



SPOTTING A FAKE JOB POSTING

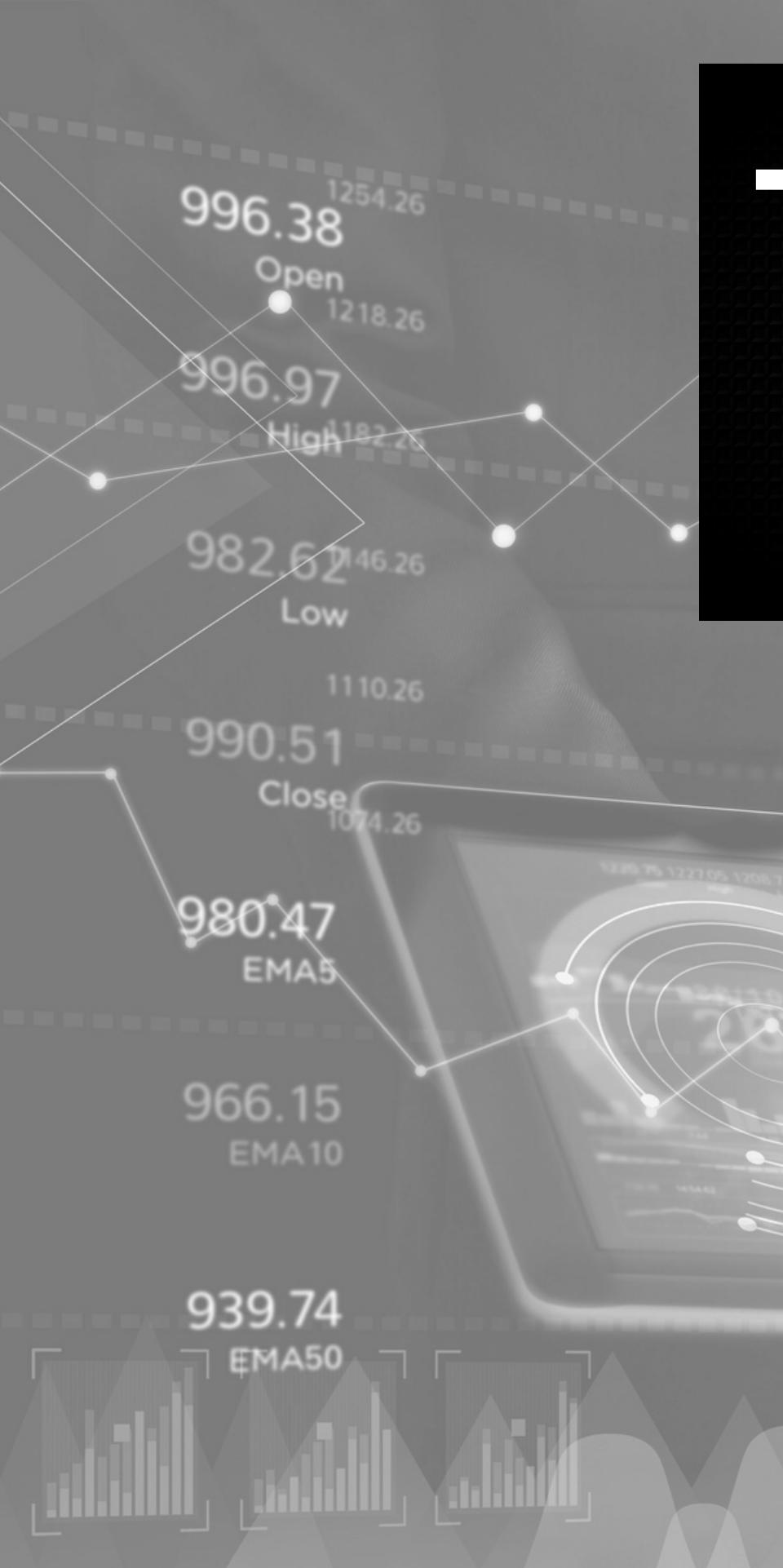
Pender Fitzgerald, Marci Morrell, Leticia Rinaldini



Outline



The Data
Project Goal
General Trends
Project Tools
Analyzation
Takeaways
Solutions to Spot a Fake



THE DATA

- Using a dataset found on Kaggle
- Consists of nearly 18,000 job postings, with a percentage of them being fraudulent postings
- Includes information on each job posting such as job title, location, department, salary, company profile, job description, requirements, benefits, company logo, employment type, whether the posting is fraudulent or not, and more

Project Goal

To create a model with a high accuracy of predicting whether or not a job posting is fraudulent.

General Trends

- 96% OF THE POSTINGS HAD AT LEAST 1 MISSING VALUE
- EVERY SINGLE JOB HAD VALUES FOR JOB DESCRIPTION AND JOB TITLE
- SALARY RANGE WAS THE MOST COMMON MISSING VALUE (84% OF THE COLUMN WAS NAN)



Real Job Postings

95%

Fraudulent Job Postings

5%

PROJECT TOOLS

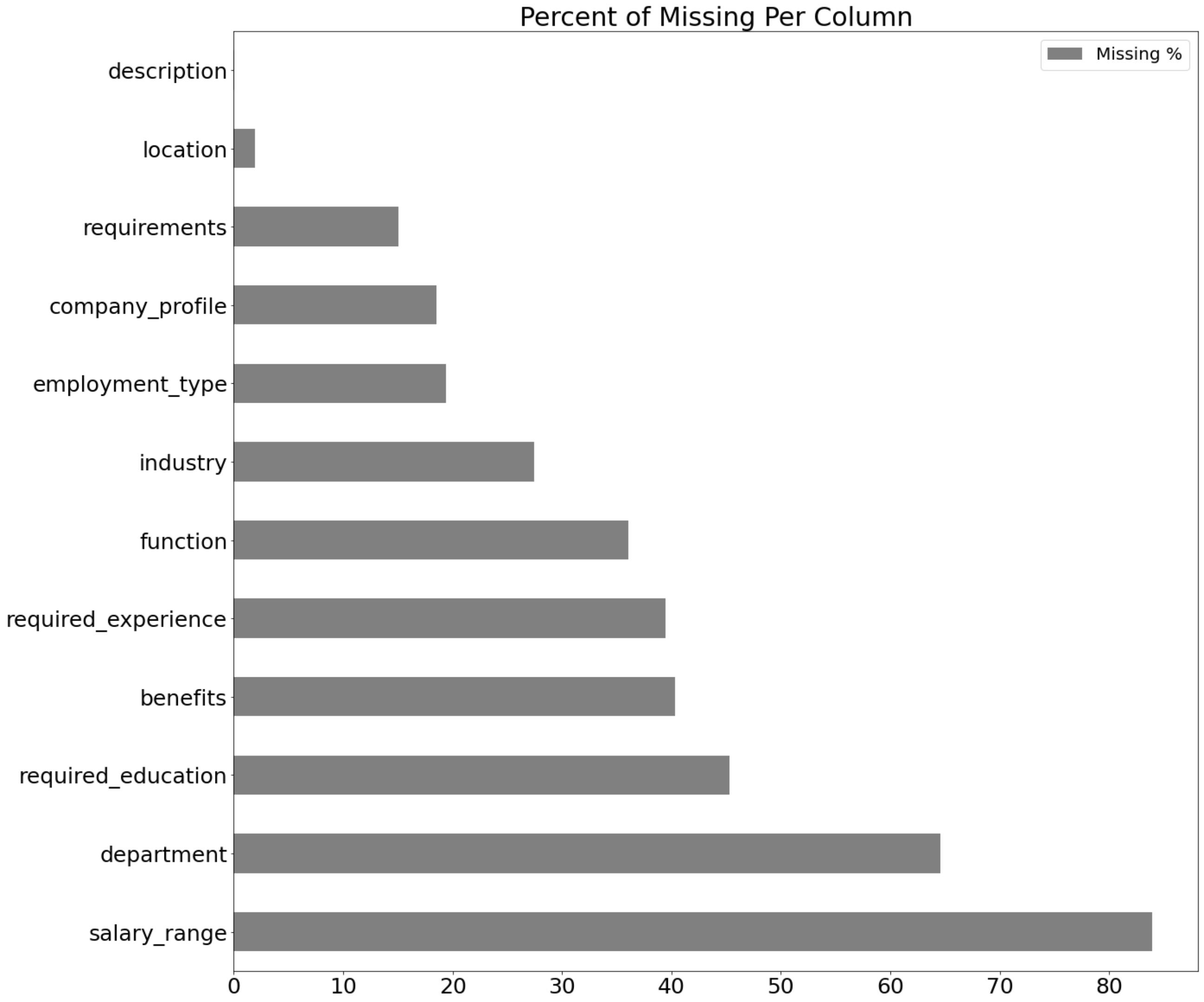


OUR PROJECT WAS ANALYZED
USING JUPYTER NOTEBOOKS TO
CLEAN AND ANALYZE DATA AS
WELL AS PLOT VISUALS AND
CREATE OUR MODEL.

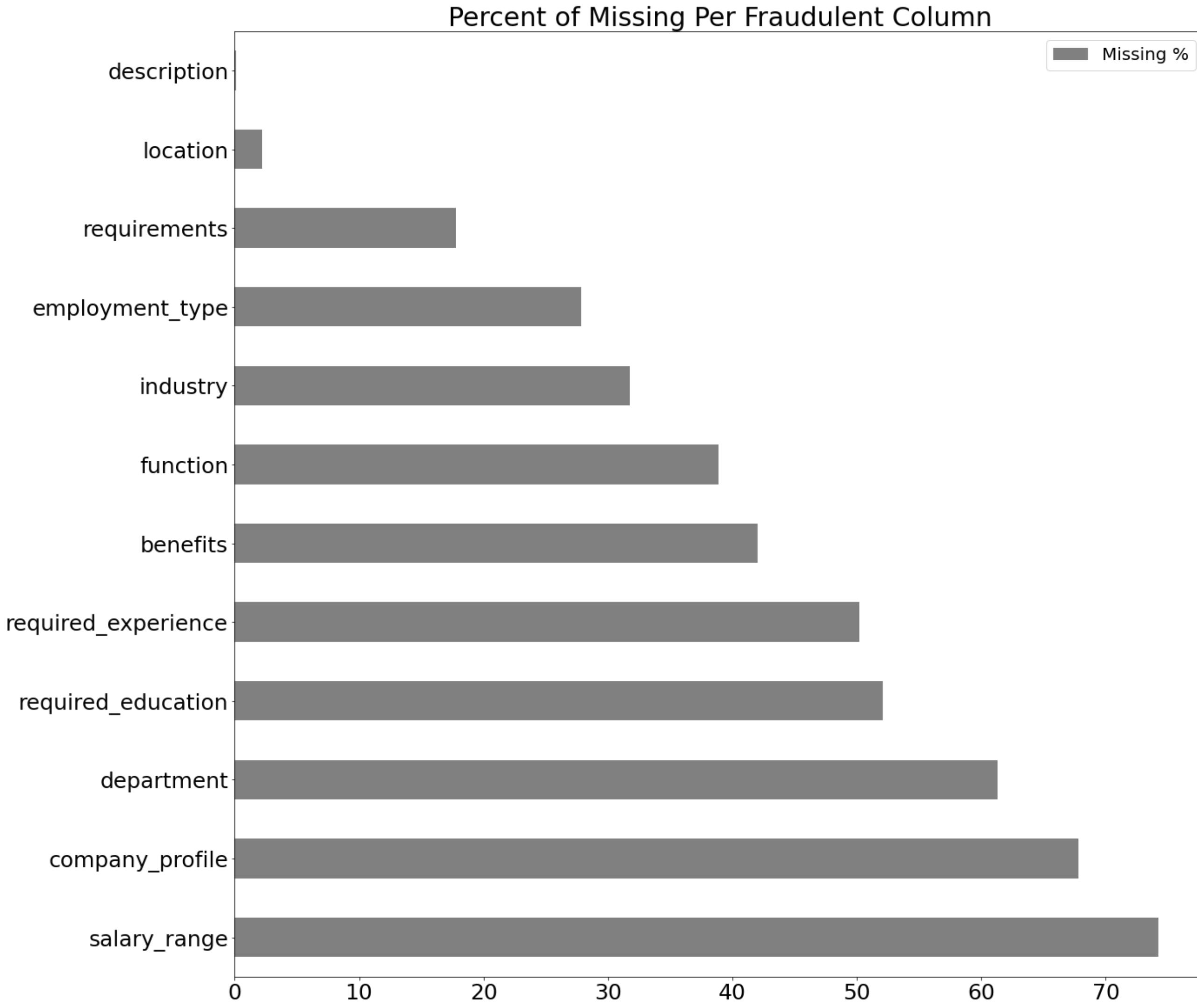


MISSING DATA

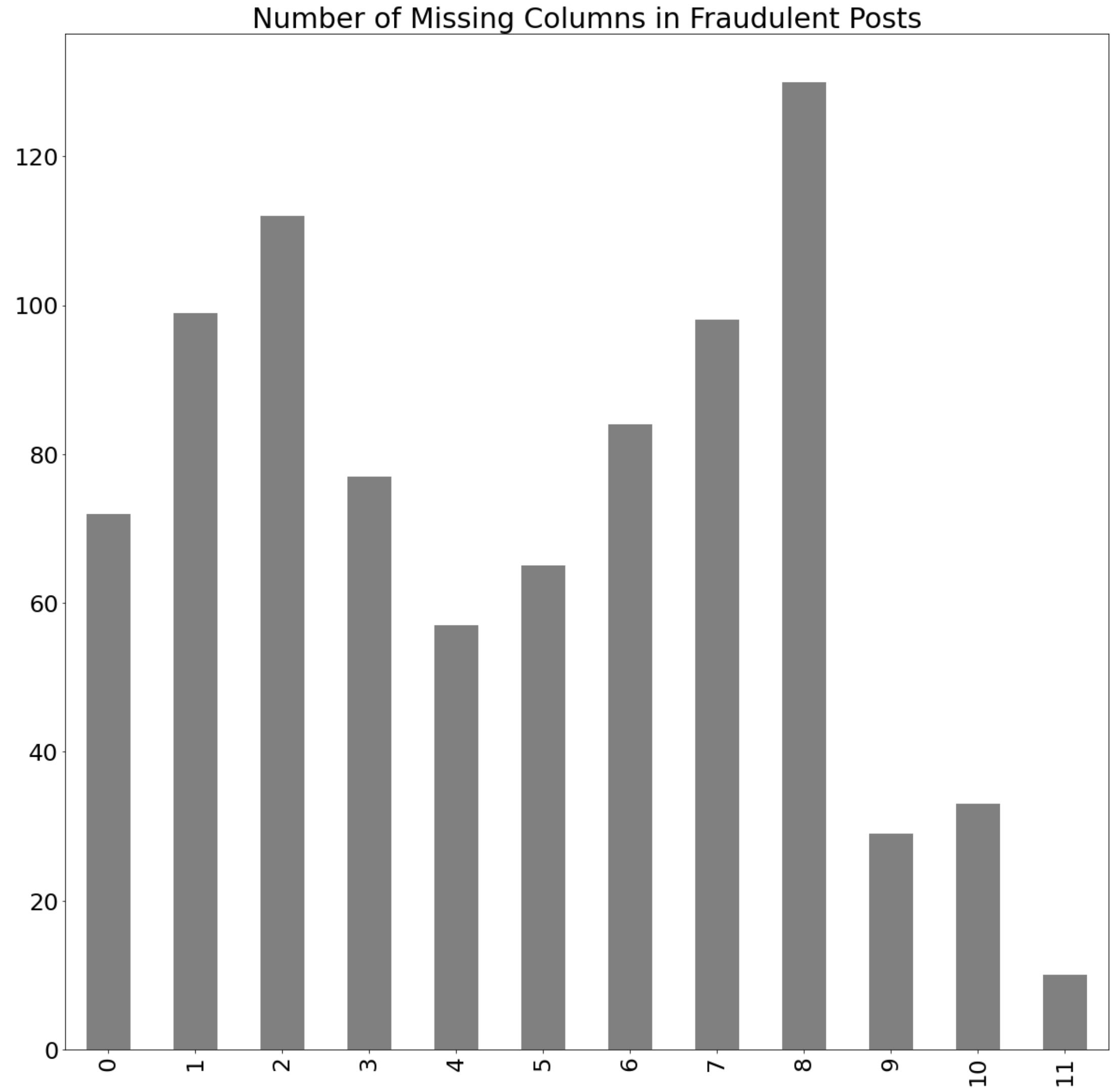
Percent of missing values per column



Percent of missing values in each column of fraudulent jobs

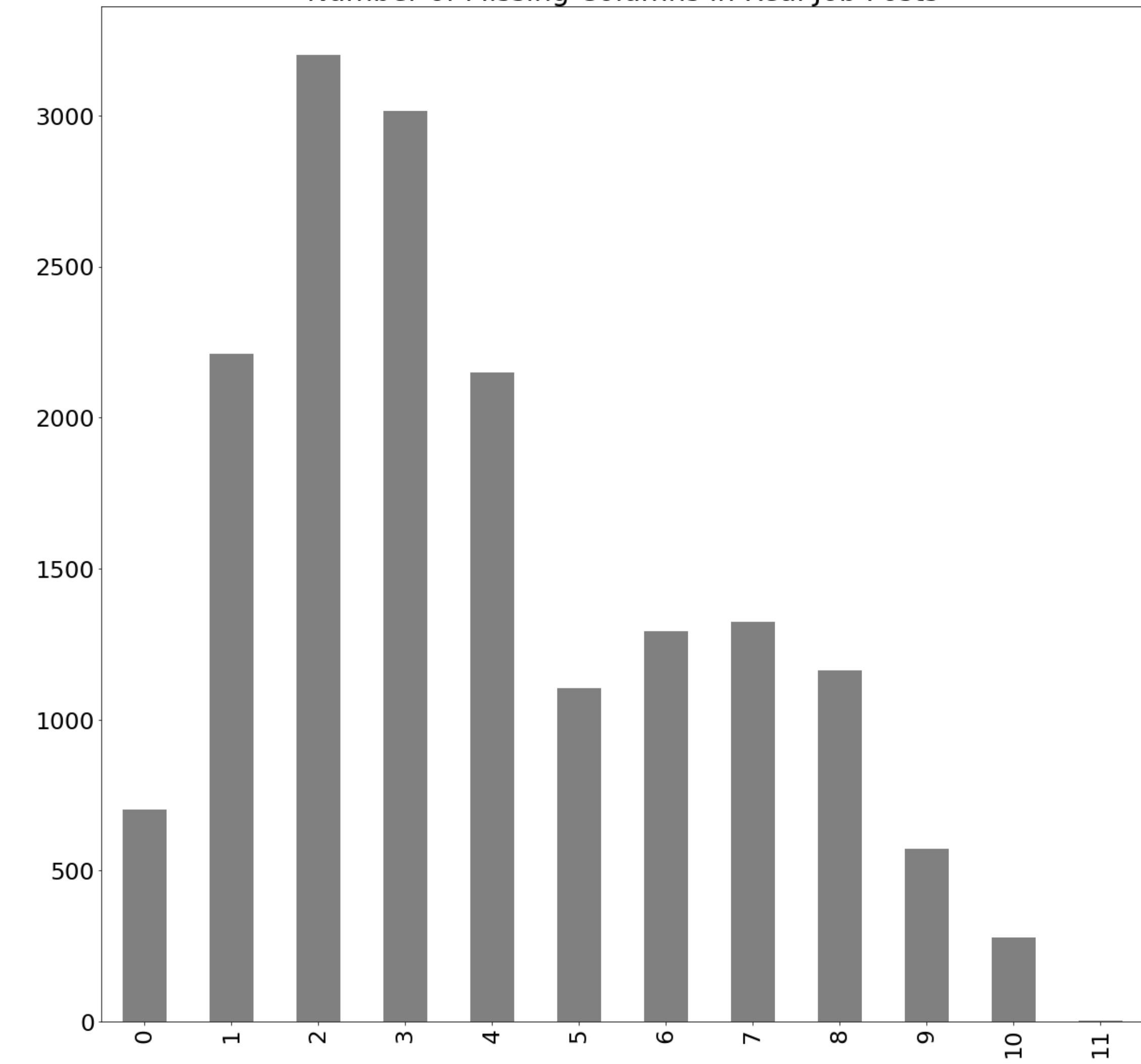


Number of missing columns in fraudulent posts



Number of missing columns in legitimate posts

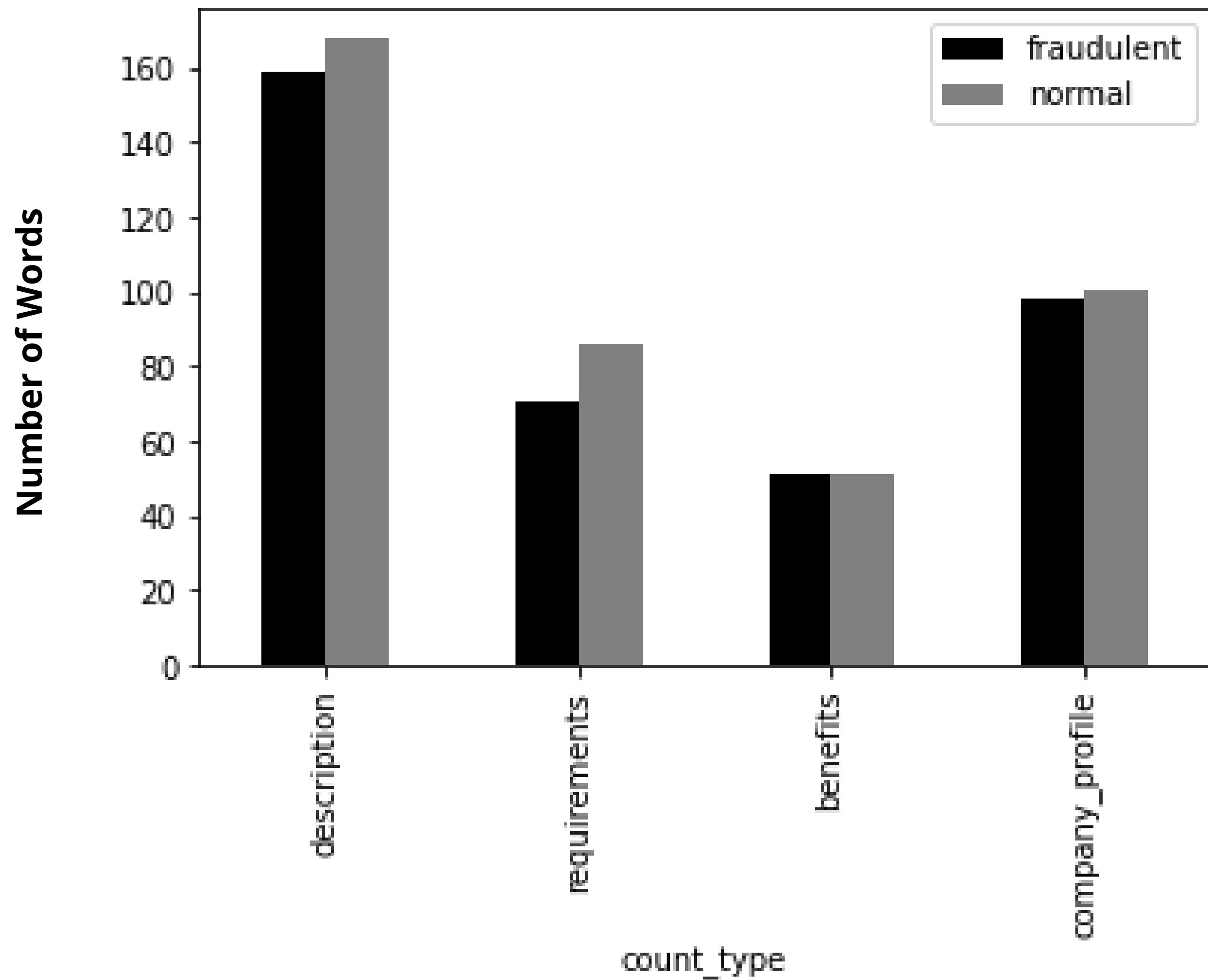
Number of Missing Columns in Real Job Posts



WORD COUNT

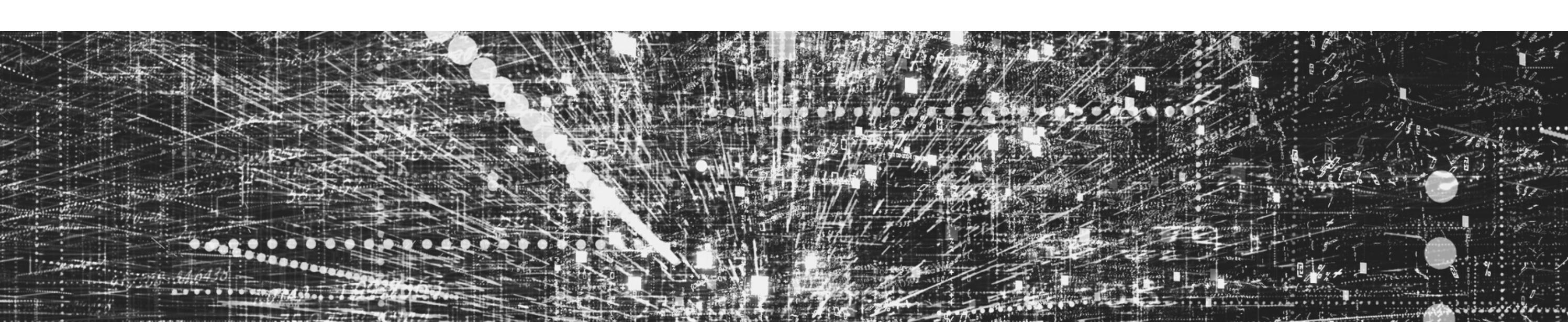
?

Word Count by Job Type



MODELS





Correlation Coefficients

company_profile	-0.390535
description	-0.040879
requirements	-0.138523
benefits	-0.025213
telecommuting	0.080746
has_company_logo	-0.489713
has_questions	-0.207180
employment_type	-0.042686
required_experience	-0.068714
required_education	0.034008
fraudulent	1.000000

LoCoPro Model

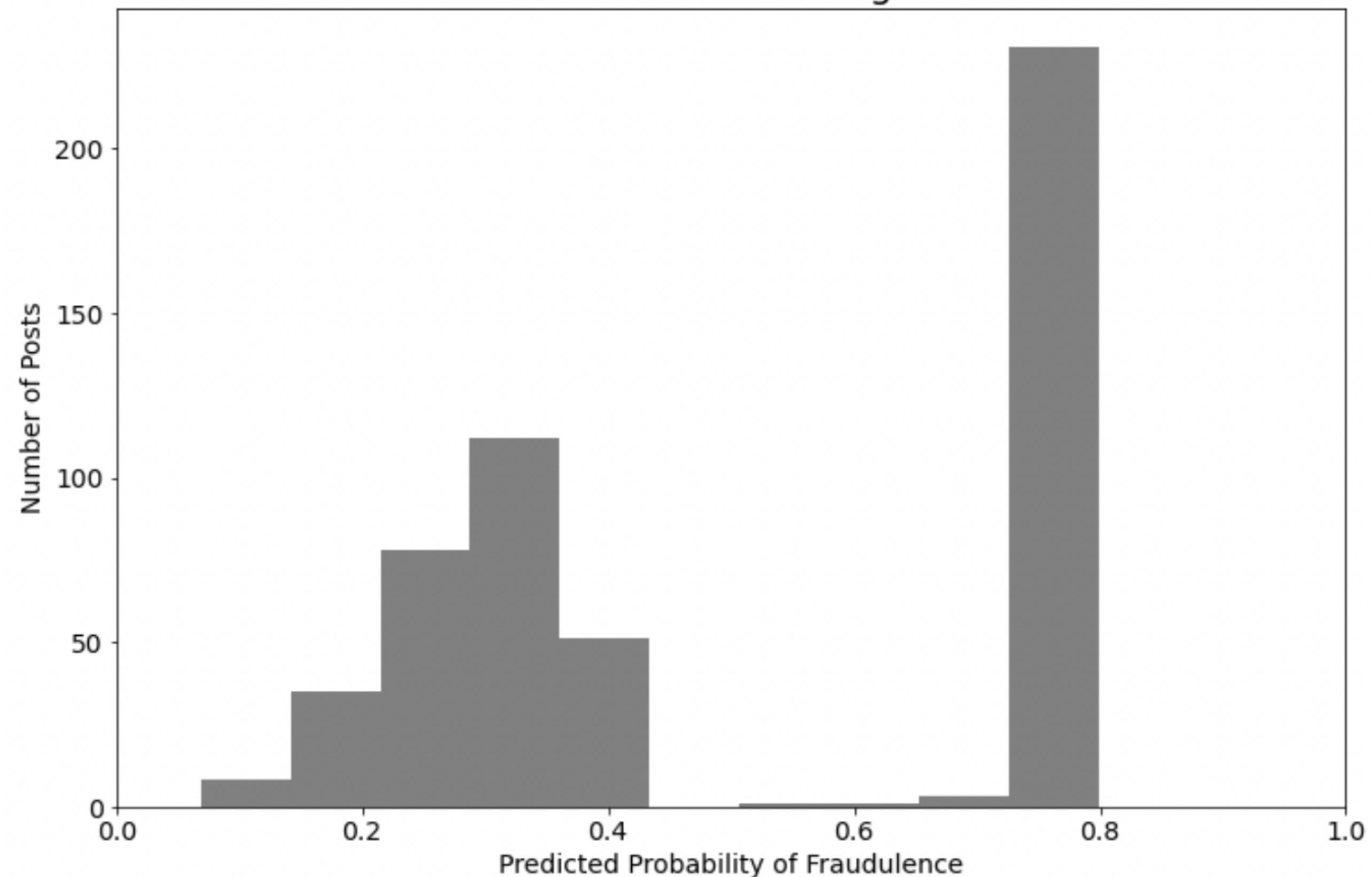
- LOGO, COMPANY PROFILE COLUMNS

Logo/Company Profile: 0.74
precision

0	0.72
1	0.78

Predicted Probabilities For Fraud. Postings

Fraudulent Postings

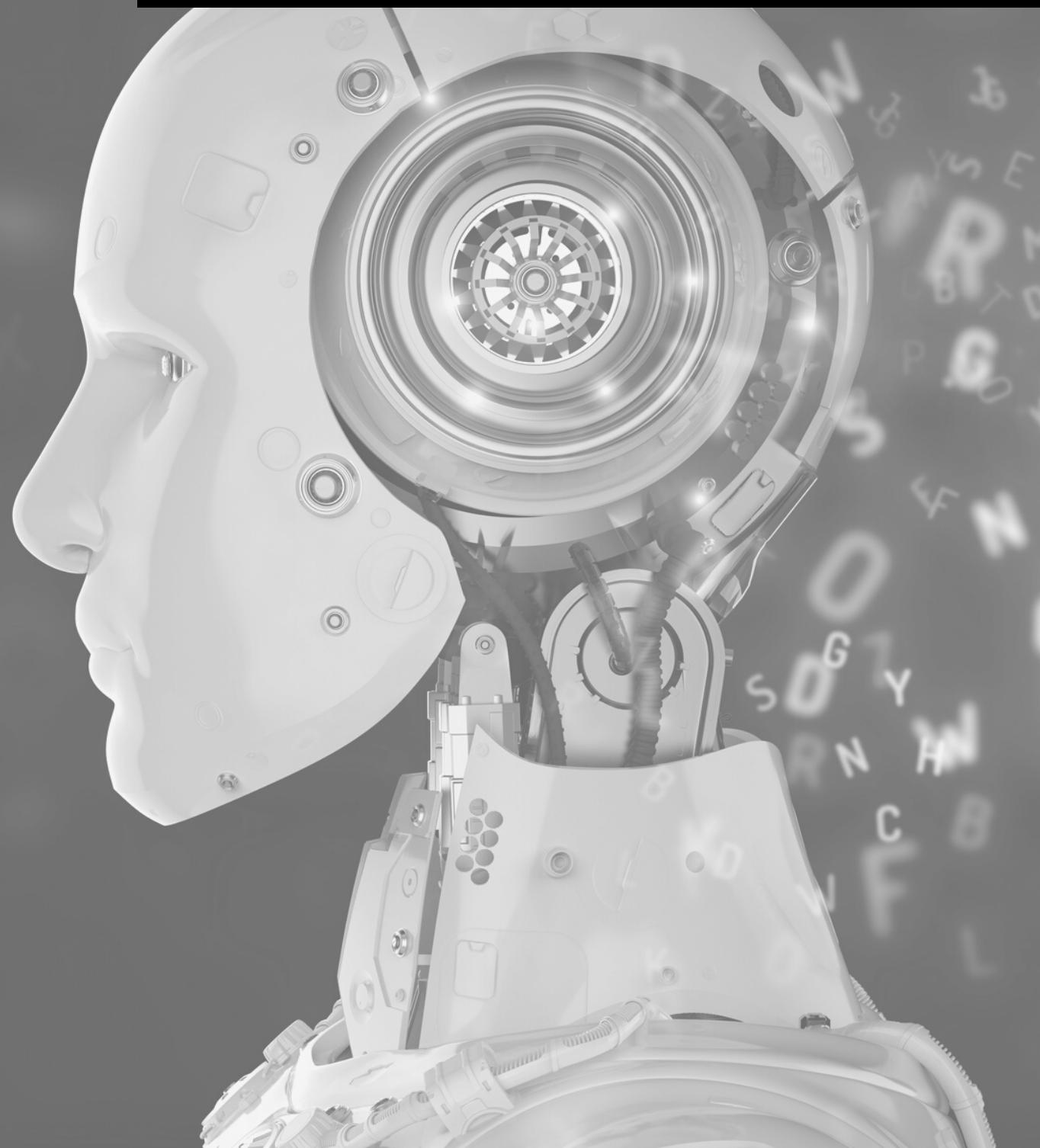


Power Set Model

- USES COLUMNS DESCRIPTION_COUNT, COMPANY_PROFILE_COUNT, HAS_SALARY_RANGE, HAS_EMPLOYMENT_TYPE, DEPARTMENT_ID

accuracy	0.75
precision	
0	0.78
1	0.73

TAKEAWAYS



LOGISTIC REGRESSION MODEL

With the data given, the best models can predict fraudulent job posts between 75% and 80% of the time

11+ MISSING

We found that if a job posting is missing 11 or more columns it is most likely fraudulent.

WORD COUNT

Most job postings that were real had a slightly higher word count for job description, company profile, and job requirements.

SALARY RANGE

Salary range was the most common missing column. Of the real job postings, 84% did not have salaries listed. The fraudulent posts had more salary ranges, and a lot were unrealistic.

Solutions

Predicting a fraudulent job posting

MISSING INFO	WORD COUNT	SALARY	COMPANY LOGO	COMPANY PROFILE
Missing information did not hugely contribute to predicting fraudulence except when reaching 11+ missing columns.	Using word count for certain columns, we were able to attain the highest accuracy (.75) using a powerset model.	Posts with unrealistic salaries listed are more likely to be fraudulent.	The Company Logo column has the highest correlation coefficient with being fraudulent.	Real jobs only had around 20% of the company profile column missing, where fake jobs had 70%.

Thank you for listening!

**We hope you can derive your own
observations and conclusions from our work.**

QUESTIONS?

ASK AWAY!

