EDA in Python

Armand Junior DONGMO NOTUE

College of Engineering and Technology, Grand Canyon University

DSC-510-O500: Advanced Probability and Statistic

Edward Ofori

04/10/2024

Task1:

Exploratory data analysis (EDA) is a crucial step in understanding the structure, relationships and patterns of a data set. EDA encompasses crucial data pre-processing tasks such as data cleaning, validation, transformation and integration, followed by statistical testing (or summarization), visualization and analysis.

Here is how to approach EDA professionally:

- Define objectives: understand the purpose of the analysis and the questions to be answered. Set clear objectives to guide exploration.
- Data collection: assess the relevance of the data set and ensure its integrity. Check that data is complete, accurate and correctly formatted.
- Initial inspection: Examine the data to understand its structure and content. Use methods such as head(), info() in Python to get an overview of the data set.
- Handle Missing Values: identify missing values and decide on an appropriate strategy for dealing with them. This may involve imputation, deletion or other methods based on the nature of the data and the objectives of the analysis.
- Exploring distributions: examine the distributions of numerical variables using histograms, whisker boxes and density diagrams.
- Analyze relationships: explore relationships between variables using scatter plots, pairwise plots and correlation analysis. Look for patterns, trends and dependencies that can inform further analysis.
- Feature engineering: create new features or transform existing ones to better capture relationships in data. This may involve clustering, coding categorical variables or creating interaction terms.
- Visualizations: use a variety of visualizations, such as bar charts, heat maps and violin charts, to effectively present results and information. Choose visualizations that best represent the data and highlight key patterns.
- Statistical tests: perform statistical tests to validate hypotheses or explore significant differences between groups. This may include t-tests, ANOVAs or chi-square tests, depending on the nature of the data and research questions.
- Iterative process: EDA is often an iterative process. Explore different angles, refine analyses and validate results to guarantee robustness and reliability.
- Document results: document your analysis, ideas and decisions throughout the EDA process. Clear documentation facilitates collaboration, reproducibility and communication of results.
- Communicate results: present your results in a clear, concise and visually appealing way. Tailor the presentation to your audience, focusing on key information and concrete recommendations.

In conclusion, it is essential to perform exploratory data analysis in a professional manner in order to gain a comprehensive understanding of the data and extract meaningful information from it. By following established best practices and using a variety of analytical tools and techniques, analysts can effectively explore the structure, patterns and relationships within a data set. EDA is a crucial step in the data analysis process, enabling analysts to identify potential sources of variation, outliers and data quality issues, as well as to formulate hypotheses for further investigation.

Task6:

Title: Exploratory Analysis of Wine Quality Dataset

Introduction:

The aim of this report is to conduct an exploratory analysis of the "Wine Quality" dataset to gain insights into the factors influencing wine quality ratings. The analysis encompasses data visualization, summary statistics, correlation analysis, and regression modeling.

Data Overview:

The dataset was downloaded from the UCI Machine Learning Repository.

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. The reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are munch more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

Two datasets were combined and few values were randomly removed.

1. Data Visualization:

Utilized histograms, box plots, and scatter plots to visualize the distributions of numerical variables and identify potential outliers.

Created bar plots to explore the distribution of categorical variables, such as wine type (red or white).

2. Summary Statistics:

Calculated summary statistics, including mean, median, standard deviation, minimum, and maximum values, for each numerical variable to understand their central tendency and variability.

3. Correlation Analysis:

Computed the correlation matrix to assess the relationships between numerical variables and their impact on wine quality ratings.

Utilized heatmaps and pair plots to visualize the correlations between variables.

4. Regression Modeling:

Selected relevant predictor variables (chemical properties) and the target variable (wine quality rating).

Split the dataset into training and testing sets.

Fitted a linear regression model to predict wine quality based on selected predictor variables.

Evaluated the model's performance using mean squared error (MSE) and validated assumptions of normality, linearity, and homoscedasticity.

Findings:

- The dataset contains a diverse range of chemical properties that may influence wine quality ratings.
- Some variables, such as alcohol content and volatile acidity, show moderate to strong correlations with wine quality.
- The linear regression model demonstrates good performance in predicting wine quality, with a low mean squared error.

- Assumptions of normality, linearity, and homoscedasticity are validated, enhancing the reliability of the regression model's predictions.

Conclusion:

The exploratory analysis of the "Wine Quality" dataset provides valuable insights into the factors affecting wine quality ratings. The findings can be leveraged by stakeholders in the wine industry to optimize production processes and enhance product quality. Further analysis and modeling techniques may be explored to deepen understanding and improve predictive accuracy.

GitHub: https://github.com/ARMAND-cod-eng/EDA-in-Python/blob/main/EDA%20in%20Python.ipynb

DataSet: https://www.kaggle.com/datasets/rajyellow46/wine-quality

References

IBM Technology(2022, September 23). Exploratory Data Analysis. [Video file]. Retrieved from https://www.youtube.com/watch?v=QiqZliDXCCg

Nabriya, P. (2024, April 5). What is Exploratory Data Analysis (EDA) and How Does it Work? Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/08/exploratory-data-analysis-and-visualization-techniques-in-data

Rogel-Salazar, J. (2023). Statistics and data visualisation with python (1st ed.). CRC Press. ISBN-13: 9781003160359

Roy, T. (2023, September 11). Exploratory Data Analysis: all you need to know - Trideep Roy - Medium. Medium. https://medium.com/@chotturoy54/exploratory-data-analysis-all-you-need-to-know-db13bbf449ef