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

PHUC (PETER) LUONG

Data Collection

1. The data comes from Foursquare (https://developer.foursquare.com/docs/places-api)
2. Five random cities are chosen.
3. Each Foursquare API search looks for nearby college departments
4. Each search function includes the following variables:
   - Radius: around 100 squared-kilometers
   - Venue categories: one of the seven “College Academic Building” subcategories from Foursquare
   - Pagination methods to extend more search results per search
5. Other data sources include Wikipedia, Google Maps and simplemaps.com

Data Summary Statistics

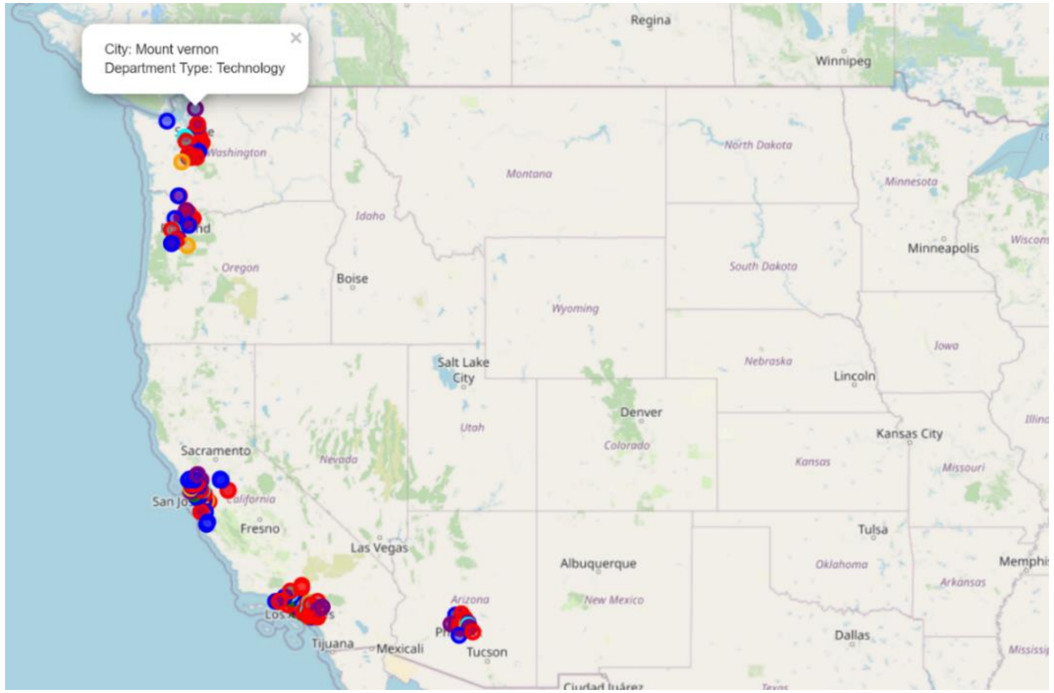
Table 1. Average Frequency by Academic Type

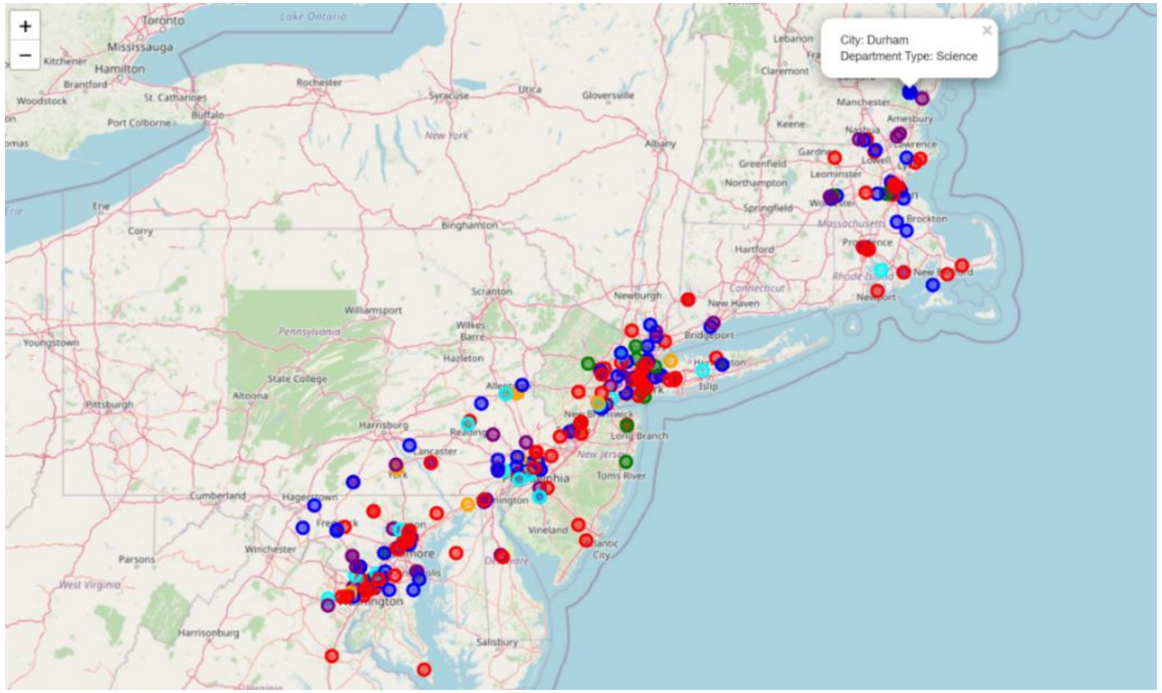
	Arts	Communications	Engineering	History	Math	Science	Technology
Frequency	0.322	0.072	0.111	0.0121	0.0335	0.326	0.122

Table 2. Cross-Tabulation Table (Contingency Table)

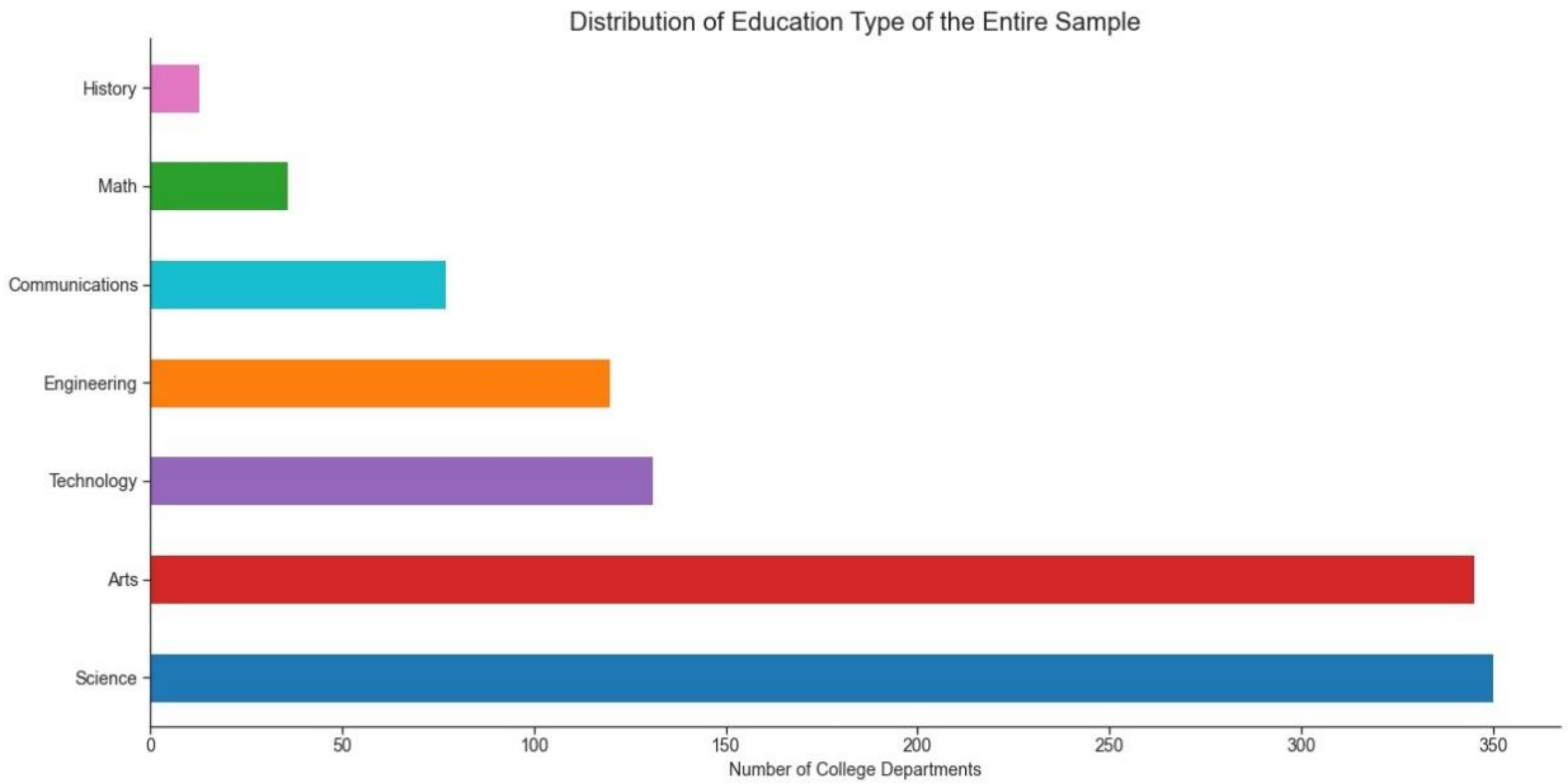
Region	Arts	Communications	Engineering	History	Math	Science	Technology	Total
East	179	40	54	6	20	174	61	534
West	166	37	66	7	16	176	70	538
Total	345	77	120	13	36	350	131	1072

Map Display of Data Points

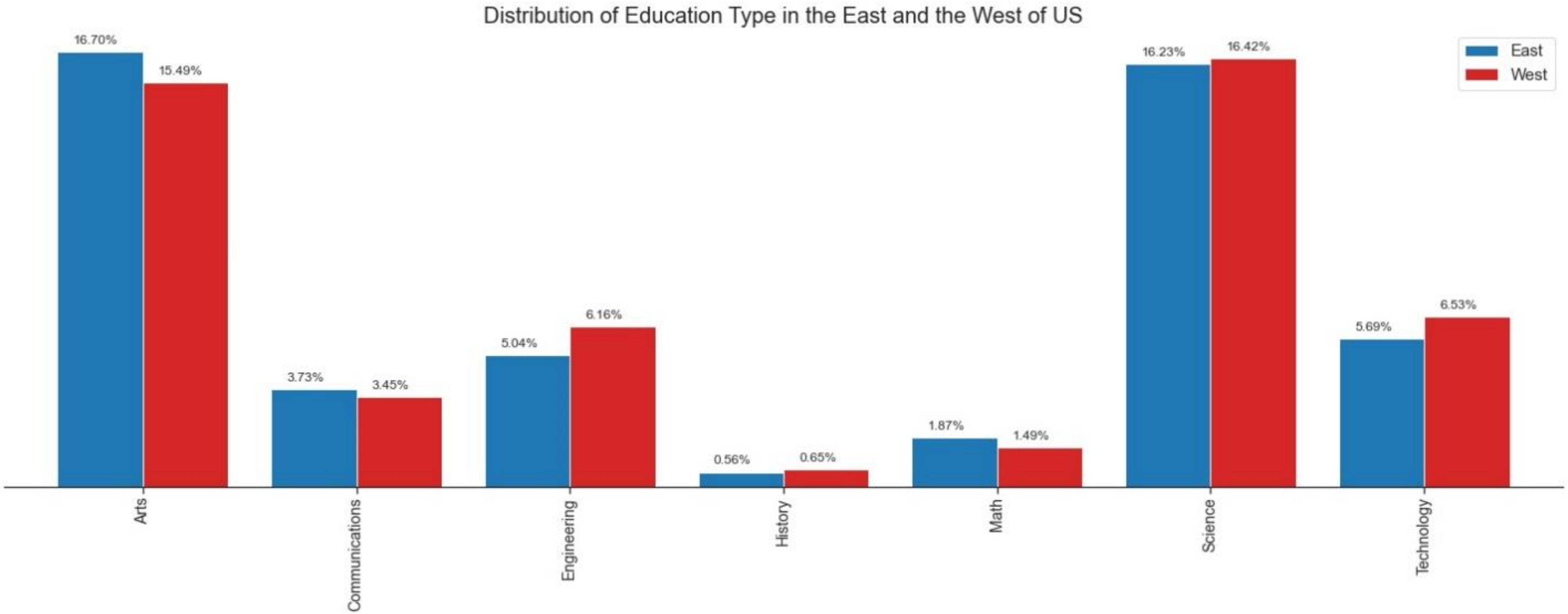




Frequency Distribution of Academic Type



Bar Chart of Frequency by Region



Chi-square Test for Independence

Table 3. Expected Frequency Table

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

Table 4. Results

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

Conclusion: not enough evidence to reject the null hypothesis: the distribution of higher education type is independent between the East and the West regions of the United States.

Multinomial Logistic Regression

Steps to perform the regression:

New data by reshaping original data it into a city-level data

Main question: can a place’s demographic characteristics predict academic type distribution of that place?

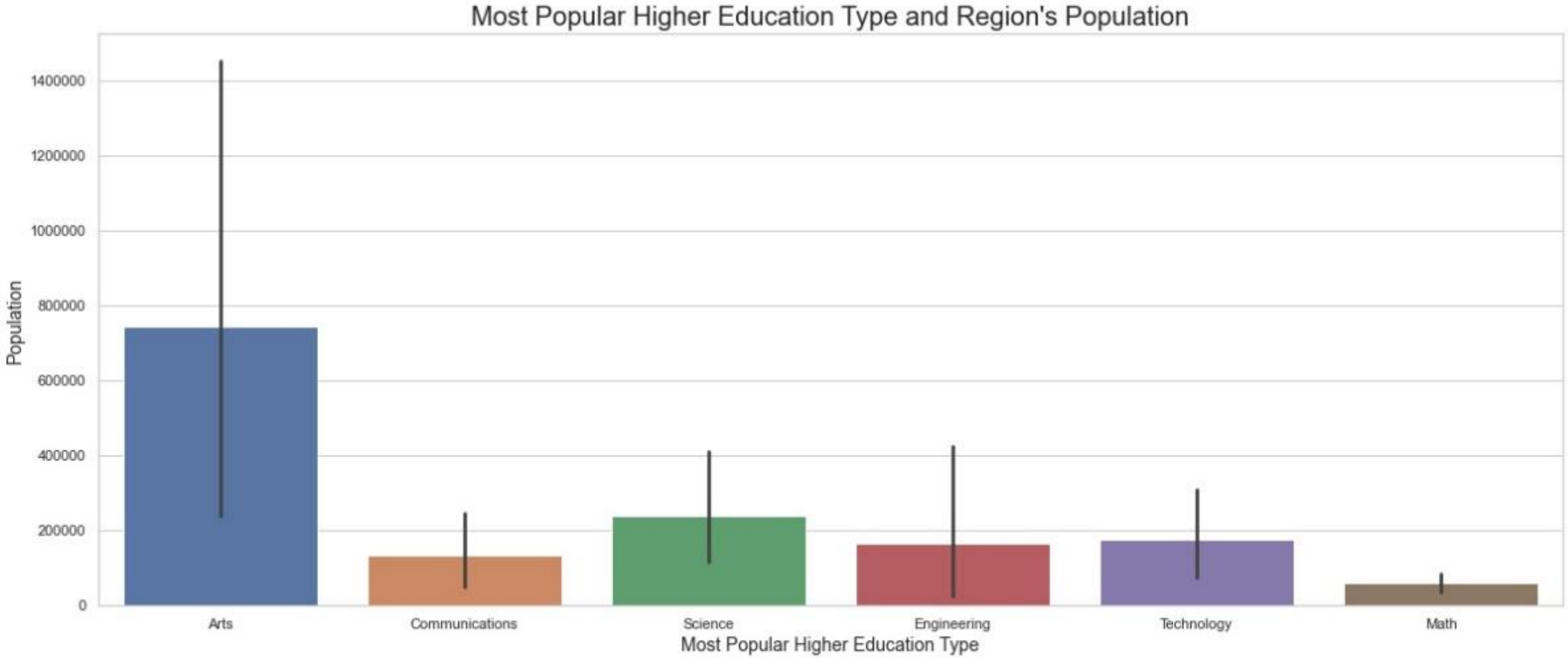
Dependent variable: type, which corresponds to the most popular academic type of a city/town

Independent variables: population (in persons), area size (in kilometers-squared) and west (a dummy which is 1 if the city is in the west).

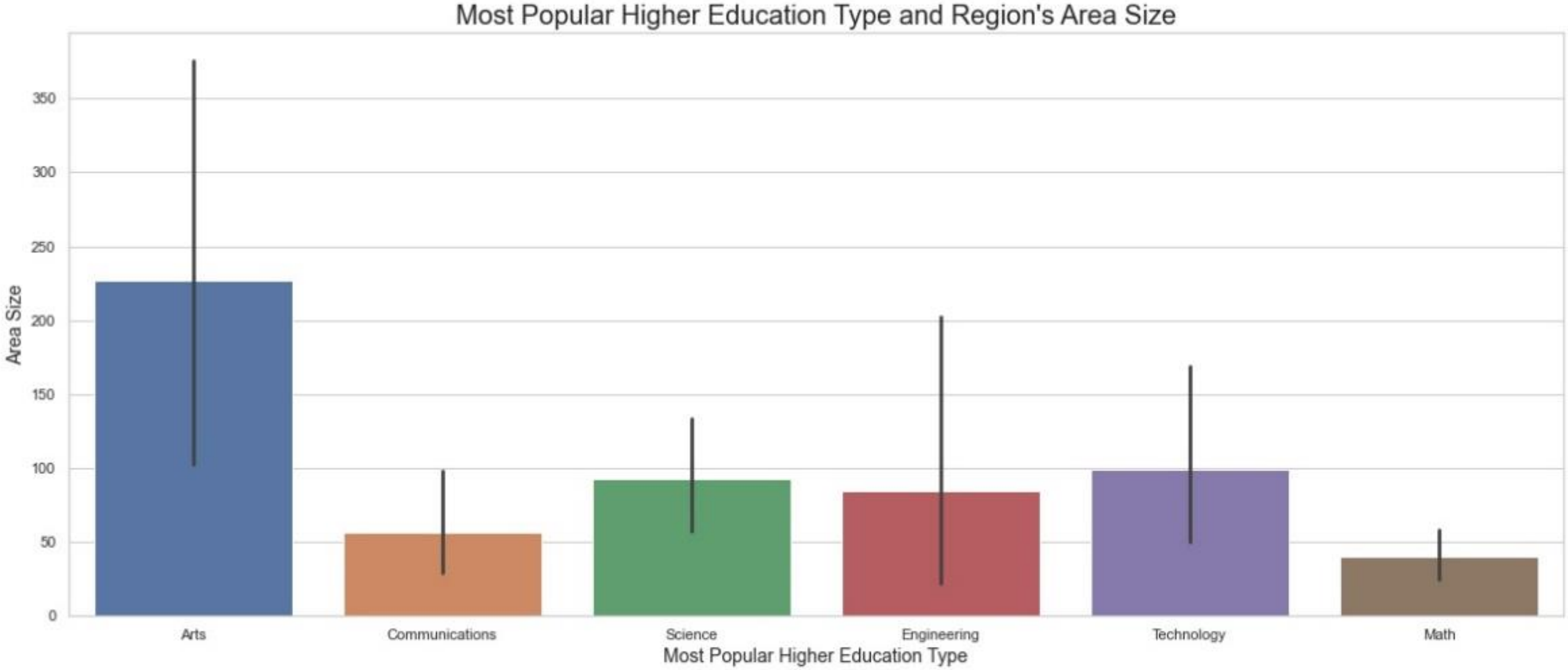
Check for multicollinearity of independent variables using Variance Inflation Factor method (VIF)

Regress dependent variable against the set of independent variables

Demographic Features



Demographic Features (con’t)



Multicollinearity?

Table 5. Variance Inflation Factor

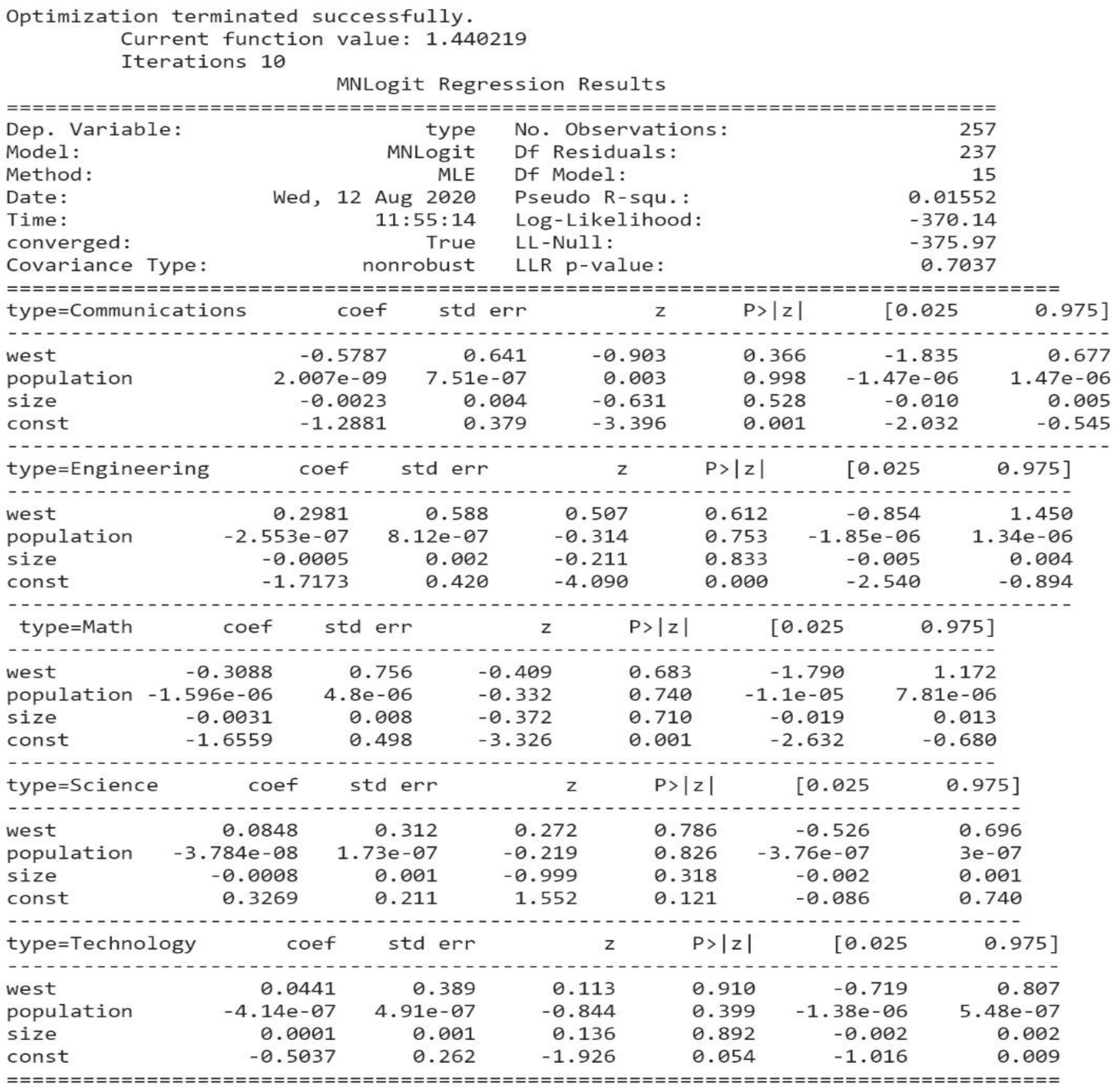
VIF Factor	features
1.024518	west
2.420938	population
2.445618	size
1.826251	const

Interpretation: a VIF value between 1 and 5 is considered “safe.” population, area size and west are not likely to be linearly related 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.

Regression Results

From the table (next slide), 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 region 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 associate between region and academic type distribution.



```
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
===================================================================================
```

Arts is used as a baseline comparison.

Each block in the regression table compares the corresponding academic category to Arts.

Discussion and limitations

1. The main downside of working with Foursquare API are the followings: The search result is inconsistent (different results for different searches), some minor incorrect sub-categorizations of venues and potentially missing venues that are not captured in the analysis
2. Sample size is the key problem in interpreting the logistic regression results. In particular, the p-value is majorly influenced by sample size. If our hypotheses are actually true (there is really a difference in education type distribution for different regions), then increasing sample size will decrease p-value and we can conclude there is indeed a difference.
3. Problems of missing important variables. 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.

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. There are two main findings:

1. Through the method of statistical inference, I find no association on the popularity of universities academic categories between the West and the East of U.S.
2. Through the method of 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. I discussed some limitations of the data set and how one should interpret the results appropriately given the scope of this research.