Final Report

Analyzing and Predict Main Industrial Distribution in Massachusetts

Names:
Shuhan Liu: U09073975
Lei Yang: U06367963

Abstract:

For this project, we decide to collect information of various jobs provided by indeed.com, and filter out jobs located in Massachusetts. Then, we will use various cluster techniques to cluster out information based on job categories and locations to get the industrial distribution in Massachusetts.

After that, we will use different regression methods to find out what is the main factor that influence the rating of a job. Therefore, when a user gives us a particular kinds of jobs he/she is looking for, this program can recommend a good place to find such a job. When a user gives us a job post he is interested in but with very few reviews, we can predict whether it is a good job or not.

Datasets:

Here is a short sample of data after cleaned up (The location of MA is dynamic sorted in the program, so in the external csv, it still has jobs with locations out of MA)

	A	B	C	D	E	F	G	H
1	Company	Title	Location	Rating	Work/Life Balance	Benefit	Security	Culture
2	Accion-Interr	Communicat	Boston	4	none	5	none	5
3	Accion-Interr	Systems Adm	Washington	4	5	none	5	none
4	Accion-Interr	Communicat	Boston	4	none	none	5	none
5	Accion-Interr	Systems Adm	Washington	4	none	5	none	5
6	Advance-Dig	Assistant Cor	Jersey City	4	none	5	5	none
7	Advance-Dig	Director of Ir	Springfield	2	5	none	none	5
8	Advance-Dig	Sales Rep	Morristown	1	4	none	none	5
9	Advance-Dig	Quality Assu	New York	5	none	4	none	4
10	Advance-Dig	Non-manage	Jersey City	3	5	none	none	none
11	Advance-Dig	Marketing D	Jersey City	4	none	5	4	none
12	Advance-Dig	Assistant Cor	Jersey City	4	5	none	none	none
13	Advance-Dig	Director of Ir	Springfield	2	none	none	5	none
14	Advance-Dig	Sales Rep	Morristown	1	none	none	5	none
15	Advance-Dig	Quality Assu	New York	5	none	5	none	4
16	Advance-Dig	Non-manage	Jersey City	3	4	4	none	none
17	Advance-Dig	Marketing D	Jersey City	4	none	none	5	5

The dataset is collected from a job search website: www.indeed.com. When you run the web scrap part of this program, this dataset will be generated automatically under the “Company” folder. It contains the information of nationwide jobs posted on indeed.com from 2014 till now. For each job, this program will record its job title, company name, ratings, and the

score for work/life balance, benefit, security and culture. For the missing attributes, the program will record that field as “none”.

Since this program focusing on jobs in Massachusetts, before using the dataset, our program will filter out jobs based on their locations. We manually assigned some cities to “Greater Boston Area” for a better visualization.

Methods:

Join and Simplify the data --- The raw data from the Internet are stored in separate files and have many jobs from places other than MA. Thus the program will parse through all csv files and only select information we needed, like company names, job titles, locations, ratings and attribute scores, and then store them into a “Pandas DataFrame” for later invoke.

Vectorize --- The data gets previously does not only contain digits, but also strings, therefore, before clustering, it needs to be vectorized. For this part, we use “TfidfVectorizor” from “sklearn” library. During vectorizing, different sections will be given appropriate weight In order to have good visualization in following steps.

K-means and Hierarchy Clustering --- To identify different industry groups (for example, software and sales are different industries group), we use K-means++, hierarchy clustering techniques and compared the result of three sets and analyze which one is the best fit for our experiment. The result of clustering should provide us the distribution of industries in Massachusetts.

Regression Model --- To decide what are the main factors that decide job ratings, this program uses regression to make predictions. In this final version, we have the overall ratings for every job and with other six attributes: category, location, work/life balance, benefit, security and culture.

To vectorize job titles and locations into serialized numbers, we will split jobs into 10 hot categories based on keywords in job titles. Then, we will label locations with different numbers. After that, we could do the linear regression on it.

Moreover, this program will transfer ratings into binary. If rating is greater or equal than four, it will be considered as good and labeled as 1, otherwise it will be labeled as 0. Then, by adding same information used in linear regression, this program can manipulate a logic regression.

Additionally, since some attributes has missing values, like some scores for “work/life balance” section. To fill these missing entries, since each job may only appear once or twice, which makes calculating the average score of the job not meaningful, we decide to fill them with their companies’ average score in that field instead of a job’s average in order to make the data more meaningful.

After using both regression techniques, the experiment conducts a conclusion that, in both cases, benefit is the most influential factor among all these six attributes, which means benefit, or in other word, salary, has the best priority.

Analysis and Results:

Clarification 1: Here we select 7 cities in Massachusetts and 1 area *Greater Boston Area" which has 6 extra cities. We will find out the leading industries of these cities.

```
MA_location = {"Greater_Boston_Area" : 0,
         "Salem" : 0,
         "Plymouth" : 0,
         "Waltham" : 0,
         "Framingham" : 0,
         "Worcester" : 0,
         "Lexington" : 0,
         "Danvers" : 0,
        }
Greater_Boston_Area = {"Boston" : 0,
                       "Providence" : 0,
                       "Lowell" : 0,
                       "Cambridge" : 0,
                       "Quincy" : 0,
                       "Newton" : 0
                      }
```

As a result of steps introduced above, we get clusters based on the location in Massachusetts.

Here is a short sample of K-Means++

Cluster 0:
patient_care_associate
greater_boston_area
critical_care_technician

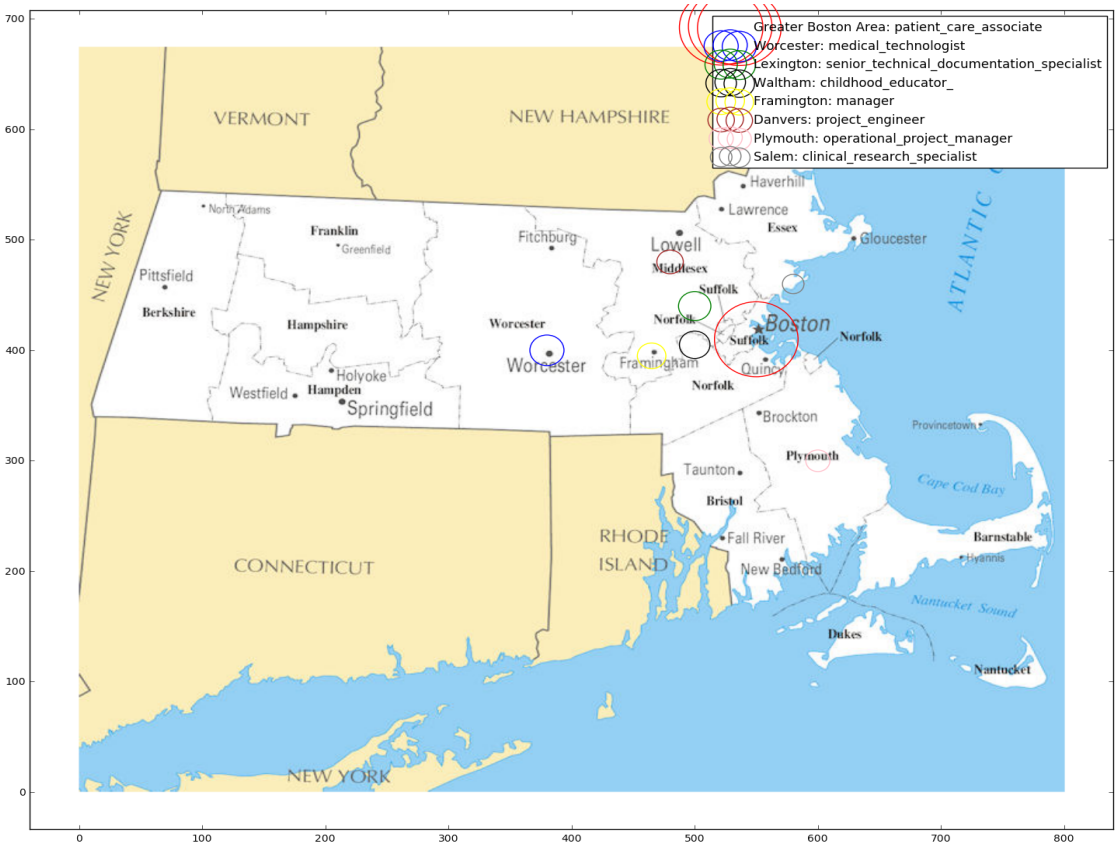
Cluster 1:
worcester
medical_technologist
administrative_assistant
executive_administrative_assistant

Cluster 2:
lexington
security_guard senior_technical_documentation_specialist

The result of K-Means++ is much better than Hierarchy (in Hierarchy, the most central attributes in cluster 0 is simply locations without job information). It's probably because K-Means++ is a hard cluster and more fit to this experiment.

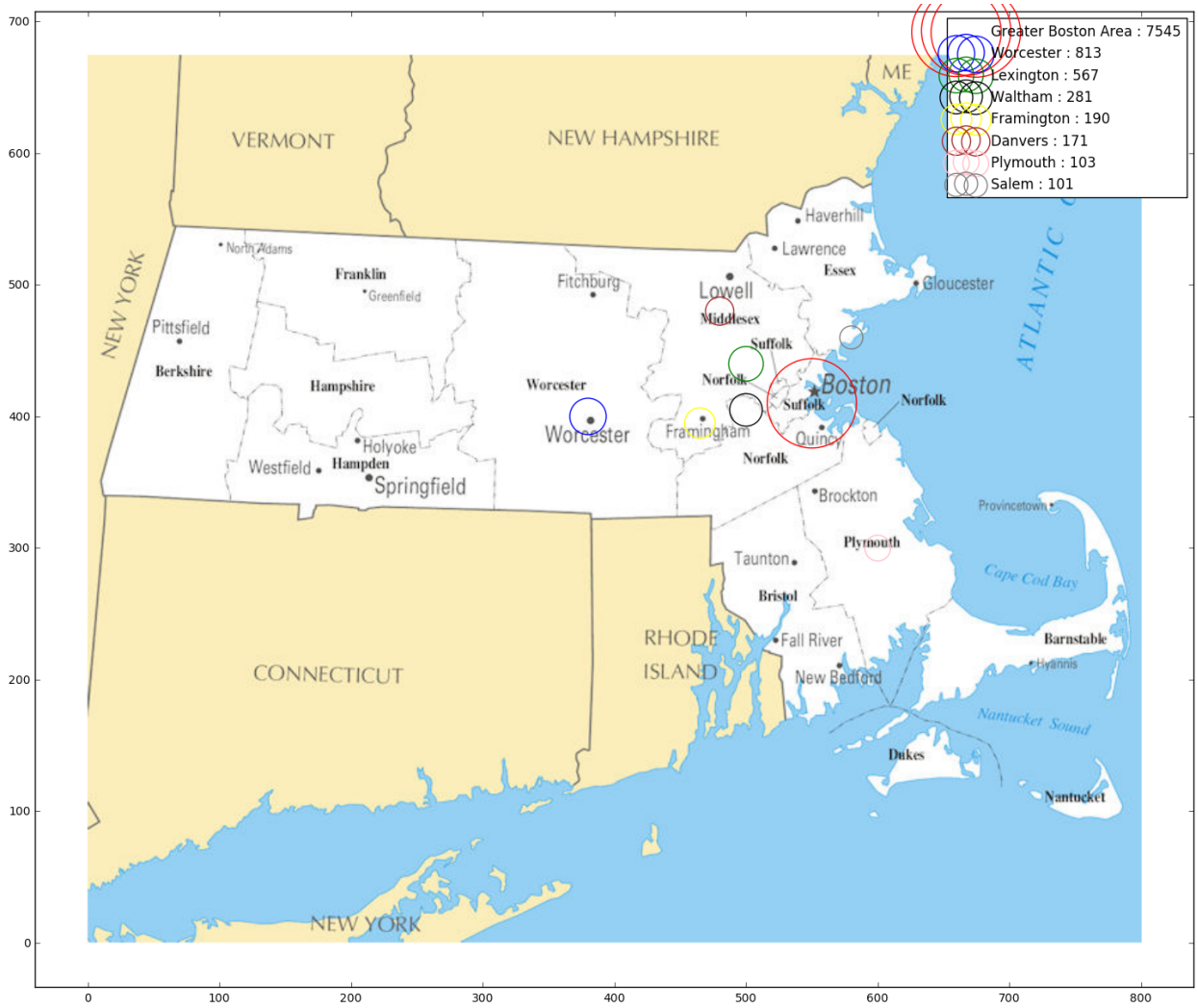
From the result of K-Means++, we can see in each cluster it has one city, which represents the area of this location, also meets our original assumption. Since we used Tfidf and the number of jobs is descending of each cities, we can also cluster 0 represents greater boston area, cluster 1 represents Worcester… Which also meets our assumption. Now we can use the result of K-Means++ and visualize the industry distribution of Massachusetts.

1. Result1: Visualize the leading industry in Massachusetts of 8 places



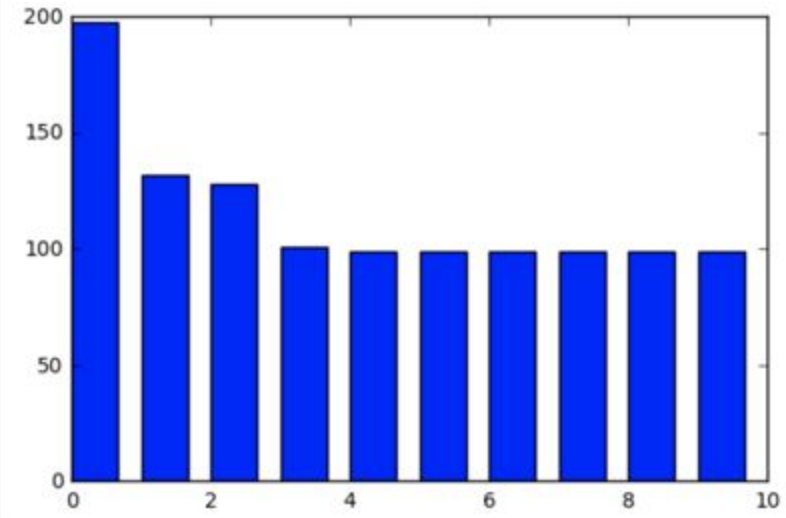
Here we can see the leading industries of each location in Massachusetts, for example, in Greater Boston Area, the leading industry is patient care associate. The size of circle represents how many jobs this area has. In the above map we can see the circle of Greater Boston Area is largest.

2. Result 2: Visualize the leading industries in Greater Boston Area



Here we can see in Greater Boston Area, there are over 7,000 jobs, almost ten times over the second Worcester (813). So we also use count vectorize to get the leading industries in Greater Boston Area.

```
Senior_Inventory_Accounting_Specialist
Patient_Services_Specialist
Treasury_Solutions_Advisor
Project_Engineer
Financial_Advisor
Brand_Ambassador
INSIDE_SALES_REP
Claims_Processor
Laboratory_Coordinator
Sales_Associate/Cashier
```



The above bar chart shows 10 leading industries in Greater Boston Area. Since we focuses on one area now, we use Count Vectorizer instead of Tfidf Vectorizer. We can see in Greater Boston Area, the top popular career is senior inventory accounting specialist.

Moreover, with both regression techniques, we find location is a more influential factor to decide ratings rather than job category. In the final report, we will include more details and more attributes, and then do the regression again to get a more precise prediction.

Result of Linear Regression

OLS Regression Results

Dep. Variable:	Rating	R-squared:	0.928
Model:	OLS	Adj. R-squared:	0.928
Method:	Least Squares	F-statistic:	2.110e+04
Date:	Tue, 13 Dec 2016	Prob (F-statistic):	0.00
Time:	16:23:54	Log-Likelihood:	-14932.
No. Observations:	9771	AIC:	2.988e+04
Df Residuals:	9765	BIC:	2.992e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Category	0.0467	0.005	9.394	0.000	0.037	0.056
Location	0.0588	0.005	11.750	0.000	0.049	0.069
Work/Life Balance	-0.0128	0.015	-0.837	0.403	-0.043	0.017
Benefit	0.4001	0.014	28.893	0.000	0.373	0.427
Security	0.2859	0.016	18.161	0.000	0.255	0.317
Culture	0.2052	0.016	12.484	0.000	0.173	0.237

Omnibus:	417.950	Durbin-Watson:	1.995
Prob(Omnibus):	0.000	Jarque-Bera (JB):	490.096
Skew:	-0.491	Prob(JB):	3.77e-107
Kurtosis:	3.489	Cond. No.	15.3

Result of Logic Regression

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
Category	0.0494	0.011	4.570	0.000	0.028 0.071
Location	0.0042	0.010	0.420	0.675	-0.016 0.024
Work/Life Balance	-0.0243	0.031	-0.786	0.432	-0.085 0.036
Benefit	0.1000	0.027	3.659	0.000	0.046 0.154
Security	0.0479	0.031	1.535	0.125	-0.013 0.109
Culture	0.0999	0.033	3.046	0.002	0.036 0.164

From the results of both linear and logic regression techniques, it's not hard to find that "benefit" is the most influential factor among all these six attributes. If we take a closer look at the confidence intervals of these attributes, we can observe that in linear regression, "work/life

balance” is considered as an irrelevant factor, while in logic regression, “location” and “security” becomes not important as well.