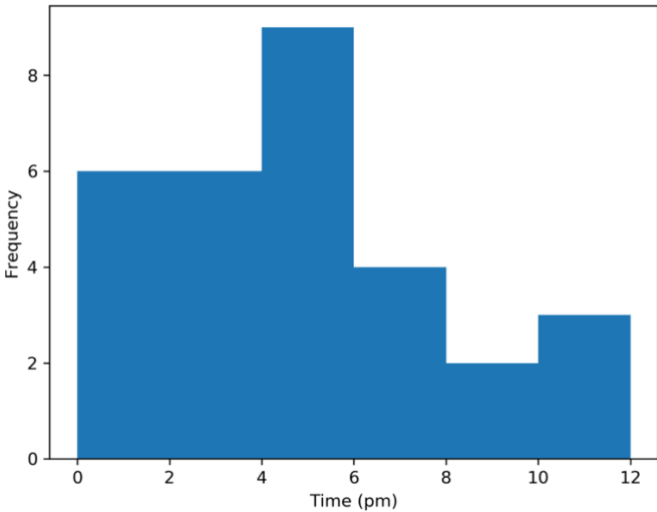


USN

Internal Assessment Test 1 – October 2024

Sub:	Data Visualization					Sub Code:	21AD71	Branch:	AInDS		
Date:	15/10/2024	Duration :	90 minutes	Max Marks:	50	Sem	VII			OBE	
<u>Answer any FIVE Questions</u>								MAR	C	RB	
								KS	O	T	
1	a	What is the need for data visualization? Explain its importance.						[6]	1	L1	
	b	What are the advantages of data visualization?						[4]			
2		What is measure of central tendency? Explain the different kinds of measures of central tendency in detail?						[10]	1	L2	
3	a	What are comparison plots? Explain the different types with diagrams.						[10]	2	L3	
4	a	What are dot maps? Mention some of the design practices followed.						[6]	2	L3	
	b	Explain indexing and splitting in NumPy.						[4]	1		
5	a	<p>1. Looking at the following histogram, can you identify the interval during which a maximum number of trains arrive?</p> <p>2. How would the histogram change if in the morning, the same total number of trains arrive as in the afternoon, and if you have the same frequencies for all time intervals?</p>  <p style="text-align: center;">Figure 2.38: Frequency of trains during different time intervals</p>						[10]	2	L3	
6	a	What are composition plots? Explain pie chart & donut charts with diagrams.						[10]	2	L2	

1.

a.

- i. Unlike machines, people are usually not equipped for interpreting a large amount of information from a random set of numbers and messages in each piece of data. Out of all our logical capabilities, we understand things best through the visual processing of information. When data is represented visually, the probability of understanding complex builds and numbers increases.
- ii. Representations can narrate a story and convey fundamental discoveries to your audience. Without appropriately modeling your information to use it to make meaningful findings, its value is reduced. Creating representations helps us achieve a more precise, more concise, and more direct perspective of information, making it easier for anyone to understand the data.
- iii. Instead of just looking at data in the columns of an Excel spreadsheet, we get a better idea of what our data contains by using visualization.

b. Visualizing data has many advantages, such as the following:

- i. Complex data can be easily understood.
- ii. A simple visual representation of outliers, target audiences, and futures markets can be created.
- iii. Storytelling can be done using dashboards and animations.
- iv. Data can be explored through interactive visualizations.

2.

Measures of Central Tendency

Measures of central tendency are often called **averages** and describe central or typical values for a probability distribution. We are going to discuss three kinds of averages in this chapter:

- **Mean:** The arithmetic average is computed by summing up all measurements and dividing the sum by the number of observations. The mean is calculated as follows:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

Figure 1.5: Formula for mean

- **Median:** This is the middle value of the ordered dataset. If there is an even number of observations, the median will be the average of the two middle values. The median is less prone to outliers compared to the mean, where outliers are distinct values in data.
- **Mode:** Our last measure of central tendency, the mode is defined as the most frequent value. There may be more than one mode in cases where multiple values are equally frequent.

For example, a die was rolled 10 times, and we got the following numbers: 4, 5, 4, 3, 4, 2, 1, 1, 2, and 1.

The mean is calculated by summing all the events and dividing them by the number of observations: $(4+5+4+3+4+2+1+1+2+1)/10=2.7$.

To calculate the median, the die rolls have to be ordered according to their values. The ordered values are as follows: 1, 1, 1, 2, 2, 3, 4, 4, 4, 5. Since we have an even number of die rolls, we need to take the average of the two middle values. The average of the two middle values is $(2+3)/2=2.5$.

The modes are 1 and 4 since they are the two most frequent events.

3.

Comparison Plots

Comparison plots include charts that are ideal for comparing multiple variables or variables over time. Line charts are great for visualizing variables over time. For comparison among items, bar charts (also called column charts) are the best way to go. For a certain time period (say, fewer than 10-time points), vertical bar charts can be used as well. Radar charts or spider plots are great for visualizing multiple variables for multiple groups.

Line Chart

Line charts are used to display quantitative values over a continuous time period and show information as a series. A line chart is ideal for a time series that is connected by straight-line segments.

The value being measured is placed on the y-axis, while the x-axis is the timescale.

Uses

- Line charts are great for comparing multiple variables and visualizing trends for both single as well as multiple variables, especially if your dataset has many time periods (more than 10).
- For smaller time periods, vertical bar charts might be the better choice.

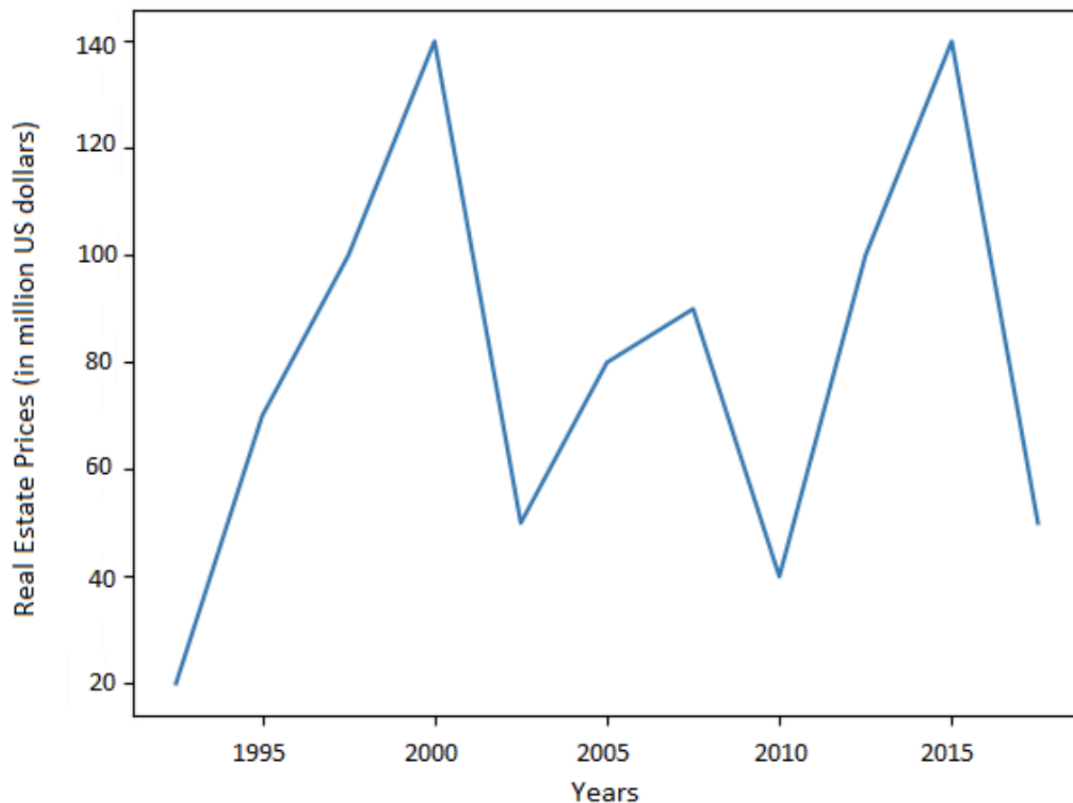


Figure 2.1: Line chart for a single variable

Design Practices

- Avoid too many lines per chart.
- Adjust your scale so that the trend is clearly visible.

Bar Chart

In a bar chart, the bar length encodes the value. There are two variants of bar charts: vertical bar charts and horizontal bar charts.

Use

While they are both used to compare numerical values across categories, vertical bar charts are sometimes used to show a single variable over time.

Don'ts of Bar Charts

- Don't confuse vertical bar charts with histograms. Bar charts compare different variables or categories, while histograms show the distribution for a single variable. Histograms will be discussed later in this chapter.
- Another common mistake is to use bar charts to show central tendencies among groups or categories. Use box plots or violin plots to show statistical measures or distributions in these cases.

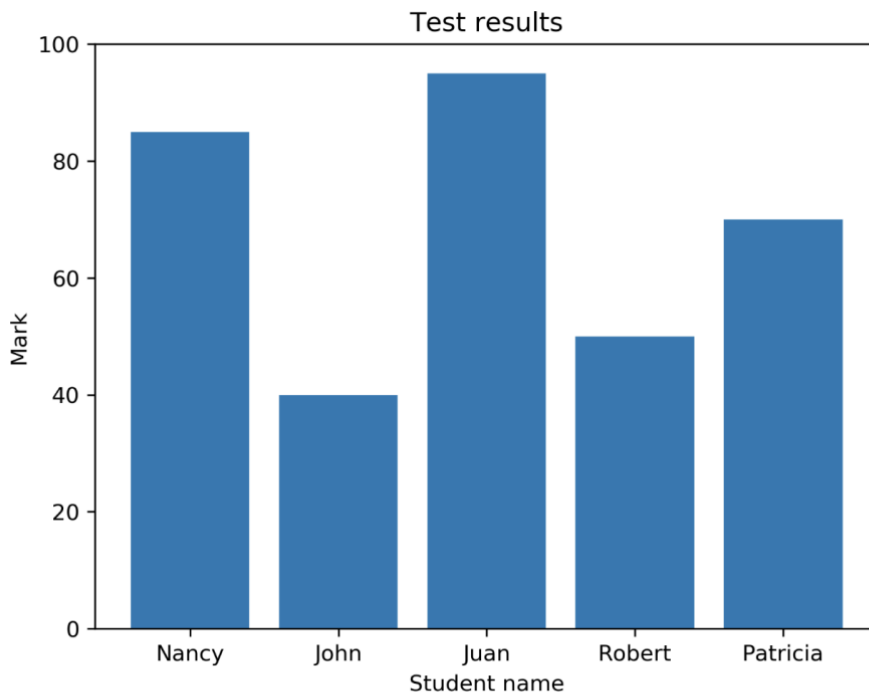


Figure 2.3: Vertical bar chart using student test data

Design Practices

- The axis corresponding to the numerical variable should start at zero. Starting with another value might be misleading, as it makes a small value difference look like a big one.
- Use horizontal labels—that is, as long as the number of bars is small, and the chart doesn't look too cluttered.
- The labels can be rotated to different angles if there isn't enough space to present them horizontally. You can see this on the labels of the x-axis of the preceding diagram.

Radar Chart

Radar charts (also known as **spider** or **web charts**) visualize multiple variables with each variable plotted on its own axis, resulting in a polygon. All axes are arranged radially, starting at the center with equal distances between one another, and have the same scale.

Uses

- Radar charts are great for comparing multiple quantitative variables for a single group or multiple groups.
- They are also useful for showing which variables score high or low within a dataset, making them ideal for visualizing performance.

Examples

The following diagram shows a radar chart for a single variable. This chart displays data about a student scoring marks in different subjects:

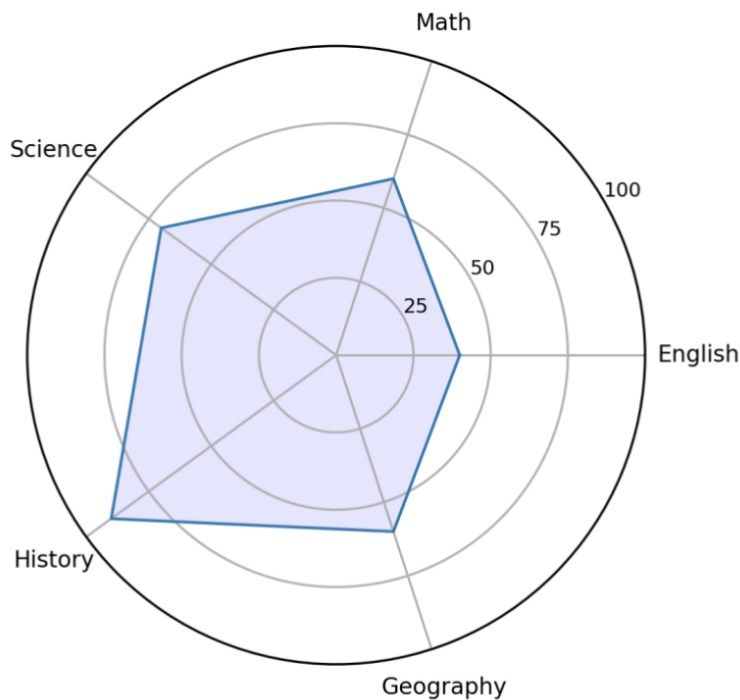


Figure 2.6: Radar chart for one variable (student)

Design Practices

- Try to display 10 factors or fewer on a single radar chart to make it easier to read.
- Use **faceting** (displaying each variable in a separate plot) for multiple variables/groups, as shown in the preceding diagram, in order to maintain clarity.

4.

a.

Dot Map

In a **dot map**, each dot represents a certain number of observations. Each dot has the same size and value (the number of observations each dot represents). The dots are not meant to be counted; they are only intended to give an impression of magnitude. The size and value are important factors for the effectiveness and impression of the visualization. You can use different colors or symbols for the dots to show multiple categories or groups.

Use

To visualize geospatial data.

Example

The following diagram shows a dot map where each dot represents a certain amount of bus stops throughout the world:



Figure 2.39: Dot map showing bus stops worldwide

Design Practices

- Do not show too many locations. You should still be able to see the map to get a feel for the actual location.
- Choose a dot size and value so that in dense areas, the dots start to blend. The dot map should give a good impression of the underlying spatial distribution.

b.

Indexing

Indexing elements in a NumPy array, at a high level, works the same as with built-in Python lists. Therefore, we can index elements in multi-dimensional matrices:

```
dataset[0]      # index single element in outermost dimension
dataset[-1]    # index in reversed order in outermost dimension
dataset[1, 1]  # index single element in two-dimensional data
dataset[-1, -1] # index in reversed order in two-dimensional data
```

Splitting

Splitting data can be helpful in many situations, from plotting only half of your time-series data to separating test and training data for machine learning algorithms.

There are two ways of splitting your data, horizontally and vertically. Horizontal splitting can be done with the `hsplit` method. Vertical splitting can be done with the `vsplit` method:

```
np.hsplit(dataset, (3)) # split horizontally in 3 equal lists
np.vsplit(dataset, (2)) # split vertically in 2 equal lists
```

5. Solution:

1. **Suggested response:** Most trains arrive at 4 p.m. and 6 p.m.
2. **Suggested response:** The histogram appears as follows:

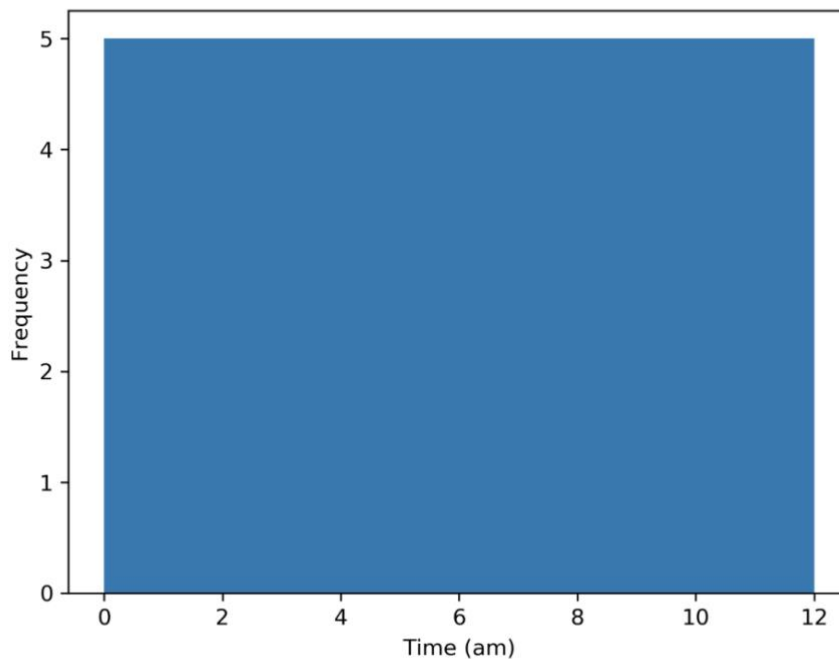


Figure 2.47: Frequency of trains in the morning

6. Composition Plots

Composition plots are ideal if you think about something as a part of a whole. For static data, you can use pie charts, stacked bar charts, or Venn diagrams. **Pie charts** or **donut charts** help show proportions and percentages for groups. If you need an additional dimension, stacked bar charts are great. Venn diagrams are the best way to visualize overlapping groups, where each group is represented by a circle. For data that changes over time, you can use either stacked bar charts or stacked area charts.

Pie Chart

Pie charts illustrate numerical proportions by dividing a circle into slices. Each arc length represents a proportion of a category. The full circle equates to 100%. For humans, it is easier to compare bars than arc lengths; therefore, it is recommended to use bar charts or stacked bar charts the majority of the time.

Use

To compare items that are part of a whole.

Examples

The following diagram shows household water usage around the world:

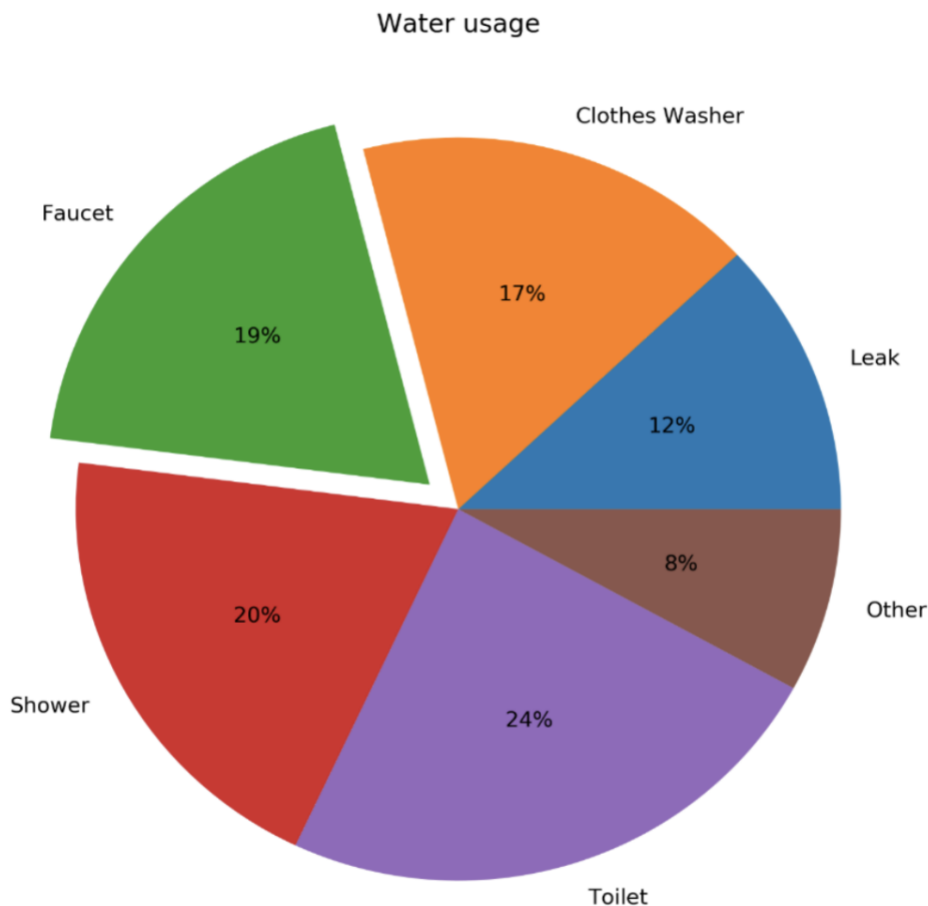


Figure 2.21: Pie chart for global household water usage

Design Practices

- Arrange the slices according to their size in increasing/decreasing order, either in a clockwise or counterclockwise manner.
- Make sure that every slice has a different color.

Variants: Donut Chart

An alternative to a pie chart is a **donut chart**. In contrast to pie charts, it is easier to compare the size of slices, since the reader focuses more on reading the length of the arcs instead of the area. Donut charts are also more space-efficient because the center is cut out, so it can be used to display information or further divide groups into subgroups.

The following diagram shows a basic donut chart:

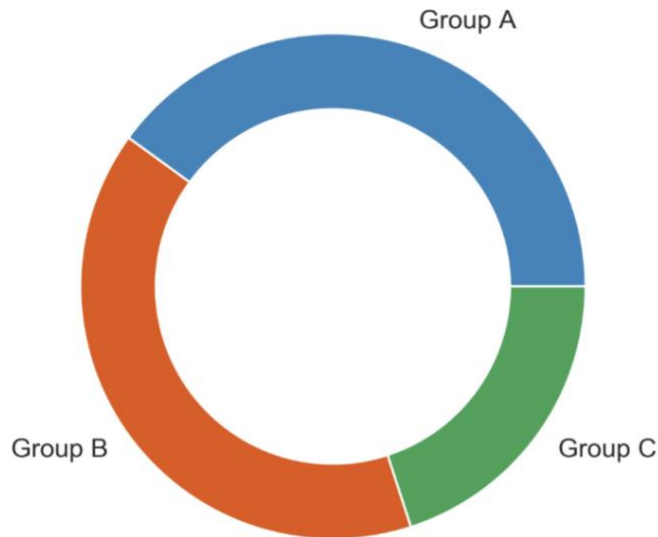


Figure 2.22: Donut chart

Design Practice

- Use the same color that's used for the category for the subcategories. Use varying brightness levels for the different subcategories.