**WATER QUALITY ANALYSIS**

**PHASE V: PROJECT DOCUMENTATION**

**Abstract:**

**The quality of water is declining at a catastrophic rate and it is having a significant influence on the environment and health. Poor water quality will have wide ranging effect on human life causing various health hazards. Anticipating the quality of water will therefore be a great boon to the society.**

**Water abstraction, water extraction, or groundwater abstraction is the process of taking water from any source, either temporarily or permanently. Most water is used for irrigation or treatment to produce drinking water**

**Objective:**

**Water quality objectives are designed for the substances or conditions of concern in a watershed so that their attainment will protect the designated uses. Based on the preceding discussions, the water uses to be protected should include drinking water, irrigation, primary-contact recreation, aquatic life and wildlife.**

**The main purpose of environmental water pollution monitoring is to continuously monitor the water pollution of water areas affecting the human health and living environment.**

**The dataset features are dissolved oxygen, PH, conductivity, biological oxygen, nitrate, fecalcoliform, and total coli form. The features of the dataset are Dissolved Oxygen by which it indicates the level of oxygen dissolved in the water, which is essential for supporting aquatic life.**

**Data visualization:**

**Data visualization is an important tool for communicating science to a broader audience. Whether you are a volunteer community scientist or a professional aquatic ecologist, there are many free tools and low-cost programs that you can use to link the scientific data to actions that can improve water quality.**

**Parameters that are frequently sampled or monitored for water quality include temperature, dissolved oxygen, pH, conductivity, ORP, and turbidity. However water monitoring may also include measuring total algae, ISEs (ammonia, nitrate, chloride), or laboratory parameters such as BOD, titration, or TOC.**

**Predictive modeling for potability:**

Water quality is dictated by interactions among geomorphic processes, vegetation characteristics, weather patterns, and anthropogenic land uses over multiple spatio-temporal scales. In order to understand how changes in climate and land use impact river water quality, a suite of data with high temporal resolution over a long period is needed. Further, all of this data must be analyzed with respect to connectivity to the river, thus requiring high spatial resolution data. Here, we present how changes in climate and land use over the past 25 years have affected water quality in the 268 sq. km Hot River catchment in New Zealand. Hydro-climatic data included daily solar radiation, temperature, soil moisture, rainfall, drought indices, and runoff at 5-km resolution. Land cover changes were measured every 8 days at 30-m resolution by fusing Land set and MODIS satellite imagery. Water quality was assessed using 15-min turbidity (2011-2014) and monthly data for a suite of variables (1990-2014). Watershed connectivity was modeled using a corrected 15-m DEM and a high-resolution drainage network. Our analyses revealed that this catchment experiences cyclical droughts which, when combined with intense land uses such as livestock grazing and plantation forest harvesting, leaves many areas in the catchment disturbed (i.e. exposed soil) that are connected to the river through surface runoff. As a result, flow-normalized turbidity was elevated during droughts and remained relatively low during wet periods. For example, disturbed land area decreased from 9% to 4% over 2009-2013, which was a relatively wet period. During the extreme drought of 2013, disturbed area increased to 6% in less than a year due mainly to slow pasture recovery after heavy stocking rates. The relationships found in this study demonstrate that high spatiotemporal resolution land cover datasets are very important to understanding the interactions between landscape and climate, and how these interaction

Code:

*# Input data files are available in the read-only "../input/" directory*

*# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory*

import os

for dirname, \_, filenames **in** os.walk('/kaggle/input'):

for filename **in** filenames:

print(os.path.join(dirname, filename))

*# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save & Run All"*

*# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session*

*# Required Functions*

def show\_distributions(columns: list, data: pd.DataFrame, nrows: int = 1, ncols: int = 3):

*# This function creates distribution subplots.*

fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 5))

axes = axes.ravel()

for index, column **in** enumerate(columns):

sns.histplot(data[column], kde=True, ax=axes[index])

axes[index].set\_title(column)

*# Adjust layout*

plt.tight\_layout()

plt.show()

def show\_boxplots(columns: list, data: pd.DataFrame, nrows: int = 3, ncols: int = 3):

*# This function creates box plot subplots.*

fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(10, 15))

axes = axes.ravel()

for index, column **in** enumerate(columns):

axes[index].boxplot(data[column])

axes[index].set\_title(column)

plt.tight\_layout()

plt.show()

def normality\_test(columns: list, data: pd.DataFrame):

*# Conducts Shapiro-Wilk test.*

for i **in** columns:

results = shapiro(data[data[i].isna() == False][i])

print(i, results.statistic)

def random\_sample\_imputation(df):

*# Random Sample Imputatuin*

cols\_with\_missing\_values = df.columns[df.isna().any()].tolist()

for var **in** cols\_with\_missing\_values:

random\_sample\_df = df[var].dropna().sample(df[var].isnull().sum(),

random\_state=0)

random\_sample\_df.index = df[

df[var].isnull()].index

df.loc[df[var].isnull(), var] = random\_sample\_df

return df

Data set:

| ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic\_carbon | Trihalomethanes | Turbidity | Potability |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | NaN | 204.890455 | 20791.318981 | 7.300212 | 368.516441 | 564.308654 | 10.379783 | 86.990970 | 2.963135 | 0 |
| 1 | 3.716080 | 129.422921 | 18630.057858 | 6.635246 | NaN | 592.885359 | 15.180013 | 56.329076 | 4.500656 | 0 |
| 2 | 8.099124 | 224.236259 | 19909.541732 | 9.275884 | NaN | 418.606213 | 16.868637 | 66.420093 | 3.055934 | 0 |
| 3 | 8.316766 | 214.373394 | 22018.417441 | 8.059332 | 356.886136 | 363.266516 | 18.436524 | 100.341674 | 4.628771 | 0 |
| 4 | 9.092223 | 181.101509 | 17978.986339 | 6.546600 | 310.135738 | 398.410813 | 11.558279 | 31.997993 | 4.075075 | 0 |

**Conclusion:**

**Throughout this article, we have aimed to review the early stages of an EDA assessment. Metadata on the imported data was initially reviewed to display early insights. A deeper dive into the summary statistics allowed us to focus on the missing values. Finally, we were able to review the histogram of the pH variable to ensure that the variable followed external expectations.**