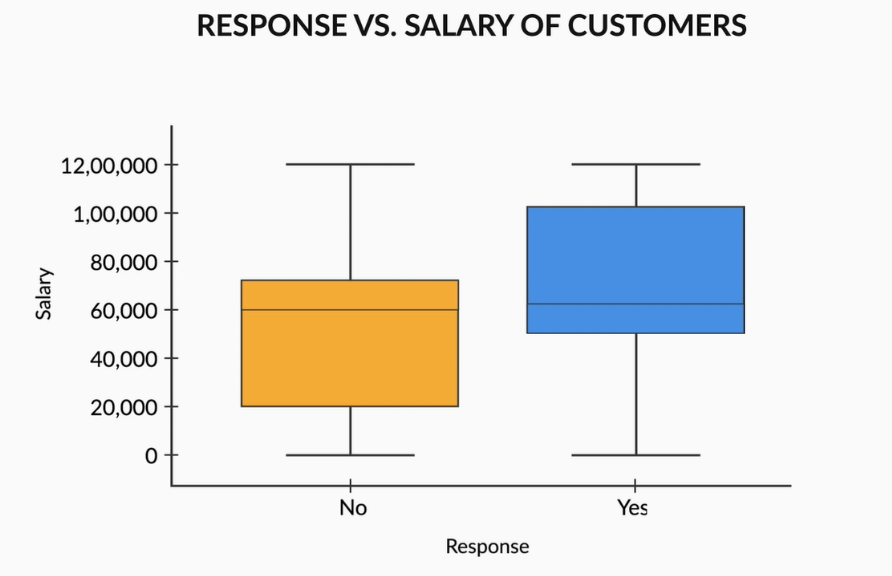
Bivariate and Multivariate Analysis

# Introduction

Welcome to the session on **'Bivariate and Multivariate Analysis'**.

So far, you have learnt how to preprocess and clean data, and then you performed a univariate analysis on the bank marketing dataset. A univariate analysis includes the analysis of individual categorical variables like job, education, response, marital status, etc. It also explains the concept of outliers, and mean, median or mode of numerical variables such as salary, balance and age.

Now, consider the following graph, which plots the 'Response' and the 'Salary' columns from the bank marketing dataset. You have already plotted this graph in the previous segments. Although the median , maximum and minimum values are the same, customers with a higher salary range show interest in opening term deposit accounts with the bank. This is nothing but a bivariate analysis, that is, the analysis of two variables/columns in the data set.



## In this session

In this session, you will learn how to analyse two or more variables at a time. You will also observe and draw better inferences on the types of customers that are showing interest in opening term deposits account with the bank. This will give you better insights and understanding of how to conduct effective marketing campaigns in the future.

Play Video

1569002

You have already learnt about the types of variables, that is, categorical and numeric variables, in the previous sessions. This session has been divided into the following topics based on the different types of variables:

* Analysis between two numeric variables
* Analysis between numeric and categorical variables
* Correlation versus causation
* Analysis between two categorical variables
* Multivariate analysis

## Guidelines for In-Module Questions

The in-video and in-content questions for this module are not graded. Note that graded questions are given on a separate page labelled 'Graded Questions' at the end of each session. The graded questions in these sessions will adhere to the following guidelines:

|  | First Attempt Marks | Second Attempt Marks |
| --- | --- | --- |
| Question with 2 Attempts | 10 | 5 |
| Question with 1 Attempt | 10 | 0 |

## People you will hear from in this module

**Subject Matter Expert**

[Mirza Rahim Baig](https://www.linkedin.com/in/rahim-baig)

Analytics Lead, Flipkart

Flipkart is one of the leading e-commerce companies in India. It started with selling books and has now expanded its business to almost every product category, including consumer electronics, fashion and lifestyle products. Rahim is currently the Analytics Lead at Flipkart. He holds a graduate degree from BITS Pilani, a premier educational institute in India.

**Subject Matter Expert**

[S Anand](https://www.linkedin.com/in/sanand0)

CEO, Gramener

Gramener is one of the most prominent data analytics and visualisation companies in India. Anand, currently the CEO, was previously the Chief Data Scientist at Gramener and also has extensive experience in management consulting and equity research.

# Numeric - Numeric Analysis

In this segment, you will learn how to analyse two numerical variables using the Bank marketing dataset. Now, there are multiple tools to analyse numerical variables. In the next video, you will learn about the different tools and plots that are helpful for extracting insights using numerical variables from a data set.

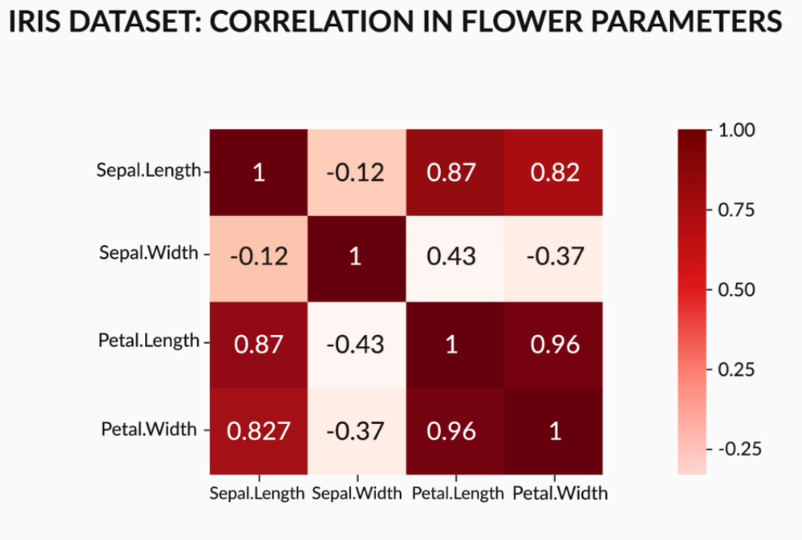
Play Video

1569002

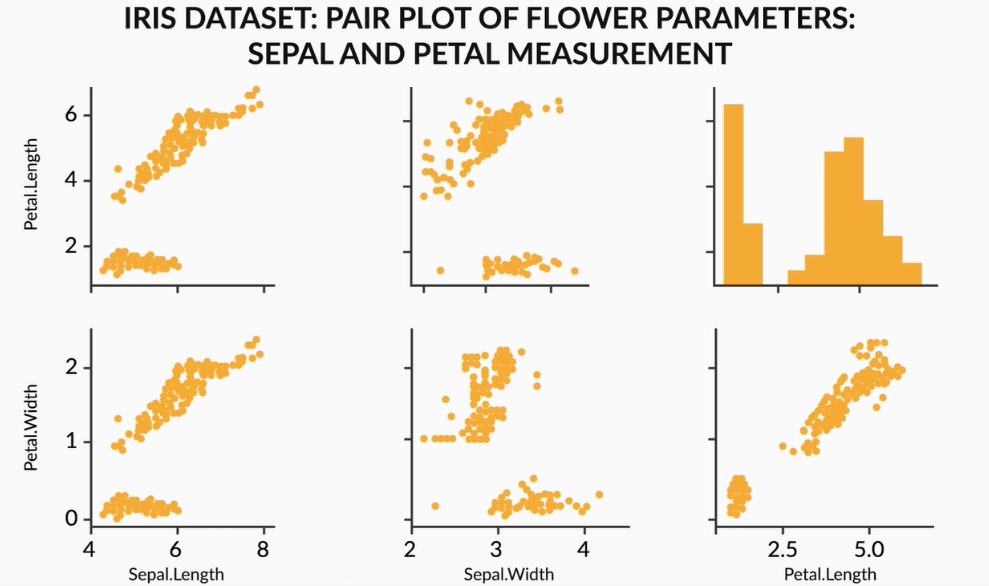
One very important concept that has been covered in the video above is that of correlation coefficient. Now, correlation coefficient depicts only a linear relationship between numerical variables. It does not depict any other relationship between variables. A zero correlation does not imply that there is no relation between variables; it merely indicates that there will no linear relationship between them. Also, there can be a negative or positive correlation between variables. A negative correlation means that if the value of one variable increases, the value of another decreases, whereas it is the opposite for a positive correlation.

Now, the higher the coefficient of correlation between numerical variables, the higher the linear relation between them.

From the **correlation matrix** below,you can observe that petal length has a high correlation with sepal length, with a correlation coefficient of 0.87. Also, there is a very high correlation coefficient of 0.96 between petal width and petal length.



However, the correlation matrix has its own limitations where you cannot see the exact distribution of a variable with another numeric variable. To solve this problem, we use **pair plots**. Pair plots are scatter plots of all numeric variables in a data set. It shows the exact variation of one variable with respect to the others. You can observe how one variable is varying with respect to another in the image below.

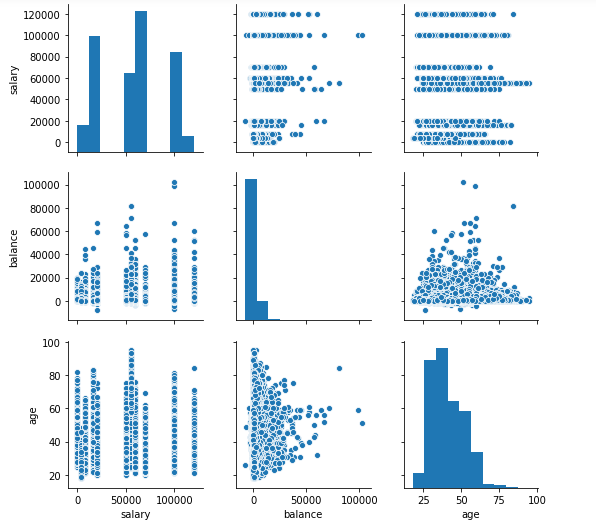


Now, in the following video, Rahim will explain how to perform a numeric bivariate analysis using the bank marketing dataset.

Play Video

1569002

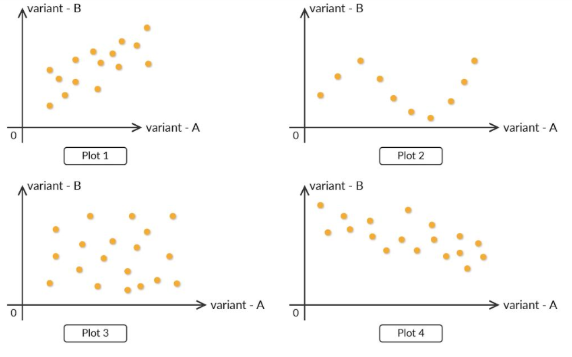
So, in the video, you saw how a pair plot can help you determine that there is no correlation between the 'age', 'balance' and 'salary' variables. Now, refer to the image below and observe how there is no correlation between these variables.



A high correlation coefficient does not imply that there will be a correlation with another numeric variable every time because there can be no causation between them. There may be cases when you will see a high correlation coefficient between two variables but there is no relation between them. You will understand this in detail in the next segment that how correlation is related to the causation.

**Comprehension: Correlation**

Consider the following four scatter plots of two variables A and B.



Based on your learning in this segment, answer the following questions.

Question 1/3

Mandatory

#### **Correlation**

Out of the four plots, which has the least correlation coefficient?

Plot-1

Plot-2

Plot-3

✓ Correct

Feedback:

*Plot-3 does not show any relation between the variables; hence, it has the least correlation coefficient.*

Your answer is Correct.

Attempt 1 of 1

Continue

Question 2/3

Mandatory

#### **Correlation**

Which of the plots above has very a low correlation coefficient but has some relation between the variables?

Plot-1

Plot-2

✓ Correct

Feedback:

*There is some relation between the variables in Plot-2, but it is not linear.*

Plot-3

Your answer is Correct.

Question 3/3

Mandatory

#### **Correlation**

Which of the plots above has a negative correlation?

Plot-1

Plot-4

✓ Correct

Feedback:

*In Plot-4, with increase in one variable, the other variable decreases; hence, there is a negative correlation between the variables.*

Plot-3

Your answer is Correct.

# **Correlation vs Causation**

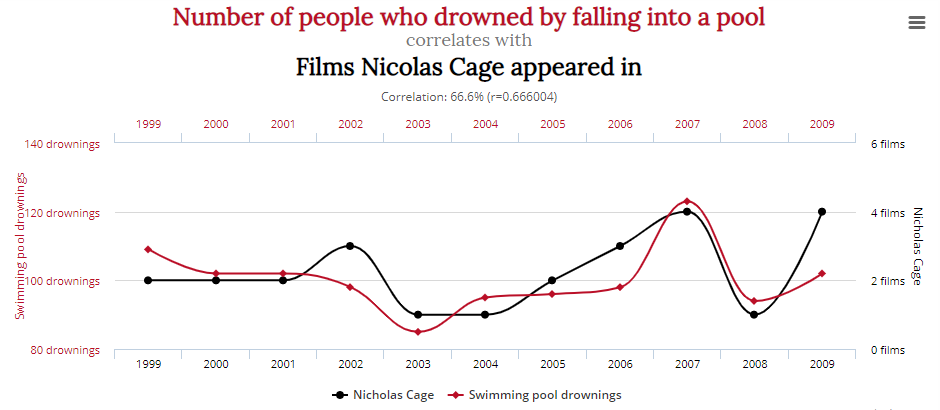
In the EDA exercise, it is very important to note that although some numerical variables can sometimes be highly correlated to each other, there may not be a cause of any relationship between them.

Let us first listen to Rahim in the next video and try to get a holistic picture of correlation among variables using some compelling examples.

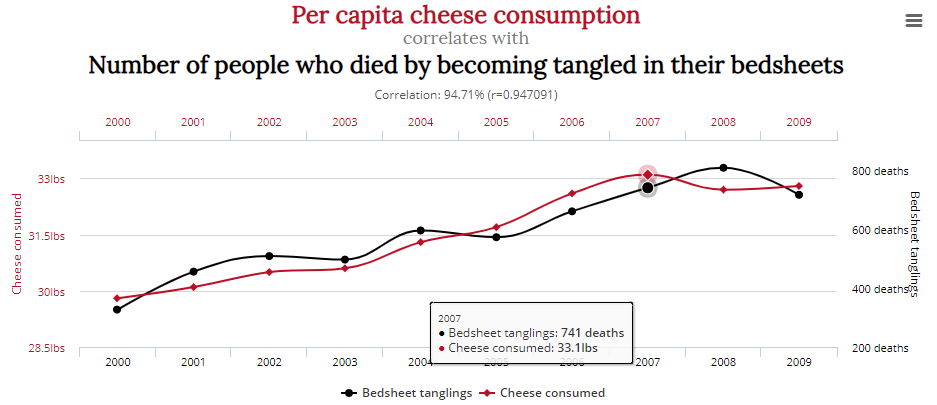
Play Video

1569002

So, the major takeaway from the video is that correlation does not imply causation. In the video, you saw that the number of people who drowned by falling into a pool is not related to movies starring Nicolas Cage. However, if you observe the plot below, you will notice that there is a very high correlation between them, as both the plots follow almost the same path.



Now, in the example below, it is quite obvious that the per capita cheese consumption has no relation with people dying from being tangled in bed sheets, although the plot shows a high relation between them.



For more such compelling examples, where causation and correlation are not related to each other, you can refer to [this](http://www.tylervigen.com/spurious-correlations) link.

In the next segment, you will learn about bivariate analysis using numerical and categorical variables.

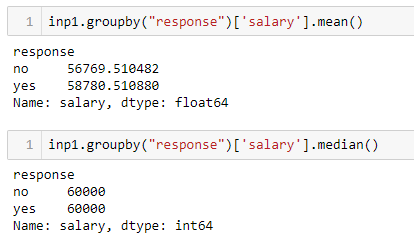
# **Numerical - Categorical Analysis**

Previously, you learnt about the bivariate analysis of numerical variables. In this segment, you will learn about the associations between numerical and categorical variables. You will learn how to apply this analysis on the same bank marketing dataset.

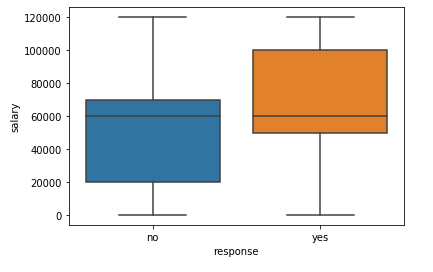
Play Video

1569002

So, in the video, you saw how the salary variable is varying with respect to the response variable. Their mean and median are the same, as shown in the image below.



However, a very different picture emerges when you plot a boxplot. The interquartile range for customers who gave a positive response is on the higher salary side. This is actually true, because people who have higher salaries are more likely to invest in term deposits.

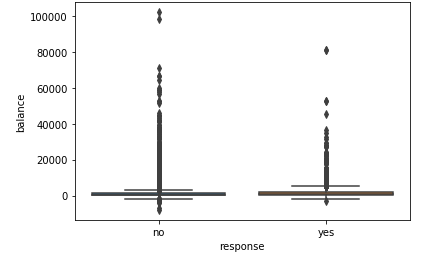


Now, in the next video, we will take a look at a different variable in the bank marketing dataset.

Play Video

1569002

In the video, you observed that after the balance versus response graph is plotted, it does not make any sense at first glance. Sometimes only a boxplot is not sufficient to draw insights, because of a high concentration of data and or because of higher values in the data set, for example, the balance variable.



In such cases, it is a good practice to analyse the data using mean, median or quartiles. In the video, you saw that the mean and median values of the balance variable are higher for customers who gave a positive response, which is again true, because people who have higher balance in their bank accounts are more likely to invest in term deposits.

In the next segment, you will get an idea about categorical versus categorical variable analysis.

Question 1/2

Mandatory

#### **Numerical - Categorical Variable Analysis**

Which of the following education levels shows the highest mean and the lowest median value for salary, respectively?

Primary and secondary

Tertiary and primary

✓ Correct

Feedback:

*Code in Python using the groupby command to find the mean and median values for salary in the different education categories.*

Secondary and tertiary

✕ Incorrect

Feedback:

*Code in Python using the groupby command to find the mean and median values for salary in the different education categories.*

Tertiary and secondary

Your answer is Wrong.

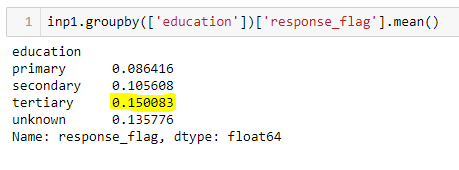
# **Categorical - Categorical Analysis**

In this segment, you will learn about the associations between two categorical variables in a bivariate analysis. Statistical analysis is essential for numerical variables, and it includes different metrics like mean, median, mode, quantiles and boxplots. Here, you will learn how to analyse categorical variables using graphs and charts, and derive maximum insights from them.

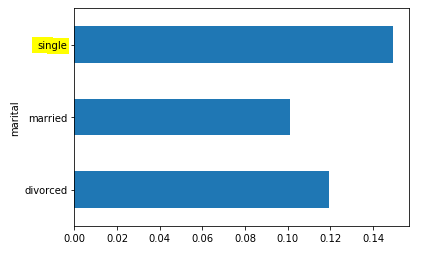
Play Video

1569002

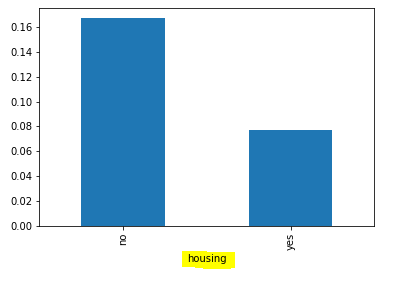
In the video, you saw that the positive response of customers to opening a term deposit with the bank increases with the education level. From this, you can infer that the bank should contact people with higher education levels to effectively increase the positive response for opening a term deposit.



Also, based on marital status analysis, you can infer that single individuals have a higher positive response rate. This could be due to various reasons: One reason could be that compared with other categories of customers, single individuals have available income to deposit in long-term savings accounts (term deposit). Hence, the campaign should target single customers.



Another very interesting inference is that people who have not purchased any housing or personal loan are more likely to open a term deposit account with the bank. This is true, probably because people who have already availed loans may not have the necessary funds to invest in a term deposit.

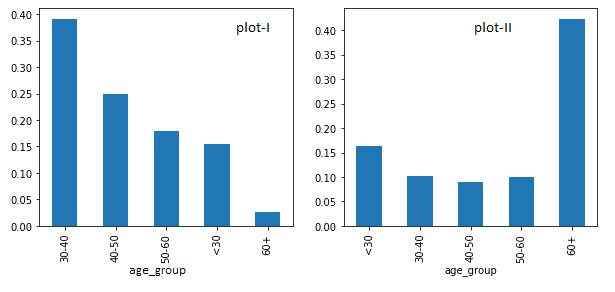


Now, let’s study the association between the age variable and response rate in the next video.

Play Video

1569002

So, age group analysis showed that people in the age group of 60+ or <30 are more likely to respond positively. It may be true for older people, since they want to invest through more secure investment methods such as term deposits to have a secure old age.



From the image above, you can observe that the bank has mostly contacted people in the age group of 30-50, and have made much less contact with people in the age group of 60+ (plot-I), although the chances of getting a positive response are higher from the people who are in the age group of 50+ or 60+ (as shown in plot-II). This is a very important insight that one can draw from this data set, i.e., the bank should target the people in the 50+ age group.

In the next segment, you will learn about multivariate analysis.

# **Multivariate Analysis**

In this segment, we will discuss the next part of EDA, i.e., multivariate analysis.

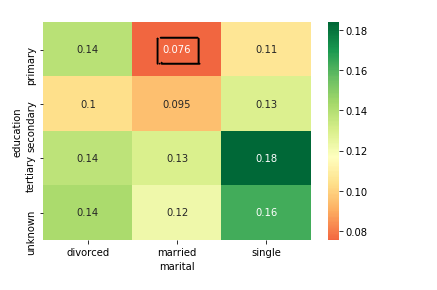
So far, you have learnt how two variables can be visualised based on their type, for example, numerical, categorical, etc. Now, let’s analyse two variables simultaneously. One of the key features of multivariate analysis is that it gives you a very precise idea about the various elements, since you are now combining multiple variables to visualise the data set. You will learn about this in more detail in the forthcoming videos.

First, let’s listen to Rahim as he explains his inferences from the bank marketing dataset in the next video.

Play Video

1569002

In the video above, you saw that our expert performed a three-variable analysis between education, marital status and response. You can see that people who are married and who have completed just their primary education are least likely to give a positive response on term deposits. This can be explained by the fact that people educated only up to the primary level are not aware of the benefits of term investments. Also, married individuals need money to fulfil their daily needs, and they require cash-on-hand to buy the daily essentials; hence, they won't prefer investing in term deposits.

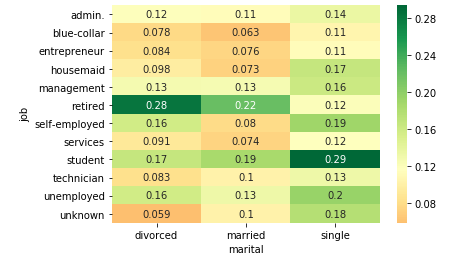


In the next video, you will see how job and marital status are varying with respect to the response variable.

Play Video

1569002

In the video, you saw that the combinations of married with blue-collar, entrepreneur and housemaid are least likely to go for term deposits. The highest rate of positive response came from students with single marital status. The bank should, therefore, consider these aspects before taking any decision.



Having gone through all these examples, you must have a clear idea about the EDA process and the various steps involved in it.

# **Graded Questions**

Suppose you are working as an analyst at an e-commerce company, and you have been given the following two data sets, which consist of details of women's apparels that have been sold out during the last year:

* **Dress Attribute Dataset:** This data set contains the different features of women's apparels.
* **Dress Sales Dataset:** This data set contains the number of sales achieved for a particular dress ID on a certain date.

So, let’s take a look at the first data set.

You can download the Dress Attribute Dataset from the link below:

Dress Attribute Dataset

Download

This data set contains a total of 13 features related to the dresses. These are the following:

* **Dress\_ID:** This represents the ID number of a particular dress; it is a unique identification number for different dresses.
* **Style:** This represents the style of a particular dress according to the occasion, for example, party, cute, vintage, etc.
* **Price:** For a particular dress ID, there are buckets of Low, Medium and High prices.
* **Rating:** This is the average rating given by the customer to a particular dress ID.
* **Size:** The size column shows the majority of the size purchased by the customers in the previous sale.
* **Season:** This depicts the season in which a particular type of dress is suitable, for example, summer, winter, etc.
* **Neckline:** This shows the type of neckline in a dress, for example, v-neck, o-neck, etc.
* **SleeveLength:** This represents the variety of sleeve length in the dresses, for example, half sleeves, full sleeves, cap sleeves, etc.
* **Material:** This represents the fabric material of the dresses, for instance, nylon, polyester, cotton, silk, etc.
* **FabricType:** This represents the type of fabric used for making the dress, for example, chiffon, broadcloth, jersey, etc.
* **Decoration:** This represents the kind of decoration around the dress like ruffles, bow, embroidery, etc.
* **Pattern Type:** This represents the kind of patterns that are printed on a particular dress. Pattern may be solid colours, geometric designs, printed or patchwork.
* **Recommendation:** This is the target variable; the recommendation is based on the features and sales of a dress in the previous year. This is either 1 (which means yes) or 0 (which means no). This shows whether a particular dress is suitable for sale to the customers or not.

Now, let's take a look at the second data set.

You can download the Dress Sales Dataset from the link below:

Dress Sales Dataset

Download

This particular data set represents the number of sales of a particular dress ID on a certain date, where the columns represent the dates on which a particular dress ID had been sold.

Now, based on the two data sets above, you are expected to perform EDA and draw meaningful insights from the data. Based on the EDA analysis, answer the graded questions.

In the previous segments, you learnt about bivariate and multivariate analysis with the example of a bank marketing dataset. Now, answer the following questions based on the steps that you have learnt about in this session.

You have been provided with a blank Jupyter Notebook with all the comments to perform t.

Jupyter Notebook\_Graded Exercise

Download

Question 1/6

Mandatory

#### **Numerical - Categorical Bivariate Analysis**

Which of the following categories of the 'Price' column has the lowest average rating?

very-high

Medium

Low

✕ Incorrect

Feedback:

*This category in the Price column has the highest average rating.*

High

✓ Correct

Feedback:

*Correct. This category in the Price column has the lowest average rating .*

print(inp0.groupby('Price')['Rating'].mean())

Your answer is Wrong.

Question 2/6

Mandatory

#### **Numerical - Categorical Bivariate Analysis**

What is the median value of rating for the 'vintage' category in the Style column?

4.6

✓ Correct

Feedback:

*This is the median value of rating for the 'vintage' category in the 'Style' column.*

print(inp0.groupby('Style')['Rating'].median())

4.7

4.55

0.00

Your answer is Correct.

Question 3/6

Mandatory

#### **Numerical - Categorical Bivariate Analysis**

Which of the following seasons shows the highest average value of sales for “Recommendation” value equals to 1.

Summer

Spring

✕ Incorrect

Feedback:

*The mean of sales in spring for recommendation value equals 1 is approximately 303, which is not the highest.*

Autumn

✓ Correct

Feedback:

*Find out the mean and median values of total sales in 'Autumn' with the 'Recommendation' variable. The mean of sales in the 'Autumn' column for recommendation equals 1 is approximately 2,870, which is the highest. You can refer to the following code to answer the question:*

print(inp0.groupby('Recommendation')['Autumn'].mean())

sns.boxplot(data=inp0, x="Recommendation",y="Autumn")

plt.show()

Winter

Your answer is Wrong.

Question 4/6

Mandatory

#### **Categorical - Categorical Bivariate Analysis**

Which of the following size categories has the most positive recommendations?

Medium and Extra Large

Extra Large and Small

Free and Small

Free and Medium

✓ Correct

Feedback:

*Code to plot the bar graph of average recommendation corresponding to the Size column in the notebook provided.*

inp0.groupby(['Size'])['Recommendation'].mean().plot.barh()

inp0.groupby('Size')['Recommendation'].sum()

Your answer is Correct.

Question 5/6

Mandatory

#### **Multivariate Analysis**

Which of the following pair of 'Style' and 'Price' categories has the highest average of positive recommendations?

Price: medium, style: vintage

Price: medium, style: cute

✓ Correct

Feedback:

*Plot the heat map of Style, Price and Recommendation. The average value for this combination is 1, which is maximum across all the values. You can refer to the code below:*

res = pd.pivot\_table(data=inp0, index="Style", columns="Price", values="Recommendation")

sns.heatmap(res, cmap="RdYlGn", annot=**True**, center=**0.427**)

plt.show()

Price: very high, style: party

Price: low, style: sexy

Your answer is Correct.

Question 6/6

Mandatory

#### **Multivariate Analysis**

Which of the following material types has no recommendation in summer and winter seasons?

Mixed and Milksilk

Nylon and Rayon

Microfiber and Silk

✕ Incorrect

Feedback:

*Silk has partial recommendation in winter.*

Milksilk and Microfiber

✓ Correct

Feedback:

*This combination does not have any recommendation in winter or summer.*

res = pd.pivot\_table(data=inp0, index="Material", columns="Season", values="Recommendation")

sns.heatmap(res, cmap="RdYlGn", annot=**True**, center=**0.427**)

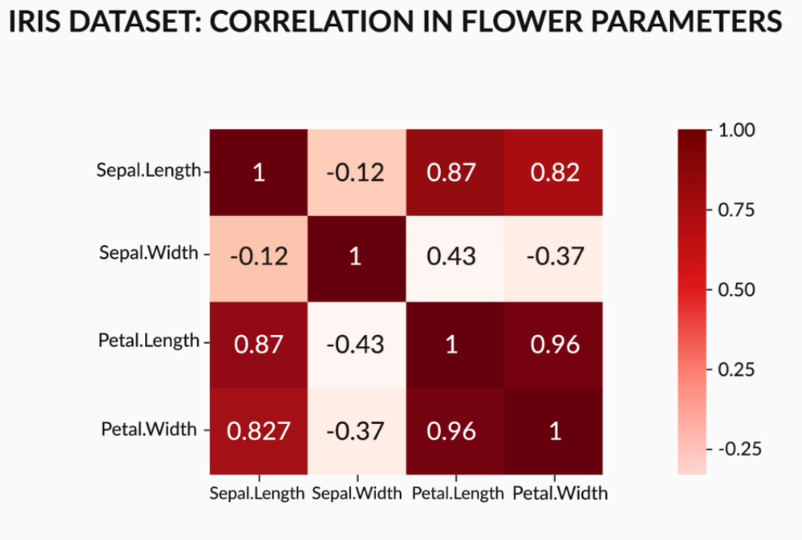
plt.show()

Your answer is Wrong.

# **Summary**

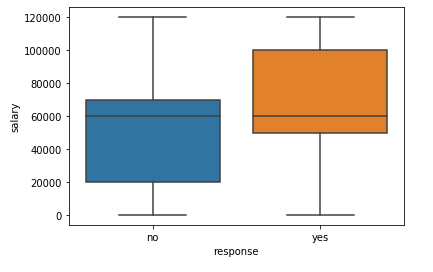
In this session, you learnt about the various types of bivariate and multivariate analysis. These include the following:

* **Analysis between two numerical variables:** The most important thing to remember is that **correlation** and **scatter plots** are the best methods to perform an analysis on numerical variables. Correlation coefficient indicates how much two numerical variables are correlated linearly. And scatter plots offer the exact visualisation between the numerical variables.



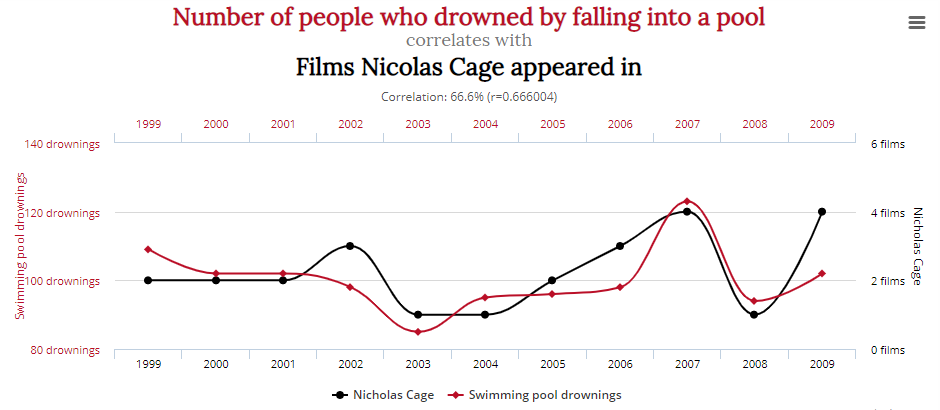
As you can observe in the correlation matrix above, among all the combinations in the data set, there is a high correlation between petal length and sepal length, and petal width and petal length.

* **Analysis between numerical and categorical variables:** This gives an idea about the variation of a particular numerical variable with respect to different categories of a categorical variable. **Boxplot** is the best way to look at a numerical variable with respect to a categorical variable. However, boxplots may sometime not be useful because of the huge difference between the maximum and minimum values in the data set, or due to the higher concentration of data in the numerical variable. Another approach could be to look into the mean/median or quartiles, which are a more efficient way to deal with a numerical variable when combined with a categorical variable. Take a look at the example shown below.

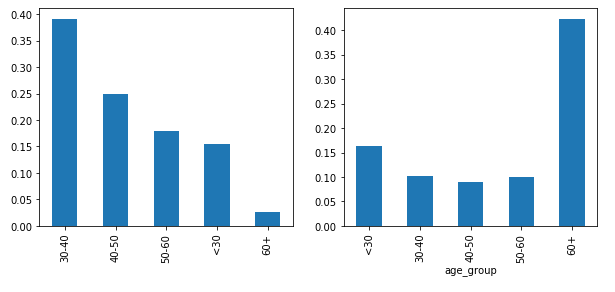


As you can see in the box plot (already explained in the bank marketing dataset) above, customers with a higher salary range are more likely to give a positive response.

* **Correlation vs causation:** This is a very important concept of data anaylsis, which states that correlation is not always related to causation. Although there may be a very high correlation between variables, there may be no causation at all.

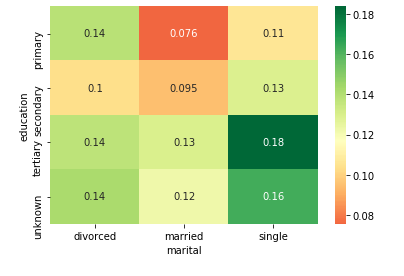


* **Analysis between two categorical variables:** A **bar graph** is the best approach to analysing two categorical variables.



One of the interesting examples, also covered in the bank marketing dataset, is that the bank has mostly contacted people in the age group of 30-50, although people in the age group of 60+ gave more positive responses among all the age groups. This is a very important inference that the bank can draw, i.e., it should contact more individuals in the age group of 60+.

* **Multivariate analysis:** Multivariate analysis yields very specific information about a data set. It basically involves the analysis of more than two variables at a time. For instance, **heat maps** are the best way to look at three variables at a time. In multivariate analysis, it is essential to look into the data by grouping the variables and infer decisions from it.



As you have seen already in the bank marketing case study, single people with tertiary education are more likely to give a positive response to term deposit. And married individuals and those who have completed up to primary education are least likely to give a positive response.

In the next segment, Rahim will summarise the module on Exploratory Data Analysis.