Data Wrangling

Rohan Kalu

Data Wrangling

- Load necessary package

```
library(dplyr)
```

```
##
## 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
```

```
library(magrittr)
library(ggplot2)
library(readr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

- Load the dataset

```
accidents <- read.csv("data/accidents.csv")
guidance <- read.csv("data/guidance.csv")
```

- Check the first 6 data of the csv

```
head(accidents)
```

```
##   Number.of.Vehicles Accident.Date Time..24hr. X1st.Road.Class Road.Surface
## 1                  2      1/1/2017        2120               U     Wet/Damp
## 2                  2      4/1/2017        1500               U          Dry
## 3                  2      5/1/2017         732            A58     Wet/Damp
## 4                  2      5/1/2017         930           A646     Wet/Damp
## 5                  2    14/01/2017         909               U    Frost/Ice
## 6                  1    15/01/2017        1659               U     Wet/Damp
##   Lighting.Conditions Daylight.Dark Weather.Conditions Local.Authority
## 1                   4          Dark                  2      Calderdale
## 2                   1      Daylight                  1      Calderdale
## 3                   4          Dark                  1      Calderdale
## 4                   1      Daylight                  1      Calderdale
## 5                   1      Daylight                  1      Calderdale
## 6                   4          Dark                  1      Calderdale
##   Type.of.Vehicle Casualty.Class Casualty.Severity Sex.of.Casualty
## 1               9              2                 2               2
## 2               9              3                 2               2
## 3               9              1                 3               1
## 4               4              1                 1               1
## 5               9              1                 3               1
## 6               9              3                 3               1
##   Age.of.Casualty
## 1              16
## 2              67
## 3              56
## 4              20
## 5              46
## 6              NA
```

- Check dimension, structure and summary of the data

```r
dim(accidents)
```

```
## [1] 2069   14
```

```r
str(accidents)
```

```
## 'data.frame':    2069 obs. of  14 variables:
##  $ Number.of.Vehicles : int  2 2 2 2 2 1 1 3 1 1 ...
##  $ Accident.Date      : chr  "1/1/2017" "4/1/2017" "5/1/2017" "5/1/2017" ...
##  $ Time..24hr.        : int  2120 1500 732 930 909 1659 1059 1849 1408 1325 ...
##  $ X1st.Road.Class    : chr  "U" "U" "A58" "A646" ...
##  $ Road.Surface       : chr  "Wet/Damp" "Dry" "Wet/Damp" "Wet/Damp" ...
##  $ Lighting.Conditions: int  4 1 4 1 1 4 1 4 1 1 ...
##  $ Daylight.Dark      : chr  "Dark" "Daylight" "Dark" "Daylight" ...
##  $ Weather.Conditions : int  2 1 1 1 1 1 1 1 1 1 ...
##  $ Local.Authority    : chr  "Calderdale" "Calderdale" "Calderdale" "Calderdale" ...
##  $ Type.of.Vehicle    : int  9 9 9 4 9 9 9 9 9 9 ...
##  $ Casualty.Class     : int  2 3 1 1 1 3 3 1 3 1 ...
##  $ Casualty.Severity  : int  2 2 3 1 3 3 3 3 3 3 ...
##  $ Sex.of.Casualty    : int  2 2 1 1 1 1 2 2 2 2 ...
##  $ Age.of.Casualty    : int  16 67 56 20 46 NA 25 50 64 22 ...
```

```
summary(accidents)
```

```
##  Number.of.Vehicles Accident.Date       Time..24hr.   X1st.Road.Class
##  Min.   :1.000      Length:2069        Min.   :   0   Length:2069
##  1st Qu.:1.000      Class :character   1st Qu.:1045   Class :character
##  Median :2.000      Mode  :character   Median :1500   Mode  :character
##  Mean   :1.906                         Mean   :1405
##  3rd Qu.:2.000                         3rd Qu.:1755
##  Max.   :7.000                         Max.   :2350
##
##  Road.Surface       Lighting.Conditions Daylight.Dark      Weather.Conditions
##  Length:2069        Min.   :1.000       Length:2069        Min.   :1.000
##  Class :character   1st Qu.:1.000       Class :character   1st Qu.:1.000
##  Mode  :character   Median :1.000       Mode  :character   Median :1.000
##                     Mean   :2.015                          Mean   :1.464
##                     3rd Qu.:4.000                          3rd Qu.:1.000
##                     Max.   :7.000                          Max.   :9.000
##
##  Local.Authority    Type.of.Vehicle  Casualty.Class  Casualty.Severity
##  Length:2069        Min.   : 1.000   Min.   :1.000   Min.   :1.000
##  Class :character   1st Qu.: 9.000   1st Qu.:1.000   1st Qu.:3.000
##  Mode  :character   Median : 9.000   Median :1.000   Median :3.000
##                     Mean   : 8.917   Mean   :1.591   Mean   :2.831
##                     3rd Qu.: 9.000   3rd Qu.:2.000   3rd Qu.:3.000
##                     Max.   :97.000   Max.   :3.000   Max.   :3.000
##
##  Sex.of.Casualty Age.of.Casualty
##  Min.   :1.000   Min.   :  1.00
##  1st Qu.:1.000   1st Qu.: 21.00
##  Median :1.000   Median : 33.00
##  Mean   :1.395   Mean   : 36.21
##  3rd Qu.:2.000   3rd Qu.: 49.00
##  Max.   :2.000   Max.   :115.00
##                  NA's   :19
```

Data cleaning

- Change the data type of accident.date into

```
accidents <- accidents %>%
  mutate(Accident.Date = as.Date(Accident.Date, format="%d/%m/%Y"))
```

- Change Time..24hr. into 24 hr format

```
accidents$`Time..24hr.` <- format(as.POSIXct(sprintf("%04d", accidents$`Time..24hr.`),
                                             format="%H%M", tz = "UTC"), format="%H:%M", usetz = FALSE)
head(accidents)
```

```
##   Number.of.Vehicles Accident.Date Time..24hr. X1st.Road.Class Road.Surface
## 1                  2    2017-01-01       21:20               U     Wet/Damp
## 2                  2    2017-01-04       15:00               U          Dry
```

```
## 3                  2    2017-01-05       07:32             A58     Wet/Damp
## 4                  2    2017-01-05       09:30            A646     Wet/Damp
## 5                  2    2017-01-14       09:09               U    Frost/Ice
## 6                  1    2017-01-15       16:59               U     Wet/Damp
##   Lighting.Conditions Daylight.Dark Weather.Conditions Local.Authority
## 1                   4          Dark                  2      Calderdale
## 2                   1      Daylight                  1      Calderdale
## 3                   4          Dark                  1      Calderdale
## 4                   1      Daylight                  1      Calderdale
## 5                   1      Daylight                  1      Calderdale
## 6                   4          Dark                  1      Calderdale
##   Type.of.Vehicle Casualty.Class Casualty.Severity Sex.of.Casualty
## 1               9              2                 2               2
## 2               9              3                 2               2
## 3               9              1                 3               1
## 4               4              1                 1               1
## 5               9              1                 3               1
## 6               9              3                 3               1
##   Age.of.Casualty
## 1              16
## 2              67
## 3              56
## 4              20
## 5              46
## 6              NA
```

EDA

- Check for missing values

```r
# Replace empty strings with NA in all columns
accidents[accidents == ""] <- NA
sapply(accidents, function(x) sum(is.na(x)))
```

```
##  Number.of.Vehicles       Accident.Date         Time..24hr.     X1st.Road.Class
##                   0                   0                   0                   0
##        Road.Surface Lighting.Conditions       Daylight.Dark  Weather.Conditions
##                   0                   0                  18                   0
##     Local.Authority     Type.of.Vehicle      Casualty.Class   Casualty.Severity
##                   0                   0                   0                   0
##     Sex.of.Casualty     Age.of.Casualty
##                   0                  19
```

- Look for the missing values

```r
missing_age <- accidents %>%
  filter(is.na(Age.of.Casualty))

# Filter rows with missing Daylight.Dark
missing_daylight.dark <- accidents %>%
  filter(is.na(Daylight.Dark))

missing_age
```

```
##    Number.of.Vehicles Accident.Date Time..24hr. X1st.Road.Class Road.Surface
## 1                   1    2017-01-15       16:59              U     Wet/Damp
## 2                   2    2017-01-27       18:35              U     Wet/Damp
## 3                   1    2017-02-04       17:30           A646          Dry
## 4                   2    2017-03-26       13:53              U          Dry
## 5                   1    2017-05-01       14:54              U          Dry
## 6                   2    2017-06-20       07:52            A58          Dry
## 7                   1    2017-07-24       23:28              U          Dry
## 8                   1    2017-10-03       08:50            A58     Wet/Damp
## 9                   1    2017-10-18       10:07              U          Dry
## 10                  2    2017-11-20       19:30              U     Wet/Damp
## 11                  2    2016-02-26       13:40              6            1
## 12                  1    2016-05-21       02:10              3            2
## 13                  4    2016-10-27       17:35              1            1
## 14                  2    2015-01-23       18:25              6            2
## 15                  2    2015-08-03       14:03              3            1
## 16                  1    2015-11-07       22:30              6            2
## 17                  2    2015-11-30       17:00              3            2
## 18                  2    2014-01-14       13:50              3            2
## 19                  1    2014-05-09       08:24              3            2
##    Lighting.Conditions Daylight.Dark Weather.Conditions Local.Authority
## 1                    4          Dark                  1      Calderdale
## 2                    4          Dark                  1      Calderdale
## 3                    4          Dark                  1      Calderdale
## 4                    1      Daylight                  1      Calderdale
## 5                    1      Daylight                  1      Calderdale
## 6                    1      Daylight                  1      Calderdale
## 7                    4          Dark                  1      Calderdale
## 8                    1      Daylight                  1      Calderdale
## 9                    1      Daylight                  1      Calderdale
## 10                   4          Dark                  1      Calderdale
## 11                   1      Daylight                  1      Calderdale
## 12                   4          Dark                  2      Calderdale
## 13                   1      Daylight                  1      Calderdale
## 14                   4          Dark                  5      Calderdale
## 15                   1      Daylight                  1      Calderdale
## 16                   4          Dark                  1      Calderdale
## 17                   4          Dark                  2      Calderdale
## 18                   1      Daylight                  1      Calderdale
## 19                   1      Daylight                  2      Calderdale
##    Type.of.Vehicle Casualty.Class Casualty.Severity Sex.of.Casualty
## 1                9              3                 3               1
## 2                9              1                 3               1
## 3                9              3                 3               1
## 4                2              1                 2               1
## 5                9              3                 2               2
## 6                4              1                 3               1
## 7                9              1                 3               2
## 8                9              3                 3               1
## 9                9              3                 3               1
## 10               9              1                 3               2
## 11               9              1                 3               1
## 12               9              2                 2               1
## 13               9              2                 3               1
```

```
## 14                9              2                 3               2
## 15                4              1                 3               1
## 16                9              3                 3               2
## 17                5              1                 2               1
## 18                1              1                 3               1
## 19                9              3                 2               1
##    Age.of.Casualty
## 1               NA
## 2               NA
## 3               NA
## 4               NA
## 5               NA
## 6               NA
## 7               NA
## 8               NA
## 9               NA
## 10              NA
## 11              NA
## 12              NA
## 13              NA
## 14              NA
## 15              NA
## 16              NA
## 17              NA
## 18              NA
## 19              NA
```

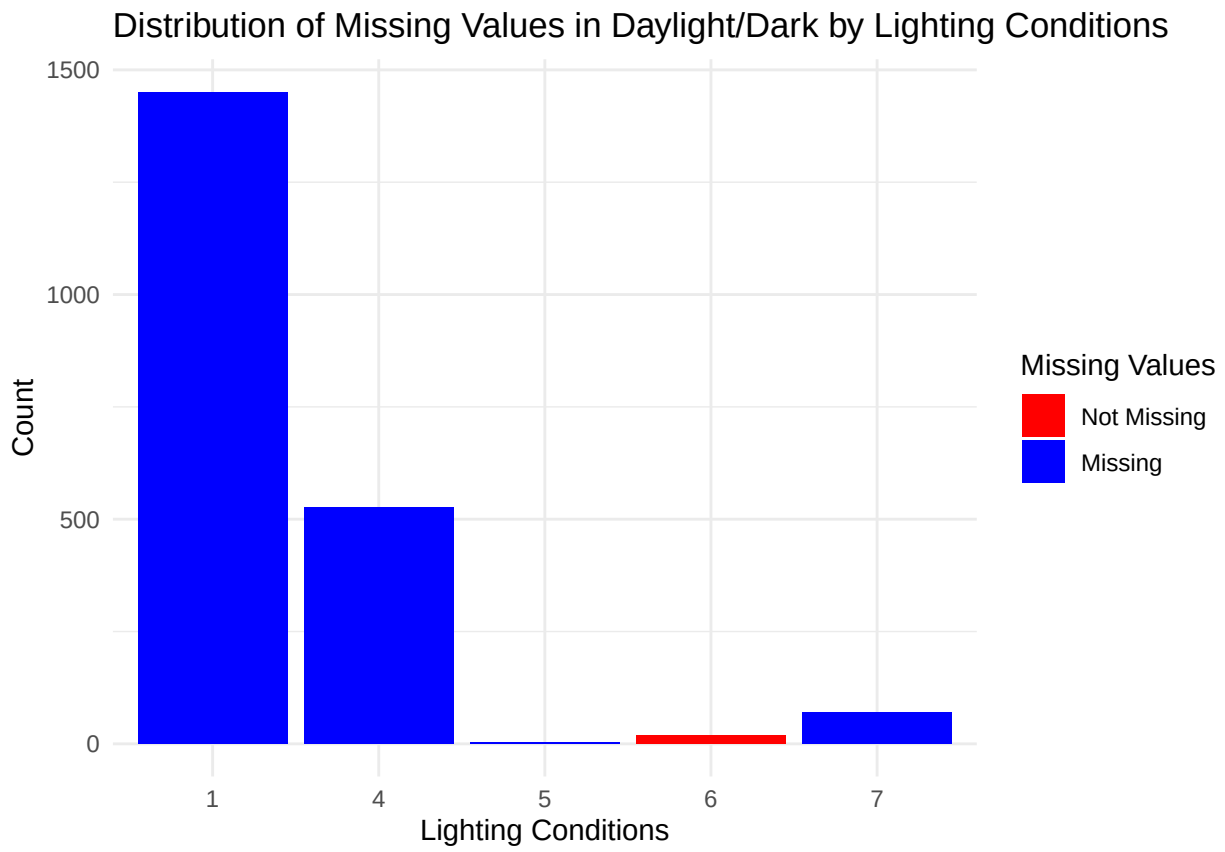
```
missing_daylight.dark
```

```
##    Number.of.Vehicles Accident.Date Time..24hr. X1st.Road.Class  Road.Surface
## 1                   1    2017-02-18       23:30               U      Wet/Damp
## 2                   1    2017-02-22       19:39            A681      Wet/Damp
## 3                   1    2017-06-01       03:25               U           Dry
## 4                   2    2017-08-18       21:00               U           Dry
## 5                   2    2017-11-08       16:51               U Wet \xa8 Damp
## 6                   1    2016-01-17       20:40               6             2
## 7                   2    2016-03-10       06:57               6             2
## 8                   1    2016-05-07       22:07               3             2
## 9                   1    2015-03-30       19:55               6             2
## 10                  1    2015-05-31       01:50               6             2
## 11                  1    2015-08-28       23:45               6             1
## 12                  1    2015-08-28       23:45               6             1
## 13                  1    2015-11-22       11:36               6             1
## 14                  1    2015-11-22       11:36               6             1
## 15                  1    2014-10-09       21:00               3             2
## 16                  1    2014-11-23       06:40               3             2
## 17                  1    2014-11-23       06:40               3             2
## 18                  1    2014-12-29       07:35               6             4
##    Lighting.Conditions Daylight.Dark Weather.Conditions Local.Authority
## 1                    6         <NA>                  1      Calderdale
## 2                    6         <NA>                  2      Calderdale
## 3                    6         <NA>                  1      Calderdale
## 4                    6         <NA>                  4      Calderdale
## 5                    6         <NA>                  1      Calderdale
```

```
## 6                    6          <NA>                  1      Calderdale
## 7                    6          <NA>                  1      Calderdale
## 8                    6          <NA>                  1      Calderdale
## 9                    6          <NA>                  5      Calderdale
## 10                   6          <NA>                  2      Calderdale
## 11                   6          <NA>                  1      Calderdale
## 12                   6          <NA>                  1      Calderdale
## 13                   6          <NA>                  1      Calderdale
## 14                   6          <NA>                  1      Calderdale
## 15                   6          <NA>                  5      Calderdale
## 16                   6          <NA>                  1      Calderdale
## 17                   6          <NA>                  1      Calderdale
## 18                   6          <NA>                  1      Calderdale
##    Type.of.Vehicle Casualty.Class Casualty.Severity Sex.of.Casualty
## 1                8              3                 2               2
## 2                3              1                 2               1
## 3                9              1                 3               1
## 4                9              1                 3               1
## 5                9              1                 3               1
## 6               90              1                 2               1
## 7                9              1                 3               1
## 8                9              2                 3               1
## 9                9              1                 3               1
## 10               8              3                 3               1
## 11               9              1                 3               1
## 12               9              2                 3               1
## 13               9              2                 3               1
## 14               9              1                 3               1
## 15               9              1                 3               1
## 16               9              2                 3               1
## 17               9              1                 3               1
## 18               9              1                 3               1
##    Age.of.Casualty
## 1               39
## 2               30
## 3               28
## 4               49
## 5               40
## 6               33
## 7               35
## 8               19
## 9               65
## 10              19
## 11              21
## 12              18
## 13              20
## 14              20
## 15              41
## 16              19
## 17              18
## 18              36
```

```r
# Convert `Daylight.Dark` to factor to handle missing values
accidents$Daylight.Dark <- factor(accidents$Daylight.Dark)
```

```r
# Plotting using ggplot
ggplot(accidents, aes(x = factor(Lighting.Conditions), fill = is.na(Daylight.Dark))) +
  geom_bar(position = "stack") +
  labs(title = "Distribution of Missing Values in Daylight/Dark by Lighting Conditions",
       x = "Lighting Conditions",
       y = "Count",
       fill = "Missing Values") +
  scale_fill_manual(values = c("TRUE" = "red", "FALSE" = "blue"),
                    labels = c("Missing", "Not Missing"),
                    guide = guide_legend(reverse = TRUE)) +
  theme_minimal()
```



Data Exploration

- Check for anamolies

```r
unique(accidents$Road.Surface)
```

```
##  [1] "Wet/Damp"      "Dry"           "Frost/Ice"     "Ice"
##  [5] "Snow"          "Wet"           "Wet \xa8 Damp" "2"
##  [9] "1"             "3"             "4"             "5"
```

```r
unique(accidents$X1st.Road.Class)
```

```
##  [1] "U"        "A58"      "A646"     "B6138"    "A629"     "A641"
```

```
##  [7] "A672"     "A6033"    "A6139"    "A644"     "A62"      "B6114"
## [13] "A6319"    "B6112"    "M62"      "A681"     "B6113"    "A629(M)"
## [19] "A643"     "A6036"    "A6025"    "A647"     "A6026(M)" "A649"
## [25] "A6026"    "3"        "6"        "1"        "4"        "2"
```

```
accidents <- accidents %>%
  mutate(
    X1st.Road.Class = case_when(
      X1st.Road.Class %in% c("U") ~ "Unclassified",
      grepl("^A\\d+", X1st.Road.Class) ~ "A",
      grepl("^A\\(M\\)$", X1st.Road.Class) ~ "A(M)",
      grepl("^A\\d+\\(M\\)$", X1st.Road.Class) ~ "A(M)",
      grepl("^B\\d+", X1st.Road.Class) ~ "B",
      X1st.Road.Class %in% c("Motorway", "A", "B", "C", "Unclassified", "A(M)") ~ X1st.Road.Class,
      X1st.Road.Class %in% c("M62") ~ "Motorway",
      TRUE ~ as.character(X1st.Road.Class)
    ),
    `Road.Surface` = case_when(
      Road.Surface %in% c("Wet/Damp", "Wet \xa8 Damp", "Wet") ~ "Wet/Damp",
      Road.Surface == "Frost/Ice" ~ "Frost / Ice",
      Road.Surface == "Ice" ~ "Ice",
      Road.Surface == "Snow" ~ "Snow",
      Road.Surface == "1" ~ "Dry",
      Road.Surface == "2" ~ "Wet / Damp",
      Road.Surface == "3" ~ "Snow",
      Road.Surface == "4" ~ "Frost / Ice",
      Road.Surface == "5" ~ "Flood (surface water over 3cm deep)",
      TRUE ~ as.character(Road.Surface)
    ),
    `Lighting.Conditions` = case_when(
      Lighting.Conditions == 1 ~ "Daylight: street lights present",
      Lighting.Conditions == 2 ~ "Daylight: no street lighting",
      Lighting.Conditions == 3 ~ "Daylight: street lighting unknown",
      Lighting.Conditions == 4 ~ "Darkness: street lights present and lit",
      Lighting.Conditions == 5 ~ "Darkness: street lights present but unlit",
      Lighting.Conditions == 6 ~ "Darkness: no street lighting",
      Lighting.Conditions == 7 ~ "Darkness: street lighting unknown",
      TRUE ~ as.character(Lighting.Conditions)
    ),
    `Weather.Conditions` = case_when(
      Weather.Conditions == 1 ~ "Fine without high winds",
      Weather.Conditions == 2 ~ "Raining without high winds",
      Weather.Conditions == 3 ~ "Snowing without high winds",
      Weather.Conditions == 4 ~ "Fine with high winds",
      Weather.Conditions == 5 ~ "Raining with high winds",
      Weather.Conditions == 6 ~ "Snowing with high winds",
      Weather.Conditions == 7 ~ "Fog or mist - if hazard",
      Weather.Conditions == 8 ~ "Other",
      Weather.Conditions == 9 ~ "Unknown",
      TRUE ~ as.character(Weather.Conditions)
    ),
    `Casualty.Class` = case_when(
      Casualty.Class == 1 ~ "Driver or rider",
      Casualty.Class == 2 ~ "Vehicle or pillion passenger",
```

```
      Casualty.Class == 3 ~ "Pedestrian",
      TRUE ~ as.character(Casualty.Class)
    ),
    `Casualty.Severity` = case_when(
      Casualty.Severity == 1 ~ "Fatal",
      Casualty.Severity == 2 ~ "Serious",
      Casualty.Severity == 3 ~ "Slight",
      TRUE ~ as.character(Casualty.Severity)
    ),
    `Sex.of.Casualty` = case_when(
      Sex.of.Casualty == 1 ~ "Male",
      Sex.of.Casualty == 2 ~ "Female",
      TRUE ~ as.character(Sex.of.Casualty)
    )
  )
```

- Check for anamolies again

```
unique(accidents$X1st.Road.Class)
```

```
## [1] "Unclassified" "A"            "B"            "Motorway"     "3"
## [6] "6"            "1"            "4"            "2"
```

```
unique(accidents$Road.Surface)
```

```
## [1] "Wet/Damp"                           "Dry"
## [3] "Frost / Ice"                        "Ice"
## [5] "Snow"                               "Wet / Damp"
## [7] "Flood (surface water over 3cm deep)"
```

```
head(accidents)
```

```
##   Number.of.Vehicles Accident.Date Time..24hr. X1st.Road.Class Road.Surface
## 1                  2    2017-01-01       21:20    Unclassified     Wet/Damp
## 2                  2    2017-01-04       15:00    Unclassified          Dry
## 3                  2    2017-01-05       07:32               A     Wet/Damp
## 4                  2    2017-01-05       09:30               A     Wet/Damp
## 5                  2    2017-01-14       09:09    Unclassified  Frost / Ice
## 6                  1    2017-01-15       16:59    Unclassified     Wet/Damp
##                      Lighting.Conditions Daylight.Dark
## 1 Darkness: street lights present and lit          Dark
## 2         Daylight: street lights present      Daylight
## 3 Darkness: street lights present and lit          Dark
## 4         Daylight: street lights present      Daylight
## 5         Daylight: street lights present      Daylight
## 6 Darkness: street lights present and lit          Dark
##           Weather.Conditions Local.Authority Type.of.Vehicle
## 1 Raining without high winds      Calderdale               9
## 2    Fine without high winds      Calderdale               9
## 3    Fine without high winds      Calderdale               9
## 4    Fine without high winds      Calderdale               4
```

```
## 5    Fine without high winds      Calderdale               9
## 6    Fine without high winds      Calderdale               9
##                 Casualty.Class Casualty.Severity Sex.of.Casualty
## 1 Vehicle or pillion passenger           Serious          Female
## 2                   Pedestrian           Serious          Female
## 3              Driver or rider            Slight            Male
## 4              Driver or rider             Fatal            Male
## 5              Driver or rider            Slight            Male
## 6                   Pedestrian            Slight            Male
##   Age.of.Casualty
## 1              16
## 2              67
## 3              56
## 4              20
## 5              46
## 6              NA
```

```
unique(accidents$Weather.Conditions)
```

```
## [1] "Raining without high winds" "Fine without high winds"
## [3] "Raining with high winds"    "Fog or mist - if hazard"
## [5] "Snowing without high winds" "Snowing with high winds"
## [7] "Fine with high winds"       "Other"
## [9] "Unknown"
```

- Check for anamolies on other columns

```
exclude_cols <- c("Age.of.Casualty", "Accident.Date", "Time..24hr.")

# Select columns except Age, Date, and Time
selected_cols <- accidents[, !names(accidents) %in% exclude_cols]

# Check unique values for each selected column using sapply
unique_values <- sapply(selected_cols, function(x) length(unique(x)))

# Print the unique values for each column
unique_values
```

```
##  Number.of.Vehicles     X1st.Road.Class        Road.Surface Lighting.Conditions
##                   7                   9                   7                   5
##       Daylight.Dark  Weather.Conditions     Local.Authority     Type.of.Vehicle
##                   3                   9                   1                  19
##      Casualty.Class   Casualty.Severity     Sex.of.Casualty
##                   3                   3                   2
```

```
accidents <- accidents %>%
  select(-Local.Authority, -Daylight.Dark)
```

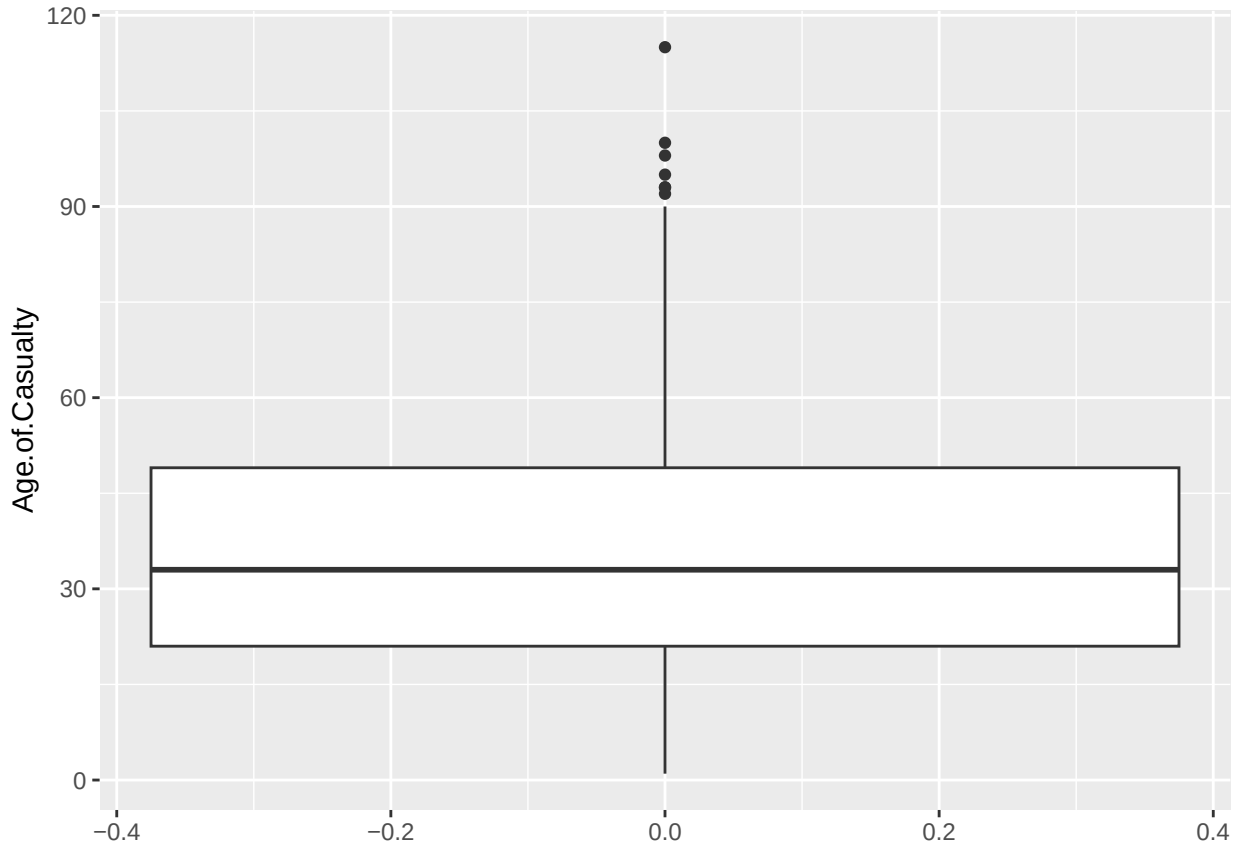
```
colnames(accidents)
```

```
##  [1] "Number.of.Vehicles"  "Accident.Date"       "Time..24hr."
##  [4] "X1st.Road.Class"     "Road.Surface"        "Lighting.Conditions"
##  [7] "Weather.Conditions"  "Type.of.Vehicle"     "Casualty.Class"
## [10] "Casualty.Severity"   "Sex.of.Casualty"     "Age.of.Casualty"
```

Outliers Detection

- Use Boxplot to detect outliers

```
filtered_data <- accidents[!is.na(accidents$Age.of.Casualty), ]

# Visual inspection with boxplot
ggplot(filtered_data, aes(y = Age.of.Casualty)) +
  geom_boxplot()
```



- Use 3-Sigma method

```
age_of_casualty <- accidents$Age.of.Casualty
mean_age = mean(age_of_casualty, na.rm = TRUE)
sd_age = sd(age_of_casualty, na.rm = TRUE)
upper_bound <- mean_age + 3 * sd_age
lower_bound <- mean_age - 3 * sd_age
outliers_dataset <- accidents %>%
  filter(Age.of.Casualty < lower_bound | Age.of.Casualty > upper_bound)
number_of_outliers_sigma <- nrow(outliers_dataset)
print(paste("Number of outliers:", number_of_outliers_sigma))
```

```
## [1] "Number of outliers: 4"
```

```r
print(outliers_dataset)
```

```
##   Number.of.Vehicles Accident.Date Time..24hr. X1st.Road.Class Road.Surface
## 1                  2    2017-01-31       18:40    Unclassified     Wet/Damp
## 2                  1    2016-06-16       12:15               6          Dry
## 3                  1    2014-02-05       17:15               6   Wet / Damp
## 4                  2    2014-06-08       16:50               3          Dry
##                        Lighting.Conditions         Weather.Conditions
## 1 Darkness: street lights present but unlit Raining without high winds
## 2           Daylight: street lights present    Fine without high winds
## 3   Darkness: street lights present and lit    Fine without high winds
## 4           Daylight: street lights present    Fine without high winds
##   Type.of.Vehicle               Casualty.Class Casualty.Severity
## 1               9              Driver or rider           Serious
## 2               9                   Pedestrian            Slight
## 3               8                   Pedestrian           Serious
## 4               9 Vehicle or pillion passenger            Slight
##   Sex.of.Casualty Age.of.Casualty
## 1          Female             115
## 2            Male             100
## 3            Male              95
## 4          Female              98
```

- IQR

```r
# Calculate interquartile range (IQR) for Age_of_Casualty
Q1 <- quantile(filtered_data$Age.of.Casualty, 0.25, na.rm = TRUE)
Q3 <- quantile(filtered_data$Age.of.Casualty, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1

# Calculate lower and upper bounds for outliers
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# Identify outliers
outliers <- filtered_data[filtered_data$Age.of.Casualty < lower_bound | filtered_data$Age.of.Casualty >

# Print count of outliers
outliers  # or use dim(outliers)[1] for the number of rows
```

```
##      Number.of.Vehicles Accident.Date Time..24hr. X1st.Road.Class Road.Surface
## 24                    2    2017-01-31       18:40    Unclassified     Wet/Damp
## 146                   1    2017-06-24       12:00               A          Dry
## 458                   1    2016-03-19       19:02               6          Dry
## 617                   1    2016-06-16       12:15               6          Dry
## 635                   1    2016-06-28       16:27               3          Dry
## 1462                  1    2014-02-05       17:15               6   Wet / Damp
## 1722                  2    2014-06-08       16:50               3          Dry
##                           Lighting.Conditions         Weather.Conditions
## 24   Darkness: street lights present but unlit Raining without high winds
## 146            Daylight: street lights present    Fine without high winds
## 458            Daylight: street lights present    Fine without high winds
```

```
## 617            Daylight: street lights present    Fine without high winds
## 635            Daylight: street lights present    Fine without high winds
## 1462   Darkness: street lights present and lit    Fine without high winds
## 1722           Daylight: street lights present    Fine without high winds
##      Type.of.Vehicle               Casualty.Class Casualty.Severity
## 24                 9              Driver or rider           Serious
## 146                9                   Pedestrian           Serious
## 458                9                   Pedestrian           Serious
## 617                9                   Pedestrian            Slight
## 635                9                   Pedestrian           Serious
## 1462               8                   Pedestrian           Serious
## 1722               9 Vehicle or pillion passenger            Slight
##      Sex.of.Casualty Age.of.Casualty
## 24            Female             115
## 146             Male              93
## 458             Male              93
## 617             Male             100
## 635           Female              92
## 1462            Male              95
## 1722          Female              98
```

- Hampel Identifier

```r
# Extract Age of Casualty data
age_casualty <- filtered_data$Age.of.Casualty

# Calculate the median and MAD
median_age <- median(age_casualty, na.rm = TRUE)
mad_age <- mad(age_casualty, na.rm = TRUE)

# Calculate Hampel's upper and lower bounds
hampel_upper_bound <- median_age + 3 * mad_age
hampel_lower_bound <- median_age - 3 * mad_age

# Identify outliers using Hampel identifier
outliers_hampel <- age_casualty[age_casualty > hampel_upper_bound | age_casualty < hampel_lower_bound]

# Print the outliers
print(outliers_hampel)
```

```
## [1] 115  93  93 100  92  95  98
```

```r
# Create a dataset for Hampel outliers
outliers_hampel_dataset <- filtered_data %>%
  filter(Age.of.Casualty > hampel_upper_bound | Age.of.Casualty < hampel_lower_bound)

# Print the number of Hampel outliers
number_of_outliers_hampel <- nrow(outliers_hampel_dataset)
print(paste("Number of Hampel outliers:", number_of_outliers_hampel))
```

```
## [1] "Number of Hampel outliers: 7"
```

```r
# Print the Hampel outliers dataset
print(outliers_hampel_dataset)
```

```
##   Number.of.Vehicles Accident.Date Time..24hr. X1st.Road.Class Road.Surface
## 1                  2    2017-01-31       18:40    Unclassified     Wet/Damp
## 2                  1    2017-06-24       12:00               A          Dry
## 3                  1    2016-03-19       19:02               6          Dry
## 4                  1    2016-06-16       12:15               6          Dry
## 5                  1    2016-06-28       16:27               3          Dry
## 6                  1    2014-02-05       17:15               6   Wet / Damp
## 7                  2    2014-06-08       16:50               3          Dry
##                        Lighting.Conditions         Weather.Conditions
## 1 Darkness: street lights present but unlit Raining without high winds
## 2           Daylight: street lights present    Fine without high winds
## 3           Daylight: street lights present    Fine without high winds
## 4           Daylight: street lights present    Fine without high winds
## 5           Daylight: street lights present    Fine without high winds
## 6   Darkness: street lights present and lit    Fine without high winds
## 7           Daylight: street lights present    Fine without high winds
##   Type.of.Vehicle               Casualty.Class Casualty.Severity
## 1               9              Driver or rider           Serious
## 2               9                   Pedestrian           Serious
## 3               9                   Pedestrian           Serious
## 4               9                   Pedestrian            Slight
## 5               9                   Pedestrian           Serious
## 6               8                   Pedestrian           Serious
## 7               9 Vehicle or pillion passenger            Slight
##   Sex.of.Casualty Age.of.Casualty
## 1          Female             115
## 2            Male              93
## 3            Male              93
## 4            Male             100
## 5          Female              92
## 6            Male              95
## 7          Female              98
```

```r
write.csv(filtered_data, file = "clean_accident.csv", row.names = FALSE)
```

Data Exploration

- Load necessary packages

```r
# Note: Please run data Wrangling first
# Unknown error preventing this rmd file to be converted into html or pdf
library(ggplot2)
library(magrittr)
library(readr)
```

- Load the cleaned csv

```
data <- read_csv("clean_accident.csv")
```

```
## Rows: 2050 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): X1st.Road.Class, Road.Surface, Lighting.Conditions, Weather.Condit...
## dbl  (3): Number.of.Vehicles, Type.of.Vehicle, Age.of.Casualty
## date (1): Accident.Date
## time (1): Time..24hr.
##
## 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.
```

- Check the first 6 rows

```
head(data)
```

```
## # A tibble: 6 x 12
##   Number.of.Vehicles Accident.Date Time..24hr. X1st.Road.Class Road.Surface
##                <dbl> <date>        <time>      <chr>           <chr>
## 1                  2 2017-01-01    21:20       Unclassified    Wet/Damp
## 2                  2 2017-01-04    15:00       Unclassified    Dry
## 3                  2 2017-01-05    07:32       A               Wet/Damp
## 4                  2 2017-01-05    09:30       A               Wet/Damp
## 5                  2 2017-01-14    09:09       Unclassified    Frost / Ice
## 6                  1 2017-01-16    10:59       A               Wet/Damp
## # i 7 more variables: Lighting.Conditions <chr>, Weather.Conditions <chr>,
## #   Type.of.Vehicle <dbl>, Casualty.Class <chr>, Casualty.Severity <chr>,
## #   Sex.of.Casualty <chr>, Age.of.Casualty <dbl>
```

- Check the structure of the data

```
str(data)
```

```
## spc_tbl_ [2,050 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Number.of.Vehicles : num [1:2050] 2 2 2 2 2 1 3 1 1 2 ...
##  $ Accident.Date      : Date[1:2050], format: "2017-01-01" "2017-01-04" ...
##  $ Time..24hr.        : 'hms' num [1:2050] 21:20:00 15:00:00 07:32:00 09:30:00 ...
##   ..- attr(*, "units")= chr "secs"
##  $ X1st.Road.Class    : chr [1:2050] "Unclassified" "Unclassified" "A" "A" ...
##  $ Road.Surface       : chr [1:2050] "Wet/Damp" "Dry" "Wet/Damp" "Wet/Damp" ...
##  $ Lighting.Conditions: chr [1:2050] "Darkness: street lights present and lit" "Daylight: street ligh
##  $ Weather.Conditions : chr [1:2050] "Raining without high winds" "Fine without high winds" "Fine wi
##  $ Type.of.Vehicle    : num [1:2050] 9 9 9 4 9 9 9 9 9 9 ...
##  $ Casualty.Class     : chr [1:2050] "Vehicle or pillion passenger" "Pedestrian" "Driver or rider" "
##  $ Casualty.Severity  : chr [1:2050] "Serious" "Serious" "Slight" "Fatal" ...
##  $ Sex.of.Casualty    : chr [1:2050] "Female" "Female" "Male" "Male" ...
##  $ Age.of.Casualty    : num [1:2050] 16 67 56 20 46 25 50 64 22 21 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Number.of.Vehicles = col_double(),
```

```
##   ..   Accident.Date = col_date(format = ""),
##   ..   Time..24hr. = col_time(format = ""),
##   ..   X1st.Road.Class = col_character(),
##   ..   Road.Surface = col_character(),
##   ..   Lighting.Conditions = col_character(),
##   ..   Weather.Conditions = col_character(),
##   ..   Type.of.Vehicle = col_double(),
##   ..   Casualty.Class = col_character(),
##   ..   Casualty.Severity = col_character(),
##   ..   Sex.of.Casualty = col_character(),
##   ..   Age.of.Casualty = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
male_accidents <- accidents %>%
  filter(Casualty.Class == "Driver or rider", Sex.of.Casualty == "Male")

female_accidents <- accidents %>%
  filter(Casualty.Class == "Driver or rider", Sex.of.Casualty == "Female")

count(male_accidents)
```

```
##     n
## 1 829
```

```
count(female_accidents)
```

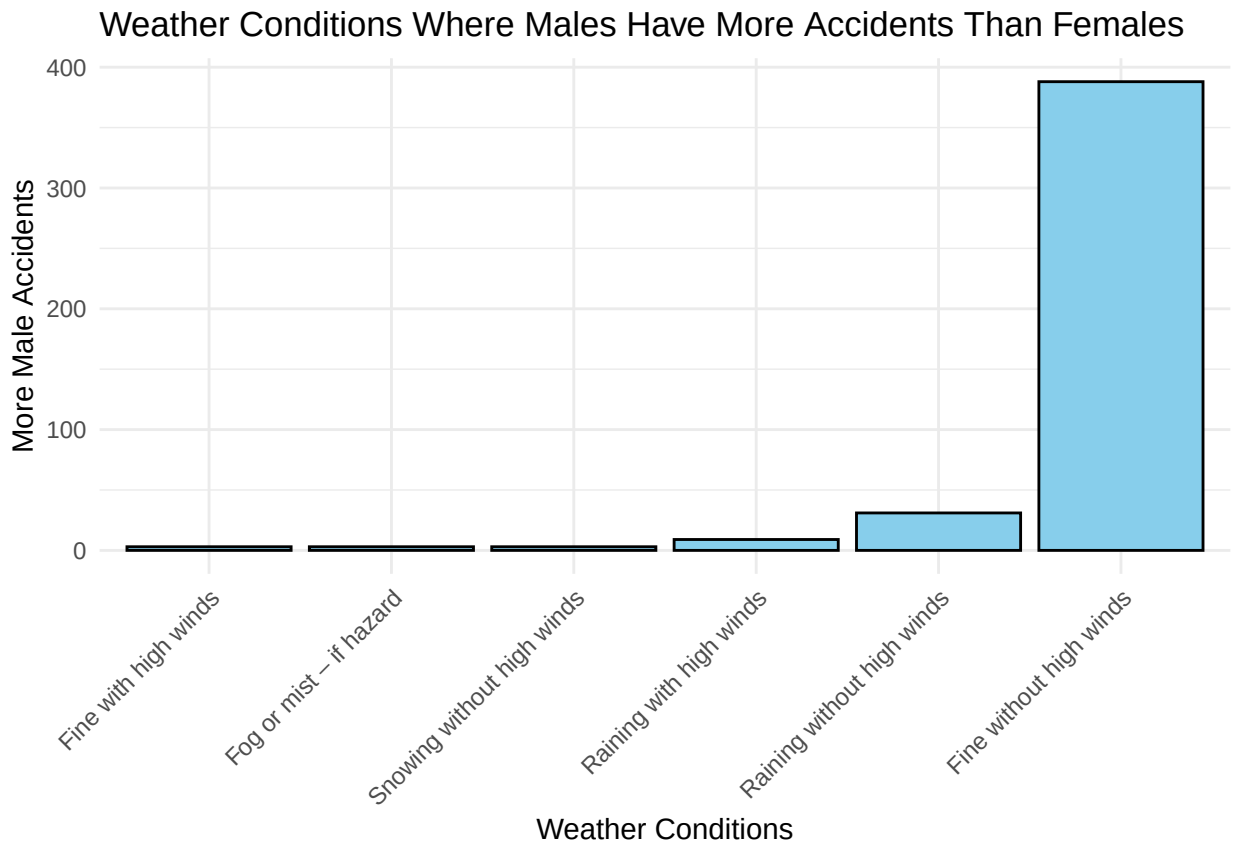
```
##     n
## 1 396
```

- Group the data by weather

```
male_counts <- male_accidents %>%
  group_by(Weather.Conditions) %>%
  summarise(total_male_accidents = n())

female_counts <- female_accidents %>%
  group_by(Weather.Conditions) %>%
  summarise(total_female_accidents = n())
```

```
accident_comparison <- merge(male_counts, female_counts, by = "Weather.Conditions", all = TRUE)

# Calculate the difference where males have more accidents than females
accident_comparison <- accident_comparison %>%
  mutate(more_male_accidents = ifelse(total_male_accidents > total_female_accidents,
                                      total_male_accidents - total_female_accidents,
                                      0))

# Filter for cases where males have more accidents than females
more_accidents <- accident_comparison %>%
  filter(more_male_accidents > 0) %>%
  arrange(desc(more_male_accidents))  # Sort by descending difference
```

```r
# Create a bar plot using ggplot2
ggplot(more_accidents, aes(x = reorder(Weather.Conditions, more_male_accidents), y = more_male_accidents
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  labs(x = "Weather Conditions", y = "More Male Accidents") +
  ggtitle("Weather Conditions Where Males Have More Accidents Than Females") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

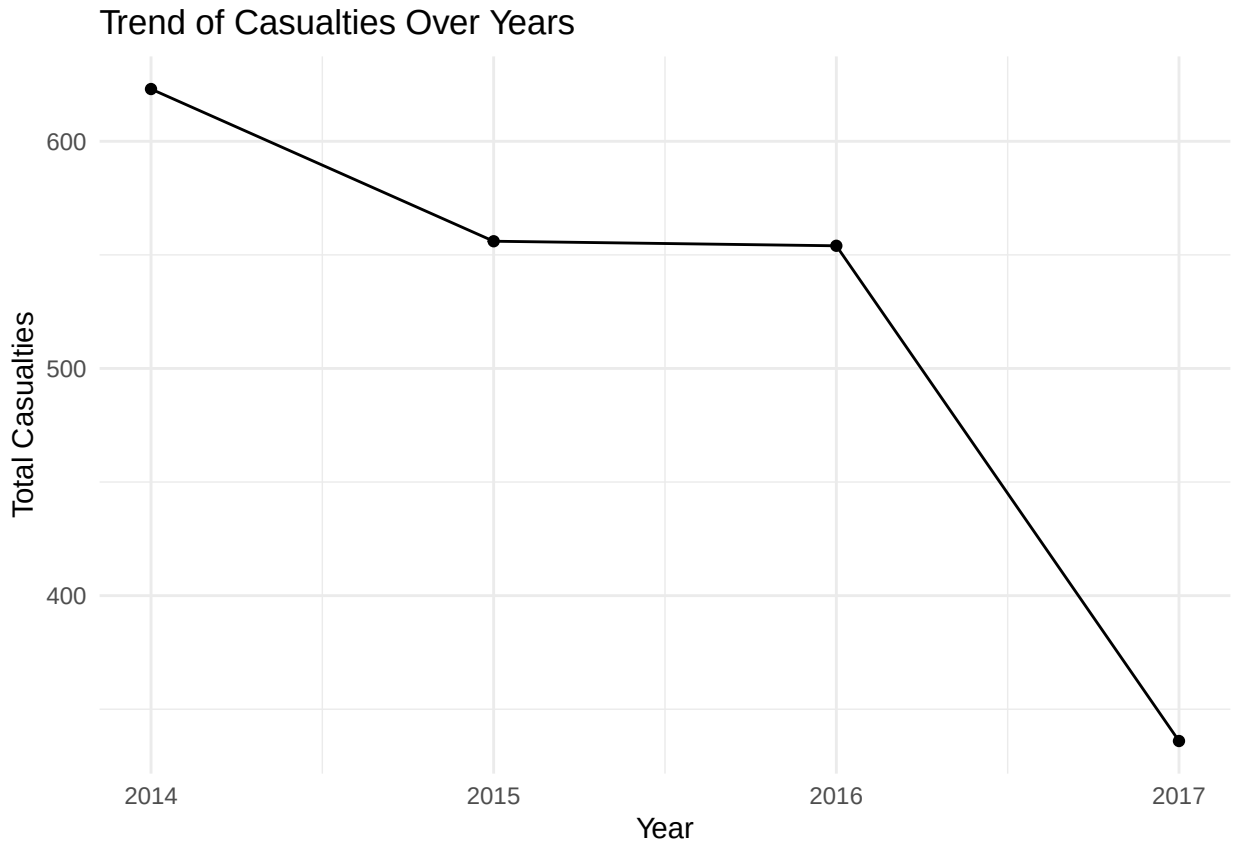



```r
number_of_cases <- nrow(more_accidents)
number_of_cases
```

```
## [1] 6
```

```r
accidents <- accidents %>%
  mutate(Year = lubridate::year(Accident.Date))

# Group by year and count casualties
casualties_by_year <- accidents %>%
  group_by(Year) %>%
  summarise(total_casualties = n())

# Find the year with the highest number of casualties
max_casualty_year <- casualties_by_year %>%
  filter(total_casualties == max(total_casualties)) %>%
```

```
  pull(Year)

# Plotting the trend of casualties over the years
ggplot(casualties_by_year, aes(x = Year, y = total_casualties)) +
  geom_line() +
  geom_point() +
  labs(title = "Trend of Casualties Over Years",
       x = "Year",
       y = "Total Casualties") +
  theme_minimal()
```



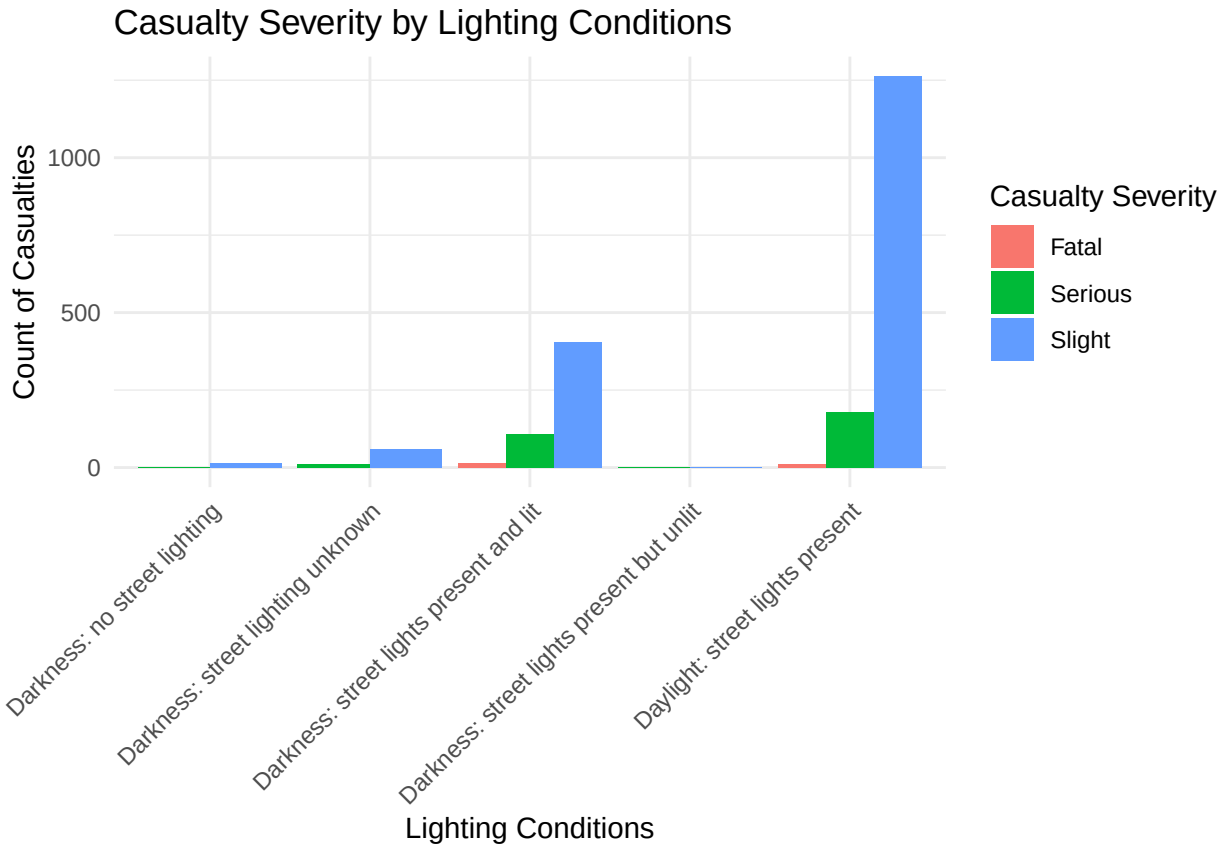

```
# Print the year with the highest number of casualties
print(paste("Year with the highest number of casualties:", max_casualty_year))
```

```
## [1] "Year with the highest number of casualties: 2014"
```

Light condition and Severity

```
# Load necessary libraries
library(ggplot2)
library(dplyr)

# Assuming 'accidents' dataset is already loaded and cleaned
```

```r
# Create the bar plot for Light Conditions and Severity
light_severity_plot <- ggplot(accidents, aes(x = Lighting.Conditions, fill = Casualty.Severity)) +
  geom_bar(position = "dodge") +  # Create a bar plot with bars side by side for each severity level
  labs(title = "Casualty Severity by Lighting Conditions",
       x = "Lighting Conditions",
       y = "Count of Casualties",
       fill = "Casualty Severity") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabili

# Print the plot
print(light_severity_plot)
```




```r
# Check the count of casualties by lighting conditions and severity
table(accidents$Lighting.Conditions, accidents$Casualty.Severity)
```

```
##
##                                           Fatal Serious Slight
##   Darkness: no street lighting                0       3     15
##   Darkness: street lighting unknown           0      10     59
##   Darkness: street lights present and lit    15     107    405
##   Darkness: street lights present but unlit   0       1      3
##   Daylight: street lights present            10     178   1263
```

Interpretation Darkness with no streetlight:

There is no fatal injury and only few seriously injured with a slightly higher number minorly injured

Darkness: street lighting unknown

Although no fatal injuries, the number of seriously and slightly injured has tripled compared to darkness with no streetlight.

Darkness: street lights present and lit

The lit streetlight at darkness has the highest number of fatal injuries (15), 107 seriously injured and 405 slightly injured. The reason might be because of low volume of traffic at night and well lit streets.

Darkness: street lights present but unlit

The unlit streetlight in darkness has the lowest number of injuries overall(0 fatal, 1 serious, 3 slight). The reason might be because of although low traffic, but low visibility as well.

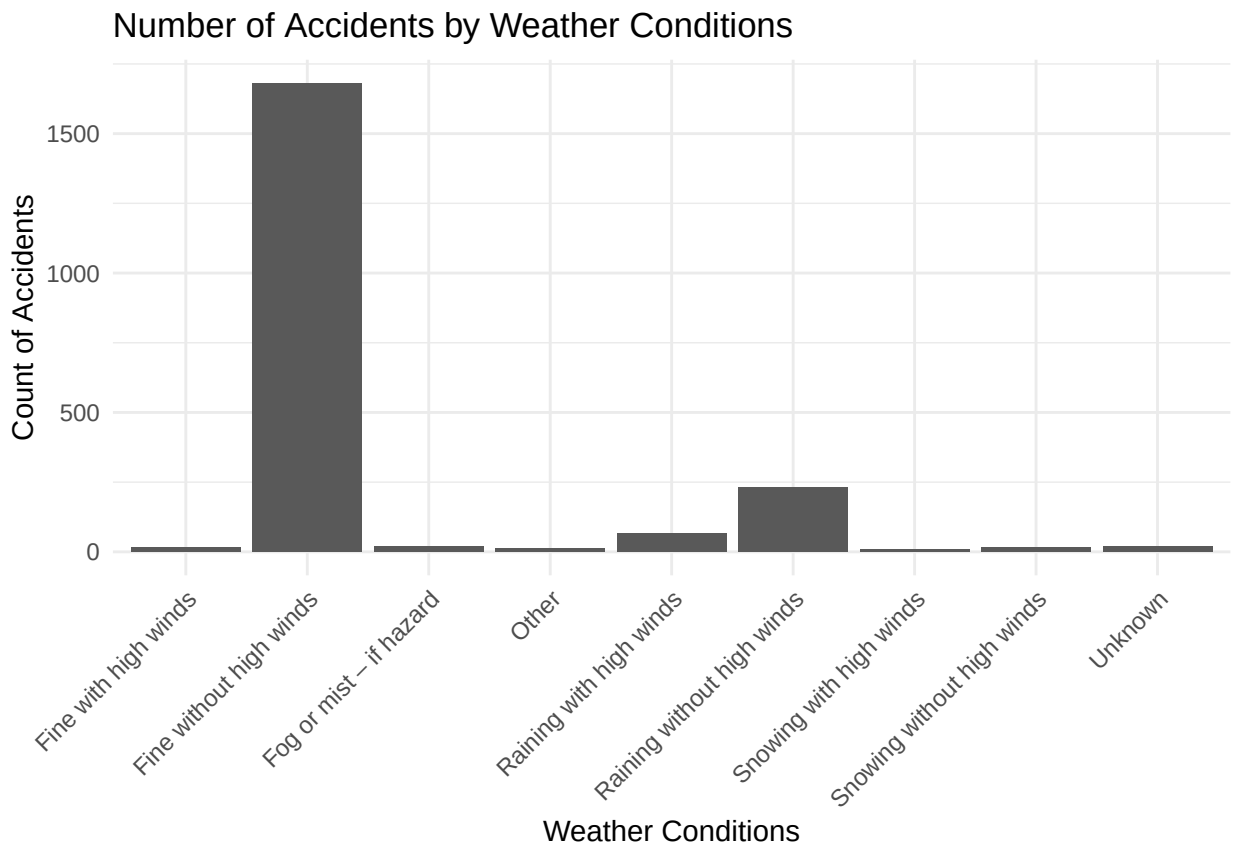
Daylight: street lights present

Shows the highest number of accidents in every category except fatality(10 fatal, 178 serious, 1263). The plausible reason might be high volume of traffic and better visibility of accidents happening on the broad daylight.

Other 2 conditions (Daylight: no street lights present and Daylight: streetlight unknown) has 0 number of casualties as the obvious reason being no streetlight being lit during the daytime.

Weather and number of vehicles involved

```
weather_vehicles_count_data <- accidents %>%
  group_by(Weather.Conditions) %>%
  summarise(Count_of_Accidents = n())

weather_vehicles_plot <- ggplot(weather_vehicles_count_data, aes(x = Weather.Conditions, y = Count_of_A
  geom_bar(stat = "identity") +
  labs(title = "Number of Accidents by Weather Conditions",
       x = "Weather Conditions",
       y = "Count of Accidents") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

weather_vehicles_plot
```



Interpretation

Fine without high winds:

The weather condition has the highest number of casualties reaching almost 2000 and is exponentially higher compared to the next in line (Raining without high winds). The likely reason might be high volume of traffic

Raining without high winds:

Although not as severe as fine without high winds, the number of vehicles involved in accident is noticeable (250). The most probable reason cold be more cautious driving during rain as the tires may slip.

Raining with high winds:

The third condition with the highest number of vehicles involved in accident but incomparable to aforementioned ones (100). The probable reason could be even more cautios driving than the rain without winds as there are 2 factors to look at.

Others (everything except aforementioned):

All the other weather conditions have less than 50 vehicles involved in an accident.

```
table(accidents$Weather.Conditions, accidents$Number.of.Vehicles)
```

```
##
##                             1   2   3   4   5   6   7
##   Fine with high winds       7  10   0   0   0   0   0
##   Fine without high winds  494 950 170  44   8   5  10
##   Fog or mist - if hazard    6  10   2   0   0   0   0
##   Other                      4   6   2   0   0   0   0
```

```
##   Raining with high winds     19  35   7   4   0   0   0
##   Raining without high winds  71 123  37   2   0   0   0
##   Snowing with high winds      2   4   2   0   0   0   0
##   Snowing without high winds   6   6   3   1   0   0   0
##   Unknown                      6  12   1   0   0   0   0
```

Linear Regression

- Load necessary libraries

```
library(dplyr)
library(magrittr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats 1.0.0     v tibble  3.2.1
## v purrr   1.0.2     v tidyr   1.3.1
## v stringr 1.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

- Import clean accident csv

```
data <- read.csv('data/accidents.csv')
head(data)
```

```
##   Number.of.Vehicles Accident.Date Time..24hr. X1st.Road.Class Road.Surface
## 1                  2      1/1/2017        2120               U     Wet/Damp
## 2                  2      4/1/2017        1500               U          Dry
## 3                  2      5/1/2017         732            A58     Wet/Damp
## 4                  2      5/1/2017         930           A646     Wet/Damp
## 5                  2    14/01/2017         909              U    Frost/Ice
## 6                  1    15/01/2017        1659              U     Wet/Damp
##   Lighting.Conditions Daylight.Dark Weather.Conditions Local.Authority
## 1                   4          Dark                  2      Calderdale
## 2                   1      Daylight                  1      Calderdale
## 3                   4          Dark                  1      Calderdale
## 4                   1      Daylight                  1      Calderdale
## 5                   1      Daylight                  1      Calderdale
## 6                   4          Dark                  1      Calderdale
##   Type.of.Vehicle Casualty.Class Casualty.Severity Sex.of.Casualty
## 1               9              2                 2               2
## 2               9              3                 2               2
## 3               9              1                 3               1
## 4               4              1                 1               1
## 5               9              1                 3               1
## 6               9              3                 3               1
```

```
##   Age.of.Casualty
## 1              16
## 2              67
## 3              56
## 4              20
## 5              46
## 6              NA
```

- Check the structure of data

```
str(data)
```

```
## 'data.frame':    2069 obs. of  14 variables:
##  $ Number.of.Vehicles : int  2 2 2 2 2 1 1 3 1 1 ...
##  $ Accident.Date      : chr  "1/1/2017" "4/1/2017" "5/1/2017" "5/1/2017" ...
##  $ Time..24hr.        : int  2120 1500 732 930 909 1659 1059 1849 1408 1325 ...
##  $ X1st.Road.Class    : chr  "U" "U" "A58" "A646" ...
##  $ Road.Surface       : chr  "Wet/Damp" "Dry" "Wet/Damp" "Wet/Damp" ...
##  $ Lighting.Conditions: int  4 1 4 1 1 4 1 4 1 1 ...
##  $ Daylight.Dark      : chr  "Dark" "Daylight" "Dark" "Daylight" ...
##  $ Weather.Conditions : int  2 1 1 1 1 1 1 1 1 1 ...
##  $ Local.Authority    : chr  "Calderdale" "Calderdale" "Calderdale" "Calderdale" ...
##  $ Type.of.Vehicle    : int  9 9 9 4 9 9 9 9 9 9 ...
##  $ Casualty.Class     : int  2 3 1 1 1 3 3 1 3 1 ...
##  $ Casualty.Severity  : int  2 2 3 1 3 3 3 3 3 3 ...
##  $ Sex.of.Casualty    : int  2 2 1 1 1 1 2 2 2 2 ...
##  $ Age.of.Casualty    : int  16 67 56 20 46 NA 25 50 64 22 ...
```

- Convert columns into something usable

```
data$Casualty.Class <- as.factor(data$Casualty.Class)
data$Casualty.Severity <- as.factor(data$Casualty.Severity)
data$Type.of.Vehicle <- as.factor(data$Type.of.Vehicle)
data$Weather.Conditions <- as.factor(data$Weather.Conditions)
str(data)
```

```
## 'data.frame':    2069 obs. of  14 variables:
##  $ Number.of.Vehicles : int  2 2 2 2 2 1 1 3 1 1 ...
##  $ Accident.Date      : chr  "1/1/2017" "4/1/2017" "5/1/2017" "5/1/2017" ...
##  $ Time..24hr.        : int  2120 1500 732 930 909 1659 1059 1849 1408 1325 ...
##  $ X1st.Road.Class    : chr  "U" "U" "A58" "A646" ...
##  $ Road.Surface       : chr  "Wet/Damp" "Dry" "Wet/Damp" "Wet/Damp" ...
##  $ Lighting.Conditions: int  4 1 4 1 1 4 1 4 1 1 ...
##  $ Daylight.Dark      : chr  "Dark" "Daylight" "Dark" "Daylight" ...
##  $ Weather.Conditions : Factor w/ 9 levels "1","2","3","4",..: 2 1 1 1 1 1 1 1 1 1 ...
##  $ Local.Authority    : chr  "Calderdale" "Calderdale" "Calderdale" "Calderdale" ...
##  $ Type.of.Vehicle    : Factor w/ 19 levels "1","2","3","4",..: 7 7 7 4 7 7 7 7 7 7 ...
##  $ Casualty.Class     : Factor w/ 3 levels "1","2","3": 2 3 1 1 1 3 3 1 3 1 ...
##  $ Casualty.Severity  : Factor w/ 3 levels "1","2","3": 2 2 3 1 3 3 3 3 3 3 ...
##  $ Sex.of.Casualty    : int  2 2 1 1 1 1 2 2 2 2 ...
##  $ Age.of.Casualty    : int  16 67 56 20 46 NA 25 50 64 22 ...
```

- Create a data to train

```r
train.data <- data
head(train.data)
```

```
##   Number.of.Vehicles Accident.Date Time..24hr. X1st.Road.Class Road.Surface
## 1                  2      1/1/2017        2120               U     Wet/Damp
## 2                  2      4/1/2017        1500               U          Dry
## 3                  2      5/1/2017         732            A58     Wet/Damp
## 4                  2      5/1/2017         930           A646     Wet/Damp
## 5                  2    14/01/2017         909               U    Frost/Ice
## 6                  1    15/01/2017        1659               U     Wet/Damp
##   Lighting.Conditions Daylight.Dark Weather.Conditions Local.Authority
## 1                   4          Dark                  2      Calderdale
## 2                   1      Daylight                  1      Calderdale
## 3                   4          Dark                  1      Calderdale
## 4                   1      Daylight                  1      Calderdale
## 5                   1      Daylight                  1      Calderdale
## 6                   4          Dark                  1      Calderdale
##   Type.of.Vehicle Casualty.Class Casualty.Severity Sex.of.Casualty
## 1               9              2                 2               2
## 2               9              3                 2               2
## 3               9              1                 3               1
## 4               4              1                 1               1
## 5               9              1                 3               1
## 6               9              3                 3               1
##   Age.of.Casualty
## 1              16
## 2              67
## 3              56
## 4              20
## 5              46
## 6              NA
```

- Create a linear model

```r
lm.model <- lm(Age.of.Casualty ~ Casualty.Class + Casualty.Severity
               + Type.of.Vehicle + Weather.Conditions, data = train.data)

summary(lm.model)
```

```
##
## Call:
## lm(formula = Age.of.Casualty ~ Casualty.Class + Casualty.Severity +
##     Type.of.Vehicle + Weather.Conditions, data = train.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.011 -13.836  -3.836  11.372  74.470
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          39.24578    4.09485   9.584  < 2e-16 ***
## Casualty.Class2     -10.33801    1.08903  -9.493  < 2e-16 ***
```

```
## Casualty.Class3      -5.74153    1.17441  -4.889 1.09e-06 ***
## Casualty.Severity2   -0.79086    3.95306  -0.200 0.841453
## Casualty.Severity3   -3.62004    3.83488  -0.944 0.345294
## Type.of.Vehicle2    -11.23831    4.42569  -2.539 0.011181 *
## Type.of.Vehicle3     -7.88001    2.83980  -2.775 0.005574 **
## Type.of.Vehicle4      1.87081    4.44414   0.421 0.673827
## Type.of.Vehicle5      6.58660    2.95297   2.231 0.025824 *
## Type.of.Vehicle8      6.55257    2.74807   2.384 0.017198 *
## Type.of.Vehicle9      4.20993    1.72319   2.443 0.014647 *
## Type.of.Vehicle10     9.20195   11.06713   0.831 0.405808
## Type.of.Vehicle11    16.29740    3.60651   4.519 6.58e-06 ***
## Type.of.Vehicle16    -4.62574   13.46905  -0.343 0.731307
## Type.of.Vehicle17   -11.62574   13.46905  -0.863 0.388160
## Type.of.Vehicle18    -6.73373   11.06074  -0.609 0.542728
## Type.of.Vehicle19     4.78531    2.89890   1.651 0.098949 .
## Type.of.Vehicle20    21.21183    6.90388   3.072 0.002151 **
## Type.of.Vehicle21    15.44571    4.45327   3.468 0.000535 ***
## Type.of.Vehicle22    14.37426   18.97931   0.757 0.448920
## Type.of.Vehicle23   -19.62574   18.97931  -1.034 0.301232
## Type.of.Vehicle90    -0.09675    6.53025  -0.015 0.988181
## Type.of.Vehicle97    45.11580   19.01769   2.372 0.017771 *
## Weather.Conditions2  -2.13510    1.33965  -1.594 0.111143
## Weather.Conditions3   2.18770    4.76285   0.459 0.646051
## Weather.Conditions4   3.89607    4.62454   0.842 0.399620
## Weather.Conditions5  -3.39275    2.42242  -1.401 0.161499
## Weather.Conditions6   1.21828    6.71970   0.181 0.856150
## Weather.Conditions7   2.08722    4.49335   0.465 0.642331
## Weather.Conditions8 -11.20213    5.48830  -2.041 0.041371 *
## Weather.Conditions9  -3.40685    4.37942  -0.778 0.436706
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.91 on 2019 degrees of freedom
##   (19 observations deleted due to missingness)
## Multiple R-squared:  0.0784, Adjusted R-squared:  0.06471
## F-statistic: 5.725 on 30 and 2019 DF,  p-value: < 2.2e-16
```

- Filter data with missing values

```
missing.data <- data %>%
  filter(is.na(Age.of.Casualty))
```

- Predict the missing values

```
predicted.age <- predict(lm.model, newdata = missing.data)

data$Age.of.Casualty[is.na(data$Age.of.Casualty)] <- predicted.age

missing.data <- data %>%
  filter(is.na(Age.of.Casualty))
missing.data
```

```
##  [1] Number.of.Vehicles  Accident.Date       Time..24hr.
```

```
##  [4] X1st.Road.Class     Road.Surface        Lighting.Conditions
##  [7] Daylight.Dark       Weather.Conditions  Local.Authority
## [10] Type.of.Vehicle     Casualty.Class      Casualty.Severity
## [13] Sex.of.Casualty     Age.of.Casualty
## <0 rows> (or 0-length row.names)
```

- Round up the predicted value and convert it to integer

```
data$Age.of.Casualty <- round(data$Age.of.Casualty)

data$Age.of.Casualty <- as.integer(data$Age.of.Casualty)
```

- Check for any remaining missing values in the dataset

```
missing.values <- colSums(is.na(data))
print(missing.values[missing.values > 0])
```

```
## named numeric(0)
```

```
dim(data)
```

```
## [1] 2069   14
```

```
write.csv(data, "regression.csv", row.names = FALSE, quote= FALSE, fileEncoding = "UTF-8")
```

```
## Warning in utils::write.table(data, "regression.csv", row.names = FALSE, :
## invalid char string in output conversion
## Warning in utils::write.table(data, "regression.csv", row.names = FALSE, :
## invalid char string in output conversion
## Warning in utils::write.table(data, "regression.csv", row.names = FALSE, :
## invalid char string in output conversion
## Warning in utils::write.table(data, "regression.csv", row.names = FALSE, :
## invalid char string in output conversion
```