1. Calculate covariance and correlation between below two columns A and B.

Ans: Covariance : If we understand the formula of covariance

∑(X-X’)(Y-Y’)/n-1

X is data point value of first variable

X’ is mean of first variable

Y is data point value of second variable

Y’ is mean of second variable

n is the total number of data points.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | X-X' | Y | Y-Y' | (X-X')\*(Y-Y') |
| 25 | -23.1429 | 52 | 6 | -138.8571429 |
| 35 | -13.1429 | 10 | -36 | 473.1428571 |
| 21 | -27.14 | 5 | -41 | 1112.74 |
| 67 | 18.85714 | 98 | 52 | 980.5714286 |
| 98 | 49.85714 | 52 | 6 | 299.1428571 |
| 27 | -21.1429 | 36 | -10 | 211.4285714 |
| 64 | 15.85714 | 69 | 23 | 364.7142857 |
|  |  |  |  | 3302.882857 |
|  |  |  |  |  |

If we substitute the above values in formula Covariance=3302.882857/7-1=550.48.

Correlation is scaled version of Covariance and formula is given as

=cov(A,B)/sd(X)sd(Y)

Cov(A,B)=550.48 from above step.Lets calculate standard deviations

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | X-X' | (X-X')^2 | Y | Y-Y' | (Y-Y')^2 | (X-X')\*(Y-Y') |
| 25 | -23.1429 | 535.5918 | 52 | 6 | 36 | -138.8571429 |
| 35 | -13.1429 | 172.7347 | 10 | -36 | 1296 | 473.1428571 |
| 21 | -27.14 | 736.5796 | 5 | -41 | 1681 | 1112.74 |
| 67 | 18.85714 | 355.5918 | 98 | 52 | 2704 | 980.5714286 |
| 98 | 49.85714 | 2485.735 | 52 | 6 | 36 | 299.1428571 |
| 27 | -21.1429 | 447.0204 | 36 | -10 | 100 | 211.4285714 |
| 64 | 15.85714 | 251.449 | 69 | 23 | 529 | 364.7142857 |
|  |  | 4984.702 |  |  | 6382 | 3302.882857 |

Sd(x)=sqrt((x-x’)^2)

Sd(y)=sqrt((y-y’)^2)

From above table sd(x)=70.6024,sd(y)=79.88

If we substitute these values in formula

550.48/70.60\*79.88=0.0975

1. What are the different ways to deal with multi collinearity?

Multicollinearity occurs when two or more variables are linearly interdependent. If we consider these variables in our model builidng the predictive model may not work as expected. One of the interdependent variable is ignored. We can find the collinearity between variables using Correlation ,Anova, Chi-Square test.

Correlation : Both the variables are numeric.

Anova test used for one categorical and one continuous.

Chi-Square test is used for two categorical variables.

1. What should be the correlation threshold value based on which we determine the highly collinear variables?

Ans: Correlation threshold value ranges from -1 to +1. It is scaled version of Covariance.

1. What are the two different types of variable we used in ANOVA?

Ans: In Anova one is categorical and one is Continuous variable.

Example : Height of students and Gender of students. Height is continuous variable where Gender is categorical variable. Is height dependent on Gender is the relation.

1. What are the null and alternate hypothesis in chi-square test?

Ans:

In null hypothesis of chi-square test the two categorical variables are independent.

Example : Relation between Rain and traffic. If rains more traffic. If no rain more traffic. Here traffic is not dependent on rain.

In alternate Hypothesis of chi-square test the variables are not independent.

Example : Same relation between Rain and traffic. If rains more traffic and if no rain less traffic. Here traffic is dependent on rain.