BlueStats – User Guide and Functionality Report

GitHub link: https://github.com/WAWQAQMAKABAKA/5243project2_team6

Introduction

BlueStats is an interactive Shiny application, which is designed to simplify the process of data exploration, cleaning, feature engineering, and visualization. By using R's Shiny framework and leveraging popular packages such as shinythemes, DT, tidyverse, and corrplot, the app enables users to efficiently work with datasets by helping them through each critical step of the data analysis workflow.

Key Functionalities

1. Welcome

Overview and Instructions:

The landing page welcomes users and provides a step-by-step guide on how to navigate the app. It outlines the primary functionalities: data upload, preprocessing, feature engineering, and visualization.

Contributors & License:

The guide also credits the contributors and states that the application is distributed under the MIT License.

2. Data Upload

Input Options:

Users can either upload their own file with supported formats include CSV, Excel (.xls/.xlsx), JSON, and RDS, with a maximum file size of 1GB, or use an example dataset, including Iris, Motor Trends Cars, and the Penguins dataset.

Data Preview:

After the data upload or selection of an example dataset, a preview will be displayed using an interactive table. A summary box shows essential dataset details, such as the number of rows and columns or dataset description.

3. Preprocessing

Data Cleaning & Transformation:

Users can apply multiple preprocessing operations:

a. Handle missing values (removal, mean, or median imputation)
b. Remove duplicate records
c. Encode categorical variables into numeric form
d. Handle outliers using standard deviation-based removal or winsorization
e. Normalize numeric features using z-score scaling

For example, values beyond ±3 standard deviations from the mean are considered outliers and can be removed or capped.

Processed Data Preview:
A data table displays the processed dataset, reflecting the changes based on selected preprocessing operations.

4. Feature Engineering
Creating New Features:
Users can select two numeric columns and choose an arithmetic operation (addition, subtraction, multiplication, or division) to create new features. If users adjust the preprocessing steps later, these new features will be recalculated automatically.

For example: Creating a new column by multiplying price and number_of_reviewers

Engineered Data Preview:
The results are displayed in a separate data table, allowing users to verify that the new features have been correctly integrated.

5. Visualization & Exploratory Data Analysis (EDA)
Data Filtering:
A dynamic filter lets users select a numeric column and set a range for data filtering, ensuring visualizations of the desired subset of data. The app supports several types of plots: Scatter Plot, Histogram, Box Plot, Correlation Matrix. Also, users could customize the color of the plot, the plot size, adding regression line/spline and download.
Summary Statistics:
The summary statistics (min, median, mean, max, standard deviation, and missing values count) for numeric columns will be displayed in a tidy table.

UI/UX

The application uses a clear layout with tab navigation and the ‘shinythemes : : cosmo’ theme for a clean and consistent design. The welcome page provides user guidance, and each module is intuitive for non-technical users.

The preprocessing module allows users to clean and transform the targeted dataset. A summary of all applied processing steps will be displayed in the “Processing Report” section, which helps users to keep track of data transformation. Additionally, each step includes a short description explaining what it does to the dataset.

Deployment and Access

The application is deployed through shinyapps.io. The application has been deployed on a hosting platform to facilitate easy access without the need for local setup. You can access the app through the github we shared at the beginning of the report and run the code from “app_314.R”.

Conclusion

Our Shiny app achieves a complete interactive data analysis pipeline. The BlueStats is a comprehensive and user-friendly tool for performing a full spectrum of data analysis tasks, from initial data upload and preprocessing to advanced visualizations and feature engineering. Its design and interactive components make it an ideal application for both beginners and experienced data scientists who wish to streamline their analysis workflow.

Team Members Contributions:
Anqi Wu: Loading dataset, Data cleaning and Preprocessing, User Interface (UI) and User Experience (UX).
Keito Taketomi: Feature Engineering, EDA, User Interface (UI) and User Experience (UX), Web Application Functionality.
Ziyue Gao: Writing the report.
Yixin Xiao: Feature engineering module, Advanced data preprocessing, UI refinement, improve the report.