510 DATA SCIENCE

Lecture 04

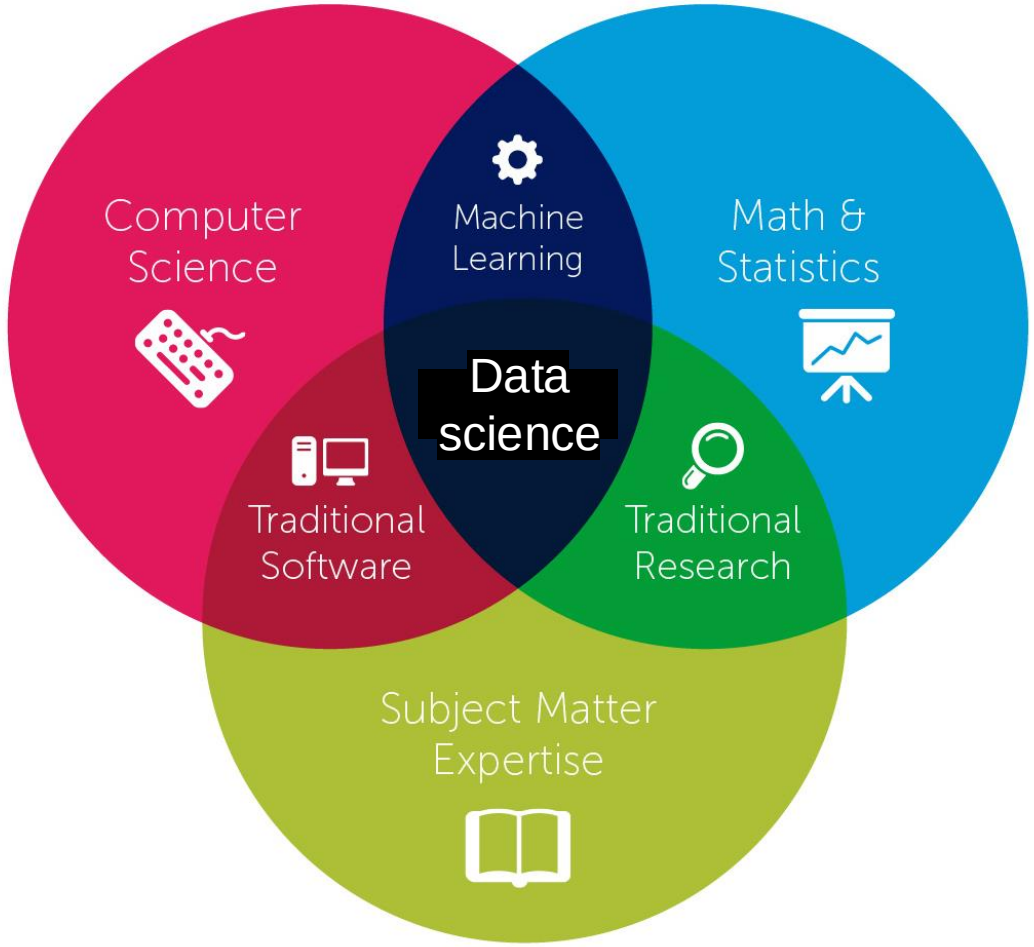
Fall 2024

Instructor: Assoc. Prof. Şener Özönder

Email: sener.ozonder@bogazici.edu.tr

Institute for Data Science & Artificial Intelligence

Boğaziçi University



Exploratory Data Analysis (EDA)

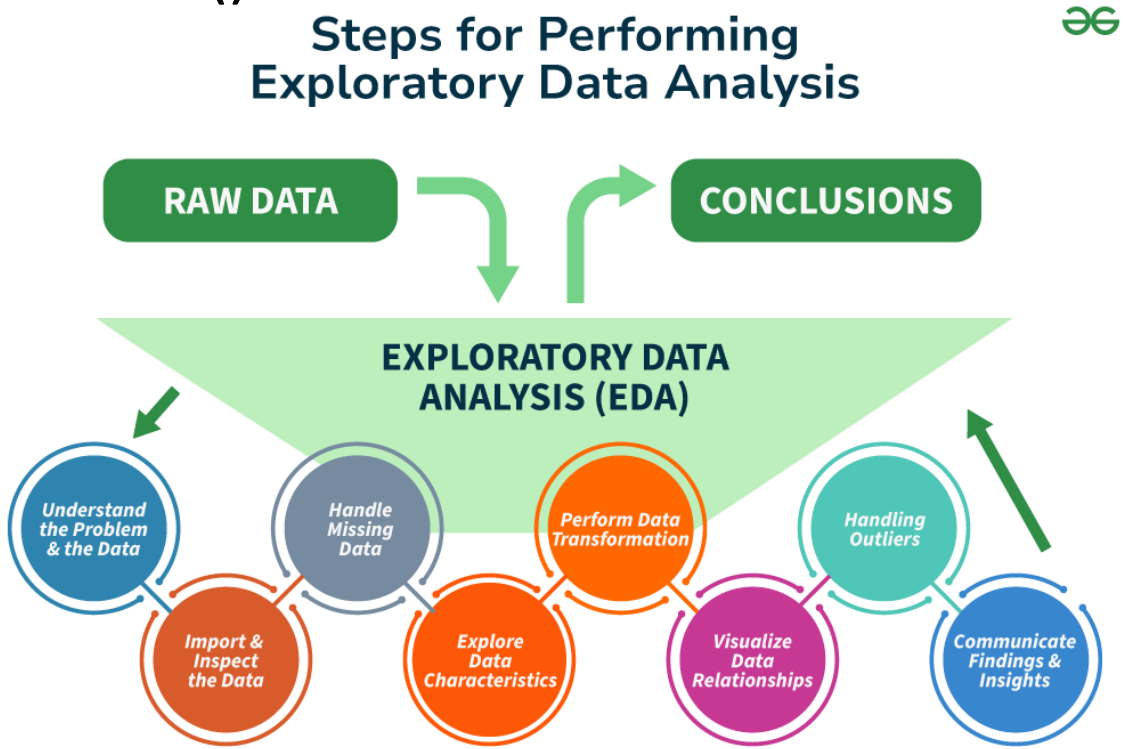
- EDA is understanding the main characteristics of a dataset, often with statistics and visual methods. The purpose is to discover patterns, spot anomalies and outliers, test hypotheses, and check assumptions.
- **Data types and structures:** df.dtypes, df.info().
- **Desciptive statistics:**

df.describe(), df.mean(), df.median(),

df.mode(), df.std().

- **Checking for Missing Values**:

df.isnull(), df.isna().sum()



Exploratory Data Analysis (EDA)

- **Univariate analysis (one variable/column at a time)**

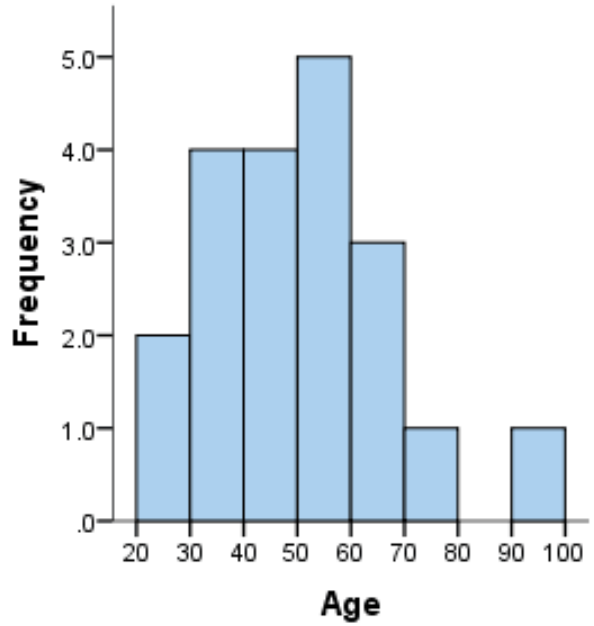
Data features (columns) are usually not independent. For example, salary is correlated with age of the employee and position in the company. Age and position are also related. Univariate ("single variable") analysis disregards these correlations and tries to analyze each column independently. For example, plotting historgrams is part of the univariate analysis.

Continuous Variables:

- Histograms: plt.hist(), df.plot(kind='hist')
- Boxplots: plt.boxplot(), df.plot(kind='box')

Categorical Variables:

- Bar plots: plt.bar(), df.value_counts().plot(kind='bar')
- Pie charts: plt.pie()



Exploratory Data Analysis (EDA)

- **Multivariate Analysis (examining relationships between columns)**

This is analysis where more than one column/variable/predictor/feature is used.

Correlation matrix: It shows correlation between predictors.

Scatter Plots: plt.scatter()

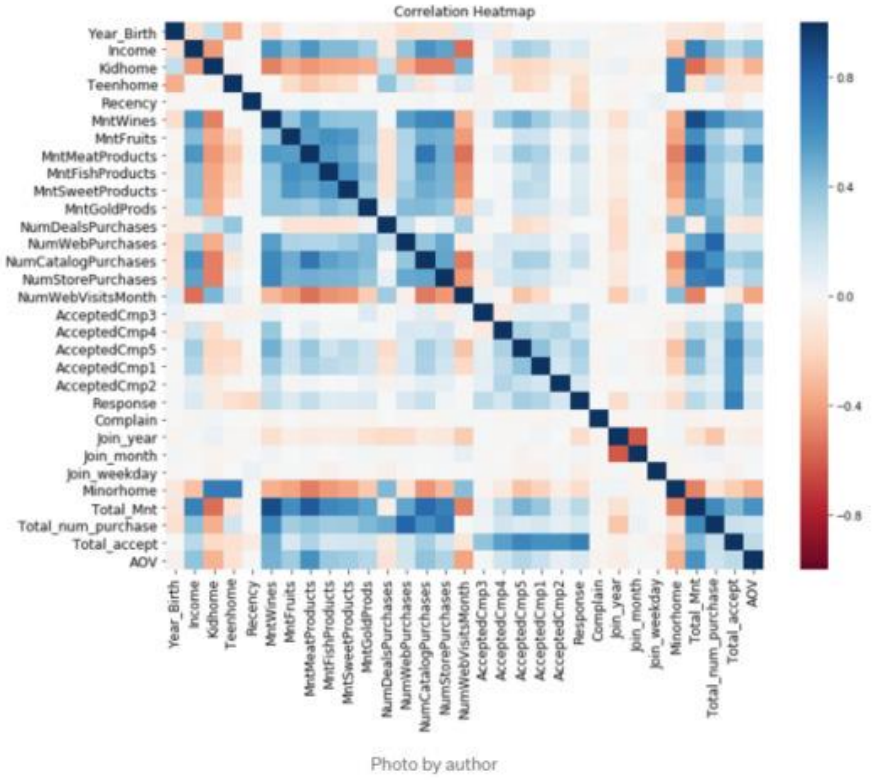


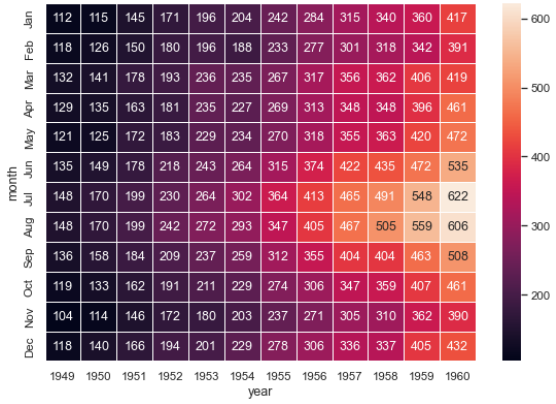
Photo by author

Exploratory Data Analysis (EDA)

- **Advanced Visualization:**

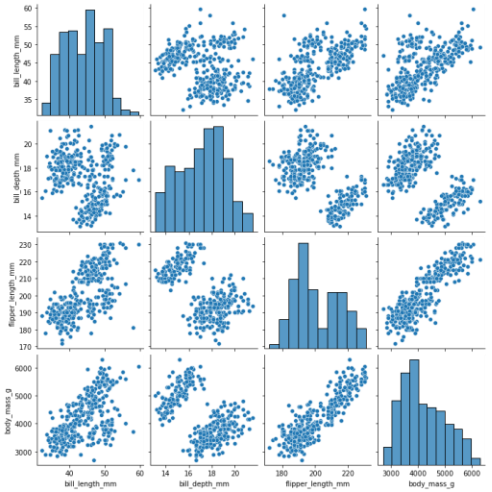
Heatmaps: Visual representation of data where values are depicted by color.

- sns.heatmap()



Faceting: Creating multiple plots which share the same variables.

- sns.FacetGrid()



Time Series Visualization:

- Line charts: plt.plot(), df.plot()

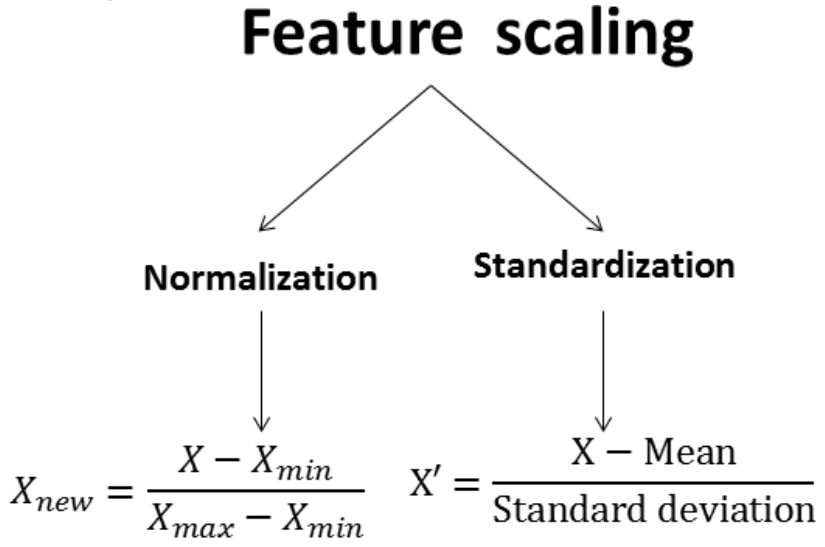
Interactive Visualization (Zoom, rotation etc):

- Libraries like Plotly, Bokeh, ...

Exploratory Data Analysis (EDA)

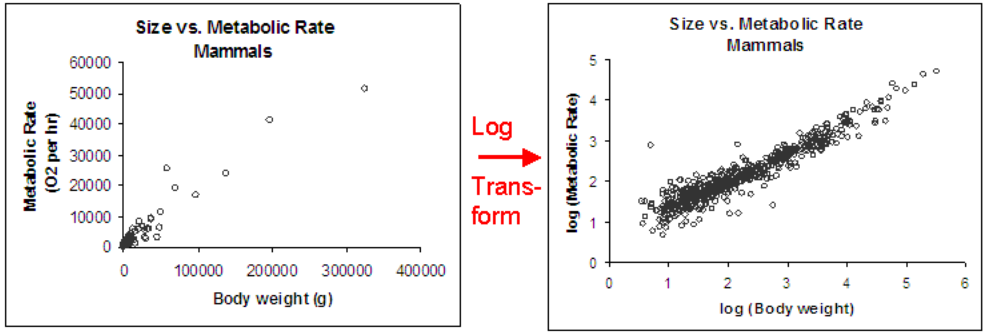
- **Data Transformation for Visualization:**

Normalization and Standardization: Most of times this helps increase model performance.
Normalization: Scale data to interval [0,1].
Standardization: Make data's μ=0 and σ=1.



Log Transformations: Useful for skewed data.

- np.log()



Binning Data: Converting continuous data into categorical.

- pd.cut(), pd.qcut()

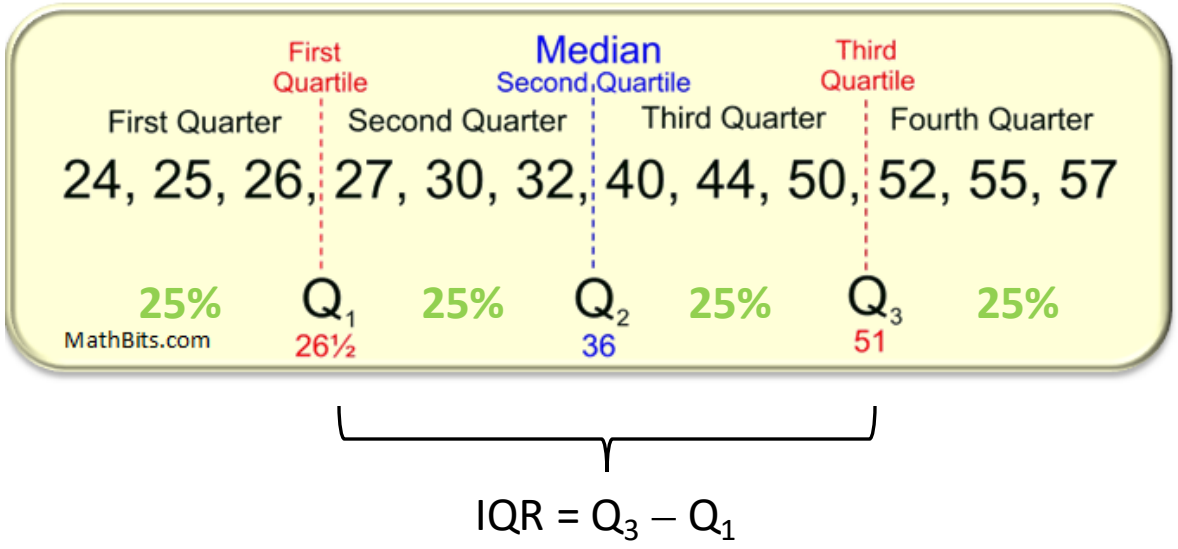
Exploratory Data Analysis (EDA)

- **Outlier Detection:**

Visual Methods: Using boxplots and scatter plots.

Statistical Methods: Z-score, IQR (interquartile range)

For the data=[24,25,26,27,30,32,40,44,50,52,55,57]



Data points below Q_1 - 1.5 IQR and above Q_3 + 1.5 IQR are considered outliers.

Exploratory Data Analysis (EDA)

Whiskers usually show
Q_1 - 1.5 IQR and Q_3 + 1.5 IQR.