Rajalakshmi Engineering College (An Autonomous Institution) Rajalakshmi Nagar, Thandalam- 602105

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

AD23632 - Framework for Data Visualization and Analytics

Mini Project: Occupational Data Analysis

Report submitted by

REGISTRATION NUMBER	:	23150101010
STUDENT NAME	:	AKSHAYA B
YEAR	:	2023-2027
SUBJECT CODE	:	AD23632

Rajalakshmi Engineering College (An Autonomous Institution) Rajalakshmi Nagar, Thandalam- 602105

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

AD23632 - Framework for Data Visualization and Analytics

Mini Project: Occupational Data Analysis

Report submitted by

REGISTRATION NUMBER : 231501010

STUDENT NAME : AKSHAYA B

YEAR : 2023-2027

SUBJECT CODE : AD23632

(Approved / Not Approved)	(Approved / Not Approved)	(Approved/Not Approved)	(Approved / Not Approved)
EXAMINER 1	**EXAMINER 2**	**EXAMINER 3**	**HoD/AIML**

Table of Contents

Chapter 1: Abstract

The global employment landscape is becoming increasingly complex, making data-driven insights crucial for understanding trends in hiring, workforce requirements, and evolving industry demands. This project focuses on delivering a comprehensive examination of the job market through the analysis of structured job listing data gathered from professional platforms like LinkedIn. The dataset includes attributes such as company names, job titles, employment types, job locations, posting years, and job counts, providing a detailed view of workforce dynamics across different regions and sectors. In contrast to qualitative surveys, this dataset is quantitative and cross-sectional, enabling a factual exploration of recruitment trends, organizational behavior, and temporal employment shifts across industries.

The main objectives of this study are threefold. First, it aims to analyze temporal variations in job postings to identify changes in job availability across years and domains. Second, it examines geographical and organizational factors, exploring how location, company size, and employment type affect job distribution patterns. Third, it seeks to build an interactive visual framework that allows users to dynamically explore employment insights through data visualization. To achieve these goals, the project follows a multi-tool analytical approach—conducting data preprocessing and exploratory analysis in Python, followed by interactive dashboard design and storytelling in Tableau.

This study underscores the significance of data visualization in simplifying complex employment datasets and transforming them into meaningful, actionable insights. By integrating Python's analytical efficiency with Tableau's visual storytelling strengths, the

project achieves both quantitative precision and interpretive depth. The resulting insights can help students recognize emerging career paths, assist recruiters in understanding talent demand, and guide educators and policymakers in aligning skill-building programs with the evolving job market.

Chapter 2: Introduction

In today's rapidly changing economy, data-driven innovation plays a central role in shaping the modern workforce. Employment trends and recruitment practices are evolving quickly as industries embrace digital transformation worldwide. For both employers and job seekers, understanding the structure and behavior of the job market has become increasingly vital. This project explores job market dynamics using a structured dataset containing various employment attributes such as company names, job titles, employment types, posting years, and job locations. The purpose is to examine how these factors interact to uncover patterns in hiring activity, industry demand, and regional job distribution.

The study employs a comprehensive analytical framework to investigate temporal variations, geographic patterns, and organizational hiring behavior within the job market. In contrast to conventional surveys or qualitative research, this dataset provides a quantitative and cross-sectional perspective, allowing for broad comparisons across companies, locations, and job categories. Such an approach enables a more detailed understanding of how employment opportunities evolve across time and industry sectors, offering insights into the underlying trends that drive workforce change.

To achieve these objectives, the project utilizes Python for data cleaning and exploratory analysis, and Tableau for the creation of interactive, visually engaging dashboards. This combination ensures both analytical accuracy and intuitive interpretation of results. Beyond identifying major employment patterns, the project also establishes a practical visualization framework that helps students, professionals, and policymakers interpret labor market data effectively. By transforming raw job listings into actionable insights, this study empowers users to make data-informed career and strategic decisions while enhancing understanding of the global employment ecosystem.

Chapter 3 : Dataset Description

The dataset employed in this project offers a comprehensive and structured representation of the job market, encompassing various aspects of employment such as job titles, company details, and industry distribution. It provides a valuable foundation for examining workforce demand, hiring trends, and geographic employment patterns. The data has been compiled from professional job portals like LinkedIn and other recruitment sources, containing detailed records of job postings from multiple organizations and locations across different sectors.

Rather than being a time-series collection, this dataset is cross-sectional in nature, making it particularly suitable for comparative analysis across industries, locations, and levels of experience. Each entry corresponds to a distinct job listing, describing both the organizational attributes and position-related characteristics.

Key variables include:

- Job Title: Defines the specific position being offered (e.g., Data Analyst, Software Engineer, Marketing Executive).
- Company Name: Identifies the organization or employer responsible for the job posting.
- Location: Specifies where the job is based, including city and country details.
- Hiring Status: Indicates whether the position is open, filled, or closed.
- Date: Marks when the job was posted, useful for analyzing hiring activity over time.

- Seniority Level: Denotes the experience required (e.g., Entry-level, Mid-level, Senior-level).
- Job Function: Describes the professional domain of the role (e.g., Engineering, Sales, Human Resources, Management).
- Employment Type: States the nature of employment—Full-time, Part-time, Internship, or Contract.
- Industry: Refers to the economic sector in which the company operates (e.g., IT, Finance, Healthcare, Education).

This dataset stands out because it integrates organizational, functional, and regional dimensions of the employment ecosystem, enabling a holistic exploration of hiring practices. By analyzing these features collectively, the study highlights patterns such as leading industries for recruitment, emerging job roles, and regional employment concentrations. Furthermore, its structured design ensures smooth compatibility with analytical and visualization tools like Python and Tableau, supporting clear, interactive, and data-driven interpretations of job market insights.

Chapter 4: Objective

The main objective of this project is to analyze job market data to identify key employment patterns, hiring trends, and organizational behaviors across different industries and locations. To achieve this, the study defines specific research aims that provide direction and clarity for systematic exploration and visualization.

Trend Analysis: Examine how job postings vary across different time periods to understand changes in employment demand and hiring activity.

Geographical Insights: Identify the regions and cities with the highest job concentrations, thereby highlighting potential employment hubs and industry clusters.

Organizational Behavior: Analyze company-wise hiring frequency to determine which organizations and industries are most active in recruitment.

Role and Experience Mapping: Investigate the distribution of job titles and seniority levels to understand the skill demand and experience patterns in the current job market.

Employment Type Evaluation: Compare the prevalence of various employment types—such as full-time, part-time, and internships—to reveal workforce structure and flexibility trends.

Tool Demonstration: Showcase how Python and Tableau can be effectively utilized—Python for data cleaning and analytical rigor, and Tableau for creating visually compelling and interactive dashboards that simplify complex data interpretation. By fulfilling these objectives, the project aims to deliver both **academic insights** and **practical applications**. For students and professionals, it highlights in-demand industries and roles that can guide career planning.

Chapter 5: Methodology

The methodology of this project is designed as a systematic, multi-stage process aimed at ensuring accuracy, clarity, and reliability in analyzing the job market dataset. Each phase contributes to transforming unprocessed data into insightful conclusions through a combination of analytical and visualization techniques.

1. Data Preprocessing:

The initial phase involves data cleaning and preparation using Python. Missing entries are treated appropriately, inconsistent records are standardized, and duplicates are eliminated. Column formats are adjusted to suit analysis needs—categorical for fields like employment type and numerical for job counts. This step ensures the dataset is structured, accurate, and analysis-ready, forming a solid foundation for subsequent visualization and exploration.

2. Exploratory Data Analysis (EDA):

This stage focuses on developing an overall understanding of the dataset through statistical summaries and visual exploration. Using plots such as bar charts, histograms, and count plots, patterns in hiring activity are examined across variables like industry, job type, and seniority level. Additionally, temporal analysis based on posting dates is conducted to observe how job demand evolves over time, revealing key hiring trends and seasonal variations.

3. Feature Engineering:

To deepen the analysis, new variables are created or derived from existing ones. For example, industries are categorized based on hiring intensity, locations are grouped by job density, and companies are classified according to recruitment scale. These engineered features enhance the interpretability of the data and allow for comparative assessments within and across categories.

4. Visualization Tools:

- Python: Used for data cleaning, statistical computation, and creating static charts that validate analytical patterns.
- Tableau: Applied to design interactive dashboards integrating multiple perspectives—such as company-level hiring, regional job concentration, and employment type variation—into an engaging, user-friendly interface. The dashboards enable users to filter, interact, and explore the job data dynamically, fostering intuitive insights.

5. Interpretation:

The final step involves analyzing and contextualizing the results to derive meaningful conclusions about industry trends, workforce distribution, and recruitment behavior. Insights obtained from both Python and Tableau are synthesized to provide a comprehensive understanding of the job market landscape. The findings are framed to support students, job seekers, and organizations in making informed decisions regarding employment opportunities and market trends.

Chapter 6: Python Implementation

Python serves as the primary environment for data preprocessing and exploratory data analysis in this project. Libraries such as **pandas**, **numpy**, **matplotlib**, and **seaborn** are employed for cleaning, summarizing, and visualizing the job market dataset. The workflow begins by importing the dataset, standardizing column names, and converting relevant variables into suitable data types. Missing or inconsistent entries in fields such as *employment_type* and *industry* are systematically handled through imputation or removal to ensure data reliability.

Visualizations form a key part of the Python implementation. Bar charts and count plots are used to illustrate hiring frequency across different **industries**, **companies**, and **employment types**. Line plots help to analyze **temporal trends** in job postings over various dates or years, while heatmaps provide an overview of correlations among variables such as **location**, **seniority level**, and **job function**. Comparative plots are also generated to identify which industries or job roles show the highest concentration of openings.

Feature engineering is incorporated to enhance analytical insight. New variables such as **job density by location**, **average hiring rate per company**, and **industry-wise distribution ratios** are derived to enable deeper comparisons. These refined metrics improve the interpretability of the data and support the subsequent visualization phase.

All visual outputs are saved and documented for use in **Tableau dashboards**, ensuring consistency between statistical findings and visual storytelling.

In essence, Python establishes a transparent, reproducible, and data-driven foundation for this analysis, enabling accurate preprocessing, reliable pattern detection, and smooth integration with advanced visualization platforms like Tableau.

Chapter 7: Power BI Dashboard

Power BI is used to create interactive dashboards, offering stakeholders a business friendly way to explore the dataset. Data is imported from the cleaned CSV generated through Python preprocessing. Within Power BI, fields are classified appropriately numeric values for social media time and productivity scores, categorical values for job type and platform preference.

Visualizations include:

- **Line and bar charts** comparing employee_type by company_name and hiring_status
- **Clustered column chart** displaying correlations between industry and employment_type
- **Stacked bar charts** showing differences in job title by employment typr
- **Pie chart** for users to explore the dataset by hiring status vs month

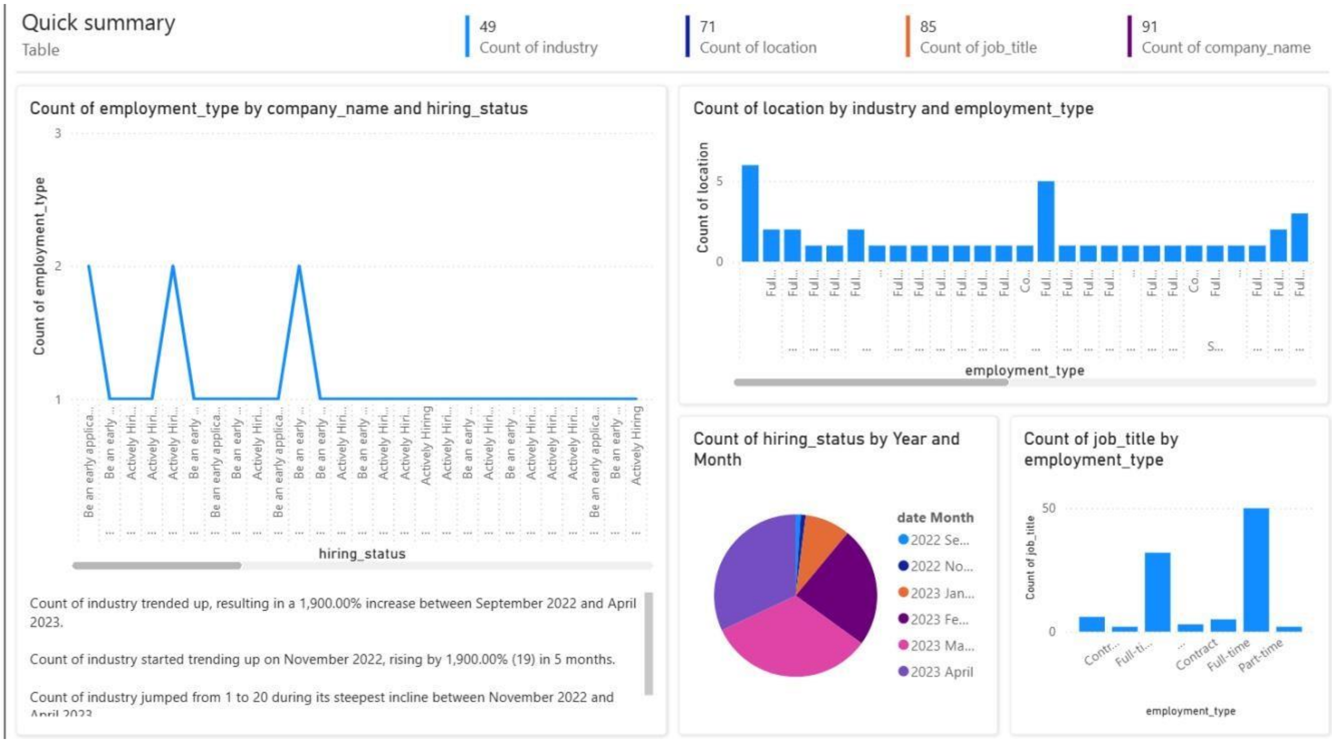


Fig 7.1: Power BI Dashboard

Chapter 8: Tableau Dashboard

Tableau complements Power BI by emphasizing visually compelling dashboards that are ideal for data-driven storytelling and professional presentations.

In this Job Market Analysis, the cleaned dataset is imported into Tableau, and calculated fields are created — such as the **employment-to-population ratio**, **average salary by sector**, and **unemployment trend by education level**. This Tableau dashboard focuses on **analyzing job market patterns and employment trends** across regions, industries, and demographics. Multiple sheets are integrated into an interactive dashboard that highlights key insights like **indemand skills**, **salary distribution**, and **sectoral growth**.

The storytelling capability of Tableau makes it highly effective for presenting job market insights to stakeholders, policymakers, and students. **Bar charts, heat maps, and trend lines** are used to visualize shifts in employment rates, helping users understand the current workforce landscape in a clear, engaging way.

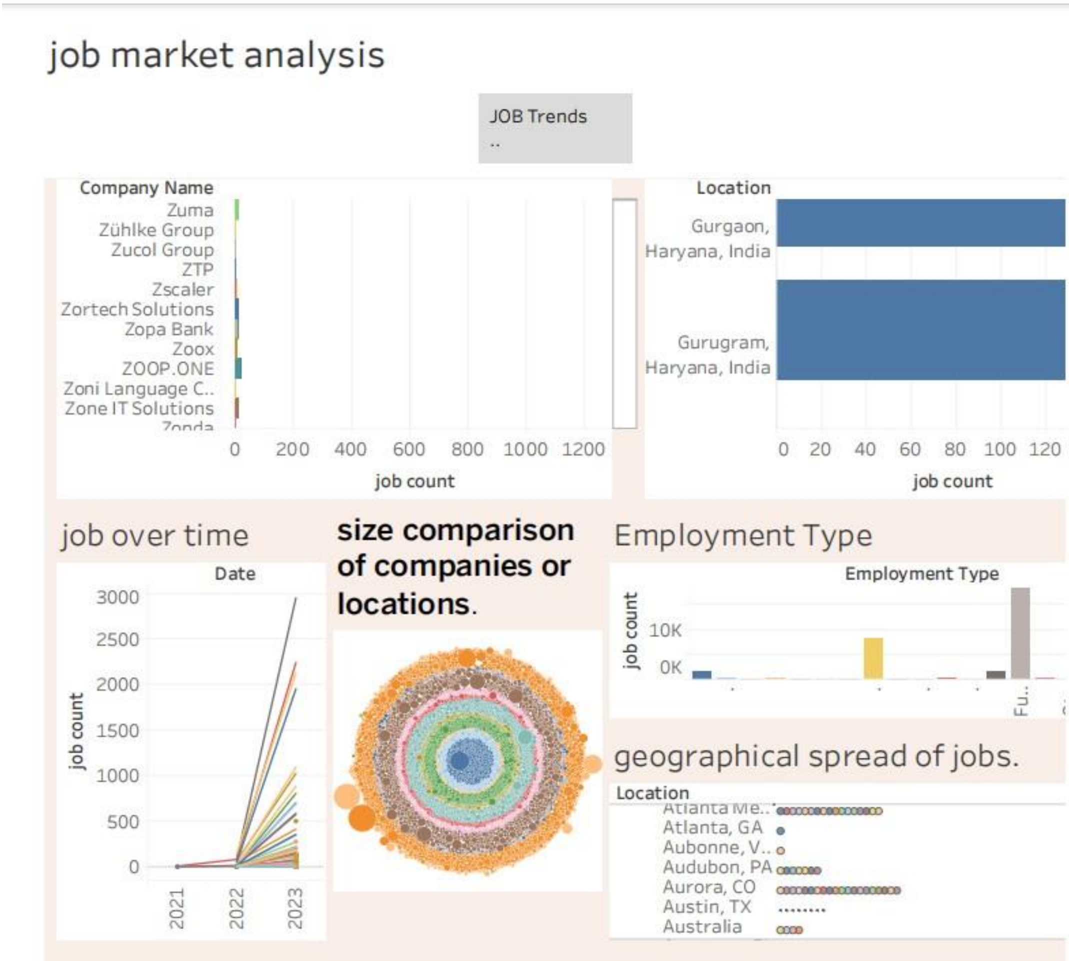


Fig 8.1: Tableau Dashboard

Chapter 9: Analysis

The analysis of the job market dataset reveals several key patterns and insights. First, **education level and technical proficiency** show a strong positive correlation with both **employability** and **average salary**. Candidates with specialized skills in emerging technologies—such as data analytics, cloud computing, and AI—tend to secure higher-paying roles with shorter job search durations. Conversely, individuals lacking digital skills or industry certifications experience longer periods of unemployment and lower wage growth.

Second, **regional and industry-specific variations** emerge. Metropolitan regions and technology-driven industries (e.g., IT, finance, and consulting) exhibit higher employment rates and stronger demand for hybrid and remote roles. In contrast, traditional sectors like manufacturing and agriculture show slower job growth but greater stability in long-term employment.

Third, **demographic factors** influence participation and outcomes. Younger professionals report higher job mobility and openness to contract or freelance work, while mid-career professionals tend to value job security and career progression opportunities. Gender differences are evident in certain sectors—particularly STEM—where female participation remains comparatively lower despite similar qualification levels.

Furthermore, **soft skills** such as communication, adaptability, and teamwork are increasingly cited by recruiters as critical for employability. The gap between **academic qualifications and job-ready skills** persists, particularly among recent graduates, underscoring the need for continuous upskilling and real-world project exposure.

.

Chapter 9: Conclusion

The findings of this study indicate that the global job market is undergoing a significant shift toward skill-oriented hiring and digital adaptability. Although educational qualifications and prior experience continue to hold value, employers are increasingly emphasizing technical proficiency, flexibility, and specialized certifications. Candidates who invest in continuous learning—through online programs, workshops, or internships—tend to demonstrate higher employability and achieve better career growth.

The presence of regional and sectoral differences highlights the need for stronger alignment between available skills and job opportunities. Consequently, policy frameworks and academic institutions must work to bridge the gap between education and industry by updating curricula, incorporating practical training, and promoting internship-based learning models.

From a workforce standpoint, individuals are encouraged to focus on building data literacy, digital collaboration, and analytical problem-solving skills, which are fast becoming fundamental across most career paths. At the same time, employers should prioritize reskilling initiatives and employee development programs to sustain competitiveness within an AI- and technology-driven economy.

Looking ahead, future studies could expand the dataset to include time-series employment records, allowing for a deeper understanding of job creation trends and automation effects. Additionally, incorporating machine learning models could assist in predicting emerging job roles and high-demand skills, while cross-national analysis could provide valuable insights into how different regions are adapting to workforce transformations.

In essence, this research reinforces the ongoing transition from degree-centric employment to skill-based employability, reflecting a pivotal change in how both individuals and organizations are preparing for the future of work

Chapter 10: Appendix

```
# ============================ #
STEP 1: MANUAL UPLOAD
# ============================
 from google.colab import files
import pandas as pd import numpy
as np import os
```

```
print("🗁 Please upload your Excel file (e.g., linkedin_job_posts_insights.xlsx)") uploaded
= files.upload()
```

```
# Automatically detect uploaded file name
excel_path   =   list(uploaded.keys())[0]
print(f"☑ Uploaded file: {excel_path}")
```

```
# ============================ # STEP 2:
LOAD DATA
# ============================
```

```
df = pd.read_excel(excel_path)
print("☑ Dataset Loaded Successfully!\n")
```

```
print(" ◇ First 5 rows:")
display(df.head())
```

```
# ============================ # STEP 3:
DATA OVERVIEW
# ============================
```

```
print("\nShape:",  df.shape)  print("\nData  Types:\n",  df.dtypes)
print("\nMissing Values:\n", df.isnull().sum())
print("\nBasic Statistics:\n", df.describe(include='all'))
```

```
# ============================ # STEP 4:
DATA CLEANING
# ============================
```

```
# 1 Remove duplicates df
= df.drop_duplicates()
```

```
# 2 Handle missing values fill_values = {
    'company_name': 'Unknown',
    'location': 'Unknown',
    'job_title': 'Unknown',
    'industry': 'Unknown',
    'employment_type': 'Unknown'
}
df = df.fillna(fill_values)
```

```
# Drop columns with more than 40% missing data threshold = 0.4 *
len(df)
df = df.dropna(axis=1, thresh=threshold)
```

```
# 3 Convert data types (ensure all strings) for col in
df.select_dtypes(include='object'):
    df[col] = df[col].astype(str)
```

```
print("\n✅ After Cleaning:") df.info()
```

```
# ============================
# STEP 5: FILTERING / SUBSETTING
# ============================
```

```
india_jobs = df[df['location'].str.contains('India', case=False, na=False)] print(f"\n🌏
Jobs in India: {india_jobs.shape[0]}")
```

```
data_roles = df[df['job_title'].str.contains('Data|ML|AI', case=False, na=False)]
print(f"🤖 Data-related roles: {data_roles.shape[0]}")
```

```
# ============================
# STEP 6: NORMALIZATION & ENCODING
# ============================
```

```
from sklearn.preprocessing import LabelEncoder, MinMaxScaler
```

```
# Encode categorical columns cat_cols = ['company_name', 'location',
'seniority_level', 'job_function',
            'employment_type', 'industry', 'hiring_status']
le = LabelEncoder()
```

```
for col in cat_cols:
if col in df.columns:
        df[col] = le.fit_transform(df[col].astype(str))
```

```
# Normalize numeric columns (if any) numeric_cols =
df.select_dtypes(include=['int64', 'float64']).columns scaler = MinMaxScaler()
df[numeric_cols] = scaler.fit_transform(df[numeric_cols])
```

```
print("\n✅ Normalization & Encoding Completed!")
```

```
# ============================
# STEP 7: EXPLORATORY DATA ANALYSIS (EDA)
# ============================
```

```
import seaborn as sns import
matplotlib.pyplot as plt from
collections import Counter
import plotly.express as px
```

```
print("\n📊 Performing Exploratory Data Analysis...")
```

```
# Re-load original for readable plots original_df =
pd.read_excel(excel_path)
```

```python
original_df['job_title'] = original_df['job_title'].astype(str)
```

```python
# --- 1 Top Hiring Companies --- top_companies =
original_df['company_name'].value_counts().head(10)
plt.figure(figsize=(10,5)) sns.barplot(x=top_companies.values,
y=top_companies.index, palette="Blues_r") plt.title("Top 10 Hiring
Companies") plt.xlabel("Number of Jobs") plt.show()
```

```python
# --- 2 Top Job Locations --- top_locations =
original_df['location'].value_counts().head(10) plt.figure(figsize=(10,5))
sns.barplot(x=top_locations.values, y=top_locations.index, palette="coolwarm")
plt.title("Top 10 Job Locations") plt.show()
```

```python
# --- 3 Word Frequency in Job Titles --- words = ' '.join(original_df['job_title']).split()
common_words = Counter(words).most_common(15) plt.figure(figsize=(10,5)) sns.barplot(x=[w[1] for
w in common_words], y=[w[0] for w in common_words], palette="Purples_r") plt.title("Most
Frequent Words in Job Titles") plt.xlabel("Frequency") plt.ylabel("Word") plt.show()
```

```python
# --- 4 Seniority Level Distribution --- if
'seniority_level' in original_df.columns:
    plt.figure(figsize=(8,4))     sns.countplot(y='seniority_level', data=original_df,
order=original_df['seniority_level'].value_counts().index, palette="Greens_r")     plt.title("Seniority Level
Distribution")     plt.show()
```

```python
# --- 5 Job Function Distribution --- if
'job_function' in original_df.columns:
    plt.figure(figsize=(8,4))     sns.countplot(y='job_function', data=original_df,
order=original_df['job_function'].value_counts().index, palette="Oranges_r")     plt.title("Job Function Distribution")
plt.show()
```

```python
# --- 6 Employment Type Distribution --- if
'employment_type' in original_df.columns:
    plt.figure(figsize=(8,4))     sns.countplot(y='employment_type', data=original_df,
order=original_df['employment_type'].value_counts().index, palette="Reds_r")     plt.title("Employment Type
Distribution")     plt.show()
```

```python
# --- 7 Industry Distribution ---
if 'industry' in original_df.columns:
    plt.figure(figsize=(8,4))     sns.countplot(y='industry', data=original_df,
order=original_df['industry'].value_counts().head(10).index, palette="mako")     plt.title("Top 10 Industries Hiring")
plt.show()
```

```python
#   ---   8   Correlation   Heatmap   ---   numeric_df   =
df.select_dtypes(include=['float64','int64'])    if    not
numeric_df.empty:
    corr = numeric_df.corr()
plt.figure(figsize=(10,8))     sns.heatmap(corr,
annot=True, cmap='viridis')
plt.title("Correlation Heatmap")     plt.show()
```

```python
# --- 9 Interactive Plots --- fig = px.bar(top_companies, title="Top
Hiring Companies (Interactive)",
labels={'index':'Company', 'value':'Job Count'}) fig.show()
```

```python
fig = px.bar(top_locations, title="Top Job Locations (Interactive)",
labels={'index':'Location', 'value':'Job Count'}) fig.show()
```

```python
# ===========================
# STEP 8: EXPORT CLEANED DATASET
# ===========================
```

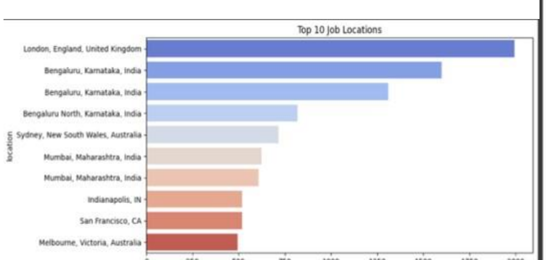
```python
clean_path = "/content/cleaned_linkedin_jobs.xlsx" df.to_excel(clean_path,
index=False)
print(f"\n✅ Cleaned dataset exported to: {clean_path}")
```

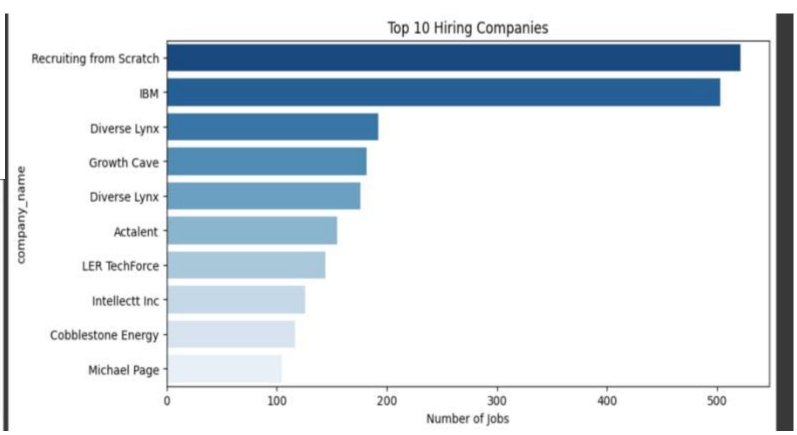
First 5 rows:

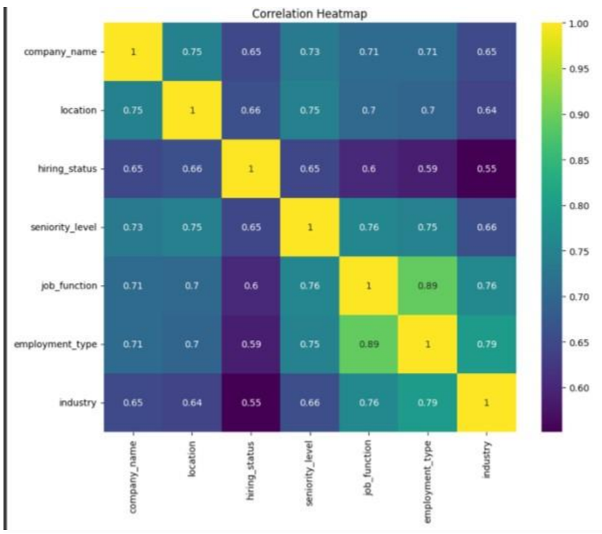
	job_title	company_name	location	hiring_status	date	seniority_level	job_function	employment_type	industry
0	Store Business Manager - DAVID JONES CHERMSIDE	M.J. Bale	Brisbane, Queensland, Australia	Be an early applicant	2023-04-13	Not Applicable	Sales and Business Development	Full-time	Government Administration
1	Full-time	Gatesman	Chicago, IL	Be an early applicant	2023-03-31	NaN	NaN	NaN	NaN
2	Senior Machine Learning Engineer	Redwolf + Rosch	Adelaide, South Australia, Australia	Be an early applicant	2023-04-25	Mid-Senior level	Engineering and Information Technology	Part-time	Internet Publishing
3	Senior Data Scientist	Bupa	Melbourne, Victoria, Australia	Be an early applicant	2023-04-29	Entry level	Engineering and Information Technology	Full-time	Technology, Information and Internet
4	\n \n Solution Architect\n ...	\n Xybion Digital\n	\n Chennai, Tamil Nadu, India\n ...	\n Be an early applicant\n	2023-01-26	\n Mid-Senior level\n	\n Engineering and Information Tech...	\n Full-time\n	\n IT Services and IT Consulting, Sof...

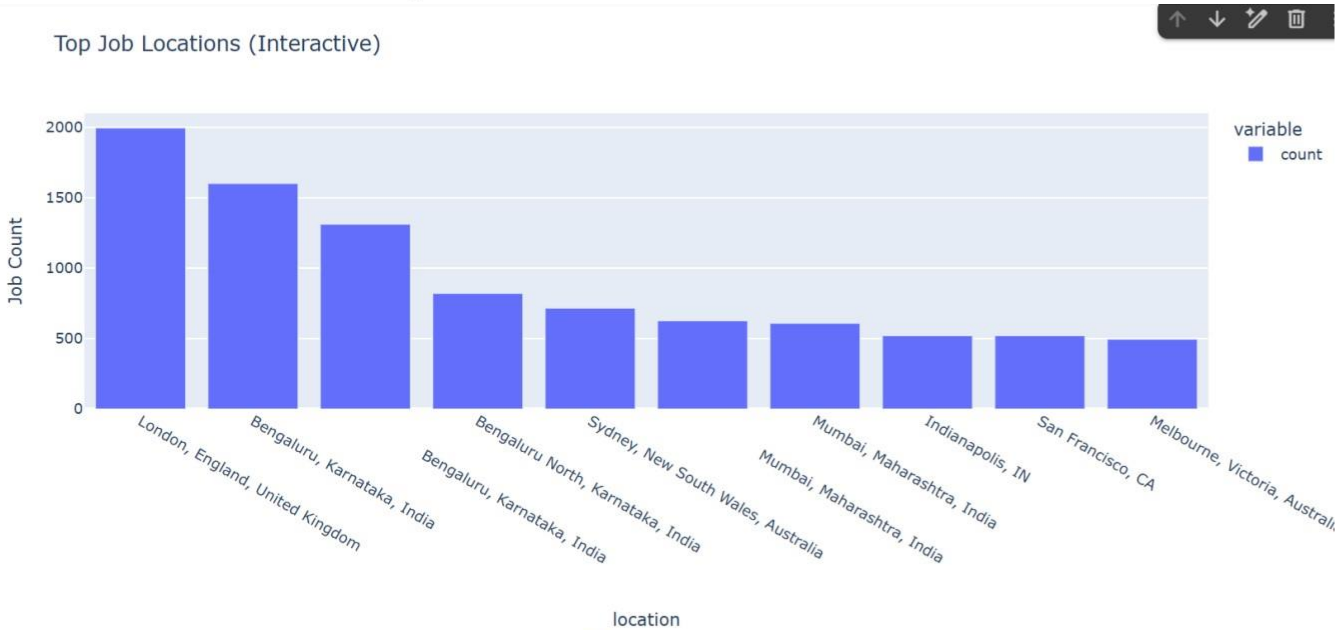
```
Shape: (31597, 9)

Data Types:
 job_title                object
company_name             object
location                 object
hiring_status            object
date             datetime64[ns]
seniority_level          object
job_function             object
employment_type          object
industry                 object
dtype: object

Missing Values:
 job_title            26
company_name        940
location              9
hiring_status         0
date                  0
seniority_level    1308
job_function       1590
employment_type    1591
industry           2011
dtype: int64
```

```
Basic Statistics:
                   job_title             company_name  \
count                  31571                    30657
unique                  6112                     7201
top     Full Stack Developer  Recruiting from Scratch
freq                     718                      521
mean                     NaN                      NaN
min                      NaN                      NaN
25%                      NaN                      NaN
50%                      NaN                      NaN
75%                      NaN                      NaN
max                      NaN                      NaN

                               location        hiring_status  \
count                             31588                31597
unique                             2043                   47
top     London, England, United Kingdom  Be an early applicant
freq                               1994                14423
mean                                NaN                  NaN
min                                 NaN                  NaN
25%                                 NaN                  NaN
50%                                 NaN                  NaN
75%                                 NaN                  NaN
max                                 NaN                  NaN

                                 date  seniority_level  \
count                           31597            30289
unique                            NaN               25
top                               NaN Mid-Senior level
freq                              NaN             8651
mean    2023-03-12 20:52:44.199132928              NaN
min               2021-05-27 00:00:00              NaN
25%               2023-02-23 00:00:00              NaN
50%               2023-03-20 00:00:00              NaN
75%               2023-04-08 00:00:00              NaN
max               2023-04-29 00:00:00              NaN

                                  job_function employment_type  \
count                                    30007           30006
unique                                     542              18
top     Engineering and Information Technology       Full-time
freq                                      7141           18305
mean                                       NaN             NaN
min                                        NaN             NaN
```










```
✅ Cleaned dataset exported to: /content/cleaned_linkedin_jobs.xlsx
```