Project Fake Finder - Progress Report 1

April 29, 2025

Project Progress Report # 1

Team Members: John DeLeon, Darshan Joshi, Dae-Breona Armour

CS 6330-DS1 Data Science | Spring B 2025

Dr. Dogdu | Angelo State University

Department of Computer Science

1 Data Exploratory Data Analysis (EDA)

1.1 Environment Setup for EDA

Before we start our Exploratory Data Analysis, there are some additional packages we will need to install.

- kagglehub - Import our Kaggle dataset
- textstat - To gain additional insight with our text fields

1.1.1 Purpose:

Additionally, we want to update some of the default pandas options and set basic logging configuration to get a better view of our dataset during our data exploration.

1.2 Data Wrangling

1.2.1 Step 1: Discovering

Discovery, also called data exploration, familiarizes the data scientist with source data in preparation for subsequent steps.

1.2.2 Purpose:

Get the first 5 rows of our dataset to make sure it loaded correctly and that there are no issues. Since the text columns has a lot of data we view those separately in an additional code cell.

	job_id	title	location	\
0	1	Marketing Intern	US, NY, New York	
1	2	Customer Service - Cloud Video Production	NZ, , Auckland	
2	3	Commissioning Machinery Assistant (CMA)	US, IA, Wever	
3	4	Account Executive - Washington DC	US, DC, Washington	

```
4       5                        Bill Review Manager  US, FL, Fort Worth

  department salary_range  telecommuting  has_company_logo  has_questions  \
0  Marketing          NaN              0                 1              0
1    Success          NaN              0                 1              0
2        NaN          NaN              0                 1              0
3      Sales          NaN              0                 1              0
4        NaN          NaN              0                 1              1

  employment_type required_experience required_education  \
0           Other          Internship                NaN
1       Full-time      Not Applicable                NaN
2             NaN                NaN                NaN
3       Full-time    Mid-Senior level  Bachelor's Degree
4       Full-time    Mid-Senior level  Bachelor's Degree

                    industry              function  fraudulent
0                        NaN             Marketing           0
1  Marketing and Advertising      Customer Service           0
2                        NaN                   NaN           0
3          Computer Software                 Sales           0
4     Hospital & Health Care  Health Care Provider           0

   job_id  \
0       1

company_profile  \
0  We're Food52, and we've created a groundbreaking and award-winning cooking
site. We support, connect, and celebrate home cooks, and give them everything
they need in one place.We have a top editorial, business, and engineering team.
We're focused on using technology to find new and better ways to connect people
around their specific food interests, and to offer them superb, highly curated
information about food and cooking. We attract the most talented home cooks and
contributors in the country; we also publish well-known professionals like Mario
Batali, Gwyneth Paltrow, and Danny Meyer. And we have partnerships with Whole
Foods Market and Random House.Food52 has been named the best food website by the
James Beard Foundation and IACP, and has been featured in the New York Times,
NPR, Pando Daily, TechCrunch, and on the Today Show.We're located in Chelsea, in
New York City.

                 description  \
0  Food52, a fast-growing, James Beard Award-winning online food community and
crowd-sourced and curated recipe hub, is currently interviewing full- and part-
time unpaid interns to work in a small team of editors, executives, and
developers in its New York City headquarters.Reproducing and/or repackaging
existing Food52 content for a number of partner sites, such as Huffington Post,
Yahoo, Buzzfeed, and more in their various content management systemsResearching
```

```
blogs and websites for the Provisions by Food52 Affiliate ProgramAssisting in
day-to-day affiliate program support, such as screening affiliates and assisting
in any affiliate inquiriesSupporting with PR &amp;amp; Events when neededHelping
with office administrative work, such as filing, mailing, and preparing for
meetingsWorking with developers to document bugs and suggest improvements to the
siteSupporting the marketing and executive staff

                                                 requirements  \
0  Experience with content management systems a major plus (any blogging
counts!)Familiar with the Food52 editorial voice and aestheticLoves food,
appreciates the importance of home cooking and cooking with the
seasonsMeticulous editor, perfectionist, obsessive attention to detail, maddened
by typos and broken links, delighted by finding and fixing themCheerful under
pressureExcellent communication skillsA+ multi-tasker and juggler of
responsibilities big and smallInterested in and engaged with social media like
Twitter, Facebook, and PinterestLoves problem-solving and collaborating to drive
Food52 forwardThinks big picture but pitches in on the nitty gritty of running a
small company (dishes, shopping, administrative support)Comfortable with the
realities of working for a startup: being on call on evenings and weekends, and
working long hours

  benefits
0      NaN
```

1.2.3 Results:

Our dataset we loaded seems to match the structure we expect from Kaggle and now we will keep doing further discovery.

1.2.4 Purpose:

Get a high-level summary of our data's structure.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17880 entries, 0 to 17879
Data columns (total 18 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   job_id               17880 non-null  int64
 1   title                17880 non-null  object
 2   location             17534 non-null  object
 3   department           6333 non-null   object
 4   salary_range         2868 non-null   object
 5   company_profile      14572 non-null  object
 6   description          17879 non-null  object
 7   requirements         15184 non-null  object
 8   benefits             10668 non-null  object
 9   telecommuting        17880 non-null  int64
```

```
 10  has_company_logo     17880 non-null  int64
 11  has_questions        17880 non-null  int64
 12  employment_type      14409 non-null  object
 13  required_experience  10830 non-null  object
 14  required_education   9775 non-null   object
 15  industry             12977 non-null  object
 16  function             11425 non-null  object
 17  fraudulent           17880 non-null  int64
dtypes: int64(5), object(13)
memory usage: 2.5+ MB
```

1.2.5 Results:

From using the "info()" function we understand that we have the 18 columns that we expect, the 17,880 rows we expect, and we understand at a high-level our data types, and how many "non-null" values we have for each column. Next let's dig deeper into the Null fields we have.

1.2.6 Purpose:

From order of greatest to least, we want to see what percentage of a given fields values are Null, we also want to start thinking about which fields we can probably use for our Data Science Model. If a field has too many Nulls or if real-world job posting don't have that data then we may suggest to drop those fields.

	Data Type	Missing Values	Missing %	Unique Values
salary_range	object	15012	83.96	874
department	object	11547	64.58	1337
required_education	object	8105	45.33	13
benefits	object	7212	40.34	6203
required_experience	object	7050	39.43	7
function	object	6455	36.10	37
industry	object	4903	27.42	131
employment_type	object	3471	19.41	5
company_profile	object	3308	18.50	1709
requirements	object	2696	15.08	11965
location	object	346	1.94	3105
description	object	1	0.01	14801
title	object	0	0.00	11231
job_id	int64	0	0.00	17880
telecommuting	int64	0	0.00	2
has_questions	int64	0	0.00	2
has_company_logo	int64	0	0.00	2
fraudulent	int64	0	0.00	2

1.2.7 Results:

Suggested that we Drop the following Columns later on: salary_range, department, required_education, benefits, required_experience, function, industry, location, job_id, has_questions, has_company_logo.

We reccommend droping the above fields because some of them aren't useful in identifying fraudulent jobs even if the columns didn't have so many nulls. For other columns, when looking at real world job posting, we don't always see this information and ideally we want the fields used in our training data to match the real-world as closely as possible.

1.2.8 Purpose:

Checking for Duplicate Rows of Data to see if we need to de-duplicate our dataset later on.

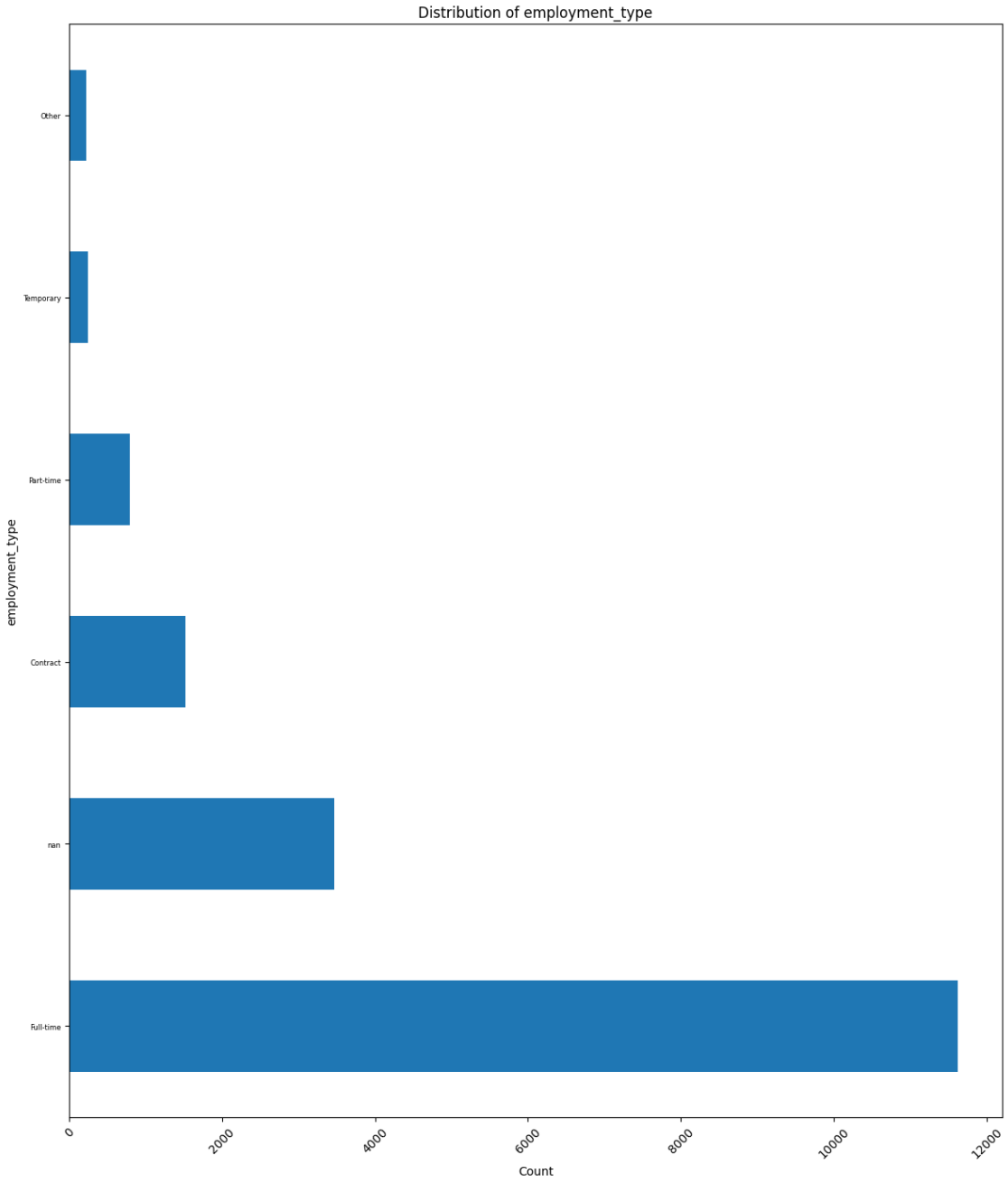
```
Has duplicates: False
Number of duplicate rows: 0
Duplicate rows:
Empty DataFrame
Columns: [job_id, title, location, department, salary_range, company_profile,
description, requirements, benefits, telecommuting, has_company_logo,
has_questions, employment_type, required_experience, required_education,
industry, function, fraudulent]
Index: []
```

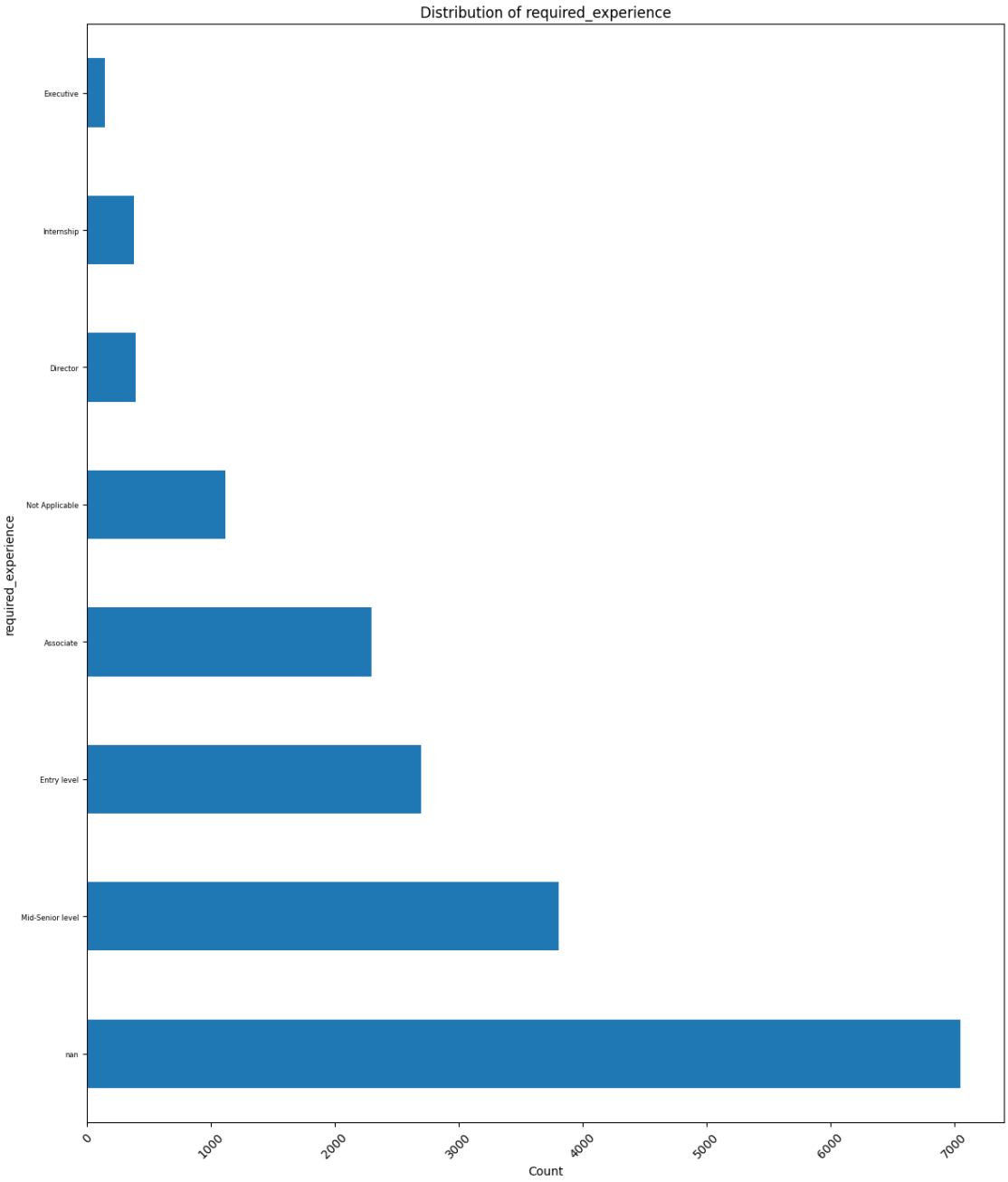
1.2.9 Results:

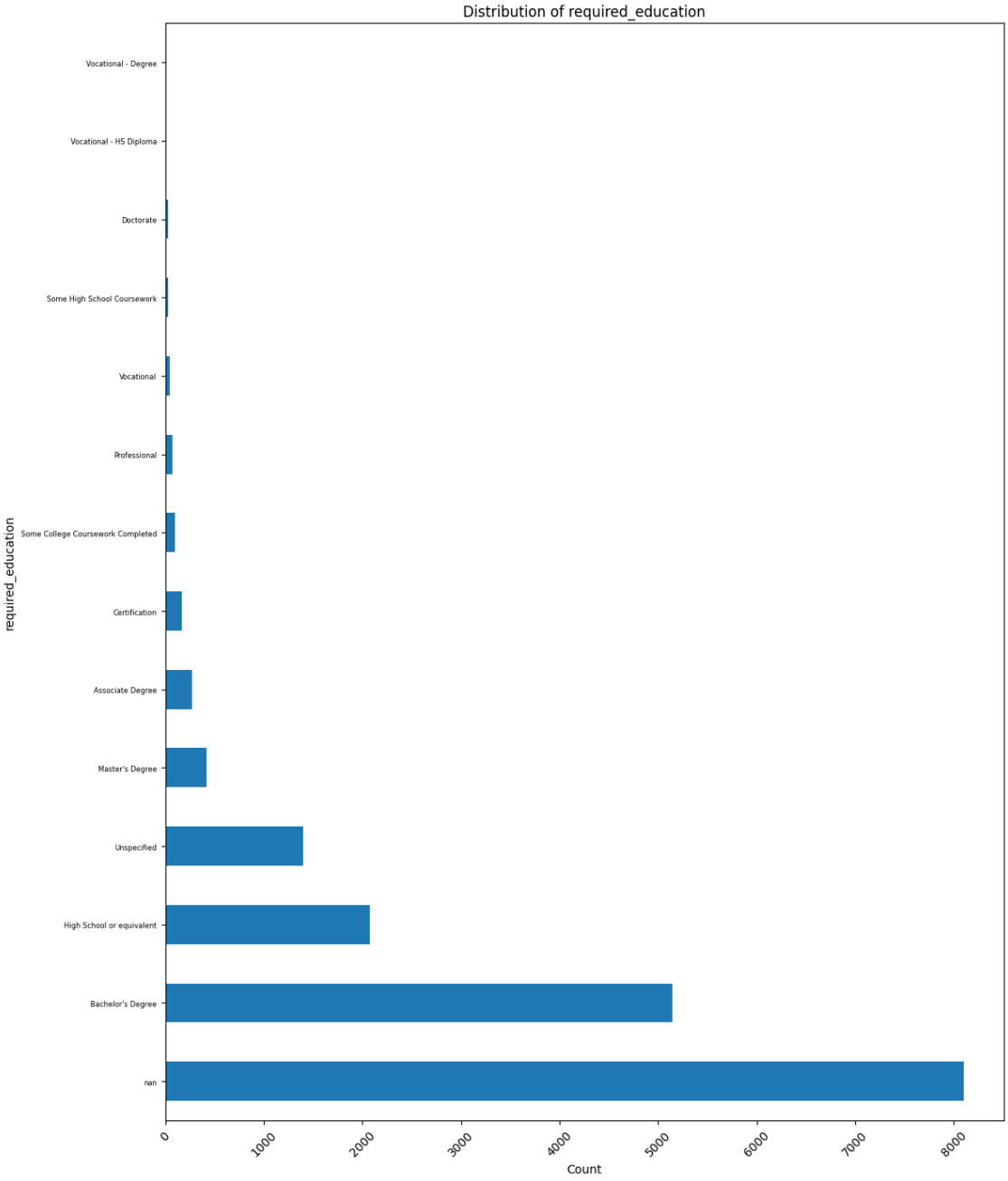
Seems like we don't have any duplicate rows of data in our dataset so we won't need to do any "de-duplication" in our "Cleaning" step.

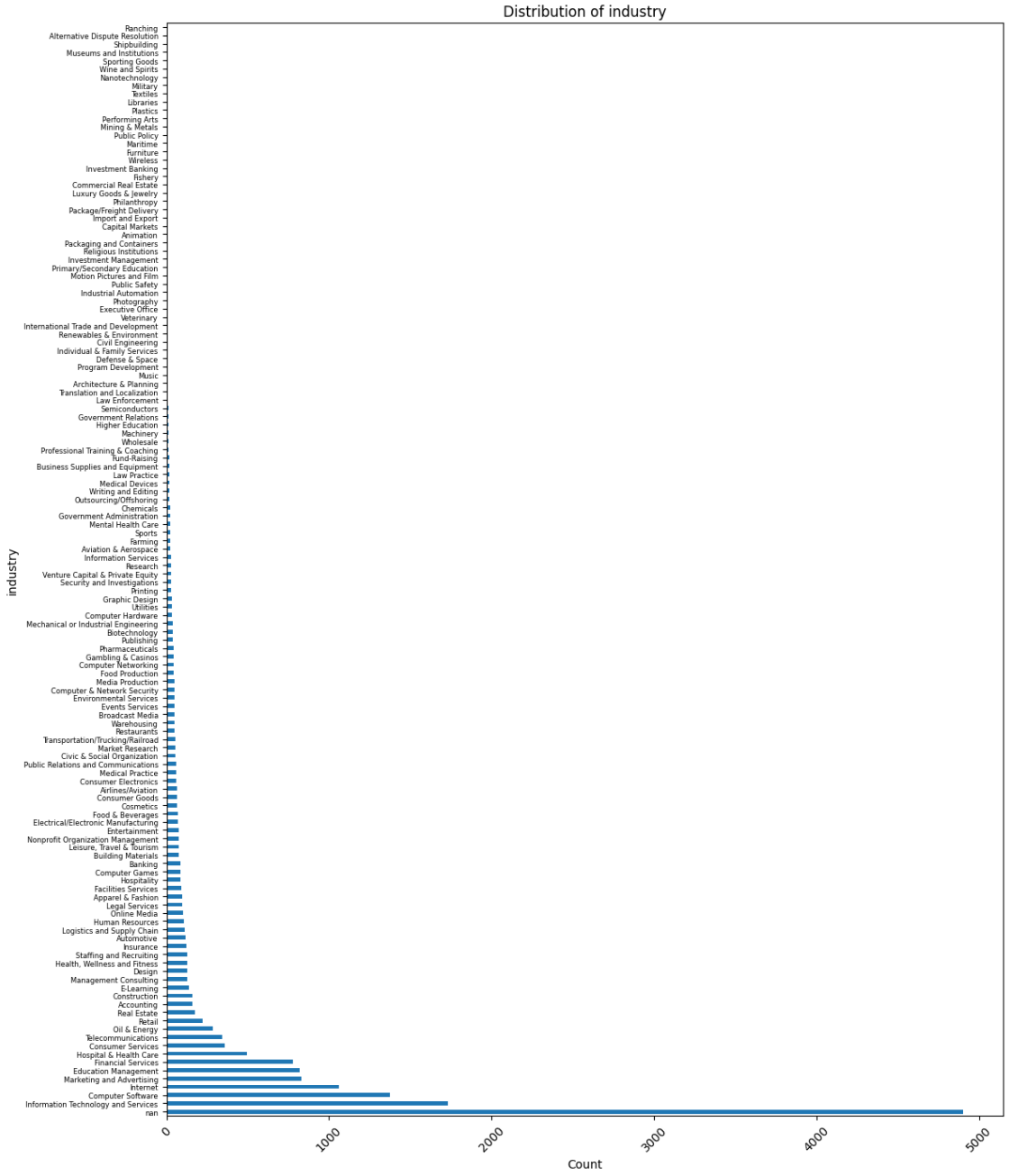
1.2.10 Purpose:

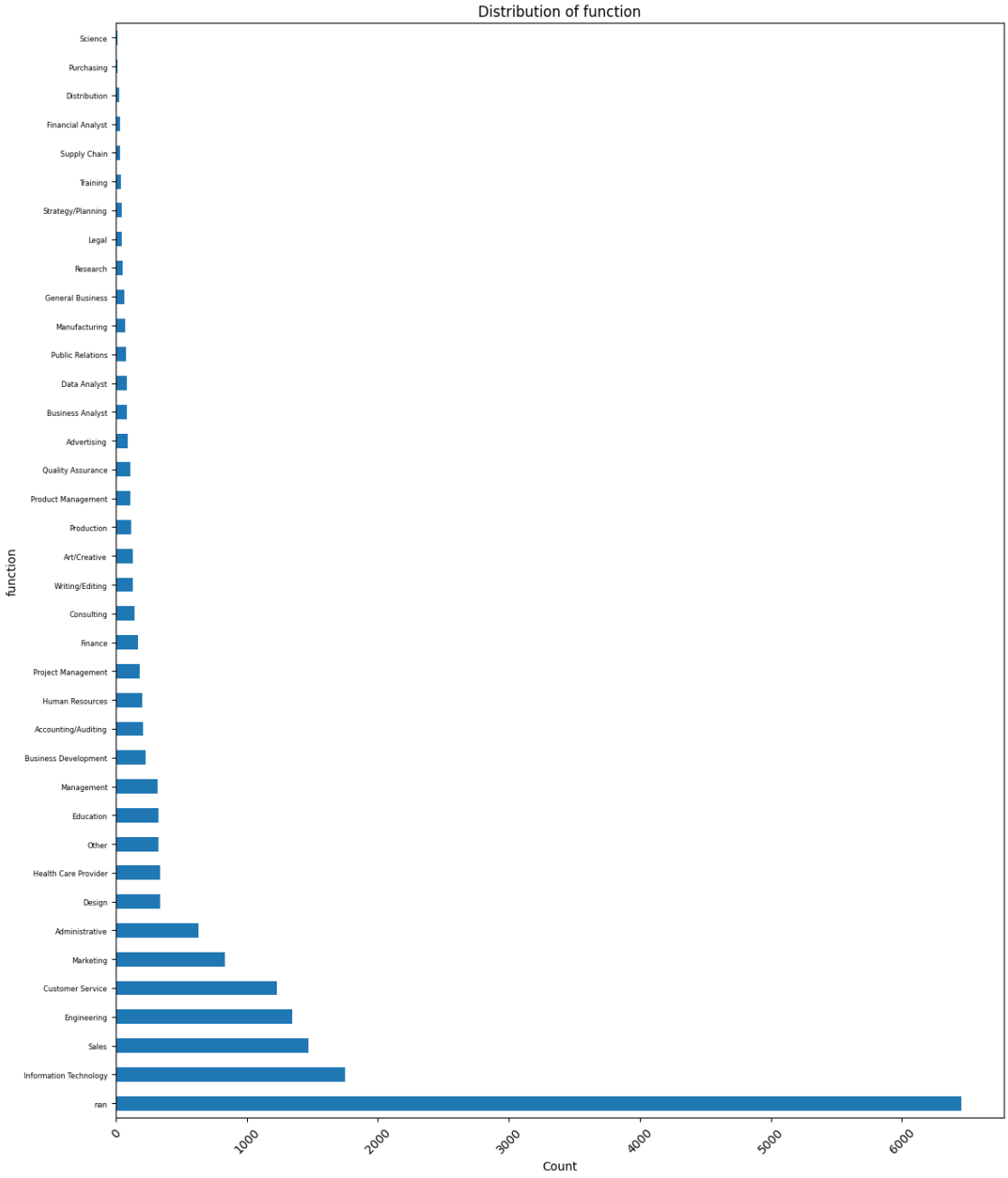
The purpose of the below function is to visualize the distribution of values in categorical and binary columns from the dataset. For each specified column, it generates a horizontal bar chart that shows how frequently each category or value appears. This helps in understanding the balance, skewness, and patterns within categorical variables, which is critical for exploratory data analysis and for identifying potential data issues (such as class imbalance).

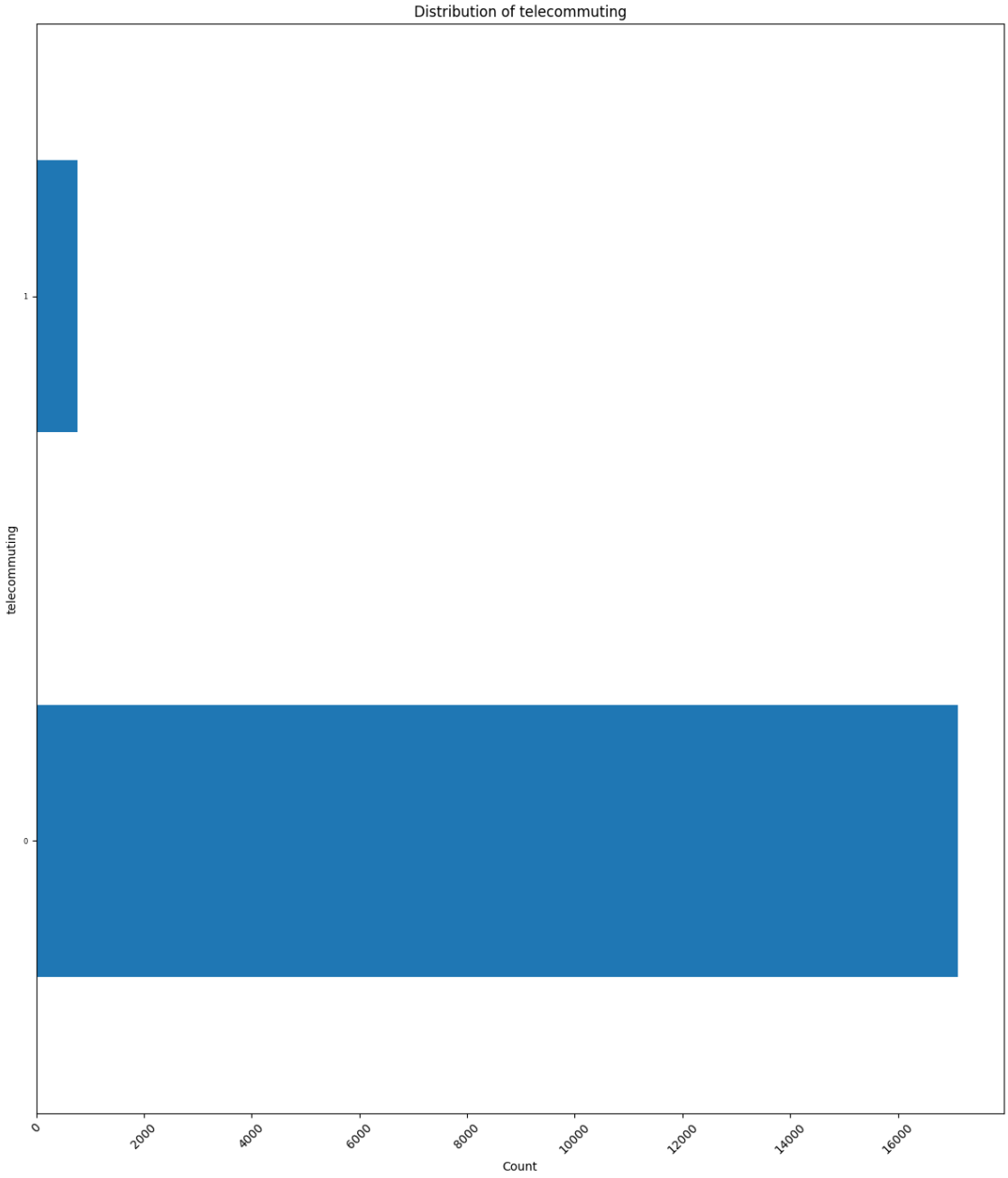

Distribution of employment_type
employment_type
Other
Temporary
Part-time
Contract
nan
Full-time
0
2000
4000
6000
8000
10000
12000
Count

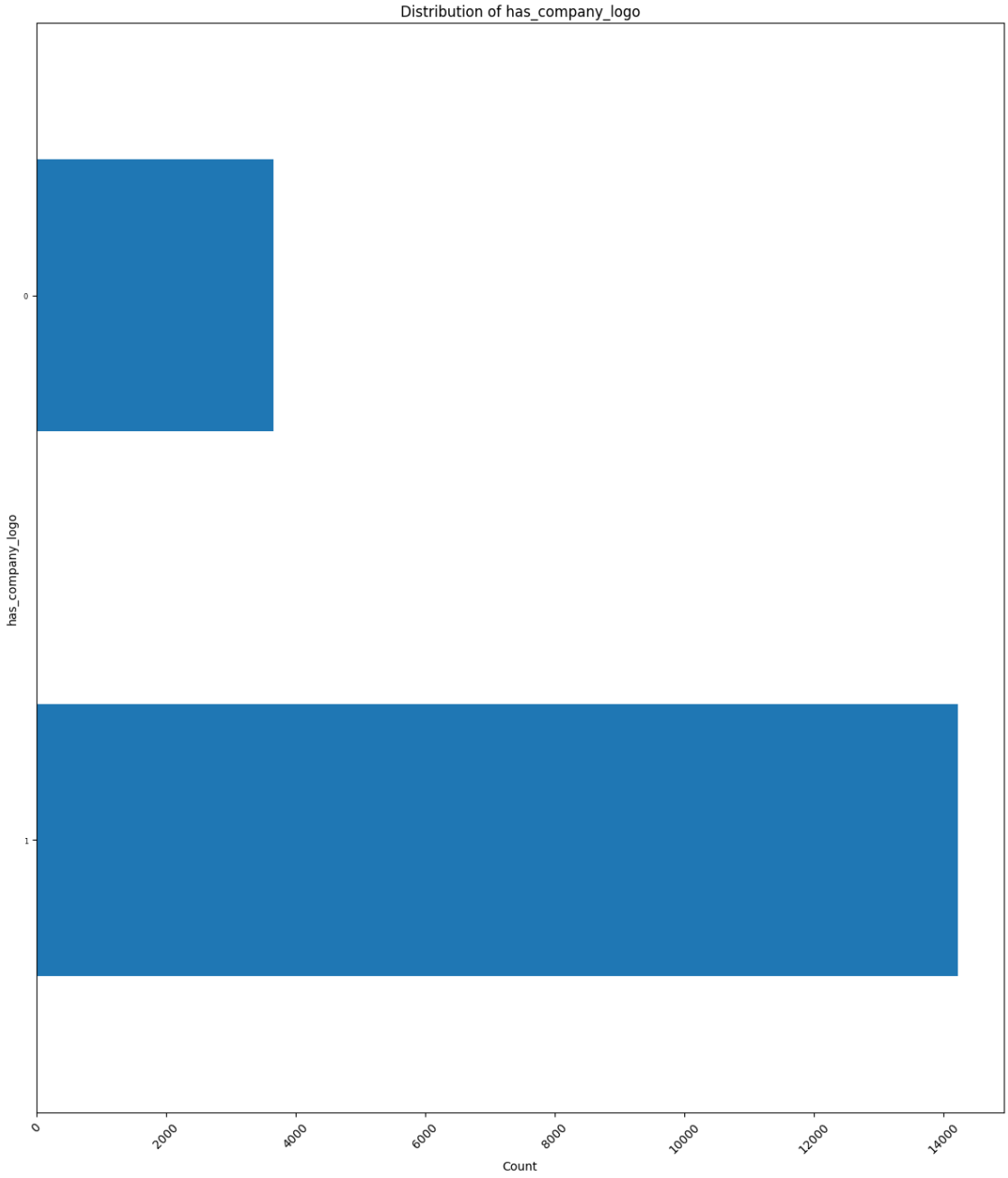

Distribution of required_experience
Executive
Internship
Director
Not Applicable
Associate
Entry level
Mid-Senior level
nan
required_experience
0
1000
2000
3000
4000
5000
6000
7000
Count

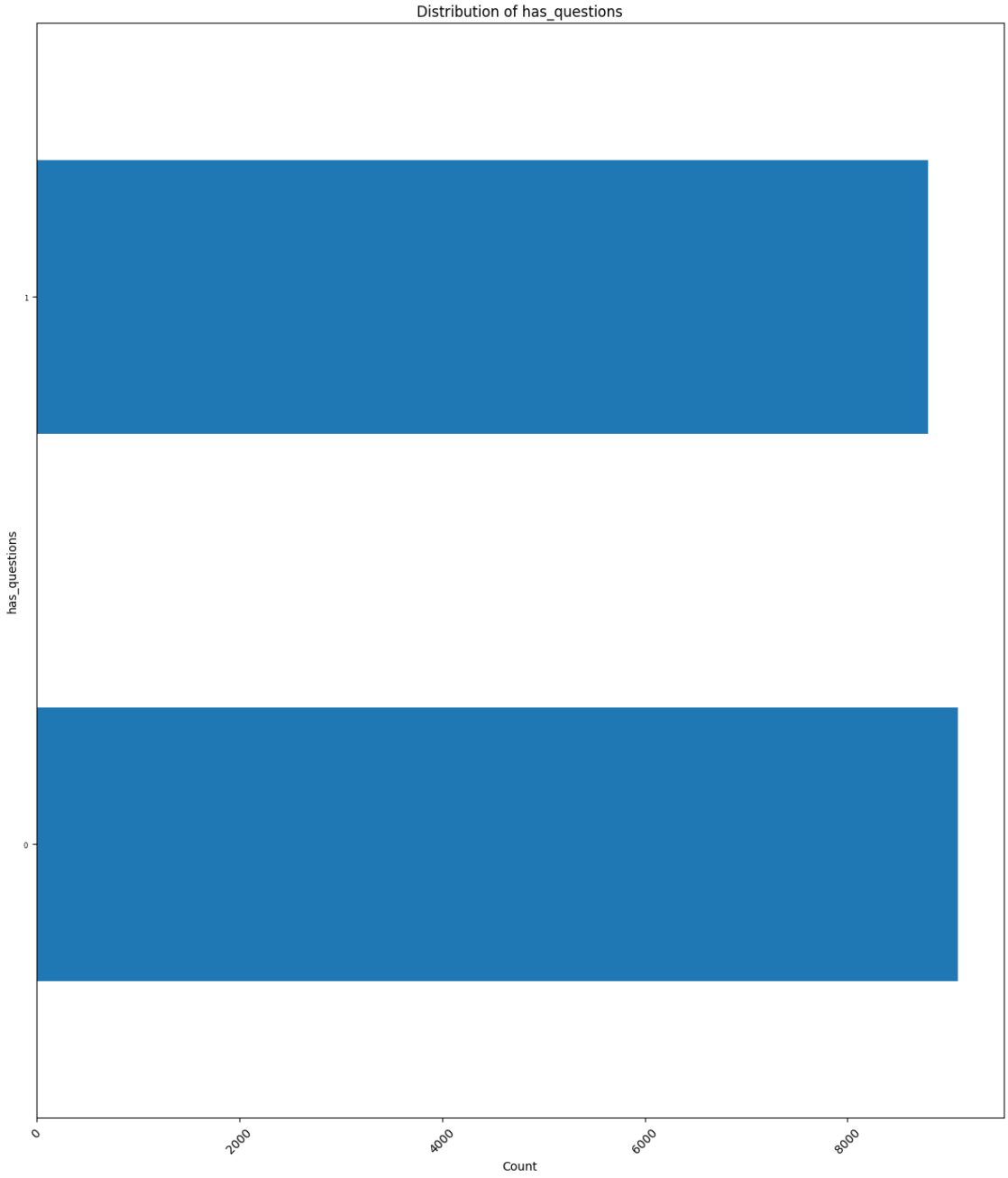

Distribution of required_education
Vocational - Degree
Vocational - HS Diploma
Doctorate
Some High School Coursework
Vocational
Professional
Some College Coursework Completed
Certification
Associate Degree
Master's Degree
Unspecified
High School or equivalent
Bachelor's Degree
nan
required_education
0
1000
2000
3000
4000
5000
6000
7000
8000
Count

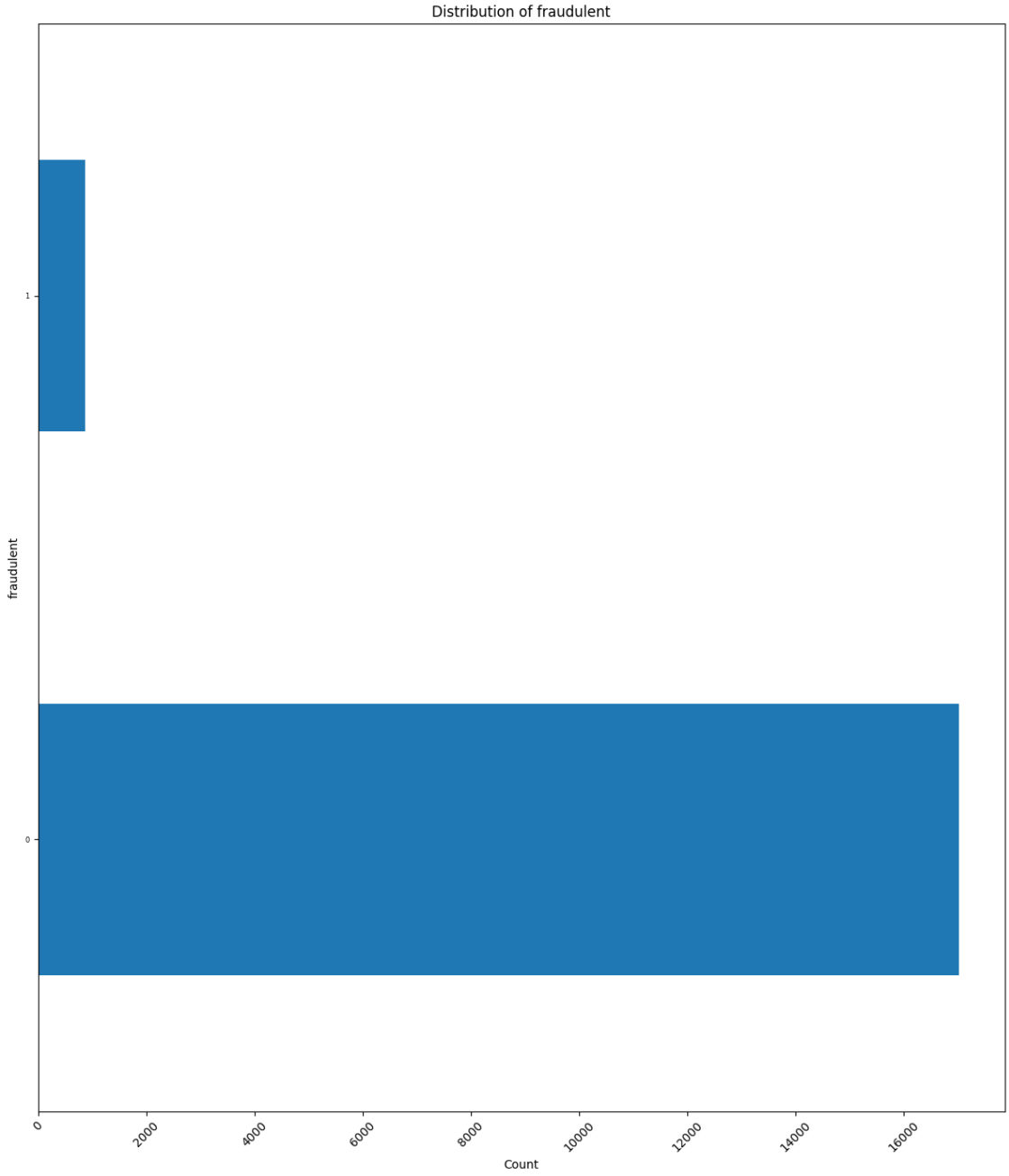

Distribution of industry
industry
Count
Ranching
Alternative Dispute Resolution
Shipbuilding
Museums and Institutions
Sporting Goods
Wine and Spirits
Nanotechnology
Military
Textiles
Libraries
Plastics
Performing Arts
Mining & Metals
Public Policy
Maritime
Furniture
Wireless
Investment Banking
Fishery
Commercial Real Estate
Luxury Goods & Jewelry
Philanthropy
Package/Freight Delivery
Import and Export
Capital Markets
Animation
Packaging and Containers
Religious Institutions
Investment Management
Primary/Secondary Education
Motion Pictures and Film
Public Safety
Industrial Automation
Photography
Executive Office
Veterinary
International Trade and Development
Renewables & Environment
Civil Engineering
Individual & Family Services
Defense & Space
Program Development
Music
Architecture & Planning
Translation and Localization
Law Enforcement
Semiconductors
Government Relations
Higher Education
Machinery
Wholesale
Professional Training & Coaching
Fund-Raising
Business Supplies and Equipment
Law Practice
Medical Devices
Writing and Editing
Outsourcing/Offshoring
Chemicals
Government Administration
Mental Health Care
Sports
Farming
Aviation & Aerospace
Information Services
Research
Venture Capital & Private Equity
Security and Investigations
Printing
Graphic Design
Utilities
Computer Hardware
Mechanical or Industrial Engineering
Biotechnology
Publishing
Pharmaceuticals
Gambling & Casinos
Computer Networking
Food Production
Media Production
Computer & Network Security
Environmental Services
Events Services
Broadcast Media
Warehousing
Restaurants
Transportation/Trucking/Railroad
Market Research
Civic & Social Organization
Public Relations and Communications
Medical Practice
Consumer Electronics
Airlines/Aviation
Consumer Goods
Cosmetics
Food & Beverages
Electrical/Electronic Manufacturing
Entertainment
Nonprofit Organization Management
Leisure, Travel & Tourism
Building Materials
Banking
Computer Games
Hospitality
Facilities Services
Apparel & Fashion
Legal Services
Online Media
Human Resources
Logistics and Supply Chain
Automotive
Insurance
Staffing and Recruiting
Health, Wellness and Fitness
Design
Management Consulting
E-Learning
Construction
Accounting
Real Estate
Retail
Oil & Energy
Telecommunications
Consumer Services
Hospital & Health Care
Financial Services
Education Management
Marketing and Advertising
Internet
Computer Software
Information Technology and Services
nan
0
1000
2000
3000
4000
5000


Distribution of function
Science
Purchasing
Distribution
Financial Analyst
Supply Chain
Training
Strategy/Planning
Legal
Research
General Business
Manufacturing
Public Relations
Data Analyst
Business Analyst
Advertising
Quality Assurance
Product Management
Production
Art/Creative
Writing/Editing
Consulting
Finance
Project Management
Human Resources
Accounting/Auditing
Business Development
Management
Education
Other
Health Care Provider
Design
Administrative
Marketing
Customer Service
Engineering
Sales
Information Technology
nan
function
0
1000
2000
3000
4000
5000
6000
Count


Distribution of telecommuting
telecommuting
1
0
0
2000
4000
6000
8000
10000
12000
14000
16000
Count


Distribution of has_company_logo
has_company_logo
0
1
0
2000
4000
6000
8000
10000
12000
14000
Count


Distribution of has_questions
has_questions
1
0
0
2000
4000
6000
8000
Count



1.2.11 Results:

After visualizing our data, we understand that our dataset has class imbalance because we have many more real jobs than fraudulent jobs. We will have to keep this in mind as we build our data science model.

Target Variable: fraudulent

Imbalanced classes:

0 (Real jobs) dominate the dataset.

1 (Fake jobs) are a minority.

We'll need to address this imbalance during modeling (e.g., SMOTE, stratified sampling).

1.2.12 Purpose:

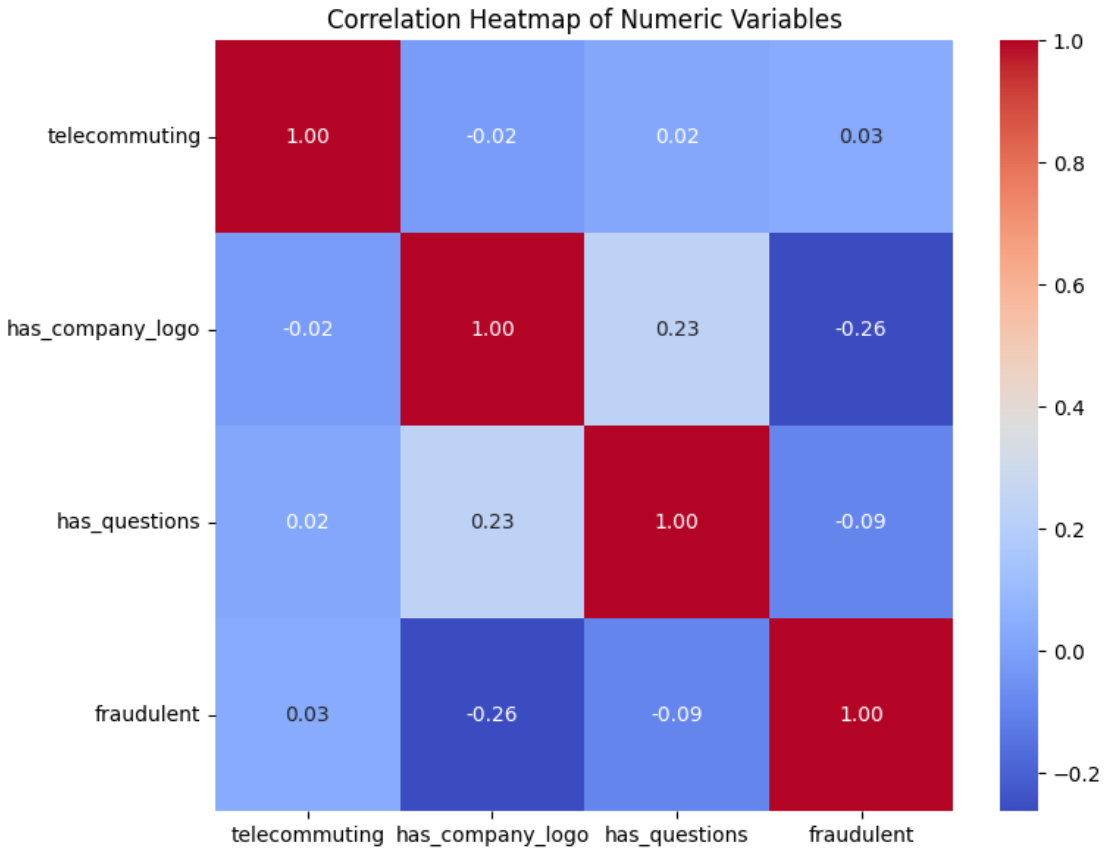
This function is designed to visualize the correlation relationships between selected numerical columns in the dataset. It computes the pairwise correlation matrix and plots it as a heatmap, where the strength and direction of the relationships are represented by color intensity. This helps in identifying strong positive or negative correlations between variables, which can inform feature selection, detect multicollinearity, or guide further analysis.

Correlation is a measure that shows the relationship between two variables. It ranges from -1 to +1:

Positive (+1) –> Strong positive correlation - as one goes up, the other also goes up.

Negative (-1) –> Strong negative correlation - as one goes up, the other goes down.

None (0) –> No correlation - they don't move together in any predictable way.



1.2.13 Results:

1. Field: has_company_logo
    - Correlation with fraudulent: negative
    - Meaning: If a job has a company logo (has_company_logo = 1), it's less likely to be fake
    - So: Fake jobs often don't have logos → makes sense.
2. Field: has_questions
    - Correlation with fraudulent: positive
    - Meaning: If has_questions = 1 (i.e., the job does have screening questions), it's more likely to be fake
    - But this seems counterintuitive? You'd think real jobs use questions more?
3. Explanation?

The opposite is more likely true in most real-world cases:

Real jobs do include screening questions. Fake jobs tend to skip them, because they want to look easy/appealing.

So, if has_questions has a positive correlation with fraudulent, it might be due to:

`3.1 - Noise or imbalance in the data.`

`3.2 - Possible feature encoding error.`

`3.3 - Or the correlation is very weak and not meaningful in practice.`

Conclusion:

A positive correlation with fraudulent means that as the other variable increases (e.g., has_questions = 1), the likelihood of being fake (fraudulent = 1) also increases.

But correlation only shows association, not causation.

1.2.14 Step 2: Structuring

Structuring data transforms features to uniform formats, units, and scales.

1.2.15 Purpose:

Format text fields in a uniform format

1.2.16 Results:

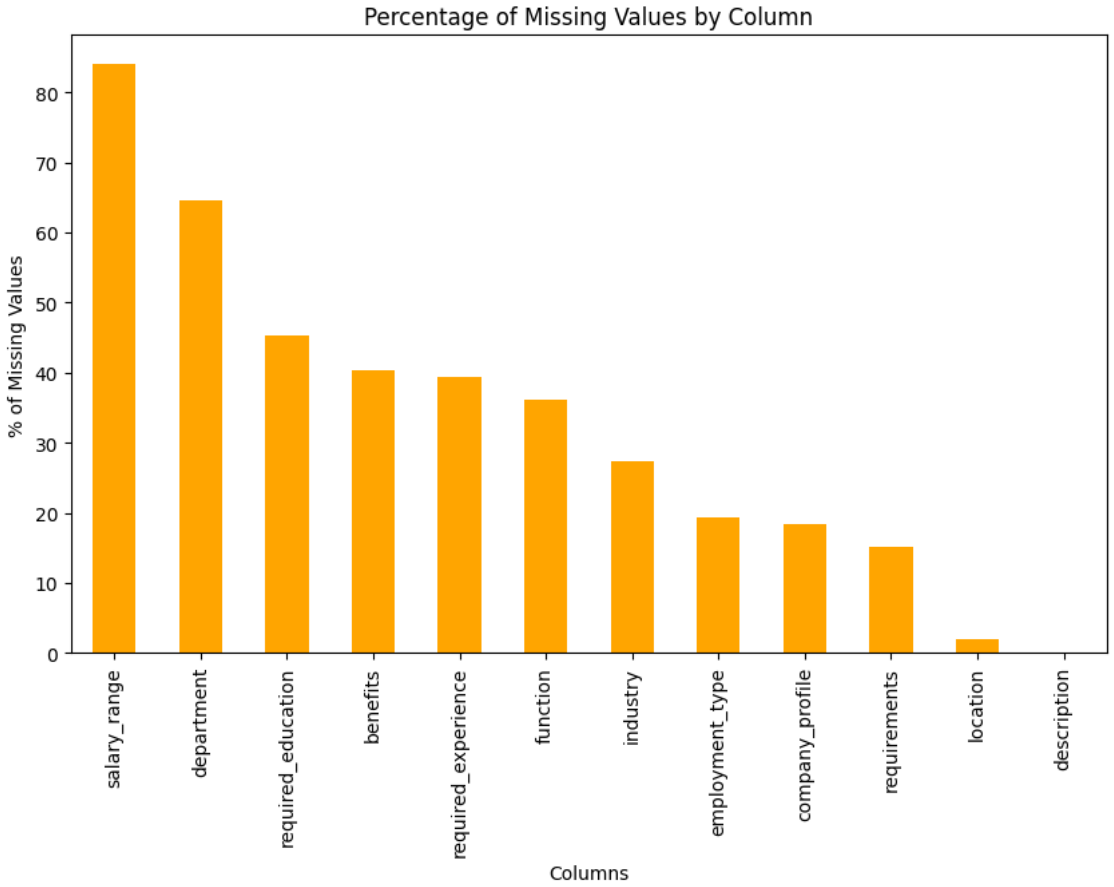
Now the text fields we care about for our model are formatted consistently. Due to the nature of our dataset we did not have to do too much structuring to it.

1.2.17 Step 3: Cleaning

Cleaning data removes or replaces missing and outlier data.

1.2.18 Purpose:

We saw earlier in a table that we have certain coumns with missing data ordered by percentage, now we either want to drop/filter those columns out of our dataset also because they have poor real-world usability for our Data Science Model.



1.2.19 Results:

We've visualized some of the fields we intend to filter in the coming steps.

1.2.20 Purpose:

We want to filter the dataframe to only have our feature and target columns since some of the other fields have too mnay nulls or are not usable for our goals and have poor real-world application.

Features:

- Text: title, description, requirements, company_profile

- Structured: employment_type (Categorical), telecommuting (Boolean)

Target: - fraudulent

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17880 entries, 0 to 17879
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   job_id           17880 non-null  int64
 1   title            17880 non-null  object
 2   description      17879 non-null  object
 3   requirements     15184 non-null  object
 4   company_profile  14572 non-null  object
 5   employment_type  14409 non-null  object
 6   telecommuting    17880 non-null  int64
 7   fraudulent       17880 non-null  int64
dtypes: int64(3), object(5)
memory usage: 1.1+ MB
```

1.2.21 Results:

Now we have only our feature and target columns along with the "job_id" field to easily identify a given row.

1.2.22 Purpose:

This function is responsible for cleaning the dataset by addressing missing or incomplete data in key text and categorical fields. Specifically, it performs two main operations:

1. Row Removal: It drops rows where all of the major text fields (title, description, requirements, company_profile) are missing, as such records provide no useful information.
2. Value Imputation: It fills missing or blank values in the employment_type column with the placeholder value 'Unknown', ensuring consistent treatment of missing employment data.

This cleaning step improves data quality by removing irrelevant records and standardizing missing values, making the dataset more reliable for analysis and modeling.

1.2.23 Results:

We now have our cleaned dataframe and can continue with the next steps.

1.2.24 Step 4: Enriching

Enriching data derives new features from existing features and appends new data from external sources.

test

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17880 entries, 0 to 17879
Data columns (total 15 columns):
```

```
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   title                          17880 non-null  object
 1   company_profile                14572 non-null  object
 2   description                    17879 non-null  object
 3   requirements                   15184 non-null  object
 4   telecommuting                  17880 non-null  int64
 5   fraudulent                     17880 non-null  int64
 6   employment_type_Contract       17880 non-null  float64
 7   employment_type_Full-time      17880 non-null  float64
 8   employment_type_Other          17880 non-null  float64
 9   employment_type_Part-time      17880 non-null  float64
 10  employment_type_Temporary      17880 non-null  float64
 11  employment_type_Unknown        17880 non-null  float64
 12  description_wordcount          17880 non-null  int64
 13  description_readability        17880 non-null  float64
 14  description_complexity_score   17880 non-null  float64
dtypes: float64(8), int64(3), object(4)
memory usage: 2.0+ MB
```

1.2.25 Step 5: Validating

Validating data verifies that the dataset is internally consistent and accurate.

1.2.26 Step 6: Publishing

Publishing data makes the dataset available to other data scientists by storing data in a database, uploading data to the cloud, or distributing data files.

1.3 Data Exploration

1.3.1 Step 1: Understand the data

Find the size of the dataset (number of rows and columns), identify and categorize the features (categorical, numerical).

Looking at the shape of the dataset to find how many records and columns we have.

```
Shape of dataset: (17880, 23)
```

From looking at the shape of the dataset, we have 17,880 rows and 18 columns.

We can identify and categorize the features using the info method of the dataset.

```
Data types and non-null counts:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17880 entries, 0 to 17879
Data columns (total 23 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   job_id                      17880 non-null  int64
```

```
 1   title                       17880 non-null  object
 2   location                    17534 non-null  object
 3   department                  6333 non-null   object
 4   salary_range                2868 non-null   object
 5   company_profile             14572 non-null  object
 6   description                 17879 non-null  object
 7   requirements                15184 non-null  object
 8   benefits                    10668 non-null  object
 9   telecommuting               17880 non-null  int64
 10  has_company_logo            17880 non-null  int64
 11  has_questions               17880 non-null  int64
 12  required_experience         10830 non-null  object
 13  required_education          9775 non-null   object
 14  industry                    12977 non-null  object
 15  function                    11425 non-null  object
 16  fraudulent                  17880 non-null  int64
 17  employment_type_Contract    17880 non-null  float64
 18  employment_type_Full-time   17880 non-null  float64
 19  employment_type_Other       17880 non-null  float64
 20  employment_type_Part-time   17880 non-null  float64
 21  employment_type_Temporary   17880 non-null  float64
 22  employment_type_Unknown     17880 non-null  float64
dtypes: float64(6), int64(5), object(12)
memory usage: 3.1+ MB
None
```

Looking at the dataset using the info method, we can see that we have 5 numerical and 13 catergorial features listed below.

Catergorical: * title * location * department * salary_range * company_profile * description * requirements * benefits * employment_type * required_experience * required_education * industry * function

Numerical: * job_id * telecommuting * has_company_logo * has_questions * fradulent

1.3.2 Step 2: Identify relationships between features

Find the direction (positive, negative) and strength (strong, moderate, weak) of correlation between the features.

1.3.3 Step 3: Describe the shape of data

Determine the shape of the distribution (symmetric, skewed).

1.3.4 Step 4: Detect outliers and missing data

Find values that are much higher or lower than the rest of the data or values that strongly affect the shape of the data.

2 ========== END OF TEMPLATE ==========

One Hot Encode Employment type. Other object types are true/false or will be handled in Text processing. TODO: Should we just dropna instead of fillna?

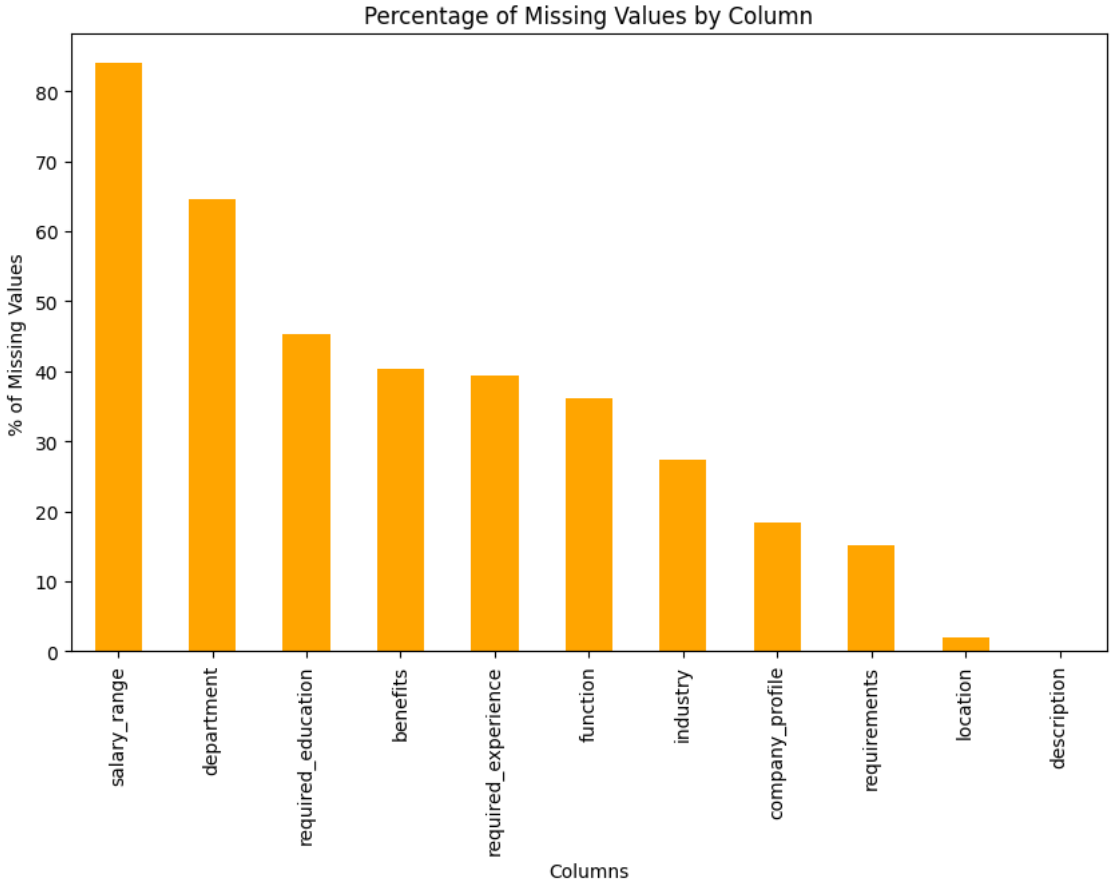
Darshan: I think this should be done later on, because we are only keeping some columns (Feature and Target), and only of those columns we need to fill Nan, Blank, Empty values with 'Unknown', I think it's better to do that.

```
employment na count: 3471
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17880 entries, 0 to 17879
Data columns (total 23 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   job_id                      17880 non-null  int64
 1   title                       17880 non-null  object
 2   location                    17534 non-null  object
 3   department                  6333 non-null   object
 4   salary_range                2868 non-null   object
 5   company_profile             14572 non-null  object
 6   description                 17879 non-null  object
 7   requirements                15184 non-null  object
 8   benefits                    10668 non-null  object
 9   telecommuting               17880 non-null  int64
 10  has_company_logo            17880 non-null  int64
 11  has_questions               17880 non-null  int64
 12  required_experience         10830 non-null  object
 13  required_education          9775 non-null   object
 14  industry                    12977 non-null  object
 15  function                    11425 non-null  object
 16  fraudulent                  17880 non-null  int64
 17  employment_type_Contract    17880 non-null  float64
 18  employment_type_Full-time   17880 non-null  float64
 19  employment_type_Other       17880 non-null  float64
 20  employment_type_Part-time   17880 non-null  float64
 21  employment_type_Temporary   17880 non-null  float64
 22  employment_type_Unknown     17880 non-null  float64
dtypes: float64(6), int64(5), object(12)
memory usage: 3.1+ MB
```

Discovery, also called data exploration, is the process of becoming familiar with the data source.

First, we will import in commonly used libraries in data wrangling, as well as update some default pandas options to get a more complete visualization of the data.

With the libraries imported and Google Drive mounted, we'll now load the data and look at the features and shape of the dataset.


Percentage of Missing Values by Column
% of Missing Values
0
10
20
30
40
50
60
70
80
salary_range
department
required_education
benefits
required_experience
function
industry
company_profile
requirements
location
description
Columns