02-Assignment

For this assignment, you'll be working with the `sc_debt.Rds` to predict earnings levels of college graduates using conditional means. You'll need to select the college-level characteristics that you think might be related to earnings levels. Please complete the following steps:

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## Warning: package 'forcats' was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(yardstick)
```

```
## Warning: package 'yardstick' was built under R version 4.0.5
```

```
## For binary classification, the first factor level is assumed to be the event.
## Use the argument `event_level = "second"` to alter this as needed.
```

```
##
## Attaching package: 'yardstick'
```

```
## The following object is masked from 'package:readr':
##
##     spec
```

```
load("sc_debt.Rdata")
sc <- sc_debt
```

1. Calculate the mean of the outcome `md_earn_wne_p6`

```
sc%>%summarize(mean_earnings=mean(md_earn_wne_p6,na.rm=TRUE))
```

```
## # A tibble: 1 x 1
##   mean_earnings
##           <dbl>
## 1        32971.
```

2. Use your mean as a prediction: Create a new variable that consists of the mean of the outcome.

```
sc<-sc%>%
  mutate(mean_earnings=mean(md_earn_wne_p6,na.rm=TRUE))
```

3. Calculate a summary measure of the errors for each observation—the difference between your prediction and the outcome.

```
sc%>%rmse(md_earn_wne_p6,mean_earnings)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard      10243.
```

4. Calculate the mean of the outcome at levels of a predictor variable.

```
sc%>%
  group_by(region)%>%
  summarize(mean_earnings_region=mean(md_earn_wne_p6,na.rm=TRUE))%>%
  arrange(-mean_earnings_region)
```

```
## # A tibble: 8 x 2
##   region          mean_earnings_region
##   <chr>                          <dbl>
## 1 New England                   36785.
## 2 Northeast                     35602.
## 3 Plains                        33702.
## 4 Great Lakes                   33374.
## 5 Far West                      32962.
## 6 Southwest                     31968.
## 7 Rocky Mountains               30454.
## 8 Soutwest                      29781.
```

5. Use these conditional means as a prediction: for every college, use the conditional mean to provide a ‘’best guess” as to that college’s level of the outcome.

```
sc<-sc%>%
  group_by(region)%>%
  mutate(mean_earnings_region=mean(md_earn_wne_p6,na.rm=TRUE))%>%
  ungroup()
```

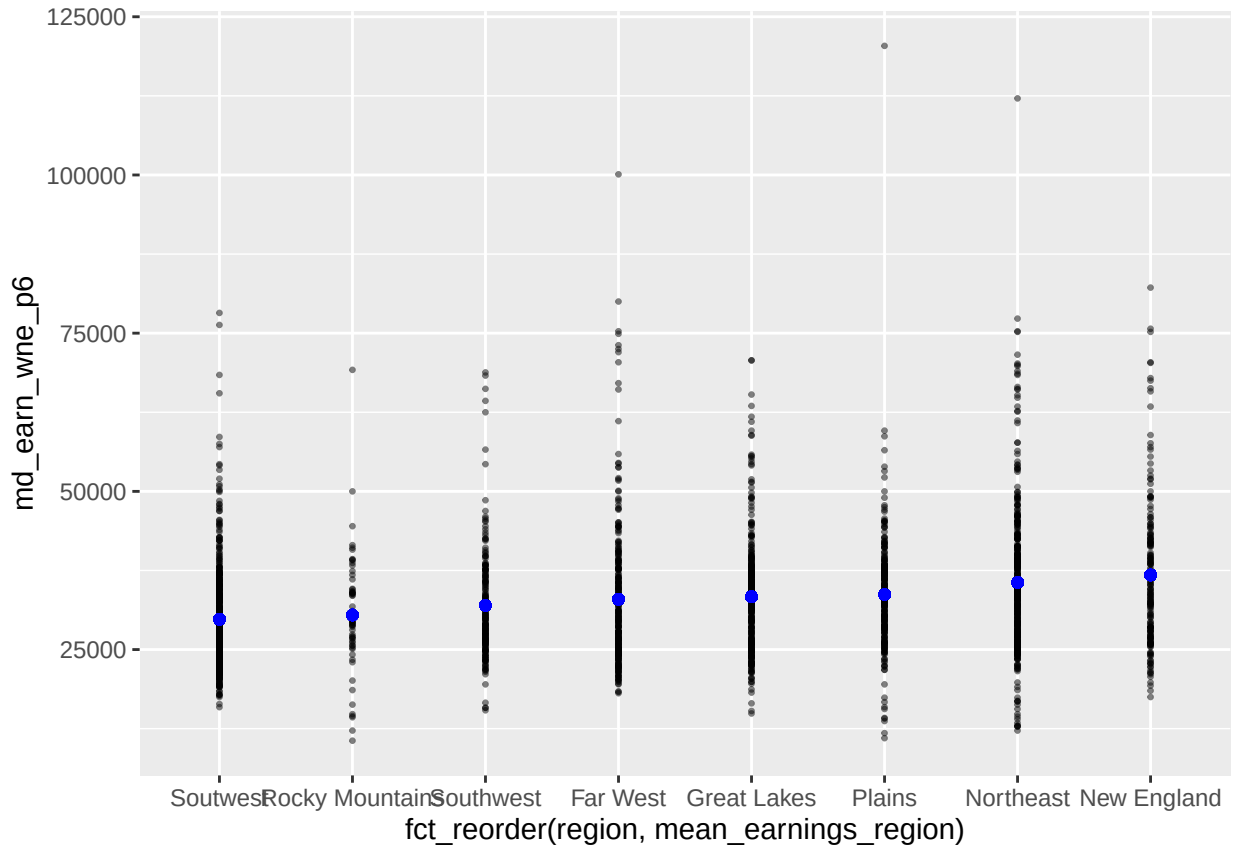
6. Calculate a summary measure of the error in your predictions.

```
sc%>%rmse(md_earn_wne_p6,mean_earnings_region)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       9987.
```

```
sc%>%
  ggplot(aes(x=fct_reorder(region,mean_earnings_region),y=md_earn_wne_p6))+
  geom_point(size=.5,alpha=.5)+
  geom_point(aes(x=region,y=mean_earnings_region),color="blue")
```

```
## Warning: Removed 228 rows containing missing values (geom_point).
```



7. Repeat the above process using the tool of conditional means, try to find 3-4 combined variables that predict the outcome with better (closer to 0) summary measures of error. Report the summary measures of error and the variables (as text in your `.Rmd` file).

```
sc%>%
  mutate(sat_level=ntile(sat_avg,4))%>%
  group_by(sat_level)%>%
  summarize(mean_sat=mean(md_earn_wne_p6,na.rm=TRUE),count=n())
```

```
## # A tibble: 5 x 3
##   sat_level mean_sat count
##       <int>    <dbl> <int>
## 1         1   31120.   308
## 2         2   34539.   308
## 3         3   36352.   307
## 4         4   44148.   307
## 5        NA   29187.  1325
```

```
sc%>%
  group_by(region)%>%
  summarize(count=n())
```

```
## # A tibble: 8 x 2
##   region          count
##   <chr>           <int>
## 1 Far West          334
## 2 Great Lakes       364
## 3 New England       201
## 4 Northeast         489
## 5 Plains            283
## 6 Rocky Mountains    77
## 7 Southwest         219
## 8 Soutwest          588
```

```
sc<-sc%>%
  mutate(sat_level=ntile(sat_avg,4))%>%
  group_by(region,control,preddeg,sat_level)%>%
  mutate(mean_earnings_lots_of_predictors=mean(md_earn_wne_p6,na.rm=TRUE))%>%
  ungroup()

sc%>%rmse(md_earn_wne_p6,mean_earnings_lots_of_predictors)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       7798.
```

Submit your assignment as `02-assignment-<yourlastname>.Rmd`, where `<yourlastname>` is your last name. (By the way, any time you see this: `<sometext>`, that indicates that you need to substitute something in, so if I were to submit the above assignment, it would be as: `02-assignment-doyle.Rmd`)

I expect that the `.Rmd` file you submit will run cleanly, and that there shouldn't be any errors. Use LOTS of text to tell me what you are doing.