DATA ANALYTICS

Unit 2:Correlation Analysis

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 2:Correlation Analysis

Mamatha H R
Department of Computer Science and Engineering

Introduction to Correlation

- *Correlation is a statistical measure of an association relationship that exists between two random variables.*
- *Correlation is not necessarily a causal relationship.*
- *Correlation is important in analytics since it helps to identify variables that may be used in the model building and also useful for identifying issues such as multi-collinearity that can destabilize regression-based models.*

Introduction to Correlation

- Correlation is a measure of the strength and direction of relationship that exists between two random variables and is measured using correlation coefficient.

- In others words, correlation is a measure of association between two variables. Correlation can assist the data scientists to choose the variables for model building that is used for solving an analytics problem

DATA ANALYTICS

In today's class…

- Correlation between two continuous random variables (ratio or interval scale)
- Correlation between two ordinal variables
- Correlation between a continuous random variable and a dichotomous (binary) random variable
- Correlation between two binary random variables

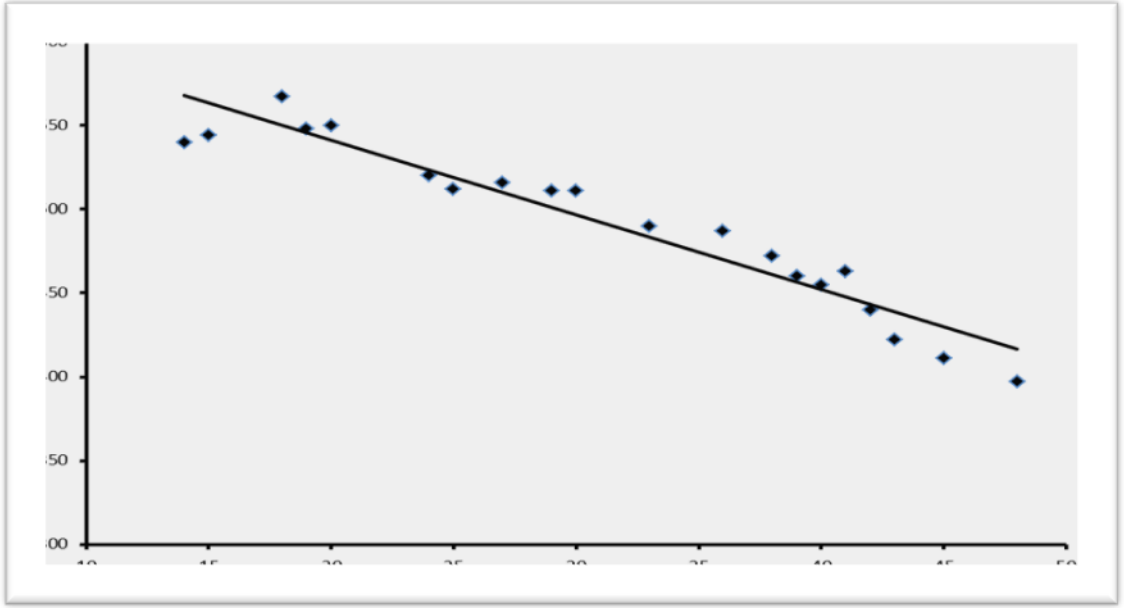
Pearson Correlation Coefficient

Pearson product moment correlation (in short Pearson correlation) is used for measuring the strength and direction of the linear relationship between two continuous random variables *X* and *Y*.

Data on age and average call duration (in seconds)

Age	14	15	18	19	20	24	25	27	29	30
Call Duration	540	544	567	548	550	520	512	516	511	511
Age	33	36	38	39	40	41	42	43	45	48
Call Duration	490	487	472	460	455	463	440	422	411	397

In the figure, we can see that the average call duration (Y) decreases as the age of the customer (X) increases. We can measure the strength of the linear association relationship using a numerical measure called correlation coefficient.



Association relationship between age and average call duration

Calculation of Pearson Product Moment Correlation Coefficient

- Pearson product moment correlation is used when we are interested in finding linear relationship between two continuous random variables (that is, the variable should be either of ratio or interval scale).

- The range of two variables can be different, thus we need to standardize the variables which can be used for measuring the correlation between two variables.

Calculation of Pearson Product Moment Correlation Coefficient

- Let X_i be different values of the variable X and Y_i be different values of Y. Then the standardized values of X and Y are given by

$$Z_X = \left(\frac{X_i - \bar{X}}{\sigma_X} \right) \qquad Z_Y = \left(\frac{Y_i - \bar{Y}}{\sigma_Y} \right)$$

- Where $\bar{X}$ and $\bar{Y}$ are mean values of random variables X and Y; σ_X and σ_Y are the corresponding standard deviations. The Pearson's correlation coefficient is given by

$$r = \frac{\sum_{i=1}^{n} Z_X Z_Y}{n} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right) \times \left(Y_i - \bar{Y} \right)}{n \sigma_X \sigma_Y}$$

Where n is the number of cases in the sample. The formula in above Eq.is also frequently used to account for the degrees of freedom and recommended when the standard deviation is calculated from sample

Calculation of Pearson Product Moment Correlation Coefficient

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\times\left(Y_i - \bar{Y}\right)}{(n-1)S_X S_Y}$$

However, here we will be using the formula in above Eq.For large samples, the correlation coefficients calculated using Eqs. will converge.

Where S_X and S_Y are the standard deviation of random variables X and Y calculate from the sample. We can note the following properties from Eq.

$$r = \frac{\sum_{i=1}^{n} Z_X Z_Y}{n} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\times\left(Y_i - \bar{Y}\right)}{n\sigma_X \sigma_Y}$$

Properties

- Whenever the value of X_i is greater than mean and if the corresponding value of Y_i is also greater than mean, then the numerator in equation will be positive.

- Whenever the value of X_i is lesser than mean and if the corresponding value of Y_i is also lesser than mean, then the numerator in equation will be positive.

- Whenever the value of X_i is lesser than mean (or greater than mean) and the corresponding value of Y_i is greater than mean (or lesser than mean), then the numerator in equation will be negative.

It is possible that we may have combinations of three cases listed above in a data set. Thus the numerator in Eq.is likely to be positive, negative, or zero.

The value of Pearson's correlation coefficient lies between –1 and +1. Equation is mathematically equivalent to below Eqs.

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} \times \sqrt{n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2}}$$

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where Cov(*X*,*Y*) is the covariance between random variables *X* and *Y* and is given by

$$\mathrm{Cov}(X, Y) = E\left(\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)\right)$$

Properties of Pearson Correlation Coefficient

- The value of correlation coefficient lies between –1 and +1. High absolute value of r, $|r|$, indicates strong relationship between the two variables.
- Positive value of r indicates positive correlation (as value of X increases, the value of Y also increases) and negative value of r indicates negative correlation (as the value of X increases, the value of Y decreases).
- The sign of correlation coefficient is same as the sign of covariance between the two random variables.

Properties of Pearson Correlation Coefficient

- Assume that the value of Pearson correlation coefficient between X and Y is r. Let Z_1 and Z_2 be the linear combinations of X and Y ($Z_1 = A + BX$ and $Z_2 = C + DY$). Then the correlation coefficient between Z_1 and Z_2 will be r when the signs of B and D are same (both are positive or negative) and $-r$ when the signs of B and D are opposite.
- Mathematically, square of correlation coefficient is equal to the co-efficient of determination (R^2) of the linear regression model, that is $r^2 = R^2$.
- Pearson correlation coefficient value may be zero even when there is a strong non-linear relationship between variables X and Y (Reed, 1917). Thus low correlation coefficient value cannot be taken as an evidence of no relationship.

Example

The average share prices of two companies over the past 12 months are shown in Table . Calculate the Pearson correlation coefficient.

X	Y
274.58	219.50
287.96	242.92
290.35	245.90
320.07	256.80
317.40	240.60
319.53	245.23
301.52	232.09
271.75	222.65
323.65	231.74
259.80	214.43
263.02	201.86
286.03	204.23

The average values are $\bar{X} = 292.9717$ and $\bar{Y} = 229.8292$

The following equation is used for calculating the correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

DATA ANALYTICS

Calculation of correlation coefficient

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
274.58	219.50	–18.39	–10.33	189.97	338.25	106.6917
287.96	242.92	–5.01	13.09	–65.61	25.12	171.3699
290.35	245.90	–2.62	16.07	–42.13	6.87	258.2717
320.07	256.80	27.10	26.97	730.86	734.32	727.4259
317.40	240.60	24.43	10.77	263.11	596.74	116.0109
319.53	245.23	26.56	15.40	409.02	705.35	237.1857
301.52	232.09	8.55	2.26	19.33	73.07	5.111367
271.75	222.65	–21.22	–7.18	152.35	450.36	51.54043
323.65	231.74	30.68	1.91	58.62	941.16	3.651284
259.80	214.43	–33.17	–15.40	510.82	1100.36	237.1343
263.02	201.86	–29.95	–27.97	837.72	897.10	782.2743
286.03	204.23	–6.94	–25.60	177.70	48.19	655.3173
Sum				3241.77	5916.89	3351.98

From Table , we have

$$\sum_{i=1}^{12}(X_i - \bar{X})(Y_i - \bar{Y}) = 3241.77$$

$$\sum_{i=1}^{12}(X_i - \bar{X})^2 = 5916.89$$

Correlation coefficient r = $\sum_{i=1}^{12}(Y_i - \bar{Y})^2 = 3351.98$

$$\frac{3241.77}{\sqrt{5916.89} \times \sqrt{3351.98}} = 0.7279$$

Spurious Correlation

One of the major problem with correlation is the possibility of spurious correlation between two random variables which in many cases is caused due to some other latent variable (hidden variable) that influences both variables for which the correlation is calculated.

Following are few examples of spurious correlation between two random variables:

Crime rate versus ice cream sale: It has been reported that the sale of ice cream and crime rates are positively correlated (Levitt and Dubner, 2009). Obviously, ice cream is not driving the crime rate. In this case the hidden variable is the temperature (summer increasing the ice cream sale) and also increasing crime (people on vacation and locked houses becomes easy target).

Spurious Correlation

Doctors and deaths: Number of doctors is positively correlated with number of deaths in villages, that is, as the number of doctors increases, the deaths also increase. We can be sure that doctors are not causing the deaths to increase (Young, 2001).

Divorce rate in Maine and per capita consumption of margarine: The divorce rate in Maine was highly correlated with per capita consumption of margarine (based on data between 1999 and 2009.The correlation was 0.9926 (Source: tylervigen.com).

Correlation coefficient: points to ponder

1. Does correlation mean two variables are related?

2. If there is no correlation between two variables, can we conclude they are not related?

Hypothesis Test for Correlation Coefficient

For any two sets of data the Pearson correlation coefficient is most likely to give a value other than zero. Many thumb rules exist to group the correlation value as no correlation, low correlation, medium correlation, and high correlation (Monroe and Stuit, 1933).

Let ρ be the population correlation coefficient. The null and alternative hypotheses are given by

H_0:	**$\rho = 0$ (there is no correlation between two random variables)**
H_A:	$\rho \neq 0$ (there is a correlation between two random variables)

- The sampling distribution of correlation coefficient *r* follows an approximate *t*-distribution with (*n* – 2) degrees of freedom (*df*) where *n* is the number of cases in the sample for calculating the correlation coefficient.
- Two degrees of freedom are lost since we estimate two mean values from the data. The mean of the sampling distribution is ρ and the corresponding standard deviation is (Ezekiel, 1941) $\sqrt{\frac{1-r^2}{n-2}}$
- The *t*-statistic for null hypothesis is given by $t_{\alpha,n-2} = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}}$
- When the null hypothesis is $\rho = 0$, the test statistic in above Eq. becomes $t_{\alpha,n-2} = r\sqrt{\frac{n-2}{1-r^2}}$

https://archive.org/details/in.ernet.dli.2015.137861/page/n51/mode/2up (Page 487)

DATA ANALYTICS

Example

The average share prices of two companies over the past 12 months are shown in Table. conduct the following two hypothesis tests at α = 0.05:

(a) The correlation between share prices of two companies is zero.

(b) The correlation between share prices of two companies is at least 0.5.

X	Y
274.58	219.50
287.96	242.92
290.35	245.90
320.07	256.80
317.40	240.60
319.53	245.23
301.52	232.09
271.75	222.65
323.65	231.74
259.80	214.43
263.02	201.86
286.03	204.23

PES UNIVERSITY ONLINE

Solution

(a) The null and alternative hypotheses are:

H_0: $\rho = 0$

H_A: $\rho \neq 0$

The corresponding *t*-statistic is $t = r\sqrt{\frac{n-2}{1-r^2}} = 0.7279\sqrt{\frac{12-2}{1-0.7279^2}} = 3.3569$

Note that this is a two-tailed test and the critical *t*-value at α = 0.05 and *df* = 10 is 2.2281

Since the calculated *t*-statistic is higher than the critical *t*-value, we reject the null hypothesis and conclude that there is a significant correlation between share prices of two companies.

The corresponding *p*-value is 0.0072

Solution

(b) The null and alternative hypotheses are given by

$H_0: \rho \leq 0.5$

$H_A: \rho > 0.5$

The corresponding *t*-statistic is $t = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.7279-0.5}{0.2168} = 1.05$

This is a right-tailed test and the corresponding *t*-critical value is 1.8124

The calculated *t*-value is less than the critical value of *t*, and thus we retain the null hypothesis and conclude that the correlation between share prices of two companies is less than 0.5.

The corresponding *p*-value is 0.1592 .

Spearman Rank Correlation

- Pearson correlation is appropriate when the random variables involved are both from either ratio scale or interval scale.
- When both random variables are of ordinal scale, we use Spearman rank correlation (also known as Spearman's rho denoted by ρ_s).
- The Spearman rank correlation, r_s, estimated from a sample is given by (Yule and Kendall 1937, Woodbury, 1940)

$$r_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2 - 1)}$$

where D_i = difference in the rank of case i under variables X and Y (that is $X_i - Y_i$).

- The sampling distribution of Spearman correlation r_s also follows an approximate t-distribution with mean ρ_s and standard deviation with $n - 2$ degrees of freedom

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

Example

Ranking of 12 countries under corruption and Gini Index (wealth discrimination) are shown in Table. Calculate the Spearman correlation and test the hypothesis that the correlation is at least 0.2 at $\alpha = 0.02$.

Countries	1	2	3	4	5	6	7	8	9	10	11	12
Corruption	1	4	12	2	5	8	11	7	10	3	6	9
Gini Index	2	3	9	5	4	6	10	7	8	1	11	12

Solution

The Spearman rank correlation calculations are shown in Table

Country	Corruption Rank (X_i)	Gini Index (Y_i)	$D = X_i - Y_i$	D^2
1	1	2	–1	1
2	4	3	1	1
3	12	9	3	9
4	2	5	–3	9
5	5	4	1	1
6	8	6	2	4
7	11	10	1	1
8	7	7	0	0
9	10	8	2	4
10	3	1	2	4
11	6	11	–5	25
12	9	12	–3	9
			$\sum_{i=1}^{12} D_i^2$	68

Solution

The Spearman rank correlation is given by

$$r_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} = 1 - \frac{6 \times 68}{12(12^2-1)} = 0.7622$$

The null and alternative hypotheses are

H_0: $\rho_s < 0.2$

H_A: $\rho_s \geq 0.2$

The corresponding *t*-statistic is

$$t = \frac{r_s - \rho_s}{\sqrt{\frac{1-r_s^2}{n-2}}} = \frac{0.7622 - 0.2}{0.2046} = 2.74$$

The one-tailed *t*-critical value for $\alpha = 0.02$ and $df = 10$ is 2.35.

Since the calculated *t*-statistic value is more than the *t*-critical value, we reject the null hypothesis and conclude that Spearman rank correlation between two countries is at least 0.2.

Point Bi-Serial Correlation

Point bi-serial correlation is used when we are interested in finding correlation between a continuous random variable and a dichotomous (binary) random variable.

Point Bi-Serial Correlation

- Assume that the random variable X is a continuous random variable and Y is a dichotomous random variable. Then the following steps are used for calculating the correlation between these two variables:
  1. Group the data into two sets based on the value of the dichotomous variable Y. That is, assume that the value of Y is either 0 or 1. Then we group the data into two subsets such that in one group the value of Y is 0 and in another group the value of Y is 1
  2. Calculate the mean values of two groups: Let $\bar{x}_0$ and $\bar{x}_1$ be the mean values of groups with $Y = 0$ and $Y = 1$, respectively.
  3. Let n_0 and n_1 be the number of cases in a group with $Y = 0$ and $Y = 1$, respectively, and S_X be the standard deviation of the random variable X.

Point Bi-Serial Correlation

The point bi-serial correlation is given by (Pearson, 1909 and Soper, 1914)

$$r_b = \frac{\bar{X}_1 - \bar{X}_0}{S_X} \sqrt{\frac{n_0 n_1}{n(n-1)}}$$

where n is the total number of cases in the sample and S_X is the standard deviation of X estimated from sample and is given by

$$S_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2}$$

Example

Ms Sandra Ruth, data scientist at Airmobile, is interested in finding the correlation between the average call duration and gender. Table provides the average call duration (measured in seconds) and gender of 30 customers of Airmobile. In Table, male is coded as 0 and Female is coded as 1. Calculate the point bi-serial correlation.

Gender	1	1	0	0	0	1	0	1	1	0
Call Duration	448	335	210	382	407	231	359	287	288	347
Gender	1	1	1	1	1	0	0	1	0	0
Call Duration	408	382	303	201	447	439	383	277	279	213
Gender	1	1	0	1	1	0	1	0	1	0
Call Duration	383	355	362	401	331	421	367	437	326	351

Solution

From the data, we can calculate the following values:

$$\bar{X} = 345.33,\ \bar{X}_0 = 353.07,\ \bar{X}_1 = 339.4118,\ S_X = 71.7189,\ n_0 = 13,\ n_1 = 17$$

Bi-serial correlation is given by

$$r_b = \frac{\bar{X}_1 - \bar{X}_0}{S_X}\sqrt{\frac{n_0 n_1}{n(n-1)}} = \frac{339.4118 - 353.07}{71.7189}\sqrt{\frac{13\times 17}{30(29)}} = -0.0960$$

There is very low negative correlation between gender and call duration.

The Phi-Coefficient

Karl Pearson recommended the use of the Phi-coefficient when both variables are binary for calculating the association relationship (Cramer, 1946). Let *X* and *Y* be two random variables both taking binary values (that is, *X* takes values 0 or 1 and similarly *Y* also takes values either 0 or 1). One can create a contingency table as shown in Table below.

	Y = 0	Y = 1	Total
X = 0	N_{00}	N_{01}	$N_{X0} = N_{00} + N_{01}$
X = 1	N_{10}	N_{11}	$N_{X1} = N_{10} + N_{11}$
Total	$N_{Y0} = N_{00} + N_{10}$	$N_{Y1} = N_{01} + N_{11}$	

Also, for contingency tables for presence or absence of two categorical variables

	Drink Coffee	Don't drink Coffee
Drink Tea	N_{11}	N_{10}
Don't drink Tea	N_{01}	N_{00}

The Phi-Coefficient

In the contingency table (Table in previous slide)

- N_{00} = Number of cases in the sample such that $X = 0$ and $Y = 0$
- N_{01} = Number of cases in the sample such that $X = 0$ and $Y = 1$
- N_{10} = Number of cases in the sample such that $X = 1$ and $Y = 0$
- N_{11} = Number of cases in the sample such that $X = 1$ and $Y = 1$
- N_{X0} = Number of cases in the sample such that $X = 0$
- N_{X1} = Number of cases in the sample such that $X = 1$
- N_{Y0} = Number of cases in the sample such that $Y = 0$
- N_{Y1} = Number of cases in the sample such that $Y = 1$
- The Phi-coefficient is given by

$$\phi = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{N_{X0}N_{X1}N_{Y0}N_{Y1}}}$$

Example

Joy Finance (JF) is a company that provides gold loans (in which gold is used as guarantee against the loan). Mr Georgekutty, Managing Director of JF, collected data to understand the relationship between loan default status (variable *Y*) and the marital status of the customer (variable *X*). Data is collected on past 40 loans and is shown in Table 8.8. Calculate the Phi-coefficient. In Table , *Y* = 0 implies non-defaulter, *Y* = 1 is defaulter, *X* = 0 is single, and *X* = 1 is married.

X	1	0	1	0	0	0	0	0	1	0
Y	0	1	0	1	0	0	0	1	1	1
X	0	1	1	0	0	1	0	0	0	1
Y	0	1	1	1	0	0	1	1	0	0
X	1	0	0	0	1	1	1	0	0	1
Y	0	0	0	1	0	0	0	0	0	0
X	1	0	0	0	1	0	1	0	1	1
Y	1	0	0	1	1	0	1	1	0	1

Solution

The contingency table for the data shown in above Table is given in below Table

		Y		Total
		0	1	
X	0	13	10	23
	1	10	7	17
Total		23	17	40

From Table , we have N00 = 13, N01 = 10, N10 =10, N11 = 7, NX0 = 23, NX1 = 17, NY0 = 23, and NY1 = 17

The Phi-coefficient is given by

$$\phi = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{N_{X0}N_{X1}N_{Y0}N_{Y1}}} = \frac{7\times13 - 10\times10}{\sqrt{23\times17\times23\times17}} = -0.0230$$

Since the Phi-coefficient is very small, we can conclude that there is not much association between the marital status and loan default.

Summary

(a) Correlation is a measure of strength and direction of linear relationship between two random variables. It can be used only when the relationship is linear.

(b) Correlation captures only association relation and not a causal relation.

(c) Pearson product moment correlation is used when two random variables are continuous. In the case of two ordinal variables the appropriate correlation is Spearman rank correlation.

Summary

(d) Bi-serial correlation is used when correlation is calculated between one continuous and one binary random variable. Phi-coefficient is used for calculating correlation between two binary random variables.

(e) Correlation is an important measure and can be used for feature selection while building regression models.

(f) One of the drawbacks of correlation is the spurious correlations, it is possible that two variables with no explainable relationship may have high correlation coefficient.

Exercise

- Mention and explain the different correlation coefficients.
- For each of the correlation coefficient, find out an application and explore how it is used in that application.

The journey so far…

- Unit 1: Exploratory Data Analysis + Visualization
  - Data acquisition and framing questions given data
  - Taking stock of data: missing values or NA, outliers, anomalies (incorrect data, inconsistent data, incomplete data, noisy data, etc.)
  - Data cleaning and pre-processing
  - Data integration (removal of redundany)
    - Dimensionality reduction (wavelets, PCA, feature subset selection, feature creation)
    - Data reduction (sampling, binning/ histograms, discretization, clustering, etc.)
  - Data transformation (z-score, range normalization, unit norm, etc.)
- Unit 2: Correlation analysis
- Coming up next week – Unit 2: Regression

References

Text Book:

"Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834