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")
```

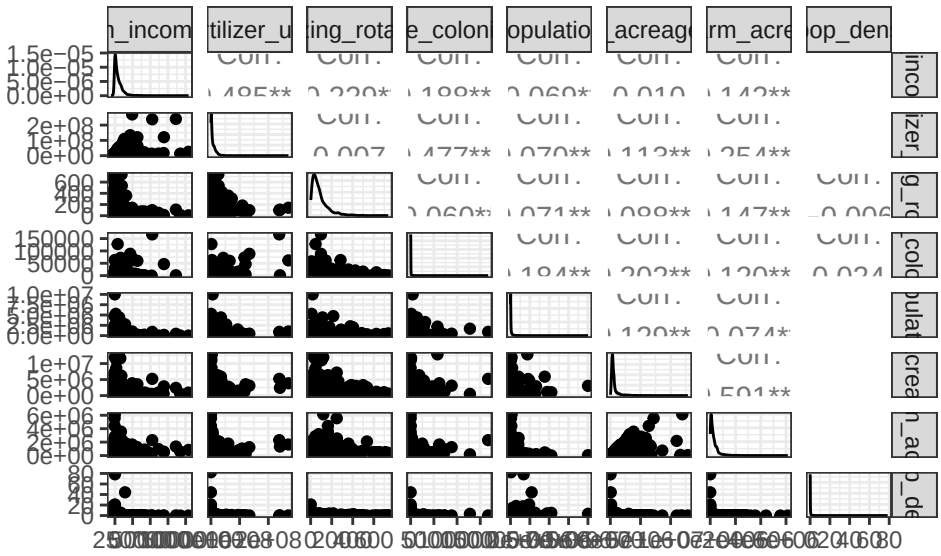
```
## Rows: 3077 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): county_code, honey_sold, state_lower, county
## dbl  (9): xmas_tree_sale, emus_sold, ornamental_fish, bison_sale, aquatic_pl...
## num (13): farm_acres, cropland_acres, 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 |>
  dplyr::select(c(farm_income_1, fertilizer_use, grazing_rotation, bee_colonies, population, county_acre

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

Exploratory Data Analysis (EDA)

Explore distributions and associations graphically and numerically.

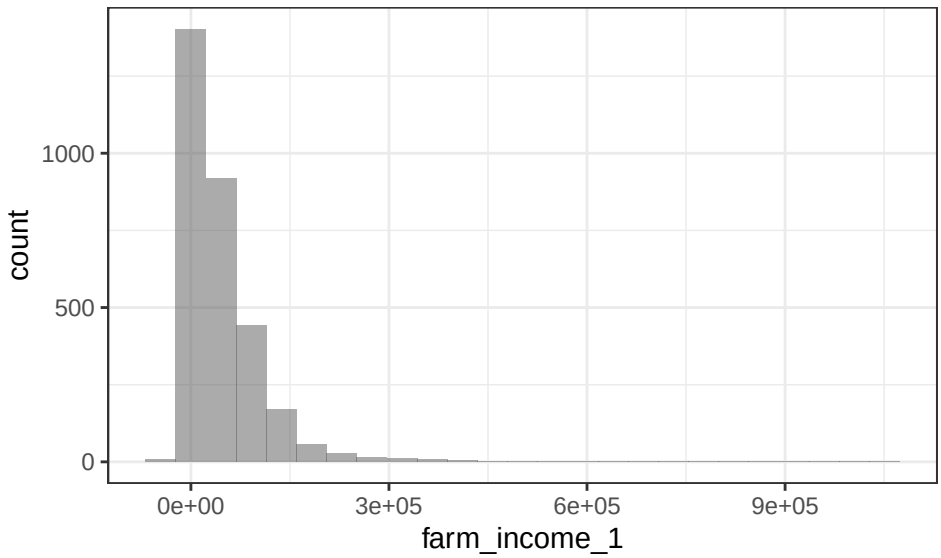
```
ggpairs(df1, columns = 1:8, progress = FALSE)
```



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

There are significant correlations with all variables, but the linearity looks pretty bad, will try some transformations.

```
gf_histogram(~farm_income_1, data=df1)
```

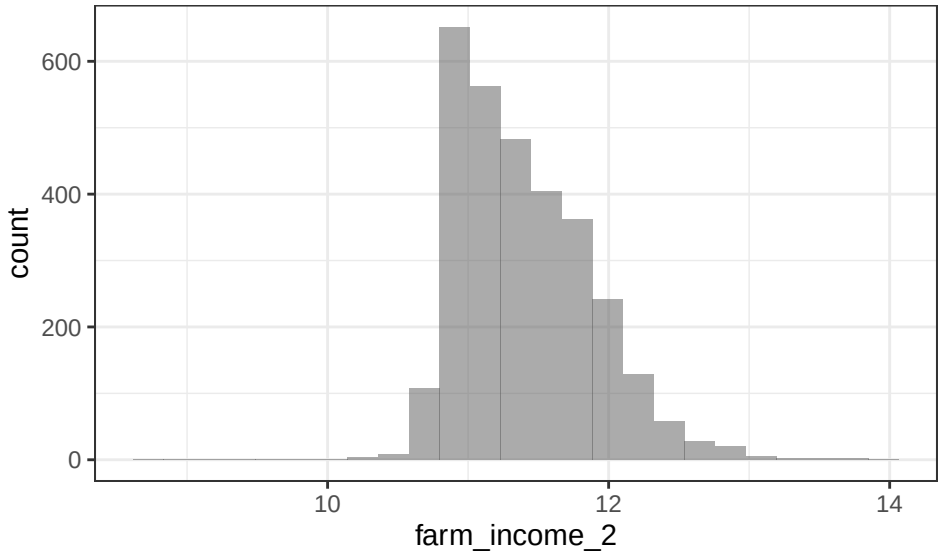


```
favstats(~farm_income_1, data=df1)
```

```
##     min   Q1 median    Q3     max  mean    sd    n missing
##  -55150 5490  27146 66897 1041275 46841 67963 3077       0
```

The mean is 46841, the median is 27146, we can also tell it looks right skewed from the histogram, and the IQR is 61407

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

#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)
gf_histogram(~farm_income_2, data=df2)
```

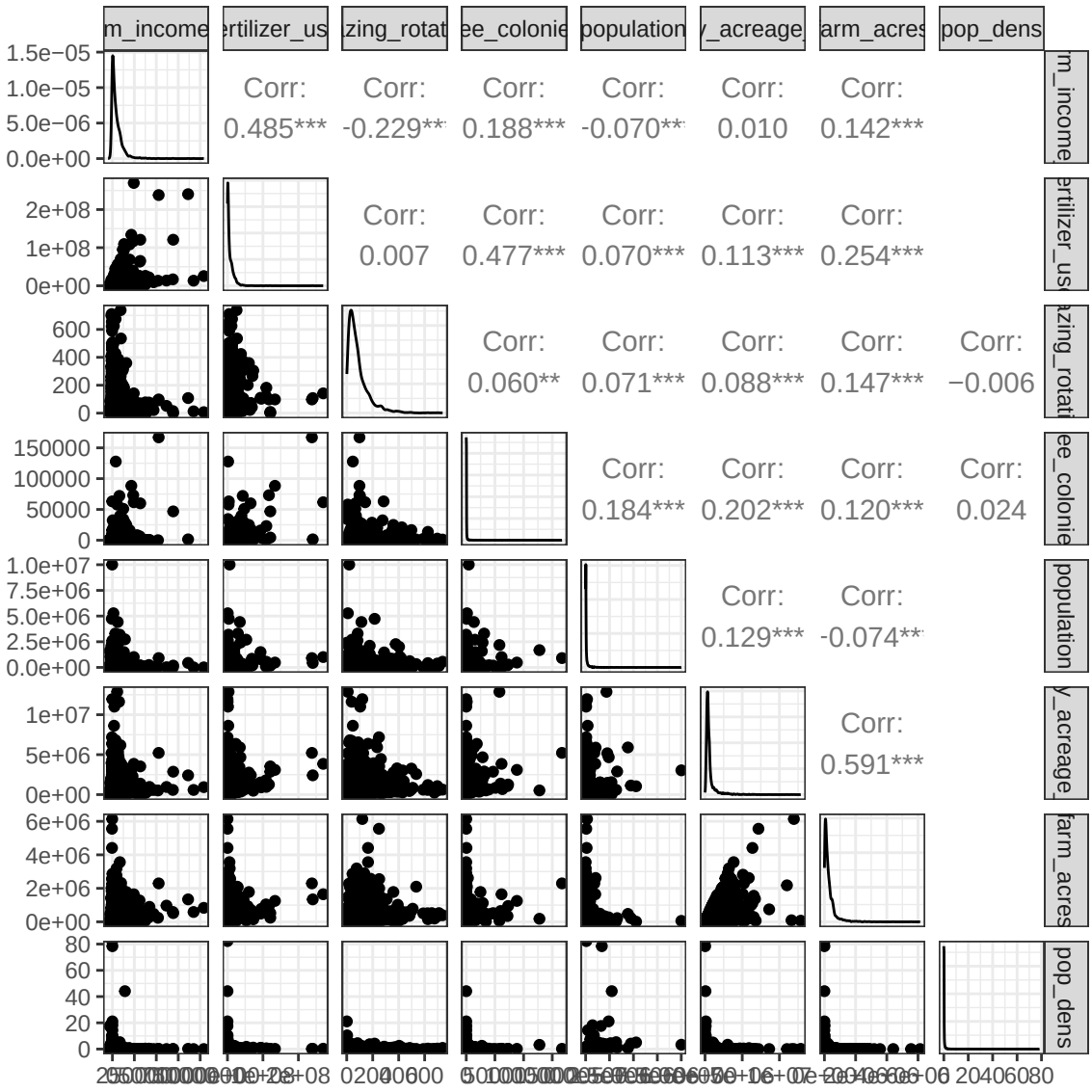



```
favstats(~farm_income_2, data=df2)
```

```
##     min     Q1 median     Q3    max mean      sd    n missing
##  8.6733 11.013 11.318 11.712 13.908 11.4 0.48012 3076       0
```

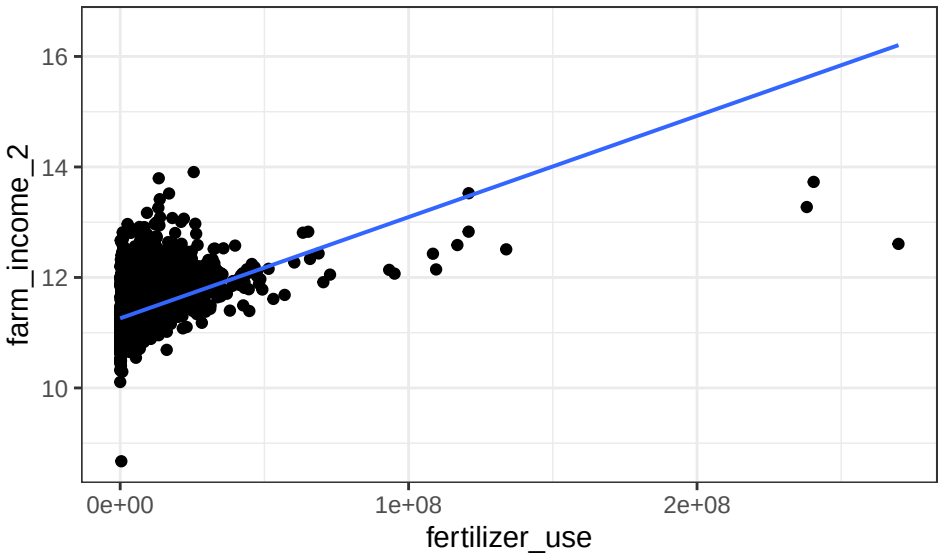
Looks much improved, the mean is 11.4, and median is 11.318, and the histogram still is not ideal, but better. The IQR = 0.699

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

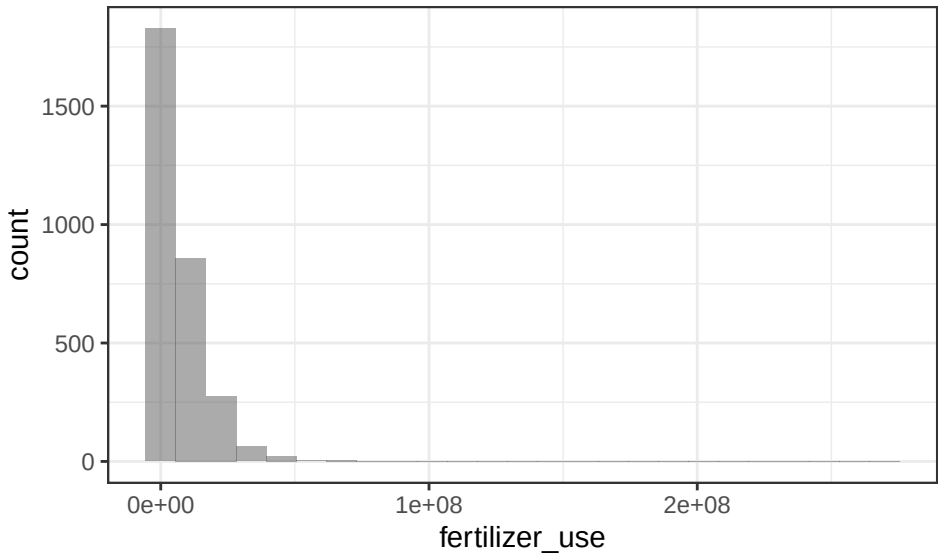


This improved linearity slightly, but still questionable, lets look at predictors.

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




```
gf_histogram(~fertilizer_use, data=df2)
```



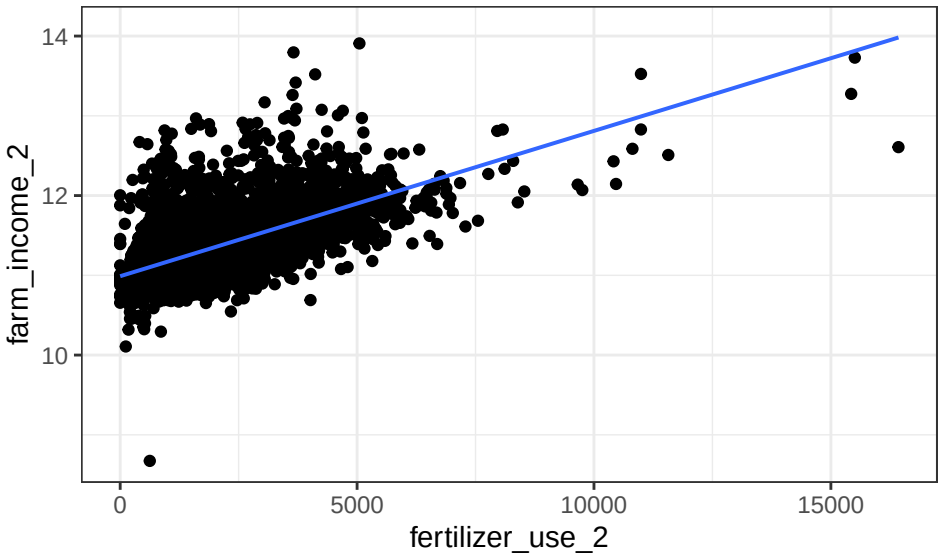

```
favstats(~fertilizer_use, data=df2)
```

```
##  min      Q1  median       Q3       max    mean       sd    n missing
##    0 1017000 3679500 10741500 269837000 7664465 12752566 3070       6
```

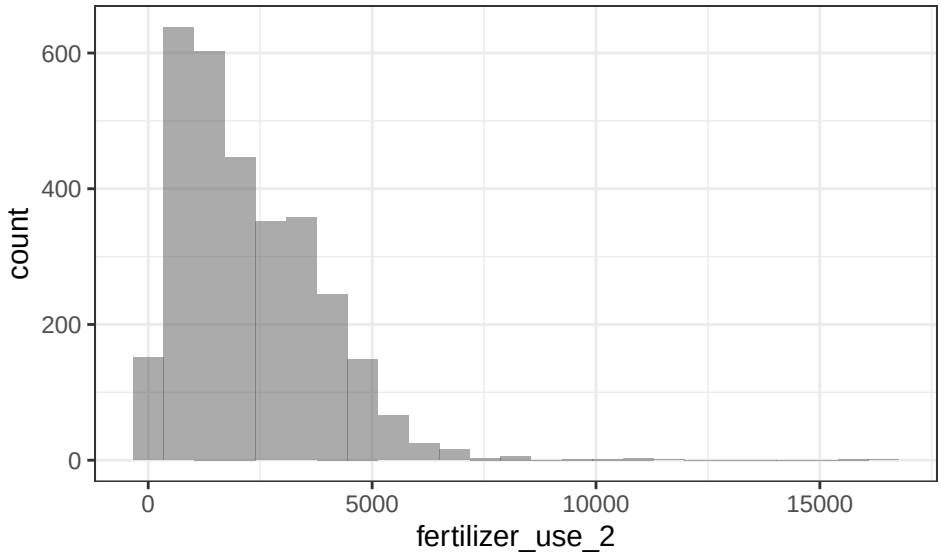
```
#fertilizer use is super right skewed and not linear w/ farm income, so trying a transformation
```

This does not look great, heavily skewed distribution and it is difficult to assess the scatterplot, the mean is 7664465, the median is 3679500, and the IQR is 9724500. Will try transforming.

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

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




```
gf_histogram(~fertilizer_use_2, data=df3)
```

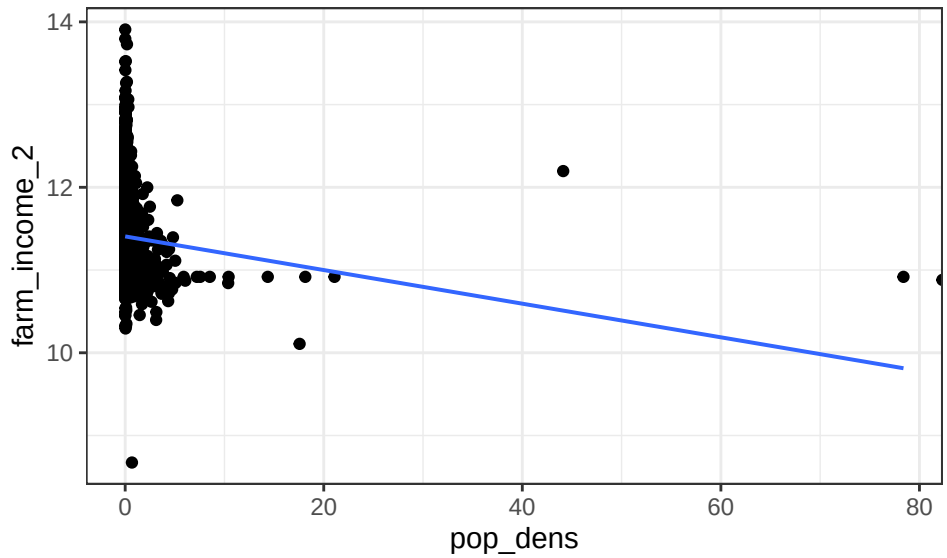



```
favstats(~fertilizer_use_2, data=df3)
```

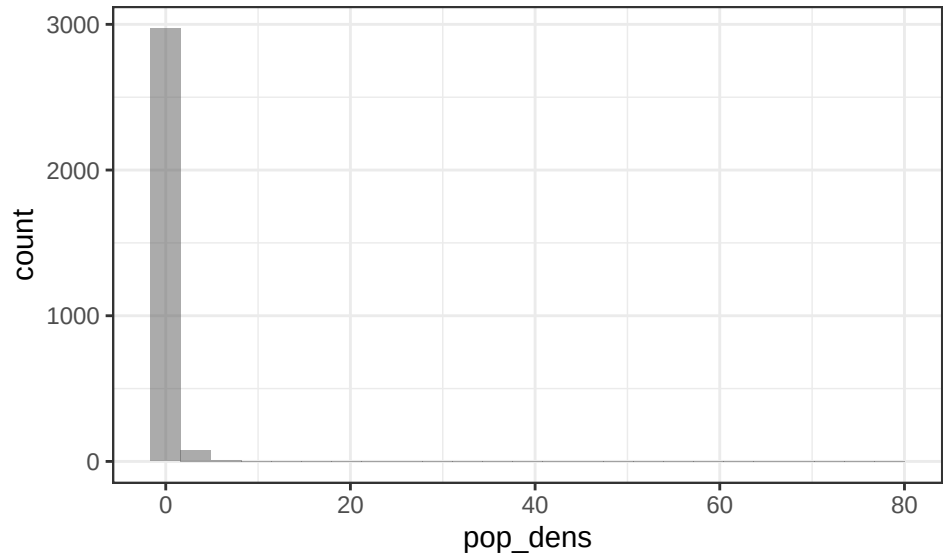
```
##  min     Q1 median     Q3   max    mean     sd    n missing
##    0 1008.5 1918.2 3277.4 16427 2260.6 1598.5 3070       6
```

Once again, this improves how the data looks. The distribution is still somewhat right skewed, with a mean of 2260.6 and a median of 1918.2, and an IQR of 2268.9, but this is an improvement, but linearity looks better based on the scatterplot.

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




```
gf_histogram(~pop_dens, data=df3)
```



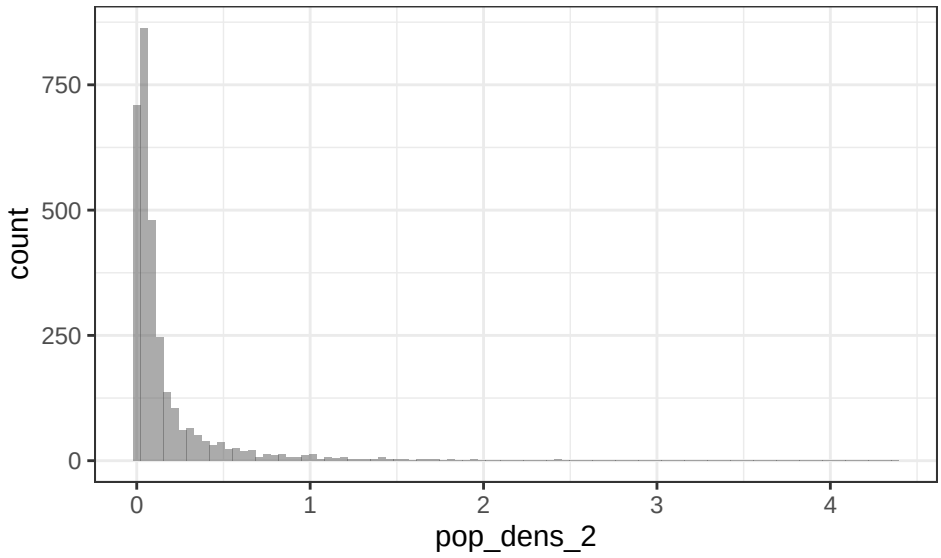

```
favstats(~pop_dens, data=df3)
```

```
##         min       Q1   median       Q3 max mean  sd    n missing
##  0.00014774 0.024769 0.064536 0.16677 Inf  Inf NaN 3066      10
```

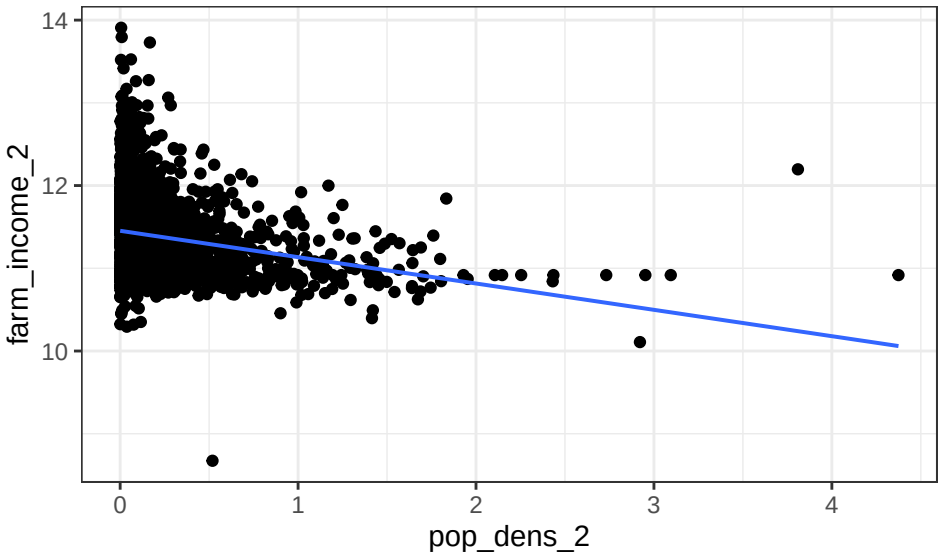
```
#population density is super right skewed and not linear w/ farm income, so trying a transformation
```

This one is also not great, scatterplot not very useful, the histogram is right skewed, the mean is 0.29915 and median is 0.064434, and the IQR is 0.14179.

```r
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)
```




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

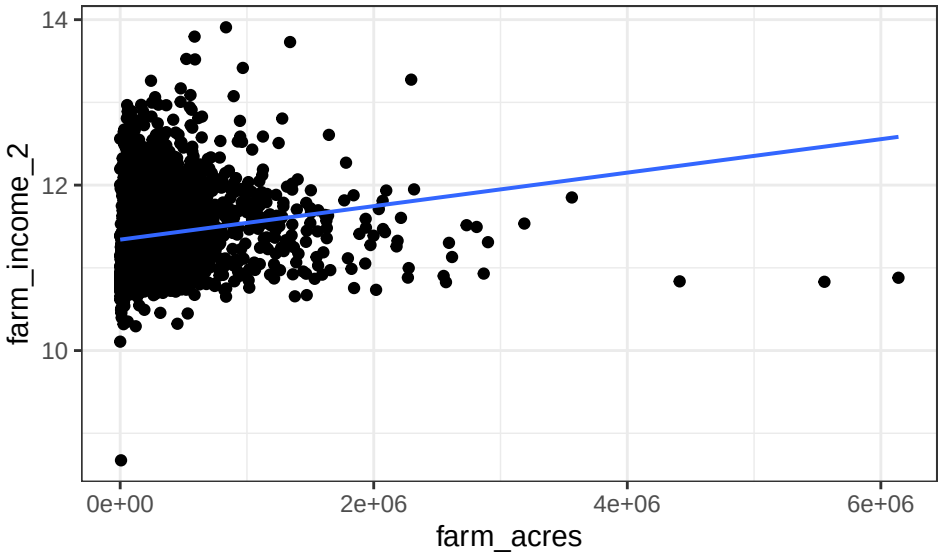


```
favstats(~pop_dens_2, data=df3)
```

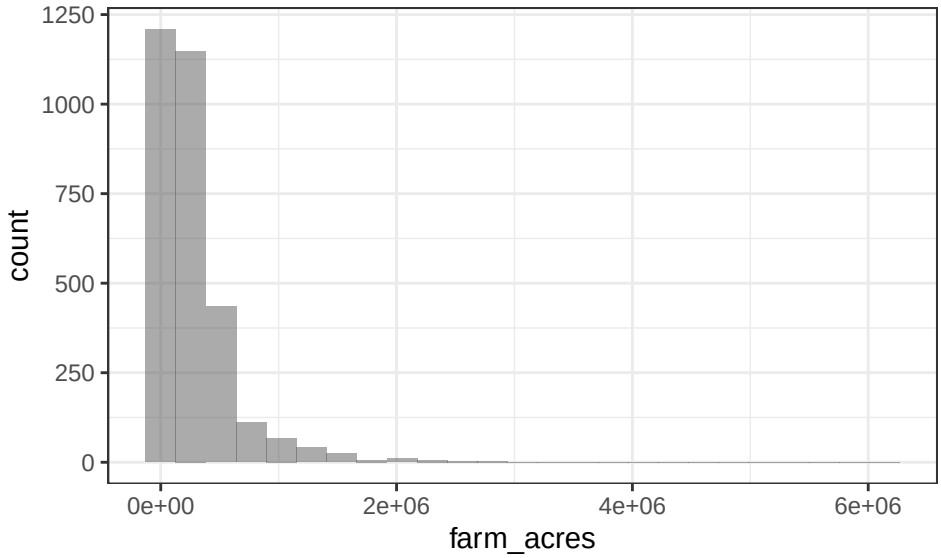
```
##         min       Q1   median       Q3    max    mean      sd    n missing
##  0.00014773 0.024438 0.062443 0.15404 4.3744 0.16482 0.30604 3065      10
```

This one still does not look great after transforming, but it probably helped. The graph is still right skewed, but not as strongly, with a mean of 0.16482 and a median of 0.062443, and IQR of 0.1296, and the relationship based on the scatter plot is not very useful.

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




```
gf_histogram(data = df3, ~ farm_acres)
```



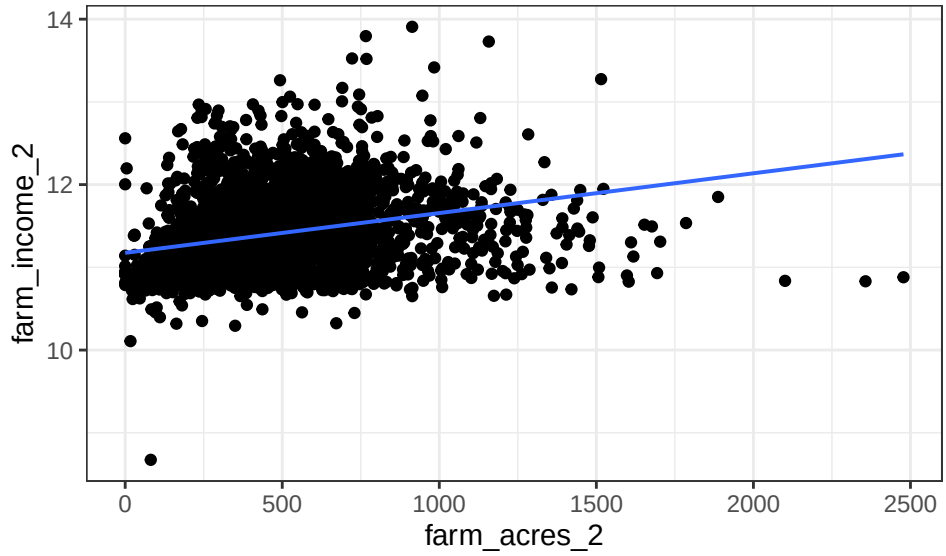
```
favstats(~farm_acres, data=df3)
```

```
##  min    Q1 median     Q3     max   mean     sd    n missing
##    0 77279 179306 359475 6139007 291893 383921 3075       0
```

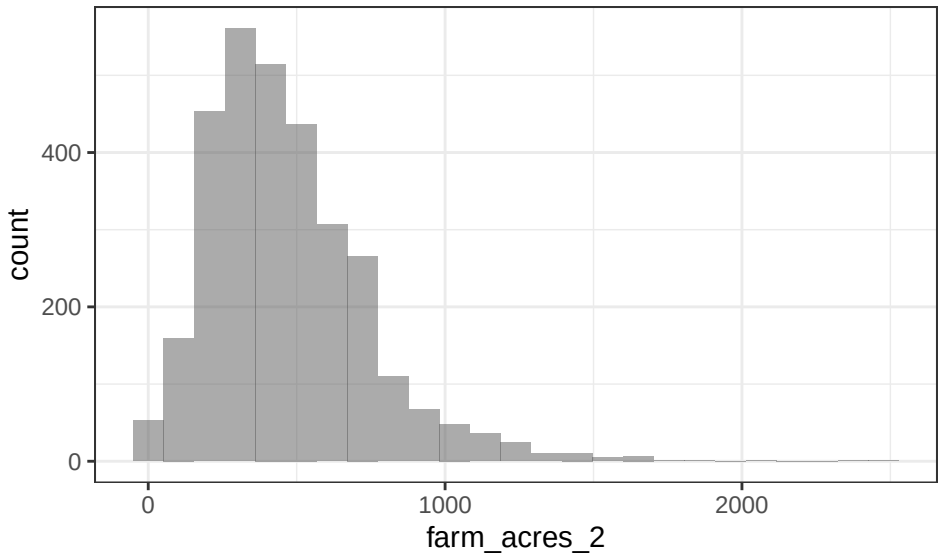
```
#farm acres is super right skewed and not linear w/ farm income, so trying a transformation
```

Once again, unhelpful scatterplot, right skewed distribution, mean of 291893, median of 179306 and IQR of 282196

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

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




```
gf_histogram(data = df3, ~ farm_acres_2)
```



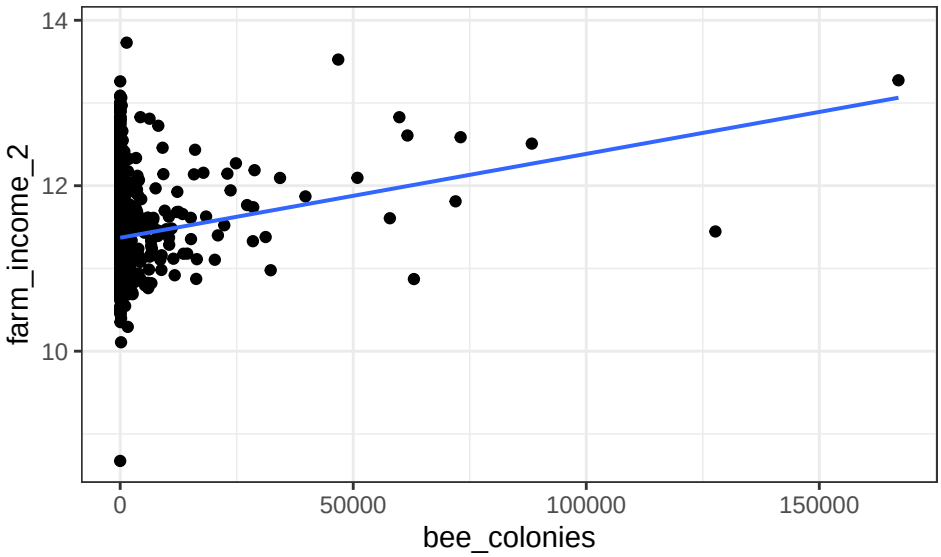

```
favstats(~farm_acres_2, data=df3)
```

```
##  min     Q1 median     Q3    max   mean     sd    n missing
##    0 277.99 423.45 599.56 2477.7 468.11 269.79 3075       0
```

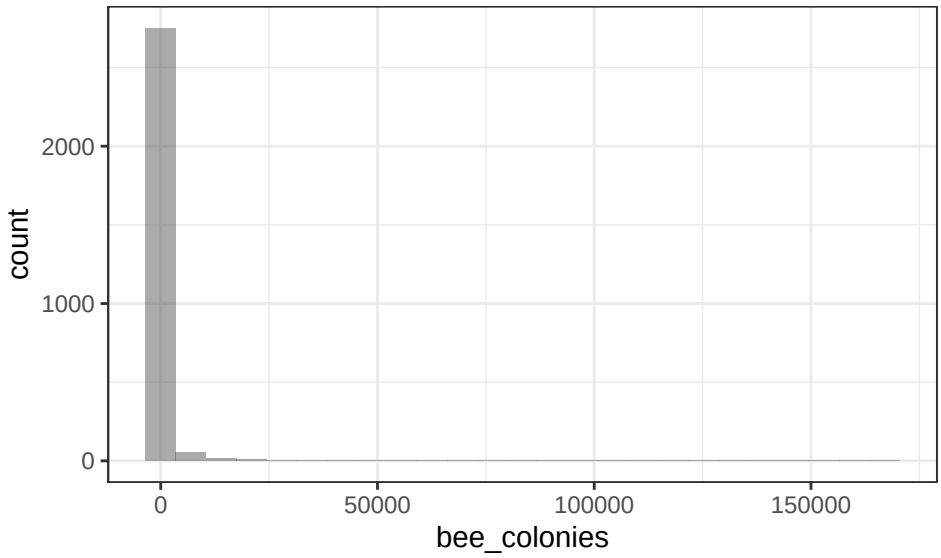
```
#looks better, but still pretty questionable
```

This transformation improved it somewhat, the histogram is less right skewed, but the relationship does not look very linear based on the scatter plot, the mean is 468.11 and the median is 423.45, and the IQR is 321.57

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



```
gf_histogram(data = df3, ~ bee_colonies)
```



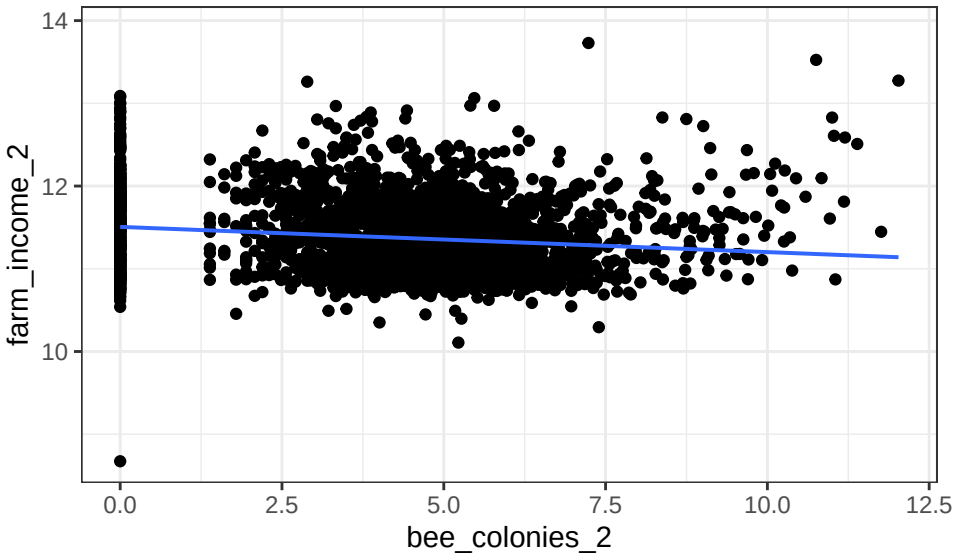

```
favstats(~bee_colonies, data=df3)
```

```
##  min Q1 median  Q3    max   mean     sd    n missing
##    0 22     93 256 166984 848.12 5847.9 2849     226
```

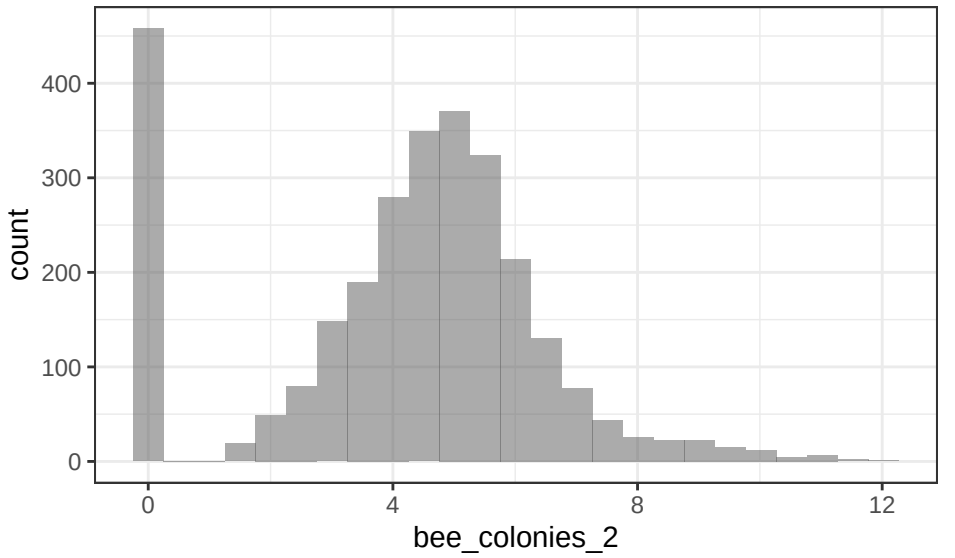
```
#bee colonies is super right skewed and not linear w/ farm income, so trying a transformation
#again, much improved
```

Very strongly right skewed histogram, scatterplot is not useful for assessing linearity, the mean is 848.12, the median is 93, and the IQR is 234

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

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




```
gf_histogram(data = df3, ~ bee_colonies_2)
```

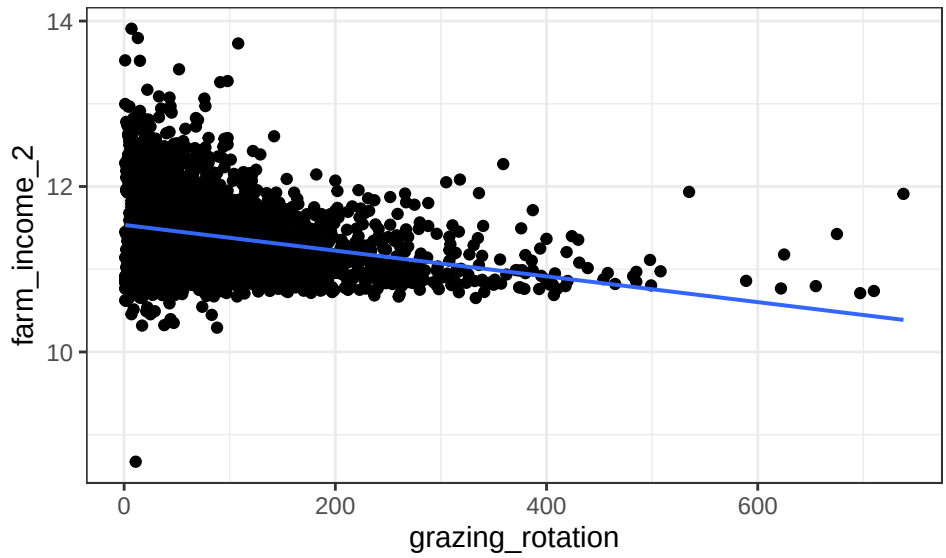



```
favstats(~bee_colonies_2, data=df3)
```

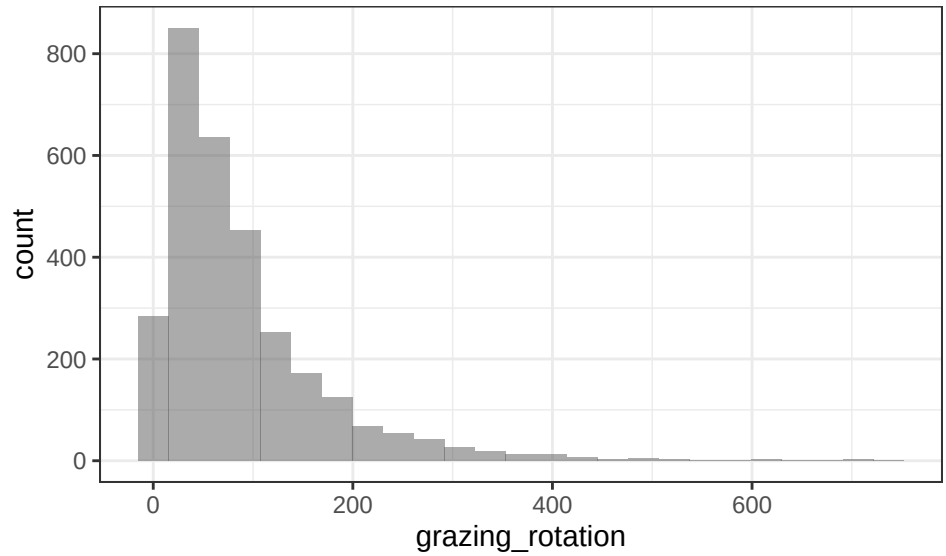
```
##  min     Q1 median     Q3    max   mean    sd    n missing
##    0 3.1355 4.5433 5.5491 12.026 4.1501 2.309 2849     226
```

After transforming the data, the histogram looks somewhat better, the main part of the distribution is very normal, but we have introduced a large number of 0s into the dataset, which represent real values, so probably shouldn't be removed. the mean is 4.1501 the median is 4.5433 and the IQR is 2.4136. The scatter plot similarlly seems to show a fairly linear relationship aside from a large number of 0 values.

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




```
gf_histogram(data = df3, ~ grazing_rotation)
```



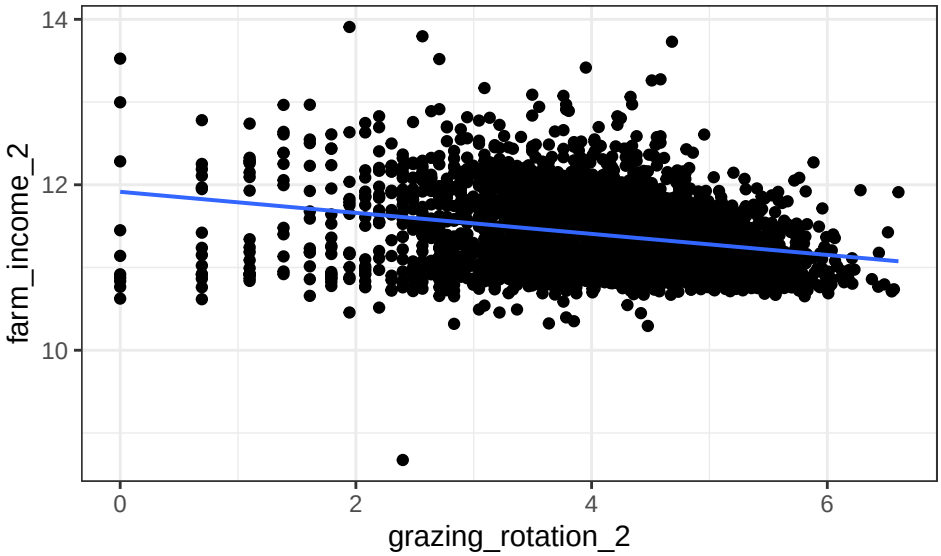

```
favstats(~grazing_rotation, data=df3)
```

```
##  min Q1 median  Q3 max   mean     sd    n missing
##    1 33     63 113 738 87.426 83.229 3034      41
```

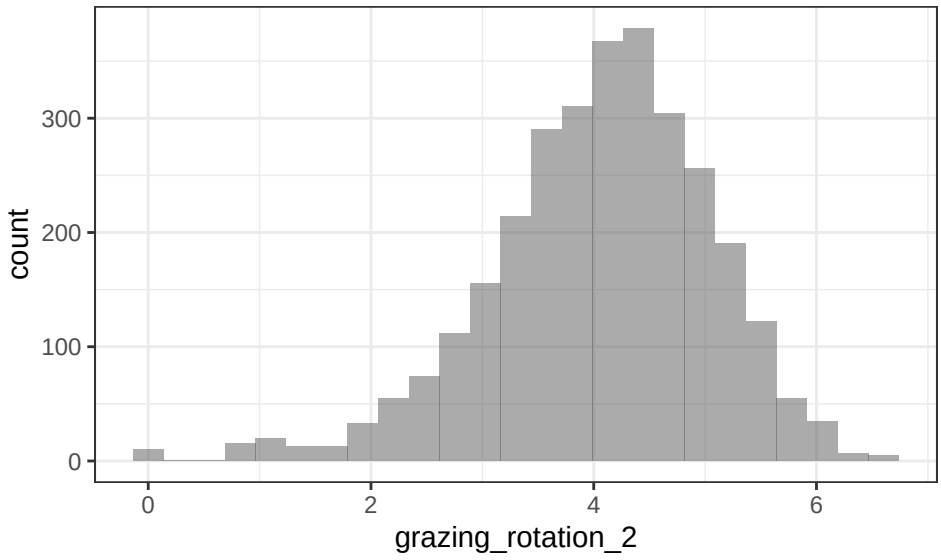
```
#looks a little problematic, give something a shot


#looks maybe a little better I guess
```

Again, a right skewed distribution, and unhelpful scatter plot for assessing linearity. The mean is 87.426, the median is 63, and the IQR is 80. worth trying a transformation.

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

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




```
gf_histogram(data = df3, ~ grazing_rotation_2)
```



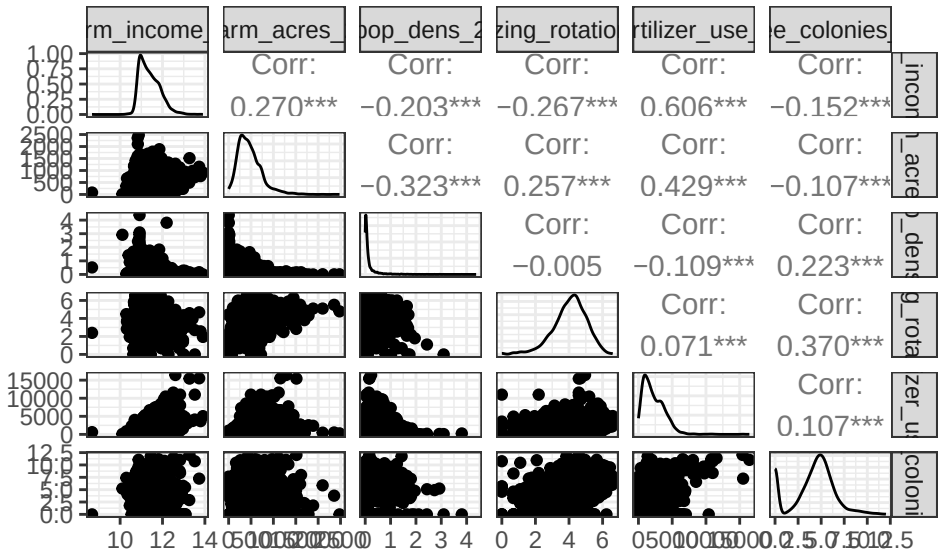

```
favstats(~grazing_rotation_2, data=df3)
```

```
##  min     Q1 median     Q3    max   mean      sd    n missing
##    0 3.4965 4.1431 4.7274 6.6039 4.0537 0.99856 3034      41
```

The scatter plot looks about the same, but the histogram looks much less skewed, the mean is 4.0537, the median is 4.1431, and the IQR is 1.2309.

After we have tried to improve all of the response variables, lets take a look at the ggpairs again

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



These generally look much better, we still have a lot of 0 values in the population density, but they are real, so the data is not actually normally distributed. The scatter plots generally look much more linear than before transforming the variables, and the correlations are similar to before transforming.

```
best.sub <- regsubsets(farm_income_2 ~ farm_acres_2 +
                         pop_dens_2 + grazing_rotation_2 + fertilizer_use_2 + bee_colonies_2, data = df

with(msummary(best.sub), data.frame(adjr2, cp, bic, rss, outmat)) %>%
  kable(digits = 3, booktabs = TRUE) %>%
  row_spec(row = 0, angle = 90)
```

	adjr2	cp	bic	rss	farm_acres_2	pop_dens_2	grazing_rotation_2	fertilizer_use_2	bee_colonies_2
1 (1)	0.386	639.071	-1359.6	363.76				*	
2 (1)	0.471	162.180	-1772.3	313.30			*	*	
3 (1)	0.495	27.609	-1896.9	298.91		*	*	*	
4 (1)	0.500	4.018	-1914.5	296.21		*	*	*	*
5 (1)	0.499	6.000	-1906.5	296.21	*	*	*	*	*