UNIT 3:
Data Engineering

Prepared by Nima Dema

Data Wrangling

GYALPOZHING
COLLEGE OF INFORMATION TECHNOLOGY

Data Wrangling

- ✓ Data wrangling, also known as data munging, is an iterative process that involves data exploration, transformation, validation, and making it available for a credible and meaningful analysis.
- ✓ Process of converting and mapping data from one raw format into another.
- ✓ Includes range of tasks involved in preparing raw data for a clearly defined purpose.

Data Cleaning

- ✓ Poor quality data weakens an organization's competitive standing and undermines critical business objectives.
- ✓ Missing, inconsistent, incorrect data leads to false conclusion and ineffective decisions.
- ✓ Data cleaning is the subset of entire data wrangling process.

Data Cleaning Workflow includes:

Inspection

Cleaning

Verification

GYALPOZHING
COLLEGE OF INFORMATION TECHNOLOGY

Data Cleaning

- ✓ Techniques depends on use case and type of issues encountered.
- ✓ Missing Values can cause unexpected or biased results.
  - Filter out records with missing data
  - Source missing information
  - Imputation, calculate the missing value based on statistical values.
- ✓ Duplicate data are data points that are repeated in your datasets
  - Need to be removed
- ✓ Irrelevant data is data that is not contextual to use case
- ✓ Data type conversion is needed to ensure that values in field are stored as the data type of that field.
- ✓ Standardizing data is needed to ensure data-time formats and units of measurement as standard across the datasets.
- ✓ Syntax errors, such as white spaces, extra spaces, typos and formats need to be fixed
- ✓ Outliers need to be examined for accuracy and inclusion in the dataset.

Questions

1. What type of data is usually found in databases and spreadsheets?
    a. Semi-Structured data
    b. Structured data
    c. Unstructured data
    d. Social media content
2. Which of these data sources is an example of semi-structured data?
    a. Social media feeds
    b. Network and web logs
    c. Document
    d. Data from APIs and web services

GYALPOZHING
COLLEGE OF INFORMATION TECHNOLOGY

Data Visualization

Data Visualization

- **Data visualization** is the techniques to present data in pictorial or graphical formats
  - Explore new patterns and hidden patterns in the data.
  - Allows decision makers and stakeholders to analyse data visually.
  - Simplifies the complex quantitative information
  - Analyze and Explore data easily
  - Identifies the area that needs improvement
  - Identifies the relationships between data points and variables.

Why Data visualization?

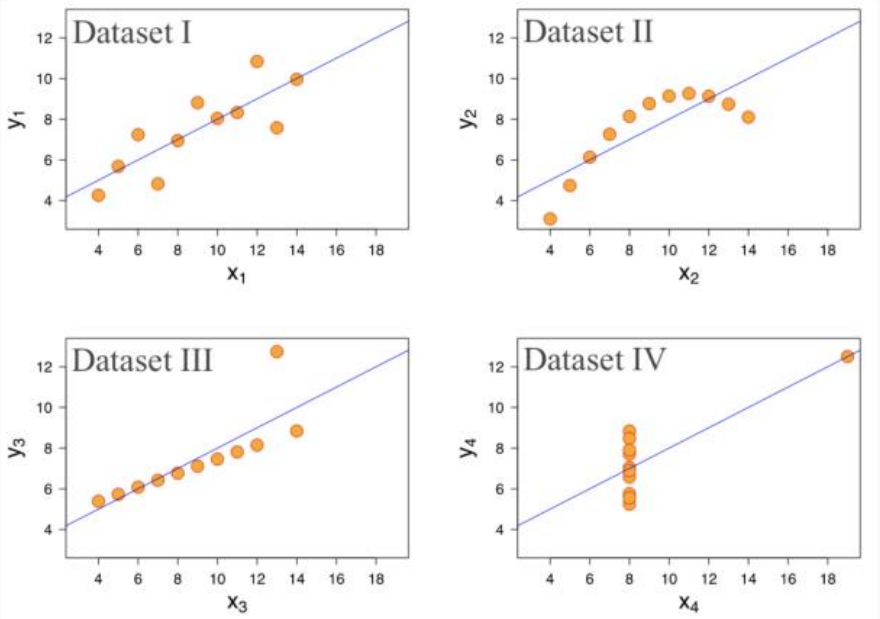
The following four data sets comprise the Anscombes Quartet; all four sets of data have identical simple summary statistics.

	Dataset I		Dataset II		Dataset III		Dataset IV	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Sum:	99.00	82.51	99.00	82.51	99.00	82.51	99.00	82.51
Avg:	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std:	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

GYALPOZHING
COLLEGE OF INFORMATION TECHNOLOGY

Why Data visualization?

Summary statistics clearly don't tell the story of how they differ.
But a picture can be worth a thousand words:



Why Data visualization?

- For your data visualization to be of value, choose the visualization that effectively delivers your findings to the audience.
  - What is the relationship that I am trying to establish?
  - Do I want to compare multiple values, such as the **`number of product sold`**, and **`revenues generated`** over last three years?
  - Is audience looking for correlation between two variables?
  - Do I want to detect anomalies in the data?

Types of visualization – Bar graph

- A bar chart is used when you want to show a distribution of data points or perform a comparison of metric values across different subgroups of your data.

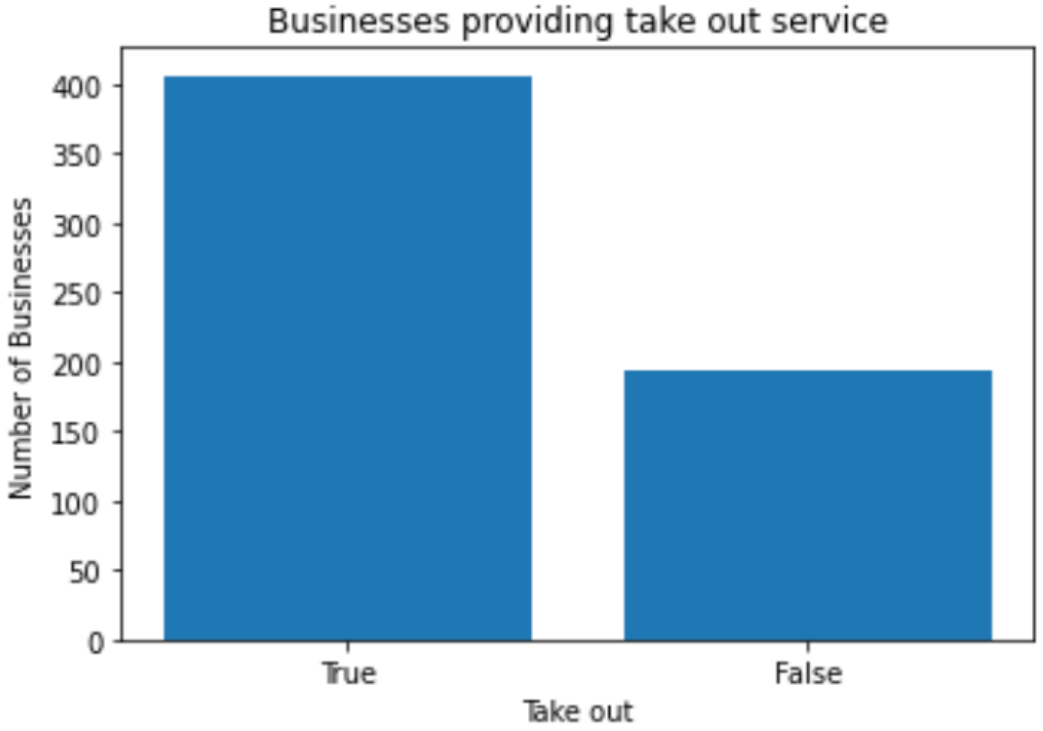
- The primary variable of a bar chart is its categorical variable. Example: take_out (True, Flase), Gender (Female, Male) etc. The important point for this primary variable is that the groups are distinct.

- In contrast, the secondary variable will be numeric in nature. The secondary variable's values determine the length of each bar. Example: can be frequency count, other measure summary measures computed for each groups.

Types of visualization – Bar graph

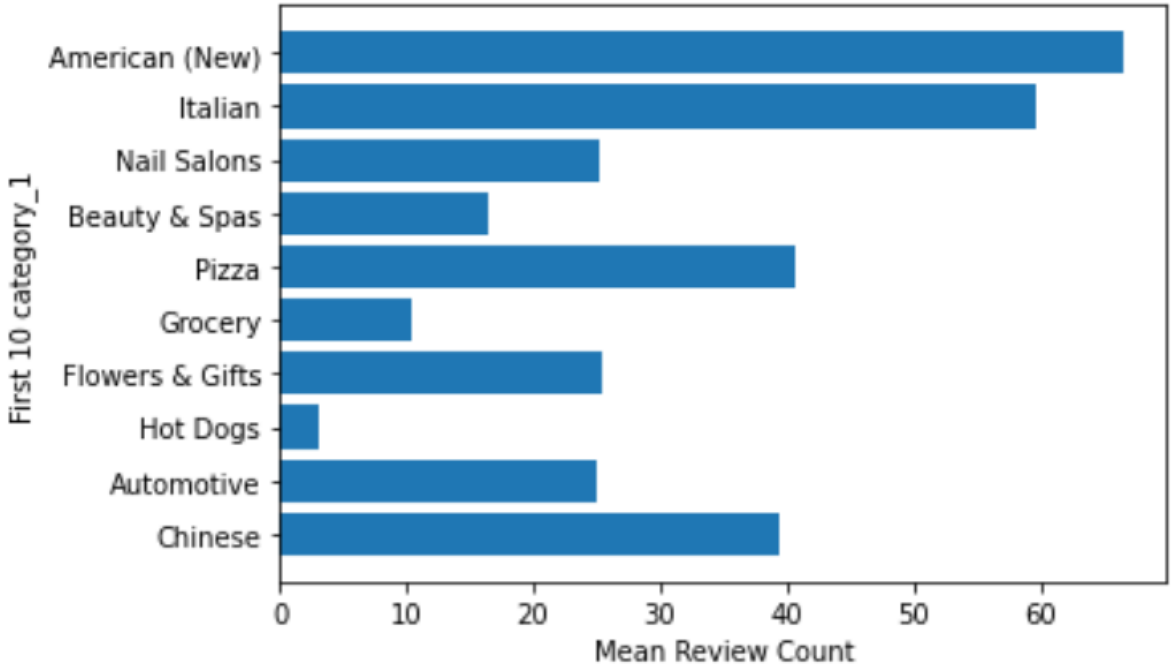
- `take_out` as Primary variable

- Frequency count in each category as secondary variable.



Types of visualization – Bar graph

- Primary variable: first ten **`category_1`** feature
- secondary variable: average **`review count`** in each category



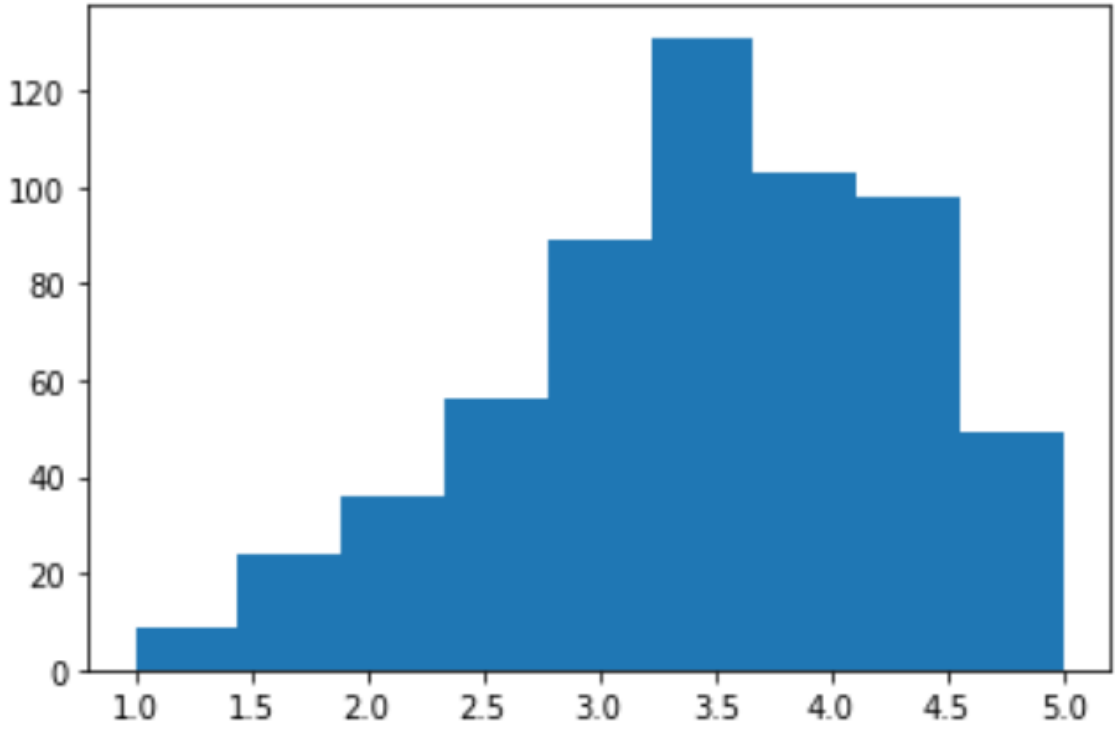
GYALPOZHING
COLLEGE OF INFORMATION TECHNOLOGY

Types of visualization – Histogram graph

- A histogram functions more or less like a bar chart in that it represents counts (or relative frequencies).
- However, histograms are used to reveal the counts of bins of continuous data. Binning is a way of discretizing continuous data so that we can better understand it.
- The bars in a histogram are typically placed right next to each other to emphasize this continuous nature:
- bar charts usually have some space between bars to emphasize the categorical nature of the primary variable.

Types of visualization – Histogram graph

Histogram of stars column of yelp datasets



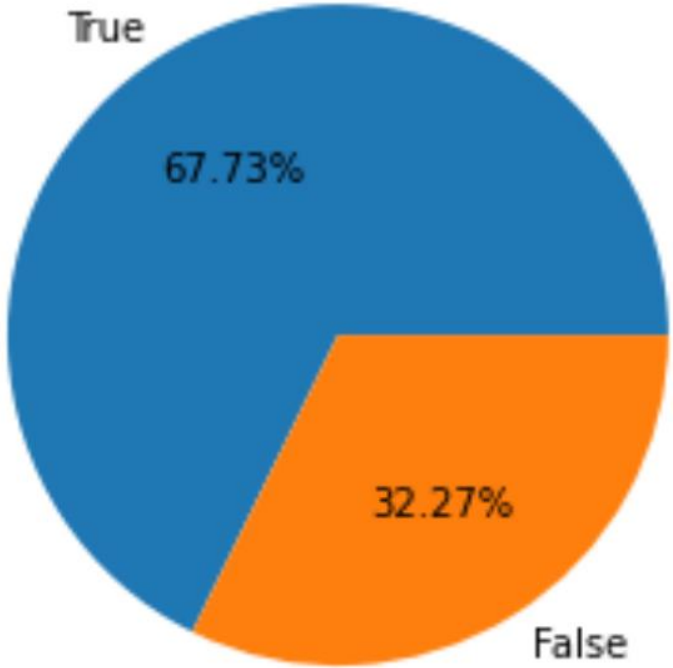
GYALPOZHING
COLLEGE OF INFORMATION TECHNOLOGY

Types of visualization – Pie Chart

- A pie chart shows how a total amount is divided between levels of a categorical variable as a circle divided into radial slices.
- Each categorical value corresponds with a single slice of the circle, and the size of each slice (both in area and arc length) indicates what proportion of the whole each category level takes.
- In order to use a pie chart, you must have some kind of whole amount that is divided into a number of distinct parts.
- Your primary objective in a pie chart should be to compare each group's contribution to the whole, as opposed to comparing groups to each other.

Types of visualization – Pie Chart

Pie chart to depicts the distribution of businesses which offers take out service and which doesn't

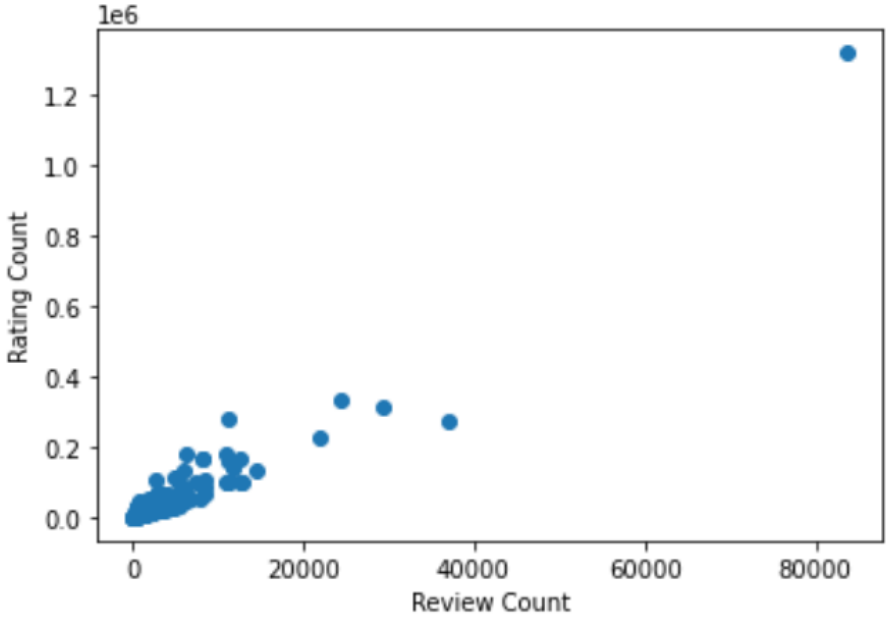


Types of visualization – Scatter Plot

- A scatter uses dots to represent values for two different numeric variables.
- Scatter plots' primary uses are to observe and show relationships between two numeric variables
- The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.
- Identification of correlational relationships are common with scatter plots

Types of visualization – Scatter Plot

- Relationships between variables can be described in many ways: positive or negative, strong or weak, linear or nonlinear.
- can divide data points into groups based on how closely sets of points cluster together
- Check if there are any outlier points

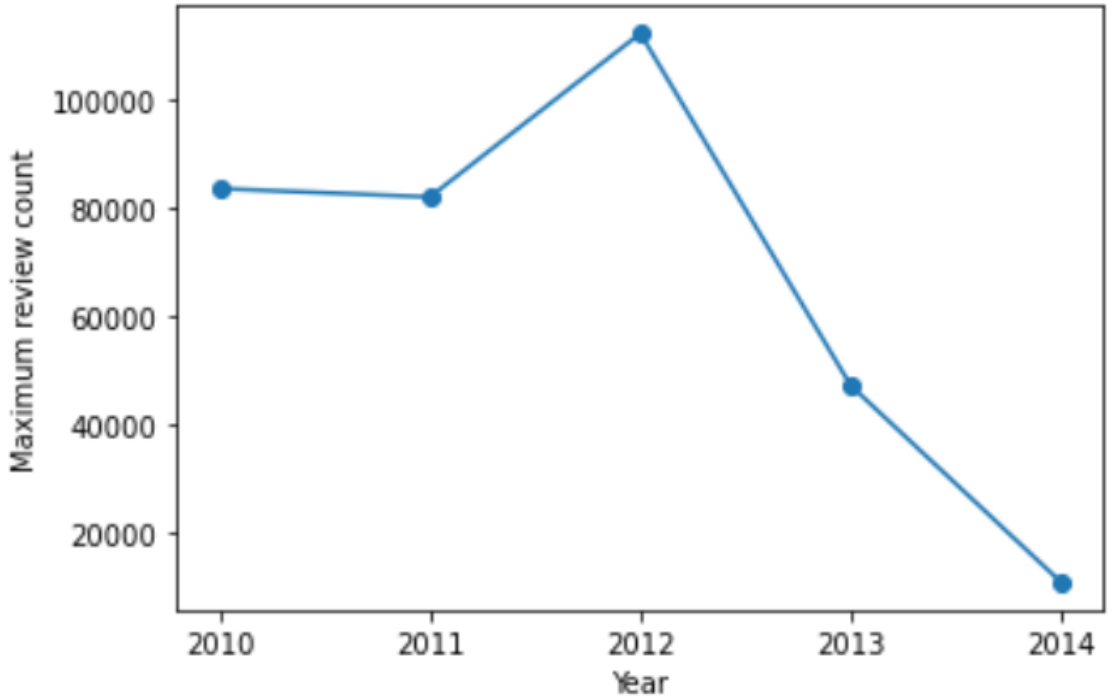


Types of visualization – Line Chart

- A line chart uses points connected by line segments from left to right to demonstrate changes in value.
- The horizontal axis depicts a continuous progression, often that of time, while the vertical axis reports values for a metric of interest across that progression.
- Use a line chart to emphasize changes in values for one variable (`plotted on the vertical axis`) for continuous values of a second variable (`plotted on the horizontal`)
- Multiple lines can also be plotted in a single line chart to compare the trend between series.

Types of visualization – Line Chart

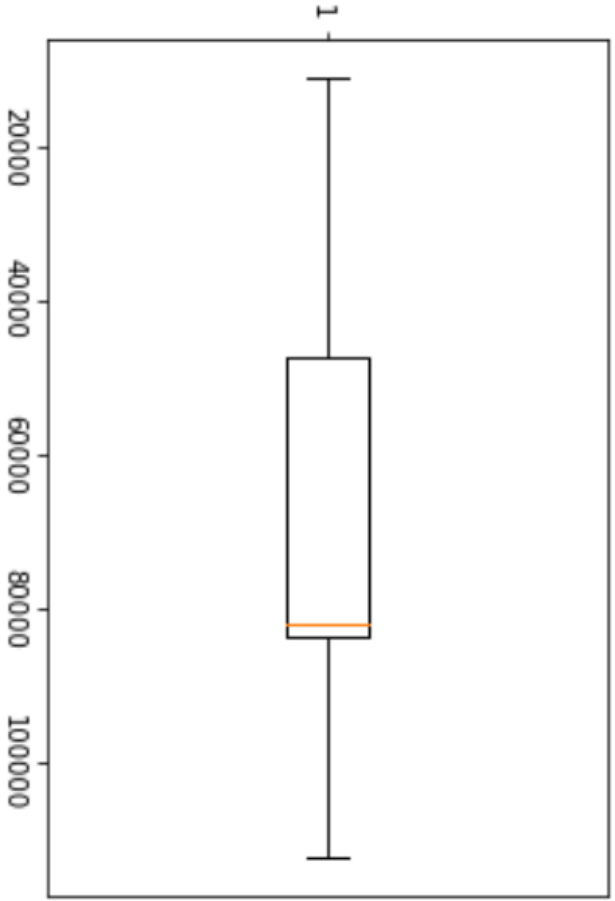
Line chart depicting maximum book reviewers from year 2010 to 2014



GYALPOZHING
COLLEGE OF INFORMATION TECHNOLOGY

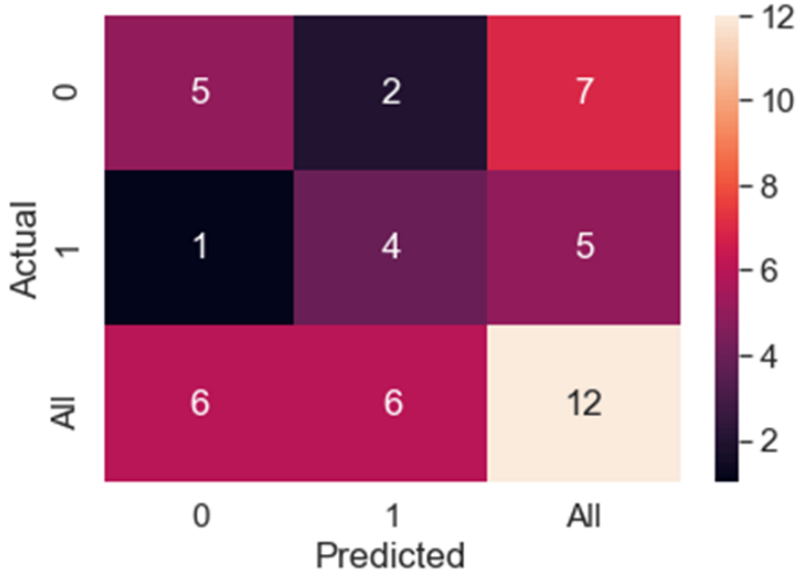
Types of visualization - boxplots

- A box plot uses boxes and lines to depict the distributions of one or more groups of numeric data
- It represent first, second and third quartile in a box and the minimum and maximum as whiskers of the data apart from outliers.
- Outlier points are defined to be any data points which are not in the interval.



Types of visualization - heatmap

- A heatmap is a graphical representation of data that uses a system of color-coding to represent different values.
  - Plot rectangular data as a color-encoded matrix.
  - Used to represent correlations between different features.
  - In machine learning heat map can be used to represent confusion matrix graphically.



THANK YOU ☺