

## Progress Report

**Potential Topic:** Determine the relationship between earnings and retention rates of institutions

**Idea:** Create a shiny app that will generate boxplots, barplots, scatterplots, OLS line and t-test results

```
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.2.1 --
## v ggplot2 3.2.1      v purrr   0.3.2
## v tibble  2.1.3      v dplyr   0.8.3
## v tidyr   1.0.0      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.4.0

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

library(broom)
```

Read the data

```
scoreboard = read_csv("./scoreboard.csv")

## Warning: 3996 parsing failures.
##   row      col expected actual      file
## 6669 LOCALE    a double  NULL './scoreboard.csv'
## 6669 LATITUDE a double  NULL './scoreboard.csv'
## 6669 LONGITUDE a double  NULL './scoreboard.csv'
## 6669 CCBASIC  a double  NULL './scoreboard.csv'
## 6669 CCUGPROF a double  NULL './scoreboard.csv'
## ....
## See problems(...) for more details.
```

Extract the names of scoreboard

```
namefile = names(scoreboard)
```

Extract the variable names for the earnings

```
data.frame(namefile) %>%
  filter(str_detect(namefile, "GT_"))
```

```
##   namefile
## 1 GT_25K_P10
## 2 GT_25K_P6
## 3 GT_25K_P7
## 4 GT_25K_P8
## 5 GT_25K_P9
## 6 GT_28K_P10
## 7 GT_28K_P8
## 8 GT_28K_P6
```

Extract the variable names for the retention rates

```
data.frame(namefile) %>%
  filter(str_detect(namefile, "RET_"))
```

```
##          namefile
## 1          RET_FT4
## 2          RET_FTL4
## 3          RET_PT4
## 4          RET_PTL4
## 5      RET_FT4_POOLED
## 6      RET_FTL4_POOLED
## 7      RET_PT4_POOLED
## 8      RET_PTL4_POOLED
## 9  RET_FT_DEN4_POOLED
## 10 RET_FT_DENL4_POOLED
## 11 RET_PT_DEN4_POOLED
## 12 RET_PT_DENL4_POOLED
## 13      POOLYRSRET_FT
## 14      POOLYRSRET_PT
## 15 RET_FT4_POOLED_SUPP
## 16 RET_FTL4_POOLED_SUPP
## 17 RET_PT4_POOLED_SUPP
## 18 RET_PTL4_POOLED_SUPP
```

Subset the scoreboard dataframe with retention rate variables and earnings variables

```
scoreboard %>%
  dplyr::select(c("INSTNM", "RET_FT4", "RET_FTL4", "RET_PT4", "RET_PTL4", "RET_FT4_POOLED", "RET_FTL4_POOLED", "RET_PT4_POOLED", "RET_PTL4_POOLED", "RET_FT_DEN4_POOLED", "RET_FT_DENL4_POOLED", "RET_PT_DEN4_POOLED", "RET_PT_DENL4_POOLED", "POOLYRSRET_FT", "POOLYRSRET_PT", "RET_FT4_POOLED_SUPP", "RET_FTL4_POOLED_SUPP", "RET_PT4_POOLED_SUPP", "RET_PTL4_POOLED_SUPP", "Fulltime-Four-Retention", "Full-Less-Four-Retention", "Part-Four-Retention", "Part-Less-Four-Retention", "Full-Cohort", "Part-Cohort", "Earnings after 6 years", "Earnings after 8 years", "Earnings after 10 years") %>%
  rename("Institution" = 1, "Fulltime-Four-Retention" = 2, "Full-Less-Four-Retention" = 3, "Part-Four-Retention" = 4, "Part-Less-Four-Retention" = 5, "Full-Cohort" = 10, "Part-Cohort" = 11, "Earnings after 6 years" = 12, "Earnings after 8 years" = 13, "Earnings after 10 years" = 14) %>%
  retention_scoreboard
```

```
head(retention_scoreboard)
```

```
## # A tibble: 6 x 14
##   Institution `Fulltime-Four-Retention` `Full-Less-Four-Retention` `Part-Four-Retention` `Part-Less-Four-Retention`
##   <chr>      <chr>                  <chr>                  <chr>
## 1 Alabama A ~ 0.5879                NULL                    0.1316
## 2 University~ 0.8436                NULL                    0.4384
## 3 Amridge Un~ 0.6667                NULL                    0.1176
## 4 University~ 0.8248                NULL                    0.375
## 5 Alabama St~ 0.5923                NULL                    0.35
## 6 The Univer~ 0.8709                NULL                    0.5556
## # ... with 10 more variables: `Part-Less-Four-Retention` <chr>,
## #   `Full-Four-Pooled-Variance` <chr>,
## #   `Full-Less-Four-Pooled-Variance` <chr>,
## #   `Part-Four-Pooled-Variance` <chr>,
## #   `Part-Less-Four-Pooled-Variance` <chr>, `Full-Cohort` <chr>,
## #   `Part-Cohort` <chr>, `Earnings after 6 years` <chr>, `Earnings after 8
## #   years` <chr>, `Earnings after 10 years` <chr>
```

Subset the data for four-year institutions

```
retention_scoreboard %>%
  select("Institution", "Fulltime-Four-Retention", "Part-Four-Retention", "Earnings after 6 years", "Earnings after 8 years", "Earnings after 10 years") %>%
  filter(`Fulltime-Four-Retention` != "NULL") %>%
  filter(`Part-Four-Retention` != "NULL") ->
  four_year_institution_retention
```

Gather the data for retention rates based on full time or part time students

```
four_year_institution_retention %>%
  rename("FullTime" = 2, "PartTime" = 3) %>%
  gather("FullTime", "PartTime", key = "Type", value = "Retention") %>%
  filter(`Earnings after 6 years` != "PrivacySuppressed") %>%
  mutate(Retention = as.numeric(Retention)) %>%
  mutate(`Earnings after 6 years` = as.numeric(`Earnings after 6 years`)) %>%
  mutate(`Earnings after 8 years` = as.numeric(`Earnings after 8 years`)) %>%
  mutate(`Earnings after 10 years` = as.numeric(`Earnings after 10 years`)) ->

four_year_institution
```

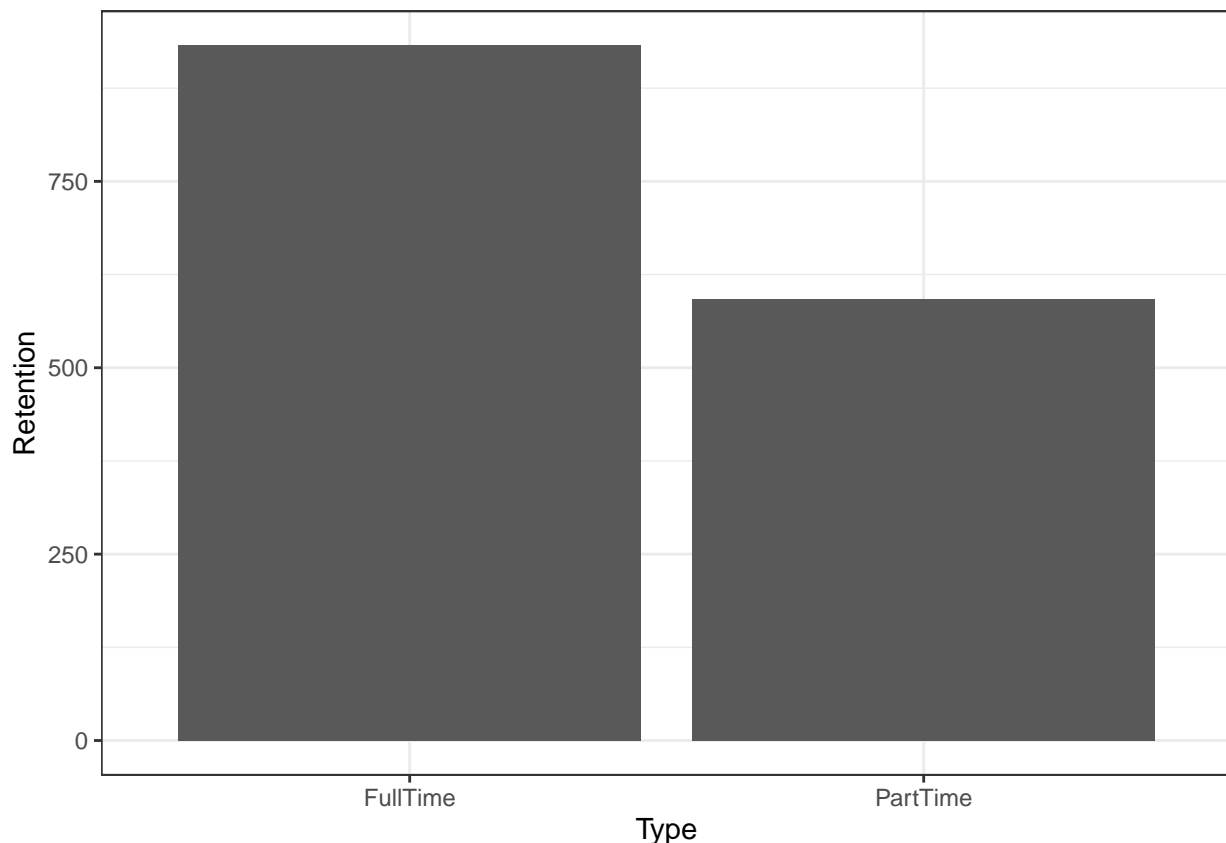
```
## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion
```

```
## Warning: NAs introduced by coercion
```

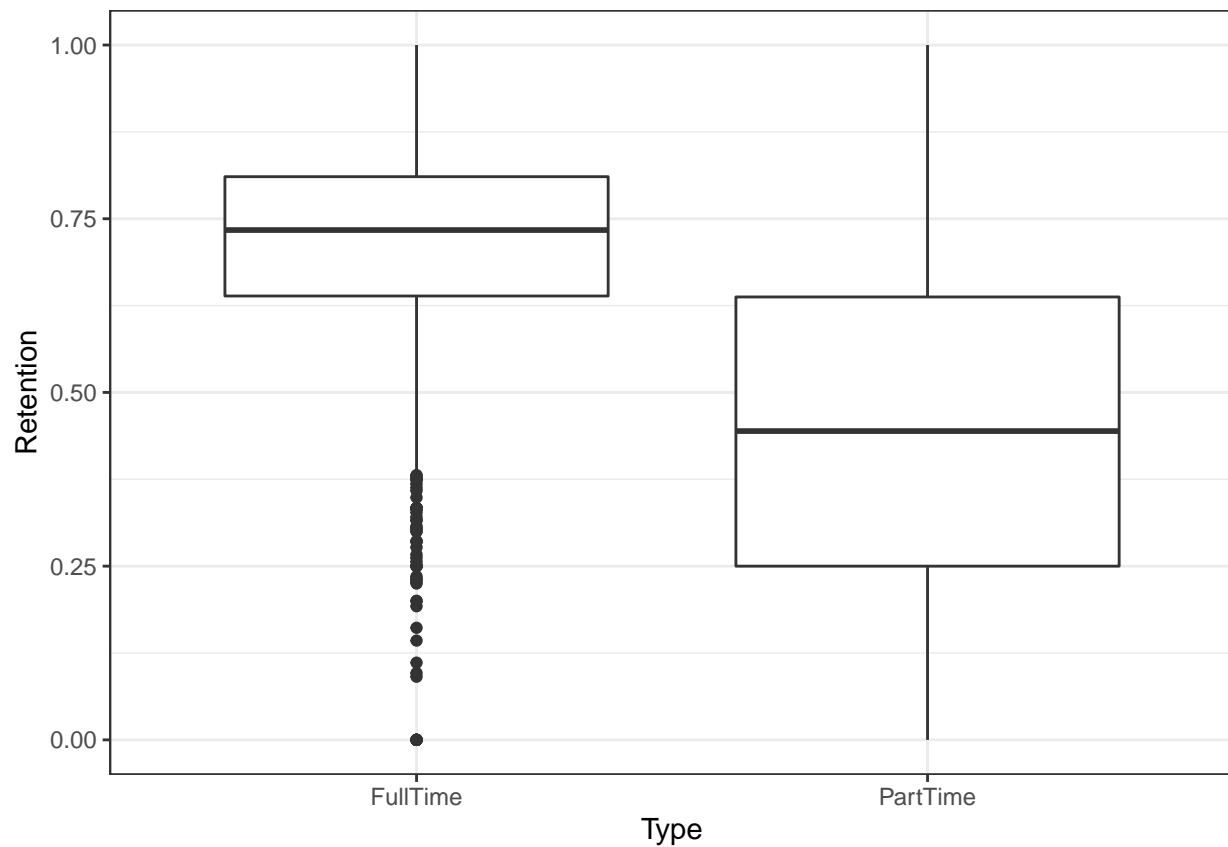
Create a barplot for retention rates based on student types for four-year institutions

```
ggplot(four_year_institution, aes(x = Type, y = Retention)) +
  geom_col() +
  theme_bw()
```



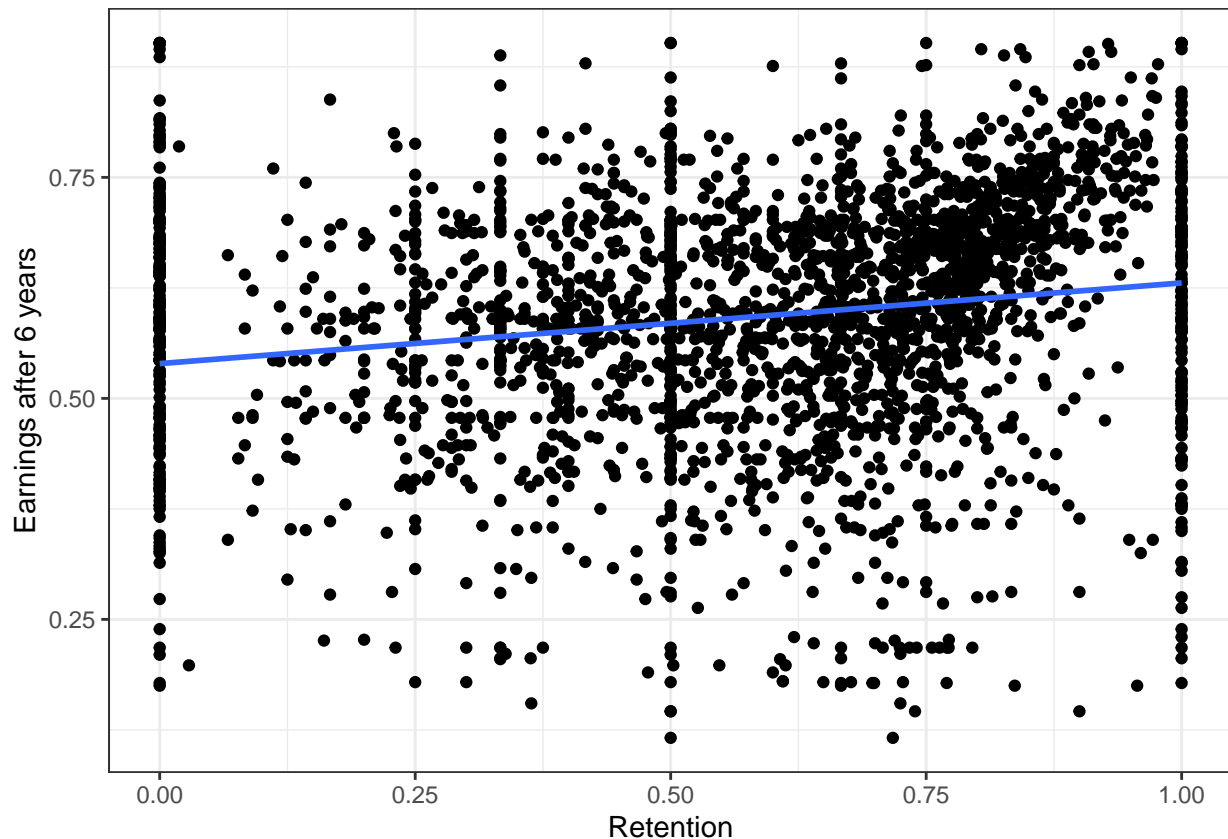
Create a bboxplot of retention rates based on types for four-year institutions

```
ggplot(four_year_institution, aes(x = Type, y = Retention)) +
  geom_boxplot() +
  theme_bw()
```



Create a scatterplot and an OLS line for Earnings after 6 years vs Retention rates

```
ggplot(four_year_institution, aes(x = Retention, y = `Earnings after 6 years`)) +  
  geom_point(na.rm = TRUE) +  
  geom_smooth(method = "lm", formula = y~x, se = F, na.rm = TRUE) +  
  theme_bw()
```



Fit a linear model for earnings after 6 years vs retention rate

```
lmout1 = lm(`Earnings after 6 years` ~ Retention, data = four_year_institution)
tidy(lmout1)
```

```
## # A tibble: 2 x 5
##   term      estimate std.error statistic  p.value
##   <chr>      <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)  0.539    0.00629    85.7    0.
## 2 Retention    0.0913   0.00983     9.29 3.11e-20
```

Run a t-test for different types of students for the retention rate

```
tout1 = t.test(Retention ~ Type, alternative = "greater", data = four_year_institution)
tidy_tout1 = tidy(tout1)
tidy_tout1$p.value
```

```
## [1] 1.876952e-140
```

Set up the hypotheses for the t-test

H0: FullTime = PartTime

HA: FullTime > PartTime

Since the p-value is extremely smaller than level of significance(0.05), we reject the null hypothesis and conclude that FullTime > PartTime

**AI:** Clean undergraduate famliy income variable.

**Jack:** Undergraduate race and gender

**Sam:** Clean the Retention Rate and Earnings

**Boxplots**

**Histograms**

**Dot plots and OLS line**

**Summary of the clean data**