

PES University, Bangalore Established under the Karnataka Act
No. 16 of 2013

UE21CS342AA2 - Data Analytics
Worksheet 1b : Correlation Analysis

Richa Shahi - shahiricha2412@gmail.com , Abhay K Iyengar – abzee2002@gmail.com

Correlation

Correlation is a measure of the strength and direction of linear relationship between two random variables in other words it is a measure of the association between two variables. This association is measured using the correlation coefficients. Correlation coefficients help us choose features in model building and leverage the association relationship between two variables.

There are different types of correlation coefficients, based on the nature of the data being compared:

- Between two continuous (interval, ratio) random variables - Pearson's Product Moment Correlation Coefficient
- Between two ordinal random variables - Spearman-Rank Correlation Coefficient
- Between a continuous RV and a dichotomous RV - Point Bi-Serial Correlation Coefficient
- Between two binary random variables - Phi Coefficient

Road Accidents

India is the world's second-most populous country with a population of around 1.2 billion people (as of July 2022). Roads are a very important mode of transport in India, spanning over 6.2 million kilometers of length, making it the country with the second-largest road network, after the United States of America. (Source: [Wikipedia](#)).

With India trying to modernize its road infrastructure, there is still the problem of frequent road accidents. Road accidents in India is a major cause of death and injury. The NCRB (National Crime Records Bureau) of India collects detailed data on traffic accidents and collisions annually. Please use the dataset provided for analysis that contains road accident data in India from 2016. You can download the dataset from [here](#).

Data Dictionary

S. No.: Serial number State/ UT: name of state/union territory in India

Fine/Clear - Total Accidents: total accidents per state/UT in Fine/Clear weather conditions

Fine/Clear - Persons Killed: total fatalities per state/UT in Fine/Clear weather conditions

Fine/Clear - Persons Injured: total injured people per state/UT in Fine/Clear weather conditions

Mist/Foggy - Total Accidents: total accidents per state/UT in Mist/Foggy weather conditions

Mist/ Foggy - Persons Killed: total fatalities perstate/UT in Mist/Foggy weather conditions

Mist/ Foggy - Persons Injured: total injured people per state/UT in Mist/Foggy weather conditions

Cloudy - Total Accidents: total accidents per state/UT in Cloudy weather conditions

Cloudy - Persons Killed: total fatalities per state/UT in Cloudy weather conditions

Cloudy - Persons Injured: total injured people per state/UT in Cloudy weather conditions

Rainy - Total Accidents: total accidents per state/UT in Rainy weather conditions

Rainy - Persons Killed: total fatalities per state/UT in Rainy weather conditions

Rainy - Persons Injured: total injured people per state/UT in Rainy weather conditions

Snowfall - Total Accidents: total accidents per state/UT in Snowfall weather conditions

Snowfall - Persons Killed: total fatalities per state/UT in Snowfall weather conditions

Snowfall - Persons Injured: total injured people per state/UT in Snowfall weather conditions

Hail/Sleet - Total Accidents: total accidents per state/UT in Hail/Sleet weather conditions

Hail/Sleet - Persons Killed: total fatalities per state/UT in Hail/Sleet weather conditions

Hail/Sleet - Persons Injured: total injured people per state/UT in Hail/Sleet weather conditions

Dust Storm - Total Accidents: total accidents per state/UT in Dust Storm weather conditions

Dust Storm - Persons Killed: total fatalities per state/UT in Dust Storm weather conditions

Dust Storm - Persons Injured: total injured people per state/UT in Dust Storm weather conditions

Others - Total Accidents: total accidents per state/UT in Other weather conditions

Others - Persons Killed: total fatalities per state/UT in Other weather conditions

Others - Persons Injured: total injured people per state/UT in Other weather conditions

Problems

- *Problem 1*
- *Problem 2*
- *Problem 3*
- *Problem 4*
- *Problem 5*
- *Problem 6*
- *Problem 7*

Problem 1

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).

Problem 2

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).

Correlation between two continuous RVs: Pearson's correlation coefficient. Pearson's correlation coefficient between two RVs x and y is given by:

$$\rho = \frac{\text{Covariance}(x, y)}{\sigma_x \cdot \sigma_y}$$

where:

ρ represents the Pearson's correlation coefficient

$\text{Covariance}(x, y)$ is the covariance between x and y

σ_x is the standard deviation of x

σ_y is the standard deviation of y.

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).

Plot the fatality rate and mist/foggy rate (see [this](#) and [this](#) for R plot customization).

Problem 3

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.

$$t = \frac{r_s - \rho_s}{\sqrt{\frac{1-r_s^2}{n-2}}}$$

where:

- t represents the t-statistic,
- r_s is the Spearman-Rank correlation coefficient,
- ρ_s value of the population correlation coefficient being tested against,
- n is the number of data points in the sample.

Problem 4

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_occur 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. You could also use the ltm package's biserial.cor function).

Problem 5

Similar to in Problem 4, 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.

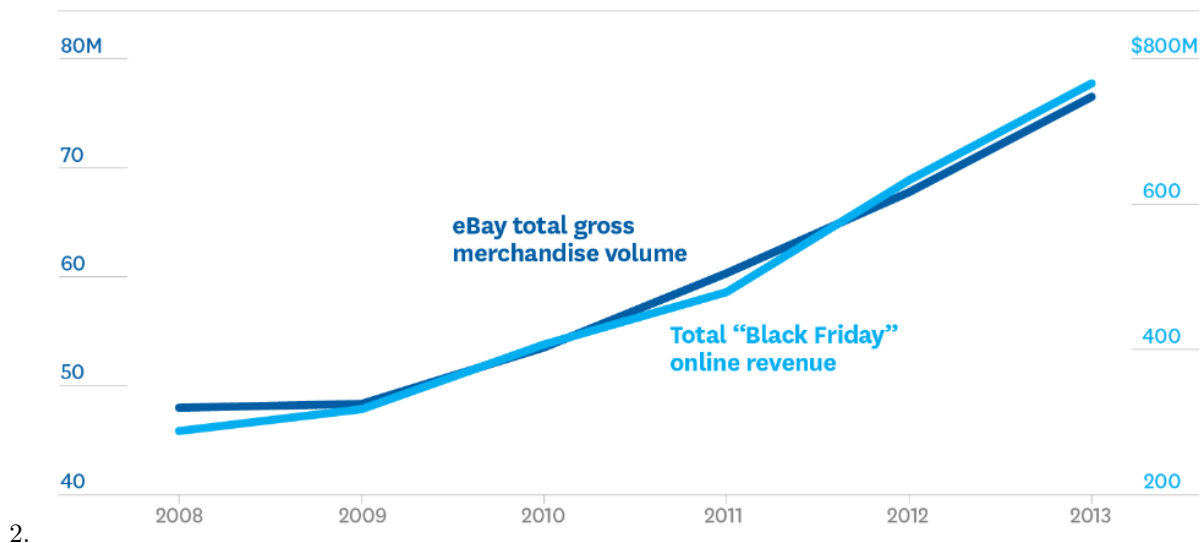
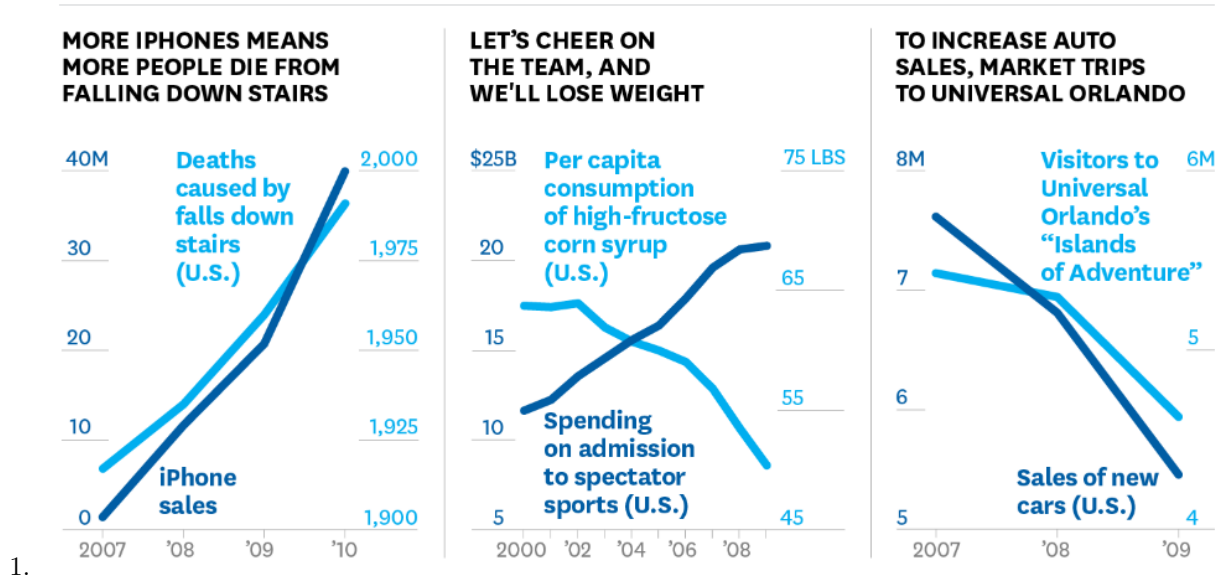
Calculate the phi coefficient of the two tables. (Hint: use the psych package).

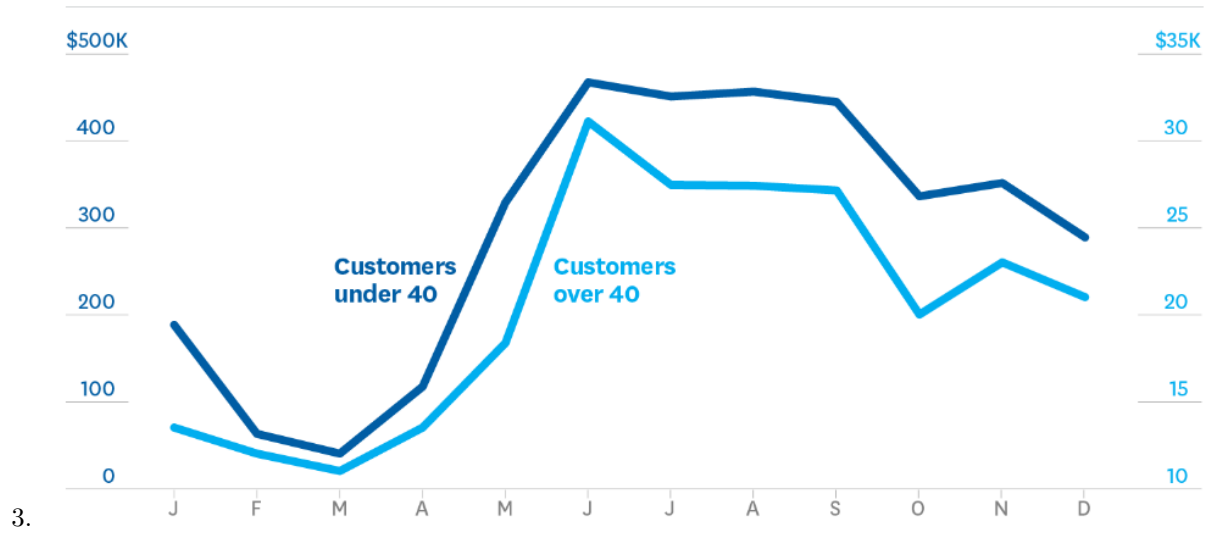
Problem 6

Read about correlation on this [website](#) and analyze the effect of sample size on correlation coefficients and spurious correlation. Are correlation coefficients affected by outliers ?

Problem 7

Look at these plots and answer What problems do they have? How do they affect correlation analysis?





The y-axis is the monthly revenue of a company and x-axis represents month.