Phishing Analysis

Data Representation:

- The dataset was represented using values of 1 and -1 for most features.
- These values likely represent binary or categorical variables, where 1 might indicate the presence of a certain feature or characteristic, while -1 might indicate its absence or a negative condition.
- Some features, such as 'URL_Length', 'web_traffic', and 'Links_pointing_to_page', have three levels (-1, 0, 1), indicating different levels or categories of those features.

Analysis Conducted:

1. Initial Data Loading and Preparation:
   - The ARFF file containing the dataset was loaded into a DataFrame.
   - Byte strings in the dataset were decoded to regular strings.
   - The DataFrame was inspected to ensure proper loading and conversion.
2. Data Exploration:
   - Basic information about the dataset was examined using `df.info()` and `df.describe()` to understand its structure, data types, and summary statistics.
   - Missing values were checked to ensure data completeness.
3. Target Variable Distribution:
   - The distribution of the target variable 'Result', representing phishing vs. non-phishing websites, was visualized using a count plot (`sns.countplot()`).
   - This analysis helped understand the balance between the two classes and whether the dataset is imbalanced.
4. Correlation Analysis:
   - A correlation matrix was computed using `df.corr()` to analyze the relationships between different features.
   - The correlation matrix was visualized as a heatmap to identify strong correlations (positive or negative) between features.
   - This analysis provided insights into potential predictors and multicollinearity in the dataset.
5. Pairplot of Selected Features:
   - Pairwise scatter plots were created for selected features ('URL_Length', 'having_At_Symbol', 'SSLfinal_State') and the target variable ('Result').
   - The pairplot was colored by the target variable to observe differences in feature relationships between phishing and non-phishing websites.

Results of the Analysis:

- The dataset contains 31 features, represented using binary or categorical values of 1 and -1.
- The target variable 'Result' indicates phishing (-1) or non-phishing (1) websites.
- Basic information and statistics about the dataset were obtained, ensuring data quality and integrity.
- The target variable distribution revealed that the dataset may be imbalanced, with more non-phishing websites than phishing websites.
- The correlation analysis identified some features that are strongly correlated, which can be considered during feature selection or model building.
- Pairplot analysis showed differences in feature relationships between phishing and non-phishing websites, indicating potential discriminatory power of these features.

Conclusion:

The analysis provided insights into the dataset's structure, quality, and relationships between features, laying the groundwork for further exploration, feature engineering, and predictive modeling to detect phishing websites accurately.