IPSI

Module 2.2

Describing and communicating data effectively: Numerical data

Trainer:.....

Module objectives

At the end of this module, participants will be able to:

1. Describe numerical data using measures of frequency, central tendency and variation
2. Frame questions that can be answered by available district health systems data
3. Apply appropriate data visualization techniques to answer questions about district health system performance

Module outline

Sr. No.	Session name	Session format	Time
1	Analyzing and visualizing numerical data Data analysis and visualization techniques • Histograms	• Facilitator led presentation and discussion • Exercise	30 min
2	• Box plots		40 min
3	• Line graphs • Scatter plots		40 min

Session 1:
Analysing and visualizing numerical data

Duration: 30 min

Data analysis

As you begin the process of analyzing and visualizing data, it is useful to ask the following questions:

- What data do you have?
- What questions do you want your data to answer?
- **What visualization methods should you use to answer your questions?**
- Did the findings from your analysis answer your question?

Adapted from: Yau, Nathan. Data Points : Visualization That Means Something / Nathan Yau. 1st edition. Indianapolis: John Wiley & Sons, Inc., 2013. Print.

Data Visualization

What visualization methods should you use to answer your questions?

There are many different ways in which data can be visualized, but typically a few types are sufficient:

- Table
- Bar chart
- Pie chart
- Histogram
- Box plot
- Line plot
- Scatterplot

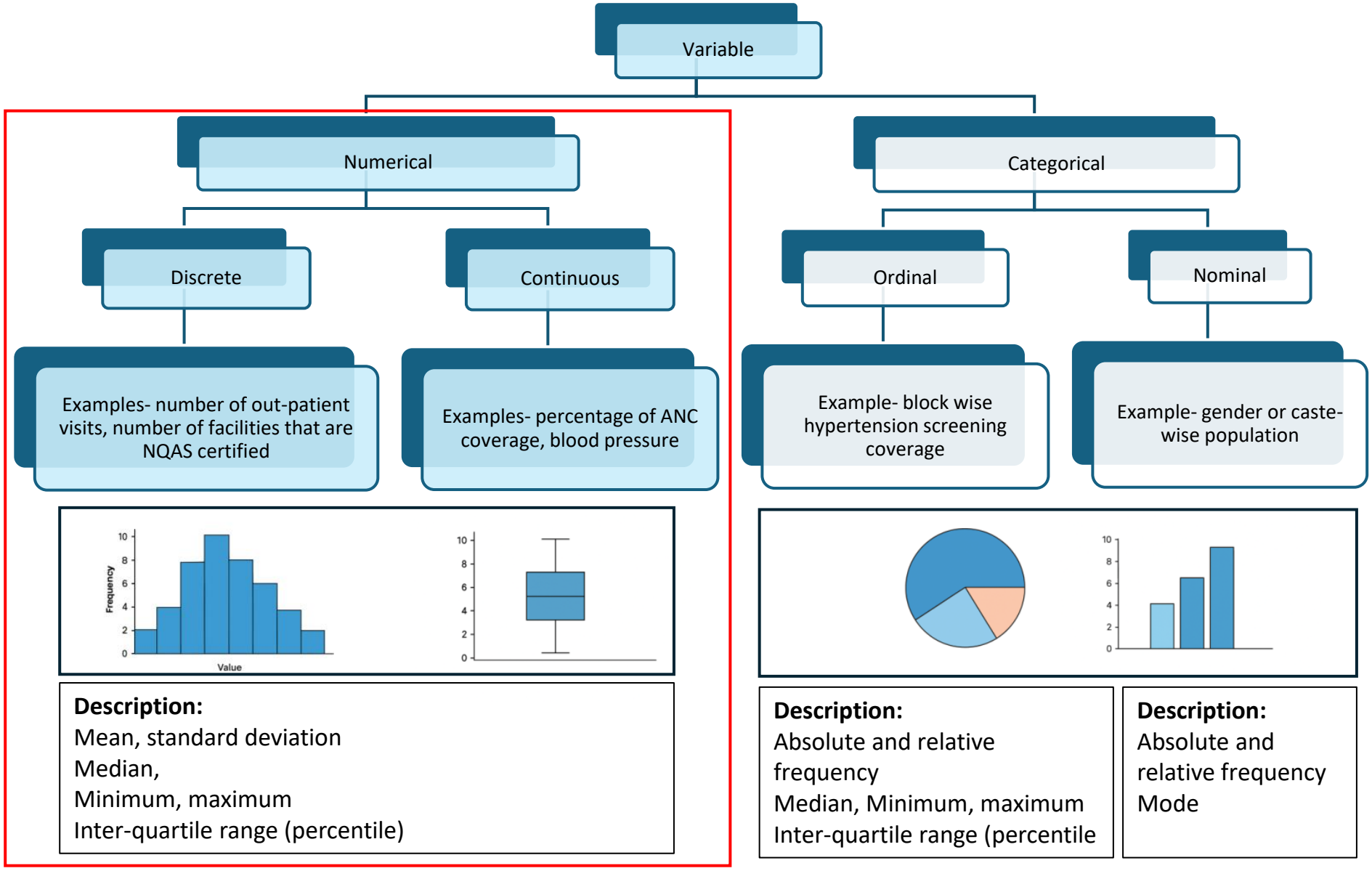
Data visualization

"The greatest value of a picture is when it forces us to notice what we never expected to see."

—John W. Tukey, Exploratory Data Analysis (1977)

- Graphs are a powerful way of visualizing data and can provide information quicker than tables.

Visualizations according the data variable type

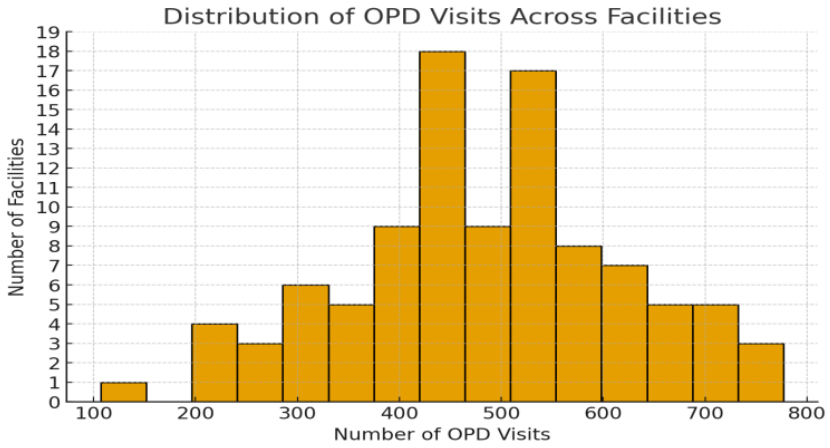


Numerical data:

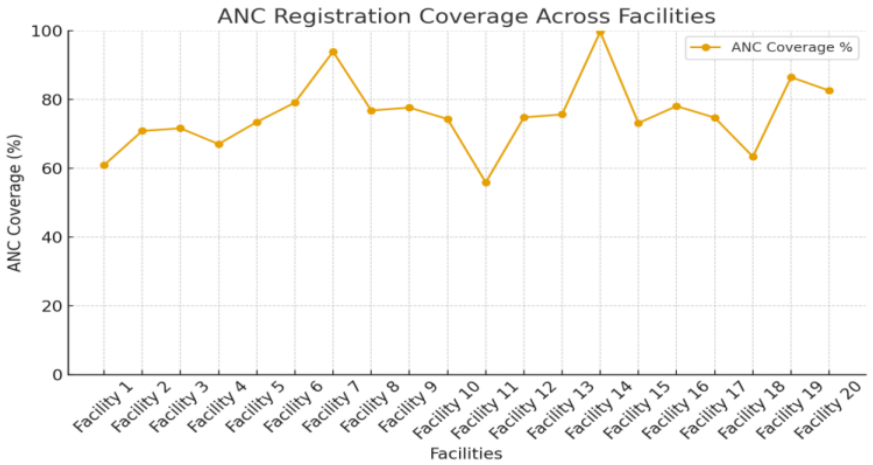
- Histograms
- Box plots
- Line charts

Adapated from- Spriestersbach A, Röhrig B, du Prel JB, Gerhold-Ay A, Blettner M. Descriptive statistics: the specification of statistical measures and their presentation in tables and graphs. Part 7 of a series on evaluation of scientific publications.

Analyzing and visualizing numerical data



Histogram – shows the distribution of OPD visits across facilities.



Line Graph – shows ANC registration coverage (%) across facilities.

- **Much of the data generated in health information systems are numerical and continuous**.
- Examples include the number of OPD visits at health facilities; the percentage of pregnant women registered for ANC in a health facility catchment area.
- Typically, data from health facilities or line-listed information from populations can be large, requiring summarizing the data and specialized methods to visualize it.

Data source: HMIS Portal, X district

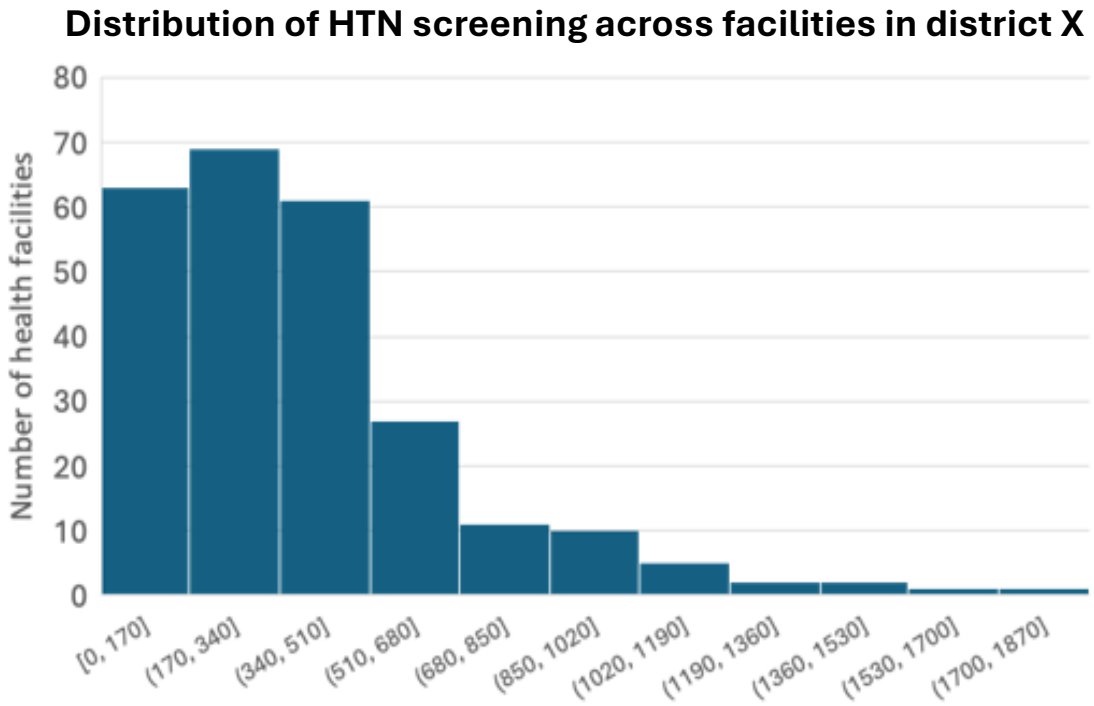
Analyzing Continuous Data

Let's take an example. In this example we have data from 252 health sub-centers from one district in India on the number of adults above 30 years of age in a facility catchment populations screened for hypertension in a 3-month period.
Computing descriptive statistics can provide some insight into the health facility performance on hypertension screening.

Mean: 376.7
Median: 319.5
Standard Deviation: 302.7
Range: 0 – 1776

- On average 377 adults were screened per facility
- There is large variation across facilities in the number screened- the standard deviation is large (302.7 individuals) particularly relative to the mean.
- There are likely some facilities that had zero screenings
- There are some facilities with a large number of screening e.g. 1776, suggestive of outliers.

Histograms



Number of individuals above 30 years of age screened for hypertension in a quarter

Data source: NCD Portal, X district

Often it is helpful to visualize numerical data using graphical methods. Visualizing data also can help in communicating findings in an interesting and easier to understand way.

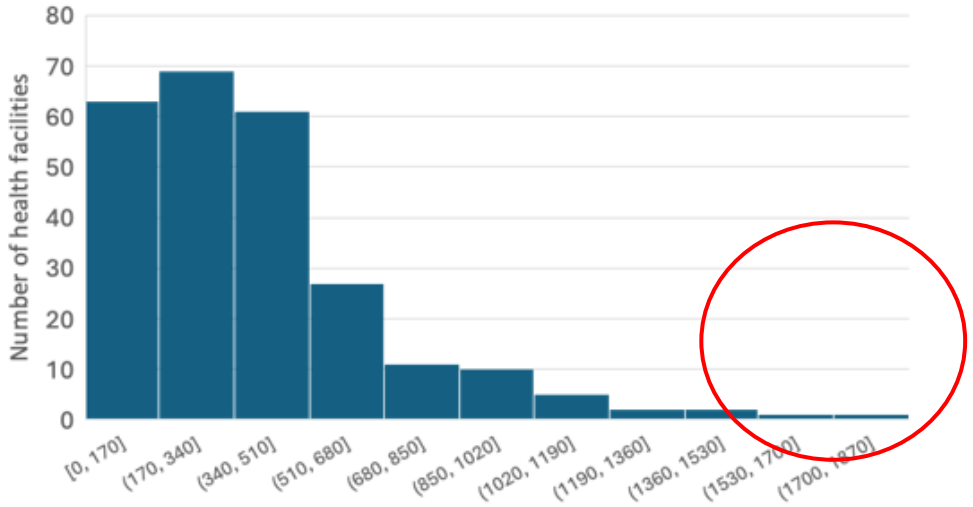
The histogram shown here presents the same facility data on hypertension screenings discussed previously.

There are several questions that this histogram can answer:

- What is the most common number of adults screened?
- Is the data skewed or symmetrically distributed?
- Are there any facilities with extreme values?

Histograms

Distribution of HTN screening across facilities in district X



Number of individuals above 30 years of age screened for hypertension in a quarter

Data source: NCD Portal, X district

There are several questions that this histogram can answer:

- What is the most common number of adults screened?
  - ➢ Screening between 170-340 adults was most common across health facilities.

- Is the data skewed or symmetrically distributed?
  - ➢ The data is highly right skewed i.e. most facilities are at the lower end of the distribution while there are a few at the high end. (the length of each bar can also represent the relative percentage of all facilities falling within a given category)

- Are there any outlier facilities?
  - ➢ Yes. The facilities at the high end of the distribution (circled) are far away from the bulk of the health facilities are likely outliers – they need to be examined more closely.

Exercise 1: Distributions- Constructing a Histogram

- Work in the same groups as before
- Open the Dataset_Module 2.2_Exercise 1
- Refer to the Instruction sheet_Module 2.2_Exercise 1

- Draw the histogram using the data.
- Interpret:
  - Which class interval has the highest number of facilities?
  - Is the data skewed?
    - Right-skew: few high-volume facilities
    - Left-skew: most are high-performing
  - Are there gaps or clusters?

Session 2:
Box Plots

Duration: 40 min

Measure of variation: Inter-quartile range

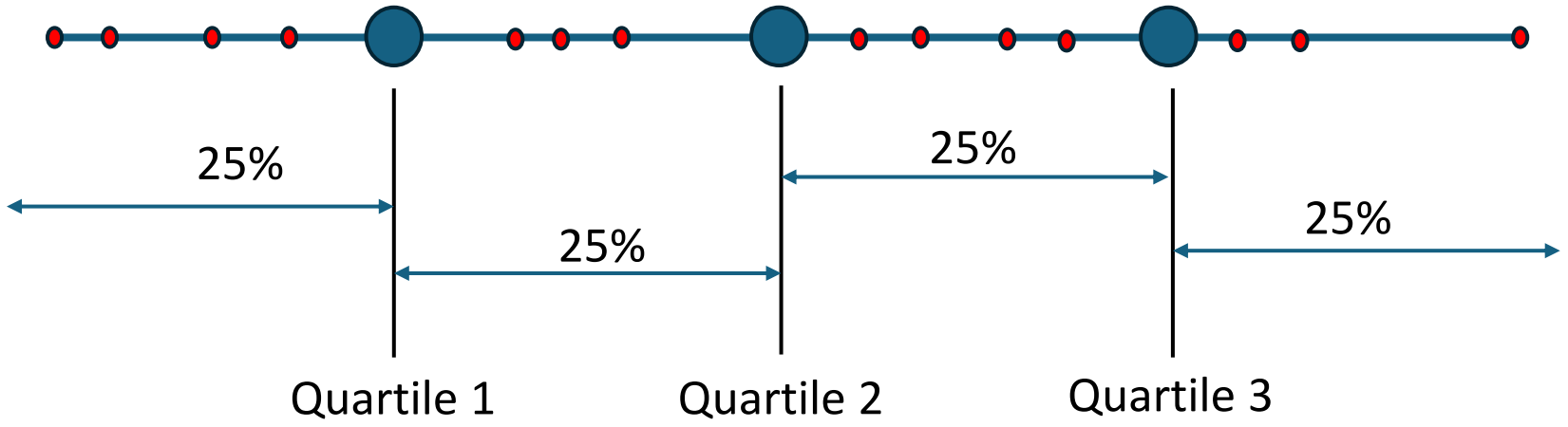
- Descriptive statistics: Summarizing data in an organized manner to enable decision-making

Measure	Example	What does it represent?
Measures of frequency	Frequency, rate, ratio, proportion, percentage	Shows how often a value occurs
Measures of central tendency	Mean, median, mode	Indicates the central/typical value in the distribution of a dataset.
Measures of dispersion/variation	Range, standard deviation, interquartile range	Indicates the spread of values

Source: Kaur, Parampreet1,; Stoltzfus, Jill1,2; Yellapu, Vikas1. Descriptive statistics. International Journal of Academic Medicine 4(1):p 60-63, Jan–Apr 2018. | DOI: 10.4103/IJAM.IJAM_7_18

Measure of variation- Interquartile range (IQR)

Quartiles- Dataset divided into four equal parts, each containing 25% of data.



	Quartile 1 (Q1)	Quartile 2 (Q2)	Quartile 3 (Q3)
Description	25% datapoints lie below this value	Median of the dataset	75% datapoints lie below this value
Also called as	25^{th} percentile	50^{th} percentile	75^{th} percentile
Formula	$\sim \left[\frac{n+1}{4}\right]^{th} term$	$\sim \left[\frac{n+1}{2}\right]^{th} term$	$\sim \left[\frac{3(n+1)}{4}\right]^{th} term$
	5^{th} term	**9^{th} term**	**14^{th} term**

Measures of variation- Interquartile range (IQR)

Interquartile range= (Q3-Q1)

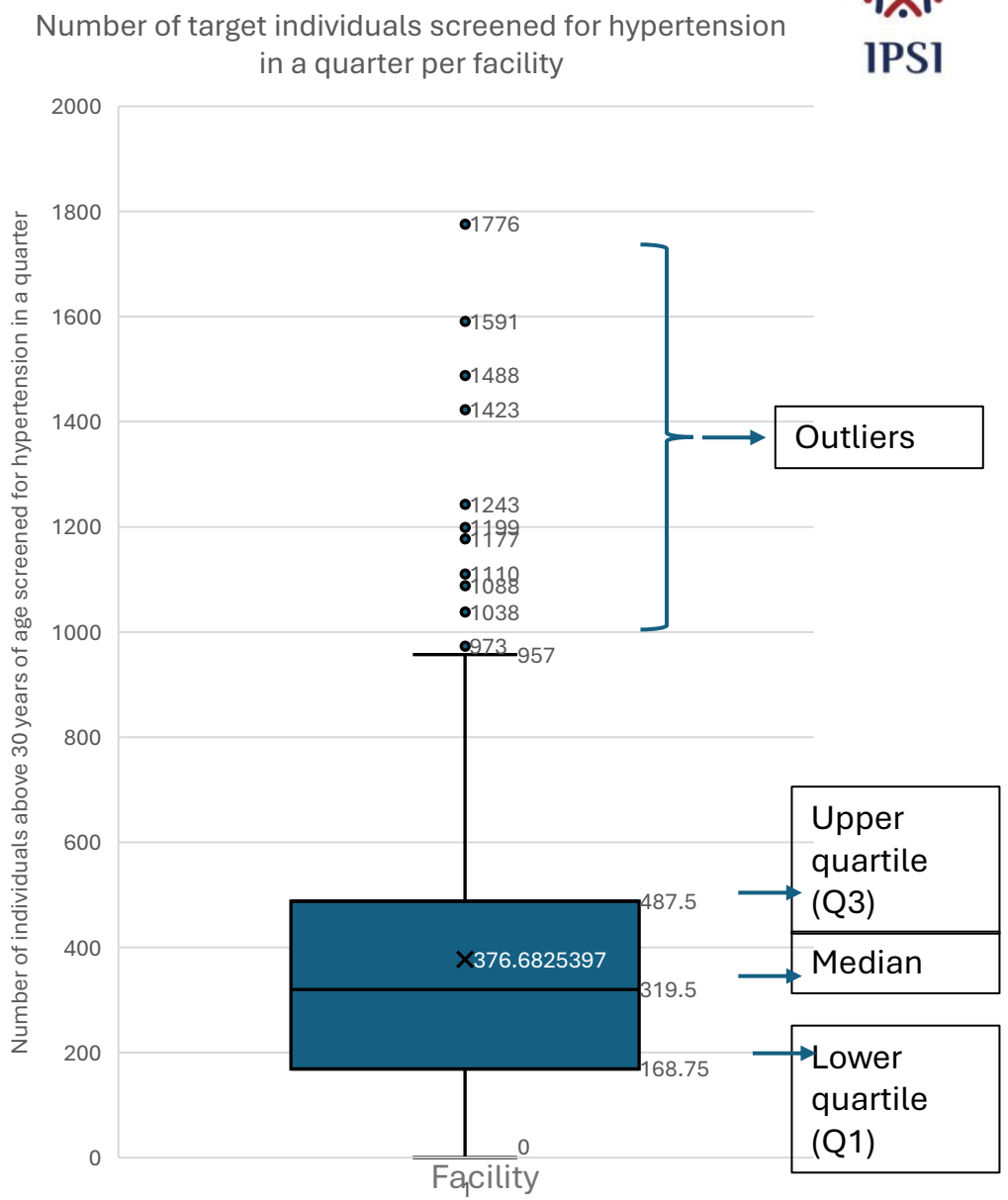
- Measures the statistical spread of the central 50%.
- Higher IQR = more spread/variability
- Lower IQR = more consistency

Identifying Outliers

- Outliers are extreme values in the dataset
- One way to identify outliers is using the IQR
  - ➢ Outliers = Data points beyond 1.5×IQR from Q1 or Q3.

Box plot

- Box Plots are another useful graphical method to analyze numerical continuous data.
- A box plot (box-and-whisker plot) is a graphical summary of a data set.
- It provides insights into
  - ➢ Central value (Median)
  - ➢ Spread (Range, Interquartile Range)
  - ➢ Skewness of data
  - ➢ Outliers



Data source: NCD Portal, X district

Box plot

Understanding the Median (Q2)

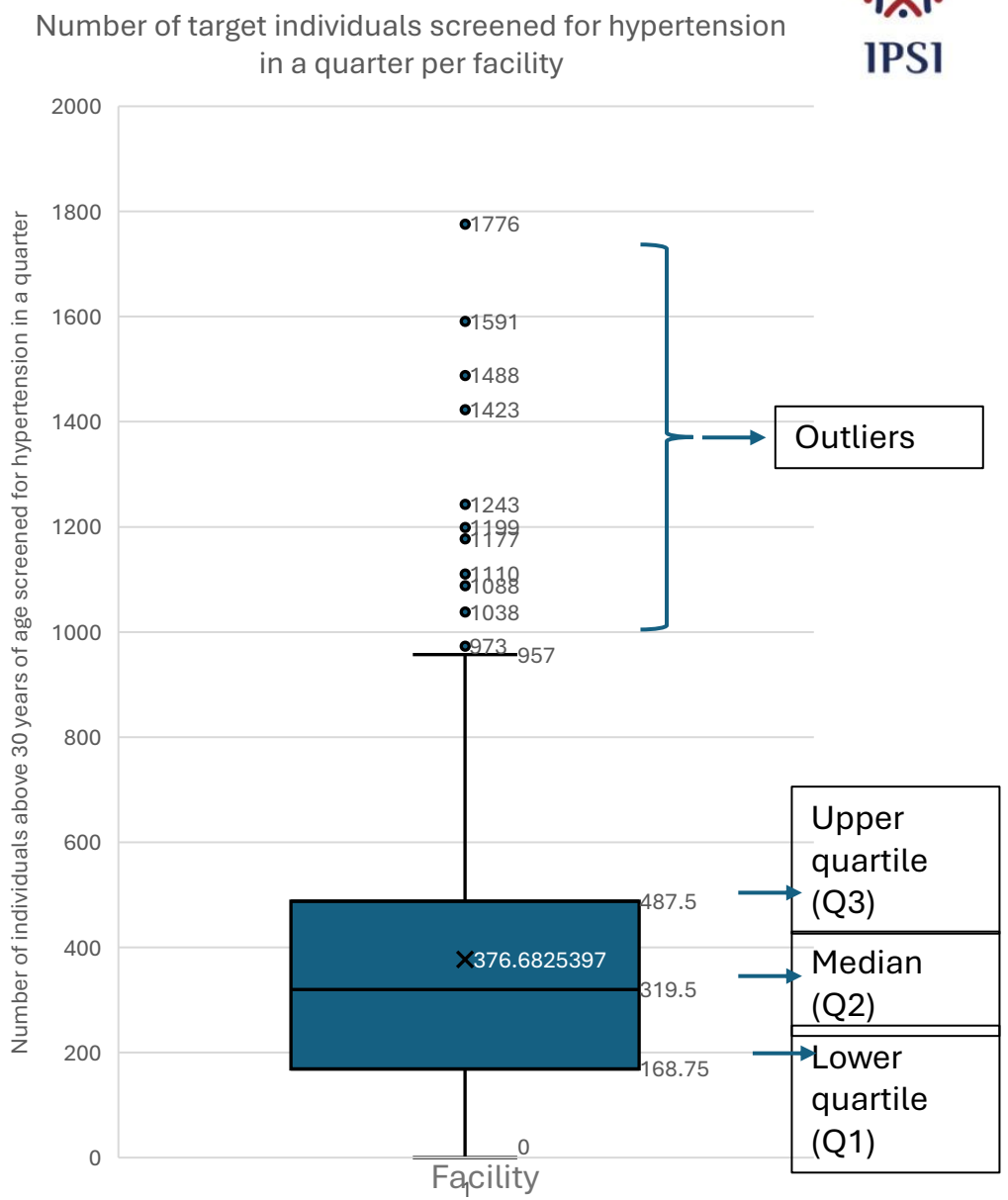
- The median is the line inside the box.
- It divides the data into two equal halves.
- More resistant to outliers than the mean.

Quartiles – Understanding the Spread

- Q1 (25th percentile): 25% of data lies below this value.
- Q3 (75th percentile): 75% of data lies below this value.
- Box represents the interquartile range (IQR = Q3 - Q1).

Importance:

- IQR shows the spread of the middle 50%.
- Helps detect data concentration and variation



Data source: NCD Portal, X district

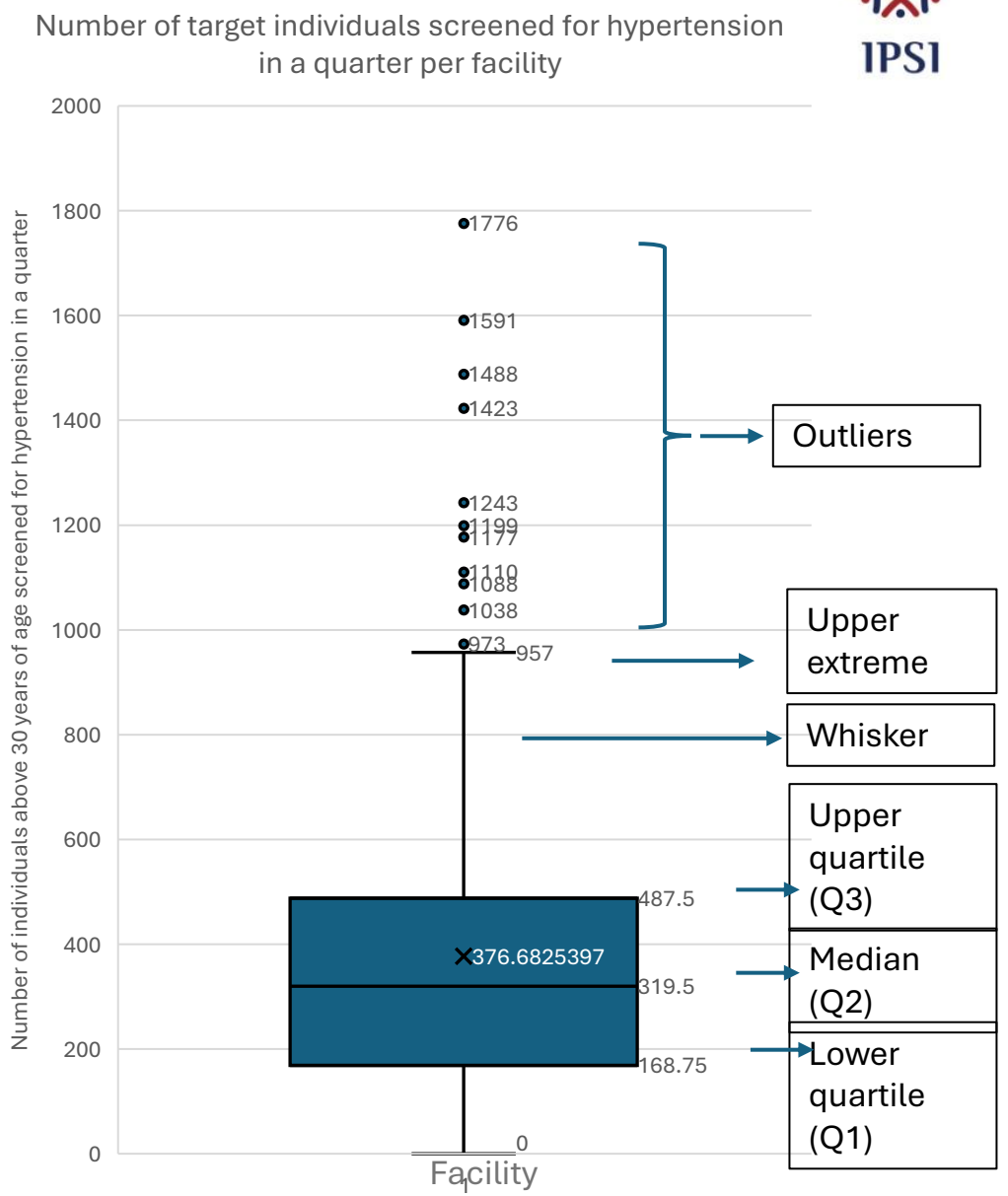
Box plot

IQR – How Variable is the Data?

- IQR = Q3 - Q1
- Measures the statistical spread of the central 50%.
- Higher IQR = more spread/variability
- Lower IQR = more consistency

Whiskers – Range Within Limits

- Whiskers extend from Q1 down to the smallest non-outlier value
- And from Q3 up to the largest non-outlier value
- Points beyond whiskers are outliers



Data source: NCD Portal, X district

Box plot

Identifying Outliers

- Outliers = Data points beyond 1.5×IQR from Q1 or Q3.
- Represented as dots, stars, or asterisks in the plot.

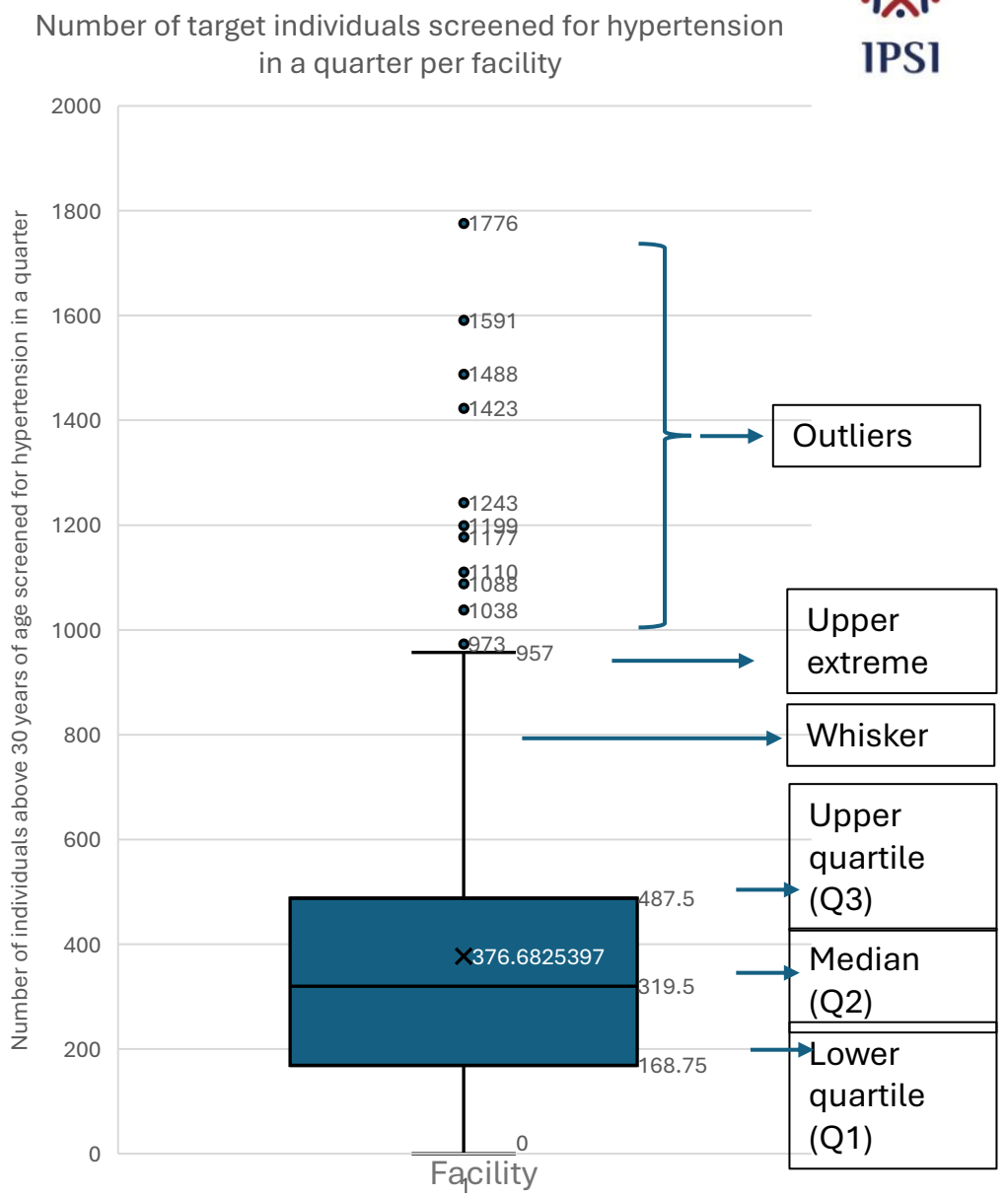
Why important?

- Helps detect unusual blocks (e.g., extremely low ANC uptake)
- Can guide further investigation

Skewness of Data : A box plot's symmetry indicates normality:

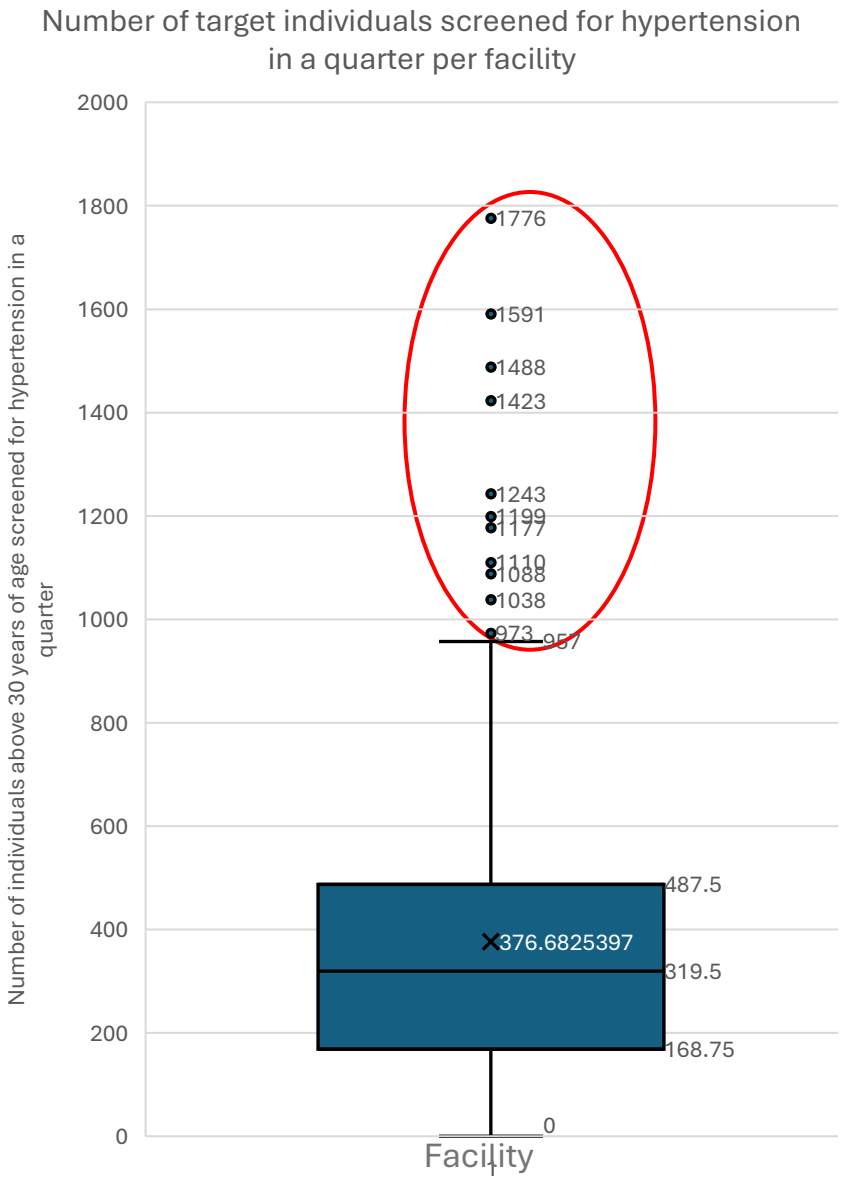
- Symmetric box and tails: Even spread
- Right-skewed: Long tail on the right; mean > median, outlying values are larger than more frequently occurring values
- Left-skewed: Long tail on the left; mean < median outlying values are smaller than most frequently occurring values

In our dataset, Mean = 377, Median = 320 - Slight right skew



Data source: NCD Portal, X district

Box plot

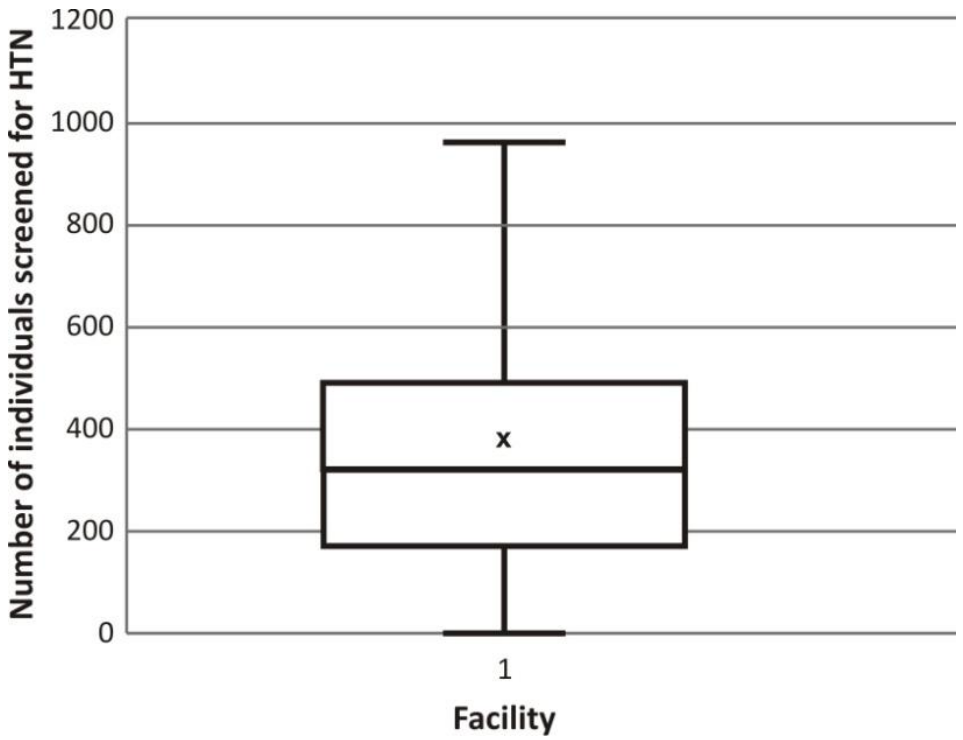


Data source: NCD Portal, X district

Box plots allow us to answer several questions about the data. Let's again look at the example of number of adults screened for hypertension at health facilities in the past three months:

- What is the median and mean level of screenings?
- ➢ The median number of screenings is around 320 and the mean is around 377.
- What is the range of screenings across facilities?
- ➢ There is a wide range in the number of screenings ranging from 0 to 1800 across health facilities.
- Is the distribution of facilities skewed?
- ➢ The distribution of facilities is skewed to lower level of screenings (right skewed)
- Are there outlier facilities?
- ➢ There are several outlier observations i.e. the dots above the upper extreme whisker circled in red.

Box plot



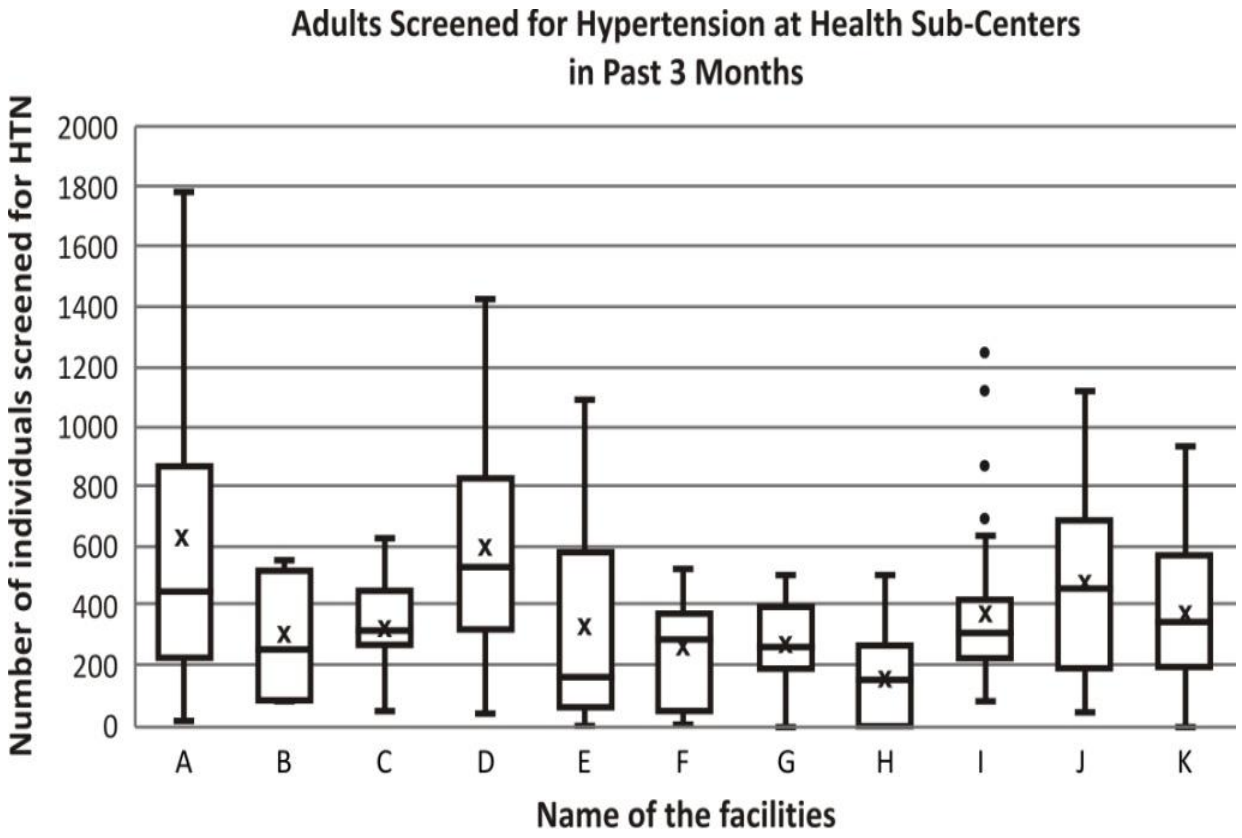
Removing the extreme ("outlier") values makes the box-plot easier to visualize.

Note that the extreme ("outlier") values represent health facilities that have unusual observations. These observations could be unusual for a number of reasons – exceptionally performing health facilities, data entry error, data reporting error, misreporting of information.

In any case, these outlier facilities need to be investigated and then a decision can be taken on removing or correcting them or not removing from the dataset.

Data source: NCD Portal, X district

Box plot



Data source: NCD Portal, X district

Question:

- How do blocks compare in terms of performance on hypertension screening?

Box plots can be used to compare groups. In the example shown, health facilities are grouped by block and box plots drawn for each block.

- Health facilities in some blocks had a wider range (e.g. A) than others as seen by the length of the box plots.
- The median health facility in D, J and A had higher number of screenings compared to others indicating better screening performance in these blocks.
- The median health facility in E and H was among the lowest across blocks.

Exercise 2: Distributions- Constructing a Box Plot

- Work in the same groups as before
- Open the Dataset_Module 2.2_Exercise 2-4
- Refer to the Instruction sheet_Module 2.2_Exercise 2

- Arrange the data in ascending order.
- Find the Five-Number Summary:
  - Minimum
  - Q1 (25th percentile)
  - Median (Q2)
  - Q3 (75th percentile)
  - Maximum
- Calculate the IQR (Q3 – Q1).
- Determine outliers, if any (beyond 1.5×IQR).
- Draw a box plot.

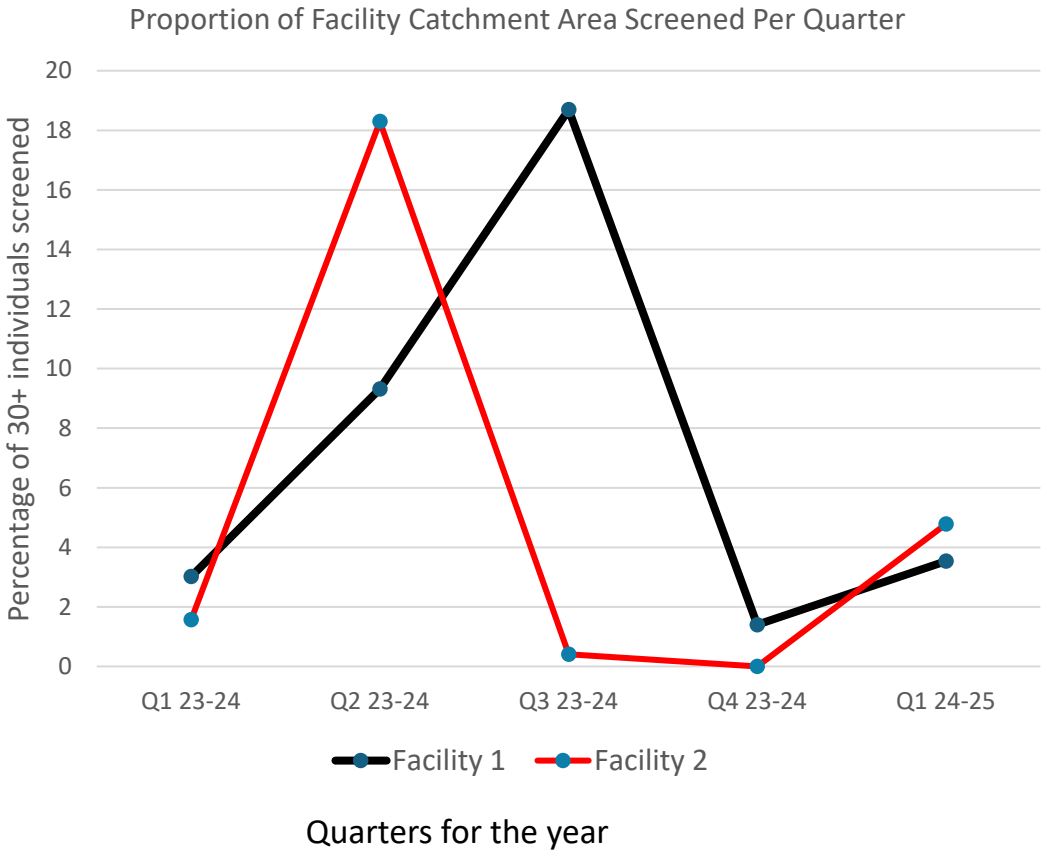
Exercise 2: Reflections

- Interpret:
  - What is the median LBW level?
  - Are there outliers?
  - Is the distribution skewed?
  - What can this tell us about LBW Status in blocks?

Session 3:
Line graphs and scatter plots

Duration: 40 min

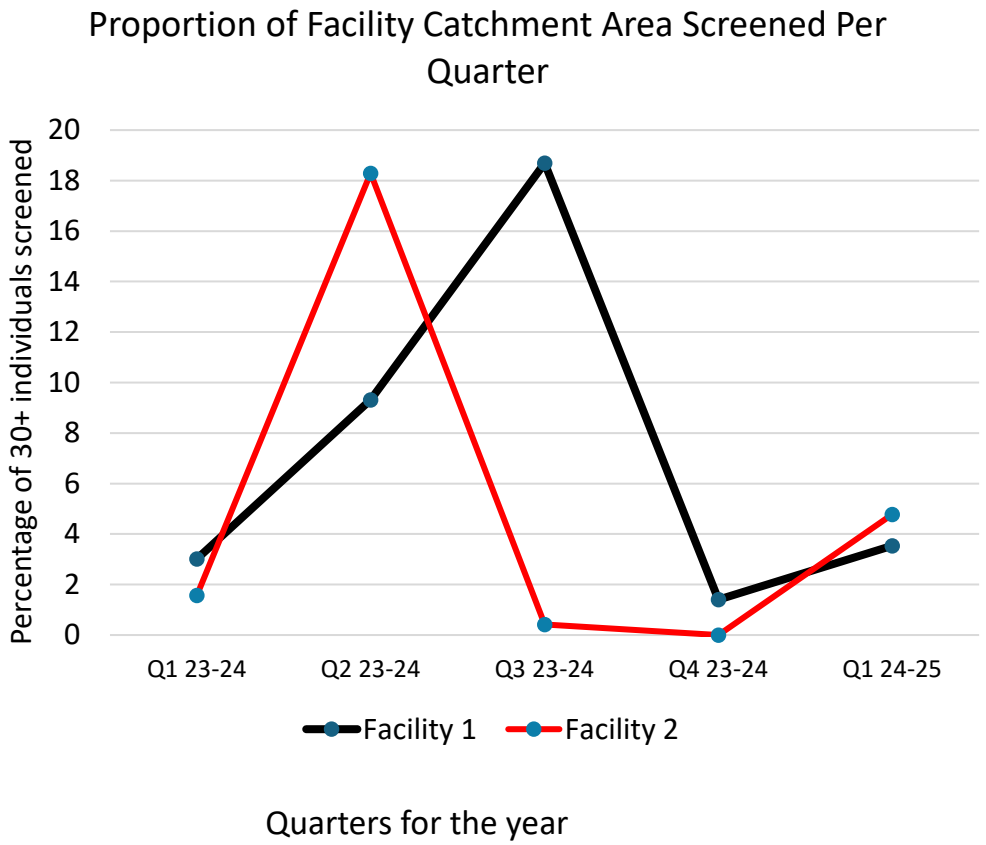
Time Trends



- Observing how units (eg. blocks, facilities) behave over time can provide useful insights into their performance
- Line graphs (see example) are typically used to show trends over time. It connects individual data points with a line to highlight the pattern.
- In the line graph shown, the average percentage of the adult population under a facility that is screened for hypertension is plotted for five quarters.

Data source: NCD Portal, Y district

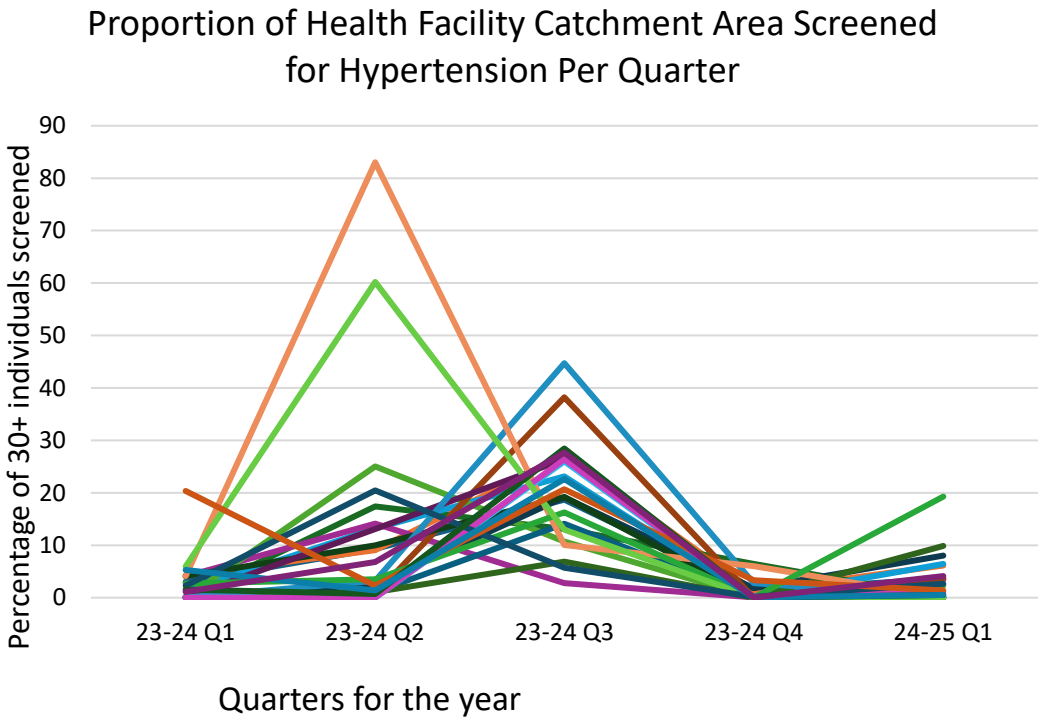
Time Trends



- There is high variability in the level of hypertension screening across quarters –
  - In both facilities the percentage population screened increases from Q1 to Q2/Q3 and then falls in Q3/Q4. It then picks up again in Q1.
- One insight from this is that the same level of effort at screening is not made in every quarter.
- In facility 2 almost no screenings were done in Q3 and Q4.

Data source: NCD Portal, Y district

Time Trends

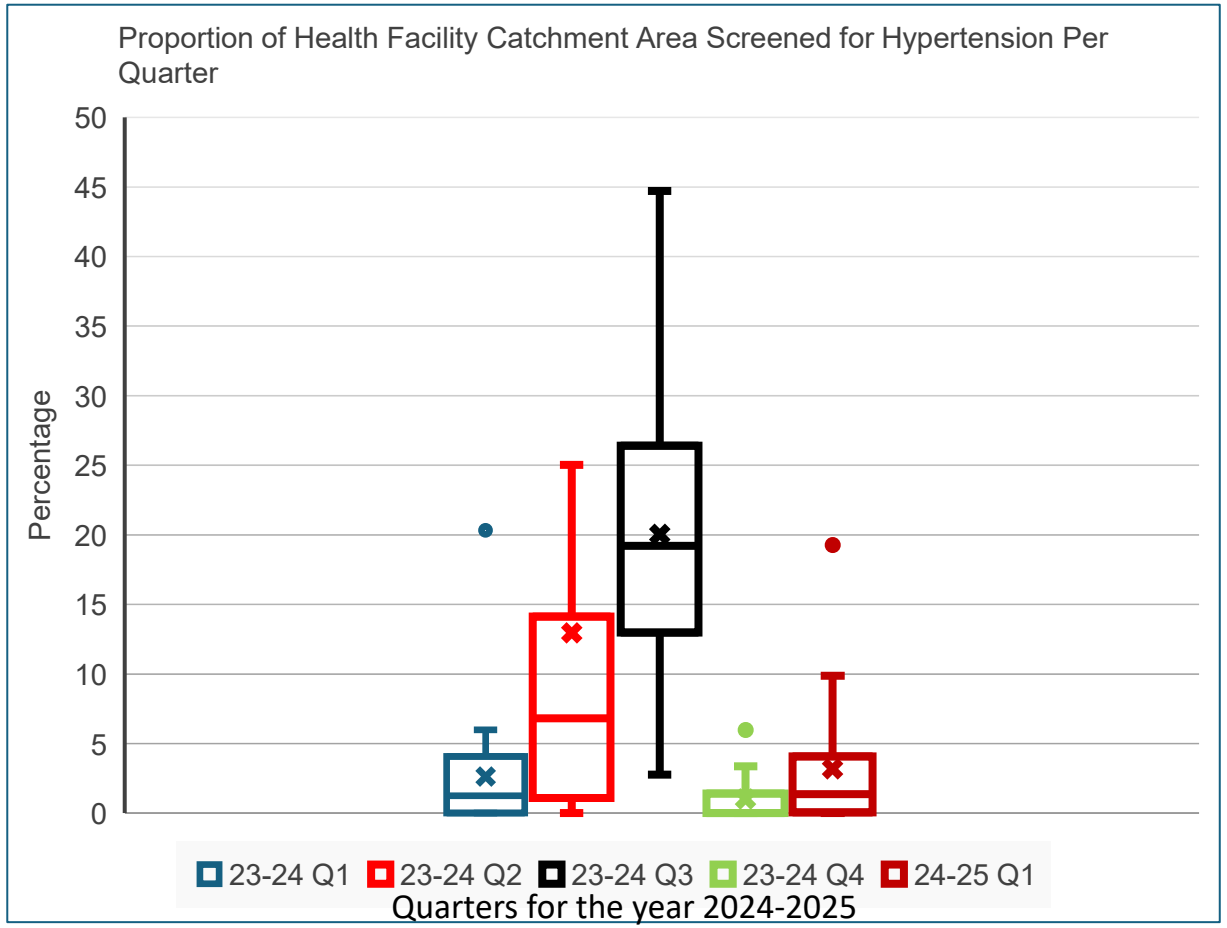


In the line graph shown, the percentage of the health facility adult population screened for hypertension is plotted for five quarters for 23 health facilities in a district

- When there are a large number of facilities/units it becomes difficult to clearly see patterns in a line graph (see example).
- One way to address this visualization problem is to use box-plots.

Data source: NCD Portal, Y district

Time Trends

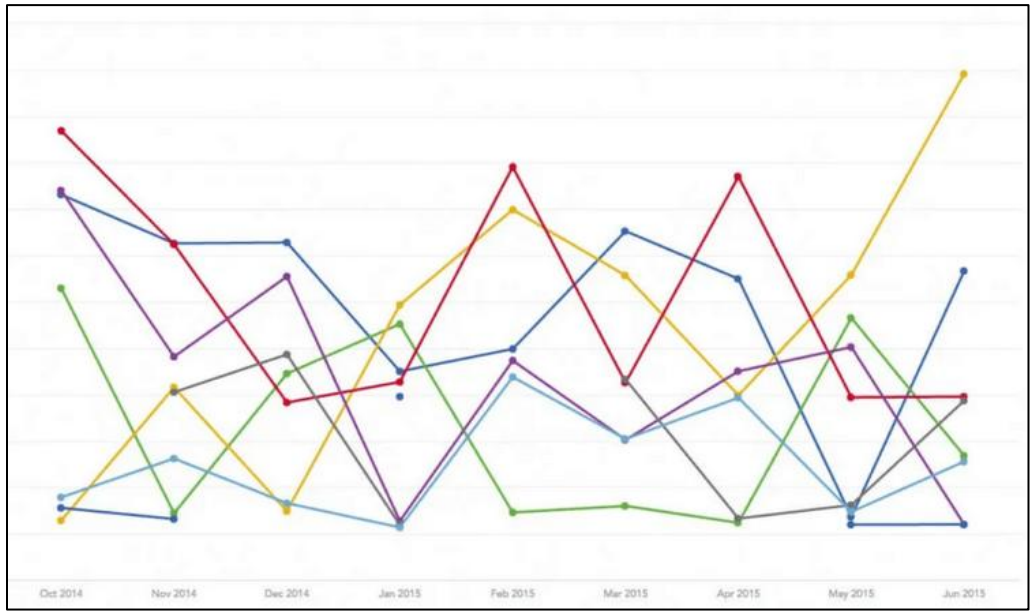


In the line graph shown, the percentage of the health facility adult population screened for hypertension is plotted for five quarters for 23 health facilities in a district.

- In this graph, box plots for each quarter are shown.
- The median level of hypertension screenings increases from Q1 to Q3 and then falls in Q4. It again increases in Q1 of the next year.
- Also, some extreme values are seen indicating potential outlier observations that need to be examined.

Data source: NCD Portal, Y district

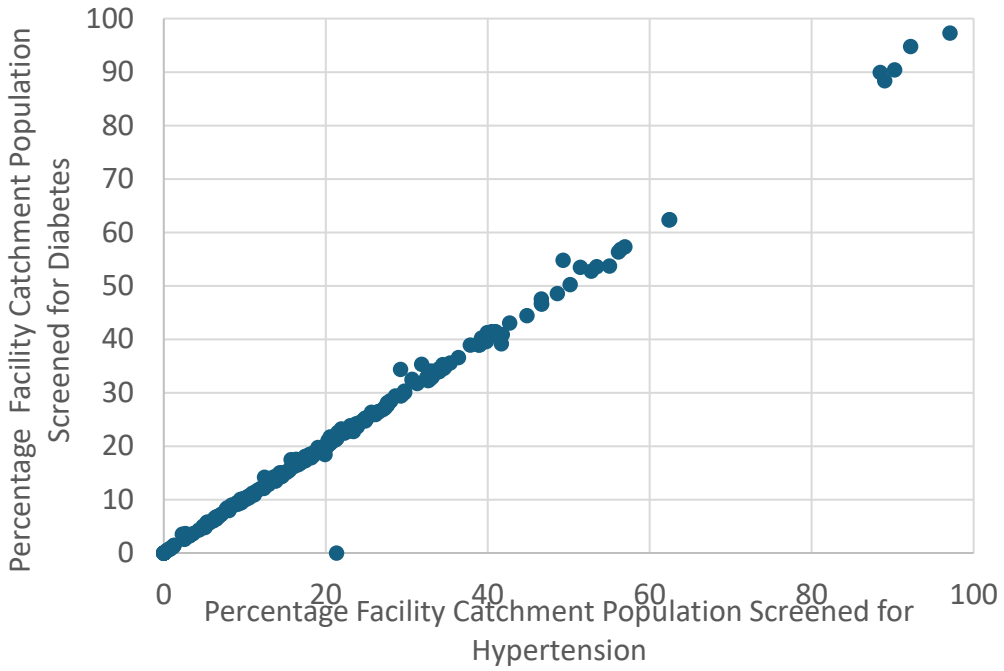
Spotting poor visualization practices: Line Chart – *Number of ANC Visits*



- **Problems:**
  - ➢ Overcrowded with too many lines, making it hard to differentiate trends.
  - ➢ Lack of clear labeling for each line (no legend or identifiers).
  - ➢ Colors are too similar
- **Better Approach:**
  - ➢ Use fewer lines or split into multiple charts.
  - ➢ Add clear legends/labels.
  - ➢ Choose distinct colors.

Image source: Storytelling with Data, Knaflic

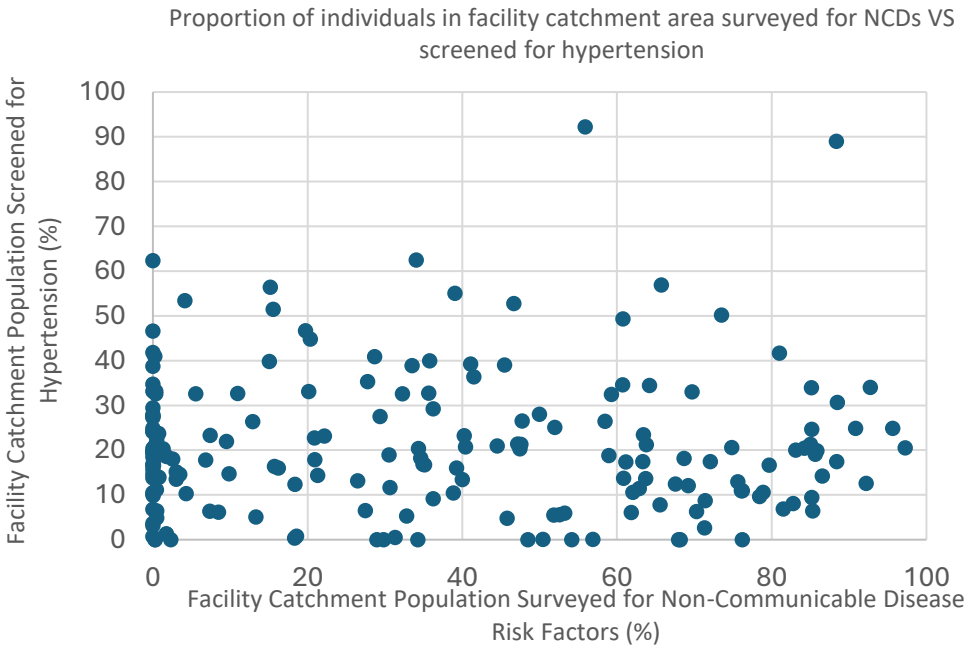
Scatter Plots



- A scatter plot is a useful way of showing the relationship between two variables.
- In the graph shown, each point represents a health facility. The percentage of the catchment population screened for diabetes (Y-axis) is plotted against the percentage screened for hypertension in the same facility catchment area.
- As one can see, there is a very strong positive association between the level of diabetes and hypertension screening i.e. the points are nearly all along a straight line.
- This strong association should not be surprising because individuals are often screened for hypertension and diabetes at the same instance.

Data source: NCD Portal, Y district

Scatter Plots



- In India, facility catchment populations are typically first surveyed for non-communicable disease risk factors (CBAC) followed by screening.
- Question: Are levels of hypertension screening associated with levels of facility catchment populations surveyed for non-communicable disease risk factors?
- The scatter plot shown, each point represents a health facility, the percentage of the catchment population screened for hypertension (Y-axis) is plotted against the percentage surveyed for non-communicable diseases in the same facility catchment area.
- One can see, there is no clear association between the level of hypertension screening and the level of non-communicable disease risk factor survey in catchment populations.

Data source: NCD Portal, Y district

Spotting poor visualization practices: Scatter Plot – *Years until retirement by gender*

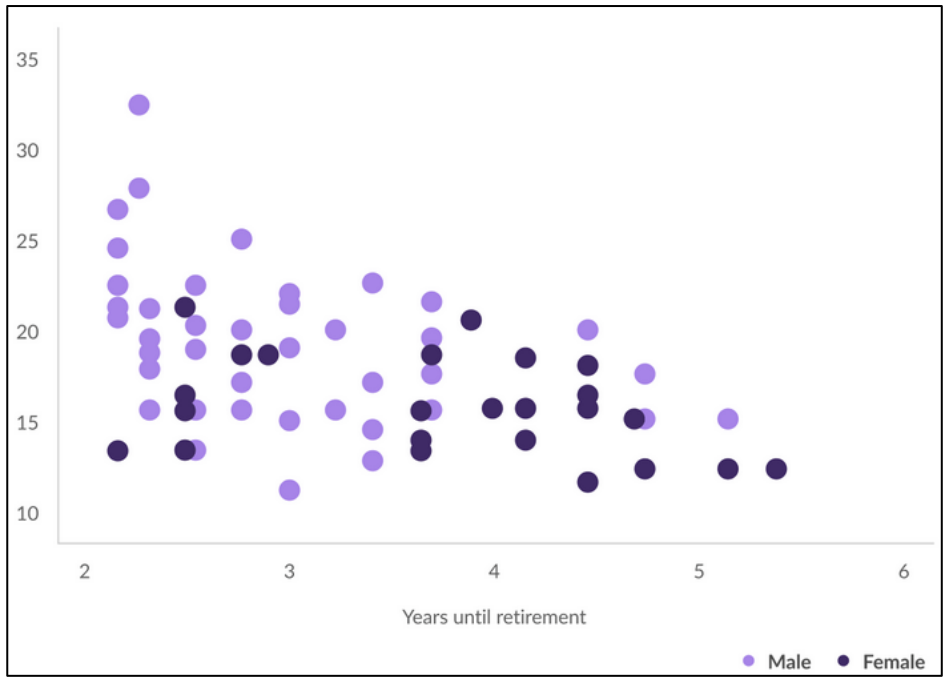


Image source: Storytelling with Data, Knaflic

Problems:

- Overlapping points make it hard to see density or exact counts (overplotting).
- Color difference between Male and Female is subtle and may not be clear to all viewers.
- No clear title or explanation of what patterns we should expect to see.

Better Approach:

- Use jittering or transparency to handle overlapping points.
- Add clear and distinct colors for categories.
- Include a descriptive title and labels to highlight insights.

Exercise 3: Line graph

- Work in the same groups as before
- Open the Dataset_Module 2.2_Exercise 2-4
- Refer to the Instruction sheet_Module 2.2_Exercise 3

- Plot a line graph: Block-wise Comparison of Total Births and Low Birth Weight (LBW) Cases
  - X-axis: Block/Facility (in order of appearance)
  - Y-axis: Number of births
- Draw two lines:
  - One for total live births
  - One for LBW cases
- Questions for Discussion:
  - Which blocks/facilities have high total births but low LBW?
  - Where are LBW numbers disproportionately high?

Exercise 4: Scatter plot

- Work in the same groups as before
- Open the Dataset_Module 2.2_Exercise 2-4
- Refer to the Instruction sheet_Module 2.2_Exercise 4

- Plot a scatter plot: Relationship Between Total Births and Low Birth Weight (LBW) Cases by Block/Facility
  - X-axis: Total live births
  - Y-axis: LBW cases
  - Each point represents one block/facility.
  - Label at least 5 points
- Questions for Discussion:
  - Do blocks/facilities with more deliveries have proportionally more LBW cases?
  - Are there any outliers (e.g., low births but high LBW)?
  - Is there a positive correlation?

Key Takeaways (1/2)

- **Analysis of continuous data** is essential for understanding facility performance- use descriptive statistics (mean, median, range, SD) to summarize large datasets.
- Data visualization transforms analysis into clear, actionable insights for decision-making
- **Design principles remain important** : choose appropriate chart type, simplify visuals, and ensure clarity.
- **Interpretation before action** : always investigate patterns and anomalies before making program or policy changes.

Key Takeaways (2/2)

- **Histograms** help visualize frequency distributions, detect skewness, and identify outliers.
- **Box plots** show central tendency, spread, skewness, and highlight unusual values useful for comparing groups (e.g., across blocks).
- **Outlier detection** is critical in investigate whether they indicate exceptional performance, errors, or data issues.

- **Time trends** (line graphs or quarterly box plots) reveal performance changes and seasonal patterns.
- **Scatter plots** display relationships between two variables which is useful for identifying associations, clusters, and anomalies.

Thank You