Data Analytics Project

Saikat Roy, Vijayesh Kumar Das, Pooja Bhatia

28 July 2018

```r
require(pacman)
```

```
## Loading required package: pacman

## Warning: package 'pacman' was built under R version 3.3.3
```

```r
p_load(tidyverse, data.table, stringr)
```

Changing directory and loading the 'Caracteristics' CSV file

```r
setwd("C:\\Users\\Saikat-PC\\Desktop\\accidents-in-france-from-2005-to-2016")
data <- read.csv("caracteristics.csv", header = TRUE)
colnames(data)
```

```
##  [1] "Num_Acc" "an"      "mois"    "jour"    "hrmn"    "lum"     "agg"
##  [8] "int"     "atm"     "col"     "com"     "adr"     "gps"     "lat"
## [15] "long"    "dep"
```

```r
#PLEASE NOTE THAT "users" and "places" dataset have been loaded later but not displayed
```

PLEASE NOTE: "users" and "places" dataset have been loaded later but not displayed

```r
#Structure of data frame
colnames(data)
```

```
##  [1] "Num_Acc" "an"      "mois"    "jour"    "hrmn"    "lum"     "agg"
##  [8] "int"     "atm"     "col"     "com"     "adr"     "gps"     "lat"
## [15] "long"    "dep"
```

```r
str(data)
```

```
## 'data.frame':    839985 obs. of  16 variables:
##  $ Num_Acc: num  2.02e+11 2.02e+11 2.02e+11 2.02e+11 2.02e+11 ...
##  $ an     : int  16 16 16 16 16 16 16 16 16 16 ...
##  $ mois   : int  2 3 7 8 12 12 5 5 9 12 ...
##  $ jour   : int  1 16 13 15 23 23 1 14 23 30 ...
##  $ hrmn   : int  1445 1800 1900 1930 1100 1115 1145 1915 1900 1030 ...
##  $ lum    : int  1 1 1 2 1 1 1 2 1 1 ...
##  $ agg    : int  2 2 1 2 2 2 2 1 2 1 ...
##  $ int    : int  1 6 1 1 3 1 1 1 1 1 ...
##  $ atm    : int  8 1 1 7 1 7 7 1 1 9 ...
##  $ col    : int  3 6 6 3 3 6 2 1 3 6 ...
##  $ com    : int  5 5 11 477 11 11 51 250 51 303 ...
##  $ adr    : Factor w/ 364689 levels "","'' En pain Chaud ''",..: 168471 102730 1 177512 346222 34585
##  $ gps    : Factor w/ 11 levels "","0","A","C",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ lat    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ long   : Factor w/ 140606 levels "","-","-100",..: 35756 35756 35756 35756 35756 35756 35756 3575
##  $ dep    : int  590 590 590 590 590 590 590 590 590 590 ...
```

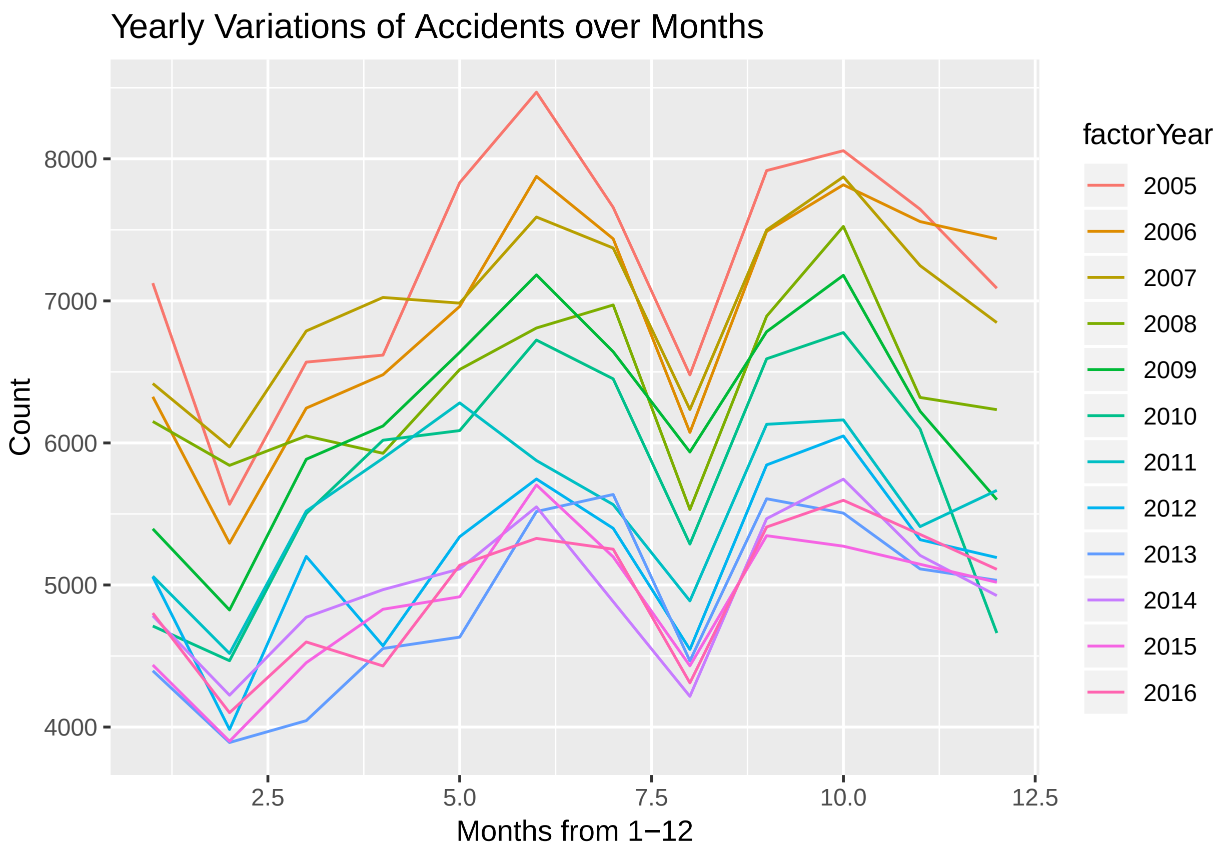
```r
# Short glimpse of data set
glimpse(data)
```

```
## Observations: 839,985
## Variables: 16
## $ Num_Acc <dbl> 2.016e+11, 2.016e+11, 2.016e+11, 2.016e+11, 2.016e+11,...
## $ an      <int> 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16...
## $ mois    <int> 2, 3, 7, 8, 12, 12, 5, 5, 9, 12, 1, 1, 2, 4, 8, 9, 11,...
## $ jour    <int> 1, 16, 13, 15, 23, 23, 1, 14, 23, 30, 25, 28, 5, 17, 1...
## $ hrmn    <int> 1445, 1800, 1900, 1930, 1100, 1115, 1145, 1915, 1900, ...
## $ lum     <int> 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 2, 3, 1, 1, 1, 1, 3, 2, ...
## $ agg     <int> 2, 2, 1, 2, 2, 2, 2, 1, 2, 1, 2, 1, 2, 2, 1, 1, 1, 2, ...
## $ int     <int> 1, 6, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 6, 1, 1, 1, ...
## $ atm     <int> 8, 1, 1, 7, 1, 7, 7, 1, 1, 9, 8, 1, 1, 1, 7, 1, 8, 5, ...
## $ col     <int> 3, 6, 6, 3, 3, 6, 2, 1, 3, 6, 6, 7, 6, 5, 5, 5, 6, 3, ...
## $ com     <int> 5, 5, 11, 477, 11, 11, 51, 250, 51, 303, 466, 197, 466...
## $ adr     <fct> 46, rue Sonneville, 1a rue du cimetière, , 52 rue vict...
## $ gps     <fct> M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, ...
## $ lat     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ long    <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ dep     <int> 590, 590, 590, 590, 590, 590, 590, 590, 590, 590, 590,...
```

Yearly Variations of Accidents over Months (1-12)

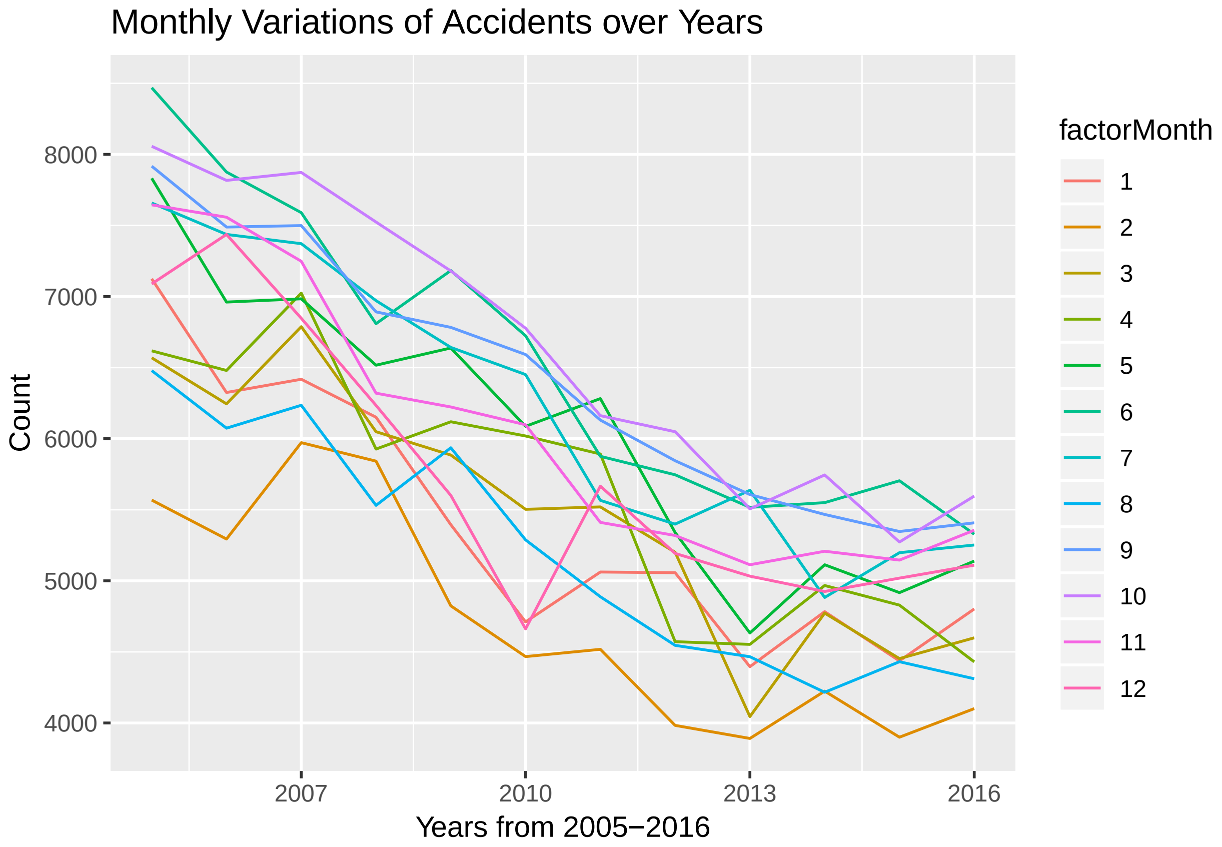
```
sample = sample_n(data,100000)

data %>% mutate(factorYear=as.factor(2000+an)) %>%
  group_by(factorYear, mois) %>% summarise(Count=n()) %>%
  ggplot(aes(x = mois, color=factorYear, y=Count))+
  ggtitle("Yearly Variations of Accidents over Months")+
  geom_line() + xlab("Months from 1-12")
```

```
## Warning: package 'bindrcpp' was built under R version 3.3.3
```



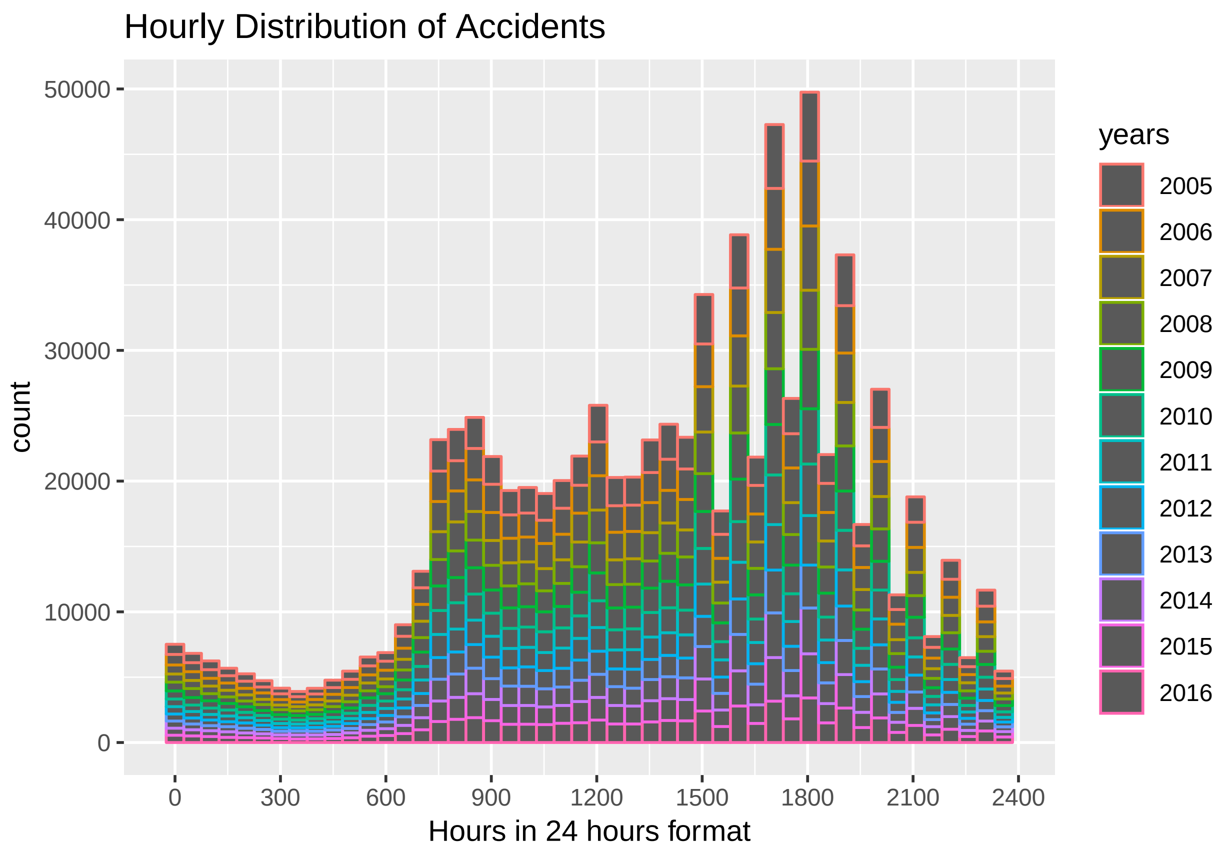
Monthly Variations of Accidents over Years (2005-2016)

```
data %>% mutate(factorMonth=as.factor(mois)) %>%
  group_by(factorMonth, an) %>% summarise(Count=n()) %>%
  ggplot(aes(x = an+2000, color=factorMonth, y=Count))+ geom_line() +
  xlab("Years from 2005-2016")+
  ggtitle("Monthly Variations of Accidents over Years")
```



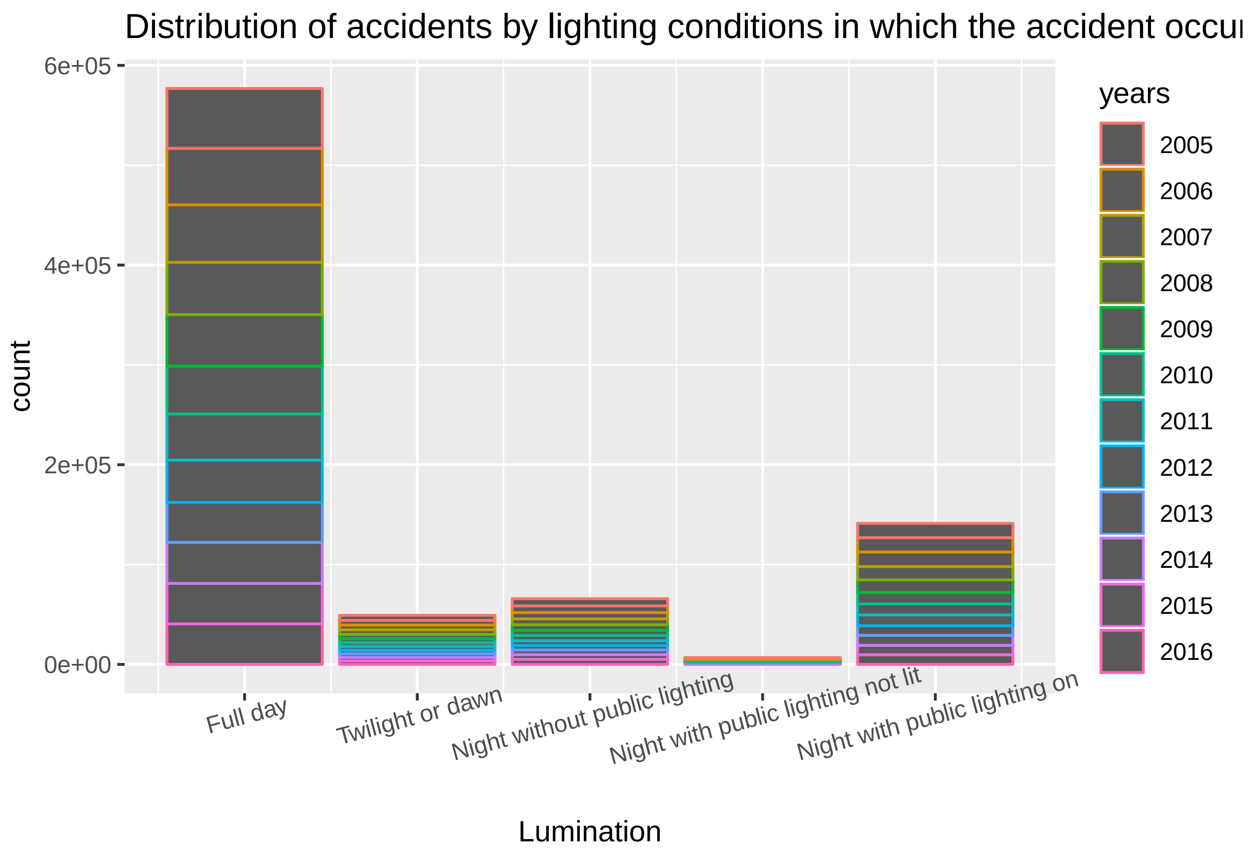
Hourly Distribution of Accidents

```
data %>% mutate(years=as.factor(2000+an)) %>%
  #group_by(factorYear, hrmn) %>% summarise(Count=n()) %>%
  ggplot(aes(x = hrmn, color=years))+
  geom_histogram(bins=48)+ xlab("Hours in 24 hours format") +
  ggtitle("Hourly Distribution of Accidents")+
  scale_x_continuous(breaks = seq(0, 2400, by = 300))+
  ggtitle("Hourly Distribution of Accidents")
```



Distribution of accidents by lighting conditions in which the accident occurred

```
data %>% mutate(years=as.factor(2000+an)) %>%
ggplot(aes(x = lum, color=years)) + ggtitle("Distribution of accidents by lighting conditions in which
scale_x_continuous(breaks = 1:5,labels=c("Full day", "Twilight or dawn", "Night without public lighting
theme(axis.text.x = element_text(angle = 15))
```

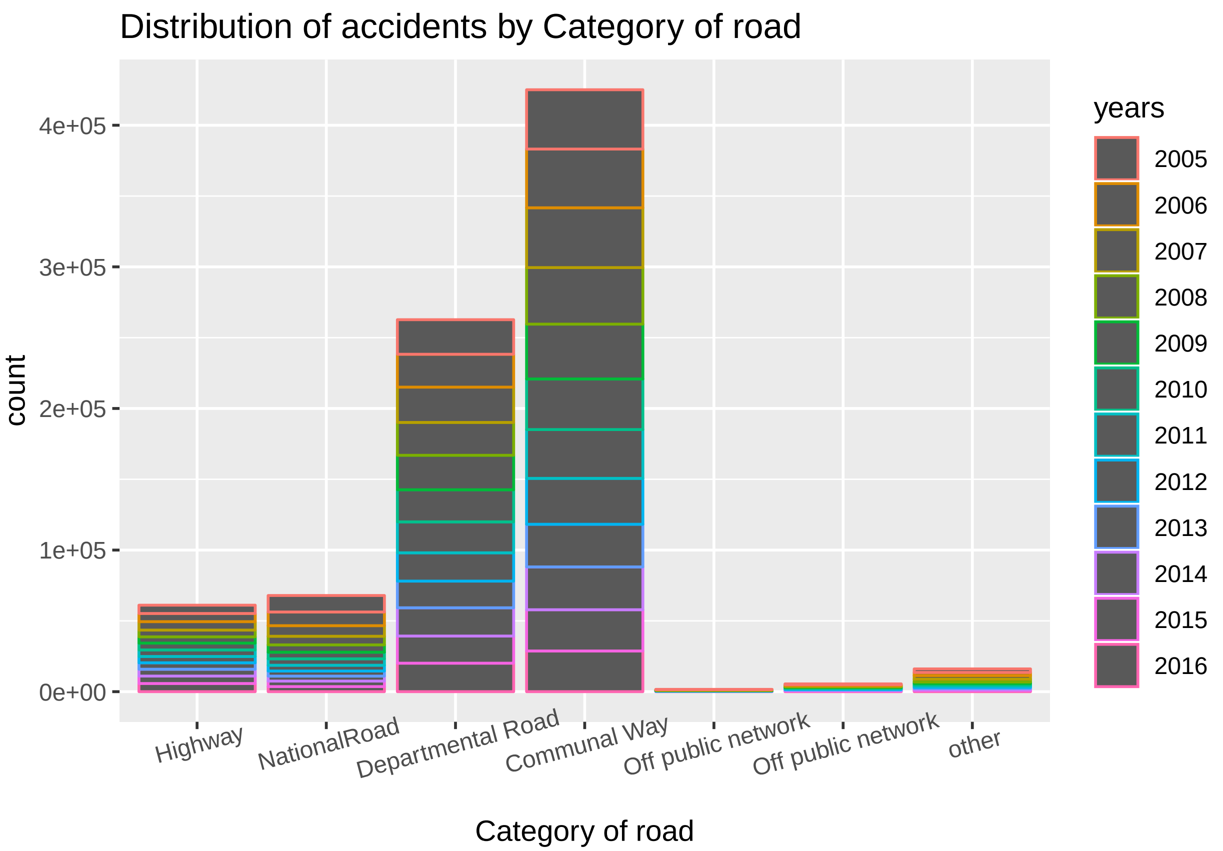


```
print(plot)
```

```
## function (x, y, ...)
## UseMethod("plot")
## <bytecode: 0x00000000182e3ba0>
## <environment: namespace:graphics>
```

Distribution of accidents by Category of road

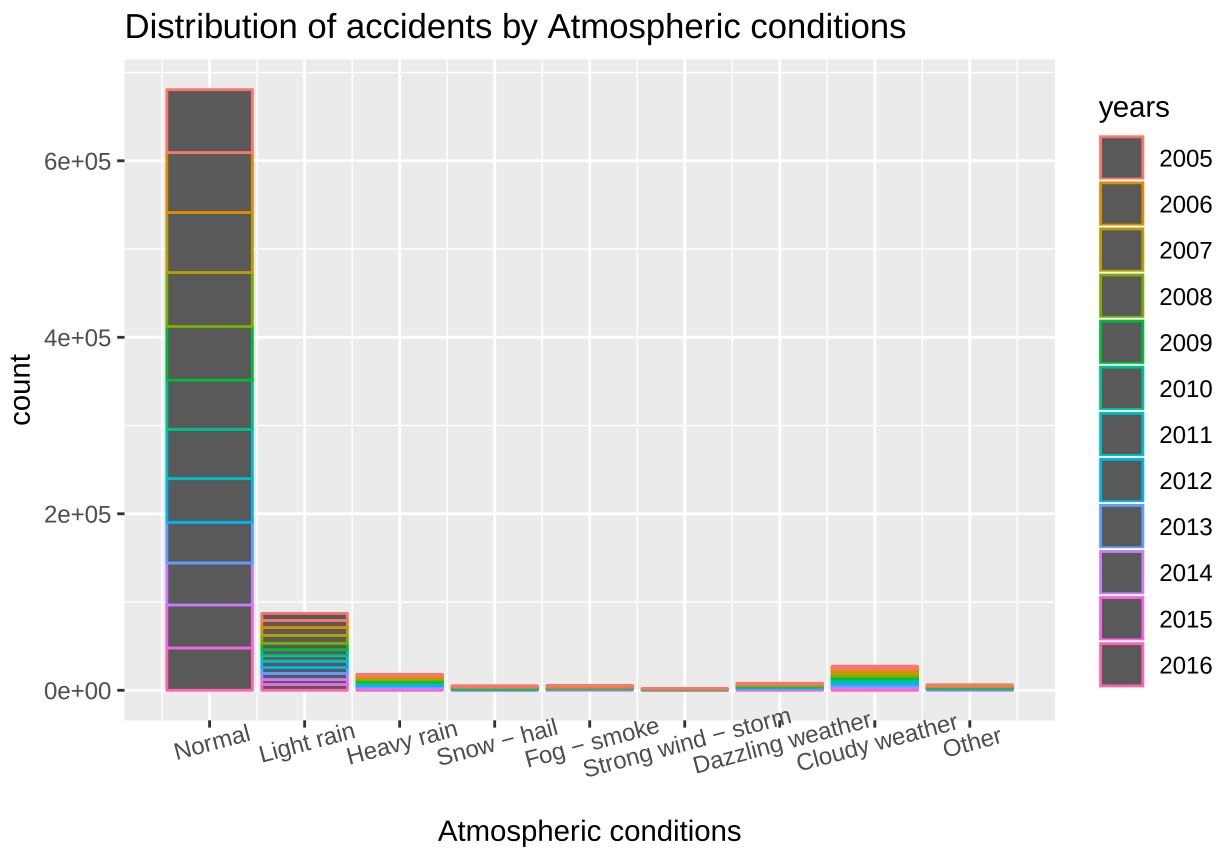
```
data1<-read.csv("places.csv", header = TRUE)
data1$year<-data[match(data$Num_Acc , data1$Num_Acc), "an"]
data1$catr<-as.factor(data1$catr)
data1%>% filter(data1$catr!=0) %>% mutate(years=as.factor(2000+year)) %>%
ggplot(aes(x =catr, color=years)) + ggtitle("Distribution of accidents by Category of road") +xlab("Cat
scale_x_discrete(breaks = c("1","2","3","4","5","6","9"),labels=c("Highway","NationalRoad","Departmenta
theme(axis.text.x = element_text(angle = 15))
```



Distribution of accidents by Atmospheric conditions

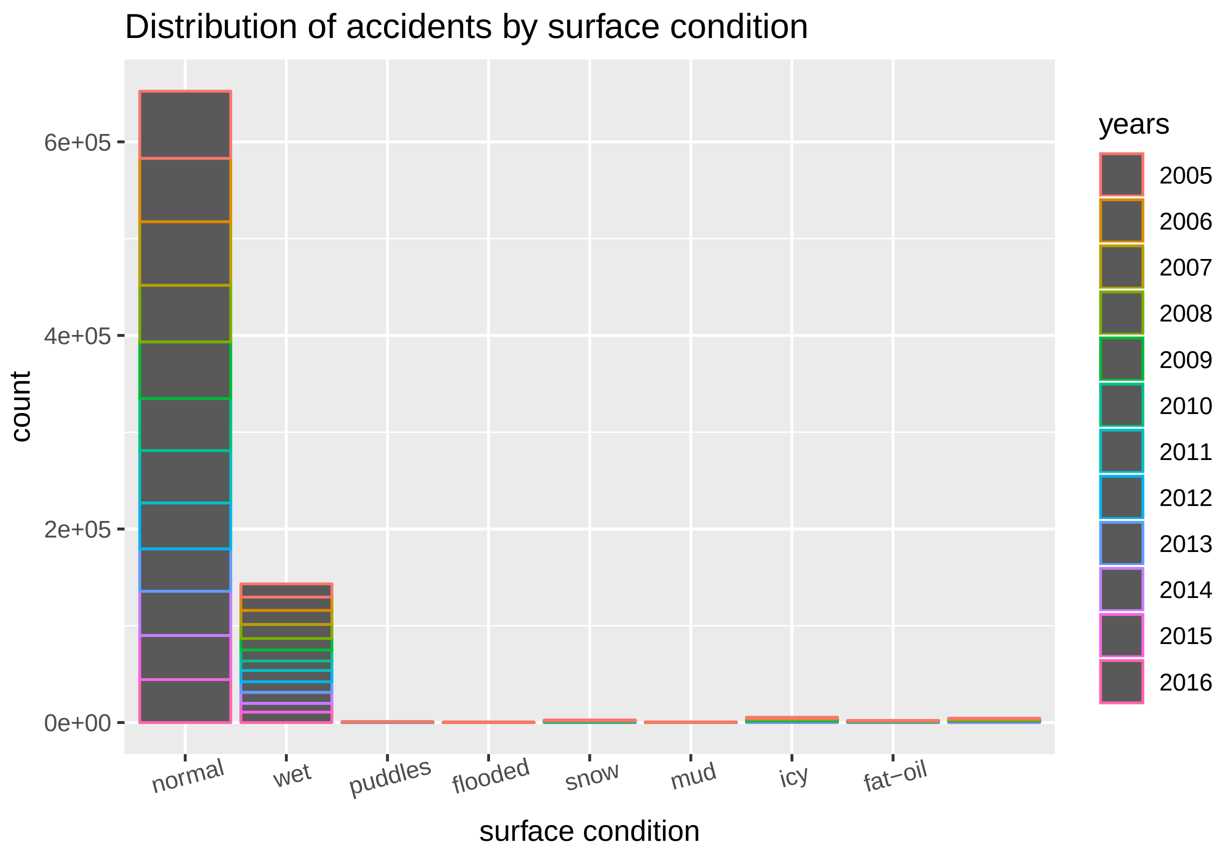
```
data %>% mutate(years=as.factor(2000+an)) %>%
ggplot(aes(x = atm, color=years)) + ggtitle("Distribution of accidents by Atmospheric conditions") +xlab
scale_x_continuous(breaks = 1:9,labels=c("Normal","Light rain","Heavy rain","Snow - hail","Fog - smoke"
theme(axis.text.x = element_text(angle = 15))
```

```
## Warning: Removed 55 rows containing non-finite values (stat_count).
```



# Distribution of accidents by surface condition

```
data1$surf<-as.factor(data1$surf)
data1 %>% filter(data1$surf!=0) %>% mutate(years=as.factor(2000+year)) %>%
ggplot(aes(x =surf, color=years)) + ggtitle("Distribution of accidents by surface condition") +
  xlab("surface condition") + ylab("count")+geom_bar()+
scale_x_discrete(breaks = c("1","2","3","4","5","6","7","8"),labels=c("normal"," wet","puddles","flooded
theme(axis.text.x = element_text(angle = 15))
```

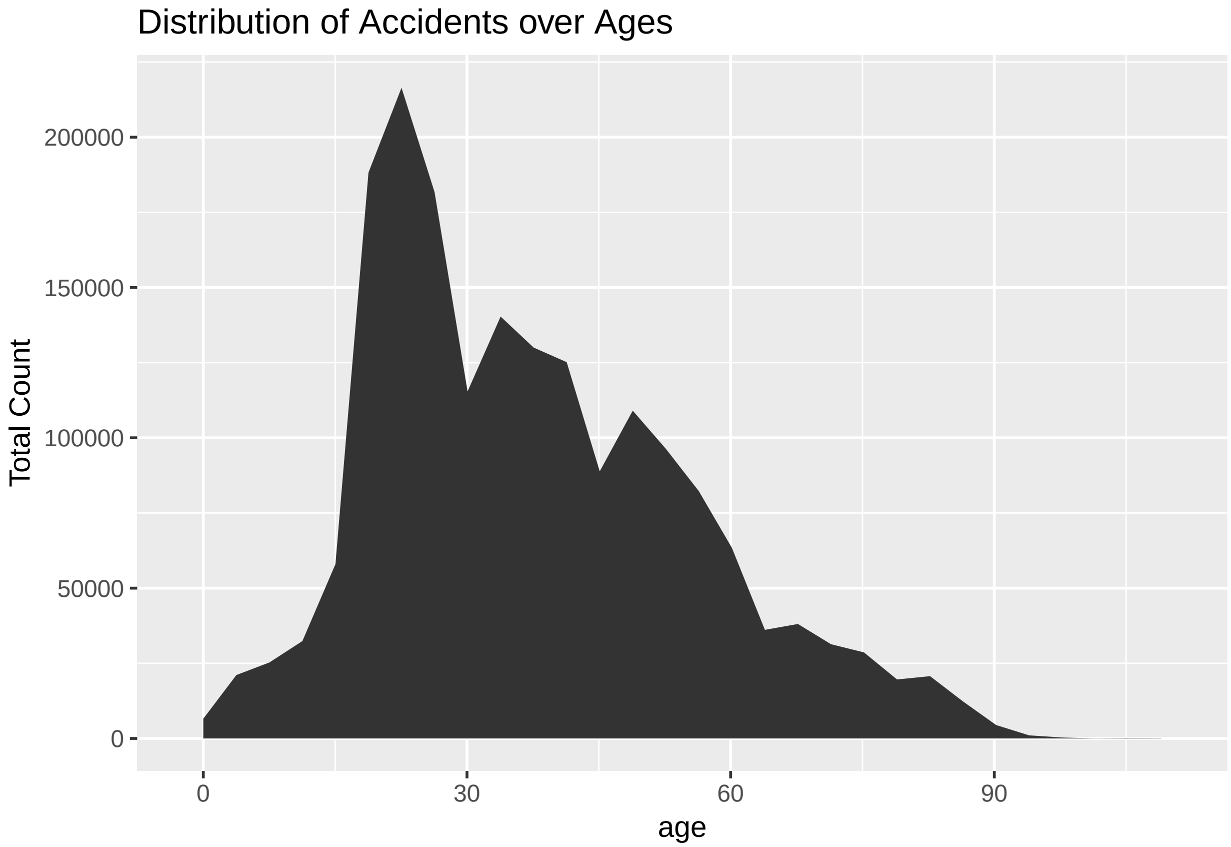


#Accidents according to the age

```r
users<-read.csv("users.csv", header = TRUE)

users$accidentYear <- data[match(users$Num_Acc , data$Num_Acc), "an"]
users$age <- (as.integer(users$accidentYear)+2000) - users$an_nais

ggplot(users, aes(x=age, fill = NULL))+
  geom_area(stat = "bin")+
  xlab("age")+
  ylab("Total Count")+ggtitle("Distribution of Accidents over Ages")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2351 rows containing non-finite values (stat_bin).
```


Distribution of Accidents over Ages
Total Count
200000
150000
100000
50000
0
0
30
60
90
age