BIG DATA ANALYTICS

Prof. Dr. Martin Theobald & Dr. Vinu Venugopal

Summer Semester 2019-20

University of Luxembourg

Final Project

Submitted By

Saddam Hossain (019027635C)

Instructions:

1. I have added all the source code to my final project folder.
2. I have used databricks as a cluster because HPC not allow me to install all packages I needed.
3. Inside the folder I added all output of the problem, databricks files, python files, report, video presentation and data sources
5. I have compiled all the code on databricks cluster . So execution may vary when you test my code.

Title

Coronavirus (COVID-19) Data Visualization Using Pyspark

Abstract

Import the data, get all the dates for the outbreak, getting daily increases and moving averages, convert integer into datetime for better visualization and graphing the number of confirmed cases, active cases, deaths, recoveries, mortality rate (CFR), and recovery rate.

Problem Definition

Coronavirus data visualization and graphing the number of confirmed cases, active cases, deaths, recoveries, mortality rate (CFR), and recovery rate. Showing country specific data.

Big Data Perspectives:

One example is how big data can play a huge role during a pandemic like Covid 19.

“To help providers detect which patients are most likely to have severe cases of COVID-19, McDevitt and his team leveraged artificial intelligence and big data to produce COVID-19 severity scores.

Utilizing data from 160 hospitalized patients in Wuhan, China, the researchers identified four biomarkers measured in blood tests that were significantly elevated in patients who died versus those who recovered, including the C-reactive protein, myoglobin, procalcitonin, and cardiac troponin I.” **Source: healthitanalytics**

Approach

I have used pyspark for doing the visualization and databricks as a cluster because our hpc platform does not allow me to install some packages and libraries I have used in this project.

Used Packages and Libraries

```
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.colors as mcolors
import pandas as pd
import random
import math
import time
from sklearn.linear_model import LinearRegression, BayesianRidge
from sklearn.model_selection import RandomizedSearchCV,
train_test_split
from sklearn.preprocessing import PolynomialFeatures
from sklearn.svm import SVR
from sklearn.metrics import mean_squared_error, mean_absolute_error
import datetime
import operator
plt.style.use('fivethirtyeight')
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")
import seaborn as sns
import sklearn
import random
import os
from pyspark.sql.functions import *
from pyspark import SparkContext
from pyspark.sql import SparkSession
from pyspark.ml  import Pipeline
from pyspark.sql import SQLContext
from pyspark.sql.functions import mean,col,split, col,
regexp_extract, when, lit
```

```
from pyspark.ml.feature import StringIndexer, VectorAssembler
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
from pyspark.ml.feature import QuantileDiscretizer
```

Spark Session initiate

```
spark = SparkSession \
.builder \
.appName("Covid 19 Data Analysis with pyspark") \
.config("spark.some.config.option", "some-value") \
.getOrCreate()
```

Data Frame created using a daily report from COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.

```
df = spark.read.format('com.databricks.spark.csv').\
options(header='true', \
inferschema='true').\
load("/FileStore/tables/*.csv",header=True)
```

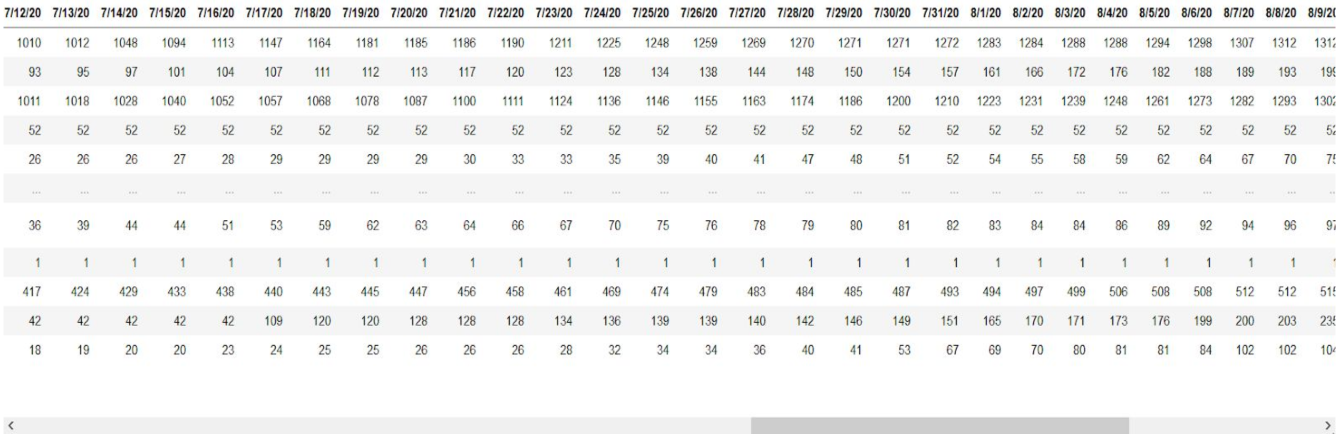
Separate Data Frame created for active, recoveries, deaths and confirmed cases

```
confirmed_dataFrame =
spark.read.format('com.databricks.spark.csv').\
options(header='true', \
inferschema='true').\
load("/FileStore/tables/confirmed_data/time_series_covid19_
confirmed_global.csv",header=True)

deaths_dataFrame =
spark.read.format('com.databricks.spark.csv').\
options(header='true', \
inferschema='true').\
load("/FileStore/tables/deaths_data/time_series_covid19_dea
ths_global.csv",header=True)


latest_data =
spark.read.format('com.databricks.spark.csv').\
options(header='true', \
inferschema='true').\
load("/FileStore/tables/latest_data/08_19_2020.csv",header=
True)

recoveries_dataFrame =
spark.read.format('com.databricks.spark.csv').\
options(header='true', \
inferschema='true').\
load("/FileStore/tables/recoveries_data/time_series_covid19
_recovered_global.csv",header=True)

apple_mobility =
spark.read.format('com.databricks.spark.csv').\
options(header='true', \
inferschema='true').\
load("/FileStore/tables/apple_mobility/applemobilitytrends_
2020_08_18.csv",header=True)
```

Convert all the data frame into pandas Library

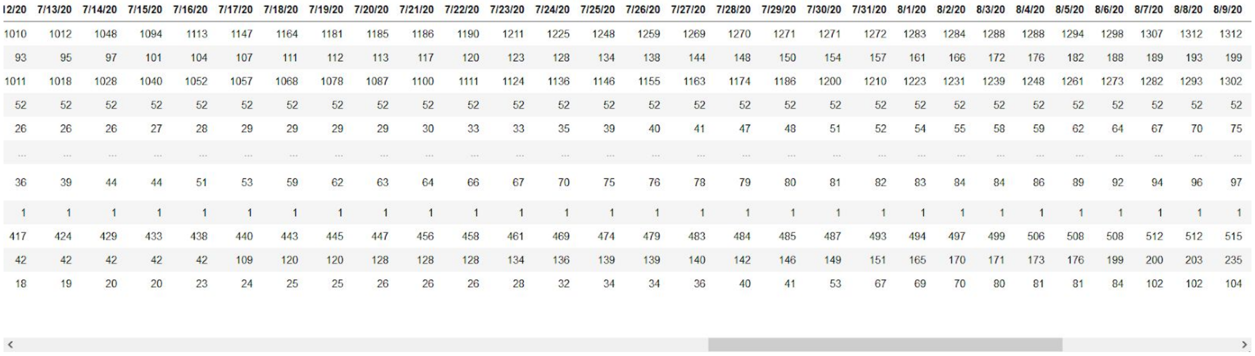
```
confirmed_data = confirmed_dataFrame.toPandas()
deaths_data = deaths_dataFrame.toPandas()
latest_cases = latest_data.toPandas()
recovered_data = recoveries_dataFrame.toPandas()
apple_data = apple_mobility.toPandas()
```

Overview of confirmed cases



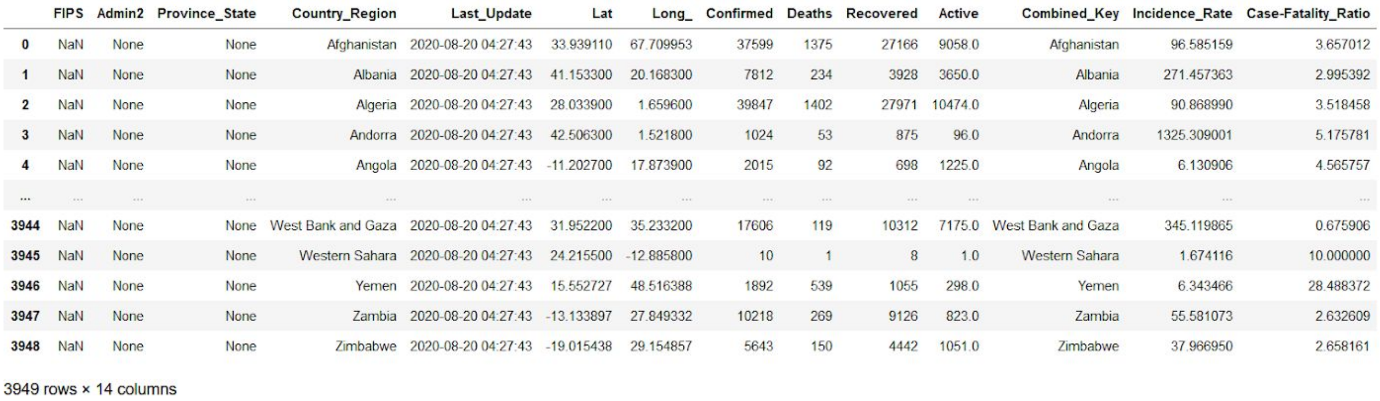
7/12/20	7/13/20	7/14/20	7/15/20	7/16/20	7/17/20	7/18/20	7/19/20	7/20/20	7/21/20	7/22/20	7/23/20	7/24/20	7/25/20	7/26/20	7/27/20	7/28/20	7/29/20	7/30/20	7/31/20	8/1/20	8/2/20	8/3/20	8/4/20	8/5/20	8/6/20	8/7/20	8/8/20	8/9/20
1010	1012	1048	1094	1113	1147	1164	1181	1185	1186	1190	1211	1225	1248	1259	1269	1270	1271	1271	1272	1283	1284	1288	1288	1294	1298	1307	1312	1312
93	95	97	101	104	107	111	112	113	117	120	123	128	134	138	144	148	150	154	157	161	166	172	176	182	188	189	193	199
1011	1018	1028	1040	1052	1057	1068	1078	1087	1100	1111	1124	1136	1146	1155	1163	1174	1186	1200	1210	1223	1231	1239	1248	1261	1273	1282	1293	1302
52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52
26	26	26	27	28	29	29	29	29	30	33	33	35	39	40	41	47	48	51	52	54	55	58	59	62	64	67	70	75
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
36	39	44	44	51	53	59	62	63	64	66	67	70	75	76	78	79	80	81	82	83	84	84	86	89	92	94	96	97
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
417	424	429	433	438	440	443	445	447	456	458	461	469	474	479	483	484	485	487	493	494	497	499	506	508	508	512	512	515
42	42	42	42	42	109	120	120	128	128	128	134	136	139	139	140	142	146	149	151	165	170	171	173	176	199	200	203	235
18	19	20	20	23	24	25	25	26	26	26	28	32	34	34	36	40	41	53	67	69	70	80	81	81	84	102	102	104

Overview of death cases in the world



12/20	7/13/20	7/14/20	7/15/20	7/16/20	7/17/20	7/18/20	7/19/20	7/20/20	7/21/20	7/22/20	7/23/20	7/24/20	7/25/20	7/26/20	7/27/20	7/28/20	7/29/20	7/30/20	7/31/20	8/1/20	8/2/20	8/3/20	8/4/20	8/5/20	8/6/20	8/7/20	8/8/20	8/9/20
1010	1012	1048	1094	1113	1147	1164	1181	1185	1186	1190	1211	1225	1248	1259	1269	1270	1271	1271	1272	1283	1284	1288	1288	1294	1298	1307	1312	1312
93	95	97	101	104	107	111	112	113	117	120	123	128	134	138	144	148	150	154	157	161	166	172	176	182	188	189	193	199
1011	1018	1028	1040	1052	1057	1068	1078	1087	1100	1111	1124	1136	1146	1155	1163	1174	1186	1200	1210	1223	1231	1239	1248	1261	1273	1282	1293	1302
52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52	52
26	26	26	27	28	29	29	29	29	30	33	33	35	39	40	41	47	48	51	52	54	55	58	59	62	64	67	70	75
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
36	39	44	44	51	53	59	62	63	64	66	67	70	75	76	78	79	80	81	82	83	84	84	86	89	92	94	96	97
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
417	424	429	433	438	440	443	445	447	456	458	461	469	474	479	483	484	485	487	493	494	497	499	506	508	508	512	512	515
42	42	42	42	42	109	120	120	128	128	128	134	136	139	139	140	142	146	149	151	165	170	171	173	176	199	200	203	235
18	19	20	20	23	24	25	25	26	26	26	28	32	34	34	36	40	41	53	67	69	70	80	81	81	84	102	102	104

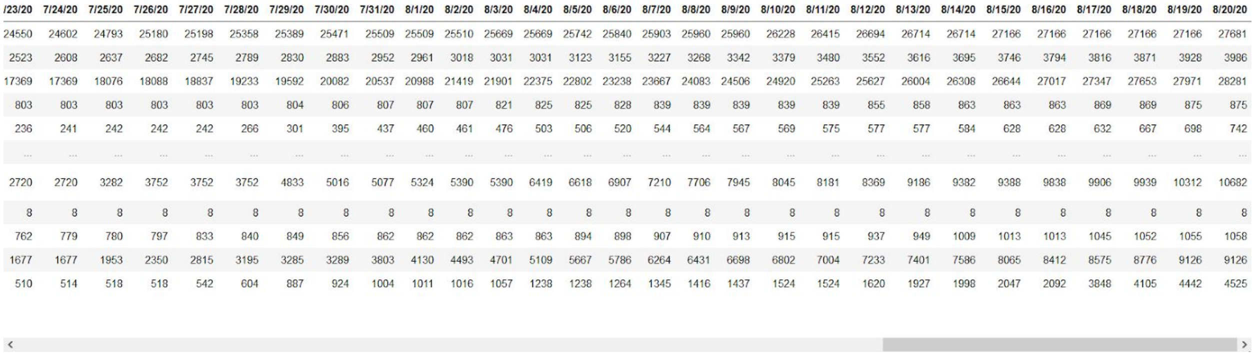
Overview of Latest cases in the world



	FIPS	Admin2	Province_State	Country_Region	Last_Update	Lat	Long_	Confirmed	Deaths	Recovered	Active	Combined_Key	Incidence_Rate	Case-Fatality_Ratio
0	NaN	None	None	Afghanistan	2020-08-20 04:27:43	33.939110	67.709953	37599	1375	27166	9058.0	Afghanistan	96.585159	3.657012
1	NaN	None	None	Albania	2020-08-20 04:27:43	41.153300	20.168300	7812	234	3928	3650.0	Albania	271.457363	2.995392
2	NaN	None	None	Algeria	2020-08-20 04:27:43	28.033900	1.659600	39847	1402	27971	10474.0	Algeria	90.868990	3.518458
3	NaN	None	None	Andorra	2020-08-20 04:27:43	42.506300	1.521800	1024	53	875	96.0	Andorra	1325.309001	5.175781
4	NaN	None	None	Angola	2020-08-20 04:27:43	-11.202700	17.873900	2015	92	698	1225.0	Angola	6.130906	4.565757
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3944	NaN	None	None	West Bank and Gaza	2020-08-20 04:27:43	31.952200	35.233200	17606	119	10312	7175.0	West Bank and Gaza	345.119865	0.675906
3945	NaN	None	None	Western Sahara	2020-08-20 04:27:43	24.215500	-12.885800	10	1	8	1.0	Western Sahara	1.674116	10.000000
3946	NaN	None	None	Yemen	2020-08-20 04:27:43	15.552727	48.516388	1892	539	1055	298.0	Yemen	6.343466	28.488372
3947	NaN	None	None	Zambia	2020-08-20 04:27:43	-13.133897	27.849332	10218	269	9126	823.0	Zambia	55.581073	2.632609
3948	NaN	None	None	Zimbabwe	2020-08-20 04:27:43	-19.015438	29.154857	5643	150	4442	1051.0	Zimbabwe	37.966950	2.658161

3949 rows × 14 columns

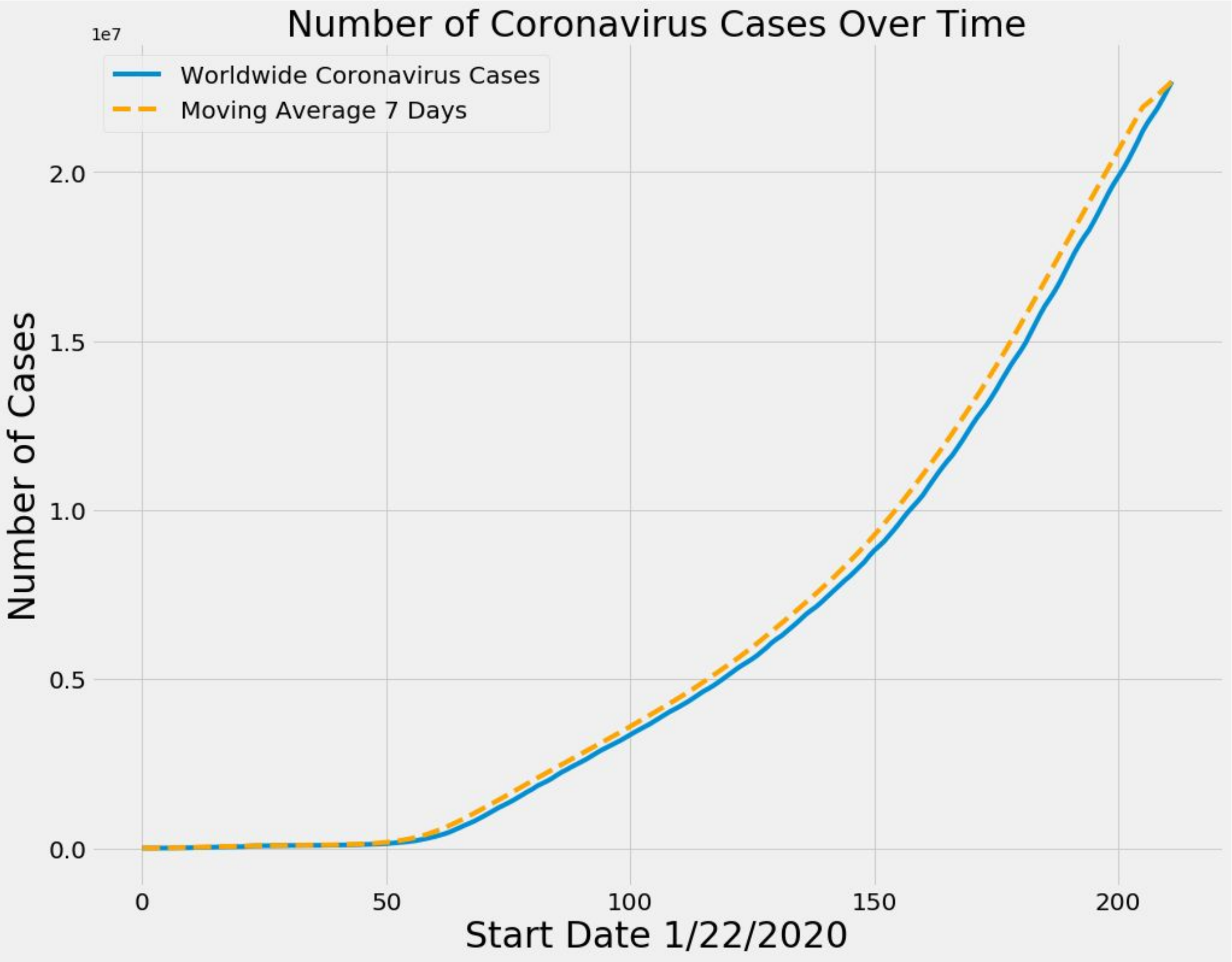
Overview of recovered cases in the world



/23/20	7/24/20	7/25/20	7/26/20	7/27/20	7/28/20	7/29/20	7/30/20	7/31/20	8/1/20	8/2/20	8/3/20	8/4/20	8/5/20	8/6/20	8/7/20	8/8/20	8/9/20	8/10/20	8/11/20	8/12/20	8/13/20	8/14/20	8/15/20	8/16/20	8/17/20	8/18/20	8/19/20	8/20/20
24550	24602	24793	25180	25198	25358	25389	25471	25509	25509	25510	25669	25669	25742	25840	25903	25960	25960	26228	26415	26694	26714	26714	27166	27166	27166	27166	27166	27681
2523	2608	2637	2682	2745	2789	2830	2883	2952	2961	3018	3031	3031	3123	3155	3227	3268	3342	3379	3480	3552	3616	3695	3746	3794	3816	3871	3928	3986
17369	17369	18076	18088	18837	19233	19592	20082	20537	20988	21419	21901	22375	22802	23238	23667	24083	24506	24920	25263	25627	26004	26308	26644	27017	27347	27653	27971	28281
803	803	803	803	803	803	804	806	807	807	807	821	825	825	828	839	839	839	839	839	855	858	863	863	863	869	869	875	875
236	241	242	242	242	266	301	395	437	460	461	476	503	506	520	544	564	567	569	575	577	577	584	628	628	632	667	698	742
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2720	2720	3282	3752	3752	3752	4833	5016	5077	5324	5390	5390	6419	6618	6907	7210	7706	7945	8045	8181	8369	9186	9382	9388	9838	9906	9939	10312	10682
8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
762	779	780	797	833	840	849	856	862	862	862	863	863	894	898	907	910	913	915	915	937	949	1009	1013	1013	1045	1052	1055	1058
1677	1677	1953	2350	2815	3195	3285	3289	3803	4130	4493	4701	5109	5667	5786	6264	6431	6698	6802	7004	7233	7401	7586	8065	8412	8575	8776	9126	9126
510	514	518	518	542	604	887	924	1004	1011	1016	1057	1238	1238	1264	1345	1416	1437	1524	1524	1620	1927	1998	2047	2092	3848	4105	4442	4525

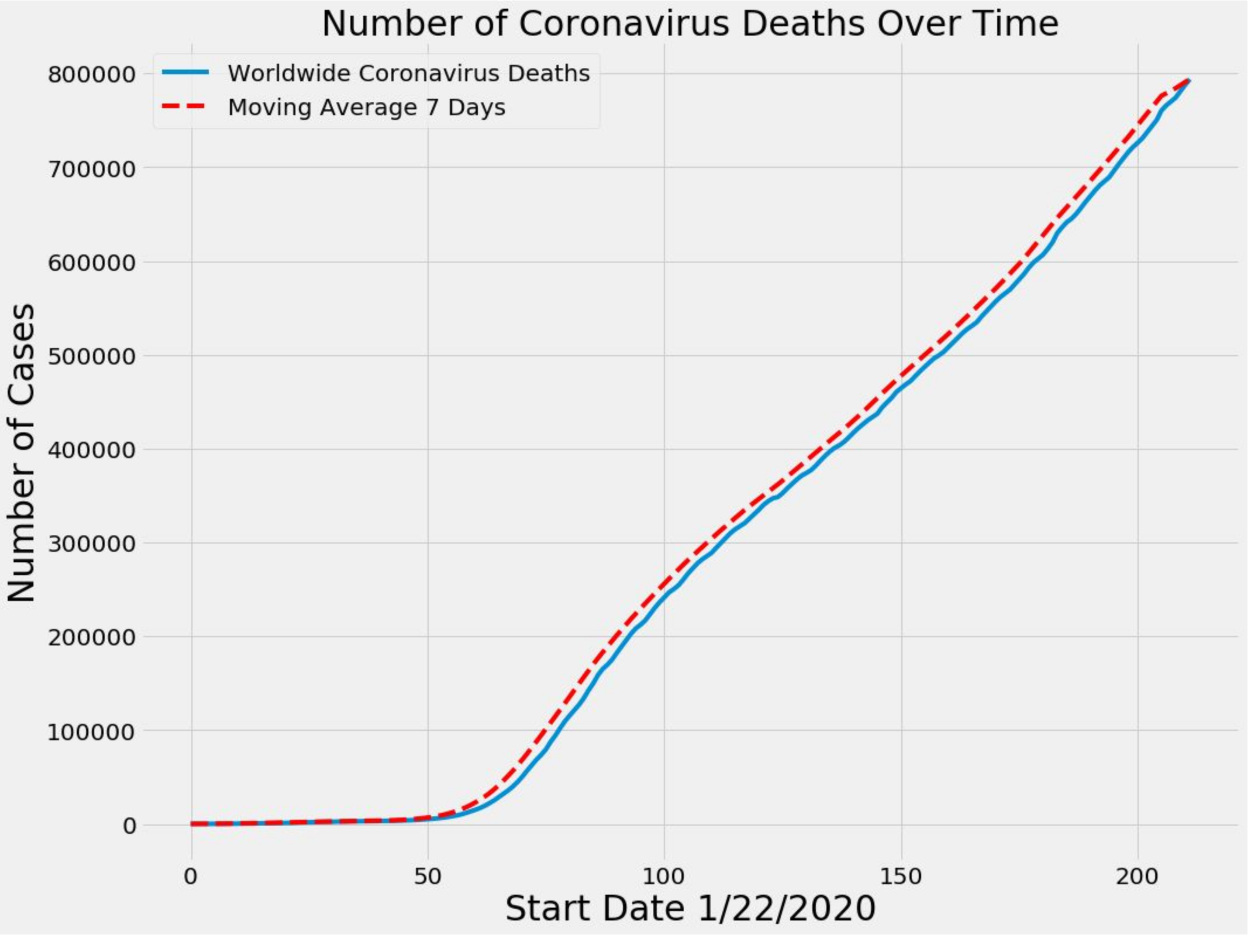
Plotting the graph for number of cases

```python
updated_date = updated_date.reshape(1, -1)[0]
plt.figure(figsize=(15, 12))
plt.plot(updated_date, total_world_cases)
plt.plot(updated_date, world_confirmed_avg, linestyle='dashed',
color='orange')
plt.title('Number of Coronavirus Cases Over Time', size=30)
plt.xlabel('Start Date 1/22/2020', size=30)
plt.ylabel('Number of Cases', size=30)
plt.legend(['Worldwide Coronavirus Cases', 'Moving Average {}
Days'.format(window)], prop={'size': 20})
plt.xticks(size=20)
plt.yticks(size=20)
plt.show()
```



Plotting the graph for number of deaths

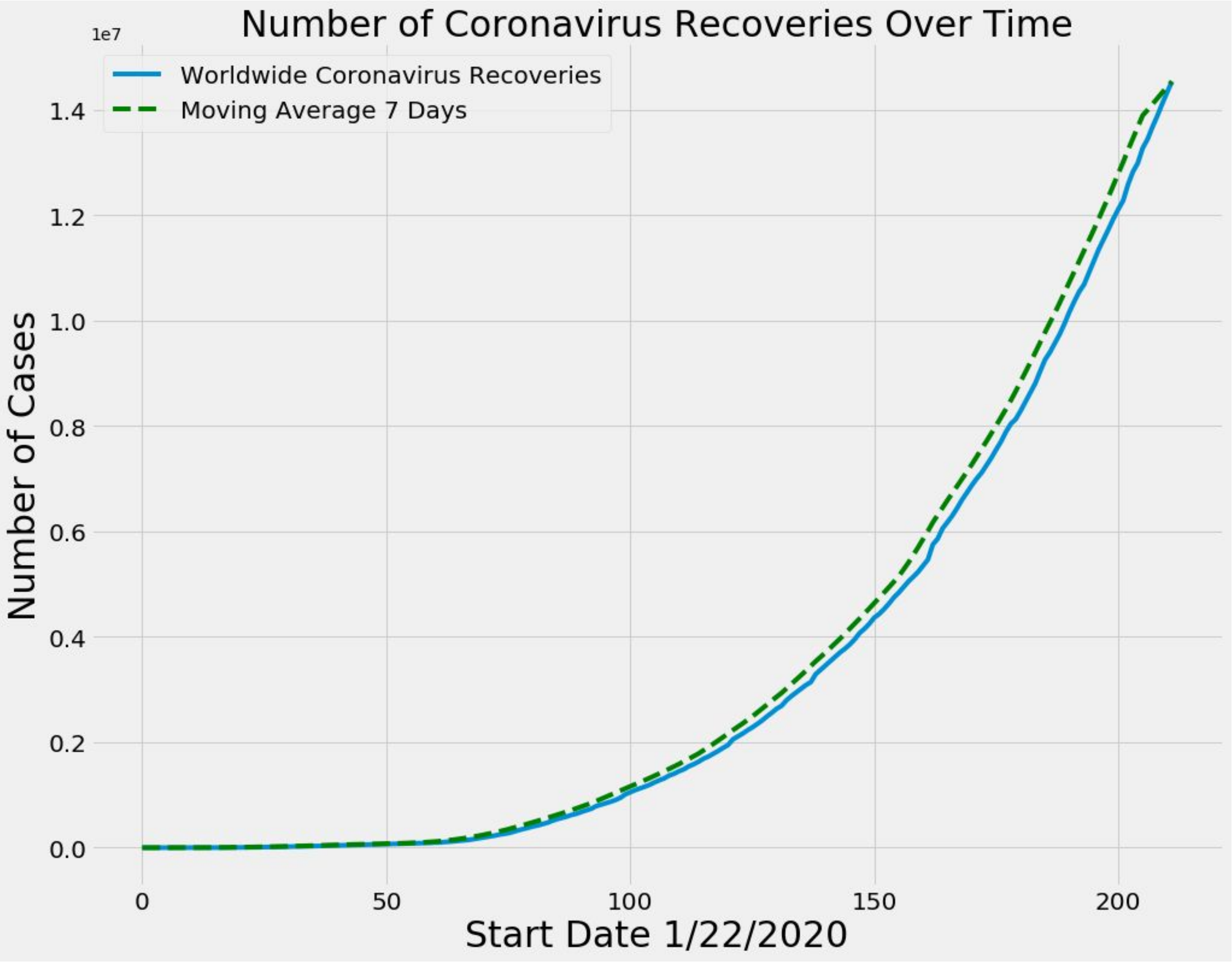
```
plt.figure(figsize=(15, 12))
plt.plot(updated_date, total_world_deaths)
plt.plot(updated_date, world_death_avg, linestyle='dashed',
color='red')
plt.title('Number of Coronavirus Deaths Over Time', size=30)
plt.xlabel('Start Date 1/22/2020', size=30)
plt.ylabel('Number of Cases', size=30)
plt.legend(['Worldwide Coronavirus Deaths', 'Moving Average {}
Days'.format(window)], prop={'size': 20})
plt.xticks(size=20)
plt.yticks(size=20)
plt.show()
```



Plotting the graph for number of Recoveries

```
plt.figure(figsize=(15, 12))
plt.plot(updated_date, total_recovered)
plt.plot(updated_date, world_recovery_avg, linestyle='dashed',
color='green')
plt.title('Number of Coronavirus Recoveries Over Time', size=30)
plt.xlabel('Start Date 1/22/2020', size=30)
plt.ylabel('Number of Cases', size=30)
plt.legend(['Worldwide Coronavirus Recoveries', 'Moving Average {}
Days'.format(window)], prop={'size': 20})
```

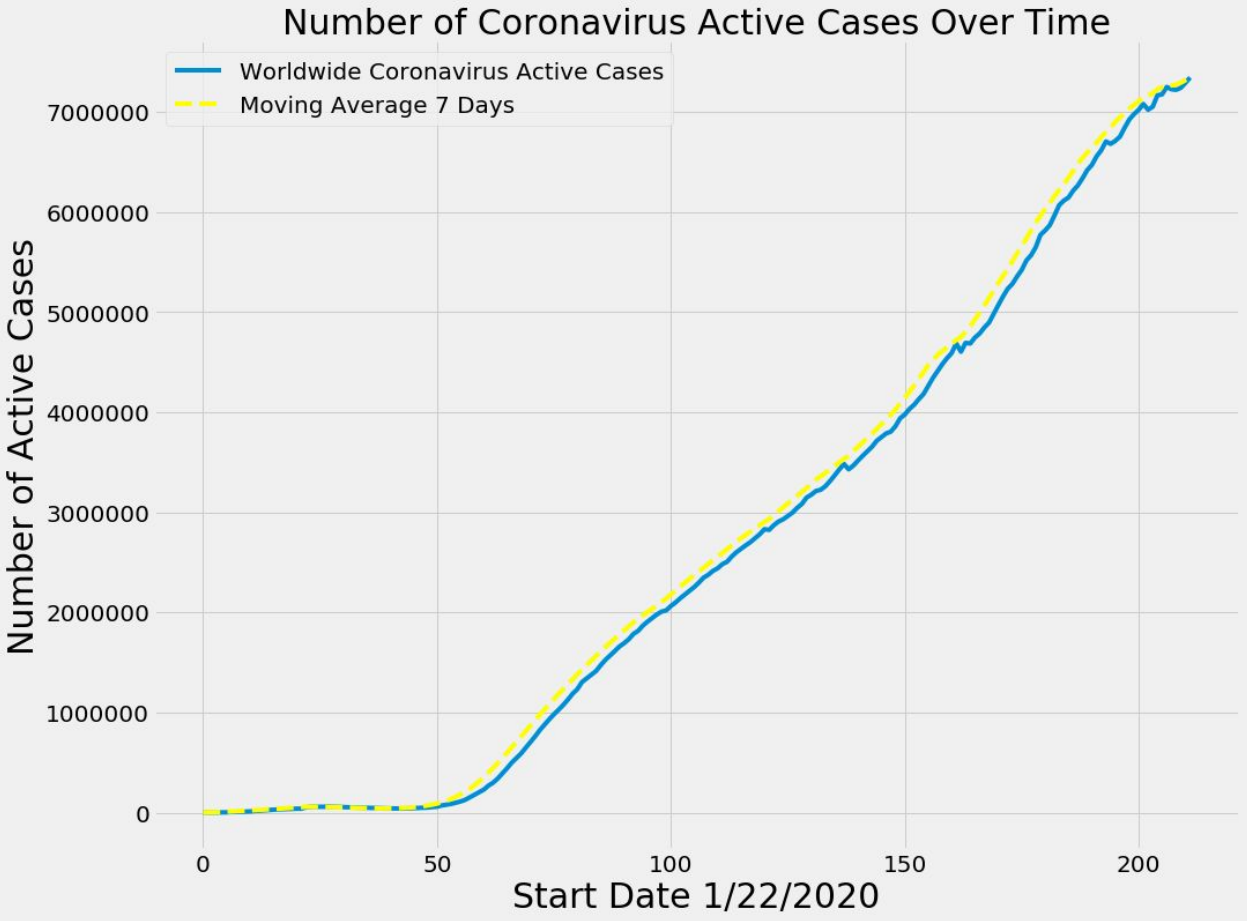
```
plt.xticks(size=20)
plt.yticks(size=20)
plt.show()
```



Plotting the graph for number of active cases

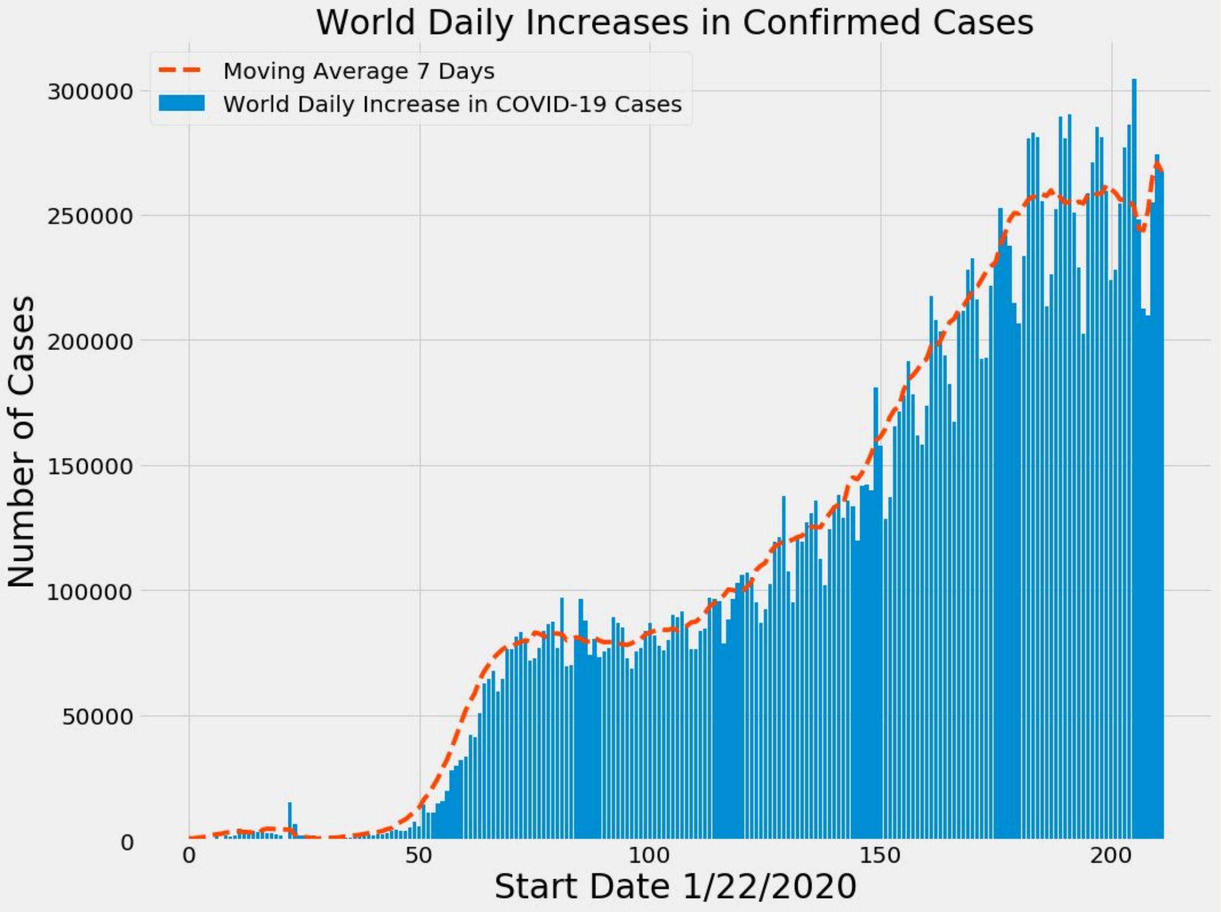
```
plt.figure(figsize=(15, 12))
plt.plot(updated_date, world_total_active)
plt.plot(updated_date, world_active_avg, linestyle='dashed',
color='Yellow')
plt.title('Number of Coronavirus Active Cases Over Time', size=30)
plt.xlabel('Start Date 1/22/2020', size=30)
plt.ylabel('Number of Active Cases', size=30)
plt.legend(['Worldwide Coronavirus Active Cases', 'Moving Average {}
Days'.format(window)], prop={'size': 20})
```

```
plt.xticks(size=20)
plt.yticks(size=20)
plt.show()
```



World Daily Confirmed Cases Increases in Bar chart

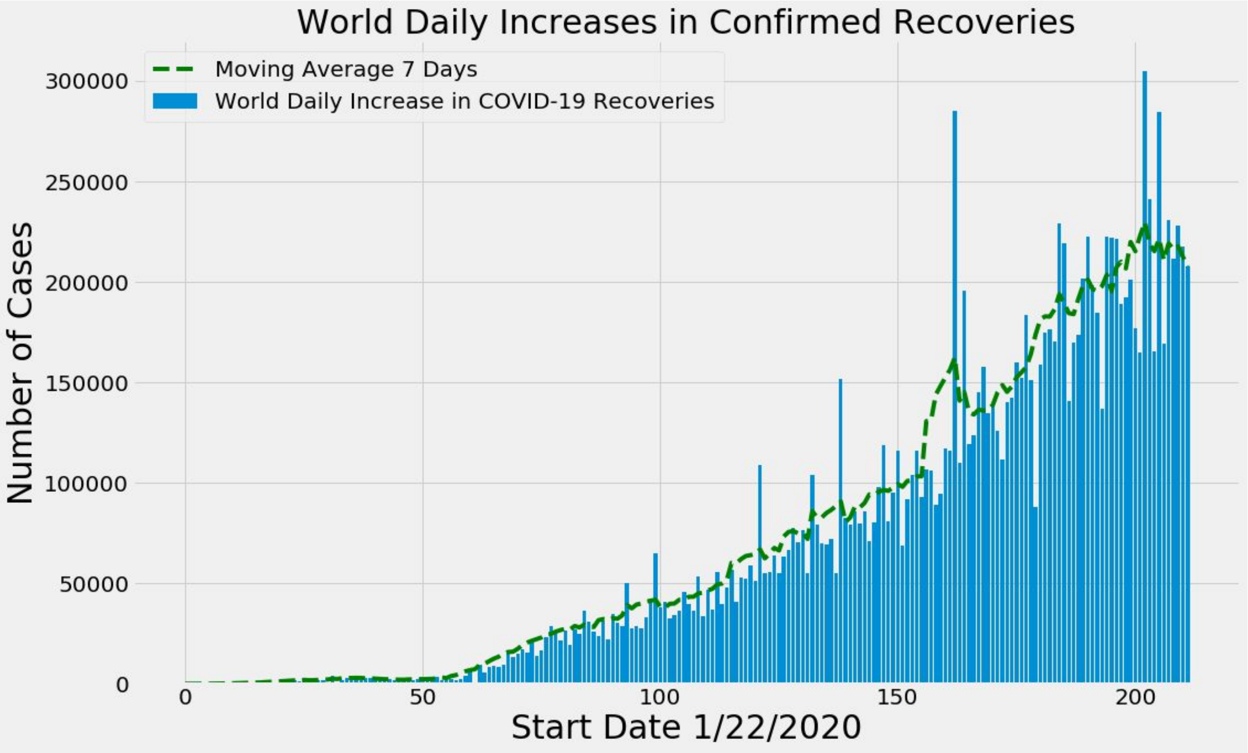
```
plt.figure(figsize=(15, 12))
plt.bar(updated_date, world_daily_increase)
plt.plot(updated_date, world_daily_increase_avg, color='OrangeRed',
linestyle='dashed')
plt.title('World Daily Increases in Confirmed Cases', size=30)
plt.xlabel('Start Date 1/22/2020', size=30)
plt.ylabel('Number of Cases', size=30)
plt.legend(['Moving Average {} Days'.format(window), 'World Daily
Increase in COVID-19 Cases'], prop={'size': 20})
plt.xticks(size=20)
plt.yticks(size=20)
plt.show()
```



World Daily Increases in Confirmed Recoveries Bar chart

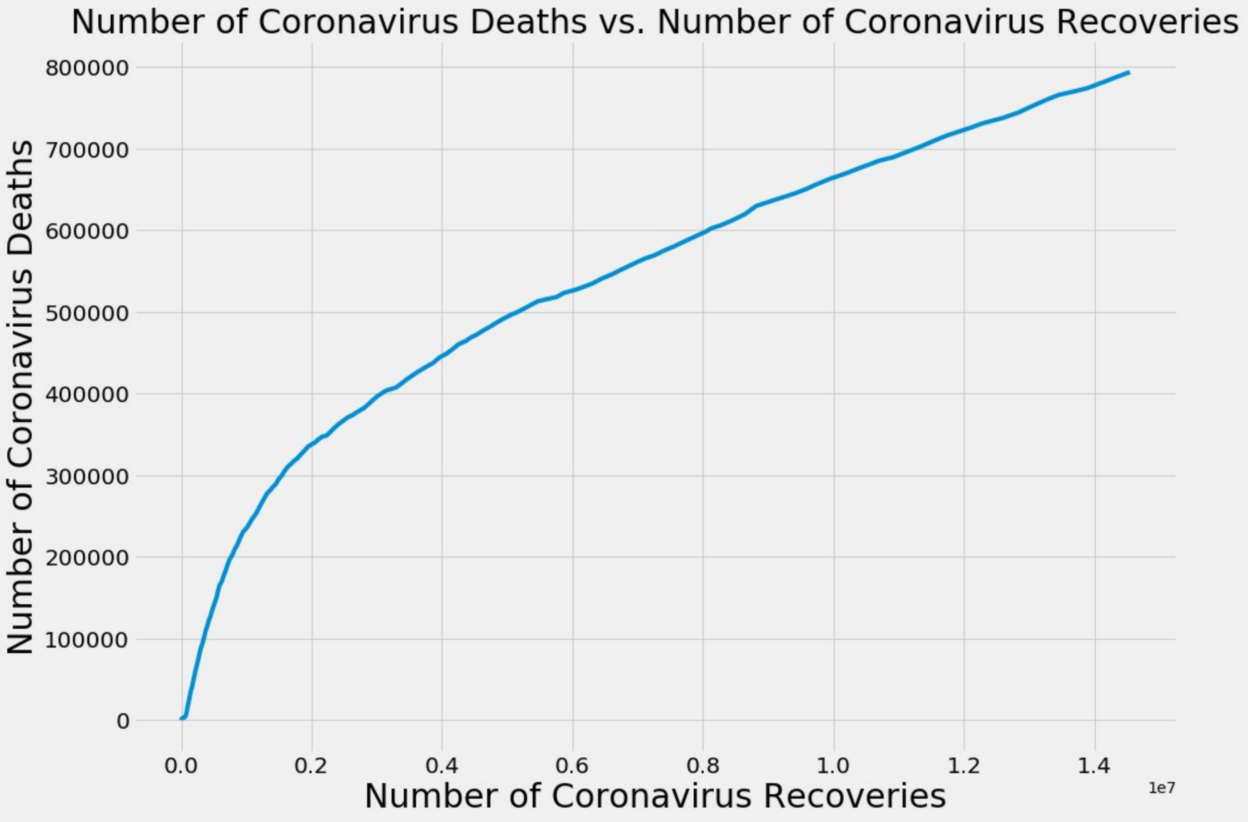
```
plt.figure(figsize=(16, 10))
plt.bar(updated_date, world_daily_recovery)
plt.plot(updated_date, world_daily_recovery_avg, color='Green',
linestyle='dashed')
plt.title('World Daily Increases in Confirmed Recoveries', size=30)
plt.xlabel('Start Date 1/22/2020', size=30)
plt.ylabel('Number of Cases', size=30)
plt.legend(['Moving Average {} Days'.format(window), 'World Daily
Increase in COVID-19 Recoveries'], prop={'size': 20})
```

```
plt.xticks(size=20)
plt.yticks(size=20)
plt.show()
```



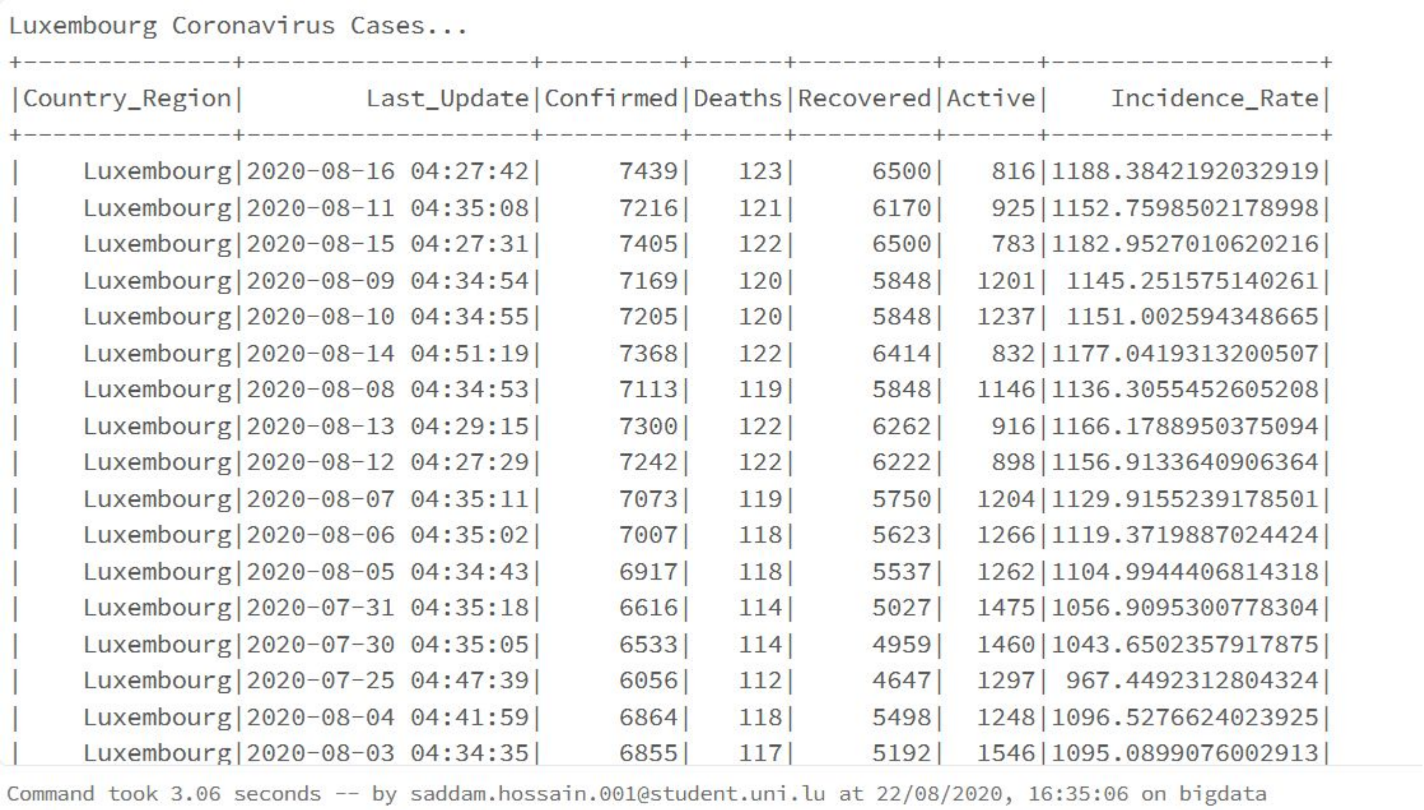
Number of Coronavirus Deaths vs Number of Coronavirus Recoveries

```
plt.figure(figsize=(15, 11))
plt.plot(total_recovered, total_world_deaths)
plt.title('Number of Coronavirus Deaths vs. Number of Coronavirus
Recoveries', size=30)
plt.xlabel('Number of Coronavirus Recoveries', size=30)
plt.ylabel('Number of Coronavirus Deaths', size=30)
plt.xticks(size=20)
plt.yticks(size=20)
plt.show()
```



Overview of Luxembourg Coronavirus Cases

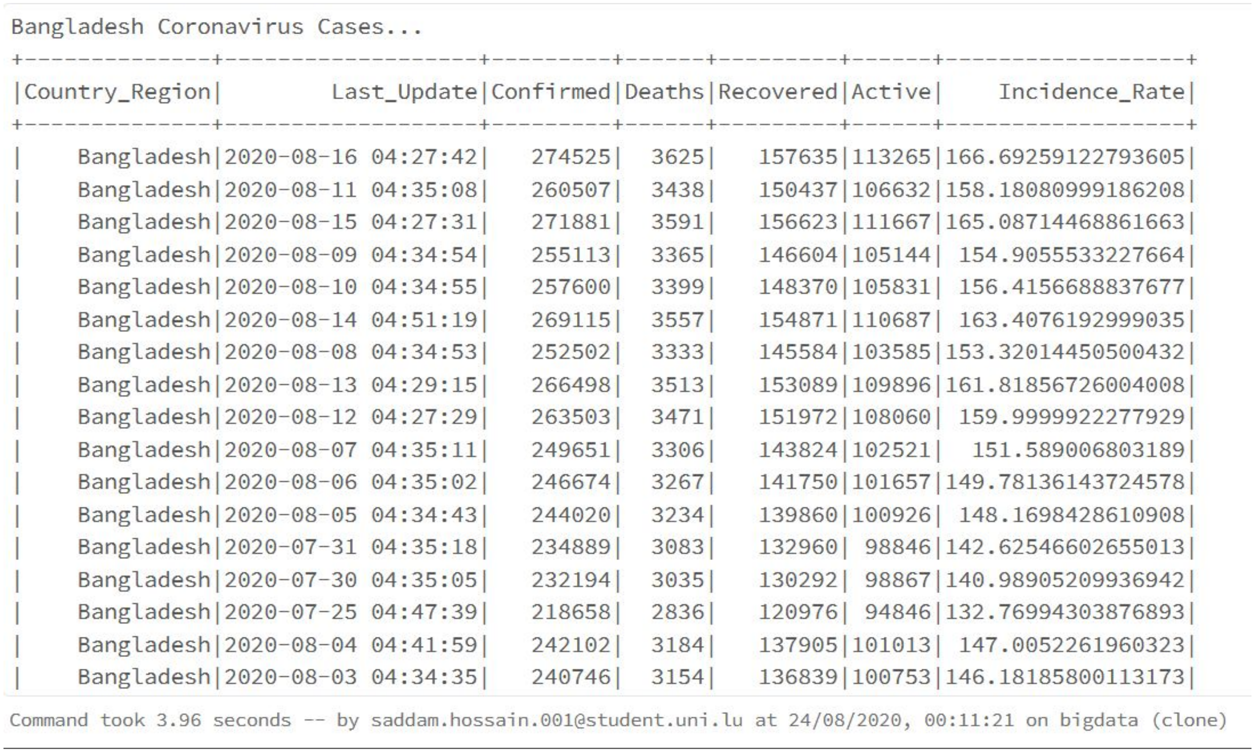
```
("Luxembourg Coronavirus Cases...")
query = """
SELECT
    Country_Region, Last_Update, Confirmed, Deaths, Recovered,
Active, Incidence_Rate

FROM
    covid

WHERE Country_Region in ('Luxembourg')
"""
spark.sql(query).show()
```



```
Luxembourg Coronavirus Cases...
+--------------+-------------------+---------+------+---------+------+------------------+
|Country_Region|        Last_Update|Confirmed|Deaths|Recovered|Active|    Incidence_Rate|
+--------------+-------------------+---------+------+---------+------+------------------+
|    Luxembourg|2020-08-16 04:27:42|     7439|   123|     6500|   816|1188.3842192032919|
|    Luxembourg|2020-08-11 04:35:08|     7216|   121|     6170|   925|1152.7598502178998|
|    Luxembourg|2020-08-15 04:27:31|     7405|   122|     6500|   783|1182.9527010620216|
|    Luxembourg|2020-08-09 04:34:54|     7169|   120|     5848|  1201| 1145.251575140261|
|    Luxembourg|2020-08-10 04:34:55|     7205|   120|     5848|  1237| 1151.002594348665|
|    Luxembourg|2020-08-14 04:51:19|     7368|   122|     6414|   832|1177.0419313200507|
|    Luxembourg|2020-08-08 04:34:53|     7113|   119|     5848|  1146|1136.3055452605208|
|    Luxembourg|2020-08-13 04:29:15|     7300|   122|     6262|   916|1166.1788950375094|
|    Luxembourg|2020-08-12 04:27:29|     7242|   122|     6222|   898|1156.9133640906364|
|    Luxembourg|2020-08-07 04:35:11|     7073|   119|     5750|  1204|1129.9155239178501|
|    Luxembourg|2020-08-06 04:35:02|     7007|   118|     5623|  1266|1119.3719887024424|
|    Luxembourg|2020-08-05 04:34:43|     6917|   118|     5537|  1262|1104.9944406814318|
|    Luxembourg|2020-07-31 04:35:18|     6616|   114|     5027|  1475|1056.9095300778304|
|    Luxembourg|2020-07-30 04:35:05|     6533|   114|     4959|  1460|1043.6502357917875|
|    Luxembourg|2020-07-25 04:47:39|     6056|   112|     4647|  1297| 967.4492312804324|
|    Luxembourg|2020-08-04 04:41:59|     6864|   118|     5498|  1248|1096.5276624023925|
|    Luxembourg|2020-08-03 04:34:35|     6855|   117|     5192|  1546|1095.0899076002913|
Command took 3.06 seconds -- by saddam.hossain.001@student.uni.lu at 22/08/2020, 16:35:06 on bigdata
```

Bangladesh Coronavirus Cases Overview

```
print("Bangladesh Coronavirus Cases...")
query = """
SELECT
    Country_Region, Last_Update, Confirmed, Deaths, Recovered,
Active, Incidence_Rate

FROM
    covid

WHERE Country_Region in ('Bangladesh')
"""
spark.sql(query).show()
```



```
Bangladesh Coronavirus Cases...
+--------------+-------------------+---------+------+---------+------+------------------+
|Country_Region|        Last_Update|Confirmed|Deaths|Recovered|Active|     Incidence_Rate|
+--------------+-------------------+---------+------+---------+------+------------------+
|    Bangladesh|2020-08-16 04:27:42|   274525|  3625|   157635|113265|166.69259122793605|
|    Bangladesh|2020-08-11 04:35:08|   260507|  3438|   150437|106632|158.18080999186208|
|    Bangladesh|2020-08-15 04:27:31|   271881|  3591|   156623|111667|165.08714468861663|
|    Bangladesh|2020-08-09 04:34:54|   255113|  3365|   146604|105144| 154.9055533227664|
|    Bangladesh|2020-08-10 04:34:55|   257600|  3399|   148370|105831| 156.4156688837677|
|    Bangladesh|2020-08-14 04:51:19|   269115|  3557|   154871|110687| 163.4076192999035|
|    Bangladesh|2020-08-08 04:34:53|   252502|  3333|   145584|103585|153.32014450500432|
|    Bangladesh|2020-08-13 04:29:15|   266498|  3513|   153089|109896|161.81856726004008|
|    Bangladesh|2020-08-12 04:27:29|   263503|  3471|   151972|108060| 159.9999922277929|
|    Bangladesh|2020-08-07 04:35:11|   249651|  3306|   143824|102521|   151.589006803189|
|    Bangladesh|2020-08-06 04:35:02|   246674|  3267|   141750|101657|149.78136143724578|
|    Bangladesh|2020-08-05 04:34:43|   244020|  3234|   139860|100926| 148.1698428610908|
|    Bangladesh|2020-07-31 04:35:18|   234889|  3083|   132960| 98846|142.62546602655013|
|    Bangladesh|2020-07-30 04:35:05|   232194|  3035|   130292| 98867|140.98905209936942|
|    Bangladesh|2020-07-25 04:47:39|   218658|  2836|   120976| 94846|132.76994303876893|
|    Bangladesh|2020-08-04 04:41:59|   242102|  3184|   137905|101013| 147.0052261960323|
|    Bangladesh|2020-08-03 04:34:35|   240746|  3154|   136839|100753|146.18185800113173|
```

Command took 3.96 seconds -- by saddam.hossain.001@student.uni.lu at 24/08/2020, 00:11:21 on bigdata (clone)

References

Data Sources:

```
https://github.com/CSSEGISandData/COVID-19

https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv
https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv

https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv
```

```
https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/08-19-2020.csv

https://covid19-static.cdn-apple.com/covid19-mobility-data/2014HotfixDev19/v3/en-us/applemobilitytrends-2020-08-18.csv
```

Cluster

```
https://community.cloud.databricks.com
```