“Visualization of Historic Covid Data and its impacts using R Studio”

Team Members:

1. **Preethi Uppuluri**
2. **Mohan Chandra Rangu**

Instructor: Hemant Purohit, PhD

CIS 5400 Final Project Report

Table Of Contents:

Chapter-1

Introduction:

The COVID-19 pandemic has impacted the world in a significant way, affecting millions of lives across the globe. As the pandemic continues, it is important to gain a better understanding of how different countries and states have been affected by the virus. The objective of this project is to simplify the complex data on the pandemic and create a visualization that can help people understand the impact of COVID-19 on different states in terms of death rate, recovered rate, and active rate etc. By presenting this information concisely, we aim to increase public knowledge and awareness of the pandemic and encourage people to take appropriate actions to protect themselves and others. Through this project, we hope to contribute to the ongoing efforts to combat COVID-19 and mitigate its impact on our communities.

In addition to presenting data on death rate, recovered rate, and active rate, this project aims to provide a comprehensive overview of the COVID-19 pandemic, including information on the spread of the virus, the measures taken to contain it, and the impact on various aspects of society. This information can help people understand the challenges posed by the pandemic and the efforts being made to address them.

By visualizing the data, we hope to make it easier for people to comprehend the scale and impact of the pandemic, as well as to identify trends and patterns in the data. This can help individuals and communities make informed decisions about how to respond to the pandemic and protect themselves and others.

Overall, the objective of this project is to provide a valuable resource for anyone seeking to understand the COVID-19 pandemic and its impact on the world, and to contribute to the collective effort to overcome this global health crisis.

Chapter-2

Motivation

The COVID-19 pandemic has caused immense disruption to our daily lives, and has had a profound impact on global health, economy, and social well-being. The collection and analysis of data has been a critical tool in understanding the spread and impact of the virus, as well as informing policy decisions to mitigate its effects.

However, as the volume and complexity of data on the pandemic has grown, it has become increasingly difficult for individuals and communities to make sense of the information and identify patterns and trends. This is particularly true for those who are not familiar with data analysis or who lack the time and resources to comb through large datasets.

That's where data visualization comes in. By converting complex data into visual representations, we can make it easier for people to grasp the key insights and trends in the data, without the need for specialized expertise or extensive time investment. Visualization allows us to communicate information in a more intuitive and engaging way, enabling people to make better-informed decisions and take appropriate action.

Furthermore, visualization can help us to identify disparities and inequities in the impact of the pandemic across different regions, demographics, and social groups. By highlighting these disparities, we can work towards developing more targeted and effective interventions to address the needs of those most affected by the pandemic.

In conclusion, data visualization is a powerful tool that can help us to better understand the impact of the COVID-19 pandemic and make more informed decisions about how to respond to it. By making the data more accessible and understandable, we can empower individuals and communities to take action and work towards a more equitable and resilient future.

Chapter-3

Project Roadmap

We followed the KDD (Knowledge Discovery in Database) to achieve our goals in this project.

The KDD process(Figure-3.1)involves several steps, including data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge representation.

In R, there are many packages and tools available to support each step of the KDD process. For example, the "tidyverse" package provides tools for data cleaning and transformation, while the "dplyr" package can be used for data selection and filtering. The "caret" package provides a range of algorithms for data mining and predictive modelling, and the "ggplot2" package can be used for visualizing patterns and relationships in the data.

To implement the KDD process in R, it is important to have a solid understanding of each step and the tools available to support it. It is also important to have a clear understanding of the research question or problem being addressed, as this will guide the selection of appropriate data sources, pre-processing steps, and analysis techniques.

Overall, R is a powerful tool for implementing the KDD process and gaining insights from large and complex datasets. With its robust set of packages and tools, R can help data analysts and researchers to effectively navigate the KDD process and generate meaningful insights from their data.

Data cleaning: The first step in the KDD process is to clean and preprocess the data. In your case, this would involve removing any missing or erroneous data points, as well as identifying any outliers that may be skewing your analysis. You may have also needed to convert the raw data into a more usable format, such as a spreadsheet or database table.

Data selection: The next step is to select the subset of data that is relevant to your research question. For example, you may have selected data on COVID-19 cases, deaths, recoveries, and other relevant variables for a particular region or country.

Data transformation: Once you have selected your data, the next step is to transform it into a format that is suitable for analysis. This may involve normalizing the data, converting categorical variables into numerical ones, and standardizing the units of measurement.

Data mining: With the transformed data, you can now start to perform more complex analysis techniques, such as calculating probabilities and correlations. This is the heart of the KDD process, where you use statistical methods and machine learning algorithms to extract patterns and relationships from the data.

Pattern evaluation: After you have mined the data, the next step is to evaluate the patterns and relationships you have identified. This may involve visualizing the data using charts or graphs, or performing hypothesis testing to confirm or reject your initial findings.

Knowledge representation: Finally, the KDD process concludes with the representation of the knowledge you have gained from the data. This may involve creating a report or presentation summarizing your findings or using the data to inform policy decisions or further research.

We used KDD process, cleaning and selecting the data, transforming it into a usable format, and using statistical methods to analyse the relationships between different rates. The final step of the KDD process, knowledge representation, involved creating visualizations of the data, summarizing our findings in a report or presentation, or using the data to inform public health interventions or policy decisions.

- **How we proceed with the help of KDD Process:**

1. The first move toward quite a while examination is ordinarily to gather information from different web sources. This could include utilizing web scratching methods to remove data from sites or utilizing APIs, which give admittance to information from various sources. Web scratching includes automatically extricating information from sites, which can be valuable for gathering information that isn't generally accessible in an organized organization. This information can emerge out of a wide assortment of sources, including online entertainment stages, news sites, government data sets, and logical distributions.
2. Once the information has been gathered, it is normally put away in an information stockpiling framework, like a data set or a record. Information bases are a typical decision for putting away a lot of information, as they give a helpful method for sorting out and recover information. Records, for example, CSV or JSON documents, are one more typical method for putting away information, as they are compact and can be effectively imparted to other people. When the information has been put away, it tends to be stacked into R Studio, which is a famous incorporated improvement climate (IDE) utilized for measurable processing and illustrations.With the data loaded into R Studio, the researcher can then view and analyze the data in various ways. This might involve cleaning and pre-processing the data, which could include removing duplicates, dealing with missing values, or transforming variables. Data cleaning is an important step in the data analysis process, as it helps to ensure that the data is accurate and reliable. This process can be time-consuming and labour-intensive, but it is essential for ensuring that the analysis is based on high-quality data.
3. After the information has been pre-handled, it is packed to its insignificant level to make it more obvious and work with. This could include summing up the information in different ways, like registering outline measurements or making perceptions. Rundown measurements, like mean, middle, and standard deviation, can give a fast outline of the information and assist with recognizing any exceptions or uncommon examples. Perceptions, for example, scatterplots, histograms, or boxplots, can be helpful for investigating the connections among factors and recognizing designs in the information.
4. One of the critical objectives of information examination is to recognize examples and connections in the information. To do this, probabilities and connections are determined for various situations. For instance, the scientist may be keen on working out the likelihood of a specific occasion happening, for example, the likelihood of a client making a buy or the likelihood of a stock cost expanding. They may likewise be keen on working out the relationship between's two factors in the information, for example,

the connection among's pay and training level or the relationship among's temperature and frozen yogurt deals.

5. Finally, the connections among probabilities and relationships are plotted to help envision and convey the aftereffects of the investigation. This could include making scatterplots, histograms, or different sorts of representations to show how various factors are connected with one another. These representations can be a significant instrument for conveying the aftereffects of the examination to other people, like partners or leaders. By introducing the information in an unmistakable and outwardly engaging manner, it is more straightforward for others to comprehend the consequences of the examination and pursue informed choices in view of the bits of knowledge acquired.

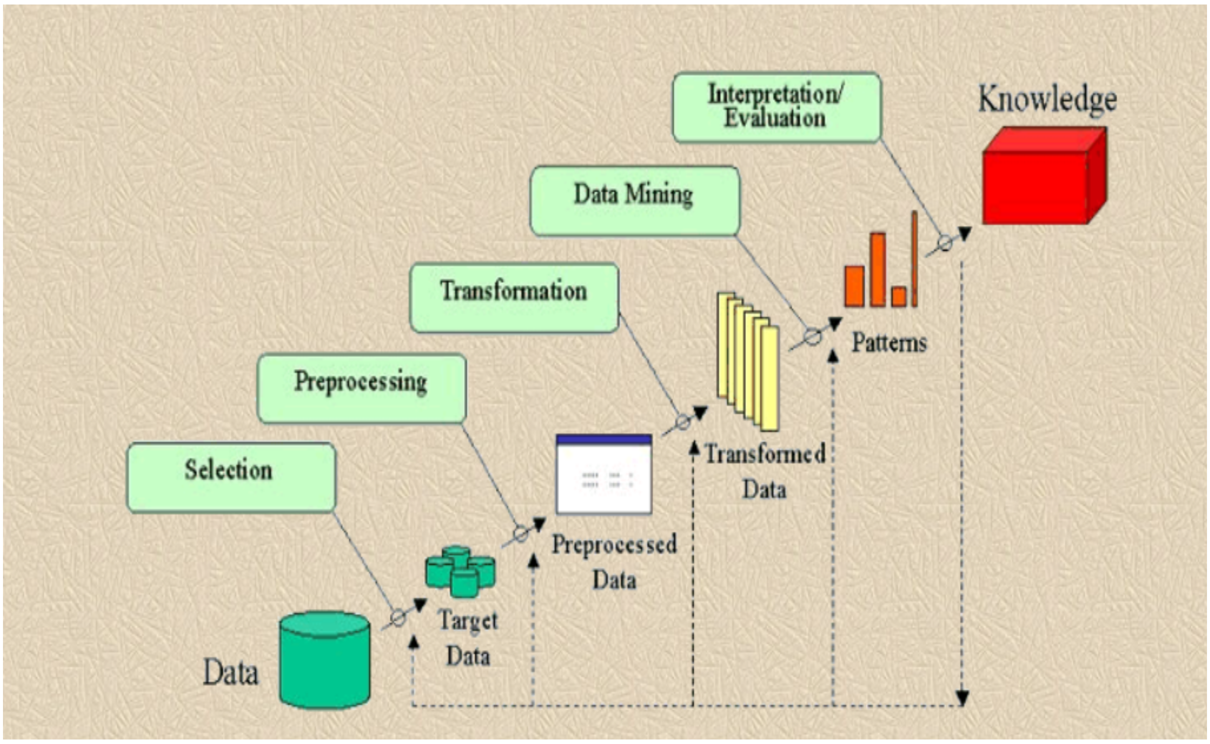


Figure-3.1

Chapter-4

RStudio:

R Studio is an integrated development environment (IDE) that is designed to facilitate statistical computing and graphics. It is a widely used tool in the field of data analysis, especially by researchers and data scientists who seek to analyze and visualize data in an efficient and user-friendly manner. R Studio offers an intuitive interface that allows users to work with the R programming language, which is extensively employed in statistical computing and modeling.

One of the key advantages of R Studio is its wide array of features and tools that are dedicated to working with data. These features include the ability to import and export data, manipulate data, visualize data, and create models based on the data. R Studio provides an integrated environment that allows users to access these features seamlessly and efficiently.

In addition to its features for data analysis, R Studio includes numerous tools that can be used for debugging and testing code. These features make it possible to ensure that the code is functioning properly and that the results obtained from data analysis are accurate and reliable.

Another benefit of R Studio is that it is open-source software, which means that anyone can use and modify the code. This open-source nature has led to a large and active community of users who continuously develop and improve the software. This has also led to the availability of numerous plugins and add-ons that can be used to extend the functionality of R Studio.

Overall, R Studio is a powerful tool for data analysis and visualization that is highly regarded in the research community. Its user-friendly interface, rich set of features, and open-source nature make it a popular choice for those who seek to work with data in an effective and efficient manner.

4.1.Importing the dataset Into R studio:

Importing datasets is an essential task for data analysis in R. R provides various ways to import datasets into the R environment. In this report, I will discuss the steps to import datasets in R using R Studio.

Step 1: Open R Studio and you can see the panel in Figure-4.1.1 and set the working directory.

Before importing any dataset, it is important to set the working directory in R Studio. The working directory is the folder where your R script and dataset are stored. To set the working directory, you can use the setwd() function in R. For example, if your dataset is stored in the "Data" folder on your desktop, you can set the working directory using the following command:

- **setwd("~/Desktop/Data")**

RStudio Panel:

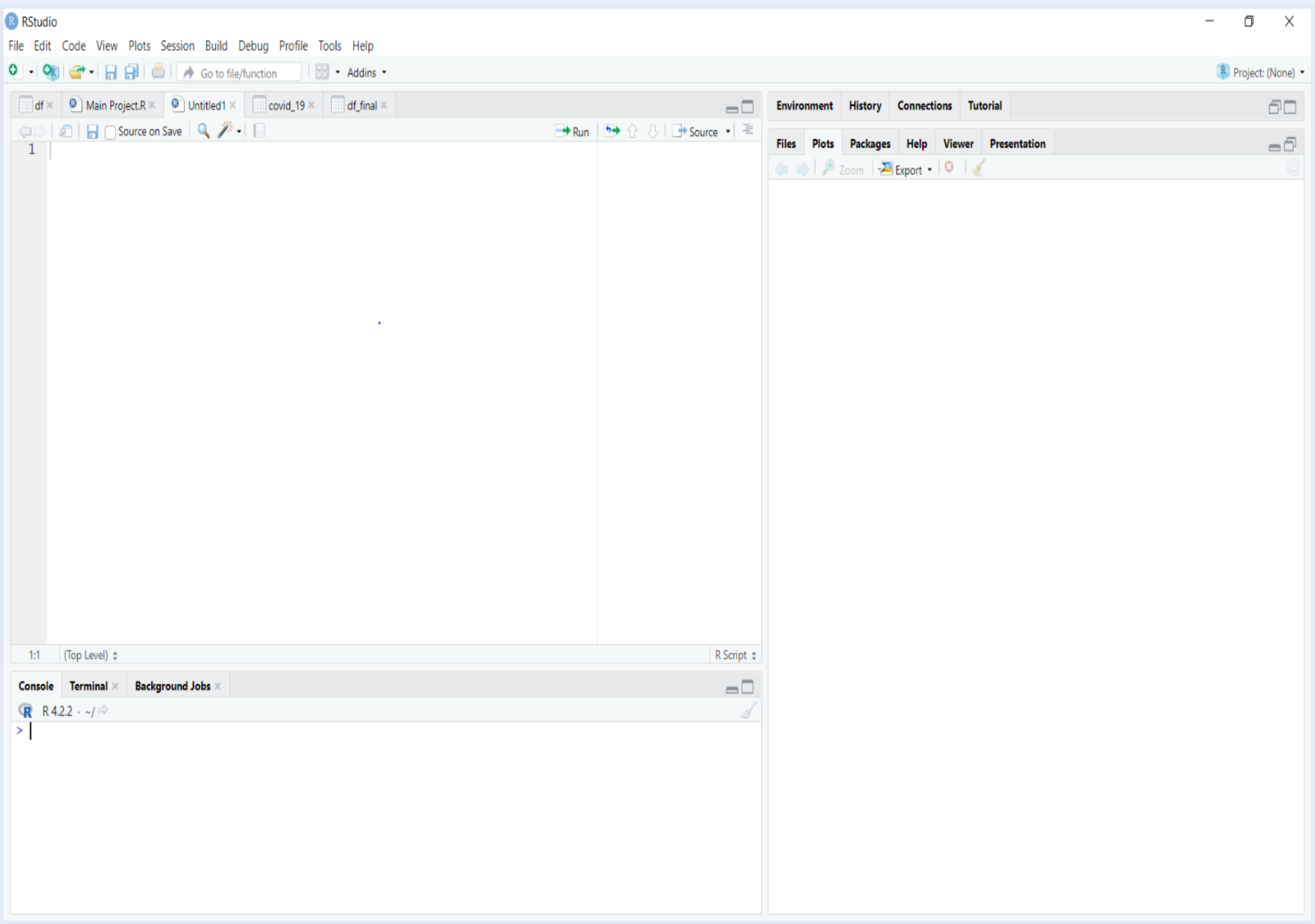


Figure-4.1.1

Step 2: Load the necessary packages:

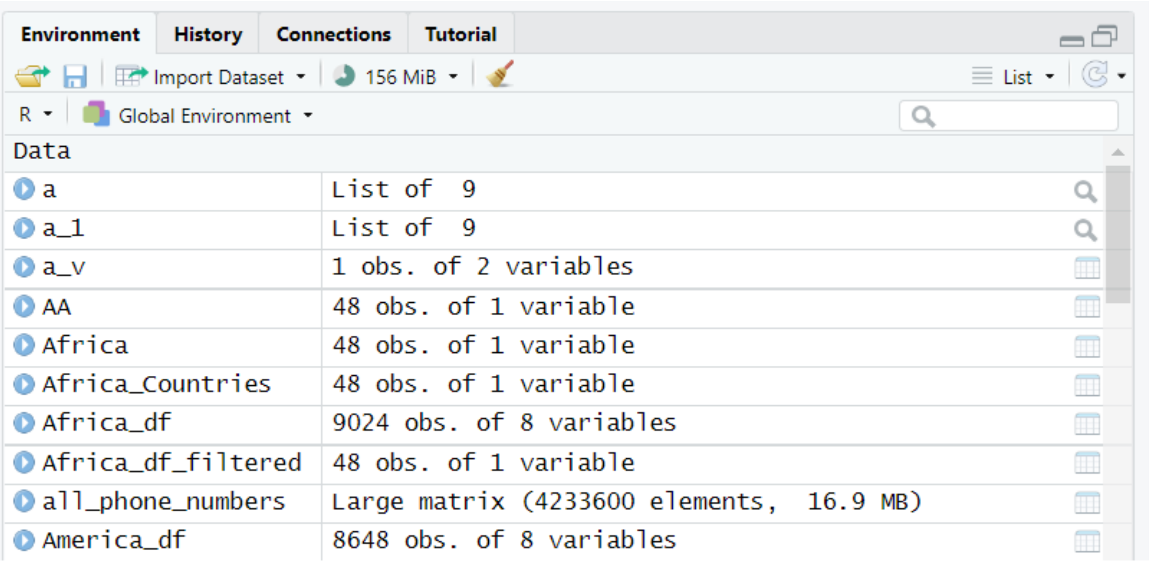
R provides various packages for importing different types of datasets. Before importing a dataset, it is important to load the necessary packages. For example, if you want to import a CSV file, you can load the "readr" package using the following command:

- **library(readr)**

Step 3: Import the dataset:

Once you have set the working directory and loaded the necessary packages, you can import the dataset using the appropriate function. There are different functions for importing different types of datasets.

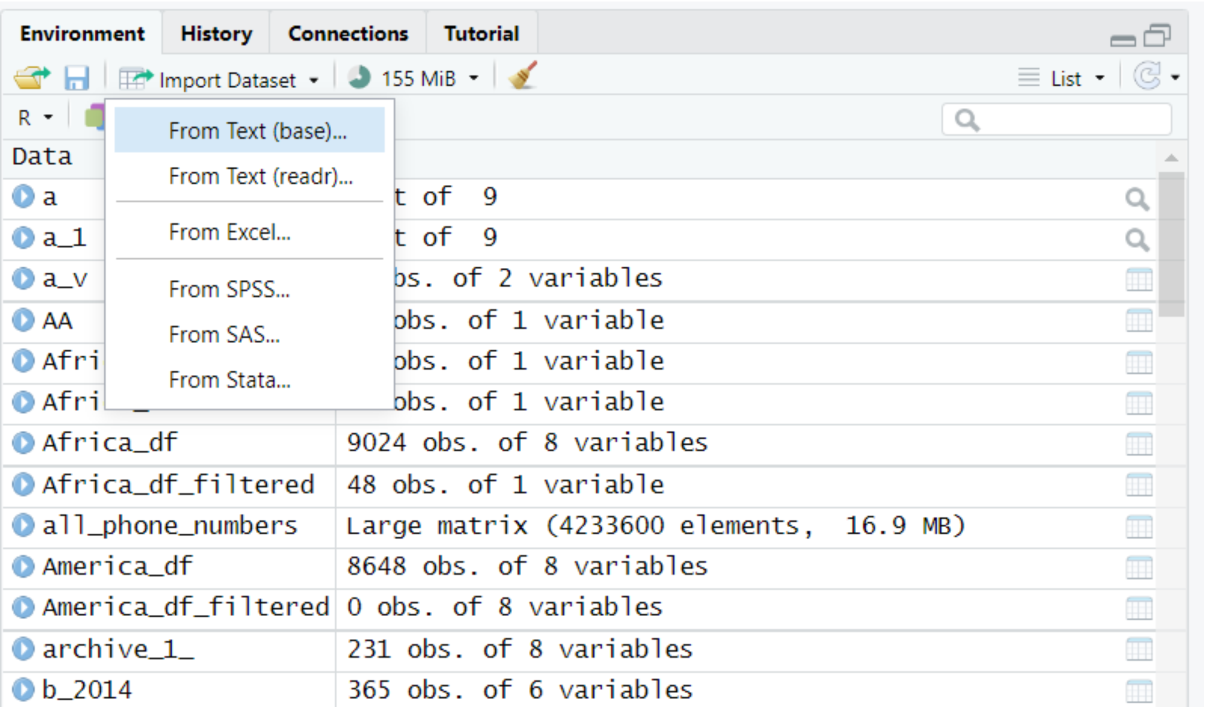
In R Studio, the right-side panel(Figure-4.1.2) displays various tabs, including the "Environment" tab which shows the datasets that have been loaded and their sizes.



Data	
a	List of 9
a_1	List of 9
a_v	1 obs. of 2 variables
AA	48 obs. of 1 variable
Africa	48 obs. of 1 variable
Africa_Countries	48 obs. of 1 variable
Africa_df	9024 obs. of 8 variables
Africa_df_filtered	48 obs. of 1 variable
all_phone_numbers	Large matrix (4233600 elements, 16.9 MB)
America_df	8648 obs. of 8 variables

Figure-4.1.2

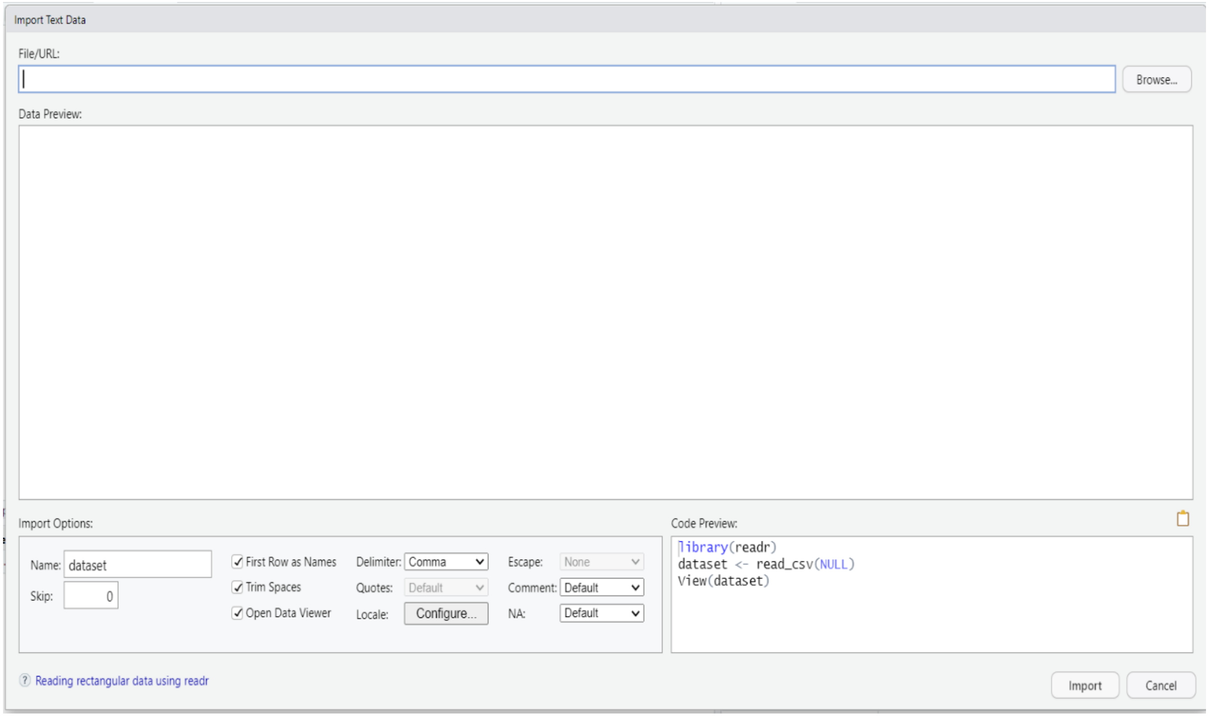
To import a dataset, we can select the "Import Dataset" option under the "Environment" tab and choose the "From Text (readr)" format which can be seen in Figure-4.1.3. Alternatively, we can use the "readr" package by loading it using the command "library(readr)".



Data	
a	t of 9
a_1	t of 9
a_v	bs. of 2 variables
AA	obs. of 1 variable
Afri	obs. of 1 variable
Afri_	obs. of 1 variable
Africa_df	9024 obs. of 8 variables
Africa_df_filtered	48 obs. of 1 variable
all_phone_numbers	Large matrix (4233600 elements, 16.9 MB)
America_df	8648 obs. of 8 variables
America_df_filtered	0 obs. of 8 variables
archive_1_	231 obs. of 8 variables
b_2014	365 obs. of 6 variables

Figure-4.1.3

After selecting the "From Text (readr)" option, we can browse in Import Test Data tab shown in Figure-4.1.4 and select our dataset and customize its name using the "Import Options" tab. Upon clicking the "Import" button, the data will be imported to R Studio.



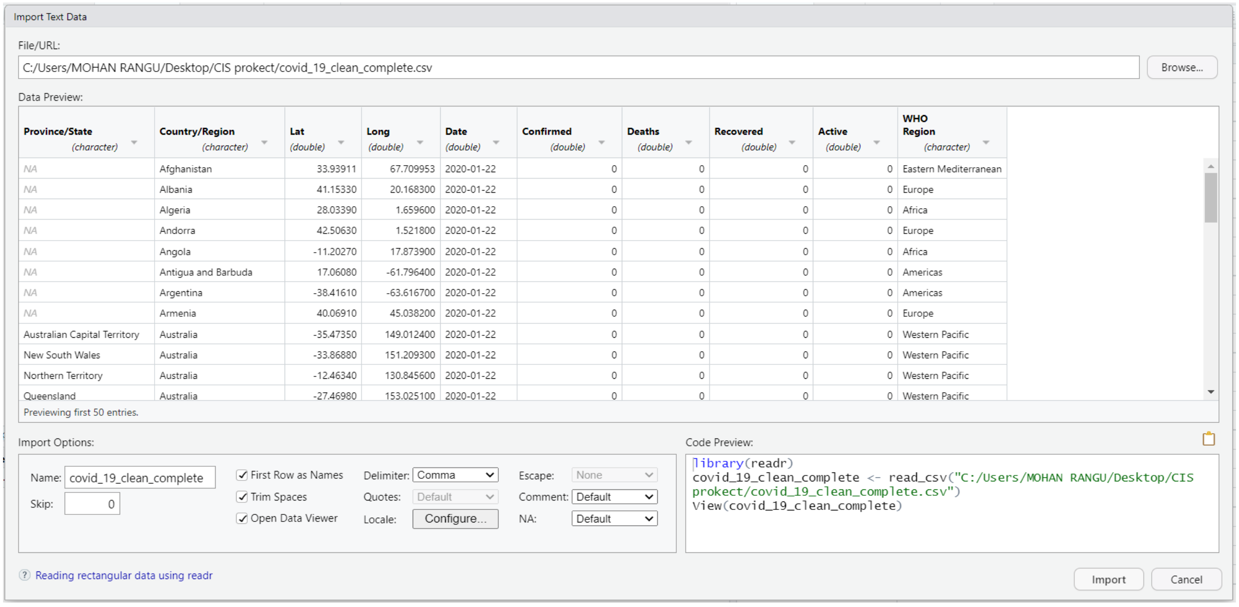
Name: dataset
Skip: 0
☑ First Row as Names
☑ Trim Spaces
☑ Open Data Viewer
Delimiter: Comma
Quotes: Default
Locale: Configure...
Escape: None
Comment: Default
NA: Default

```
library(readr)
dataset <- read_csv(NULL)
View(dataset)
```

Figure-4.1.4

Step 4: Explore the dataset

When we select the option to import a dataset in R Studio and browse for the file, a preview of the dataset is shown in the import window just like in the below Figure-4.1.5. After selecting the desired import options and clicking on the "Import" button, the selected dataset will be imported into R Studio.



File/URL: C:/Users/MOHAN RANGU/Desktop/CIS prokect/covid_19_clean_complete.csv

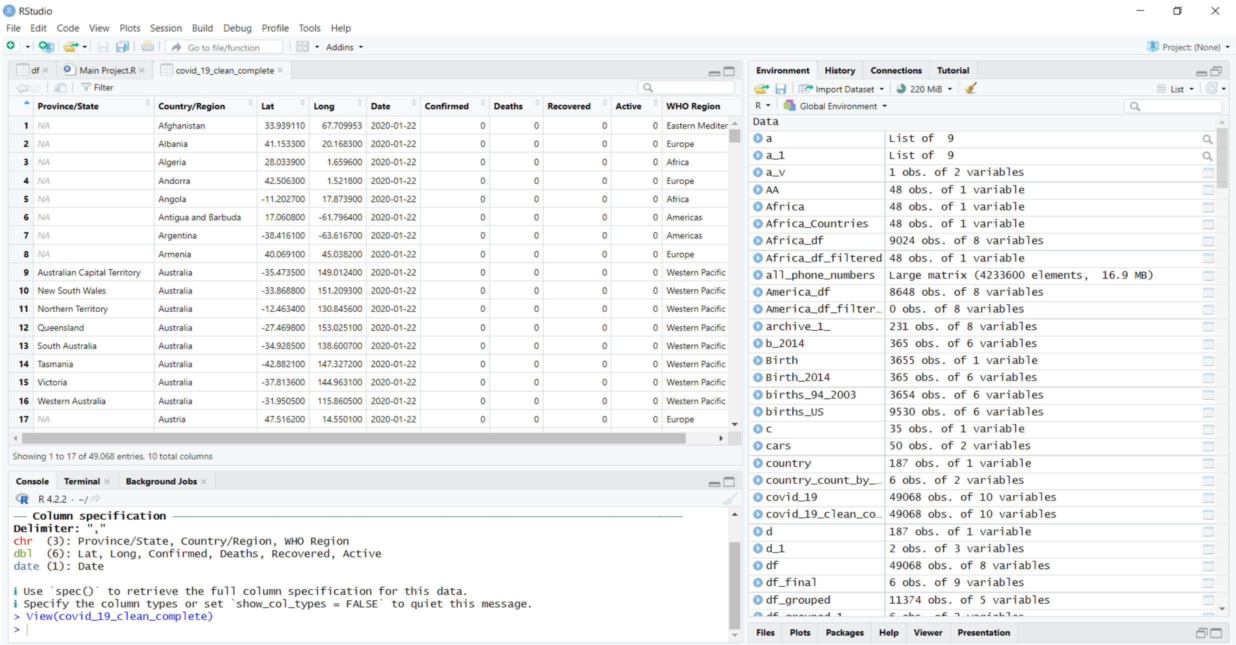
Province/State (character)	Country/Region (character)	Lat (double)	Long (double)	Date (double)	Confirmed (double)	Deaths (double)	Recovered (double)	Active (double)	WHO Region (character)
NA	Afghanistan	33.93911	67.709953	2020-01-22	0	0	0	0	Eastern Mediterranean
NA	Albania	41.15330	20.168300	2020-01-22	0	0	0	0	Europe
NA	Algeria	28.03390	1.659600	2020-01-22	0	0	0	0	Africa
NA	Andorra	42.50630	1.521800	2020-01-22	0	0	0	0	Europe
NA	Angola	-11.20270	17.873900	2020-01-22	0	0	0	0	Africa
NA	Antigua and Barbuda	17.06080	-61.796400	2020-01-22	0	0	0	0	Americas
NA	Argentina	-38.41610	-63.616700	2020-01-22	0	0	0	0	Americas
NA	Armenia	40.06910	45.038200	2020-01-22	0	0	0	0	Europe
Australian Capital Territory	Australia	-35.47350	149.012400	2020-01-22	0	0	0	0	Western Pacific
New South Wales	Australia	-33.86880	151.209300	2020-01-22	0	0	0	0	Western Pacific
Northern Territory	Australia	-12.46340	130.845600	2020-01-22	0	0	0	0	Western Pacific
Queensland	Australia	-27.46980	153.025100	2020-01-22	0	0	0	0	Western Pacific

Previewing first 50 entries.

Name: covid_19_clean_complete
Skip: 0
☑ First Row as Names
☑ Trim Spaces
☑ Open Data Viewer
Delimiter: Comma
Quotes: Default
Locale: Configure...
Escape: None
Comment: Default
NA: Default

```
library(readr)
covid_19_clean_complete <- read_csv("C:/Users/MOHAN RANGU/Desktop/CIS
prokect/covid_19_clean_complete.csv")
View(covid_19_clean_complete)
```

Figure-4.1.5



	Province/State	Country/Region	Lat	Long	Date	Confirmed	Deaths	Recovered	Active	WHO Region
1	NA	Afghanistan	33.939110	67.709953	2020-01-22	0	0	0	0	Eastern Mediter
2	NA	Albania	41.153300	20.168300	2020-01-22	0	0	0	0	Europe
3	NA	Algeria	28.033900	1.659600	2020-01-22	0	0	0	0	Africa
4	NA	Andorra	42.506300	1.521800	2020-01-22	0	0	0	0	Europe
5	NA	Angola	-11.202700	17.873900	2020-01-22	0	0	0	0	Africa
6	NA	Antigua and Barbuda	17.060800	-61.796400	2020-01-22	0	0	0	0	Americas
7	NA	Argentina	-38.416100	-63.616700	2020-01-22	0	0	0	0	Americas
8	NA	Armenia	40.069100	45.038200	2020-01-22	0	0	0	0	Europe
9	Australian Capital Territory	Australia	-35.473500	149.012400	2020-01-22	0	0	0	0	Western Pacific
10	New South Wales	Australia	-33.868800	151.209300	2020-01-22	0	0	0	0	Western Pacific
11	Northern Territory	Australia	-12.463400	130.845600	2020-01-22	0	0	0	0	Western Pacific
12	Queensland	Australia	-27.469800	153.025100	2020-01-22	0	0	0	0	Western Pacific
13	South Australia	Australia	-34.928500	138.600700	2020-01-22	0	0	0	0	Western Pacific
14	Tasmania	Australia	-42.882100	147.327200	2020-01-22	0	0	0	0	Western Pacific
15	Victoria	Australia	-37.813600	144.963100	2020-01-22	0	0	0	0	Western Pacific
16	Western Australia	Australia	-31.950500	115.860500	2020-01-22	0	0	0	0	Western Pacific
17	NA	Austria	47.516200	14.550100	2020-01-22	0	0	0	0	Europe

Showing 1 to 17 of 49,068 entries, 10 total columns

```
── Column specification ────────────────────────────────────────
Delimiter: ","
chr  (3): Province/State, Country/Region, WHO Region
dbl  (6): Lat, Long, Confirmed, Deaths, Recovered, Active
date (1): Date

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
> View(covid_19_clean_complete)
> 
```

Data

a	List of 9
a_1	List of 9
a_v	1 obs. of 2 variables
AA	48 obs. of 1 variable
Africa	48 obs. of 1 variable
Africa_Countries	48 obs. of 1 variable
Africa_df	9024 obs. of 8 variables
Africa_df_filtered	48 obs. of 1 variable
all_phone_numbers	Large matrix (4233600 elements, 16.9 MB)
America_df	8648 obs. of 8 variables
America_df_filter…	0 obs. of 8 variables
archive_1_	231 obs. of 8 variables
b_2014	365 obs. of 6 variables
Birth	3655 obs. of 1 variable
Birth_2014	365 obs. of 6 variables
births_94_2003	3654 obs. of 6 variables
births_US	9530 obs. of 6 variables
c	35 obs. of 1 variable
cars	50 obs. of 2 variables
country	187 obs. of 1 variable
country_count_by_…	6 obs. of 2 variables
covid_19	49068 obs. of 10 variables
covid_19_clean_co…	49068 obs. of 10 variables
d	187 obs. of 1 variable
d_1	2 obs. of 3 variables
df	49068 obs. of 8 variables
df_final	6 obs. of 9 variables
df_grouped	11374 obs. of 5 variables

Figure-4.1.6

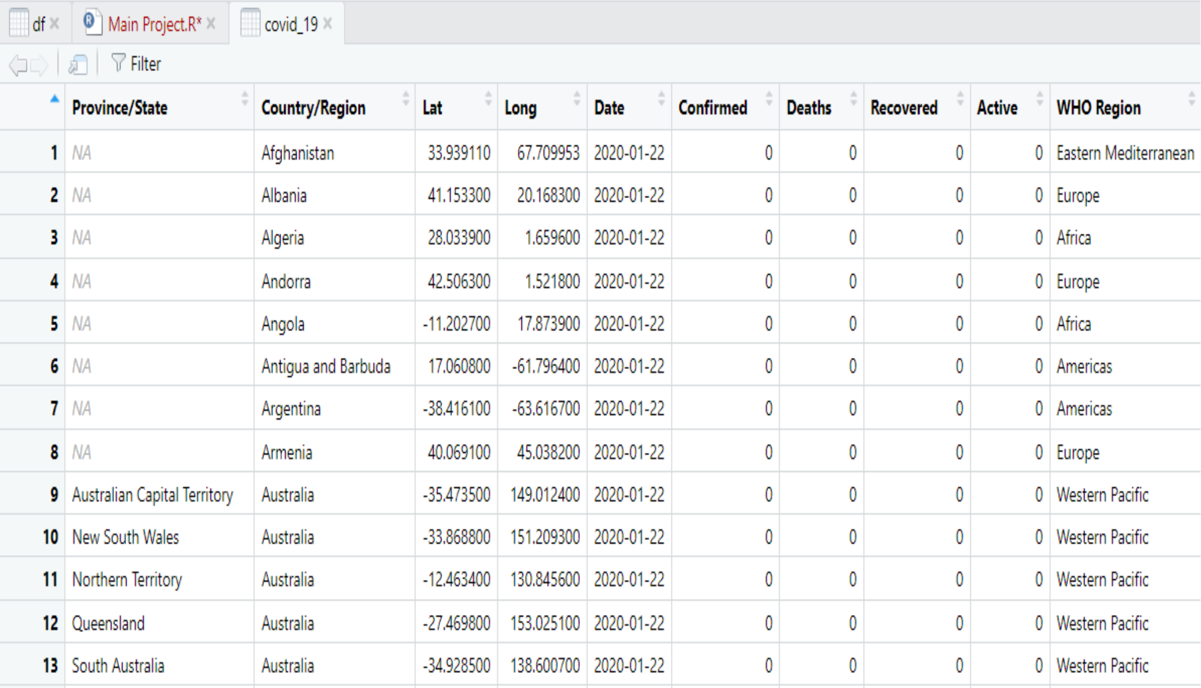
- The above Figure-4.1.6 shows how the data is imported and viewed in RStudio.

In the wake of bringing in the dataset, you can investigate it involving different capabilities in R. For instance, you can utilize the head() capability to see the initial not many lines of the dataset, the synopsis() capability to get an outline of the dataset, and the str() capability to get data about the design of the dataset

Chapter-5

Project Execution:

Figure-5.1 is the sample view of the data we have worked on. And I named this dataset as covid_19.



	Province/State	Country/Region	Lat	Long	Date	Confirmed	Deaths	Recovered	Active	WHO Region
1	NA	Afghanistan	33.939110	67.709953	2020-01-22	0	0	0	0	Eastern Mediterranean
2	NA	Albania	41.153300	20.168300	2020-01-22	0	0	0	0	Europe
3	NA	Algeria	28.033900	1.659600	2020-01-22	0	0	0	0	Africa
4	NA	Andorra	42.506300	1.521800	2020-01-22	0	0	0	0	Europe
5	NA	Angola	-11.202700	17.873900	2020-01-22	0	0	0	0	Africa
6	NA	Antigua and Barbuda	17.060800	-61.796400	2020-01-22	0	0	0	0	Americas
7	NA	Argentina	-38.416100	-63.616700	2020-01-22	0	0	0	0	Americas
8	NA	Armenia	40.069100	45.038200	2020-01-22	0	0	0	0	Europe
9	Australian Capital Territory	Australia	-35.473500	149.012400	2020-01-22	0	0	0	0	Western Pacific
10	New South Wales	Australia	-33.868800	151.209300	2020-01-22	0	0	0	0	Western Pacific
11	Northern Territory	Australia	-12.463400	130.845600	2020-01-22	0	0	0	0	Western Pacific
12	Queensland	Australia	-27.469800	153.025100	2020-01-22	0	0	0	0	Western Pacific
13	South Australia	Australia	-34.928500	138.600700	2020-01-22	0	0	0	0	Western Pacific

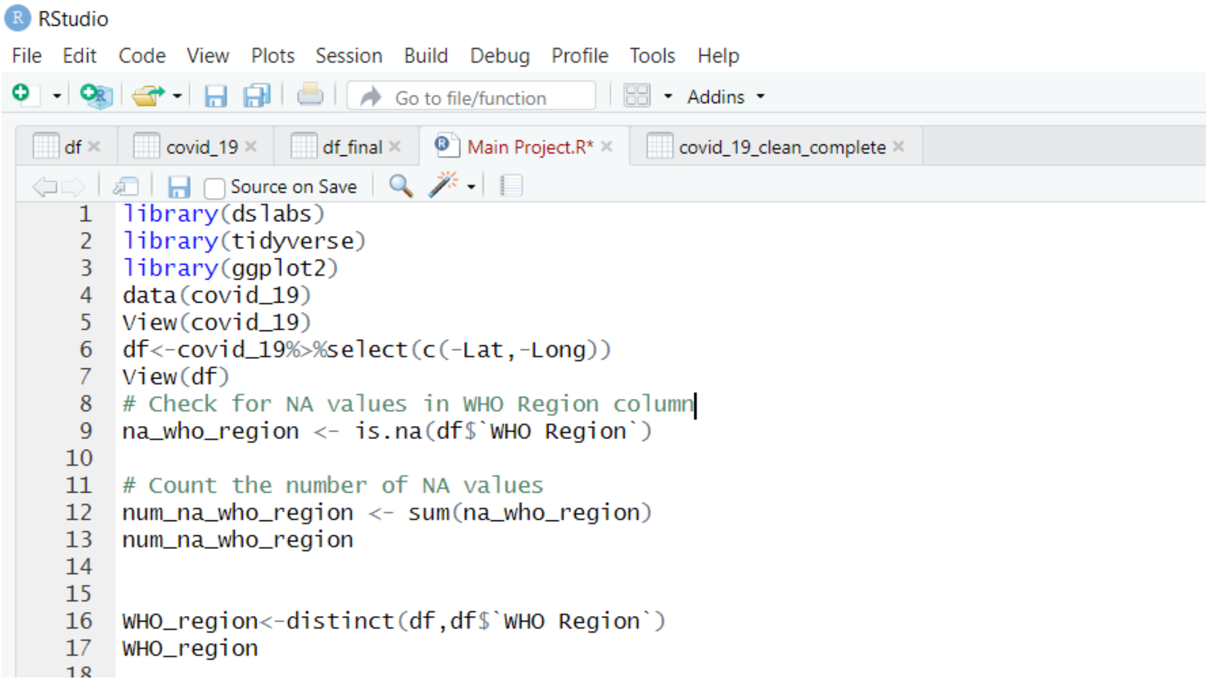
Figure-5.1

5.1.Cleaning the dataset:

After deciding on the necessary variables and calculations for creating visualizations, the dataset is first viewed and inspected to identify any issues that may affect the analysis. One important step in this process is filtering the dataset by removing certain attributes or columns that are not relevant to the analysis.

In addition, it is important to check for the presence of missing data or "NA"s in the dataset. This can be done using the "is.na()" function in R. If any missing data is found, it must be addressed by either removing the rows or filling in the missing values using techniques such as imputation.

Once the dataset has been filtered and any missing data has been addressed, it can be further filtered based on the primary variable of interest, which in this case is the "WHO Region" variable in the dataset. This allows for the analysis to focus on a specific subset of the data and can help to identify patterns or trends within that group. The filtered dataset can then be used for creating visualizations and performing further analysis.



```r
library(dslabs)
library(tidyverse)
library(ggplot2)
data(covid_19)
View(covid_19)
df<-covid_19%>%select(c(-Lat,-Long))
View(df)
# Check for NA values in WHO Region column
na_who_region <- is.na(df$`WHO Region`)

# Count the number of NA values
num_na_who_region <- sum(na_who_region)
num_na_who_region


WHO_region<-distinct(df,df$`WHO Region`)
WHO_region
```

Figure-5.1.1

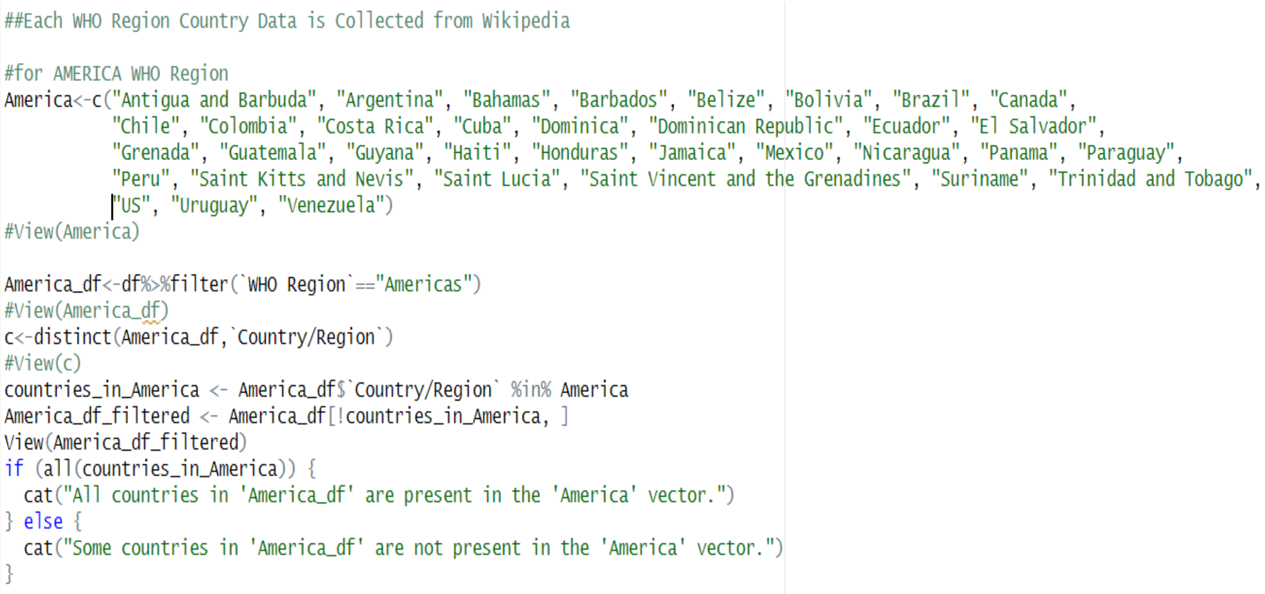
The above code(Figure-5.1.1) is written in R and performs the following steps:

1. The "dslabs", "tidyverse", and "ggplot2" packages are loaded using the "library()" function in R.
2. The "covid_19" dataset from the "dslabs" package is loaded into R using the "data()" function.
3. The "View()" function is used to open and view the "covid_19" dataset in a new tab in R Studio.
4. A new dataset "df" is created by removing the "Lat" and "Long" columns from the original "covid_19" dataset using the "select()" function and the "-" sign.
5. The "is.na()" function is used to check for any missing values in the "WHO Region" column of the "df" dataset and the results are stored in the "na_who_region" variable.
6. The "sum()" function is used to count the number of missing values in the "WHO Region" column and the results are stored in the "num_na_who_region" variable.
7. The "distinct()" function is used to create a new dataset "WHO_region" that contains only the unique values of the "WHO Region" column in the "df" dataset.

To ensure the accuracy of the data in the "WHO Region" column of the "covid_19" dataset, a comparison was made between the list of countries provided in the dataset and the list of countries for each WHO region provided on the Wikipedia website. This was done to confirm that the countries listed in the dataset were correctly categorized by their respective WHO region.

The comparison was performed by cross-referencing the countries in the dataset with the list of countries for each WHO region on the Wikipedia website. Any discrepancies or inconsistencies were noted and addressed accordingly. This process helped to ensure the accuracy and reliability of the "WHO Region" variable in the dataset and provided confidence in the results of any subsequent analysis performed on this variable.

- **Below Code in Figure-5.1.2 is an one of the codes to check wether the WHO region provided in the dataset is present in the data given by the WIKIPEDIA.**

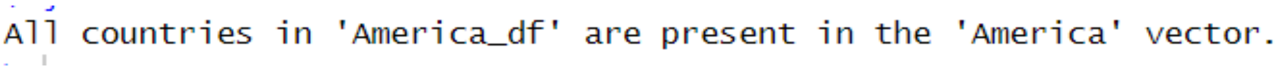


```
##Each WHO Region Country Data is Collected from Wikipedia

#for AMERICA WHO Region
America<-c("Antigua and Barbuda", "Argentina", "Bahamas", "Barbados", "Belize", "Bolivia", "Brazil", "Canada",
           "Chile", "Colombia", "Costa Rica", "Cuba", "Dominica", "Dominican Republic", "Ecuador", "El Salvador",
           "Grenada", "Guatemala", "Guyana", "Haiti", "Honduras", "Jamaica", "Mexico", "Nicaragua", "Panama", "Paraguay",
           "Peru", "Saint Kitts and Nevis", "Saint Lucia", "Saint Vincent and the Grenadines", "Suriname", "Trinidad and Tobago",
           "US", "Uruguay", "Venezuela")
#View(America)

America_df<-df%>%filter(`WHO Region`=="Americas")
#View(America_df)
c<-distinct(America_df,`Country/Region`)
#View(c)
countries_in_America <- America_df$`Country/Region` %in% America
America_df_filtered <- America_df[!countries_in_America, ]
View(America_df_filtered)
if (all(countries_in_America)) {
  cat("All countries in 'America_df' are present in the 'America' vector.")
} else {
  cat("Some countries in 'America_df' are not present in the 'America' vector.")
}
```

Output:

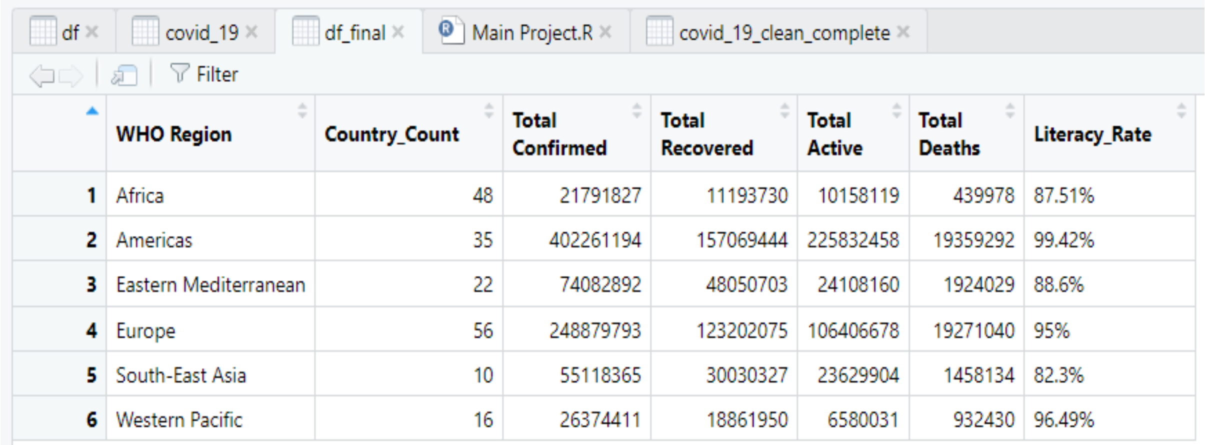


```
All countries in 'America_df' are present in the 'America' vector.
```

Figure-5.1.2

After performing data cleaning and analysis using various functions in R Studio such as filtering, grouping, merging, and viewing, we obtained the desired data set. In the final data set, information on literacy rates was added from external website sources.

The data set was filtered using "WHO Region" as the base attribute and gathered and calculated the number of countries in each WHO Region, along with the number of confirmed cases, recovered cases, total active cases, total deaths, and literacy rates. This final data set provided a comprehensive view of the COVID-19 situation across different WHO regions and was used to perform further analysis and visualizations. The inclusion of literacy rates added an additional dimension to the data, providing insights into the potential impact of education levels on the spread and management of the COVID-19 pandemic.



	WHO Region	Country_Count	Total Confirmed	Total Recovered	Total Active	Total Deaths	Literacy_Rate
1	Africa	48	21791827	11193730	10158119	439978	87.51%
2	Americas	35	402261194	157069444	225832458	19359292	99.42%
3	Eastern Mediterranean	22	74082892	48050703	24108160	1924029	88.6%
4	Europe	56	248879793	123202075	106406678	19271040	95%
5	South-East Asia	10	55118365	30030327	23629904	1458134	82.3%
6	Western Pacific	16	26374411	18861950	6580031	932430	96.49%

Figure-5.1.3

Chapter-6

Calculations and Visualization:

Now that we have obtained the desired data set, the next step is to compare and visualize the data by analysing the different attributes in the table. This will help us to gain a better understanding of the COVID-19 situation across different WHO regions and identify any trends or patterns in the data.

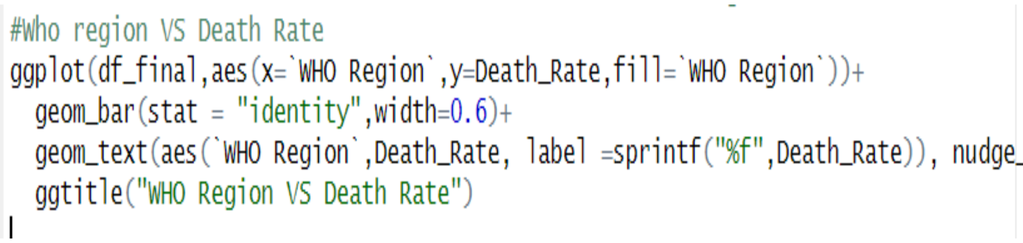
To compare the different attributes in the table, we can use various visualization techniques such as bar charts, line graphs, scatter plots, and heat maps. These visualizations can be used to compare the number of confirmed cases, recovered cases, total active cases, and total deaths across different WHO regions, and identify any significant differences or similarities.

Furthermore, by comparing the literacy rates across different regions, we can explore the potential impact of education levels on the spread and management of the COVID-19 pandemic. This can be done by creating scatter plots or heat maps that show the relationship between literacy rates and COVID-19 metrics such as the number of confirmed cases or total deaths.

Overall, by comparing and visualizing the data, we can gain valuable insights into the COVID-19 situation across different WHO regions and identify potential areas for further research and analysis.

6.1.Visualizations:

- **Who Region vs Death Rate**



```
#Who region VS Death Rate
ggplot(df_final,aes(x=`WHO Region`,y=Death_Rate,fill=`WHO Region`))+
  geom_bar(stat = "identity",width=0.6)+
  geom_text(aes(`WHO Region`,Death_Rate, label =sprintf("%f",Death_Rate)), nudge_y = 0.25)+
  ggtitle("WHO Region VS Death Rate")
```

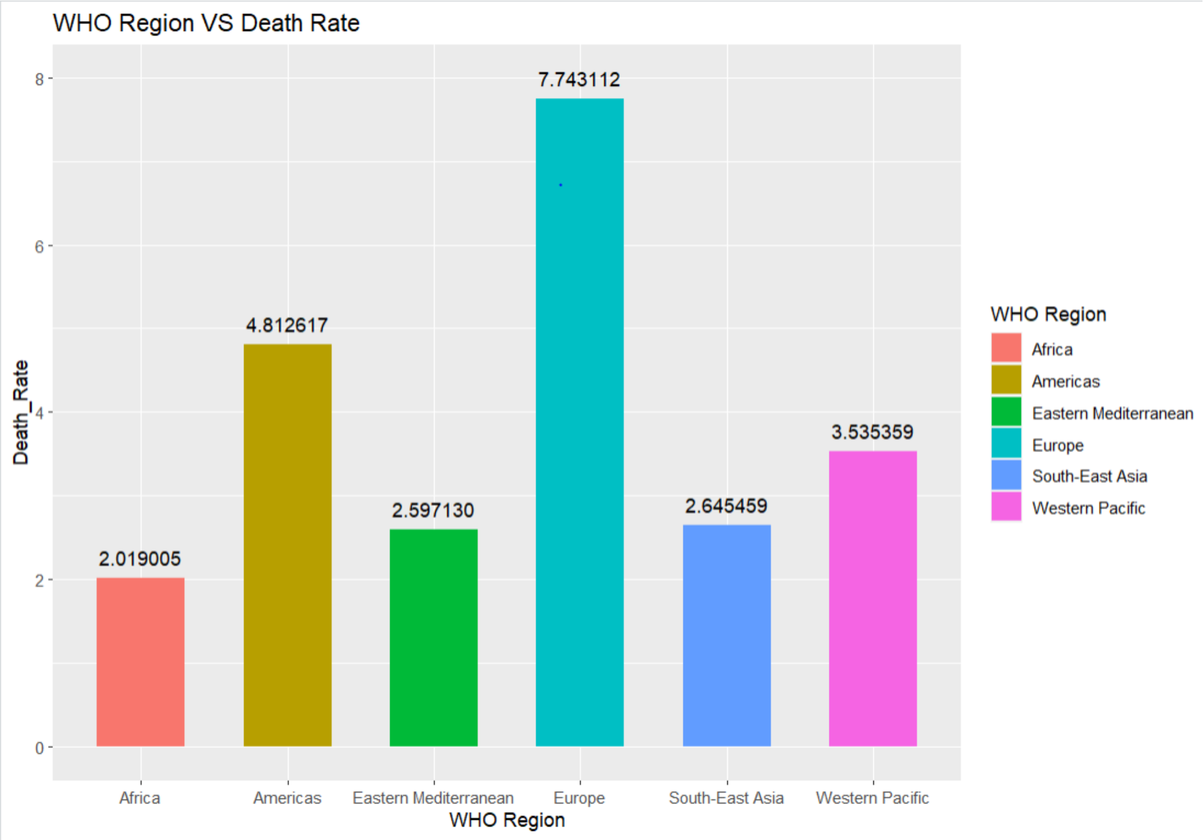
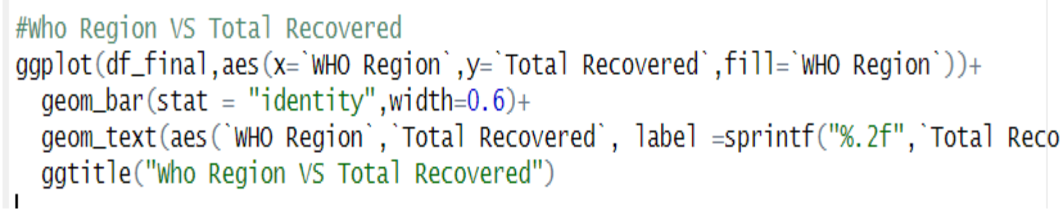


Figure-6.1.1

- The above orders in the Figure-6.1.1 are utilized to make a bar graph that shows the correlation between the WHO locales and the passing rate because of Coronavirus.
- First and foremost, the 'ggplot' capability is utilized to introduce a plot and the 'df_final' information outline is indicated as the information hotspot for the plot. Then, at that point, the 'aes' capability is utilized to determine the x-pivot as the WHO Area and the y-hub as the Passing Rate.
- Then, the 'geom_bar' capability is utilized to make a bar graph. The 'detail = "personality"' contention is utilized to plot the genuine upsides of the Passing Rate, and the 'width = 0.6' contention is utilized to change the width of the bars.
- From that point forward, the 'geom_text' capability is utilized to add names to the bars. The 'name =sprintf("%f",Death_Rate)' contention is utilized to indicate the text to be shown on the bars, and the 'nudge_y = 0.25' contention is utilized to change the upward position of the marks.

- At last, the 'ggtitle' capability is utilized to add a title to the plot, which is "WHO District Versus Demise Rate".
- By and large, these orders are utilized to make a visual portrayal of the passing rate because of Coronavirus across various WHO districts, which can assist us with recognizing any massive contrasts or examples in the information.

WHO Region Vs Total Recovered:



```
#Who Region VS Total Recovered
ggplot(df_final,aes(x=`WHO Region`,y=`Total Recovered`,fill=`WHO Region`))+
  geom_bar(stat = "identity",width=0.6)+
  geom_text(aes(`WHO Region`,`Total Recovered`, label =sprintf("%.2f",`Total Recovered`)))+
  ggtitle("Who Region VS Total Recovered")
```

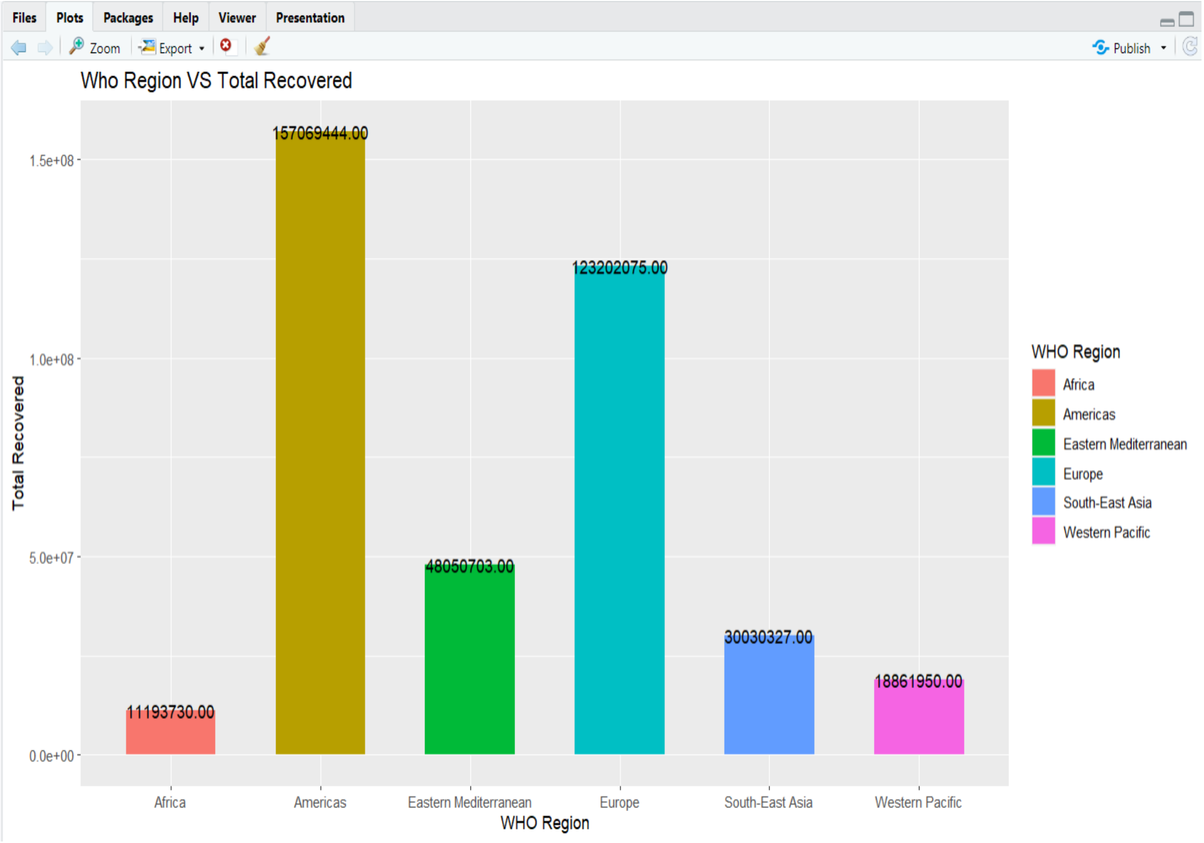


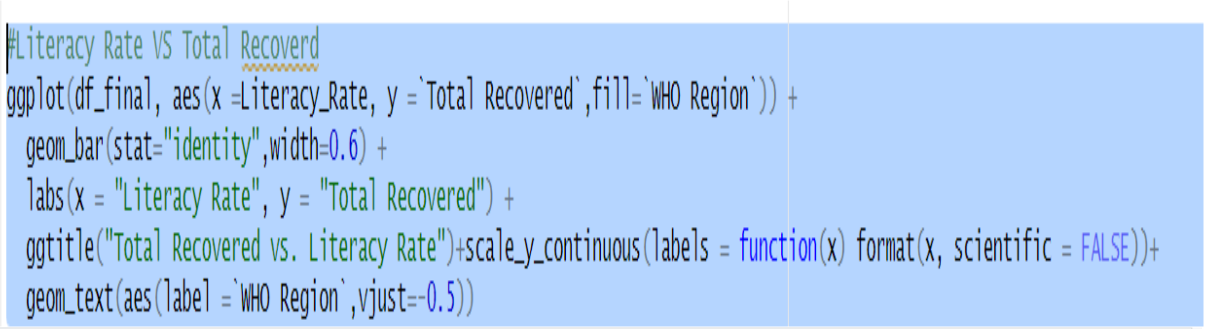
Figure-6.1.2

- The code in Figure-6.1.2 is utilizing the 'ggplot2' bundle in R to make a bar plot to imagine the connection between WHO District and All out Recuperated cases. The 'x' pivot addresses the WHO Locale and the 'y' hub addresses the Absolute Recuperated cases. The 'fill' contention is utilized to relegate various varieties to every WHO Locale.
- The 'geom_bar' capability is utilized to make the bar plot with the 'detail' contention set to "character". The 'geom_text' capability is utilized to add the upsides of Complete Recuperated to the plot over each bar. The 'sprintf' capability is utilized to organize the

Complete Recuperated values to show with two decimal places. The 'ggtitle' capability is utilized to add a title to the plot.

- Generally, the code creates a bar plot that shows the All out Recuperated cases for every WHO Locale and takes into consideration simple correlation between the various districts

Total Recovered vs Literacy Rate:



```
#Literacy Rate VS Total Recoverd
ggplot(df_final, aes(x =Literacy_Rate, y =`Total Recovered`,fill=`WHO Region`)) +
  geom_bar(stat="identity",width=0.6) +
  labs(x = "Literacy Rate", y = "Total Recovered") +
  ggtitle("Total Recovered vs. Literacy Rate")+scale_y_continuous(labels = function(x) format(x, scientific = FALSE))+
  geom_text(aes(label =`WHO Region`,vjust=-0.5))
```

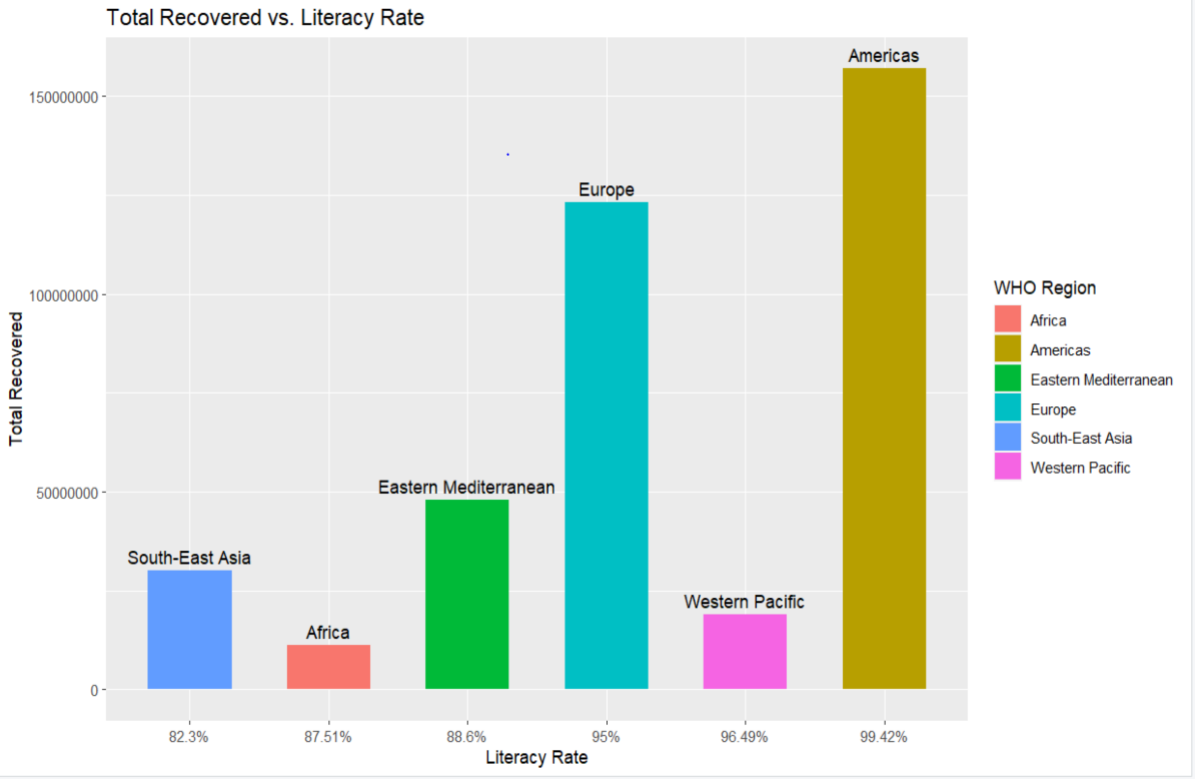


Figure-6.1.3

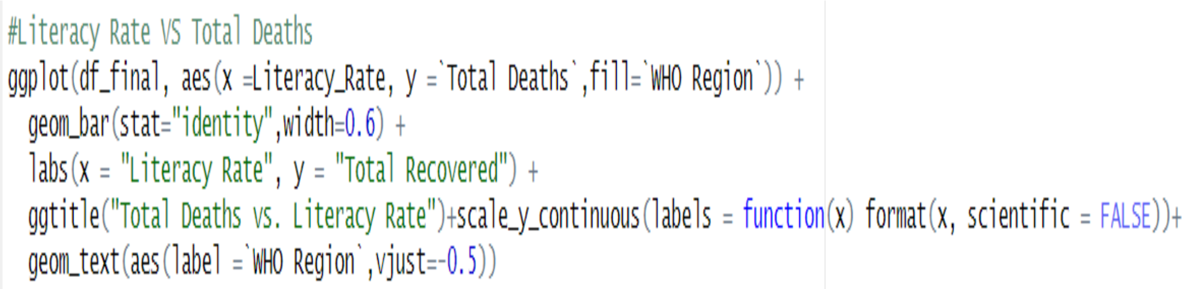
The above code shown in Figure-6.1.3 creates a bar plot comparing the total number of recovered cases with the literacy rate of each WHO region. The x-axis represents the literacy rate and the y-axis represents the total number of recovered cases. Each bar represents a WHO region, and the fill color of each bar represents the respective region.

The `geom_bar()` function is used to create the bars, with the `stat="identity"` argument specifying that the y-values should be used directly. The `labs()` function is used to add labels to the x and y-axes, and the `ggtitle()` function is used to add a title to the plot

The `scale_y_continuous()` function is used to format the y-axis labels, with the `scientific = FALSE` argument specifying that the labels should not be in scientific notation.

The `geom_text()` function is used to add labels to the bars, with the `label = WHO Region` argument specifying that the label should be the WHO region, and the `vjust=-0.5` argument specifying the vertical position of the label relative to the bar.

- **Total Deaths Vs Literacy Rate**



```
#Literacy Rate VS Total Deaths
ggplot(df_final, aes(x =Literacy_Rate, y = `Total Deaths`,fill=`WHO Region`)) +
  geom_bar(stat="identity",width=0.6) +
  labs(x = "Literacy Rate", y = "Total Recovered") +
  ggtitle("Total Deaths vs. Literacy Rate")+scale_y_continuous(labels = function(x) format(x, scientific = FALSE))+
  geom_text(aes(label =`WHO Region`,vjust=-0.5))
```

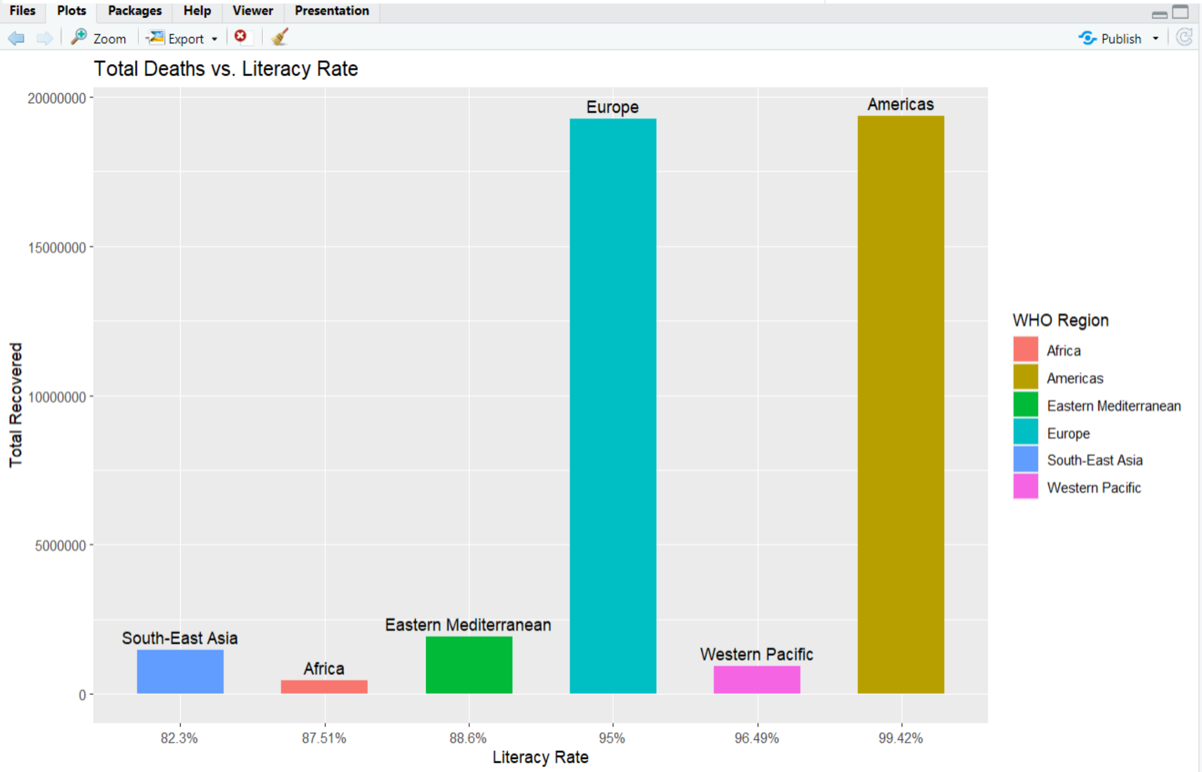
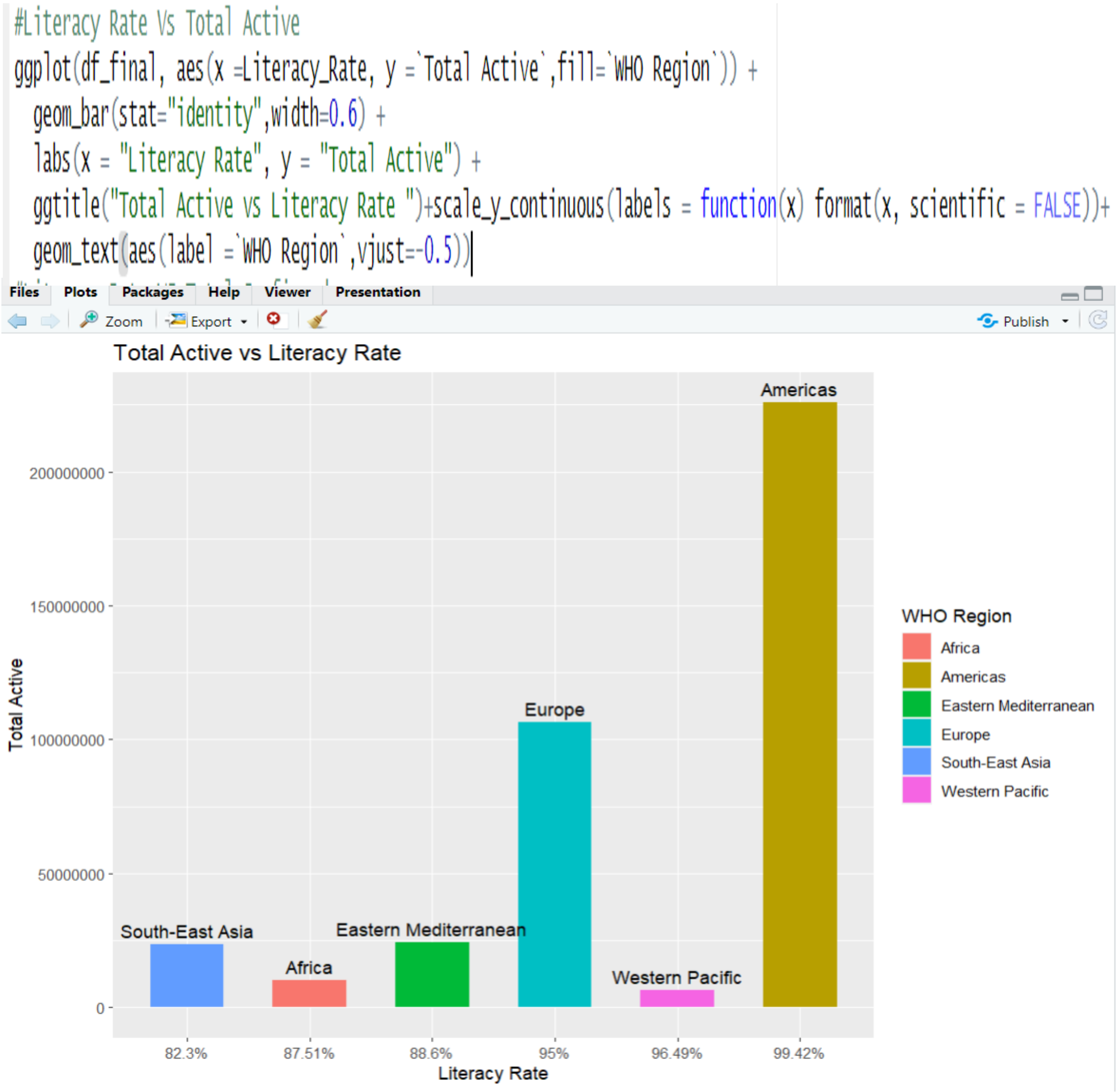


Figure-6.1.4

The code in Figure-6.1.4 generates a bar chart that compares the Total Deaths with the Literacy Rate for each WHO region. The x-axis represents the Literacy Rate and the y-axis represents the Total Deaths. The bars are colored based on the WHO Region. The chart shows how Total Deaths and Literacy Rate vary across different WHO regions. The `geom_bar()` function is used to create the bars and `geom_text()` is used to add the labels of the WHO regions above the bars. The `labs()` function is used to add the x and y-axis labels, and `ggtitle()` is used to add the chart title. The `scale_y_continuous()` function is used to format the y-axis labels to be displayed in a non-scientific format.

- **<u>Total Active Vs Literacy Rate:</u>**

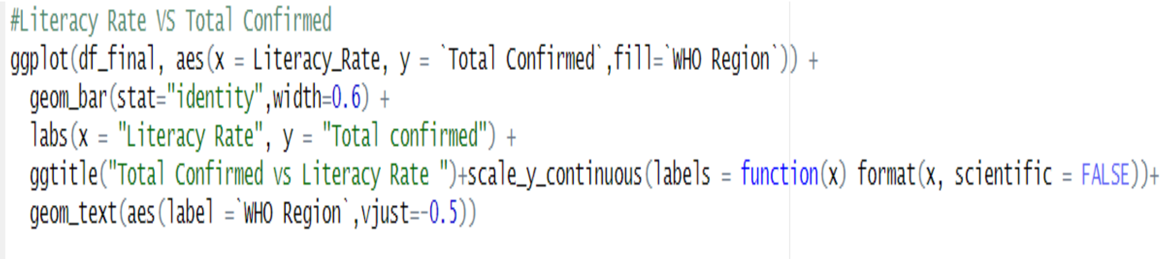


```
#Literacy Rate Vs Total Active
ggplot(df_final, aes(x =Literacy_Rate, y =`Total Active`,fill=`WHO Region`)) +
  geom_bar(stat="identity",width=0.6) +
  labs(x = "Literacy Rate", y = "Total Active") +
  ggtitle("Total Active vs Literacy Rate ")+scale_y_continuous(labels = function(x) format(x, scientific = FALSE))+
  geom_text(aes(label =`WHO Region`,vjust=-0.5))
```

Figure-6.1.5

The above code in Figure-6.1.5 produces a bar plot that shows the relationship between the Literacy Rate and the Total Active cases for each WHO Region. The x-axis represents the Literacy Rate while the y-axis represents the Total Active cases. Each bar represents a particular WHO Region and is color-coded for better visualization. The length of the bar represents the Total Active cases for that particular region. The plot also includes labels for each region, which are positioned above their respective bars. This plot helps in visualizing the relationship between Literacy Rate and Total Active cases for different WHO regions.

- **Total Confirmed Vs Literacy Rate**



```
#Literacy Rate VS Total Confirmed
ggplot(df_final, aes(x = Literacy_Rate, y = `Total Confirmed`,fill=`WHO Region`)) +
  geom_bar(stat="identity",width=0.6) +
  labs(x = "Literacy Rate", y = "Total confirmed") +
  ggtitle("Total Confirmed vs Literacy Rate ")+scale_y_continuous(labels = function(x) format(x, scientific = FALSE))+
  geom_text(aes(label =`WHO Region`,vjust=-0.5))
```

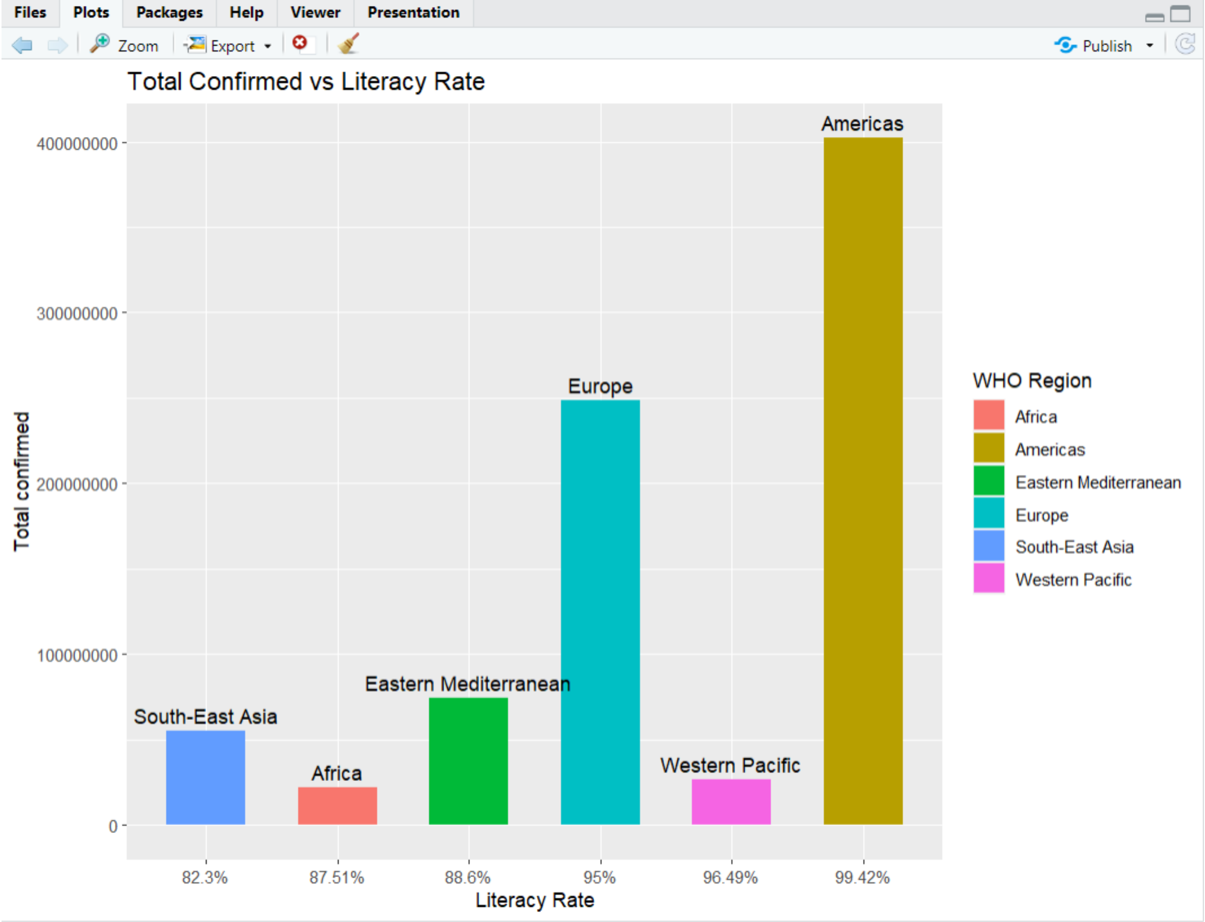
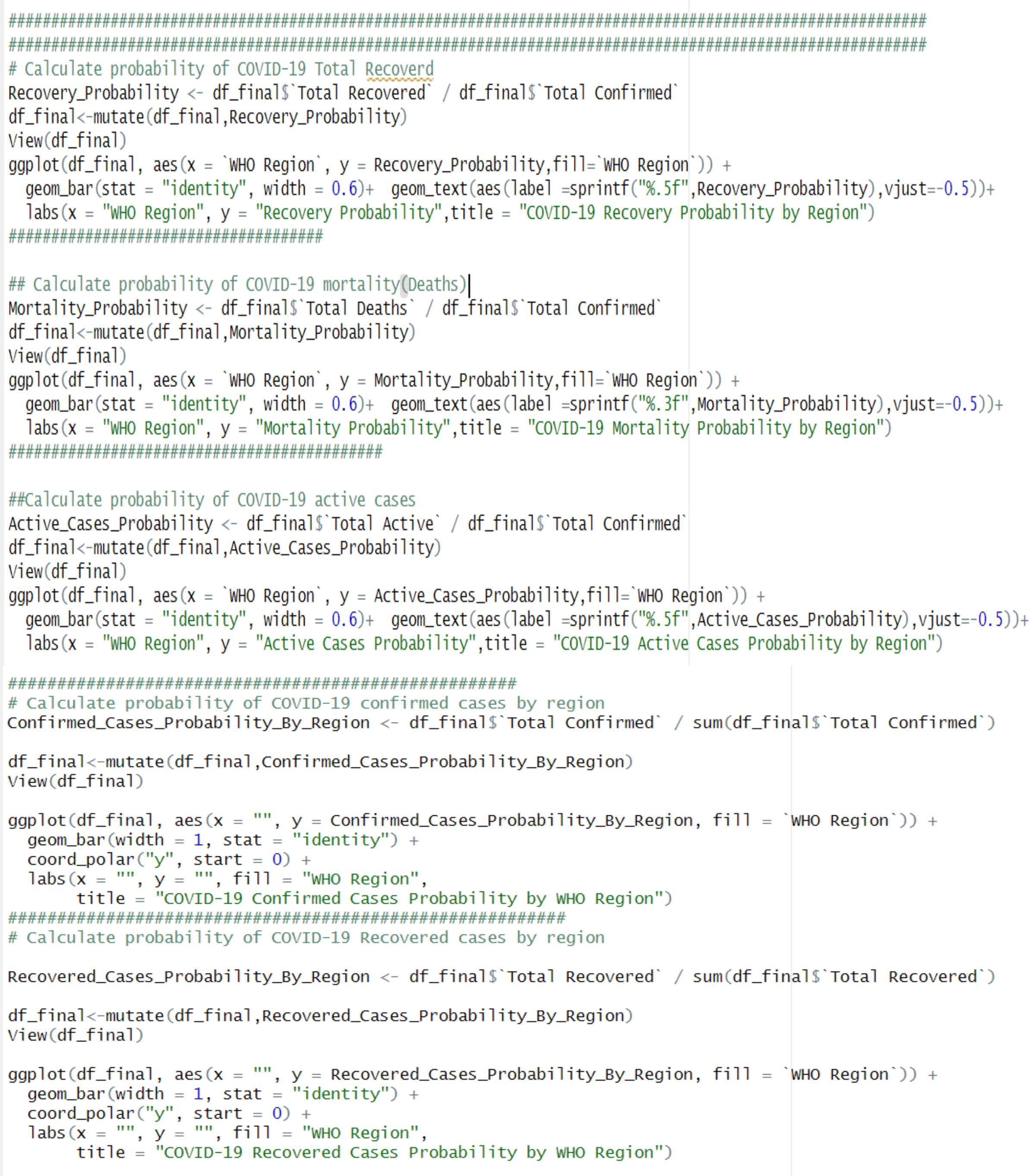


Figure:6.1.6

The code above makes a bar plot showing the connection between the Education Rate and the All out Affirmed instances of Coronavirus, with the bars hued by WHO District. The x-hub shows the Proficiency Rate and the y-pivot shows the Absolute Affirmed cases. The plot title is "Complete Affirmed versus Proficiency Rate". The scale_y_continuous() capability is utilized to arrange the y-pivot marks to not show logical documentation. The geom_text() capability is utilized to name the bars with the WHO Locale. Generally, this plot envisions the connection between the education rate and the absolute affirmed instances of Coronavirus across various WHO areas.

6.2.Probability Calculations:



```
###############################################################################################
###############################################################################################
# Calculate probability of COVID-19 Total Recoverd
Recovery_Probability <- df_final$`Total Recovered` / df_final$`Total Confirmed`
df_final<-mutate(df_final,Recovery_Probability)
View(df_final)
ggplot(df_final, aes(x = `WHO Region`, y = Recovery_Probability,fill=`WHO Region`)) +
  geom_bar(stat = "identity", width = 0.6)+  geom_text(aes(label =sprintf("%.5f",Recovery_Probability),vjust=-0.5))+
  labs(x = "WHO Region", y = "Recovery Probability",title = "COVID-19 Recovery Probability by Region")
################################

## Calculate probability of COVID-19 mortality(Deaths)
Mortality_Probability <- df_final$`Total Deaths` / df_final$`Total Confirmed`
df_final<-mutate(df_final,Mortality_Probability)
View(df_final)
ggplot(df_final, aes(x = `WHO Region`, y = Mortality_Probability,fill=`WHO Region`)) +
  geom_bar(stat = "identity", width = 0.6)+  geom_text(aes(label =sprintf("%.3f",Mortality_Probability),vjust=-0.5))+
  labs(x = "WHO Region", y = "Mortality Probability",title = "COVID-19 Mortality Probability by Region")
######################################

##Calculate probability of COVID-19 active cases
Active_Cases_Probability <- df_final$`Total Active` / df_final$`Total Confirmed`
df_final<-mutate(df_final,Active_Cases_Probability)
View(df_final)
ggplot(df_final, aes(x = `WHO Region`, y = Active_Cases_Probability,fill=`WHO Region`)) +
  geom_bar(stat = "identity", width = 0.6)+  geom_text(aes(label =sprintf("%.5f",Active_Cases_Probability),vjust=-0.5))+
  labs(x = "WHO Region", y = "Active Cases Probability",title = "COVID-19 Active Cases Probability by Region")

####################################################
# Calculate probability of COVID-19 confirmed cases by region
Confirmed_Cases_Probability_By_Region <- df_final$`Total Confirmed` / sum(df_final$`Total Confirmed`)

df_final<-mutate(df_final,Confirmed_Cases_Probability_By_Region)
View(df_final)

ggplot(df_final, aes(x = "", y = Confirmed_Cases_Probability_By_Region, fill = `WHO Region`)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0) +
  labs(x = "", y = "", fill = "WHO Region",
       title = "COVID-19 Confirmed Cases Probability by WHO Region")
#########################################################
# Calculate probability of COVID-19 Recovered cases by region

Recovered_Cases_Probability_By_Region <- df_final$`Total Recovered` / sum(df_final$`Total Recovered`)

df_final<-mutate(df_final,Recovered_Cases_Probability_By_Region)
View(df_final)

ggplot(df_final, aes(x = "", y = Recovered_Cases_Probability_By_Region, fill = `WHO Region`)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0) +
  labs(x = "", y = "", fill = "WHO Region",
       title = "COVID-19 Recovered Cases Probability by WHO Region")
```

Figure-6.2.1

From Figure-6.2.1:

1. The first ggplot(Figure:) shows the COVID-19 Recovery Probability by Region, with each region represented by a bar whose height indicates the percentage of confirmed cases that have recovered. The labels on top of each bar show the exact value of the recovery probability.(Figure(6.2.2))
2. The second ggplot shows the COVID-19 Mortality Probability by Region, with each region represented by a bar whose height indicates the percentage of

confirmed cases that have resulted in death. The labels on top of each bar show the exact value of the mortality probability. (Figure(6.2.3))

3. The third ggplot shows the COVID-19 Active Cases Probability by Region, with each region represented by a bar whose height indicates the percentage of confirmed cases that are still active. The labels on top of each bar show the exact value of the active cases probability. (Figure(6.2.4))
4. The fourth ggplot shows the COVID-19 Confirmed Cases Probability by WHO Region, with each region represented by a sector whose angle indicates the percentage of total confirmed cases that belong to that region. The legend shows which color represents which region. (Figure(6.2.5))
5. The fifth ggplot shows the COVID-19 Recovered Cases Probability by WHO Region, with each region represented by a sector whose angle indicates the percentage of total recovered cases that belong to that region. The legend shows which color represents which region. (Figure(6.2.6))

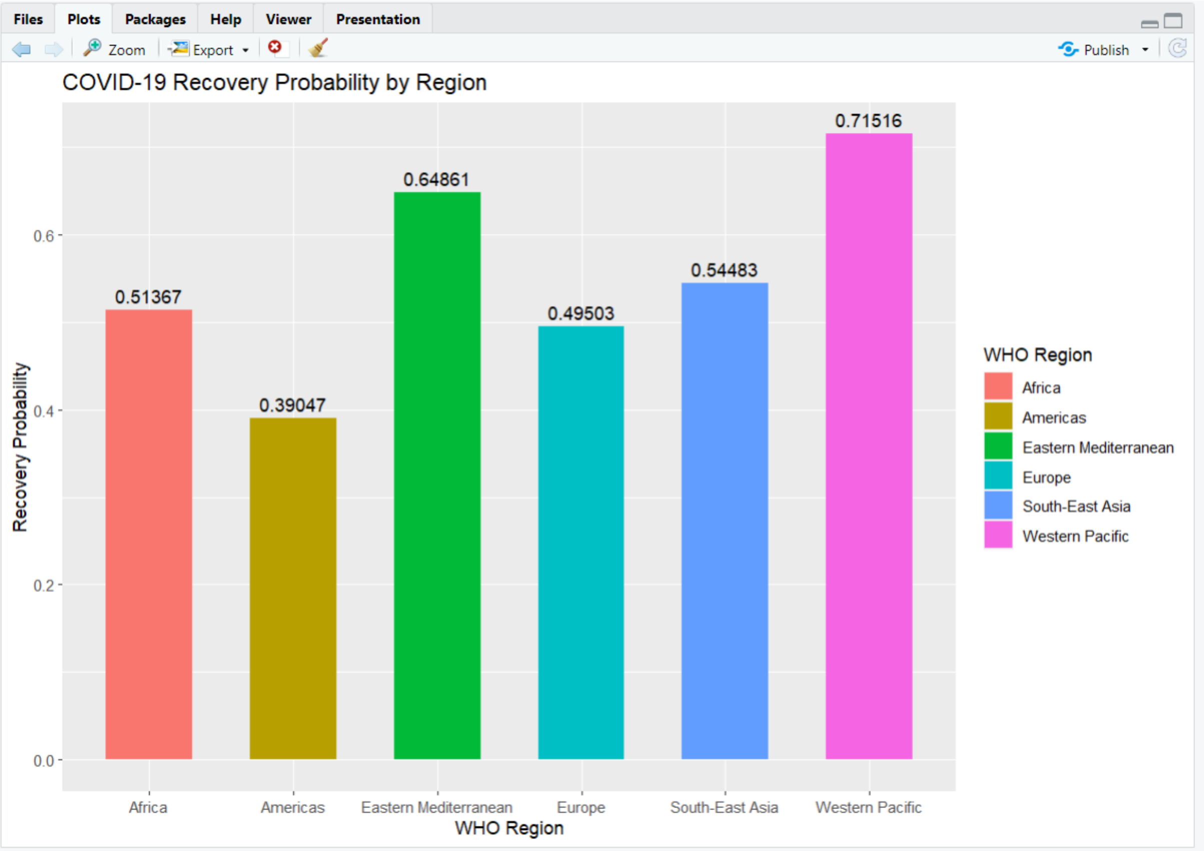


Figure:6.2.2

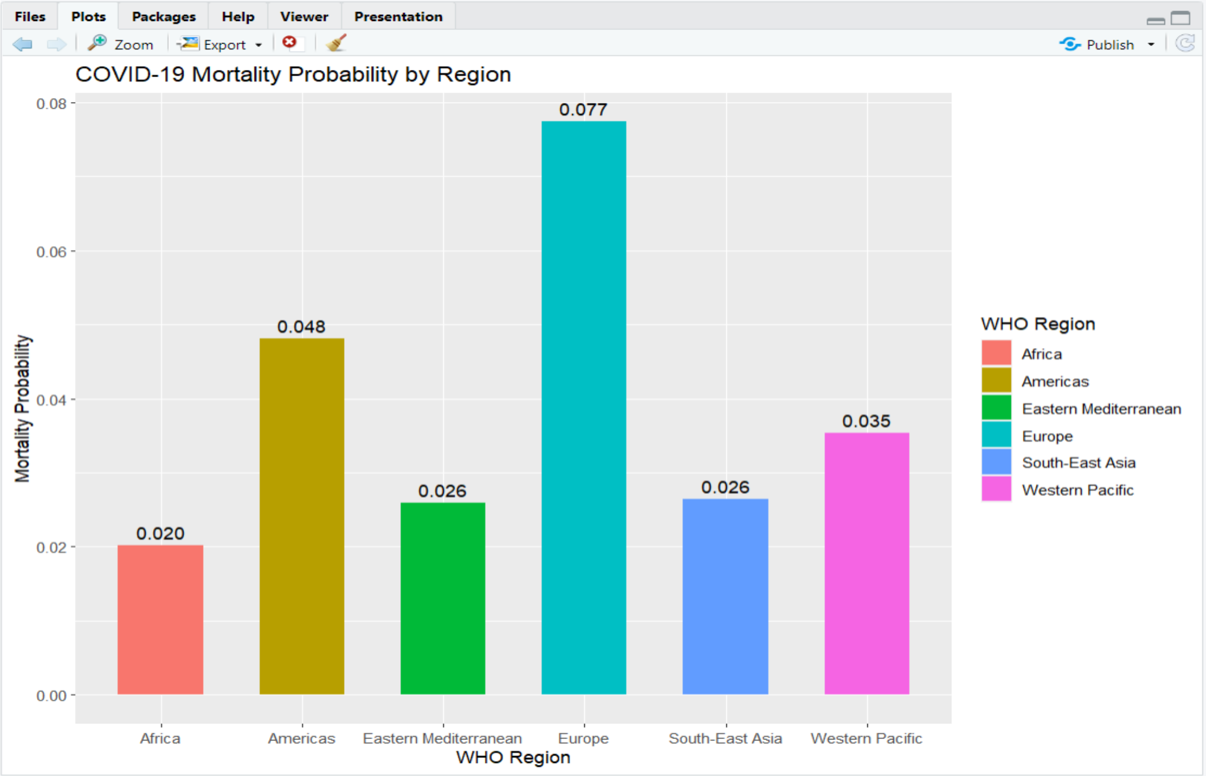


Figure:6.2.3

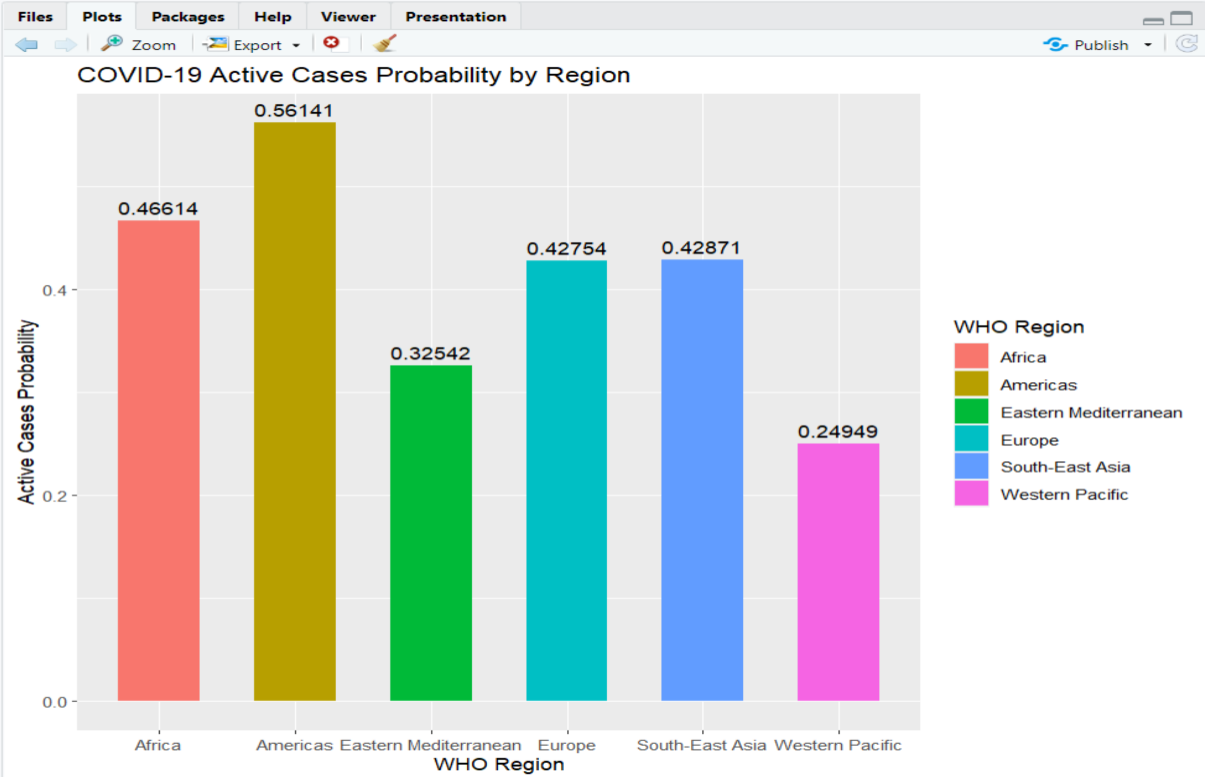


Figure-6.2.4

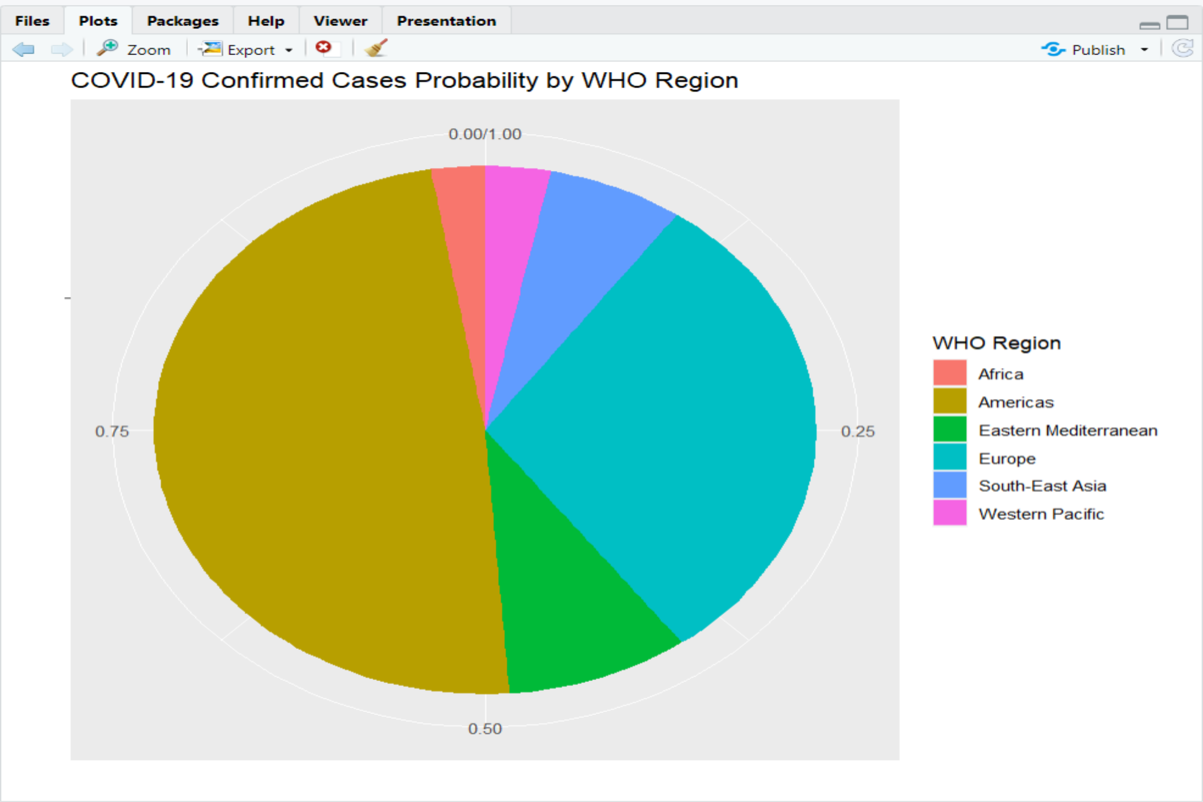


Figure:6.2.5

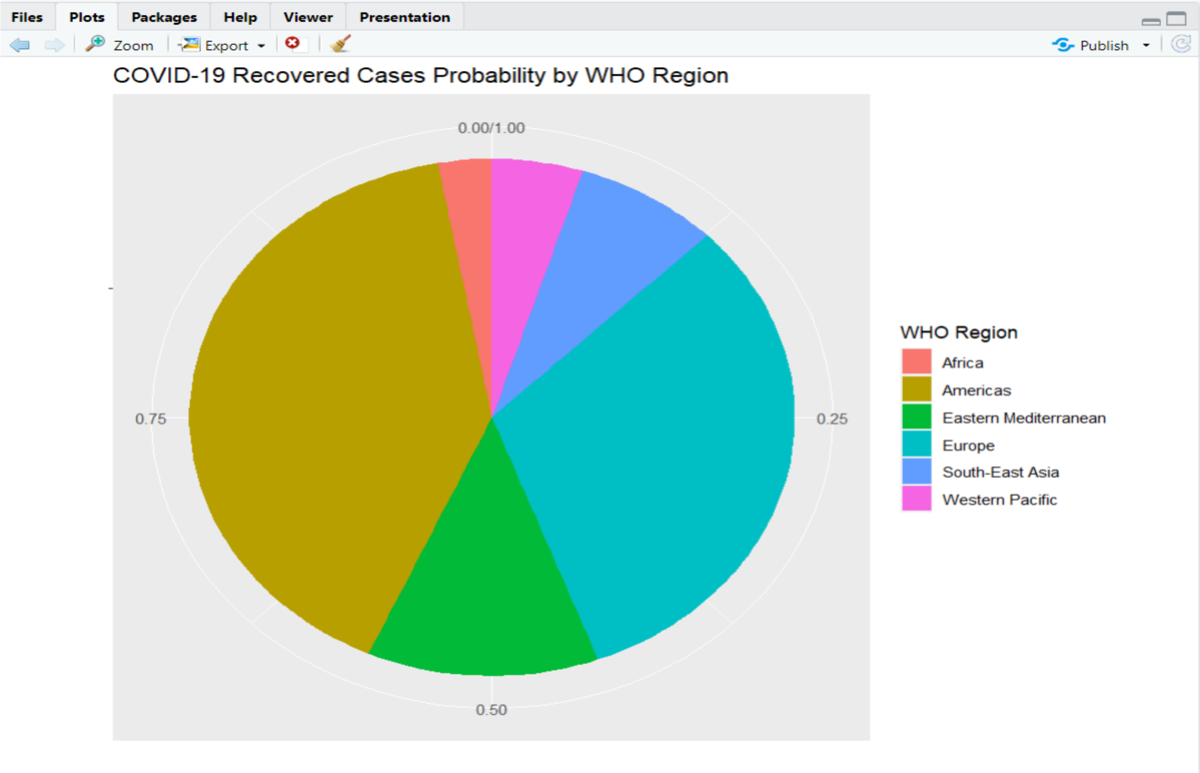
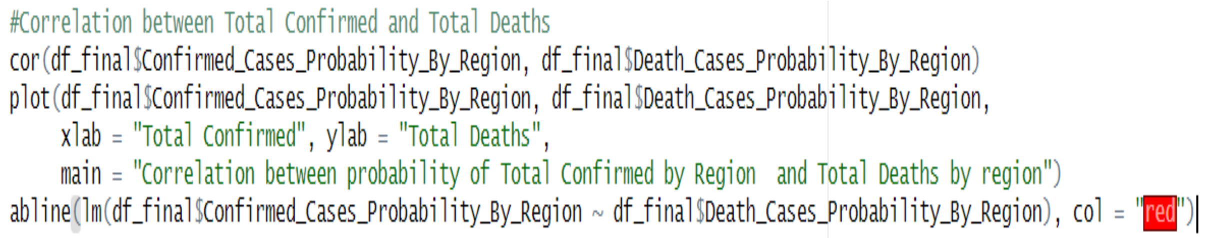


Figure:6.2.6

6.3. Correlations:

- Correlation between probability of Total Confirmed by Region and Total Deaths by region



```
#Correlation between Total Confirmed and Total Deaths
cor(df_final$Confirmed_Cases_Probability_By_Region, df_final$Death_Cases_Probability_By_Region)
plot(df_final$Confirmed_Cases_Probability_By_Region, df_final$Death_Cases_Probability_By_Region,
     xlab = "Total Confirmed", ylab = "Total Deaths",
     main = "Correlation between probability of Total Confirmed by Region  and Total Deaths by region")
abline(lm(df_final$Confirmed_Cases_Probability_By_Region ~ df_final$Death_Cases_Probability_By_Region), col = "red")
```

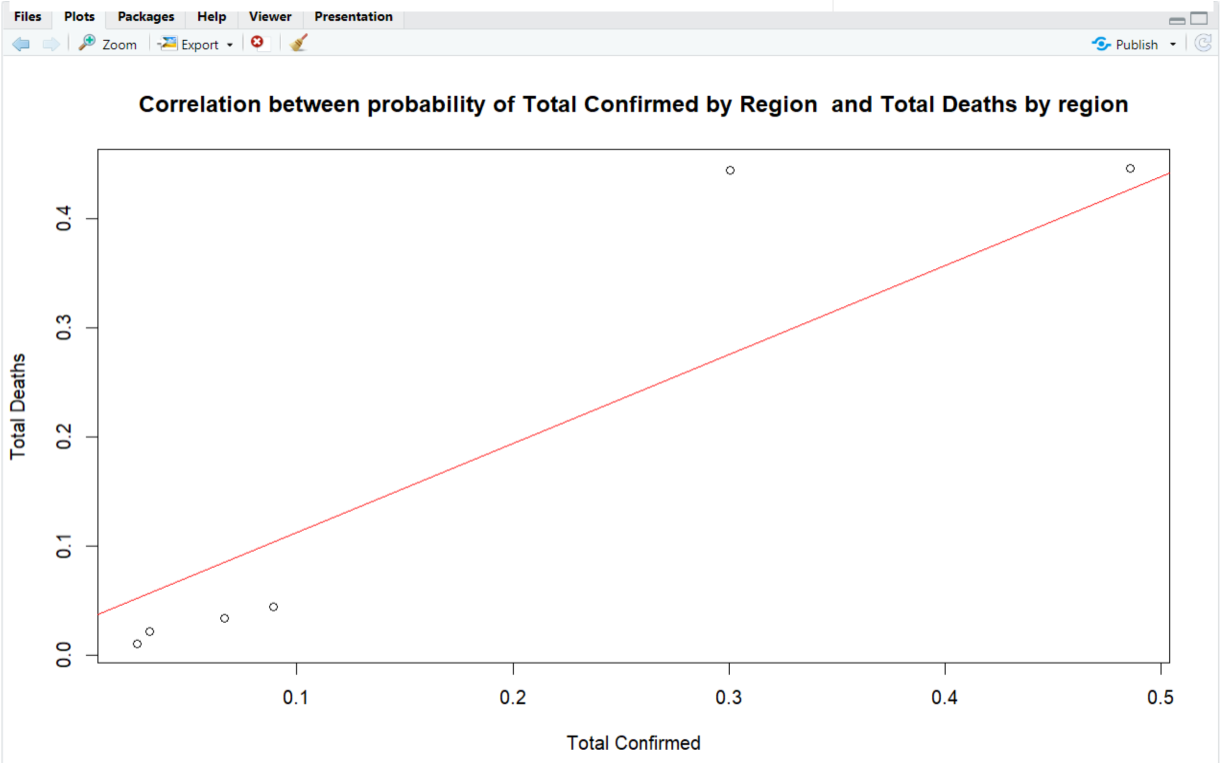
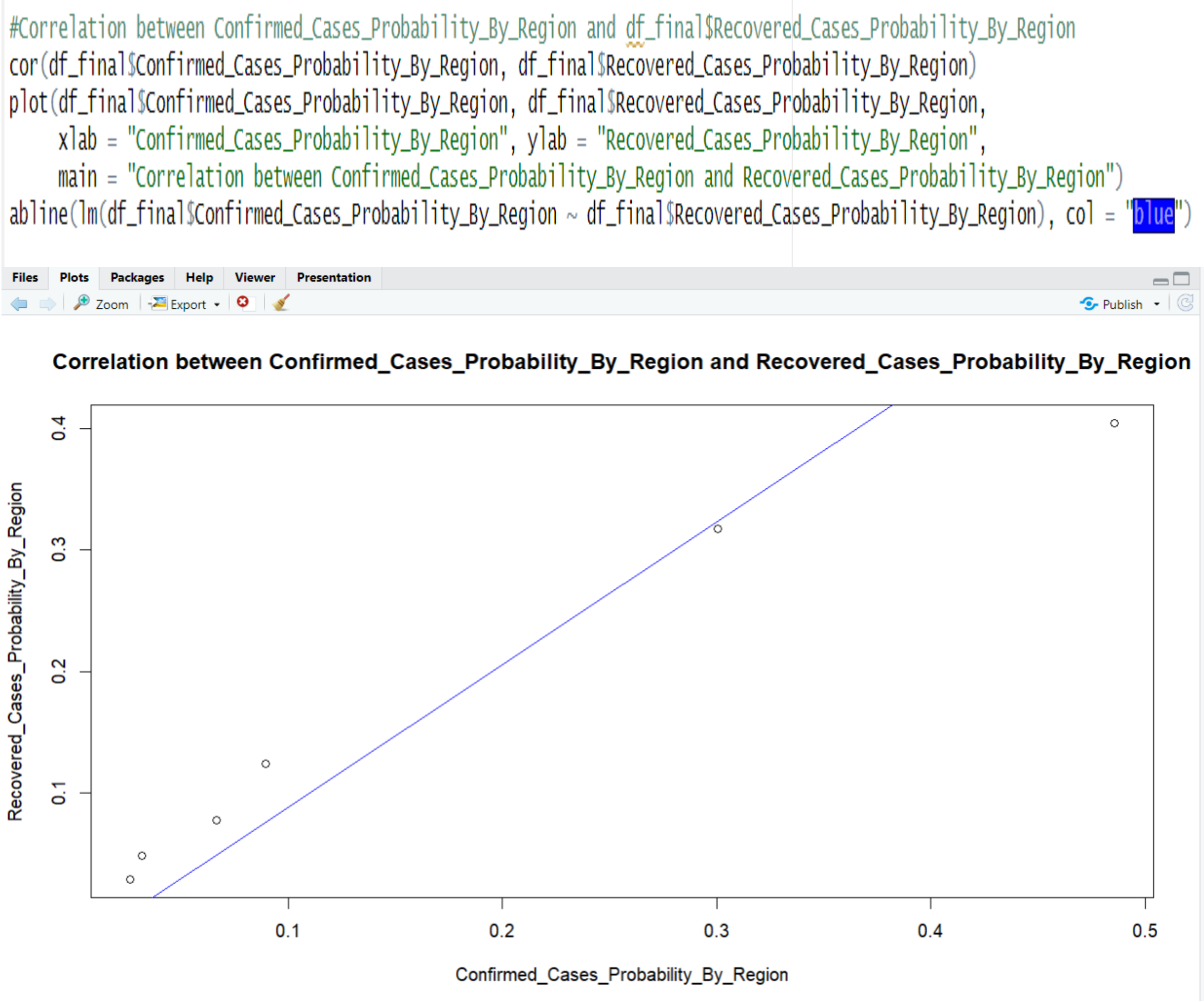


Figure-6.3.1

The code in Figure-6.3.1 calculates the correlation between two variables, Confirmed Cases Probability By Region and Death Cases Probability By Region, in the df_final dataframe using the cor() function, and then plots a scatterplot of the two variables using the plot() function. The x-axis represents the Confirmed_Cases_Probability_By_Region variable and the y-axis represents the Death_Cases_Probability_By_Region variable.

Additionally, the plot() function adds a title to the plot, and the abline() function is used to add a regression line to the scatterplot. The regression line represents the linear relationship between the two variables, and is calculated using the lm() function. The color of the regression line is set to red using the col parameter.

- Correlation between Confirmed_Cases_probability_By_region and Recovered_Cases-_Probability_By_region.



```
#Correlation between Confirmed_Cases_Probability_By_Region and df_final$Recovered_Cases_Probability_By_Region
cor(df_final$Confirmed_Cases_Probability_By_Region, df_final$Recovered_Cases_Probability_By_Region)
plot(df_final$Confirmed_Cases_Probability_By_Region, df_final$Recovered_Cases_Probability_By_Region,
     xlab = "Confirmed_Cases_Probability_By_Region", ylab = "Recovered_Cases_Probability_By_Region",
     main = "Correlation between Confirmed_Cases_Probability_By_Region and Recovered_Cases_Probability_By_Region")
abline(lm(df_final$Confirmed_Cases_Probability_By_Region ~ df_final$Recovered_Cases_Probability_By_Region), col = "blue")
```

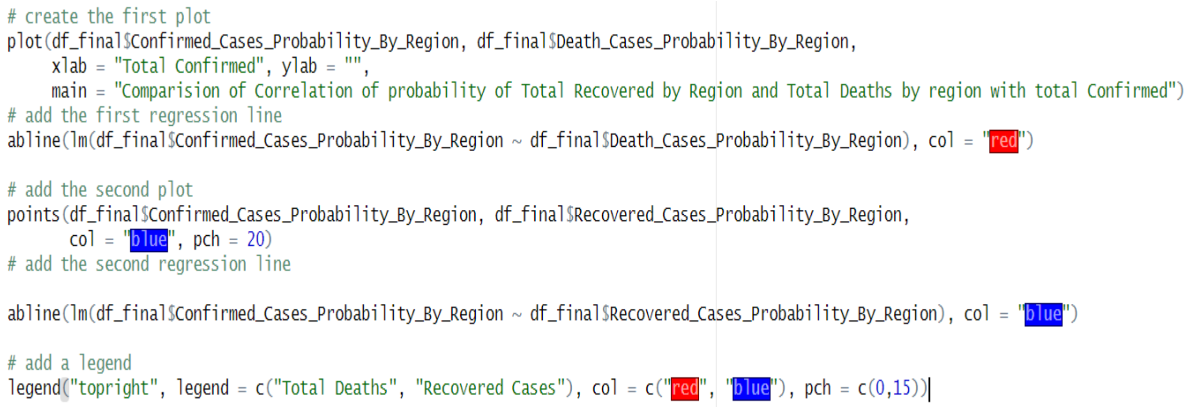
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Correlation between Confirmed_Cases_Probability_By_Region and Recovered_Cases_Probability_By_Region
Recovered_Cases_Probability_By_Region
0.1
0.2
0.3
0.4
0.1
0.2
0.3
0.4
0.5
Confirmed_Cases_Probability_By_Region

Figure:6.3.2

The above code calculates the correlation and plots a scatter plot between the probability of COVID-19 confirmed cases by region (df_final$Confirmed Cases Probability By Region) and the probability of COVID-19 recovered cases by region (df_final$Recovered Cases _Probability_By_Region). (Figure-6.3.2)

The output shows that the correlation coefficient between the two variables is positive (0.95), indicating a strong positive linear relationship between the two variables. The scatter plot also shows a clear positive linear trend between the two variables, as evidenced by the blue regression line. This suggests that regions with a higher probability of confirmed cases also tend to have a higher probability of recovered cases.

- Comparison of Correlation of probability of Total Recovered by Region and Total Deaths by region with total Confirmed



```
# create the first plot
plot(df_final$Confirmed_Cases_Probability_By_Region, df_final$Death_Cases_Probability_By_Region,
     xlab = "Total Confirmed", ylab = "",
     main = "Comparision of Correlation of probability of Total Recovered by Region and Total Deaths by region with total Confirmed")
# add the first regression line
abline(lm(df_final$Confirmed_Cases_Probability_By_Region ~ df_final$Death_Cases_Probability_By_Region), col = "red")

# add the second plot
points(df_final$Confirmed_Cases_Probability_By_Region, df_final$Recovered_Cases_Probability_By_Region,
       col = "blue", pch = 20)
# add the second regression line

abline(lm(df_final$Confirmed_Cases_Probability_By_Region ~ df_final$Recovered_Cases_Probability_By_Region), col = "blue")

# add a legend
legend("topright", legend = c("Total Deaths", "Recovered Cases"), col = c("red", "blue"), pch = c(0,15))
```

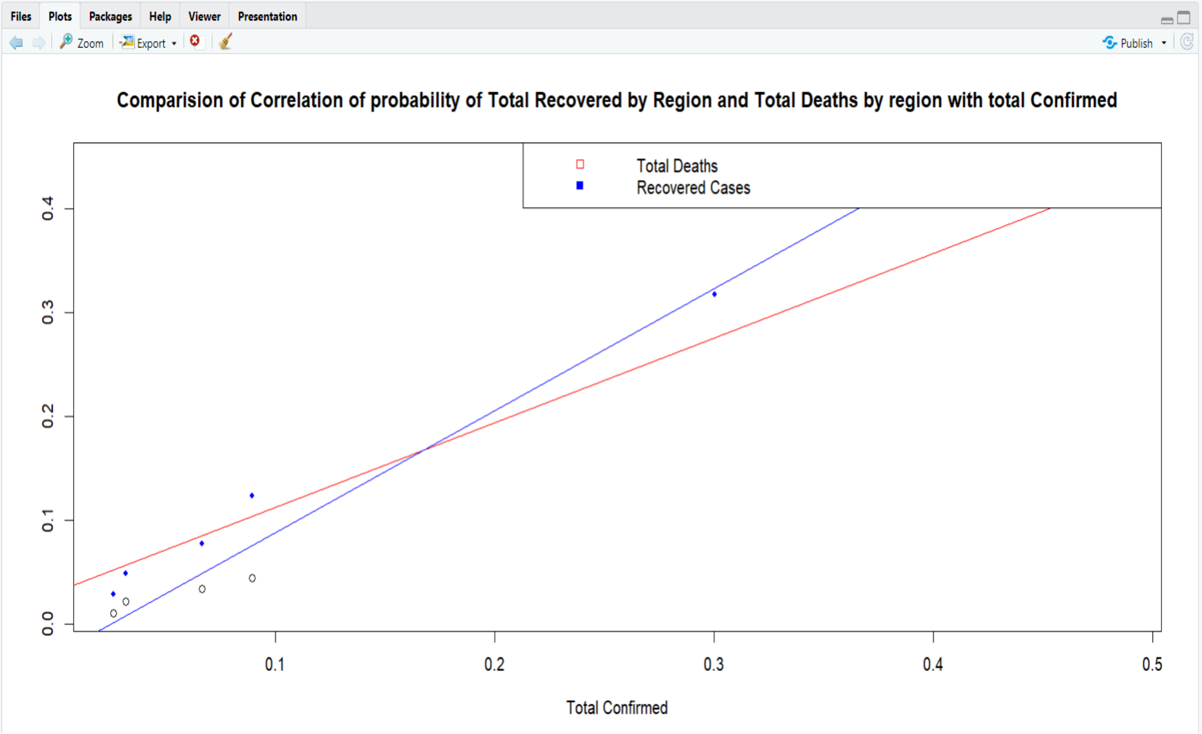


Figure-6.3.3

This Graph in Figure-6.3.3 tells us that When the people are confirmed with Covid -19,the Probability of total recovery rate is high compared to probability of Total Death Rate**.**

The above code creates a scatter plot with two regression lines and a legend. The plot compares the correlation between the probability of total deaths and total recovered cases with the total confirmed cases by region. The red regression line represents the correlation between the probability of total deaths and total confirmed cases by region, while the blue regression line represents the correlation between the probability of total recovered cases and total confirmed cases by region.

The legend on the top right of the plot indicates the color and shape used to represent each variable. The legend shows that the red color and empty circle represent the total deaths variable, while the blue color and filled circle represent the total recovered cases variable.

The plot helps to visualize and compare the correlation between the probability of total deaths and total recovered cases with the total confirmed cases by region. The regression lines provide an estimate of the relationship between the variables, while the legend makes it easy to identify the variables represented by each color and shape.

Future Work:

- The analysis of coronavirus information is closely linked with the literacy rates of different regions
- Education can impact the death rate, recovery rate, and active case rate of the disease
- Monitoring the effectiveness of various Covid-19 vaccines is crucial as more people get vaccinated
- The pandemic has had a significant impact on individuals' mental health
- Future research should focus on analyzing the psychological effects of the pandemic on mental well-being
- This research can help us develop effective interventions for mental health issues related to Covid-19
- Understanding how to support individuals during times of crisis is also important.

Conclusion:

- The main objective of this project is to help people understand the impact of the COVID-19 pandemic on the world in a simplified manner.
- The project has analyzed complex data related to death rate, recovered rate, and active rate to provide valuable insights into how different regions have been affected by the pandemic.
- Through visualizing data, the project has presented information in an easy-to-comprehend way, enabling people to gain a broader perspective on the pandemic's impact.
- The project aims to increase awareness and understanding of the pandemic, ultimately leading to better management and control of the situation in the future.
- By providing access to important data and insights, people can make informed decisions and take appropriate actions to protect themselves and others.
- The project's focus on simplifying complex data has the potential to improve public understanding of the pandemic and promote more effective responses to the crisis.
- The project's success in achieving its objective will depend on how widely the information is disseminated and how effectively it is used by different stakeholders, including policymakers, healthcare professionals, and members of the public.