Assignment 1: Pearson's Correlation Analysis on Penguins Dataset

1. Objective

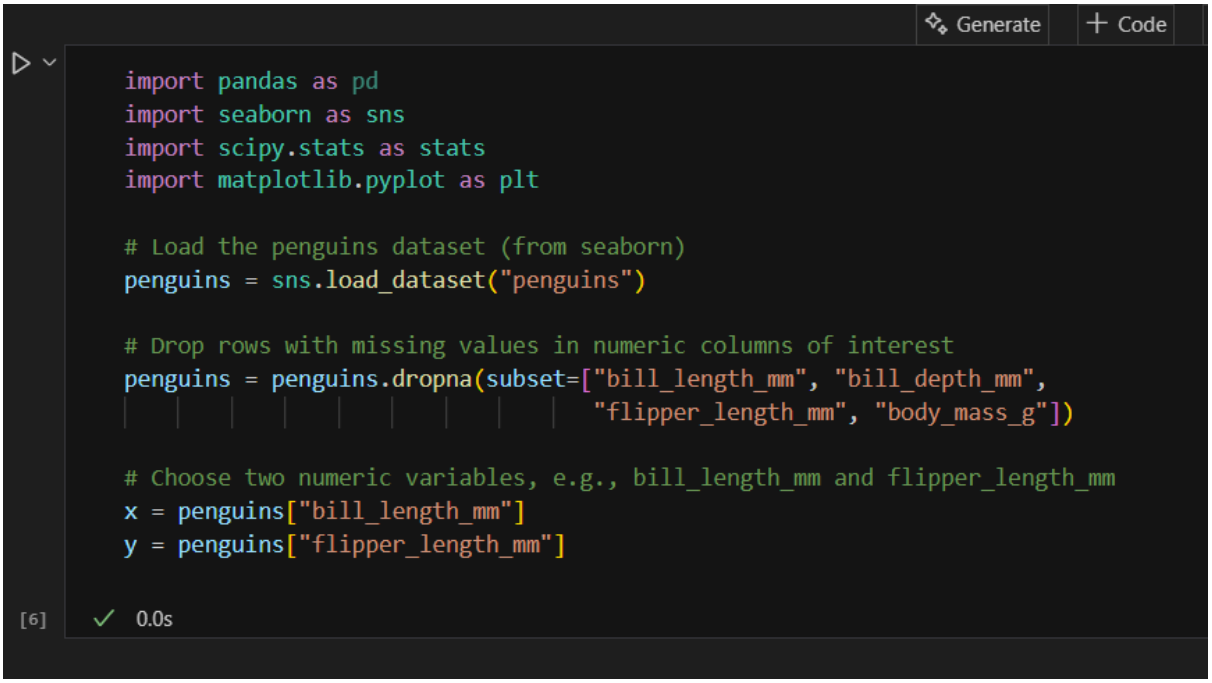
The objective of this assignment is to perform a **Pearson's correlation analysis** on the "penguins" dataset to explore the linear relationships between several numerical variables: `bill_length_mm`, `bill_depth_mm`, `flipper_length_mm`, and `body_mass_g`.

2. Methodology

The analysis was conducted using Python with the `pandas`, `seaborn`, and `scipy.stats` libraries.

- **Data Preparation**: The penguins dataset was loaded, and any rows with missing values in the relevant columns were dropped to ensure accurate calculations.
- **Pearson's r and p-value**: The Pearson correlation coefficient (`r`) and the corresponding p-value were calculated for the relationship between `bill_length_mm` and `flipper_length_mm` using `scipy.stats.pearsonr`.
- **Correlation Matrix**: A correlation matrix was generated for all four numeric variables using `pandas.DataFrame.corr()`. This matrix provides a comprehensive view of the pairwise linear relationships.
- **Visualizations**:
    - A **pairplot** was created to visualize the scatter plots for each variable pair and the distribution of each individual variable.
    - A **heatmap** was generated from the correlation matrix to provide a color-coded visual summary of the correlation coefficients.

3. Code and outputs



```python
# Compute Pearson's correlation
corr_coef, p_value = stats.pearsonr(x, y)
print(f"Pearson correlation between bill length and flipper length: {corr_coef:.4f}")
print(f"P-value: {p_value:.4e}")

# Correlation matrix among all numeric features
numeric_cols = ["bill_length_mm", "bill_depth_mm", "flipper_length_mm", "body_mass_g"]
corr_matrix = penguins[numeric_cols].corr(method="pearson")
print("\nCorrelation matrix:")
print(corr_matrix)
```

5] ✓ 0.0s

```
Pearson correlation between bill length and flipper length: 0.6562
P-value: 1.7440e-43

Correlation matrix:
                   bill_length_mm  bill_depth_mm  flipper_length_mm  \
bill_length_mm           1.000000      -0.235053           0.656181
bill_depth_mm           -0.235053       1.000000          -0.583851
flipper_length_mm        0.656181      -0.583851           1.000000
body_mass_g              0.595110      -0.471916           0.871202

                   body_mass_g
bill_length_mm        0.595110
bill_depth_mm        -0.471916
flipper_length_mm     0.871202
body_mass_g           1.000000
```

APR ASSIGN

assign.ipynb

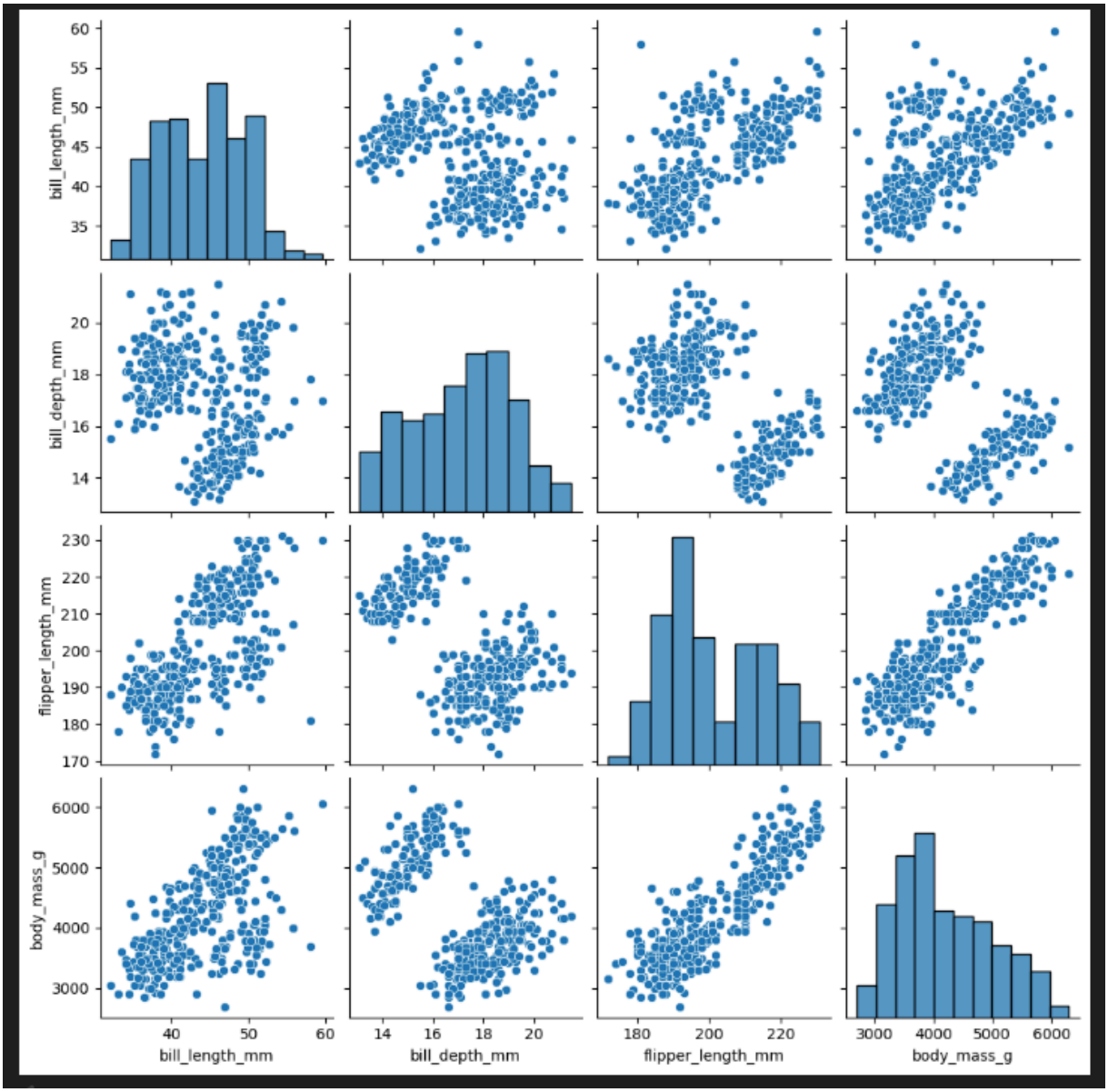
assign.ipynb > import pandas as pd

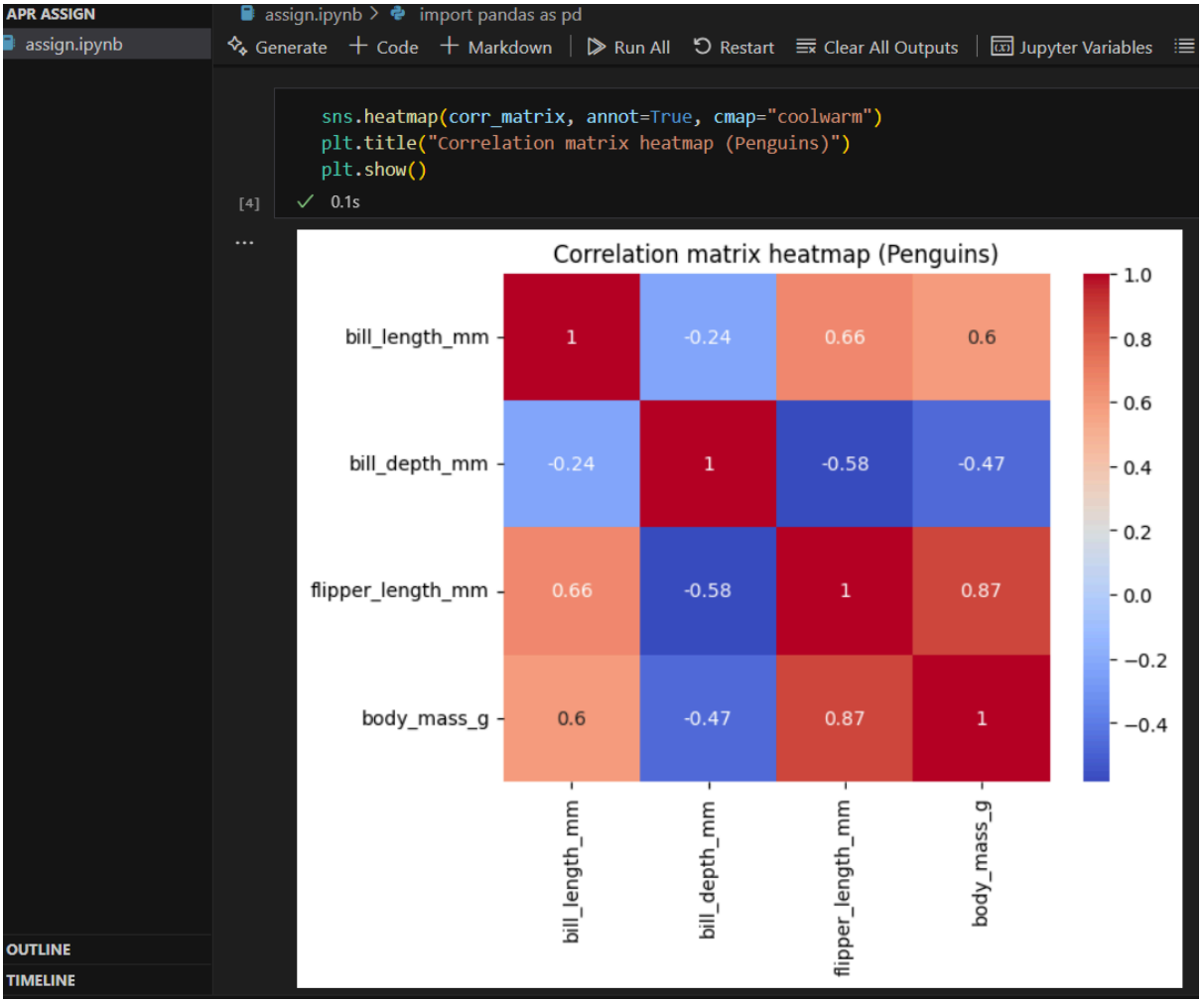
Generate | Code | Markdown | Run All | R

More...

```
# Visualize
sns.pairplot(penguins[numeric_cols])
plt.show()
```

[3] ✓ 1.1s





4. Results and Analysis

The analysis yielded the following results, which should be included as screenshots in your report.

- **Pearson's Correlation and P-value**: Pearson's correlation coefficient between `bill_length_mm` and `flipper_length_mm` is **0.6562**. The p-value is **1.7440×10−43**. This extremely small p-value indicates that the correlation is statistically significant.
- **Correlation Matrix**: The correlation matrix provides a complete overview of the linear relationships. Key findings include:
    - A strong positive correlation between `flipper_length_mm` and `body_mass_g` (**0.871202**). This makes sense as larger penguins (higher body mass) tend to have longer flippers.
    - A moderate positive correlation between `bill_length_mm` and `flipper_length_mm` (**0.656181**).
    - A weak negative correlation between `bill_length_mm` and `bill_depth_mm` (**-0.235053**).
- **Pairplot**: The pairplot visually confirms these relationships. The scatter plot for `flipper_length_mm` vs. `body_mass_g` shows a clear upward trend, indicative of

a strong positive correlation. The histograms on the diagonal show the distribution of each variable.

- **Heatmap**: The heatmap visually represents the correlation matrix, using a color gradient to show the strength and direction of the correlations. Warm colors (reds) indicate positive correlations, while cool colors (blues) indicate negative correlations. The annotated values on the heatmap match the correlation matrix, confirming the accuracy of the visualization.

These sections provide all the necessary components for your report. You can use this structure and the provided information to create a comprehensive and well-documented assignment report.