Data Science Job Postings Analysis

University of Chicago

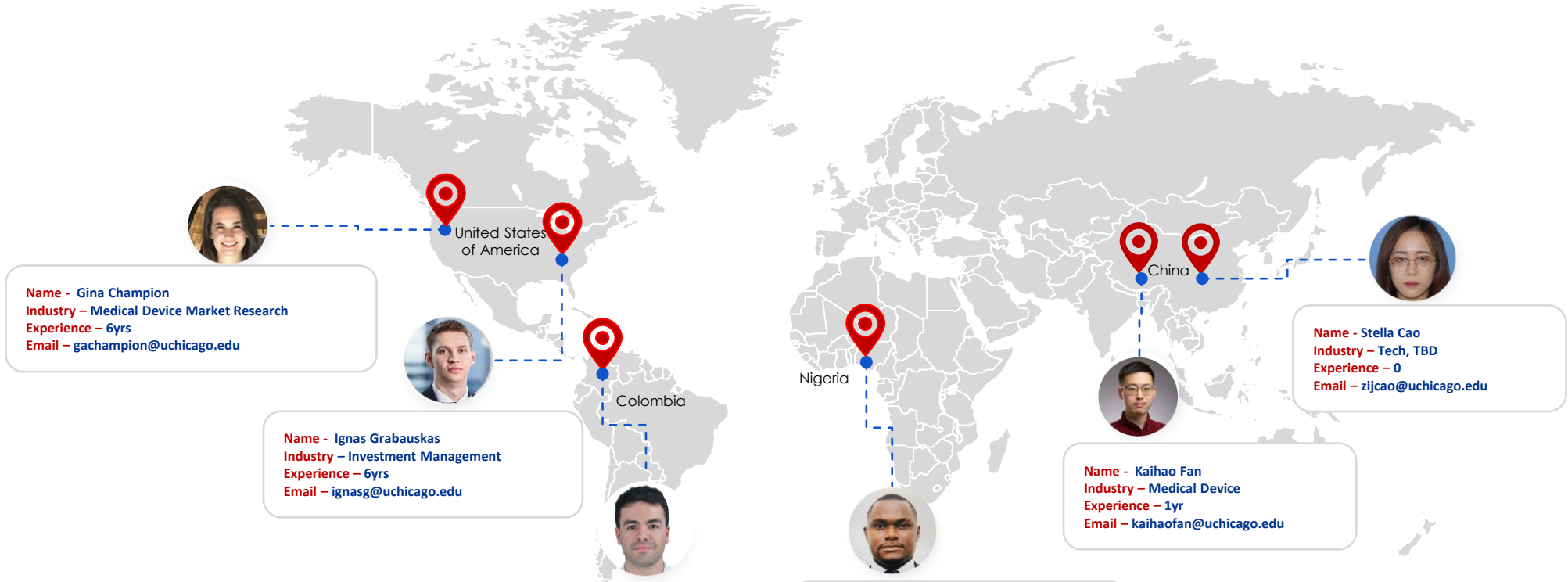
Data Mining Principles – Final Presentation

March 2021

Contents

Executive Summary > Introduction

Our team consisted of interesting individuals with diverse cultures, across different continents, and different time zones …



Name - Gina Champion
Industry – Medical Device Market Research
Experience – 6yrs
Email – gachampion@uchicago.edu

Name - Ignas Grabauskas
Industry – Investment Management
Experience – 6yrs
Email – ignasg@uchicago.edu

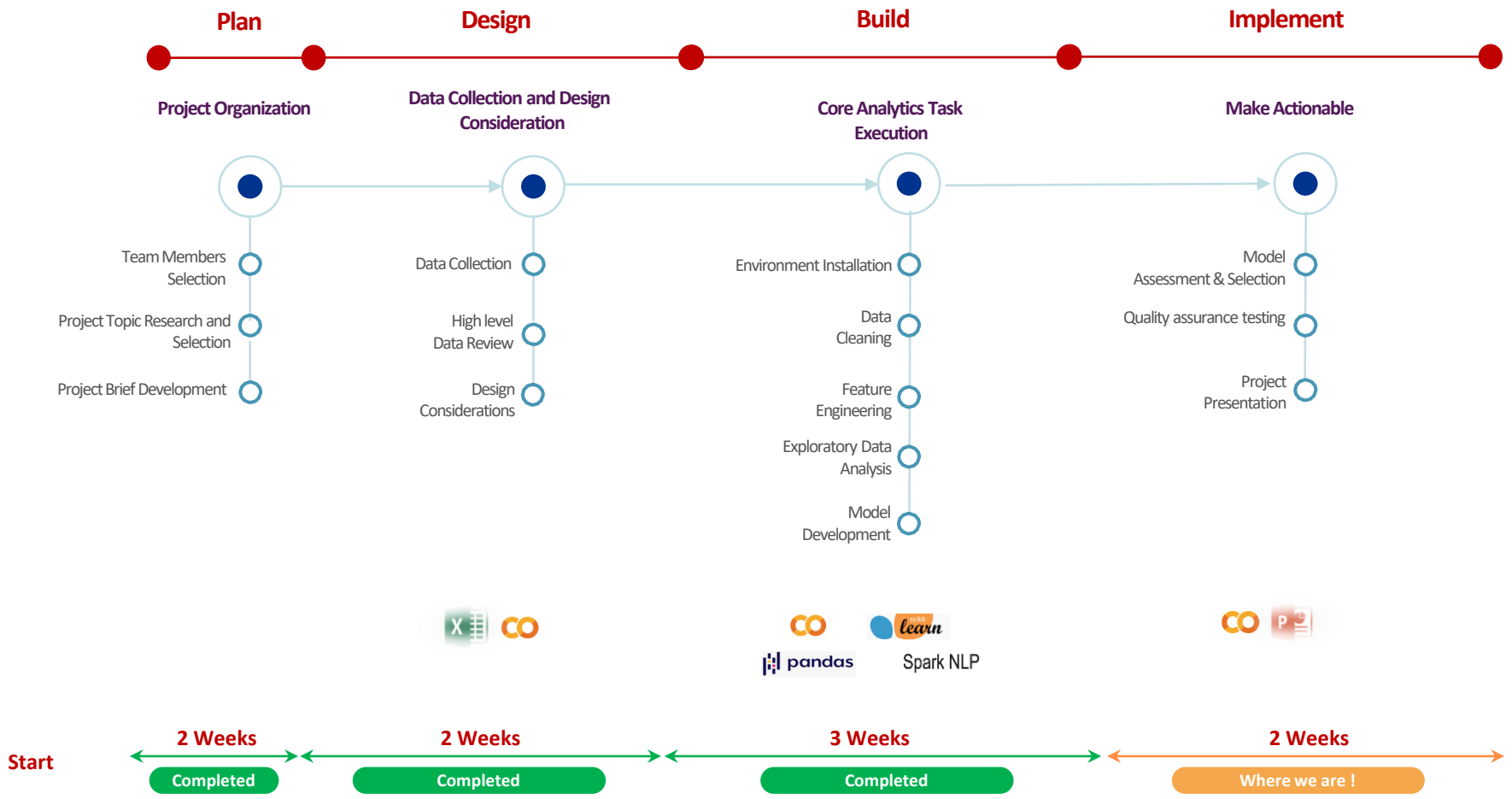
Name - Nicolas Yanez
Industry – Healthcare
Experience – 4yrs
Email – nyanez@uchicago.edu

Name - Oluwafemi Fabiyi
Industry – Big4 Consulting
Experience – 11yrs
Email – fabiyio@uchicago.edu

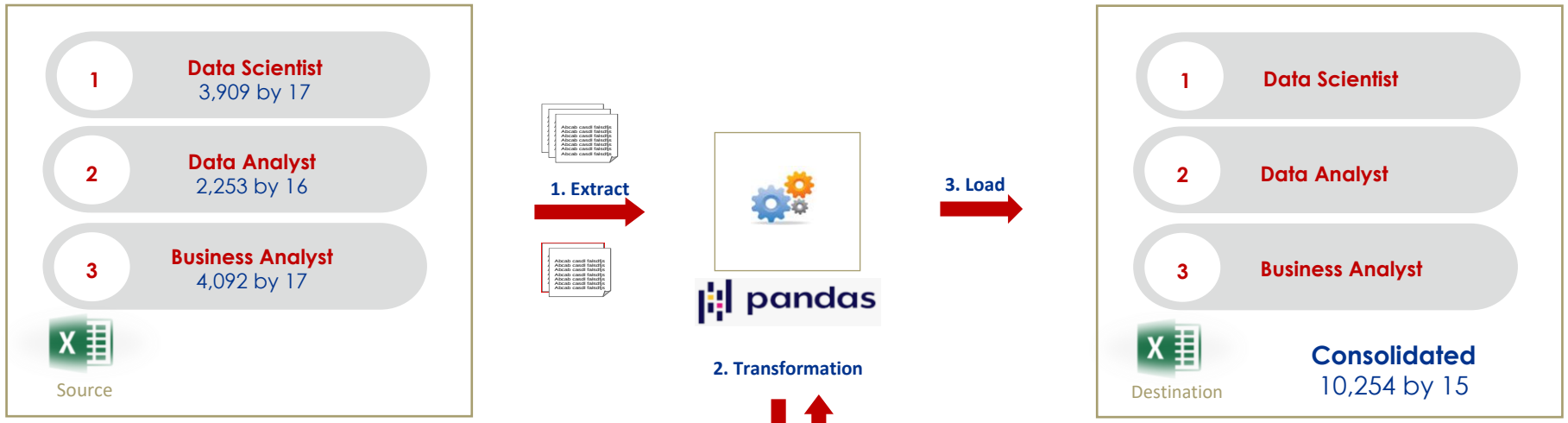
Name - Kaihao Fan
Industry – Medical Device
Experience – 1yr
Email – kaihaofan@uchicago.edu

Name - Stella Cao
Industry – Tech, TBD
Experience – 0
Email – zijcao@uchicago.edu

Data availability and its demonstrable value have created demand amongst employers for individuals with the necessary skills and experience. Our objective is to utilize the job descriptions of job postings for three commonly sought-after positions in data analytics in order to extract insights in their similarities and differences.



Our Work Approach > Data Cleaning



https://www.kaggle.com/andrewmvd

1 - Addressing Extra Columns

- Each data set had some non-descript columns that needed investigating.
- For the **Business Analyst** dataset, the non-descript columns were the result of a subset of rows being shifted.

2 - Combining Datasets

- Before combining datasets, a **target column** in each dataset was created to indicate the job type: **business analyst, data analyst, or data scientist**
- After combining datasets, subsequent cleaning was done

3 - Imputing Founded Year

- KDE was used to impute the missing values for the **Founded Year** column
- Grid search was applied to find the optimal bandwidth
- A sample was drawn from the fitted KDE to fill missing values

4 - Imputing Rating

- KDE was used to impute the missing values of the **Rating** column
- Grid search was applied to find optimal bandwidth
- A sample was drawn from the fitted KDE to fill missing values

5 - Miscellaneous

- Cleaned up **Job Description** by **replacing meta characters**
- Cleaned up **Company Name** column by **removing** embedded ratings
- Dropped columns that weren't needed for analysis e.g., Revenue (categorical) which had mostly missing values

```
'Company - Private': 1,
'Company - Public': 1,
'-1': 1,
'Nonprofit Organization': 0,
'Subsidiary or Business Segment': 1,
'Government': 0,
'College / University': 0,
```

```
'1 to 50 employees': 0,
'51 to 200 employees': 1,
'201 to 500 employees': 2,
'501 to 1000 employees': 3,
'1001 to 5000 employees': 4,
```

02 - Company & Size columns

- Company Ownership and Size columns were both string type
- New features were created of integer type to reflect the categorical nature of the strings

03 – Text Data Processing

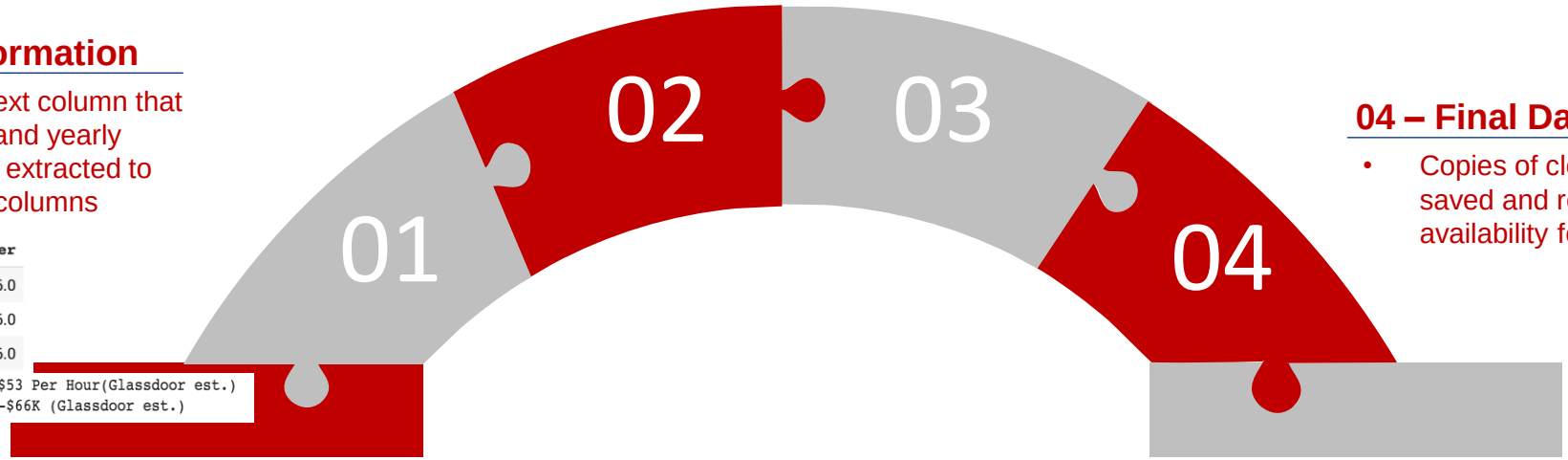
- Job Description was taken through various transformations to create a feature space suitable for NLP work

01 - Salary Information

- Was originally a text column that contained hourly and yearly rates. Rates were extracted to create numerical columns

	SalaryLower	SalaryUpper
0	37.0	66.0
1	37.0	66.0
2	37.0	66.0

```
Example of hourly syntax: $34-$53 Per Hour(Glassdoor est.)
Example of salary syntax: $37K-$66K (Glassdoor est.)
```

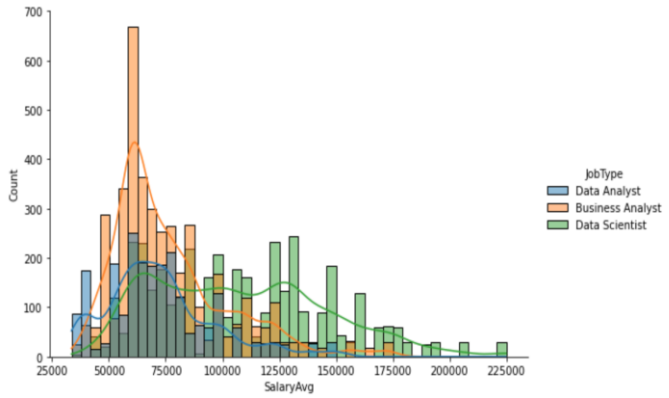


04 – Final Dataset

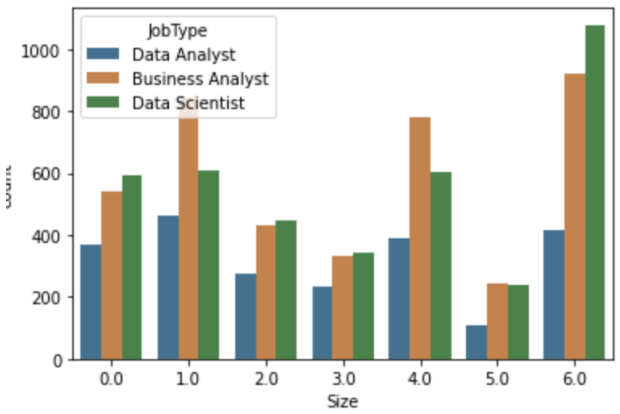
- Copies of cleaned data were saved and read to ensure availability for re-use/backup

	JobTitle	JobDescription	Rating	CompanyName	Location	Headquarters	Size	Industry	Sector	JobType	OrganizationAge	SalaryLower	SalaryUpper	SalaryAvg	IsBusiness
0	Data Analyst, Center on Immigration and Justic...	Are you eager to roll up your sleeves and harn...	3.2	Vera Institute of Justice	New York, NY	New York, NY	2.0	Social Assistance	Non-Profit	Data Analyst	60	37000.0	66000.0	51500.0	0
1	Quality Data Analyst	Overview Provides analytical and technical su...	3.8	Visiting Nurse Service of New York	New York, NY	New York, NY	6.0	Health Care Services & Hospitals	Health Care	Data Analyst	128	37000.0	66000.0	51500.0	0

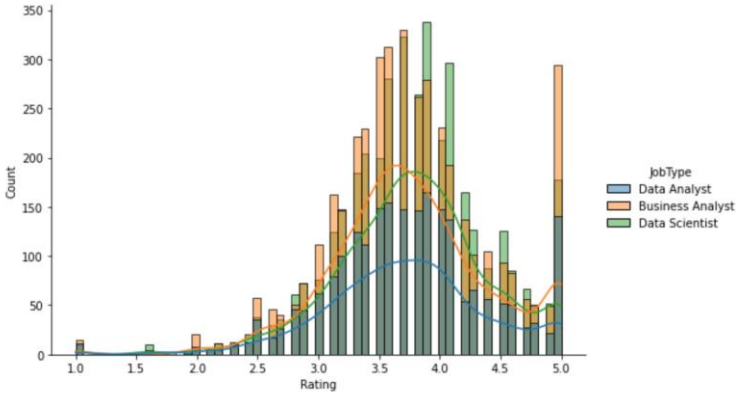
Salary Distribution



Company Size Distribution



Rating Distribution

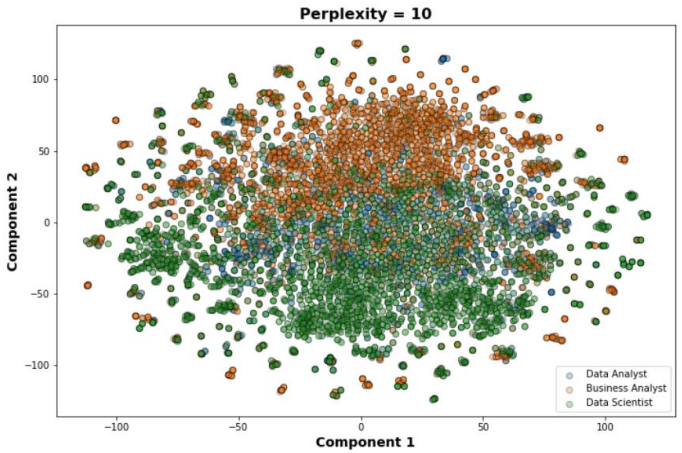


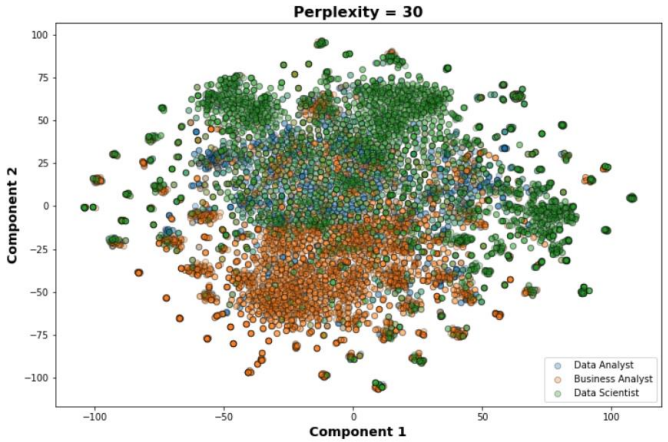
Data Scientist

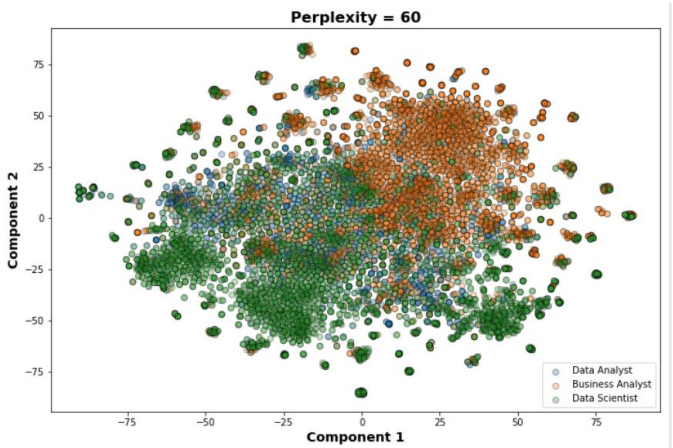
Data Analyst

Business Analyst

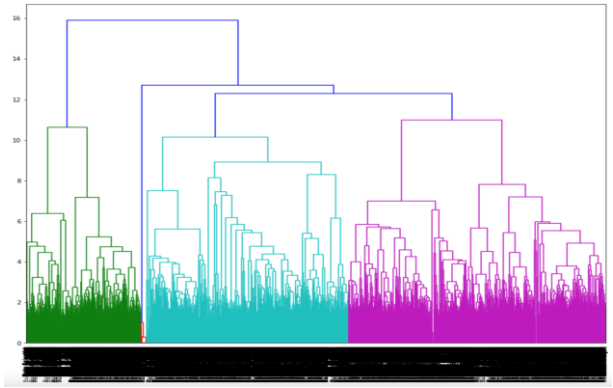
Job Description – t-SNE



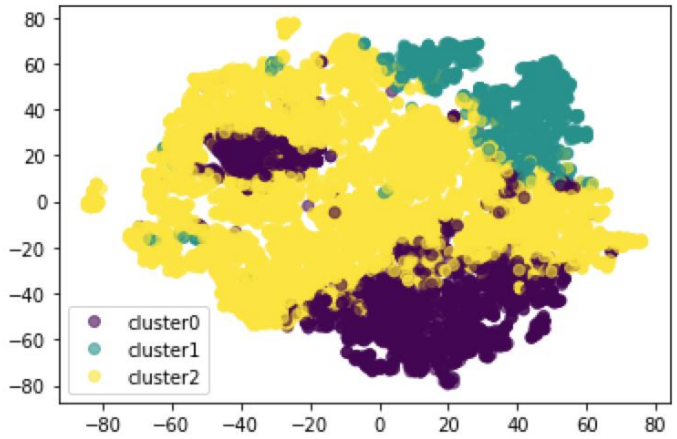




Hierarchical Clustering + TFIDF



K-Means + NMF



Actual vs K-Means Label Comparison

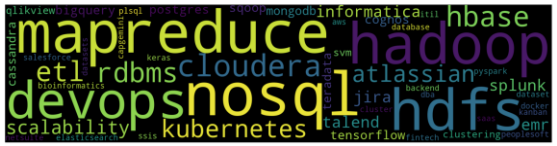
index	cluster	
Business Analyst	0	1874
	1	18
	2	2200
Data Analyst	0	316
	1	80
	2	1857
Data Scientist	0	326
	1	1162
	2	2421

Top2Vec – Overall Dataset

["Are you eager to roll up your sleeves and harness data to drive policy change? Do you enjoy sifting through complex datasets to illuminate trends
'Overview Provides analytical and technical support for the integration of multiple data sources used to prepare internal and external reporting f
'We're looking for a Senior Data Analyst who has a love of mentorship, data visualization, and generating actionable insights from raw data. In thi
'Requisition NumberRR-0001939 Remote:Yes We collaborate. We create. We innovate. Intrigued? You're a business professional with an innate curiosi
"ABOUT FANDUEL GROUP FanDuel Group is a world-class team of brands and products all built with one goal in mind – to give fans new and innovative

Topic 0

Topic 1



Topic 2

Latent Dirichlet Allocation (LDA)

The topic would be 0
['analyst', 'analysis', 'ability', 'management', 'team', 'requirements', 'skills', 'work', 'business', 'data']

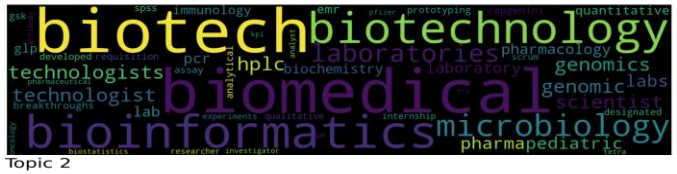
The topic would be 1
['ibm', 'hadoop', 'engineer', 'python', 'data', 'ml', 'aws', 'spark', 'learning', 'machine']

The topic would be 2
['lab', 'molecular', 'gs', 'scientist', 'biology', 'cell', 'scientific', 'research', 'clinical', 'laboratory']

Data Scientist

Topic 0

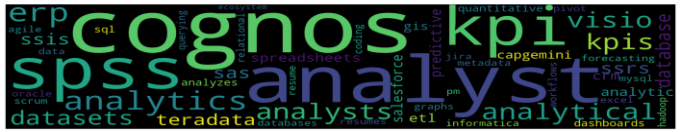
Topic 1



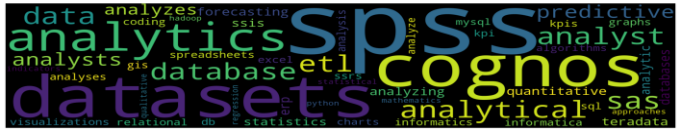
Topic 2

Data Analyst

Topic 0

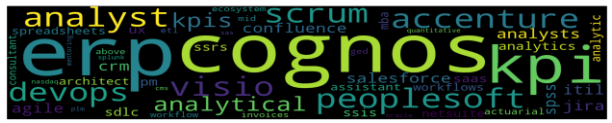


Topic 1

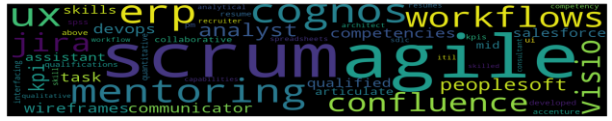


Business Analyst

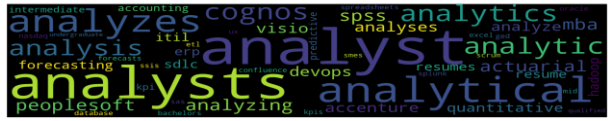
Topic 0



Topic 1



Topic 2



Job Title Classification Accuracy	Naïve Bayes	SVM
• No preprocessing	0.822	0.8639
• Stopwords removed	0.8235	0.8644
• Stopwords removed + stemming	0.8203	0.8634
Job Description Classification Accuracy		
• No preprocessing	0.7172	0.7252
• Stopwords removed	0.7248	0.7471
• Stopwords removed + stemming	0.733	0.7503

Job Title SVM (Stopwords Removed + stemmed)

Parameters (after Grid Search)	
Loss	Hinge
Penalty	12
Alpha	0.001
Random State	1
Ngram	(1, 2)

	precision	recall	f1-score	support
Business Analyst	0.97	0.91	0.94	822
Data Analyst	0.67	0.95	0.79	470
Data Scientist	0.97	0.78	0.86	759
accuracy			0.87	2051
macro avg	0.87	0.88	0.86	2051
weighted avg	0.90	0.87	0.88	2051

Job Description SVM (Stopwords Removed + stemmed)

Parameters (after Grid Search)	
Loss	Hinge
Penalty	12
Alpha	0.001
Random State	1
Ngram	(1, 2)

	precision	recall	f1-score	support
Business Analyst	0.76	0.88	0.82	822
Data Analyst	0.62	0.31	0.42	470
Data Scientist	0.74	0.83	0.78	759
accuracy			0.73	2051
macro avg	0.70	0.68	0.67	2051
weighted avg	0.72	0.73	0.71	2051

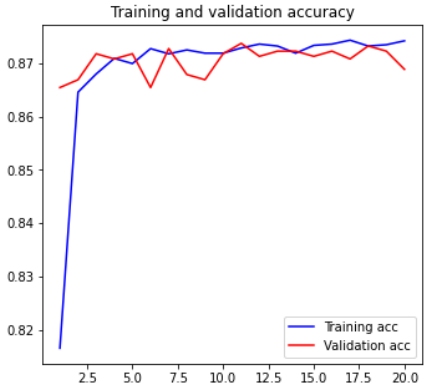
Job Title - Neural Network

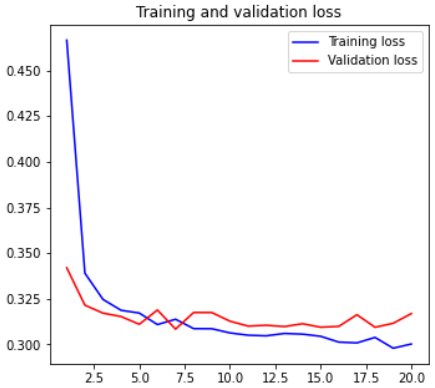
Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 50, 100)	216500
global_max_pooling1d (Global	(None, 100)	0
dense (Dense)	(None, 25)	2525
dropout (Dropout)	(None, 25)	0
dense_1 (Dense)	(None, 3)	78

Total params: 219,103
Trainable params: 219,103
Non-trainable params: 0

Training Accuracy: 0.8753
Testing Accuracy: 0.8688





	precision	recall	f1-score	support
0	0.98	0.90	0.94	3258
1	0.66	0.98	0.79	1788
2	0.99	0.79	0.88	3157
accuracy			0.88	8203
macro avg	0.87	0.89	0.87	8203
weighted avg	0.91	0.88	0.88	8203

	precision	recall	f1-score	support
0	0.97	0.89	0.93	834
1	0.66	0.98	0.79	465
2	0.98	0.77	0.86	752
accuracy			0.87	2051
macro avg	0.87	0.88	0.86	2051
weighted avg	0.90	0.87	0.87	2051

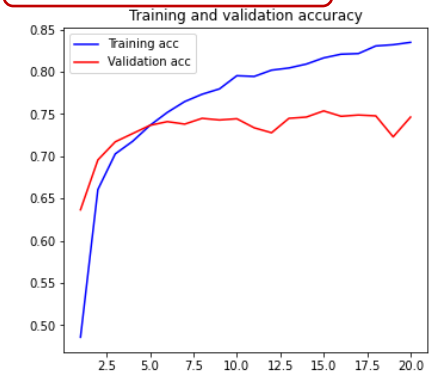
Job Description - Neural Network

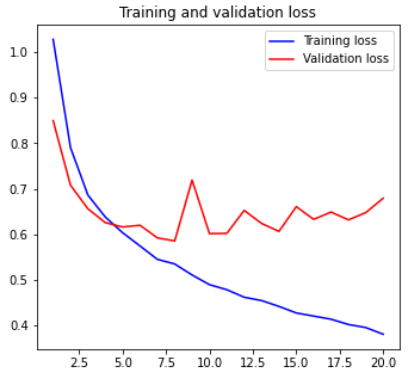
Model: "sequential_1"

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 800, 100)	4158700
global_max_pooling1d_1 (Glob	(None, 100)	0
dense_2 (Dense)	(None, 50)	5050
dropout_1 (Dropout)	(None, 50)	0
dense_3 (Dense)	(None, 25)	1275
dropout_2 (Dropout)	(None, 25)	0
dense_4 (Dense)	(None, 3)	78

Total params: 4,165,103
Trainable params: 4,165,103
Non-trainable params: 0

Training Accuracy: 0.8522
Testing Accuracy: 0.7465





	precision	recall	f1-score	support
0	0.99	0.86	0.92	3258
1	0.61	0.97	0.75	1788
2	0.97	0.78	0.86	3157
accuracy			0.85	8203
macro avg	0.86	0.87	0.84	8203
weighted avg	0.90	0.85	0.86	8203

	precision	recall	f1-score	support
0	0.85	0.76	0.80	834
1	0.51	0.78	0.61	465
2	0.90	0.72	0.80	752
accuracy			0.75	2051
macro avg	0.75	0.75	0.74	2051
weighted avg	0.79	0.75	0.76	2051

Our Work Approach > Modeling (4 of 4)

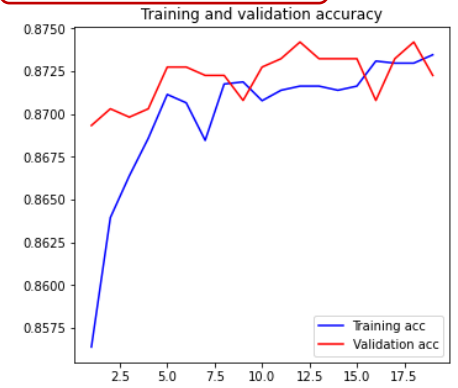
Job Title + Salary - Neural Network

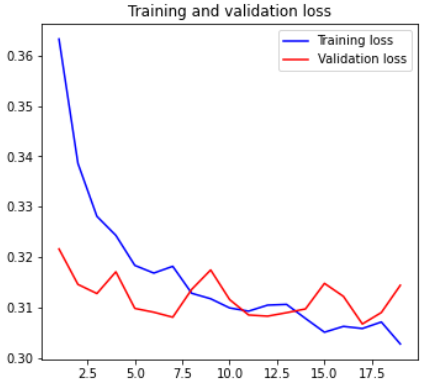
Model: "model"

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 50)]	0	
embedding_2 (Embedding)	(None, 50, 100)	216500	input_1[0][0]
input_2 (InputLayer)	[(None, 1)]	0	
global_max_pooling1d_2 (GlobalM	(None, 100)	0	embedding_2[0][0]
dense_5 (Dense)	(None, 5)	10	input_2[0][0]
concatenate (Concatenate)	(None, 105)	0	global_max_pooling1d_2[0][0] dense_5[0][0]
dense_6 (Dense)	(None, 50)	5300	concatenate[0][0]
dropout_3 (Dropout)	(None, 50)	0	dense_6[0][0]
dense_7 (Dense)	(None, 25)	1275	dropout_3[0][0]
dropout_4 (Dropout)	(None, 25)	0	dense_7[0][0]
dense_8 (Dense)	(None, 3)	78	dropout_4[0][0]

Total params: 223,163
Trainable params: 223,163
Non-trainable params: 0

Training Accuracy: 0.8742
Testing Accuracy: 0.8723





	precision	recall	f1-score	support
0	0.98	0.90	0.94	3258
1	0.65	1.00	0.79	1788
2	1.00	0.78	0.88	3157
accuracy			0.87	8203
macro avg	0.87	0.89	0.87	8203
weighted avg	0.91	0.87	0.88	8203

	precision	recall	f1-score	support
0	0.97	0.90	0.93	834
1	0.66	0.99	0.79	465
2	0.99	0.77	0.87	752
accuracy			0.87	2051
macro avg	0.88	0.89	0.86	2051
weighted avg	0.91	0.87	0.88	2051

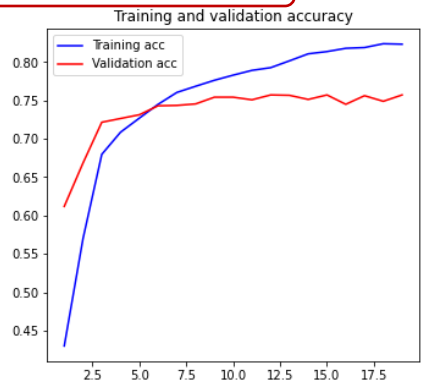
Job Description + Salary - Neural Network

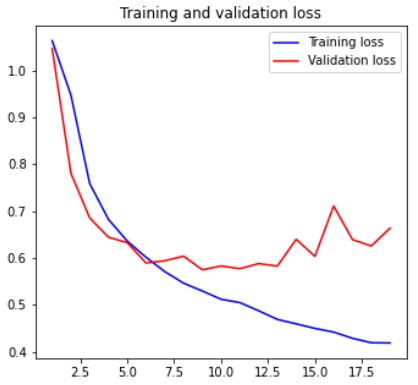
Model: "model_1"

Layer (type)	Output Shape	Param #	Connected to
input_3 (InputLayer)	[(None, 800)]	0	
embedding_3 (Embedding)	(None, 800, 100)	4158700	input_3[0][0]
input_4 (InputLayer)	[(None, 1)]	0	
global_max_pooling1d_3 (GlobalM	(None, 100)	0	embedding_3[0][0]
dense_9 (Dense)	(None, 5)	10	input_4[0][0]
concatenate_1 (Concatenate)	(None, 105)	0	global_max_pooling1d_3[0][0] dense_9[0][0]
dense_10 (Dense)	(None, 50)	5300	concatenate_1[0][0]
dropout_5 (Dropout)	(None, 50)	0	dense_10[0][0]
dense_11 (Dense)	(None, 25)	1275	dropout_5[0][0]
dropout_6 (Dropout)	(None, 25)	0	dense_11[0][0]
dense_12 (Dense)	(None, 3)	78	dropout_6[0][0]

Total params: 4,165,363
Trainable params: 4,165,363
Non-trainable params: 0

Training Accuracy: 0.8497
Testing Accuracy: 0.7572





	precision	recall	f1-score	support
0	0.92	0.90	0.91	3258
1	0.63	0.91	0.74	1788
2	0.99	0.76	0.86	3157
accuracy			0.85	8203
macro avg	0.85	0.86	0.84	8203
weighted avg	0.88	0.85	0.86	8203

	precision	recall	f1-score	support
0	0.80	0.84	0.82	834
1	0.54	0.68	0.60	465
2	0.92	0.72	0.81	752
accuracy			0.76	2051
macro avg	0.75	0.74	0.74	2051
weighted avg	0.78	0.76	0.76	2051

Conclusion

Model Performance
NB, SVM, and NN both arrived at similar accuracy (SVM outperformed NB)

Ambiguity
Difficult to differentiate between various job descriptions

Other Features
Salary seemed somewhat promising based on EDA but showed little improvement for modeling

Challenges

Modeling
Steep learning curve given inexperience in NLP

Orchestration
Aggregating individual development into final notebook (each person doing their work in a decentralized fashion)
Working on final notebook in parallel

Future Improvements

Other NLP Techniques
Choose techniques that suit the task

Job Title
Better validation that the datasets include only the jobs they are supposed to

Model Selection
Given the EDA, utilize other models

Questions