Data Science Toolkit

An interactive R Shiny web application for data exploration, cleaning, feature engineering, and analysis.

Overview

The Data Science Toolkit provides a user-friendly interface for common data science workflows:

- **Data Loading**: Upload datasets in various formats (CSV, Excel, JSON, RDS) or use built-in sample datasets
- **Data Cleaning**: Handle missing values, remove duplicates, and apply transformations
- **Feature Engineering**: Create new features, extract components from dates and text, and generate interaction terms
- **Exploratory Data Analysis**: Visualize data distributions, correlations, and relationships with interactive plots
- **Filtering**: Apply dynamic filters to focus on specific data subsets

Setup Instructions

Prerequisites

- R (version 4.0.0 or higher)
- RStudio (recommended for local development)

Installation

1. Clone this repository or download the source files
2. Open the project in RStudio or your preferred R environment
3. Install the required packages by running:

```
# Install required packages
install.packages(c(
  "shiny", "shinydashboard", "shinycssloaders", "DT", "dplyr",
  "tidyr", "ggplot2", "plotly", "readr", "readxl", "jsonlite",
  "data.table", "corrplot", "caret", "shinyjs", "shinyWidgets",
  "GGally", "moments", "lubridate"
))
```

Running Locally

To run the application locally:

1. Open the `app.R` file
2. Click "Run App" in RStudio, or run the following command:

```
shiny::runApp()
```

Deployment to shinyapps.io

To deploy the application to shinyapps.io (https://www.shinyapps.io/):

1. Create an account on shinyapps.io if you don't have one
2. Install the `rsconnect` package:

```
install.packages("rsconnect")
```

3. Set up your account credentials:

```
rsconnect::setAccountInfo(
  name = "your-account-name",
  token = "your-token",
  secret = "your-secret"
)
```

4. Deploy the application:

```
rsconnect::deployApp(
  appName = "data-science-toolkit",
  appTitle = "Data Science Toolkit"
)
```

Project Structure

- `app.R`: Main application file integrating all modules
- `global.R`: Global variables and functions
- `www/`: Static assets
    - `custom.css`: Custom styling
- `modules/`: Shiny modules
    - `data_loading.R`: Module for loading datasets
    - `data_cleaning.R`: Module for data cleaning and preprocessing
    - `feature_eng.R`: Module for feature engineering
    - `eda.R`: Module for exploratory data analysis
    - `help.R`: Module for help documentation
- `data/`: Built-in datasets (automatically created at runtime)

Features

Data Loading

- Support for multiple file formats (CSV, Excel, JSON, RDS)
- Built-in sample datasets
- Data preview and summary statistics

Data Cleaning

- Variable selection
- Duplicate row handling
- Missing value treatment (removal, imputation)
- Data transformations (scaling, encoding, outlier handling)

Feature Engineering

- Creation of new variables using arithmetic operations

- Polynomial feature generation
- Interaction terms
- Binning for continuous variables
- Date component extraction
- Text feature extraction

Exploratory Data Analysis

- Summary statistics
- Distribution analysis (histograms, density plots, box plots)
- Correlation analysis
- Scatter plots with trend lines
- Multivariate analysis
- Custom visualizations
- Dynamic data filtering

Contributions

Alex Friedman – Led the whole project and worked on the main Shiny framework and app structure.
Liu Yang – Focused on data handling, cleaning, and transformations.
Anxin Yi – Designed the interactive visualizations and UI elements.
Thomas Bordino – Added feature engineering tools and some machine learning components.

License