Sub-Task 3: Summary of Key Findings and Data Augmentation Suggestions

Key Findings:

A. Churn Rate: Approximately 9.7% of the customers have churned, while the majority, about 90%, have not churned. This indicates an imbalanced dataset, which may require special handling in model development.
B. Consumption Data: The consumption data is significantly positively skewed with a long right tail. This means that a few customers have very high consumption, which might be an important feature in predicting churn.
C. Outliers: There are several columns with a significant number of outliers. These outliers could be addressed during feature engineering to improve model performance.
D. Correlation with Price Sensitivity: While the association with churn is generally modest, a correlation plot reveals a higher degree of correlation between churn and other price sensitivity variables. This suggests that price sensitivity and churn may have a complex, nonlinear relationship. This finding highlights the need to develop the price sensitivity feature differently to make it a more significant predictor of churn.

Data Augmentation Suggestions:

1. Competitor Pricing Data: Access to competitor pricing and market share data can provide valuable insights into the competitive landscape and how it influences customer churn.
2. Economic Indicators: Incorporate economic indicators for the regions where customers are located. Economic conditions can influence customer behavior and should be considered in the churn prediction model.
3. Customer Interaction Data: Data on customer interactions, such as customer support calls, emails, or website interactions, can provide insights into customer engagement and satisfaction.
4. Social Media Data: Monitor social media platforms for customer sentiment and feedback related to the utility industry. Social media data can capture public opinions and concerns that may impact churn.

These data augmentation suggestions will enrich the dataset and provide a more comprehensive view of the factors influencing customer churn. By incorporating these data sources, we can improve the accuracy and effectiveness of the churn prediction model.