Data Visualization with R

"The greatest value of a picture is when it forces us to notice what we never expected to see."-John Turkey, founder of EDA.

What is DataViz?

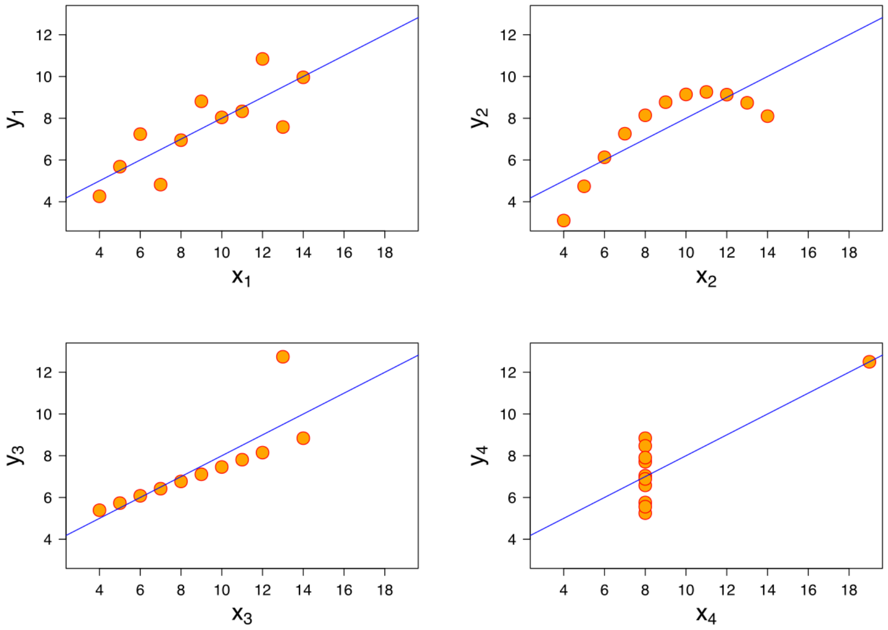
- It is statistic and design combined in meaningful and appropriate ways.
- A form of graphical data analysis emphasising accurate representation and interpretation of data, while also relying on good design principles to not only make the plots attractive but meaningful as well.
- Visualisation is a creative process that involves some amount of trial and error.

Exploratory vs Explanatory Plotting

It is important to understand the distinction between exploratory and explanatory plots

- Exploratory visualisations: are easily generated, data-heavy and intended for a small specialist audience, (you and your collegaues), primary focus is on graphical data analysis (plots should be meaningful).
- Explanatory visualisation are labor intensive, data specific, intended for a broader audience (publications and presentations).

Importance of Visualisation



Grammar of graphics

The graphics are built upon an underlying grammar, that can help us think creatively about data visualisation.The grammar of graphics is a plotting framework developed by Leland Wilkilson, published in 1999

Two underlying principles of grammar of graphics:

- Graphics are made of distinct layers of grammatical elements
- Meaningful plots are built around aesthetic mappings

Essential grammatical elements

Element	Description
Data	The dataset being plotted
Aesthetics	The scales onto which we ma p our data
Geometrics	The visual elements used for our data
Facets	Plotting small multiples
Statistics	Representations of our data t o aid understanding
Coordinates	The space on which the data will be plotted
Themes	All non-data ink

Base Package vs ggplot2

Ggplot2 is a very flexible way of making complex plots. Base package is good for straightforward plots, the syntax is mostly straightforward for univariate and bivariate data.

Some limitations of using base package:

- Plot doesn’t get redrawn: If we want to match with a new data, it can dangerous as

information can be lost in this way.
- Plot is drawn as an image: The plot is not an image that we can manipulate ones it's made.
- We need to add legend manually which is a potential entry point for errors.
- No unified framework for plotting - there are different functions for different types of plots

Data Layer

The structure of your data will dictate how you construct plots in ggplot2. Making your data conform to a structure that matches the plot in mind will make the task of visualization much easier through several R data visualization examples.

Lets discuss about two distinct types of variables that will help us understand our data:

- Variables that defined by a small number of groups are called categorical data.
    - Sub-grouped to ordinal and non-ordinal. Example Gender: male or female, Spiciness: mild, medium, hot
- Variables that can take any value if measured with enough precision.
    - Sub-grouped to continuous and discrete. Example: pair of twins with heights 68.12 inches and 68.11 inches respectively, population sizes are discrete have to be in round numbers.
    - Discrete numeric data can be considered ordinal, this usually can happen when there is a small set of groups with each group having a lot of members.

Understanding distributions

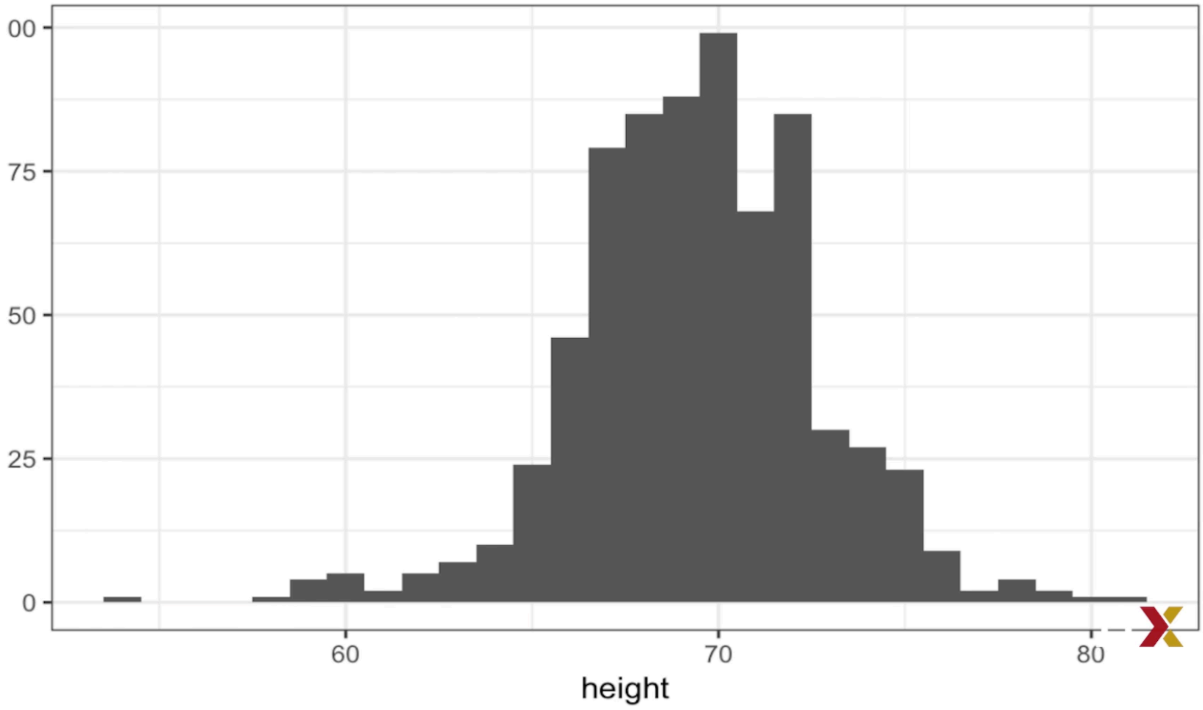
- One of the useful outputs of data visualisation is that we can learn distributions of variables.
- Our first data visualization building block is learning to summarize lists of factors or numeric vectors.
- The most basic statistical summary of a list of objects or numbers is its distribution.
- To learn the distribution of categorical data we calculate the frequency counts of each unique value.
- When data is not categorical, reporting the frequency of each unique entry is not an effective summary since most entries are unique.
- We use CDF, eCDF, histograms or normal distributions.

Lets check out some examples now

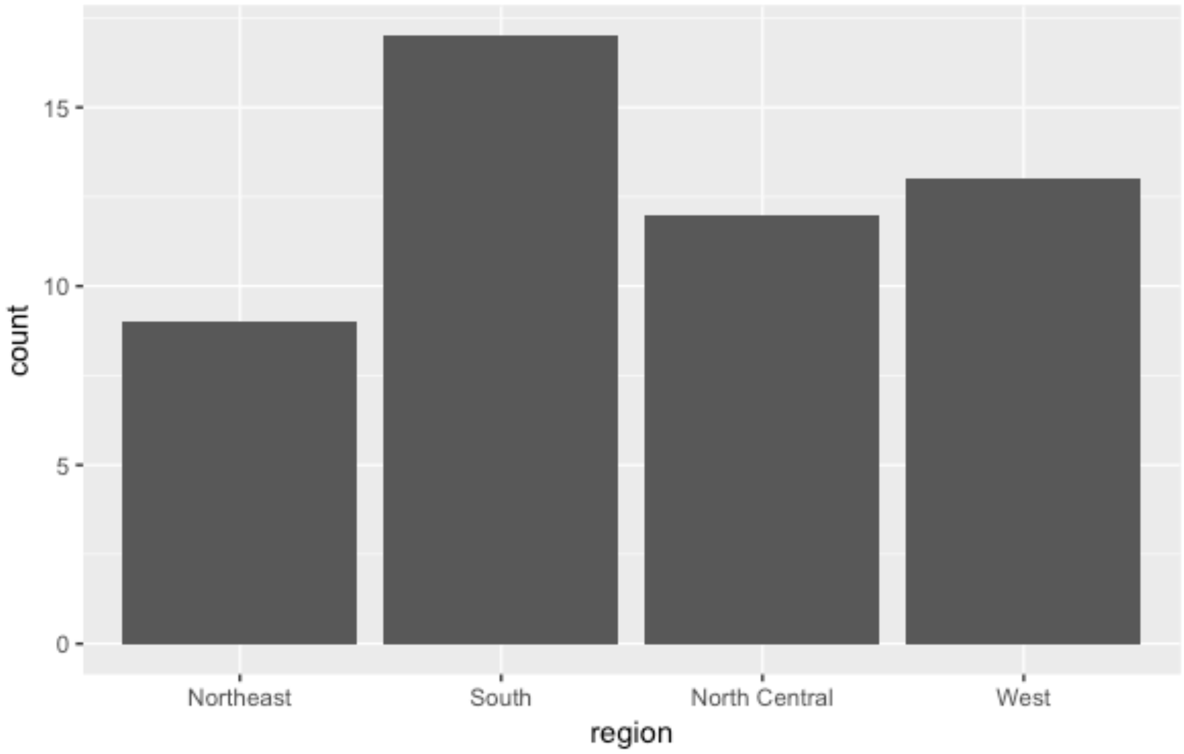
Histograms

- Histograms are much preferred as they sacrifice only a bit of information
- They demonstrate the central tendency of a distribution more intuitively
- They produce plots that are easy to interpret.
- They are plotted by dividing a data into non overlapping bins of the same size.
- Each bin giving the count of values falling in that interval.
- The histogram plots these counts as bars with the base of the bar, the interval.
- The information we lose are the values that can't be distinguished as they are part of the bin with the same interval.

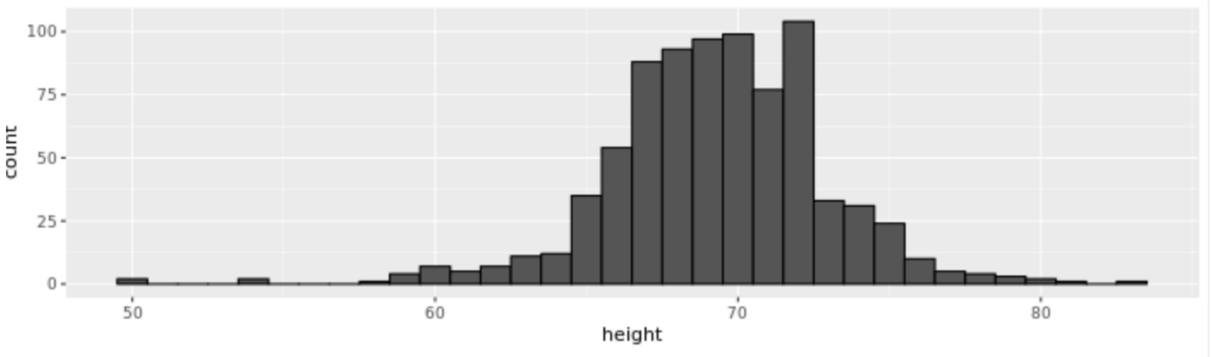
The heights are close to symmetrical around 69 inches.



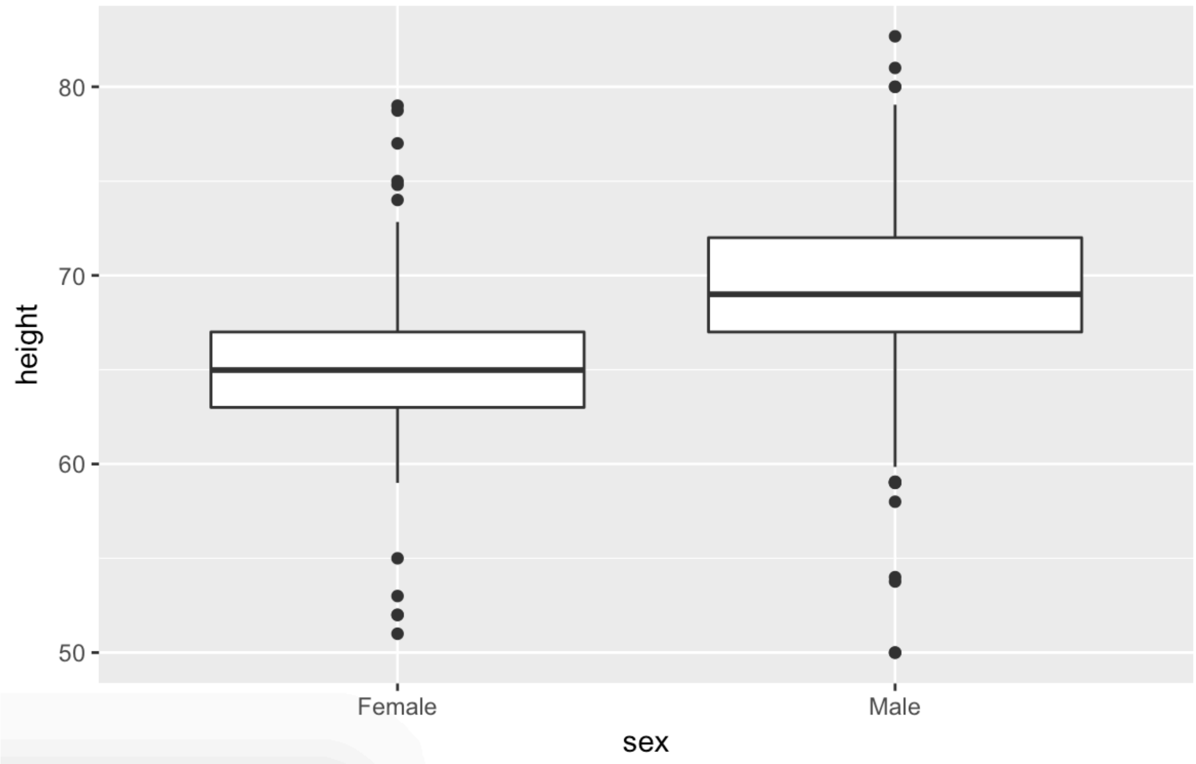
Is this a histogram?



Based on this plot, how many males are between 62.5 and 65.5?



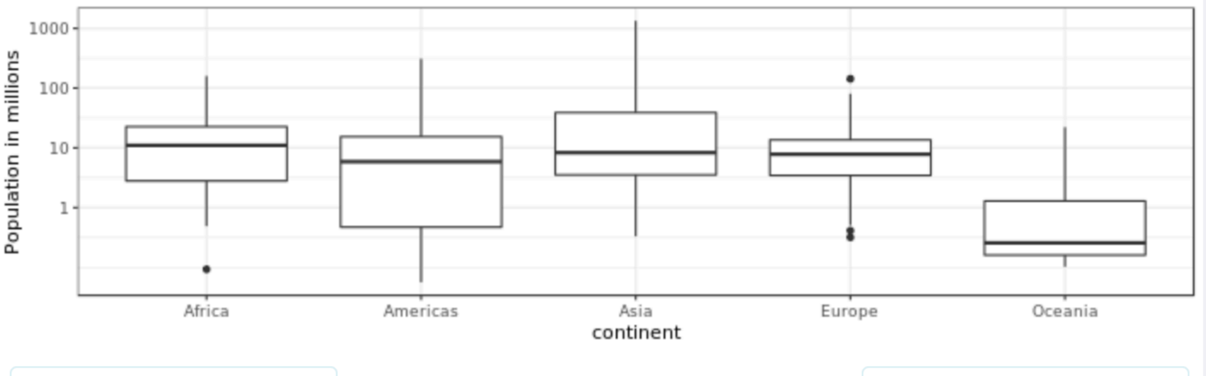
Boxplots are used for comparing two or more distributions, which has greater size of people?



Box-plotting across more than two categories

Lets answer a few questions based on the above diagram

1. Which continent has population of the largest size?
2. Which continent has median country with the largest population?
3. Median of Africa?
4. What proportion of countries in Europe have populations below 14 million: 0.75
5. Largest interquartile range?



A hint of Exploratory Data Analysis

- Generate questions about your data.

- Search for answers by visualising, transforming and modelling your data.
- Use what you learn to refine your questions and/or generate new questions.

Lets start coding !