EDA

Exploratory Data Analysis-1 (EDA-1)

- Data Cleaning,
- Imputation Techniques,
- Data Analysis and Visualization(Scatter Diagram, Correlation Analysis),
- Transformations
- Auto EDA libraries

EDA

1) **Describe a dataset**: Number of rows/columns, missing data, data types, preview.
2) **Clean data :** Handle missing data, invalid data types, incorrect values and outliers
3) **Visualize data distributions:** Bar charts, histograms, box plots.
4) **Calculate and visualize: C**orrelations (relationships) between variables, Heat map

Data Cleaning

Data Cleaning

Data cleaning or cleansing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

Data Quality:

- Validity,
- Accuracy,
- Completeness,
- Consistency,
- Uniformity.

Validity

- **Data-Type Constraints:** values in a particular column must be of a particular datatype, e.g., boolean, numeric, date, etc.
- **Range Constraints:** typically, numbers or dates should fall within a certain range.
- **Mandatory Constraints**: certain columns cannot be empty.
- **Set-Membership constraints:** values of a column come from a set of discrete values,. For example, Blood Groups – Fixed set of discrete values

Accuracy

The degree to which the data is close to the true values.
While defining all possible valid values allows invalid values to be easily spotted, it does not mean that they are accurate.
A *valid* street address mightn't actually exist.

Another thing to note is the difference between accuracy and precision. Saying that you live on the earth is, actually true. But, not precise. Where on the earth?. Saying that you live at a particular street address is more precise.

Consistency & Uniformity

The degree to which the data is consistent, within the same data set or across multiple data sets.
Inconsistency occurs when two values in the data set contradict each other.

A valid age, say 3, mightn't match with the marital status, say divorced.
A customer is recorded in two different tables with two different Genders.
Which one is true?.

The degree to which the data is specified using the same unit of measure.
The weight may be recorded either in pounds or kilos. The date might follow the USA format or European format. The currency is sometimes in USD and sometimes in Euros.
And so data must be converted to a single measure unit.

Outliers

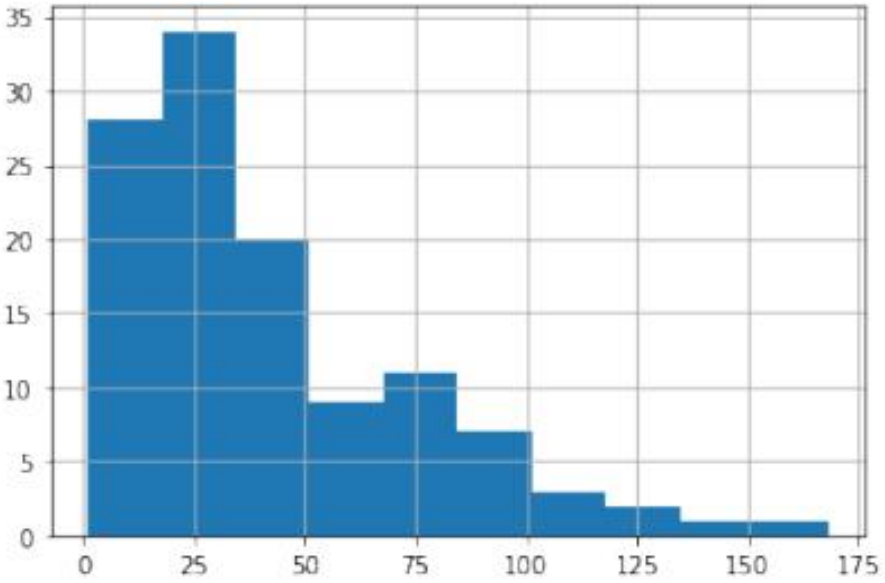
Outliers are data that is distinctively different from other observations. They could be real outliers or mistakes.

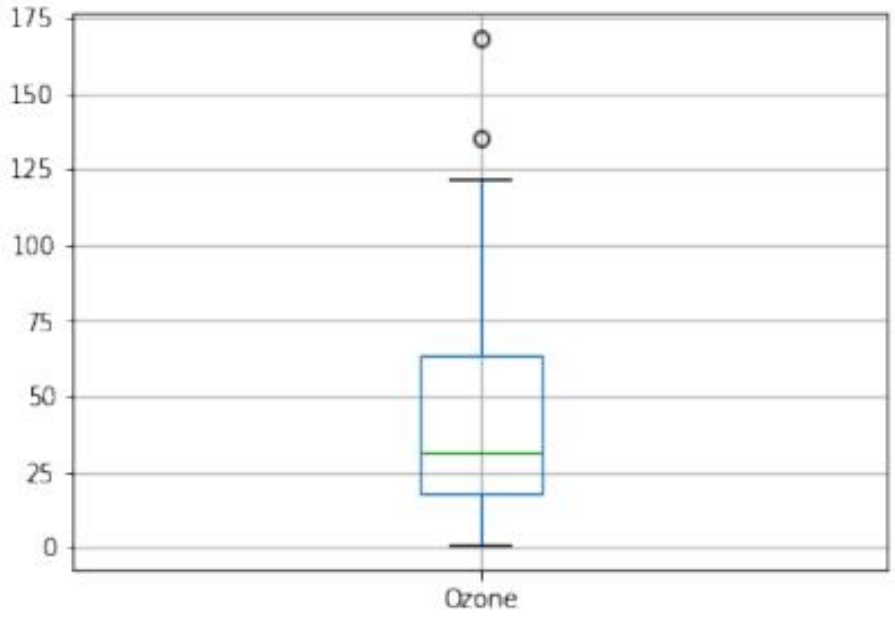
How to find out?
Depending on whether the feature is numeric or categorical, we can use different techniques to study its distribution to detect outliers.

Outliers

Histogram/Box Plot:

When the feature is numeric, we can use a histogram and box plot to detect outliers. From Histogram, if the data is highly skewed then there is a possibility of outliers and confirm with the boxplot





Outliers:

Descriptive Statistics:

Also, for numeric features, the outliers could be too distinctive. We can look at their descriptive statistics.

For example, for the feature Ozone, we can see that the maximum value is 168, while the 75% quartile is only 68. The 168 value could be an outlier.

```
#Descriptive stat
data_cleaned3['Ozone'].describe()
```

```
count    116.000000
mean      42.129310
std       32.987885
min        1.000000
25%       18.000000
50%       31.500000
75%       63.250000
max      168.000000
Name: Ozone, dtype: float64
```

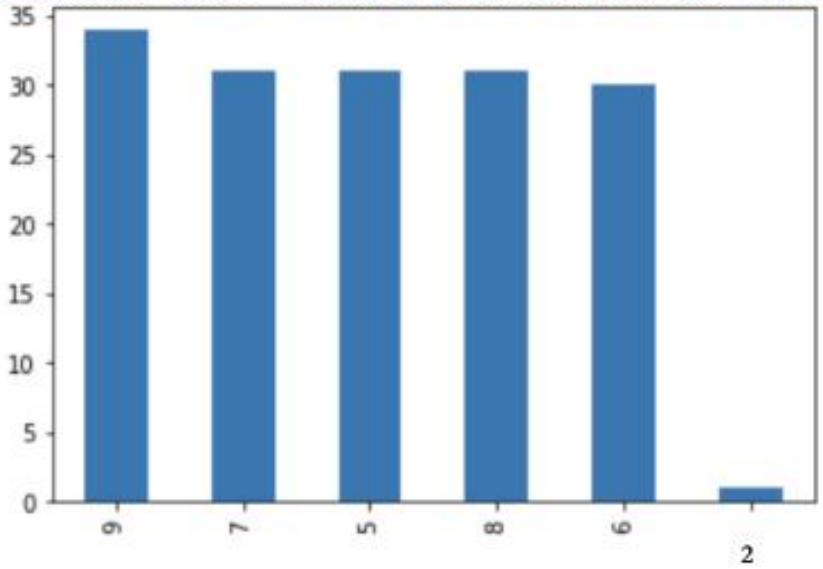
Outliers:

Bar Chart:

When the feature is categorical. We can use a bar chart to learn about its categories and distribution.

For example, the feature Month has a reasonable distribution except for category 2 .

2nd Month column has only one value , this can be an outlier.



Outliers:

What to do?
While outliers are not hard to detect, we have to determine the right solutions to handle them. It highly depends on the dataset and the goal of the project.

The methods of handling outliers are somewhat similar to missing data. We either drop or adjust or keep them. We can refer to the missing data section for possible solutions.

Data Cleaning Steps:

Duplicate rows

Rename the columns

Drop unnecessary columns

Convert data types to other types

Remove strings in columns

Change the data types

Outliers

Refer ipython notebook for data cleaning steps

Missing Values:

- Detection
- Treatment

What is Missing Values:

Some of the values will be missed in the data set because of various reasons such as human error, machine failures etc

```
data_cleaned3[data_cleaned3.isnull().any(axis=1)]
```

	Ozone	Solar	Wind	Month	Day	Year	Temp	Date
4	NaN	NaN	14.3	5.0	5	2010	56	2010-05-05
5	28.0	NaN	14.9	5.0	6	2010	66	2010-05-06
9	NaN	194.0	8.6	5.0	10	2010	69	2010-05-10
10	7.0	NaN	6.9	5.0	11	2010	74	2010-05-11
23	32.0	92.0	12.0	NaN	24	2010	61	NaT
24	NaN	66.0	16.6	5.0	25	2010	57	2010-05-25
25	NaN	266.0	14.9	5.0	26	2010	58	2010-05-26
26	NaN	NaN	8.0	5.0	27	2010	57	2010-05-27
31	NaN	286.0	8.6	6.0	1	2010	78	2010-06-01

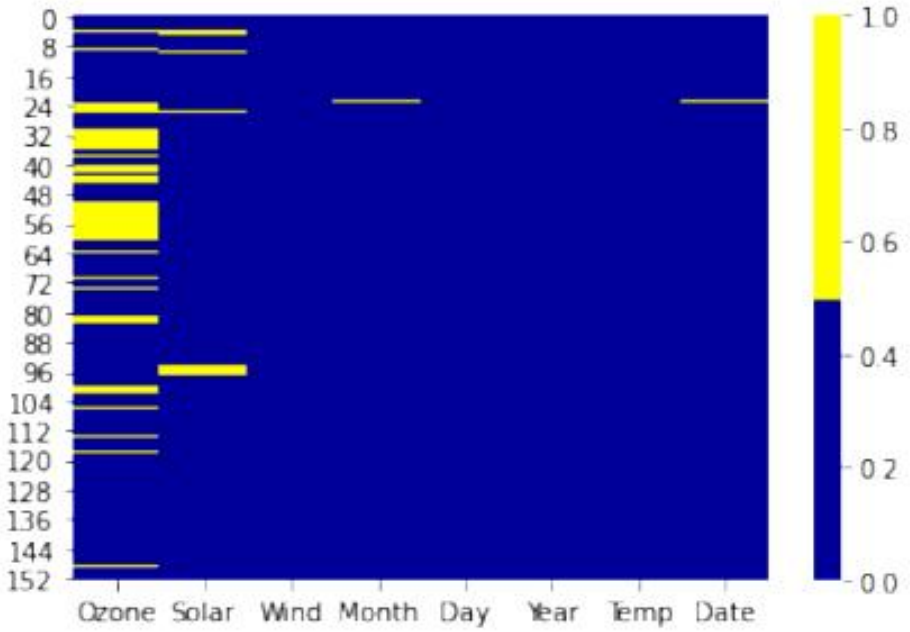
Missing Values: How to find ?

Missing Data Heatmap:
When there is a smaller number of features, we can visualize the missing data via heatmap.

```
data_cleaned3[data_cleaned3.isnull().any(axis=1)]
```

	Ozone	Solar	Wind	Month	Day	Year	Temp	Date
4	NaN	NaN	14.3	5.0	5	2010	56	2010-05-05
5	28.0	NaN	14.9	5.0	6	2010	66	2010-05-06
9	NaN	194.0	8.6	5.0	10	2010	69	2010-05-10
10	7.0	NaN	6.9	5.0	11	2010	74	2010-05-11
23	32.0	92.0	12.0	NaN	24	2010	61	NaT
24	NaN	66.0	16.6	5.0	25	2010	57	2010-05-25
25	NaN	266.0	14.9	5.0	26	2010	58	2010-05-26
26	NaN	NaN	8.0	5.0	27	2010	57	2010-05-27
31	NaN	286.0	8.6	6.0	1	2010	78	2010-06-01

The horizontal axis shows the feature name; the vertical axis shows the number of observations/rows; the yellow colour represents the missing data while the blue colour otherwise.



Treat Missing Values:

Drop the Observation

In statistics, this method is called the listwise deletion technique. In this solution, we drop the entire observation if it contains a missing value.
Only if we are sure that the missing data is not informative, we perform this. Otherwise, we should consider other solutions

Treat Missing Values:

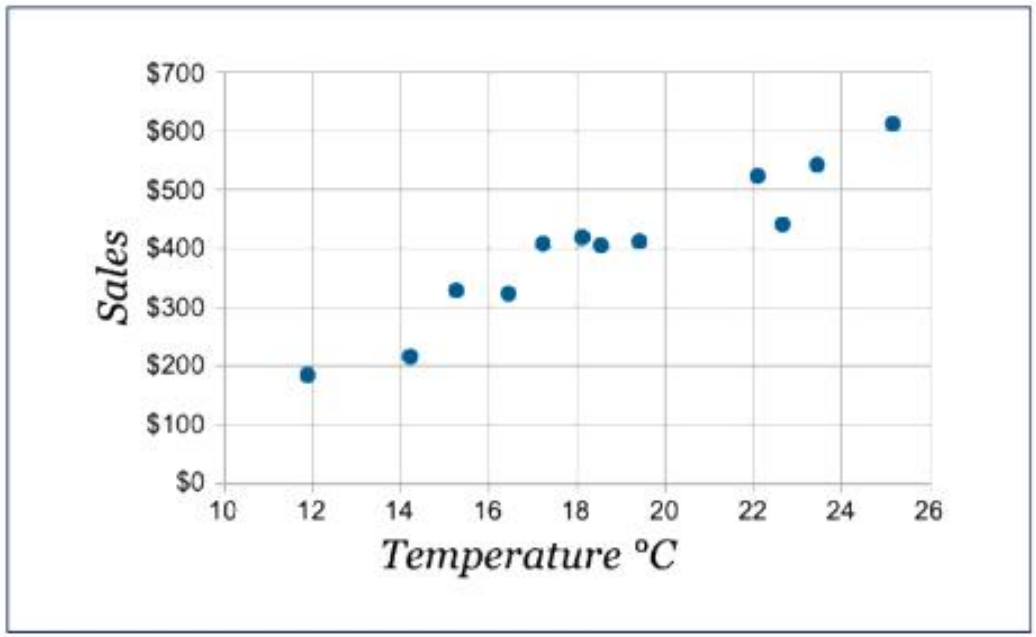
df.method()	description
dropna()	Drop missing observations
dropna(how='all')	Drop observations where all cells is NA
dropna(axis=1, how='all')	Drop column if all the values are missing
dropna(thresh = 5)	Drop rows that contain less than 5 non-missing values
fillna(0)	Replace missing values with zeros
isnull()	returns True if the value is missing
notnull()	Returns True for non-missing values

Scatter Plot

Scatter Plot:

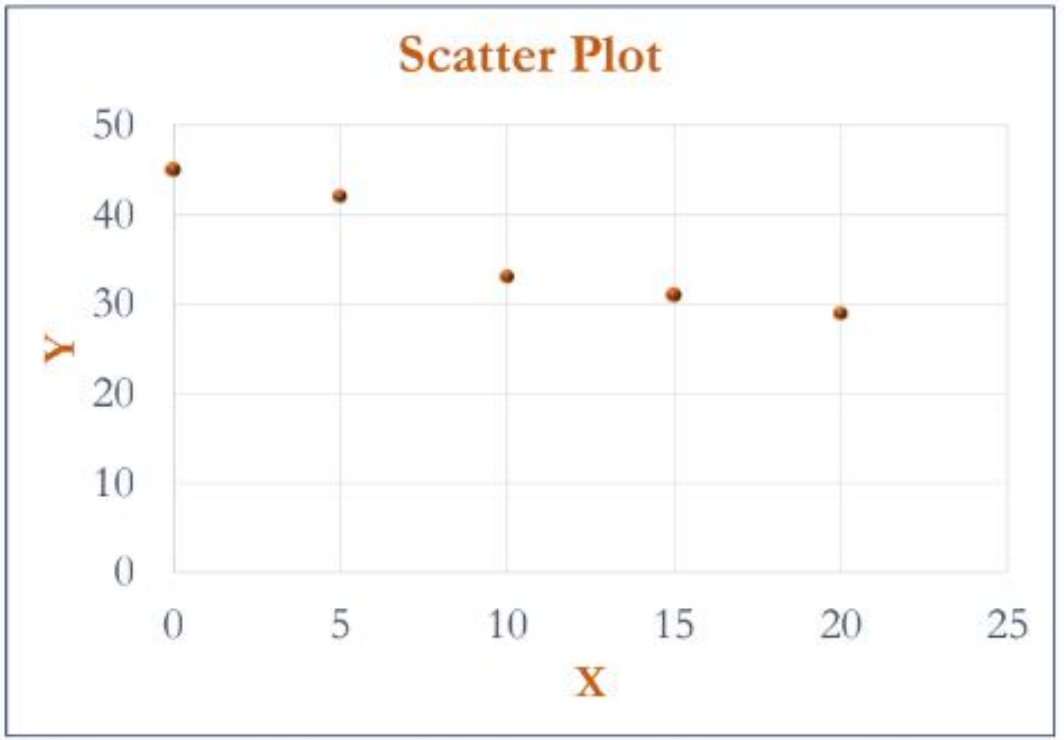
A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. Scatter plots are used to observe relationships between variables.

Temperature °C	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408



Which variable affects which one?

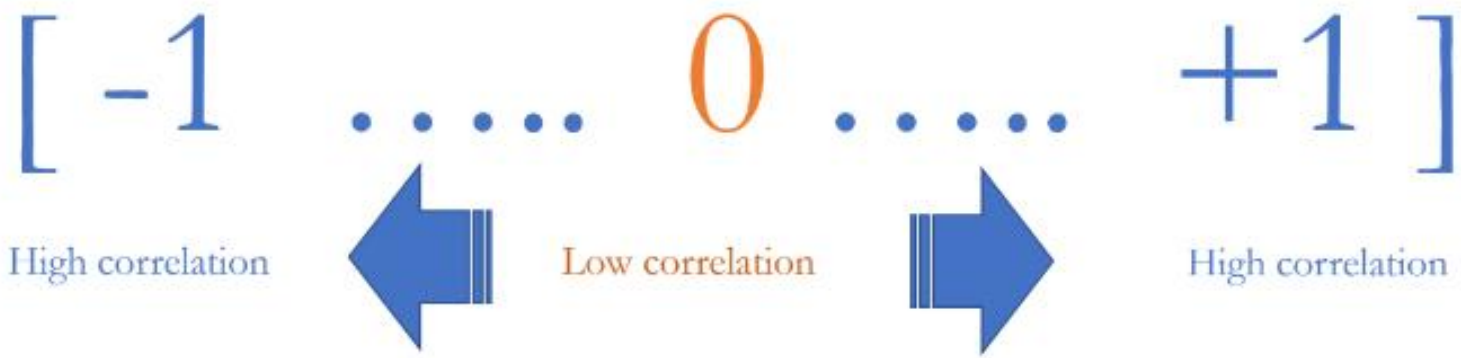
Cigarettes (X) in Years	Lung Capacity (Y)
0	45
5	42
10	33
15	31
20	29



Correlation

Pearson Correlation:

Correlation is a bi-variate analysis that measures the strength of linear association between two variables and the direction of the relationship. Correlation is a statistical technique used to determine the degree to which two variables are linearly related.



$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2}\sqrt{(Y-\bar{Y})^2}}$$

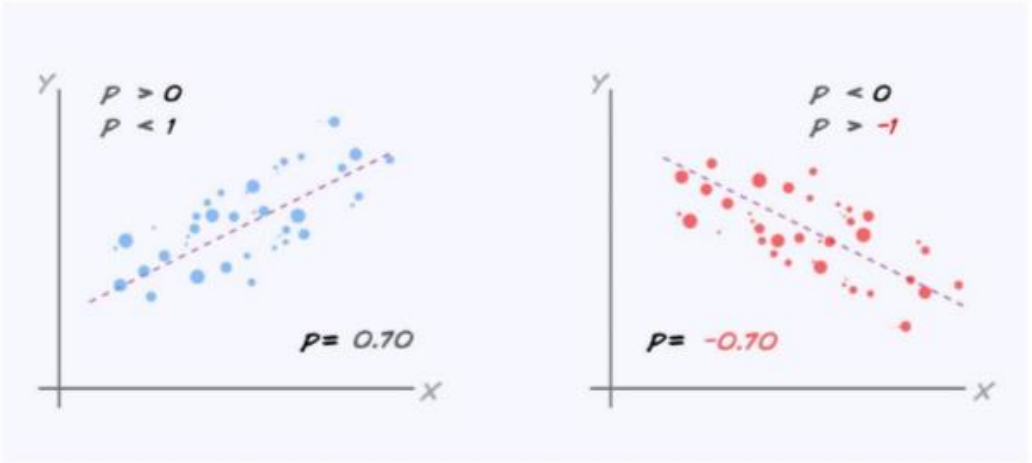
Where, $\bar{X}$ - mean of X variable

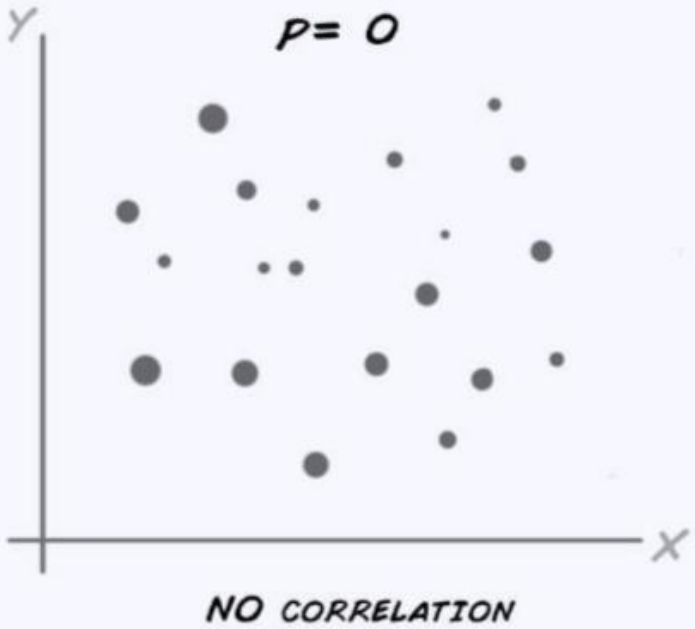
$\bar{Y}$ - mean of Y variable

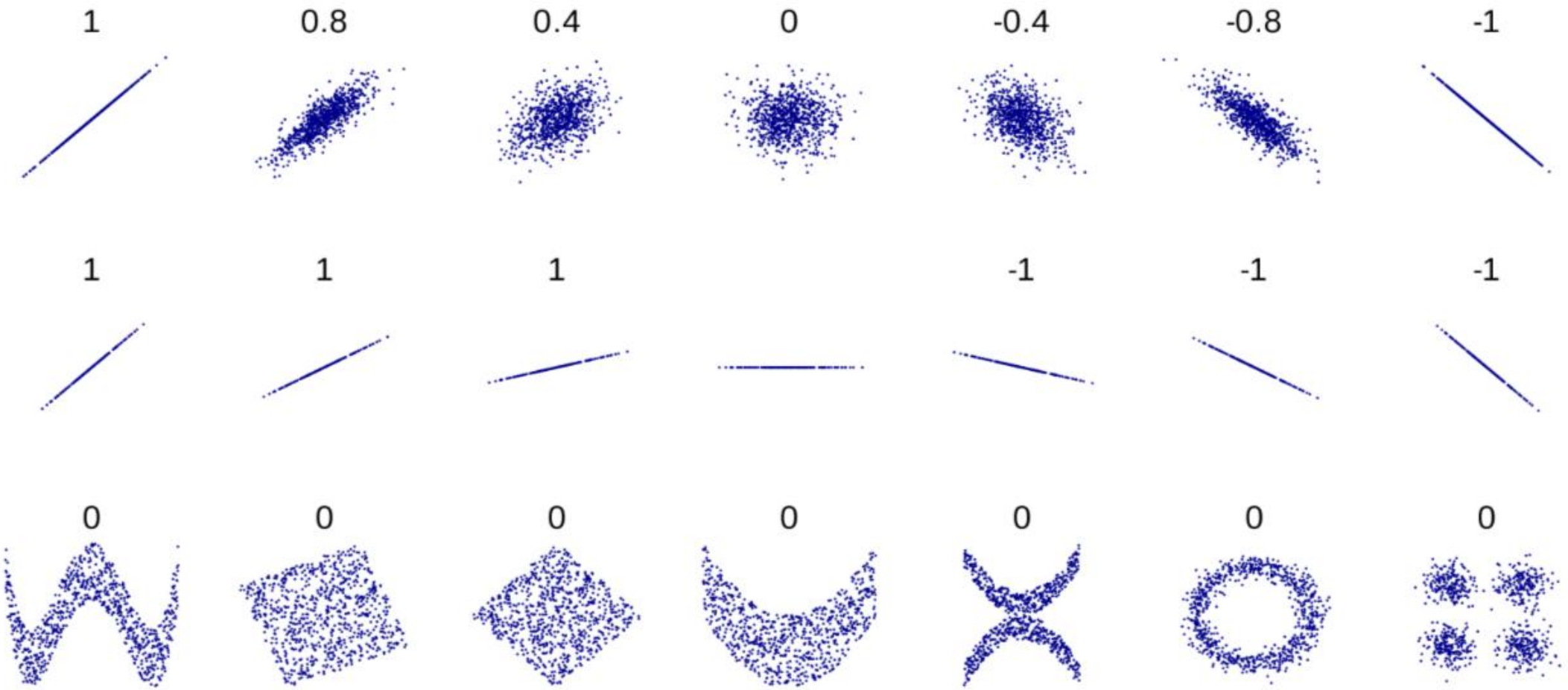
Correlation r - Interpretation

- Positive r indicates positive linear association between x and y or variables, and negative r indicates negative linear relationship
- r – always between -1 and +1
- The strength increases as r moves away from zero toward wither -1 or +1
- The extreme values +1 and -1 indicate perfect linear relationship (points lie exactly along a straight line)
- Graded interpretation : r 0.1-0.3 = weak; 0.4-0.7 = moderate and 0.8-1.0=strong correlation


Y
p= 1
POSITIVE CORRELATION
X
MORE X MORE Y
LESS X LESS Y
Y
p= -1
NEGATIVE CORRELATION
X
MORE X LESS Y
LESS X MORE Y


Y
p > 0
p < 1
p= 0.70
X
Y
p < 0
p > -1
p= -0.70
X


Y
P= 0
X
NO CORRELATION


1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

Transformations

Dummy Variable

Feature Scaling

Dummy variables

Categorical variables have to be converted to numerical using a method called One-hot encoding

Pd.get_dummies(df)

Ozone	Solar.R	Wind	Temp C	Month	Day	Year	Temp	Weather
41	190	7.4	67	5	1	2010	67	S
36	118	8	72	5	2	2010	72	C
12	149	12.6	74	5	3	2010	74	PS
18	313	11.5	62	5	4	2010	62	S
NA	NA	14.3	56	5	5	2010	56	S

	Ozone	Solar	Wind	Month	Day	Year	Temp	Weather_C	Weather_PS	Weather_S
0	41.0	190.0	7.4	5.0	1	2010	67	0	0	1
1	36.0	118.0	8.0	5.0	2	2010	72	1	0	0
2	12.0	149.0	12.6	5.0	3	2010	74	0	1	0
3	18.0	313.0	11.5	5.0	4	2010	62	0	0	1
4	31.0	238.0	14.3	5.0	5	2010	56	0	0	1

Feature Scaling:

Some machine learning algorithms are sensitive to feature scaling means results will vary based on the units of the features so remove of the effect of scaling, it is required to go for feature scaling

Standardization

Standardization is scaling technique where the values are cantered around the mean with a unit standard deviation.

$$Z = \frac{x - \mu}{\sigma}$$

Normalization

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}.$$

Automatic EDA methods

Exploratory data analysis (EDA) is an essential early step in most data science projects and it often consists of taking the same steps to characterize a dataset (e.g. find out data types, missing information, distribution of values, correlations, etc.).

Given the repetitiveness and similarity of such tasks, there are a few libraries that automate and help speed up the process

Libraries:
pandas_profiling
sweetviz

Thank you