TEAM 17: PROJECT REPORT

COVID 19 STATISTICS : A CROSS COUNTRY COMPARISON

1.1 BACKGROUND:

The COVID-19 pandemic has deeply impacted on global health systems, economies, and everyday life since its emergence in 2019. Reliable and regularly updated data have been essential for tracking the virus's spread, severity, and recovery trends worldwide. Publicly available statistics have enabled comparisons of country-level responses and guided critical healthcare and policy decisions.

1.2 DATA COLLECTION:

• COVID-19 data was gathered by scraping the Worldometers website https://www.worldometers.info/coronavirus/ using rvest package in R.
• The primary data table includes country-wise statistics on cases, deaths, recoveries, active cases, and tests.

1.3 DATA DESCRIPTION:

Variable Name	Description
country_other	Country name
total_cases	Total covid-19 Cases
new_cases	New Cases
total_deaths	Deaths due to covid-19
new_deaths	New Deaths reported
total_recovered	Total Recovered cases from covid
active_cases	Total Active cases of covid
serious_critical	Critical cases
total_cases_per_million	Cases per Million of population
deaths_per_million	Deaths per Million of population
total_tests	Tests conducted to detect covid
tests_per_million	Tests per Million of population
population	Population size of the country

2. OBJECTIVES:

• To identify the most and least affected countries in terms of total cases, deaths, active cases and per-capita rates
• To assess the variability of case fatality and recovery rates across countries
• To examine the relationship between testing rates and fatality or recovery outcomes
• To investigate whether population size influences the number of cases and deaths reported
• To identify potential underreporting patterns based on inconsistencies between testing, case counts, and death rates

3. METHODOLOGY

3.1 DATA CLEANING

- The COVID-19 dataset was scraped from Worldometer, extracting country-level data on cases, deaths, recoveries, and tests.
- Column names were standardized by converting to lowercase and replacing spaces and special characters with underscores. Whitespace was trimmed from all data points.
- Numeric columns were cleaned by removing formatting symbols like commas and plus signs, and data types were converted appropriately from character to numeric type.

- Missing values and non-country entries were handled appropriately to prepare the data for accurate analysis.

3.2 EXPLORATORY DATA ANALYSIS

Step	Description
Creation of Derived Metrics	Created CFR, Recovery Rate, Active Case Rate, and Tests per Case; grouped by population and testing intensity.
Computation of Descriptive Statistics	Calculated mean, median, standard deviation, and quartiles for key metrics.
Comparative Rankings	Ranked countries by total cases, deaths, CFR, recovered cases, and testing rates (total and per capita).
Category-wise Analysis	Categorized countries by population size (Small, Medium, Large, Very Large) and testing intensity (Low, Medium, High) using conditional thresholds, comparing outcome averages across categories.
Correlation Analysis	Computed correlation matrix to analyze relationships between major variables.
Outlier Detection	Detected outliers in total cases, deaths, CFR, and tests per million using the inter-quartile range (IQR) method.
Analysis of Missing Values	Conducted missing data analysis to assess the extent and distribution of incomplete records across variables.
Data Visualization	Generated visual summaries, ranking plots, and correlation heatmaps for interpretation.

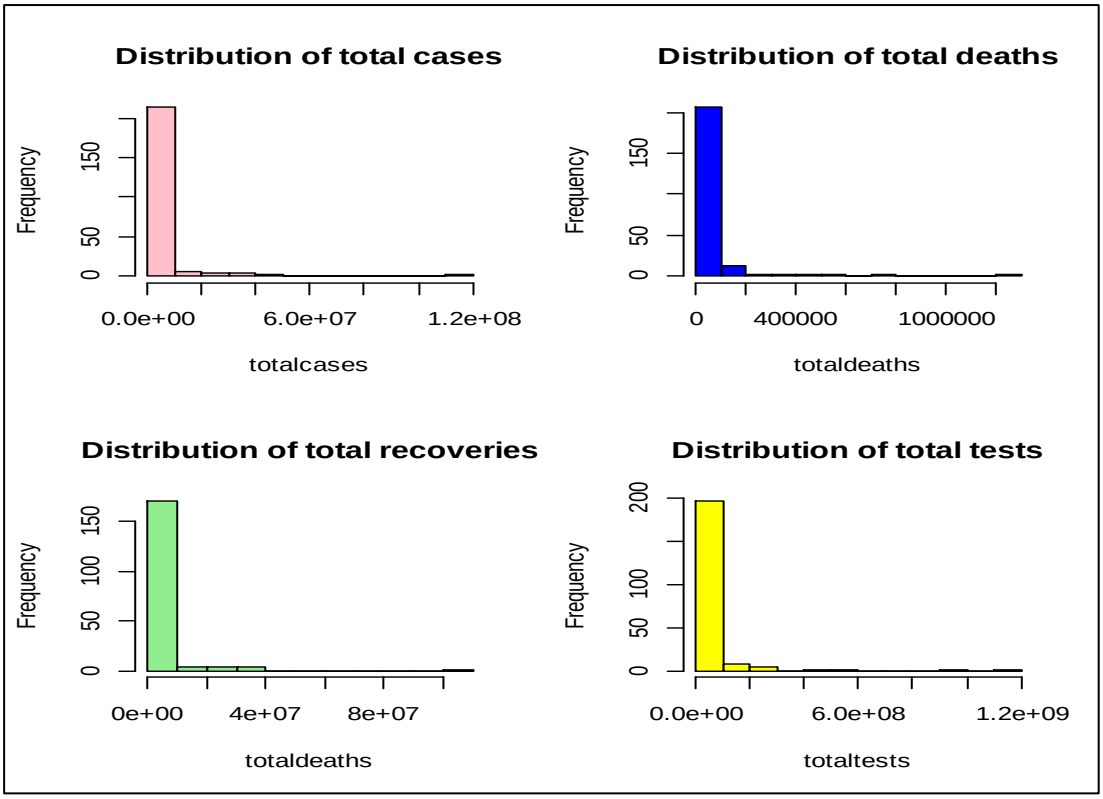
3.3 SHINY APP

Tab	Description
1: Descriptive Statistics	Constructed histograms and displayed key descriptive insights for each variable of the data. The user can choose the variable of interest.
2: Visualizing Variables Across Countries	Users can select a variable. The countries are ranked according to the selected variable and the top 10 countries are displayed along with their respective values in a horizontal bar plot and a table.
3: Correlation Analysis	Scatterplots are displayed to show the relationships between key variables (e.g., tests vs deaths).
4: Country-wise Data	The Country Explorer tab provides detailed statistics and ranks for any selected country.
5: About Section	The About section summarizes data descriptions and methodology.

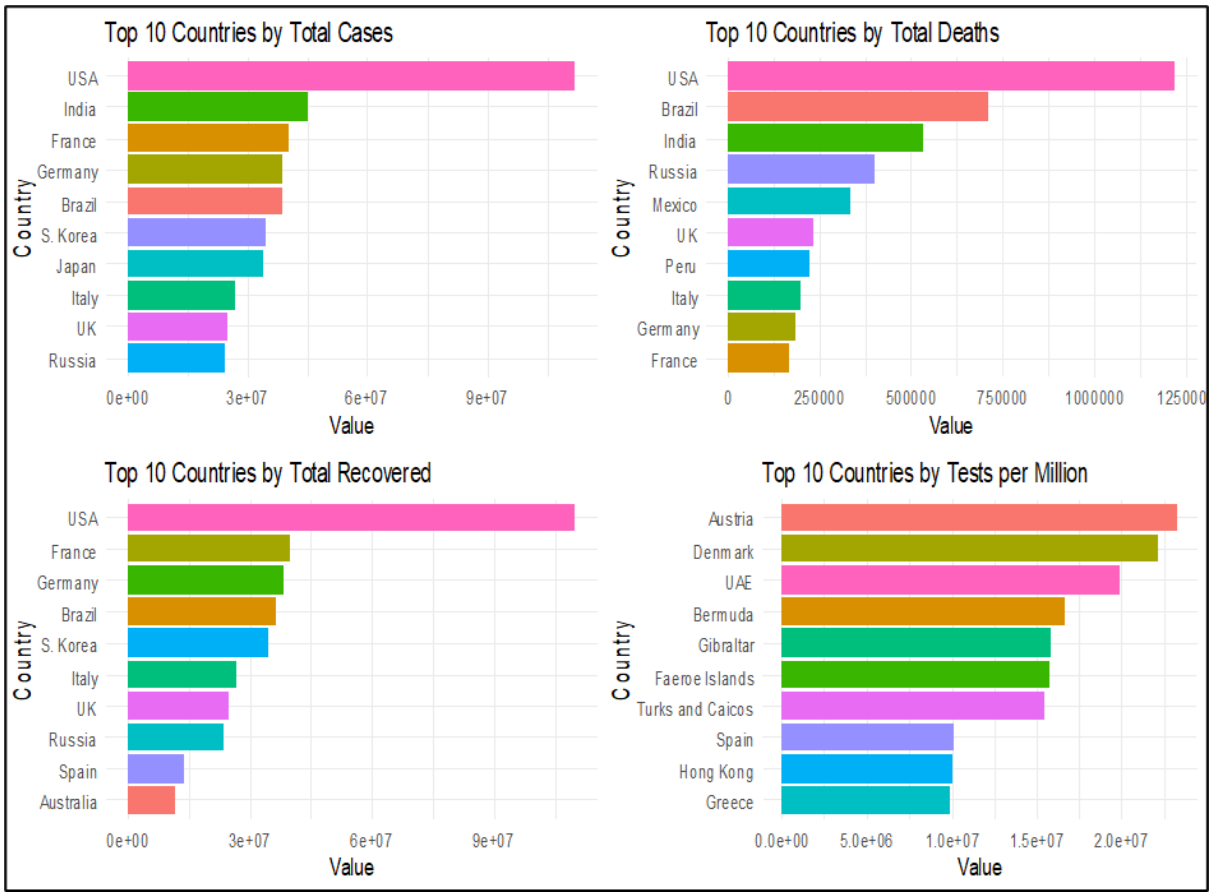
4. KEY FINDINGS:

4.1 DESCRIPTIVE INSIGHTS

The distribution of the key metrics is highly positive-skewed, with a few countries showing extremely high case counts while most have much lower numbers. The large gap between the mean and median highlights the influence of outliers and global disparities in testing and population size.



4.2 RANKING OF COUNTRIES BASED ON KEY METRICS



METRIC	**MOST AFFECTED COUNTRIES**	**COMMENTS**
TOTAL CASES	USA, India, France, Germany, Brazil, South Korea, Japan, Italy, UK, Russia	• Most populous and/or highly urbanized, early hit, high mobility. • The USA's case count far exceeds all other nations, highlighting its extensive population exposure during the pandemic. \High case counts generally correlate with population size, testing capacity, and early outbreak severity.
TOTAL DEATHS	USA, Brazil, India, Russia, Mexico, UK, Peru, Italy, Germany, France	• Overlaps with top cases; high total deaths also reflect severity, age structure, and possibly, reporting • High fatality counts are strongly associated with countries reporting high total cases. • The mortality burden was especially severe in the Americas (USA, Brazil, Mexico). • Differences in healthcare capacity, vaccination rates, and reporting accuracy likely contributed to variations between countries
TOTAL RECOVERED	USA, France, Germany, Brazil, S. Korea, Italy, UK, Russia, Spain, Australia	• Recovery figures closely align with total case counts, implying that most of the infected individuals recovered successfully. • Countries with efficient healthcare systems and robust vaccination programs likely achieved faster recovery rates and lower active case burdens.
TESTING PER MILLION	Austria, Denmark, UAE, Bermuda, Gibraltar, Faeroe Islands, Turks and Caicos, Spain, Hong Kong, Greece	Many European countries and small territories lead in testing rate
CASE FATALITY RATE(CFR)	MS Zaandam, Yemen, Western Sahara, Sudan, Syria, Somalia, Peru, Egypt, Mexico, Bosnia & Herzegovina	Mixture of small populations (outlier effect), conflict/low-resource settings (Yemen, Sudan, Syria), and places with under-testing

Comparative Insights :

Across all three variables – cases, deaths, and recoveries, the United States consistently dominates, reflecting both the magnitude of its outbreak and the scale of its medical response.

India and Brazil also appear prominently, illustrating pandemic intensity in high-population nations.

European countries such as France, Germany, and Italy show strong recovery ratios, possibly reflecting effective healthcare infrastructure and early vaccination strategies.

4.3 CATEGORY WISE ANALYSIS

(i) **By Population Category**

Population category	No. of countries	Average no. of cases	Average no. of deaths	Average CFR	Average tests per million
Large (10-100M)	79	4553194.038	36883.31646	1.823328	1382251.714
Medium (1-10M)	65	972642.4769	7226.846154	1.210342	2275113.111
Small (<1M)	71	59302.85915	312.0606061	0.712759	3405981.814
Very Large (>100M)	14	19829898.29	257610.2143	1.873237	607489.5

Comments :-

- Testing is highest per capita in small populated countries, and lowest in very large populated countries.
- CFR is lowest in small countries, highest in very large countries, possibly reflecting access to care and under-reporting in large populations.

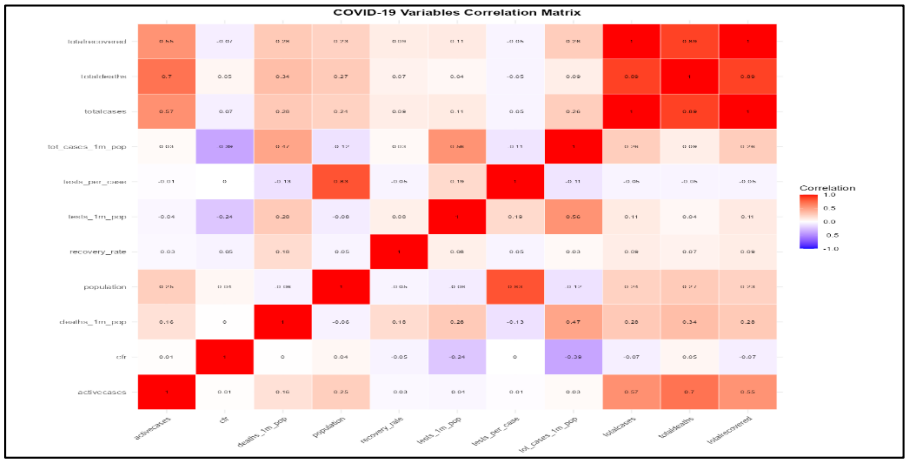
(ii) By Testing Intensity

Testing category	No. of countries	Average CFR	Average deaths per million	Average cases per million
High Testing	103	0.698075	1974.794118	363847.7476
Low Testing	42	2.415737	72.02380952	5275.547619
Medium Testing	68	1.546147	1126.147059	89610.32353

Comments :-

- Higher testing is associated with lower CFR and higher detected cases per million, reflecting greater general awareness regarding covid 19.
- Low-testing countries report fewer cases per capita, but a much higher fatality among those detected, suggesting significant under-testing and under-reporting.

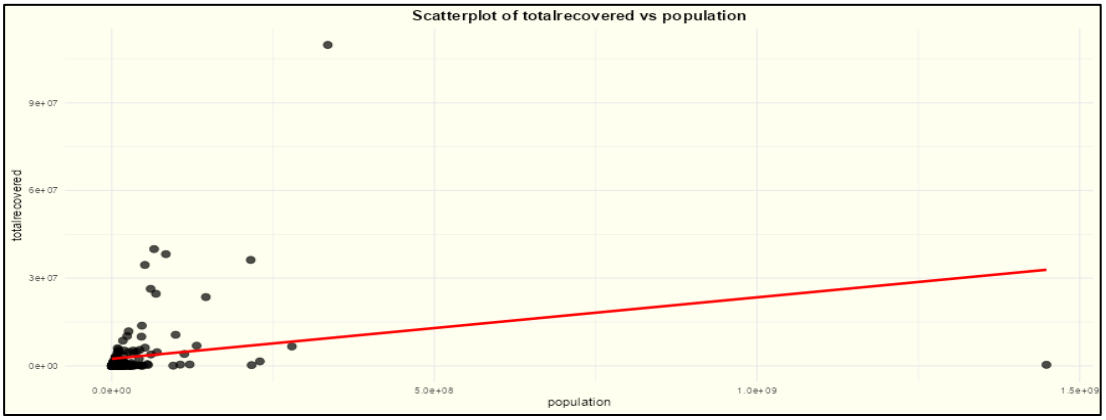
4.4 CORRELATION ANALYSIS:



- Total Cases ↔ Total Deaths: r = 0.89• Total Cases ↔ Total Recovered: r = 1.00 (almost perfect). Countries with more cases have more deaths and more recovered as expected.
- Population ↔ Tests per Case: r = 0.83 : Larger countries tend to have more tests per case, possibly due to proactive policies or higher detection in testing-rich environments.
- No major negative correlations, indicating most relationships are direct (more of one variable means more of another).

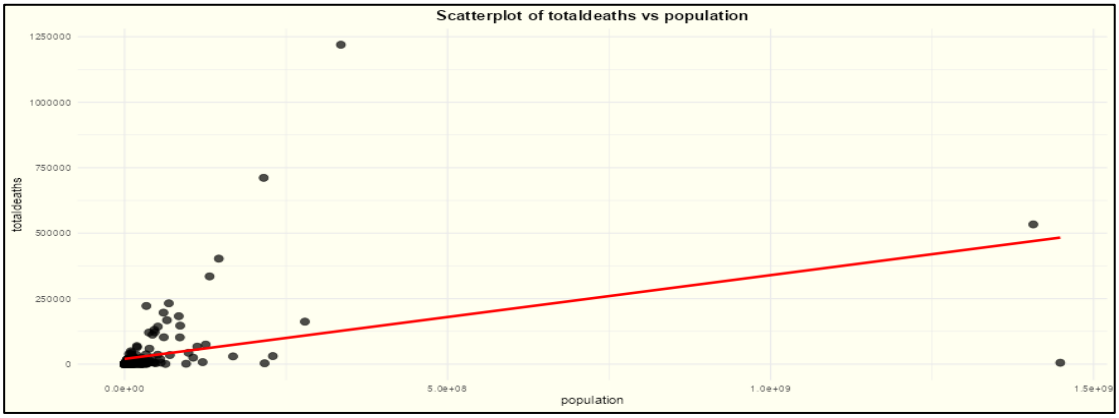
Specific Important cases for the study of correlation:

1. **Relationship between Population and Total Recoveries**



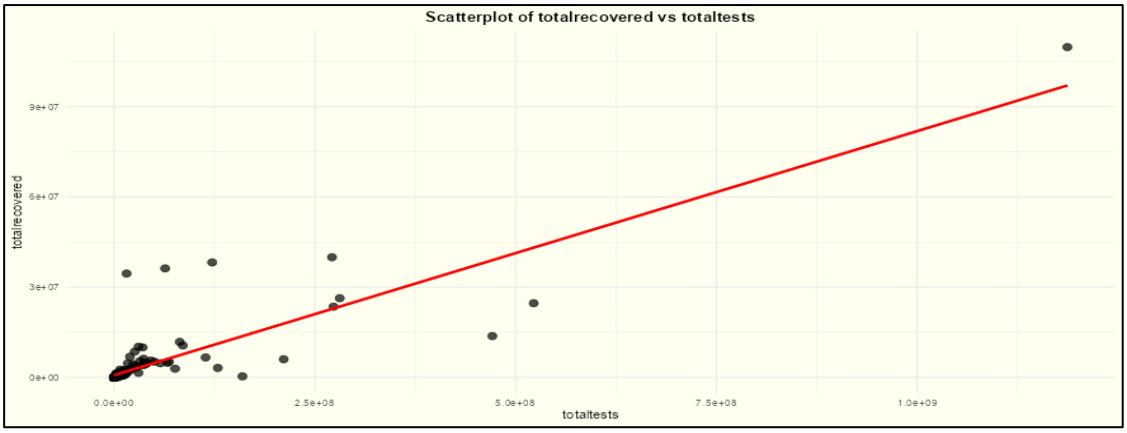
Interpretation:
Countries with larger populations tend to report higher total recoveries, but the relationship is not strong. This suggests that while population size contributes to recovery numbers, other factors—such as infection rate, healthcare capacity, and public health interventions—play a more dominant role. The correlation coefficient between population and total recoveries is **0.24**, indicating a weak positive relationship.

2. **Relationship between Population and Total Deaths**



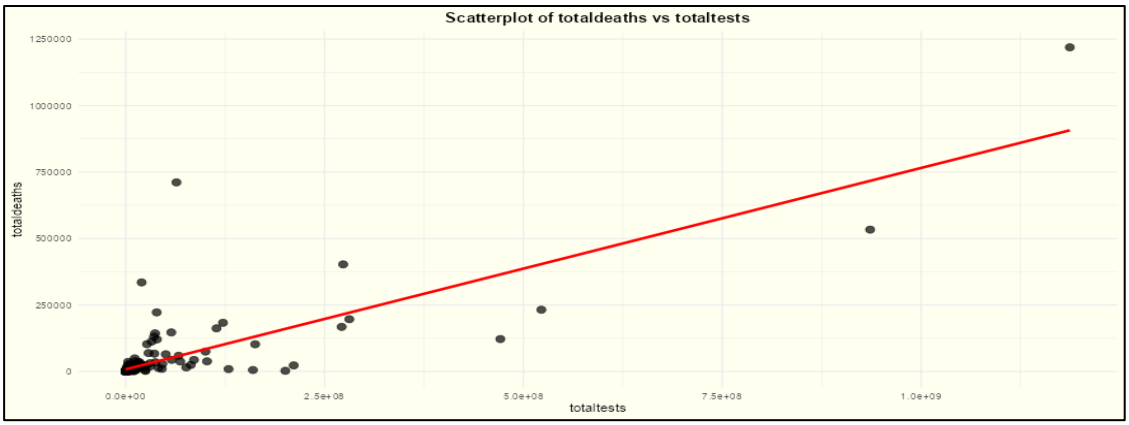
Interpretation:
Larger populations are somewhat associated with greater death counts, which is expected as more people potentially face exposure to infection. However, the correlation coefficient between population and total deaths is **0.40**, indicating a moderate positive linear relationship. The moderate correlation implies variability due to differences in healthcare systems, containment policies, and vaccination coverage across countries.

3. **Relationship between Total Tests and Total Recoveries**



Interpretation:
Higher testing volumes are strongly linked with higher reported recoveries. Increased testing enables early detection and treatment, contributing to better recovery outcomes. This finding underscores the effectiveness of extensive testing campaigns in controlling disease spread and improving recovery rates. The correlation coefficient between total tests and total recoveries is **0.86**, representing a strong positive linear relationship.

4. **Relationship between Total Tests and Total Deaths**



Interpretation:
Countries conducting more tests also tend to record more deaths, likely because higher testing capacity allows for more accurate case and fatality detection. This relationship reflects reporting accuracy rather than a direct causal link between testing and mortality. The correlation coefficient between total tests and total deaths is **0.80**, showing another strong positive linear relationship.

Overall Insight:
Testing activity shows a much stronger relationship with both recoveries and deaths compared to population size. This emphasizes the critical role of widespread testing in identifying and managing COVID-19 cases effectively. These scatterplots collectively illustrate that testing capacity is a stronger predictor of reported COVID-19 outcomes than population size, highlighting the importance of surveillance and diagnostic infrastructure during global health crises.

4.5 DETECTION OF OUTLIER COUNTRIES:

- Outliers span the largest and hardest-hit countries (USA, India, France, Germany, Brazil, South Korea, Japan, Italy, UK, Russia, Turkey, Spain).
- There are outliers on total cases, deaths, but not on CFR or testing intensity, except forcertain European microstates on certain axes.
- Some small countries are testing outliers (high testing per capita).

5. LIMITATIONS, ETHICS AND REFERENCE

Limitations:

- Data may be incomplete or inconsistent across countries.
- Analysis reflects only data available at the time of download.
- Simple visualizations; results depend on reporting quality.

Ethics:

- Source credited; no individual details or private info.

References:

- Worldometers (https://www.worldometers.info/coronavirus/)
- R tidyverse, dplyr, rvest, shiny documentation
- R documentation(https://cran.r-project.org)

6. REPRODUCIBILTY NOTES:

6.1 Required Softwares and packages

- R(version 4.1.0 or later recommended)
- Rstudio(optional) for convenient interface
- The following R packages must be installed:l
  - Tidyverse
  - Dplyr
  - Rvest
  - Shiny
  - Shinythemes
  - Ggplot2

6.2 Running the app:

- Open R or RStudio and set the working directory to the location containing the app folder.
- Lauch the Shiny app by running
  shiny::runApp("app")
- The app should open in your web browser presenting interactive control, plots and the "About" tab.

6.3 Data scraping and cleaning

- The app automatically scrapes the latest COVID-19 data from Worldometers each time it is run, or uses the provided dataset if implemented.
- All data cleaning and processing steps are self contained within the R scripts-no external manual processing is required.
- If the data source website changes format, minor adjustments to the code may be needed.

TEAM MEMBERS

Isha Goel (251080068)	Mridul Garg (240662)
Sumedha Banerjee (251080103)	Suropriya Chakrabarty (251080105)