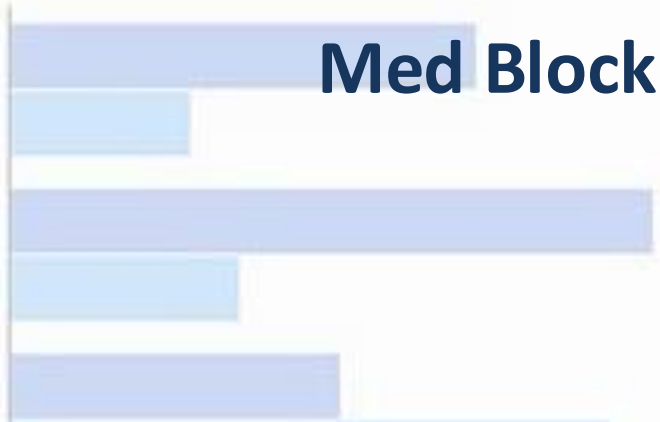
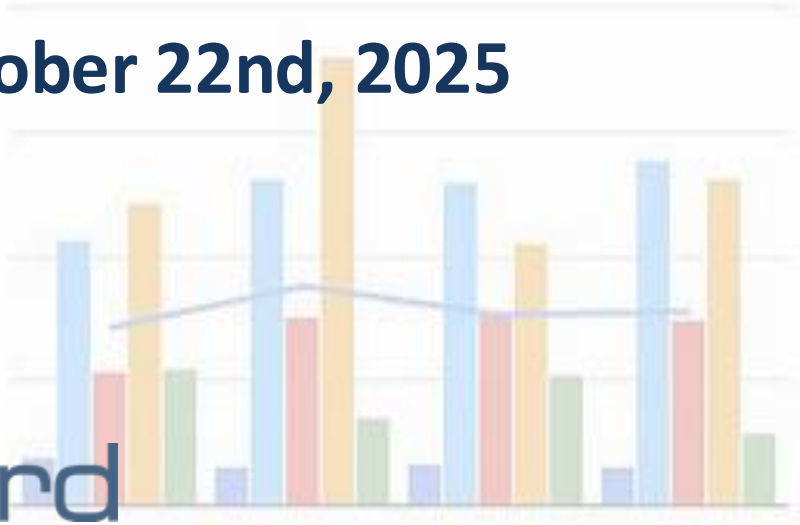
Data Visualization: Effective Plotting & Plot Interpretation

UBC Bioinformatics and Statistics Workshop #2

Med Block C, October 22nd, 2025





THE UNIVERSITY OF BRITISH COLUMBIA

BC Children's Hospital

Centre for Heart Lung Innovation UBC and St. Paul's Hospital

icord

Workshop Outline

- What is data visualization?
- What are the different types of plots we use to visualize data?
- How do we choose the right plot to display our findings?

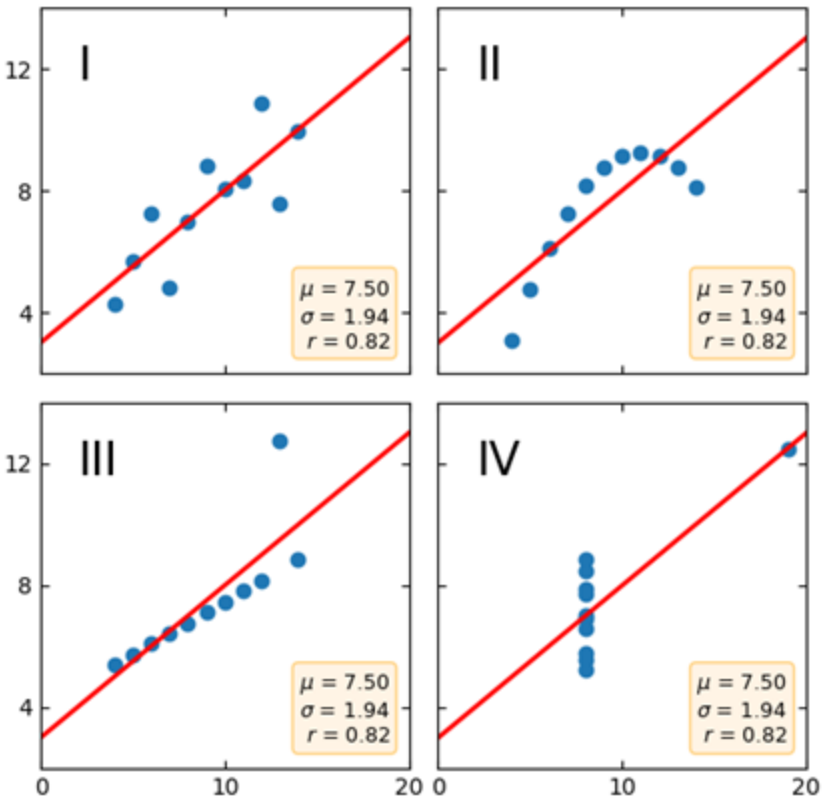
If you have any questions following the workshop please feel free to reach out to session instructors: **Jaycee Farmer** (jayfarm@student.ubc.ca) and **Cindy Wang** (xwang75@student.ubc.ca)

What is Data Visualization?

Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.

Data-Visualization tools and techniques offer executives and other knowledge workers new approaches to dramatically improve their ability to grasp information hiding in their data.

Importance of Data Visualization: Anscombe's Quartet



Types of Plots

Basic Visualization:

➔ Histogram
➔ Bar / Line Chart
➔ Box plot
➔ Scatter plot

Advanced Visualization:

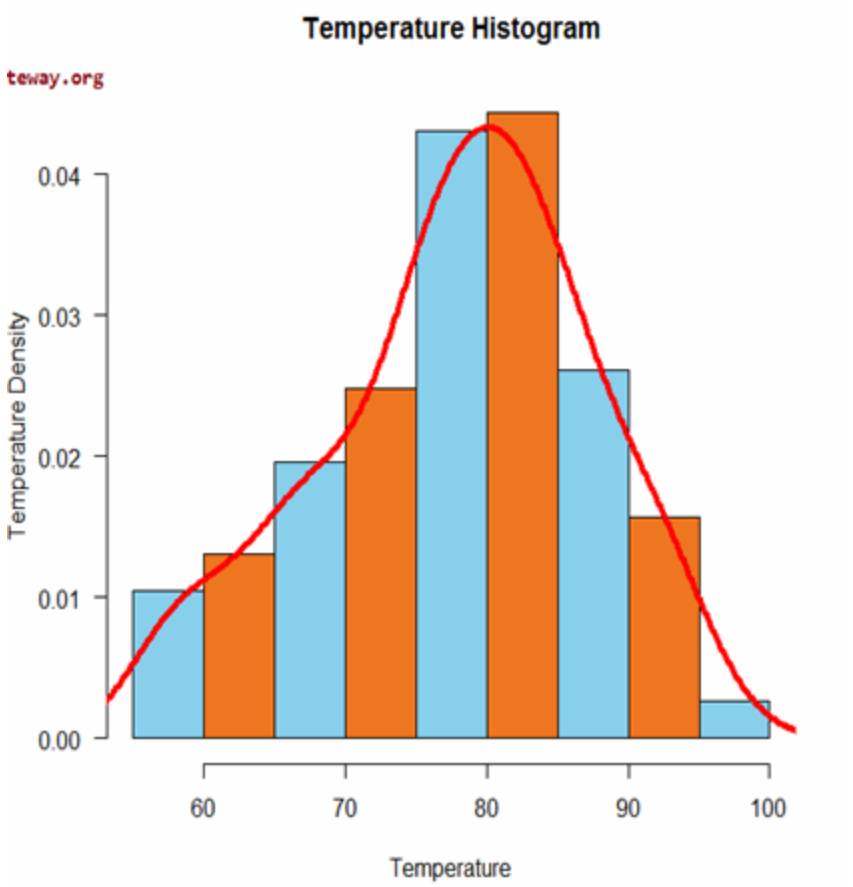
➔ Heat Map
➔ Forest Plots
➔ Volcano Plots

Histograms

Histogram is basically a plot that breaks the data into bins (or breaks) and shows frequency distribution of these bins. You can change the breaks also and see the effect it has data visualization in terms of understandability.

Below is a simple example of the code for a line chart:

```
hist(airquality$temp, freq=FALSE, main=
"Temperature histogram",
        xlab= "Temperature", ylab=
"Temperature density",
    las=1, col=c("skyblue", "chocolate2"))
```

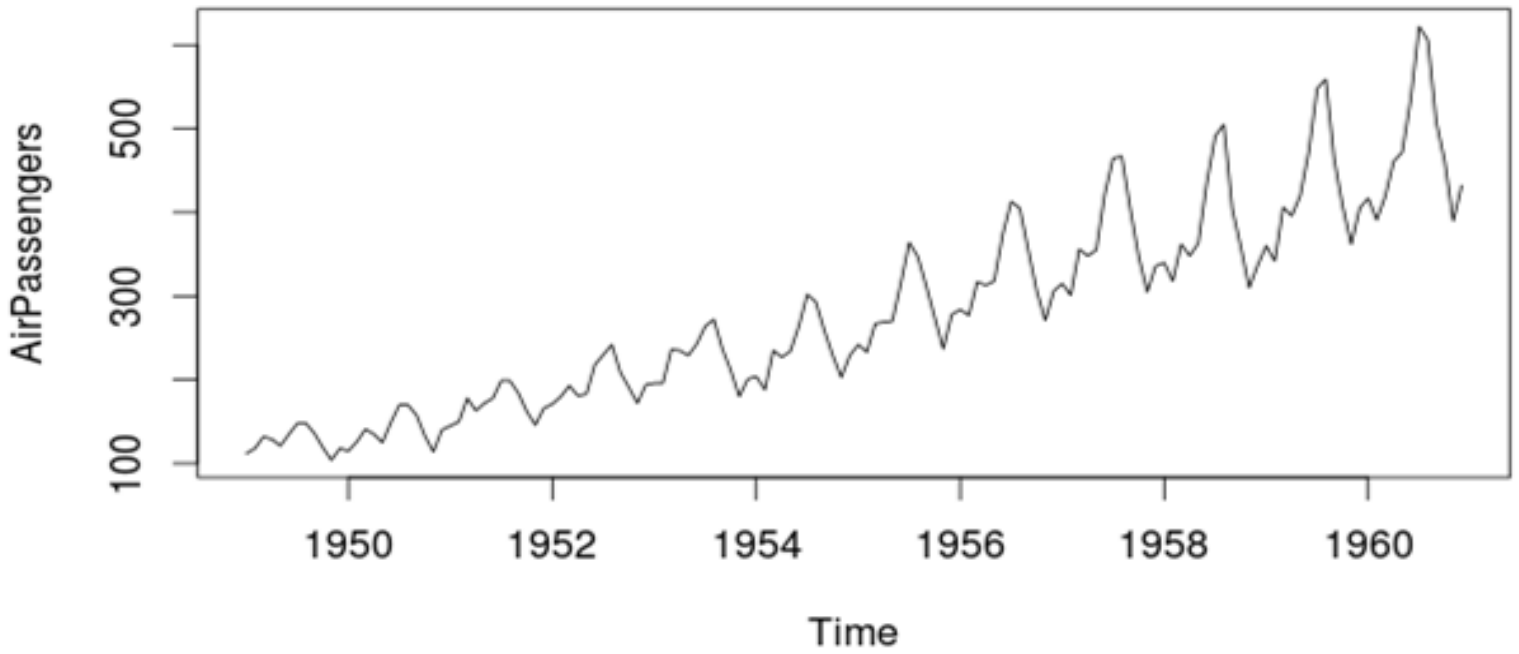


Line Charts

A line chart is a graphical representation that displays data points connected by straight lines to show trends, changes, and relationships over a continuous interval, most commonly time.

Below is a simple example of the code for a line chart:

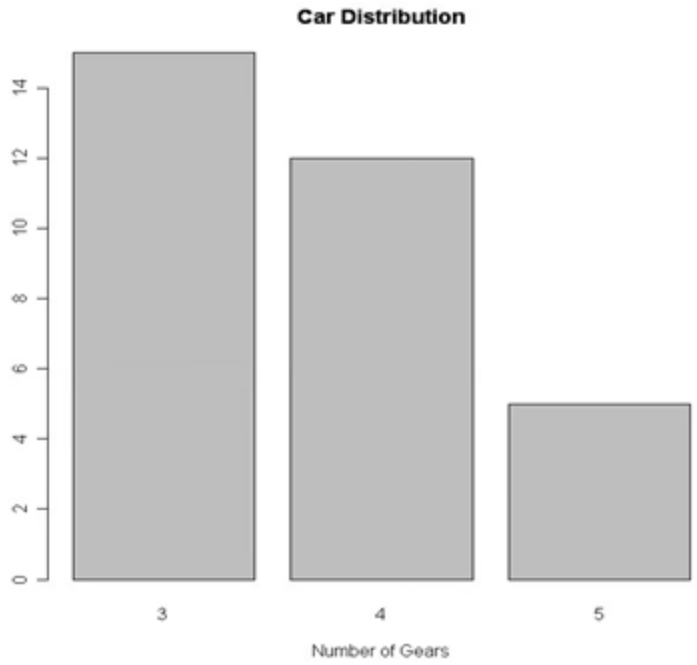
```
plot(AirPassengers,type="l")
```



Bar Chart

Bar Plots are suitable for showing comparison between cumulative totals across several groups. Stacked Plots are used for bar plots for various categories. A simple example of code for a bar plot would be:

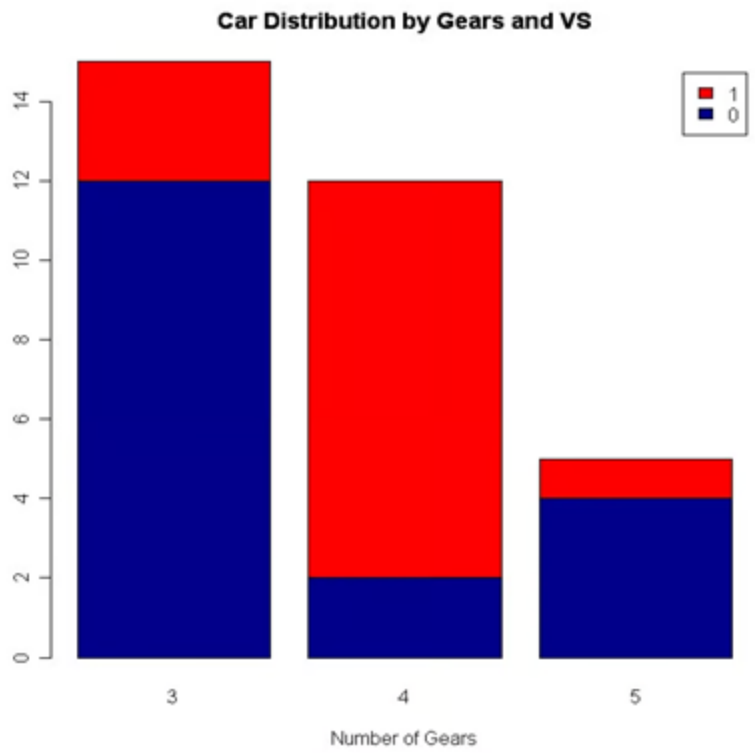
```
counts <- table(mtcars$gear)
barplot(counts, main="Car Distribution",
  xlab="Number of Gears")
```



Stacked Bar Chart

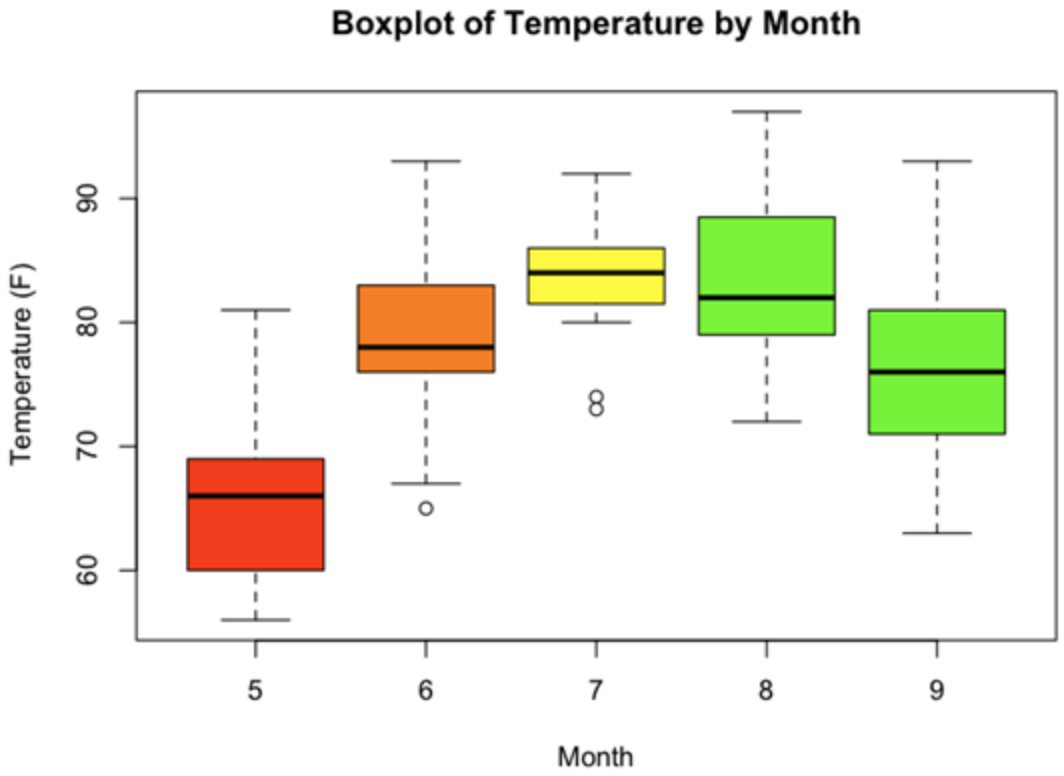
Stacked Bar Chart extends the standard bar chart from looking at numeric values across one categorical variable to two. Each bar in a standard bar chart is divided into a number of sub-bars stacked end to end, each one corresponding to a level of the second categorical variable.

```
counts <- table(mtcars$vs, mtcars$gear)
barplot(counts, main="Car Distribution by Gears and VS",
  xlab="Number of Gears", col=c("darkblue"
  legend = rownames(counts))
```



Box Plots

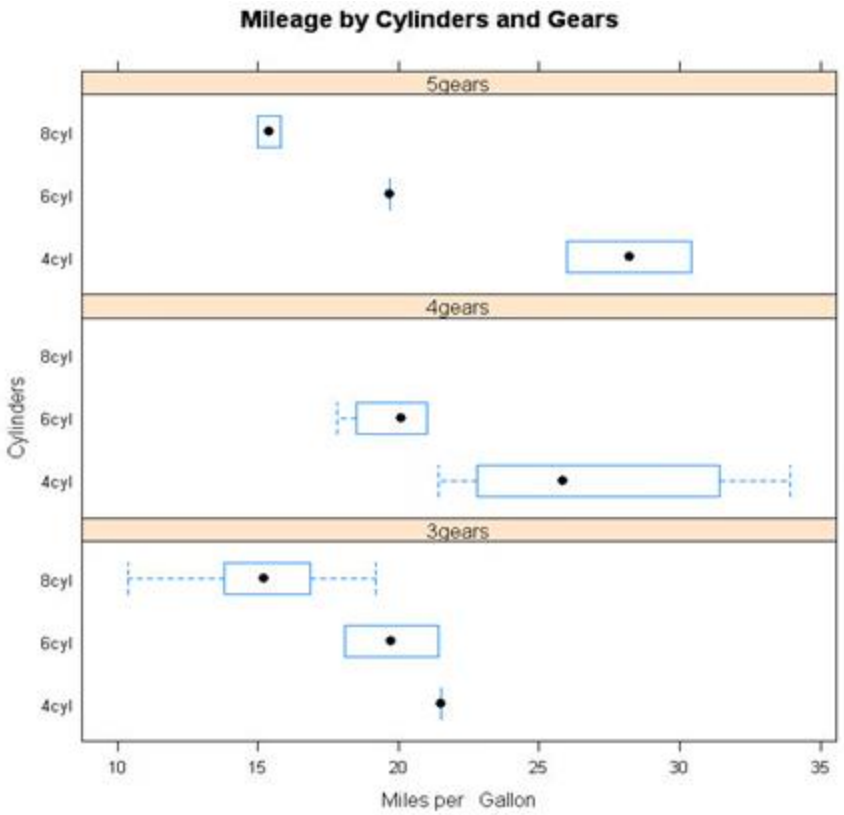
Box Plot shows 5 statistically significant numbers - the minimum, the 25th percentile, the median, the 75th percentile and the maximum. It is thus useful for visualizing the spread of the data is and deriving inferences accordingly.



Box Plots (including group-by option)

Here is a simple example of code that could be used to generate a box plot in R:

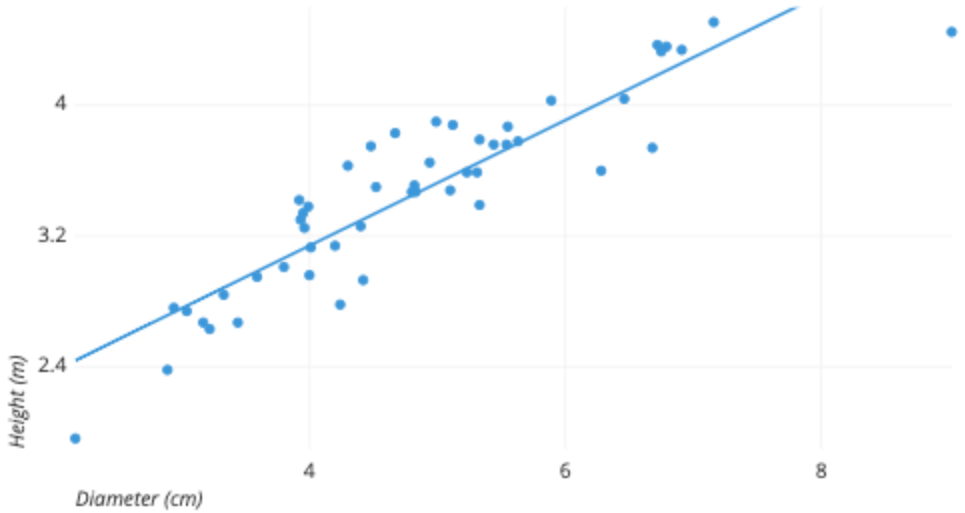
boxplot(mpg~cyl,data=mtcars, main="Mileage by Cylinders and Gears", xlab="Miles Per Gallon", ylab="Cylinders")



Scatterplot

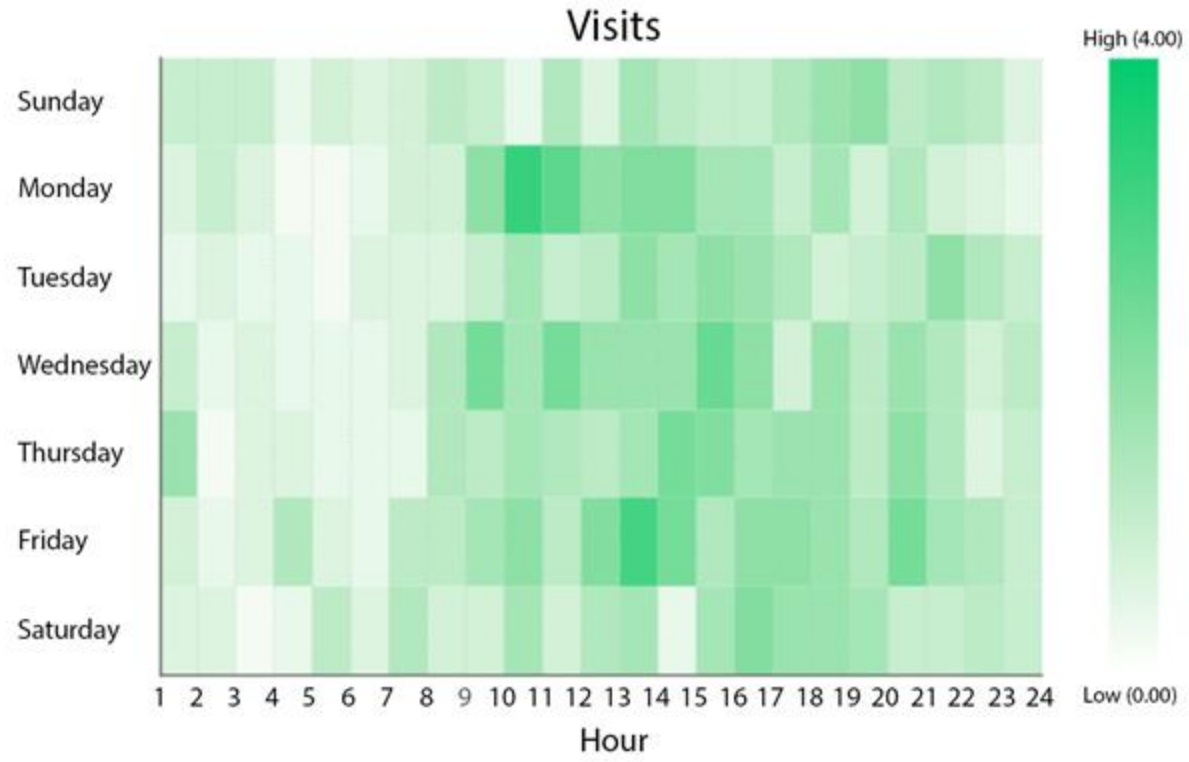
A scatter plot uses dots to represent values for two **numeric** variables. Each dot indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

```
plot(diameter, height,
        main = "My Scatterplot",
        xlab = "X-axis Label",
        ylab = "Y-axis Label",
        pch = 19,
        col = "blue")
```



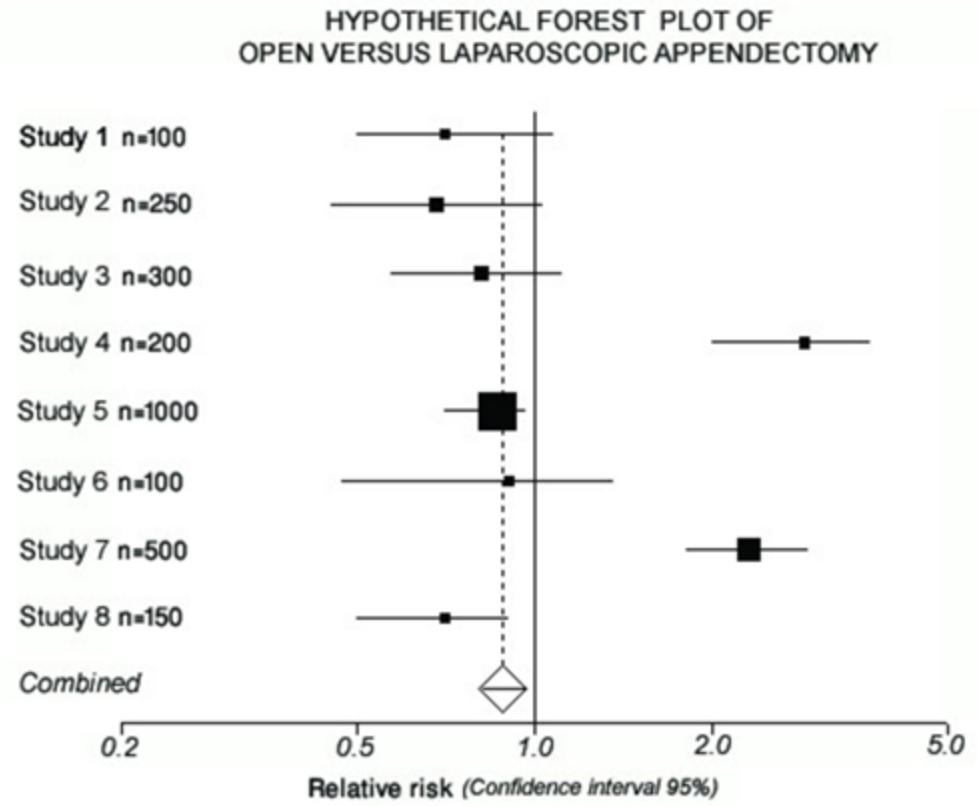
Heatmap

A heatmap depicts values for a main variable of interest across two axis variables as a grid of colored squares. The axis variables are divided into ranges. Heatmaps are used to show relationships between the 2 variables. By observing how cell colors change across each axis, you can observe if there are any patterns in value for one or both variables.



Forest Plot

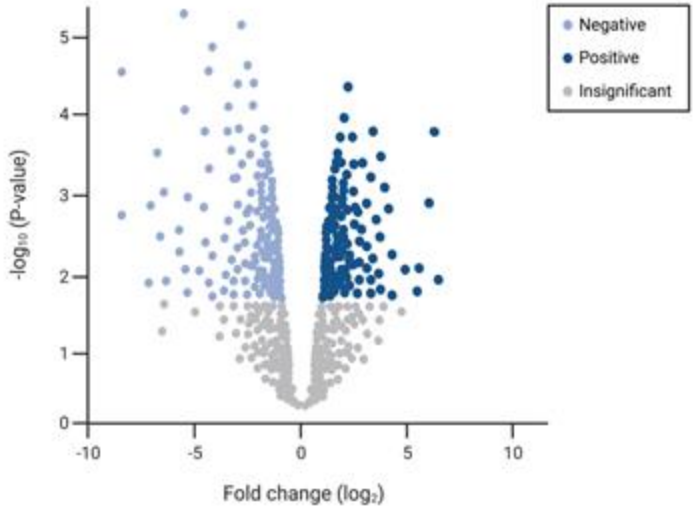
A forest plot is a graphical way to show estimates of effect sizes (such as odds ratios, relative risks, or hazard ratios) along with their confidence intervals. It's most commonly used in meta-analyses, to visualize and compare results from multiple studies or subgroups at once.



Volcano Plot

A volcano plot is a scatter plot used in data analysis (esp. in genomics, proteomics, transcriptomics) to show both statistical significance and magnitude of change for each data point — usually genes, proteins, or metabolites.

You have a group of cells A and a group of cells B. Group B was treated with a drug. Now you want to see what effect the drug has in gene expression. Does the drug cause some genes to be upregulated? Does it downregulate the expression of other genes? By how much?



Mini Exercise

1. In a cohort study, you measured blood glucose levels and insulin resistance (HOMA-IR) among 500 participants. You want to assess whether there's a correlation between these two continuous variables.
   → Which visualization would best depict this relationship?

1. You're studying average blood pressure measurements taken monthly over a year for two treatment groups in a longitudinal study. You want to visualize how these averages change over time.
   → Which plot would best communicate the trend and group differences?

1. You are analyzing RNA-seq data comparing gene expression levels across three different cancer subtypes. You want to display both the median expression and the spread (variability) of expression for a set of 50 genes.
   → Which plot would best show the distribution and variability of expression values across subtypes?

4. You collected data on the length of hospital stay (in days) for 1,200 patients and want to identify whether the data are skewed or if there are outliers.
→ Which plot would best illustrate the shape of this distribution and potential outliers?

5. You conducted a meta-analysis of 12 clinical trials evaluating the efficacy of a new antihypertensive drug. Each study reports an odds ratio and a 95% confidence interval.
→ Which plot would best summarize and compare these study results visually?

6. You've identified thousands of genes from a transcriptomic study, each with a log2 fold change and a p-value when comparing tumor versus normal tissue.
→ Which visualization technique would best highlight significantly upregulated and downregulated genes?

Mini Exercise - Answer

1. In a cohort study, you measured blood glucose levels and insulin resistance (HOMA-IR) among 500 participants. You want to assess whether there's a correlation between these two continuous variables.
   → Which visualization would best depict this relationship?
   → scatterplot
1. You're studying average blood pressure measurements taken monthly over a year for two treatment groups in a longitudinal study. You want to visualize how these averages change over time.
   → Which plot would best communicate the trend and group differences?
   → line chart
1. You are analyzing RNA-seq data comparing gene expression levels across three different cancer subtypes. You want to display both the median expression and the spread (variability) of expression for a set of 50 genes.
   → Which plot would best show the distribution and variability of expression values across subtypes?
   → boxplot

4. You collected data on the length of hospital stay (in days) for 1,200 patients and want to identify whether the data are skewed or if there are outliers.

→ Which plot would best illustrate the shape of this distribution and potential outliers?

→ histogram

5. You conducted a meta-analysis of 12 clinical trials evaluating the efficacy of a new antihypertensive drug. Each study reports an odds ratio and a 95% confidence interval.

→ Which plot would best summarize and compare these study results visually?

→ forest plot

6. You've identified thousands of genes from a transcriptomic study, each with a log2 fold change and a p-value when comparing tumor versus normal tissue.

→ Which visualization technique would best highlight significantly upregulated and downregulated genes?

→ volcano plot

Thank you! Questions?

The workshop recording and relevant materials will be available in the UBC Bioinformatics and Statistics Workshops Teams Channel following the session.

For additional questions, contact: Jaycee Farmer (jayfarm@student.ubc.ca); Cindy Wang (xwang75@student.ubc.ca)

Post-session Survey:
https://ubc.ca1.qualtrics.com/jfe/form/SV_39upHAA9rcviot0