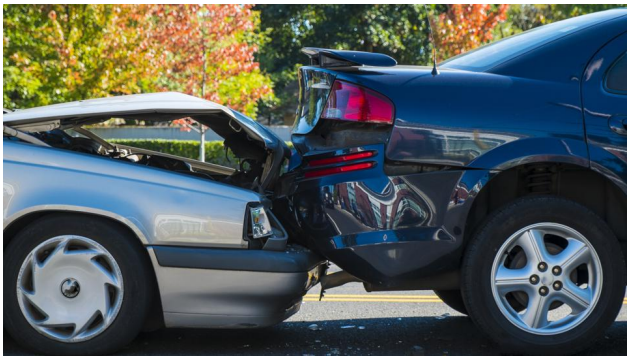
USA Motor Vehicles Crash Reporting

2023-03-11



Data Description:

This dataset profiles the drivers of motor vehicles who were involved in collisions on municipal and state roads in the USA. The dataset contains details on every traffic accident that was reported by the Automated Crash Reporting System on county and local roads in the United States (ACRS). This dataset displays the drivers engaged in each collision along with its statistics.

Link to Dataset:

https://catalog.data.gov/dataset/crash-reporting-drivers-data/resource/9851a37f-4f32-464e-8ba6-c23023653a7f

Project Statement:

We will be able to learn useful information by utilizing the provided dataset, such as which year saw the greatest number of reported cases, what weather conditions result in the majority of accidents, what severity level was present at the time of the accident, and what types of crashes are most frequently reported. Consequently, carrying out some studies and figuring out the outcome.

Research Results:

To analyze the data for the following situations and determine statistics, linear regression, polynomial regression, and clustering.

The following scenarios are presented for each year:

Scenario 1: Types of Accidents reported. Scenario 2: Total Cases reported. Scenario 3: Total Cases Injury Severity. Scenario 4: Weather At The Time Of An Incidents For Each Year.

Installing Packages And Libraries:

```
install.packages("tidyr", repos = "http://cran.us.r-project.org")
install.packages("factoextra", repos="http://cran.us.r-project.org")
install.packages("dplyr", repos = "http://cran.us.r-project.org")
install.packages("knitr", repos = "http://cran.us.r-project.org")
install.packages("ggplot2", repos = "http://cran.us.r-project.org")
library(factoextra)
library(tidyr)
library(dplyr)
library(lubridate)
library(stringr)
library(knitr)
library(ggplot2)
```

Main Source Code Logic Begin Here :

```
#Read into a data frame:

usa_crashdf =  read.csv("C:/Users/Public/USA_Motor_Vehicles_Crash_Reporting.csv")

#Changing the column's names to more appropriate ones:

usa_crashmdf =  select_all(usa_crashdf,(~gsub("\\s+|\\.", "_", .)))
colnames(usa_crashmdf)
```

```
##  [1] "Report_Number"                  "Local_Case_Number"
##  [3] "Agency_Name"                    "ACRS_Report_Type"
##  [5] "Crash_Date_Time"                "Route_Type"
##  [7] "Road_Name"                      "Cross_Street_Type"
##  [9] "Cross_Street_Name"              "Off_Road_Description"
## [11] "Municipality"                   "Related_Non_Motorist"
## [13] "Collision_Type"                 "Weather"
## [15] "Surface_Condition"              "Light"
## [17] "Traffic_Control"                "Driver_Substance_Abuse"
## [19] "Non_Motorist_Substance_Abuse"   "Person_ID"
## [21] "Driver_At_Fault"                "Injury_Severity"
## [23] "Circumstance"                   "Driver_Distracted_By"
## [25] "Drivers_License_State"          "Vehicle_ID"
## [27] "Vehicle_Damage_Extent"          "Vehicle_First_Impact_Location"
## [29] "Vehicle_Second_Impact_Location" "Vehicle_Body_Type"
## [31] "Vehicle_Movement"               "Vehicle_Continuing_Dir"
## [33] "Vehicle_Going_Dir"              "Speed_Limit"
## [35] "Driverless_Vehicle"             "Parked_Vehicle"
## [37] "Vehicle_Year"                   "Vehicle_Make"
## [39] "Vehicle_Model"                  "Equipment_Problems"
## [41] "Latitude"                       "Longitude"
## [43] "Location"
```

```r
#Replacing all spaces with NAs:

usa_crashmdf[usa_crashmdf == ""] = NA

#Removing the duplicates in the data frame:

usa_crashmdf = unique(usa_crashmdf)

#Selecting required columns:

usa_crashmdf1 = select(usa_crashmdf,Local_Case_Number,Agency_Name,ACRS_Report_Type,Crash_Date_Time,Route

#Removal of the NA values from the data frame:

usa_crashmdf2 = na.omit(usa_crashmdf1)

#Creating a Year column from Crash Date/Time:

usa_crashmdf2$Year = as.POSIXct(usa_crashmdf2$Crash_Date_Time, format = "%m/%d/%Y %H:%M:%S")

usa_crashmdf2$Year = format(usa_crashmdf2$Year , format="%Y")

usa_crashmdf2$Year = as.numeric(usa_crashmdf2$Year)

#Dataframe structure :

str(usa_crashmdf2)
```

```
## 'data.frame':    133419 obs. of  22 variables:
##  $ Local_Case_Number     : chr  "200023865" "200016465" "200016526" "200016305" ...
##  $ Agency_Name           : chr  "Montgomery County Police" "Montgomery County Police" "Montgomery Cou
##  $ ACRS_Report_Type      : chr  "Property Damage Crash" "Property Damage Crash" "Injury Crash" "Prope
##  $ Crash_Date_Time       : chr  "06/18/2020 02:00:00 AM" "04/19/2020 03:39:00 PM" "04/20/2020 09:15:0
##  $ Route_Type            : chr  "County" "County" "County" "Municipality" ...
##  $ Cross_Street_Type     : chr  "County" "County" "County" "Municipality" ...
##  $ Weather               : chr  "CLOUDY" "CLEAR" "CLOUDY" "N/A" ...
##  $ Surface_Condition     : chr  "DRY" "DRY" "DRY" "DRY" ...
##  $ Light                 : chr  "UNKNOWN" "DAYLIGHT" "DAYLIGHT" "DAYLIGHT" ...
##  $ Driver_Substance_Abuse: chr  "UNKNOWN" "ALCOHOL PRESENT" "NONE DETECTED" "NONE DETECTED" ...
##  $ Injury_Severity       : chr  "NO APPARENT INJURY" "NO APPARENT INJURY" "POSSIBLE INJURY" "NO APPA
##  $ Circumstance          : chr  "N/A" "N/A" "N/A" "N/A" ...
##  $ Driver_Distracted_By  : chr  "UNKNOWN" "UNKNOWN" "UNKNOWN" "NOT DISTRACTED" ...
##  $ Vehicle_Damage_Extent : chr  "UNKNOWN" "DISABLING" "DISABLING" "SUPERFICIAL" ...
##  $ Vehicle_Body_Type     : chr  "UNKNOWN" "VAN" "PASSENGER CAR" "PASSENGER CAR" ...
##  $ Vehicle_Continuing_Dir: chr  "Unknown" "East" "North" "North" ...
##  $ Vehicle_Going_Dir     : chr  "Unknown" "East" "North" "North" ...
##  $ Speed_Limit           : int  35 25 25 25 40 35 35 35 40 40 ...
##  $ Driverless_Vehicle    : chr  "No" "No" "No" "No" ...
##  $ Vehicle_Year          : int  2020 2004 2006 2011 2018 2017 2014 2018 2003 1996 ...
##  $ Vehicle_Make          : chr  "UNK" "DODGE" "HONDA" "TOYOTA" ...
##  $ Year                  : num  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
##  - attr(*, "na.action")= 'omit' Named int [1:17476] 1 3 11 17 23 28 33 41 47 54 ...
##   ..- attr(*, "names")= chr [1:17476] "1" "3" "11" "17" ...
```

```r
#Selecting columns for applying rules:

usa_crash_col = select(usa_crashmdf2,Local_Case_Number,Agency_Name,Route_Type,Weather,ACRS_Report_Type,

#Scenario-1 :Types of Crashes reported for each year:

usa_rule1 = select(usa_crash_col,ACRS_Report_Type,Year)

#Scenario-1 :Groupby count of required column:

usa_crash_rule1 = aggregate(usa_rule1$ACRS_Report_Type, by=list(usa_rule1$ACRS_Report_Type,usa_rule1$Yea

usa_crashgrp = rename(usa_crash_rule1,"Total_Report_Type" = "x", "ACRS_Report_Type" = "Group.1","Year" =

#Scenario-1 :Types Of Crashes Reported For Each Year Statistics:

kable(summary(select(usa_crashgrp,Total_Report_Type,Year)),row.names = FALSE,caption = "Types Of Crashes
```

Table 1: Types Of Crashes Reported For Each Year Statistics

Total_Report_Type	Year
Min. : 37.0	Min. :2015
1st Qu.: 52.5	1st Qu.:2017
Median : 7339.0	Median :2018
Mean : 5558.5	Mean :2018
3rd Qu.: 8683.8	3rd Qu.:2020
Max. :11316.0	Max. :2022

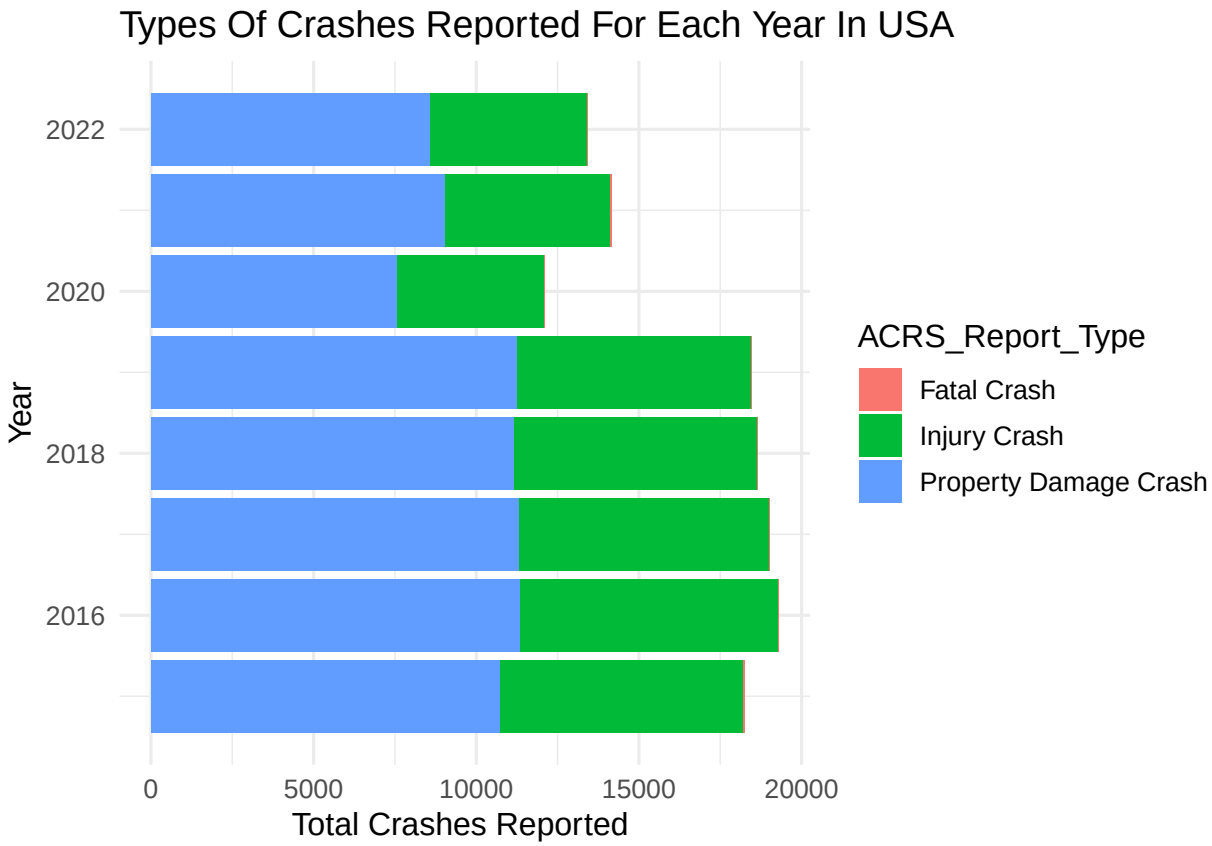
```r
#Scenario-1 :Linear regression of Types Of Crashes Reported For Each Year In USA:

summary(lm(Total_Report_Type~Year, data = usa_crashgrp))
```

```
##
## Call:
## lm(formula = Total_Report_Type ~ Year, data = usa_crashgrp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6629  -4791   1192   4072   5826
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 648090.3   798147.6   0.812    0.425
## Year          -318.3      395.4  -0.805    0.429
##
## Residual standard error: 4439 on 22 degrees of freedom
## Multiple R-squared:  0.02861,    Adjusted R-squared:  -0.01554
## F-statistic: 0.6481 on 1 and 22 DF,  p-value: 0.4294
```

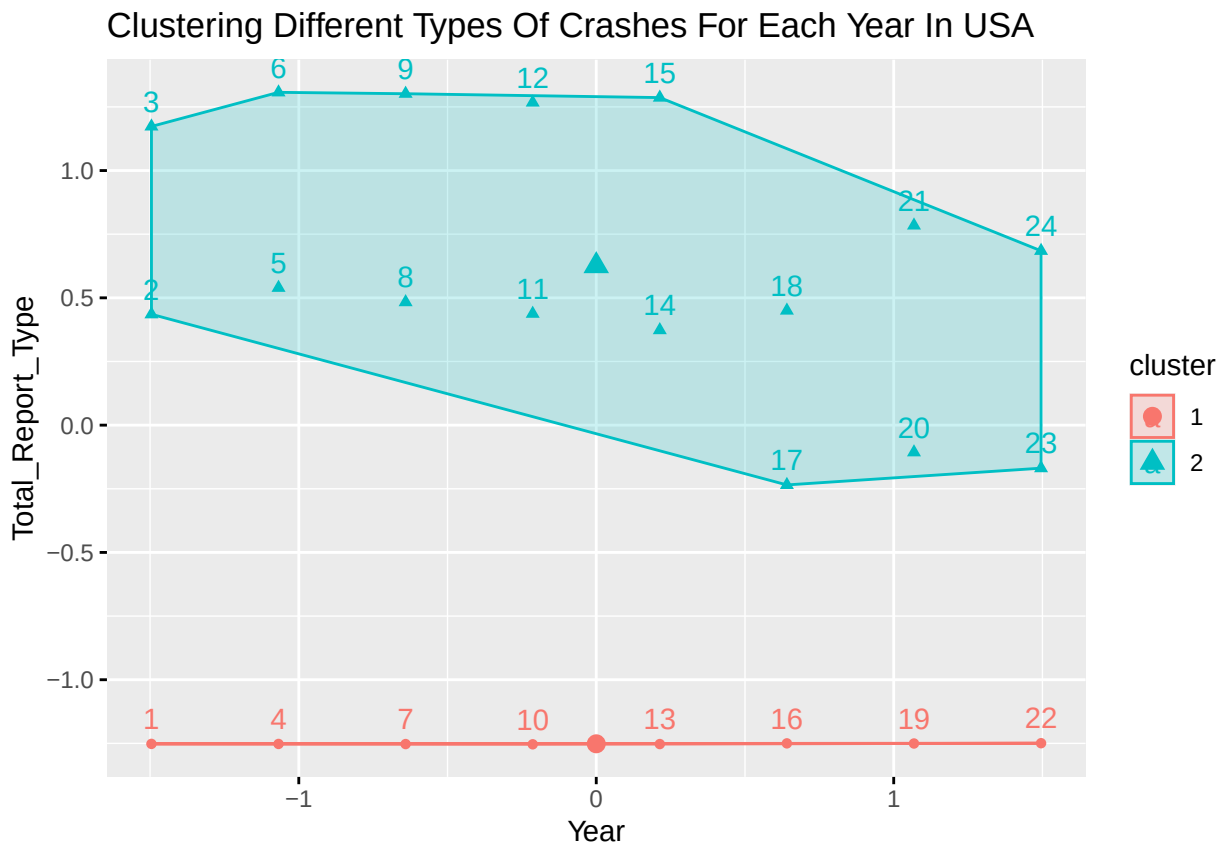
```r
#Scenario-1 :Polynomial regression of Types Of Crashes Reported For Each Year In USA:

summary(lm(Total_Report_Type ~ poly(Year, 2, raw = TRUE),data = usa_crashgrp))
```

```
##
## Call:
## lm(formula = Total_Report_Type ~ poly(Year, 2, raw = TRUE), data = usa_crashgrp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6298  -4791   1321   4319   5590
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -1.917e+08  8.234e+08  -0.233    0.818
## poly(Year, 2, raw = TRUE)1  1.902e+05  8.159e+05   0.233    0.818
## poly(Year, 2, raw = TRUE)2 -4.720e+01  2.021e+02  -0.234    0.818
##
## Residual standard error: 4537 on 21 degrees of freedom
## Multiple R-squared:  0.03113,    Adjusted R-squared:  -0.06114
## F-statistic: 0.3374 on 2 and 21 DF,  p-value: 0.7174
```

```r
#Scenario-1 :Plotting Types Of Crashes Reported For Each Year:

ggplot(usa_crashgrp,aes(x=Year, y=Total_Report_Type,fill=ACRS_Report_Type)) +
  ggtitle("Types Of Crashes Reported For Each Year In USA") +
  xlab("Year") +
  ylab("Total Crashes Reported ") +
  theme_minimal(base_size = 12) +
  geom_bar(stat="identity") +
  coord_flip() +
  scale_color_discrete(name = "Different Types Of Crashes")
```




```
#Scenario-1 :Clustering for Types Of Crashes Reported For Each Year In USA:

fviz_cluster((kmeans((select(usa_crashgrp,Year,Total_Report_Type)), centers =2)), data = (select(usa_cra
```




```
#Scenario-2 :Total Cases Reported For Each Year In USA:

usa_crash_rule2 = select(usa_crash_col,Year)

#Scenario-2 : Groupby count of required column:

usa_crashgrp_rule2 = aggregate(usa_crash_rule2$Year, by=list(usa_crash_rule2$Year), FUN=length)

usa_crashgrp_rule2 = rename(usa_crashgrp_rule2,"Total_Cases" = "x","Year" = "Group.1")

#Scenario-2 :Total Cases Reported For Each Year In USA Statistics:

kable(summary(select(usa_crashgrp_rule2,Total_Cases,Year)),row.names = FALSE,caption = "Total Cases Rep
```

Table 2: Total Cases Reported For Each Year Statistics

Total_Cases	Year
Min. :12118	Min. :2015
1st Qu.:13976	1st Qu.:2017
Median :18357	Median :2018
Mean :16676	Mean :2018
3rd Qu.:18752	3rd Qu.:2020
Max. :19294	Max. :2022

```r
#Scenario-2 :Linear regression of Total Cases Reported For Each Year In USA:

summary(lm(Total_Cases~Year, data = usa_crashgrp_rule2))
```

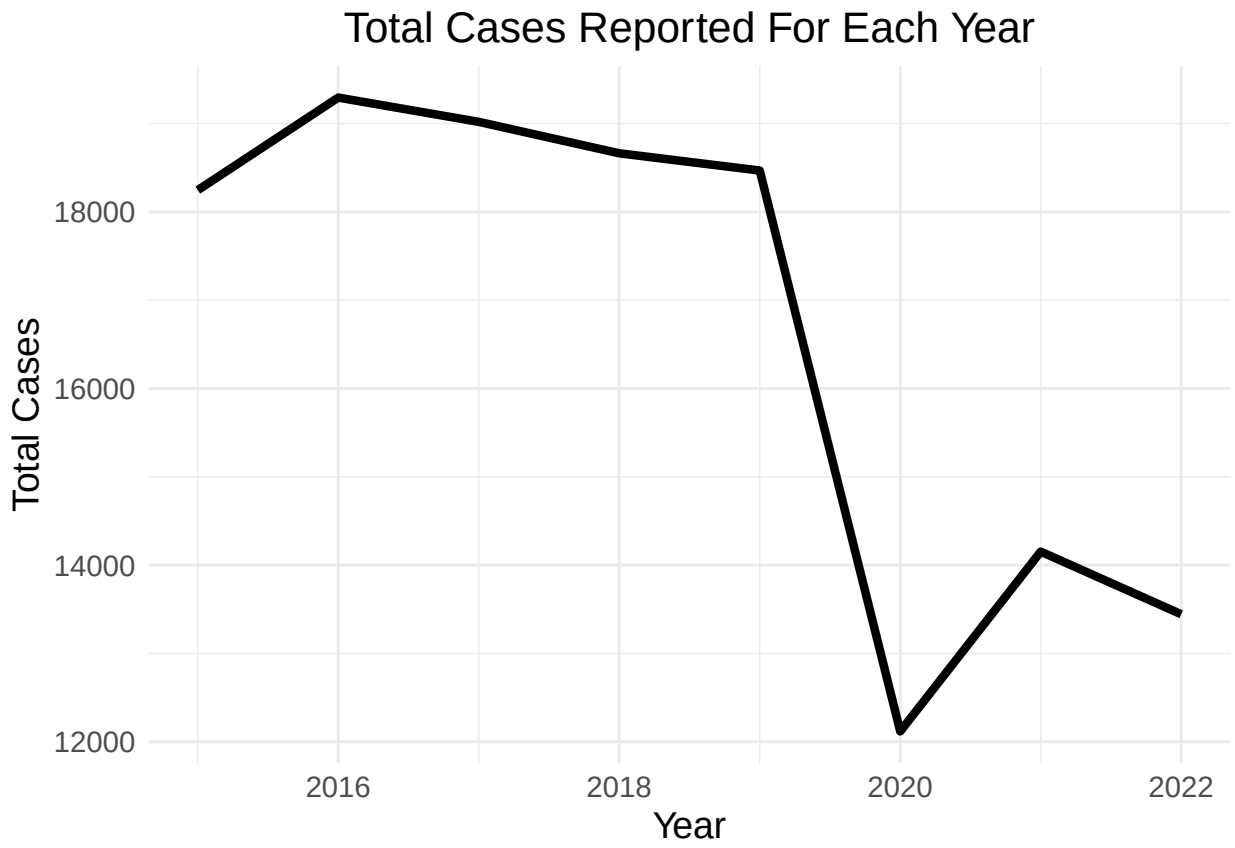
```
##
## Call:
## lm(formula = Total_Cases ~ Year, data = usa_crashgrp_rule2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3125.2  -544.4   170.9  1060.7  2269.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1944271.0   586182.9   3.317   0.0161 *
## Year           -955.0      290.4  -3.288   0.0166 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1882 on 6 degrees of freedom
## Multiple R-squared:  0.6431, Adjusted R-squared:  0.5837
## F-statistic: 10.81 on 1 and 6 DF,  p-value: 0.01665
```

```r
#Scenario-2 :Polynomial regression of Total Cases Reported For Each Year In USA:

summary(lm(Total_Cases ~ poly(Year, 2, raw = TRUE),data = usa_crashgrp_rule2))
```

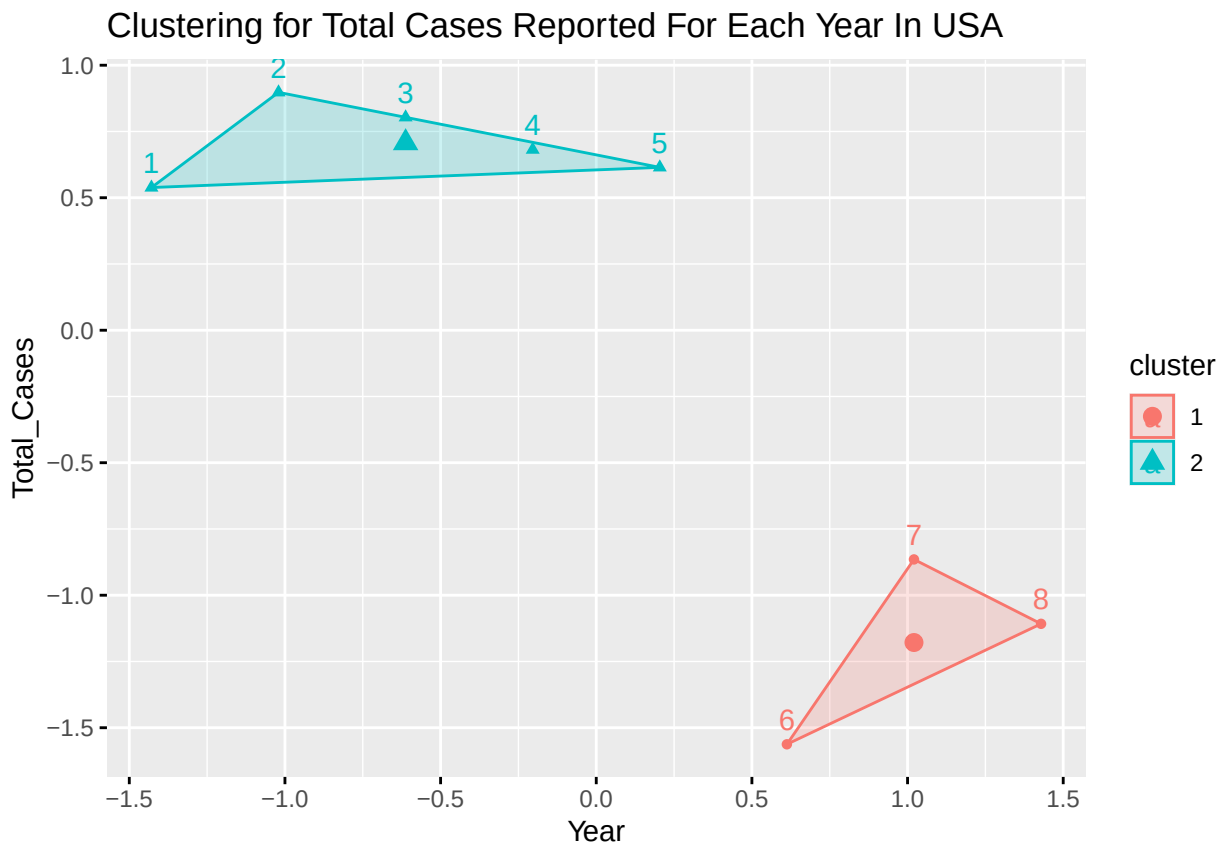
```
##
## Call:
## lm(formula = Total_Cases ~ poly(Year, 2, raw = TRUE), data = usa_crashgrp_rule2)
##
## Residuals:
##        1        2        3        4        5        6        7        8
##  -780.792   372.565   486.125   801.887  1561.851 -3549.982     6.387  1101.958
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -5.750e+08  5.945e+08  -0.967    0.378
## poly(Year, 2, raw = TRUE)1  5.707e+05  5.890e+05   0.969    0.377
## poly(Year, 2, raw = TRUE)2 -1.416e+02  1.459e+02  -0.970    0.376
##
## Residual standard error: 1891 on 5 degrees of freedom
## Multiple R-squared:  0.6997, Adjusted R-squared:  0.5796
## F-statistic: 5.825 on 2 and 5 DF,  p-value: 0.04942
```

```r
#Scenario-2 :Plotting Total Cases Reported For Each Year In USA:

ggplot(usa_crashgrp_rule2,aes(x=Year,y=Total_Cases)) +geom_line(size=1.5) +labs(x = "Year", y = "Total C
    ggtitle(paste0("Total Cases Reported For Each Year"))+
    theme_minimal()+
    theme(legend.position = "right",
          plot.title = element_text(hjust = 0.5,size=16),
```

```
        text = element_text(size=14))+
    scale_fill_brewer(palette="Set3")
```




```
#Scenario-2 :Clustering for Total Cases Reported For Each Year In USA:

fviz_cluster((kmeans((select(usa_crashgrp_rule2,Year,Total_Cases)), centers =2)), data = (select(usa_cra
```




```
#Scenario-3 :Total Cases Injury Severity:

USA_crash_rule3 = select(usa_crash_col,Injury_Severity,Year)

#Scenario-3 : Groupby count of required column:

usa_crashgrp_rule3 = aggregate(USA_crash_rule3$Injury_Severity, by=list(USA_crash_rule3$Injury_Severity

usa_crashgrp_rule3 = rename(usa_crashgrp_rule3,"Total_Injury_Severity" = "x","Injury_Severity" = "Group

#Scenario-3 :Total Cases Injury Severity Statistics For Each Year In USA :

kable(summary(select(usa_crashgrp_rule3,Total_Injury_Severity,Year)),row.names = FALSE,caption = "Total
```

Table 3: Total Cases Injury Severity Statistics

Total_Injury_Severity	Year
Min. : 9	Min. :2015
1st Qu.: 128	1st Qu.:2017
Median : 1240	Median :2018
Mean : 3335	Mean :2018
3rd Qu.: 2236	3rd Qu.:2020
Max. :15305	Max. :2022

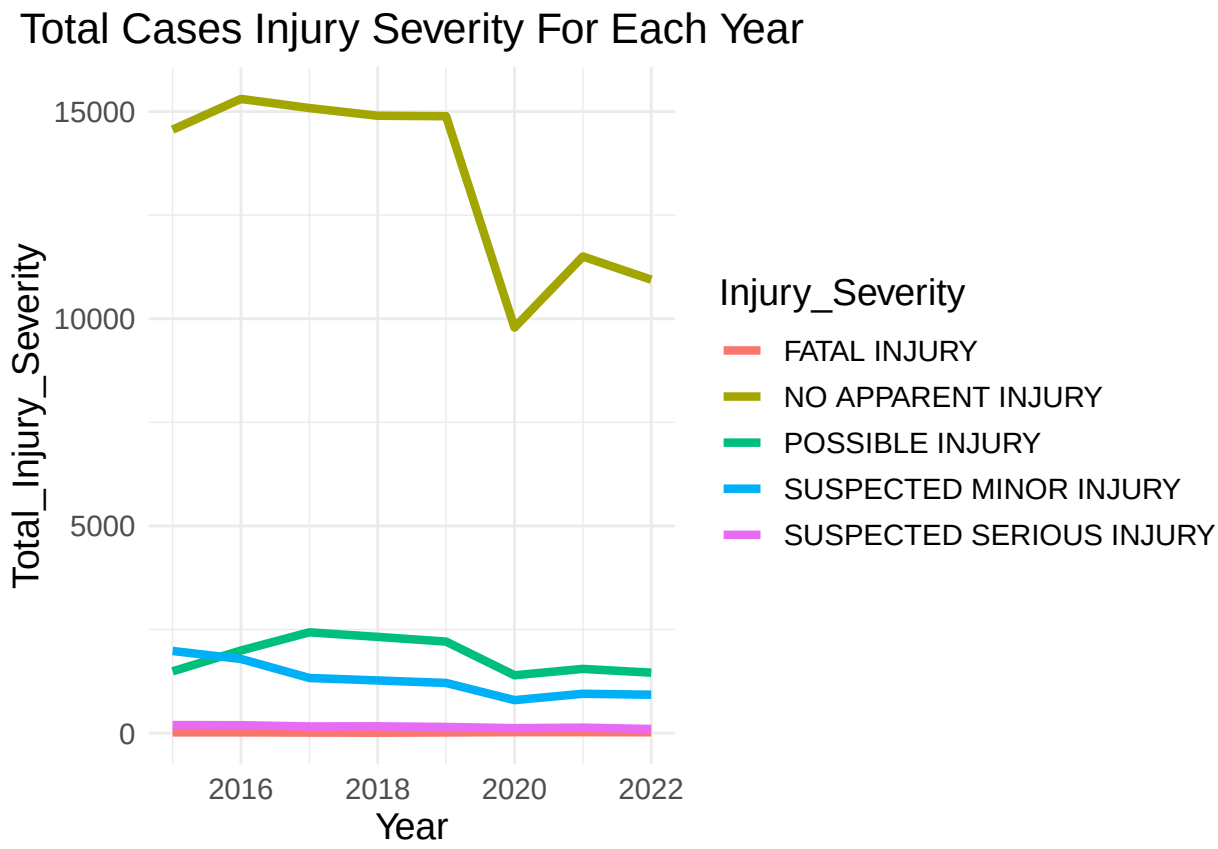
```r
#Scenario-3 :Linear regression Of Total Cases Injury Severity For Each Year :

summary(lm(Total_Injury_Severity~Year, data = usa_crashgrp_rule3))
```

```
##
## Call:
## lm(formula = Total_Injury_Severity ~ Year, data = usa_crashgrp_rule3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3986  -3043  -2095  -1169  11648
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 388854.2   734330.6   0.530    0.600
## Year          -191.0      363.8  -0.525    0.603
##
## Residual standard error: 5272 on 38 degrees of freedom
## Multiple R-squared:  0.007201,   Adjusted R-squared:  -0.01893
## F-statistic: 0.2756 on 1 and 38 DF,  p-value: 0.6026
```

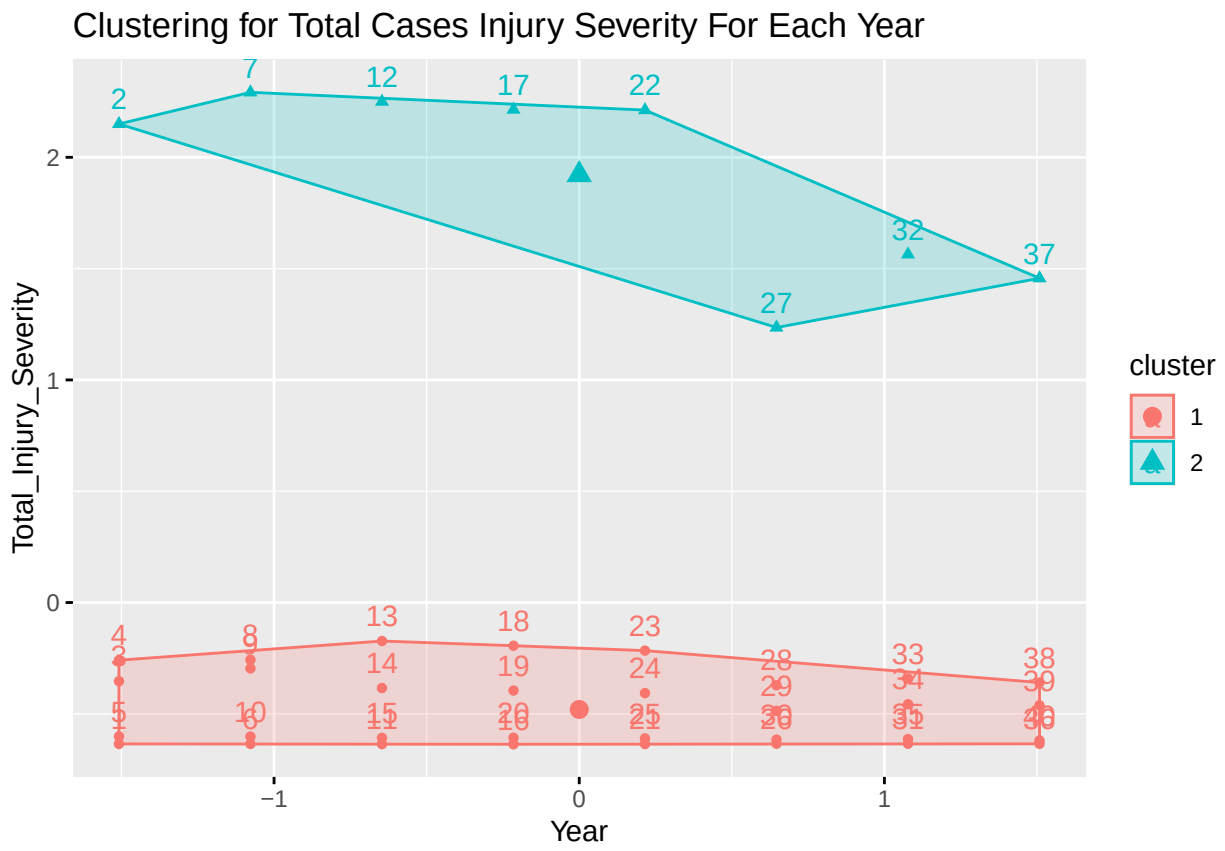
```r
#Scenario-3 :Polynomial regression Of Total Cases Injury Severity For Each Year In USA:

summary(lm(Total_Injury_Severity ~ poly(Year, 2, raw = TRUE),data = usa_crashgrp_rule3))
```

```
##
## Call:
## lm(formula = Total_Injury_Severity ~ poly(Year, 2, raw = TRUE),
##     data = usa_crashgrp_rule3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3787  -3142  -2237  -1231  11521
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -1.150e+08  7.508e+08  -0.153    0.879
## poly(Year, 2, raw = TRUE)1  1.141e+05  7.440e+05   0.153    0.879
## poly(Year, 2, raw = TRUE)2 -2.832e+01  1.843e+02  -0.154    0.879
##
## Residual standard error: 5341 on 37 degrees of freedom
## Multiple R-squared:  0.007834,   Adjusted R-squared:  -0.0458
## F-statistic: 0.1461 on 2 and 37 DF,  p-value: 0.8646
```

```r
#Scenario-3 :Plotting Total Cases Injury Severity For Each Year In USA:

ggplot(usa_crashgrp_rule3,aes(x=Year,y=Total_Injury_Severity,col = Injury_Severity)) +geom_line(size=1.5
    ggtitle(paste0("Total Cases Injury Severity For Each Year"))+
    theme_minimal()+
    theme(legend.position = "right",
          plot.title = element_text(hjust = 0.5,size=16),
          text = element_text(size=14))+
    scale_fill_brewer(palette="Set3")
```




```
#Scenario-3 :Clustering for Total Cases Injury Severity For Each Year In USA:

fviz_cluster((kmeans((select(usa_crashgrp_rule3,Year,Total_Injury_Severity)), centers =2)), data = (sel
```




```
#Scenario-4 :Weather at the time of an each incident:

usa_crash_rule4 = select(usa_crash_col,Weather,Year)
usa_crash_rule4 = filter(usa_crash_rule4, Weather != "N/A")

#Scenario-4 : Groupby count of required column:

usa_crashgrp_rule4 = aggregate(usa_crash_rule4$Weather, by=list(usa_crash_rule4$Weather,usa_crash_rule4$

usa_crashgrp_rule4 = rename(usa_crashgrp_rule4,"Total_Incidents" = "x","Weather" = "Group.1","Year" = "

#Scenario-4 :Weather At The Time Of An Each Incident Statistics For Each Year :

kable(summary(select(usa_crashgrp_rule4,Total_Incidents,Year)),row.names = FALSE,caption = "Weather At
```

Table 4: Weather At The Time Of An Each Incident Statistics

Total_Incidents	Year
Min. : 1	Min. :2015
1st Qu.: 24	1st Qu.:2016
Median : 47	Median :2018
Mean : 1387	Mean :2018
3rd Qu.: 1237	3rd Qu.:2020
Max. :13734	Max. :2022

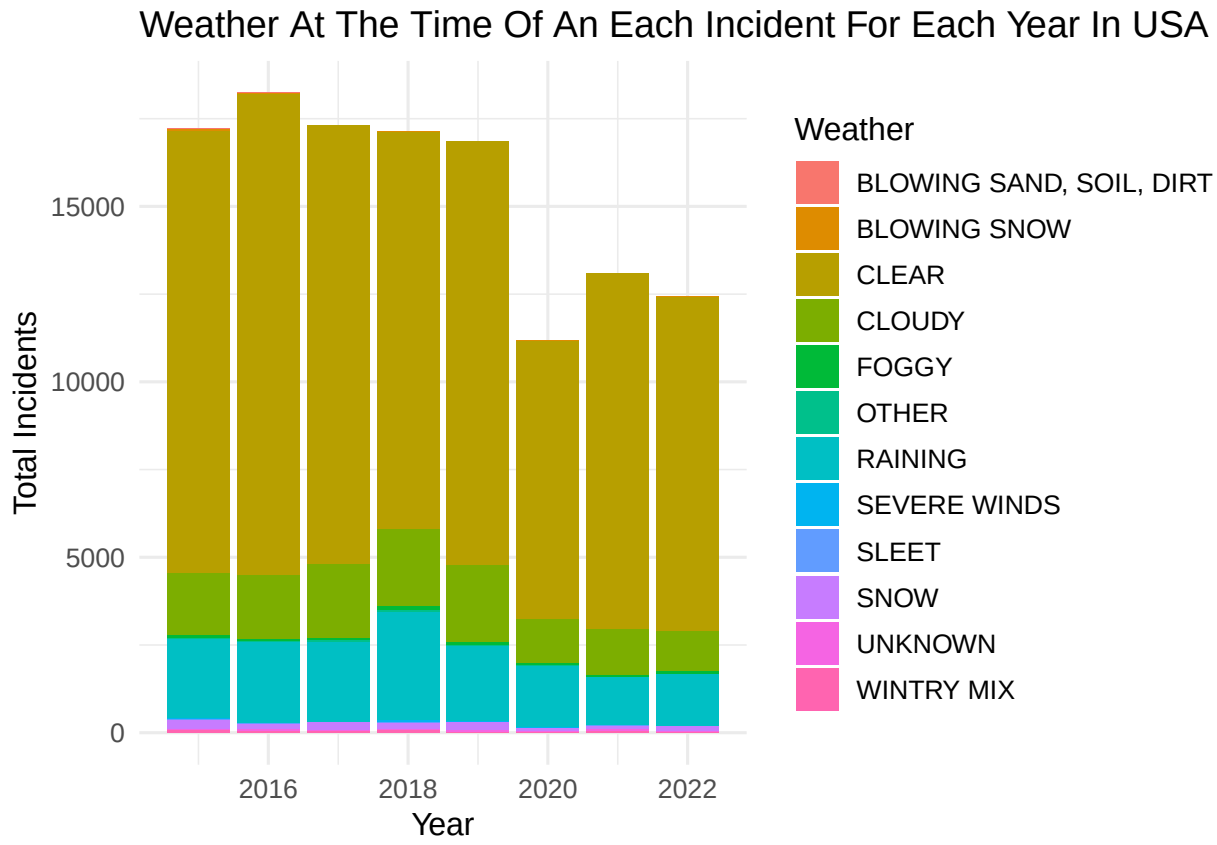
```r
#Scenario-4 :Linear regression Of Weather At The Time Of An Each Incident For Each Year In USA :

summary(lm(Total_Incidents~Year, data = usa_crashgrp_rule4))
```

```
##
## Call:
## lm(formula = Total_Incidents ~ Year, data = usa_crashgrp_rule4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1570.6 -1404.4 -1271.5   -18.6 12208.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 117295.11  302355.85   0.388    0.699
## Year           -57.43     149.80  -0.383    0.702
##
## Residual standard error: 3259 on 87 degrees of freedom
## Multiple R-squared:  0.001686,   Adjusted R-squared:  -0.009789
## F-statistic: 0.147 on 1 and 87 DF,  p-value: 0.7024
```

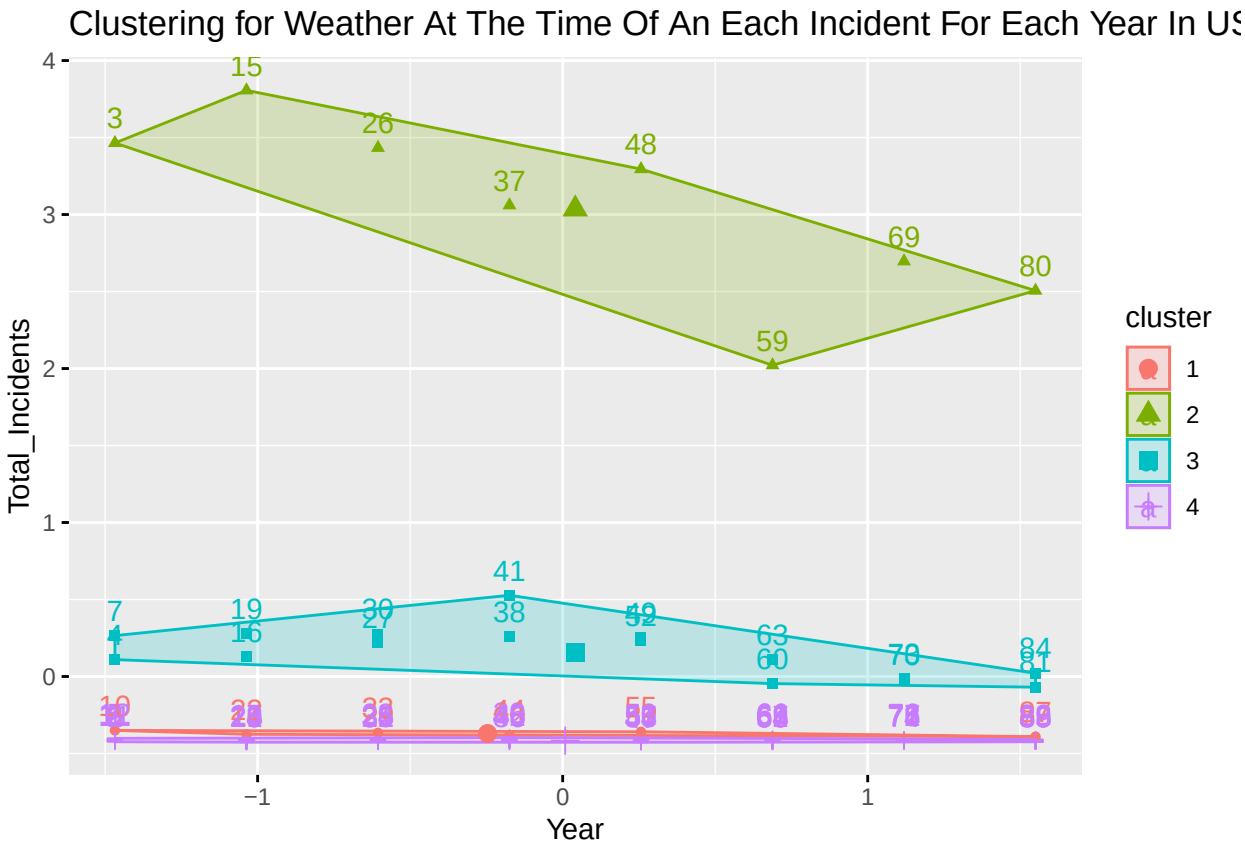
```r
#Scenario-4 :Polynomial regression Of Weather At The Time Of An Each Incident For Each Year In USA :

summary(lm(Total_Incidents ~ poly(Year, 2, raw = TRUE),data = usa_crashgrp_rule4))
```

```
##
## Call:
## lm(formula = Total_Incidents ~ poly(Year, 2, raw = TRUE), data = usa_crashgrp_rule4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1510.8 -1441.6 -1315.5    86.7 12220.2
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -6.000e+07  3.068e+08  -0.196    0.845
## poly(Year, 2, raw = TRUE)1  5.951e+04  3.040e+05   0.196    0.845
## poly(Year, 2, raw = TRUE)2 -1.476e+01  7.530e+01  -0.196    0.845
##
## Residual standard error: 3278 on 86 degrees of freedom
## Multiple R-squared:  0.002132,   Adjusted R-squared:  -0.02107
## F-statistic: 0.09187 on 2 and 86 DF,  p-value: 0.9123
```

```r
#Scenario-4 :Plotting Weather At The Time Of An Each Incident For Each Year :

ggplot(usa_crashgrp_rule4,aes(x=Year, y=Total_Incidents,fill=Weather)) +
  ggtitle("Weather At The Time Of An Each Incident For Each Year In USA") +
  xlab("Year") +
  ylab("Total Incidents") +
  theme_minimal(base_size = 12) +
  geom_bar(stat="identity") +
  scale_color_discrete(name = "Weather Type")
```




```
#Scenario-4 :Clustering for Weather At The Time Of An Each Incident For Each Year In USA:

fviz_cluster((kmeans((select(usa_crashgrp_rule4,Year,Total_Incidents)), centers =4)), data = (select(usa
```


Clustering for Weather At The Time Of An Each Incident For Each Year In US
Total_Incidents
Year
cluster
1
2
3
4