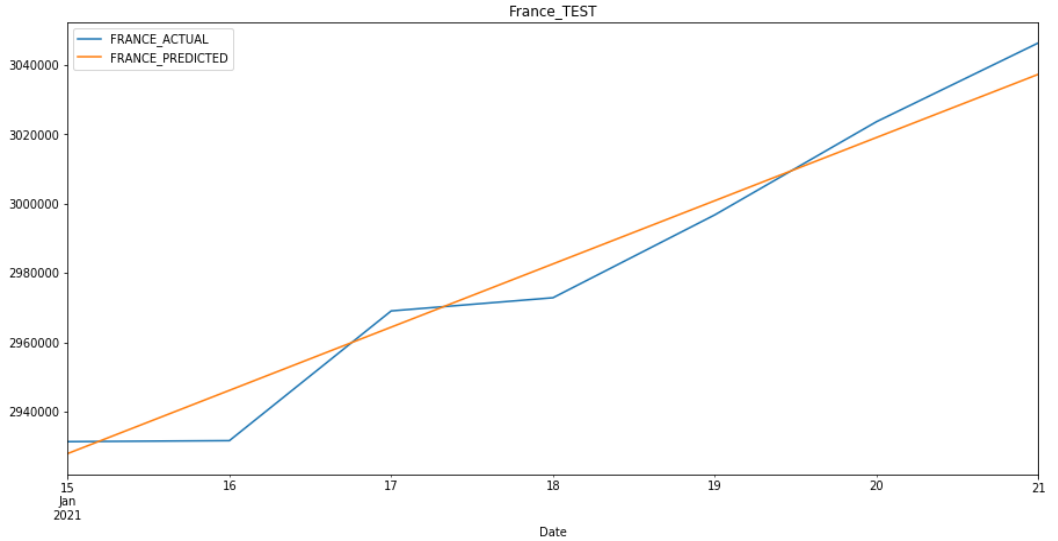

```
{'_id': ObjectId('60119ae5c9cc5683e5fa9a82'), 'Country': 'Germany', 'Prediction': {'2021-01-15 00:00:00': 2035510,
'2021-01-16 00:00:00': 2055786, '2021-01-17 00:00:00': 2076062, '2021-01-18 00:00:00': 2096337, '2021-01-19 00:00:
00': 2116613, '2021-01-20 00:00:00': 2136889, '2021-01-21 00:00:00': 2157164}, 'Actual_Values': {'2021-01-15 00:0
0:00': 2023828, '2021-01-16 00:00:00': 2038645, '2021-01-17 00:00:00': 2050129, '2021-01-18 00:00:00': 2059382, '2
021-01-19 00:00:00': 2071615, '2021-01-20 00:00:00': 2100618, '2021-01-21 00:00:00': 2108895}}
```

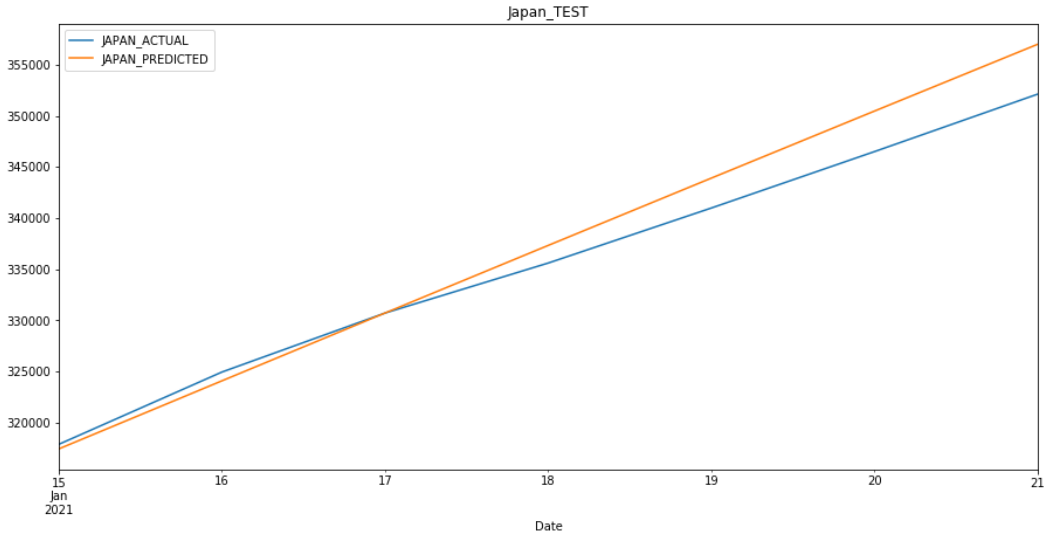
BOX-PLOT DEVELOPMENT

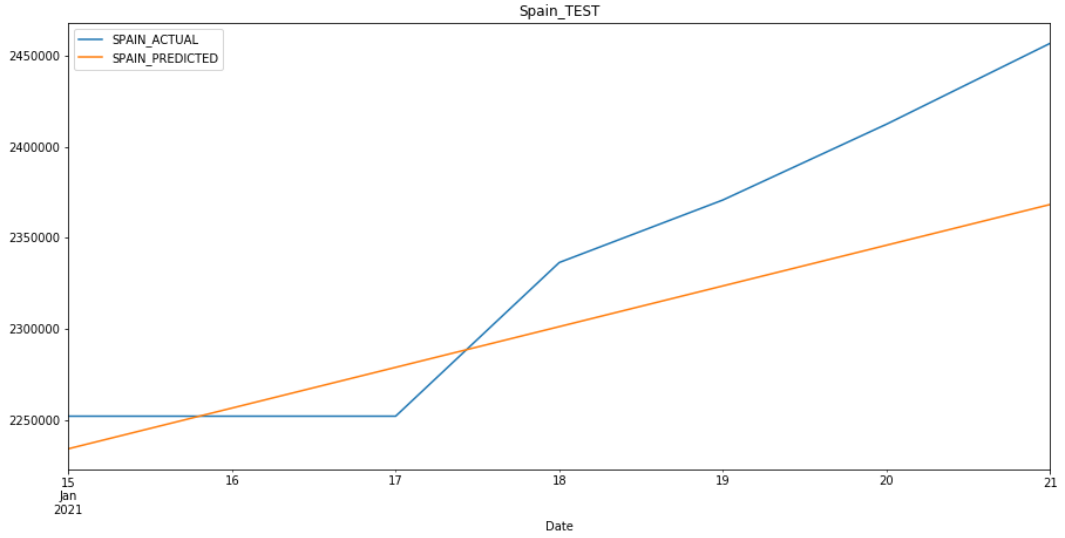


PREDICTION FOR OTHER COUNTRIES



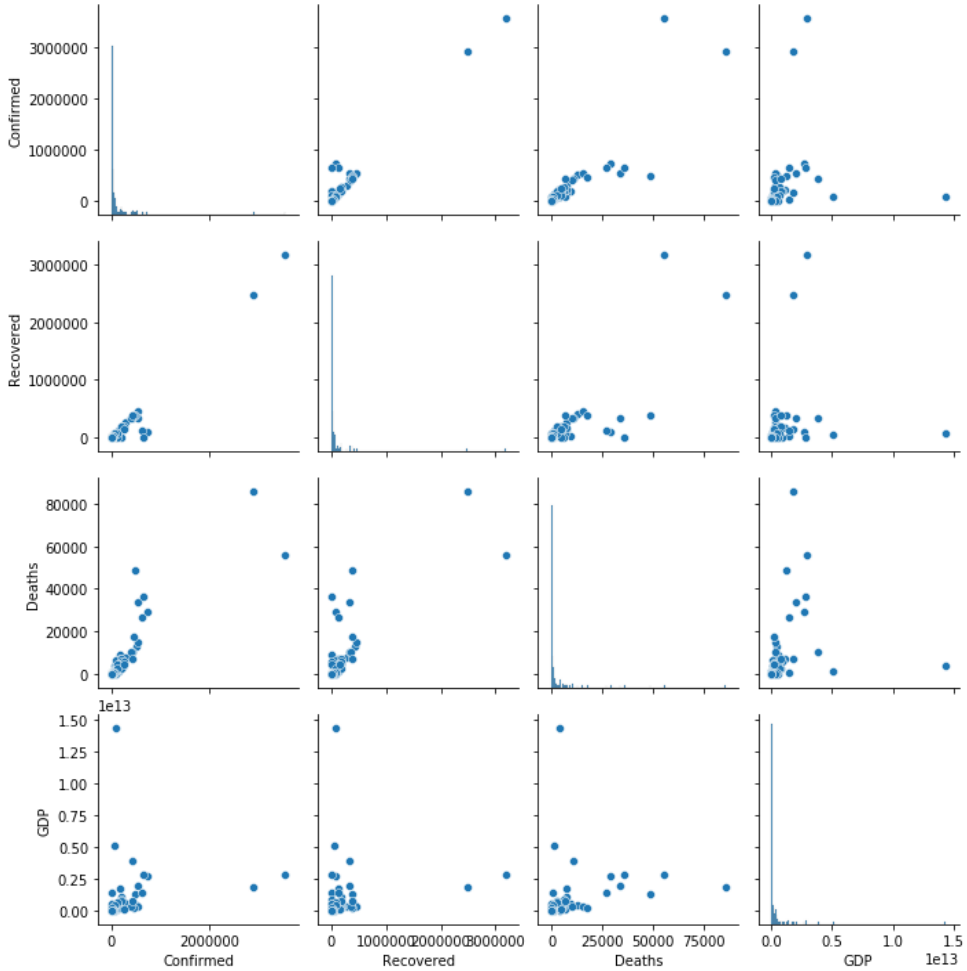




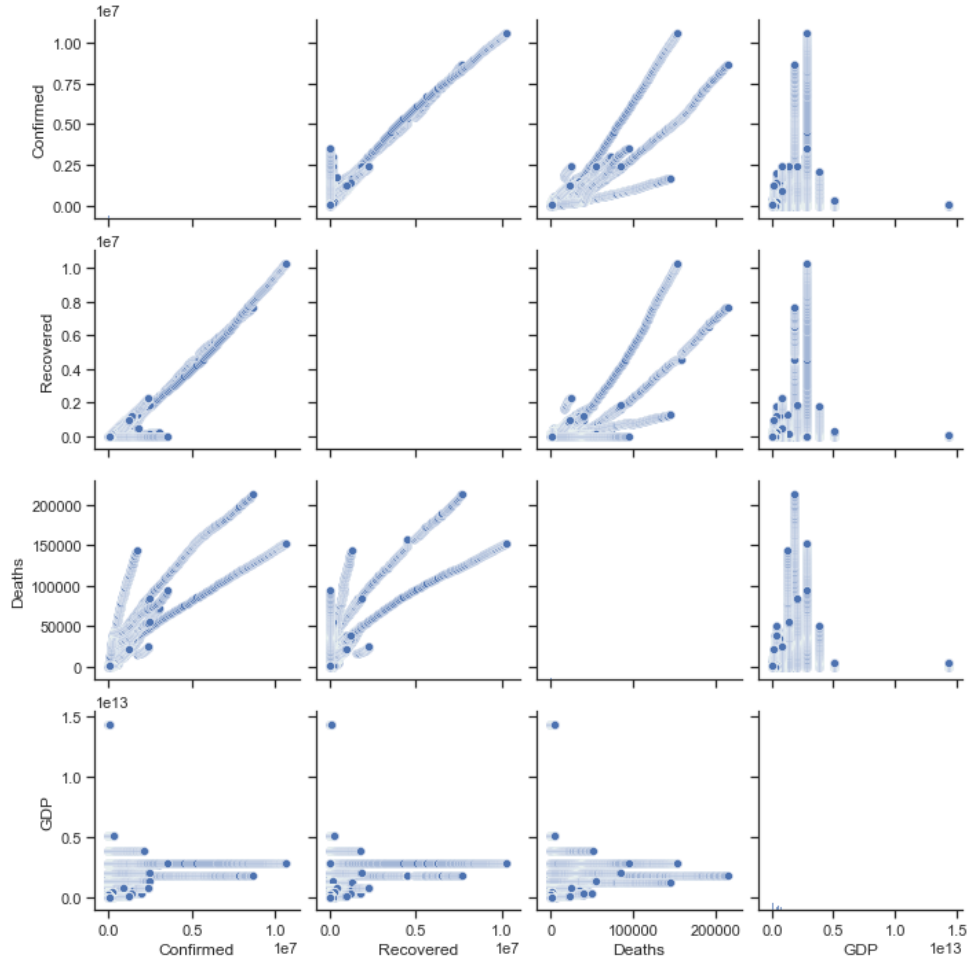


INTERESTING INSIGHTS AND CONCLUSION

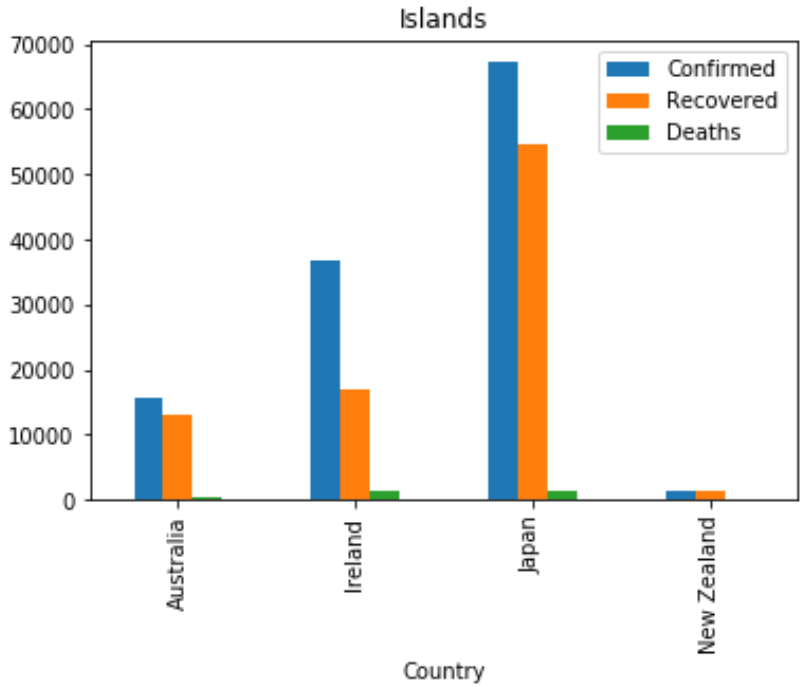
INSIGHTS

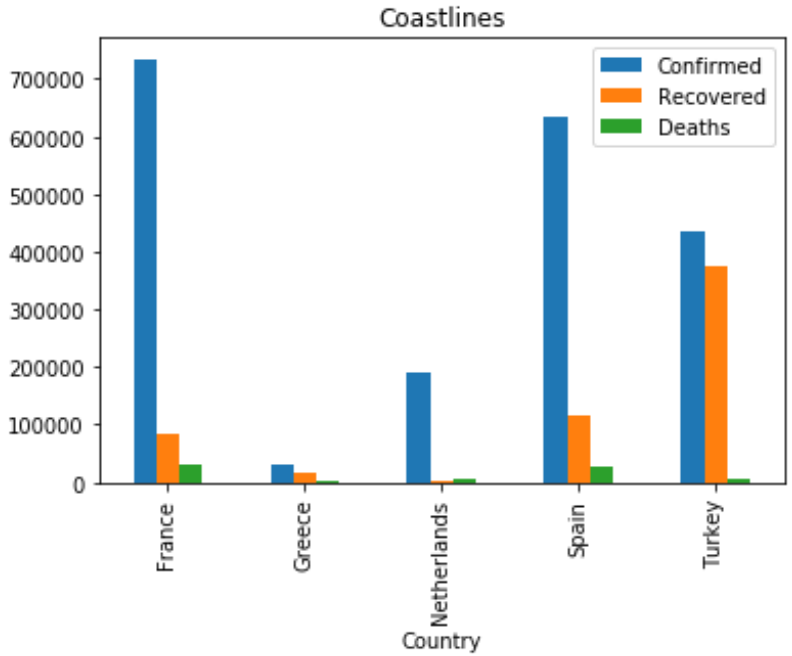


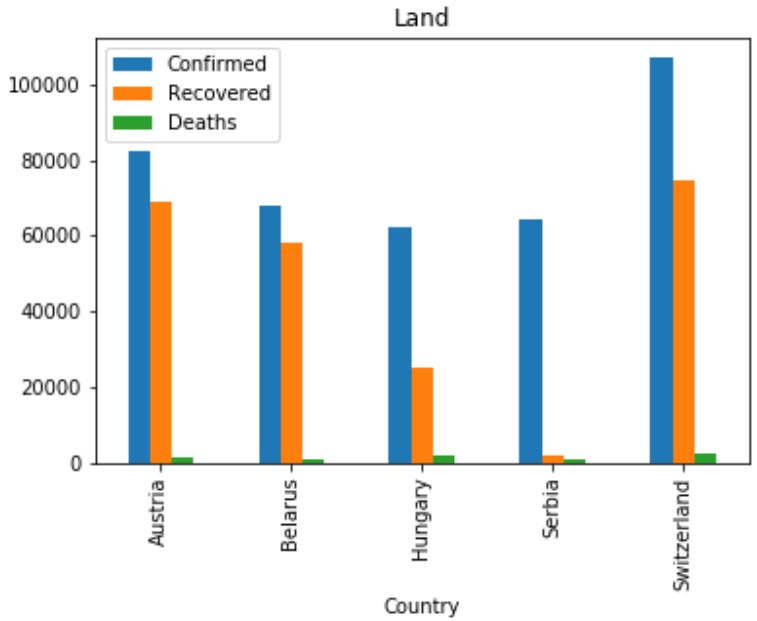
ON THE GDP



INSIGHTS ON REGIONS







ADJUSTING THE MODEL

```
11  def check_country(country):
12
13      if(country not in set(df["Country"])):
14          raise NoSuchCountry("Sorry the country was not found in the dataset.")
15
16      data = df[df["Country"]==country]
17
18      # Does not work
19      # Q3 = data.max().Confirmed*0.75
20      # data = data[data.Confirmed >= Q3]
21
22
23      data_train = data[:len(data)-7]
24      data_test = data[-7:]
```

CHANGING ORDER IN MODEL

CHANGING #TREES AND RAND_STATE

FILTERING OUT ZEROS

CONCLUSION

RANDOM FORESTS NEED TO BE TRAINED ON THE WHOLE DATA

IF SEASONAL DATA USE SARIMAX

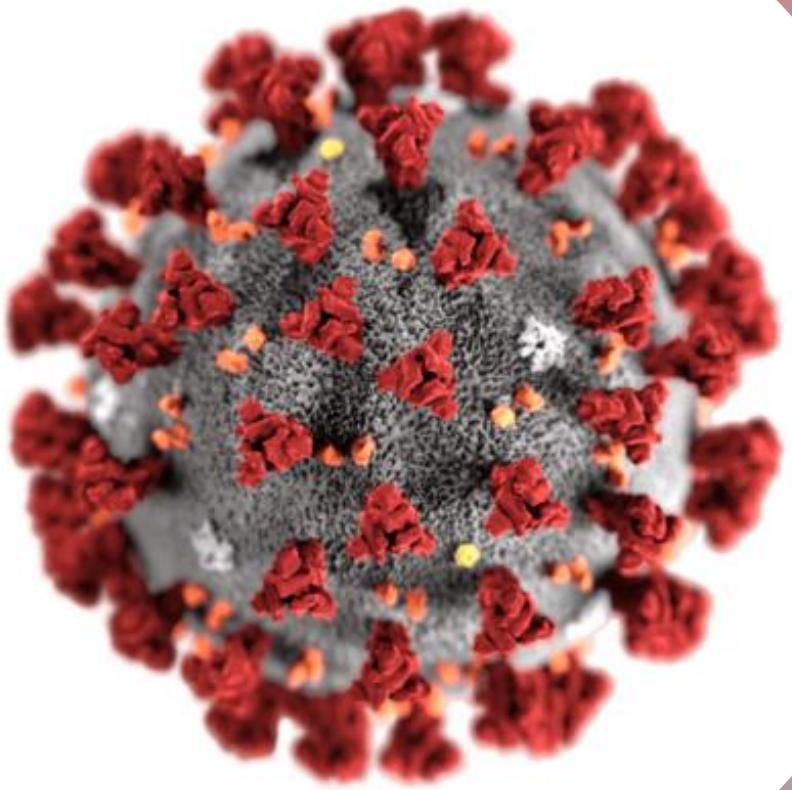
GDP DOES NOT HAVE AN INFLUENCE ON THE FIRST GLANCE

DATA CLEANING AND FILTERING ARE A HUGE PROCESS

REGION OF A COUNTRY CAN BE INFLUENTIAL

WE CAN DERIVE A HYPOTHESIS BUT CAN NOT BE 100% SURE WHEN INTERPRETING DATA

FUTURE APPROACHES



FUTURE WORK

As phase two of the project, the team would like to study the financial impacts within the health industry that has been raised due to the COVID deaths and recovered cases.

FUTURE WORK

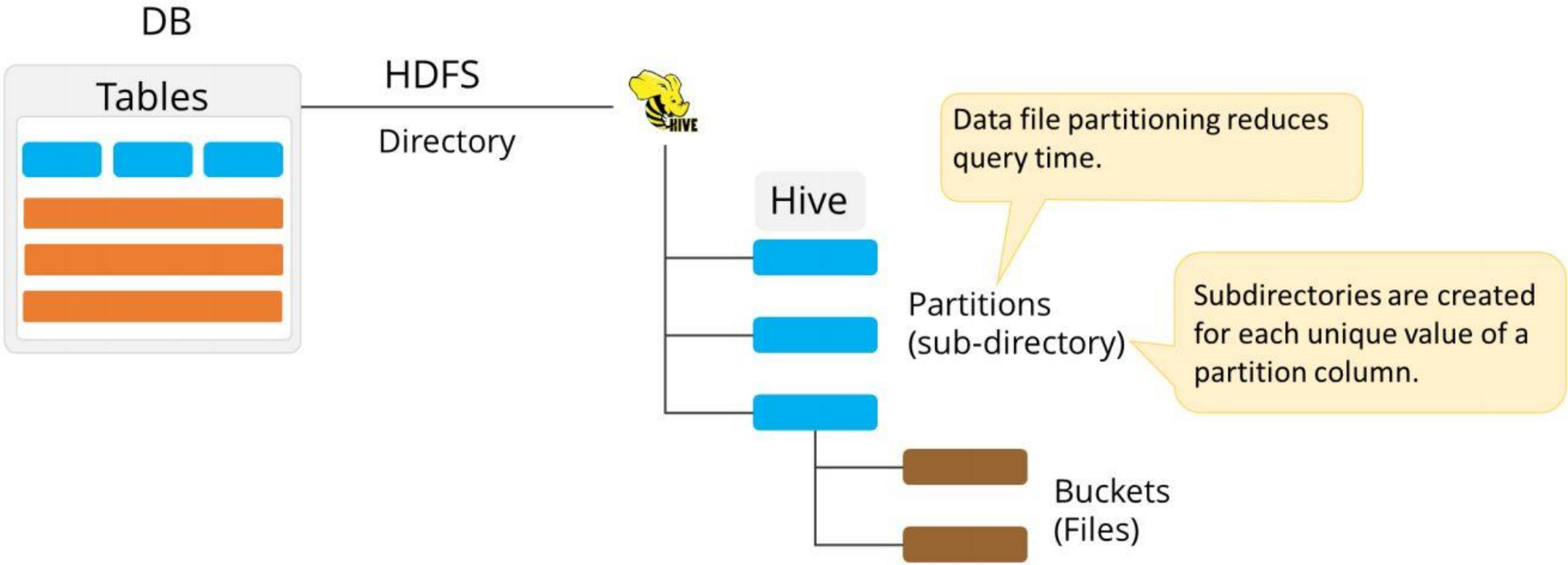
REAL TIME DATA & STREAM PROCESSING

ANALYZING CORONA POLICIES FOR EFFECTIVITY

VISUALIZING CORONA HOTSPOTS BASED ON GEOINFORMATION

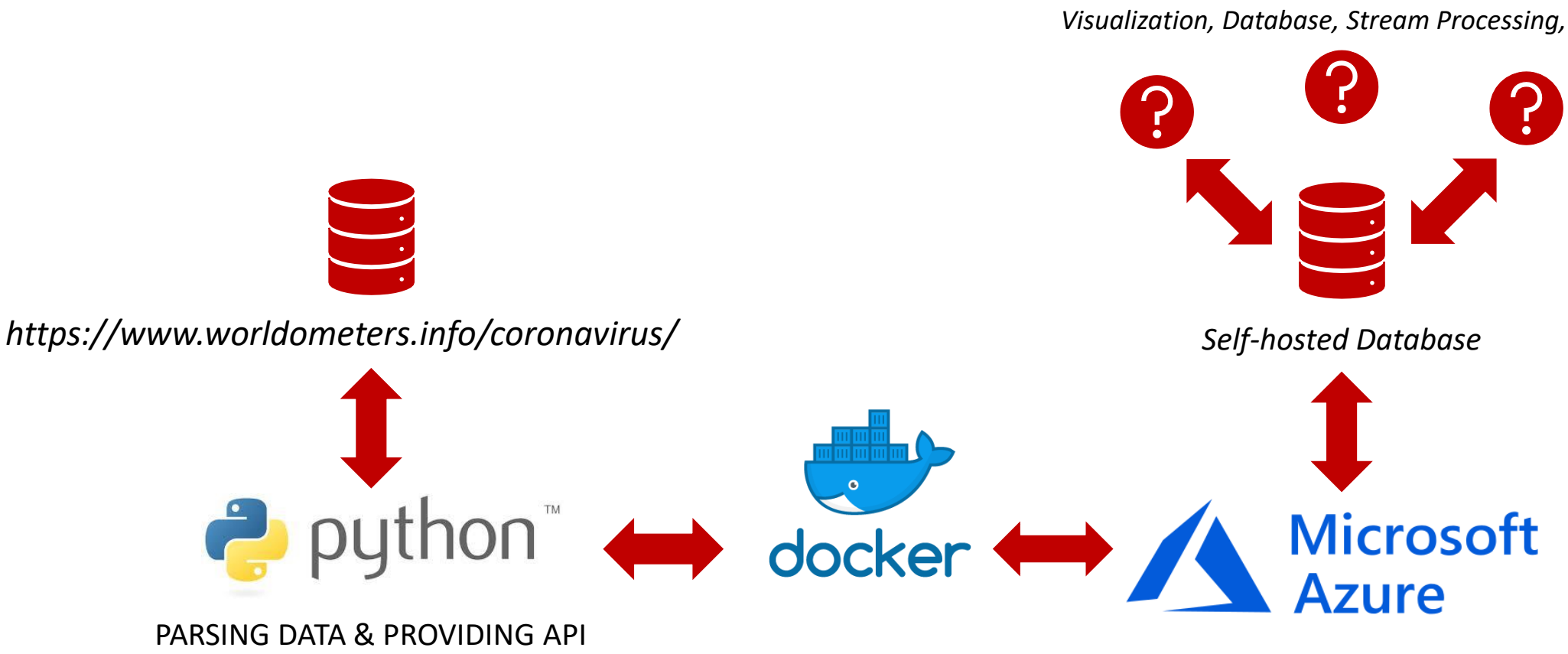
ORGANIZING DATA IN HADOOP

1. **Directory:** which makes it easier to navigate where the location of the data
2. **Hive:** used to perform data analysis, querying on data, and data summarization on large volume datasets
3. **Partitions:** aids in distributing the execution load horizontally and helps in faster execution of queries
4. **Buckets:** dividing hive partition into number of equal culsters, making it easy to retrieve data



USING DOCKER

Receiving live COVID-19 data from a data source for different countries



VISUALIZATION TOOLS

COMPASS:

One of the most powerful tools used to visualize data that are stored on MongoDB because:

- Enables comprehensive analysis via Graphical User Interface (GUI)
- Provides real-time view of data
- Represents data in the form of histograms

mongoDB. COMPASS:

NAVIGATE AND VISUALIZE

https://www.mongodb.com/products/charts

VISUALIZATION TOOLS

https://www.mongodb.com/products/charts

STREAM PROCESSING

APACHE KAFKA:

Apache Kafka is a database that acts as a streaming platform for messaging, storage. Processing,. Its main advantage is that it integrates real time data with 0 downtime and 0 data loss.

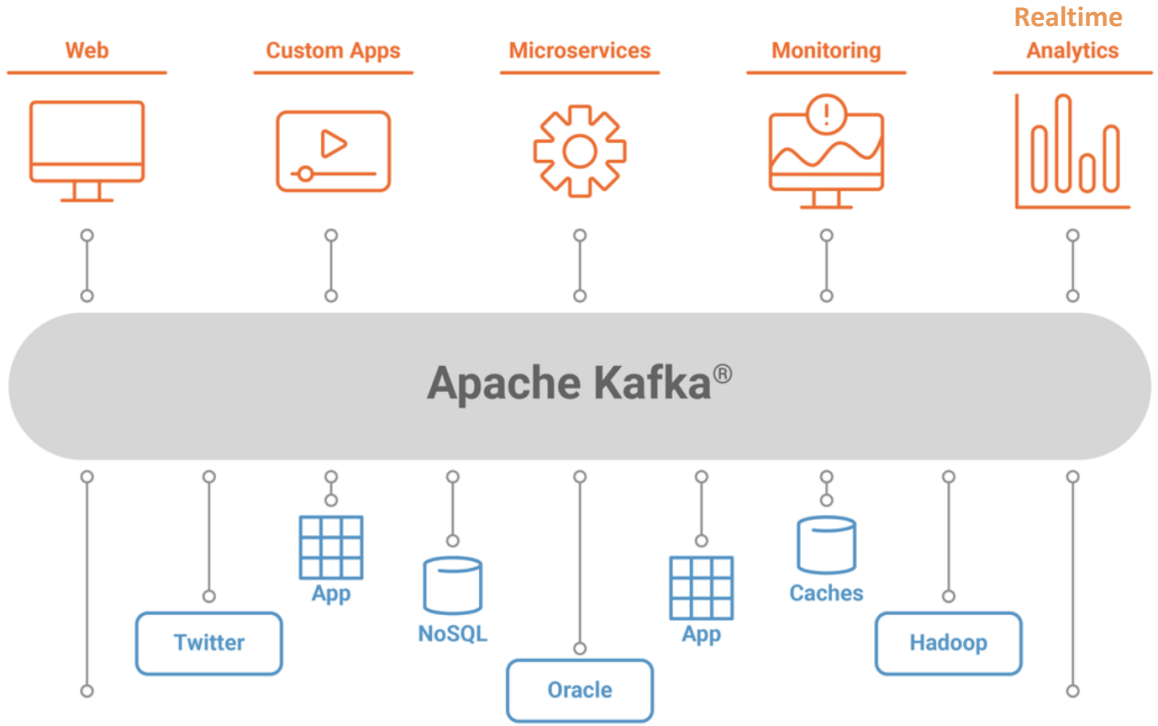
Project Application:

Can be used in our project because it stores data in the form of logs and manages these logs, meaning that the retrieval of data and events regarding time is very precise.

Advantages:

- Open source
- Free of cost
- Cloud hosting on multiple datacenters

https://kafka.apache.org/



=> Web-Interfaces on insights for people of interest

Joining, Querying, Monitoring

THANK YOU FOR YOUR ATTENTION

ANY QUESTIONS?

ALL LINKS HAVE LAST BEEN ACCESSED AT THE 28TH OF JANUARY 2021