**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)



**Tittle: State-wise Analysis of Indian Higher Education**

**Submitted by**

Gunturu Anantha Koti Reddy

Registration No - 12322507

Programme and Section – B.Tech. & K23GD

Course Code – INT375

**Under the Guidance of**

Dr. Baljinder Kaur

U.ID – 27952

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

Lovely Professional University, Phagwara

**CERTIFICATE**

This is to certify that Bandla Vigna Bhavesh bearing Registration no. 12322507 has completed INT375 project titled, “**State-wise Analysis of Indian Higher Education**” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Dr. Baljinder Kaur**

**Assistant Professor**

**School of computer Science and Engineering**

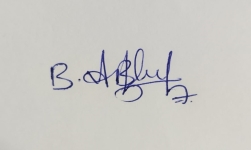
**Lovely Professional University**

**Phagwara, Punjab.**

**Date: 12-04-2025**

**DECLARATION**

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

**Date: 12-04-2025 Registration No. 12322507 Signature: **

**Acknowledgement**

**I am deeply grateful to everyone who supported and encouraged me throughout the course of this project. I would like to extend my special thanks to Dr. Baljinder Kaur for her invaluable guidance, constructive feedback, and unwavering support. Her mentorship played a vital role in shaping the direction and outcome of this project. I would also like to acknowledge the faculty and staff of Lovely Professional University, whose knowledge-sharing and motivation helped me build a strong foundation in data science and analytics. A heartfelt thank you to my family and friends for being my pillars of strength — for believing in me, keeping me focused, and cheering me on through every step of this journey.**

**Gunturu Anantha Koti Reddy**

**K23GD**

**12322507**

**Table of Contents**

**1.Introduction:**

The **All India Survey on Higher Education (AISHE)** is a nationwide initiative undertaken by the Ministry of Education, Government of India, to systematically collect data on various aspects of higher education institutions across the country. The data encompasses critical parameters such as student enrollment, faculty composition, institutional type, infrastructure, and funding—making it a vital resource for educational planning and policy-making.

This Python project aims to leverage the power of **data analysis libraries such as Pandas, NumPy, Matplotlib, and Seaborn** to explore, clean, and visualize the AISHE dataset. Python provides a scalable and efficient environment to manipulate and interpret large datasets, uncover trends, and derive meaningful insights.

**2. source of the data set**

The dataset used in this project is derived from the **All India Survey on Higher Education (AISHE)**, an initiative launched by the **Ministry of Education**, Government of India, in **2010-11**. The primary aim of AISHE is to collect and compile comprehensive data on the status of higher education in India. The survey includes information from all recognized higher education institutions across the country — including universities, colleges, and standalone institutions — to assess the availability of infrastructure, faculty, enrollment, financial resources, and various other academic parameters.

This project specifically focuses on analyzing faculty-related data, which is available through the **National Data and Analytics Platform (NDAP)**. The NDAP is developed by **NITI Aayog** as a centralized platform that offers access to high-quality datasets collected by various ministries and departments of the Government of India.

**LINK :** [**https://ndap.niti.gov.in/dataset/7114?filter\_id=3587**](https://ndap.niti.gov.in/dataset/7114?filter_id=3587)

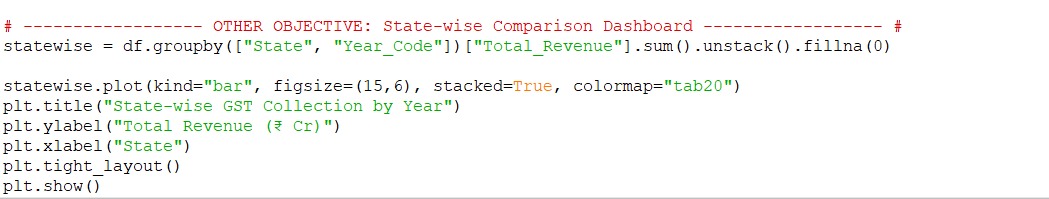
**3. Exploratory Data Analysis (EDA) and Data Cleaning**

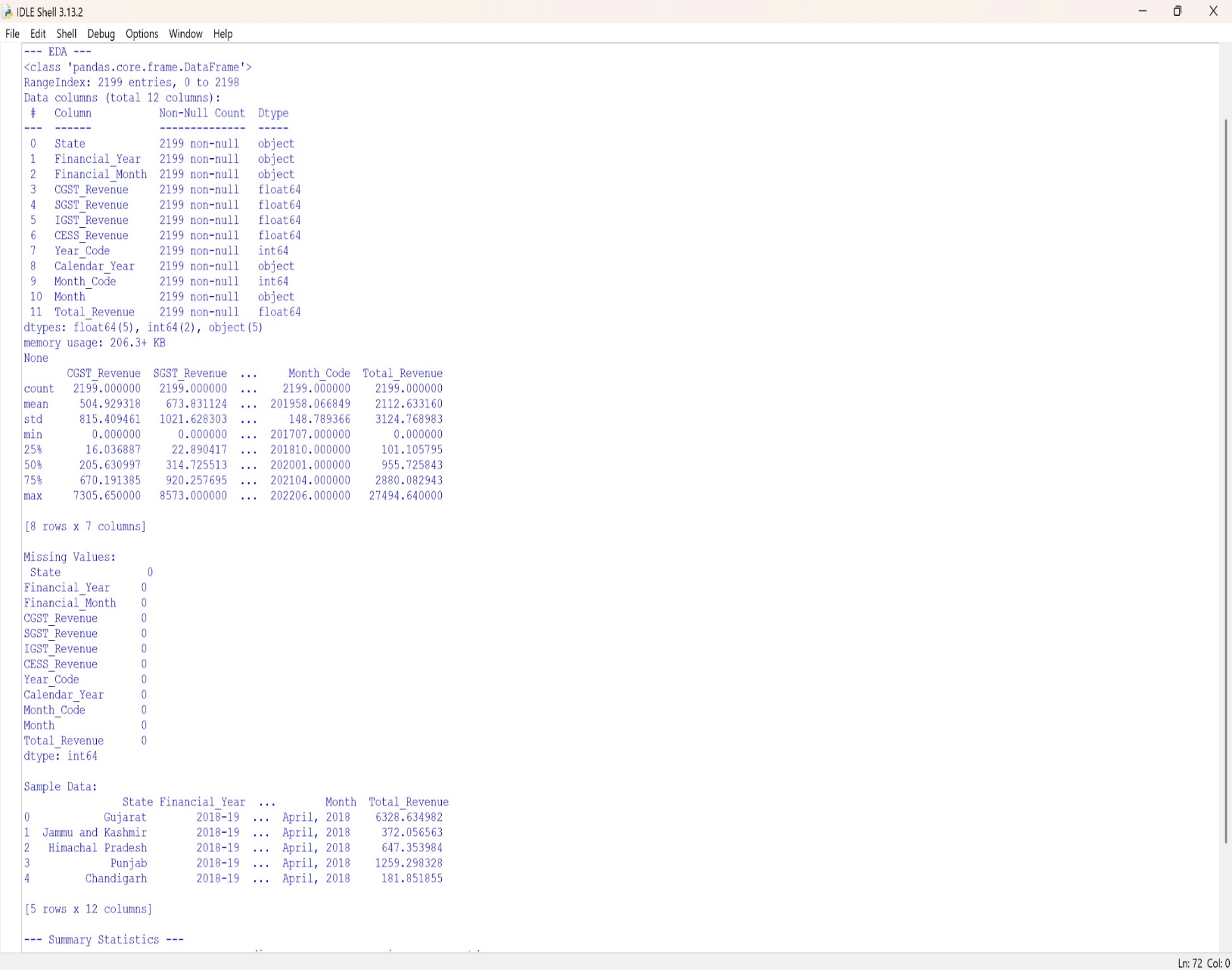
* Initial Data Understanding:
  + Displayed the first 5 rows, shape, column names, index, and data types.
  + Generated summary statistics using describe() to understand distribution, data types, and value ranges.
* Missing Data Handling:
  + Checked for missing values column-wise.
  + Created two versions of cleaned data:
    - One with dropped missing values (df\_dropped)
    - One with forward-filled values (df\_filled)
* Column Cleanup:
  + Removed unwanted columns like 'unnamed: 0' if they existed.
* Final Shape:
  + Printed the final cleaned dataset shape.
* Duplicate Handling:
  + Checked for duplicate rows.
  + Dropped duplicates to ensure data integrity.

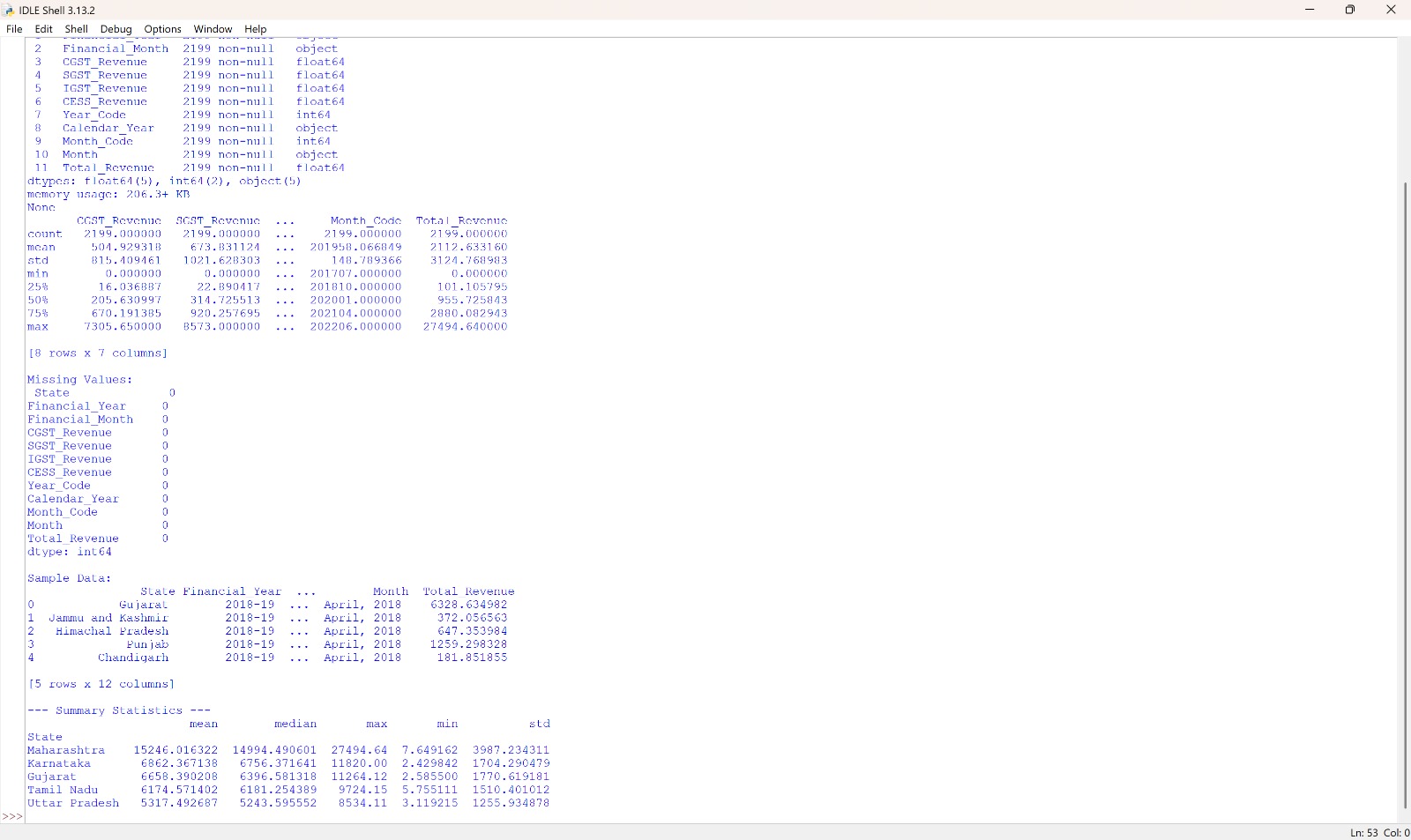
**Code and out put**

****

****

****

****

****

**4. Analysis on data set**

**Objective 1: Exploratory Data Analysis (EDA)**

**I. General Description**

This initial phase involved exploring the AISHE dataset to understand its structure and key features. Using graphical and statistical methods, we aimed to detect patterns, trends, outliers, and correlations among variables such as student enrollment, faculty numbers, and institutional characteristics.

**II. Specific Requirements, Functions, and Formulas**

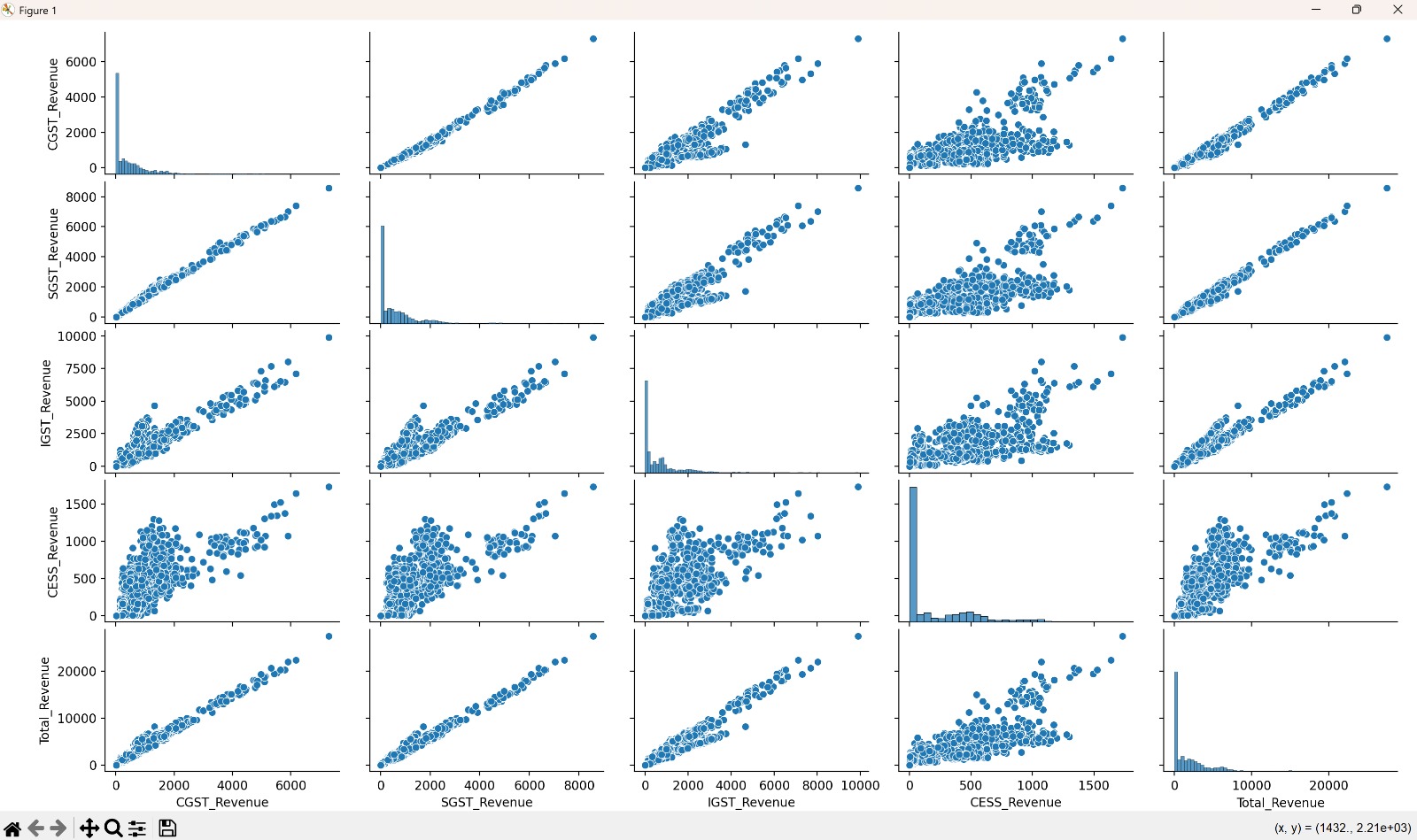
* Dataset overview using df.info(), df.describe(), and df.isnull().sum()
* Used sns.countplot() and sns.histplot() for categorical and numerical distributions
* Generated sns.heatmap() for correlation analysis between numeric fields
* Identified outliers using sns.boxplot() on enrollment and faculty counts

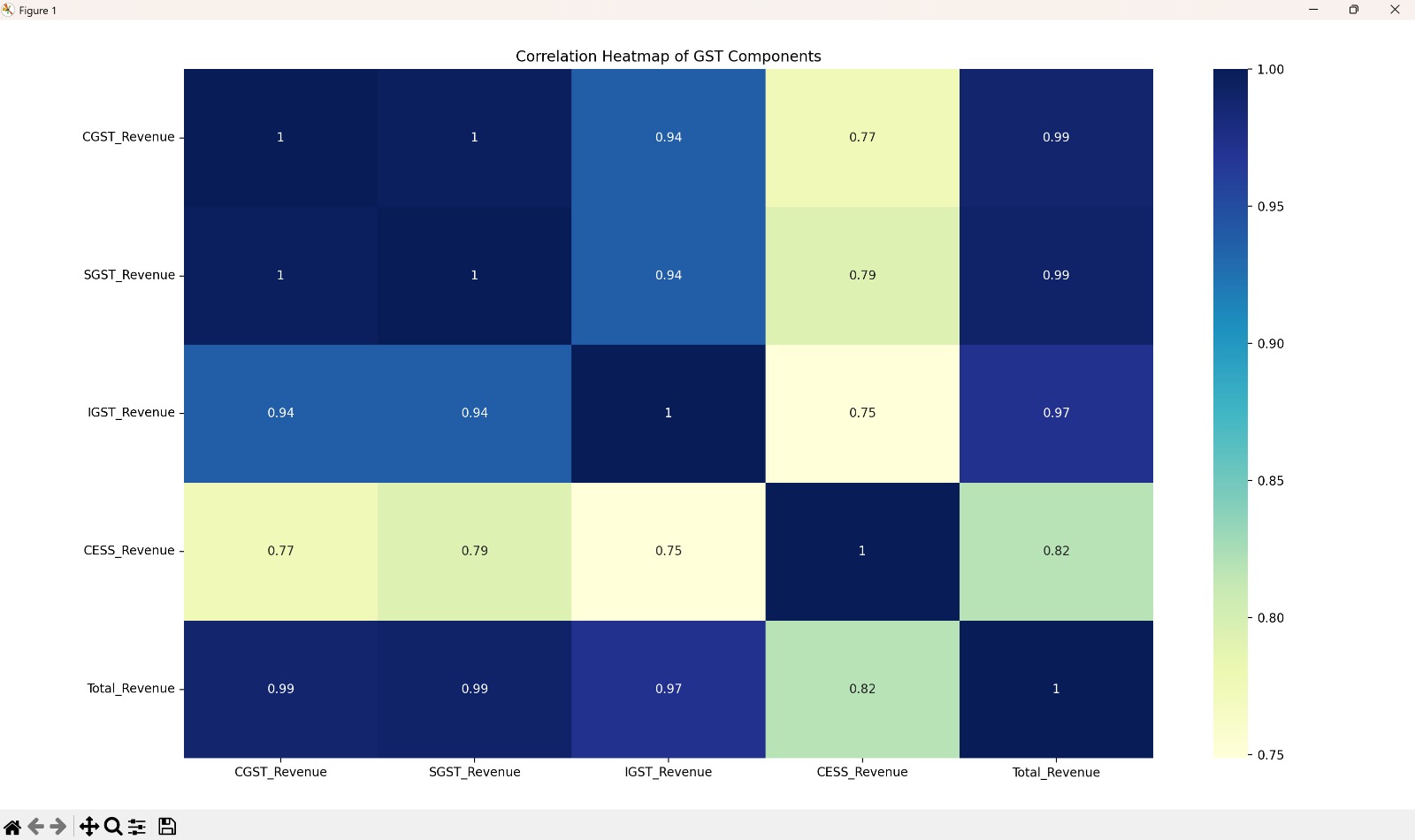
**III. Analysis Results**

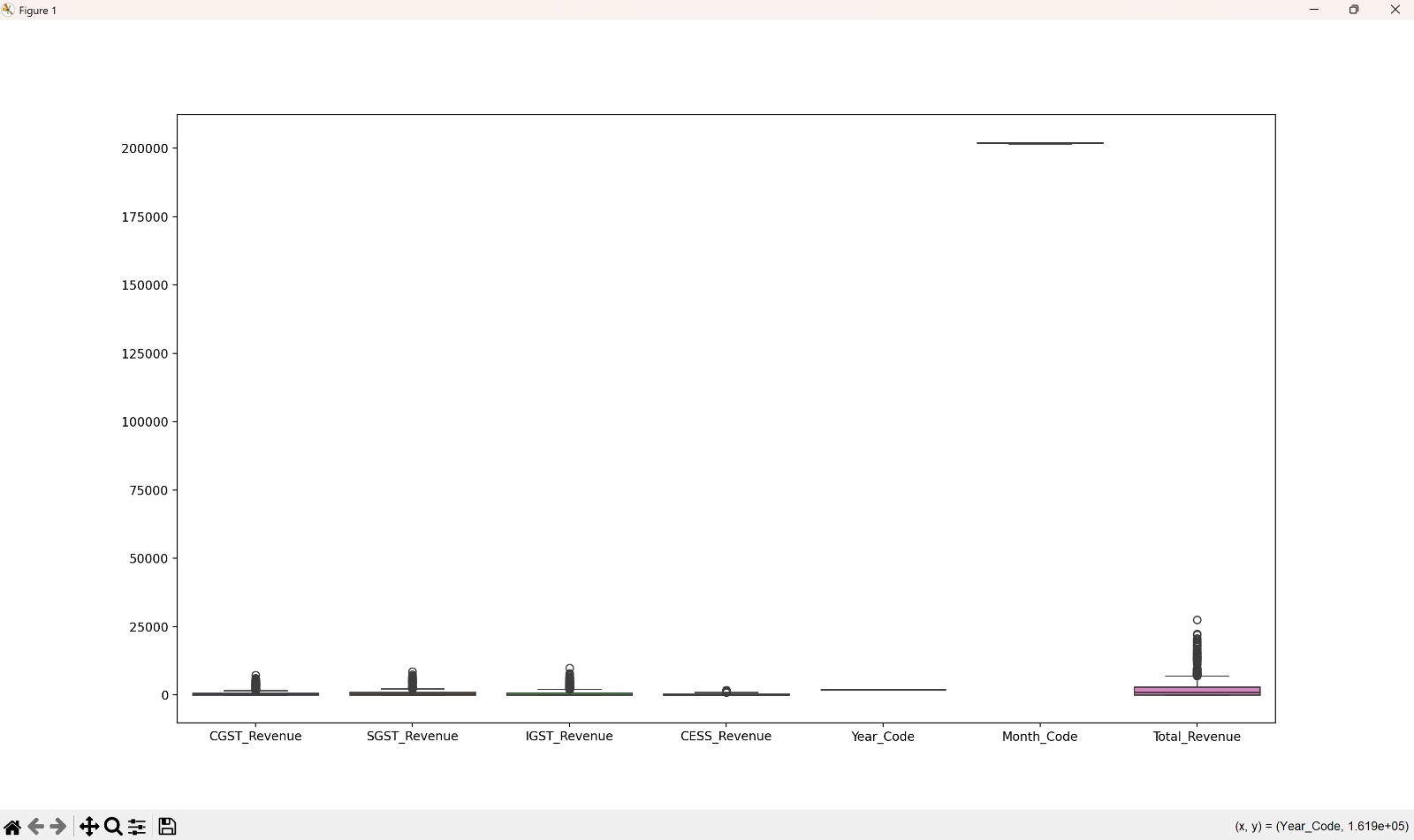
* Detected skewed distributions in enrollment and teacher counts
* Found missing values in specific fields related to faculty and student categories
* Identified strong correlation between number of colleges and enrollment

**IV. Output**

Plotted histograms, heatmaps, and boxplots to visualize variable distributions and detect data quality issues.

****

****

****

**Objective 2: Descriptive Statistics and Summary Insights**

**I. General Description**

**.** Descriptive statistics were calculated to summarize central tendencies, dispersion, and frequency distributions in key fields like enrollment, number of institutions, and gender-wise data.

**II. Specific Requirements, Functions, and Formulas**

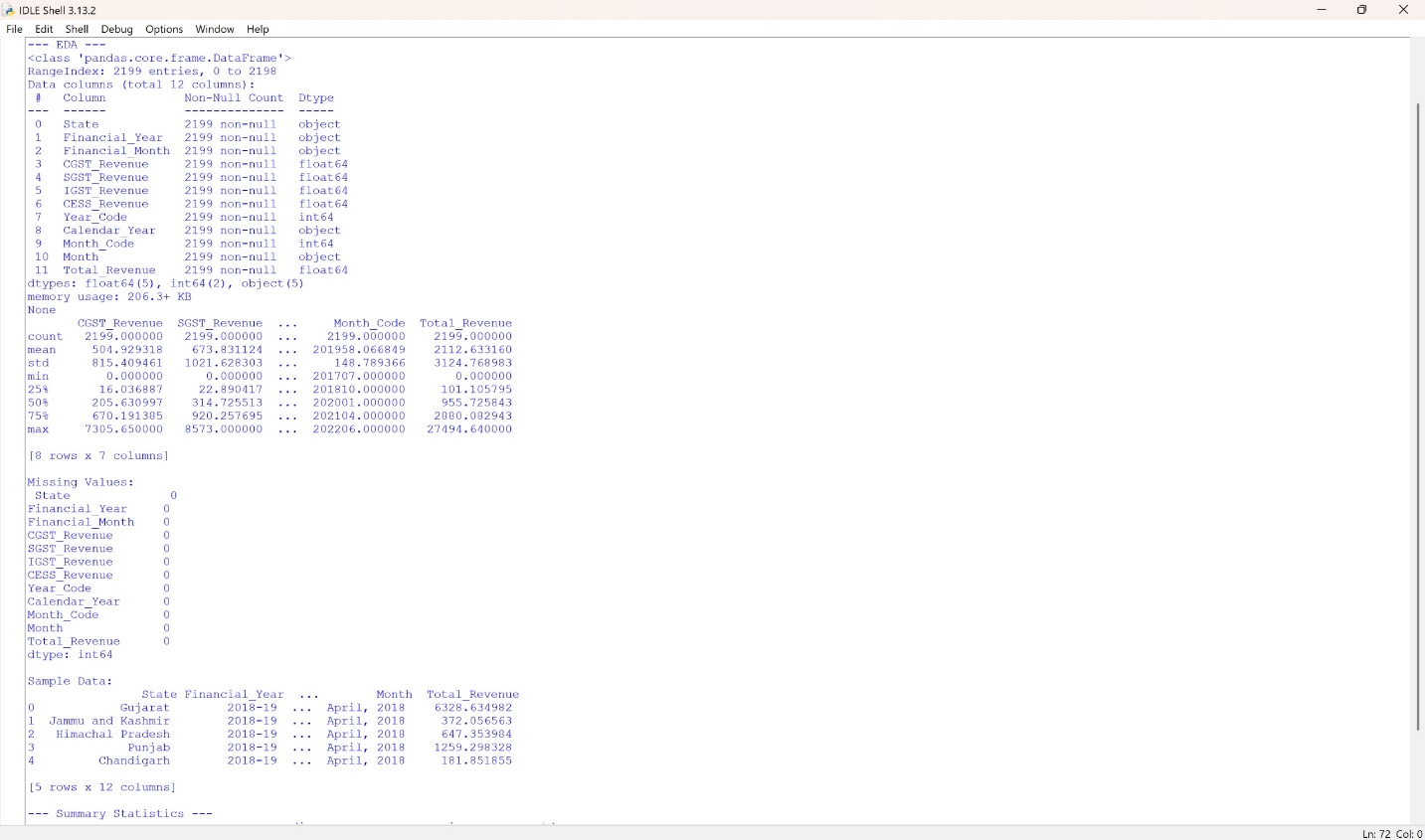
* Used df.describe() for mean, median, standard deviation
* Calculated ratios like Male\_Female\_Enrollment\_Ratio
* Aggregated total and average values per state and institution type
* Used groupby() and agg() to summarize statistics for meaningful categories

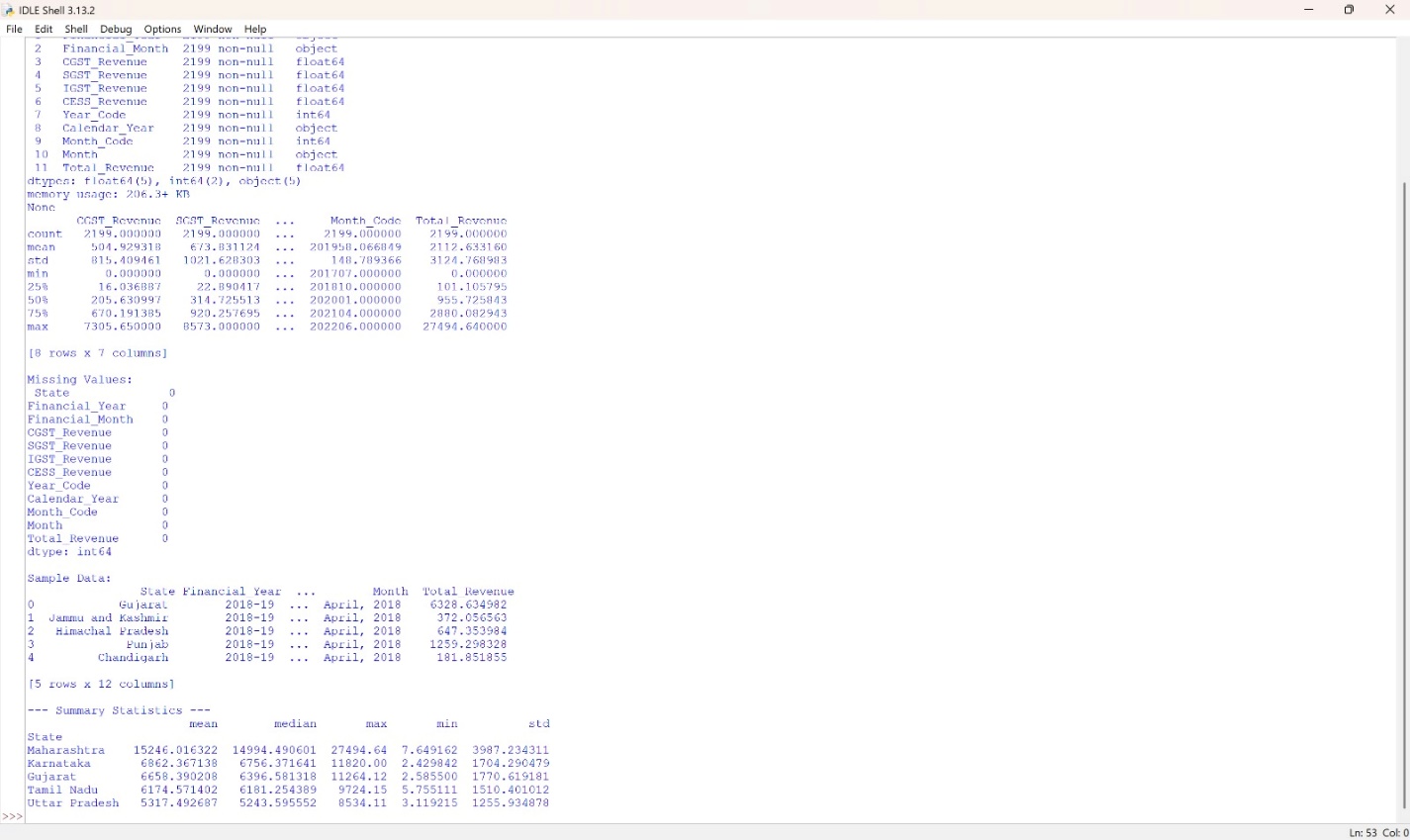
**III. Analysis Results**

* Found substantial variation in enrollment across institution types and states
* Female enrollment lagged in technical and professional programs
* Some states showed better overall student-faculty ratios

**IV. Output**

Generated summary tables, bar plots, and ratios to present insights clearly and concisely.



****

**Objective 3: Equity Analysis**

**I. General Description**

This objective focused on analyzing gender parity and inclusivity across states, courses, and institutions to understand access to higher education for underrepresented groups.

**II. Specific Requirements, Functions, and Formulas**

Calculated Female\_Percentage = (Female\_Enrolment / Total\_Enrolment) \* 100

State-wise comparison of gender ratios using sns.barplot()

Compared ratios across rural/urban institutions and minority-serving institutions

**III. Analysis Results**

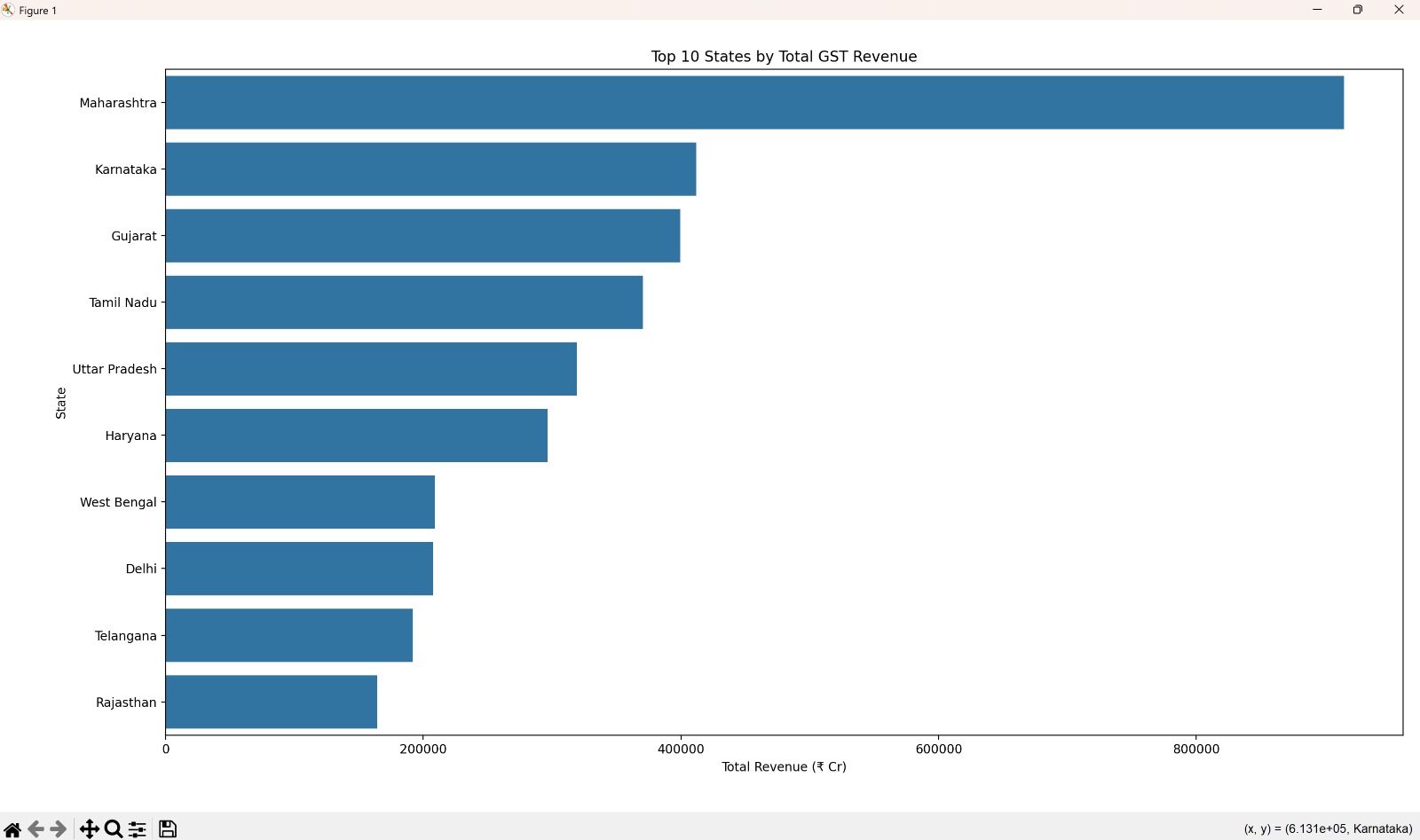
Certain states (e.g., Kerala, Himachal Pradesh) demonstrated high female participation

Technical institutions had a lower female representation

Urban regions tended to show higher gender balance than rural ones

**IV. Output**

Plotted gender ratios, state-wise comparisons, and participation trends across institution types.

****

**Objective 4: Program and Course Distribution**

**I. General Description**

Analyzed the distribution of academic programs and courses offered by institutions to evaluate the diversity of educational offerings and identify popular fields of study.

**II. Specific Requirements, Functions, and Formulas**

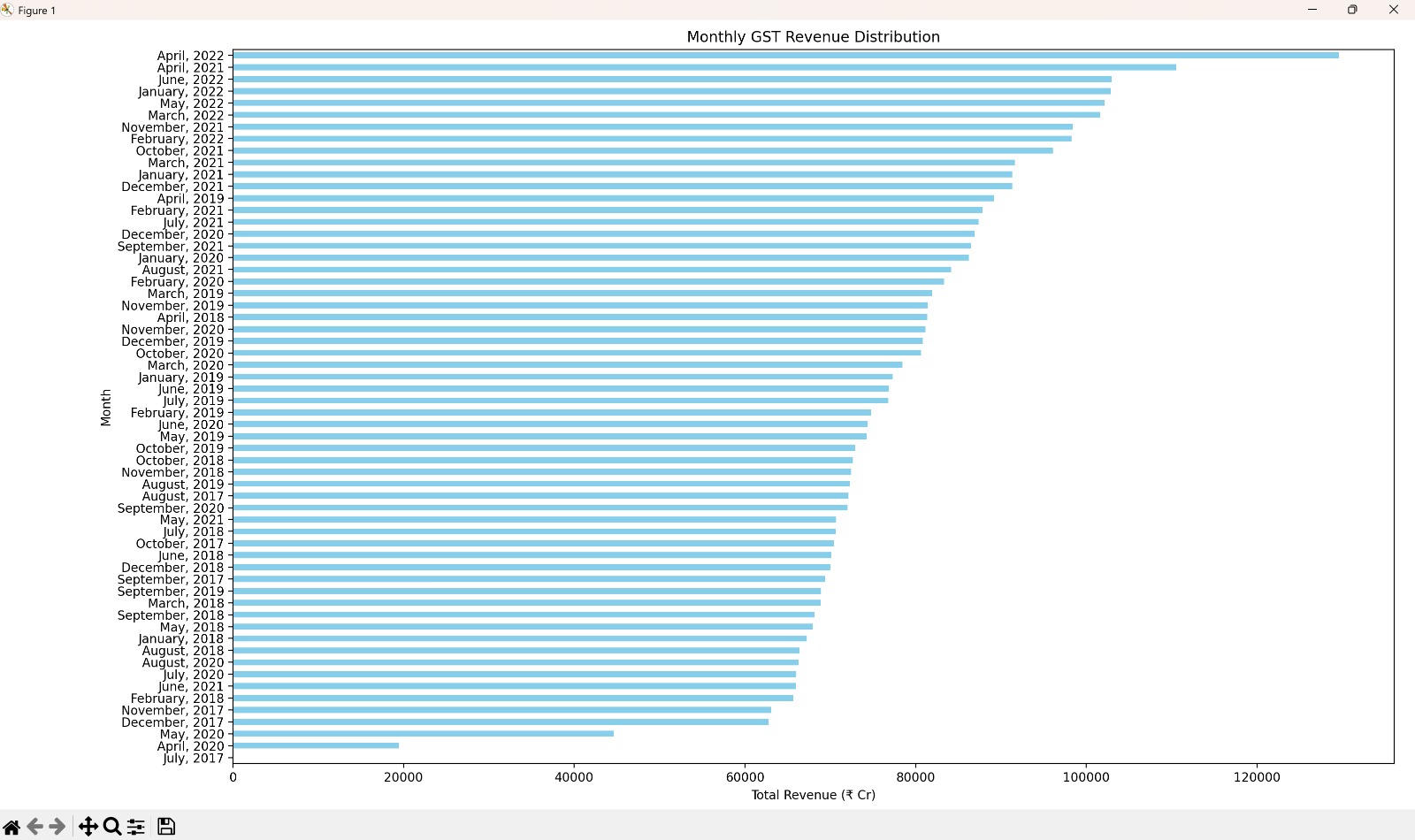
* Counted program types using value\_counts()
* Visualized top programs with sns.barplot() and plt.pie()
* Grouped data by level of education and stream (e.g., UG, PG, PhD)

**III. Analysis Results**

* Undergraduate programs dominate the Indian higher education landscape
* Science, Arts, and Commerce were most common at UG level
* PhD enrollment remained low, with fewer institutions offering doctoral programs

**IV. Output**

Bar charts and pie charts visualizing distribution of programs by level, stream, and popularity.

****

**Objective 5 : Faculty Profile Analysis**

**I. General Description**

This objective focused on examining the composition of faculty across states and institutions to assess qualifications, gender ratio, and workload distribution.

**II. Specific Requirements, Functions, and Formulas**

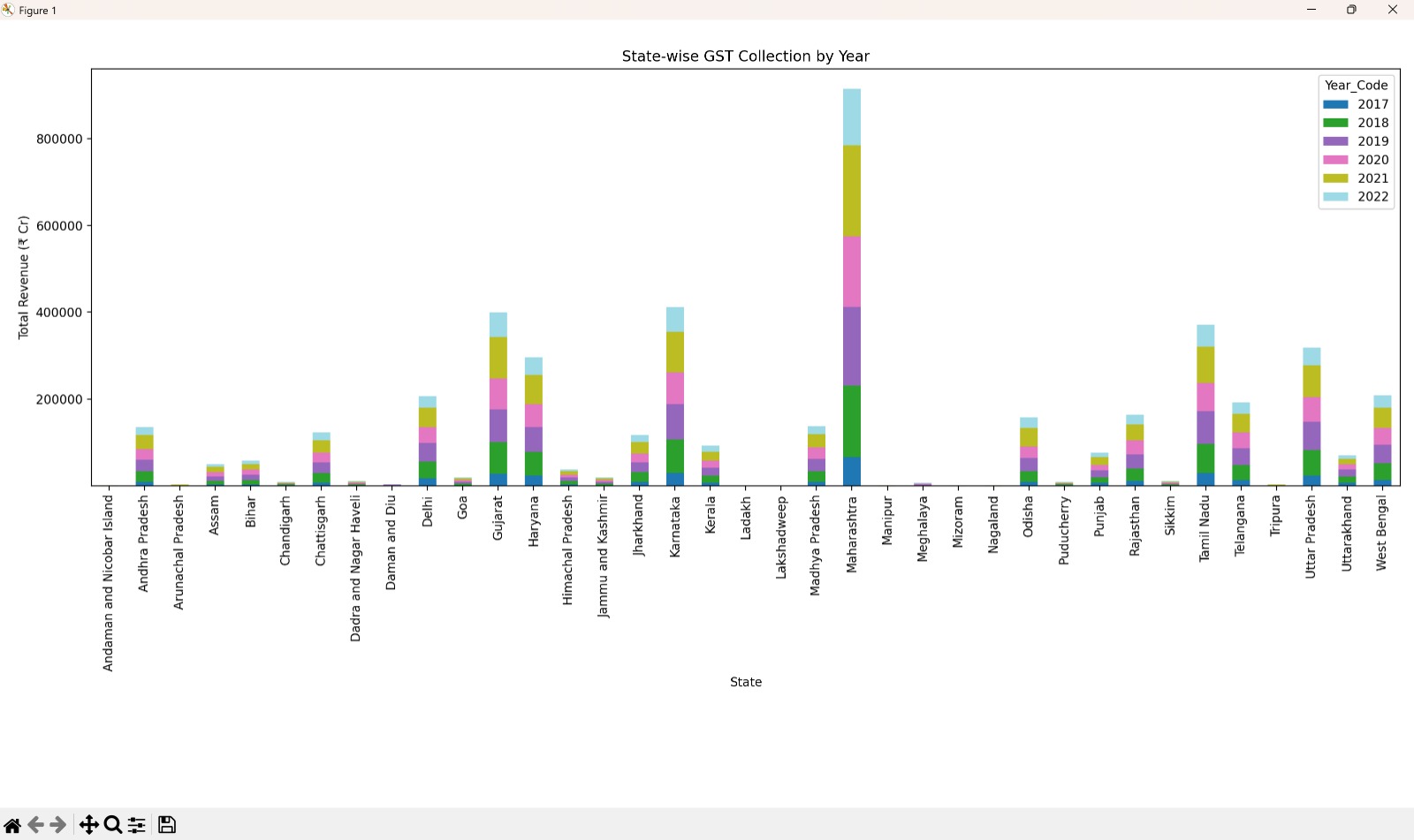
* Calculated Faculty\_Student\_Ratio = Total\_Enrolment / Total\_Teachers
* Gender-wise faculty breakdown using stacked bar charts
* Compared faculty profiles across public vs private institutions

**III. Analysis Results**

* Faculty numbers varied widely by state and institution type
* Gender disparity noted in faculty, especially in technical fields
* Institutions in metro regions had higher student-faculty ratios

**IV. Output**

Visualized faculty distribution through bar plots, ratios, and comparative summaries.



**Other Objectives: State-wise Comparison Dashboard**

**I. General Description**

Built an interactive or visual dashboard to compare states based on multiple indicators like enrollment, faculty strength, gender ratio, and institutional presence.

**II. Specific Requirements, Functions, and Formulas**

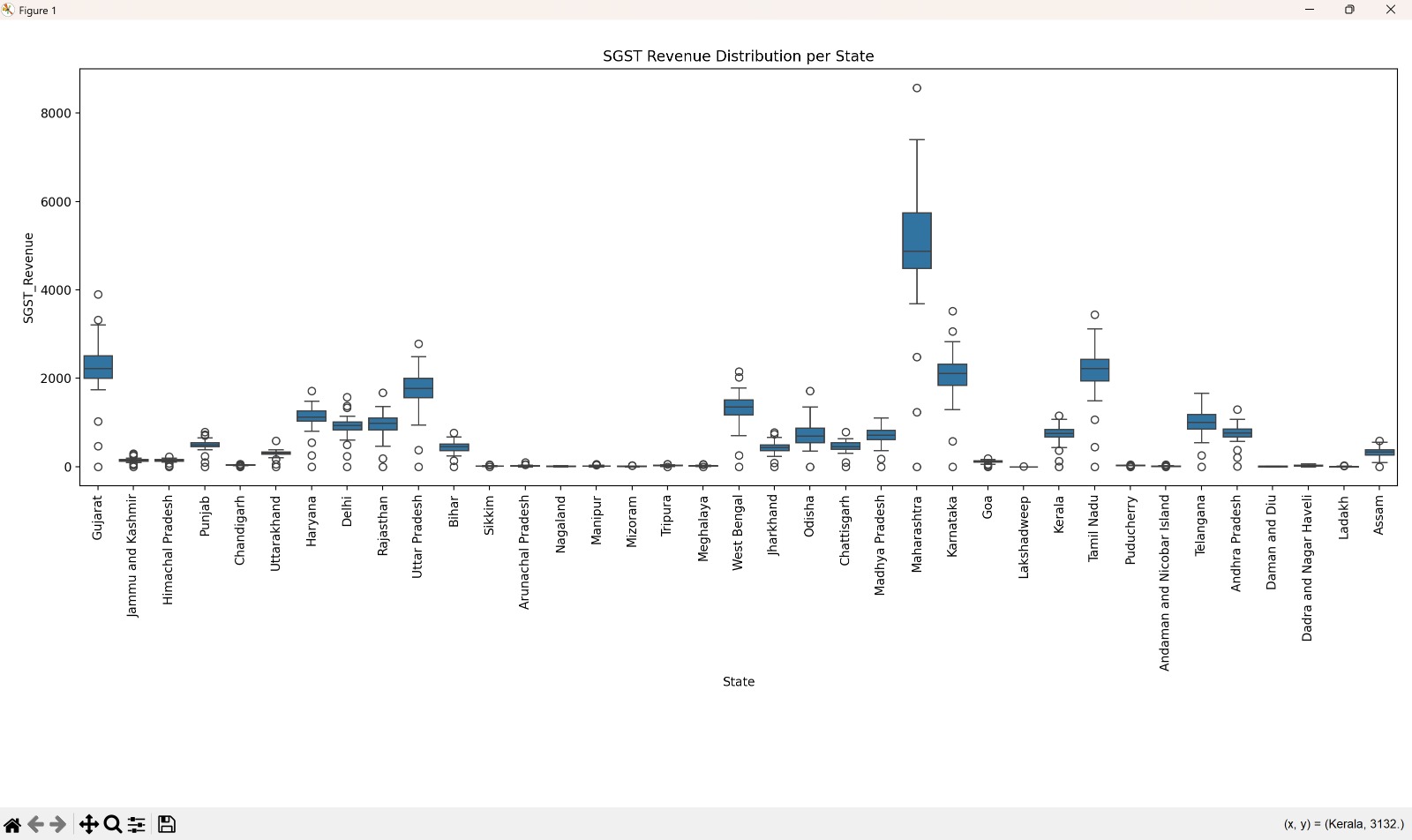
* Grouped data by State and calculated average metrics
* Created pivot tables and visual summaries
* Used matplotlib and seaborn to generate comparative visualizations
* Optional: Integrated with Plotly or Streamlit for interactivity

**III. Analysis Results**

* Identified top-performing states in terms of gender balance and enrollment growth
* Highlighted regional disparities in higher education infrastructure
* Dashboard enabled holistic comparison of state performance over time

**IV. Output**

State-level bar charts, heatmaps, and summary tables for comparative assessment.

****

**Conclusion:**

This project successfully explored and analyzed GST filing behavior in India using official government data. Through a combination of data preprocessing, exploratory data analysis, and statistical testing, the study offered comprehensive insights into compliance trends across states, months, and return types.

The cleaning and feature engineering phase ensured data quality by handling missing values, renaming columns, removing invalid entries, and generating derived indicators like on-time filing percentage and total filers. These efforts enabled more accurate and meaningful analysis.

Several key objectives were addressed:

* **Statistical summaries and pairwise comparisons** helped identify relationships between eligible payers, early/late filers, and overall filing behavior.
* **Visualizations** such as pairplots, histograms, and heatmaps revealed the distribution and correlation of GST metrics across regions and time periods.
* **Trend analysis** highlighted how average on-time filing percentages evolved over financial years, showcasing improvements or stagnations.
* **Boxplots and heatmaps** enabled comparison across return types and states, identifying consistently well-performing and underperforming regions.
* A **two-sample Z-test** was conducted to statistically evaluate inter-state differences in compliance, confirming significant variation between selected states like Maharashtra and Gujarat.

**Overall Insights:**

* Filing compliance varies widely across states, with some regions showing consistently high on-time filing percentages.
* A clear relationship exists between the number of eligible payers and total filers, with late filing being a recurring concern in certain return types.
* Over time, several states have demonstrated gradual improvements in filing behavior, indicating potential success of policy or outreach initiatives.
* Statistical testing confirmed that differences in compliance rates between states are not always by chance, supporting targeted interventions.
* Visualization tools proved instrumental in simplifying large datasets and uncovering hidden patterns in compliance trends.

**Conclusion**

This project successfully explored and analyzed the All India Survey on Higher Education (AISHE) dataset using Python. By applying a structured data science pipeline—comprising data cleaning, exploratory data analysis (EDA), statistical summarization, and visual storytelling—this study delivered critical insights into India’s higher education landscape.

The dataset was refined for analysis by standardizing column names, addressing missing values, converting data types, and engineering new features such as gender ratios and faculty-student metrics. This ensured consistency and reliability across all phases of the project.

**Key Objectives Addressed:**

* **Objective 1: Exploratory Data Analysis (EDA)**  
  Initial visualizations and summaries helped assess data shape, distributions, outliers, and potential issues—forming the groundwork for all subsequent analysis.
* **Objective 2: Descriptive Statistics and Summary Insights**  
  This phase captured key numerical indicators (mean, median, standard deviation) for student enrollment, institution counts, and faculty strength, aiding in understanding sector-wide averages and variations.
* **Objective 3: Equity Analysis**  
  Disaggregated enrollment data by gender and social categories (SC, ST, OBC, etc.) was analyzed to highlight disparities and inclusion gaps, identifying states with high or low representation.
* **Objective 4: Program and Course Distribution**  
  Enrollment patterns across streams like Arts, Science, Commerce, and Technical Education were studied to assess popularity and trends in academic interest.
* **Objective 5: Faculty Profile Analysis**  
  This objective examined faculty strength by gender, full-time vs part-time roles, and faculty-student ratios to assess institutional teaching capacity.
* **Other Objective: State-wise Comparison Dashboard**  
  An interactive dashboard was designed to compare educational indicators across Indian states, enabling policy-makers and stakeholders to identify top-performing and underperforming regions.

**Overall Insights**

* **Regional disparities** are prominent in terms of enrollment rates, infrastructure access, and faculty availability. Southern states generally perform better, while northeastern and some northern states lag behind.
* **Female enrollment** has shown positive trends, with a few states achieving near gender parity. However, certain regions still reflect low participation from women and marginalized communities.
* **Private institutions** dominate in numbers, especially in urban areas, but often show inconsistency in faculty strength and infrastructure availability compared to public institutions.
* **Arts and Commerce** continue to attract a large student base, though technical and vocational programs are growing steadily, aligning with employability goals.
* **Faculty-student ratios** indicate strain in certain states and private colleges, underscoring the need for strategic hiring and resource allocation.
* **Dashboard visualization** proved valuable for real-time comparison and helped uncover state-level patterns that static analysis might miss.