

Data Analytics

Mrinal Das

Quiz-3: Part A

Raw Data

Temperature

Speed

Defect

80

120

No

?

115

No

95

140

Yes

95

140

Yes

How will you preprocess this dataset ?

Answer

Preprocessing Steps

1. Replace missing temperature values.
2. Remove duplicate rows.
3. Normalize speed values.
4. Encode defect labels:

Defect

No = 0

Yes = 1

Processed Data:

Temperature	Speed	Defect
80	0.25	0
85	0.18	0
95	0.85	1

Feature Selection

Food	Calories	Sugar (g)	Fat (g)	Protein (g)	Fiber (g)	Vitamin C (mg)	Package Color	Barcode Number	Manufacturer ID	Category
Apple	95	19	0.3	0.5	4.4	8	Red	345678	M101	Healthy
Banana	105	14	0.4	1.3	3.1	10	Yellow	456781	M102	Healthy
Pizza	285	4	10	12	2	1	Brown	673421	M205	Moderate
Burger	354	8	17	20	1	0	Brown	876512	M310	Unhealthy
Chocolate Cake	410	38	18	5	1	0	Brown	975643	M412	Unhealthy

Evaluate Features

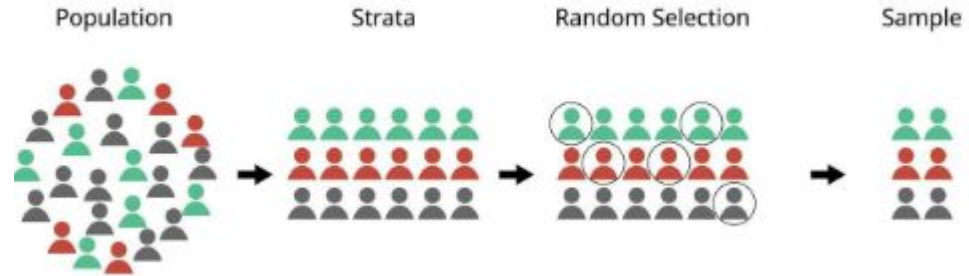
Feature	Keep?	Reason
Calories	✓	High calories often indicate energy-dense foods.
Sugar	✓	Helps distinguish sugary foods.
Fat	✓	Important nutritional feature.
Protein	✓	Indicates nutritional quality.
Fiber	✓	High-fiber foods are often healthier.
Vitamin C	✓	Useful nutritional indicator.
Package Color	✗	Packaging color does not determine nutritional value.
Barcode Number	✗	Unique identifier only.
Manufacturer ID	✗	Does not describe the food itself.

Sampling Methods

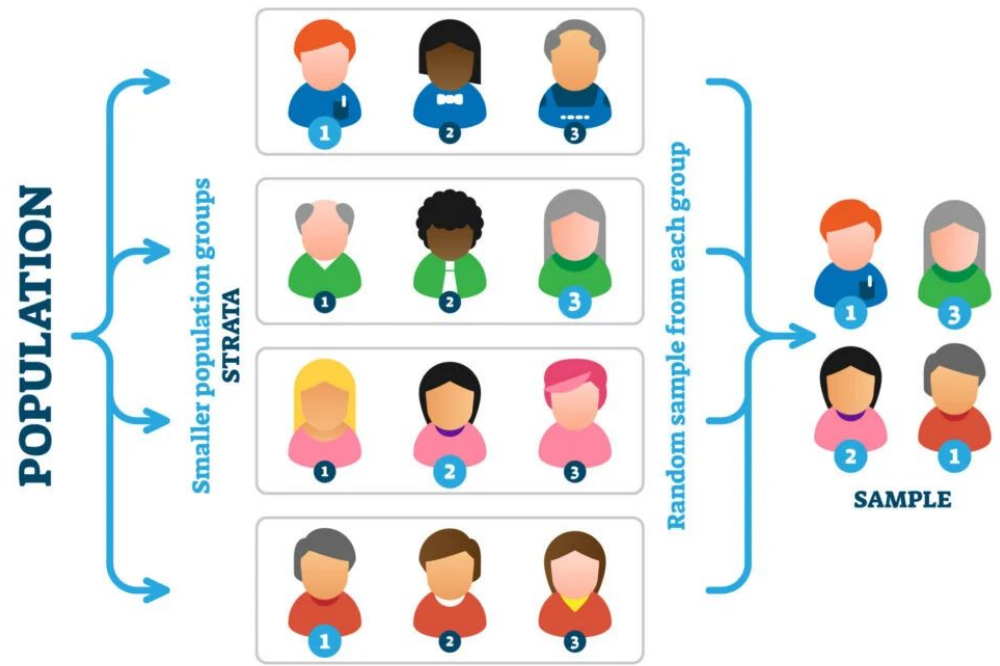
Random Sampling

Stratified Sampling

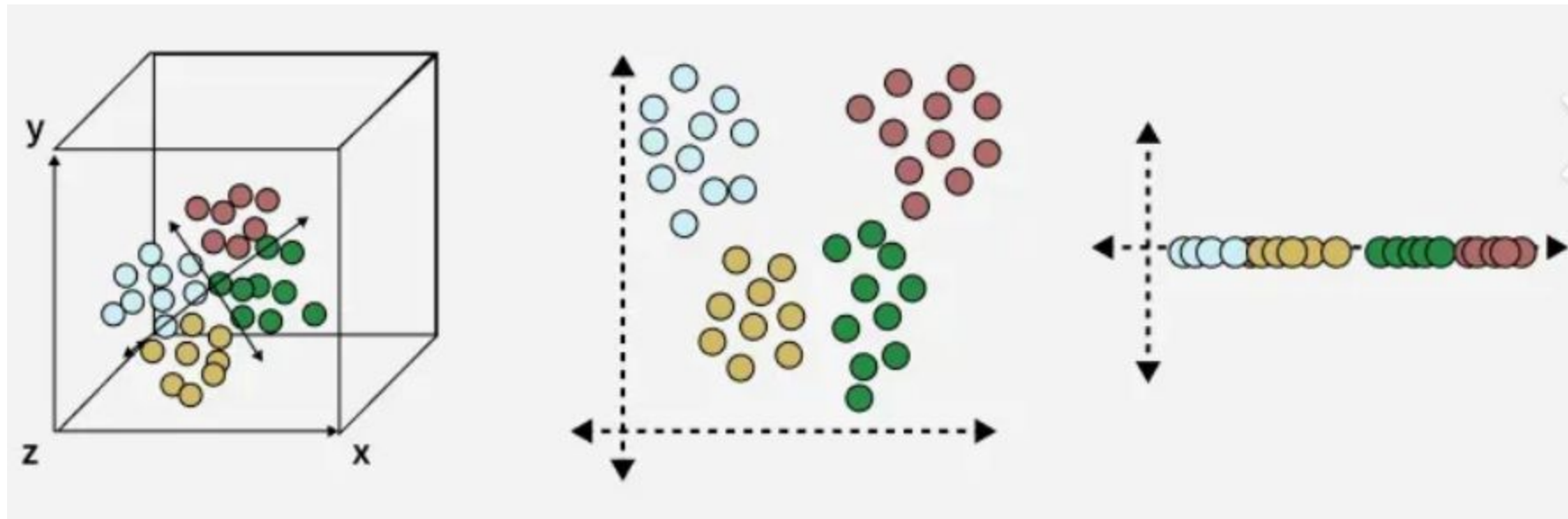
Reservoir Sampling



STRATIFIED SAMPLING



Reduce Dimension



What is Visualization

Data visualization is the graphical representation of data to reveal patterns, trends, relationships, and insights.

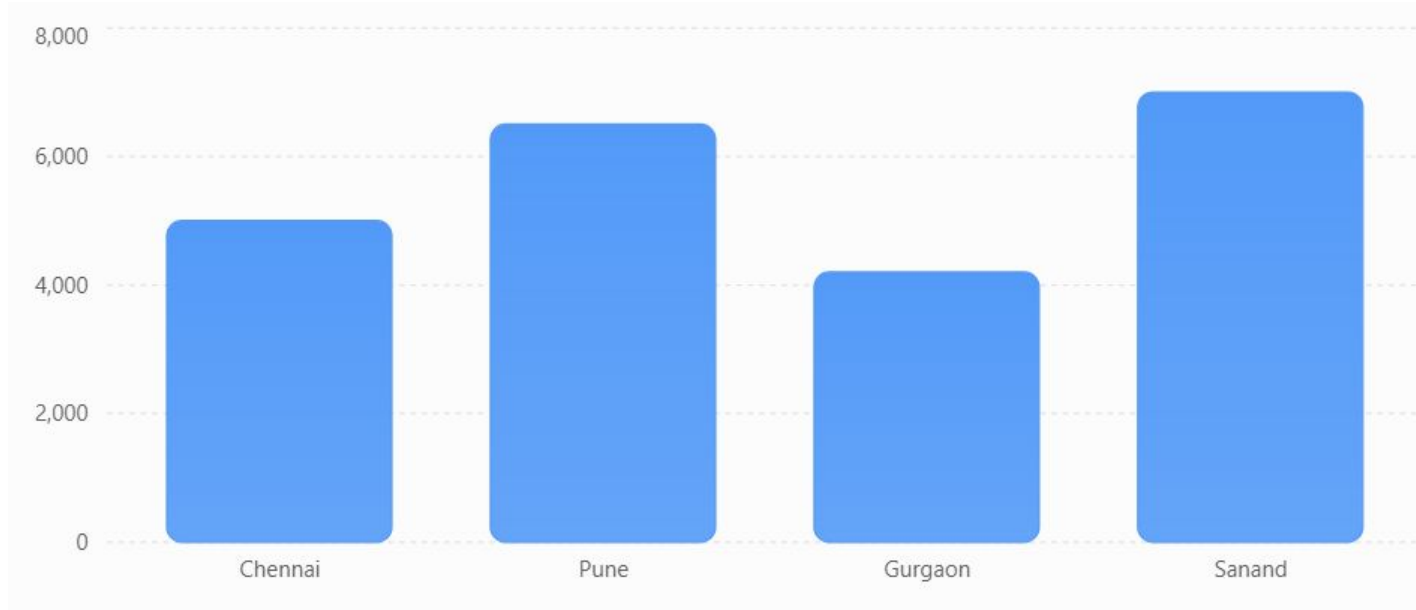
Example

Lamborghini India has following data

Raw Data	
Plant	Cars Produced
Chennai	5000
Pune	6500
Gurgaon	4200
Sanand	7000

Which plant produces the most ?

Example



Visual presentation is easier to comprehend.

Principles of Effective Visualization

Principle 1: Know your audience

Managers - Dashboard

Engineers - Detailed plots

Customers - Simple charts

Principle 3: Highlight

What is the point

Principle 2: Simplicity

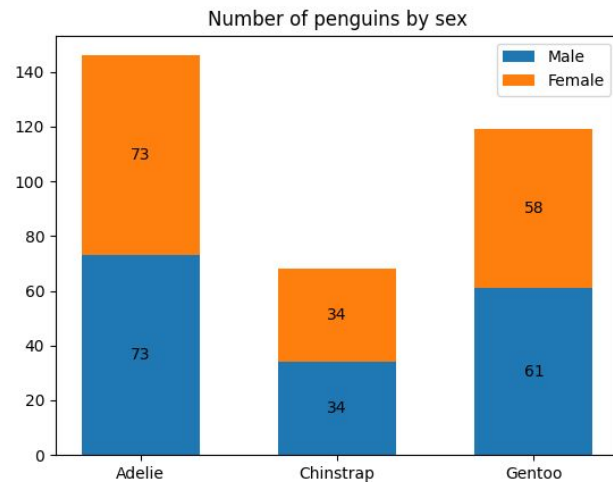
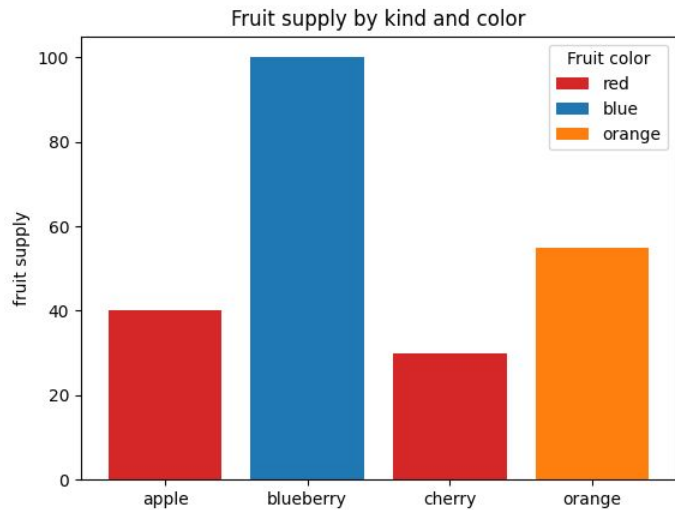
Simple, clear, focused

Too many colors, too many charts, too many labels

Principle 4: Accuracy

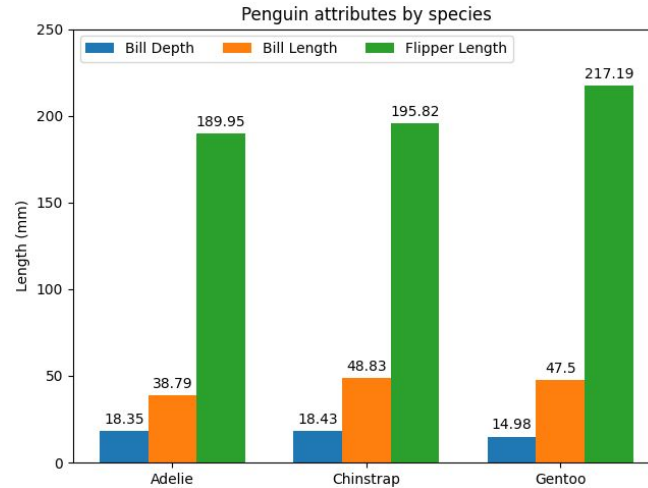
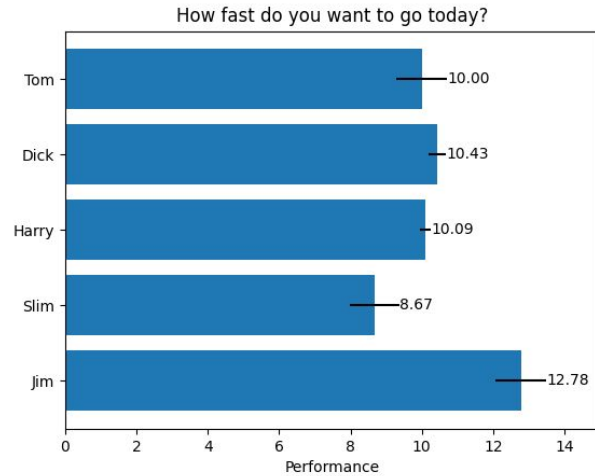
Precise scales, numbers

Bars

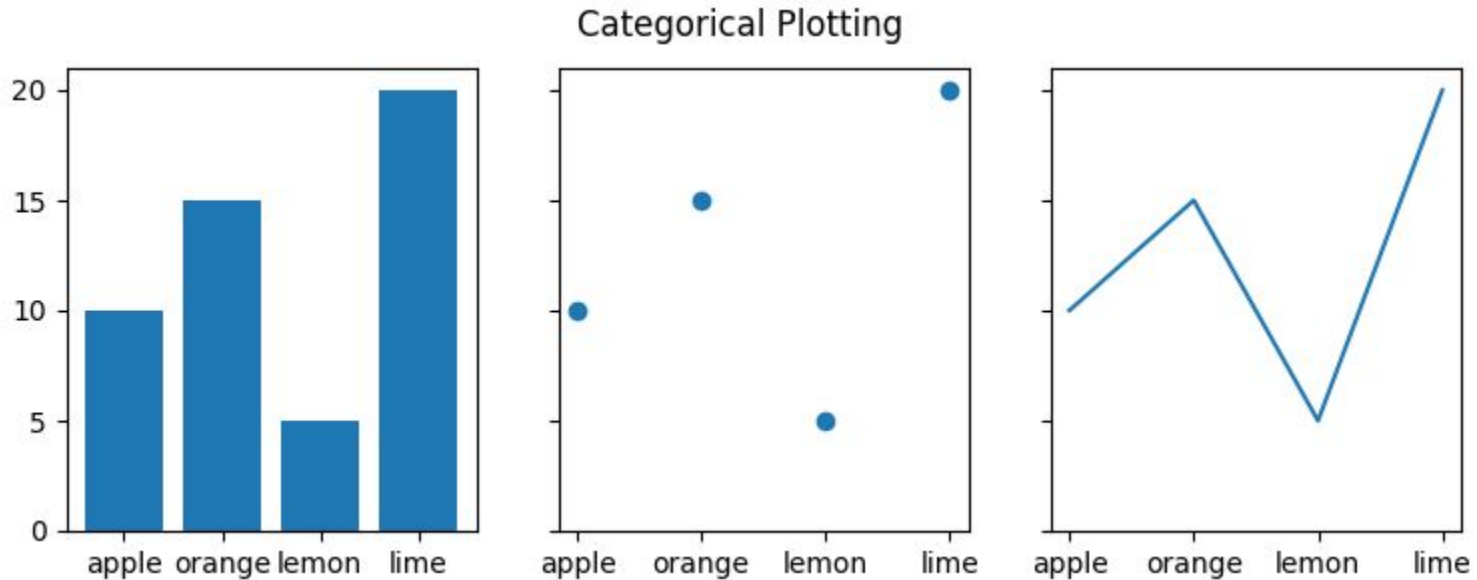


Categories are discrete and objective is to compare

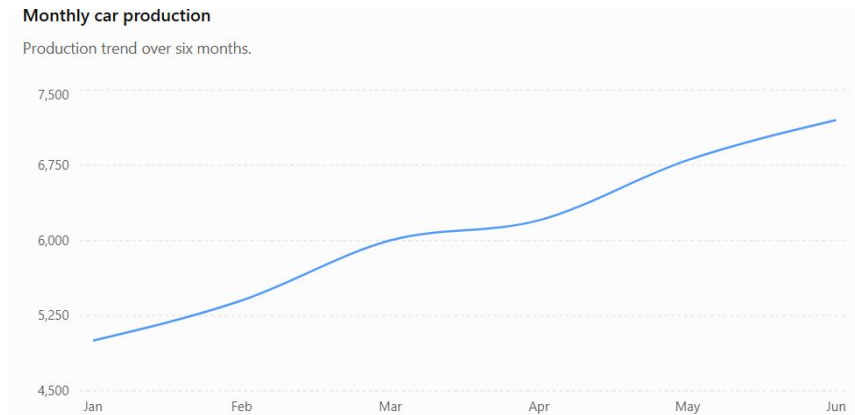
Bars with Numbers



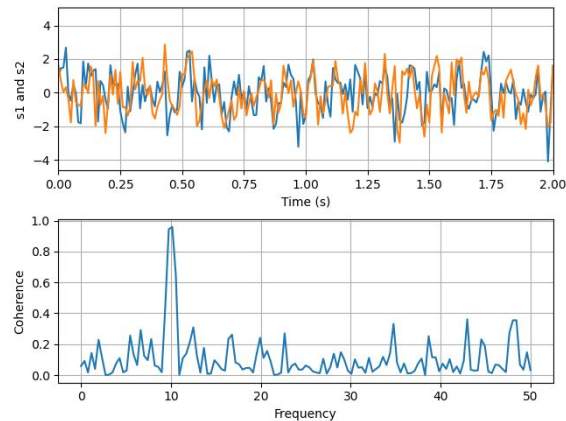
Different ways of categorical plotting



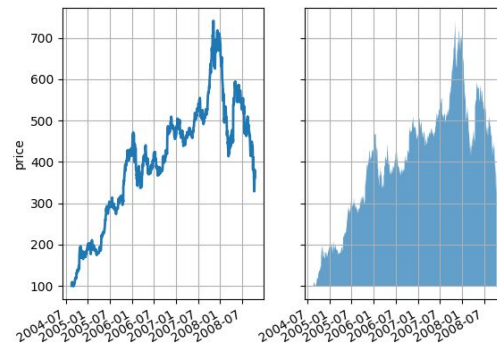
Line Curves



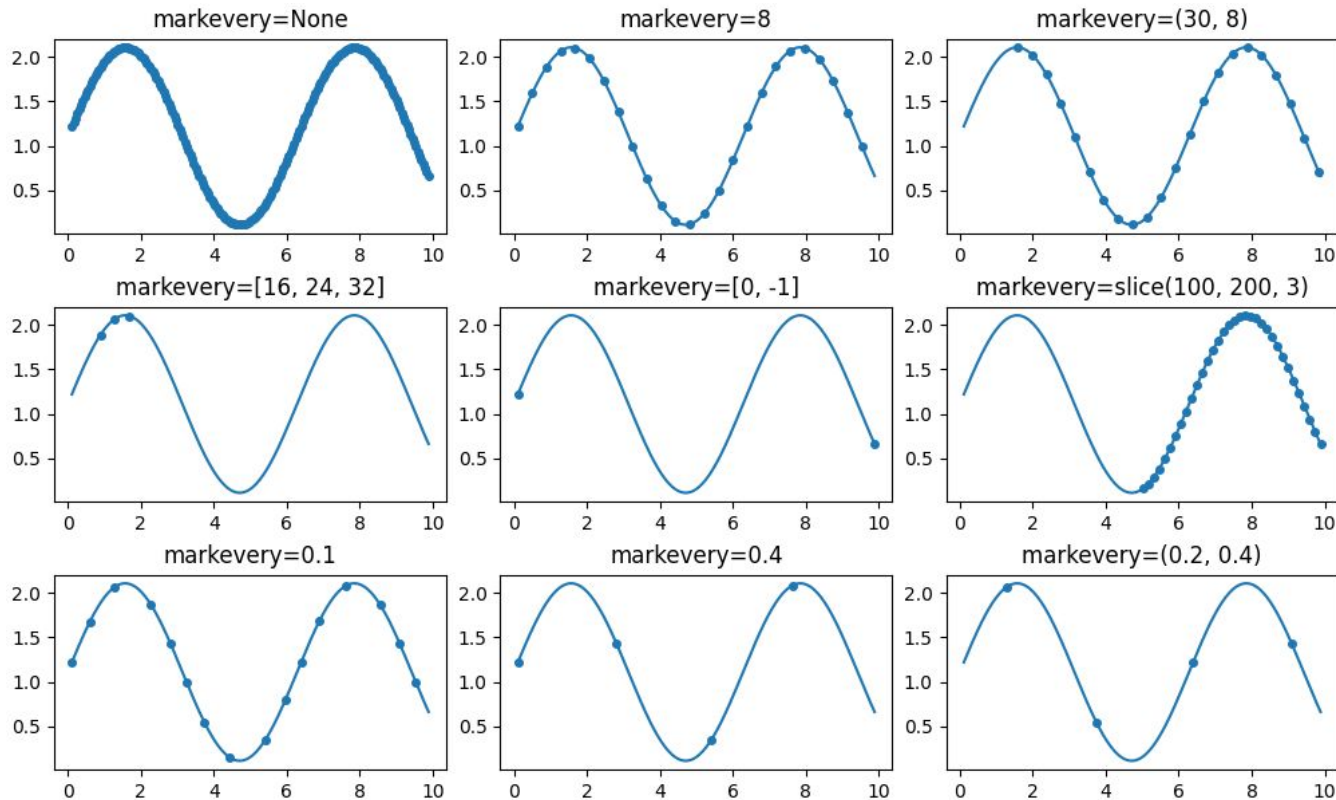
Show trends over time



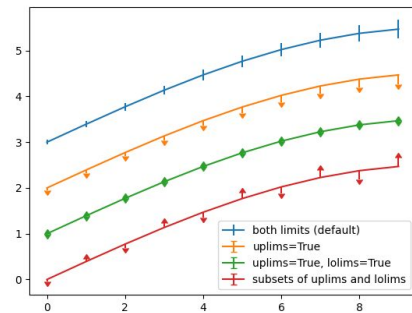
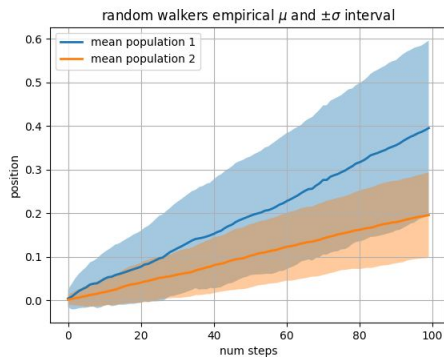
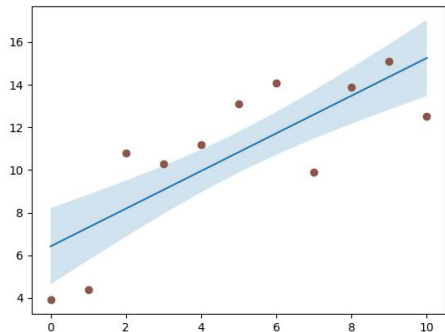
Google (GOOG) daily closing price



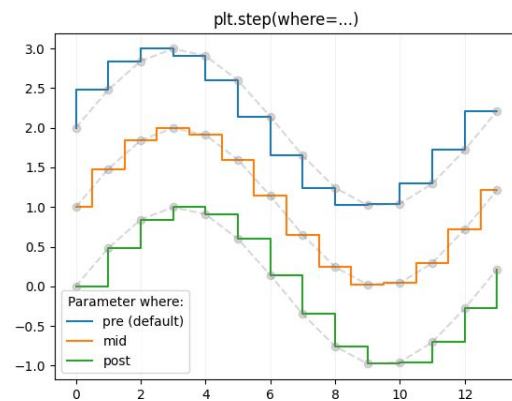
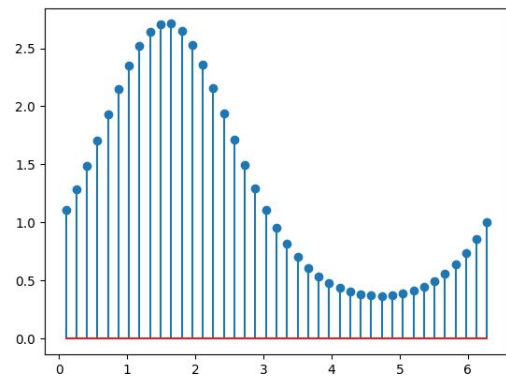
Markers to Show Specific Points



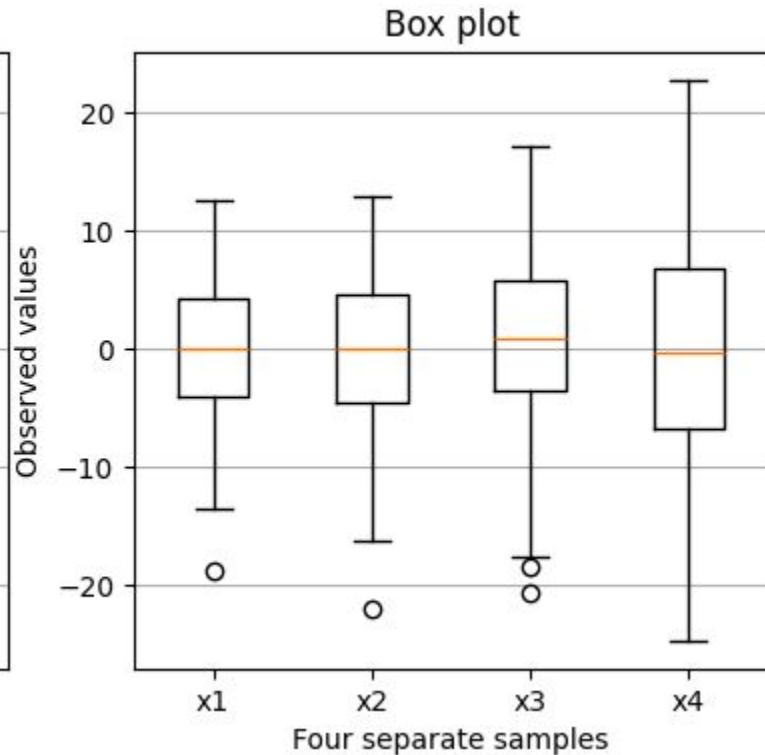
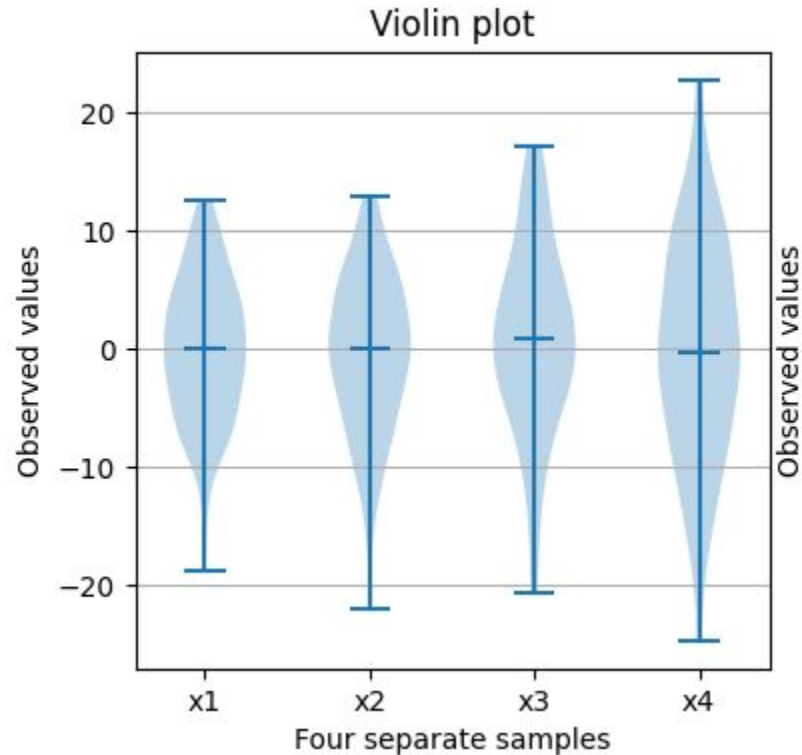
Line Plot with Interval



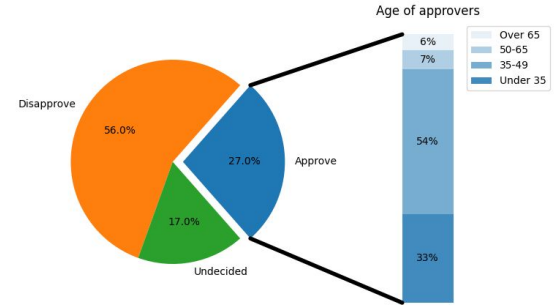
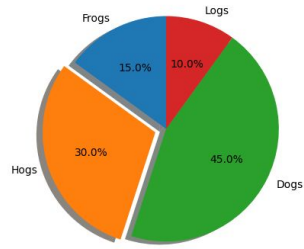
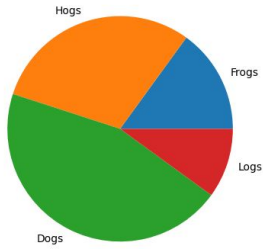
Can show confidence or uncertainty



Statistical Properties of Data



Pie Chart



Show proportions

Quiz time: Car Manufacturing

A factory manager asks:

1. Which plant produces the most cars?
2. Is production increasing?
3. Which defect type is most common?
4. Does temperature affect defects?
5. Are there unusual machine readings?

Case Study: Car Manufacturing

1. Which plant produces the most cars?
 - Bar chart
2. Is production increasing?
 - Line chart
3. Which defect type is most common?
 - Pie chart
4. Does temperature affect defects?
 - Scatter plot
5. Are there unusual machine readings?
 - Box plot

Quiz time

Month	Cars Produced
Jan	1000
Feb	1200
Mar	1500
Apr	1300
May	1700

Create a presentation

Summary

- Visualization transforms data into insights.
- Choose charts based on the question being asked.
- Keep visualizations simple and accurate.
- Avoid misleading designs.
- Visualization is a communication tool, not just a plotting tool.
- Effective storytelling turns data into actionable decisions.

Storytelling with Data

Data analytics with visualization is not only about showing plots.

It can have larger impact:

You want to establish a point, you can use data visualization tactfully.

It can be about procuring a product, or arguing for a change,

Data visualization as a storytelling can be useful.

Example

Suppose there is a free snack in your canteen.

The canteen serves banana whereas you want cake.

How will you convince the management to replace banana with cake ?

Can we use data analytics for this purpose, and how ?

Strategy

You need to collect some data before you can use that

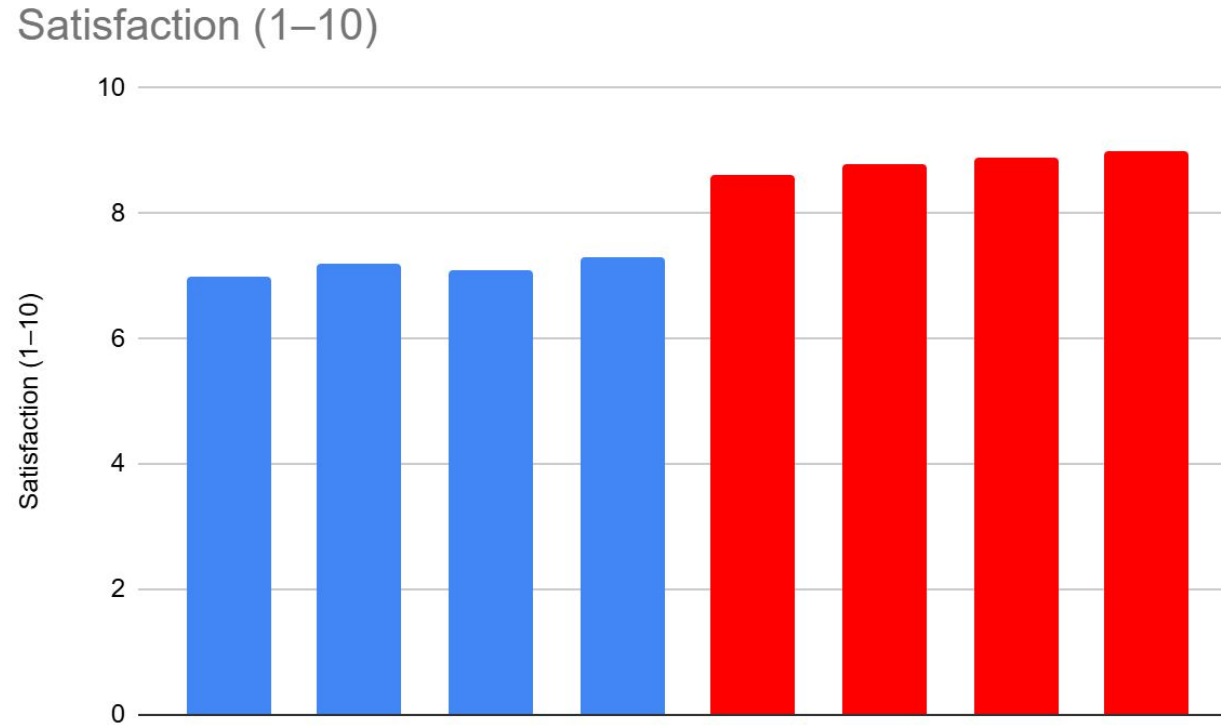
So you do a pilot study

4 weeks for banana and 4 weeks for cake and collect some data

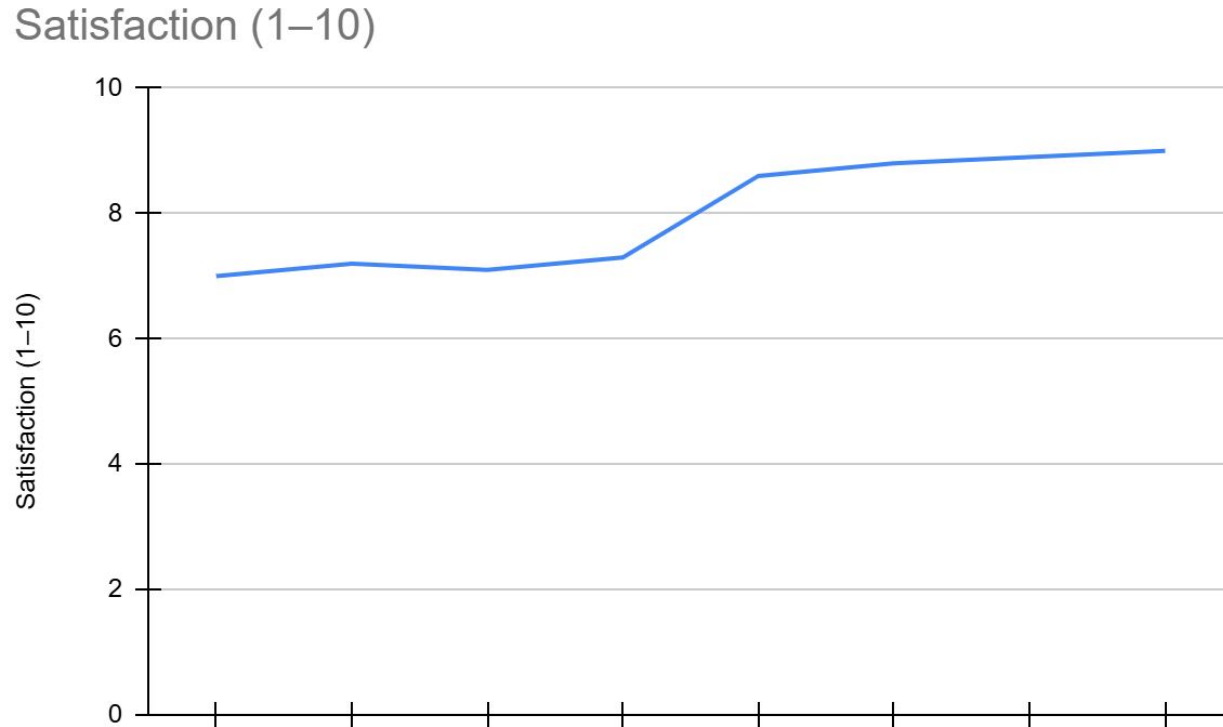
Data Collected

Week	Snack	Satisfaction (1-10)	Avg Break (min)	Cross-Team Conversations	Event Participation (%)	Employees Returning Late	Snack Cost (₹)
1	Banana	7	11	130	32	5	4,500
2	Banana	7.2	12	140	35	4	4,550
3	Banana	7.1	11	135	33	6	4,520
4	Banana	7.3	12	142	36	5	4,530
5	Cake	8.6	17	295	74	7	8,900
6	Cake	8.8	18	315	81	6	9,100
7	Cake	8.9	18	330	84	8	9,050
8	Cake	9	19	340	86	7	9,200

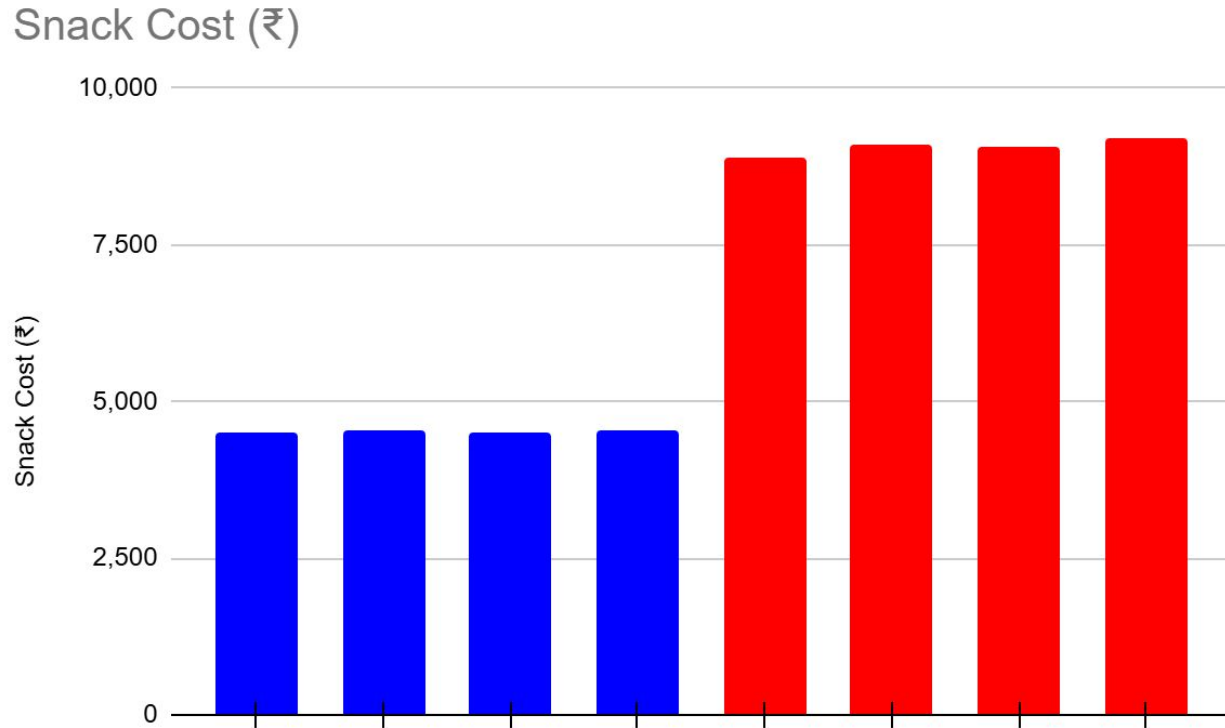
Plot-1: Satisfaction level



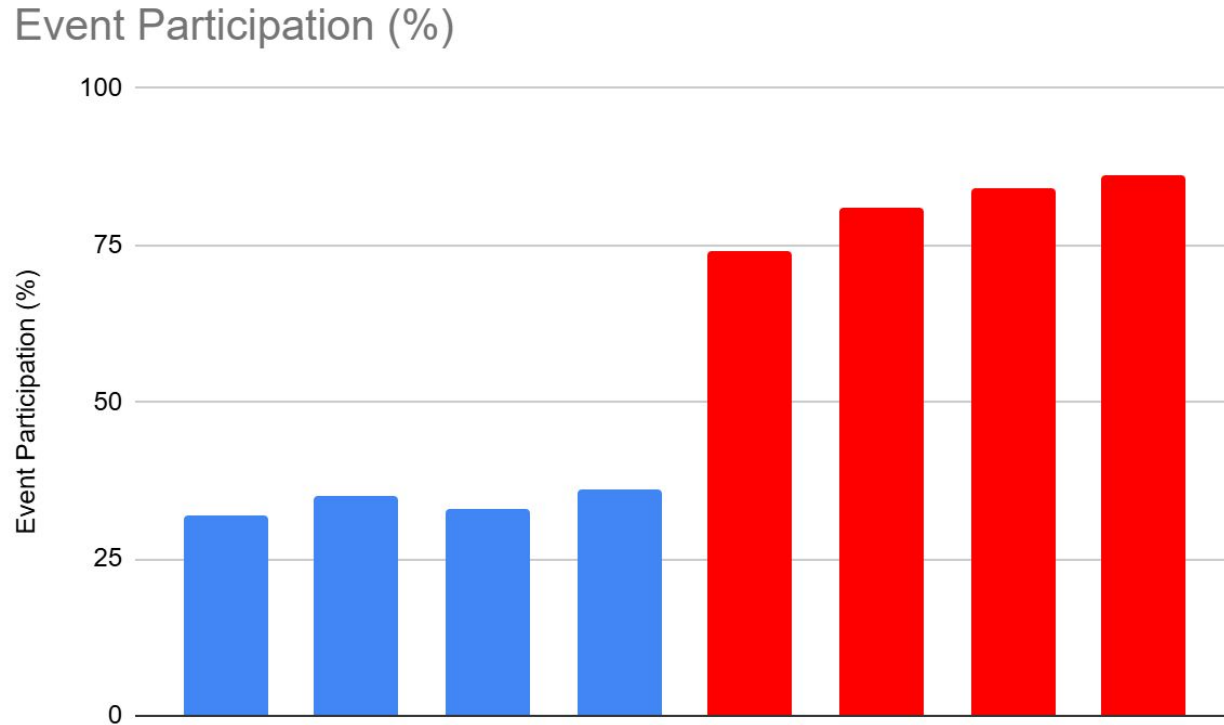
Plot-2: Satisfaction over time



Plot-3: Snack Cost

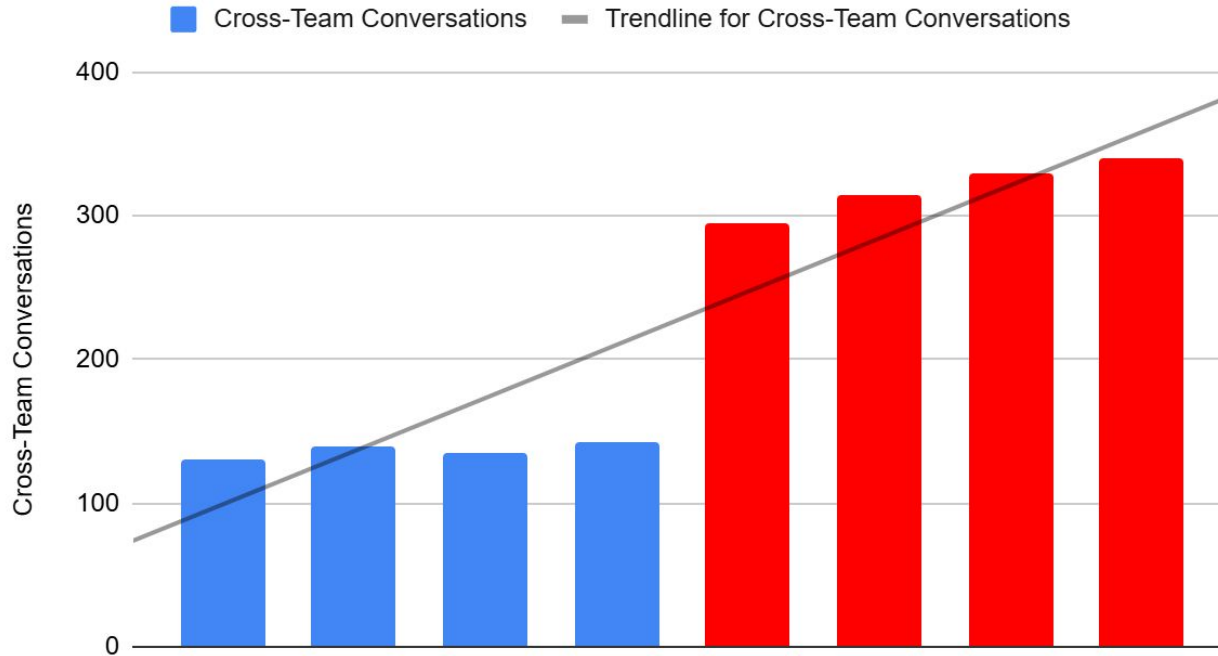


Plot-4: Event Participation



Plot-5: Cross team conversation

Cross-Team Conversations



Data Analytics

Descriptive Analytics

Employee satisfaction has increased.

Cross team conversation has increased.

Event participation has increased.

Prescriptive Analytics

Upon various trials or simulation one can conclude that Cake increases employee happiness but also increases cost. So management can decide to have cake once in a week.

Diagnostic Analytics

Change in the snack is the only factor.

Predictive Analytics

We can go further and fit a ML model to predict satisfaction level.

Room for further exploration

Is the increase in cost due to increase in participation ?

Is increase in satisfaction is due to a change in snack ? Will there be same effect if banana is replaced by apple ?

Can be there be some other alternative that will increase satisfaction but will not increase the cost ?

Is there any important factor that we are missing ?

Quiz-4

1. Write 3 visualization techniques or plots ?
2. In order to show proportions what type of plot will you use ?
3. Write 2 things that you liked/understood today.
4. Write 2 things that you did not like/understand today.
5. Consider the case study you picked in the last session.
 - a. Identify the visualizations that will be useful.
 - b. Create a data story out of the case study.