Analysis of the UK accident data

Xue Yu

December 8, 2017

Introduction

Causalty from road accidents is one the most reasons for human mortality in the modern world. For the case of a developed and highly populated country such as the United Kindom, study on the patterns and causes of road accidents is of special imperial importance.

In this project, we are aimed at recognize patterns of the road accidents in UK and explore possible ways to reduce the rate of the accidents.

Data collection

The Department of Transport in the United Kingdom has extensively recorded road accidents over the years. For this project, we used data on road accidents occurring in the UK from 2009 to 2014, archived at Kaggle. The dataset has 33 columns, including the location (2 columns for postal co-ordinates, 2 columns for geographic co-ordinates) and the index of the accident.

The following code is used to import data for local directory.

```r
### library import
library(ggplot2)
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
```

```r
library(xts)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```r
library(date)
library(scales)
library(forcats)
library(lubridate)
```

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

## The following object is masked from 'package:base':
##
##     date
```

```r
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year

## The following objects are masked from 'package:xts':
##
##     first, last

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

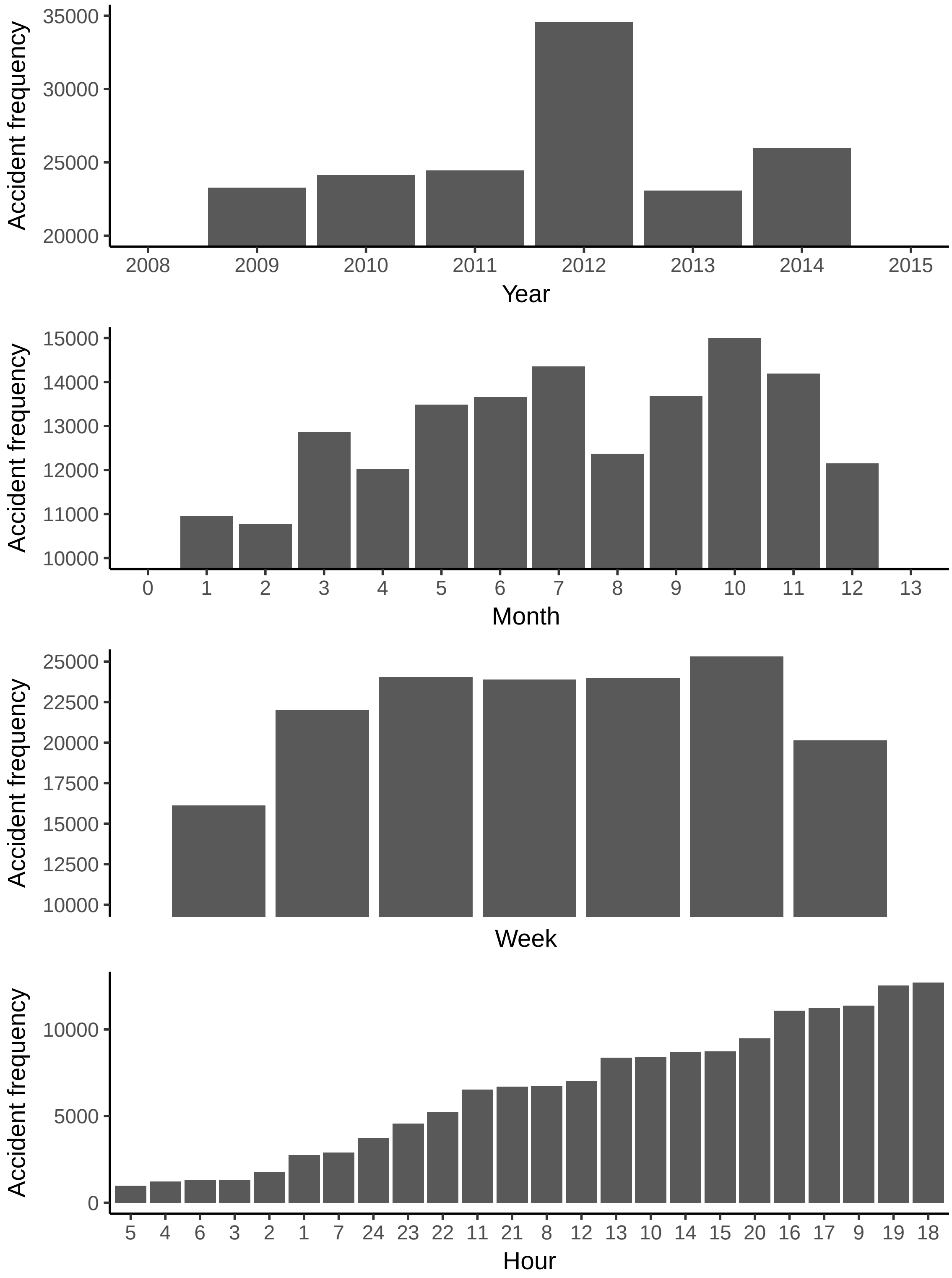
```r
### Data import
path = "F:/Study/courses/DS 5230/project"
data_09_11 <- paste(path, "accident_09_11.csv",sep="/")
data_12_14 <- paste(path, "accident_12_14.csv",sep="/")


London_2009_2011 <- read.csv(data_09_11,head=TRUE, na.strings=c("","NA"))
London_2012_2014 <- read.csv(data_12_14,head=TRUE, na.strings=c("","NA"))

accident <- rbind(London_2009_2011 ,London_2012_2014)

#clean variable names
variable.names <- c("Accident_Index","Location_Easting_OSGR" ,"Location_Northing_OSGR" , "Longitude","La

colnames(accident) <- variable.names
```

Data process

```
########################
# ggplot themes
themeblank_twolines <- theme_bw()+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"))+
  theme(panel.border = element_blank(),
        axis.line.x = element_line(size = 0.5, linetype = "solid", colour = "black"),
        axis.line.y = element_line(size = 0.5, linetype = "solid", colour = "black"))

themeblank_twolines1 <- theme_bw()

themeblank <- theme_bw()+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"))

## basic data

accident <- accident %>%
  mutate(hour=substr(Time,1,2),minute=substr(Time,4,5)) %>%
  mutate(hour=as.numeric(hour)+1,minute=as.numeric(minute),
         Date= as.Date(Date,origin='1900-01-01'),
         month=month(Date))
```

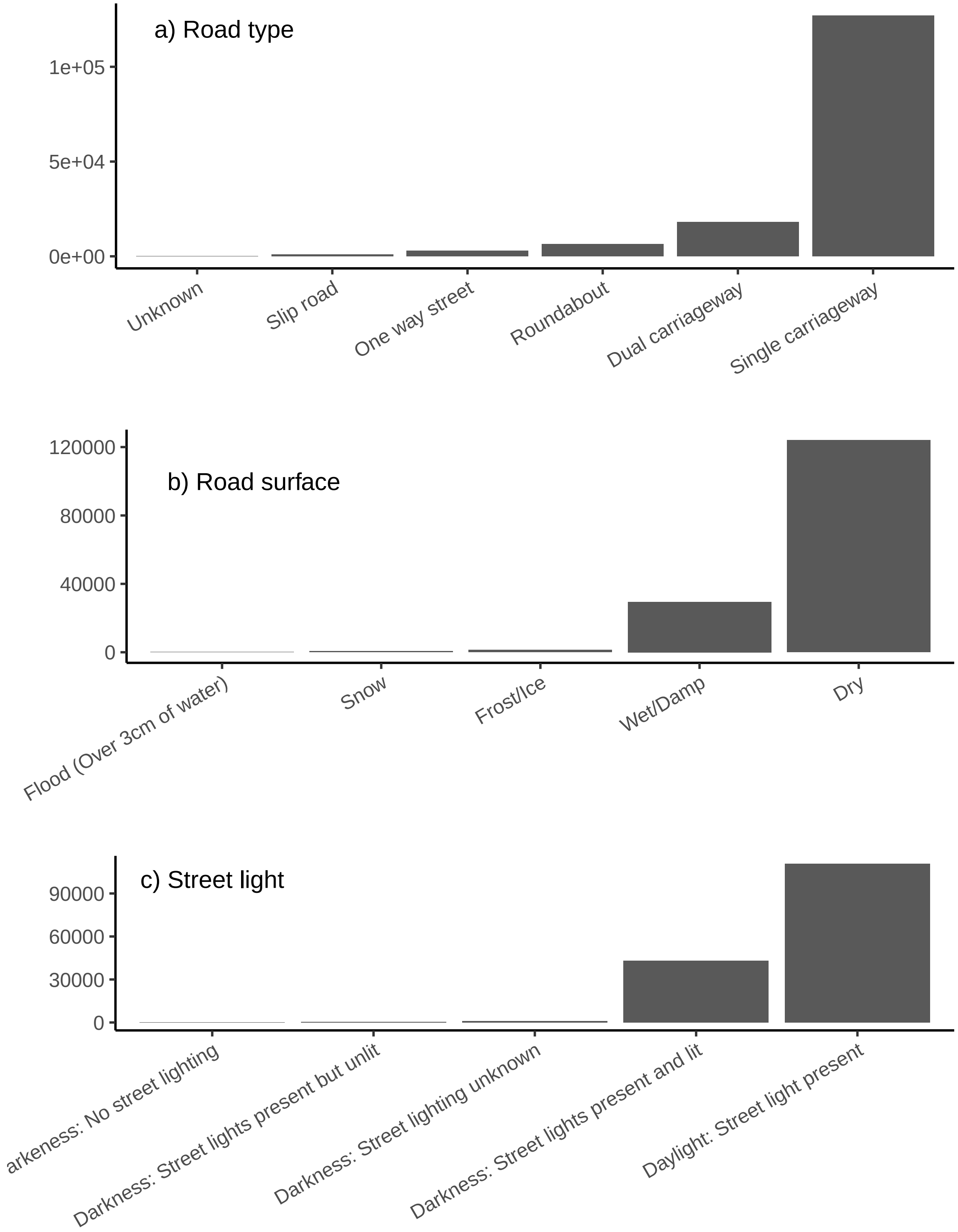
Analysis of time patterns

```
p_hour <- ggplot(accident,aes(fct_rev(fct_infreq(factor(hour)))))+
  geom_bar()+
  labs(x='Hour',y='Accident frequency')+
  themeblank_twolines

p_week <- ggplot(accident,aes(Day_of_Week))+
  geom_bar()+
  labs(x='Week ',y='Accident frequency')+
  #scale_x_discrete(limits=c(2005,2014),breaks=seq(2005,2014,2))+
  scale_y_continuous(limits=c(0,26000),breaks=seq(0,26000,2500))+
  coord_cartesian(ylim = c(10000, 25000)) +
  scale_x_discrete(labels=c("1" = "Mon", "2" = "Tue",'3'='Wed','4'='Thu',
                            "5" = "Fri",'6'='Sat','7'='Sun'))+
  themeblank_twolines

p_month <- ggplot(accident,aes(month))+
  geom_bar()+
  labs(x='Month ',y='Accident frequency')+
  scale_y_continuous(limits=c(0,15000),breaks=seq(0,15000,1000))+
  coord_cartesian(ylim = c(10000, 15000))+
  scale_x_continuous(limits=c(0,13),breaks=seq(0,13,1))+
  themeblank_twolines

p_year <- ggplot(accident,aes(Year))+
  geom_bar()+
  labs(x='Year ',y='Accident frequency')+
  scale_y_continuous(limits=c(0,35000),breaks=seq(0,35000,5000))+
  coord_cartesian(ylim = c(20000, 35000))+
  scale_x_continuous(limits=c(2008,2015),breaks=seq(2008,2015,1))+
  themeblank_twolines

grid.arrange(p_year,p_month,p_week,p_hour,ncol=1)
```



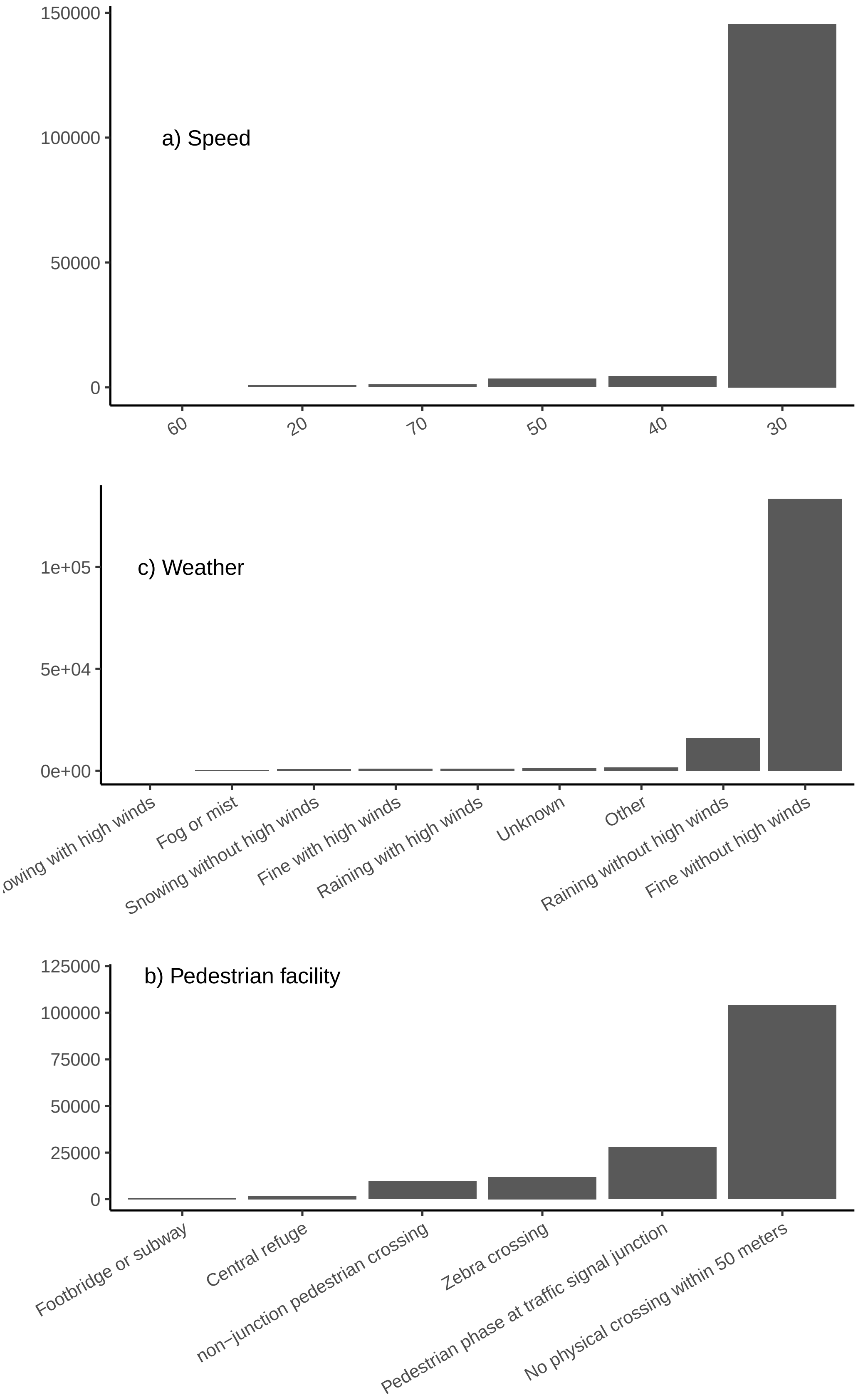
Analysis of main variable impact on road accident frequencies

```
####################################################
#### Analysis of the frequencies of the main covariates

p_road_type <- ggplot(accident,aes(fct_rev(fct_infreq(factor(Road_Type)))))+
  geom_bar()+
  themeblank_twolines+
  labs(x="",y='')+
  annotate("text", x = 1.2, y = 120000, label = "a) Road type")+
  theme(axis.text.x = element_text(angle = 30, hjust = 1))


p_road_surface <- ggplot(subset(accident,Road_Surface_Conditions !='NA'),
                         aes(fct_rev(fct_infreq(factor(Road_Surface_Conditions)))))+
  geom_bar()+
  labs(x="",y='')+
  annotate("text", x = 1.2, y = 100000, label = "b) Road surface")+
  themeblank_twolines+
  theme(axis.text.x = element_text(angle = 30, hjust = 1))

p_light <- ggplot(accident,aes(fct_rev(fct_infreq(factor(Light_Conditions)))))+
  geom_bar()+
  labs(x="",y='')+
  annotate("text", x = 1, y = 100000, label = "c) Street light")+
  themeblank_twolines+
  theme(axis.text.x = element_text(angle = 30, hjust = 1))

grid.arrange(p_road_type,p_road_surface,p_light,ncol=1)
```



Analysis of drive speed, weather and pedestrian plot

```r
##### drive speed, weather and pedestrian plot

p_speed <- ggplot(subset(accident,Speed_limit!='NA'),
                  aes(fct_rev(fct_infreq(factor(Speed_limit)))))+
  geom_bar()+
  labs(x="",y='')+
  annotate("text", x = 1.2, y = 100000, label = "a) Speed")+
  themeblank_twolines+
  theme(axis.text.x = element_text(angle = 30, hjust = 1))

p_pedestrian_human <- ggplot(subset(accident,Pedestrian_Crossing.Physical_Facilities!='NA'),
                             aes(fct_rev(fct_infreq(factor(Pedestrian_Crossing.Physical_Facilities)))))
  geom_bar()+
  labs(x="",y='')+
  annotate("text", x = 1.5, y = 120000, label = "b) Pedestrian facility")+
  themeblank_twolines+
  theme(axis.text.x = element_text(angle = 30, hjust = 1))


p_weather <- ggplot(subset(accident,Weather_Conditions !='NA'),
                 aes(fct_rev(fct_infreq(factor(Weather_Conditions)))))+
  geom_bar()+
  labs(x="",y='')+
  annotate("text", x = 1.5, y = 100000, label = "c) Weather")+
  themeblank_twolines+
  theme(axis.text.x = element_text(angle = 30, hjust = 1))

grid.arrange(p_speed,p_weather,p_pedestrian_human,ncol=1)
```


a) Speed
150000
100000
50000
0
60
20
70
50
40
30
c) Weather
1e+05
5e+04
0e+00
owing with high winds
Fog or mist
Snowing without high winds
Fine with high winds
Raining with high winds
Unknown
Other
Raining without high winds
Fine without high winds
b) Pedestrian facility
125000
100000
75000
50000
25000
0
Footbridge or subway
Central refuge
non−junction pedestrian crossing
Zebra crossing
Pedestrian phase at traffic signal junction
No physical crossing within 50 meters

Prepare the data for PCA and cluster analysis

Through the above analysis, several key variables are selected, including: Latitude, Longitude, Day_of_Week,Road_Type, Speed_limit, Light_Conditions,Weather_Conditions,Road_Surface_Conditions, hour, and month. We perform the factor analysis to evaluate the score of each of these variable's contribution to the occurrence of road accident frequencies.

Light conditions

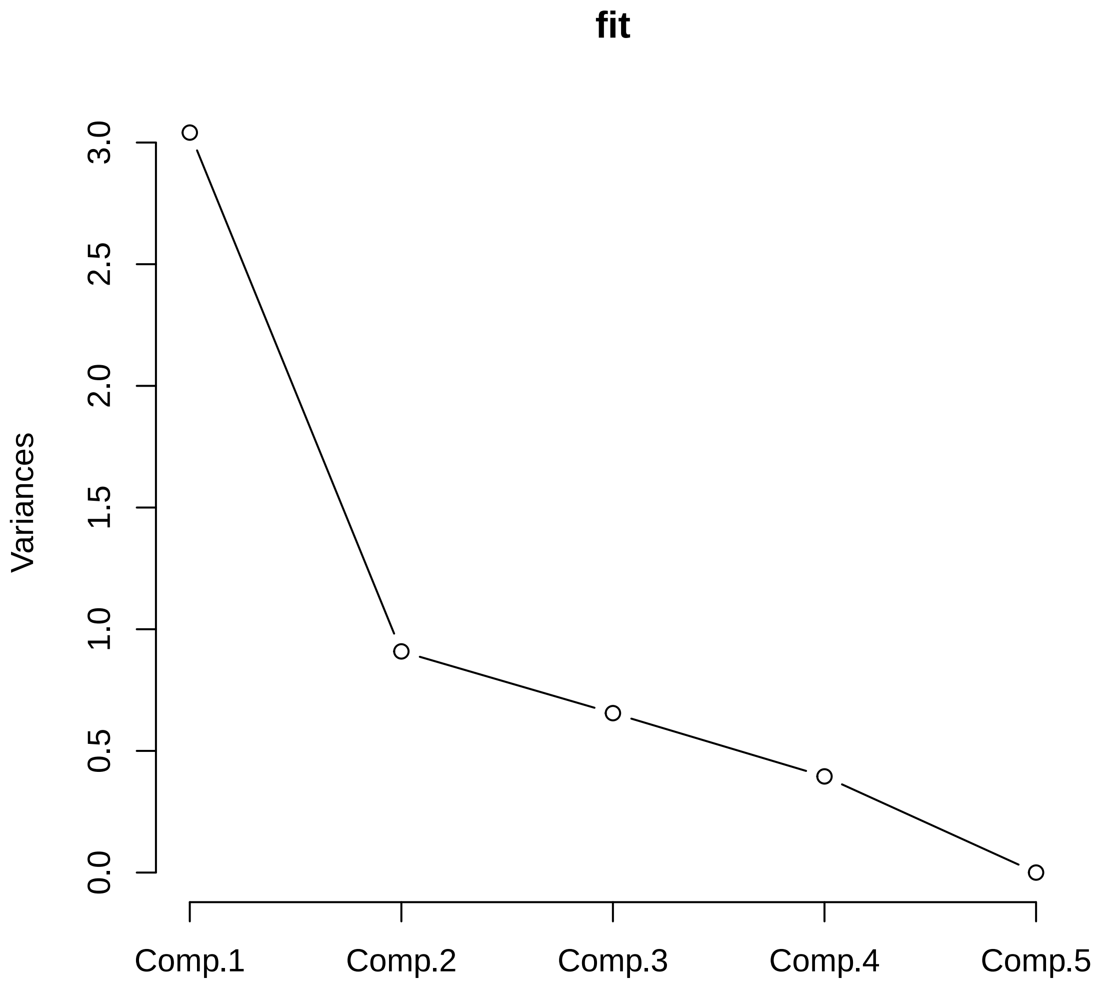
```
variables <- c('Latitude', 'Longitude', 'Day_of_Week','Road_Type','Speed_limit',
               'Light_Conditions','Weather_Conditions','Road_Surface_Conditions', 'hour','month','Year'

accident1 <- accident[,variables]

accident1 <-  accident1[!rowSums((is.na(accident1))),]

accident1 <- accident1 %>%
  mutate(date2=as.Date(paste(Year,month,'01',sep='/')))

Daylight1 <- accident1 %>%
  filter(Light_Conditions==unique(accident$Light_Conditions)[1]) %>%
  group_by(date2) %>%
  summarise(light1=n())

Daylight2 <- accident1 %>%
  filter(Light_Conditions==unique(accident$Light_Conditions)[2]) %>%
  group_by(date2) %>%
  summarise(light2=n())

Daylight3 <- accident1 %>%
  filter(Light_Conditions==unique(accident$Light_Conditions)[3]) %>%
  group_by(date2) %>%
  summarise(light3=n())

Daylight4 <- accident1 %>%
  filter(Light_Conditions==unique(accident$Light_Conditions)[4]) %>%
  group_by(date2) %>%
  summarise(light4=n())

Daylight5 <- accident1 %>%
  filter(Light_Conditions==unique(accident$Light_Conditions)[5]) %>%
  group_by(date2) %>%
  summarise(light5=n())

light <- Daylight1 %>%
  left_join(Daylight2,by=c('date2')) %>%
  left_join(Daylight3,by=c('date2')) %>%
  left_join(Daylight4,by=c('date2')) %>%
  left_join(Daylight5,by=c('date2'))

light <- light %>%
  mutate(incidence=light1+light2+light3+light4+light5) %>%
  mutate(light1r=light1/incidence,light2r=light2/incidence,light3r=light3/incidence,light4r=light4/inci


colnames(light)
```

```
##  [1] "date2"     "light1"    "light2"    "light3"    "light4"
##  [6] "light5"    "incidence" "light1r"   "light2r"   "light3r"
## [11] "light4r"   "light5r"
```

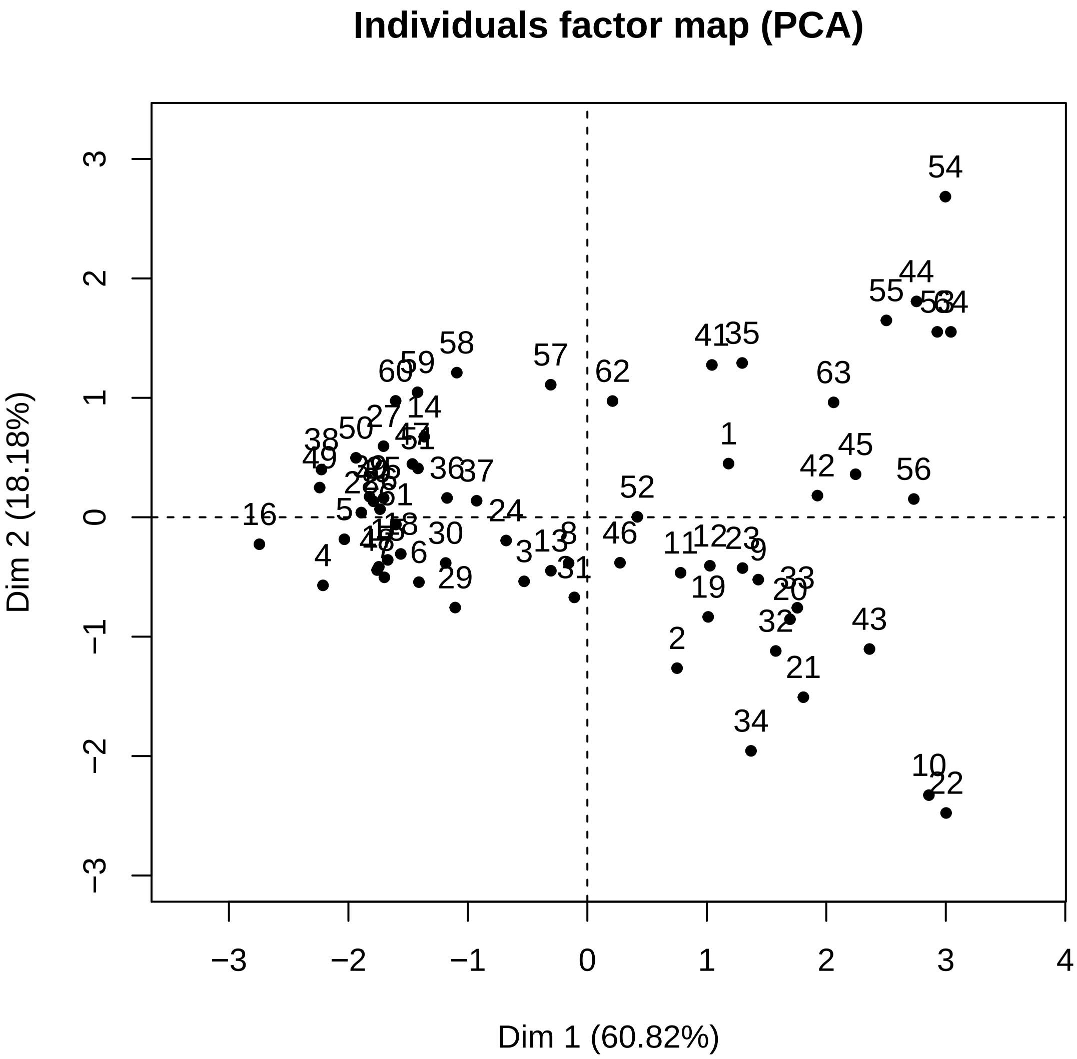
```
light <- light[,c(1,7,8:12)]

colnames(light) <- c('date2','incidence','Daylight: Street light present','Darkness: Street lights prese
                     'Darkness: Street lighting unknown','Darkness: Street lights present but unlit',
                     'Darkeness: No street lighting' )

# standardize data
light <-  light[!rowSums((is.na(light))),]
light_stan <- as.data.frame(scale(light[,c(3:7)]))

# Pricipal Components Analysis
# entering raw data and extracting PCs
# from the correlation matrix
fit <- princomp(light_stan, cor=TRUE)
summary(fit) # print variance accounted for
```

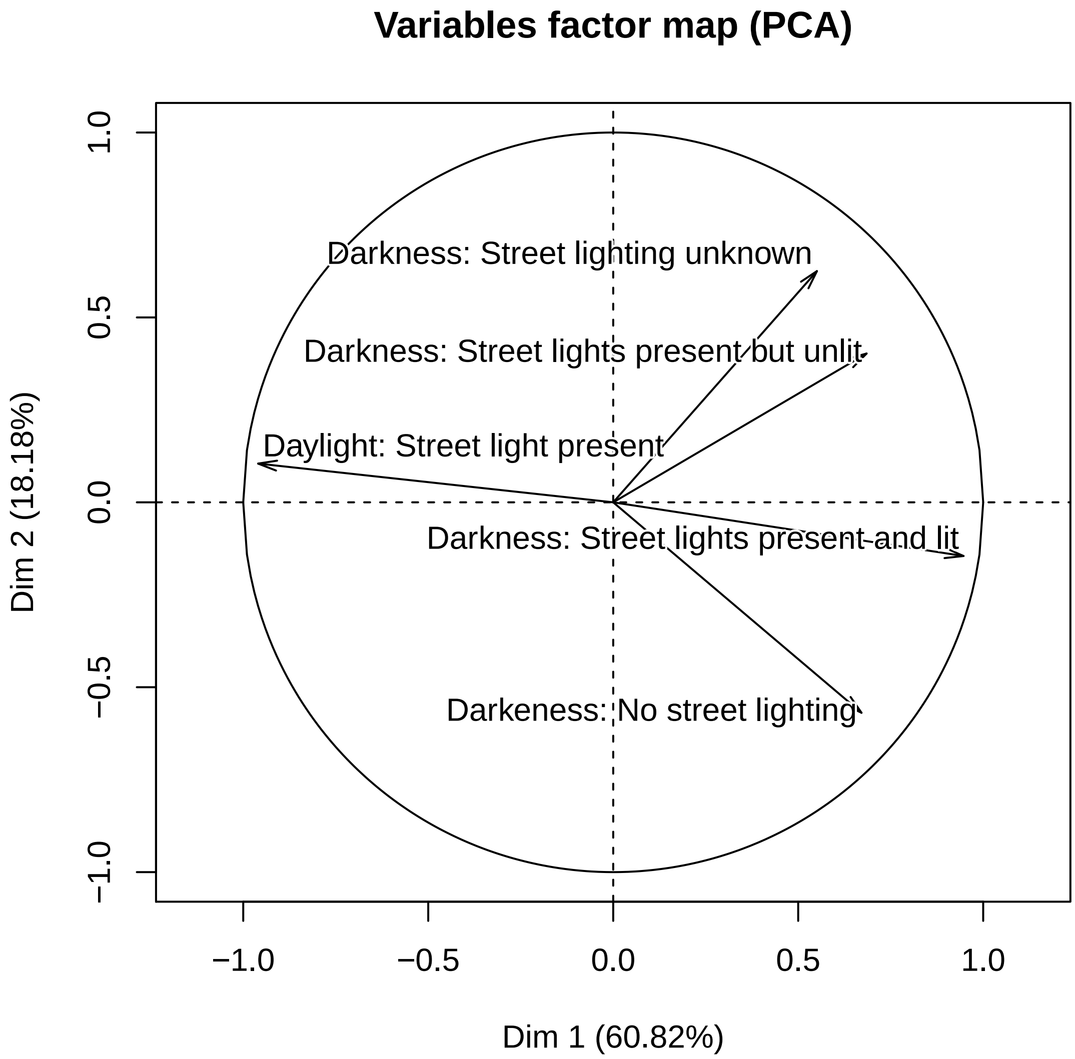
```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4 Comp.5
## Standard deviation     1.7438165 0.9533253 0.8094059 0.62859926      0
## Proportion of Variance 0.6081792 0.1817658 0.1310276 0.07902741      0
## Cumulative Proportion  0.6081792 0.7899450 0.9209726 1.00000000      1
```

```
#loadings(fit) # pc loadings
plot(fit,type="lines") # scree plot
```



```
# PCA Variable Factor Map
library(FactoMineR)

result <- PCA(light_stan) # graphs generated automatically
```


Individuals factor map (PCA)
Dim 1 (60.82%)
Dim 2 (18.18%)
−3
−2
−1
0
1
2
3
4
54
44
55
53
64
41
35
63
58
57
62
60
59
14
27
50
47
51
38
49
1
45
42
56
36
37
52
16
5
24
4
30
8
46
11
12
23
9
3
13
31
6
29
19
33
20
32
43
2
21
34
10
22



Here perform a hierarchical cluster analyis

```
variable_chose <- c("Location_Easting_OSGR" ,"Location_Northing_OSGR" ,
                  "Accident_Severity" ,
                     "Number_of_Vehicles",  "Day_of_Week",
                    "X1st_Road_Class",     "Road_Type"  ,"Speed_limit" ,'Number_of_Casualties',
                     "Junction_Detail"  , "Junction_Control"  , "X2nd_Road_Class",
                     "Light_Conditions" ,"Weather_Conditions"  ,"Road_Surface_Conditions" ,
                     "Special_Conditions_at_Site", "Carriageway_Hazards")


a14 <- accident %>%
filter(Year==2014)

# clean and group some obvious very skewed data
```

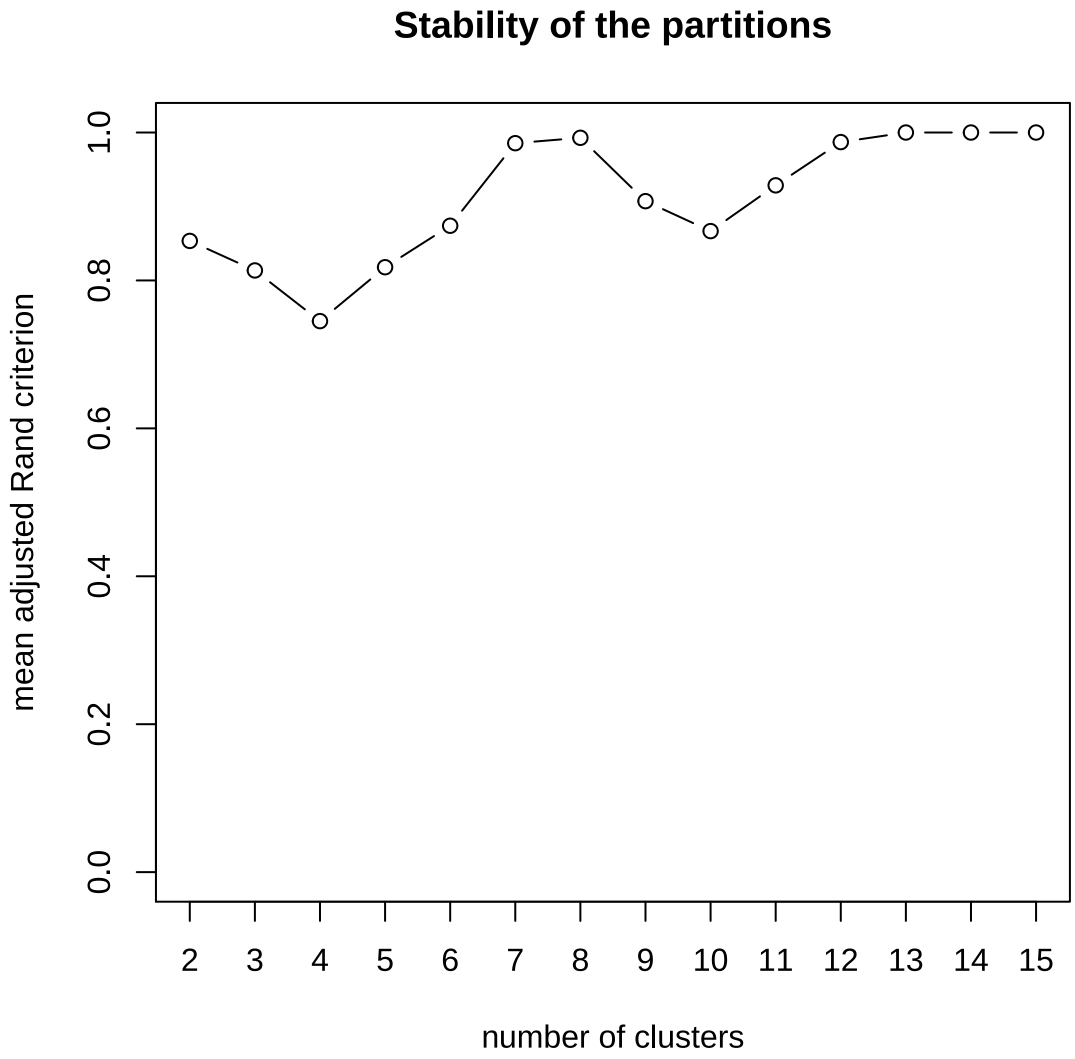
```
a14 <- a14[,variable_chose] %>%
  mutate(casualties = ifelse(Number_of_Casualties==1,Number_of_Casualties,
                             ifelse(Number_of_Casualties==2,Number_of_Casualties,3)))

a14_data <- a14[,-which(names(a14) %in% c('casualties'))] # remove redundant variable


require('ClustOfVar')
```

```
## Loading required package: ClustOfVar
```

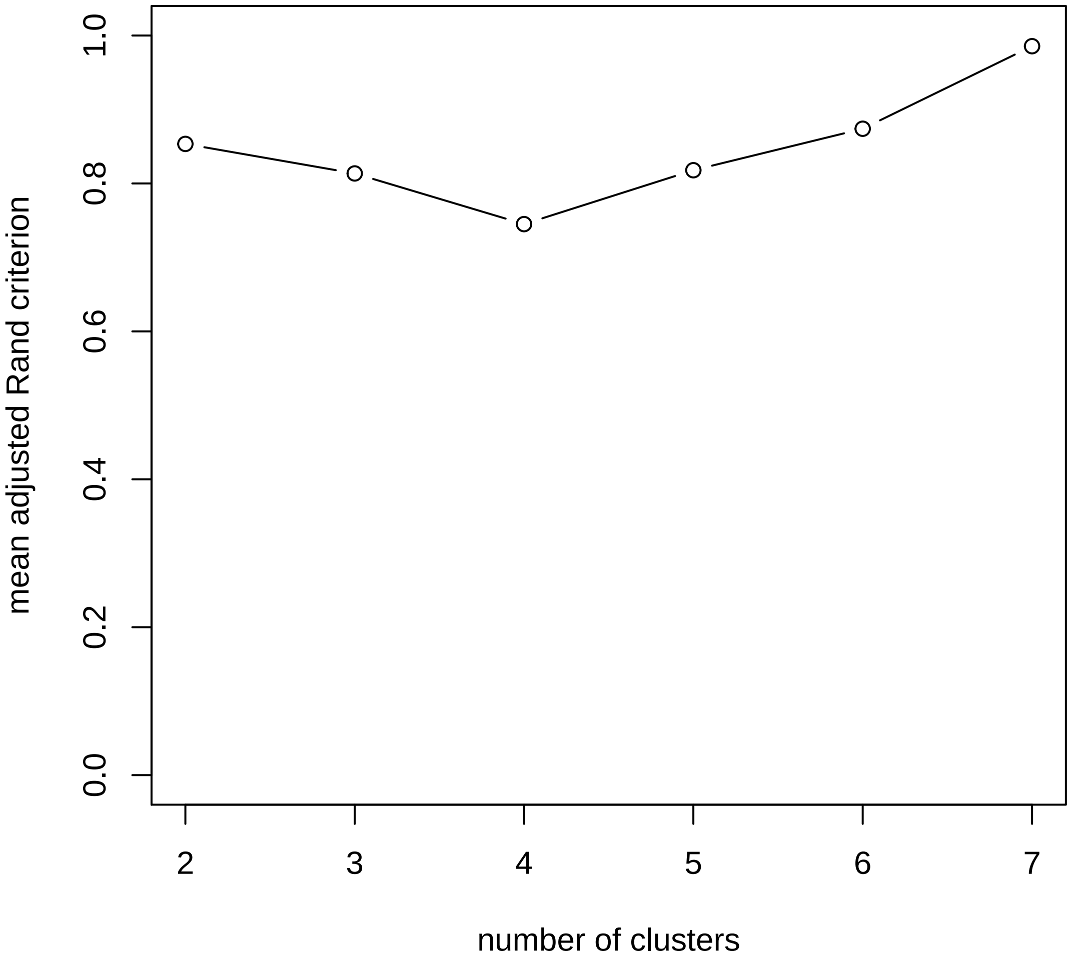
```
library(ClustOfVar)
X.quanti <-PCAmixdata::splitmix(a14_data)$X.quanti # qunatitave matrix

X.quali <- PCAmixdata::splitmix(a14_data)$X.quali # qualitative matrix
```

Evaluate the number of clusters on the mean adjusted R values

```
tree <-  hclustvar(X.quanti,X.quali)

stab <- stability(tree,B=40)
```

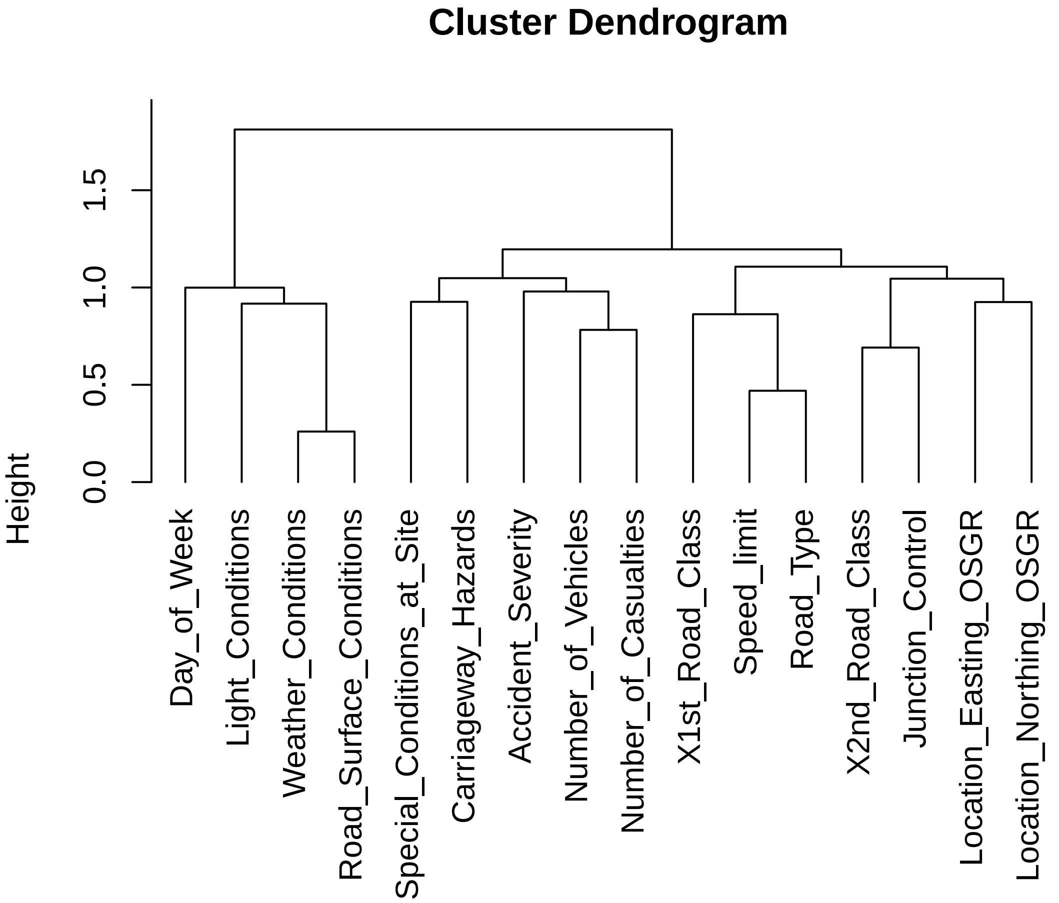


```
plot(stab,nmax=7)
```



Visualize the calculated hierarchical clusters of the variables.

```
plot(tree)
```



The result generate three clusters of the variables, namely, cluster 1: Day of week, street light condition, weather condition, and road surface condition; cluster 2: special conditions at site, carriageway hazards,accident severity, number of vehicles, and number of casualities; and cluster 3: first road class, speed limit, road type, second road class, junction control, location easting OSGR, and location northing OSGR. These three clusters of variables maybe termed as: weather condition, accident condition, and road condition.