DS203: Exercise 6.1

Nirav Bhattad

Last updated November 7, 2024

Report 7

Introduction

This section compares the approaches taken in two distinct reports on the DS203 Assignment - E6. The reports, written by Me and Aditya Anand Gupta, focus on data analysis, handling missing values, outlier treatment, and dimensionality reduction for three main problems. Each report provides valuable insights and employs slightly different methodologies. This comparative analysis highlights differences and evaluates the efficacy of each approach.

Problem 1: Handling Outliers, Missing, and Incorrect Data

Overview

Problem 1 focuses on exploring, identifying, and improving data quality by addressing missing values and outliers within the dataset.

A. Exploratory Data Analysis (EDA)

Both authors performed an initial data inspection, but there are some key differences:

- **My Report:** The EDA includes calculations of mean, standard deviation, and a visual line plot to detect instability. The report uses `plotly` for interactive plots, emphasizing early visualization of data patterns.
- **Aditya Anand Gupta's Report:** Aditya provides detailed column statistics with a summary of missing data and identifies that the `Timestamp` column requires reformatting. The EDA also emphasizes the skewness of data, with many values close to zero.

B. Identification of Unstable Period

- **My Report:** Analyzes a two-week period from July 30 to August 14, 2019, with line plots to illustrate fluctuations. The period is selected based on visual inspection of wild fluctuations.
- **Aditya Anand Gupta's Report:** Chooses a different period, from May 15 to May 29, 2019, identified through daily averaged plots. This choice is justified with more comprehensive fluctuation analysis.

C. Outlier Handling and Data Imputation

Each report applies distinct methods to handle missing values and outliers:

- **My Report:** Implements multiple methods, including median imputation, trimming, capping, RANSAC regression, and Loess smoothing, to treat outliers and smooth data. The choice of median imputation is well-justified based on its robustness.

- **Aditya Anand Gupta's Report:** Aditya uses Z-score filtering for outlier removal, Winsorization, and moving average smoothing. Imputation is performed using the mean, and Winsorization caps extreme values. Aditya's approach emphasizes consistency and the preservation of data distribution.

D. Global Trend Analysis for Local Adjustments

Both reports suggest leveraging global trends:

- **My Report:** Mentions regression-based imputation to address unstable regions.
- **Aditya Anand Gupta's Report:** Expands on regression imputation, implementing a time series decomposition approach to capture trends and seasonality.

Problem 2: Column Processing and Dimensionality Reduction

A. Column Processing and Reduction

Both reports discuss criteria for retaining columns, but the methods differ:

- **By Report:** Retains columns based on variance and correlation thresholds, selecting 86 columns.
- **Aditya Anand Gupta's Report:** Employs a low variance threshold to discard irrelevant columns. Aditya's report also identifies specific redundant columns and provides a list of columns dropped.

B. Outlier Handling, Normalization, and Standardization

- **My Report:** Uses the Interquartile Range (IQR) method for outliers, filling NaN values with column means and standardizing data.
- **Aditya Anand Gupta's Report:** Aditya uses similar methods, noting normalization with `StandardScaler`. This subsection is concise, focusing on bringing the dataset into a compatible format for further analysis.

C. Correlation Analysis and VIF

- **My Report:** Applies correlation and VIF analysis, setting thresholds of 0.8 for correlation and 10 for VIF.
- **Aditya Anand Gupta's Report:** Identifies 32 columns with high VIF values and provides specific column names. This subsection is more granular, reflecting detailed analysis of multicollinearity.

D. Principal Component Analysis (PCA)

Both authors implement PCA but use different interpretations:

- **My Report:** Identifies an elbow point around 10 to 15 components.
- **Aditya Anand Gupta's Report:** Concludes an elbow point at around 8 to 10 components post-VIF analysis, suggesting dimensionality reduction to 54 columns.

Problem 3: PCA and t-SNE on MNIST and E6 Datasets

A. PCA Analysis on MNIST Dataset

- **My Report:** Conducts PCA and provides an elbow plot, identifying 100 components as optimal.
- **Aditya Anand Gupta's Report:** Aditya provides further insight into clustering behavior of different digit groups in the PCA-reduced space.

B. t-SNE Analysis on MNIST and E6 Datasets

Both authors use t-SNE to visualize data structure in two dimensions:

- **My Report:** Provides a clear visualization of digit clustering but doesn't elaborate extensively on overlapping regions.
- **Aditya Anand Gupta's Report:** Adds interpretation of overlapping clusters in MNIST, suggesting similarities among certain digits.

C. t-SNE on E6 Dataset

- **My Report:** Offers basic observations on t-SNE clusters in E6 without extensive commentary on structure.
- **Aditya Anand Gupta's Report:** Observes continuous structures and arch-like formations, indicating sequential relationships within the data.

Conclusion

In summary, both reports successfully tackle the assignment's requirements with slightly different focuses. My Report emphasizes multiple data treatment methods and interactive visualizations, whereas Aditya Anand Gupta provides a systematic, step-by-step approach with a focus on statistical rigor. Together, these reports illustrate the application of comprehensive data analysis techniques in solving real-world data challenges.

§1. Report 139

Introduction

This report compares two analyses of datasets in the context of data exploration, cleaning, outlier handling, and dimensionality reduction. The primary document, *DS203: Exercise 6*, will be compared with the alternative document, *139.pdf*, to highlight methodological, analytical, and interpretative differences. This comparison is broken down by each major problem subsection* and sub-task, with additional details provided to further enhance understanding.

Problem 1: Outlier Handling and Data Imputation

Exploratory Data Analysis (EDA)

Summary Statistics

- **My Report:** The analysis begins with a thorough EDA, calculating key statistics like mean, median, standard deviation, and data range. A nine-month period is covered with frequency information and outlier visualization using Plotly.
- **139 Report:** Provides fewer statistical details, mentioning only a discrepancy between count and shape due to NA and null values. After cleaning, it visualizes the data with scatter, box plots, and histograms. Specific observations are made, such as a high concentration of values around 0.5 and outliers above an upper fence of 71.33k.

Key Difference

My report provides a more statistically rich description, while 139.pdf focuses on a specific outlier threshold. Additionally, DS203 discusses frequency (1 record every 5 minutes), which is absent in 139.pdf.

Unstable Period Identification

- **My Report:** Identifies the unstable period from `2019-07-30` to `2019-08-14`, based on observed fluctuations and gaps.

- **139 Report:** Selects a different period, `2019-07-16` to `2019-07-30`, using a 20k-minute duration. Fluctuations and zeros in this interval are highlighted as reasons.

Key Difference

My report provides a longer period for analysis and appears to base the selection on visual observation of instability, while 139 uses an approximate minute duration to define the period.

Outlier Handling Techniques

Methods Used

- **My Report:** Applies five techniques: median imputation, trimming, capping, RANSAC regression, and Loess smoothing.
- **139 Report:** Uses three methods: mean imputation, capping at 71.33k, and trimming values at 0 and above 71.33k.

Key Difference

My report implements a wider array of techniques, including advanced regression and smoothing. In contrast, 139 applies simpler imputation and capping without exploring non-parametric or robust regression techniques.

Global Trend Information Usage

- **My Report:** Proposes a seasonal decomposition approach to capture global trends and guide corrections for local instabilities.
- **139 Report:** Recommends calculating a moving average across the dataset to replace missing or volatile values in the identified period, without seasonal analysis.

Key Difference

My report leverages seasonal decomposition, which provides a more structured approach than a simple moving average, potentially yielding insights into cyclic trends in the dataset.

Problem 2: Data Cleaning and Feature Selection

Handling Null and Low Variance Columns

- **My Report:** Columns with a variance below 0.05 are removed, non-numeric values are replaced, and missing values are imputed with column means.
- **139 Report:** Drops specific columns (56, 58 for nulls; 2, 82 for low unique values). Non-numeric columns are dropped with a less systematic approach.

Key Difference

My report follows a threshold-based approach, providing a quantitative criterion for column removal. The 139 report uses individual column selection based on specific null and unique value observations.

Outlier Detection and Removal

- **My Report:** Detects outliers using the Interquartile Range (IQR) method and standardizes columns before PCA.
- **139 Report:** Applies a 10x IQR threshold for outlier removal, assuming minimal data context and caution in outlier removal. High-skew columns are also normalized.

Key Difference

My report employs a stricter, statistical-based IQR outlier detection, while 139 uses an exaggerated IQR threshold to minimize impact on data integrity, reflecting a more cautious approach.

Correlation and VIF Analysis

- **My Report:** Drops columns with correlations above 0.8 and calculates VIF, dropping columns with VIFs over 10.
- **139 Report:** Drops columns with correlations from 0.6 to 1 and removes columns with VIF exceeding 100.

Key Difference

My Report employs a more conservative correlation threshold (0.8), potentially retaining more features than 139, which uses a lower correlation threshold (0.6) and a very high VIF threshold (100), which may allow some multicollinearity to persist.

PCA Analysis

- **My Report:** Conducts PCA before and after removing correlated features, determining the optimal component count using elbow plots.
- **139 Report:** Uses the `kneed` library to find an elbow at 5 components without progressive PCA stages.

Key Difference

DS203's iterative PCA analysis provides more insight into dimensionality, allowing a more dynamic component count adjustment, while 139 selects components based on a single elbow plot.

Problem 3: Dimensionality Reduction with PCA and t-SNE

PCA Analysis on MNIST Dataset

- **My Report:** Provides an elbow diagram, cumulative variance analysis, and a scatter plot (PC2 vs. PC1) to evaluate variance retention.
- **139 Report:** Selects 40 principal components based on the `kneed` library, noting that PC1 holds more information. Two PCs only capture 10% variance, while 40 PCs capture 55%.

Key Difference

DS203 emphasizes component interpretability and optimal variance capture with fewer components, whereas 139 chooses more components (40) for variance capture, potentially prioritizing information retention over dimensionality reduction.

t-SNE Analysis

- **My Report:** Applies t-SNE to both MNIST and E6 datasets, visualizing the results with scatter plots and assessing separability.
- **139 Report:** Evaluates clustering quality using K-means with t-SNE, calculating Silhouette Score and Davies-Bouldin Index, and applies t-SNE to both datasets.

Key Difference

The DS203 report lacks quantitative clustering metrics (Silhouette and Davies-Bouldin), which the 139 report includes to assess cluster quality. This adds an extra level of analysis to 139, particularly useful for evaluating cluster distinctiveness.

Conclusion

The DS203 report generally offers a more rigorous approach to statistical analysis, particularly with outlier handling, PCA, and correlation analysis. The 139 report provides valuable clustering metrics during t-SNE and a simplified outlier handling approach, which can be beneficial in cases requiring conservative assumptions. These differences reflect two contrasting data handling and dimensionality reduction strategies, each with its own merits and potential applications.

§2. Report 128

Introduction

This report compares two data analysis and preprocessing reports, each detailing the analysis of the HT R Phase Current dataset and another dataset referred to as E6 Run. Both reports apply data cleaning techniques, outlier handling, and dimensionality reduction but differ in terms of approach, level of detail, and methodology. In this comparison, we highlight the key differences and provide an explanation for each.

Initial Data Loading and Structure

My Report

The first report mentions the data loading process with a detailed description of converting the `Timestamp` column into a datetime format and indexing it for time series analysis.

128 Report

While the second report also mentions loading and indexing the data based on `Timestamp`, it provides less technical detail compared to the first report.

Explanation

The first report provides more granular detail on how the data was structured, which would be beneficial for reproducibility. The second report keeps it high-level and focuses on summarizing the steps.

Descriptive Statistics

My Report

The first report computes a wide range of statistical measures, including the mean, median, mode, standard deviation, and quantiles.

128 Report

The second report only provides basic statistics such as the mean, maximum, and minimum HT R Phase Current values.

Explanation

The first report gives a more comprehensive statistical summary, making it more useful for understanding the distribution of the data. The second report provides only the essential statistics.

Outlier Detection

My Report

The first report applies Z-scores to detect outliers, with a threshold set at ± 3 for identifying and removing outliers.

128 Report

The second report also mentions using Z-scores but does not specify the threshold or how exactly the outliers were identified.

Explanation

The first report provides more details about the threshold for Z-scores, making the methodology clearer. The second report lacks the specificity needed for understanding how outliers were detected.

Smoothing Techniques

My Report

The first report applies multiple outlier handling techniques such as imputation, trimming, robust regression, and Loess smoothing to deal with fluctuations.

128 Report

In contrast, the second report only applies a Simple Moving Average (SMA) with a window size of 5.

Explanation

The first report's exploration of multiple outlier handling techniques makes it more thorough. The second report's reliance on a single method (SMA) simplifies the process but may miss other effective methods for smoothing.

Global Trend Adjustment

My Report

The first report details how global trend information was used to adjust the local 2-week fluctuation period.

128 Report

The second report also mentions global trend adjustment but does not elaborate on how this adjustment was calculated.

Explanation

The first report provides a more detailed explanation of how global trends were used to smooth local fluctuations, making it a more robust approach to handling anomalies.

Statistical Measures After Smoothing

My Report

The first report computes various statistical measures post-smoothing, including mean, variance, median, and mode.

128 Report

The second report only includes mean and variance after smoothing.

Explanation

The first report's inclusion of a broader set of statistics offers more insight into how the smoothing process affected the dataset.

PCA and Dimensionality Reduction

My Report

The first report covers Principal Component Analysis (PCA) in detail, including steps such as standardization, correlation analysis, and multicollinearity handling via VIF. It also explains the elbow method for determining the number of components to retain.

128 Report

The second report includes PCA but without VIF analysis or detailed discussion on multicollinearity. It also does not provide an elbow diagram for PCA interpretation.

Explanation

The first report offers a more comprehensive approach to dimensionality reduction by integrating VIF analysis and detailed PCA interpretation. The second report is more concise, focusing on PCA without the added details on multicollinearity.

Conclusion

In conclusion, while both reports tackle data preprocessing effectively, the first report provides a more detailed and comprehensive approach, especially in terms of outlier handling, smoothing techniques, and PCA analysis. The second report simplifies some of the processes but could benefit from additional details in specific areas such as outlier detection and dimensionality reduction.