STAT5243 Project 2

Team 8

2025-03-14

Introduction

The Interactive Data Analysis App ("IDA") is a one-stop solution designed to make your data workflow seamless and intuitive. Whether you're a beginner or an experienced analyst, this app equips you with everything you need to load, preprocess, analyze, and visualize datasets—all in one interactive platform.

IDA incorporates an intuitive workflow that allows for:

- Data Cleaning: Easily handle missing values, remove duplicates, and preprocess data with automated tools.
- Feature Engineering: Apply various transformations, including PCA, feature selection, and custom feature creation.
- Data Visualization: Create interactive graphs, dynamically select variables and plot types, and visualize numerical, categorical, and mixed data.
- Statistical Tests for Model Assumptions: Conduct normality, independence, and stationarity tests to validate analytical models.
- Time Series Analysis: Explore patterns over time using ACF, PACF, and rolling mean visualizations.

STEP 1 - Load your dataset

Load your dataset on the landing page of our app. It was designed to cater to multiple file formats, of which are CSV, XLSX, JSON and RDS files. You can also choose one of our two built-in datasets to get acquainted with the features of the IDA.

STEP 2 - Clean your dataset and pre-process variables of interest

The Data Cleaning and Preprocessing modules will allow you to initially clean, transform, and enrich your raw data using the following functions:

- Missing Value Strategy: to correct for missing and duplicated values.
- Columns to Scale: to scale/standardize the values of a numeric variable, allowing for better statistical testing down the line.
- Categorical Columns to Encode: to transform the categorical variables of your choosing into factor levels or dummy variables, depending on your use case.
- Outlier Handling Strategy: winsorize or impute outliers in chosen variables to allow for more precise analysis down the line.
- Select Data Variables: to ensure time variables keep their date format for future time series analysis on the app.

STEP 3 - Conduct Statistical Analysis

3.a. Feature Engineering

Once your raw data ready, you can explore a full suite of functionalities on IDA, starting with feature engineering.

On one hand, our app allows for multiple forms of feature selection, whether it be through Principal Component Analysis, LASSO regression, Elastic Net, or Backward Stepwise Regression, among other options.

On the other hand, you can also generate new features from both numeric panel data and time series data. This includes algebraic transformations like variable multiplication and logarithmic transformations, as well as time series adjustments such as differencing and rolling means. These tools make it easy to experiment with feature engineering and immediately see the results.

3.b. Exploratory Data Analysis - Visualization

The EDA section provides several ways to explore data visually. IDA enables you to generate insightful visual representations of variable distributions through:

- Univariate Analysis: Displaying histograms, boxplots, and other visual tools to understand individual variable distributions.
- Bivariate Analysis: Utilizing correlation heatmaps, line plots, and scatter plots to explore relationships between two variables.
- Time Series Analysis: Examining temporal patterns using time distribution plots, autocorrelation function (ACF), and partial autocorrelation function (PACF) plots to detect potential cyclical trends.

These interactive visualizations make data exploration more intuitive and will allow you to uncover key insights efficiently.

3.c. Exploratory Data Analysis - Statistical Tests

The final section of IDA helps validate potential relationships between variables and provides initial diagnostics for commonly used model assumptions. The app includes:

- Shapiro-Wilk Test – Evaluates whether a dataset follows a normal distribution.
- Pearson's Chi-Squared Test – Tests independence between two categorical variables.
- Spearman Correlation Coefficient – Assesses non-linear relationships between two numeric variables beyond the assumption of linearity.
- Wilcoxon Rank-Sum Test – Compares distributions of two independent groups within a categorical variable concerning a numeric outcome.
- Kruskal-Wallis Test – Extends the Wilcoxon test to compare distributions across three or more groups.
- Augmented Dickey-Fuller Test – Determines whether a time series dataset exhibits stationarity.

These statistical tools will allow you to explore and validate patterns in their data before proceeding to more advanced modeling.