Master of Computer Application(MCA) Major Project Report

On

Analyzing the Impact of Digital Learning on Rural Education

Submitted in Partial Fulfillment of the Requirements for the Award of the Degree

of

Master of Computer Applications

Submitted by:

Anjali Kundliya

Student ID: O23MCA110251

Specialization: Data Analytics

Centre for Distance & Online Education

Chandigarh University

Year: 2025

Certificate

This is to certify that the Data Analytics major project entitled **"Analyzing the Impact of Digital Learning on Rural Education"**, submitted in partial fulfillment of the requirements for the award of the degree of Master of Computer Applications (MCA) in Data Analytics by Anjali Kundliya, Student ID: O23MCA110251 of Chandigarh University, under the supervision of [Guide's Name].

This project has not been submitted for the award of any other degree or diploma and reflects the student's independent effort and commitment to research.

Name of Student: Anjali Kundliya
Student ID: O23MCA110251
Date: 28th May, 2025

Guide Name: [Guide's Name]
Designation: [Guide's Designation]
Department: [Guide's Department]

Centre for Distance & Online Education
Chandigarh University
Year: 2025

Declaration

I, Anjali Kundliya, Student ID: O23MCA110251, hereby declare that the major project entitled **"Analyzing the Impact of Digital Learning on Rural Education"** submitted to Chandigarh University in partial fulfillment of the requirements for the award of the degree of Master of Computer Applications (MCA), is my original work and has not been submitted previously for the award of any degree or diploma in any other university or institute.

I further declare that all sources of information used have been duly acknowledged in the project report.

Name of Student: Anjali Kundliya
Student ID: O23MCA110251

Acknowledgment

I would like to express my sincere gratitude to [Guide's Name], my project guide, for their valuable guidance, encouragement, and support throughout the development of this project, **"Analyzing the Impact of Digital Learning on Rural Education"**. Their expertise and insights were instrumental in shaping the direction of this research.

I also thank the Centre for Distance & Online Education, Chandigarh University faculty members, for their constant support and for providing me with the necessary resources and environment to complete this project.

Finally, I thank my family and friends for their unwavering encouragement and moral support throughout this journey.

Name of Student: Anjali Kundliya
Student ID: O23MCA110251

Abstract

This project, titled **"Analyzing the Impact of Digital Learning on Rural Education,"** aims to assess the role of digital education in enhancing learning outcomes in rural India, aligning with Sustainable Development Goal 4 (SDG 4) - Quality Education. By leveraging datasets such as Education in India (which provides insights into enrollment trends, dropout rates, and teacher-student ratios), Internet Usage Rural VS Urban (highlighting digital connectivity disparities), Govt of India Literacy Rate (offering a state-wise literacy rate comparison), this study explores the digital divide and its implications for education in rural areas. The data will be stored and managed in a MySQL database for efficient querying, data preparation, and integration with Power BI. The analysis will focus on how digital access influences literacy rates, student enrollment, and retention, while also addressing regional disparities in internet access and infrastructure. The findings from these datasets will be visualized through an interactive Power BI dashboard to facilitate actionable insights for policymakers, educators, and stakeholders working towards achieving equitable and quality education in rural India, as outlined in SDG 4.

Table of Contents

Chapter 1.

Introduction

Education is a cornerstone of societal progress, and its quality significantly impacts a nation's economic growth and social development. In recent years, digital learning has emerged as a powerful tool to bridge educational gaps, offering flexible, scalable, and accessible learning solutions. However, the digital divide remains a critical barrier in India, where rural areas often struggle with inadequate infrastructure, limited digital literacy, and socio-economic challenges that hinder the effective adoption of digital learning.

This project, titled "Analyzing the Impact of Digital Learning on Rural Education," seeks to explore the impact of digital education on learning outcomes in rural India. By utilizing MySQL as the database for efficient data management and analysis, this study draws insights from key datasets such as Education in India (covering enrollment trends, dropout rates, and teacher-student ratios), Internet Usage Rural VS Urban (highlighting digital connectivity disparities), and Govt of India Literacy Rate (providing state-wise literacy comparisons). These datasets form the foundation for understanding how digital access influences literacy rates, student enrollment, and retention, while also shedding light on regional disparities in digital infrastructure.

Furthermore, this project aims to identify regional disparities in digital learning adoption, uncover patterns affecting student retention, and assess the impact of digital access on educational equity. The insights generated from this analysis will be presented through an interactive Power BI dashboard, providing policymakers, educators, and stakeholders with actionable information to drive targeted interventions and bridge the digital divide in education. This approach aligns with the

broader goal of ensuring quality education for all, as outlined in SDG 4, and contributes to building a more equitable and digitally inclusive education system in India.

Chapter 2.

SDLC of Project

The Software Development Life Cycle (SDLC) is a structured, step-by-step approach to planning, developing, testing, and deploying software projects. It provides a systematic way to deliver high-quality software within a defined timeline and budget. For this project, 'Analyzing the Impact of Digital Learning on Rural Education,' the SDLC ensures that all stages, from initial planning to final deployment, are carefully managed to produce reliable and insightful data analytics outputs.

1. **Project Planning and Requirement Analysis:** This phase involves understanding the project's objectives, defining the scope, and identifying the specific requirements for successful completion.
    - *Key Steps:*
        - Define project goals aligned with the title 'Analyzing the Impact of Digital Learning on Rural Education.'
        - Identify data requirements, including datasets like 'Education in India,' 'Internet Usage Rural vs Urban,' and 'Govt of India Literacy Rate.'
        - Establish success criteria, such as generating meaningful insights into digital learning gaps in rural education.
2. **System Design:** The design phase focuses on creating a blueprint for the project's architecture, data flow, and user interface.
    - *Key Steps:*
        - Plan data processing workflows and Power BI dashboard layout.

        - Design a data model that integrates multiple datasets for comprehensive analysis.
        - Outline key performance indicators (KPIs) like literacy rates, enrollment trends, and internet penetration.
        - Plan for integrating MySQL as the primary database for structured data storage and management.
3. **Data Collection and Preparation:** This phase involves gathering, cleaning, and transforming raw data to ensure accuracy and consistency.
    - *Key Steps:*
        - Collect data from identified Kaggle datasets.
        - Clean and preprocess the data to remove inconsistencies and handle missing values.
        - Import the cleaned CSV files into MySQL for efficient querying and data management.
        - Integrate data sources for seamless Power BI visualization through MySQL connections.
4. **Development:** This is the stage where the actual building of data pipelines, dashboards, and analytical models takes place.
    - *Key Steps:*
        - Build data pipelines to automate data extraction and processing.
        - Use MySQL to store and manage structured data, enabling complex data joins and aggregations.
        - Design Power BI dashboards with interactive charts, maps, and filters for data exploration.
        - Implement DAX functions for advanced data analysis.
5. **Testing:** Testing ensures that the developed components work as expected and meet the project requirements.
    - *Key Steps:*

        - Validate data accuracy, dashboard functionality, and user interactions.
        - Test for data consistency and responsiveness across different devices.
        - Ensure the MySQL database connections are stable and queries are optimized for performance.
        - Fix any bugs or issues identified during this phase.
6. **Deployment:** This phase involves publishing the final product for end-users or stakeholders to access.
    - *Key Steps:*
        - Publish the Power BI dashboard for public or organizational access.
        - Ensure data security and scalability for real-world usage.
        - Set up regular database updates to keep the dashboard data current and relevant.
7. **Maintenance and Improvement:** After deployment, ongoing maintenance is necessary to keep the dashboard updated and relevant.
    - *Key Steps:*
        - Regularly update the MySQL database with new data for continuous insights.
        - Gather user feedback for ongoing improvements and feature enhancements.
        - Optimize Power BI reports based on user interactions and changing data patterns.

This SDLC approach provides a structured, efficient, and outcome-focused development process, supporting the successful delivery of impactful educational insights through data analytics and visualization.

Chapter 3.

Design

The design should focus on effectively integrating multiple datasets, ensuring data quality, and creating impactful visualizations. Given the use of MySQL as the primary database, the design will also account for efficient data storage, querying, and management. Here's a structured design approach:

1. **Project Architecture**
    - *Data Sources:*
        a) Education in India Dataset (Enrollment, dropout rates, teacher-student ratios)
        b) Internet Usage Rural vs Urban (Digital connectivity disparities)
        c) Govt of India Literacy Rate (State-wise literacy levels)
    - *Data Storage:*
        a) MySQL Database for structured data storage and efficient querying
        b) Google Drive / Kaggle Datasets
    - *Data Processing:*
        a) Python in Google Colab (Pandas, NumPy) for data cleaning and transformation
        b) Import processed data to MySQL for centralized data management
    - *Visualization:* Power BI for interactive dashboards connected to MySQL for real-time data updates

2. Data Flow and Integration

- *Data Ingestion:*
    a) Import raw datasets from Kaggle to local storage
    b) Preprocess the data in Google Colab (handling missing values, data type conversions)
    c) Load cleaned data into MySQL for centralized data management
- *Data Preprocessing:*
    a) Clean, transform, and merge datasets for consistency
    b) Create structured tables in MySQL with primary and foreign keys for efficient data linking
- *Data Modeling:*
    a) Design relational tables in MySQL to integrate multiple datasets
    b) Establish relationships between tables for seamless Power BI integration
- *Data Analysis:*
    a) Use MySQL for complex data aggregations and filtering
    b) Apply statistical methods to identify key patterns and correlations before visualization

3. Dashboard Design (Power BI)

- *Pages:*
    - Overview of Digital Education in Rural India
    - Internet Penetration vs Education Outcomes
    - Gender Disparity in Digital Learning
    - Regional Analysis (State-wise comparisons)
- *Visual Elements:*
    - Bar Charts (Enrollment, dropout rates)
    - Line Charts (Trends over time)

    - Maps (Regional distribution of literacy rates)
    - KPIs (Digital Access Index, Student-Teacher Ratio)

4. Key Performance Indicators (KPIs)

- Student Enrollment Rates
- Dropout Rates (Male vs Female)
- Literacy Rates by State (Rural vs Urban)
- Internet Penetration Rates
- Teacher-Student Ratios

5. Analytical Approach

- *Descriptive Analysis:* Summary statistics, data distributions using MySQL queries
- *Correlation Analysis:* Impact of internet access on educational outcomes
- *Comparative Analysis:* Urban vs Rural Education Metrics
- *Trend Analysis:* Changes in digital learning adoption over time

Chapter 4.

Coding and Implementation

To effectively implement this project, a combination of Google Colab (for data preprocessing and analysis), MySQL (for structured data storage), and Power BI (for visualization) will be used. The coding and implementation can be broadly broken down into the following steps:

1. **Setting Up the Environment:**
    - Use Google Colab for Python coding, leveraging libraries like Pandas, NumPy, Matplotlib, Seaborn, and Plotly for data analysis and visualization.
    - Ensure integration with Google Drive for easy data access.
    - Set up a MySQL database on your local machine or cloud server for efficient data management and querying.
    - Install the required libraries and MySQL Connector for seamless integration between Python and MySQL.
2. **Data Import and Cleaning:**
    - Load the 'Education in India,' 'Internet Usage Rural VS Urban,' and 'Govt of India Literacy Rate' datasets.
    - Perform data cleaning in Google Colab:
        i. Handle missing values, remove duplicates, and format columns appropriately.
        ii. Standardize column names for consistency across datasets.
    - Import the cleaned data into MySQL to create structured tables for centralized storage and efficient querying.

- Use **MySQL** to define primary keys, foreign keys, and relationships between datasets for easier data integration.

3. **Data Analysis and Transformation:**
    - Perform exploratory data analysis (EDA) in Google Colab to identify key patterns and trends.
    - Use correlation analysis to measure the impact of digital learning on education outcomes.
    - Implement data aggregation for state-level and district-level insights.
    - Store the aggregated and transformed data in MySQL for faster access during visualization.
    - Use SQL queries to prepare data summaries and calculated fields required for Power BI dashboards.
4. **Data Visualization:**
    - Connect Power BI directly to the MySQL database for real-time data updates.
    - Create interactive dashboards with charts, maps, and filters for dynamic data exploration.
    - Implement DAX functions for advanced calculations, such as enrollment trends, dropout rates, and gender disparities.
    - Design user-friendly interfaces that highlight key insights, including student enrollment, literacy rates, and internet penetration.
5. **Reporting and Documentation:**
    - Document the code and analysis steps for project transparency.
    - Maintain a MySQL data schema document to outline the structure of the database.
    - Prepare a comprehensive project report summarizing findings and insights from the data analysis.
6. **Final Review and Presentation:**

- Review the code, MySQL database design, and Power BI dashboard for accuracy and completeness.
- Test data consistency and responsiveness across different devices.
- Prepare a presentation for the final project submission, including key findings, visualizations, and recommendations.

Below is the Python code executed in Google Colab:

Digital Learning Impact Analysis for Rural Education in India (SDG 4)

```
# Importing required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go
import warnings
warnings.filterwarnings('ignore')

# Mounting Google Drive for data access
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
# File Paths (replace with your actual file paths from Google Drive)
education_file =
'/content/drive/MyDrive/Major_Project/Education_in_India.csv'
internet_file =
'/content/drive/MyDrive/Major_Project/Internet_Usage_Rural_vs_Urban.cs
v'
literacy_file =
'/content/drive/MyDrive/Major_Project/Literacy_Rate_India.csv'
enrollment_file =
'/content/drive/MyDrive/Major_Project/Student_Enrollment.csv'
dropout_file =
'/content/drive/MyDrive/Major_Project/Dropout_Rates.csv'

# Loading the datasets
education_df = pd.read_csv(education_file)
internet_df = pd.read_csv(internet_file)
literacy_df = pd.read_csv(literacy_file)
enrollment_df = pd.read_csv(enrollment_file)
dropout_df = pd.read_csv(dropout_file)

# Previewing the datasets
print('Education Dataset:')
display(education_df.head())

print('\nInternet Usage Dataset:')
display(internet_df.head())

print('\nLiteracy Rate Dataset:')
display(literacy_df.head())
```

```
print('\nStudent Enrollment Dataset:')
display(enrollment_df.head())

print('\nDropout Rates Dataset:')
display(dropout_df.head())
```

```
Education Dataset:
```


```
{"type":"dataframe"}
```


```
Internet Usage Dataset:
```


```
{"summary":"{\n  \"name\": \"display(dropout_df\",\n  \"rows\": 5,\n
\"fields\": [\n    {\n      \"column\": \"States/UTs\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 2,\n        \"samples\": [\n          \"Andaman
& Nicobar Islands\",\n          \"India\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Area\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 3,\n
\"samples\": [\n          \"Urban\",\n          \"Rural\"\n        ],\
n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Women (age 15-49)  who have ever
used the internet (%)\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 5,\n        \"samples\":
[\n          \"24.6 \",\n          \"27.9 \"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Men (age 15-49)  who have ever
used the internet (%)\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 5,\n        \"samples\":
[\n          \"48.7 \",\n          \"41.1 \"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe"}
```


```
Literacy Rate Dataset:
```


```
{"summary":"{\n  \"name\": \"display(dropout_df\",\n  \"rows\": 5,\n
\"fields\": [\n    {\n      \"column\": \"Category\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 2,\n        \"samples\": [\n
\"State\",\n          \"Country\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Country/ States/ Union Territories
Name\",\n      \"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 5,\n        \"samples\": [\n          \"Andhra
Pradesh\",\n          \"Bihar\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Literacy Rate (Persons) - Total -
```

```
2001\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 7.336007088328091,\n        \"min\": 47.0,\n        \"max\":
64.8,\n        \"num_unique_values\": 5,\n        \"samples\": [\n
60.5,\n          47.0\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Literacy Rate (Persons) - Total - 2011\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 4.710838566539933,\n
\"min\": 61.8,\n        \"max\": 73.0,\n        \"num_unique_values\":
5,\n        \"samples\": [\n          67.0,\n          61.8\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Literacy Rate (Persons) - Rural - 2001\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 6.88200552164847,\n
\"min\": 43.9,\n        \"max\": 59.7,\n        \"num_unique_values\":
5,\n        \"samples\": [\n          54.5,\n          43.9\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Literacy Rate (Persons) - Rural - 2011\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 4.700319138101156,\n
\"min\": 59.8,\n        \"max\": 69.3,\n        \"num_unique_values\":
5,\n        \"samples\": [\n          60.4,\n          59.8\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Literacy Rate (Persons) - Urban - 2001\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 4.933558553417602,\n
\"min\": 71.9,\n        \"max\": 85.3,\n        \"num_unique_values\":
5,\n        \"samples\": [\n          76.1,\n          71.9\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Literacy Rate (Persons) - Urban - 2011\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 4.3543082114154466,\n
\"min\": 76.9,\n        \"max\": 88.5,\n        \"num_unique_values\":
5,\n        \"samples\": [\n          80.1,\n          76.9\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\n}","type":"dataframe"}
```


Student Enrollment Dataset:


```
{"summary":"{\n  \"name\": \"display(dropout_df\",\n  \"rows\": 5,\n
\"fields\": [\n    {\n      \"column\": \"ISO3\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 5,\n        \"samples\": [\n          \"AGO\",\
n          \"BGD\",\n          \"ARG\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Countries and areas\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 5,\n        \"samples\": [\n
\"Angola\",\n          \"Bangladesh\",\n          \"Argentina\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
```

```
}\n    },\n    {\n      \"column\": \"Region\",\n      \"properties\":
{\n        \"dtype\": \"string\",\n        \"num_unique_values\": 5,\n
\"samples\": [\n          \"SSA\",\n          \"SA\",\n
\"LAC\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Sub-region\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 5,\n        \"samples\":
[\n          \"ESA\",\n          \"SA\",\n          \"LAC\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Income Group\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 2,\n        \"samples\":
[\n          \"Lower middle income (LM)\",\n          \"Upper middle
income (UM)\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Total\",\n      \"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 5,\n        \"samples\": [\n          \"19%\",\
n          \"34%\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Residence Rural\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 4,\n        \"samples\":
[\n          \"2%\",\n          \"31%\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Residence Urban\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 4,\n        \"samples\": [\n          \"27%\",\
n          \"48%\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Wealth quintile Poorest\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 4,\n
\"samples\": [\n          \"0%\",\n          \"8%\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Wealth quintile Richest\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 4,\n        \"samples\": [\n          \"62%\",\
n          \"73%\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Data source\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 2,\n        \"samples\":
[\n          \"Demographic and Health Survey\",\n          \"Multiple
Indicator Cluster Survey\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Time period\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 4,\n
\"samples\": [\n          \"2015-16\",\n          \"2019\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\n}","type":"dataframe"}
```


Dropout Rates Dataset:

```
{"summary":"{\n  \"name\": \"display(dropout_df\",\n  \"rows\": 5,\n
\"fields\": [\n    {\n      \"column\": \"Rank\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
1,\n        \"min\": 1,\n        \"max\": 5,\n
\"num_unique_values\": 5,\n        \"samples\": [\n          2,\n
5,\n          3\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"State\",\n      \"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 5,\n        \"samples\": [\n
\"Maharashtra\",\n          \"Andhra Pradesh\",\n          \"Bihar\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Capital\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 5,\n        \"samples\": [\n
\"Mumbai\",\n          \"Hyderabad\",\n          \"Patna\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Population\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 5,\n        \"samples\":
[\n          \"112,374,333\",\n          \"84,580,777\",\n
\"104,099,452\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"%
of Total Population\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 3.8601321220911595,\n        \"min\":
6.99,\n        \"max\": 16.5,\n        \"num_unique_values\": 5,\n
\"samples\": [\n          9.28,\n          6.99,\n          8.6\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Males\",\n      \"properties\":
{\n        \"dtype\": \"string\",\n        \"num_unique_values\": 5,\n
\"samples\": [\n          \"58,243,056\",\n          \"42,442,146\",\n
\"54,278,157\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Females\",\n      \"properties\": {\n        \"dtype\": \"string\",\
n        \"num_unique_values\": 5,\n        \"samples\": [\n
\"54,131,277\",\n          \"42,138,631\",\n          \"49,821,295\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Sex Ratio\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 5,\n        \"samples\": [\n          \"929\",\
n          \"993\",\n          \"918\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Literacy Rate (%)\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
8.181747979496805,\n        \"min\": 61.8,\n        \"max\": 82.34,\n
\"num_unique_values\": 5,\n        \"samples\": [\n          82.34,\n
67.02,\n          61.8\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Rural Population\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 5,\n        \"samples\":
[\n          \"61,545,441\",\n          \"56,361,702\",\n
```

\"92,075,028\"\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"Urban Population\",\n   \"properties\": {\n   \"dtype\":
\"string\",\n   \"num_unique_values\": 5,\n   \"samples\":
[\n   \"50,827,531\",\n   \"28,219,075\",\n
\"11,729,609\"\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"Area (km*km)\",\n   \"properties\": {\n   \"dtype\":
\"string\",\n   \"num_unique_values\": 5,\n   \"samples\":
[\n   \"307,713\",\n   \"275,045\",\n
\"94,163\"\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"Density (1/km*km)\",\n   \"properties\": {\n   \"dtype\":
\"string\",\n   \"num_unique_values\": 5,\n   \"samples\":
[\n   \"365\",\n   \"308\",\n   \"1,102\"\n
],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   },\n   {\n   \"column\": \"Decadal Growth (%)\",\n
\"properties\": {\n   \"dtype\": \"string\",\n
\"num_unique_values\": 5,\n   \"samples\": [\n
\"16.0%\",\n   \"10.98%\",\n   \"25.1%\"\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   }\n]\n}","type":"dataframe"}

```
# Basic Data Information
print('\nBasic Information about Datasets:')
education_df.info()
internet_df.info()
literacy_df.info()
enrollment_df.info()
dropout_df.info()
```

```
Basic Information about Datasets:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 680 entries, 0 to 679
Columns: 819 entries, AC_YEAR to TOTCLS7G
dtypes: float64(13), int64(803), object(3)
memory usage: 4.2+ MB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 111 entries, 0 to 110
Data columns (total 4 columns):
 #   Column                                                Non-Null
Count  Dtype
---  ------
--------------  -----
 0   States/UTs                                            111 non-
null    object
 1   Area                                                  111 non-
null    object
 2   Women (age 15-49)  who have ever used the internet (%)  111 non-
```

```
null    object
 3   Men (age 15-49)  who have ever used the internet (%)    111 non-
null    object
dtypes: object(4)
memory usage: 3.6+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 8 columns):
 #   Column                                     Non-Null Count  Dtype
---  ------                                     --------------  -----
 0   Category                                   36 non-null     object
 1   Country/ States/ Union Territories Name  36 non-null     object
 2   Literacy Rate (Persons) - Total - 2001   36 non-null     float64
 3   Literacy Rate (Persons) - Total - 2011   36 non-null     float64
 4   Literacy Rate (Persons) - Rural - 2001   36 non-null     float64
 5   Literacy Rate (Persons) - Rural - 2011   36 non-null     float64
 6   Literacy Rate (Persons) - Urban - 2001   36 non-null     float64
 7   Literacy Rate (Persons) - Urban - 2011   36 non-null     float64
dtypes: float64(6), object(2)
memory usage: 2.4+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 87 entries, 0 to 86
Data columns (total 12 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   ISO3                     87 non-null     object
 1   Countries and areas      87 non-null     object
 2   Region                   87 non-null     object
 3   Sub-region               87 non-null     object
 4   Income Group             87 non-null     object
 5   Total                    87 non-null     object
 6   Residence Rural          77 non-null     object
 7   Residence Urban          80 non-null     object
 8   Wealth quintile Poorest  70 non-null     object
 9   Wealth quintile Richest  68 non-null     object
 10  Data source              87 non-null     object
 11  Time period              87 non-null     object
dtypes: object(12)
memory usage: 8.3+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35 entries, 0 to 34
Data columns (total 14 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Rank                   35 non-null     int64
 1   State                  35 non-null     object
 2   Capital                34 non-null     object
 3   Population             35 non-null     object
 4   % of Total Population  35 non-null     float64
```

```
 5   Males                   35 non-null     object
 6   Females                 35 non-null     object
 7   Sex Ratio               35 non-null     object
 8   Literacy Rate (%)       35 non-null     float64
 9   Rural Population        35 non-null     object
 10  Urban Population        35 non-null     object
 11  Area (km*km)            35 non-null     object
 12  Density (1/km*km)       35 non-null     object
 13  Decadal Growth (%)      35 non-null     object
dtypes: float64(2), int64(1), object(11)
memory usage: 4.0+ KB
```

```
# Checking for missing values
print('\nMissing Values in Datasets:')
print('Education Dataset:', education_df.isnull().sum().sum())
print('Internet Usage Dataset:', internet_df.isnull().sum().sum())
print('Literacy Rate Dataset:', literacy_df.isnull().sum().sum())
print('Student Enrollment Dataset:',
enrollment_df.isnull().sum().sum())
print('Dropout Rates Dataset:', dropout_df.isnull().sum().sum())
```

```
Missing Values in Datasets:
Education Dataset: 569
Internet Usage Dataset: 0
Literacy Rate Dataset: 0
Student Enrollment Dataset: 53
Dropout Rates Dataset: 1
```

```
# Initial Data Cleaning (Dropping NaNs and duplicates)
education_df.dropna(inplace=True)
education_df.drop_duplicates(inplace=True)

internet_df.dropna(inplace=True)
internet_df.drop_duplicates(inplace=True)

literacy_df.dropna(inplace=True)
literacy_df.drop_duplicates(inplace=True)

enrollment_df.dropna(inplace=True)
enrollment_df.drop_duplicates(inplace=True)

dropout_df.dropna(inplace=True)
dropout_df.drop_duplicates(inplace=True)

print('\nData cleaning completed.')
```

```
Data cleaning completed.
```

```python
# Basic Summary Statistics
print('\nBasic Summary Statistics:')
print('\nEducation Data:')
print(education_df.describe())

print('\nInternet Usage Data:')
print(internet_df.describe())

print('\nLiteracy Rate Data:')
print(literacy_df.describe())

print('\nStudent Enrollment Data:')
print(enrollment_df.describe())

print('\nDropout Rates Data:')
print(dropout_df.describe())
```

```
Basic Summary Statistics:

Education Data:
           STATCD       DISTCD  DISTRICTS      BLOCKS     VILLAGES  \
count  625.000000   625.000000      625.0  625.000000   625.000000
mean    17.123200  1728.449600        1.0   11.196800   902.617600
std      9.585606   957.614955        0.0    9.844854   634.078639
min      1.000000   101.000000        1.0    1.000000     6.000000
25%      9.000000   928.000000        1.0    6.000000   399.000000
50%     18.000000  1817.000000        1.0    9.000000   819.000000
75%     24.000000  2410.000000        1.0   13.000000  1234.000000
max     36.000000  3610.000000        1.0   66.000000  3963.000000

         CLUSTERS    TOTPOPULAT   P_URB_POP  POPULATION_0_6
GROWTHRATE  ...  \
count  625.000000  6.250000e+02  625.000000    6.250000e+02
625.000000  ...
mean   124.952000  1.888816e+06   24.721488    2.487723e+05
17.700576  ...
std     96.306507  1.518239e+06   18.988409    1.963243e+05
11.293000  ...
min      1.000000  7.948000e+03  -25.350000    1.103980e+03  -
58.390000  ...
25%     60.000000  8.240590e+05   11.370000    1.092593e+05
12.080000  ...
50%    105.000000  1.563107e+06   19.400000    2.040779e+05
17.090000  ...
75%    167.000000  2.585962e+06   32.400000    3.374011e+05
22.780000  ...
max    592.000000  1.105413e+07  100.000000    1.256855e+06
111.010000  ...
```

```
            UUNI_ALL        UUNI_SC       UUNI_ST      TOTCLS1G
TOTCLS2G  \
count     625.000000     625.000000    625.000000    625.000000
625.000000
mean    47976.334400   11857.083200   5388.812800   4718.425600
3251.440000
std     59291.695536   17020.437047   9558.063282   3957.180655
3349.302481
min         0.000000       0.000000      0.000000     24.000000
0.000000
25%      7408.000000     715.000000    158.000000   2007.000000
712.000000
50%     33599.000000    7164.000000   1175.000000   3712.000000
2074.000000
75%     63080.000000   15701.000000   5766.000000   6183.000000
4935.000000
max    511213.000000  153047.000000  68723.000000  24728.000000
25677.000000

           TOTCLS3G     TOTCLS4G     TOTCLS5G      TOTCLS6G
TOTCLS7G
count    625.000000   625.000000   625.000000    625.000000
625.000000
mean     800.372800   926.480000   414.646400    733.729600
410.059200
std     1335.467712  1463.457884   848.541153   1142.826513
970.904971
min        0.000000     0.000000     0.000000      0.000000
0.000000
25%       79.000000    16.000000    20.000000    160.000000
22.000000
50%      277.000000   111.000000    84.000000    439.000000
120.000000
75%      955.000000  1414.000000   496.000000    866.000000
341.000000
max    16053.000000  8590.000000  8074.000000  15458.000000
11904.000000

[8 rows x 816 columns]

Internet Usage Data:
       States/UTs   Area  \
count         111    111
unique         37      3
top         India  Urban
freq            3     37

       Women (age 15-49)  who have ever used the internet (%)  \
count                                                  111
unique                                                 106
```

```
top                                                    30.8
freq                                                      2

       Men (age 15-49)  who have ever used the internet (%)
count                                                   111
unique                                                   99
top                                                    79.7
freq                                                      2

Literacy Rate Data:
       Literacy Rate (Persons) - Total - 2001  \
count                                36.000000
mean                                 69.433333
std                                  10.647709
min                                  47.000000
25%                                  62.975000
50%                                  68.700000
75%                                  77.225000
max                                  90.900000

       Literacy Rate (Persons) - Total - 2011  \
count                                36.000000
mean                                 77.711111
std                                   8.497723
min                                  61.800000
25%                                  71.725000
50%                                  76.600000
75%                                  85.850000
max                                  94.000000

       Literacy Rate (Persons) - Rural - 2001  \
count                                36.000000
mean                                 64.669444
std                                  11.459440
min                                  43.900000
25%                                  57.425000
50%                                  63.300000
75%                                  74.275000
max                                  90.000000

       Literacy Rate (Persons) - Rural - 2011  \
count                                36.000000
mean                                 73.413889
std                                   9.190497
min                                  59.800000
25%                                  65.875000
50%                                  71.900000
75%                                  80.875000
max                                  93.000000
```

```
       Literacy Rate (Persons) - Urban - 2001  \
count                                36.000000
mean                                 82.161111
std                                   5.463138
min                                  69.800000
25%                                  79.275000
50%                                  81.850000
75%                                  84.925000
max                                  96.100000

       Literacy Rate (Persons) - Urban - 2011
count                                36.000000
mean                                 86.283333
std                                   4.982856
min                                  75.100000
25%                                  83.175000
50%                                  86.250000
75%                                  89.650000
max                                  97.600000

Student Enrollment Data:
       ISO3 Countries and areas Region Sub-region                Income
Group  \
count    68                   68     68         68
68
unique   68                   68      6          7
4
top     DZA              Algeria    SSA        WCA  Upper middle income
(UM)
freq      1                    1     26         16
25

       Total Residence Rural Residence Urban Wealth quintile
Poorest  \
count     68              68              68                       68

unique    46              40              48                       33

top       4%              1%             89%                       0%

freq       3               9               3                       24


       Wealth quintile Richest                          Data source Time
period
count                        68                                  68
68
unique                       43                                  12
18
top                         99%  Multiple Indicator Cluster Survey
```

```
2018
freq                          6                                    42
13

Dropout Rates Data:
            Rank  % of Total Population  Literacy Rate (%)
count  34.000000              34.000000          34.000000
mean   17.970588               2.910294          78.257353
std    10.399549               3.713324           8.517944
min     1.000000               0.010000          61.800000
25%     9.250000               0.112500          72.360000
50%    17.500000               1.740000          78.620000
75%    26.750000               5.035000          86.000000
max    35.000000              16.500000          94.000000
```

```
# Data Preprocessing and Merging

# Renaming columns for consistency (if needed)
education_df.columns =
education_df.columns.str.strip().str.lower().str.replace(' ', '_')
internet_df.columns =
internet_df.columns.str.strip().str.lower().str.replace(' ', '_')
literacy_df.columns =
literacy_df.columns.str.strip().str.lower().str.replace(' ', '_')
enrollment_df.columns =
enrollment_df.columns.str.strip().str.lower().str.replace(' ', '_')
dropout_df.columns =
dropout_df.columns.str.strip().str.lower().str.replace(' ', '_')

print("\nEducation Dataset Columns:\n", education_df.columns)
print("\nLiteracy Rate Dataset Columns:\n", literacy_df.columns)
print("\nInternet Usage Dataset Columns:\n", internet_df.columns)
print("\nStudent Enrollment Dataset Columns:\n",
enrollment_df.columns)
print("\nDropout Rates Dataset Columns:\n", dropout_df.columns)
```

```
Education Dataset Columns:
 Index(['ac_year', 'statcd', 'distcd', 'statname', 'distname',
'districts',
       'blocks', 'villages', 'clusters', 'totpopulat',
       ...
       'uuni_all', 'uuni_sc', 'uuni_st', 'totcls1g', 'totcls2g',
'totcls3g',
       'totcls4g', 'totcls5g', 'totcls6g', 'totcls7g'],
      dtype='object', length=819)

Literacy Rate Dataset Columns:
 Index(['category', 'country/_states/_union_territories_name',
       'literacy_rate_(persons)_-_total_-_2001',
```

```
       'literacy_rate_(persons)_-_total_-_2011',
       'literacy_rate_(persons)_-_rural_-_2001',
       'literacy_rate_(persons)_-_rural_-_2011',
       'literacy_rate_(persons)_-_urban_-_2001',
       'literacy_rate_(persons)_-_urban_-_2011'],
      dtype='object')

Internet Usage Dataset Columns:
 Index(['states/uts', 'area',
       'women_(age_15-49)__who_have_ever_used_the_internet_(%)',
       'men_(age_15-49)__who_have_ever_used_the_internet_(%)'],
      dtype='object')

Student Enrollment Dataset Columns:
 Index(['iso3', 'countries_and_areas', 'region', 'sub-region',
'income_group',
       'total', 'residence_rural', 'residence_urban',
       'wealth_quintile_poorest', 'wealth_quintile_richest',
'data_source',
       'time_period'],
      dtype='object')

Dropout Rates Dataset Columns:
 Index(['rank', 'state', 'capital', 'population',
'%_of_total_population',
       'males', 'females', 'sex_ratio', 'literacy_rate_(%)',
       'rural_population', 'urban_population', 'area_(km*km)',
       'density_(1/km*km)', 'decadal_growth_(%)'],
      dtype='object')
```

```
education_df.rename(columns={"STATNAME": "state"}, inplace=True)
literacy_df.rename(columns={"country/_states/_union_territories_name":
"state"}, inplace=True)
internet_df.rename(columns={"states/uts": "state"}, inplace=True)
enrollment_df.rename(columns={"countries_and_areas": "state"},
inplace=True)
dropout_df.rename(columns={"state": "state"}, inplace=True)

print(set(education_df.columns).intersection(set(literacy_df.columns))
)
print(set(education_df.columns).intersection(set(internet_df.columns))
)
print(set(education_df.columns).intersection(set(enrollment_df.columns
)))
print(set(education_df.columns).intersection(set(dropout_df.columns)))
```

```
set()
set()
set()
set()
```

```
# Merging the datasets using common keys ("state", "region", or
"state/ut")
possible_keys = ['statname', 'region', 'state/ut']

common_key = None
for key in possible_keys:
    if key in education_df.columns:
        common_key = key
        print(f"\nUsing '{common_key}' as the common key for
merging.")
        break

if common_key is None:
    raise KeyError('No common key ("statname", "region", or
"state/ut") found in Education dataset.')

merged_df = education_df

for df in [literacy_df, internet_df, enrollment_df, dropout_df]:
    if common_key in df.columns:
        print(f"\nMerging on column: {common_key}")
        merged_df = merged_df.merge(df, on=common_key, how='inner')
    else:
        print(f"\nWarning: Common key '{common_key}' not found in
dataset. Skipping this dataset.")

print('\nMerged Dataset Preview:')
display(merged_df.head())
```

```
Using 'statname' as the common key for merging.

Warning: Common key 'statname' not found in dataset. Skipping this
dataset.

Warning: Common key 'statname' not found in dataset. Skipping this
dataset.

Warning: Common key 'statname' not found in dataset. Skipping this
dataset.

Warning: Common key 'statname' not found in dataset. Skipping this
dataset.

Merged Dataset Preview:
```

```
{"type":"dataframe"}
```

```
# Checking for any remaining missing values after merging
print('\nMissing Values after Merging:',
merged_df.isnull().sum().sum())
```

```
Missing Values after Merging: 0
```

```
# Final Clean-Up
merged_df.dropna(inplace=True)
merged_df.drop_duplicates(inplace=True)

print('\nFinal Dataset Shape:', merged_df.shape)
print('\nFinal Dataset Columns:', merged_df.columns)
```

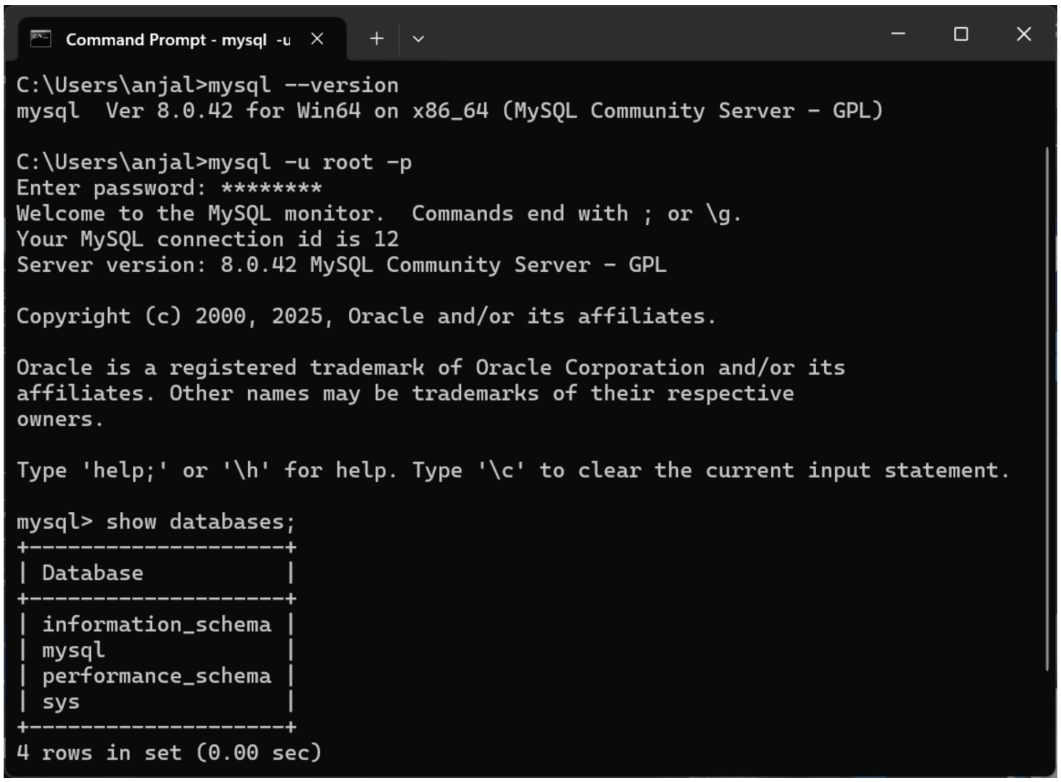
```
Final Dataset Shape: (625, 819)
```

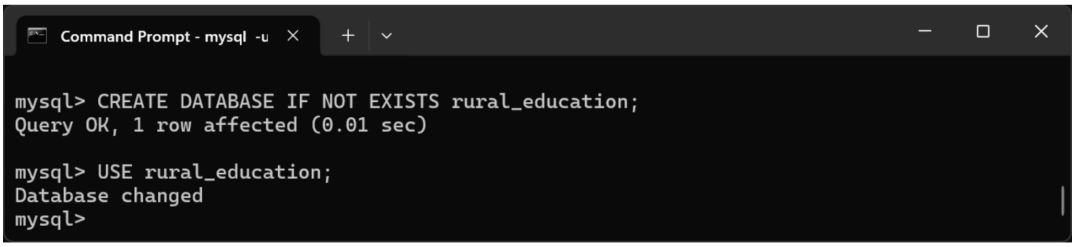
```
Final Dataset Columns: Index(['ac_year', 'statcd', 'distcd',
'statname', 'distname', 'districts',
       'blocks', 'villages', 'clusters', 'totpopulat',
       ...
       'uuni_all', 'uuni_sc', 'uuni_st', 'totcls1g', 'totcls2g',
'totcls3g',
       'totcls4g', 'totcls5g', 'totcls6g', 'totcls7g'],
      dtype='object', length=819)
```

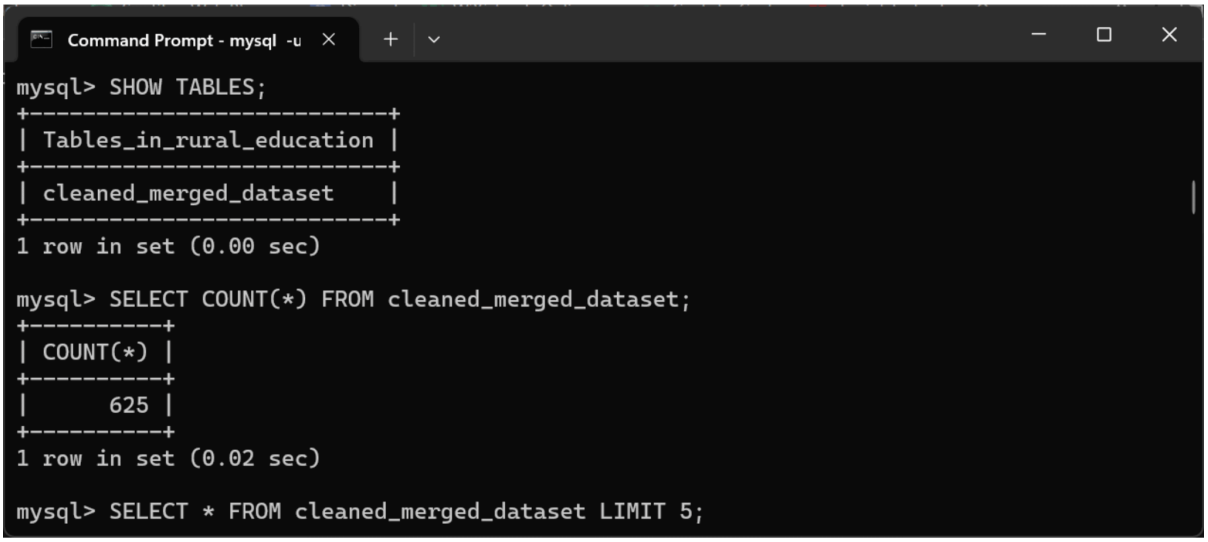
```
# Saving the cleaned and merged dataset for Power BI
merged_df.to_csv('/content/drive/MyDrive/Major_Project/Cleaned_Merged_
Dataset.csv', index=False)
print('\nCleaned and Merged Dataset saved successfully.')
```

```
Cleaned and Merged Dataset saved successfully.
```

Below are the step-by-step MySQL procedures to create a database and connect it to Power BI for creating charts:






Command Prompt - mysql -u
mysql> SHOW TABLES;
+---------------------------+
| Tables_in_rural_education |
+---------------------------+
| cleaned_merged_dataset |
+---------------------------+
1 row in set (0.00 sec)
mysql> SELECT COUNT(*) FROM cleaned_merged_dataset;
+----------+
| COUNT(*) |
+----------+
| 625 |
+----------+
1 row in set (0.02 sec)
mysql> SELECT * FROM cleaned_merged_dataset LIMIT 5;

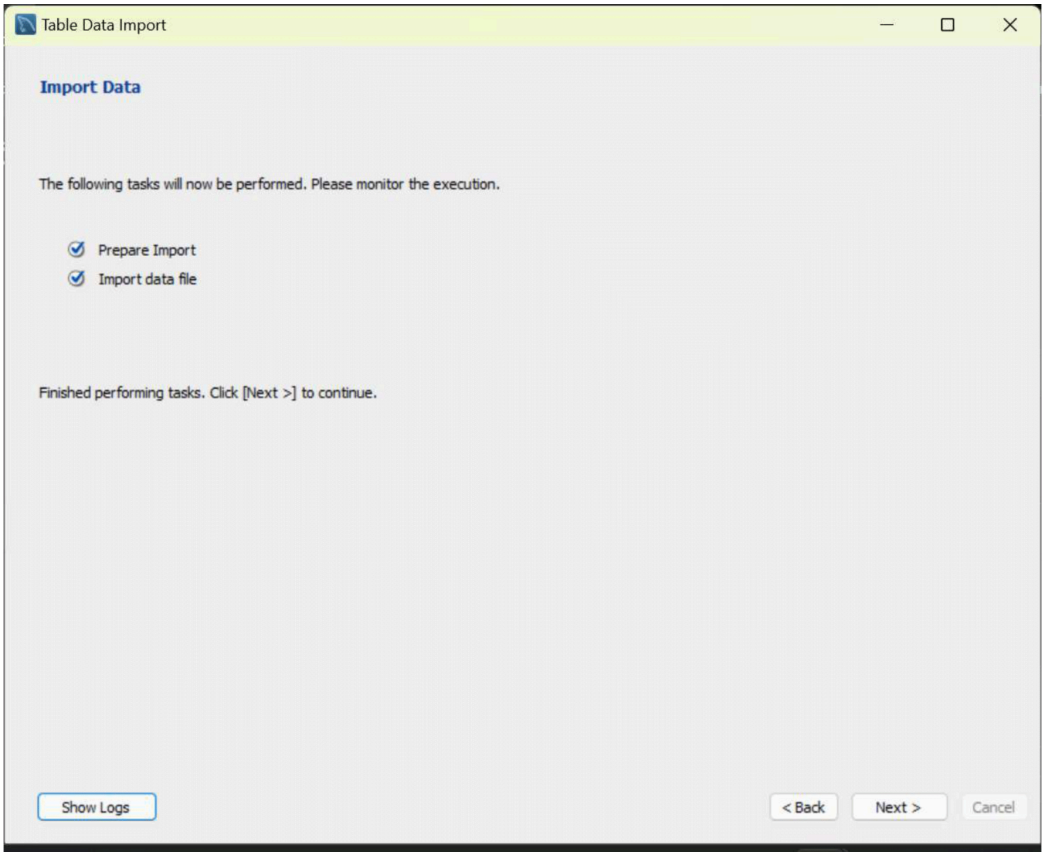
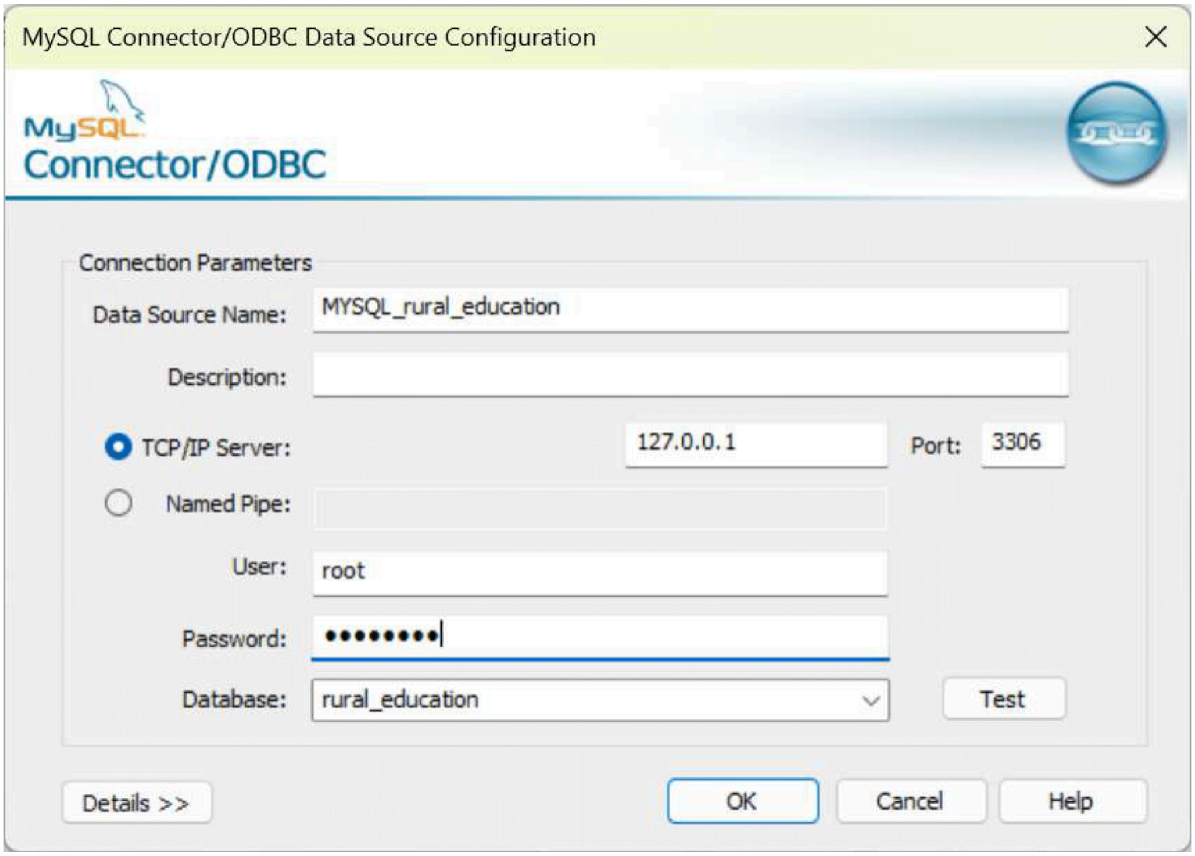
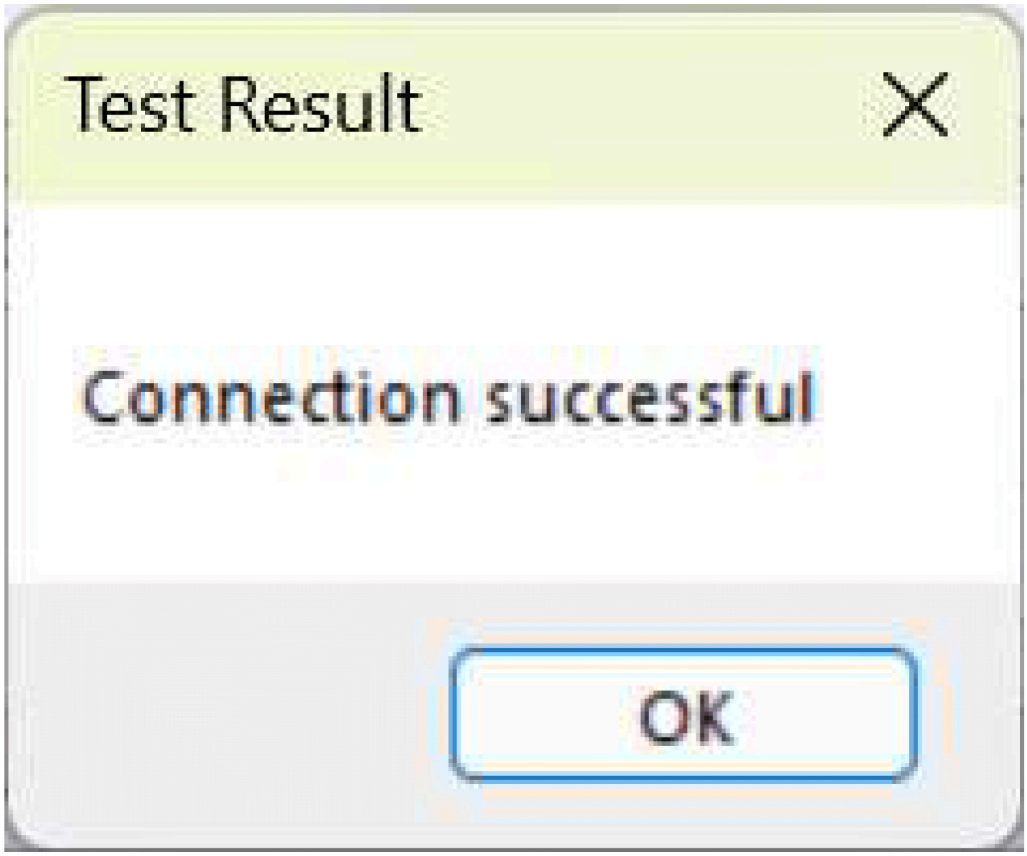
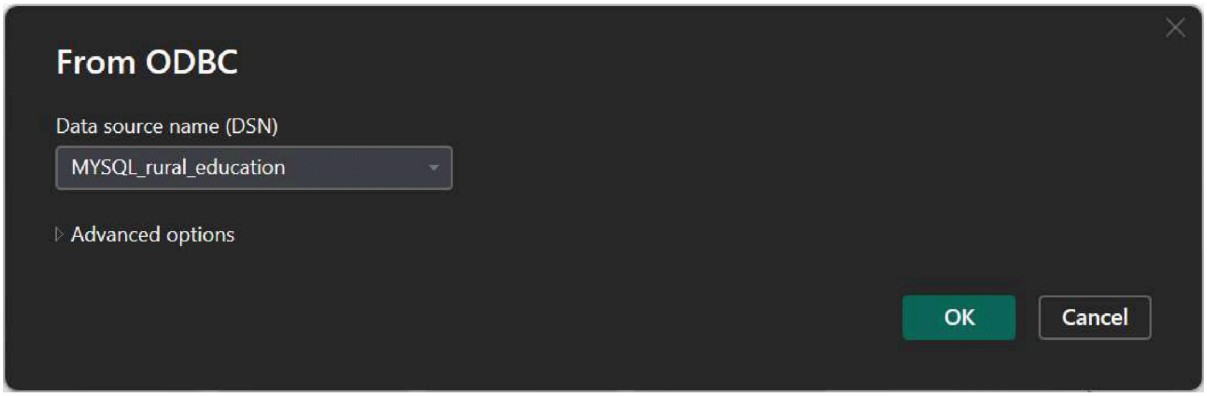


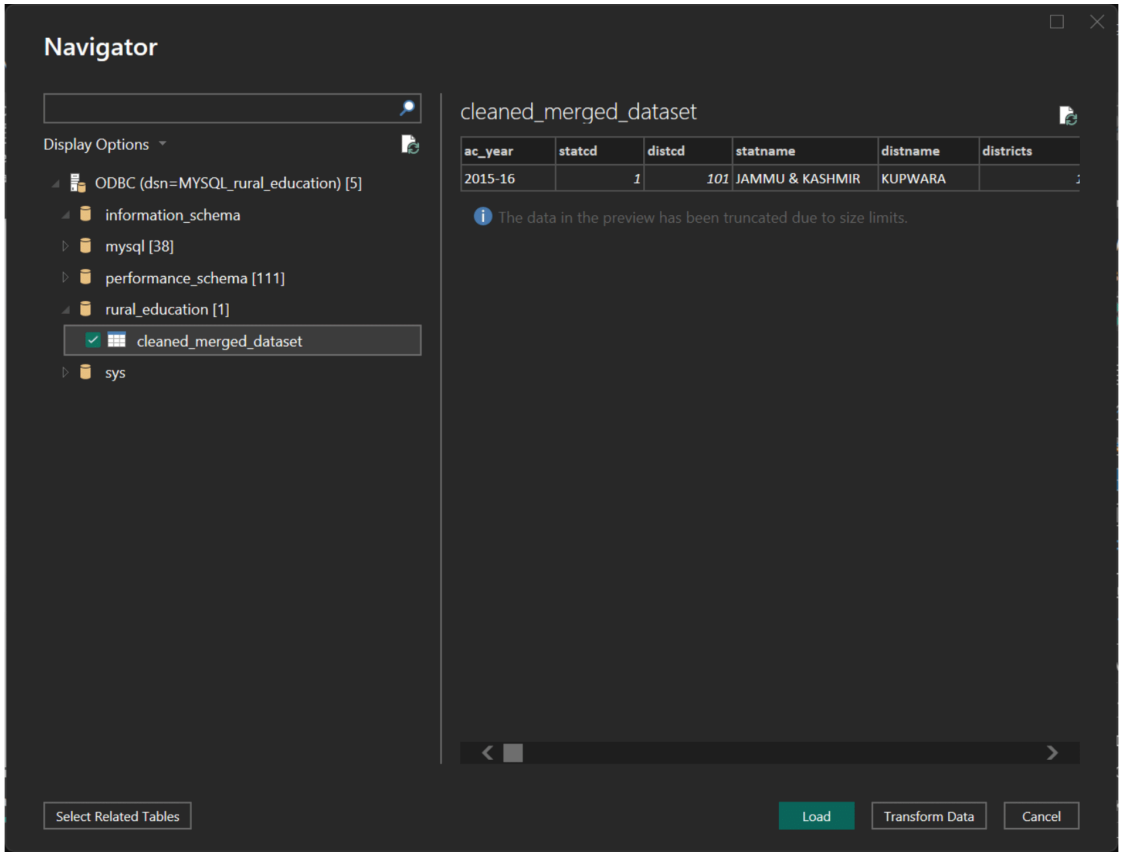
Table Data Import
Import Data
The following tasks will now be performed. Please monitor the execution.
Prepare Import
Import data file
Finished performing tasks. Click [Next >] to continue.
Show Logs
< Back
Next >
Cancel


Command Prompt - mysql -u
mysql> CREATE USER 'root'@'%' IDENTIFIED BY 'AK06@mca';
Query OK, 0 rows affected (0.04 sec)
mysql> GRANT ALL PRIVILEGES ON rural_education.* TO 'root'@'%';
Query OK, 0 rows affected (0.01 sec)
mysql> FLUSH PRIVILEGES;
Query OK, 0 rows affected (0.01 sec)
mysql>

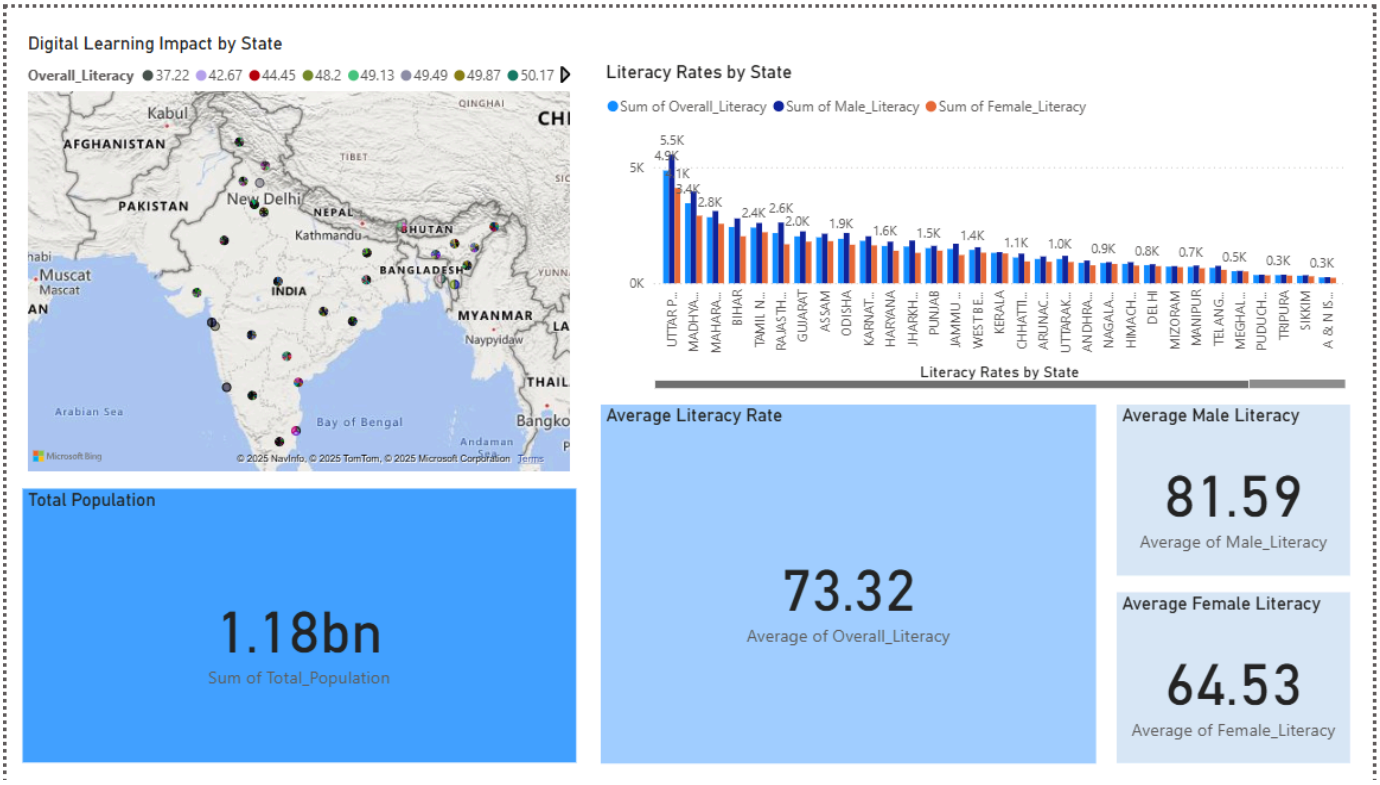

MySQL Connector/ODBC Data Source Configuration
MySQL
Connector/ODBC
Connection Parameters
Data Source Name:
MYSQL_rural_education
Description:
TCP/IP Server:
127.0.0.1
Port:
3306
Named Pipe:
User:
root
Password:
Database:
rural_education
Test
Details >>
OK
Cancel
Help

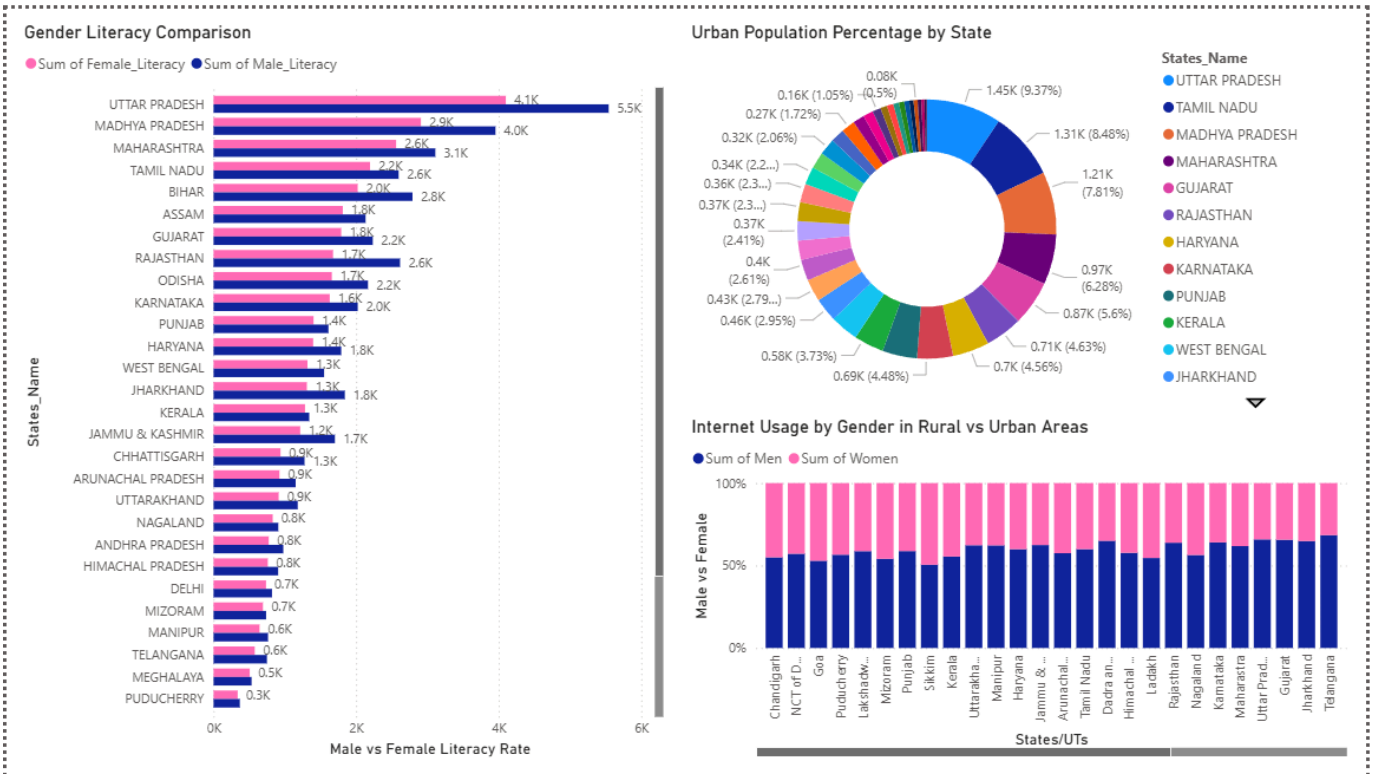

Test Result
Connection successful
OK


From ODBC
Data source name (DSN)
MYSQL_rural_education
Advanced options
OK
Cancel


Navigator
Display Options
ODBC (dsn=MYSQL_rural_education) [5]
information_schema
mysql [38]
performance_schema [111]
rural_education [1]
cleaned_merged_dataset
sys
cleaned_merged_dataset
ac_year
statcd
distcd
statname
distname
districts
2015-16
1
101
JAMMU & KASHMIR
KUPWARA
The data in the preview has been truncated due to size limits.
Select Related Tables
Load
Transform Data
Cancel

The following are screenshots of the Power BI dashboard:




Gender Literacy Comparison
Sum of Female_Literacy
Sum of Male_Literacy
States_Name
UTTAR PRADESH 4.1K 5.5K
MADHYA PRADESH 2.9K 4.0K
MAHARASHTRA 2.6K 3.1K
TAMIL NADU 2.2K 2.6K
BIHAR 2.0K 2.8K
ASSAM 1.8K
GUJARAT 1.8K 2.2K
RAJASTHAN 1.7K 2.6K
ODISHA 1.7K 2.2K
KARNATAKA 1.6K 2.0K
PUNJAB 1.4K
HARYANA 1.4K 1.8K
WEST BENGAL 1.3K
JHARKHAND 1.3K 1.8K
KERALA 1.3K
JAMMU & KASHMIR 1.2K 1.7K
CHHATTISGARH 0.9K 1.3K
ARUNACHAL PRADESH 0.9K
UTTARAKHAND 0.9K
NAGALAND 0.8K
ANDHRA PRADESH 0.8K
HIMACHAL PRADESH 0.8K
DELHI 0.7K
MIZORAM 0.7K
MANIPUR 0.6K
TELANGANA 0.6K
MEGHALAYA 0.5K
PUDUCHERRY 0.3K
0K 2K 4K 6K
Male vs Female Literacy Rate
Urban Population Percentage by State
States_Name
UTTAR PRADESH
TAMIL NADU
MADHYA PRADESH
MAHARASHTRA
GUJARAT
RAJASTHAN
HARYANA
KARNATAKA
PUNJAB
KERALA
WEST BENGAL
JHARKHAND
0.08K (0.5%)
0.16K (1.05%)
0.27K (1.72%)
0.32K (2.06%)
0.34K (2.2...)
0.36K (2.3...)
0.37K (2.3...)
0.37K (2.41%)
0.4K (2.61%)
0.43K (2.79...)
0.46K (2.95%)
0.58K (3.73%)
0.69K (4.48%)
0.7K (4.56%)
0.71K (4.63%)
0.87K (5.6%)
0.97K (6.28%)
1.21K (7.81%)
1.31K (8.48%)
1.45K (9.37%)
Internet Usage by Gender in Rural vs Urban Areas
Sum of Men
Sum of Women
Male vs Female
100%
50%
0%
Chandigarh
NCT of D...
Goa
Puducherry
Lakshadw...
Mizoram
Punjab
Sikkim
Kerala
Uttarakha...
Manipur
Haryana
Jammu & ...
Arunachal...
Tamil Nadu
Dadra an...
Himachal ...
Ladakh
Rajasthan
Nagaland
Karnataka
Maharastra
Uttar Prad...
Gujarat
Jharkhand
Telangana
States/UTs
Gender Literacy Gap by States_Name
Gender Literacy Gap
30
20
10
0
RAJASTHAN
JHARKHAND
JAMMU & KA...
CHHATTISGARH
MADHYA PRA...
UTTAR PRADE...
UTTARAKHAND
DADRA & NA...
BIHAR
ODISHA
HARYANA
TELANGANA
GUJARAT
ANDHRA PRA...
MAHARASHTRA
KARNATAKA
HIMACHAL PR...
ARUNACHAL ...
TAMIL NADU
MANIPUR
DAMAN & DIU
WEST BENGAL
ASSAM
SIKKIM
GOA
PUNJAB
A & N ISLANDS
TRIPURA
DELHI
States Name
17.07
Gender Literacy Gap
73.32
Average Literacy Rate
16.94
Gender_Internet_Gap
Male vs Female Internet Users by Area
Avg_Male_Internet
Avg_Female_Internet
60
40
20
0
Urban
Total
Rural
Area

The following screenshots illustrate the DAX functions used in Power BI:

```
Avg_Male_Internet = AVERAGE('Internet_Usage_Rural_vs_Urban'[men])
```

```
Average Literacy Rate = AVERAGE('cleaned_merged_dataset'[Overall_Literacy])
```

```
Gender_Internet_Gap =
[Avg_Male_Internet] - [Avg_Female_Internet]
```

```
Gender Literacy Gap = AVERAGE('cleaned_merged_dataset'[Male_Literacy]) - AVERAGE('cleaned_merged_dataset'
[Female_Literacy])
```

Chapter 5.

Testing

Testing is a critical phase in the Software Development Life Cycle (SDLC) that ensures the reliability, accuracy, and effectiveness of the data analytics project. It involves validating the data, checking the functionality of MySQL database queries, Power BI dashboards, and ensuring the overall integrity of the analysis. The key testing phases include:

1. **Data Validation and Integrity Testing:**
    - Ensure data accuracy and consistency across different datasets (e.g., Education in India, Internet Usage Rural vs Urban, Govt of India Literacy Rate).
    - Verify data types, missing values, and outliers.
    - Check for proper data merging, aggregation, and relationship integrity within MySQL before connecting to Power BI.
    - Validate that data imported from MySQL into Power BI is correctly mapped and formatted.
2. **Functional Testing:**
    - Test each Power BI dashboard element (charts, graphs, maps, filters) to ensure they display the correct data from the MySQL database.
    - Validate that interactive features like slicers, drill-through, and filters work as intended with live MySQL data.
    - Ensure accurate calculations for KPIs like literacy rates, student-teacher ratios, and dropout rates, including verifying MySQL queries for data accuracy.

3. **Performance Testing:**
    - Assess the loading time and responsiveness of Power BI dashboards connected to the MySQL database.
    - Optimize MySQL data queries to reduce processing time and improve dashboard speed.
    - Test for scalability with larger datasets in the MySQL database to ensure future data expansion is supported.
4. **Usability Testing:**
    - Check the overall design, layout, and user experience of the Power BI dashboards.
    - Ensure data is presented in a clear, understandable, and visually appealing format, with effective MySQL data integration.
    - Collect user feedback to identify areas for improvement in the dashboard design and MySQL data organization.
5. **Security and Data Privacy Testing:**
    - Verify that sensitive data in the MySQL database is protected, and access controls are properly configured.
    - Ensure data sources are secure and comply with data privacy regulations.
    - Test MySQL user roles and permissions to prevent unauthorized access.
6. **Cross-Device Compatibility Testing:**
    - Test the dashboard on different devices (PC, tablet, mobile) to ensure consistent performance with live data from MySQL.
    - Check for responsive design and mobile-friendliness in the Power BI dashboards.
7. **Regression Testing:**
    - After each update, ensure that previously tested functionalities still work as expected.

- Validate that new features or data updates in the MySQL database do not break existing functionality in Power BI.

8. **User Acceptance Testing (UAT):**
    - Conduct final testing with real-world users, such as students, teachers, or education policymakers.
    - Gather feedback to ensure the project meets user expectations and project goals, including verifying MySQL data accuracy and Power BI dashboard effectiveness.

Chapter 6.

Application

This project aims to bridge the digital divide in education by leveraging data to uncover insights that can drive meaningful change in rural education systems. The application of this project extends across multiple domains:

1. **Educational Policy and Planning:**
    - *Data-Driven Decisions:* Helps policymakers identify regions with low digital penetration and high dropout rates, guiding resource allocation for digital infrastructure.
    - *Impact Assessment:* Assesses the effectiveness of government initiatives like Digital India and PM eVidya in rural areas.
    - *Funding Prioritization:* Supports data-backed funding decisions for digital education projects in underserved regions.
2. **Educational Institutions and NGOs:**
    - *Student Retention Programs:* Identifies factors contributing to dropout rates, helping schools and NGOs design targeted retention strategies.
    - *Curriculum Enhancement:* Provides insights into the digital readiness of students, guiding curriculum adjustments to include more digital literacy components.
    - *Teacher Training:* Highlights regions with high student-teacher ratios, encouraging training programs for educators in digital teaching methods.

3. **Technology Providers and EdTech Startups:**
    - *Market Expansion:* Identifies potential markets for digital learning tools and platforms.
    - *Product Development:* Guides the development of region-specific digital learning solutions based on connectivity and digital readiness.
4. **Academic Research and Advocacy:**
    - *Research Studies:* Provides a foundation for further research on digital learning's impact on educational outcomes.
    - *Advocacy and Awareness:* Supports NGOs and advocacy groups in promoting digital literacy and bridging the rural-urban education gap.
5. **Student and Teacher Empowerment:**
    - *Skill Development:* Encourages digital skill development among rural students, preparing them for a tech-driven job market.
    - *Digital Inclusion:* Promotes educational equity by providing insights into areas with limited digital access.

Project GitHub Link

Chapter 7.

Conclusion

This project successfully highlights the critical role of digital learning in transforming rural education in India, aligning with Sustainable Development Goal 4 (SDG 4) - Quality Education. Through comprehensive data analysis using datasets like Education in India, Internet Usage Rural vs Urban, and Govt of India Literacy Rate, it provides valuable insights into the challenges and opportunities for digital education in rural areas.

The analysis reveals significant disparities in digital access and educational outcomes, emphasizing the need for targeted interventions to bridge the digital divide. It underscores the importance of internet connectivity, digital literacy, and educational infrastructure in improving enrollment rates, reducing dropout rates, and enhancing overall learning outcomes. MySQL has been used as the database to centralize the data management, allowing for efficient querying and aggregation of datasets to draw meaningful insights.

The project's Power BI dashboard offers an interactive platform for visualizing these insights, empowering policymakers, educators, and NGOs to make data-driven decisions. The real-time connection to MySQL ensures that the dashboard is continuously updated with the latest data, providing up-to-date information for stakeholders.

By identifying key performance indicators (KPIs) such as student enrollment rates, teacher-student ratios, and digital access levels, this project provides a roadmap for enhancing digital education in rural India. It also highlights the potential for data

analytics to drive impactful educational reforms, fostering a more inclusive and equitable learning environment for all.

Moving forward, continuous data updates, real-time monitoring via MySQL integration, and localized policy interventions can further strengthen the impact of digital learning, ensuring that no student is left behind in the digital era. Through ongoing analysis and dashboard updates, the project lays the foundation for future educational initiatives aimed at bridging gaps and improving digital access across rural India.

Chapter 8.

Bibliography

1. **National Sample Survey Office (NSSO).** (2021). *Education in India - NSS 75th Round (2017-18)*. Ministry of Statistics and Programme Implementation, Government of India.
2. **Internet and Mobile Association of India (IAMAI).** (2022). *Internet Usage Rural vs Urban*. Internet and Mobile Association of India.
3. **Government of India.** (2021). *Literacy Rate in India - Census 2011 and 2021 (Provisional)*. Office of the Registrar General & Census Commissioner, India.
4. **United Nations.** (2015). *Sustainable Development Goals - Goal 4: Quality Education*. United Nations Department of Economic and Social Affairs.
5. **World Bank.** (2020). *Digital Learning for Every Child in India - Challenges and Opportunities*. World Bank Publications.
6. **Power BI Documentation.** (2024). *Getting Started with Power BI*. Microsoft Power BI.