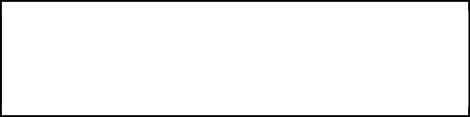
Bansilal Ramnath Agarwal Charitable Trust’s

**Vishwakarma Institute of Information Technology, Pune-48**

(An Autonomous Institute affiliated to Savitribai Phule Pune University)

Department of Computer Science and Engineering (Artificial Intelligence)



**Lab Assignment Submission**

**Machine Learning Lab**

**SY**

***Semester II Academic Year 2023-24***

***Guided by,***

Dr. Anuradha Yenkikar

***Student name: Neha Sawant***

***Batch:A3***

***Roll no: 281060***

***PRN:22210877***

**ASSIGNMENT NO: 1**

1.1] Problem Statement -

Perform the following operations using R/Python on suitable data sets:

a) read data from different formats (like csv, xls)

b) indexing and selecting data, sort data,

c) describe attributes of data, checking data types of each column,

d) counting unique values of data, format of each column, converting variable data type

(e.g. from long to short, vice versa),

e) identifying missing values and fill in the missing values

1.2] S/W Packages and Libraries used:

For the following assignment, the interpreter used was Google Collab and the Libraries used were

* Pandas - It is a powerful data manipulation library in Python, that provides data structures and functions for working with structured data.
* Numpy - NumPy is a fundamental package for scientific computing with Python, providing support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays.

1.3] Theory-

Methodology:

* For Reading Data from Different Formats:
  + Utilize libraries like Pandas to read data from various formats such as CSV, Excel (xls/xlsx), etc
  + Use of appropriate functions like read\_csv() or read\_excel() to load the data into Python data structures like DataFrames.
* For Indexing and Selecting Data, Sorting Data:
  + Pandas provides powerful indexing and selection mechanisms using methods like .loc[] and .iloc[].
  + Sorting data can be done using the sort\_values() function, specifying the column(s) to sort by.
* For Describing Attributes of Data, Checking Data Types:
  + Use the describe() method to generate descriptive statistics of the data.
  + Check data types of each column using the dtypes attribute of the DataFrame.
* For Counting Unique Values, Formatting Columns, Converting Variable Data Types:
  + Employ functions like value\_counts() to count unique values in a column.
  + Convert variable data types using functions like astype() to cast one data type to another.
* Identifying Missing Values and Filling Them:
  + Pandas provides functions like isna() or isnull() to identify missing values.
  + Use methods like fillna() or dropna() to handle missing data by filling them with appropriate values or dropping rows/columns with missing values.

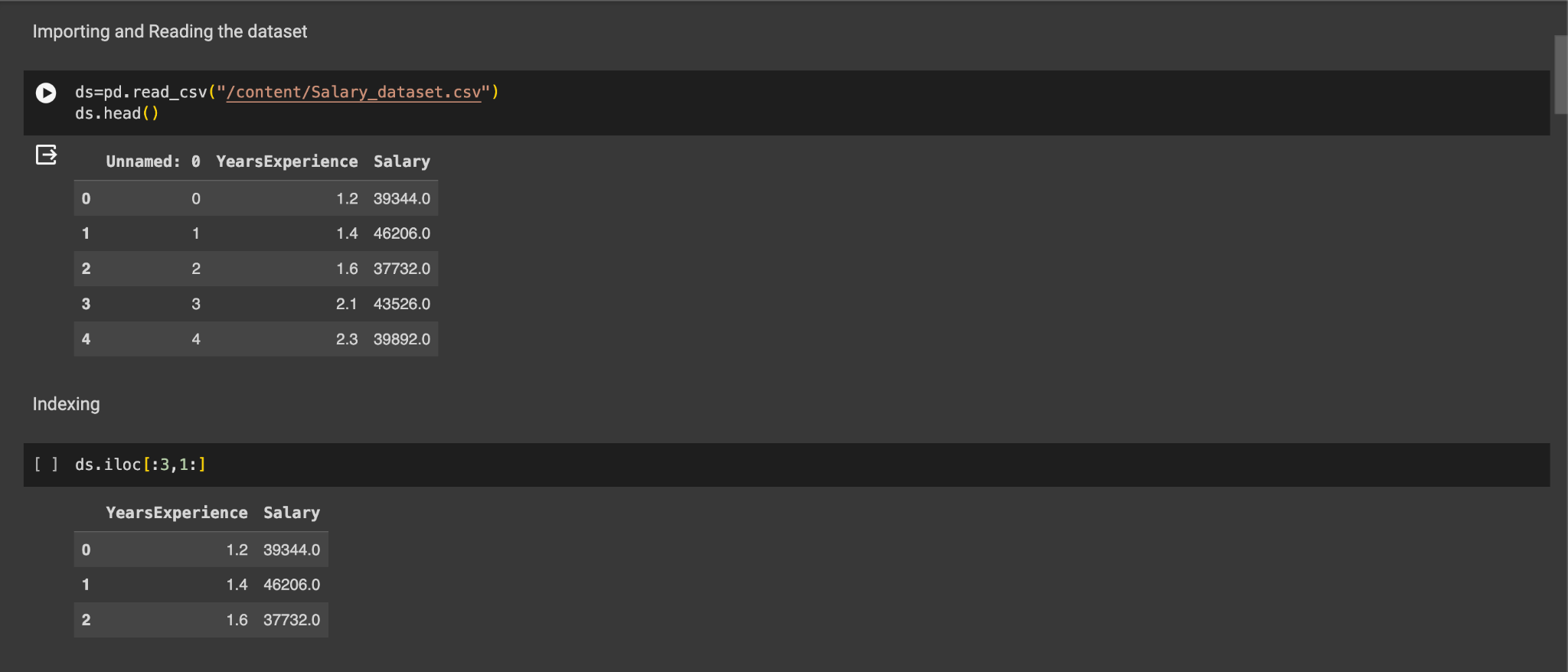
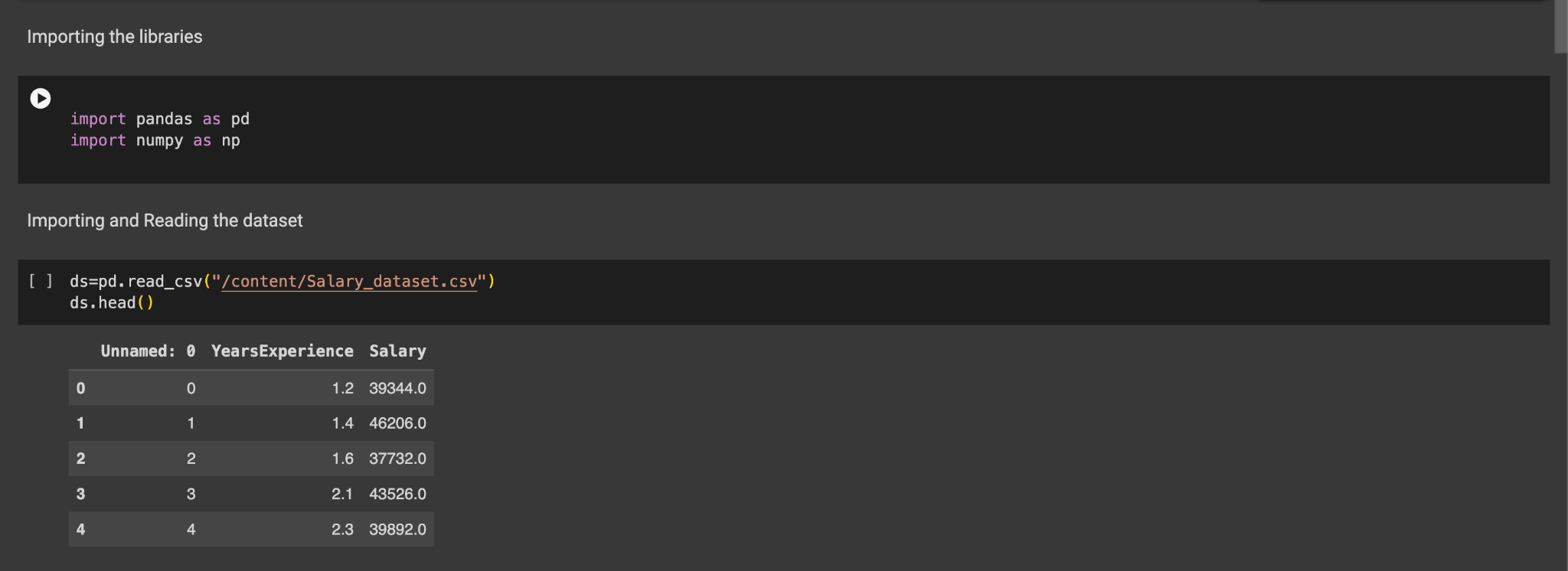
1.4] Applications:

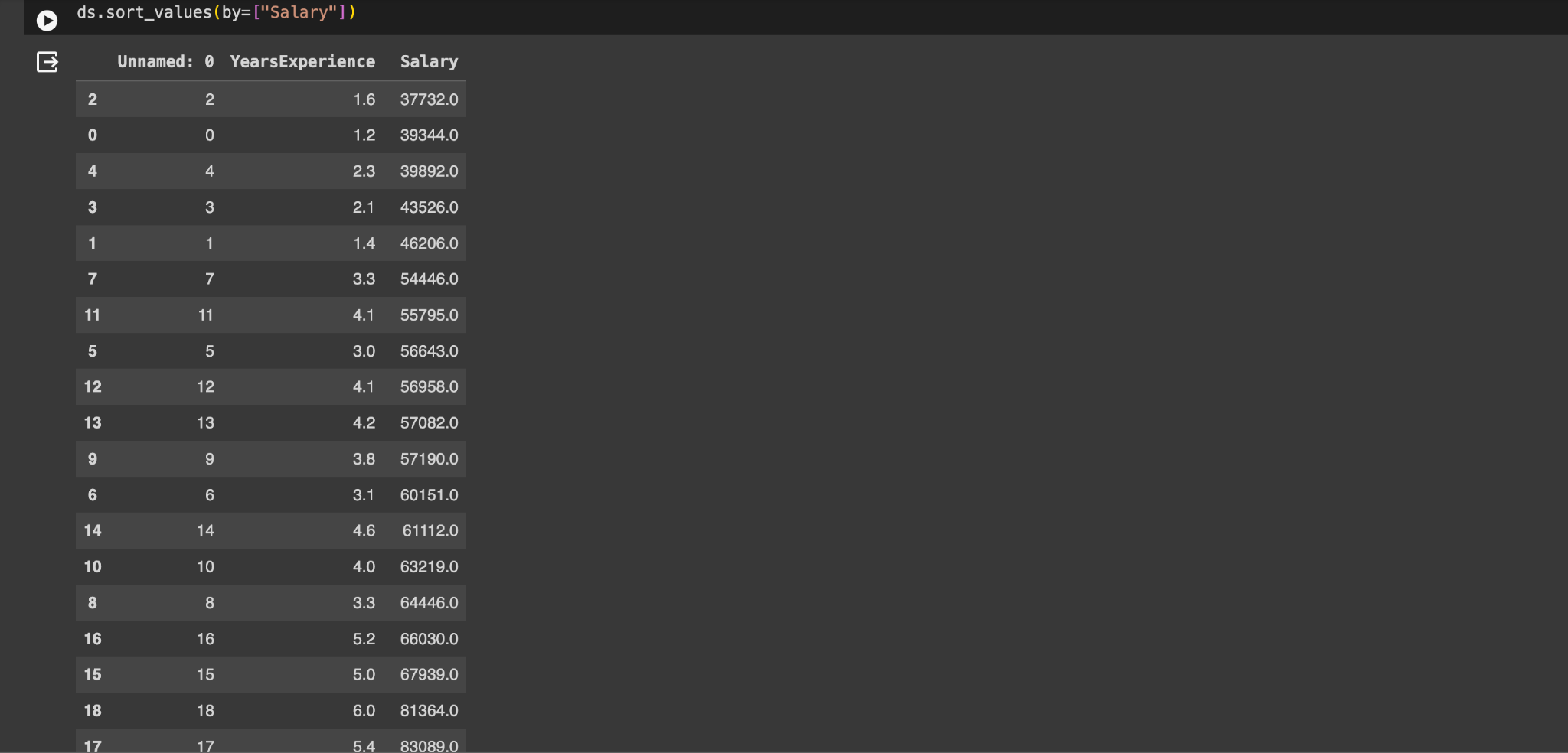
* Data Analysis: These operations are fundamental to exploratory data analysis (EDA), a crucial step in any data science or machine learning project.
* Data Cleaning: Handling missing values, formatting data, and converting data types are essential steps in data cleaning pipelines.

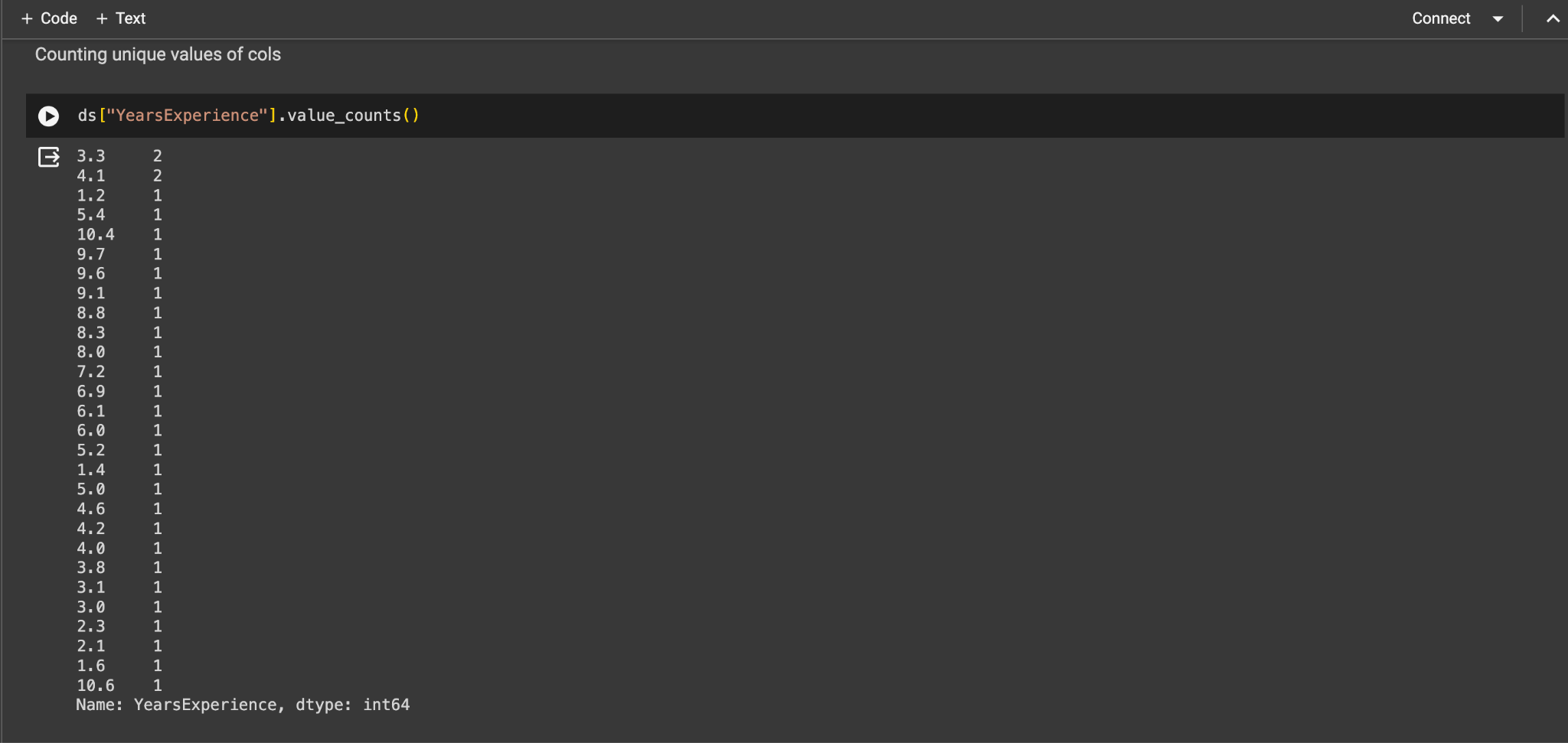
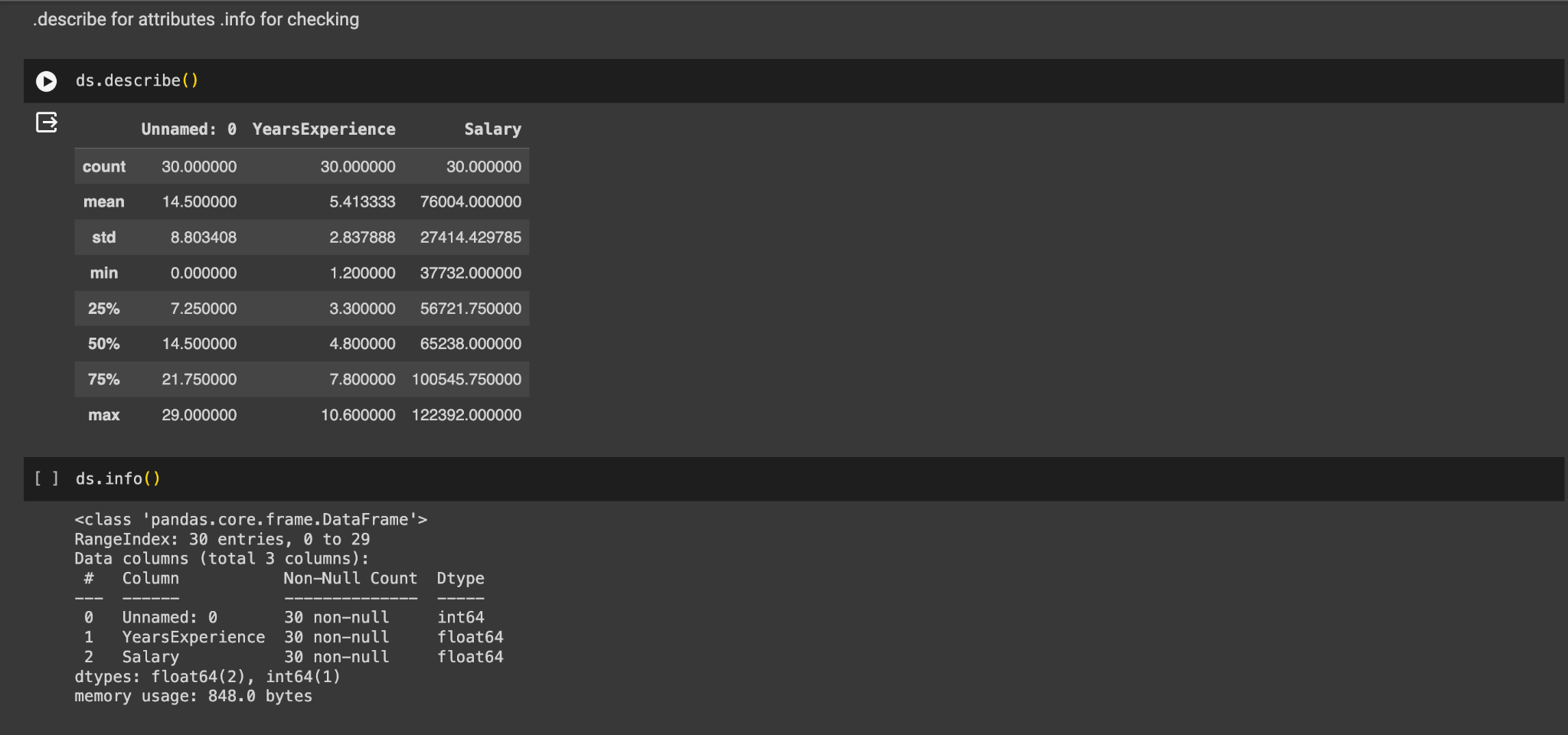
1.5] Limitations:

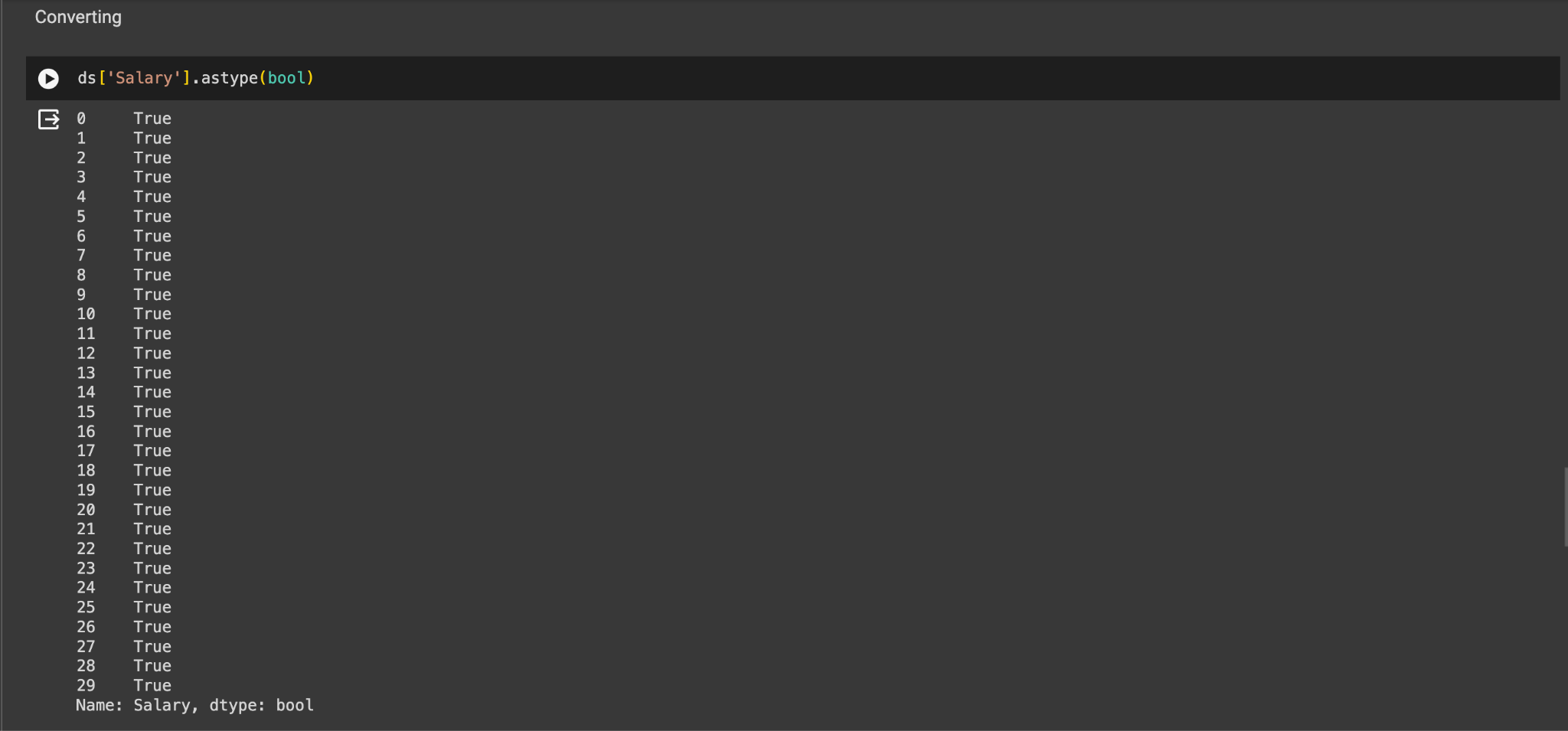
* Performance: While Pandas and NumPy are powerful libraries, they may not be optimal for very large datasets due to memory limitations.
* Learning Curve: These libraries have a steep learning curve, especially for beginners, which might hinder the quick implementation of complex operations.

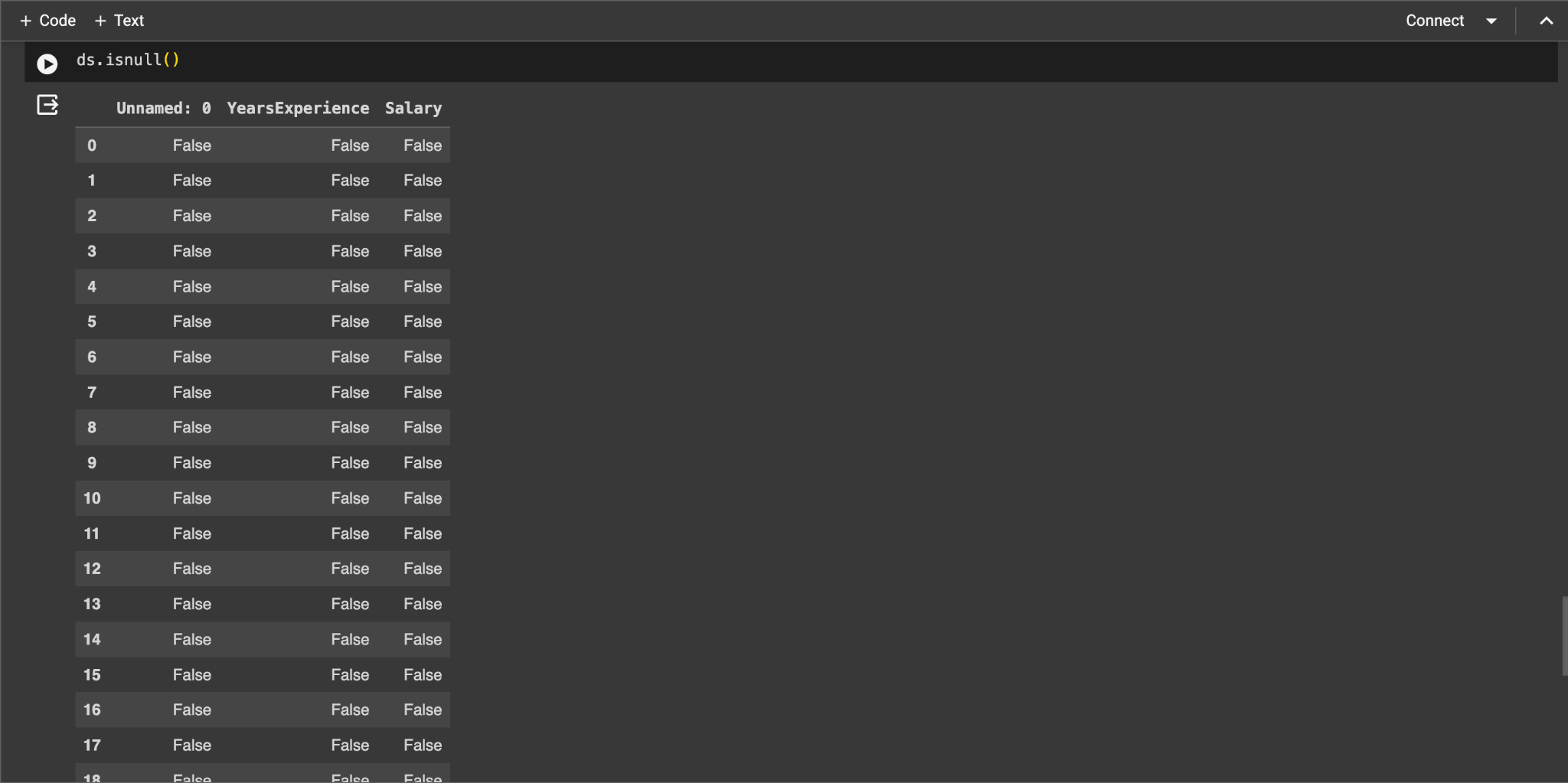
1.6] Working:

****

****

****

****



1.7] Conclusion:

In conclusion, the assignment showcased the effective use of Python libraries like Pandas and NumPy for diverse data manipulation tasks, from reading various formats to handling missing values. These operations are vital for data analysis, cleaning, and visualization, offering essential insights into structured data. Despite their benefits, it's important to recognize the limitations of these libraries, such as performance constraints with large datasets. Nevertheless, with a solid understanding of their theoretical framework, Pandas and NumPy empower proficient data handling, supporting informed decision-making across domains.

**ASSIGNMENT NO: 2**

2.1] Problem Statement -

Perform the following operations using Python on the data sets:

a) Compute and display summary statistics for each feature available in the dataset. (e.g minimum value, maximum value, mean, range, standard deviation, variance and percentiles

b) Data Visualization-Create a histogram for each feature in the dataset to illustrate the feature distributions.

c) Data cleaning, Data integration, Data transformation, Data model building (e.g. Classification)

2.2] S/W Packages and Libraries used:

For the following assignment, the interpreter used was Google Collab and the Primary Library used was-

* Matplotlib: Matplotlib is a versatile plotting library in Python, offering a wide range of visualization capabilities, including histograms for illustrating feature distributions.

2.3] Theory-

* For Computing Summary Statistics:
  + Utilize libraries like Pandas to load the dataset into a DataFrame.
  + Use Pandas' describe() method to generate summary statistics such as minimum, maximum, mean, standard deviation, variance, and percentiles for each feature.
* For Data Visualization - Histogram Creation:
  + Employ Matplotlib, a widely-used plotting library in Python, to create histograms for each feature.
  + Iterate through each feature in the dataset, plot its histogram using Matplotlib's hist() function, and customize the plot as necessary to illustrate feature distributions effectively.
* Data Cleaning:
  + Identify and handle missing values using techniques like imputation or removal.
  + Detect and handle outliers that may affect the integrity of the dataset.
* Data Integration:
  + Merge or concatenate multiple datasets if required for analysis.
  + Ensure consistency in data formats and representations across integrated datasets.

2.4] Applications:

Descriptive Analysis: Summary statistics provide insights into the central tendency, variability, and distribution of features, aiding in descriptive analysis.

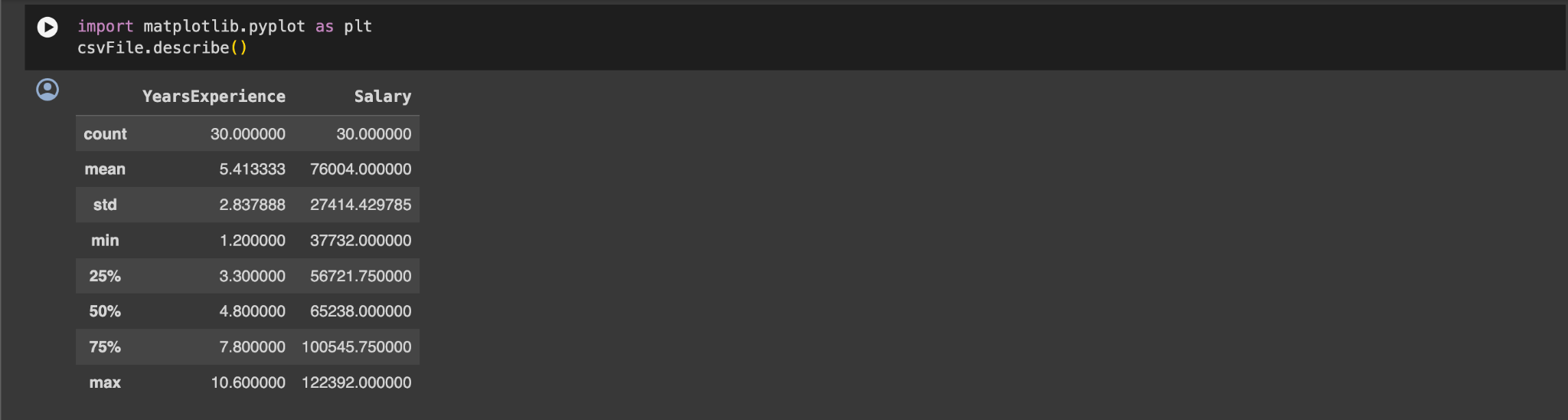
Exploratory Data Analysis (EDA): Histograms visualize the distribution of individual features, revealing patterns, trends, and potential outliers.

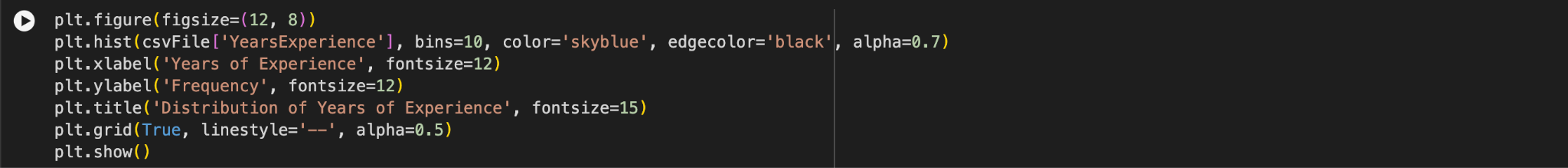
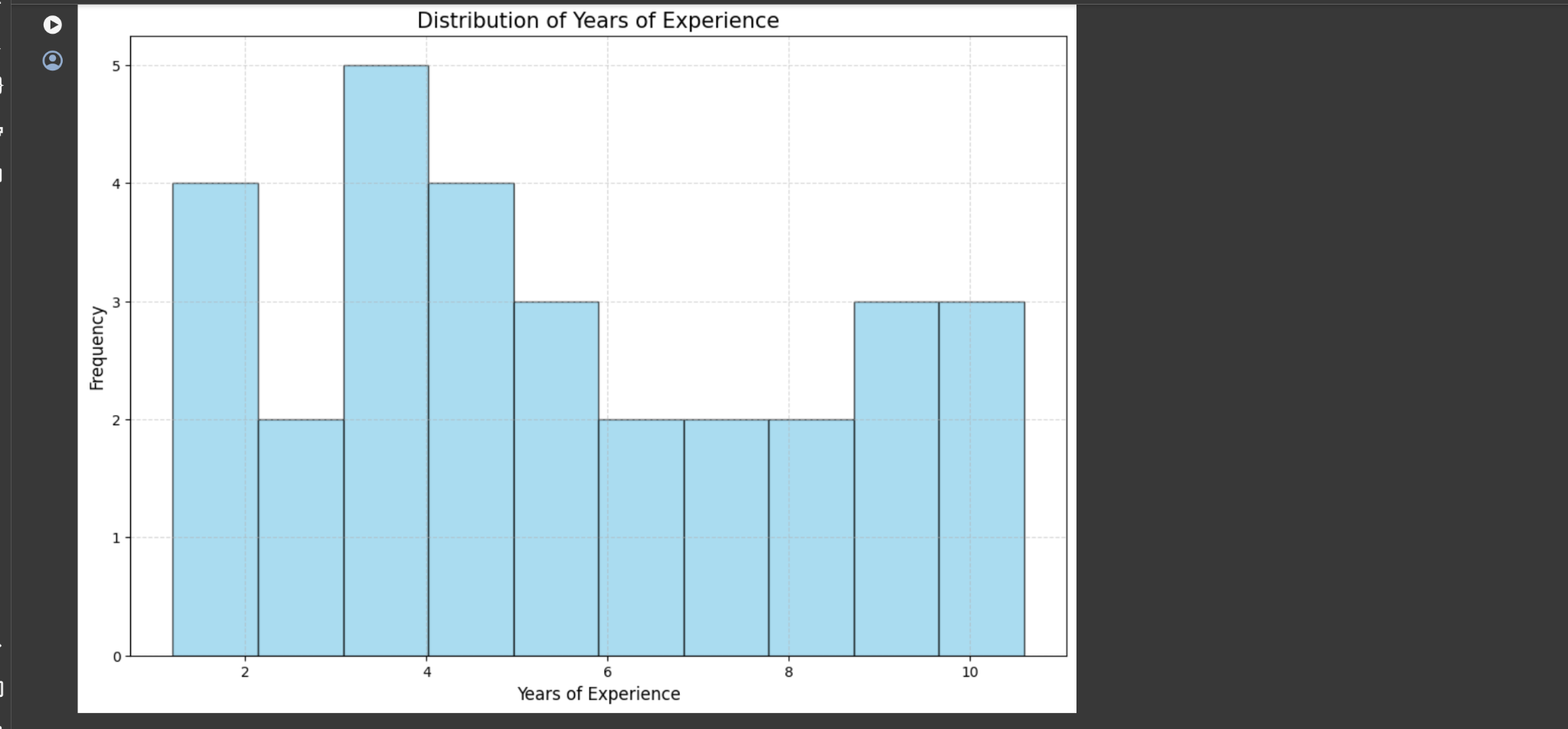
2.5] Limitations:

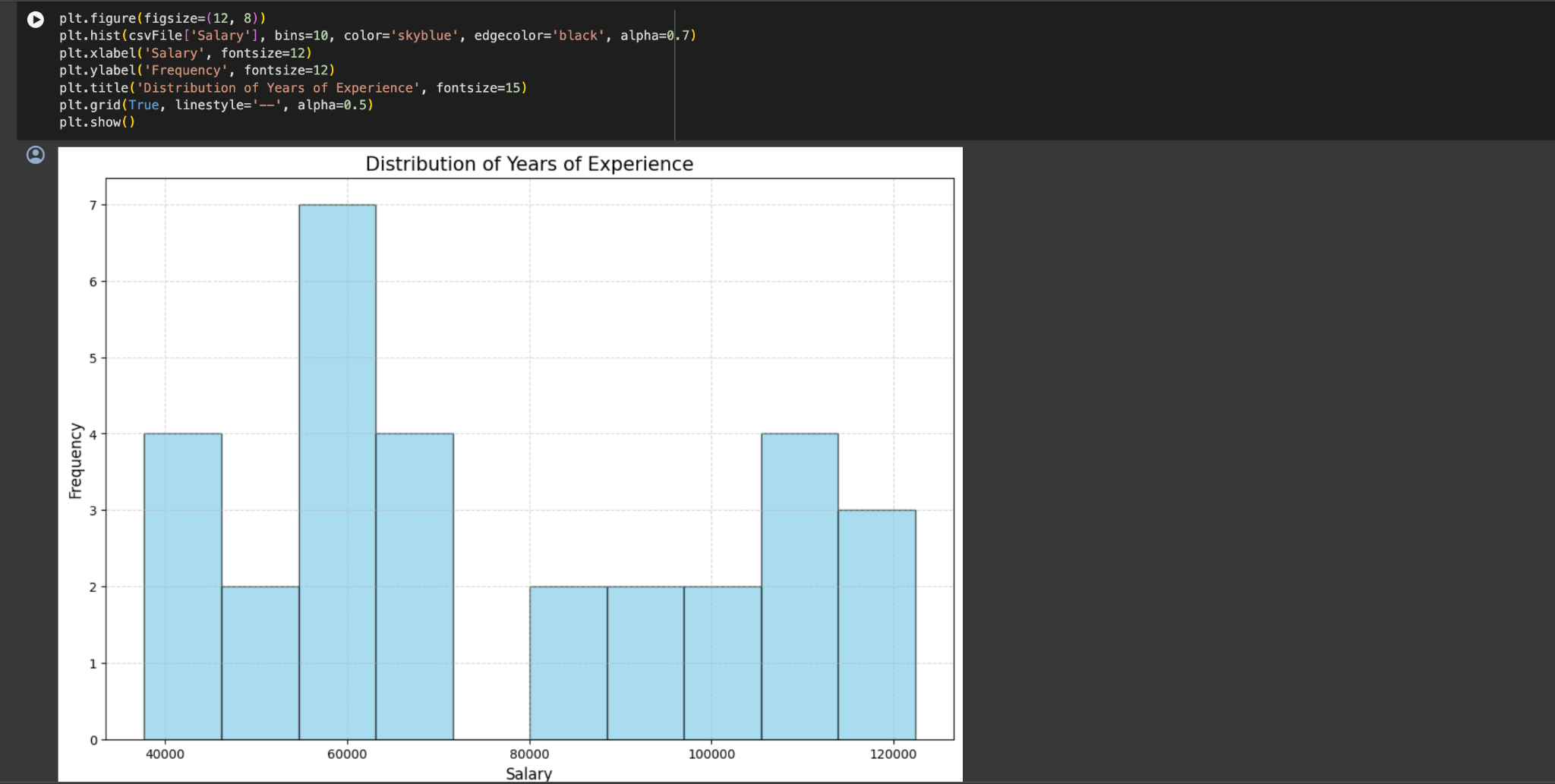
Data Quality: Incomplete or inaccurate data can lead to biased summary statistics and misleading visualizations.

Model Performance: The success of classification models depends on factors like feature selection, model choice, and hyperparameter tuning, which may require iterative experimentation.

2.6] Working:

****

****



2.7] Conclusion:

By computing summary statistics, visualizing feature distributions, and executing data cleaning, integration, transformation, and model-building processes, it addresses essential aspects of exploratory data analysis and classification tasks. While offering valuable insights and enabling the construction of predictive models, it's important to acknowledge potential limitations such as data quality issues and computational constraints. Nevertheless, leveraging these libraries empowers practitioners to navigate datasets efficiently, derive meaningful insights, and develop robust classification models for informed decision-making.

**ASSIGNMENT NO: 3**

3.1] Problem Statement -

Apply appropriate ML algorithm on a dataset collected in a cosmetics shop showing details of customers to predict customer response for special offers.

3.2] S/W Packages and Libraries used:

For the following assignment, the interpreter used was Google Collab and the Primary Library used was-

* Scikit-learn: Scikit-learn offers a wide range of machine learning algorithms and evaluation metrics, enabling the implementation and evaluation of the KNN classifier for predicting customer responses to special offers.

3.3] Theory-

* Data Preparation:
  + Load the dataset using Pandas to a DataFrame, ensuring it contains relevant customer details and responses to special offers.
  + Preprocess the data by handling missing values, encoding categorical variables, and splitting the dataset into features (input variables) and target variables (customer response).
* Feature Scaling:
  + Standardize the features using StandardScaler from Scikit-learn to ensure all features are on the same scale, preventing any particular feature from dominating the model training process.
* Train-Test Split:
  + Split the dataset into training and testing sets using train\_test\_split from Scikit-learn. This ensures the model is trained on one portion of the data and evaluated on another to assess its generalization performance.
* Model Selection and Training:
  + Apply the K-Nearest Neighbors (KNN) algorithm from Scikit-learn to predict customer responses to special offers. KNN is chosen for its simplicity and effectiveness in classification tasks.
  + Train the KNN classifier on the training data, tuning hyperparameters such as the number of neighbors (K) if necessary.

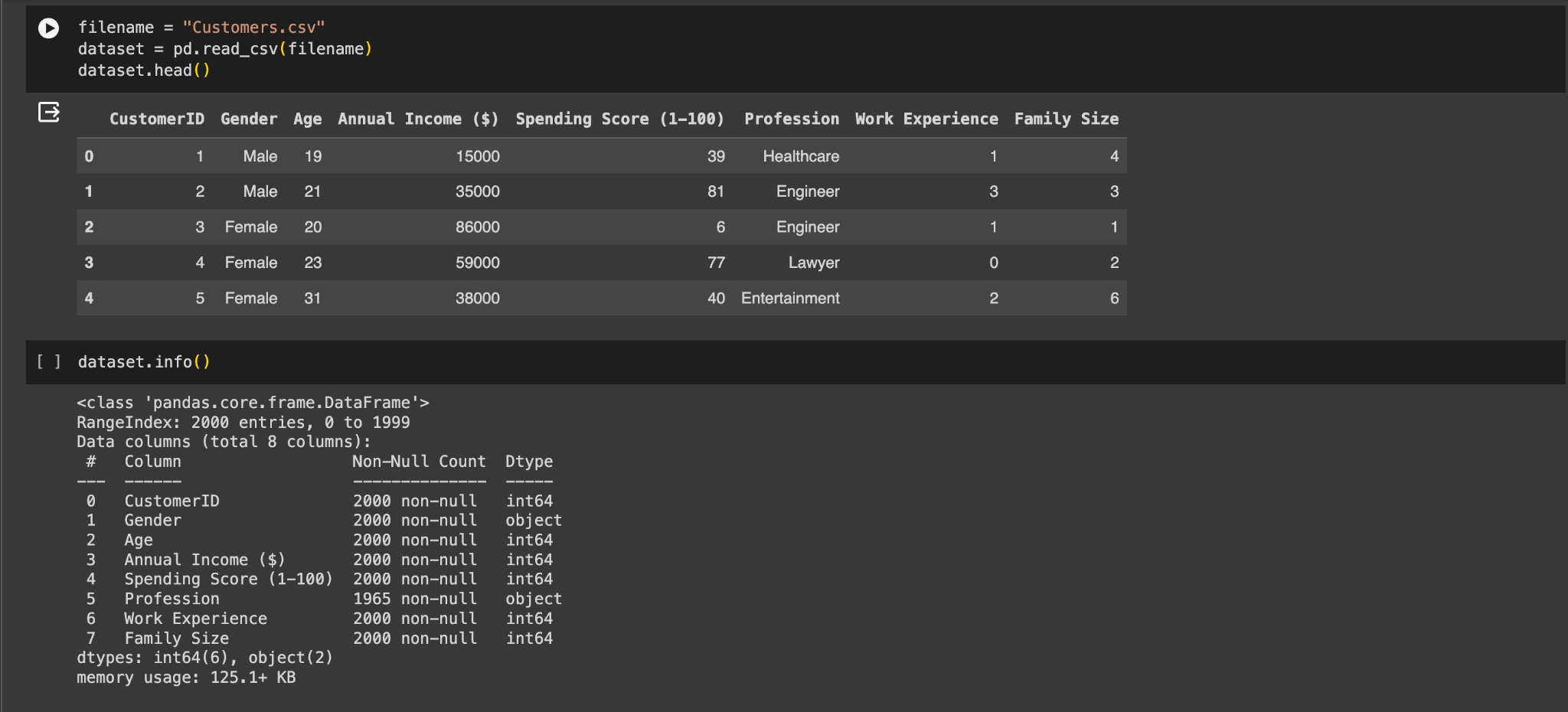
3.4] Applications:

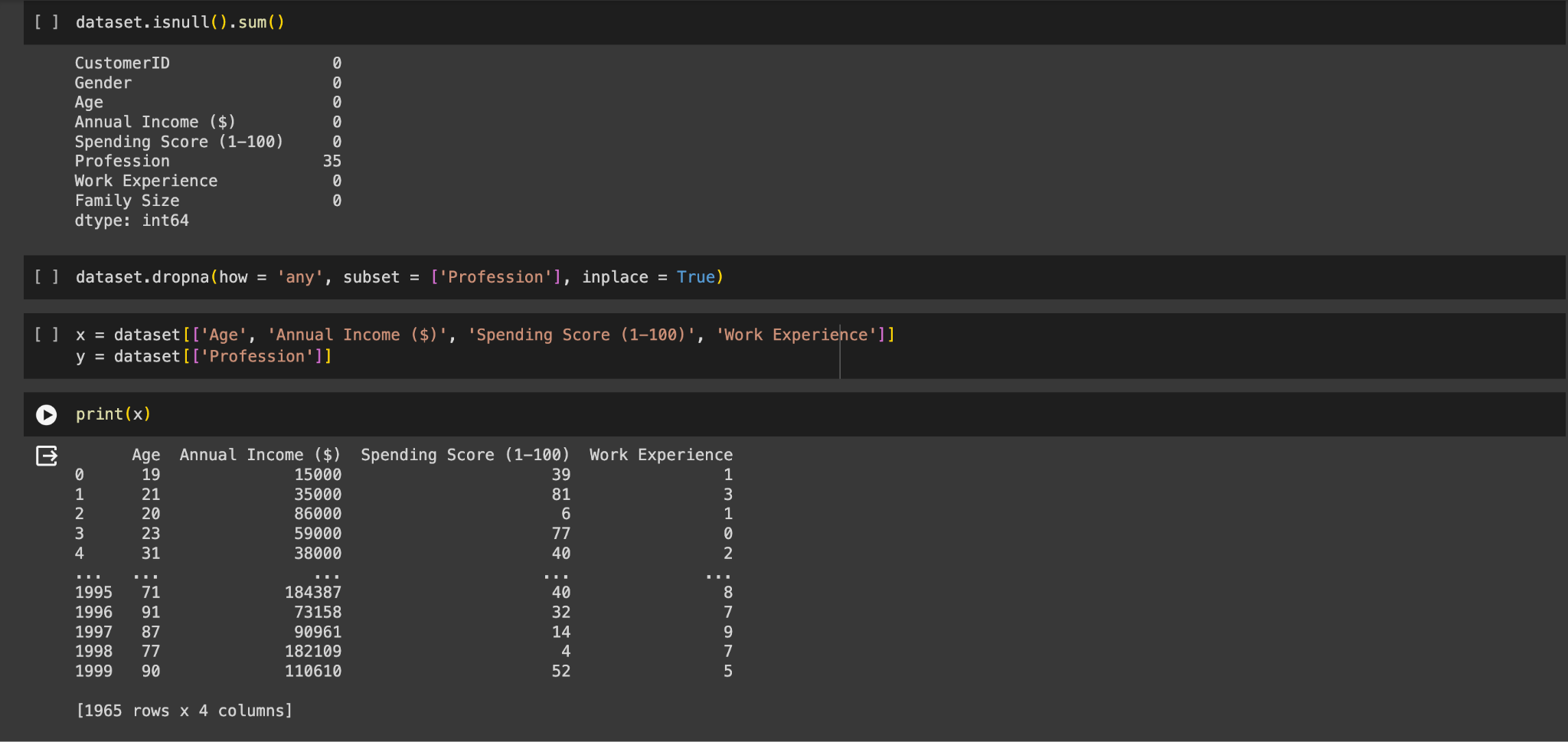
* Customer Relationship Management: Predicting customer responses to special offers can aid in personalized marketing strategies, enhancing customer satisfaction and loyalty.
* Sales Optimization: Understanding customer preferences enables the optimization of product offerings and promotional campaigns, leading to increased sales and revenue.

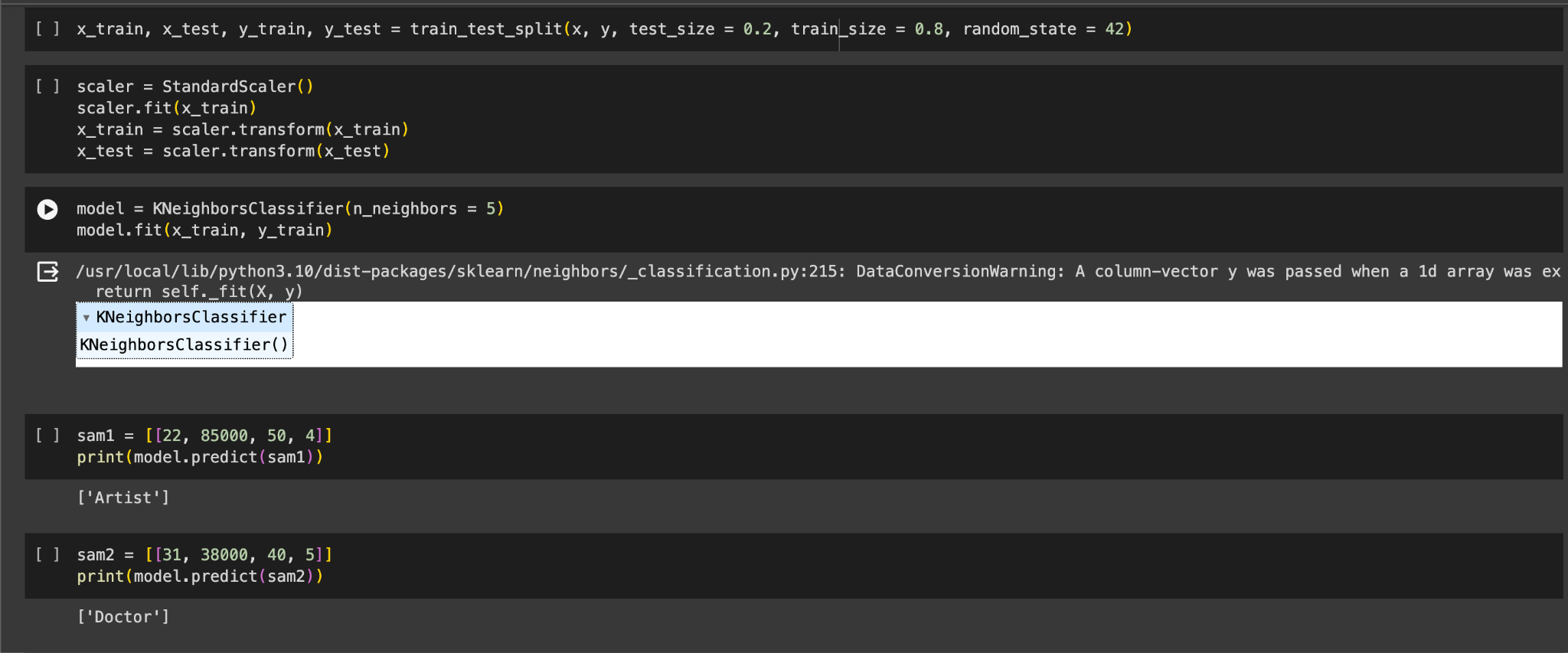
3.5] Limitations:

* Data Quality: The effectiveness of the predictive model heavily relies on the quality and representativeness of the dataset. Inaccurate or biased data may lead to unreliable predictions.
* Model Complexity: While KNN is simple and intuitive, it may not perform optimally with large datasets or in high-dimensional feature spaces. Other algorithms like Decision Trees or Random Forests could be explored for better performance in such scenarios.

3.6] Working:

****





3.7] Conclusion:

By following this methodology and understanding the applications and limitations of the assignment, practitioners can effectively apply machine learning techniques to predict customer responses in a cosmetics shop setting, leveraging the capabilities of libraries like Pandas, NumPy, and Scikit-learn.

**ASSIGNMENT NO: 4**

4.1] Problem Statement -

a) Apply Data pre-processing (Label Encoding, Data Transformation) techniques if necessary.

b) Perform data preparation (Train-Test Split)

c) Apply Machine Learning Algorithm

d) Evaluate Model.

e) Apply Cross-Validation and Evaluate Model

4.2] S/W Packages and Libraries used:

For the following assignment, the interpreter used was Google Collab and the Primary Library used were Pandas, Numpy, Matplotlib, and sklearn which we have discussed earlier.

4.3] Theory-

* Data Pre-processing:
* If the dataset contains categorical variables, apply Label Encoding or One-Hot Encoding using libraries like scikit-learn's preprocessing module to convert them into numerical representations suitable for machine learning algorithms.
* Data Preparation (Train-Test Split):
* Split the pre-processed dataset into training and testing sets using train\_test\_split from the scikit-learn library. This ensures that the model's performance is evaluated on unseen data.
* Apply Machine Learning Algorithm:
* Choose an appropriate machine learning algorithm based on the nature of the problem (classification, regression, etc.) and the characteristics of the dataset. Common algorithms include Decision Trees, Random Forests, Support Vector Machines (SVM), or Gradient Boosting Machines (GBM).
* Evaluate Model:
* Evaluate the trained model's performance on the testing dataset using evaluation metrics such as accuracy, precision, recall, F1-score, or mean squared error (for regression).
* Visualize the model's performance using appropriate plots such as confusion matrix, ROC curve, or learning curve.
* Apply Cross-Validation and Evaluate Model:
* Apply cross-validation techniques such as k-fold cross-validation to assess the model's generalization performance and robustness.
* Compute cross-validated performance metrics to get a more reliable estimate of the model's performance.

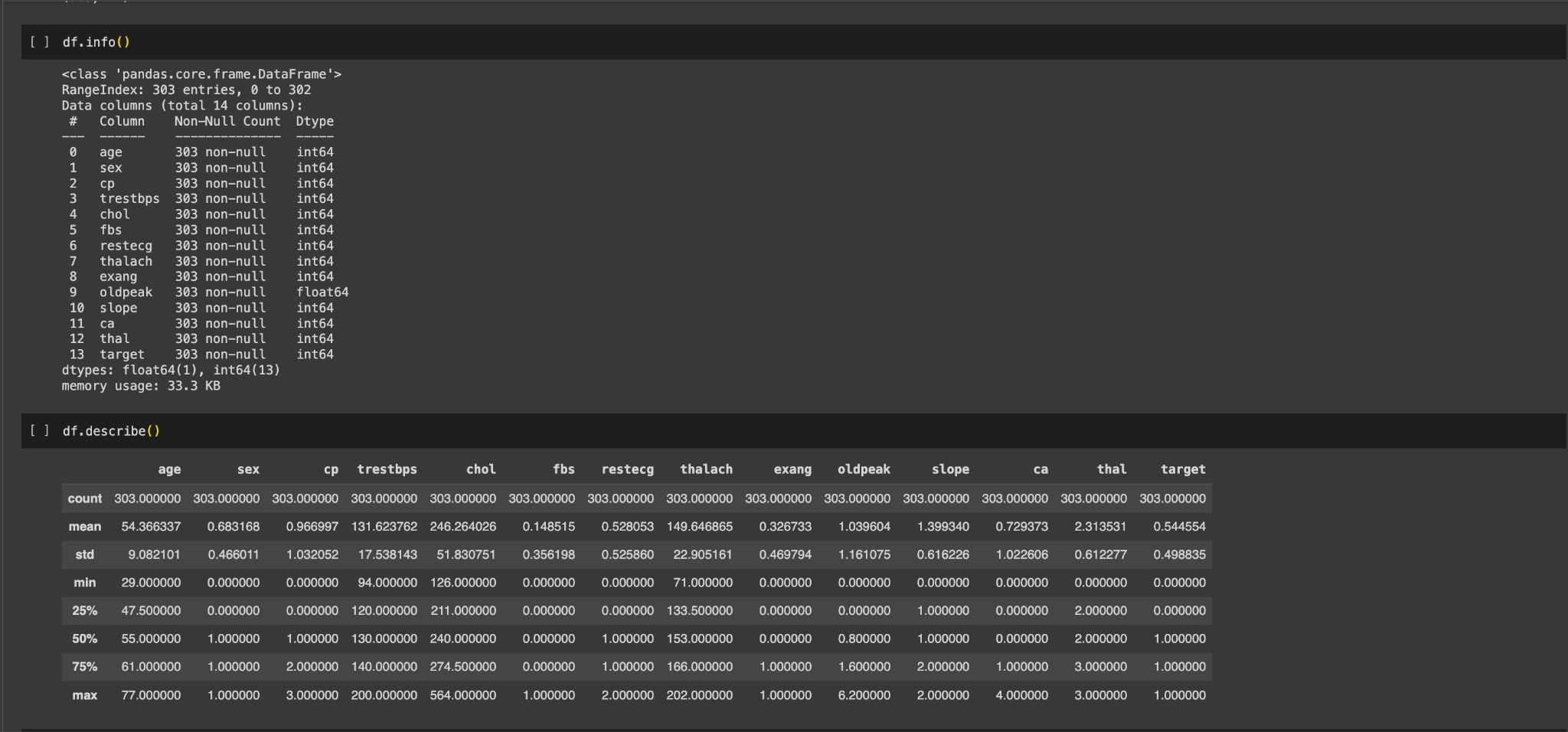
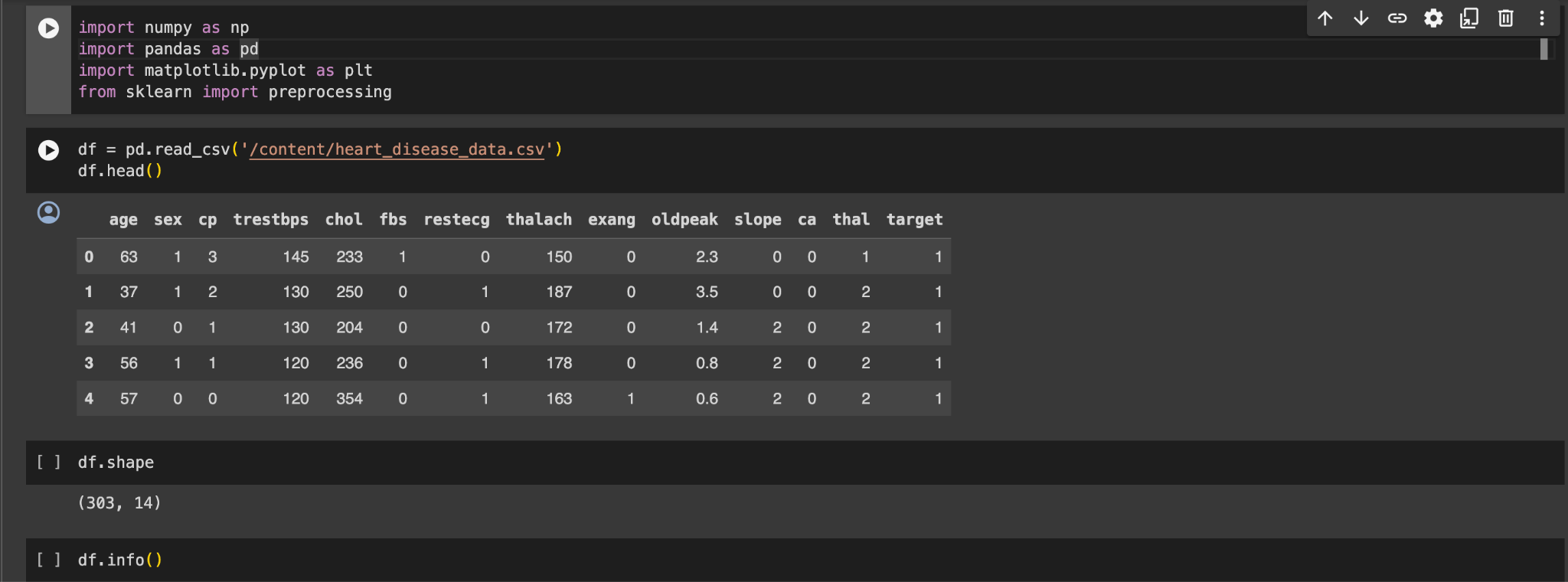
4.4] Applications:

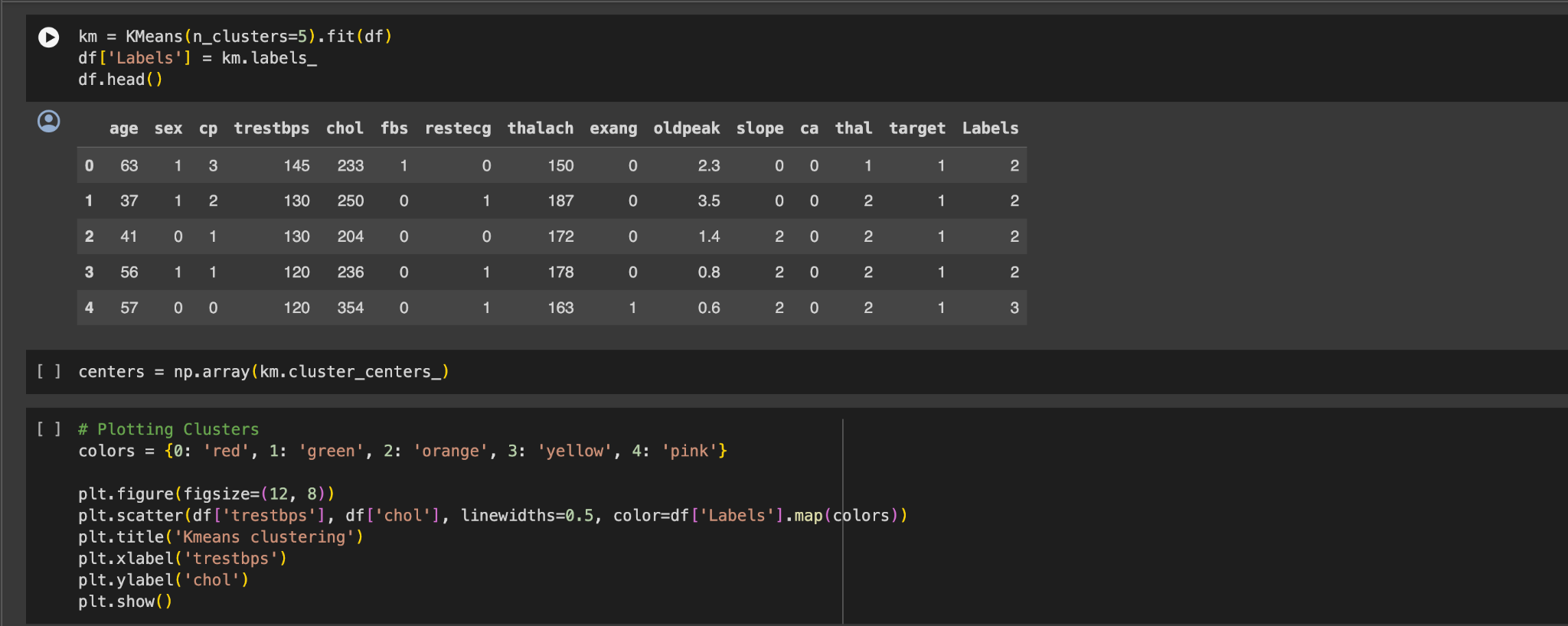
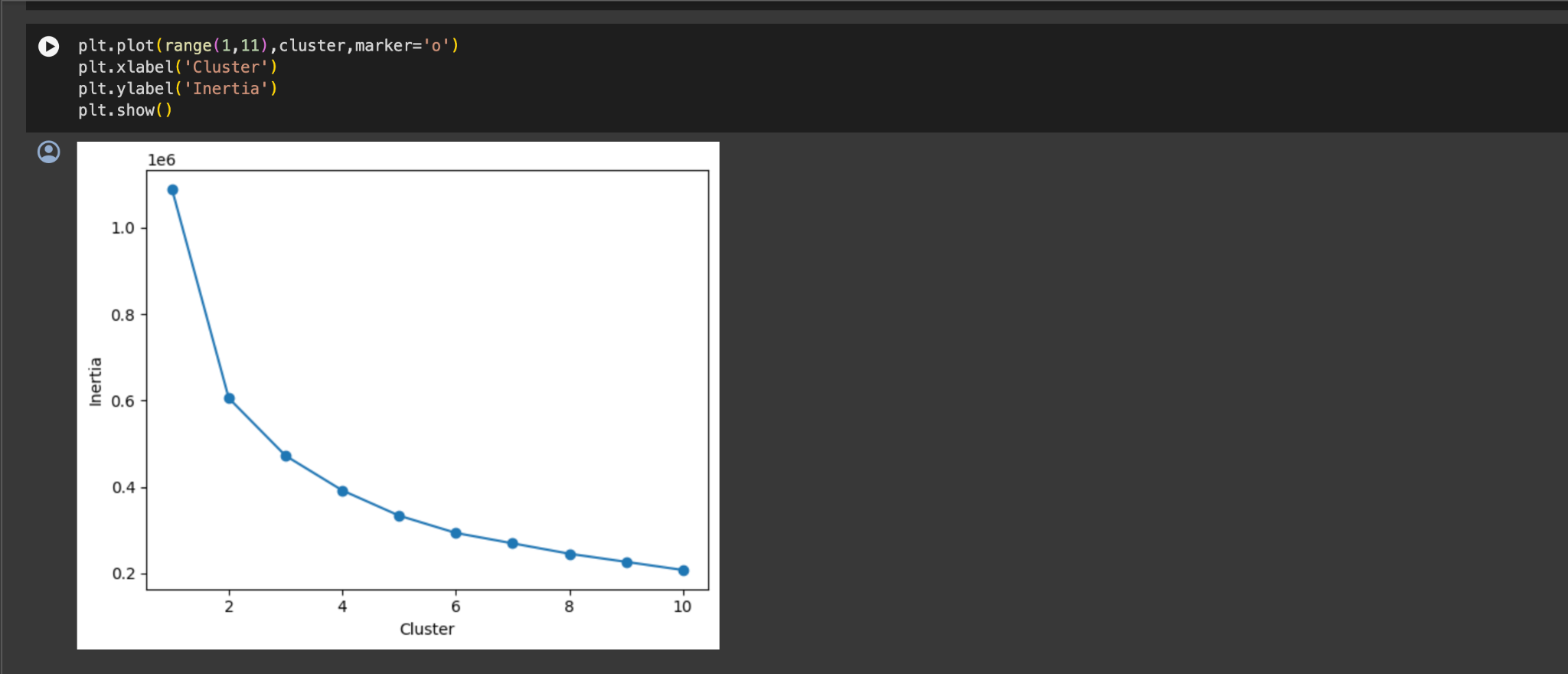
* Predictive Analytics: By applying machine learning algorithms to pre-processed data, businesses can make predictions and identify patterns in their datasets, leading to informed decision-making.
* Customer Segmentation: Machine learning models can help segment customers based on their behavior or characteristics, allowing businesses to tailor marketing strategies and offerings to specific customer groups.

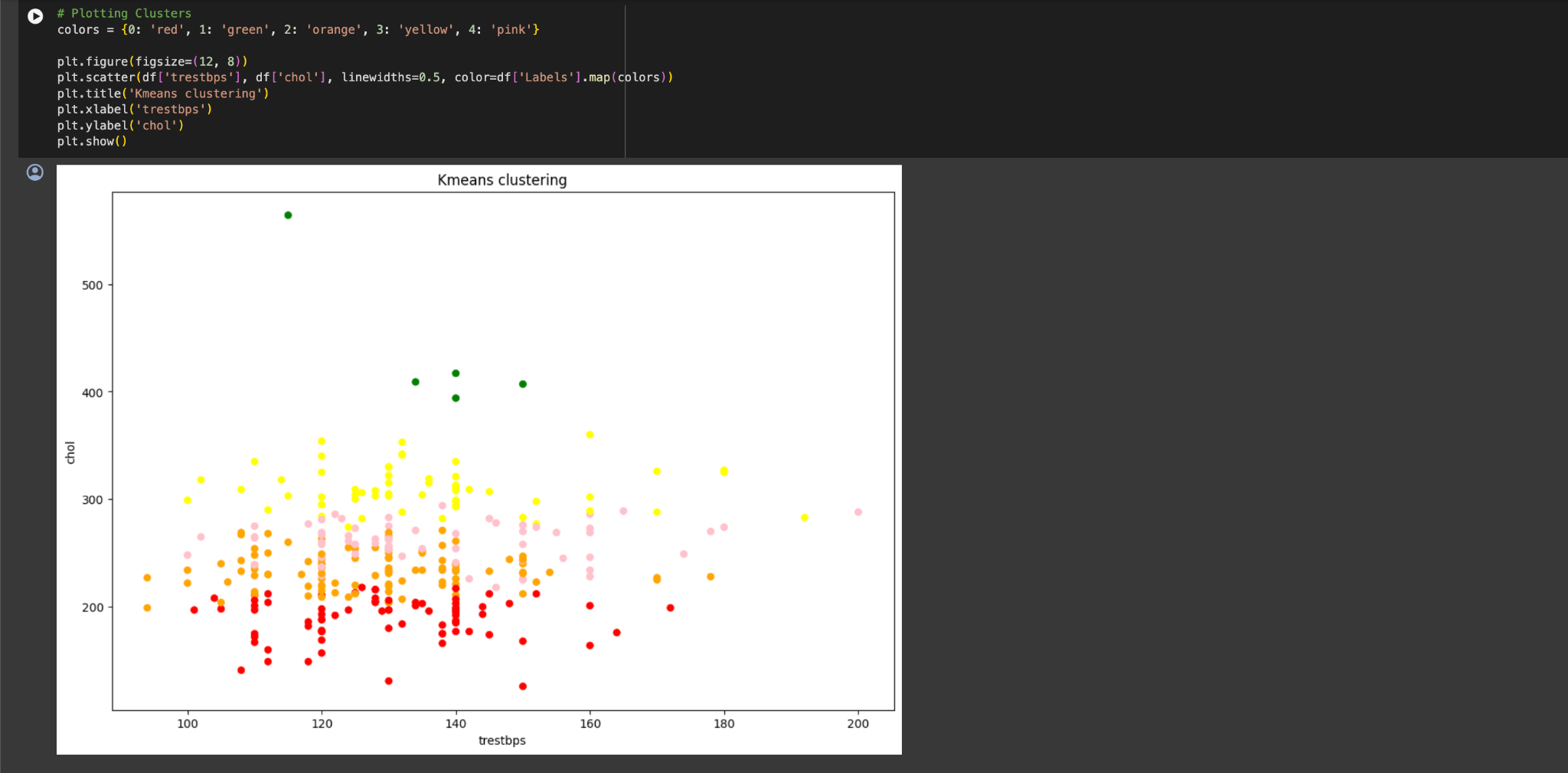
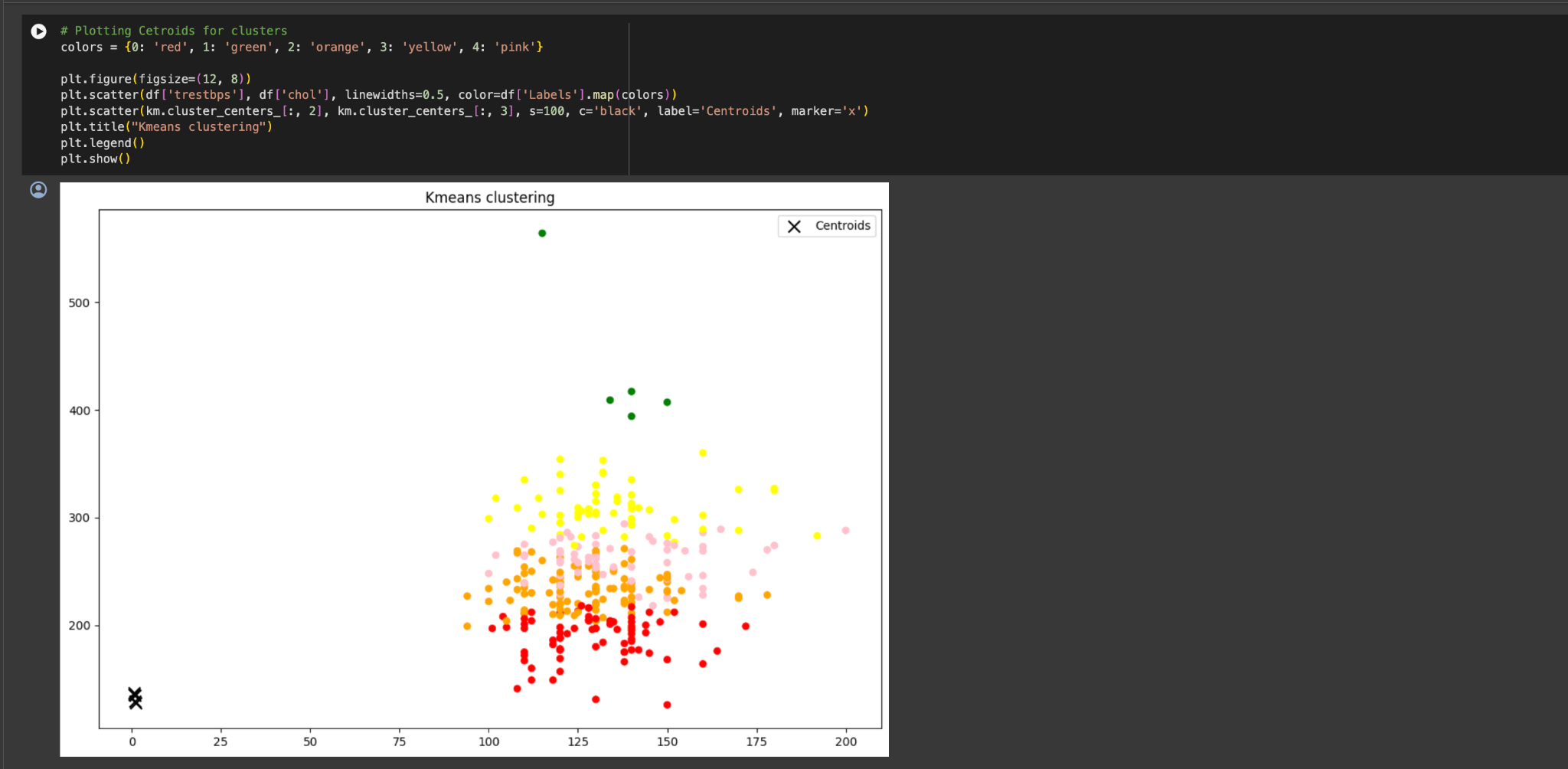
4.5] Limitations:

* Data Quality: The effectiveness of machine learning models heavily relies on the quality and representativeness of the data. Poor-quality or biased data can lead to inaccurate predictions and unreliable model performance.
* Model Selection: Choosing the most suitable machine learning algorithm for a given problem requires domain knowledge and experimentation. There is no one-size-fits-all solution, and different algorithms may perform differently on different datasets.

4.6] Working:







4.7] Conclusion:

By adhering to this methodology and understanding the assignment's applications and limitations, practitioners can effectively apply machine learning techniques to preprocess data, prepare it for modeling, train machine learning models, evaluate their performance, and assess their generalization capabilities through cross-validation.

**ASSIGNMENT NO: 5**

5.1] Problem Statement -

Visualize the data using Python by plotting the graphs for assignment no. 1 and 2. Consider a suitable data set.

a) Use Scatter plot, bar plot, Box plot and Histogram.

5.2] S/W Packages and Libraries used:

For the following assignment, the interpreter used was Google Collab and the Primary Library used was

* Seaborn: Seaborn is a statistical data visualization library in Python that provides a high-level interface for creating attractive and informative visualizations. It works well with Pandas DataFrames and offers a variety of plot types and customization options.

5.3] Theory-

* Data Selection:
  + Choose a suitable dataset that contains numerical and categorical variables, allowing for the creation of different types of plots such as scatter plots, bar plots, box plots, and histograms.
* Data Preprocessing:
  + Load the dataset into Python using Pandas.
  + Perform any necessary data cleaning and preprocessing steps, such as handling missing values or encoding categorical variables.
* Visualization:
* Utilize the Seaborn library for creating visualizations.
* Create scatter plots to visualize the relationship between two numerical variables.

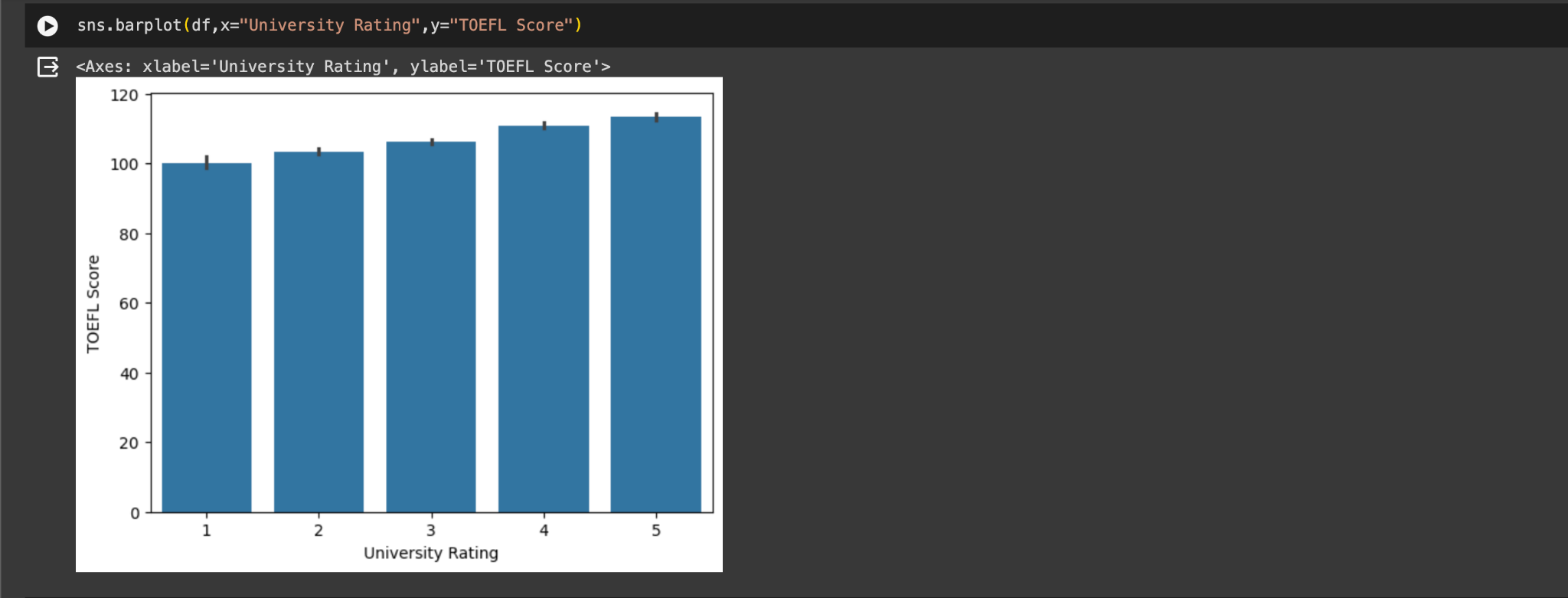
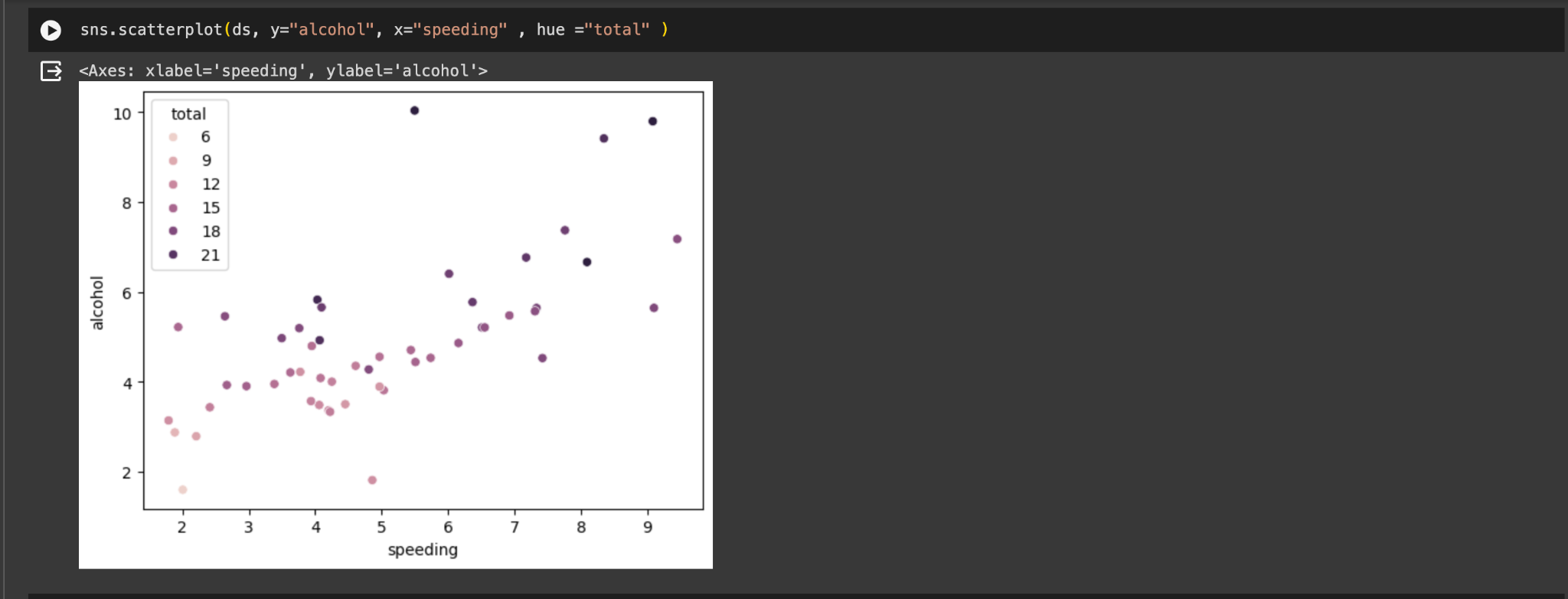
5.4] Applications:

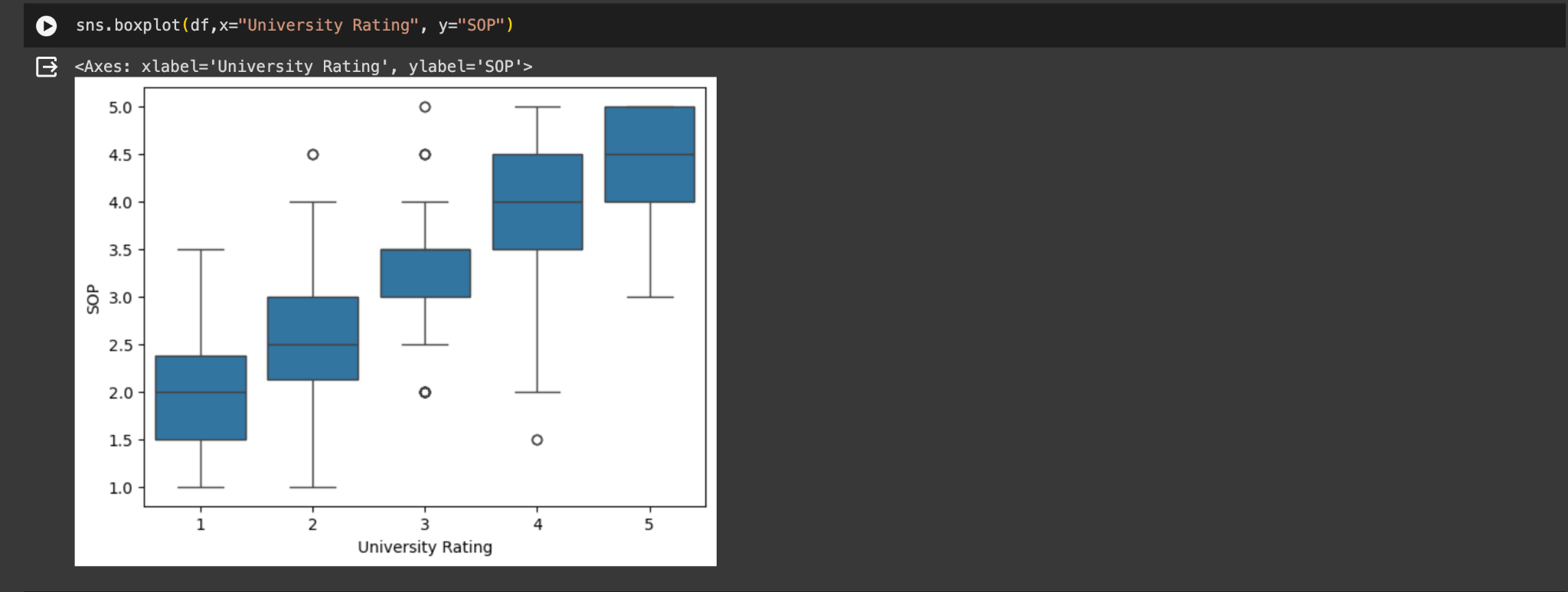
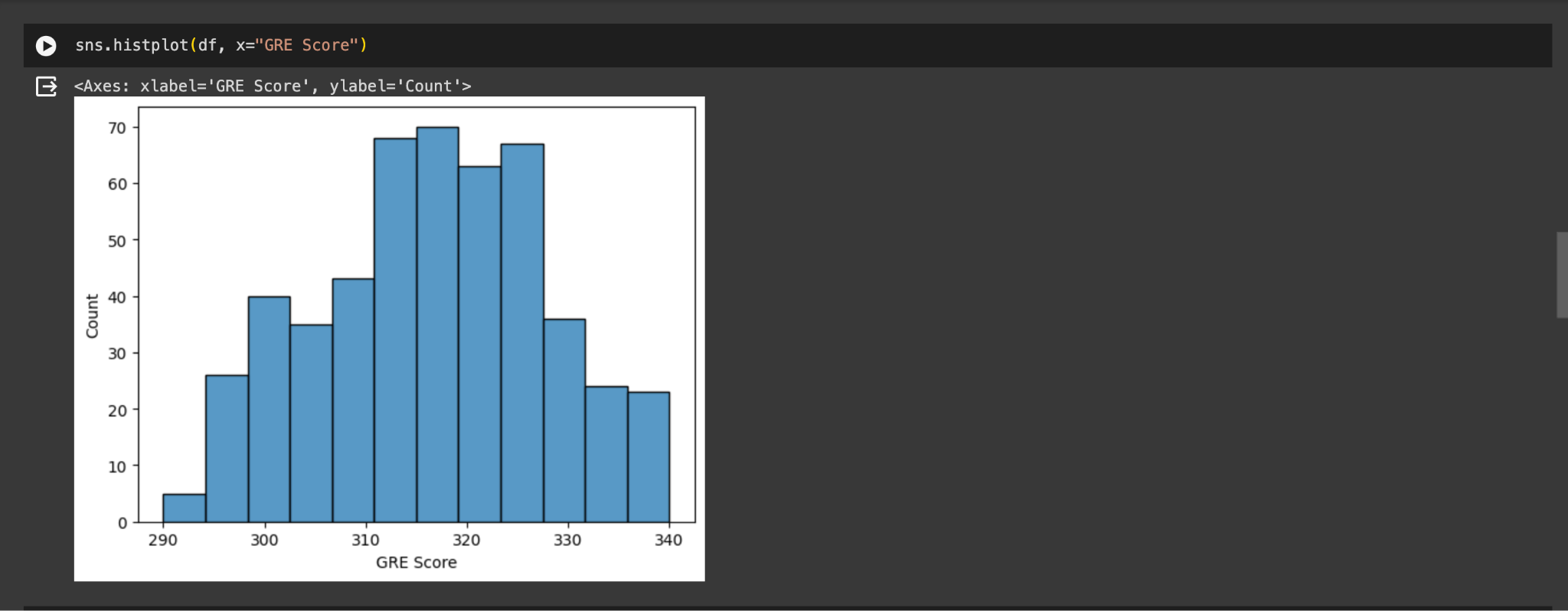
* Exploratory Data Analysis (EDA): Visualization techniques provide insights into the underlying structure and patterns within the data, aiding in understanding relationships and identifying trends.
* Data Communication: Visualizations serve as powerful tools for communicating findings and insights to stakeholders or decision-makers in a clear and concise manner.

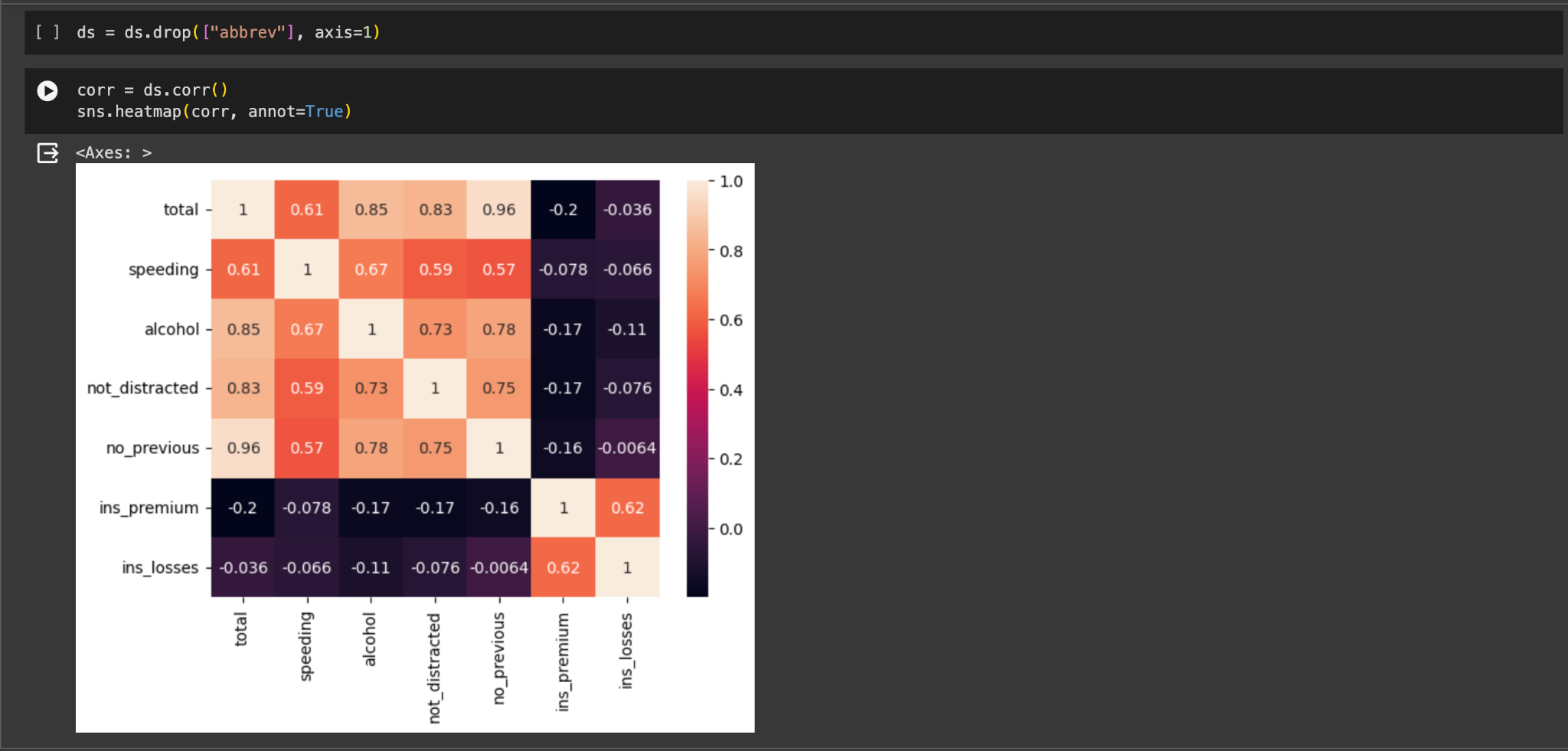
5.5] Limitations:

* Data Quality: Visualizations are only as good as the quality of the underlying data. Inaccurate or incomplete data may lead to misleading visualizations and erroneous conclusions.
* Interpretation: While visualizations can highlight patterns and relationships in the data, they may not always imply causation. Care must be taken to interpret visualizations within the context of the data and domain knowledge.

5.6] Working:

****

****

****

5.7] Conclusion:

By following the outlined methodology and utilizing the Seaborn library, you can effectively visualize the selected dataset using scatter plots, bar plots, box plots, and histograms. These visualizations will help in exploring relationships, understanding distributions, and identifying patterns within the data, thereby facilitating further analysis and decision-making processes.

**ASSIGNMENT NO: 6**

6.1] Problem Statement-

This data consists of temperatures of INDIA averaging the temperatures of all places month wise. Temperatures values are recorded in CELSIUS.

a) Apply Linear Regression using suitable library function and predict the Month-wise

temperature.

b) Assess the performance of regression models

6.2] S/W Packages and Libraries used-

Software Package: Python

Libraries Used:

pandas: For data manipulation and analysis.

scikit-learn: For implementing machine learning algorithms, including Linear Regression.

matplotlib: For data visualization.

6.3] Theory-

* Data Preprocessing:
  + Load the temperature data.
  + Prepare the data by separating features (months) and target variable (monthly temperatures).
  + Split the data into training and testing sets.
* Model Training:
  + Initialize a Linear Regression model.
  + Fit the model on the training data.
* Model Evaluation:
  + Predict the month-wise temperatures using the trained model.
  + Evaluate the model's performance using Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-Square metrics.

6.4] Application:

* The application of this regression technique is to predict the month-wise temperatures in India based on historical data.
* It can be used by meteorologists, climatologists, and environmental scientists for various purposes such as weather forecasting, climate trend analysis, and planning for agriculture and infrastructure.

6.5] Limitations:

* Linear Regression assumes a linear relationship between the independent and dependent variables, which may not always hold true in real-world scenarios.
* It may not capture complex patterns or non-linear relationships present in the data.
* The accuracy of predictions can be affected by outliers and noise in the data.
* The model's performance heavily depends on the quality and representativeness of the training data.

6.6] Working:

A white surface with a white line

Description automatically generated with medium confidence

A calendar with numbers and a few days

Description automatically generated with medium confidence

A table with numbers and letters

Description automatically generated

A screenshot of a computer

Description automatically generated

A white background with black text

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph with blue dots

Description automatically generated

A screenshot of a computer program

Description automatically generated

A white rectangular object with a black border

Description automatically generated

A graph with green dots and lines

Description automatically generated

6.7] Conclusion:

* Linear Regression can be effectively used to predict month-wise temperatures in India.
* The model's performance can be assessed using metrics like MSE, MAE, and R-Square.
* Despite its limitations, Linear Regression provides a simple and interpretable approach for temperature prediction, which can be valuable for various applications in climate science and related fields.

**ASSIGNMENT NO: 7**

7.1] Problem Statement-

To implement Classification techniques for following scenario:

Every year many students give the GRE exam to get admission in foreign Universities. The

data set contains GRE Scores (out of 340), TOEFL Scores (out of 120), University Rating

(out of 5), Statement of Purpose strength (out of 5), Letter of Recommendation strength (out

of 5), Undergraduate GPA (out of 10), Research Experience (0=no, 1=yes), Admitted (0=no,

1=yes). Admitted is the target variable.

The counselor of the firm is supposed to check whether the student will get an admission or

not based on his/her GRE score and Academic Score. So to help the counselor to take

appropriate decisions, build a machine learning model classifier using a Decision tree to

predict whether a student will get admission or not.

a) Apply Data pre-processing (Label Encoding, Data Transformation....) techniques if

necessary.

b) Perform data-preparation (Train-Test Split)

c) Apply Machine Learning Algorithm

d) Evaluate Model.

7.2] S/W Packages and Libraries used-

Software Package: Python

Libraries Used:

pandas: For data manipulation and analysis.

scikit-learn: For implementing machine learning algorithms, including Decision Tree Classifier.

matplotlib: For data visualization.

7.3] Theory-

* Data Preprocessing:
  + Load the data.
  + Perform any necessary data transformations (e.g., label encoding for categorical variables).
  + Check for missing values and handle them if necessary.
* Data Preparation (Train-Test Split):
  + Split the data into training and testing sets to train the model on one set and evaluate its performance on another.
* Model Training:
  + Initialize a Decision Tree Classifier model.
  + Fit the model on the training data.
* Model Evaluation:
  + Predict admission outcomes for the test data.
  + Evaluate the model's performance using metrics like accuracy, precision, recall, and F1-score.
  + Visualize the results if necessary.

7.4] Applications:

* The application of this classification technique is to predict whether a student will get admission to a foreign university based on their GRE score and academic performance.
* It can be used by education consultants, university admission offices, and students themselves to assess their chances of admission and make informed decisions.

7.5] Limitations-

* Decision Tree models are prone to overfitting, especially on noisy data or data with a large number of features.
* They may not capture complex relationships between variables as effectively as other algorithms.
* Decision Trees are sensitive to small variations in the data, which can lead to different tree structures and results.
* The interpretability of the model might be limited if the tree becomes too large and complex.

7.6] Working-

A white background with black text

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A white background with black text

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A diagram of a diagram

Description automatically generated

A screenshot of a computer code

Description automatically generated

A red and blue squares

Description automatically generated

A blue and red text on a white background

Description automatically generated

A collage of graphs

Description automatically generated

A white rectangular object with blue lines

Description automatically generated

7.7] Conclusion-

* The Decision Tree Classifier can be used effectively to predict admission outcomes based on GRE scores and academic performance.
* It provides a simple and interpretable model that can help counselors and students make informed decisions about university admissions.
* However, it's important to be aware of its limitations and consider other machine learning techniques for more accurate predictions, especially in complex scenarios with a large number of features.