Stony Brook University

# Data Exploration

Vibha Mane
The COSINE Lab
Department of Electrical and Computer Engineering

# What is Machine Learning?

- Machine Learning is the task of **learning from data**, that is, generating models from data, and then **making predictions**.
- In particular, we want to **unveil possible hidden patterns** and structures from data, and use this information for analysis and understanding of the nature of data.
- We want to develop **mathematical algorithms** to train a model from data, and utilize it to make predictions or decisions, in an automated manner.

# Machine Learning Example - Classification

- As an example, we have the **Iris Flower Data Set** or Fisher's Iris Data Set (Fisher, 1936) which consists of **50 samples or observations** from each of the three species of Iris: Iris Setosa, Iris Virginica and Iris Versicolor. The **three species** are called **class labels** or **target features**.
- For each of the 150 samples, **four features** were measured: sepal length, sepal width, petal length and petal width. These four features are called **descriptive features or** just **features**
- The objective here is to **learn the model** from the given data set, so that we can identify which species a new observation belongs to. This task is known as **classification**.

# Iris Flower Data

| sepal_length | sepal_width | petal_length | petal_width | species |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| 4.9 | 3 | 1.4 | 0.2 | Iris-setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| 5 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa |
| 4.6 | 3.4 | 1.4 | 0.3 | Iris-setosa |

A subset of the Iris Data Set:

- Each of the first four columns contains a **feature**.
- The last column is the **class label**, for the three different species.
- Each row is an observation of a different iris flower; these observations are known as **samples**.

# Types of Data in Machine Learning

- **Numerical Data:** Also known as quantitative data can be continuous or discrete.
  - **Continuous Data** can take any numerical value, such as 3.1, 1000.4, 22.5. Some examples are height of a person or salary of a person.
  - **Discrete Data** are whole numbers such as 22, 31, 110. Some examples are age of a person or the number of people taking a Python course.
- **Categorical Data:** This type of data is assigned a label based on a **qualitative property**. There are two types of categorical data:
  - **Nominal Data** is named data which does not has any numerical value. An example is the state a person resides in - New York, Texas, California. Another example is the species of Iris flower in the above data set.
  - **Ordinal Data** has some sense of order in its values. An example is the financial status of a person with three categories: low, medium, high.

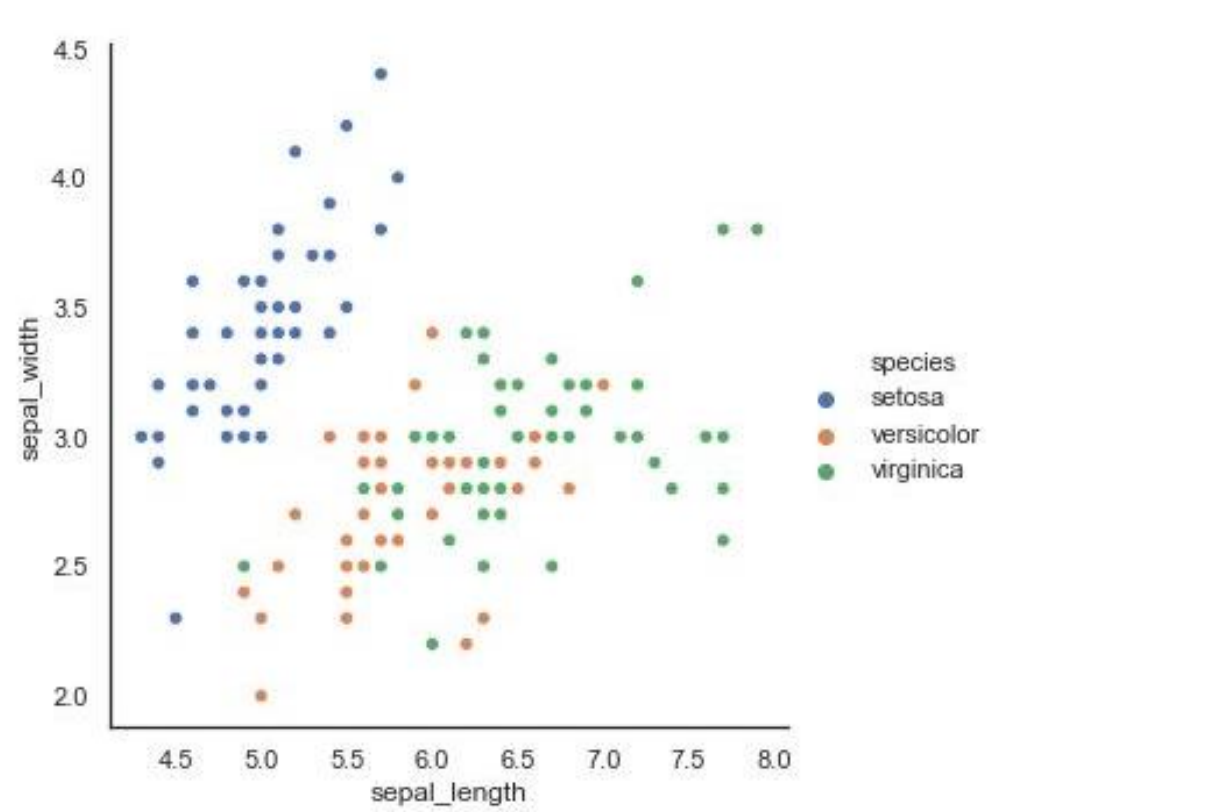
# Exploratory Data Analysis

- **Exploratory Data Analysis** is a data analysis approach to reveal important characteristics of a data set through visualization.
- Some methods of representing data visually are:
  - Scatter Plots
  - Histogram Plots
  - Density Plots
  - Bar Plots
  - Box-and-whisker plot
  - Hexbinplot

# Data Visualization: Feature Space

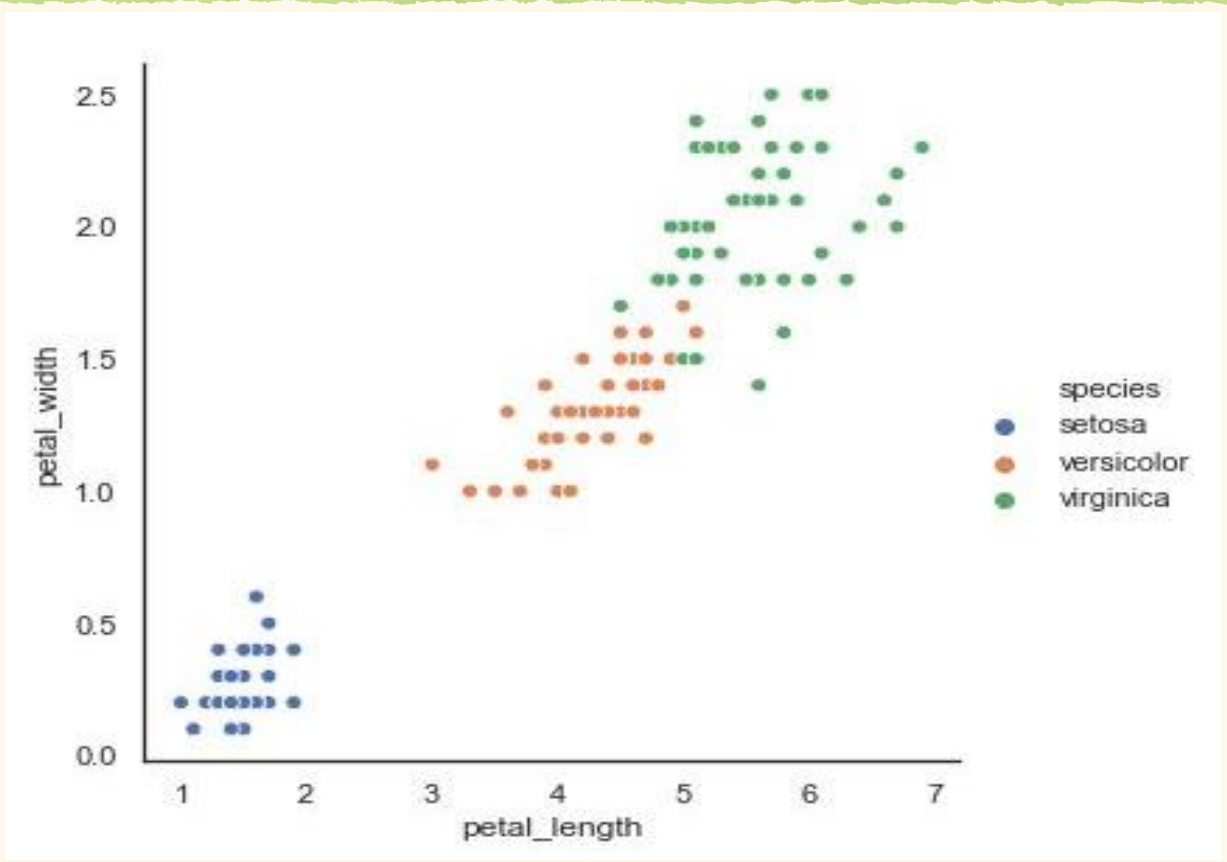
- A useful way to detect patterns in data and relationship between features is through exploratory data analysis with visualization.
- In this example, the **feature space** refers to the four features of the Iris data sets.
- A **scatter plot** displays the data points with two features chosen for the horizontal and vertical axis.
- In the following figures, we plot the Iris data set in feature space.

# Data Visualization: Scatter Plot I




- This is a scatter plot of the Iris Data Set in the sepal_length vs. sepal_width **feature space**.
- The three species are shown in three different colors.
- The species setosa is well separated, however the species versicolor and virginica can't be discerned based on these features.

# Data Visualization: Scatter Plot II

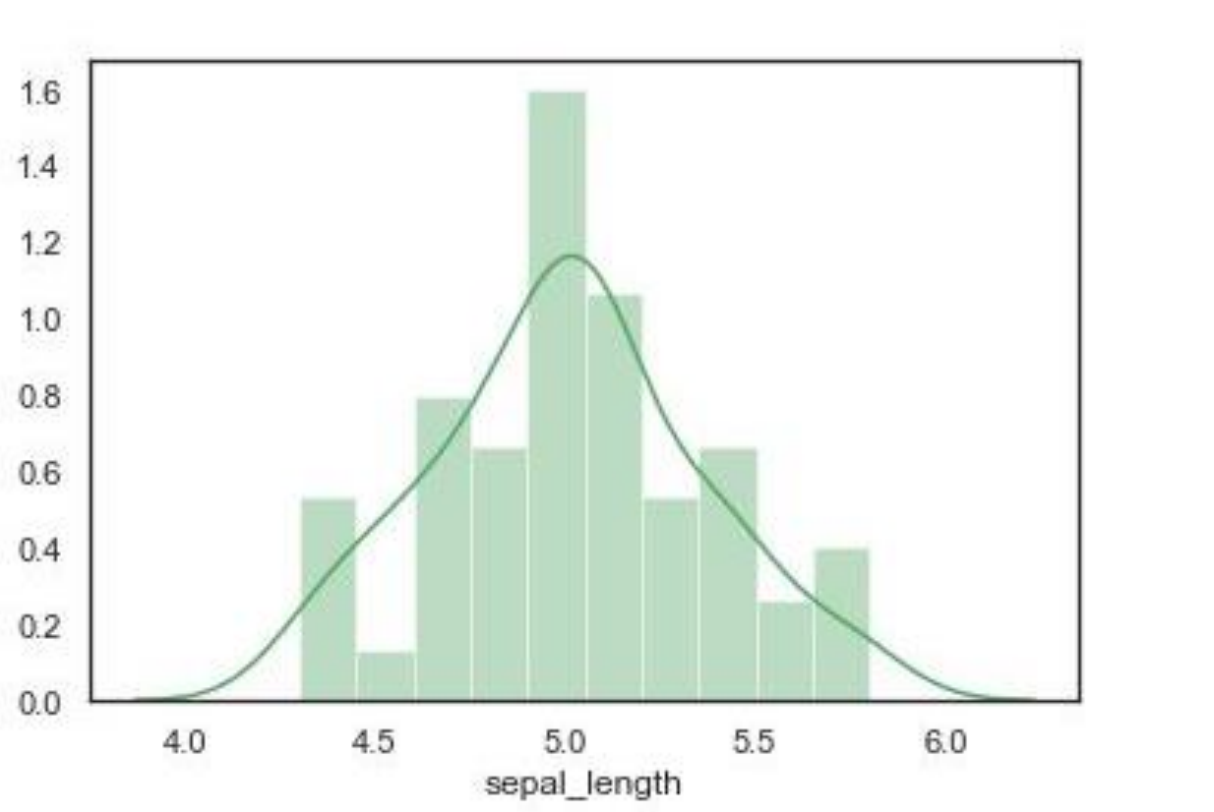



- This is another scatter plot of the Iris Data Set in the petal_length vs. petal_width **feature space**.
- In this feature space, all three species are well separated.

# Data Visualization: Histogram

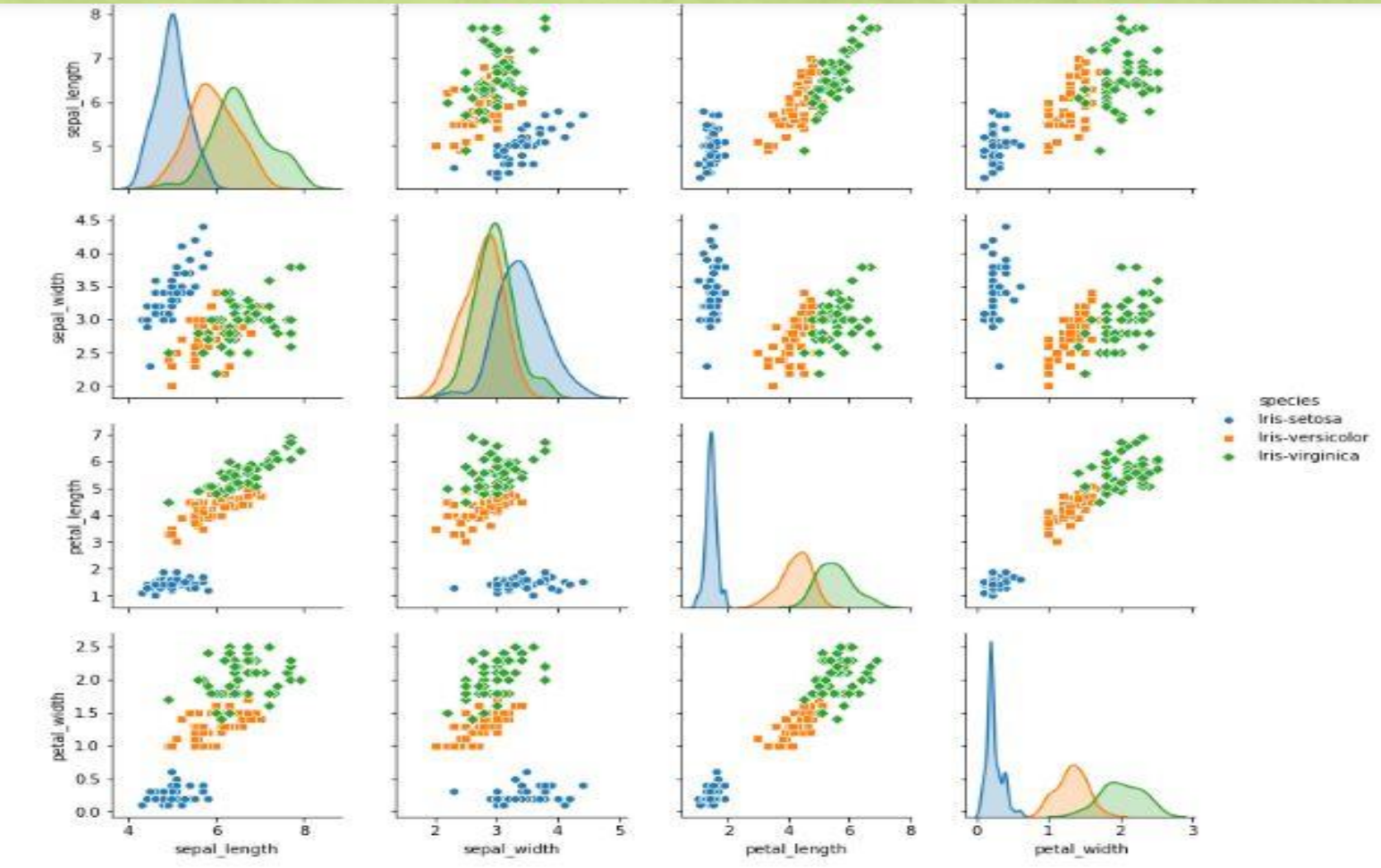
- A histogram is a plot that tells you the **frequency distribution** of a set of continuous data.
- The plot on next page shows the histogram of a single continuous variable from the Iris data set: sepal_length.
- To construct a histogram, the entire range of values is divided into intervals, also known as **bins**.
- For each bin, a vertical bar with height  proportional to the number of data points in that bin is plotted.

# Data Visualization: Histogram & Density Plot




- The vertical bars give a histogram plot of the sepal_length feature of the Iris data set.
- Also shown is a probability density plot fitted to the data (solid curve).
- The density plot will be studied in more details in a later module.

# Scatter Plot Matrix of the Iris Data Set

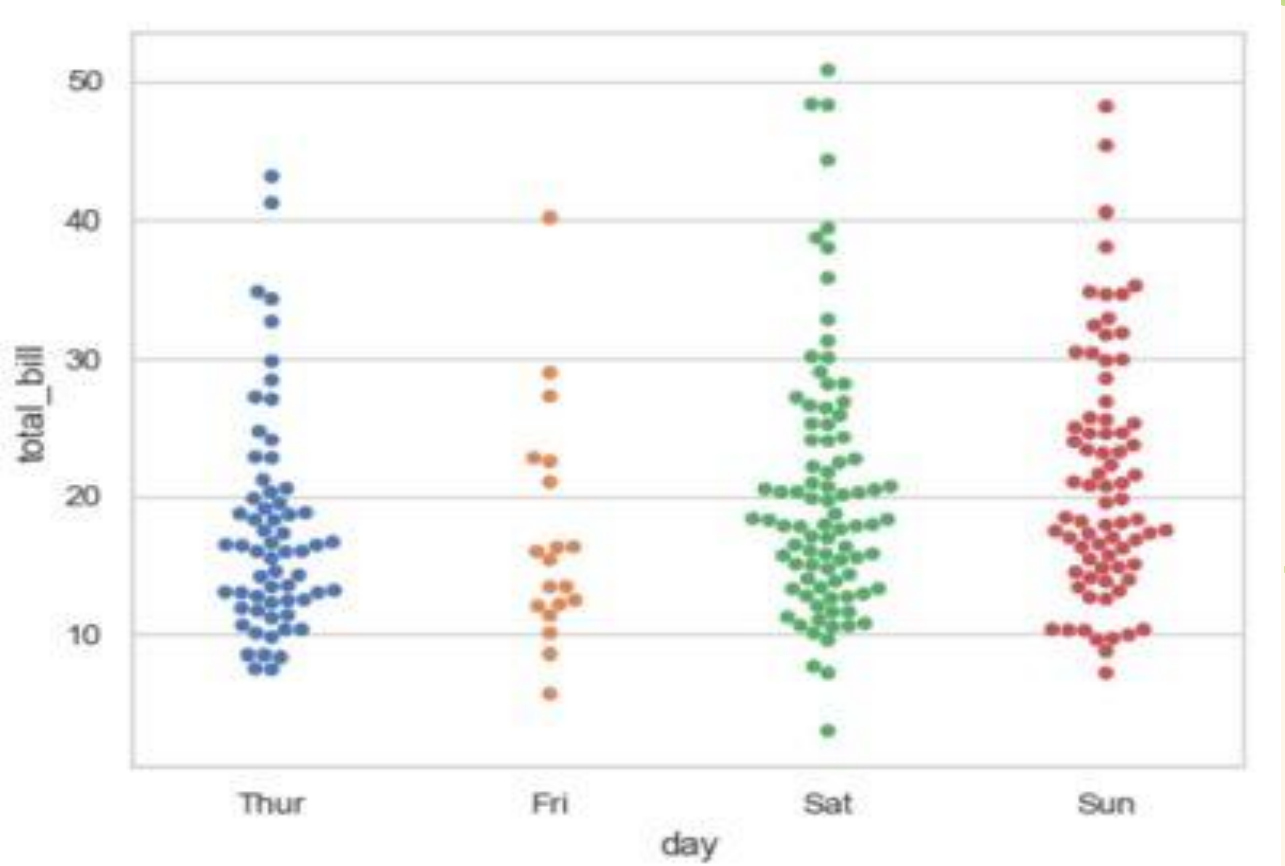



- **Scatter plot matrix** of the Iris data set with 4 features.
- It can visually represent multiple features of a data set to explore their relationship and discover hidden patterns.
- The plots on the diagonal are density plots, where each color represents a different species.

# What the Scatter Plot Matrix Shows

- The scatter plots show that the species versicolor and virginica share similar sepal and petal lengths.
- The petal lengths of all setosa  are about the same, and remarkably shorter than that of the other two species.
- These plots also show that the species are well separated in some features.
- This separation of the species in feature space facilitates the classification problem.

# Swarm Plot

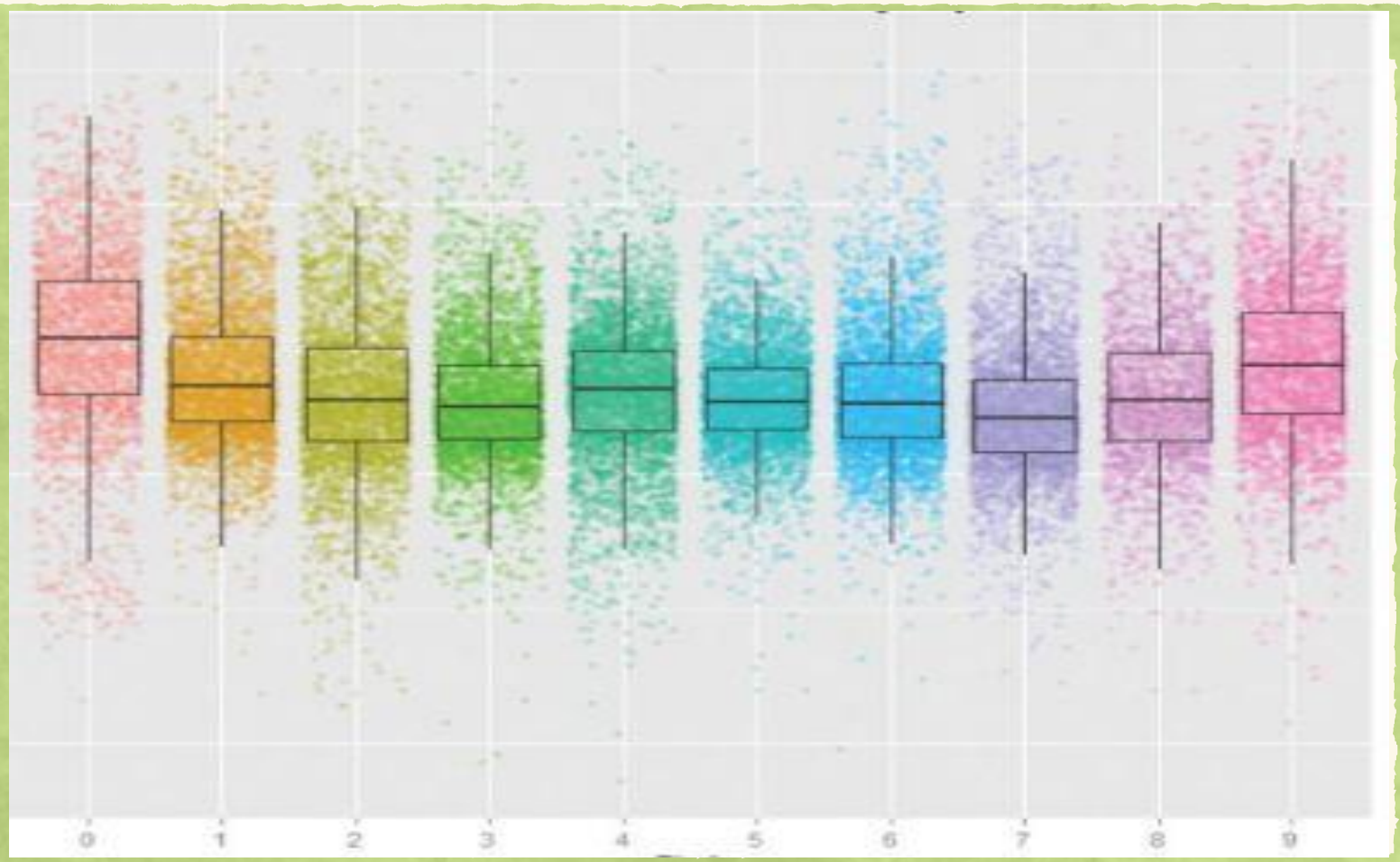



- A **swarm plot** compares data between different groups of categorical variables.
- It applies random jitter to move data points away from each other to avoid overlap.
- This figure shows restaurant bills of customers for different days of the week
- The days of the week here are **categorical variables**.

# Box-and-Whiskers Plot

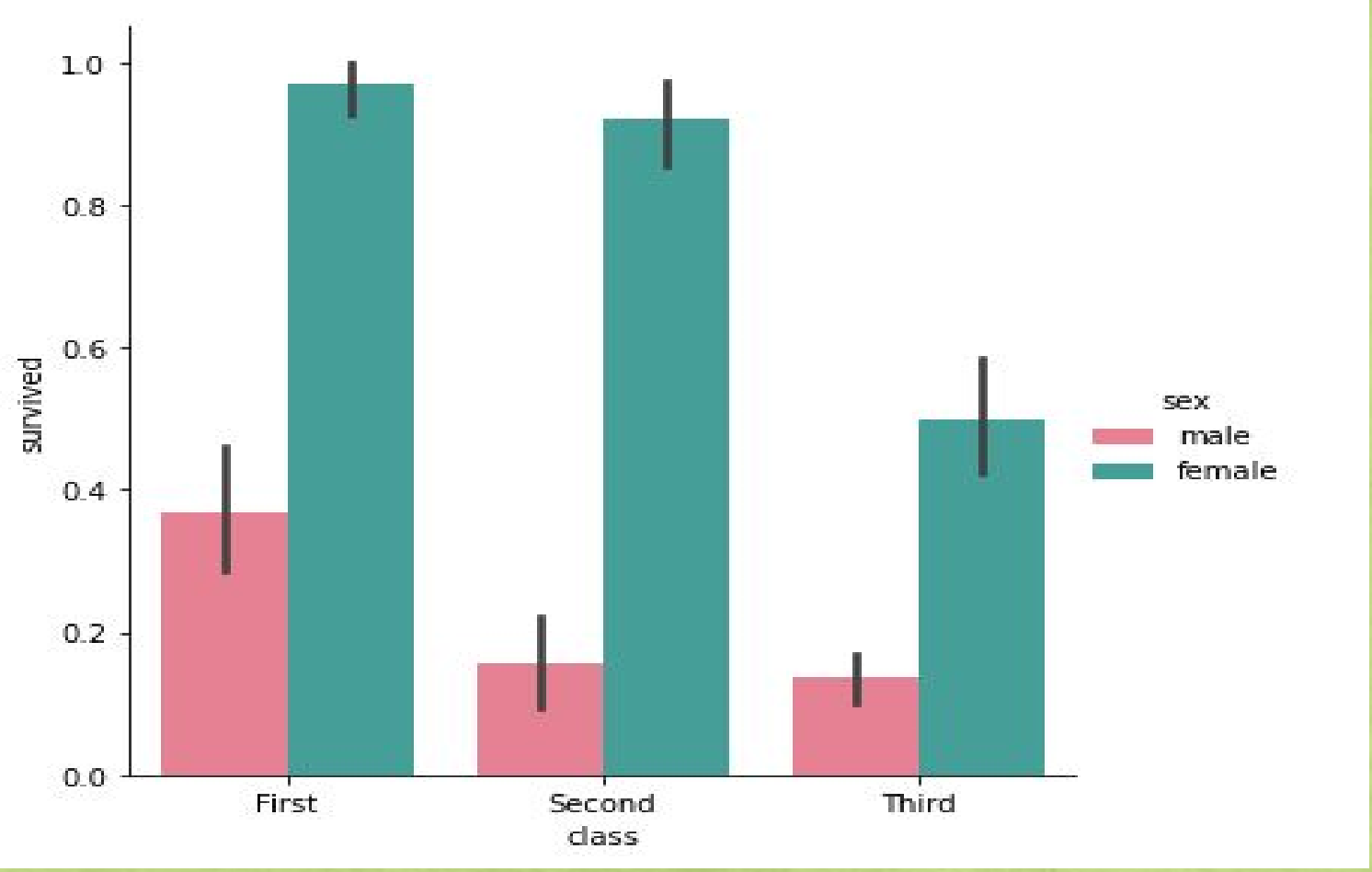
- The **box-and-whiskers** plot shows the distribution of numerical data in a way that facilitates comparison between different levels of a categorical variable.
- The **box** shows the range that contains the central 50% of the data. The horizontal line inside the box is the median of the data.
- The upper and lower hinges (lines) of the box correspond to the first and the third quartiles of the data, so the box spans **interquartile range (IQR)**.
- The **whiskers** are the two lines outside the box. The upper whisker extends from the hinge to the highest value that is within 1.5*IQR of the hinge. The lower whisker extends from the hinge to the lowest value that is within 1.5*IQR of the hinge

# Example - Box-And-whiskers Plot




- This plot shows the mean household income (on a log scale) versus zip code [1].
- The horizontal axis represents the first digit of the zip code, ranging from 0 through 9.
- The zip code 0 includes Maine, Vermont and Massachusetts. The zip code 9 includes California and Hawaii.
- The highest mean incomes are in the zip codes 0 and 9.

# Example - Bar Plot



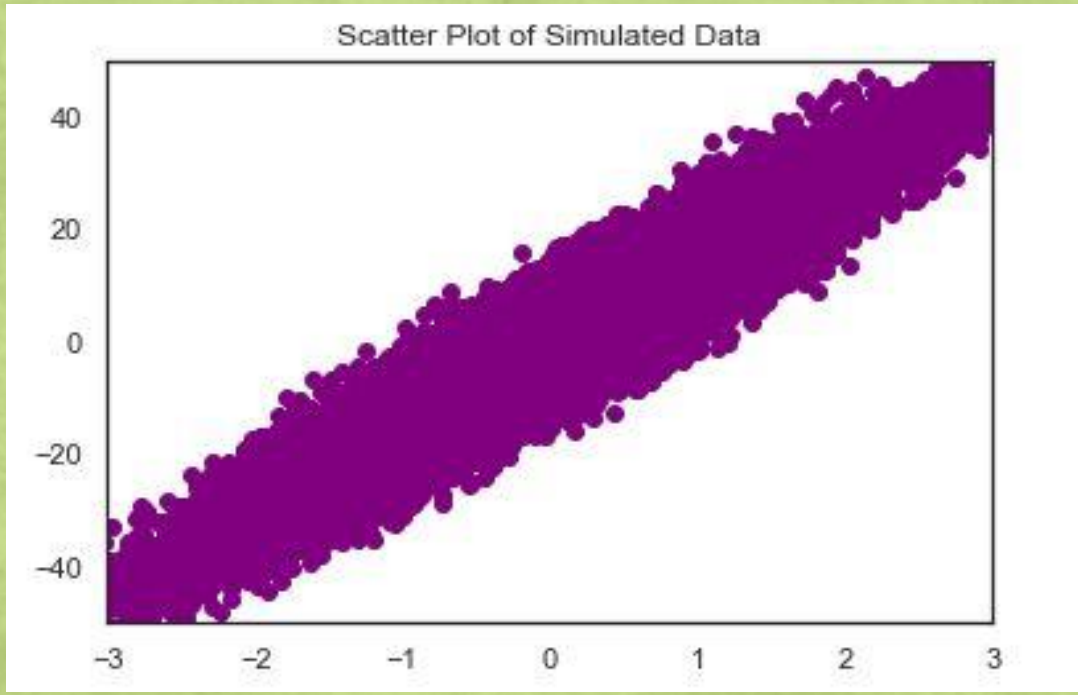

- A **bar plot** represents categorical data with rectangular bars, with height proportional to the values which they represent.
- The bar plot in this figure from **Titanic data** shows passenger survival for the three classes: first, second and third.
- As can be seen, the survival was much lower for the third class, compared to the other classes, and much lower for male compared to female.

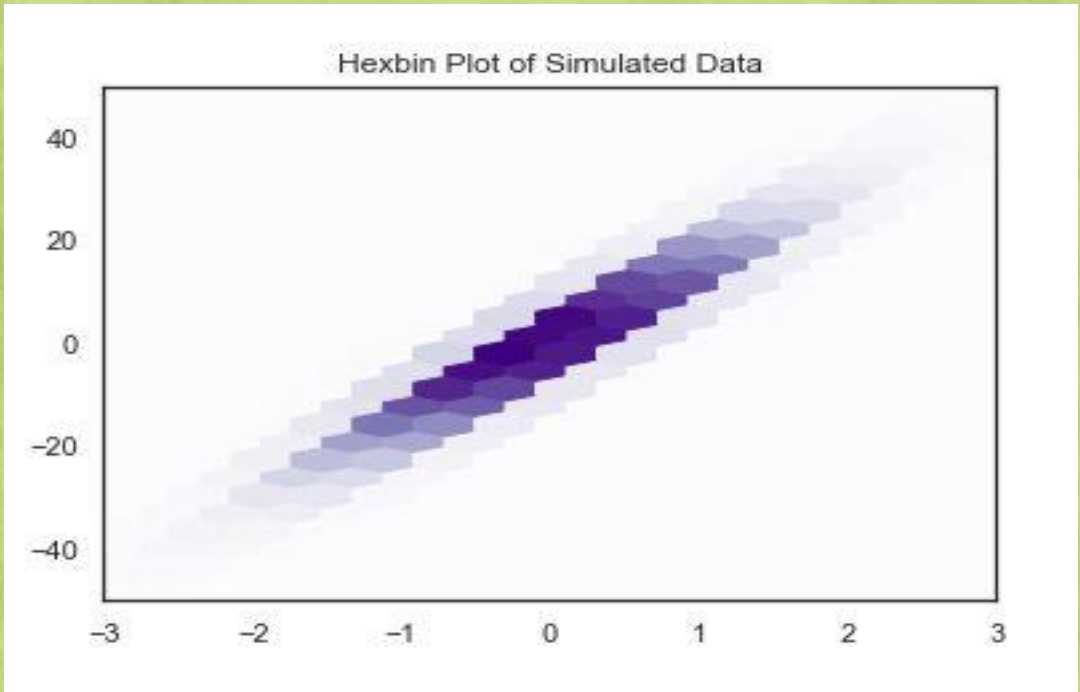
# Hexbinplot for Large Data Sets

- Scatter plot is a popular visualization tool. However, if the amount of data becomes too large, for example, with millions of data points, then it is difficult to see the structure of data in a scatter plot.
- In this case, a better alternative is the **hexbinplot**. Similar to a scatter plot, a hexbinplot visualizes data in the x-axis and the y-axis.
- Instead of **overlapping**, the plotting window is **split into several hexbins**. Data is placed in hexbins and **shading** is utilized to represent the concentration of data in each hexbin.

# Example - Hexbinplot

- The following plots show scatter plot (left) and hexbinplot (right), with simulated data and 50,000 data points.



# References

1. EMC Education Services, *Data Science and Big Data Analytics*, Wiley, 2015.