Craft a Story with a Dataset
By Sarah Carrier

Job Ads - Are they real or fake?

The Jupyter notebook for EDA on this dataset is here:https://github.com/atomlattice/Springboard_practice/blob/bb18a8f351e2e49d8a87e6fe6eb9bc8b77654f79/Craft_A_Story/Craft%20a%20Story%20from%20a%20Dataset%20-%20Real%20or%20Fake%20Job%20Posting%20Prediction.ipynb

Problem Statement:
What's the harm in applying to a fake job ad? Perhaps it's just a waste of time. But the perpetrators may be able to steal your identity or your money. This is becoming more of a concern than ever as fake job postings proliferate across multiple job seeking and career development platforms online. What can we learn from a dataset where fraudulent job ads have been identified - what patterns and signs can be identified to warn job seekers against applying to fake jobs?

Dataset Description:
The dataset used for this project is the Fake/Real Job Posting dataset and this is available via Kaggle. It originated from The University of the Aegean | Laboratory of Information & Communication Systems Security http://emscad.samos.aegean.gr/

The dataset, named The Employment Scam Aegean Dataset (EMSCAD) by researchers of University of the Aegean, consists of 17,800 job ads posted between 2012 to 2014 through Workable, a recruiting software, whose 866 fraudulent job ads were manually annotated by employees of Workable. The criteria of inclusion is said to be including "client's suspicious activity on the system, false contact or company information, candidate complaints and periodic meticulous analysis of the clientele". So on one hand there may be a small number of mislabeled job ads, on the other hand the annotation may include factors not contained in the dataset.

And the dataset includes structured and unstructured data, open text fields include "title", "company profile", "description", "requirements", "benefits", and to some extent 'location", "department" and "salary range". Structured fields include "employment type", "required experience", "required education", "industry" and "function", and there are binary fields indicating whether the ad has "company logo" and screening "questions", and whether the job involves "telecommuting". The following table lists the details of the feature fields of the dataset:

String Data

1. Title: The job advertisement header
2. Location: The location of the job adviser
3. Department: Job relevant department like sales

4. Salary range: Suggested Salary Range such as $50,000 - 60,000

HTML Fragment

1. Company Profile: A brief description of the company
2. Description: Advertised Job details
3. Requirement: Required list for job
4. Benefits: Benefits list offered by employer

Binary

1. Telecommuting: True for Telecommuting positions
2. Company Logo: True if company logo exists
3. Questions: True if screening question exists
4. Fraudulent: Classification attribute

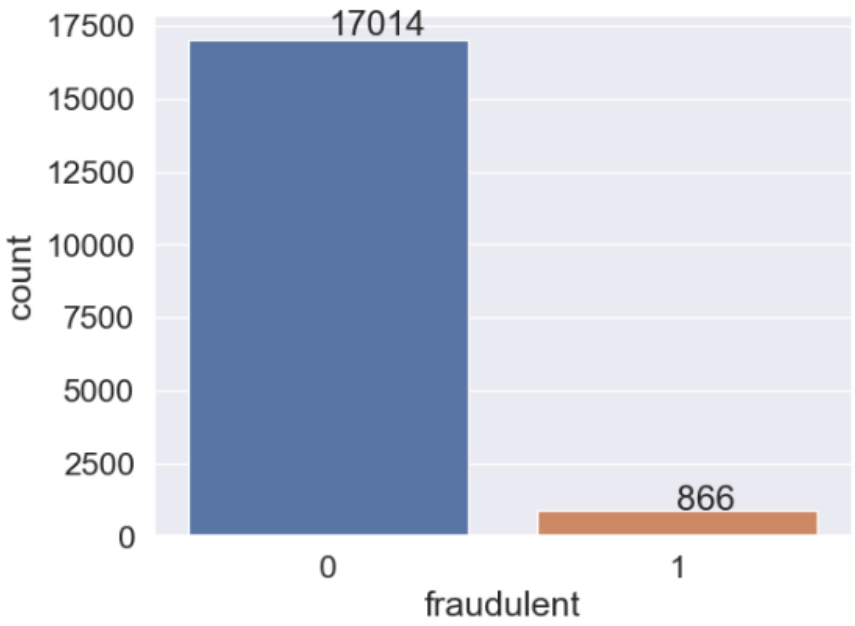
Nominal

1. Employment Type: Full-type, Part-time, Contract, etc.
2. Required Experience: Executive, Entry level, Intern, etc.
3. Required Education: Doctorate, Master's Degree, Bachelor's, etc.
4. Industry: Automotive, IT, Health care, Real estate, etc.
5. Function: Consulting, Engineering, Research, Sales etc.

Data Wrangling and Analysis:

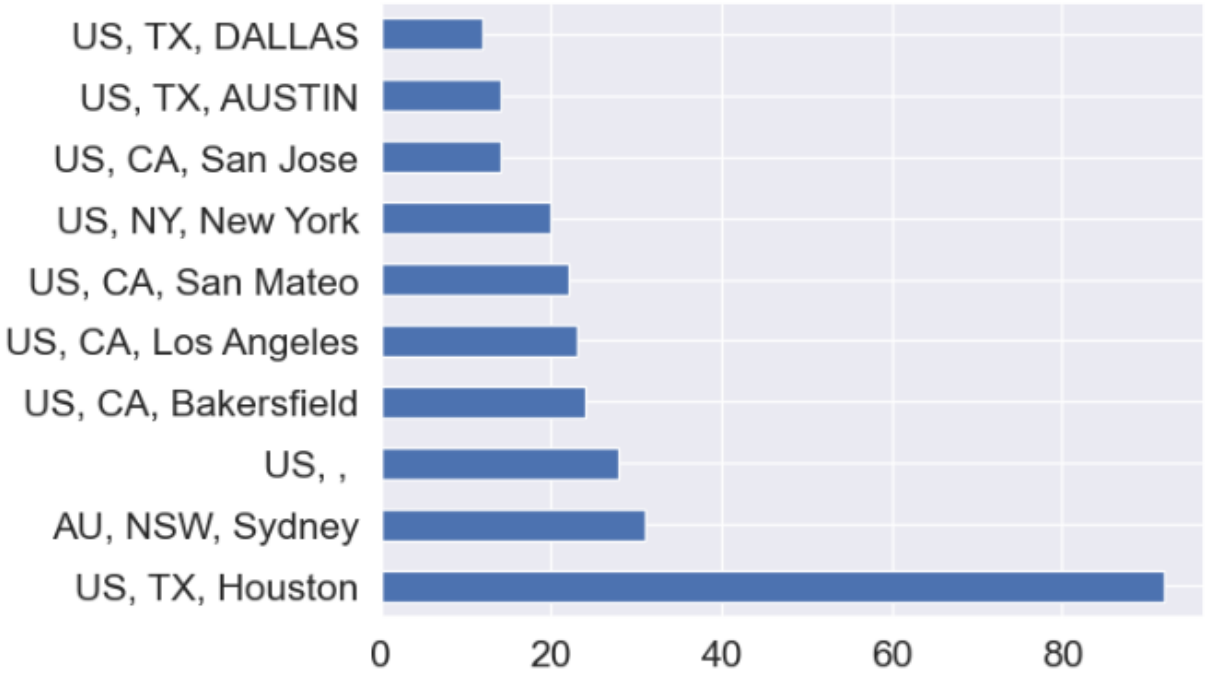
There are a lot of NaNs in this dataset HOWEVER they are in telling places, so it is actually necessary to leave them. The variable in question FRAUDULENT (0 or 1) has no missing values. Nor are there missing values in other designations by the creators of the dataset such as the presence or absence of a logo, whether telecommuting is available, etc. If a salary isn't listed in the ad, or benefits, or etc., that is actually to be expected, and its absence is a red flag that the ad may be fraudulent.

Most of the job ads in this dataset ARE NOT fraudulent:



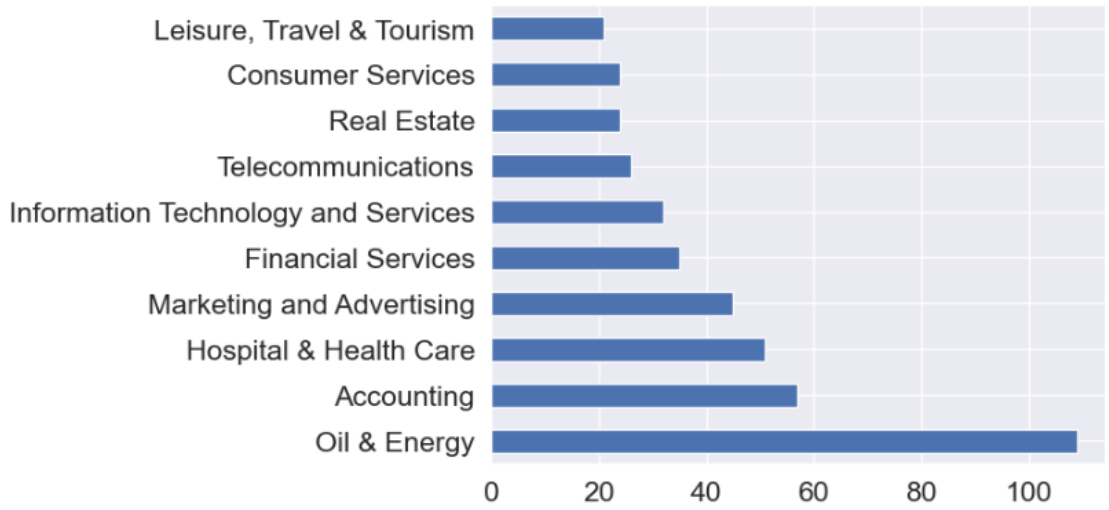
I decided to just pull out the ads that are fraudulent in order to get an idea of what patterns are going on in order to build a list of warning signs for job seekers. 866 rows are still more than enough to pull out some patterns.

1. What are the most common locations for these fake job ads?



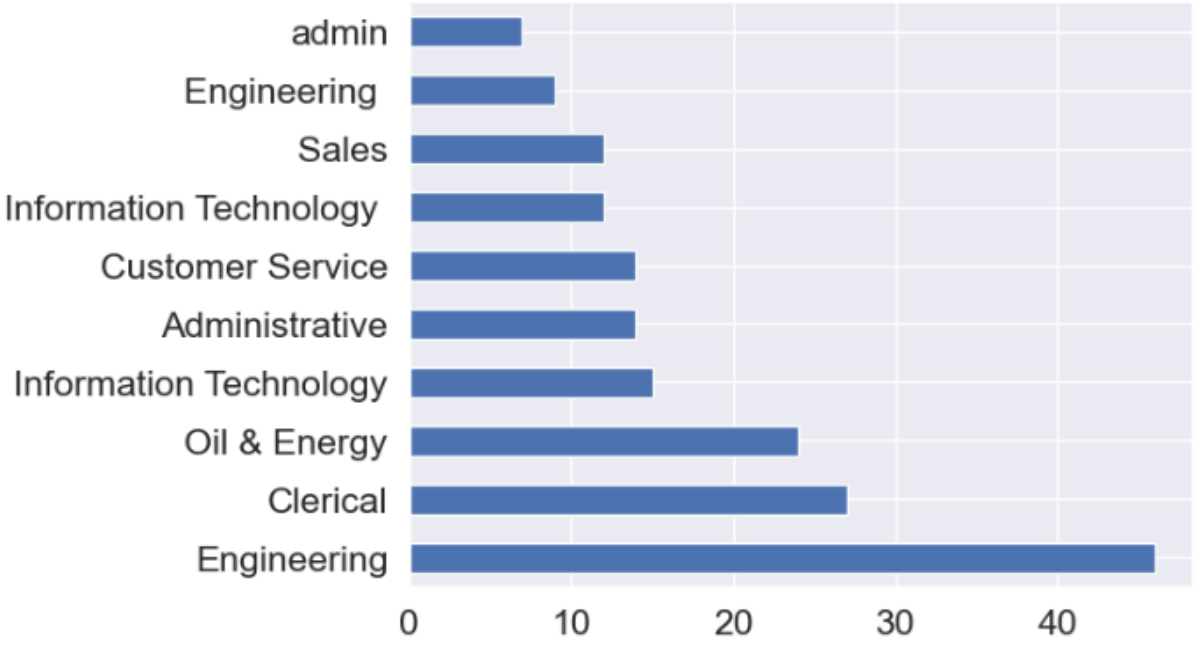
Of the fraudulent ads, by far the most common location for a fake job ad is HOUSTON TEXAS. Taking the next feature, industry, into account tells a bit about why:

2. What are the most common industries in fake job ads?



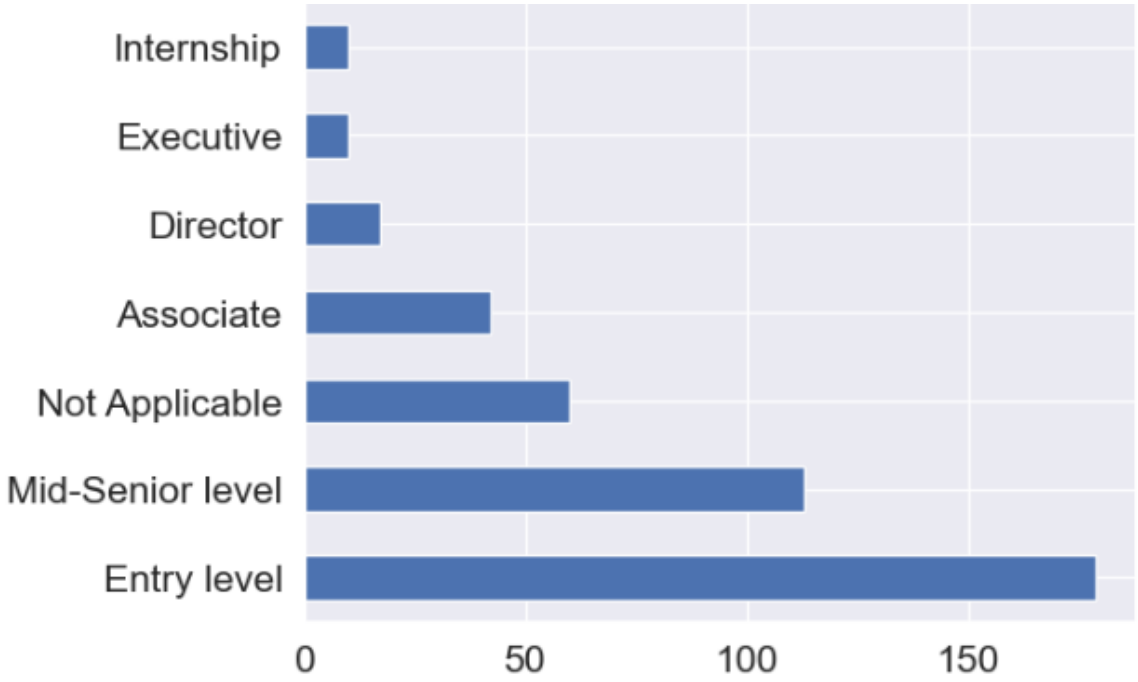
The most common INDUSTRY in fake jobs ads is OIL/ENERGY. Now, this makes sense because of the prevalence of the oil industry in Houston and Texas as a whole.

3. What is the most common department?



The most common DEPARTMENT listed in jobs ads is ENGINEERING.

4. What is the most common experience level required in fake job ads?



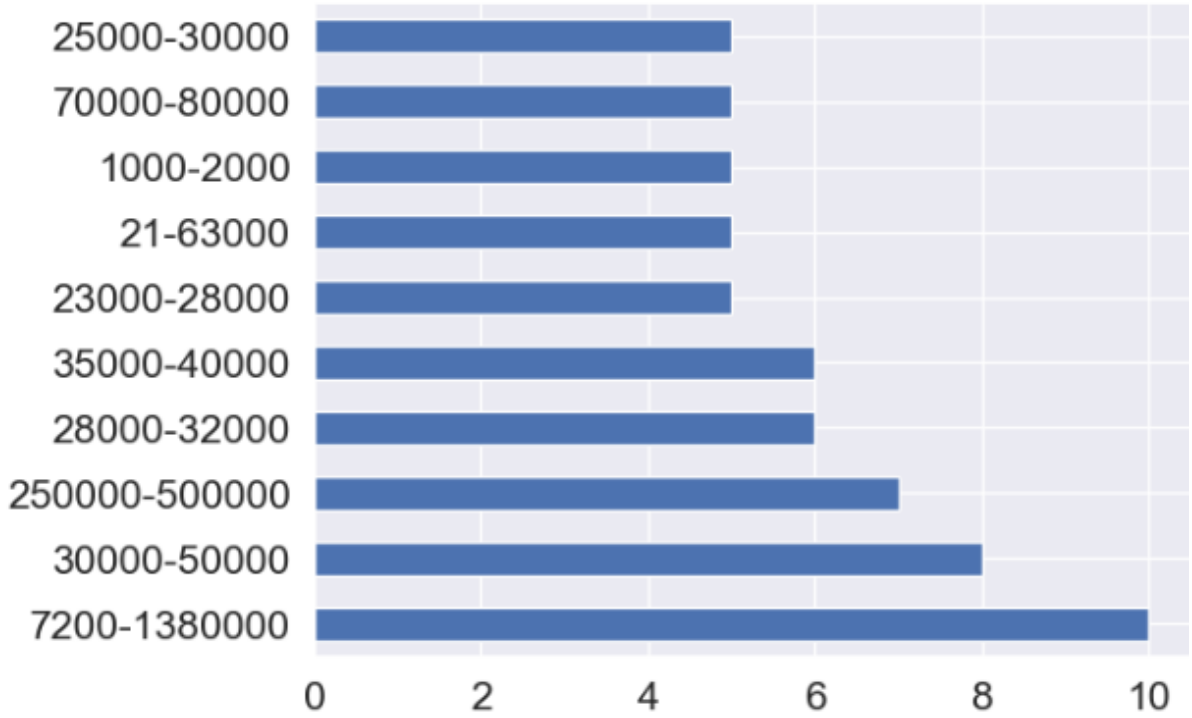
Entry Level is the most common experience level in fake jobs ads.

5. What about job titles?

The most common fake job titles are:

- Data Entry Admin/Clerical Positions - Work From Home
- Home Based Payroll Typist/Data Entry Clerks Positions Available
- Cruise Staff Wanted *URGENT*

6. What about salaries?



The most common salary range is, implausibly, 7200-1380000.

I looked at intersections of different features.

1. For each job with a common title, what is the experience level required? We know that ENTRY LEVEL is the most common.

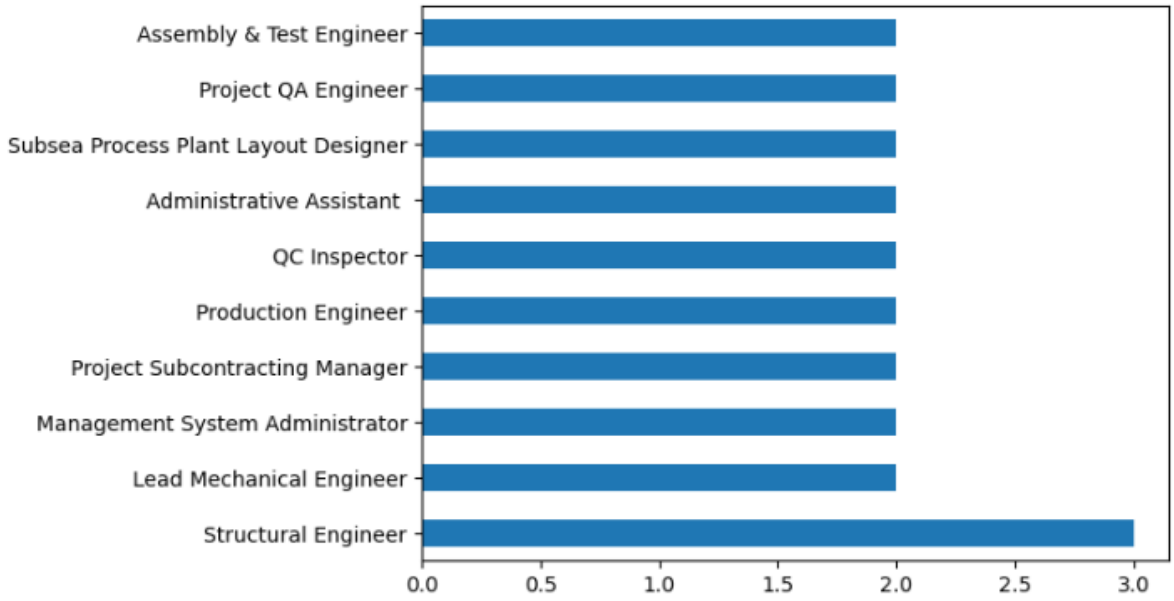
This revealed that there is NO experience level listed for any of the ads with job title “Data Entry Admin/Clerical Positions - Work From Home,” which is one of the 3 most common fake job titles. This is why I didn’t get rid of the NaNs - the absence of this information is actually most telling.

2. Another interesting result occurs when looking at the salary ranges listed for the most common department in fake ads, Engineering.

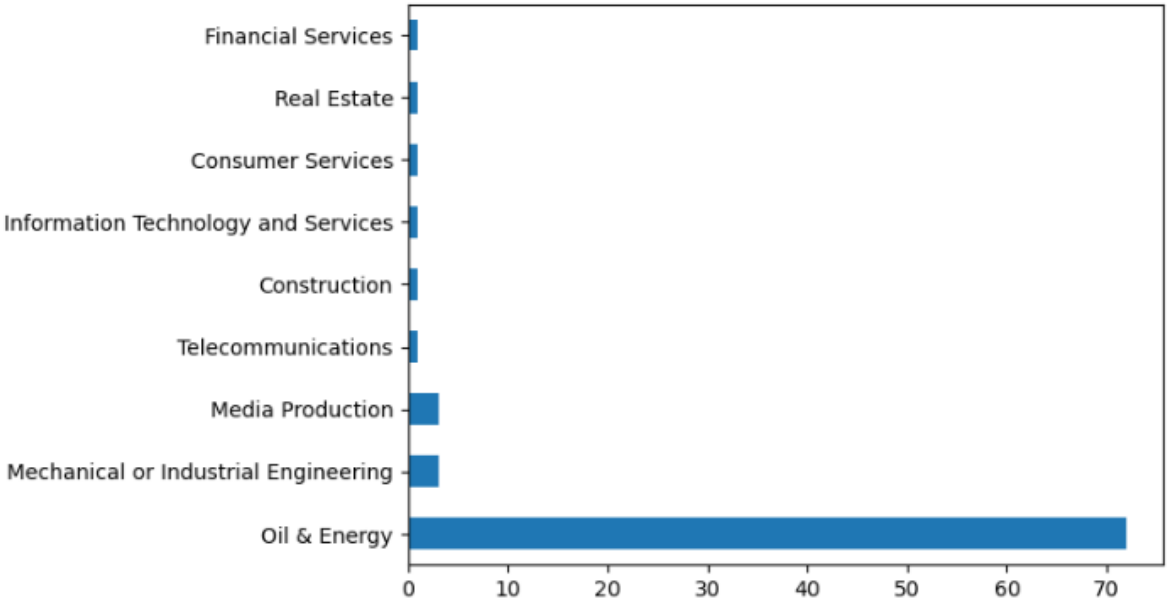
This reveals that MOST of the Engineering fake jobs had NO salary even listed. And the two the did were odd: 120000-180000 and the other 60000-100000.

I also went back to fake jobs located in **Houston** so I could confirm the most common industries, titles, and other features.

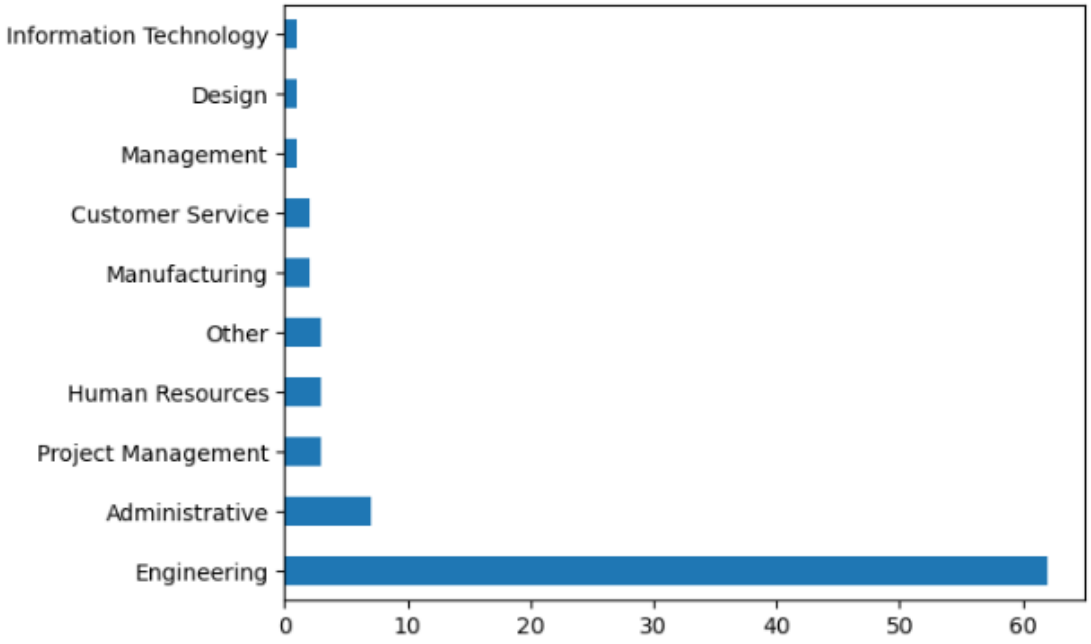
Most common fake job title in Houston:



Most common fake job industry in Houston:



Most common FUNCTION listed in fake Houston ads:



SUMMARY AND CONCLUSION:

Things to watch out for in jobs ads, just based on features listed here, not even yet doing text analysis on the descriptions of each job:

1. Most fake jobs ads list a US location, with Houston, TX being the most common.
2. For all fake jobs, and for all fake jobs in Houston, there are some commonalities:
    a. The most common industry all around plus in Houston is OIL AND ENERGY.
    b. Function and Department are very closely defined in this dataset and overall, plus overall in Houston, fake job ads are most likely to be Engineering (function and department).
3. Of all fake jobs, within OIL AND ENERGY industry, most of the jobs are in Engineering.
4. The most common fake job title in Houston is STRUCTURAL ENGINEER, followed closely by LEAD MECHANICAL ENGINEER.

Some summaries - things for job seekers to note:

1. Most fake ads in this dataset are for *Engineering* jobs in the *Oil and Energy* industry and located in *Houston, TX*.
    a. It does make sense since there are lots of oil companies in Houston and in Texas.
2. Most of the fake jobs are entry level.
3. Fake job ads have very odd and contradictory salary information listed, sometimes not making sense like “21-63000”
4. Most fake jobs ads are missing A LOT of information from the ads, oftentimes missing very important information like what city, and etc.

5. The most common job titles for all the fraudulent ads are used, word for word, over and over again. This should be suspicious to job seekers if the same job title is seen over and over again.