Team members

Name	BU Email
Janaki Riji Nair	janaki@bu.edu
Hongyi Yu	yuhy@bu.edu
Corey Evans	ceboston@bu.edu

Introduction:

Our project is an attempt towards predicting the tuition range of universities in the United States using machine learning. We have used decision tree, a model learned through machine learning algorithm, to predict the range of tuition incurred by students going to U.S. universities based on factors such as ranking of the university, the geographic location and setting of the university, type of school (public/private) and the number of students enrolled in undergraduate programs in the universities.

We use a publicly available dataset called the 2017 National Universities Rankings for our project. Since we use Classification Learning, a type of machine learning that classifies entities based on their characteristics, we transformed numeric attributes in our dataset into nominal attributes using Python. This included discretizing both the input variable of 'location' to place each university into East, West, North, South categories and the output variable of 'tuition' into ranges like less than $25,000; $25000 to $35,000; $35,000 to $45,000 and, above $45,000.

We then used SQL to find out the average tuition charged in various geographic regions in the United States. This is visualized in Figure 1 (in the Results section of this report) which shows that universities in the East had the highest tuition on average in the United States. We also visualized the highest tuition charged in various cities on the U.S. map (in the Results section) to compare the tuition costs across cities in the United States. After preparing the dataset, we used WEKA to learn a decision tree model that predicted the tuition range of universities. On testing, our model showed more than 76% accuracy in predicting the tuition range of U.S. universities.

***Dataset description** :*

Our dataset published by the U.S. News & World Report in 2017 was obtained as a .csv file from the data.world database.
(Dataset accessible from:
https://datasetsearch.research.google.com/search?query=Best%20National%20University%20Rankings&docid=L2cvMTFqb pnbHg1dg%3D%3D)

The following table summarizes the attributes of the dataset:

Attribute	Description
Input attribute	
Rank	This is a numeric attribute that refers to the national ranking of universities in the United States according to the U.S. News & World Report.
Setting	This is a nominal attribute that gives information on whether the university is situated in a city, urban, suburban or rural setting.
Location	This is a nominal attribute that refers to the city and state in which a university is located.
School type	This is a nominal attribute that describes whether the university is a public or private university.
Undergraduate enrollment	This is a numeric attribute that refers to the total number of students enrolled in undergraduate programs in a university.
Output attribute	
Tuition	This is a numeric attribute that would show the tuition charged by a university over the course of an academic year.

The relational table we created (included in the Appendix of this report) based on our dataset provides data on ranking, setting, location, (school) type, (undergraduate) enrollment and tuition charged by 151 universities located in the United States. While creating the relational table, we removed punctuation symbols, like commas and quotes, from the data. We also modified the attribute names so that they comprise single words; i.e. School type became ‘Type’ and Undergraduate enrollment became ‘Enrollment’. As the names of the universities are unique and irrelevant to our data analysis, we removed this attribute from the relational table. This was

followed by transforming some of the variables in our relational table to prepare our data for data mining.

***Data preparation**:*

We intended to use Classification learning to predict the range of tuition charged by universities in the United States. This meant our output variable which was available as a numeric attribute in our dataset had to be transformed into a nominal attribute. We used Python to discretize the tuition into nominal tuition ranges such as less than $25000, $25000 to $35000, $35000 to $45000, $45000 and higher. The Python program (given in the Appendix of the report) classified the data on tuition into the aforementioned ranges using for loops and if-elif statements.

We also classified the attribute of location by region. Data on the location attribute was available in the dataset in the form of city and state for each university. We classified this attribute into different geographic regions of the United States such as North, South, East, and West. We wrote a Python program for this that sifted through the dataset using the following functions: loops, splitting, join, lists, and elif statements. The Python program (given in the Appendix of the report) classified each location into one of the four regions.

WEKA was used to randomize the dataset prior to performing data mining on the discretized dataset. We randomized the dataset, then split the data into training and testing sets using an 80/20 ratio. As the entire dataset was 151 tuples, this equated to 121 rows in the training set, and 30 in the set for testing.

***Data analysis**:*

We used SQL on the discretized dataset to compute the average tuition of universities across different geographical regions in the United States. The SQL code used can be referred to in the Appendix of the report.

We tried to learn a classification analysis Decision Tree model from our training data on WEKA. Decision tree models select the best rules for each attribute, building a decision tree from those rules, and then choose less accurate rules to split into further rules that “best divide” the data. Each rule represents a “branch” on the tree, and the algorithm creates a decision path which routes through each branch until ending on a nominal output. The Decision Tree model showed an accuracy of 71.0744% on the training dataset. The model reported a higher accuracy of 76.6667% on the test dataset.

We also used the 1R algorithm on the training dataset. 1R, a machine learning algorithm designed by R.C. Holte, formulates one rule for each attribute in the attribute, and the most

accurate rule is chosen for the model. The 1R model, as processed through WEKA, was only 50.4132% accurate at classifying instances.

***Results**:*

Average tuition of U.S. Universities across geographic regions

The SQL program, which computed the average tuition of U.S. universities by geographic region, revealed the Eastern region of the United States to have the highest average tuition in the country. This is illustrated in Figure 1 below.

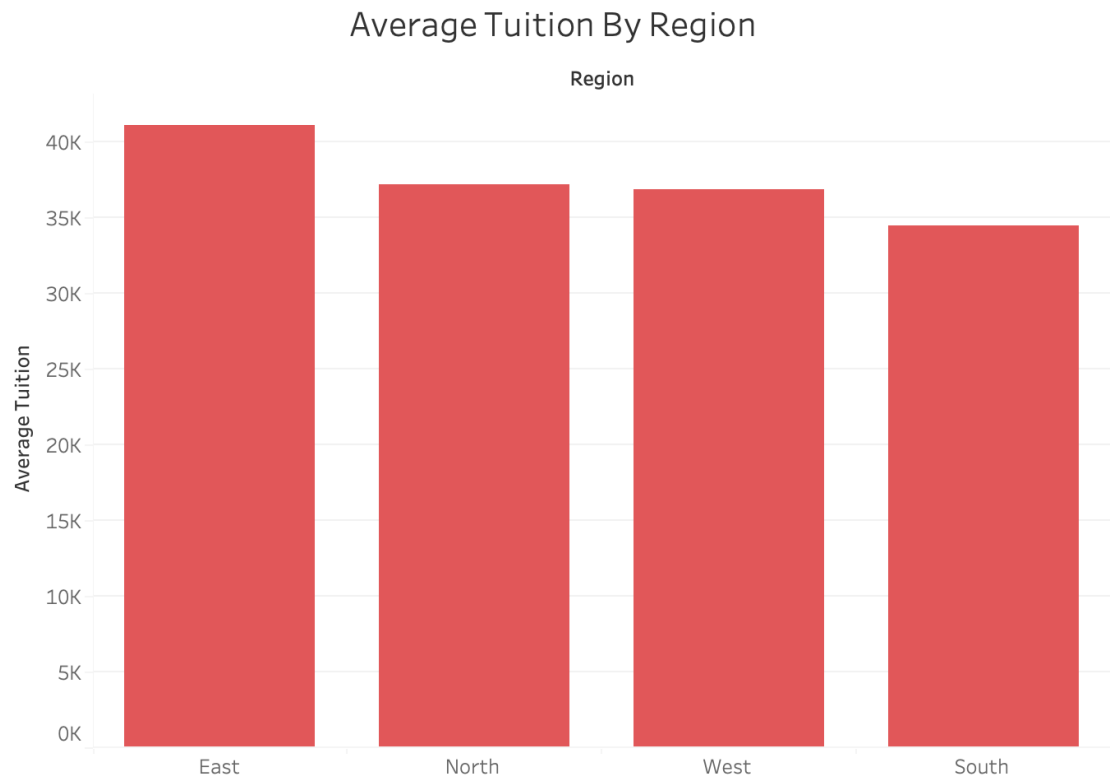
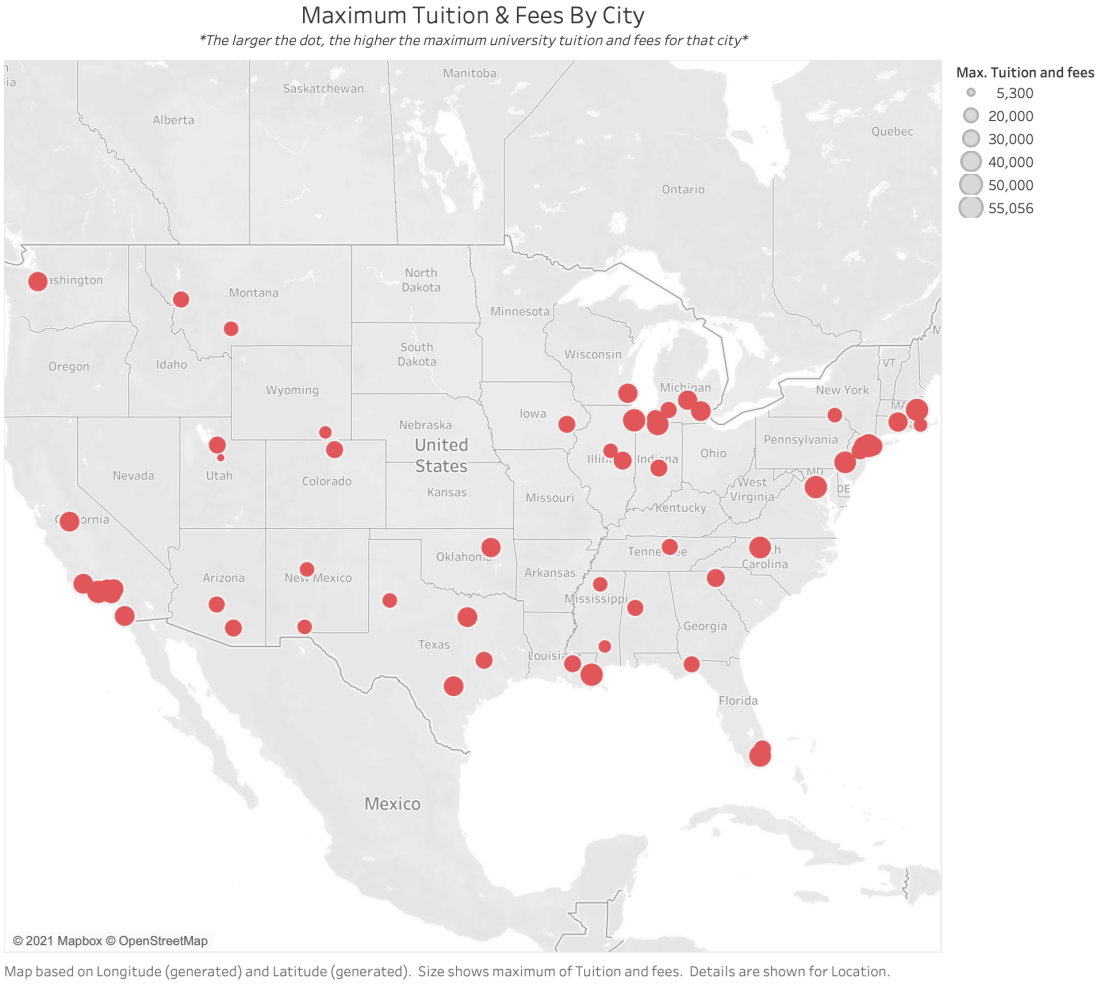


Figure 1: Bar Graph showing average tuition of U.S. universities by geographic region

To visualize the above figure, we plotted the averages and regions into a spreadsheet, and uploaded the Excel file to, data visualization software, Tableau. The results were displayed in a bar graph chart, organized in descending order. This highlights the most expensive region in contrast to the least expensive region to attend college in the United States.

Maximum tuition of U.S. Universities across cities

We also used Tableau to illustrate the highest tuition charged in each city in our dataset. To put this into perspective the entire raw dataset when fed through Tableau revealed the maximum college tuition and fees charged by universities in various cities in Figure 2 given below. To follow Tufte's design principles and use "less ink", tuition was distinguished by bubble size: the larger the bubble, the higher the tuition charged by a university in that city. The results reflect the costliness of universities in each region displayed.



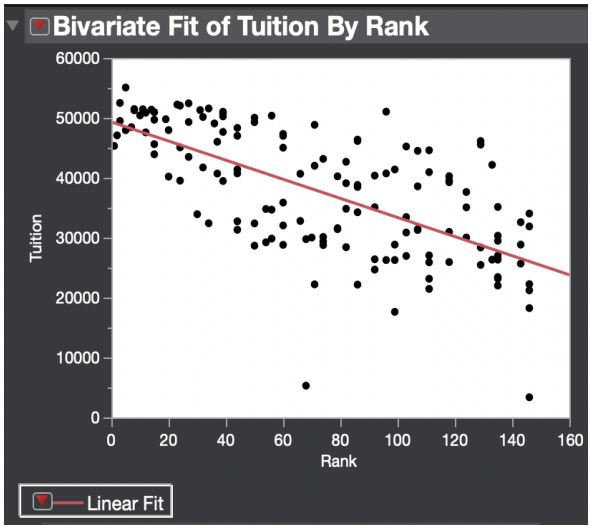
Map based on Longitude (generated) and Latitude (generated). Size shows maximum of Tuition and fees. Details are shown for Location.

Figure 2: Map of United States showing Maximum University Tuition by U.S. city

Relationship between U.S. University tuition and Ranking/Enrollment

We also explored the relationship between numeric attributes in our dataset using the linear regression model and visualized the linear relationship fit through the regression lines shown in Figure 3 and Figure 4 given below. We found an obvious tendency in the scattered plot and concluded that both rank of university and undergraduate enrollment have a strong relationship

with tuition charged by the university. The tuition tends to be higher when the rank of the university is higher and the enrollment of students is smaller. The graph also shows the outliers that do not quite fit the rules.



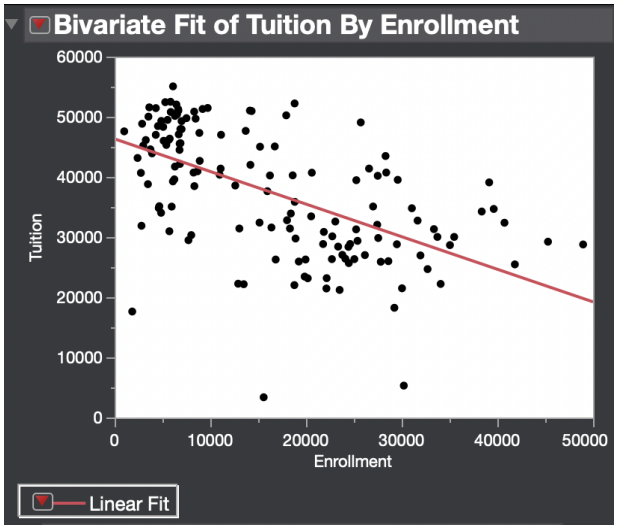


Figure 3 (left): Bivariate Fit of University Tuition by Rank. Figure 4 (right): Bivariate Fit of University Tuition by Enrollment

Predicting tuition range of U.S. Universities

The decision tree model generated on WEKA by the training dataset generalizes moderately well, which was indicated by the approximate accuracy of 77% when used on the test dataset. The results of the Decision Tree suggest that higher education institution's tuition may be influenced by the following attributes: School Type (Public/Private), Rank, Location, and Enrollment. Intuitively, this makes sense as a university's published rank and type may influence the administration's decision on the monetary amount to charge students. Additionally, if a school is located in an expensive city like Boston, it could be logically presumed to have a costlier price tag. When the Decision Tree Model was run on the Training dataset, it resulted in the Confusion Matrix given in the following Table 1.

	PREDICTED			
ACTUAL	$45,000 and Up	$35,000 to $45,000	$25,000 to $35,000	$25,000 and Lower
$45,000 and Up	28	5	1	0
$35,000 to $45,000	3	20	8	1
$25,000 to $35,000	0	2	38	0
$25,000 and Lower	1	1	13	0

Table 1: Confusion Matrix on Training Dataset (Decision Tree model)

When the Decision Tree Model was used on the Testing dataset, it resulted in the Confusion Matrix given in the following Table 2.

	PREDICTED			
ACTUAL	$45,000 and Up	$35,000 to $45,000	$25,000 to $35,000	$25,000 and Lower
$45,000 and Up	15	1	0	0
$35,000 to $45,000	0	1	2	0
$25,000 to $35,000	0	2	7	1
$25,000 and Lower	0	0	1	0

Table 2: Confusion Matrix of Test Dataset (Decision Tree model)

Given below in Figure 5 is the decision tree learned from the Training dataset used in this project.

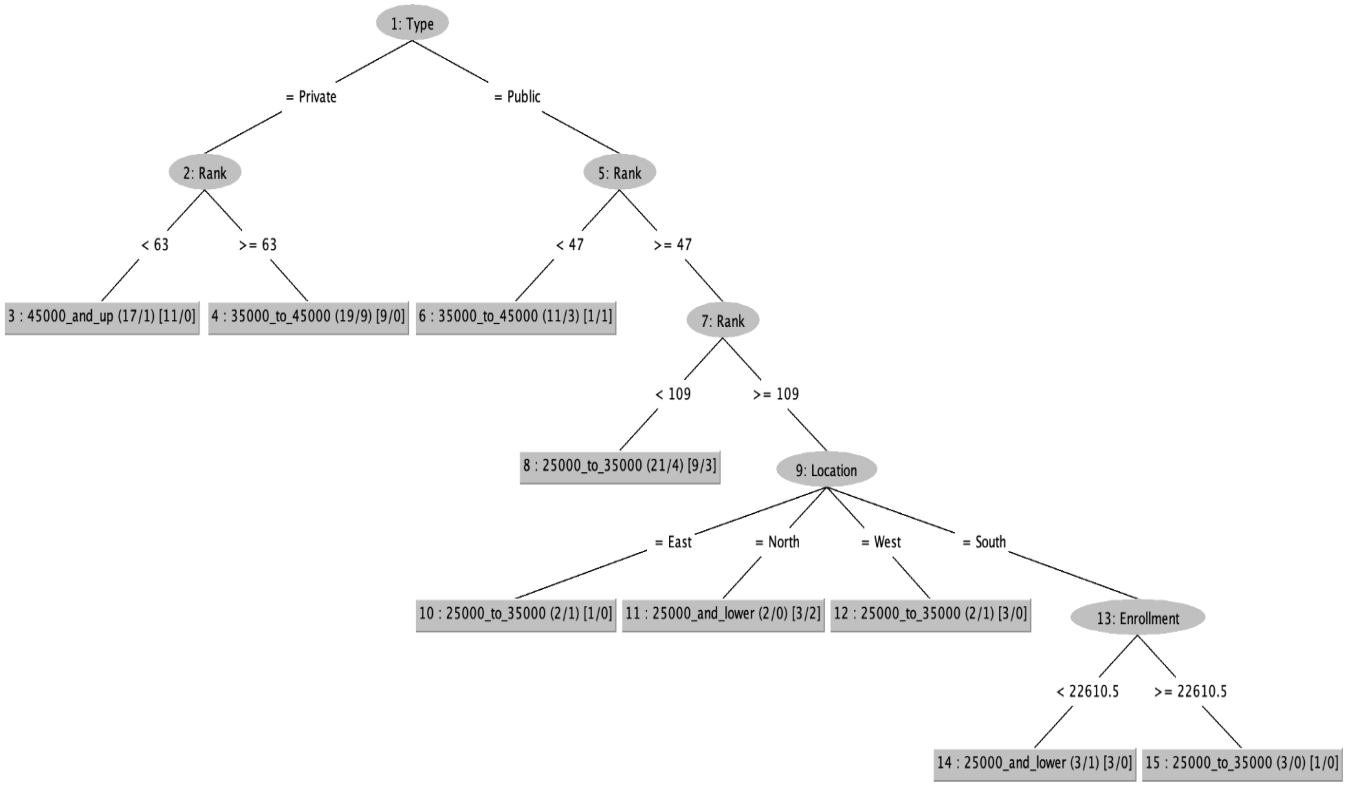


Figure 5: Decision Tree Model derived from Training Dataset

***Conclusions**:*

Our team used the National Universities Ranking Database from *US News*, carefully discretized the variables through Python programming, and sorted through the data with SQL and WEKA. Upon applying the classification analysis Decision Tree machine learning algorithm, we were able to predict a U.S. university's tuition range on a test set of data using a set of attributes - Type (Public/Private), Rank, Location, and Enrollment - with an approximate accuracy of 77%.

***Appendix**:*

Python program used for classifying 'location' attribute and discretizing 'tuition' attribute

```
infile = open('National-Universities-Rankings.txt', 'r')
outfile = open('National-Universities-Rankings_Classified.txt', 'w')

header = infile.readline()
for line in infile:
    line = line[:-1]
    fields = line.split(',')
    state = fields[0][-2:]
    if state == 'MT' or state == 'WY' or state == 'ND' or state == 'SD' or state == 'NE' or state ==
'MN' or state == 'IA' or state == 'WI' or state == 'IL' or state == 'MI' or state == 'IN' or state ==
'WV' or state =='PA' or state == 'AK' or state == 'OH':
        fields[0] = 'North'
    elif state == 'NM' or state == 'TX' or state == 'LA' or state == 'MS' or state == 'AL' or state ==
'FL' or state == 'GA' or state == 'SC' or state == 'NC' or state == 'VA' or state == 'TN' or state ==
'MO' or state == 'KS' or state == 'KY' or state == 'CO' or state == 'OK' or state == 'AR':
        fields[0] = 'South'
    elif state == 'ME' or state == 'VT' or state == 'NH' or state == 'MA' or state == 'CT' or state ==
'NJ' or state == 'DE' or state == 'MD' or state == 'NY' or state == 'RI' or state == 'DC':
        fields[0] = 'East'
    elif state == 'WA' or state == 'OR' or state == 'CA' or state == 'AZ' or state == 'ID' or state ==
'NV' or state == 'UT' or state == 'HI':
        fields[0] = 'West'

    tuition = int(fields[-1])
    if tuition <= 25000:
        fields[-1] = '25000_and_lower'
    elif tuition > 25000 and tuition <= 35000:
        fields[-1] = '25000_to_35000'
    elif tuition > 35000 and tuition <= 45000:
        fields[-1] = '35000_to_45000'
    elif tuition >= 45000:
        fields[-1] = '45000_and_up'
```

```
    print(fields[0], fields[1], fields[2], fields[3], fields[4], fields[5], file = outfile)

infile.close()
outfile.close()
```

SQL program to determine average university tuition by geographical region

```
SELECT AVG(Tuition)
FROM University
WHERE Location = 'North';

SELECT AVG(Tuition)
FROM University
WHERE Location = 'West';

SELECT AVG(Tuition)
FROM University
WHERE Location = 'East';

SELECT AVG(Tuition)
FROM University
WHERE Location = 'South';
```

Cross Validation Analysis of Decision Tree Model of Training Dataset (10 folds)

```
Size of the tree : 15
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          86               71.0744 %
Incorrectly Classified Instances        35               28.9256 %
Kappa statistic                          0.5865
Mean absolute error                      0.2075
Root mean squared error                  0.3337
Relative absolute error                 57.0255 %
Root relative squared error             78.2489 %
Total Number of Instances              121
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.824	0.046	0.875	0.824	0.848	0.792	0.937	0.855	45000_and_up
0.625	0.090	0.714	0.625	0.667	0.560	0.807	0.626	35000_to_45000
0.950	0.272	0.633	0.950	0.760	0.638	0.843	0.620	25000_to_35000
0.000	0.009	0.000	0.000	0.000	-0.034	0.684	0.234	25000_and_lower
Weighted Avg. 0.711	0.128	0.644	0.711	0.666	0.577	0.840	0.640	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
28	5	1	0	a = 45000_and_up
3	20	8	1	b = 35000_to_45000
0	2	38	0	c = 25000_to_35000
1	1	13	0	d = 25000_and_lower

Figures

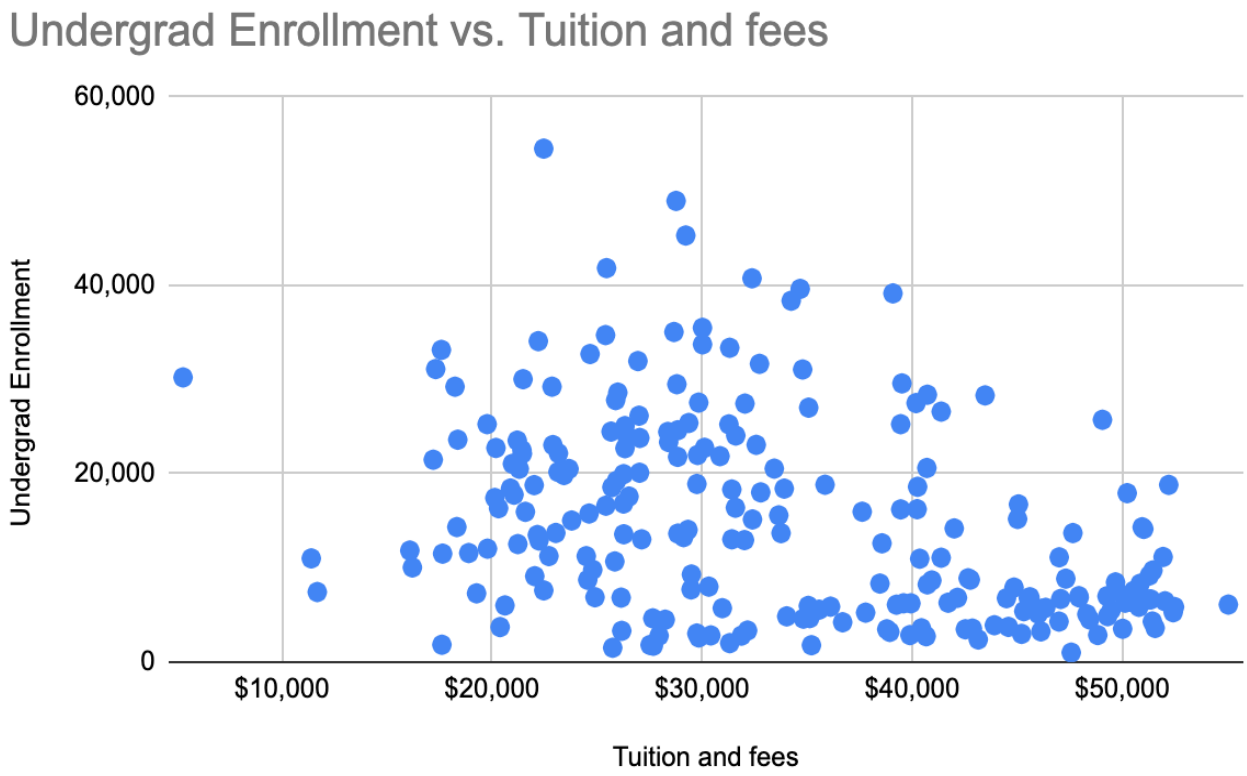


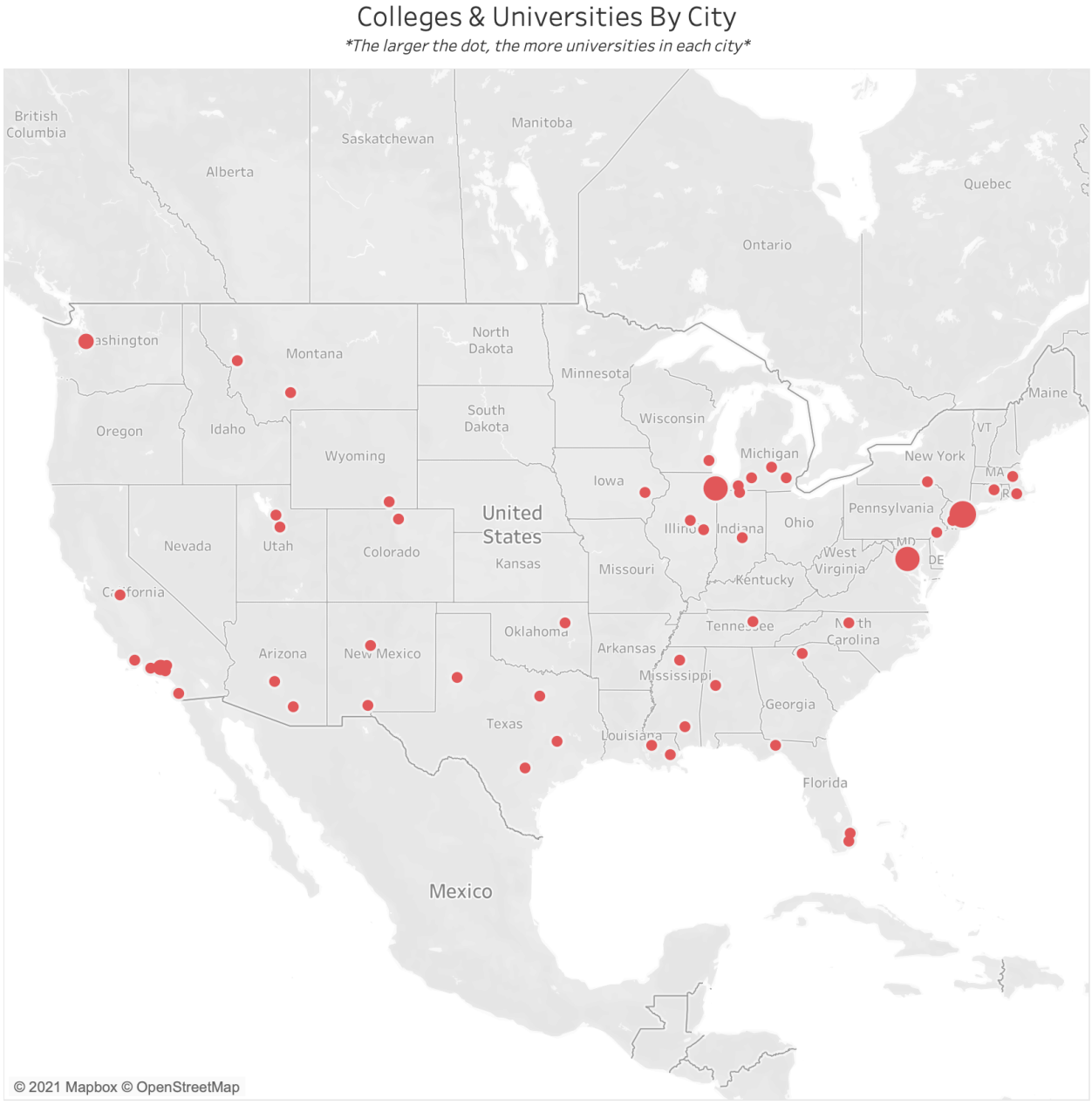
Figure 6: Scatter Plot of Undergraduate Enrollment & University Tuition/Fees

College Tuition & Fees By City

The larger the dot, the more expensive the tuition

Map based on Longitude (generated) and Latitude (generated). Size shows sum of Tuition and fees. Details are shown for Location.

Figure 7: Sum of University Tuition/Fees by City



Map based on Longitude (generated) and Latitude (generated). Size shows count of Tuition and fees. Details are shown for Location.

Figure 8: Count of Universities in the United States by City

Relational table

Location	Rank	Setting	Type	Enrollment	Tuition
East	1	Suburban	Private	5402	45000_and_up
East	2	Urban	Private	6699	45000_and_up
North	3	Urban	Private	5844	45000_and_up
East	3	City	Private	5532	45000_and_up
East	5	Urban	Private	6102	45000_and_up
West	5	Suburban	Private	6999	45000_and_up
East	7	Urban	Private	4527	45000_and_up
South	8	Suburban	Private	6639	45000_and_up
North	8	Urban	Private	9726	45000_and_up
East	10	Urban	Private	6524	45000_and_up
East	11	Rural	Private	4307	45000_and_up
West	12	Suburban	Private	1001	45000_and_up
North	12	Suburban	Private	8314	45000_and_up
East	14	City	Private	6652	45000_and_up
East	15	Rural	Private	14315	45000_and_up
South	15	Urban	Private	3910	35000_to_45000
North	15	Suburban	Private	8462	45000_and_up
South	15	Urban	Private	6883	45000_and_up
South	19	City	Private	7504	45000_and_up
South	20	City	Private	6867	45000_and_up
East	20	Urban	Private	7562	45000_and_up
West	20	City	Public	27496	35000_to_45000
West	23	Urban	Private	18810	45000_and_up
North	24	Urban	Private	6454	45000_and_up
West	24	Urban	Public	29585	35000_to_45000
South	24	Suburban	Public	16736	45000_and_up
East	27	Suburban	Private	5290	45000_and_up
North	27	City	Public	28312	35000_to_45000
South	27	Suburban	Private	4871	45000_and_up
South	30	Suburban	Public	18415	25000_to_35000
East	31	Suburban	Private	9192	45000_and_up
South	32	Suburban	Public	6301	35000_to_45000

East	32	Urban	Private	6304	45000_and_up
East	34	Suburban	Private	3621	45000_and_up
South	34	Urban	Public	15142	25000_to_35000
East	36	Urban	Private	25722	45000_and_up
North	37	Urban	Private	5121	45000_and_up
West	37	Suburban	Public	20607	35000_to_45000
East	39	Urban	Private	17932	45000_and_up
East	39	Urban	Private	13697	45000_and_up
East	39	Suburban	Private	5864	45000_and_up
South	39	Urban	Private	6662	45000_and_up
West	39	Suburban	Public	25256	35000_to_45000
North	44	City	Private	5075	45000_and_up
West	44	City	Public	28384	35000_to_45000
West	44	Urban	Public	26590	35000_to_45000
North	44	City	Public	33368	25000_to_35000
South	44	Suburban	Private	11122	45000_and_up
North	44	City	Public	31662	25000_to_35000
West	50	Suburban	Private	3533	45000_and_up
South	50	Suburban	Public	35043	25000_to_35000
North	50	Suburban	Private	6994	45000_and_up
North	50	City	Public	40742	25000_to_35000
North	54	Urban	Public	45289	25000_to_35000
West	54	Urban	Public	31063	25000_to_35000
East	56	Urban	Private	11157	45000_and_up
South	56	Urban	Private	6411	45000_and_up
South	56	City	Public	27547	25000_to_35000
South	56	Suburban	Public	39619	25000_to_35000
East	60	Urban	Private	8855	45000_and_up
North	60	City	Public	29497	25000_to_35000
East	60	City	Private	15196	45000_and_up
East	60	Rural	Public	18826	35000_to_45000
East	60	Suburban	Public	27443	25000_to_35000
East	60	City	Private	4299	45000_and_up
South	66	Suburban	Public	18016	25000_to_35000
East	66	Urban	Private	2744	35000_to_45000

West	68	City	Private	30221	25000_and_lower
North	68	Urban	Public	18908	25000_to_35000
East	70	City	Public	35484	25000_to_35000
South	71	City	Private	14189	35000_to_45000
East	71	City	Private	2873	45000_and_up
North	71	Urban	Public	34071	25000_and_lower
East	74	Suburban	Private	7909	35000_to_45000
East	74	City	Private	2397	35000_to_45000
South	74	City	Public	48960	25000_to_35000
East	74	Suburban	Public	22748	25000_to_35000
South	74	Suburban	Public	25384	25000_to_35000
North	79	Rural	Public	16387	25000_to_35000
West	79	Suburban	Public	16231	35000_to_45000
East	79	Suburban	Public	18322	25000_to_35000
South	82	Suburban	Public	4608	25000_to_35000
North	82	Suburban	Public	39143	35000_to_45000
South	82	Suburban	Private	8894	35000_to_45000
North	82	City	Public	23357	25000_to_35000
East	86	Suburban	Public	13491	25000_and_lower
North	86	City	Public	38364	25000_to_35000
North	86	Urban	Private	8334	35000_to_45000
South	86	City	Private	5758	45000_and_up
West	86	Urban	Private	5647	45000_and_up
South	86	City	Private	3478	35000_to_45000
South	92	City	Public	32706	25000_and_lower
South	92	City	Public	24111	25000_to_35000
South	92	City	Public	27010	35000_to_45000
East	92	Suburban	Public	10973	35000_to_45000
North	96	Urban	Private	14138	45000_and_up
South	96	Urban	Private	8248	35000_to_45000
East	96	Suburban	Public	16831	25000_to_35000
South	99	Suburban	Public	21786	25000_to_35000
North	99	City	Private	11079	35000_to_45000
East	99	City	Public	1839	25000_and_lower
East	99	Suburban	Public	19951	25000_to_35000

North	103	Urban	Private	2991	45000_and_up
South	103	Suburban	Public	31958	25000_to_35000
West	103	City	Public	20538	25000_to_35000
South	103	Urban	Public	21863	25000_to_35000
East	107	Suburban	Private	12607	35000_to_45000
East	107	Suburban	Public	13034	25000_to_35000
West	107	Urban	Private	6782	35000_to_45000
South	107	City	Public	25237	25000_to_35000
North	111	City	Public	30034	25000_and_lower
North	111	Suburban	Private	8665	35000_to_45000
South	111	City	Public	27812	25000_to_35000
North	111	City	Public	20182	25000_and_lower
South	111	City	Public	22132	25000_and_lower
West	111	City	Private	3735	35000_to_45000
West	111	Urban	Public	23794	25000_to_35000
North	118	Rural	Public	5721	25000_to_35000
East	118	Suburban	Private	6090	35000_to_45000
North	118	Urban	Public	28609	25000_to_35000
West	118	City	Public	18608	35000_to_45000
South	118	City	Public	19245	25000_to_35000
North	118	Urban	Private	6240	35000_to_45000
East	124	Urban	Private	3480	35000_to_45000
North	124	Urban	Private	15961	35000_to_45000
North	124	Urban	Private	5961	35000_to_45000
East	124	Urban	Private	6883	25000_and_lower
West	124	City	Public	33732	25000_to_35000
West	129	Urban	Public	41828	25000_to_35000
East	129	Rural	Private	3257	45000_and_up
South	129	City	Public	24433	25000_to_35000
East	129	Urban	Private	6792	45000_and_up
East	133	Suburban	Private	6824	35000_to_45000
South	133	City	Public	22705	25000_to_35000
South	135	City	Public	19859	25000_and_lower
South	135	Urban	Public	26156	25000_to_35000
South	135	City	Private	4667	35000_to_45000

East	135	Urban	Public	8008	25000_to_35000
East	135	Urban	Public	7713	25000_to_35000
South	135	City	Public	22159	25000_and_lower
North	135	Urban	Public	25054	25000_to_35000
South	135	Rural	Public	18785	25000_and_lower
South	143	Suburban	Public	23062	25000_to_35000
West	143	City	Public	24612	25000_to_35000
West	143	Rural	Public	24470	25000_to_35000
East	146	Suburban	Private	4852	25000_to_35000
North	146	Suburban	Public	23513	25000_and_lower
West	146	Urban	Public	29234	25000_and_lower
East	146	Suburban	Private	2805	25000_to_35000
East	146	Suburban	Public	12908	25000_and_lower
South	146	Suburban	Public	15575	25000_and_lower

Table 3: Discretized National Universities Ranking Relation