**Water Consumption Analysis Report**

**Introduction**

This report summarizes the analysis of water consumption data, focusing on predicting current charges, identifying patterns through clustering, and conducting time series analysis. The objective is to explore the interrelationships among features while employing various statistical and machine learning techniques.

**Dataset Overview**

The dataset comprises **50,315** entries with **25 columns**, including features related to water consumption, charges, and account details. Key attributes include:

* **Consumption (HCF)**: The volume of water consumed.
* **Current Charges**: The billing amount associated with consumption.
* **Estimated**: Indicates whether a charge is estimated.
* **Cluster**: Derived from K-Means clustering to segment data.

**Methodology**

1. **Data Cleaning**:
   * Addressed missing values and formatted date fields.
   * Introduced new features such as Consumption\_per\_day to enhance analysis.
2. **Exploratory Data Analysis (EDA)**:
   * Visualized data distributions and relationships.
   * Generated a correlation matrix to understand feature interdependencies.
3. **Modeling Approaches**:
   * **Linear Regression**: Used to predict Current Charges based on Consumption (HCF).
   * **Decision Tree Classifier**: Classified estimated vs. non-estimated charges.
   * **K-Means Clustering**: Segmented the dataset into distinct clusters based on consumption patterns.
   * **Time Series Analysis**: Decomposed consumption data to analyze trends and seasonality.

**Key Findings**

**1. Correlation Analysis**

* Strong positive correlation (0.99) between **Consumption (HCF)** and **Water & Sewer Charges**, indicating that higher consumption leads to higher charges.
* **Current Charges** and **Consumption (HCF)** also have a high correlation (0.99).

**2. Time Series Decomposition**

* **Trend**: Indicates a general upward movement in water consumption.
* **Seasonal**: Displays fluctuations, suggesting cyclical consumption patterns, possibly related to seasonal changes.
* **Residual**: Shows random fluctuations around a zero mean, supporting the assumption of a stationary process post-decomposition.

**3. K-Means Clustering Results**

* Identified three clusters, with most entries falling into one or two clusters, while a few extreme values represent outliers.
* Clustering helps to differentiate between high and low consumption behavior, which could inform targeted interventions.

**4. Linear Regression Findings**

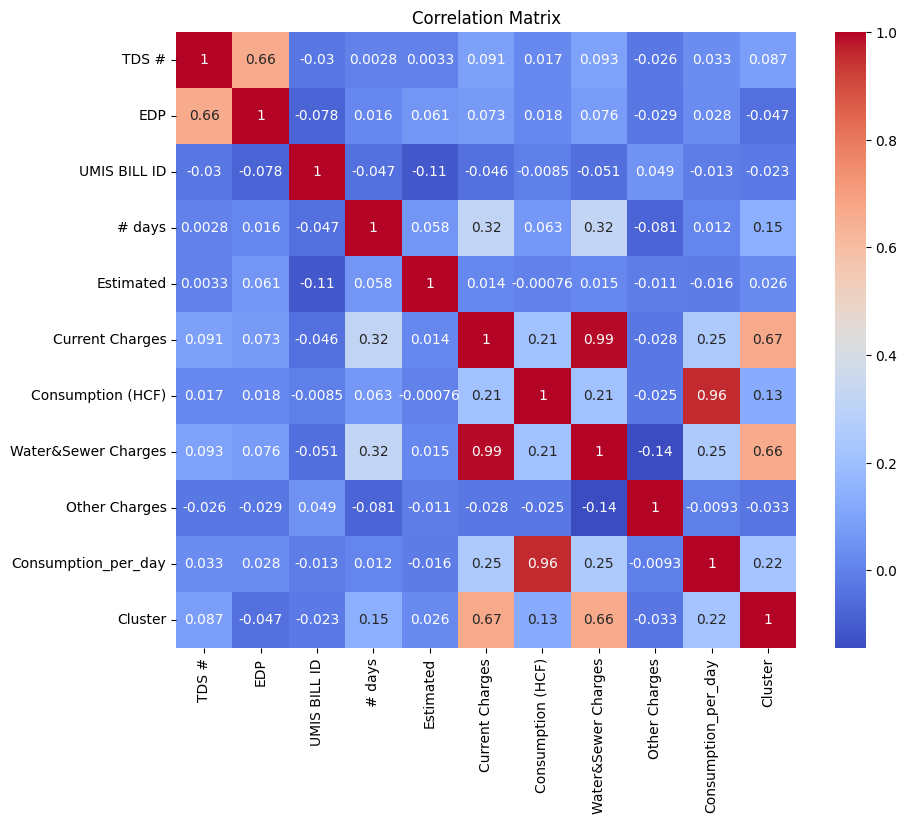
* The linear regression model predicted Current Charges with reasonable accuracy for lower consumption levels, but struggled with higher charges, indicating the need for more complex modeling approaches.
* **Model Metrics**:
  + MAE: 3,217
  + MSE: 37,815,205
  + RMSE: 6,149

**Visualizations**

The following visualizations were generated to support the analysis:

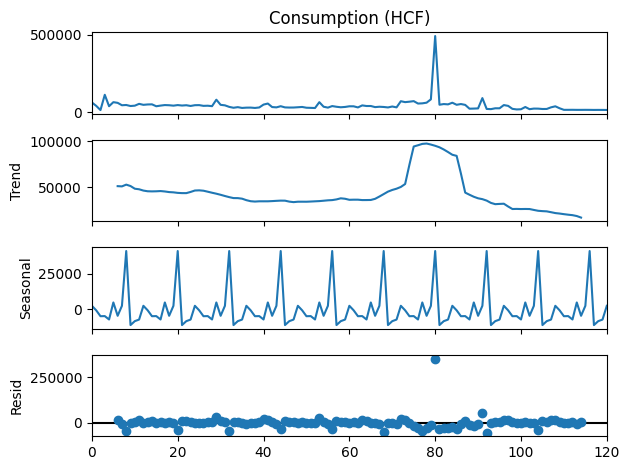
1. **Correlation Matrix**:

Displays relationships between variables clearly, identifying key interactions.



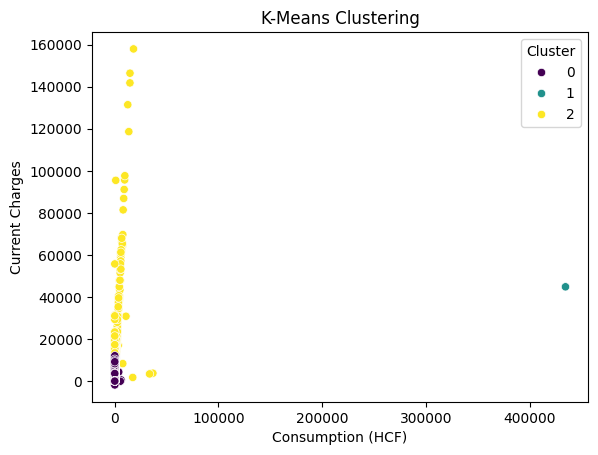
1. **Time Series Decomposition Plot**:

Illustrates the overall trend, seasonal patterns, and residual noise in the consumption data.



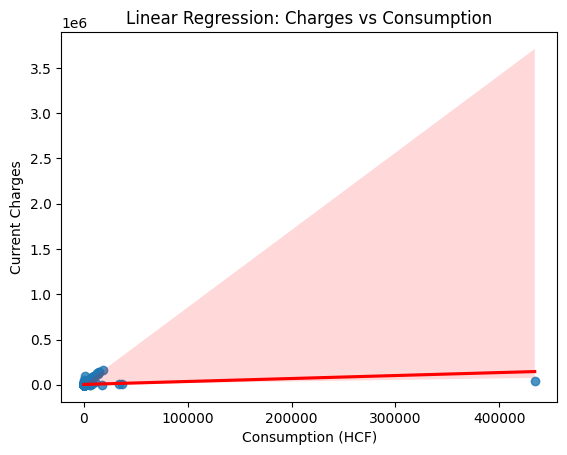
1. **K-Means Clustering Scatter Plot**:

Visualizes clusters of water consumption and current charges, highlighting distinct patterns and outliers.



1. **Linear Regression Plot**:

Indicates the fit of the regression model and presents the challenges in predicting high consumption charges.



**Recommendations**

1. **Model Enhancement**: Consider using more advanced regression techniques (e.g., Ridge, Lasso, and ensemble methods) to improve predictions, especially for high-charge scenarios.
2. **Feature Engineering**: Introduce additional features, such as demographic information or weather data, to enrich the dataset and improve model performance.
3. **Imbalance Handling**: Address class imbalance in the target variable through techniques like SMOTE to enhance decision tree performance.
4. **Further Analysis**: Conduct additional clustering analyses or segmentation strategies to gain insights into different consumer behaviors and preferences.

**Conclusion**

This report presents an extensive analysis of water consumption data utilizing predictive modeling, clustering, and time series techniques. The findings illustrate significant interconnectedness among consumption metrics and charges, providing a pathway for better understanding and forecasting of water usage patterns. Future work should focus on refining methods and exploring additional data sources for deeper insights.