**Data Processing Documentation**

Notation: file name that are mentioned for the first time is highlighted in bold; file name that are subsequently mentioned is set to italics.

1. Original datasets from data sources:
   1. Credit ratings data of U.S. companies from 2006-2015
      1. Extracted from WRDS;
      2. Initial cleanup in Excel: drop unnecessary columns and keep only: gvkey, rating, company name, date, year, month, and state;
      3. File name: **Ratings 2006-2015 original.csv** (which is the already cleaned up version after the step described under a(ii).
   2. Weather data from 2006-2015
      1. Extracted from NOAA
      2. Initial processing in Excel: filtered to get only data for years 2006-2015 and remove unnecessary columns (such as county name, event ID)
      3. File name: **Weather data 2006-2015.csv** (which is the already cleaned up version after the step described under b(ii).
   3. Economic data from 2006-2015 on state level
      1. Extracted from Bureau of Economic Analysis
      2. File name: **Economic data.csv**
2. Initial processing of ratings and economic datasets
   1. Method: Python (file name: **Initial processing before SQL.py**)
   2. Processing of *Ratings 2006-2015 original.csv*:
      1. Convert the column of state initials into state full names to help better match with the corresponding column in the weather and economic datasets
      2. Extract gvkey of the companies in the dataset for further query in WRDS which is output as **gvkey.txt**
   3. Processing of *Economic data.csv*
      1. Clean up the column of geographical area names
      2. Output file: **Modified economic data.csv**
   4. More detailed descriptions can be found in the comments of the python file
3. Query for fiscal year end information and financial ratios of companies in WRDS using *gvkey.txt*
   1. The output files from WRDS have been initially cleaned up in Excel by dropping unnecessary columns
   2. The financial ratios selected include: Total Debt/Total Assets (Total liabilities as a fraction of total assets), Total Debt/Total Assets (Total debt as a fraction of total assets), Total Debt/EBITDA, and Cash Flow/Total Debt
      1. Reason for choosing: query results from WRDS indicate inconsistency in the availability of company rating data and company ratio data. These ratios are used because while being credit-related, they also enable a relatively large number of original observations to be kept.
   3. The files include: **Fiscal year end original.csv**, and **Financial ratios.csv**
   4. The files in 3c are then further cleaned up using Python:
      1. *Fiscal year end original.csv* is cleaned up using **Clean fiscal year end.py** to drop unnecessary columns;
      2. *Financial ratios.csv* is cleaned up using **Financial ratios.py** to drop unnecessary columns and remove duplicates.
      3. The output files (cleaned up versions) are: **Fiscal cleaned.csv**, and **Ratios cleaned.csv**
4. Merge datasets (except for weather dataset) using SQL
   1. More detailed descriptions and codes can be found in the file **SQL code and description.docx**
   2. First, the fiscal data and ratings data are merged using gvkey, fiscal year end, and year information
   3. Then, the output table from 4b is merged with the economic data using geography information (i.e. state name)
   4. Lastly, the output table from 4c is merged with the ratios data using gvkey and year information
   5. The final table from 4d is then exported as **Rating\_fiscal\_econ\_ratio.csv**
5. Merge *Rating\_fiscal\_econ\_ratio.csv* with *Weather data 2006-2015.csv*
   1. The merge is completed in Python using the script file **Merge weather data.py:** 
      1. To begin with, the file is *Weather data 2006-2015.csv* cleaned up by converting its columns that express damages of weather events using strings (e.g. “0.5k”) into numerical values;
      2. Then the column in *Weather data 2006-2015.csv* expressing tornado scale using strings (e.g. “EF0”) is also converted into numerical values;
      3. After the weather data is cleaned up, pivot tables that summarize the weather events (using aggregation of “count”) and severity measure (using aggregation of “sum” except for the variable of tornado scale, in which case “average” is used)
      4. Afterwards, Rating\_fiscal\_econ\_ratio.csv is merged with weather data using pivot tables mentioned above in 5a(iii)
      5. Additionally, dummy variables for each weather event (for indicating their existence) are added to the dataset
      6. Lastly, the “Rating” column is converted into numerical values
   2. More detailed descriptions can be found in the comments of the python file
   3. Output: file **Final rating dataset.csv**