Wk2: Numerical Summaries

- Population distributions:
    - Theoretically where our sample came from
    - Sample shape roughly matches population shape (histogram)
    - Larger the sample (n), better the fit.

- Population dist. incl. measures of:
    - Centre: 'middle of data'
        - Median ($\tilde{y}$ or $\tilde{\mu}$) or mean ($\bar{y}$ or μ)
        - Median is not as influenced by outliers and so is 'more robust'- buzz word!
    - Spread/variation: 'how spread out data is'
        - Inter quartile range (IQR): middle 50% of data
        - Range: (highest-lowest data value) easily influenced by outliers.
    - Shape: 'the shape the histogram follows'.
        - Symmetric: normal, uniform, triangular
        - Skewed: left/ right
        - Bi/multi modal: more than one peak
    - Standard deviation:
        - Deviations of data from the mean: residuals
        - Another special average: for how far the observed scores vary from the mean score. The average difference in observed values from the mean value.
        - Calculated by formula:

$$s = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n - 1}}$$

        - Where residuals = $(y_1 - \bar{y})$ = obs. Value – sample mean.
        - YOU DON'T HAVE TO KNOW THIS FORMULA, IT'S JUST GOOD TO UNDERSTAND. USE CALCULATOR FOR S.D.
- Summarising categorical data:
    - Mean/median not appropriate
    - Make a table with:
        - Variable name
        - Category name

- Count, proportion and % of each observation in each category
- Sample size (n).