Description of my work:

* Performed EDA on the dataset to help in finding best ways to answer our problem
* Performed first stage data preprocessing which included the following;
  + Creating new column by extracting data from different column (Year, Time duration…etc.) which assisted in analysis.
  + Imputed missing values and verified the authenticity of the imputation
  + Collapsed subcategories to prepare variables for modeling (weather condition & wind speed)
* Performed second stage data preprocessing which included the following;
  + Performing preprocessing on different data set and combined it with the original data to enhance model performance that was weakened by the imbalanced data.
  + Tried different missing value imputation methods (regression) to impute values for model features with highest missing values (precipitation and Wind chill)
* Assisted with modeling requirements (providing different clean datasets and brainstorm meetings)
* Provided weather conditions EDA Plots for the GUI.
* Performed initial logistic regression model that included weather condition variables only
* Wrote description of data set, preprocessing techniques and appendix tables in the group project report.

Please reference the following files:

* Data Preprocessing (first stage data preprocessing)
* Data\_Pre\_2 (second stage data preprocessing)
* EDA\_Plots (weather EDA plots)
* Wind\_Chill\_Imputation (Windchill missing value regression imputation)
* Weather\_model

Note that the previous files were used for preprocessing different datasets, Therefore, it was edited multiple times and generated different files. It was not intended to be run by user. However, the code can be referenced to review the preprocessing process. In addition, some imputations were neglected as a result of authenticity check (second & third imputation). It was intended to be left in the code files for documentation.

Also please reference the following zip files;

* us\_data\_2019\_criteria\_1 (preprocessed 2019 data using first imputation)
* us\_data\_2019\_criteria\_2 (preprocessed 2019 data using second imputation)
* us\_data\_2018 (preprocessed 2018 data using first and second imputation)
* us\_data\_2017 (preprocessed 2017 data using first and second imputation)
* us\_data\_2016 (preprocessed 2016 data using first and second imputation)
* final\_data ( preprocessed dataset that used 2019 & 2018 using first imputation method and columns for imputed windchill using regression imputation)