UE20CS312 - Data Analytics - Worksheet 1b - Correlation Analysis

PES University

‘SUNDEEP A, Dept. of CSE - PES1UG20CS445’

2022-08-26

Correlation

Road Accidents

Problem 1 (2 points)

Find the total number of accidents in each state for the year 2016 and display your results. Make sure to display all rows while printing the dataframe. Print only the necessary columns. (Hint: use the grep command to help filter out column names).

```
library(ggpubr)
```

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

```
df <- read.csv('road_accidents_india_2016.csv', row.names=1)
accident <- grep("Total.Accidents$",colnames(df),ignore.case = T,value = TRUE)
total_accidents<-data.frame(state.ut=df$State..UT,total_acc=rowSums(df[,c(accident)],na.rm=TRUE))
print.data.frame(total_accidents)
```

```
##             state.ut total_acc
## 0     Andhra Pradesh     24888
## 1  Arunachal Pradesh       249
## 2              Assam      7435
## 3              Bihar      8222
## 4       Chhattisgarh     13580
## 5                Goa      4304
## 6            Gujarat     21859
## 7            Haryana     11234
## 8   Himachal Pradesh      3168
## 9    Jammu & Kashmir      5501
## 10         Jharkhand      4932
## 11         Karnataka     44403
## 12            Kerala     39420
## 13    Madhya Pradesh     53972
## 14       Maharashtra     39878
## 15           Manipur       538
## 16         Meghalaya       620
## 17           Mizoram        83
## 18          Nagaland        75
## 19            Orissa     10532
## 20            Punjab      6952
## 21         Rajasthan     23066
## 22            Sikkim       210
## 23        Tamil Nadu     71431
## 24         Telangana     22811
## 25           Tripura       557
## 26       Uttarakhand      1591
## 27     Uttar Pradesh     35612
## 28       West Bengal     13580
## 29     A & N Islands       238
## 30        Chandigarh       428
## 31      D & N Haveli        70
## 32       Daman & Diu        71
## 33             Delhi      7375
## 34       Lakshadweep         1
## 35        Puducherry      1766
```

Problem 2 (2 points)

Find the (fatality rate $= \frac{\text{total number of deaths}}{\text{total number of accidents}}$) in each state. Find out if there is a significant linear correlation at a significance of $\alpha = 0.05$ between the *fatality rate* of a state and the *mist/foggy rate* (fraction of total accidents that happen in mist/foggy conditions).

Plot the fatality rate against the mist/foggy rate. (Hint: use the `ggscatter` library to plot a scatterplot with the confidence interval of the correlation coefficient).

```
col_death <- grep("Persons.Killed$",colnames(df),ignore.case=T,value=TRUE)
total_accidents$total_deaths <- rowSums(df[,c(col_death)])

total_accidents$fatality_rate <-total_accidents$total_deaths/total_accidents$total_acc

total_accidents$mist_rate <- df$Mist..Foggy...Total.Accidents/total_accidents$total_acc
head(total_accidents[,c("state.ut","total_deaths","fatality_rate","mist_rate")])
```

```
##             state.ut total_deaths fatality_rate  mist_rate
## 0    Andhra Pradesh         8541    0.34317743 0.04222919
## 1 Arunachal Pradesh          149    0.59839357 0.12449799
## 2             Assam         2572    0.34593141 0.06603900
## 3             Bihar         4901    0.59608368 0.21515446
## 4      Chhattisgarh         3908    0.28777614 0.02120766
## 5               Goa          336    0.07806691 0.00000000
```

```
 co_relation_factor<- cor(total_accidents$fatality_rate,total_accidents$mist_rate)
 sprintf("The co-relation factor is : %f",co_relation_factor)
```

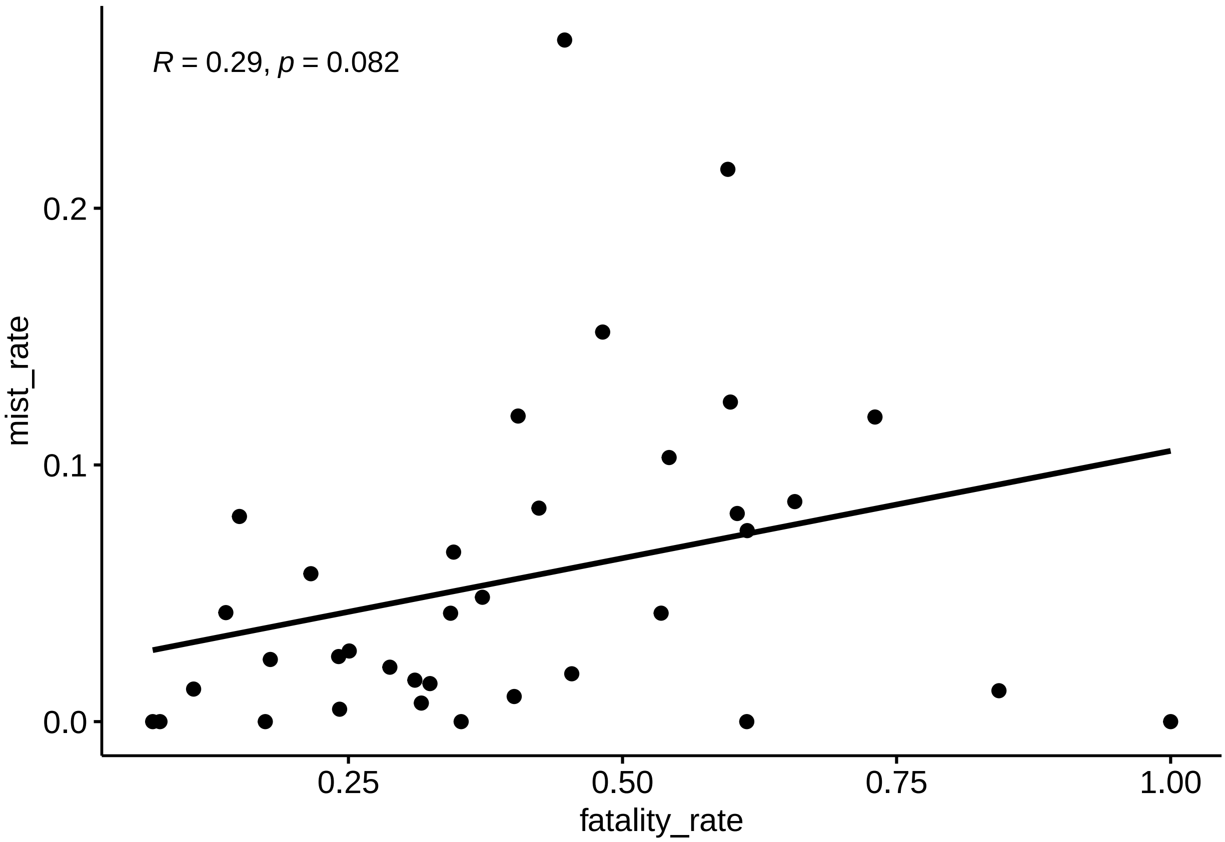
```
## [1] "The co-relation factor is : 0.293516"
```

```
print("Yes, there is a significant co-relation between fatality rate of a state and the mist/foggy rat
```

```
## [1] "Yes, there is a significant co-relation between fatality rate of a state and the mist/foggy rat
```

```
ggscatter(total_accidents, x = "fatality_rate", y = "mist_rate",
   add = "reg.line",
   conf.int.level =0.95,
   cor.coef = TRUE,
   cor.coeff.args = list(method = "pearson")
   )
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Problem 3 (3 points)

Rank the states based on total accidents and total fatalities (give a rank of 1 to the state that has the highest value of a property). You are free to use any tie-breaking method for assigning ranks.

Find the Spearman-Rank correlation coefficient between the two rank columns and determine if there is any statistical significance at a significance level of $\alpha = 0.05$. Also test the hypothesis that the correlation coefficient is at least 0.2.

```
total_accidents$acc_rank <- rank(desc(total_accidents$total_acc),ties.method='random')
total_accidents$death_rank <- rank(desc(total_accidents$total_deaths),ties.method='random')
head(total_accidents[,c("state.ut","death_rank","acc_rank")])
```

```
##             state.ut death_rank acc_rank
## 0     Andhra Pradesh          7        7
## 1 Arunachal Pradesh        28       29
## 2             Assam        18       16
## 3             Bihar        13       15
## 4      Chhattisgarh        16       12
## 5               Goa        23       21
```

```
spearman_coefficient <- cor(total_accidents$acc_rank,total_accidents$death_rank,method="spearman")
sprintf("The spearman coefficient is : %f",spearman_coefficient)
```

```
## [1] "The spearman coefficient is : 0.958044"
```

```
print("There is a positive co-relation between death rank and accident rank")
```

```
## [1] "There is a positive co-relation between death rank and accident rank"
```

```
degrees <- nrow(total_accidents)-2
sprintf("The no of degrees is : %d",degrees)
```

```
## [1] "The no of degrees is : 34"
```

```
t_stat<-(spearman_coefficient-0.2)/sqrt(1-spearman_coefficient*spearman_coefficient)/(nrow(total_accider
t_stat
```

```
## [1] 0.07778679
```

```
2*pt(q=t_stat,df=degrees,lower.tail=FALSE)
```

```
## [1] 0.9384536
```

Problem 4 (1.5 points)

Convert the column `Hail.Sleet...Total.Accidents` to a binary column as follows. If a hail/sleet accident has occurred in a state, give that state a value of 1. Otherwise, give it a value of 0. Once converted, find out if there is a significant correlation between the `hail_accident_occcur` binary column created and the number of rainy total accidents for every state.

Calculate the point bi-serial correlation coefficient between the two columns. (Hint: it is equivalent to calculating the Pearson correlation between a continuous and a dichotomous variable.).

```
hail_acc <- ifelse(df$Hail.Sleet...Total.Accidents>0, 1, 0)
rs <- cor.test(df$Rainy...Total.Accidents,hail_acc)
rs
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$Rainy...Total.Accidents and hail_acc
## t = 0.84232, df = 34, p-value = 0.4055
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1947090  0.4503544
## sample estimates:
##       cor
## 0.1429725
```

```
print("There is a positive co-relation between Total accidents due ot Hail and Total number of accients
```

```
## [1] "There is a positive co-relation between Total accidents due ot Hail and Total number of accients
```

```
sprintf("The point bi-serial corelation coefficient between the two columns is : %f" ,cor(df$Rainy...To
```

```
## [1] "The point bi-serial corelation coefficient between the two columns is : 0.142973"
```

Problem 5 (1.5 points)

create a binary column to represent whether a dust storm accident has occurred in a state (1 = occurred, 0 = not occurred). Convert the two columns into a contingency table. Calculate the phi coefficient of the two tables.

```
library("psych")
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
dust_storm <- ifelse(df$Dust.Storm...Total.Accidents>0, 1, 0)

conti_table <- table(hail_acc,dust_storm)
conti_table
```

```
##         dust_storm
## hail_acc  0  1
##        0 14  5
##        1  2 15
```

```
Phi_coefficient<- phi(conti_table)
sprintf("The phi coefficient is : %f",Phi_coefficient)
```

```
## [1] "The phi coefficient is : 0.620000"
```