Steps Prior to Data Analysis
Prepared by Statistics Canada
August 2022

- **Get familiar with the data**
  - Read all available documentation: User's Guide, Data Dictionary, Metadata, etc.
    - What is the data source/type (survey, census, administrative, linked data)?
    - What are the observed population, the observation unit and the reference period?
    - Is it microdata or aggregate data?
    - Are there suppression rules to apply?
  - If using multiple cycles, make sure the concepts were defined and measured in a consistent manner. Pay special attention to whether response categories are consistent across cycles.

- **Clean the data**
  - VIMO (Valid, Invalid, Missing, Outlier)
    - **Valid**: the data values are valid, so not blank or missing and the values are within a valid range.
    - **Invalid:** the data values are impossible (e.g., age of 301 years old).
    - **Missing**: the variable is left blank.
      - Is it a built-in valid skip or a true missing?
      - Imputation?
      - Use only complete cases?
      - Are there too many missing data for a given variable? E.g. you may decide to only include variables will less than 5-10% of non-response.
    - **Outlier:** the values are extremely small or extremely large compared to what we would expect.
      - Look at the distribution
      - Remember that some outlier values may be rare but actually plausible/true (e.g., age of 102 years old)
  - Re-format the variables needed if necessary
    - Re-code missing values
    - "Don't know" response: may make sense to keep as a valid response and not a missing response
  - Define your concepts
    - Derive new variables if necessary (e.g., new age range)

- **Play with the data**
    - Descriptive statistics
        - Look at the frequency distributions of categorical variables.
        - Look at means, medians, histograms of continuous variables.
        - For some pairs of key related categorical variables, two-way frequency distribution can be examined, or a scatter plot for a pair of key related continuous variables.
        - For survey data, the unweighted (sample) statistics and weighted (estimated population) statistics can be examined
        - Explore the correctness of the data by comparing it to other related information.
            - Can you replicate estimates in an existing/published table?
            - Do the numbers make sense?
        - If you are analyzing a sub-population (e.g., unemployed women aged 25 to 34 in rural areas), is your sample size adequate according to what the user guide suggests?
    - Data visualization (e.g., histogram, scatterplot etc.)
        - Can help you detect weird patterns and outliers to investigate
    - Look at the distribution of the survey weights if applicable

Document every step!
A good best practice is to document decisions in a word document or within notes in your code.

Learn more by watching the following videos from the data literacy learning catalogue!

- **Types of Data**: Understanding and Exploring Data https://www.statcan.gc.ca/eng/wtc/data-literacy/catalogue/892000062020004
- **Data Accuracy and Validation**: Methods to ensure the quality of data https://www.statcan.gc.ca/eng/wtc/data-literacy/catalogue/892000062020008
- **Statistics 101**: Exploring measures of central tendency https://www.statcan.gc.ca/eng/wtc/data-literacy/catalogue/892000062020002
- **Statistics 101**: Exploring measures of dispersion https://www.statcan.gc.ca/eng/wtc/data-literacy/catalogue/892000062020003
- **Statistics 101**: Proportions, ratios and rates https://www.statcan.gc.ca/eng/wtc/data-literacy/catalogue/892000062021003