**Data Scientist Job Data Analysis**

**IEOR E4523 Data Analytics**

**Professor Uday Menon**

**Tofu and Chili**

**Yuhao Chen(yc3982)**

**Tianhao Han(th2939)**

**Jingyi He(jh4445))**

**Xiaoling Pang(xp2185)**

**Janey Yang(gy2300)**

# **1. Abstract**

Amidst the pandemic and fast development of tech companies, there is an increasing demand for data scientists. Under this circumstance, we would like to carry out this project for a deep understanding towards seeking and preparing for data-related jobs. In this project, we aim to answer a series of questions mainly centered at the following three topics: What are differences in roles among data scientists, data engineers, and data analysts? What skills are the most useful when preparing for data-related roles? In terms of salary, which cities and what types of companies are suggested to pursue data-related jobs? Thus, we gather job posting data from glassdoor to perform the following approaches: Unsupervised learning on how to cluster the data to gain more insight on job hunting; Supervised learning on how to predict the salary with other data as features and Text mining on Job description to see if there exists some pattern between salary and texts. Through the analysis, many useful insights are extracted. To give one piece of advice to young data-related job-seekers, we would suggest companies with one or more of the following features: located in CA, within the IT industry, being privately owned, and small to medium size.

# **2.Data Description and Basic EDA**

The data is web scrape data scientist job items from Glassdoor.com, acquired through Kaggle. The dataset contains the three types of data with the following 14 columns: job information, job description, and company information. Details can be seen in (Appendix 2a).

As for exploratory data analysis, firstly, we take a look at the location and salary (Appendix 2b). We get geojson data of US states, and map the salary data. We find that CA and NY are two highest salary states.Secondly, we explore the sector distribution (Appendix 2c). IT, business services, Biotech are top distributed in data scientist jobs. Finally, we discover an interesting fact about size and salary (Appendix 2d). The median salary of DS jobs in companies with size around 5001 to 10000 employees is the lowest.

# 3.Data Preprocessing

**3.1 Data preprocessing for Supervised learning**

The majority of the variables in the original dataset are categorical, even the salary, which represents a range of salaries. As a result, if we wish to use linear regression models to generate numerical salary forecasts, we must first convert that column into numerical. We selected to run three times uniformly depending on the salary range and pick the average of the simulations. Then we discovered that the elements in job title column are distinct but comparable: so we we split them into 8 types: 'Job Title\_Data Engineer','Job Title\_Data Scientist', 'Job Title\_Junior', 'Job Title\_Manager','Job Title\_Quant', 'Job Title\_Senior', 'Job Title\_others'. In addition, we included certain columns at the end, such as age, which is dependent on the column year the company was created. If a skill appears in the job description section of each row, the value of that specific skill column becomes 1, otherwise 0. Finally, we created a correlation matrix by converting all category features into dummy variables and removing features with correlations greater than 0.85 to eliminate collinearity(Appendix 3.1a). All of the preceding preparations are for linear models. We changed the columns into categorical ones for the categorization models.

**3.2 Text Cleaning for Job Description**

The original ‘Job Description’ contains a series of sentences and symbols. WIth NLTK, we split the raw text into sentences and then into words. We normalize cases and remove company websites, symbols, and punctuations. Finally, we fill out stopwords and stem words to get a cleaned tokenized text data.

**4. Supervised learning with Linear Models and Classification Models**

**4.1 Linear Models**

In linear models, salary is a numerical and dependent variable.

Linear Regression model: The root mean square error is 34.344078. R-squared is 0.265652 and adjusted r-squared is 0.23737 which is not good.

Lasso: We use the least negative mean absolute error and cross validation to find the best alpha in lasso regression. The alpha is 0.03 in our best lasso model (Appendix 4.1a). The RMSE is 34.100945, R-squared is 0.276013 and adjusted R-squared is 0.24813 in the lasso regression model, which is better than the previous model.

**4.2 Classification Models**

Decision Tree classification: First, we attempted DT with the default parameters. 0.236934 turns out to be the accuracy. The accuracy is 0.353659 after pruning the tree by constraining max depth and max-leaf nodes, which is significantly better than the baseline.

Random Forest classification: We started with the default parameters and found that the accuracy was 0.292683. The optimal values for max depth, max samples, and n estimators were then determined using GridSearchCV.We also utilized cross-validation with 5 folds to train the models due to the shortage of the dataset. In addition, the model's accuracy for the RF with better parameters is 0.355401.

Boosting: Next, we explored the both boosting algorithms XGBoost and Catboost. Also, GridSearchCV was used to choose parameters and perform cross-validation in both models. In addition, for feature engineering, we choose the best 10 features with the highest feature value based on Gini importance (Appendix 4.2a). XGBoost is our final model (Appendix 4.2b), which is slower, but it has the best accuracy, compared to Catboost’s 0.353659

For our best model, we identify Location, Job Title, and Sectors as most important factors based on Gini. As a result, they are the most significant factor to consider when looking for work.

# 5.Text Mining

**5.1 Word Frequency Analysis**

To have an intuitive view, we generated four word clouds (5.1a) using the job descriptions of all job titles, data scientists, data engineers, and data analysts, respectively.Many of the words such as “experience” “team” “analysis” are addressed in all four pictures. Then, we did text cleaning such as removing stop-words and lemmatization using the NLTK package, and counted word frequencies. From 5.1b we can see the frequent occurrence of “data” and “experience” is universal and significant. Thus, we decided to remove these two words in the following analysis to have a clear view of other possible frequent words.

The following figures (5.1c-5.1e) show the results of the top frequency words for data scientists, data engineers, and data analysts, respectively. What we have discovered is the job descriptions were similar regarding all three data-related roles. However, some nuances do exist. The job descriptions of data scientists emphasize on the “science” part, while data engineers mention more about specific tools such as SQL and data analysts focus most on the “business” sense.

**5.2 Latent Dirichlet Allocation**

We also applied Latent Dirichlet Allocation (LDA), trying to see if we could extract three main topics of the text. However, the three topics we got were not necessarily corresponding to the previous three job titles. The results of LDA for three topics are shown in Appendix 5.2a-c. Our lesson is that LDA has some limitations for the type of text. Due to innate characteristics of job descriptions for each role, the topics we extracted might not be significant. Applying topic modeling in other contexts such as novels or news might yield better results.

**5.3 K-means Clustering and POS tagging**

We vectorize the words and conduct a part-of-speech tagging analysis, where we categorize each word into word classes or lexical categories. We find that nouns and adjectives are the two most common word types in the job description, which is within our expectation. After removing stop words, we find this trend still holds (appendix 5.3a).

We then perform a k-means clustering on the job description data with TF-IDF word embedding. We choose k = 3 because there are three main job categories: data scientists, data engineers, and data analysts. After the k-means clustering, we count the number of data points in each cluster and visualize the top keywords with highest TF-IDF score (appendix 5.3b). We notice that cluster 0 only has 61 observations, while cluster 1 and 2 have 2532 and 1316, which means that cluster 0 differs from other two clusters a lot. Cluster 0 has two interesting keywords: quantum and IBM. Cluster 1 features some specific skills like machine learning. Cluster 2 has a topic related to research with keywords like laboratory, research and development. We also conduct PCA to reduce the dimension of vectorized words into 2 dimensions and plot a scatter plot (appendix 5.3d), which further confirms that cluster 0 differs from other two clusters in a larger extent regarding TF-IDF. However, as we further examine, different clusters do not have significantly different salaries (appendix 5.3c). Based on the analysis above, we doubt that data scientists, data engineers, and data analysts have similar job descriptions to some extent.

# 6.Unsupervised learning

**6.1 Cluster analysis**

To perform clustering with categorical data and numerical data, the K-prototype algorithm is a good choice. After some tuning, a number of 3 clusters is the best choice (Appendix 6.1b)

We chose 3 numerical data and 3 categorical data to perform clustering. According to figure 6.2c, 6.2d, 6.2e shows, the three categorical variables cluster shows the following characteristics of different groups of companies: Group 1 has a majority in Texas, and is privately owned, has a sector major in Business Service. Group 2 has a majority in California, is mostly private owned, and is hiring IT service most. Group 3 has a majority in TX and CA, a public-owned company, and hires most in the IT and biotech sectors.

According to figure 6.2f, the three numerical variables show the following characteristics of different clusters: Group 1’s feature is big size(10000+ employees), but has a medium performance on rating and salary. Group 2’s features are middle and small size, but with good rating and salary. Group 3’s features are middle and small size companies, but with poor rating and salary.

Finally, we can summarize three groups with different companies tags: Group 1:"Big, nice, but stingy"; Group 2:"Small, nice, and generous"; Group 3:"Public, mean and stingy".

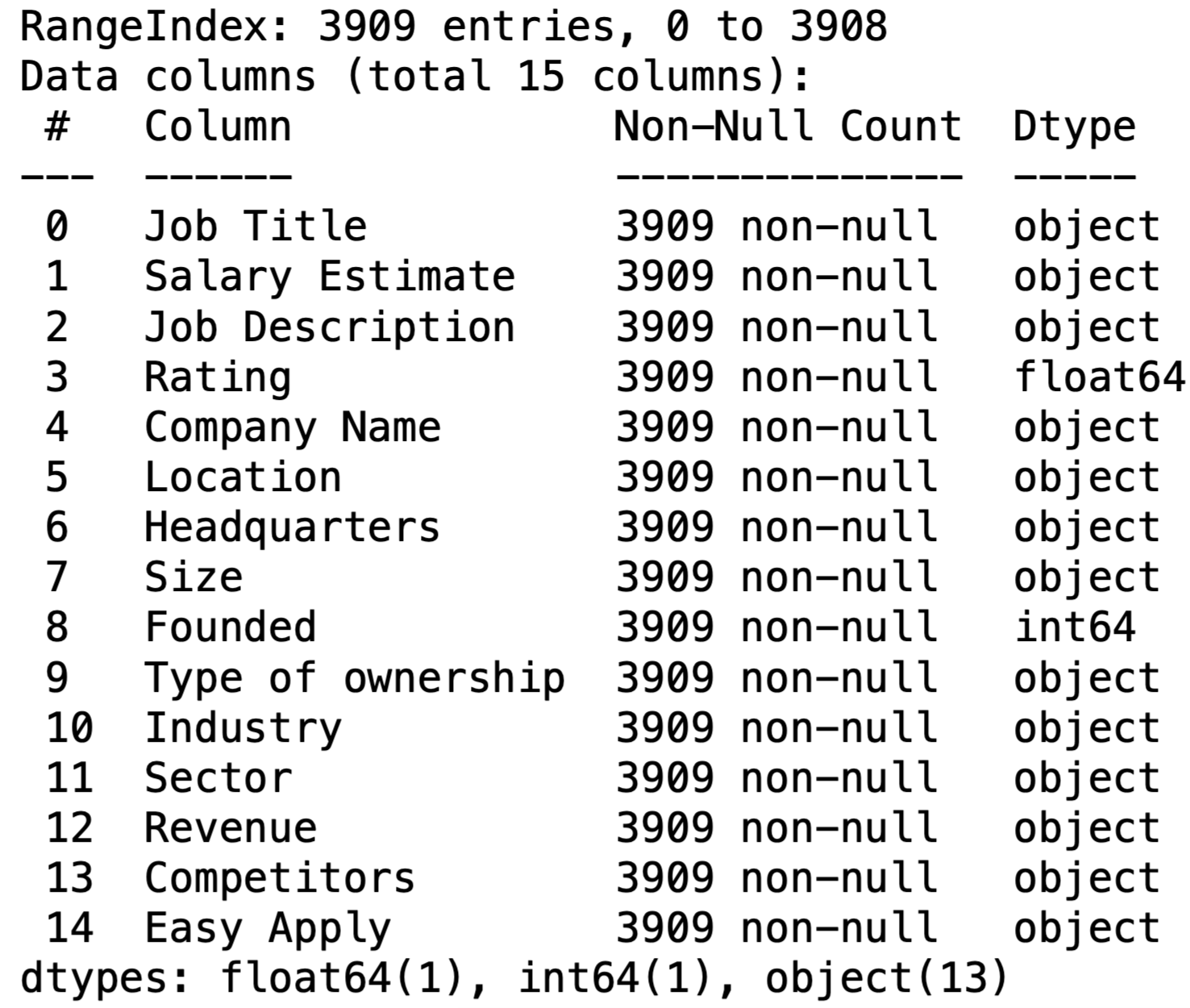
**7. Conclusion and future improvements**

Through cluster analysis, we can answer the question: “Where can we find the best data scientist job?” It should be located in CA, IT industry, privately owned, with medium or small company size.Through a supervised approach, we learned the 3 most important features based on Gini for our best tree model: Location, Job title, and Sectors. Through a text mining approach, we learned that Skills/Tasks across different job titles are quite similar, but different titles showed a slight difference in emphasis, such as Analysts highly emphasize business sense, while Engineers focus more on specific tools, such as SQL.

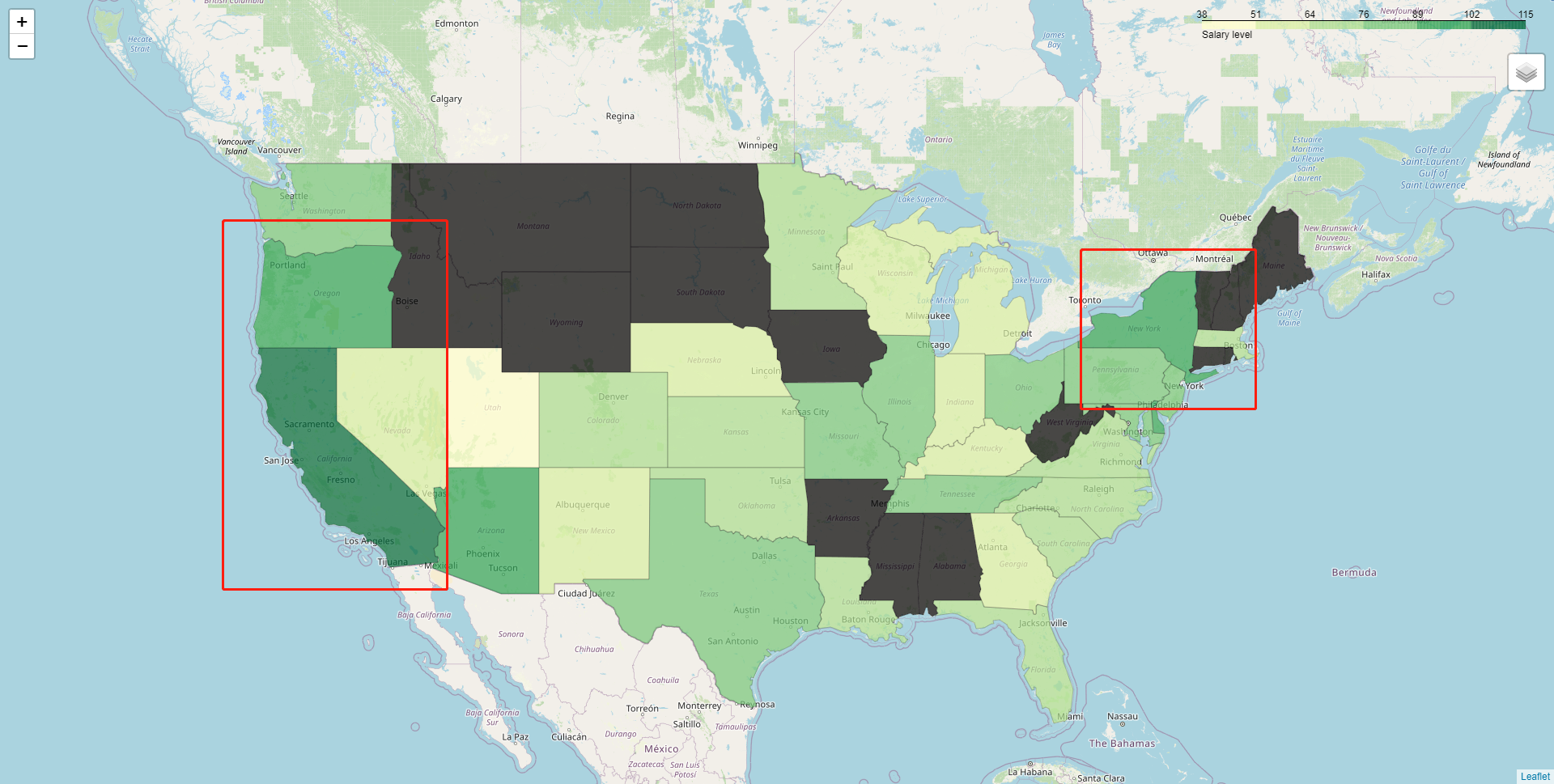
For future improvements, we should use better web scraping tools to gain a better data abundance, plus we can try other word embedding techniques(e.g.word2vec)

# **Appendix**

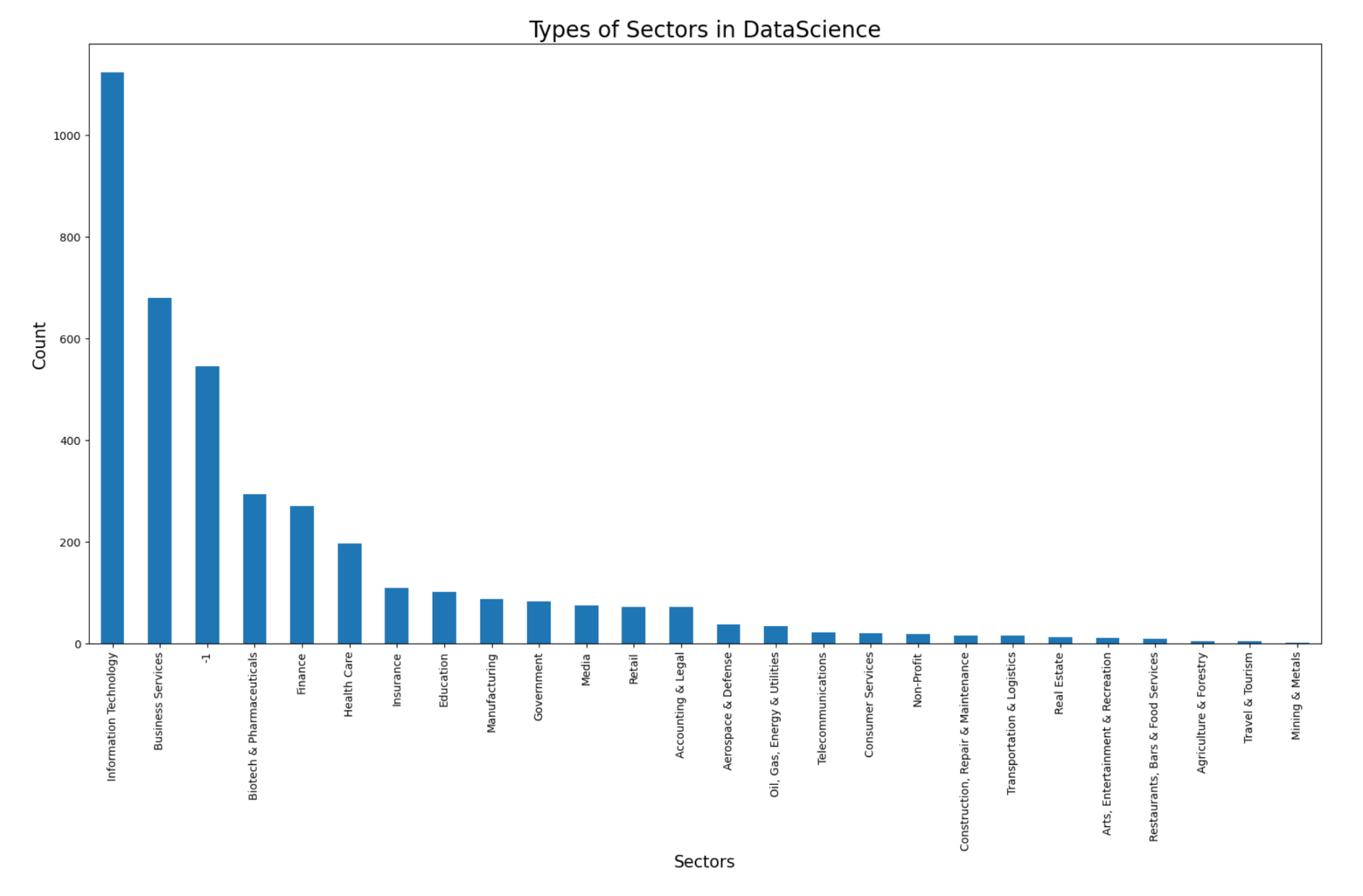
2a Dataset overview



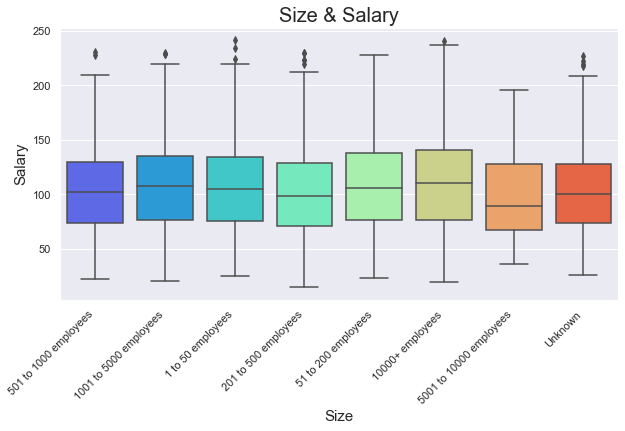
2b Location vs Salary



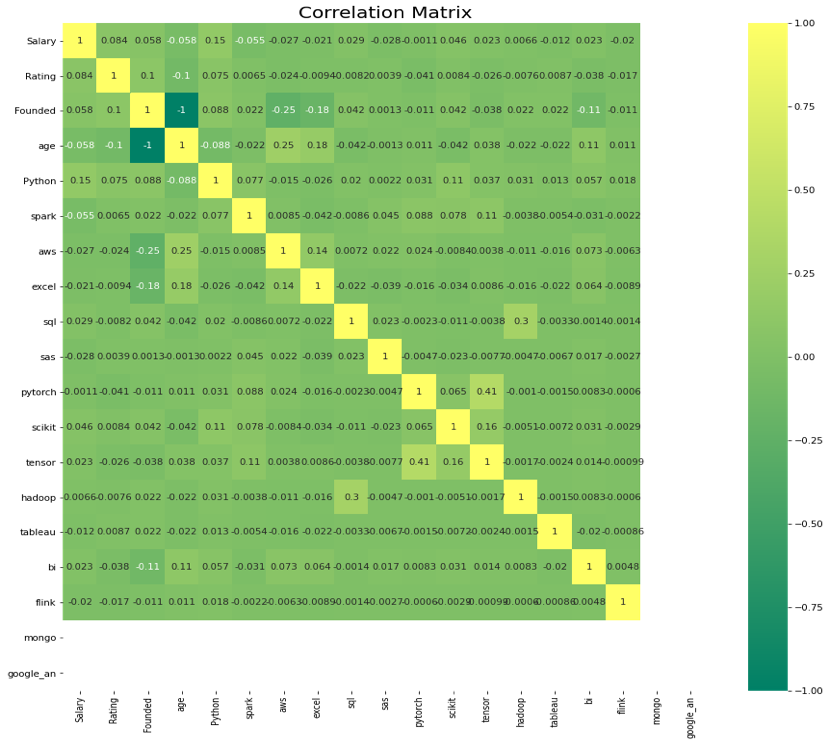
2c Sector Distribution



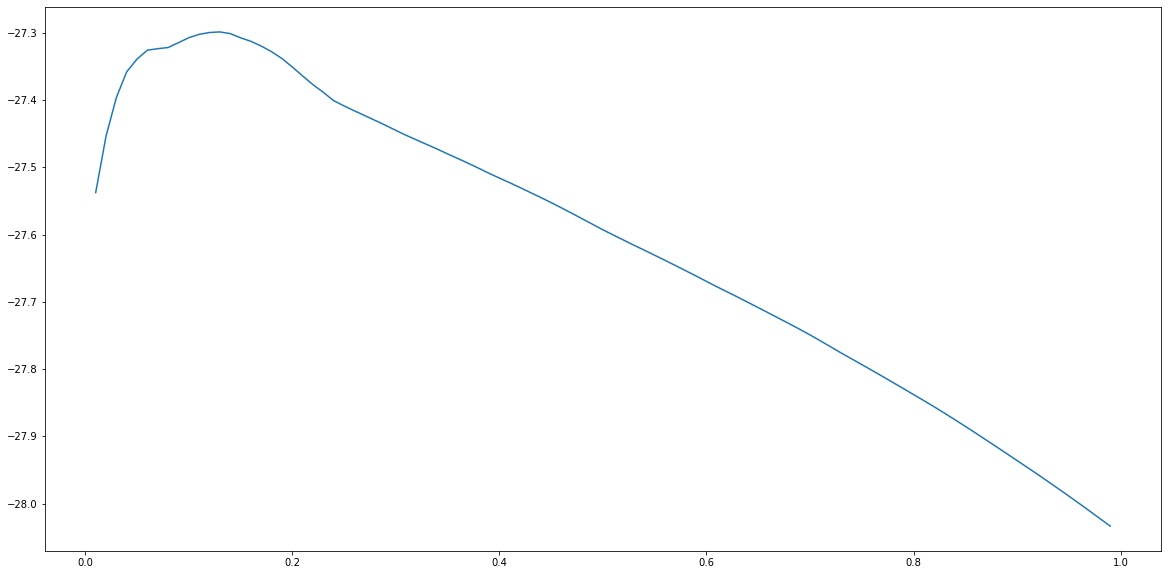
2d Size vs Salary

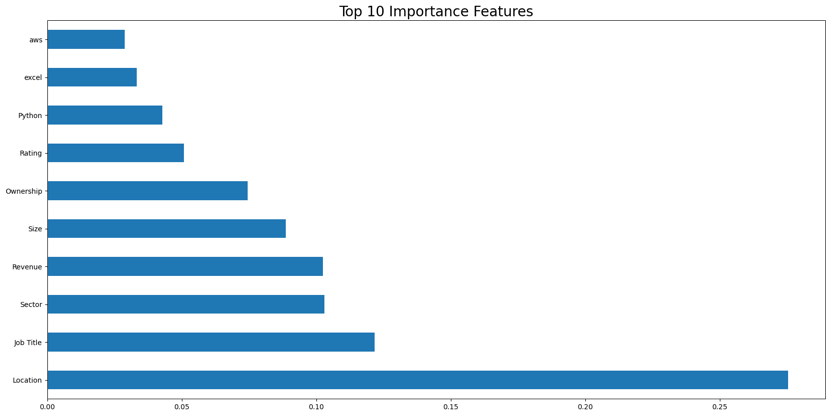


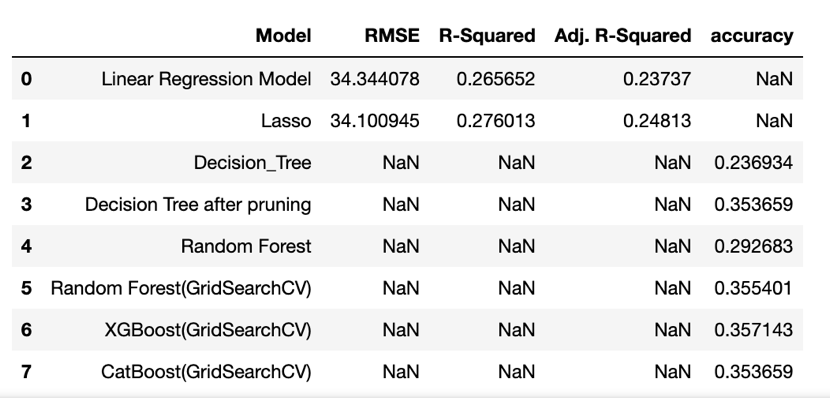
3.1a Correlation matrix for features

****

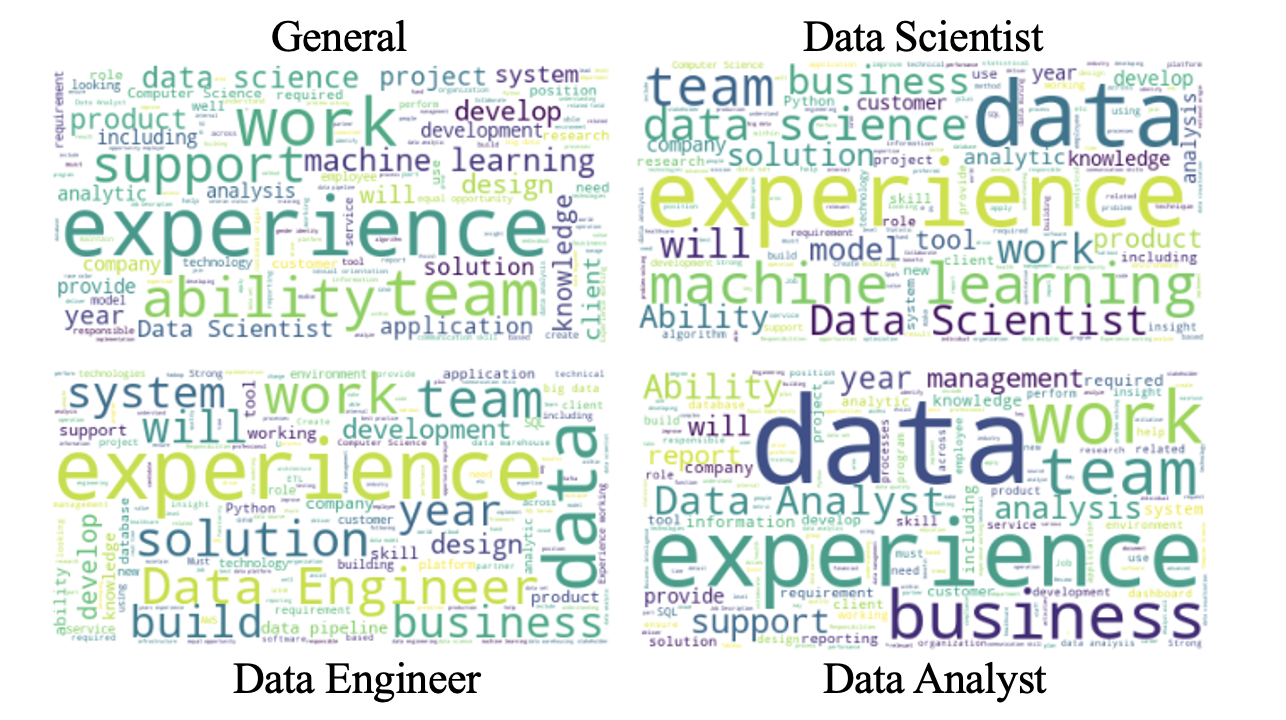
4.1a Alpha and corresponding negative mean absolute error in lasso models

****

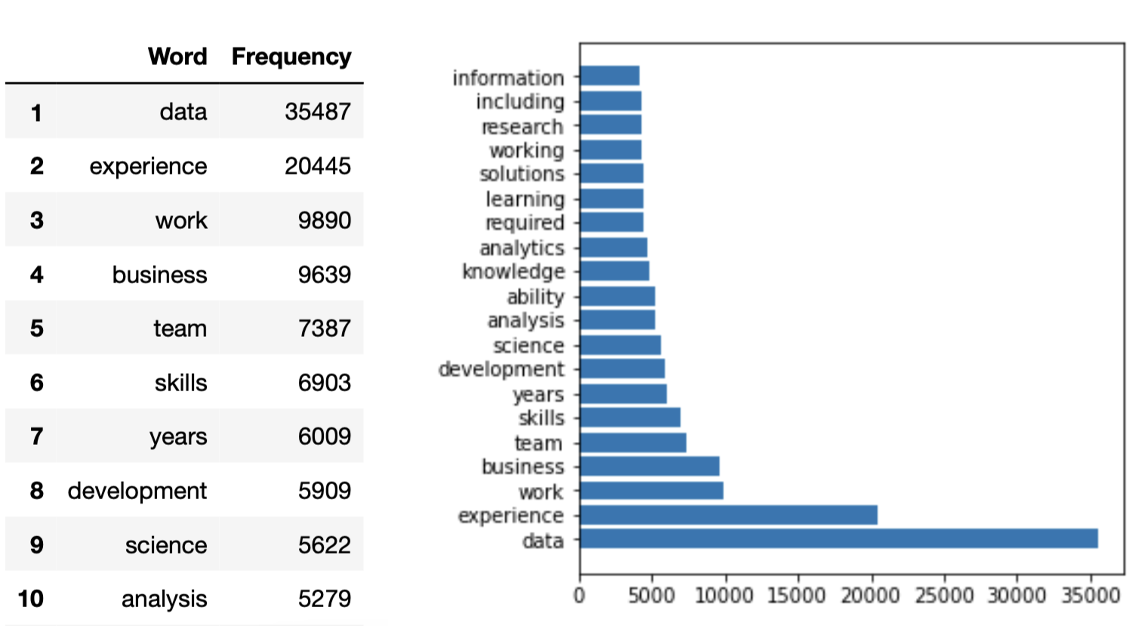
4.2a The best 10 features with the highest feature value based on Gini importance****

4.2b Score matrix for all model****

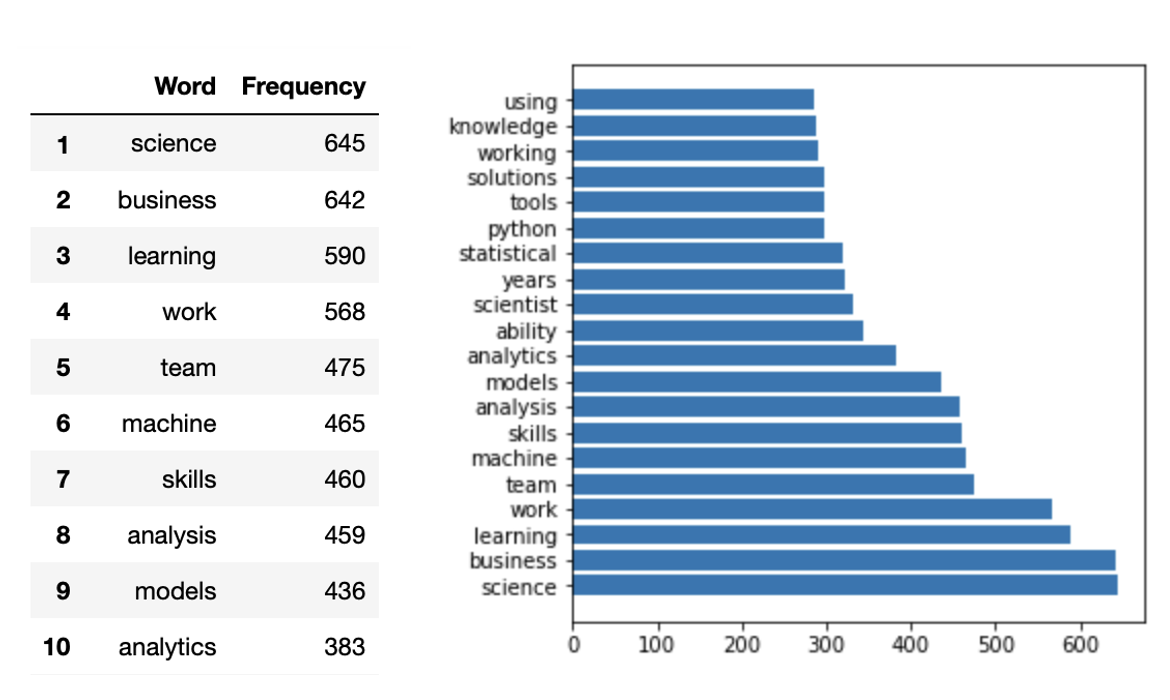
5.1a Word clouds

****

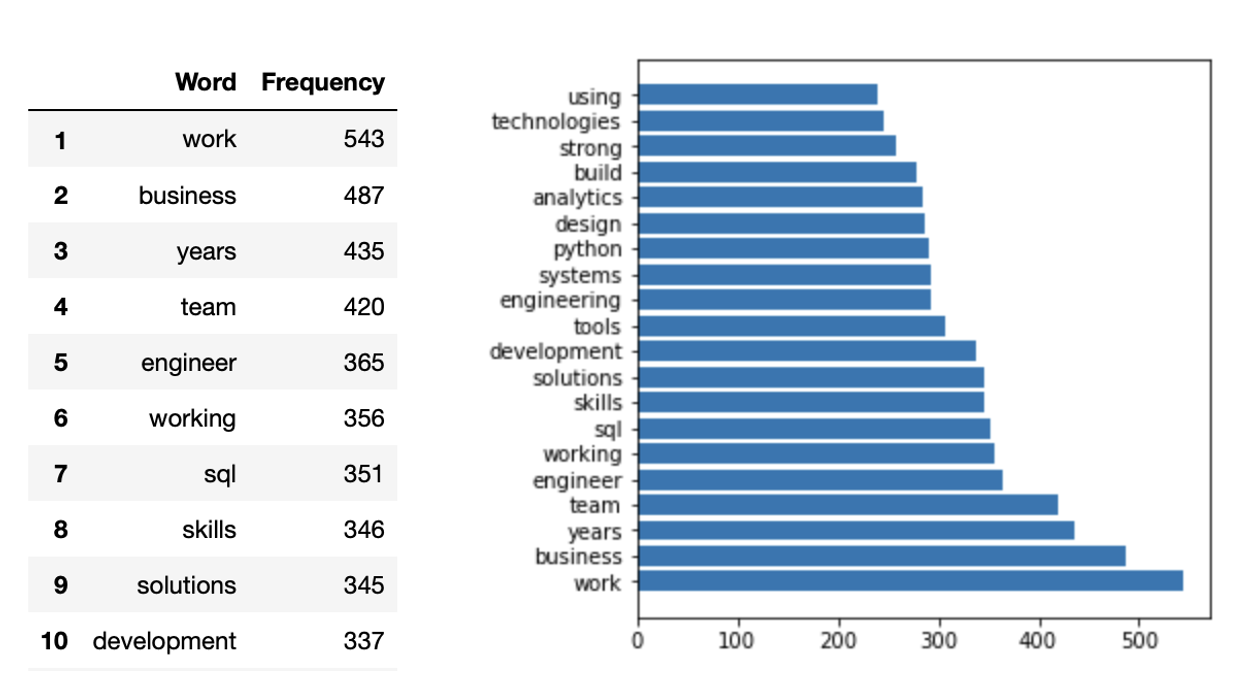
5.1b Word frequency

****

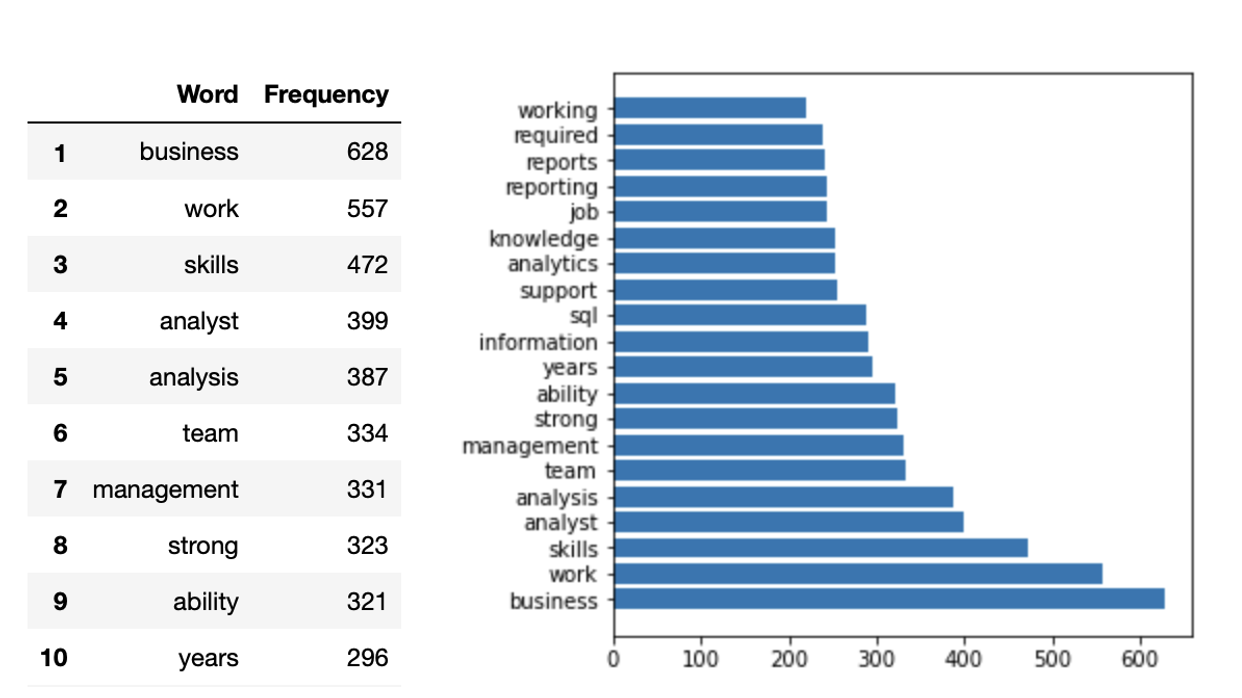
5.1c Word frequency for “data scientist”



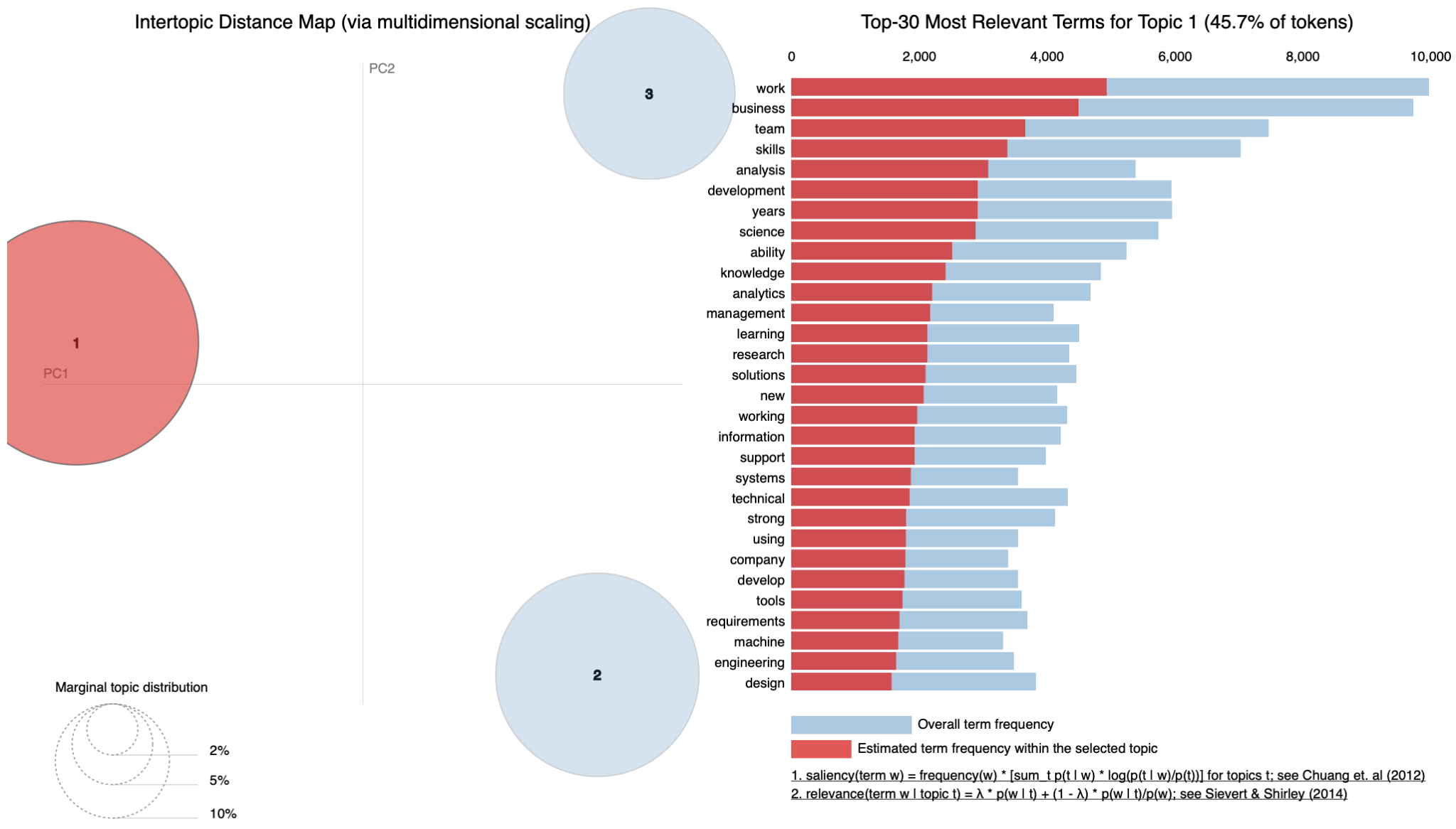
5.1d Word frequency for “data engineer”



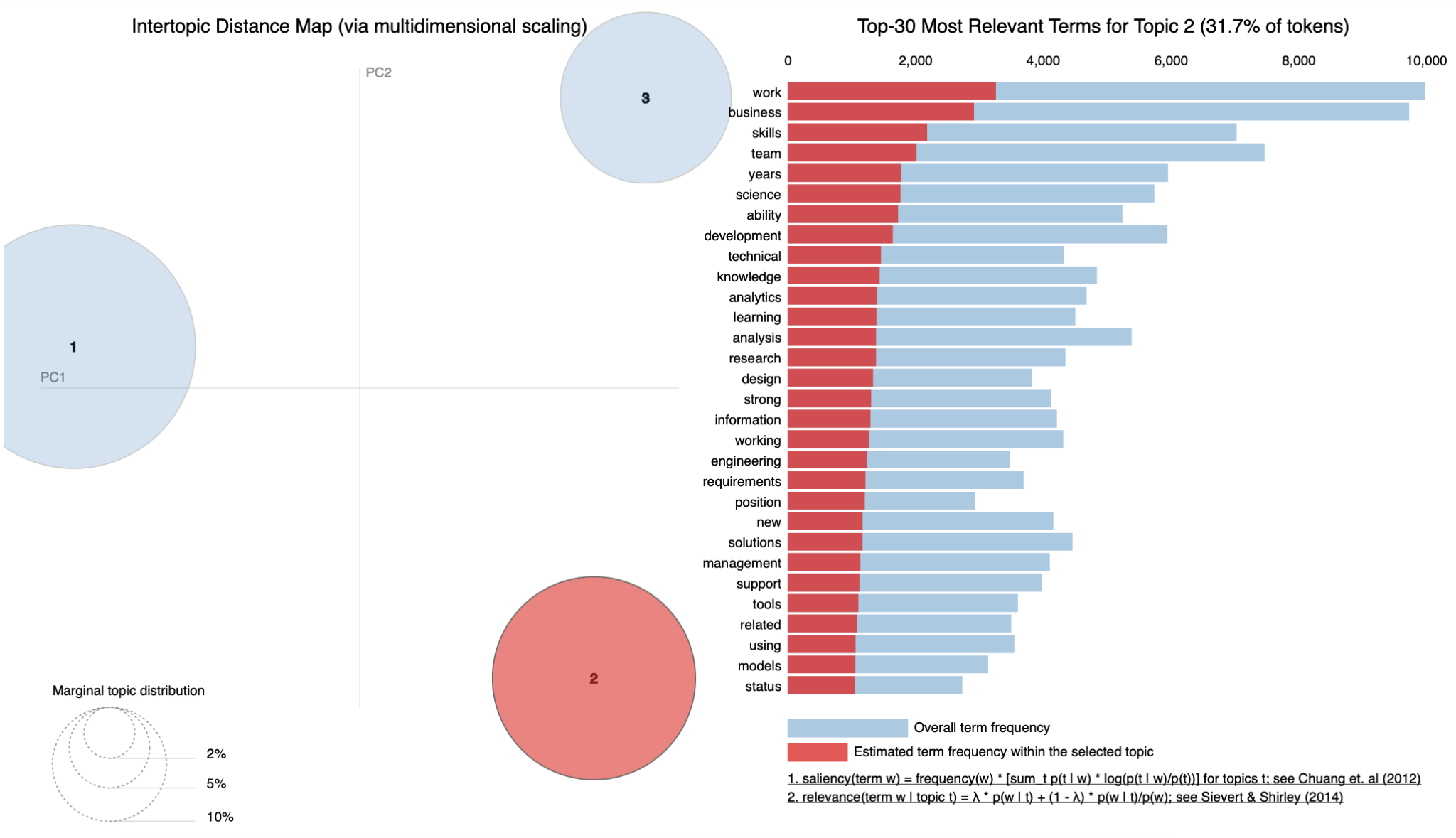
5.1e Word frequency for “data analyst”



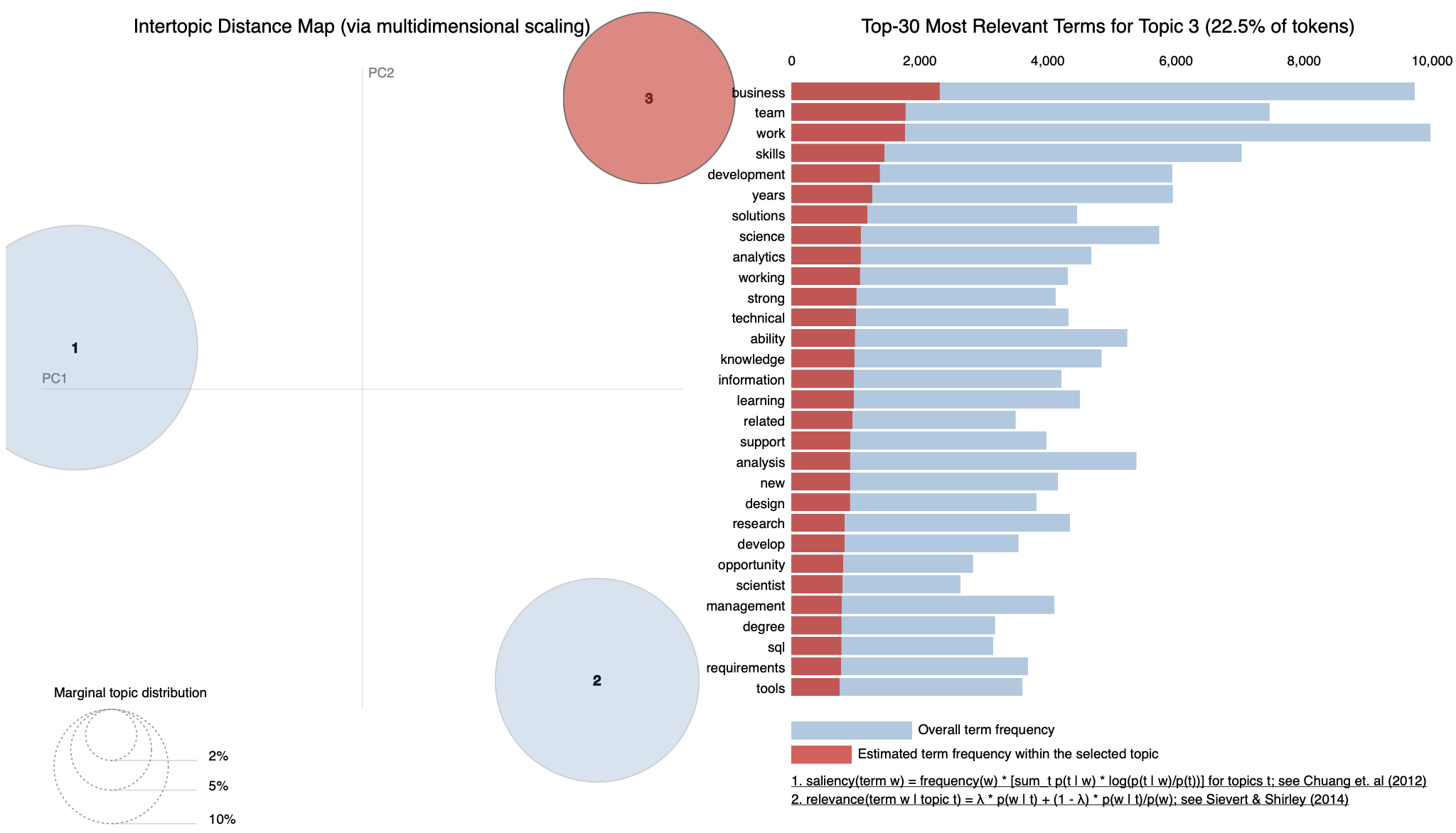
5.2a Topic 1 extracted by LDA



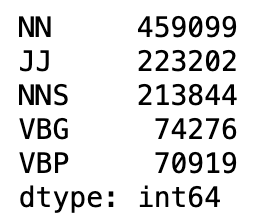
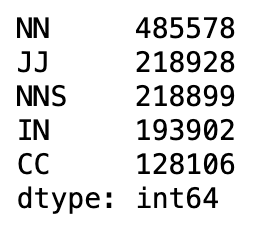
5.2b Topic 2 extracted by LDA

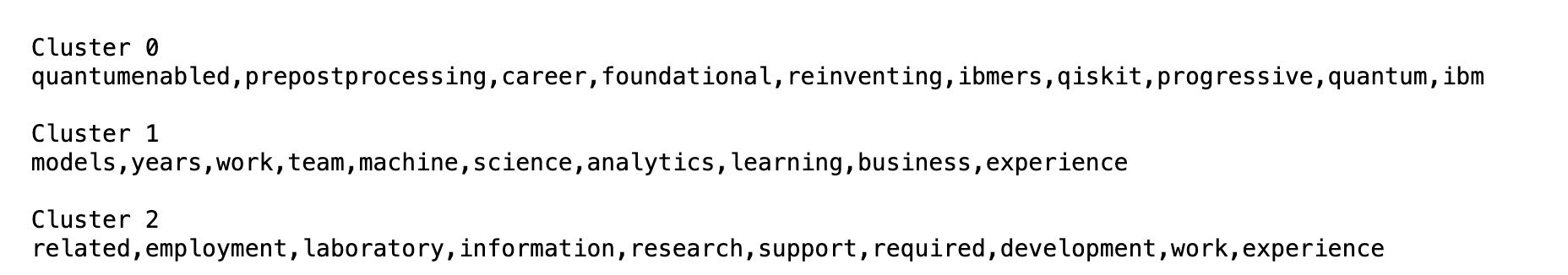


5.2c Topic 3 extracted by LDA

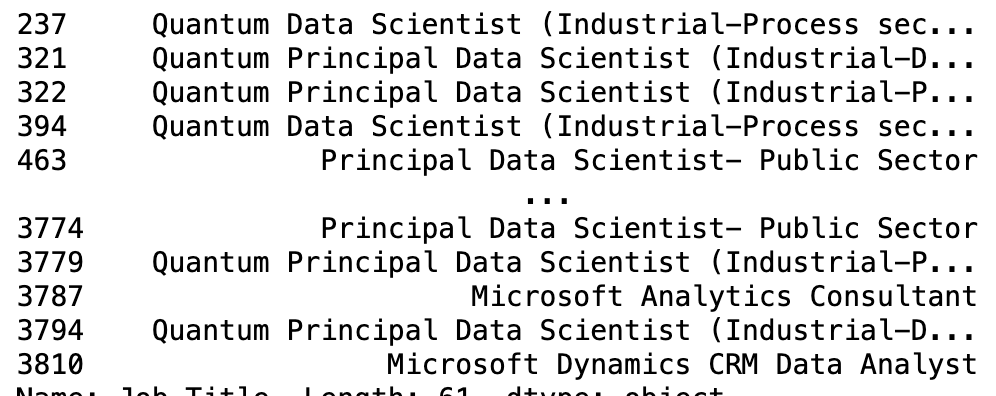


5.3a POS Tagging with (left) and without (right) Stopwords

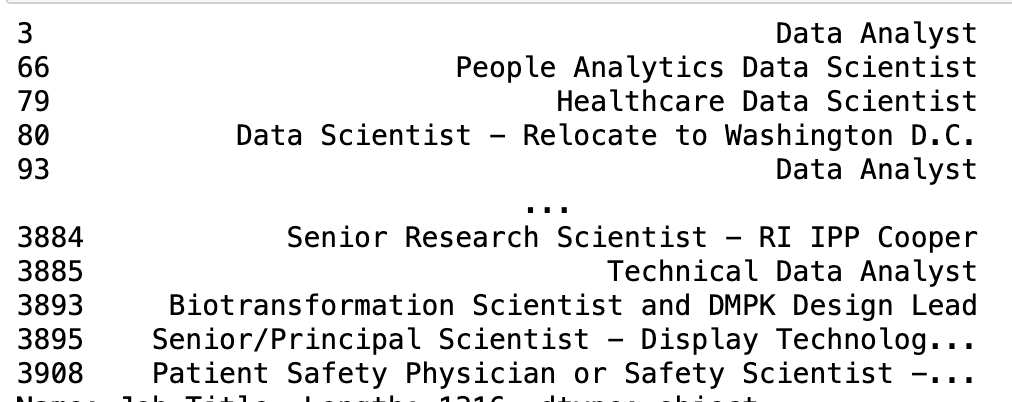


5.3b K-means Clustering Keywords for each cluster 

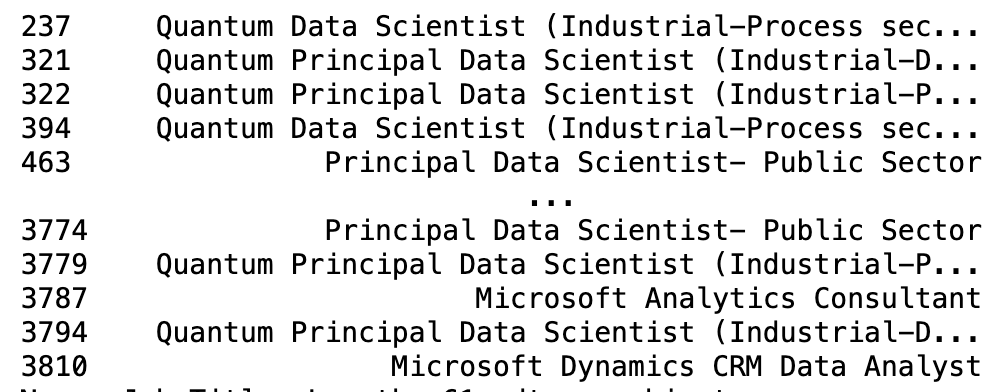
5.3c Job Title for for each cluster



Cluster 0

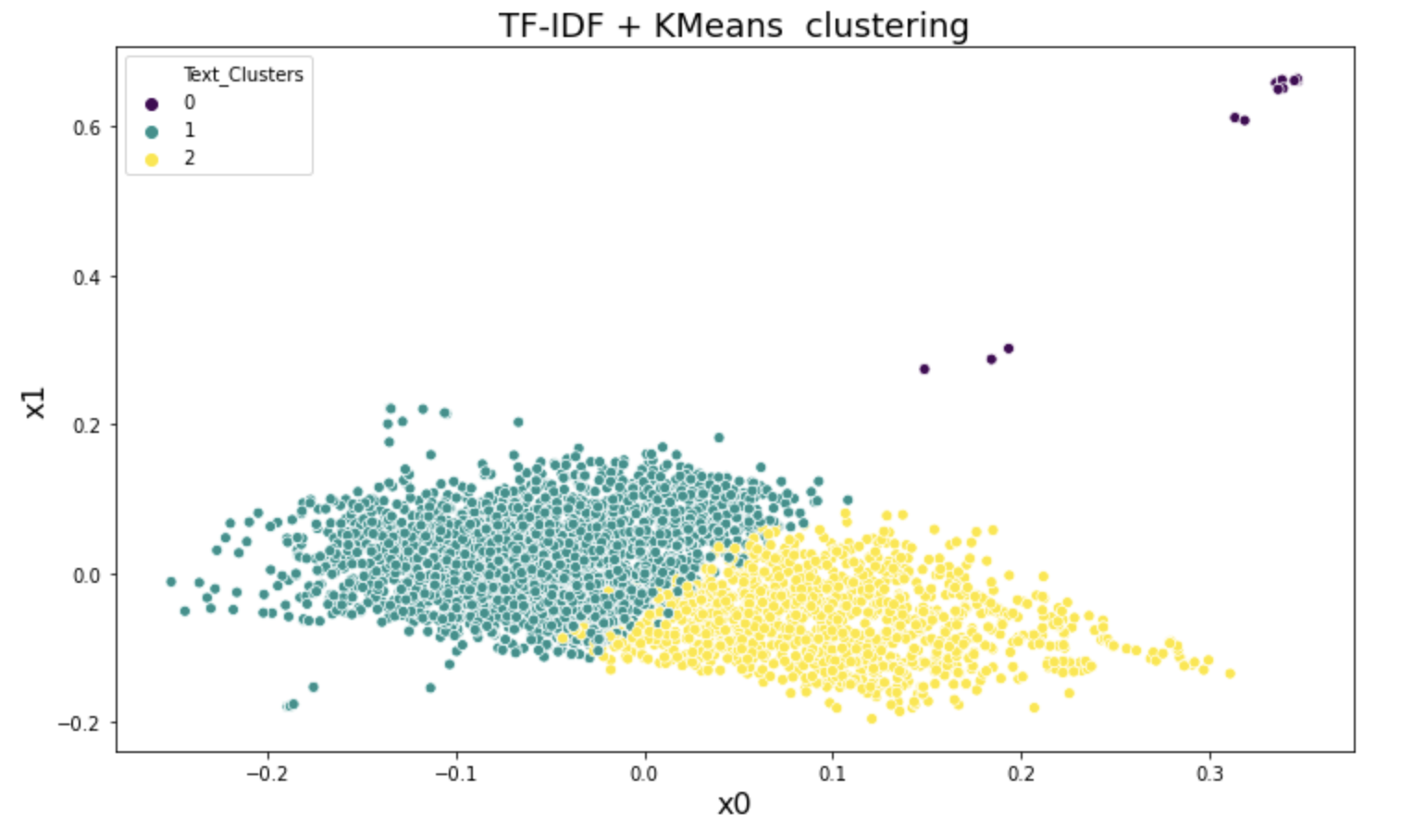


Cluster 1

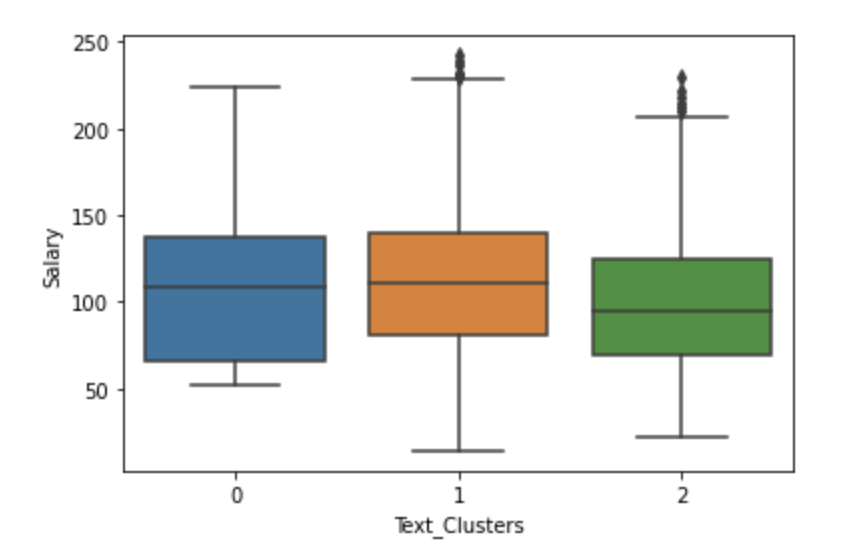


Cluster 2

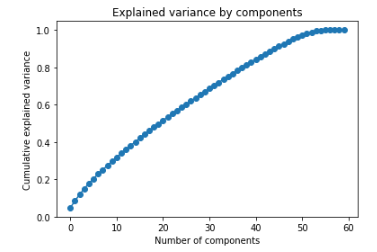
5.3d Scatterplot for clusters after PCA



5.3e Boxplot of Salaries for each cluster



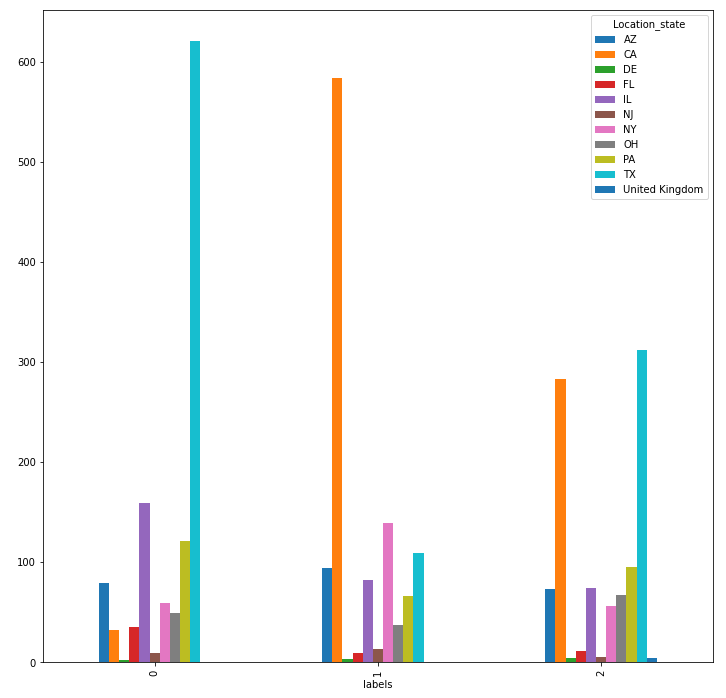
6.1a PCA on One-hot encoding data



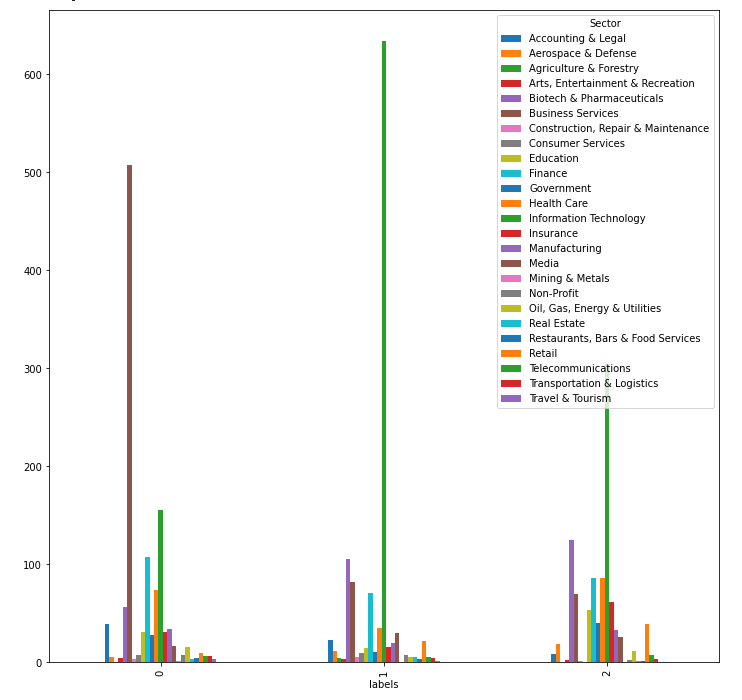
6.1b tuning the K-prototype algorithm



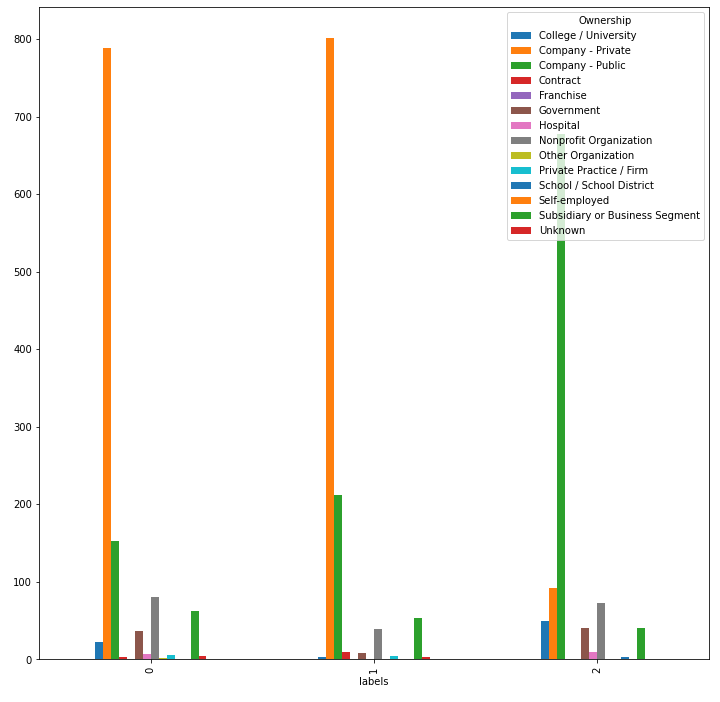
6.1c Cluster upon location\_state



6.1d Cluster upon sector



6.1e Cluster upon ownership



6.1f Cluster upon salary, size, rating

