Hands On: Exploratory Data Analysis with Python I

Thu, 27th April 2023

Data Analytics

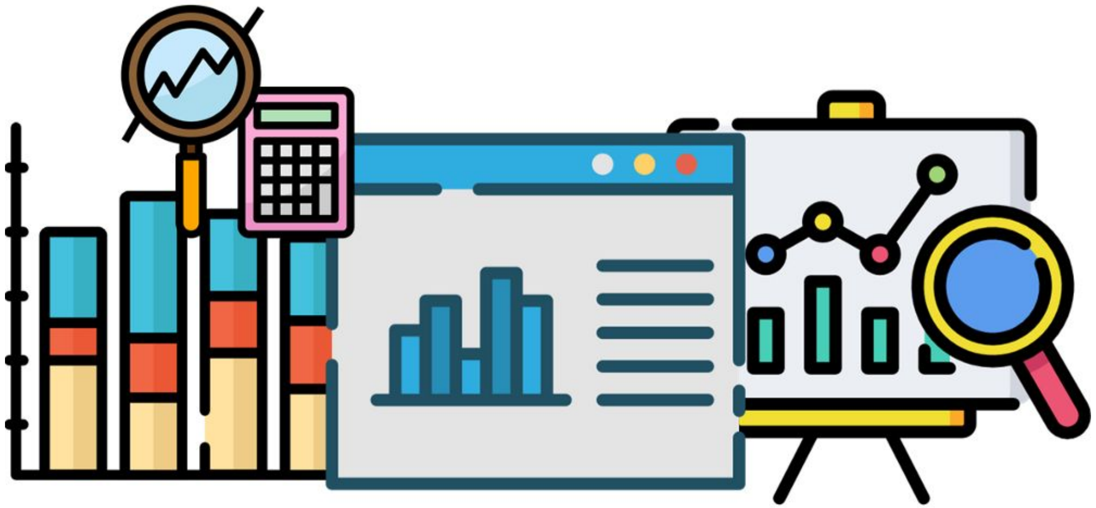
Program Zenius Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka

1. Exploratory Data Analysis Definition & Purpose
2. Data Understanding
3. Data Cleaning
4. Univariate & Bivariate Data Analysis
5. Visualization for Exploratory Data Analysis

Exploratory Data Analysis

Exploratory Data Analysis

an approach to analyze the data to discover trends, patterns, or to find assumptions with the help of statistical summary and graphical representations



Why is EDA important?

- Helps identify errors in data sets.
- Gives a better understanding of the data set.
- Helps detect outliers or anomalous events.
- Helps understand data set variables and the relationship among them.

Exploratory Data Analysis: Data Understanding

Data Understanding

- First step of EDA is to understand the current dataset which includes the knowledge you have about data, the needs the data will satisfy, its content and location
- Exploring data glossaries such as data dictionaries, metadata and any information about the data will ease the data understanding
- The step is crucial to seek a better context and value of current data and business problem

Data Understanding

In Python, there are several functions which can be used on dataframe to better understand the data. Those functions are :

Functions	
read_csv(), read_excel(), read_json	Read csv/excel/json file and transform into a dataframe
head()	return *n* rows from the top of dataframe (default is 5)
shape()	returns the number of rows by the number of columns for my dataset.
columns	returns the name of all of your columns in the dataset.
describe()	summarizes the count, mean, standard deviation, min, and max for numeric variables
nunique()	returns the number of unique values for each variable.
unique()	Returns all unique value for each column

Exploratory Data Analysis: Data Cleaning

Data Cleaning

Insights and analysis are only as good as the data you are using, Garbage in garbage out.

Thus, data cleaning is essential as part of exploratory data analysis

Data Cleaning

You can perform data cleaning on your dataset by following these steps :

1. Remove duplicate or irrelevant observations

- Duplicated values happens mostly during data collections.
- When you combine data from multiple sources, department, clients, there are the risk of getting duplicated data.
- Let's say we are going to analyze data from millennials, we need to exclude elderly data from the dataset to remove irrelevant data

Data Cleaning

You can perform data cleaning on your dataset by following these steps :

2. Fix inconsistent structure and errors

- **Structural error consist of strange naming convention, typos, incorrect capitalization**
- **i.e *San Fransisco and SF, N/A and blank***

Data Cleaning

You can perform data cleaning on your dataset by following these steps :

3. Outlier Handling

- Outliers will introduce variability in data which will decrease statistical power, it mostly happens because of improper data input.
- If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.It will improve the performance of data you are working with
- However, sometimes it is the appearance of an outlier that will prove a theory you are working on

Data Cleaning

You can perform data cleaning on your dataset by following these steps :

4. Handle Missing data
Missing data can not be ignored. Most analysis and algorithms are not able to interpret empty value.
You can drop missing values or alter its value into assumptions such as mean, modes

Exploratory Data Analysis: Data Visualizing

Data Visualizing

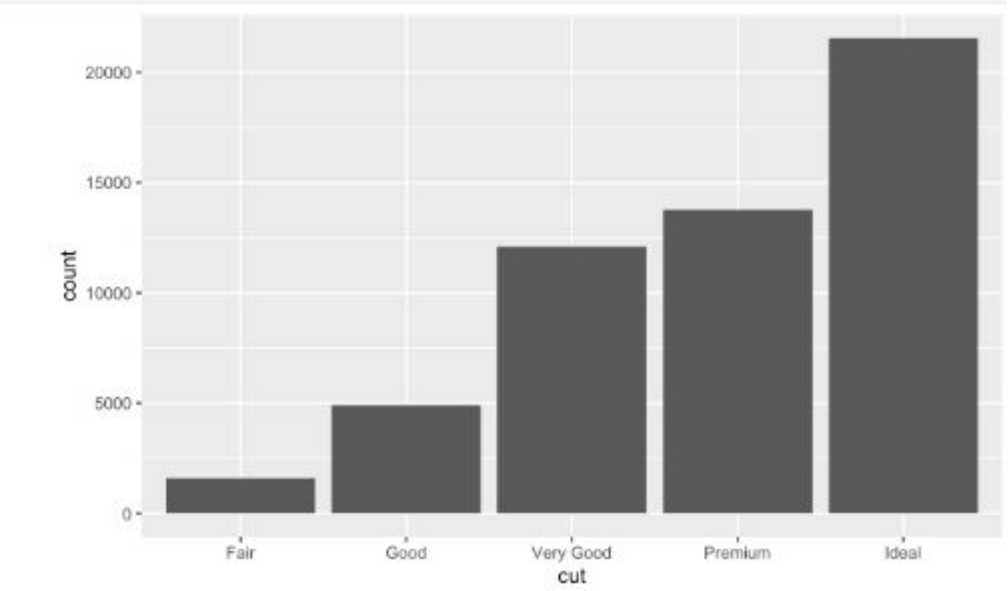
- After the data cleaning process, We can start generating insight by visualizing the dataset.

- How you visualize the variable will depend on whether the variable is **categorical or continuous**

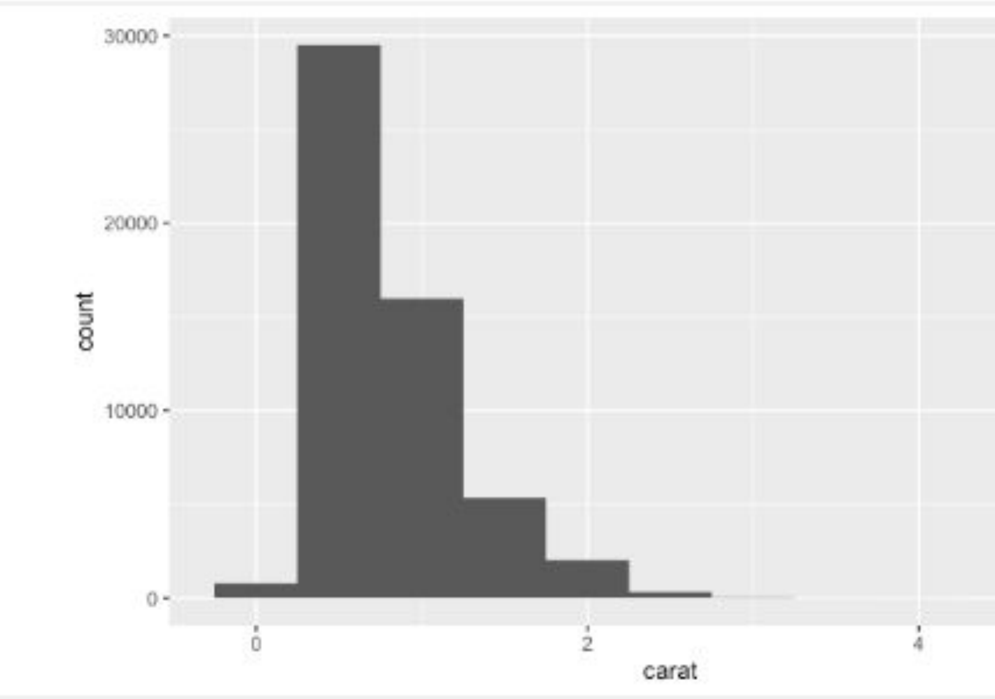
- Categorical : Gender, Product Type, Post Code, Location
- Continuous : Wages, Product Price, Quantity, Temperature

Data Visualizing : Variable Type

Categorical



Continuous



Univariate & Multivariate Analysis

- Univariate Analysis: summarize only <u>**one variable**</u> at a time. i.e Weight, Speed.
    - Histograms, Pie Chart, Bar Chart

- Bivariate Analysis : compare <u>**two variables**</u>. i.e Sales & Temperature.
    - Correlation Coefficient, Scatterplot, Heatmap

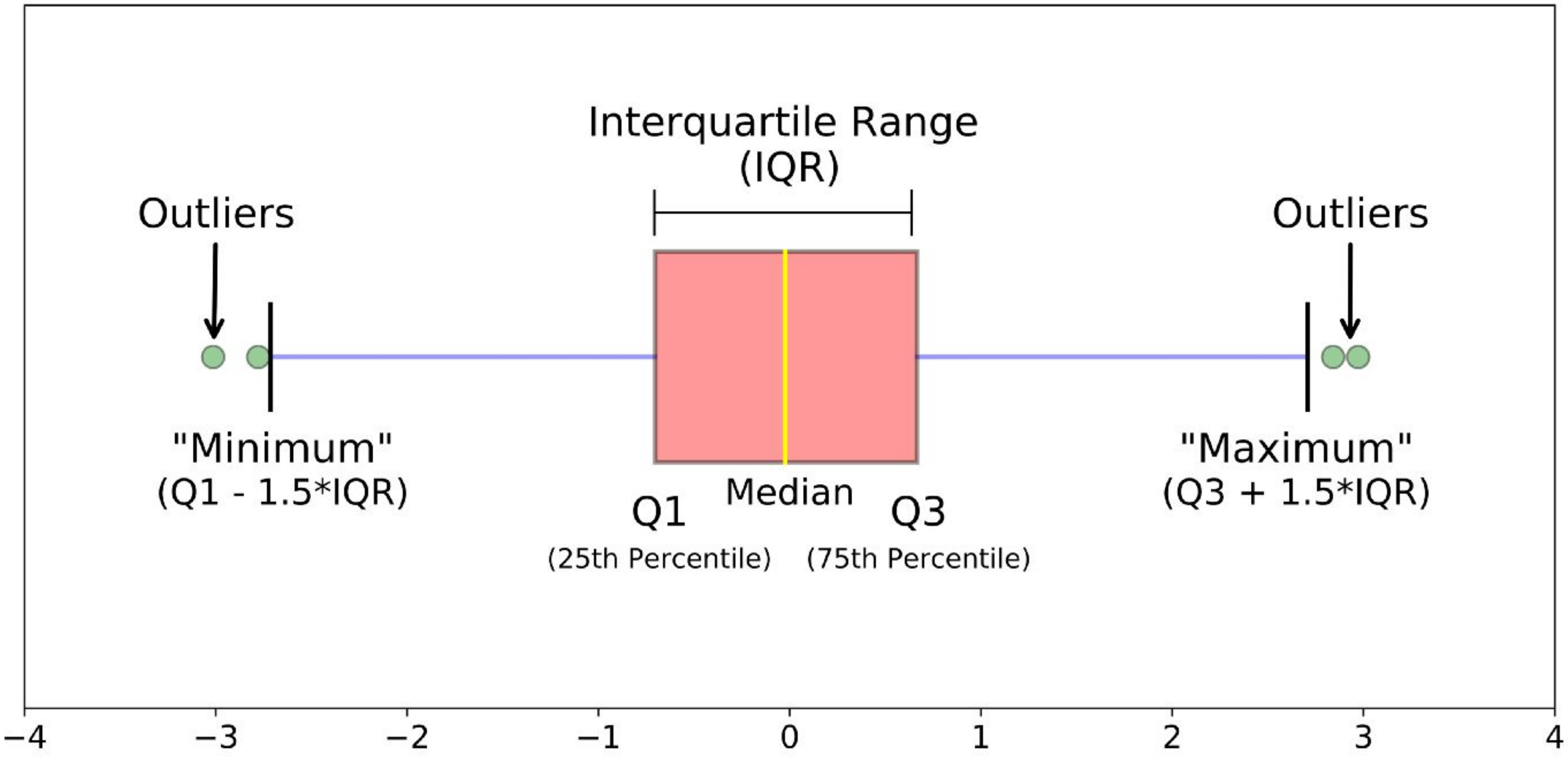
Univariate Analysis : Descriptive Statistics

- **Central Tendency**: Central tendency refers to the location of a distribution. It represents the typical value that normally expect from the variable. We can describe the central tendency by using these :
    - **Mean** : Average of value in the distribution
    - **Median (Q2)** : The Fiftieth percentile of a distribution
    - **Mode** : Most frequent value that occurs in the distribution

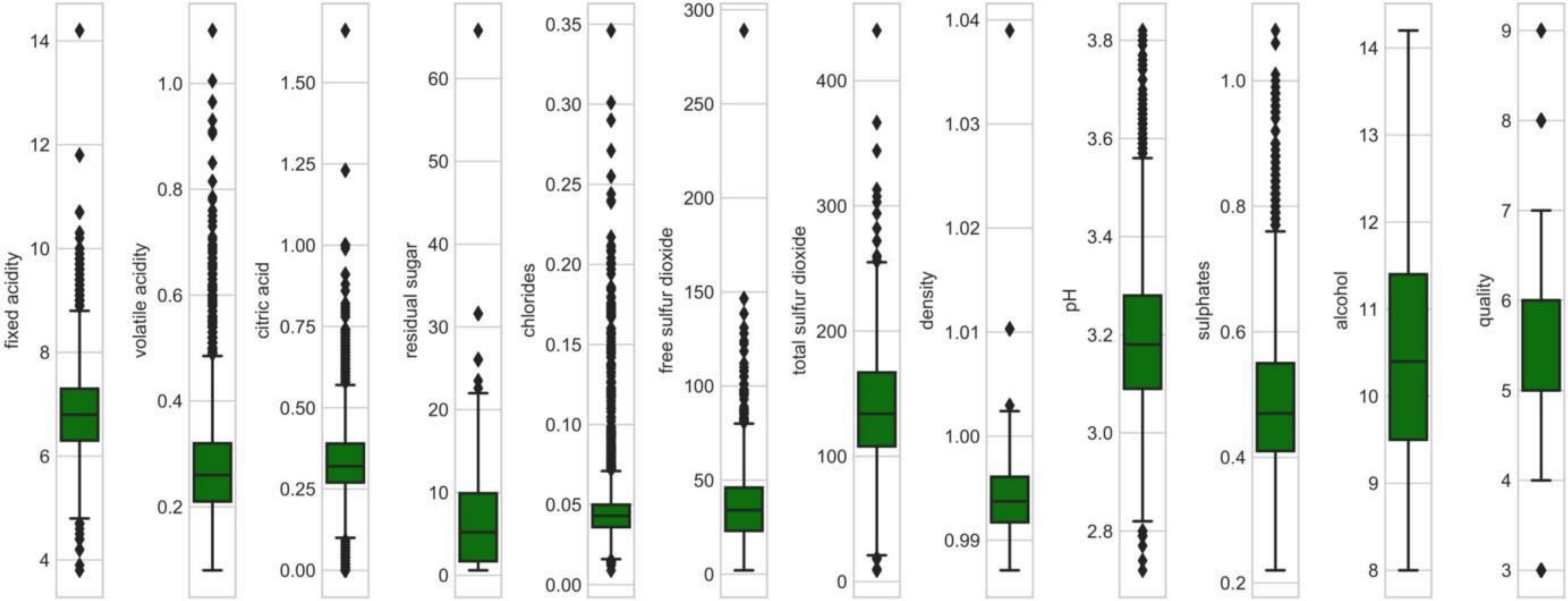
Univariate Analysis : Descriptive Statistics

- **Variance:**
- refers to the variance of a distribution.
    - **Range** : Difference between highest and lowest value
    - **Standard Deviation**: How much a value of each variables differ from the mean of the group
    - **Interquartile Range** : Range between third quartile (Q3) and first quartile (Q1)

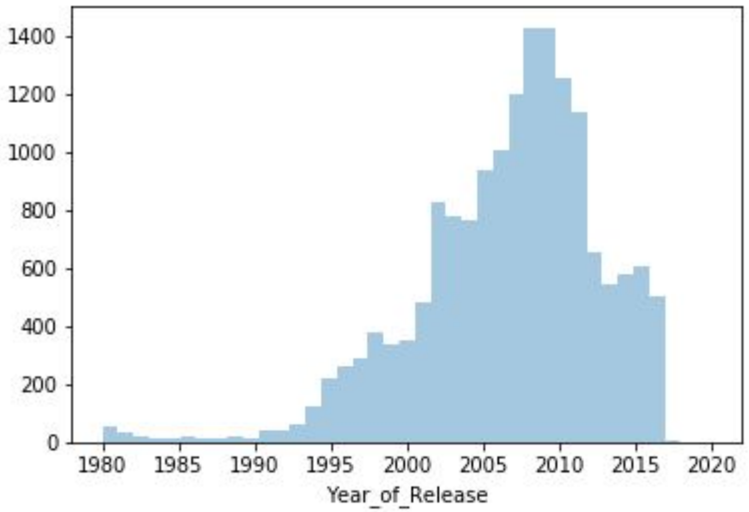
Univariate Analysis : Boxplot



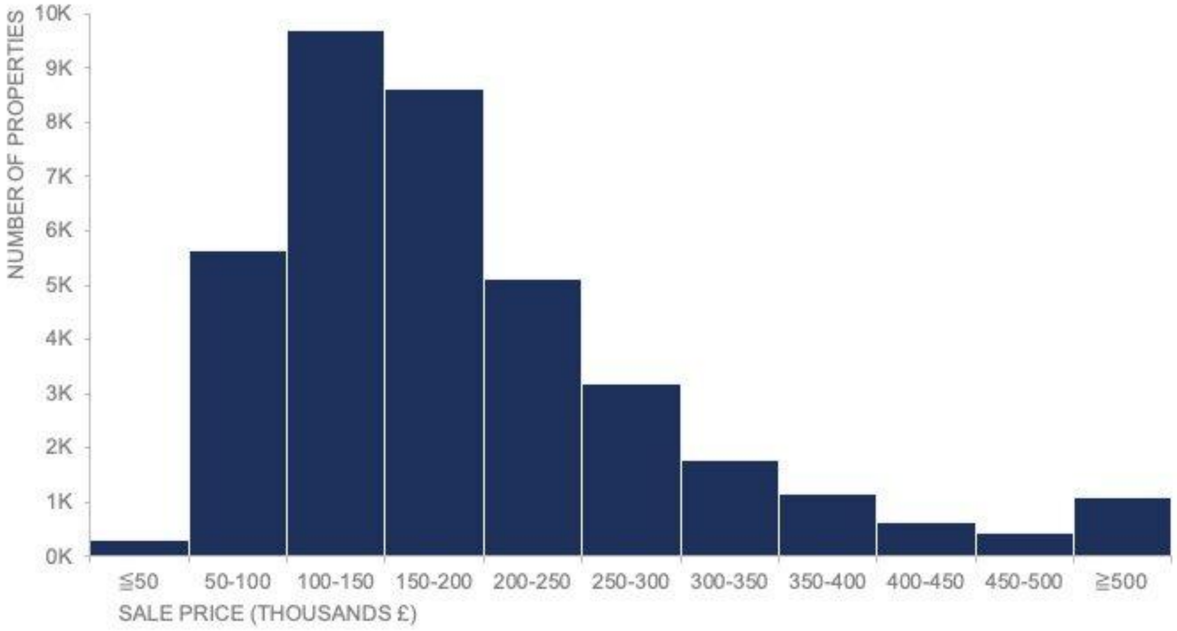
Univariate Analysis : Boxplot



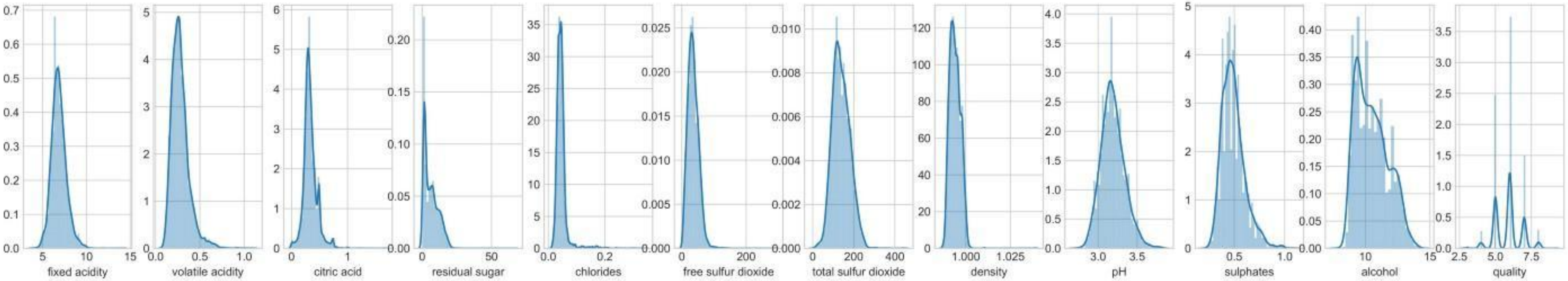
Univariate Analysis: Histogram



Distribution of property sales: January 2013 to September 2019



Univariate Analysis : Distplot

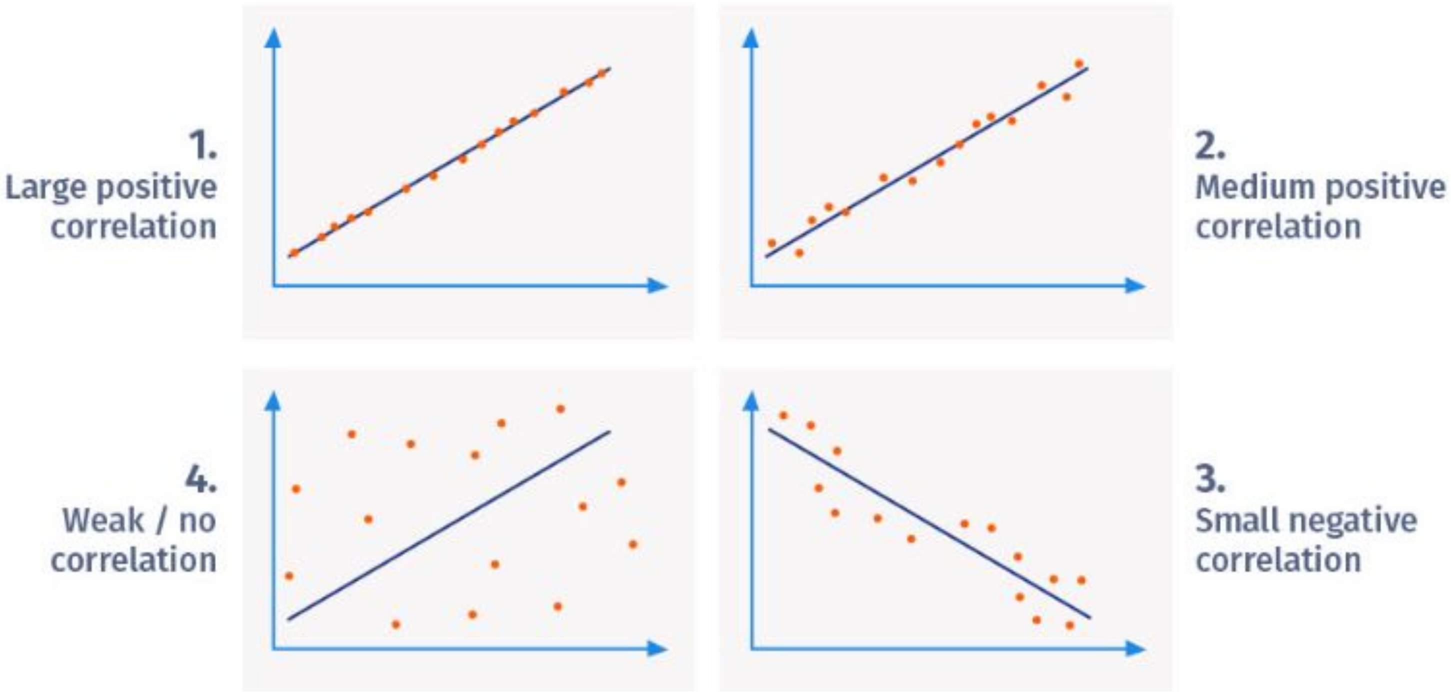


Bivariate Analysis : Relationship

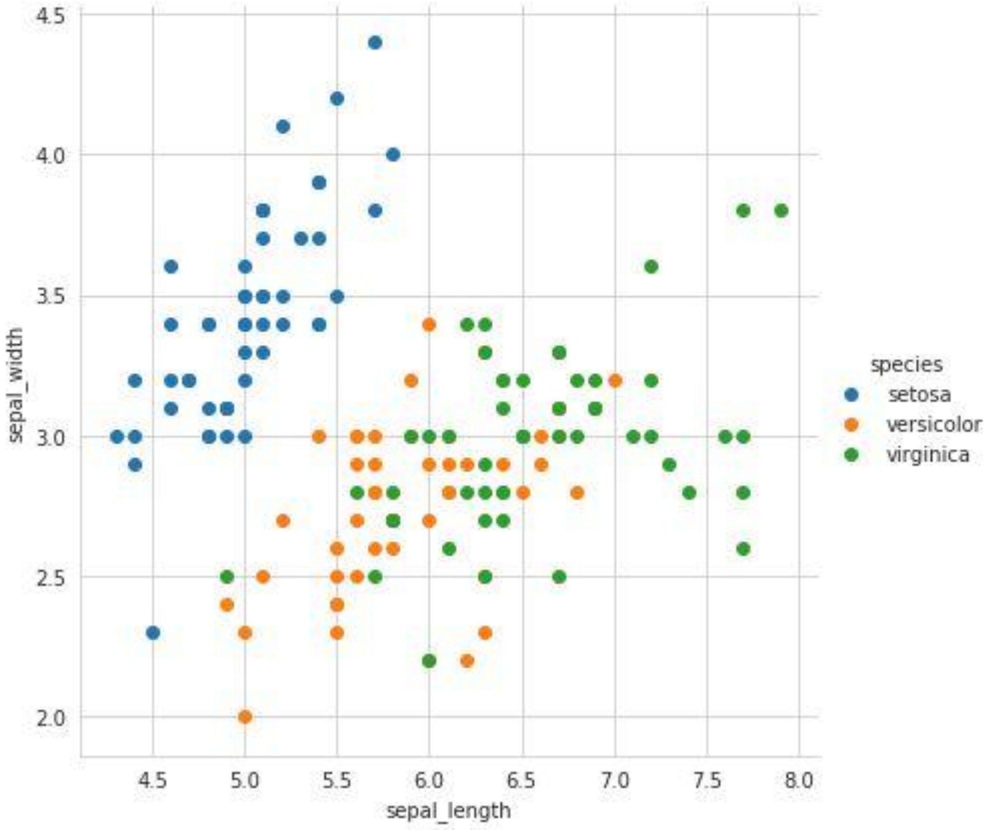
During EDA, we learn about the relationship between two variables. Some questions to answers are:

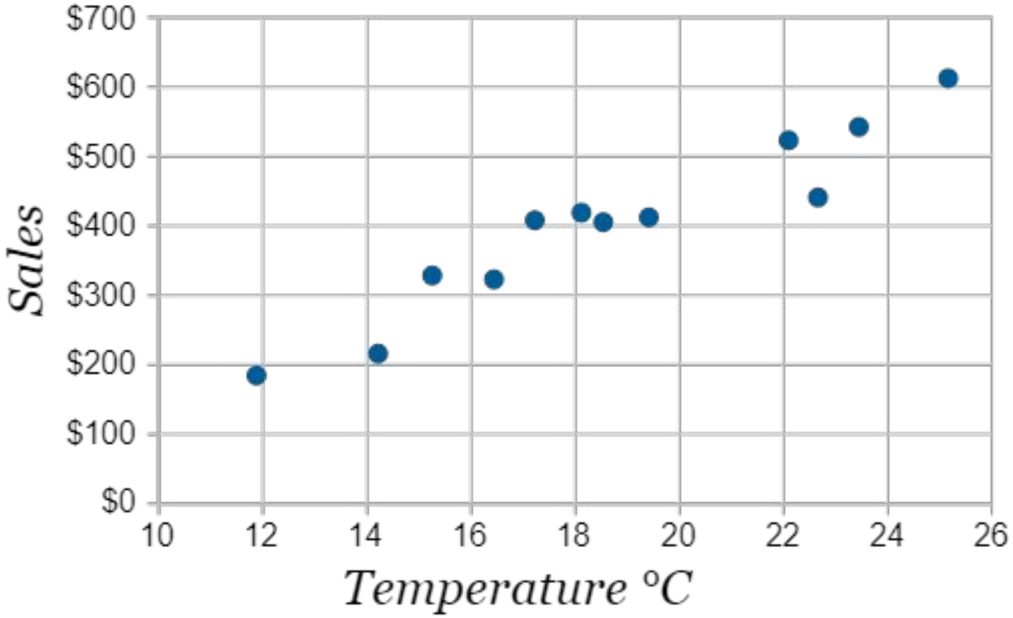
1. How correlated is one variable with another variable?
2. Does lower value on one variable corresponds to a lower value on another variable?
3. Does higher value on one variable corresponds to a higher value on another variable?
4. What type of relationship do the two features follow?

Bivariate Analysis : Pearson Correlation

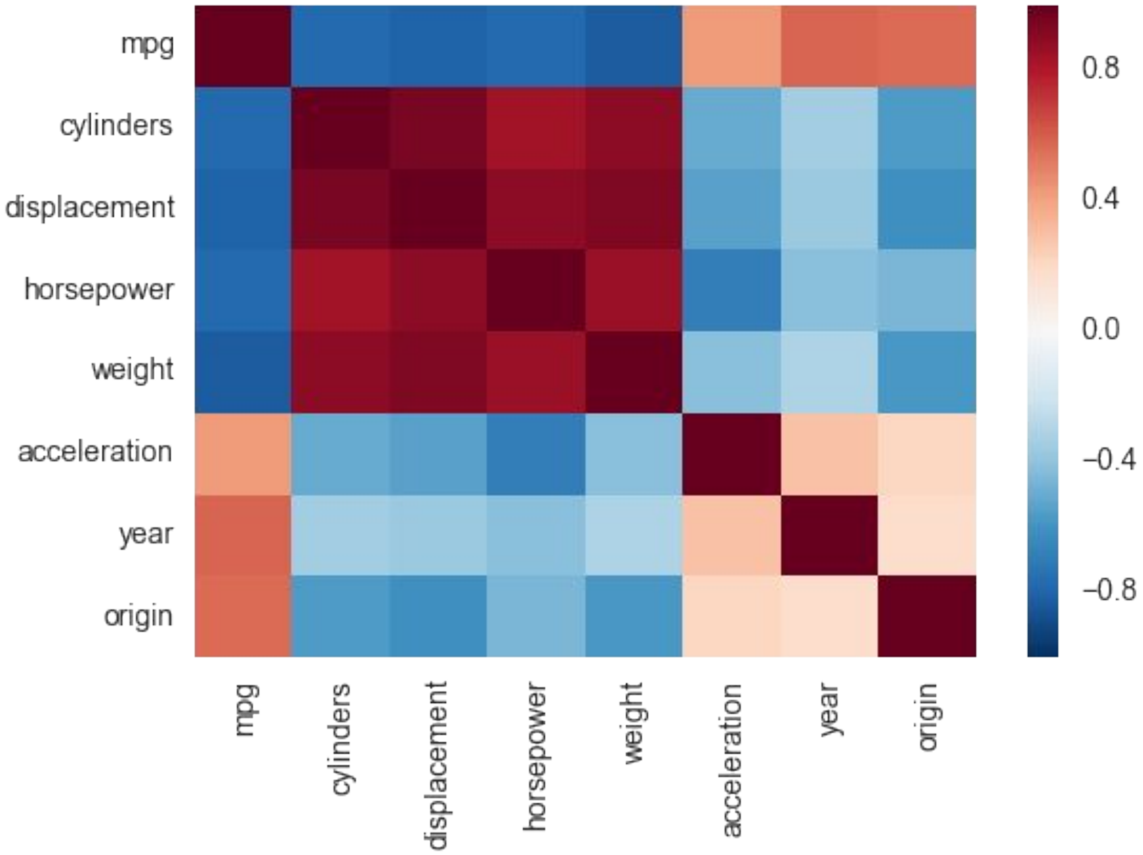


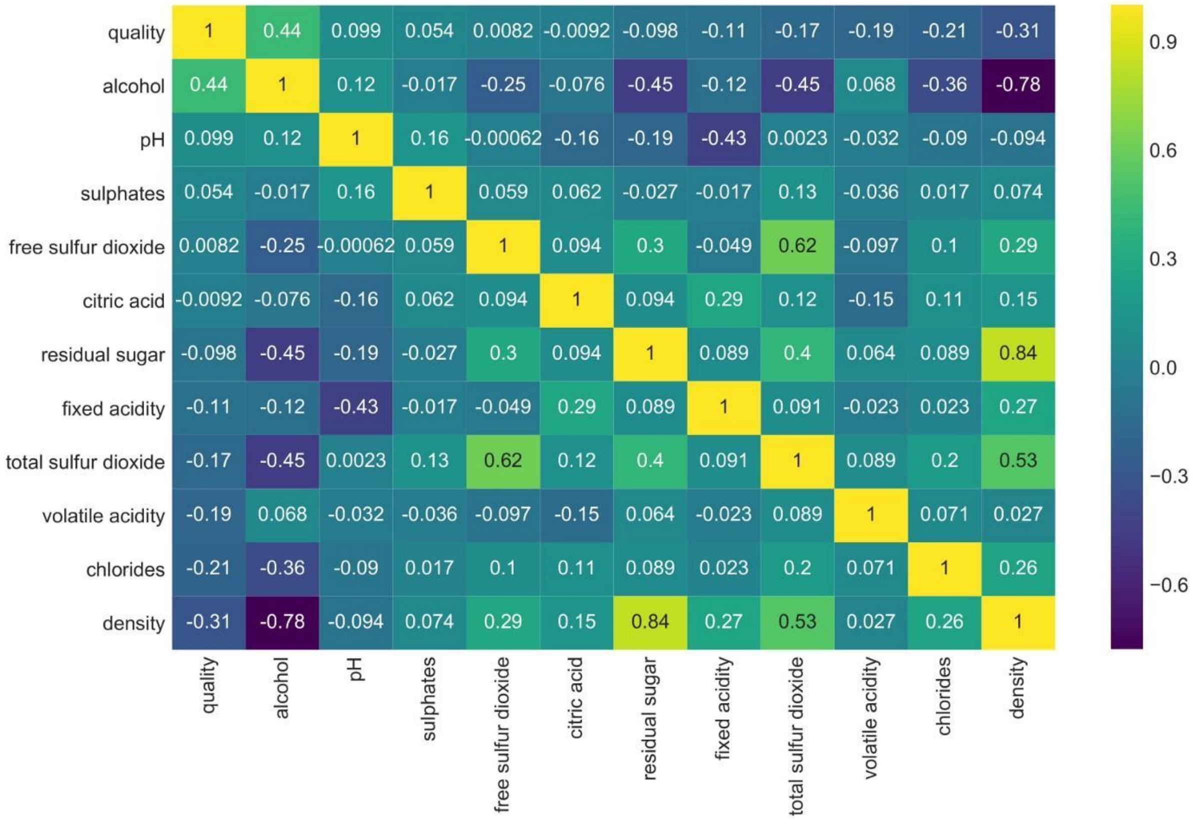
Bivariate Analysis : Scatterplot





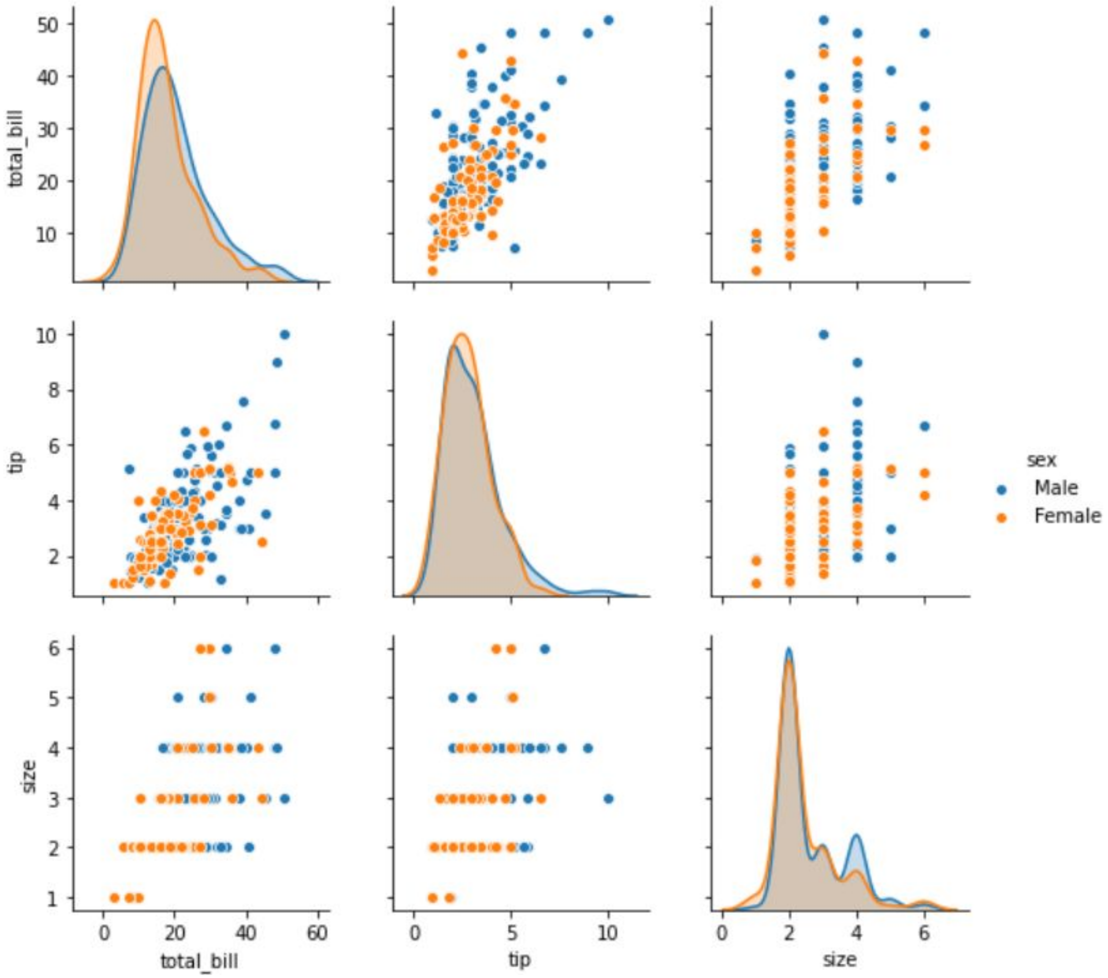
Bivariate Analysis : Heatmap





	quality	alcohol	pH	sulphates	free sulfur dioxide	citric acid	residual sugar	fixed acidity	total sulfur dioxide	volatile acidity	chlorides	density
quality	1	0.44	0.099	0.054	0.0082	-0.0092	-0.098	-0.11	-0.17	-0.19	-0.21	-0.31
alcohol	0.44	1	0.12	-0.017	-0.25	-0.076	-0.45	-0.12	-0.45	0.068	-0.36	-0.78
pH	0.099	0.12	1	0.16	-0.00062	-0.16	-0.19	-0.43	0.0023	-0.032	-0.09	-0.094
sulphates	0.054	-0.017	0.16	1	0.059	0.062	-0.027	-0.017	0.13	-0.036	0.017	0.074
free sulfur dioxide	0.0082	-0.25	-0.00062	0.059	1	0.094	0.3	-0.049	0.62	-0.097	0.1	0.29
citric acid	-0.0092	-0.076	-0.16	0.062	0.094	1	0.094	0.29	0.12	-0.15	0.11	0.15
residual sugar	-0.098	-0.45	-0.19	-0.027	0.3	0.094	1	0.089	0.4	0.064	0.089	0.84
fixed acidity	-0.11	-0.12	-0.43	-0.017	-0.049	0.29	0.089	1	0.091	-0.023	0.023	0.27
total sulfur dioxide	-0.17	-0.45	0.0023	0.13	0.62	0.12	0.4	0.091	1	0.089	0.2	0.53
volatile acidity	-0.19	0.068	-0.032	-0.036	-0.097	-0.15	0.064	-0.023	0.089	1	0.071	0.027
chlorides	-0.21	-0.36	-0.09	0.017	0.1	0.11	0.089	0.023	0.2	0.071	1	0.26
density	-0.31	-0.78	-0.094	0.074	0.29	0.15	0.84	0.27	0.53	0.027	0.26	1

Bivariate Analysis : Pairplot



Assignment

Assignment 6

Instruksi Assignment 6

- Lengkapi dan kerjakan kembali syntax yang dibuat pada saat **Hands On: Exploratory Data Analysis with Python I dan II**.
- Pelajari syntax yang dibuat pada saat **Hands On: Exploratory Data Analysis with Python I dan II** dengan cara :
  - membuat summary terkait langkah-langkah yang dikerjakan
  - membuat storytelling terkait data analysis yang sudah dikerjakan dan hasil visualisasi
  - (Dikerjakan pada ipynb dan diletakkan diatas code yang sudah ditulis - dapat ditulis sebagai command atau text cell - agar lebih mudah dinilai)

Kumpulkan link google colab dan beri nama notebook dengan format **Topik 11 12 - [Nama Lengkap]**

Available from	Until
Apr 27 at 09.00 PM	May 6 at 18.00PM

Thank you!

Any Questions?