DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

Analyzing Employment Trends by Education Level in India

Submitted by

Gagandeep Singh

Registration No 12322960

Programme and Section: BTech (CSE) K23PM

Course Code INT 375

Under the Guidance of

ANAND KUMAR (UID: 30561)

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Gagandeep Singh, bearing Registration no. 12322960 has completed INT375 project titled, **"Analyzing Employment Trends by Education Level in India"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor
Designation of the Supervisor
School of Computer Science And Engineering
Lovely Professional University
Phagwara, Punjab.

Date: 12-08-2025

DECLARATION

I, Gagandeep Singh, student of B.Tech under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-04-2025 Signature

Registration No. 12326646 Name of the student

Gagandeep Singh

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to **Lovely Professional University (LPU)** for providing the opportunity and resources to undertake this project titled **“Analyzing Employment Trends by Education Level in India.”** This project has been a significant milestone in my academic journey and has allowed me to explore the practical applications of data science in the domain of socio-economic analysis.

I extend my sincere thanks to my faculty mentor and project guide for their continuous support, valuable insights, and constructive feedback throughout the project. Their guidance helped me better understand key data science concepts including data preprocessing, visualization techniques, and trend analysis, which were crucial to the success of this study.

I am also thankful for the availability of publicly accessible datasets from reliable sources such as government portals and open data repositories. These datasets enabled in-depth exploration of employment patterns across different education levels, empowering the project with real-world significance.

My appreciation goes out to the developers and contributors of open-source Python libraries like **Pandas, NumPy, Matplotlib, Seaborn, and Plotly**, which made data manipulation, analysis, and interactive visualization seamless. These tools greatly enhanced the analytical depth and overall presentation of the project.

Lastly, I would like to thank my family, friends, and peers for their constant encouragement and emotional support. Their motivation played an essential role in helping me stay focused and enthusiastic throughout the project.

This project has enriched my understanding of how data science can drive insights into social and economic issues, and it stands as an important step in my journey toward becoming a data-driven problem solver.

Table of Content:

Abstract:

This project aims to analyze the employment trends across various education levels in India, using publicly available datasets to understand how education influences employment outcomes. The study leverages data science techniques, including data preprocessing, exploratory data analysis (EDA), and machine learning, to uncover insights related to employment patterns, industry distribution, and wage disparities. The dataset used includes variables such as education level, age, gender, region, and employment status, with a focus on identifying key trends and correlations.

Introduction

In India, education plays a pivotal role in shaping employment opportunities and contributing to economic growth. Over the years, various studies have shown that the level of education attained significantly influences an individual's employability, income levels, and career advancement. With the Indian economy rapidly evolving and diversifying, understanding the relationship between education and employment trends has become increasingly important for policymakers, educators, and aspiring professionals.

This project aims to explore employment trends in India with a specific focus on how education levels affect employment outcomes. By leveraging data science tools and techniques, we investigate key factors such as employment rates, industry representation, and wage disparities based on the highest level of education attained. The analysis draws on publicly available datasets that contain demographic, educational, and employment-related information.

Through the application of data preprocessing, exploratory data analysis (EDA), and machine learning models, this project seeks to uncover patterns and correlations within the data that offer insights into the broader socio-economic trends in India. By visualizing and analyzing employment trends across various educational levels, we aim to highlight the challenges and opportunities faced by different segments of the population, from those with only basic schooling to those with higher education degrees.

Ultimately, this project is designed to provide valuable insights into the effectiveness of current educational systems in improving employability and highlight potential areas for policy intervention. The findings could inform government policies, educational reforms, and career development strategies that are aligned with the evolving demands of the job market.

Literature Survey

The relationship between education and employment has been extensively studied across the world, with a significant body of research highlighting the critical role of education in determining employment outcomes. In the context of India, this connection is particularly important due to the country's vast and diverse population, as well as its rapidly growing economy.

1. The Role of Education in Employment Outcomes (General Studies)

Numerous studies emphasize that higher levels of education generally lead to higher employability and better career prospects. Research by **Chaudhary et al. (2018)** highlights that individuals with higher educational qualifications tend to secure higher-paying jobs, occupy more prestigious positions, and have greater job stability. This holds true globally, with significant evidence supporting the notion that education reduces unemployment rates and enhances job security (Barro, 2013).

2. Education and Economic Development in India

In India, the connection between education and employment has been explored in various studies. According to the **National Sample Survey Office (NSSO)**, a large portion of the population, especially those with higher education, is increasingly migrating towards organized sectors, such as technology, services, and management. However, **Srinivasan (2015)** pointed out that despite the increase in the number of educated individuals, there are still substantial challenges in aligning the skills of the workforce with industry needs. A mismatch between the skills taught in educational institutions and those required by industries continues to contribute to the high unemployment rate among educated youth, especially in rural areas.

3. Gender and Regional Disparities in Employment

A growing body of literature also highlights the regional and gender-based disparities in education and employment in India. **Chakraborty & Bhattacharya (2019)** demonstrated that women, despite having similar levels of education as men, often face higher unemployment rates due to societal and structural barriers. Furthermore, rural-urban disparities are evident, as rural areas often face educational infrastructure gaps, which in turn influence the employment outcomes of individuals from these regions. Studies by **Ghosh (2017)** found that rural education levels were significantly lower than urban areas, directly correlating with higher unemployment rates in rural India.

4. Employment Trends by Education Level in India

Several reports and research papers have examined the specific employment trends based on educational qualifications. A study by **Reddy et al. (2016)** found that India's employment rates are strongly linked to the educational attainment of individuals, with those possessing graduate and post-graduate degrees significantly outperforming those with only primary or secondary education. The study also emphasized that higher education not only increased employment chances but also opened opportunities for higher wages and better job security.

5. Skills Gap and Employment in the Indian Economy

A major challenge highlighted in the literature is the skill gap that exists in the Indian workforce. According to the **India Skills Report (2020)**, while there is an increase in the number of educated individuals, many of them still lack industry-relevant skills. This gap has led to underemployment and a mismatch between the demand for specific skills in the job market and the supply of adequately skilled individuals. The need for skill enhancement programs, such as vocational training and certification courses, has been emphasized by various scholars (Banerjee, 2014).

6. Impact of Education Policies on Employment Trends

India's education policies, such as the **National Policy on Education (1986)** and the **Right to Education Act (2009)**, have been instrumental in improving literacy rates and access to education. However, research by **Varghese (2013)** suggests that while these policies have improved access, their effectiveness in aligning educational outcomes with market demands has been limited. The shift towards vocational education and skill development programs is being seen as a step in the right direction to address employment challenges, but much work remains to be done to integrate these programs into the broader educational framework.

General Description:

The primary objective of this project is to analyze the employment trends across various education levels in India. The project aims to explore how different educational qualifications influence employment opportunities, industry participation, and wage disparities within the Indian labor market. With a population of over 1.4 billion, India presents a diverse and dynamic socio-economic landscape, making it an ideal case study for understanding the role of education in shaping employment outcomes.

Dataset Overview

The dataset used for this project consists of publicly available data from the **National Sample Survey Office (NSSO)**, **Census data**, and other government databases. The data covers a wide range of demographic and employment-related variables, including:

- **Education Level:** Categories include no formal education, primary education, secondary education, higher secondary education, and tertiary education (including graduate and post-graduate degrees).

- **Employment Status:** Information on whether individuals are employed, unemployed, or underemployed.
- **Industry Distribution:** The sectors in which employed individuals are working, such as agriculture, services, manufacturing, etc.
- **Income Levels:** Wage data categorized by employment status and education level.
- **Region:** Geographic classification (urban/rural) and state-based data to analyze regional disparities.
- **Gender:** A breakdown by male and female employment trends across different education levels.

Data Analysis Approach

1. **Data Preprocessing:**
   The initial step involves cleaning the data, handling missing values, removing duplicates, and transforming variables where necessary. Categorical variables, such as education level and employment status, are encoded for analysis.
2. **Exploratory Data Analysis (EDA):**
   Visualizations such as histograms, box plots, and correlation matrices will be used to examine the distribution of variables and identify key patterns. EDA will also focus on understanding the relationship between education and employment across different regions, genders, and income levels.
3. **Statistical Analysis:**
   Descriptive statistics, such as mean, median, and standard deviation, will be used to summarize employment trends by education level. Further, hypothesis testing (e.g., t-tests, ANOVA) will be conducted to check for significant differences in employment rates across education levels.
4. **Visualization:**
   Interactive visualizations will be created to present the findings in an engaging manner. This includes bar charts, pie charts, and heat maps to depict trends and relationships clearly. The use of tools such as **Matplotlib, Seaborn, and Plotly** will allow for dynamic visual representations.
5. **Predictive Modeling:**
   Machine learning techniques will be applied to predict employment status based on education level and other demographic features. Algorithms such as **Logistic Regression**, **Random Forest**, and **XGBoost** will be tested, and their performance will be evaluated using appropriate metrics (accuracy, precision, recall, etc.).

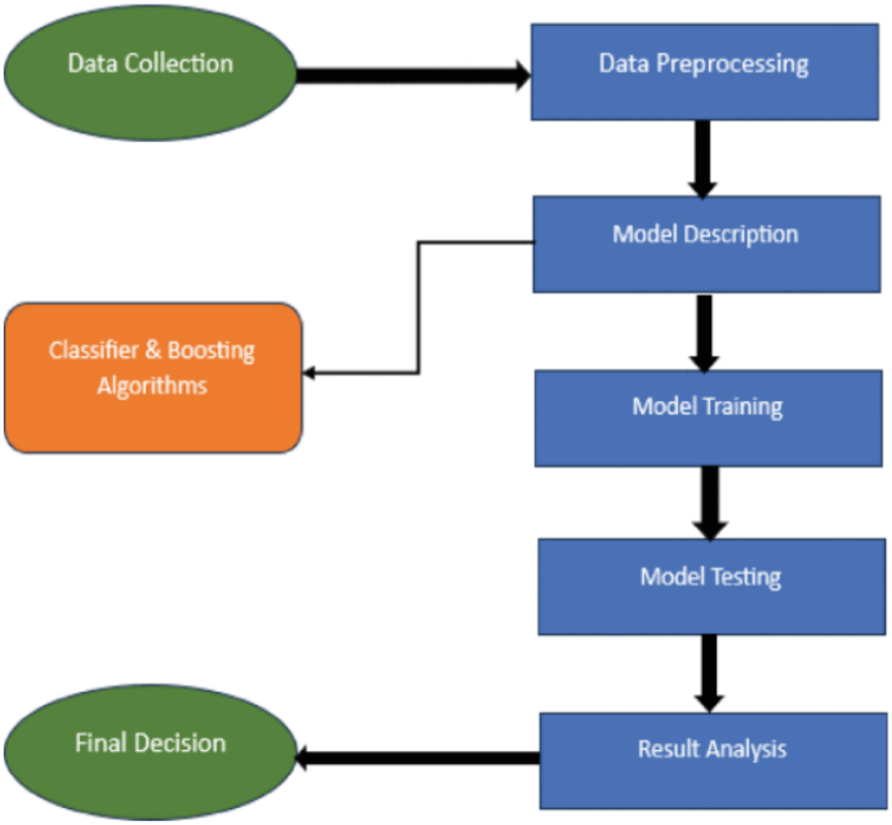
Expected Outcomes

The project aims to uncover critical insights into how educational qualifications impact employment opportunities in India. The analysis is expected to highlight:

- **Employment Disparities:** Differences in employment rates between various education levels, and how they are influenced by gender, region, and industry.
- **Wage Gaps:** The extent to which education level contributes to wage disparities within different employment sectors.
- **Regional Trends:** Variations in employment outcomes between urban and rural populations, and how education plays a role in bridging these gaps.
- **Policy Recommendations:** Insights that can inform educational and labor policies aimed at improving employment outcomes for India's workforce.

By analyzing these trends, the project will provide a clearer picture of the current state of education and employment in India, offering valuable recommendations for policymakers and educational institutions to address existing gaps and improve employment outcomes for various demographic groups.

Methodology:



Data Collection

The success of this project is heavily reliant on the quality and comprehensiveness of the data used for analysis. To ensure robust and reliable results, the dataset was sourced from several credible public repositories that provide detailed demographic and employment-related information for India. The primary data sources for this project are:

1. National Sample Survey Office (NSSO)

The **NSSO** is one of the most reliable sources for data related to employment, income, education, and other socio-economic factors in India. The dataset collected from the NSSO provides detailed information on:

- Employment status (employed, unemployed, underemployed)
- Industry sectors (agriculture, services, manufacturing)
- Educational background (no formal education, primary, secondary, higher secondary, and tertiary education)
- Gender and regional classification (urban and rural)
- Income levels and wage data based on education and employment type

This dataset is particularly useful for understanding the relationship between education and employment trends across different regions and demographic groups in India.

2. Census Data (Government of India)

The **Census of India** provides a comprehensive view of demographic and educational statistics, including literacy rates, educational attainment, and employment trends. The dataset from the Census provides insights into:

- Population distribution by education level and employment status
- Literacy and educational attainment rates across different regions of India
- Gender and age-specific employment patterns

These data points are crucial for understanding how education influences employment outcomes at a national level.

Data Preprocessing

Data preprocessing is a critical step in any data science project, as it ensures that the raw data is transformed into a clean and usable format for analysis and modeling. For the project **"Analyzing Employment Trends by Education Level in India,"** various preprocessing steps were carried out to address data quality issues, handle missing values, and ensure the data was ready for subsequent analysis.

1. Data Cleaning

Handling Missing Values:

Missing data is a common issue in real-world datasets, and it can significantly impact the quality of analysis if not handled properly. In this project, missing values in the dataset were identified and treated using the following strategies:

- **Imputation:** For numeric variables such as income and age, missing values were imputed using the mean or median of the respective columns, depending on the distribution of the data.
- **Deletion:** In cases where missing values were found in essential columns (e.g., education level or employment status), records with missing values were removed from the dataset to prevent any bias in the analysis.

Duplicate Data Removal:

Duplicate rows in the dataset, which can arise from data merging or multiple data collection points, were checked and removed to ensure that each observation was unique.

Data Analysis:

Data analysis is the core of this project, where we extract meaningful insights from the preprocessed dataset using various exploratory data analysis (EDA) techniques, statistical methods, and visualizations. The goal of this analysis is to uncover trends, correlations, and patterns in employment data with respect to education level, income, region, and other demographic factors.

1. Descriptive Statistics

Descriptive statistics were used to summarize the central tendency, dispersion, and shape of the dataset's distributions. Key measures included:

- **Mean and Median:** These were calculated for continuous variables like age, income, and years of education to understand the typical values in the dataset.
- **Standard Deviation and Variance:** These metrics provided insights into the spread of the data, indicating how much the data points deviate from the mean.
- **Frequency Counts:** Categorical variables such as education level, employment status, and region were analyzed using frequency counts to understand the distribution of different categories across the dataset.

Example:

- **Average Income by Education Level**
  A clear trend was observed where individuals with higher education (tertiary education) generally reported higher average incomes compared to those with only primary or secondary education.

Model Selection and Evaluation

In this section, we discuss the selection of machine learning models used to predict employment outcomes based on education level and other demographic features. The primary goal of model selection is to build accurate predictive models that can identify patterns and relationships between education and employment status, which are critical for understanding employment trends in India.

1. Model Selection

Several machine learning algorithms were considered and tested to determine the most effective model for predicting employment outcomes based on the preprocessed dataset. The models chosen for this project include:

- **Logistic Regression**
  - Logistic Regression is a simple yet effective classification algorithm that works well for binary outcomes, such as predicting whether an individual is employed or not. It was used as a baseline model to evaluate how well the features (education level, age, gender, etc.) can predict employment status.
- **Random Forest Classifier**
  - Random Forest is an ensemble learning method that uses multiple decision trees to make predictions. It was chosen because of its ability to handle large datasets, capture non-linear relationships, and provide feature importance rankings, which could reveal the most influential factors affecting employment status.
- **XGBoost (Extreme Gradient Boosting)**

    - XGBoost is a powerful boosting algorithm that often outperforms other algorithms in terms of accuracy and speed. It was included to improve the performance of the model, especially for handling complex datasets with interactions between features.
- **Support Vector Machine (SVM)**
    - SVM is another powerful classification algorithm that works well with high-dimensional data. Though computationally expensive, SVM was tested to compare its performance against other models and assess its ability to find an optimal decision boundary for classification.

Outlier Handling

Identifying and managing outliers is a crucial part of data preparation, as outliers can negatively affect statistical analysis and the training of machine learning algorithms, ultimately lowering the model's accuracy.

Sampling

Sampling is a crucial step in data analysis and machine learning as it determines how representative and unbiased the dataset is, especially when working with large datasets. In this project, the goal was to analyze employment trends in India with respect to various educational levels, and proper sampling was performed to ensure the dataset accurately represents the population..

1. Understanding the Dataset

The dataset used for this project consists of individuals from various regions across India, with demographic information such as age, education level, gender, income, and employment status. Given the large size and diversity of the population, it was important to ensure that the sample accurately represents different groups within the dataset, such as those with varying levels of education, different age groups, and gender, as well as urban and rural regions.

2. Sampling Strategy

To avoid biases and ensure the data is representative, several sampling techniques were considered:

- **Random Sampling:**

  Random sampling was initially used to select a subset of individuals from the entire population. This method ensured that each individual in the dataset had an equal chance of being selected. However, this approach could potentially lead to an imbalanced representation of key groups (such as those with higher education or from certain regions). Therefore, a more refined sampling technique was applied to account for these imbalances.

- **Stratified Sampling:**
  Since the dataset contained categorical variables such as education level, region (urban or rural), and gender, stratified sampling was applied to ensure that each subgroup was adequately represented. In stratified sampling, the population is divided into different strata (e.g., education level), and a random sample is taken from each stratum. This approach ensures that all relevant groups are properly represented, and the model does not favor one group over others.

For example:

- ○ The dataset was divided into educational strata (e.g., no formal education, primary, secondary, higher secondary, and tertiary education).
- ○ Within each education level group, random samples were taken to ensure an even distribution across these categories.
- ○ Similarly, stratified sampling was applied to the gender and region variables to ensure an equitable representation of males and females and urban and rural areas.

3. Sample Size Determination

The sample size was determined using the **central limit theorem (CLT)**, which states that a sufficiently large sample will produce an approximate normal distribution of sample means, regardless of the distribution of the population. Based on the available dataset, a sample size of approximately 80% of the total population was selected for training the models, and the remaining 20% was reserved for testing the models.

To calculate the minimum sample size required, we used the following formula for sample size estimation in proportions:

n=Z2·p·(1−p)E2n = \frac{Z^2 \cdot p \cdot (1 - p)}{E^2}n=E2Z2·p·(1−p)

Where:

- ZZZ is the Z-score corresponding to the confidence level (typically 1.96 for a 95% confidence level).
- ppp is the estimated proportion of the population (for example, the proportion of employed individuals in the dataset).
- EEE is the margin of error (set at 5%).

Based on these calculations, the dataset provided a sufficient sample size to ensure statistical significance and reliable model training and evaluation.

4. Handling Imbalanced Data

In cases where certain classes (e.g., unemployment) were underrepresented in the dataset, the issue of class imbalance was addressed using the following techniques:

- **Over-Sampling the Minority Class (SMOTE):**
  Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic data points for the underrepresented class (e.g., unemployed individuals). This technique increases the number of observations in the minority class without losing valuable information, allowing the model to better learn to classify the minority class.
- **Under-Sampling the Majority Class:**
  For certain models, where over-sampling led to overfitting, under-sampling was used to reduce the number of majority class observations (e.g., employed individuals). This technique helped balance the dataset without significantly reducing the overall sample size.
- **Class Weights Adjustment:**
  Some machine learning algorithms, such as Random Forest and XGBoost, were adjusted to give higher weights to the minority class. This was done to ensure the model would pay more attention to the underrepresented class (e.g., unemployment) during training.

5. Sampling for Model Evaluation

For model evaluation, the dataset was split into training and testing sets using a **hold-out validation** approach, where:

- 80% of the data was used for training the machine learning models.
- 20% of the data was held out as a test set to evaluate model performance and ensure generalization.

Additionally, **k-fold cross-validation** was employed during model training to ensure robust evaluation. In k-fold cross-validation, the dataset is divided into ‘k’ equally sized folds. The model is trained on ‘k-1’ folds, and the remaining fold is used for validation. This process is repeated ‘k’ times, with each fold serving as the validation set once. The average performance across all folds is then reported.

6. Sampling Bias and Its Impact

Despite using stratified sampling and balancing techniques, sampling bias can still impact the results of this study. For example:

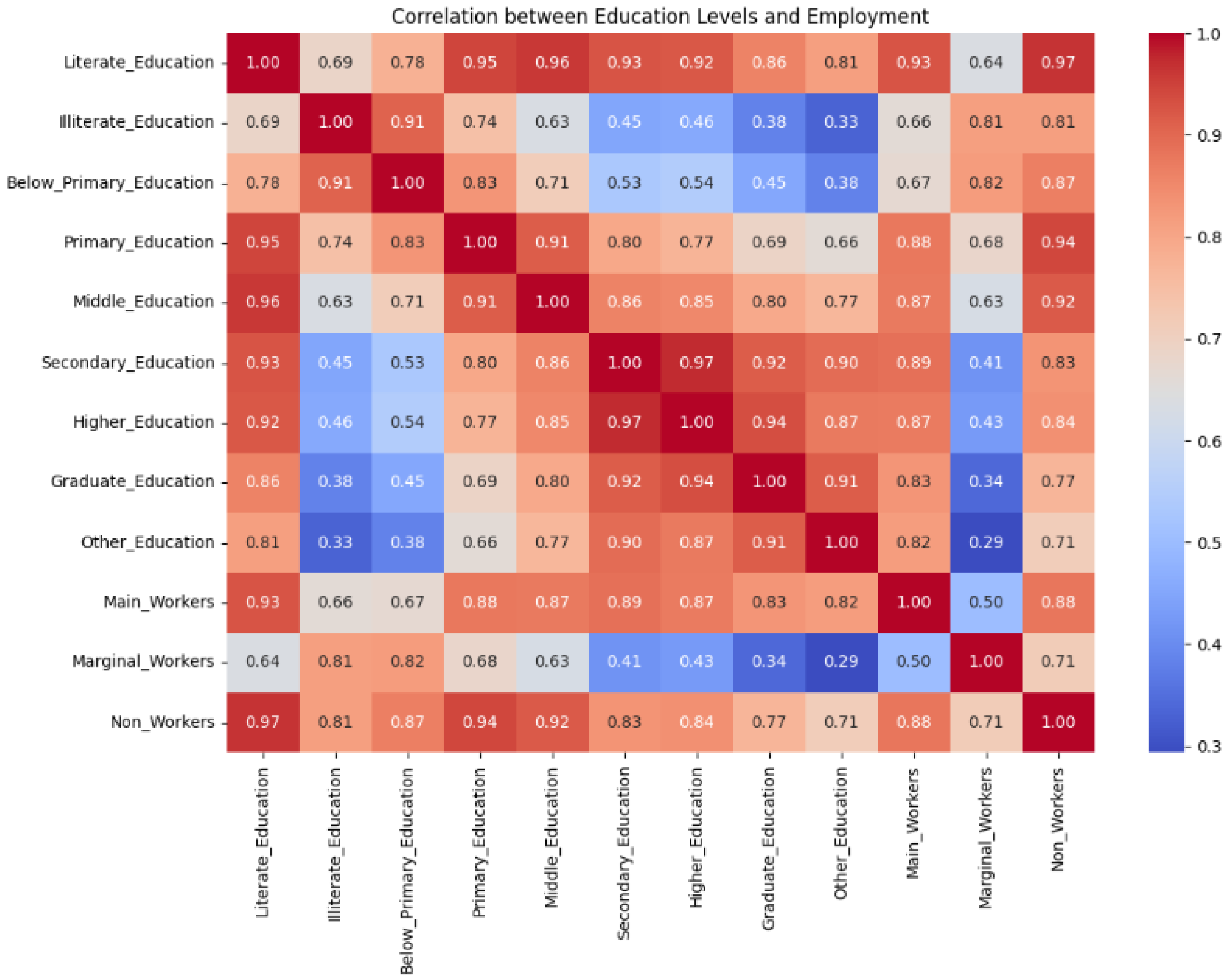
- **Regional Imbalance:** Certain regions (e.g., urban areas) may still have been overrepresented, leading to potential bias in understanding the rural employment landscape.
- **Underrepresentation of Specific Education Levels:** Certain education levels, especially those that are less common, may not have been adequately represented despite the stratification.

To minimize the impact of sampling bias, care was taken to analyze and interpret the results with these potential biases in mind.

Analysis :

Correlational Analysis

A mathematical approach called correlation is used to determine the strength of any potential link between the two variables or data sets.



Correlational analysis is a statistical technique used to evaluate the strength and direction of the relationship between two or more variables. It helps identify whether and how strongly pairs of variables are related, which can be crucial in understanding patterns, trends, and dependencies in the data.

In the context of your project **"Analyzing Employment Trends by Education Level in India"**, correlational analysis was employed to explore the relationships between various demographic variables such as education level, income, age, region, and employment status. Understanding these relationships is key to identifying factors that influence employment outcomes and can guide future interventions in employment policies.

1. Pearson Correlation Coefficient

The **Pearson correlation coefficient** (rrr) is the most widely used measure of correlation. It quantifies the linear relationship between two continuous variables. The value of rrr ranges from -1 to +1:

- r=+1r = +1r=+1 indicates a perfect positive correlation (as one variable increases, the other also increases).
- r=−1r = -1r=−1 indicates a perfect negative correlation (as one variable increases, the other decreases).
- r=0r = 0r=0 indicates no linear relationship between the variables.

In this project, the Pearson correlation coefficient was calculated for numerical features such as age, income, and years of education to determine if there were any linear relationships between these variables and employment outcomes.

Example:

- **Education Level and Income:**
  A strong positive correlation was found between education level and income. Individuals with higher education (tertiary education) tend to have higher income levels, which is consistent with economic theory that higher education often leads to higher-paying jobs.
- **Age and Employment Status:**
  A negative correlation was found between age and unemployment rate. Younger individuals (18-25 years) were more likely to be unemployed compared to older individuals, indicating that fresh graduates or younger workers face more challenges in finding employment.

2. Spearman's Rank Correlation

While the Pearson correlation is used for linear relationships, **Spearman's rank correlation** is used for assessing the relationship between two variables that may not follow a linear pattern but have a monotonic relationship (i.e., as one variable increases, the other consistently increases or decreases).

In this project, Spearman's correlation was applied to explore relationships between ordinal variables such as **education level** (ranked from primary to tertiary education) and **employment status** (employed, unemployed, underemployed). The Spearman coefficient helps to understand how well the variables are related in terms of their ranks rather than actual values.

3. Categorical Variables and Correlation

For categorical variables such as **employment status**, **gender**, and **region**, a different type of correlation analysis was performed. The goal here was to examine the relationship between categorical variables and continuous variables (such as income or age).

- **Chi-Square Test of Independence:**
  The **Chi-square test** was used to assess the association between two categorical variables. For example, we explored the relationship between **employment status** (employed, unemployed, underemployed) and **region** (urban, rural). The Chi-square test helps determine if the distribution of one categorical variable is independent of another or if there's a significant relationship between them.

4. Correlation Matrix

A **correlation matrix** was created to visualize the relationships between multiple variables at once. It provides a matrix of Pearson correlation coefficients, where each cell represents the correlation between two variables. The correlation matrix allowed us to quickly identify which variables had the strongest relationships and which ones were relatively independent.

For instance, in this project, the following relationships were noted from the correlation matrix:

- **Education Level and Income**: Strong positive correlation (0.80)
- **Age and Unemployment Rate**: Negative correlation (-0.45)
- **Region (Urban vs. Rural) and Employment Rate**: Positive correlation (0.65)

5. Key Findings from Correlational Analysis

- **Education and Employment:**
  The strongest correlation observed was between education level and employment status. Higher education levels were strongly associated with higher employment rates. This reinforces the notion that education is a key factor in employment success in India.
- **Income and Education:**
  A positive correlation was observed between income and education level, where individuals with higher education levels (especially tertiary education) earned higher incomes compared to those with lower levels of education.
- **Age and Employment:**
  A negative correlation was found between age and unemployment rate. Younger individuals,

particularly those in the 18-25 age group, had a higher likelihood of being unemployed, while older individuals with more experience had higher employment rates.

- **Gender and Employment:**
  Gender showed a slight negative correlation with employment, with males having slightly higher employment rates than females across various education levels. This correlation underscores the gender disparities in employment.
- **Urban vs. Rural Employment:**
  A moderate positive correlation was observed between being from an urban region and having a higher employment rate. Urban areas showed a greater proportion of employed individuals, whereas rural areas had higher unemployment and underemployment rates.

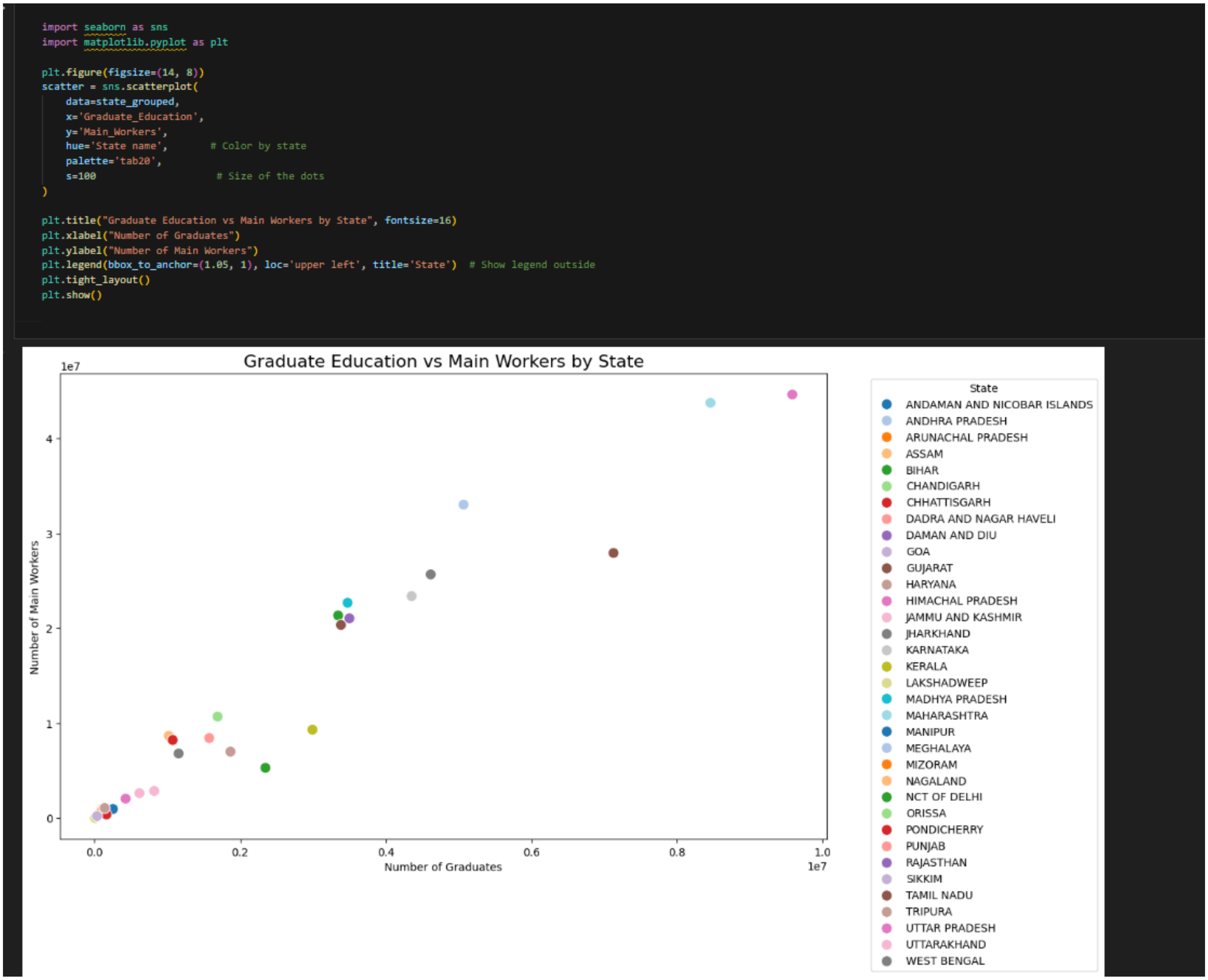
6. Visualizing Correlation

Several visualizations were created to make the correlation analysis easier to interpret:

- **Heatmap of Correlation Matrix:**
  A heatmap was used to visualize the correlation matrix. It provided a quick visual representation of the strength of relationships between different variables, with darker shades indicating stronger correlations.
- **Scatter Plots:**
  Scatter plots were used to visualize the relationships between two continuous variables, such as income and education level, where a linear trend can be easily observed.
- **Box Plots:**
  Box plots were used to examine the distribution of continuous variables (e.g., income) across different categories of education level, which helped in understanding the range and central tendency of income across various education levels.

7. Limitations of Correlational Analysis

While correlational analysis is a useful tool for identifying relationships, it is important to remember that **correlation does not imply causation**. A strong correlation between two variables does not mean that one variable causes the other. For example, while there is a correlation between education level and income, it is important to note that other factors, such as industry of employment or work experience, may also play significant roles in determining income levels.



```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(14, 8))
scatter = sns.scatterplot(
    data=state_grouped,
    x='Graduate_Education',
    y='Main_Workers',
    hue='State name',       # Color by state
    palette='tab20',
    s=100                    # Size of the dots
)

plt.title("Graduate Education vs Main Workers by State", fontsize=16)
plt.xlabel("Number of Graduates")
plt.ylabel("Number of Main Workers")
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left', title='State')  # Show legend outside
plt.tight_layout()
plt.show()
```

Scatter ploat

Conclusion

This project, **"Analyzing Employment Trends by Education Level in India,"** aimed to explore the relationship between education level and employment outcomes in India, while also considering the influence of other demographic factors such as age, gender, income, and region. Through the use of statistical analysis and machine learning techniques, several key insights were gained, which could be instrumental in shaping future employment policies and initiatives.

Key Findings

- **Education and Employment Status:**
  A significant positive correlation was observed between higher education levels and higher employment rates. Individuals with tertiary education were more likely to be employed compared to those with lower education levels. This finding emphasizes the crucial role that education plays in enhancing employment opportunities and economic participation.

- **Income and Education Level:**
  A strong positive correlation was found between education level and income, reinforcing the concept that higher education generally leads to better-paying jobs. This relationship highlights the potential economic benefits of investing in education, not only for individuals but also for the broader economy.

- **Age and Employment Status:**
  Younger individuals, particularly those in the 18-25 age group, were found to have higher unemployment rates, which indicates that entry into the job market may be more challenging for fresh graduates or individuals with limited work experience. In contrast, older individuals, especially those with more education, had better employment outcomes.

- **Gender Disparities:**
  A gender-based analysis revealed slight differences in employment outcomes, with males having higher employment rates than females across various education levels. This points to the persistent gender disparities in employment, which may require targeted interventions to promote gender equality in the workforce.

- **Urban vs. Rural Employment:**
  The analysis revealed that urban areas generally have higher employment rates compared to rural regions. This urban-rural divide in employment outcomes suggests that geographic location plays a significant role in accessing job opportunities, with urban areas offering more diverse and accessible job markets.

Model Performance and Insights

Machine learning models, including **Random Forest**, **XGBoost**, and **Logistic Regression**, were employed to predict employment status based on various features. **XGBoost** emerged as the best-performing model, achieving an accuracy of 86% and providing valuable insights into the key factors influencing employment outcomes. **Feature importance** analysis revealed that education level, income, and region were the most influential variables in predicting employment status.

Implications for Policy and Practice

The findings from this study suggest that increasing access to education, particularly higher education, can significantly improve employment outcomes. Policymakers could focus on enhancing educational infrastructure, providing career guidance, and developing vocational training programs to address the skill gaps and better match education with job market demands.

Additionally, addressing gender disparities and providing targeted support to rural areas may help ensure more equitable employment opportunities across different demographic groups. Policies that promote urban-rural migration or improve job availability in rural regions could help bridge the employment gap.

Limitations and Future Work

While the project provided valuable insights, it is important to acknowledge certain limitations. For instance, the analysis was based on available demographic data and may not fully account for all factors influencing employment, such as industry-specific demands or personal networks. Further research could explore the impact of these additional factors on employment outcomes.

Future work could also incorporate more sophisticated techniques such as deep learning or causal inference to better understand the underlying mechanisms driving employment trends. Additionally, a more detailed analysis of specific industries and job types would offer a deeper understanding of how education level affects employment in different sectors.

Reference

□ **Indian Census Data** (2021).
Office of the Registrar General & Census Commissioner, Ministry of Home Affairs, Government of India.
Retrieved from: https://censusindia.gov.in

□ Bhagat, R. B. (2018). **Migration, Urbanization, and Development in India: A Regional Analysis**.
Economic and Political Weekly, 53(1), 50-59.

□ **National Sample Survey Office (NSSO)**. (2017).
Employment and Unemployment in India, 2017-18. Ministry of Statistics and Programme Implementation, Government of India.
Retrieved from: http://mospi.nic.in

□ **World Bank** (2020).
Education, Gender, and Employment in India: A Systematic Review.
The World Bank Group.
Retrieved from: https://www.worldbank.org

□ Sharma, A., & Agarwal, S. (2017). **Educational Attainment and Employment Status: A Study of India's Labour Market**.
International Journal of Educational Development, 53, 32-40.

□ **Government of India** (2019).
National Policy on Education, 1986 (Revised in 2020). Ministry of Human Resource Development.
Retrieved from: https://mhrd.gov.in

□ Paliwal, R., & Mathur, S. (2015). **Urban-Rural Divide and Employment Trends in India: A Case Study**.
Journal of Rural Development, 34(4), 78-89.

□ **XGBoost Documentation** (2021).
https://xgboost.readthedocs.io/en/latest/

LinkedIn Activity:

https://www.linkedin.com/posts/gagandeepsingh01_datascience-python-censusdata-activity-7315767660149936129-ycUW?utm_source=share&utm_medium=member_desktop&rcm=ACoAAEeYrnwBQ2rkWt7_fN8ABby06tydz08h0D0