Code Comprehension User Study Grading Rubric

Task 0

Main Intension of the Program

- Taking in messy datasets
- Removing empty rows and columns
- Cleaning the dataset to make it ready for later analysis

Programming Actions

- Dropping all the rows with values that are all NaN
- Replacing all the text “——-” with np.nans
- Dropping rows with the string “TIME” as the entity in the “TIME” column
- Dropping all the NaN columns and resetting the index
- Convert all the values in the numerical column into the actual numerical value.

Task 1

Main Intension of the Program

- Checking the missingness based on the column “T1”
- verifying the significance of the missingness of the columns with the chi2 contingency test

Programming Actions

- Finding a new DataFrame with only columns that have at least one missing value
- Counting the number of non-missing values of the columns in the newly created DataFrame based on the categories in the “T1” column
- Using the “melt” function to present the count of the non-missing value of each column in each T1 category in a cleaner way
- Use the transform function to find the count of the missing value of each column in each T1 category and name it with “Missing Column.”
- The chi2_contingency method was used to assess the missingness in each column(except T1), with 0.0001 as the significant level.
- Finding and plotting the proportion of non-missing values in each category

Task 2

Main Intension of the Program

- filter out certain rows and columns based on the information
- Imputing the DataFrame using the mean imputation of the groups categorized by time quantile and T1 column

Programming Actions

- Filtering out T1 categories with more than 200 rows
- For all the other columns(all columns except “T1”), filter out the entire column if the minimum number of rows of one T1 category is greater than 200
- Using pivot to reset the DataFrame
- Adding T1 column back the to cleaning DataFrame
- Giving quantiles to each rows based on the time column\
- Replacing the missing the value of the by the meaning of the group that that value belong to in terms of quantiles and T1 column

Task 3

Main Intension of the Program

- Adjusting the wind direction so it can be categorized into certain category in the later part
    - Clustering them
    - None of them would go over 360
- visualizing the change in the distribution of present wind direction before and after imputation

Programming Actions

- Concatenating the original and imputed DataFrame together
- Filtering out rows outside of certain time period
- Adjusting the value in WDIR
- Converting time to full number
- converting WDIR from numerical to categorical value