

# data151eda2

Tyler Bontrager

2022-10-25

```
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
?ggplot2
```

## *# IMPORTING DATASETS*

```
tuition_cost <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/d
```

```
## Rows: 2973 Columns: 10
## -- Column specification -----
## Delimiter: ","
## chr (5): name, state, state_code, type, degree_length
## dbl (5): room_and_board, in_state_tuition, in_state_total, out_of_state_tuit...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
tc = tuition_cost
```

```
tuition_income <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master
```

```
## Rows: 209012 Columns: 7
## -- Column specification -----
## Delimiter: ","
## chr (4): name, state, campus, income_lvl
## dbl (3): total_price, year, net_cost
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ti = tuition_income
```

```
salary_potential <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/mast
```

```
## Rows: 935 Columns: 7
## -- Column specification -----
```

```

## Delimiter: ","
## chr (2): name, state_name
## dbl (5): rank, early_career_pay, mid_career_pay, make_world_better_percent, ...
##
## i Use `spec()`` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
sp = salary_potential

historical_tuition <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/07/data/historical_tuition.csv')

## Rows: 270 Columns: 4
## -- Column specification -----
## Delimiter: ","
## chr (3): type, year, tuition_type
## dbl (1): tuition_cost
##
## i Use `spec()`` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
ht = historical_tuition

diversity_school <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/07/data/diversity_school.csv')

## Rows: 50655 Columns: 5
## -- Column specification -----
## Delimiter: ","
## chr (3): name, state, category
## dbl (2): total_enrollment, enrollment
##
## i Use `spec()`` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
ds = diversity_school

# Time to explore the data!

table(tc$state,tc$degree_length)

##
##           2 Year 4 Year Other
## Alabama         21    33    0
## Alaska           1     5    0
## Arizona          23    11    0
## Arkansas         24    22    0
## California      119   135    0
## Colorado         18    20    0
## Connecticut      14    22    0
## Delaware         4     5    0
## Florida          33    55    0
## Georgia          29    50    0
## Hawaii           8     6    0
## Idaho            4     9    0
## Illinois         52    73    0
## Indiana          18    44    0
## Iowa             18    34    0
## Kansas           25    27    0

```

```
## Kentucky      15    29    0
## Louisiana      8    26    0
## Maine          9    18    0
## Maryland       16    29    0
## Massachusetts  21    72    0
## Michigan       30    48    0
## Minnesota      33    38    0
## Mississippi    15    17    0
## Missouri       23    50    0
## Montana        11    11    0
## Nebraska       10    23    0
## Nevada         4     6    0
## New Hampshire  7     14   0
## New Jersey     21    33    0
## New Mexico     14    10    0
## New York       58   163    0
## North Carolina  59    58    0
## North Dakota   9     9     0
## Ohio           47    80    0
## Oklahoma       15    25    0
## Oregon         15    25    0
## Pennsylvania   31   129    0
## Rhode Island   1     10    0
## South Carolina  23    34    0
## South Dakota   5     13    0
## Tennessee      17    45    0
## Texas          67    82    1
## Utah           4     10    0
## Vermont        3     16    0
## Virginia       30    49    0
## Washington     33    27    0
## West Virginia  9     21    0
## Wisconsin      31    36    0
## Wyoming        7     1     0
```

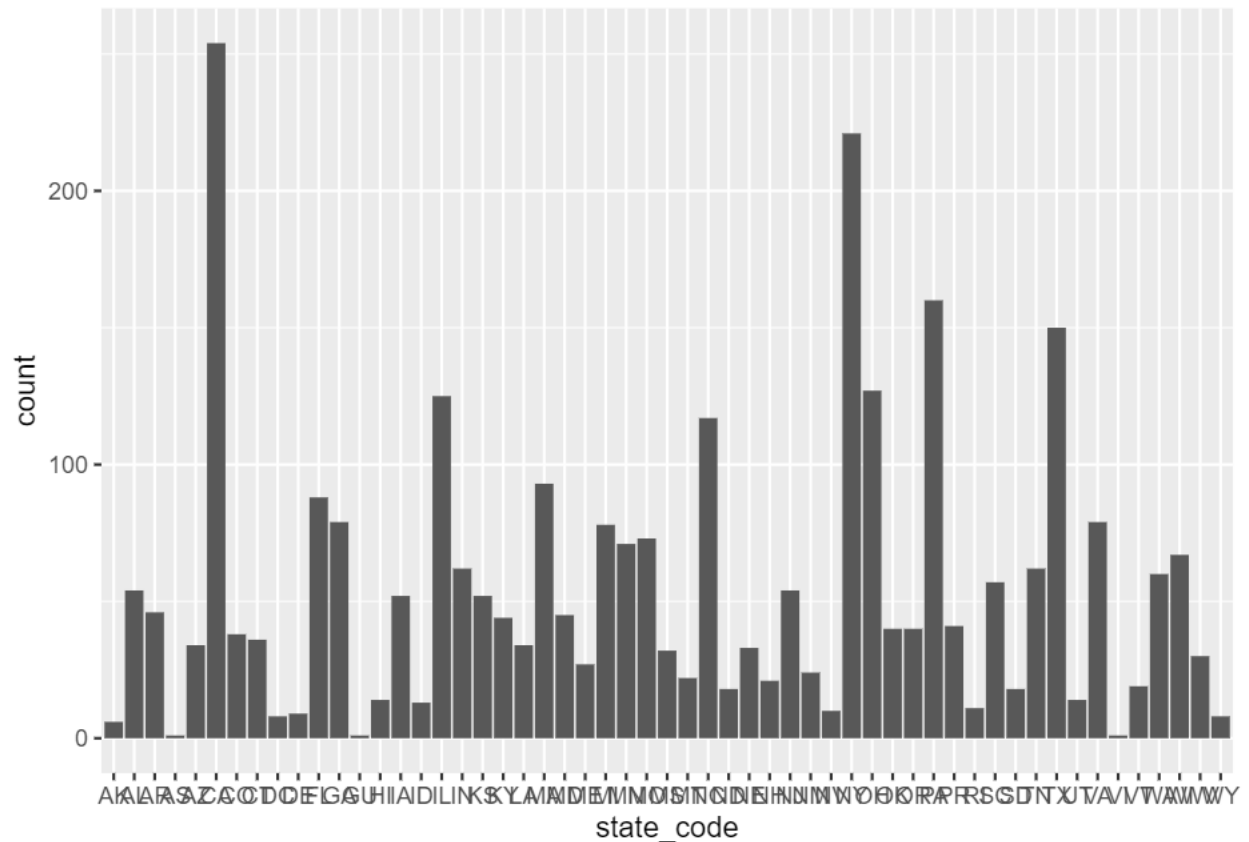
```
table(tc$state)
```

```
##
##      Alabama      Alaska      Arizona      Arkansas      California
##      54           6         34         46           254
##      Colorado    Connecticut    Delaware      Florida      Georgia
##      38           36          9         88           79
##      Hawaii      Idaho      Illinois      Indiana      Iowa
##      14           13        125         62           52
##      Kansas      Kentucky    Louisiana      Maine      Maryland
##      52           44         34         27           45
##      Massachusetts    Michigan    Minnesota    Mississippi    Missouri
##      93           78         71         32           73
##      Montana      Nebraska      Nevada    New Hampshire    New Jersey
##      22           33          10         21           54
##      New Mexico    New York    North Carolina    North Dakota      Ohio
##      24           221        117         18           127
##      Oklahoma      Oregon    Pennsylvania    Rhode Island    South Carolina
##      40           40        160         11           57
##      South Dakota    Tennessee      Texas         Utah      Vermont
```

```
##          18          62          150          14          19
##   Virginia   Washington West Virginia   Wisconsin   Wyoming
##          79          60          30          67          8
```

This is a graph of the number of higher-education schools in each U.S. territory and state.

```
ggplot(tc, aes(x=state_code))+
  geom_bar()
```



```
#indexing each row in the table to have a unique identifier
tc$index <- 1:nrow(tc)
```

```
tc
```

```
## # A tibble: 2,973 x 11
##   name      state state-1 type  degree-2 room_~3 in_st-4 in_st-5 out_o-6 out_o-7
##   <chr>    <chr> <chr>  <chr> <chr>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 Aaniiih ~ Mont~ MT      Publ~ 2 Year      NA    2380    2380    2380    2380
## 2 Abilene ~ Texas TX       Priv~ 4 Year   10350  34850  45200  34850  45200
## 3 Abraham ~ Geor~ GA       Publ~ 2 Year    8474   4128  12602  12550  21024
## 4 Academy ~ Minn~ MN       For ~ 2 Year      NA  17661  17661  17661  17661
## 5 Academy ~ Cali~ CA       For ~ 4 Year  16648  27810  44458  27810  44458
## 6 Adams St~ Colo~ CO       Publ~ 4 Year    8782   9440  18222  20456  29238
## 7 Adelphi ~ New ~ NY       Priv~ 4 Year  16030  38660  54690  38660  54690
## 8 Adironda~ New ~ NY       Publ~ 2 Year  11660   5375  17035   9935  21595
## 9 Adrian C~ Mich~ MI       Priv~ 4 Year  11318  37087  48405  37087  48405
## 10 Advanced~ Virg~ VA       For ~ 2 Year      NA  13680  13680  13680  13680
## # ... with 2,963 more rows, 1 more variable: index <int>, and abbreviated
## #   variable names 1: state_code, 2: degree_length, 3: room_and_board,
```

```
## # 4: in_state_tuition, 5: in_state_total, 6: out_of_state_tuition,
## # 7: out_of_state_total

## Joint distributions
#tc2way<-tc %>%
# group_by(state_code, degree_length)%>%
# mutate(freq=sum(Freq))

#tc2way

tc_with_count2 = tc %>%
  group_by(state_code) %>%
  mutate(school_count = n())

tc_with_count2

## # A tibble: 2,973 x 12
## # Groups:   state_code [55]
##   name      state state-1 type  degre-2 room_~3 in_st~4 in_st~5 out_o~6 out_o~7
##   <chr>      <chr> <chr>  <chr> <chr>      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 Aaniiih ~ Mont~ MT      Publ~ 2 Year      NA      2380    2380    2380    2380
## 2 Abilene ~ Texas TX      Priv~ 4 Year    10350   34850   45200   34850   45200
## 3 Abraham ~ Geor~ GA      Publ~ 2 Year     8474    4128   12602   12550   21024
## 4 Academy ~ Minn~ MN      For ~ 2 Year      NA   17661   17661   17661   17661
## 5 Academy ~ Cali~ CA      For ~ 4 Year   16648   27810   44458   27810   44458
## 6 Adams St~ Colo~ CO      Publ~ 4 Year     8782    9440   18222   20456   29238
## 7 Adelphi ~ New ~ NY      Priv~ 4 Year   16030   38660   54690   38660   54690
## 8 Adironda~ New ~ NY      Publ~ 2 Year   11660    5375   17035    9935   21595
## 9 Adrian C~ Mich~ MI      Priv~ 4 Year   11318   37087   48405   37087   48405
## 10 Advanced~ Virg~ VA      For ~ 2 Year      NA   13680   13680   13680   13680
## # ... with 2,963 more rows, 2 more variables: index <int>, school_count <int>,
## # and abbreviated variable names 1: state_code, 2: degree_length,
## # 3: room_and_board, 4: in_state_tuition, 5: in_state_total,
## # 6: out_of_state_tuition, 7: out_of_state_total

ggplot(tc_with_count2, aes(x=state_code, y=school_count, fill=degree_length))+
  geom_bar(stat = "identity",
    position="dodge")
```

