## AIML

Project Report

Semester-IV (Batch-2022)

**AQI Prediction Model**

A red and white sign

Description automatically generated with low confidence

**Supervised By: Submitted By:**

Dr Jatin Arora Dhruv Kundrai (2210990276) (G7)

Dhruv Walia (2210990280) (G7)

Dipanshu Saini(2210990288) (G7)

**Department of Computer Science and Engineering**

Chitkara University Institute of Engineering & Technology,Chitkara University, Punjab

**Abstract:**

This project focuses on developing an Air Quality Index (AQI) prediction model utilizing advanced Artificial Intelligence and Machine Learning (AIML) techniques. Leveraging popular data science libraries such as Pandas, NumPy, Seaborn, and Matplotlib, we performed comprehensive data analysis and visualization. The project involved importing an AQI dataset, splitting it into training and testing subsets, and using various statistical methods to analyze the data. Initially, we employed linear regression to create our prediction model; however, due to inaccuracies in the R² values, we ultimately adopted logistic regression. Through graphical representations such as pair plots and bar plots, we visualized data trends and validated our model's accuracy. This report details our methodology, findings, and the effectiveness of logistic regression in predicting AQI, highlighting its potential for real-world applications in environmental monitoring and public health.

**Implementation**

1. Data Collection and Preparation:

- Data Import: The AQI dataset was imported from a reputable online source. This dataset included various environmental parameters such as pollutant levels, temperature, and humidity.

- Data Cleaning: We utilized Pandas for data cleaning, which involved handling missing values, removing duplicates, and ensuring consistent data types across the dataset.

- Data Splitting: The dataset was divided into training and testing subsets using an 80-20 split to ensure robust model evaluation.

2. Exploratory Data Analysis (EDA):

- Statistical Analysis: Descriptive statistics were computed to understand the basic characteristics of the data.

- Visualization: Using Seaborn and Matplotlib, we created visualizations to explore data relationships and trends:

- Pair Plots: Illustrated relationships between different features.

- Bar Plots: Showcased AQI variations across different categories or time periods.

3. Model Development:

- Initial Linear Regression:

- We began with linear regression to predict AQI levels. However, the model's performance was unsatisfactory due to inaccurate R² values.

- Logistic Regression:

- To address the issues with linear regression, we switched to logistic regression, which is more suitable for classification tasks.

- Feature Engineering: Important features were selected based on correlation analysis and domain expertise.

- Model Evaluation: The model's performance was assessed using accuracy, precision, recall, and confusion matrix metrics on the testing dataset.

4. Tools and Libraries:

- Pandas: Used for data manipulation and analysis.

- NumPy: Facilitated numerical computations.

- Seaborn and Matplotlib: Enabled comprehensive data visualization.

- Scikit-learn: Provided machine learning algorithms and evaluation metrics.

5. Challenges and Solutions:

- Data Imbalance: Tackled through techniques such as oversampling and undersampling, and by using balanced accuracy metrics.

- Model Selection: The initial challenges with linear regression led to adopting logistic regression, which provided more reliable classification results.

**TABLE OF CONTENT**

**CHAPTER.NO Title**

**1. INTRODUCTION**

**1.1 Background**

**1.2 Objectives**

**1.3 Significance**

**2. Problem Definition and Requirements**

**2.1 Problem Statement**

**2.2 Software Requirements**

**2.3 Hardware Requirements**

**2.4 Data Sets**

**3. Proposed Design / Methodology**

**3.1 Algorithms**

**4. Results**

**4.1 Screenshots**

**Introduction:**

Air quality is a critical concern due to increasing pollution from industrialization and urbanization. The Air Quality Index (AQI) is crucial for assessing pollution levels and their health impacts. Accurate AQI predictions help public health agencies, environmentalists, and citizens take proactive measures.

This project uses Artificial Intelligence and Machine Learning (AIML) to develop an AQI prediction model. Using libraries like Pandas, NumPy, Seaborn, and Matplotlib, we analyzed an extensive AQI dataset. Initially, we used linear regression, but due to inaccuracies in R² values, we switched to logistic regression, which provided more reliable classification results. Detailed data visualizations, such as pair plots and bar plots, helped validate our model’s accuracy.

This report details our approach, including data collection, exploratory analysis, model development, validation, and visualization, demonstrating how AIML techniques can enhance environmental monitoring and public health.

**Background:**

Air pollution is a growing global issue, significantly impacting public health and the environment. The Air Quality Index (AQI) is a standardized indicator used worldwide to communicate the quality of air in a region, quantifying the concentration of various pollutants and their potential health effects. As urbanization and industrial activities increase, the need for accurate AQI prediction models becomes imperative to inform and protect the public.

Traditionally, AQI forecasting has relied on statistical methods and historical data trends. However, with the advent of Artificial Intelligence and Machine Learning (AIML), more sophisticated and accurate prediction models can be developed. These models can handle large datasets, uncover hidden patterns, and provide more precise forecasts.

Our project leverages AIML techniques to build an AQI prediction model. Utilizing data science libraries such as Pandas for data manipulation, NumPy for numerical computations, and Seaborn and Matplotlib for data visualization, we aimed to create a comprehensive and reliable model. Initial attempts with linear regression were inadequate, leading us to adopt logistic regression, which better suited our classification needs.

**Objectives:**

1. Develop a Robust AQI Prediction Model:

- Create an accurate and reliable model using advanced AIML techniques to predict the Air Quality Index based on historical and real-time data.

2. Utilize Comprehensive Data Analysis:

- Perform detailed exploratory data analysis (EDA) using libraries like Pandas, NumPy, Seaborn, and Matplotlib to understand data patterns and relationships.

3. Implement and Compare Machine Learning Algorithms:

- Evaluate the performance of different machine learning algorithms, starting with linear regression and ultimately selecting logistic regression for its superior accuracy in classification tasks.

4. Visualize Data Effectively:

- Use data visualization tools to present findings clearly, including pair plots, bar plots, heatmaps, and accuracy plots to validate the model and communicate results.

5. Enhance Public Health Monitoring:

- Provide a tool that can aid in the timely dissemination of air quality information, enabling proactive measures to protect public health and improve environmental monitoring practices.

**Significance:**

The development of an accurate AQI prediction model holds substantial significance for public health, environmental monitoring, and urban planning. As air pollution continues to rise due to industrialization and urban expansion, the need for precise and timely air quality forecasts becomes critical. This project addresses this need by leveraging AIML techniques to predict AQI with higher accuracy and reliability.

By providing accurate AQI predictions, the model can help public health agencies issue timely warnings and advisories, enabling the public to take necessary precautions. This is particularly important for vulnerable populations such as children, the elderly, and individuals with respiratory conditions. Moreover, the insights gained from this model can guide policymakers and urban planners in implementing effective pollution control measures and strategies for sustainable urban development.

The use of advanced data visualization techniques in this project also enhances the interpretability of the data, making it easier for stakeholders to understand complex air quality trends and patterns. This can foster better communication and engagement among the community, researchers, and policymakers.

Overall, this project contributes to the ongoing efforts in environmental protection and public health by providing a sophisticated tool for AQI prediction. The application of AIML in this context demonstrates the potential of modern technology to address critical environmental challenges, paving the way for more informed decision-making and improved quality of life.

**Problem Statement:**

In the era of increasing pollution and urbanization, accurately predicting air quality is a crucial public health challenge. Despite the availability of extensive environmental data, existing methods for AQI forecasting often fall short in terms of accuracy and reliability. This inadequacy is particularly evident in traditional statistical methods that fail to capture complex patterns and interactions among various pollutants and environmental factors.

The primary issue is the lack of advanced predictive models that can efficiently utilize available data to forecast AQI levels with high precision. Current models often struggle with data variability and are not robust enough to provide timely and actionable insights. Moreover, many prediction models do not effectively incorporate visualization tools that can help stakeholders easily interpret and utilize the forecasted data.

Therefore, the problem statement of this project is to develop a sophisticated AQI prediction model using advanced AIML techniques that can address these shortcomings. The model should:

1. Accurately predict AQI levels based on historical and real-time environmental data.

2. Utilize comprehensive data analysis and visualization techniques to uncover hidden patterns and trends.

3. Compare the effectiveness of different machine learning algorithms to identify the most suitable approach.

4. Provide a user-friendly interface for stakeholders to access and interpret the predictions.

By tackling these challenges, the project aims to enhance the reliability of AQI forecasts, thereby supporting better decision-making for public health protection and environmental management.

**Software Requirements**

1. Python (Programming Language):

- Python serves as the primary programming language for this project, providing a versatile and powerful environment for data manipulation, analysis, and machine learning.

2. Jupyter Notebook or Integrated Development Environment (IDE):

- Jupyter Notebook or an IDE like PyCharm or Visual Studio Code has been used for writing and executing Python code, facilitating interactive development and experimentation.

3. Python Libraries:

- Pandas: Required for data manipulation and analysis, providing data structures and functions for efficient handling of tabular data.

- NumPy: Essential for numerical computations and array operations, enabling efficient manipulation of multi-dimensional arrays and matrices.

- Scikit-learn: A comprehensive machine learning library that provides a wide range of algorithms for classification, regression, clustering, and more.

- Seaborn and Matplotlib: Used for data visualization, these libraries offer a variety of plotting functions to create insightful and visually appealing charts and graphs.

4. Data Source:

- Access to a reliable and comprehensive dataset containing historical and real-time air quality data is necessary for training and evaluating the AQI prediction model. This dataset can be sourced from reputable environmental agencies or research institutions.

6. Operating System Compatibility:

- The software should be compatible with the operating system used by the development team, whether it be Windows, macOS, or Linux.

Ensuring compatibility and availability of these software components is essential for the successful implementation and execution of the AQI prediction model project.

**Hardware Requirements**

1. Computer System with:

* Modern Processor
* Minimum 8GB RAM
* Adequate Storage Space

2. Display Monitor (High-Resolution Recommended)

3. Input Devices:

* Keyboard
* Mouse (or preferred input device)

4. Internet Connection (Stable and High-Speed)

5. Testing Devices (Optional)

6. External Storage (Optional)

7. Printer (Optional)

8. Comfortable Workspace

**Dataset Structure for AQI Prediction Model**

Data Description

Our project utilized a dataset obtained from an online source. The dataset contained various features potentially influencing air quality, with the following key elements:

* Pollutant Concentrations**:** The dataset included concentrations of key pollutants directly contributing to AQI calculations:
  + pm2\_5 (fine particulate matter)
  + spm (suspended particulate matter)
  + so2 (sulfur dioxide)
  + no2 (nitrogen dioxide)
* Monitoring Station Information**:** The data included details about the monitoring stations:
  + location\_monitoring\_station ( indicating the station name or location)
  + stn\_code ( a unique identifier for each station)
  + agency (the agency responsible for data collection)

Analysis of Feature Relevance

While the dataset provided valuable information on pollutant concentrations, some features had uncertain relevance for direct AQI prediction:

* stn\_code and agency might be useful for understanding data collection specifics but require additional context (e.g., station type linked to stn\_code) for direct prediction.
* rspm (respirable suspended particulate matter) might be redundant with spm depending on the data collection methods.

Missing Features and Considerations

For a more comprehensive AQI prediction model, the following features were ideally present but not included in this specific dataset:

* Meteorological Data: Weather variables like temperature, humidity, wind speed/direction significantly impact air quality and could enhance model accuracy.
* AQI**:** The target variable for prediction (Air Quality Index values) was not directly available. This is essential for model training and evaluation.

**Proposed Design/Methodology**

Data Collection and Preprocessing

- Sourced dataset from reputable online repositories focusing on air quality indices (AQI).

- Curated dataset to ensure data integrity, addressing missing values, outliers, and inconsistencies.

- Preprocessed data through normalization and feature engineering for analysis.

Exploratory Data Analysis (EDA)

- Conducted comprehensive EDA using Pandas, NumPy, Seaborn, and Matplotlib.

- Utilized descriptive statistics and visualizations (e.g., histograms, scatter plots, correlation matrices) to unveil key insights into data distributions, relationships between variables, and potential trends.

-Model Development

- Partitioned dataset into training and testing subsets for model evaluation.

- Initially explored Linear Regression for AQI prediction but pivoted towards Logistic Regression due to challenges with R-squared values.

- Logistic Regression emerged as a more suitable approach, showcasing superior predictive accuracy and robustness.

Performance Evaluation

- Evaluated Logistic Regression model performance using metrics such as accuracy, precision, recall, and F1-score.

- Visualized model performance through confusion matrices and ROC curves to assess predictive capability accurately.

Conclusion

- Methodology facilitated construction of a reliable predictive model for air quality prediction using Logistic Regression.

- Despite challenges, approach yielded promising results, emphasizing the importance of robust methodologies in environmental research.

Future Work

- Explore alternative modeling techniques and incorporate additional environmental variables.

- Validate model across diverse geographical regions and conduct longitudinal studies to analyze temporal trends and seasonal variations in air quality dynamics.

**Algorithms**

1. Linear Regression

- Description: Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, aiming to find the best-fitting straight line that minimizes the sum of squared residuals.

- Application: Initially considered for AQI prediction in your project, it's commonly used in various fields for regression tasks.

2. Logistic Regression

- Description: Despite its name, Logistic Regression is used for classification tasks, not regression. It models the probability of a binary outcome by applying a logistic function to a linear combination of predictor variables. It's particularly suitable for binary classification problems.

- Application: Adopted as the primary algorithm for AQI prediction after challenges with Linear Regression. It's widely used in fields like healthcare, finance, and marketing for binary classification tasks.

3. Performance Evaluation Metrics

- Description: Various performance evaluation metrics were employed to assess the efficacy of the models. These include:

- Accuracy: Measures the proportion of correct predictions out of total predictions made by the model.

- Precision: Measures the proportion of true positive predictions out of all positive predictions made by the model.

- Recall: Measures the proportion of true positive predictions out of all actual positive instances.

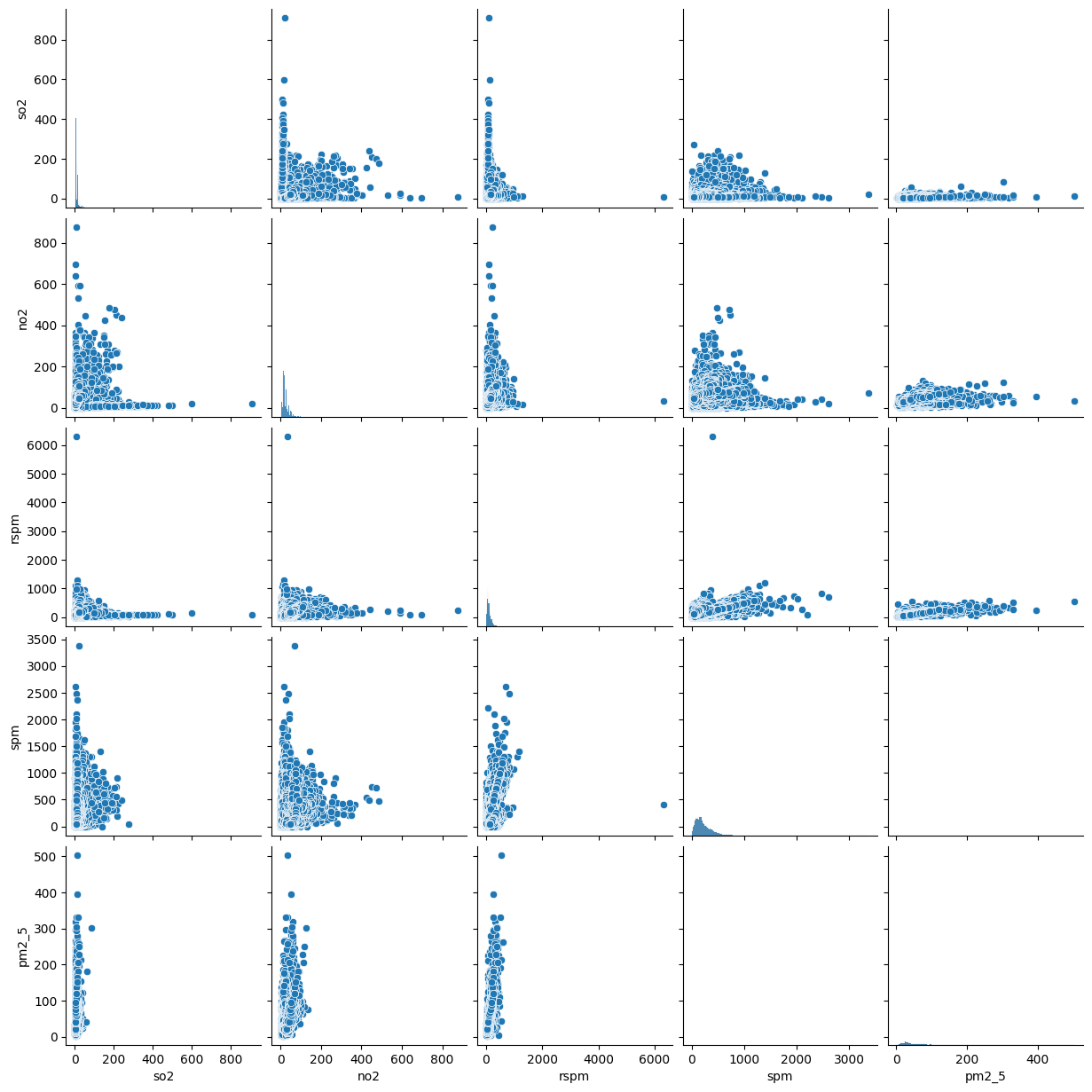
- F1-score: The harmonic mean of precision and recall, providing a balance between the two metrics.

- Application: Used to quantitatively evaluate the performance of both Linear and Logistic Regression models in predicting AQI levels.

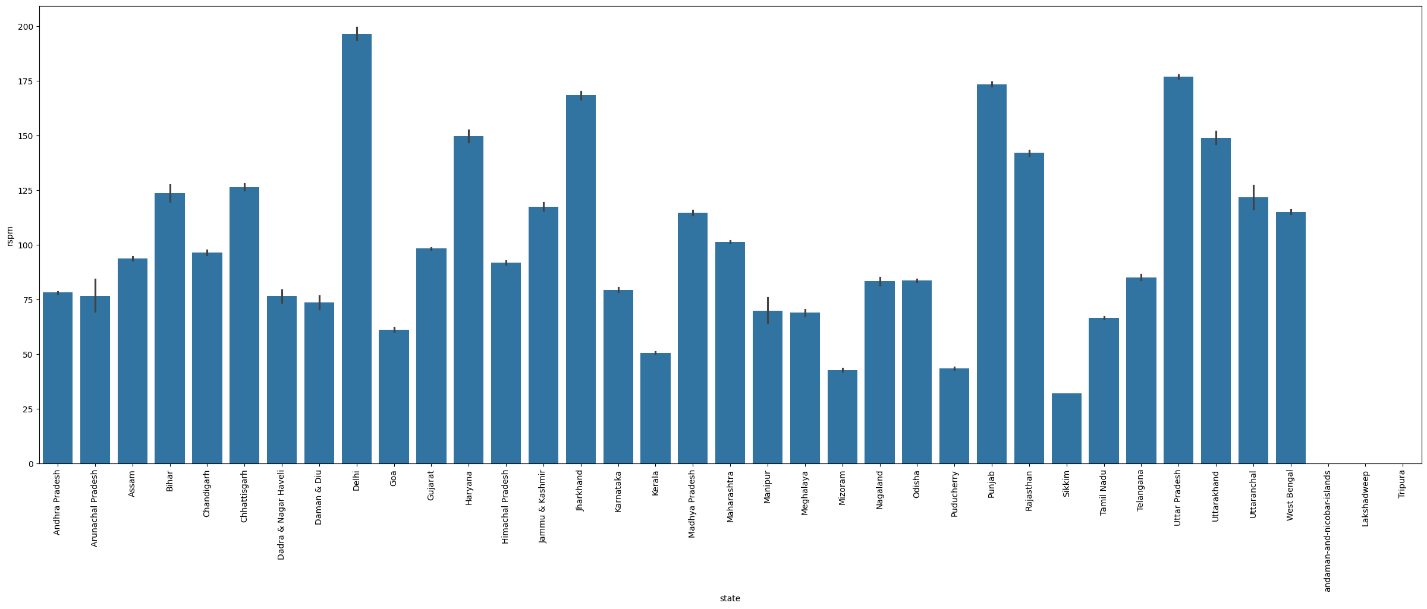
**RESULTS**

**Snap-Shots**

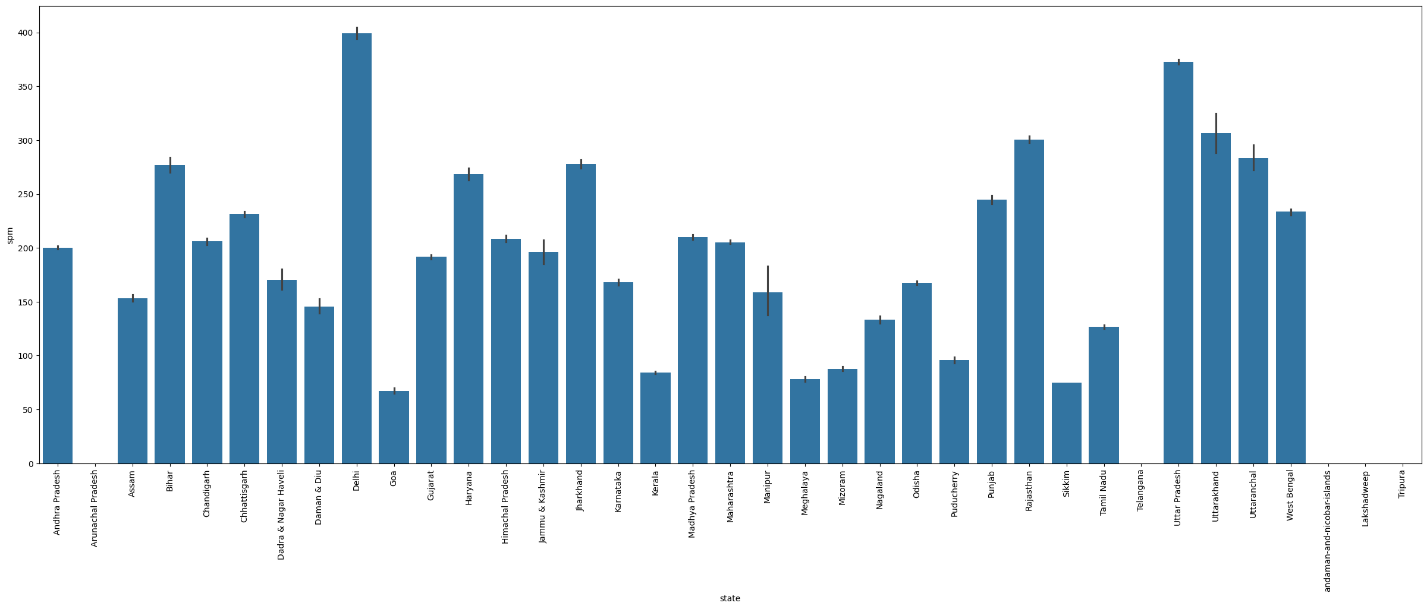
**1.Pair Plot**

****

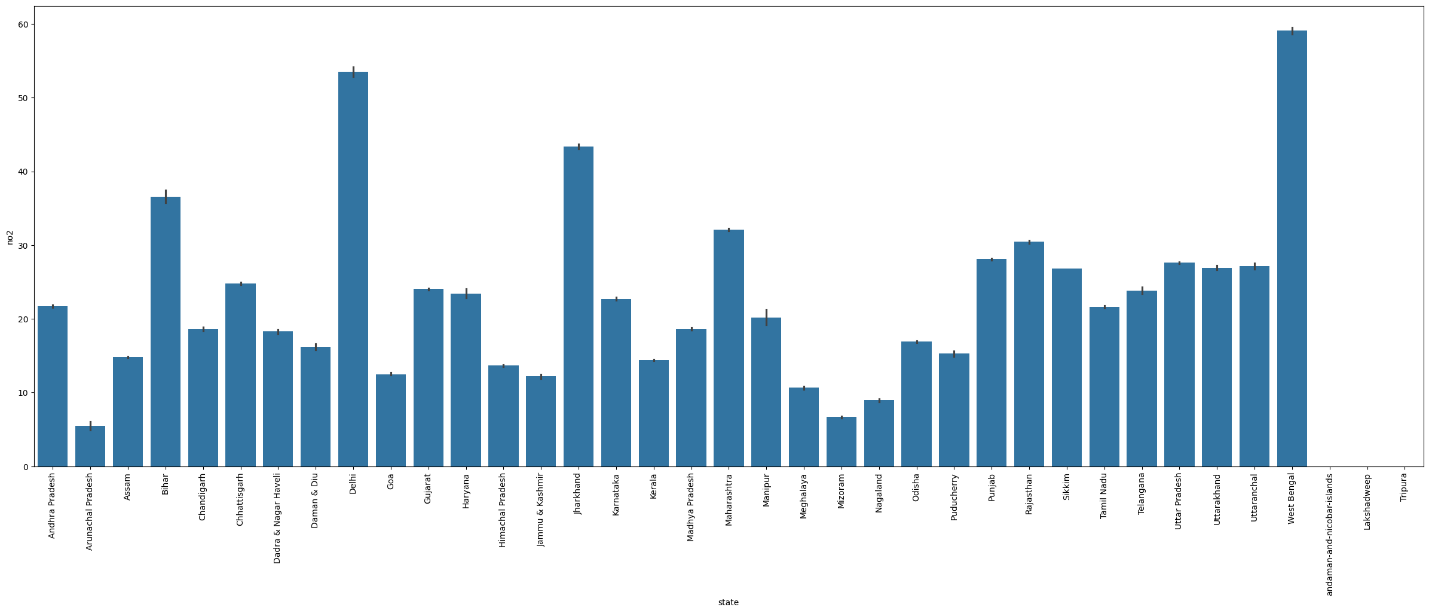
**2.Bar plot**

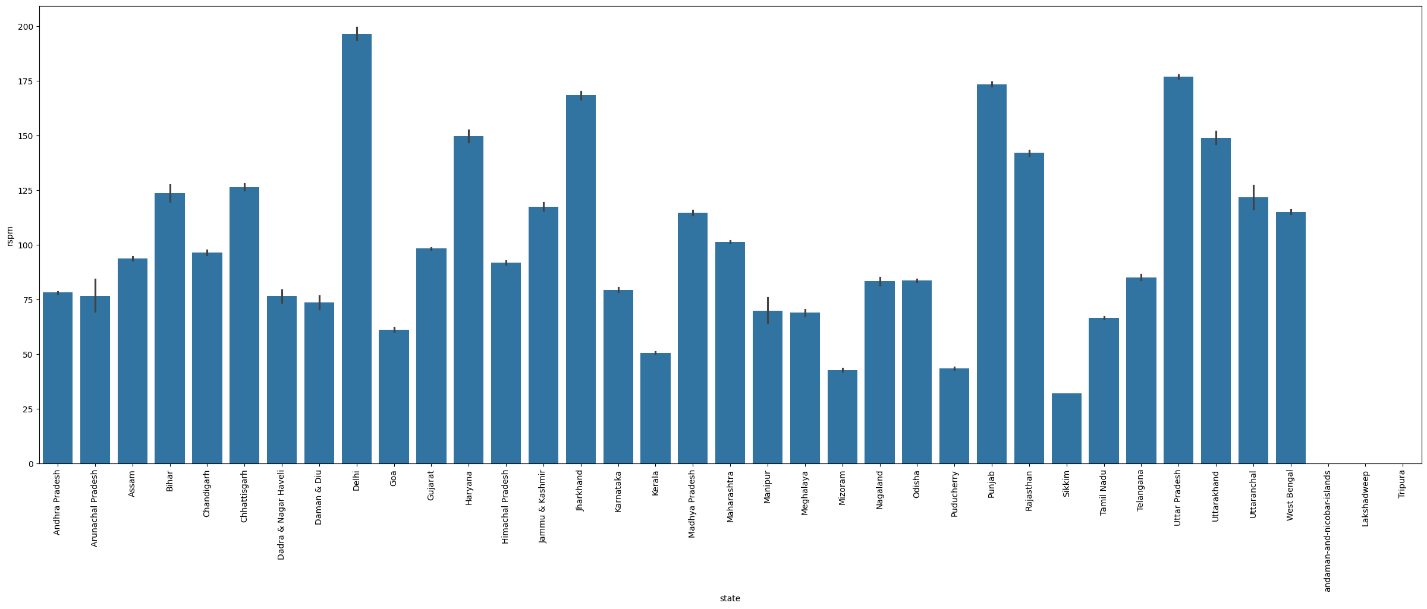
****

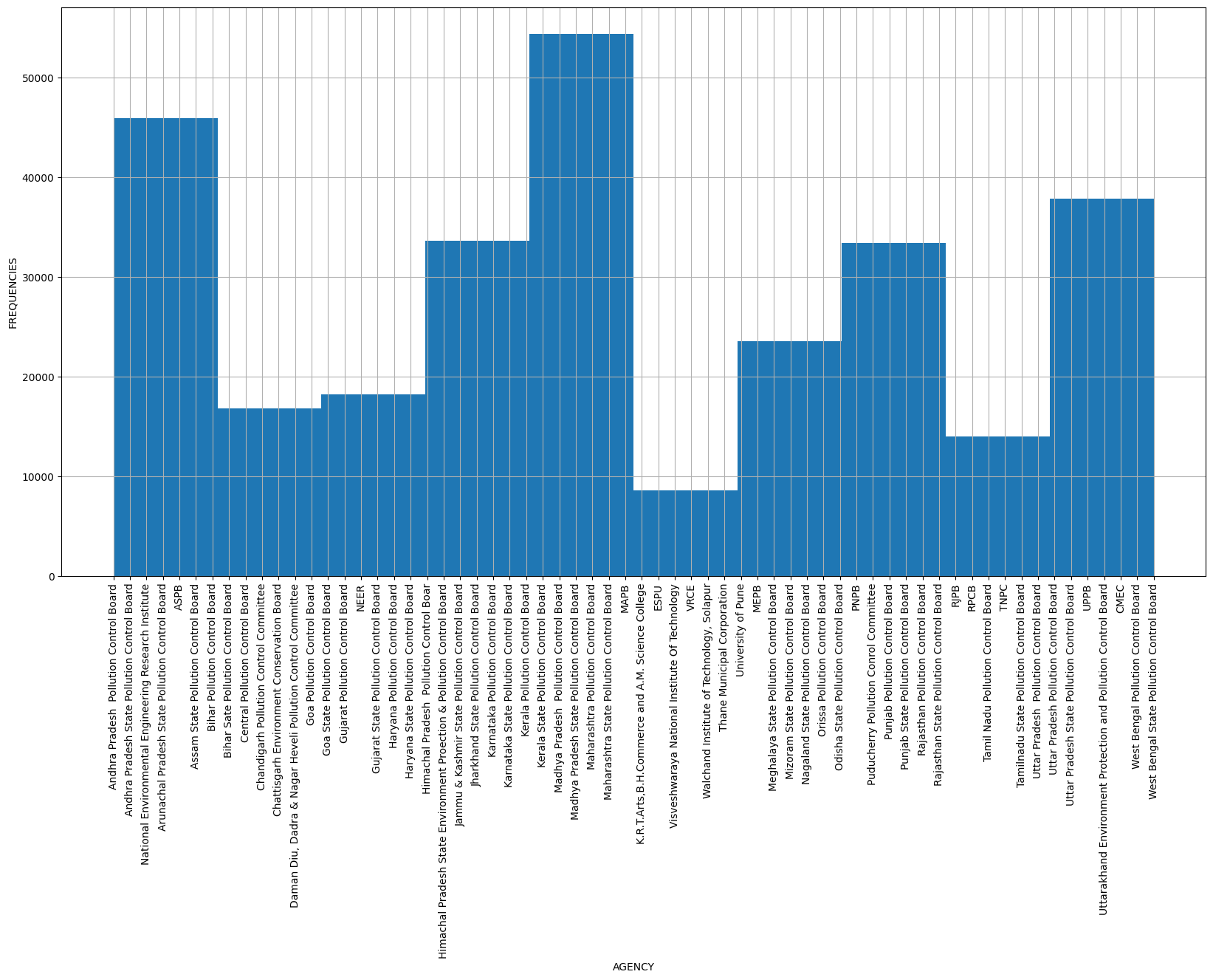
**3.Bar plot**

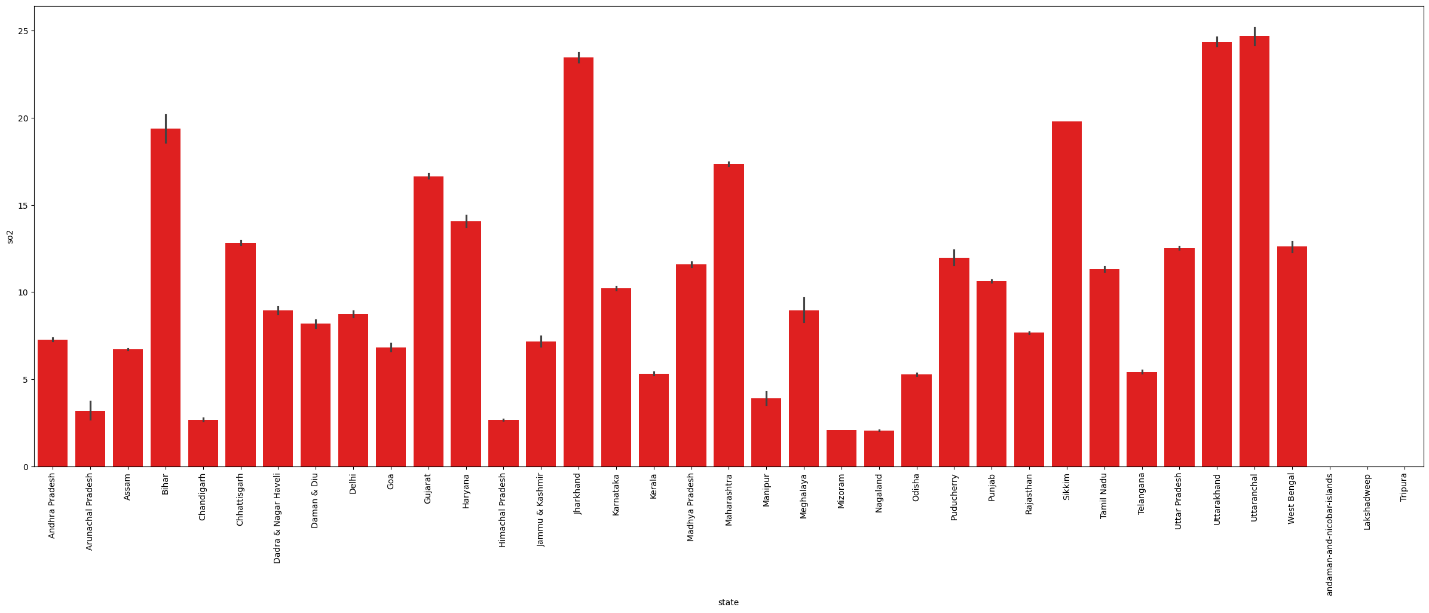
****

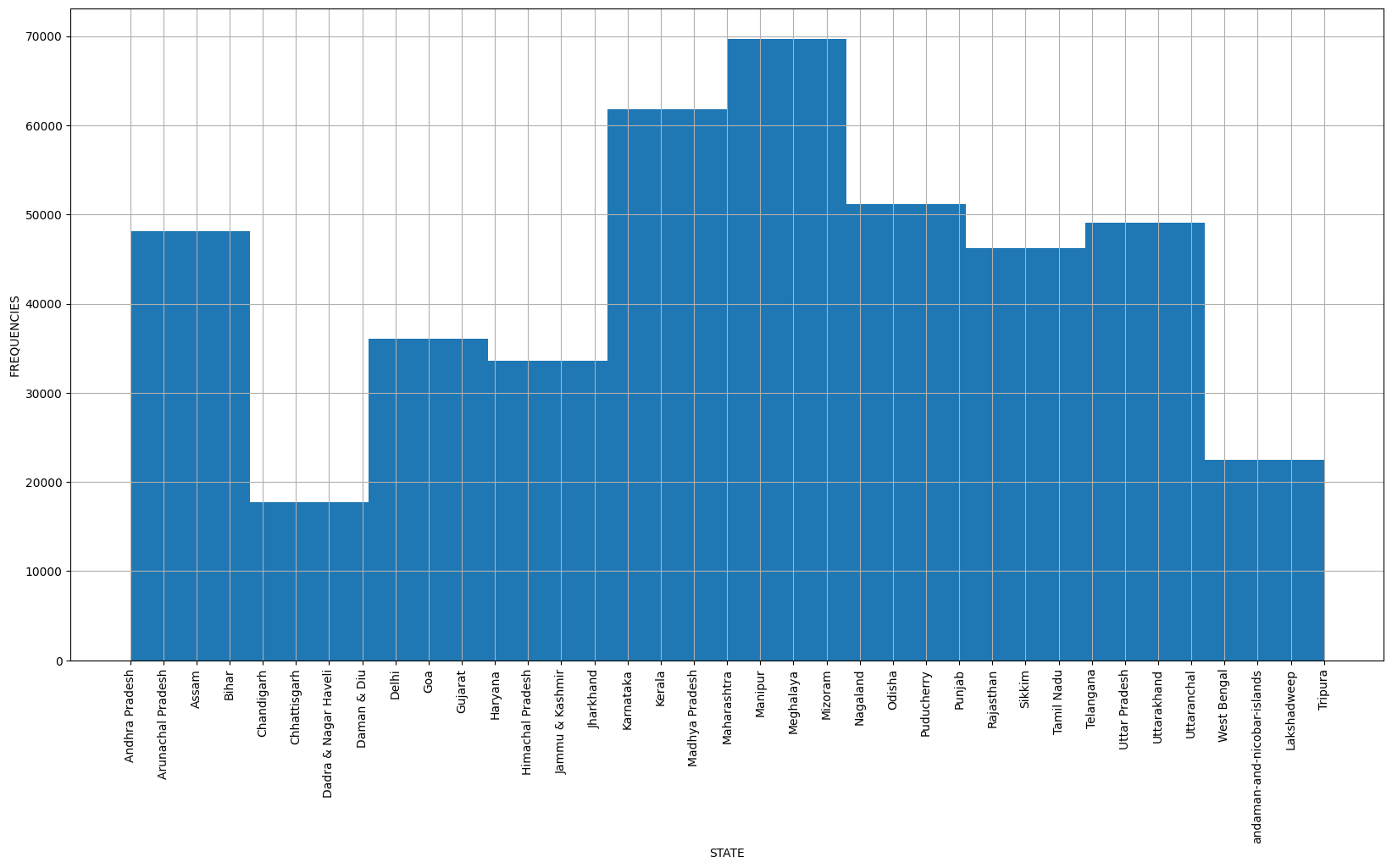
**4.Bar plot**

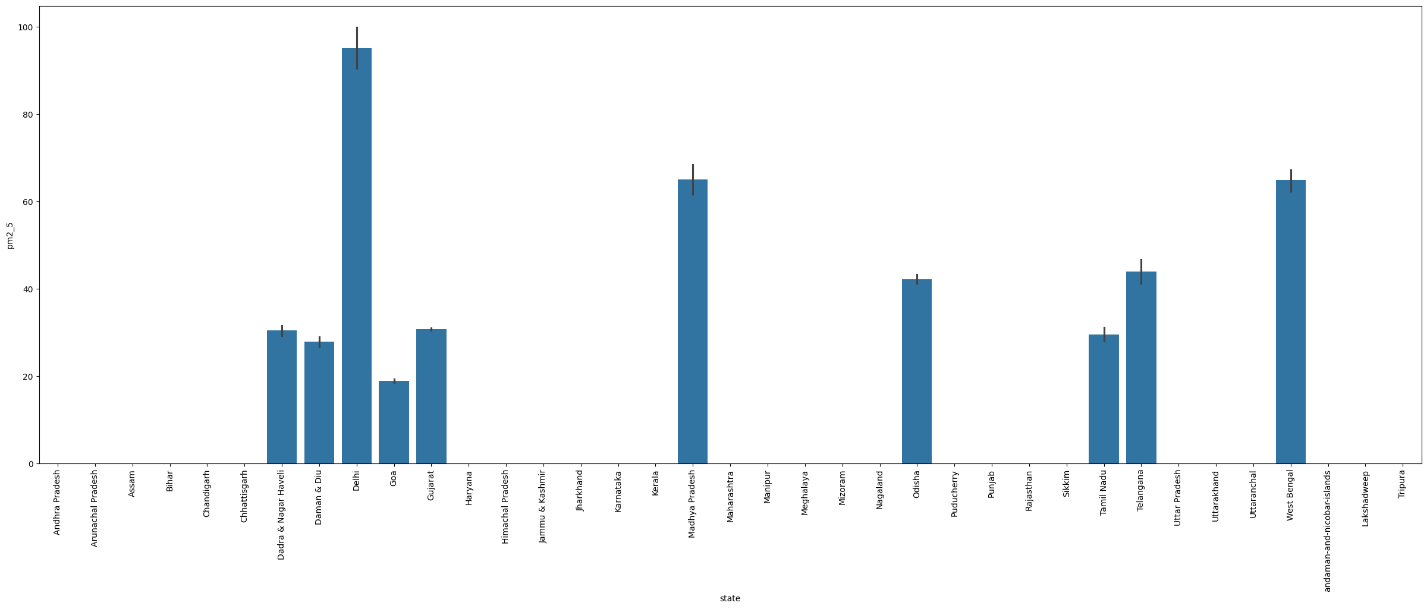
****

**5.Bar plot**

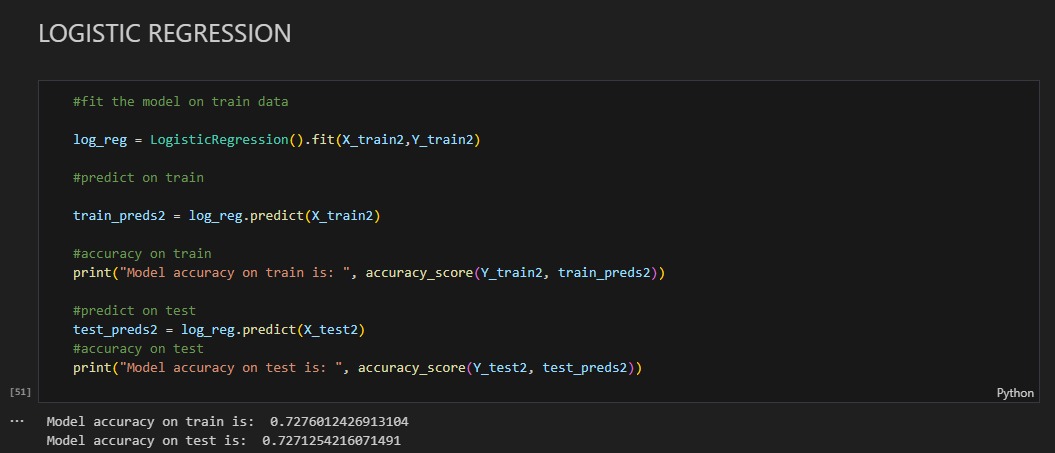
**6.Histogram**

**7.Bar plot**

**8.Histogram**

**9.Histogram **

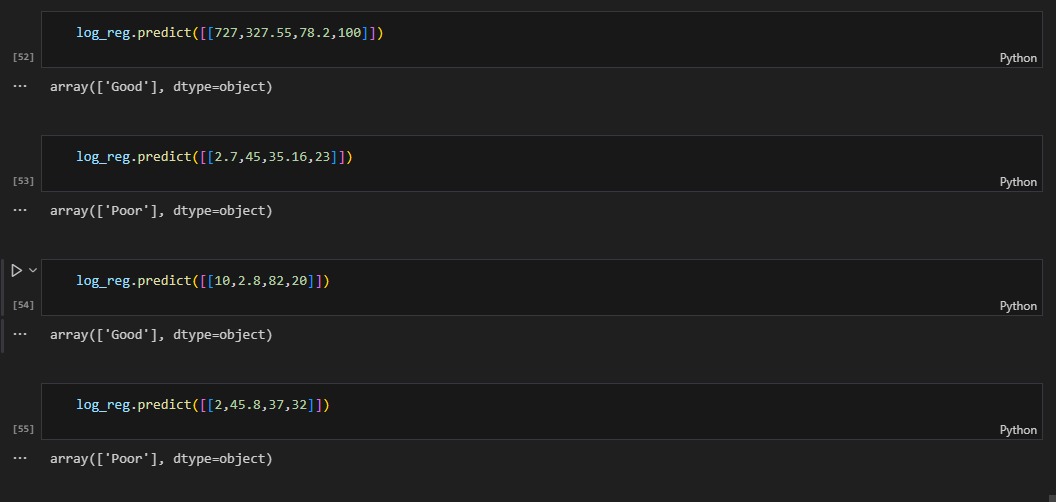
**Logistic Regression**

****

**Linear Regression**

****

**Final Prediction**

****