**Project Report**

1. **Title and authors.**
   1. Title: Correlations Between Student Earnings and College Features in US – CSE 163 Final Project
   2. Author: Tianyi/Flora Tang, Yinxuan/Owen Wang
2. **Summary of research questions and results.**
   1. What are the yearly changes of median earnings of former students (working and not enrolled 10 years after entry) based on different college control types (public, private nonprofit, private for-profit)?
      1. Result:

图片包含 屏幕截图

描述已自动生成

*produced by main.py*

* 1. What are the median earnings of former college students (working and not enrolled 10 years after entry) in different states?
     1. Result:

图片包含 文字, 地图

描述已自动生成

*produced by main.py*

* 1. What is the relationship between colleges’ median SAT/ACT score percentages (median credit / full credit) and median earnings of former students (working and not enrolled 10 years after entry)?
     1. Result:

图片包含 屏幕截图

描述已自动生成

*produced by main.py*

1. **Motivation and background.**

Since the median earnings of former students are considered as an important indicator of colleges’ competency, we are trying to connect it with various college features to help perspective students choose colleges.

* 1. The first question would help perspective students choose colleges based on control types that correspond to high median earnings of former students.
  2. The second question would help perspective students choose collages based on states that correspond to high median earnings of former students.
  3. The third question would help perspective students choose colleges based on median SAT/ACT score percentages that correspond to high median earnings of former students.

1. **Dataset.**
   1. Description: our code will download and unpack a US college dataset and its documentation file, as well as a US geospatial dataset. All useful files will be cached for future use.
      1. The college dataset includes data from different academic years in separate files, and we will merge them into a single file. We will also delete data we don’t need and convert missing/privacy-suppressed data to NaN.
      2. We will perform no further preprocessing for the geospatial dataset.
   2. URL:
      1. College Dataset:
         1. Homepage: <https://collegescorecard.ed.gov/data/>
         2. Dataset Download:

<https://ed-public-download.app.cloud.gov/downloads/CollegeScorecard_Raw_Data.zip>

* + - 1. Documentation Download: <https://collegescorecard.ed.gov/assets/CollegeScorecardDataDictionary.xlsx>
    1. Geospatial Data:
       1. Homepage:

<https://www.naturalearthdata.com/downloads/50m-cultural-vectors/50m-admin-0-countries-2/>

* + - 1. Dataset Download: <https://www.naturalearthdata.com/http//www.naturalearthdata.com/download/110m/cultural/ne_110m_admin_1_states_provinces.zip>

1. **Methodology.**
   1. Question 1: Use a titled line plot to visualize the yearly changes of the average median earnings of former students (working and not enrolled 10 years after entry) per college control type (public, private nonprofit, private for-profit). The plot should be colored by control types, with a legend indicating the meaning of colors.

The result would be based on the graph. Our goal is to find college control types that correspond to high median earnings of former students.

* 1. Question 2: Merge the college dataset with the geospatial dataset. Plot a titled map of United States colored by the average median earnings of former students (working and not enrolled 10 years after entry) per year per state, with a legend indicating the meaning of colors. For aesthetic and readability purposes, mainland, Alaska, and Hawaii would be plotted as separate subplots in one figure.

The result would be based on the graph. Our goal is to find states that correspond to high median earnings of former students.

* 1. Question 3: Plot two quadratic regression lines in the same titled plot for all colleges’ average median SAT/ACT score percentages per year with respect to their average median earnings of former students (working and not enrolled 10 years after entry) per year. The plot should be colored by test types, with a legend indicating the meaning of colors. To calculate the median SAT score percentage for a college in one year, add up median SAT reading, math, and writing scores and divide by 2400 (the full score of SAT before January 2016). To calculate the median ACT score percentage for a college in one year, divide ACT cumulative score by 36 (the full score of ACT).

The result would be based on the graph. Our goal is to find the median SAT/ACT score percentages that correspond to high median earnings of former students.

1. **Results.**
   1. Question 1:

图片包含 屏幕截图

描述已自动生成

*produced by main.py*

Control types ranking from highest median earnings of former students to lowest: private nonprofit, public, private for-profit.

From 2007-2015, median earnings of former private nonprofit college students are relatively constant around 42500 dollars, median earnings of former public college students are relatively constant around 35000 dollars, while median earnings of former private for-profit college students have decreased from around 30000 dollars to around 27500 dollars.

* 1. Question 2:

图片包含 文字, 地图

描述已自动生成

*produced by main.py*

States with highest median earnings of former college students: District of Columbia, Massachusetts, New York, Connecticut, Maryland.

The median earnings of former college students in Washington State are lower than expected, despite that several huge companies are in the zone. This may be a result of the lower overall economy of west states than that of east-north states.

* 1. Question 3:

图片包含 屏幕截图

描述已自动生成

*produced by main.py*

From 0.0 to about 0.3 college score percentage, median earnings of former students decrease as score percentage increases. From about 0.3 to 1.0 college score percentage, median earnings of former students increase as score percentage increases. The highest median earnings correspond to highest score percentages.

The result is partly different from our prediction. We thought higher score percentage leads to higher earnings, but some colleges with very low score percentages surprisingly have higher student median earnings than those of colleges with slightly higher score percentages. One explanation could be that, test score dominates student earnings, but only to some degrees. At the lower end of score percentage, other abilities in students like leadership and creativity dominate their earnings.

1. **Reproducing results.**
   1. Activate cse163 environment
   2. Install xlrd and GeoPandas 0.5.0 (pip install xlrd geopandas==0.5.0)
   3. Execute "main.py"
   4. Results are under "results" folder in the current working directory
2. **Work plan evaluation.**
   1. Data Preprocessing (details indicated in “Dataset” part) – 13h
      1. Underestimated (17h). Downloading and unpacking files was harder than expected. Also, the college dataset has undescriptive column names so we have to combine it with its documentation.
   2. Question 1 (details indicated in “Methodology” part) – 2h
      1. Accurate. Only includes normal Pandas operations.
   3. Question 2 (details indicated in “Methodology” part) – 3h
      1. Underestimated (7h). Plotting Alaska, Hawaii, and mainland separately was harder than expected. We also included the functionality of highlighting an outlier.
   4. Question 3 (details indicated in “Methodology” part) – 2h
      1. Accurate. Only includes normal Pandas operations.
3. **Testing.** We test our code using 3 small college datasets generated by us, along with the original college documentation and geo dataset. assert\_equals is used to compare our hand calculations to objects returned by processs\_data\_\* functions in main.py. (To support our test, we modified check\_approx\_equals in hw7 cse163\_utils.py to ignore element order when comparing two instances of numpy.ndarray).
4. **Live Presentation or Video?** We’ll give a 2-minute presentation about our project on Tuesday June 5, 2018 2:30-4:20pm.
5. **Collaboration.** Nobody helped us in our project.