CHECKLIST FOR THE DATA CLEANING AND TIDYING

1. ~~Examine the variables and their data types.~~
2. ~~Examine the head and tail of the data frame. Make sure that you import your data correctly. Check for any separation argument problem (";" or ",") of the data, the existence of header in the dataset as well as the existence of NAs.~~
3. ~~Check whether~~
    a. ~~Column headers are values, not variable names.~~
    b. ~~Multiple variables are stored in one column.~~
    c. ~~Variables are stored in both rows and columns.~~
    d. ~~Multiple types of observational units are stored in the same table.~~
    e. ~~A single observational unit is stored in multiple tables.~~

    ~~If so, apply data tidying techniques such as stack/unstack, melt, and pivot. Examine the head and tail of the tidy data frame.~~
4. ~~Fix the column names if you detect any typos.~~
5. ~~Drop unnecessary columns.~~
6. ~~Remove the duplicates if it is not the nature of the data.~~
7. ~~Get rid of any unnecessary strings in the values.~~
8. ~~Remove the white spaces in the string values.~~
9. ~~Be sure that all strings are in the same format (e.g. all in lower case). If not, correct them.~~
10. Look at the value counts of strings and be sure that all levels of the categories are unique. If not, correct them.
11. If you have year, month, and/or day columns, combine them and create a date column.
12. Examine the data types again and be sure that numeric variables are float, categorical ones are object, and date is in date format. If not, correct it.
13. Examine the descriptive statistics of numerical variables. Search for any unusual behavior. Are the variables in the correct range? If not, find the locations and correct them.
14. Search for possible outliers. If there are outliers, replace them with the mean.
15. ~~Search for uniformity. The units in the numeric columns are in the same format or not. That is, examine whether some data are in meters but some in centimeters. If they are not consistent, convert them into the same units.~~
16. ~~Search for the missing values. Examine their percentage in each column. If the percentage is low, fill them with mean/median/mode. If the percentage is high (e.g >60% 65%), you can drop the column.~~

After you clean your data set, create several research questions (at least 5) to explore your data. Draw figures using all necessary tools to make them perfect, hence you can answer your questions and interpret the results.