P4 - Data Analysis Proposal

Ram Pangaluri, Ashok Shanker, Fabian Leon

How have Texas' sorghum yields improved over time?

Advances in breeding, technology, production, and management all led to higher sorghum yields for our state. How have sorghum yields changed over time and what time periods saw the greatest increases in yield?

This analysis will be implemented as a *line graph* with range sliders for the year range. The `plotly` package will be used to execute this visualization. The line graph is a justified approach because metrics of sorghum performance (i.e. grain yield) are continuous variables along a time series.

What physiological dependencies are there between agronomic traits?

For example, is there a relationship between a sorghum hybrid's height and how long it will take to flower? Do taller hybrids yield more? Do hybrids which reach maturity faster have higher yields?

This visualization will consist of an interactive *scatterplot* window that plots and displays *pearson's R* for whichever two variables the person selects from a drop down menu of relevant agronomic traits like yield, plant height, exsertion, days to flowering, etc. The visualization will be implemented using the `ggplot2`, `shiny`, and `shinydashboard` packages.

What counties of Texas are the highest yielding environments for sorghum hybrids?

This analysis will consist of a *Chloropleth Map* of average yields per county. The map would be dynamic and allow the person to hover over counties to get more information on the county's yield numbers and the datapoints which led to those figures. The chloropleth map will be implemented with the `plotly` package. The chloropleth map approach is appropriate since it can efficiently display the continuous yield data on geospatial points which correspond to counties in our dataset.

Which brands have released the most hybrids over the years?

The dataset is representative of 180 company brands which have produced the ~4000 hybrids in the trials. Which brands were most profilic in releasing hybrids over the years?

A *racing bar chart* would allow us to visualize the accumulation of hybrids (a continuous variable) from each brand over the years. The racing bar chart would be implemented using the `ddplot` package in R.

Which major environmental and/or management scenarios lead to higher yields? (Prediction)

To understand what environmental or management factors affect sorghum yields, we will conduct a *linear mixed model* analysis with explanatory variables fitted to yield as follows:

$$Yield = Location + Year + Hybrid + Hybrid * Location + Irrigation + PreviousCrop + Rainfall + PlantingDensity$$

Significance of the effects and the variance explained by the predictor variables will be reported with an *ANOVA* table. Additionally, the model will then be plotted as a *scatterplots* to visualize the effects and their interactions. The linear mixed model framework is justified since many of the environmental and management factors have the potential for second or third order interactions and will need to be evaluated for their contributions to yield variation.