# STAT230 Work Team Project: Proposal & EDA

Work Team 7: Evan, Becca, Laith

due Thursday, October 12th

## Introduction & Motivation

We are interested in how we can attempt to predict the profitability of farming by different farming practices. Specifically we are curious about sustainable agriculture decisions and decisions about crop types. From a conservationist perspective, it is important that sustainable choices are economically viable for farmers.

https://www.nass.usda.gov/AgCensus/

Our data is sourced from the United States Census of Agriculture, between the years 2009 and 2017, along with county demographic information from Wikipedia, sourced from the U.S Census.

Our response variable in reported farm income, and we are planning to predict it with a number of variables (pop_density, farm acres, bee colonies, grazing rotation, fertilizer use)

## 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.

```
## wrangle!!
df0 <- read_csv("gooddata.csv")
```

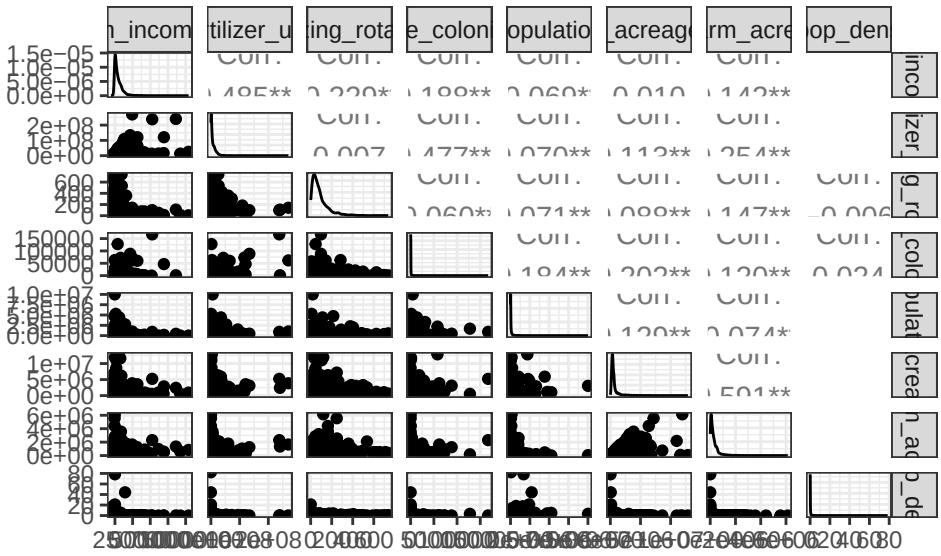
```
## Rows: 3077 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): county_code, honey_sold, state_lower, county
## dbl  (9): xmas_tree_sale, emus_sold, ornamental_fish, bison_sale, aquatic_pl...
## num (13): farm_acres, cropland_acres, 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.
```

```
df0 <- df0 |>
  clean_names()

df <- df0 |>
  select(c(farm_income_1, fertilizer_use, grazing_rotation, bee_colonies, population, county_acreage_20

df1 <- df |>
  mutate(pop_dens = population/county_acreage_2000)
```

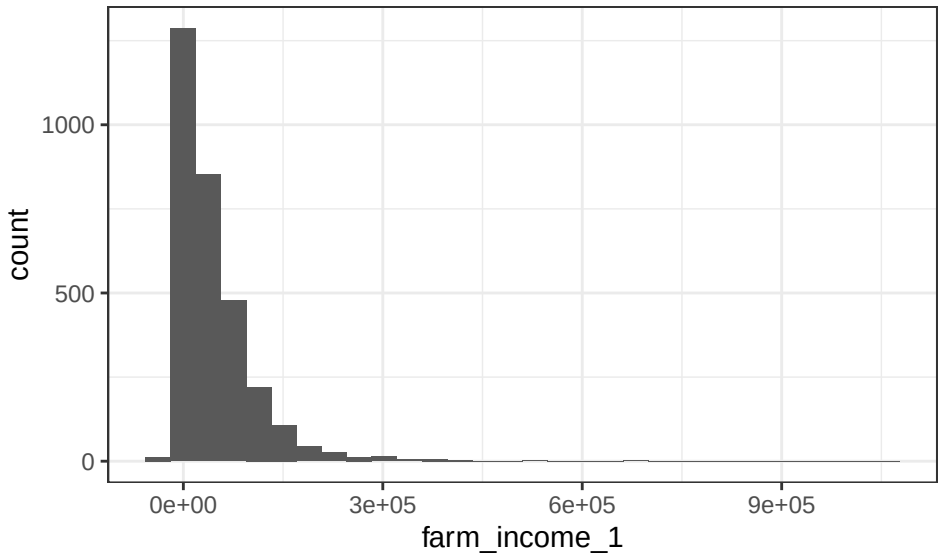
## Exploratory Data Analysis (EDA)

Explore distributions and associations graphically and numerically.

```
ggpairs(df1, columns = 1:8, progress = FALSE)
```




```
ggplot(data = df1, aes(x = farm_income_1))+
  geom_histogram()
```

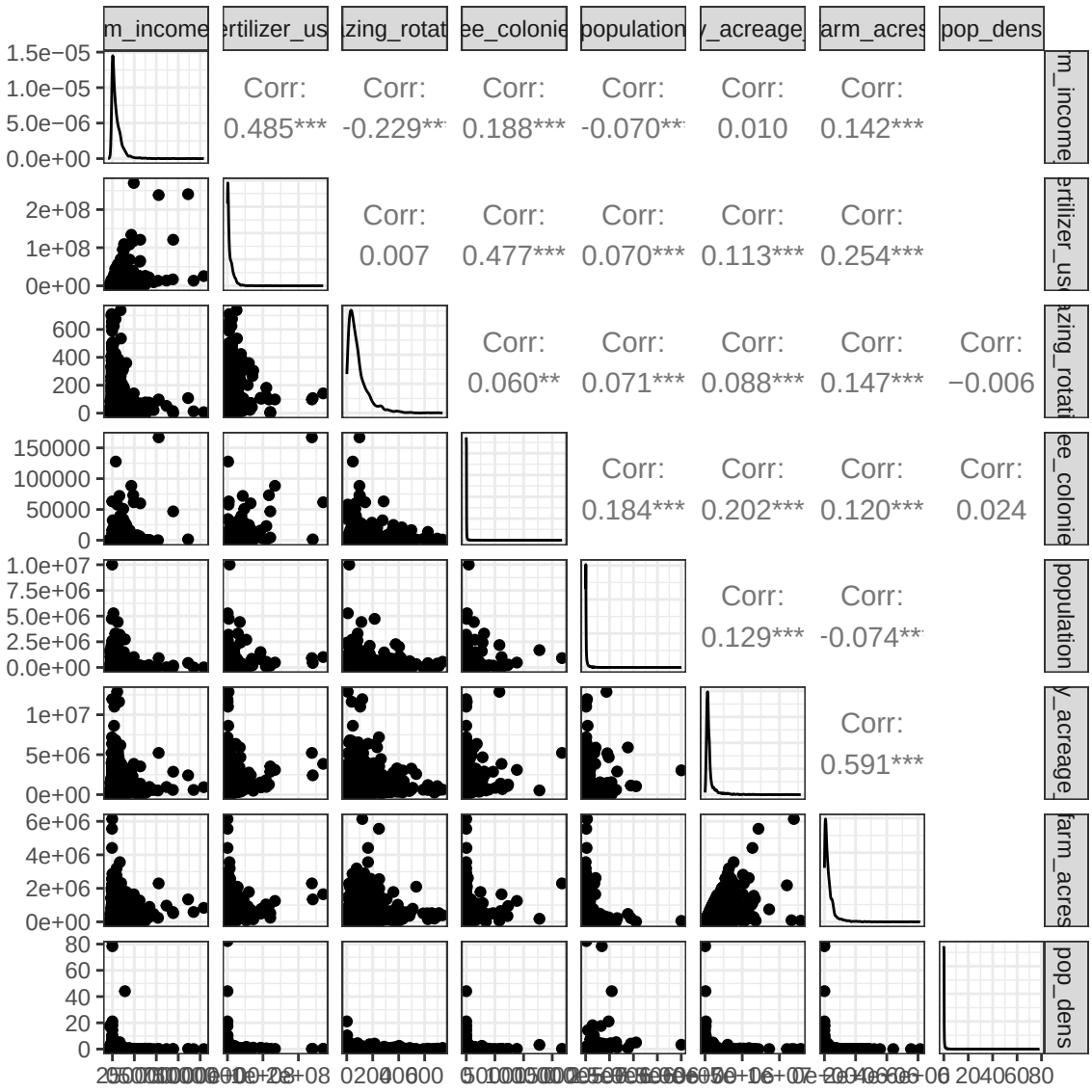



```
#looks like farm income could maybe use a transformation to improve linearity.
```

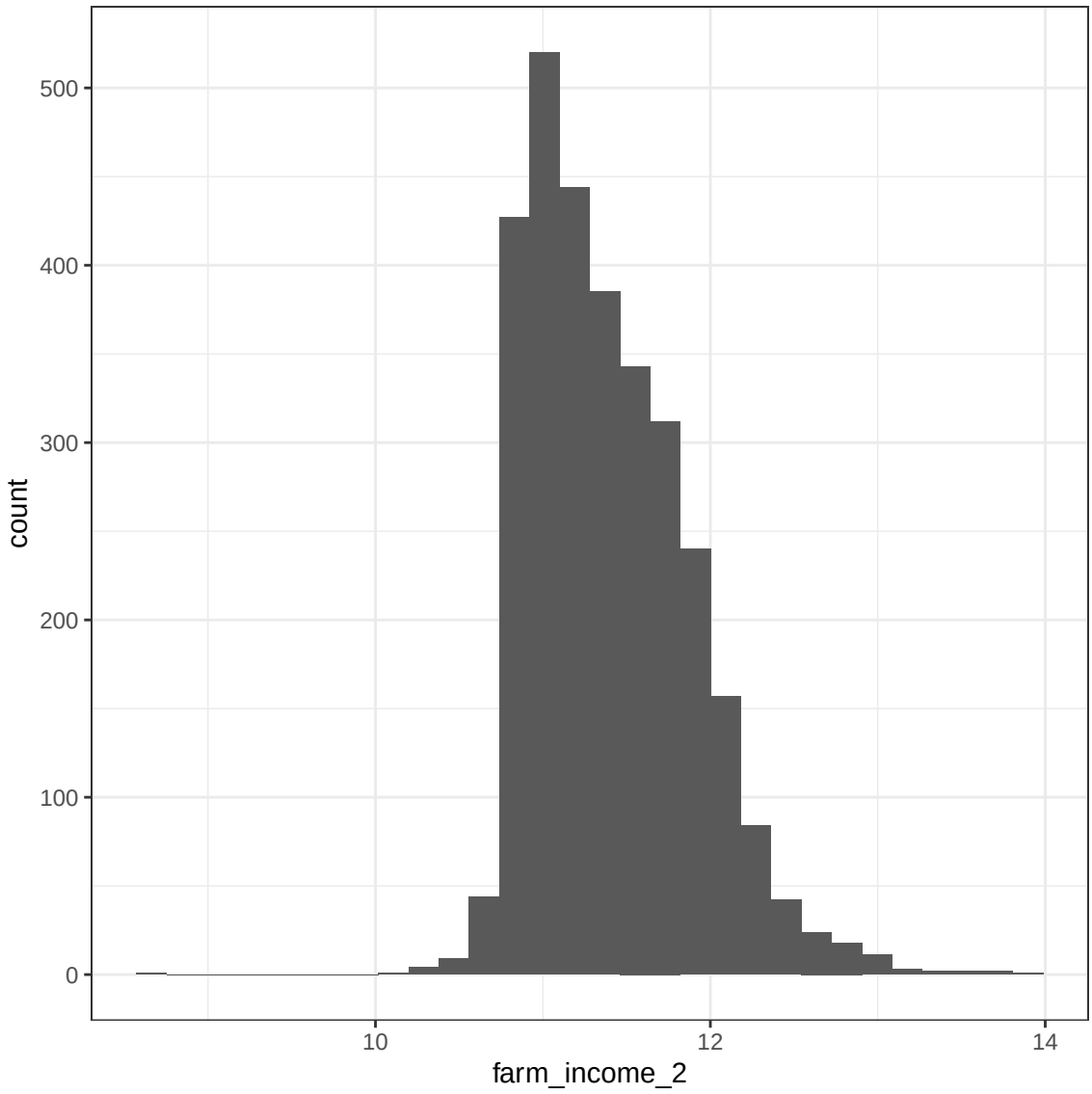
```
df2 <- df1 |>
  mutate(farm_income_2 = log(farm_income_1+55152))
```

```r
#we had one value that was very low because it was by far the smallest, even after we transposed the da
df2 <- df2 |>
  filter(farm_income_2 > 5)
```

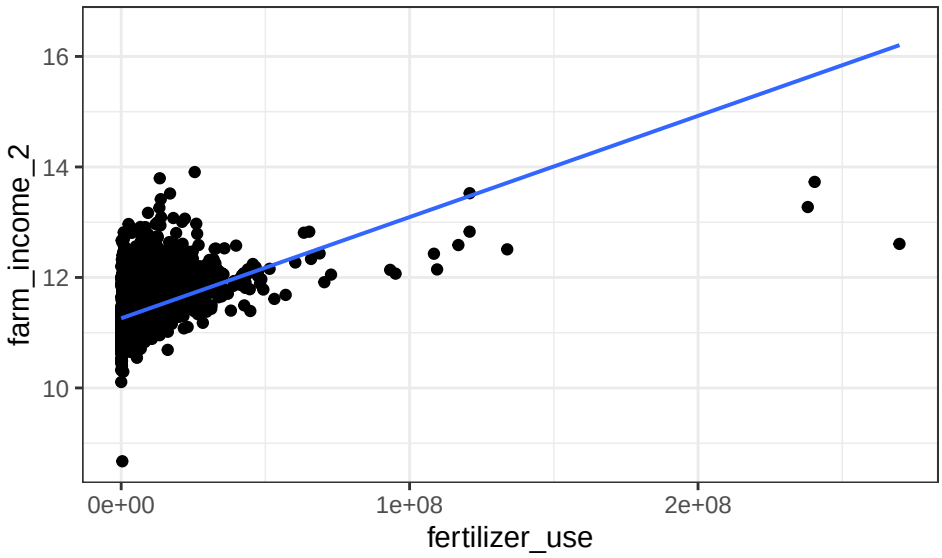
```r
ggpairs(df2, columns = 1:8, progress = FALSE)
```



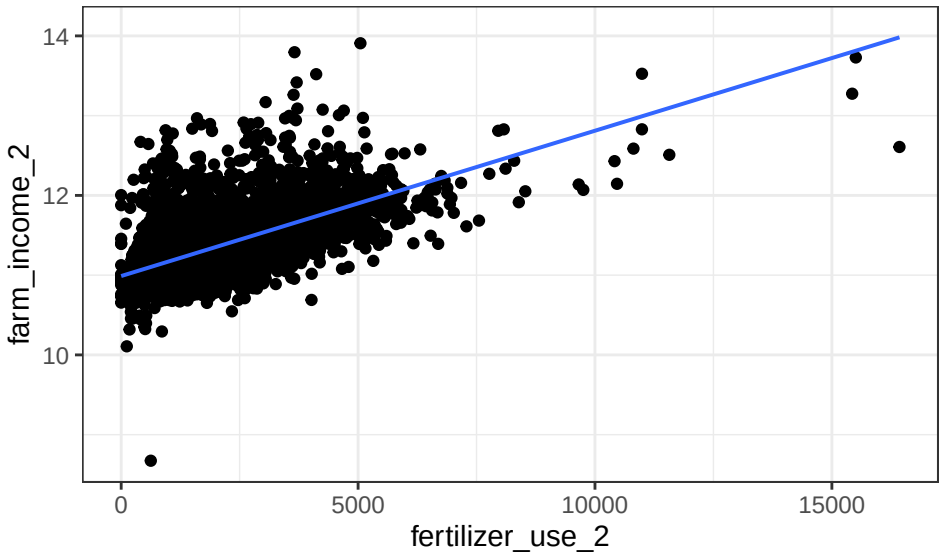

```r
ggplot(data = df2, aes(x = farm_income_2))+
  geom_histogram(bins = 30)
```




```
gf_point(data = df2, farm_income_2 ~ fertilizer_use) |>
  gf_lm()
```

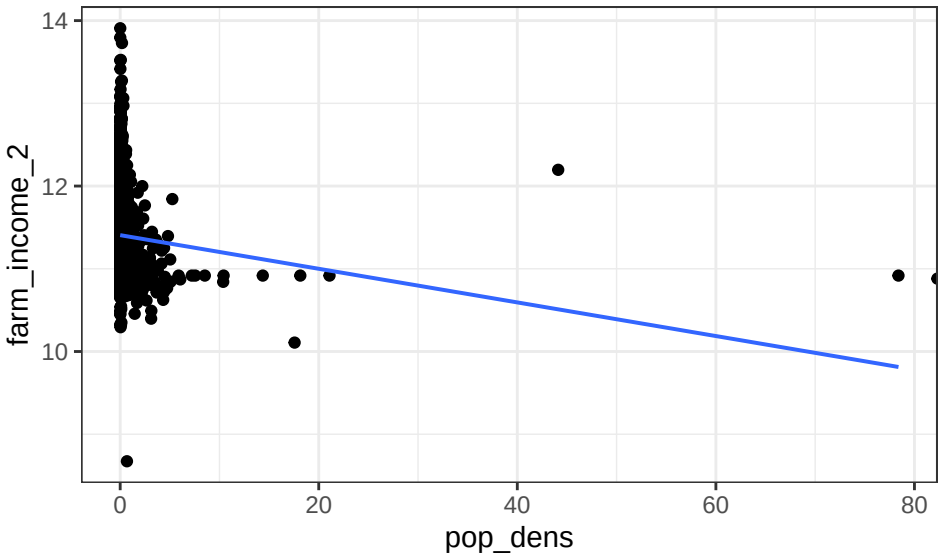



```
#fertilizer use is super right skewed and not linear w/ farm income, so trying a transformation

df3 <- df2 |>
  mutate(fertilizer_use_2 = sqrt(fertilizer_use))

gf_point(data = df3, farm_income_2 ~ fertilizer_use_2) |>
  gf_lm()
```

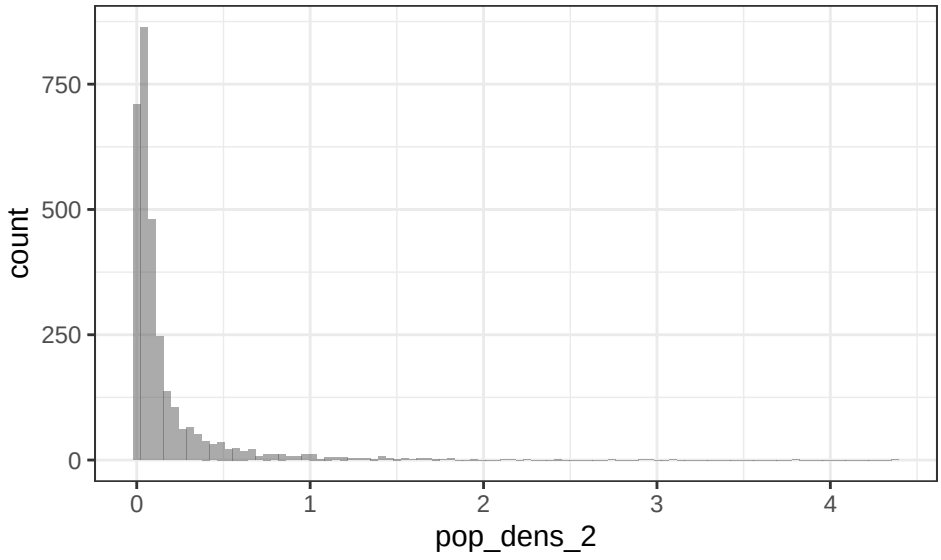



```
#looks much better now
```

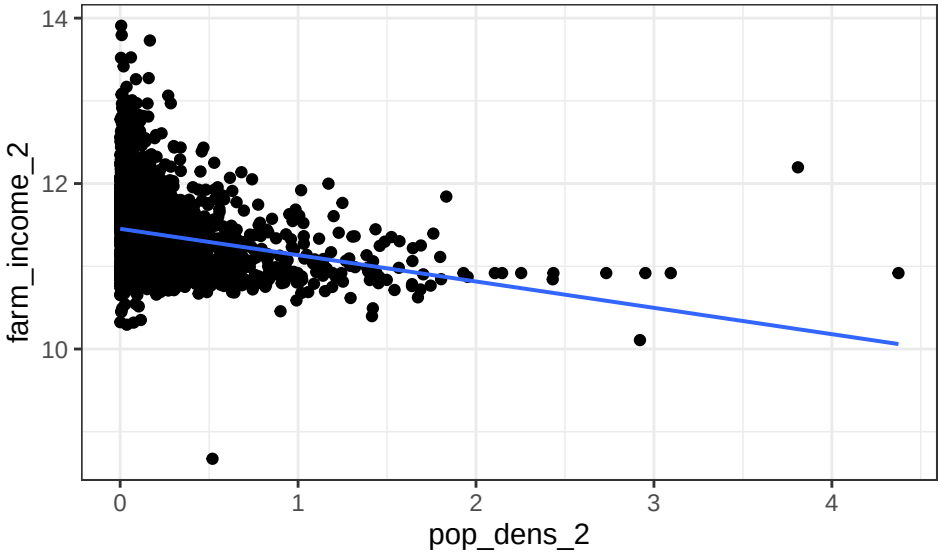
```
gf_point(data = df3, farm_income_2 ~ pop_dens) |>
  gf_lm()
```



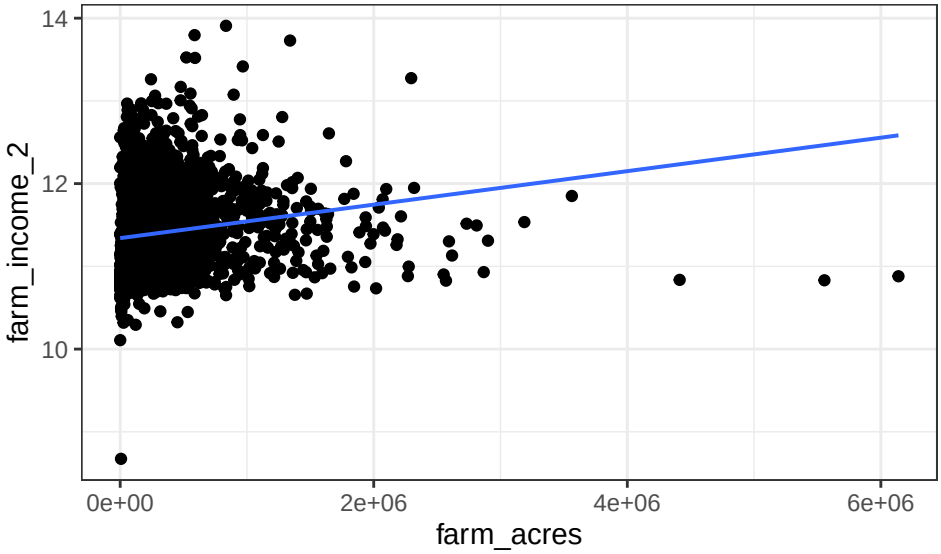

```
#population density is super right skewed and not linear w/ farm income, so trying a transformation

df3 <- df3 |>
  mutate(pop_dens_2 = log(pop_dens+1))


#again looks a lot better

#looks like we still have a value that is inf because there is somehow 0 acres in the county, just goin

df3 <- df3 |>
  filter(!is.infinite(pop_dens_2))

gf_histogram(data = df3, ~ pop_dens_2, bins = 100)
```



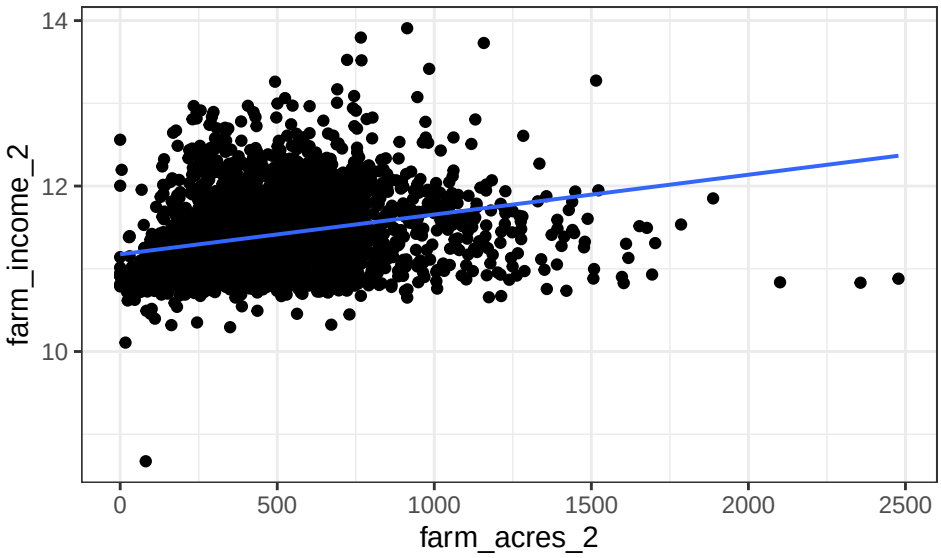
```
gf_point(data = df3, farm_income_2 ~ pop_dens_2) |>
  gf_lm()
```




```
gf_point(data = df3, farm_income_2 ~ farm_acres) |>
  gf_lm()
```

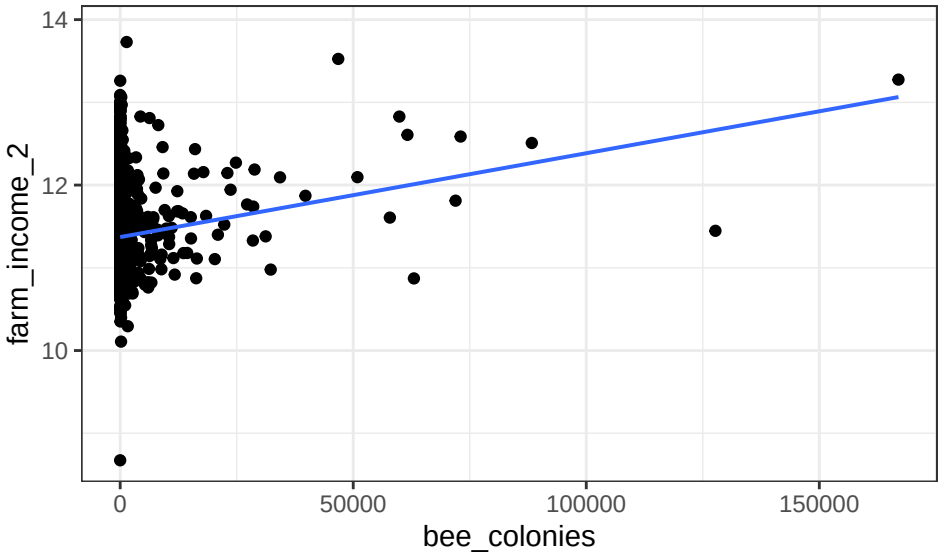



```
#farm acres is super right skewed and not linear w/ farm income, so trying a transformation

df3 <- df3 |>
  mutate(farm_acres_2 = sqrt(farm_acres))

gf_point(data = df3, farm_income_2 ~ farm_acres_2) |>
  gf_lm()
```

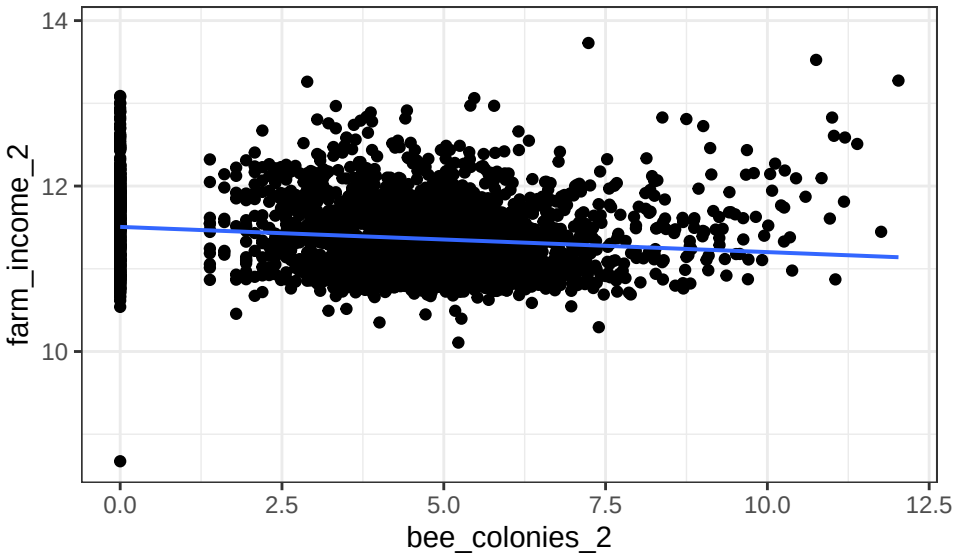



```
#looks better, but still pretty questionable
```

```
gf_point(data = df3, farm_income_2 ~ bee_colonies) |>
  gf_lm()
```

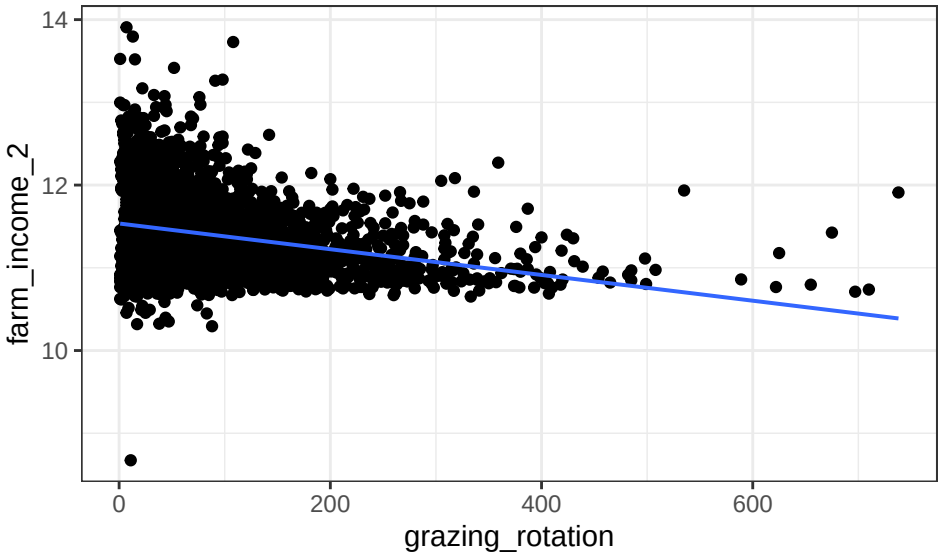



```
#bee colonies is super right skewed and not linear w/ farm income, so trying a transformation

df3 <- df3 |>
  mutate(bee_colonies_2 = log(bee_colonies+1))

gf_point(data = df3, farm_income_2 ~ bee_colonies_2) |>
  gf_lm()
```

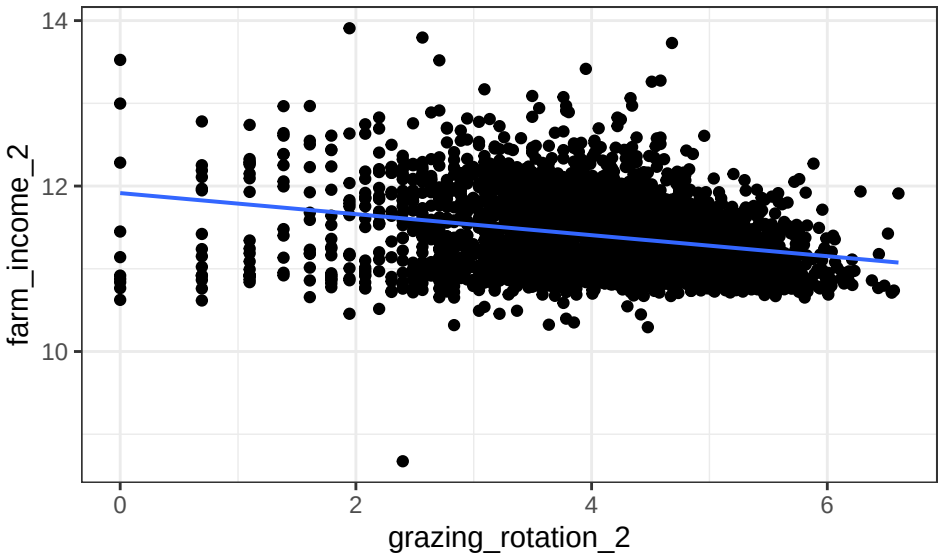



```
#again, much improved
```

```
gf_point(data = df3, farm_income_2 ~ grazing_rotation) |>
  gf_lm()
```




```
#looks a little problematic, give something a shot

df3 <- df3 |>
  mutate(grazing_rotation_2 = log(grazing_rotation))

gf_point(data = df3, farm_income_2 ~ grazing_rotation_2) |>
  gf_lm()
```




```
#looks maybe a little better I guess
```

```
df4 <- df3 |>
  select(c(farm_income_2, farm_acres_2, pop_dens_2, grazing_rotation_2, fertilizer_use_2, bee_colonies_2))
ggpairs(df4, columns = 1:6, progress = FALSE)
```