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

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

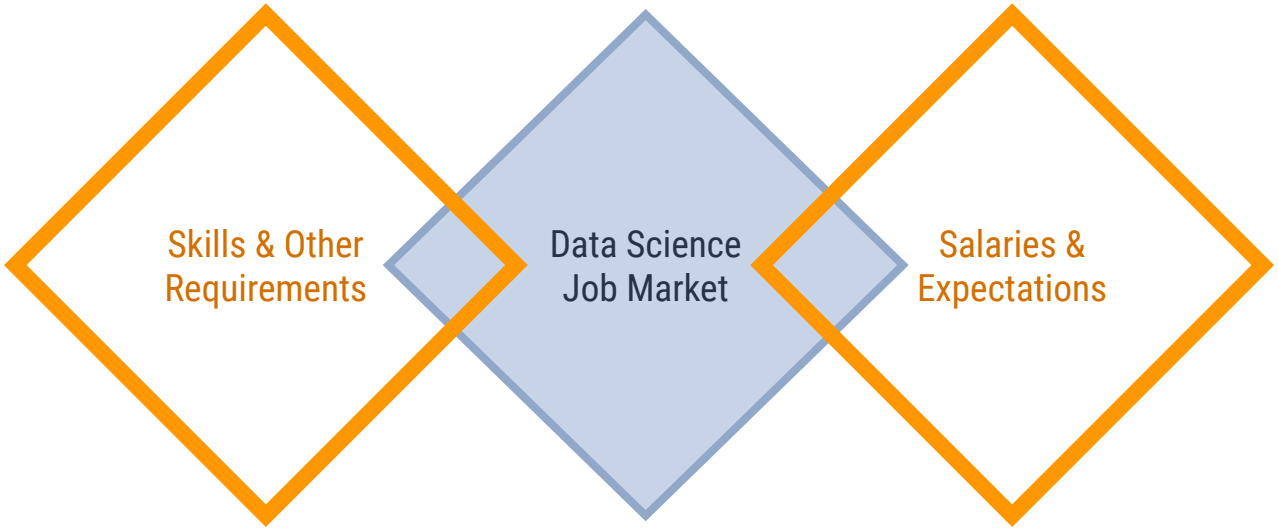
Agenda

1. Project Goals & Dataset Overview
2. Exploratory Data Analysis (EDA)
3. Data Cleaning
4. Unsupervised Machine Learning
5. Supervised Machine Learning
6. Conclusion & Recommendation
7. Limitations & Challenges

1

Project Goals & Dataset Overview

Our Project Goal



Data Sources

- **Data Scientist Job Market in the U.S.** (Updated in 2019)

https://www.kaggle.com/sl6149/data-scientist-job-market-in-the-us#alldata.csv

- **Indeed dataset - Data Scientist/Analyst/Engineer** (Updated in 2019)

https://www.kaggle.com/elroyggj/indeed-dataset-data-scientistanalystengineer

Attributes of Datasets

Data Scientist Job Market (6,964 observations)	Indeed Dataset (5,715 observations)
Company	Job Title
Position	Job Type
Job Description	Skills
Number of Reviews	Queried Salary
Company Location	Company Location

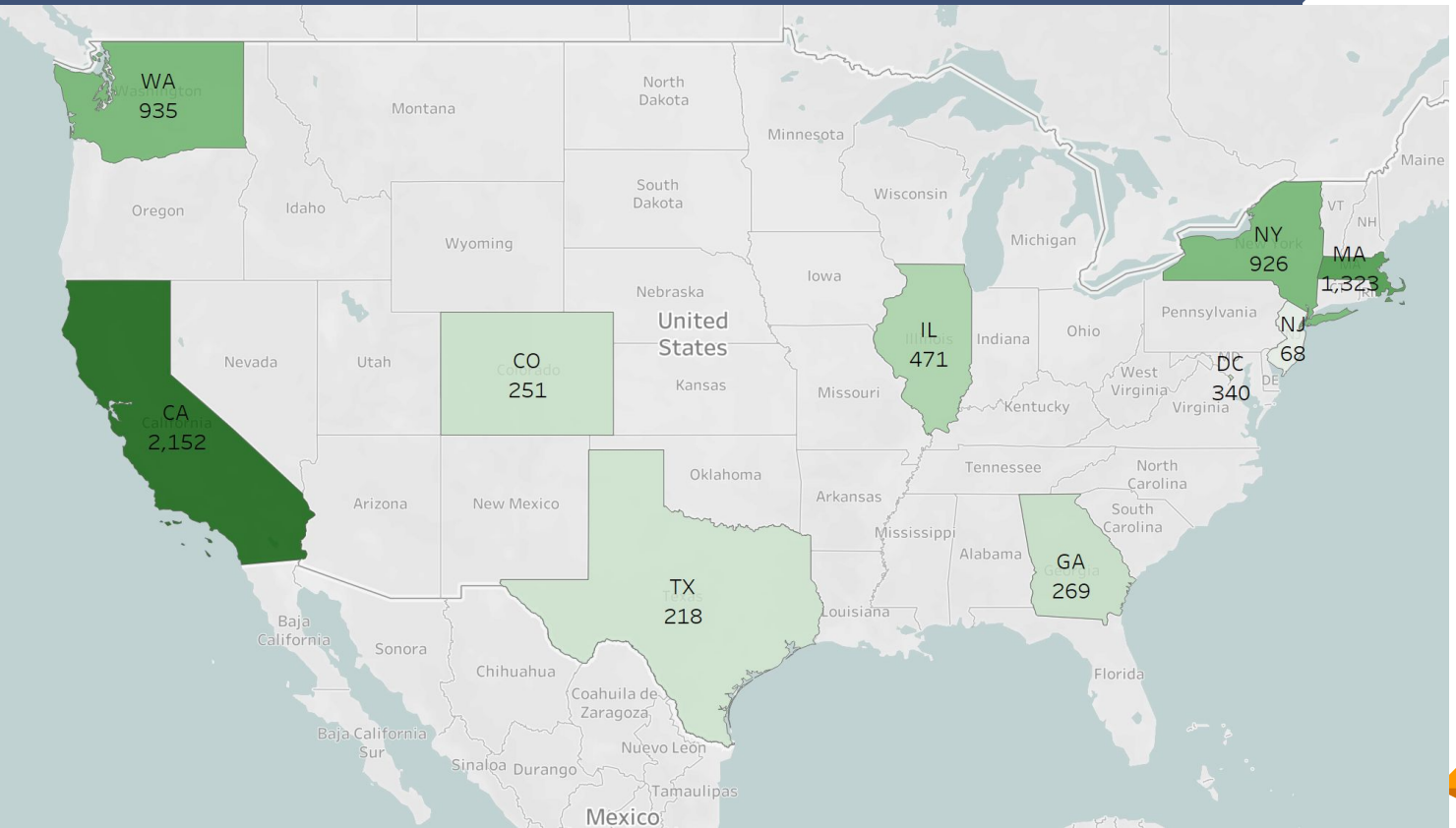
Business Problems

- What type of talents do employers want regarding tools, skills, degrees and majors?
- How much would you make if you work in a data related industry in U.S. given your knowledge on technical tools and degrees?
- What possible future career paths there could be?

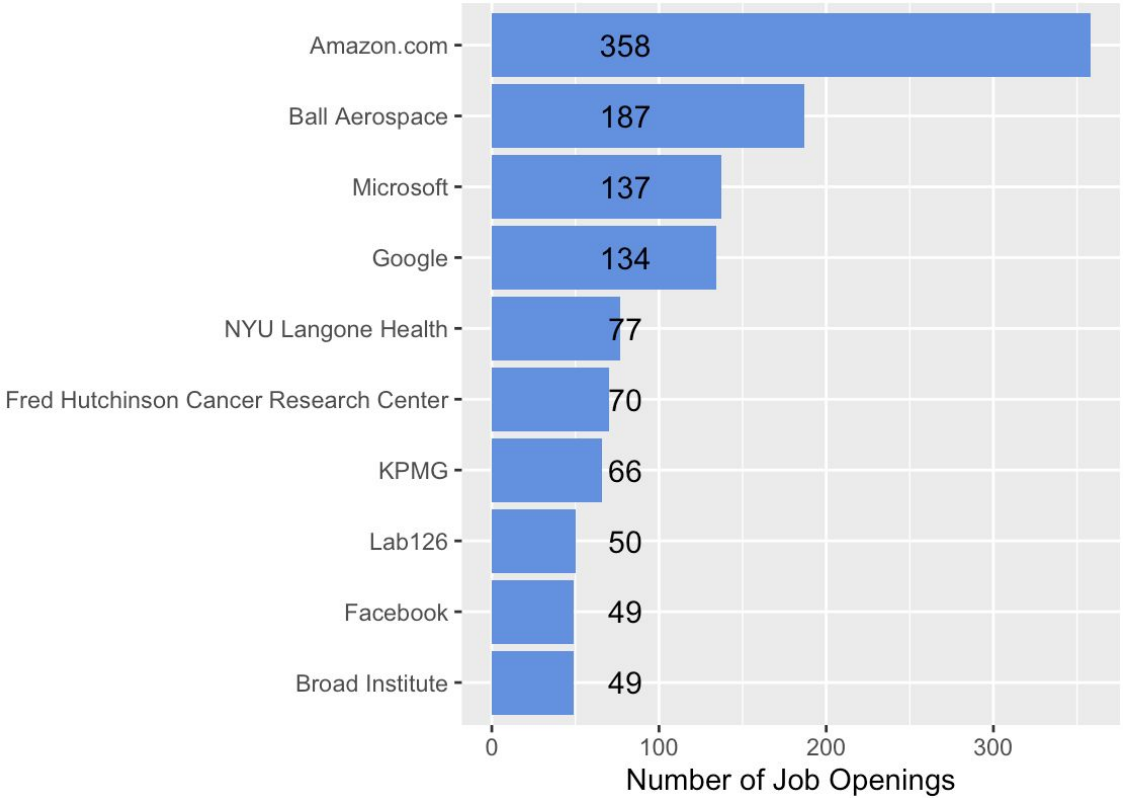
2

Exploratory Data Analysis (EDA)

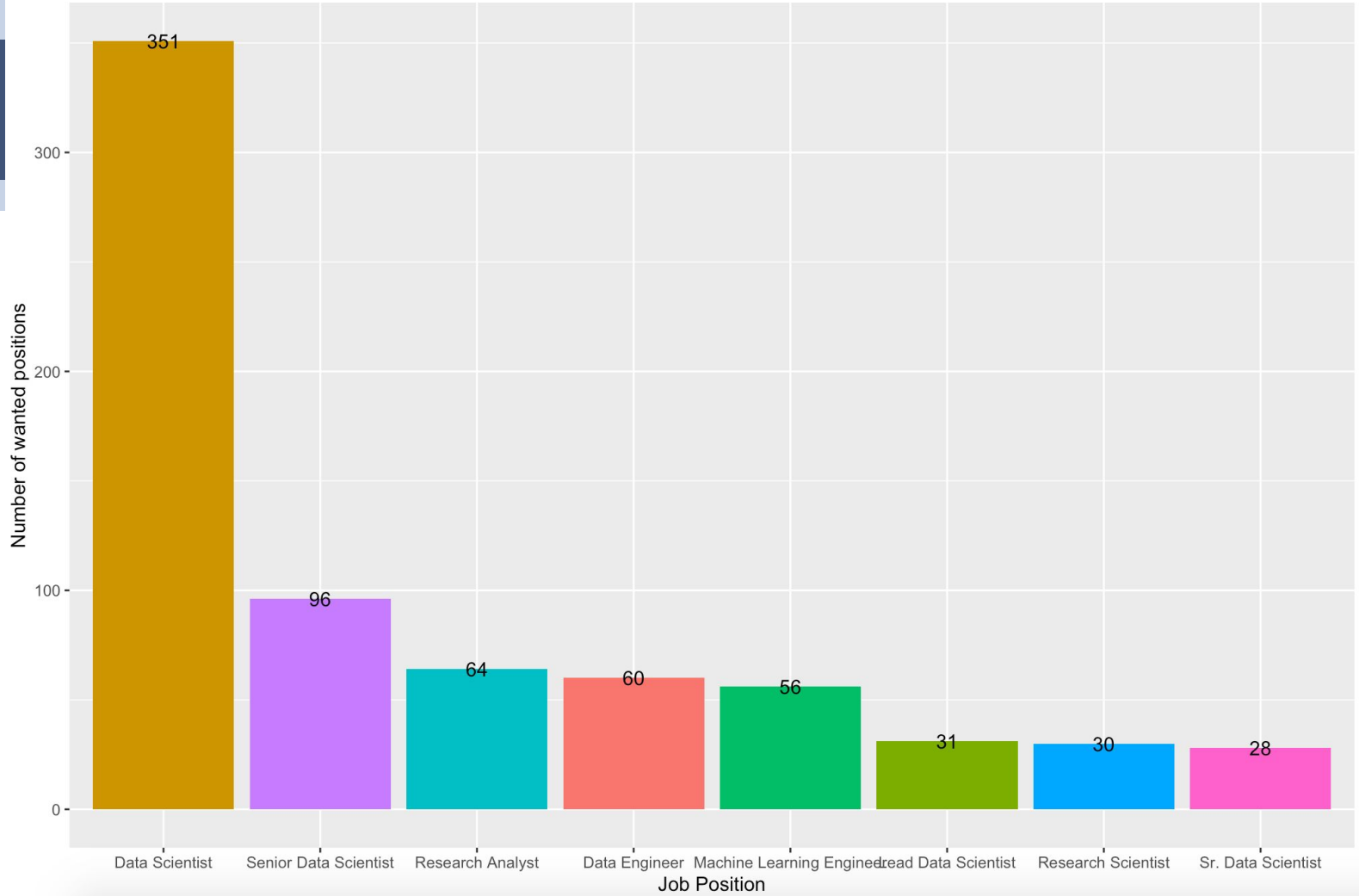
Top 10 States with Largest Number of Data Scientist Jobs



Companies With Most Data Science Job Openings



Top 8 Most Wanted Job Positions in Data Science Industry



3

Data Cleaning

Split the Location To City and State

location
Atlanta, GA 30301
Atlanta, GA
Atlanta, GA
Atlanta, GA 30303
Atlanta, GA

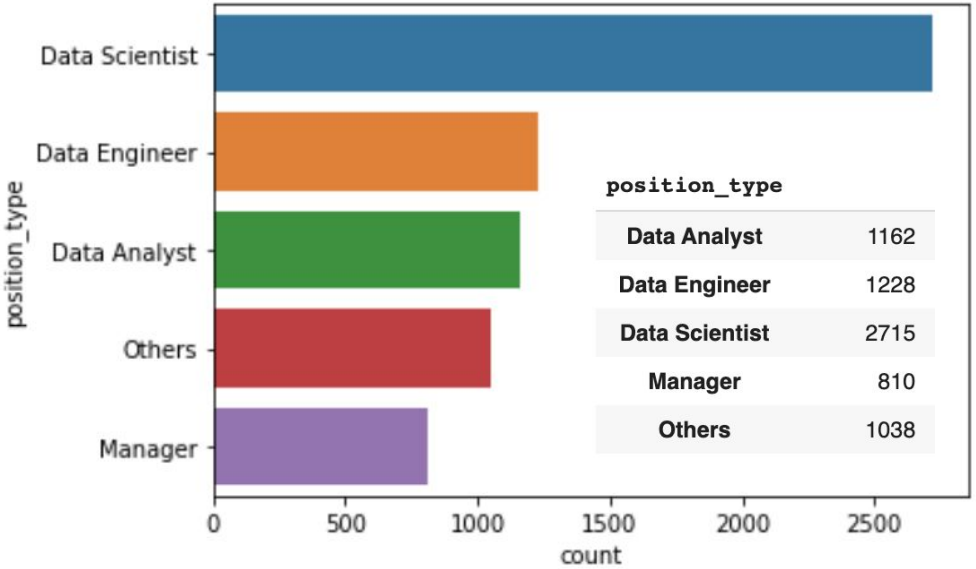
city	state
Atlanta	GA
Atlanta	GA
Atlanta	GA
Atlanta	GA
Atlanta	GA

- Remove zip code
- Split the “location” to “city” and “state”

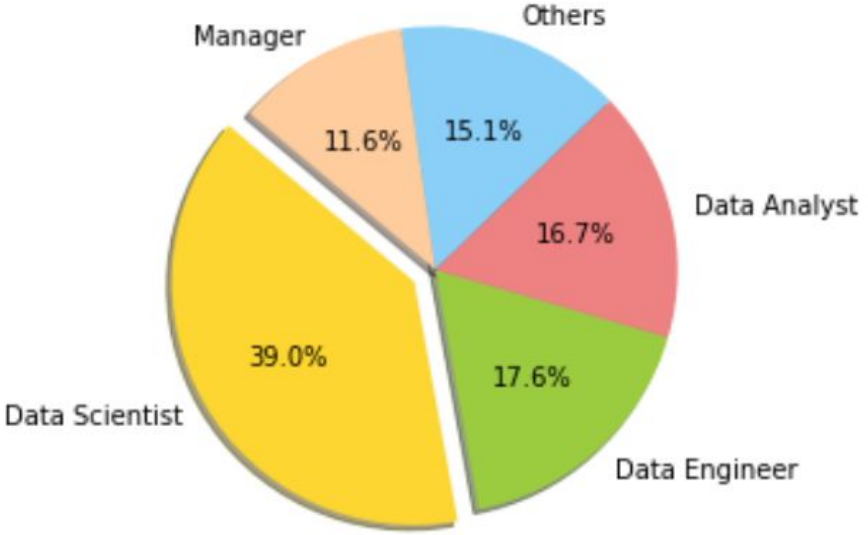
Categorize Job Positions into 5 Main Types

1. Data Scientist
   - “Scientist”, “Science”, “Data science”
2. Data Analyst
   - “Analyst”, “Analytics”, “Analysis”, “Analytical”, “Intelligent”, “Statistician”
3. Data Engineer
   - “Machine learning”, “Machine”, “Learning”, “Engineer”, “Programmer”, “Developer”, “Principal statistical programmer”
4. Manager
   - “Manager”, “Director”, “Senior”
5. Others
   - All other positions that don’t fit the above search criteria

Basic Count for Each Position Type



position_type	
Data Analyst	1162
Data Engineer	1228
Data Scientist	2715
Manager	810
Others	1038



Let's Focus on 3 Main Job Positions

Data Analyst

Data Scientist

Data Engineer

How Are They Different?

4

Unsupervised Machine Learning - Text Analysis

Word Cloud → Data Analyst

Top 3 Keywords:

- **Team**
- **Data**
- **Client**

Word Cloud → Data Scientist

Data Scientist Job Description

Top 3 Keywords:

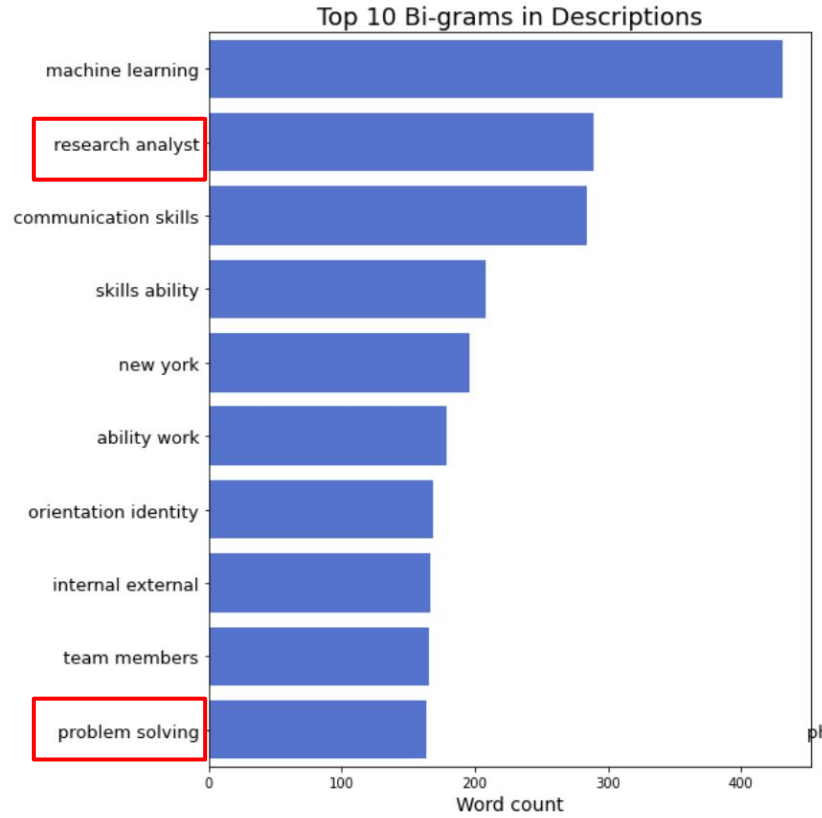
- **Machine Learning**
- **Communication skill**
- **Data Science**

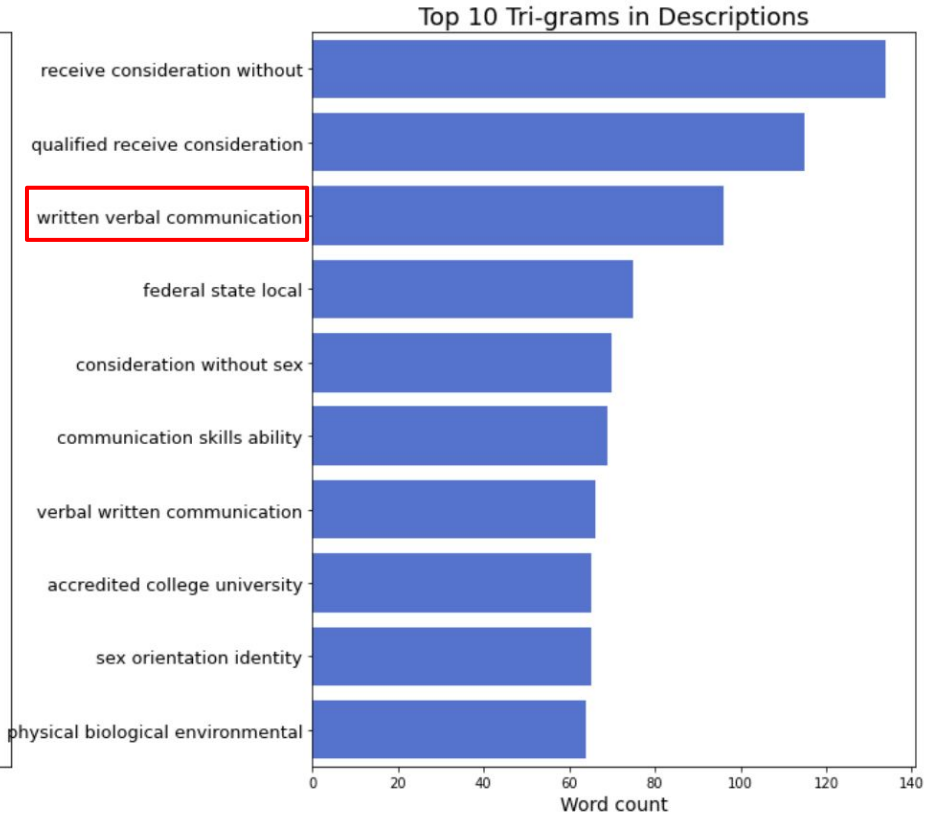
Word Cloud → Data Engineer

Top 3 Keywords:

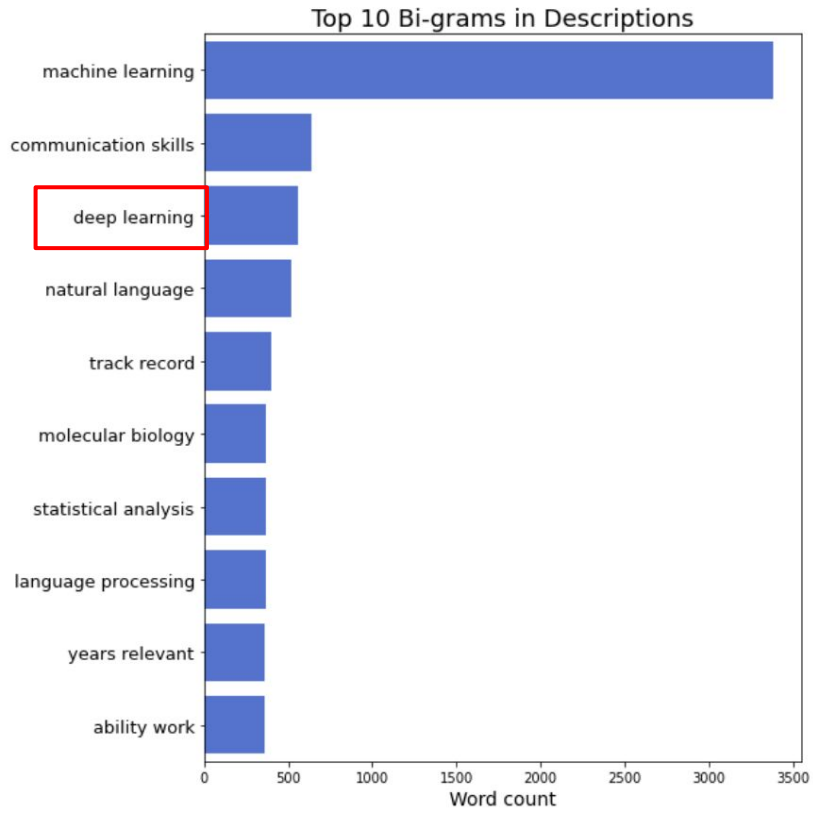
- **Machine Learning**
- **Computer Science**
- **Application**

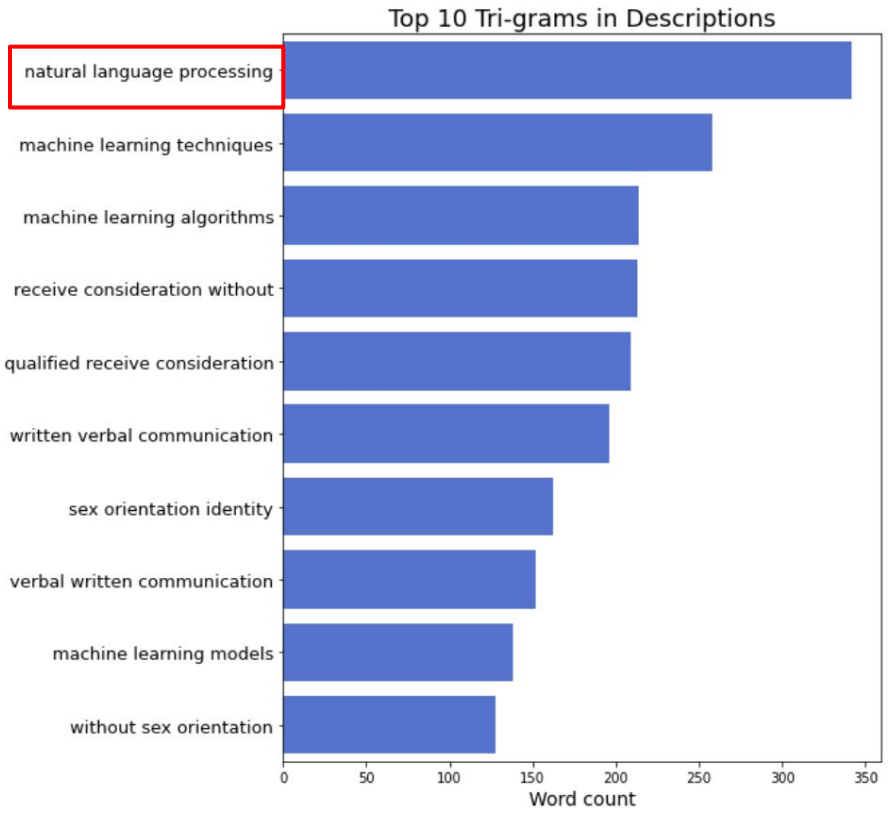
N-gram → Data Analyst



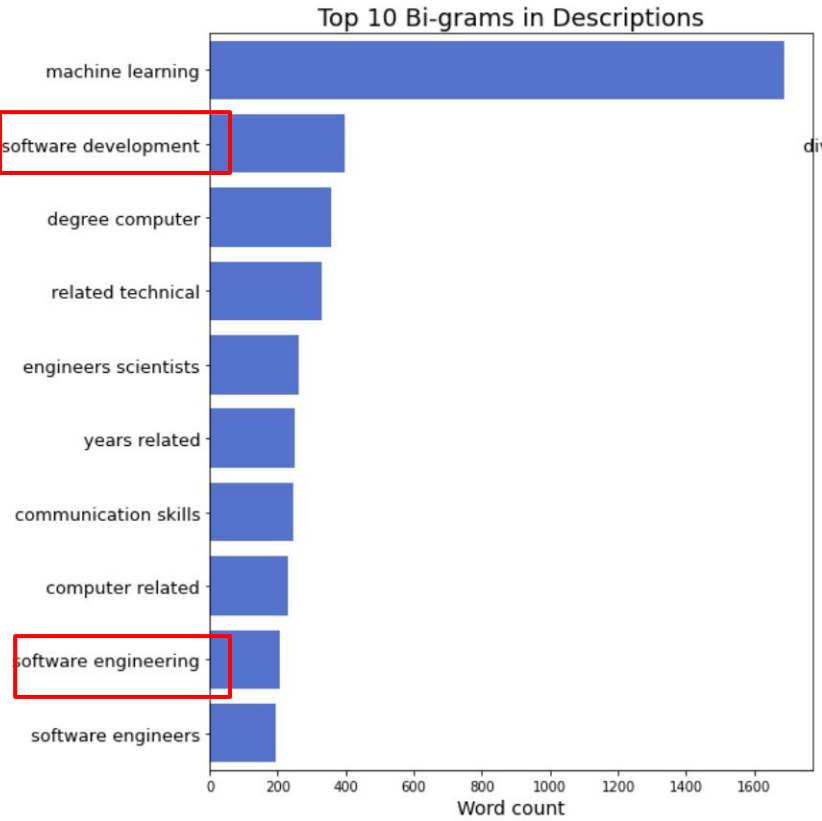


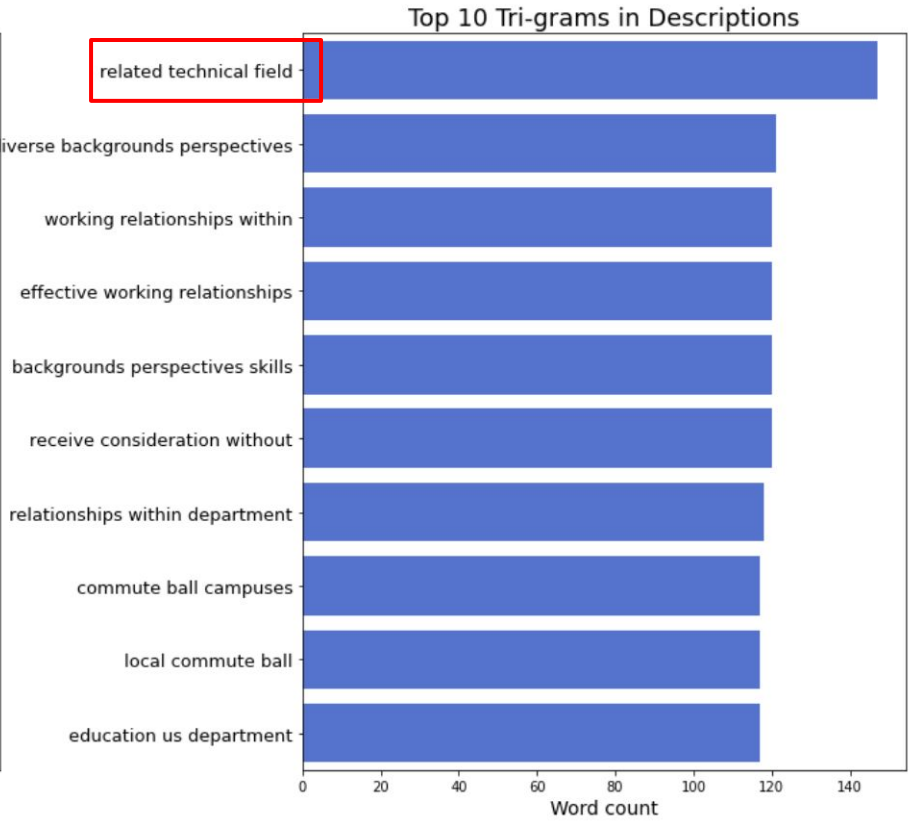
N-gram → Data Scientist





N-gram → Data Engineer





Comparison Between Bi-gram & Tri-gram

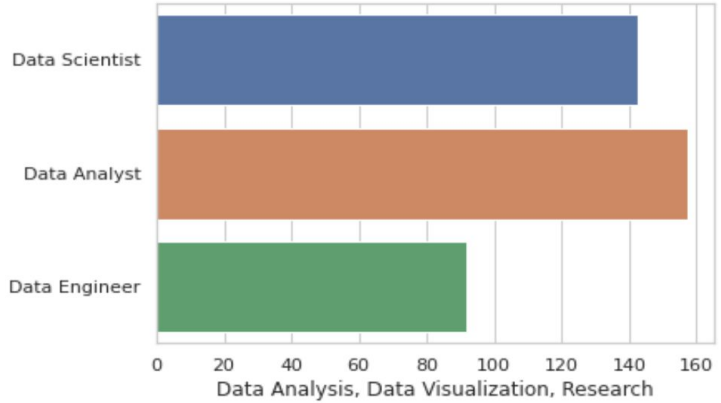
Data Scientist	Data Analyst	Data Engineer
machine learning	machine learning	machine learning
communication skills	research analyst	software development
deep learning	communication skills	degree computer
natural language	skills ability	related technical
track record	new york	engineers scientists
molecular biology	ability work	years related
statistical analysis	orientation identity	communication skills
language processing	internal external	computer related
years relevant	team members	software engineering
ability work	problem solving	software engineers

Data Scientist	Data Analyst	Data Engineer
natural language processing	receive consideration without	related technical field
machine learning techniques	qualified receive consideration	diverse backgrounds perspectives
machine learning algorithms	written verbal communication	working relationships within
receive consideration without	federal state local	effective working relationships
qualified receive consideration	consideration without sex	backgrounds perspectives skills
written verbal communication	communication skills ability	receive consideration without
sex orientation identity	verbal written communication	relationships within department
verbal written communication	accredited college university	commute ball campuses
machine learning models	sex orientation identity	local commute ball
without sex orientation	physical biological environmental	education us department

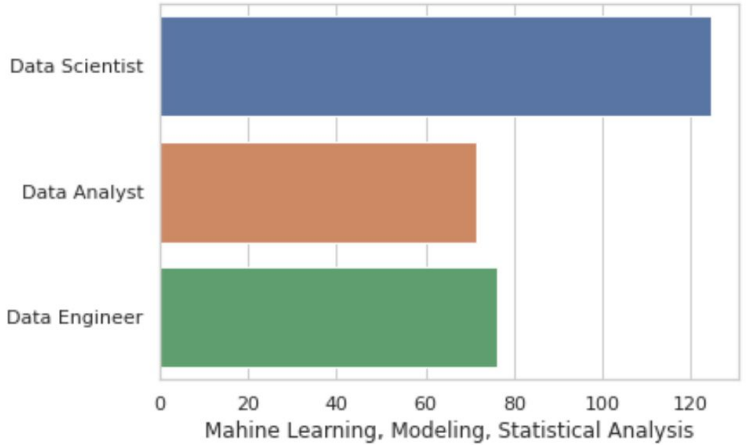
Importance of Technical Skills For Each Job Position

	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%

Data Scientist vs. Data Analyst vs. Data Engineer

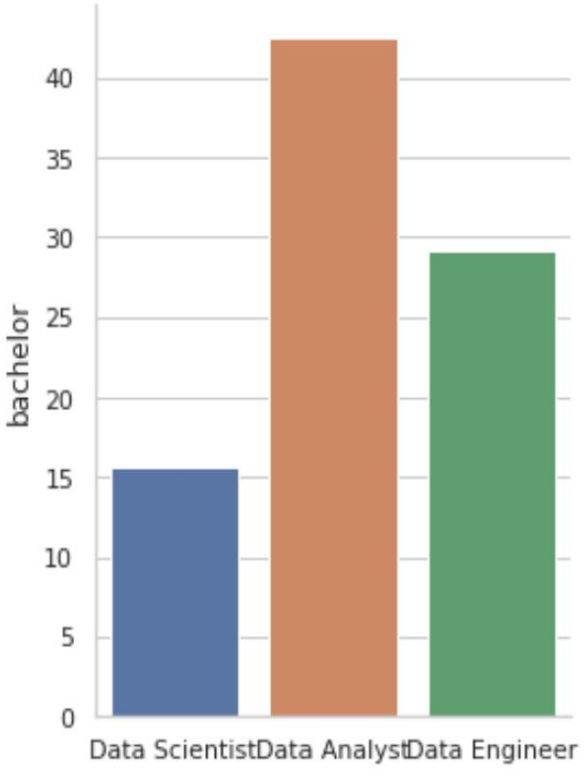


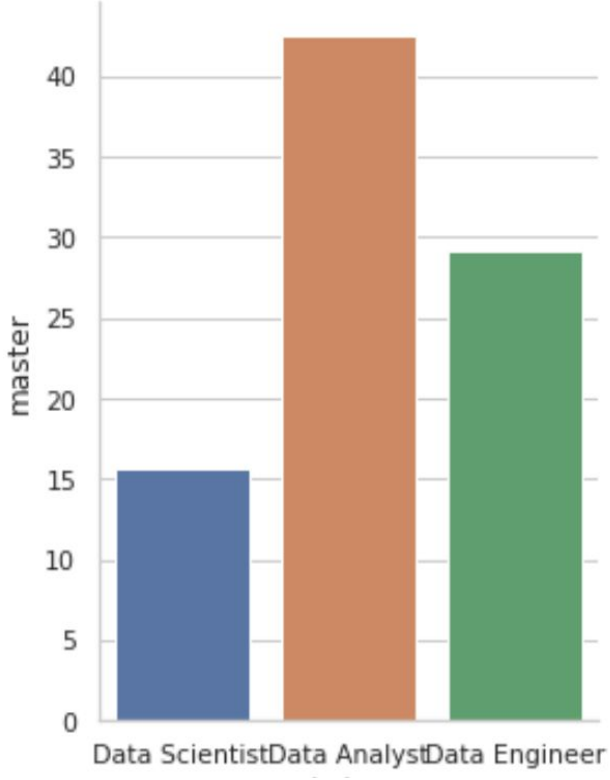
Sum of Keywords Frequency:
Data Analysis, Data Visualization, Research

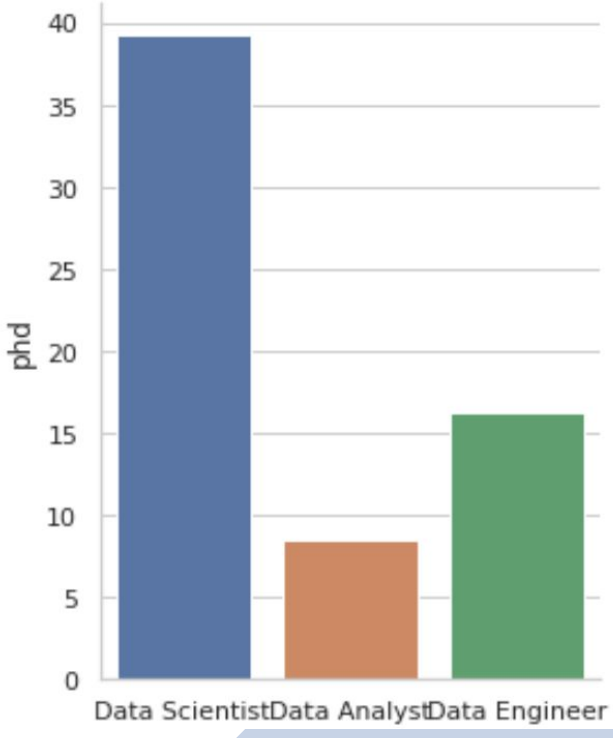


Sum of Keywords Frequency:
Machine Learning, Modeling, Statistical Analysis

Degree Requirement For Each Job Position







5

Supervised Machine Learning - Predictive Model

Preprocessing Steps Before Running Models

- We are switching to use **“indeed dataset”** for predictive models
  - Three job position types (Data Analyst/Data Scientist/Data Engineer)
- Preprocessing steps:
  - Remove unnecessary columns
  - Analysis on different salary range
- Data Cleaning
  - Fill blank values with NAs for Company Revenue/Company Employees
  - Replace NAs with 0 for Number of Reviews/Number of Stars
  - Convert to dummy variables
  - Rename column names

Supervised Machine Learning

- Split Train/Test Set ⟶ 80%/20%
- Use 10-fold cross validation
- Run the following models:
  - Linear Discriminant Analysis (LDA)
  - Stepwise Regression
  - K-Nearest Neighbors (KNN)
  - Support Vector Machines (SVM) with a linear Kernel
  - Random Forest (RF)
  - Boosted Trees
- Check RMSE and accuracy score for each model

Best Model - Random Forest

- Estimate the best model on a testing dataset
- Best Model: **Random Forest**
  - Lowest RMSE score
  - Highest accuracy of 86%

Create Prediction Based on Survey Responses

- Anonymous survey
- Received 60 Responses
- Students who studied Business Analytics/Data Science
  - Technical Skills
  - Location Preference
  - Industry Preference
  - Most fitted job title
  - A company that they want to work for

Salary Prediction (2 minutes)

Hi Everyone!

Thanks in advance for your time.

We are currently working on a data analyst/scientist/engineer salary predictive model for our capstone project. Since many of us are looking for jobs in the field of data science, we would love to invite you to take the following anonymous survey to help us predict MSBA students' prospective salary.

MSBA Cohort B Team 3

* Required

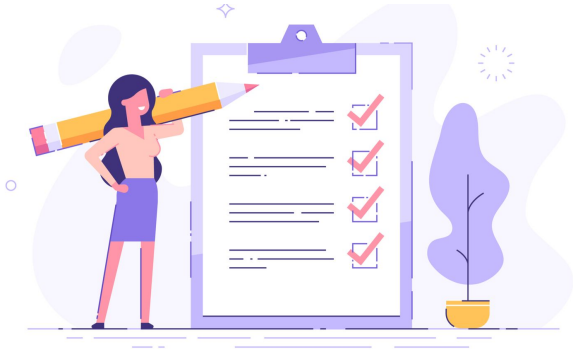
What technical skills do you have? (Choose more than 1) If you chose Other, please list all the other technical skills you have (

Request edit access

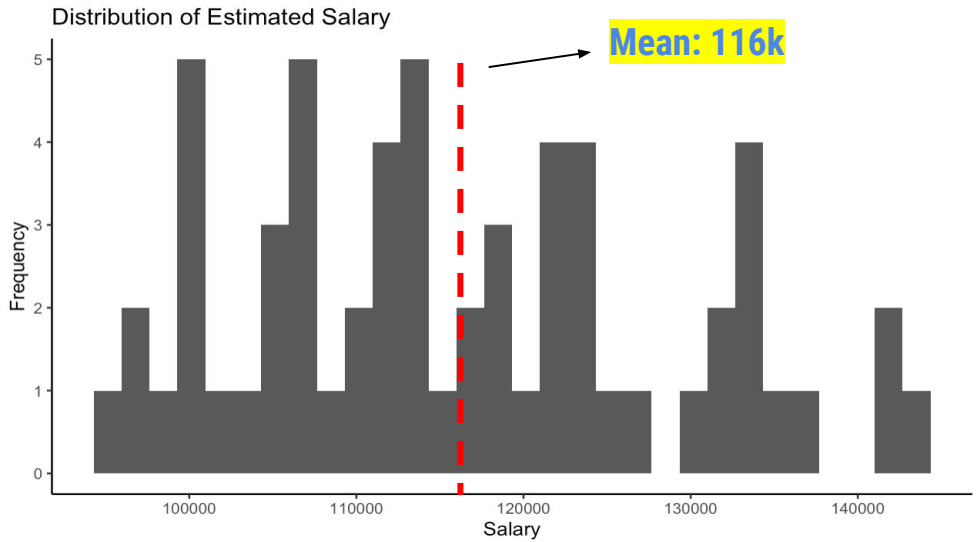
Overview on Survey Responses

Preferences are clear among our survey takers:

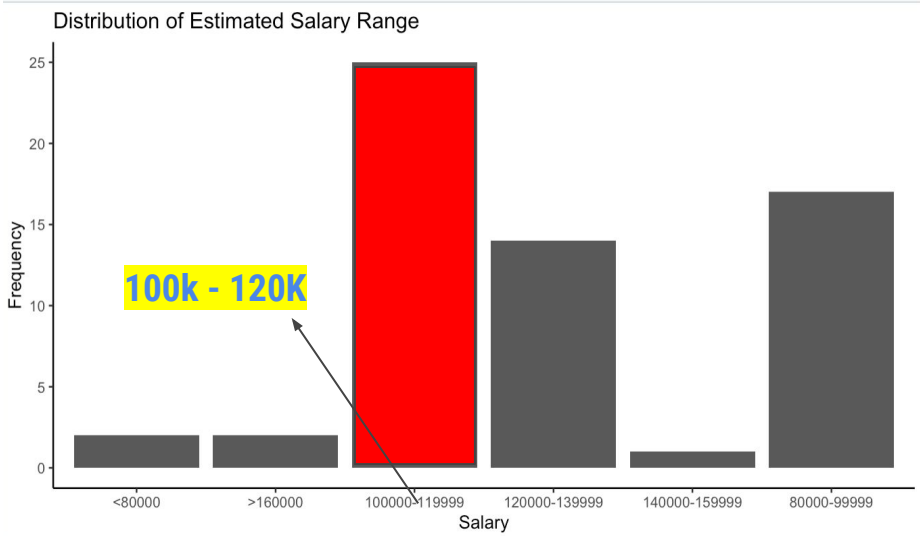
- Skills
  - 73.33% have at least 5 technical skills
  - Python, SQL, Machine Learning, R, Tableau
- Locations
  - NY, MA, CA
- Industries
  - Consulting & Business Services
  - Internet & Software
  - Banking & Financial Services



Results of prediction model based on survey



- Regression: Salary Value ($)
  - Range: 90K - 142K



- Classification: Salary Range ($)

6

Conclusion & Recommendation

Salary Prediction for MSBA students

- Profile
  - Skills: Python, R, SQL, Tableau, ML
  - Location: NY
  - Industry: Consulting
  - Positon: Data Analyst
  - Company: NA
- Prediction
  - Numeric value: $ 114, 208.33
  - Range: $ 100,000 - 119,999

Conclusion

- 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

7

Limitations & Challenges

😐 Challenges We've Faced Throughout This Process

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

THANKS!

Any Questions?