Introduction:

In this report, I'll outline the process and findings from a comprehensive data analysis assignment. The assignment involved tackling various data quality issues, performing exploratory data analysis (EDA), and applying advanced techniques like Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction and visualization.

Problem 1: Dealing with outliers / missing / incorrect data

Process:

- Conducted EDA on the dataset (e5-htr-current.csv) to understand its nature and quality.
- Identified a 2-week period with unstable data.
- Implemented outlier removal, data smoothing, and missing data imputation techniques.
- Utilized global data trend information to guide local changes.

Learnings:

- Importance of thorough EDA in understanding data characteristics.
- Critical role of data preprocessing techniques in enhancing data quality.
- Significance of leveraging global trends to guide local data treatment.

Problem 2: Outliers, missing values, scaling / normalization, correlation analysis, VIF analysis, PCA analysis

Process:

- Conducted EDA on e5-Run2-June22-subset-100-cols.csv dataset.
- Processed columns to resolve outliers, missing values, and incorrect data.
- Standardized or normalized columns as needed.
- Identified correlated columns and performed VIF analysis.

- Applied PCA before and after preprocessing to reduce dimensionality.

Learnings:

- Need for a systematic approach to handle outliers, missing values, and correlation.
- Importance of dimensionality reduction techniques like PCA in handling high-dimensional data.
- Understanding multicollinearity and its impact on data analysis.

Problem 3: PCA and t-SNE

Process:

- Applied PCA and t-SNE on mnist_test.csv dataset.
- Visualized and interpreted results using scatter plots and elbow diagrams.
- Applied t-SNE on e5-Run2-June22-subset-100-cols.csv dataset for further analysis.

Learnings:

- Differences between PCA and t-SNE in dimensionality reduction and visualization.
- Interpretation of PCA components and t-SNE clusters.
- Importance of choosing appropriate dimensionality reduction techniques based on data characteristics.

Problem 4: Major Learnings

Key Insights:

- Data preprocessing is crucial for ensuring data quality and improving analysis outcomes.
- EDA provides valuable insights into data characteristics and helps in making informed decisions.
- Dimensionality reduction techniques like PCA and t-SNE are powerful tools for visualizing high-dimensional data and identifying patterns.
- Understanding correlation and multicollinearity is essential for accurate analysis and model building.

Conclusion:

The assignment provided valuable hands-on experience in data analysis techniques, ranging from data preprocessing to advanced dimensionality reduction methods. By systematically addressing data quality issues and leveraging exploratory and visualization techniques, I gained insights into the datasets and learned important principles applicable in real-world data analysis scenarios.

Reflection:

Reflecting on the assignment, I realized the importance of a structured approach to data analysis, encompassing data preprocessing, exploratory analysis, and advanced techniques for dimensionality reduction and visualization. Moving forward, I intend to apply these learnings in future data analysis projects to derive meaningful insights and make informed decisions.

Overall, the assignment was instrumental in enhancing my data analysis skills and deepening my understanding of various techniques and methodologies in data science.