Chapter 3

Data Visualization in python™

Contents:

Line graph, more lines in a single graph, Scatter plots, Bar graphs, Histograms, Pie-charts, Grouped barcharts Heatmap visualization, computing statistical quantities and plotting them;

You need to install: matplotliib module

Then, you need to insert the following two lines at the top of your code, if you are using IDLE

```
matplotlib.use('TkAgg')
 # sometimes the above line may not be needed!
import matplotlib.pyplot as plt
```

Reference: https://matplotlib.org/users/pyplot_tutorial.html

Data Visualization

Data visualization is an important component of data engineering and science and the whole aim is to make the user understand what data reveals in the required context. Hence, we need to know how to represent a given data in a given context.

Remember this always! *Data is different from Information!*

Let's start exploring a few basic representative functions!

Simple Line/Curve plotting

You need X, Y values, stored in the form of **lists**.

Example 3.1

Let us try plotting Y = 2X graph

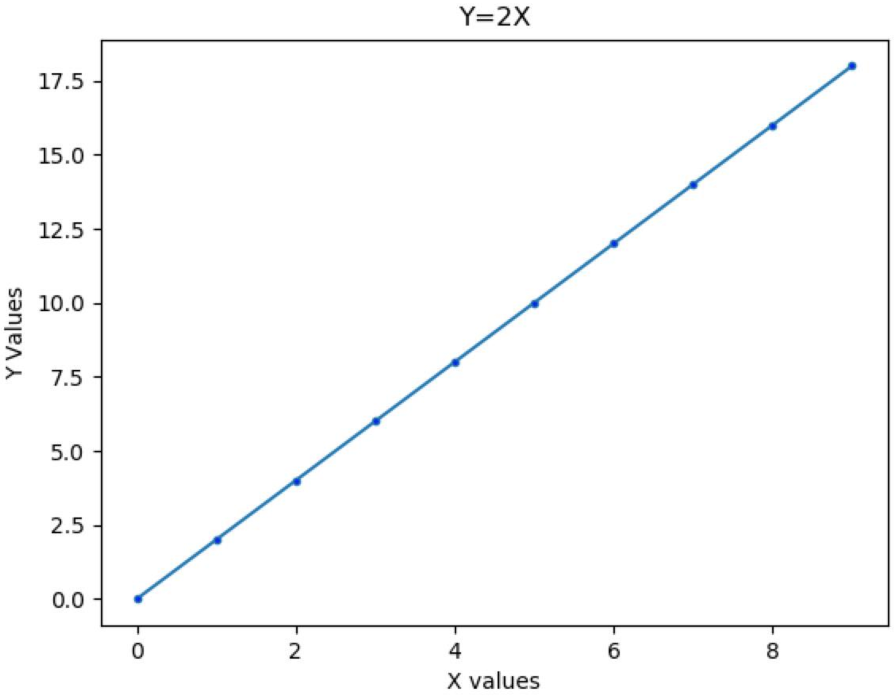
We need a plotting function and that is given by matplotlib module - plot() function – Following can be controlled in the plot function:

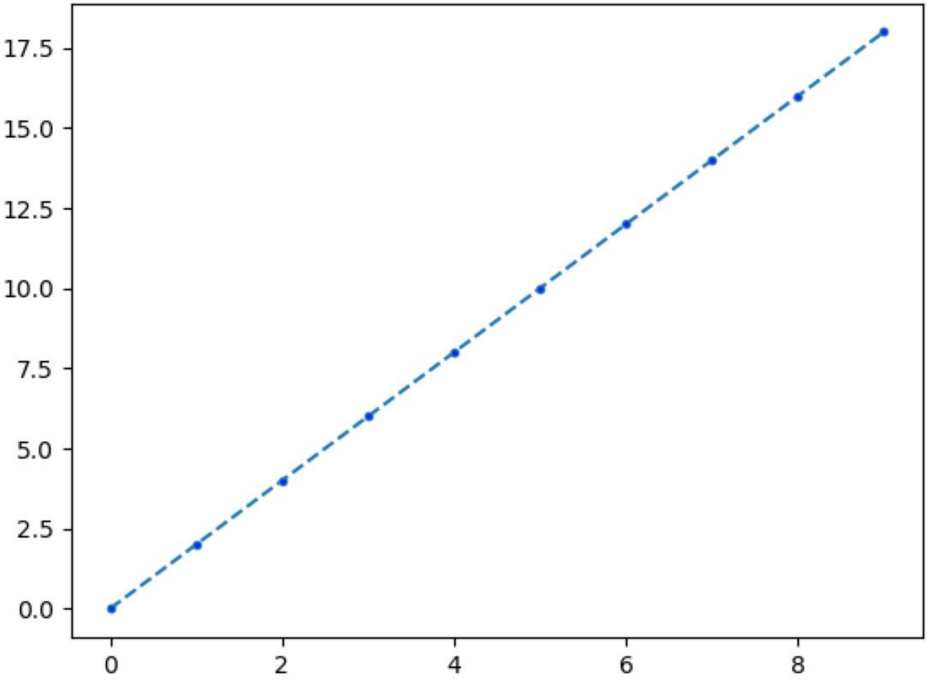
- Type of line, color, width of the line, labeling x and y axes, defining range of x and y values, legends, etc

See how to use this function via an example.

Detailed documentation:
https://matplotlib.org/api/_as_gen/matplotlib.lines.Line2D.html#matplotlib.lines.Line2D

You should see a plot like this….





Simple Line/Curve plotting

Plotting more than one line on the same graph

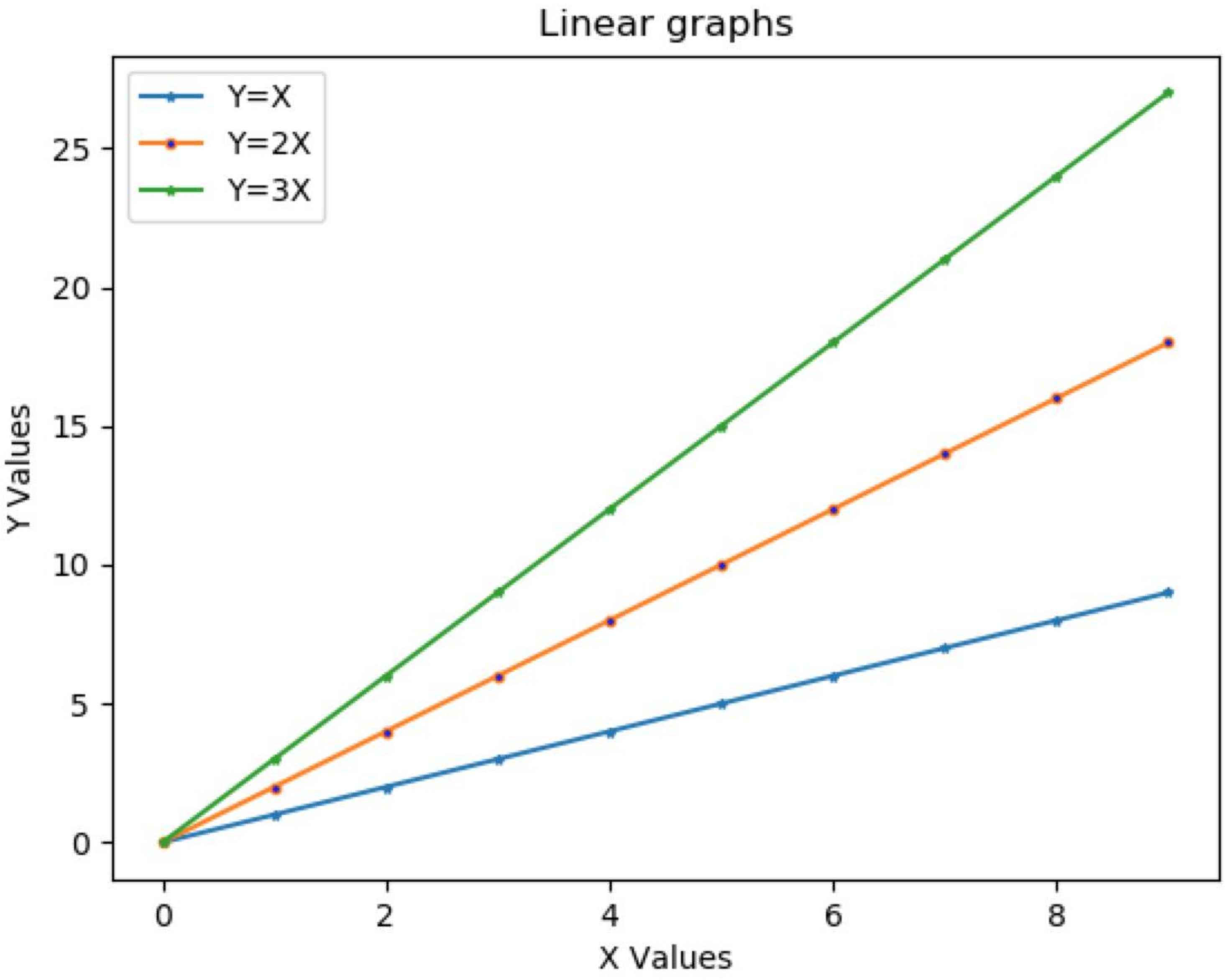
Example 3.2

Let us try plotting Y = 3X, Y = 2X, Y = X graphs

- You need to color each line separately, put labels for x and y axes, give your plot a title, identify the lines with respective legends.

Detailed documentation:
https://matplotlib.org/api/_as_gen/matplotlib.lines.Line2D.html#matplotlib.lines.Line2D

You should see a plot like this….



Simple Lines/Curves plotting

Plotting more curves on the same graph

Example 3.3

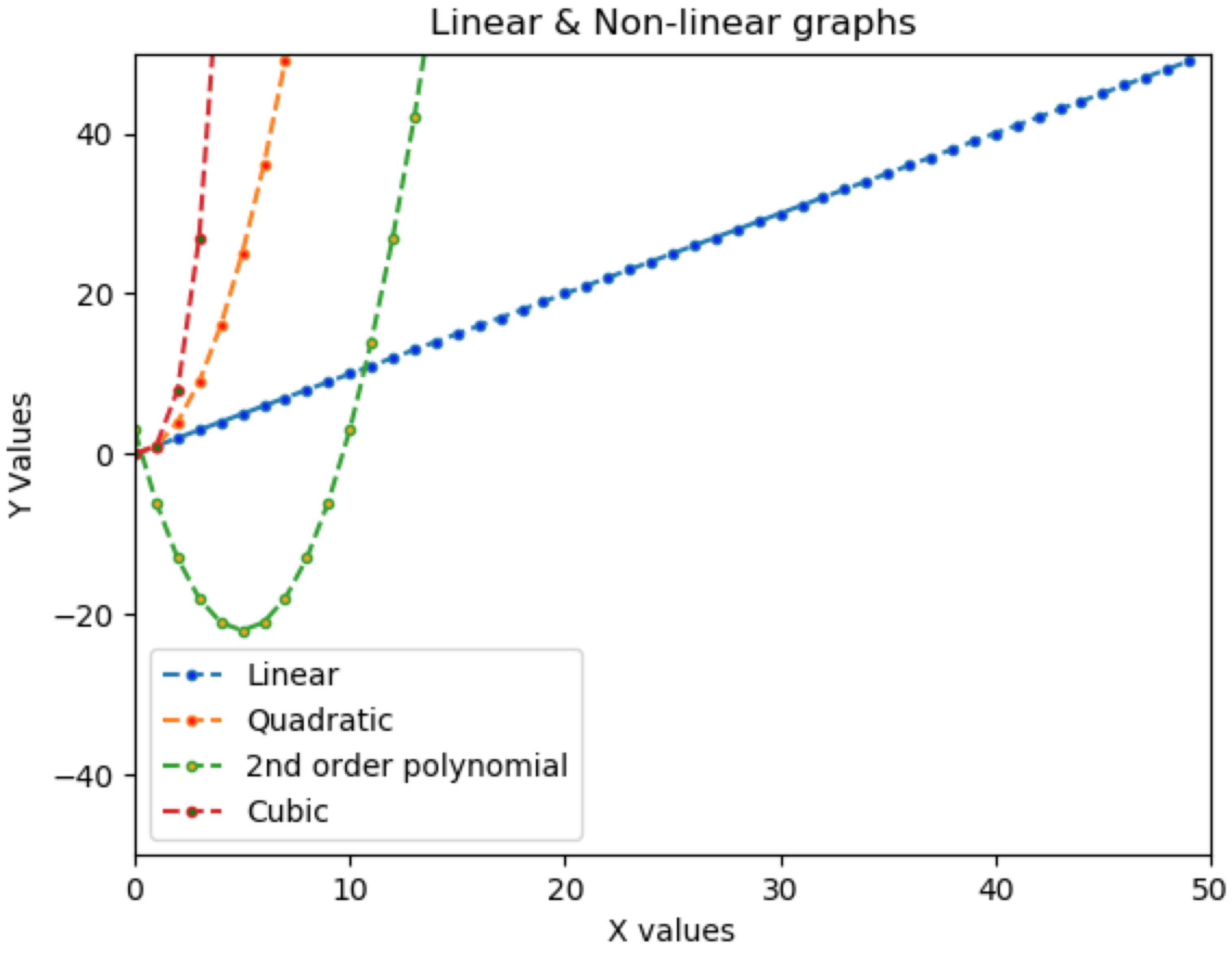
Let us try plotting $Y = X$, $Y = X^2$, $Y = X^3$ and $Y = X^2-10X+3$ graphs. Try to adjust the x range and y range and get a decent display.

You need to color each curve separately, put labels for x and y axes, give your plot a title, identify the curves with respective legends.

Detailed documentation:

https://matplotlib.org/api/_as_gen/matplotlib.lines.Line2D.html#matplotlib.lines.Line2D

You should see a plot like this….



(c) Bharadwaj V Dept of ECE NUS (2023)

Scatter Plots

Example 3.4

Generate a set of 150 random (x,y) values and present as a scatter plot. Scatter plots are useful to see the “spread” and “trend” in the data set. We may identify outliers!

As before, put labels for x and y axes, give your plot a meaningful title

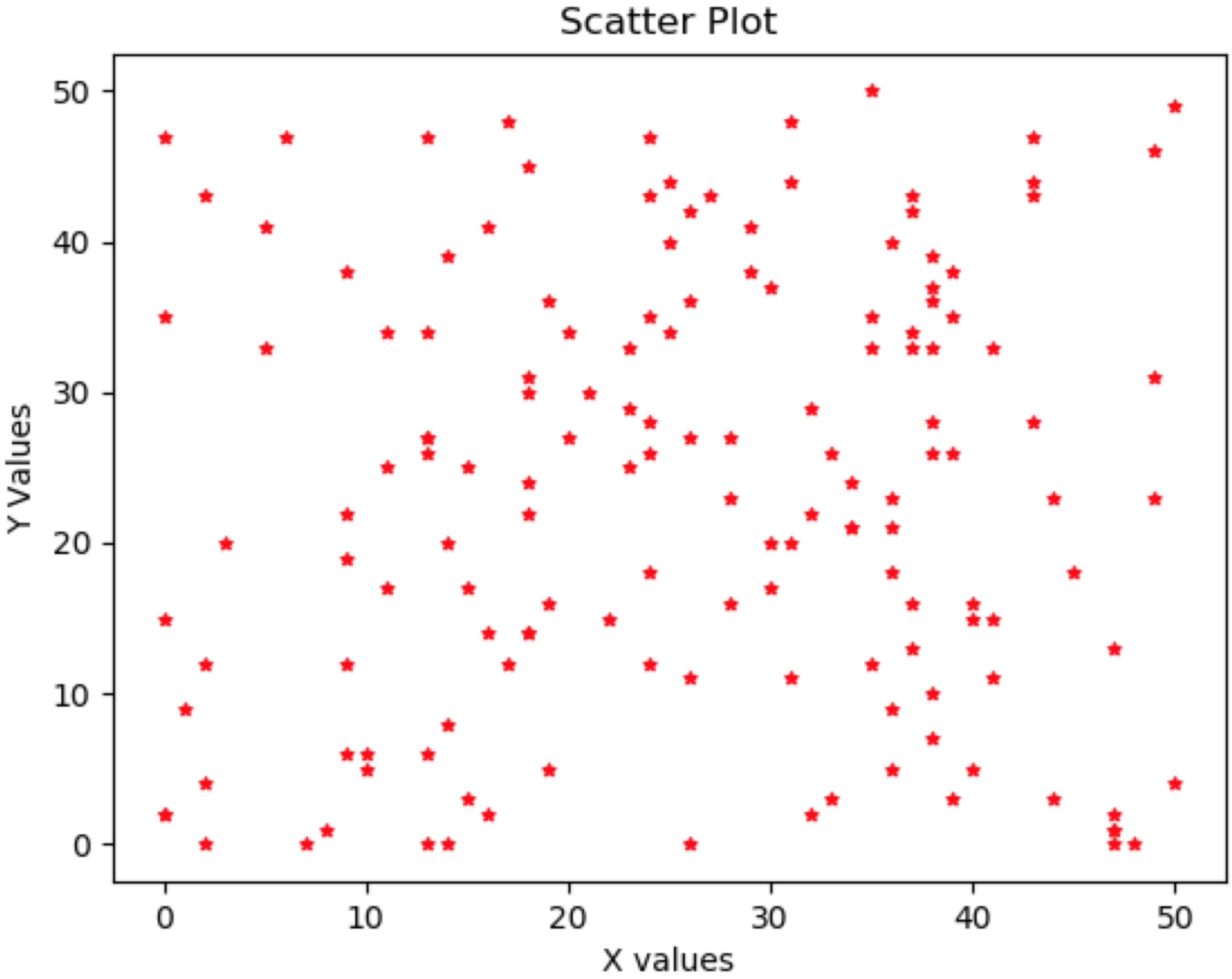
Example 3.5 – Plot the following as a Scatter plot and see

GirlsMarks = [89, 90, 70, 89, 100, 80, 42, 100, 80, 35]
BoysMarks = [30, 29, 73, 48, 100, 48, 38, 45, 81, 30]

MarksRange = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]

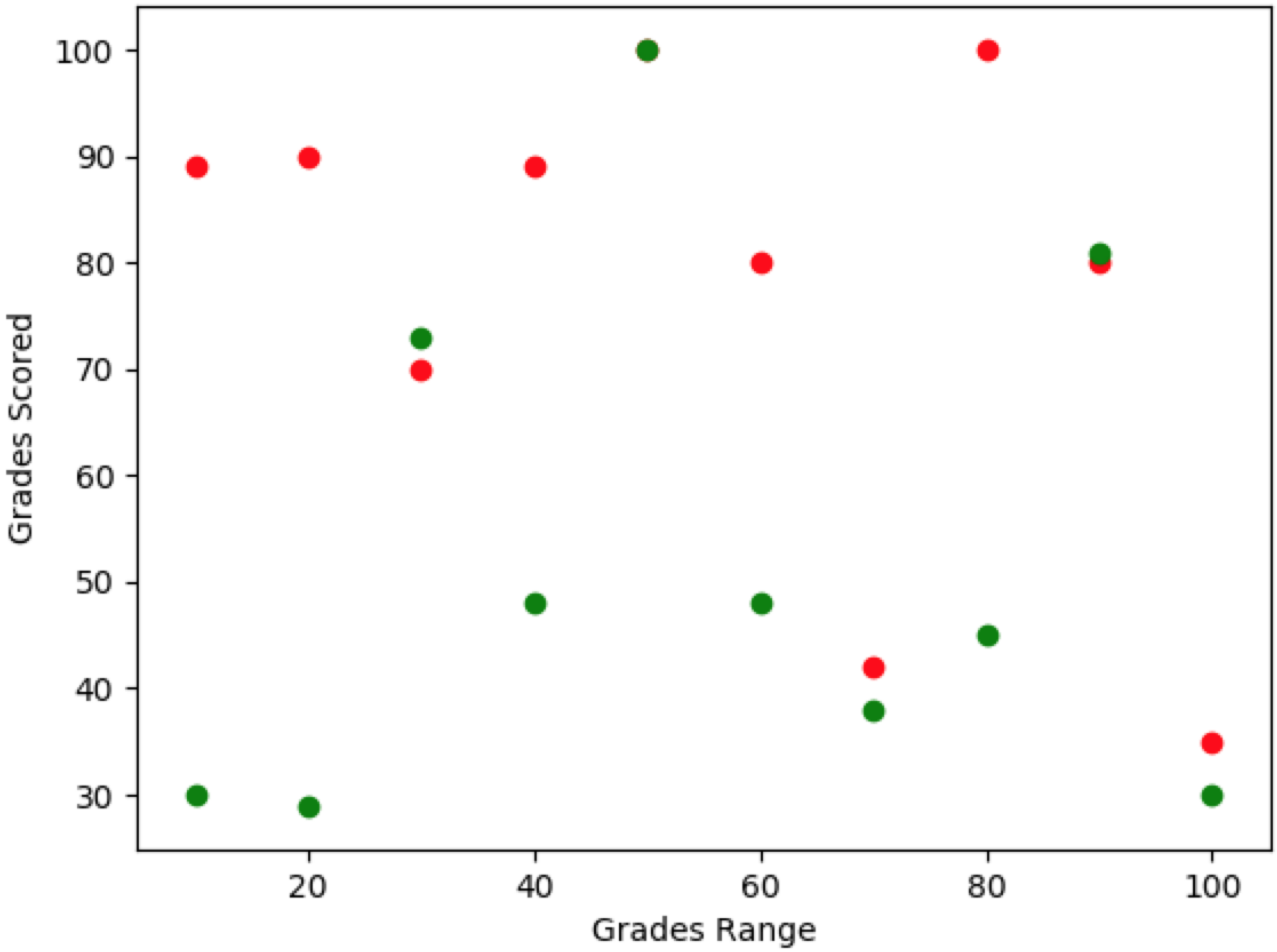
You should see a plot like this….

Example 3.4



You should see plot like this….

Example 3.5



Can you put a legend for this plot to Identify the categories?

Bar Graphs

Bar graph plotting is a better way to get insights into the frequency of distribution of a set of items. It is straightforward to plot a bar graph *with a few input parameters*. We quote here a few.

- Number of X values (given)
- Frequency of items for each x, in the y-axis (given or to be computed)
- Correct labels for X and Y axes and a title
- Decide vertical or horizontal bar representation
- Other cosmetic items – color, width of the bars, indication of peak values, etc.

Bar Graphs

Example 3.6

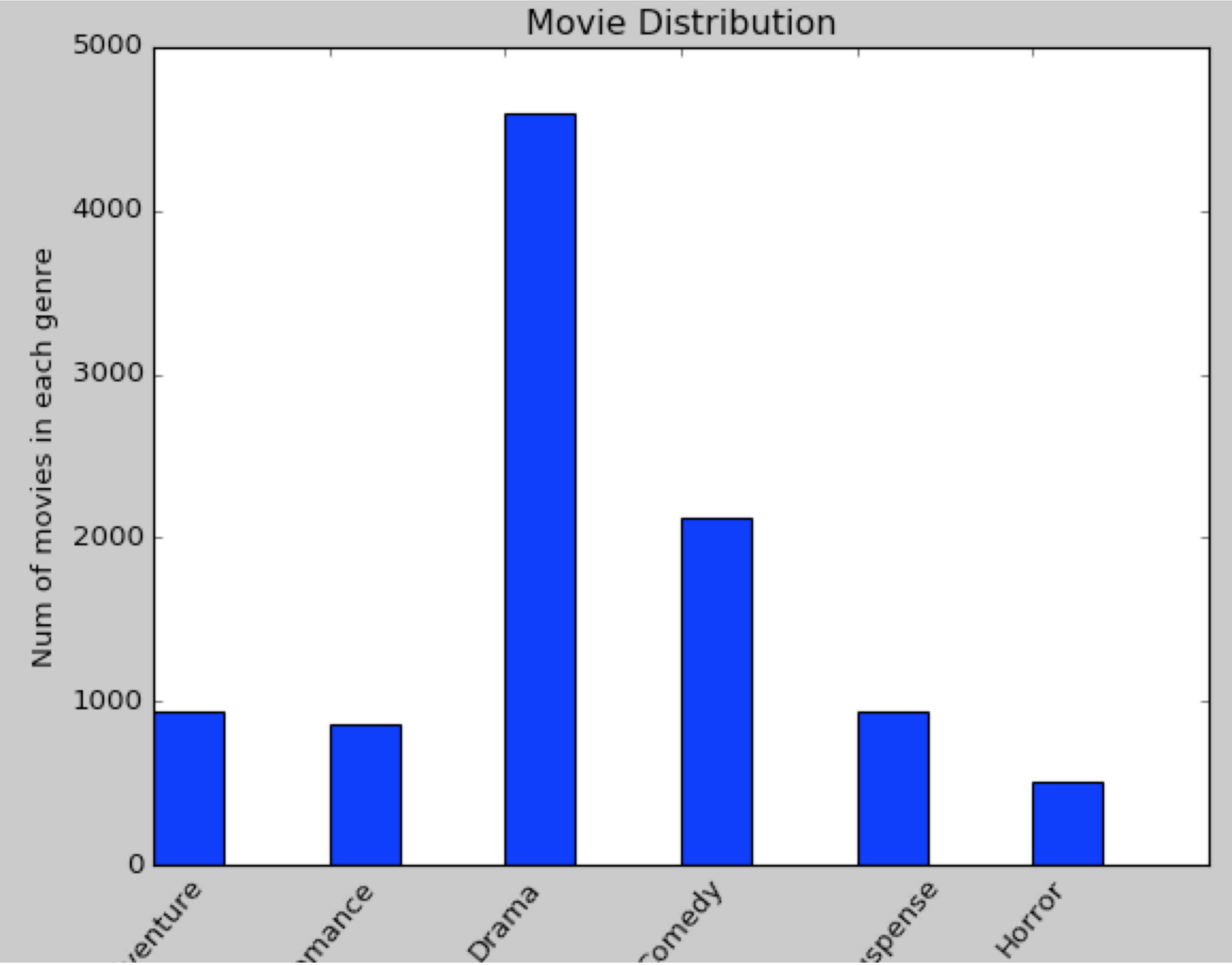
It will be interesting to see the distribution of data pertaining to movie genres over number of years. We are given this data. We can plot a bar graph to analyse the movie distribution.

The categories of movies are: 'Adventure', ’Romance', 'Drama', 'Comedy', 'Thriller/Suspense', 'Horror’

The respective number of movies is given by: 941, 854, 4595, 2125,942,509

Note: This data may be presented as a table or in any other form.

You should see a chart like this….

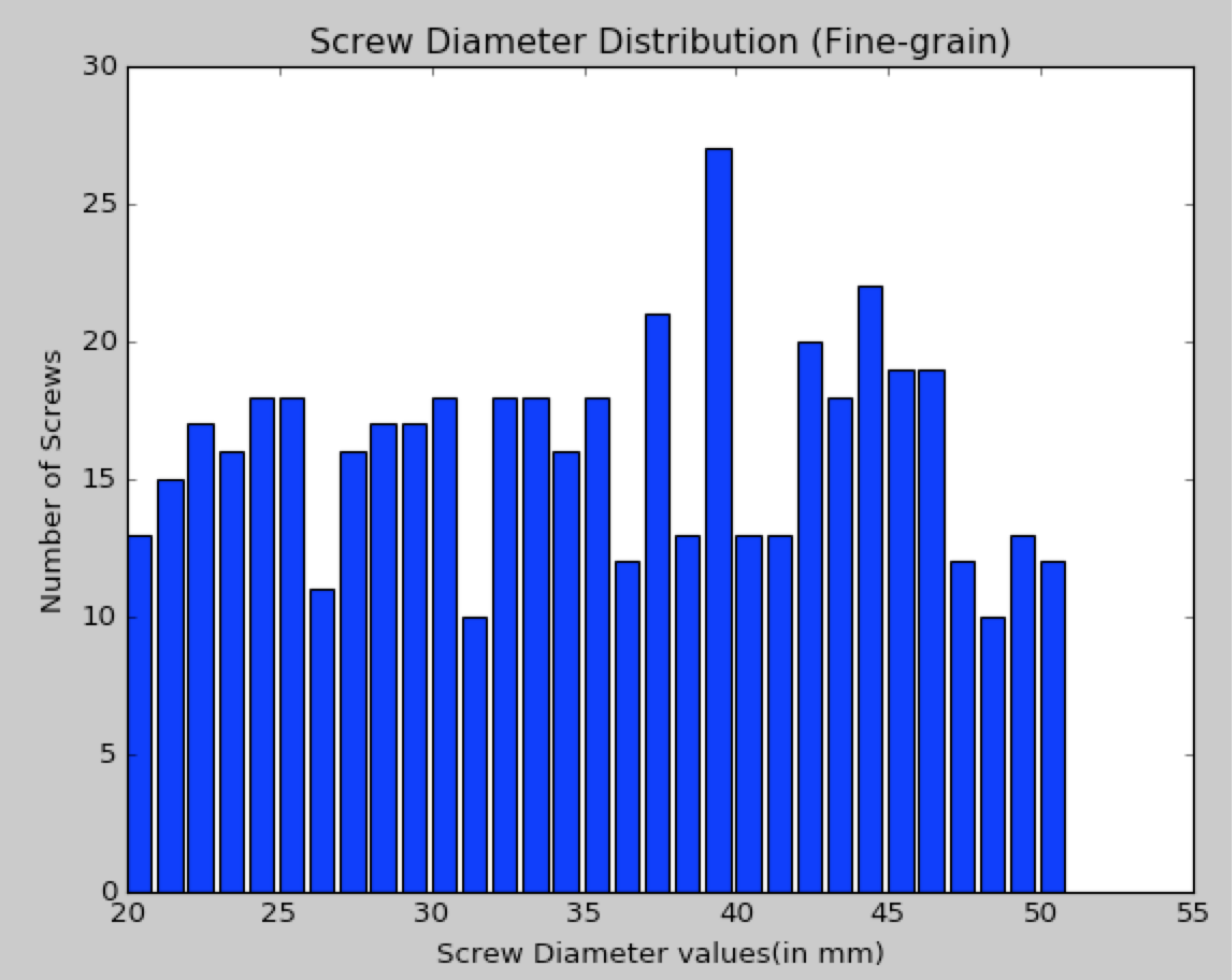


Example 3.7 - *A real-life application!*

A manufacturing unit that produces screws of varying diameters was subject to testing the samples by the auditors. 500 samples were taken from a lot and the range of screw diameter varies from 20mm to 50mm in the 500 size sample. Generate a bar graph and show the distribution of the screw diameters.

Data preparation - First generate random frequency list in the above range of diameters. Note that you need to derive the frequency and then plot a bar graph to infer.

You should see a chart like this….



Histograms

Pertaining to histograms there are number of parameters you can modify so that presentation looks cleaner, meaningful and easy to interpret. We quote here a few.

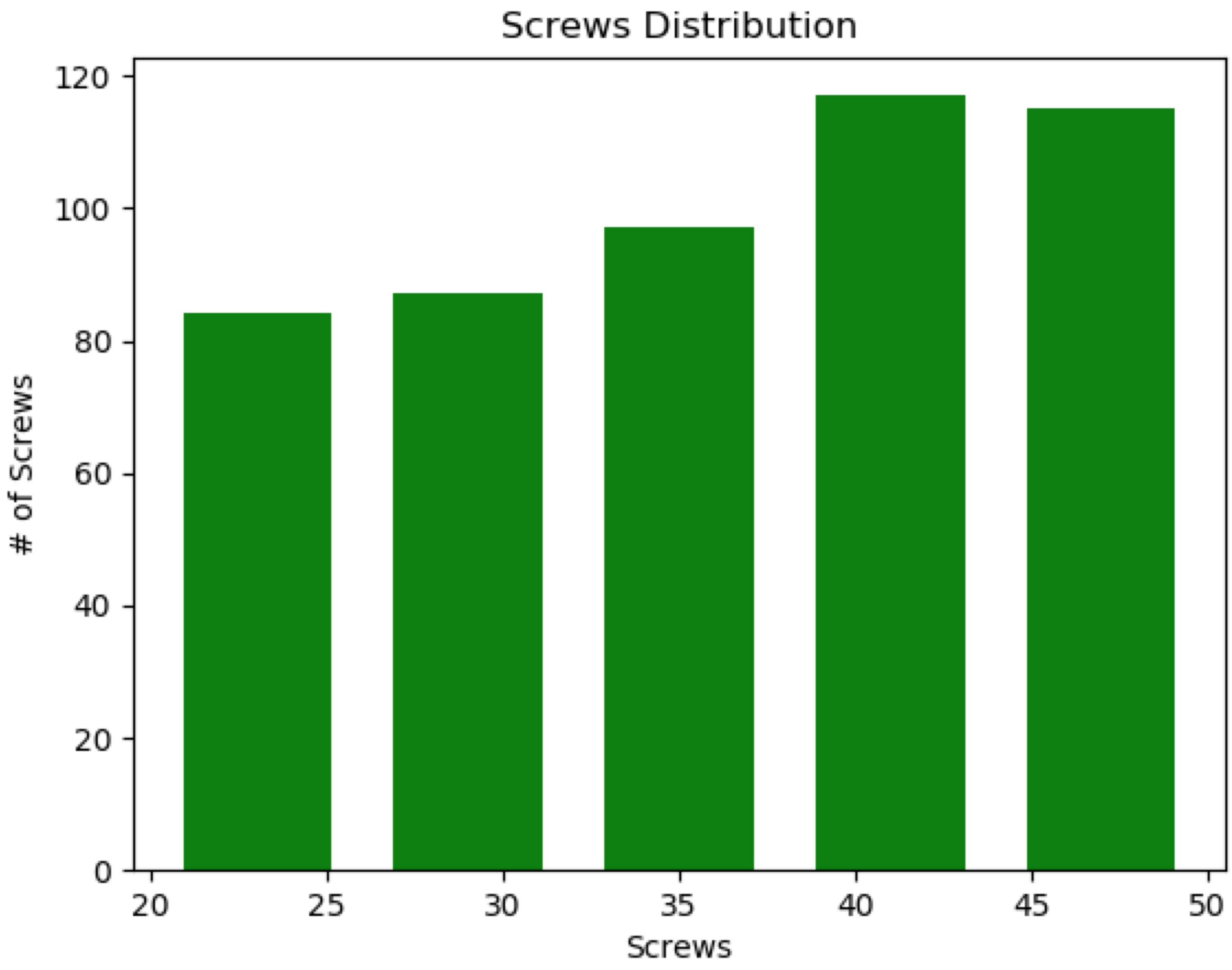
- Number of bins(classes) in the x axis (can be varied)
- Frequency of items in the y-axis (given or to be computed)
- Correct labels for X and Y axes and a title
- Decide vertical or horizontal bars
- Other cosmetic items – color, width of the bars, indication of peak values, etc

Example 3.8 - For the Example 3.7, plot a histogram.

- Demonstrate how bins and the range of X values can be adjusted so that visualization is coarse.
- Set bins to 5 and 10 see the differences.
- How this variation in number of bins offer different interpretations?

Observe that based on the parameters range and bins, the size(width) of each bin will be decided automatically

You should see a histogram like this….



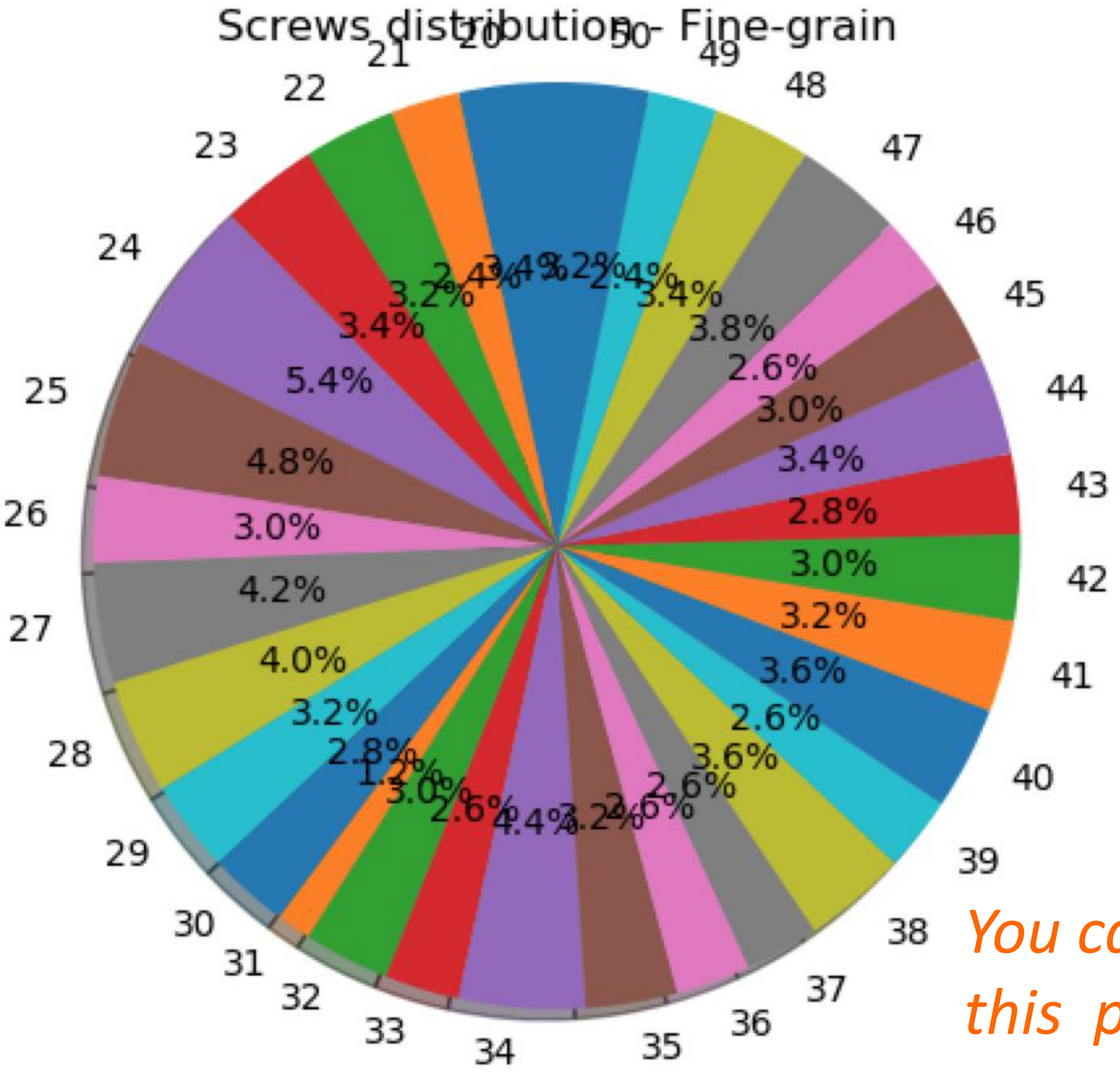
Pie Chart

Sometimes it will be useful to get a direct feel of distribution by computing the percentage of certain items and reporting them. A useful representation of this requirement is via plotting a Pie Chart. Use a Pie chart when the data items (features) are less, or else it will appear cluttered and purpose will not be served. We quote here a few.

- Number of Slices (classes)
- Radius of the circle
- Starting angle
- Representation using % values
- Other cosmetic items – color, shadow effect, etc

You should see a pie-chart like this….

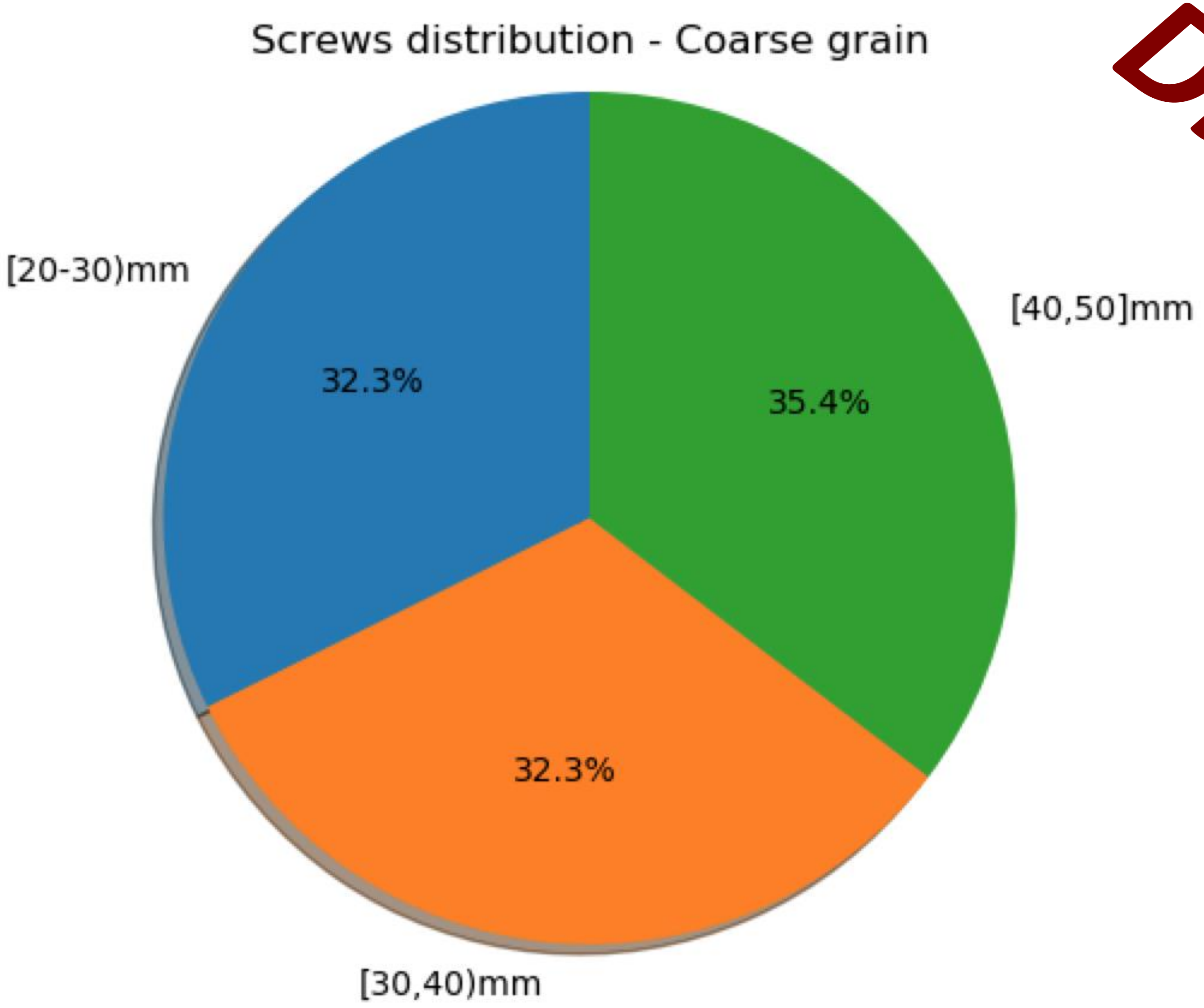
Example 3.9



(Pie-Chart for Example 3.7)

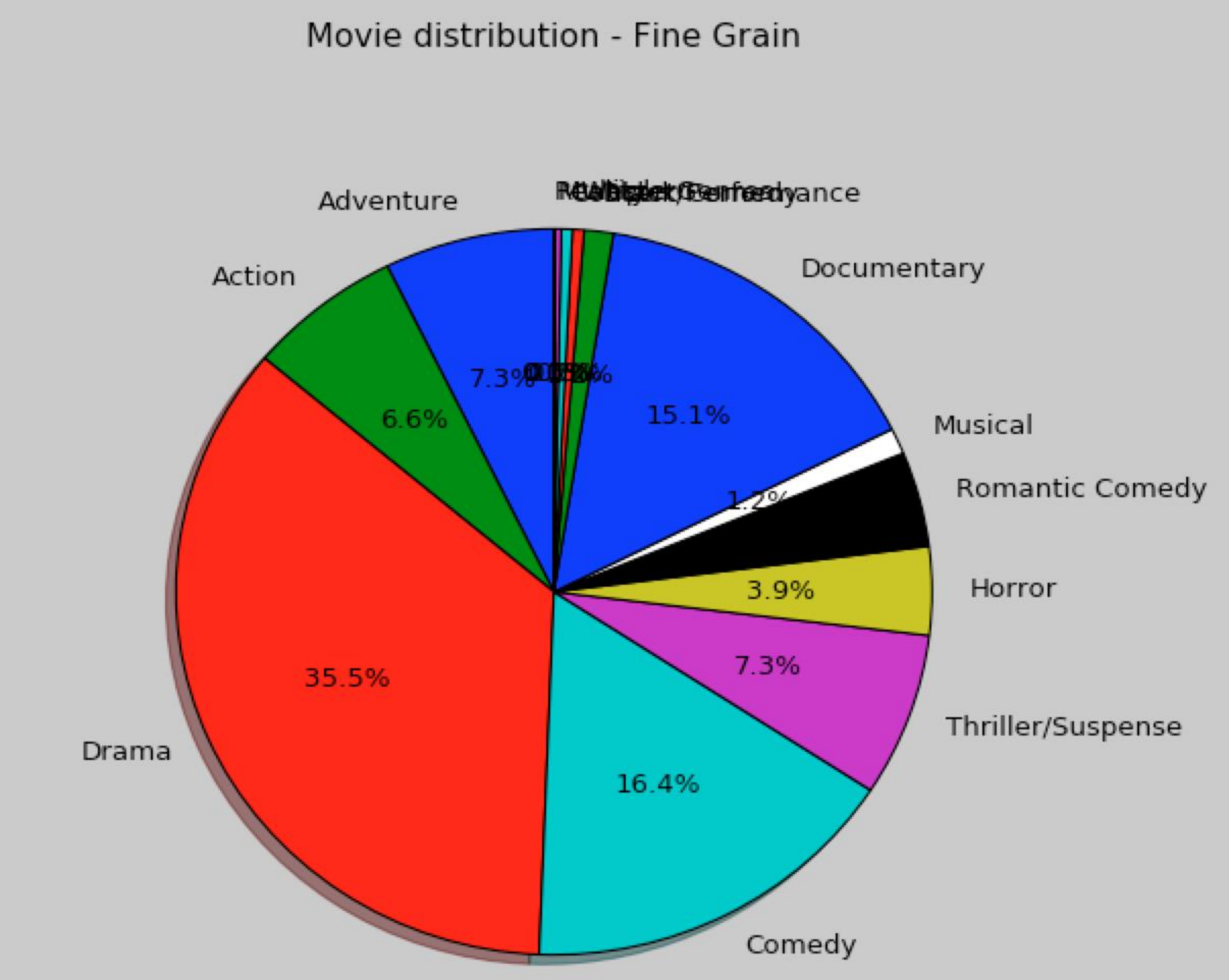
You can make this plot coarse!

You should see a pie-chart like this….



DIY!

You should see a pie-chart like this….



(Pie-Chart for Example 3.6)

DIY! – Try to make this coarse!

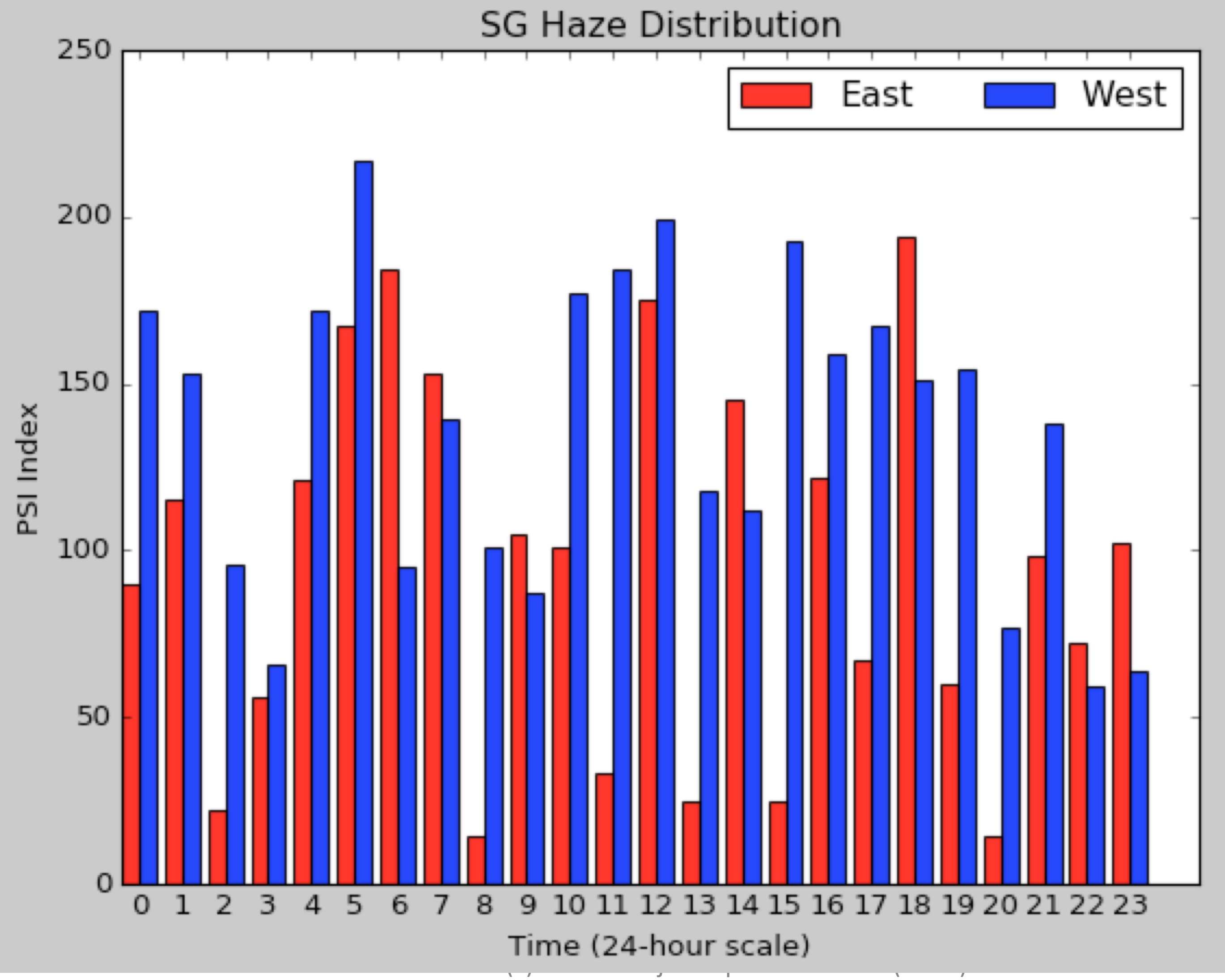
Grouped Bar Charts

Example 3.10:

Application: Consider Singapore's haze data over a period of 24 hours in 4 zones (East, West, North, South). The data are available in separate text files, one for each zone. Do the following:

- Read the data from the files
- Plot a grouped bar chart
- Label X and Y axes clearly, put a legend to your plot, color each zone bar differently,

You should see something like this! (Plotted only for 2 zones for clarity)



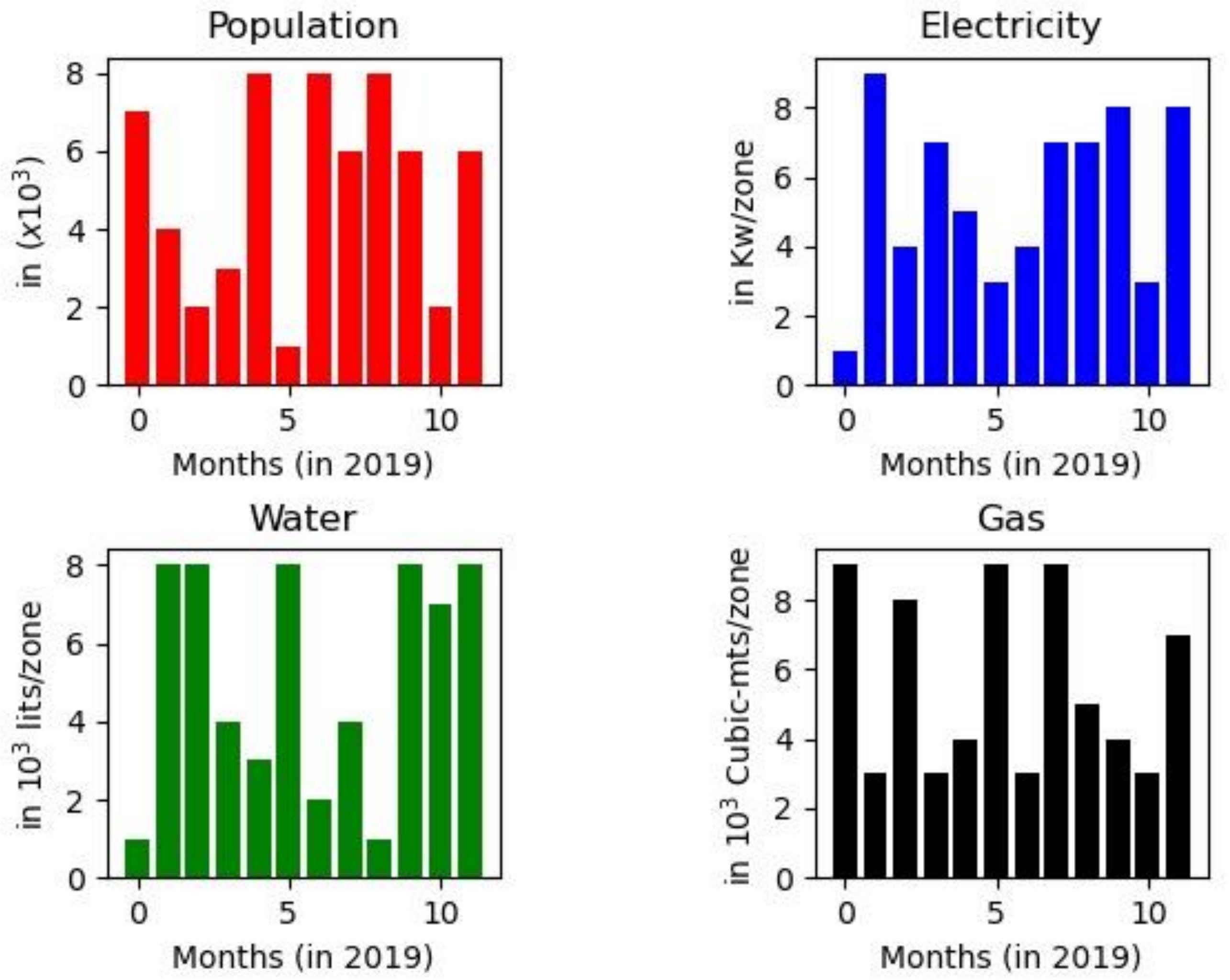
Multiple Bar Charts: Sub-plots

Example 3.11:

Application: If you wish to generate 4 different plots using same DF on different quantities and present as a collective view, you can use this representation.

- Given a population distribution and their consumption on Water, Gas, and Electricity, plot bar charts as sub-plots
- Use meaningful legends and axes labels and present the charts

You should see something like this!



Note: Plot generated for demo purposes. Plots used synthetic data and not real data;

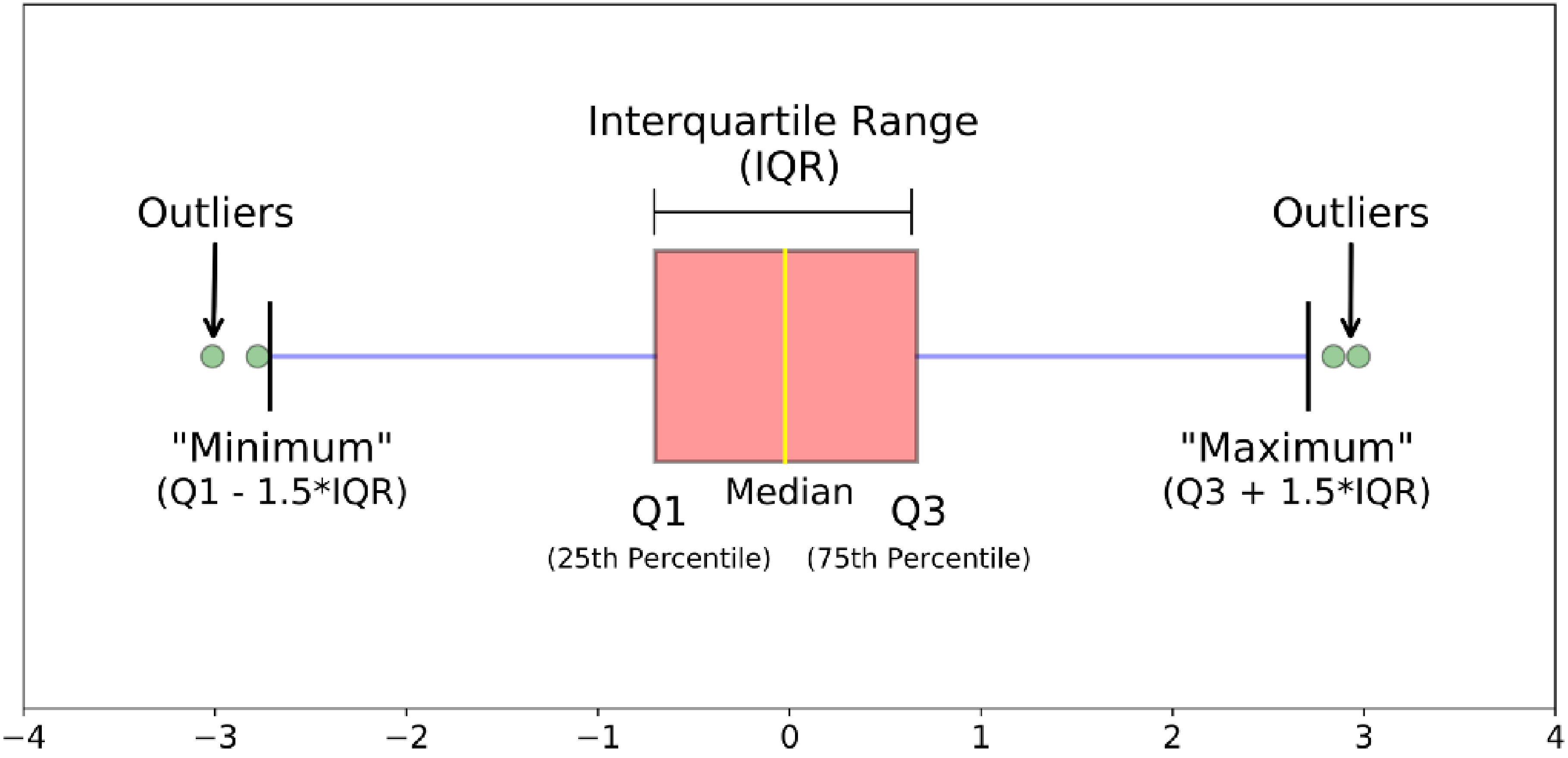
Box Plots

Useful plot in statistics to understand the distribution of the data

Boxplot is also used for *detect the outlier in data set*. It captures the summary of the data efficiently with a simple box and whiskers and allows us to compare easily across groups. *Boxplot summarizes a sample data using 25th, 50th and 75th percentiles.* These percentiles are also known as the lower quartile, median and upper quartile.

Box plot is constructed with 5 quantities

- Minimum
- First Quartile or 25%
- Median (Second Quartile) or 50%
- Third Quartile or 75%
- Maximum


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
(25th Percentile)
Median
Q3
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4
−3
−2
−1
0
1
2
3
4

Example:

Data on the heights of 40 students in a class

[59 60 61 62 62 63 63 64 64 64 65 65 65 65 65 65 65 65 65 66 66 67 67 68 68 69 70 70 70 70 70 71 71 72 72 73 74 74 75 77]

Construct a box plot with the following properties:

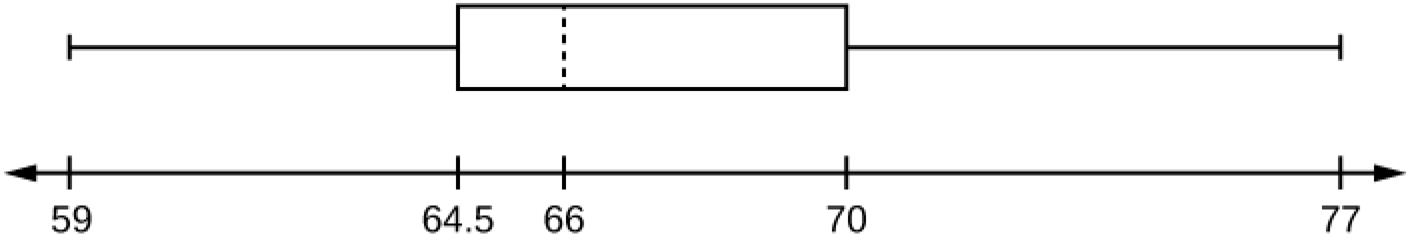
Minimum value = 59

Maximum value = 77

*Q*1: First quartile = 64.5

*Q*2: Second quartile or median= 66

*Q*3: Third quartile = 70



Example 3.12:

Create a box plot for the bodyfat data - bodyfat_Example 2.2.csv; With respect to ‘Category’ & ‘bodyfat’ features, plot a boxplot and identify the respective statistical quantities.

You may use tips.csv data (via kaggle or Github) and try plotting a boxplot w.r.t the total bill spent on different days.

Data correlation – Heatmap visualization

- If we have a DF with numerical values for the features, we can perform a correlation operation:

>> DF.corr(method ='kendall') # also try “spearman” (These methods use rank computations)

Note: If your dataset has any **naNs**, you need to do fill them up before using **corr();**

The output of this **corr()** is also a DF and hence any cell can be accessed using its indices; As we expect, all diagonal entries will be 1;

After generating the correlation matrix (your DF), we can plot a ***heatmap*** for visualization.

Example 3.13: Generate a correlation matrix for *bodyfat_Example_Correlation* , print and visualize using a heatmap.

Final Remarks:

Data visualization is an important component of data engineering and science and the whole aim is to make the user understand what data reveals in the required context. Hence, appropriate plotting functions need to be chosen!

From now on, if you look at any real-life data pertaining to any information/situation/context, identify which plot would give the best representation to extract the underlying information!

Remember this always! *Data is different from Information!*

Thank you!