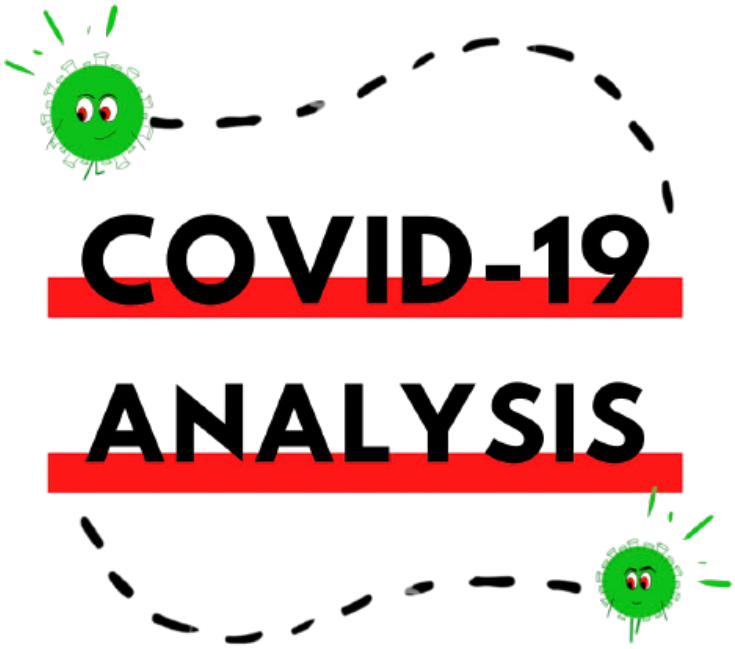
Srinidhi Shukla
Data Analysis and Knowledge discovery
Section 203
Midterm Assessment

COVID-19 ANALYSIS



Srinidhi Shukla
Instructor- Sahar Behpour
Data Analysis and Knowledge Discovery, INFO 5810, Section Number (203)
Midterm Assessment
Date 19/08/2020

1. Introduction

For almost a year now, the Novel Coronavirus has taken hold of the planet. In so many ways, it has devastated the world that it will take a long time to recover from the distress it has caused, surpassing the deaths caused by any other flu, though there are a significant number of recovered cases. Whether it is the world's economy or the global population's health. Let us see the impact of the disease during the months of April to September on the world and the USA.

1.1. Data

The dataset, 'dataset2.csv' used for this assignment, is a collection of statistical data related to the impact caused by the disease Coronavirus. It has been extracted from the World Meter. It consists of information across various countries across the world such as Total cases, Total deaths, Total recovered etc.,

- Why this dataset?
  This dataset consists of reliable and precise data with consistent records of COVID-19 related numbers. It is sorted by date, country and continent, making it simple to perform the analysis.

- Where is the data collected from?
  The dataset is generated from : https://www.worldometers.info/coronavirus . To crawl the website, a script was written and BeautifulSoup was used to extract the interest element.

- Why was the data collected?
  The data was collected to keep track of and perform analysis based on the number of people infected, died and recovered by COVID-19.

- Who collected the data?
  The data for research on the COVID-19 records was obtained by a data source hub-Kaggle user named Tanvish Aggarwal. The main objective of obtaining this information was to know whether or not the lockdown implementation worked and, if so, which lockdown technique works best? The data has also been used for the potential prediction of cases for the country in order to take precautionary steps for the country.

- What are the attributes used in this dataset?
  The following is the list of attributes used in this dataset-
  1. Country
  2. Total Cases
  3. New Cases
  4. Total Deaths
  5. New Deaths
  6. Total Recovered
  7. Active Cases
  8. Serious/Critical
  9. Total Cases/1M pop
  10. Deaths /1M pop
  11. Total Tests

12. Tests /1M pop
13. Continent
14. Date

- What type of data is that?
    1. Country - Qualitative (Specific country name)
    2. Total Cases - Quantitative (Measured)
    3. New Cases - Quantitative (Measured)
    4. Total Deaths - Quantitative (Measured)
    5. New Deaths - Quantitative (Measured)
    6. Total Recovered - Quantitative (Measured)
    7. Active Cases - Quantitative (Measured)
    8. Serious/Critical - Quantitative (Measured)
    9. Total Cases/1M pop- Quantitative (Measured)
    10. Deaths /1M pop- Quantitative (Measured)
    11. Total Tests- Quantitative (Measured)
    12. Tests /1M pop- Quantitative (Measured)
    13. Continent- Qualitative (Specific continent name)
    14. Date - Quantitative (Measured)

- What is the data about?
  The data is about the impact Coronavirus has on the human lives. It tells about the total number of people who got infected, died, recovered, serious / critical from April to September and also the daily details about new infected, death and recovered cases between that time.

- What is the time interval for the collected data?
  This dataset contains the data from 04/11/2020 to 09/30/2020.

- Basic statistics about the data set.
  There is a lot of scope for performing statistical functions on this data such as associated death rates, average deaths, tests, cures etc.,

- Is the data unstructured? If so, why? What are the specific things that make it to be called an unstructured dataset?
  This dataset is well structured with specified rows and columns and threshold values of the categories specified.

- Are the attributes normalized?
  The attributes are normalized, for example the attribute Date is specified in the date range for months April to September.

1.2.Background Research and Related Publications

- Further study on Early Chinese statistics which indicate that the majority of deaths from coronavirus disease 2019 (COVID-19) occurred among adults aged 60 years or older and people with significant underlying health conditions has been done by many researchers for CDC in the

following paper “Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) — United States, February 12–March 16, 2020”

- Another such insightful publication is “Covid-19 — Navigating the Uncharted” by Anthony S.Fauci, M.D., H. Clifford Lane, M.D., and Robert R. Redfield, M.D.

1.3.Objectives

The objectives set for this assignment are listed below. We perform analysis to determine the following

- To extract the name of the month from given date column
- Total cases in USA
- To find New cases in USA
- To find Total deaths in USA
- To find Total recovered in USA
- To find Active cases in USA
- To find Total Cases/1M pop in the World
- To find Deaths /1M pop in the World
- To find Total Cases/1M pop in USA by 9/30/20
- To find Average New cases per day in USA in September
- To find the country name with maximum number of Tests /1M pop and the date
- To perform data validation
- To plot a chart to visualize the Month wise Total cases in USA (using slicers and filters)
- To plot a chart to visualize the Deaths per day
- To plot a chart to visualize the New cases per day
- To plot a chart to visualize the Criticality per day
- To plot a chart to visualize the Total Cases VS Total Deaths in USA
- To plot a chart to visualize the continent-wise Cases
- To plot a chart to visualize the continent-wise Death rate
- To plot a chart to visualize the continent-wise Recovery rate
- To plot a chart to visualize the Worldwide COVID-19 statistics
- To plot a chart to visualize the Worldwide spread of COVID-19

2. Methods

2.1.Tools

- The tool used in this assignment for the analysis is Microsoft Excel.

2.2. Data pre-processing

- The dataset I am working on in this assignment does not provide any scope for omitting missing values or duplicates because, if done so, there might be a loss of significant data. The rest of the data is well sorted and processed for the analysis.
- But the list of countries contains some names of the Cruise ships(such as below “Diamond Princess” is a name of a Cruise ship) which were irrelevant, so they have been removed.

Diamond Princess	19224	17577	351

2.3. Data Analysis

The analysis methods used in this assignment are

- Index Functions
- Match Functions
- Filtering Data
- Date- time functions
- Pivot Tables
- Pivot Charts
- Array Formulas
- Text Functions
- Lookup Functions
- Visualizing the results (Pareto, Line, Pie, Column charts etc.,)

3. Results

- To extract the name of the month from given date column

Function:

```
=TEXT(N2:N34933, "mmmm")
```

Result:

Date	Month
4/11/20	April

- Total cases in USA

Function:

```
=INDEX(B2:B34933, MATCH(A3&N34721, A2:A34933&N2:N34933,0))
```

Result:

USA Total cases	7447282

- To find New cases in USA

Function:

```
=INDEX(C2:C34933, MATCH(A3&N34721, A2:A34933&N2:N34933,0))
```

Result:

USA New cases	40929

- To find Total deaths in USA

Function:

```
=INDEX(D2:D34933, MATCH(A3&N34721, A2:A34933&N2:N34933,0))
```

Result:

USA Total deaths	211740

- To find Total recovered in USA

  Function:

  ```
  =INDEX(F2:F34933, MATCH(A3&N34721, A2:A34933&N2:N34933,0))
  ```

  Result:

USA Total recovered	4699706

- To find Active cases in USA

  Function:

  ```
  =INDEX(G2:G34933, MATCH(A3&N34721, A2:A34933&N2:N34933,0))
  ```

  Result:

USA Active cases	2535836

- To find Total Cases/1M pop in the World

  Function:

  ```
  =INDEX(I2:I34933, MATCH(A2&N34721, A2:A34933&N2:N34933,0))
  ```

  Result:

World Total Cases/1M pop	4381

- To find Deaths /1M pop in the World

  Function:

  ```
  =INDEX(J2:J34933, MATCH(A2&N34721, A2:A34933&N2:N34933,0))
  ```

  Result:

World Deaths /1M pop	130.6

- To find Total Cases/1M pop in USA by 9/30/20

  Function:

  ```
  =INDEX($B$2:$B$34933, SMALL(IF(COUNTIF(D34721, D2:D34933)*COUNTIF(A3, A2:A34933), ROW($A$2:$D$34933)-MIN(ROW($A$2:$D$34933))+1), ROW(A1)), COLUMN(A1))
  ```

  Result:

Total Cases/1M pop in USA by 9/30/20	22466

- To find Average New cases per day in USA in September

  Function:

  ```
  =ROUNDUP(AVERAGEIFS(C2:C34933,A2:A34933,"USA",E2:E34933,"September"),1)
  ```

  Result:

Average New cases per day in USA in September	24575.3

- To find the country name with maximum number of Tests /1M pop and the date

Function:

```
=VLOOKUP(MAX(C2:C34933),C2:D34933,2,0)
```

```
=TEXT(VLOOKUP(MAX(C2:C34933),C2:E34933,3,0), "MM/DD/YY")
```

Result:

Country with maximum number of Tests /1M pop	Faeroe Islands
By date	09/30/20

- To perform data validation

Function:

```
=ISNUMBER(B2:B34933)
```

Result:

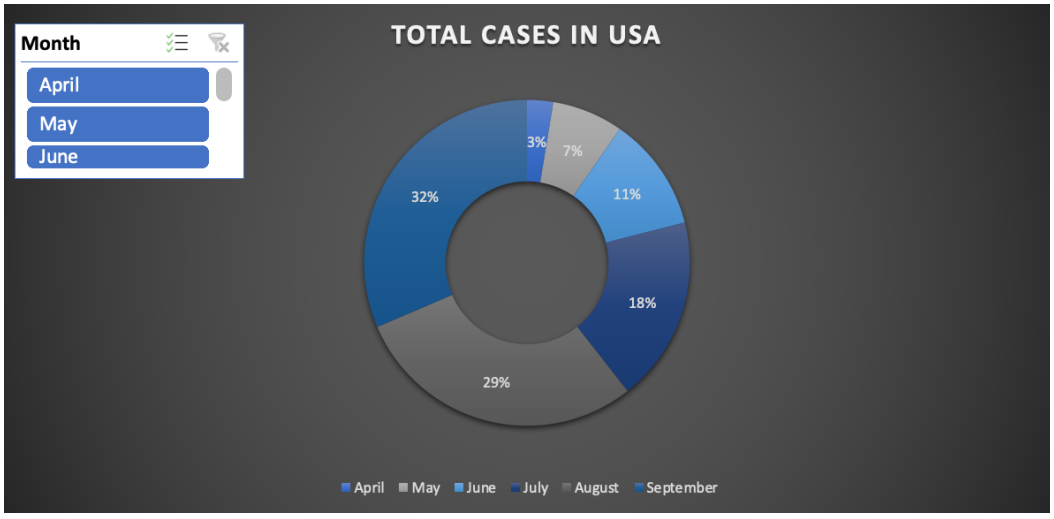
Country	Total Cases	Result
World	1741818	TRUE
USA	508575	TRUE
Spain	161852	TRUE

- To plot a chart to visualize the Month wise Total cases in USA (using slicers and filters)

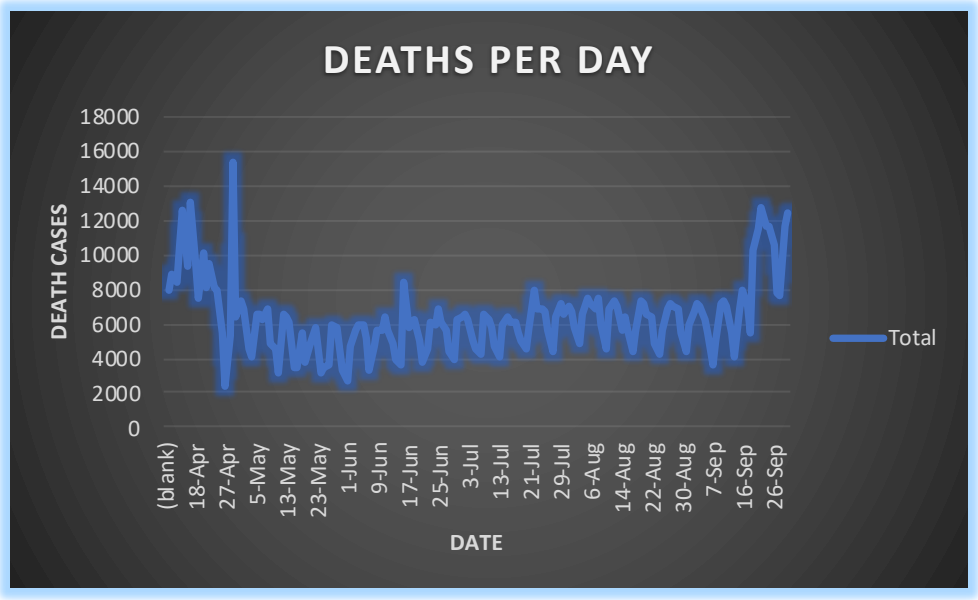
Function: Pie chart, Filters, Slicers, Pivot tables, Pivot charts

Result:

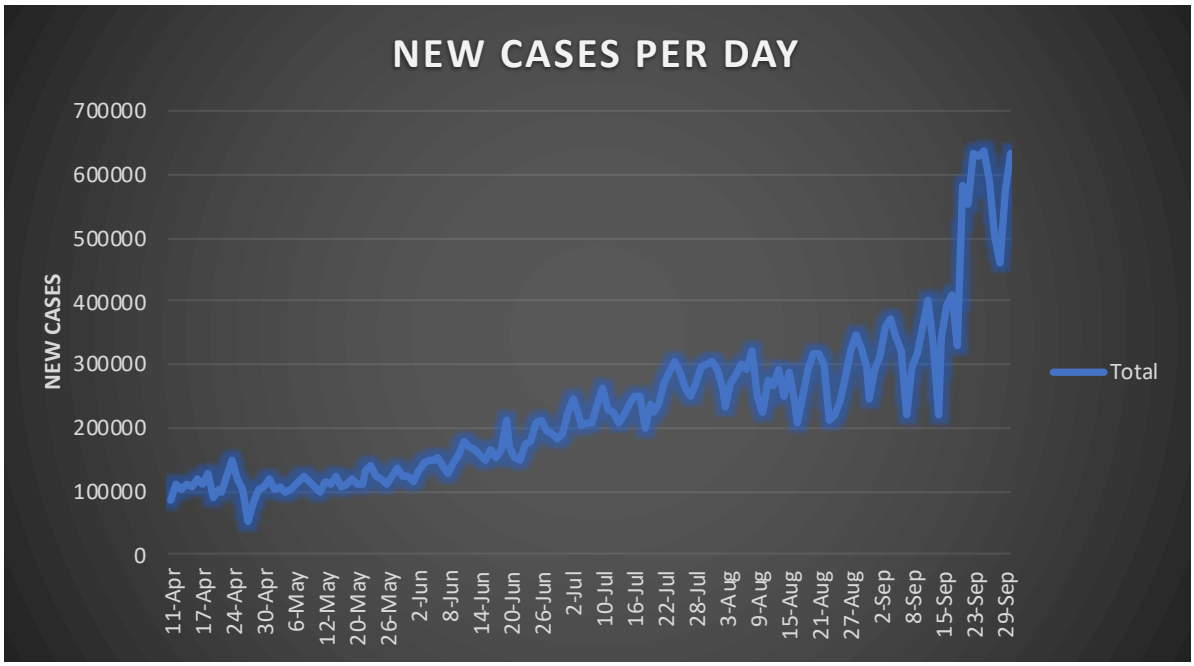
Country	USA
Sum of Total Cases	
Month	Total
April	15173640
May	41258447
June	66508178
July	107552312
August	170528508
September	183842444



- To plot a chart to visualize the Deaths per day
  Function: Line chart, Pivot tables, Pivot charts
  Result:



- To plot a chart to visualize the New cases per day
  Function: Line chart, Pivot tables, Pivot charts
  Result:

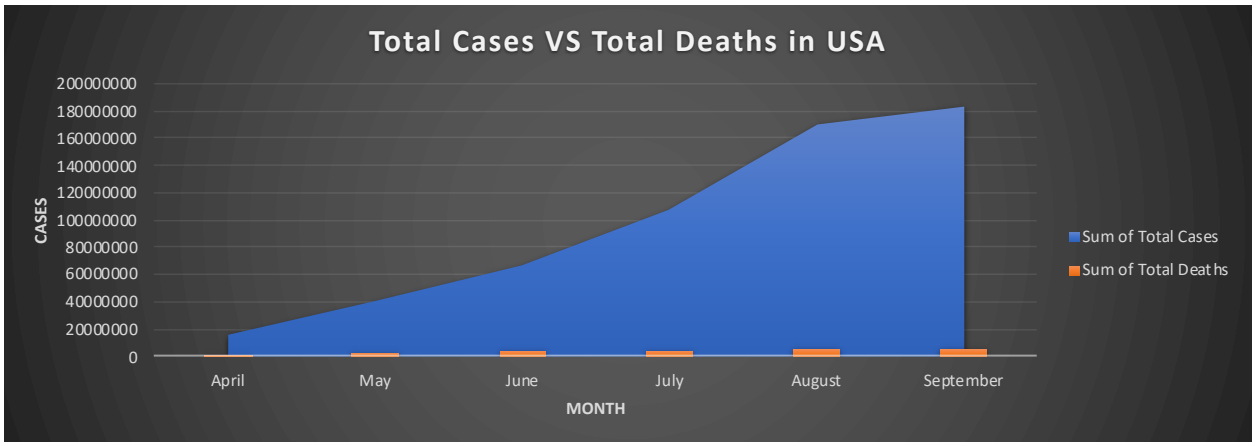


- To plot a chart to visualize the Criticality per day
  Function: Line chart, Pivot tables, Pivot charts
  Result:

- To plot a chart to visualize the Total Cases VS Total Deaths in USA
  Function: Combination chart, Pivot tables, Pivot charts
  Data:

Month	Sum of Total Cases	Sum of Total Deaths
April	15173640	800287
May	41258447	2430643
June	66508178	3549198
July	107552312	4111556
August	170528508	5348028
September	183842444	5402848

Result:



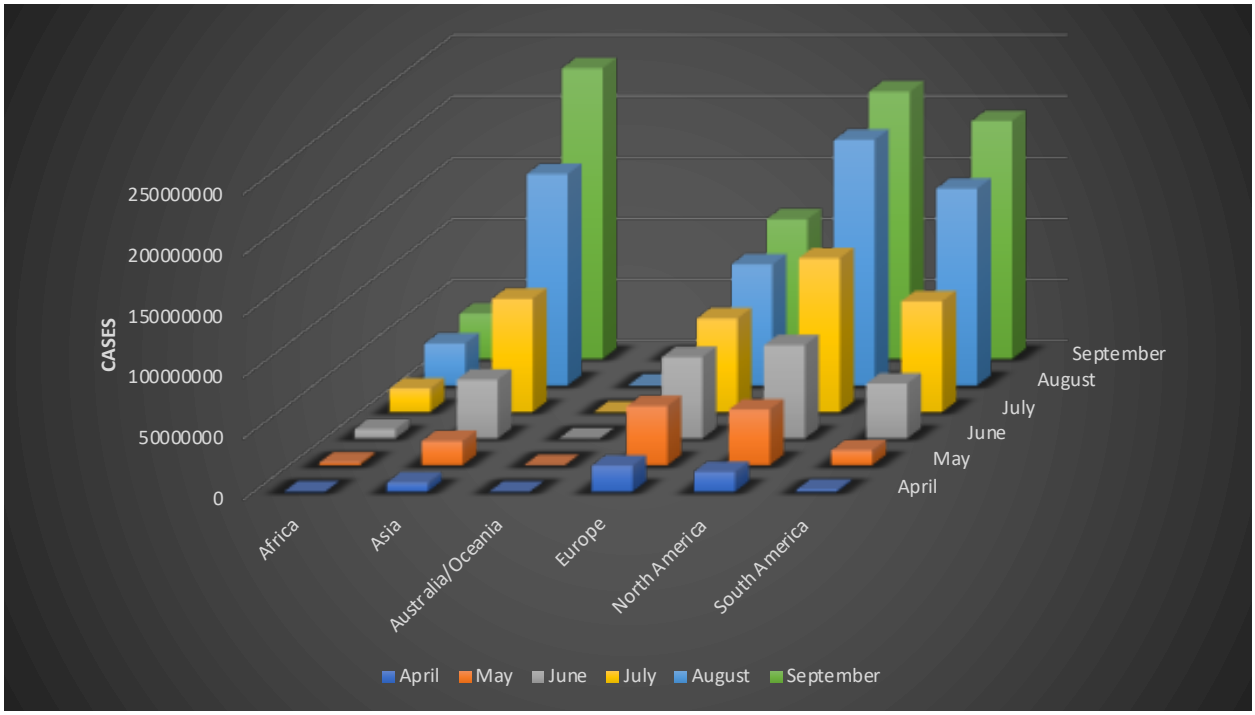
- To plot a chart to visualize the continent-wise Cases

  Function: Column chart, Pivot tables, Pivot charts

  Data:

Continent	April	May	June	July	August	September
Africa	475808	2346586	7774585	19329924	34375626	37070074
Asia	7560341	19773701	48256934	92698897	173118626	237898596
Australia/Oceania	153871	240846	270502	383761	770140	812471
Europe	21282146	48355443	66791293	76672800	99141776	114336389
North America	16323824	45549773	76506030	125949825	201087651	218813610
South America	1819799	12231397	45090603	90726568	161225491	194738087

  Result:



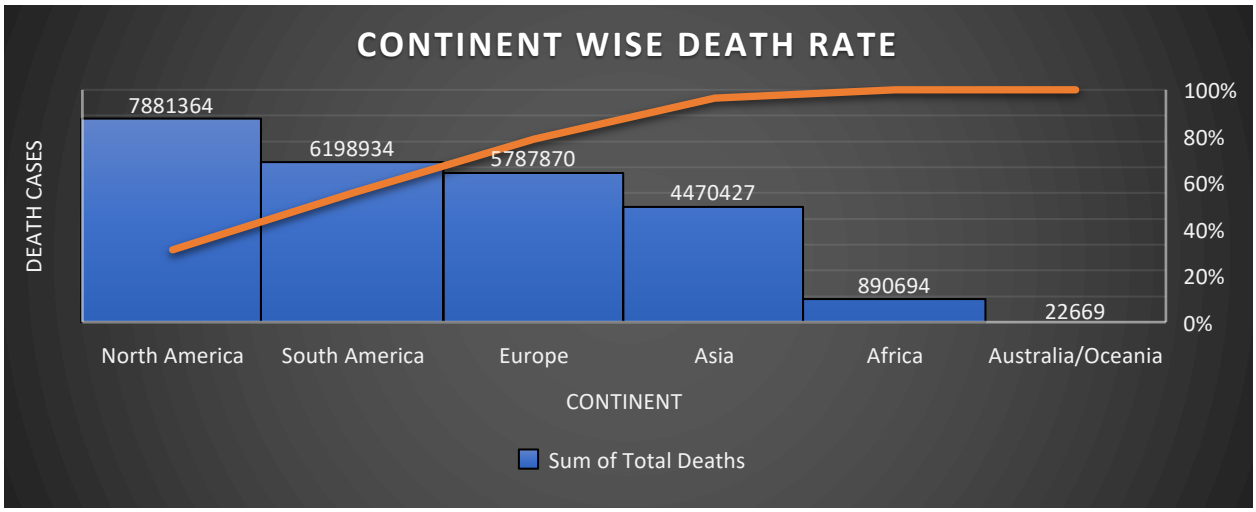
- To plot a chart to visualize the continent-wise Death rate

  Function: Pareto chart, Pivot tables

  Data:

Continent	Sum of Total Deaths
Africa	890694
Asia	4470427
Australia/Oceania	22669
Europe	5787870
North America	7881364
South America	6198934

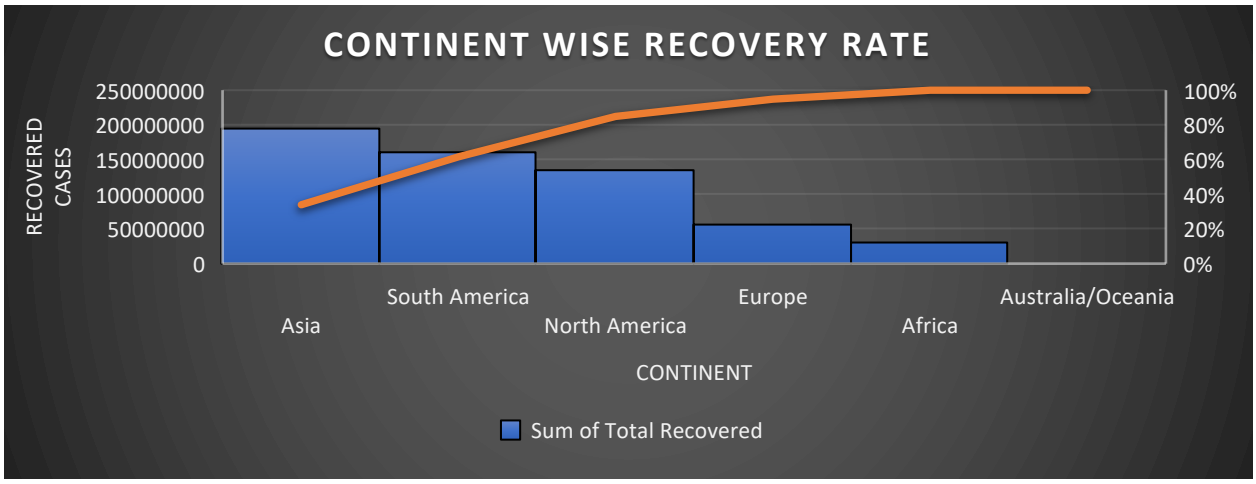
  Result:



- To plot a chart to visualize the continent-wise Recovery rate
  Function: Pareto chart, Pivot tables
  Data:

Continent	Sum of Total Recovered
Africa	30090443
Asia	194274852
Australia/Oceania	709811
Europe	56139731
North America	134320043
South America	160468883

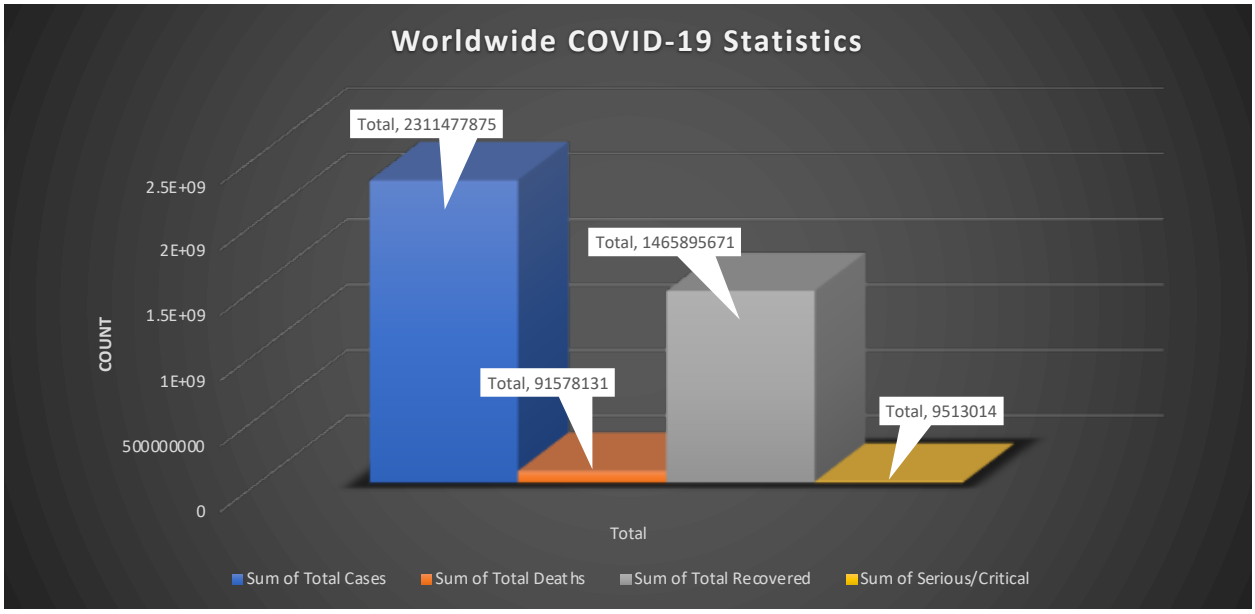
Result:



- To plot a chart to visualize the Worldwide COVID-19 statistics
  Function: Column chart, Pivot tables, Pivot charts
  Data:

	Sum of Total Cases	Sum of Total Deaths	Sum of Total Recovered	Sum of Serious/Critical
Total	2311477875	91578131	1465895671	9513014

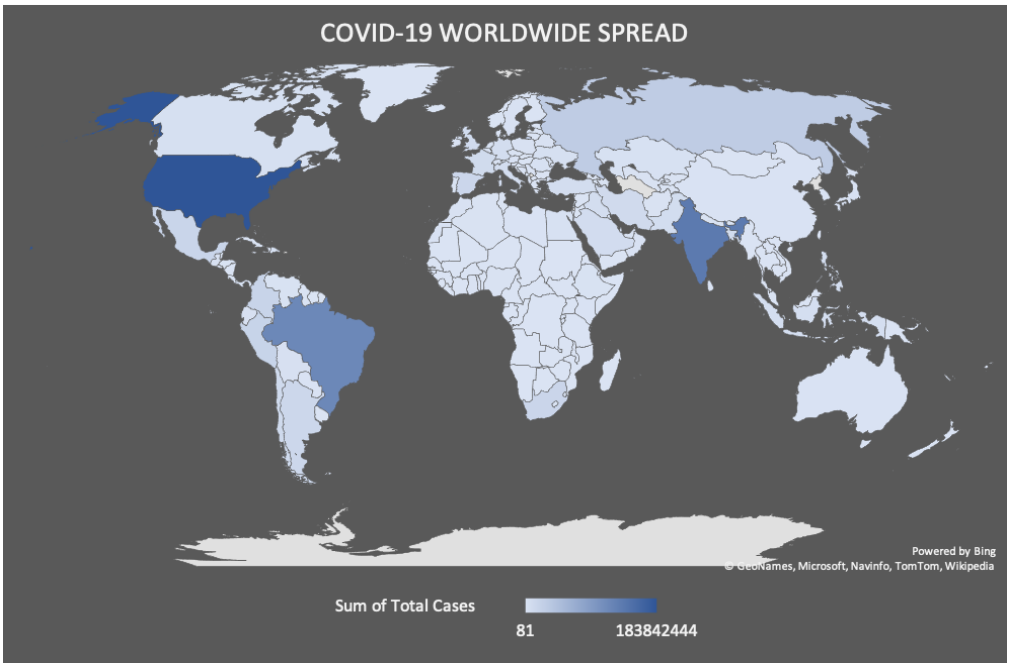
Result:



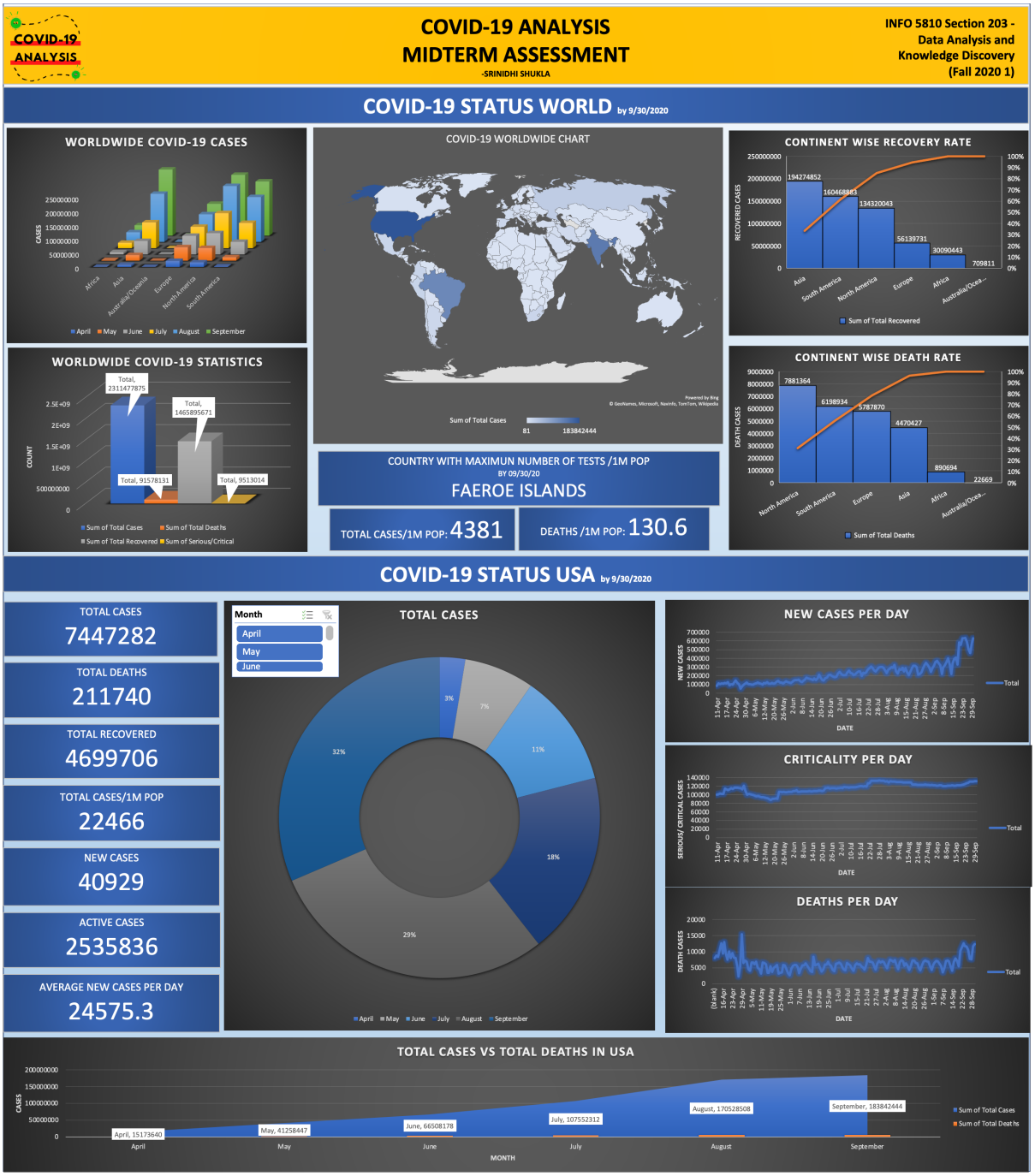
- To plot a chart to visualize the Worldwide spread of COVID-19
  Function: Map chart, Pivot tables
  Data:

Row Labels	Sum of Total Cases	Sum of Total Recove	Sum of Total Deaths
Afghanistan	1047092	852718	38634
Albania	313715	179897	9221
Algeria	1309498	922184	43985

Result:



- Dashboard:



4. Discussion

- The date and time functions are very efficient and easy to use to deal with data containing the dates and times. Some of the functions which were very helpful for me are- YEAR, MONTH etc.,
- The pivot tables and charts are excellent for visualization of the analyzed data.
- The Filter option makes it easy to filter out the unnecessary data while performing the analysis.
- For complex analysis, Array formulas work the best. We can perform multiple calculations in a single formula.
- The 'Data Analysis' button provided in the Analysis ToolPak makes the statistical calculation quick and simple.

5. Evaluation and Conclusion

- This assignment allowed me to come up with new things with the help of learnings in the previous assignments.
- I managed to build a dashboard using the limited data that is in the dataset.
- Microsoft Excel has many interesting features and is capable of performing amazing analysis and creating beautiful visualizations.
- I always wondered how I could infuse my artistic skills into Data science. I created a logo using my hand drawn illustrations for this assignment.

References

Array Formulas in Excel. Excel-easy.com. (2020). Retrieved 16 October 2020, from https://www.excel-easy.com/functions/array-formulas.html.

Bialek, S., Boundy, E., Bowen, V., Chow, N., Cohn, A., & Dowling, N. et al. (2020). Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) — United States, February 12–March 16, 2020. *MMWR. Morbidity And Mortality Weekly Report*, *69*(12), 343-346. https://doi.org/10.15585/mmwr.mm6912e2

Coronavirus Update (Live): 40,612,487 Cases and 1,122,254 Deaths from COVID-19 Virus Pandemic - Worldometer. Worldometers.info. (2020). Retrieved 17 October 2020, from https://www.worldometers.info/coronavirus/.

COVID-19 Dataset. Kaggle.com. (2020). Retrieved 15 October 2020, from https://www.kaggle.com/tavishaggarwal/covid-19-dataset-on-country-level.

Fauci, A., Lane, H., & Redfield, R. (2020). Covid-19 — Navigating the Uncharted. *New England Journal Of Medicine*, *382*(13), 1268-1269. https://doi.org/10.1056/nejme2002387

How to use the Excel AVERAGEIFS function | Exceljet. Exceljet.net. (2020). Retrieved 17 October 2020, from https://exceljet.net/excel-functions/excel-averageifs-function.

INDEX and MATCH in Excel. Excel-easy.com. (2020). Retrieved 17 October 2020, from https://www.excel-easy.com/examples/index-match.html.