# Part-1

# Dataset: -



* *The source I used for the dataset is https://*[*www.kaggle.com/datasets/shivamb/real-or-fake-fake-*](http://www.kaggle.com/datasets/shivamb/real-or-fake-fake-) *jobposting-prediction”.*

# Dataset Description: -



* *The "Real or Fake: Fake Job Posting Prediction" dataset is a binary classification problem that aims to classify job postings into two categories: real or fake. The dataset was collected from various sources and was compiled by Shivam Bansal. It consists of 18,000 job postings with a total of 19 variables.*

# Dataset Dictionary: -

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Name** | **Definition of the variable** | **DataType** |
| 1. | **jobId :** | A unique identifier for each job posting. | numeric, range: 1-18,000 |
| 2. | **title :** | The title of the job posting. | string |
| 3. | **location :** | The location of the job. | String |
| 4. | **department :** | The department the job posting belongs to. | String |
| 5. | **salary\_range :** | The salary range for the job. | String |
| 6. | **company\_profile :** | A brief description of the company. | String |
| 7. | **description :** | A detailed description of the job responsibilities and  requirements. | String |
| 8. | **requirements :** | A list of requirements for the job. | String |
| 9. | **benefits :** | A list of benefits offered by the company. | String |
| 10. | **telecommuting :** | Indicates whether the job allows telecommuting. | Binary (0 = No,1 = Yes) |
| 11. | **has\_company\_logo :** | Indicates whether the job posting has a company logo. | Binary (0 = No,1 = Yes) |
| 12. | **has\_questions :** | Indicates whether the job posting includes interview  questions. | Binary (0 = No,1 = Yes) |
| 13. | **employment\_type :** | The type of employment (e.g. full-time, part-time, contract,  etc.) | String |
| 14. | **required\_experience :** | The required level of experience for the job. | String |
| 15. | **required\_education :** | The required level of education for the job. | String |
| 16. | **Industry :** | The industry the job belongs to. | String |
| 17. | **Function :** | The function of the job. | String |
| 18. | **fraudulent :** | Indicates whether the job posting is fraudulent or not. | Binary (0 = No,1 = Yes) |
| 19. | **Date :** | The date the job was posted on. | created (date, format:  YYYY-MM-DD) |
| 20. | **NA** | Not Applicable | - |

 **FINER Questions: -**

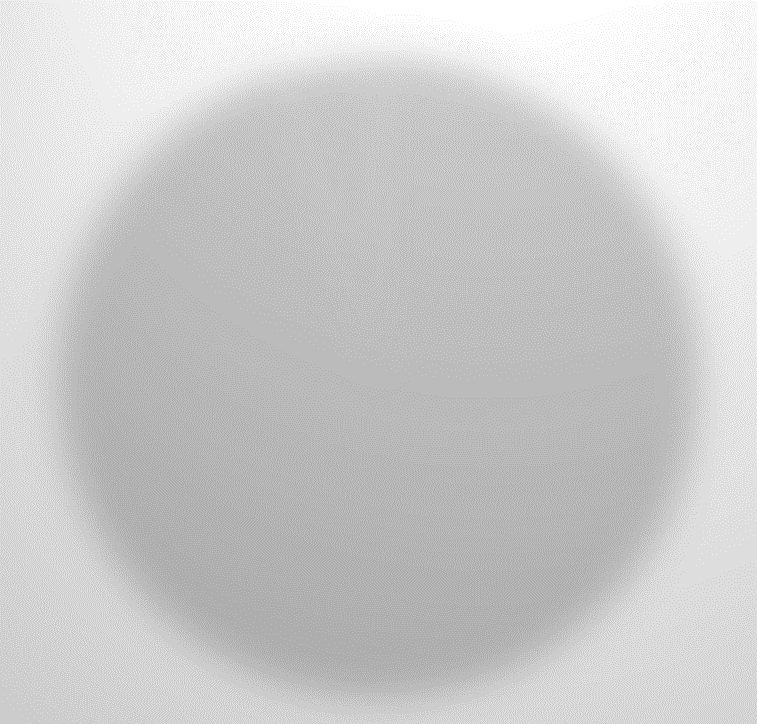
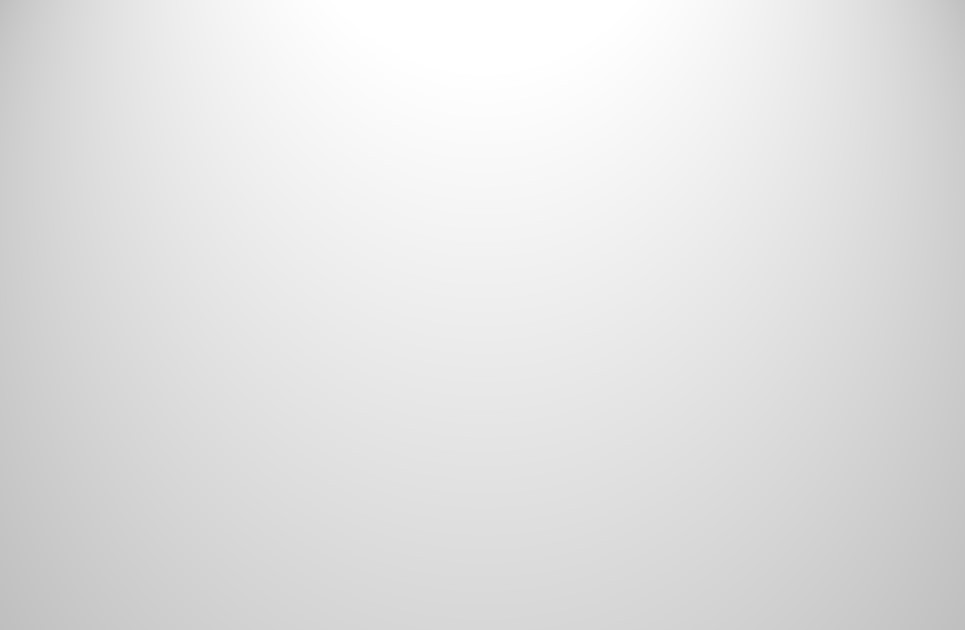
* 1. What is the relevance of the dataset and why to study the dataset?
  2. What is the main focus or feature of the dataset?
  3. What contributions do these feature give to the dataset?
  4. How do the numerical data pattern affect the conclusion of dataset?
  5. What is the relationship between different variables of the dataset and how do the affect a data observers understanding?
  6. What is the limitation of the following dataset?

# Tracking Analysis: -

* The aim to study this dataset was for me is to focus and address an important issue going on specially in Canada about the fraudulent job postings and scams doing the rounds of the internet, but I was able to find a dataset that focused on a global platform.
* The particular reason for selecting the following dataset was that it matched with the rubric criteria od data length and it helps us to look at the bigger picture more affectively.
* I have explored the data, customized the title row and froze the panes to make the dataset more readable and accessible.

# Data Visuals: -

## Pie Chart



**Chart Title**

**866, 5%**

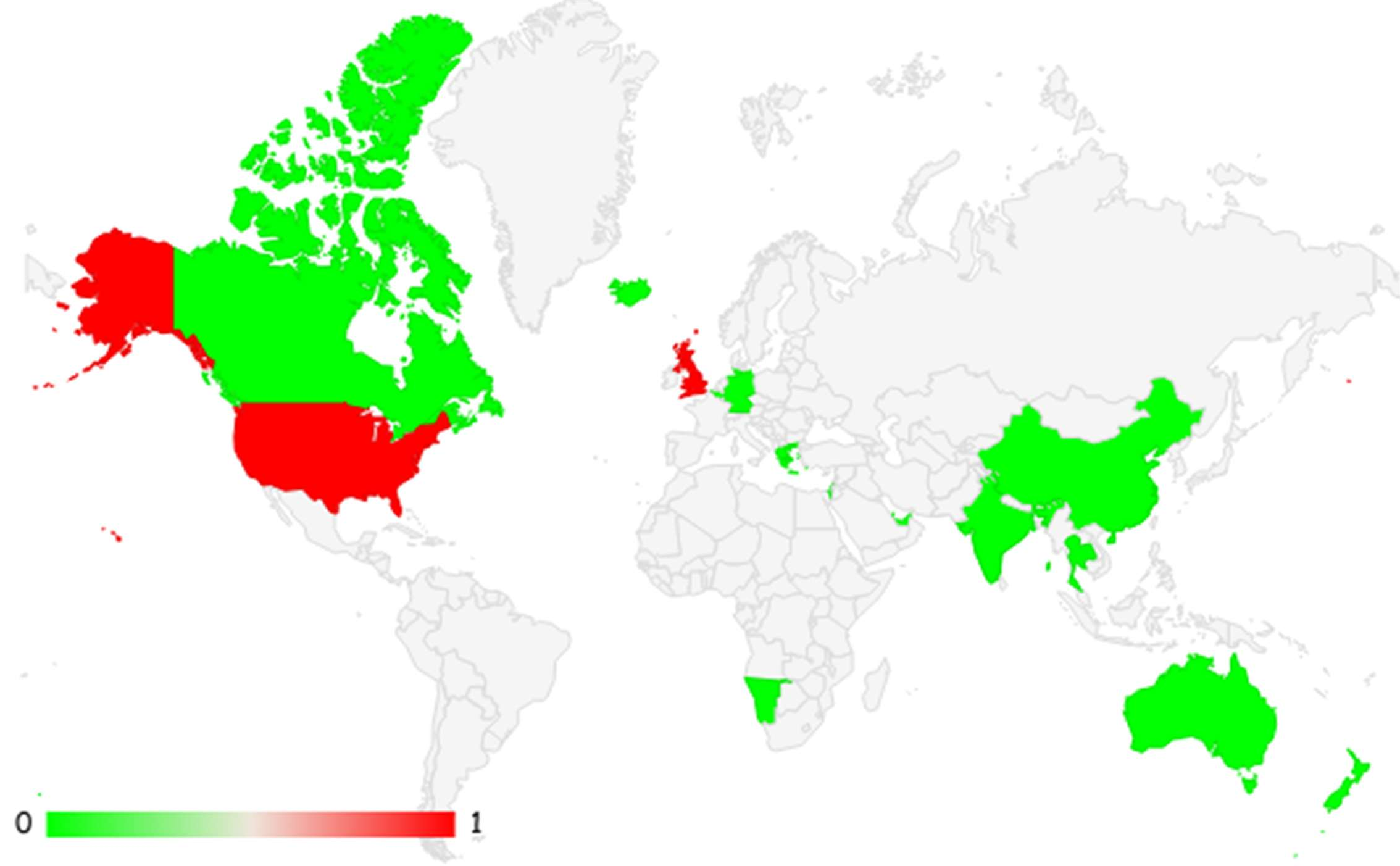
Fraud Job

Legitmate Jobs

**17014, 95%**

* The above visuals show the percentage of fraud job advertisement making the rounds in the market.

## Geo Graph



1. Not Fraudulent
2. Fraudulent

* The above chart describes what part of world has the most cases of fraudulent job. The red has the most cases while the green has lesser significantly.
* These entire graphical representation and data usage is based on the assumption of that the data is valid and 100 percent accurate.
* The 19 variables make it difficult that all the data is accumulated, the independent variables such as the country and the fraudulent status is used in making of this graphical representation.
* The data manipulation for the raw data included eliminating blank cells.
* I broke down the location column in to country, state and city.
* There is various city who are also state and many countries do not recognize the term of “state”, so there are many blank cells.
* There are long, ambiguous and useless texts which was wrapped and organized.

# Part -2

# Univariate Analysis: -

# employment\_type

# 

|  |  |
| --- | --- |
| **Row Labels** | **Count of employment\_type** |
| Contract | 1524 |
| Full-time | 11620 |
| NA | 3471 |
| Other | 227 |
| Part-time | 797 |
| Temporary | 241 |
| **Grand Total** | **17880** |

# industry

# Data before cleaning

# 

# Data after cleaning

# 

# Visualization

# 

# has\_company\_logo

# 0 signifies no logo while 1 signifies logo.

# 

# 

# required\_education

# before cleaning data

|  |  |
| --- | --- |
| **Row Labels** | **Count of required\_education** |
| Associate Degree | 274 |
| Bachelor's Degree | 5145 |
| Certification | 170 |
| Doctorate | 26 |
| High School or equivalent | 2080 |
| Master's Degree | 416 |
| NA | 8105 |
| Professional | 74 |
| Some College Coursework Completed | 102 |
| Some High School Coursework | 27 |
| Unspecified | 1397 |
| Vocational | 49 |
| Vocational - Degree | 6 |
| Vocational - HS Diploma | 9 |
| **Grand Total** | **17880** |
|  |  |

# 

# After cleaning data

|  |  |
| --- | --- |
| **Row Labels** | **Count of required\_education** |
| Associate Degree | 274 |
| Bachelor's Degree | 5145 |
| Certification | 170 |
| Doctorate | 26 |
| High School or equivalent | 2080 |
| Master's Degree | 416 |
| Professional | 74 |
| Some College Coursework Completed | 102 |
| Some High School Coursework | 27 |
| Vocational | 49 |
| Vocational - Degree | 6 |
| Vocational - HS Diploma | 9 |
| **Grand Total** | **8378** |

# Difference between data before and after cleaning visualization.

# functions

# salary\_range

|  |
| --- |
| **Max of salary\_range** |
| 45992 |

|  |
| --- |
| **Average of salary\_range** |
| 44763.03704 |

|  |
| --- |
| **Min of salary\_range** |
| 40000 |

# 

# 

# has-questions

|  |  |  |  |
| --- | --- | --- | --- |
| **Count of job\_id** | **Column Labels** |  |  |
| **Row Labels** | **0** | **1** | **Grand Total** |
| **Contract** | **523** | **1001** | **1524** |
| **Full-time** | **6175** | **5445** | **11620** |
| **Other** | **81** | **146** | **227** |
| **Part-time** | **402** | **395** | **797** |
| **Temporary** | **110** | **131** | **241** |
| **Grand Total** | **7291** | **7118** | **14409** |
|  |  |  |  |

# telecommuting

# Part -3

# Hypotheses: -

# The following are the five hypotheses that I could devise for the above data.

# Customer happiness is directly impacted by the caliber of customer service, according to the hypothesis.

# Explanation: Customers are more likely to be pleased with the firm overall if they feel they are receiving high-quality customer service. On the other hand, they could be less inclined to patronize the business again or refer others if they experience poor customer service.

# The frequency of social media posts and the degree of client involvement are believed to be positively correlated.

# Explanation: The visibility and reach of a company's content might grow by posting more frequently on social media sites. More customer interaction, including likes, shares, and comments, may result from this.

# Giving customers customized promos may boost their likelihood of making repeat purchases and loyalty.

# Explanation: Customers may feel more valued and are more inclined to make subsequent purchases when businesses give tailored promotions, such as discounts on goods they have already purchased or recommendations based on their purchase history. Customer loyalty and purchasing frequency may rise as a result.

# Hypothesis: Providing easy-to-use and informative product documentation can decrease the number of customer support inquiries.

# Explanation: If customers have access to clear and concise product documentation, they may be less likely to require assistance from customer support. This can save time and resources for the company while also improving the customer experience.

# Hypothesis: The implementation of a customer feedback system can lead to improvements in overall customer satisfaction.

# Explanation: By providing customers with a way to provide feedback on their experiences with the company, the company can gain insights into areas where they may be falling short and make improvements accordingly. This can lead to increased customer satisfaction and loyalty.

# Bi-variate/Multi-Variate Statistical Analysis: -