STAT230 Work Team Project: Proposal & EDA

Work Team 7: Evan, Becca, Laith

due Thursday, October 12th

[DELETE & DON'T INCLUDE the following instructions in your proposal]

- Propose a proposal: The proposal part should include at least the following:
  - Work Team number, all members' names, and Work Team Project Title (if there is one).
  - Introduction & Motivation - What do you plan to study? What is the main (research) question of interest? WHY is this question interesting or important to your team?
  - Dataset - Where does it come from (including a citation of the data source)? What variables are you planning to use? Which is the response variable for your analysis?

Since this is a warm-up project, I would recommend each work team to choose at least 4 variables but no more than 6 variables (including the response variable) from your dataset for the Work Team project.

- Perform an EDA (including graphs, numerical summaries, AND verbal descriptions) on the chosen variables that will be used in your modeling. Use appropriate graphical displays to justify that MLR is a reasonable choice for modeling (i.e. the CHOOSE step in 4-step modeling). You may find the document **Modeling in R: EDA** under *Useful Resources* particularly helpful for this deliverable. Additional Notes -

This document should also include all the R code used to *wrangle* data, as well as those used to create the *EDA* output (plots, tables, numerical summaries, etc). Recall that you've learned various `R` functions for data wrangling back in **R Activity 2** (Part 3) and **R Activity 4** (Part 2), as well as from **R Tutorials** (under *Useful Resources*). Don't hesitate to ask for help from me or SDS fellows on this front.

[— DELETE until here: End of Instructions —]

Introduction & Motivation

Dataset & Wrangling

On RStudio, be sure to save your dataset file in *the SAME folder* you saved this RMD file. I would strongly suggest that you create a specific folder on RStudio for this Project. When you wrangle your data, keep in mind that it's usually safer to save the mutated/filtered dataset as a *new* dataset.

```
setwd("C:/Users/emaca/Documents/F2023 Classes/Stats/datasets")
```

```
## wrangle!!
df <- read_csv("gooddata.csv")
```

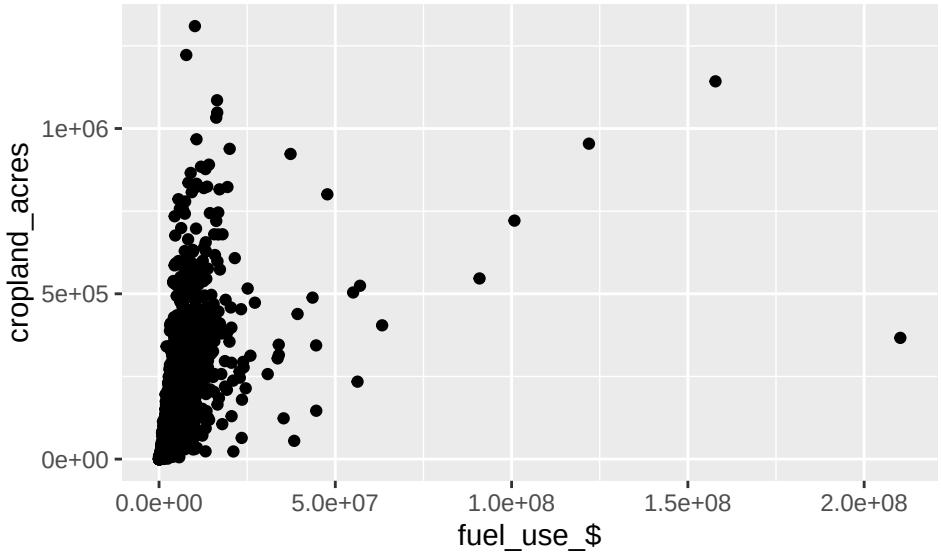
```
## Rows: 3077 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): county_code, honey_sold, state_lower, county
## dbl  (9): xmas_tree_sale, emus_sold, ornamental_fish, bison_sale, aquatic_pl...
## num (13): farm_acres, cropland_acres, irrigated_acres, crop_totals_sales_mea...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Exploratory Data Analysis (EDA)

Explore distributions and associations graphically and numerically.

```
p1 <- ggplot(data = df, aes(x = `fuel_use_$`, y = cropland_acres))+
  geom_point()
p1
```

```
## Warning: Removed 3 rows containing missing values (`geom_point()`).
```




```
ggpairs(df, columns = 2:12)
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 3 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 37 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 4 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 1790 rows containing missing values
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 6 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 9 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 7 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 31 rows containing missing values

## Warning: Removed 3 rows containing missing values (‘geom_point()‘).

## Warning: Removed 3 rows containing non-finite values (‘stat_density()‘).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 37 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 4 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 3 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 1790 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 6 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 10 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 3 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 9 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 31 rows containing missing values

## Warning: Removed 37 rows containing missing values (‘geom_point()‘).
## Removed 37 rows containing missing values (‘geom_point()‘).

## Warning: Removed 37 rows containing non-finite values (‘stat_density()‘).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 38 rows containing missing values
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 37 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 1794 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 39 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 42 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 37 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 40 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 59 rows containing missing values

## Warning: Removed 4 rows containing missing values (‘geom_point()‘).
## Removed 4 rows containing missing values (‘geom_point()‘).

## Warning: Removed 38 rows containing missing values (‘geom_point()‘).

## Warning: Removed 4 rows containing non-finite values (‘stat_density()‘).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 4 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 1790 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 6 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 10 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 4 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 10 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 32 rows containing missing values

## Warning: Removed 3 rows containing missing values (‘geom_point()‘).
```

```
## Warning: Removed 37 rows containing missing values (‘geom_point()‘).

## Warning: Removed 4 rows containing missing values (‘geom_point()‘).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 1790 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 6 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 9 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 7 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 31 rows containing missing values

## Warning: Removed 1790 rows containing missing values (‘geom_point()‘).
## Removed 1790 rows containing missing values (‘geom_point()‘).

## Warning: Removed 1794 rows containing missing values (‘geom_point()‘).

## Warning: Removed 1790 rows containing missing values (‘geom_point()‘).
## Removed 1790 rows containing missing values (‘geom_point()‘).

## Warning: Removed 1790 rows containing non-finite values (‘stat_density()‘).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 1790 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 1790 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 1790 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 1790 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 1795 rows containing missing values

## Warning: Removed 6 rows containing missing values (‘geom_point()‘).
## Removed 6 rows containing missing values (‘geom_point()‘).

## Warning: Removed 39 rows containing missing values (‘geom_point()‘).

## Warning: Removed 6 rows containing missing values (‘geom_point()‘).
## Removed 6 rows containing missing values (‘geom_point()‘).
```

```
## Warning: Removed 1790 rows containing missing values (‘geom_point()‘).

## Warning: Removed 6 rows containing non-finite values (‘stat_density()‘).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 10 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 6 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 10 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 32 rows containing missing values

## Warning: Removed 9 rows containing missing values (‘geom_point()‘).

## Warning: Removed 10 rows containing missing values (‘geom_point()‘).

## Warning: Removed 42 rows containing missing values (‘geom_point()‘).

## Warning: Removed 10 rows containing missing values (‘geom_point()‘).

## Warning: Removed 9 rows containing missing values (‘geom_point()‘).

## Warning: Removed 1790 rows containing missing values (‘geom_point()‘).

## Warning: Removed 10 rows containing missing values (‘geom_point()‘).

## Warning: Removed 9 rows containing non-finite values (‘stat_density()‘).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 9 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 13 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 33 rows containing missing values

## Warning: Removed 3 rows containing missing values (‘geom_point()‘).

## Warning: Removed 37 rows containing missing values (‘geom_point()‘).

## Warning: Removed 4 rows containing missing values (‘geom_point()‘).

## Warning: Removed 1790 rows containing missing values (‘geom_point()‘).
```

```
## Warning: Removed 6 rows containing missing values (‘geom_point()‘).

## Warning: Removed 9 rows containing missing values (‘geom_point()‘).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 7 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 31 rows containing missing values

## Warning: Removed 7 rows containing missing values (‘geom_point()‘).

## Warning: Removed 9 rows containing missing values (‘geom_point()‘).

## Warning: Removed 40 rows containing missing values (‘geom_point()‘).

## Warning: Removed 10 rows containing missing values (‘geom_point()‘).

## Warning: Removed 7 rows containing missing values (‘geom_point()‘).

## Warning: Removed 1790 rows containing missing values (‘geom_point()‘).

## Warning: Removed 10 rows containing missing values (‘geom_point()‘).

## Warning: Removed 13 rows containing missing values (‘geom_point()‘).

## Warning: Removed 7 rows containing missing values (‘geom_point()‘).

## Warning: Removed 7 rows containing non-finite values (‘stat_density()‘).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 32 rows containing missing values

## Warning: Removed 31 rows containing missing values (‘geom_point()‘).
## Removed 31 rows containing missing values (‘geom_point()‘).

## Warning: Removed 59 rows containing missing values (‘geom_point()‘).

## Warning: Removed 32 rows containing missing values (‘geom_point()‘).

## Warning: Removed 31 rows containing missing values (‘geom_point()‘).

## Warning: Removed 1795 rows containing missing values (‘geom_point()‘).

## Warning: Removed 32 rows containing missing values (‘geom_point()‘).

## Warning: Removed 33 rows containing missing values (‘geom_point()‘).
```

```
## Warning: Removed 31 rows containing missing values (‘geom_point()‘).

## Warning: Removed 32 rows containing missing values (‘geom_point()‘).

## Warning: Removed 31 rows containing non-finite values (‘stat_density()‘).
```