Introduction

Water scarcity is an important problem for İstanbul as the city is always growing as well as environmental factors are constantly changing due to yearly cycles or global warming. Therefore, understanding the balance between natural water supply and human demand is essential for ensuring long-term water sustainability. In this project, the relationships between rainfall, water consumption, population growth, and reservoir water storage will be examined in Istanbul over the period from 2011 to 2024.

The analysis tries to answer four main hypotheses: the effect of rainfall on reservoir storage, the impact of water consumption on reservoir levels, the relationship between population growth and water consumption, and the combined effect of rainfall and water consumption on long-term water sustainability which is the reservoir water storage over the period from 2001 to 2024.

To test these hypotheses, the study employs exploratory data analysis (EDA) alongside statistical and machine learning models. The first three hypotheses are evaluated using Pearson and Spearman correlation analyses. Additionally, for the first hypothesis one-month and two-month lagged rainfall effects are analyzed to account for environmental delay using Pearson and Spearman correlation analyses. For second, third, and fourth hypotheses Ordinary Least Squares (OLS) regression is applied to quantify causal relationships. Lastly, fourth hypothesis is assessed through machine learning models which are linear, ridge, and random forest regression.

Data & Data Collection

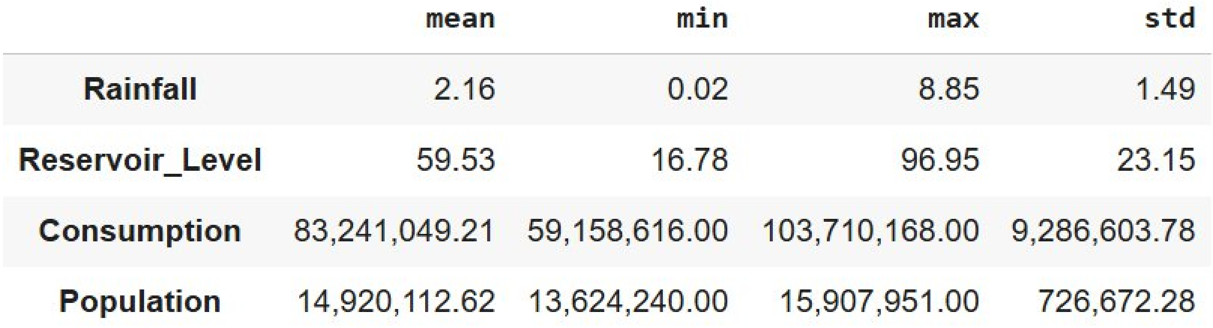
Reservoir levels, rainfall, and water consumption data were obtained from the Kaggle: Istanbul 2011–2024 Dam, Precipitation and Consumption Dataset, while population growth statistics was sourced from nufusu.com which is based on official Turkish Statistical Institute (TUİK) records.

While reservoir storage data were available from October 2000 to February 2024, the primary analysis period was taken between 2011 to 2024 to ensure temporal alignment with the other variables. On the other hand, for the fourth hypothesis regarding long-term sustainability the full range from 2001 to 2024 was used. Firstly, all date columns were transformed into a standardized datetime format and ensured that there was not a missing date. Then, the numbers were transformed to float as they were object, and the wrong rates were normalized by dividing by 100 to correct the scaling. Finally, daily data were converted to monthly averages.

Rainfall and daily consumption were in one dataset therefore after date columns were transformed into a standardized datetime format, and controlled that there was no date was missing, daily consumption was converted into monthly totals and daily rainfall was aggregated into monthly averages.

Population data were in yearly format between 2009 to 2024, but since it was manually curated, it was processed directly between 2011 to 2024. Date columns were changed into a standardized datetime format, and the population values were stored as objects turned into integers. Finally, the annual data were redistributed to include each month of that year.

Exploratory Data Analysis (EDA)



	mean	min	max	std
Rainfall	2.16	0.02	8.85	1.49
Reservoir_Level	59.53	16.78	96.95	23.15
Consumption	83,241,049.21	59,158,616.00	103,710,168.00	9,286,603.78
Population	14,920,112.62	13,624,240.00	15,907,951.00	726,672.28

Table 1

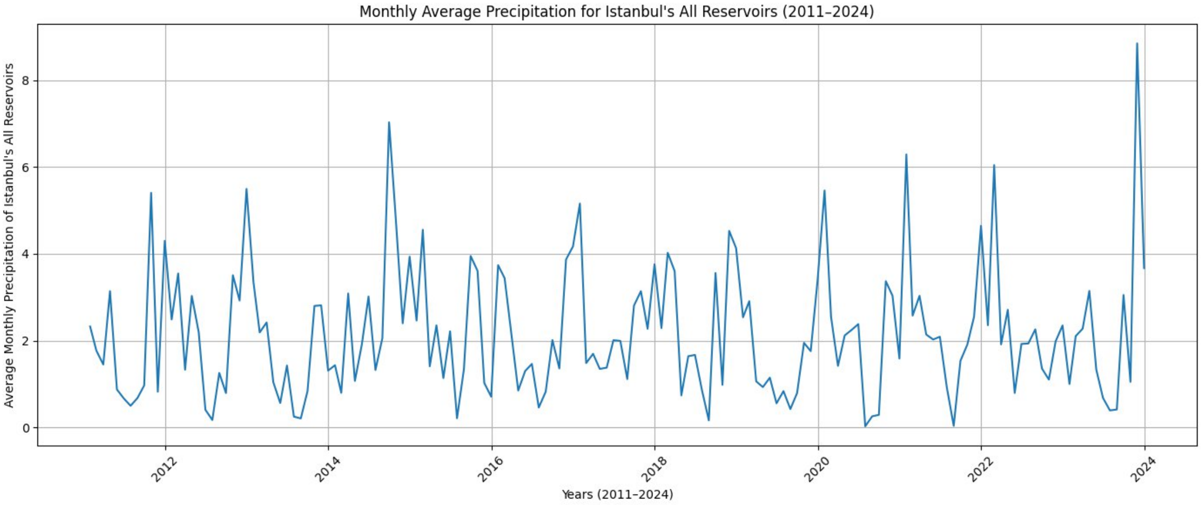


Figure 1

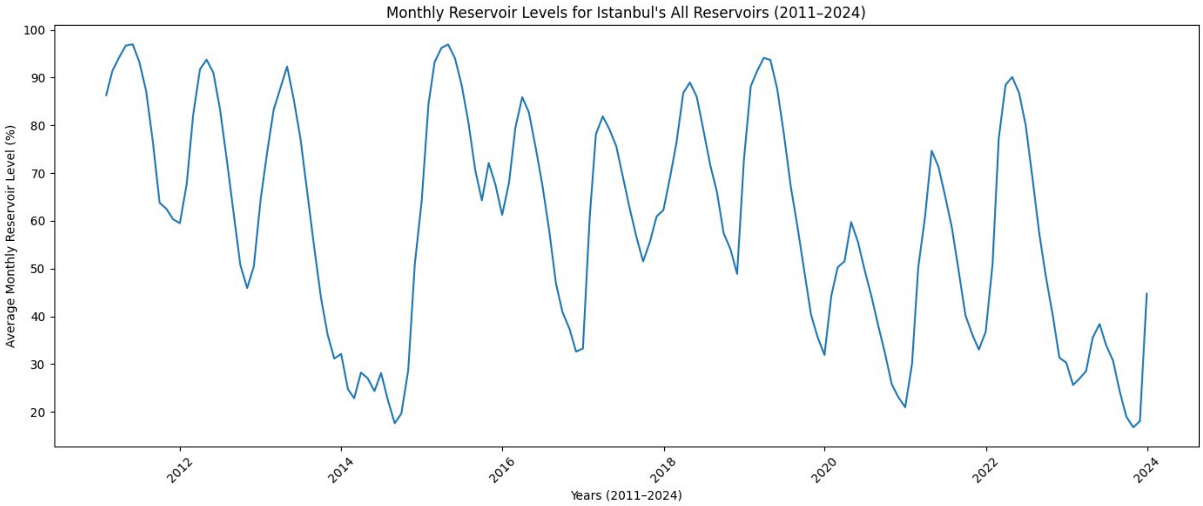


Figure 2

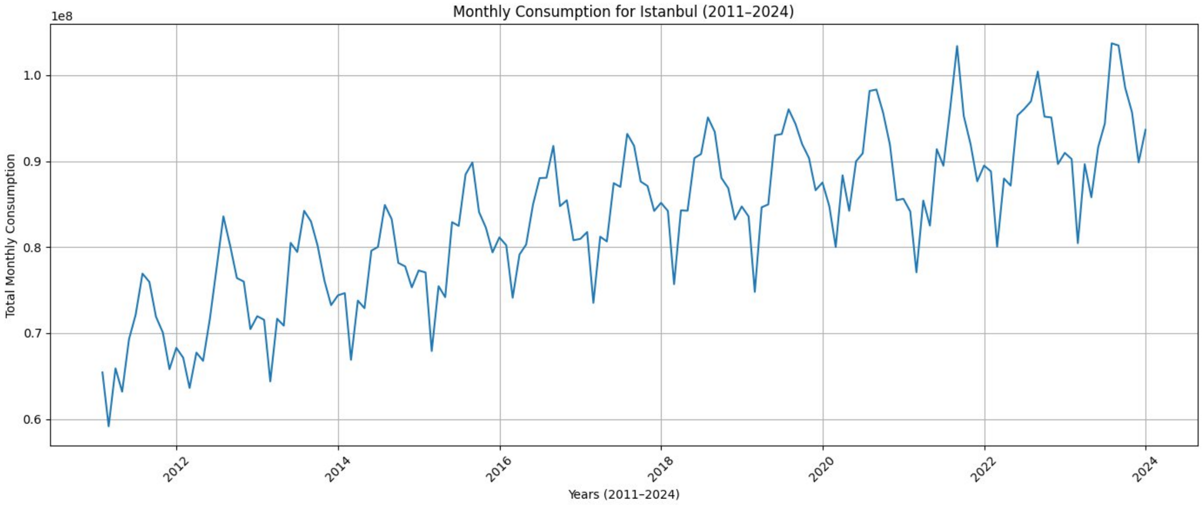


Figure 3

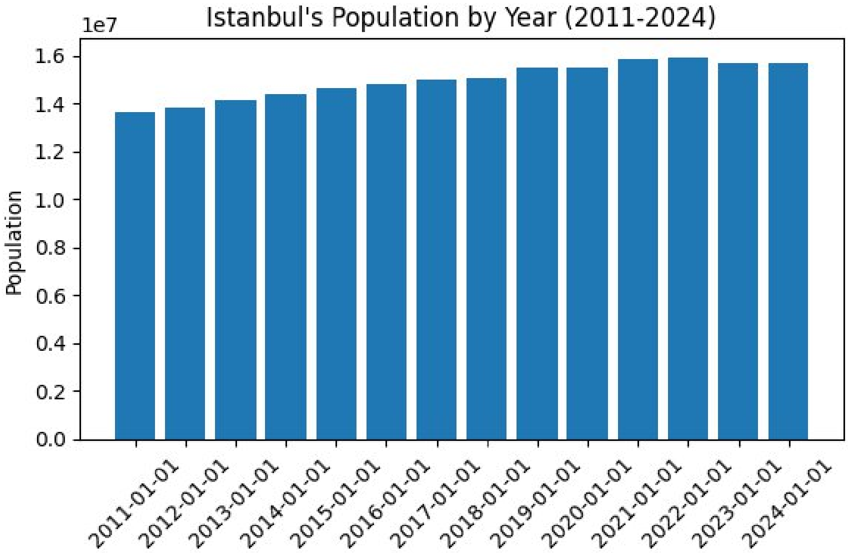


Figure 4

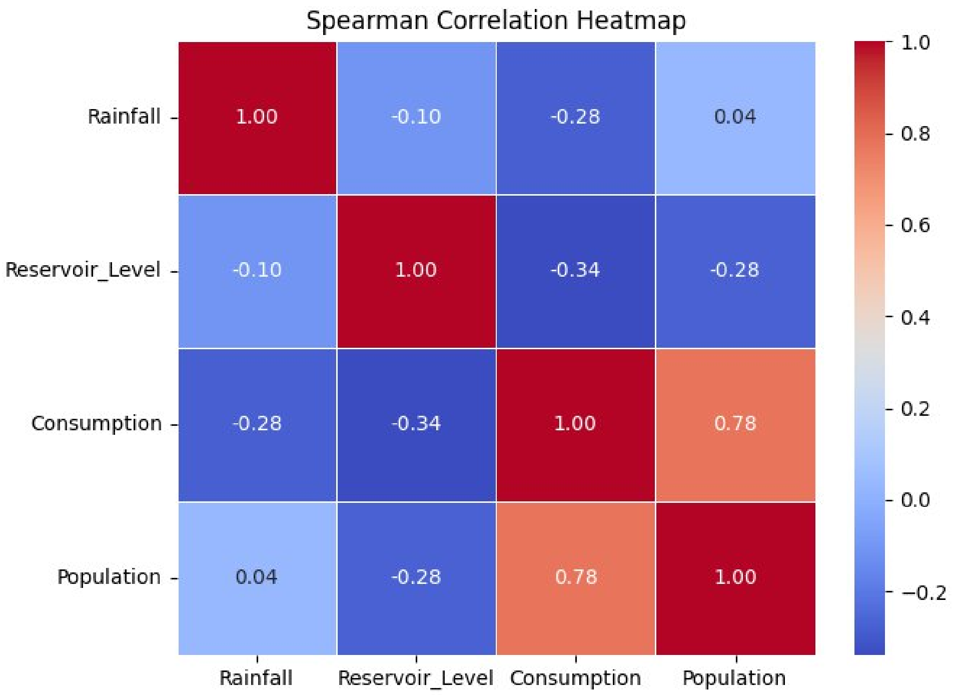


Figure 5

Monthly average rainfall in Istanbul exhibits high variability, as shown in Table 1, the standard deviation of 1.49 mm is high compared to the mean of 2.16 mm. Additionally, demonstrated in Figure 1, city does not follow a stable rainfall regime, and tends to alter between dry periods and heavy downpours which also influences reservoir levels over time. As shown in heatmap (Figure 5), the lack of a strong contemporaneous correlation between reservoir levels and rainfall indicates that the rainfall's impact on reservoir levels is not immediate and needs to be looked at the delayed effects.

Monthly average reservoir levels show inconsistencies similar to the rainfall as the standard deviation of 23.15% against the mean of 59.5% reveals in Table 1. For instance, as seen in Figure 2, reservoir levels reached their maximum levels in 2015 after the 2014 drought and continue to constantly fluctuate throughout years. Moreover, the wide range between minimum of 16.78% and maximum of 96.95% throughout the study period demonstrates that reservoir levels are highly sensitive to both environmental changes and rising urban consumption.

Water consumption shows similar seasonal periodicity throughout years with water demand increasing in summer months and receding in winter as demonstrated in Figure 3. Additionally, despite these seasonal cycles, there is a consistent upward trajectory in water demand as water consumption starting with approximately 60 million m³ increase to more than 100 million m³ throughout the study period is shown in Table 1. The increase of water consumption highly correlated with Istanbul's population growth as the Spearman correlation coefficient of 0.78 shown in Figure 5.

Istanbul's population growth is consistent with the low standard deviation relative to its scale with respect to Table 1. Moreover, according to Figure 4, while the population increased from 13.6 million in 2011 to a peak with nearly 15.9 million in 2022, the last two years shows a slight decline, although the water demand continues to rise in Figure 3 despite the decrease in population.

Methodology

To test research hypotheses, this study integrates statistical analysis with machine learning techniques. The first three hypotheses are evaluated using Pearson and Spearman correlation analyses. Both analyses chosen to evaluate the differences as Pearson correlation is used to measure the strength of linear relationships and is not good with outliers while the Spearman correlation evaluates monotonic relationships and used more with the data contains more outliers. Additionally, for the first hypothesis, one-month and two-month lagged rainfall effects are evaluated with Pearson and Spearman correlation analyses because the effects of rainfall on reservoir levels are delayed as mentioned above.

For the second, third, and fourth hypotheses, Ordinary Least Squares (OLS) regression is implemented. Especially for second and third hypotheses, while correlation analyses evaluate whether values acted together, OLS allows for estimating the specific magnitude of impact through regression coefficients, therefore, it is applied to obtain more detailed results. The fourth hypothesis contains three variables, and OLS has an advantage of multiple regression, therefore it is chosen. Moreover, OLS is applied to the two-month lagged version of first hypothesis as this version has the strongest correlation results making it the most suitable for further modeling.

In the final stage of the study, three different machine learning algorithms are compared to evaluate fourth hypothesis regarding long-term water sustainability. Linear Regression is used as a baseline model as its linearity identifies overall trends while it can be inadequate in predicting sudden changes. On the other hand, Ridge Regression is used as it is more controlled, uses regularization to reduce the noise as well as to obstruct overfitting. Finally, Random Forest Regression is implemented to capture non-linear relationships that other models could miss as it is more flexible by averaging the results of multiple decision trees.

Results

Hypothesis 1: the impact of rainfall on reservoir levels

The initial correlation shows a weak negative relationship between rainfall and reservoir levels in the same month, with a Pearson coefficient of -0.15 ($p = 0.04$) and a Spearman coefficient of -0.10 ($p = 0.19$). These results suggest that rainfall does not have a strong immediate effect on reservoir levels.

When a one-month lag is applied, the relationship becomes weakly positive, with both Pearson ($r = 0.11$, $p = 0.05$) and Spearman ($\rho = 0.15$, $p = 0.05$) correlations indicate that rainfall has a minimal effect after one-month.

The strongest results with positive relationships are observed with a two-month lag. The Pearson correlation increases to 0.31 ($p < 0.001$) and the Spearman correlation reaches 0.32 ($p < 0.001$) which means rainfall affecting the reservoir levels the most after two-months. Therefore, OLS model is used for further research with a coefficient of 5.14 indicating that each 1 mm increase in monthly average precipitation two months prior leads to a 5.14% increase in current reservoir levels, and a F-statistic $p < 0.001$ showing strong correlation aligning with previous correlation results. Moreover, the R-squared value for this model is 0.097, suggesting that lagged rainfall explains approximately 10% of the variance in Istanbul's water storage which has a strong impact as there are other environmental and urban factors that affect reservoir levels.

Hypothesis 2: the impact of water consumption on reservoir levels

The correlation of second hypothesis shows a strong negative relationship between water consumption and reservoir levels, with a Pearson coefficient of -0.35 ($p < 0.001$) and a Spearman coefficient of -0.33 ($p < 0.001$). Moreover, The OLS regression model further supports these findings with a F-statistic $p < 0.001$. Although the regression coefficient -8.771e-07 appears small due to the scale difference between total consumption (in cubic meters) and occupancy rates (0–100%), the negative sign confirms that as consumption increases, reservoir levels experience a decline. The R-squared value of 0.124 suggests that water consumption alone explains approximately 12.4% of the fluctuations in reservoir levels which indicates that urban factors have a big impact on Istanbul’s reservoir levels, comparable to the influence of environmental factors.

Hypothesis 3: the impact of population on water consumption

The correlation of third hypothesis shows a strong positive relationship between population growth and water consumption, with both Pearson ($r = 0.78$, $p < 0.001$) and Spearman ($\rho = 0.77$, $p < 0.001$). Moreover, the OLS regression model further supports these findings with a F-statistic $p < 0.001$ and regression coefficient 10.0652 indicates that for every additional person added to the city's population, daily water consumption increases by approximately 10.07 m³. The R-squared value of 0.620 is the highest among all tested models. This suggests that population alone accounts for 62% of the variance in Istanbul's water consumption as one of the most significant factors in long-term demand growth.

Hypothesis 4: the impact of rainfall and water consumption on water sustainability

The OLS regression model is used as multiple regression as evaluates the combined impact of rainfall and water consumption on Istanbul's water sustainability. Model indicates that both variables remain statistically significant, $p = 0.001$ for rainfall and $p < 0.001$ for consumption, and both have negative relationship with regression coefficient -3.96 for rainfall and -1.028e-06 for consumption. However, the reason for the negative coefficient in rainfall is attributed to a two-month time lag identified in Hypothesis 1, and consumption coefficient is small because of differences of the scales. The R-squared value is 0.186 which indicates that combination of rainfall and water consumption have an impact of 18.6% on water sustainability.

Three machine learning methods, Linear, Ridge, and Random Forest regression, are evaluated and all the methods showed negative R-square scores as -0.0703 for linear regression, -0.0701 for ridge regression, and -0.3934 for random forest. Moreover, Root Mean Square Errors (RMSE) are 22.21 for linear regression, 22.20 for ridge regression and 25.34 for random forest which indicates a high prediction error relative to the 0-100% scale of reservoir levels. Even though ridge regression is slightly better with bigger R-square score and smaller RMSE, the negative R-square score suggests that the relationship between precipitation, consumption, and reservoir levels is highly non-linear and volatile.

Discussion

The results of the first hypothesis explain the dynamic between rainfall and Istanbul's reservoir levels. The fact that the strongest correlation occurs with a two-month lag suggests that the rainfall does not directly transfer to the reservoir storage as it could be that groundwater and soil moisture levels must first reach saturation before dam occupancy can increase. This perspective is supported by the distinction between meteorological and hydrological drought; while a lack of rain is immediate, its impact on dams develops only after a significant reduction in river flow and underground water (Kurnaz, 2014). Moreover, due to various factors such as Istanbul's soil structure and climate, this transition period appears to take approximately two months. Therefore, while further research could enable more effective decision-making in water management, this delay shows that ensuring long-term water policies is essential as decisions made hastily may not be sufficient.

The R-squared of 12% in the second hypothesis, compared to approximately 10% in the first hypothesis, demonstrates that Istanbul reservoir levels are influenced to a larger extent by urban consumption than by rainfall. This highlights how important policies related to water consumption are since human-induced depletion is more manageable and controllable pressure than natural factors like rainfall.

This urban pressure is mostly connected with the population growth as the results of the third hypothesis show that Istanbul's population explains 62% of the variance in water consumption. However, while official population figures decreased in last two years (Figure 3), water consumption continued to rise (Figure 4). This trend could be attributed to the high volume of tourists and unregistered residents who consume the city's resources but are not reflected in official census data. Moreover, this increase might be caused by a lack of public awareness regarding water conservation, leading to overuse in urban areas. Therefore, water reduction strategies in domestic environments might be a part of the solution such as usage of water efficient equipment in toilets, dishwashers, kitchen tabs, etc. (Yalçıntaş, et al., 2015).

The final stage of the study, which integrated multiple regression and machine learning models, reveals the inherent difficulty in forecasting Istanbul's water sustainability. While R-squared of 18.6% in OLS model indicates that the rainfall and water consumption has a visible effect on water sustainability, the negative R-squared scores across all machine learning models show that the relationship between environmental factors and reservoir levels is not simply linear or easily predictable. This lack of predictive accuracy, combined with a relatively high RMSE of 22.20 (Ridge Regression), suggests that the system is highly volatile.

This unpredictability can be attributed to the other factors discussed in previous hypothesis, such as unregistered population with tourists or meteorological and hydrological drought. Moreover, while dams are the primary source of water for Istanbul, the consequences of climate change might affect long-term water sustainability as reservoir levels are highly sensitive to climatic conditions (Yalçıntaş, et al., 2015). Therefore, the negative results of the models do not mean failure but rather shows a scientific proof that Istanbul's water sustainability is in a fragile state, and immediate actions on water management is a need, as it can be finding different ways for water resources such as rainwater harvesting or protection of already existing water resources by implementing new regulations that cause wiser consumption such as incentivized pricing on water bills (Daloğlu Çetinkaya, et al., 2023).

Conclusion

This study has revealed that water sustainability in Istanbul is a complex structure where both natural cycles and urban dynamics are intertwined. The analysis results revealed that rainfall was reflected in the reservoir levels with a two-month delay, water consumption had a greater impact than rainfall, and that most of the consumption originated from population growth. Moreover, machine learning models indicate that Istanbul's water sustainability is sensitive and unpredictable as well as that more comprehensive research needs to be done to capture the non-linear nature of Istanbul's water levels. Overall, ensuring long-term sustainability requires a multifaceted approach that both urban and environmental policies need to implement such as promoting water-efficient appliances in households and finding new water sources.

References

Daloğlu Çetinkaya, I., Yazar, M., Kılınç, S., & Güven, B. (2023). Urban climate resilience and water insecurity: future scenarios of water supply and demand in Istanbul. *Urban Water Journal*, *20*(10), 1336-1347.

Kurnaz, L. (2014). Drought in Turkey. *İstanbul Policy Center, Sabancı Üniversitesi-İstanbul*.

nufusu.com. (2024). İstanbul Population. URL: https://www.nufusu.com/il/istanbul-nufusu

Özçelik, B. (2024). Istanbul 2011–2024 Dam, Precipitation and Consumption Dataset. Kaggle. URL: https://www.kaggle.com/datasets/noepinefrin/istanbul-2011-2024-dam-precipitation-and-consumption?resource=download

Yalçıntaş, M., Bulu, M., Küçükvar, M., & Samadi, H. (2015). A framework for sustainable urban water management through demand and supply forecasting: The case of Istanbul. *Sustainability*, *7*(8), 11050-11067.