# Explore Data Science Job Market in the U.S.

MSBA Cohort B Team 3
Jingcheng Huang, Minna Tang, Youming Qiu, Yishuang Song, Yue Gong

## 1. Introduction

We suppose all students in the MSBA program want to get jobs after graduation in order to apply what we've learned at school to the real business environment.

Here are some business problems we want to solve:

1. What type of talents do employers want concerning tools, skills, degrees and majors?
2. What is the difference between different job roles(e.g. Engineer vs. Data Scientist vs. Analyst) in the data science job market?
3. How much would you make as a data scientist with your specific educational backgrounds and experiences?

## 2. Data Sources

In order to address our business problems, we have decided to utilized the following datasets :

- Data Scientist Job Market in the U.S. https://www.kaggle.com/sl6149/data-scientist-job-market-in-the-us#alldata.csv
- Indeed dataset - Data Scientist/Analyst/Engineer https://www.kaggle.com/elroyggj/indeed-dataset-data-scientistanalystengineer

## 3. Text Analysis

We use the following analytics methods during the project:

**1. Text Cleaning**

Text cleaning was used to break down all the job descriptions and requirements into text chunks, so we can identify important factors that would affect the salary.

For example, we find that most job postings say that they require their applicants to at least have a college degree. Some other important factors are leadership, teamplayer, years of experience, programming language, and etc.

**2. Word Cloud**

Word cloud showed us important words from the job description. The larger the word appears on the Word Cloud, the more important it is for the specific job position. Data analysts focus on team and research. Data scientist and data engineer focus more on machine learning. These information give us some first sights of jobs requirements for each job position.

**3. N-gram**

As for top 10 unigrams of descriptions, three data jobs have lots of same word, such as data, experience,skills , business and etc.

For Bi-grams, data scientists and data engineer focus on machine learning and computer science. Data analyst focus on research and years experience.

As for Tri-grams, data scientists and data analyst have same top three words such as equal opportunity employer and without regard race. However, Data engineers are more care about degree of computer science.

| | Data Scientist | Data Analyst | Data Engineer |
|---|---|---|---|
| Python | 53.29% | 29.10% | 54.05% |
| R | 45.04% | 33.58% | 17.79% |
| SQL | 29.97% | 34.54% | 27.03% |
| Java | 18.64% | 7.57% | 36.79% |
| C/C++ | 19.75% | 12.15% | 34.61% |
| Tableau | 7.65% | 14.82% | 3.38% |
| Excel | 8.50% | 37.21% | 5.26% |
| Matlab | 10.43% | 6.72% | 8.71% |

## 4. Predictive Supervised ML Model

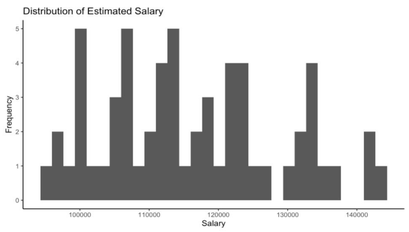
We used the indeed dataset to train a supervised machine learning model to help predict a candidate's future salary in the data science field. The model was trained in various advanced methods, such as boosted trees and KNN.

The input variables are factors that indicate candidates' and company's backgrounds; the output variable is a numeric salary value or a salary range.

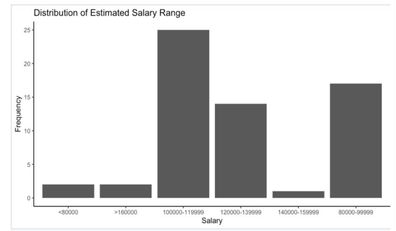
The dataset was divided into 80% training set, 20% testing set. 10-fold cross validation was applied on the training set to improve our model. Out of all the models, Random forest is the best with the lowest RMSE value and accuracy score.

We asked our networks to fill out a salary prediction survey. Out of the 60 survey respondents:

- 73.33% have at least 5 technical skills
  - Python, R, SQL, Tableau, Machine Learning
- Popular locations: New York, Massachusetts, and California
- Most popular industries: Consulting, Internet, and Banking
- Predicted salaries range from $94, 424 - $142,819 with a mean of $116, 252.




Regression: Salary Value




Classification: Salary Range

## 5. Results and Recommendation

Who gets hired:

- **Data Scientist:** Python, R, SQL
- **Data Analyst:** Excel, SQL, R
- **Data Engineer:** Java, C/C++, Python
- **Essential Skills:** Machine Learning, Data/Statistical Analysis, Data Visualization
- **Preferred Education Level:** Master's degree (or higher)

Resume tips:

- Include your technical skills
- Read job descriptions carefully

## 6. Limitations and Challenges

- Unverified title categorization
- For survey responses
  - Small sample size
  - Maybe not very representative
  - Inflated responses
- Unexpected epidemic
  - Unable to communicate face-to-face

## 7. Acknowledgements

This project would not be possible without the support of Professor Mohammad Soltanieh-ha. His dedication to students was remarkable because we would have not completed this project without bi-weekly project meetings with him.

Professors at the MS in Business Analytics program have provided their expertise to help us navigate through the project. The staff member at Questrom School of Business assisted us to make this project possible logistically.