Course: Artificial Intelligence and Machine Learning

Code: 20CS51I, WEEK - 4

SESSION – 5, Multivariate analysis

Finding relationship in data

- Covariance

- Correlation

Finding relationship in data

Covariance and Correlation are two mathematical concepts which are commonly used in the field of probability and statistics. Both concepts describe the relationship between two variables.

Covariance

1. It is the relationship between a pair of random variables where change in one variable causes change in another variable.
2. It can take any value between -infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.
3. It is used for the linear relationship between variables.
4. It gives the direction of relationship between variables.

Formula –

For Population:

$$Covri(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{n}$$

For Sample

$$Covari(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{n-1}$$

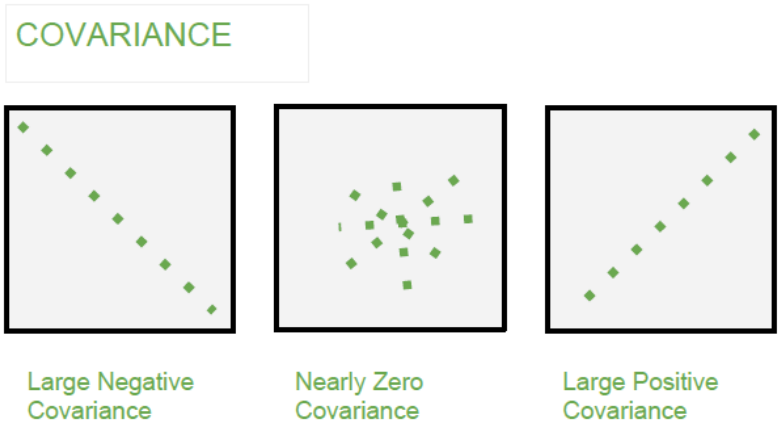
Here,

x' and y' = mean of given sample set

n = total no of sample

xi and yi = individual sample of set

Example –



Formula for Covariance

For example, the covariance between two random variables X and Y can be calculated using the following formula (for population):

$$\text{Cov}(X,Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

For a sample covariance, the formula is slightly adjusted:

$$\text{Cov}(X,Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n - 1}$$

Where:

- **Xi** – the values of the X-variable
- **Yj** – the values of the Y-variable
- $\bar{X}$ – the mean (average) of the X-variable

- $\bar{\mathbf{Y}}$ – the mean (average) of the Y-variable
- **n** – the number of data points

The relationship between the two concepts can be expressed using the formula below:

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

Where:

- **ρ(X,Y)** – the correlation between the variables X and Y
- **Cov(X,Y)** – the covariance between the variables X and Y
- **σX** – the standard deviation of the X-variable
- **σY** – the standard deviation of the Y-variable

Example of Covariance

John is an investor. His portfolio primarily tracks the performance of the S&P 500 and John wants to add the stock of ABC Corp. Before adding the stock to his portfolio, he wants to assess the directional relationship between the stock and the S&P 500.

John does not want to increase the unsystematic risk of his portfolio. Thus, he is not interested in owning securities in the portfolio that tend to move in the same direction. He can calculate the covariance between the stock of ABC Corp. and S&P 500 by following the steps below:

1. Obtain the data.

First, John obtains the figures for both ABC Corp. stock and the S&P 500. The prices obtained are summarized in the table below:

	S&P 500	ABC Corp.
2013	1,692	68
2014	1,978	102
2015	1,884	110
2016	2,151	112
2017	2,519	154

2. Calculate the mean (average) prices for each asset.

$$\text{Mean (S\&P 500)} = \frac{1{,}692 + 1{,}978 + 1{,}884 + 2{,}151 + 2{,}519}{5} = 2{,}044.80$$

$$\text{Mean (ABC Corp.)} = \frac{68 + 102 + 110 + 112 + 154}{5} = 109.20$$

3. For each security, find the difference between each value and mean price.

			Step 3		Step 4
	S&P 500	ABC Corp.	a	b	a x b
2013	1,692	68	-352.80	-41.20	14,535.36
2014	1,978	102	-66.80	-7.20	480.96
2015	1,884	110	-160.80	0.80	-128.64
2016	2,151	112	106.20	2.80	297.36
2017	2,519	154	474.20	44.80	21,244.16
Mean	**2,044.80**	**109.20**	**Sum**		**36,429.20**

4. Multiply the results obtained in the previous step.

5. Using the number calculated in step 4, find the covariance.

$$\text{Cov(S\&P 500, ABC Corp.)} = \frac{36{,}429.20}{5 - 1} = 9{,}107.30$$

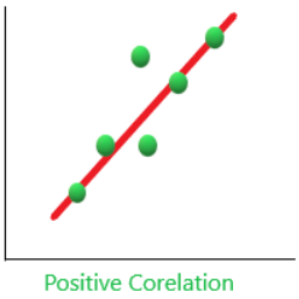
In such a case, the positive covariance indicates that the price of the stock and the S&P 500 tend to move in the same direction.

Correlation

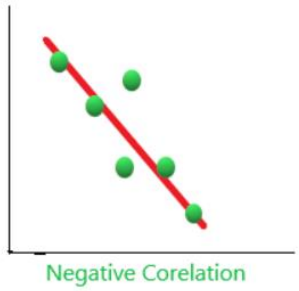
Correlation means an association, It is a measure of the extent to which two variables are related.

1. It show whether and how strongly pairs of variables are related to each other.
2. Correlation takes values between -1 to +1, wherein values close to +1 represents strong positive correlation and values close to -1 represents strong negative correlation.
3. In this variable are indirectly related to each other.
4. It gives the direction and strength of relationship between variables.

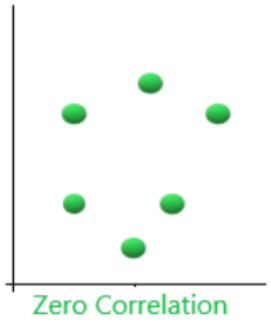
1. Positive Correlation: When two variables increase together and decrease together. They are positively correlated. '1' is a perfect positive correlation. For example – demand and profit are positively correlated the more the demand for the product, the more profit hence positive correlation.



2. Negative Correlation: When one variable increases and the other variable decreases together and vice-versa. They are negatively correlated. For example, If the distance between magnet increases their attraction decreases, and vice-versa. Hence, a negative correlation. '-1' is no correlation



3. Zero Correlation(No Correlation): When two variables don't seem to be linked at all. '0' is a perfect negative correlation. For Example, the amount of tea you take and level of intelligence.



Plotting Correlation matrix using Python

Step 1: Importing the libraries.

- Python3

```
import sklearn

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd
```

Step 2: Finding the Correlation between two variables.

- Python3

```
y = pd.Series([1, 2, 3, 4, 3, 5, 4])

x = pd.Series([1, 2, 3, 4, 5, 6, 7])

correlation = y.corr(x)

correlation
```

Output:

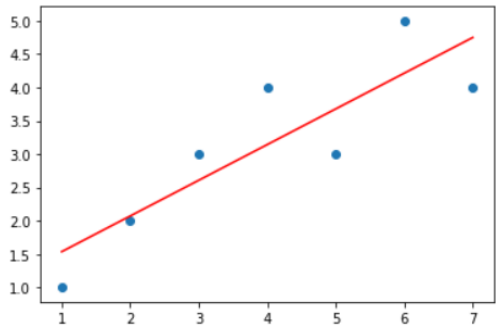
```
0.8603090020146067
```

Step 3: Plotting the graph. Here we are using scatter plots. A scatter plot is a diagram where each value in the data set is represented by a dot. Also, it shows a relationship between two variables.

- Python3

```
# plotting the data

plt.scatter(x, y)

 # This will fit the best line into the graph

plt.plot(np.unique(x), np.poly1d(np.polyfit(x, y, 1))

        (np.unique(x)), color='red')
```

Output:



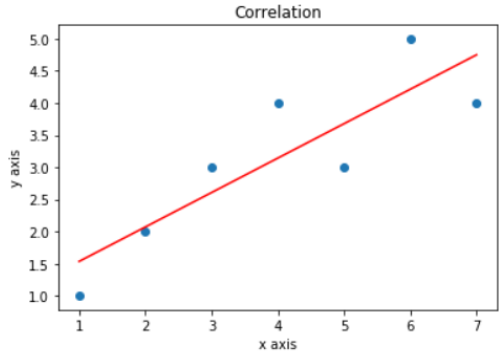
Observe both the images you will find similarity Also, observe the value of the correlation is near to 1, hence the positive correlation is reflected.

Adding title and labels in the graph

- Python3

```
# adds the title
plt.title('Correlation')
 # plot the data
plt.scatter(x, y)
 # fits the best fitting line to the data
plt.plot(np.unique(x),
         np.poly1d(np.polyfit(x, y, 1))
         (np.unique(x)), color='red')
 # Labelling axes
plt.xlabel('x axis')
plt.ylabel('y axis')
```

Output:



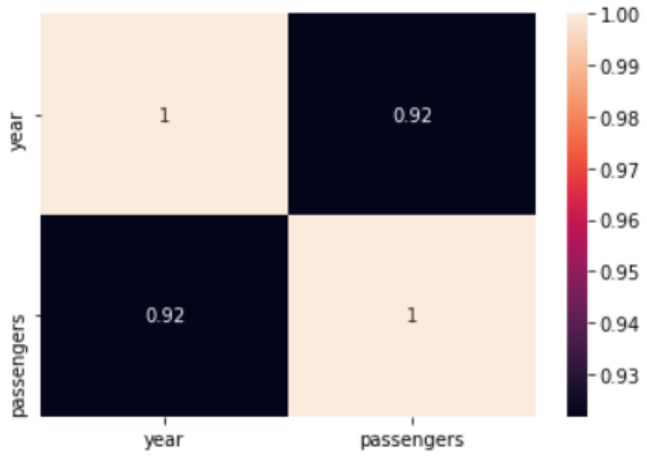
Plot using Heatmaps

There are many ways you can plot correlation matrices one efficient way is using the heatmap. It is very easy to understand the correlation using **heatmaps** it tells the correlation of one feature(variable) to every other feature(variable). In other words, A correlation matrix is a tabular data representing the 'correlations' between pairs of variables in a given data.

- Python3

```
import seaborn as sns

 # checking correlation using heatmap

#Loading dataset

flights = sns.load_dataset("flights")

 #plotting the heatmap for correlation

ax = sns.heatmap(flights.corr(), annot=True)
```

Output:



Covariance versus Correlation –

Covariance	Correlation
Covariance is a measure of how much two random variables vary together	Correlation is a statistical measure that indicates how strongly two variables are related.
involve the relationship between two variables or data sets	involve the relationship between multiple variables as well
Lie between -infinity and +infinity	Lie between -1 and +1
Measure of correlation	Scaled version of covariance
provide direction of relationship	provide direction and strength of relationship

Covariance	Correlation
dependent on scale of variable	independent on scale of variable
have dimensions	dimensionless

Multivariate Analysis for Correlation

Heat map plots graphically the actual correlation values using color and measure of the linear relationships.

Functions to use:

- sns.heatmap() —axes-level plot

First, we run df.corr() to get a table with the correlation coefficients. This table is also known as a correlation matrix.

```
cars.corr()
```

	year	selling_price	km_driven	mileage_kmpl	engine_cc	max_power_bhp	seats
year	1.000000	0.414092	-0.418006	0.329145	0.018848	0.226320	-0.009144
selling_price	0.414092	1.000000	-0.225534	-0.126054	0.455734	0.748489	0.041358
km_driven	-0.418006	-0.225534	1.000000	-0.173073	0.205914	-0.038075	0.227336
mileage_kmpl	0.329145	-0.126054	-0.173073	1.000000	-0.575831	-0.374621	-0.452085
engine_cc	0.018848	0.455734	0.205914	-0.575831	1.000000	0.703975	0.610309
max_power_bhp	0.226320	0.748489	-0.038075	-0.374621	0.703975	1.000000	0.191999
seats	-0.009144	0.041358	0.227336	-0.452085	0.610309	0.191999	1.000000

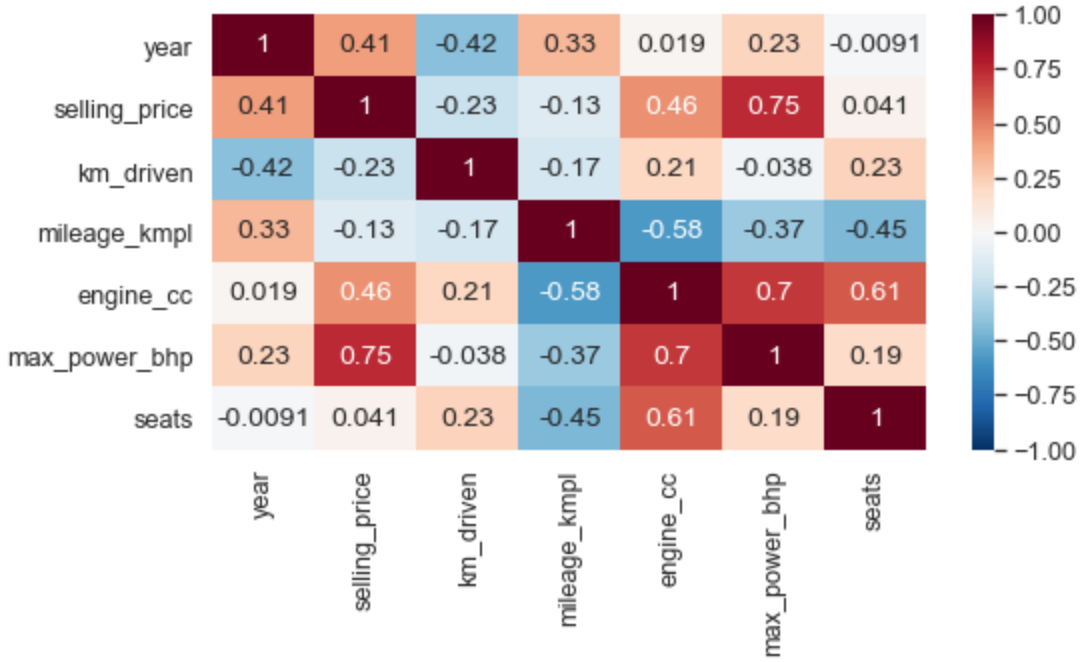
Correlation matrix

sns.heatmap()

```
sns.set(font_scale=1.15)
plt.figure(figsize=(8,4))sns.heatmap(
   cars.corr(),
```

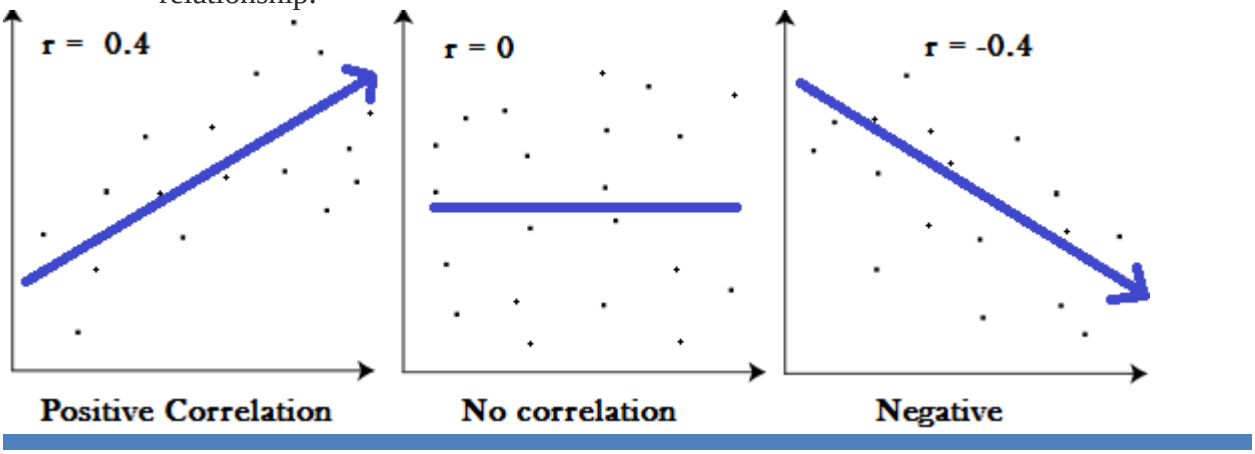
```
    cmap='RdBu_r',
    annot=True,
    vmin=-1, vmax=1);
```

cmap='RdBu_r' sets the color scheme, annot=True draws the values inside the cells, and vmin and vmax ensures the color codes start at -1 to 1.



What to look out for:

- Highly correlated features. These are the dark-red and dark-blue cells. Values close to 1 mean a high positive linear relationship, while close to -1 show a high negative relationship.



https://www.geeksforgeeks.org/create-a-correlation-matrix-using-python/