A Project Report

On

Data Science in Area and Population

Submitted in partial fulfillment of the

requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Session 2023-24

in

CSE(Cyber Security)

By

Name of Student: Shruti Kapoor **Name of Student: Arnab Srivastava**
Admission No.: 20SCSE1290083 **Admission No.:20SCSE1290002**

Under the guidance of

Ms. Indervati
Designation: Assistant Professor

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA

Apr, 2024

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

We hereby certify that the work which is being presented in the project, entitled **"DATA SCIENCE IN AREA AND POPULATIONS"** in partial fulfillment of the requirements for the award of the Bachelor of Technology submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of February, 2024 to April, 2024, under the supervision of Ms Indervati (Assistant Professor), Department of Computer Science and Engineering, Galgotias University, Greater Noida.

The matter presented in the project has not been submitted by us for the award of any other degree of this or any other places.

Arnab Srivastava

20SCSE1290002

Shruti Kapoor

20SCSE1290083

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Name: Ms. Indervati

Designation Assistant

Professor

CERTIFICATE

This is to certify that Project Report entitled “Data Science in Area and Population” which is submitted by Arnab Srivastava(20SCSE1290002), Shruti Kapoor(20SCSE1290083) in partial fulfillment of the requirement for the award of degree B. Tech. in Department of School of Computer Science and Engineering,

Galgotias University, Greater Noida, India is a record of the candidate own work carried out by them under my supervision. The matter embodied in this thesis is original and has not been submitted for the award of any other degree

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Program Chair **Signature of Dean**

Date: 23 April 2024

Place: Greater Noida

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Professor Indervati, Department of Computer Science & Engineering, Galgotias University, Greater Noida, India *for his constant support and guidance throughout the course of our work. Her sincerity, thoroughness and perseverance have beena constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.*

We also take the opportunity to acknowledge the contribution of Professor (Dr.) Sunita Yadav, Dean, Department of Computer Science & Engineering, Galgotias University, Greater Noida, India *for his full support and assistance during the development of the project.*

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Signature:
Name: Shruti Kapoor
Roll No.: 20131290018
Date:

Signature:
Name: Arnab Srivastava
Roll No.: 20131290002
Date :

Abstract

The distribution and composition of populations vary geographically, which has an impact on corporate development, government changes, urban development, and other areas. Even though these kinds of statistics are widely applicable and significant, it can be difficult to obtain local census estimates in a timely and accurate manner due to the dynamic nature of population counts, their political content, and logistical and administrative difficulties. Given these difficulties, the main goal is to use data science approaches to close the knowledge gap between the dynamic demographic picture and wise decision-making.

It seeks to offer thorough, current, and region-specific population insights by utilizing cutting-edge analytical methodologies. The need to support fact-based business strategies, facilitate responsive government policies, and direct successful urban development projects is what spurred this research. Data science can be helpful in identifying areas where necessities are not distributed fairly.

By performing an exploratory data analysis (EDA), it can categorize the areas that require more attention in terms of development and employment. It can be used to generate graphs and plots to track the rise and fall of different factors, which can help to identify unstable regions. This approach enables categorizing problems and identifying the people facing them.

As a result, it can save time and resources while also providing accurate information. The strategy aims to equip key players with the information they need to adjust to the shifting demographic dynamics, improving the standard of living, economic success, and social well-being for various populations.

TABLE OF CONTENTS

List of Table

List of Figures

CHAPTER-1
INTRODUCTION

1.1 Introduction

Emerging technological advancements have become the determining attribute in the consolidation of data science, machine learning, government, and business expansion. The utilization of data science in various fields has helped with the proper use of resources in the successful growth of businesses and government policies. Implementing data science in area andpopulation analysis will lend a hand to various sectors in their advancements and reshape the ways they operate. Now, this will raise the question of how data science can be exploited to aid effective governance and business expansion, explicitly in the context of development using areaand population analysis.

This issue comes with great weight as the globe brawls with numerous trials and tribulations in the execution of ideas through the collaboration of business allies with the government. So to detect the places where there's a gap in the proposed strategies implemented by the government and to have it's better results, data science appears as a potent instrument that addresses issues that lack attributes such as unemployment, residency, area growth, and poverty and helpsgovernment and businesses evolve and accomplish these objectives.

This paper's goal is to examine the complex interrelationships between data science, governance, and business expansion, with the target of examining how these applications relate to development and thus evolving the area standards leading to the betterment of society. By rendering the uses of data science, we can build a framework that not only gets associated with government operations but also expands business growth by analyzing the superficial areas of establishing firms.

The link between governance and business expansion is built on the foundation of data science.It provides decision-makers with previously unreachable insights, predictions, and solutions by acknowledging its capacity to process and analyze enormous datasets. Governments and businesses can both make informed decisions, increase productivity, and spur innovation by embracing data-driven initiatives.

Effective governance is a fundamental aspect of development. There are many challenges thatthe government has to confront, such as offering services (e.g., travel and tourism, providing jobs), maintaining infrastructure (e.g., residential issues, counts of hospitals, schools, etc.), and strengthening economic growth around the globe. Data science has the tools to track progress, identify problems, and create effective policies that promote confidence among investors and citizens by enabling openness and accountability.

Due to their function as generators of employment and economic progress, businesses also playa crucial role in development. Their interests must, however, be in line with the welfare of society as a whole. Businesses may use data science to better understand their markets, take ethical risks, and implement CSR programs that not only improve their reputation but also advance global development.

In this paper, we will look at examples and best practices from around the globe where data science has effectively closed the gap between governance and commercial growth, leading to observable gains in development indices. It intends to offer insights, suggestions, and a thoroughknowledge of how these dynamics might be utilized cooperatively to drive global growth by weaving the threads of data science, governance, and economic expansion. The way toward a society that is more just, sustainable, and prosperous begins at the nexus of these areas, and it is here that we can create a future by working together.

1.2 Formulation of Problem

In this paper, we have focused on overcoming various drawbacks. One such drawbacks involves Lack of Knowledge of Government Programs. Lack of data is one of the main obstacles that governments confront in helping recipients receive the advantages of their policies and programs. Many people are just uninformed of the programs that are available and how to use them, especially those who reside in rural areas or who are members of marginalizedgroups. Low literacy rates, limited access to information, and linguistic hurdles are only a few causes of this. As a result, many people are unable to take advantage of government programs that could enhance their quality of life. For instance, research by the Indian governmentdiscovered that 22% of recipients claimed that it took time for their applications to be reviewed or accepted, and 17% discovered it to be difficult.

It is challenging to obtain the documentation needed to receive the benefits.Another drawback that we encountered is Lack of knowledge regarding market demand. It is another issue that both enterprises and governments encounter. To decide where to sell their goods and how much to charge for them, businesses need to know which products are popularin various markets. Additionally, governments require this information to create effective programs and policies. However, it's frequently hard to find or impossible to get this information. Businesses, for instance, might not have access to information about the population, income levels, and buying power parity of various regions. Additionally, governments might not have access to real-time market trend data. As a result, companies can end up selling the incorrect products in the incorrect locations, and Governments may create plans and programs that are not in line with what the market requires. This may result in inefficiency and resource waste. Based on employment rate, PPP, and literacy rate, government programs and business ventures do not provide any benefits. Data on the employment rate, PPP,and literacy rate can be used to further understand why government programs have not been successful. These three metrics represent, respectively, income level, economic development, and human capital. The World Bank estimates that India's employment rate was 47.4% in 2022. This indicates that either the unemployed or underemployed make up more than half of the workforce in India. This presents a serious problem since it indicates that a sizable population could gain from government initiatives that offer work or educational opportunities. In 2022, India's PPP per person was $2,332. This is less than both the South Asian average of $2,702 andthe global average of $17,953. India is a nation with a low per capita income, which indicates that its people have fewer disposable cash to spend on products and services. People may find itdifficult to afford to take part in government programs that require them to make a charge or co-payment as a result. India had a 78.7% literacy rate in 2022. Although higher than the 86.6% global average, this is still lower than many other nations. People may find it challenging to comprehend and apply for government programs due to low literacy rates. This data suggests that a significant portion of the population in India is not taking advantage of government programs because they are either unemployed or underemployed, have poor incomes, or have low literacy rates or have no residential space or proper resources available.

Factors	Uses
Transportation	• Predict demand for public transportation services in differentareas. • Identify areas where new transportation infrastructure is needed. • Match people with transportationservices that are most convenient forthem.
Employment rate	• Predict future employment trends. • Identify areas with highunemployment rates. • Match people with job training programs and other employment assistance services.
Literacy rate	• Identify areas with low literacy rates. • Match people with literacy programsand other education and training services. • Personalize recommendations for educational resources based on a person's individual needs.
Purchasing power parity	• Identify areas with high and low levels of economic activity. • Match people with government services and programs that are most relevant to their income level. • Personalize recommendations for government services and programs based on a person's purchasing power parity.

Table 1.2.1 Factors and Uses: It will be used during decision making by government and corporate.

1.3 Tools and Technology used

We are using python language to deal with the different datasets and perform the different data visualizations that are meant for our requirements. We took two of the datasets targeting Unemployment and illiteracy to focus on the areas where the efforts need to be made to develop and improve the condition for the time being. We took a live dataset from Centre of Monitoring Indian Economy to see the current situation of the people. We also took dataset from Kaggle to train the model and to look if the dataset is potential enough to see the rate of employment in their different part of age. Firstly, the essential libraries of python which are required to plot graphs and charts are:

- Matplotlib- Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. Create publication quality plots. Make interactive figures that can zoom, pan, update.
- Seaborn- Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data.
- Jovian Opendataset- Jovian is a platform for sharing and collaboraring on Jupyter notebooks and data science projects. jovian-py is an open-source Python package for uploading your data science code, Jupyter notebooks, ML models, hyperparameters, metrics etc. to your Jovian account.
- Pandas- Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.
- Numpy- NumPy can be used to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays andmatrices and it supplies an enormous library of high-level mathematical functions that operateon these arrays and matrices.
- Statsmodels- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics is available for each estimator.

All these libraries play a huge role in working with the datasets with all the libraries having their different targets working for the same objective and that is data visualization. Jovian requires Collaboratory work hence its required to work on google collab whereas the EDA of poverty in India is made on vscode.

CHAPTER-2

LITERATURE REVIEW

To address the problems of market demand, government program and service ignorance,and the resulting inefficiencies and resource waste, this research paper proposes a machine learning-powered platform. In order to predict a person's eligibility for government services and programs and identify areas that are in high demand by both governments and businesses, theplatfrm would gather data on a variety of topics, including residence, mode of transportation, andeducational attainment. Machine learning would be used to train models to make these predictions. It would then use this information to provide users with recommendations in atailored manner.

The platform has the potential to have a profound impact on the lives of an enormous amount ofpeople by:

- Reducing the time and effort required to apply for services and programs offered by thegovernment
- Improving accessibility to nearby companies and services
- Increasing public knowledge of government initiatives and services
- Encouraging economic growth and job creation
- Proper utilization of resources
- Offering a market-based approach for residential space

This work is intriguing because it offers a fresh approach to a pressing issue. The platform has the ability to significantly improve the lives of people who might not be aware of the resources that are available to them by automating the process of matching people with government programs and services that they qualify for through the use of machine learning. Giving information about the best times and intervals between which to provide transportation services in a certain area: more services will be offered if the area is densely populated and in need of public transportation than in areas where less services are needed.

The platform may also be used to assist companies in finding new markets and creating goods and services that are customized to meet the demands of particular localities. The platform has the potential to enhance firms' operational efficiency and effectiveness by offering them insights into market demand.

Overall, this research study offers a promising new strategy for tackling issues related to government program and service ignorance, market demand, and the resulting inefficiencies and resource waste. Both individuals and businesses could experience major improvements in their lives as a result of the platform.

This study attempts to identify the benefits of data science (DS) for businesses and government schemes by visualizing various factors such as employment status, literacy rate, transportation services intervals, etc. using graph and prediction. It also outlines the potential and challenges associated with establishing this competence.

A combination of people, technology, shared information (language, rules, regulations), and organizations (internal and external) linked by value propositions makes up all service systems [1]. Since all of these fields are concerned with comprehending, developing, enhancing, and expanding service systems, it is imperative that we recognize the significance of technology and make a clear distinction between its broad spectrum.

One of the subjects with the most business importance is corporate governance. Professional investors place a high value on good corporate governance; McKinsey reports that these investors are willing to pay a premium for investments in businesses that uphold high standards of governance [2].

The principles of corporate governance were released by the Organization for Economic Cooperation and Development (OECD) in 1999. They describe corporate governance as the framework for setting organizational goals and keeping track of performance to make sure those goals are met [3]. Key resources included in the enterprise's corporate governance are required tocarry out all business-related tasks. [4]. Under the leadership of top executives, these vital resources together enable the company's business operations and activities to be carried out in a way that meets its goals.Businesses that use DS and BD technology can increase productivity by 5–6%. A more recent study indicated that investments in BD and DS are linked to a 3–7% increase in corporate productivity. The researchers examined a data set of 814 organizations from 2008 to 2014. Furthermore, examining expert sports organizations compared the anticipated yearly growth of the sector (3%) compared to the growth of companies who used analytics (7.2%) in the year after adoption. firms that announce the implementation of BA have more positive stock market reactions than other firms, according to the findings of a research conducted by Teo et al. (2016).[6–5]. According to Carillo (2017), data scientists are not the onlyones who should be developing their analytical abilities. According to the author, managers need to develop into management scientists with a multidisciplinary skill set that includes business management, analytical and modeling tools and methodologies, knowledge, and data management abilities. A rising number of companies are creating the post of chief data officer, which positions a new generation of leaders to investigate data value, since business strategies depend more and more on data (Lee et al., 2014).[7]. Our goal with the content analysis was to establish a connection between the research insights found in the literature study and the information collected, together with the categories and situations found in the qualitative research. The combination of theoretical components and practical data produced the coding system. Procedures for open, axial, and selective coding were used.[8] Furthermore, new specialized positions in data science, modeling, analysis, translation, management, and governance—such as chief data officer, data scientist, and management scientist—aredeveloping in the organizational field. As a result, businesses must be prepared to capture and harnessing this workforce, as well as defining training strategies. [9]

Nesta (2013b) claims that these adaptive technologies have the ability to create "intelligent online platforms" and "digital tutors" that can respond to students and use learner data to become intelligent enough to predict students' progress and then provide the necessary support and evaluation. [10]Additionally, the Innovation Unit supports learning analytics software that may automatically and algorithmically create "playlists" of customized pedagogies for every single student.[11].

Policy innovation laboratories in education employ a variety of digital technologies and data analytics techniques that represent new versions of "public policy instruments," as defined by political sociologists Lascoumes and le Gales (2007). Specific policy concepts and objectives can be operationalized and realized via the use of technological tools and material procedures known as policy instruments.[12] However, policy instruments are more than just impartial tools. Rather, "all instruments constituting "specific effects" that organize public policy in accordance with its own logic are produced by "a condensed form of knowledge about social control and ways of exercising it" (Lascoumes and le Gales 2007, p. 3).

Particular tools, like the adaptive learning analytics platforms they support or Nesta's evaluative index, are digital policy instruments that are loaded with presumptions and condensed knowledge about the social reality of education that they are intended to measure. These tools also have specificconsequences on the structure and shape of judgments made about how to alter that reality.

Since then, the term "Data Science" has been embraced globally to denote an interdisciplinary discipline. However, there have been other discussions centered on different aspects of data science in an attempt to define how it differs from statistics or how to standardize it [13]. To provide just one example, in 1997 [14], C. F. Jeff Wu stated in his inauguration speech for the H. C. Carver Professorship in Statistics at the University of Michigan that statistics should be renamed Data Science and Statisticians Data Scientists. The new modern methodologies, however, are poolingthe two disciplines of statistics and computer science as in the interaction of computational algorithms with cognitive science in artificial intelligence and the viewpoint of machine learning as a marriage of statistics and knowledge representation [15].

The demand for Data Scientists has led Universities to rush to offer new curricula aimed at training Data Scientists. Most of the courses offered are at post graduate level, although undergraduate courses are also emerging. Universities are racing to create new curriculum geared at teaching Data Scientists because of the need for these professionals. Although undergraduate courses are starting to appear, the majority of courses offered are at the postgraduate level. Although they differ in focus and substance, they all have some pertinent characteristics, such as strong theoretical backgrounds in computer science and statistics together with real-world application experience. Whether transdisciplinary inside the institution or with corporate and industrial partners, there has to be a strong two-way interchange between academics and applications [16]. By incorporating the supervision of postgraduate projects and placements as part of Data Science training, universities may effectively foster a vital relationship between business and industry. In today's world, data scientists are highly valued and sought after. The connection between data science and statistics has been hotly debated [17]. Data science relies heavily on statistics, thus it's important to think about how statisticians shouldhandle that responsibility. Here, data science is being used by a small- to medium-sized enterprise (SME) software company to optimize the insights gleaned from their extensive collection of fascinating data on rent balances, property maintenance, and vacant or void properties [18].

Textbooks are a little slower to catch up, but there are several great books—like James et al.

and the more comprehensive Data Science Handbook—that provide statistical techniques in a Data Science framework.

We have high hopes for the future, both for the continuous advancement of statistical applications in data science and the growing significance of data science as a business strategy. The significant financial rewards guarantee that research is adequately funded, and new conceptsand techniques are always being developed.[19]. According to Chang (2014), BIaaS is a Software as a Service (SaaS) approach to business intelligence in which computation takes place in a private cloud and applications are provided as a service. Applications started moving to the ud, which gave rise to BIaaS; nonetheless, it has taken some time for BIaaS to gain traction. The main reason that organizations use BIaaS is to reduce the time it takes to deliver analytics. The main motivations for enterprises to switch to BIaaS include accelerating time to value and reducing infrastructure and installation costs (Chang, 2014).[20–21] Cloud computing is the foundation of services. D computing offers both people and companies shared processing resources as a service. Cloud computing is utilized by services, and Big Data is a topic of research because of MapReduce, a framework for efficiently processing Big Data (Chang & Wills, 2016).Organizational Sustainability Modeling (OSM) was utilized by Chang and Willis (2016) to compare the performance of cloud and non-cloud platforms. They suggested that whilecloud solutions offer the most constant performance, non-cloud solutions could be better for protecting sensitive information but might not be the most reliable. According to Newman et al. (2016), corporate performance may be improved by applying the corporate Data Science (BDS) strategy. They concentrate on introducing their system architecture and experimental techniquetoenable companies to consider implementing their idea because BDS lacks clear rules.[22].

CHAPTER-3

WORKING OF PROJECT

3.1 Implementation of Poverty Factor

As the first step, we imported all the libraries which were essential or required for visualization.

```
import numpy as np # linear algebra
import pandas as pd # data processing


import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# Any results you write to the current directory are saved as output.

import pandas as pd, numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

Figure 3.1.1 Libraries for Visualization

Now, we will merge the data set in a single data frame to have meaningful data.

```
df1 = pd.read_csv("/kaggle/input/predicting-poverty/train_values_wJZrCmI.csv")
df2 = pd.read_csv("/kaggle/input/predicting-poverty/train_labels.csv")

df = df1.merge(df2, on='row_id')
```

Figure 3.1.2 Merging dataset.

After this work, we try to get a clear information of the dataset with the help of different functions that specify the nature and properties of it (.Head(), .isNull(), .dtypes(), etc). We then perform "**age_grouping"** hat takes a DataFrame as input and categorizes individuals into age groups based on their age. his categorization is done using specified age conditions and age bins. After calling the function with a DataFrame, a new column, 'age_group', is added to theDataFrame to store the age group for everyone, which can be used for further data analysis or reporting.

```python
def age_grouping(data):
    age_condition = [
    (data['age'] < 30 ),
    (data['age'] >= 30) & (data['age'] < 45),
    (data['age'] >= 45) & (data['age'] < 60),
    (data['age'] >= 60)
    ]
    age_bins = ['< 30', '30 to 44', '45 to 60', '> 60']
    data['age_group'] = np.select(age_condition, age_bins)

age_grouping(df)
```

Figure 3.1.3 Age Grouping Analysis

Then, a function, **count_unique**, that counts and prints the unique values and their counts for a list of specified columns in a DataFrame. This function can be used to gain insights into the distribution of categorical data in the dataset. After calling the function with the DataFrame and the list of columns, the unique value counts for each specified column are printed to the console.

```python
def count_unique(df, cols):
    for col in cols:
        print(df[col].value_counts())

categ_cols = ['age_group','country','is_urban','female','married','religion','relationship_to_hh_head',
 'education_level','literacy','can_add','can_divide','can_calc_percents','can_calc_compounding',
 'employed_last_year','employment_category_last_year','employment_type_last_year',
 'income_ag_livestock_last_year','income_friends_family_last_year','income_government_last_year',
 'income_own_business_last_year','income_private_sector_last_year','income_public_sector_last_year',
 'borrowing_recency','formal_savings','informal_savings','cash_property_savings',
 'has_insurance','has_investment','borrowed_for_emergency_last_year','borrowed_for_daily_expenses_last_year',
 'borrowed_for_home_or_biz_last_year','phone_technology','can_call','can_text','can_use_internet',
 'can_make_transaction','phone_ownership','advanced_phone_use','reg_bank_acct',
 'reg_mm_acct','reg_formal_nbfi_account','financially_included','active_bank_user',
 'active_mm_user','active_formal_nbfi_user','active_informal_nbfi_user','nonreg_active_mm_user', 'share_hh_income_provided',
'num_times_borrowed_last_year','num_shocks_last_year','num_formal_institutions_last_year',
            'num_informal_institutions_last_year']

count_unique(df, categ_cols)
```

Figure 3.1.4 Distributing categorial Data in Dataset

After that, we perform data imputation for missing values in the columns 'education_level' and 'share_hh_income_provided' in the DataFrame **df**. Data imputation is a common datapreprocessing step to handle missing values in a dataset.

```python
for column in df[['education_level', 'share_hh_income_provided']]:
    mode = df[column].mode()
```

Figure3.1.5 Data Imputation for missing values

After these steps we create a function, **df_converted**, that takes a DataFrame, a source data type (**convert_from**), and a target data type (**convert_to**). It then selects columns in the DataFrame with the source data type and converts them to the target data type, updating the DataFrame in place. This is useful when you need to change the data type of specific columns in a DataFrame for further analysis or processing.

```
def df_converted(df, convert_from, convert_to):
    cols = df.select_dtypes(include=[convert_from]).columns
    for col in cols:
        df[col] = df[col].values.astype(convert_to)
    return df

df = df_converted(df, np.bool, np.int64)
```

Figure3.1.6 Converting data into target data type

Then for classification we use the code calculates a binary value (1 or 0) for each individual in the Data Frame **df** based on the 'poverty probability' column. If the 'poverty probability' is greater than or equal to 0.5, the individual is considered to be in poverty (assigned a value of 1), and if it's less than 0.5, the individual is not considered to be in poverty (assigned a value of 0). This isa common technique for converting probabilities or scores into binary classification results, such as in the case of determining whether an individual is in poverty or not.

```
df['Poverty_conditional'] = [1 if poverty_probability>=0.5
                             else 0 for poverty_probability in df['poverty_probability']]
```

Figure 3.1.7 Binary classification of poverty condition

After all these steps, we will perform plotting and graphing to assume the future of poverty or employment in India.

These plots can be used to study and have a discussion to see the future problems which may be corrected in the present time so that later the results could vary or improve from today.

3.2 Implementation of Unemployment Factor

The dataset URL to download is defined in the first line:

```
dataset_url = 'https://www.kaggle.com/gokulrajkmv/unemployment-in-india'
```

Next, the opendatasets library is imported:

```
import opendatasets as od
```

The dataset is downloaded from Kaggle using the od.download() function:

The contents of the directory where the dataset was downloaded may be listed using the os.listdir()function:

```
od.download(dataset_url)
```

The contents of the directory where the dataset was downloaded may be listed using the os.listdir()function:

```
data_dir =
'./unemployment-
in-india'import os
```

It is defined that the variable project_name:

```
project_name = "analysis-of-unemployment-india"
```

Importing the Jovian Library

```
import jovian
```

To commit dataset to a Jovian project, we use the jovian.commit() function.

```
jovian.commit(project=project_name)
```

Importing the pandas library

```
import pandas as pd
```

A Pandas DataFrame is loaded with the dataset using the pd.read_csv() function:

```
unemployment_df = pd.read_csv('/content/unemployment-in-india/Unemployment in India.csv')
```

A terminal printout of the unemployment_df DataFrame appears:

```
unemployment_df
```

3.3 Effects of Poverty and Unemployment as Factors on Government Policies andBusiness Expansions

Two of the biggest issues facing governments and businesses worldwide are poverty and unemployment. Thesetwo elements may significantly affect public policy and corporate growth in a number of ways, including:

Policies of the government:

- Government spending: In order to reduce unemployment and poverty, governments may need to spend more money on social programs and other initiatives. Higher debt levels and budget deficits may result from this.

- Taxation: In order to generate more money for social programs and other initiatives, governments may need to modify their tax laws. The competitiveness of businesses may suffer as a result.

- Regulation: To address the underlying causes of poverty and unemployment, governments may enact new rules or amend those that already exist. This can increase operating expenses and make it harder for companies to grow.

Business growth:

- Decreased demand: The need for products and services may decline as a result of poverty and unemployment. Businesses may find it more challenging to grow and add new employees as a result.

- Cost increases: Because of unemployment and poverty, businesses may have to pay more for security, insurance, and training, among other things. Additionally, this may make it more challenging forcompanies to grow.

- Scarcity of skilled labor: Unemployment and poverty may contribute to a lack of trained labor. Because of this, it might be challenging for companies to hire the skilled labor they need to grow.

How gender disparities, population, employment rate, literacy rate, unemployment, and poverty may be used toinform company development and government policy:

Data on gender disparities, population, employment rate, literacy rate, poverty, and unemployment may be usedby businesses and governments to guide their policy decisions. As an illustration:
Governments:

- Governments may utilize data to pinpoint social programs and interventions to the areas and persons most impacted by unemployment and poverty. More effective targeting of social programs and interventions is possible with the use of this information.

- Create evidence-based policies: Data may be used by governments to assess the efficacy of various initiatives and policies. In the future, new evidence-based policies can be created using this knowledge.

- Encourage inclusive economic growth: Data may be used by governments to pinpoint and remove obstacles that keep some groups of people—like women and minorities—from fully engaging in the economy.

Businesses:

- Find unmet needs in communities and areas: Companies may utilize data to find unmet requirements in communities and populations. Developing new goods and services or entering new markets can be accomplished with the use of this knowledge.

- Boost operational effectiveness: Companies may utilize data to pinpoint areas in which they can increasetheir operational effectiveness. Profits might increase and expenses can decrease as a result.

- Improved decision-making around investments: Organizations may leverage data to make better- informed investment choices. For instance, they can utilize data to pinpoint regions with high demandfor products and services or those with a labor shortage.

Here are some instances of how corporations and governments use statistics on gender disparities, population,employment rate, literacy rate, poverty, and unemployment:

- Data is being used by the Indian government to focus its social initiatives and actions. For instance, low- income households are the focus of the government's Pradhan Mantri Jan Dhan Yojana (PMJDY) financial inclusion program.

- Data are being used by the World Bank to assess the success of its initiatives aimed at reducing poverty. For instance, the World Bank's Impact Evaluation of the Rural Roads initiative in India discovered that household incomes and poverty reduction were positively impacted by the initiative.

- Data is being used by The Coca-Cola Company to find new business potential in emerging nations. For instance, the business uses data to pinpoint locations with strong bottled water demand.

- Walmart Corporation is leveraging data to increase the effectiveness of its operations in underdeveloped nations. For instance, the business uses data to pinpoint places where inventory expenses might be cut.

Two of the biggest issues facing governments and businesses worldwide are poverty and unemployment.Nonetheless, data may be utilized to guide choices and policies in ways that support inclusive economic growth, lower unemployment, and lessen poverty.

CHAPTER-4

RESULT

4.1 Result using Poverty Factor

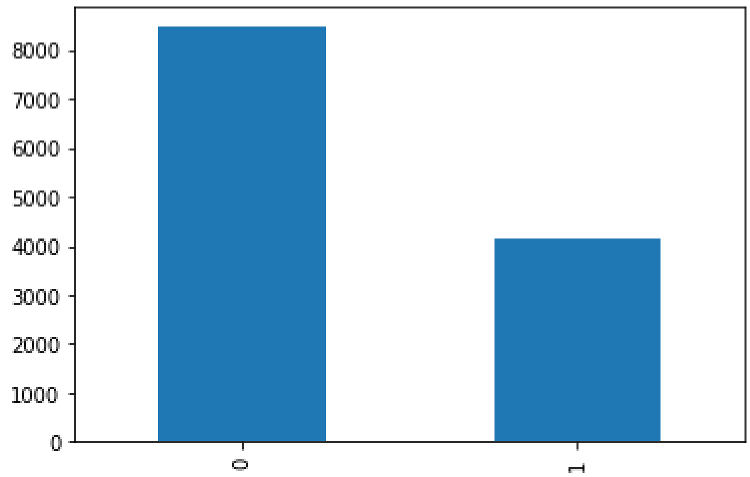


Figure-4.1.1: Shedding light on the proportion of urban and non-urban observations in the dataset.

By examine this graph, It is helpful for examining and displaying how a categorical variable ('is_urban') is distributed throughout the dataset, providing information on the relative frequencies of various urbanization levels.

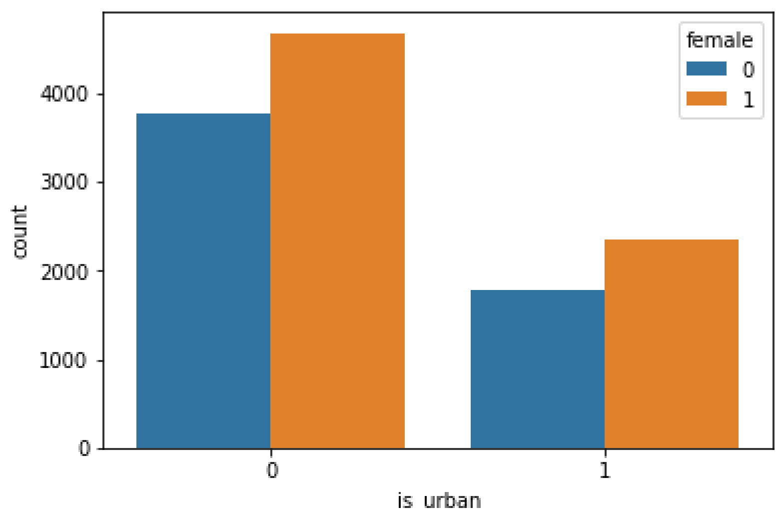


Figure-4.1.2: Differences in gender distribution between urban and rural region

By examining the gender distribution throughout the various urbanization levels is made easier to see with the use of this count graphic. It sheds light on the differences in gender distribution between urban and rural regions as bar will be included in the count plot for every distinct value of ‘is_urban” such as 0 and 1. There will be sub-bars within each bar that correspond to the two genders (o for male and 1 for female). The number of observations for a given combination of ‘is_urban’ and ‘female’ values is shown by the height of each sub-bar.

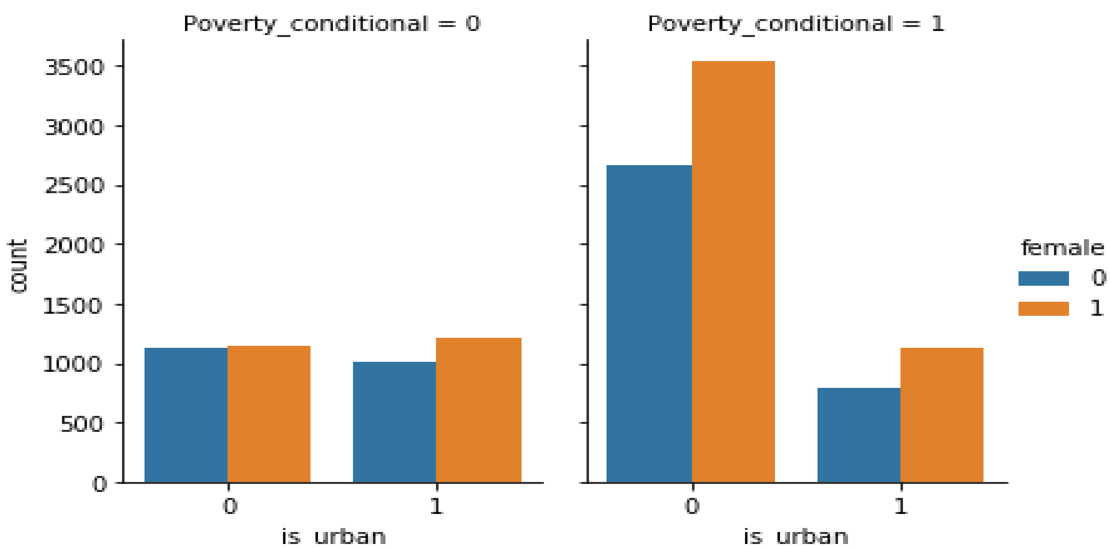


Figure-4.1.3: The distribution of 'female' and 'is_urban' changes with various poverty-related circumstances.

By examine with the conditional state associated with poverty taken into consideration, this catplot enables avisual examination of the gender distribution across various urbanization levels.

It becomes easier to see how the distribution of 'female' and 'is_urban' changes with various poverty-related circumstances.

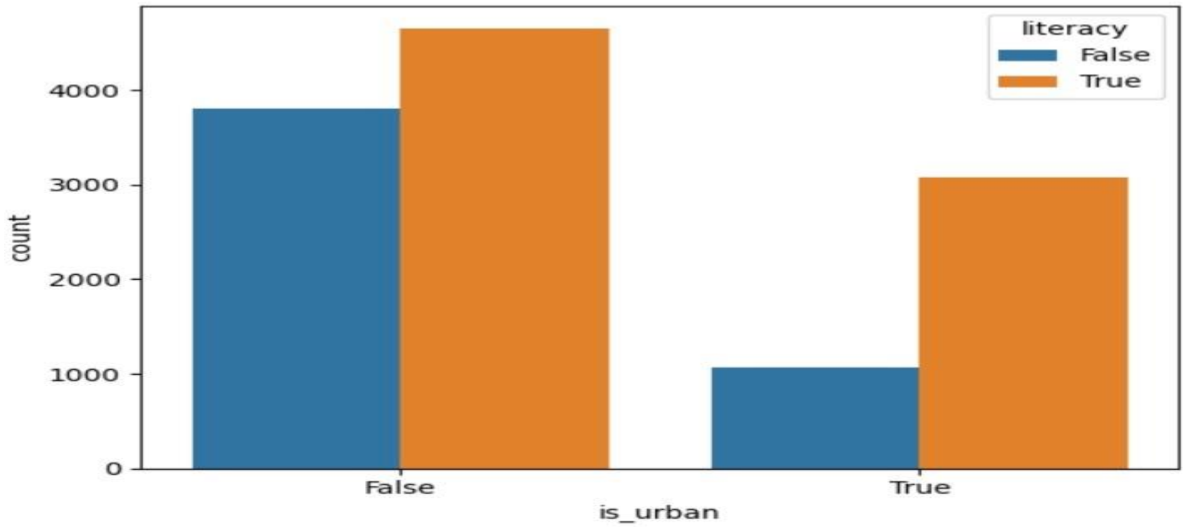


Figure-4.1.4: Distribution of literacy levels among the urban and non-urban groups

By observing this count plot, the distribution of literacy levels among the urban and non-urban groups may be visually evaluated. It enables you to find any patterns or trends by comparing the literacy levels of various groups. You may check, for instance, if the distribution of literacy levels varies across the two groups, or whether literacy levels are greater or lower in urban regions when compared to non-urban areas.

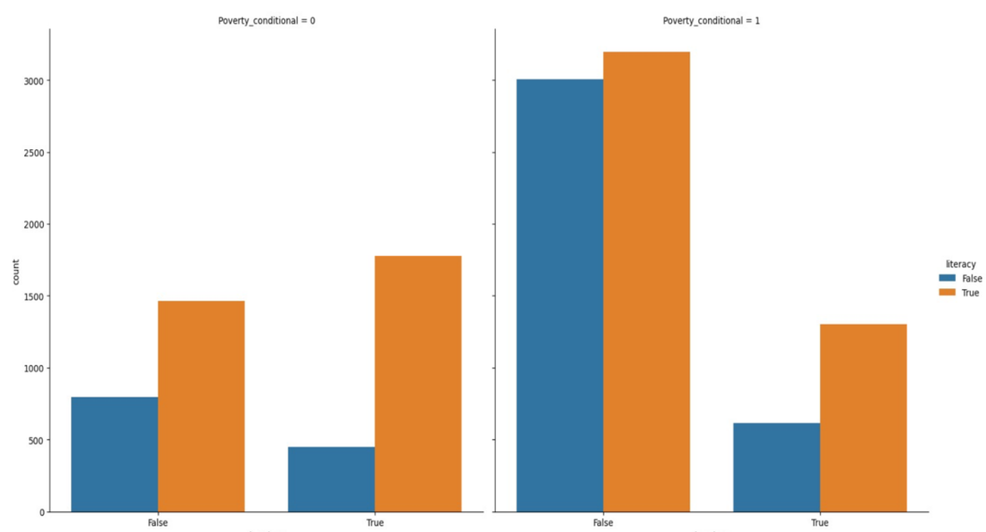


Figure- 4.1.5 Difference in literacy distribution with varying degree of poverty

By examining this plot, you may learn more about the differences in literacy distribution across urban and non-urban areas with varying degrees of poverty. Additionally, you may watch to see whether the correlations between these variables exhibit any prominent trends or patterns.

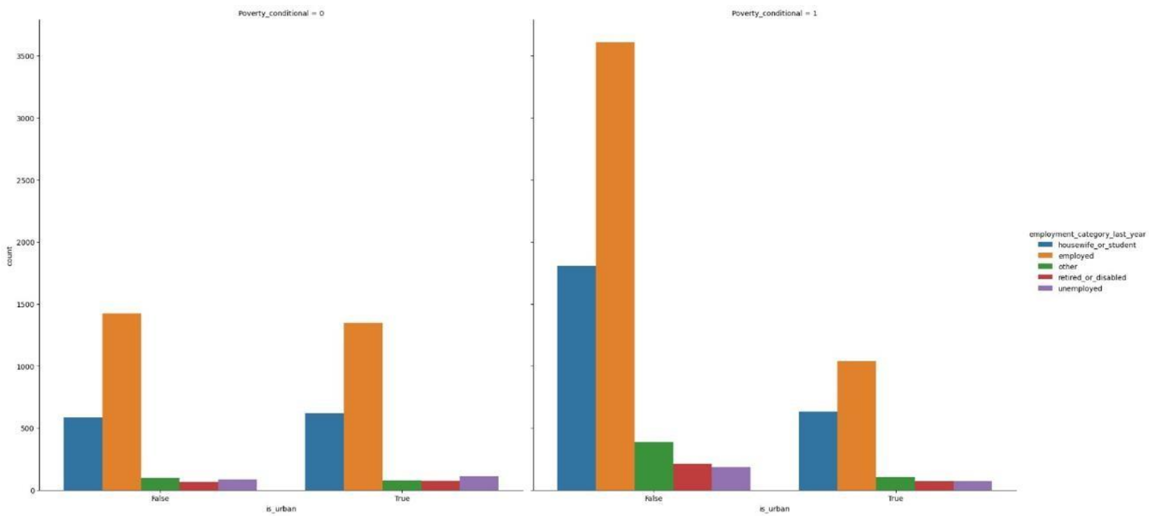


Figure-4.1.6: Difference in employment with varying degree of poverty

By examining this plot, you may learn more about the differences in employment category distribution between urban and non-urban areas at varying degrees of poverty.

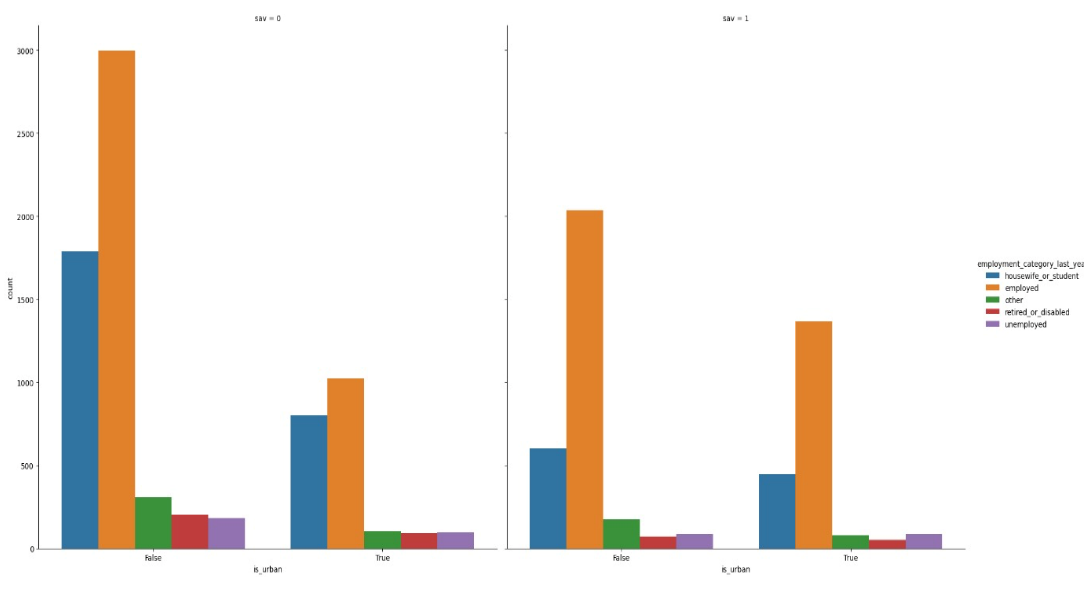


Figure-4.1.7: Analysis of how occupation categories are distributed

By examining this plot, the count of observations based on the binary savings indicator ("sav"), employment categories ("employment_category_last_year"), and urbanization status ("is_urban") are shown. It facilitates the investigation of how occupation categories are distributed throughout various savings behavior groups in urban and non-urban settings.

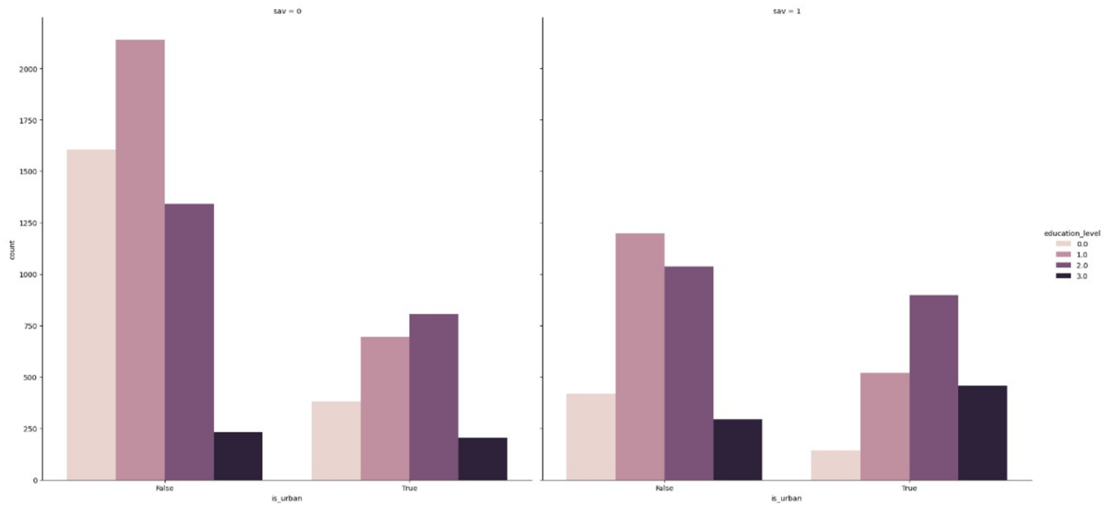


Figure-4.1.8: Different savings behavior groups' distributions of education levels.

By examining this plot, you may learn more about how different savings behavior groups' distributions of education levels change across urban and non-urban locations. The distribution of education levels is shown for every combination of "is_urban" and "sav."

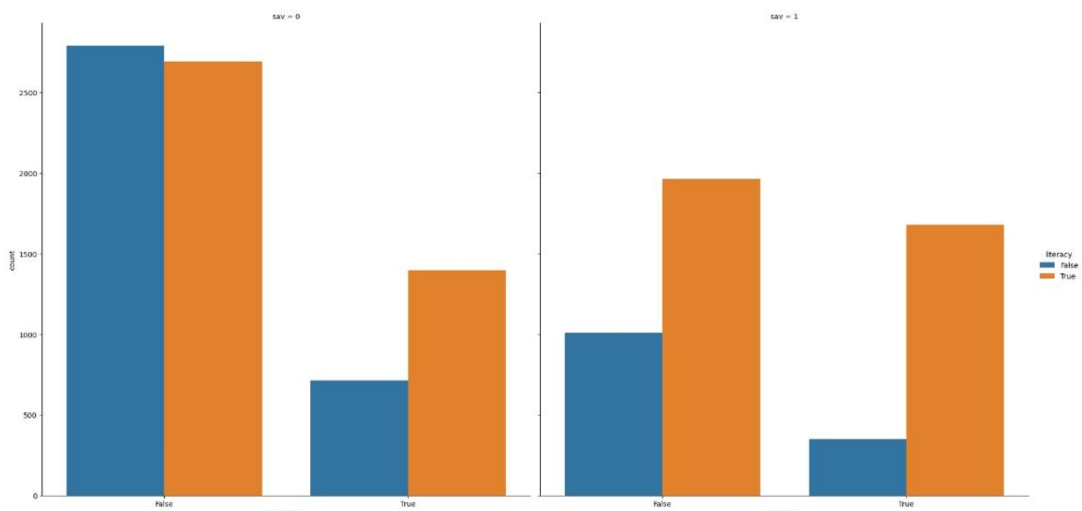


Figure-4.1.9: Differences in literacy distribution among various saving behavior

By examining this plot, you may learn more about the differences in literacy distribution between urban and non-urban locations among various savings behavior categories. You can see the distribution of literacy levels for each combination of "is_urban" and "sav."

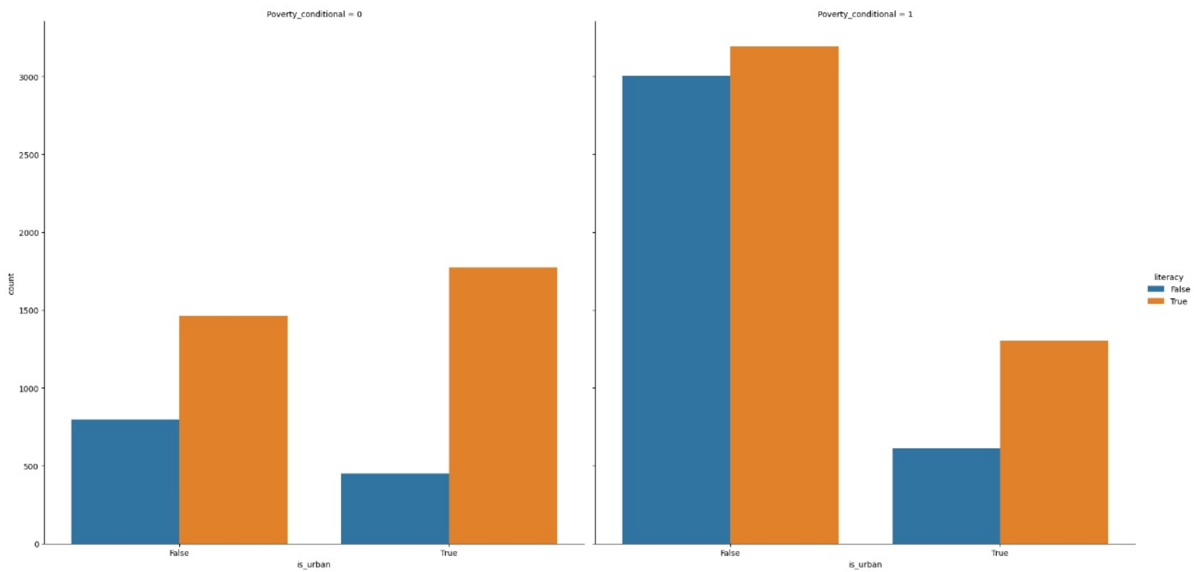


Figure- 4.1.10: Varied degree of poverty within urban and non-urban areas

By examining this plot, you may learn more about how varied degrees of poverty within urban and non-urban areas affect the distribution of literacy levels. The literacy distribution is shown for every combination of "is_urban" and "Poverty_conditional."

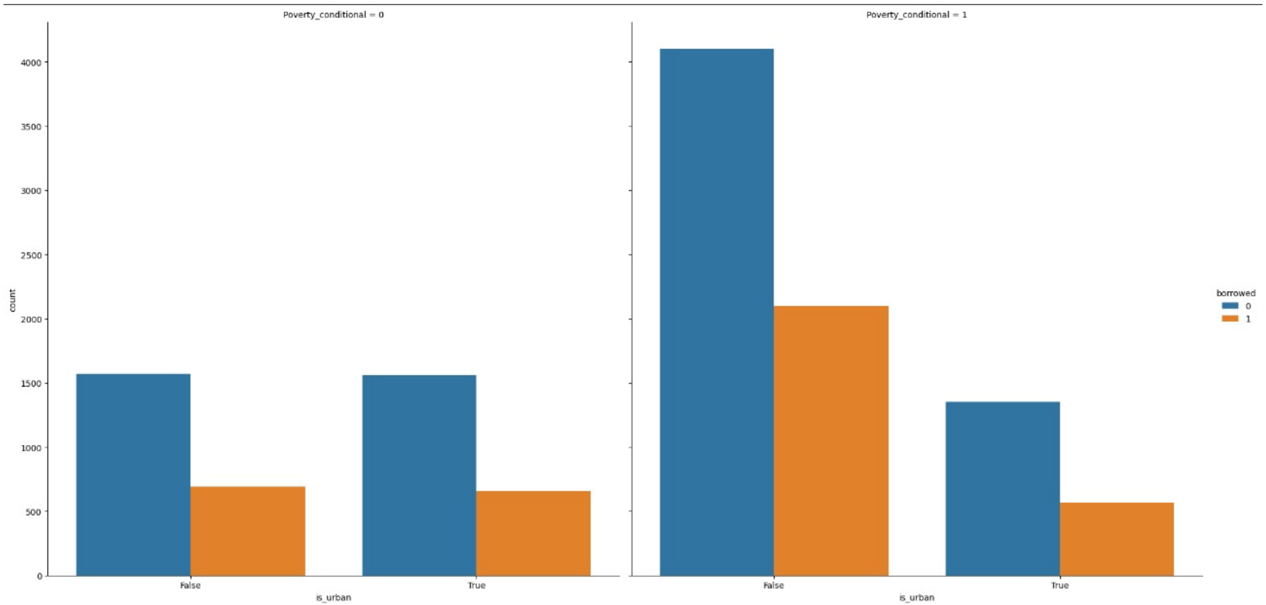


Figure-4.1.11: Differences in borrowing behavior at varying levels of poverty

By examining this plot, you will learn more about the differences in borrowing behavior between urban and non-urban locations at varying levels of poverty. You can see the distribution of people who have borrowed money ("borrowed" = 1) or who have not borrowed money ("borrowed" equals0) for each combination of "is_urban" and "Poverty_conditional."

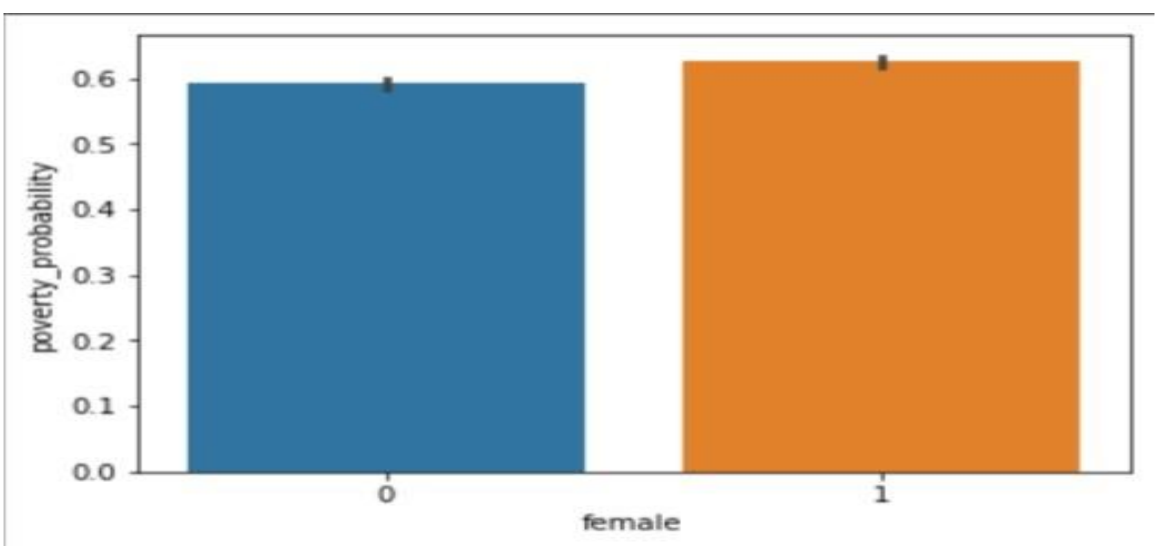


Figure-4.1.12: Distribution of various saving behaviors groups religious connections

By examining this plot, you may learn more about how various savings behavior groups' religious connections are distributed in urban and non-urban locations. You can see the distribution of religiousaffiliations for each combination of "is_urban" and "sav."

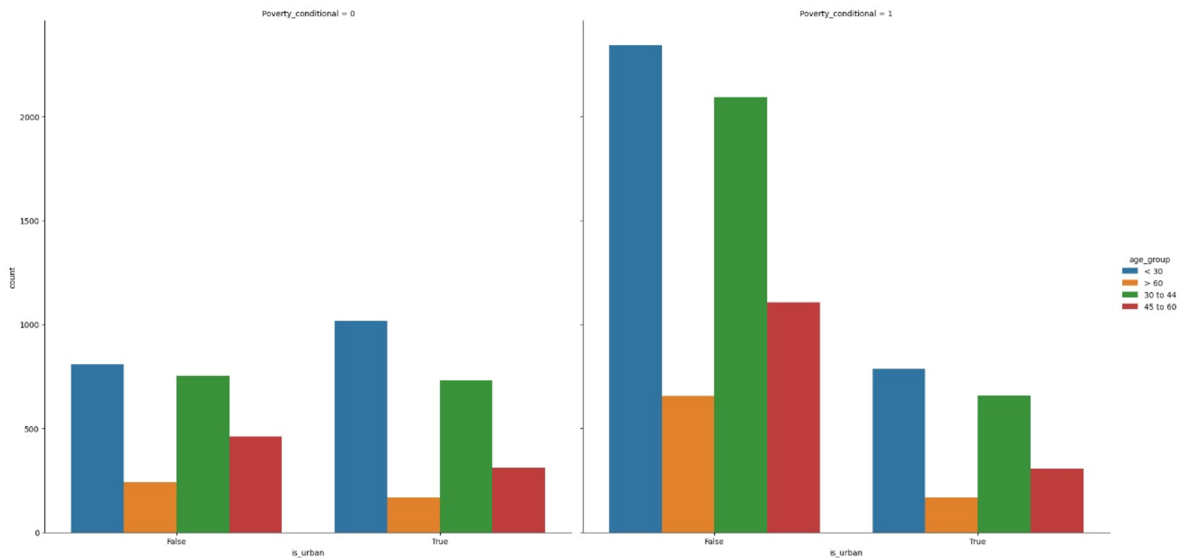


Figure-4.1.13: Differences in age distribution between urban and non-urban location.

By examining this plot, you may learn more about the differences in age group distribution between urban andnon-urban locations with varying degrees of poverty. The age group distribution is shown for every combination of "is_urban" and "Poverty_conditional."

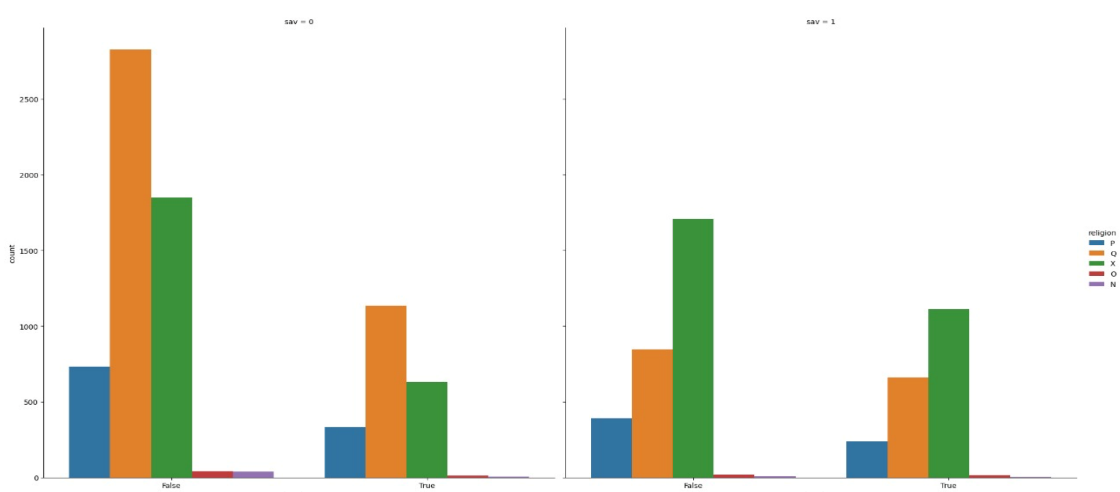


Figure- 4.1.14: Poverty Probability among female variable

By examining this bar plot, the average poverty likelihood between various "female" variable levels may be graphically compared. You can observe how the mean poverty likelihood varies by gender if "female" is binary. The plot offers a simple means of examining and contrasting the "poverty_probability" variable's central tendency among the "female" variable's categories.

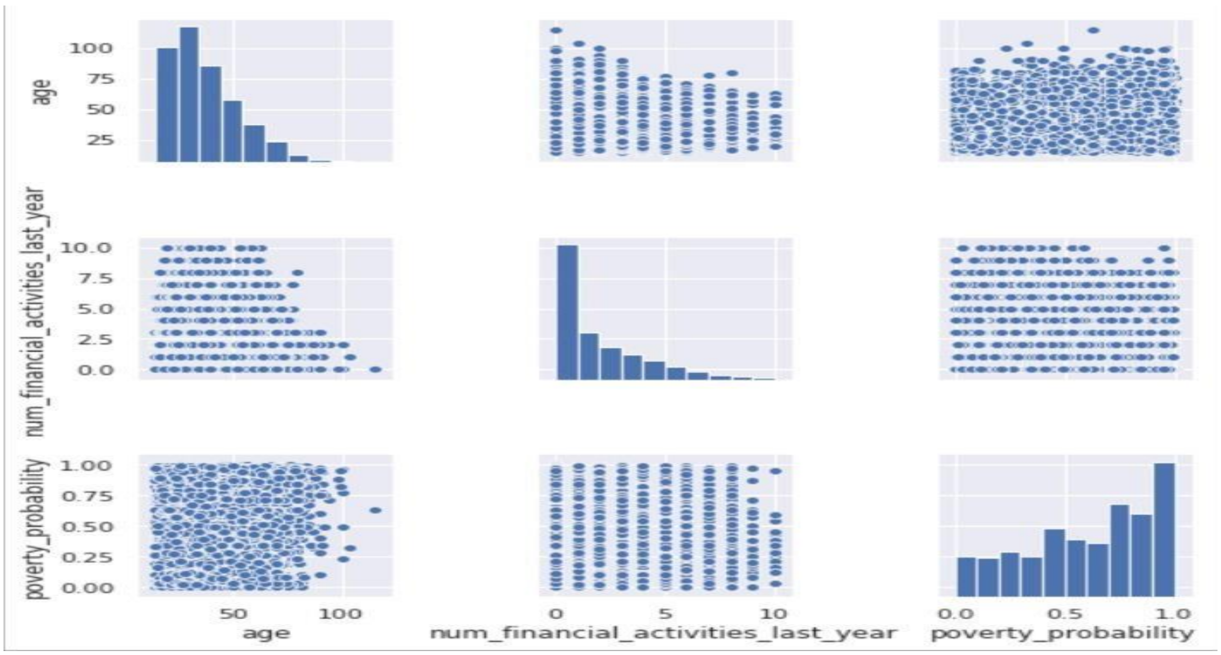


Figure-4.1.15: Quantitative and visual information on the correlations and relationships found.

By examining how to interpret a Pair Plot, you may examine the correlations between pairs of numerical variables graphically with the use of the scatterplots on the off-diagonal. For instance, you can look for any patterns or trends in the Relationship between 'age' and 'num_financial_activities_last_year' or 'poverty_probability.' TheDistribution of each individual variable is displayed by the histograms on the diagonal. This sheds light on the distributional spread and form of each variable.

“How to Interpret a Correlation Matrix?”

The quantitative measurements of the linear relationship between variables are provided by the correlation matrix (corrs). The correlation coefficients for each entry in the matrix range from -1 to 1. A linear connection is said to be positive (close to 1) if the correlation is positive, and negatively correlated (close to -1) if it is negative. If there is no linear relationship, the correlation value is 0. Finding the pairings of variables with strong or weak linear correlations may be accomplished by looking at the correlation matrix.

The combination of the graphs enables the investigation of correlations and links between the chosen numerical variables in the DataFrame in a quantitative and visual manner. A graphical summary may be obtained from the pair plot, while numerical insights into the linear relationships can be obtained via the correlation matrix.

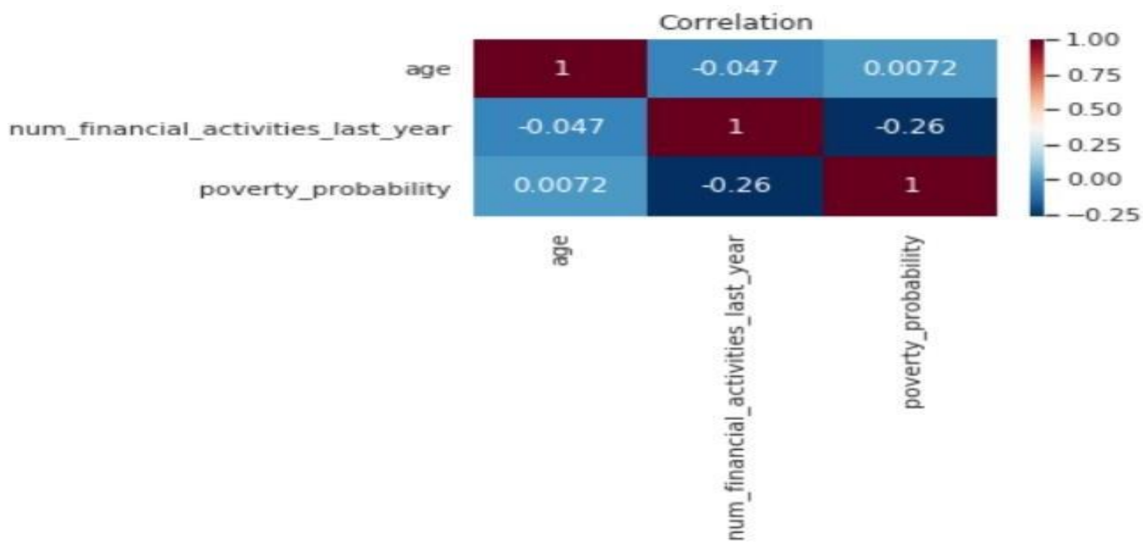


Figure-4.1.16: Heat map indicating Positive correlation.

By examining this graph, we can interpret the heat map as: Positive Correlations: Warmer-colored cells—that is, those that are closer to red—indicate positive correlations. The 'age' and 'poverty_probability' cells, for instance, may be red, indicating a positive correlation between higher age and a higher probability of poverty.

Negative Correlations: Cooler-colored cells, such as those that are closer to blue, suggest negative correlations. The 'num_financial_activities_last_year' and 'poverty_probability' boxes, for instance, may be blue, indicating a negative correlation between the likelihood of poverty and the number of financial activities completed in the previous year.

Strength of Correlations: The correlations' strength may be evaluated with the use of numerical valuesand color intensity. Values closer to 1 or -1 reflect stronger correlations, whereas values closer to 0 suggest weaker correlations.

4.2 Result using Unemployment Factor

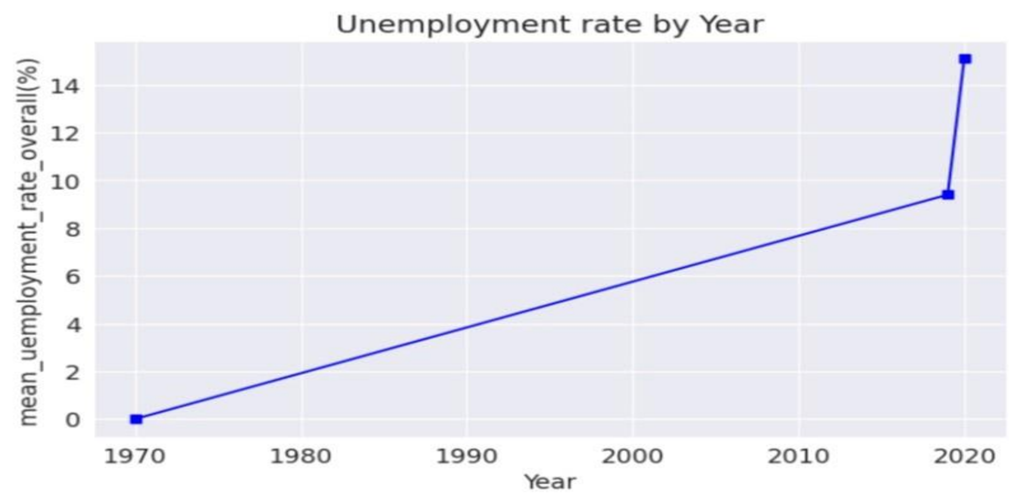


Figure-4.2.1: Relationship between time (years) and the mean unemployment rate.

By examine a line plot of unemployment rates over time and uses Seaborn to produce a dark grid style. A blue square ('s') designates each data point, and lines connect them together. Years are plotted on the x-axis, while the total mean unemployment rate is plotted on the y-axis. To give the story context and meaning, labels and a title are added., it shows the link between time (years) and the mean unemployment rate is shown graphically in the graph,which offers insights into the general trend or pattern in the data throughout the given years.

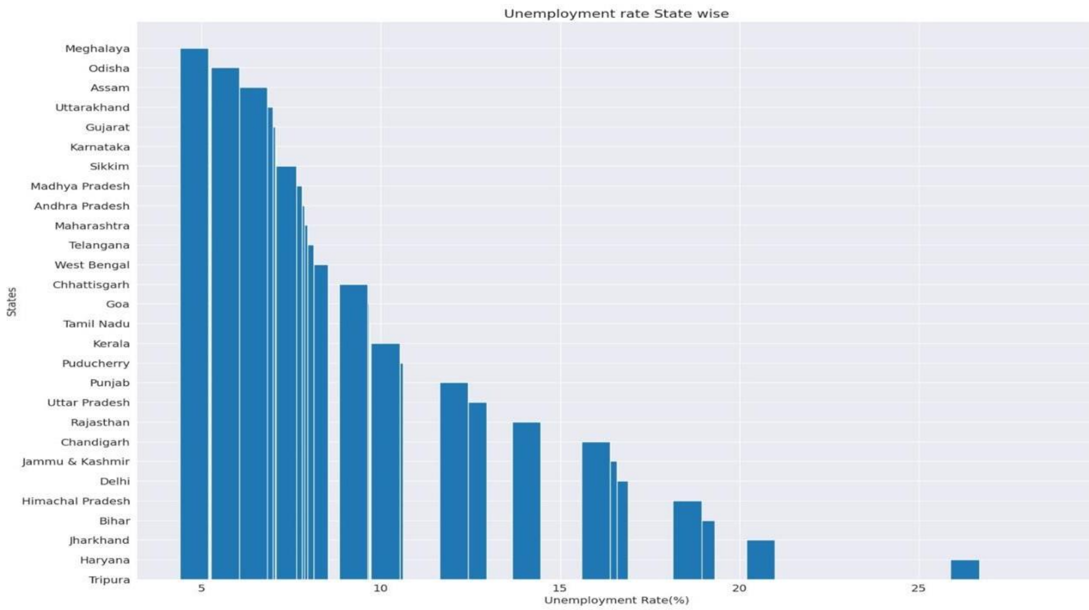


Figure-4.2.2: The unemployment rates for the various states.

By examine a thorough grasp of the link between states and their various unemployment rates is facilitated by the code, which creates a horizontal bar plot that illustrates unemployment rates by state. Both business executives and government leaders may make critical decisions with the help of this depiction.

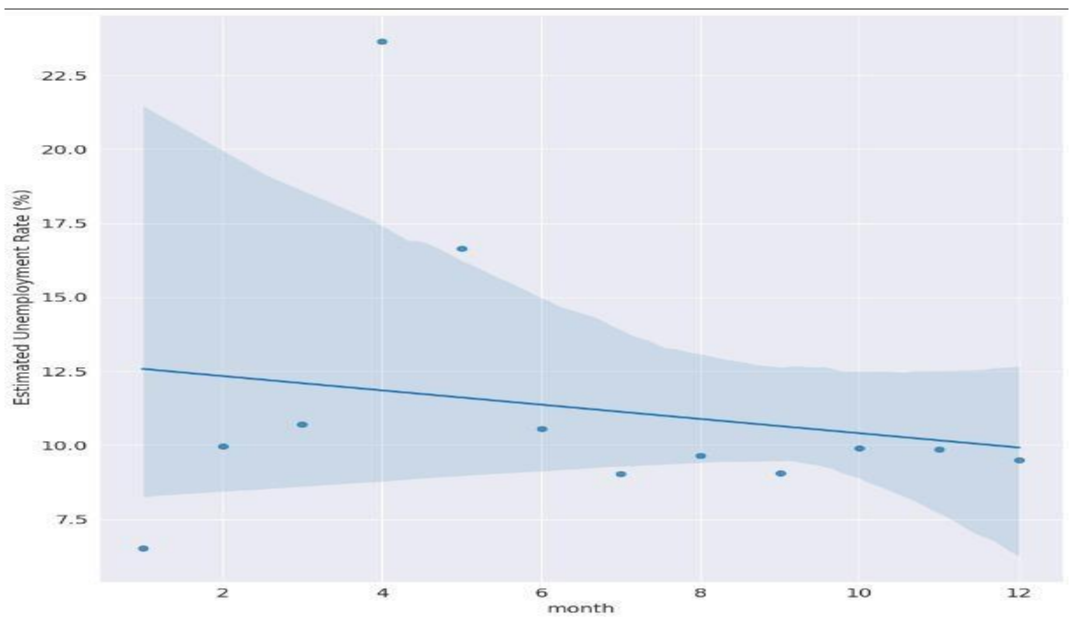
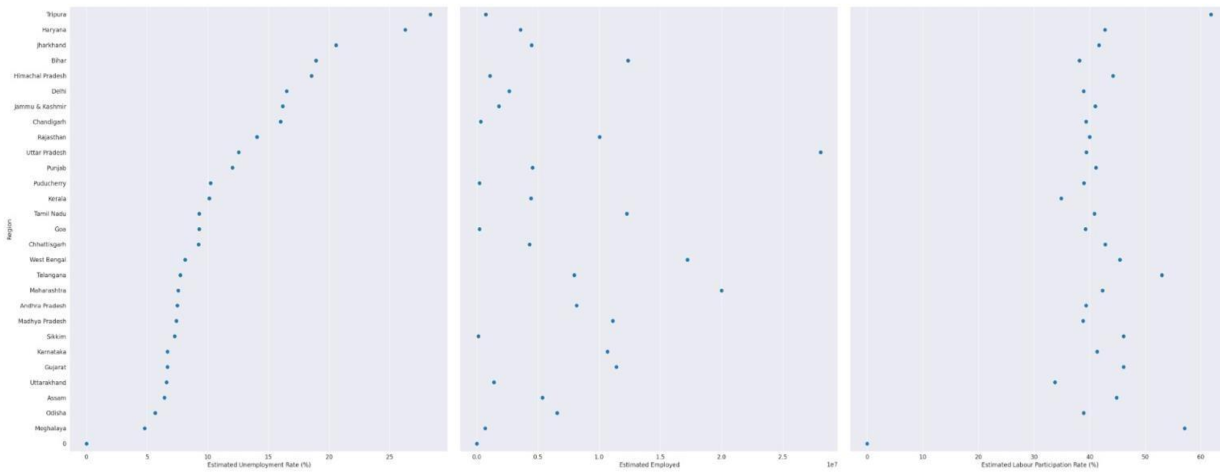


Figure-4.2.3: Relationship between the month and the predicted unemployment rate.

By examining this graph, we can comprehend how the projected unemployment rate varies over the course of the month, this kind of plot is helpful. Insights into the general trend are provided by the linear regression line, which shows if the unemployment rate changes in a predictable way as the months go by.

Figure-4.2.4: Relationship between unemployment rate and other factor



By examining this graph, the visualization approach offers a thorough understanding of the connections, accounting for the regional component, between the projected unemployment rate and other factors. It contributes to a better understanding of the variables driving unemployment rates by assisting in the identification of probable patterns, correlations, or variances in socioeconomic indicators across different areas. Unemployment and poverty are complicated issues influenced by several social and economic variables. To effectively address these concerns and devise interventions and policies, it is imperative to comprehend the complex linkages that exist between these components.

This essay investigates the relationship between poverty and unemployment rates and the population, employment rate, gender disparities, and literacy rate.

1. Rate of Literacy: The rate of literacy is an important measure of human capital and has a big impact on reducing unemployment and poverty.

1.1. Effect on Deprivation:

- Reduced rates of poverty are positively connected with higher literacy rates. This is so that they may acquire the information and skills needed to get higher-paying employment, which raises their standard of life and increases their income.

- By enabling people to make knowledgeable decisions regarding their family planning, money, and health, literacy helps to further reduce poverty.

- However, depending on the situation, literacy may or may not reduce poverty. Even those with greater education may find it difficult to obtain work in areas with a dearth of job prospects, which might result in ongoing poverty.

1.2. Effect on Joblessness:

- Raising the proportion of skilled workers in the labor force can lower unemployment rates through increased literacy. A workforce with higher levels of literacy is better able to adjust to shifting work requirements and support economic expansion.

- Through the provision of training and skills relevant to the labor market, literacy programs can increase an individual's employability.

- However, if additional variables like economic stagnation or a mismatch in skill sets exist, unemployment rates may still be high despite high literacy rates.

2. Demographics: The effects of population expansion on unemployment and poverty can be intricate and multidimensional.

2.1. Effect on Deprivation:

- Particularly in areas with limited resources, rapid population expansion can put a pressure on resources and make poverty worse. This may result in more people competing for housing, employment, and other necessities.

- But in other cases, rising population may also boost economic expansion by raising demand and spending. This may lead to the creation of new jobs and a decrease in poverty.

2.2. Effect on Joblessness:

- A bigger work force may result from rapid population increase, which might raise unemployment rates if job creation does not keep up with the demand.

- The population's age distribution is another factor. While an elderly population with decreased labor force participation might result in labor shortages, a young population entering the workforce can put pressure on the employment market.

3. Rate of Employment: The employment rate has a direct bearing on unemployment and poverty and is a crucial measure ofeconomic activity.

3.1. Effect on Deprivation:

- Because more people have access to income and can raise their standard of life, higher employment rates are associated with lower rates of poverty.

- Opportunities for employment enable people to become financially independent and improve the standard of living for their families and communities.

- But employment quality is also a major factor in reducing poverty. It's possible that low-paying employment with little perks won't pay enough to get out of poverty.

3.2. Effect on Joblessness:

- Higher unemployment rates are a direct result of low employment rates. Political instability and societal discontent may rise as a result of this.

- To lower unemployment rates, governments can enact measures that encourage job creation and economic expansion. These might involve making investments in training, education, and infrastructure.

4. Disparities in Gender: Poverty and unemployment rates are greatly impacted by gender disparities in social standards,educational attainment, and work possibilities.

4.1. Effect on Deprivation:

- In comparison to men, women experience poverty at a disproportionate rate. Numerous reasons, including poorer earnings, restricted access to education and training, and discriminatory cultural norms, are to blame for this.

- Combating poverty requires funding for girls' education as well as providing economic possibilities for women.

4.2. Effect on Joblessness:

- Due to factors including daycare insufficiency, maternity leave fines, and discriminatory hiring practices, women experience greater unemployment rates than males.

- Reducing the overall unemployment rate and accelerating economic development are two major benefits of closing the gender gap in the workforce.

5. Techniques and Suggestions: Drawing on the study, the following are some crucial tactics and suggestions fortackling unemployment and poverty:

- Putting money into education and skill development is essential for raising literacy rates, giving people the tools they need for the workforce, and increasing their employability.

- Encouraging economic expansion and job creation: Governments have the power to enact laws that draw in capital, assist small companies, and open up employment prospects across a range of industries.

- Reducing poverty and unemployment requires addressing gender inequality by putting laws and programs in place that support women's equitable access to resources, work opportunities, and education.

- Population management: It's important to put policies into place that encourage sustainable population increase and deal with the problems brought on by rapid population growth.

- Social safety nets: Enough social safety nets for the most disadvantaged groups can help reduce poverty and lessen the effects of unemployment.

CHAPTER-5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

This research study has presented a machine learning-powered platform to address the issues of market demand, government program and service ignorance, and the ensuing inefficiencies and resource waste.

Using machine learning to train models, the platform would gather information on a range of topics, such as residence, mode of transportation, and educational attainment. The models would then be used to predict a person's eligibility for government services and programs and to pinpoint areas that are in high demand by both governments and businesses. When making recommendations to users, the platform would use this data in a personalized way.

By lowering the time and effort needed to apply for government programs and services, enhancing access to local businesses and services, raising awareness of government programsand services, and encouraging economic growth and job creation, the platform has the potential to have a profound impact on the lives of a large number of people.

5.2 Future Scope

There are other areas where further research on this project could be done in the future in addition to the above-described next steps. Future studies might look into the following, for instance:

- To increase the precision of the platform's predictions and suggestions, new machine learning models are being developed connecting the platform to other current infrastructure, such company directories andgovernmental websites.
- Assessing how the platform affects consumers, companies, and governments.
- Creating additional functionality and features for the platform.

The study endeavor possesses the capability to yield a noteworthy impact on the domains of public policy and electronic government. The research project can assist to enhance the lives of many people, foster economic growth, and create jobs by creating a machine learning-powered platform to solve the issue of lack of awareness about government programs and services as well as lack of knowledge regarding market demand.

References

[1] J. Sphorer, Service Science: The next frontier in Service Innovation, Singapore: IBM CORPORATION, June 2017.

[2] R. N. a. G. Wilson, A Premiun for Good Governance, McKinsey Quarterly, 2002.

[3] "Organization for Economic Cooperation and Development, Directorate Financial, Fiscal and Enterprise affairs, OECD," Principles of Corporate Governance, p. 219, April 1999.

[4] P. W. a. J. W. Ross, "IT Governance: How Top Performers Manage IT Decisions for Superior Results," Harvard Business School Press, 2004, pp. 6-7.

[5] D. a. C. D. Barton, "Making advanced analytics work for you," Harvard Business Review, pp. 78-83, 2012.

[6] E. H. M. K. L. a. H. H. Brynjolfsson, "Strength in numbers: how does data-driven decisionmaking affect firm performance?," 8 December 2018. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486.

[7] K. Carillo, "Let's stop trying to be 'sexy'–preparing managers for the (big) data-drivenbusiness era," Business Process Management Journal, pp. 598-622, 2017.

[8] J. a. S. A. Corbin, Basics of Qualitative Research: Techniques and Procedures forDeveloping Grounded Theory, Los Angels: SAGE Publications, 2015.

[9] M. P. A. a. S. L. Hampson, 10 ideas for 21st century education, London: Innovation Unit, 2012.

[10] P. a. l. G. P. Lascoumes, "Introduction: understanding public policy through its instruments – from the nature of instruments to the sociology of public po," Governance, pp. 1-21, 2007.

[11] Nesta., "Decoding Learning," 28 May 2013. [Online]. Available: http://www.nesta.org.uk/areas_of_work/public_services_lab/digital_education/assets/features/more_and_better_learning_using_technology.

[12] H. A. C. a. V. R. Joseph, "A Conversation with Jeff Wu. Statistical Science," 2016, pp. 624-636.

[13] C. Wu, Statistics = Data Science?, 1997.

[14] P. Flach, "Machine Learning: The Art and Science of Algorithms that Make Sense of Data, Cambridge University Press, 2012.

[15] S. a. K. R. Coleman, "The Information Quality Framework for Evaluating Data Science Programs.," https://ssrn.com/abstract=2911557., 2017.

[16] "kdnuggets discussion," 2013. [Online]. Available:ttps://www.kdnuggets.com/2013/04/data-science-end-statistics-discussion.html.

[17] F. C. S. B. J. Mustafazade, "Application of machine learning for decision support in social housing. Statistics and Data Science," in new Developments for Business and Industrial Applications conference, turin.

[18] G. e. a. James, "The Introduction to Statistical Learning with Applications in R.," 2017.

[19] F. Cady, The Data Science Handbook, Wiley., wiley, 2017.

[20] V. Chang, "The Business Intelligence as a Service in the Cloud, Future Generation," 2014, pp. 37,512-514.

[21] V. &. W. G. Chang, "A Model to Compare Cloud and non-Cloud Storage of Big Data.," Future Generation Computer Systems, 2016, pp. 57,56-76.

[22] R. C. V. W. R. J. &. W. G. B. Newman, "Model and experimental development for Business Data Science," International Journal of Information Management, pp. 607-617, 2016.

Project Report

ORIGINALITY REPORT

12%	11%	5%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	4%
2	eprint.ncl.ac.uk Internet Source	1%
3	Submitted to University of Salford Student Paper	1%
4	Submitted to Liverpool John Moores University Student Paper	1%
5	Mauricius Munhoz de Medeiros, Norberto Hoppen, Antonio Carlos Gastaud Maçada. "Data science for business: benefits, challenges and opportunities", The Bottom Line, 2020 Publication	<1%
6	pypi.org Internet Source	<1%
7	Submitted to University of Bradford Student Paper	<1%

8	vdocuments.pub Internet Source	<1%
9	www.tandfonline.com Internet Source	<1%
10	tudr.thapar.edu:8080 Internet Source	<1%
11	Alonso, Igor Aguilar, José Carrillo Verdún, and Edmundo Tovar Caro. "Bridging the Gap between Technology and Business within Corporate Governance of Enterprise: Focus Areas of Special Attention in the Technology Domain", Advanced Materials Research, 2012. Publication	<1%
12	sourceforge.net Internet Source	<1%
13	eprints.soton.ac.uk Internet Source	<1%
14	Submitted to ABES Engineering College Student Paper	<1%
15	Submitted to Higher Education Commission Pakistan Student Paper	<1%
16	cse.anits.edu.in Internet Source	<1%
17	worldofusefulanswers.com Internet Source	

18	Submitted to th-koeln Student Paper	<1 %
19	hdl.handle.net Internet Source	<1 %
20	Orit Hazzan, Koby Mike. "Guide to Teaching Data Science", Springer Science and Business Media LLC, 2023 Publication	<1 %
21	economictimes.indiatimes.com Internet Source	<1 %
22	www.kluniversity.in Internet Source	<1 %
23	glisc.info Internet Source	<1 %
24	www.researchgate.net Internet Source	<1 %
25	dokumen.pub Internet Source	<1 %
26	technodocbox.com Internet Source	<1 %
27	www.4tad.org Internet Source	<1 %

28	www.slideshare.net Internet Source	<1 %
29	wz.pcz.pl Internet Source	<1 %