P2: Data Acquisition and Cleaning

Ram Pangaluri, Fabian Leon, Ashok Shanker

Dataset Review

For this project we plan to primarily use two main datasets. The first is the TAES dataset from Texas A&M Agrilife, which contains data from 2021 Texas Sorghum Variety Trials. The second is the dataset from the USDA National Agricultural Statistics website. This dataset is helpful because we can adjust this data for different time periods, counties, etc.

TAES Dataset:
This dataset contains sorghum statistics from 1970-2021 for different counties/cities around Texas, centering around College Station.

Extract
The source of this data is Texas A&M Agrilife. Original file was in csv file format.

Transform
As csv files are hard to work with, we imported the file to excel and saved it as an excel file. Additionally several unnecessary columns with ancillary variables were deleted, so that our analysis could focus on the factors that we believe impact sorghum growth the most. Empty values were left in.

Load
Final file saved as excel spreasheet.

USDA Dataset:
This dataset is more queryable and contains additional information on not only sorghum but other crops as well. There are many columns.

Extract
The source of this data is the USDA National Agricultural Statistics Website. Original file was downloaded as excel file.

Transform

Load