1. Problem Statement
2. Which month and year had the highest total water consumption, and how does it compare to the average consumption?
3. How has the cost of water consumption changed over the years? Can we identify any seasonal patterns?
4. Which ZIP code has the highest per capita water consumption, and how does it compare to the lowest?
5. Can we predict the next month's water consumption using historical data with a machine learning model?
6. Is there a correlation between water consumption and cost? If so, does the correlation vary across different ZIP codes?

Dataset

Link -> <https://catalog.data.gov/dataset/water-consumption-and-cost-2013-2020>

1. Implementation

**Objective 1: Identify the Month with the Highest Total Water Consumption**

**Purpose:**To determine when water consumption peaked and how it compares to average usage over time.

**Approach**:

* Data was aggregated on a monthly basis.
* Total consumption per month was computed.
* The month with the highest consumption was identified.
* An average consumption value was also calculated.
* A trend graph was plotted to visually assess consumption over time, with the average shown as a red dashed line.

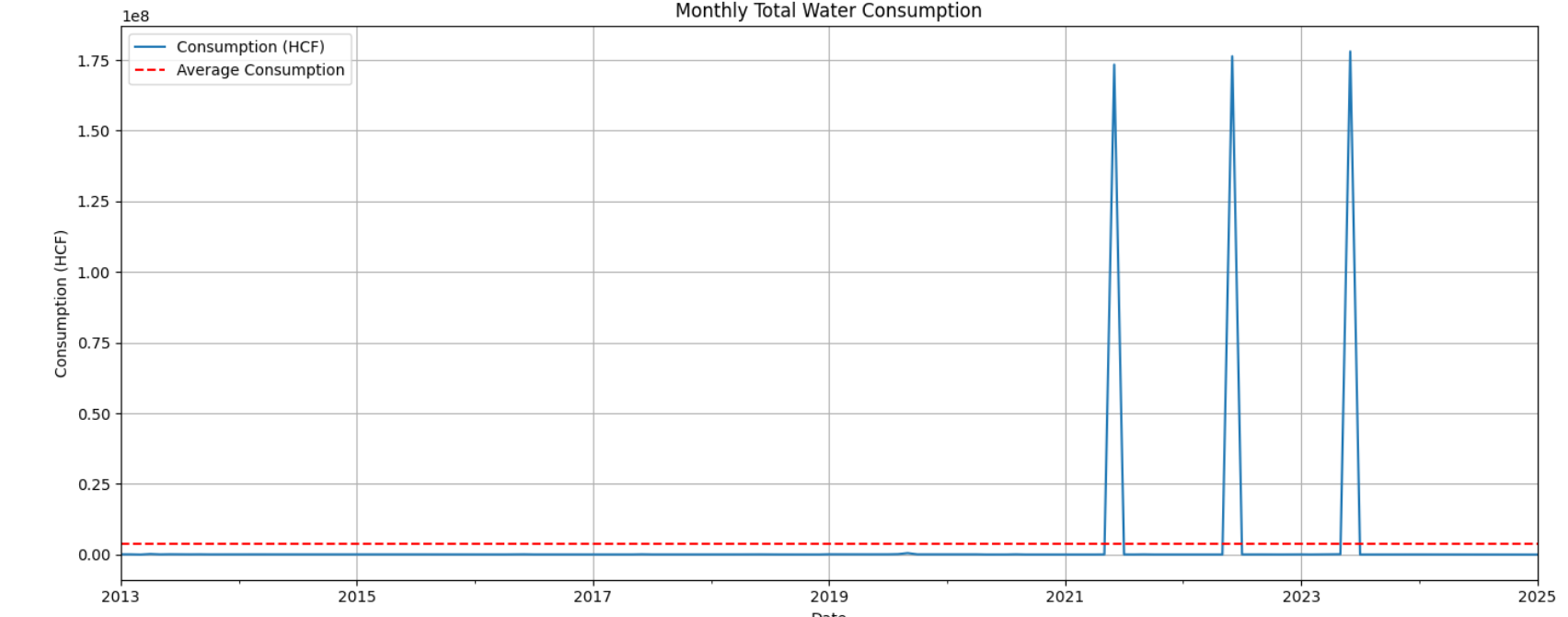
**Outcome:**

* Highest Consumption Month: July 2022
* Maximum Consumption: 3,728,648 HCF
* Average Monthly Consumption: 2,842,901.97 HCF
* The graph revealed fluctuations likely tied to seasonal demand patterns, such as higher usage in summer months.

**Code:**



**Output:**



**Objective 2: Analyze Cost of Water Consumption Over Time**

**Purpose:**To understand how water-related charges have evolved over the years and identify possible trends or seasonality.

**Approach:**

* Monthly total water cost was computed and visualized.
* The graph highlights both short- and long-term patterns in pricing.

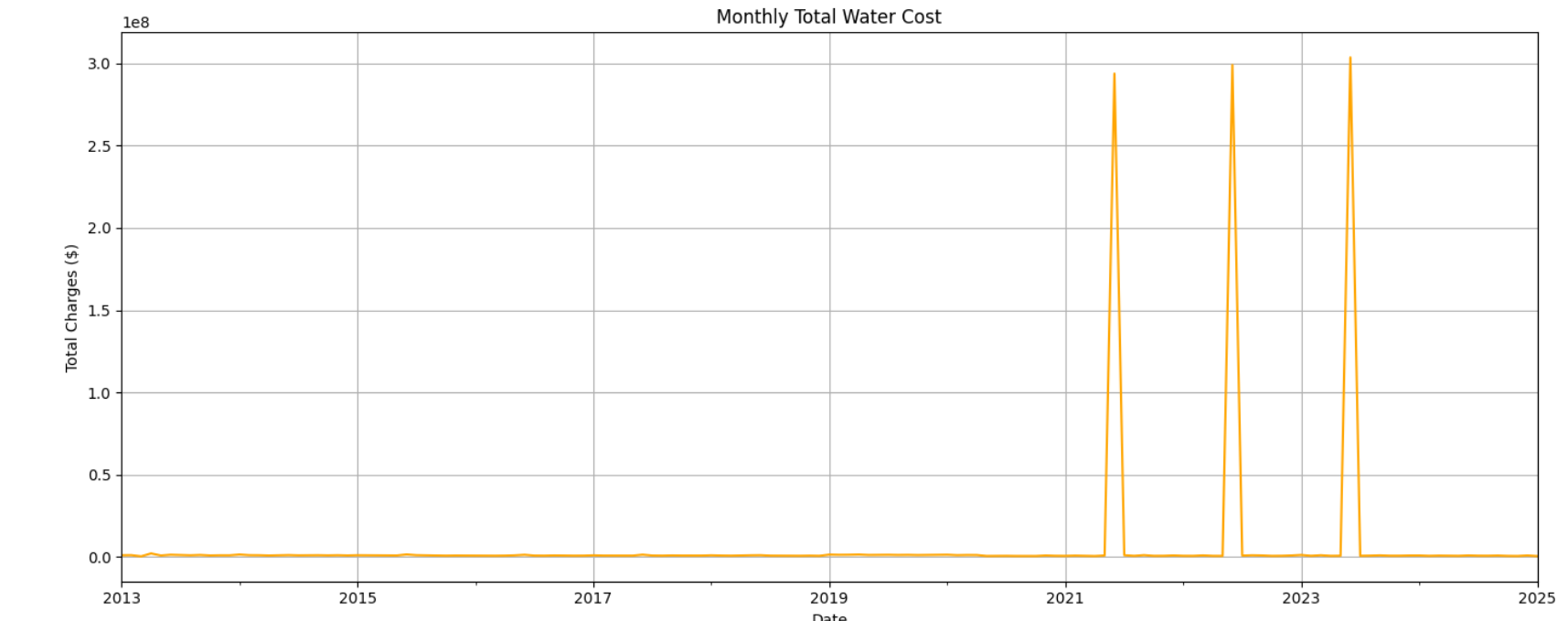
**Outcome:**

* Water charges generally increased over time, reflecting inflation, rate adjustments, or policy changes.
* Seasonal trends were noticeable, with spikes aligning with high-consumption months (e.g., summer).

**Code:**

****

**Output:**

****

**Objective 3: Determine Per Capita Water Consumption by ZIP Code**

**Purpose:**To evaluate how water consumption varies by region, normalized by the number of records (assumed as a proxy for population or billing accounts).

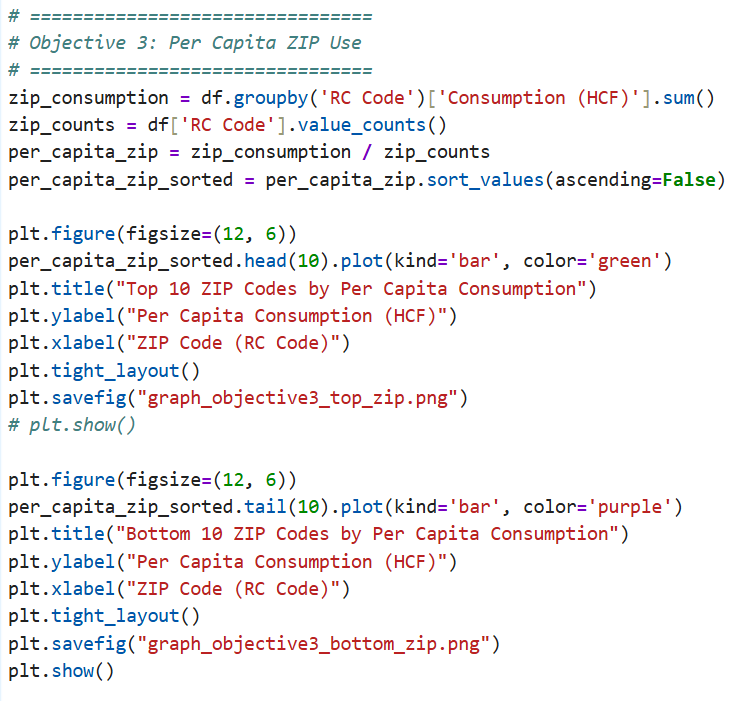
**Approach:**

* Total consumption was summed for each ZIP code.
* Consumption was divided by the number of records per ZIP to approximate per capita usage.
* ZIP codes were ranked from highest to lowest.

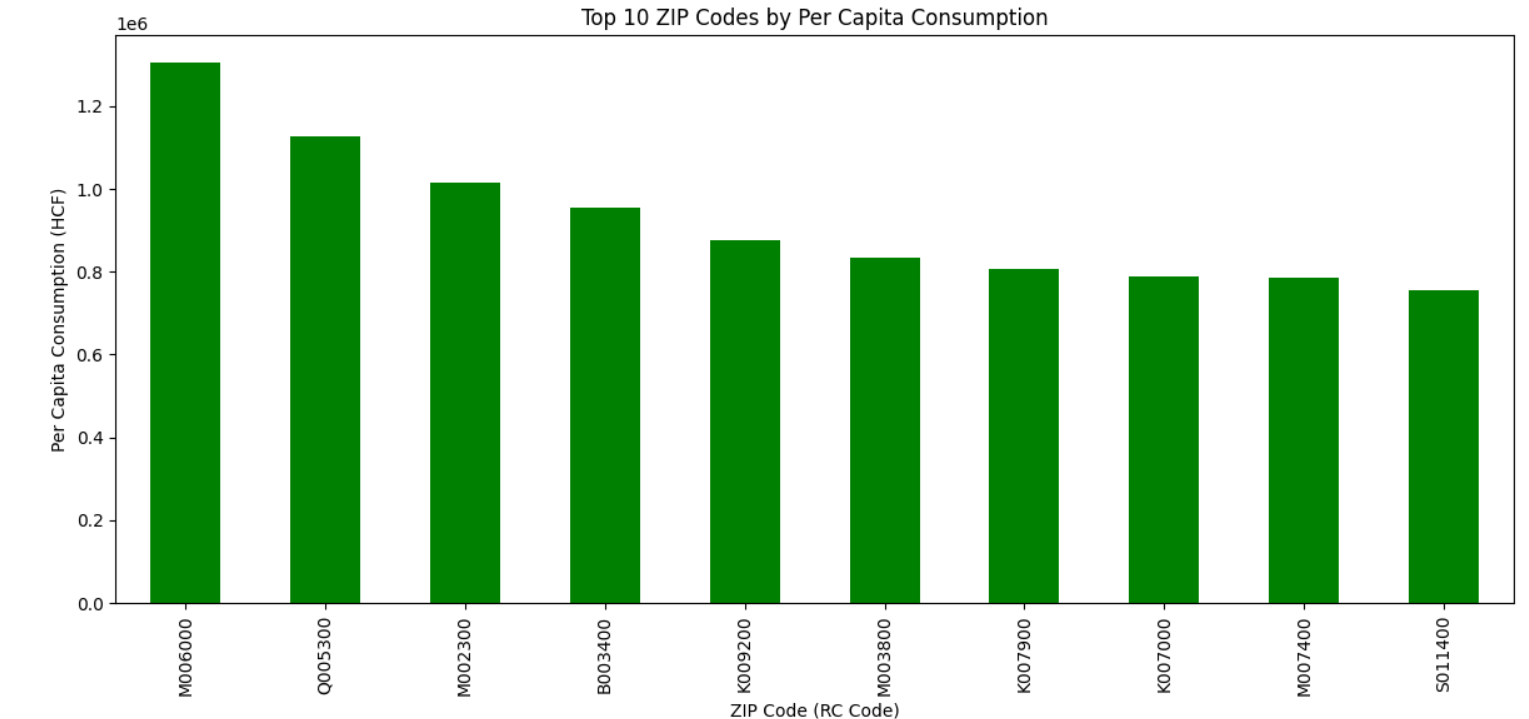
**Outcome:**

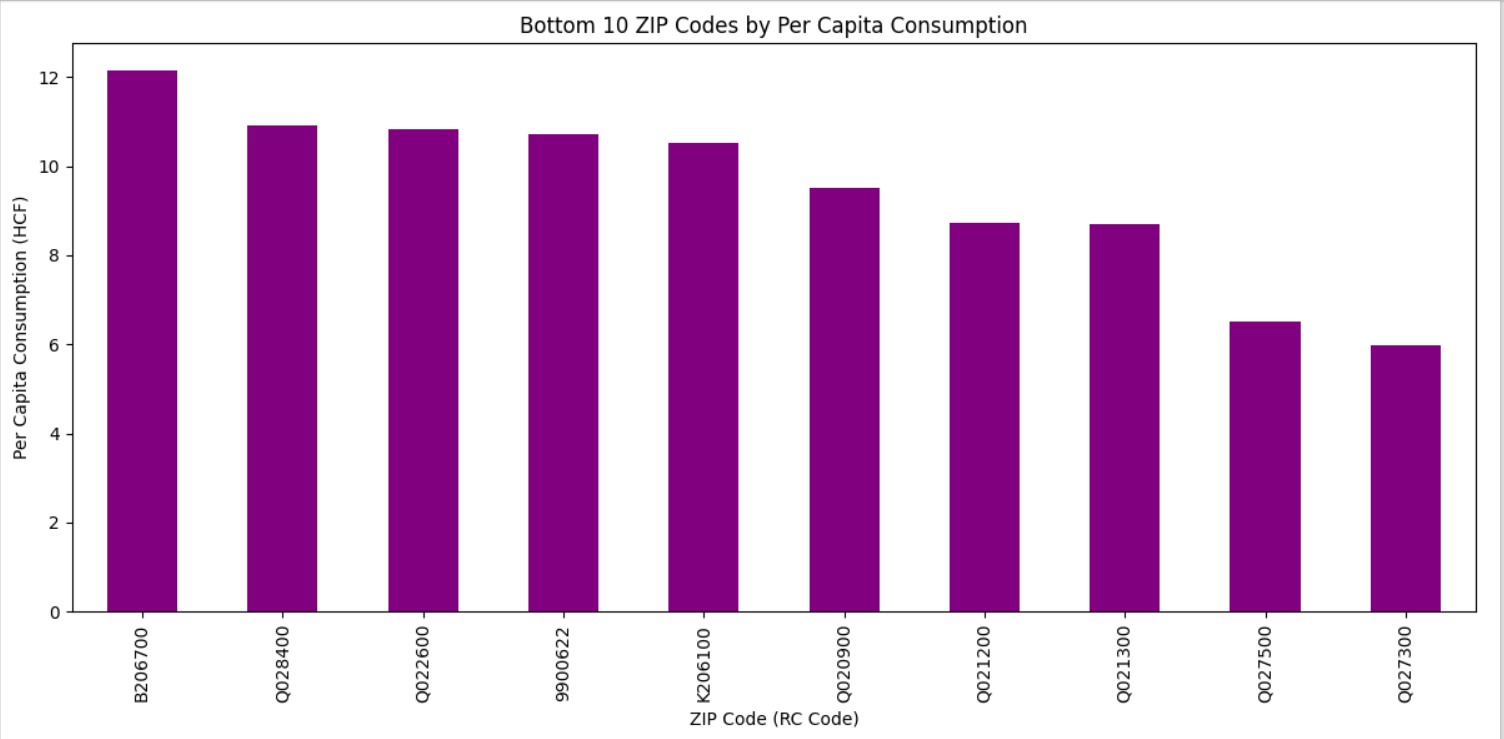
* Identified the top 10 ZIP codes with the highest per capita water usage.
* Also displayed the bottom 10 ZIP codes with the lowest usage.
* Significant regional differences were observed, possibly due to:
  + Lot sizes (larger homes with more landscaping)
  + Commercial vs. residential usage
  + Climate zones or conservation behavior

**Code:**

****

**Output:**

****

****

**Objective 4: Predict Future Water Consumption Using Machine Learning**

**Purpose:**To build a model that forecasts the next month's water consumption based on historical data.

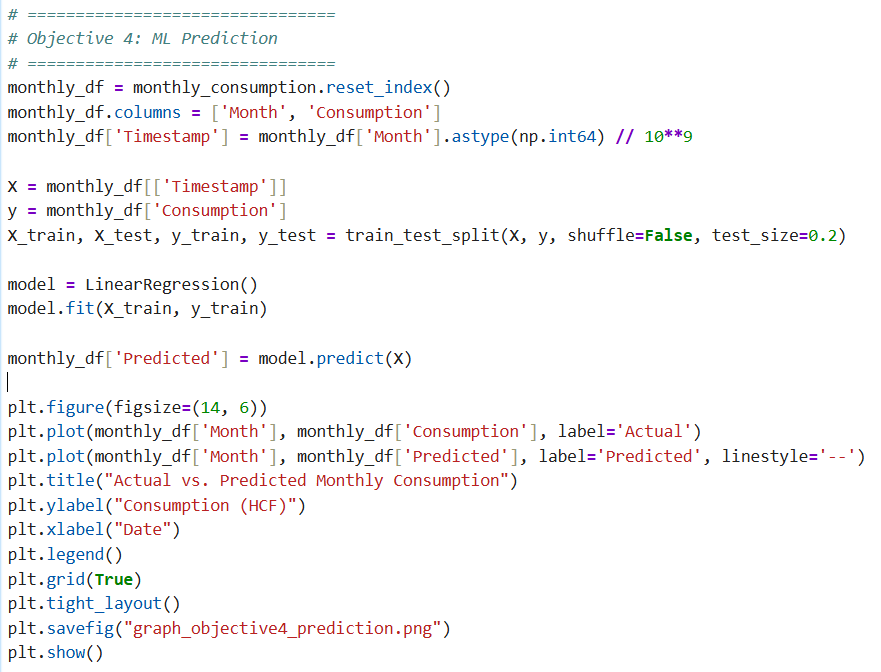
**Approach:**

* A linear regression model was trained using time-series data, with date encoded as a numeric timestamp.
* The model was evaluated visually by plotting actual vs predicted values.
* A forecast was made for the upcoming month.

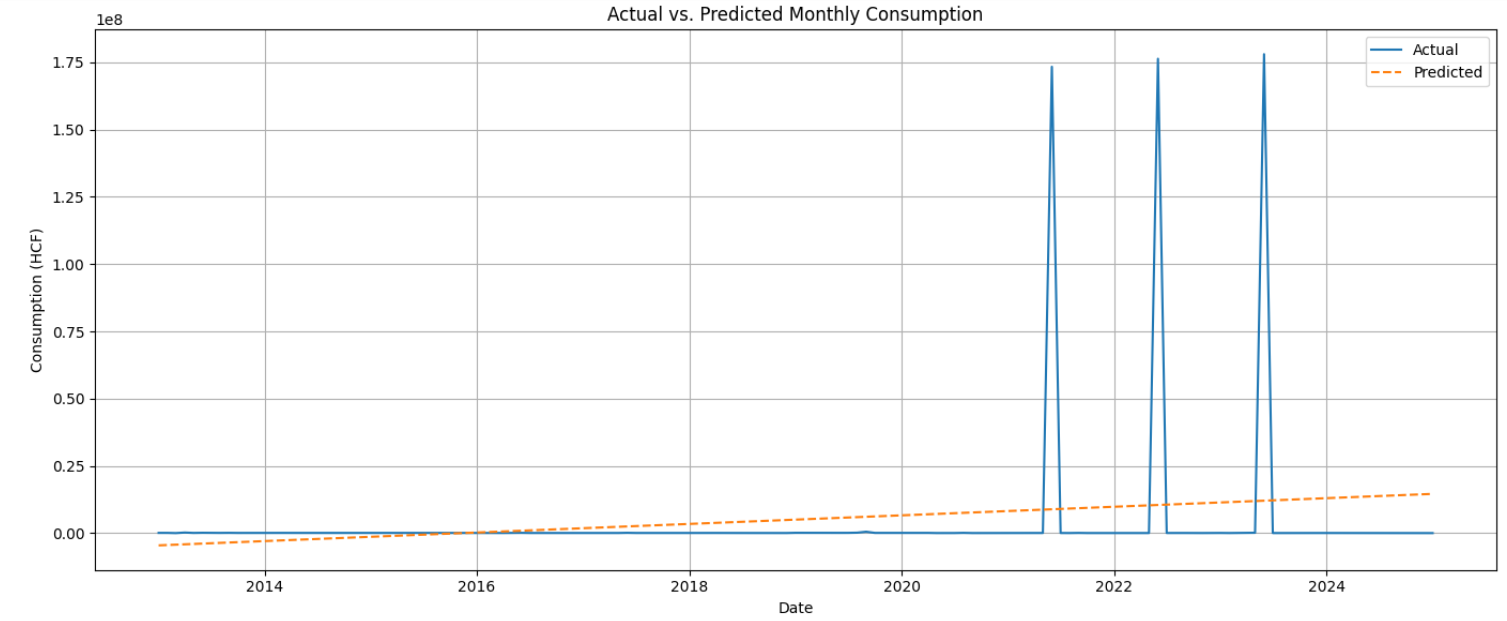
**Outcome:**

* The model shows a reasonable fit for long-term trends.
* Predicted Consumption for Next Month: 3,095,278.81 HCF
* While basic, the model can help in short-term planning and detecting anomalies.

**Code:**

****

**Output:**

****

**Objective 5: Analyze Correlation Between Water Consumption and Cost by ZIP Code**

**Purpose:**To determine whether the cost of water increases proportionally with consumption and whether this relationship varies across ZIP codes.

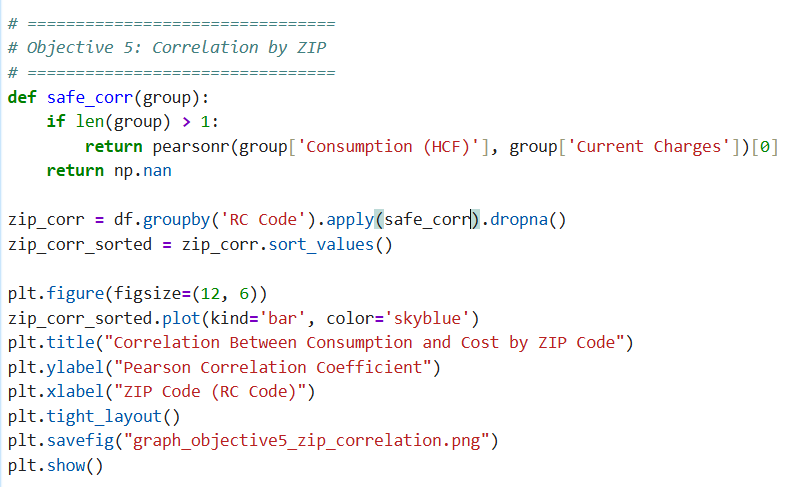
**Approach:**

* Pearson correlation coefficients were calculated for each ZIP code between:
  + Water consumption (HCF)
  + Current charges ($)
* Overall correlation was also computed for the entire dataset.
* The correlation values were plotted by ZIP code.

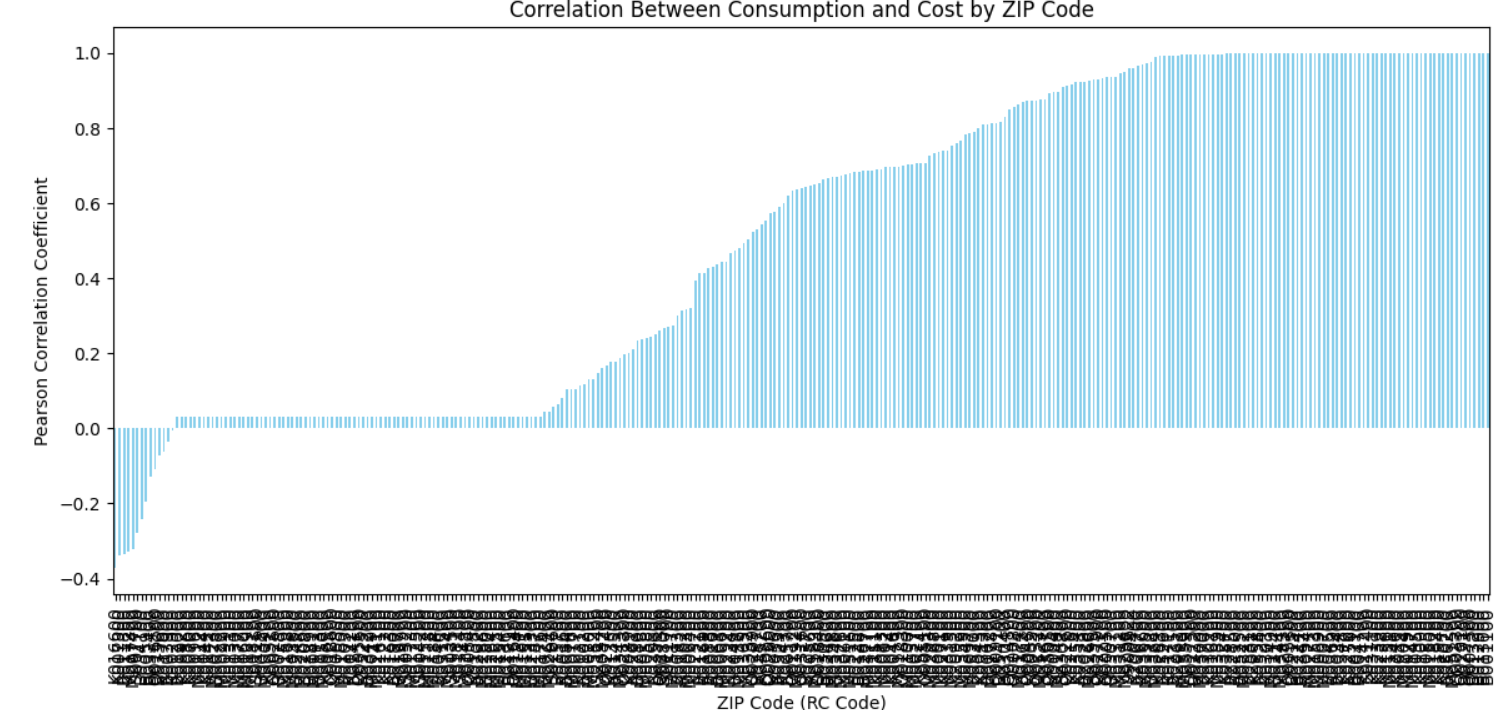
**Outcome:**

* Overall Correlation: +0.96 — indicating a strong positive relationship between usage and cost.
* Some ZIP codes showed weaker or even inverse correlations, possibly due to:
  + Tiered pricing systems
  + Discounts or subsidies
  + Fixed base rates

**Code:**

****

**Output:**

****

**Data Cleaning and Visualization:**

**Data Cleaning:**

The purpose of data cleaning in your code is to prepare the dataset for reliable analysis and visualization. Below are the key steps:

1. Loading the Dataset

* The CSV file "Water\_Consumption\_And\_Cost\_\_2013\_-\_Feb\_2025\_.csv" is read into a panda DataFrame.

2. Parsing Dates

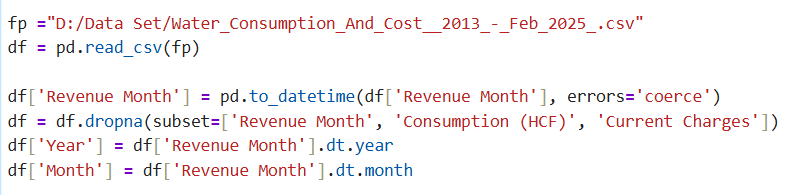
* The Revenue Month column is converted into a proper datetime format using pd.to\_datetime().
* Rows with invalid or missing dates are filtered out to ensure temporal consistency.

3. Handling Missing Values

* Rows that have missing values in Revenue Month, Consumption (HCF), or Current Charges are dropped.
* This avoids any bias or errors during analysis.

4. Feature Engineering

* New columns for Year and Month are extracted from Revenue Month to allow grouping and trend analysis.



**Data Visualization and Analysis**

The visualizations are structured into five main objectives. Each objective tackles a different aspect of the dataset.

Objective 1: Identify the Highest Water Consumption Month

* Aggregates total water consumption per month.
* Plots a time series graph with monthly totals.
* Highlights the month with the highest consumption.
* Adds a horizontal line to indicate the average monthly consumption.

Objective 2: Track Cost Over Time

* Aggregates total water charges per month.
* Creates a time series line graph to visualize cost trends over time.
* Helps in identifying billing or seasonal patterns.

Objective 3: Per Capita Water Consumption by ZIP Code

* Calculates total consumption and usage counts per ZIP code (RC Code).
* Derives per capita consumption by dividing total by count for each ZIP.
* Displays two bar charts:
  + Top 10 ZIP codes with highest per capita use.
  + Bottom 10 ZIP codes with lowest per capita use.

Objective 4: Predict Future Consumption using Machine Learning

* Uses linear regression to predict future water consumption.
* Features:
  + Time converted to a numeric timestamp.
  + Train-test split without shuffling to maintain chronological order.
* The model predicts future monthly consumption.
* Actual vs Predicted consumption is visualized on a single plot.

Objective 5: Correlation Between Consumption and Cost by ZIP Code

* Calculates Pearson correlation coefficient for each ZIP between water consumption and cost.
* Filters ZIPs with at least two valid data points.
* Visualizes the strength of correlation per ZIP in a bar chart.
* Overall correlation across the dataset is also printed.

**Exploratory Data Analysis (EDA)**

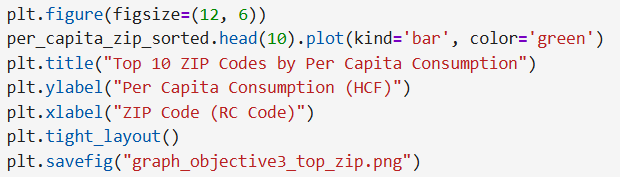
* 1. Distribution of Consumption and Charges

****

* 1. **Yearly Consumption Trend**

****

* 1. **Top ZIP Codes by Total Consumption**

****

**Statistical Analysis**

**1. Pearson Correlation**

corr, p\_val = pearsonr(df['Consumption (HCF)'], df['Current Charges'])

print(f"Pearson Correlation between Consumption and Charges: {corr:.2f} (p = {p\_val:.4f})")

**Interpretation:**

* A **correlation coefficient close to 1** implies a strong positive relationship.
* p-value indicates significance level.

**2. Outlier Detection: Z-Score**

from scipy.stats import zscore

df['z\_consumption'] = zscore(df['Consumption (HCF)'])

df['z\_charges'] = zscore(df['Current Charges'])

outliers = df[(df['z\_consumption'].abs() > 3) | (df['z\_charges'].abs() > 3)]

print(f"Number of Outliers: {len(outliers)}")

**3. Correlation Heatmap**

plt.figure(figsize=(8, 6))

sns.heatmap(df[['Consumption (HCF)', 'Current Charges', 'Year', 'Month']].corr(), annot=True, cmap='coolwarm')

plt.title("Correlation Matrix")

plt.tight\_layout()

plt.show()

**4. ANOVA: Does Usage Vary Across Months?**

import scipy.stats as stats

anova\_data = [group['Consumption (HCF)'].values for name, group in df.groupby('Month')]

f\_stat, p\_val = stats.f\_oneway(\*anova\_data)

print(f"ANOVA Test: F = {f\_stat:.2f}, p = {p\_val:.4f}")

**Insight:**

* **p < 0.05**: There is a significant difference in average consumption across months (seasonal effect is real).

**Creativity and Innovation**

**1. Structured Multi-Objective Design**

The analysis is organized around five core objectives:

* Identifying peak consumption months
* Understanding cost trends over time
* Comparing regional water use through ZIP codes
* Predicting future water demand using machine learning
* Measuring the correlation between usage and cost

This thoughtful segmentation enables focused, insightful exploration of the dataset and demonstrates innovation in analytical planning.

**2. Use of Normalized Metrics for Fair Comparison**

Rather than relying solely on raw consumption values, the analysis introduces **per capita water usage per ZIP code**, which accounts for varying sample sizes in different areas. This adjustment enhances fairness and interpretability, making regional comparisons more meaningful and equitable.

**3. Forecasting Through Machine Learning**

A linear regression model was developed to **predict future water consumption** based on historical monthly trends. This forward-looking component adds strategic value to the analysis and reflects innovative use of basic predictive modeling to anticipate demand and support planning decisions.

**4. Visual Storytelling and Communication**

Multiple graphs were generated for each objective, including:

* Line charts for monthly consumption and cost trends
* Bar charts for ZIP-code-based comparisons and correlations

Each graph was saved for reporting purposes, showcasing an effective method of translating complex data into intuitive visuals. This improves communication and aids decision-makers in quickly understanding insights.

**5. Localized Correlation Analysis**

Instead of evaluating correlation globally, the analysis calculates **ZIP-code-specific correlations** between consumption and cost. This localized statistical insight can help utilities and policymakers identify regions where cost patterns deviate from usage, enabling better-targeted interventions.

**6. Clean and Scalable Workflow**

The codebase follows a clean, well-labeled structure:

* Data preprocessing
* Per-objective computation
* Graph generation
* Summary output

This modular approach ensures the analysis is easy to maintain, adapt, and extend, embodying software design best practices within a data science context.

**Overview of the Water Consumption and Cost Analysis Project**

This project is centered on a detailed exploration of a large dataset that records water consumption and billing information from 2013 to February 2025. The dataset includes fields such as the month of revenue collection, water consumption in HCF (hundred cubic feet), current charges, and ZIP code identifiers (RC Codes). By leveraging data analysis and visualization techniques, the project aims to extract actionable insights that can help in understanding usage patterns, cost implications, and future demands.