U.K. Accidents- Ten Years History.

Sheetal Munjewar

2023-02-19

Contents

Introduction 2

Research Questions 2

Approach 2

How your approach addresses (fully or partially) the problem. 2

Data 2

function declarations 2

Load Data 3

Merge data (Three datasets into one) 3

Populate column code with meaningful descriptios using Excel file Road-Accident-Safety-Data-Guide.xls into new column. 3

Get rid of excess data columns. 5

Change Date column to date format. 5

Adding new columns for aggregation and summerization. 5

Display final data set and save it in seperate file. 5

What do you not know how to do right now that you need to learn to import and cleanup your dataset? 8

Discuss how you plan to uncover new information in the data that is not self-evident. 8

What are different ways you could look at this data to answer the questions you want to answer? 8

Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain. 8

How could you summarize your data to answer key questions? 8

Required Packages 8

Plots and Table Needs 8

What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.). 8

Questions for future steps 9

********* NOTE - Scratch work *************** 9

THis section is not to demonstrate any thing, its scratch section to play around dataset to understand variables and its co-relations and try out plot function to learn about it. it can be consider as a prep work for final week project. 9

Introduction

Road safety is the common concern around the world, As a part of this exercise we are going to explore U.K road safety data about the circumstances of personal injury road accidents in GB from 2005 to 2014,

Data Source link : https://www.kaggle.com/datasets/benoit72/uk-accidents-10-years-history-with-many-variables

Different data Sources files (cvs):

Accident file: main data set contains information about accident severity, weather, location, date, hour, day of week, road type… Vehicle file : contains information about vehicle type, vehicle model, engine size, driver sex, driver age, car age… Casualty file: contains information about casualty severity, age, sex social class, casualty type, pedestrian or car passenger… Lookup file : contains the text description of all variable code in the three files

License - http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

Research Questions

- Accidents are on rise or decline over the years ?
- Co-relation between weather with number or severity of an accident?
- Does driver age has an effect on the number of accident?
- What is the relation between hour, day, week, month with number of fatal accident?
- Are certain car models safer than others?
- Is the social class of a casualty dependent of the accident severity?

Approach

Data must be collected from legal source (Publicly available), Check for missing data, merge the different data sources/files into one data frame. In out case we have four data sources. Map column codes with text string for look up table, map and assign column names. map log/lat into the countries. Filter required columns to address research questions and use graphs for visualizations.

How your approach addresses (fully or partially) the problem.

Project approach is the address following future forcast :

Can you forecast the future daily/weekly/monthly accidents? Action that can prevent future accident based on variable relationship and predictions ? Fatal accidents can be predict or avoided ? Variables contributing rise in fatal accidents ?

Data

Four data Sources(cvs):

Accident file: main data set contains information about accident severity, weather, location, date, hour, day of week, road type… Vehicle file : contains information about vehicle type, vehicle model, engine size, driver sex, driver age, car age… Casualty file: contains information about casualty severity, age, sex social class, casualty type, pedestrian or car passenger… Lookup file : contains the text description of all variable code in the three files

Sources : https://www.kaggle.com/datasets/benoit72/uk-accidents-10-years-history-with-many-variables

function declarations

```
##
## Attaching package: 'dplyr'
## The following objects are masked from ’package:stats’:
##
##     filter, lag
```

```
## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
## corrplot 0.92 loaded
```

```
## Loading required package: lattice
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: ’plyr’
```

```
## The following objects are masked from ’package:dplyr’:
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

Load Data

- Total three data sources and one label index excel.
- Accident_Index field, unique identifier that refers to one accident and common to link all data sets.

Merge data (Three datasets into one)

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(warning = FALSE)
knitr::opts_chunk$set(fig.width = 12, fig.height = 10)
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 70), tidy = TRUE)

df <- merge(Accidents, Casualties, by='Accident_Index')
df <- merge(df, Vehicles, by='Accident_Index')
rm(Accidents, Casualties, Vehicles)
# str(df)
# head(df)
```

Populate column code with meaningful descriptios using Excel file Road-Accident-Safety-Data-Guide.xls into new column.

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(warning = FALSE)
knitr::opts_chunk$set(fig.width = 12, fig.height = 10)
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 70), tidy = TRUE)

setwd("E:\\Data_Science_DSC510\\DSC520-Statistics\\dsc520")

Location_code <- read_excel("assignments/Final-Project/Road-Accident-Safety-Data-Guide.xls", sheet="Police Force
df <- left_join(df, Location_code, by=c("Police_Force"="code"))
df <- dplyr::rename(df, Location=label)
rm(Location_code)

setwd("E:\\Data_Science_DSC510\\DSC520-Statistics\\dsc520")
Junction_type <- read_excel("assignments/Final-Project/Road-Accident-Safety-Data-Guide.xls", sheet="Junction Det
df <- left_join(df, Junction_type, by=c("Junction_Detail"="code"))
df <- dplyr::rename(df, Junction=label)
rm(Junction_type)

Light_conditions <- read_excel("assignments/Final-Project/Road-Accident-Safety-Data-Guide.xls", sheet="Light Con
df <- left_join(df, Light_conditions, by=c("Light_Conditions" = "code"))
df <- dplyr::rename(df, Lighting = label)
rm(Light_conditions)

Weather_conditions <- read_excel("assignments/Final-Project/Road-Accident-Safety-Data-Guide.xls", sheet="Weather
df <- left_join(df, Weather_conditions, by=c("Weather_Conditions"="code"))
df <- dplyr::rename(df, Weather = label)
rm(Weather_conditions)

Surface_conditions <- read_excel("assignments/Final-Project/Road-Accident-Safety-Data-Guide.xls", sheet="Road Su
df <- left_join(df, Surface_conditions, by = c("Road_Surface_Conditions" = "code"))
df <- dplyr::rename(df, Surface = label)
rm(Surface_conditions)

Vehicle_type <- read_excel("assignments/Final-Project/Road-Accident-Safety-Data-Guide.xls", sheet="Vehicle Type'
df <- left_join(df, Vehicle_type, by = c("Vehicle_Type" = "code"))
df <- dplyr::rename(df, Vehicle = label)
rm(Vehicle_type)

Vehicle_manoeuvre <- read_excel("assignments/Final-Project/Road-Accident-Safety-Data-Guide.xls", sheet="Vehicle
df <- left_join(df, Vehicle_manoeuvre, by = c("Vehicle_Manoeuvre" = "code"))
df <- dplyr::rename(df, Manoeuvre = label)
rm(Vehicle_manoeuvre)

Skidding <- read_excel("assignments/Final-Project/Road-Accident-Safety-Data-Guide.xls", sheet="Skidding and Over
df <- left_join(df, Skidding, by = c("Skidding_and_Overturning" = "code"))
df <- dplyr::rename(df, Skidding = label)
rm(Skidding)

Journey_purpose <- read_excel("assignments/Final-Project/Road-Accident-Safety-Data-Guide.xls", sheet="Journey Pu
df <- left_join(df, Journey_purpose, by = c("Journey_Purpose_of_Driver" = "code"))
df <- dplyr::rename(df, Journey = label)
rm(Journey_purpose)

Age_band <- read_excel("assignments/Final-Project/Road-Accident-Safety-Data-Guide.xls", sheet="Age Band")
df <- left_join(df, Age_band, by = c("Age_Band_of_Driver" = "code"))
df <- dplyr::rename(df, Age_Band = label)
rm(Age_band)

Casualty_severity <- read_excel("assignments/Final-Project/Road-Accident-Safety-Data-Guide.xls", sheet="Accident
df <- left_join(df, Casualty_severity, by=c("Casualty_Severity"="code"))
df <- dplyr::rename(df, Casualty_Outcome=label)
rm(Casualty_severity)
```

Get rid of excess data columns.

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(warning = FALSE)
knitr::opts_chunk$set(fig.width = 12, fig.height = 10)
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 70), tidy = TRUE)

df <- df[!names(df) %in% c("Driver_Home_Area_Type","Driver_IMD_Decile","Propulsion_Code","Age_Band_of_Driver","1

#dim(df)
#head(df)
```

Change Date column to date format.

```
df$Date<- as.Date(df$Date, "%m/%d/%Y")
#str(df$Date)
#head(df$Date)
```

Adding new columns for aggregation and summerization.

```
df$Year <- format(as.Date(df$Date), "%Y")
df$Month <- format(as.Date(df$Date), "%m")
```

Display final data set and save it in seperate file.

```
head(df)
```

```
##   Accident_Index Location_Easting_OSGR Location_Northing_OSGR Longitude
## 1  200501BS00001                525680                 178240 -0.191170
## 2  200501BS00002                524170                 181650 -0.211708
## 3  200501BS00003                524520                 182240 -0.206458
## 4  200501BS00003                524520                 182240 -0.206458
## 5  200501BS00004                526900                 177530 -0.173862
## 6  200501BS00005                528060                 179040 -0.156618
##   Latitude Police_Force Accident_Severity Number_of_Vehicles
## 1 51.48910            1                 2                  1
## 2 51.52007            1                 3                  1
## 3 51.52530            1                 3                  2
## 4 51.52530            1                 3                  2
## 5 51.48244            1                 3                  1
## 6 51.49575            1                 3                  1
##   Number_of_Casualties       Date Day_of_Week  Time Local_Authority_.District.
## 1                    1 2005-04-01           3 17:42                         12
## 2                    1 2005-05-01           4 17:36                         12
## 3                    1 2005-06-01           5 00:15                         12
## 4                    1 2005-06-01           5 00:15                         12
## 5                    1 2005-07-01           6 10:35                         12
## 6                    1 2005-10-01           2 21:13                         12
##   Local_Authority_.Highway. X1st_Road_Class X1st_Road_Number Road_Type
## 1                 E09000020               3             3218         6
## 2                 E09000020               4              450         3
## 3                 E09000020               5                0         6
## 4                 E09000020               5                0         6
## 5                 E09000020               3             3220         6
## 6                 E09000020               6                0         6
##   Speed_limit X2nd_Road_Class X2nd_Road_Number
## 1          30              -1                0
## 2          30               5                0
## 3          30              -1                0
## 4          30              -1                0
## 5          30              -1                0
## 6          30              -1                0
##   Pedestrian_Crossing.Human_Control Pedestrian_Crossing.Physical_Facilities
## 1                                 0                                       1
## 2                                 0                                       5
## 3                                 0                                       0
## 4                                 0                                       0
## 5                                 0                                       0
## 6                                 0                                       0
##   Light_Conditions Weather_Conditions Road_Surface_Conditions
## 1                1                  2                       2
## 2                4                  1                       1
## 3                4                  1                       1
## 4                4                  1                       1
## 5                1                  1                       1
## 6                7                  1                       2
##   Special_Conditions_at_Site Vehicle_Reference.x Casualty_Class Sex_of_Casualty
## 1                          0                   1              3               1
## 2                          0                   1              2               1
## 3                          0                   2              1               1
## 4                          0                   2              1               1
## 5                          0                   1              3               1
## 6                          0                   1              1               1
##   Age_of_Casualty Age_Band_of_Casualty Casualty_Severity Car_Passenger
## 1              37                    7                 2             0
## 2              37                    7                 3             0
## 3              62                    9                 3             0
## 4              62                    9                 3             0
## 5              30                    6                 3             0
## 6              49                    8                 3             0
##   Bus_or_Coach_Passenger Casualty_Type Casualty_Home_Area_Type
## 1                      0             0                       1
## 2                      4            11                       1
## 3                      0             9                       1
## 4                      0             9                       1
## 5                      0             0                       1
## 6                      0             3                      -1
##   Vehicle_Reference.y Vehicle_Type Vehicle_Manoeuvre
## 1                   1            9                18
## 2                   1           11                 4
## 3                   1           11                17
## 4                   2            9                 2
## 5                   1            9                18
## 6                   1            3                18
##   Vehicle_Location.Restricted_Lane Skidding_and_Overturning
## 1                                0                        0
## 2                                0                        0
## 3                                0                        0
## 4                                0                        0
## 5                                0                        0
## 6                                0                        1
##   X1st_Point_of_Impact Was_Vehicle_Left_Hand_Drive. Journey_Purpose_of_Driver
## 1                    1                            1                        15
## 2                    4                            1                         1
## 3                    4                            1                         1
## 4                    3                            1                        15
## 5                    1                            1                        15
## 6                    1                            1                        15
##   Sex_of_Driver Age_of_Driver Engine_Capacity_.CC. Age_of_Vehicle
## 1             2            74                   -1             -1
## 2             1            42                 8268              3
## 3             1            35                 8300              5
## 4             1            62                 1762              6
## 5             2            49                 1769              4
## 6             1            49                   85             10
##              Location                            Junction
## 1 Metropolitan Police Not at junction or within 20 metres
## 2 Metropolitan Police                          Crossroads
## 3 Metropolitan Police Not at junction or within 20 metres
## 4 Metropolitan Police Not at junction or within 20 metres
## 5 Metropolitan Police Not at junction or within 20 metres
## 6 Metropolitan Police Not at junction or within 20 metres
##                       Lighting               Weather     Surface
## 1                     Daylight Raining no high winds Wet or damp
## 2        Darkness - lights lit    Fine no high winds         Dry
## 3        Darkness - lights lit    Fine no high winds         Dry
## 4        Darkness - lights lit    Fine no high winds         Dry
## 5                     Daylight    Fine no high winds         Dry
## 6 Darkness - lighting unknown    Fine no high winds Wet or damp
##                                 Vehicle                   Manoeuvre Skidding
## 1                                   Car           Going ahead other     None
## 2 Bus or coach (17 or more pass seats)         Slowing or stopping     None
## 3 Bus or coach (17 or more pass seats) Going ahead right-hand bend     None
## 4                                   Car                      Parked     None
## 5                                   Car           Going ahead other     None
## 6           Motorcycle 125cc and under           Going ahead other  Skidded
##                      Journey Age_Band Casualty_Outcome Year Month
## 1 Other/Not known (2005-10)  66 - 75          Serious 2005    04
## 2   Journey as part of work  36 - 45           Slight 2005    05
## 3   Journey as part of work  26 - 35           Slight 2005    06
## 4 Other/Not known (2005-10)  56 - 65           Slight 2005    06
## 5 Other/Not known (2005-10)  46 - 55           Slight 2005    07
## 6 Other/Not known (2005-10)  46 - 55           Slight 2005    10
```

```
# write.csv(df, file = "filtered_eported_data.csv")
```

What do you not know how to do right now that you need to learn to import and cleanup your dataset?

- In above steps data sets are merged and final draft has been displayed using head() command above.
- Date column must have "NA" values, during next step use appropriate filter to pick selected data.

Discuss how you plan to uncover new information in the data that is not self-evident.

- Data columns not relevant to address problem questions and with no co-relations ewith variables has already been eliminated in steps above.

What are different ways you could look at this data to answer the questions you want to answer?

- Very first step is to look at data, can trust data source and set, data populated correctly, noise is data (missing,NA or NULL), identify variables and co-rleations among them, will it address out problem questions ?

Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.

- Identify joining column to merge data from all sources, used description excel to replaced selected columns code with meaningful description given in excel. Change data type od Date filed from string to date. Create new valirable Month and Year for future aggregation and reporting.

How could you summarize your data to answer key questions?

- Once data is ready, next steps will be to use different plotting functions and understand co-orelations between various variables, and try to get problem questions addresses and visualize relationship using various plotting options in next steps.

Required Packages

Base packages plus "ggplot2", "dplyr", "magrittr", "tidyverse", "broom", "purrr", "GGally", "scales", "reshape", "moments", "ggpubr", "readxl" .. and more on need basis.

Plots and Table Needs

scatter plots, time-series plot and histograms to analyze and visualize the data patterns.

What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

- In addtion to packages, I have create seperate function plotHistogram() - Mentions above in function declaration chunk. with all input parameters it will plot histogram with x and y axis label with title.

- Will Standardize same for scatterplots and more based on project need in coming weeks.

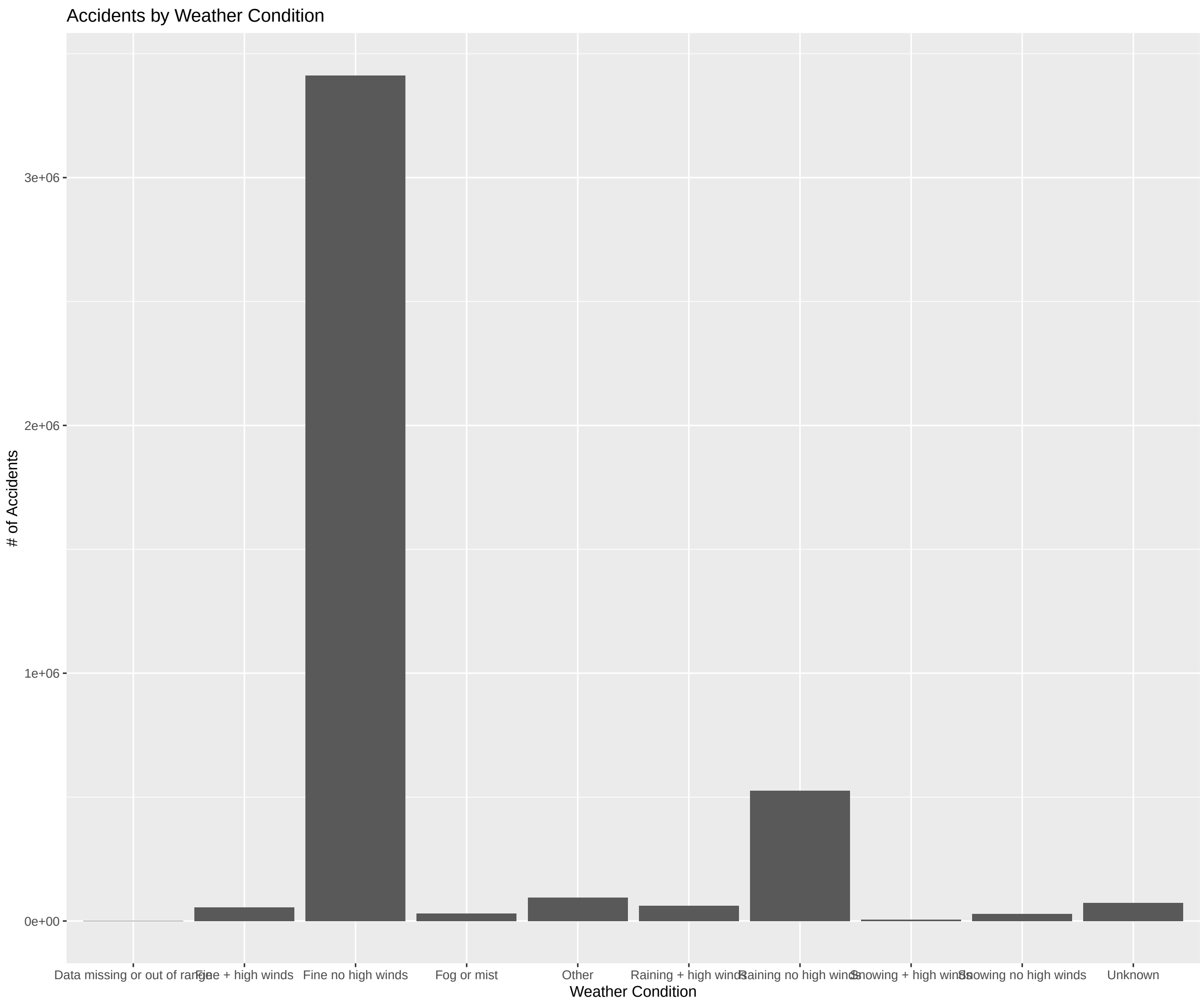
Questions for future steps

- Wide data set, wrangling will be challenging to all together at one place and pick selective columns to address out research questions, In addition fear of unknown as we move forward.
- Will focus on flitering to eliminate no needed data to generate meaningful outcome to address problem questions.

*********** NOTE - Scratch work *****************

THis section is not to demonstrate any thing, its scratch section to play around dataset to understand variables and its co-relations and try out plot function to learn about it. it can be consider as a prep work for final week project.

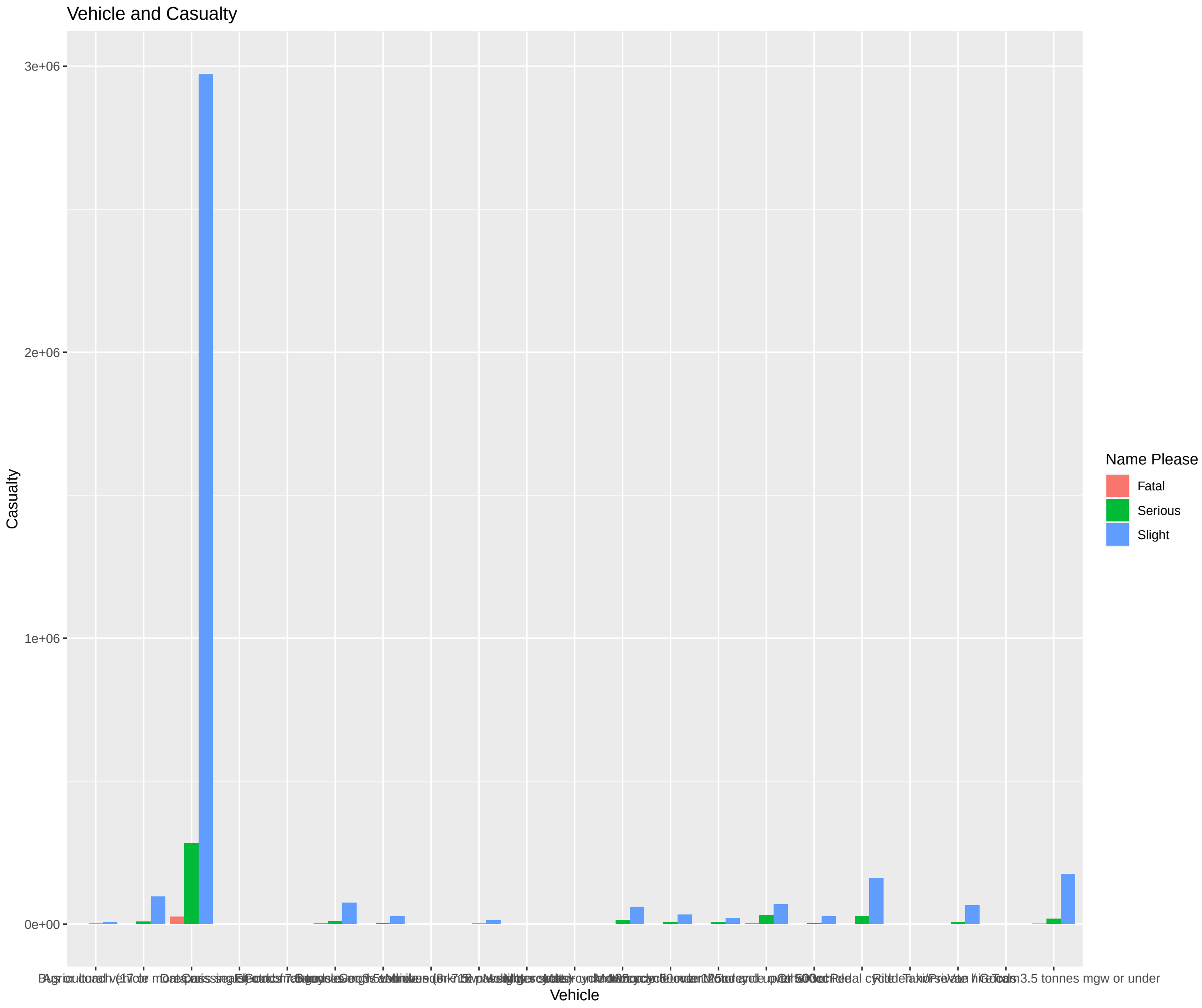
```
# unique(df$Weather)
# any(is.na(df$Vehicle))
# any(is.na(df$Vehicle))

plotHistogram(df, df$Weather, df$Accident_Severity, "Weather Condition", "# of Accidents", "Accidents by Weather
```



```
unique(df$Vehicle)
```

```
##  [1] "Car"
##  [2] "Bus or coach (17 or more pass seats)"
##  [3] "Motorcycle 125cc and under"
##  [4] "Other vehicle"
##  [5] "Motorcycle over 500cc"
##  [6] "Pedal cycle"
##  [7] "Van / Goods 3.5 tonnes mgw or under"
##  [8] "Motorcycle over 125cc and up to 500cc"
##  [9] "Taxi/Private hire car"
## [10] "Goods 7.5 tonnes mgw and over"
## [11] "Goods over 3.5t. and under 7.5t"
## [12] "Motorcycle 50cc and under"
## [13] "Minibus (8 - 16 passenger seats)"
## [14] "Agricultural vehicle"
## [15] "Tram"
## [16] "Ridden horse"
## [17] "Data missing or out of range"
## [18] "Motorcycle - unknown cc"
## [19] "Mobility scooter"
## [20] "Goods vehicle - unknown weight"
## [21] "Electric motorcycle"
```

```
plotHistogram(df, df$Vehicle, df$Casualty_Outcome, "Vehicle", "Casualty", "Vehicle and Casualty", "Name Please")
```



```
any(is.na(df$Year))
```

```
## [1] TRUE
```

```
any(is.null(df$Year))
```

```
## [1] FALSE
```

```
any(is.na(df$Date))
```

```
## [1] TRUE
```

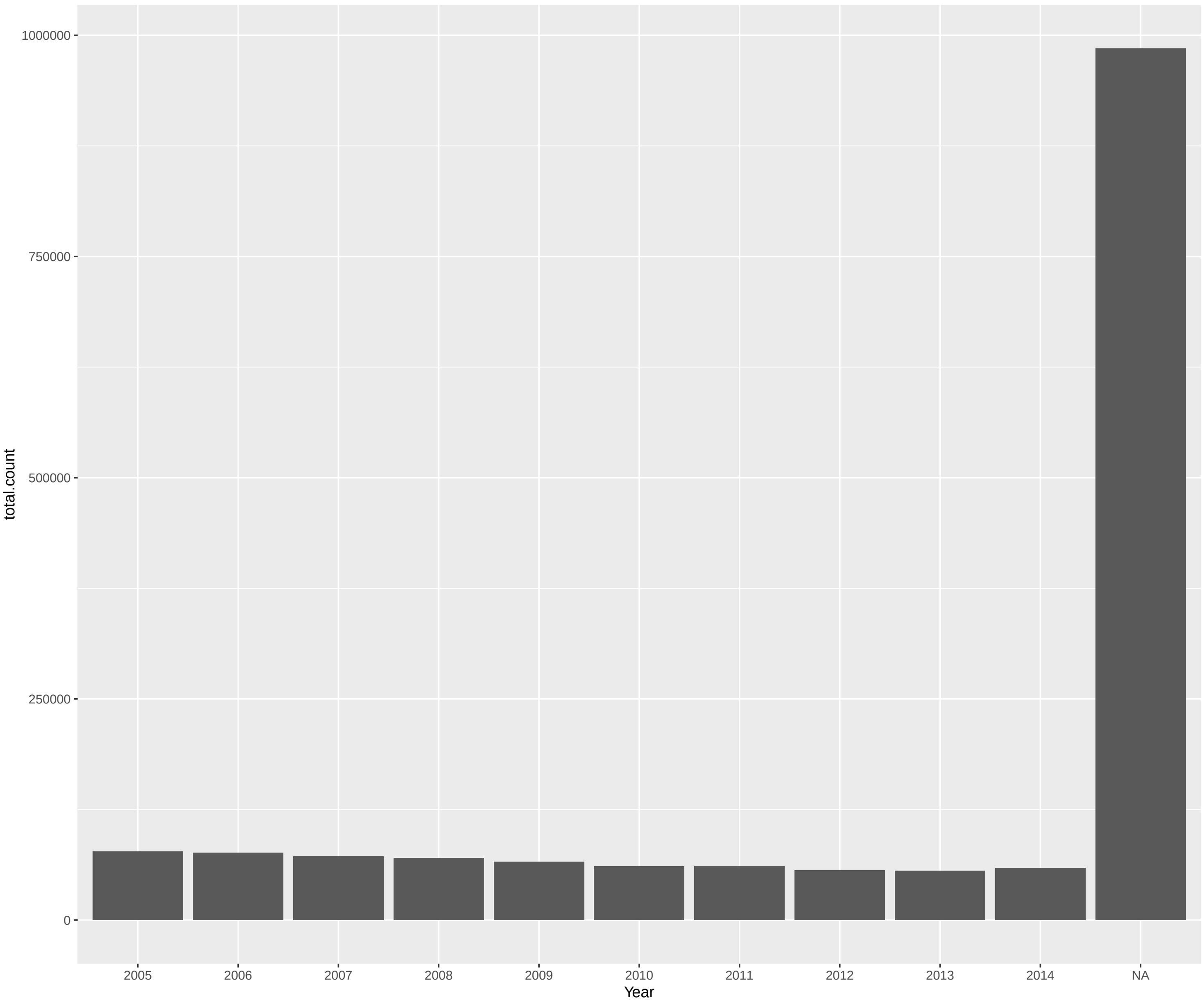
```
colSums(is.na(df))
```

```
##                          Accident_Index                   Location_Easting_OSGR
##                                       0                                     256
##                  Location_Northing_OSGR                               Longitude
##                                     256                                     256
##                                Latitude                            Police_Force
##                                     256                                       0
##                       Accident_Severity                      Number_of_Vehicles
##                                       0                                       0
##                    Number_of_Casualties                                    Date
##                                       0                                 2578146
##                             Day_of_Week                                    Time
##                                       0                                       0
##              Local_Authority_.District.               Local_Authority_.Highway.
##                                       0                                       0
##                         X1st_Road_Class                        X1st_Road_Number
##                                       0                                       0
##                               Road_Type                             Speed_limit
##                                       0                                       0
##                         X2nd_Road_Class                        X2nd_Road_Number
##                                       0                                       0
##       Pedestrian_Crossing.Human_Control Pedestrian_Crossing.Physical_Facilities
##                                       0                                       0
##                        Light_Conditions                      Weather_Conditions
##                                       0                                       0
##                 Road_Surface_Conditions              Special_Conditions_at_Site
##                                       0                                       0
##                     Vehicle_Reference.x                          Casualty_Class
##                                       0                                       0
##                         Sex_of_Casualty                         Age_of_Casualty
##                                       0                                       0
##                    Age_Band_of_Casualty                       Casualty_Severity
##                                       0                                       0
##                           Car_Passenger                  Bus_or_Coach_Passenger
##                                       0                                       0
##                           Casualty_Type                 Casualty_Home_Area_Type
##                                       0                                       0
##                     Vehicle_Reference.y                            Vehicle_Type
##                                       0                                       0
##                       Vehicle_Manoeuvre        Vehicle_Location.Restricted_Lane
##                                       0                                       0
##                Skidding_and_Overturning                    X1st_Point_of_Impact
##                                       0                                       0
##            Was_Vehicle_Left_Hand_Drive.               Journey_Purpose_of_Driver
##                                       0                                       0
##                         Sex_of_Driver                          Age_of_Driver
##                                     0                                      0
##                   Engine_Capacity_.CC.                        Age_of_Vehicle
##                                     0                                      0
##                              Location                               Junction
##                                     0                                      0
##                              Lighting                                Weather
##                                     0                                      0
##                               Surface                                Vehicle
##                                     0                                      0
##                             Manoeuvre                               Skidding
##                                     0                                      0
##                               Journey                               Age_Band
##                                     0                                      0
##                      Casualty_Outcome                                   Year
##                                     0                                2578146
##                                 Month
##                               2578146
```

```
sum(is.na(df$Date))
```

```
## [1] 2578146
```

```
df1 <- df %>% select(Accident_Index, Location, Accident_Severity, Number_of_Vehicles, Number_of_Casualties, Date

by_year_count <- df1 %>% select(Accident_Index, Year) %>% group_by(Year) %>% dplyr::summarise(total.count = n())

chart1 <- ggplot(data=by_year_count, aes(x=Year, y=total.count)) + geom_bar(stat="identity")
chart1
```



1000000
750000
500000
250000
0
total.count
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
NA
Year