INTRODUCTION TO DATA MANAGEMENT PROJECT REPORT

On

Corona World And Indian States Analysis

Submitted By: -

Sarvjeet Singh

11912936

B.Tech [CSE]

Under The Guidance Of

Komal Arora: 17783

School Of Computer Science & Engineering

Lovely Professional University, Phagwara

(August-December 2021)

DECLARATION

I, Sarvjeet Singh, student of Computer Science & Engineering under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 11/16/2021

Sarvjeet Singh

Registration No: 11912936

Signature: Sarvjeet Singh

ACKNOWLEDGEMENT

Primarily I'd thank god for being able to complete my project with success. Then I'd like to thank my mentor Ms. Komal Arora, whose valuable guidance has been the ones that helped me patch this project and make it full proof success in contribution towards the completion of this project. Last but not least I'd rather thanks to Lovely Professional University, and my parent’s inspiration, who gave me this golden opportunity to learn many new things, to learn another aspects of life.

Sarvjeet Singh

CONTENTS

INTRODUCTION

- Data management is important because the data your organization creates is a very valuable resource.

- The last thing you want to do is spend time and resources collecting data and business intelligence, only to lose or misplace that information.

- In that case, you would then have to spend time and resources again to get that same business intelligence you already had.

- And on that data analysis is carried out which show visualization of our problems in efficient way.

- Data Analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision- making.

- This project is based on such data analysis on Covid data from Jan 2020 to May 2021.

- This dataset has mostly all countries date wise Confirmed, Death and Recovered Cases.

- This dataset contains 8 data fields.

OBJECTIVES/SCOPE OF ANALYSIS

After analysis of the dataset, the aim of this project is to give answer of given objectives in easy way:

1. To display top 5 countries which are affected the most
2. To Show Top 5 country having highest number of death due to Corona
3. To Show Five Least Affected Countries (Confirmed and Death Rate)
4. To Show Top Five Countries with Most Recovered Cases
5. Second Wave in India
6. Confirmed Cases World Map
7. Indian State Wise Confirmed, Recovered and Death Cases
8. Tree Map of Confirmed Cases in States
9. Confirmed Cases State Wise Map
10. Indian Month Wise Confirmed, Recovered and Death Cases

SOURCE OF DATASET

Source of dataset: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

The dataset is based on the Corona Cases From around the Globe Day Wise From Jan 2020-May 2021

The columns included in the dataset are given below:

- SNo
  - Serial No for every row
- ObservationDate
  - Date of recording the Data
- Province/State
  - Name of the State of the Country
- Country
  - Name of the Country
- Last Updated
  - Date of Updating the Data
- Confirmed
  - No of Confirmed Cases on the Given Date
- Deaths
  - No of Death Cases on the Given Date
- Recovered
  - No of Recovered Cases on the Given Date

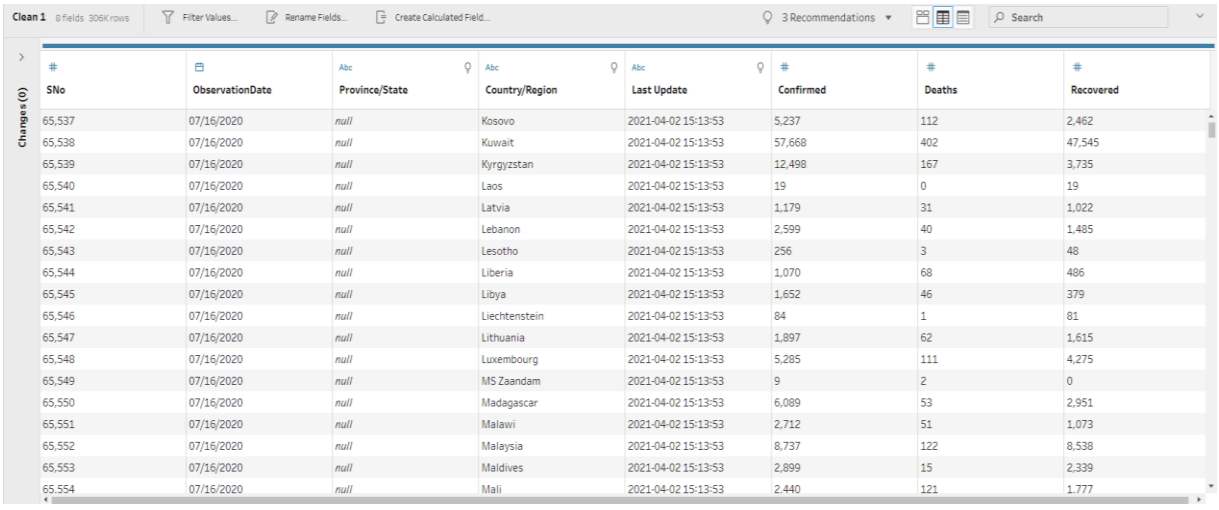
Sample of Dataset with data fields is given below:

	A	B	C	D	E	F	G	H
1	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
2	1	1/22/2020	Anhui	Mainland China	1/22/2020 17:00	1	0	0
3	2	1/22/2020	Beijing	Mainland China	1/22/2020 17:00	14	0	0
4	3	1/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6	0	0
5	4	1/22/2020	Fujian	Mainland China	1/22/2020 17:00	1	0	0
6	5	1/22/2020	Gansu	Mainland China	1/22/2020 17:00	0	0	0
7	6	1/22/2020	Guangdong	Mainland China	1/22/2020 17:00	26	0	0
8	7	1/22/2020	Guangxi	Mainland China	1/22/2020 17:00	2	0	0
9	8	1/22/2020	Guizhou	Mainland China	1/22/2020 17:00	1	0	0
10	9	1/22/2020	Hainan	Mainland China	1/22/2020 17:00	4	0	0
11	10	1/22/2020	Hebei	Mainland China	1/22/2020 17:00	1	0	0
12	11	1/22/2020	Heilongjiang	Mainland China	1/22/2020 17:00	0	0	0
13	12	1/22/2020	Henan	Mainland China	1/22/2020 17:00	5	0	0
14	13	1/22/2020	Hong Kong	Hong Kong	1/22/2020 17:00	0	0	0
15	14	1/22/2020	Hubei	Mainland China	1/22/2020 17:00	444	17	28
16	15	1/22/2020	Hunan	Mainland China	1/22/2020 17:00	4	0	0
17	16	1/22/2020	Inner Mongolia	Mainland China	1/22/2020 17:00	0	0	0
18	17	1/22/2020	Jiangsu	Mainland China	1/22/2020 17:00	1	0	0
19	18	1/22/2020	Jiangxi	Mainland China	1/22/2020 17:00	2	0	0
20	19	1/22/2020	Jilin	Mainland China	1/22/2020 17:00	0	0	0
21	20	1/22/2020	Liaoning	Mainland China	1/22/2020 17:00	2	0	0
22	21	1/22/2020	Macau	Macau	1/22/2020 17:00	1	0	0
23	22	1/22/2020	Ningxia	Mainland China	1/22/2020 17:00	1	0	0
24	23	1/22/2020	Qinghai	Mainland China	1/22/2020 17:00	0	0	0
25	24	1/22/2020	Shaanxi	Mainland China	1/22/2020 17:00	0	0	0
26	25	1/22/2020	Shandong	Mainland China	1/22/2020 17:00	2	0	0
27	26	1/22/2020	Shanghai	Mainland China	1/22/2020 17:00	9	0	0
28	27	1/22/2020	Shanxi	Mainland China	1/22/2020 17:00	1	0	0
29	28	1/22/2020	Sichuan	Mainland China	1/22/2020 17:00	5	0	0

ETL PROCESS

- ETL is a process that extracts the data from different source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system.
- Full form of ETL is Extract, Transform and Load.
- The triple combination of ETL provides crucial functions that are many times combined into a single application or suite of tools that help in the following areas:
    - Enhances Business Intelligence solutions for decision making.
    - Allows verification of data transformation, aggregation and calculations rules.
    - Allows sample data comparison between source and target system.
    - Helps to improve productivity as it codifies and reuses without additional technical skills.

- Initially, the ra dataset was arranged as shown in given picture:



Clean 1 8 fields 306K rows | Filter Values... | Rename Fields... | Create Calculated Field... | 3 Recommendations | Search

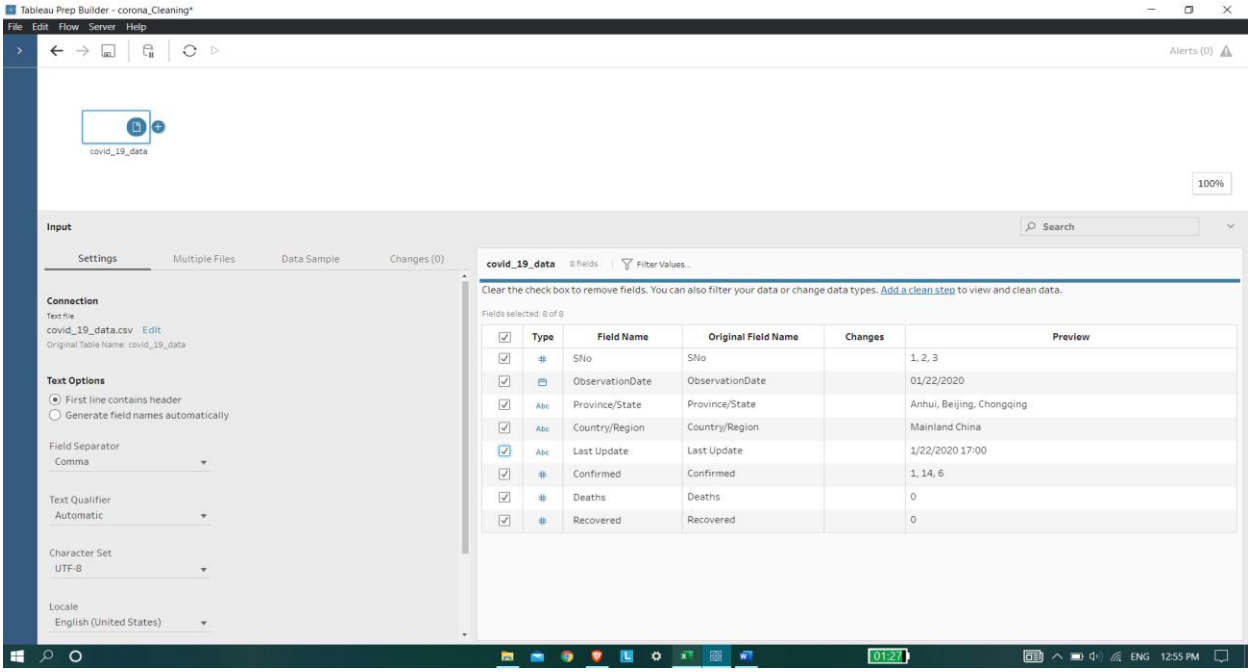
Changes (0)

SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
65,537	07/16/2020	null	Kosovo	2021-04-02 15:13:53	5,237	112	2,462
65,538	07/16/2020	null	Kuwait	2021-04-02 15:13:53	57,668	402	47,545
65,539	07/16/2020	null	Kyrgyzstan	2021-04-02 15:13:53	12,498	167	3,735
65,540	07/16/2020	null	Laos	2021-04-02 15:13:53	19	0	19
65,541	07/16/2020	null	Latvia	2021-04-02 15:13:53	1,179	31	1,022
65,542	07/16/2020	null	Lebanon	2021-04-02 15:13:53	2,599	40	1,485
65,543	07/16/2020	null	Lesotho	2021-04-02 15:13:53	256	3	48
65,544	07/16/2020	null	Liberia	2021-04-02 15:13:53	1,070	68	486
65,545	07/16/2020	null	Libya	2021-04-02 15:13:53	1,652	46	379
65,546	07/16/2020	null	Liechtenstein	2021-04-02 15:13:53	84	1	81
65,547	07/16/2020	null	Lithuania	2021-04-02 15:13:53	1,897	62	1,615
65,548	07/16/2020	null	Luxembourg	2021-04-02 15:13:53	5,285	111	4,275
65,549	07/16/2020	null	MS Zaandam	2021-04-02 15:13:53	9	2	0
65,550	07/16/2020	null	Madagascar	2021-04-02 15:13:53	6,089	53	2,951
65,551	07/16/2020	null	Malawi	2021-04-02 15:13:53	2,712	51	1,073
65,552	07/16/2020	null	Malaysia	2021-04-02 15:13:53	8,737	122	8,538
65,553	07/16/2020	null	Maldives	2021-04-02 15:13:53	2,899	15	2,339
65,554	07/16/2020	null	Mali	2021-04-02 15:13:53	2,440	121	1,777

Steps taken to clean dataset through Tableau Prep:

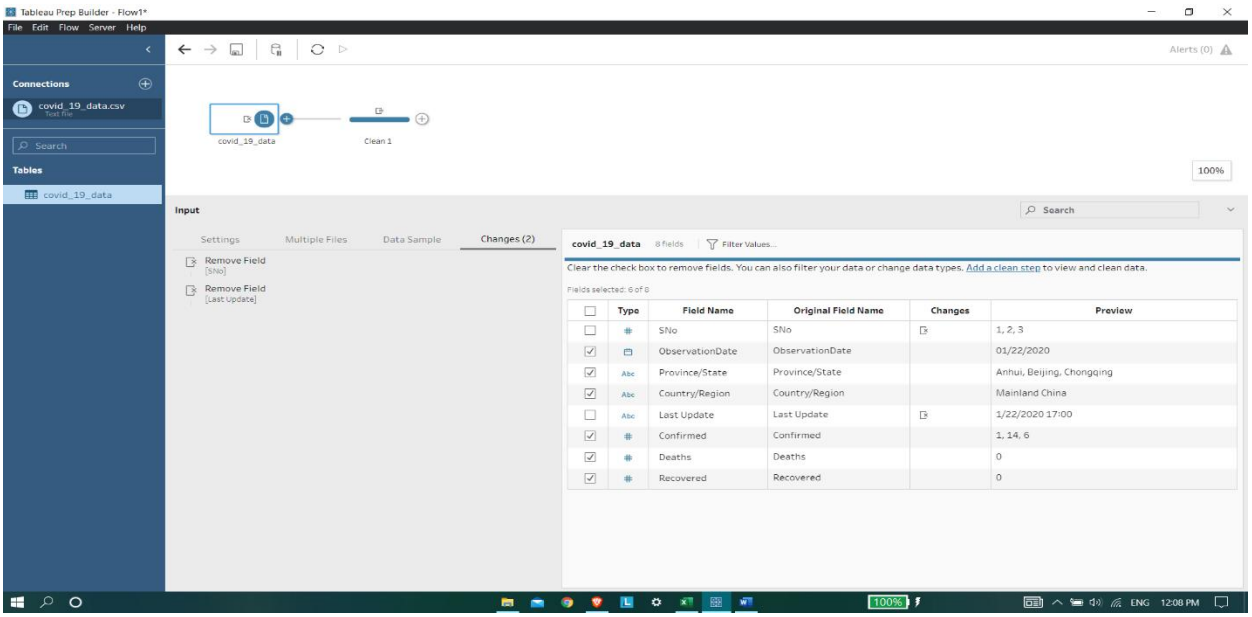
Step 1:

- Firstly, open Tableau Prep and the Click on connect to data then import the csv file.



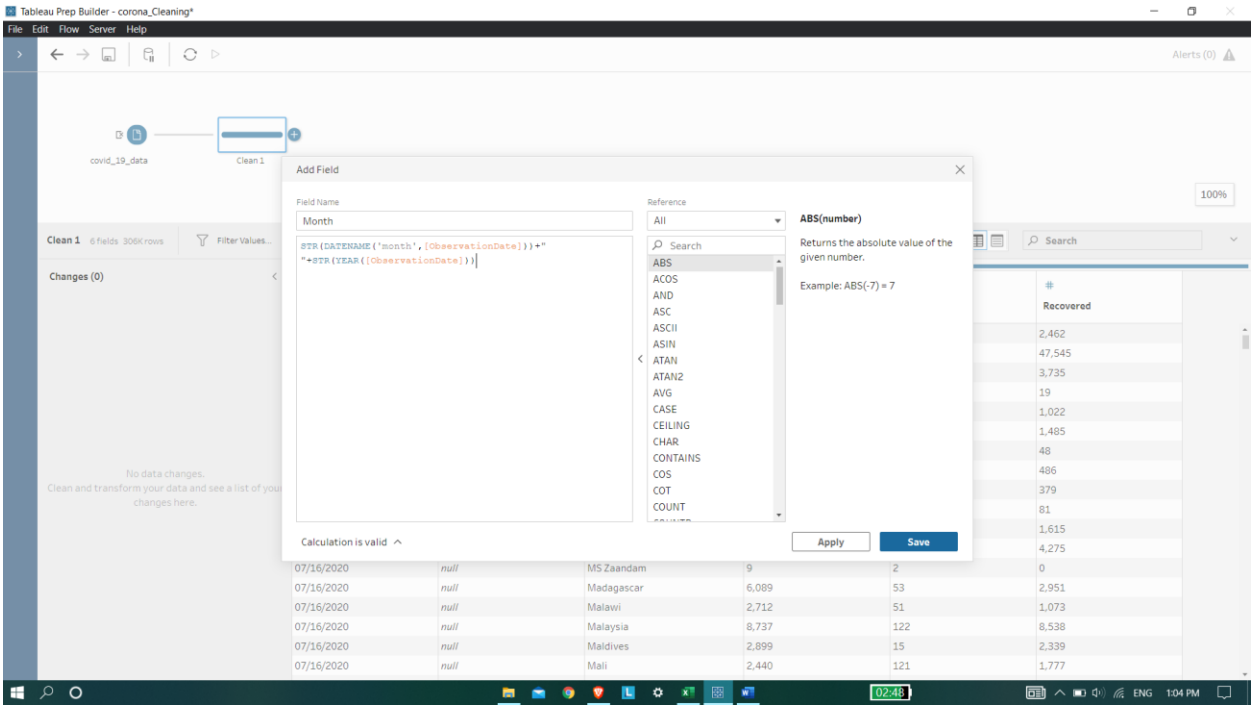
Step 2:

- Remove the unwanted column from the dataset (SNo and Last Update)

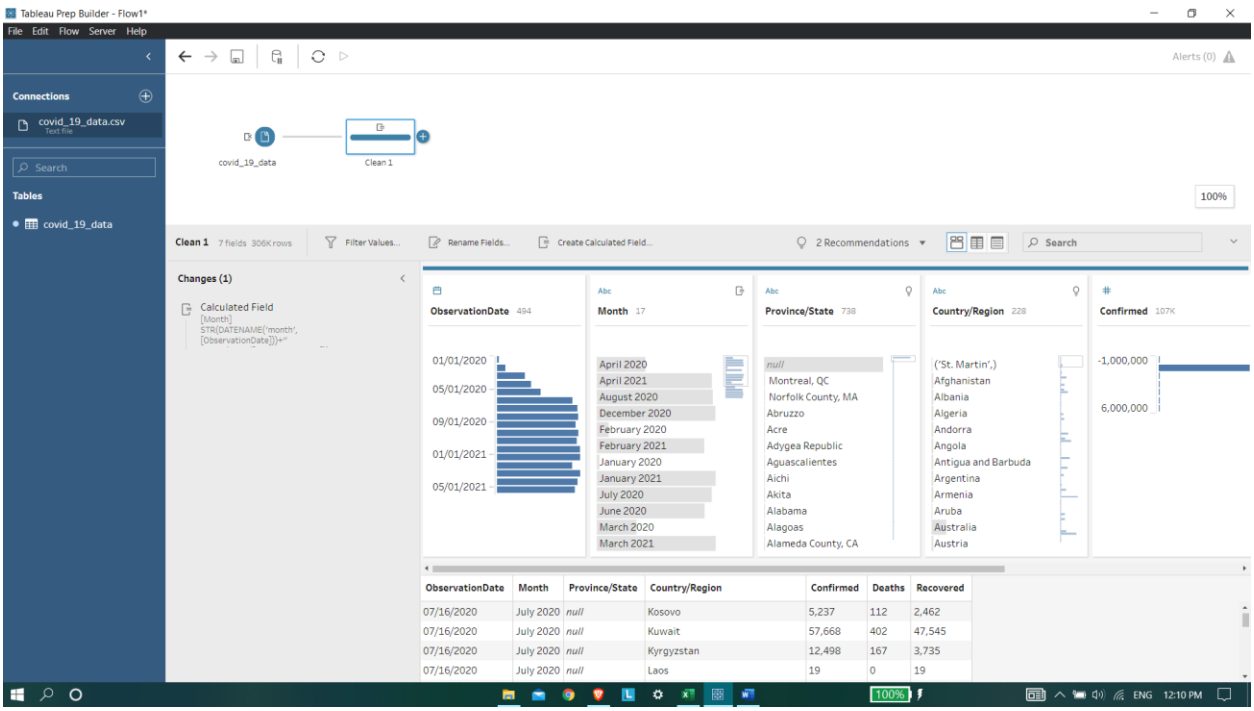


Step 3:

- Create a Calculated Field For Extracting Month Name and Year from the ObservationDate Column.

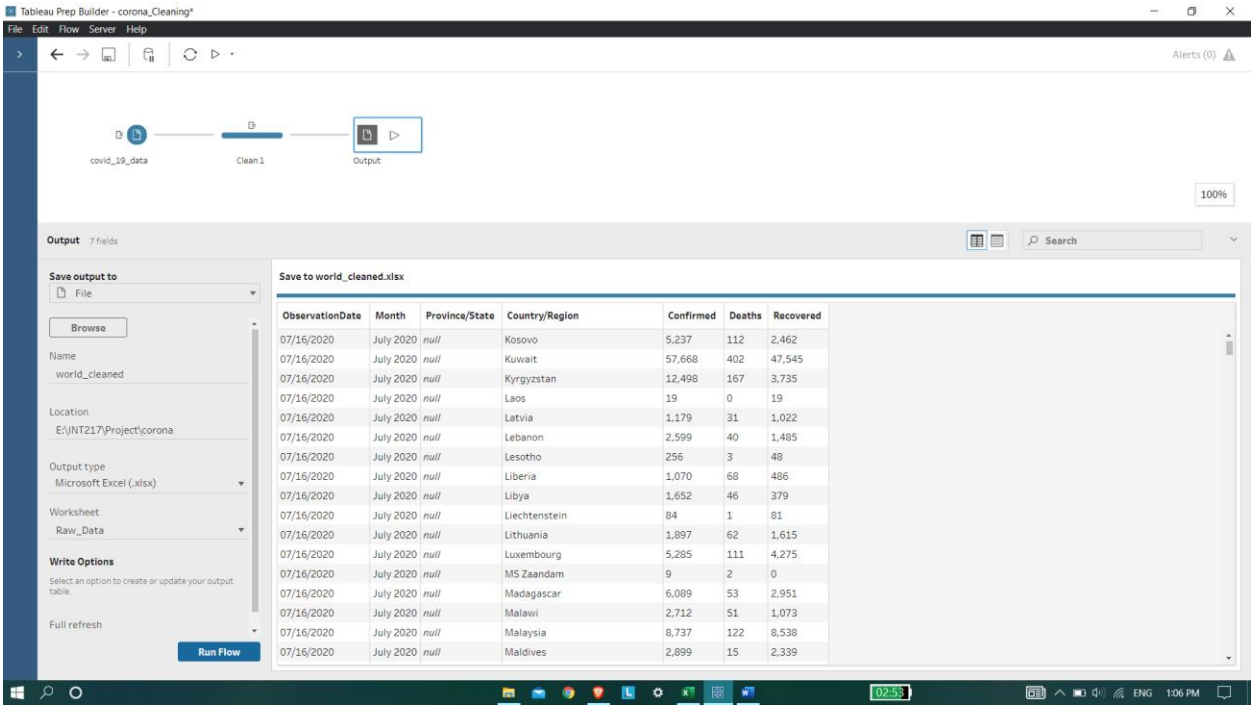


- After adding:



Step 4:

- Then save the output file in xlsx format



Step 5:

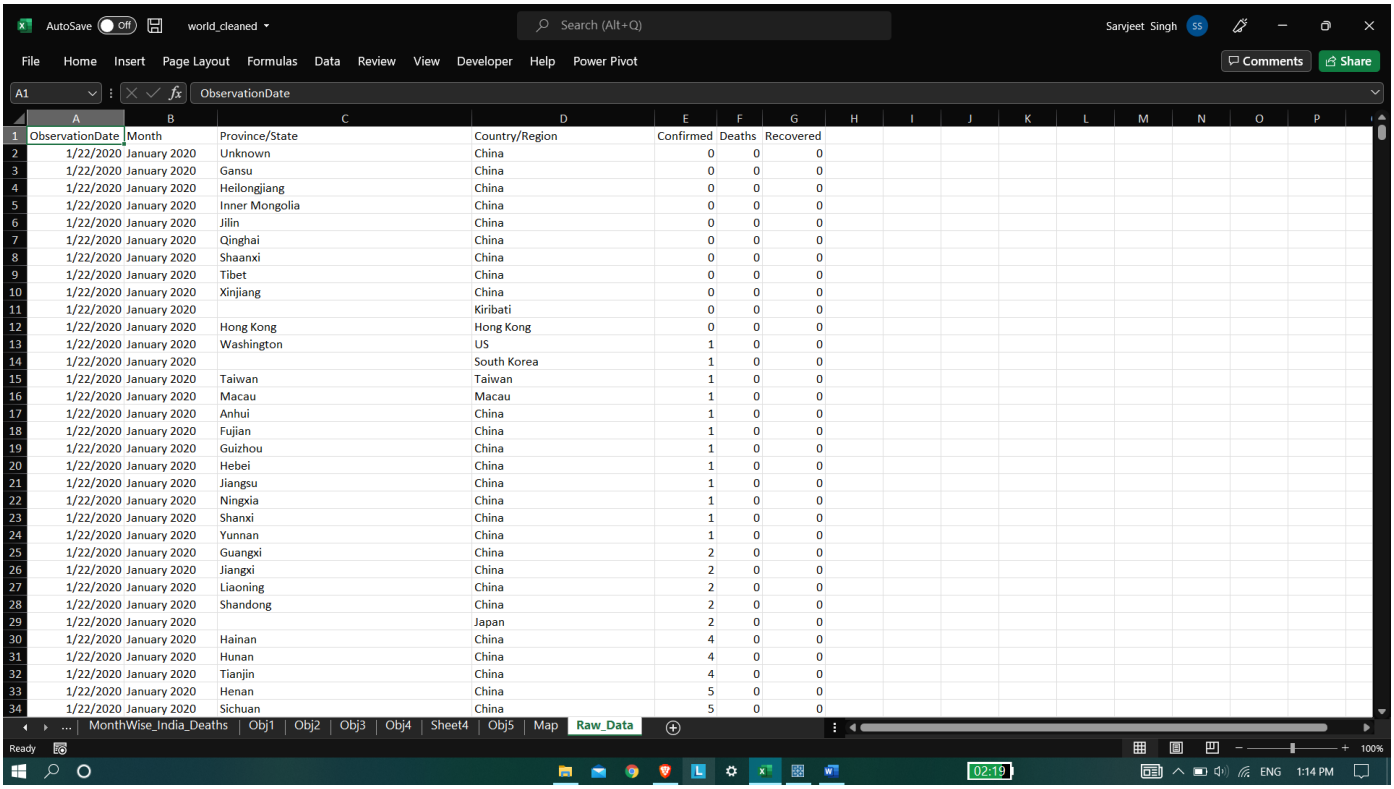
- Some data in Cases column were in negative so I changed them to positive

	A	B	C	D	E	F	G	H
1	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
2	147525	11/2/2020	Unknown	Colombia	4/2/2021 15:13	-302844	0	0
3	5	1/22/2020	Gansu	Mainland China	1/22/2020 17:00	0	0	0

	A	B	C	D	E	F	G	H
1	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
2	118364	9/24/2020	Unknown	Colombia	4/2/2021 15:13	0	-178	-12684
3	141535	10/25/2020	Unknown	Colombia	4/2/2021 15:13	0	-154	-8072
4	147525	11/2/2020	Unknown	Colombia	4/2/2021 15:13	302844	0	0

	A	B	C	D	E	F	G	H
1	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
2	145278	10/30/2020	Unknown	Colombia	4/2/2021 15:13	0	505	-854405
3	118364	9/24/2020	Unknown	Colombia	4/2/2021 15:13	0	-178	-12684
4	141535	10/25/2020	Unknown	Colombia	4/2/2021 15:13	0	-154	-8072
5	1	1/22/2020	Anhui	Mainland China	1/22/2020 17:00	1	0	0

After Cleaning:



AutoSave Off | world_cleaned | Search (Alt+Q) | Sarvjeet Singh

File Home Insert Page Layout Formulas Data Review View Developer Help Power Pivot | Comments | Share

A1 | ObservationDate

	A	B	C	D	E	F	G
1	ObservationDate	Month	Province/State	Country/Region	Confirmed	Deaths	Recovered
2	1/22/2020	January 2020	Unknown	China	0	0	0
3	1/22/2020	January 2020	Gansu	China	0	0	0
4	1/22/2020	January 2020	Heilongjiang	China	0	0	0
5	1/22/2020	January 2020	Inner Mongolia	China	0	0	0
6	1/22/2020	January 2020	Jilin	China	0	0	0
7	1/22/2020	January 2020	Qinghai	China	0	0	0
8	1/22/2020	January 2020	Shaanxi	China	0	0	0
9	1/22/2020	January 2020	Tibet	China	0	0	0
10	1/22/2020	January 2020	Xinjiang	China	0	0	0
11	1/22/2020	January 2020		Kiribati	0	0	0
12	1/22/2020	January 2020	Hong Kong	Hong Kong	0	0	0
13	1/22/2020	January 2020	Washington	US	1	0	0
14	1/22/2020	January 2020		South Korea	1	0	0
15	1/22/2020	January 2020	Taiwan	Taiwan	1	0	0
16	1/22/2020	January 2020	Macau	Macau	1	0	0
17	1/22/2020	January 2020	Anhui	China	1	0	0
18	1/22/2020	January 2020	Fujian	China	1	0	0
19	1/22/2020	January 2020	Guizhou	China	1	0	0
20	1/22/2020	January 2020	Hebei	China	1	0	0
21	1/22/2020	January 2020	Jiangsu	China	1	0	0
22	1/22/2020	January 2020	Ningxia	China	1	0	0
23	1/22/2020	January 2020	Shanxi	China	1	0	0
24	1/22/2020	January 2020	Yunnan	China	1	0	0
25	1/22/2020	January 2020	Guangxi	China	2	0	0
26	1/22/2020	January 2020	Jiangxi	China	2	0	0
27	1/22/2020	January 2020	Liaoning	China	2	0	0
28	1/22/2020	January 2020	Shandong	China	2	0	0
29	1/22/2020	January 2020		Japan	2	0	0
30	1/22/2020	January 2020	Hainan	China	4	0	0
31	1/22/2020	January 2020	Hunan	China	4	0	0
32	1/22/2020	January 2020	Tianjin	China	4	0	0
33	1/22/2020	January 2020	Henan	China	5	0	0
34	1/22/2020	January 2020	Sichuan	China	5	0	0

... | MonthWise_India_Deaths | Obj1 | Obj2 | Obj3 | Obj4 | Sheet4 | Obj5 | Map | Raw_Data

Ready | 100% | 02:19 | ENG 1:14 PM

Analysis on Dataset:

1. To display top 5 countries which are affected the most

Introduction

By performing this analysis, we will get top 5 countries having highest no of cases.

Description

Analysis is based on Country and Confirmed Cases.

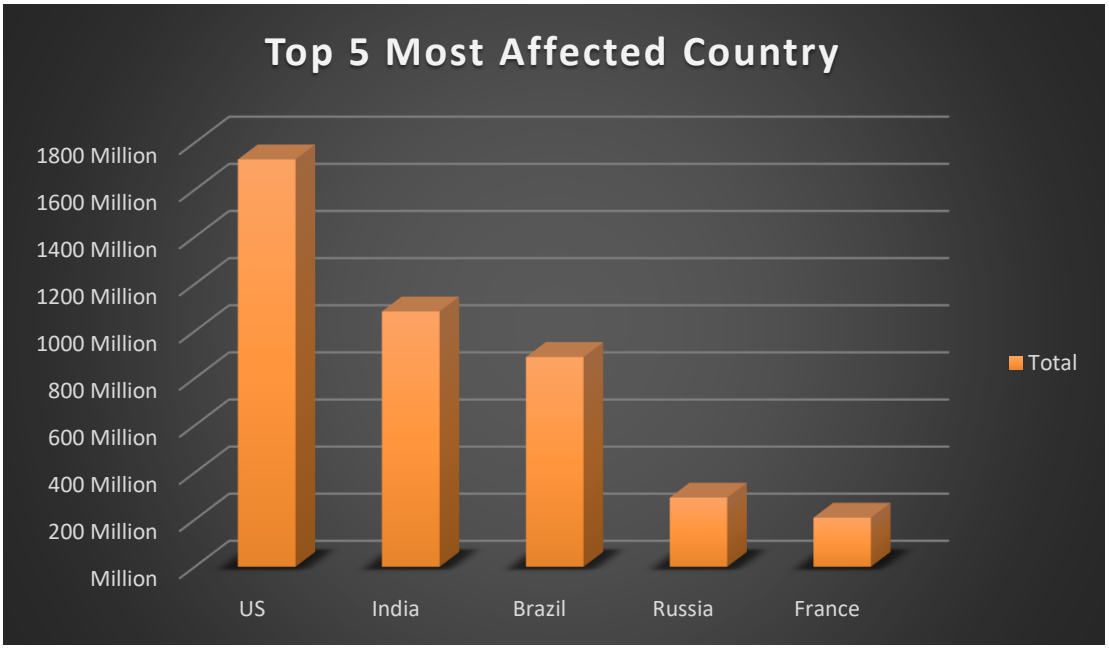
Specific requirements:

Used Pivot table and chart on column Country and Confirmed Cases.

Then Filtered The table based on Total Confirmed Cases to show only top 5.

Analysis Results:

Confirmed	Confirmed_Cases
US	6049 Million
India	3227 Million
Brazil	2654 Million
Russia	931 Million
France	855 Million
Grand Total	**13715 Million**



2. To Show Top 5 country having highest number of death due to Corona

Introduction

By performing this analysis, we will get top 5 countries having highest no of death

Description

Analysis is based on Country and Death Cases and Death Rate.

Specific requirements:

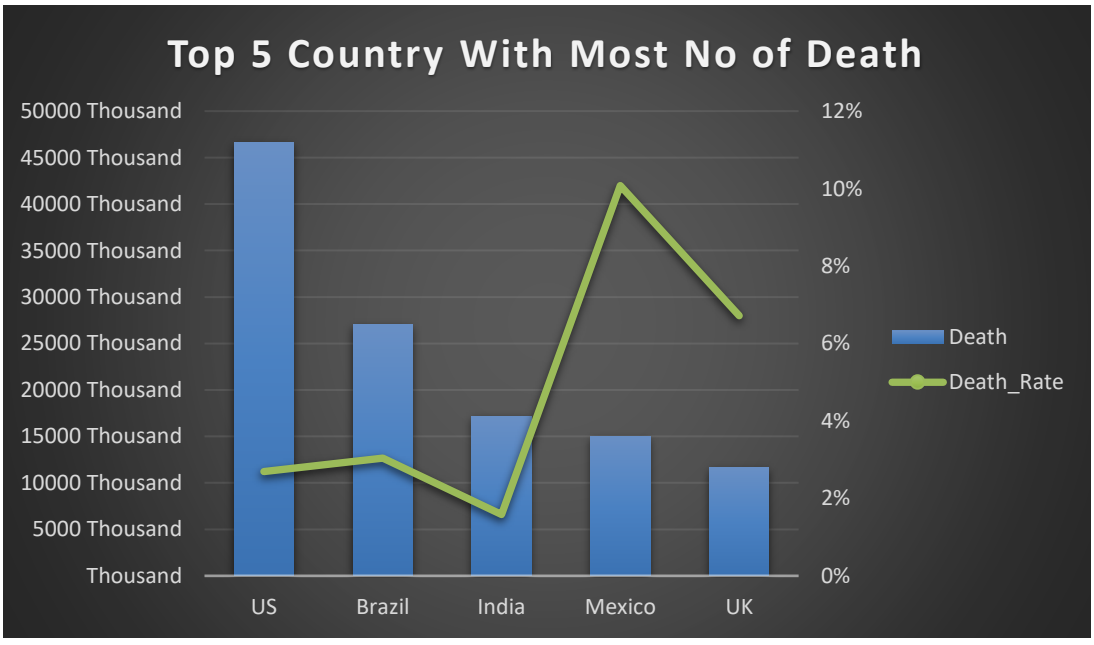
Used Pivot table and chart on column Country and Death Cases.

Made a Calculated Column for Death Rate=(Deaths/Confirmed)

Then Filtered The table based on Total Death Cases to show only top 5.

Analysis Results:

Confirmed	Death	Death_Rate
US	123304 Thousand	2%
Brazil	72625 Thousand	3%
India	44425 Thousand	1%
Mexico	43006 Thousand	9%
UK	29172 Thousand	4%
Grand Total	**312531 Thousand**	**2%**



3. To Show Five Least Affected Countries

Introduction

By performing this analysis, we will get top 5 countries which are least affected.

Description

Analysis is based on Country and Confirmed

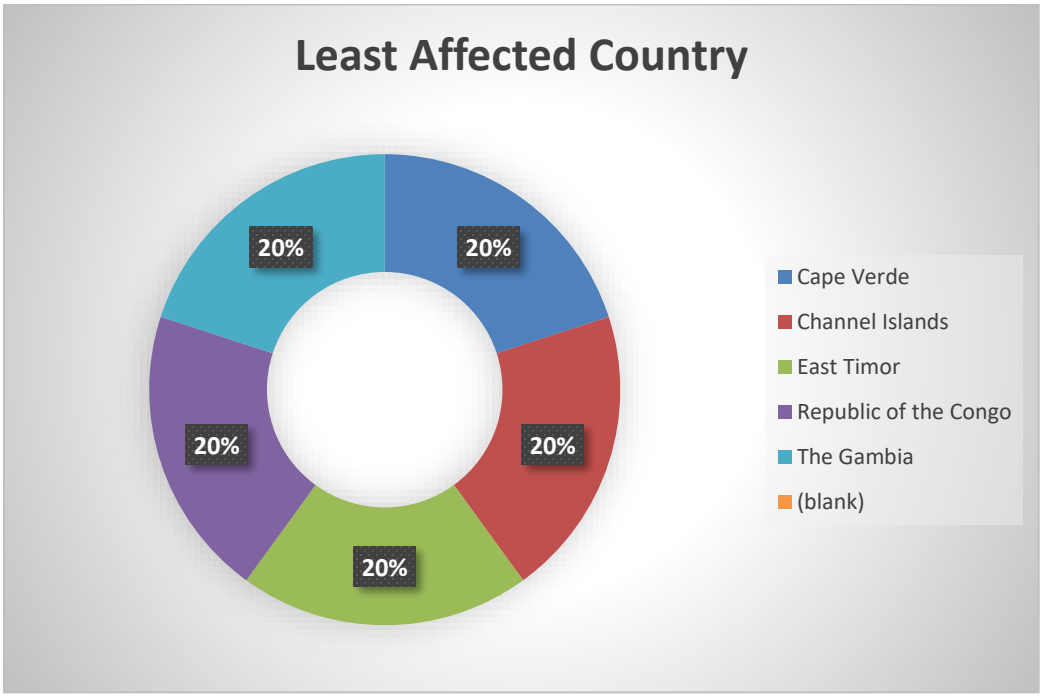
Specific requirements:

Used Pivot table and chart on column Country and Confirmed.

Then Filtered The table based on Total Confirmed Cases to show only top 5.

Analysis Results:

Country	Confirmed_Cases
Cape Verde	1
Channel Islands	1
East Timor	1
Republic of the Congo	1
The Gambia	1
(blank)	



4. To Show Top Five Countries with Most Recovered Cases

Introduction

By performing this analysis, we will get top 5 countries with Most Recovered Cases.

Description

Analysis is based on Country and Recovered and Recovered Rate.

Specific requirements:

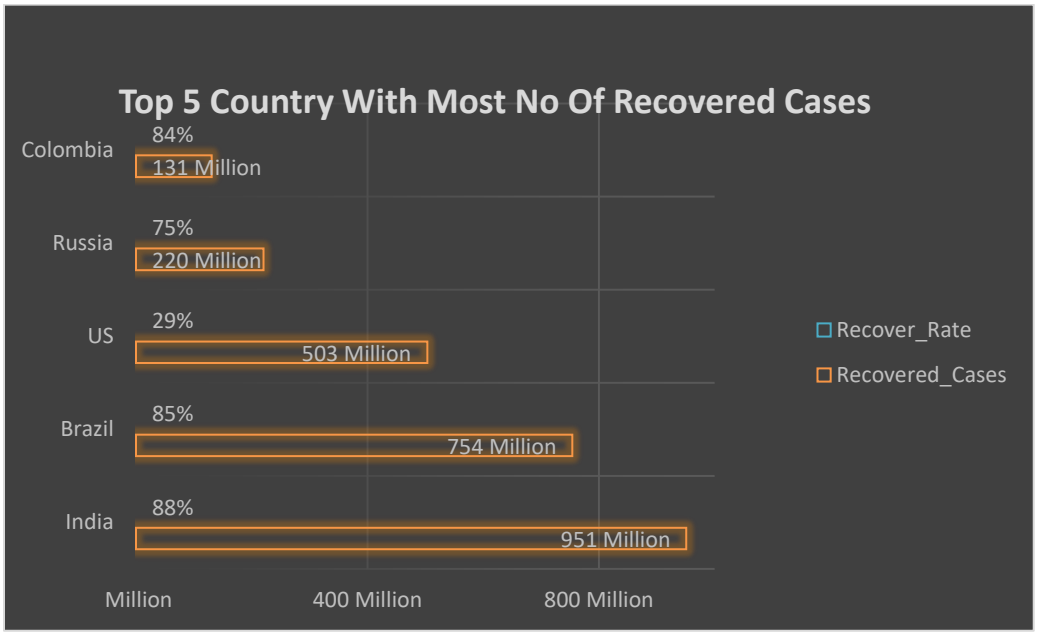
Used Pivot table and chart on column Country and Death Cases.

Made a Calculated Column for Recovered Rate=(Recovered/Confirmed)

Then Filtered The table based on Total Death Cases to show only top 5.

Analysis Results:

Country	Recovered_Cases	Recover_Rate
India	2901 Million	90%
Brazil	2314 Million	87%
Russia	791 Million	85%
Turkey	564 Million	91%
US	503 Million	8%
Grand Total	**7073 Million**	**52%**



5. Second Wave in India

Introduction

By performing this analysis, we will get to know how India's second wave happened.

Description

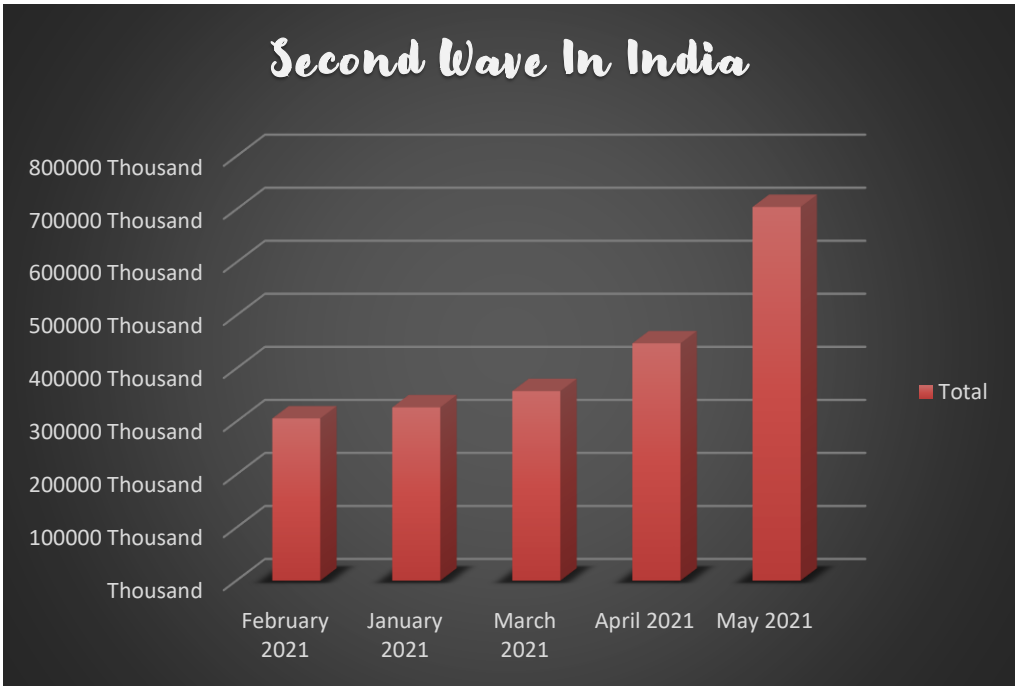
Analysis is based on Country and Confirmed.

Specific requirements:

Used Pivot table and chart on column Month and Confirmed Cases and in filtered on Country(India)

Analysis Results:

Country/Region	India
Month	**Confirmed_Cases**
February 2021	305984 Thousand
January 2021	326909 Thousand
March 2021	357487 Thousand
April 2021	447599 Thousand
May 2021	704714 Thousand
Grand Total	**2142693 Thousand**



6. Confirmed Cases World Map

Introduction

By performing this analysis, we will get to know which countries and its neighbor are how much affected.

Description

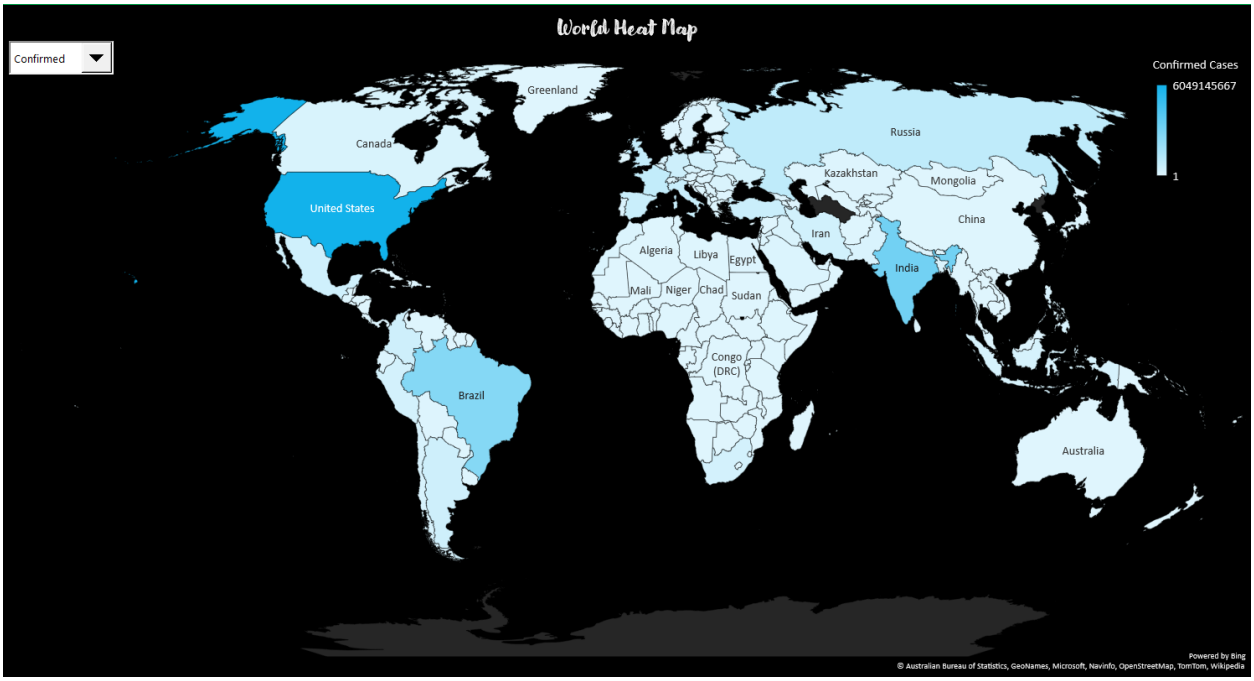
Analysis is based on Country and Confirmed.

Specific requirements:

Used Pivot table and chart on column Country and Confirmed Cases, Death, Recovered and a Calculated Field Active Cases then used offset formula to Copy the data to another sheet as Map chart can't be made on Pivot Table.

And Used Drop down column to select the either of these fileds

Analysis Results:



7. Indian State Wise Confirmed, Recovered and Death Cases

Introduction

By performing this analysis, we will get idea about the states Confirmed, Recovered and Death cases of India due to Covid.

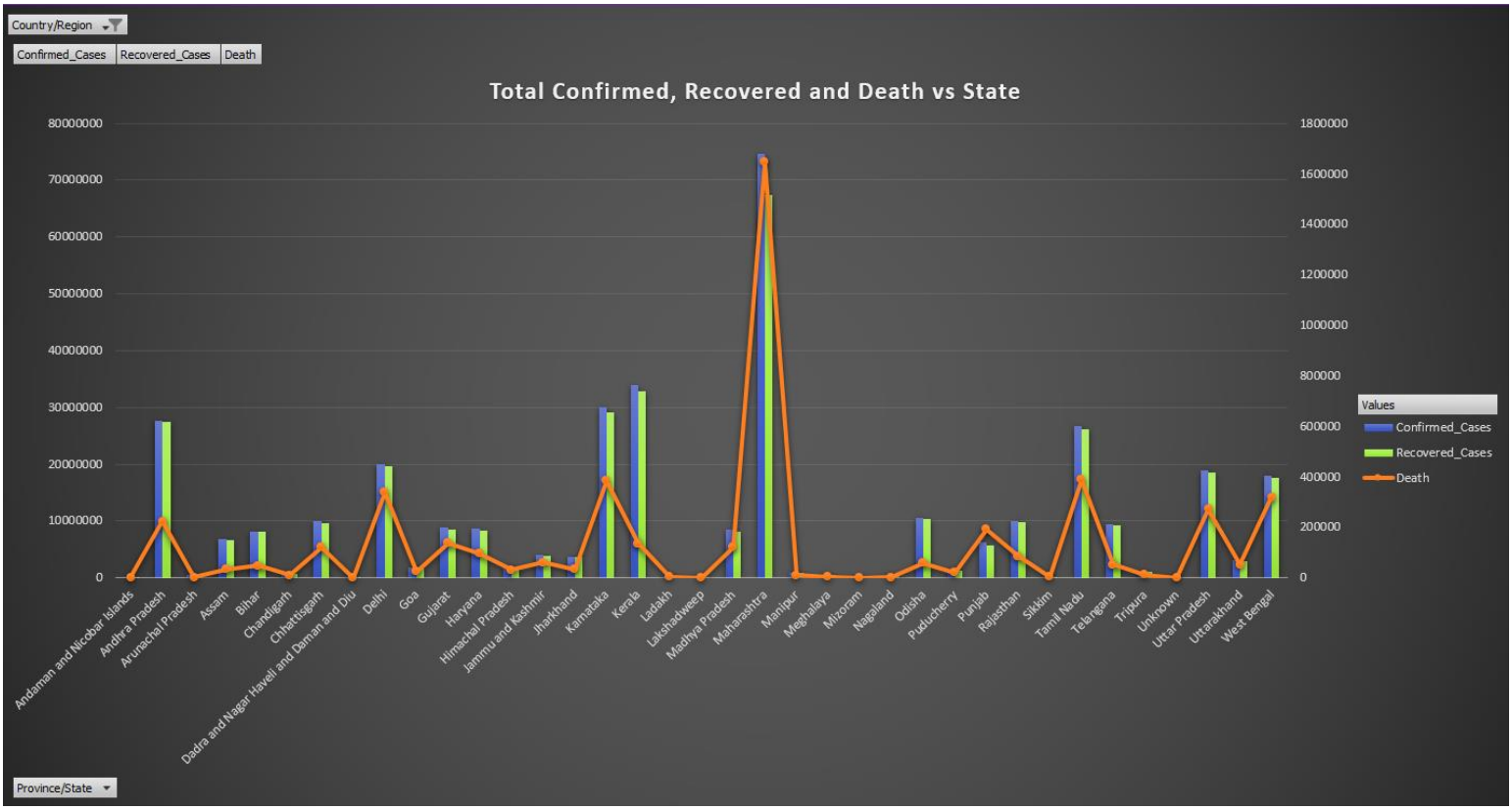
Description

Analysis is based on State and Confirmed and Recovered and Death.

Specific requirements:

Used Pivot table and chart on column State and Confirmed, Recovered and Death Cases. And filtered on India

Analysis Results:



8. Tree Map of Confirmed Cases in States

Introduction

By performing this analysis, we will get idea about the states Confirmed cases of India due to Covid.

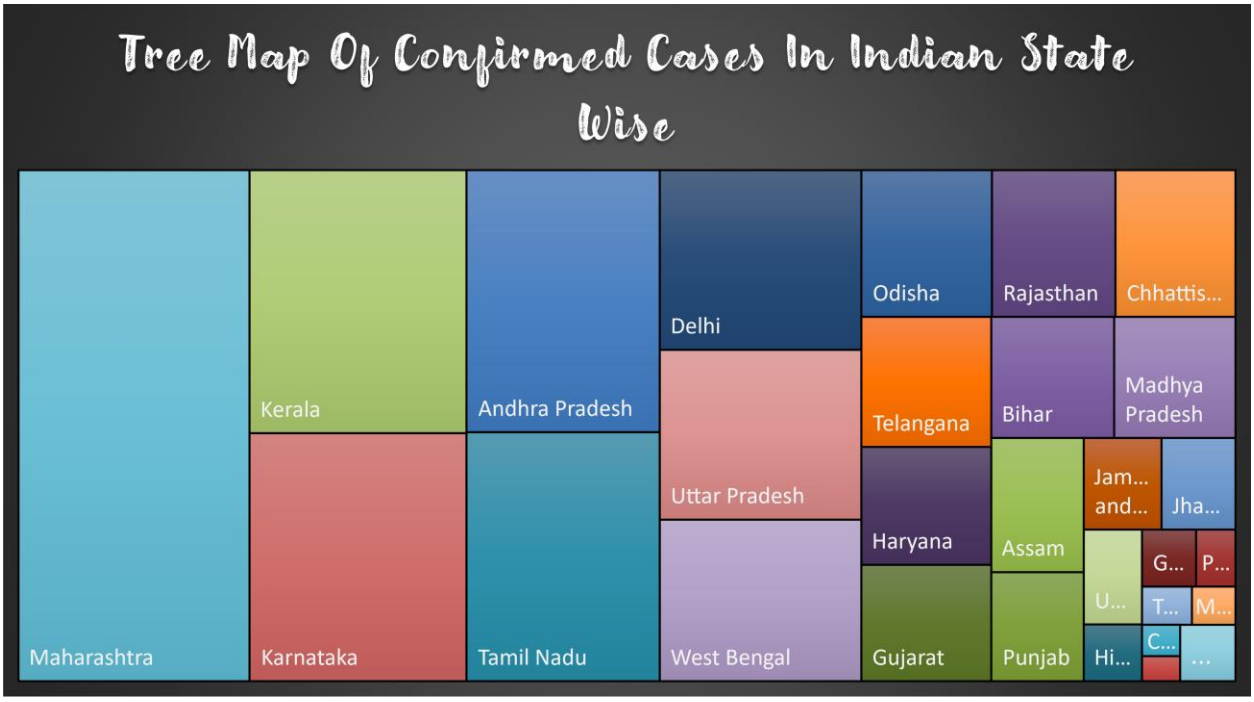
Description

Analysis is based on State and Confirmed.

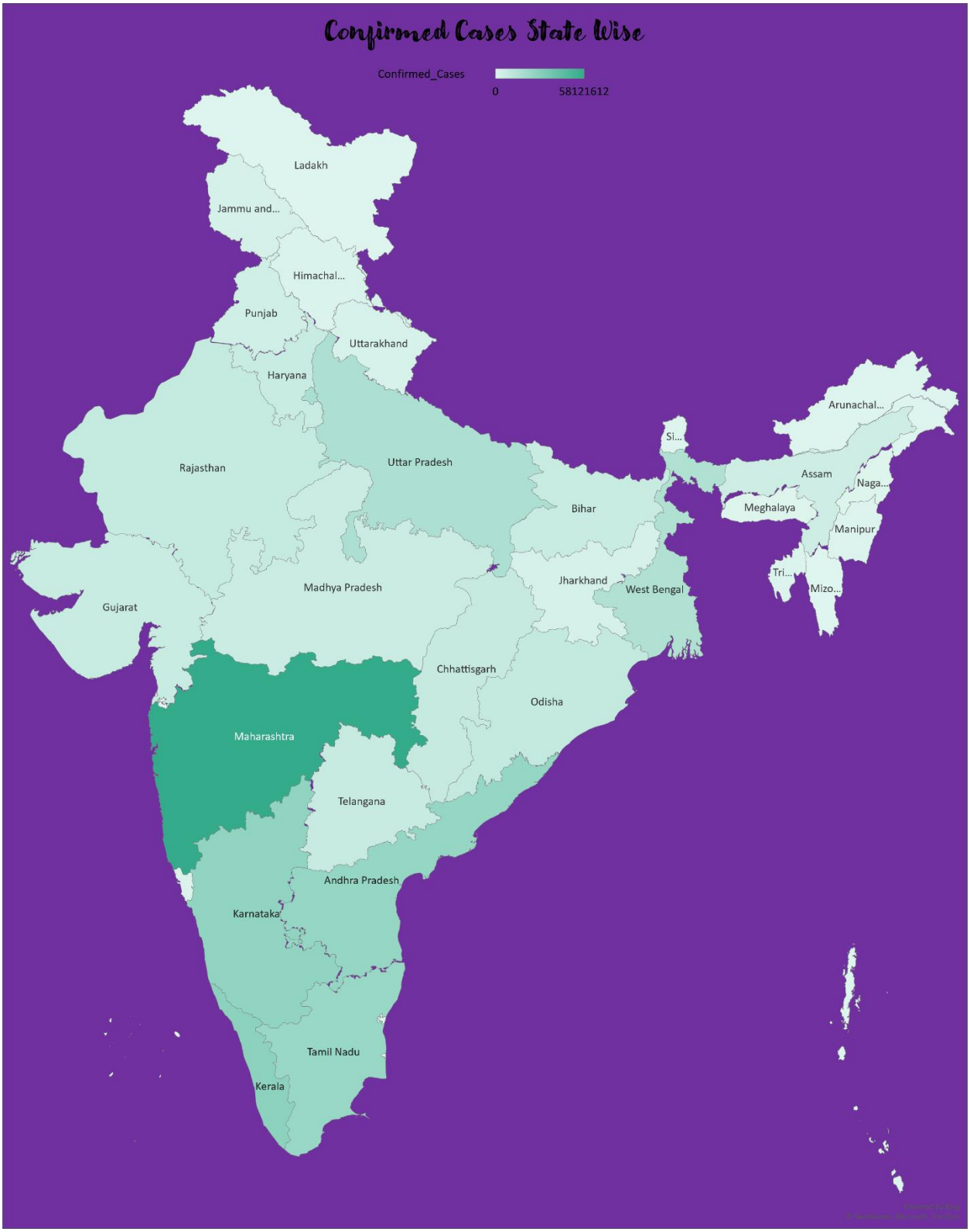
Specific requirements:

Used Pivot table and chart on column State and Confirmed, Recovered and Death Cases. And filtered on India

Analysis Results:



9. Confirmed Cases State Wise Map



10. Indian Month Wise Confirmed, Recovered and Death Cases

Introduction

By performing this analysis, we will get India's Confirmed, Recovered and Death Monthly

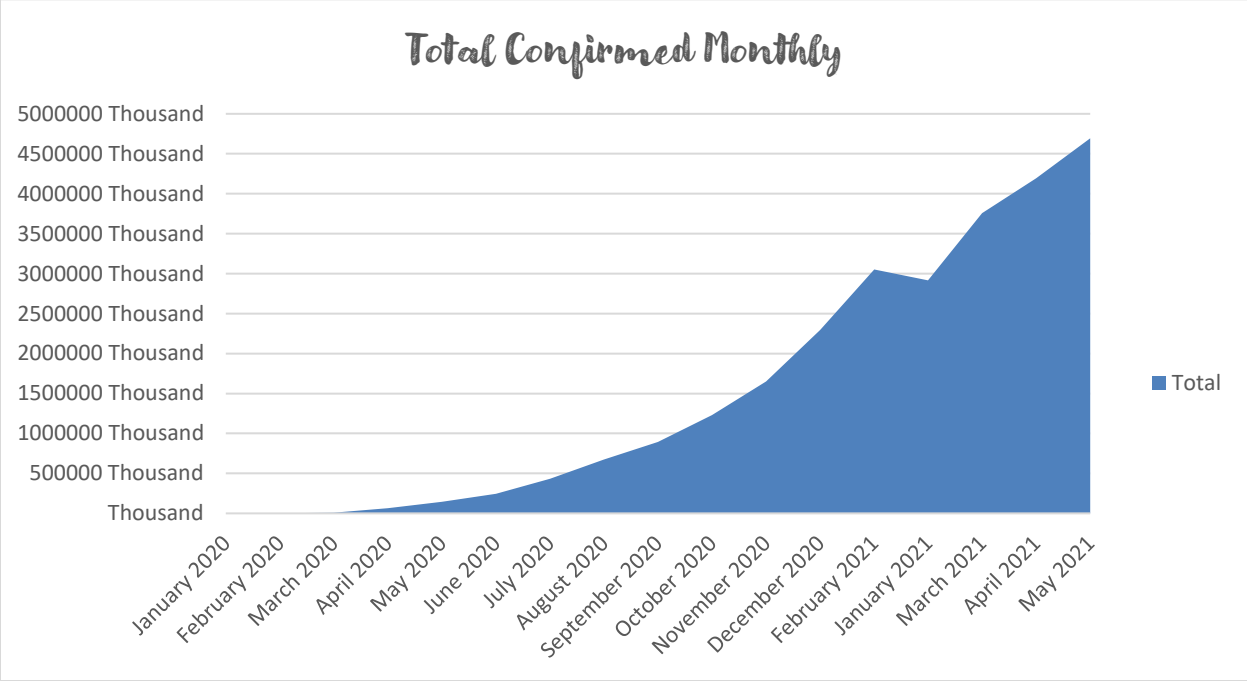
Description

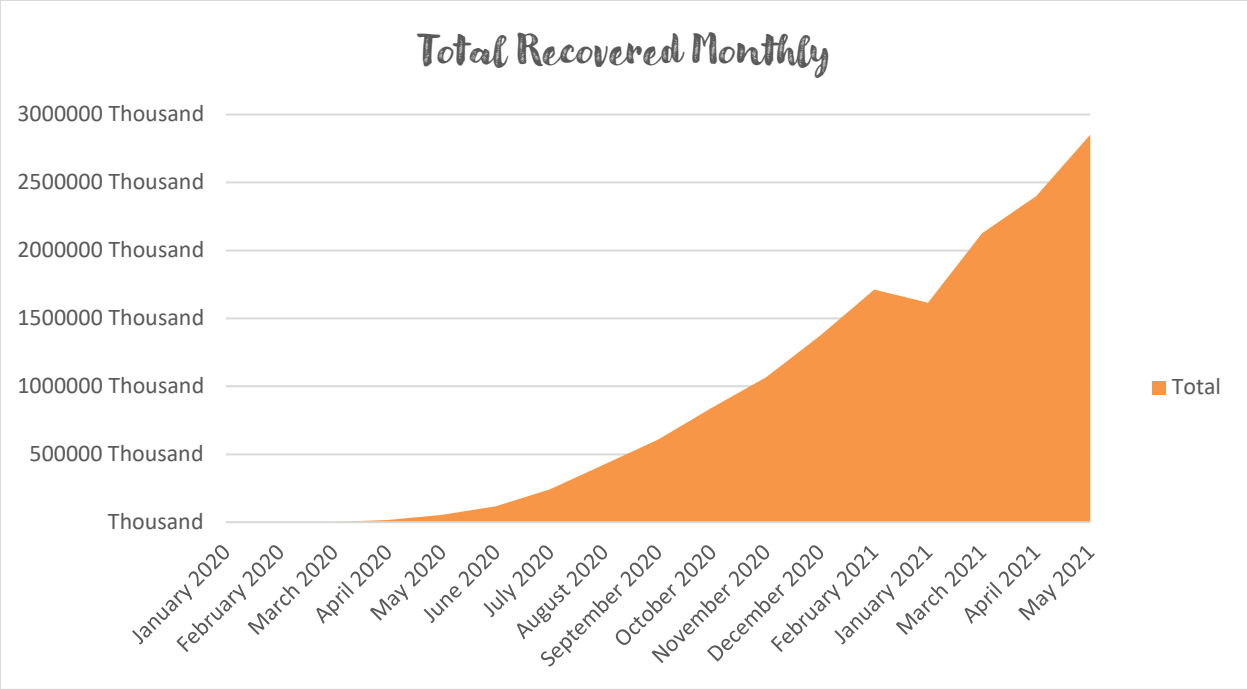
Analysis is based on Country and Confirmed, Recovered and Death Cases.

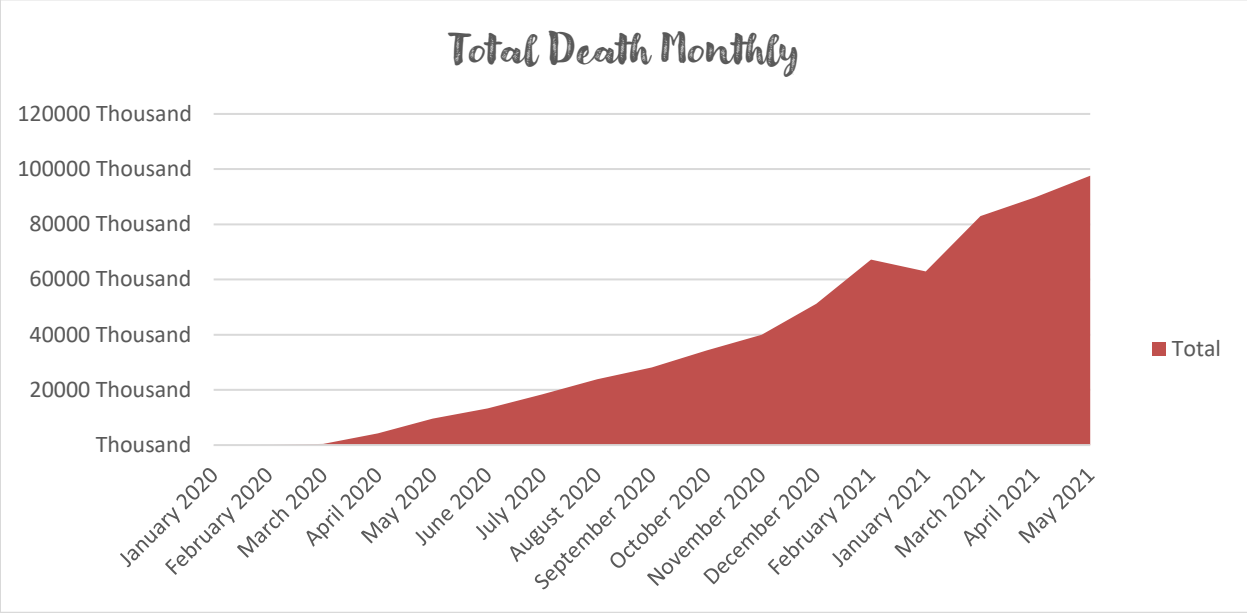
Specific requirements:

Used Pivot table and chart on column Country, Confirmed, Recovered and Death. And filtered on India

Analysis Results:




Total Recovered Monthly
3000000 Thousand
2500000 Thousand
2000000 Thousand
1500000 Thousand
1000000 Thousand
500000 Thousand
Thousand
January 2020
February 2020
March 2020
April 2020
May 2020
June 2020
July 2020
August 2020
September 2020
October 2020
November 2020
December 2020
February 2021
January 2021
March 2021
April 2021
May 2021
Total


Total Death Monthly
120000 Thousand
100000 Thousand
80000 Thousand
60000 Thousand
40000 Thousand
20000 Thousand
Thousand
January 2020
February 2020
March 2020
April 2020
May 2020
June 2020
July 2020
August 2020
September 2020
October 2020
November 2020
December 2020
February 2021
January 2021
March 2021
April 2021
May 2021
Total

FINAL DASHBOARDS