Nathan Beebe

Professor Holtman

STAT 201 – 940

September 5th, 2020

Intro to Statistics Final Project

Summary

For the project, Kaggle was a great resource to find some data regarding college tuition and salary out of college. The goal was to understand the relation between STEM major percentage at a university and salary out of college in early and mid-career.

The dataset I found was from a website called tuitiontracker.org, on Kaggle some users were looking to practice their data science skills on a set of data that was on that website. I decided that I wanted to do some analysis on some of the data. The CSV files that I worked with had two different sets of data. The salary potential CSV features 936 observations of 7 variables of salary potential data from United States universities. The tuition cost CSV has 2974 observations of 10 variables of tuition cost data. A random sample of 10 observations was taken from each file.

I decided to use Python 3.8 with some data science libraries such as Pandas, Numpy, and Matplotlib. I have worked with Python before, and the project gave me an opportunity to get used to these popular data science libraries that are used by a lot of people. It was challenging to get used to the syntax for some of the functions, but I managed to get the hang of it by looking at some examples.

Some of the visualizations are hard to see in the document, but they are also included as separate PNG files in the submission.

Descriptive Statistics

Tuition Description

```
Tuition description:

       in_state_tuition  out_of_state_tuition
count          10.000000             10.000000
mean        23068.400000          24472.000000
std         17202.438653          15884.435226
min          2356.000000           8116.000000
25%         11350.750000          12151.000000
50%         16820.000000          19288.000000
75%         32297.750000          32297.750000
max         56382.000000          56382.000000
```

Salary Potential Description

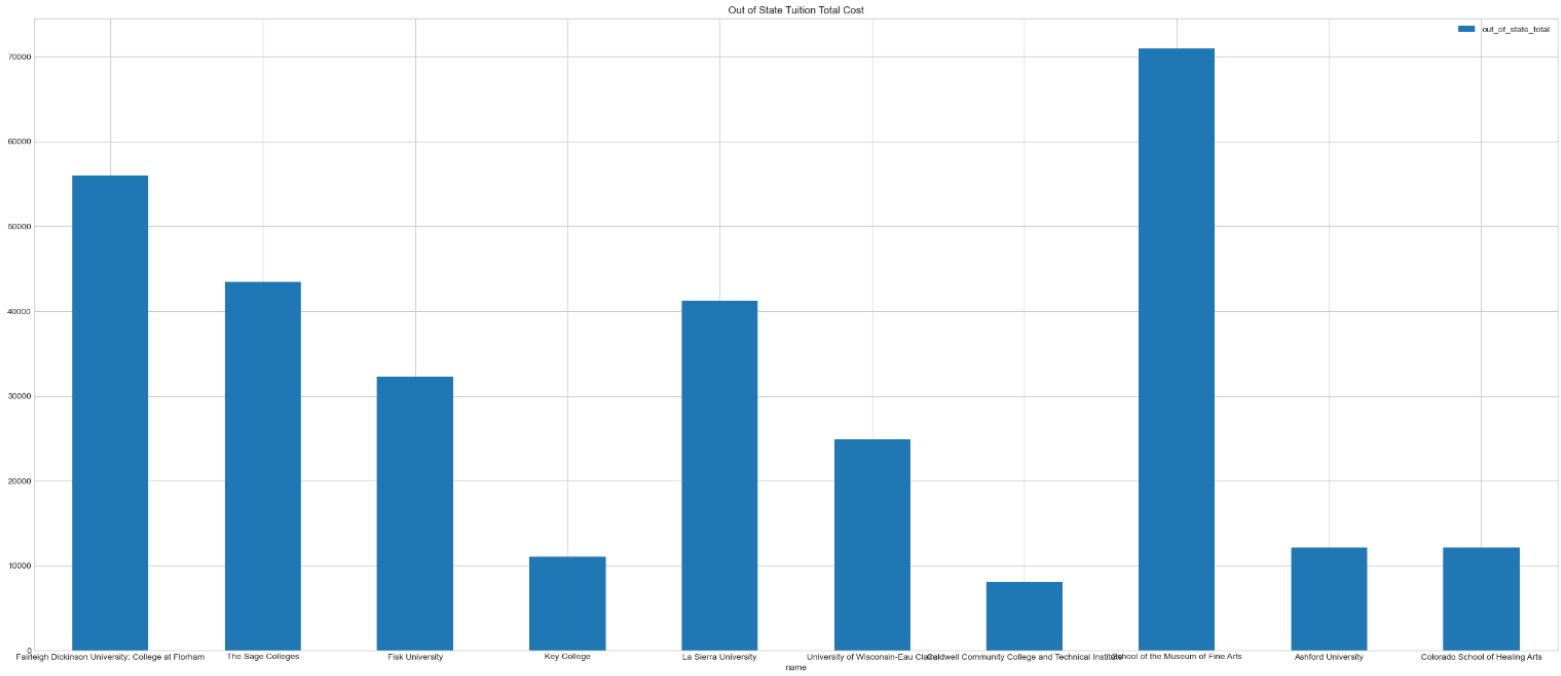
```
Salary description:

       early_career_pay  mid_career_pay
count          10.000000       10.000000
mean        52950.000000    96240.000000
std          9389.030242    18853.305045
min         42300.000000    74500.000000
25%         46500.000000    83025.000000
50%         51100.000000    92800.000000
75%         56650.000000   103200.000000
max         75200.000000   139400.000000
```

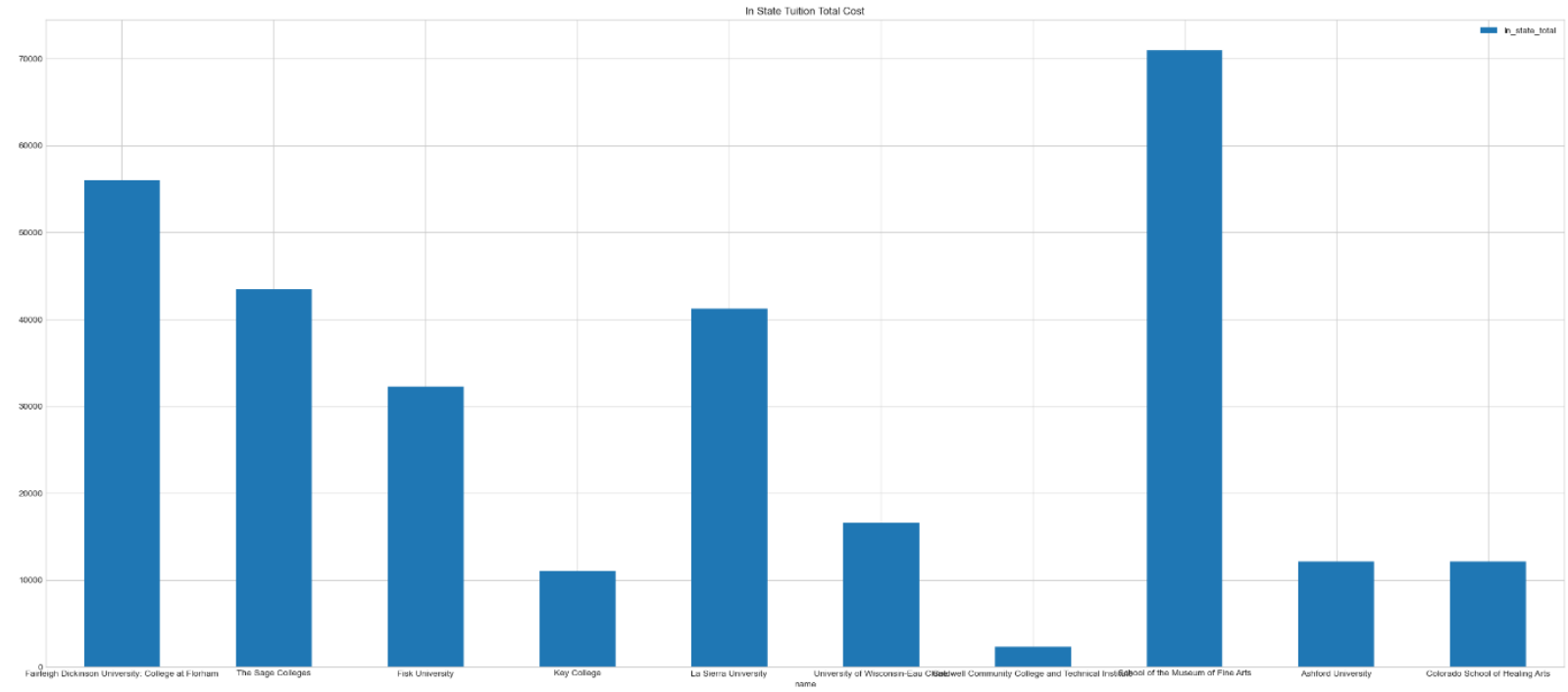
Data Visualization

Included all visualizations as PNG images with my submission, so they can be viewed at higher resolution.

Out of State Tuition Bar Graph



In State Tuition Bar Graph



STEM Percentage Versus Early Salary Amount Scatter Plot