Group 1 Report

March 2024

Introduction

This report provides an in-depth analysis of a collaborative initiative focusing on the FairWater project, involving Northumbrian Water and Newcastle University, with the aim of advancing water efficiency in domestic environments. Employing the CRISP-DM methodology, the investigation encompasses three complete cycles, thoroughly addressing the stages of business and data understanding, data preparation, modeling, evaluation, and deployment. The objective of this study is to uncover patterns in water consumption and identify opportunities for technological innovations that can significantly improve water efficiency. In doing so, it aims to offer actionable insights and recommendations to stakeholders involved in the project. These findings are intended to guide the development of solutions that not only reduce water consumption and lower utility bills for households, especially those that are low-income, elderly, or vulnerable, but also contribute to the overall sustainability of water resources. Ultimately, the goal is to enhance the project's impact on society by fostering more efficient water use and encouraging positive behavior changes among consumers.

Cycle 1

1.Business Understanding - Water Consumption

Business objectives

Our stakeholders in the FairWater project want to understand the water usage across a single household in Naples to have some useful insights to reduce the water consumption of households.

What Investigations Are We Carrying Out?

To achieve our business objective we will investigate the following:

- In our first cycle, we aim to analyze the water consumption of the whole single household.
- Moving into the second cycle, our focus shifts towards investigating the monthly, weekly, and hourly water usage of each appliance within the household, aiming to discern patterns across these time frames.
- Finally, in the third cycle, our objective is to predict the forthcoming daily water flow of the household.

2.Data Understanding

In this phase, we focus on evaluating the data requirements in relation to the report's objectives, data sources and the reliability of the data. This step is of paramount importance in increasing the likelihood of achieving the project's goals and generating valuable results. Once the initial dataset is collected, the following step is conducting a thorough review of the business objectives to ensure that they are realistically aligned with the available data resources. This alignment is crucial for the successful execution of the project.

Initial Data Collection

This report's dataset originates from the Water End USE Dataset and TOols (WEUSEDTO) in the github at: https://github.com/Water-End-Use-Dataset-Tools/WEUSEDTO, which collects 1 year of monitoring between 2019 and 2020 (March to November 2019 and July to October 2020). The dataset records 7 water appliances' flow time series, With 1 year's water flow included in every appliance. Each appliance's flow dataset was available in CSV format.

Data Description

After collecting initial dataset, we need to explore and give a description for our data which we import to this project. In the data directory, Water End USE Dataset contains 7 CSV format files which storing water flow time series in different appliance.

File Name	File Type	Description
aggregatedWholeHouse	CSV	Water flow end use time series in the whole house
feedBidet	CSV	Water flow end use time series in bidet
feedDishwasher	CSV	Water flow end use time series in dishwasher
feedKitchenfaucet	CSV	Water flow end use time series in kitchen faucet
feedWashingmachine	CSV	Water flow end use time series in washingmachine
feedShower	CSV	Water flow end use time series in shower
feedToilet	CSV	Water flow start and end use time series in toilet
feedWashbasin	CSV	Water flow end use time series in washbasin

Dataset Summary for each water appliance csv file. These being:

Table 2: Dataset Summary

FileName	Rows_num	Columns_num	NAs
aggregatedWholeHouse.csv	166082	2	0
feedBidet.csv	121993	2	0
feedDishwasher.csv	53	3	0
feedKitchenfaucet.csv	167065	2	0
feedShower.csv	170181	2	0
feedToilet.csv	1188	3	0
feedWashbasin.csv	179933	2	0
feedWashingmachine.csv	12055	2	0

Column name for each water appliance csv file. These being:

Column name	Type	Description
Uinx	int	The timestamp point at which the flow of water begins
flow	num	The volumn of Water flow at each recording point
Endflow	int	The timestamp point at which the flow of water ends

In examining each of the CSV data files, there were several issues that arose during the import process that required attention. A thorough review of these files revealed that there were a lot of 0 values in the flow column in all the csv datasets, which may have some impact on the subsequent time series prediction results, and the number of columns varied from one csv file to another, with only the presence of the Endflow column in feedToilet.csv and feedDishwasher.csv It may be necessary to pay attention to the subsequent processing of the data. In addition, the time series are recorded in timestamp format, and the timestamps need to be converted to datetime format for further analysis of the time series, and appropriate preprocessing steps should be taken to address these issues.

3.Data Preparation

Based on the aim of the first CRISP-DM, we will process the *aggregatedWholeHouse.csv* file, which includes the following tasks:

- Check if the dataset's sampling rules are consistent with the actual ones.
- Convert the data types of the dataset fields.
- Data quality assessment

3.1 Sampling rules

According to the dataset introduction provided by the coursework, the data in this file is calculated according to the following rules:

- When water usage is continuous, the water volume is recorded once every minute.
- If no water usage is detected, then the water volume is recorded once every five minutes.
- The volumes of seven types of water usage (Washbasin, Bidet, Kitchen Faucet, Shower, Washing Machine, Dishwasher, Toilet) are combined to obtain the total water volume at a certain time.
- Missing data filled in through interpolation.

To verify whether the dataset's sampling method is consistent with the standard rules, it is necessary first to convert the *unix* column into *datetime* type. Then, calculate the time interval for each sampling. The following results can be obtained through calculation:

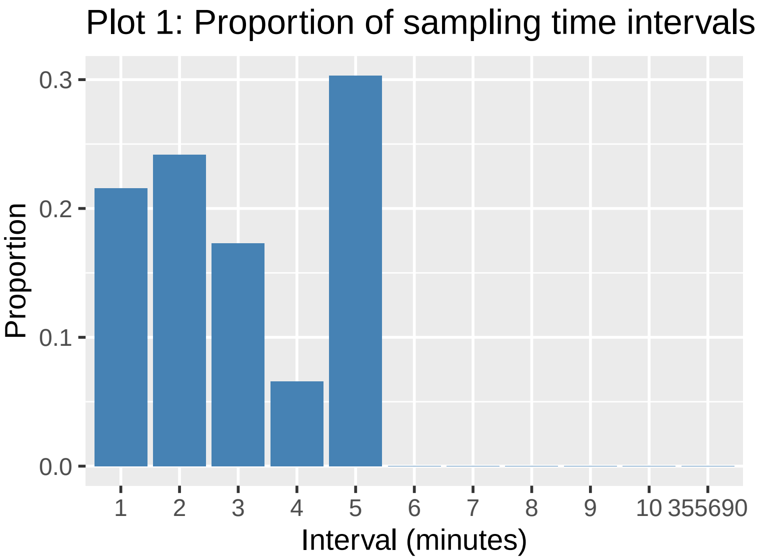


Figure 1: Proportion of sampling intervals

From the above figure, it can be seen that:

- About 30% of the water usage records in the dataset have a sampling interval of 5 minutes, indicating that the user's water usage is not frequent.
- The sampling time intervals are mainly distributed between 1-5 minutes, which conforms to the standard sampling strategy.
- Sampling intervals greater than 5 minutes do not comply with the standard sampling strategy and are considered outliers, which will be further analyzed.

3.2 Data quality assessment

Further, we analyze the data with sampling intervals greater than 5 minutes:

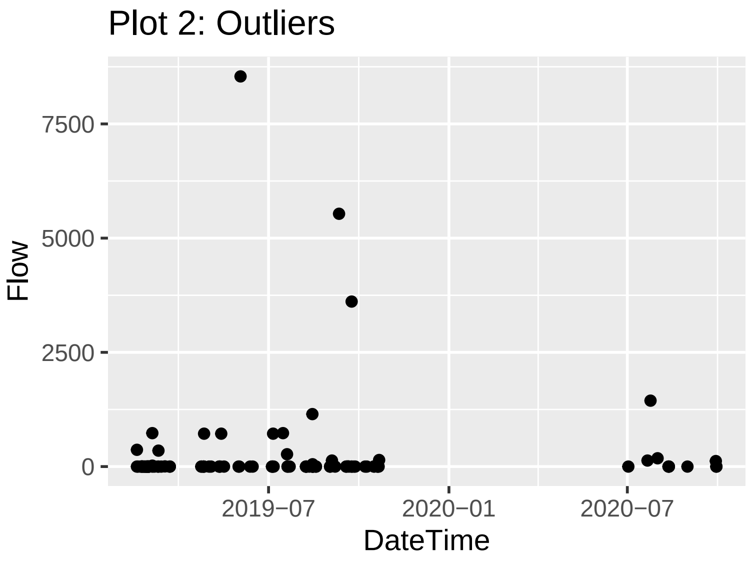


Figure 2: Outliers in the aggregated household

Based on the analysis above, we can observe that:

- The outliers are evenly distributed across two time periods, divided into from 2019-02-13 08:57:09 to 2019-10-29 07:53:06, and from 2020-07-20 08:03:00 to 2020-11-03 17:21:21.
- The records with a sampling interval of 355,690 minutes are due to the existence of these two periods, which are considered normal records.

Therefore, we need to divide the *aggregatedWholeHouse* dataset into two segments. Since the first segment covers most months of 2019, containing more temporal features, it is more suitable for building the model.

Regarding outliers, since the reasons causing these data are unknown and their quantity is small, and considering that in the third cycle a time series model is required for predicting water usage, it's essential to ensure the continuity of the time series data. Therefore, outliers will not be processed. And attempts will be made later to complete the prediction tasks using models that are less affected by outliers.

Table 4: First segment of clean aggregated whole house time series data

unix	flow	datetime	time_diff
1550048229	0	2019-02-13 08:57:09	NA
1550048372	177	2019-02-13 08:59:32	2
1550048612	0	2019-02-13 09:03:32	4
1550048912	0	2019-02-13 09:08:32	5
1550049212	0	2019-02-13 09:13:32	5
1550049512	0	2019-02-13 09:18:32	5

4.Modeling - identifying periodic patterns

In identifying data patterns, we attempted to explore the data both **qualitatively** and **quantitatively** based on **multiple time features**. In the qualitative analysis, we conducted visual Exploratory Data Analysis (EDA) to try to summarize the overall water usage situation of households, explore periodic patterns, and the stability of these patterns. In the quantitative analysis, we performed ACF and PACF analyses on the time series to precisely identify stable periodic patterns. This approach not only provides high-quality time series data for the modeling task in the third cycle but also aids in model selection and parameter setting.

4.1 Household water usage

knitr::kable(head(raw_sink)) The description of the current state of overall household water usage primarily unfolds from the following aspects:

- Household water usage at different hours of the day
- Household water usage on different days of the week
- Monthly household water usage trends
- Quarterly household water usage trends

The analysis of water usage during different hours of the day reveals:

- The daily average water usage peaks at 6 AM, with the lowest usage occurring in the early hours of the morning.
- The average water usage throughout the day exhibits a bimodal distribution, with peaks at 6 AM and 8 PM.

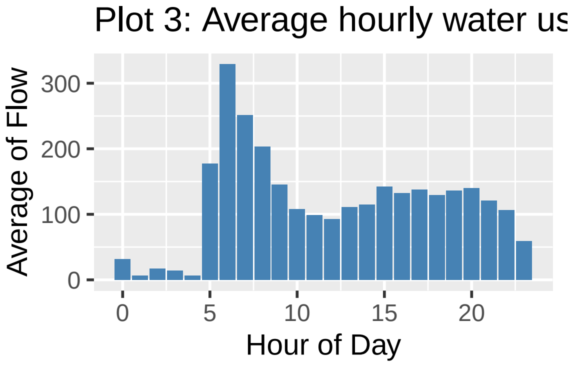


Figure 3: Average hourly water usage throughout the day

The water usage on different days of the week within a household is as follows:

- Household water usage is higher on weekends.
- Household water usage decreases progressively from Tuesday to Friday, with Tuesday's average water usage significantly higher than Monday's.

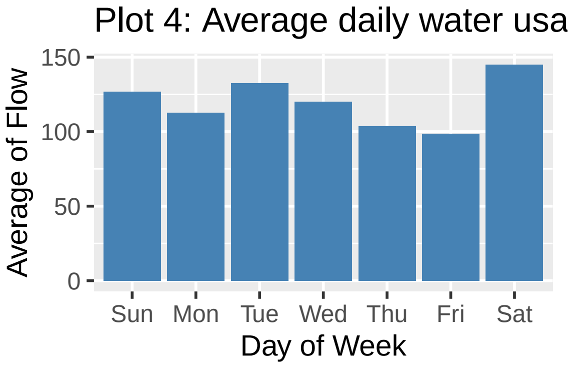


Figure 4: Average daily water usage throughout the week

The trend in average monthly household water usage is as follows:

- The average water usage reaches its maximum in October, while a decrease is observed in August.
- The monthly average water usage shows a fluctuating trend, necessitating further investigation to confirm if it is related to seasonal changes.

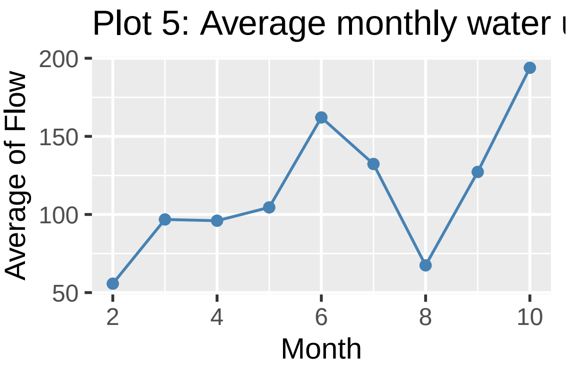


Figure 5: Average monthly water usage

The analysis of the average household water usage by quarter reveals the following:

- The average water usage fluctuates in the first three quarters, but the changes are relatively minor.
- The average water usage increases rapidly in the fourth quarter.
- Given that there is only one year of data available, it is temporarily uncertain whether there is a seasonal variation.

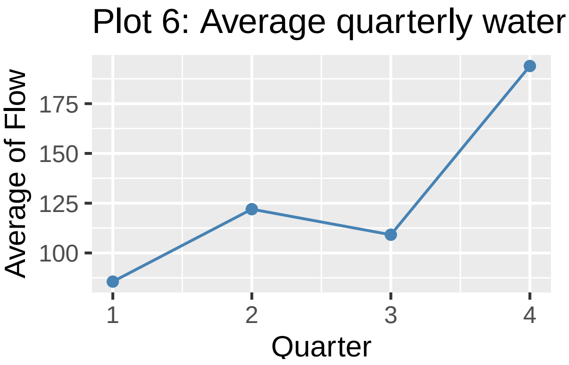


Figure 6: Average quarterly water usage

4.2 Periodic patterns and stability

After analyzing the current state of overall household water consumption, we have observed that there might be periodic changes across some time feature dimensions. Next, we will focus on exploring the periodic changes in water usage based on different time features. When identifying periodic patterns in time series data, we followed these steps sequentially:

- Selection of time features
- Identification of periodic patterns
- Assessment of pattern stability

In the dataset, new time-related features are added, including *year*, *quarter*, *month*, *week*, *day of the week*, and *hour*. We plotted the time series graph of water usage and the first-order difference time series graph. Based on the hourly and daily time series graphs, similar water usage characteristics were observed:

- Water usage was more frequent and the volume was higher from February to March 2019.
- From April to August 2019, the data showed periodic patterns with increasing instability.
- Extreme values of water usage appeared from September to October 2019, making the series even more unstable.
- The hourly time series was more unstable than the daily time series.(Please refer to the appendix)

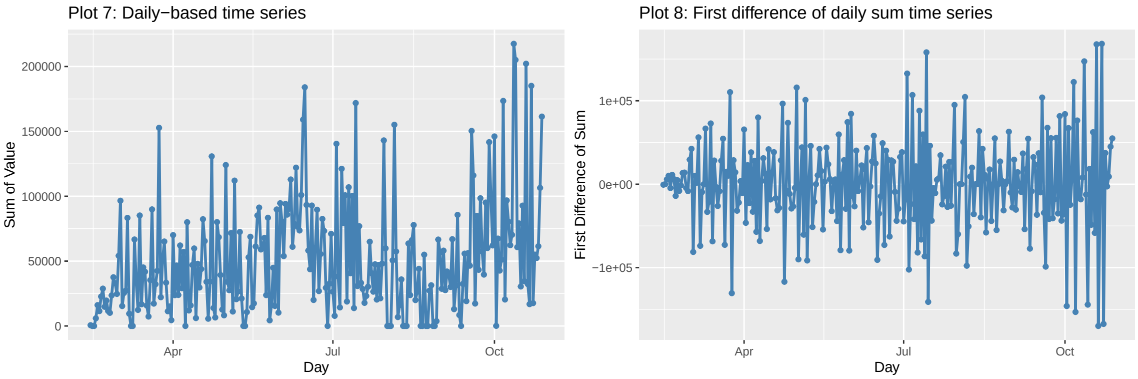


Figure 7: a) Daily-based time series. b) First difference of daily sum time series

Based on the above analysis, it is clear that household water usage is closely related to the hour and day, exhibiting similar patterns. Therefore, selecting the day as a feature not only avoids the influence of extreme values in the hourly series but also eliminates the impact of random events.

Next, we will continue to explore whether the dataset exhibits periodic patterns related to the day of the week, month, and quarter (for the monthly and quarterly time series graphs, please see the appendix).

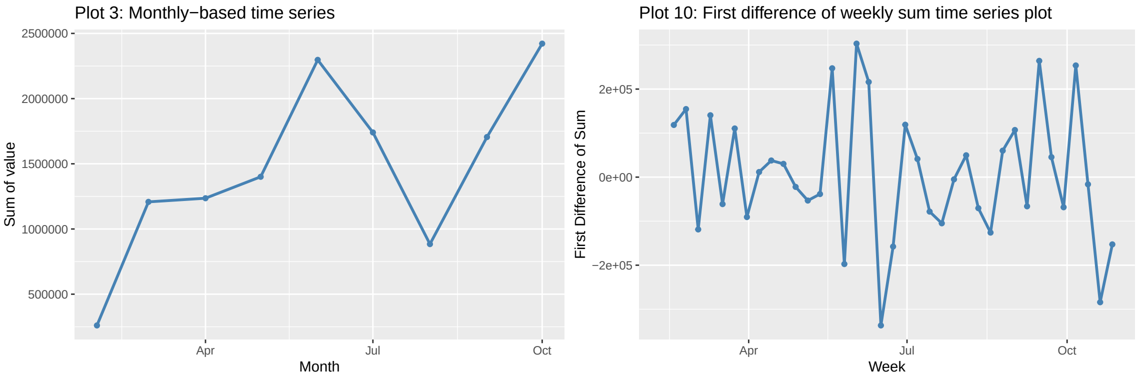


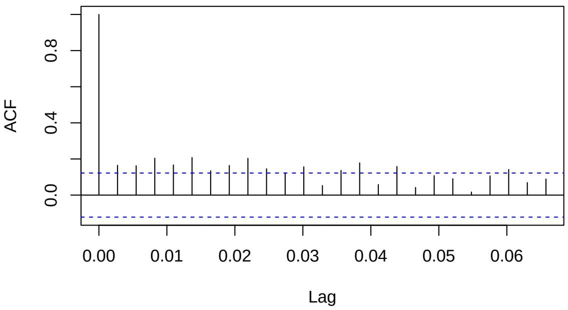
Figure 8: a) Monthly-based time series. b) First difference of weekly sum time series plot

knitr::kable(head(raw_dishwasher)) Based on the observation of the plots, the time series based on months and quarters, due to their smaller data spans, do not definitively confirm the existence of periodic patterns, but a certain trend can be observed:

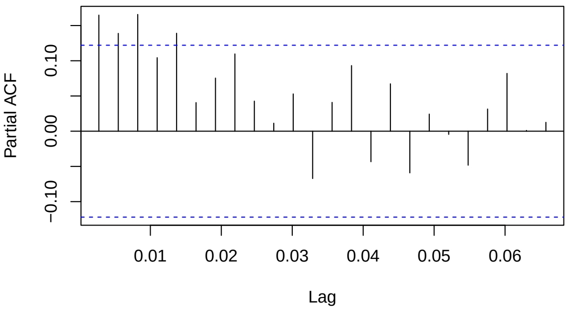
- The time series data based on months shows some periodic tendencies. For example, the change in water usage presents a V-shaped pattern.
- The time series data based on quarters shows that water usage gradually increases from the first quarter, peaking in the third quarter. The rapid decline in water usage is due to the fourth quarter's data being incomplete and thus not accounted for.
- Based on the time series data for days of the week, there is an overall presence of periodic patterns, and it is relatively stable.

Through visual data exploration, we observed that the dataset exhibits periodic changes in hourly, daily, and weekly water usage, while the variations in monthly and quarterly water usage are less pronounced. Next, we will attempt a quantitative exploration of the periodic and stable time series data using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF).

Plot 11: ACF for daily data



Plot 12: PACF for Daily Data



From the observation of the above results, we can find (for the ACF and PACF graphs based on hourly and weekly data, please refer to the appendix):

- The hourly time series lacks significant autocorrelation and there is no strong evidence to support an AR component.
- The daily time series does not exhibit significant autocorrelation. However, it shows significant autocorrelation at the first lag, indicating that an **AR(1)** model might be suitable for this time series.
- The weekly time series lacks significant autocorrelation and there is no strong evidence to support an AR component.

5.Evaluation

The analysis leads to the following conclusions:

- Data from the time period 2019-02-13 08:57:09 to 2019-10-29 07:53:06 is selected for the time series analysis.
- The dataset contains outliers, recommending the use of models that are insensitive to outliers.
- The time series data of the dataset is relatively unstable, suggesting the use of models suitable for non-stationary time series for forecasting tasks.
- The daily total water usage of households exhibits stable periodic patterns, showing significant autocorrelation at the first lag, making it suitable for AR(1) modeling.

Cycle 2

Business Understanding - Water Consumption

The primary objective of the second cycle of our report is to look further into understanding the water consumption patterns within a household in Naples. Building upon the insights gained from the aggregated data analysis in the first cycle, our focus now shifts towards a detailed examination of individual appliances' water usage.

By dissecting the water flow data from washing machines, showers, dish washers and other household fixtures, we aim to uncover insights into usage patterns, identify potential inefficiencies and possibly pinpoint opportunities for optimisation. This phase of analysis is can be crucial in helping to mitigate the climate change crisis.

Our approach involves a comprehensive assessment of the data to extract meaningful insights and trends. We will explore various factors such as time of day, day of the week, and hourly variations. Additionally, we will investigate any anomalies or irregularities in the data that may indicate leaks, malfunctions or other issues requiring attention.

Data Understanding

Since we covered gained a rigorous understanding of the data in the previous cycle, there is no need to recover this however, we will take a look at the toilet water usage.

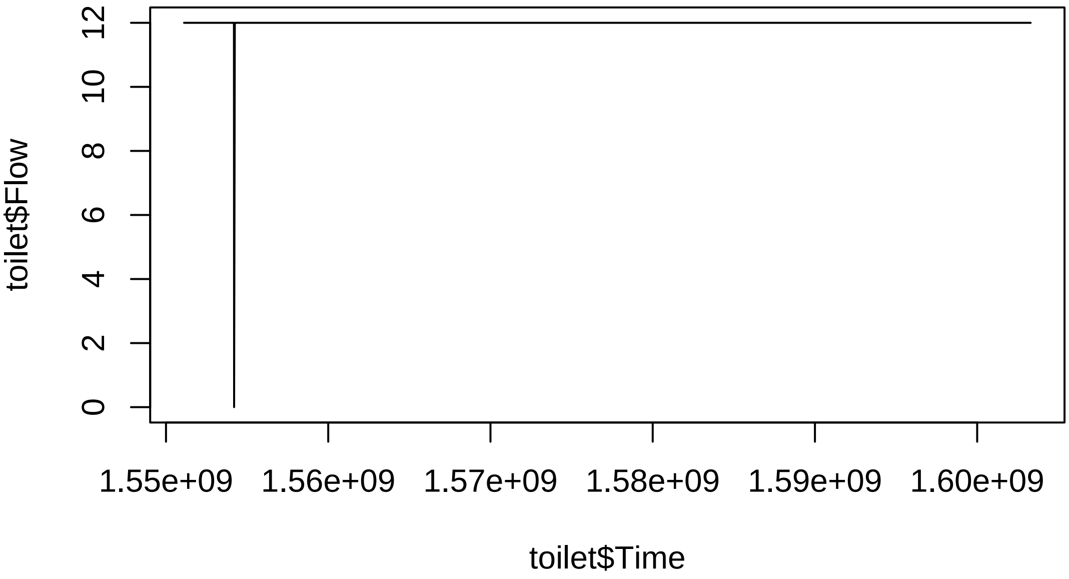


Figure 9: Toilet water usage

From this we can see that there is not much to analyse, therefore we shall not be exploring this further.

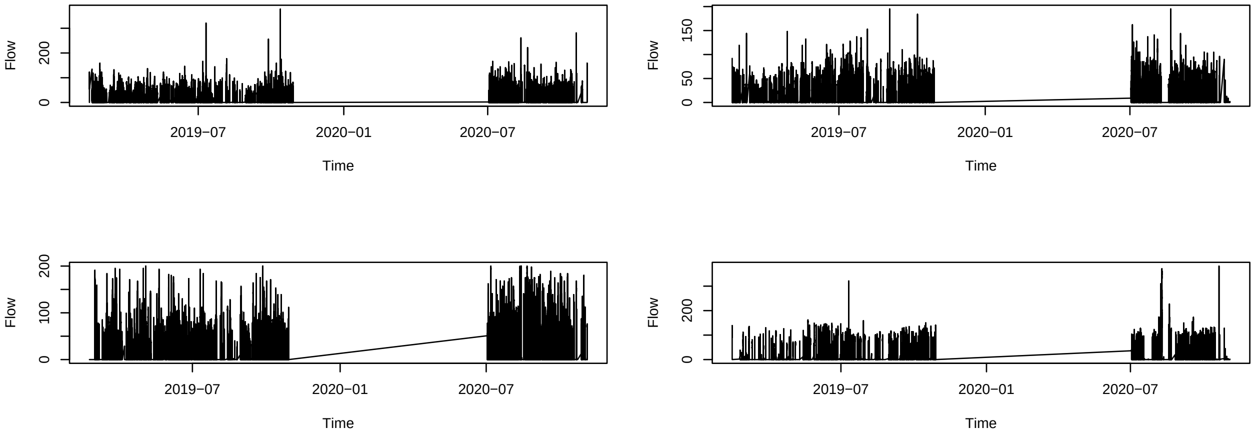


Figure 10: Outliers for a) Sink b) Bidet c) Kitchen Faucet d) Shower

It was difficult to find outliers since the dataset for each appliance was so sparse so we thought a visual check would be appropriate and upon inspection it seems that there are no obvious issues.

Data Preparation

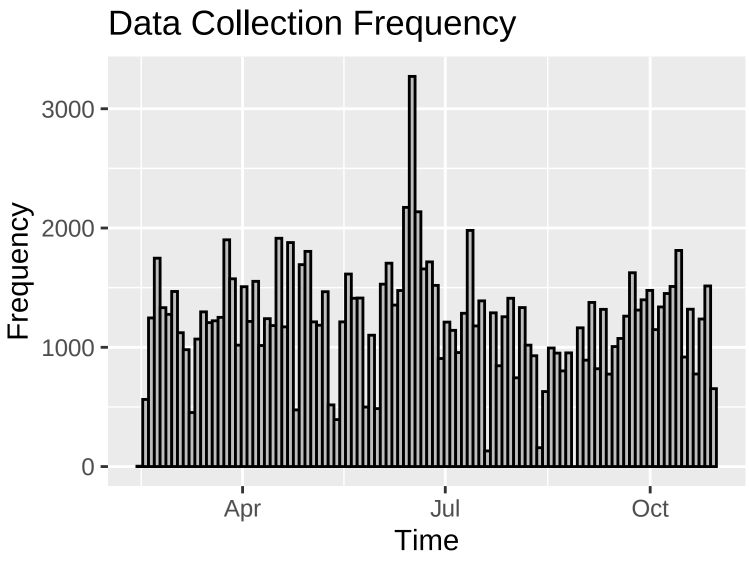
As mentioned previously, the datasets were very similar and hence we will perform virtually the same transformations to get the desired features. We first converted to datetime, added additional features based on the Time variable to use in the modelling section and removed the outliers.

Time	Flow	Month	Week	Day	Hour	Datetime	Year	Weekday
2019-02-13 08:56:09	0	Feb	7	Wednesday	8	2019-02-13 08:56:09	2019	Wed
2019-02-13 08:58:31	123	Feb	7	Wednesday	8	2019-02-13 08:58:31	2019	Wed
2019-02-13 08:58:32	54	Feb	7	Wednesday	8	2019-02-13 08:58:32	2019	Wed
2019-02-16 16:55:26	135	Feb	7	Saturday	16	2019-02-16 16:55:26	2019	Sat
2019-02-16 16:55:27	11	Feb	7	Saturday	16	2019-02-16 16:55:27	2019	Sat
2019-02-16 16:55:34	6	Feb	7	Saturday	16	2019-02-16 16:55:34	2019	Sat

Modelling & Evaluation - Visualising Water Consumption

Visualising water usage.

Before we visualise the data it is important to look at how much data is being collected each day and whether or not it is consistent.



We will now look at the individual appliances, splitting the data into two parts, pre 2020 and post 2020 since the data is missing about 8 months.

Washbasin Usage

Summed Pre 2020 flow:

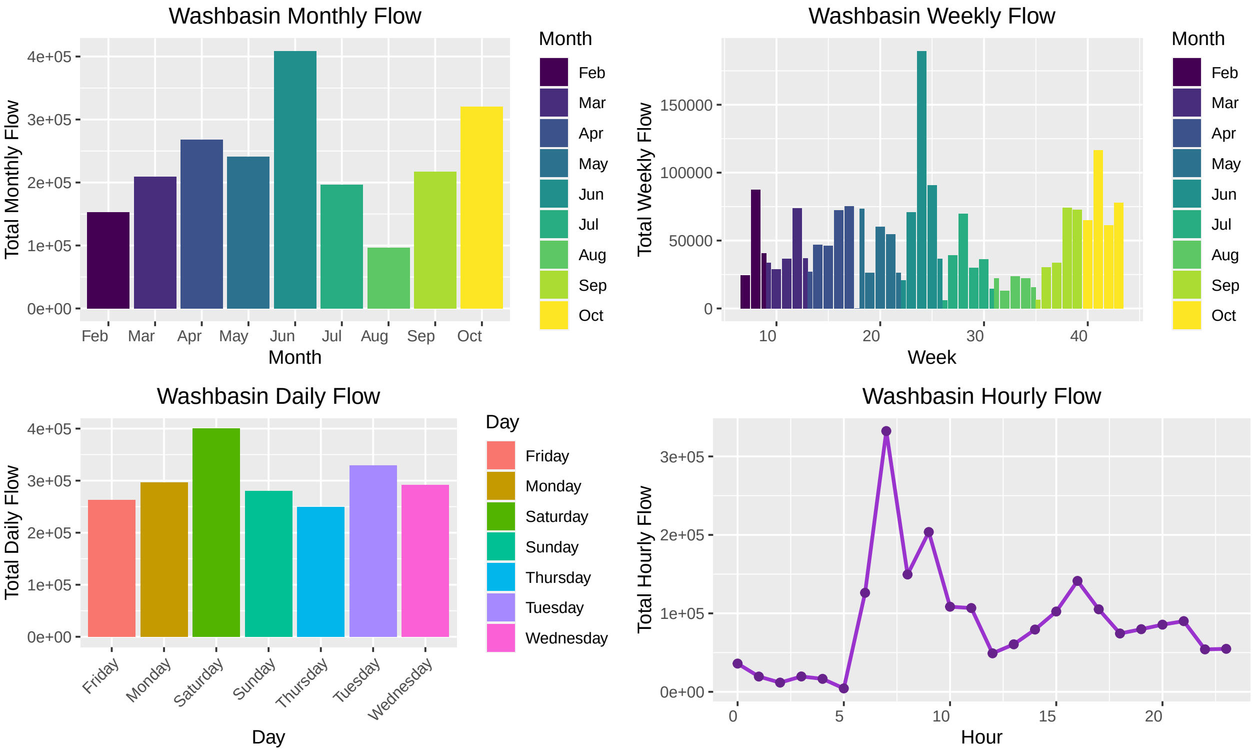


Figure 11: Monthly, Weekly, Daily and Hourly flows for Sink appliance (2019)

Summed Pre 2020 flow (complex):

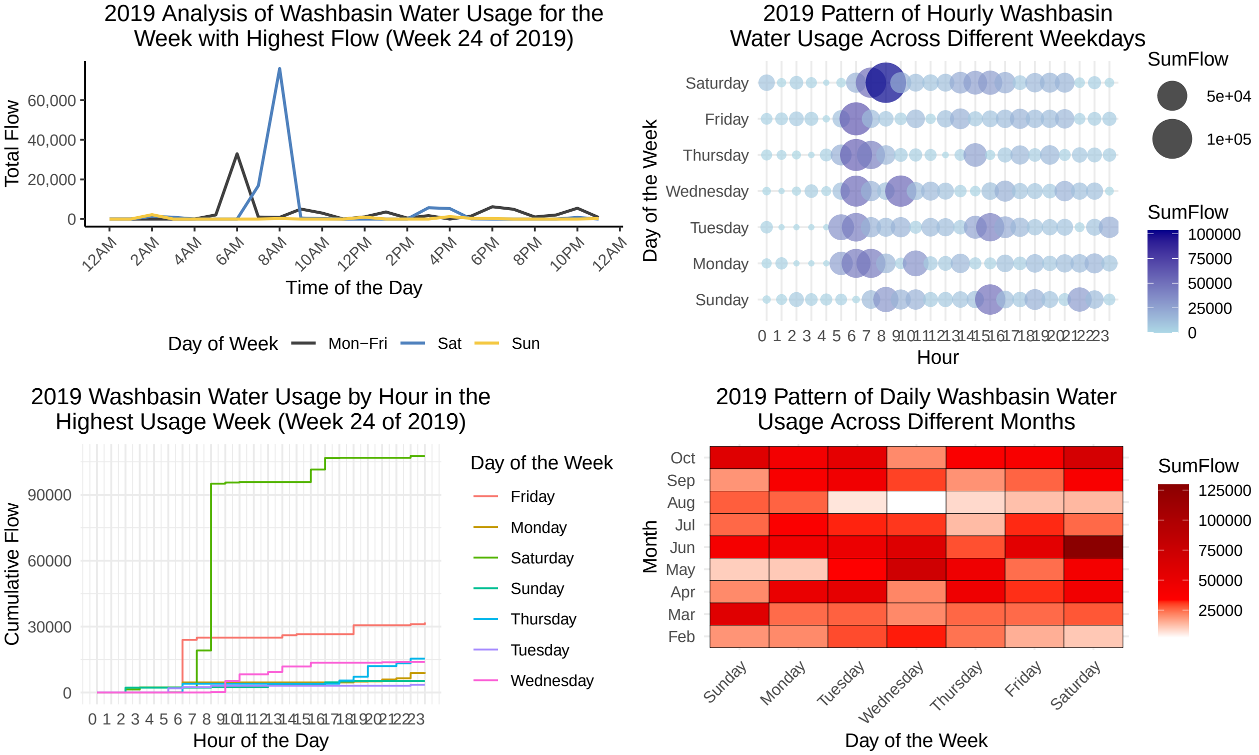


Figure 12: Complex plots for Sink appliance (2019)

Notes on the plots:

-
-
-

Summed Post 2020 flow:

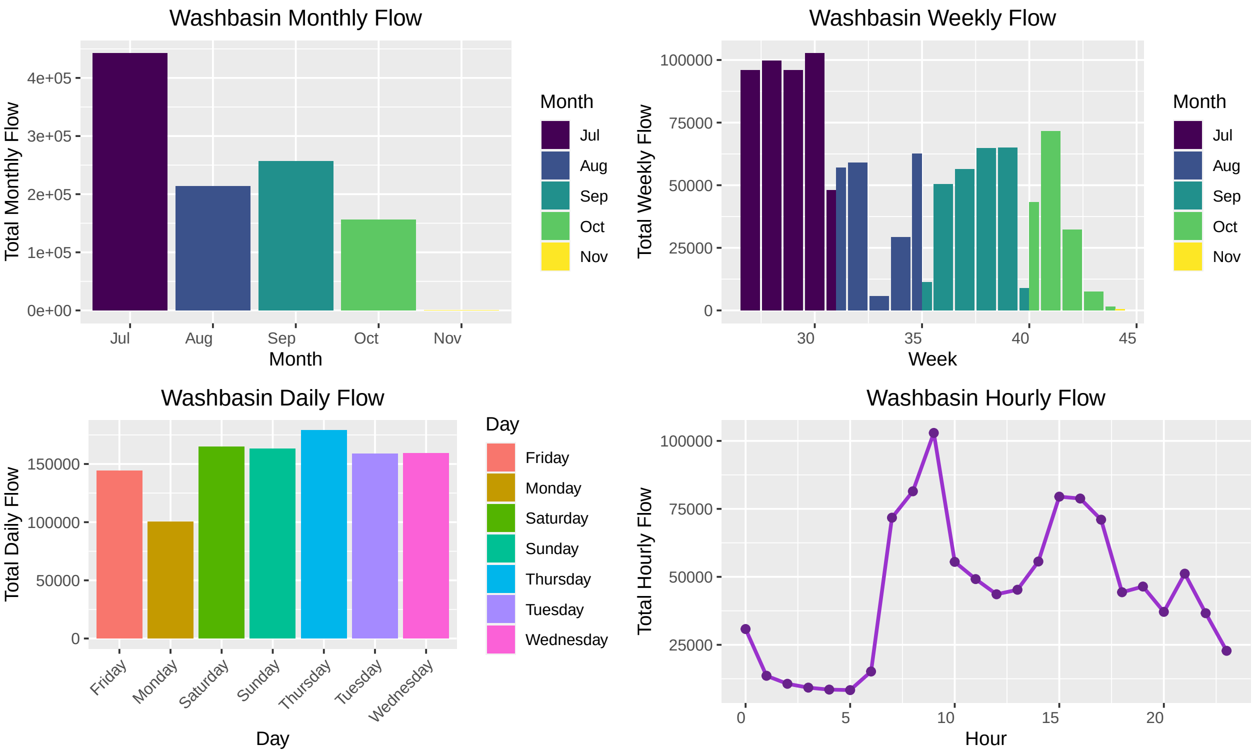


Figure 13: Monthly, Weekly, Daily and Hourly flows for Sink appliance (2020)

Summed Post 2020 flow (complex):

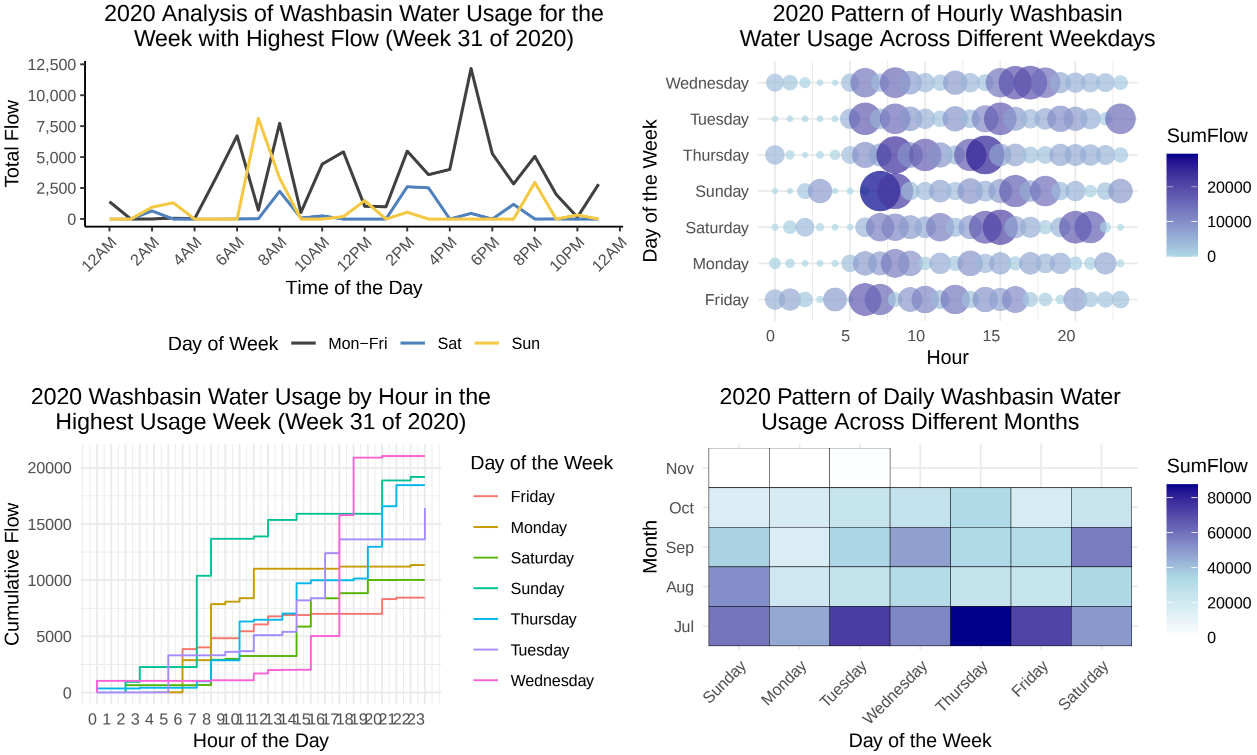


Figure 14: Complex plots for Sink appliance (2020)

Notes on the plots:

-
-
-

Bidet Usage

Summed Pre 2020 flow:

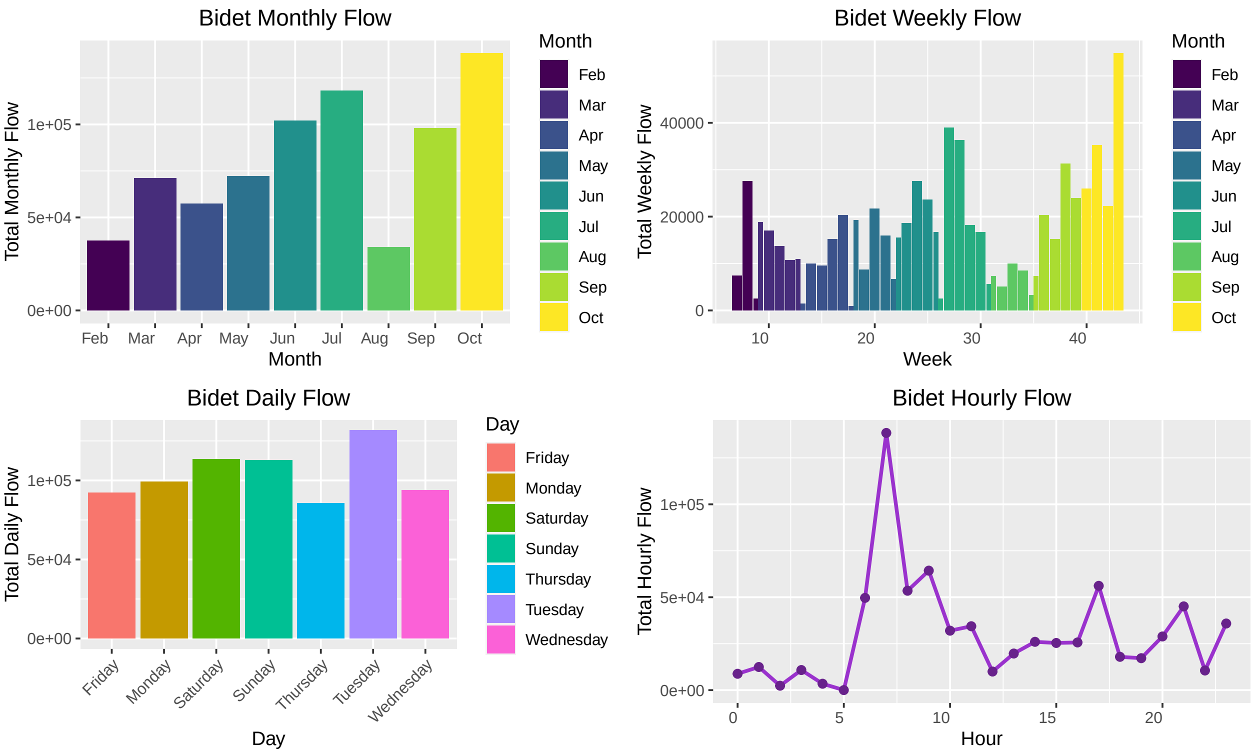


Figure 15: Monthly, Weekly, Daily and Hourly flows for Bidet appliance (2019)

Summed Pre 2020 flow (complex):

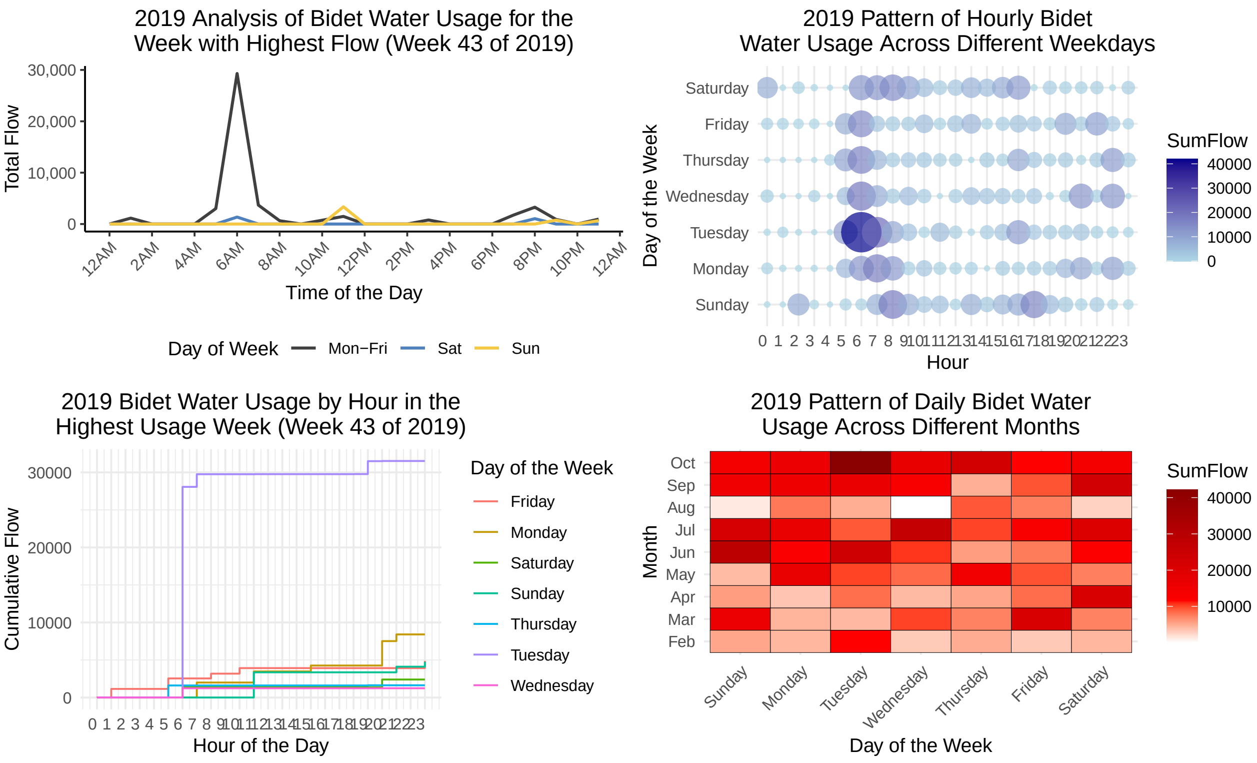


Figure 16: Complex plots for Bidet appliance (2019)

Summed Post 2020 flow:

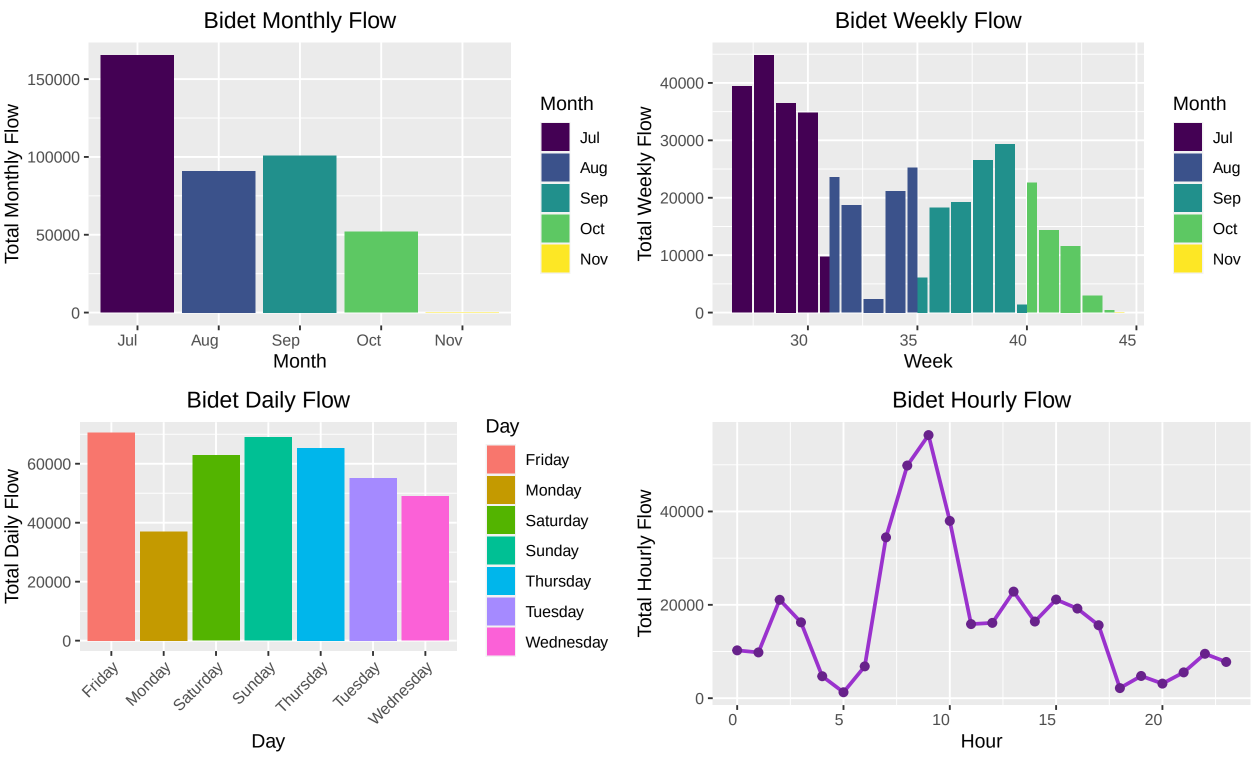


Figure 17: Monthly, Weekly, Daily and Hourly flows for Bidet appliance (2020)

Summed Post 2020 flow (complex):

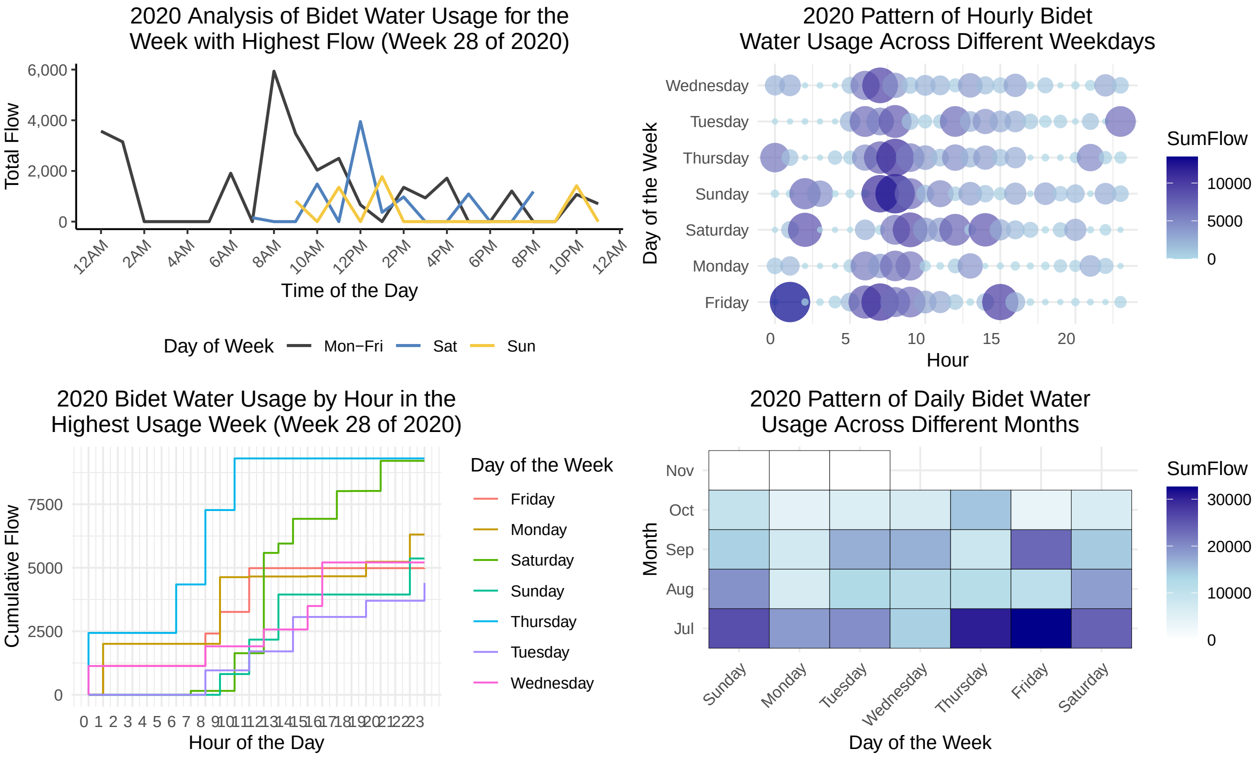


Figure 18: Complex plots for Bidet appliance (2020)

Notes on the plots:

1. In August 2019, the Bidet water usage significantly dropped and then increased, suggesting that the resident might have been away for a lengthy period, leading to reduced water consumption at home.
2. On October 22, 2019, from 6:28 to 7:28, the Bidet was in continuous use for one hour, which is unusually long, with erratic water flow, indicating a potential data anomaly. This event is also the cause for the abnormally high water flow on Tuesday in October 2019.
3. Throughout 2019, there was a peak in Bidet water usage around 7 a.m., likely correlating with the resident's morning routine, particularly for bowel movements, resulting in increased water flow.
4. By 2020, the time of peak Bidet usage shifted to 9 a.m. This change could be due to the COVID-19 pandemic's impact, as the resident, now working from home, had no commute, thereby delaying the morning routine.
5. July 2020 experienced the peak in Bidet usage, probably because the pandemic led the resident to minimize outings, increasing the time spent at home.

Kitchen Faucet Usage

Summed Pre 2020 flow:

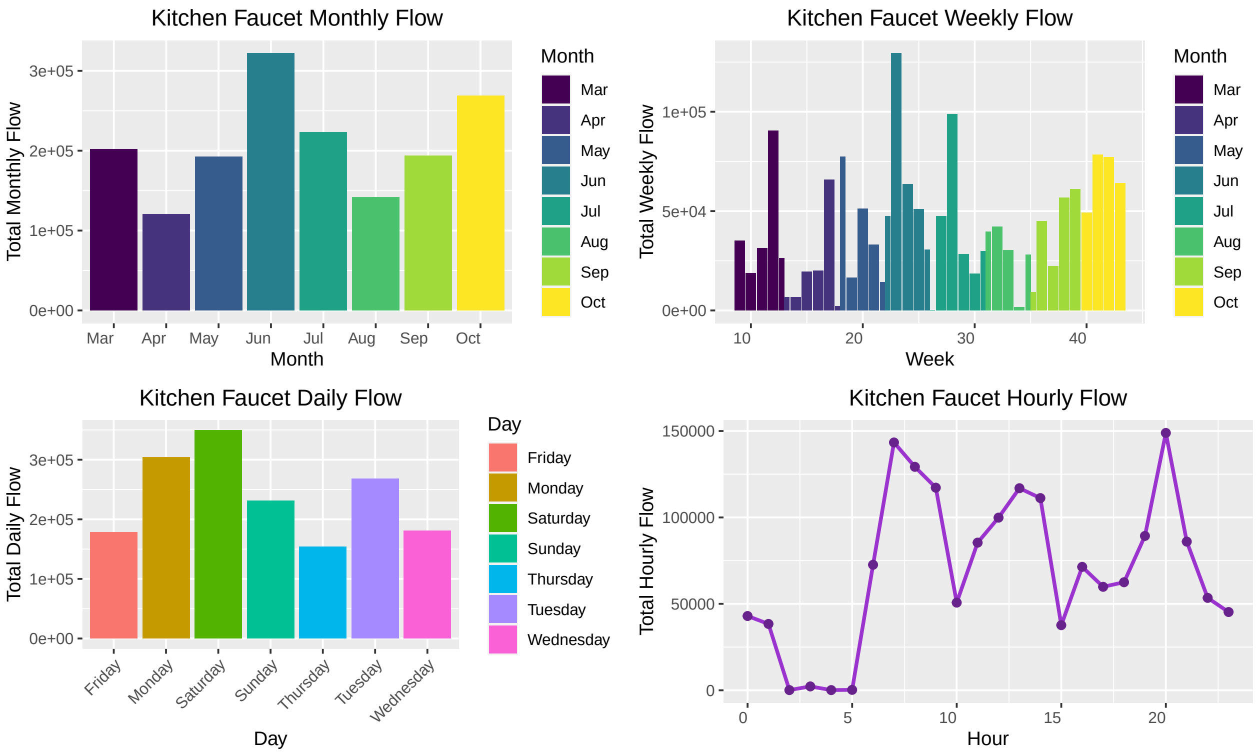


Figure 19: Monthly, Weekly, Daily and Hourly flows for Kitchen Faucet (2019)

Summed Pre 2020 flow (complex):

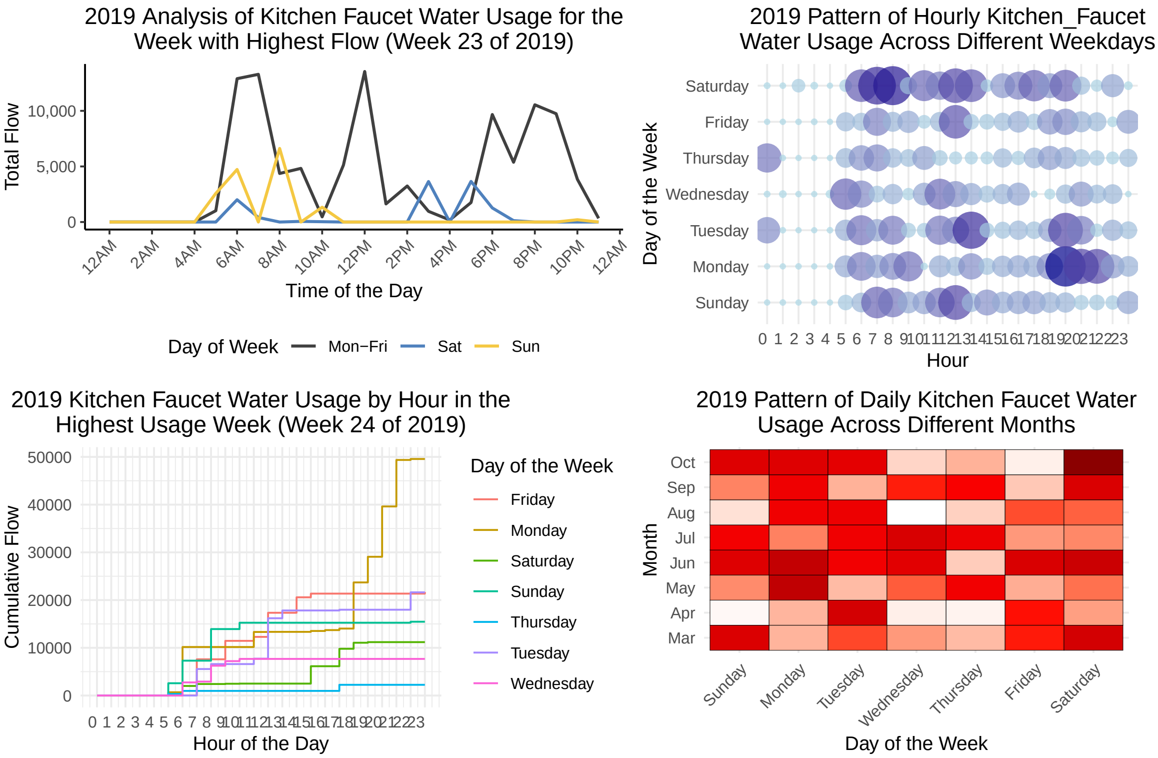


Figure 20: Complex plots for Kitchen Faucet (2019)

In Figure 19, our analysis reveals notable trends in the water usage of the Kitchen Faucet appliance in 2019. June is the month with the highest recorded water flow, surpassing 300,000 liters, while April has the lowest consumption, totaling nearly 125,000 liters. It is clear that the peak flowing week is in week 23. Moreover, there are distinctive patterns across weekdays: Mondays and Saturdays register the highest water usage, hitting 300,000 and 350,000 liters, respectively, while Thursdays just reach 150,000 liters. Notably, a clear diurnal pattern emerges, with the faucet experiencing rapid flow escalation from 6 AM, peaking at 7 AM, followed by a secondary peak at 8 PM, gradually tapering off until 11 PM.

In Figure 20, we dive deeper in further combinations within our data. Notably, during the highest flowing week (week 23 in June 2019), we observe that the water flow from the Kitchen Faucet tends to be higher on weekdays compared to Saturdays and Sundays. Within weekdays, we have the following peaks: firstly, from 6 AM to 7 AM, followed by a notable surge from 10 AM to 12 PM, with a subsequent peak occurring at 8 PM before gradually declining overnight. Moreover, Monday is the day in which the appliance is more used. During the weekends, there is a different pattern: Sunday exhibit more use during the morning while Saturday showcase more activity during the afternoon before 6PM.

In Figure 20, we can identify a monthly water usage pattern by days of the week. We can highlight that in all months, Saturdays have the highest flow being September the month with a total flow of more than 150,000 liters. In October, Wednesday and Friday had the lowest flows with less than 25000 liters.

In Figure 20, we discern a recurring monthly water usage pattern correlated with specific days of the week. We can highlight that in October, Saturdays exhibit the highest flow rates. Across all months, we consistently observe elevated flows predominantly between Sundays and Wednesdays. However, in March, the flows from Monday to Thursday are remarkably low, not surpassing 25,000 liters. Moreover, we investigate the hourly water usage pattern throughout the week: Saturday mornings have the highest flow rates before 9 AM, while Monday nights experience surges post 5 PM. Additionally, a trend of sustained flow rates characterizes weekday mornings, with a notable decrease in Kitchen Faucet usage observed during nighttime hours.

Summed Post 2020 flow:

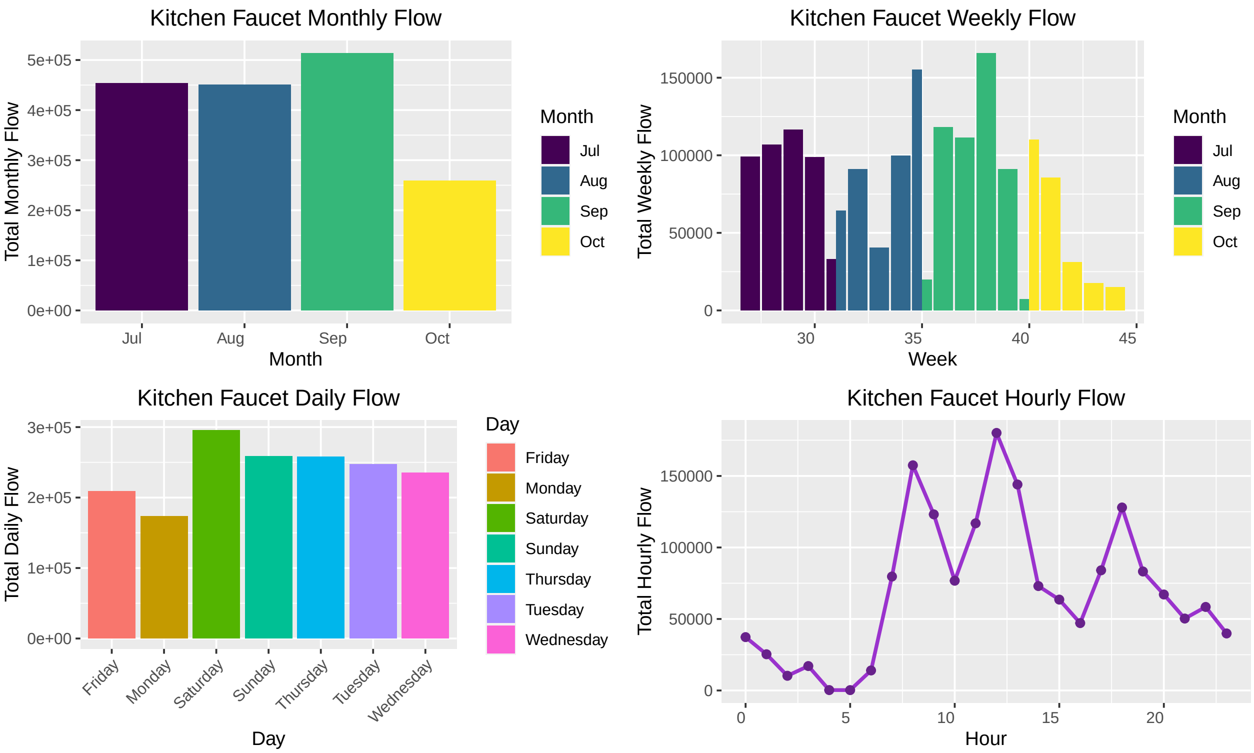


Figure 21: Monthly, Weekly, Daily and Hourly flows for Kitchen Faucet (2020)

Summed Post 2020 flow (complex):

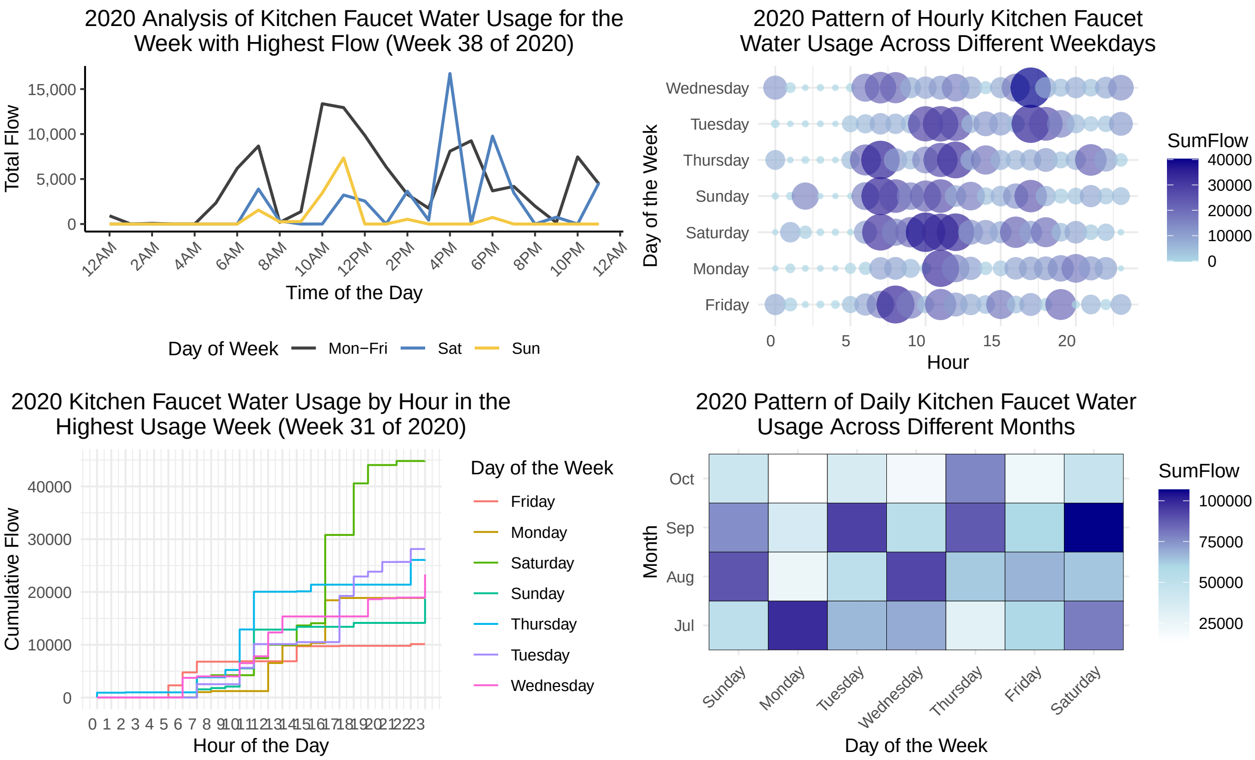


Figure 22: Complex plots for Kitchen Faucet (2020)

In Figure 19, our analysis reveals notable trends in the water usage of the Kitchen Faucet appliance in 2019. June is the month with the highest recorded water flow, surpassing 300,000 liters, while April has the lowest consumption, totaling nearly 125,000 liters. It is clear that the peak flowing week is in week 23. Moreover, there are distinctive patterns across weekdays: Mondays and Saturdays register the highest water usage, hitting 300,000 and 350,000 liters, respectively, while Thursdays just reach 150,000 liters. Notably, a clear diurnal pattern emerges, with the faucet experiencing rapid flow escalation from 6 AM, peaking at 7 AM, followed by a secondary peak at 8 PM, gradually tapering off until 11 PM.

In Figure 20, we dive deeper in further combinations within our data. Notably, during the highest flowing week (week 23 in June 2019), we observe that the water flow from the Kitchen Faucet tends to be higher on weekdays compared to Saturdays and Sundays. Within weekdays, we have the following peaks: firstly, from 6 AM to 7 AM, followed by a notable surge from 10 AM to 12 PM, with a subsequent peak occurring at 8 PM before gradually declining overnight. Moreover, Monday is the day in which the appliance is more used. During the weekends, there is a different pattern: Sunday exhibit more use during the morning while Saturday showcase more activity during the afternoon before 6PM.

In Figure 20, we can identify a monthly water usage pattern by days of the week. We can highlight that in all months, Saturdays have the highest flow being September the month with a total flow of more than 150,000 liters. In October, Wednesday and Friday had the lowest flows with less than 25000 liters.

In Figure 20, we discern a recurring monthly water usage pattern correlated with specific days of the week. We can highlight that in October, Saturdays exhibit the highest flow rates. Across all months, we consistently observe elevated flows predominantly between Sundays and Wednesdays. However, in March, the flows from Monday to Thursday are remarkably low, not surpassing 25,000 liters. Moreover, we investigate the hourly water usage pattern throughout the week: Saturday mornings have the highest flow rates before 9 AM, while Monday nights experience surges post 5 PM. Additionally, a trend of sustained flow rates characterizes weekday mornings, with a notable decrease in Kitchen Faucet usage observed during nighttime hours.

Shower Usage

Summed Pre 2020 flow:

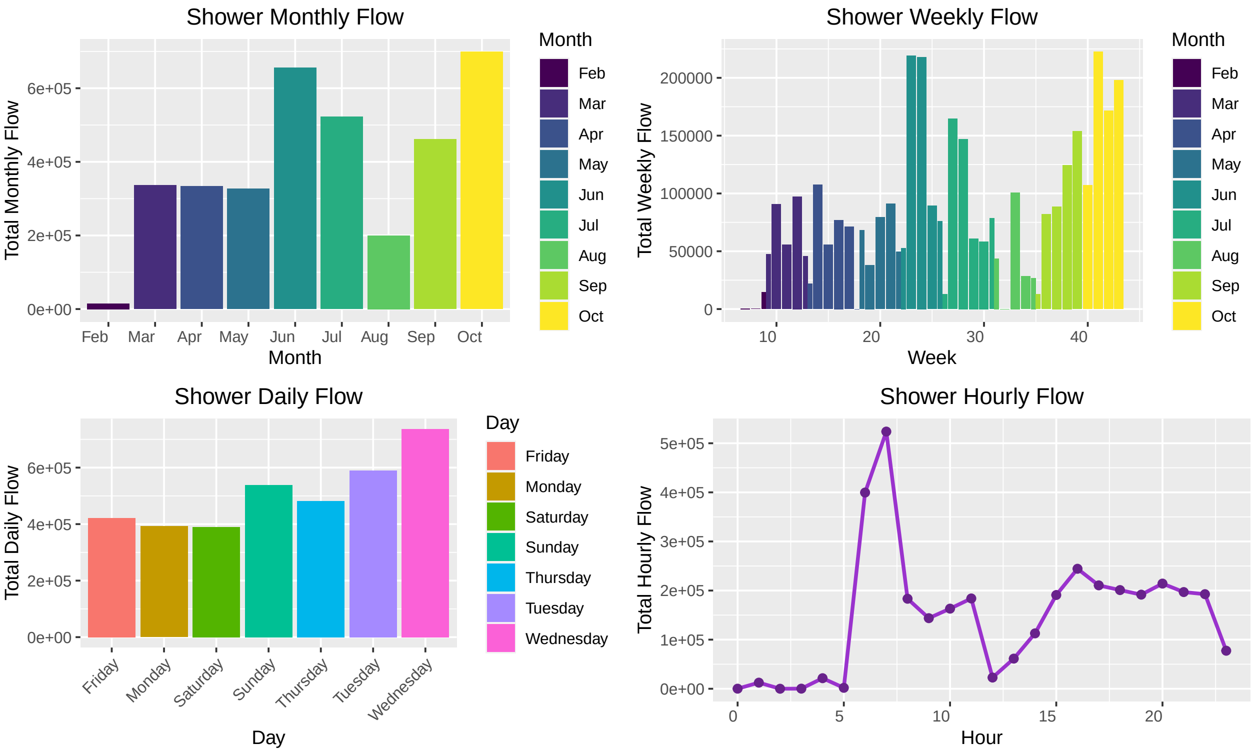


Figure 23: Monthly, Weekly, Daily and Hourly flows for Shower appliance (2019)

Summed Pre 2020 flow (complex):

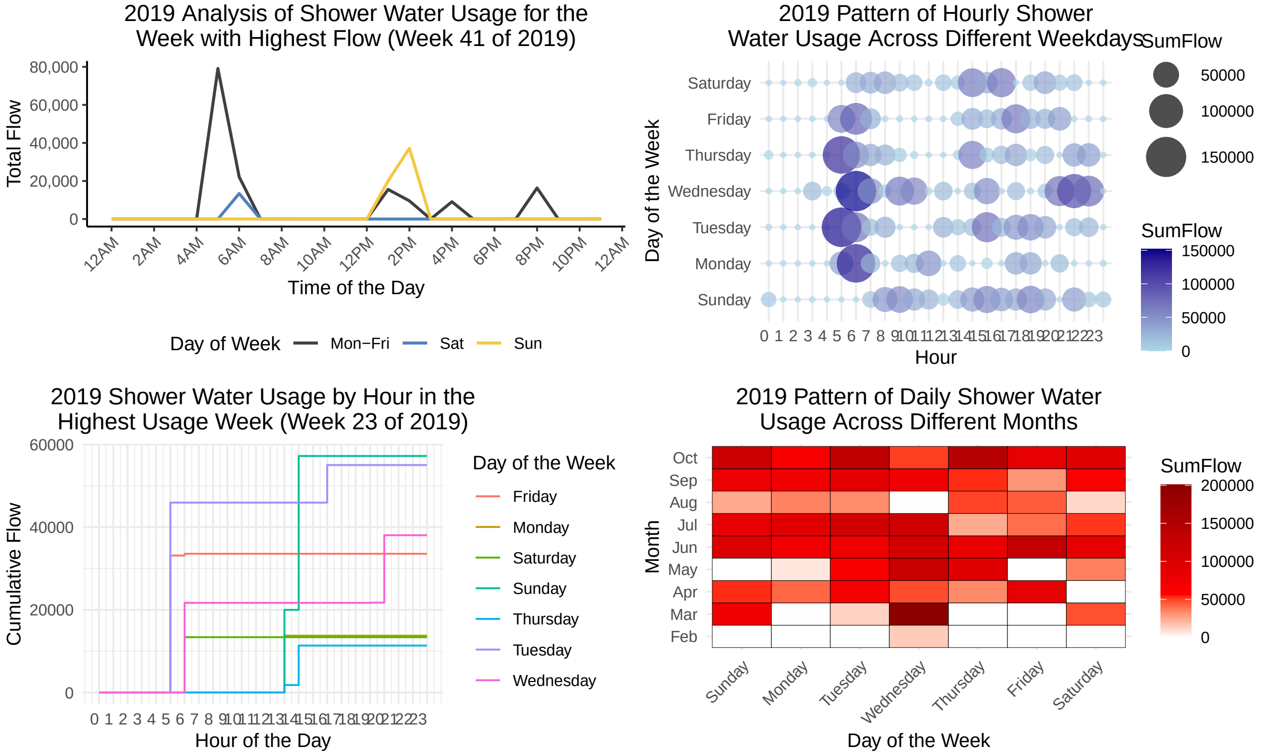


Figure 24: Complex plots for Shower appliance (2019)

Notes on the plots:

-
-
-

Summed Post 2020 flow:

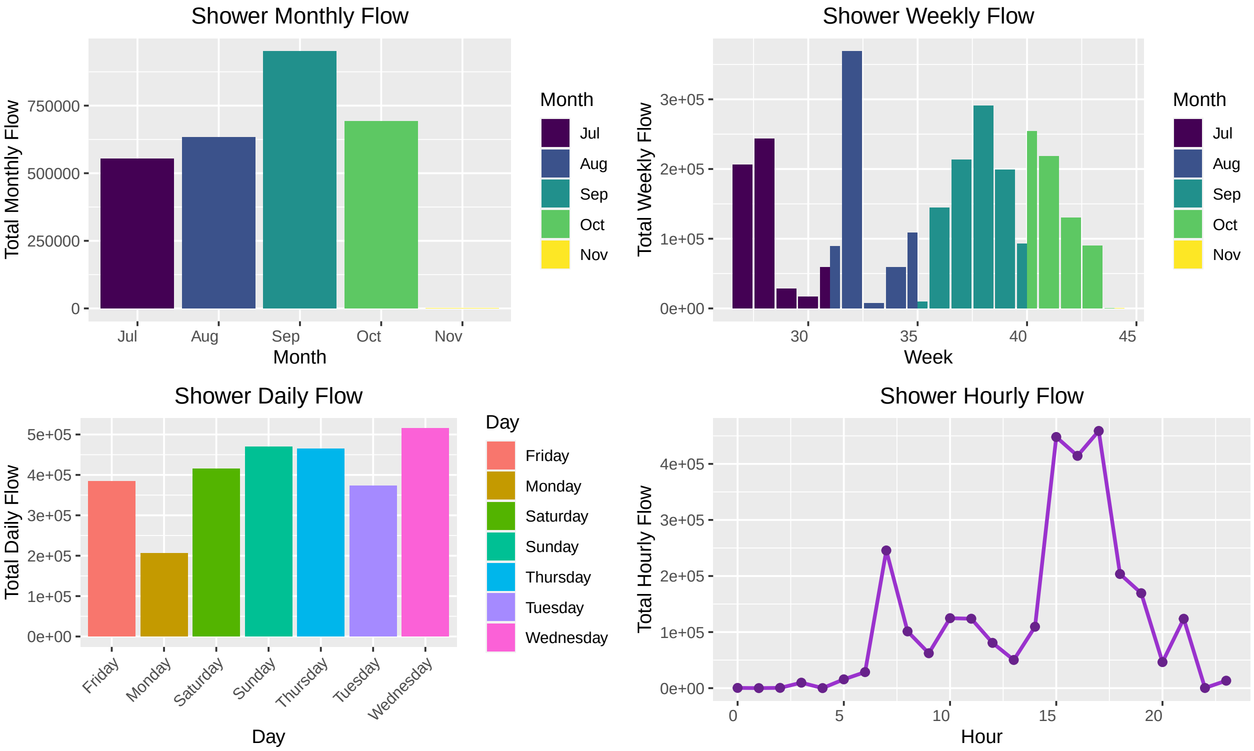


Figure 25: Monthly, Weekly, Daily and Hourly flows for Shower appliance (2020)

Summed Post 2020 flow (complex):

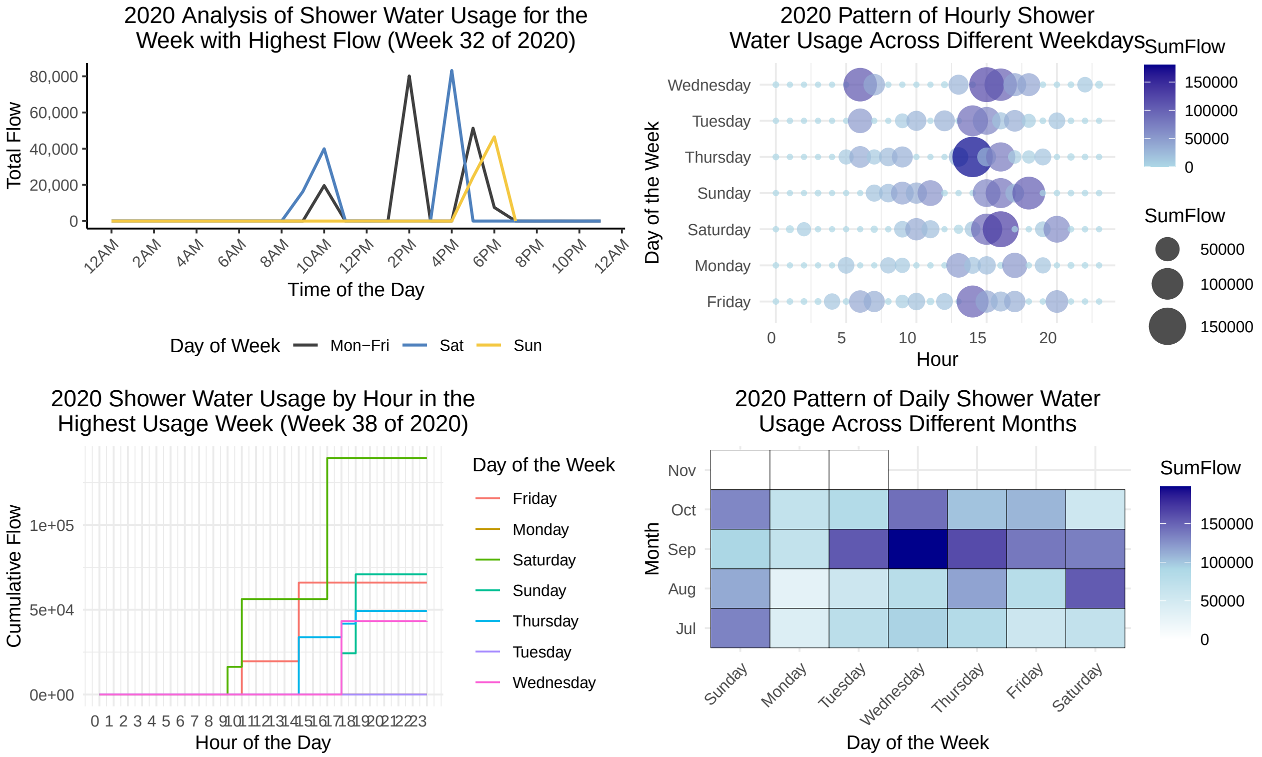


Figure 26: Complex plots for Shower appliance (2020)

Notes on the plots:

-
-
-

Washing Machine Usage

Summed Pre 2020 flow:

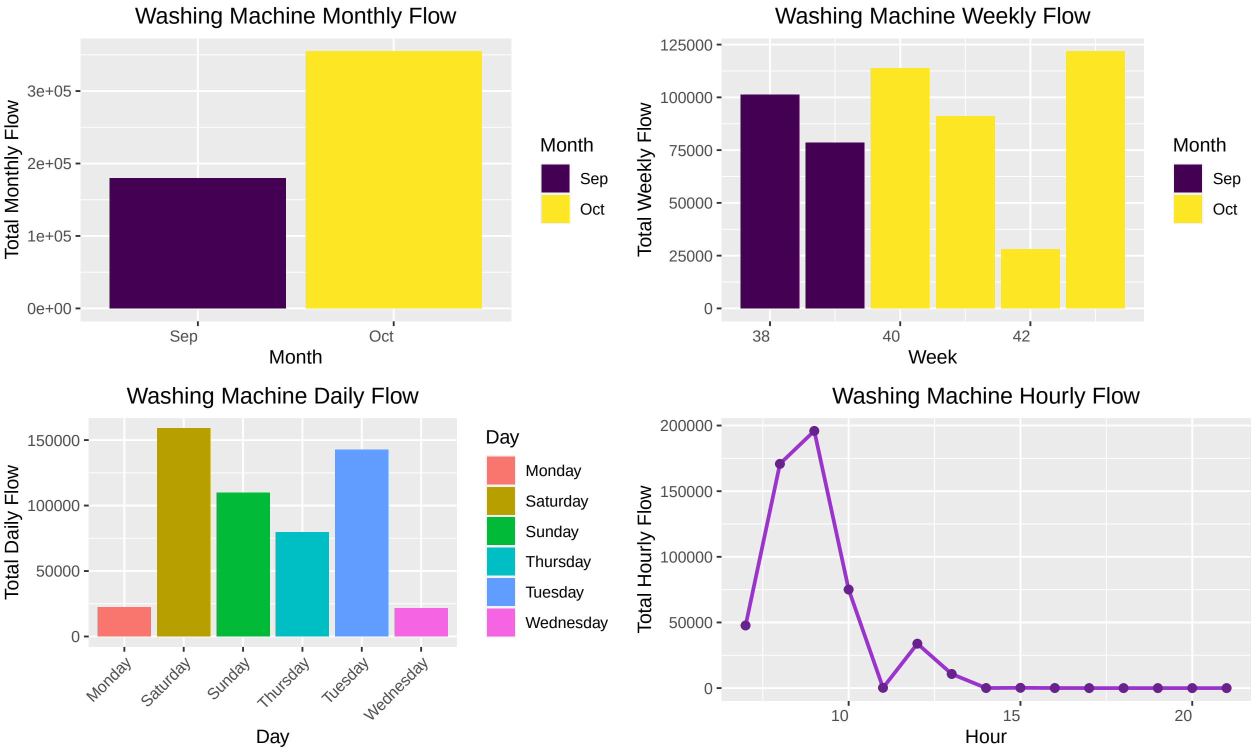


Figure 27: Monthly, Weekly, Daily and Hourly flows for Washing Machine (2019)

Summed Pre 2020 flow (complex):

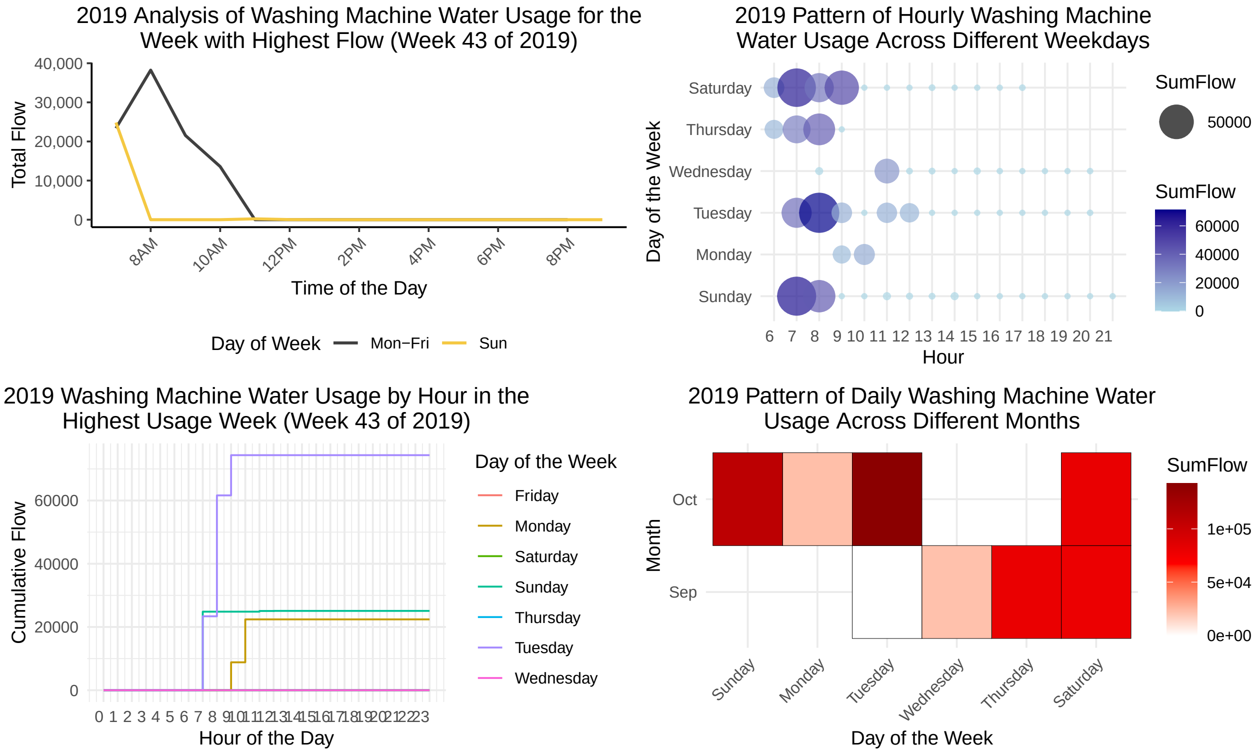


Figure 28: Complex plots for Washing Machine (2019)

Notes on the plots:

-
-
-

Summed Post 2020 flow:

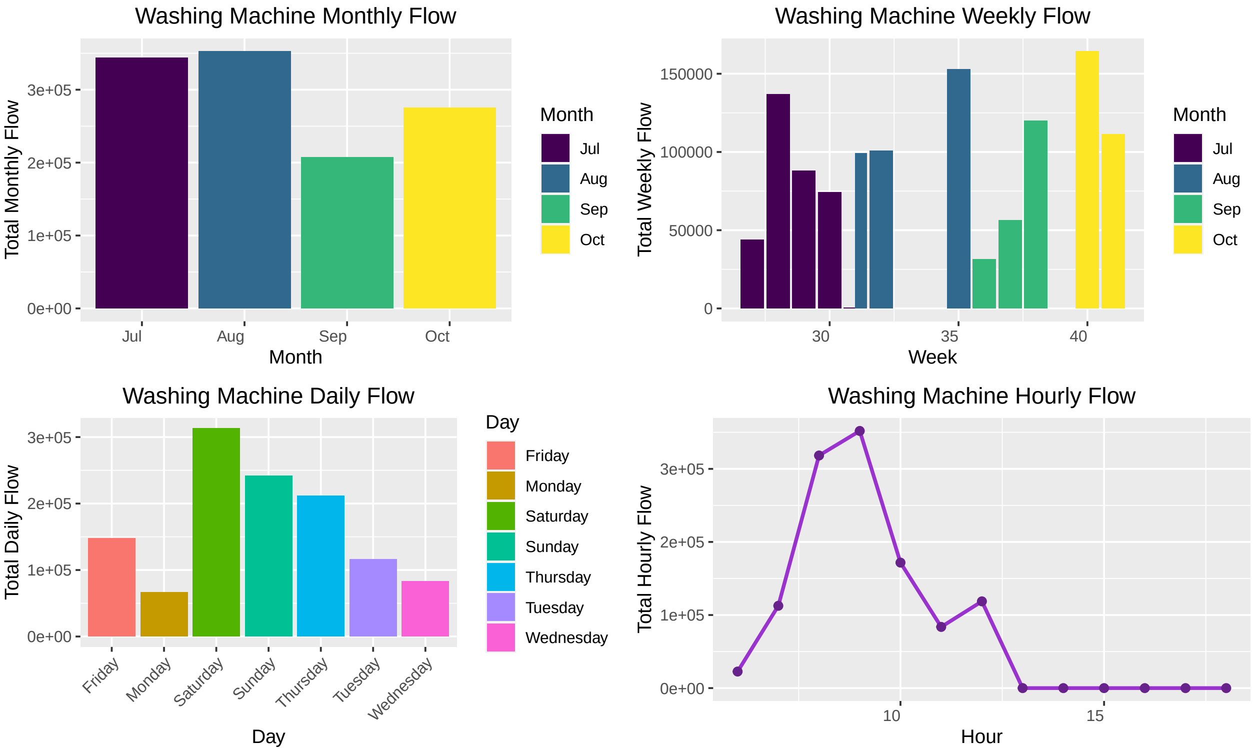


Figure 29: Monthly, Weekly, Daily and Hourly flows for Washing Machine (2020)

Summed Post 2020 flow (complex):

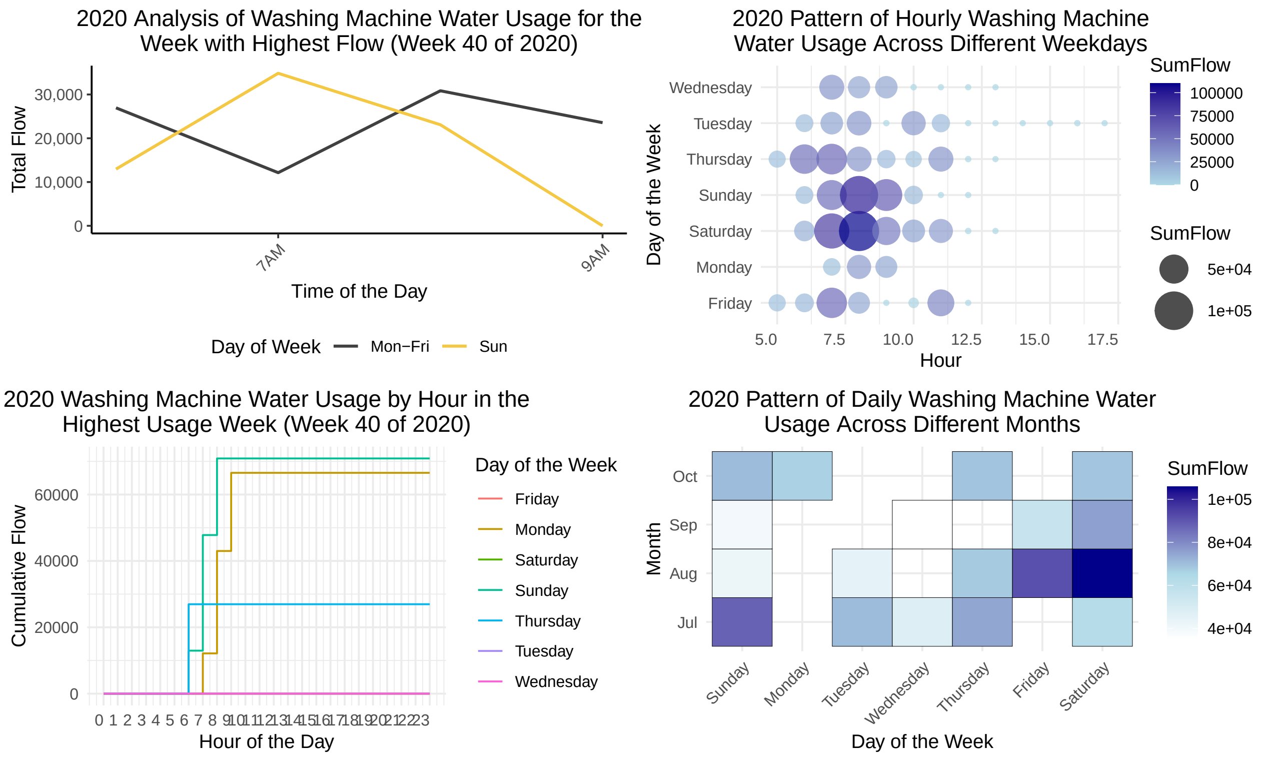


Figure 30: Complex Plots for Washing Machine (2020)

Notes on the plots:

-
-
-

Toilet Usage

Summed Pre 2020 flow:

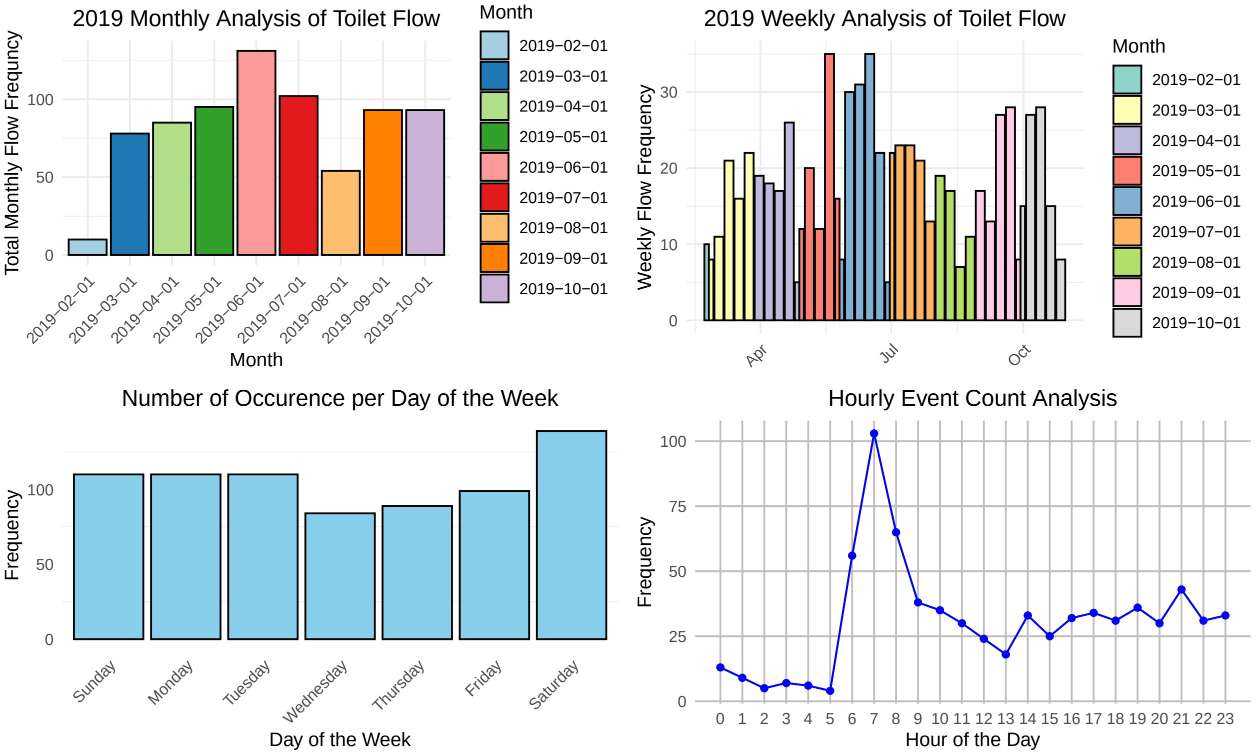


Figure 31: TITLE FOR THE PLOT (2019)

Summed Post 2020 flow:

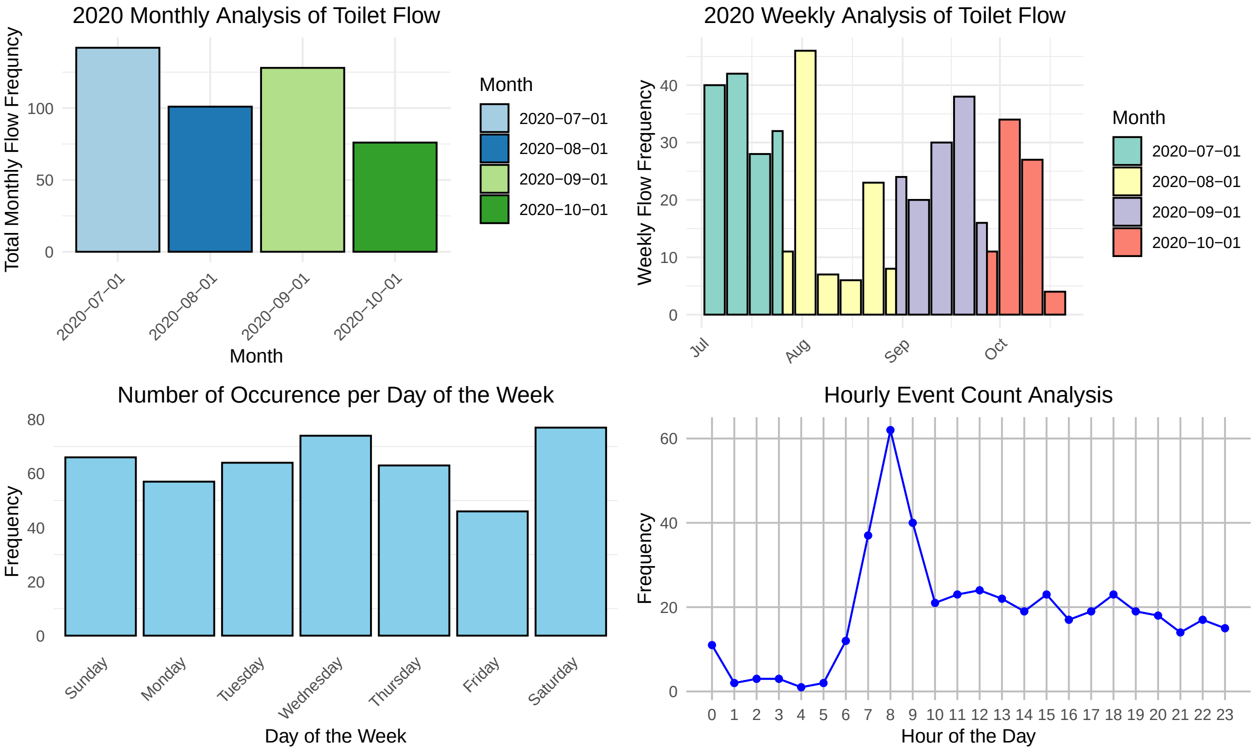


Figure 32: TITLE FOR THE PLOT (2020)

Dish Washer Usage

Summed Pre 2020 flow:

Summed Pre 2020 flow (complex):

Summed Post 2020 flow:

Summed Post 2020 flow (complex):

Cycle 3 - Water Flow Prediction

Business Understanding - Predicting Water Usage

The primary business objective of this cycle is to enhance the efficiency and sustainability of water management practices for this single household in Naples, as facilitated by the local water utility company. Specifically, the aim is to develop a predictive model for the aggregated water flow leveraging various algorithms and statistical techniques, enabling the utility company to anticipate demand and allocate resources more effectively. By understanding and forecasting water consumption patterns at the household level, the utility company seeks to minimise waste, optimise resource allocation and promote environmental conservation efforts.

The available household water usage data provides insights into consumption patterns and trends over the specified time period. Analysis reveals fluctuations in water usage, deemed to be influenced by seasonal variations, demographic changes and appliance usage patterns. Understanding the context of water usage within the household, including lifestyle habits and socio-economic factors, is crucial for developing accurate predictive models. Additionally, consideration of external factors such as regulatory requirements and environmental concerns provides essential context for the project's objectives and constraints.

The project plan encompasses several key phases, including data preparation, model development, evaluation and finally deployment of all 3 cycles. In the data preparation phase, historical water flow data will be cleaned and preprocessed appropriately according to model demands. The model development phase involves selecting time series methods, training the predictive model on the data and aiming for the best performance. Evaluation metrics such as r-squared, AMSE and RMSE/MSE will be used to assess the predictive model's performance. Finally, the deployment phase involves implementing the predictive model and other parts of the analysis into operational processes and providing stakeholders with actionable insights for decision-making.

Data Preparation

Preparing the data for modelling.

Modelling & Evaluation

The data available poses a particular problem, its very sparse. It contains valid zeros for every time no water flow occurs within the household, this particular timeseries data is known as a intermitant timeseries. We need to find a differenrt way to modelling the data for this reason as the usual ARIMA models will not be as effective as predicting a trend with this structure is more difficult.

Specialised techniques designed to handle time series data with excess zeros:

- Croston's Method (and Adjusted)
- Discrete ARMA
- INARMA
- Bootsrapping
- Temporal Aggregation

Croston's Method

Separates data into two components, non-zero demand and inter-demand timeseries.

https://medium.com/swlh/forecasting-an-intermittent-time-series-1461de7616fe https://www.youtube.com/watch?v=M0cJwmdzOu4

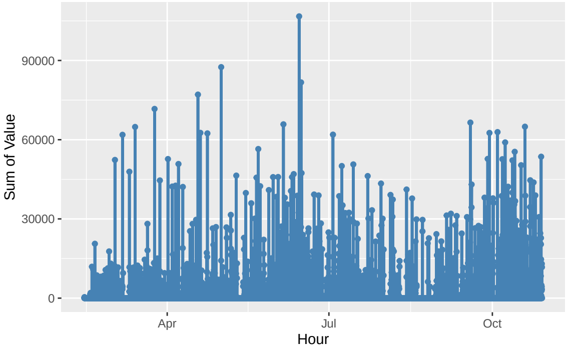
Deployment

Using findings, explain how the action for the company.

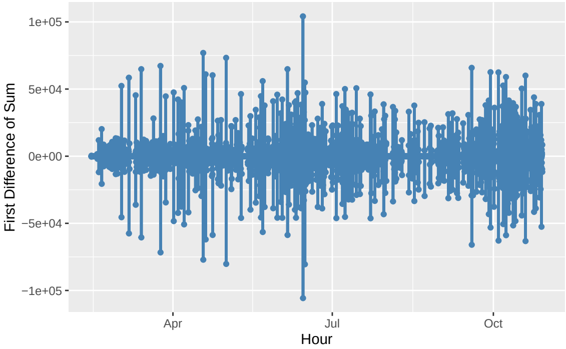
Appendix

Cycle 1

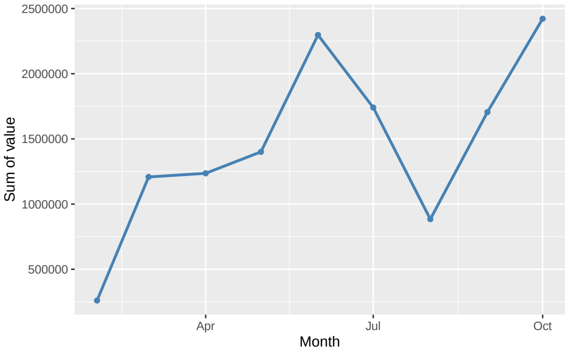
Plot 1: Hourly−based time series



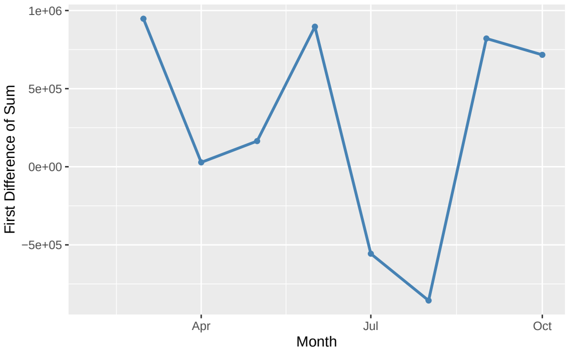
Plot 2: First difference of hourly time series



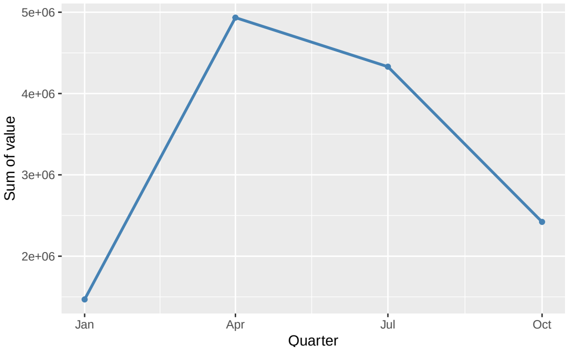
Plot 3: Monthly−based time series



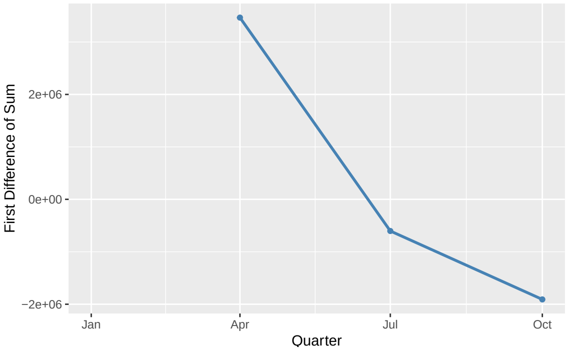
Plot 4: First difference of monthly sum time series



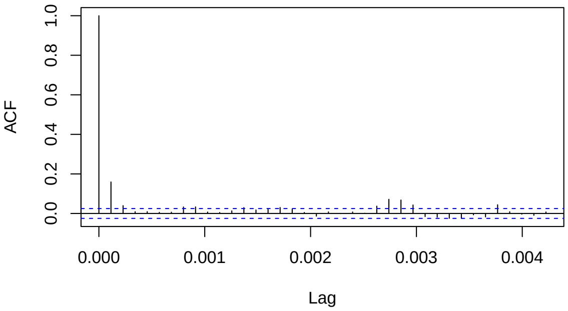
Plot 5: Quarterly−based time series



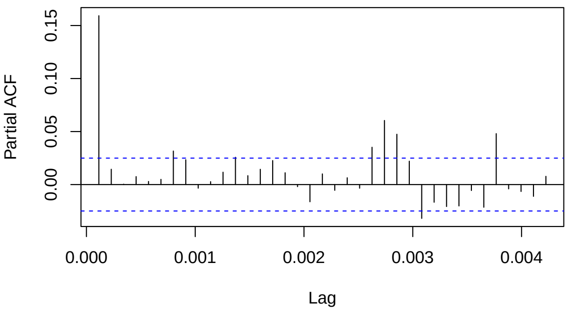
Plot 6: First difference of quarterly sum time series



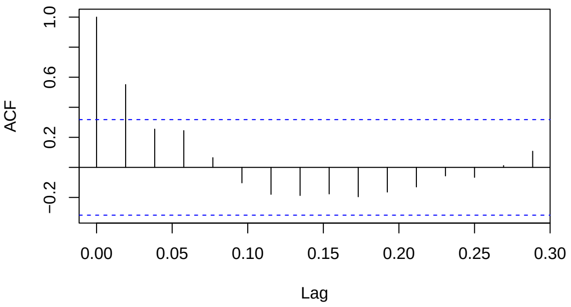
Plot 7: ACF for hourly data



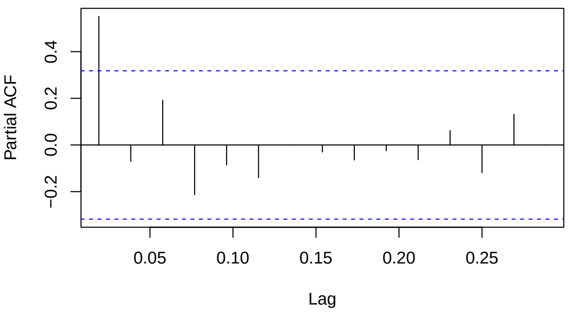
Plot 8: PACF for hourly data



Plot 9: ACF for weekly data



Plot 10: PACF for weekly data


Partial ACF
Lag