STAT480 Project Title: College tuition, diversity, and pay

Vahid Azizi, Seyedzahra Khoshmanesh, and Saba Moeinizade

4/21/2020

Abstract: Many people are interested to know about the tuition, costs, diversity and potential salary when searching for college. In this project, we want to analyze tuition costs across different states, and explore diversity in different schools. The data set includes different variables such as school name, state, type of school, in-state/out-of-state tuition, group/racial/gender category, early/mid-career pay, stem percent and historical tuition information. We are also interested in trends of tuition over time. We will use different summary statistics and visualizations in R to address these problems.

Getting familaiar with data

```
address<-c('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-03-10/')

tuition_cost <- readr::read_csv(paste(address,'tuition_cost.csv',sep=""))

tuition_income <- readr::read_csv(paste(address,'tuition_income.csv',sep=""))

salary_potential <- readr::read_csv(paste(address,'salary_potential.csv',sep=""))

historical_tuition <- readr::read_csv(paste(address,'historical_tuition.csv',sep=""))

diversity_school <- readr::read_csv(paste(address,'diversity_school.csv',sep=""))
```

```
tuition_cost%>%glimpse()
```

```
## Rows: 2,973
## Columns: 10
## $ name                 <chr> "Aaniiih Nakoda College", "Abilene Christian U...
## $ state                <chr> "Montana", "Texas", "Georgia", "Minnesota", "C...
## $ state_code           <chr> "MT", "TX", "GA", "MN", "CA", "CO", "NY", "NY"...
## $ type                 <chr> "Public", "Private", "Public", "For Profit", "...
## $ degree_length        <chr> "2 Year", "4 Year", "2 Year", "2 Year", "4 Yea...
## $ room_and_board       <dbl> NA, 10350, 8474, NA, 16648, 8782, 16030, 11660...
## $ in_state_tuition     <dbl> 2380, 34850, 4128, 17661, 27810, 9440, 38660, ...
## $ in_state_total       <dbl> 2380, 45200, 12602, 17661, 44458, 18222, 54690...
## $ out_of_state_tuition <dbl> 2380, 34850, 12550, 17661, 27810, 20456, 38660...
## $ out_of_state_total   <dbl> 2380, 45200, 21024, 17661, 44458, 29238, 54690...
```

```
tuition_income%>%glimpse()
```

```
## Rows: 209,012
## Columns: 7
## $ name        <chr> "Piedmont International University", "Piedmont Internat...
## $ state       <chr> "NC", "NC", "NC", "NC", "NC", "NC", "NC", "NC", "NC", "...
## $ total_price <dbl> 20174, 20174, 20174, 20174, 20514, 20514, 20514, 20514,...
## $ year        <dbl> 2016, 2016, 2016, 2016, 2017, 2017, 2017, 2017, 2017, 2...
## $ campus      <chr> "On Campus", "On Campus", "On Campus", "On Campus", "On...
## $ net_cost    <dbl> 11475.00, 11451.00, 16229.00, 15592.00, 11668.39, 11643...
## $ income_lvl  <chr> "0 to 30,000", "30,001 to 48,000", "48_001 to 75,000", ...
```

```
salary_potential%>%glimpse()
```

```
## Rows: 935
## Columns: 7
## $ rank                     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13...
## $ name                     <chr> "Auburn University", "University of Alaba...
## $ state_name               <chr> "Alabama", "Alabama", "Alabama", "Alabama...
## $ early_career_pay         <dbl> 54400, 57500, 52300, 54500, 48400, 46600,...
## $ mid_career_pay           <dbl> 104500, 103900, 97400, 93500, 90500, 8910...
## $ make_world_better_percent <dbl> 51, 59, 50, 61, 52, 53, 48, 57, 56, 58, 6...
## $ stem_percent             <dbl> 31, 45, 15, 30, 3, 12, 27, 17, 17, 20, 8,...
```

```
historical_tuition%>%glimpse()
```

```
## Rows: 270
## Columns: 4
## $ type         <chr> "All Institutions", "All Institutions", "All Instituti...
## $ year         <chr> "1985-86", "1985-86", "1985-86", "1985-86", "1985-86",...
## $ tuition_type <chr> "All Constant", "4 Year Constant", "2 Year Constant", ...
## $ tuition_cost <dbl> 10893, 12274, 7508, 4885, 5504, 3367, 13822, 16224, 74...
```

```
diversity_school%>%glimpse()
```

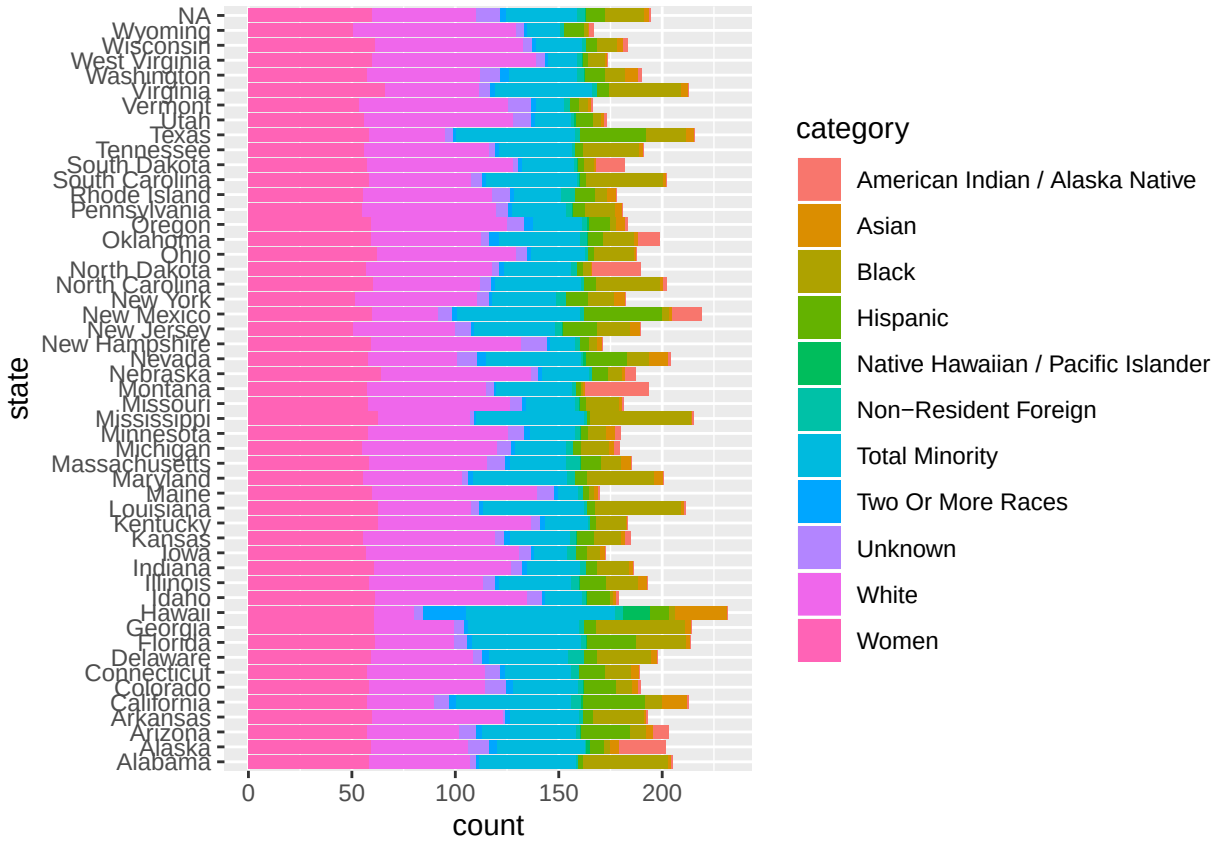
```
## Rows: 50,655
## Columns: 5
## $ name             <chr> "University of Phoenix-Arizona", "University of Ph...
## $ total_enrollment <dbl> 195059, 195059, 195059, 195059, 195059, 195059, 19...
## $ state            <chr> "Arizona", "Arizona", "Arizona", "Arizona", "Arizo...
## $ category         <chr> "Women", "American Indian / Alaska Native", "Asian...
## $ enrollment       <dbl> 134722, 876, 1959, 31455, 13984, 1019, 58209, 1903...
```

The relationship between college tuition and campus diversity

```
#merged<-dplyr::left_join(tuition_cost, diversity_school, by = c("name", "state"))
```

First, we look at the availabe diverse categories captured in schools across the US.

```
diversity_school %>%
  group_by(state,category) %>%
  mutate(enrollment_rate=round((enrollment/total_enrollment),2)*100) %>%
  summarise(mean_enrollment_rate_per_state=mean(enrollment_rate)) %>%
  arrange(desc(mean_enrollment_rate_per_state)) %>%
    ggplot(aes(x=state,weight=mean_enrollment_rate_per_state,fill=category)) +
  geom_bar()+
  coord_flip()
```



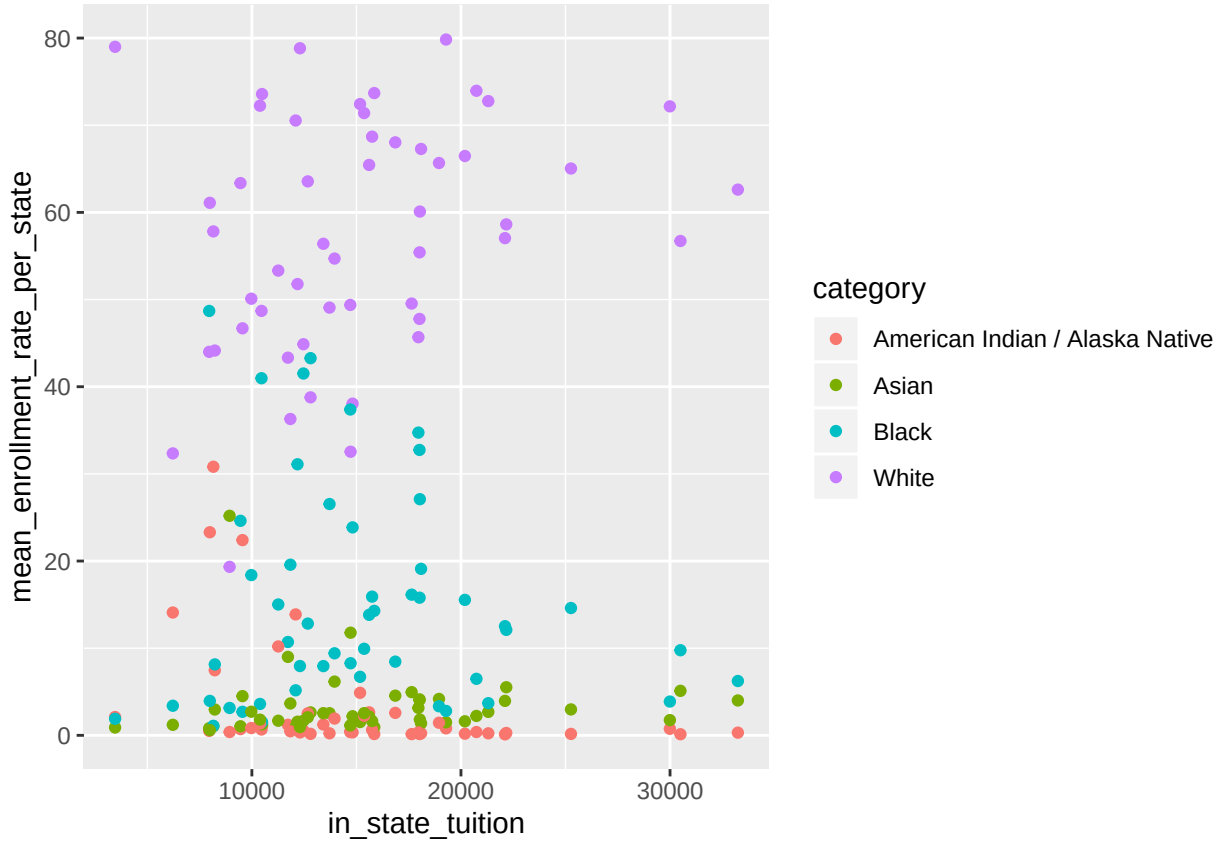
Let us create a new variable which indicates the enrollment rate per each category for each state in the US.

```
diversity_enroll_rate <- diversity_school %>%
  group_by(state,category) %>%
  mutate(enrollment_rate=round((enrollment/total_enrollment),2)*100) %>%
  summarise(mean_enrollment_rate_per_state=mean(enrollment_rate))
diversity_enroll_rate %>% glimpse()
```

```
## Rows: 561
## Columns: 3
## Groups: state [51]
## $ state                          <chr> "Alabama", "Alabama", "Alabama", "Al...
## $ category                       <chr> "American Indian / Alaska Native", "...
## $ mean_enrollment_rate_per_state <dbl> 0.67123288, 1.26027397, 40.97260274,...
```

Now, let us explore any relationship between the enrollment rate and tuition for each diverse category. We merge tuition_cost dataset with diversity dataset.

```
tuition_cost %>%
  group_by(state) %>%
  summarise(in_state_tuition=mean(in_state_tuition),
            out_of_state_tuition=mean(out_of_state_tuition)) %>%
   merge(diversity_enroll_rate) %>%
   filter(category %in% c("Asian","White","Black","American Indian / Alaska Native")) %>%
  ggplot(aes(x=in_state_tuition,y=mean_enrollment_rate_per_state, color=category)) +
  geom_point()
```



The historical trend of tuition

```
historical_tuition%>%distinct(year)%>%arrange()
```

```
## # A tibble: 19 x 1
##    year
##    <chr>
##  1 1985-86
##  2 1995-96
##  3 2000-01
##  4 2001-02
##  5 2002-03
##  6 2003-04
##  7 2004-05
```

```
##  8 2005-06
##  9 2006-07
## 10 2007-08
## 11 2008-09
## 12 2009-10
## 13 2010-11
## 14 2011-12
## 15 2012-13
## 16 2013-14
## 17 2014-15
## 18 2015-16
## 19 2016-17
```

```
historical_tuition%>%distinct(type)
```

```
## # A tibble: 3 x 1
##   type
##   <chr>
## 1 All Institutions
## 2 Public
## 3 Private
```

```
historical_tuition%>%distinct(tuition_type)
```

```
## # A tibble: 6 x 1
##   tuition_type
##   <chr>
## 1 All Constant
## 2 4 Year Constant
## 3 2 Year Constant
## 4 All Current
## 5 4 Year Current
## 6 2 Year Current
```

```
table(historical_tuition$type)
```

```
##
## All Institutions          Private           Public
##              114               78               78
```

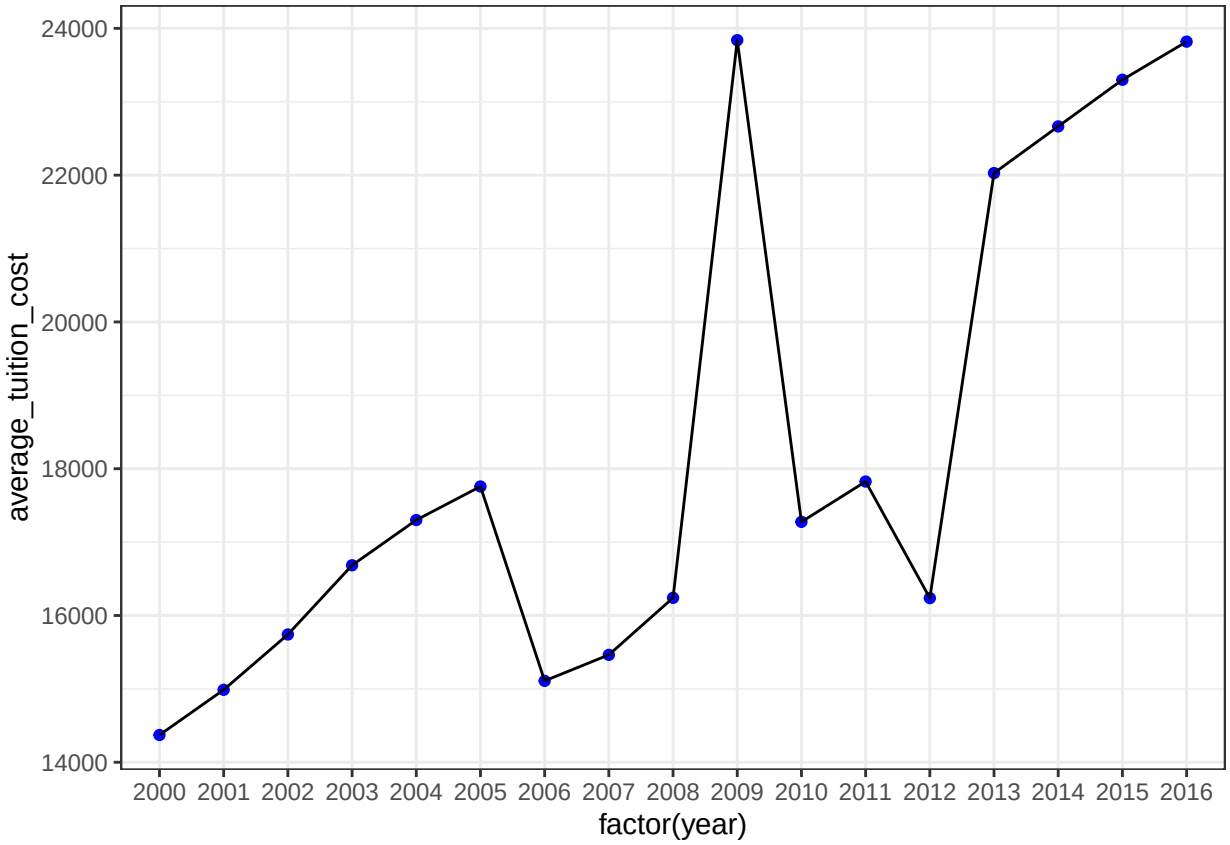
```
table(historical_tuition$tuition_type)
```

```
##
## 2 Year Constant  2 Year Current 4 Year Constant  4 Year Current    All Constant
##              45              45              45              45              45
##     All Current
##              45
```

The historical tuition is provided across three types of All Institutions, Public, and Private universities. We observe that the data is available consistantly from 2000 till 2017. For convinience, we will manupliate this column to have the starting year only (e.g., 2000-01 will be 2000).

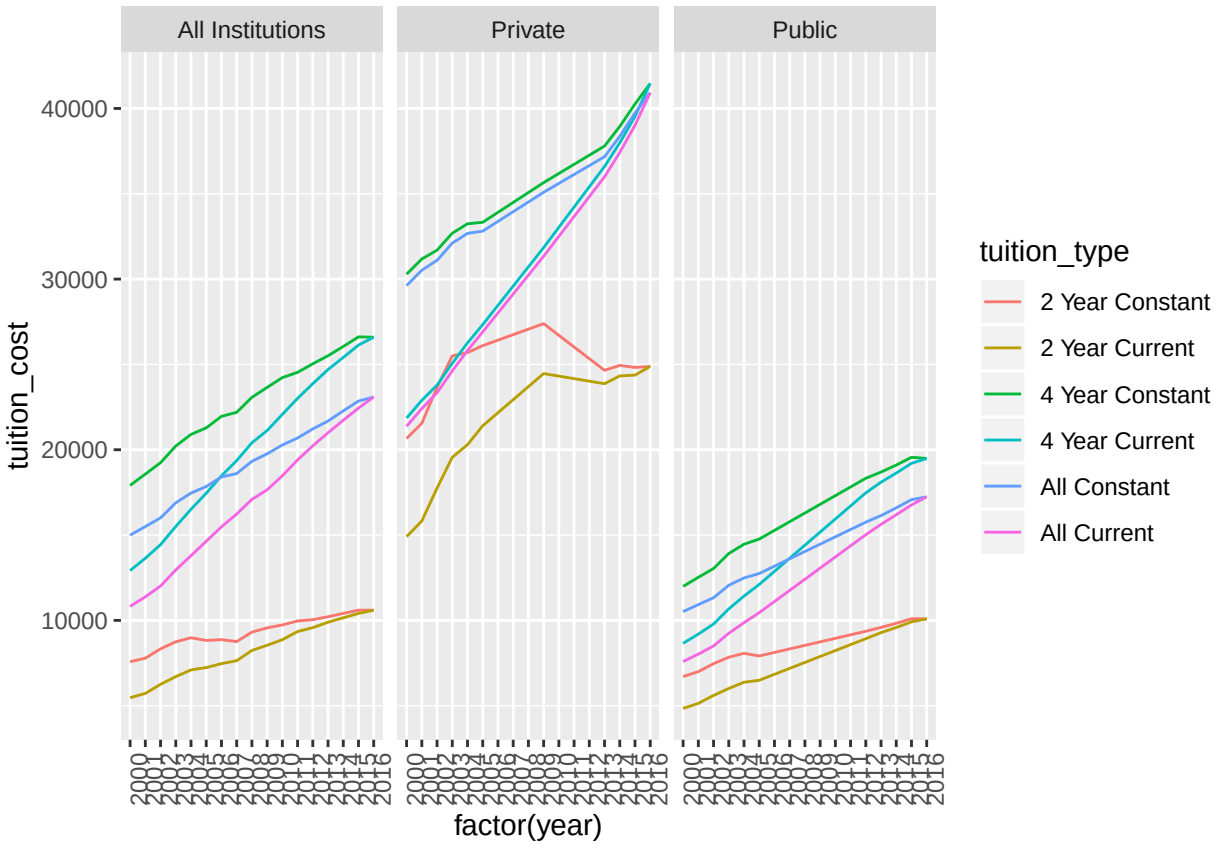
```
historical_tuition<-separate(data=historical_tuition,col = 2,c("year","year_next"))%>%select(-`year_nex
```

```
historical_tuition%>%filter(year>=2000)%>%group_by(year)%>%summarise(average_tuition_cost=mean(tuition_
 ggplot(aes(x=factor(year),y=average_tuition_cost,group=1))+geom_point(color='blue')+geom_line()+theme_
```



The average tuition is increased from 2000 till 2005, then there is a sudden decrease and after that we see a huge jump from 2008 to 2009. This follows by a sudden drop and then again increases in 2012.

```
historical_tuition%>%filter(year>=2000)%>%
 ggplot(aes(x=factor(year),y=tuition_cost,group=tuition_type,color=tuition_type))+geom_line()+facet_wra
  theme(axis.text.x = element_text(angle = 90,hjust=1))
```

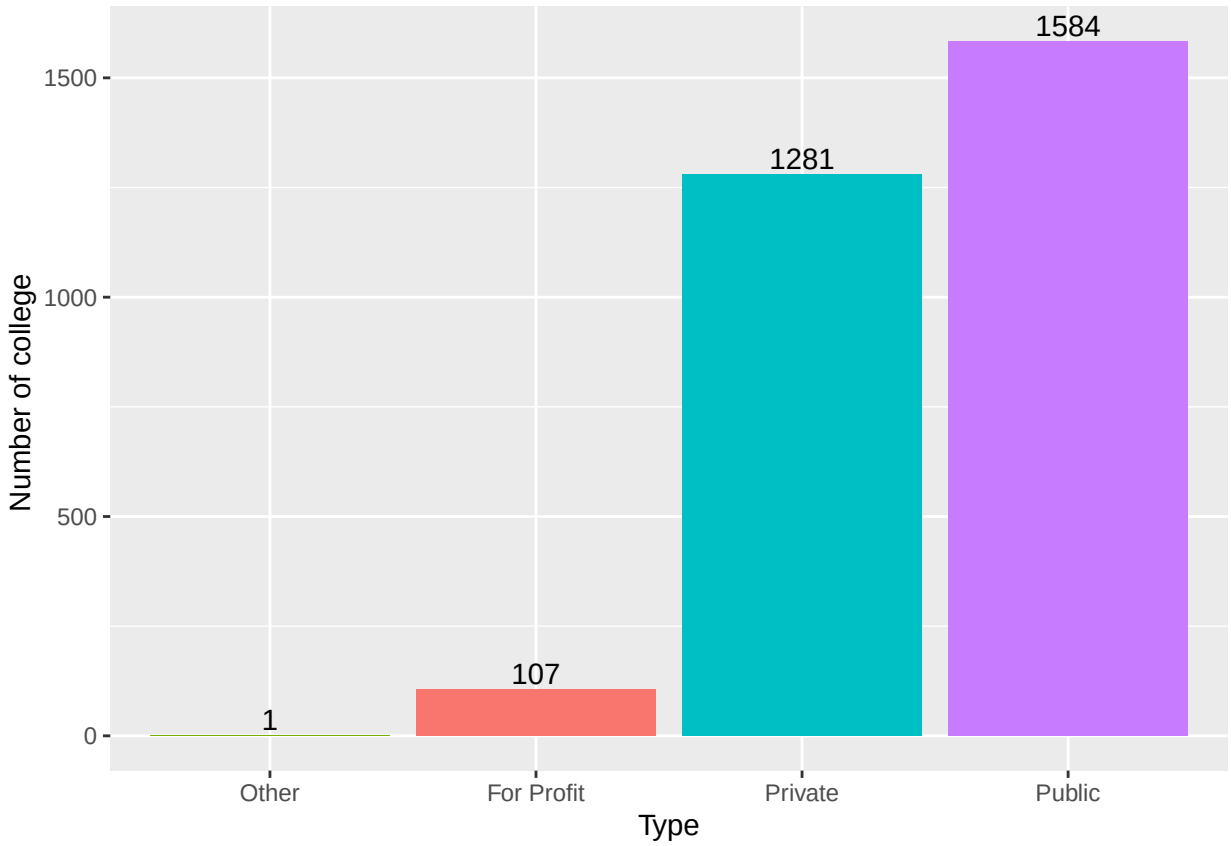


Cost of college tuition in US by geographic area

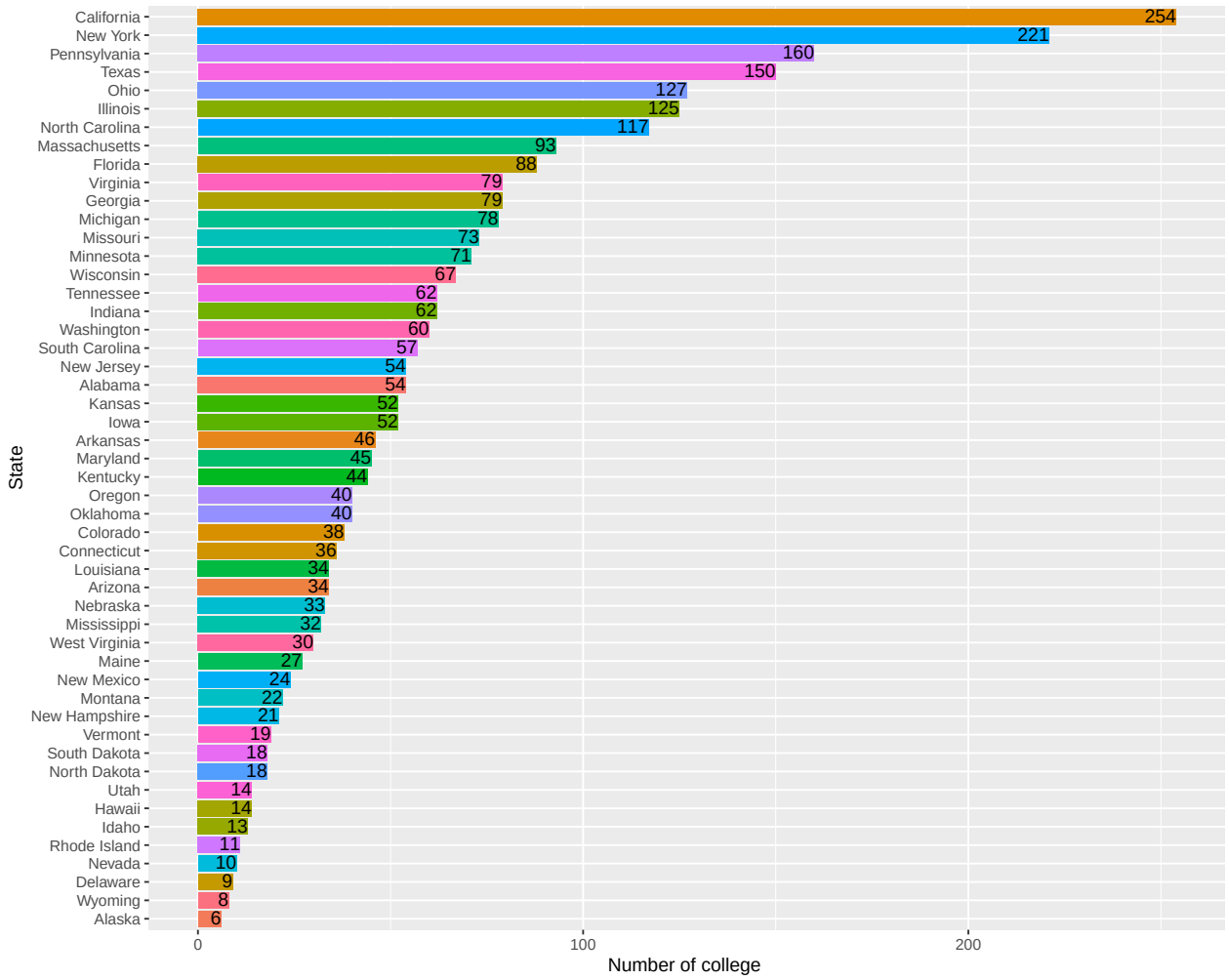
```
head(tuition_cost)
```

```
## # A tibble: 6 x 10
##   name  state state_code type  degree_length room_and_board in_state_tuition
##   <chr> <chr> <chr>      <chr> <chr>                  <dbl>            <dbl>
## 1 Aani~ Mont~ MT         Publ~ 2 Year                    NA             2380
## 2 Abil~ Texas TX         Priv~ 4 Year                 10350            34850
## 3 Abra~ Geor~ GA         Publ~ 2 Year                  8474             4128
## 4 Acad~ Minn~ MN         For ~ 2 Year                    NA            17661
## 5 Acad~ Cali~ CA         For ~ 4 Year                 16648            27810
## 6 Adam~ Colo~ CO         Publ~ 4 Year                  8782             9440
## # ... with 3 more variables: in_state_total <dbl>, out_of_state_tuition <dbl>,
## #   out_of_state_total <dbl>
```

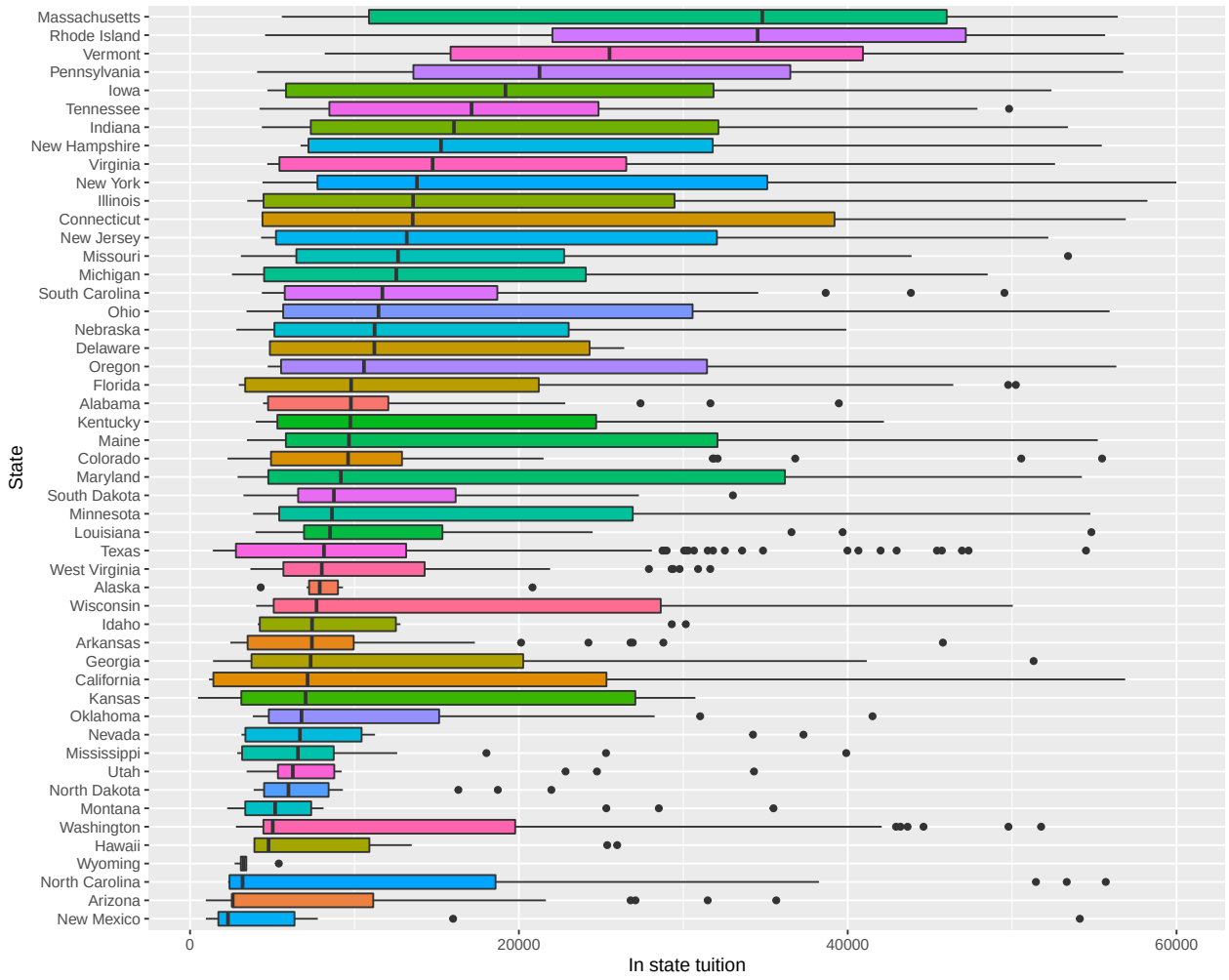
```
library(forcats)

tuition_cost %>% group_by(type) %>% summarise(n=n()) %>% ggplot(aes(x=fct_reorder(type,n), y=n,fill=as.
  geom_text(aes(label=n), position=position_dodge(width=0.9), vjust=-0.25)+
  theme(legend.position = "none")
```



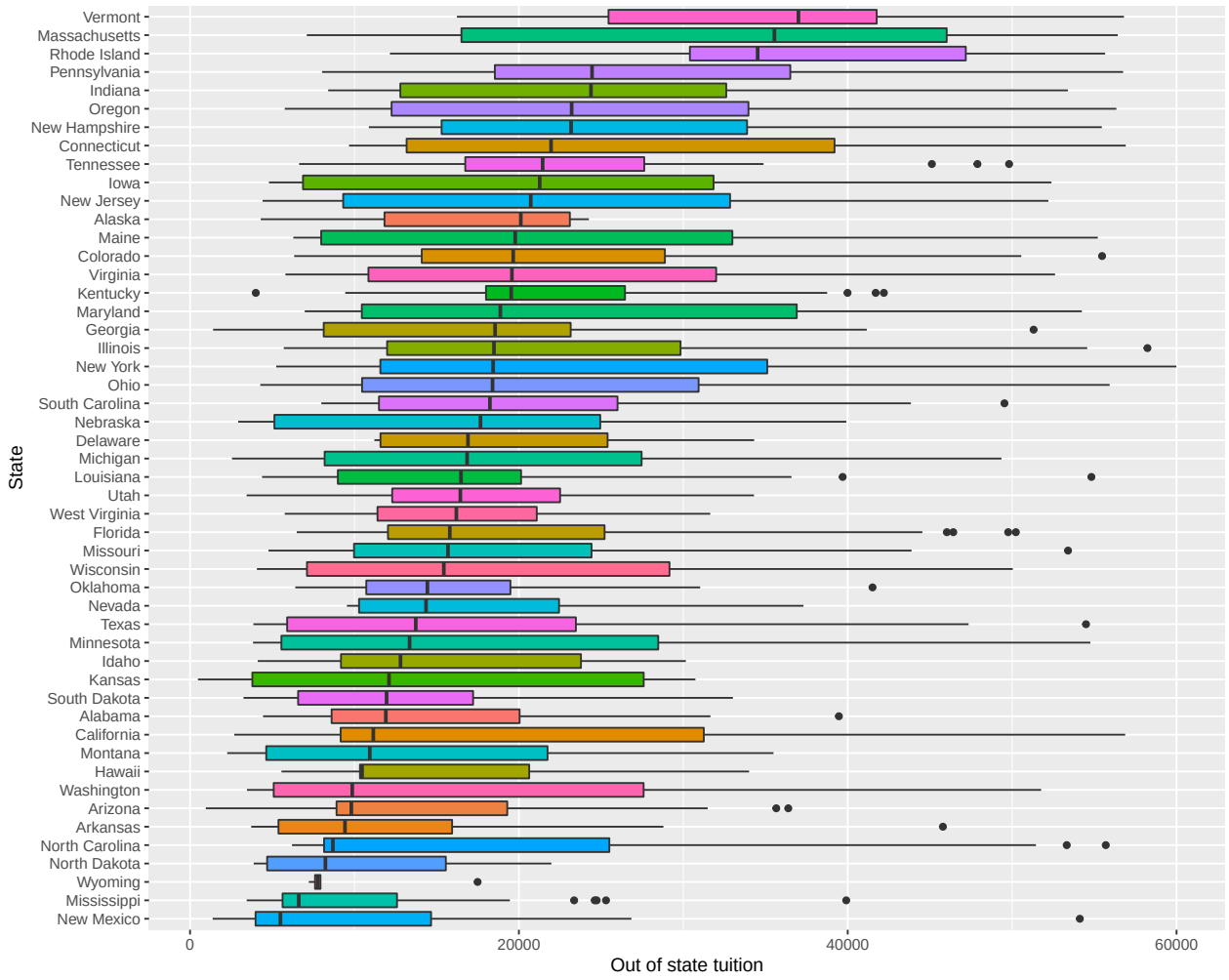

```
tuition_cost[!is.na(tuition_cost$state),] %>% group_by(state) %>% summarise(n=n()) %>% ggplot(aes(x=fct_
  geom_text(aes(label=n), position=position_dodge(width=.3),hjust=1, vjust=.4)+
  coord_flip()+
  theme(legend.position = "none")
```



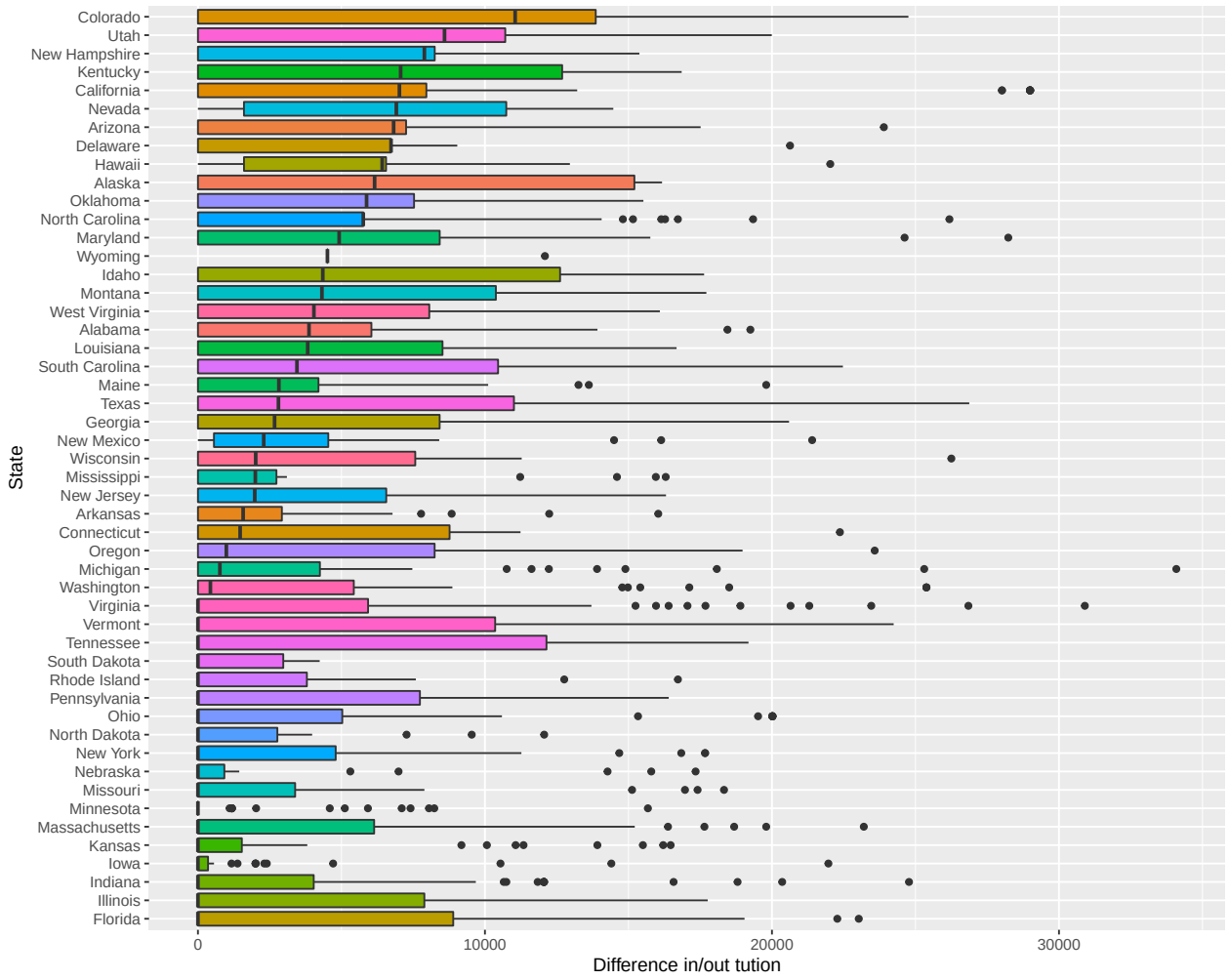

```
tuition_cost[!is.na(tuition_cost$state),] %>% group_by(state) %>% ggplot(aes(x=fct_reorder(state,in_sta
  geom_boxplot()+
  coord_flip() +
  labs(x="State",y="In state tuition")+
  theme(legend.position = "none")
```



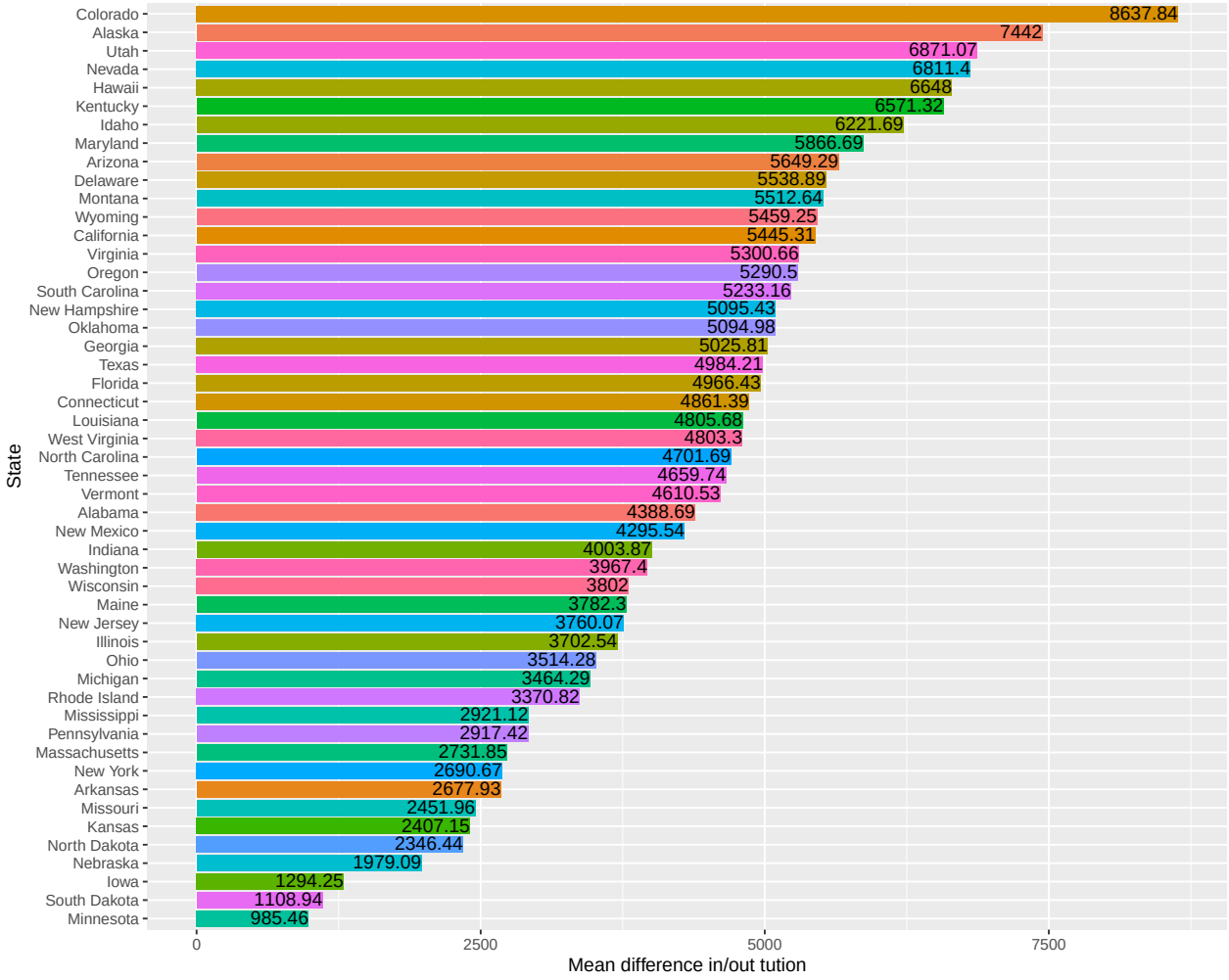

```
tuition_cost[!is.na(tuition_cost$state),] %>% group_by(state) %>% ggplot(aes(x=fct_reorder(state,out_of_
  geom_boxplot()+
  coord_flip() +
  labs(x="State",y="Out of state tuition")+
  theme(legend.position = "none")
```




```
tuition_cost[!is.na(tuition_cost$state),] %>% mutate(tuition_diff_in_out=out_of_state_tuition-in_state_
  geom_boxplot()+
  coord_flip() +
  labs(x="State",y="Difference in/out tution")+
  theme(legend.position = "none")
```




```
tuition_cost[!is.na(tuition_cost$state),] %>% mutate(tuition_diff_in_out=out_of_state_tuition-in_state_
  summarise(mean_diff=round(mean(tuition_diff_in_out),2)) %>%
  ggplot(aes(x=fct_reorder(state,mean_diff), y=mean_diff,fill=as.factor(state))) +
  geom_bar(stat="identity")+
  geom_text(aes(label=mean_diff), position=position_dodge(width=.3),hjust=1, vjust=.4)+
  coord_flip() +
  labs(x="State",y="Mean difference in/out tution")+
  theme(legend.position = "none")
```



Average potential salary in early/mid-career

```
salary_potential%>%distinct(state_name)%>%count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    50
```

This data includes early/mid career pay across 50 states. Let's take a look at universities in Iowa and compare their potential slary.

```
salary_potential%>%filter(state_name=="Iowa")%>%arrange(name)
```

```
## # A tibble: 25 x 7
##     rank name  state_name early_career_pay mid_career_pay make_world_bett~
##    <dbl> <chr> <chr>                <dbl>          <dbl>            <dbl>
##  1     5 Alle~ Iowa                 51600          91300               NA
```

```
##  2     9 Cent~ Iowa                    49000          88800                55
##  3    20 Clar~ Iowa                    44900          80200                52
##  4    13 Coe ~ Iowa                    47200          86900                41
##  5     6 Corn~ Iowa                    49600          90600                47
##  6     8 Dord~ Iowa                    51000          89600                57
##  7     2 Drak~ Iowa                    52800          99900                49
##  8    10 Grac~ Iowa                    48300          88700                66
##  9     4 Grin~ Iowa                    53400          96500                49
## 10     1 Iowa~ Iowa                    56100         101300                48
## # ... with 15 more rows, and 1 more variable: stem_percent <dbl>
```

```
salary_potential%>%filter(state_name=="Iowa")%>%select(name,early_career_pay,mid_career_pay)%>%reshape2
  geom_bar(aes(x=factor(name),y=value,fill=variable),stat='identity',position = 'dodge')+coord_flip()
```

```
## Using name as id variables
```