Lecture 4 Notes

2024-01-23

# Getting Started

Loading data

```
require(tidyverse)
```

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

```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
```

```
## ── 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
```

```
df <- read_rds("https://github.com/jbisbee1/DS1000_S2024/raw/main/data/sc_debt.Rds")
```

`group_by()` and `summarize()`

```
df %>%
  mutate(sat_category = ifelse(sat_avg > 1200,
                               "high sat",
                               "low sat")) %>%
  group_by(sat_category) %>%
  summarise(mean_earnings = mean(md_earn_wne_p6,na.rm=T))
```

```
## # A tibble: 3 × 2
##   sat_category mean_earnings
##   <chr>                <dbl>
## 1 high sat            43703.
## 2 low sat             33960.
## 3 <NA>                29250.
```

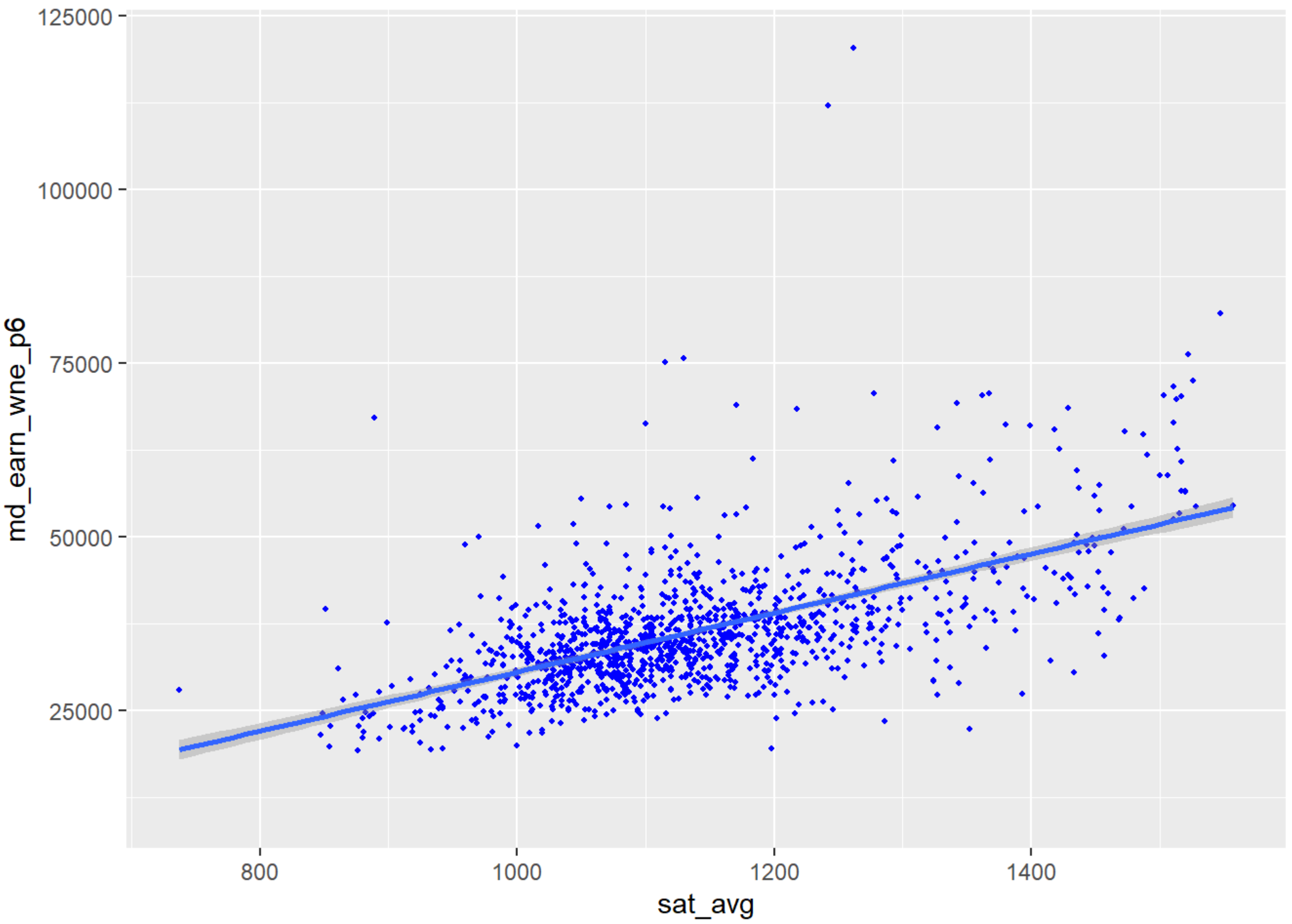
Visualization

```
df %>%
  ggplot(aes(x = sat_avg,
             y = md_earn_wne_p6)) +
  geom_point(size = .8,
             color = 'blue') +
  geom_smooth(method = "lm")
```

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

```
## Warning: Removed 1348 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 1348 rows containing missing values (`geom_point()`).
```



```
colnames(df)
```

```
##  [1] "unitid"       "instnm"       "stabbr"       "grad_debt_mdn"
##  [5] "control"      "region"       "preddeg"      "openadmp"
##  [9] "adm_rate"     "ccbasic"      "sat_avg"      "md_earn_wne_p6"
## [13] "ugds"         "costt4_a"     "selective"    "research_u"
```

```
df %>%
  ggplot(aes(x = sat_avg,
             y = md_earn_wne_p6,
             color = region,
             size = ugds,
             group = 1)) +
  geom_point() +
  geom_smooth(method = "lm")
```

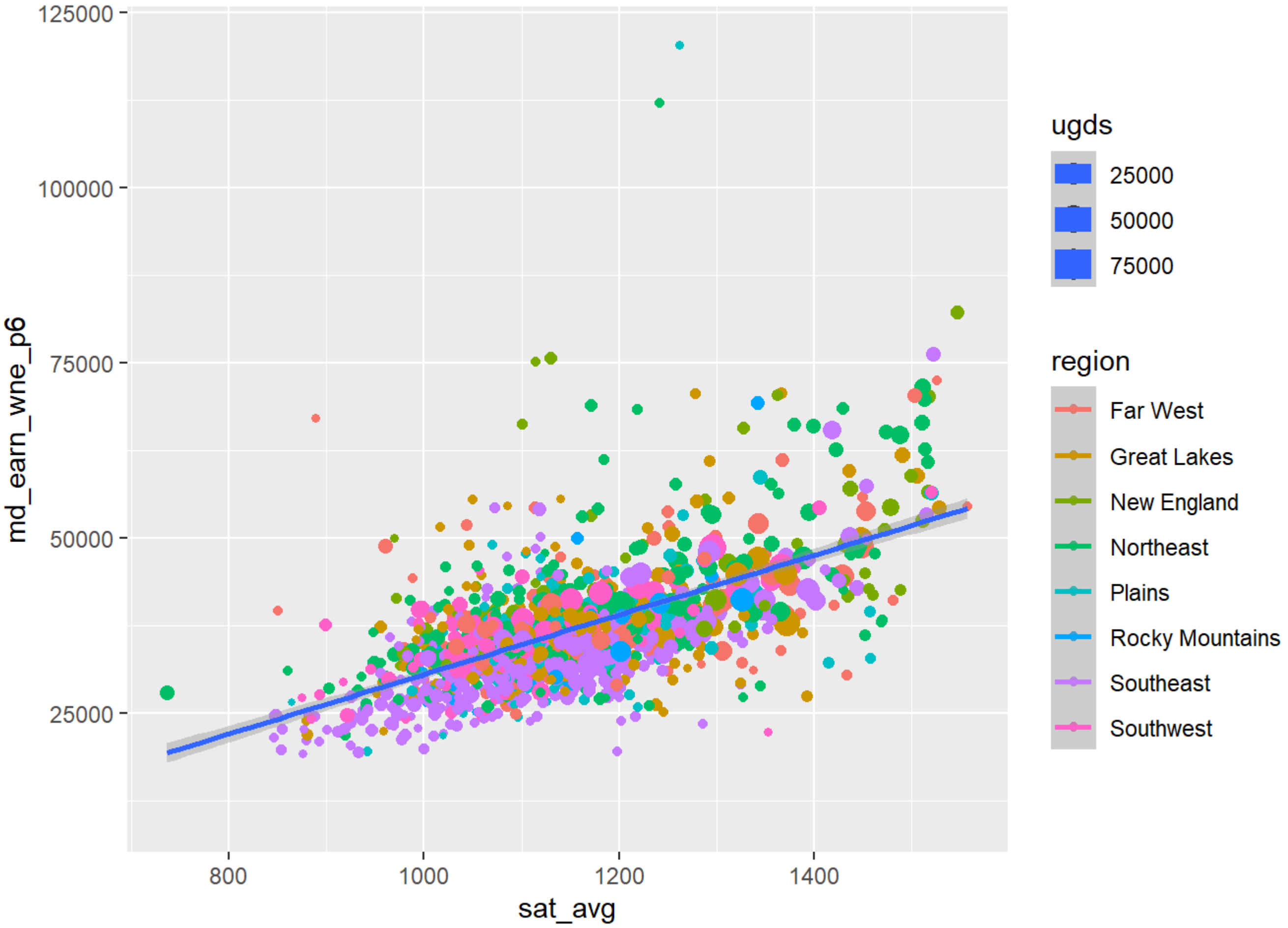
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

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

```
## Warning: Removed 1348 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour, size
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

```
## Warning: Removed 1348 rows containing missing values (`geom_point()`).
```



```r
require(plotly)
```

```
## Loading required package: plotly
```

```
## Warning: package 'plotly' was built under R version 4.3.2
```

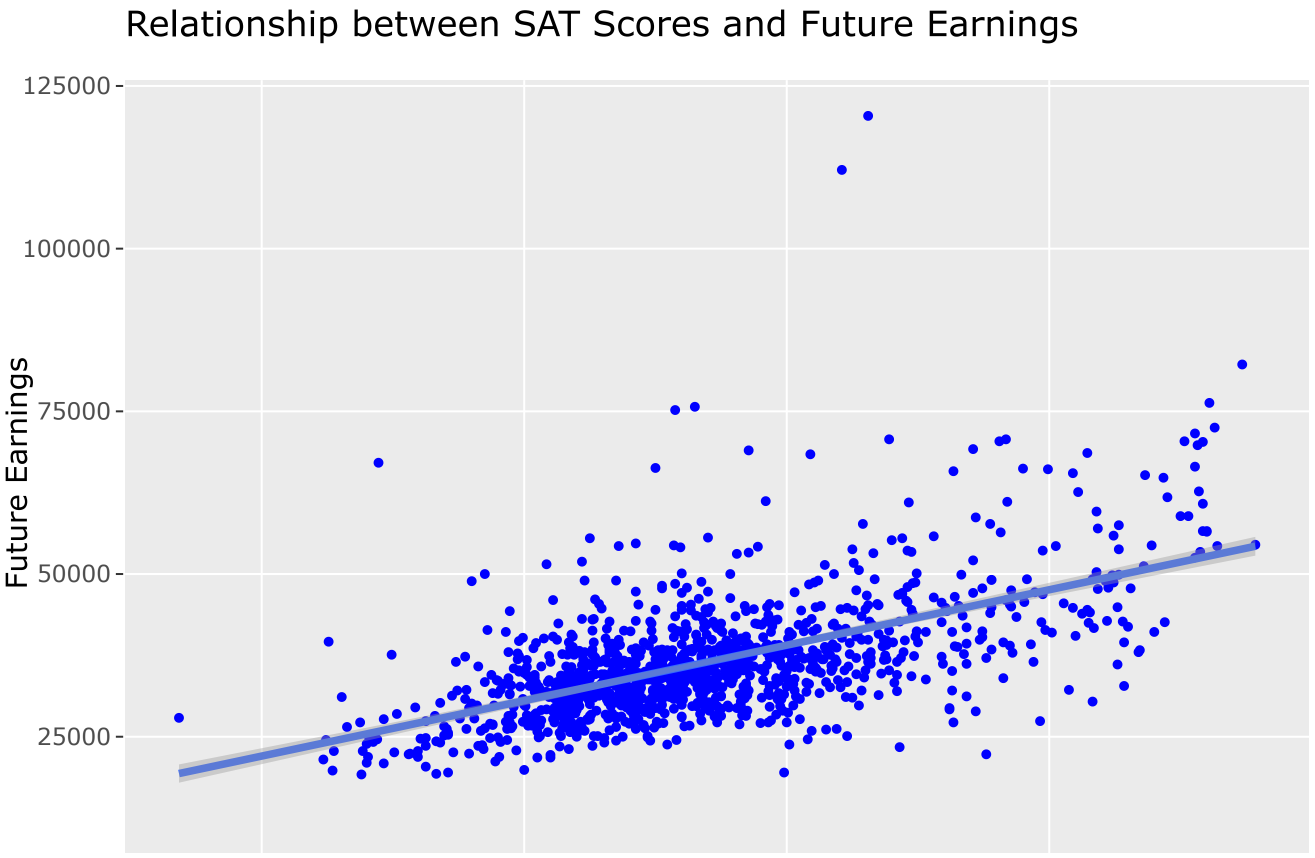
```
##
## Attaching package: 'plotly'
##
## The following object is masked from 'package:ggplot2':
##
##     last_plot
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following object is masked from 'package:graphics':
##
##     layout
```

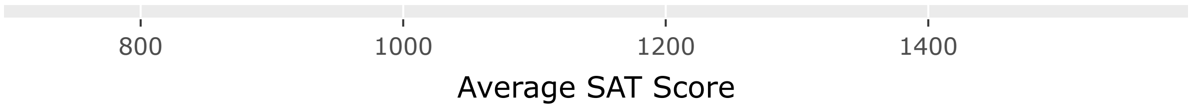
```
scatter_plot <- df %>%
  ggplot(aes(x = sat_avg,
             y = md_earn_wne_p6,
             text=instnm,
             group = 1)) +
  geom_point() +
  geom_point(size = .8,
             color = 'blue') +
  geom_smooth(method = "lm") +
  labs(x = "Average SAT Score",
       y = "Future Earnings",
       title = "Relationship between SAT Scores and Future Earnings",
       subtitle = "Sample of 2,546 schools in 2019")

ggplotly(p = scatter_plot,tooltip = 'text')
```

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

```
## Warning: Removed 1348 rows containing non-finite values (`stat_smooth()`).
```

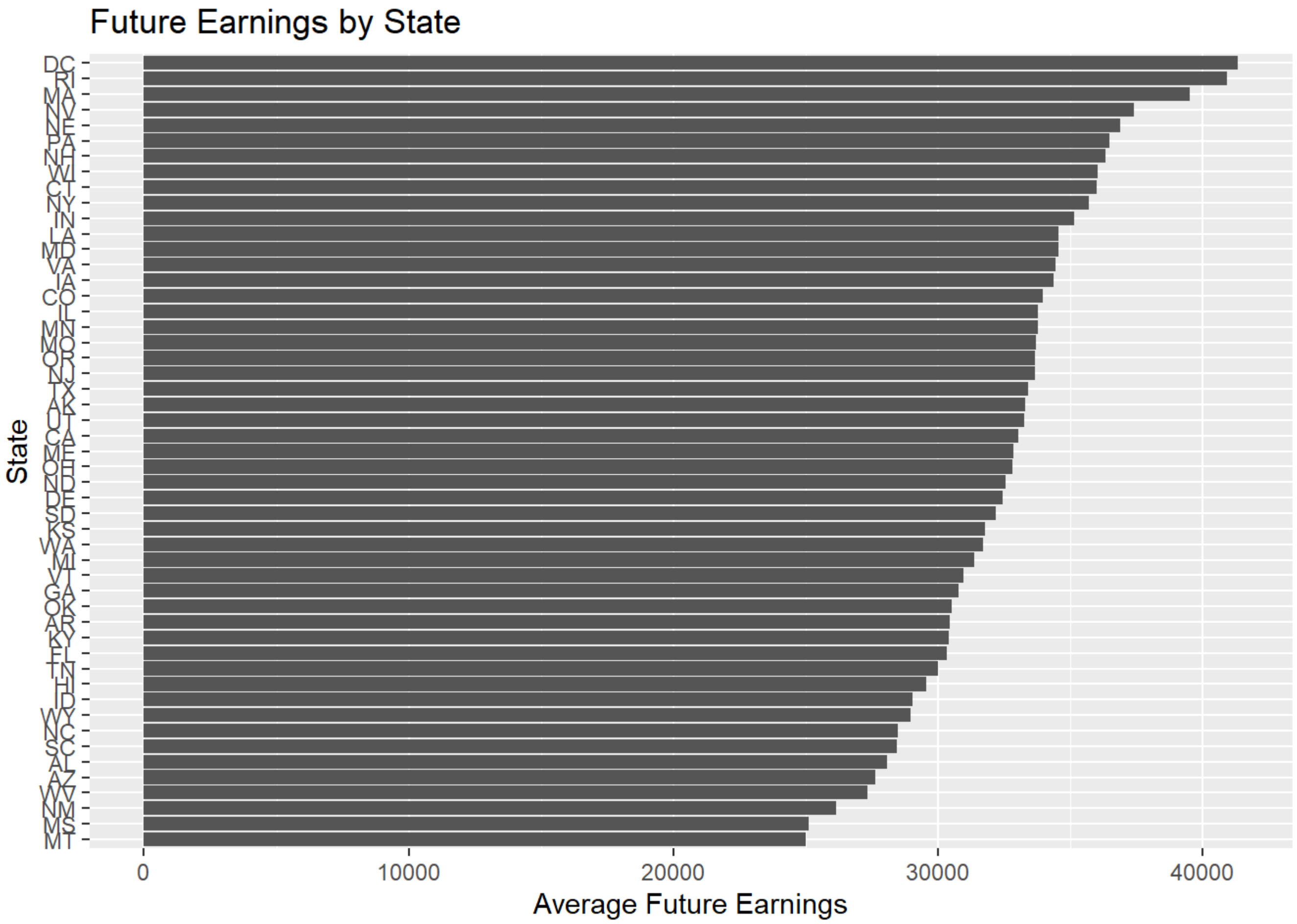
```
## Warning: The following aesthetics were dropped during statistical transformation: tex
t
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```





Poor choices of `geom_...()`

```
df %>%
  group_by(stabbr) %>%
  summarise(mean_earn = mean(md_earn_wne_p6,na.rm=T)) %>%
  ggplot(aes(y = reorder(stabbr,mean_earn),
             x = mean_earn)) +
  geom_bar(stat = "identity") +
  labs(x = "Average Future Earnings",
       y = "State",
       title = 'Future Earnings by State')
```

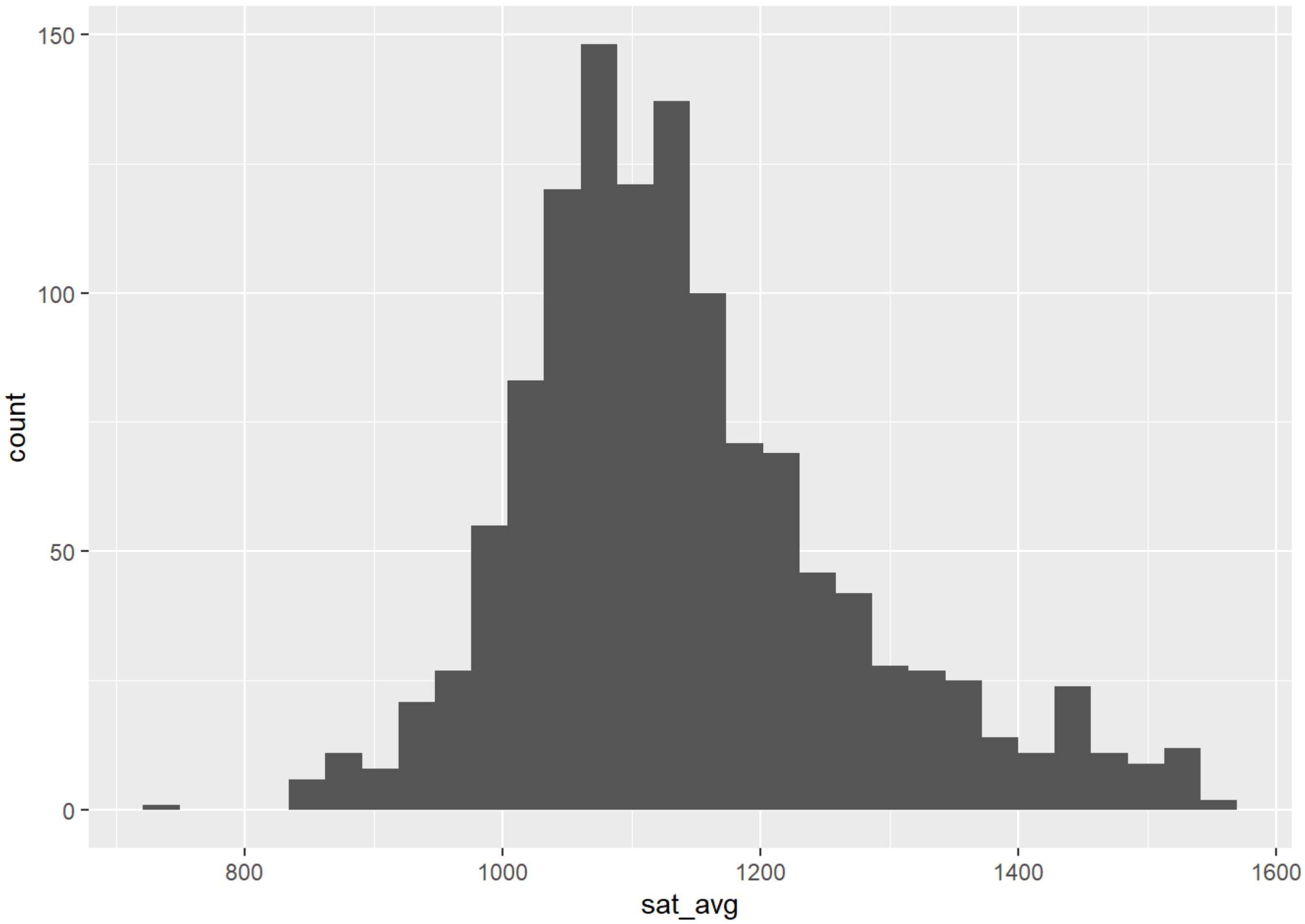


Using `geom_histogram()` or `geom_density()`

```
df %>%
  ggplot(aes(x = sat_avg)) +
  geom_histogram()
```

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

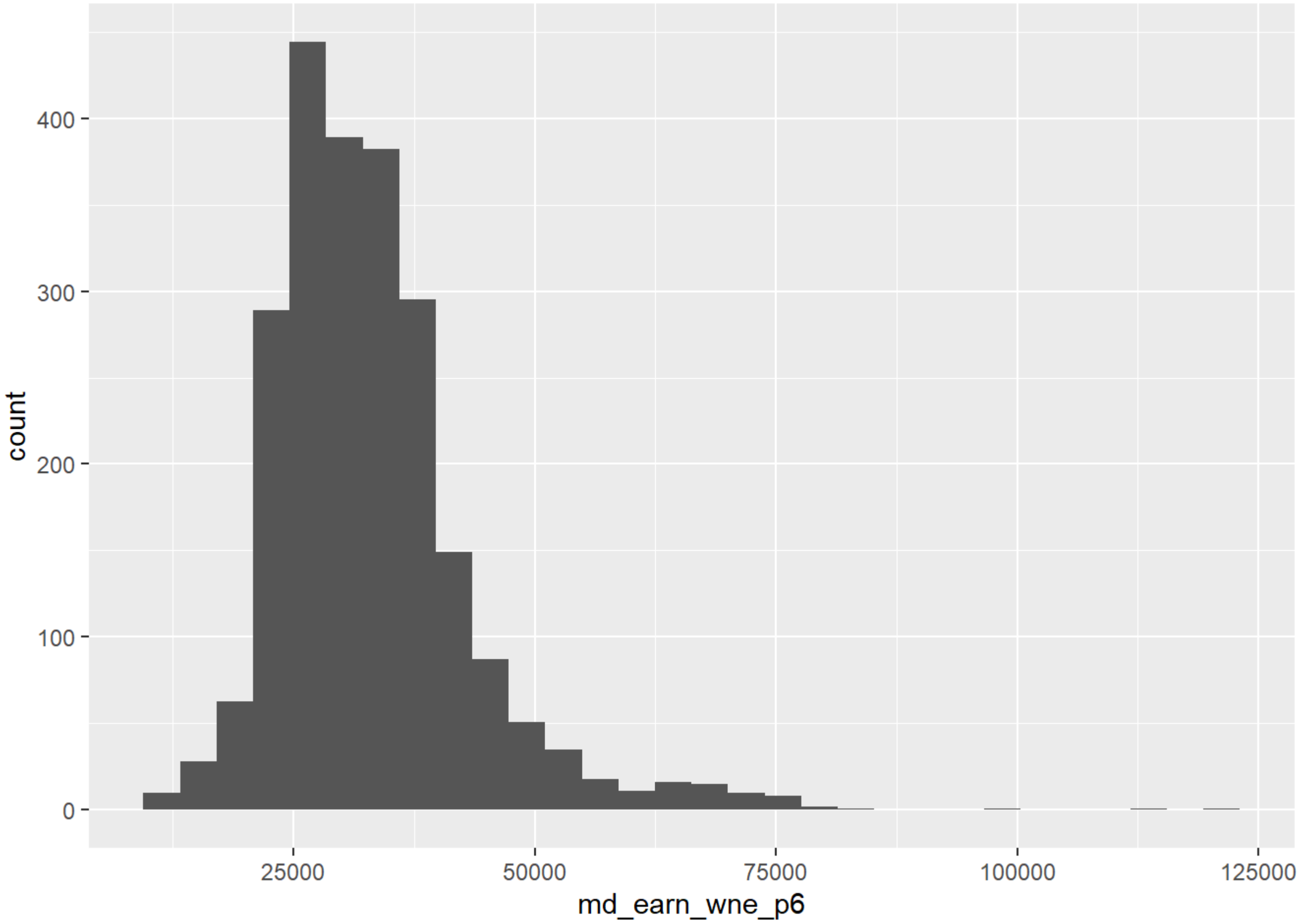
```
## Warning: Removed 1317 rows containing non-finite values (`stat_bin()`).
```



```
df %>%
  ggplot(aes(x = md_earn_wne_p6)) +
  geom_histogram()
```

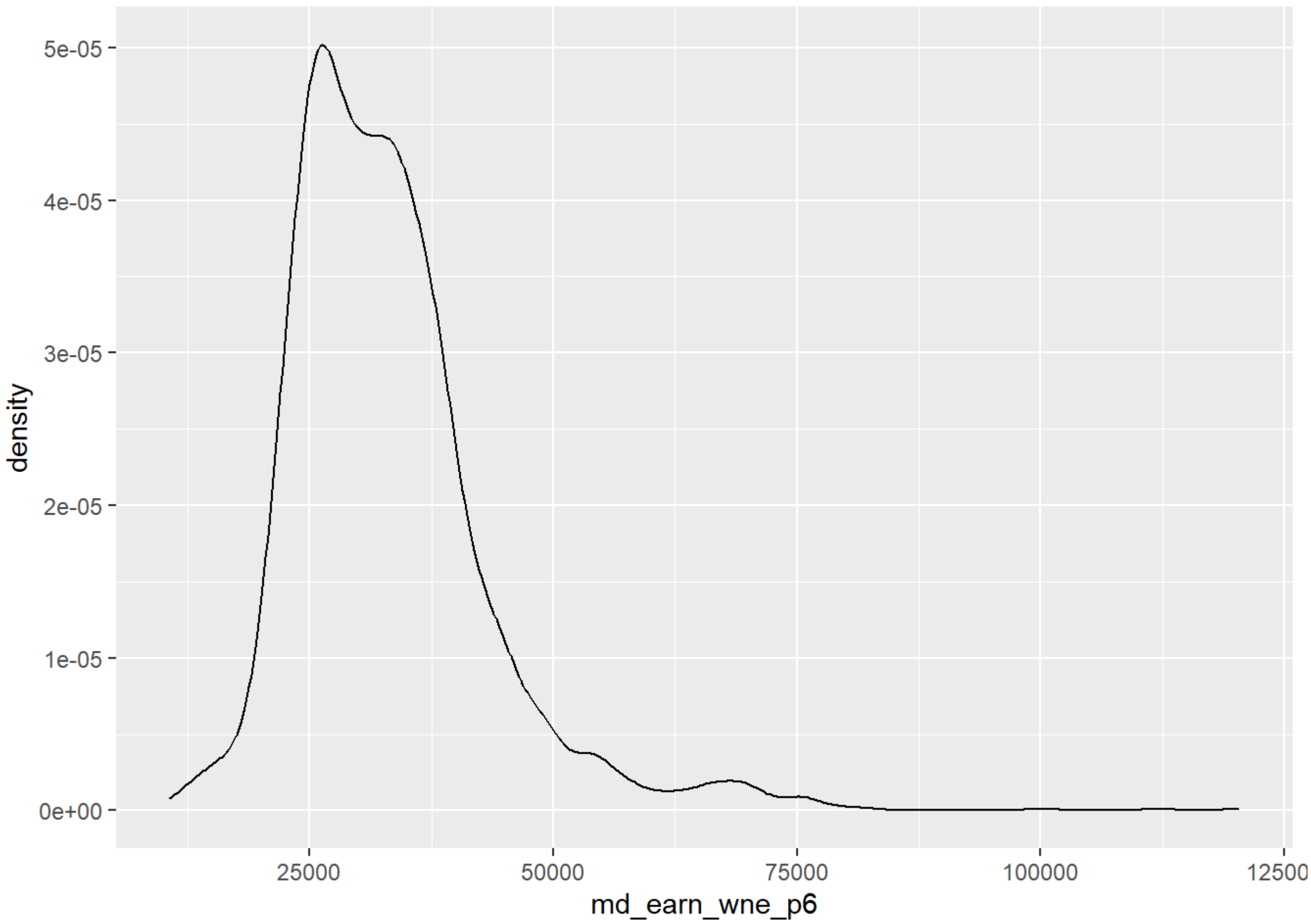
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 240 rows containing non-finite values (`stat_bin()`).
```



```
df %>%
  ggplot(aes(x = md_earn_wne_p6)) +
  geom_density()
```

```
## Warning: Removed 240 rows containing non-finite values (`stat_density()`).
```


5e-05
4e-05
3e-05
2e-05
1e-05
0e+00
density
25000
50000
75000
100000
12500
md_earn_wne_p6