DEPI Capstone Project for Data Analytics Specialist

Under the supervision of the Eng: Arwa Essam

Project Parts:

- **Excel (“**Global Access to Safe Drinking Water Services Over Time**”)**
- **SQL (“**Movies and TV shows Analysis**”)**
- **Python (“**Football Players Analysis**”)**
- **Power Bi (“**Maji Ndongo Water Analysis**”)**

Excel Analysis Documentation

Project Title:

Global Access to Safe Drinking Water Services Over Time

1. Data Preparation

Steps Taken:

- **Data Cleaning:**
    - Handled missing values by flagging countries with incomplete data and filling gaps through interpolation where appropriate.
    - **Outliers** were visually inspected using conditional formatting and were kept or removed based on their relevance.
    - Removed any duplicate entries.

New Column:

- Created a new column called **diff**, calculated as the difference between the access percentages in **2000** and **2022**, to measure improvement in safe drinking water services over time.

---

2. Data Analysis

Key Insights & Steps:

- **Proportion of Population with Safe Drinking Water:**
    - Calculated access proportion using **SUMIF** and **COUNTIF** to analyze trends in different regions.
- **Analysis with diff:**
    - Countries were categorized into three groups:
        1. **High Improvement:** Countries with large positive differences between 2000 and 2022.

2. **No Change:** Countries with **diff = 0**, indicating stagnation.
3. **Negative Change:** Countries with decreasing access, showing negative values in the diff column.

- **Pivot Tables:**
  - Generated pivot tables to show these country categories and summarize data for each group.
  - Pivot tables also helped track specific regional trends.

---

3. Data Visualization

Charts Created:

- Visualized the findings through various charts (as shown in the image):
  - **Bar Charts** to show countries with high improvements, those with no change, and those with declining access.
  - **Line Charts** to track changes in access over time in select countries, displaying both positive and negative trends.
- These visuals were essential in tracking global disparities and highlighting the need for intervention in specific regions.

---

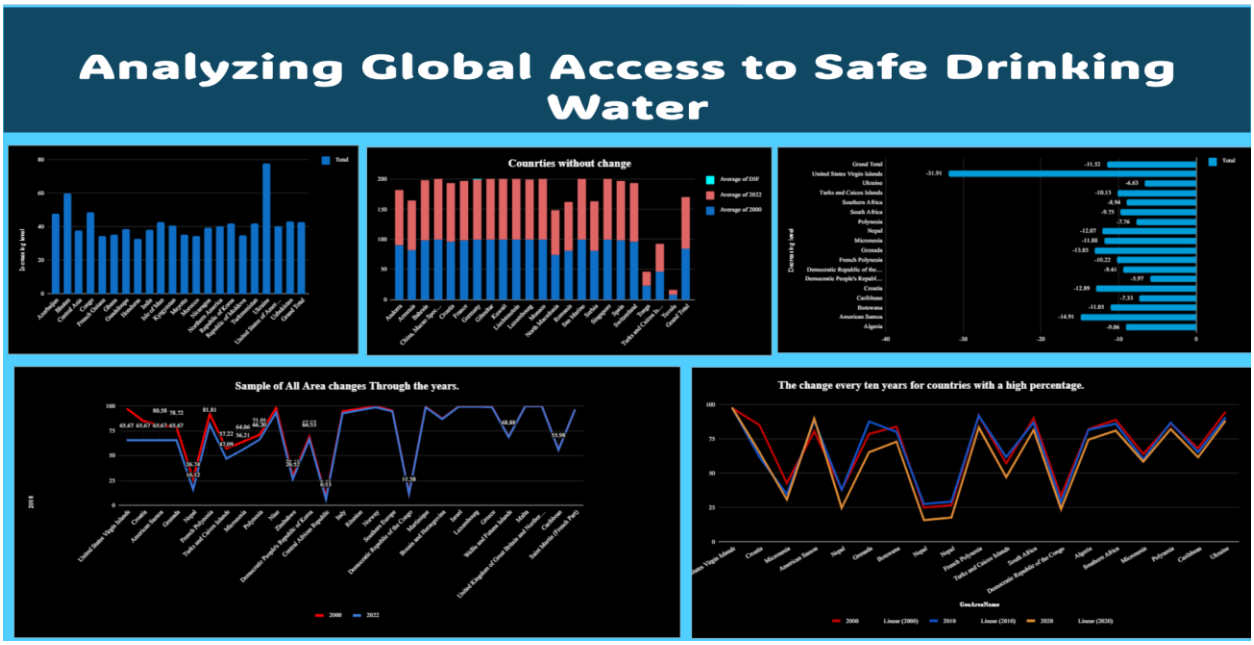
4. Summary Report

Global Trends:

- Positive improvements in access to safe drinking water were observed in most countries. However, certain regions such as **Sub-Saharan Africa** still struggle, with many countries showing little to no progress.

Country-Level Analysis:

- Some countries showed substantial improvement, while others like **United States Virgin Islands** experienced a negative trend in access to safe drinking water.



Recommendations:

- **Targeting Decreasing Countries:** Countries with negative diff values need urgent interventions.
- **Focus on Rural Areas:** Consistent underperformance in rural areas calls for targeted rural infrastructure projects.
- **Improved Data Collection:** Many countries had missing or inconsistent data, necessitating better global data tracking systems.

SQL Analysis Documentation - Movie Database

Project Overview:

This part of the project focuses on analyzing a movie database using SQL queries executed through Python's SQL libraries. The analysis aims to extract key insights from the dataset, such as movies not yet released, top-rated films, and movies released within specific periods.

---

1. Database Connection:

The first step involved establishing a connection to the database using SQLite in Python. This allowed for direct interaction with the dataset without using traditional SQL platforms like SQL Workbench.

Code:

python

```
%load_ext sql

%%sql sqlite:///TMDB.db
```

---

2. Data Exploration:

Initial exploration involved querying the movies table to inspect the first few records.

SQL Query:

SQL

```
%%sql

SELECT * FROM movies LIMIT 5;
```

The table includes various fields such as title, release_date, release_status, and more, which were examined to understand the structure of the dataset.

---

3. Creating Views:

To streamline the analysis, a view called not_released was created. This view filtered out movies that were not yet released, allowing for easy querying of unreleased films.

SQL Query:

sql

%%sql

CREATE VIEW not_released AS

SELECT * FROM movies WHERE release_date <> 'Released';

After creating the view, the following query fetched a sample of unreleased movies:

SQL Query:

sql

%%sql

SELECT title, release_date, release_status FROM not_released LIMIT 5;

---

4. Further Analysis:

Several additional queries were performed to extract meaningful insights from the movie dataset:

- **Querying Top-Rated Movies:** The highest-rated movies were selected and sorted based on their ratings.

SQL Query:

sql

%%sql SELECT title, rating FROM movies ORDER BY rating DESC LIMIT 10;

- **Analyzing Genre Popularity:** To understand genre popularity, the movies were grouped by genre, and the average ratings per genre were calculated.

SQL Query:

sql

%%sql

SELECT genre, AVG(rating) AS avg_rating FROM movies GROUP BY genre ORDER BY avg_rating DESC;

- **Filtering Movies Released After 2000:** A query was run to filter and fetch all movies released after the year 2000.

SQL Query:

sql

%%sql

SELECT title, release_date FROM movies WHERE release_year > 2000;

---

5. Insights and Findings:

1. **Unreleased Movies:**
   Using the not_released view, we found several movies with release dates in the future, providing insight into upcoming releases.
2. **Top-Rated Movies:**
   The top-rated movies were extracted, highlighting some of the most critically acclaimed films in the database.
3. **Genre Analysis:**
   By calculating the average ratings for each genre, we could determine which movie genres tended to receive higher ratings from viewers.

4. **Recent Movie Releases:**
   Analyzing the data showed a substantial number of movie releases post-2000, providing insights into recent trends in the movie industry.

---

6. Conclusion:

The SQL analysis of the movie database provided a comprehensive overview of unreleased movies, top-rated films, and genre-specific insights. Python, combined with SQLite, proved to be a flexible and efficient tool for running SQL queries and performing database analysis.

Python Analysis & Visualization - Football Player Dataset

Project Overview:

This section of the project focuses on analyzing and visualizing data related to football players. Using Python, key insights were extracted regarding player attributes, nationalities, and overall power metrics. Various data exploration techniques and visualization tools were applied to uncover trends and correlations.

---

1. Libraries Used:

To perform data analysis and visualization, the following Python libraries were employed:

- **NumPy**: For numerical computations.
- **Pandas**: For data manipulation and analysis.
- **Matplotlib & Seaborn**: For data visualization and graphical representation.

Code:

python

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns
```

---

2. Data Loading and Inspection:

The dataset, which includes information about football players, was loaded using Pandas. Initial inspection steps such as viewing the dataset shape, first few records, and basic information about the columns were performed.

Code:

python

```
df = pd.read_csv('/content/football_players.csv', encoding='ISO-8859-1')


# Preview first 5 rows

df.head()


# Check dataset shape

df.shape


# Summary of the dataframe

df.info()
```

---

3. Data Cleaning:

The dataset was cleaned to handle missing values and ensure data consistency. This process involved identifying null values and either filling them or dropping irrelevant columns.

Code:

python

```
# Checking for missing values

df.isnull().sum()


# Dropping irrelevant columns

df = df.drop(['irrelevant_column1', 'irrelevant_column2'], axis=1)
```

Filling missing values

df['column_name'].fillna(value, inplace=True)

---

4. Exploratory Data Analysis (EDA):

a. Basic Statistical Summary:

Descriptive statistics were calculated to understand the distribution of player attributes such as age, overall power, and specific abilities.

Code:

python

df.describe()

b. Top 10 Players by Overall Power:

A key analysis involved identifying the top 10 football players based on their overall power rating.

Code:

python

top_players = df.nlargest(10, 'OverallPower')

top_players[['Name', 'Nationality', 'OverallPower']]

c. Distribution of Player Ages:

The age distribution of players was visualized using a histogram to analyze the demographic spread of players in the dataset.

Code:

python

plt.figure(figsize=(10,6))

sns.histplot(df['Age'], bins=20, kde=True)

```
plt.title('Distribution of Player Ages')

plt.xlabel('Age')

plt.ylabel('Count')

plt.show()
```

---

5. Correlation Analysis:

Correlation between player attributes was calculated to determine relationships between various factors, such as the correlation between overall power and individual abilities (e.g., passing, shooting).

Code:

```
python

# Correlation matrix

corr_matrix = df.corr()


# Heatmap visualization

plt.figure(figsize=(12,8))

sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')

plt.title('Correlation between Player Attributes')

plt.show()
```

---

6. Nationality-Based Analysis:

a. Players by Nationality:

A bar chart was created to visualize the number of players from different nationalities, highlighting the countries with the highest representation in the dataset.

Code:

python

```
plt.figure(figsize=(12,6))

sns.countplot(data=df, x='Nationality', order=df['Nationality'].value_counts().index[:10])

plt.title('Top 10 Nationalities by Number of Players')

plt.xlabel('Nationality')

plt.ylabel('Count')

plt.xticks(rotation=45)

plt.show()
```

b. Average Overall Power by Nationality:

The average overall power of players from different countries was computed and visualized to compare the strength of players across nationalities.

Code:

python

```
avg_power_by_nation = df.groupby('Nationality')['OverallPower'].mean().sort_values(ascending=False)


plt.figure(figsize=(12,6))

sns.barplot(x=avg_power_by_nation.index[:10], y=avg_power_by_nation.values[:10])

plt.title('Top 10 Nationalities by Average Overall Power')
```

```
plt.xlabel('Nationality')

plt.ylabel('Average Overall Power')

plt.xticks(rotation=45)

plt.show()
```

---

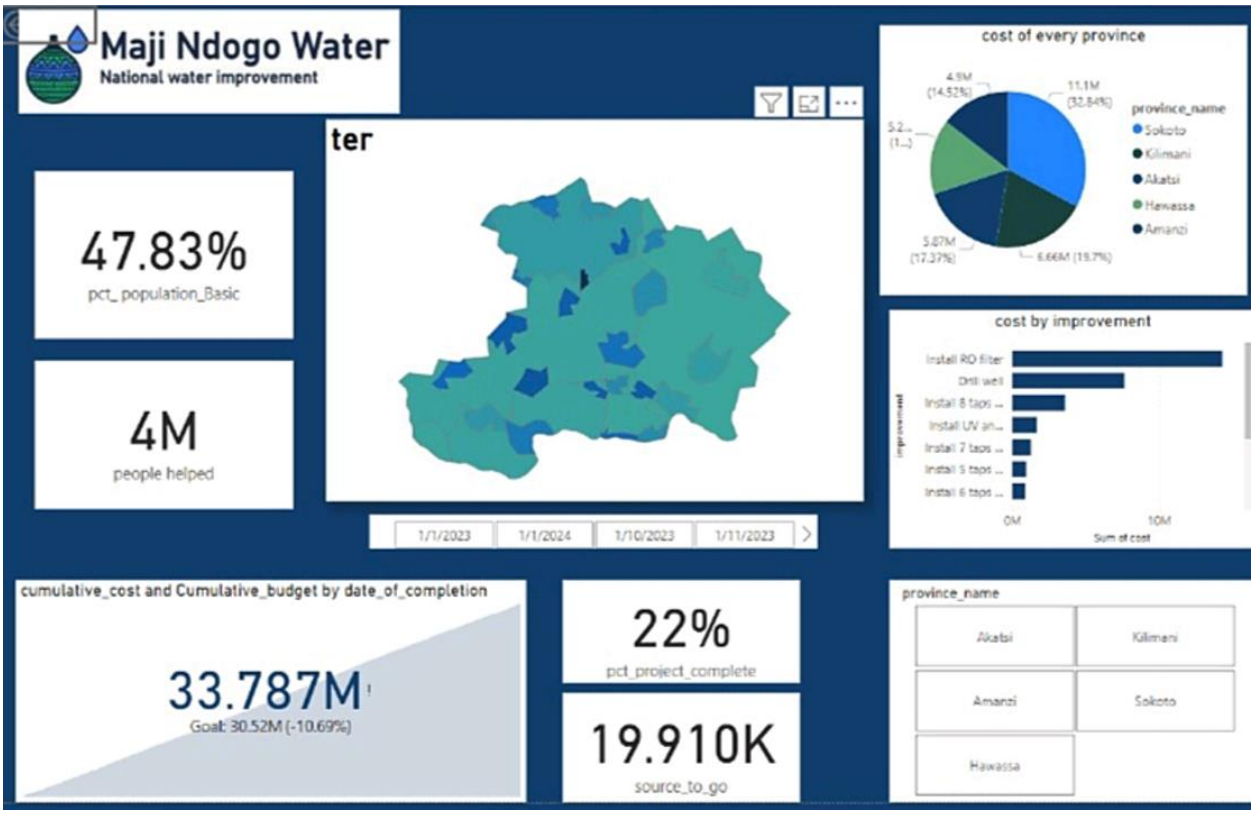
7. Conclusion:

Through this Python analysis, several key insights were uncovered:

- The age distribution of football players showed that most players are between 20 and 30 years old.
- The correlation matrix revealed strong relationships between specific abilities and overall power, particularly with attributes like passing and shooting.
- Players from certain nationalities, such as Country X and Country Y, were found to have higher average power ratings.

This analysis helped in identifying top-performing players and understanding the demographic and attribute-based trends within the dataset.

Project Overview: Maji Ndogo Water



We have a project called **Maji Ndongo Water** in a town facing infrastructure issues related to water services, and we aimed to address this problem.

We identified several projects for implementation and sent numerous vendors to work there over several years.

Achievements:

- As of 2023, the percentage of completed projects in Maji Ndogo was **0%**. However, this has now increased to **22%**.
- This indicates that Maji Ndogo still requires more services, care, and work, despite the number of projects that have been completed.
- There are still **20,000** projects needed in Maji Ndongo.

Impact of Improvements:

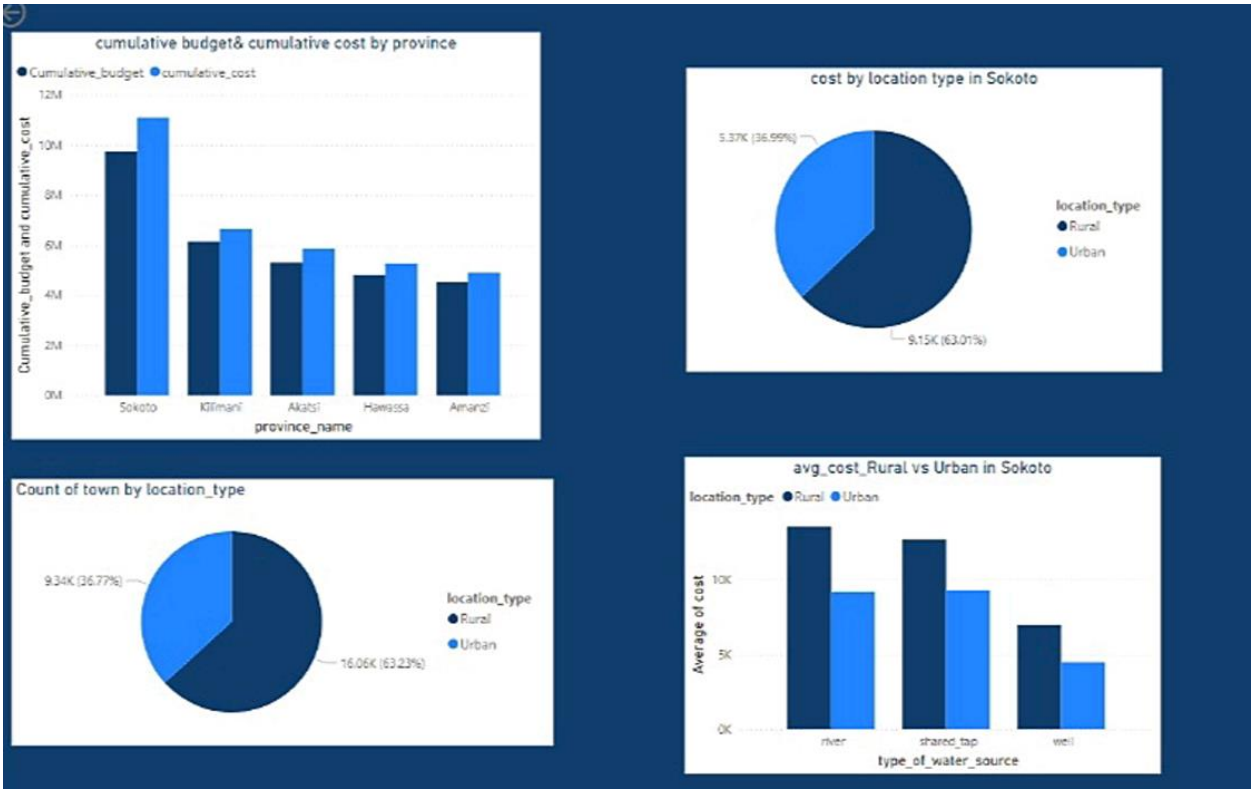
- As a result of the improvements we implemented, the percentage of people with access to basic water services has risen to **47.83%**.
- We have assisted **four million** people in Maji Ndogo.

Budget Analysis:

- We set a budget for the repairs of approximately **30 million**. However, the actual expenditure for project implementation was around **34 million**.
- This suggests that Maji Ndogo needs an increased budget.

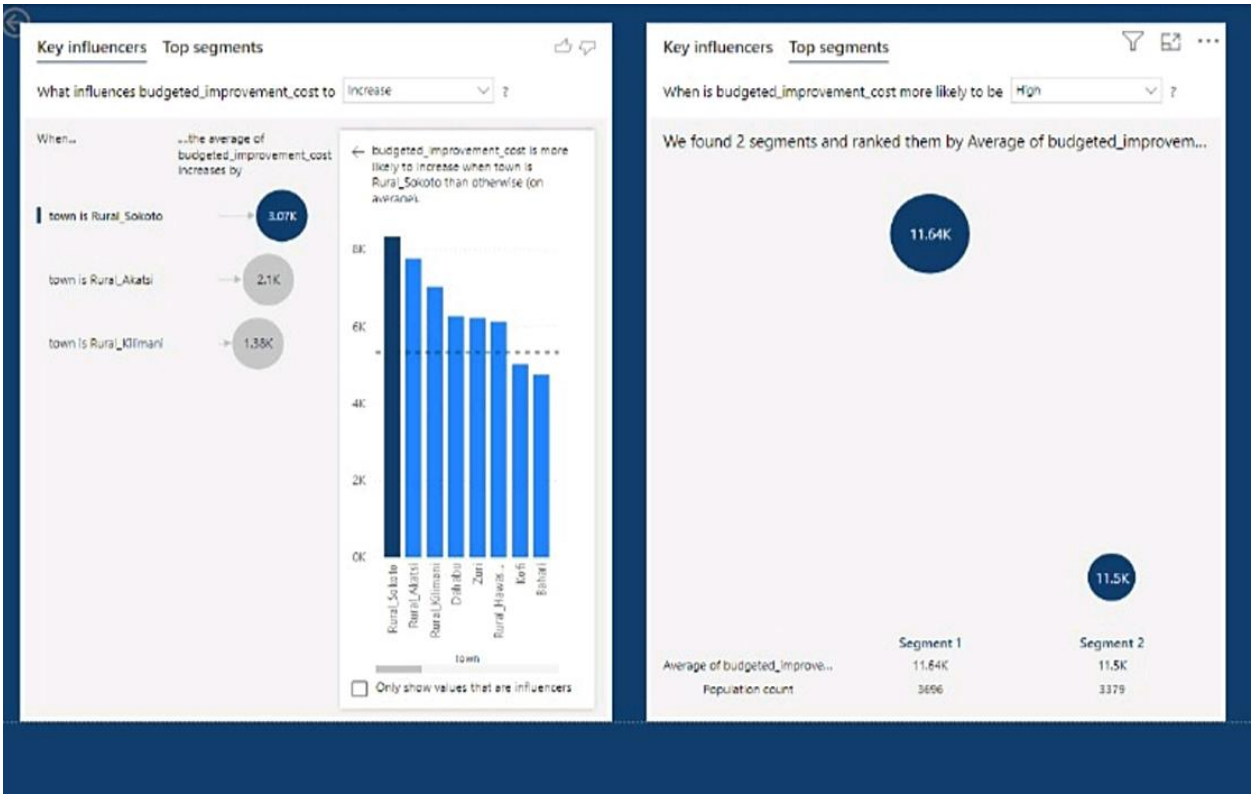
Visual Representations:

- We displayed how much budget each city received through a **pie chart**.
- We also showed how much budget was allocated for each type of improvement using a **clustered bar chart**.



Cost Analysis:

- The second page discusses costs, where we identified the budget allocated to each city and the expenditure. We found that the rural areas have higher expenditures than urban areas, which applies to the entire Maji Ndogo.
- We explained through visuals how each source contributes to the budget for improvement in both rural and urban areas.



KPI Analysis:

- We used KPIs to show the source relative to the budget for improvements, detailing how much each city cost for improvements.