U.K.Accidents-Ten Years History.

Sheetal Munjewar

2023-03-03

Contents

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/

Problem statement / 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 ?

Insights from Data

- Reduction in the number of reported accidents between 2005 and 2014.
- Accidents reported on Friday certainly in lead compared to other week days.
- Accidents tend to occur on the business hours when people commute to work.
- Contingency table show, proportion of fatal accidents is higher than during the day, while we observe the opposite result for the slight accidents, results proved our conclusion using chi-square test.
- Casualty Outcome proportion conclude that probability of an accident to be fatal is higher when it's foggy or misty.
- More "slight" accidents happen in urban areas; however, "fatal" accidents ratio is more in a rural area compared to total accidents occurs in both areas.
- Probability of an accident to be fatal is higher on road that are "Not a junction or within 20 metres of a junction". On the contrary an accident happening on a roundabout is much more likely to be a slight accident and not likely at all to be a fatal accident.
- Road surface with "Oil or Diesel" can cause more accidents (Slight,Serious and Fatal)
- Death rate of drivers aged over 75 is much higher probably because they are more vulnerable to injuries, or they are driving old car.

Limitations

- Data set is based on Rural and Urban categorization, however actual location and popluation, country geo region specific data and transparency is limited.
- Data set findings are limited and sampling based, however actual vechicle and its categorization with model and approximate count across is missing.
- Road conditions and changes across the decade are limited.
- New traffic rule and regulation like change in speed limit on highway, expansion of roads and manu more changes over the decade are limited.

Improvements

Data set used for analysis does have missing information and it doesn't cover all the facts, further data correction can improve accuracy and evaluate more observation in future.

Concluding Remarks

The total number of vehicle accidents are the decline, likelihood of being involved in a collision is higher on Fridays while the lowest is Saturday and Sunday. A severe injury or terminal outcome are a lot more likely to occur if the vehicle comes to an abrupt stop rather than skidding. Young people are more likely to drive recklessly and be involved in a collision calls for more data,the full moon has no effect whatsoever on the number of vehicle collisions.

Test of Independence, shows data findings are 95% CI statistically correct as we always have a p-value < 0.05

Required Packages

Base packages plus
"ggplot2", "dplyr", "broom", "purrr", "GGally", "scales", "caret", "moments", "ggpubr", "readxl", "corrplot"

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)

```r
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.

```r
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")

Day_of_Week <- read_excel("assignments/Final-Project/Road-Accident-Safety-Data-Guide.xls",
    sheet = "Day of Week")
df <- left_join(df, Day_of_Week, by = c(Day_of_Week = "code"))
df <- dplyr::rename(df, day_of_Week = label)
rm(Day_of_Week)

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)

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

```r
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 Manoeuvre")
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 Overturning")
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 Purpose")
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 Severity")
df <- left_join(df, Casualty_severity, by = c(Casualty_Severity = "code"))
df <- dplyr::rename(df, Casualty_Outcome = label)
rm(Casualty_severity)

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

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

Get rid of excess data columns.

```r
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(warning = FALSE)
```

```r
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", "1st_Point_of_Impact", "Hit_Object_off_Carriageway",
    "Vehicle_Leaving_Carriageway", "Hit_Object_in_Carriageway", "Junction_Location",
    "Vehicle_Location-Restricted_Lane", "Towing_and_Articulation", "Pedestrian_Road_Maintenance_Worker",
    "Pedestrian_Movement", "Pedestrian_Location", "Casualty_Reference",
    "LSOA_of_Accident_Location", "Did_Police_Officer_Attend_Scene_of_Accident",
    "Carriageway_Hazards", "Pedestrian_Crossing-Physical_Facilities", "Pedestrian_Crossing-Human_Control",
    "2nd_Road_Class", "2nd_Road_Number", "Junction_Control", "Junction_Detail",
    "Local_Authority_(Highway)")]

# dim(df) head(df) unique(df$Skidding)
```

Change Date column to date format.

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

Adding new columns for aggregation and summerization.

```r
df$Year <- format(as.Date(df$Date), "%Y")
df$Month <- format(as.Date(df$Date), "%m")
df$time_slot <- as.numeric(substr(df$Time, 0, 2))
```

Check "NA" counts in dataset

```r
# Check 'NA' count in dataset.
sort(sapply(df, function(x) sum(is.na(x))), decreasing = TRUE)
```

```
##                                     Date                                    Year
##                                  2578146                                 2578146
##                                    Month                            Road_Surface
##                                  2578146                                   20880
##                                time_slot                   Location_Easting_OSGR
##                                      264                                     256
##                   Location_Northing_OSGR                               Longitude
##                                      256                                     256
##                                 Latitude                          Accident_Index
##                                      256                                       0
##                             Police_Force                       Accident_Severity
##                                        0                                       0
##                       Number_of_Vehicles                    Number_of_Casualties
##                                        0                                       0
##                              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
##                     Urban_or_Rural_Area                     Vehicle_Reference.x
##                                       0                                       0
##                          Casualty_Class                         Sex_of_Casualty
##                                       0                                       0
##                         Age_of_Casualty                    Age_Band_of_Casualty
##                                       0                                       0
##                       Casualty_Severity                           Car_Passenger
##                                       0                                       0
##                  Bus_or_Coach_Passenger                           Casualty_Type
##                                       0                                       0
##                 Casualty_Home_Area_Type                     Vehicle_Reference.y
##                                       0                                       0
##                            Vehicle_Type                       Vehicle_Manoeuvre
##                                       0                                       0
##        Vehicle_Location.Restricted_Lane                Skidding_and_Overturning
##                                       0                                       0
##                    X1st_Point_of_Impact            Was_Vehicle_Left_Hand_Drive.
##                                       0                                       0
##               Journey_Purpose_of_Driver                           Sex_of_Driver
##                                       0                                       0
##                           Age_of_Driver                    Engine_Capacity_.CC.
##                                       0                                       0
##                          Age_of_Vehicle                             day_of_Week
##                                       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                             Urban_Rural
##                                       0                                       0
```

Display final data set and save it in seperate file.

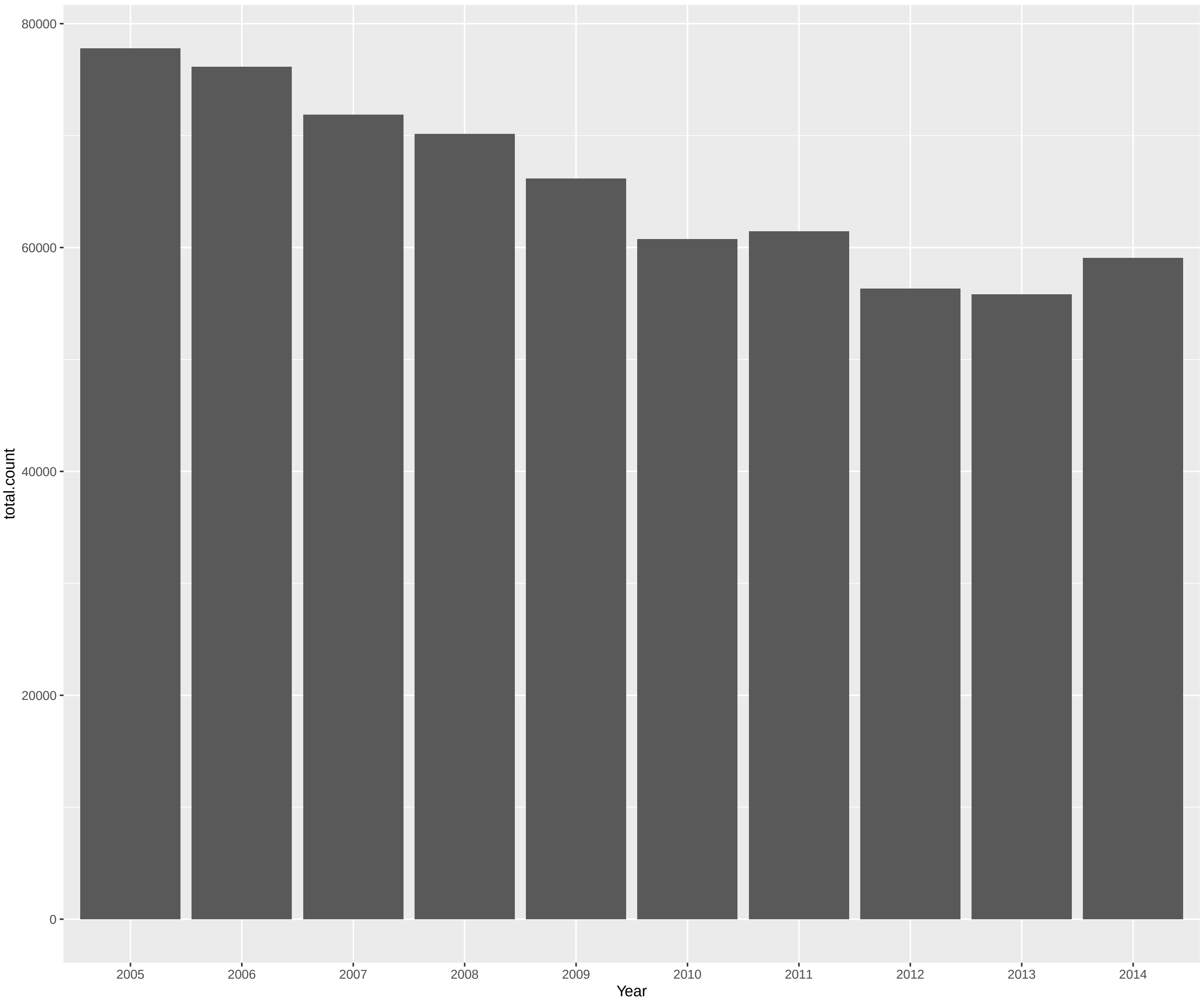
```
# write.csv(df, file = 'filtered_eported_data.csv')
head(df)
```

```
##   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 Urban_or_Rural_Area Vehicle_Reference.x
## 1                          0                   1                   1
## 2                          0                   1                   1
## 3                          0                   1                   2
## 4                          0                   1                   2
## 5                          0                   1                   1
## 6                          0                   1                   1
##   Casualty_Class Sex_of_Casualty Age_of_Casualty Age_Band_of_Casualty
## 1              3               1              37                    7
## 2              2               1              37                    7
## 3              1               1              62                    9
## 4              1               1              62                    9
## 5              3               1              30                    6
## 6              1               1              49                    8
##   Casualty_Severity Car_Passenger Bus_or_Coach_Passenger Casualty_Type
## 1                 2             0                      0             0
## 2                 3             0                      4            11
## 3                 3             0                      0             9
## 4                 3             0                      0             9
## 5                 3             0                      0             0
## 6                 3             0                      0             3
##   Casualty_Home_Area_Type Vehicle_Reference.y Vehicle_Type Vehicle_Manoeuvre
## 1                       1                   1            9                18
## 2                       1                   1           11                 4
## 3                       1                   1           11                17
## 4                       1                   2            9                 2
## 5                       1                   1            9                18
## 6                      -1                   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 day_of_Week
## 1             2            74                   -1             -1     Tuesday
## 2             1            42                 8268              3   Wednesday
## 3             1            35                 8300              5    Thursday
## 4             1            62                 1762              6    Thursday
## 5             2            49                 1769              4      Friday
## 6             1            49                   85             10      Monday
##              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  Road_Surface Urban_Rural
## 1 Other/Not known (2005-10)  66 - 75          Serious Oil or diesel       Urban
## 2   Journey as part of work  36 - 45           Slight          Snow       Urban
## 3   Journey as part of work  26 - 35           Slight Oil or diesel       Urban
## 4 Other/Not known (2005-10)  56 - 65           Slight Oil or diesel       Urban
## 5 Other/Not known (2005-10)  46 - 55           Slight Oil or diesel       Urban
## 6 Other/Not known (2005-10)  46 - 55           Slight Oil or diesel       Urban
##   Year Month time_slot
## 1 2005    04        17
## 2 2005    05        17
## 3 2005    06         0
## 4 2005    06         0
## 5 2005    07        10
## 6 2005    10        21
```

Data Analysis and Visualization Section :

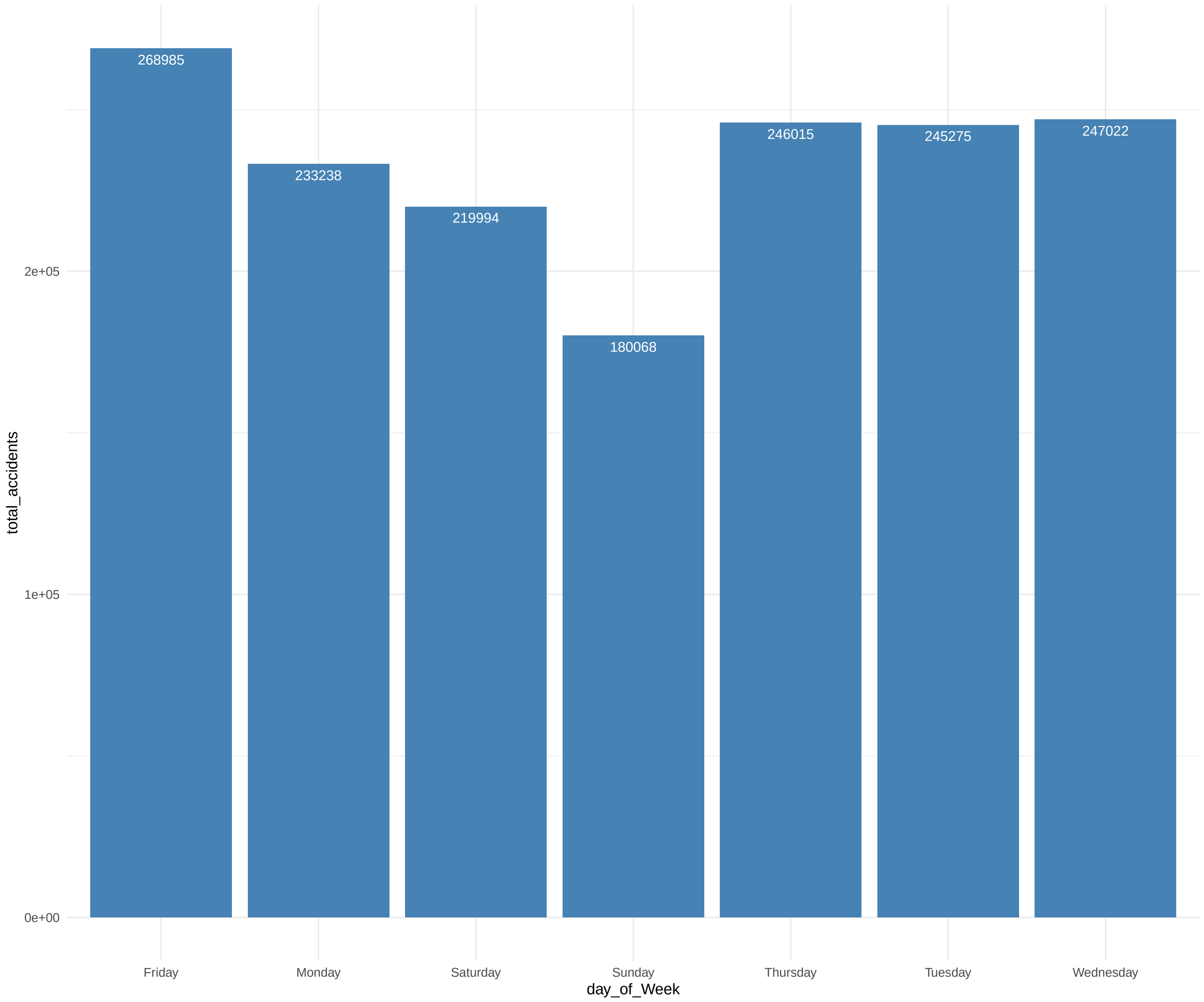
Graph indicate reduction in the number of reported accidents between 2005 and 2014.

```
df1 <- df %>%
    select(Accident_Index, Location, Accident_Severity, Number_of_Vehicles,
        Number_of_Casualties, Date, Day_of_Week, Month, Year, Time, Road_Type,
        Lighting, Weather, Surface, Skidding, ) %>%
    group_by(Accident_Index) %>%
    filter(row_number() == 1)

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

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



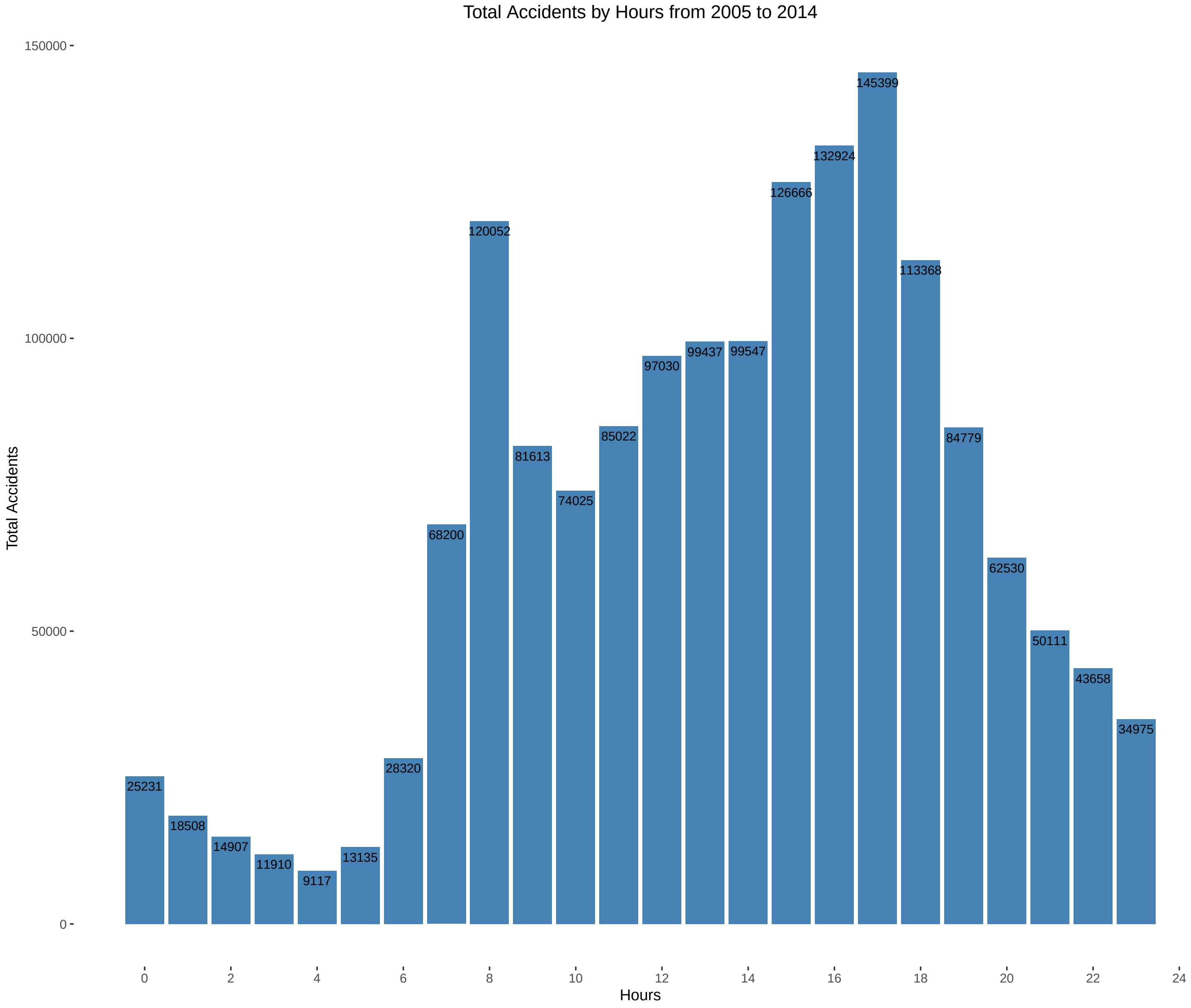
Accident count on specific day of the week.

```
df %>%
    group_by(day_of_Week) %>%
    dplyr::summarize(total_accidents = n_distinct(Accident_Index)) %>%
    ggplot(aes(x = day_of_Week, y = total_accidents)) + geom_bar(stat = "identity",
    fill = "steelblue") + geom_text(aes(label = total_accidents), vjust = 1.6,
    color = "white", size = 3.5) + theme_minimal()
```



Accident by hours .

```
df %>%
    group_by(time_slot) %>%
    dplyr::summarize(total_accidents = n_distinct(Accident_Index)) %>%
    ggplot(aes(x = time_slot, y = total_accidents)) + geom_bar(stat = "identity",
    fill = "steelblue") + geom_text(aes(label = total_accidents), vjust = 1.6,
    color = "black", size = 3) + scale_x_continuous(breaks = round(seq(0,
    24, by = 2), 0)) + ggtitle("Total Accidents by Hours from 2005 to 2014") +
    xlab("Hours") + ylab("Total Accidents") + theme(plot.title = element_text(hjust = 0.5),
    panel.background = element_blank())
```

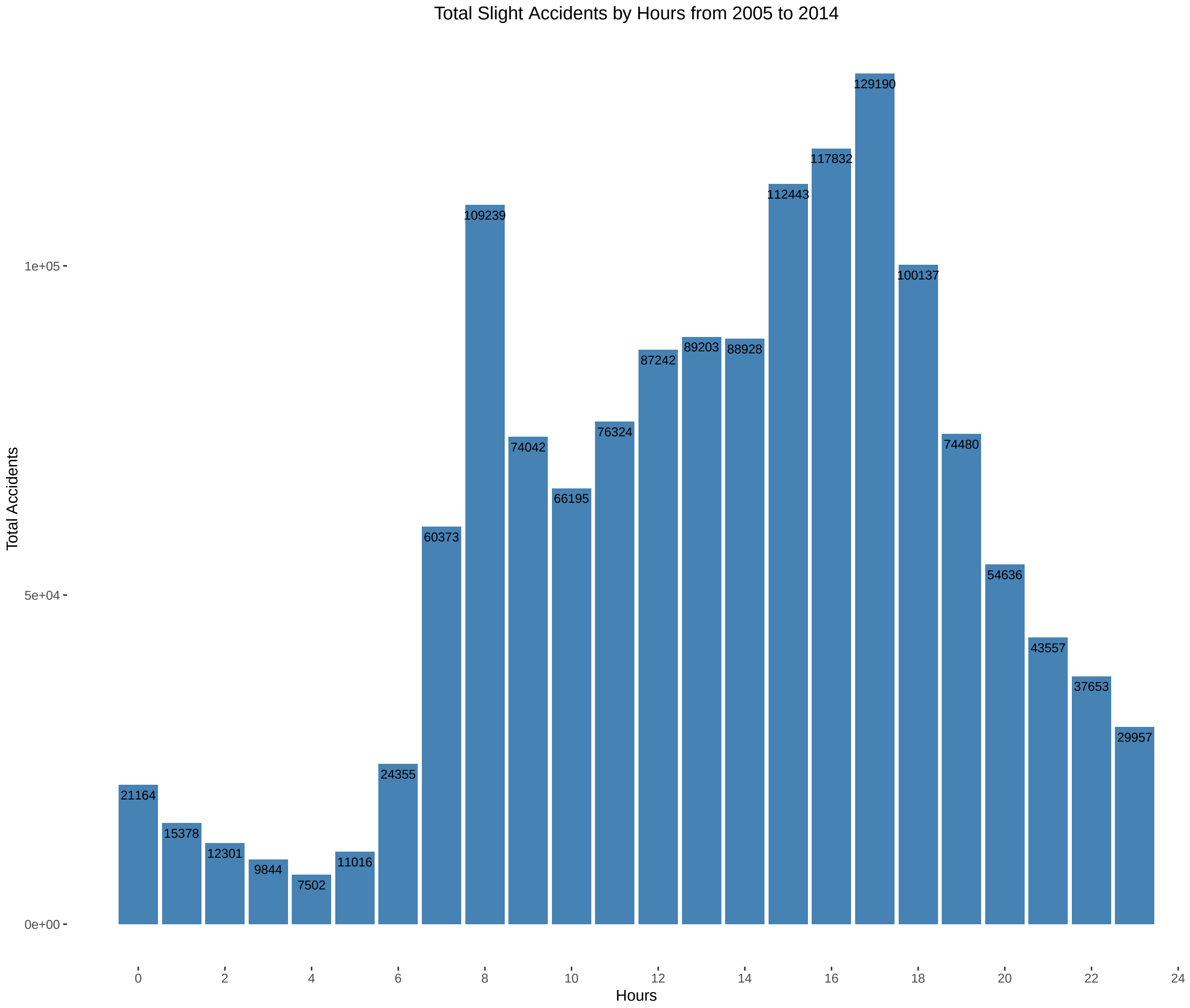


Slight Accident by hours

```
head(df$Casualty_Severity)
```

```
## [1] 2 3 3 3 3 3
```

```
df %>%
    filter(Casualty_Outcome == "Slight") %>%
    group_by(time_slot) %>%
    dplyr::summarize(total_accidents = n_distinct(Accident_Index)) %>%
    ggplot(aes(x = time_slot, y = total_accidents)) + geom_bar(stat = "identity",
    fill = "steelblue") + geom_text(aes(label = total_accidents), vjust = 1.6,
    color = "black", size = 3) + scale_x_continuous(breaks = round(seq(0,
    24, by = 2), 0)) + ggtitle("Total Slight Accidents by Hours from 2005 to 2014") +
    xlab("Hours") + ylab("Total Accidents") + theme(plot.title = element_text(hjust = 0.5),
    panel.background = element_blank())
```

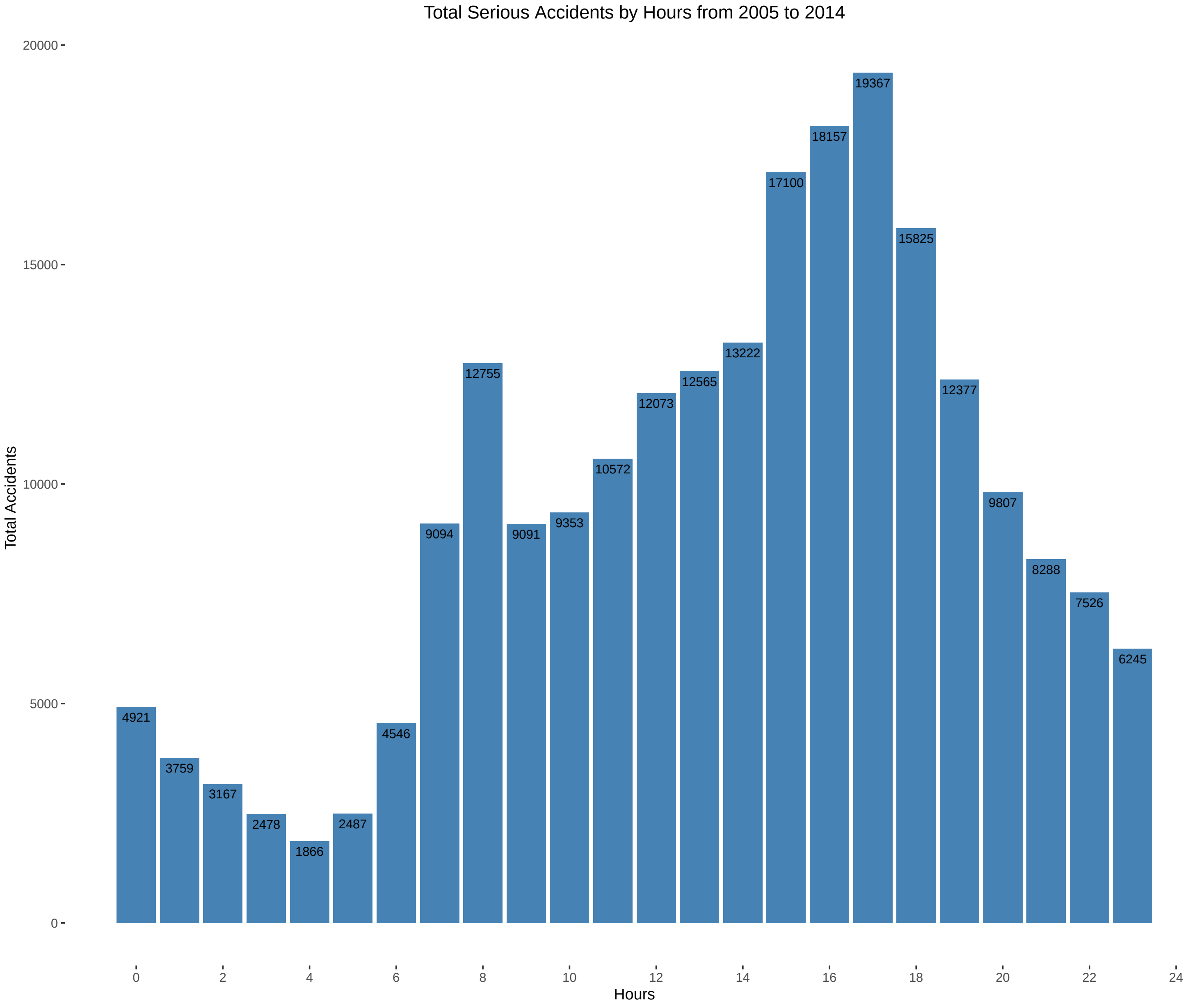


Serious Accident by hours

```
head(df$Casualty_Severity)
```

```
## [1] 2 3 3 3 3 3
```

```
df %>%
    filter(Casualty_Outcome == "Serious") %>%
    group_by(time_slot) %>%
    dplyr::summarize(total_accidents = n_distinct(Accident_Index)) %>%
    ggplot(aes(x = time_slot, y = total_accidents)) + geom_bar(stat = "identity",
    fill = "steelblue") + geom_text(aes(label = total_accidents), vjust = 1.6,
    color = "black", size = 3) + scale_x_continuous(breaks = round(seq(0,
    24, by = 2), 0)) + ggtitle("Total Serious Accidents by Hours from 2005 to 2014") +
    xlab("Hours") + ylab("Total Accidents") + theme(plot.title = element_text(hjust = 0.5),
    panel.background = element_blank())
```

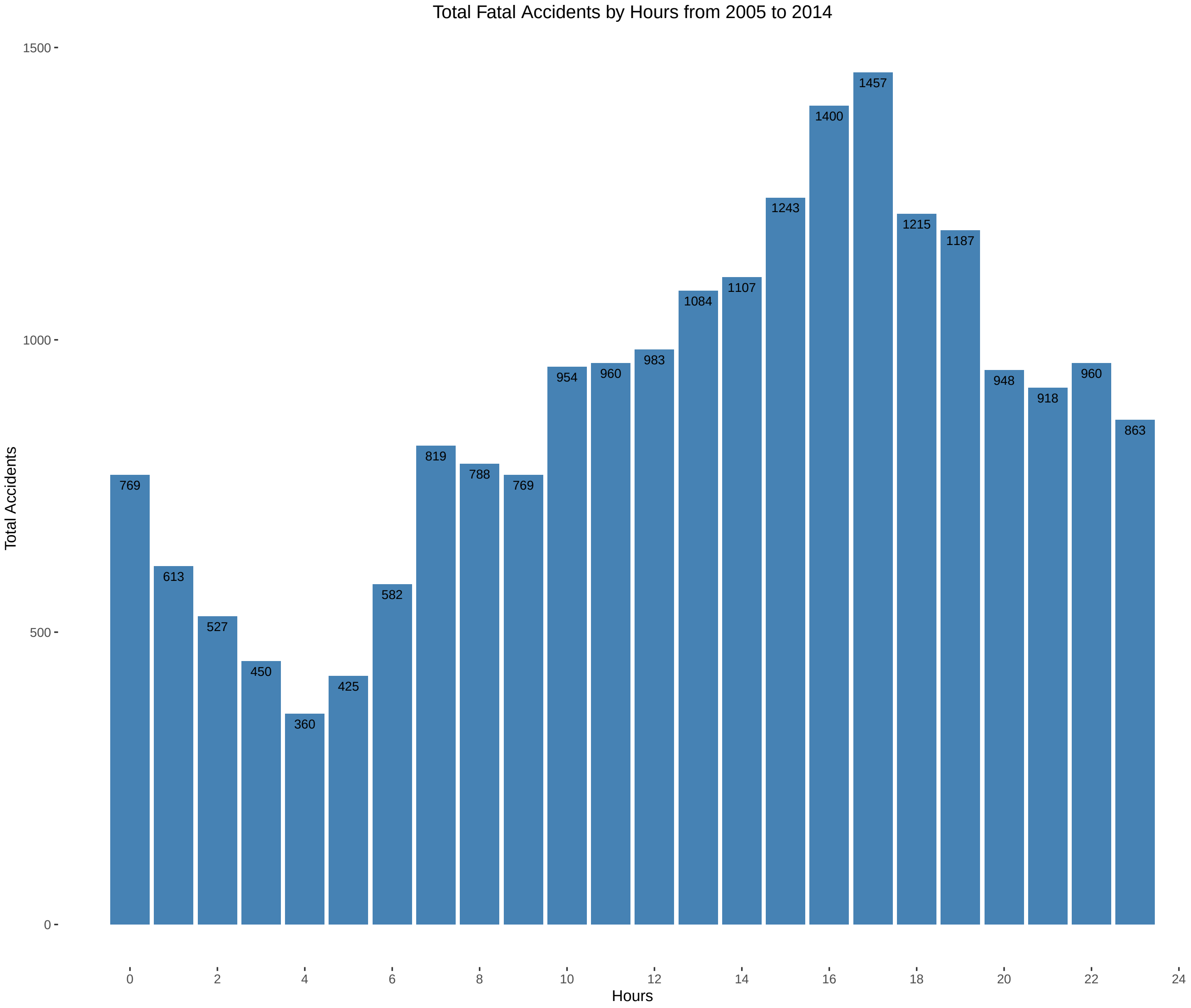


Fatal Accidents by hours

```
head(df$Casualty_Severity)
```

```
## [1] 2 3 3 3 3 3
```

```
df %>%
    filter(Casualty_Outcome == "Fatal") %>%
    group_by(time_slot) %>%
    dplyr::summarize(total_accidents = n_distinct(Accident_Index)) %>%
    ggplot(aes(x = time_slot, y = total_accidents)) + geom_bar(stat = "identity",
    fill = "steelblue") + geom_text(aes(label = total_accidents), vjust = 1.6,
    color = "black", size = 3) + scale_x_continuous(breaks = round(seq(0,
    24, by = 2), 0)) + ggtitle("Total Fatal Accidents by Hours from 2005 to 2014") +
    xlab("Hours") + ylab("Total Accidents") + theme(plot.title = element_text(hjust = 0.5),
    panel.background = element_blank())
```



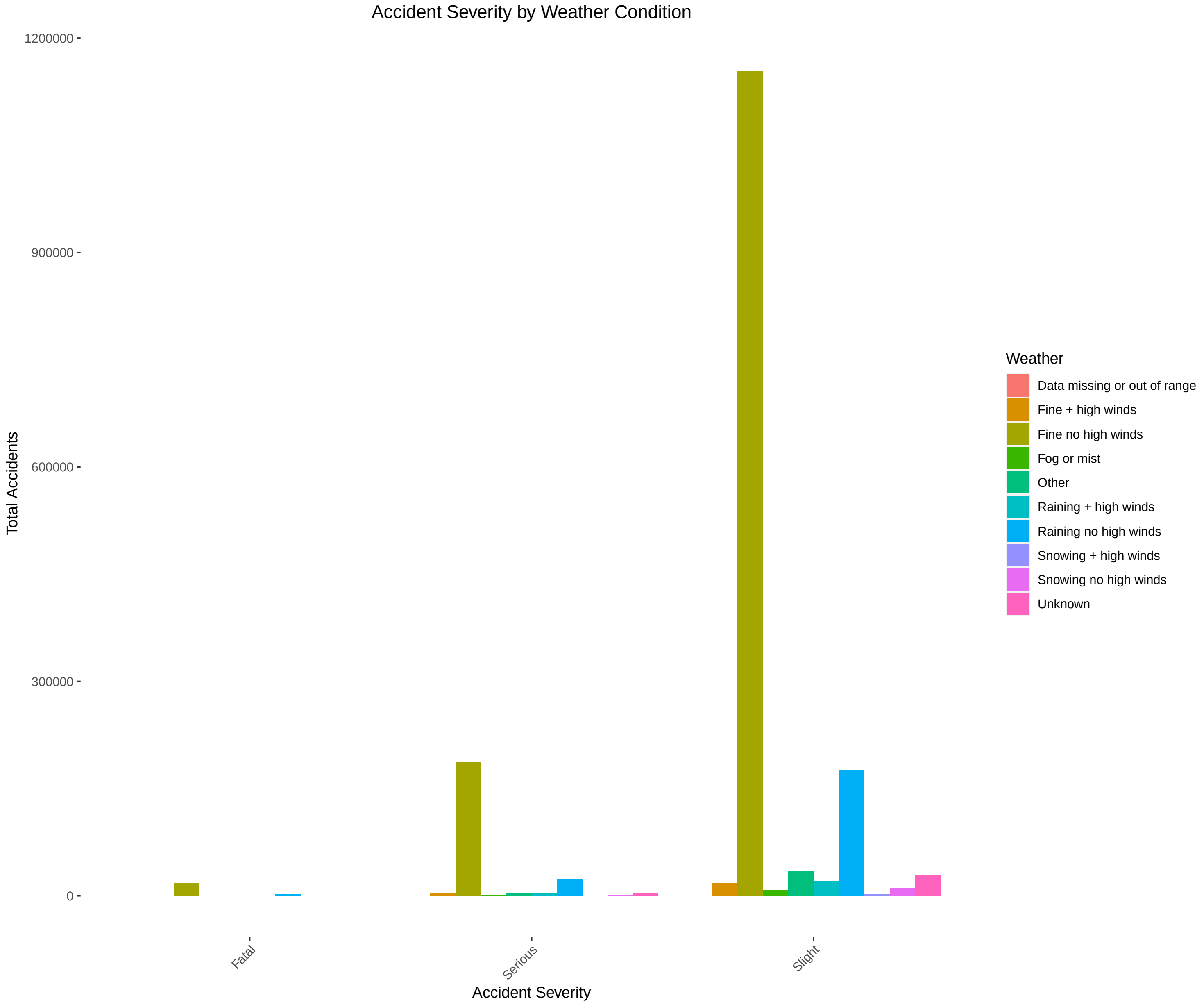
Contingency Table and Row Percentage

```
# Looking at the proportion table it seems that the hour when the
# accident occurs has an impact on the accident severity. We can
# observe that during the night the proportion of fatal accidents is
# higher than during the day while we observe the opposite result for
# the slight accidents. Result can be conclude using chi-square test.

acc_time_severity <- table(df$time_slot, df$Casualty_Outcome)
prop.table(acc_time_severity, 1)
```

```
##
##             Fatal     Serious      Slight
##   0  0.021861896 0.144985121 0.833152983
##   1  0.024825576 0.150697697 0.824476728
##   2  0.026132889 0.163191549 0.810675563
##   3  0.027645340 0.157643802 0.814710858
##   4  0.029419036 0.155451174 0.815129790
##   5  0.028080761 0.154692836 0.817226403
##   6  0.017307967 0.132727793 0.849964240
##   7  0.009280456 0.103431637 0.887287907
##   8  0.005340104 0.079921145 0.914738752
##   9  0.007100083 0.083528678 0.909371239
##   10 0.009902410 0.090959827 0.899137763
##   11 0.008340327 0.089898937 0.901760736
##   12 0.007638865 0.088334297 0.904026838
##   13 0.007935690 0.090285037 0.901779273
##   14 0.008880752 0.096156473 0.894962775
##   15 0.007749175 0.096827816 0.895423009
##   16 0.007897682 0.097013114 0.895089205
##   17 0.007280241 0.096227233 0.896492526
##   18 0.007905668 0.100456973 0.891637358
##   19 0.010719464 0.105627039 0.883653498
##   20 0.012939036 0.116994773 0.870066191
##   21 0.013592608 0.120276801 0.866130591
##   22 0.016509455 0.124017241 0.859473304
##   23 0.018486344 0.129439155 0.852074501
```

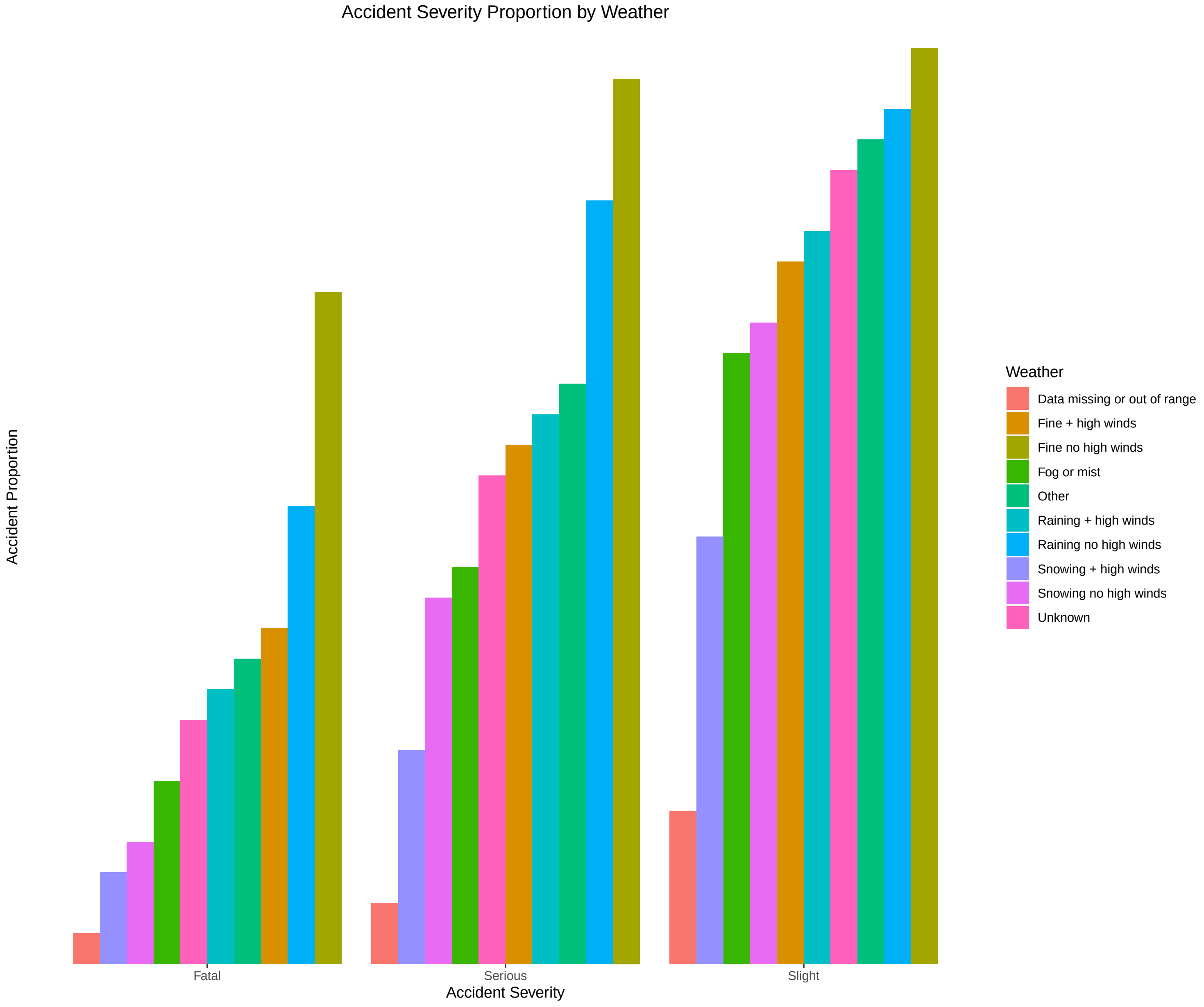
Accident Severity by Weather COndition

```
df %>%
    group_by(Weather, Casualty_Outcome) %>%
    dplyr::summarize(total_accidents = n_distinct(Accident_Index)) %>%
    ggplot(aes(x = Casualty_Outcome, y = total_accidents, fill = Weather)) +
    geom_bar(stat = "identity", position = "dodge") + ggtitle("Accident Severity by Weather Condition") +
    xlab("Accident Severity") + ylab("Total Accidents") + theme(plot.title = element_text(hjust = 0.5),
    panel.background = element_blank(), axis.text.x = element_text(angle = 45,
        hjust = 1))
```



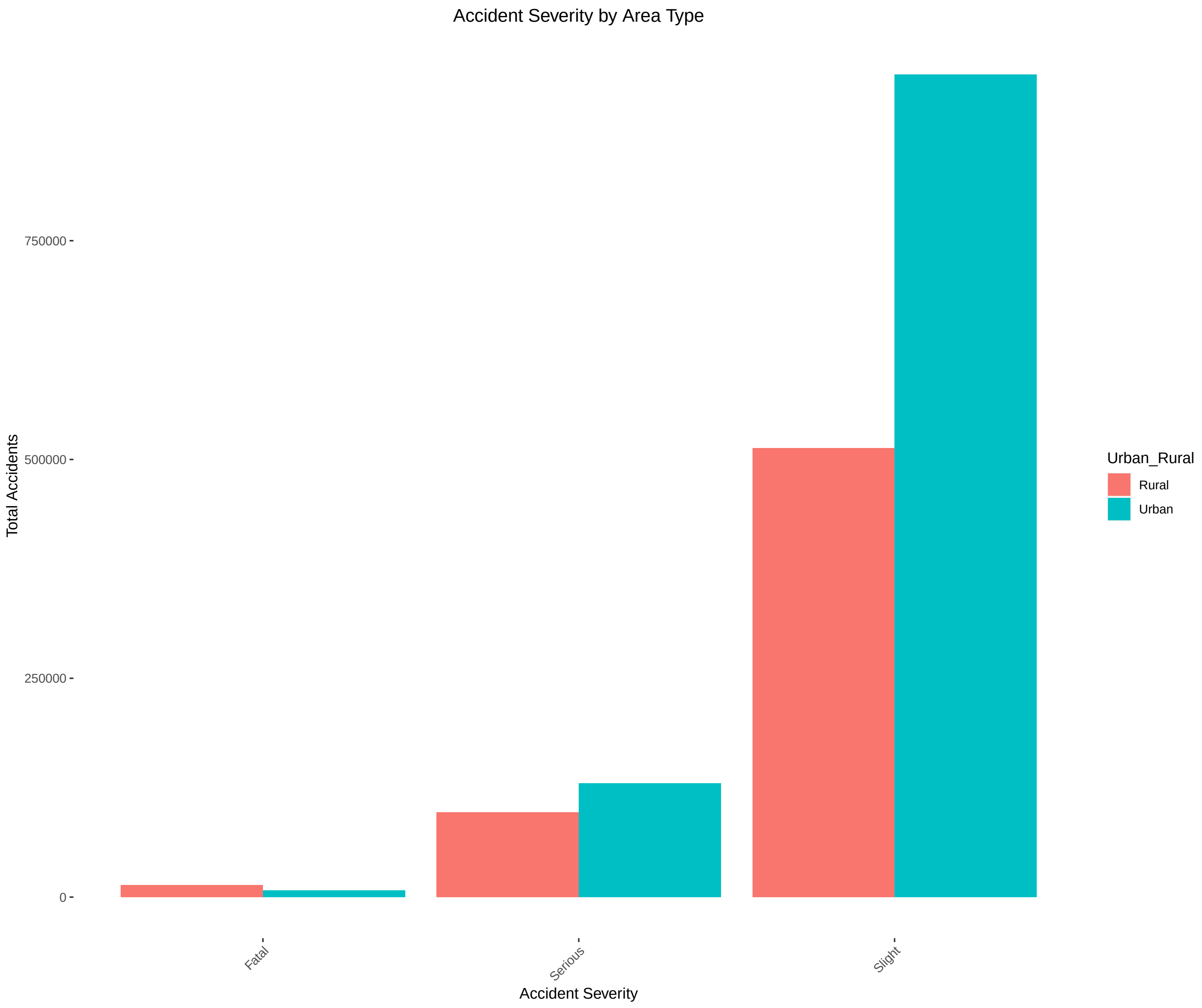
Accident Severity Proportion by Weather Condition

```
df %>%
    group_by(Weather, Casualty_Outcome) %>%
    dplyr::summarize(total_accidents = n_distinct(Accident_Index)) %>%
    mutate(freq = percent(total_accidents/sum(total_accidents))) %>%
    ggplot(aes(x = Casualty_Outcome, y = freq, fill = Weather)) + geom_bar(stat = "identity",
    position = "dodge") + ggtitle("Accident Severity Proportion by Weather") +
    xlab("Accident Severity") + ylab("Accident Proportion") + theme(plot.title = element_text(hjust = 0.5),
    panel.background = element_blank(), axis.text.y = element_blank(),
    axis.ticks.y = element_blank())
```



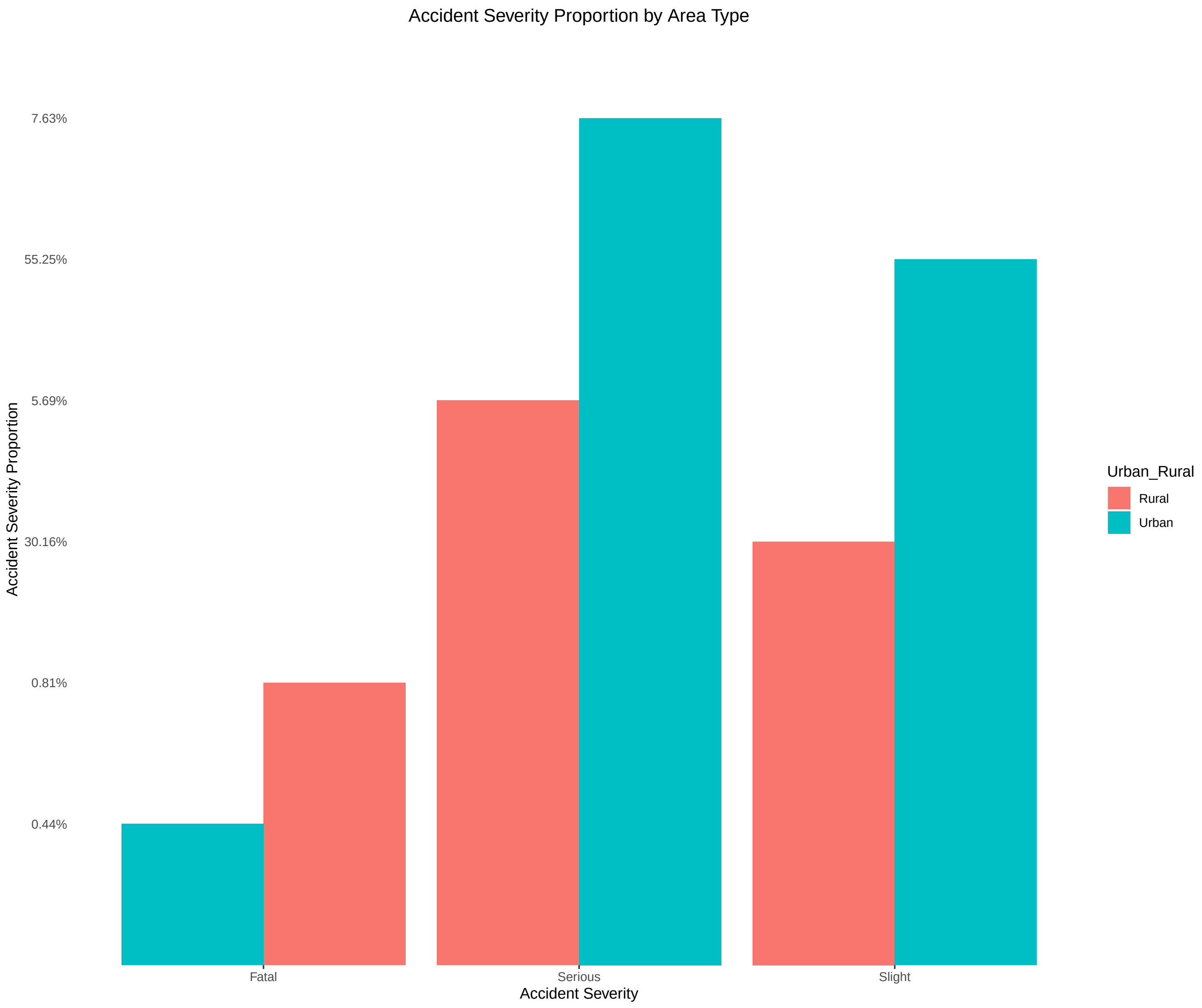
Accident Severity by Area Type

```
df %>%
    filter(Urban_Rural != "Unallocated") %>%
    group_by(Urban_Rural, Casualty_Outcome) %>%
    dplyr::summarize(total_accidents = n_distinct(Accident_Index)) %>%
    ggplot(aes(x = Casualty_Outcome, y = total_accidents, fill = Urban_Rural)) +
    geom_bar(stat = "identity", position = "dodge") + ggtitle("Accident Severity by Area Type") +
    xlab("Accident Severity") + ylab("Total Accidents") + theme(plot.title = element_text(hjust = 0.5),
    panel.background = element_blank(), axis.text.x = element_text(angle = 45,
        hjust = 1))
```



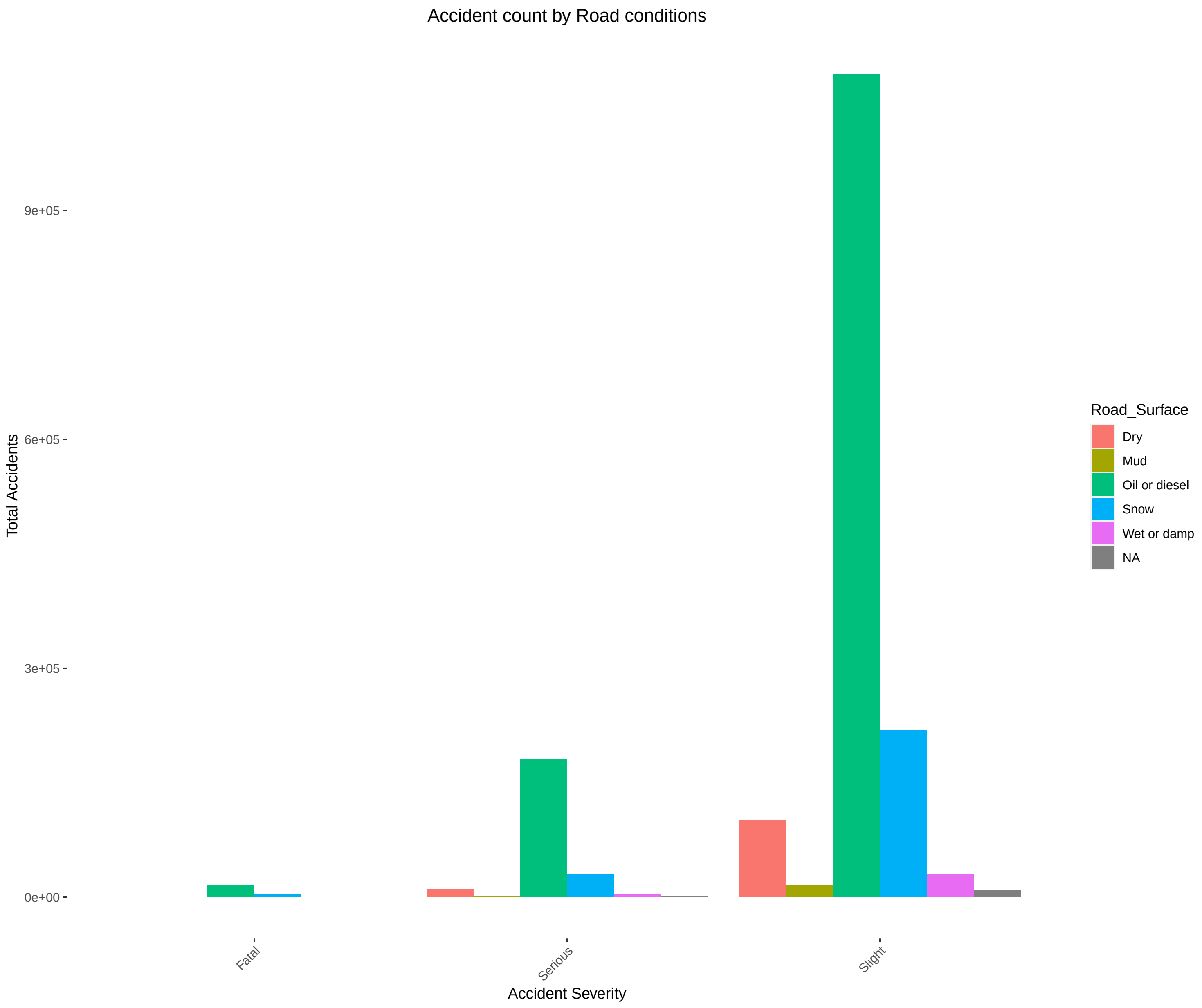
Accident Severity Proportion by Area Type

```
df %>%
    group_by(Urban_Rural, Casualty_Outcome) %>%
    filter(Urban_Rural != "Unallocated") %>%
    dplyr::summarize(total_accidents = n_distinct(Accident_Index)) %>%
    mutate(freq = percent(total_accidents/sum(total_accidents))) %>%
    ggplot(aes(x = Casualty_Outcome, y = freq, fill = Urban_Rural)) + geom_bar(stat = "identity",
    position = "dodge") + ggtitle("Accident Severity Proportion by Area Type") +
    xlab("Accident Severity") + ylab("Accident Severity Proportion") +
    theme(plot.title = element_text(hjust = 0.5), panel.background = element_blank(),
        axis.ticks.y = element_blank())
```

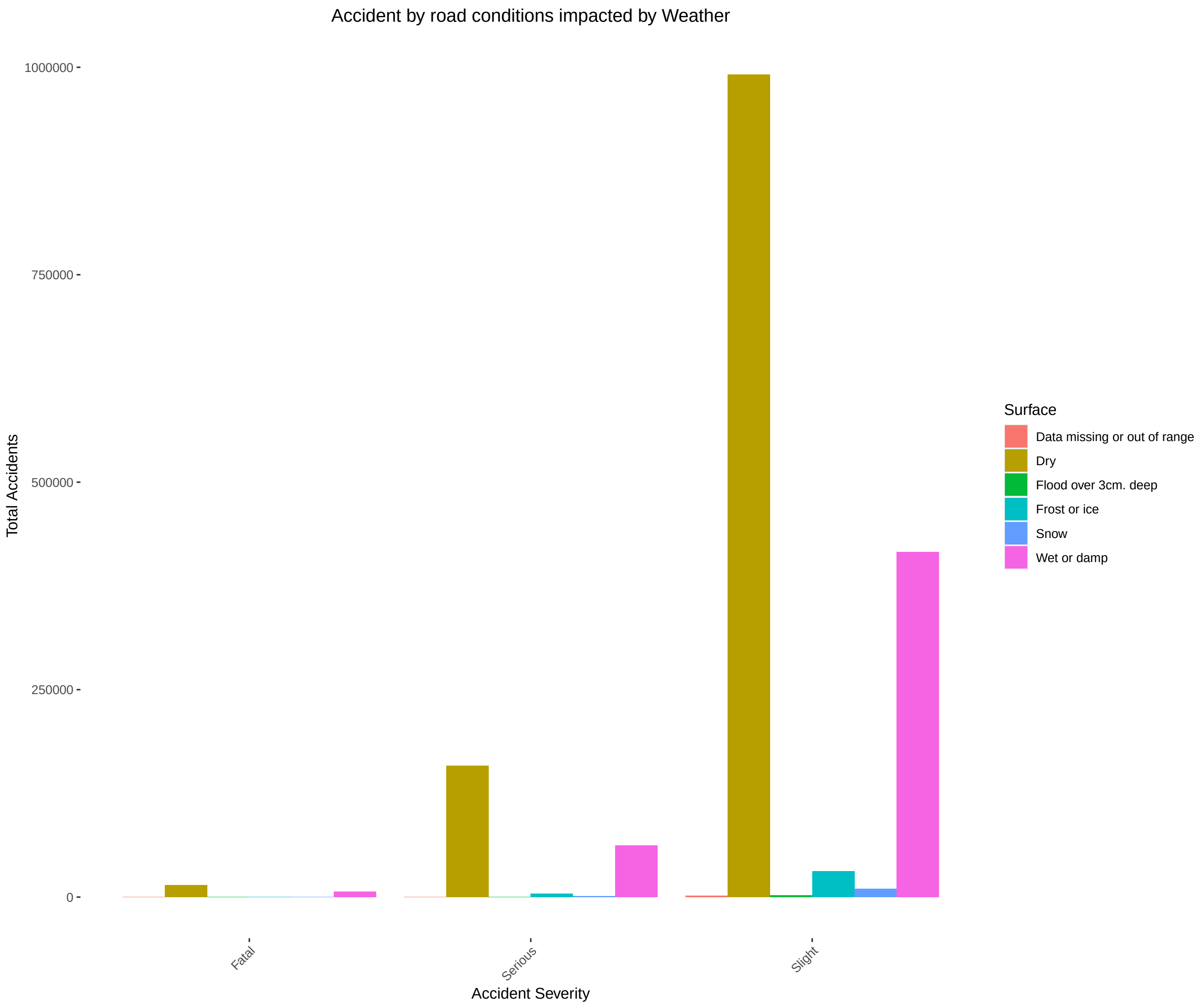


Road conditions contributing accidents.

```
df %>%
    group_by(Road_Surface, Casualty_Outcome) %>%
    dplyr::summarize(total_accidents = n_distinct(Accident_Index)) %>%
    ggplot(aes(x = Casualty_Outcome, y = total_accidents, fill = Road_Surface)) +
    geom_bar(stat = "identity", position = "dodge") + ggtitle("Accident count by Road conditions") +
    xlab("Accident Severity") + ylab("Total Accidents") + theme(plot.title = element_text(hjust = 0.5),
    panel.background = element_blank(), axis.text.x = element_text(angle = 45,
        hjust = 1))
```

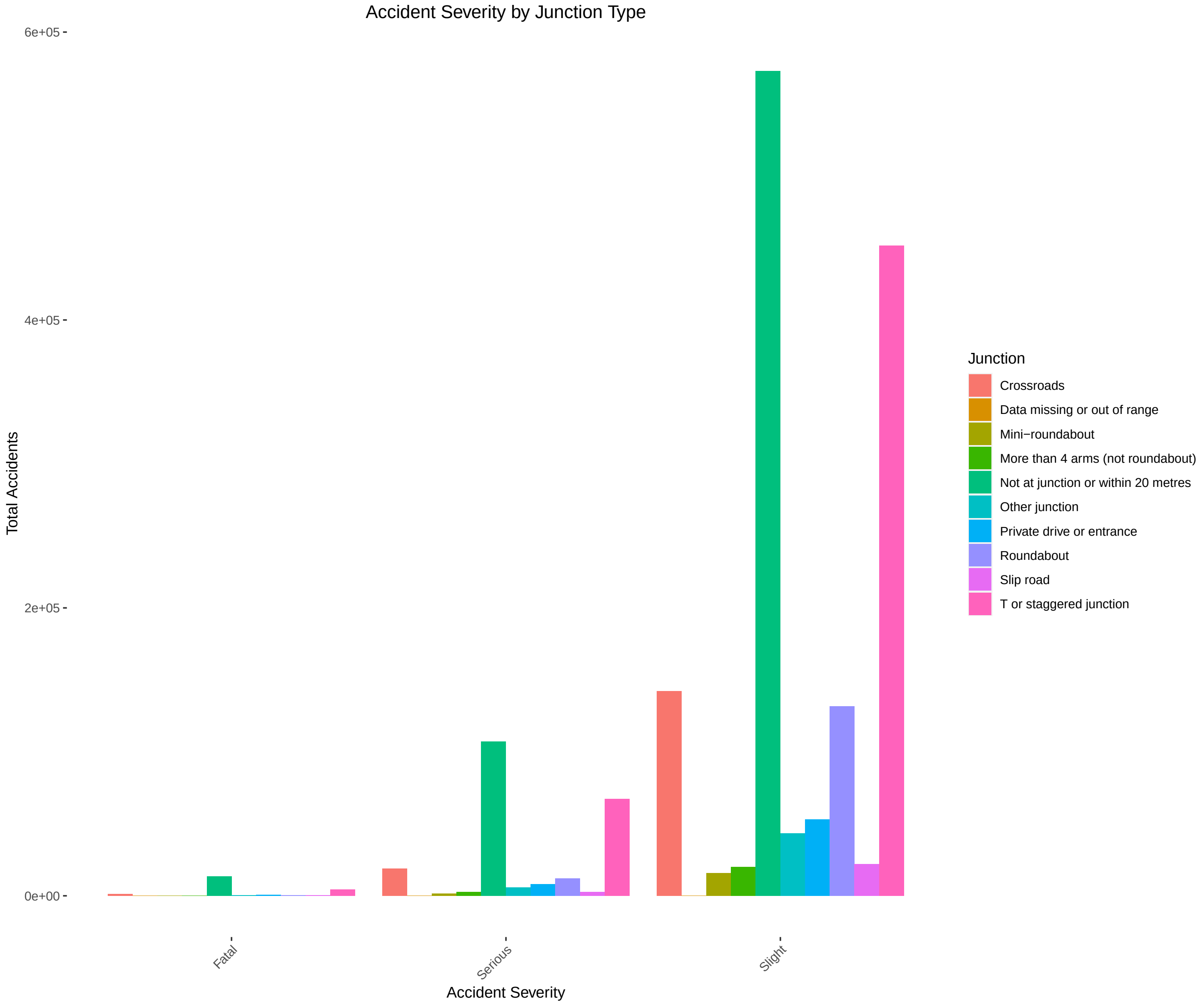


```
df %>%
    group_by(Surface, Casualty_Outcome) %>%
    dplyr::summarize(total_accidents = n_distinct(Accident_Index)) %>%
    ggplot(aes(x = Casualty_Outcome, y = total_accidents, fill = Surface)) +
    geom_bar(stat = "identity", position = "dodge") + ggtitle("Accident by road conditions impacted by Weather")
    xlab("Accident Severity") + ylab("Total Accidents") + theme(plot.title = element_text(hjust = 0.5),
    panel.background = element_blank(), axis.text.x = element_text(angle = 45,
        hjust = 1))
```



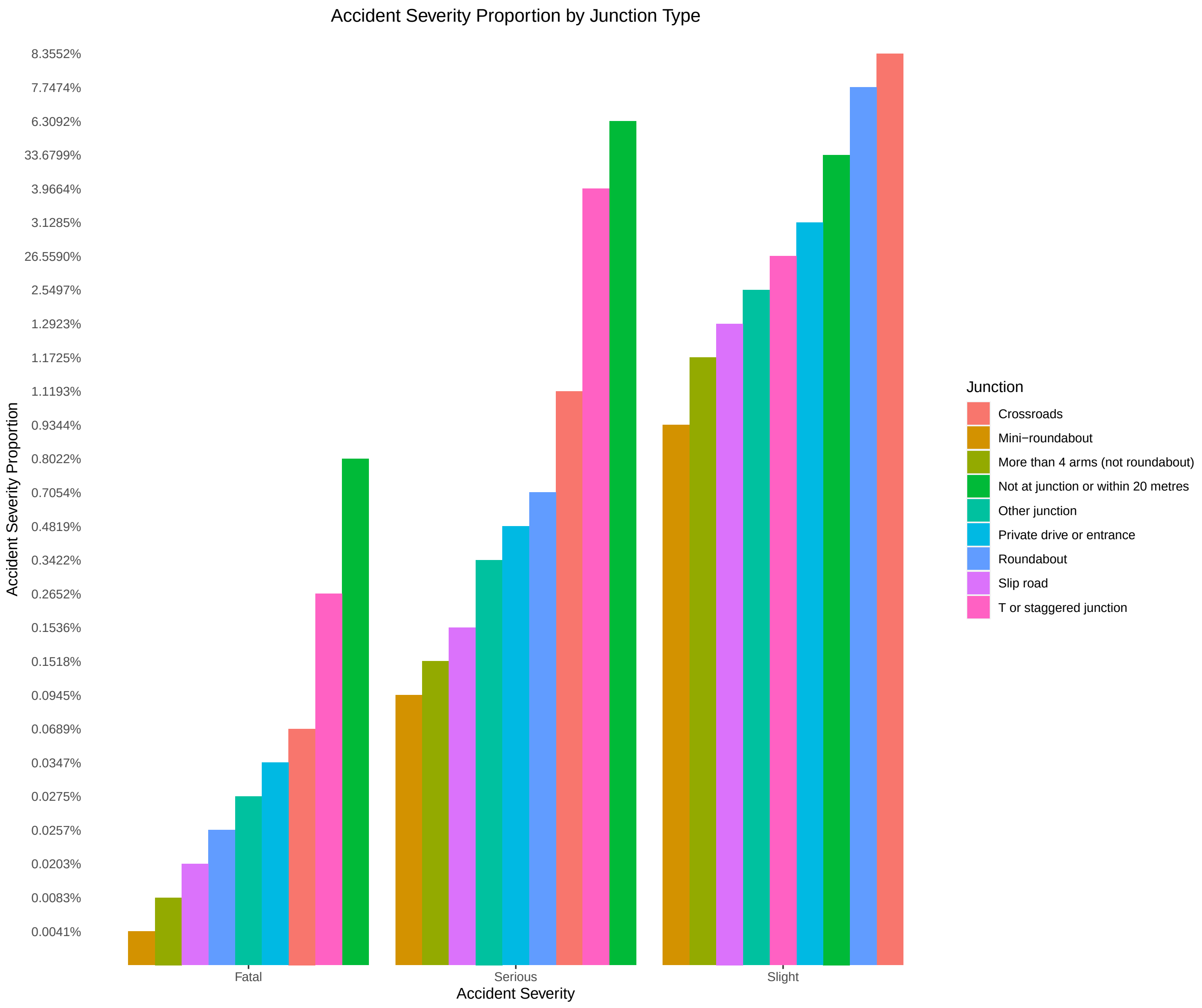
Accident Severity by Junction Type

```
df %>%
    group_by(Junction, Casualty_Outcome) %>%
    dplyr::summarize(total_accidents = n_distinct(Accident_Index)) %>%
    ggplot(aes(x = Casualty_Outcome, y = total_accidents, fill = Junction)) +
    geom_bar(stat = "identity", position = "dodge") + ggtitle("Accident Severity by Junction Type") +
    xlab("Accident Severity") + ylab("Total Accidents") + theme(plot.title = element_text(hjust = 0.5),
    panel.background = element_blank(), axis.text.x = element_text(angle = 45,
        hjust = 1))
```



Accident Severity Proportion by Junction Type

```
df %>%
    group_by(Junction, Casualty_Outcome) %>%
    filter(Junction != "Data missing or out of range") %>%
    dplyr::summarize(total_accidents = n_distinct(Accident_Index)) %>%
    mutate(freq = percent(total_accidents/sum(total_accidents))) %>%
    ggplot(aes(x = Casualty_Outcome, y = freq, fill = Junction)) + geom_bar(stat = "identity",
    position = "dodge") + ggtitle("Accident Severity Proportion by Junction Type") +
    xlab("Accident Severity") + ylab("Accident Severity Proportion") +
    theme(plot.title = element_text(hjust = 0.5), panel.background = element_blank(),
        axis.ticks.y = element_blank())
```



Accurancy Matrix

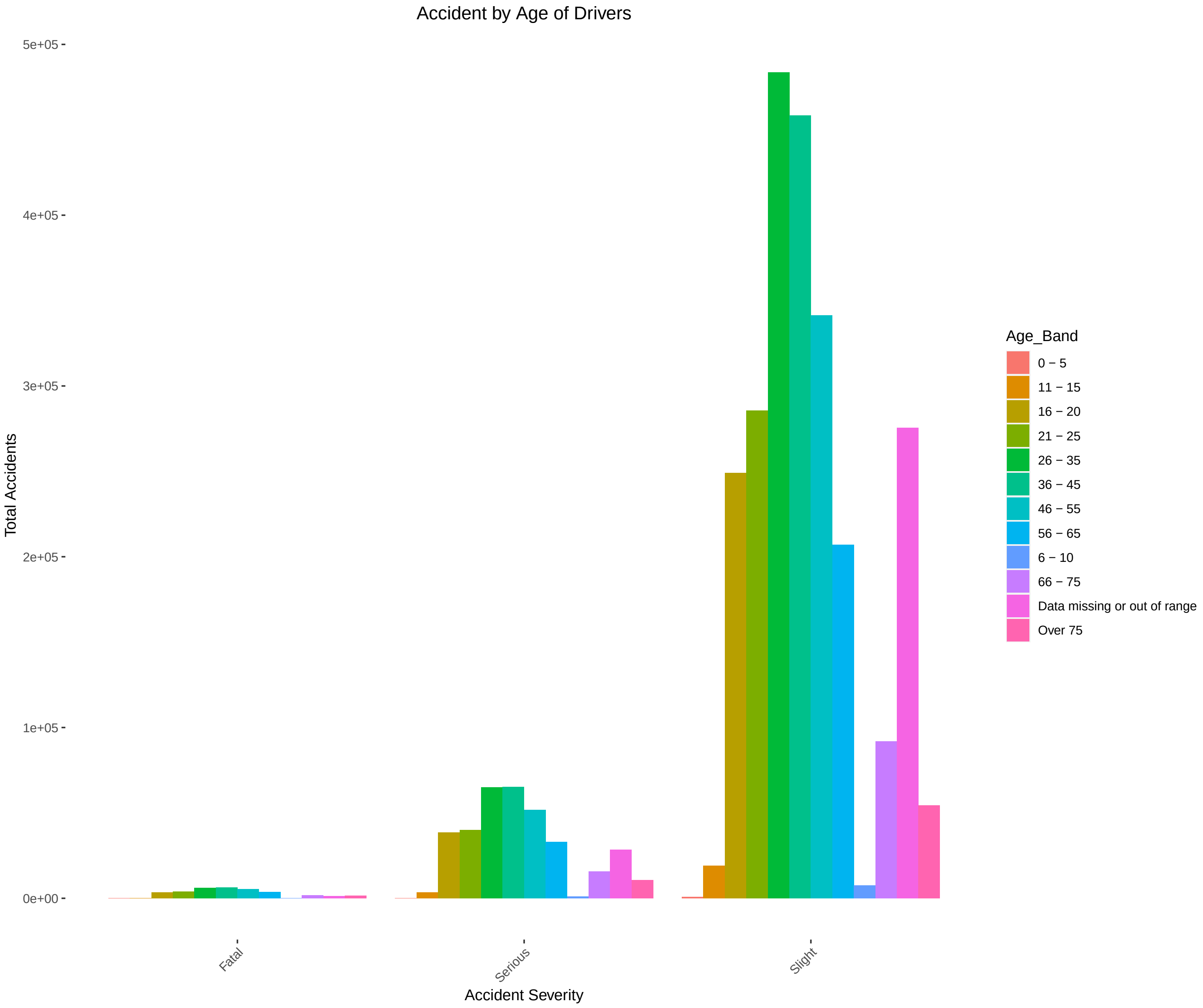
```
# We can see that the probability of an accident to be fatal is
# higher on road that ar enot a junction or within 20 metres of a
# junction. On the contrary an accident happening on a roundabout is
# much more likely to be a slight accident and not likely at all to
# be a fatal accident.

# Why I removed the rows labelled as 'Data missing or out of range'?
# There's only 26 rows with missing information over million rows so
# it is safe to remove them. And also as we can see in the below
# frquency table the proportion of the fatal accident for 'Data
# missing or out of range' would be missleading in our plot 5/26~19%
# while the second highest proportion is just 3%.
```

```
tt <- table(df$Junction, df$Casualty_Outcome)
prop.table(tt, 1)
```

```
##
##                                             Fatal     Serious      Slight
##   Crossroads                          0.004971314 0.084453076 0.910575610
##   Data missing or out of range        0.037735849 0.056603774 0.905660377
##   Mini-roundabout                     0.002679957 0.068573981 0.928746062
##   More than 4 arms (not roundabout)   0.003852718 0.077426114 0.918721168
##   Not at junction or within 20 metres 0.015837429 0.118661959 0.865500612
##   Other junction                      0.007350642 0.086299909 0.906349448
##   Private drive or entrance           0.007660691 0.103079321 0.889259988
##   Roundabout                          0.001941470 0.060376874 0.937681656
##   Slip road                           0.009421106 0.073704762 0.916874132
##   T or staggered junction             0.006421158 0.096987396 0.896591446
```

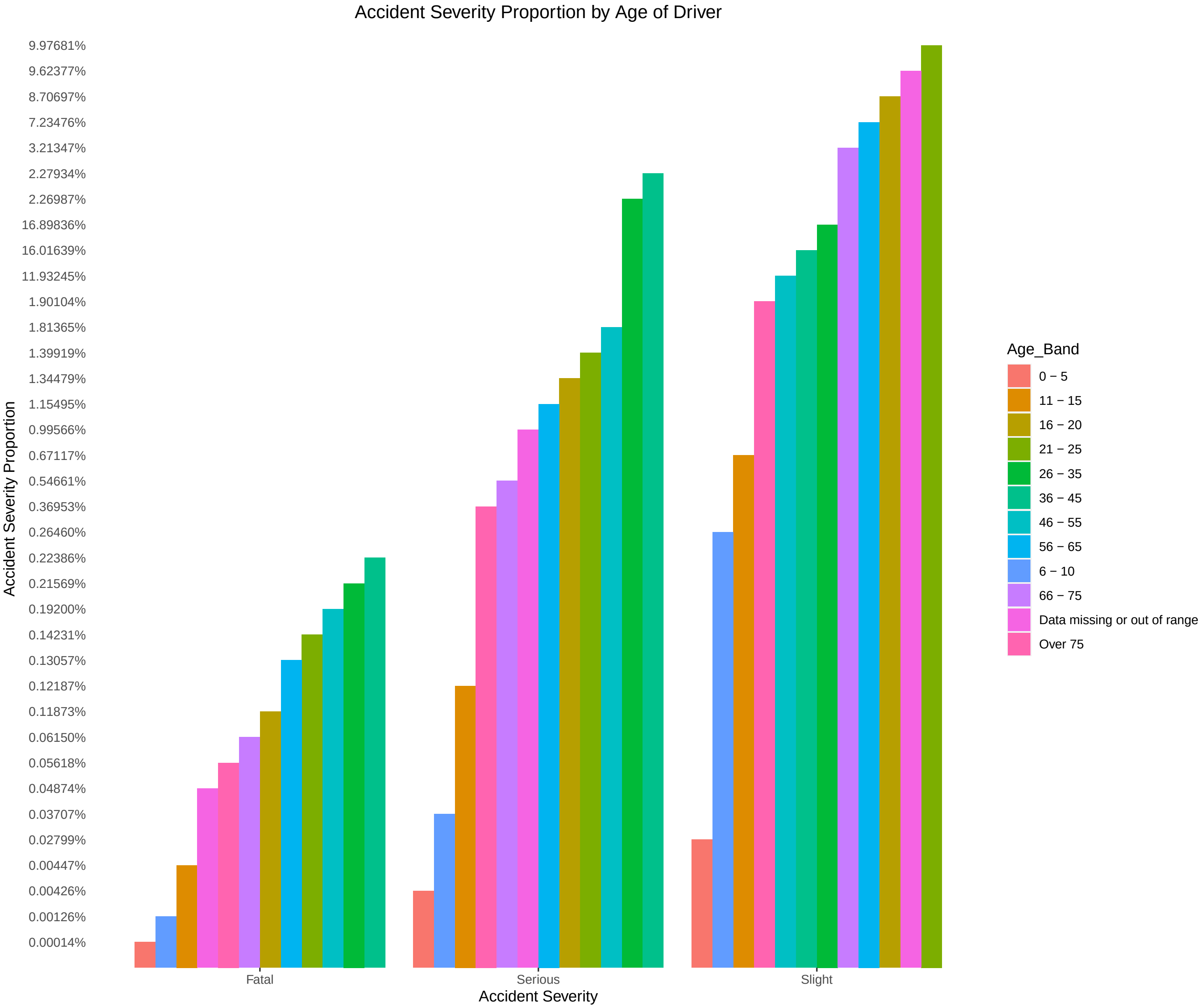
Accident Severity by Age of Drivers

```
df %>%
    group_by(Age_Band, Casualty_Outcome) %>%
    dplyr::summarize(total_accidents = n_distinct(Accident_Index)) %>%
    ggplot(aes(x = Casualty_Outcome, y = total_accidents, fill = Age_Band)) +
    geom_bar(stat = "identity", position = "dodge") + ggtitle("Accident by Age of Drivers") +
    xlab("Accident Severity") + ylab("Total Accidents") + theme(plot.title = element_text(hjust = 0.5),
    panel.background = element_blank(), axis.text.x = element_text(angle = 45,
        hjust = 1))
```



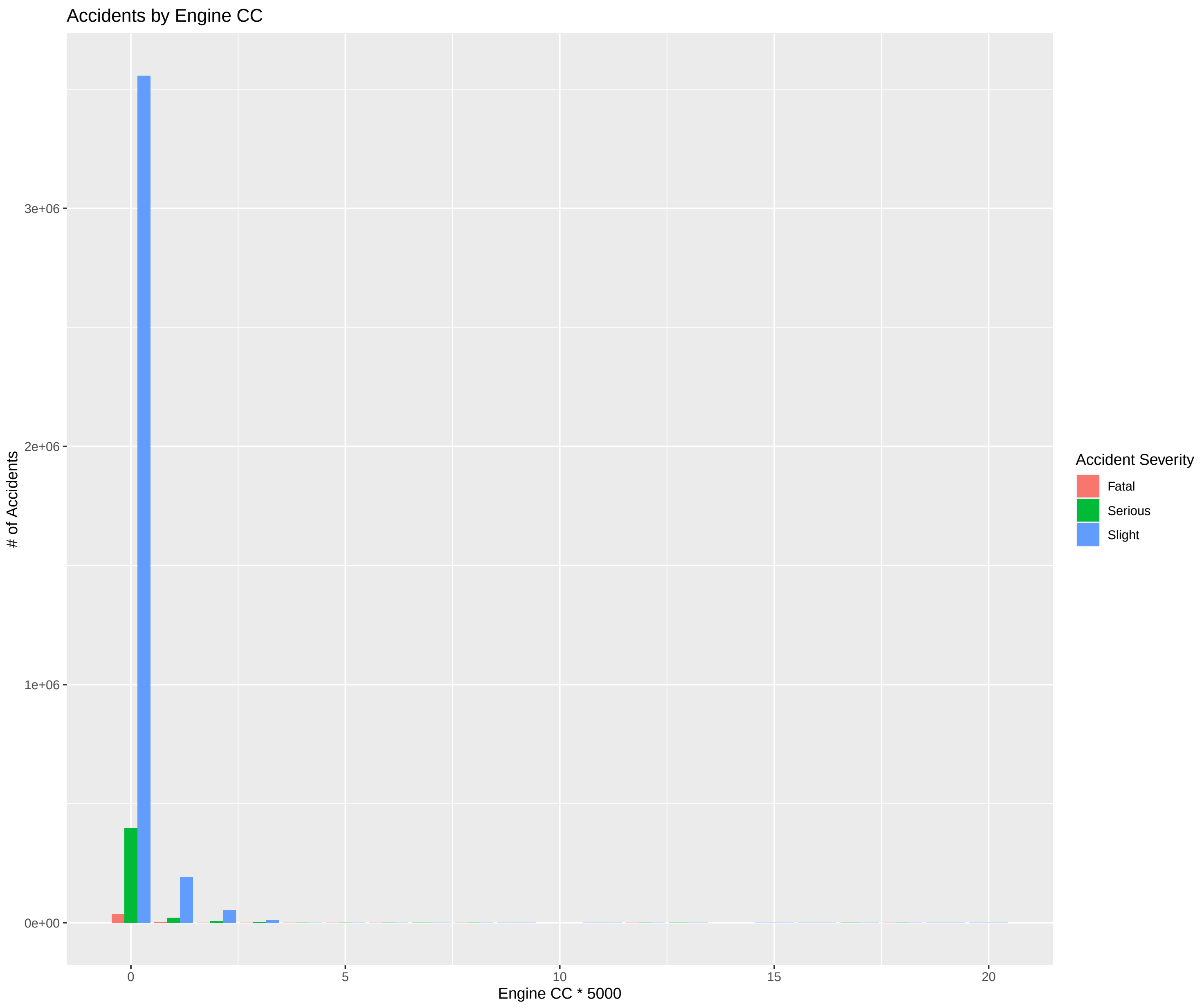
Accident Severity Proportion by Age of Driver

```
df %>%
    group_by(Age_Band, Casualty_Outcome) %>%
    dplyr::summarize(total_accidents = n_distinct(Accident_Index)) %>%
    mutate(freq = percent(total_accidents/sum(total_accidents))) %>%
    ggplot(aes(x = Casualty_Outcome, y = freq, fill = Age_Band)) + geom_bar(stat = "identity",
    position = "dodge") + ggtitle("Accident Severity Proportion by Age of Driver") +
    xlab("Accident Severity") + ylab("Accident Severity Proportion") +
    theme(plot.title = element_text(hjust = 0.5), panel.background = element_blank(),
        axis.ticks.y = element_blank())
```



Comparison by Engine CC

```
func_plotHistogram(df, round(df$Engine_Capacity_.CC./5000), df$Casualty_Outcome,
    "Engine CC * 5000", "# of Accidents", "Accidents by Engine CC", "Accident Severity")
```



Inferential Statistics

Test of Independence: Accident Severity vs Hours

```
# As the p-value is significantly less than 0.05, we reject with the
# Null hypothesis that the accident severity is independent of the
# hours.
chisq.test(acc_time_severity)
```

```
##
##  Pearson's Chi-squared test
##
## data:  acc_time_severity
## X-squared = 23317, df = 46, p-value < 2.2e-16
```

Test of Independence: Accident Severity vs Weekend night

- Again we reject with the Null hypothesis that the accident severity is independent of Weekend night hours.

Test of Independence: Accident Severity vs Weather, Area Type and Junction Type

```
# All our previous findings are with 95% CI statistically correct as
# we always have a p-value < 0.05

acc_weather_severity <- table(df$Weather, df$Casualty_Outcome)
acc_area_severity <- table(df$Urban_Rural, df$Casualty_Outcome)
acc_junction_severity <- table(df$Junction, df$Casualty_Outcome)
chisq.test(acc_weather_severity)
```

```
##
##  Pearson's Chi-squared test
##
## data:  acc_weather_severity
## X-squared = 3284.1, df = 18, p-value < 2.2e-16
```

```
chisq.test(acc_area_severity)
```

```
##
##  Pearson's Chi-squared test
##
## data:  acc_area_severity
## X-squared = 27715, df = 4, p-value < 2.2e-16
```

```
chisq.test(acc_junction_severity)
```

```
##
##  Pearson's Chi-squared test
##
## data:  acc_junction_severity
## X-squared = 28573, df = 18, p-value < 2.2e-16
```