POPULAR HIGHER EDUCATION TYPES IN THE EAST AND THE WEST OF THE U.S.

Phuc (Peter) Luong

Section 1: Introduction

International students and out-of-state domestic students often would like to have a thorough understanding of the U.S. university system at a local level before enrolling. Besides ranks and popularity, specific majors and academic types play important roles in the students' decisions to attend college.

In addition to students, policy makers and educators are also interested in learning more about the academic distribution of a region before establishing new institutions. This knowledge helps not only compete with other universities and colleges, but also to have an overall understanding of the general demand of the local areas and nearby places.

However, one often wonders if there is a difference between different U.S. regions regarding specific education subjects. In other words, the paper examines whether there is a significant difference in academic type distribution of colleges and universities between the eastern and western regions of the U.S. "Are we more likely to see Technology-focused majors in the West coast relative to the East coast?" is an example of one of the interested questions.

For this project, I am going to utilize Foursquare API to find location data of a random selection of higher education departments throughout the East and West of the U.S. By using Foursquare API, I am also able to find location of university departments and their respective categories. The choices of region are completely arbitrary. Furthermore, five cities are drawn randomly for each region and Foursquare API will help obtain information about the academic type distribution of nearby cities and towns that house the corresponding universities. Later on, I will test formally, using Chi-square test for independence and multinomial logistic regression, to understand whether demographic characteristics of a region can predict the categorization of the higher education subjects of that region.

Section 2: Data Collection and Features

2.1. Data Sources and Collection Methods

As mentioned before, the main data source for this analysis comes from Foursquare API (for developers), a global database containing rich information of venues from many regions across the globes. The venues are also sorted to more than 900 categories, which are categorized mainly by the Foursquare consumer community. This aspect is very important in my analysis since all the assumptions about the population distribution (classes/categories of all universities in the U.S.) are based on the accuracy of venues' categorical classification. For instance, I will use the labelling for a university's department based solely on the data given by API Foursquare search. Using Foursquare API, I limited my search to the seven subcategories of 'College Academic Building' section of Foursquare. I performed ten different searches using five personal favorite cities from the West and the other five from the East of U.S. Each search has a radius perimeter of around a little less than 100 kilometers. Since the limit for each search only returns at most 100 results, I performed 'pagination' method where I added another variable in the API search function to display more results.

The raw data table (the main data for this paper) combines all the search results above. Since the search radius is large, overlapping results occur. However, I take into account this issue and the final data set is coded so that there are no duplicates. For the locations and demographic data of unincorporated cities and towns, I scraped data and manually fill some data points in an Excel table containing relevant information. In addition to Foursquare API, I use sources from the following websites: Wikipedia, Google Maps and https://simplemaps.com/. The data tables can also be accessed through my GitHub account.

2.2. Data Description and Summary

The first data set, a raw dataset from Foursquare search results, concatenates different results obtained from Foursquare API data search. Relevant variables include name of the institution's department, coordinates (latitude and longitude), city, state, region and department type, which is the main dependent variable for this entire analysis. For the rest of this report, I will refer to this dataset as the main data. Later on, I will also merge the main data with data from simplemaps.com, a city/town-level data that includes basic demographic information about the regions such as population, density and coordinates (cities and town level). This will be covered more in the later regression analysis.

For the main dataset, there are 1072 observations for the entire sample. Two interested regions are East and West. The number of unique cities/towns of the two regions is 257. The number of unique academic categories is 7: Arts, Communications, Engineering, History, Math, Science and Technology. Their respective frequencies are as follows:

Table 1. Average Frequency by Academic Type

	Arts	Communications	Engineering	History	Math	Science	Technology
Frequency	0.322	0.072	0.111	0.0121	0.0335	0.326	0.122

The following cross-tabulation table (or contingency table) displays the number of observations based on interested categories and the total of data records in the sample.

Table 2. Cross-Tabulation Table

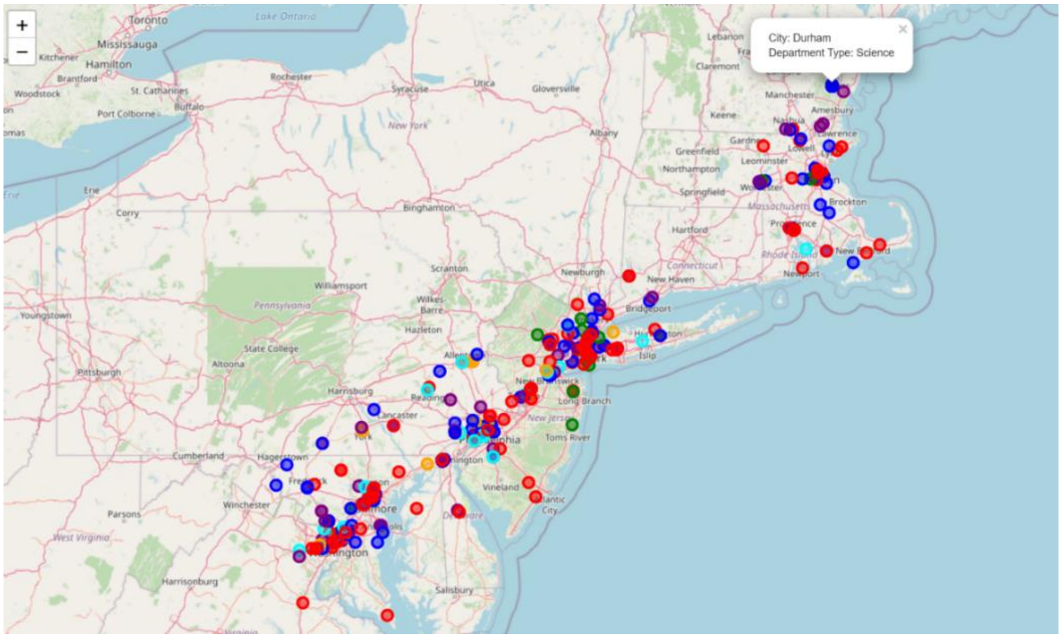
Region	Arts	Communications	Engineering	History	Math	Science	Technology	Total
East	179	40	54	6	20	174	61	534
West	166	37	66	7	16	176	70	538
Total	345	77	120	13	36	350	131	1072

2.3. Data Visualization

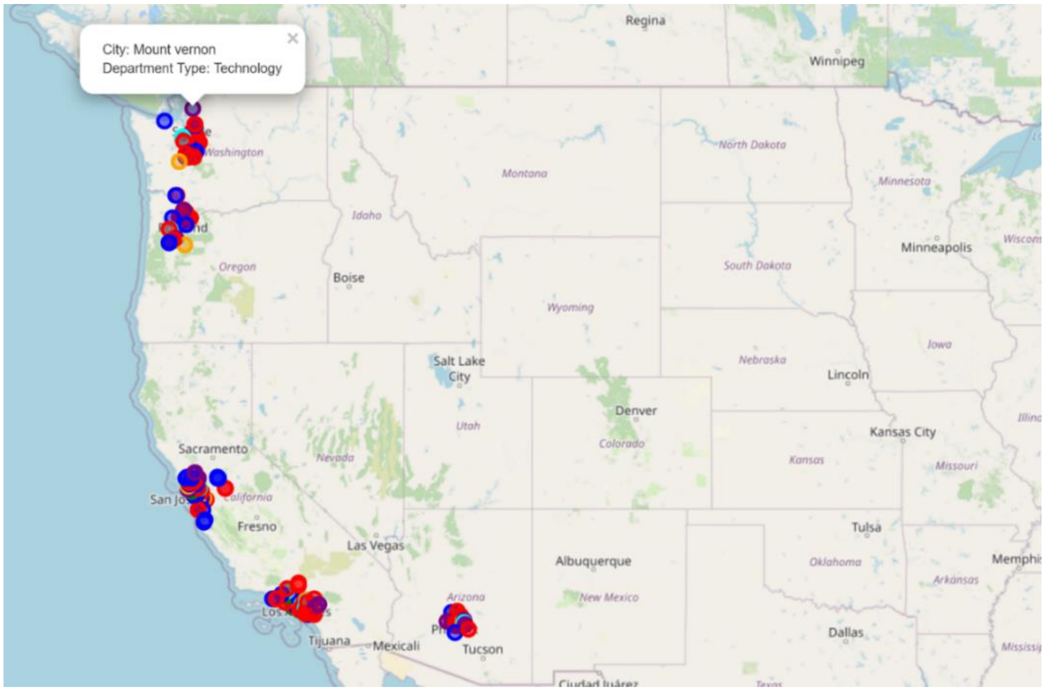
Let's first examine the geospatial visualization of the main dataset to understand where the sample comes from. As noted before, the data is divided into two interested regions: the western and eastern U.S. regions. Please also note that the decision to categorize data based on regions depends entirely on the map visualization of the data point as there is no official document to decide whether a city or town is in the East or the West of the U.S. Each data point is an institution's department and is color-coded according to an academic category, which is one of the seven types categorized by Foursquare. By examining the following maps, there is no systematic difference in academic type distribution between the two regions. All data points and their respective colors are scattered randomly.

Figure 1. Map of Sample Data Points

Eastern U.S.



Western U.S.



Note that the zoom level is different

Figure 2.

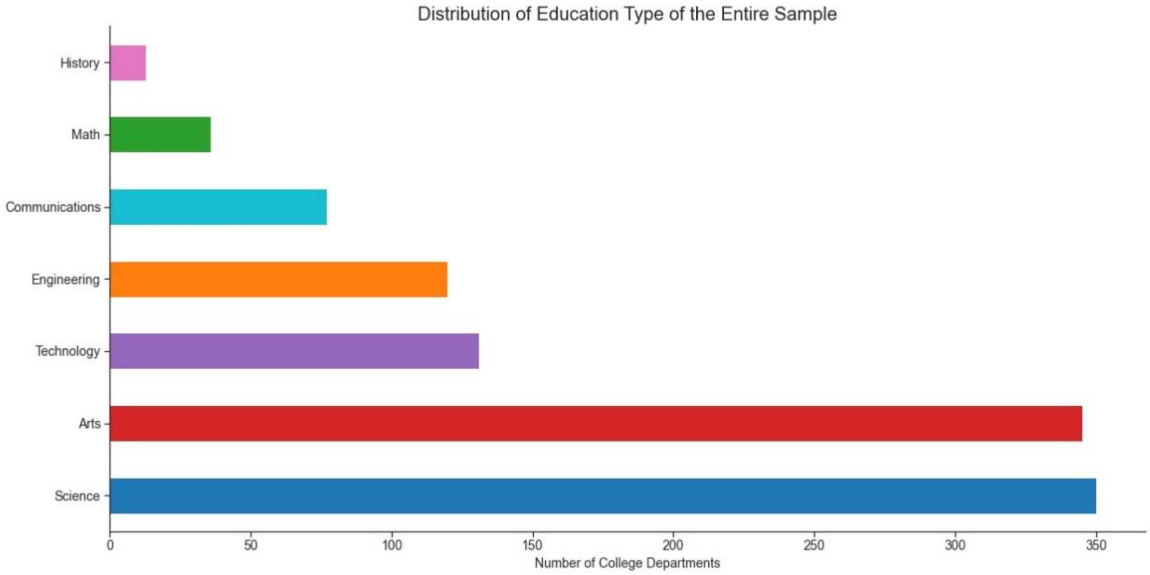
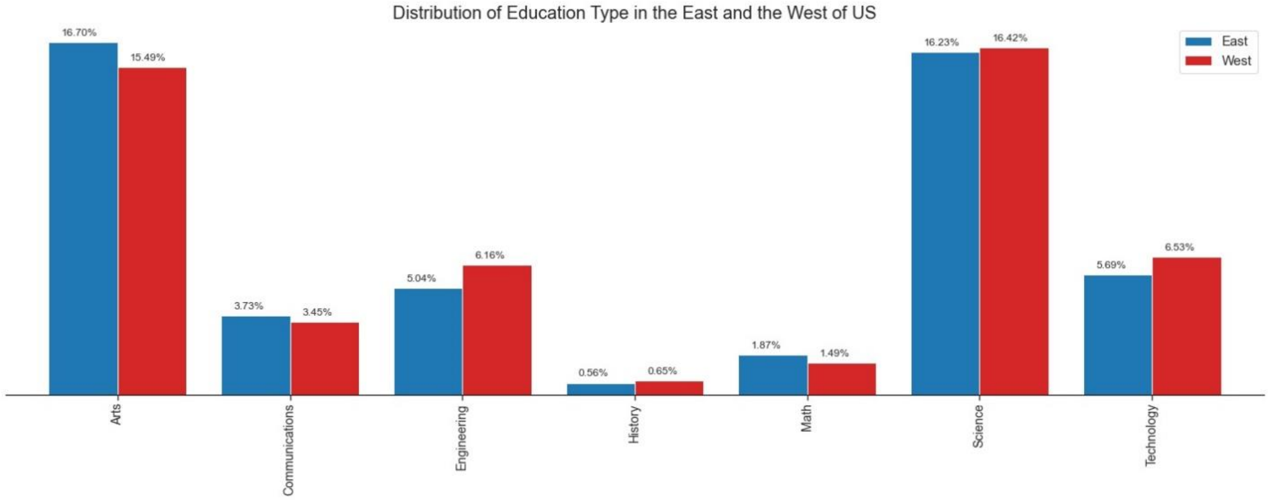


Figure 3.



Figures 2 and 3 visualize the frequency of education type in the sample. Figure 2 shows the distribution of education type of the entire sample. Each bar represents the number of observations found in the sample, corresponding to the type. Arts, Science and Technology are among the most popular academic categories, accounting for 32.2%, 32.6% and 12.2% of the total sample respectively. History has the lowest distribution, accounting only 1.2% of the total sample.

Alternatively, Figure 3 shows the distribution of academic categories between the eastern and western regions of the U.S. Each bar represents the frequency of the corresponding academic type relative to the total sample size of 1072. Arts and Science are still two of the most popular academic categories for two regions. History has the lowest frequency in both regions. The regional trend in academic type frequency is very similar to that of the whole sample. By simply eyeballing the data, there is also a similarity between the academic type distribution between the two regions. For later parts of this report, I will test if there is truly an association between the two regions' academic types.

Section 3: Research Methodology and Results

3.1. Objectives

In addition to analyzing the main features of data through data visualization methods such as bar charts to identify trends of academic type distribution between the two U.S. regions, this section will test formally whether such relationship exists. The data is split into two regions, which is convenient to perform Chi-square test for independence. More precisely, I will test whether there is an association in academic categories distribution between the eastern and western regions of the U.S. Finally, using a modified version of the original data set, I will perform a basic multinomial logistic regression to test whether one can predict academic type of a region (at city/town level) by using data on the region's features such as population, area size and location.

The selection of cities is random regarding the main objective of this research. In particular, the choices are not in any way related to the distribution of higher education type. For example, I did not choose New York because of the its academic type distribution, but simply because it is one of my favorite cities. Therefore, the selections will not likely bias the final results.

3.2. Chi-square test for Independence

The data set obtained from Foursquare API fits into two main categorical variables. The two regions are East and West and type of education is one of the seven categories. From the previous sections, I have also manipulated the main data into an appropriate contingency table displaying the frequency distribution of the interested variables, which is shown in Table 2.

Before diving deep into city-level observations, this section first tests whether there is a difference in the distribution of higher education type across two western and eastern regions of the U.S. Therefore, it is appropriate to conduct the Chi-square test for independence with the main categorical variables of interest as region and higher education type.

Before proceeding with the Chi-square test for independence, let's ensure that all of the following three conditions are satisfied:

- The sampling method follows simple random sampling.
- The studied variables are all categorical.
- For the sample data displayed in the contingency table, the expected frequency count cannot be less than 5.

Next, we check if our data sample truly satisfies the assumptions before continuing with the Chi-square test.

- As stated in the motivation section, I chose these cities randomly. There is no reason to believe that the sampling process is not random. More precisely, the selection of cities is arbitrary and does not concern with the specific academic type distribution of that region. Therefore, my sampling method of choosing college buildings/departments is random and will unlikely bias the analysis results.
- It is natural that our data has a set of categorical variables: regions and education type
- From the contingency table, obtained by manipulating the original cross-tabular table (***Table 3***) into relative frequency distribution, it seems that all the expected frequency cells are larger than 5.

Table 3. Expected Frequency

Region	Arts	Communications	Engineering	History	Math	Science	Technology	Total
East	171.856	38.3563	59.776	6.475	17.932	174.347	65.255	534
West	173.143	38.6436	60.223	6.524	18.067	175.653	65.744	538
Total	345	77	120	13	36	350	131	1072

Therefore, with all the assumptions satisfied, we can continue with the Chi-square test for independence between the two U.S. regions.

Hypothesis

The main hypothesis I am interested in testing is to study whether academic type distribution of the universities from the two regions, East and West of U.S., is independent. Precisely, the chi-square test of independence will tell us if we can understand the frequency of education type of one region knowing the academic distribution of the other region. The hypotheses are as follows:

- $\boldsymbol{H_o}$: There is an association of education type distribution between the East and West regions
- $\boldsymbol{H_a}$: There is no association of education type distribution between the East and West regions

Analysis plan:

The degree of freedoms has the following formula: DF = (r - 1) * (c - 1), where:

- r is the number of categories of the first variable (region)
- c is the number of categories of the second variable (education type)

so in our sample: DF = (2 - 1)*(7 - 1) = 6
The expected frequency cell has the following formula: $E_{r,c} = \frac{n_r * n_c}{n}$ and the formula for Chi-square statistics is: $X^2 = \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}}$, where *r* and *c* are row and column of the contingency/expected frequency table. All values of $O_{r,c}$ and $E_{r,c}$ are shown in Table 2 and Table 3 respectively.

Decision rule:

We will reject the null hypothesis H_o if the calculated test statistic is larger the critical value based on the degrees of freedom and level. Alternatively, we can also reject the null hypothesis if the p-value obtained from the test statistics is sufficiently small relative to the chosen significance level.

Chi-square Test Results:

Table 4. Chi-square test results

Chi-square Test	Results
Pearson Chi-square (6.0)	2.9430
p-value	0.8160
Cramer’s V	0.0524

Using the Chi-square package ‘researchpy.crosstab’ from Python, the Chi-square statistics is 2.943 and p-value is 0.82. Therefore, for all significance levels: 0.01, 0.05 and 0.1, we do not have sufficient evidence to reject the null hypothesis in favor of the alternative. In other words, there is enough evidence to suggest that the distribution of higher education type is independent between the East and the West regions of the United States.

3.4. Multinomial Logistic Regression

For this section, I am interested in building a simple model to predict academic frequency of a region using its demographic data. This is a classification problem which extends the basic logistic regression to multiclass problems with the dependent variable having more than two discrete outcomes. Instead of using the main data, I modify the original version by reshaping it into a new city-level data which contains the city’s most popular academic type as a multi-class dependent variable and demographic features as independent variables.

One of the first assumptions of multinomial logistic regression is the unique case for each independent variable. Because of non-unique mapping in cities and academic type of the original data in our original data (a city has multiple universities’ departments and hence multiple academic type), I first created a rank variable which displays an academic type that has the highest frequency. If there is a tie, sorting and selection will be random between ties. After that, I merge the “rank” data with their respective cities’ demographic data, including population (in persons), area size (in squared kilometers) and west (a dummy which is 1 if the city is in the west).

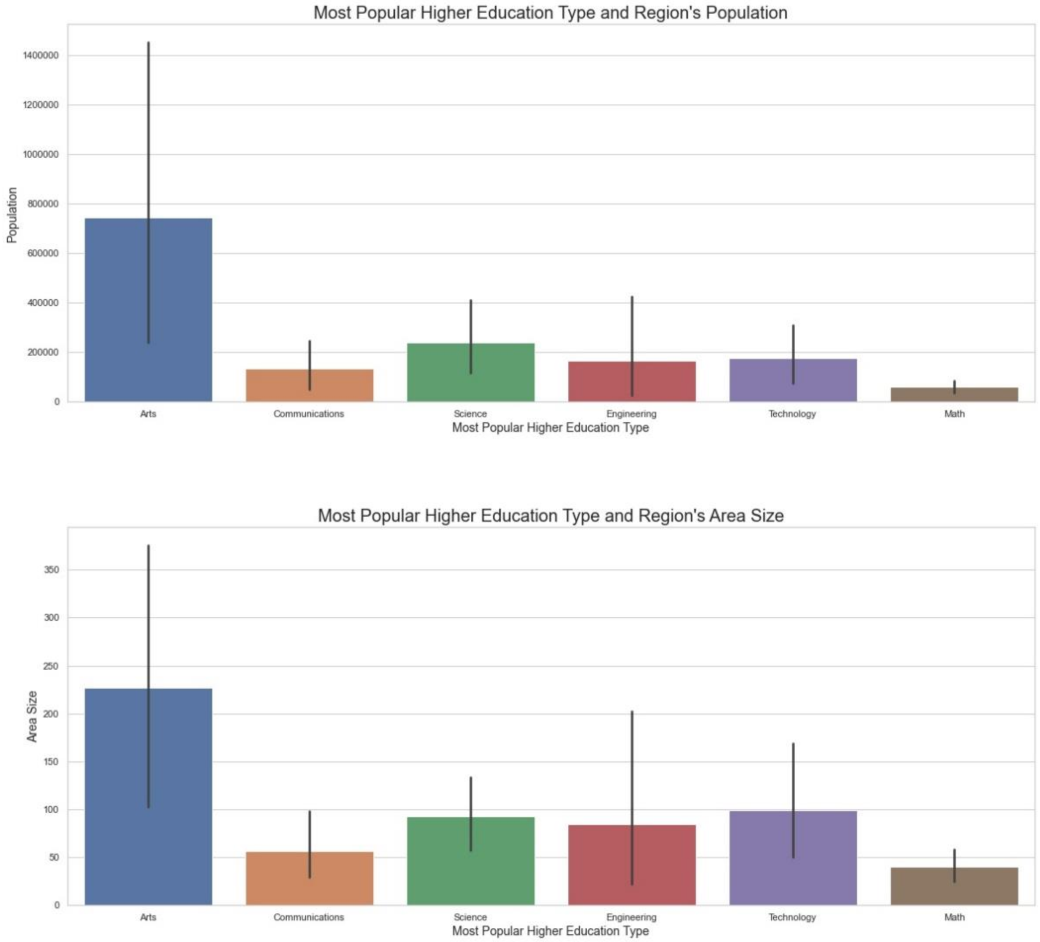
The other important assumption of this model is non-multicollinearity between independent variables. In other words, one independent variable in the model cannot be predicted by the set of other independent variables. I used the Variance Inflation Factor (VIF) method to check linearly dependence of independent variables in the regression model. As a rule of thumb, a VIF value between 1 and 5 is considered “safe.” All the correlation coefficients of population, area size and west are relatively low since all of their respective VIFs are less than 3. Therefore, it is appropriate to proceed with the multinomial logistic regression on our data sample.

Table 5. Variance Inflation Factor (VIF)

VIF Factor	features
1.024518	west
2.420938	population
2.445618	size
1.826251	const

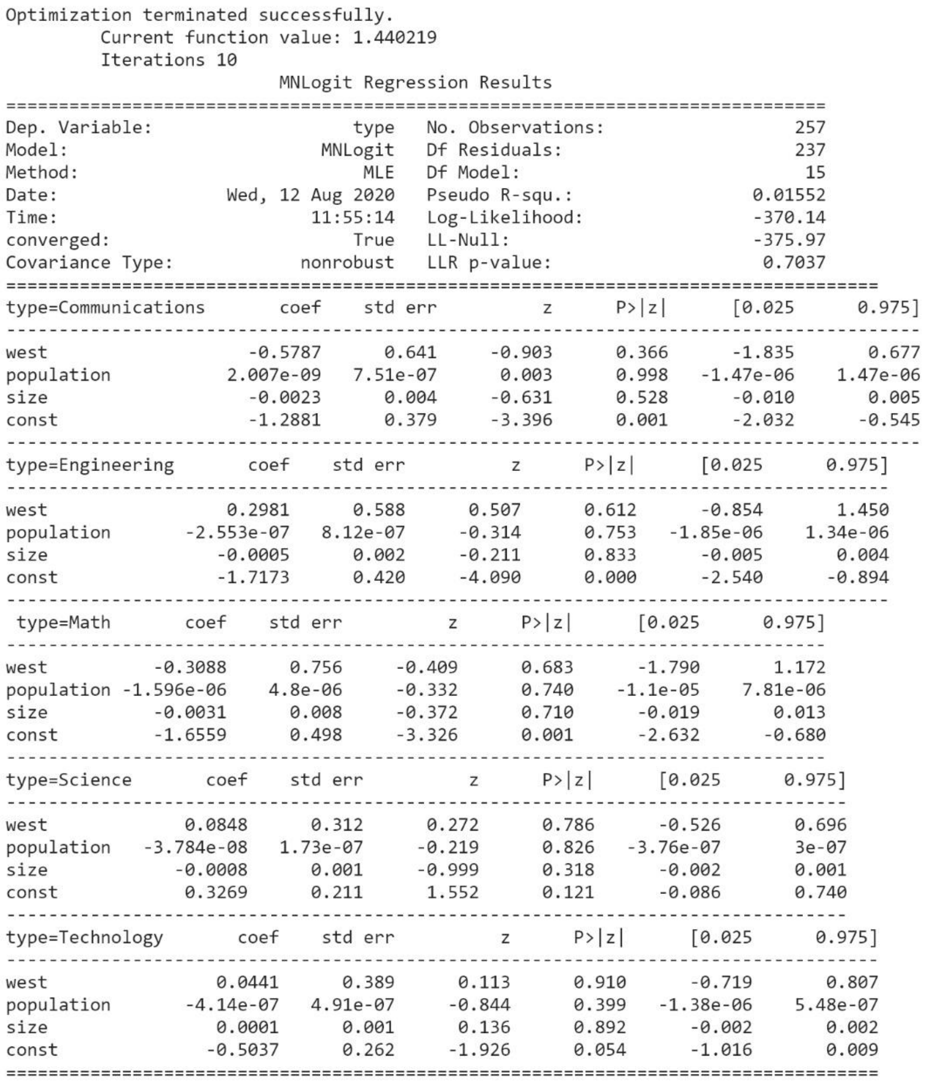
The following figures show the relationship between a city or town's most popular academic type and its population and area size.

Figure 4. City/Town's demographic features and Academic Type Frequency



Since History type does not have many data points in the main dataset and is the least popular type from previous analysis, new variable "rank" does not have History as one of its elements. There is a trend in both population and area size and academic categories. Arts seems to be influenced by places that have higher population and larger area sizes.

As the sample size is small, only containing 257 unique cities, I will only perform basic multinomial logistic regression package using MNLogit from statsmodels package. The following table provides the result of the multinomial logistic regression.



```
Optimization terminated successfully.
         Current function value: 1.440219
         Iterations 10
                          MNLogit Regression Results
==============================================================================
Dep. Variable:                   type   No. Observations:                  257
Model:                        MNLogit   Df Residuals:                      237
Method:                           MLE   Df Model:                           15
Date:                Wed, 12 Aug 2020   Pseudo R-squ.:                 0.01552
Time:                        11:55:14   Log-Likelihood:                -370.14
converged:                       True   LL-Null:                       -375.97
Covariance Type:            nonrobust   LLR p-value:                    0.7037
===================================================================================
type=Communications       coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
west                   -0.5787      0.641     -0.903      0.366      -1.835       0.677
population           2.007e-09   7.51e-07      0.003      0.998   -1.47e-06    1.47e-06
size                   -0.0023      0.004     -0.631      0.528      -0.010       0.005
const                  -1.2881      0.379     -3.396      0.001      -2.032      -0.545
---------------------------------------------------------------------------------------
type=Engineering       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
west                0.2981      0.588      0.507      0.612      -0.854       1.450
population      -2.553e-07   8.12e-07     -0.314      0.753   -1.85e-06    1.34e-06
size               -0.0005      0.002     -0.211      0.833      -0.005       0.004
const              -1.7173      0.420     -4.090      0.000      -2.540      -0.894
------------------------------------------------------------------------------------
 type=Math       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
west          -0.3088      0.756     -0.409      0.683      -1.790       1.172
population -1.596e-06     4.8e-06     -0.332      0.740     -1.1e-05    7.81e-06
size          -0.0031      0.008     -0.372      0.710      -0.019       0.013
const         -1.6559      0.498     -3.326      0.001      -2.632      -0.680
------------------------------------------------------------------------------
type=Science       coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
west            0.0848      0.312      0.272      0.786      -0.526       0.696
population  -3.784e-08   1.73e-07     -0.219      0.826   -3.76e-07       3e-07
size           -0.0008      0.001     -0.999      0.318      -0.002       0.001
const           0.3269      0.211      1.552      0.121      -0.086       0.740
--------------------------------------------------------------------------------
type=Technology       coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
west               0.0441      0.389      0.113      0.910      -0.719       0.807
population       -4.14e-07   4.91e-07     -0.844      0.399   -1.38e-06    5.48e-07
size               0.0001      0.001      0.136      0.892      -0.002       0.002
const             -0.5037      0.262     -1.926      0.054      -1.016       0.009
===================================================================================
```

The baseline for comparison is type Arts. For example, the fourth block of the table, where type is equal to Science, compares the log-likelihood ratio (or non-technically, the chance of seeing more) of Science relative to Arts. *west* indicates the difference of seeing two types between two regions (a multinomial logit of 0.0848 points): we are more likely to encounter Science type in the West relative to the East. Inversely, there is a negative trend for *population*. Holding other factors constant, for a one person increase in *population*, the chances of encountering Science type relative to Arts type decrease (by a very small point, about $3.78 * 10^{-8}$). Similarly, there is a negative trend in *size*. For one squared-kilometer increase in a region's size, we are less likely to see type Science relative to type Arts (a multinomial logit estimate of 0.0008 units).

For the remaining blocks of the result table, interpretations are similar using different education types against Arts. However, there is no apparent trends across different independent variables. More precisely, we cannot predict the frequency of academic type given the region's population, area size and location.

From the table, all of the results are statistically insignificant as all p-values are larger than any given significance level (either alpha of 0.1, 0.5 and 0.01). In addition, the R-squared is low, which is only 0.015. Therefore, we cannot conclude that there is a difference between education type distribution by basic demographic variables. In other words, the population, size and location of regions do not predict the academic type frequency of colleges and universities. The results are also consistent with the Chi-square test for independence from the previous section that there is no association between region and academic type distribution.

Section 4: Discussion and Limitations

The analysis is mainly for learning and practicing. If interested in interpreting the conclusion of this paper in a meaningful way, please first take into consideration of several limitations of the methodology. The limitation is mainly working with Foursquare API to fetch location data and geospatial analysis. The main downside of working with Foursquare API are the followings:

1. The search result is inconsistent

- The search algorithm of Foursquare API seems to change as time goes. For example, the results now are different from those a week ago. Due to time and resource constraints, I will leave the results as they were the first time I searched.

2. Some minor incorrect sub-categorizations of venues

- Even with detailed documentations such as particular '*categoryId*' for each venue, irrelevant results are usually mixed in. For example: I got Museum in my search for academic building. For the clarity of this analysis, I will simply drop such results.
- Do also note that Foursquare is entirely built by the community which means anybody can suggest edits or changes to the venues. Hence the labeling is not always correct.

3. Missing venues

- This may be an indirect consequence of the two reasons above: some venues are actually there without being documented. However, one good way to remedy this is to input several nearby location's latitudes and longitudes which can somewhat add more results (albeit with potentially heavy overlapping between results)

Considerations when interpreting multinomial logistic regression results

- Sample size is the key problem in interpreting the logistic regression results. In particular, the p-value is majorly influenced by sample size. The only clear solution is to gather more data. If the hypothesis that there is really a difference in education type distribution for different regions is true, then increasing sample size will decrease p-value and we can conclude there is indeed a difference.
- Missing important variables is problematic. By omitting other vital demographic factors influencing academic type distribution such as population age, income and gender, the multinomial logit estimates are likely biased. This is entirely due to missing data of many unincorporated towns/cities, or places that do not have a government, which do not have data on the aforementioned variables.

Section 5: Conclusion

In summary, I analyzed popular higher education academic type between the western and eastern U.S. using location data provided by Foursquare API and other external resources. Through methods of statistical inference and classification, I find no relationship between a region's demographic features such as population, area size and location on the academic type frequency of that region. There is also no association on the popularity of universities academic categories between the West and the East of U.S. I discussed some limitations of the data set and how one should interpret the results appropriately given the scope of this research.

Section 6: References and Resources

The following references have been very helpful to this report for providing resources and insights:

- IBM Applied Data Science Capstone lecture notes and discussion threads with many outstanding resources
- Foursquare API. For a list of available categories at Foursquare API, please visit: https://developer.foursquare.com/docs/build-with-foursquare/categories/
- Stack Overflow for technical suggestions
- statcompute.com for guidance on regression analysis for Python