Module 1D

Visualization

Module 1 : Descriptive Statistics

Data Visualization

Agenda:

- Data types and their common visualizations:
  - Scatterplots
  - Mosaic and bar plots
  - Histograms
  - Box and Violin plots
  - Cumulative Frequency Distributions
- Interpretation of popular plots in genomics

Types of data:

Categorical Variable

- AKA Class variables or Nominal variables
- They do not have magnitude on a numerical scale
- **Nominal**
  - Lack inherent order
  - Ex: blood type, genotype, sex, state, survival (live or die), drug treatment (aspirin vs ibuprofen)
- **Ordinal**
  - Inherent order
  - Ex. **age,** education level/degree

Quantitative Variables

- AKA Numerical variables
- Random Variable is a Quantitative variable
- **Continuous**
  - Ability to take any value ex.. Human weight, **age**
  - **They can be measured**
- **Discrete**
  - Spaces between possible values ex. Number of offspring, **age**
  - **They can be counted**

Data type determines plot type

- https://www.data-to-viz.com/ ←(and their code in Python and R)
- https://statisticsbyjim.com/graphs/
- https://piktochart.com/blog/types-of-graphs/
- https://www.sciencedirect.com/science/article/pii/S2666389920301896
- https://www.nature.com/articles/d41586-023-03393-9

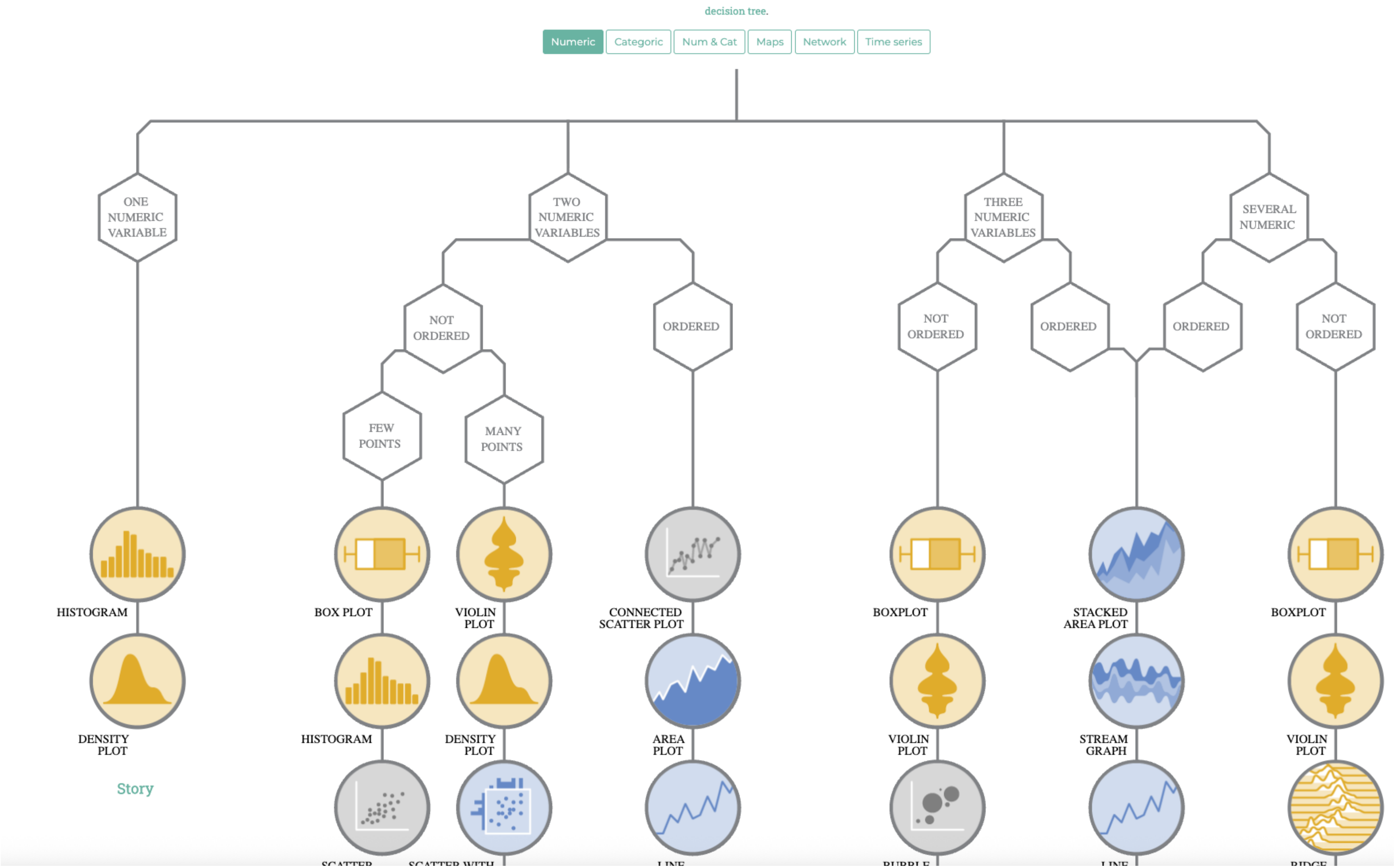
https://www.edwardtufte.com/tufte/

https://monachalabi.com/

The plots we will examine:

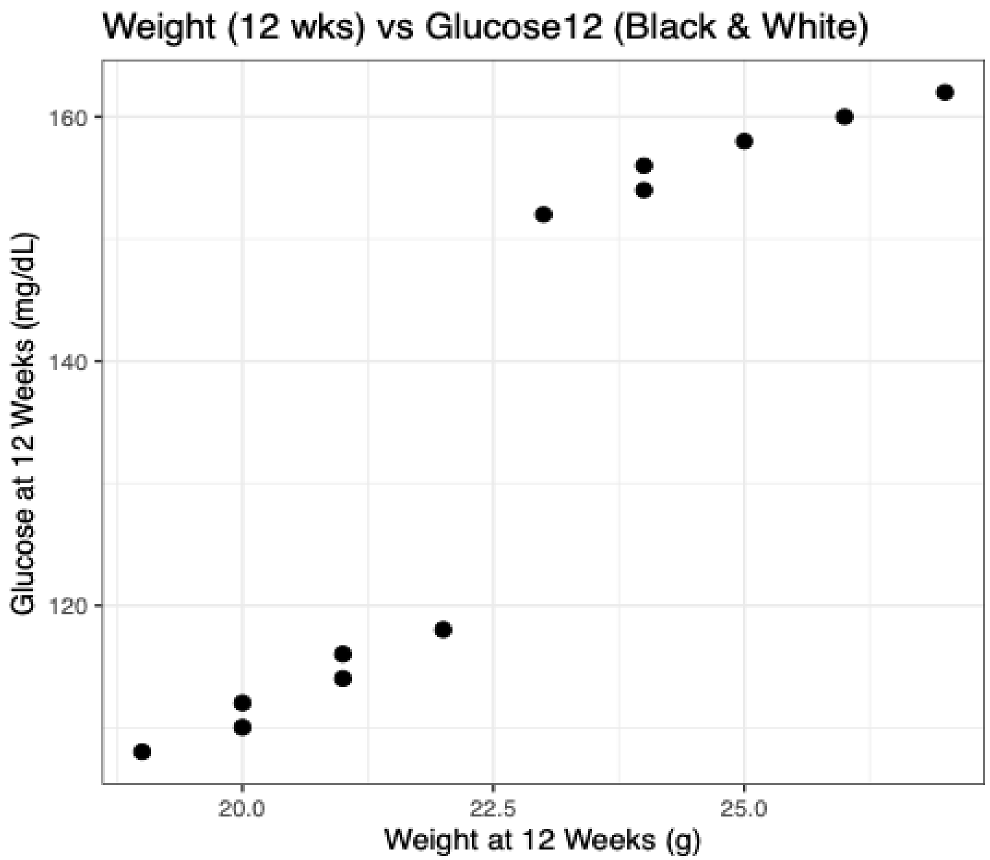
- Scatterplots
- Histograms
- Mosaic plots, Bar plots
- Boxplots & Violin plots
- Cumulative Frequency Plots

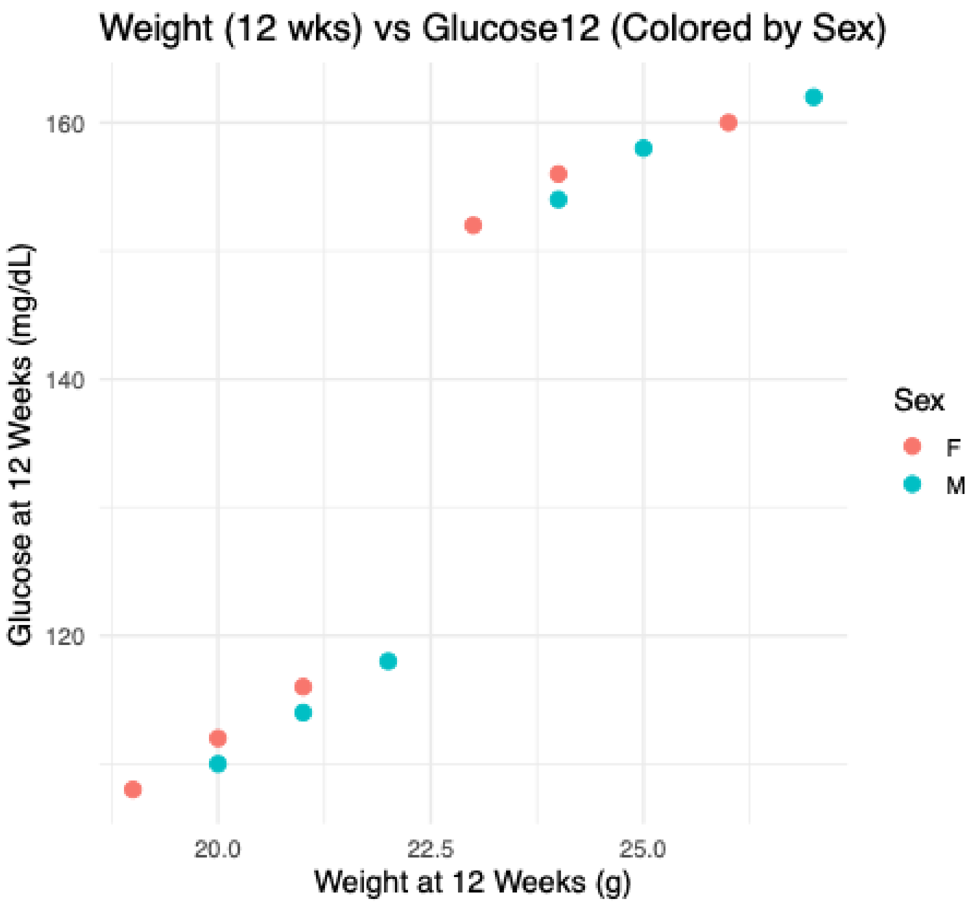
Data type determines plot type



Types of Data	Graphical Method
Two numeric variables	Scatter
Two categorical variables	Grouped Bar Plot
	Mosaic Plot
One numeric variable, **One categorical variable**	Violin plot/Boxplot
	Cumulative Frequency Distribution
	Multiple Histograms

Scatterplot

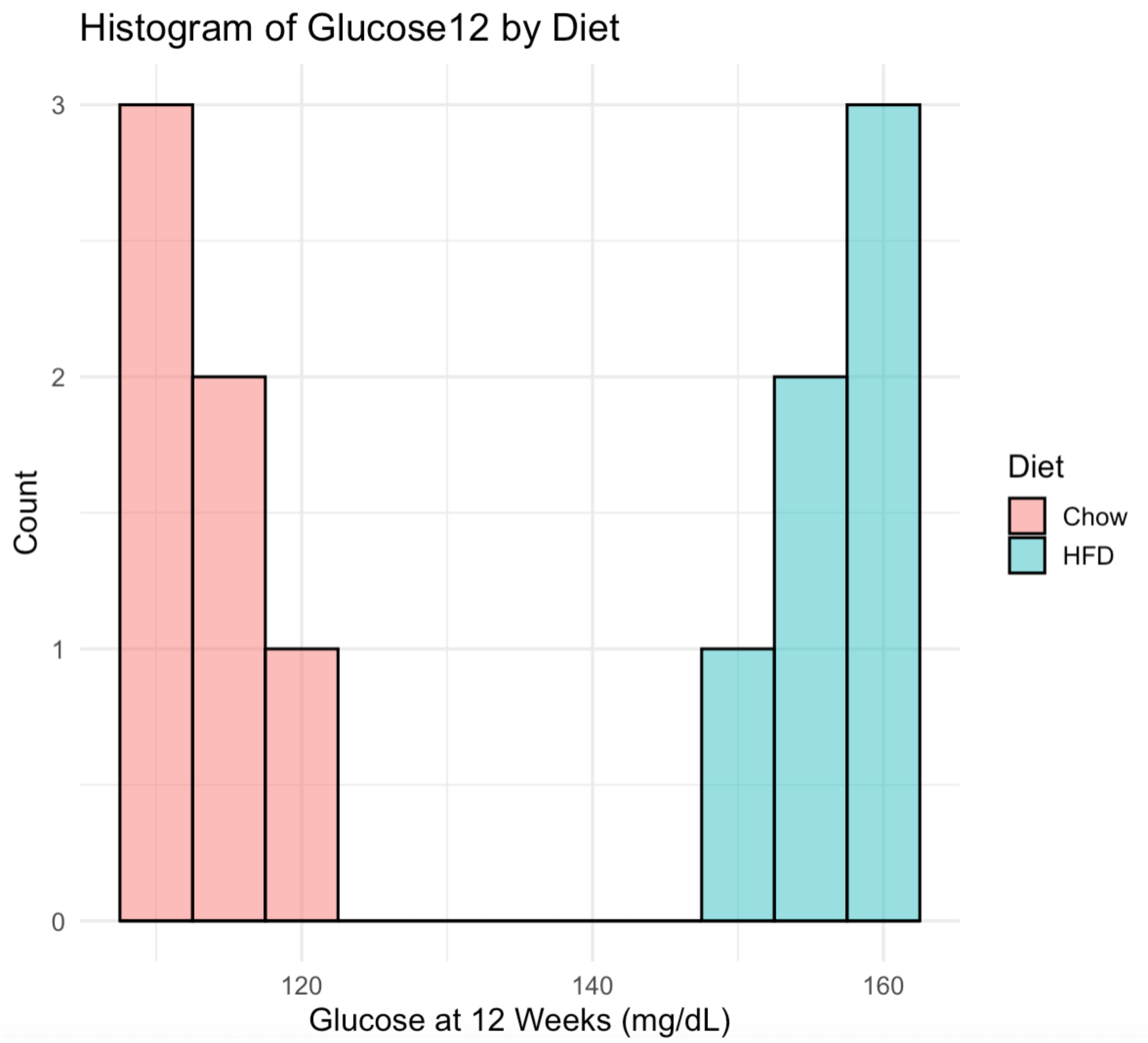




Free online textbook that gives r code!
https://bookdown.org/dli/rguide/scatterplots-and-best-fit-lines-two-sets.html

Hans Rosling ted talk (his website has data visualizations – scatterplots that move!- and datasets):
https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen

Histogram

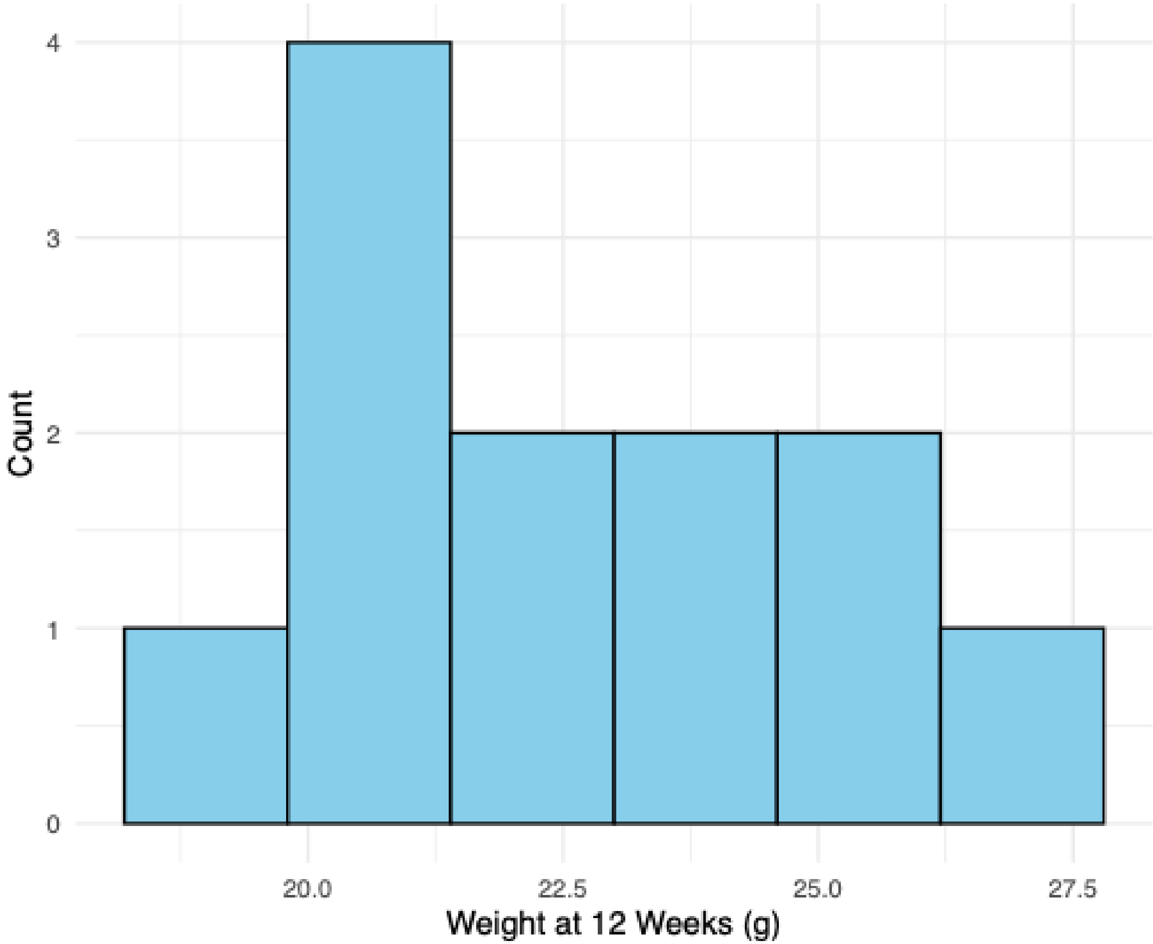


One warning about histograms:
Be careful about “bin” size; you can introduce artefacts!

Histogram

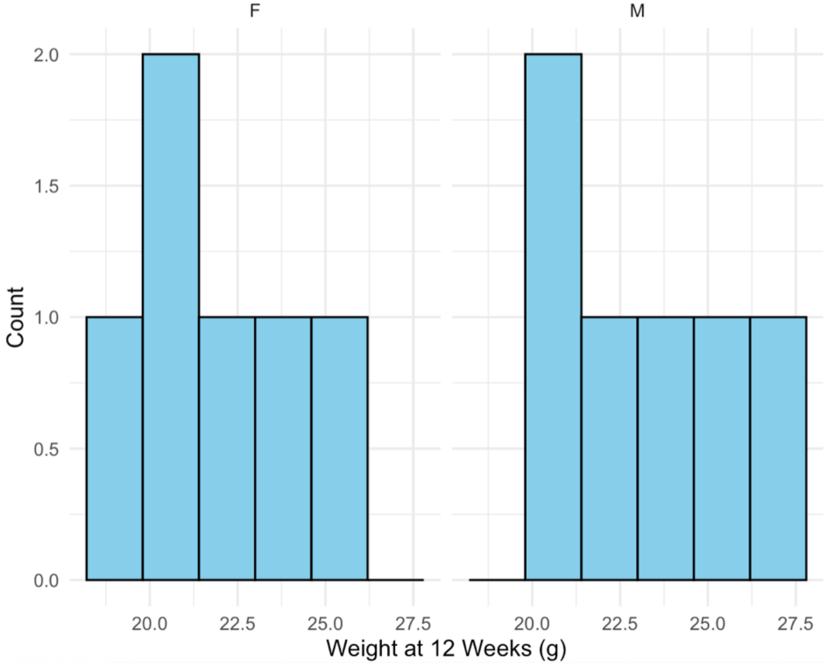
Histogram of Weight12

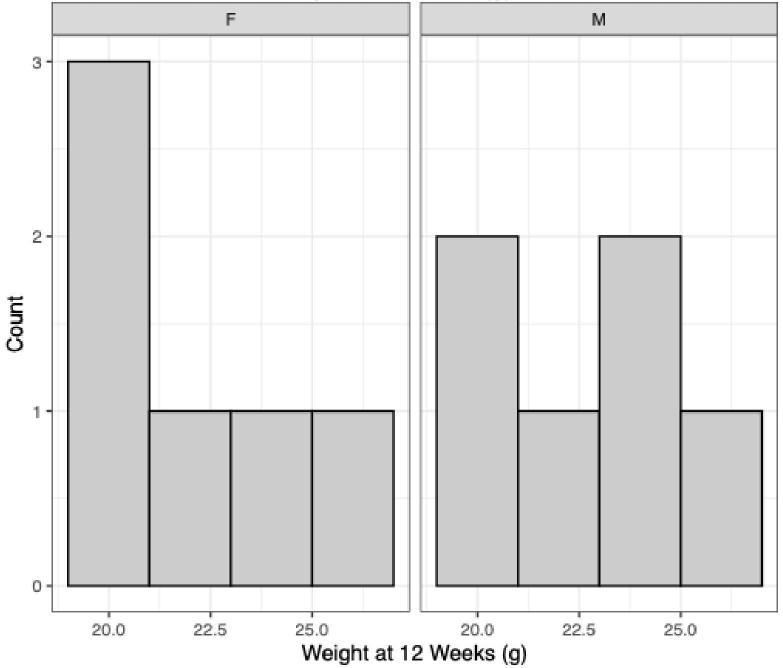
A small dataset but approximately bell–shaped



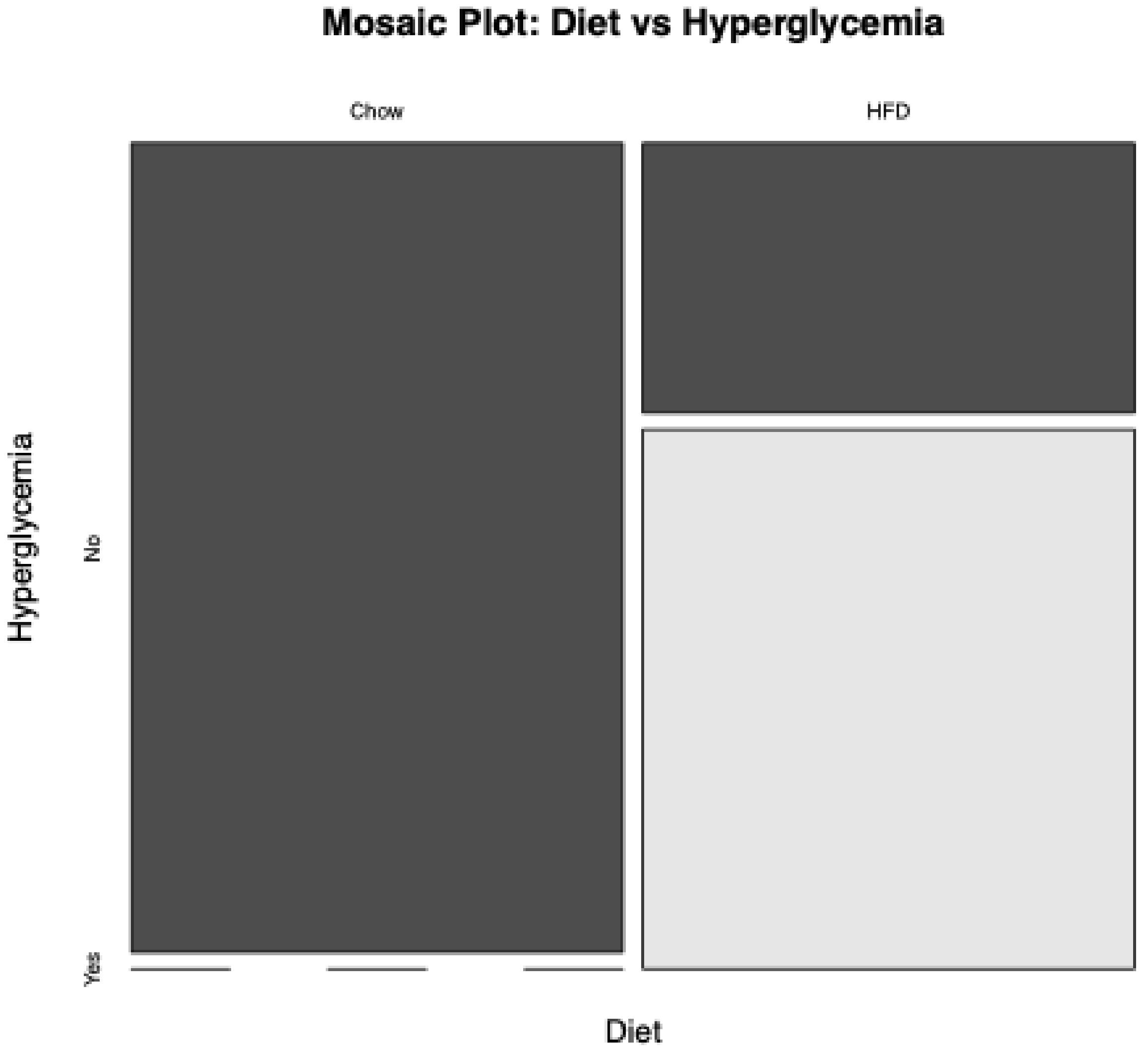
Weight12 Histograms by Sex

Separating males and females produces more normal-like distributions

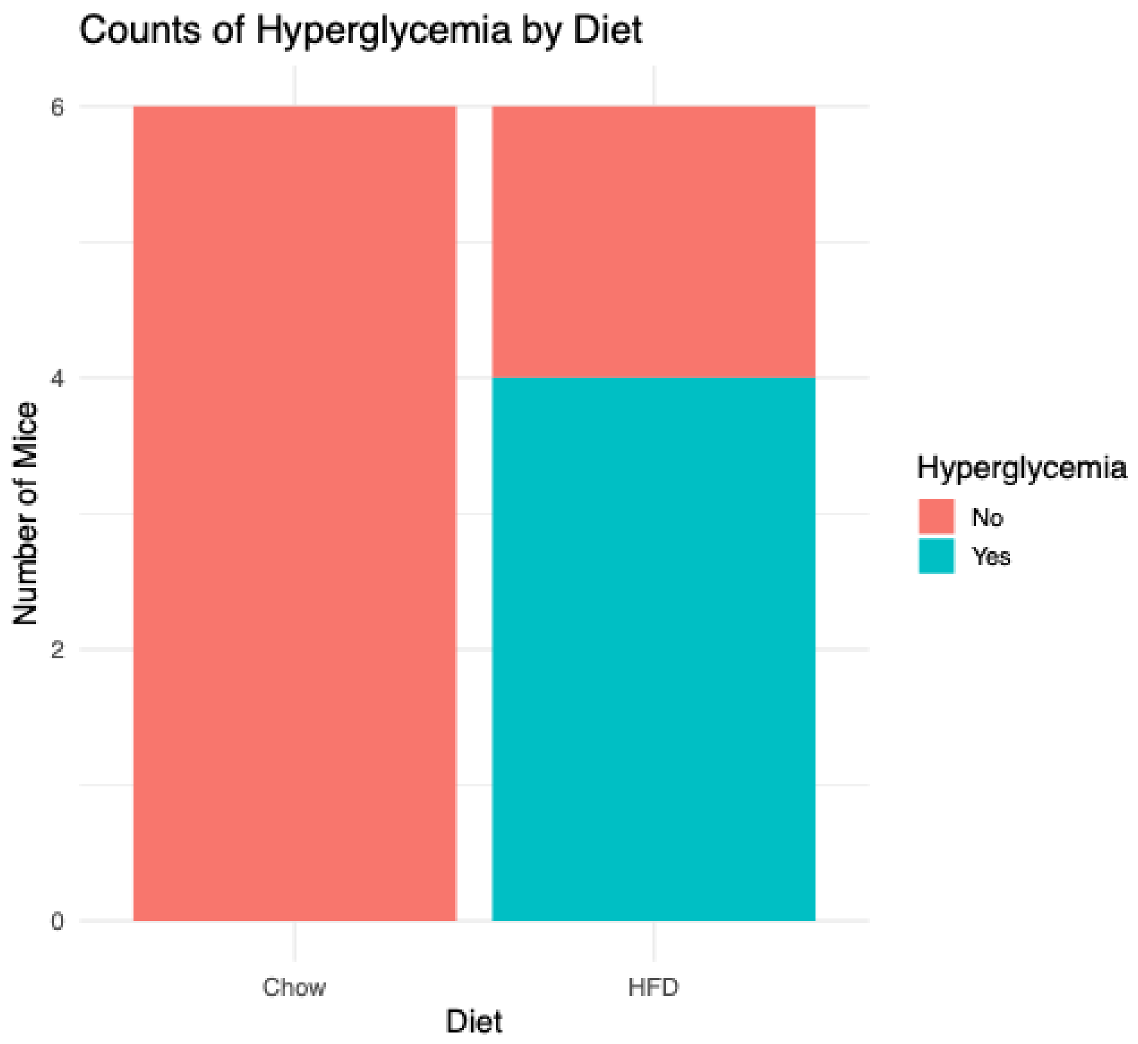




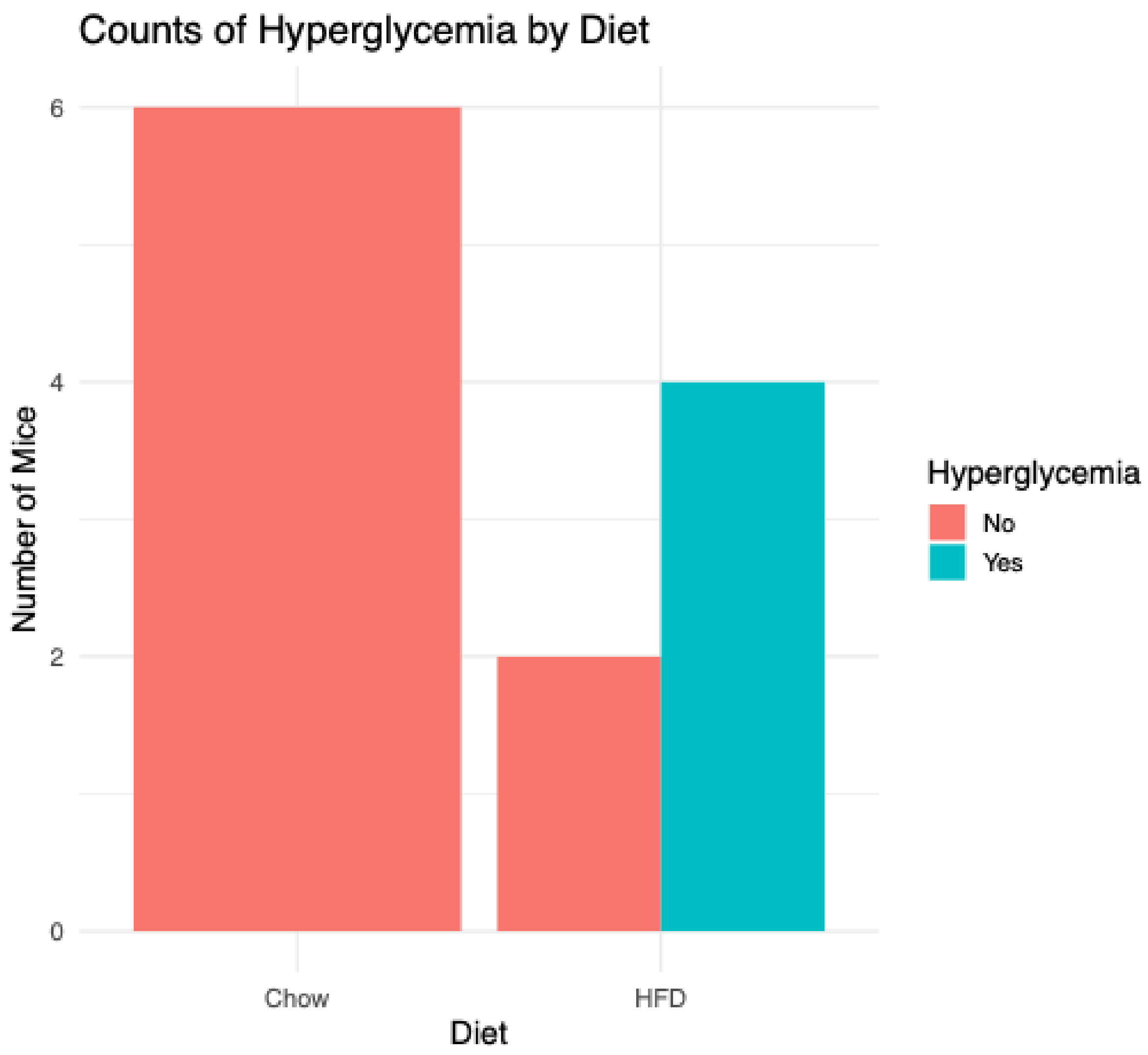
Mosaic Plot


Mosaic Plot: Diet vs Hyperglycemia
Chow
HFD
Hyperglycemia
No
Yes
Diet

Bar Plot

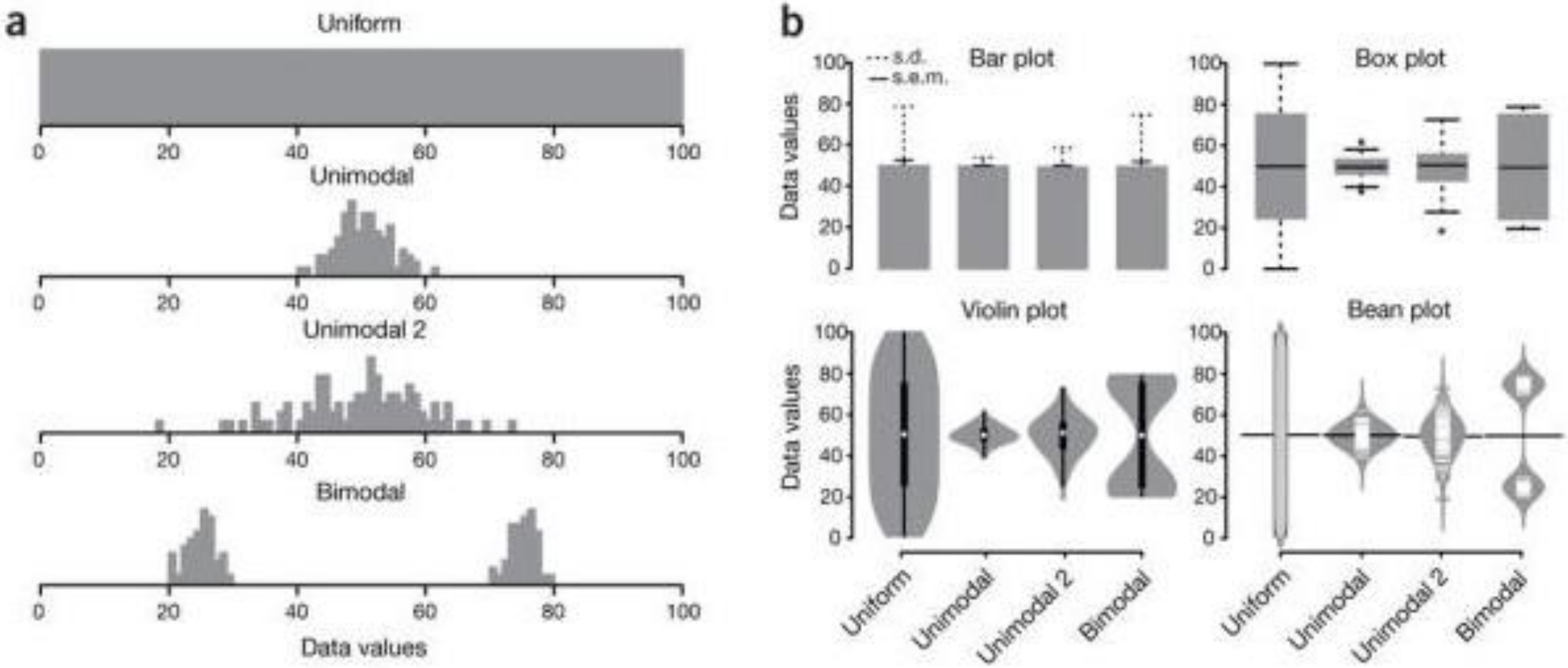


Counts of Hyperglycemia by Diet



Counts of Hyperglycemia by Diet

Boxplots & Violin plots



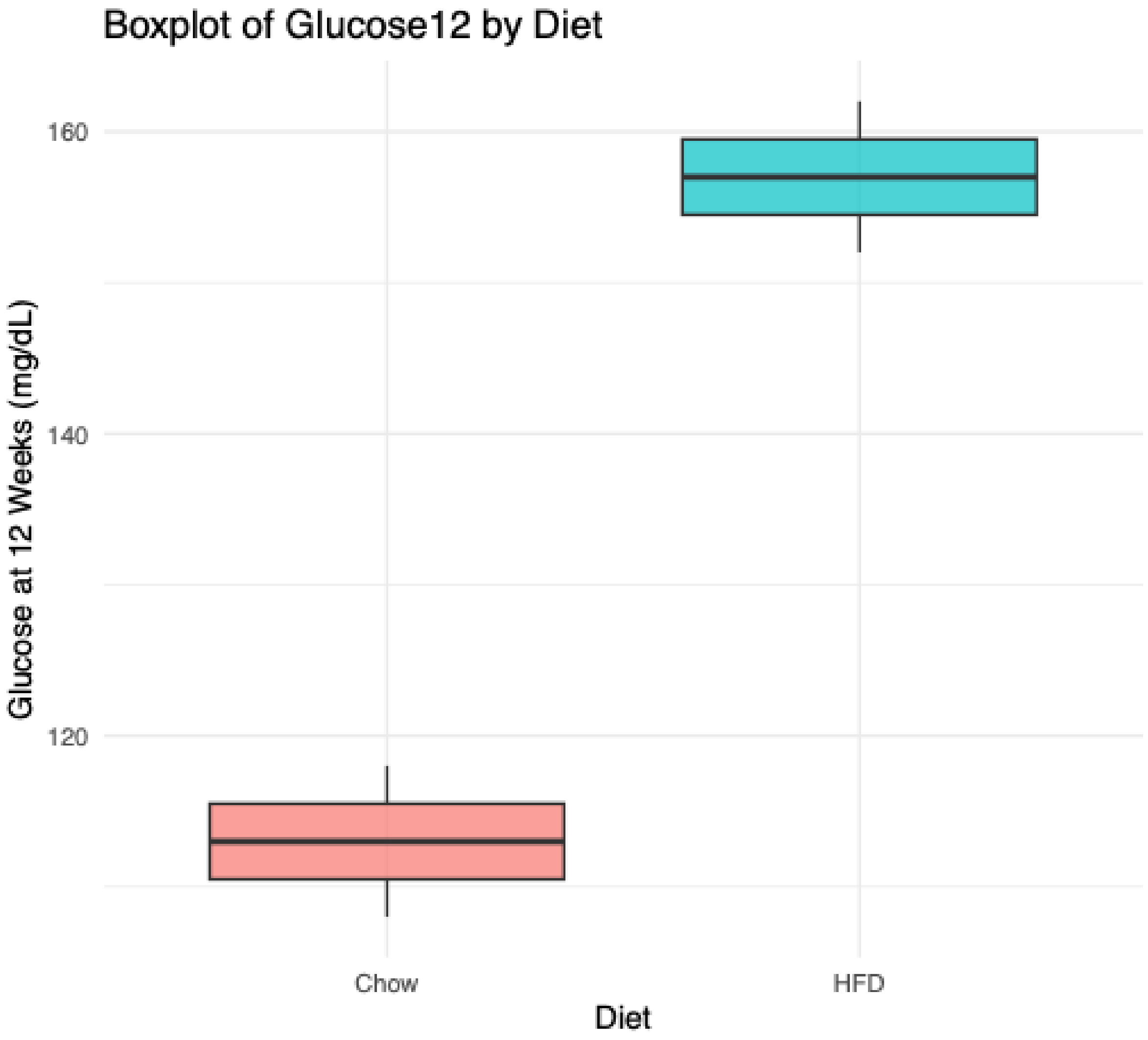
Data visualization with box plots(a)

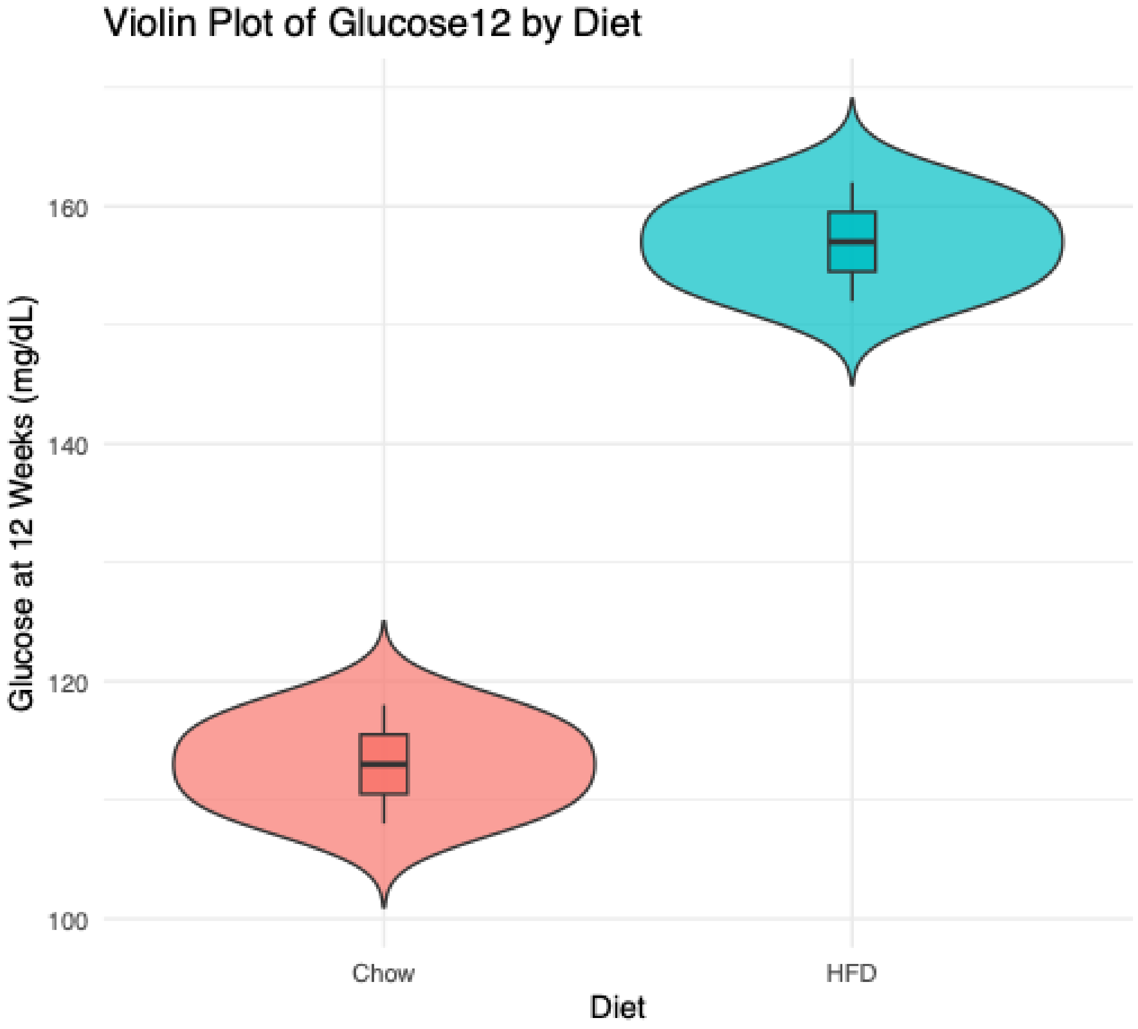
Hypothetical sample data sets of 100 data points each that are uniform, unimodal with one of two different variances or bimodal. Simple bar plot representations and statistical parameters may obscure such different data distributions.

(b) Comparison of data visualization methods. Bar plots typically represent only the mean and s.d. or s.e.m. Box plots visualize the five-number summary of a data set (minimum, lower quartile, median, upper quartile and maximum). Violin and bean plots represent the actual distribution of the individual data sets.

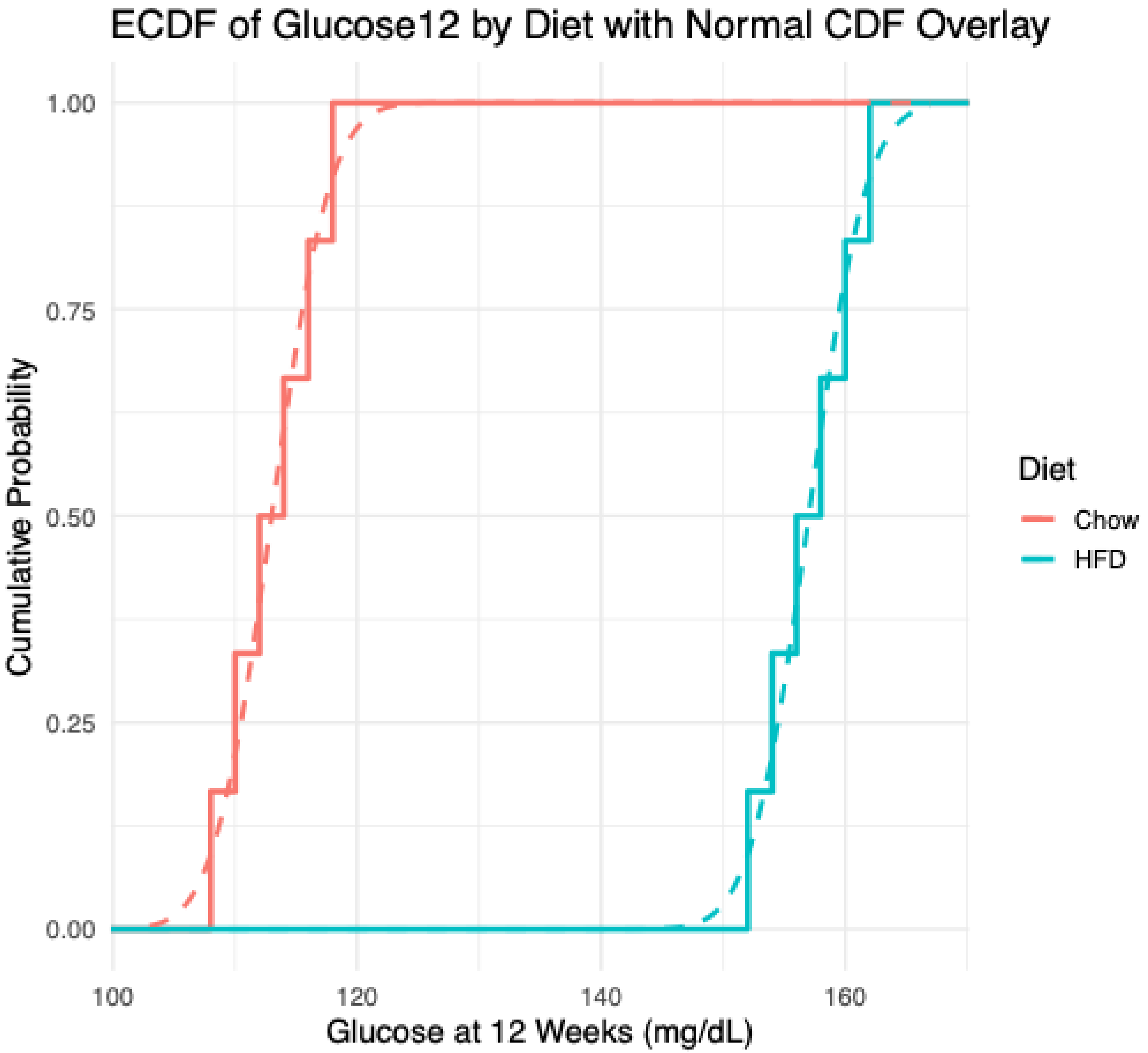
https://pubmed.ncbi.nlm.nih.gov/24481215/

Boxplots & Violin plots





Cumulative Frequency Distribution



Summary

1. The appropriate visualization will depend on the type of variable(s) you are graphing

# variables	Variable Type	Recommended Plots	Use Case
1 (univariate)	Categorical	Bar Chart~~, Pie Chart~~	Comparing category frequencies
	Numerical	Histogram, Boxplot, Density Plot	Understanding distributions
2 (Bivariate)	Categorical & Categorical	Grouped Bar Chart, Mosaic Plot	Comparing proportions of two groups
	Numerical & Categorical	Boxplot, Violin Plot, Strip Plot	Comparing distributions across categories
	Numerical & Numerical	Scatter Plot, Line Plot, Hexbin Plot	Examining relationships or trends
3+ (Multivariate)	Multiple Categorical	Stacked Bar Chart	Analyzing categorical interactions
	Multiple Numerical	Scatterplot Matrix	Comparing multiple numeric relationships
	Mixed	Faceted Plots, Heatmap, Bubble Chart	Visualizing mixed data relationships

2. Everything else is (mostly) artistry and **being clear** in what you are revealing to your audience (See: Edward Tufte for “rules”)