# **Week 11: Lab pre-read**

### 1. Exploratory Data Analysis via Data Visualisation

Large datasets are complex entities. They may capture a huge range of parameters for a large set of individuals over a long period of time. Take the example of weather data. It may contain measurements on a range of parameters - humidity, temperature, wind speed, pressure, cloud cover, for multiple different geographical locations for each day of the last *n* years. This data presented in a tabular form is unintelligible for most. But put it in the form of a line chart tracing the temperature over the past year, and we will be able to promptly identify the hottest week of the year.

#### 1.1 Working with an Image

Load an HD image in your python script. Select a patch of the image and show it in tabular form. Can you make out what the image contains? Now plot the same patch and share what it contains.

import matplotlib.pyplot as plt

import matplotlib.image as img

path = '/content/image.png'

plak = img.imread(path)

# image data numerically

print(plak.shape)

print(plak)

# displaying the image

plt.figure(figsize=(8, 6))

plt.imshow(plak)

plt.show()

#### 1.2 Data visualisation tools

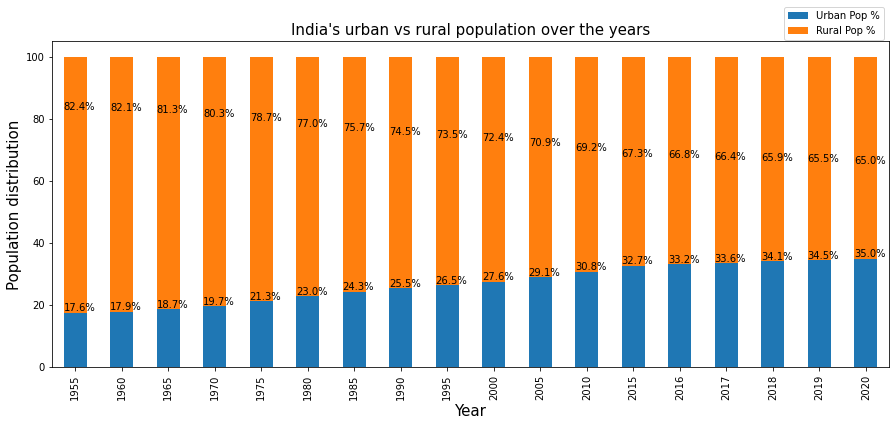
Visualisation is a great tool for building intuition about the data and for illuminating aspects of it that may require further investigation. Thus, while visualisation is handy in presenting results, it becomes even more critical in the exploratory phase of any new, sizeable dataset. It gives us a framework for asking conceptual questions, then querying and showcasing the dataset to shed light on the answer.

We can thus organise the visualisation tools based on the types of questions they can help illuminate. Questions about:

##### Composition

Pie chart: A disc which is sliced into as many sections as the constituent factors and whose sizes are proportionate to the relative sizes or amounts of associated factors.

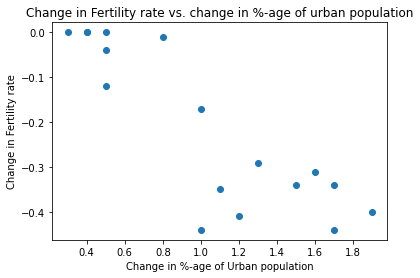
Stacked bar graphs: In order to compare the proportions of factors across individuals or over time, we unroll the pie charts, and visualise each pie chart as a ‘stacked bar’ in this visualisation.



##### Distribution

Histogram: For observing distributions in 1D i.e. the size of each group in a serialisation of groups within a population, we use a histogram. (line density plot).

Scatterplots (2D, 3D): Histograms in higher dimensions take the form of scatter plots. A concentration of points in a given region of the high dimensional space corresponds to a high.



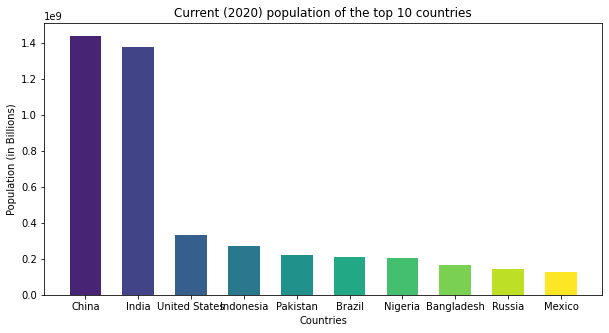
##### Comparison (in time)

Heatmaps: Heat map is a 2-D data visualisation matrix, where the values of a given variable of interest are divided into ranges across the horizontal axis and each cell’s colour indicates the value of the main variable in the corresponding cell range.

This is used for observing the variance across each subgroup/variables and identifying the pattern of correlations.

Bar graphs/line charts: A bar chart is used for comparing the distribution (count/frequency) of a categorical variable/feature to understand which subgroups/categories are highest or lowest, and how each group compares against each other.

Here each categorical value will be plotted as one bar, and the length of each bar represents the bar’s value.



Comparison (over time)

Line charts: Line charts are used to observe the changes in the value of a variable (Plotted on the Y-axis) with respect to continuous ordered points on the X-axis. Here data points are connected by a straight line.

It is used to observe the changes (trend, acceleration, deceleration, and volatility of the data) over time at given equal intervals of time. Multiple lines can also be plotted in a single line chart to compare the changes over time.

