I: DEPLOYED APPLICATION LINK

https://clarice-chen.shinyapps.io/final_version/

The app can be opened by the above link.

II: GITHUB LINK

https://github.com/meghapola/Applied-Data-Science-Project-2

The entire project, including all related files, is available at this GitHub link. The COMBINED CODE FINAL.R file is the final R script, and the README.md file explains how the website operates.

III: EXPLAINING APP'S FUNCTIONALITIES AND HOW TO USE IT

Jiaqi: This section of the Shiny application facilitates dynamic data import and exploration by allowing users to upload datasets in multiple formats, including .csv, .xlsx, .json, and .rds. Upon file selection, the application automatically detects the file type and applies the appropriate loading function to parse and store the dataset in a reactive object. Additionally, users can opt to load pre-defined sample datasets such as mtcars or iris for demonstration purposes. The uploaded or selected dataset is then rendered as an interactive data table (DT), enabling users to inspect raw data before proceeding to further preprocessing and analysis. This functionality ensures flexibility in data handling while providing an intuitive interface for preliminary data exploration within the Shiny framework.

Safia: Upon uploading a dataset, the data cleaning and preprocessing section of the app provides users with an interactive interface to clean and prepare their selected datasets before further analysis. This step is important as it will allow a user to ensure consistency, remove inconsistencies, and enhance the quality of their dataset. The app allows users to dynamically select cleaning steps and instantly preview changes in the dataset. Importantly, because the cleaning steps can be manually selected by clicking checkboxes, users can create different versions of clean datasets and select cleaning steps that give them the clean dataset they desire. Specifically, users can play with the following preprocessing functionalities:

Standardizing Column Names: Converts column names to lowercase and replaces spaces with underscores to maintain consistency.

Removing Duplicates: Detects and removes duplicate rows to avoid redundancy.

Handling Missing Values (Mean & Median): Users can opt for either mean or median imputation, replacing missing numeric values with their respective statistical measures.

Text Cleaning: Options include trimming whitespace, converting text to lowercase, and removing special characters to ensure uniform formatting.

Date Standardization: Converts date columns into a consistent YYYY-MM-DD format and allows for the extraction of features such as year, month, and day.

Outlier Handling (Z-Score & Mean): Users can either remove outliers using a Z-score threshold (greater than 3) or replace them with the mean value.

Fixing Inconsistent Categorical Labels: Standardizes text-based categorical values to ensure consistency.

Dropping High-Missing Columns: Removes columns where a significant portion of values are missing, ensuring a cleaner dataset.

Once a user selects their desired pre-processing steps, they can click the 'Apply Cleaning' button, which will update the dataset preview and allow users to inspect changes automatically. Then, the new dataset can be exported and used for further analysis.

Megha: The next stages of data preprocessing allows for users to continue formatting the dataset before analysis. For numeric data, the user will have the option to either standardize the data using z-scores or normalize the numeric columns using min-max scaling. The next functionality is an option to encode categorical features. When this option is chosen, the user can convert the categorical features using one-hot encoding. As a result, each of these features will be transformed into a set of binary columns, where each column represents a unique category with values of 0 or 1 that correspond to either the absence or presence of that category. There are also functionalities that allow the user to detect and handle outliers. The first two options, as mentioned, are to replace the outliers with the mean or use a cap of a Z-score > 3, removing any data points with higher Z-score values. If using the Z-score cap is chosen, any outliers will be replaced with a maximum or minimum value, depending on if the outlier is above or below the upper or lower limit respectively, rather than removing the outliers entirely. The user can use this option when they do not want to lose any data points but want to handle the influence of any extreme values on the data analysis. The last option is to remove the outliers using the Interquartile Range (IQR). In this case, the lower bound is Q1 - 1.5 * IQR, and the upper bound is Q3 + 1.5 * IQR. Any points outside these bounds are removed.

Jingxi: In the Scaling and Encoding tabs, an instantly updating preview table is added showing the changes after the operations. Other than the one-hot encoding, label encoding is also added to the Encoding tab. Label encoding replaces distinct categories with integer values. These transformations are particularly helpful when preparing datasets for machine learning algorithms that require numeric inputs. Additional transformations are accessible through the Polynomial and Interaction tabs. The polynomial feature generator raises a selected numeric column to a specified power (e.g., squaring, cubing) to capture non-linear relationships. Meanwhile, the interaction feature creator multiplies two numeric columns, creating a new column that may reveal interaction effects between variables. Both transformations add new columns to the existing dataset, instantly updating the preview to reflect the augmented feature space. The Datetime tab provides functionalities for date-based feature engineering. Users select a date or date-like column, and if necessary, the app will attempt to parse it into a valid Date format (YYYY-MM-DD). Once successfully parsed, columns for year, month, and day are automatically generated, which can be crucial for time-series analysis or seasonal pattern detection. Finally, the Save tab offers a simple way to download the fully processed dataset as a CSV file. The downloaded filename includes the current date for easy version tracking, ensuring that changes made during the entire workflow are saved permanently.

Yixuan: The main purpose of the EDA section is to gain a deep understanding of the data structure, characteristics, and potential patterns through visualization and statistical methods, providing a basis for subsequent modeling or decision-making. Therefore, in this section of the web application, users can view data previews, data information, categorical variable analysis, and numerical variable analysis in the display area below. This section allows users to observe basic data information such as dimensions and variable types. Additionally, users can see the categories present in the data and statistical summaries of numerical variables, including mean, median, variance, and standard

deviation. To visually display data characteristics and multivariate relationships, users can generate charts using options available on the left panel. Specific chart types include scatter plots, histograms, box plots, and time series plots. Users can manually select the x and y variables, choose the chart type, and input custom labels for the x-axis, y-axis, and chart title. Furthermore, to enhance clarity, users can manually adjust axis intervals. For time-series data, users can also select the desired date format.

IV: EACH MEMBER'S CONTRIBUTION

Loading Datasets & Data Cleaning and Preprocessing: Safia, Megha, Jiaqi
Feature Engineering and EDA: Jingxi, Yixuan
Report: All