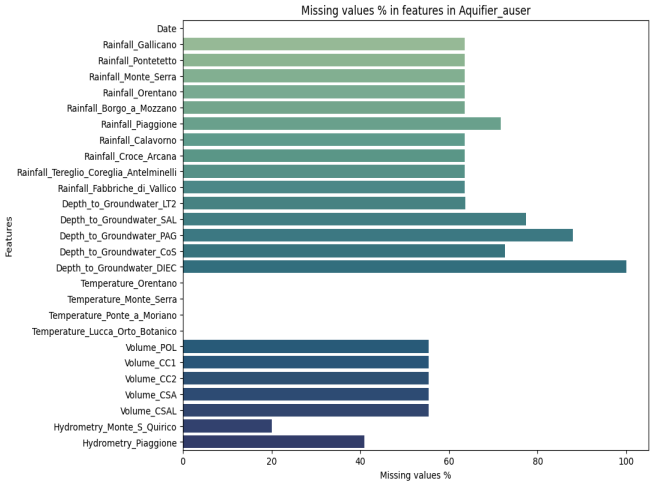
Our roadmap to the solution

EDA:

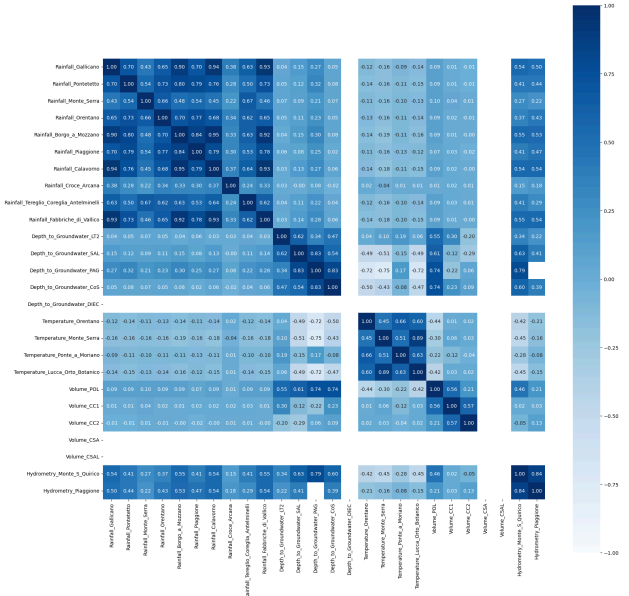
The following insights are derived when features from the dataset and checked as per their respective target feature:

- There is a large proportion of missing values in the dataset

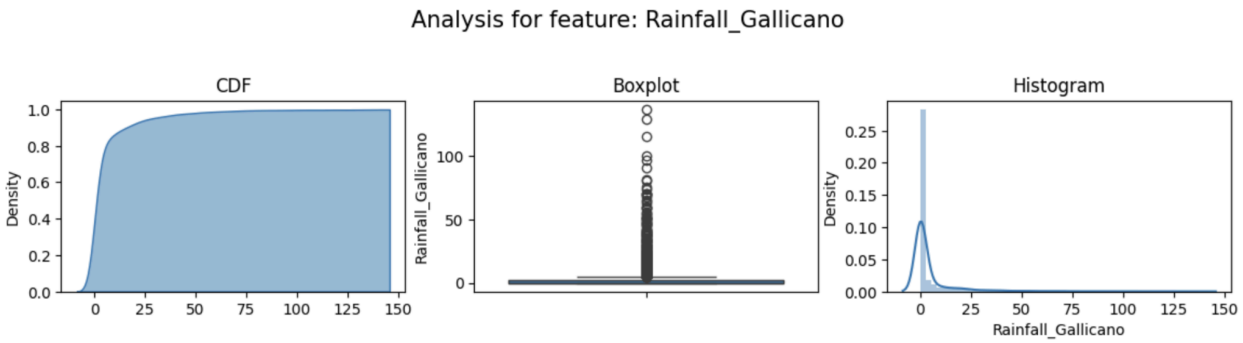
- graph of the percentage of **missing values** in each feature:



- Aquifer_auser **correlation** heat map:

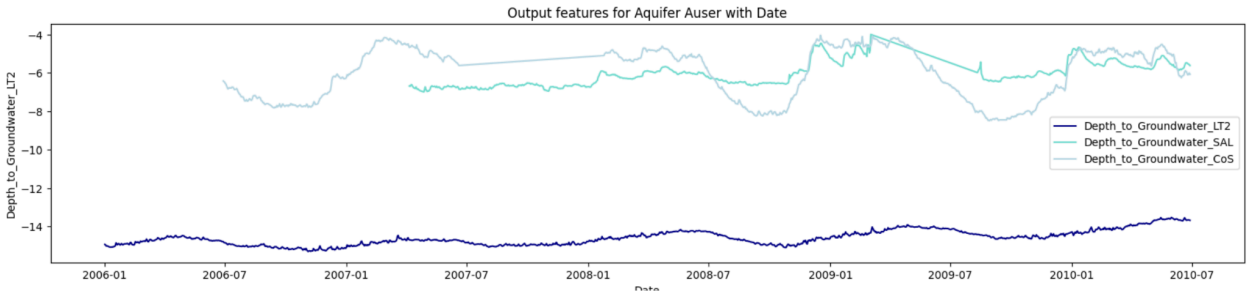


- **univariate analysis** for one feature in Aquifer_auser:



- There are a considerable number of outliers in rainfall features from 100mm to 300mm and few in the hydrometry, volume features in Aquifers.

- **yearly** and **monthly** observation:



Metrics Used:

- Median Absolute Error (MAE)
- Root Mean Square Log Error(RMSLE)
- R Squared (R^2)

Data Preprocessing:

1. **Handling Missing Values:**
   - Initially, consider only the rows where target features are not null.
   - Use KNN imputation to fill in missing values in independent features:
   - Plot the selected metrics (MAE, RMSLE, R2 score) for different k values.

- Choose the k value with the least error as determined by the metrics.

2. **Impute Missing Target Values:**
   - Use the complete dataset (with filled independent features) to predict missing values in the respective target features.
   - Consider baseline models such as Random Forest, Linear Regression, Decision Tree, or KNN for imputation.

KNN Imputation for Independent Features:

- **Subsetting Data**:
  - Consider only rows where the target features are not null.

- **Selecting k Value:**
  - Plot the metrics (MAE, RMSLE, R2 score) for different k values in KNN imputation.
  - Choose the k value that minimizes the error based on the selected metrics.

- **Applying KNN Imputation:**
  - Impute missing values in the independent features using the chosen k value.

Target Value Imputation Using Baseline Models:

- **Preparing Data:**
  - Use the complete dataset with filled independent features and target features.

- **Selecting Baseline Model:**
  - Choose a baseline model for imputing missing target values. Options include Random Forest, Linear Regression, Decision Tree, or KNN.

- **Training and Predicting:**
  - Train the selected baseline model on the dataset.
  - Use the trained model to predict missing values in the target features.

- **Impute Missing Values:**
  - Replace the missing target values with the predictions from the baseline model.

These steps aim to address missing values in both independent and target features using a combination of KNN imputation and baseline models. Adjust parameters and models based on the characteristics of your data and the performance of the chosen metrics.

Feature Engineering Steps:

1. **Averaging:**
   -Calculate the average for each feature based on different time durations: daily, weekly, monthly, and yearly.

2. **Latency Consideration:**
   - Recognize the possible latency in variables due to the effect of rainfall and temperature.

3. **Shifted Variables:**
   - Experiment with shifting variables by days, weeks, or months to capture lag effects.

4. **Model Training and Evaluation:**
   - Train models using the integrated and shifted datasets for each waterbody.
   - Evaluate models using Median Absolute Error (MedAE) to capture robust performance.

5. **Trend Analysis:**
   - Analyze trends in MedAE for different time durations and shifts.
   - Identify the optimal time duration and shift that minimizes MedAE for each waterbody.

6. **Final Dataset Preparation:**
   - Prepare the final dataset with actual values of target variables.
   - Include the effects of rainfall and temperature based on the selected time duration and shift.

7. **Iterative Refinement:**
   - Iteratively refine models and feature engineering based on the observed trends and performance.

8. **Documentation and Reporting:**
   - Document the chosen time durations, shifts, and any insights gained during the feature engineering process.
   - Report the final models, their configurations, and performance metrics.

9. **Cross-Validation:**
    - Implement cross-validation techniques to ensure the models' generalizability and robustness.

10. **Validation with Test Set:**
    - Validate the final models using a separate test set to ensure their performance on unseen data.

These steps aim to create a robust model, considering the effects of rainfall and temperature with appropriate latency considerations. The focus is on capturing and optimizing for the temporal patterns in the data.

In each of these new datasets, we shift each feature value by 1 to 31 days, 1 to 52 weeks, and 1 to 12 months. Then using a Random Forest Regressor, note the time duration after which the MAE is lowest to predict that shifted feature (i.e. the rainfall and temp effect is most appropriate).

Observing the best results at 0 days, we made the final aquifers dataset by shifting the values by 0 days backward.

The Model:

1. **Data Normalization:**
   - Normalize the datasets to remove negative values, ensuring consistent scaling for the subsequent modeling steps.

2. **Model Exploration:**
   - Experiment with various machine learning models, including KNN, Linear Regression, Random Forest, Support Vector Regression (SG regression), Decision Trees, and XGBoost.

3. **Error Metric Evaluation:**
   - Evaluate the performance of each model using appropriate error metrics to determine the one with the least errors and the best predictions. Consider metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or others suitable for the specific forecasting task.

Time series analysis

1. **Objective and Importance:**
   - The project aimed to predict water level depth over the next 3 years using historical data for crucial applications like water resource management and environmental monitoring.

2. **Methodology**:
   - Implemented the Prophet time series forecasting model for accurate predictions, considering seasonality components and potential holidays.
   - Conducted thorough data preparation, including feature selection and cleaning, to ensure the model's effectiveness.

3. **Results and Insights:**
   - Successfully predicted water level depth for the specified 3-year period, revealing insights into trends, seasonality patterns, and noteworthy changes.

4. **Conclusion and Future Considerations:**
   - Concluded with the model's overall success in capturing water level dynamics.
   - Identified areas for future improvements, such as incorporating additional data sources or refining model parameters.

References:

- https://www.kaggle.com/code/sanggusti/timeseries-analysis-acea-water-prediction
- https://medium.com/mlearning-ai/water-availability-prediction-using-historical-data-b9044bfddd76#2282