P2: Data Acquisition And Cleaning

Ram Pangaluri, Fabian Leon, Ashok Shanker

February 13, 2022

Data Sources

The sources of data that you will extract from.

Texas A&M AgriLife Research (TXAR) Dataset

Texas A&M AgriLife Research conducts the grain sorghum performance tests each year to provide growers in Texas with accurate and unbiased information on hybrid performance at locations across the state. Selection of superior hybrids that are well adapted for a given region is essential for maximizing yield and profit. The TAES dataset contains data from multi-environment trials of commercially released sorghum hybrids grown from 1970-2021 across different counties/cities around Texas.

```
############## Read dataset  ##############
TXAR = read.csv("data/TXAR_data.csv", na.strings = c(".","", "NA"))
#Get rid of the observations from Curry County, New Mexico
TXAR = TXAR[-which(TXAR$County == ’Curry’),]

TXAR[,2:9] = lapply(TXAR[,2:9], factor)
TXAR[,10:19] = lapply(TXAR[,10:19],as.numeric)

# To display the contents and range of this dataset
year_range = c(min(TXAR$Year), max(TXAR$Year))
nyears = length(unique(TXAR$Year))
nlocations = length(unique(TXAR$Location))
ncounties = length(unique(TXAR$County))
nhybrids = length(unique(TXAR$Hybrid))
ncompanies = length(unique(TXAR$Brand))
```

Characteristics/Attributes: This dataset is extensive and has 50+ columns describing hybrid sorghum performance and the management practices used to grow each trial. Some important variables include…

- Year - Year survey data point was obtained
- County - Texas County Name
- Hybrid - Specific sorghum genotype
- Irrigation - Irrigation level/amount
- Brand - Company brand of sorghum plant/seed
- Grain Yield, Plant height, Days to Flowering, etc.

USDA Dataset

The USDA's National Agricultural Statistics Service (NASS) conducts hundreds of surveys every year and prepares reports covering virtually every aspect of U.S. agriculture. Production of grain sorghum, prices paid and received by farmers, farm labor and wages, farm finances, chemical use, and changes in the demographics of U.S. producers are only a few examples!

Transformation

The type of transformation needed for this data (cleaning, handling missing/incomplete data, etc.).

```r
# Required Transformations:
# Get rid of the observations from Curry County, New Mexico
TXAR = TXAR[-which(TXAR$County == ’Curry’),]

# Make variables appropriate
TXAR[,2:9] = lapply(TXAR[,2:9], factor)
TXAR[,10:19] = lapply(TXAR[,10:19],as.numeric)
```

Final File

The type of final file to load the data

The final tables or collections that will be used in the project. & tables/collections/observations/columns, and why this was chosen.