DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

Hospital General Information

Submitted by

Name: Katakam Mohan Sai

Registration No: 12313034

Programme and Section: Data Science, K23EV

Course Code: INT375

Under the Guidance of

Assistant Professor. Sandeep Kaur (UID:23614)

Discipline of CSE/IT

Lovely School of Computer Science

Lovely Professional University, Phagwara

# CERTIFICATE

This is to certify that Katakam Mohan Sai bearing Registration no. 12313034 has completed INT375 project titled, “Hospital General Information” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

Signature and Name of the Supervisor

Designation of the Supervisor

School of Computer Science Lovely Professional University Phagwara, Punjab.

Date:15-04-2025

# DECLARATION

I, Katakam Mohan Sai, student of Data Science, 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: 15-04-2025 | Signature |
| Registration No. 12313034 | Katakam Mohan Sai |

# ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my project guide ,Sandeep Kaur, for their valuable guidance and support throughout this project, “Hospital General Information*”*. I am thankful to the Department of Computer Science and Engineering, Lovely Professional University, for providing the necessary resources and environment. I also acknowledge the Government of India for making the dataset publicly available, enabling this research.

Name: Katakam Mohan Sai

Registration number: 12313034

**Table of Contents**

1. Introduction
2. Source of Dataset
3. EDA Process
4. Analysis on Dataset

i. Introduction ii. General Description iii. Specific Requirements, Functions, and Formulas iv. Analysis Results v. Visualization

1. Conclusion
2. Future Scope
3. References

# INTRODUCTION

Hospital General Information provides a comprehensive overview of essential details about a hospital's structure, services, and operational standards. This includes data on hospital type, location, ownership, emergency services availability, and certifications. Such information is crucial for patients, healthcare professionals, and administrative bodies to understand the hospital's capabilities, accessibility, and compliance with healthcare regulations. Accurate general information also supports decision-making for treatment, referrals, and collaborations.

# Source of Dataset

* File Name: Hospital General Information
* Source: <https://data.cms.gov/provider-data/dataset/xubh-q36u>
* Attributes:

1. **Facility ID**  
   A unique identifier for each hospital.
2. **Facility Name**  
   The official name of the hospital.
3. **Address**  
   Street address of the hospital.
4. **City/Town**  
   The city or town where the hospital is located.
5. **State**  
   The U.S. state abbreviation (e.g., AL for Alabama).
6. **ZIP Code**  
   Postal code of the hospital's location.
7. **County/Parish**  
   The county or parish in which the hospital is situated.
8. **Telephone Number**  
   Contact number for public inquiries.
9. **Hospital Type**  
   Classification of the hospital, such as:
   * Acute Care
   * Critical Access
   * Psychiatric
10. **Hospital Ownership**  
    Type of organization that owns or operates the hospital:

* Government
* Proprietary (for-profit)
* Voluntary non-profit

1. **Emergency Services**  
   Indicates if emergency services are available (Yes or No).
2. **Meets criteria for birthing friendly designation**  
   Specifies if the hospital meets standards for being birth-friendly (e.g., maternity care practices).
3. **Hospital overall rating**  
   CMS star rating (1 to 5 stars) that summarizes performance across several areas.
4. **Hospital overall rating footnote**  
   Additional notes or clarifications about the rating.
5. **MORT Group Measure Count**  
   Total number of mortality-related measures evaluated.
6. **Count of Facility MORT Measures**  
   Number of mortality measures this facility reported.
7. **Count of MORT Measures Better**  
   How many mortality measures are better than the national average.
8. **Count of MORT Measures No Different**  
   Measures that are statistically similar to the national average.
9. **Count of MORT Measures Worse**  
   Measures worse than the national average.
10. **MORT Group Footnote**  
    Notes or exceptions related to MORT data.
11. **Safety Group Measure Count**  
    Total number of safety-related metrics.
12. **Count of Facility Safety Measures**  
    Number of safety measures this hospital reported.
13. **Count of Safety Measures Better**  
    Safety metrics that outperform the national average.
14. **Count of Safety Measures No Different**  
    Similar to the national average.
15. **Count of Safety Measures Worse**  
    Safety metrics worse than national benchmarks.
16. **Safety Group Footnote**  
    Comments on safety data or availability.
17. **READM Group Measure Count**  
    Total number of readmission-related measures.
18. **Count of Facility READM Measures**  
    How many such measures were reported by the facility.
19. **Count of READM Measures Better**  
    Readmission rates better than average.
20. **Count of READM Measures No Different**  
    No significant difference from national values.
21. **Count of READM Measures Worse**  
    Poorer than national standards.
22. **READM Group Footnote**  
    Explanations or special cases in the READM category.
23. **Pt Exp Group Measure Count**  
    Total number of patient experience metrics.
24. **Count of Facility Pt Exp Measures**  
    Number of patient experience metrics reported.
25. **Pt Exp Group Footnote**  
    Notes on patient survey responses or data gaps.
26. **TE Group Measure Count**  
    Number of timely/effective care metrics grouped.
27. **Count of Facility TE Measures**  
    How many of those metrics the facility reported.
28. **TE Group Footnote**  
    Additional comments on TE data.

# EDA Process (Exploratory Data Analysis Process)

The Exploratory Data Analysis (EDA) on the "Hospital General Information" dataset, which contains 5,396 entries and 38 columns, reveals detailed attributes of hospitals across various U.S. regions. The dataset includes general facility information like name, address, and contact details, as well as critical operational metrics such as hospital type, ownership, and availability of emergency services. It also encompasses a range of performance indicators across domains like mortality (MORT), safety, readmission (READM), patient experience (Pt Exp), and timely and effective care (TE), although several of these fields contain substantial missing values. Notably, only about 41% of entries include data on whether the hospital meets the birthing-friendly designation criteria. Most performance-related columns are stored as objects, suggesting potential formatting or parsing issues. Overall, the dataset offers a rich foundation for assessing hospital quality and accessibility, but it also requires preprocessing to handle missing and inconsistent data before deeper analysis. ​

# ANALYSIS ON DATASET

This dataset contains information on 5,396 hospitals across the U.S., covering both structural details and performance metrics. Most hospitals are categorized as Acute Care Hospitals, and ownership types vary among Government, Voluntary non-profit, and Proprietary institutions. All hospitals have complete entries for basic attributes like address, city, state, ZIP code, and emergency services availability.

A significant portion of the data is dedicated to evaluating hospital quality through grouped measures such as Mortality (MORT), Safety, Readmission (READM), Patient Experience (Pt Exp), and Timely & Effective Care (TE). Each group provides the number of measures evaluated and the counts of outcomes that were "Better," "No Different," or "Worse" than national benchmarks. However, many of these performance columns are stored as text and include placeholders like “Not Available,” indicating the need for cleaning and conversion to numeric types.

Additionally, the dataset has missing values, particularly in columns such as Meets criteria for birthing friendly designation, Hospital overall rating footnote, and group footnotes across categories. This indicates inconsistency in data reporting or availability across hospitals.

Overall, this dataset is a valuable resource for assessing hospital accessibility, performance, and quality variation across different states and ownership models. It is well-suited for analysis such as geographic performance comparisons, ownership impact on quality, or clustering hospitals based on their service and quality profiles—but will require preprocessing steps like handling missing values and converting data types for deeper insights

**4.2 Histogram:**

**4.2.1 Introduction**  
The ZIP Code histogram visualizes the geographic distribution of hospitals in the Hospital General Information dataset across the United States.

**4.2.2 General Description**

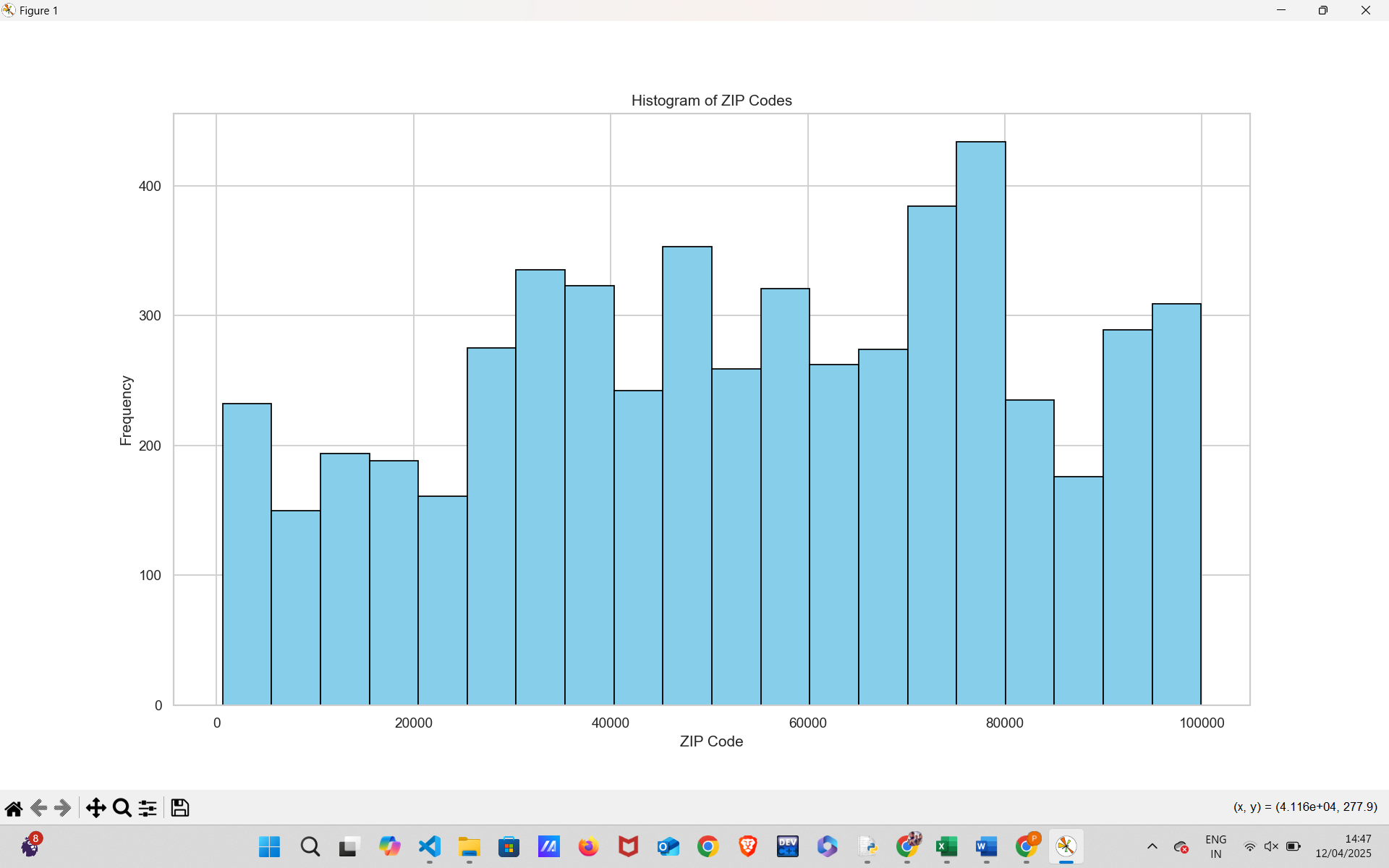
The graph represents hospital counts based on their ZIP Codes, which range from low to high values on the x-axis, with frequency (number of hospitals) on the y-axis. Each bar groups a range of ZIP Codes, showing how hospitals are spread across different geographic areas.

**4.2.3 Specific Requirements**

To create this visualization, ZIP Codes needed to be converted to numeric format and cleaned of any missing or non-standard entries. The data was then grouped into bins to observe concentration patterns across the ZIP Code spectrum.

**4.2.4AnalysisResults**  
The histogram reveals an uneven distribution of hospitals across ZIP Code regions. Higher concentrations appear in ZIP ranges from 60,000 to 80,000, indicating denser hospital presence in those geographic areas—likely corresponding to more urbanized or healthcare-focused regions. Meanwhile, fewer hospitals are observed in lower ZIP Code ranges, suggesting sparser coverage in certain parts of the country. This insight can guide further geographic or accessibility analysis.

**4.2.5 Visualization**



**4.3 Bar Chart**

**4.3.1 Introduction**

This bar chart displays the top 10 U.S. states with the highest number of hospitals based on the Hospital General Information dataset.

**4.3.2 General Description**

Each bar represents a state, with its height indicating the number of hospitals located there. States are sorted in descending order to highlight those with the greatest hospital density.

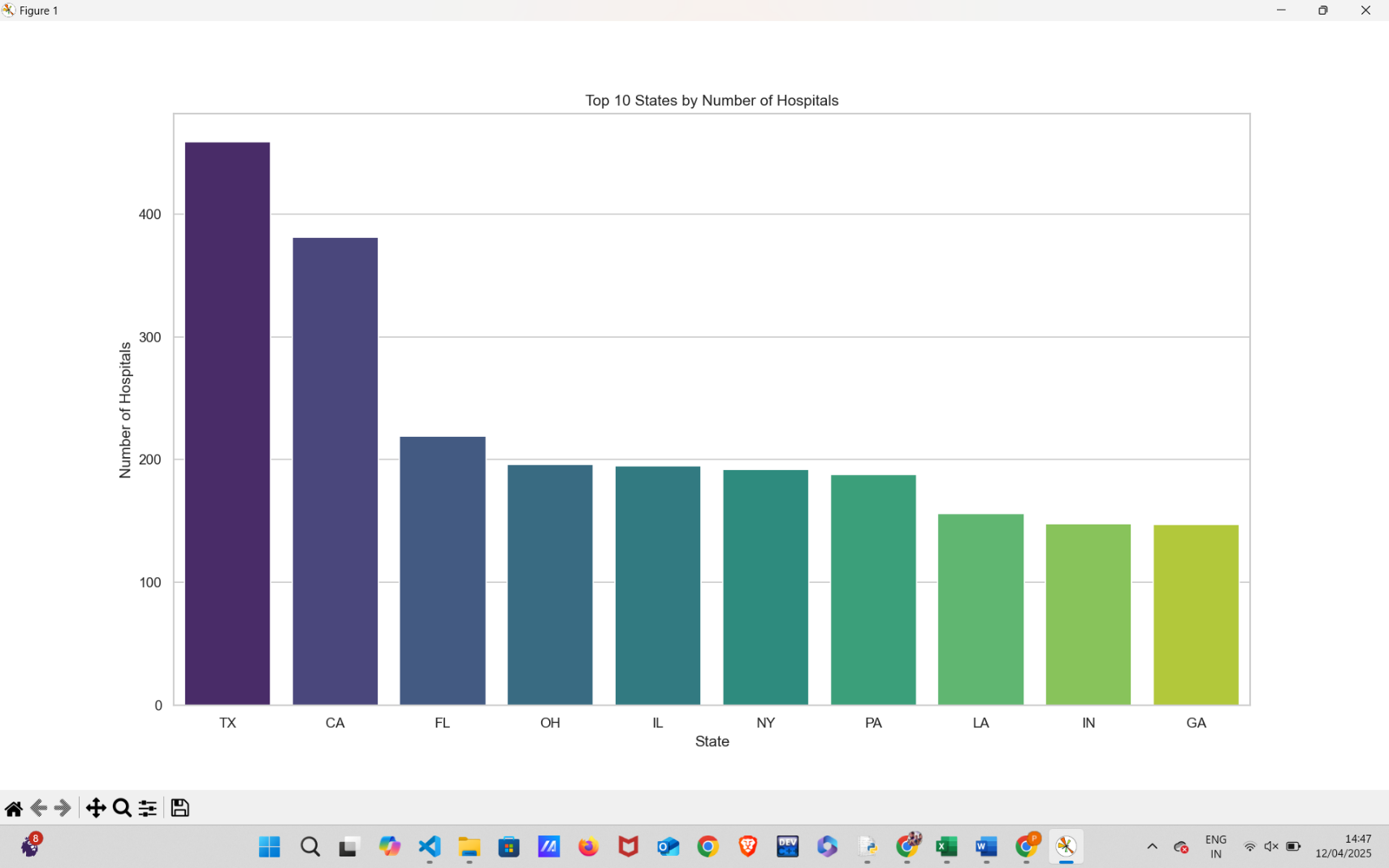
**4.3.3 Specific Requirements**

To generate this chart, the dataset was grouped by the "State" column, and the number of hospitals per state was counted. The top 10 states were then selected and visualized using a color-gradient bar chart for clarity.

**4.3.4 Analysis Results**

Texas (TX) leads with the highest number of hospitals, followed by California (CA) and Florida (FL). Other states like Ohio (OH), Illinois (IL), and New York (NY) also show substantial hospital counts. This distribution reflects population density and healthcare infrastructure, helping identify key regions with major hospital networks.

**4.3.5 Visualization**



**4.4 Line Plot**

The line chart illustrates the average overall hospital ratings for the top 10 U.S. states based on hospital count.

**4.4.1 Introduction**

The line chart illustrates the average overall hospital ratings for the top 10 U.S. states based on hospital count.

**4.4.2 General Description**

The x-axis represents states, while the y-axis shows their corresponding average hospital ratings. The data is connected using a line to visualize the trend in rating performance across these high-density states.

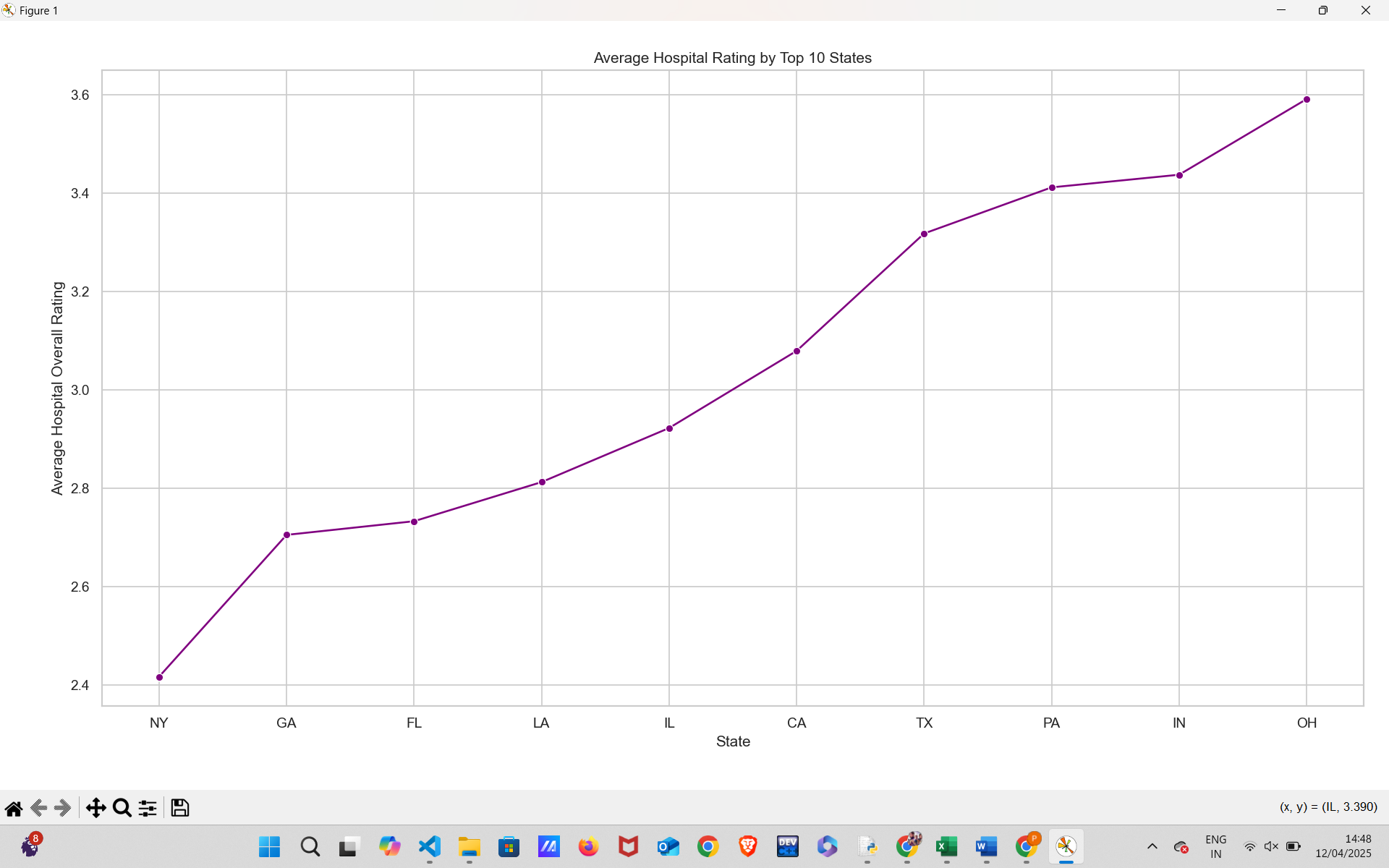
**4.4.3 Requirements**

This visualization was generated by calculating the mean overall rating of hospitals within each of the top 10 states identified earlier by hospital count. Only valid numerical ratings were included in the averaging process.

**4.4.4 Analysis Results**

Ohio (OH) stands out with the highest average rating, followed by Indiana (IN) and Pennsylvania (PA), indicating better hospital quality in these states. Conversely, New York (NY) has the lowest average rating, with Georgia (GA) and Florida (FL) also scoring on the lower end. This highlights a possible quality gap despite high hospital presence in some states.

**4.4.5 Visualization**



**4.6 Box plot**

**4.6.1 Introduction**

The box plot displays the distribution of hospital overall ratings categorized by different ownership types.

**4.6.2 General Description**

The x-axis lists various hospital ownership categories, including government, private, voluntary, and special entities. The y-axis represents hospital ratings, ranging from 1 to 5. Each box shows the interquartile range, with lines (whiskers) extending to show variability outside the upper and lower quartiles.

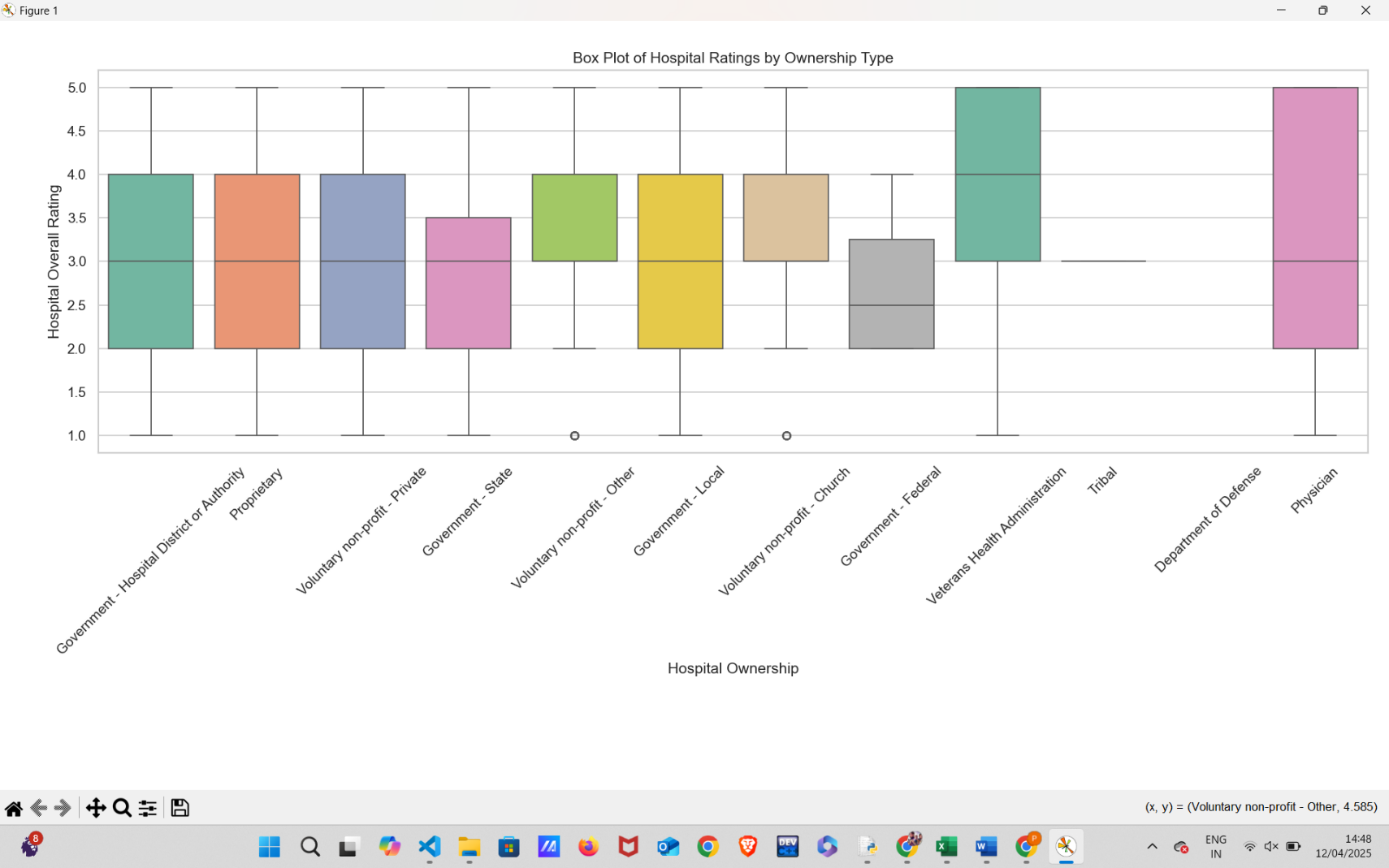
**4.6.3 Requirements**

This visualization requires hospital rating data and corresponding ownership information. It is used to compare how ownership type might influence hospital performance based on patient ratings.

**4.6.4 Analysis Results**

Hospitals owned by "Voluntary non-profit – Other" and "Veterans Health Administration" tend to have higher median ratings and smaller variability, indicating consistent quality. In contrast, "Government – Federal" and "Physician-owned" hospitals show more spread and lower medians, hinting at varying performance levels. Some ownership types like "Tribal" and "Department of Defense" show limited spread, possibly due to fewer data points.

**4.6.5 Visualization**



**4.6 Heatmap**

**4.6.1 Introduction**

The heatmap illustrates the correlation between various numerical features within the dataset, primarily focusing on healthcare-related group footnotes and ZIP codes.

**4.6.2 General Description**

On both axes, the heatmap includes features such as ZIP Code, MORT (Mortality), Safety, READM (Readmission), Pt Exp (Patient Experience), and TE (Timely and Effective Care) Group Footnotes. The color gradient from blue to red represents correlation strength, with red indicating strong positive correlation (1.0) and blue indicating negative correlation.

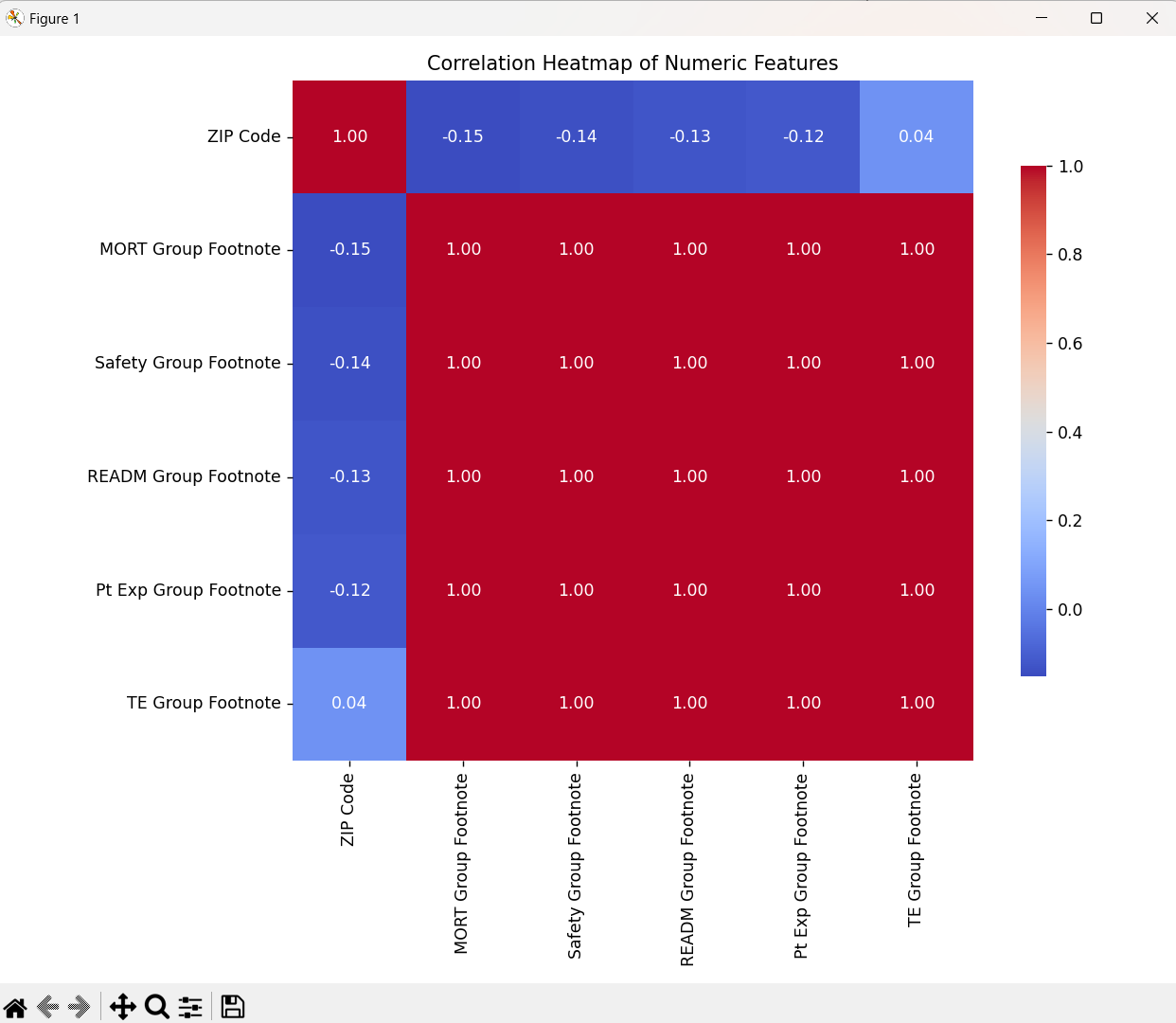
**4.6.3 Requirements**

This visualization requires a dataset with numeric values across these metrics to compute Pearson correlation coefficients. It helps in identifying multicollinearity or feature redundancy in the dataset.

**4.6.4 Analysis Results**

There is a perfect correlation (1.00) among the healthcare group footnotes—MORT, Safety, READM, Pt Exp, and TE—indicating they may be duplicates or highly dependent on one another. ZIP Code shows weak negative correlation with these variables, suggesting geographic location has minimal linear relationship with hospital group scores. This insight calls for further data cleaning or dimensionality reduction if modeling is to be performed.

**4.6.5 Visualization**

****

**4.7 T-test**

**4.7.1Introduction**:  
The T-test is a statistical method used to compare the means of two independent groups to determine whether there is a significant difference between them.

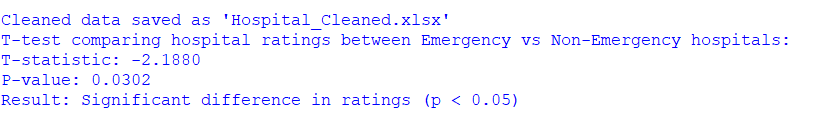
**4.7.2 General Description**  
In this case, the T-test compares the average hospital ratings between hospitals with emergency services and those without. The variable Hospital overall rating is converted to numeric, and hospitals with missing values are excluded from the analysis.

**4.7.3 Specific Requirements**  
The independent variable is Emergency Services (Yes/No), and the dependent variable is the hospital’s overall rating. The test assumes unequal variances (equal\_var=False).

**4.7.4 Analysis Results**

The T-statistic and P-value are computed. If the P-value is less than 0.05, it indicates a statistically significant difference in average hospital ratings between the two groups.If the P-value is greater than or equal to 0.05, it means there is no significant difference in ratings based on the presence of emergency services.

**4.7.5 Visualization**



**4.8 Chi-square Test Analysis**

**4.8.1 Introduction**:  
The Chi-square test is a non-parametric test used to determine whether there is a significant association between two categorical variables.

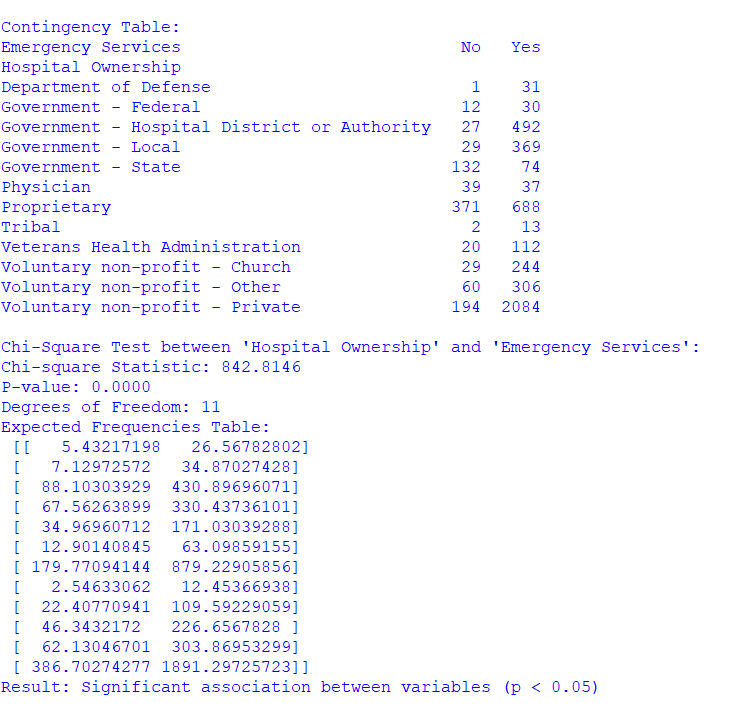
**4.8.2 General Description**  
This test evaluates the relationship between hospital ownership type and the availability of emergency services. A contingency table is created to show the distribution of ownership types across emergency service categories.

**4.8.3 Specific Requirements**  
The categorical variables are Hospital Ownership and Emergency Services. The chi2\_contingency function calculates the Chi-square statistic, P-value, degrees of freedom, and expected frequencies.

**4.8.4 Analysis Results**

The Chi-square test outputs a Chi-square statistic, P-value, and expected values for each cell.A P-value < 0.05 indicates a significant association between hospital ownership and emergency service availability.A P-value ≥ 0.05 suggests no significant association between the two variables.

**4.8.5 Visualization**

****

# FUTURE SCOPE

This dataset provides valuable insights into hospital distribution, ownership, and ratings across different states. In the future, it can be used to predict hospital performance using machine learning models. With ZIP code data, we can also perform geographic analysis to identify areas with poor healthcare access. The ownership type’s impact on ratings can guide health policy decisions. Since some features are perfectly correlated, data cleaning and feature selection can improve model accuracy. If additional data like time or patient feedback is added, we can study trends over time and patient satisfaction patterns. Overall, this dataset has strong potential for deeper healthcare analytics and decision-making support.

**REFERENCES**

1. Centers for Medicare & Medicaid Services (CMS). *Hospital General Information Dataset*. <https://data.cms.gov>
2. U.S. Department of Health & Human Services. *Hospital Compare – Overall Ratings*. Retrieved from <https://www.medicare.gov/hospitalcompare>
3. HealthData.gov. *Hospital Performance and Quality Measures*. [https://healthdata.go](https://healthdata.gov)