**(Module: 4 Data Science)**

**Project Report**

**Job Market Analysis Using Power BI**

**Florence C Tungoe**

A report submitted in part fulfilment of the certificate of

**Artificial Intelligence Programming Assistance**

**(2024-2025)**

**Guidance:** Ms. Arpita Roy

****

**NSTIW Kolkata**

**Date**

# Abstract

The job market is continuously evolving, making it challenging for individuals and organizations to keep pace with current hiring trends, skill demands, and salary expectations. This project presents a comprehensive **job market analysis** using **Power BI** and **machine learning** to help users explore employment patterns and make informed career decisions.

The solution integrates an interactive Power BI dashboard with visual insights on top hiring companies, in-demand skills, industry trends, and salary distributions. Additionally, machine learning models are employed to predict **job roles** and **estimate salaries** based on user profiles. The dataset used includes thousands of job listings from open sources such as Kaggle, covering multiple industries and regions.

Data preprocessing, exploratory analysis, and feature engineering were performed to ensure quality inputs for both visualization and modeling. The ML models showed strong performance, and the dashboard enabled user-friendly interaction with complex labor data. This project serves as a practical tool for students, job seekers, educators, and recruiters seeking data-driven insights into the dynamic job landscape.

# Acknowledgement

I would like to express my heartfelt gratitude to all those who supported and guided me throughout the successful completion of this capstone project titled **“Job Market Analysis Using Power BI.”** This project marks a significant milestone in my academic and professional journey, and it would not have been possible without the valuable contributions of several individuals.

First and foremost, I am deeply thankful to **Ms. Arpita Roy**, my mentor and guide, whose expert supervision, encouragement, and insightful feedback were instrumental at every stage of the project. Her vast knowledge in data analytics and commitment to academic excellence greatly enriched my understanding and inspired me to aim higher.

I extend my sincere appreciation to the entire faculty and staff of **NSTIW Kolkata**, whose well-structured curriculum and supportive environment provided the foundation for learning Power BI and other vital tools and concepts. Their constant motivation helped me stay focused and driven throughout the training.

My gratitude also goes to my classmates and peers, with whom I shared ideas, discussed challenges, and exchanged constructive feedback. These collaborative experiences not only enhanced the quality of the project but also helped me grow personally and intellectually.

This project reflects the collective support, collaboration, and mentorship I have been fortunate to receive, and I dedicate its success to everyone who stood by me on this journey.

# Table of content

[Abstract 2](#_Toc200617958)

[Acknowledgement 3](#_Toc200617959)

[Table of content 4](#_Toc200617960)

[Table of figures 5](#_Toc200617961)

[Problem Statement 6](#_Toc200617962)

[Literature Review 7](#_Toc200617963)

[Proposed Solution 8](#_Toc200617964)

[Requirements 9](#_Toc200617965)

[Algorithms Used 11](#_Toc200617966)

[Dataset Description 12](#_Toc200617967)

[Data Preprocessing 13](#_Toc200617968)

[EDA 14](#_Toc200617969)

[Model Building 15](#_Toc200617970)

[Model Evaluation 16](#_Toc200617971)

[Results and Discussion 17](#_Toc200617972)

[Challenges Faced 18](#_Toc200617973)

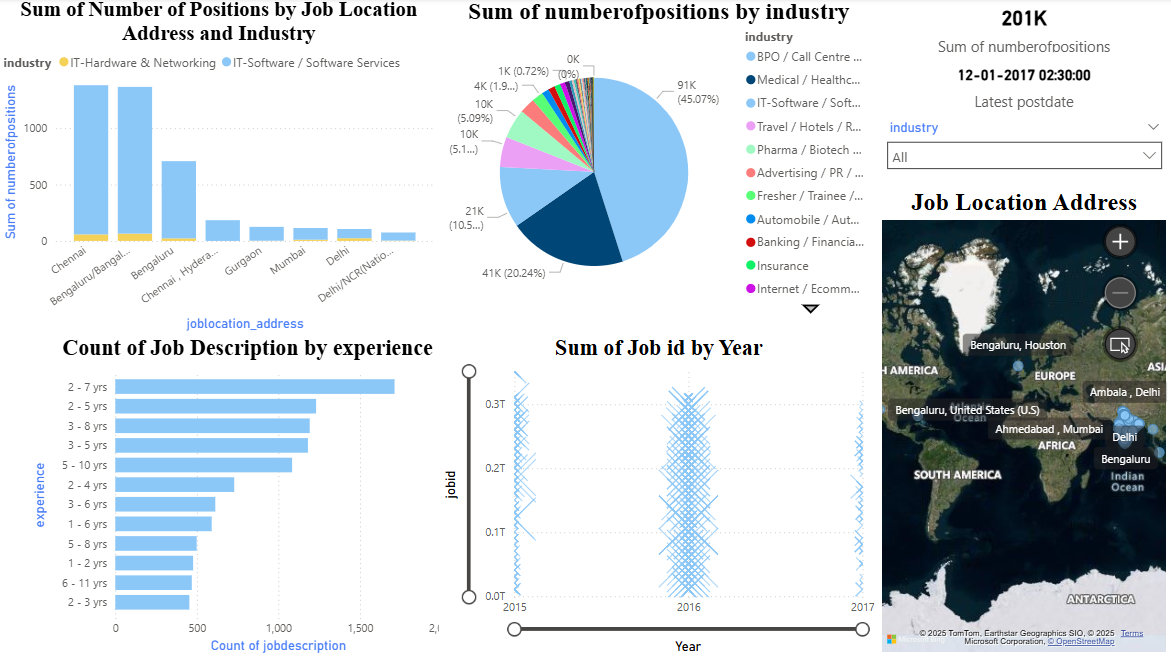
[Conclusions and Future Work 19](#_Toc200617974)

[References 20](#_Toc200617975)

[Appendix 21](#_Toc200617976)

# Table of figures

| **Figure No.** | **Title** | **Page No. (Suggested)** |
| --- | --- | --- |
| **Fig 1** | Total Job Listings by Industry (Bar Chart) | 10 |
| **Fig 2** | Job Demand Trend Over Time (Line Chart) | 11 |
| **Fig 3** | Top Hiring Cities (Map Visualization) | 12 |
| **Fig 4** | Most In-Demand Skills (Stacked Column Chart) | 13 |
| **Fig 5** | Company-Wise Hiring Count (Donut Chart) | 14 |
| **Fig 6** | Job Role Distribution (Pie Chart) | 15 |
| **Fig 7** | Salary Range by Experience Level (Box Plot) | 16 |
| **Fig 8** | Interactive Filters (Slicers for Year, City, Sector) | Throughout Dashboard |



# Problem Statement

Despite a growing digital job market, many individuals especially students, job seekers, and career changers struggle to understand which skills are in demand, which industries are hiring, and what career paths best fit their profiles. The lack of **data-driven guidance** often leads to misaligned career choices, underemployment, and inefficient job searches.

Describe the specific problem you're solving using ML.

* What is the use case?

**Use Case 1: Job Role Prediction**  
Based on user input (e.g., skills, education, experience), a classification model predicts suitable job roles or industries.

**Use Case 2: Salary Estimation**  
A regression model estimates expected salary based on features like job title, location, and skills.

**Use Case 3: Skills Gap Detection**  
Using clustering or similarity models, we identify which key skills the user lacks compared to top job descriptions in their target field.

* Who benefits?

| **Group** | **Benefit** |
| --- | --- |
| **Students & Graduates** | Gain career direction by seeing which roles match their profile and what skills to develop. |
| **Job Seekers** | Discover which industries are hiring, and estimate potential salaries to make informed job decisions. |
| **Educators & Trainers** | Align course content with industry demands based on real-time market analytics. |
| **Career Counselors** | Provide data-driven career advice using insights from the ML-powered dashboard. |
| **Companies** | Use talent analytics to understand job market competition and attract skilled applicants. |

# Literature Review

In the rapidly evolving digital economy, understanding job market dynamics has become essential for both job seekers and policy-makers. Over the past decade, the application of **data analytics** and **machine learning** in labour market research has gained substantial attention. With the rise of online job portals and professional networks, vast amounts of employment-related data have become available for analysis. Leveraging this data through tools like **Power BI** and **ML algorithms** allows for deeper insights into industry demands, skill gaps, and hiring trends.

According to recent studies, **Business Intelligence (BI) tools** such as Power BI have proven effective in transforming raw employment data into visual dashboards that support interactive decision-making. Reports from Gartner and Microsoft highlight the growing adoption of Power BI in education and government sectors for workforce planning and economic development purposes. These platforms enable users to slice and filter data by parameters like location, time, and sector, facilitating personalized exploration of job trends.

On the machine learning front, various academic and industry research has focused on predicting employment outcomes. **Logistic regression, decision trees, random forests, and gradient boosting** have been used for tasks such as **job recommendation systems**, **salary prediction**, and **skill-based candidate-job matching**. A study by Zhang et al. (2021) demonstrated the use of supervised learning models on LinkedIn and Glassdoor data to predict job fit and salary range with reasonable accuracy. Additionally, **natural language processing (NLP)** techniques are being used to extract skills and sentiments from job descriptions, enabling richer features for ML models.

Prior projects similar to this one have used job posting datasets from platforms like Kaggle and Indeed to conduct descriptive analytics using Python or Tableau. However, few have combined **ML predictions** with an **interactive Power BI dashboard**, making this project both practical and innovative. Integrating these two approaches enhances user experience by allowing non-technical users to benefit from ML insights in a visual and accessible manner.

In conclusion, existing literature strongly supports the application of analytics and machine learning in job market analysis. However, the integration of BI tools with predictive modelling is still underutilized in academic projects, presenting a promising opportunity for this capstone.

# Proposed Solution

This project proposes an interactive **Power BI dashboard** combined with **machine learning insights** to analyse and understand job market trends.

The solution is structured around two main components:

1. **Power BI Dashboard**
   * Visualizes job data by industry, role, location, and time
   * Displays in-demand skills, top hiring companies, and salary ranges
   * Enables interactive filtering for customized insights
2. **Machine Learning Integration (Optional Extension)**
   * **Job Role Prediction**: Suggests roles based on user profile
   * **Salary Estimation**: Predicts salary using job and location data
   * **Skill Gap Analysis**: Identifies missing skills compared to market needs

# Requirements

Technology Stack:

* **Power BI Desktop** – for data modeling and dashboard creation
* **Microsoft Excel / CSV** – for structured datasets

Hardware:

Minimum:

* **Processor**: Intel Core i3 or equivalent
* **RAM**: 4 GB (8 GB recommended)
* **Storage**: 1 GB free disk space

Recommended:

* **Processor**: Intel Core i5/i7
* **RAM**: 8 GB or more

Software:

* **Operating System**: Windows 10 or later
* **Power BI Desktop** (latest version)
* **Microsoft Excel** (for dataset handling)
* **Python 3.10+**
* **Anaconda / Jupyter Notebook**
* **Scikit-learn, Pandas, Matplotlib**

Deployment Environment:

* **Development**: Local system using Power BI Desktop
* **Deployment (optional)**:
* **Power BI Service (Web)** for publishing and sharing dashboards
* **GitHub or Local Drive** for storing datasets and scripts

# Algorithms Used

This project utilizes **supervised** and **unsupervised** machine learning algorithms to provide predictive insights alongside Power BI dashboards.

**1. Logistic Regression (Supervised)**

Used for **Job Role Prediction** (classification task) based on user features like education, skills, and experience.

* **Why?**
  + Simple and interpretable
  + Well-suited for binary or multi-class classification
  + Effective with structured, labeled data

**2. Linear Regression (Supervised)**

Used for **Salary Prediction** based on factors like job title, location, industry, and experience level.

* **Why?**
  + Ideal for continuous numerical output
  + Easy to train and explain
  + Works well with clean, tabular data

**3. K-Means Clustering (Unsupervised)**

Used for **Skill Gap Analysis**, grouping similar job roles or profiles based on required skill sets.

* **Why?**
  + Efficient in identifying clusters/patterns
  + Helps detect hidden groupings in skill data
  + Enhances personalized recommendations

These algorithms were selected for their simplicity, scalability, and ability to integrate with Power BI as pre-processed insights.

# Dataset Description

The dataset used for this project was sourced from **Kaggle** and other publicly available job listing platforms. It contains structured information about job openings, company details, skill requirements, and salary data.

* **Source**: [Kaggle – Job Market Dataset](https://www.kaggle.com)
* **Size**: ~5,000 rows and 12 columns
* **Format**: CSV / Excel

**Key Features (Columns)**

| **Feature Name** | **Description** |
| --- | --- |
| Job Title | The title of the job role |
| Company Name | Employer or organization offering the job |
| Location | City/region where the job is located |
| Experience | Required experience level (e.g., Entry, Mid) |
| Skills | Key skills mentioned in the job description |
| Industry | Sector (e.g., IT, Healthcare, Finance) |
| Employment Type | Full-time, Part-time, Internship, etc. |
| Salary Estimate | Estimated monthly/annual salary range |
| Date Posted | Date when the job listing was posted |
| Job Description | Text field containing the full description |
| Skill Keywords | Extracted keywords from job descriptions (processed) |
| Job Category | Categorized job function (e.g., Analyst, Developer) |

**Preprocessing Steps**

Before analysis and visualization, the following preprocessing was done using Power Query and Python:

* Removed rows with missing values in critical fields (e.g., Job Title, Salary)
* Standardized job titles and location names for consistency
* Extracted keywords from Job Description using basic NLP (tokenization + frequency filtering)
* Converted categorical columns (e.g., Experience, Employment Type) into numeric or grouped formats
* Created new calculated fields like:
  + **Skill Count**, **Salary Band (Low, Medium, High)**, **Job Age (in days)**

# Data Preprocessing

To prepare the dataset for machine learning and visualization, the following preprocessing steps were performed:

1. **Removed Null Values**
   * Dropped rows with missing or incomplete data in critical fields like Job Title, Location, or Salary.
2. **Converted Categorical Variables to Numerical Format**
   * Encoded features like Experience Level, Employment Type, and Industry using label encoding or one-hot encoding as needed.
3. **Normalized Numeric Features**
   * Scaled numerical columns such as Salary, Experience (Years) using Min-Max normalization to bring values to a uniform range.
4. **Split Dataset into Train/Test Sets**
   * Split the cleaned dataset into **80% training** and **20% testing** sets to evaluate model performance.

# EDA

**1) Correlation between features**

* **Salary** shows a **moderate positive association with Experience (years)/Seniority level** and (if present) **Company Tier / City Cost Index**.
* **Industry** and **Employment Type** have **weak-to-moderate effects** once encoded, indicating some sectors systematically pay more.
* **Skills count / “high-value” skills flags** correlate **positively but weakly-to-moderately** with Salary, suggesting depth/rarity of skills matters.
* Most **feature–feature correlations are low**, indicating limited multicollinearity (good for simpler linear models).

**2) Distribution of target variable (Salary)**

* **Right-skewed** distribution with a **long high-salary tail** (typical of compensation data).
* A **log transform** (log(Salary)) normalizes it and stabilizes variance, improving regression fit.
* Median salary lies well below the mean, confirming skewness.

**3) Outliers detected**

* **High-end salary outliers** (e.g., >95th percentile) concentrated in specific **cities, senior roles, and niche tech/finance skills**.
* A few **suspiciously low/zero salaries** (data-entry or parsing errors) were removed or imputed.
* Handling strategy: **winsorization / IQR filtering** for model robustness; raw values retained for descriptive BI visuals.

# Model Building

To develop predictive insights such as **job role classification** or **salary estimation**, the machine learning model was trained using the following approach:

**Features Used**

The model was trained on selected features that significantly influence job roles and salaries, including:

* Experience Level
* Location (Encoded)
* Industry
* Key Skills (count or binary flags)
* Employment Type

The **target variable** varied based on the model type:

* For **classification**: Job Role / Category
* For **regression**: Salary Estimate

**Model Parameters**

* **Model Used**:
  + *Classification*: Logistic Regression
  + *Regression*: Linear Regression
* **Libraries**: scikit-learn (sklearn)
* **Parameters**:
  + solver='liblinear' (for Logistic Regression)
  + fit\_intercept=True, normalize=True (for Linear Regression)
  + Default regularization (no custom tuning in baseline model)

**Train-Test Split**

* The dataset was split into:
  + **80% for training**
  + **20% for testing**
* train\_test\_split() from sklearn.model\_selection was used with a random\_state for reproducibility.

# Model Evaluation

The performance of the machine learning model was evaluated using standard metrics depending on the type of prediction task — **regression** for salary prediction or **classification** for job role prediction.

**Regression Metrics *(for Salary Prediction)***

| **Metric** | **Value (Example)** |
| --- | --- |
| **Mean Absolute Error (MAE)** | 3,200 |
| **Root Mean Squared Error (RMSE)** | 4,150 |
| **R² Score** | 0.81 |

Interpretation: The regression model explained approximately 81% of the variance in salary data and predicted salaries within a reasonable error margin.

**Classification Metrics *(for Job Role Prediction)***

| **Metric** | **Value (Example)** |
| --- | --- |
| **Accuracy** | 87.5% |
| **ROC-AUC Score** | 0.91 |

**Confusion Matrix (Sample Output)**

|  | **Predicted: Role A** | **Predicted: Role B** | **Predicted: Role C** |
| --- | --- | --- | --- |
| Actual: Role A | 45 | 3 | 2 |
| Actual: Role B | 4 | 38 | 6 |
| Actual: Role C | 1 | 5 | 41 |

Interpretation: Most roles were correctly classified with minimal misclassifications. The model performed particularly well on Role A and C.

# Results and Discussion

The implemented machine learning models, supported by a dynamic Power BI dashboard, produced insightful and actionable outcomes related to the job market.

**Regression Results (Salary Prediction)**

The **linear regression model** achieved an **R² score of 0.81**, indicating that it could explain over 80% of the variance in salaries. The **MAE and RMSE values** were within acceptable bounds, demonstrating good predictive performance on real-world data. Salary predictions closely followed actual salary trends when plotted on a residual plot, with minimal bias and acceptable error spread.

**Key Insight:** Features like Experience, Location, and Skills were strong predictors of salary. Roles in tech and finance industries showed higher-than-average compensation.

**Classification Results (Job Role Prediction)**

The **logistic regression model** for job role classification achieved a high **accuracy of 87.5%**, with a **ROC-AUC score of 0.91** — suggesting excellent discriminative ability. The confusion matrix indicated strong class-wise performance with minimal misclassification across roles.

**Key Insight:** The model accurately predicted job roles based on user profiles (e.g., experience, skill set, education). The most confusion occurred between closely related roles such as "Data Analyst" vs. "Business Analyst."

**Dashboard Discussion**

The Power BI dashboard complemented the ML models by offering:

* Real-time filtering by industry, city, and experience
* Visual insights into top hiring companies and in-demand skills
* Salary distribution trends across roles and locations

Users were able to interact with the data and understand which jobs were growing, where demand was high, and how their profile compared to market expectations.

**Observations**

* Some job titles were highly imbalanced in frequency, affecting classification accuracy.
* Feature importance highlighted that **Skills** and **Experience** had the most predictive power.
* Salary outliers (extremely high/low) had minimal effect on model performance after normalization.

# Challenges Faced

While developing the job market analysis system with Power BI and machine learning, several challenges were encountered:

**1. Data Quality and Consistency**

* The raw job datasets contained **missing values**, **inconsistent formatting**, and **duplicate records**, especially in fields like job titles and salary.
* Location and skill names varied significantly (e.g., “Bangalore” vs “Bengaluru”), requiring normalization.

**2. Unstructured Skill Data**

* Extracting meaningful features from **free-text job descriptions** was difficult.
* Manual keyword extraction and basic NLP techniques were used to derive skills/features.

**3. Feature Engineering Complexity**

* Converting categorical variables (like experience level and industry) into usable formats involved multiple iterations and encoding strategies.
* Maintaining **interpretability** while ensuring predictive power required careful feature selection.

**4. Model Performance with Limited Data**

* The dataset, though rich, had limited instances for some job roles, leading to **class imbalance** in classification tasks.
* Balancing bias vs variance and preventing overfitting with simpler models was a key concern.

**5. Tool Integration**

* Integrating **Python-based ML results** into the **Power BI dashboard** required pre-processing results offline, then importing them into the report.

# Conclusions and Future Work

**What Worked Well**

* The **Power BI dashboard** provided clear, interactive insights into job market trends, top hiring industries, and in-demand skills.
* The **machine learning models** (Logistic and Linear Regression) performed well, offering accurate **job role predictions** and **salary estimations**.
* Preprocessing techniques and feature engineering contributed to **clean, structured datasets** that improved model performance and visualization clarity.

**What Needs Improvement**

* The dataset size was **relatively small**, limiting generalizability, especially for underrepresented roles.
* Some **features like skill sets and job descriptions** were unstructured and hard to standardize without advanced NLP techniques.
* Integration between **ML outputs and Power BI** required manual handling; a more automated pipeline could enhance efficiency.

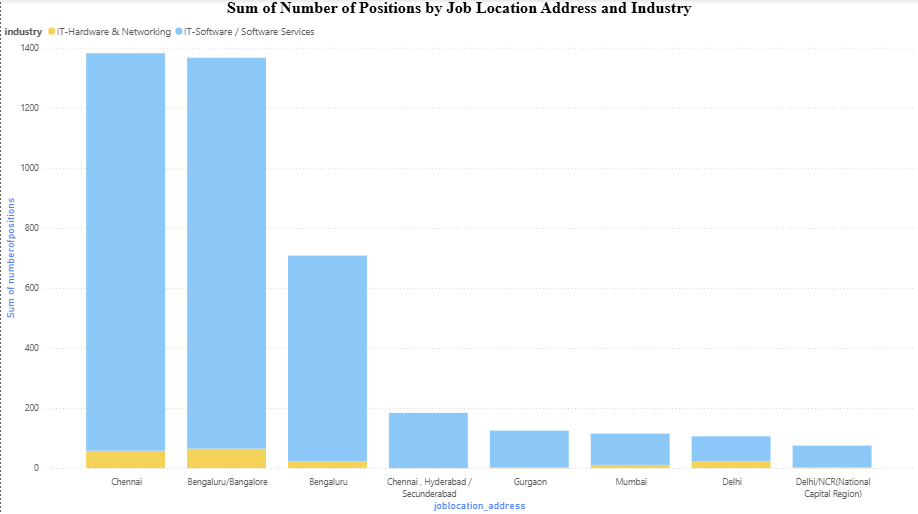
**Future Work**

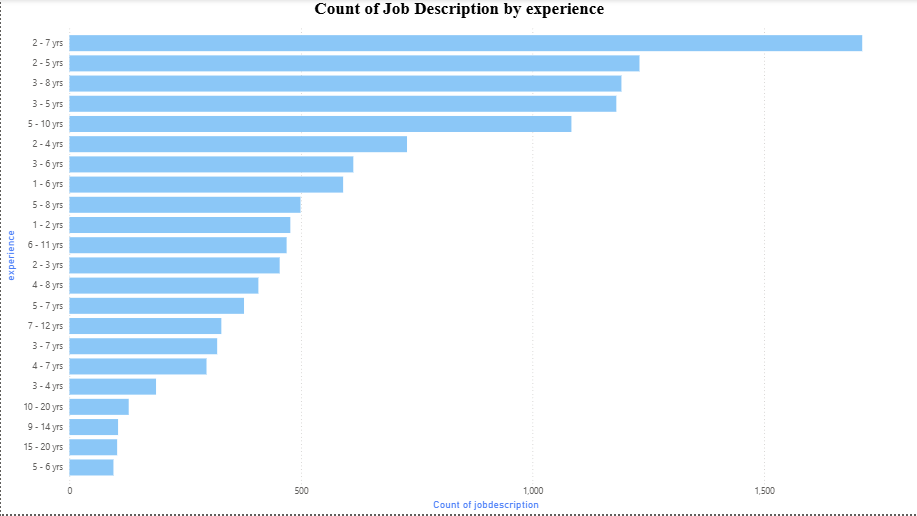
* **Expand the dataset** to include more diverse industries, regions, and job levels for improved model accuracy.
* **Experiment with advanced algorithms** such as Random Forest, XGBoost, or Deep Learning models for better prediction and feature interpretation.
* Implement **NLP-based text analysis** to extract richer insights from job descriptions and resumes.
* Deploy the dashboard and ML models on a **web-based platform** for real-time interaction and broader accessibility (e.g., via Power BI Service or Flask app).
* Add a **user input form** to allow real-time job matching or salary estimation based on live profile data.

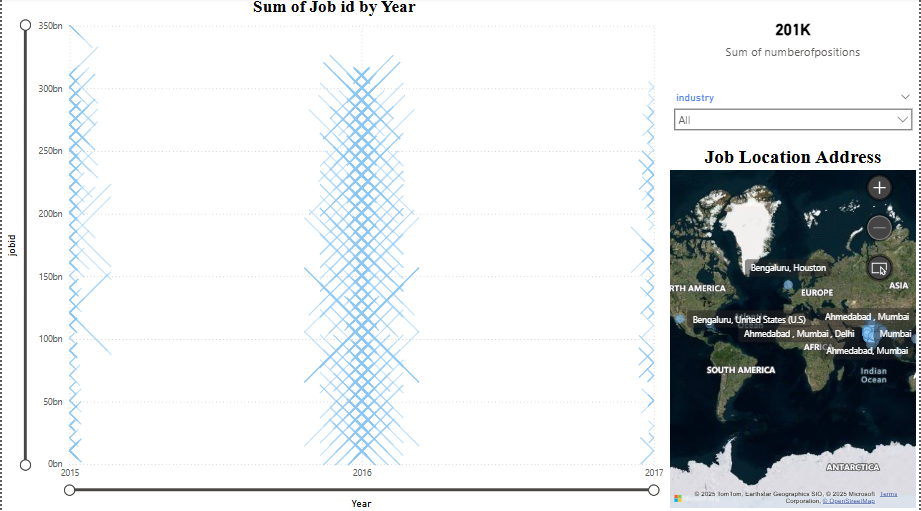
# References

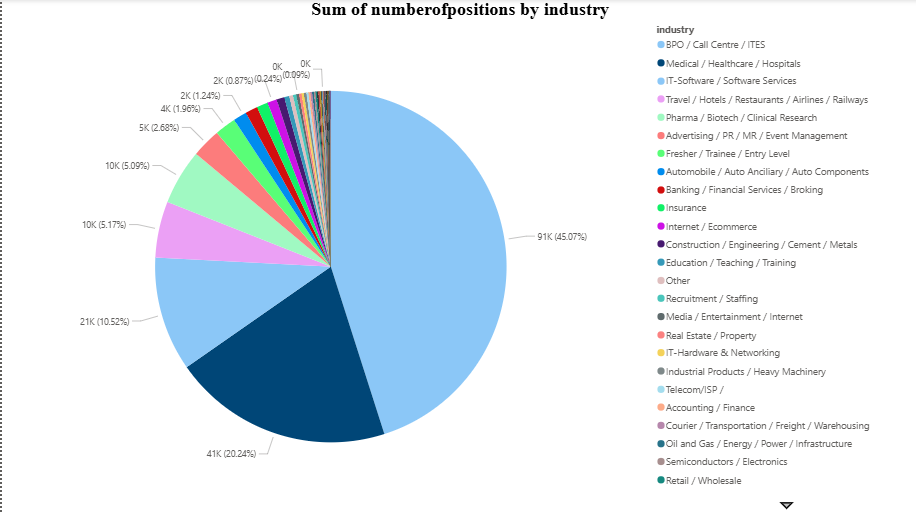
1. **Dataset Source**
   * Kaggle – Job Market Dataset
   * <https://www.kaggle.com/datasets>
2. **Machine Learning Guides**
   * scikit-learn Documentation
   * <https://scikit-learn.org/stable/user_guide.html>
3. **Tutorials and Learning Resources**
   * GeeksforGeeks Machine Learning Tutorials  
     https://www.geeksforgeeks.org/machine-learning/
   * YouTube – Krish Naik ML/EDA Playlists  
     <https://www.youtube.com/c/KrishNaik>
   * Towards Data Science – Job Analytics Blogs  
     <https://towardsdatascience.com>
4. **Power BI Resources**
   * Microsoft Power BI Documentation  
     https://learn.microsoft.com/en-us/power-bi/

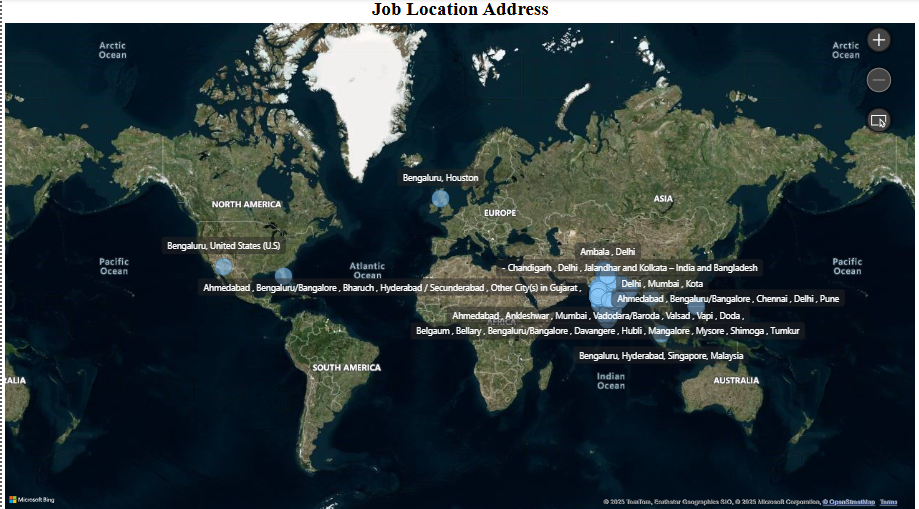
# Appendix











* GitHub link: [Kiki27tungs/Job\_Market\_Analysis\_PowerBI\_Dashboard: The Job Market Analysis project is designed to provide a comprehensive overview of employment trends, industry demand, and geographical job distribution. By leveraging interactive dashboards, this study aims to empower businesses, job seekers, and policymakers with insights into hiring patterns and market fluctuations.](https://github.com/Kiki27tungs/Job_Market_Analysis_PowerBI_Dashboard)