FairWater Analysis Report

Group - 9

Semester 2 - 23/24

Abstract

Context

The escalating demands for water in urban and residential settings, compounded by the challenges of climate change, necessitate efficient water management strategies. Understanding household water consumption patterns is crucial for developing sustainable water usage practices and infrastructure planning.

Objectives

The project aimed to elucidate water usage patterns across different household appliances and times, identify peak usage hours, and determine which appliances consume the most water on a weekly basis. Additionally, we sought to predict future water flow rates using statistical modeling, thereby offering actionable insights for optimizing water consumption.

Results

Our hourly analysis identified three main usage peaks—early morning (6-8 AM), early afternoon (1-3 PM), and evening (7-9 PM)—across appliances like Bidet, KitchenFaucet, Shower, and Washbasin, with dishwashers and washing machines showing distinct patterns. Weekly trends revealed higher consumption during weekends, especially noticeable in aggregated data and certain appliances, with Showers and Washing Machines noted as the highest water users. Predictive modeling provided accurate future water flow forecasts, enhancing demand anticipation capabilities.

Novelty

The novel aspect of this work lies in its comprehensive examination of water usage across multiple household appliances, combined with the application of both Linear Regression and ARIMA models for prediction. This dual-model approach, especially in the context of residential water usage, offers a new perspective by not only identifying consumption patterns but also predicting future demands with significant accuracy. Our findings contribute actionable insights, paving the way for more informed water conservation strategies and infrastructure decisions.

Introduction

The FairWater project represents a pioneering initiative led by Northumbrian Water in collaboration with Newcastle University, aimed at revolutionizing the way we understand and manage domestic water consumption. In the face of escalating environmental concerns and the pressing need for sustainable resource management, the project seeks to harness the power of data science to unveil new technologies and methodologies. These innovations are designed to meticulously monitor water usage across various household appliances and activities, identify inefficiencies, and foster behavioral changes among consumers. By doing so, FairWater aspires to significantly mitigate water wastage, reduce the carbon footprint, and alleviate the financial burdens of utility bills, particularly for low-income families, the elderly, and other vulnerable groups.

Embarking on this project, we adopt the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, which provides a comprehensive and flexible roadmap for tackling complex data-driven challenges. The CRISP-DM cycle encompasses six critical phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This structure guides our approach in dissecting the multifaceted problem of water consumption and inefficiency within domestic settings.

Round 1 of the CRISP-DM Cycle

1. Business Understanding

1.1 Business Objectives

Aligned with the FairWater project's mission to revolutionize domestic water consumption understanding and management, our business objectives are centered on creating sustainable water resource management strategies in response to escalating environmental concerns. By leveraging data science, we aim to develop innovations that are not merely technological but also promote behavioral change, enhancing water use efficiency across various household appliances and activities. Our purpose extends beyond environmental impact to address socio-economic factors, seeking to alleviate the financial burden of utility bills, especially for low-income families, the elderly, and other vulnerable groups.

In keeping with the initiative's pioneering spirit, led by Northumbrian Water in collaboration with Newcastle University, we intend to dive into a multifaceted analysis of water usage patterns. By identifying inefficiencies and driving conservation behaviors, we aspire to significantly reduce water wastage, cut down the carbon footprint, and create cost-effective solutions that can be implemented in existing housing stocks to benefit a broad spectrum of society.

1.2 Data Mining Goals

1. **Which hour of the day is the most usage happening across different appliance types?**

To alleviate the current situation of water stress, understanding household water use patterns and identifying the peak times of the day for different appliances are key to improving water efficiency, reducing energy consumption, and reducing financial burdens. We intend to explore during which periods of time households use the most water, and which household appliances' water consumption during these periods contributes to the peak in total water consumption. This knowledge will enable us to better design customized water-saving measures, such as promoting the use of efficient appliances, optimizing water usage habits, and even using certain appliances during off-peak hours to reduce overall water demand.

2. **Which appliance is used the most in weekly bases?**

Understanding and improving the water usage habits of household appliances can effectively enhance water resource utilization efficiency. This question aims to analyze and identify the types of appliances that consume the most water in homes. Such an analysis can help us understand the key factors driving household water use, whether specific household appliance water inefficiencies exist, and whether it is possible to reduce overall water use by replacing equipment with more efficient models or adjusting usage habits.

3. **How can we predict water usage by training a regression model on historical data, excluding the latest 2 months for model validation?**

Accurate predictions of future household water consumption are critical for enabling effective management and planning of water resources. This question explores how historical water use data can be utilized to train a regression model to predict future water demand. As a preliminary idea, we will select appropriate features, address seasonal and trend factors in the data, and choose and tune models suitable for predicting water consumption. Successful implementation of accurate forecasts not only assists households in managing water use more efficiently but also provides water utilities with valuable information to optimize water allocation and reduce waste.

1.3 Project Plan

The timeline for carrying out analyses and reporting the outcomes for this project is as follows: The project was released on 11th March 2024 including all the datasets with the deadline for the project completion being 22nd March 2024. The two weeks of available time was planned to be used roughly as follows:

- **Week 1** Create the Project using project template and load the data and analysis report. Going through the datasets - exploring data connections and possible relations, checking for data quality and missing data. Outlining the CRISP-DM layout. Clean and munge the data. Carry out Exploratory Data Analysis and draw insights about the success criteria defined.
- **Week 2** Modelling and Evaluation of the analysis carried out so far. Complete the project and the report with all the required files needed for submission. Complete the presentation & Dashboard as well.

Throughout the timeline, commits are to made via git and reflective log to be filled with observations and encounters. Our team members divided the tasks among us and took small steps together to achieve our collective goals.

1.4 Choices of Research Tools/Models/software

In this project, we carefully selected a series of tools and software to enable us to effectively process and analyze data, and further propose solutions to support our research and development of sustainable water management strategies. Below is an overview of the tools we selected and their application in the project. We will use Excel to conduct initial exploration of data, identify and deal with missing values, outliers, and data consistency issues. We will use R to perform preliminary exploratory data analysis (EDA), as well as data preprocessing.

We will use R for more in-depth statistical analysis and data visualization in the report to facilitate our discovery of problems and to further propose our decision-making solutions as well as for the predictive model. For the presentation purposes we are going to use PowerBI to create our dashboard and PowerPoint for the slides of our presentation. To ensure project collaboration efficiency and code version control, we chose to use Git as the version control system and GitHub as the code storage and team collaboration platform. This allows our team members to collaborate more effectively and also ensures the transparency and traceability of the project's continuous integration and deployment process.

Through the combined use of these tools, we aim to build a powerful analytical framework to support our business objectives and research needs. Our aim is to improve the efficiency and sustainability of water resources management through these technical means, taking into account socio-economic impacts, especially supporting vulnerable groups.

1.5 Constrains and Challenges

To face the challenges of water resources management, we must consider multiple dimensions, including data quality and availability, technological constraints, the difficulty of changing behavior, economic and policy factors, and the difficulty of technological development and innovation.

First, for data quality and availability:

- Data may have missing values or incorrect records. For example, water use data from November 2019 to July 2020 is missing in the total water use data. This affects the quality of the data and the accuracy of the analysis.
- Regarding data representativeness, the data collected only accounts for water use from different categories of households and does not reflect the water use of different households, including those in various regions, at different income levels, and with different household sizes.

Second, regarding technological constraints:

- The adoption of smart water meters and the need for high-performance computing resources highlight technological constraints, and in some regions, the coverage of smart water meters may not be widespread enough.
- Data processing power: High-performance computing resources may be required to process large amounts of data and to enable efficient data analysis and model training.

Also, regarding the difficulty of behavior change, it is a significant challenge to motivate users to change their deep-rooted water use habits; people in different regions and cultures may have different water use habits, and it may take time and patience to change these habits.

Finally, from an economic and policy perspective, the initial investment in water-saving technologies may be burdensome for low-income households, and a lack of policies and incentives may prevent the widespread implementation of water-saving measures.

2. Data Understanding

2.1 Data Collection

Sources

- The FairWater project's data was sourced exclusively from sophisticated smart water meters installed in residential properties. These meters provided detailed insights into household water consumption patterns. Our partnerships with local water utility companies and academic institutions, notably Newcastle University, enriched our dataset, enabling a comprehensive analysis of water use efficiency.

Types of Data

Our datasets is composed entirely of quantitative data:

- **Quantitative Data**: This includes high-resolution measurements from smart water meters, such as water flow rates, usage times, and volumes. This data is crucial for analyzing consumption patterns across different appliances and activities within the households.

Data Collection Methods

- **Smart Meters**: The cornerstone of our data collection, these devices were installed across a diverse range of domestic settings. They automatically recorded detailed water usage data, transmitting this information in real-time to the data collection system for analysis.
- **Smarphone Application**: An application was installed on the user's smartphone to manually detect usage events for dishwashers and toilet as well.

2.2 Exploring the data

The datasets for the FairWater project include detailed records of water usage for various household appliances and overall household consumption. Each dataset comprises a combination of the following columns:

- Time: The timestamp indicating when the water usage was recorded. This is a critical component for analyzing water consumption patterns over different periods.
- Flow: The amount of water used during the recorded period, typically measured in liters or gallons. This column is key to quantifying water usage across different appliances and activities.
- EndTime/EndFlow: Present in certain datasets (feedDishwasher and feedToilet), these columns provide additional context for the duration of water usage events or the completion of a usage cycle.

The specific datasets and their columns are as follows and include unix (a timestamp in Unix format) and flow (water flow rate):

- Aggregated Whole House Water Usage:
  - **aggregatedWholeHouse.csv**: “Time”, “Flow”
- Appliance-Specific Datasets:
  - **feedBidet.csv**: “Time”, “Flow”
  - **feedDishwasher.csv**: “Time”, “Flow”, “EndTime”
  - **feedKitchenfaucet.csv**: “Time”, “Flow”
  - **feedShower.csv**: “Time”, “Flow”
  - **feedToilet.csv**: “Time”, “Flow”, “EndFlow”
  - **feedWashbasin.csv**: “Time”, “Flow”
  - **feedWashingmachine.csv**: “Time”, “Flow”

It is notable that all datasets consistently track the time and flow of water usage, with the EndTime and EndFlow columns offering additional granularity for specific appliances. This uniformity facilitates a comprehensive analysis of water consumption patterns both at the appliance and household levels.

Initial exploration of these datasets focuses on assessing their quality, summarizing key statistics, and identifying preliminary patterns or insights. The uniform structure across most datasets simplifies this process, allowing for a cohesive analysis of water consumption trends. Detailed examination of the Time and Flow columns will enable the identification of peak usage times, efficiency opportunities, and the impact of specific appliances on overall water use. The additional EndTime and EndFlow columns in the dishwasher and toilet datasets provide a deeper understanding of water usage durations and cycles, crucial for pinpointing inefficiencies and devising targeted interventions.

Starting by checking if there are NA values in the datasets:

```
## NA values in df_Bidet: 0
```

```
## NA values in df_Kitchenfaucet: 0
```

```
## NA values in df_Shower: 0
```

```
## NA values in df_Washbasin: 0
```

```
## NA values in df_Washingmachine: 0
## NA values in df_AggregatedWholeHouse: 0
```

```
## NA values in df_Dishwasher: 0
```

```
## NA values in df_Toilet: 0
```

As we can see there are no NA values in the datasets which means there is no need for further adjustment in the preprocessing phase. Next, we will visualize the datasets to examine their contents thoroughly. This crucial step will provide us with a deeper understanding of the data's structure and characteristics, guiding us on how to best preprocess the data for our forthcoming analysis. Focusing on visualizing the outliers in our datasets is a strategic approach that enables us to identify and understand the extremities within our data. Outliers, which are data points significantly different from the majority of the data, can have a profound impact on the overall analysis, affecting statistical tests, models, and conclusions drawn from the data.

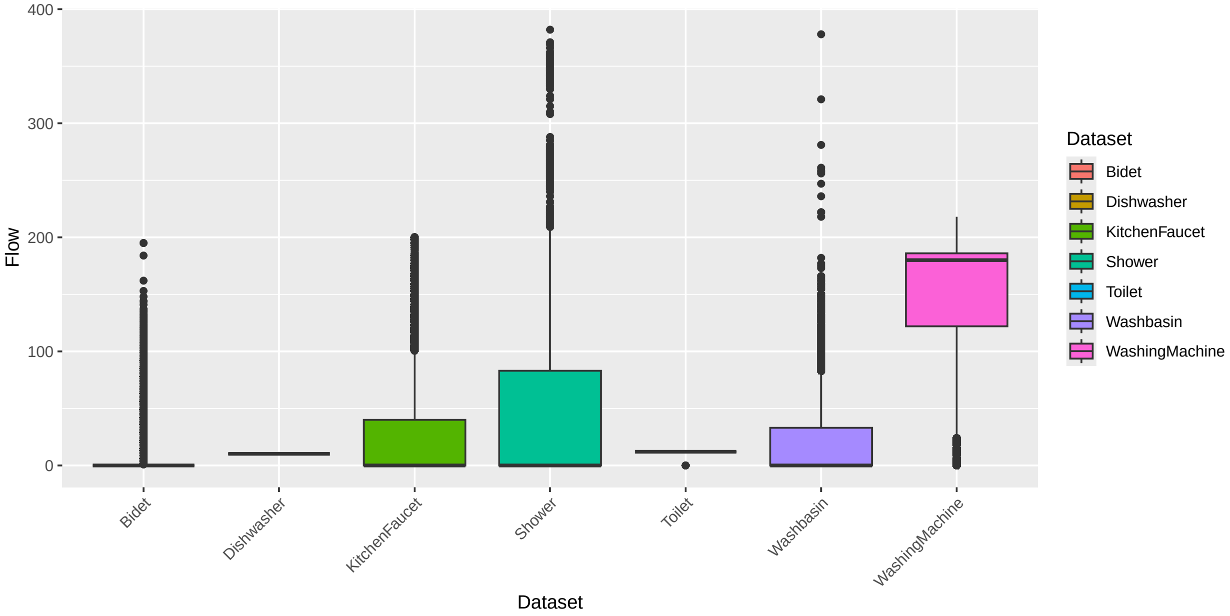


Figure 1: Boxplot of Flow for each Dataset.

In our analysis, we have chosen not to remove outliers from our time series data. This decision is rooted in the unique characteristics of time series analysis, where outliers can represent significant, real-world events or seasonal variations that are crucial for understanding the underlying patterns. Rather than viewing these outliers as anomalies to be discarded, we recognize their value in providing a comprehensive view of the data's behavior over time. Modern analytical techniques are well-equipped to accommodate such extremes, allowing us to extract meaningful insights without compromising the integrity of our analysis.

3. Data Preparation

During the data preparation phase, we focus on three essential steps: data selection, cleaning, and wrangling. These steps are crucial for creating a structured and reliable dataset that can be further used for modeling and evaluation. The ultimate objective is to construct a dataframe that contains the most useful information about water usage, facilitating insightful analysis and accurate modelling.

3.1 Data Selection and Cleaning

We start our project by loading individual datasets for different household appliances into separate DataFrames. Each DataFrame has the name of the appliance as a suffix:

- df_Bidet
- df_Kitchenfaucet
- df_Shower
- df_Washbasin
- df_Washingmachine
- df_AggregatedWholeHouse
- df_Dishwasher
- df_Toilet

These DataFrames will be the basis of our analysis, giving us detailed insights into water usage patterns for each household activity.

As part of our data understanding, the dataframes were checked for missing values. But we could see that there no missing NA values present in the dataset. This indicates that our dataset is clean and could be used for data wrangling steps.

3.2 Data Wrangling

After selecting and cleaning our data, the next crucial step is data wrangling or transformation. One fundamental transformation we perform is converting Unix timestamps into more readable datetime objects. This enables us to better understand the data.

To achieve this, we take each dataset and convert the Unix timestamps into datetime objects. We then store these datetime objects in separate columns, allowing for easier interpretation and analysis of date and time-related information.

In the data wrangling phase, we not only convert Unix timestamps but also address outliers and standardize dates across our datasets. During our analysis, we identified instances where some datasets contained dates from the 1970s, which is likely to be inconsistent with our expected timeframe. Finally, we are implementing feature engingeering to extract as much information we can from DateTime column.

To ensure consistency and reliability in our analysis, we apply a uniform date range filter to all datasets. We set the start date to September 1, 2019. Any dates falling outside this range are considered outliers and will be removed from the datasets.

Thus we do a filter on the datasets based on the condition below:

- **date >= start_date**

The filtered datasets which are having unifrom date ranges are used for further modelling and analysis. After applying the filtering to our datasets we have implemented feature engineering to extract some useful information out of the date and time. As part of our data transformation strategy, we identified the need for linear interpolation to address gaps in our data. This technique was particularly useful in datasets with time series information, where continuity and completeness of data are crucial for accurate analysis. Linear interpolation allowed us to estimate missing values based on linear relationships between existing data points, thereby ensuring our datasets were both comprehensive and representative of the real-world phenomena they aim to model. This is how the final datasets will look like :

```
##                DateTime Hour_of_Day Day_of_Week Month Season Flow
## 1 2019-02-13 11:18:14          11           4     2 Winter   90
## 2 2019-02-13 11:18:15          11           4     2 Winter  139
## 3 2019-02-13 11:18:16          11           4     2 Winter  102
## 4 2019-02-13 11:18:17          11           4     2 Winter   65
## 5 2019-02-13 11:18:18          11           4     2 Winter    2
```

The implementation of linear interpolation involved identifying instances where consecutive time stamps exhibited gaps indicative of missing data. By calculating the midpoints between existing data points, we were able to insert estimated values that maintained the integrity of the dataset's temporal progression. By integrating linear interpolation into our data wrangling phase, we enhanced the robustness of our datasets, making them more suitable for the sophisticated modeling and analysis that followed. This approach ensured that our data retained its temporal continuity, allowing for more accurate predictions and insights from our subsequent modeling efforts.

4. Modelling

Having meticulously prepared the data to meet our goals, we now move into the Exploratory Data Analysis (EDA) and Modelling stages of the CRISP-DM framework. This step is crucial for blending the data effectively with our success criteria. In the upcoming section, we will embark on a detailed exploration of the data to uncover insights that are critical to achieving our broader objectives and also to be able to move to the modelling phase of this cycle.

4.1 Exploratory Data Analysis (EDA)

In our Exploratory Data analysis, we inspected all eight datasets to gain some insights with the help of some plots.

We initially crafted a plot to check the summary of flows in individual appliances. We check the total flow and the mean flows for each appliance.

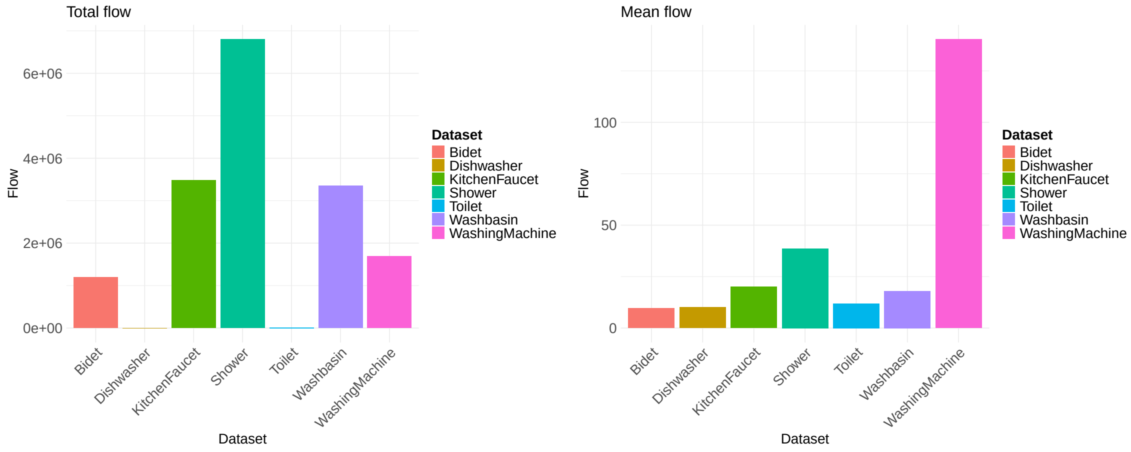


Figure 2: Flow summary of each dataset

We see from *Figure 2* that the total flow in datasets such as Shower, KitchenFaucet and Washbasin were significantly larger than in the other remaining datasets. So we can optimise the waterflow in these appliances to save the most water / energy. We also see that the mean flow for WashingMachine is much larger than the mean flow for other datasets. This can be attributed to the fact that a Washing machine cycle lasts much longer and hence the mean water usage is also greater than the average water used in other appliances.

We then created a plot to check the top 500 flow values for the entire date range across our datasets.

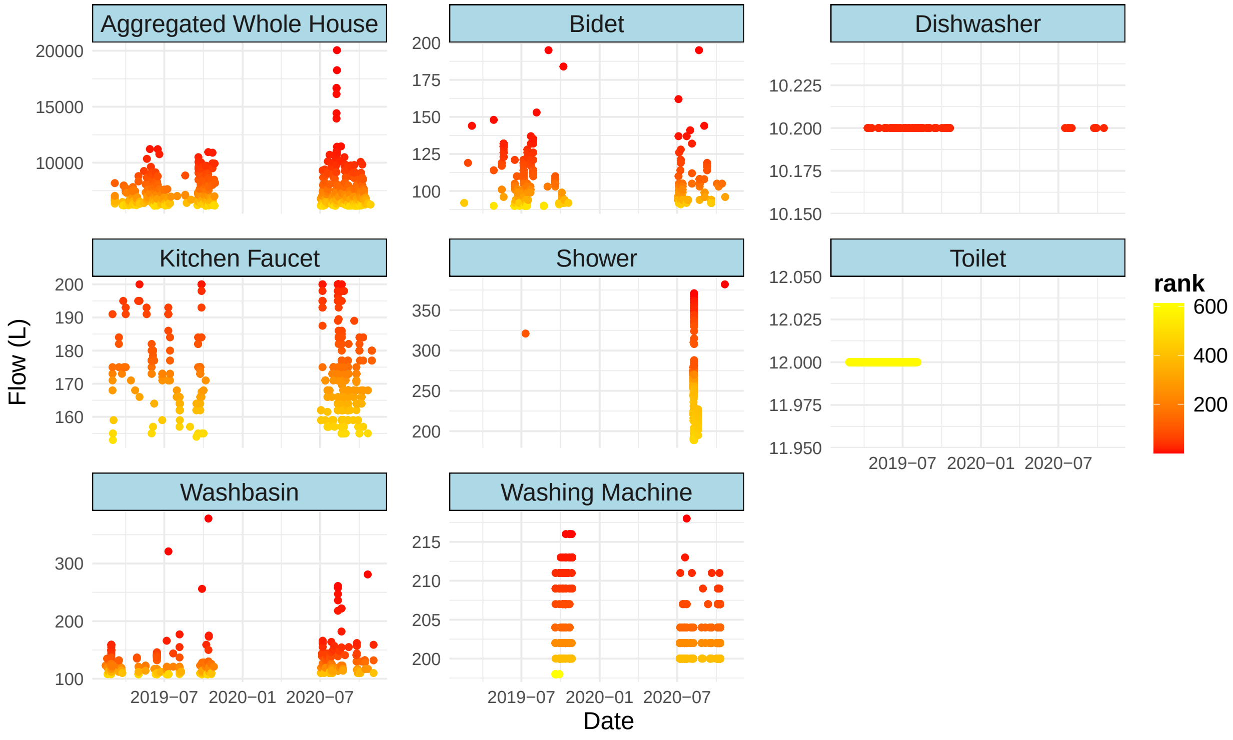


Figure 3: Top 500 Flows for each Dataset for the entire Date Range.

Our examination yielded noteworthy observations [*Figure 3*]. Across datasets such as aggregatedWholeHouse, Shower, and Washbasin, the month of August in 2020 emerged with the highest frequency of top flow values. Additionally, the flow values for Dishwasher and Toilet exhibited a consistent trend over the months. Interestingly, Bidet demonstrated a peak in top flow values during July. Kitchen Faucet and Washing Machine had pretty much a uniform spread of top flow values. Upon scrutinizing these findings, we see that the month of August 2020 is characterized by notable instances of elevated flow values, warranting further investigation to check the credibility of the pattern.

We plot the average flow values per month to check if there is any substantial claims that support our previous finding and further deduce seasonality in flow using actual seasons. The data is color-coded to represent different seasons.

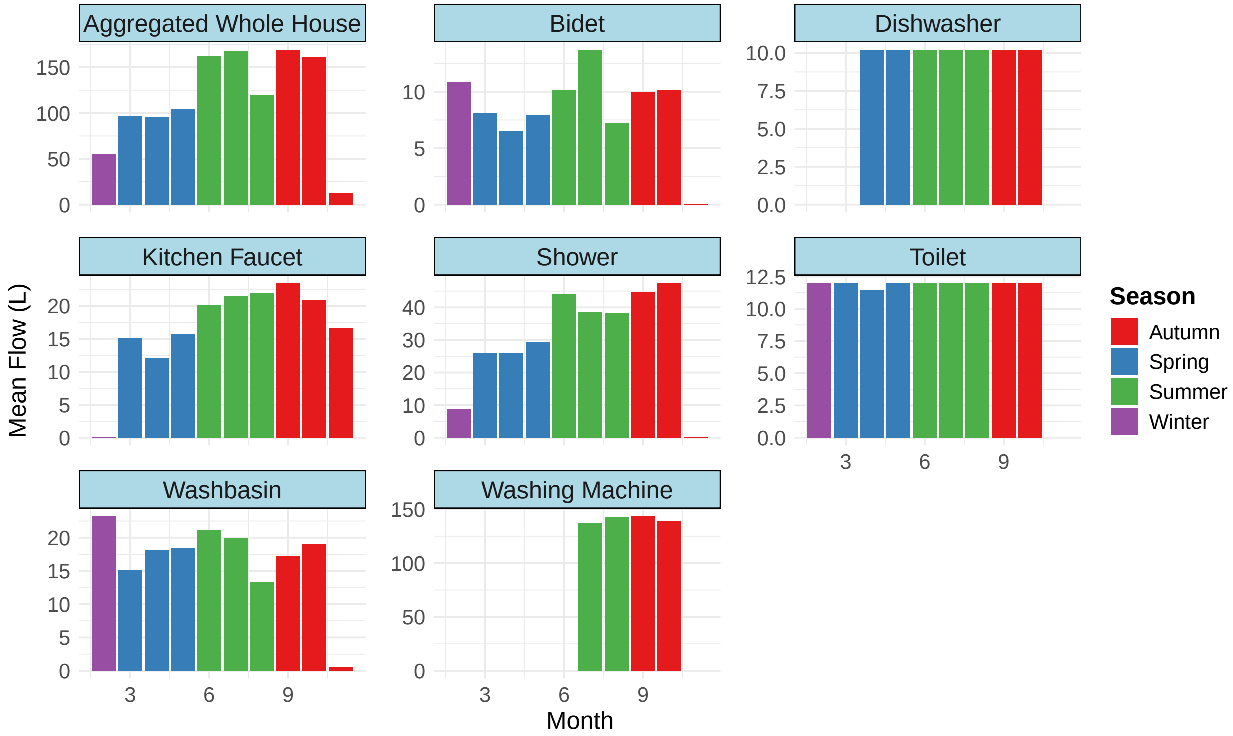


Figure 4: Mean Water Flow for Each Month.

From *Figure 4* we can observe that the mean water flow increases for the Summer and Autumn season as compared to Winter and Spring in most of the appliances. Also there is no clear evidence which would force us to further investigate the underlying reason for the huge consumption of water flow values in August 2020.

4.1.1 Which hour of the day is the most usage happening across different appliance types?

To answer our first business question, we created a plot that illustrates the mean water flow for each hour of the day across different appliance types in the house. It represents water usage data for different household appliances across various hours of the day, with the data segmented by seasons - Autumn, Spring, Summer, and Winter and without segmentation.

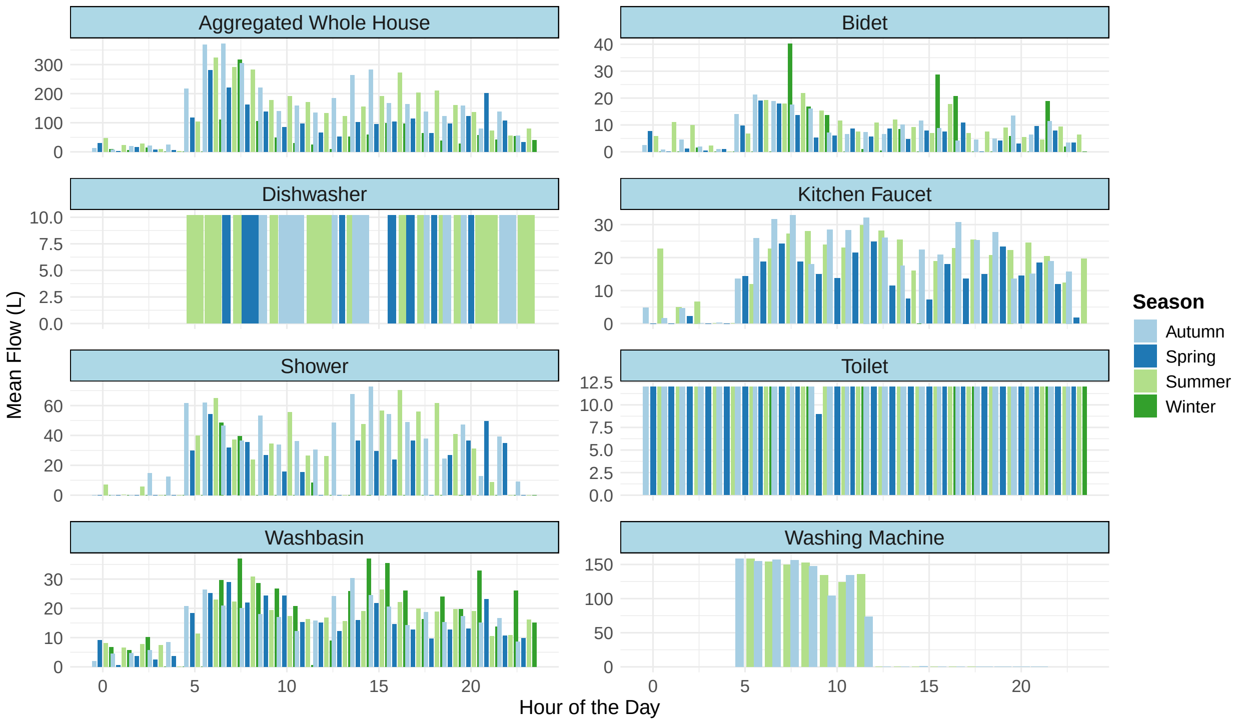


Figure 5: Mean Water Flow for Each Hour of the Day, segmented by seasons.

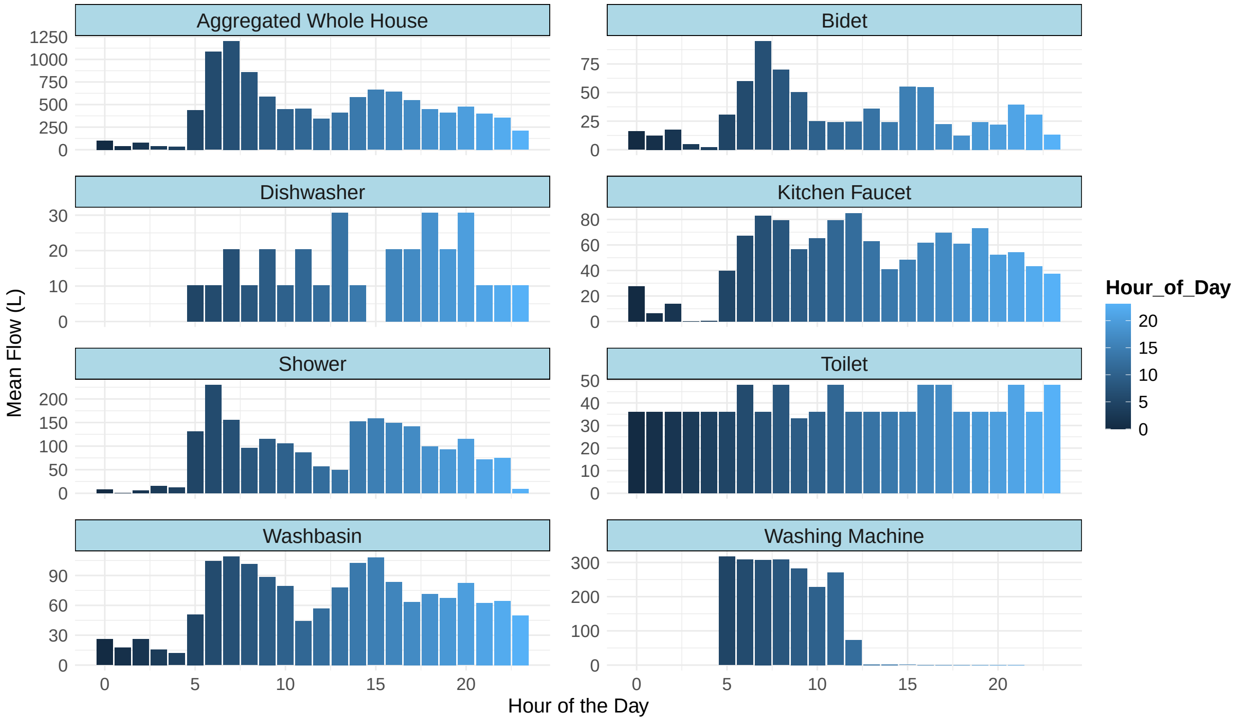


Figure 6: Mean Water Flow for Each Hour of the Day.

From *Figure 5* & *Figure 6*, we can infer the below observations:

- The majority of datasets, including Aggregated, Bidet, KitchenFaucet, Shower, and Washbasin, exhibit three peaks. These peaks occur during morning rush hours, typically between 6 AM to 8 AM, followed by a second peak in the afternoon, around 1 PM to 3 PM, and finally, another peak in the evening, approximately from 7 PM to 9 PM.
- For Aggregated Whole House, KitchenFaucet, and Shower data, usage spans across all waking hours, from 5 AM to 11 PM, with a decline during sleeping hours. Notably, the peaks during Summer and Autumn are more pronounced in the morning hours compared to those in Spring and Winter.
- Dishwasher usage reveals distinct peaks, indicating specific times when it is typically operated, while Toilet usage remains relatively consistent throughout the day.
- Washing Machine usage is predominantly concentrated during daytime hours, from 5 AM to 12 PM.

4.1.2 Which appliance is used the most in weekly bases?

To answer our second business question, we plotted graphs to display the average water usage for different appliances in the household across the days of the week, segmented by the seasons - Autumn, Spring, Summer, and Winter and without the segmentation.

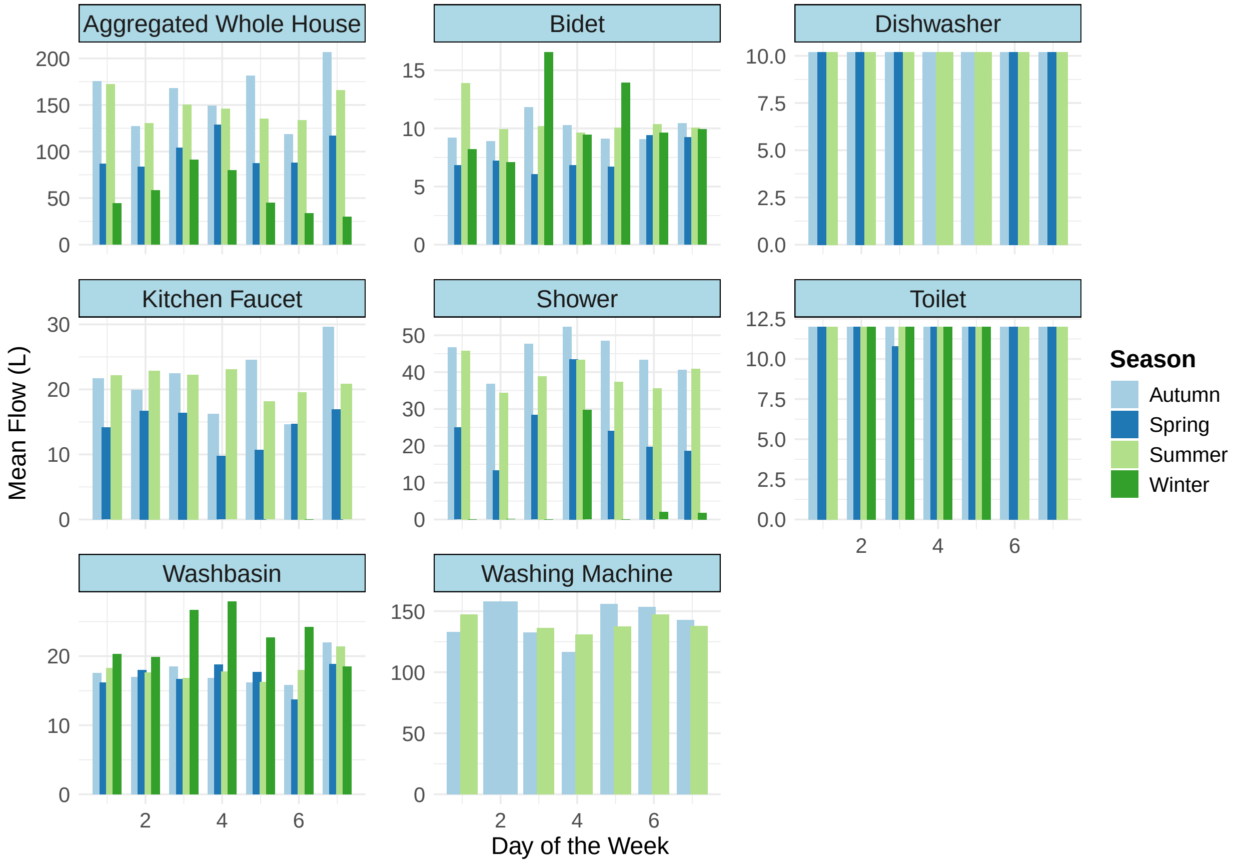


Figure 7: Mean Water Flow for Each Day of the Week, segmented by seasons.

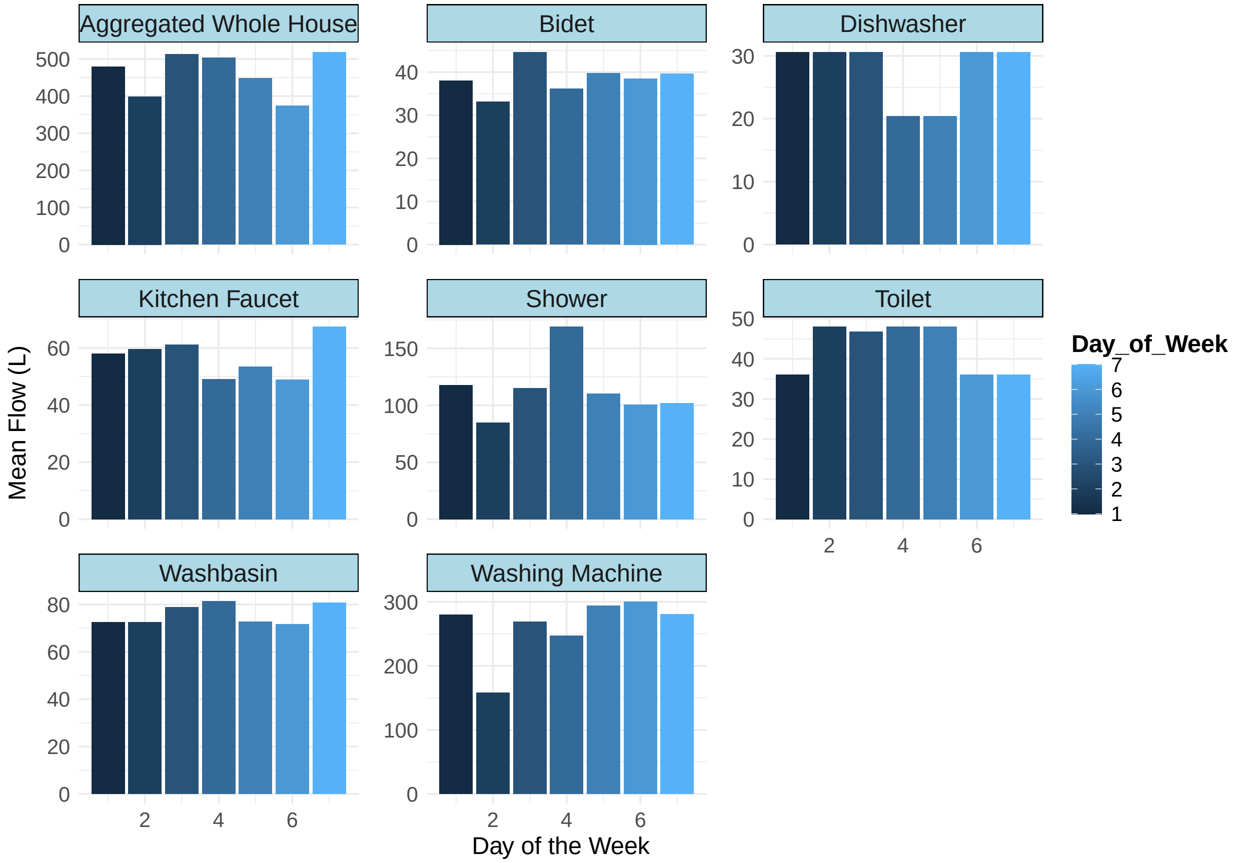


Figure 8: Mean Water Flow for Each Day of the Week.

From *Figure 7* & *Figure 8*, we can draw the below conclusions:

- As indicated by the plots in the preceding subsection, even when grouped weekly, Aggregated, KitchenFaucet, Shower, and WashingMachine exhibit higher flow rates during the Summer and Autumn seasons. Conversely, Bidet and Washbasin demonstrate increased daily flows during the Winter season.
- In the case of Aggregated data, notably higher mean consumption is observed during weekends, a trend that aligns with intuitive expectations. This pattern is similarly reflected in the KitchenFaucet and Washbasin datasets.
- Flow patterns for Dishwasher and Toilet remain relatively consistent throughout the week for each seasons. But mean usage for Toilet is more during weekdays for overall data.
- It is evident that Shower and WashingMachine exhibit the highest water usage throughout the week.

4.2 Models

In this phase of the CRISP-DM cycle, we are going to develop and evaluate two predictive models with the primary goal of understanding and forecasting water flow within a given context. More specifically, based on the previous EDA we saw that the most used appliances are: Kitchenfaucet, Shower,and Washbasin. With this in mind we are going to create these models to predict the flow of this appliances plus the flow of the whole house dataset. This predictive capability is crucial for a variety of applications, ranging from optimizing water usage to enhancing efficiency in water distribution systems.

Linear Regression Model

The first model we explored is a Linear Regression Model. Linear regression is a fundamental statistical approach that models the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables) by fitting a linear equation to observed data. The simplicity of the linear model makes it a good initial approach, as it provides a clear and interpretable model that can easily highlight significant predictors of water flow. Despite its simplicity, when the relationships between variables are indeed linear, this model can provide robust predictions and valuable insights into the factors influencing water flow.

AutoRegressive Integrated Moving Average (ARIMA)

Our second model utilizes the ARIMA (AutoRegressive Integrated Moving Average) algorithm, a classical approach in time series forecasting that combines autoregression, differencing, and moving average components. This model is specifically designed to capture the autocorrelation within time series data, making it highly effective for analyzing and predicting trends and cycles in data that follow a temporal sequence. ARIMA models are adept at handling data with trends, seasonal patterns, and other complexities inherent in time series, providing a robust tool for forecasting future values based on historical observations. Given its ability to model various aspects of temporal data, ARIMA stands as an invaluable choice for predicting water flow fluctuations over time, accommodating both seasonal variations and longer-term trends that are critical for effective water resource management.

Below are the datasets we have used for our modelling:

Aggregated Whole House

```
##              DateTime Hour_of_Day Day_of_Week Month Season Flow
## 1 2019-02-13 08:57:09           8           4     2 Winter    0
## 2 2019-02-13 08:59:32           8           4     2 Winter  177
## 3 2019-02-13 09:03:32           9           4     2 Winter    0
## 4 2019-02-13 09:08:32           9           4     2 Winter    0
## 5 2019-02-13 09:13:32           9           4     2 Winter    0
```

Kitchenfaucet

```
##              DateTime Hour_of_Day Day_of_Week Month Season Flow
## 1 2019-02-13 08:57:09           8           4     2 Winter    0
## 2 2019-02-13 08:59:32           8           4     2 Winter  177
## 3 2019-02-13 09:03:32           9           4     2 Winter    0
## 4 2019-02-13 09:08:32           9           4     2 Winter    0
## 5 2019-02-13 09:13:32           9           4     2 Winter    0
```

Shower

```
##              DateTime Hour_of_Day Day_of_Week Month Season Flow
## 1 2019-02-13 11:18:14          11           4     2 Winter   90
## 2 2019-02-13 11:18:15          11           4     2 Winter  139
## 3 2019-02-13 11:18:16          11           4     2 Winter  102
## 4 2019-02-13 11:18:17          11           4     2 Winter   65
## 5 2019-02-13 11:18:18          11           4     2 Winter    2
```

Washbasin

```
##              DateTime Hour_of_Day Day_of_Week Month Season Flow
## 1 2019-02-13 08:56:09           8           4     2 Winter    0
## 2 2019-02-13 08:58:31           8           4     2 Winter  123
## 3 2019-02-13 08:58:32           8           4     2 Winter   54
## 4 2019-02-16 16:55:26          16           7     2 Winter  135
## 5 2019-02-16 16:55:27          16           7     2 Winter   11
```

The primary purpose of these models is to predict the flow of water in a given household and aiming to:

- **Optimize Resource Allocation**: By accurately predicting water flow, we can optimize the allocation of water resources, reducing waste and ensuring that water is available where and when it's needed.
- **Inform Policy and Infrastructure Decisions**: Predictive insights can guide policy decisions and infrastructure investments, such as upgrades to plumbing systems or changes in water distribution strategies.
- **Enhance Sustainability**: Improved water flow predictions contribute to the sustainability of water systems, ensuring that water usage is efficient and that the environmental impact is minimized.
- **Emergency Preparedness**: Predictive models can help in anticipating water demand spikes or identifying potential issues before they become critical, contributing to better emergency preparedness and response strategies.

5. Evaluation

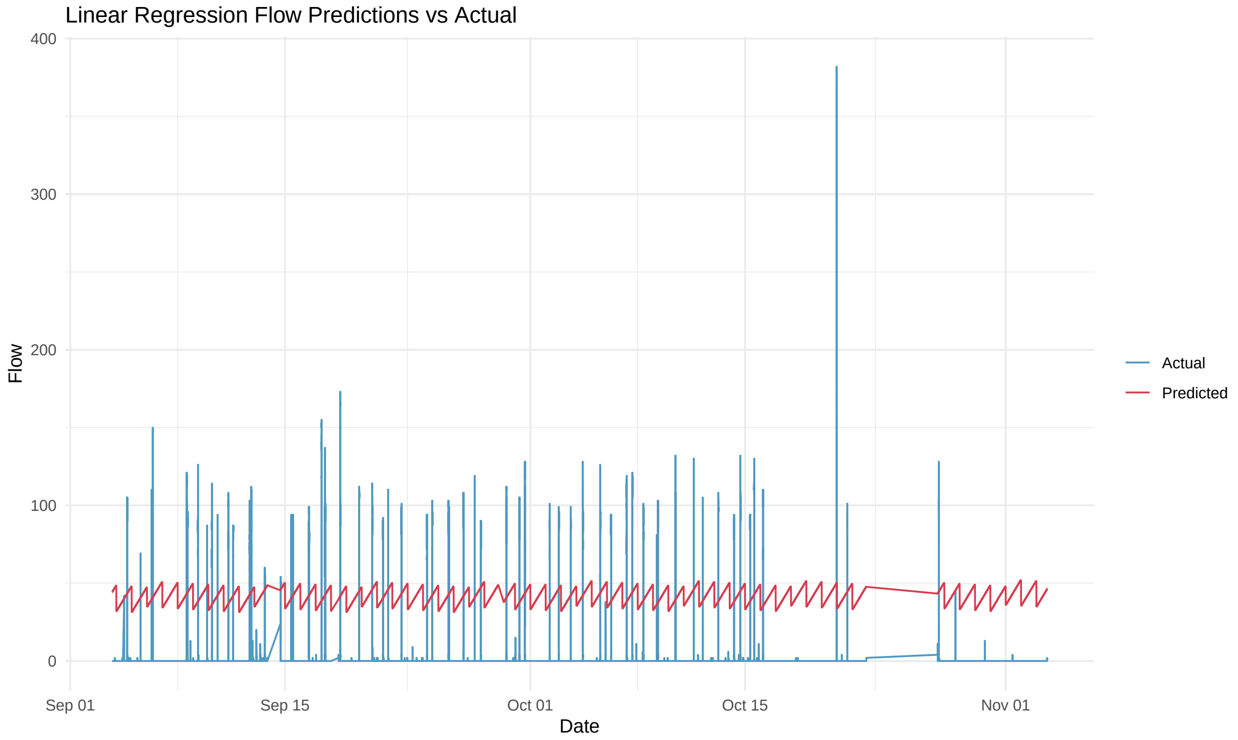
As we embark on the evaluation of our models, we delve into a meticulous comparison between the predicted outcomes and the actual data. This process is instrumental in revealing the accuracy, reliability, and overall effectiveness of our predictive tools. It's here that we employ various metrics, such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which serve as our guides in quantifying the performance of our models. These metrics, grounded in statistical rigor, offer us a clear lens through which we can view the predictive capabilities of our models, allowing us to make informed decisions about their application.

```
##
## --- Kitchenfaucet ---
## Linear Regression Model - MAE: 26.64913 , RMSE: 32.91554
## ARIMA Model - MAE: 26.09165 , RMSE: 33.70124
##
## --- Shower ---
## Linear Regression Model - MAE: 45.80695 , RMSE: 47.34741
## ARIMA Model - MAE: 48.70899 , RMSE: 66.3265
##
## --- Washbasin ---
## Linear Regression Model - MAE: 21.09371 , RMSE: 23.7045
## ARIMA Model - MAE: 16.85178 , RMSE: 25.31306
##
## --- AggregatedWholeHouse ---
## Linear Regression Model - MAE: 287.8537 , RMSE: 759.7015
## ARIMA Model - MAE: 312.7548 , RMSE: 760.496
```

Mean Absolute Error (MAE)

MAE measures the average magnitude of errors in a set of predictions, without considering their direction. It's the mean over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

- **Kitchenfaucet**: Both models show similar MAE scores (around 26), indicating that, on average, both models predict the flow with a deviation of about 26 units from the actual values. The performance is fairly comparable between the two models.
- **Shower**: The Linear Regression model slightly outperforms the ARIMA model, with a lower MAE of 45.80 compared to 48.70. This suggests that Linear Regression is somewhat more accurate on average for the Shower dataset.
- **Washbasin**: Linear Regression has a higher MAE (21.09) compared to the ARIMA model (16.85), indicating that ARIMA provides more accurate average predictions for the Washbasin dataset.
- **AggregatedWholeHouse**: Both models exhibit relatively high MAE values, with Linear Regression slightly outperforming ARIMA. This might indicate challenges in accurately modeling the aggregated flow, potentially due to increased variability or complexity in the data.

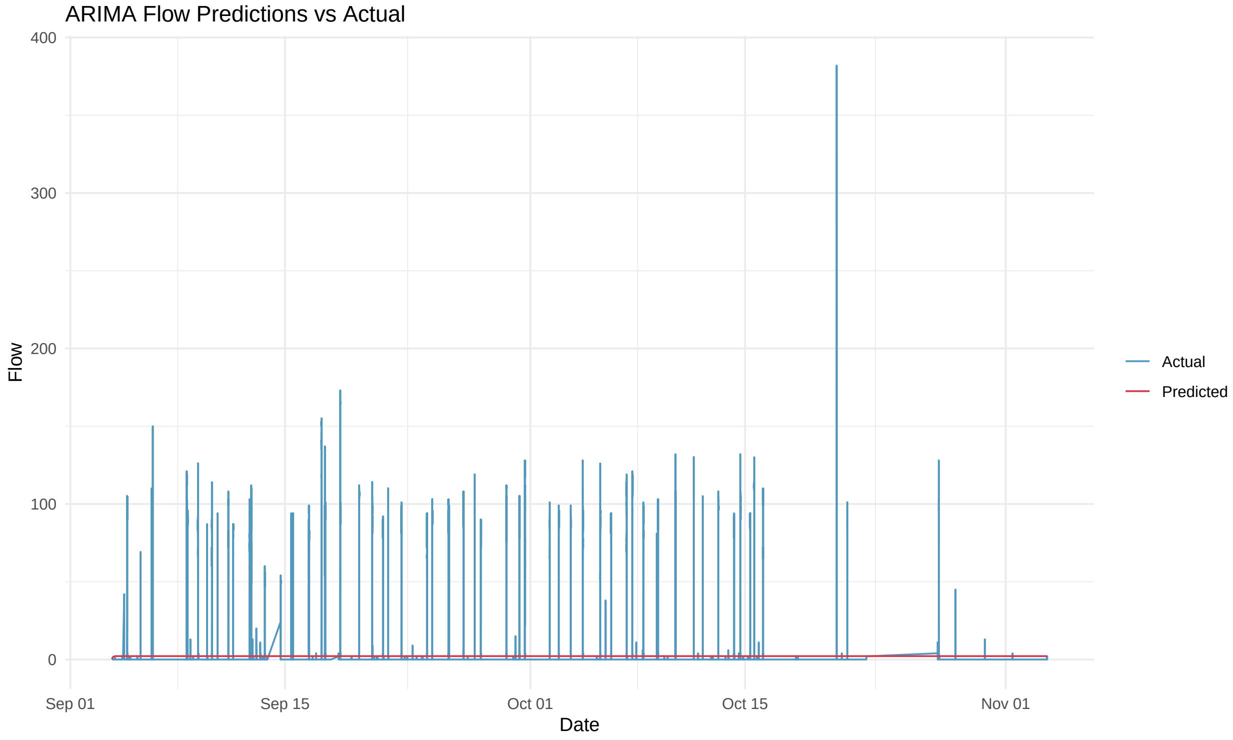


Root Mean Squared Error (RMSE)

RMSE measures the square root of the average of squared differences between prediction and actual observation. It gives a relatively high weight to large errors, meaning the RMSE is most useful when large errors are particularly undesirable.

- **Kitchenfaucet**: The RMSE scores are close for both models, with Linear Regression slightly edging out ARIMA. This indicates that both models similarly handle large errors in their predictions.
- **Shower**: Linear Regression shows a significantly lower RMSE (47.34) compared to ARIMA (66.32), suggesting that Linear Regression is better at minimizing large prediction errors for the Shower dataset.
- **Washbasin**: Here, the ARIMA model has a slightly higher RMSE than Linear Regression, suggesting that while ARIMA on average makes more accurate predictions (as indicated by MAE), it might occasionally make larger errors compared to Linear Regression.

- **AggregatedWholeHouse**: The RMSE values are quite close, indicating both models similarly handle large errors in predicting aggregated flow. However, the high RMSE values suggest that both models may struggle with accurately forecasting larger deviations in the aggregated dataset.



Technical Evaluation Summary

- **Consistency**: The Linear Regression model shows consistent performance across datasets, making it a reliable choice for different scenarios. However, its simplicity may limit its ability to capture complex patterns in certain datasets (e.g., AggregatedWholeHouse).
- **Model Suitability**: The ARIMA model's performance suggests it is well-suited for datasets with clear temporal patterns and seasonality (e.g., Washbasin), but it might be less effective for more complex or aggregated data.
- **Error Sensitivity**: Considering RMSE, Linear Regression tends to be less prone to large errors than ARIMA, making it a safer choice in applications where large prediction errors are particularly problematic.
- **Model Selection**: The choice between Linear Regression and ARIMA should consider the specific characteristics of the dataset and the application's requirements. For datasets with strong temporal patterns, ARIMA might offer advantages, while Linear Regression provides a solid baseline model that's easier to interpret and can be surprisingly effective across various scenarios.

6. Deployment

The deployment phase is a critical component of the CRISP-DM cycle, where the predictive models transition from development to real-world application. Successful deployment means integrating our Linear Model (LM) and AutoRegressive Integrated Moving Average (ARIMA) into the existing water management infrastructure to utilize their predictive capabilities for enhancing water flow efficiency and sustainability. This section outlines our approach to deploying these models and ensuring their insights deliver tangible benefits.

Integration into Existing Systems

The deployment process begins with the integration of our predictive models into the existing water management systems. This involves:

- **Data Pipeline Configuration**: Establishing automated data pipelines that feed real-time and historical water usage data into our models, ensuring that the predictions are based on the most current information.
- **API Development**: Creating APIs (Application Programming Interfaces) that allow the models to communicate seamlessly with water management software, enabling real-time predictions and analyses.
- **User Interface Integration**: Updating user interfaces of management software to display predictive insights and recommendations, making them accessible to decision-makers.

Operationalisation

Once integrated, the models enter the operationalisation phase, where they begin to influence decision-making processes:

- **Resource Allocation**: Utilizing model predictions to dynamically allocate water resources more efficiently, reducing waste, and addressing demand more accurately.
- **Policy Implementation**: Informing policy decisions and strategic planning with predictive insights, guiding investments in infrastructure improvements and conservation initiatives.
- **Alert Systems**: Incorporating model predictions into alert systems that notify administrators of potential issues, such as unusual patterns that could indicate leaks or impending shortages.

Monitoring and Maintenance

Deployment is not the final step; ongoing monitoring and maintenance are crucial to ensure the models remain accurate and relevant:

- **Performance Monitoring**: Continuously tracking the performance of deployed models against real-world outcomes, identifying any deviations or reductions in accuracy over time.
- **Model Updating**: Regularly retraining models with new data to adapt to changes in water usage patterns, ensuring that the models stay current and effective.
- **Feedback Loops**: Implementing feedback mechanisms that allow users to report discrepancies, feeding this information back into the development cycle for continuous improvement.

7. Conclusion and Future Work

Conclusion

In our business research, we explored the water consumption patterns of different household appliances, pinpointing when usage typically peaks. We also identified which appliances consume the most water, targeting these for optimization efforts to significantly reduce overall water use. Moreover, through predictive modeling, we gained valuable forecasts on future water consumption. This advancement has empowered us with the ability to better anticipate demand fluctuations, enabling us to optimise water consumption more effectively.

Future Work

As we look ahead, there's a clear path to improve and refine our predictive models, building on what we've achieved with our Linear Regression and ARIMA models. Starting with advanced feature engineering, we plan to dig deeper into our data to find more powerful signals that can make our predictions more accurate. This step will lead us to experiment with hybrid models that combine the strengths of both traditional and modern approaches, like mixing ARIMA's ability to handle time series with the pattern recognition power of machine learning algorithms such as Random Forests and Gradient Boosting Machines. This combination could provide a more comprehensive understanding of our data. To ensure our models perform well across different times and conditions, we're looking into better cross-validation techniques that are specially designed for time series data. This way, we can be more confident that our models will be reliable. We're also excited to explore more advanced models, including machine learning and deep learning, which have the potential to uncover complex patterns our current models might miss. Of course, finding the best settings for these models through tuning and optimization will be key to maximizing their performance.

Another important area we're focusing on is expanding our datasets, either by adding more data or creating synthetic data. This could help overcome challenges like data sparsity or noise and improve our models' accuracy. But advancing our models isn't just a technical task; it's also about working closely with stakeholders to ensure the models meet real-world needs. By continually getting feedback and adjusting our approach, we aim to make our models not only more accurate but also more useful for practical decisions.

8. Dashboard

Power BI Dashboard: Link (Ctrl + Click)