Report Code

Surya Menon

12/5/2018

```
library(tidyverse)
library(modelr)
library(fiftystater)
```

```
# loading a year of data
col_08_09 <- read_csv("MERGED2016_17_PP.csv")
col_08_09 <- col_08_09 %>%
  mutate("Year" = "2008-09")

# join together years of data
college_08_15 <- rbind(col_08_09, col_09_10, col_10_11, col_11_12, col_12_13, col_13_14, col_14_15)

# split data into train/valid/test sets for modeling
set.seed(1)

college_parts <- resample_partition(colleges ,c(train = 0.6, valid = 0.2, test = 0.2))

college_parts_train <- as_tibble(college_parts$train)
college_parts_test <- as_tibble(college_parts$test)
college_parts_valid <- as_tibble(college_parts$valid)

# create STEM variable
colleges %>%
  # combine percentages of degrees in STEM-related fields
  mutate(stem_pct = PCIP11 + PCIP14 + PCIP15 + PCIP26 + PCIP27 +
           PCIP40 + PCIP41)

# create debt-to-earnings variable
colleges %>%
  mutate(GRAD_DEBT_MDN = as.numeric(GRAD_DEBT_MDN),
         MD_EARN_WNE_P8 = as.numeric(MD_EARN_WNE_P8),
         # median debt/median earnings 8 years after graduation
         DEBT_TO_EARN = GRAD_DEBT_MDN/MD_EARN_WNE_P8)

# combine parental education level into 1 variable
colleges %>%
  gather(PAR_ED_PCT_MS, PAR_ED_PCT_HS,PAR_ED_PCT_PS,
         key="ParentEdu", value = "Percent")

# average cost of attendance - top states
loan_1617 %>%
  filter(CONTROL %in% c("1", "2"),PREDDEG == "3") %>%
  mutate(
    cost = as.numeric(COSTT4_A)
  ) %>% select(INSTNM, STABBR,cost) %>%
  group_by(STABBR) %>% summarise(tot_stu = mean(cost, na.rm = TRUE)) %>% arrange(desc(tot_stu))
```

```r
# merge
col_08_09 <- read.csv("MERGED2016_17_PP.csv")
col_08_09 <- col_08_09 %>%
  mutate("Year" = "2008-09")

col_09_10 <- read.csv("MERGED2009_10_PP.csv")
col_09_10 <- col_09_10 %>%
  mutate("Year" = "2009-10")

col_10_11 <- read.csv("MERGED2010_11_PP.csv")
col_10_11 <- col_10_11 %>%
  mutate("Year" = "2010-11")

col_11_12 <- read.csv("MERGED2011_12_PP.csv")
col_11_12 <- col_11_12 %>%
  mutate("Year" = "2011-12")

col_12_13 <- read.csv("MERGED2012_13_PP.csv")
col_12_13 <- col_12_13 %>%
  mutate("Year" = "2012-13")

col_13_14 <- read.csv("MERGED2013_14_PP.csv")
col_13_14 <- col_13_14 %>%
  mutate("Year" = "2013-14")

col_14_15 <- read.csv("MERGED2014_15_PP.csv")
col_14_15 <- col_14_15 %>%
  mutate("Year" = "2014_15")

college_08_13 <- rbind(col_08_09, col_09_10, col_10_11, col_11_12, col_12_13, col_13_14, col_14_15)

# subset
colleges <- select(college_08_13, Year, COSTT4_A, ICLEVEL, INSTNM, CITY, REGION, STABBR, LATITUDE, LONGI

# make tibble  - easier to work with
colleges <- as_tibble(colleges)

mass <- c("Massachusetts Institute of Technology","Harvard University",
"Brandeis University", "Boston College", "Tufts University",
"University of Massachusetts-Amherst", "University of Massachusetts-Lowell",
"University of Massachusetts-Boston", "Massachusetts College of Liberal Arts",
"University of Massachusetts-Dartmouth")
```

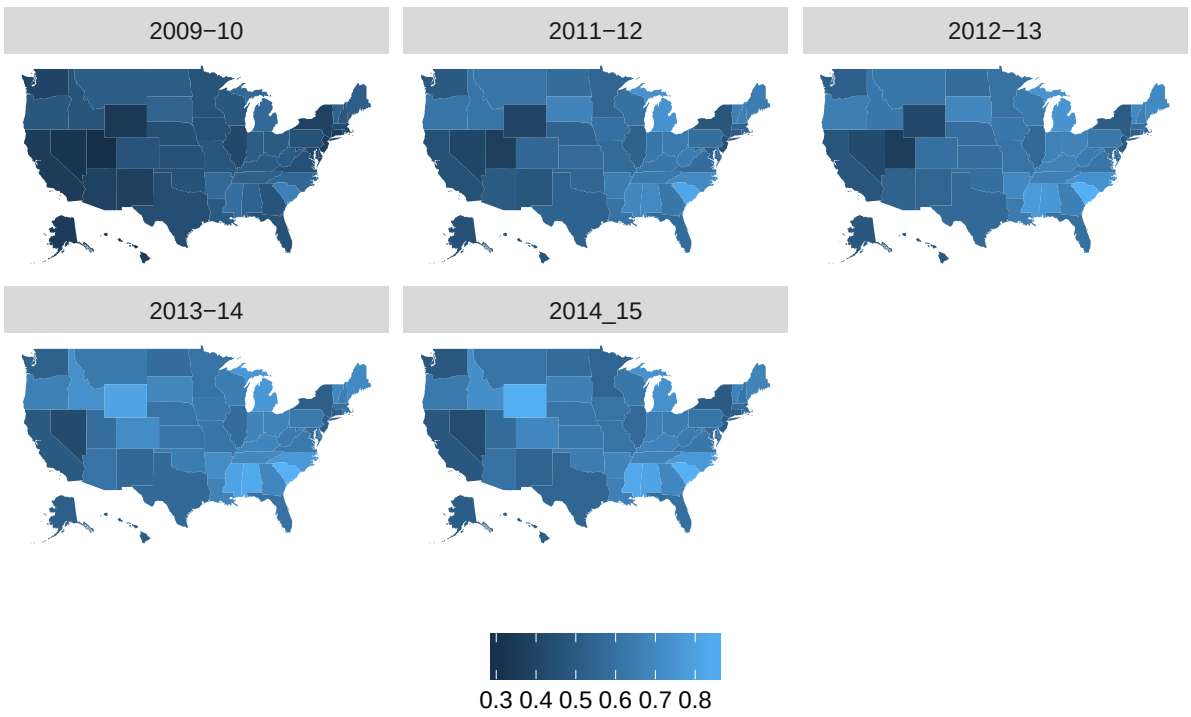
```r
# average debt to earnings ratio by state - excude null data
colleges %>%
  filter(CONTROL %in% c("1", "2"), ICLEVEL == "1") %>%
  filter(Year != "2008-09", Year != "2010-11") %>%
  select(INSTNM, STABBR,MD_EARN_WNE_P8, Year,
         GRAD_DEBT_MDN, MD_FAMINC, COSTT4_A) %>%
  mutate(GRAD_DEBT_MDN = as.numeric(as.character(GRAD_DEBT_MDN)),
         MD_EARN_WNE_P8 = as.numeric(as.character(MD_EARN_WNE_P8)),
         DEBT_TO_EARN = GRAD_DEBT_MDN/MD_EARN_WNE_P8
  ) %>%
  group_by(STABBR, Year) %>%
```

```
  summarise(avg_earn = mean(DEBT_TO_EARN , na.rm = TRUE)) %>%
  ungroup() %>%
  mutate(STABBR = ifelse(STABBR == "DC", "district of columbia",
                tolower(state.name[match(STABBR,state.abb)]))
  ) %>%
  ggplot(aes(map_id = STABBR)) +
  geom_map(aes(fill = avg_earn), map = fifty_states) +
  scale_fill_gradient(name = " ") +
  expand_limits(x = fifty_states$long, y = fifty_states$lat) +
  coord_map() + facet_wrap(~Year) + scale_x_continuous(breaks = NULL) +
  scale_y_continuous(breaks = NULL) +
  labs(x = "", y = "", title = "Average debt-to-earnings") +
  theme(legend.position = "bottom", panel.background = element_blank())
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
## Warning in evalq(as.numeric(as.character(GRAD_DEBT_MDN)), <environment>):
## NAs introduced by coercion
```

```
## Warning in evalq(as.numeric(as.character(MD_EARN_WNE_P8)), <environment>):
## NAs introduced by coercion
```

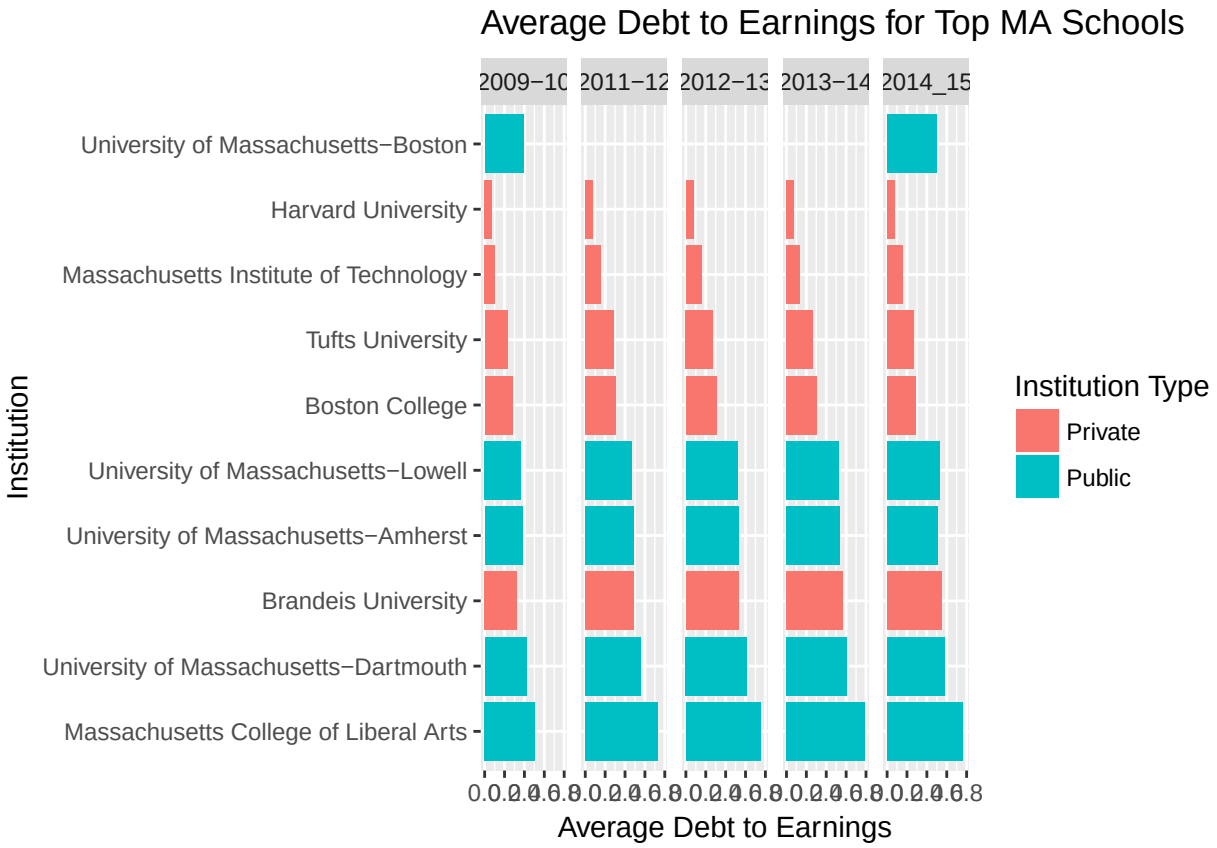



```
# average debt-to-earnings MA schools (Figure 1 in report)
colleges %>%
  filter(INSTNM %in% mass, Year != "2008-09", Year != "2010-11") %>%
  mutate(GRAD_DEBT_MDN = as.numeric(as.character(GRAD_DEBT_MDN)),
    MD_EARN_WNE_P8 = as.numeric(as.character(MD_EARN_WNE_P8)),
    DEBT_TO_EARN = GRAD_DEBT_MDN/MD_EARN_WNE_P8,
    CONTROL = recode(CONTROL, "1" = "Public","2" = "Private")
  ) %>%
  group_by(Year, INSTNM, CONTROL) %>%
```

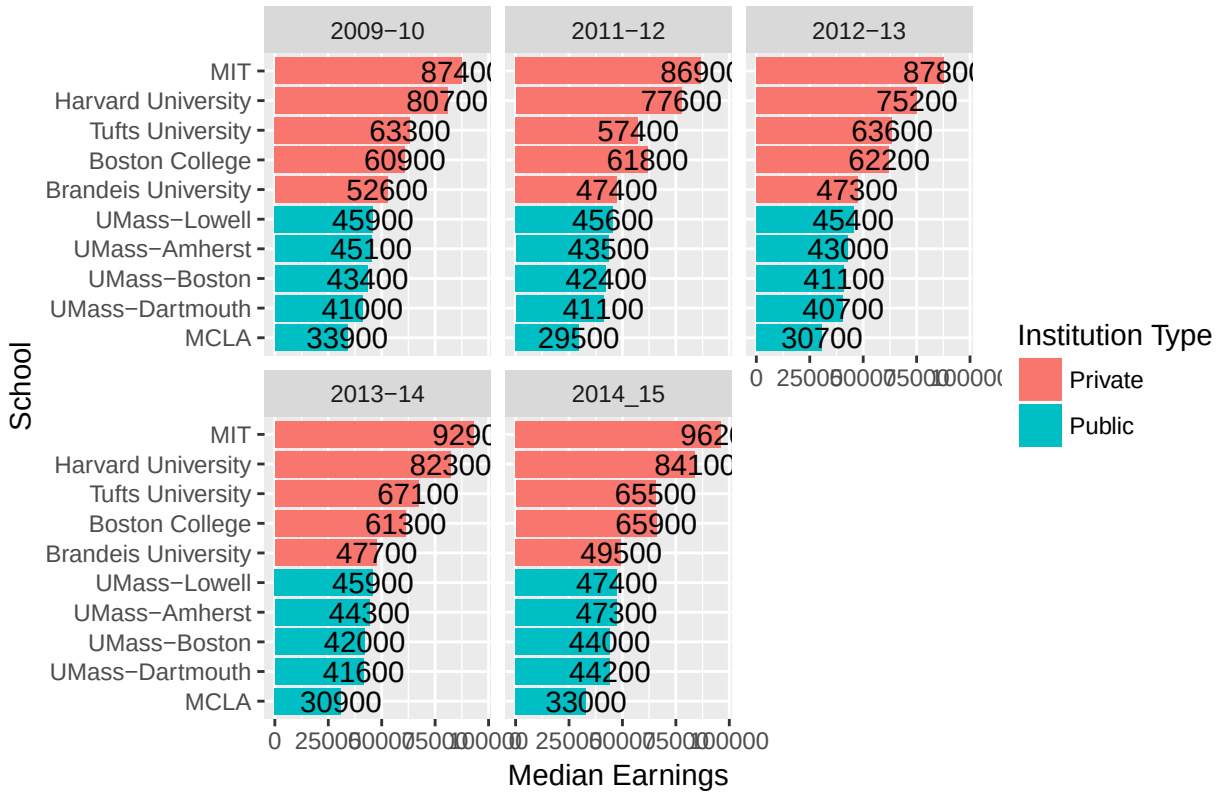
```
  summarise(avg_dte = mean(DEBT_TO_EARN , na.rm = TRUE)) %>% ungroup() %>%
  mutate(INSTNM = reorder(INSTNM, desc(avg_dte))) %>%
  ggplot() + geom_col(aes(x = INSTNM, y = avg_dte, fill = CONTROL), position = "dodge")  + facet_grid(~
labs(title = "Average Debt to Earnings for Top MA Schools", y = "Average Debt to Earnings", x = "Institu
```

```
## Warning in evalq(as.numeric(as.character(GRAD_DEBT_MDN)), <environment>):
## NAs introduced by coercion
```

```
## Warning: Removed 3 rows containing missing values (geom_col).
```




```
# median earnings MA schools (Figure 2 in report)
colleges %>% filter(INSTNM %in% mass, Year != "2008-09", Year != "2010-11") %>%
  mutate(
    MD_EARN_WNE_P8 = as.numeric(as.character(MD_EARN_WNE_P8)),
    CONTROL = recode(CONTROL, "1" = "Public","2" = "Private"), INSTNM = recode(INSTNM, "Massachusetts In
  labs(y = "Median Earnings", x = "School",
title = "Median Earnings Students 8 Years Post-Graduation") + scale_fill_discrete(name  ="Institution Ty
```




```
# cost of attendance - 2016-17 (Figure 3 in report)
loan_1617 <- read_csv("MERGED2013_14_PP.csv", na=(""))
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   UNITID = col_integer(),
##   MAIN = col_integer(),
##   NUMBRANCH = col_integer(),
##   PREDDEG = col_integer(),
##   HIGHDEG = col_integer(),
##   CONTROL = col_integer(),
##   ST_FIPS = col_integer(),
##   REGION = col_integer(),
##   CIP01CERT1 = col_integer(),
##   CIP01CERT2 = col_integer(),
##   CIP01ASSOC = col_integer(),
##   CIP01CERT4 = col_integer(),
##   CIP01BACHL = col_integer(),
##   CIP03CERT1 = col_integer(),
##   CIP03CERT2 = col_integer(),
##   CIP03ASSOC = col_integer(),
##   CIP03CERT4 = col_integer(),
##   CIP03BACHL = col_integer(),
##   CIP04CERT1 = col_integer(),
##   CIP04CERT2 = col_integer()
##   # ... with 180 more columns
```

```
## )
```

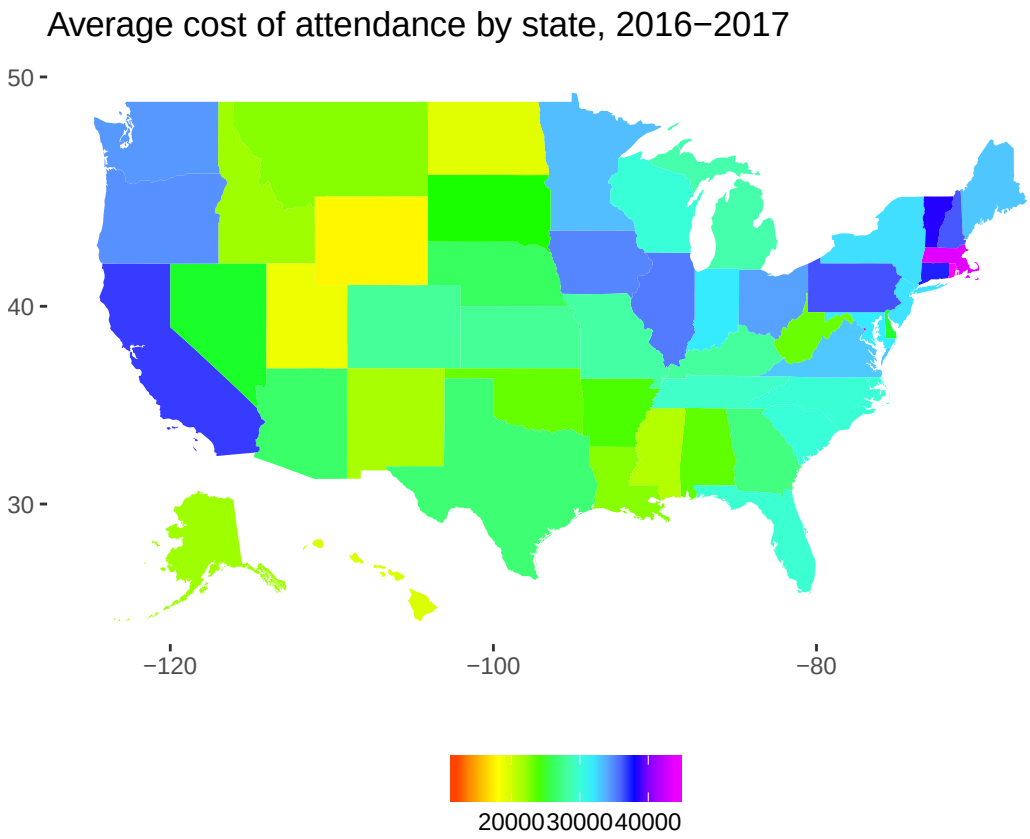
```
## See spec(...) for full column specifications.
```

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 1)
```

```
## Warning: 80411 parsing failures.
## row # A tibble: 5 x 5 col     row col        expected   actual file                    expected   <int
## ... ................. ... ..................................................... ........ .....
## See problems(...) for more details.
```

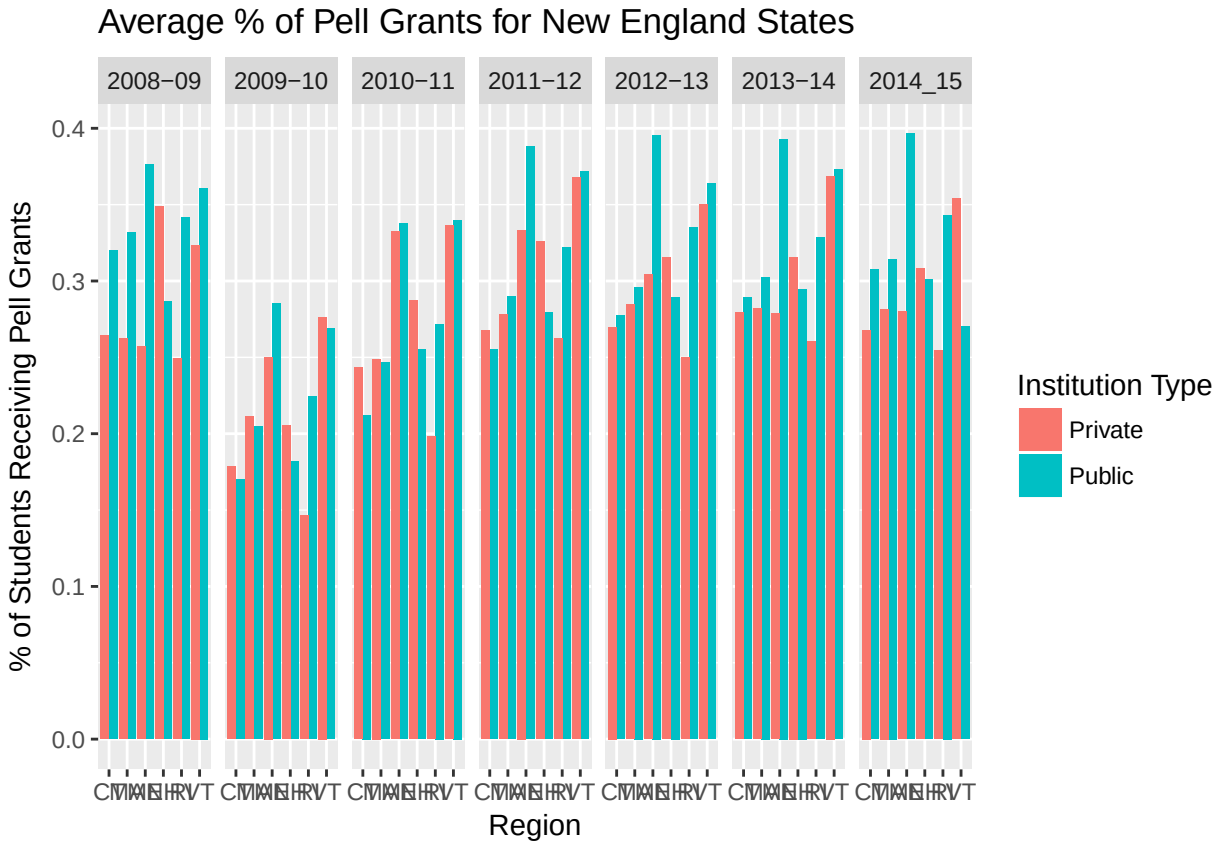
```
loan_1617 %>%
  filter(CONTROL %in% c("1", "2"),PREDDEG == "3") %>%
  mutate(
    cost = as.numeric(COSTT4_A)
  ) %>% select(INSTNM, STABBR,cost) %>%
  group_by(STABBR) %>% summarise(tot_stu = mean(cost, na.rm = TRUE)) %>%
  mutate(
    STABBR = ifelse(STABBR == "DC", "district of columbia", tolower(state.name[match(STABBR,state.abb)]
  ) %>%
  ggplot(aes(map_id = STABBR)) + geom_map(aes(fill = tot_stu), map = fifty_states) + expand_limits(x =
  coord_map() + labs(x = "", y = "", title = "Average cost of attendance by state, 2016-2017") +
  theme(legend.position = "bottom",
        panel.background = element_blank()) + scale_fill_gradientn(name = "", colours=rainbow(6))
```

```
## Warning in evalq(as.numeric(COSTT4_A), <environment>): NAs introduced by
## coercion
```

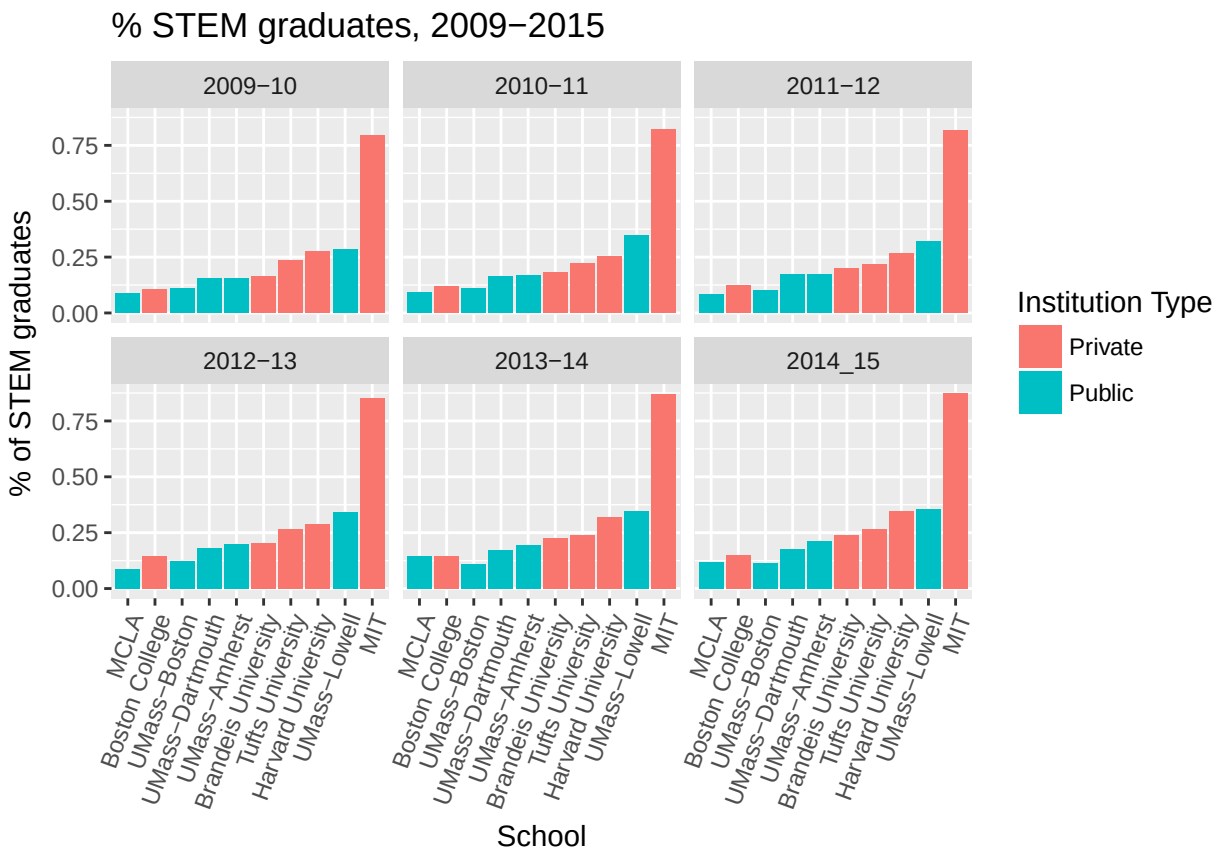


```
# Average percent of Pell Grants by New England state (Figure 4 in report)
colleges %>% filter(CONTROL %in% c("1", "2"), ICLEVEL == "1") %>%
  filter(REGION == "1") %>%
  mutate(PCTPELL = as.numeric(as.character(PCTPELL)),
    CONTROL = recode(CONTROL, "1" = "Public","2" = "Private")) %>%
group_by(Year, CONTROL, STABBR) %>% summarise(avg_loan = mean(PCTPELL, na.rm=TRUE)) %>% ggplot() + geom_
labs(title = "Average % of Pell Grants for New England States", y = "% of Students Receiving Pell Grants
```

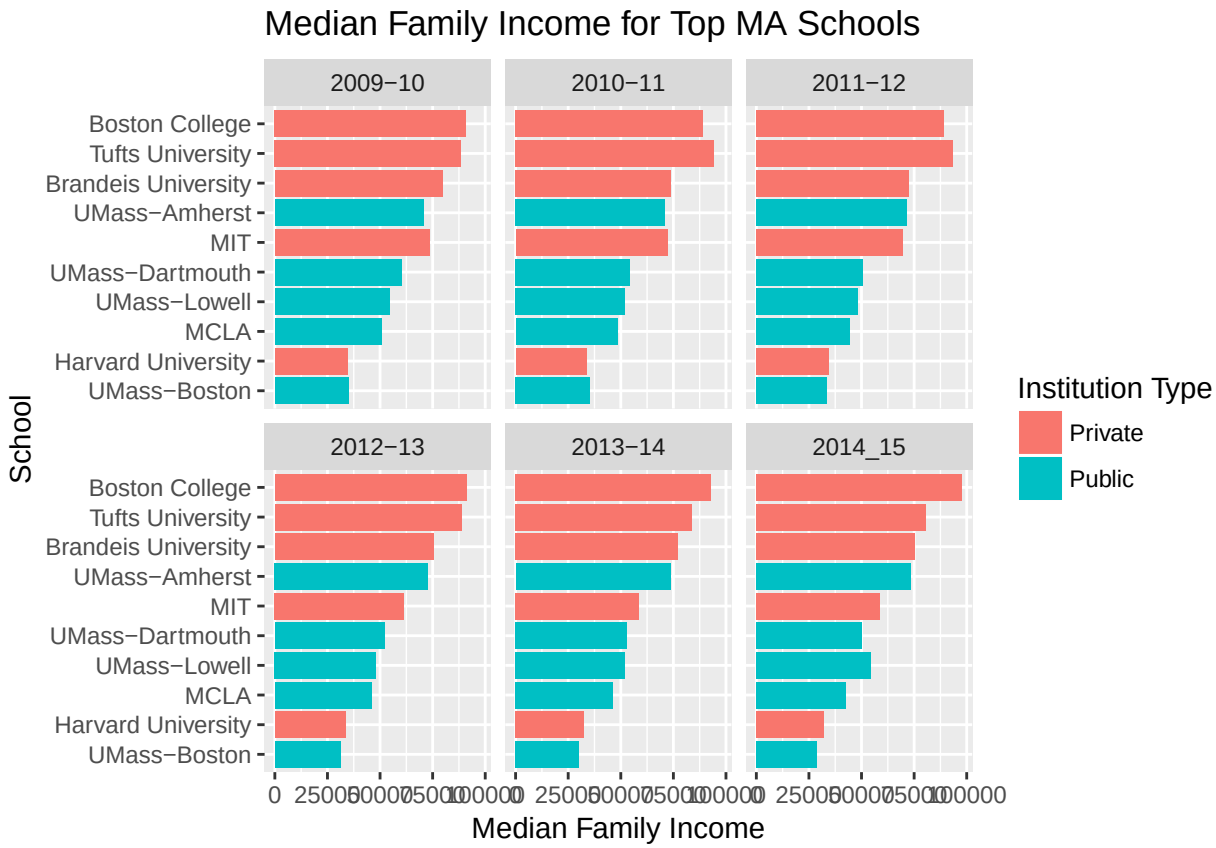
```
## Warning in evalq(as.numeric(as.character(PCTPELL)), <environment>): NAs
## introduced by coercion
```




```
# % of STEM graduates (Figure C in Appendix)
colleges %>% filter(INSTNM %in% mass) %>%
  filter(Year != "2008-09") %>%
  mutate_at(
    vars(starts_with("PCIP")) ,funs(as.numeric(as.character(.)))) %>%
  group_by(Year) %>%
  mutate(stem_pct = PCIP11 + PCIP14 + PCIP15 + PCIP26 + PCIP27 +
           PCIP40 + PCIP41,
         CONTROL = recode(CONTROL, "1" = "Public","2" = "Private"),
         INSTNM = recode(INSTNM, "Massachusetts Institute of Technology" = "MIT", "University of Massach
         INSTNM = reorder(INSTNM, stem_pct)) %>%
  ggplot() + geom_col(aes(x = INSTNM, y = stem_pct, fill = CONTROL)) +
  theme(axis.text.x = element_text(angle = 70, hjust = 1)) +
  facet_wrap(~Year) + labs(y = "% of STEM graduates", x = "School",
                           title = "% STEM graduates, 2009-2015") +
  scale_fill_discrete(name  ="Institution Type")
```




```
# MA median family income (Figure D in Appendix)
colleges %>%
  filter(INSTNM %in% mass, Year != "2008-09") %>%
  select(INSTNM, MD_FAMINC, CONTROL, Year) %>%
  mutate(
    MD_FAMINC = as.numeric(as.character(MD_FAMINC)),
    CONTROL = recode(CONTROL, "1" = "Public","2" = "Private"), INSTNM = recode(INSTNM, "Massachusetts In
  ggplot(aes(INSTNM, MD_FAMINC, fill = CONTROL)) + geom_col() + coord_flip() + labs(title="Median Family
```

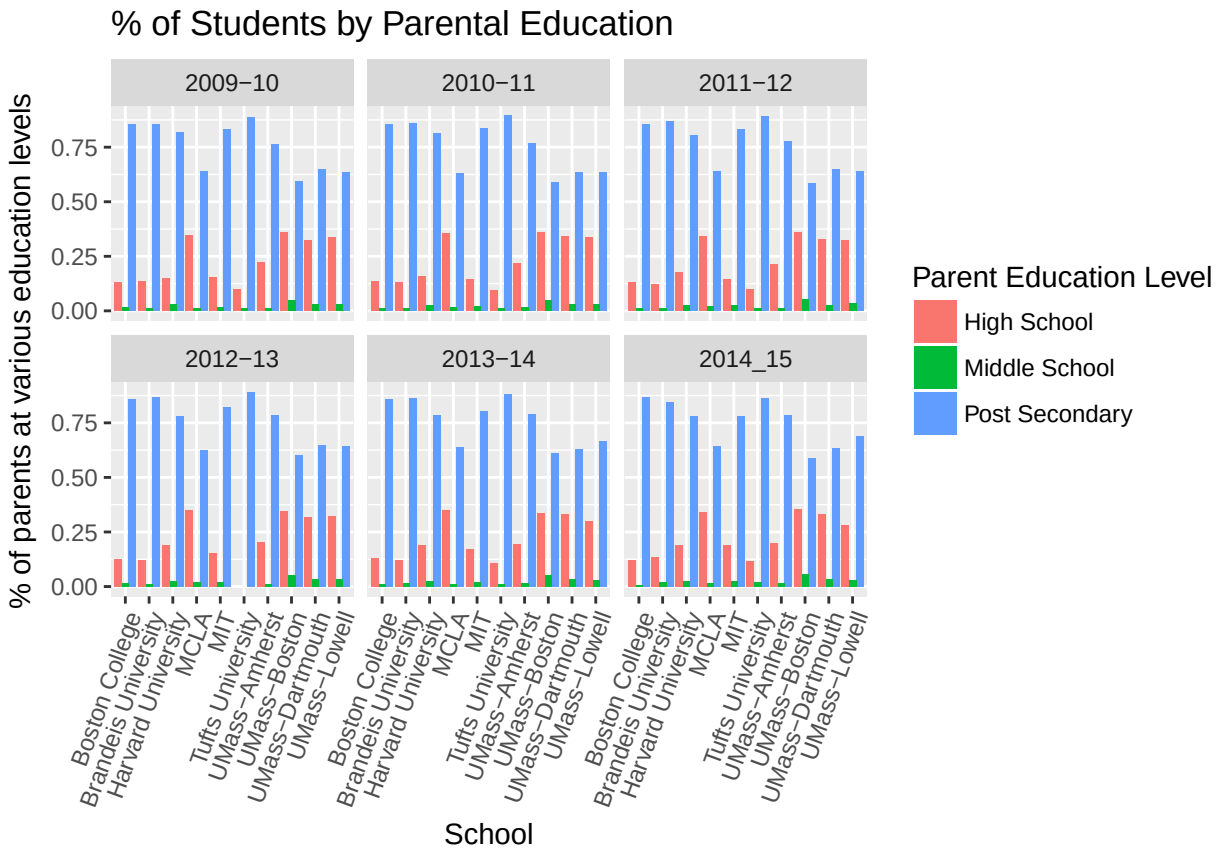



```
# MA family education level (Figure E in Appendix)
colleges %>%
  filter(INSTNM %in% mass, Year != "2008-09") %>%
  gather(PAR_ED_PCT_MS, PAR_ED_PCT_HS,PAR_ED_PCT_PS, key="ParentEdu", value = "Percent") %>%
  select(INSTNM, CONTROL, ParentEdu, Percent, Year) %>%
  mutate(Percent = as.numeric(as.character(Percent)),
         ParentEdu = recode(ParentEdu, PAR_ED_PCT_MS = "Middle School",
                            PAR_ED_PCT_HS = "High School",
                            PAR_ED_PCT_PS = "Post Secondary"),
         INSTNM = recode(INSTNM, "Massachusetts Institute of Technology" = "MIT", "University of Massach
labs(title="% of Students by Parental Education", y = "% of parents at various education levels", x = "
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
## Warning in evalq(as.numeric(as.character(Percent)), <environment>): NAs
## introduced by coercion
```

```
## Warning: Removed 2 rows containing missing values (geom_col).
```


% of Students by Parental Education
2009−10
2010−11
2011−12
2012−13
2013−14
2014_15
% of parents at various education levels
0.00
0.25
0.50
0.75
Boston College
Brandeis University
Harvard University
MCLA
MIT
Tufts University
UMass−Amherst
UMass−Boston
UMass−Dartmouth
UMass−Lowell
School
Parent Education Level
High School
Middle School
Post Secondary