**UNDERSTANDING DATA VISUALIZATION**

More than 90% of the world’s data has appeared in the past 2 years. We are moving from digitalization to datafication. Real time data is used to make both important and non-important decisions in various sectors of the world like the finance sector, medical sector, mechanical sector and the business sector. It is with data that we better understand the world, and that means we can make better decisions.

In an increasingly data-driven world, the need to make data more accessible and understandable is more important than ever.

Data literacy is the ability to read, work with, analyze and argue with data. And to do this we need to understand the data we have which can be achieved by visualization.

Visualization is a way of representing information and data by using visual tools like graphs, maps and charts to understand the patterns, trends and outliers present in the data. It provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

Using charts or graphs to visualize vast amounts of complex information is more straightforward than digging spreadsheets or reports.Visualization is a preferred way to understand the data as our eyes are drawn to colours and patterns easily. Our eyes can easily grab the trend, outliers and patterns in the data when we see a graph or a chart. Some advantages of data visualization include:

- Easily sharing information

- Interactively explore opportunities

- Visualize patterns and relationships

Data analysis is a methodical approach of applying statistical measures to describe, analyze, and evaluate data. The most common types of analysis done on data:

- Univariate analysis

- Bivariate analysis

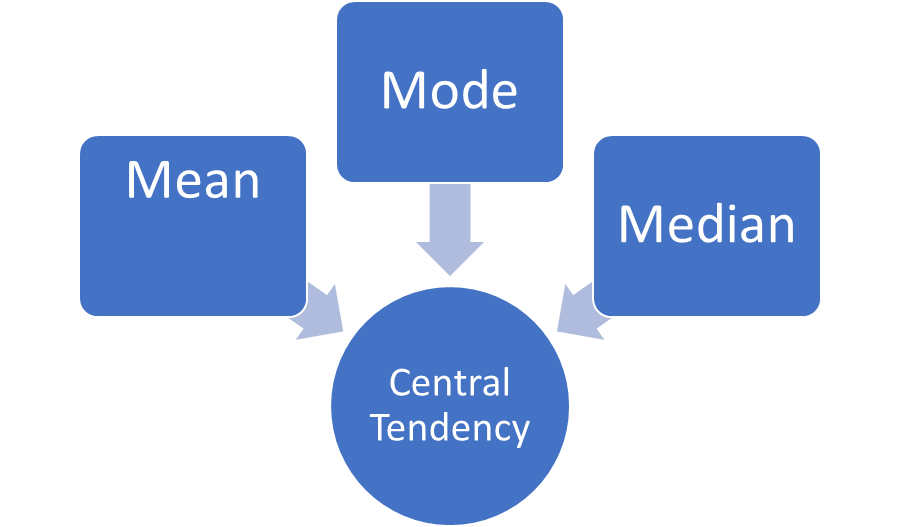
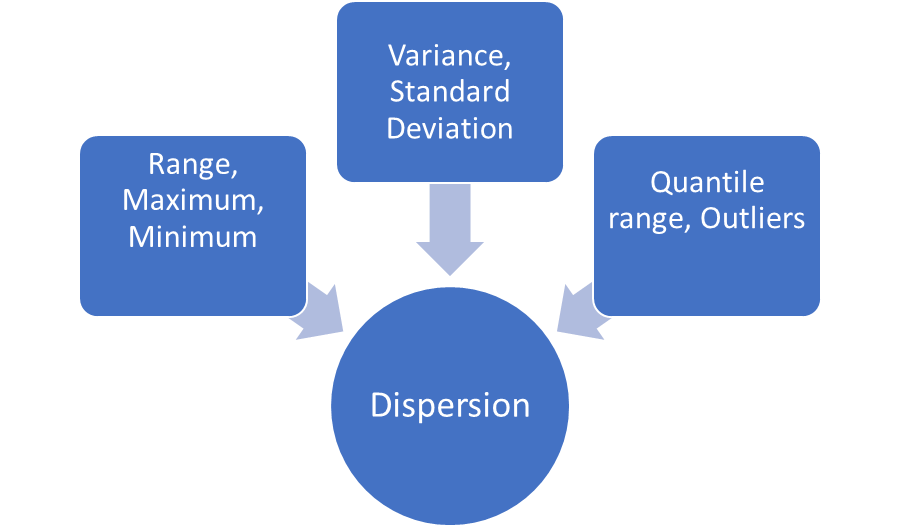
- Multivariate analysis

**UNIVARIATE ANALYSIS** (*Uni:one*)

**Univariate analysis** is the simplest form of analyzing data. Univariate analysis refers to analysis of one variable.

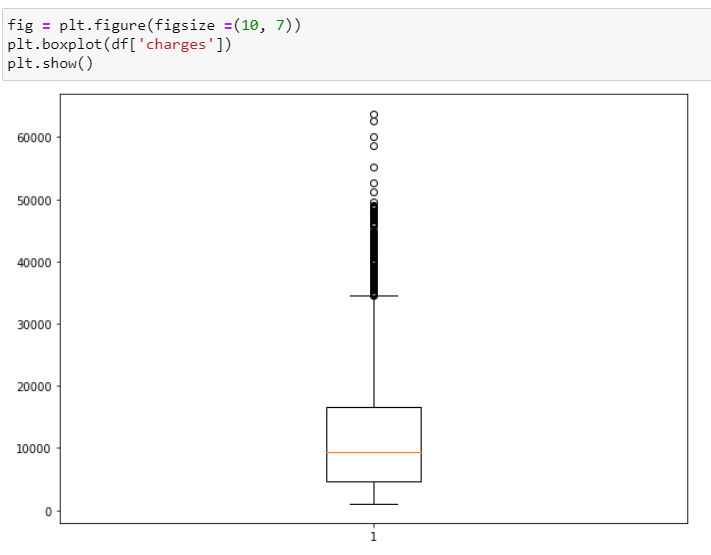
The main objective of the univariate analysis is to describe the data in order to find out the patterns,trends and outliers in the data.

1. **Summary statistics**: The most common way to perform univariate analysis is to describe a variable using [summary statistics](https://www.statology.org/descriptive-inferential-statistics/).The following are the two most common types of summary statistics: - - -**Measures of central tendency**: these numbers describe where the center of a dataset is located. -**Measures of Dispersion**: these numbers describe the spread of the data or how the data varies from the central point.

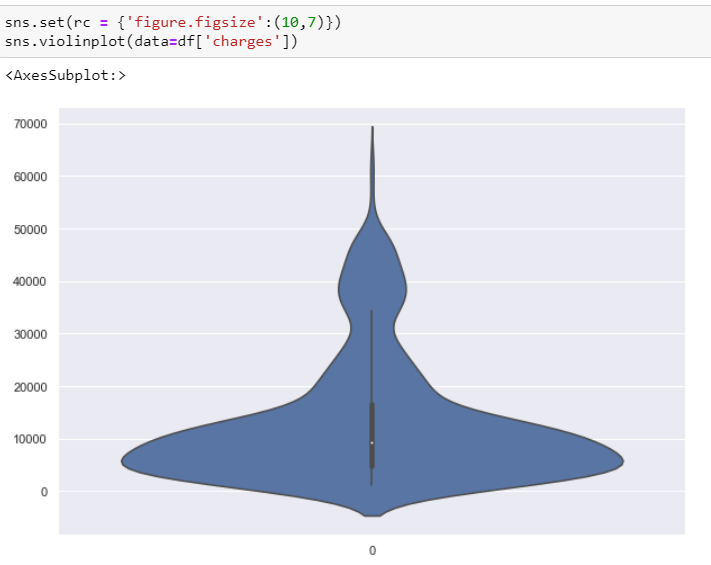
Measures of Dispersion Measures of Central Tendency

1. **Frequency distributions** is a way to understand how often different values occur in a dataset. These come in handy in the case of categorical variables.
2. **Visualization:** Another way to perform univariate analysis is to create charts to visualize the distribution of values for a certain variable.

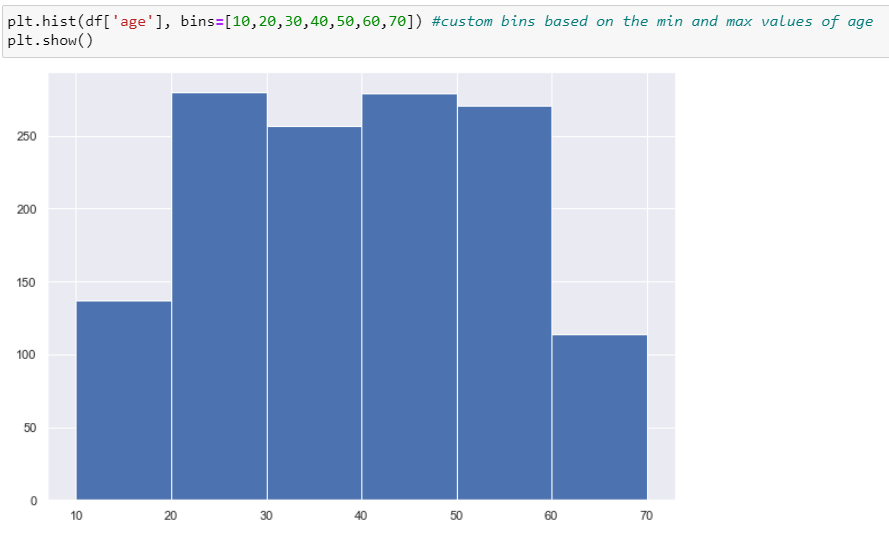
* **Boxplot:**
  + Helps us understand the distribution of the data.
  + Shows 5 number summary of the dataset i.e., minimum value, first quantile, median, third quantile, maximum value.
  + It helps visualize outliers.



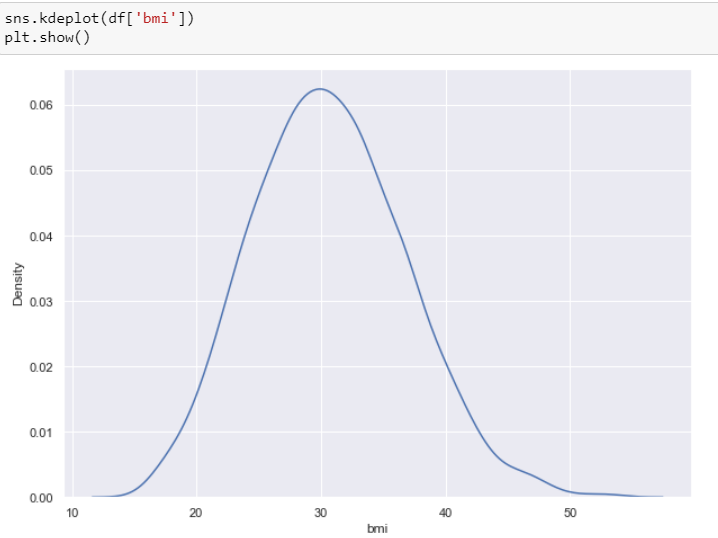
* **Violin plots:**
  + It is a combination of boxplot and density plot.
  + It shows us the distribution shape of the data.



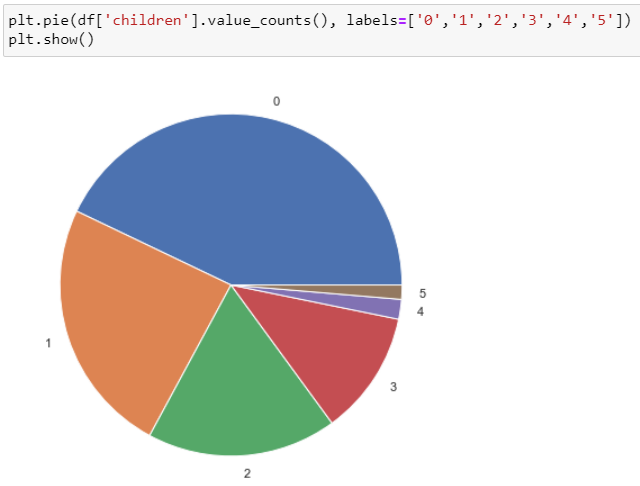
* **Histogram:**
  + A histogram is a type of chart that uses vertical bars to display frequencies.
  + This type of chart is a useful way to visualize the distribution of values in a dataset.
  + It also helps us visualize the existence of outliers.



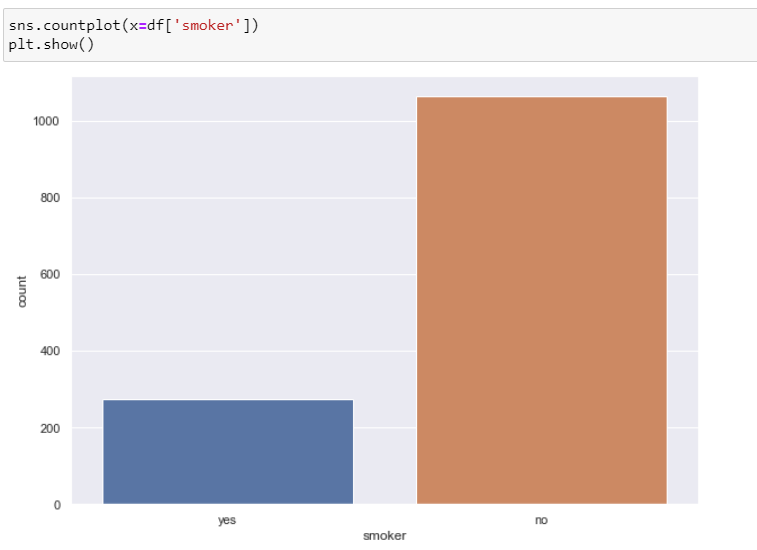
* **Density plots:**
  + A density plot is a representation of the distribution of a numeric variable.
  + Also called Kernel density plot(hence kdeplot).



* **Pie chart:**
  + Helps us understand distribution for categorical columns.
  + Pie Chart is a graph that displays the data in a circular format.
  + It is a type of pictorial representation of data that shows the relative sizes through the slices of pie.



* **Countplot:**
  + Count plot helps us to visualize the frequency of the categorical columns.



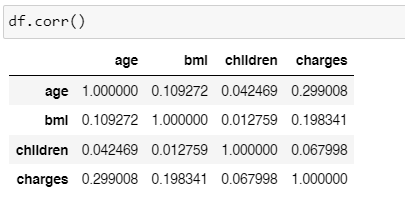
**BIVARIATE ANALYSIS***(bi:two)*

It is the analysis of 2 variables. It helps us determine the relationship between the two variables.

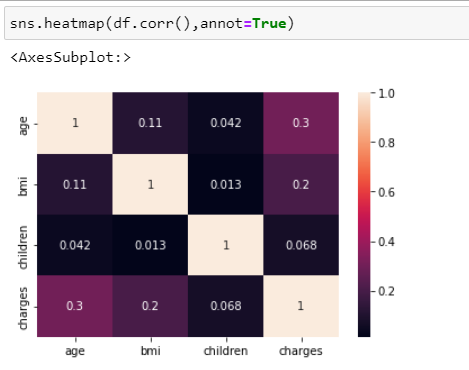
1. **Correlation:** Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

* 1 indicates a strong positive relationship.
* -1 indicates a strong negative relationship.
* A result of zero indicates no relationship at all.

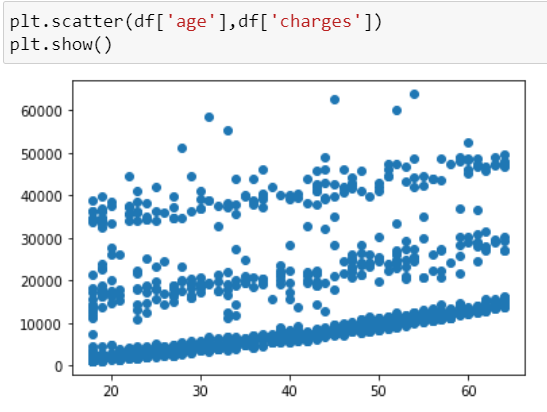
A correlation matrix showing all the correlation values can be obtained as below:



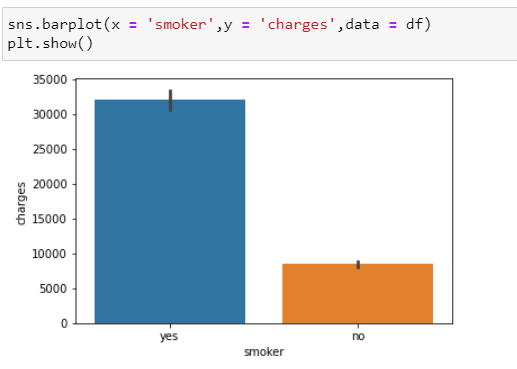
* **Heat map:** The correlation matrix mentioned above can be visualized with the help of a heat map. A heatmap is a graphical representation of data that uses a system of color-coding to represent different values. A heat map uses a warm-to-cool color spectrum to show us the correlation in the dataset



* **Scatter Plot:** Scatter plots are used when both the variables are numerical and it helps us understand the relation.

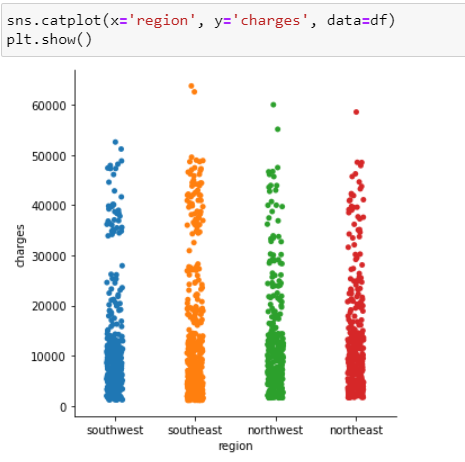


* **Bar Plot:** Barplots are used when we want to find the relation between one categorical and one numerical variable.The height of each rectangle gives us an estimate of the central tendency for a numeric variable.



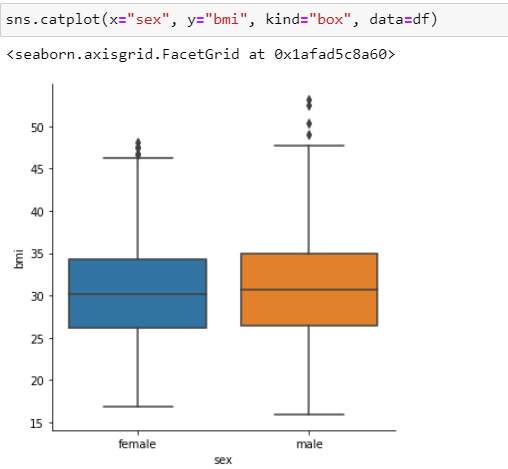
* **Cat Plot:** Shows relationship between a numerical and categorical column.

By default, the visual representation will be a jittered(similar to scatter) strip plot.

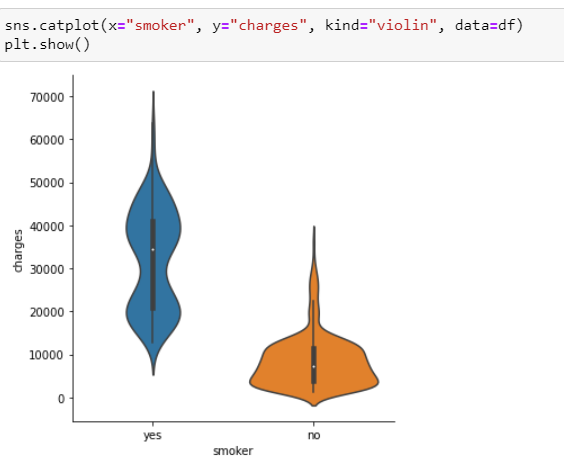


By adding an extra keyword argument to catplots i.e., kind=”box”/”violin”, we can get the below plots to understand the relations better.

* **Box plot using cat plot:**



* **Violin plot using cat plot:**



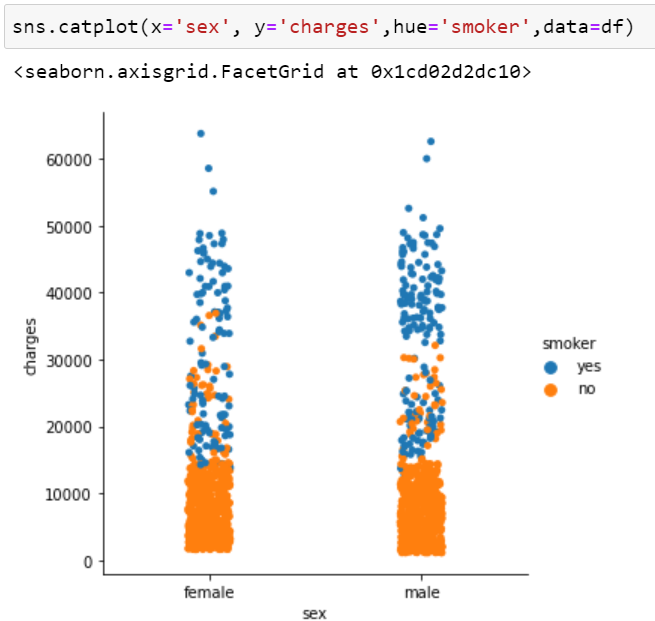
**MULTIVARIATE ANALYSIS**

Multivariate analysis is the simultaneous observation and analysis of more than one variable.There are more than 20 different methods to perform multivariate analysis and which method is best depends on the type of data and the problem you are trying to solve.

We are going to focus on some simple visualizations that can be used for multivariate analysis.

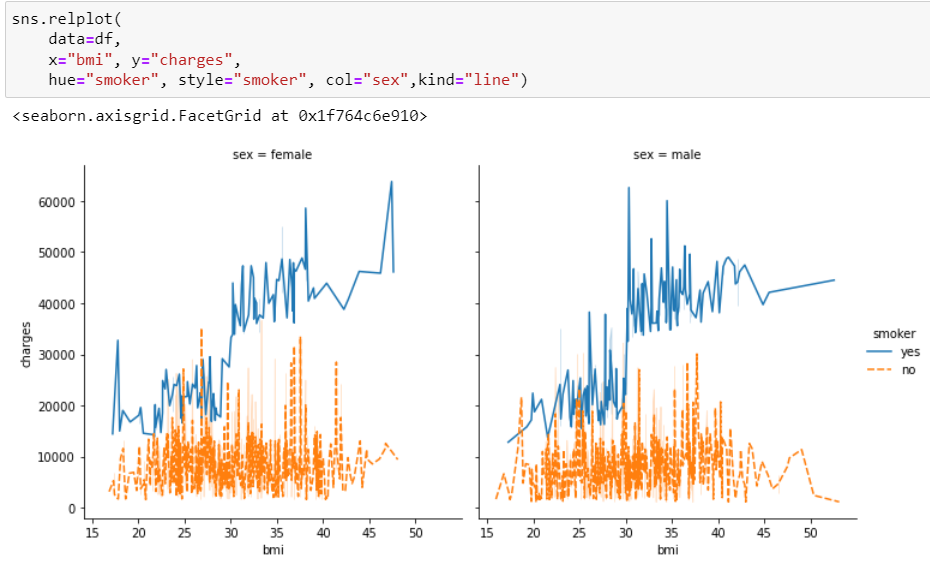
Plots like relpot and catplot from the seaborn library can be used for multivariate analysis by using different subsets of the data like hue, size, and style parameters. These parameters control what visual semantics are used to identify the different subsets.

* **Catplot:**



* **Relplot:**





* **Pairplot:**

A pairplot along with one column as hue can be used to understand the relation of 3 variables



**TOOLS FOR DATA VISUALIZATION:**

There are a number of tools available to help us with visualisations. Our focus for this blog has been Python. These are some of the widely used and easy to understand tools:

**Python:** Python has multiple libraries for data visualization. Two of the majorly used libraries are: matplotlib and seaborn.

**Tableau and Power BI:** These are data visualization tools that help us visualize and represent data in the form of graphs, dashboards, etc.

**Microsoft excel:** It is a data visualization tool that has an easy interface, so it doesn’t have to be difficult to work with.

Github link with the coding for these visualizatuins:

**The csv file and the details of the data used for explanation of concepts can be found here:** [**https://www.kaggle.com/datasets/mirichoi0218/insurance**](https://www.kaggle.com/datasets/mirichoi0218/insurance) **(insurance data)**

**Checkout the matplotlib and seaborn documentation to explore the visualizations provided by these libraries:**

[**https://matplotlib.org/stable/plot\_types/index.html**](https://matplotlib.org/stable/plot_types/index.html)

[**https://seaborn.pydata.org/api.html**](https://seaborn.pydata.org/api.html)