**Water Quality Analysis Project Design Document**

**Phase 3: Development Part 1 - Data Preprocessing and Exploratory Data Analysis (EDA)**

**Data Preprocessing**

In this phase, we will focus on preparing the water quality dataset for analysis. Data preprocessing is a crucial step to ensure that the data is clean, structured, and ready for further exploration and modeling. Here's an outline of the data preprocessing tasks:

1. Handling Missing Data:

Identify and address missing values in the dataset.

Decide on an appropriate strategy for handling missing data, such as imputation or removal of rows/columns with missing values.

2. Handling Outliers:

Use the anomaly detection techniques introduced in Phase 2 to identify outliers.

Decide how to handle outliers, whether to keep, transform, or remove them based on their impact on the analysis.

3. Data Transformation:

Normalize or scale numerical features as necessary. Standardization (mean = 0, std = 1) is a common technique.

Encode categorical variables if present in the dataset using techniques like one-hot encoding or label encoding.

4. Feature Engineering:

Create new informative features if needed, based on domain knowledge or EDA insights.

Feature selection: Determine which features are relevant for the analysis.

5. Data Splitting:

Split the dataset into training and testing sets for model development and evaluation. A common split ratio is 80-20 or 70-30, depending on the dataset size.

**Exploratory Data Analysis (EDA)**

Once the data preprocessing is complete, we will proceed with Exploratory Data Analysis (EDA). EDA is a critical step for gaining insights into the dataset, identifying patterns, and understanding the relationships between variables. Here's how we plan to conduct EDA:

1. Summary Statistics:

Calculate summary statistics for numerical features, including mean, median, standard deviation, minimum, maximum, etc.

Summarize categorical variables by counting unique values.

2. Data Visualization:

Create visualizations to represent the data distribution and relationships:

* + Histograms and density plots for numerical features to understand their distributions.
  + Bar charts for categorical features to visualize their distributions.
  + Scatter plots to explore relationships between numerical features.
  + Correlation matrices or heatmaps to visualize correlations between variables.

3. Potability Analysis:

Analyze the distribution of potable and non-potable water samples.

Visualize the relationship between water quality parameters and potability.

4. Anomaly Visualization (If Applicable):

Visualize anomalies identified during data preprocessing using anomaly detection techniques. This will help us understand the nature of unusual patterns.

5. Hypothesis Testing (If Relevant):

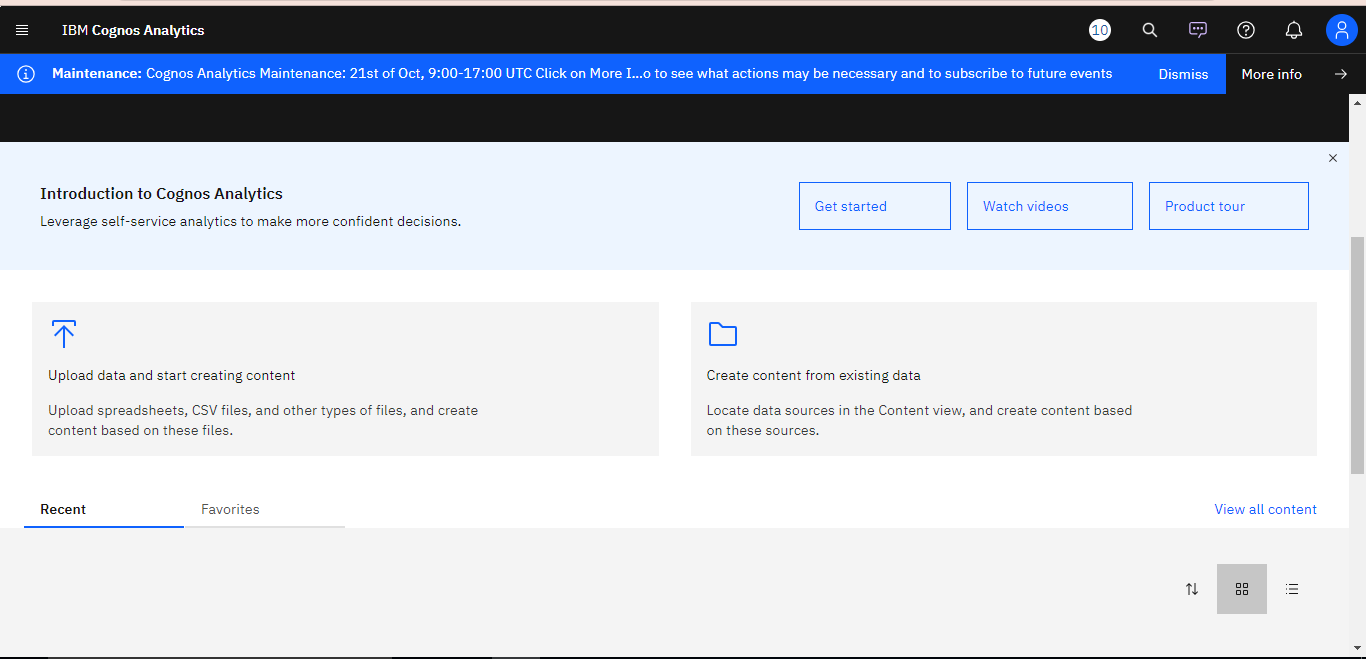
If specific hypotheses or questions arise during EDA, conduct statistical tests to validate or reject them.

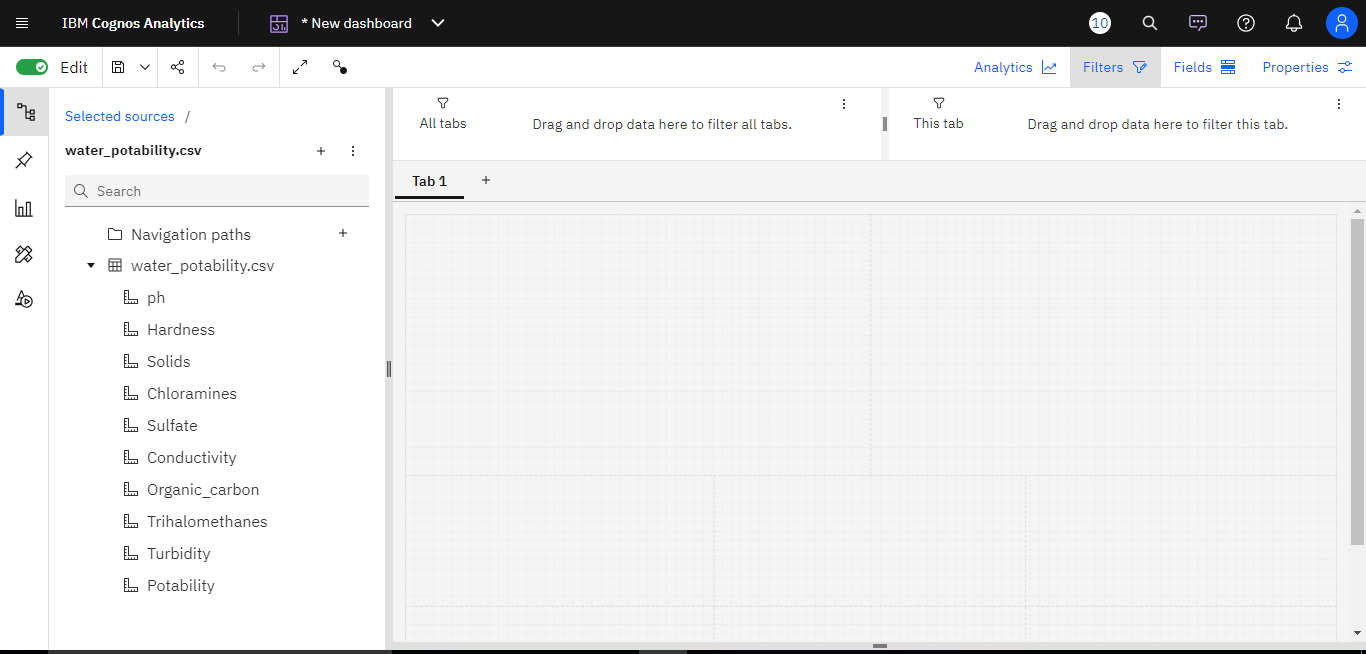
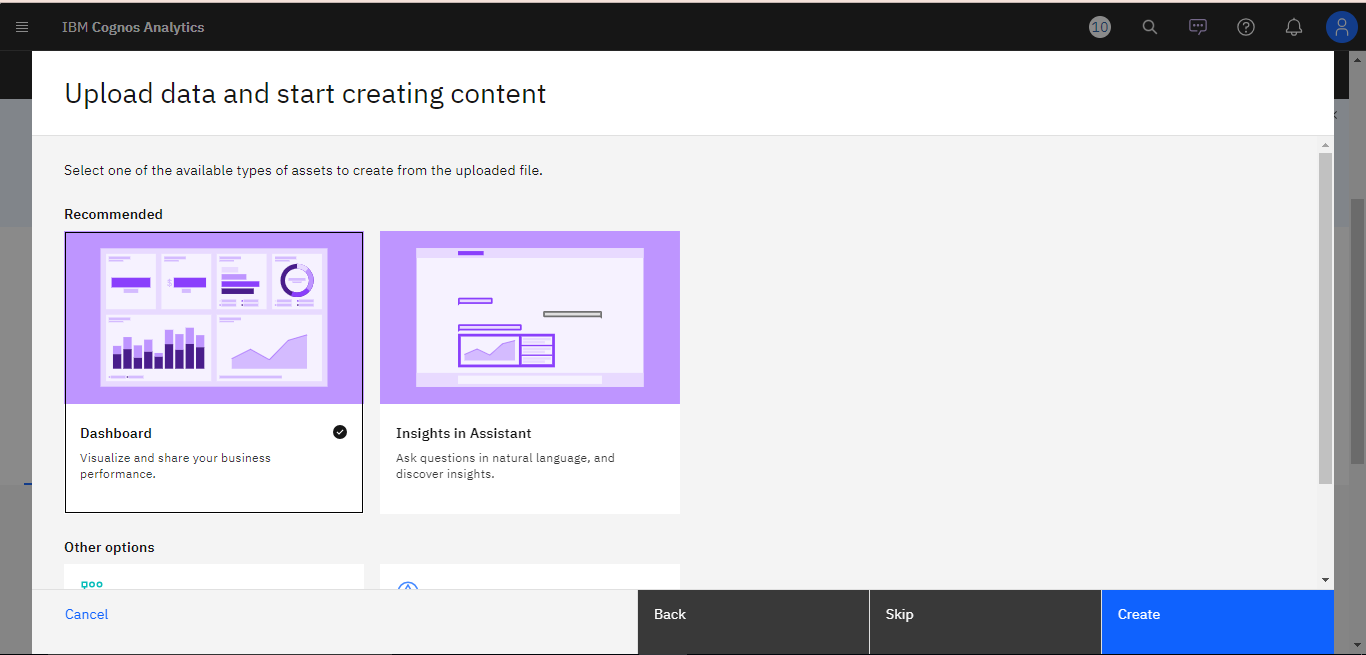
6. Insights and Documentation:

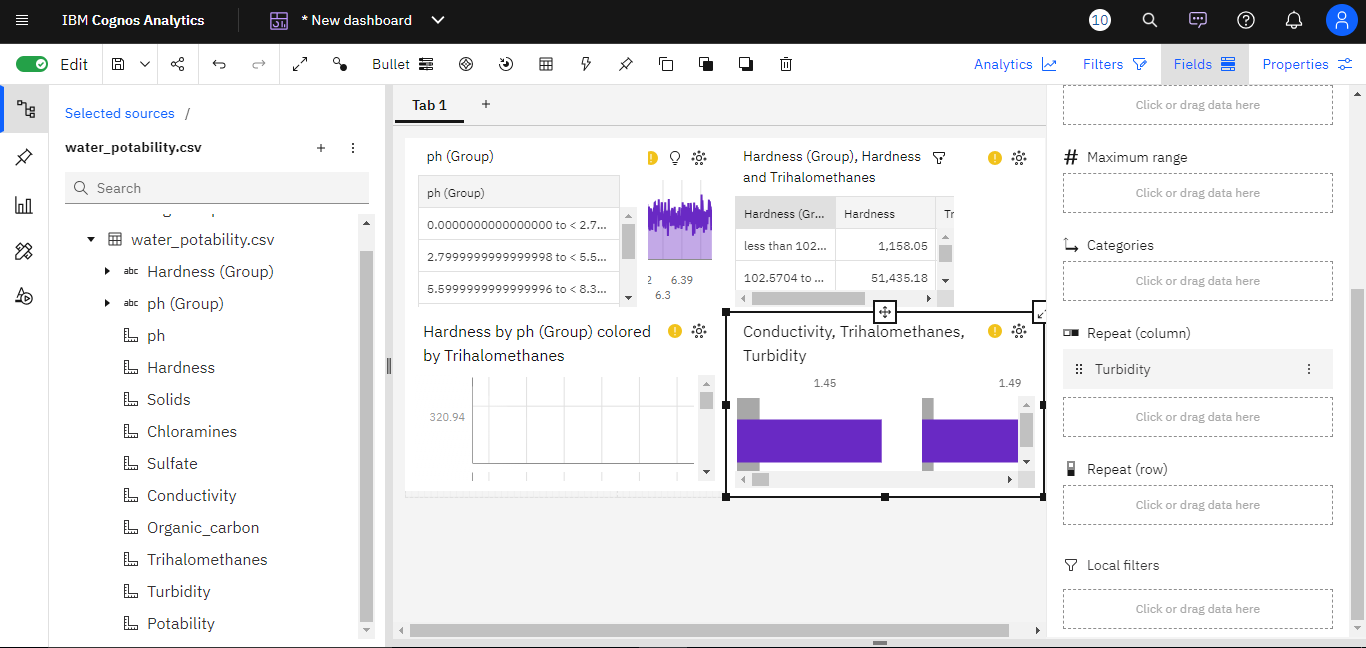
Document the key findings and insights from EDA. These insights will guide our subsequent modeling and analysis steps.

The output of this phase will be a well-preprocessed dataset and a clear understanding of the data's characteristics and relationships. These will serve as the foundation for building our water quality analysis model in the next phases.

**Account create in IBM Cognos**

****

**Data is uploaded**

**Creating Visual:**