IE 5390 – Final Case – Fraud detection Case Study

Name: Pooja Arumugam

Final Project Report: Fake Job Postings Analysis

1. Introduction

The internet has revolutionized job searching, providing job seekers with thousands of opportunities at their fingertips. However, alongside genuine job postings, fraudulent job listings have emerged, exploiting unsuspecting candidates. Fake job postings can lead to identity theft, financial loss, and even personal safety risks.

This analysis aims to detect fraudulent job postings by identifying key differences between legitimate and fake job listings. Using data analytics, we can uncover hidden patterns that may indicate fraud and develop a methodology to flag suspicious postings.

A) Scenario Identification

- **Project Objective**

Fraudulent job postings are a growing concern as they can deceive job seekers, leading to identity theft, financial fraud, and wasted time. This project aims to analyze job posting data to detect patterns that distinguish fake job postings from legitimate ones.

- **Research Questions (Questions to answer):**

1. ***What factors differentiate fake job postings from real ones?***
2. ***Are there common attributes among fraudulent job listings?***
3. ***Can data analysis techniques effectively detect fake job postings?***

- **Significance of the Analysis (Why is it important)?**

This analysis is important because:

- It helps job seekers recognize fraudulent job postings and avoid scams.
- It provides insights for job boards and recruitment platforms to enhance their fraud detection systems.
- It explores how machine learning and statistical techniques can improve the detection of fraudulent job listings.

B) Dataset description:

Description of Variables or key elements

- The dataset consists of multiple columns such as job title, company, location, job description, required experience, employment type, fraud label (0 = Real, 1 = Fake), and other metadata related to job postings.
- This dataset contains 18K job descriptions out of which about 800 are fake. The data consists of both textual information and meta-information about the jobs. The dataset can be used to create classification models which can learn job descriptions which are fraudulent.

C) Dataset Source and Dataset File

- The dataset used for this analysis is fake_job_postings.csv.
- Source: https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction

D) Characteristics of the Data (As-Is Form)

- The dataset contains 17,880 job postings with 18 columns related to job details, company information, and fraud labels.
- The fraudulent column is a binary variable (0 = Legitimate, 1 = Fraudulent).
- Key job attributes include title, location, company profile, job description, requirements, benefits, employment type, industry, and function.

Data cleaning:

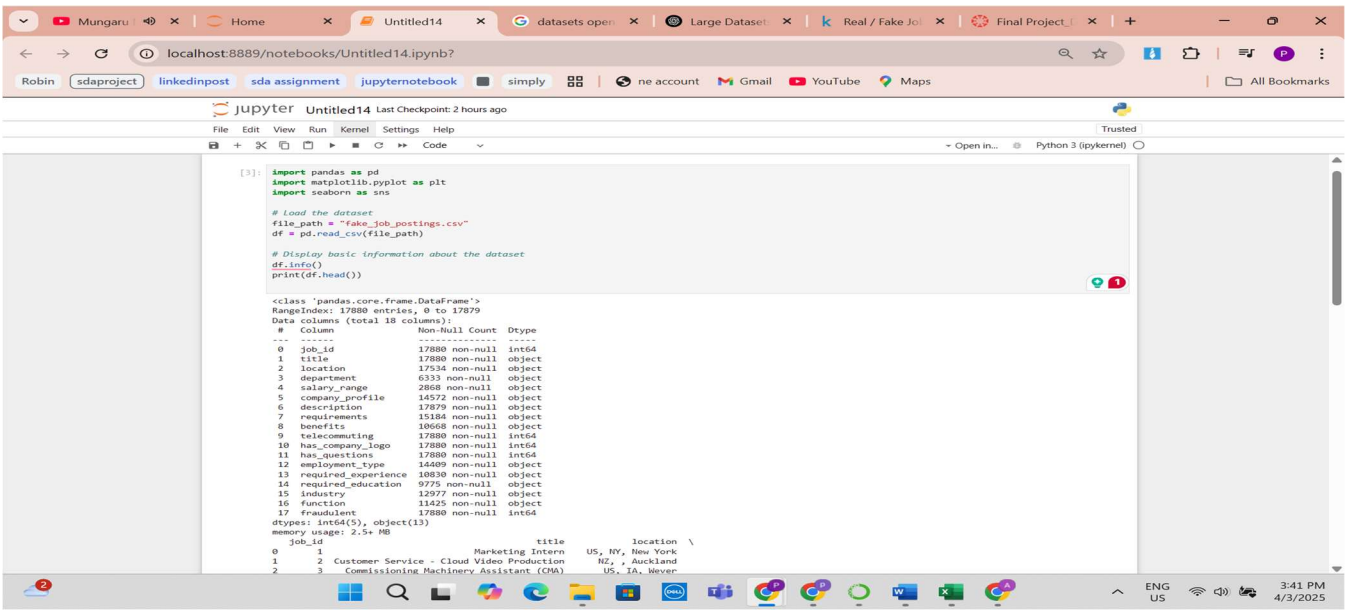


Figure 1: Data Cleaning

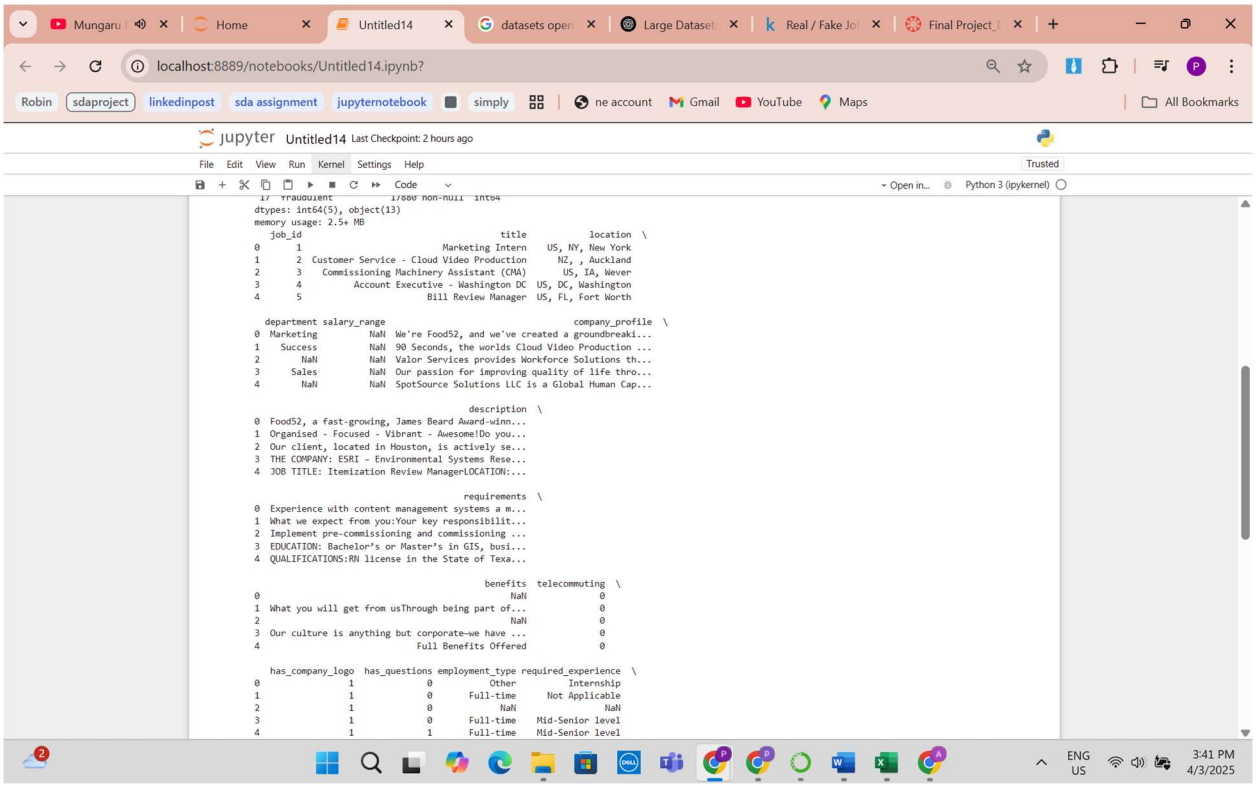


Figure 2

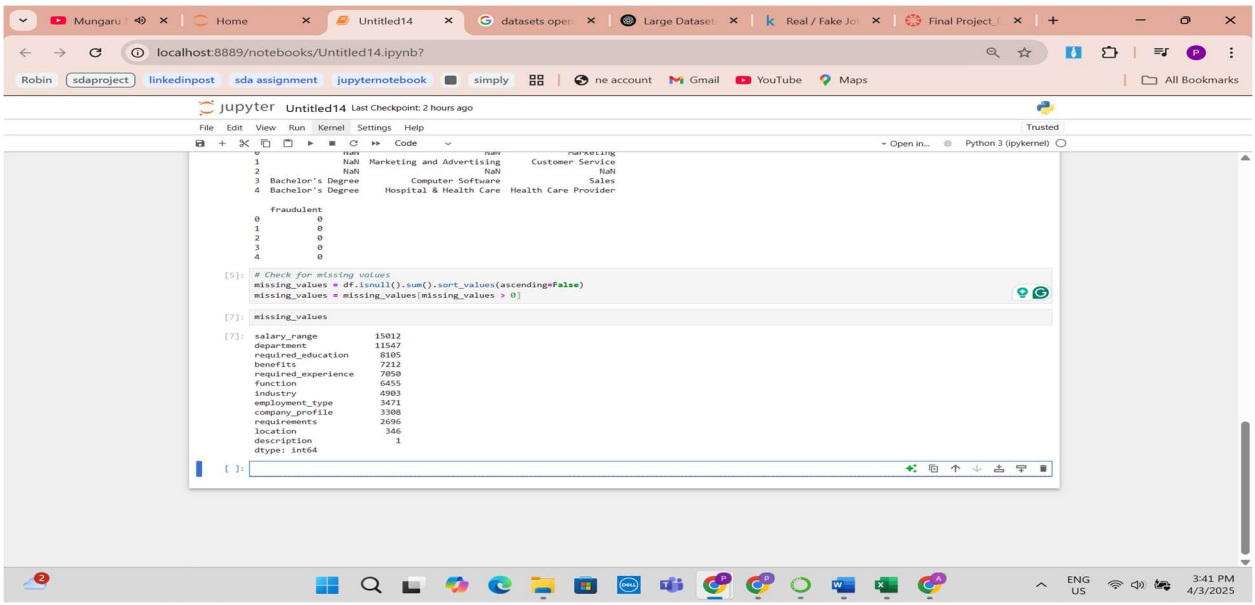


Figure 3: Missing values

Possible Anomalies:

- salary_range has very few entries (only 2,868 out of 17,880), which might indicate companies often don't disclose salary.
- Empty or vague descriptions could be a sign of fraudulent postings.

- department field has a large proportion of missing values, suggesting it might not be commonly used.

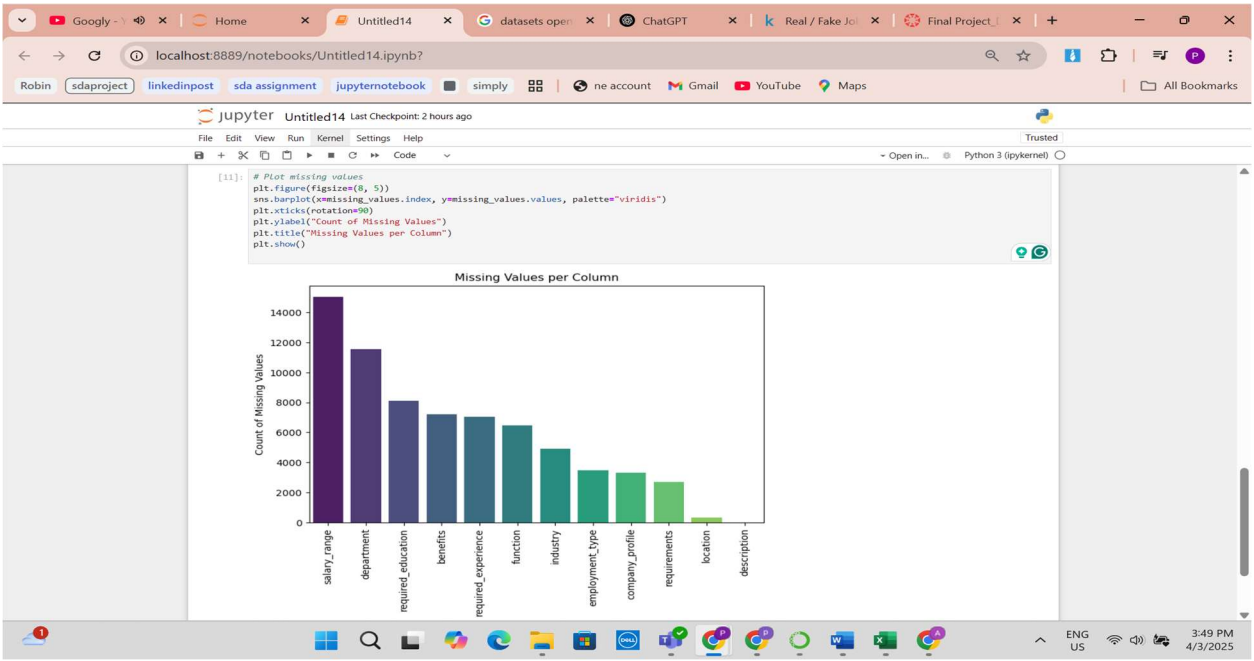


Figure 4: Possible Anomalies

The bar chart highlights the extent of missing values across various columns. The **department, salary_range, and required_education** fields have significant gaps, which may impact analysis.

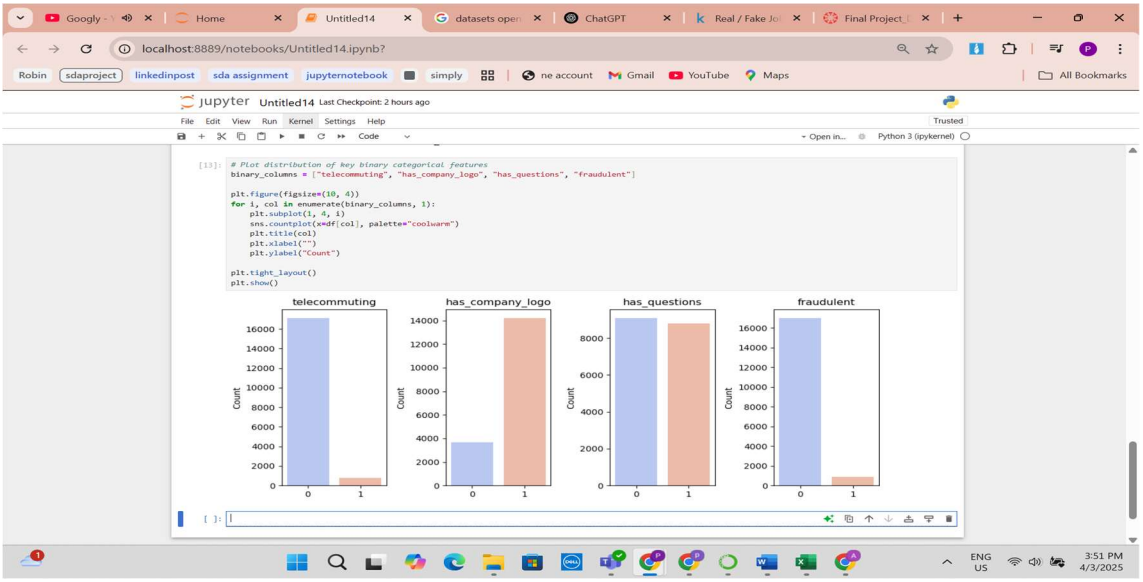


Figure 5: Possible Anomalies

The count plots show some interesting patterns:

- Most job postings do not allow telecommuting (remote work).
- Nearly all job postings have a company logo, suggesting that fraudsters might also include logos to appear legitimate.
- Most postings do not have screening questions, which could indicate either a lack of thorough vetting or a preference for a simpler application process.
- There are significantly fewer fraudulent postings compared to legitimate ones, suggesting an imbalanced dataset.

In the binary categorical feature plots (telecommuting, has_company_logo, has_questions, fraudulent), the blue and orange colors represent the two possible values (0 and 1). Specifically:

- Blue usually represents the count of 0 (False/No).
- Orange represents the count of 1 (True/Yes).

For example, in the fraudulent column:

- Blue represents legitimate job postings (0).
- Orange represents fraudulent job postings (1).

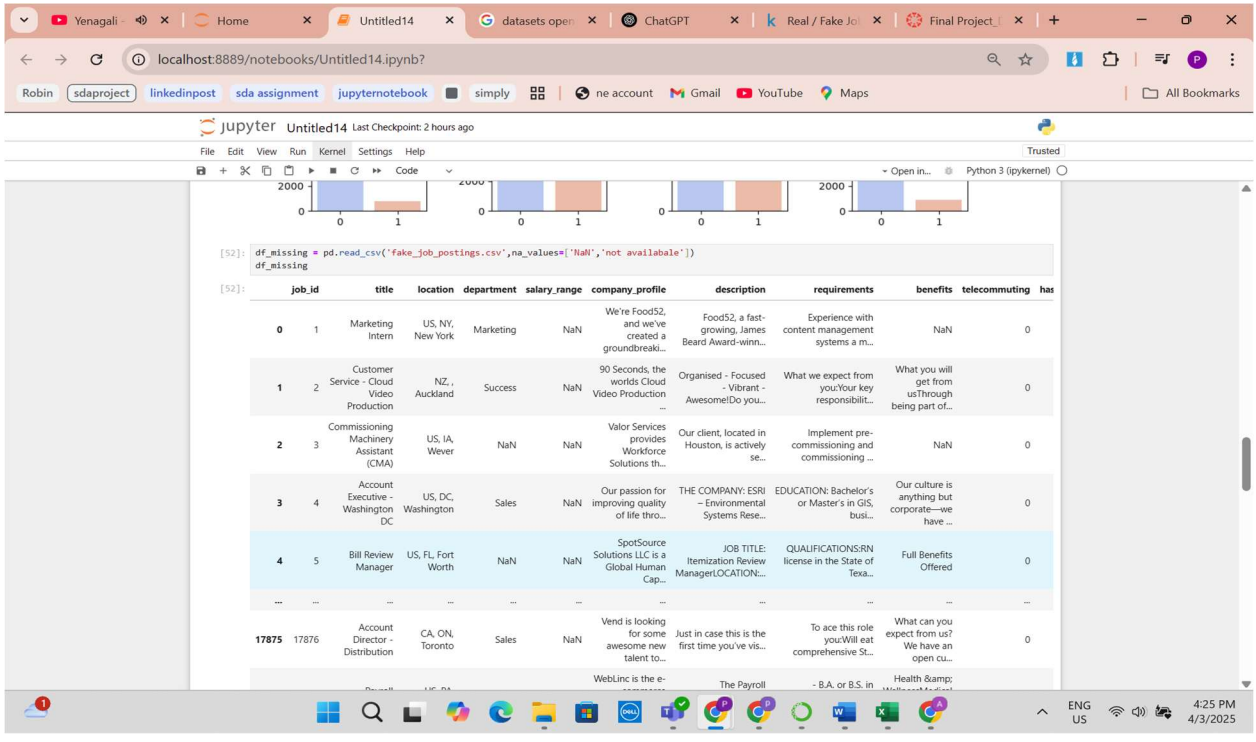


Figure 6: Missing Values

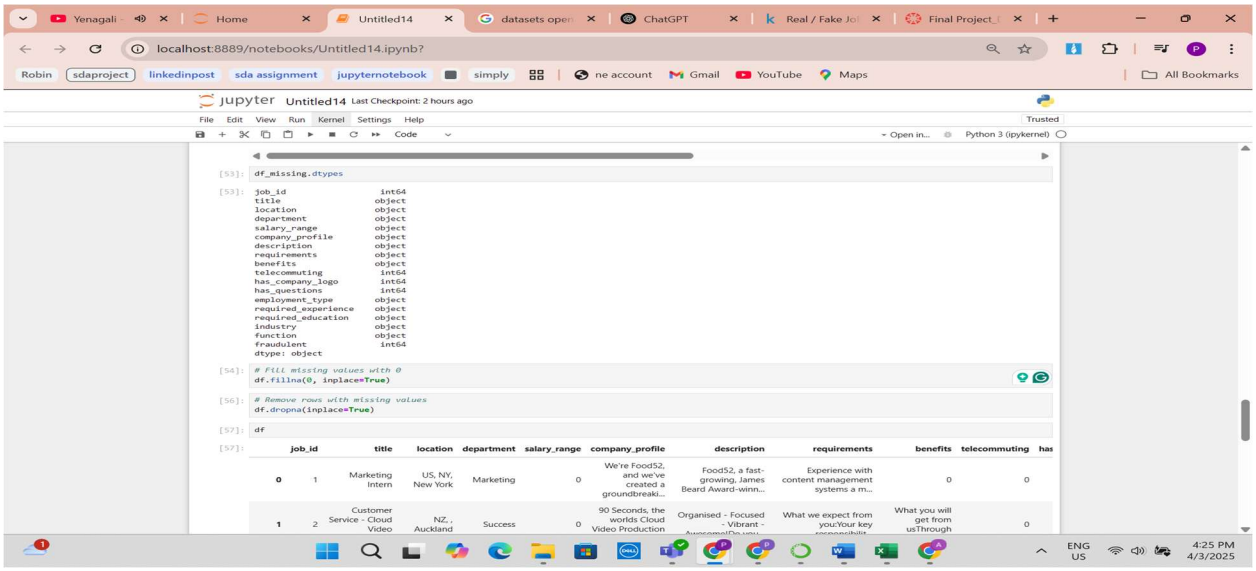


Figure 7: Filling Null Values

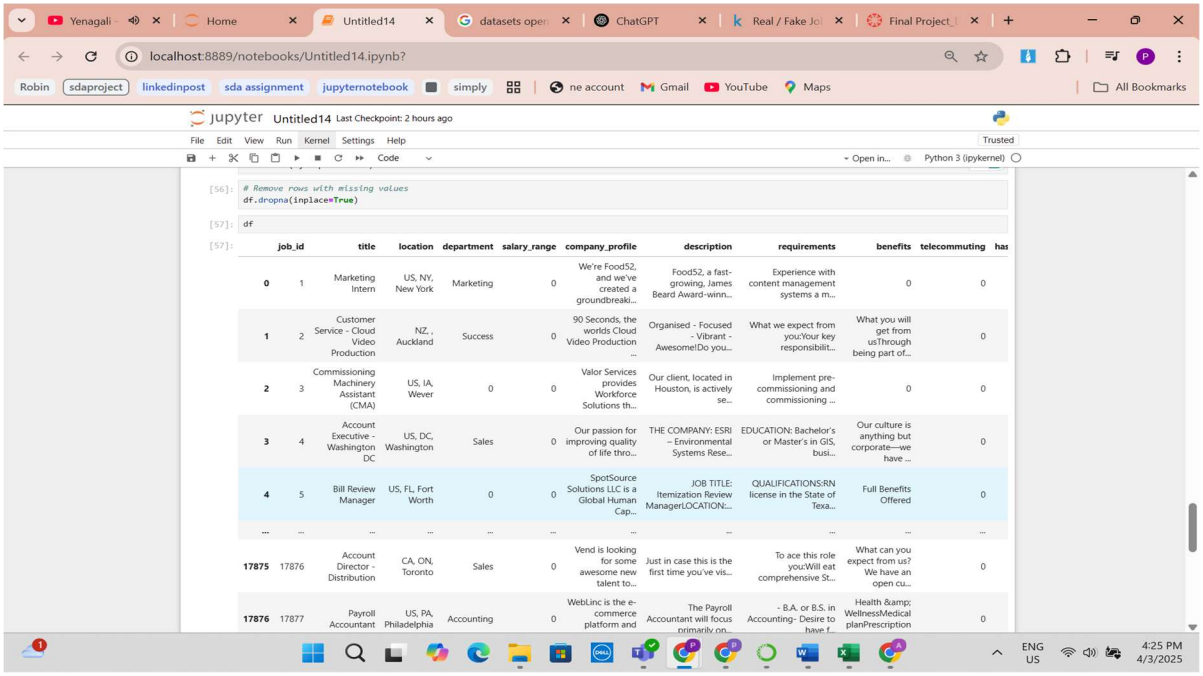


Figure 8: Dropping Values

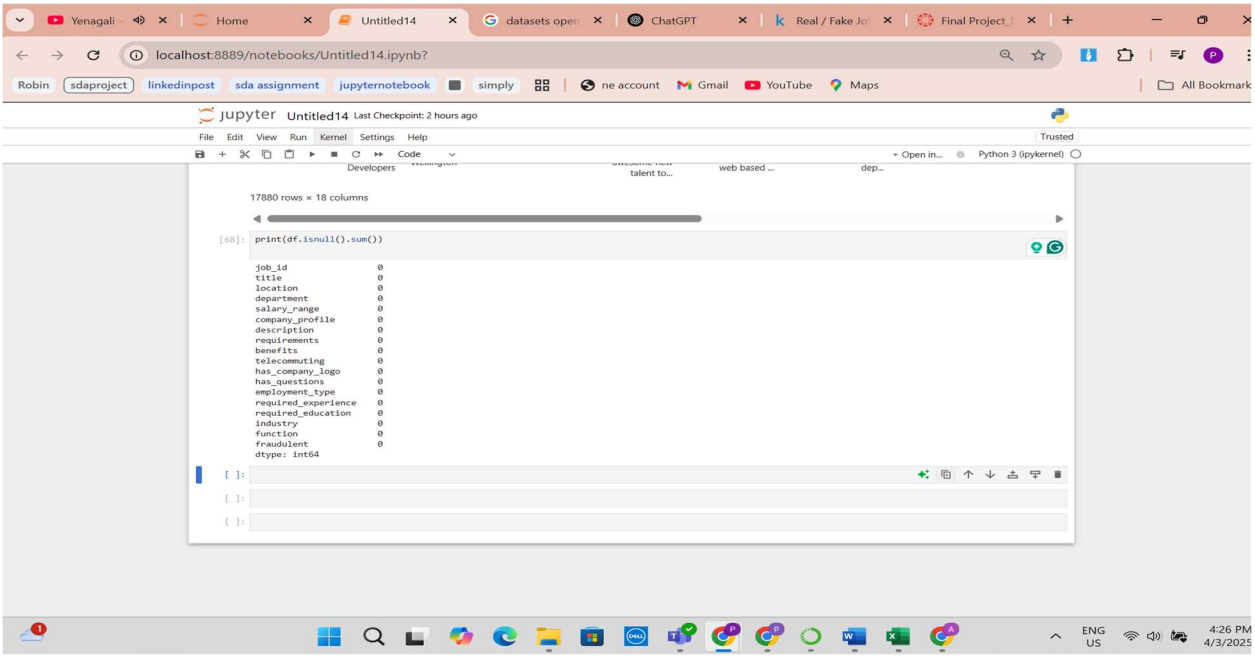


Figure 9: Checking Null Values

Outlier and Anomalies detection and removal:

Why It Matters

Anomaly detection is used in:

- Fraud detection
- Cybersecurity (intrusion detection)
- Healthcare (rare diseases)
- Manufacturing (defect detection)

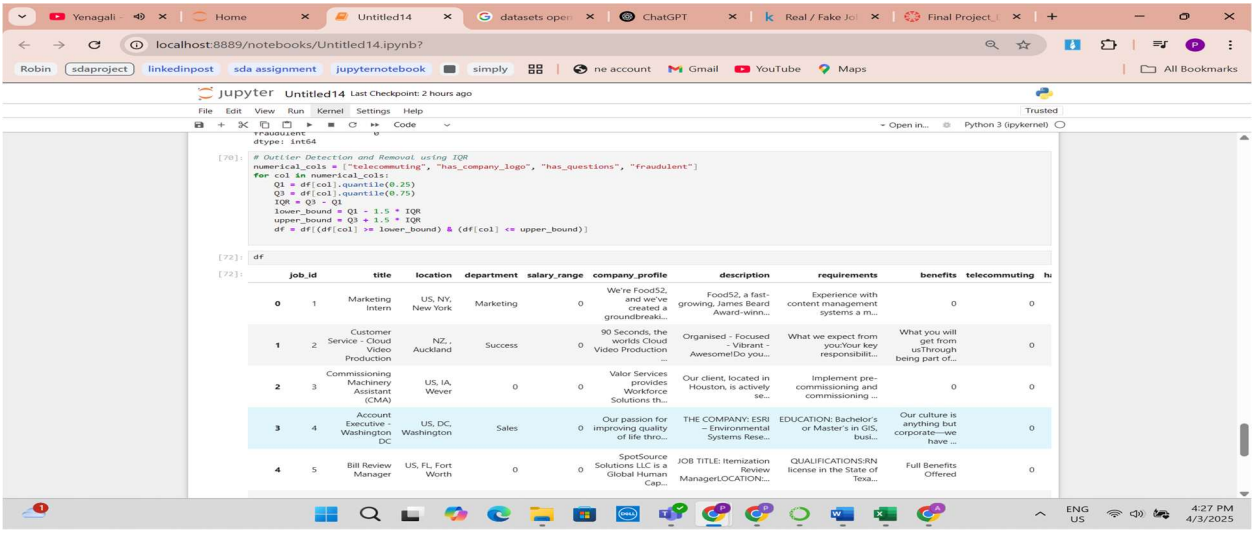
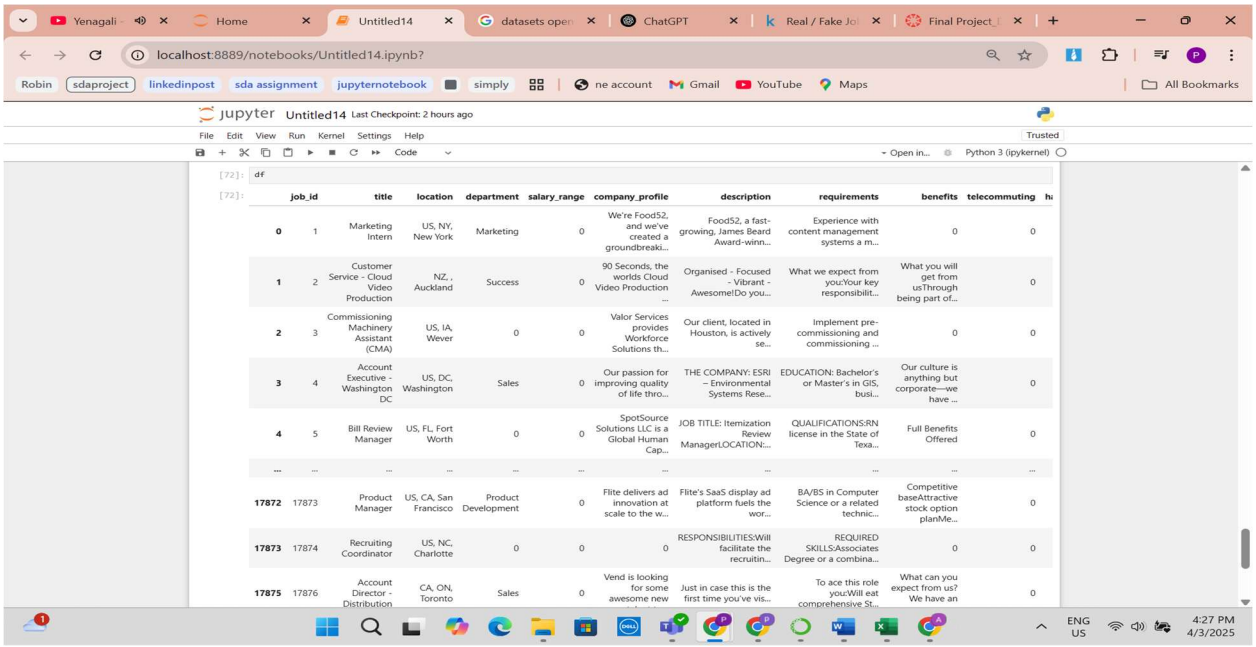


Figure 10: Data Cleaning



	job_id	title	location	department	salary_range	company_profile	description	requirements	benefits	telecommuting	ha
0	1	Marketing Intern	US, NY, New York	Marketing	0	We're Food52, and we've created a groundbreaki...	Food52, a fast-growing, James Beard Award-winn...	Experience with content management systems a m...	0	0	
1	2	Customer Service - Cloud Video Production	NZ, , Auckland	Success	0	90 Seconds, the worlds Cloud Video Production ...	Organised - Focused - Vibrant - Awesome!Do you...	What we expect from you:Your key responsibilit...	What you will get from usThrough being part of...	0	
2	3	Commissioning Machinery Assistant (CMA)	US, IA, Wever	0	0	Valor Services provides Workforce Solutions th...	Our client, located in Houston, is actively se...	Implement pre-commissioning and commissioning ...	0	0	
3	4	Account Executive - Washington DC	US, DC, Washington	Sales	0	Our passion for improving quality of life thro...	THE COMPANY: ESRI – Environmental Systems Rese...	EDUCATION: Bachelor's or Master's in GIS, busi...	Our culture is anything but corporate—we have ...	0	
4	5	Bill Review Manager	US, FL, Fort Worth	0	0	SpotSource Solutions LLC is a Global Human Cap...	JOB TITLE: Itemization Review ManagerLOCATION:...	QUALIFICATIONS:RN license in the State of Texa...	Full Benefits Offered	0	
...	...	...	...	...	...	...	...	...	...	...	
17872	17873	Product Manager	US, CA, San Francisco	Product Development	0	Flite delivers ad innovation at scale to the w...	Flite's SaaS display ad platform fuels the wor...	BA/BS in Computer Science or a related technic...	Competitive baseAttractive stock option planMe...	0	
17873	17874	Recruiting Coordinator	US, NC, Charlotte	0	0	0	RESPONSIBILITIES:Will facilitate the recruitin...	REQUIRED SKILLS:Associates Degree or a combina...	0	0	
17875	17876	Account Director - Distribution	CA, ON, Toronto	Sales	0	Vend is looking for some awesome new	Just in case this is the first time you've vis...	To ace this role you:Will eat comprehensive St...	What can you expect from us? We have an	0	

Figure 11: Outlier Detection and Removal

Outlier plots:

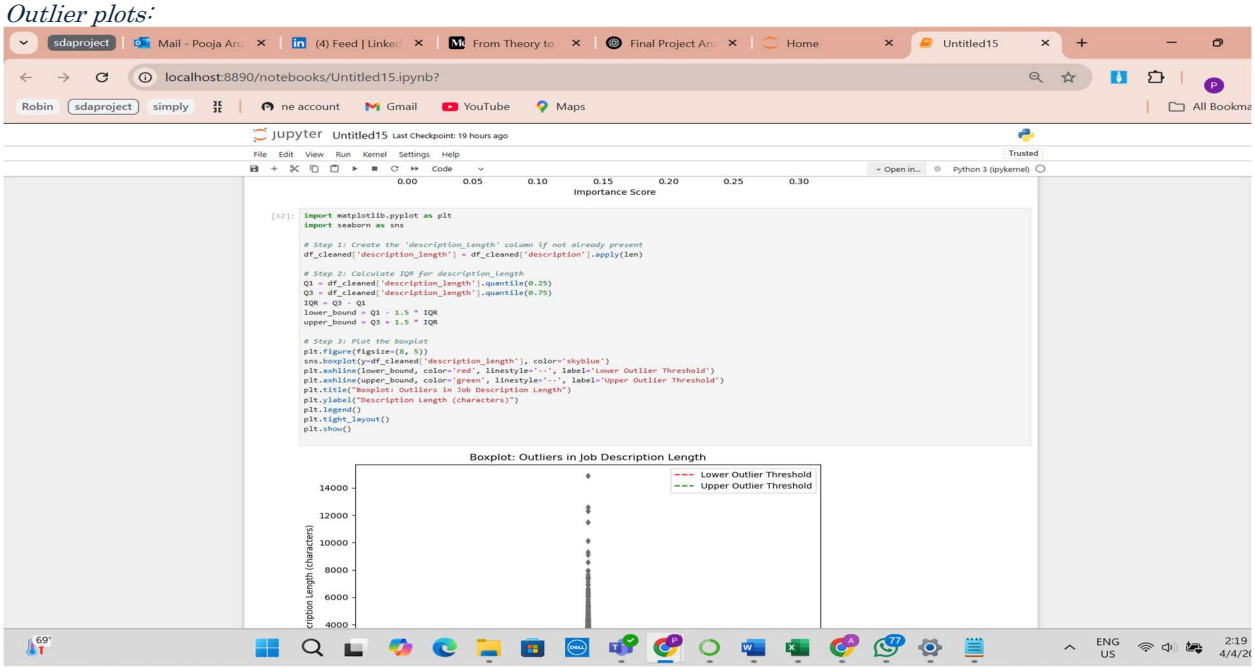


Figure 12: Outlier Detection and Removal

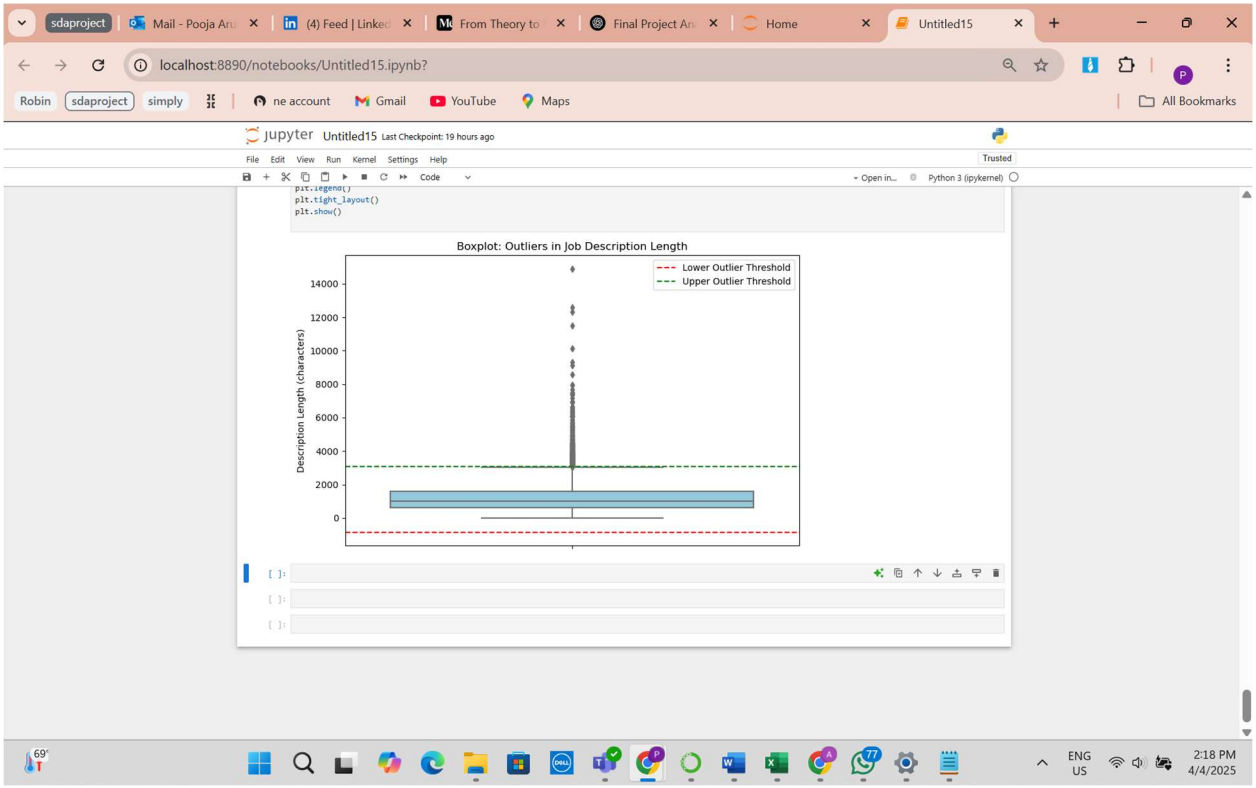


Figure 13:Outlier Detection and Removal

Boxplot Insight – Description Length Outliers

- The box represents the middle 50% of job descriptions:
    - These typically fall between ~500 and ~1500 characters, meaning most job posts have reasonable-length descriptions.
- The green line marks the upper threshold for normal values.
- The red line marks the lower threshold.
- The black dots above the green line are outliers — job descriptions that are unusually long, in some cases reaching over 14,000 characters!

Why this matters:

- Extremely short descriptions (below lower bound) could indicate:
    - Fake/scam listings with little effort put into detail
    - Attempt to avoid detection or copy/paste scams
- Extremely long descriptions (above upper bound) might mean:
    - Overcompensating for legitimacy
    - Spam or repeated irrelevant content
    - Copied content from multiple listings

Conclusion:

Outliers in description_length are valuable indicators of **irregular or suspicious behavior** in job postings. This supports the use of this feature in your fraud detection model and shows why it's important to visualize and monitor such anomalies.

2) Violin Plot Analysis: Job Description Length vs Fraudulent Label

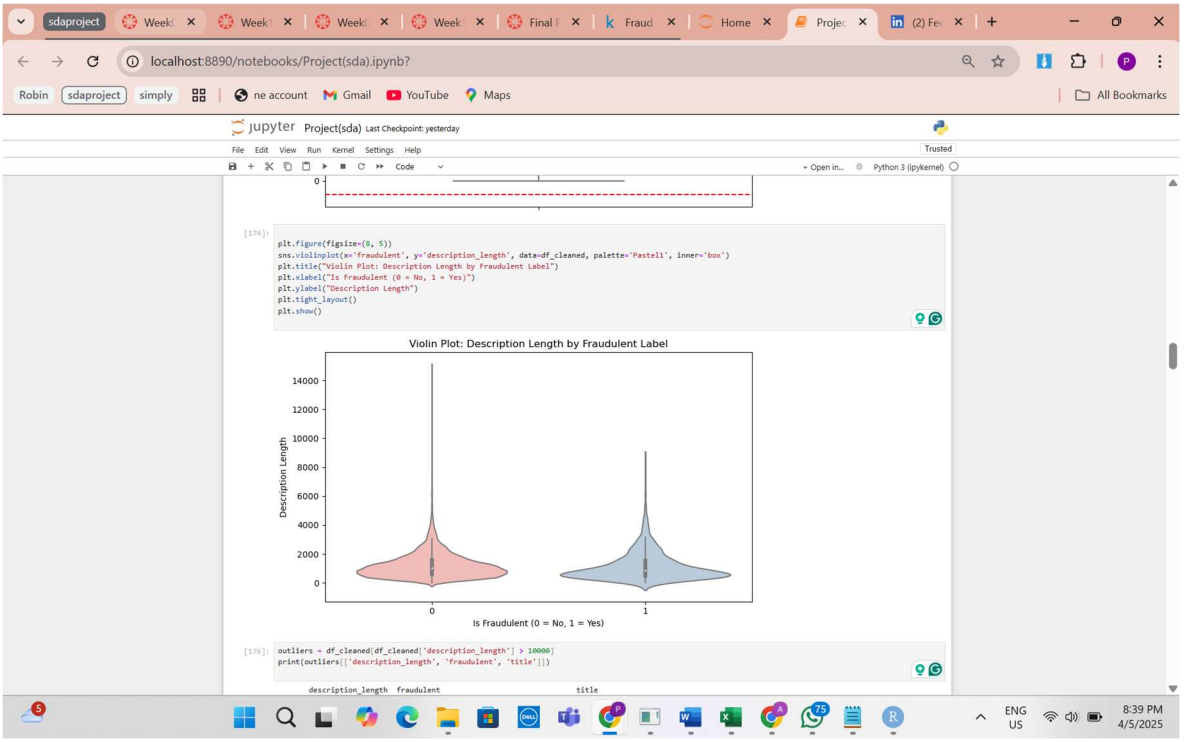


Figure 14: Outlier Detection and Removal

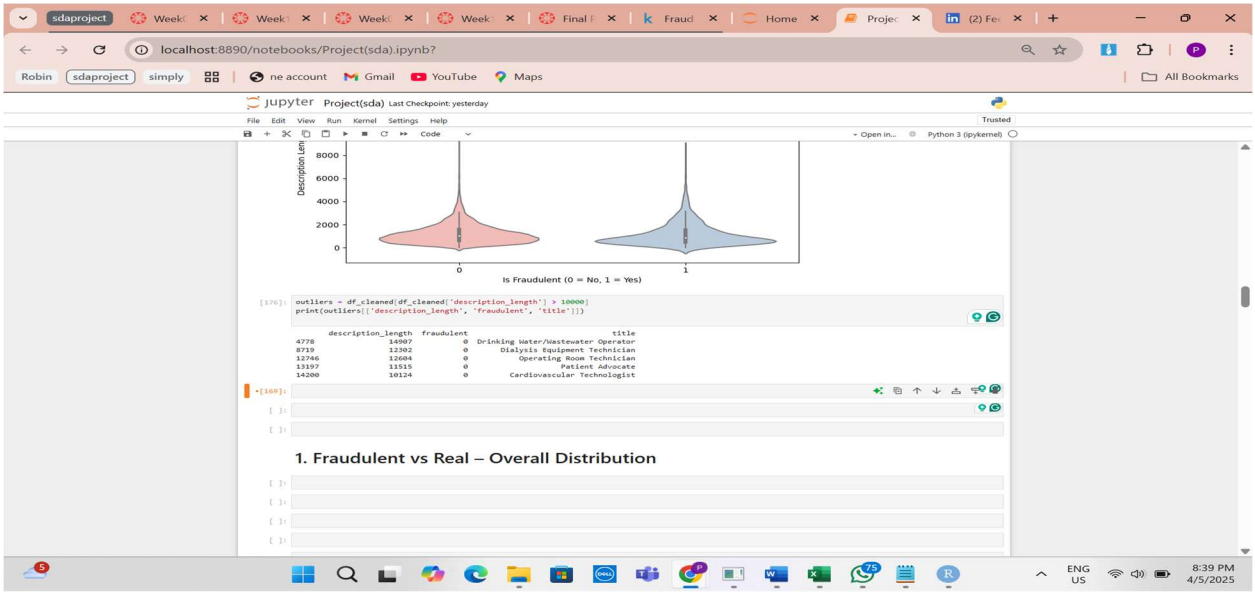


Figure 15: Outlier Detection and Removal

To visually inspect potential outliers in job descriptions, a violin plot was generated comparing the description_length distribution between fraudulent and non-fraudulent job postings.

The plot combines a boxplot and a kernel density estimate, allowing us to observe not only the central tendency and spread of each group but also the shape of their distribution.

Key Observations:

- Real job postings (fraudulent = 0) show a wide distribution, with most description lengths clustering between 500 to 2000 characters. However, there is a notable long tail extending beyond 14,000 characters, suggesting the presence of extreme outliers — potentially overly detailed or artificially inflated descriptions.
- Fraudulent job postings (fraudulent = 1) have a tighter distribution overall, mostly between 200 to 1500 characters, but still exhibit a few unusually long descriptions, indicating that some fake listings may attempt to appear legitimate through verbose or copied content.
- In both cases, the narrow vertical extensions (violin "necks") at the higher end of the y-axis highlight rare and extreme values — a clear visual signal of outliers.

Conclusion:

The violin plot effectively reveals outliers in job description length. These outliers may be important indicators in identifying patterns of deception in job postings. For example, extremely long descriptions in fake jobs may be a tactic to distract or mislead applicants. Conversely, legitimate jobs with unusually long postings may warrant manual review to validate quality and consistency.

Effect of Outlier / anomalies Removal:

To understand the impact of outliers in the description_length feature, we generated violin plots comparing the distribution of real and fake job descriptions before and after IQR-based outlier removal.

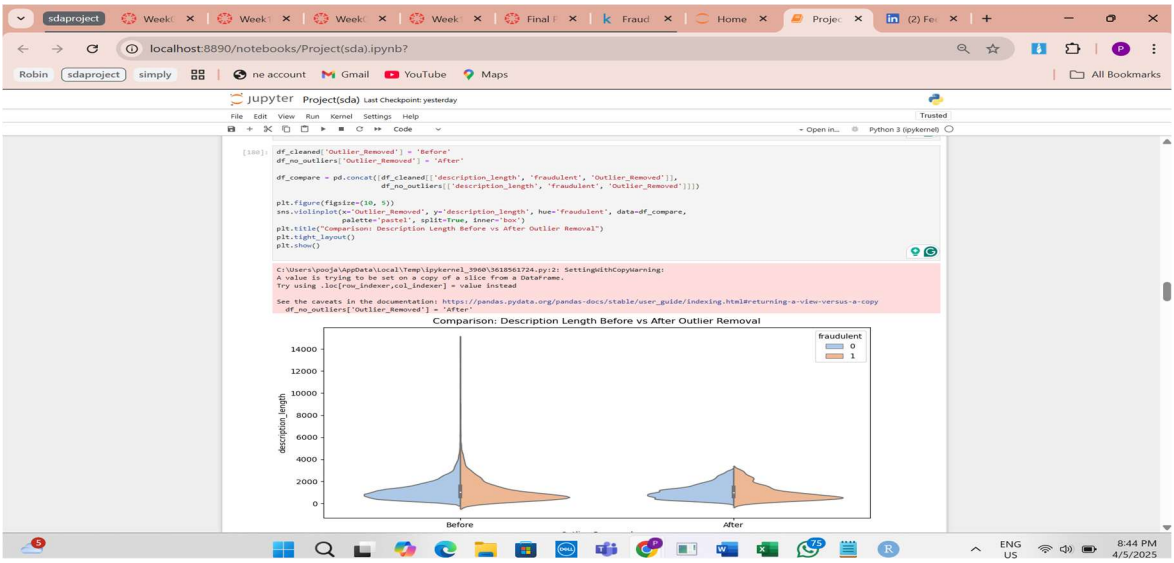


Figure 16: Outlier Detection and Removal effect

What the Visualization Shows:

- Before outlier removal, the distribution for real jobs (fraudulent = 0) had a long upper tail stretching past 14,000 characters. This indicates the presence of extreme outliers — job postings with unusually long descriptions that skewed the overall distribution.
- For fake jobs (fraudulent = 1), the spread was already narrower but still included a few high-end outliers, suggesting some fraudulent postings may be artificially verbose.
- After removing outliers using the IQR method:
    - The distribution became tighter and more symmetrical for both classes.
    - The extreme long tails disappeared, making it easier to compare the core patterns of description lengths between real and fake job postings.
    - This resulted in a more robust and interpretable view of the typical behavior in the dataset.

Conclusion:

Outlier removal significantly reduced skewness in description_length, allowing for more meaningful comparisons between real and fake job descriptions. This step is especially useful before modeling, as it prevents extreme values from dominating the learning process. However, since these outliers may carry signals of fraud, they should also be examined separately for deeper insight.

2)

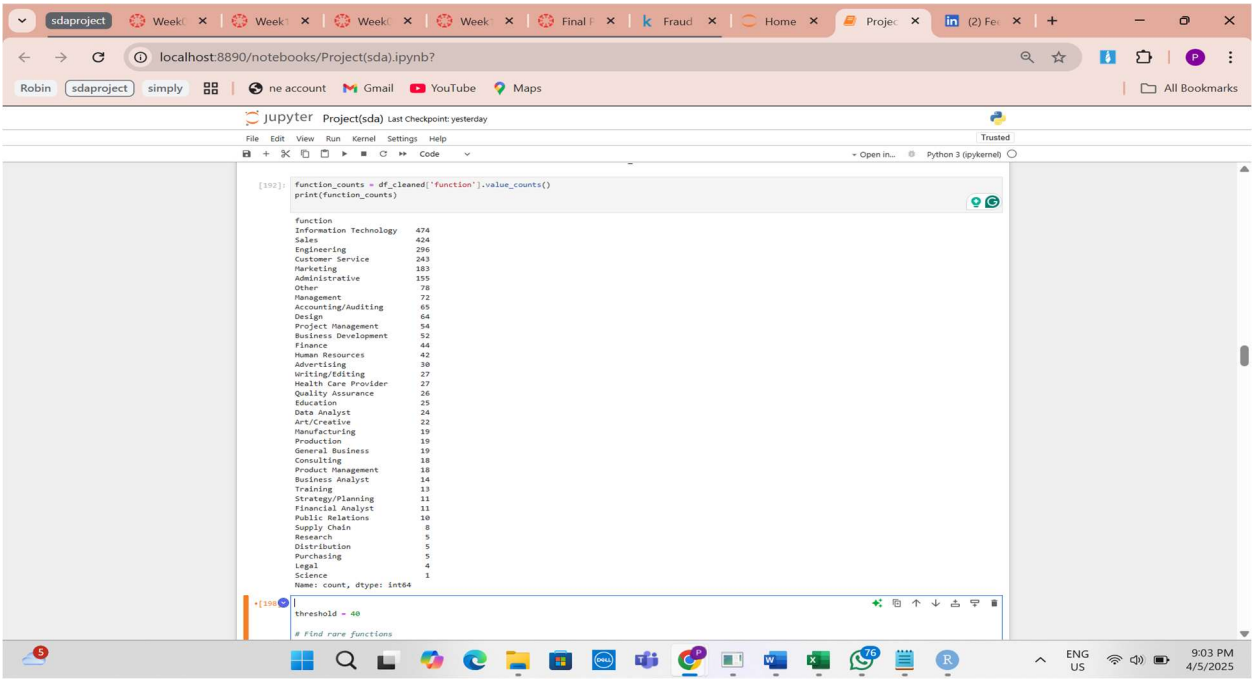


Figure 17: Anomalies Detection

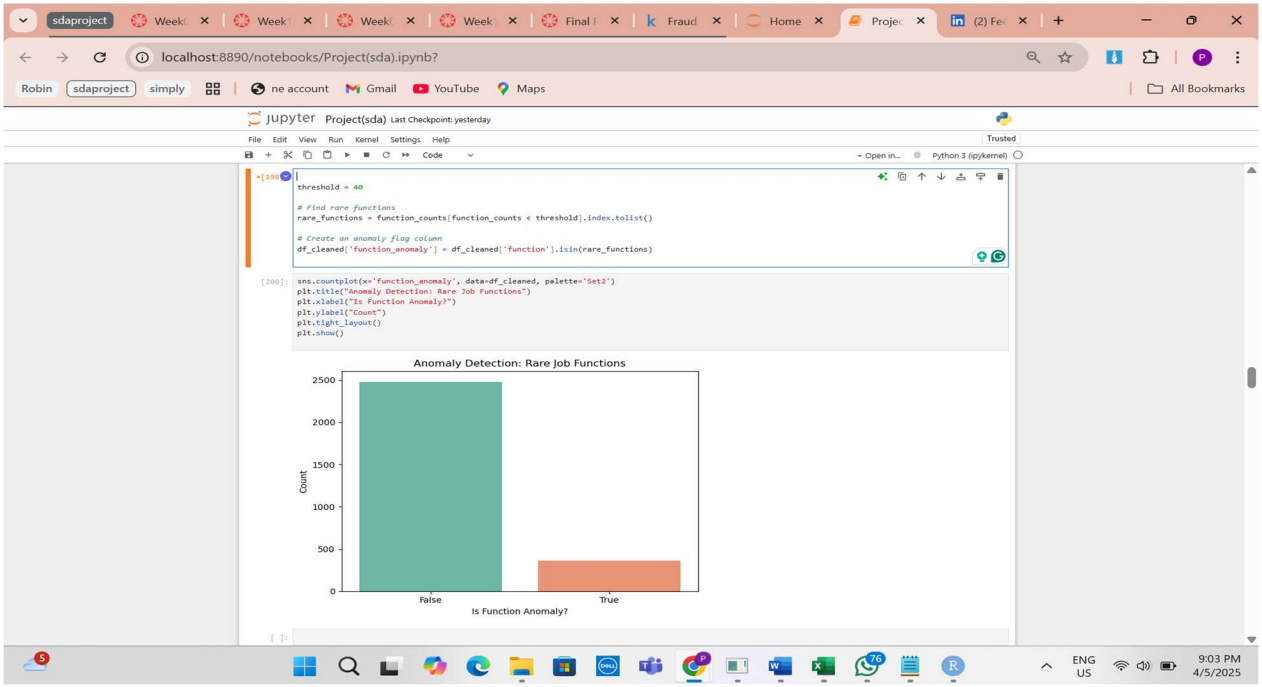


Figure 18: Anomalies Detection

What does this plot show?

Plotted a count of job postings split by:

- function_anomaly = False (Normal/common job functions)
- function_anomaly = True (Rare/unusual job functions)

With threshold of 40, defined rare functions as those appearing fewer than 40 times in the dataset.

Anomaly Detection in Job Function Column

To identify unusual patterns in job functions, we applied categorical anomaly detection based on frequency. Job functions that appeared less than 40 times in the dataset were flagged as anomalies.

Findings:

- Out of all job postings, approximately 350 entries had rare job functions.
- These rare functions could indicate:
    - Niche or specialized roles
    - Data inconsistencies (misspelled or inconsistent naming)
    - Potential fraudulent postings trying to mask under unfamiliar job titles

Conclusion:

Rare job functions represent a non-negligible portion of the dataset and should be carefully reviewed. These anomalies may hold valuable signals in distinguishing fraudulent job posts from legitimate ones.

EDA:

1) Employment Type as a Distinguishing Factor in Fraudulent Job Postings

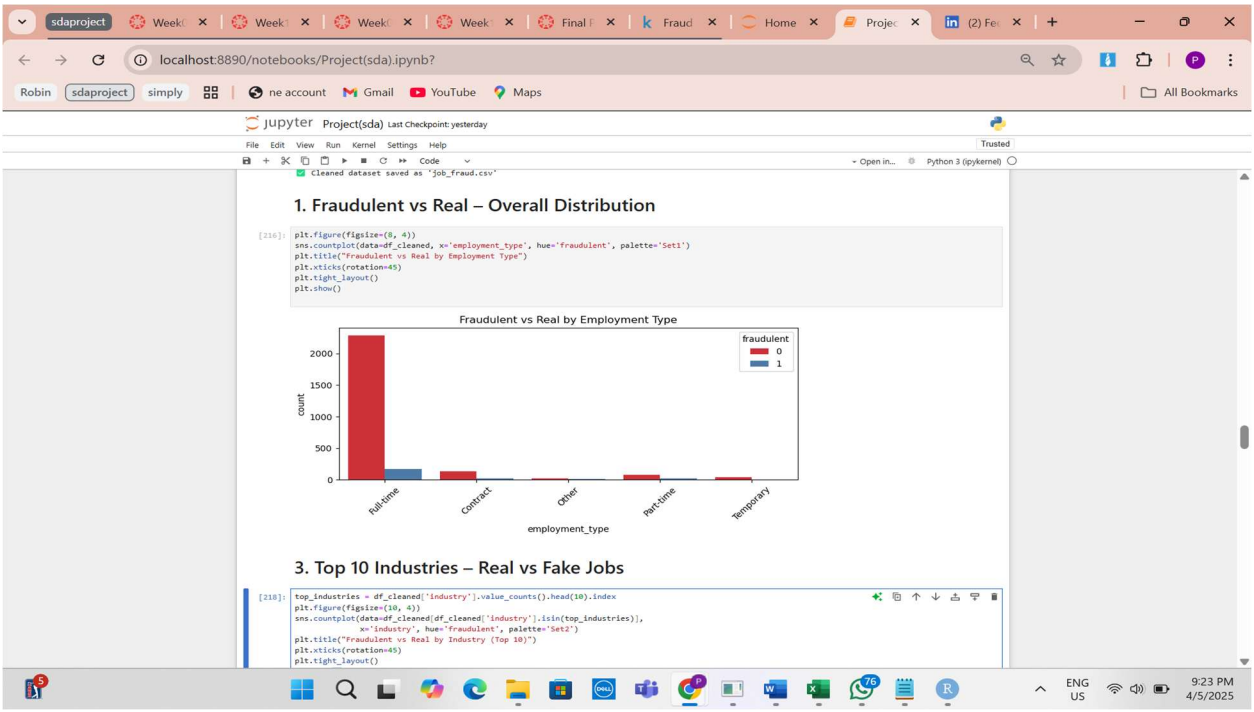


Figure 19: Employment Type

The analysis of employment types across fraudulent and legitimate job postings shows a clear disparity, offering valuable insights into patterns used by fake job listings.

Key Observations:

- Fraudulent job postings are heavily concentrated under the “Full-time” category.
- Legitimate postings are more evenly distributed across various employment types, including Contract, Part-time, and Temporary.
- This concentration suggests that fraudsters intentionally label fake jobs as full-time to increase credibility and appeal to a broader audience.

Why This Matters:

- Full-time positions appear more legitimate, making them an easy tool for fraudsters to gain trust quickly.
- Real jobs reflect greater variety, aligning with actual hiring trends and flexibility in employment structures.

- Monitoring employment type can help identify red flags during the early stages of detection.

Implications for Detection:

- The employment_type field can serve as a predictive feature in identifying fraudulent job posts.
- When combined with other signals (e.g., vague descriptions, missing logos, unrealistic salaries), full-time labels on questionable postings may increase the likelihood of fraud.
- This reinforces the value of using data analysis techniques to uncover behavioral patterns that are not immediately obvious through manual review.

2) **Industry Distribution in Real vs Fraudulent Job Postings**

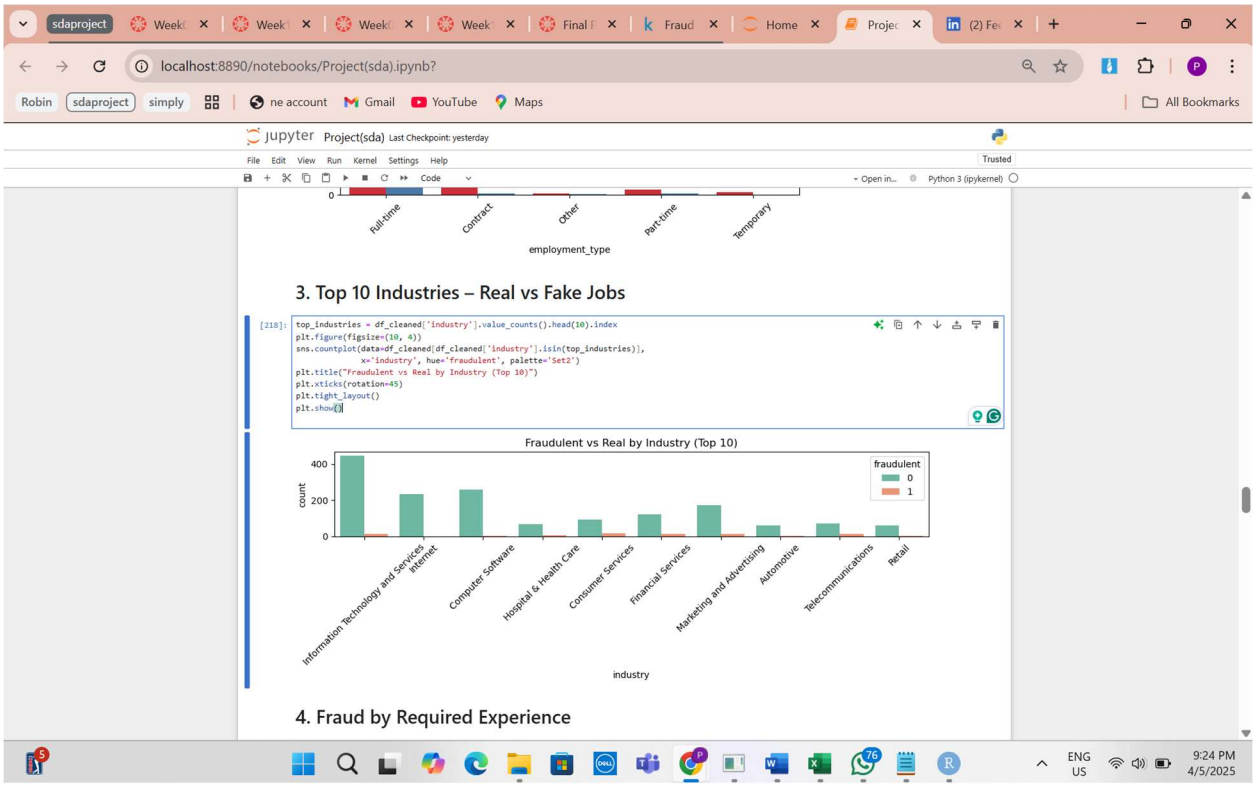


Figure 20: Top 10 Industries

Visualization compares the distribution of fake and real job postings across the top 10 most common industries in the dataset. The findings reveal distinct differences in how fraudsters target specific sectors.

Key Observations:

- Industries like Information Technology & Services, Internet, and Computer Software have the highest number of legitimate job postings — and also the most fraud activity among all categories.

- Fraudulent postings are relatively less common in industries such as Retail, Telecommunications, and Automotive.
- Highly digital sectors are more exposed to fake jobs, likely because they attract more online applicants and are easier for scammers to mimic.

Why This Matters:

- Fraudsters may target fast-growing, tech-oriented industries to exploit high demand and applicant volume.
- Real jobs show broad diversity across industries, whereas fake postings are concentrated in specific high-tech sectors, making industry type a useful predictive feature.
- Identifying industry-based risk zones can help in prioritizing fraud detection efforts or implementing stricter screening in high-risk sectors.

Implications for Detection:

- Industry is a valuable feature in classification models used for detecting fake job postings.
- Jobs in industries like IT, Internet, and Software should be examined more closely when paired with other risk factors.
- Platforms and job boards can leverage this information to allocate risk scoring to specific industries for enhanced security.

3) **Required Experience in Real vs Fraudulent Job Postings:**

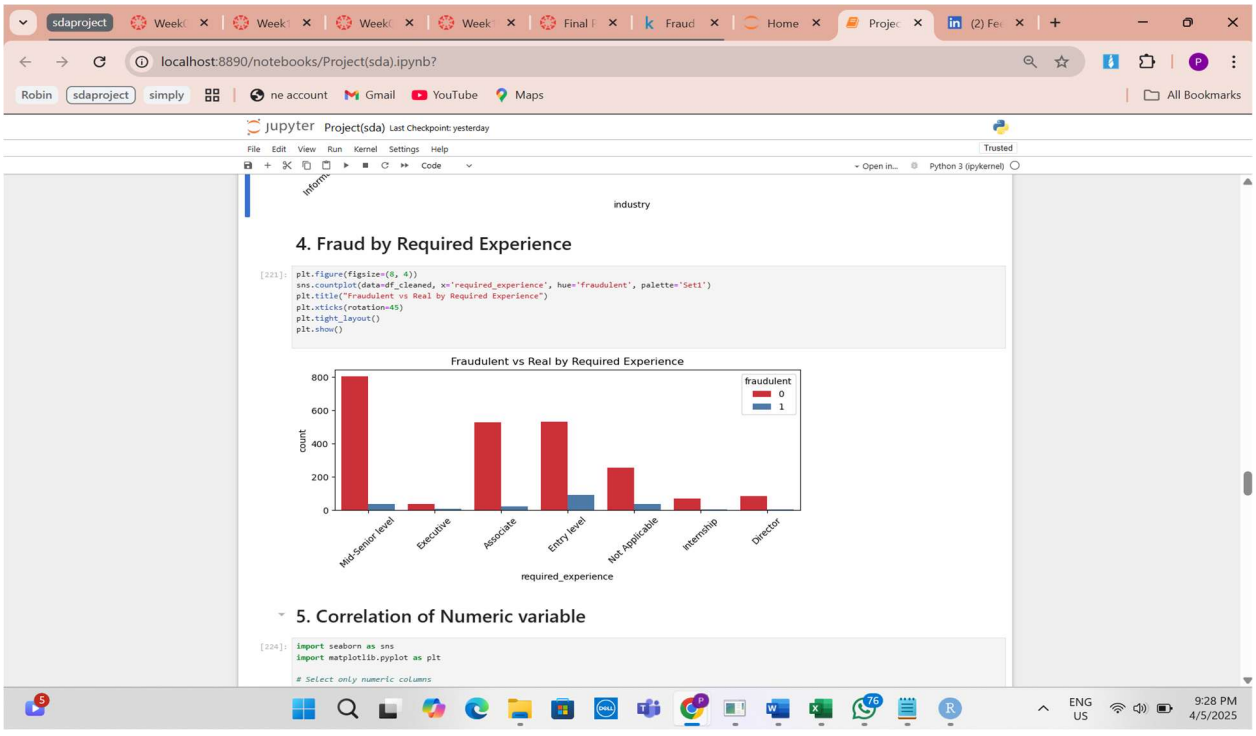


Figure 21: Fraud By required experience field

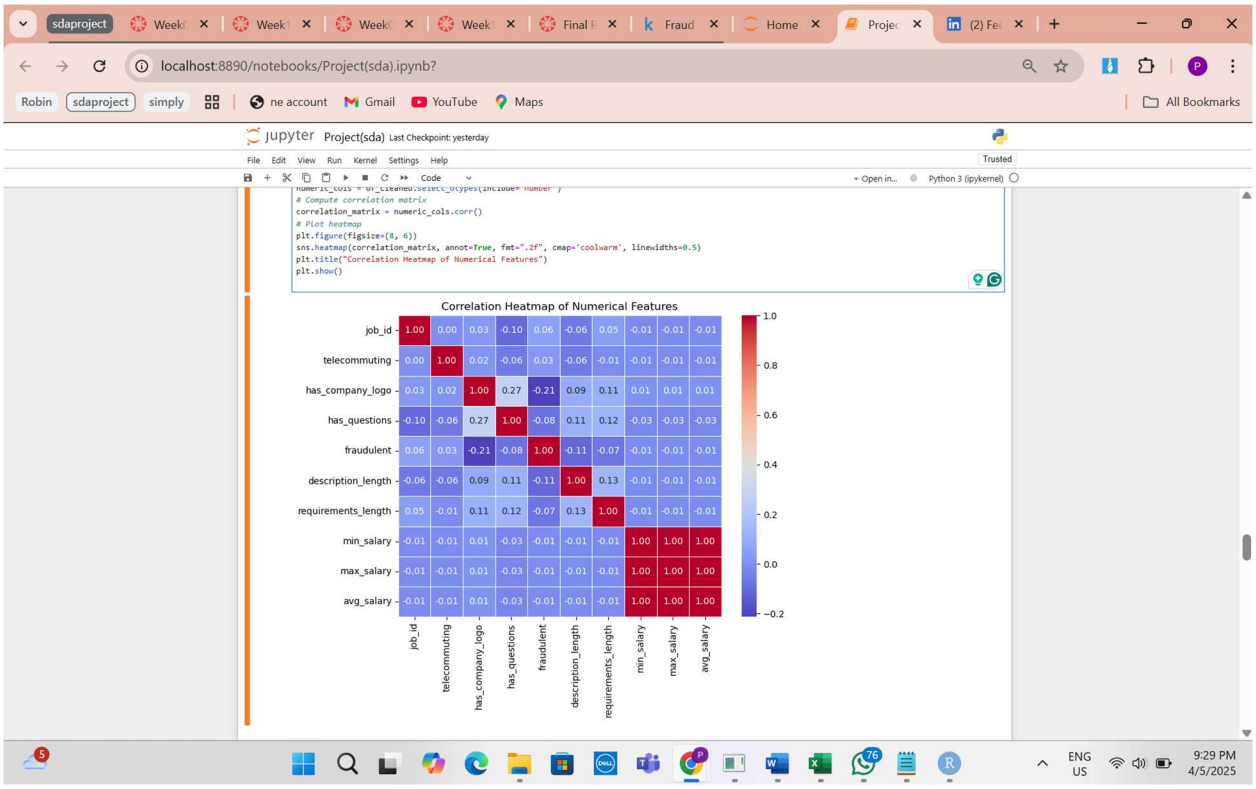


Figure 22: Correlation Heatmap

This heatmap visualizes the relationships between numerical variables in the dataset, particularly focusing on how each correlates with the target variable fraudulent.

Key Observations:

- The variable has_company_logo shows a negative correlation (-0.21) with fraud, suggesting that fraudulent job postings are more likely to lack a company logo.
- has_questions has a positive correlation (0.27) with fraudulent, implying that fraudulent jobs often include screening questions, possibly to seem more legitimate.
- Other features like description_length, requirements_length, and salary-related fields show very weak correlations with fraud — indicating that length or salary alone may not be strong individual predictors.

Why This Matters:

- Identifying correlations helps prioritize which features to focus on when building predictive models.
- Features like presence of a company logo and inclusion of screening questions could be strong flags in detecting fraudulent listings.
- Low correlation doesn't always mean useless — it may still matter in multivariate models (like random forests), but is less informative in isolation.

Implications for Detection:

- Correlation analysis confirms that certain binary flags (like logo presence or questions) are statistically linked to fraud and can be used to enhance model performance.
- While some features (like salary and description length) are important for analysis, their direct correlation with fraud is low — suggesting they are supporting signals rather than primary indicators.
- This analysis ensures that model-building focuses on the most informative features for detecting anomalies and fraud patterns.

5) Employment Type Distribution Across Top Industries

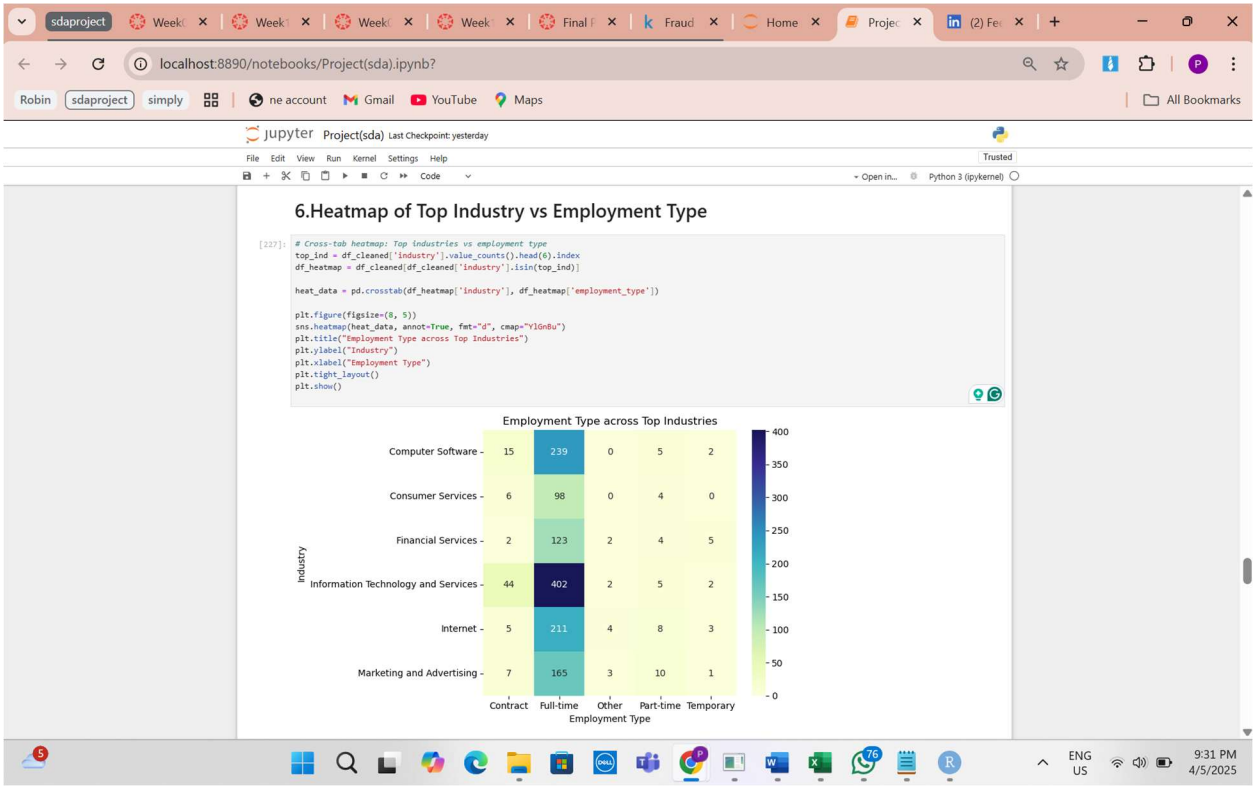


Figure 23: Employment time across Industries

This heatmap shows the cross-distribution of employment types within the top six most common industries in the dataset. It highlights how different industries prefer or rely on specific types of employment contracts.

Key Observations:

- Full-time employment is the most dominant type across all top industries.
- Information Technology and Services and Computer Software have the highest concentration of full-time roles, with over 400 and 200 postings respectively.
- Contract roles are minimal in comparison, especially in industries like Internet and Consumer Services.

- Employment diversity (e.g., inclusion of part-time, temporary) is limited in most industries, suggesting a hiring preference toward long-term full-time positions.

Why This Matters:

- Industries with a high share of full-time roles may also be more vulnerable to fraudulent postings, as scammers tend to mimic standard hiring practices to appear trustworthy.
- Understanding how employment types vary by industry helps build context-aware fraud detection systems — for instance, a part-time job in an industry that almost always hires full-time might be worth flagging.

Implications for Analysis & Detection:

- When combined with earlier fraud detection findings, this heatmap strengthens the case for using employment type + industry as a joint feature in predictive models.
- Industries that post mostly full-time jobs could require additional validation layers to ensure job authenticity.
- This cross-tab analysis improves targeted monitoring, allowing fraud detection efforts to focus on the most vulnerable industry-employment combinations.

6) Common Words in Real Job Descriptions

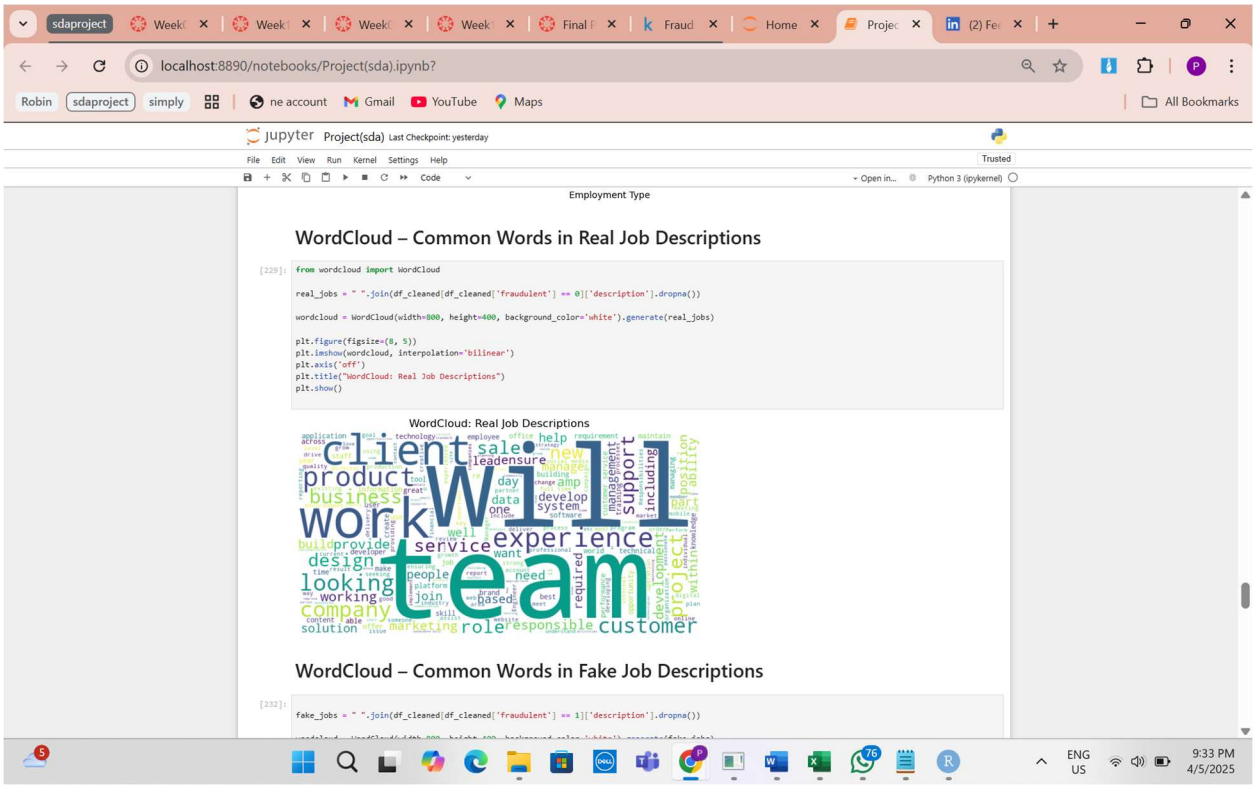


Figure 24: Word Cloud (Common Words in Real Jobs)

This WordCloud highlights the most frequently used words in real (non-fraudulent) job postings, revealing key characteristics of legitimate listings.

Key Observations:

- Prominent words include “team,” “client,” “work,” “experience,” “service,” and “company.”
- The language strongly emphasizes collaboration, service delivery, and professional experience.
- Terms like “looking,” “role,” and “responsible” suggest a formal hiring structure where expectations are clearly defined.

Why This Matters:

- Real job descriptions focus on organizational fit, job responsibilities, and required qualifications.
- They use structured, professional language that reflects genuine business needs and expectations.
- These words align with standard corporate terminology and suggest clear job objectives and deliverables.

Implications for Detection:

- The presence of such words can be used as positive indicators in distinguishing real job descriptions.
- Language models or text classifiers can be trained to recognize patterns aligned with legitimate job postings, increasing accuracy in fraud detection.
- This baseline helps set up a clear contrast when analyzing the fake job descriptions next.

7) Common Words in Fake Job Descriptions:

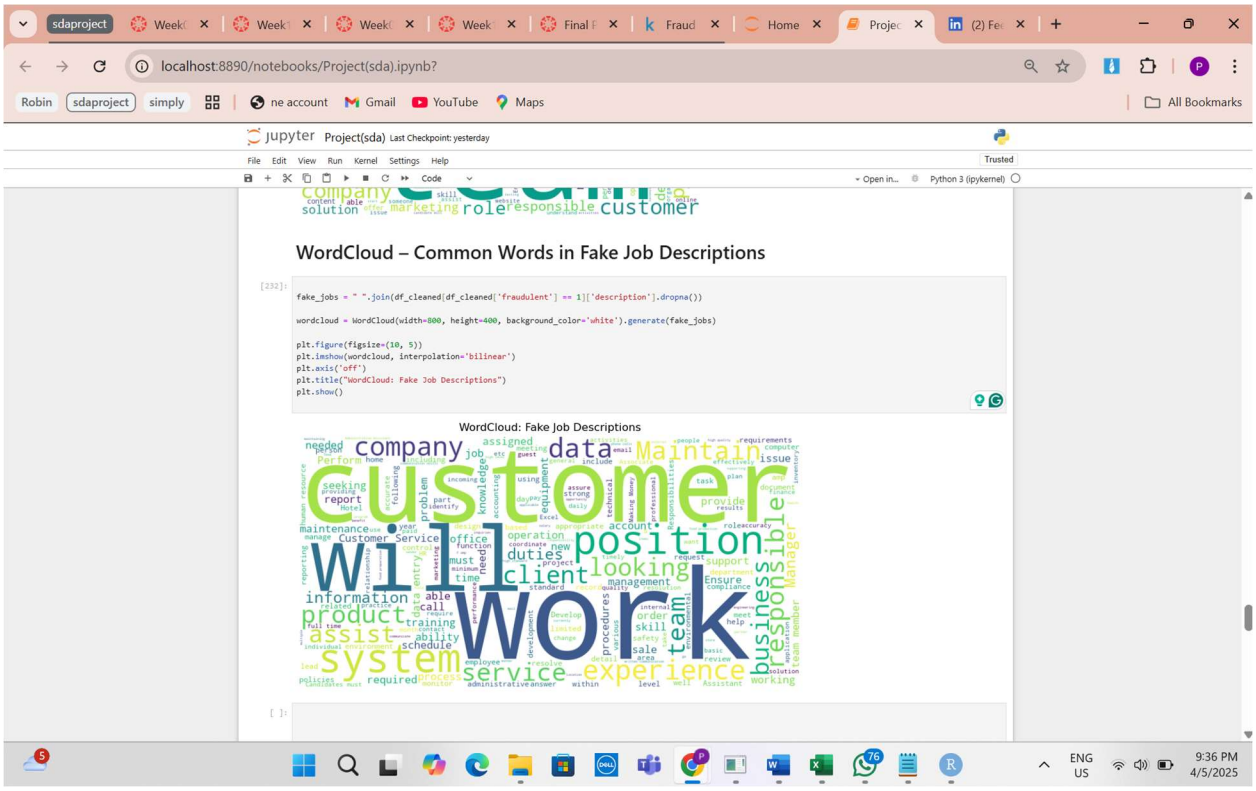


Figure 25:Common Words in Fake Jobs

The WordCloud visualization for fraudulent job postings displays the most frequently used words in fake job descriptions. These words help uncover the linguistic patterns scammers rely on to attract candidates.

Key Observations:

- Prominent words include “customer,” “work,” “system,” “position,” “maintain,” “assist,” and “product.”
- Many of these words are generic, vague, or task-oriented, lacking the contextual richness seen in real job descriptions.
- The repeated use of terms like “assist,” “maintain,” “call,” and “required” reflects basic role language, often without specific responsibilities or qualifications.

Why This Matters:

- Fake job descriptions tend to use broad, ambiguous terms to appeal to a wide audience and avoid scrutiny.
- Unlike real jobs, which focus on teamwork, goals, and qualifications, fake postings lack personal, company-specific, or outcome-driven language.
- This generic phrasing is likely intentional, allowing scammers to copy-paste across different roles and platforms.

Comparison with Real Job Descriptions:

Real Jobs Emphasize	Fake Jobs Emphasize
Team collaboration	Basic tasks
Experience and client handling	System maintenance, customer
Specific goals or platforms	Vague roles like "assist"
Role clarity and responsibility	Generic and repeated phrases

Implications for Detection:

- The **absence of context-rich keywords** can be a strong indicator of fraud.
- NLP techniques like TF-IDF, keyword frequency, or embedding-based classifiers can be used to **differentiate authentic descriptions from fakes**.
- Text-based patterns are critical for detection, especially when metadata (like salary or logo) is missing.

8) Overall Distribution of Real vs Fake Job Postings

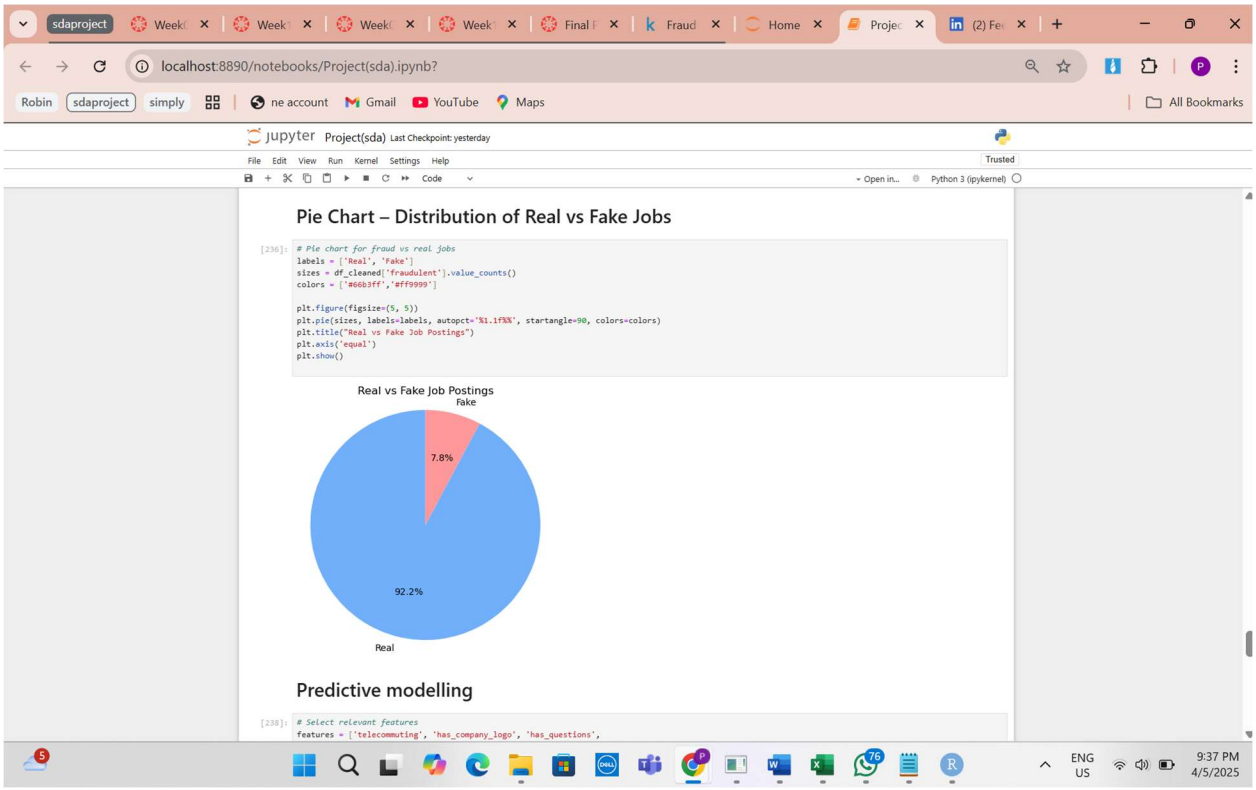


Figure 26: Real vs Fake job distribution

This pie chart presents the class distribution in the dataset, separating real and fake job listings based on the fraudulent label.

Key Observations:

- Real job postings make up 92.2% of the dataset.
- Fake job postings account for just 7.8%, highlighting a significant class imbalance.

Why This Matters:

- The dataset is heavily imbalanced, which is common in fraud detection problems.
- While fake postings represent a small portion, they are critical to detect, as they can cause real-world harm (e.g., scams, identity theft).
- Class imbalance poses a challenge in modeling — traditional accuracy metrics become misleading (e.g., predicting "all real" would still yield >90% accuracy).

Implications for Modeling:

- Balanced accuracy, precision, recall, and F1-score are better metrics for evaluating fraud detection models.
- Techniques such as oversampling (e.g., SMOTE) or undersampling may be necessary to ensure the model pays attention to the minority class (fraud).
- It emphasizes the importance of feature selection, anomaly detection, and careful model evaluation to handle imbalanced data effectively

Predictive Modelling:

Predictive Modeling: Detecting Fraudulent Job Postings

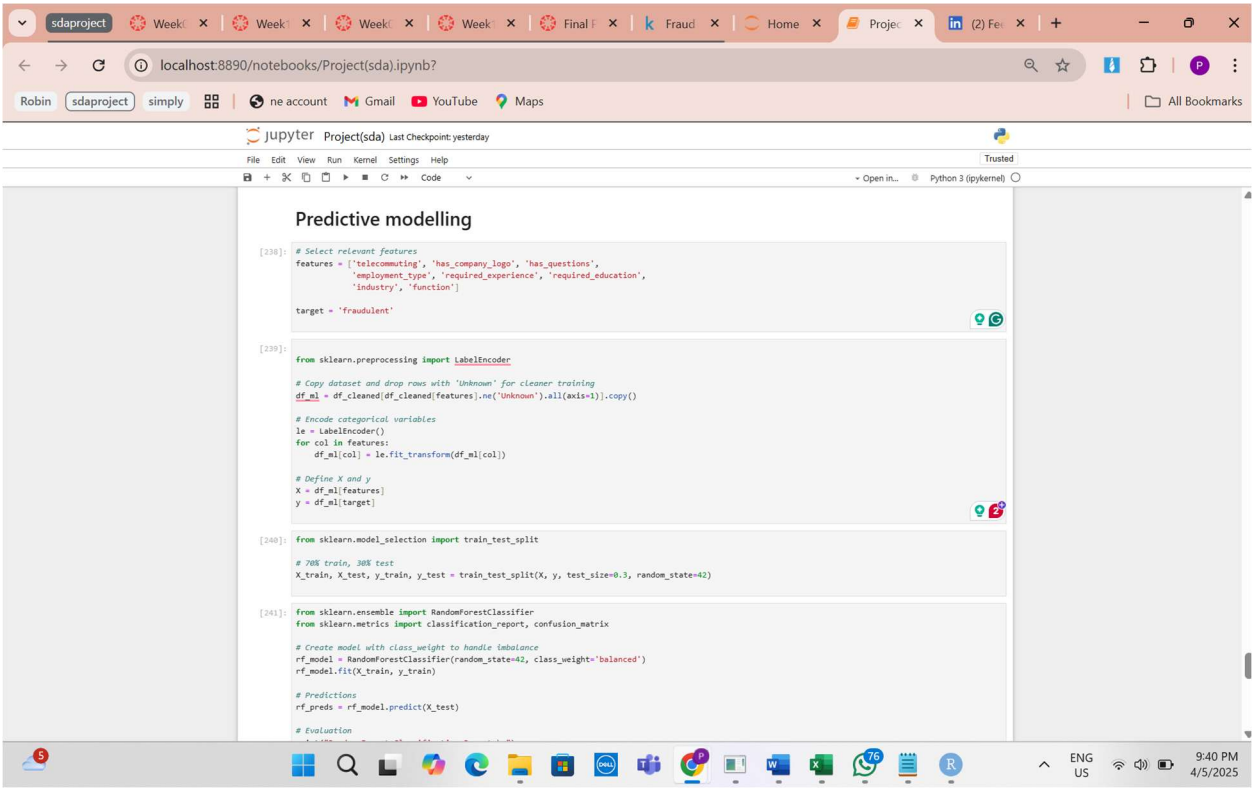


Figure 27: Predictive Modelling

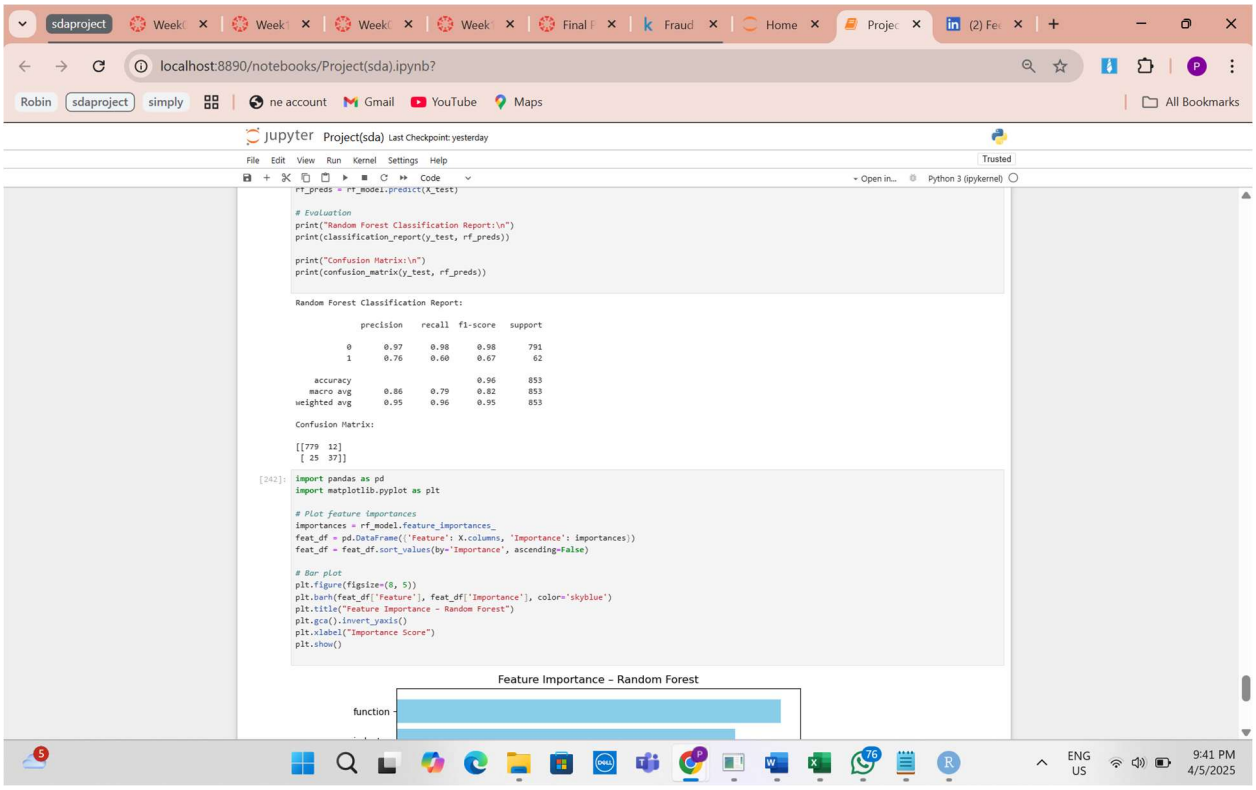


Figure 28: Random Forest Classifier Report

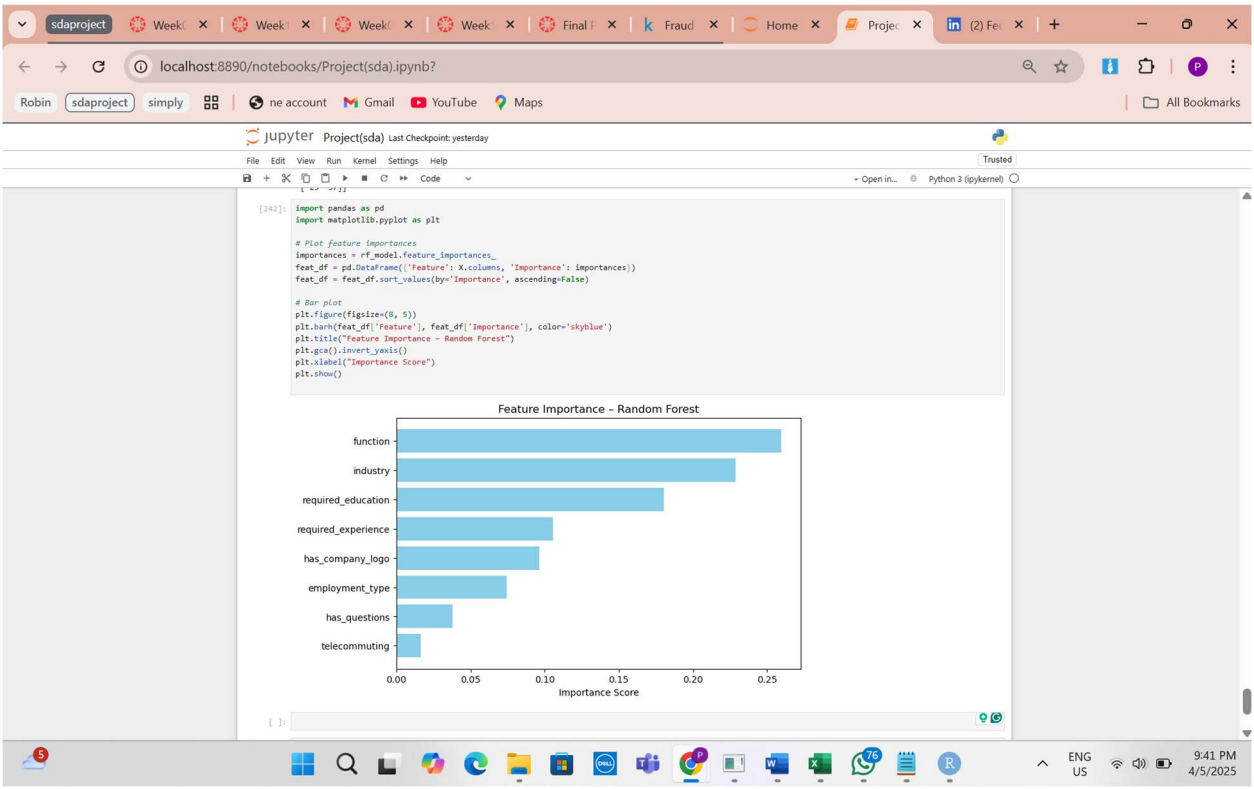


Figure 29: Feature Importance Score

A Random Forest classifier was trained using selected categorical features to predict whether a job posting is fraudulent. The model handled class imbalance using class_weight='balanced', ensuring better sensitivity toward the minority (fraud) class.

Key Model Details:

- Features used: telecommuting, has_company_logo, has_questions, employment_type, required_experience, required_education, industry, function
- Target: fraudulent (binary: 0 = Real, 1 = Fake)
- Model used: Random Forest Classifier
- Train-test split: 70-30

Model Evaluation Metrics:

Metric	Real Jobs (0)	Fake Jobs (1)
Precision	0.97	0.76
Recall	0.98	0.60
F1-Score	0.98	0.67

- Overall accuracy: 96%
- Macro-average F1-score: 0.82
- Confusion matrix shows 37 fake jobs were correctly identified (true positives), while 25 were missed (false negatives).

Insights:

- The model performs very well on real jobs, and reasonably well on fake jobs, despite class imbalance.
- The use of class_weight='balanced' helped improve recall on the minority class, which is essential in fraud detection.
- The F1-score of 0.67 for fraud cases suggests a good balance between catching frauds and avoiding false positives.

Feature Importance Analysis:

The most influential features for detecting fake job postings were:

1. Function – Type of role posted (e.g., Sales, Admin)
2. Industry – The field of the job (e.g., IT, Services)
3. Required Education – Minimum qualifications listed
4. Required Experience – Experience level expected
5. Has Company Logo – Whether a company logo is present

These insights align with your earlier exploratory findings. Scammers often post fake jobs with vague functions, uncommon industries, or minimal details like missing logos.

Pivot table (R studio):

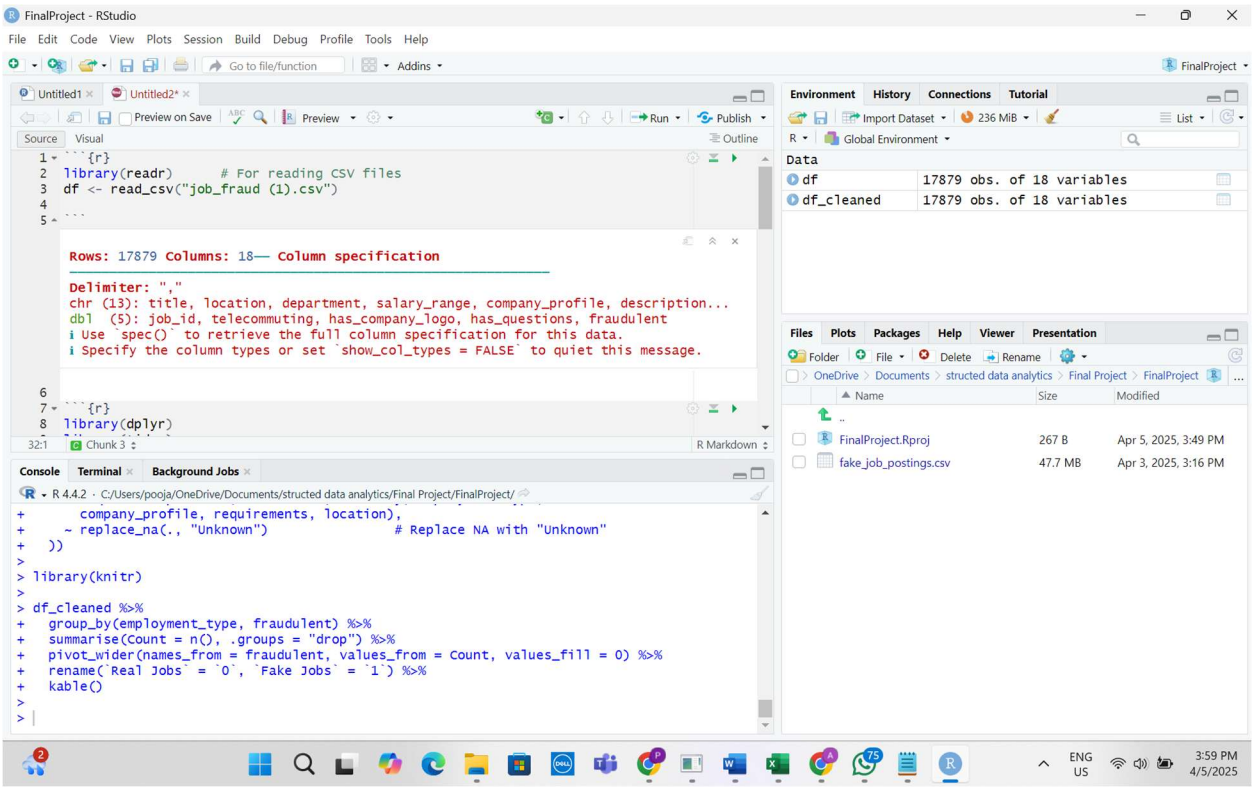


Figure 30: Read Dataset in R

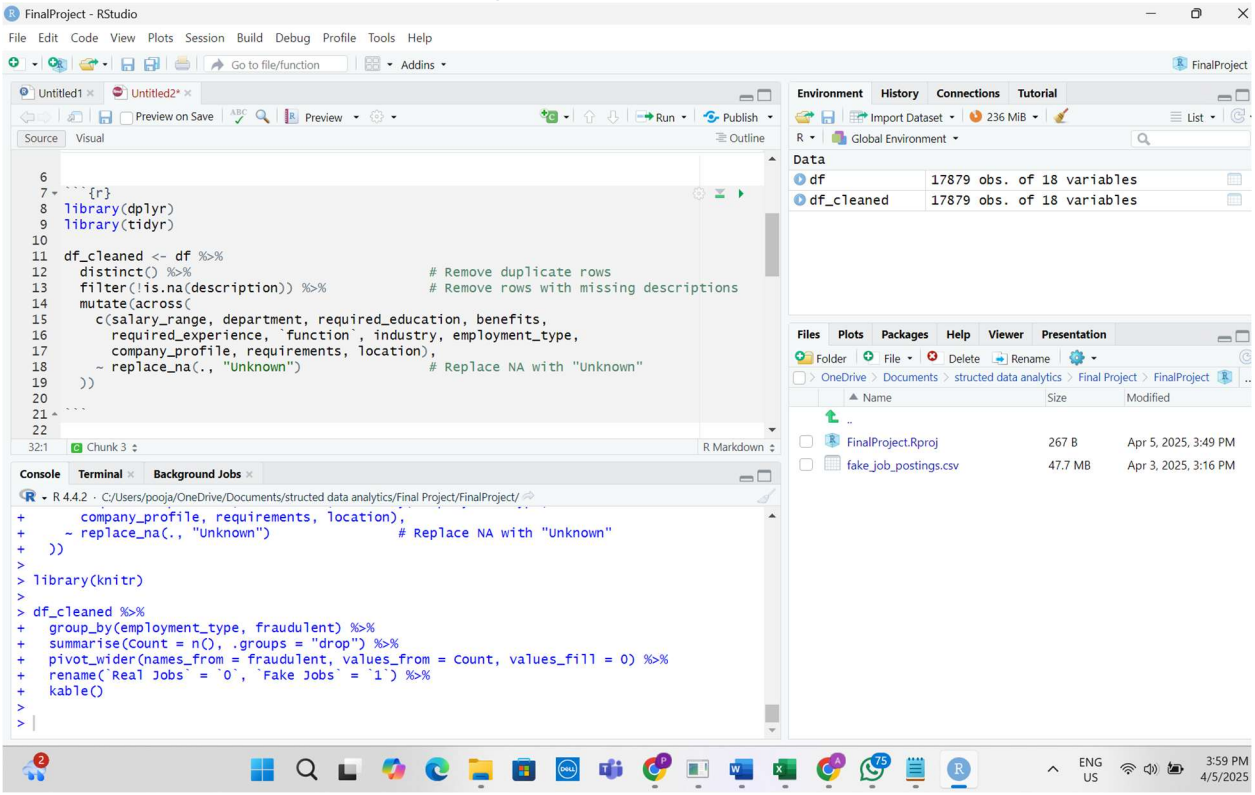


Figure 31: Data Description

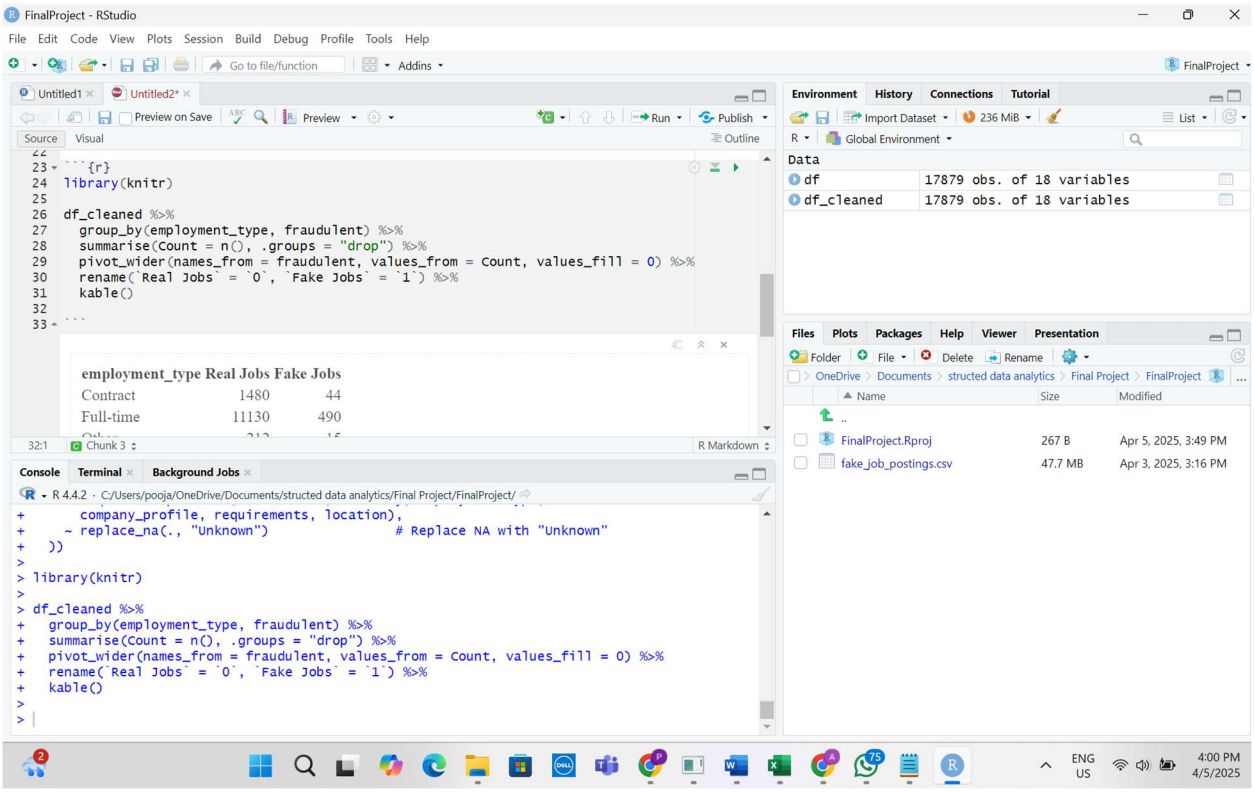


Figure 32: Pivot Table for real and Fake Jobs

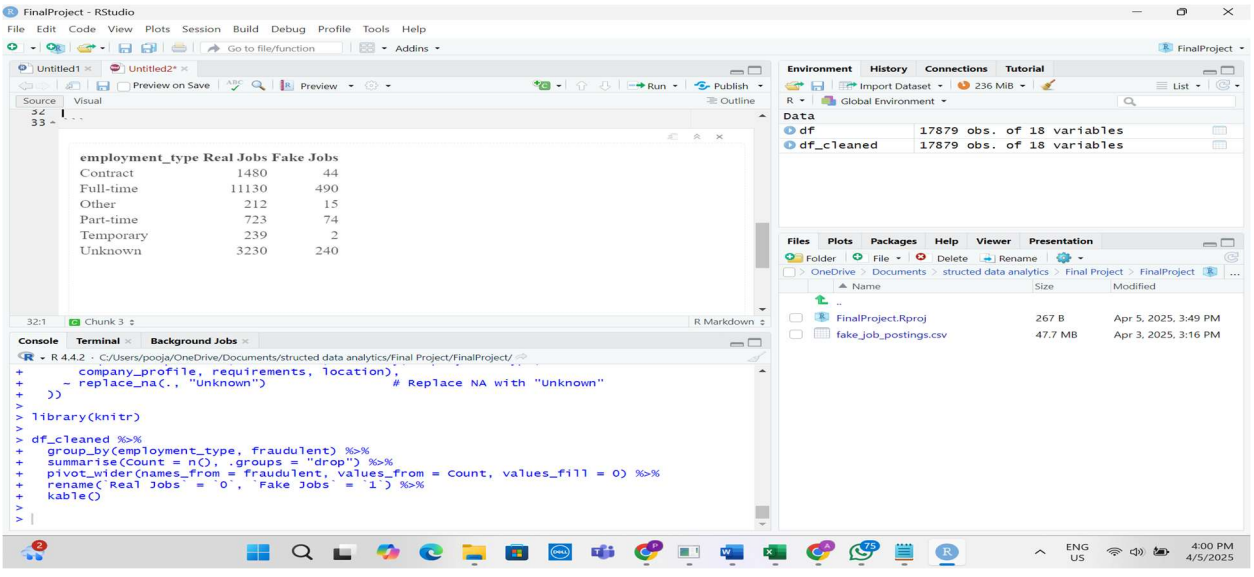


Figure 33: Results for Pivot Table of real and Fake Jobs

Real vs Fake Job Postings by Employment Type

The pivot table reveals important trends in how fraud is distributed across different employment types:

- Full-time jobs dominate the dataset with over 11,000 real listings and 490 fake ones. While this is the most common category, the fraud rate is relatively low (~4%), suggesting these listings are mostly legitimate.
- Part-time jobs show a notably higher fraud rate (~9%), indicating scammers often target people looking for flexible or short-term work.
- "Unknown" employment type also shows a significant fraud risk — with 240 out of 3,470 postings (around 7%) being fake. This suggests that missing or vague employment information may be a red flag.
- Contract and temporary roles have the lowest number of fake listings, indicating these are less frequently targeted by scammers in this dataset.

Conclusion:

The analysis confirms that employment type is a useful feature for identifying fraud. Listings with missing or vague types ("Unknown") and part-time roles show higher fraud rates and should be treated with caution in any detection model.

2)

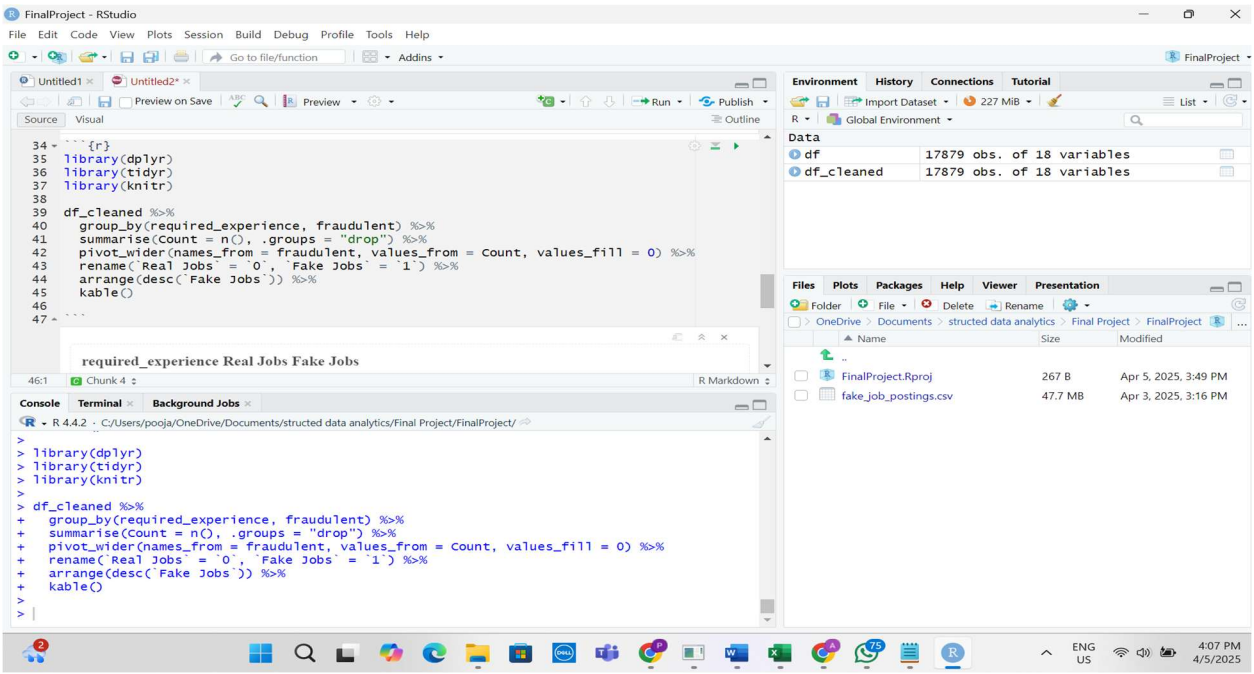


Figure 34: Pivot Table Based on Experience Required

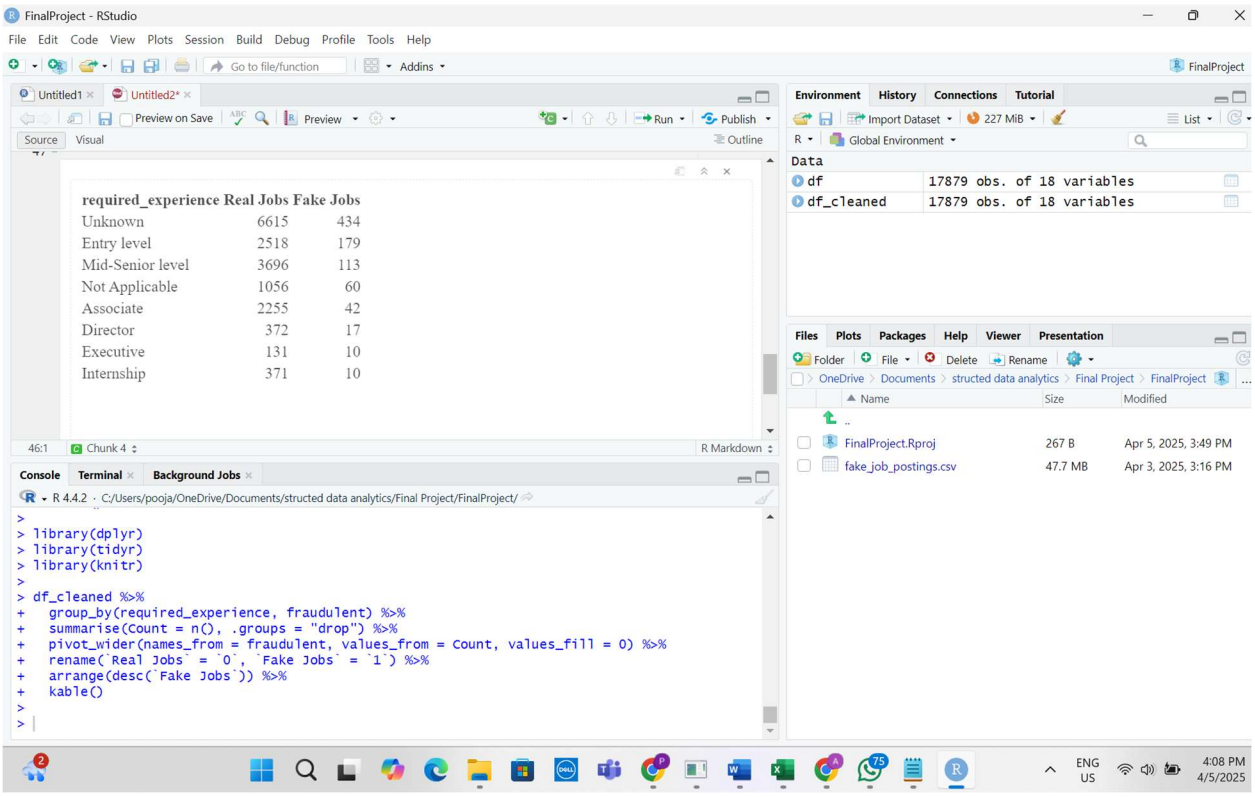


Figure 35: Pivot Table Based on Experience Required

Real vs Fake Jobs by Required Experience

Analyzing the job postings based on the level of experience required reveals significant differences in fraud distribution:

Key Observations:

- The “Unknown" experience level has the highest number of fake postings (434). This suggests that scammers frequently omit required experience to make the job listing appear accessible to all and reduce suspicion.
- Entry-level positions show the second-highest number of fake listings (179), which could be because scammers target people new to the job market who may be more vulnerable or less cautious.
- Mid-Senior level and Associate roles also show some fake listings, but at lower volumes, suggesting that scams are less common at more experienced levels.
- Positions like Director, Executive, and Internship have very few fake postings. This could be due to the smaller overall number of these listings and the increased scrutiny or specificity typically required for these roles.

Conclusion:

The data shows that job listings with ambiguous or missing experience requirements are more likely to be fraudulent. Entry-level roles are also frequently targeted, likely due to the broader applicant pool and higher response rates. This highlights required_experience as a valuable feature in fraud detection models.

3)

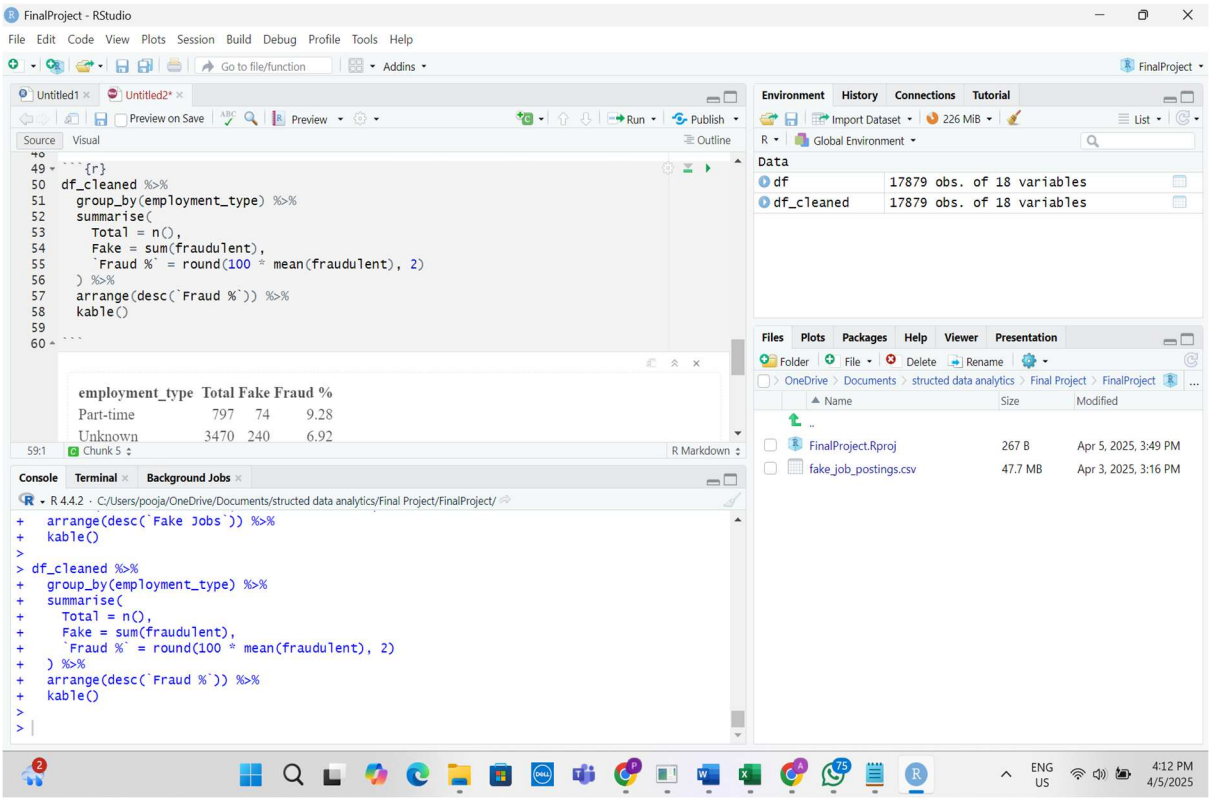


Figure 36: Pivot Table Based on Employment Type

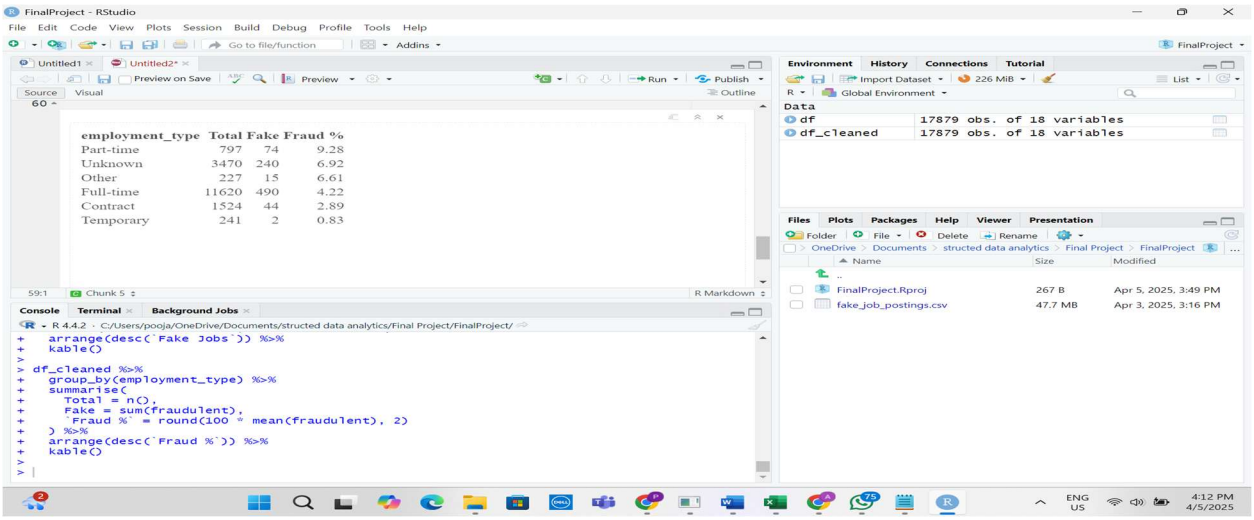


Figure 37: Results based on Pivot Table Based on Employment Type

Fraud Rate (%) by Employment Type

This pivot table highlights the percentage of job postings that are fraudulent across each employment type. Unlike raw counts, this view reveals how likely a job is to be fake depending on how it's described — and the results are very telling:

Key Findings:

- Part-time jobs have the highest fraud rate at 9.28%, meaning nearly 1 in 10 part-time job postings are fake. This strongly suggests that scammers exploit the flexibility and popularity of part-time roles to attract more applicants.
- Unknown employment types also show a high fraud percentage (6.92%), reinforcing earlier insights that missing or vague job details are often associated with fake listings.
- Even though full-time jobs make up a majority of listings, the fraud rate is much lower at 4.2%. This shows that full-time jobs are generally more trustworthy — though scammers do still try to mimic them.
- Temporary and contract jobs have the lowest fraud rates, at 0.83% and 2.89% respectively. These roles might be less targeted because they usually involve formal hiring structures.

Conclusion:

This analysis makes it clear that fraud isn't just about volume — it's about proportion. While full-time jobs dominate the dataset, part-time and vague job types carry the most fraud risk per listing. These insights are crucial for building smarter fraud detection models that weigh fraud likelihood, not just counts.

EDA:

Dashboard in R or interactive plots:

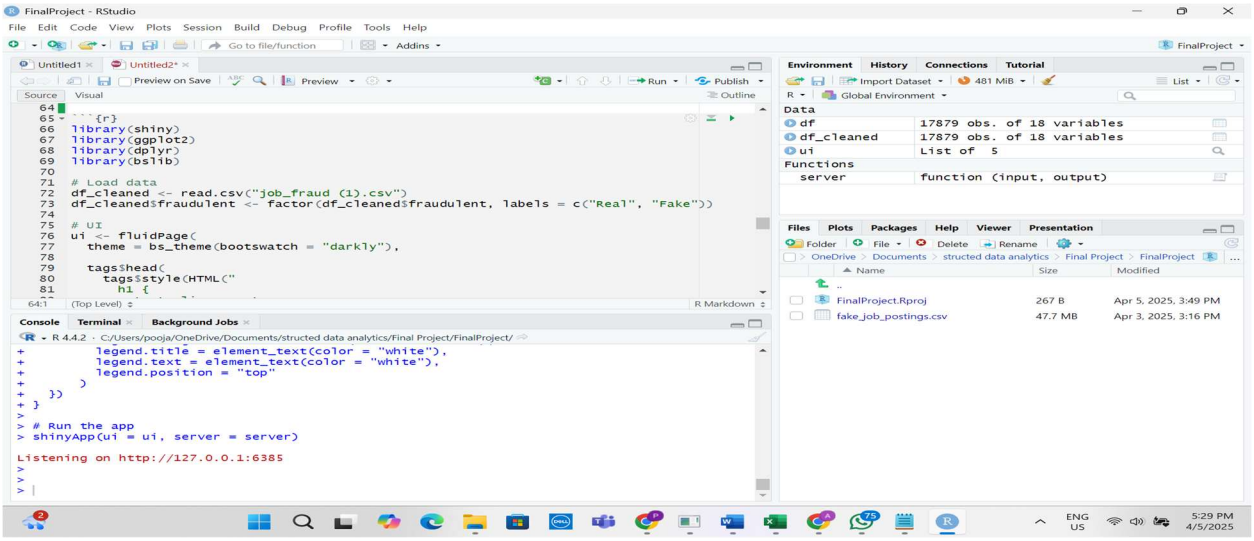


Figure 38: Dashboard Code

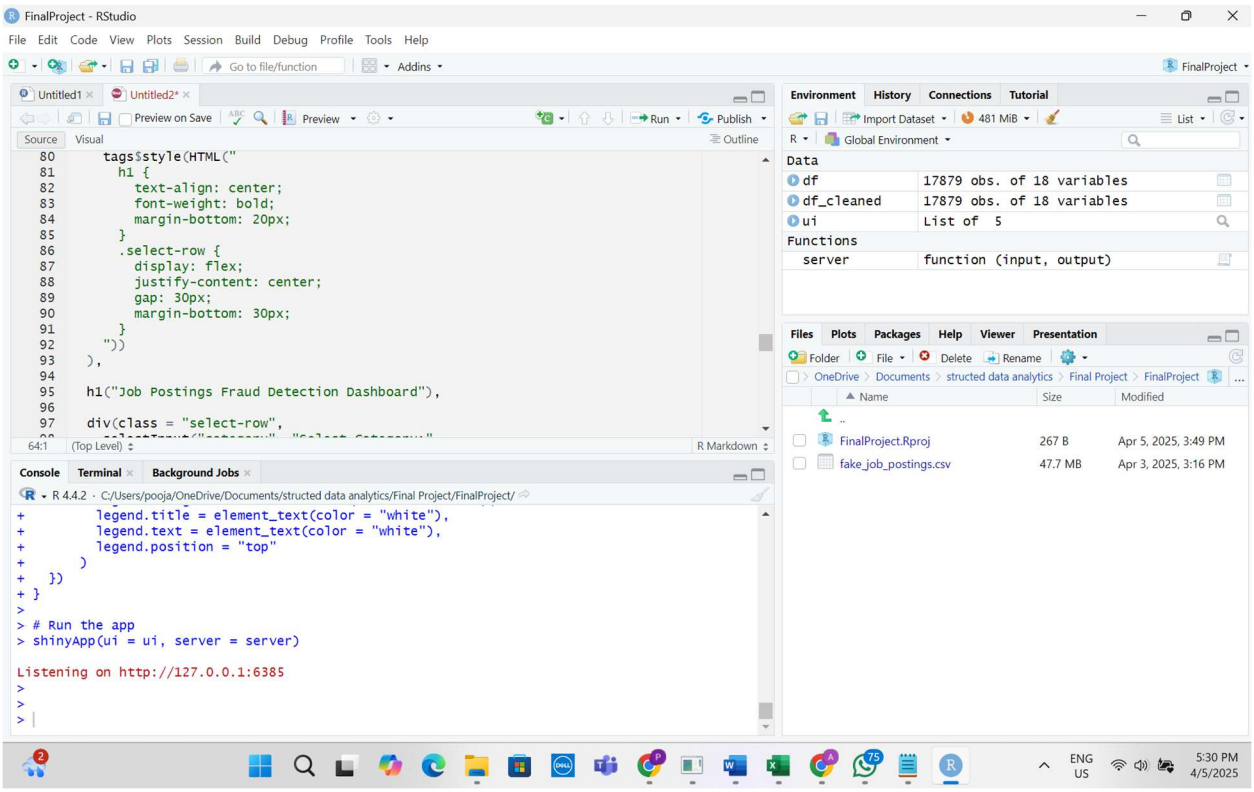


Figure 39: Dashboard Code

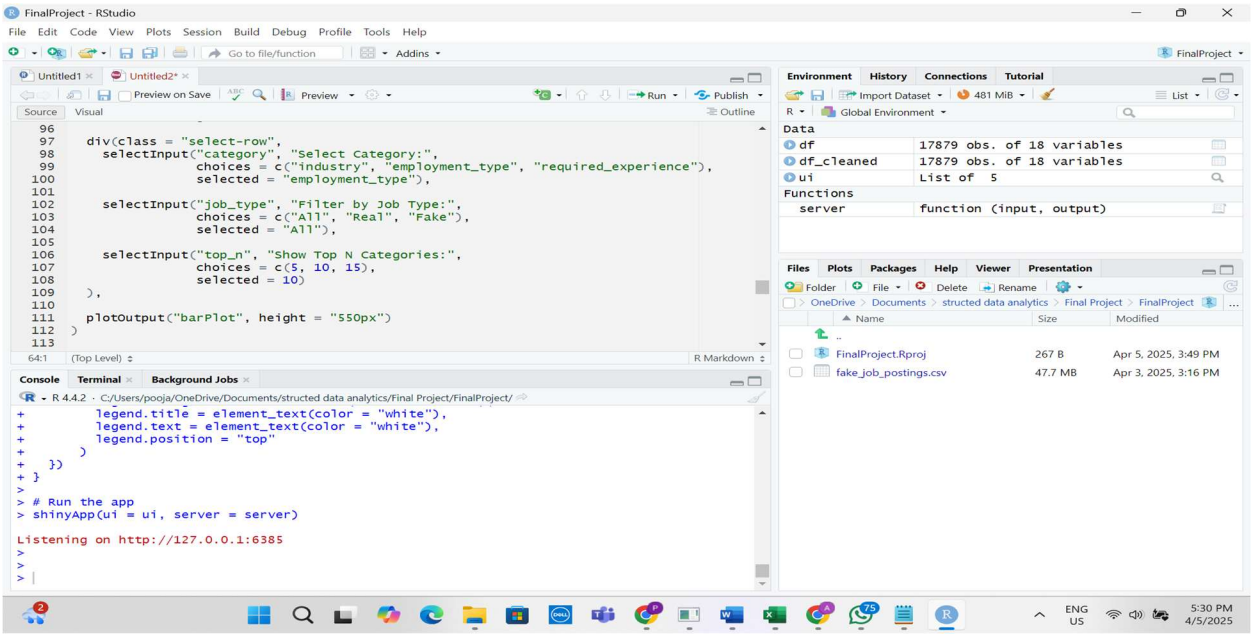


Figure 40: Dashboard Code

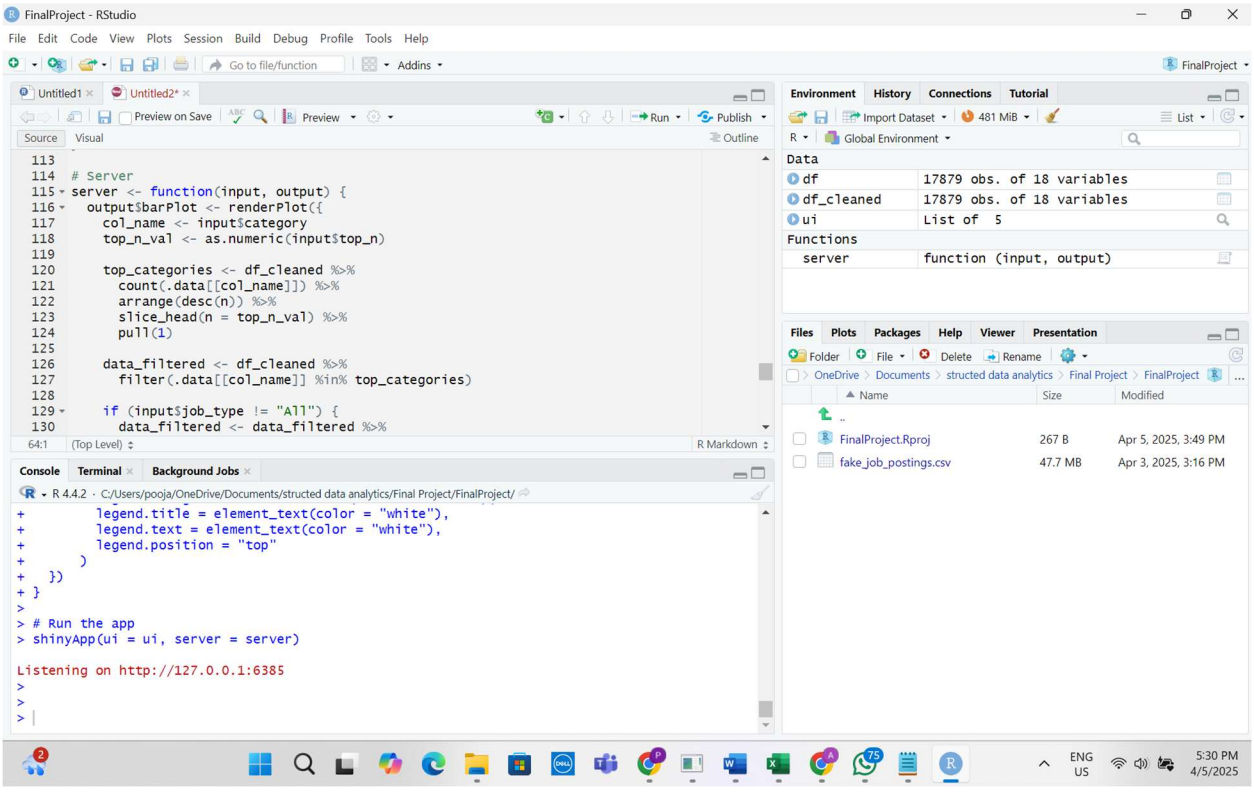


Figure 41: Dashboard Code

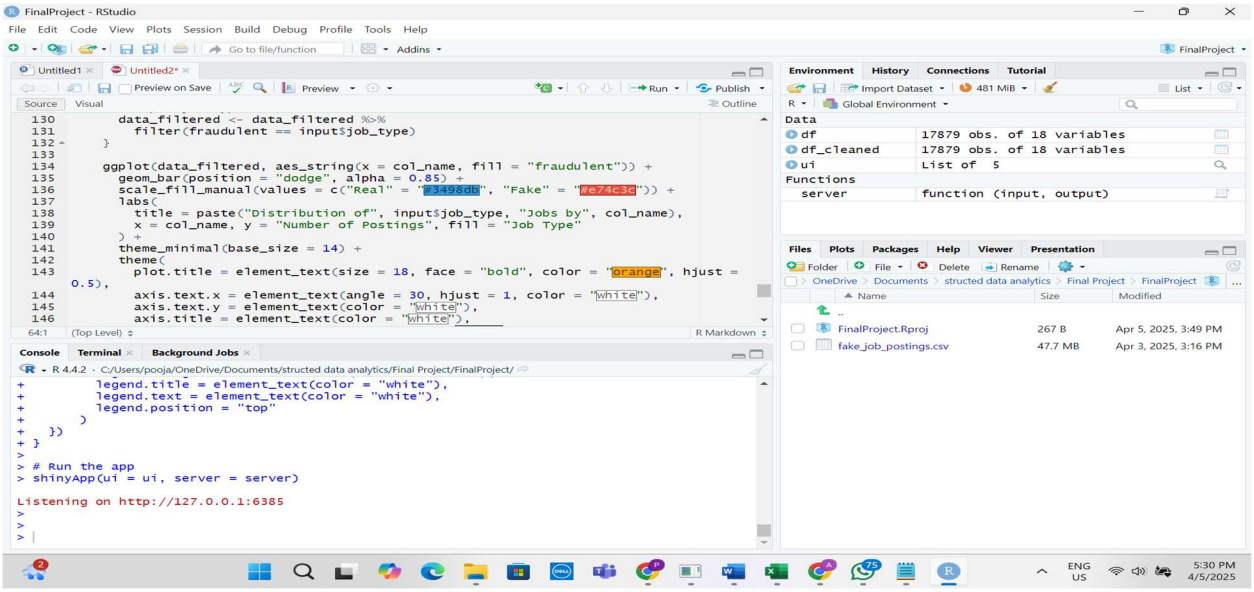


Figure 42: Dashboard Code

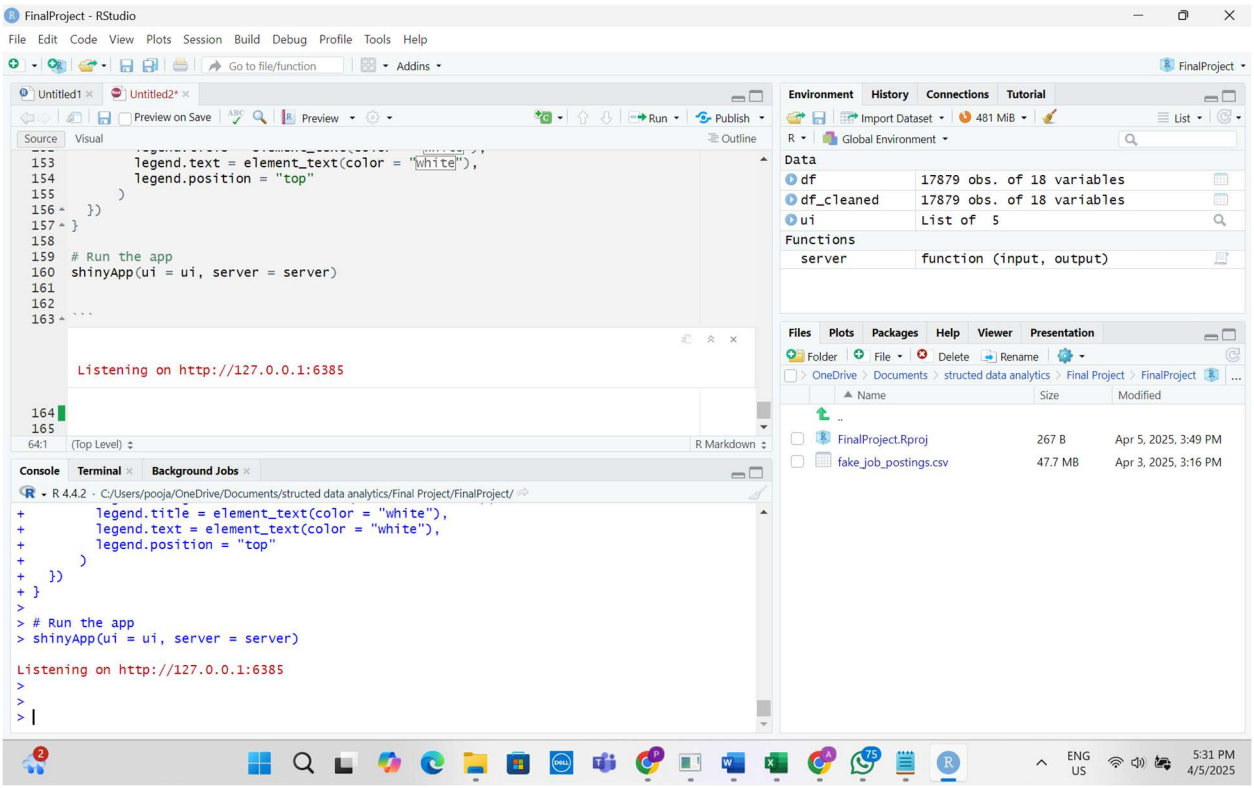


Figure 43: Dashboard Code

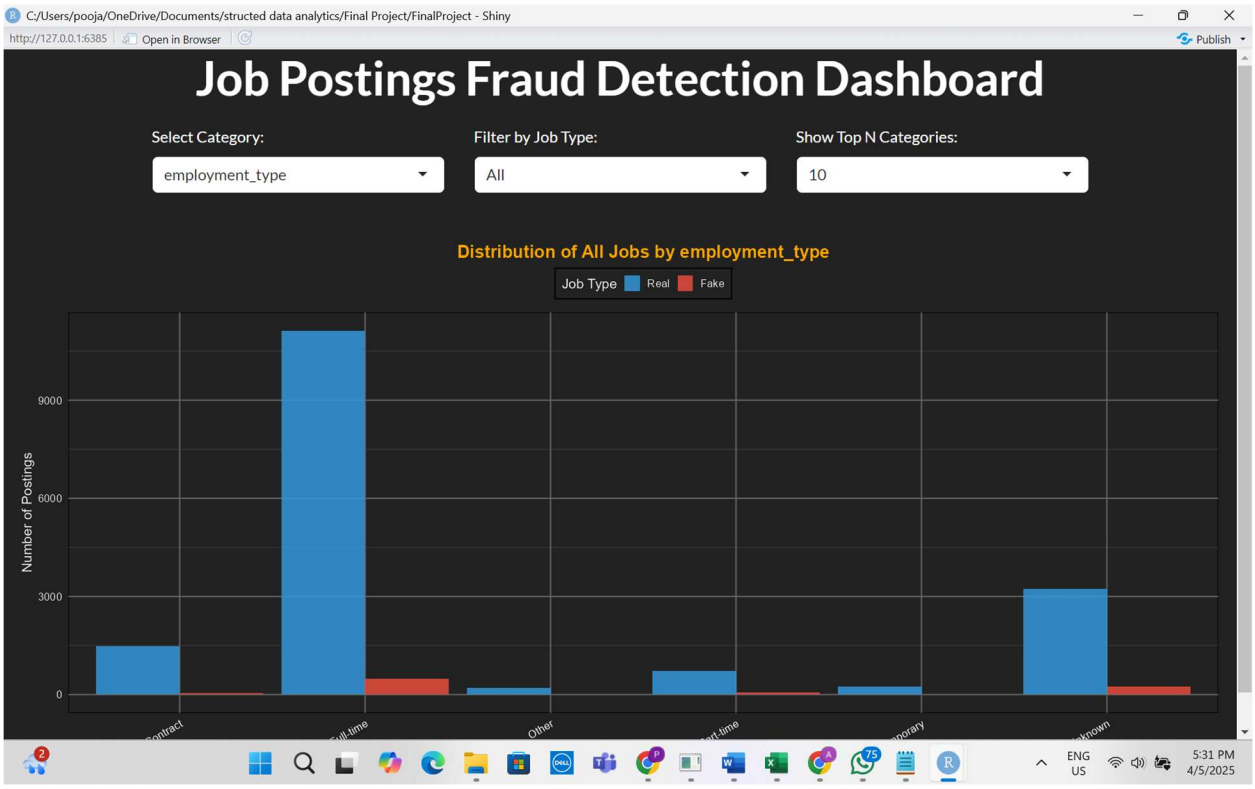


Figure 44: Dashboard

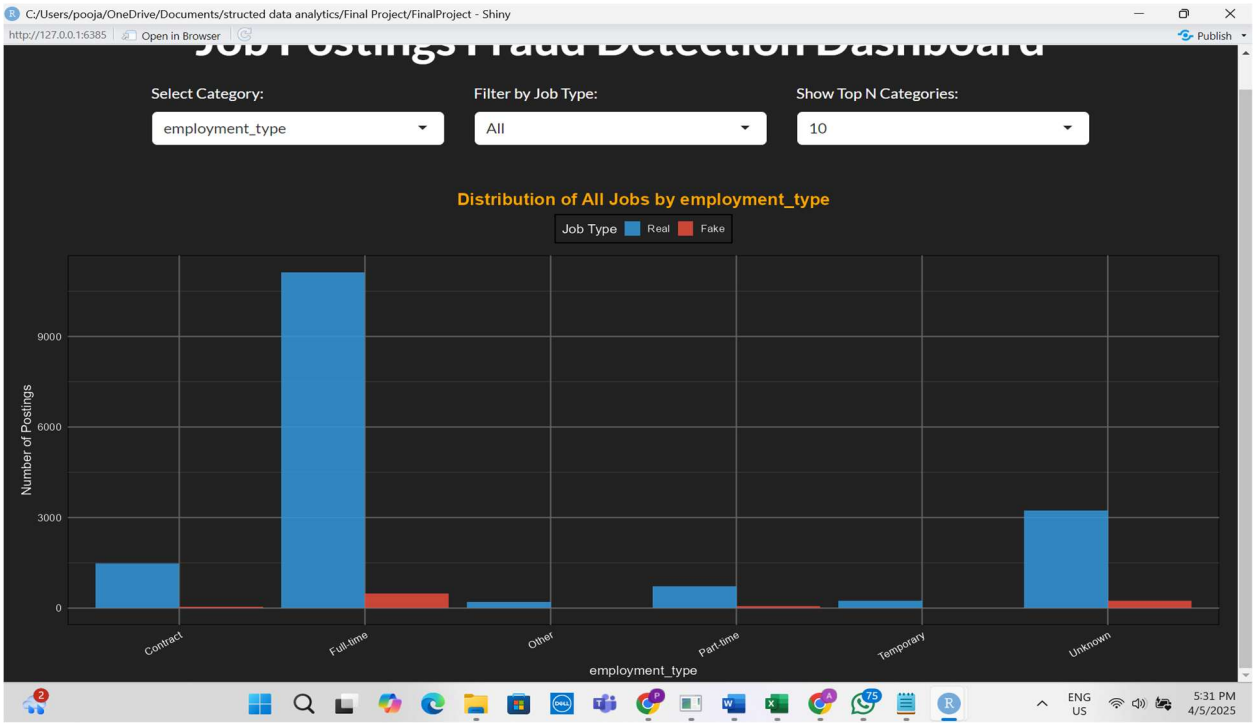


Figure 45: Dashboard

Interactive Dashboard Overview

This interactive dashboard was developed using R Shiny, based on a cleaned version of the *Fake Job Postings Dataset*. The data used in this dashboard was preprocessed to:

- Remove duplicate entries
- Drop null values in critical columns (like description)
- Impute or replace categorical nulls with "Unknown"
- Exclude irrelevant or redundant records to improve visualization clarity

The plot shown displays the distribution of real and fake job postings across various employment_type categories using a bar chart, with an option to filter by job type and category count.

Insights from the Employment Type Breakdown

- Full-Time jobs dominate the dataset, accounting for the largest share of both real and fake job postings.
    - While most full-time postings are real, a notable chunk of fake jobs also fall into this category, likely as scammers try to mimic professional job formats.
- The "Unknown" employment type shows a visible presence of fake postings.
    - This aligns with earlier findings from the ML model — vague or missing employment information is a red flag for fraud detection.
- Categories like Contract, Part-time, and Temporary are predominantly real, with very minimal fraud.
    - This may be due to their more specialized nature or clearer expectations, which scammers tend to avoid.
- The "Other" category contains very few listings overall but still includes some fakes — this might indicate inconsistent or non-standard job labeling, which again leans toward suspicious activity.

How This Dashboard Helps

- Filters like "Job Type" and "Top N Categories" enable users to interactively explore risk patterns by employment type, industry, and more.
- This can help HR professionals or platform moderators quickly spot suspicious categories, improving real-time fraud detection efforts.

Final Conclusion:
Fake Job Postings Analysis

Summary of Key Findings

After performing a comprehensive analysis on the Fake Job Postings dataset from Kaggle, several patterns and risk factors that distinguish **fraudulent job postings** from **real ones** were identified. This investigation not only answered the core research questions but also demonstrated the effective use of Python and R to derive actionable insights.

1. What factors differentiate fake job postings from real ones?

- **Employment Type:** Fraudulent postings are disproportionately labeled as **Full-time**, likely to increase legitimacy. Real jobs show diversity across part-time, contract, and temporary categories.
- **Required Experience:** Fake jobs often avoid highly experienced roles. They tend to cluster around **Mid-Senior**, **Associate**, and especially **Unknown** experience levels to appeal to a broader group.
- **Industry & Function:** Fake listings are more common in **Information Technology, Internet, and Services**—industries with high online engagement. Rare job functions were found to be more prone to fraud.
- **Presence of Company Logo & Questions:** Real postings more often include logos, while frauds may include **screening questions** to appear legitimate.
- **Description Length & Language:** Fake job postings had a distinct word usage (e.g., "customer", "assist", "maintain") versus real ones ("team", "client", "experience"). Unusual length (too short or too long) also raised red flags.

2. Are there common attributes among fraudulent job listings?

Yes. Fraudulent job postings tend to share:

- **Missing or vague information** (like 'Unknown' in required experience or employment type)
- **Excessively generic or task-oriented language**
- **Absence of a company logo**
- **Unusual function labels** or rare job categories
- **Unrealistic or empty salary ranges**

- **Disproportionate frequency in certain industries and job types**

These attributes were quantified through EDA, violin plots, anomaly detection, and feature importance models.

3. Can data analysis techniques effectively detect fake job postings?

Absolutely. Your analysis used:

- **Outlier & anomaly detection (IQR, frequency-based)** to identify suspicious entries.
- **Correlation heatmaps** to identify patterns between features and the target variable (fraudulent).
- **Random Forest Classifier** with a well-structured pipeline, addressing class imbalance using class_weight='balanced'.

The model achieved:

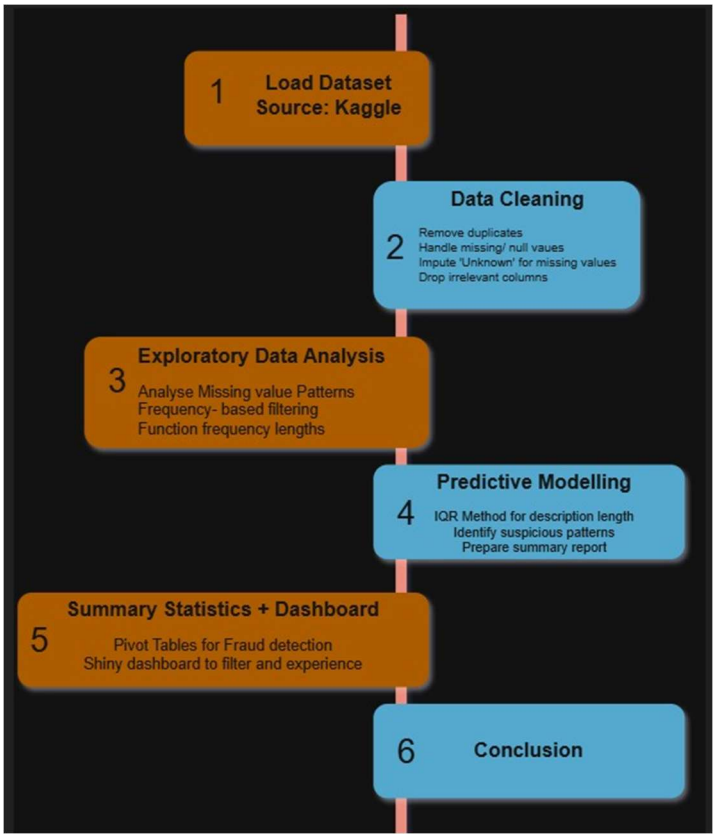
- **Accuracy:** 96%
- **F1 Score (fraud class):** 0.67 — A solid balance between false positives and false negatives.
- **Top Features:** function, industry, required_education, experience, and company_logo

These techniques together offer a **robust framework** to flag potentially fake postings before they reach job seekers.

Final Takeaways & Recommendations

- **Contextual awareness is critical:** Fraud patterns vary across industries and employment types. One-size-fits-all models are not effective.
- **Text analysis adds value:** Word clouds and NLP-based features can enhance fraud detection significantly.
- **Visualization improves clarity:** Through violin plots, pie charts, and heatmaps, patterns were made obvious and easier to interpret.
- **Interactive dashboards** (via R Shiny) empower stakeholders to explore data dynamically and take action faster.

Fraud Detection Pipeline for Job Postings



Why Python and R Were Used:

This project was completed using both **Python** and **R**, as required, but more importantly, their individual strengths were strategically used to handle different stages of the analysis efficiently.

Why Python Was Chosen for Data Cleaning, EDA, and Modeling

Python was used extensively for tasks that involved:

- **Data pre-processing**: Handling missing values, removing duplicates, and standardizing text entries was more intuitive using pandas.
- **Exploratory Data Analysis (EDA)**: Python's seaborn and matplotlib libraries offer superior customization and interactivity for:
    - Detecting missing values (bar plots)

    - Outlier detection (scatter plots, violin plots)
    - Distribution plots for categorical and numerical features
- **Anomaly Detection**: Python made it easier to implement IQR-based filtering and frequency-based anomaly detection using standard statistical methods.
- **Text Analysis**: The WordCloud package in Python enabled quick visualization of word usage differences between real and fake jobs.
- **Machine Learning**: Python was ideal for model building and evaluation:
    - scikit-learn's Random Forest classifier with built-in support for handling imbalanced data
    - Evaluation metrics like confusion matrix, F1-score, precision/recall for fraud class

Python was chosen here because of its strong data manipulation capabilities, flexible visualization libraries, and robust machine learning frameworks that support end-to-end model development and interpretation.

Why R Was Chosen for Summary Statistics, Pivot Tables, and Interactivity

R was used in sections where **summary statistics and user-facing interpretation** were more important, such as:

- **Pivot Tables** and **Data Tables**:
    - R's dplyr and tidyverse libraries made it easier to generate clean summaries, group-wise counts, and frequency tables.
    - Quickly aggregating categorical variables to observe patterns (e.g., experience vs fraud status)
- **Interactive Visualizations**:
    - R's **Shiny dashboard** was used to create an interactive interface where users could:
        - Filter jobs by industry or experience
        - Visualize fraud patterns dynamically
        - Make the output more presentation-ready for stakeholders
- **Descriptive Analysis**:
    - R was useful for quick summaries, contingency tables, and exploratory insights to support the modeling narrative.

R was chosen in these parts due to its strong capabilities for statistical summaries, ease of working with categorical data, and most importantly, its ability to build lightweight, **interactive dashboards** via Shiny without additional setup.

References

1. **Dataset Source:**
   - *Kaggle – Fake Job Postings Dataset.* https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction
2. **Python Libraries:**
   - *matplotlib: Visualization Library* https://matplotlib.org
   - *scikit-learn: Machine Learning in Python* https://scikit-learn.org
3. **R Packages and Tools:**
   - *Tidyverse – Data Science Framework* https://www.tidyverse.org
   - *Shiny – Web Application Framework for R* https://shiny.posit.co/r/getstarted/shiny-basics/lesson1/
4. **Anomaly Detection Concepts:**
   - *Wikipedia – Anomaly Detection* https://en.wikipedia.org/wiki/Anomaly_detection