**INTRODUCTION TO DATA MANAGEMENT PROJECT REPORT**

August-December 2020

***COVID-19 Data Analysis***

Submitted by-

Bhavesh Gujjula

Registration No- 11801452

Section- KM071

Course Code- INT217

Under the Guidance of

**Ms. Savleen Kaur – 18306**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**



**DECLARATION**

I, Bhavesh Gujjula student of B. Tech CSE 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: 17-12-2020 Registration No.11801452 Bhavesh Gujjula

**ACKNOWLEDGEMENT**

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. A special gratitude I give to my **mentor** whose contribution in stimulating suggestions and encouragement helped me to coordinate my project especially in writing this report. I express my thanks to my institution **Lovely Professional University** for giving me an opportunity to learn this interesting topic. I also convey my regards to my faculty assistance all through this training named **“Data Science Project on COVID-19 Data Analysis”.** Once again, I would like to thank all my supporters from the core of my heart.

|  |  |  |
| --- | --- | --- |
| **S.NO** | **TITLE** | **Page. No** |
| 1 | INTRODUCTION | 5 |
| 2 | OBJECTIVE / SCOPE OF ANALYSIS | 6 |
| 3 | SOURCE OF DATA | 7 |
| 4 | ETL PROCESS | 8 |
| 5 | ANALYSIS ON DATASET | 9 |
| 6 | VISUALIZATION | 12 |
| 7 | REFERENCE | 12 |

**INTRODUCTION**

Data Analysis is a process of collecting, transforming, cleaning, and modeling data with the goal of discovering the required information. The results so obtained are communicated, suggesting conclusions, and supporting decision-making. Data visualization is at times used to portray the data for the ease of discovering the useful patterns in the data. The terms Data Modeling and Data Analysis mean the same.

The World Health Organisation (WHO) has declared the coronavirus disease 2019 (COVID-19) a pandemic. A global coordinated effort is needed to stop the further spread of the virus. A pandemic is defined as “occurring over a wide geographic area and affecting an exceptionally high proportion of the population.” The last pandemic reported in the world was the H1N1 flu pandemic in 2009.

On 31 December 2019, a cluster of cases of pneumonia of unknown cause, in the city of Wuhan, Hubei province in China, was reported to the World Health Organisation. In January 2020, a previously unknown new virus was identified, subsequently named the 2019 novel coronavirus, and samples obtained from cases and analysis of the virus’ genetics indicated that this was the cause of the outbreak. This novel coronavirus was named Coronavirus Disease 2019 (COVID-19) by WHO in February 2020. The virus is referred to as SARS-CoV-2 and the associated disease is COVID-19.

**Insights of data:**

dateRep – Date on which the cases are reported

day – On which day, the cases were reported

month – In which month, the cases were reported

year – In which year, the case was reported

cases – Count of the positive Covid cases

deaths – Count of the Deaths due to the virus

countries and Territories – Countries which hare effected by virus

geoId – Geographical Id of the countries

countryterritoryCode – Territory code of various countries

continentExp – Name of the continents effected by the virus

**SCOPE OF ANALYSIS**

The current coronavirus outbreak represents a formidable challenge to governments and healthcare professionals around the world. The speed at which the virus has spread has left most countries scrambling for a response and often highlights how poorly prepared their institutions and healthcare systems are equipped to deal with a pandemic of this proportion.

This project on COVID-19 provides the overall Statistics details on the daily number of new reported COVID-19 cases and deaths worldwide from 31 Dec 2019 to 29 Nov 2020.

**Objectives of this project:**

1. To analyse the positive cases and deaths across countries.
2. To analyse the positive cases and deaths across continents.
3. To find the Top 5 countries which are most effected by the virus.
4. To visualize the deaths and positive cases on monthly bases.
5. To analyse the positive cases and deaths in India.

**SOURCE OF DATASET**

This data set is downloaded from ECDC, It is an EU agency aimed at strengthening Europe's defences against infectious diseases. The core functions cover a wide spectrum of activities: surveillance, epidemic intelligence, response, scientific advice, microbiology, preparedness, public health training, international relations, health communication, and the scientific journal *Eurosurveillance*.

ECDC disease programmes cover antimicrobial resistance and healthcare-associated infections; emerging and vector-borne diseases; food- and waterborne diseases and zoonoses; HIV, sexually transmitted infections and viral hepatitis; influenza and other respiratory viruses; tuberculosis; and vaccine-preventable diseases.

Link - <https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide>

**ETL PROCESS**

In computing, extract, transform, load (ETL) is a process in database usage to prepare data for analysis, especially in data warehousing. Data extraction involves extracting data from homogeneous or heterogeneous sources, while data transformation processes data by transforming them into a proper storage format/structure for the purposes of querying and analysis; finally, data loading describes the insertion of data into the final target database such as an operational data store, a data mart, or a data warehouse. A properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that separate sources can be used together, and finally delivers data in a presentation-ready format so that application developers can build applications and end users can make decisions.

**Step 1: Removing the blank cells from the dataset.**

For this, select the whole dataset. Go to Find and Select in the Home tab of excel. Select Go to Special from the drop-down menu and then tick the blank option. All the blank cells will be selected. Then go to Delete option in the home tab again and select Delete Rows from the drop-down menu. This will remove any rows with blank cells.

**Step 2: Removing columns which are not properly defined or not crucial to our analysis.**

For this we will have columns which are redundant like the column with just the index numbers. For this we will select that particular column and then go to delete option in the home tag and then select Delete Columns from the drop-down menu.

**Step 3: Giving proper and appropriate column names.**

The dataset does not have proper columns so our next step would be to give proper column names to the columns wherever required.

**Step 4: Excluding the NULL values from the data.**

We’ll be using Tableau prep for this work as it’ll make the work simple and faster because we might not know how many null values could be there in this huge data set. Tableau helps us do one step cleaning with ease.

**Step 5: Improvising Proper Data Formatting**

Without proper Data Formatting, proper analysis will not take place. So, we will bring down certain columns to their proper format. For example, the dates should be in the date format and price and sales should be in currency format for better results.

**Step 6: Removing Duplicate Values**

It might be possible that our data may be containing duplicate values which may hinder in precise analysis. So, our last task in ETL will be removing duplicate values and making our data perfect for analysis.

**ANALYSIS ON DATASET**

1. **Countries wise victims**

**Description:** We will calculate the total number of positive cases in each country.

**Requirements:** PowerPivot, pivot table and pivot charts

**Visualization:** Line Chart

**Result**:

1. **Continents wise victims**

**Description:** We will calculate the total number of positive cases in each continent.

**Requirements:** PowerPivot, pivot table and pivot charts

**Visualization:** Clustered Column

**Result**:

1. **Deaths and Cases on monthly base**

**Description:** Recording the highest positive case and death month wise.

**Requirements:** PowerPivot, pivot table and pivot charts

**Visualization:** Line with markers

**Result**:

1. **Top 5 Major Infected Countries**

**Description:** We will be top 5 most infected countries

**Requirements:** PowerPivot, pivot table and pivot charts

**Visualization:** Clustered Bar

**Result**:

1. **Continent wise Deaths**

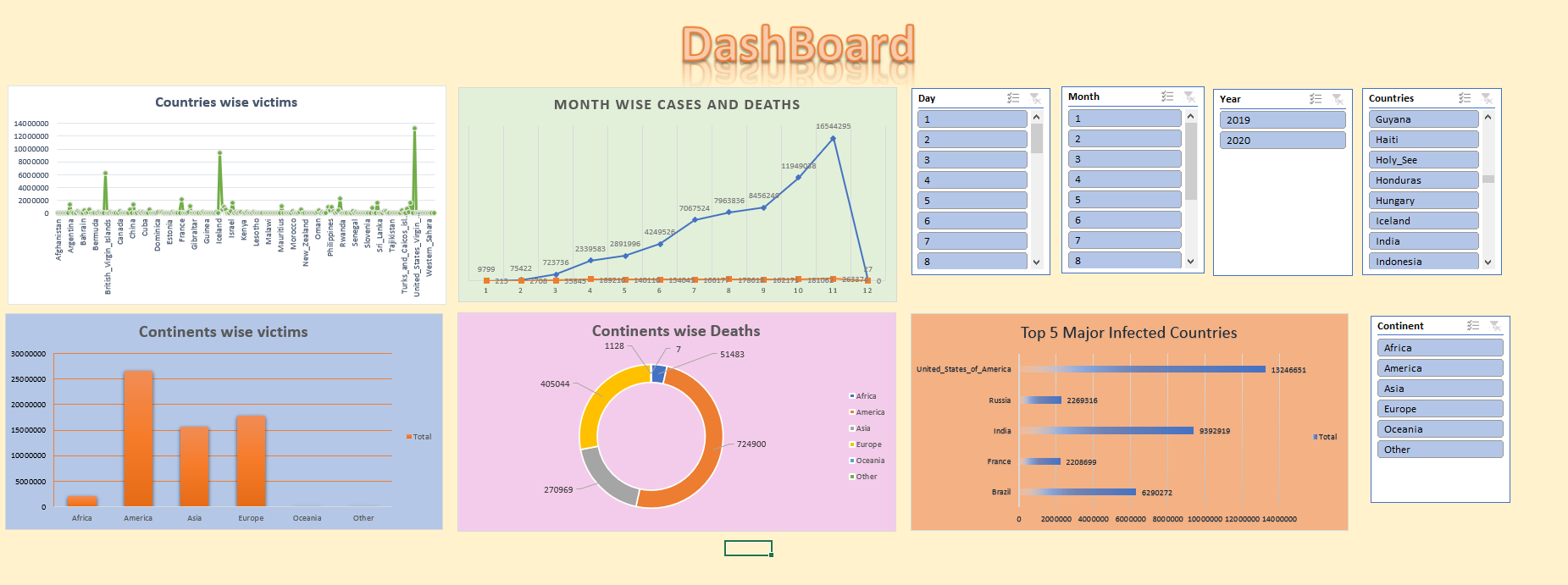
**Description:** We will be calculating number of deaths continent wise

**Requirements:** PowerPivot, pivot table and pivot charts

**Visualization:** Doughnut

**Result:**

**DASHBOARD**



**REFERENCE AND BIBILOGRAPHY**

1. European Centre for Disease Prevention and Control
2. YouTube
3. World Health Organization