POLS0008
INTRODUCTION TO QUANTITATIVE RESEARCH METHODS

EXAMINING DATA (PART II)

Dr Anwar Musah (a.musah@ucl.ac.uk)
Lecturer in Social and Geographic Data Science
UCL Geography

QUICK RECAP OF WEEK 2

Summary measures

What kinds of analysis and summary statistics can you perform on a particular type of dataset?

1. Categorical Data

You can group the data by according to categories and perform the following:

- Compute the Frequencies (counts)
- Compute the Percentages (or Relative Frequency)
- Calculate the Cumulative Frequencies or Cumulative Percentages
- Graphical approaches also include bar plots and pie charts
- The Mode (category with that occurs most)

2. Numerical Data

You can perform the following analysis:

- Compute the mode (value that occur most)
- Compute the median
- Compute the mean
- Lowest (Minimum) & Highest (Maximum)
- Percentiles
- Variance
- Standard deviation
- Range
- Quartiles and Interquartile ranges

Calculation of these summary measures

Summary measure	Type	Formula
Mean	Central Tendency	$\bar{x} = \frac{\sum x_i}{n}$
Median	Central Tendency	$\frac{n+1}{2}$
Lower Quartile (25%)	Range value	$\frac{n+1}{4}$
Upper Quartile (75%)	Range value	$\frac{3(n+1)}{4}$
Interquartile Range	Derived range value from Q1 & Q3	
Range	Derived range value from min and max	
Variance	Dispersion measure	$\frac{\sum(x_i - \bar{x})^2}{n-1}$
Standard Deviation	Derived dispersion measure from variance	$\sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$

Data Visualisation

What is data visualisation?

- Data visualisation gives us a clear idea of what information means by giving it context through maps, infographics, statistical charts and many more

- Data visualisation makes quantitative (or qualitative) data more natural for the human mind to comprehend through pictorial means

- Data visualisation makes it easier to identify trends, patterns and outliers within large data sets

Why is it important?

- It's a rapid and efficient approach for summarising raw data in the most efficient way

- The most important thing is communication – it uses simplified visuals of raw data that's "crunched" by models to communicate findings that are intuitive & effective

- Visual outputs can serve help academics, stakeholders, policy makers etc., for decision making as well as prediction for outcomes.

Benefits of data visualisation?

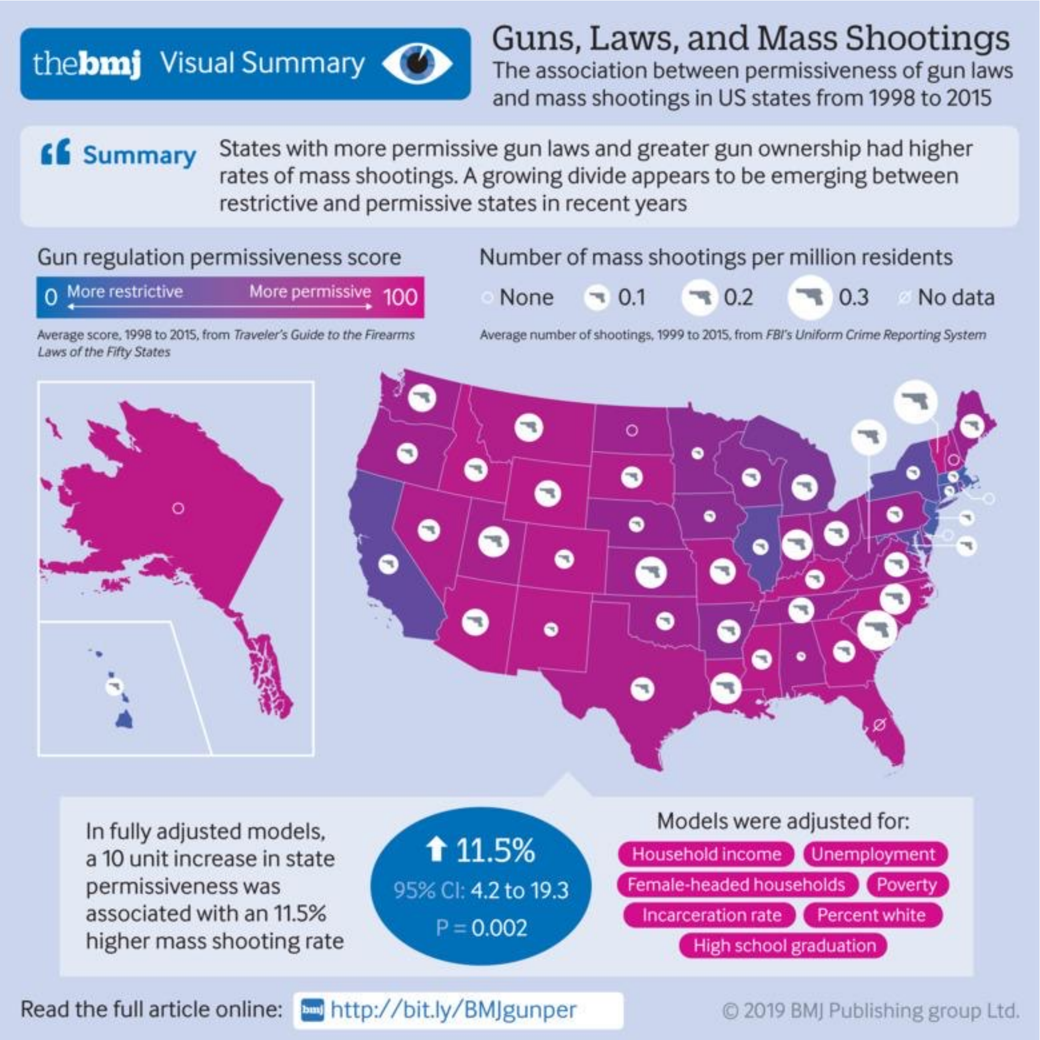
- To support evidence-based research in establishing correlations in relationships
- To demonstrate trends of time which, turn in, is used for making predictions (aka forecast) ahead in time
- To show frequency of events represented in pictorial form, and how data is centred around a value and how its spread out around that value as well as over an interval etc.

IMPORTANT NOTE: The third bullet uses very useful visualisation techniques to display how data are distributed

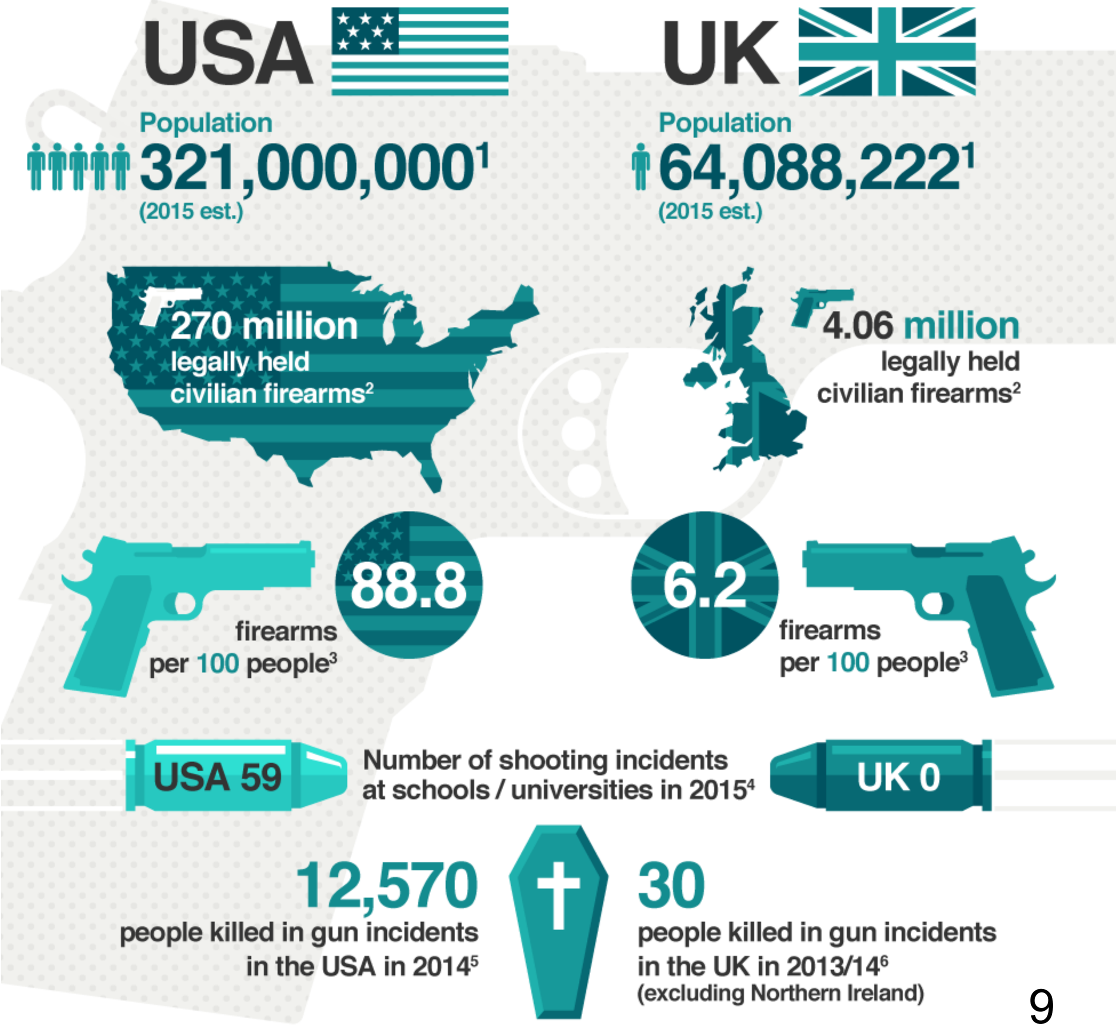
General techniques for data visualisation [1]

Infographics

1.

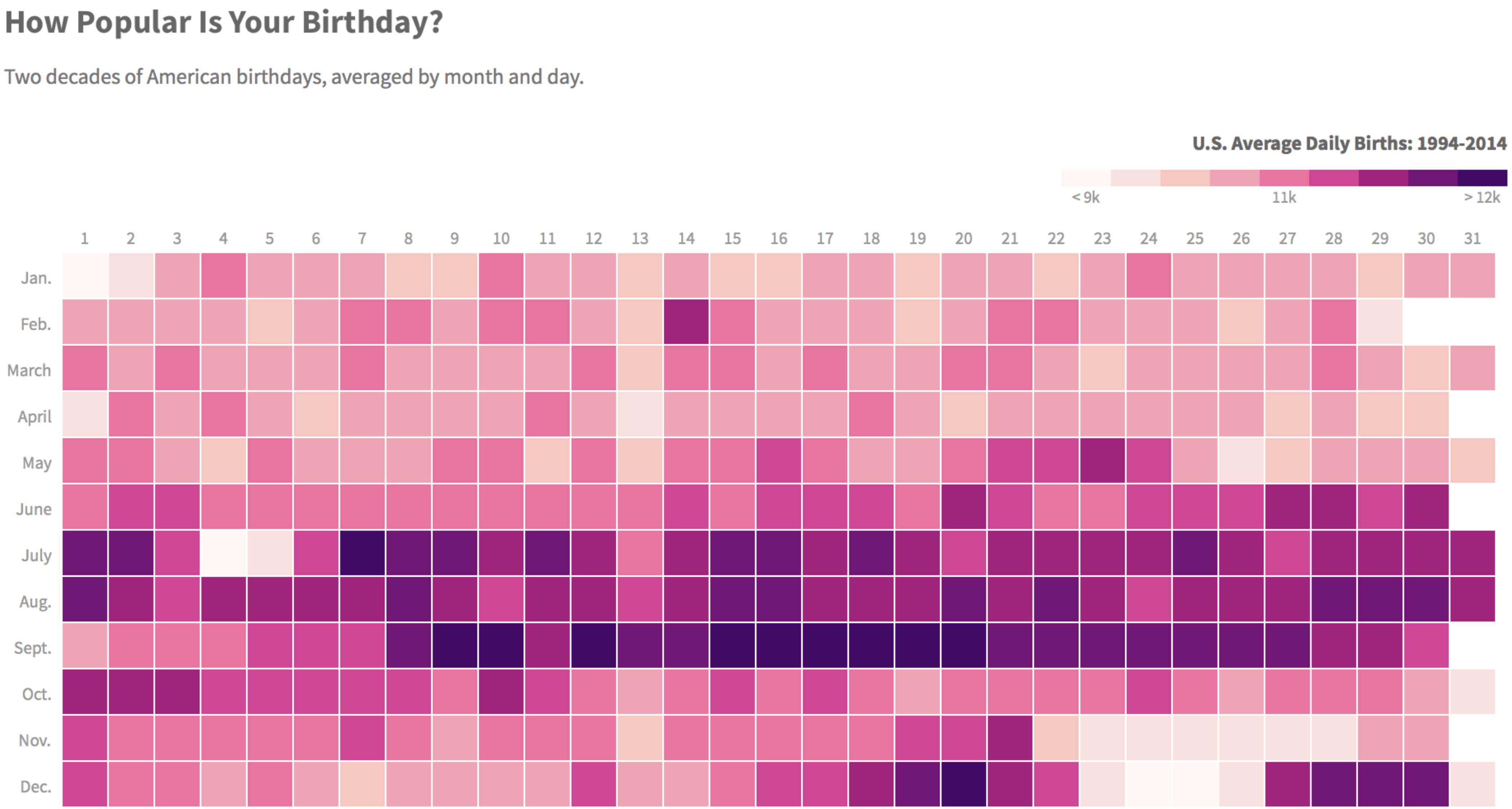


2.



General techniques for data visualisation [2]

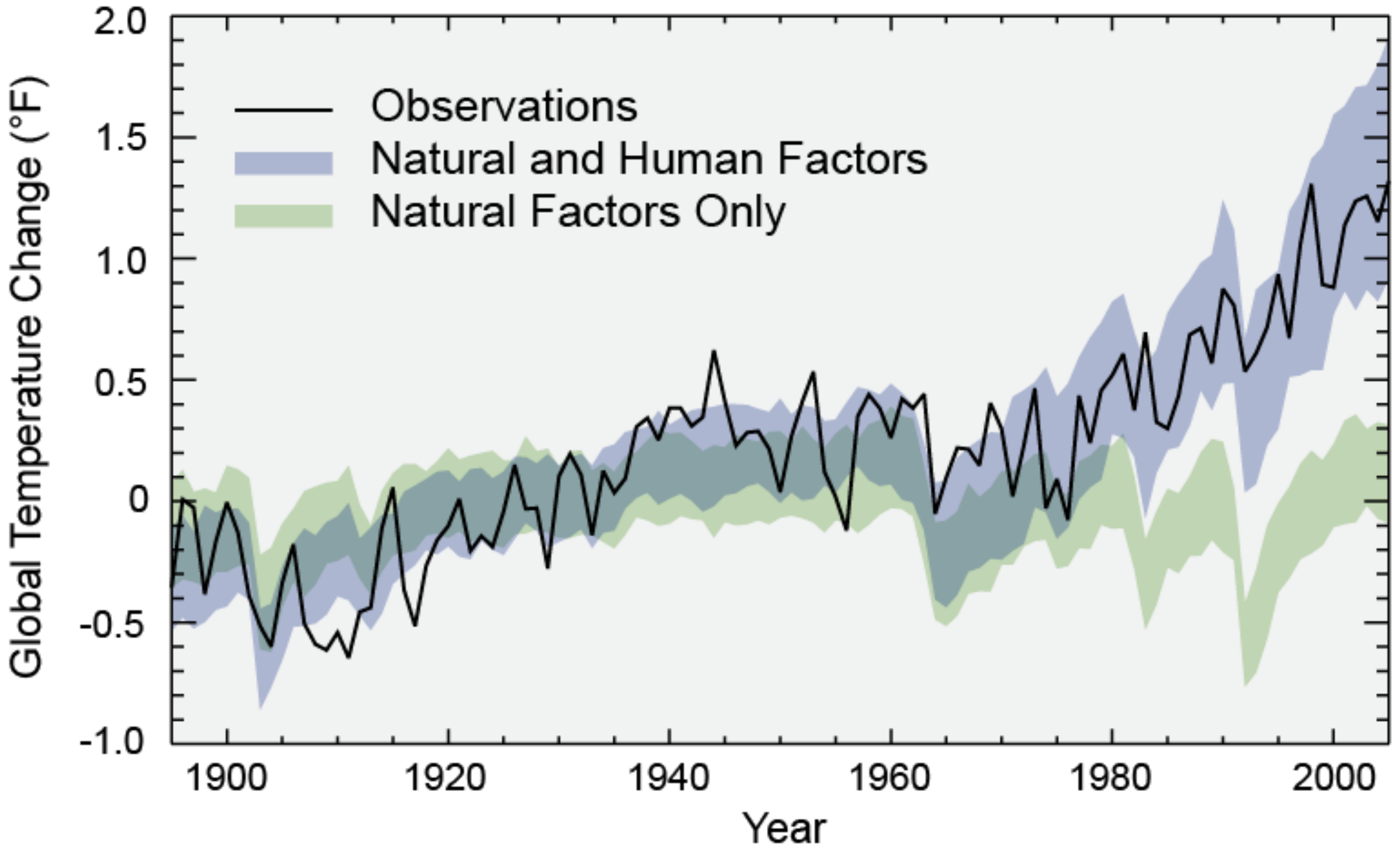
Heatmap visualisation



Source: https://thedailyviz.com/2016/09/17/how-common-is-your-birthday-dailyviz/

General techniques for data visualisation [3]

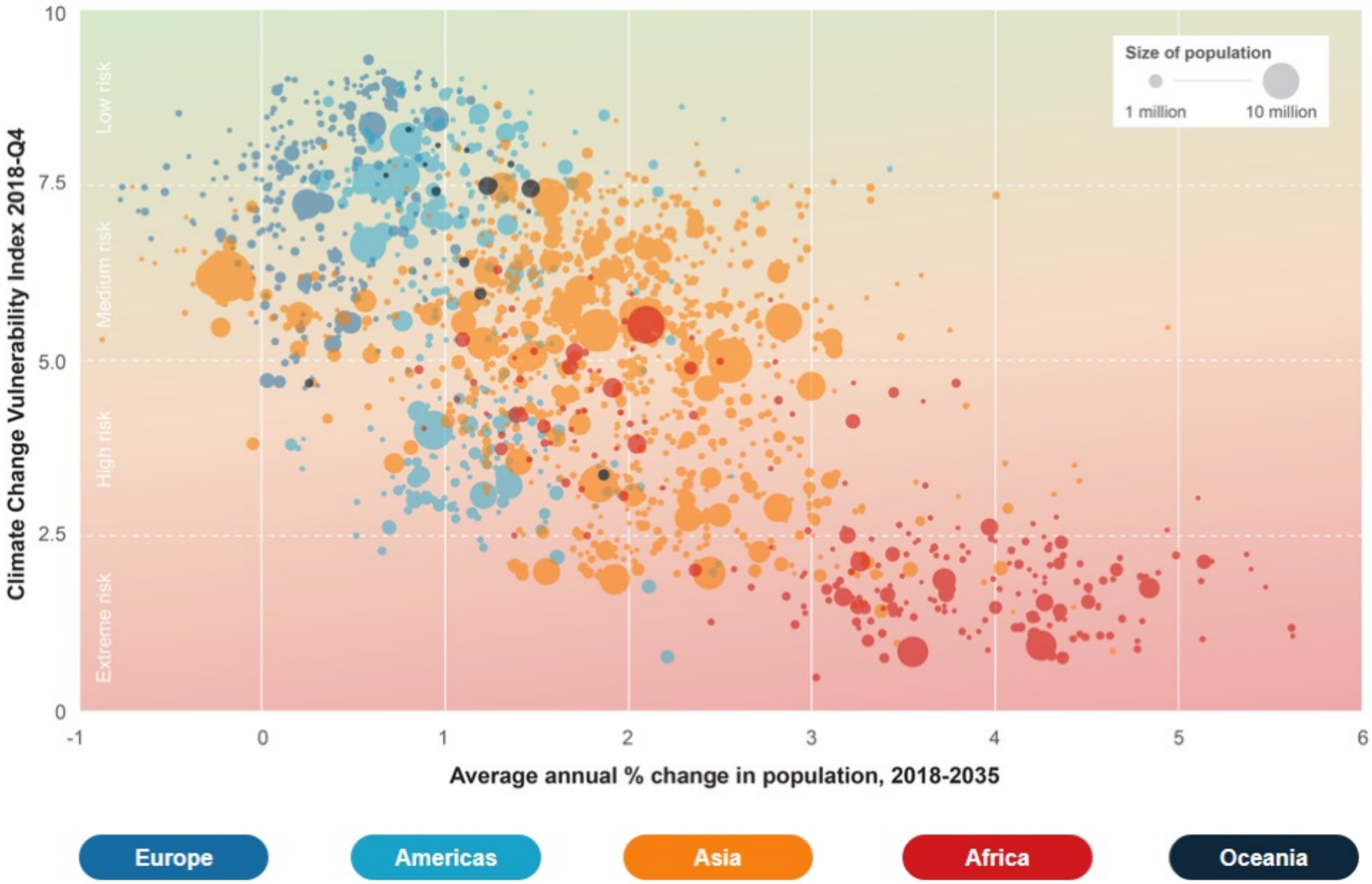
Trend-based visualisation



General techniques for data visualisation [4]

Bi-, or multi-variable visualisation

Source: [LINK]



General techniques for data visualisation [5]

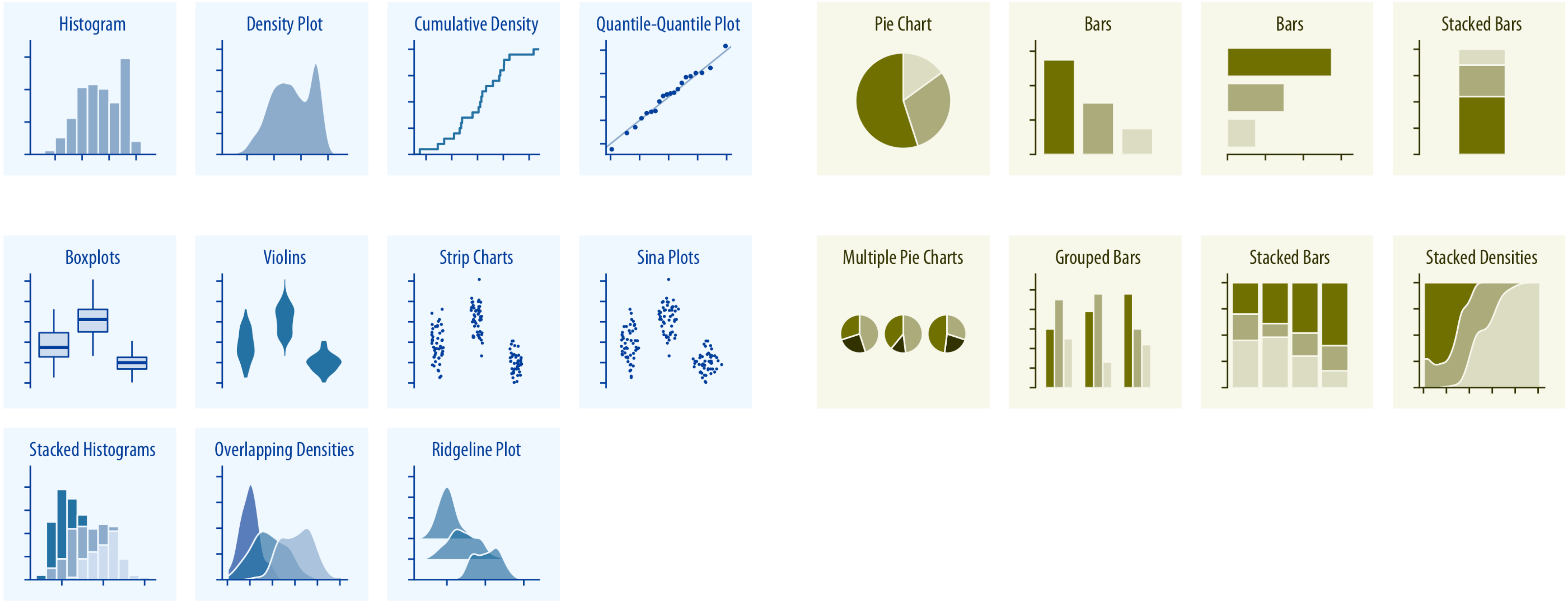
Visualisations representing densities & distributions of numerical data

1. Plots for densities and distributions (numerical data)

2. Plots for proportions (qualitative or categorical data)

Single variable

Multiple variables



Source: Fundamentals of Data Visualization [LINK]

Graphical representation of distributions

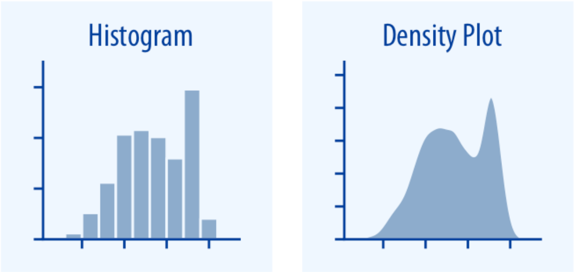
Histograms and/or **Density** plots are **BEST** for this type of visualization of numeric data

- Histogram is the most commonly used graph to show the frequency of observations (or data points)

- These observations (or data points) are counted in user specified ranges (aka groups) called “bins”

- This bins are stacked adjacently to each other along the x-axis to visualise the frequency distribution which read from the y-axis.

- Don’t confuse this with a bar plot!



Graphical representation of distributions

Example: Monitoring a new born puppy's growth within the 1st 4-weeks. These are readings of the *changes* in its weight (kg)

Step 1: Data set: -0.2,-0.4, 0, 0.1, 0.1, 0.3, 0.12, 0.4, 0.5, 0.8, 0.9, 0.5, 0.6, 0.7, 0.6, 1.3, -0.2, 0, 0.1, 0.1, 0.3, 0.4, 0.5, 0.5, 0.6, 1.2

Data set (sorted): -0.40, -0.20, -0.20, 0.00, 0.00, 0.10, 0.10, 0.10, 0.10, 0.12, 0.30, 0.30, 0.40, 0.40, 0.50, 0.50, 0.50, 0.50, 0.60, 0.60, 0.60, 0.70, 0.80, 0.90, 1.20, 1.30

Step 2: Create set of user-specified ranges (or bins) to group the values along the x-axis

- Bins are based on interval of 0.25 starting from -0.5 to 1.5 (-0.5. -0.25, 0, 0.25, 0.5, 0.75, 1, 1.25 and 1.5)

Step 3: Count the observed values that fall within each of its corresponding interval and plot as a histogram

Graphical representation of distributions

Example: Monitoring a puppy's growth within the 1st 4-weeks. These are readings of the *changes* in its weight (kg)

Step 1: Data set: -0.2,-0.4, 0, 0.1, 0.1, 0.3, 0.12, 0.4, 0.5, 0.8, 0.9, 0.5, 0.6, 0.7, 0.6, 1.3, -0.2, 0, 0.1, 0.1, 0.3, 0.4, 0.5, 0.5, 0.6, 1.2

Data set (sorted): -0.40, -0.20, -0.20, 0.00, 0.00, 0.10, 0.10, 0.10, 0.10, 0.12, 0.30, 0.30, 0.40, 0.40, 0.50, 0.50, 0.50, 0.50, 0.60, 0.60, 0.60, 0.70, 0.80, 0.90, 1.20, 1.30

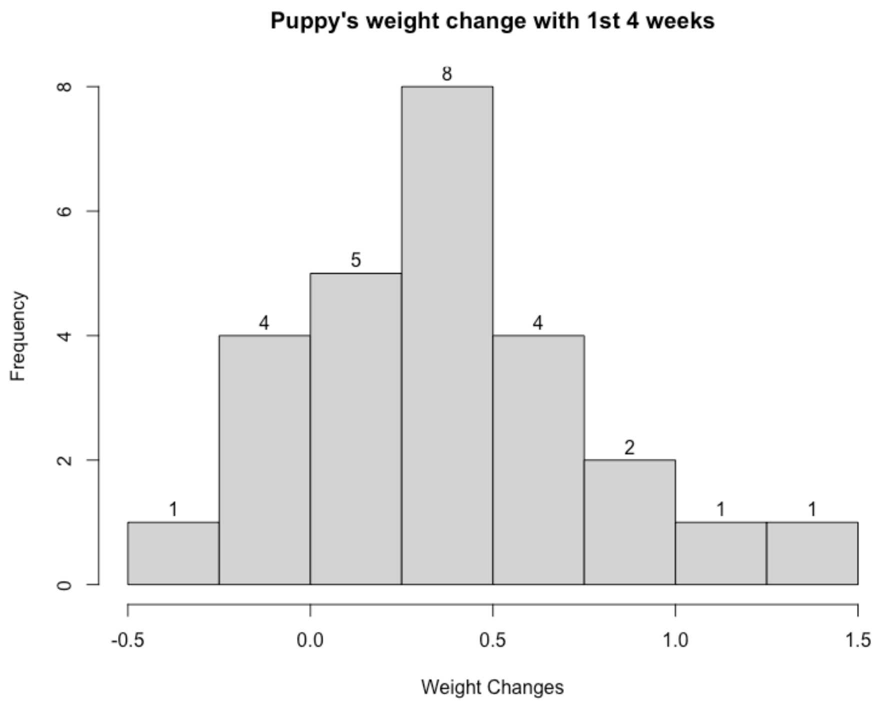
Step 2: Create set of user-specified ranges (or bins) to group the values along the x-axis

- Bins are based on interval of 0.25 starting from -0.5 to 1.5 (-0.5. -0.25, 0, 0.25, 0.5, 0.75, 1, 1.25 and 1.5)

Step 3: Count the observed values that fall within each of its corresponding interval and plot as a histogram

Code:

```
weightChanges <- c(-0.2,-0.4, 0, 0.1, 0.1, 0.3, 0.12, 0.4, 0.5, 0.8, 0.9, 0.5, 0.6, 0.7, 0.6, 1.3, -0.2, 0, 0.1, 0.1, 0.3, 0.4, 0.5, 0.5, 0.6, 1.2)

bins <- c(-0.5, -0.25, 0, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5)

hist(weightChanges, breaks = bins, main="Puppy's weight change with 1st 4 weeks", xlab = "Weight Changes", labels = TRUE)
```



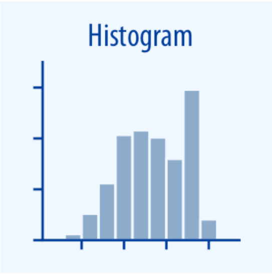
Graphical representation of distributions

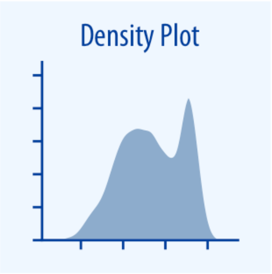
Histograms and/or Density plots are BEST for this type of visualization of numeric data

- Density plots shows the same thing as a histogram i.e., frequency of observations (or data points)

- However, these observations (or data points) are counted on a continuous interval

- Basically, it's a smoothed curve, and because its plotted over on a continuous scale, we read the density as a proportion or percentage instead of counts.





Graphical representation of distributions

Example: Monitoring a puppy's growth within the 1st 4-weeks. These are readings of the *changes* in its weight (kg)

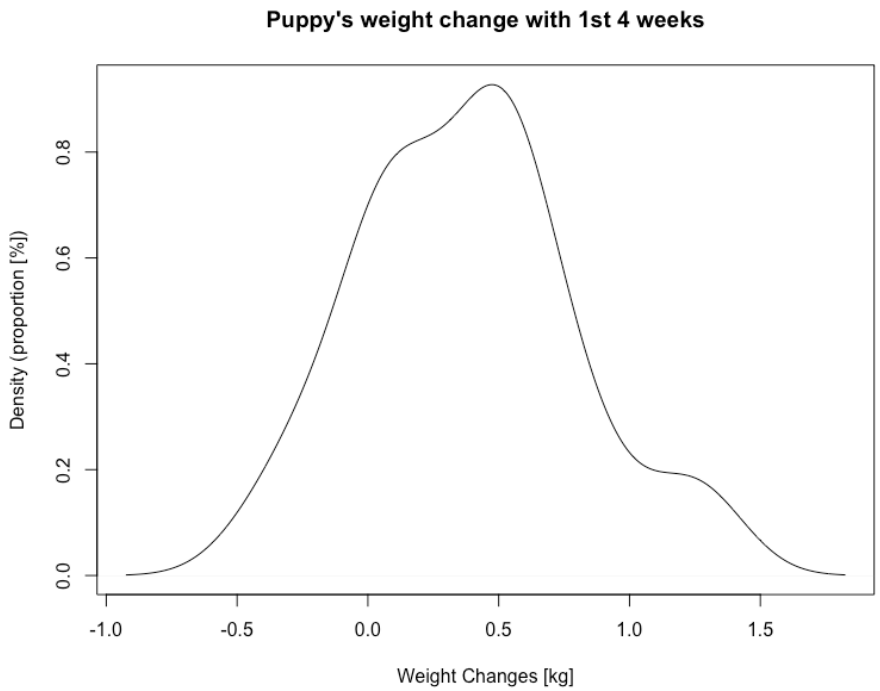
Step 1: Data set: -0.2,-0.4, 0, 0.1, 0.1, 0.3, 0.12, 0.4, 0.5, 0.8, 0.9, 0.5, 0.6, 0.7, 0.6, 1.3, -0.2, 0, 0.1, 0.1, 0.3, 0.4, 0.5, 0.5, 0.6, 1.2

Data set (sorted): -0.40, -0.20, -0.20, 0.00, 0.00, 0.10, 0.10, 0.10, 0.10, 0.12, 0.30, 0.30, 0.40, 0.40, 0.50, 0.50, 0.50, 0.50, 0.60, 0.60, 0.60, 0.70, 0.80, 0.90, 1.20, 1.30

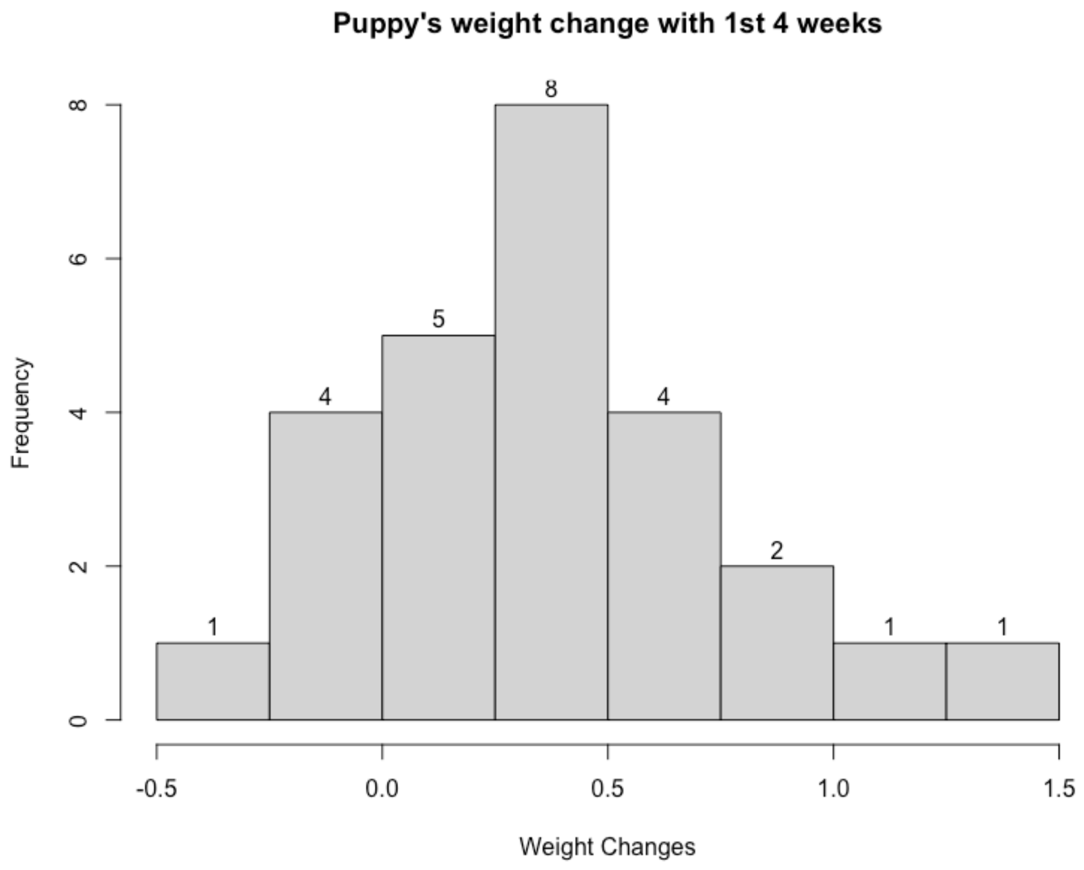
Step 2: To create a density plot is quite easy!

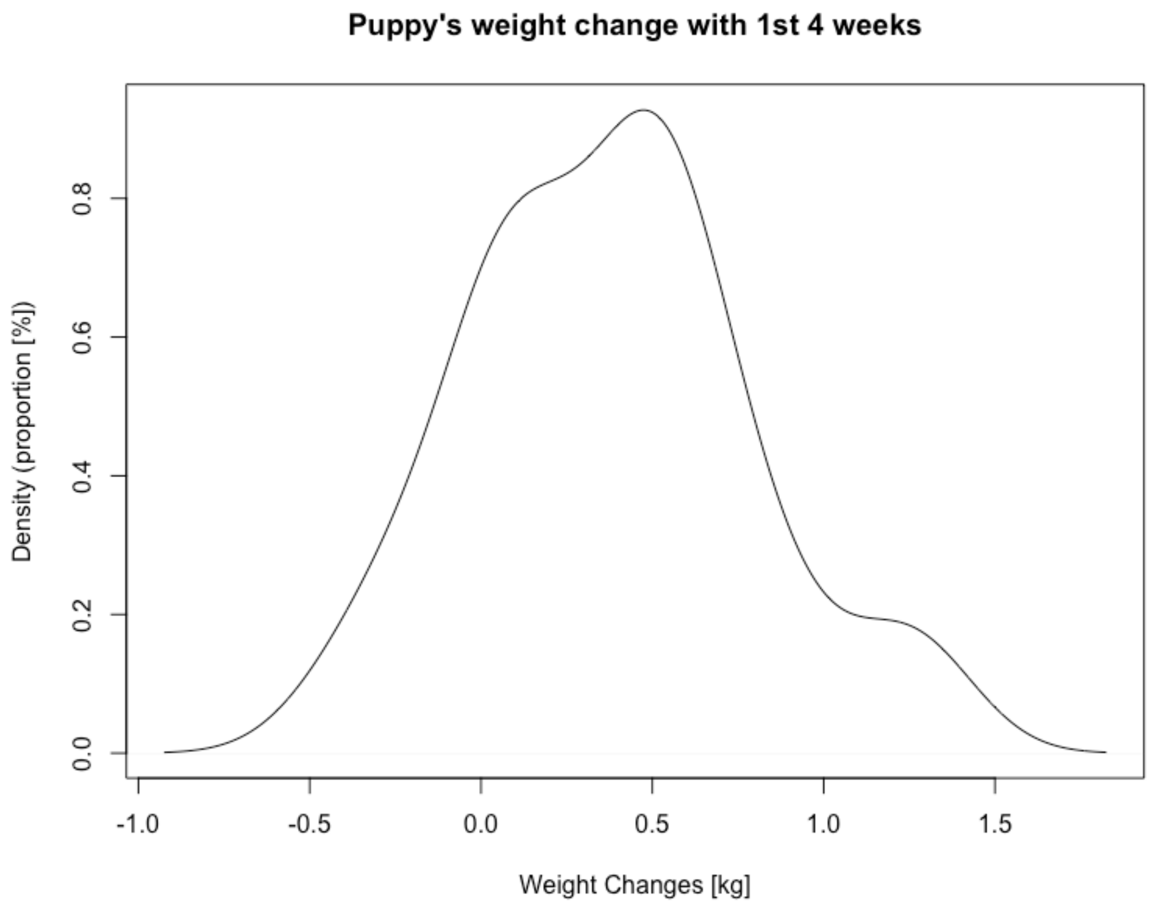
Code:

```
weightChanges <- c(-0.2,-0.4, 0, 0.1, 0.1, 0.3, 0.12, 0.4, 0.5, 0.8, 0.9, 0.5, 0.6, 0.7, 0.6, 1.3, -0.2, 0, 0.1, 0.1, 0.3, 0.4, 0.5, 0.5, 0.6, 1.2)

den <- density(weightChanges)

plot(den, main="Puppy's weight change with 1st 4 weeks", xlab = "Weight Changes [kg]", ylab = "Density (proportion [%])")
```



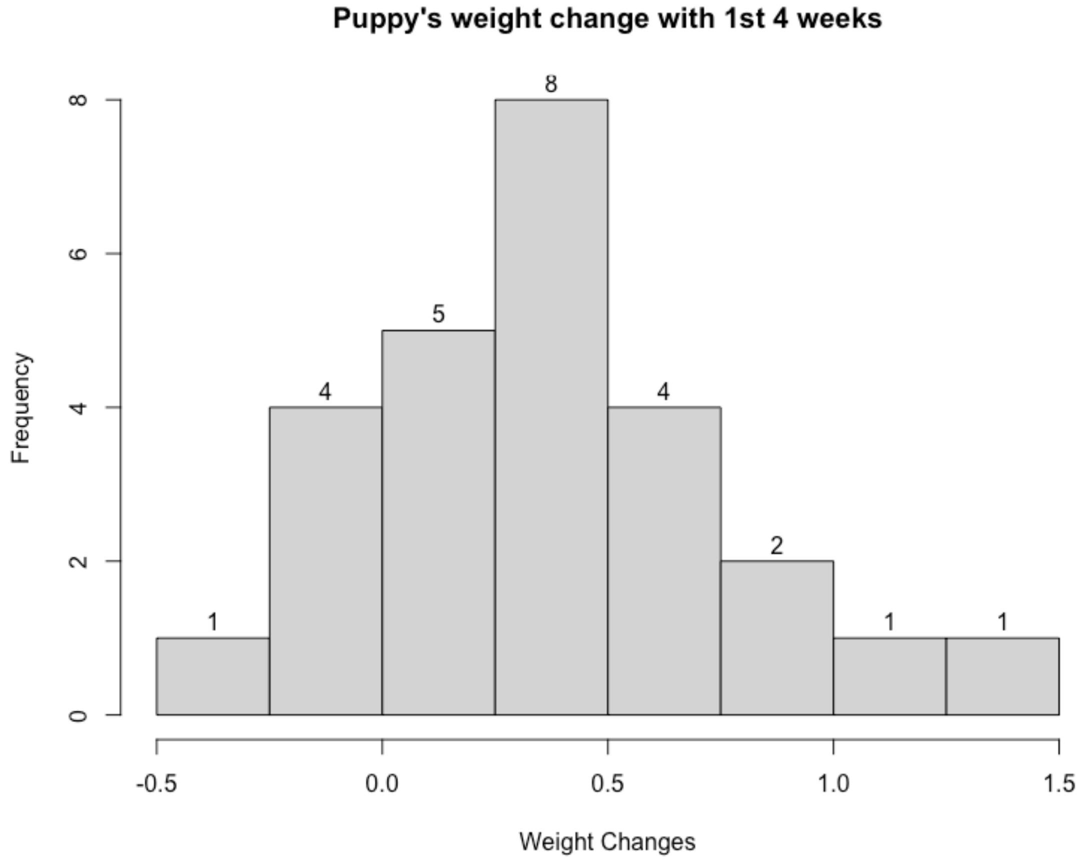
Histogram/Density plots for assessing the distributions of data [1]

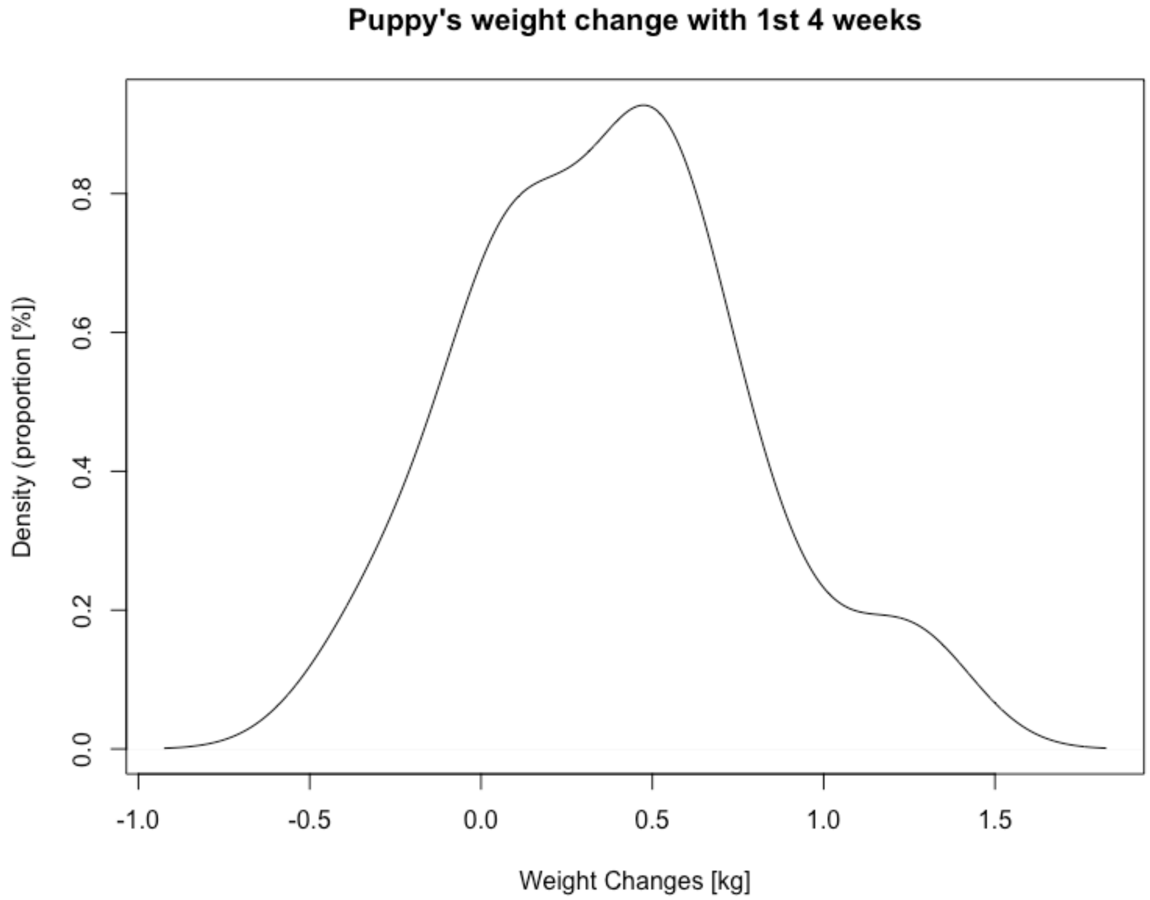




How does this link with summary statistics learnt from week 2?

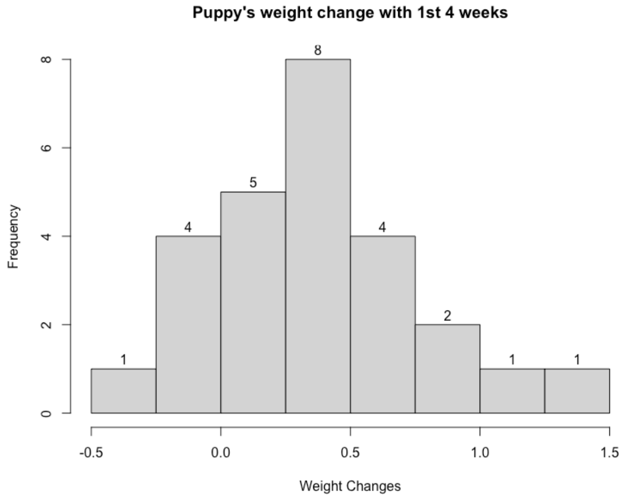
Histogram/Density plots for assessing the distributions of data [2]

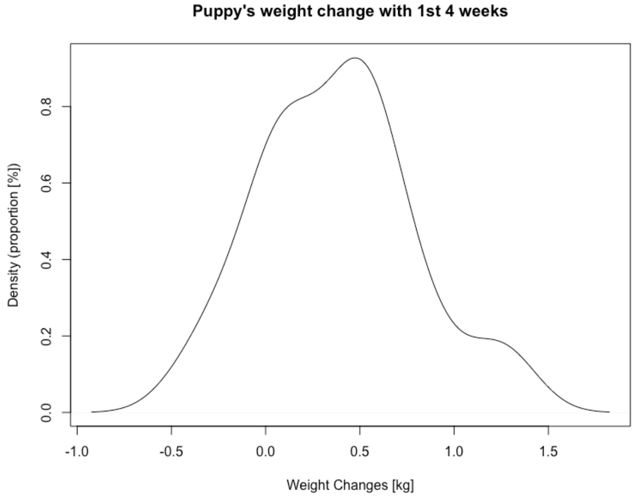




It's a nice way to visually understand the distribution of a continuous variable. It gives you a feel for its central tendency and variability.

Histogram/Density plots for assessing the distributions of data [3]





The central tendencies of the weight measures (see slide 15) are as follows:

- The overall mean weight change was 0.378 kg which means on average within that 8-week period, the puppy's weight increased by 0.378 kg
- The median weight change was 0.4 kg

The dispersion values of the weight measures (see slide 15) are as follows:

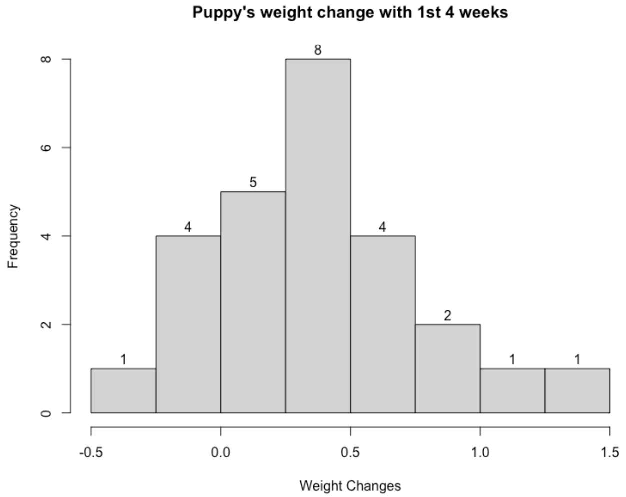
- The overall standard deviation was ±0.412 kg. The change in weight of puppy varied between 0.378 ± 0.412 kg (i.e., -0.034kg to +0.79kg)

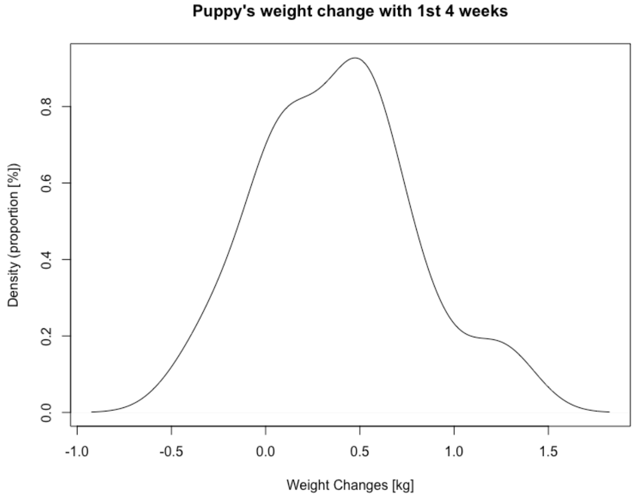
IMPORTANT: Notice how the mean and median lay at the centre of the histogram or density plot.

Also, notice how the shape of the histogram is near symmetric.

This pattern is what we called a **"Normal distribution"** or **"Bell-shaped curved"**

Histogram/Density plots for assessing the distributions of data [4]





The central tendencies of the weight measures (see slide 15) are as follows:

- The overall mean weight change was 0.378 kg which means on average within that 8-week period, the puppy's weight increased by 0.378 kg
- The median weight change was 0.4 kg

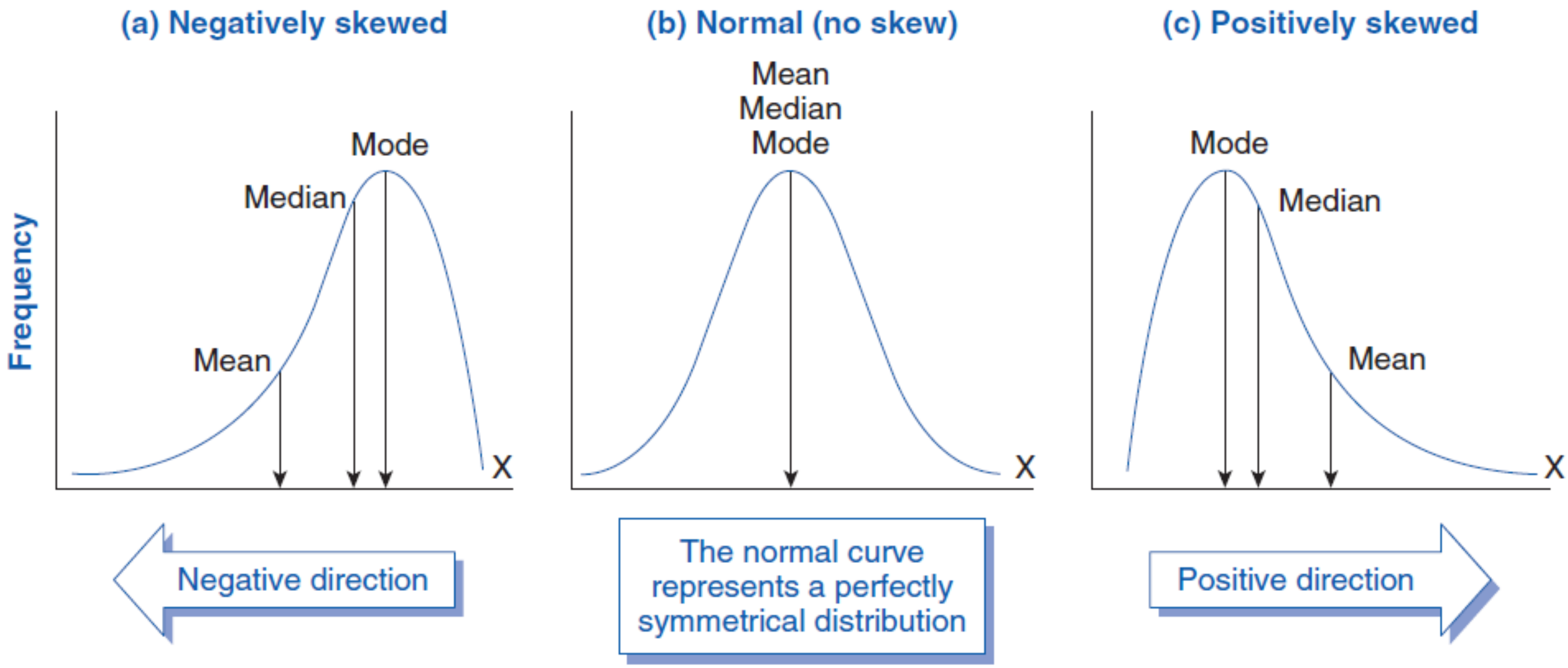
The dispersion values of the weight measures (see slide 15) are as follows:

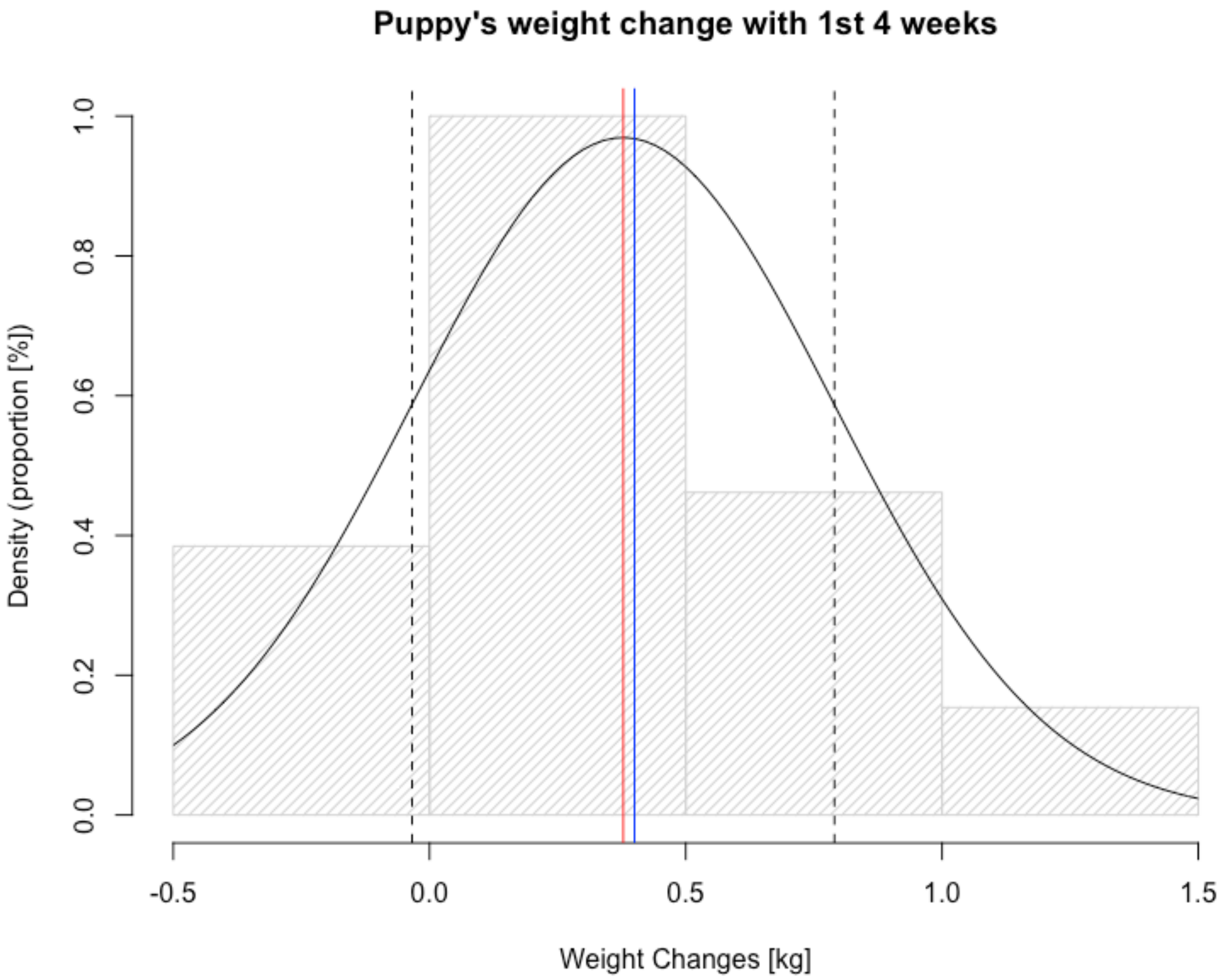
- The overall standard deviation was ±0.412 kg. The change in weight of puppy varied between 0.378 ± 0.412 kg (i.e., -0.034kg to +0.79kg)

IMPORTANT: Notice how the mean and median lay at the centre of the histogram or density plot.

Also, notice how the shape of the histogram is near symmetric.

This pattern is what we called a **"Normal distribution"** or **"Bell-shaped curved"**


(a) Negatively skewed
(b) Normal (no skew)
(c) Positively skewed
Frequency
Mode
Median
Mean
X
Mean
Median
Mode
X
Mode
Median
Mean
X
Negative direction
The normal curve represents a perfectly symmetrical distribution
Positive direction



Code:

```
# enter data
weightChanges <- c(-0.2,-0.4, 0, 0.1, 0.1, 0.3, 0.12, 0.4, 0.5, 0.8, 0.9,
0.5, 0.6, 0.7, 0.6, 1.3, -0.2, 0, 0.1, 0.1, 0.3, 0.4, 0.5, 0.5, 0.6, 1.2)

# next, extract mean and standard deviation from data
m<-mean(weightChanges)
std<-sd(weightChanges)

# plot histogram with normal curve
hist(weightChanges, density = 20, prob=TRUE, main="Puppy's weight
change with 1st 4 weeks", xlab = "Weight Changes [kg]", ylab = "Density
(proportion [%])")

# adds the normal curve
curve(dnorm(x, mean=m, sd=std), add=TRUE)

# add red line for mean
abline(v = 0.378, col = "red")

# add blue line for median
abline(v = 0.400, col = "blue")

# add black dashed line for -sd
abline(v = -0.034, lty = "dashed", col = "black")

# add black dashed line for +sd
abline(v = 0.79, lty = "dashed", col = "black")
```

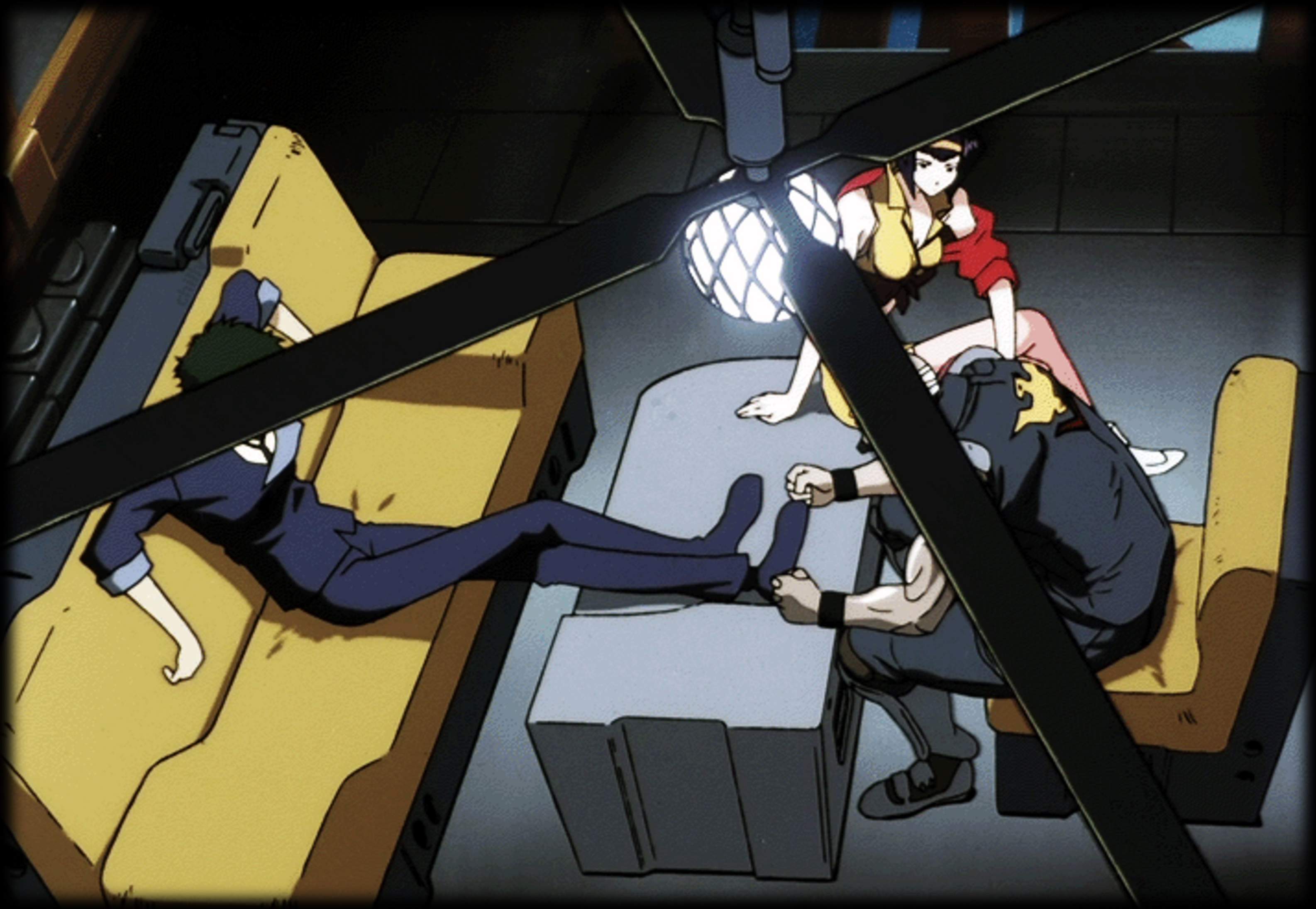
Visualisation in RStudio

Creating impressive visualisation

- Base functions for creating graphics in R: plot( )
- R Packages for creating impressive plots: ggplot2( )
  - You will have to first install 'ggplot2' package first with the install.package( ) function
  - After its installed, you will need to load the package into R with library( ) function
- All this will be become clear in the practical session

原作　矢立肇

Breaktime

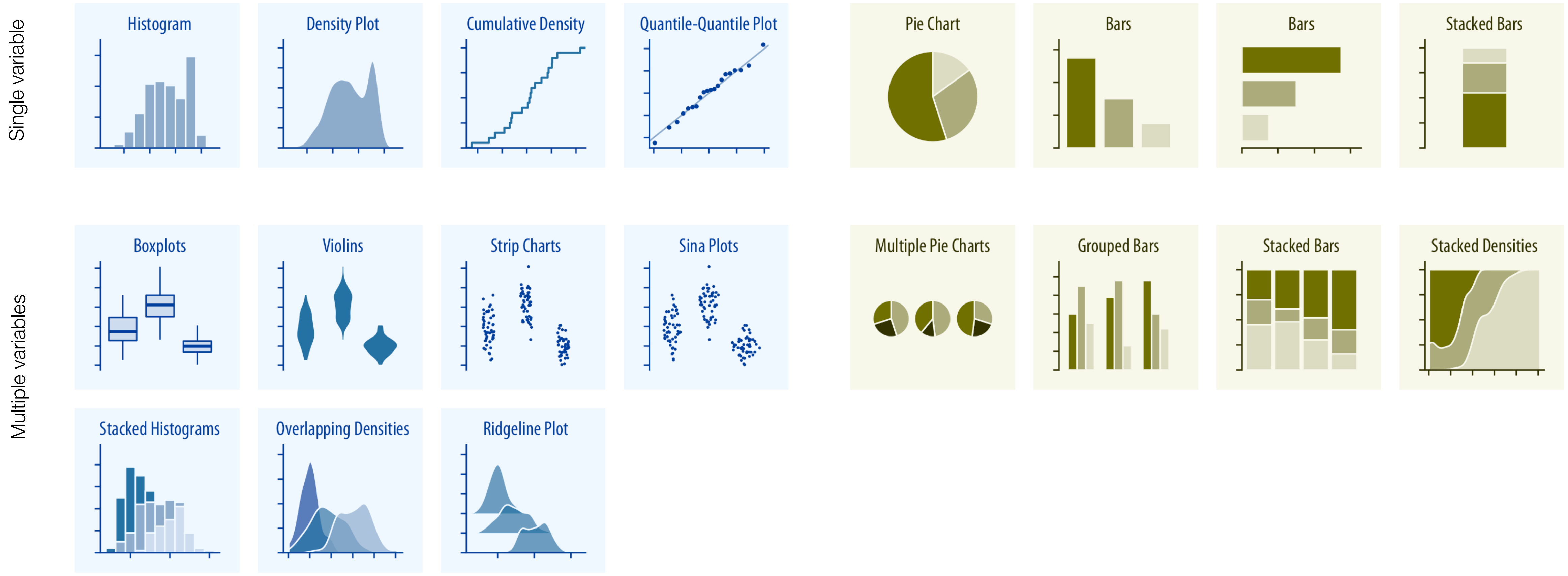


Visualisation: Types of graphs & scenarios

Graph types for data visualisation

1. Plots for densities and distributions (numerical data)

2. Plots for proportions (qualitative or categorical data)



Source: Fundamentals of Data Visualization [LINK]

Visualisation of data with amounts & categories [1]

Single variable (W/categories)

Multiple variables (W/categories)

A single variable with a set of categories

The most commonest approach to visualising data the corresponds to amounts (i.e., numerical values [or proportions]) for some of categories in a categorical variable is the use of bars.

Alternatively, the bars can removed and replace with dot at the location where the corresponding bar would end.

Multiple variables each with a set of categories

If there are two or more sets of categories for which we want to show amounts, we can group or stack the bars. However, we can also map the categories onto the *x* and *y* axis and show amounts by colour, via a heatmap.

Personal thoughts (especially when dealing with singular variables with categories):

- In a **nominal** case – the ranking i.e., categories high to low (and vice versa) matters in the visualisation
- In an **ordinal** case – the order of the categories matters in the visualisation

Let's see examples and potential pitfalls!

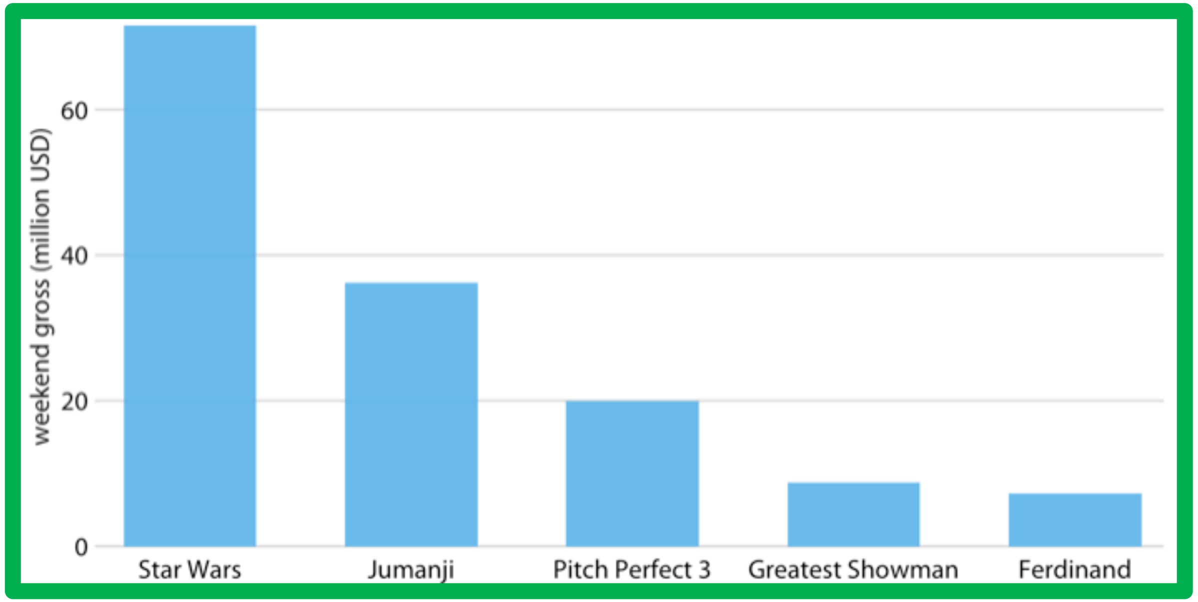
Visualisation of data with amounts & categories [2]

The table below contains movies with the weekend grossing values (Source: http://www.boxofficemojo.com/)

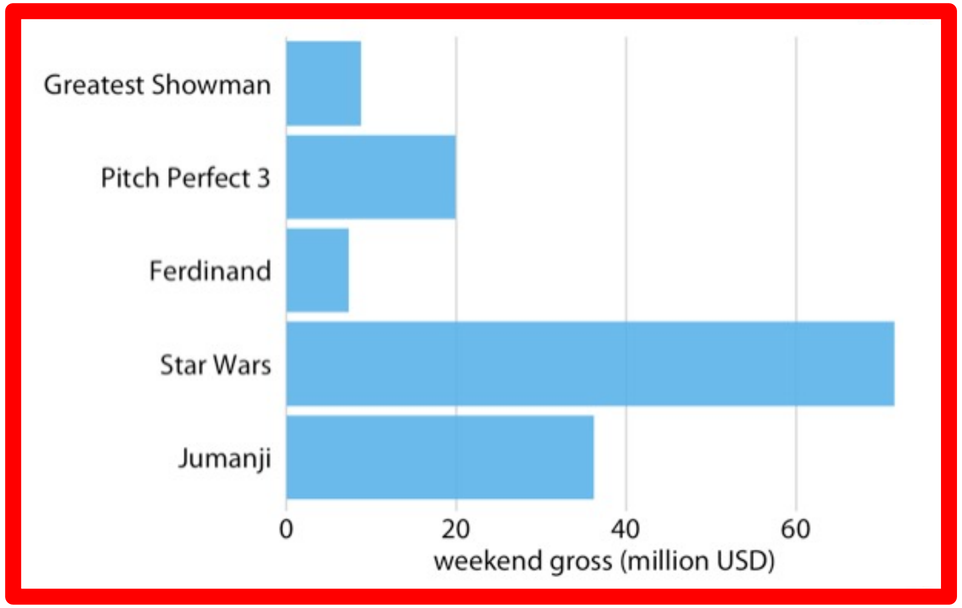
Rank	Title	Weekend gross
1	Star Wars: The Last Jedi	$71,565,498
2	Jumanji: Welcome to the Jungle	$36,169,328
3	Pitch Perfect 3	$19,928,525
4	The Greatest Showman	$8,805,843
5	Ferdinand	$7,316,746

Notes: Remember in a **nominal** case - the ranking i.e., categories from high to low (or vice versa) matters in the visualisation

Good



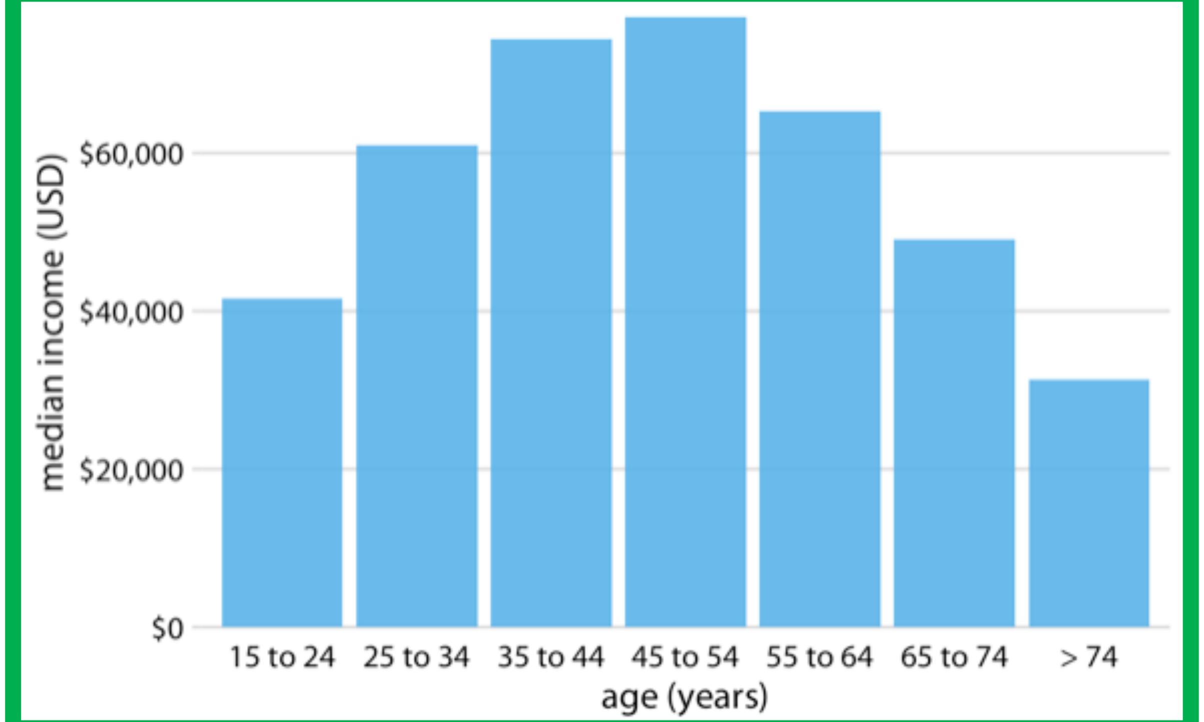
Bad



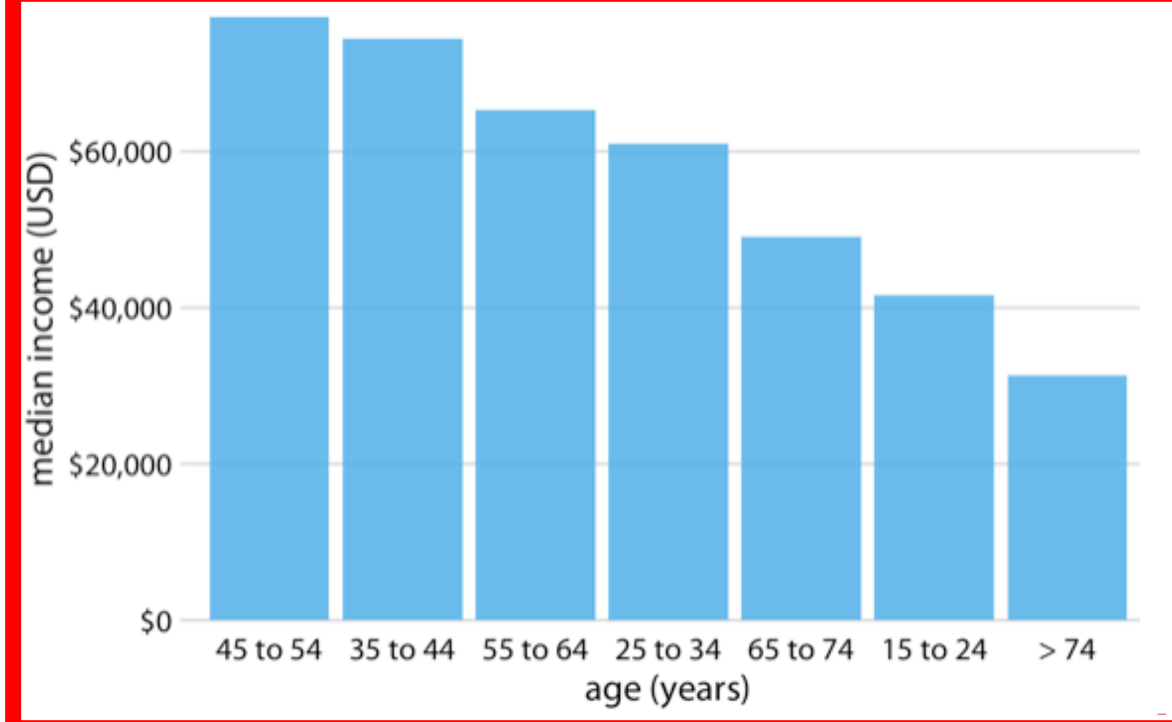
Visualisation of data with amounts & categories [3]

The graphed data represents 2016 median U.S. annual household income versus age group

Good

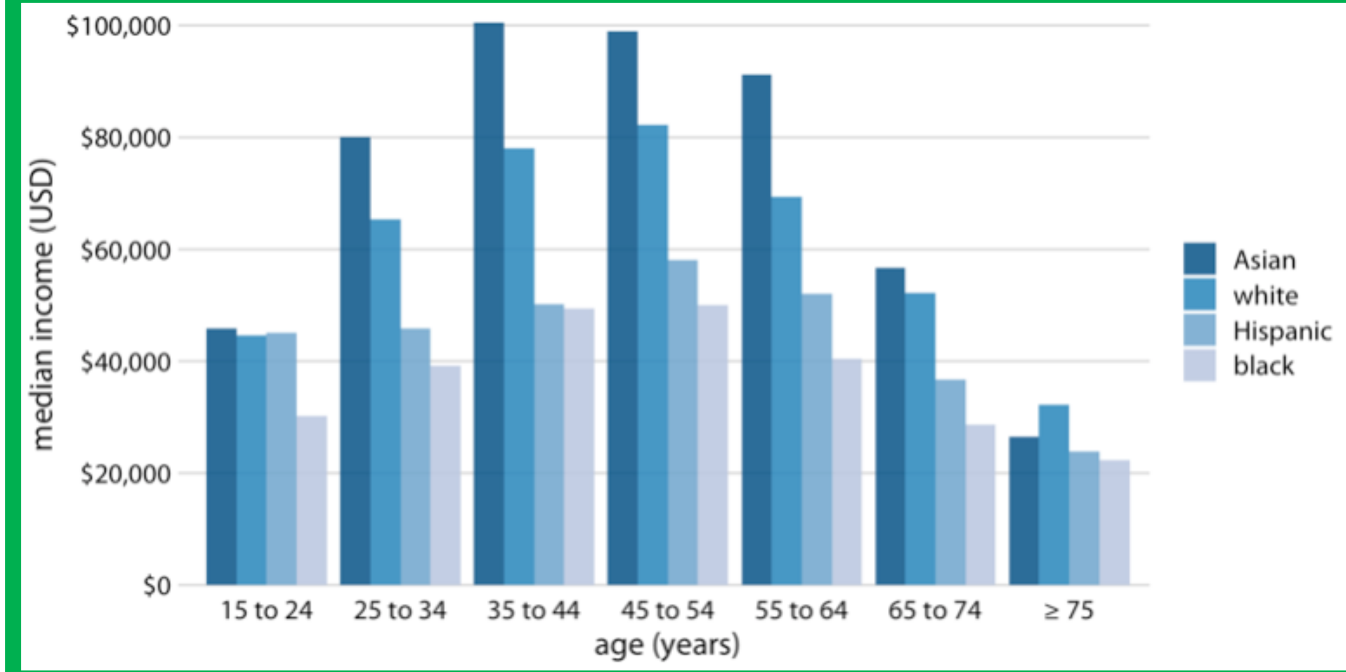


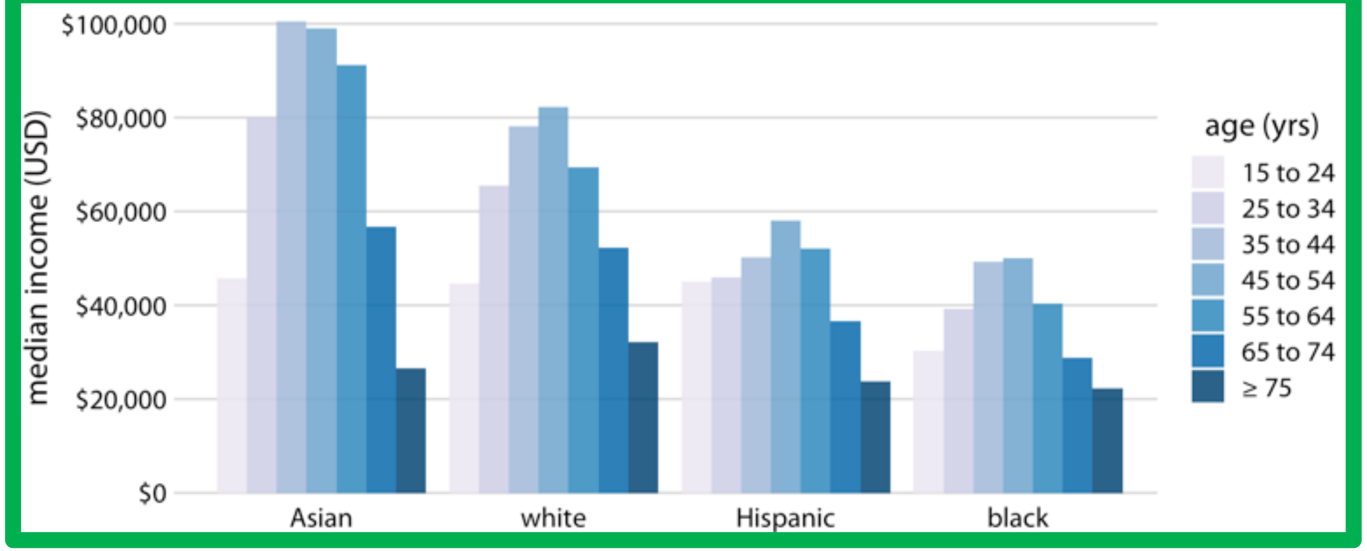
Bad



Notes: Remember in an **ordinal** case - the ordering of the categories **matters** in the visualisation

Visualisation of data with amounts & categories [4]





This is an example of a grouped bar plot. Where the outcome variable [income (in USD)] is visualised across two other independent categorical variables:

- Age group (years) [ordered]
- Ethnicity [nominal]

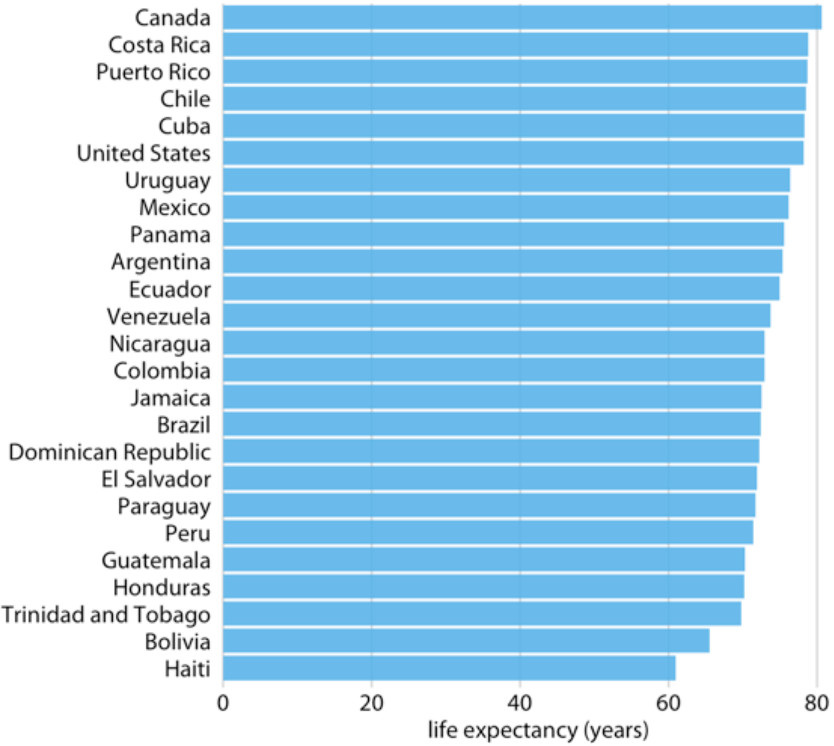
Top graph: Main variables of interest in this case is **income** versus **age groups** (which the age groups are broken down by ethnicity). Ordering in the age groups are maintained here on the x-axis.

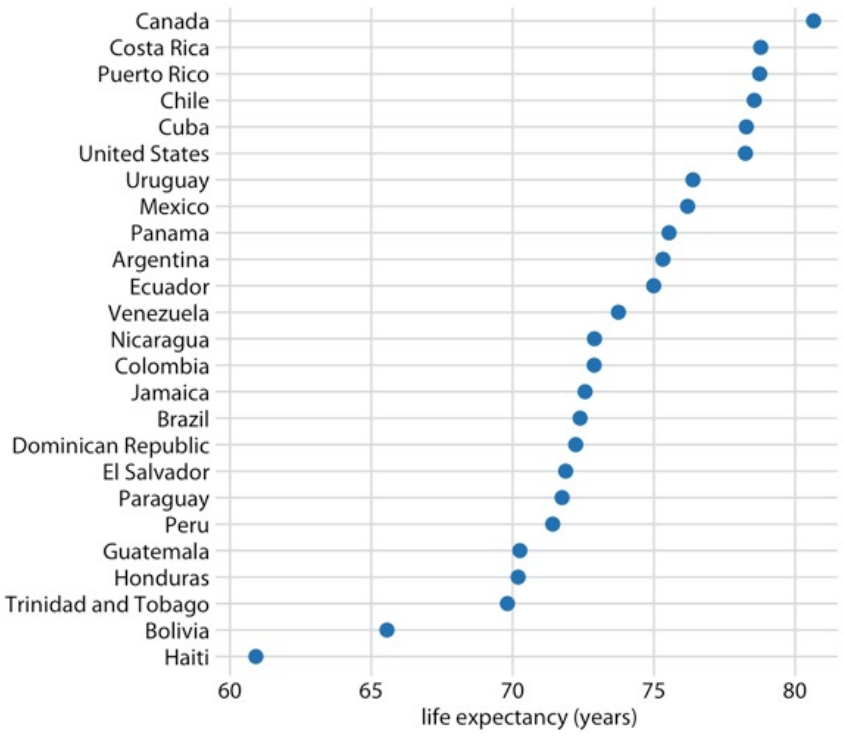
Bottom graph: Main variables of interest in this case is **income** versus **Ethnicity groups** (which the ethnic groups are broken down by age). Ordering in the ethnic categories on the x-axis is not a issue here - **but the ordering of the age groups within ethnic categories must be maintained here on the x-axis.**

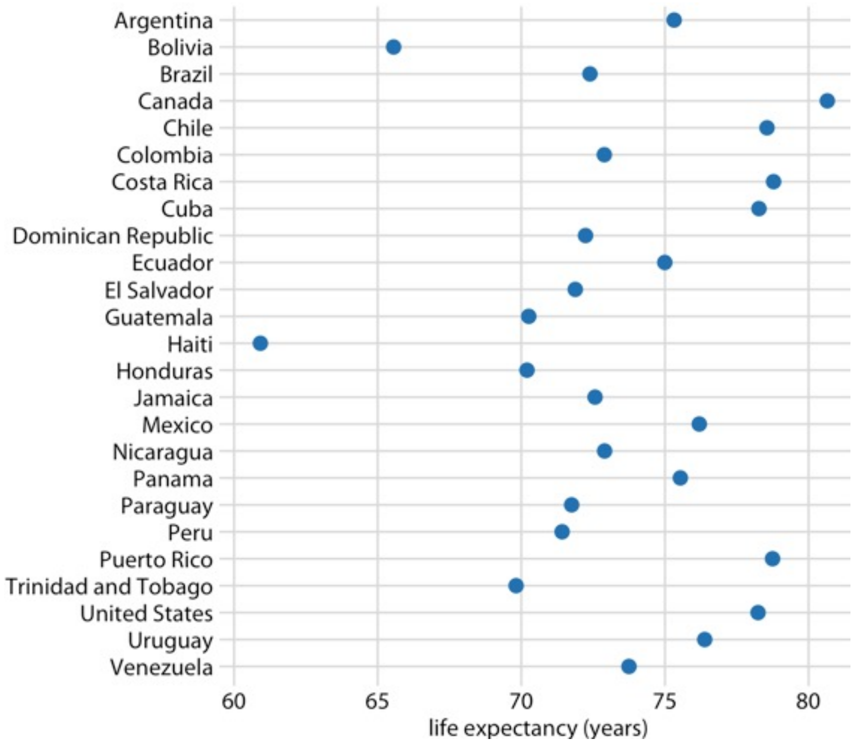
Visualisation of data with amounts & categories [5]

Dot Plots: An alternate version to the bars are dot plots. The bar can removed and replace with dot at the location where the corresponding bar would end.

The three images at the bottom represents data on life expectancies of countries in Central and South America in 2007 [Data source: [Gapminder project]]

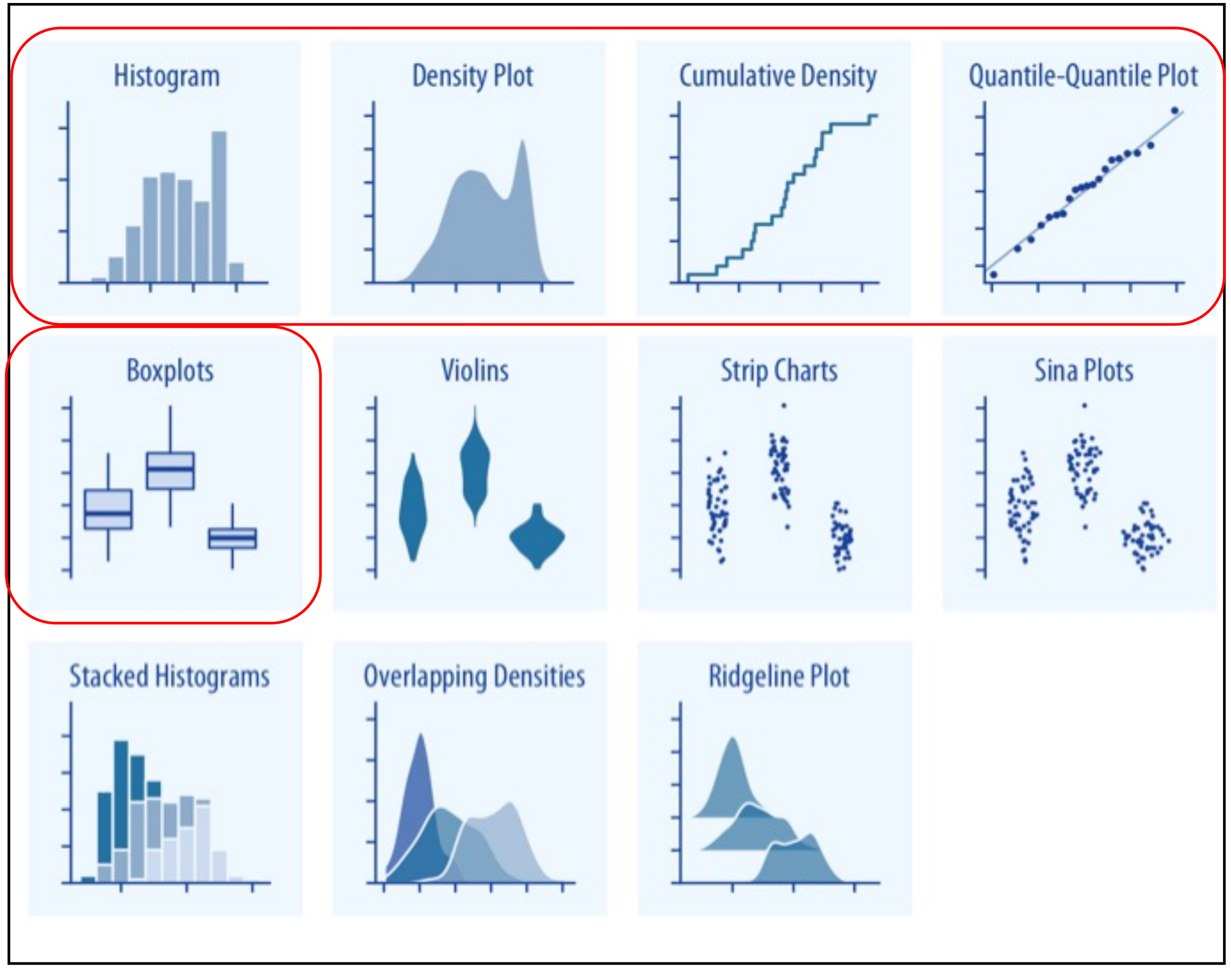






QUIZ: Which graph is correct – LEFT, MIDDLE OR RIGHT one?

Visualisation of data concerning distributions

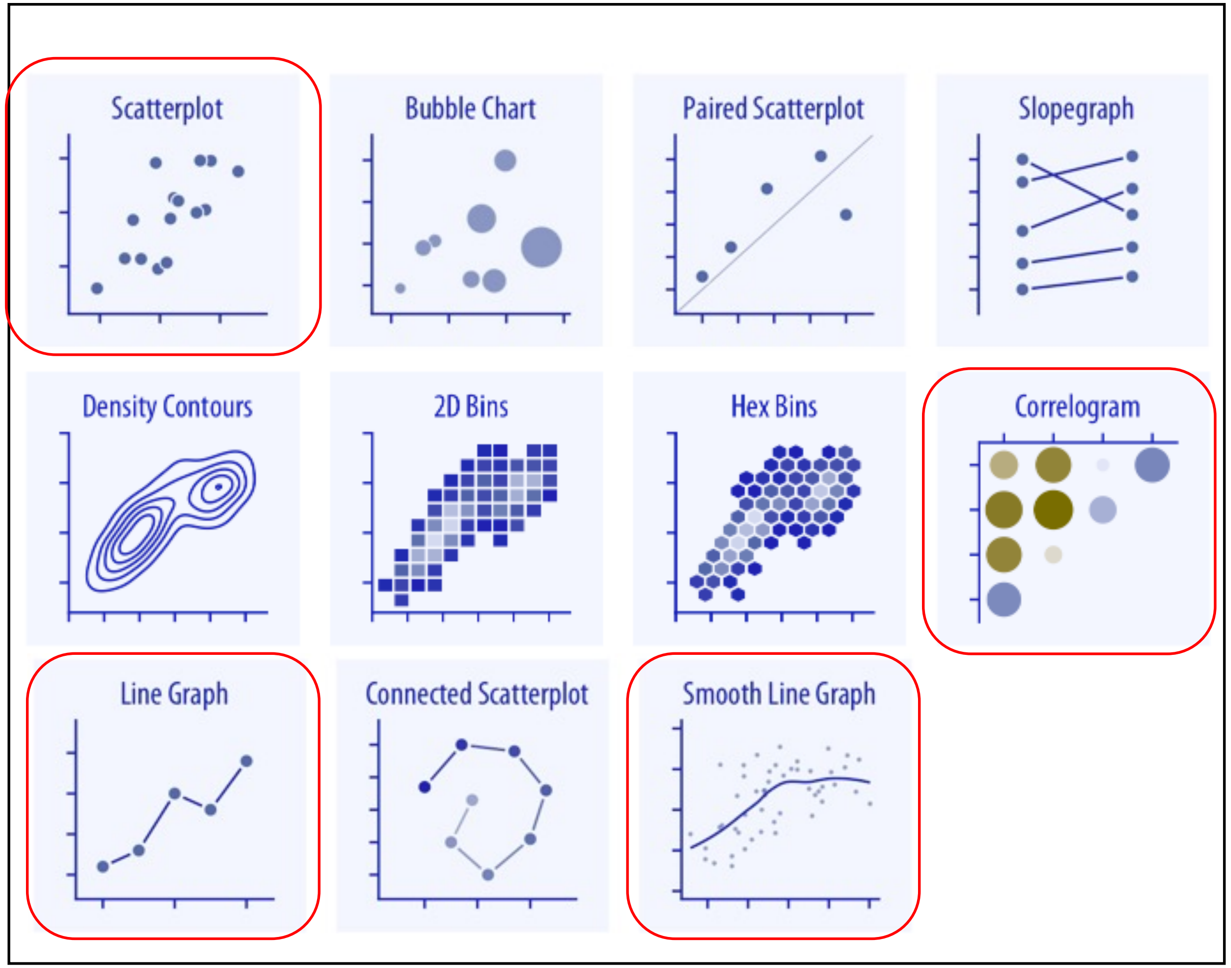


These visualisation methods are the most common approach to represent or display the frequency of observations, as well as how data is spread out over an interval, or how its grouped/clustered around a central point.

You can't really go wrong here – your go-to choice for visualising continuous data for examining its distribution are:

- Histogram
- Density plot
- Cumulative density plots
- Box plots
- Quantile-Quantile plot (used in regression… a lot!)

Visualisation of data concerning x-y relationships [1]



These visualisation methods are the most common approach to represent the relationship between one continuous variable to another. **The y-axis is always the dependent variable, and the x-axis is always the independent variable where we assess its effect or impact on the dependent variable!**

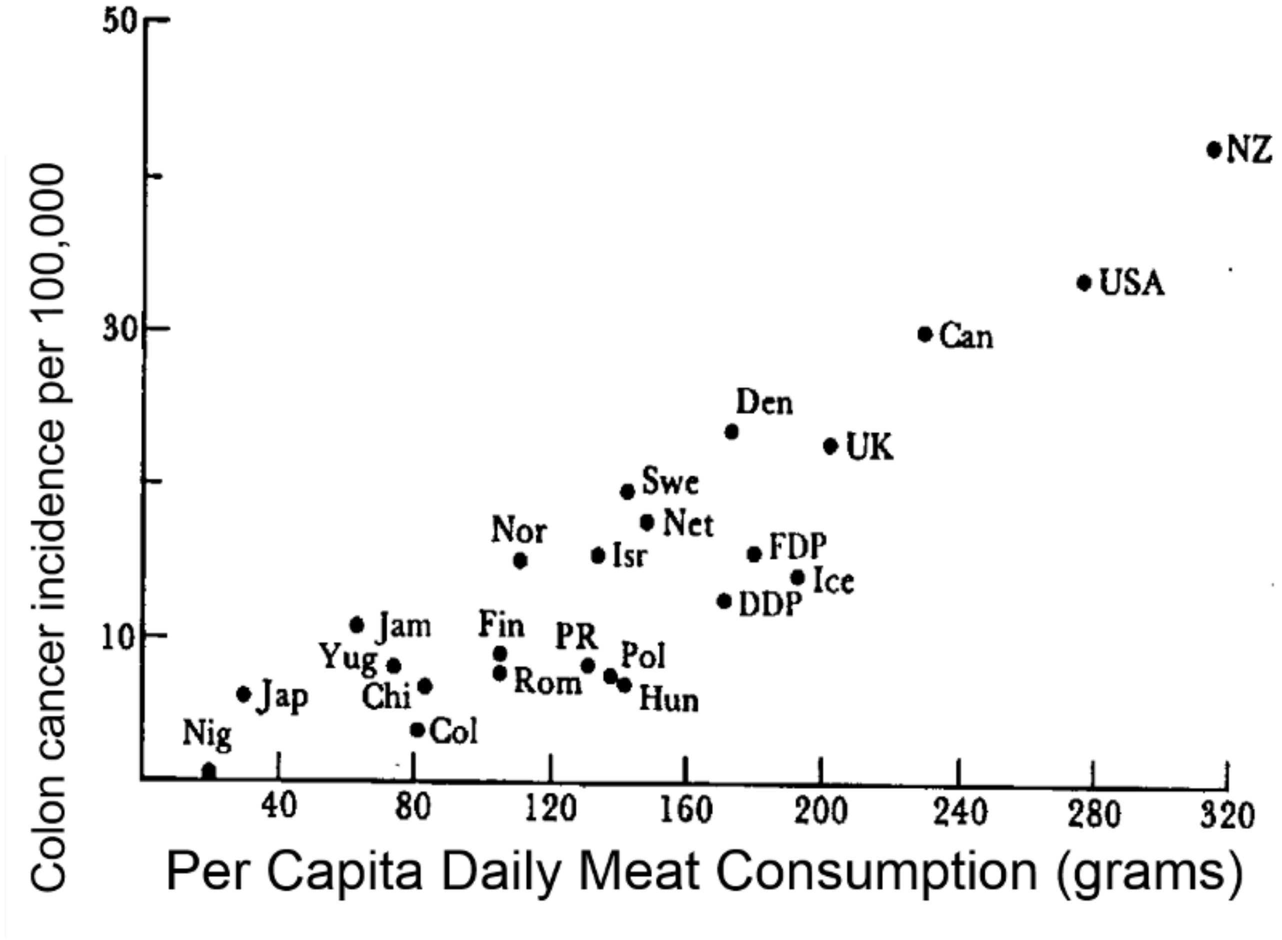
No never make this mistake of flipping the positions - as this will be considered as a critical error

Your go-to choice for visualising two continuous data for examining its relations are:

- Scatter plot (correlation & regression)
- Line graph (time series [outcome versus time]])
- Smooth line graph (non-linear regression)
- Correlogram (correlations between multiple pairs of variables.

Visualisation of data concerning x-y relationships [2]

Assessing the relationship between meat consumption and colon cancer, a country-level analysis



Don't worry – you'll be taught correlations soon. This is an example to show its use on scatter plots

R code:

```
cor.test(df$xvar, df$yvar)
```

Example output:

```
cor.test(cancerdata$meatcon, cancerdata$incidence)

OUTPUT:
Pearson's product-moment correlation
data: cancerdata$meatcon and cancerdata$incidence
df = 2, p-value = 0.0015

95 percent confidence interval:
0.6451325 0.9963561
sample estimates:

Cor
0.8315218
```

Interpretation: There is a very strong positive correlation between levels of meat consumption and incidence of colon cancer in general, and such relationship is statistically significant (p = 0.0015 < 0.05)

Best practices & advice when it comes to data visualisation [1]

Everything on your graph should be labelled accordingly:

- **Title** – a clear short title letting the reader know what they are looking at should be present. Or alternatively, a figure legend for that image will do.

- **Axis labels/titles** – clear labels for the x and y axes must be present
  - ❖ Should include in the labelling the units of measurements [height (m), soil arsenic (mg/kg) etc.]
  - ❖ These labels should be short and descriptive

- **Legends** – for categories in categorical variables which keys/colour codes must be present and labelled accordingly
  - ❖ Male and Female, and not 0 and 1.

- **Captions** – If the graphics are **NOT** yours (i.e., its ripped from a source). Take the opportunity to apply a caption on the graph (on or beneath it) providing source attribution for the data.

- **Colour scheme** – Use of colour scheme matters
  - ❖ Sequential colours – for plotting quantitative variable that goes from low to high (vice versa)
  - ❖ Diverging – for contrasting the extremes (low, medium and high) of a quantitative variable
  - ❖ Qualitative – e.g., nominal categories. Use to distinguish between different categories in a categorical variable

Best practices & advice when it comes to data visualisation [2]

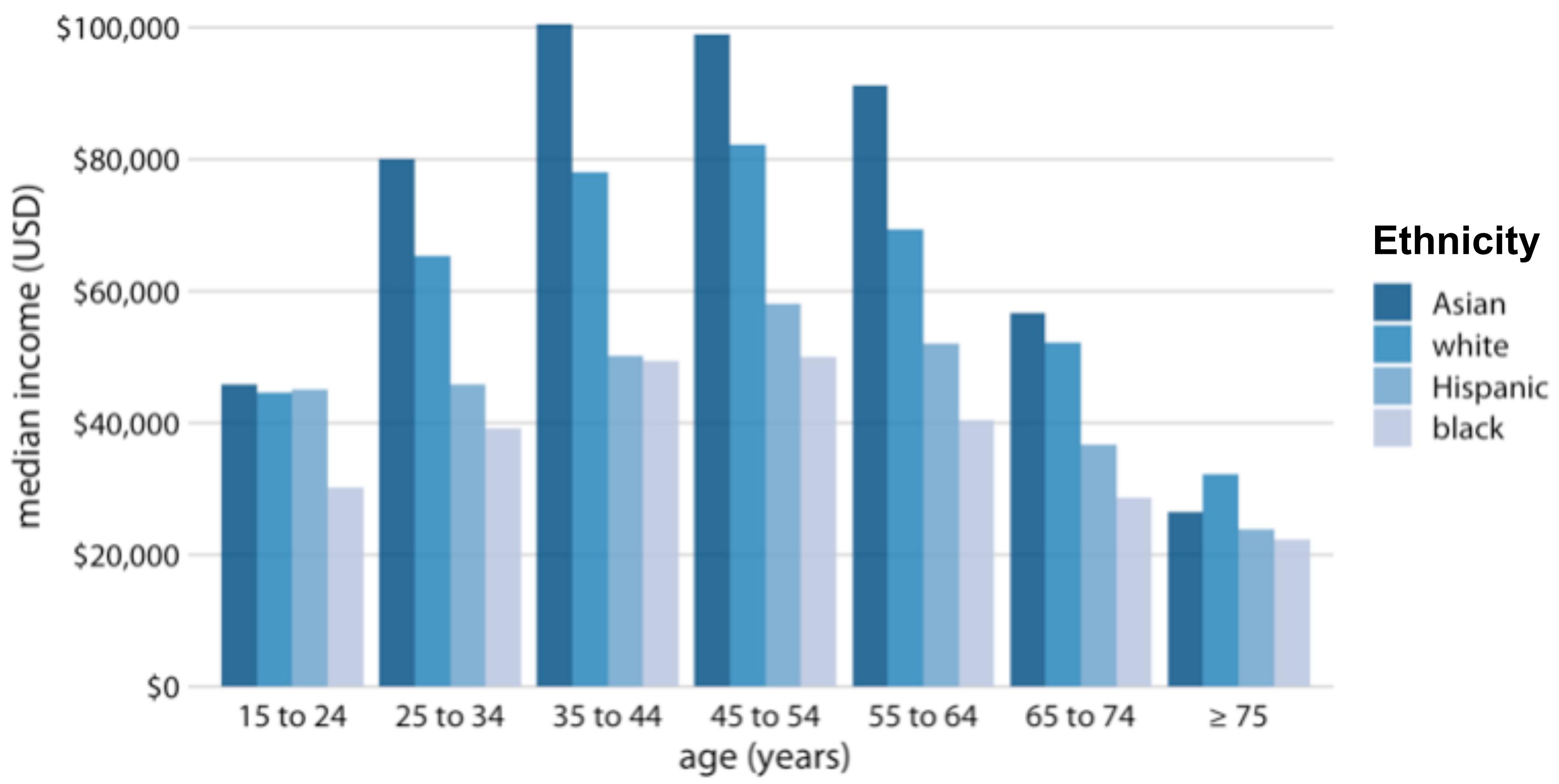


Figure 1: Descriptive analysis shows the overall median income among various age groups broken down ethnic categories in the United States of America.

- The x and y axis are labelled with the current units of measurements (i.e., years and USD)
- Legend for ethnic, which is labelled accordingly has been provided, and colour coded too.
- A title was not given but a figure legend was added at the base of the image.
- Example of very good visualisation

Any questions?