Final Project: Is College worthy? (Tuition VS Salary)

Junjie Yang

5/2/2020

Import first data:

First data source, tuition_cost, is from "College tuition, Diversity, and Pay" in rfordatascience/tidetuesday/2020-03-10, which is originally acquired from the US Department of Education and the Chronicle of Higher Education.

Data Cleaning for tuition_cost data

In the tuition_cost data, relevant columns are selected (name of the school, state, state code, type of the school, length of the degree). Also, room and board fee and tuition are combined as total tuition and fee.

Snippet of tuition_cost data

```
head(tuition_cost,10)
```

```
## # A tibble: 10 x 7
##    name   state state_code type  degree_length in_state_tuitio~ out_of_state_tu~
##    <chr>  <chr> <chr>      <chr> <chr>                    <dbl>            <dbl>
##  1 Aanii~ Mont~ MT         Publ~ 2 Year                    2380             2380
##  2 Abile~ Texas TX         Priv~ 4 Year                   45200            45200
##  3 Abrah~ Geor~ GA         Publ~ 2 Year                   12602            21024
##  4 Acade~ Minn~ MN         For ~ 2 Year                   17661            17661
##  5 Acade~ Cali~ CA         For ~ 4 Year                   44458            44458
##  6 Adams~ Colo~ CO         Publ~ 4 Year                   18222            29238
##  7 Adelp~ New ~ NY         Priv~ 4 Year                   54690            54690
##  8 Adiro~ New ~ NY         Publ~ 2 Year                   17035            21595
##  9 Adria~ Mich~ MI         Priv~ 4 Year                   48405            48405
## 10 Advan~ Virg~ VA         For ~ 2 Year                   13680            13680
```

Import second data:

Second data source, student_diversity by college/university, along with school type, degree length, state, in-state vs out-of-state is from the Chronicle of Higher Education.

Data Cleaning for student_diversity data

In the student_diversity data, the main data cleansing task is to modify name of institution to match the "name" column and "state" column in the tuition_cost data, in order to combine dataset. Several data wrangling steps were applied. First is to change the column name "INSTITUTION" to "name". After that, convert any abbreviation of University from "U." to "University". From the first glance, the name of state is located at the very end of the name of institution. The next step is to extract state from school name with the help of state.name which contains the list of all the state name and column "state" is created. Last but

not least, state name inside the name of institution needed to remove. Using str_count to count the letters within state in each observation and str_sub help to keep the name of school only in the “name” column. Str_trim and str_squish are used to remove unnecessary spaces in “name”.

Combine tuition_cost and student_diversity data based on “name” and “state”

So far, student_diversity and tuition_cost are modified to share two common column, “name” – name of the school and “state” – the state that the school is located. Thus, student_diversity and tuition_cost datasets are merged for late development. There are a few schools appears in the tuition_cost dataset but not in the student_diversity and “NA” value appear. It is reasonable and schools with “NA” value are removed from the combined dataset. The combined dataset is arranged by state and the name of the school.

Import third data: Best_School

Third data source, best_school is html data, acquired from the from the payscale.com. It contains all the schools in United States that are arranged by various measurement of career performance, such as “Early Career Pay” and "Mid Career Pay.

Data Cleaning for data