Report on Exploratory Data Analysis (Report)

Mathematical Foundation of Computer Science
Masters in Computer Science (2022-2024)
Department of Computer Science, University of Delhi

Submitted To: Dr. Vasudha Bhatnagar

Submitted By: Arpit Kumar Mishra, Bhushan, Mukesh Kumar

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a data analysis approach that utilizes visual techniques to analyze data. EDA is primarily used to discover trends and patterns, and to check assumptions using statistical summary and graphical representations. By visually exploring the data, EDA enables a better understanding of the data's main features, the relationships between variables, and the identification of important variables for problem-solving. EDA often involves the use of descriptive statistics to gain quick insights into the characteristics of the data. Ultimately, EDA serves as a powerful tool for gaining a comprehensive understanding of data and identifying important trends and patterns.

It is a phenomenon under data analysis used for gaining a better understanding of data aspects like:

- main features of data
- variables and relationships that hold between them
- identifying which variables are important for our problem

Descriptive statistics: It refers to a set of techniques and measures used to summarize and describe the main characteristics of a dataset. These statistics provide a clear and concise summary of the data, including measures of central tendency.

Description of Data Set(1/2)

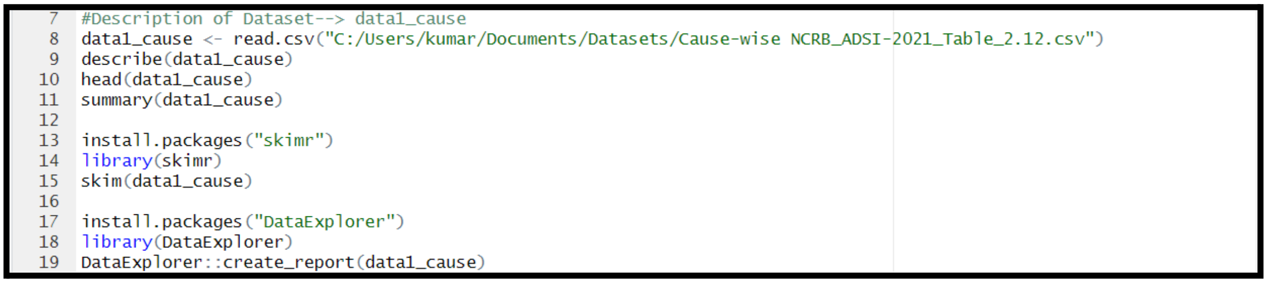
Name of the Data set: Cause-wise Distribution of Suicides by Means/Mode Adopted in India during 2021
Link to the data set:
https://data.gov.in/resource/cause-wise-distribution-suicides-meansmode-adopted-india-during-2021
Source: Open Government Data(OGD) Platform India
Released under: National Data Sharing and Accessibility Policy(NDSAP)
Contributor: Ministry of Home Affairs , Department of States , National Crime Records Bureau (NCRB)

Description: The Cause-wise Distribution of Suicides by Means/Mode Adopted in India during 2021 dataset provides information on the number of suicides committed in India in 2021, broken down by means/mode of suicide. The data was collected by the National Crime Records Bureau (NCRB), a department of the Indian Ministry of Home Affairs, and released under the National Data Sharing and Accessibility Policy (NDSAP).

The dataset includes information on the number of suicides committed by various means/modes such as hanging, poisoning, drowning, and others.

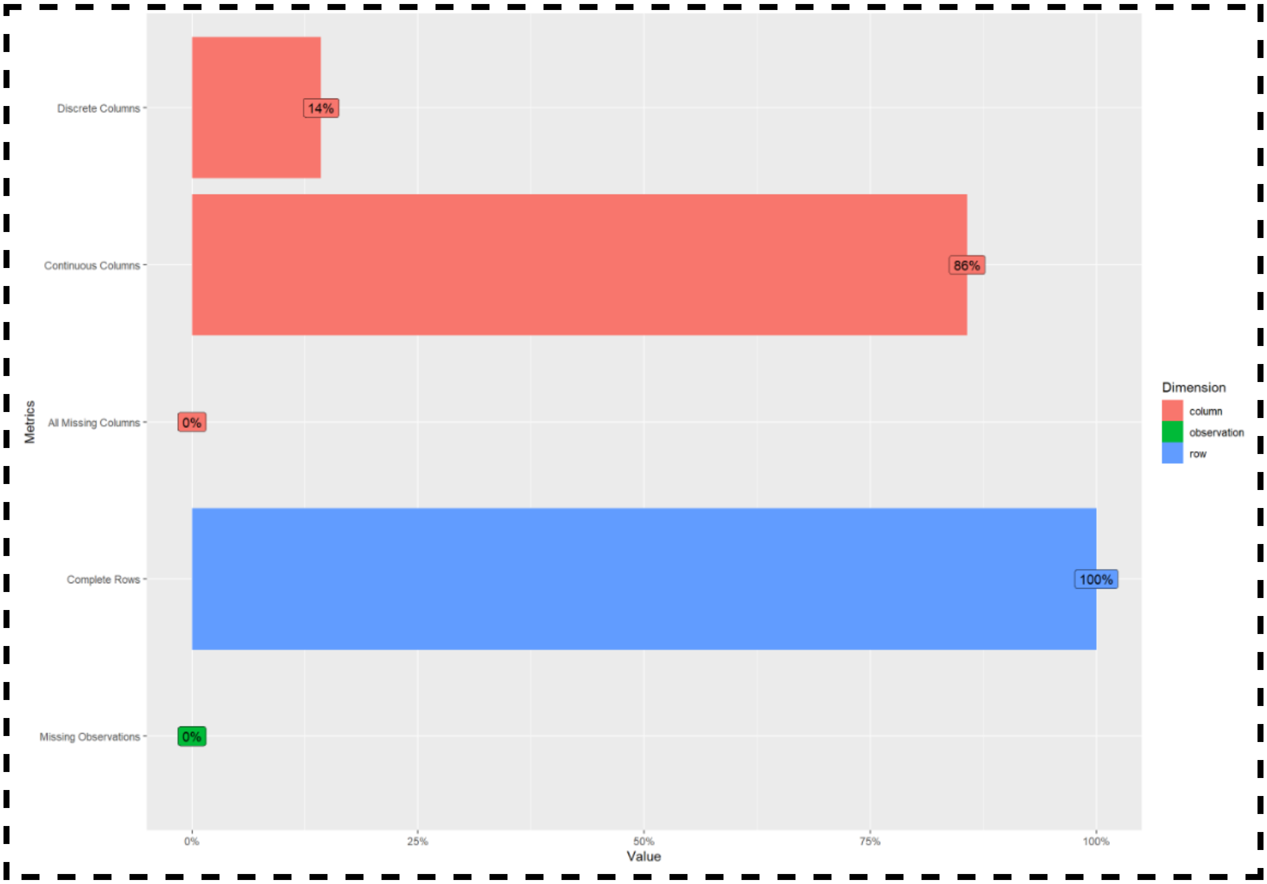
Dataset Statistics(1/2)

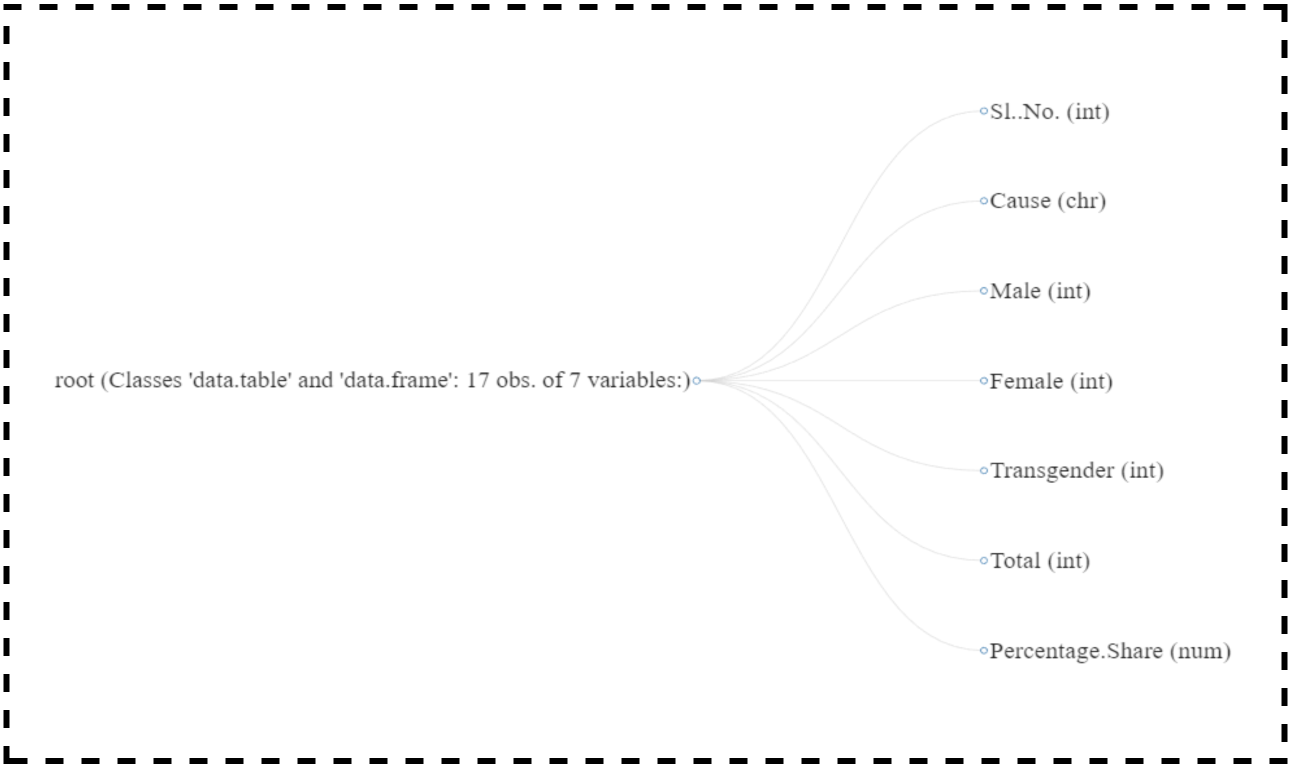
Code:



```
#Description of Dataset--> data1_cause
data1_cause <- read.csv("C:/Users/kumar/Documents/Datasets/Cause-wise NCRB_ADSI-2021_Table_2.12.csv")
describe(data1_cause)
head(data1_cause)
summary(data1_cause)

install.packages("skimr")
library(skimr)
skim(data1_cause)

install.packages("DataExplorer")
library(DataExplorer)
DataExplorer::create_report(data1_cause)
```

Output:




root (Classes 'data.table' and 'data.frame': 17 obs. of 7 variables:)
Sl..No. (int)
Cause (chr)
Male (int)
Female (int)
Transgender (int)
Total (int)
Percentage.Share (num)

Description of Data Set(2/2)

Name of the Data set: State/UTs-wise Educational Status Distribution of Suicides during 2021

Link to the data set: https://data.gov.in/resource/stateuts-wise-educational-status-distribution-suicides-during-2021

Source: Open Government Data(OGD) Platform India

Released under: National Data Sharing and Accessibility Policy(NDSAP)

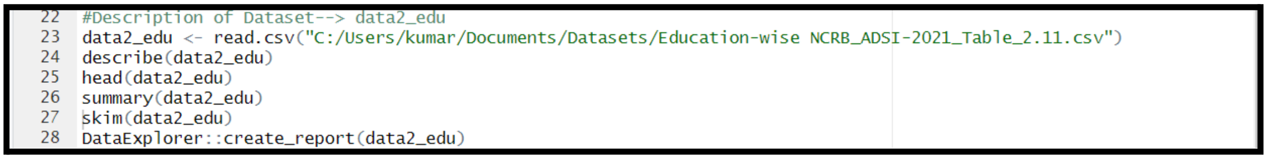
Contributor: Ministry of Home Affairs , Department of States , National Crime Records Bureau (NCRB)

Description: The State/UTs-wise Educational Status Distribution of Suicides during 2021 dataset provides information on the number of suicides committed in India in 2021, broken down by state/UT and educational level. The data was collected by the National Crime Records Bureau (NCRB), a department of the Indian Ministry of Home Affairs, and released under the National Data Sharing and Accessibility Policy (NDSAP).

The dataset includes information on the number of suicides committed by individuals with different educational backgrounds, such as those with no education, primary education, secondary education, and others. The data is further classified by state/UT, and gender.

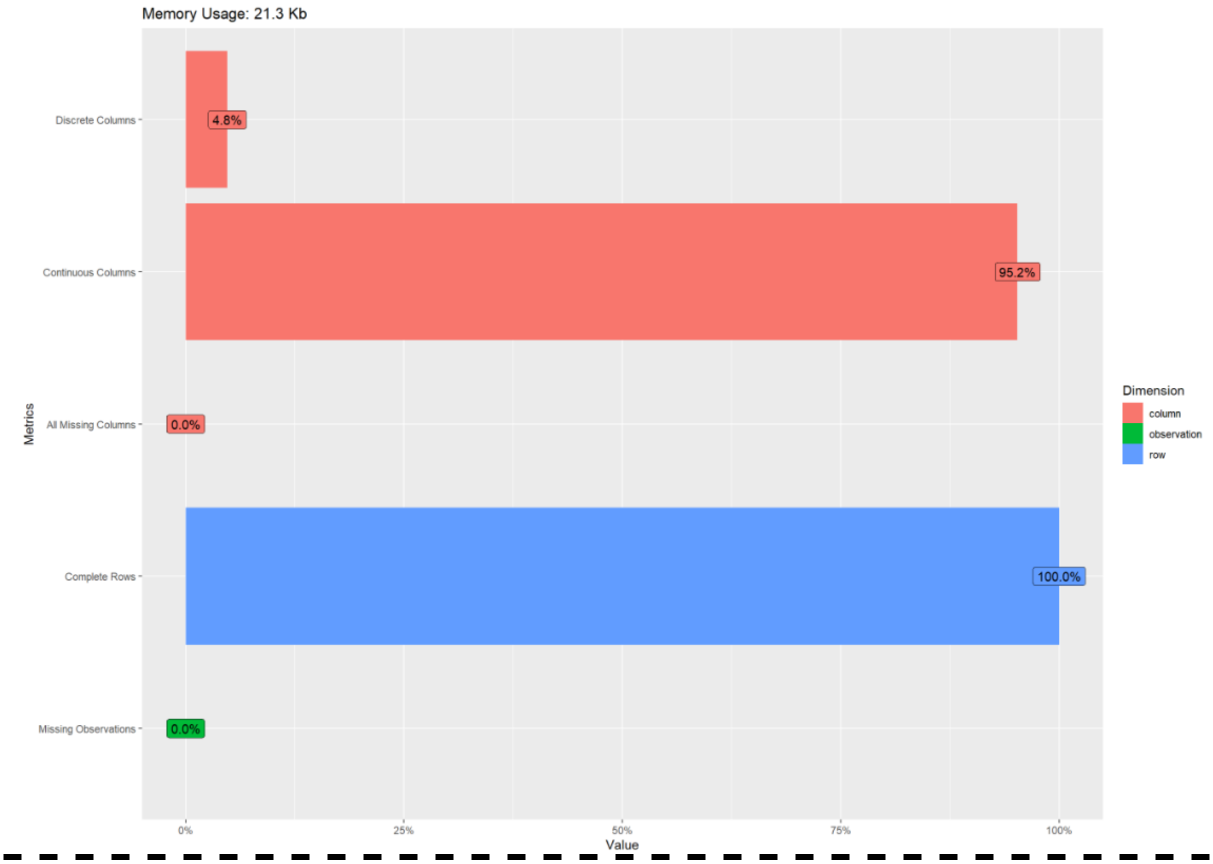
Dataset Statistics(2/2)

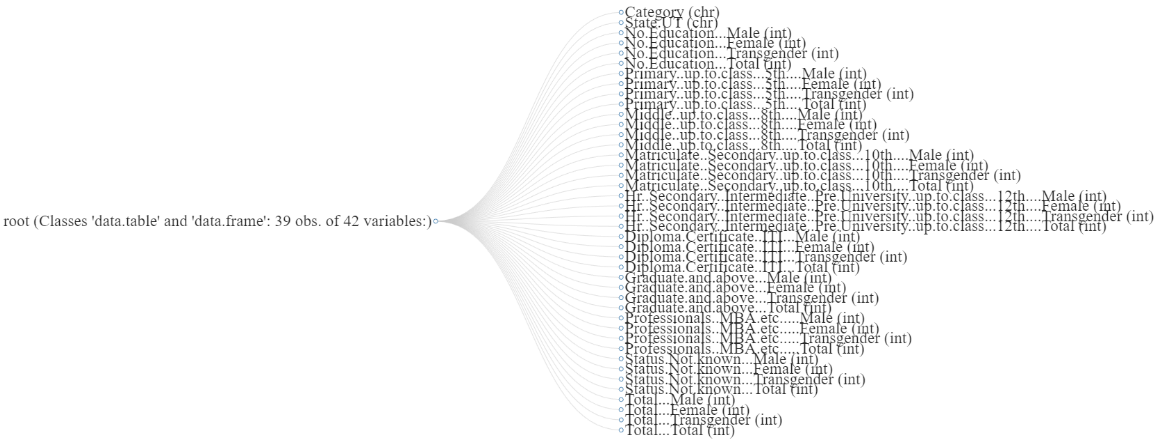
Code:



```
#Description of Dataset--> data2_edu
data2_edu <- read.csv("C:/Users/kumar/Documents/Datasets/Education-wise NCRB_ADSI-2021_Table_2.11.csv")
describe(data2_edu)
head(data2_edu)
summary(data2_edu)
skim(data2_edu)
DataExplorer::create_report(data2_edu)
```

Output:

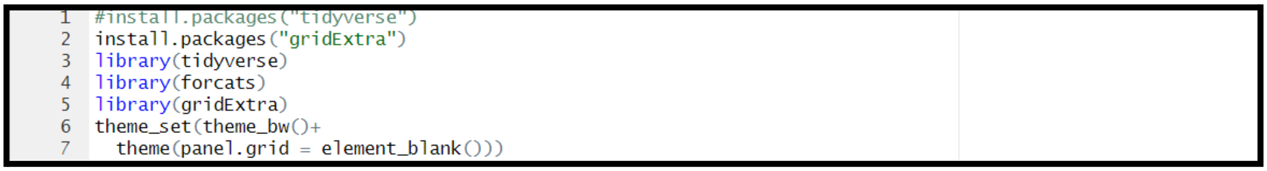



root (Classes 'data.table' and 'data.frame': 39 obs. of 42 variables:)
Category (chr)
State.UT (chr)
No.Education...Male (int)
No.Education...Female (int)
No.Education...Transgender (int)
No.Education...Total (int)
Primary..up.to.class...5th....Male (int)
Primary..up.to.class...5th....Female (int)
Primary..up.to.class...5th....Transgender (int)
Primary..up.to.class...5th....Total (int)
Middle..up.to.class...8th....Male (int)
Middle..up.to.class...8th....Female (int)
Middle..up.to.class...8th....Transgender (int)
Middle..up.to.class...8th....Total (int)
Matriculate..Secondary..up.to.class...10th....Male (int)
Matriculate..Secondary..up.to.class...10th....Female (int)
Matriculate..Secondary..up.to.class...10th....Transgender (int)
Matriculate..Secondary..up.to.class...10th....Total (int)
Hr..Secondary..Intermediate..Pre.University..up.to.class...12th....Male (int)
Hr..Secondary..Intermediate..Pre.University..up.to.class...12th....Female (int)
Hr..Secondary..Intermediate..Pre.University..up.to.class...12th....Transgender (int)
Hr..Secondary..Intermediate..Pre.University..up.to.class...12th....Total (int)
Diploma.Certificate..ITI...Male (int)
Diploma.Certificate..ITI...Female (int)
Diploma.Certificate..ITI...Transgender (int)
Diploma.Certificate..ITI...Total (int)
Graduate.and.above...Male (int)
Graduate.and.above...Female (int)
Graduate.and.above...Transgender (int)
Graduate.and.above...Total (int)
Professionals..MBA.etc.....Male (int)
Professionals..MBA.etc.....Female (int)
Professionals..MBA.etc.....Transgender (int)
Professionals..MBA.etc.....Total (int)
Status.Not.known...Male (int)
Status.Not.known...Female (int)
Status.Not.known...Transgender (int)
Status.Not.known...Total (int)
Total...Male (int)
Total...Female (int)
Total...Transgender (int)
Total...Total (int)

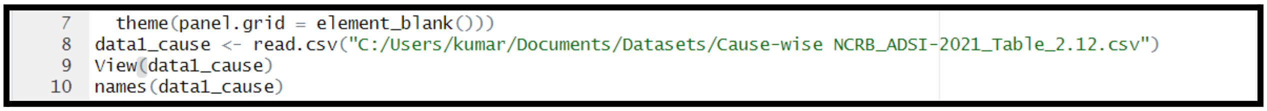
Data Collection

(Dataset 1/2)

Code:

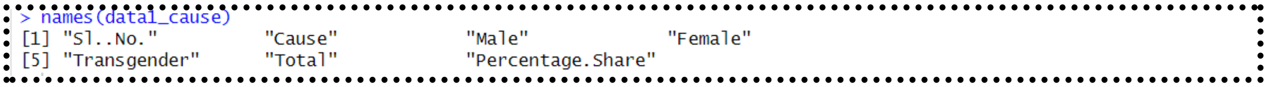


```r
#install.packages("tidyverse")
install.packages("gridExtra")
library(tidyverse)
library(forcats)
library(gridExtra)
theme_set(theme_bw()+
  theme(panel.grid = element_blank()))
```



```r
  theme(panel.grid = element_blank()))
data1_cause <- read.csv("C:/Users/kumar/Documents/Datasets/Cause-wise NCRB_ADSI-2021_Table_2.12.csv")
View(data1_cause)
names(data1_cause)
```

Output:



```
> names(data1_cause)
[1] "Sl..No."           "Cause"             "Male"              "Female"
[5] "Transgender"       "Total"             "Percentage.Share"
```

Inferences:

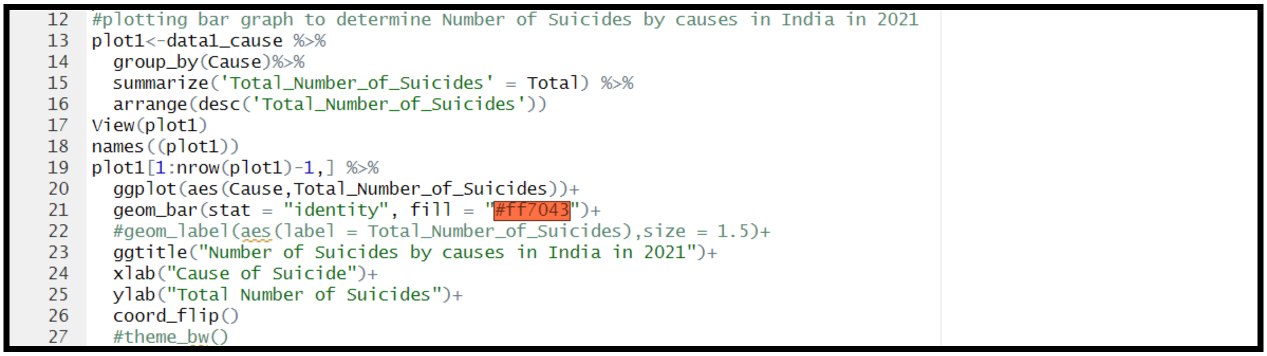
- **No. of Attributes: 7**
- **No. of Records: 17**
- **No. of “discrete” attributes: 1**
- **No. of “continuous” attributes: 6**
- **Total No of observations: 119**

Data Visualization

Data visualization is the graphical representation of information and data. It uses statistical graphics, plots, information graphics and other tools to communicate information clearly and efficiently.

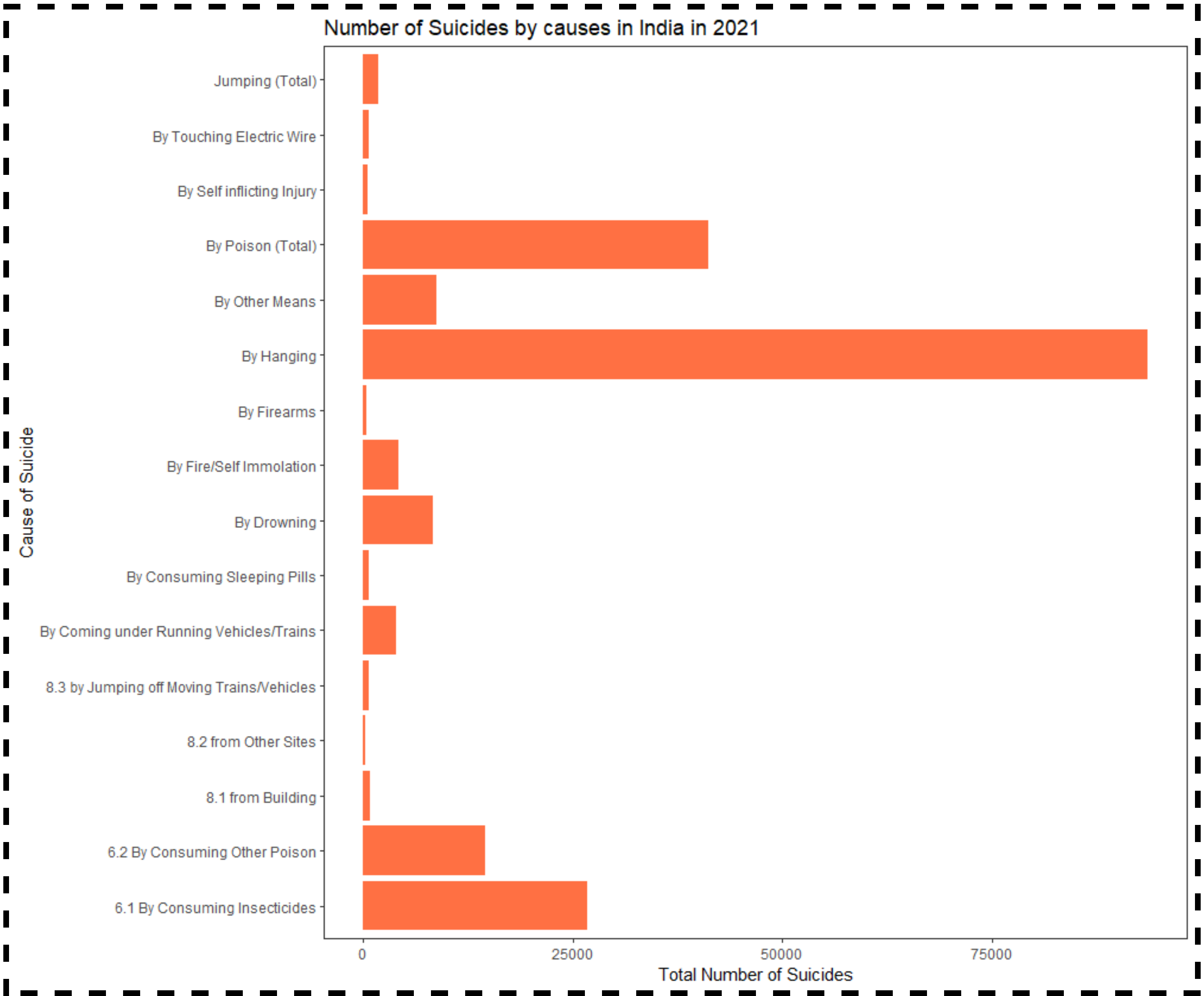
****

Code:



```
#plotting bar graph to determine Number of Suicides by causes in India in 2021
plot1<-data1_cause %>%
  group_by(Cause)%>%
  summarize('Total_Number_of_Suicides' = Total) %>%
  arrange(desc('Total_Number_of_Suicides'))
View(plot1)
names((plot1))
plot1[1:nrow(plot1)-1,] %>%
  ggplot(aes(Cause,Total_Number_of_Suicides))+
  geom_bar(stat = "identity", fill = "#ff7043")+
  #geom_label(aes(label = Total_Number_of_Suicides),size = 1.5)+
  ggtitle("Number of Suicides by causes in India in 2021")+
  xlab("Cause of Suicide")+
  ylab("Total Number of Suicides")+
  coord_flip()
  #theme_bw()
```

Output:

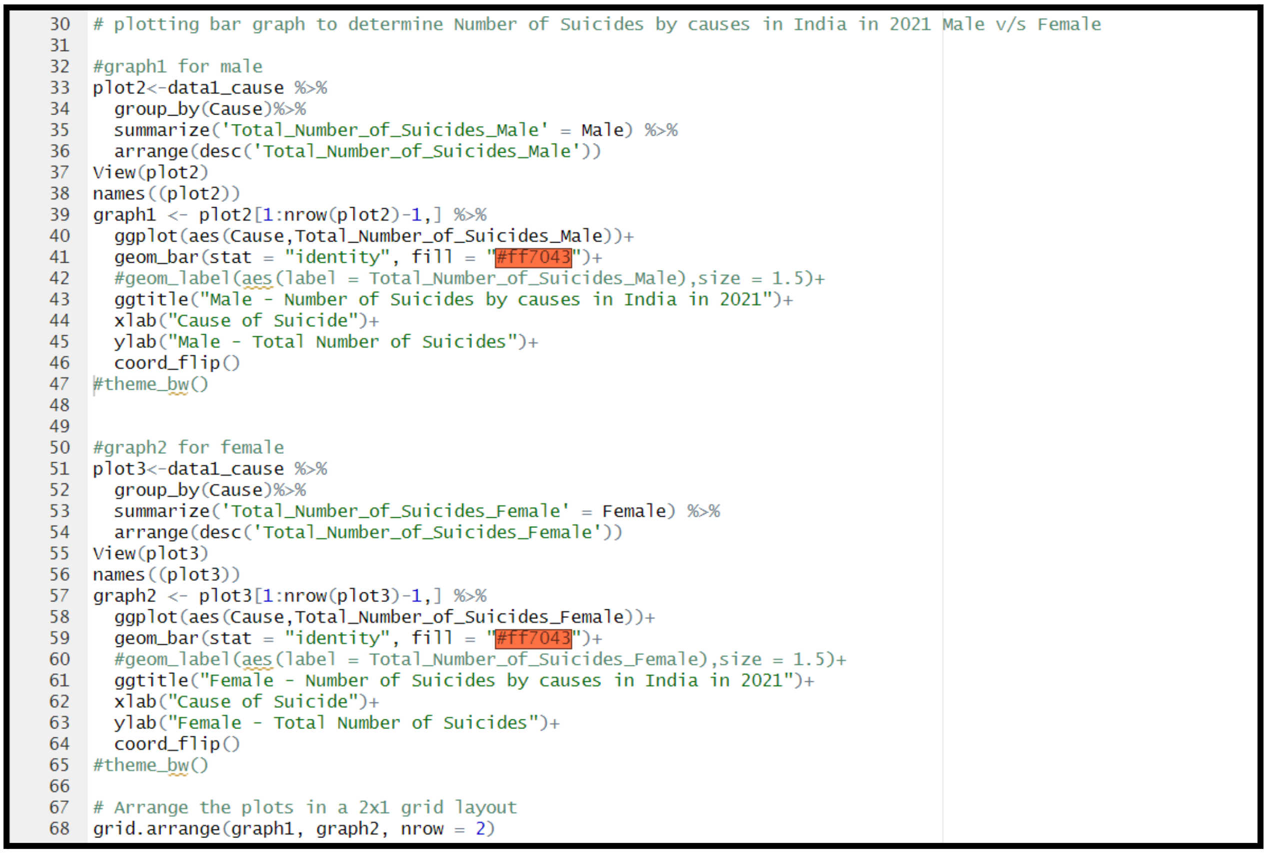


Inferences:

The above graph represents the number of suicides in India in 2021 categorised by the cause by which the suicide was commınted. By the bar graph we can make out that “Hanging to death” is the most practiced mode of suicide. And “jumping off Moving Trains/Vehicles”, “suicide from building”, “suicide from other sites” are the least practiced mode of suicide.

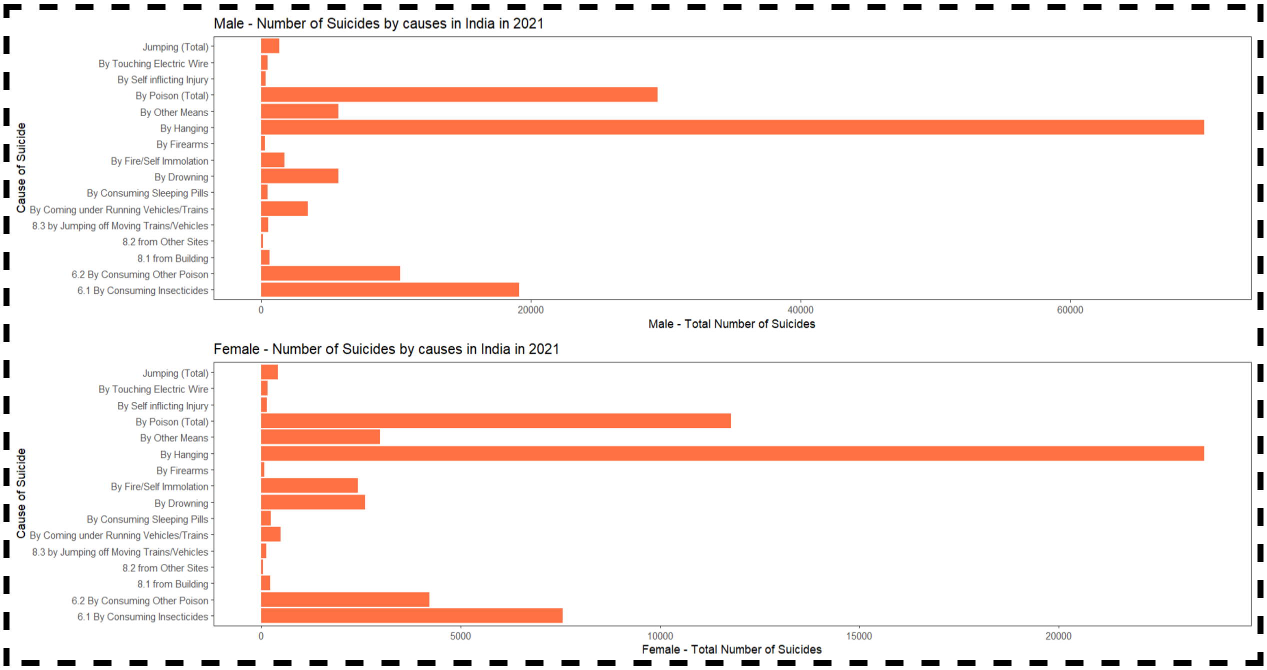
****

Code:

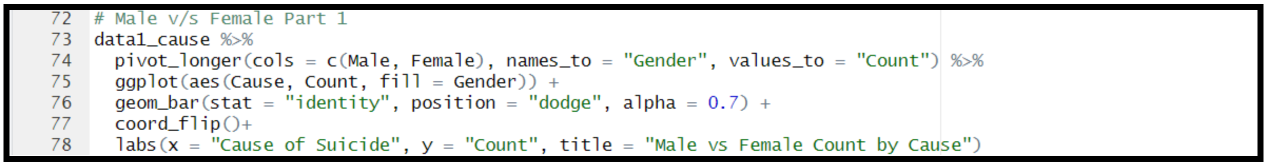


```
# plotting bar graph to determine Number of Suicides by causes in India in 2021 Male v/s Female

#graph1 for male
plot2<-data1_cause %>%
  group_by(Cause)%>%
  summarize('Total_Number_of_Suicides_Male' = Male) %>%
  arrange(desc('Total_Number_of_Suicides_Male'))
View(plot2)
names((plot2))
graph1 <- plot2[1:nrow(plot2)-1,] %>%
  ggplot(aes(Cause,Total_Number_of_Suicides_Male))+
  geom_bar(stat = "identity", fill = "#ff7043")+
  #geom_label(aes(label = Total_Number_of_Suicides_Male),size = 1.5)+
  ggtitle("Male - Number of Suicides by causes in India in 2021")+
  xlab("Cause of Suicide")+
  ylab("Male - Total Number of Suicides")+
  coord_flip()
#theme_bw()


#graph2 for female
plot3<-data1_cause %>%
  group_by(Cause)%>%
  summarize('Total_Number_of_Suicides_Female' = Female) %>%
  arrange(desc('Total_Number_of_Suicides_Female'))
View(plot3)
names((plot3))
graph2 <- plot3[1:nrow(plot3)-1,] %>%
  ggplot(aes(Cause,Total_Number_of_Suicides_Female))+
  geom_bar(stat = "identity", fill = "#ff7043")+
  #geom_label(aes(label = Total_Number_of_Suicides_Female),size = 1.5)+
  ggtitle("Female - Number of Suicides by causes in India in 2021")+
  xlab("Cause of Suicide")+
  ylab("Female - Total Number of Suicides")+
  coord_flip()
#theme_bw()

# Arrange the plots in a 2x1 grid layout
grid.arrange(graph1, graph2, nrow = 2)
```

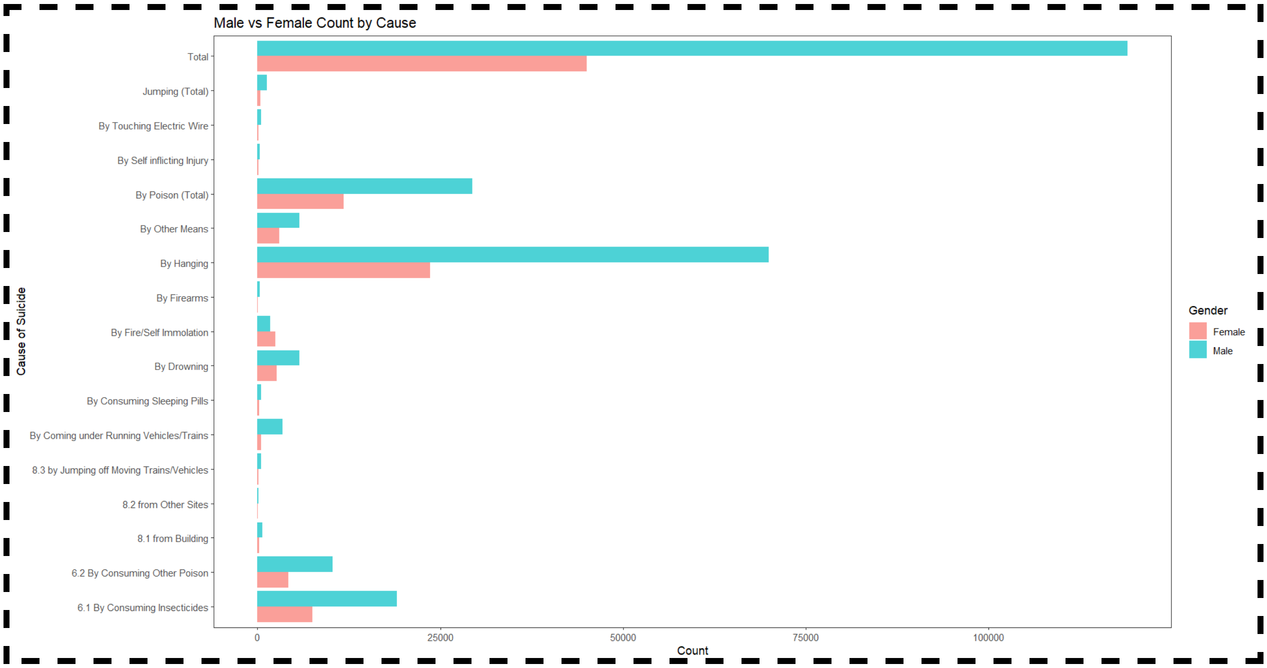
Output:



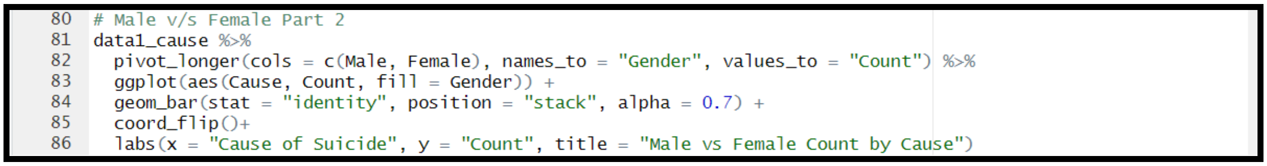
Code:



```r
# Male v/s Female Part 1
data1_cause %>%
  pivot_longer(cols = c(Male, Female), names_to = "Gender", values_to = "Count") %>%
  ggplot(aes(Cause, Count, fill = Gender)) +
  geom_bar(stat = "identity", position = "dodge", alpha = 0.7) +
  coord_flip()+
  labs(x = "Cause of Suicide", y = "Count", title = "Male vs Female Count by Cause")
```

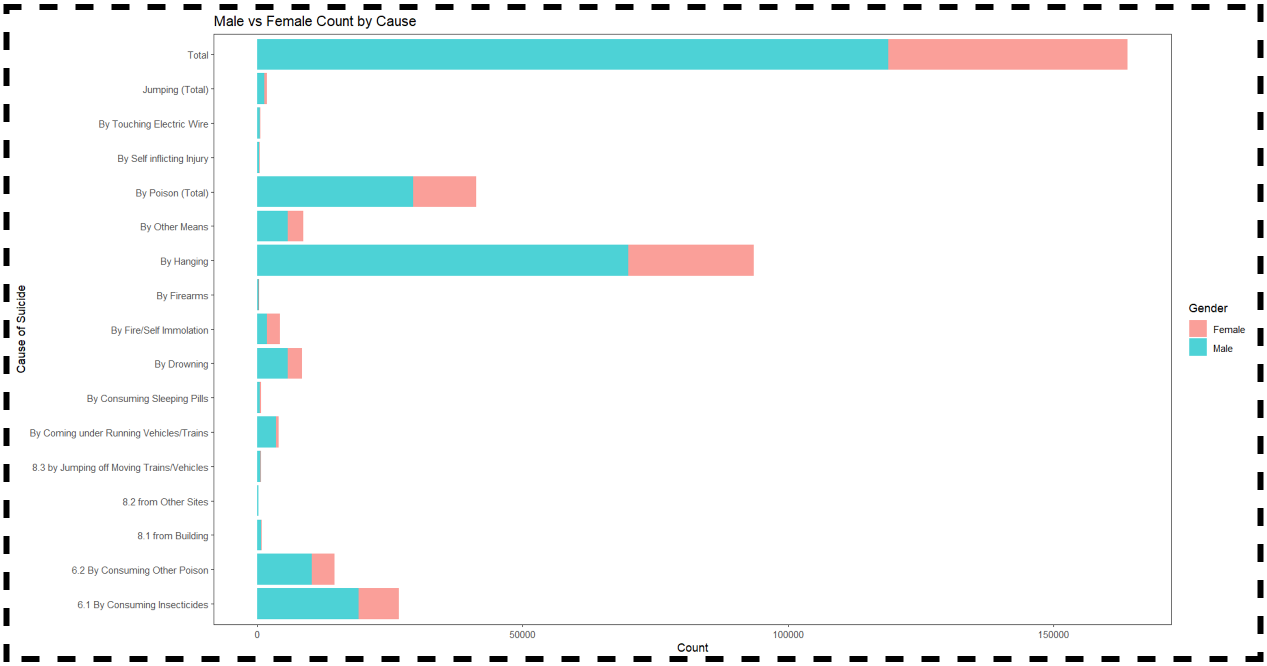


Code:



```
# Male v/s Female Part 2
data1_cause %>%
  pivot_longer(cols = c(Male, Female), names_to = "Gender", values_to = "Count") %>%
  ggplot(aes(Cause, Count, fill = Gender)) +
  geom_bar(stat = "identity", position = "stack", alpha = 0.7) +
  coord_flip()+
  labs(x = "Cause of Suicide", y = "Count", title = "Male vs Female Count by Cause")
```

Output:



Inferences:

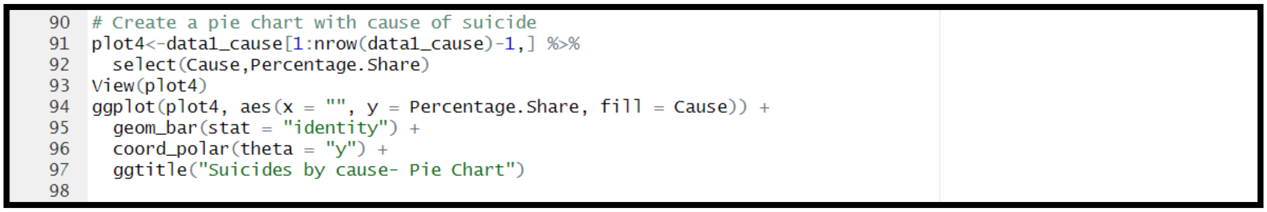
The above three graphs compares the Male and Female suicides with three different perspectives. The first graph shows Male v/s Female Suicide cause side by side, the second graph shows Male v/s Female Suicide cause compared individually and the last graph shows the comparison of Male v/s Female suicide cause, Male and Female count stacked over eachother.

We can make out the following inferences from the graphs:

- Both male and female have committed the most suicides by “hanging”.
- In males, “suicide by jumping off Moving Trains/Vehicles”, “suicide from building”, “suicide from other sites” are the least practiced mode of suicide.
- Whereas in females, “suicide by firearms” and “suicide by touching electric wires” are the least practiced mode of suicide.
- There are less half females suicides than male suicides.

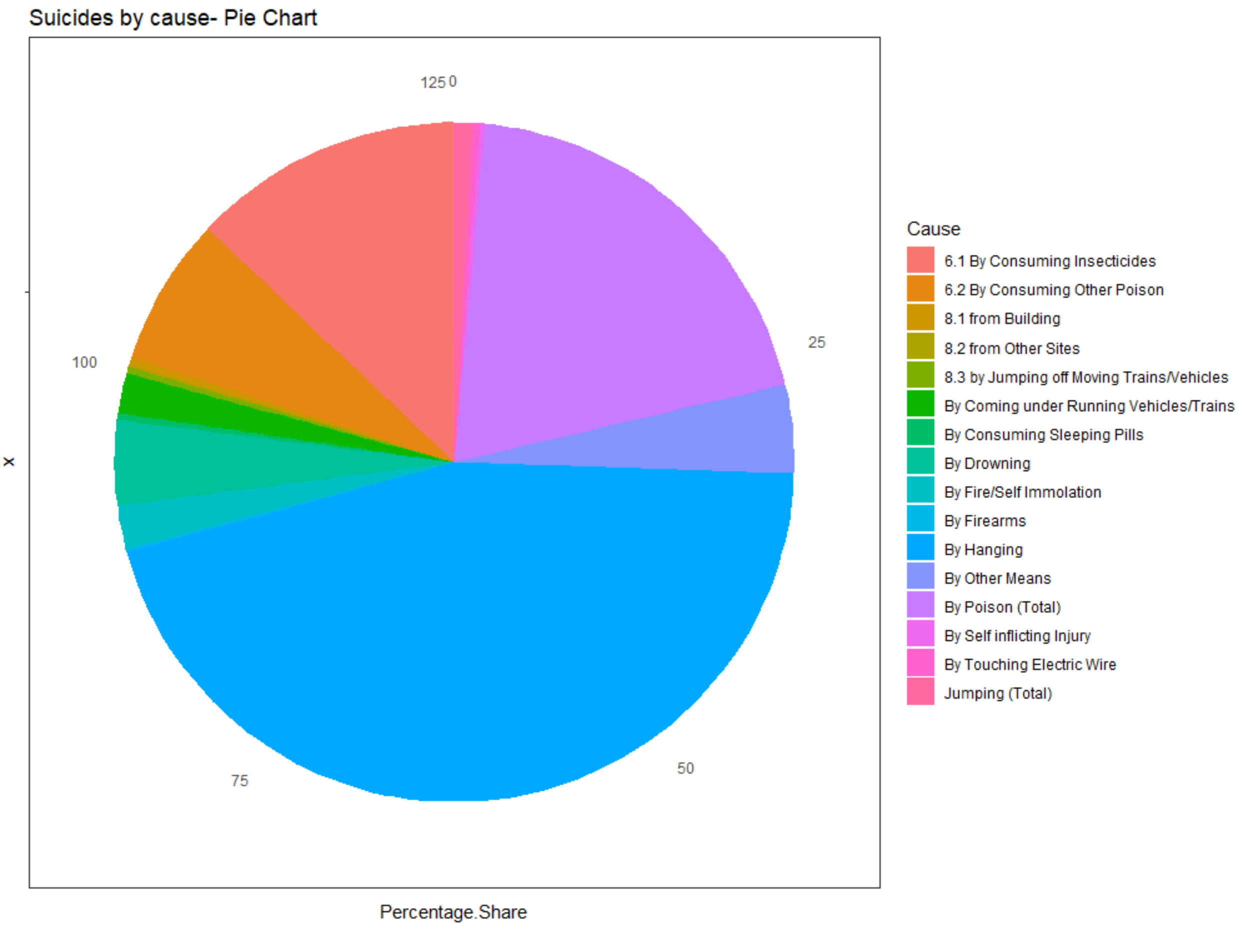
****

Code:



```
# Create a pie chart with cause of suicide
plot4<-data1_cause[1:nrow(data1_cause)-1,] %>%
  select(Cause,Percentage.Share)
View(plot4)
ggplot(plot4, aes(x = "", y = Percentage.Share, fill = Cause)) +
  geom_bar(stat = "identity") +
  coord_polar(theta = "y") +
  ggtitle("Suicides by cause- Pie Chart")

```

Output:



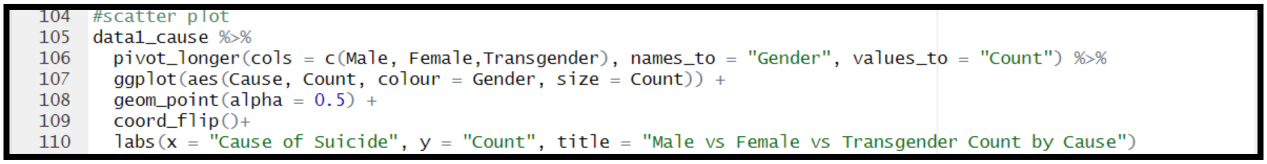
Inferences:

The above figure is the Pie Chart of the Suicides Causes in India in 2021.

We can make out from the above Pie chart that “suicides by hanging” is the most practiced mode of suicide followed by “suicide by poision” and “suicide by insecticides”.

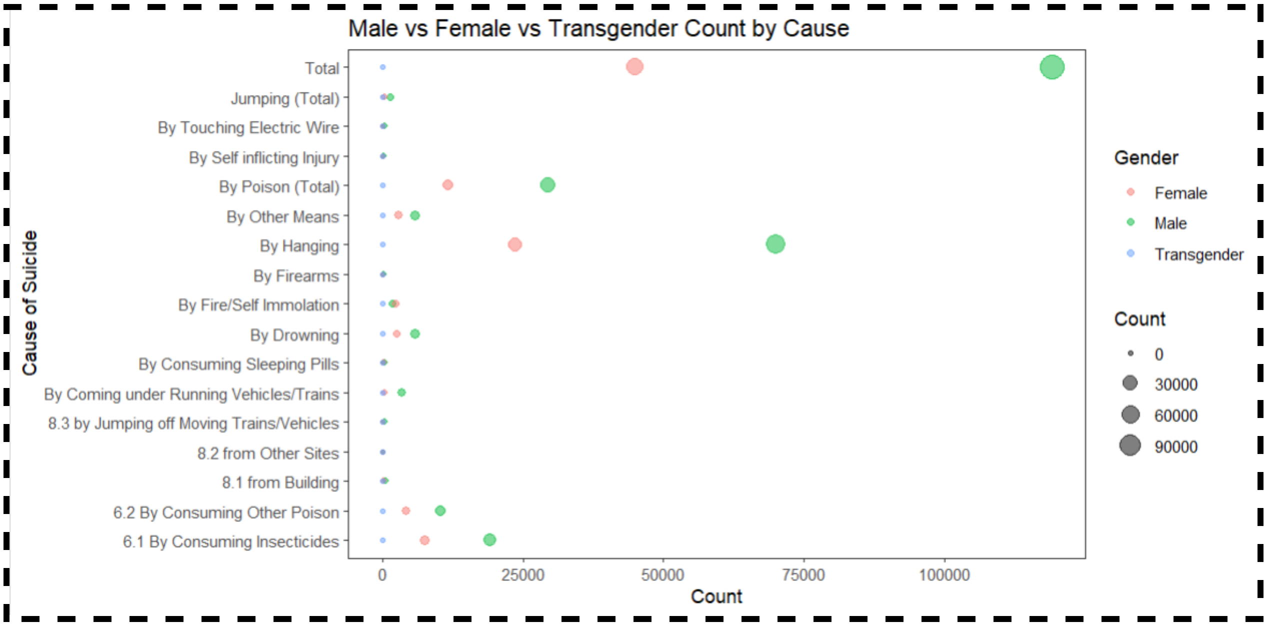
****

Code:



```
#scatter plot
data1_cause %>%
  pivot_longer(cols = c(Male, Female,Transgender), names_to = "Gender", values_to = "Count") %>%
  ggplot(aes(Cause, Count, colour = Gender, size = Count)) +
  geom_point(alpha = 0.5) +
  coord_flip()+
  labs(x = "Cause of Suicide", y = "Count", title = "Male vs Female vs Transgender Count by Cause")
```

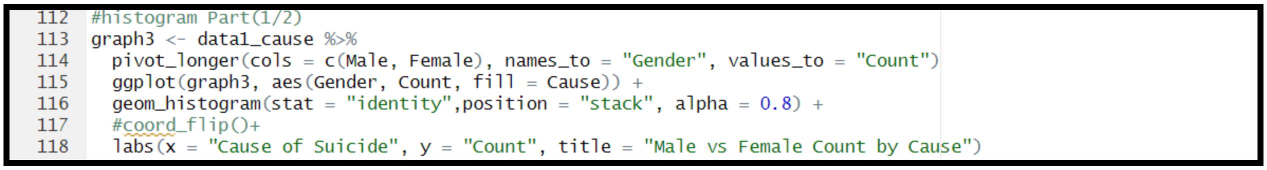
Output:



The above scatter-plot graph represents the number of cases in each category of cause of the suicide. The comparison in the above plot is done among the Male, female and transgender. The Point Size of the points are determined by the Count of cases under each category of the ‘Cause of suicide’.

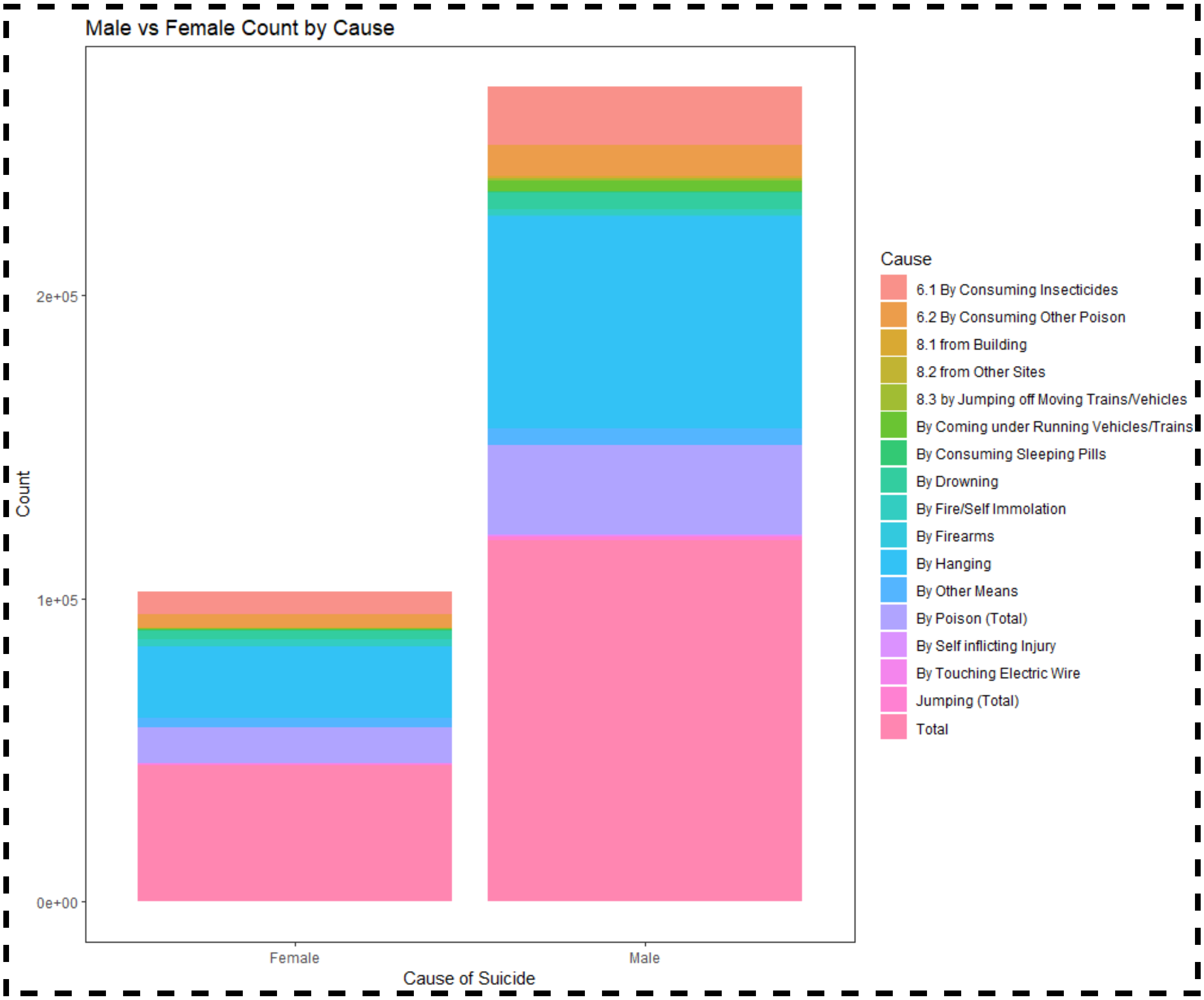
****

Code:



```
#histogram Part(1/2)
graph3 <- data1_cause %>%
  pivot_longer(cols = c(Male, Female), names_to = "Gender", values_to = "Count")
  ggplot(graph3, aes(Gender, Count, fill = Cause)) +
  geom_histogram(stat = "identity",position = "stack", alpha = 0.8) +
  #coord_flip()+
  labs(x = "Cause of Suicide", y = "Count", title = "Male vs Female Count by Cause")
```

Output:

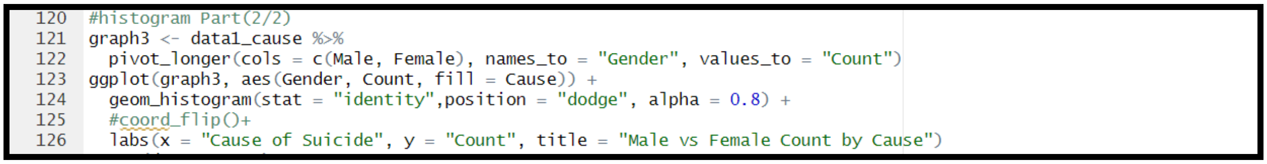


Inferences:

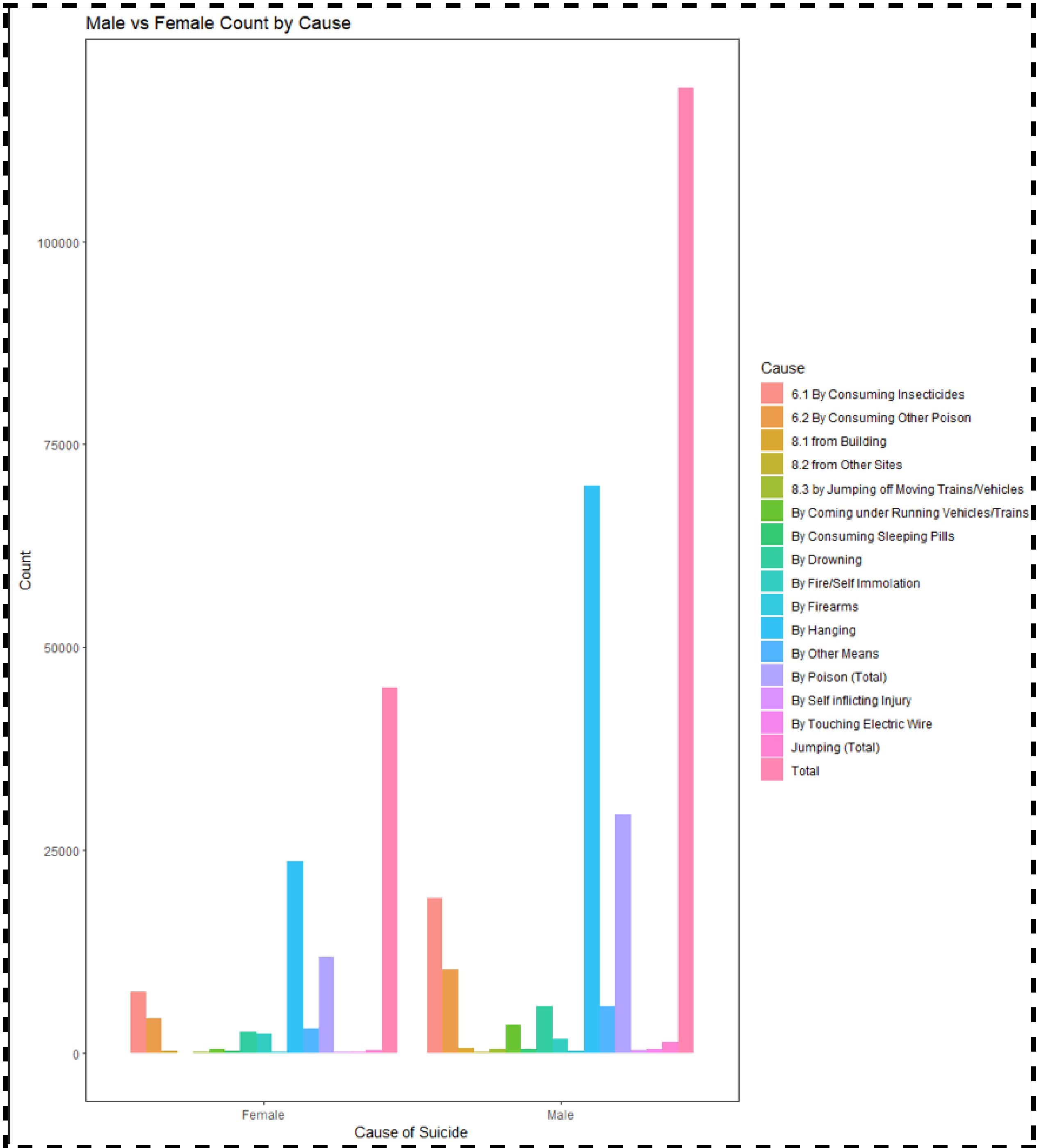
The above graph shows the histogram plot of the suicide cases in India in 2021 Male v/s Female. In the graph the cause of the suicide are differentiated by the colour of the stacked bars. Comparison can made through the graph b/w the Male and Female suicide cases. The graph clearly shows that male total suicides are more than double the female suicides. We can also compare the other category of cause of suicide.

****

Code:



```
#histogram Part(2/2)
graph3 <- data1_cause %>%
  pivot_longer(cols = c(Male, Female), names_to = "Gender", values_to = "Count")
ggplot(graph3, aes(Gender, Count, fill = Cause)) +
  geom_histogram(stat = "identity",position = "dodge", alpha = 0.8) +
  #coord_flip()+
  labs(x = "Cause of Suicide", y = "Count", title = "Male vs Female Count by Cause")
```

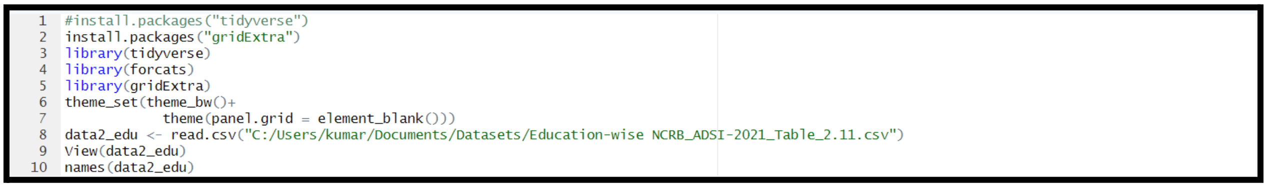


The above graph is the ‘dodge’ version of the previous histogram graph.

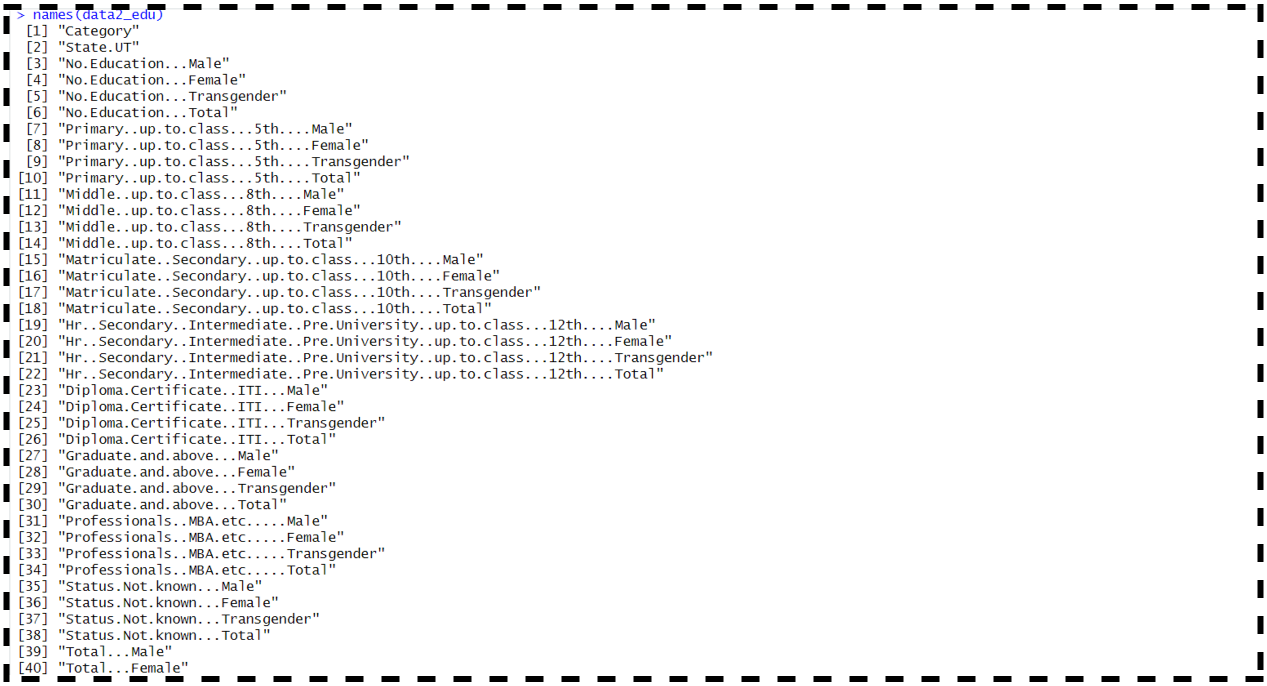
****

Data Collection

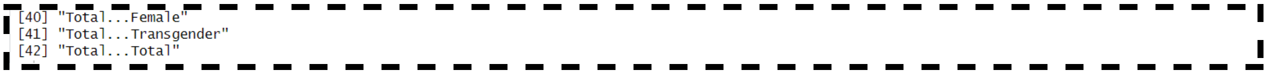
(Dataset 2/2)



```r
#install.packages("tidyverse")
install.packages("gridExtra")
library(tidyverse)
library(forcats)
library(gridExtra)
theme_set(theme_bw()+
            theme(panel.grid = element_blank()))
data2_edu <- read.csv("C:/Users/kumar/Documents/Datasets/Education-wise NCRB_ADSI-2021_Table_2.11.csv")
View(data2_edu)
names(data2_edu)
```



```
> names(data2_edu)
 [1] "Category"
 [2] "State.UT"
 [3] "No.Education...Male"
 [4] "No.Education...Female"
 [5] "No.Education...Transgender"
 [6] "No.Education...Total"
 [7] "Primary..up.to.class...5th....Male"
 [8] "Primary..up.to.class...5th....Female"
 [9] "Primary..up.to.class...5th....Transgender"
[10] "Primary..up.to.class...5th....Total"
[11] "Middle..up.to.class...8th....Male"
[12] "Middle..up.to.class...8th....Female"
[13] "Middle..up.to.class...8th....Transgender"
[14] "Middle..up.to.class...8th....Total"
[15] "Matriculate..Secondary..up.to.class...10th....Male"
[16] "Matriculate..Secondary..up.to.class...10th....Female"
[17] "Matriculate..Secondary..up.to.class...10th....Transgender"
[18] "Matriculate..Secondary..up.to.class...10th....Total"
[19] "Hr..Secondary..Intermediate..Pre.University..up.to.class...12th....Male"
[20] "Hr..Secondary..Intermediate..Pre.University..up.to.class...12th....Female"
[21] "Hr..Secondary..Intermediate..Pre.University..up.to.class...12th....Transgender"
[22] "Hr..Secondary..Intermediate..Pre.University..up.to.class...12th....Total"
[23] "Diploma.Certificate..ITI...Male"
[24] "Diploma.Certificate..ITI...Female"
[25] "Diploma.Certificate..ITI...Transgender"
[26] "Diploma.Certificate..ITI...Total"
[27] "Graduate.and.above...Male"
[28] "Graduate.and.above...Female"
[29] "Graduate.and.above...Transgender"
[30] "Graduate.and.above...Total"
[31] "Professionals..MBA.etc.....Male"
[32] "Professionals..MBA.etc.....Female"
[33] "Professionals..MBA.etc.....Transgender"
[34] "Professionals..MBA.etc.....Total"
[35] "Status.Not.known...Male"
[36] "Status.Not.known...Female"
[37] "Status.Not.known...Transgender"
[38] "Status.Not.known...Total"
[39] "Total...Male"
[40] "Total...Female"
```



```
[40] "Total...Female"
[41] "Total...Transgender"
[42] "Total...Total"
```

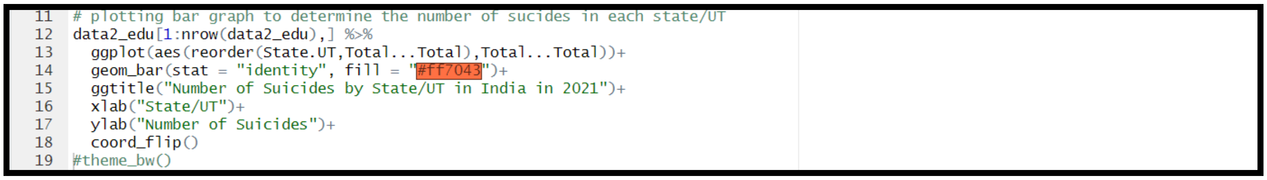
Inferences:

- **No. of Attributes: 42**
- **No. of Records: 39**
- **No. of “discrete” attributes: 2**
- **No. of “continuous” attributes: 2**

- **Total No of observations: 1638**

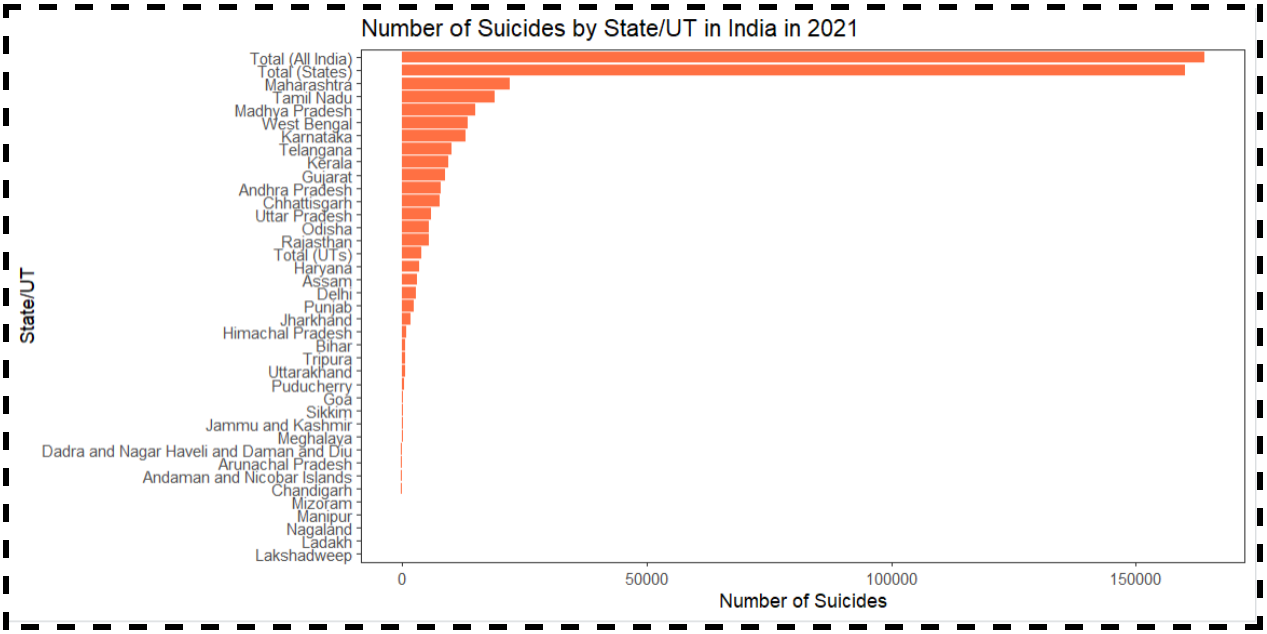
****

Code:



```
# plotting bar graph to determine the number of sucides in each state/UT
data2_edu[1:nrow(data2_edu),] %>%
  ggplot(aes(reorder(State.UT,Total...Total),Total...Total))+
  geom_bar(stat = "identity", fill = "#ff7043")+
  ggtitle("Number of Suicides by State/UT in India in 2021")+
  xlab("State/UT")+
  ylab("Number of Suicides")+
  coord_flip()
#theme_bw()
```

Output:

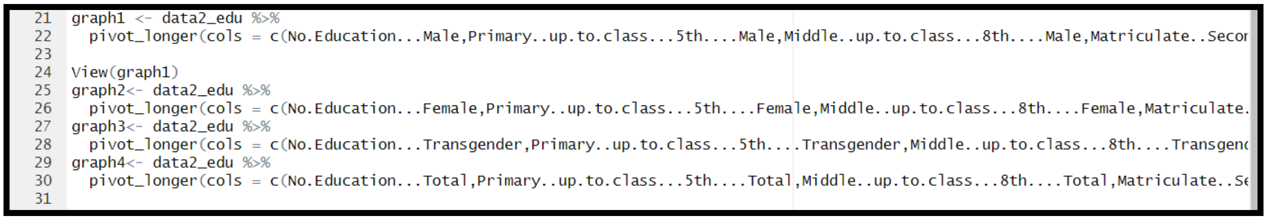


Inferences:

The above graph shows the number of suicides against each states/union territory in India in 2021. From the bar graph we can make out following inferences:

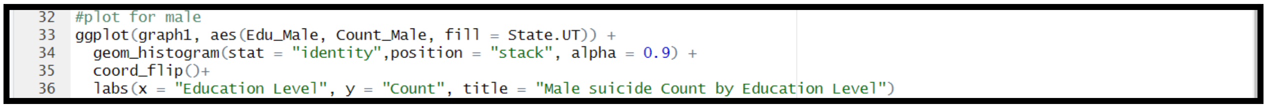
- Maharashtra state has the most number of suicide cases in 2021, followed by Tamil Nadu, Madhya Pradesh, West Bengal, Karnataka...
- Among the states Nagaland has the least number of suicides, followed by Mizoram.
- Among the UTs Delhi has the most number of suicides.
- Among the UTs Lakshadweep has the least number of suicides, followed by Ladakh and Chandigarh.

****



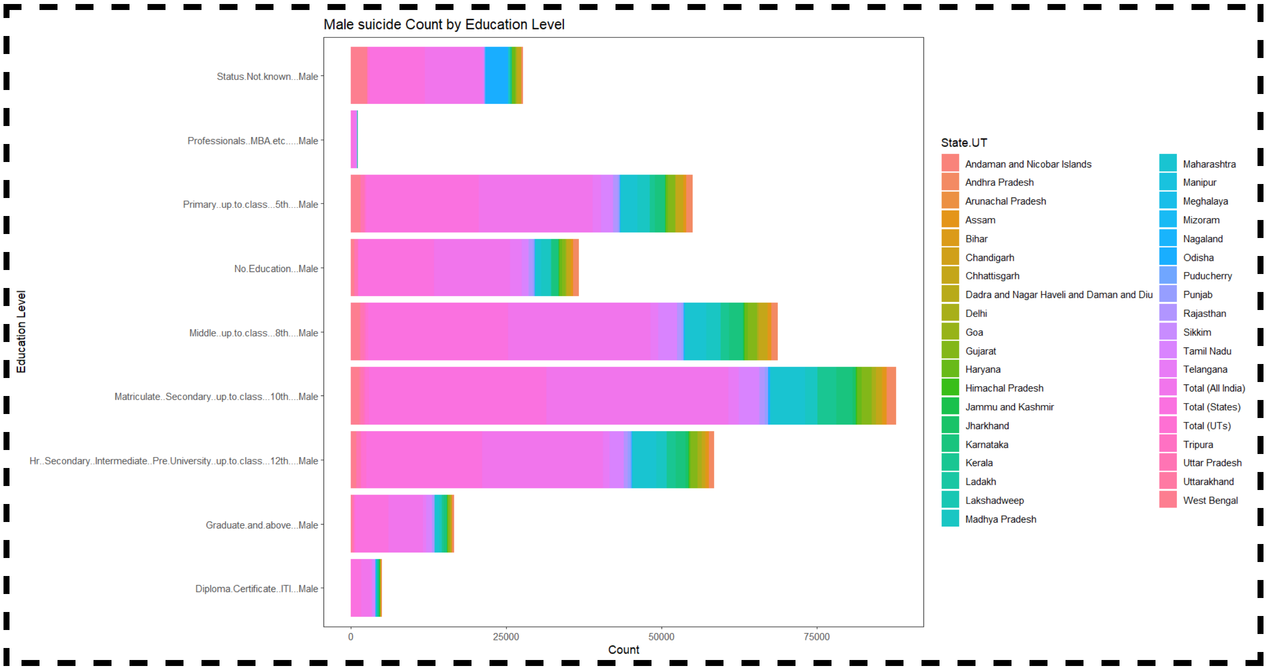
```
graph1 <- data2_edu %>%
  pivot_longer(cols = c(No.Education...Male,Primary..up.to.class...5th....Male,Middle..up.to.class...8th....Male,Matriculate..Secor

View(graph1)
graph2<- data2_edu %>%
  pivot_longer(cols = c(No.Education...Female,Primary..up.to.class...5th....Female,Middle..up.to.class...8th....Female,Matriculate.
graph3<- data2_edu %>%
  pivot_longer(cols = c(No.Education...Transgender,Primary..up.to.class...5th....Transgender,Middle..up.to.class...8th....Transgenc
graph4<- data2_edu %>%
  pivot_longer(cols = c(No.Education...Total,Primary..up.to.class...5th....Total,Middle..up.to.class...8th....Total,Matriculate..Se
```

Code:



```
#plot for male
ggplot(graph1, aes(Edu_Male, Count_Male, fill = State.UT)) +
  geom_histogram(stat = "identity",position = "stack", alpha = 0.9) +
  coord_flip()+
  labs(x = "Education Level", y = "Count", title = "Male suicide Count by Education Level")
```

Output:



Inferences:

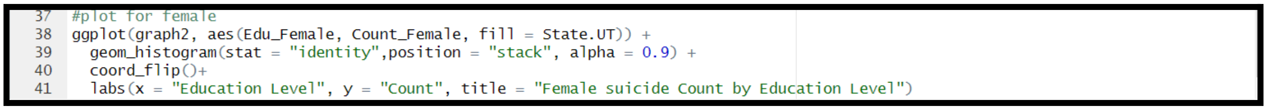
The graph depicting suicide counts among males by education level suggests that those who have completed their secondary education up to class 10th are at a greater risk of suicide than males with higher levels of education. Conversely, males with professional degrees, such as an MBA, appear to have the lowest risk of suicide.

It is important to note that the graph displays variations in suicide rates among states, which are represented by different colors. These variations may reflect underlying socio-economic, cultural, and other contextual factors that influence suicidal behavior in different regions. Further research is necessary to explore the complex interplay of these factors and their relationship with male suicide rates by education level.

****

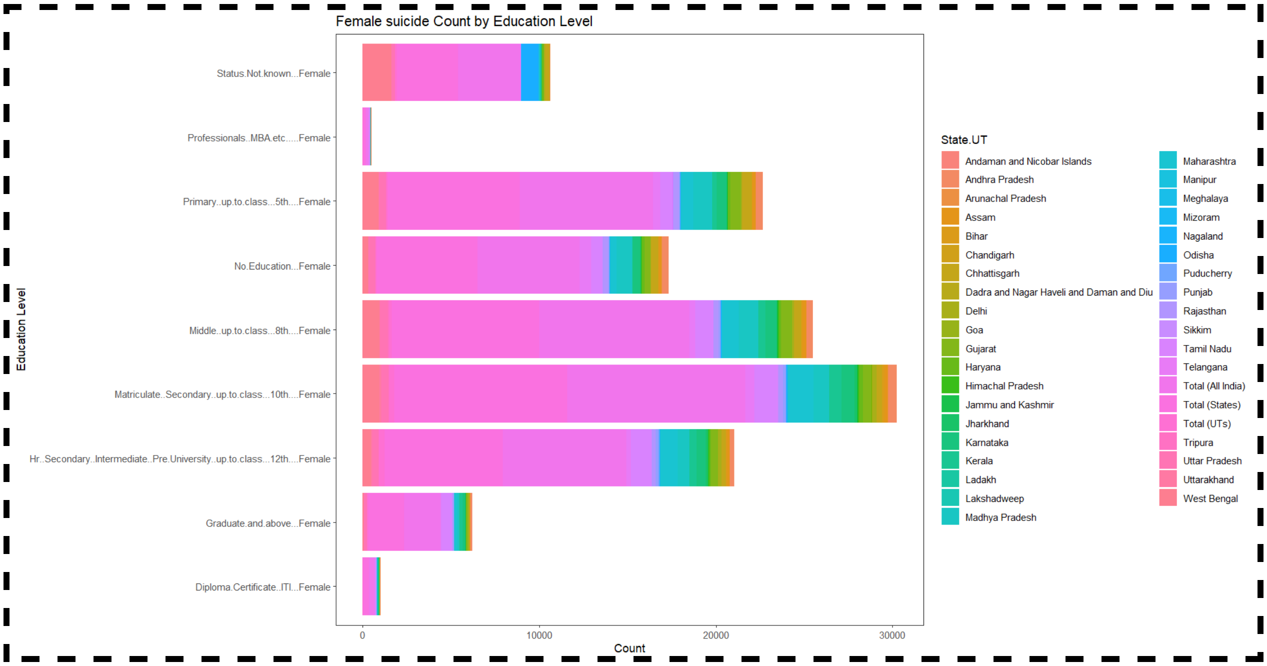
****

Code:



```
#plot for female
ggplot(graph2, aes(Edu_Female, Count_Female, fill = State.UT)) +
  geom_histogram(stat = "identity",position = "stack", alpha = 0.9) +
  coord_flip()+
  labs(x = "Education Level", y = "Count", title = "Female suicide Count by Education Level")
```

Output:



Inferences:

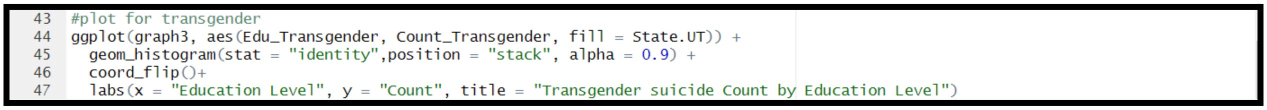
The above graph represents the Female suicide count by Eductaion level. In the graph the various states suicides are represented by the various colours. The following inferences can be made from the above data:

- Most suicides are with those females whose education level is ‘Mariculate Seconday up to class 10th’.
- Least suicides are by the ‘Professionals MBA etc…’ Females.
- We can observe a decreasing trends in suicides as a female goes for higher education or have education level more than ‘Mariculate Seconday up to class 10th’.

****

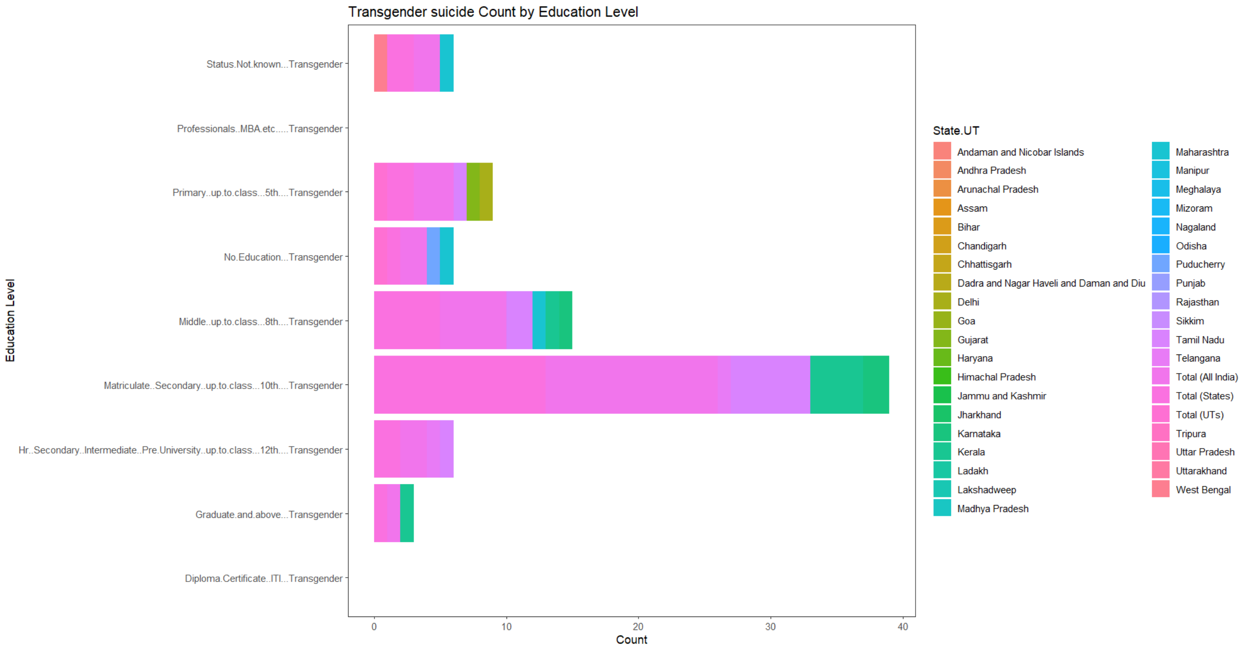
****

Code:



```
#plot for transgender
ggplot(graph3, aes(Edu_Transgender, Count_Transgender, fill = State.UT)) +
  geom_histogram(stat = "identity",position = "stack", alpha = 0.9) +
  coord_flip()+
  labs(x = "Education Level", y = "Count", title = "Transgender suicide Count by Education Level")
```

Output:



Inferences:

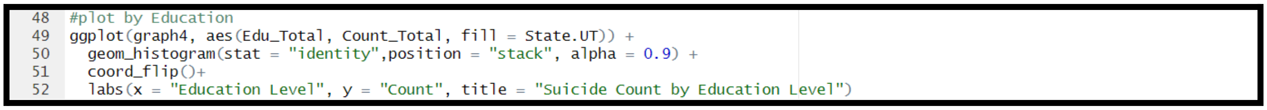
Based on the graph depicting suicide counts among transgender individuals by education level, it can be inferred that those with a secondary education level up to class 10th have the highest incidence of suicide. Conversely, transgender individuals with graduate-level education or higher have the lowest incidence of suicide.

The graph also indicates that suicide rates vary among states, with each state represented by a different color.

****

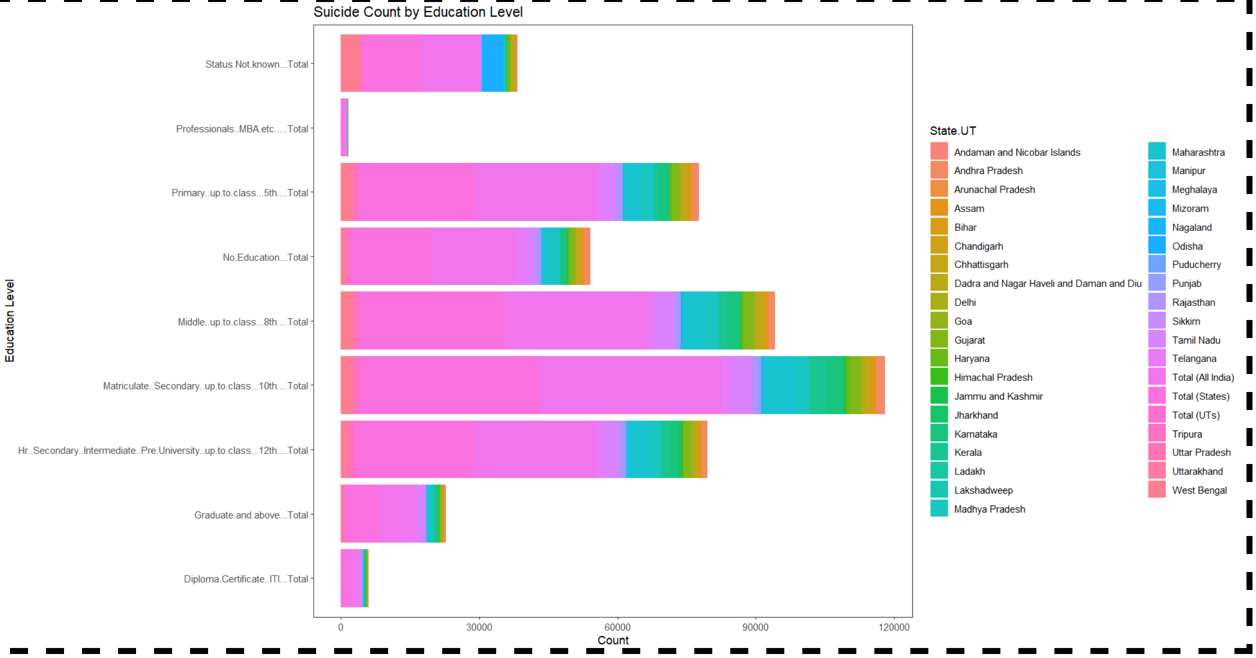
****

Code:



```
#plot by Education
ggplot(graph4, aes(Edu_Total, Count_Total, fill = State.UT)) +
  geom_histogram(stat = "identity",position = "stack", alpha = 0.9) +
  coord_flip()+
  labs(x = "Education Level", y = "Count", title = "Suicide Count by Education Level")
```

Output:



Inferences:

According to the graph representing suicide counts by education level, it can be inferred that individuals who have completed their secondary education up to class 10th have the highest suicide rates among all education levels. Conversely, individuals with professional degrees such as an MBA have the lowest suicide rates.

Additionally, the graph indicates that suicide rates vary by state, with each state having a different color. Further analysis is required to understand the reasons for these variations.

****