Lecture 7 Notes

2024-07-15

Regression

- Starting with old school data

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts t
o become errors
```

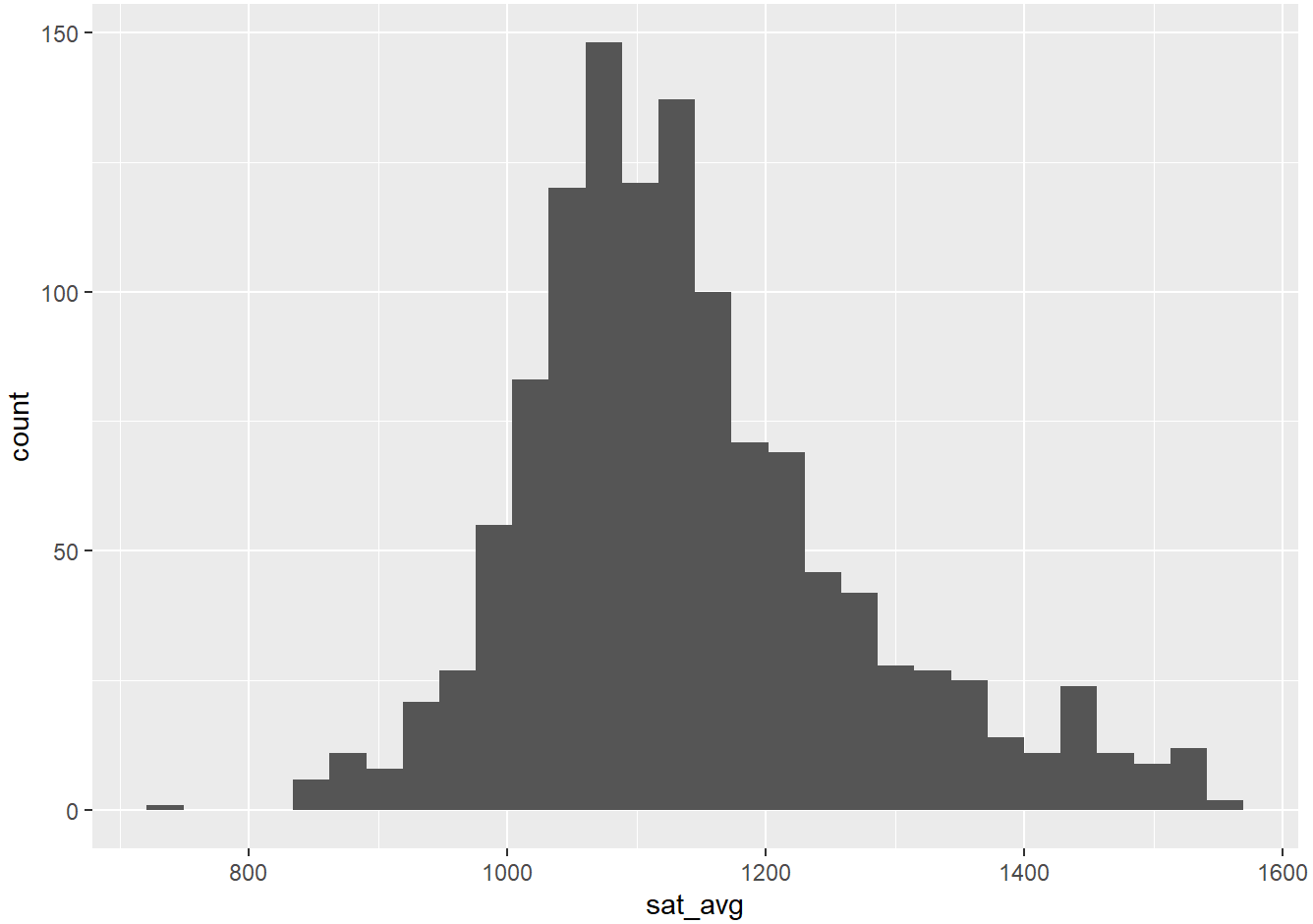
```
debt <- read_rds("https://github.com/jbisbee1/ISP_Data_Science_2024/raw/main/data/sc_deb
t.Rds")
```

Univariate VIsualization

```
# Looking at X variable first
debt %>%
  ggplot(aes(x = sat_avg)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

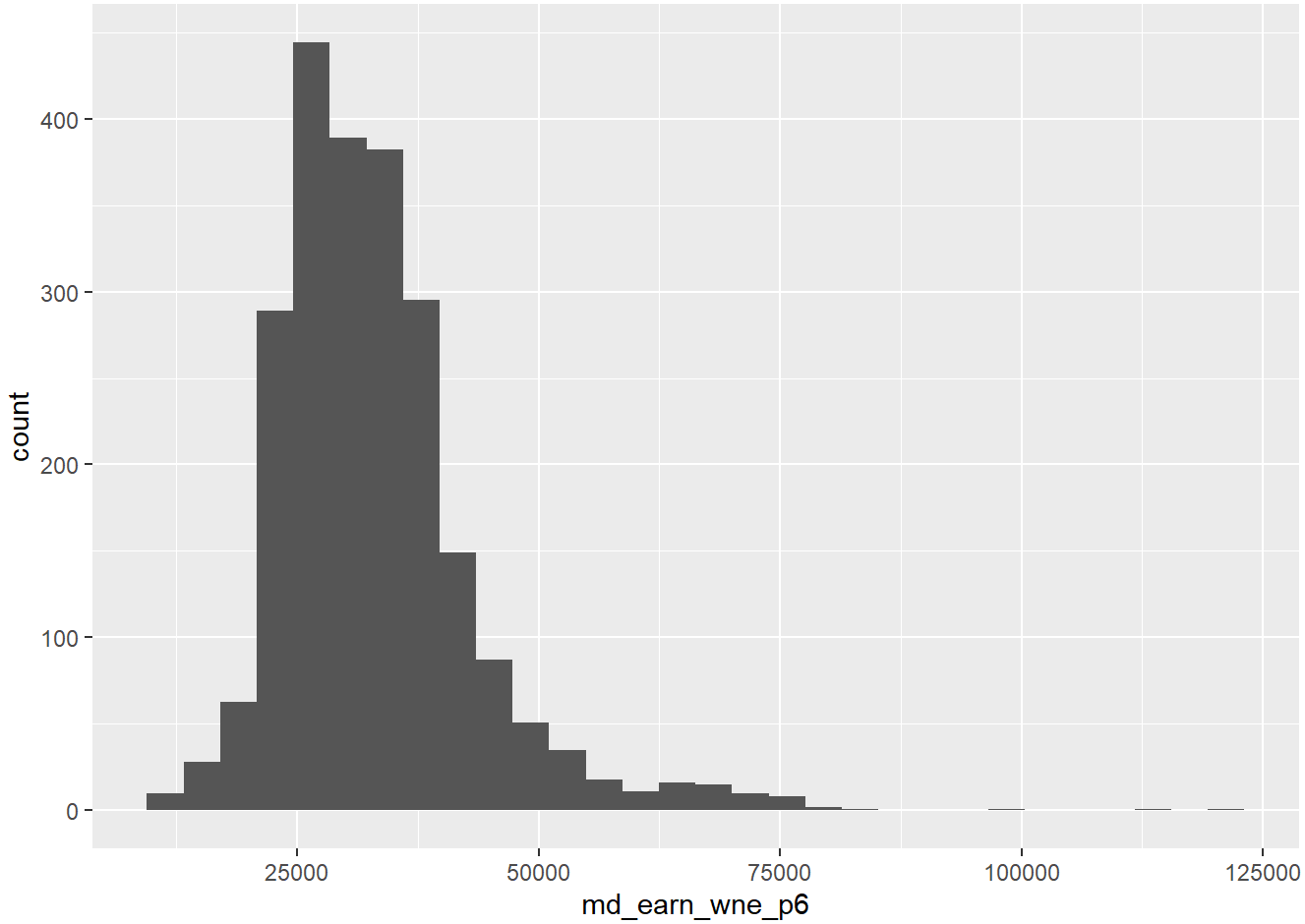
```
## Warning: Removed 1317 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```
# Look at Y variable second
debt %>%
  ggplot(aes(x = md_earn_wne_p6)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 240 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



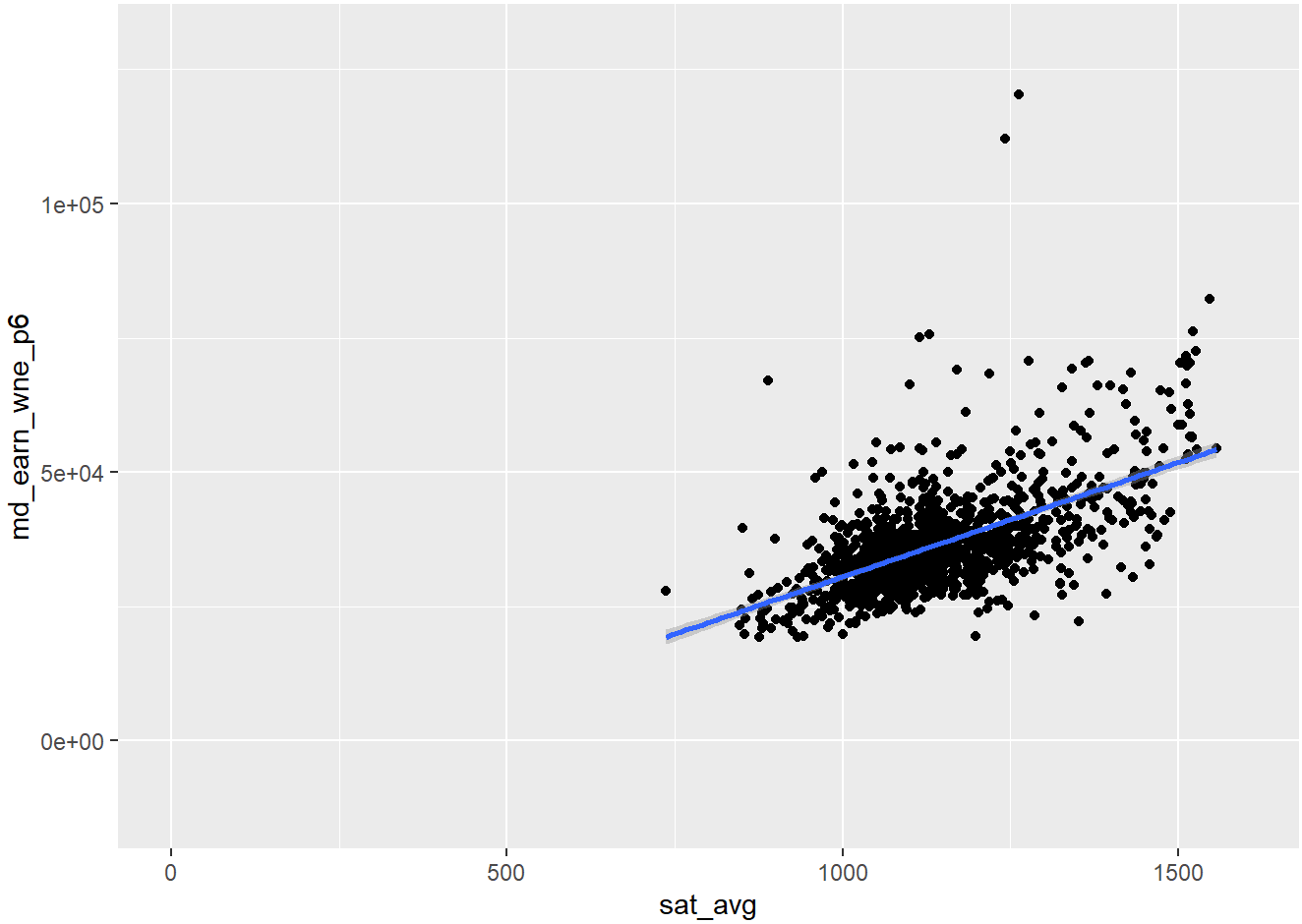
Multivariate Visualization

```
debt %>%
  ggplot(aes(x = sat_avg,
             y = md_earn_wne_p6)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  xlim(c(0,1600)) +
  ylim(c(-13000,130000))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 1348 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1348 rows containing missing values or values outside the scale rang
e
## (`geom_point()`).
```



Introducing lm() function

- Two inputs:

1. `formula`
2. `data`

```
model_1 <- lm(md_earn_wne_p6 ~ sat_avg,debt)

summary(model_1)
```

```
##
## Call:
## lm(formula = md_earn_wne_p6 ~ sat_avg, data = debt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23239  -4311   -852   2893  78695
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12053.87    1939.80  -6.214 7.12e-10 ***
## sat_avg         42.60       1.69  25.203  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7594 on 1196 degrees of freedom
##   (1348 observations deleted due to missingness)
## Multiple R-squared:  0.3469, Adjusted R-squared:  0.3463
## F-statistic: 635.2 on 1 and 1196 DF,  p-value: < 2.2e-16
```

Better summary of lm() function results

- use `broom` package

```
require(broom)
```

```
## Loading required package: broom
```

```
tidy(model_1)
```

```
## # A tibble: 2 × 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) -12054.    1940.       -6.21 7.12e- 10
## 2 sat_avg         42.6      1.69     25.2  9.09e-113
```

Introducing Movies!

```
mv <- read_rds("https://github.com/jbisbee1/ISP_Data_Science_2024/raw/main/data/mv.Rds")
```

of observations / unit of analysis / # of variables

```
summary(mv)
```

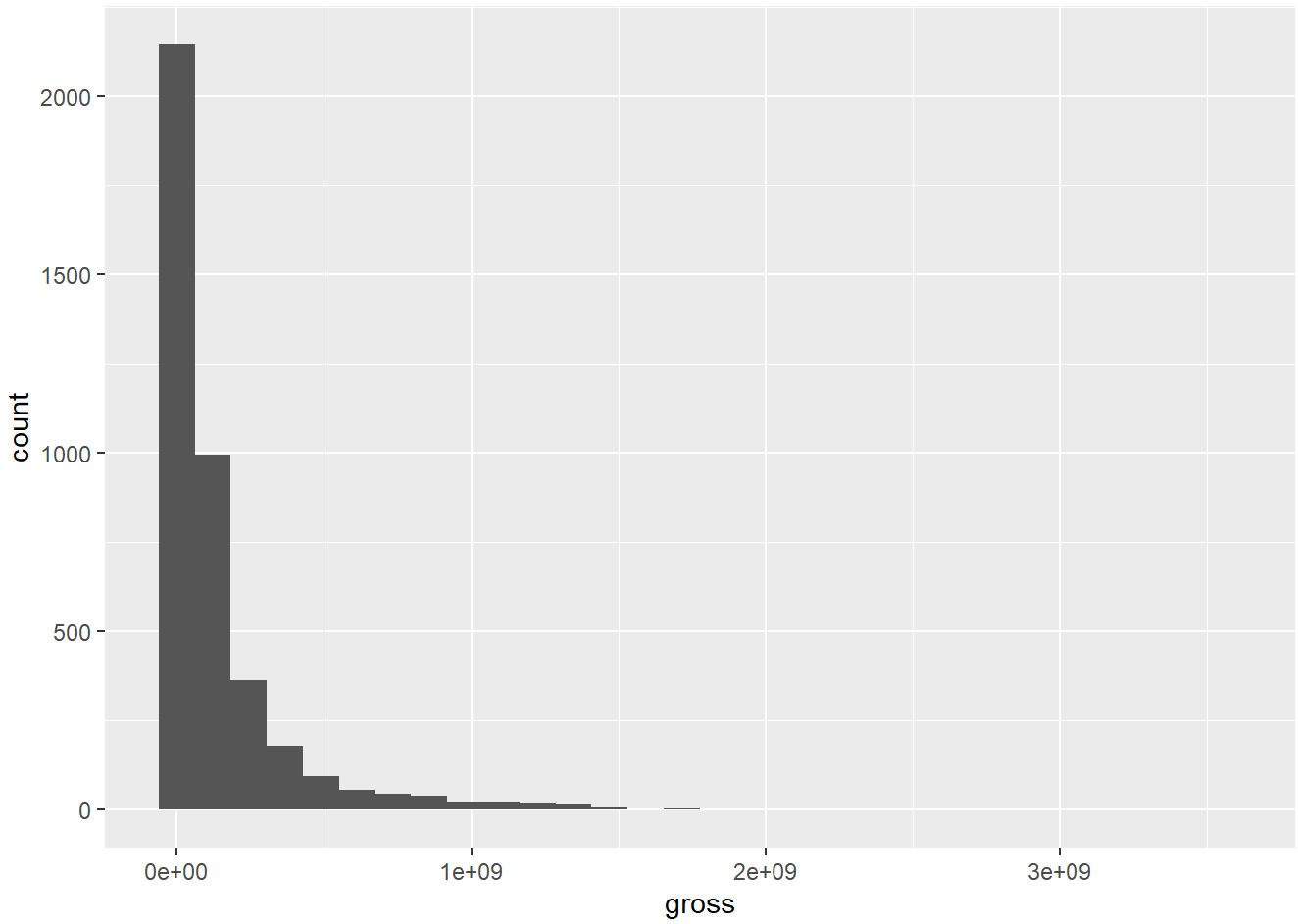
```
##     title              rating             genre                year
##  Length:7673        Length:7673        Length:7673        Min.   :1980
##  Class :character   Class :character   Class :character   1st Qu.:1991
##  Mode  :character   Mode  :character   Mode  :character   Median :2000
##                                                           Mean   :2000
##                                                           3rd Qu.:2010
##                                                           Max.   :2020
##
##    released             score           votes            director
##  Length:7673        Min.   :1.900   Min.   :      7   Length:7673
##  Class :character   1st Qu.:5.800   1st Qu.:   9100   Class :character
##  Mode  :character   Median :6.500   Median :  33000   Mode  :character
##                     Mean   :6.391   Mean   :  88255
##                     3rd Qu.:7.100   3rd Qu.:  93000
##                     Max.   :9.300   Max.   :2400000
##                     NA's   :3       NA's   :3
##     writer             star             country              budget
##  Length:7673        Length:7673        Length:7673        Min.   :     5172
##  Class :character   Class :character   Class :character   1st Qu.: 16865322
##  Mode  :character   Mode  :character   Mode  :character   Median : 37212044
##                                                           Mean   : 57420173
##                                                           3rd Qu.: 77844746
##                                                           Max.   :387367903
##                                                           NA's   :4482
##      gross              company             runtime            id
##  Min.   :7.140e+02   Length:7673        Min.   : 55.0   Min.   :   2
##  1st Qu.:1.121e+07   Class :character   1st Qu.: 95.0   1st Qu.:1668
##  Median :5.178e+07   Mode  :character   Median :104.0   Median :4043
##  Mean   :1.402e+08                      Mean   :107.3   Mean   :4223
##  3rd Qu.:1.562e+08                      3rd Qu.:116.0   3rd Qu.:6636
##  Max.   :3.553e+09                      Max.   :366.0   Max.   :9502
##  NA's   :3668                           NA's   :4       NA's   :4162
##    imdb_id          bechdel_score    boxoffice_a          language
##  Length:7673        Min.   :0.000   Min.   :8.100e+02   Length:7673
##  Class :character   1st Qu.:1.000   1st Qu.:7.822e+06   Class :character
##  Mode  :character   Median :3.000   Median :3.563e+07   Mode  :character
##                     Mean   :2.171   Mean   :6.738e+07
##                     3rd Qu.:3.000   3rd Qu.:8.406e+07
##                     Max.   :3.000   Max.   :1.058e+09
##                     NA's   :4162    NA's   :4318
```

Univariate Visualization

```
# Y variable
mv %>%
  ggplot(aes(x = gross)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

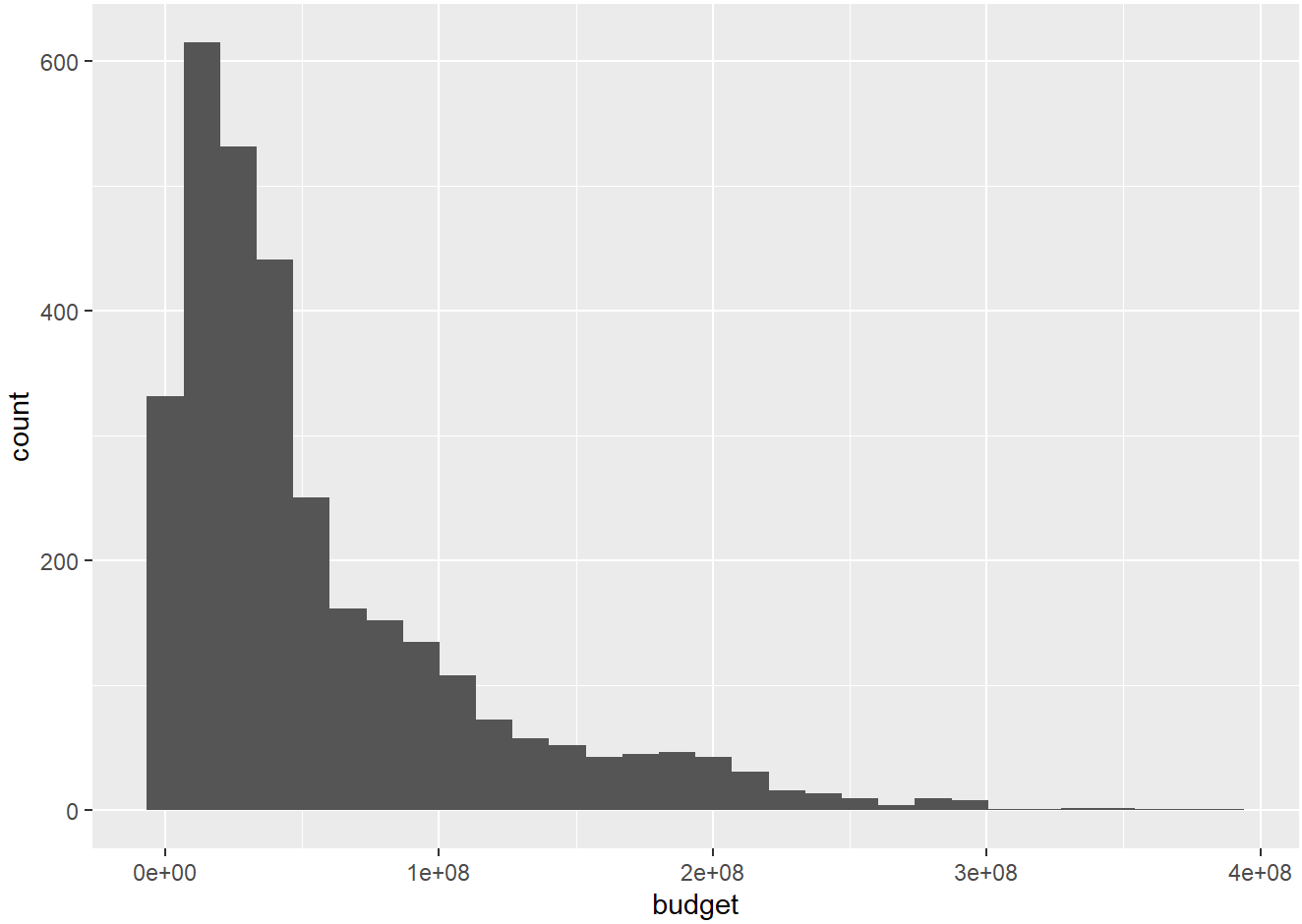
```
## Warning: Removed 3668 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```
# X Variable
mv %>%
  ggplot(aes(x = budget)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4482 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



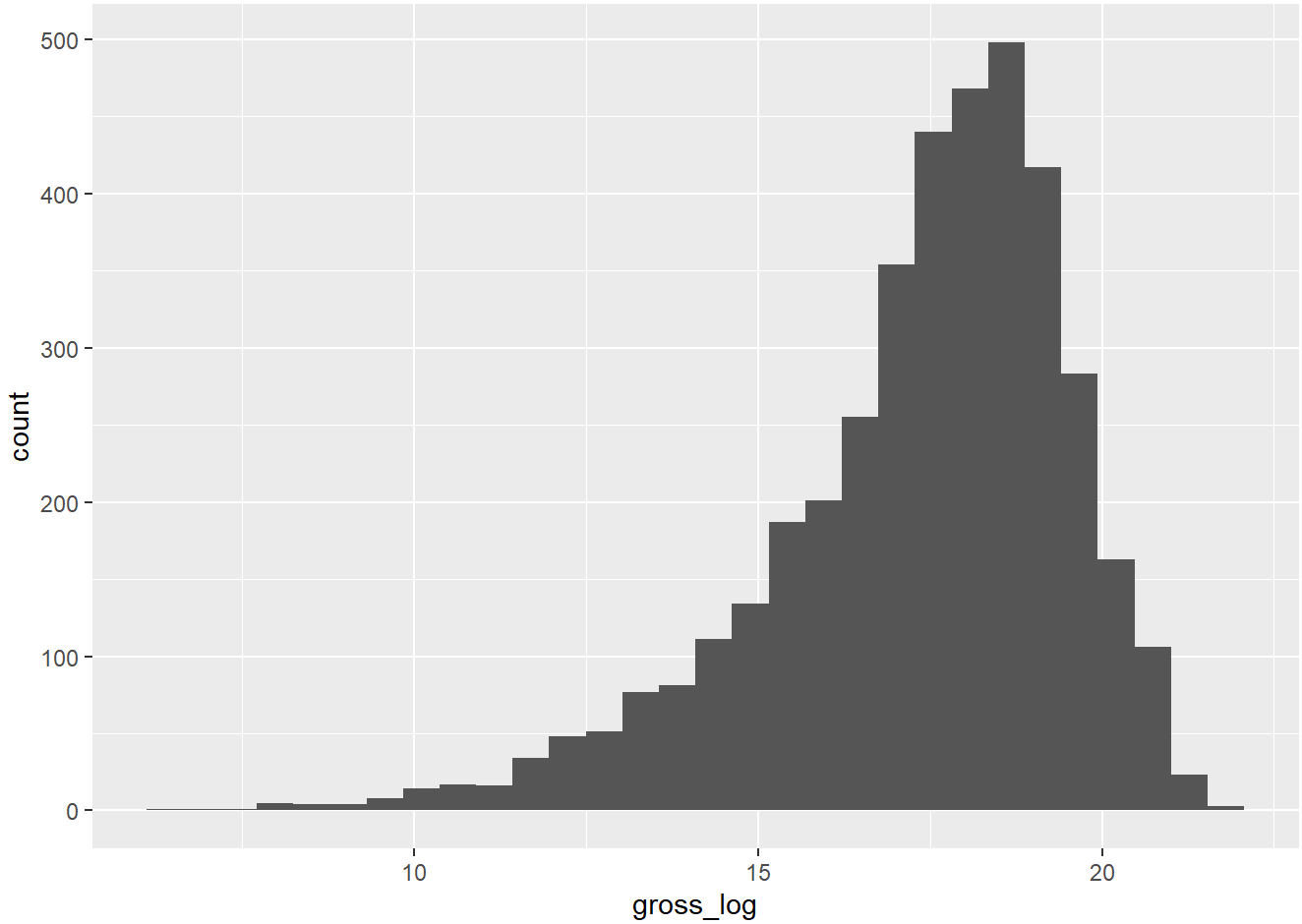
Using log() to fix skew

```
mv <- mv %>%
  mutate(gross_log = log(gross),
         budget_log = log(budget))

# Look at univariate again
mv %>%
  ggplot(aes(x = gross_log)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

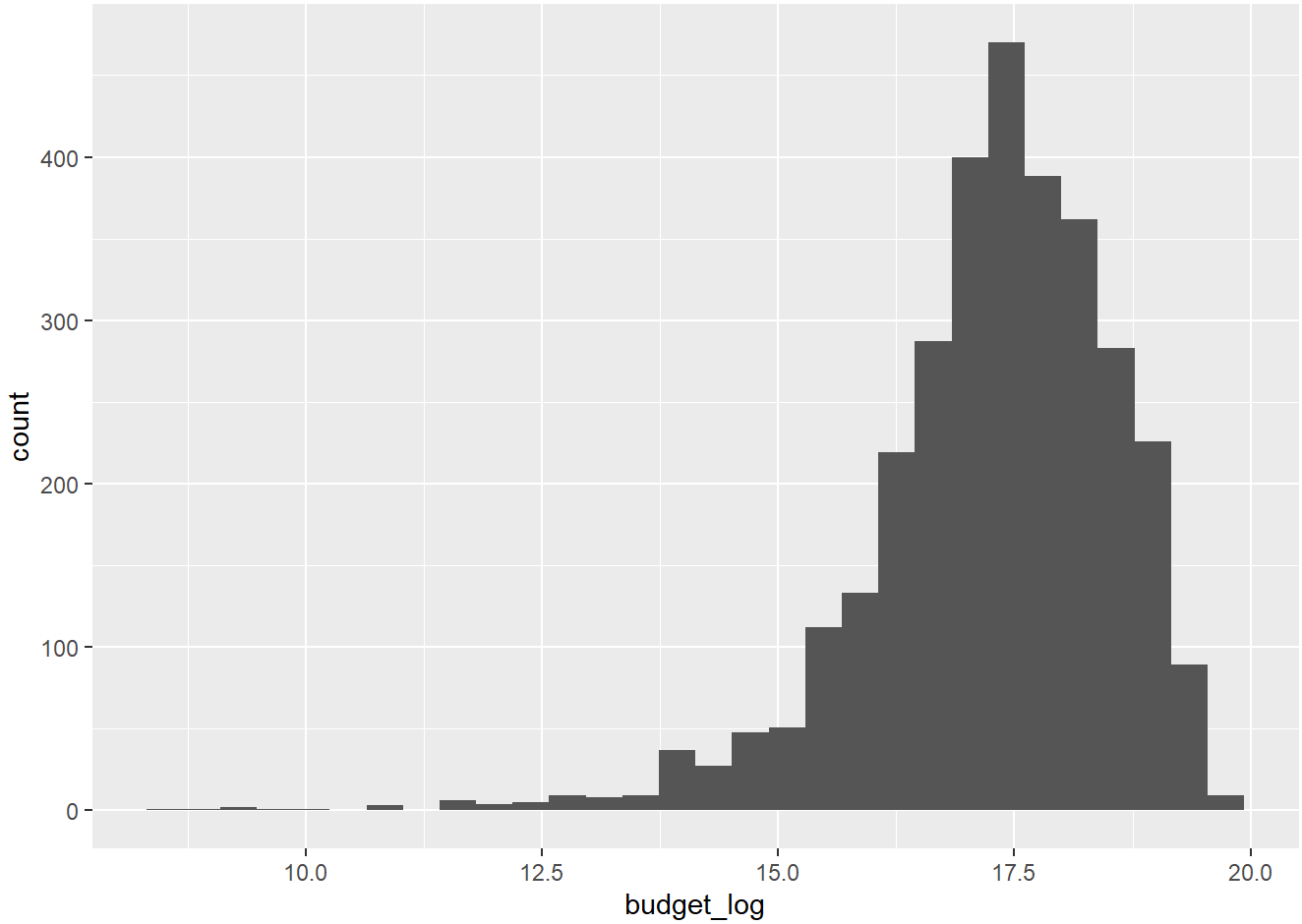
```
## Warning: Removed 3668 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```
mv %>%
  ggplot(aes(x = budget_log)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4482 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



Multivariate