CS838 - Stage 5 Report

Trang Ho, Thomas Ngo, Qinyuan Sun

1. Dataset

In this project stage, we will do analysis on the merged table E of two tables AOM and IPEDS. The CSV file storing table E can be found here:
https://github.com/TrangHo/cs838-spring2017/blob/master/stage5/data/_aom_mapped_v2.csv?raw=true

This merged table consists merges the person ID (*pid*) of AOM to his/her associated affiliation in IPEDS. The AOM has data of individuals submitting papers from 2006 - 2014. Hence, each tuple in the merged table represents a conference attendance of an individual at the specific year, together with the individual's affiliation information.

Below is the schema of table E:

Attribute Name	Description
year	The year the individual attended the conference
pid	The individual ID
ipeds_aid	The individual's affiliation ID
ipeds_name	The individual's affiliation name
ipeds_alias	The individual's affiliation alias
ipeds_city	The individual's affiliation city
ipeds_prov	The individual's affiliation province
ipeds_web	The individual's affiliation website
GROFFER	The individual's affiliation graduate offering
CCSIZSET	The individual's affiliation Size and Setting by Carnegie Classification 2010
INSTSIZE	The individual's affiliation institution size category
CBSATYPE	The individual's affiliation CBSA Type Metropolitan or Micropolitan

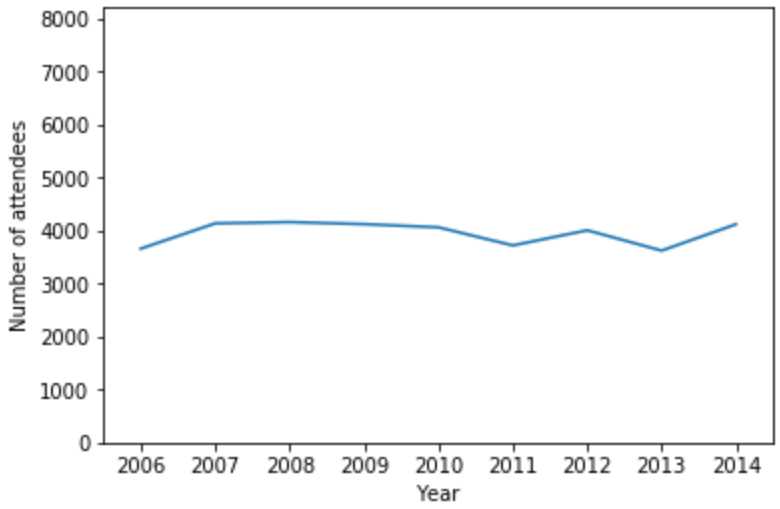
The possible categorical values of GROFFER, CCSIZSET, INSTSIZE, CBSATYPE attributes can be found here:
https://github.com/TrangHo/cs838-spring2017/blob/master/stage5/data/ipeds_variable%20details.xlsx

There are total of 35585 tuples in table E. Below are 5 sample tuples from E:

year	pid	ipeds_aid	ipeds_name	ipeds_alias	ipeds_city	ipeds_prov	ipeds_web	GROFFER	CCSIZSET	INSTSIZE	CBSATYPE
2007	296447	100663	university of alabama at birmingham	0	birmingham	alabama	www.uab.edu	1	15	4	1
2014	653507	100663	university of alabama at birmingham	0	birmingham	alabama	www.uab.edu	1	15	4	1
2008	287610	100663	university of alabama at birmingham	0	birmingham	alabama	www.uab.edu	1	15	4	1
2014	18785	100663	university of alabama at birmingham	0	birmingham	alabama	www.uab.edu	1	15	4	1
2009	66090	100663	university of alabama at birmingham	0	birmingham	alabama	www.uab.edu	1	15	4	1

2. Analysis

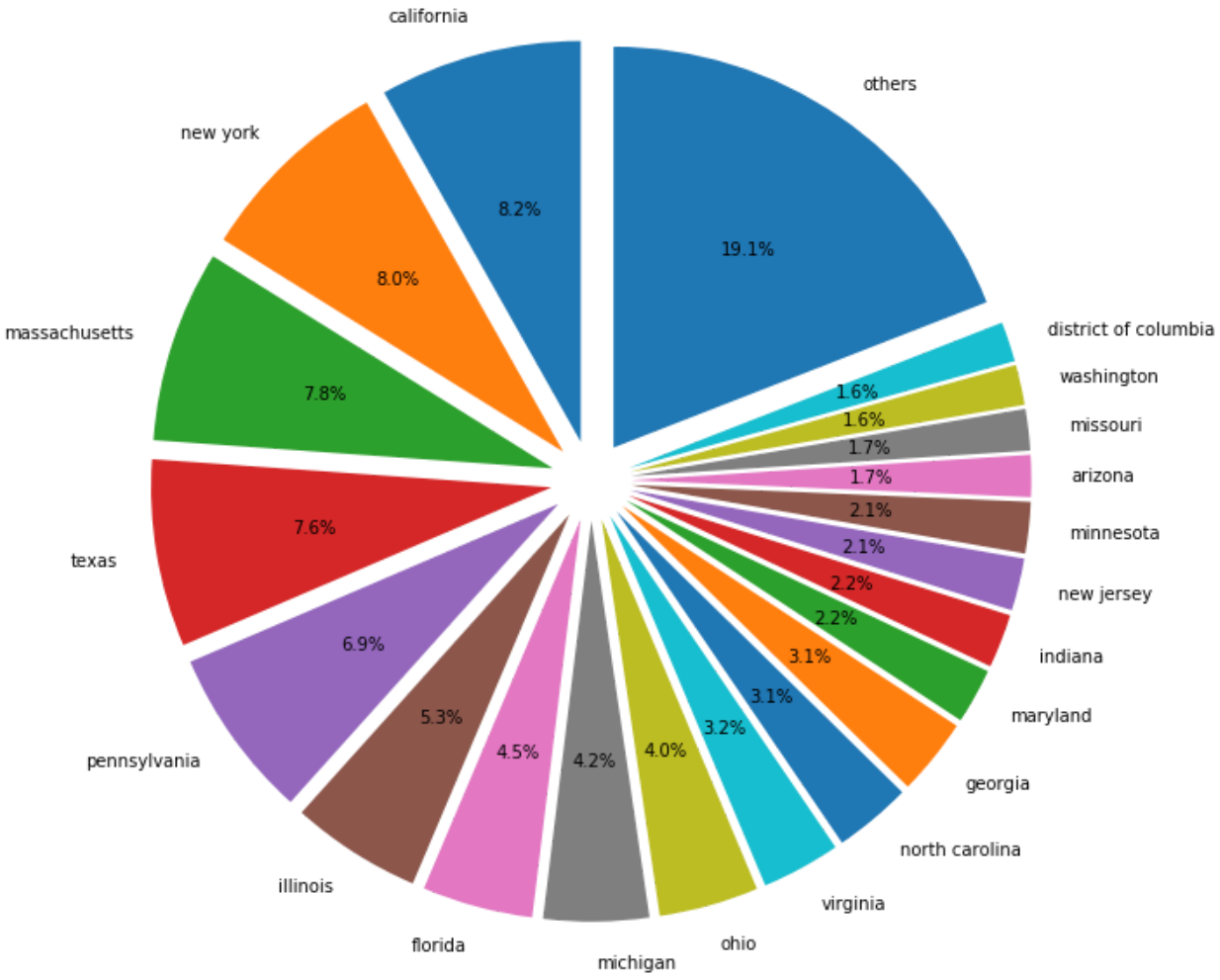
2.1 Trend in Number of Conference Attendees by Year



First analysis we did is to check how attendance of the conference varies. We calculate the attendance for the conference each year by grouping by year and aggregating on pid. We can see that even there are a few fluctuations in attendance, the number of attendees by year seems to be stable. The detail numbers can be found in the table below.

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014
#Attendees	3657	4134	4158	4119	4059	3718	4003	3620	4117

2.2 Conference Attendees Grouped by State



The chart above displays how the conference attendees are distributed by affiliations from different states from 2006 to 2014. The chart only lists the first 20 states with the most conference attendees. We have also done the same analysis for each year and found that the distribution is the same each year from 2006 to 2014.

To obtain this chart, we first grouped the data by state (*ipeds_prov*) and aggregated on the count of person ID (*pid*). After that we sorted the obtained data by the count of *pid* to have the list of top 20 states.

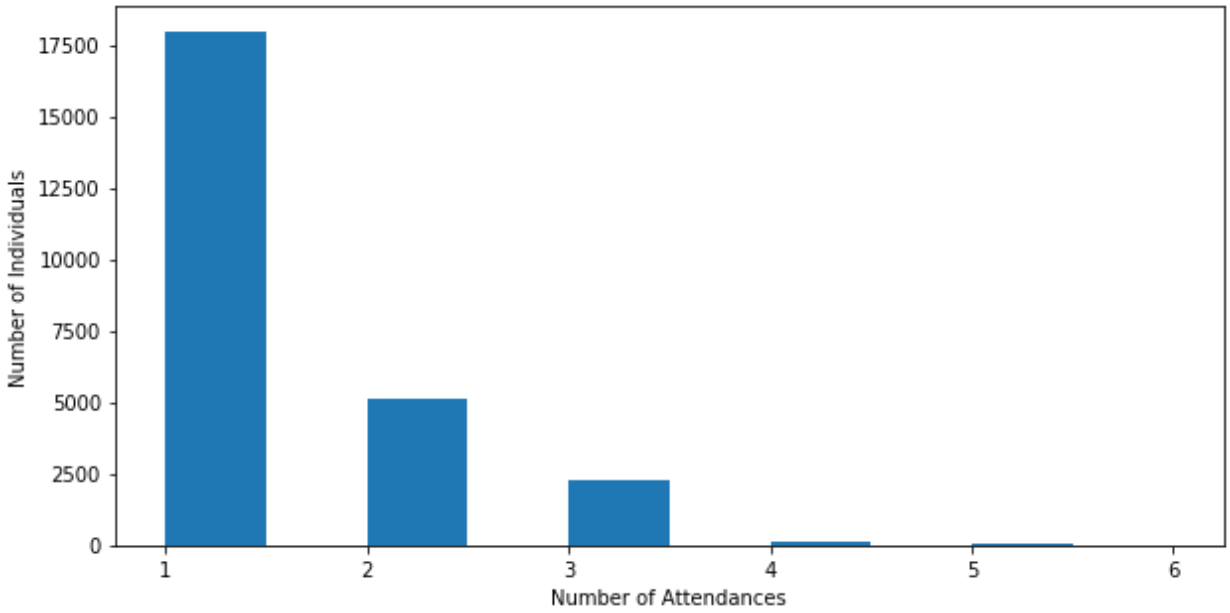
To obtained the top states for each year, before we group the data by the state, we first grouped the data by *year*. After that, the same process is applied to get the list of top 20 states for each year.

The data of the chart can be found in the table below.

State	Total # of Attendees from 2006 to 2014
California	2907
New York	2833
Massachusetts	2758
Texas	2689
Pennsylvania	2456
Illinois	1873
Florida	1590
Michigan	1512
Ohio	1430
Virginia	1132
North Carolina	1113
Georgia	1106
Maryland	794
Indiana	772
New Jersey	750
Minnesota	735
Arizona	599
Missouri	594
Washington	576
District of Columbia	572
Others	6794

2.3 Conference Attendance Frequency from 2006 to 2014

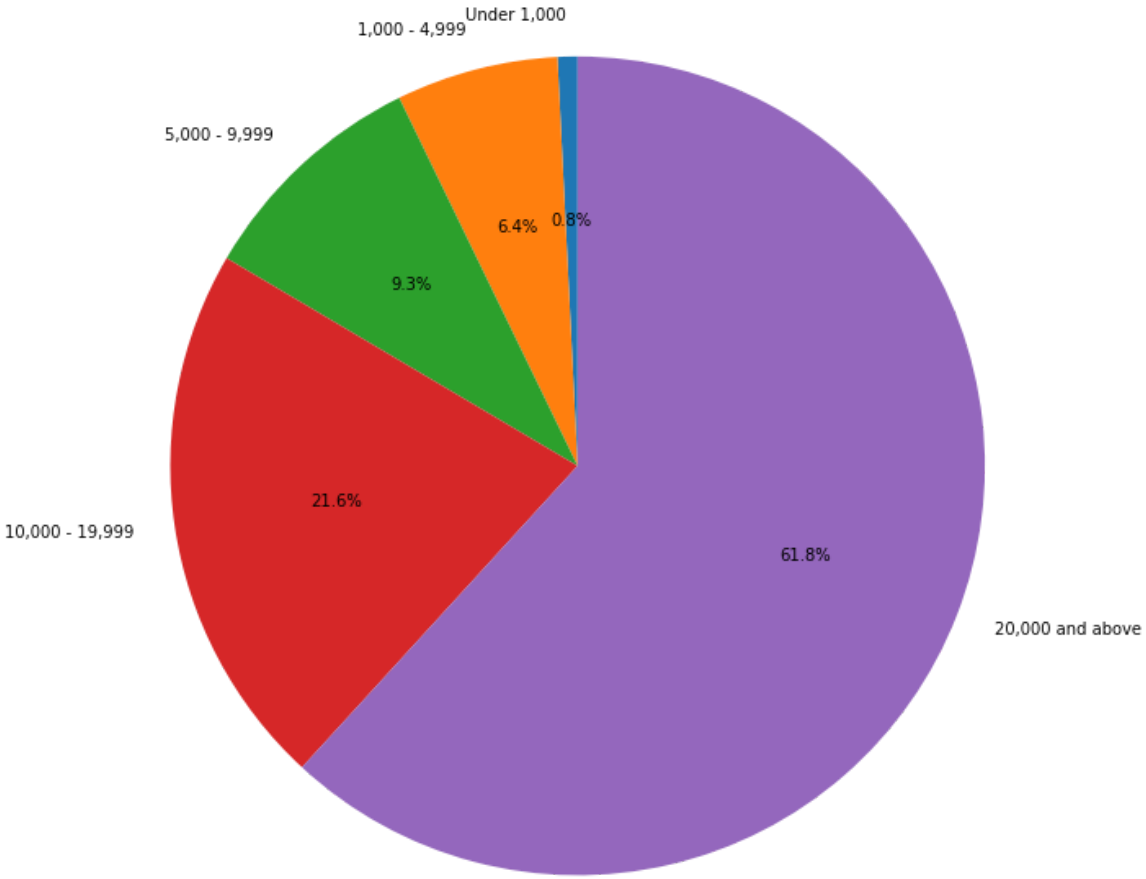
The histogram chart below shows the frequency of conference attendance from 2006 to 2014. As we can see, most individuals attended only one conference from 2006 to 2014 and more than 99% individuals attended at most three conferences from 2006 to 2014. There is only two individuals who attended six conferences from 2006 to 2014.



To obtain the data, we first grouped the data by person ID (*pid*), and then count the number of tuples that share the same *pid*. After that, we use the Python matplotlib to plot the histogram of the obtained data. Since there are too many individuals (25492 unique individuals) to list here, we will display the data of the histogram in the following table.

#Attendances	1	2	3	4	5	6
#Attendees	17972	5132	2244	105	37	2

2.5 Conference Attendances Grouped by Institution Size Category (INSTSIZE) from 2006 to 2014



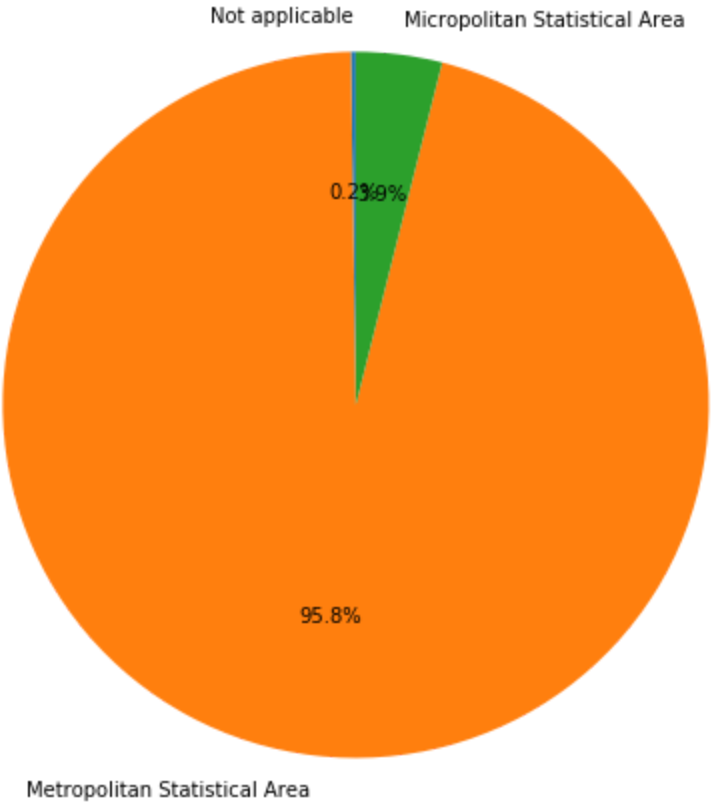
The chart above displays the proportion of conference attendances of each Institution Size Category. As we can see, the bigger the size of the affiliation, the more conference attendances the affiliation has. This chart shows a correlation between the number of attendances and the affiliation size.

To obtain the data for this chart, we first grouped the data by the institution size category (*INSTSIZE*) and then aggregated on the count of attendees (person ID *pid*). Below is the data table for the chart. We then use the Python matplotlib to plot the pie chart from these data.

Please notice that the institution size category is displayed in number, for the full understanding of these number, please refer to the possible categorical values link we share above.

INSTSIZE	1	2	3	4	5
#Attendees	281	2287	3314	7702	22001

2.6 Conference Attendances Grouped by CBSA Type of Affiliations (CBSATYPE) from 2006 to 2014



As we can see from the chart, most of the attendees were from affiliations that are in metropolitan statistical area. It seems that people from affiliations in area with low population are in more disadvantage regarding doing research and attending conferences than people from affiliations in area with higher population.

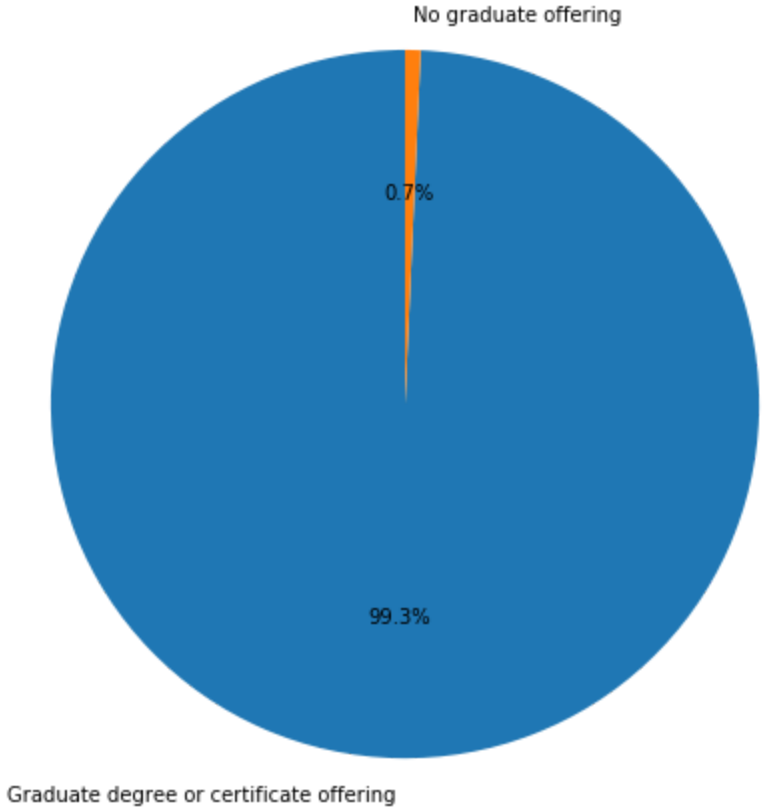
To obtain the data for this chart, we first grouped the data by CBSA type (*CBSATYPE*) and then aggregated on the count of attendees (person ID *pid*). We then use the Python matplotlib to plot the pie chart from these data.

Below is the data table for the chart. Please notice that the CBSA type is displayed in number, for the full understanding of these number, please refer to the possible categorical values link we share above.

CBSATYPE	-2	1	2
#Attendees	80	34105	1400

2.7 Conference Attendances Grouped by Whether Affiliations Offer Graduate Degree (GROFFER) from 2006 to 2014

The chart below shows the ratio of conference attendances between affiliations that offer graduate degree and affiliations that do not. As we can see, the ratio is significantly skewed to affiliations that offer graduate degree. This makes sense since there most of the research is often done at graduate level.



To obtain the data for this chart, we first grouped the data by whether affiliations offer graduate degree (*GROFFER*) and then aggregated on the count of attendees (person ID *pid*). We then use the Python matplotlib to plot the pie chart from these data.

Below is the data table for the chart. Please notice that the *GROFFER* is displayed in number, for the full understanding of these number, please refer to the possible categorical values link we share above.

GROFFER	1	2
#Attendees	35334	251

2.8 Multivariate Linear Regression

In this section, we ran a multivariate linear regression analysis using sklearn. The independent variable is number of attendees for each school in from 2006 to 2014. The covariates are GROFFER, CCSIZSET, INSTSIZE, and CBSATYPE. We want to determine which covariates are important in determining the attendance. Therefore, we have 4 null hypotheses, each of which hypothesizes that the corresponding covariate has no correlation with conference attendance. In order to test the hypotheses, we use Ordinary Least Square regression method.

Null Hythotheses:

H1: Affiliation's graduate course offer is not correlated with the number of affiliation's conference attendants
H2: Affiliation's size is not correlated with the number of affiliation's conference attendants
H3: Affiliation's institute size is not correlated with the number of affiliation's conference attendants
H4: Affiliation's CBSA type is not correlated with the number of affiliation's conference attendants

Covariate	GROFFER	CCSIZSET	INSTSIZE	CBSATYPE
p-value	4.48e-10	0.81	0	6.52e-04

We can see that INSTSIZE, GROFFER, and CBSATYPE have p-values less than 0.001. Therefore, we can reject the null hypotheses H1, H2, and H4 at 1% significance level. Put differently, the factors are significantly correlated with the number of attendants in conference. On the other hand, CCSIZSET has p-value of 0.81 greater than 0.1. Therefore, we does not have enough evidence to suggest any correlation between the affiliation size and the number of attendants. Furthermore, as a robustness check, we make independent variable to be attendance for each school each year and we have the same findings.

3. Conclusion

In this project stage, we did some OLAP style analysis and a statistical analysis. We grouped five different attributes and checked how these attributes impact the attendance to conference. In the process, we got familiar with the concept of OLAP.

Furthermore, we did a multivariate linear regression analysis to determine which attributes are influential to conference attendance statistically. We discovered that size of the institution, whether the institution offers graduate degree and affiliation CBSA type are highly correlated with conference attendance. In the future, we can examine other interesting covariates to discover what contributes to conference attendance.