**INTRODUCTION TO DATA MANAGEMENT**

**PROJECT REPORT**

5TH Semester August-December 2021

**DATA ANLYSIS ON OLYMPICS**

Submitted by

Garima Kehar

**Registration No.:** 11902683

**Programme:** Bachelor’s in Technology

**Section:** KM002

**Course Code:** INT 217

**Under the Guidance of**

**Name of faculty: Maneet Kaur**

**U Id: 15709**

**A picture containing text

Description automatically generatedDiscipline of CSE/IT**

**Lovely School of Computer Science & Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that I Garima Kehar having registration no 11902683 has completed INT 217 project titled, **“COVID-19 ANALYSIS ON GLOBAL AND INDIAN DATASET”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab.

Date: 8/12/2021

**DECLARATION**

I, Garima Kehar, student of bachelor’s in technology 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: 08/12/2021

Name of the student: Garima Kehar

Registration No.: 11902683

**ACKNOWLEDGEMENT**

I would like to express my gratitude towards all the people who have contributed their precious time and effort to help me. Without whom it would not have been possible for me to understand and complete the project.

I would like to thank Professors of Department of Computer Science and Engineering, my Project Supervisor for his guidance, support, motivation and encouragement through out the period this work was carried out. His readiness for consultation at all times, his educative comments, his concern and assistance even with practical things have been invaluable.

We are grateful to Department of Computer Science and Engineering for providing necessary facilities in the department.

**Name:** Garima Kehar

**Registration No.:** 11902683

**Table of Content**

|  |  |  |
| --- | --- | --- |
| **SL NO.** | **Content** | **Page No.** |
| 1 | Cover Page | **1** |
| 2 | Certificate | **2** |
| 3 | Declaration | **3** |
| 4 | Acknowledgement | **4** |
| 5 | Index | **5** |
| 6 | Introduction | **6** |
| 7 | Objective/Scope of Analysis | **7** |
| 8 | Source of dataset | **8** |
| 9 | ETL Process | **9-10** |
| 10 | Analysis on dataset | **11-12** |
| 11 | Objective 1 | **12-15** |
| 12 | Objective 2 | **15-19** |
| 13 | Objective 3 | **19-21** |
| 14 | Objective 4 | **21-24** |
| 15 | Objective 5 | **24-29** |
| 16 | Objective 6 | **29-31** |
| 17 | List of analysis | **32-33** |
| 18 | Bibliography | **34** |
|  |  |  |

**INTRODUCTION**

Coronaviruses are a family of viruses that can cause respiratory illness in humans. They are called “corona” because of crown-like spikes on the surface of the virus. Severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS) and the common cold are examples of coronaviruses that cause illness in humans.

The new strain of coronavirus — COVID-19 — was first reported in Wuhan, China in December 2019. The virus has since spread to all continents.

The number of people infected changes daily. As of this writing (10/11/2021), more than 196,910,000 people in the world have been infected. Over 4,850,000 people have died. Some 192 countries and territories on all continents have now reported cases of COVID-19.

COVID-19 enters your body through your mouth, nose or eyes (directly from the airborne droplets or from transfer of the virus from your hands to your face). The virus travels to the back of your nasal passages and mucous membrane in the back of your throat. It attaches to cells there, begins to multiply and moves into lung tissue. From there, the virus can spread to other body tissues.

COVID-19 is likely spread:

* The virus travels in respiratory droplets released into the air when an infected person coughs, sneezes, talks, sings or breathes near you (within 6 feet). You may be infected if you inhale these droplets.
* You can also get COVID-19 from close contact (touching, shaking hands) with an infected person and then touching your face.

It’s considered possible to get COVID-19 after touching a contaminated surface and then touching your eyes, mouth, or nose before washing your hands. But it’s thought to be unlikely.

Coronaviruses are often found in bats, cats and camels. The viruses live in but do not infect the animals. Sometimes these viruses then spread to different animal species. The viruses may change (mutate) as they transfer to other species. Eventually, the virus can jump from animal species and begins to infect humans. In the case of COVID-19, the first people infected in Wuhan, China are thought to have contracted the virus at a food market that sold meat, fish and live animals. Although researchers don’t know exactly how people were infected, they already have evidence that the virus can be spread directly from person to person through close contact.

**OBJECTIVE**

The main objective to perform analysis on this data set is because I want to study that how many countries are effected by corona and which country have most number of death cases or active cases

Versus which state has most no of active and death cases

2nd objective is to know which country get vaccinated vs which state get both vaccination dose.

3RD main objective is who unemployment in India due to Covid-19 effect the people.

**DATA PROCESSING**

**DATA CLEANING**

**COLLECTION OF RAW DATA**

**EXPLORATORY DATA ANALYSIS**

**(VISUALIZATION)**

**DATA**

**SOURCES**

**LINE GRAPGH, RADAR GRAPH**

**PIE CHART, DONUT GRAPH**

**COLUMN GRAPH, BAR GRAPH**

**SOURCE OF DATASET:**

<https://www.mygov.in/covid-19/>

<https://unemploymentinindia.cmie.com/>

<https://www.worldometers.info/coronavirus/>

**ETL PROCESS**

ETL, which stands for Extraction, Transformation and Loading, is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or other target system. It is the foundation of data analytics and a process for integrating and loading data for computation and analysis, eventually becoming the primary method to process data for data warehousing projects.

**Diagram

Description automatically generated**

* **Extraction**: Extraction is the operation of extracting information from a source system for further use in a data warehouse environment. This is the first stage of the ETL process. It is often one of the most time-consuming tasks in the ETL. The source systems might be complicated and poorly documented, and thus determining which data needs to be extracted can be difficult. The data has to be extracted several times in a periodic manner to supply all changed data to the warehouse and keep it up to date.
* **Cleansing:** The cleansing stage is crucial in a data warehouse technique because it is supposed to improve data quality. The primary data cleansing features found in ETL tools are rectification and homogenization. They use specific dictionaries to rectify typing mistakes and to recognize synonyms, as well as rule-based cleansing to enforce domain-specific rules and defines appropriate associations between values.
* **Transformation**: Transformation is the core of the reconciliation phase. It converts records from its operational source format into a particular data warehouse format. If we implement a three-layer architecture, this phase outputs our reconciled data layer.
* **Load:** The **Load** is the process of writing the data into the target database. During the load step, it is necessary to ensure that the load is performed correctly and with as little resources as possible. Loading can be carried in two ways:  
  + **Refresh:** Data Warehouse data is completely rewritten. This means that older file is replaced. Refresh is usually used in combination with static extraction to populate a data warehouse initially.
* **Update:** Only those changes applied to source information are added to the Data Warehouse. An update is typically carried out without deleting or modifying pre-existing data. This method is used in combination with incremental extraction to update data warehouses regularly.

The whole ETL process for this project is done in Microsoft Excel.

**ANALYSIS ON DATASET**

In this analysis of COVID-19 dataset, I have collected the dataset or can raw data from 3 different – different sites. And then examine the data and process for cleaning. Cleaning like removing the unnecessary rows and column and the NA values that are present in the dataset by the help of filter and sort functions based upon the requirement of the analysis.

The analysis is divided into 6 objectives with its sub-objectives. The six main objective and their sub objectives are explained and visualized with their results below. There are 6 different dashboard which give summary of the objectives

**LET’S START WITH MAIN DASHBOARD**

In this dashboard I had attached the 6 different links to move to the observations dashboard and give a short summary about how many observations and sub observation

are there in my project.

Graphical user interface, application

Description automatically generated

**Observation 1**

In my 1st observation I had taken the country data and do the analysis that which country is having how many cumulative active cases and the cumulative death cases.

As the country data set is so long so to summarize that I had used the pivot table and the slicers.

And with the help of the slicer I can get the country name and how many active cases and death cases are there.

Graphical user interface, application, Excel

Description automatically generated

**Observation 1.2**

In my 2nd observation I had taken the state data set and do the analysis that which state is having how many active cases and the death cases.

Similarly, the state data set is so long so to summarize that I had used the pivot table and the slicers.

And with the help of the slicer, I can get the state name and how many active cases and death cases are there.

Graphical user interface, application

Description automatically generated

**Observation 1.3**

In this observation I had chosen the state data set and get the active, death and discharge ratio of the states with the help of pivot chart and the slicer.

Graphical user interface, application, table, PowerPoint

Description automatically generated

**Observation 2**

In my second observation I had taken the country data set and see which country is have most no of active cases and for that I had just drag the data and do some sorting and we are ready to go and use the graph and we found that united states of America, Brazil and India has most no of active cases in the month of November.

Graphical user interface, application

Description automatically generated

**Observation 2.1**

In this observation I had done the analysis and find 10 active cases over 1 lakh of population and for that analysis I had used the pivot table and used the filter and find the top 10 active case and draw the graph accordingly and for the graph I had used the line graph. And after the analysis from the top 10 found that **MONTENEGRO** is the country which has most no of active cases around 24k.

Graphical user interface, application, table, Excel

Description automatically generated

**Observation 2.2**

In my this observation I had taken the country data set and see which country is have most no of death cases and for that I had just drag the data and do some sorting and we are ready to go and use the graph and we found that same united states of America, Brazil and India has most no of death cases also in the month of November.

Graphical user interface, application

Description automatically generated

**Observation 2.3**

In this observation I had done the analysis and find 10 death cases over 1 lakh of population and for that analysis I had used the pivot table and used the filter and find the top 10 death case and draw the graph accordingly and for the graph I had used the line graph. And after the analysis from the top 10 found that **PERU** is the country which has most no of active cases around 24k.

Graphical user interface

Description automatically generated

**Observation 3**

In my this observation I had taken the state data set and see which state is have most no of death cases and for that I had just used the pivot table and chart and we found that **KERALA**, **MAHARASHTRA** and **TAMIL NADU** has most no of death cases also in the month of November.

Graphical user interface, application, table, Excel

Description automatically generated

**Observation 3.1**

In this analysis I had looked top 10 most affected states of India which had most no of active cases due to covid-19.

Graphical user interface, application

Description automatically generated

**Observation 4**

In this analysis we had looked around the top 10 active ratios of the state and for that analysis I had used the pivot table and the pie chart. And used some sorting.

**MIZORAM** has 36% of active ratio.

Graphical user interface, application, table, Excel

Description automatically generated

**Observation 4.1**

In this analysis we had looked around the top 7 discharge ratios of the state and for that analysis I had used the pivot table and the pivot chart. And used some sorting.

And found that **DADRA AND NAGAR HAVELI AND DAMAN AND DIIU** has most number of discharge ratio.

Graphical user interface, application

Description automatically generated

**Observation 4.2**

In this analysis we had looked around the top 15 death ratios of the state and for that analysis I had used the pivot table and 3-D bar graph. And used some sorting here also.

And found that **PUNJAB** has most number of death cases in the month of November .

Graphical user interface, application

Description automatically generated

**Observation 5**

Let’s talk about the vaccination

In this analysis we had taken the country data set of **WORLD HEALTH ORGANIZATION** of the month of November.

In this observation had done the analysis on the top 20 countries which had taken maximum number of vaccination.

And surprisingly found that there are six countries that had used 8 different-different type of vaccinations.

Graphical user interface

Description automatically generated

**Observation 5.1**

In this observation found which state has taken highest number of first dose and for that I had used the pivot table and pivot chart.

Graphical user interface

Description automatically generated

**Observation 5.2**

In this observation found which state has taken highest number of second dose and for that I had used the pivot table and pivot chart.

Graphical user interface

Description automatically generated

**Observation 5.3**

In this analysis we had taken top 10 most vaccinated countries, and to find that I had used add operator means adding the first dose and second doses of the countries and find the total. Just find 1 country fully vaccinated data and then used the fill handler and it automatically fill rest of the countries data.

Similarly, I had found the bottom 10 countries which had taken less no of both the vaccinations.

**China and India** are the most vaccinated countries taken 46% and 22% of both the vaccination respectively.

**Pitcairn island and Burundi** is the 2 bottom countries which had taken a smaller number of both doses.

And this analysis is done using pivot table, pivot graphs and some sorting.

Graphical user interface, chart, application

Description automatically generated

Chart, line chart

Description automatically generated

**Observation 5.3**

Here we are comparing with states means top 10 and bottom 10 most vaccinated and less vaccinated states of India.

After the analysis, found that **MADHAYA PRADESH** is the most vaccinated state of India and

**Sikkim and Lakshadweep** are the least vaccinated state of india.

Graphical user interface, chart, application, table, Excel

Description automatically generated

Chart, pie chart

Description automatically generated

**Observation 6.1**

Let us talk about the unemployment that is caused by covid -19

In this analysis we had observe that in May 2021 there is unemployment in both the rural and the urban area.

Graphical user interface, application

Description automatically generated

**Observation 6.2**

In this observation we had done the analysis on the top 5 month which has highest amount of unemployment in India.

And we found that mostly in the month of May and June there is most of the unemployment.

Graphical user interface, application, table

Description automatically generated

**LIST OF ANALYSIS**

**OBSERVATION 1**

1. WHICH CUNTRY HAS HOW MANY CUMULATIVE ACTIVE AND DEATH CASES
2. WHICH STATE HAS HOW MANY ACTIVE AND DEATH CASES.
3. ACTIVE VS DISCHARGE VS DEATH RATTIO OF STATES.

**OBSERVATION 2**

1. MOST EFFECTED COUNTRIES
2. TOP 10 ACTIVE CASES OVER THE 1 LAKH POPULATION IN THE COUNTRY DATASET.
3. MOST EFFECTED COUNTRY WITH MORE NO OF DEATH CASES
4. TOP 10 DEATH CASES OVER THE 1 LAKH POPULATION IN THE COUNTRY DATASET.

**OBSERVATION 3**

1. MOST EFFECTED STATE OF INDIA
2. TOP 10 STATE HAVE MORE NO OF ACTIVE CASES.

**OBSERVATION 4**

1. TOP 10 STATES HAVE THE MOST NO OF ACTIVE RATIO.

2.TOP 7 STATES HAVE MOST NO OF DISCHARGE RATIO.

1. TOP 15 STATE HAVE MORE NUMBER OF DEATH CASES.

**OBSERVATION 5**

1. NO OF VACCINATION TAKEN BY COUNTRIES
2. WHICH STATE TAKEN HIGHEST NO OF FIRST DOSE AND THE SECOND DOSE.
3. TOP 10 AND BOTTOM 10 COUNTRIES WITH HIGHEST NO OF BOTH VACCINATION AND THE LEAST NO OF BOTH VACCINATION.
4. TOP 10 AND BOTTOM 10 STATES WITH HIGHEST AND LEAST NO OF BOTH VACCINATION DOSES.

**OBSERVATION 6**

1. UNMEMPLOYMENT IN RURAL AND URBAN AREA OF INDIA.
2. TOP 5 MONTHS HWHICH HAS HIGHEST NO OF UNEMPLOYMENT IN INDIA.

**BIBLIOGRAPHY**

<https://www.mygov.in/covid-19/>

<https://unemploymentinindia.cmie.com/>

<https://www.worldometers.info/coronavirus/>

[**https://my.clevelandclinic.org/health/diseases/21214-coronavirus-covid-19**](https://my.clevelandclinic.org/health/diseases/21214-coronavirus-covid-19)