CMP7205 — Applied Statistics

England' and Wales' Water: Investigating Water Companies' Performance

Coursework Assignment Report

Alexander Samuel Roberts

MSc Big Data Analytics

Student No. 21158995

Submitted: December 2024

Word Count: 3,102

Module lead: Professor Atif Azad

BIRMINGHAM CITY
University

School of Computing and Digital Technology
Faculty of Computing, Engineering and the Built Environment
Birmingham City University

Contents

Executive Summary

Over recent years, water companies across England and Wales have come under increasing scrutiny from the media and consumers for poorer performance and rising costs. But there are little recent academic sources to help substantiate these claims and concerns raised. This report aims to help verify to what extent concerns regarding the Water Industry are valid. This will be done namely by selecting and analysing a dataset supplied by the UK Government as part of the Water Resource Management Plan (WRMP), which contains official statistics regarding several performance indicators for the Water Industry across England and Wales. This report will help determine to what extent the Water Industry must adapt to changing markets and consumer expectations.

List of Tables

List of Figures

1 Introduction

1.1 Problem Domain

Within the past few years, water companies across England and Wales have come under increasing scrutiny for their performance and environmental impacts (Thomas, 2019) (Masud, 2024) (Stallard, 2024). This includes key topics such as the quality of water supplies, commitments to maintaining and improving England's waterways, filtering sewage waste, and the value for money for customers. This report will help investigate to what extent these claims and concerns are justified by analysing and evaluating the significance of certain performance metrics and trends of relevant data.

1.2 Aim and Objectives

The aim of this report is to conclude (from available information) whether the overall water industry's performance has degraded from previous years to present. The objectives of this report are as follows:

1. Perform a review of related works.
2. Establish statistical questions, relevant to the report aim, which should be particularly investigated.
3. Research and select a free and open-source (FOSS) dataset(s) that should be analysed for helping investigate this report's aim.
4. Determine a methodology that should be followed to help process and analyse available data to assist in answering statistical questions stated in Objective 2.
5. Carry out the methodology on the selected dataset(s) and document the results.
6. Establish conclusions regarding the implementation results, and future works.

2 Related Works

Walker, Styles et al. (2021) leveraged six productivity metrics (including the Hicks-Moorsteen productivity index) to investigate and benchmark twelve water and sewage companies (WaSCs) in the United Kingdom during the 2010s. Their results highlighted a slight overall decrease in the overall water industry performance. Although, productivity in the industry did increase by 1.8% during 2014-2018, but Walker, Styles et al. (2021) noted this is likely due to wider economic developments rather than specifically the industry itself. It was also noted WaSCs that serve more densely populated areas had performed better compared to other WaSCs with less densely populated areas. This is supported by Molinos-Senante and Maziotis (2018) and Villegas, Molinos-Senante and Maziotis (2019). It is important to note these results by Walker, Styles et al. (2021) are from a small sample size, and therefore, might not wholly represent the industry. However, their results also take into account factors such as customer satisfaction and renewable energy — implying their findings are more comprehensive compared to other studies. Walker, Williams and Styles (2020) concludes there is substantial room for improvement for both energy and efficiency for water companies across England and Wales. Their use of hypothesis testing supported this conclusion.

As well as efficiency, WaSCs are supposedly under strict environmental commitments. However, Purnell et al. (2021) found rising sewage dumps for the Water Industry in 2019 were statistically significantly higher compared to previous years. But on the other hand, these numbers were still less compared to 2010. Sala-Garrido et al. (2021) found that in general, water companies produced excess greenhouse gas emissions for the same amount of output during the 2010s. This was particularly prominent for larger water companies. Sala-Garrido et al. (2021) iteratively addresses limitations of their research and experiments, such as by using multiple models and cross-efficiency techniques — helping validate their results.

Overall, there is limited recent research material on the UK Water Industry, with much of the research done during the 2000s and 2010s (as indicated by Goh and See (2021)). This report will contribute to helping fill this gap.

3 Identified Dataset

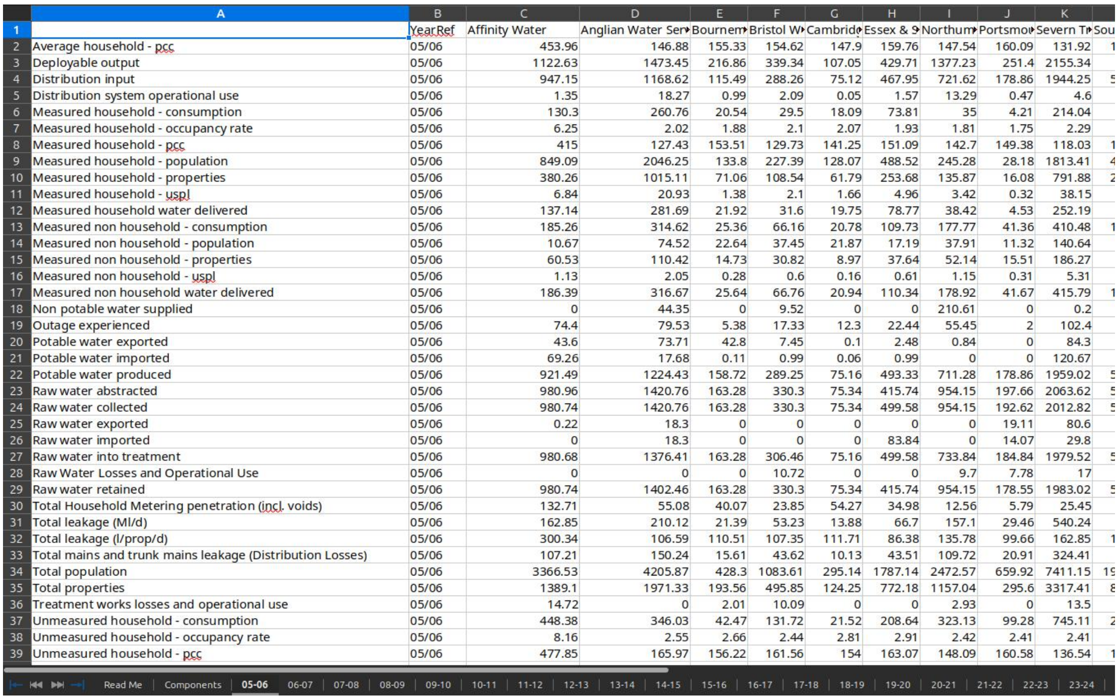
According to the UK Government (2024a), every year, water companies across England must publish relevant data regarding their progress towards achieving a joint performance agreement; the Water Resource Management Plan (WRMP). This data is open-source and is readily available by the UK Government (2024a). Moreover, the dataset was updated in October 2024, making it fairly recent and therefore relevant for use. The metrics should be fairly objective (as this is enforced by UK standards authorities such as OfWat), but as the data relies on being supplied by the companies themselves, this could provide a window for potential bias to be introduced.

The dataset consists of a single Microsoft Excel file, with a number of Excel sheets representing the annual period the data relates to. For example, the sheet '07-08' represents the collected data between 1st April 2007 to 31st March 2008, according to the metadata described in the Excel file. This is illustrated in Figure 3.1. The various attributes of each sheet are typically the same, but there are some differences, such as fewer or additional attributes. All attributes of the dataset are described in Table A.1, with the respective descriptions retrieved from the UK Government (2024b). Due to the number of attributes, detailed descriptions will not be included in this report, but further information can be found in an official report by the UK Government (2024b). Moreover, all attributes, except for 'YearRef' (as this is an ordinal label), are numerical and have a ratio statistical type.

Some keys terms used within the dataset include:

- 'Measured': refers to properties fitted with a meter, and customers pay for water on a volumetric basis.
- 'Unmeasured': refers to properties where no meter is fitted, and so, demand is estimated.
- 'Void': properties which are connected to water supplies but do not receive a charge — which typically includes properties with no occupants.
- 'Households': residential/domestic buildings.

- ’Non-households’: non-domestic buildings (e.g., hospitals, universities, etc.)
- ’USPL’: underground supply pipe leakage.
- ’Ml/d’: Megalitres per day.
- ’l/h/d’: Litres per household per day.
- ’PCC’: per capita consumption.



	A	B	C	D	E	F	G	H	I	J	K	
1		YearRef	Affinity Water	Anglian Water Ser	Bournem	Bristol W	Cambrid	Essex & S	Northum	Portsmo	Severn T	Sou
2	Average household - pcc	05/06	453.96	146.88	155.33	154.62	147.9	159.76	147.54	160.09	131.92	1
3	Deployable output	05/06	1122.63	1473.45	216.86	339.34	107.05	429.71	1377.23	251.4	2155.34	
4	Distribution input	05/06	947.15	1168.62	115.49	288.26	75.12	467.95	721.62	178.86	1944.25	5
5	Distribution system operational use	05/06	1.35	18.27	0.99	2.09	0.05	1.57	13.29	0.47	4.6	
6	Measured household - consumption	05/06	130.3	260.76	20.54	29.5	18.09	73.81	35	4.21	214.04	
7	Measured household - occupancy rate	05/06	6.25	2.02	1.88	2.1	2.07	1.93	1.81	1.75	2.29	
8	Measured household - pcc	05/06	415	127.43	153.51	129.73	141.25	151.09	142.7	149.38	118.03	1
9	Measured household - population	05/06	849.09	2046.25	133.8	227.39	128.07	488.52	245.28	28.18	1813.41	4
10	Measured household - properties	05/06	380.26	1015.11	71.06	108.54	61.79	253.68	135.87	16.08	791.88	2
11	Measured household - uspl	05/06	6.84	20.93	1.38	2.1	1.66	4.96	3.42	0.32	38.15	
12	Measured household water delivered	05/06	137.14	281.69	21.92	31.6	19.75	78.77	38.42	4.53	252.19	
13	Measured non household - consumption	05/06	185.26	314.62	25.36	66.16	20.78	109.73	177.77	41.36	410.48	1
14	Measured non household - population	05/06	10.67	74.52	22.64	37.45	21.87	17.19	37.91	11.32	140.64	
15	Measured non household - properties	05/06	60.53	110.42	14.73	30.82	8.97	37.64	52.14	15.51	186.27	
16	Measured non household - uspl	05/06	1.13	2.05	0.28	0.6	0.16	0.61	1.15	0.31	5.31	
17	Measured non household water delivered	05/06	186.39	316.67	25.64	66.76	20.94	110.34	178.92	41.67	415.79	1
18	Non potable water supplied	05/06	0	44.35	0	9.52	0	0	210.61	0	0.2	
19	Outage experienced	05/06	74.4	79.53	5.38	17.33	12.3	22.44	55.45	2	102.4	
20	Potable water exported	05/06	43.6	73.71	42.8	7.45	0.1	2.48	0.84	0	84.3	
21	Potable water imported	05/06	69.26	17.68	0.11	0.99	0.06	0.99	0	0	120.67	
22	Potable water produced	05/06	921.49	1224.43	158.72	289.25	75.16	493.33	711.28	178.86	1959.02	5
23	Raw water abstracted	05/06	980.96	1420.76	163.28	330.3	75.34	415.74	954.15	197.66	2063.62	5
24	Raw water collected	05/06	980.74	1420.76	163.28	330.3	75.34	499.58	954.15	192.62	2012.82	5
25	Raw water exported	05/06	0.22	18.3	0	0	0	0	0	19.11	80.6	
26	Raw water imported	05/06	0	18.3	0	0	0	83.84	0	14.07	29.8	
27	Raw water into treatment	05/06	980.68	1376.41	163.28	306.46	75.16	499.58	733.84	184.84	1979.52	5
28	Raw Water Losses and Operational Use	05/06	0	0	0	10.72	0	0	9.7	7.78	17	
29	Raw water retained	05/06	980.74	1402.46	163.28	330.3	75.34	415.74	954.15	178.55	1983.02	5
30	Total Household Metering penetration (incl. voids)	05/06	132.71	55.08	40.07	23.85	54.27	34.98	12.56	5.79	25.45	
31	Total leakage (Ml/d)	05/06	162.85	210.12	21.39	53.23	13.88	66.7	157.1	29.46	540.24	
32	Total leakage (l/prop/d)	05/06	300.34	106.59	110.51	107.35	111.71	86.38	135.78	99.66	162.85	1
33	Total mains and trunk mains leakage (Distribution Losses)	05/06	107.21	150.24	15.61	43.62	10.13	43.51	109.72	20.91	324.41	
34	Total population	05/06	3366.53	4205.87	428.3	1083.61	295.14	1787.14	2472.57	659.92	7411.15	19
35	Total properties	05/06	1389.1	1971.33	193.56	495.85	124.25	772.18	1157.04	295.6	3317.41	8
36	Treatment works losses and operational use	05/06	14.72	0	2.01	10.09	0	0	2.93	0	13.5	
37	Unmeasured household - consumption	05/06	448.38	346.03	42.47	131.72	21.52	208.64	323.13	99.28	745.11	2
38	Unmeasured household - occupancy rate	05/06	8.16	2.55	2.66	2.44	2.81	2.91	2.42	2.41	2.41	
39	Unmeasured household - pcc	05/06	477.85	165.97	156.22	161.56	154	163.07	148.09	160.58	136.54	1

Figure 3.1: Illustration of the identified dataset in LibreOffice Calc (The Document Foundation, 2024).

3.1 Statistical Questions

The questions formulated in Table 3.1 below were identified as being the most relevant to this report’s aim.

Table 3.1: Statistical Questions

Question ID	**Question**	**Rationale for the relevance of the question**
1	How has the demand for water changed over time?	Water demand can be used to help explain a number of trends. For instance, it can be used to measure how well water companies have scaled or descaled to meet a changing market.
2	How has leakages changed over time?	Leakages in water systems can give a good indication of both the state of waste the system currently has how the company has handled leakages overall over time.
3	How does water availability differ across the various regions?	This can help evaluate to what extent geographical differences affect company performance.
4	How has individual water consumption changed over time?	This can give insight into a more granular view of how much water individual consumers use in their everyday lives.
5	How have raw water imports changed over the past decade?	This will help determine how many water companies have become more or less dependent on foreign suppliers for viable water.
6	How has the efficiency of water treatment changed over time?	This can be used as a base method of seeing how efficient water companies have become over time.
7	How has the levels of water outage changed over time?	This can be very important for consumers, as it shows how reliable the water service by the company is.

4 Methodology

Any implementation or analysis will be completed mainly using R and RStudio (Posit, 2024) — both of which are widely used for data analysis and visualisation.

4.1 Data Ingestion

Online Analytical Processing (OLAP) databases are often preferred for storing data for analysis, as they support largely improved query processing performance compared to Online Transactional Processing (OLTP) databases. One such example is ClickHouse (Clickhouse Inc., 2024), a columnar-oriented OLAP database. ClickHouse, like many OLAP databases, is excellent for storing structured data, rather than semi- or unstructured data. This database mainly leverages Structured Query Language (SQL) for queries. This database will be hosted using Docker (Docker Inc., 2024a), a containerisation software that specialises in creating virtual and isolated environments — which makes them independent and portable.

Moreover, a React application, developed by Ricciuti (2024), can be used to manage the ClickHouse database graphically. This is also available as a Docker image. Docker Compose (Docker Inc., 2024b) will be used to create and manage the aforementioned Docker images — which makes container creation easier and more manageable.

Godard (2024) has developed an R package called 'ClickHouseHTTP', which enables communication between R and a ClickHouse database. This will be used, in conjunction with a process called Extract, transform, Load, Transform (EtLT), to write the relevant data to the database. EtLT is a hybrid approach of Extract, Load, Transform (ELT) and Extract, Transform, Load (ETL), where minimal transformations are performed on raw data so that it can be stored in a database. The main benefits of this is that raw data can be stored almost directly in a structured data-oriented database (which is often not the case with ELT) without having the drawbacks of large amounts of intermediate processing as ETL does.

It is worth noting some data transformations for the ingestion process will likely include:

- Data frame capture: where the appropriate Excel sheet is converted into a data-frame, so the data can be manipulated and transformed further.
- Semantic schema standardisation: removing semantic heterogeneity from the column names.
- Data type conversion(s): the data types are updated to reflect appropriateness.

4.2 Data Analytics

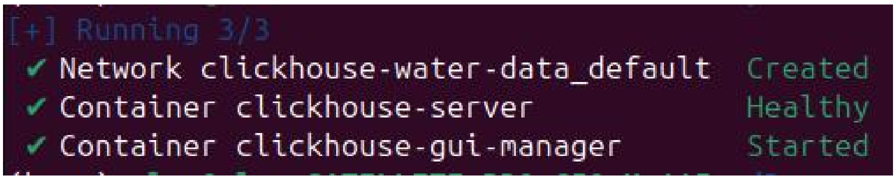
There are several forms of analytics that could be leveraged to investigate the statistical questions in Table 3.1. This includes:

- Time-series analysis: for seeing how data has changed over time. This will be particularly useful for looking at temporal changes.
- Exploratory Data Analysis (EDA): helps to discover trends and patterns in the data.
- Statistical analysis: specifically looks at how the data is distributed, including its Measurements of Central Tendency and Dispersion.
- Inferential analysis: where techniques such as hypothesis testing can be leveraged to help verify (or disregard) assumptions, for example, regarding differences between samples.
- Predictive analytics: leveraging predictive models (such as linear regression or other machine learning algorithms) to predict future trends.

5 Experimental Setup

The first step is to create the ClickHouse database. The Docker Compose YAML code for this can be found in Appendix C. The code beneath is the necessary BASH command to initiate the containers. The '-d' will force the containers to run in detached mode (in the background). Figure 5.1 illustrates the successful start-up of these containers using Docker v27.2.1.

```
docker compose --file compose-clickhouse.yaml up -d
```



```
[+] Running 3/3
 ✔ Network clickhouse-water-data_default  Created
 ✔ Container clickhouse-server            Healthy
 ✔ Container clickhouse-gui-manager       Started
```

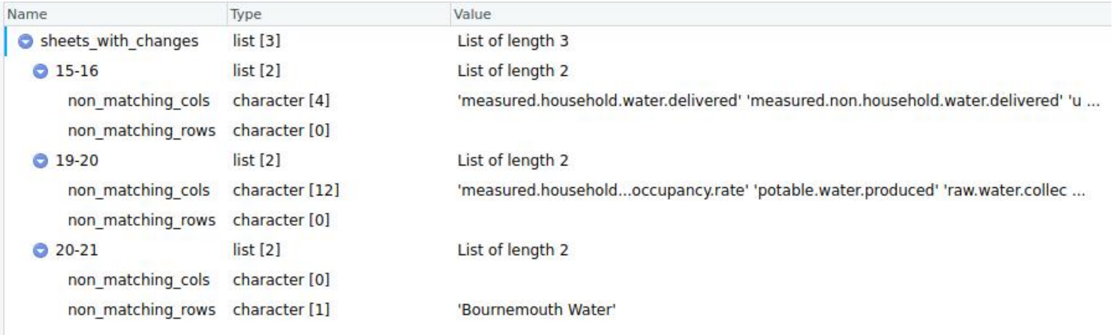
Figure 5.1: Successful start-up of the Docker containers.

5.1 Data Ingestion

The code for the Ingestion stage can be found in Appendix B.1. This section will only discuss key data wrangling steps.

Once the raw data in the Excel sheet is downloaded, each sheet is converted into its own dataframe. Each dataframe was then transposed so that instead of the company names being the column (feature) names, the actual attributes became the column names instead. Subsequently, all columns are converted into the numeric type, rather than the default character type. Finally, each sheet (dataframe) is added to a named vector; providing dictionary-like functionality. This essentially means each dataframe represents each respective time period.

Each dataframe was programmatically checked for semantic schema differences (using the code in Appendix B.1.2). Figure 5.2 illustrates the results of this. This does not account for more attributes added in later entries.

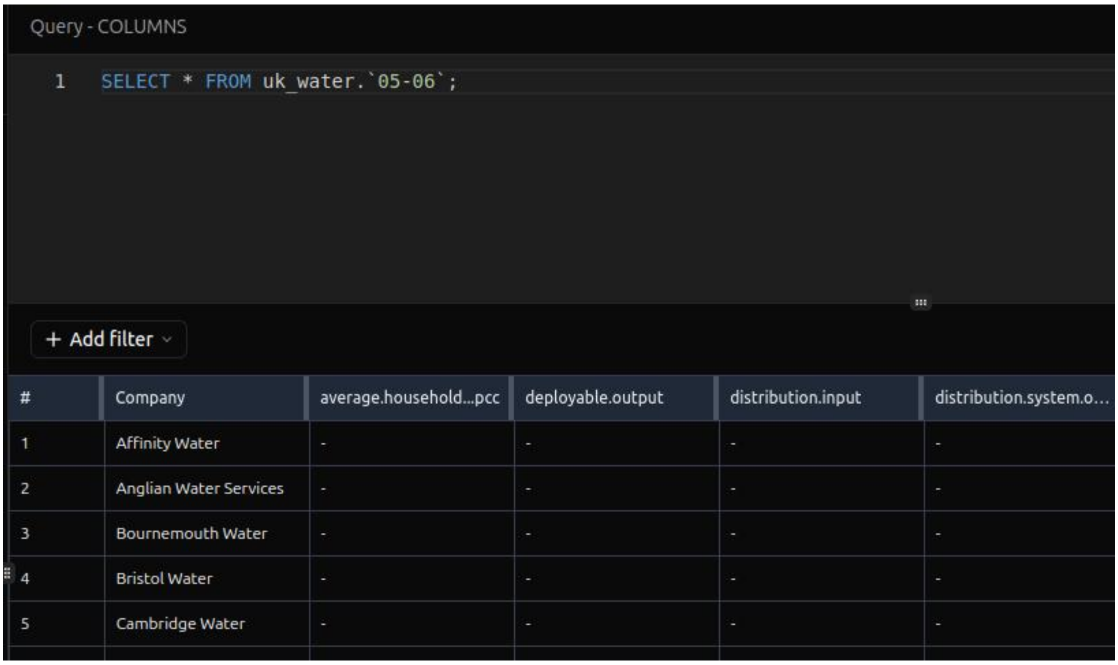


Name	Type	Value
sheets_with_changes	list [3]	List of length 3
15-16	list [2]	List of length 2
non_matching_cols	character [4]	'measured.household.water.delivered' 'measured.non.household.water.delivered' 'u ...
non_matching_rows	character [0]	
19-20	list [2]	List of length 2
non_matching_cols	character [12]	'measured.household...occupancy.rate' 'potable.water.produced' 'raw.water.collec ...
non_matching_rows	character [0]	
20-21	list [2]	List of length 2
non_matching_cols	character [0]	
non_matching_rows	character [1]	'Bournemouth Water'

Figure 5.2: Summary of semantic schema differences of the raw data.

As suggested in Figure 5.2, there are some column differences between some sheets. This should be taken into consideration when performing an analysis of the data (as an attribute in one sheet might not be present in another). However, it is problematic that Bournemouth Water is not present after '19-20'. To help prevent dataframe dimensionality inconsistencies, the Ingestion stage will be updated so that latter dataframes (after '19-20') include a Bournemouth Water row, which is made up of NA values. These values should be removed during calculations and on charts.

The attribute 'YearRef' was removed as the year the data refers to will be the table name within the ClickHouse database (as shown in Figure 5.3).

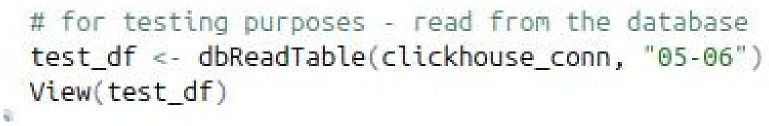


```
SELECT * FROM uk_water.`05-06`;
```

#	Company	average.household...pcc	deployable.output	distribution.input	distribution.system.o...
1	Affinity Water	-	-	-	-
2	Anglian Water Services	-	-	-	-
3	Bournemouth Water	-	-	-	-
4	Bristol Water	-	-	-	-
5	Cambridge Water	-	-	-	-

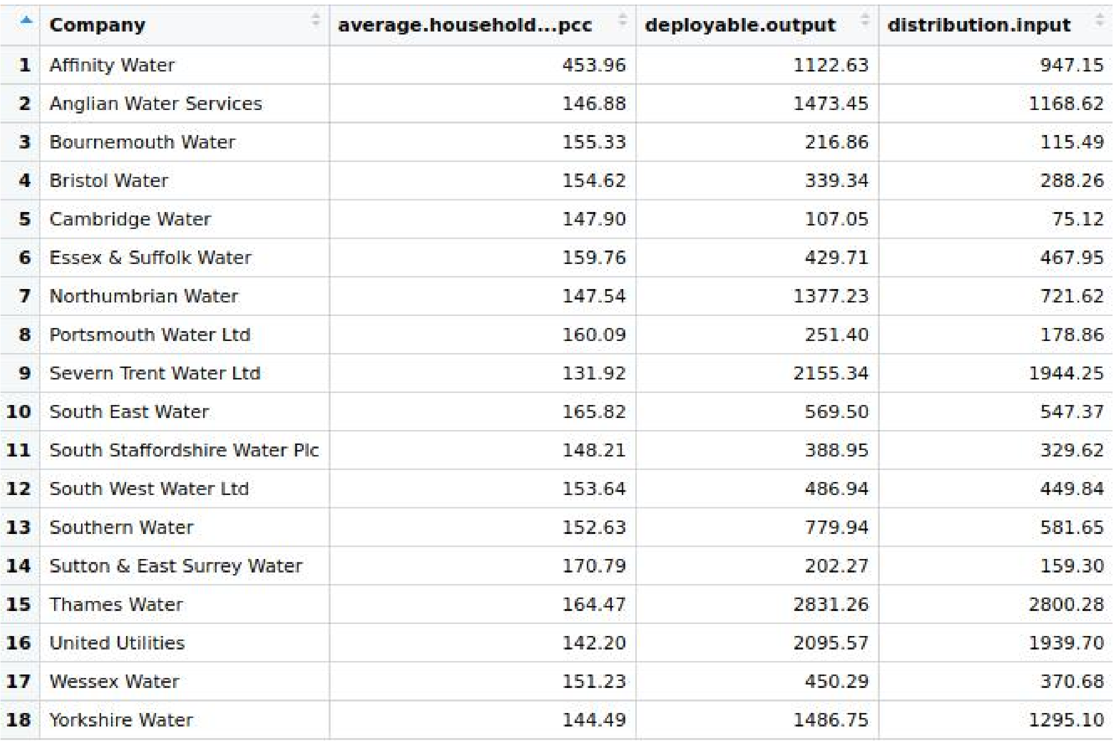
Figure 5.3: CH-UI view of a simple SQL query for the ’05-06’ table.

As indicated in Figure 5.3, due to some bugs with CH-UI, it was not able to show the stored data. However, Figure 5.4 confirms the data is stored within the database.



```
# for testing purposes - read from the database
test_df <- dbReadTable(clickhouse_conn, "05-06")
View(test_df)
```

(a) R code used to retrieve sample data from ClickHouse.



	Company	average.household...pcc	deployable.output	distribution.input
1	Affinity Water	453.96	1122.63	947.15
2	Anglian Water Services	146.88	1473.45	1168.62
3	Bournemouth Water	155.33	216.86	115.49
4	Bristol Water	154.62	339.34	288.26
5	Cambridge Water	147.90	107.05	75.12
6	Essex & Suffolk Water	159.76	429.71	467.95
7	Northumbrian Water	147.54	1377.23	721.62
8	Portsmouth Water Ltd	160.09	251.40	178.86
9	Severn Trent Water Ltd	131.92	2155.34	1944.25
10	South East Water	165.82	569.50	547.37
11	South Staffordshire Water Plc	148.21	388.95	329.62
12	South West Water Ltd	153.64	486.94	449.84
13	Southern Water	152.63	779.94	581.65
14	Sutton & East Surrey Water	170.79	202.27	159.30
15	Thames Water	164.47	2831.26	2800.28
16	United Utilities	142.20	2095.57	1939.70
17	Wessex Water	151.23	450.29	370.68
18	Yorkshire Water	144.49	1486.75	1295.10

(b) Table view of the results in RStudio (Posit, 2024).

Figure 5.4: Confirming the data upload to ClickHouse was successful.

6 Results and Discussion

The following subsections directly correspond to the questions in Table 3.1. As well, all code can be found in Appendix B.2.

6.1 QID 1: Water Demand

One way to determine water demand is to add the various ”consumption” recordings (refer to Table A.1), including measured; unmeasured; household; and non-household, together. Then, the mean average can be taken of these values for each time period, providing an overall estimated daily water demand for England and Wales. This is shown in Figure 6.1.

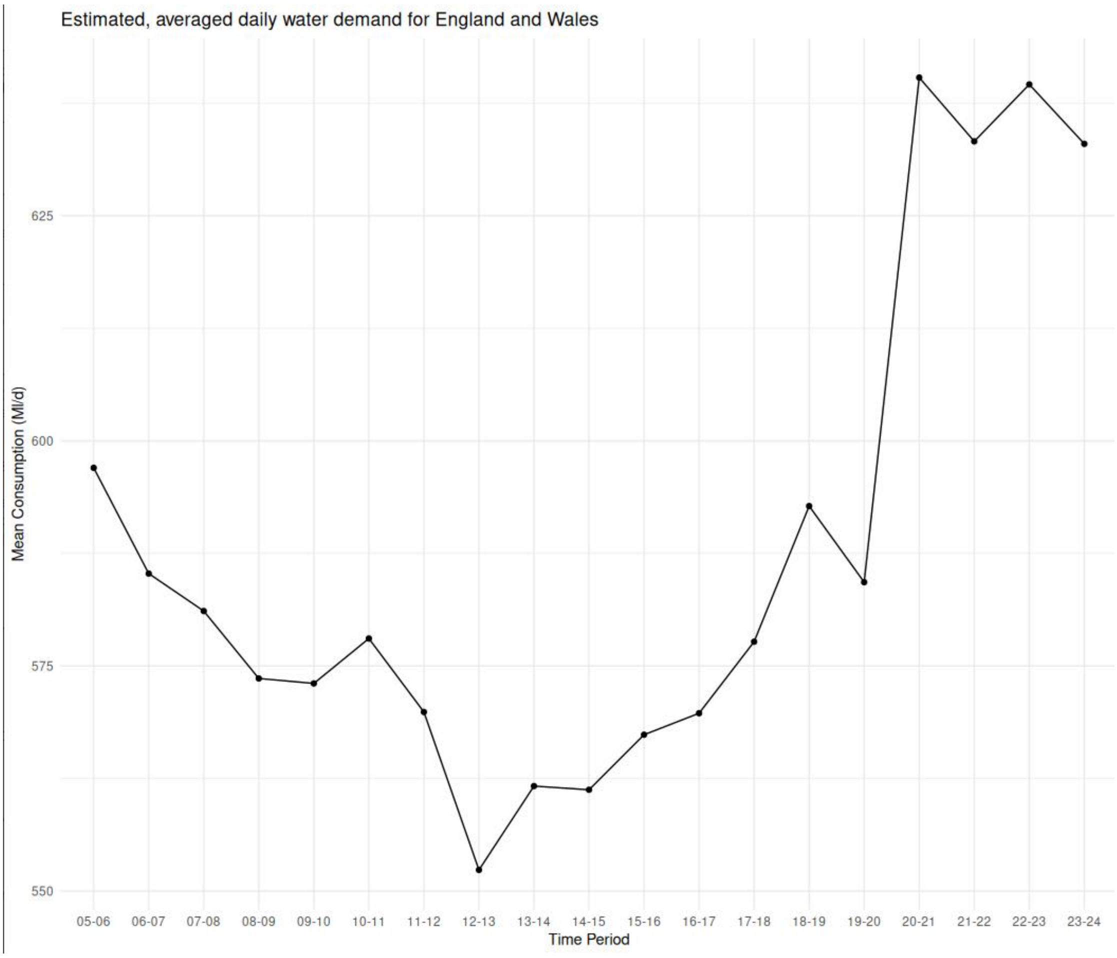
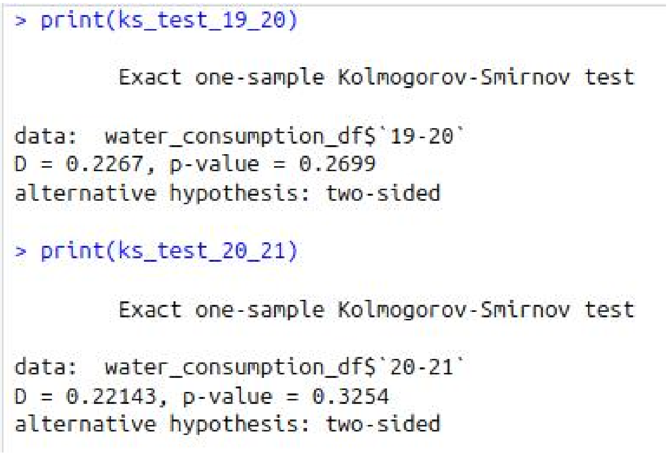


Figure 6.1: Estimated, averaged daily water demand for England and Wales.

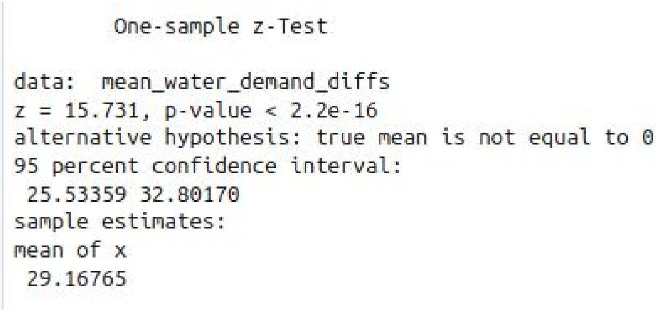
As shown in Figure 6.1, it is clear there is a steep increase in mean water usage after '19-20'. However, it is important to note these values do not include Bournemouth Water, as this information is missing. To help verify the statistical significance of this trend, hypothesis testing can be used. To confirm whether the samples are normally distributed, the Kolmogorov-Smirnov (K-S) Test will be used — which tests for the goodness-of-fit for a given distribution (e.g., normal distribution). The null hypothesis is that the distribution is normally distributed, while the alternative hypothesis is that the distribution is not normally distributed. Figure 6.2 shows the results for the sheets '19-20' and '20-21'.



```
> print(ks_test_19_20)

        Exact one-sample Kolmogorov-Smirnov test

data:  water_consumption_df$`19-20`
D = 0.2267, p-value = 0.2699
alternative hypothesis: two-sided

> print(ks_test_20_21)

        Exact one-sample Kolmogorov-Smirnov test

data:  water_consumption_df$`20-21`
D = 0.22143, p-value = 0.3254
alternative hypothesis: two-sided
```

Figure 6.2: K-S Test results for sheets '19-20' and '20-21'.

As indicated in Figure 6.2, both K-S Test results for both time periods have a p-value above the significance level (0.05). Additionally, the maximum difference for both does not exceed 0.23 (suggesting the cumulative difference to a normal distribution is not large). Hence, it has failed to reject the null hypothesis. This means the data is independent and identically distributed (IID). As the distributions are normally distributed and both the means and standard deviations are known, a Z-test can be used; specifically a paired Z-test as these samples are temporally related. For this, rows for "Bournemouth Water" will be removed, so these values do not affect the test. The results of this are shown in Figure 6.3.



```
        One-sample z-Test

data:  mean_water_demand_diffs
z = 15.731, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 25.53359 32.80170
sample estimates:
mean of x
 29.16765
```

Figure 6.3: Paired Z-test for evaluating the significance of mean differences between sheets '19-20' and '20-21'.

As indicated in Figure 6.3, the p-value is 2.2×10^{-16}, far less than the 5% significance value. Therefore, the null hypothesis is to be rejected. This implies the mean differences between the samples is significant, helping verify water demand within the past 4 years to date has increased significantly across England and Wales.

6.2 QID 2: Leakages Over Time

Figure 6.4 illustrates the mean daily leakage across all water companies in England and Wales.

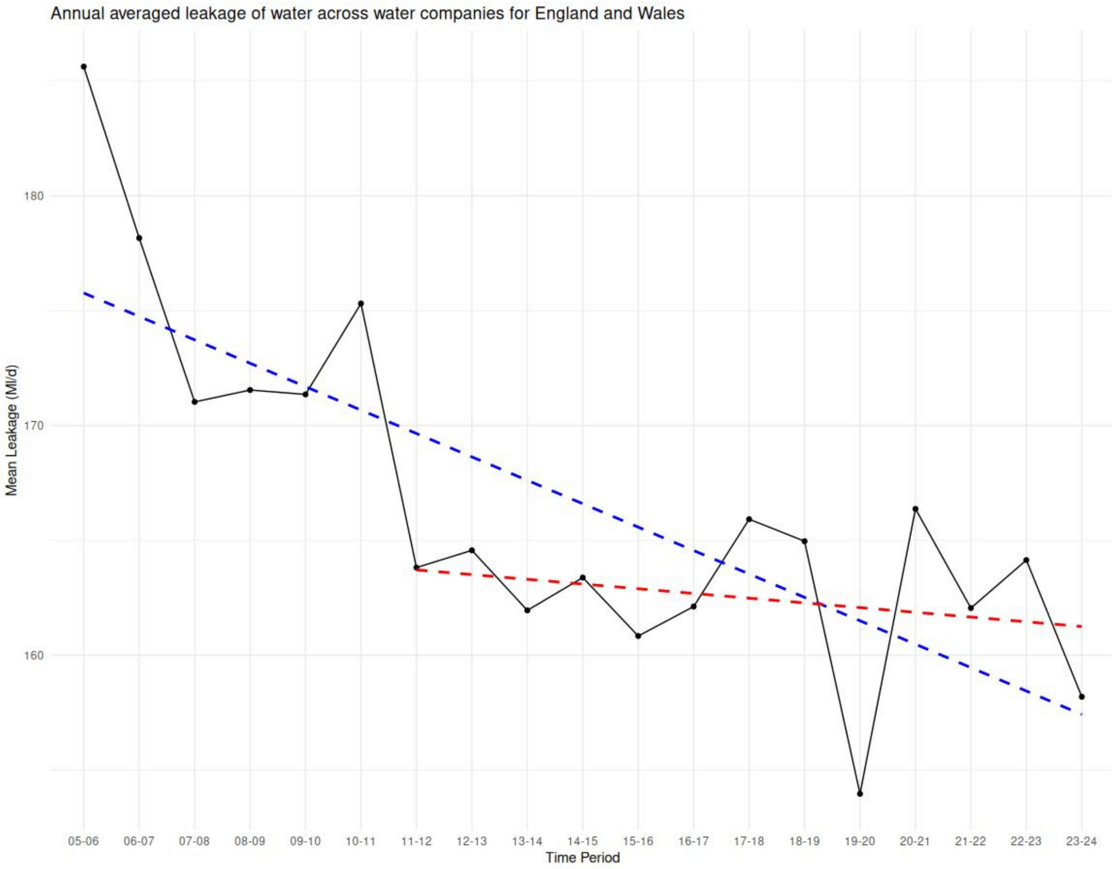


Figure 6.4: Annual mean leakage of consumable water across the whole Water Industry.

The blue regression line in Figure 6.4 shows that all water companies across England and Wales have gradually decreased the amount of consumable water wasted due to leakages. This does not include sewage handling (as this is not available in the dataset). However, the red regression line focusing on the time periods ’11-12’ to ’23-24’ shows that progress has been slow over the past decade by the Water Industry for reducing leakage further. Table 6.1 regards summary data for the time period ’23-24’.

Table 6.1: Summary statistics for leakages for the sheet '23-24'.

Measure of Central Tendency or Dispersion	**Value (2 d.p.)**
Minimum	12.29
1st quartile	50.82
Median	107.48
Mean	158.19
3rd quartile	182.87
Max	572.86

Table 6.1 demonstrates that there is a large discrepancy between the highest leakage-prone (maxima) companies and the lowest leakage-prone companies (minima). One relationship which could be key in explaining this difference is between the daily water leakage and the respective populations for each company, as shown in Figure 6.5.

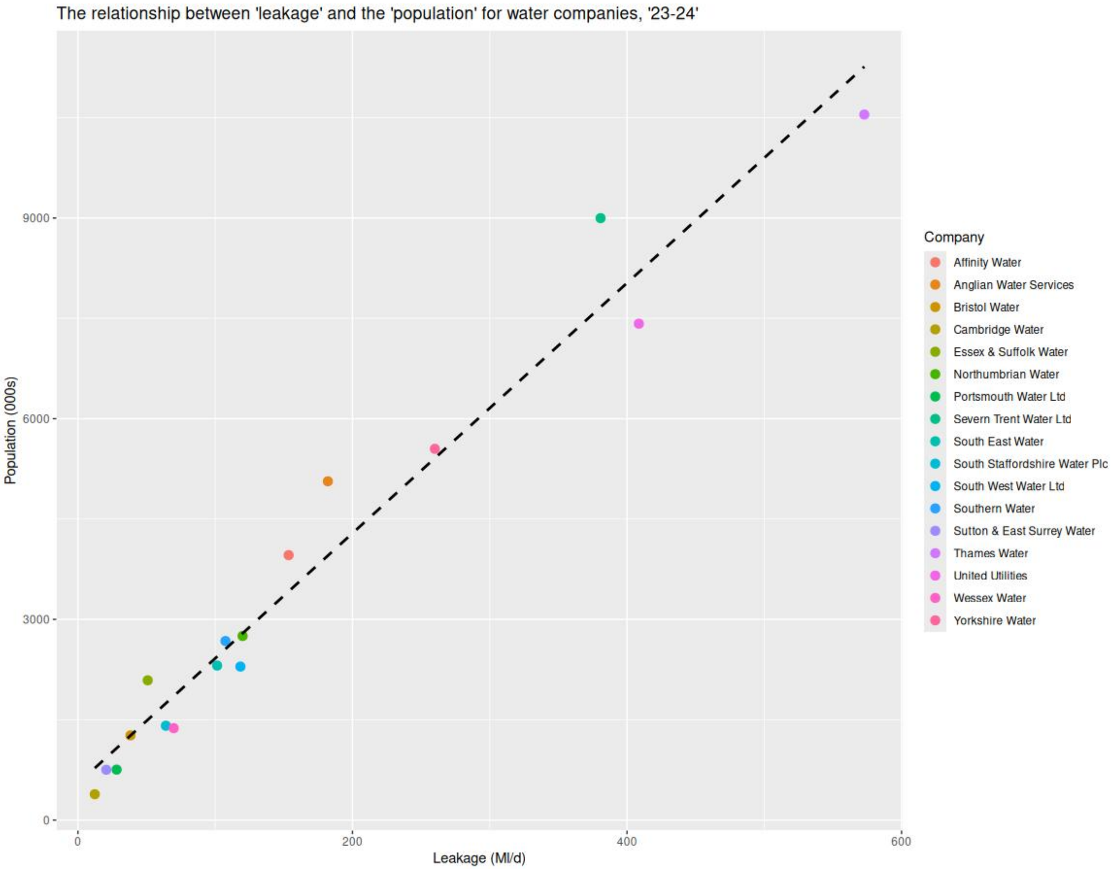


Figure 6.5: The relationship between 'leakage' and 'total population' for each company, '23-24'.

The black dashed linear regression line in Figure 6.5 illustrates a strong positive correlation between 'leakage' and 'total population'. This suggests that the bigger the population (number of customers) a company has, the more likely they are to experience more leakage of water per day.

6.3 QID 3: Water Availability By Company

Water availability on its own can provide an idea of the scale of the company, but together with the respective population really puts this scale into perspective — of how well companies make water available depending on its obligations to its customers. This is shown in Figure 6.6.

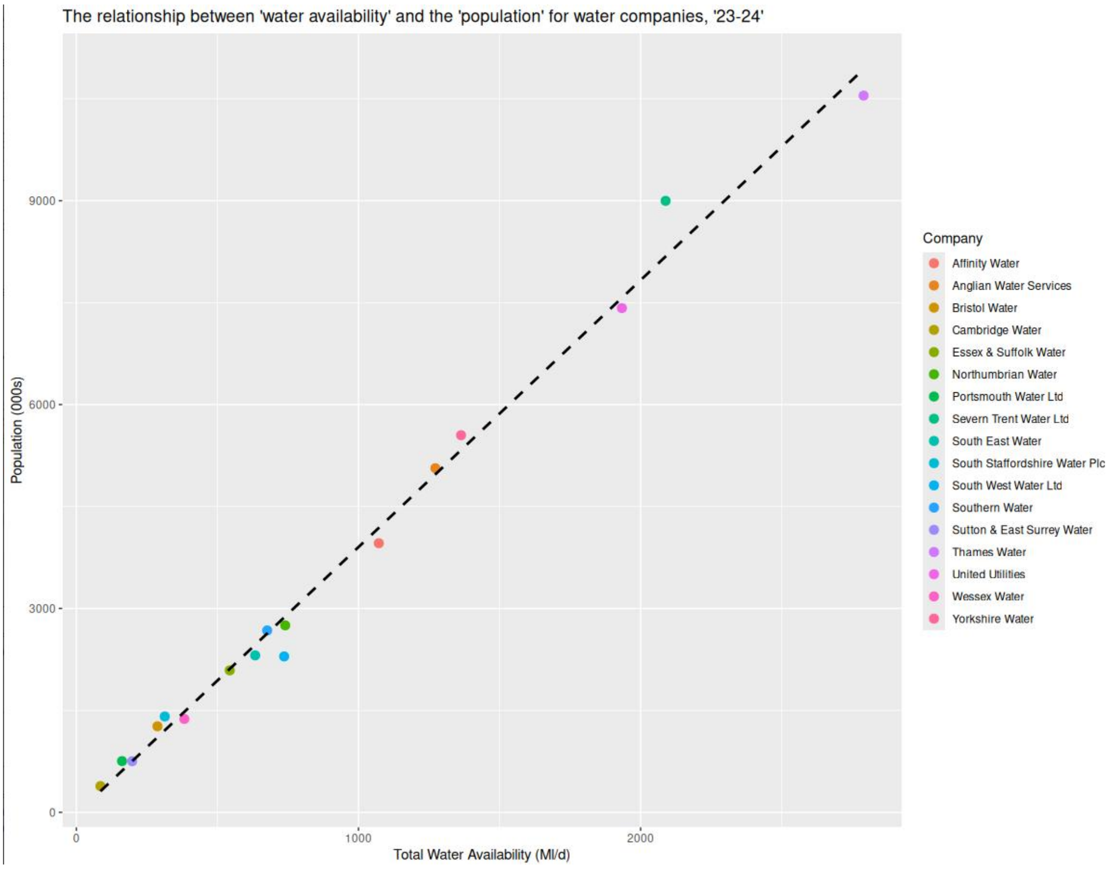
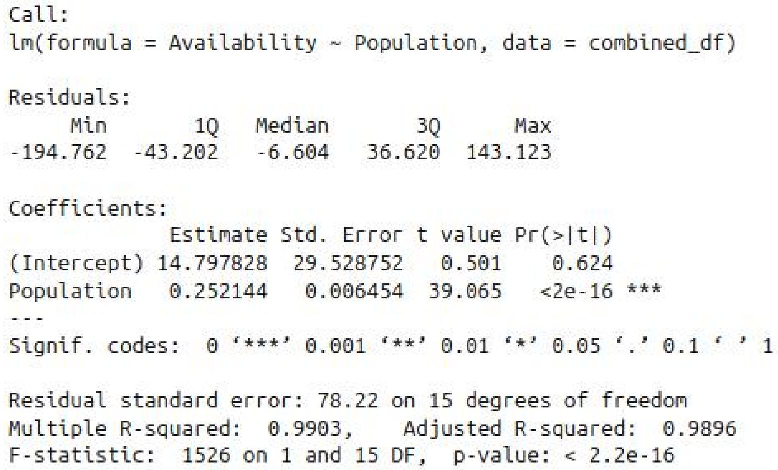


Figure 6.6: The relationship between 'total water availability' and 'total population' for each company, '23-24'.

As indicated in Figure 6.6, there is a strong linear relationship between the total water available and the size of the population (of customers) for every company. This is confirmed, in part, by the high Adjusted R^2 score and very low p-value in Figure 6.7, which summarises the evaluation metrics of the linear line shown in Figure 6.6. This linear line could be used for purposes of predicting how much water should be available for a given population. Although, it is important to note that this trend across the industry does not necessarily mean the respective company has reached the ideal levels of availability for the consumer.



```
Call:
lm(formula = Availability ~ Population, data = combined_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-194.762  -43.202   -6.604   36.620  143.123 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 14.797828  29.528752   0.501    0.624    
Population   0.252144   0.006454  39.065   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 78.22 on 15 degrees of freedom
Multiple R-squared:  0.9903,	Adjusted R-squared:  0.9896 
F-statistic:  1526 on 1 and 15 DF,  p-value: < 2.2e-16
```

Figure 6.7: Summary of the linear model statistics for plotting ’total water availability’ and ’total population’ across companies.

6.4 QID 4: Individual Water Consumption

Figure 6.8 shows the mean averaged daily water consumption PCC across England and Wales, but this also includes two water companies from different location: Thames Water (South East) and United Utilities (North West).

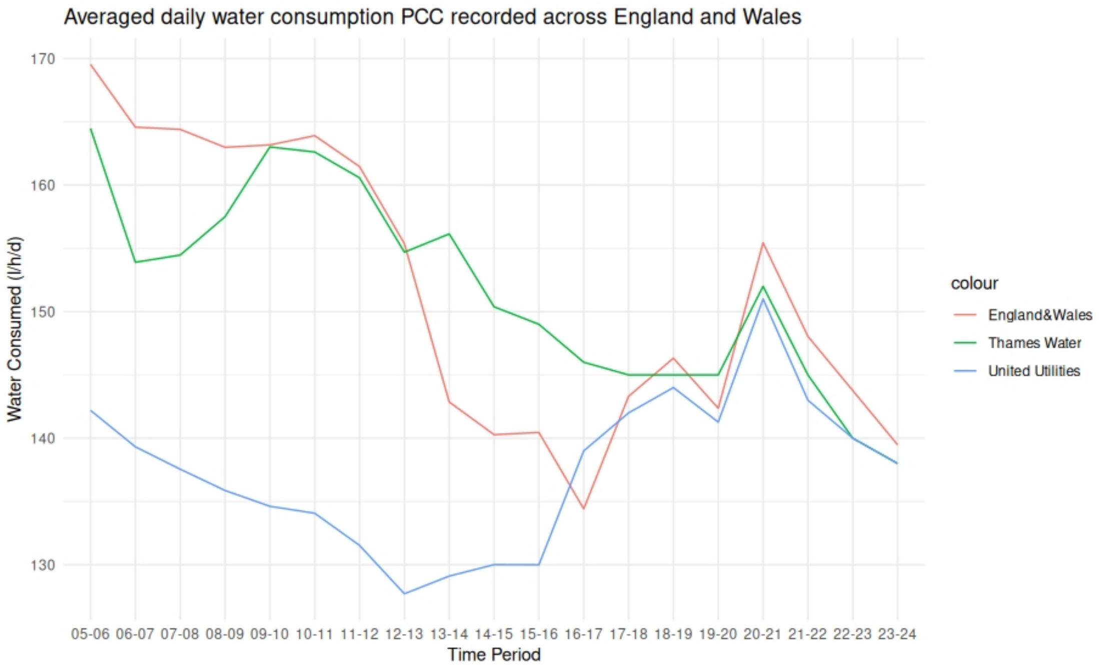


Figure 6.8: Mean water consumption PCC for England and Wales, Thames Water, and United Utilities; since '05-06'.

Interestingly, the levels of consumption in Figure 6.8 vary significantly before '20-21'. More north of England, consumers used significantly less water compared to consumers down south. In '20-21', the consumption converges towards the nationwide average, likely due to the 2021 heatwave (Met Office, 2021). This demonstrates how water companies might have to supply varying amounts of water and operate differently depending on which part of the country they supply with water.

6.5 QID 5: Raw Water Imports

Figure 6.9 shows the raw water imported for '13-14', while Figure 6.10 shows the raw water imported for '23-24'. Companies who imported no raw water were removed from the charts to help make them more concise and easier to compare. Table 6.2 shows the summary statistics for '13-14', while Table 6.3 shows the summary statistics for '23-24'.

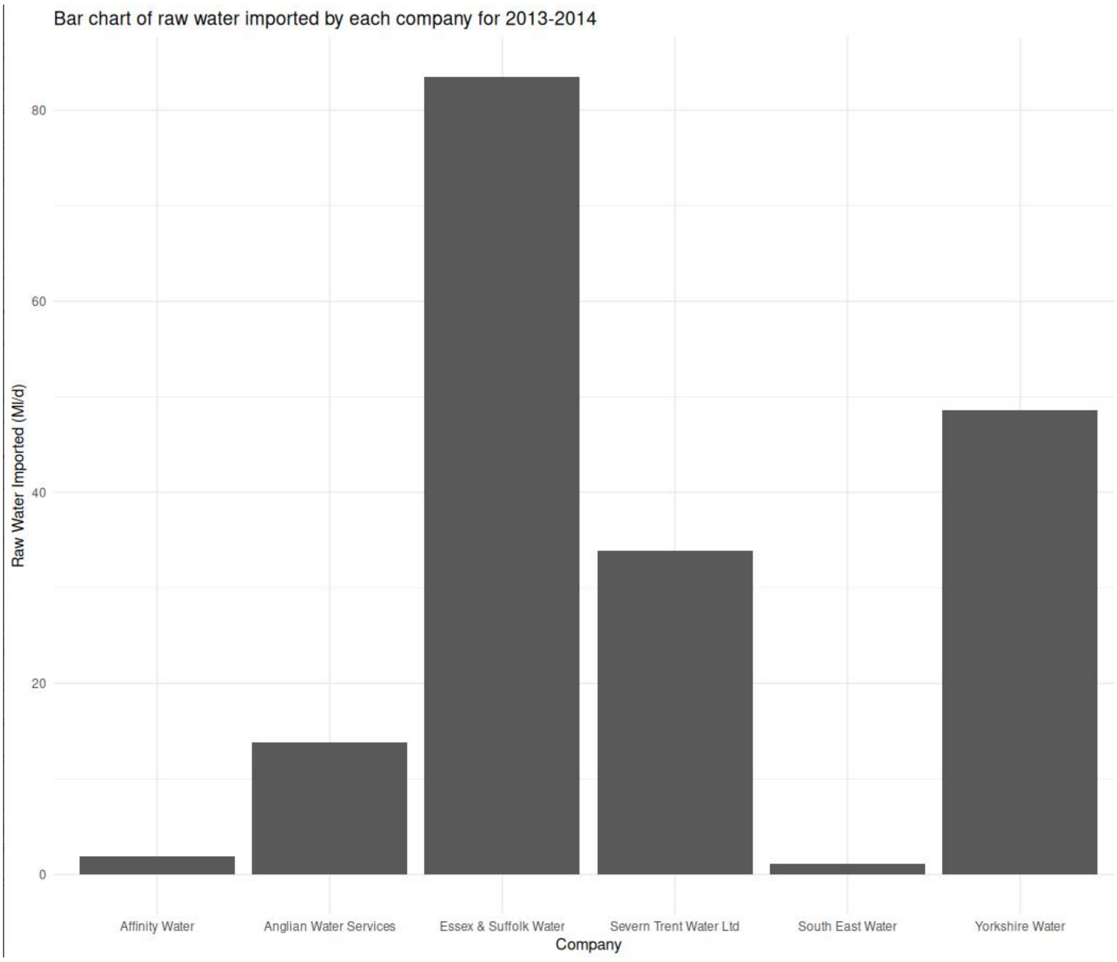


Figure 6.9: Raw water imported for the sheet '13-14'.

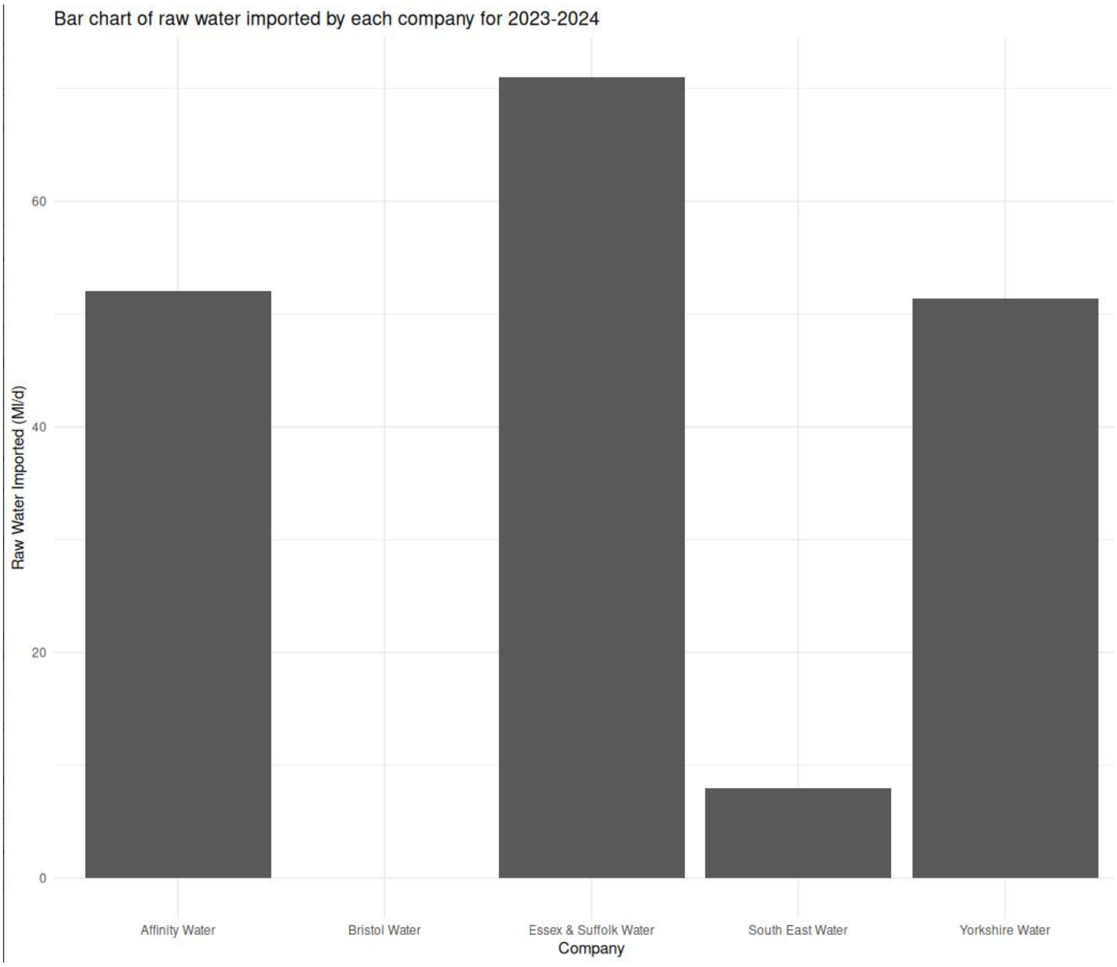


Figure 6.10: Raw water imported for the sheet ’23-24’.

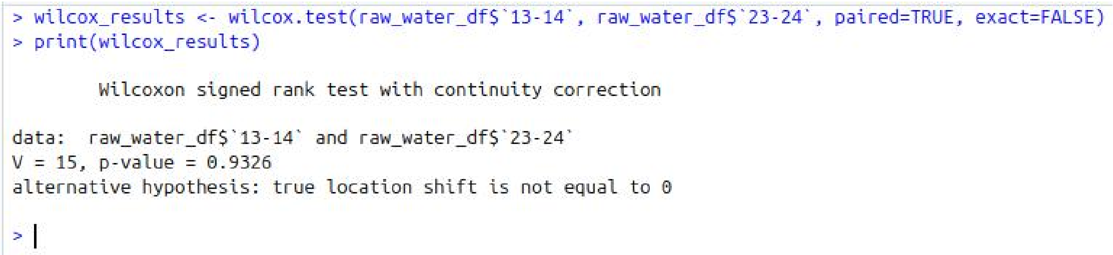
Table 6.2: Summary statistics for raw water imported for the sheet '13-14'.

Measure of Central Tendency or Dispersion	**Value (3 d.p.)**
Minimum	0.000
1st quartile	0.000
Median	0.000
Mean	10.156
3rd quartile	1.700
Max	83.480

Table 6.3: Summary statistics for raw water imported for the sheet '23-24'.

Measure of Central Tendency or Dispersion	**Value (3 d.p.)**
Minimum	0.000
1st quartile	0.000
Median	0.000
Mean	10.73
3rd quartile	0.82
Max	71.00

Based on Figures 6.9 and 6.9, it is clear the distributions are non-parametric. Moreover, based on Tables 6.2 and 6.3, it appears only a select number of companies import fairly substantial amounts of raw water. To confirm the differences between '13-14' and '23-24', is not statistically significant, the Wilcoxon Signed-Rank Test will be used. This test can be applied on non-parametric distributions and is meant for comparing paired samples. Additionally, the null hypothesis is that the median difference between paired observations is zero, while the alternative hypothesis is that the median difference between them is not zero. However, to account for the zero values in the data for some companies, continuity correction will be applied to the calculation, which enables for an approximate p-value, but unfortunately can have a large difference to the actual p-value. Figure 6.11 shows these results.



```
> wilcox_results <- wilcox.test(raw_water_df$`13-14`, raw_water_df$`23-24`, paired=TRUE, exact=FALSE)
> print(wilcox_results)

        Wilcoxon signed rank test with continuity correction

data:  raw_water_df$`13-14` and raw_water_df$`23-24`
V = 15, p-value = 0.9326
alternative hypothesis: true location shift is not equal to 0

> |
```

Figure 6.11: Wilcox Signed-Rank Test results for raw water imported data, for sheets '13-14' and '23-24'.

Figure 6.11 indicates the approximate p-value to be 0.9326, which is well above the required significance level (0.05), and therefore, this has failed to reject the null hypothesis. This suggests the differences in raw water imported a decade ago compared to present is not statistically significant. Therefore, it can be assumed that the Water Industry as a whole in England and Wales are not more dependent on foreign nations for raw water today compared to a decade ago.

6.6 QID 6: Water Treatment Efficiency

Figure 6.12 illustrates the annual mean daily average of water lost due to treatment works on raw water and operational use (e.g., cleaning machines, transportation of raw water). It is clear that after '10-11', more water was being lost for treatment works — an estimated increase of over 130%. This trend then generally continued to present. However, it may only be a minority of companies that had drastically increased water loss (possibly due to historic or faulty infrastructure), rather than a small increase in all of the companies. To confirm this, some hypothesis testing steps taken in Section 6.1 were repeated here. Figure 6.13 illustrates the results of performing the K-S Test on sheets '10-11' and '11-12'.

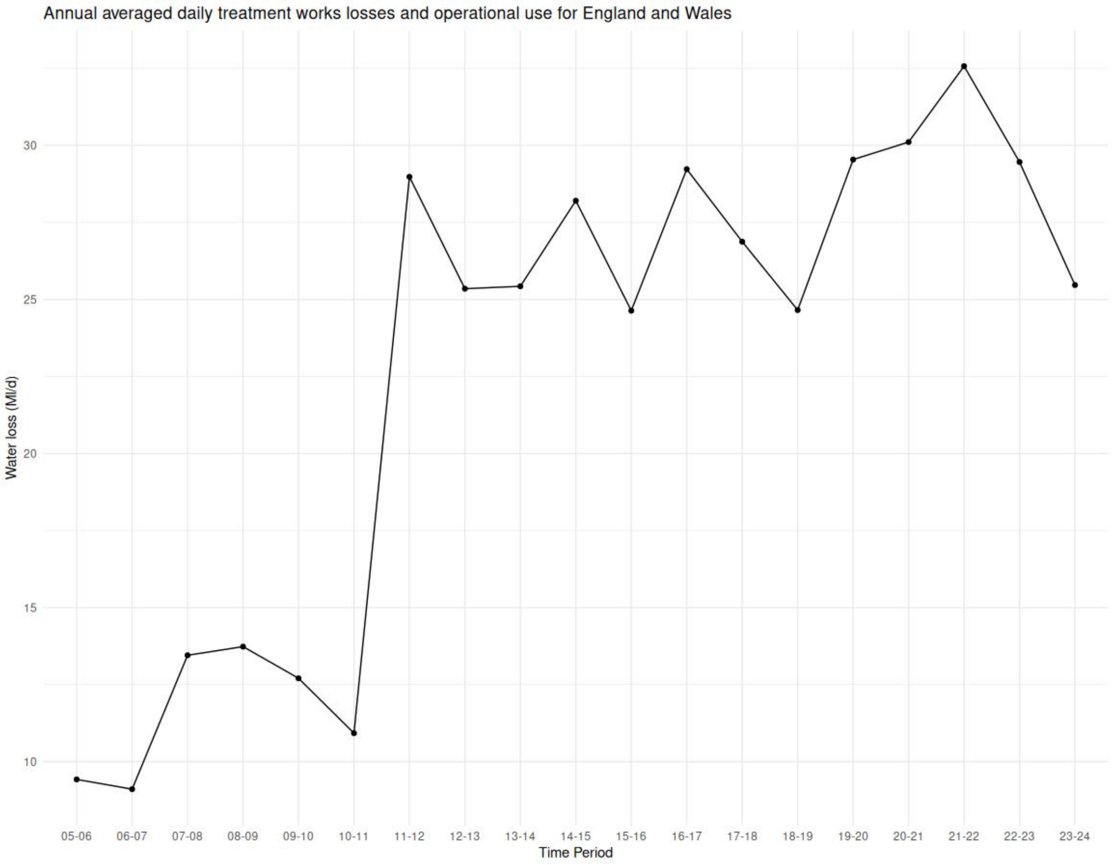
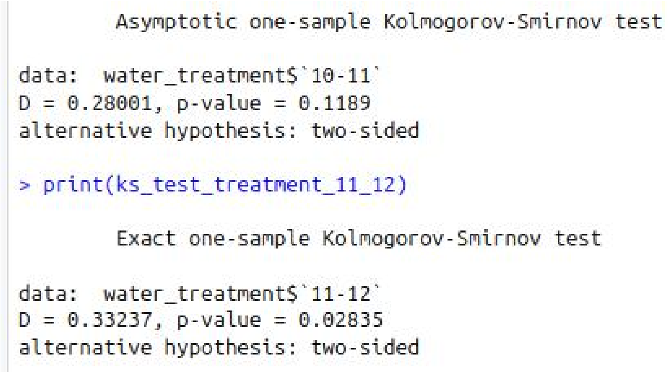


Figure 6.12: Annual averaged daily water loss due to treatment works losses and operational use.



```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  water_treatment$`10-11`
D = 0.28001, p-value = 0.1189
alternative hypothesis: two-sided

> print(ks_test_treatment_11_12)

        Exact one-sample Kolmogorov-Smirnov test

data:  water_treatment$`11-12`
D = 0.33237, p-value = 0.02835
alternative hypothesis: two-sided
```

Figure 6.13: K-S Test results for sheets '10-11' and '11-12' for treatment losses and operational works.

Figure 6.13 confirms the distributions are likely not normally distributed (as the p-value for '10-11' is less than the 2.5% significance level for a two-tailed test, and the maximum difference is about 0.3 in both). Therefore, it is likely that rather than the average water company having increased treatment losses, it is a specific selection of them which exacerbated the mean averages. Overall, this suggests general water companies have maintained fairly consistent treatment works losses, which is good considering the rise in demand over time (Section 6.1).

6.7 QID 7: Water Outage Predictions

Normally, companies strive to maintain their services as much as possible. But this is not always possible. For instance, Figure 6.14 shows the mean averaged daily water outage recorded for all water companies across England and Wales.

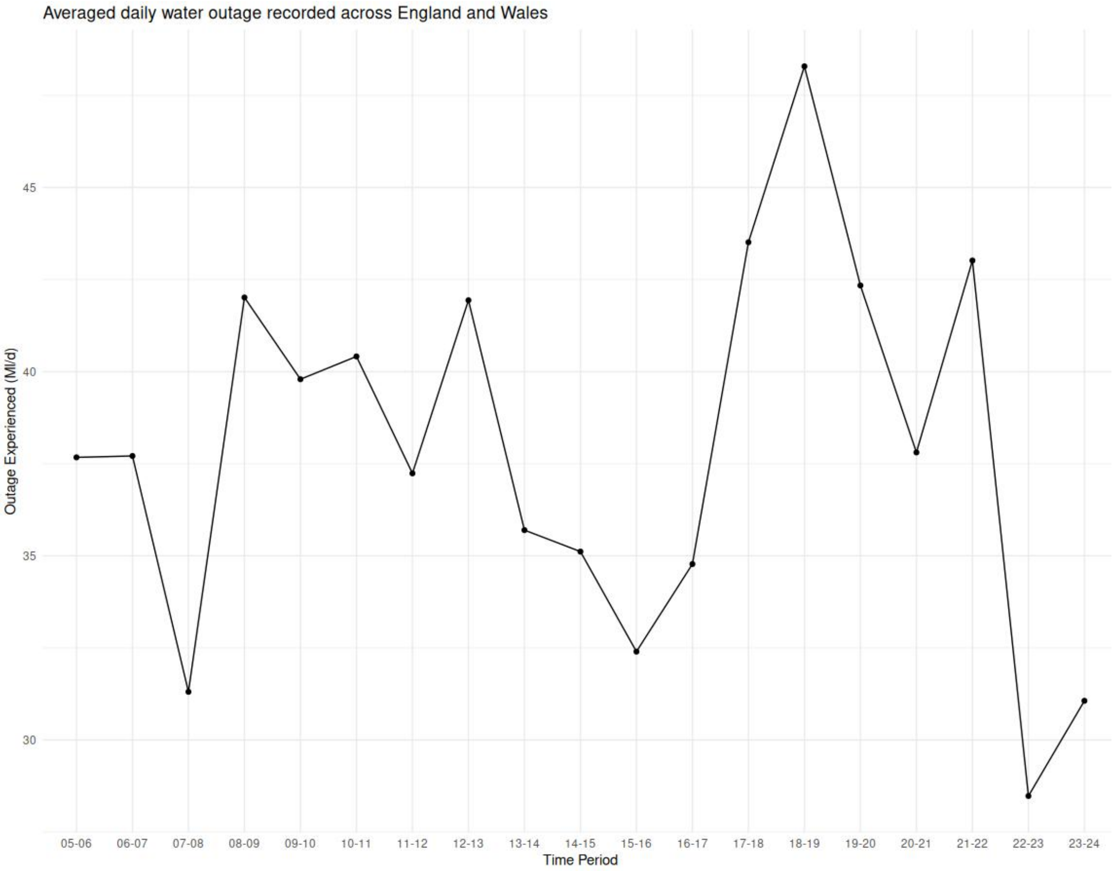


Figure 6.14: Mean averaged daily water outage recorded across England and Wales.

The results in Figure 6.14 are fairly erratic — which makes sense, as there are numerous hidden factors for each company that influences these results. Nonetheless, techniques such as Machine Learning can be applied to this data to make "best guess" predictions. One model in particular is Autoregressive Integrated Moving Average (ARIMA), which is commonly used for making future predictions using past time-series data. ARIMA takes into account the temporal relationship in the training data. Figure 6.15 shows Figure 6.14 but with the ARIMA predictions for '25-26'.

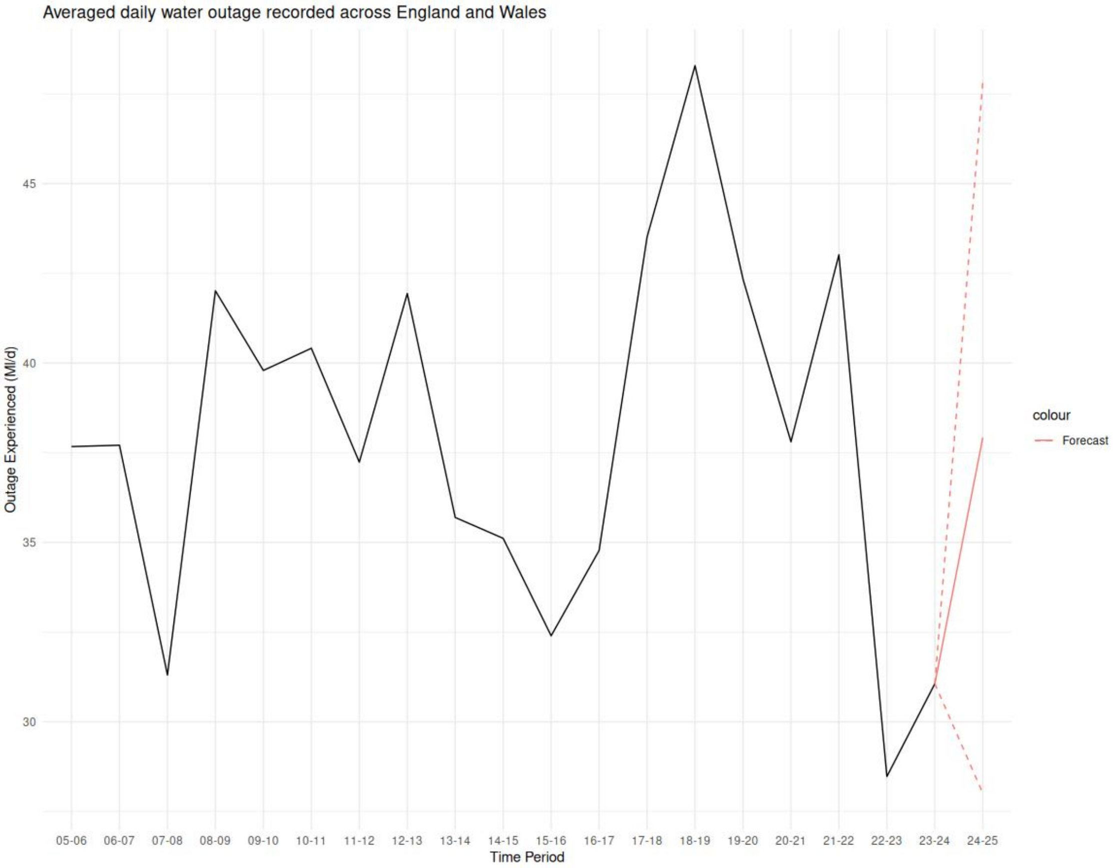


Figure 6.15: Water outage prediction for '25-26' using ARIMA.

The solid red line in Figure 6.15 is the forecast result for applying the ARIMA model on the water outage data. The red dashed lines is the lower and upper bound of the 95% prediction interval (so it is 95% likely that the next value will be within the prediction interval). Based on this model, it seems water outages are likely to increase nationally over the next year or so.

7 Conclusions and Future Work

In summary, it is apparent water companies have made significant progress in improving services and their performance since 2005. Although, over the past decade, progress has slowed. Along with increased demand, this has lead to issues in water leakages, raw water treatment losses, and water outages across for many consumers across the Water Industry.

Future work includes investigating how some water companies handle sewage and waste water, with respect to the environment and efficiency. Additionally, customer satisfaction can also be investigated with respect to a water company's performance.

Reference List

Clickhouse Inc. (2024). *Fast open-source OLAP database*. Available at: https://clickhouse.com/ [Accessed on 4th Dec. 2024].

Docker Inc. (2024a). *Accelerated containerisation software*. Available at: https://www.docker.com/ [Accessed on 4th Dec. 2024].

Docker Inc. (2024b). *Docker Compose*. Available at: https://docs.docker.com/compose/ [Accessed on 15th Dec. 2024].

Godard, P. (2024). *ClickHouseHTTP*. Available at: https://github.com/patzaw/ClickHouseHTTP [Accessed on 15th Dec. 2024].

Goh, K. H. and See, K. F. (2021). Twenty years of water utility benchmarking: A bibliometric analysis of emerging interest in water research and collaboration. *Journal of Cleaner Production*, 284, p. 124711. Available at: 10.1016/j.jclepro.2020.124711.

Masud, F. (2024). All water firms now investigated for sewage spills. *BBC*. 16 July. Available at: https://www.bbc.co.uk/news/articles/crg5ev7e46mo#:~:text=All%2011%20water%20and%20wastewater%20companies%20in%20England,the%20regulator%20said%20it%20was%20expanding%20its%20investigation. [Accessed on 6th Nov. 2024].

Met Office (2021). *Heatwave helps mark fifth warmest July on record*. Available at: https://www.metoffice.gov.uk/blog/2021/heatwave-helps-mark-fifth-warmest-july-on-record [Accessed on 30th Dec. 2024].

Molinos-Senante, M. and Maziotis, A. (2018). Assessing the influence of exogenous and quality of service variables on water companies´ performance using a true-fixed stochastic frontier approach. *Urban Water Journal*, 15 (7), pp. 682–691. Available at: 10.1080/1573062X.2018.1539502.

Posit (2024). *RStudio Desktop*. Available at: https://posit.co/download/rstudio-desktop/ [Accessed on 15th Dec. 2024].

Purnell, S., Mills, N., Davis, K. et al. (2021). Assessment of the pollution incident performance of water and sewerage companies in England. *PLoS One*, 16 (10), pp. 1–18. Available at: 10.1371/journal.pone.0251104.

Ricciuti, C. (2024). *CH-UI*. Available at: https://github.com/caioricciuti/ch-ui [Accessed on 15th Dec. 2024].

Sala-Garrido, R., Mocholi-Arce, M., Molinos-Senante, M. et al. (2021). Eco-Efficiency of the English and Welsh Water Companies: A Cross Performance Assessment. *International Journal of Environmental Research and Public Health*, 18, p. 2831. Available at: 10.3390/ijerph18062831.

Stallard, E. (2024). Water firms to be punished for years of sewage leaks. *BBC*. 6 August. Available at: https://www.bbc.co.uk/news/articles/c5ypp032le0o [Accessed on 6th Nov. 2024].

The Document Foundation (2024). *LibreOffice (24.8.3.2)*. [computer software]. Available at: https://www.libreoffice.org/ [Accessed on 1st Dec. 2024].

Thomas, S. (2019). Is the ideal of independent regulation appropriate? Evidence from the United Kingdom. *Competition and Regulation in Network Industries*, 20 (3), pp. 218–228. Available at: https://doi-org.bcu.idm.oclc.org/10.1177/1783591719836875.

UK Government (2024a). *Water Resource Management Plan Annual Review Data*. Available at: https://www.data.gov.uk/dataset/87b59684-3da3-45cf-8881-e4727cfd1415/water-resource-management-plan-annual-review-data [Accessed on 30th Nov. 2024].

UK Government (2024b). *Water resource management plan supply-demand balance forecast collated data*. Available at: https://www.data.gov.uk/dataset/4eccde92-ca38-4e7e-b4b7-45a95079caf5/water-resource-management-plan-supply-demand-balance-forecast-collated-data [Accessed on 30th Nov. 2024].

Villegas, A., Molinos-Senante, M. and Maziotis, A. (2019). Impact of environmental variables on the efficiency of water companies in England and Wales: a double-bootstrap approach. *Environment Science and Pollution Research*, 26, pp. 31014–31025. Available at: 10.1007/s11356-019-06238-z.

Walker, N. L., Styles, D., Gallagher, J. et al. (2021). Aligning efficiency benchmarking with sustainable outcomes in the United Kingdom water sector. *Journal of Environmental Management*, 287, p. 112317. Available at: 10.1016/j.jenvman.2021.112317.

Walker, N. L., Williams, A. P. and Styles, D. (2020). Key performance indicators to explain energy & economic efficiency across water utilities, and identifying suitable proxies. *Journal of Environmental Management*, 269, p. 110810. Available at: 10.1016/j.jenvman.2020.110810.

Appendix A Dataset Attributes

Table A.1: Dataset attributes.

Attribute name	Description	Unit
Average household - pcc	Estimated average per capita consumption for household use, both measured and unmeasured.	l/h/d
Deployable output	The available supply of water output, constrained by certain factors (e.g., environment, drought measures, treatment, etc.)	Ml/d
Distribution input	The amount of water entering the distribution system at the point of production.	Ml/d
Distribution system operational use	The water knowingly used by a water company to meet its obligations.	Ml/d
Measured households - consumption	The average measured volume of water supplied to households.	Ml/d
Measured households - occupancy rate	The average number of occupants per household, supplied with measured water.	N/A
Measured households - PCC	An estimated per capita consumption of households that are supplied with measured water.	Ml/d
Measured households - population	The population of the covered domestic properties that consume measured water. Measured in thousands.	N/A
Measured households - properties	The number of properties that are supplied with measured water. Measured in thousands.	Ml/d
Measured household - USPL	Estimated underground supply pipe leakage for households that are supplied with measured water.	Ml/d
Measured household water delivered	The average volume of water delivered to households that are supplied with measured water. This should include USPL.	Ml/d
Measured non-households - consumption	The average measured volume of water supplied to non-households.	Ml/d

Continuation of Table A.1

Attribute name	**Description**	**Unit**
Measured non-households - population	The population of the covered non-domestic properties that consume measured water. Measured in Thousands.	Ml/d
Measured non-households - properties	The number of non-domestic properties that are supplied with measured water. Measured in thousands.	N/A
Measured non-households - USPL	Estimated underground supply pipe leakage for non-households that are supplied with measured water.	Ml/d
Measured non-households water delivered	The average volume of water delivered to non-households that are supplied with measured water. This should include USPL.	Ml/d
Non-potable water supplied	The volume of water supplied that does not meet drinking standards.	Ml/d
Outage experienced	The average volume of water inaccessible during a period of outage. This is typically somewhat recoverable.	Ml/d
Potable water exported	The volume of drinking water exported to foreign countries/nations.	Ml/d
Potable water imported	The volume of drinking water imported from foreign countries/nations.	Ml/d
Potable water produced	The total volume of drinking water produced from raw water (both imported and exported).	Ml/d
Raw water abstracted	The amount of raw water collected, including imported raw water minus exported raw water.	Ml/d
Raw water collected	The amount of raw water collected.	Ml/d
Raw water exported	The amount of raw water exported to foreign countries/nations.	Ml/d
Raw water imported	The amount of raw water imported from foreign countries/nations.	Ml/d
Raw water into treatment	The amount of raw water that has been treated.	Ml/d

Continuation of Table A.1

Attribute name	**Description**	**Unit**
Raw water Losses and Operational Use	The volume of raw water lost during its transportation, and the volume of raw water used to clean and maintain raw water extraction mechanisms.	Ml/d
Raw water retained	The volume of raw water that was treated minus the volume of raw water lost or already used.	Ml/d
Total Household Metering penetration (incl. voids)	The percentage of total households that are measured with a water meter and pay for water on a volumetric basis. This includes void properties.	Ml/d
Total leakage	The sum of distribution losses underground and supply pipe losses. Measured in Megalitres per day.	Ml/d
Total leakage	The sum of distribution losses underground and supply pipe losses. Measured in Litres per property per day.	l/prop/d
Total mains and trunk mains leakage (Distribution Losses)	The volume of water lost via trunk mains, service reservoirs, etc. Distribution losses are distribution input minus the water taken.	Ml/d
Total population	An estimated total population that the water company supplies water to.	N/A
Total properties	An estimated total number of properties that the water company supplies water to.	N/A
Treatment works losses and operational use	The amount of processed raw water lost during treatment and for operational purposes.	Ml/d

Continuation of Table A.1

Attribute name	**Description**	**Unit**
Unmeasured household - consumption	An estimated average volume of water supplied to households that do not have a water meter fitted.	Ml/d
Unmeasured household - occupancy rate	An estimated average occupancy of households that do not have a water meter fitted.	Ml/d
Unmeasured household - pcc	An estimated per capita consumption of households without a water meter, that are supplied with water.	Ml/d
Unmeasured household - population	An estimation of the population of the covered area that are supplied with water, from households that do not have a water meter. Measured in Thousands.	Ml/d
Unmeasured household - properties	The number of non-domestic properties that are supplied with water but do not have a water meter installed.	Ml/d
Unmeasured household - USPL	Estimated underground supply pipe leakage for households that are supplied with unmeasured water.	Ml/d
Unmeasured household water delivered	Estimated average volume of water delivered to households that do not have a water meter. This should include USPL.	Ml/d
Unmeasured non-household - consumption	An estimated average of the volume of water supplied to non-households that do not have a water meter fitted.	Ml/d
Unmeasured non-household - properties	The number of non-domestic properties that are supplied with unmeasured water.	Ml/d
Unmeasured non-household - population	The population of the non-domestic properties that are supplied with unmeasured water. Measured in Thousands.	Ml/d

Continuation of Table A.1

Attribute name	Description	Unit
Unmeasured non-household - water delivered	The average volume of water delivered to non-households that do not have a water meter fitted. This should include USPL.	Ml/d
Unmeasured non-household - USPL	Estimated underground supply pipe leakage for non-households that are supplied with unmeasured water.	Ml/d
Unmeasured void households - properties	The number of non-occupied domestic properties that are supplied with water, but do not have a water meter.	Ml/d
Void non-households - properties	The number of non-occupied, non-domestic properties that are supplied with water.	Ml/d
Void properties - USPL	Estimated underground supply pipe leakage for void households and non-household.	Ml/d
Water available for use	The actual amount of water available for use minus losses and leakages.	Ml/d
Water delivered	The average total volume of water delivered, which should also include USPL.	Ml/d
Water delivered billed	How much of the total volume of water delivered was billed.	Ml/d
Water taken	The water delivered plus any additional water consumption.	Ml/d
Water taken unbilled	Water taken legally unbilled (e.g., for uses by firefighters) plus water taken illegally unbilled. A measure of how much of the water taken was unbilled.	Ml/d

End of Table

Appendix B R Code

B.1 Data Ingestion

B.1.1 Main

```r
#'
#' Download and load water data
#'

#'
#' Setup
#'

# imports
require("readxl") # for reading Excel file
require("dplyr")
require("DBI") # database communication
require("devtools") # for downloading external packages (e.g., from
↪  GitHub)

# download external ClickHouse package by patzaw
devtools::install_github("patzaw/ClickHouseHTTP")

#'
#' Data Ingestion
#'

load_data <- function(sheets_to_read) {
  # named list to act as dictionary
  water_df_lst <- list()

  for(sheet in seq_along(sheets_to_read)) {
    print(sheets_to_read[sheet])
    water_df <- read_excel("./water-data.xlsx",
    ↪  sheet=sheets_to_read[sheet])

    # remove the YearRef column as this is not needed
    water_df$YearRef <- NULL

    #' transpose the dataframe so attributes such as 'Potable water'
    ↪  are as columns rather
    #' than as rows
    water_df <- t(water_df)
    # set the column names - transpose function does not do this by
    ↪  default
    colnames(water_df) <- water_df[1,]

    # remove column names from rows
    water_df <- water_df[-1,]

    # convert back to dataframe rather than matrix - caused by
    ↪  transpose
    water_df <- as.data.frame(water_df)

    # help standardise column names by converting them all to lower
    ↪  case
    colnames(water_df) <- tolower(colnames(water_df))

    water_df_lst[[sheets_to_read[sheet]]] <- water_df
  }

  water_df_lst
}

convert_cols_to_numeric <- function(df) {
  data.frame(sapply(df, function(col) as.numeric(col)), row.names =
  ↪  row.names(df), stringsAsFactors = FALSE)
}

transfer_to_clickhouse <- function(water_df_lst, conn) {
  for(df_year in names(water_df_lst)) {
    data <- as_tibble(water_df_lst[[df_year]], rownames = "Company")
    dbWriteTable(conn, df_year, data, overwrite = TRUE)
  }
  invisible() # return nothing once function has completed
}

# try-catch exception handling to handle errors in case the desired
↪  data is not available or found
tryCatch(
  expr = {
    #'Downloads the relevant data then saves it to the current R
    ↪  working directory in an Excel file
    #'called 'water-data'
    download.file("https://environment.data.gov.uk/api/
    file/download?fileDataSetId=7b48642b-7bd6-4d8b-bc7a-
    16221a0d2948&fileName=
    Annual%20Review%20Historic%20Data%20-%20For%20publication
  %20031024.xlsx",
         mode = 'w',
         destfile = './water-data.xlsx')
  print("Download successful!")
 },
 error = function(e) {
  print("Download failed!")
  print(e)
 },
 finally = {
  water_df_lst <- load_data(c("05-06", "06-07", "07-08", "08-09",
  ↪  "09-10",
                "10-11", "11-12", "12-13", "13-14",
                ↪  "14-15",
                "15-16", "16-17", "17-18", "18-19",
                ↪  "19-20",
                "20-21", "21-22", "22-23", "23-24"))

  # convert all columns into numeric types instead of chr (character)
  ↪  types
  water_df_lst <- lapply(water_df_lst, function(df)
  ↪  convert_cols_to_numeric(df))

  # add NA for Bournemouth Water - make dataframe dimensions
  ↪  consistent
  sheets_to_update <- c("20-21", "21-22", "22-23", "23-24")
  for (sheet in sheets_to_update) {
   current_sheet <- water_df_lst[[sheet]]
   index_of_row_before <- which(row.names(current_sheet) ==
   ↪  "Anglian Water Services")
   water_df_lst[[sheet]] <- current_sheet %>%
   ↪  add_row(!!!setNames(rep(NA,

                            ↪  ncol(current_sheet)),

                            ↪  names(current_sheet)),
                    .after=index_of_row_before)
   # update new row's name
   rownames(water_df_lst[[sheet]])[index_of_row_before+1] <-
   ↪  "Bournemouth Water"
  }

  # connect to the ClickHouse server - assumes the server is active
  ↪  and healthy
  clickhouse_conn <- dbConnect(ClickHouseHTTP::ClickHouseHTTP(),
                                 host = "localhost",
                                 port = 8123,
                                 user = "alroberts",
                                 db = "uk_water")

    transfer_to_clickhouse(water_df_lst, clickhouse_conn)

    # for testing purposes - read from the database
    # test_df <- dbReadTable(clickhouse_conn, "05-06")
    # View(test_df)

    dbDisconnect(clickhouse_conn)
  }
)
```

B.1.2 Investigating semantic schema differences

```
#'
#' Supplimentary code used to confirm data integrity and ingestion
#'

# check for semantic schema differences
sheets_with_changes = list()
# iterate through each dataframe (sheet)
for(index in seq_along(water_df_lst)[-length(water_df_lst)]) {
  current_sheet <- water_df_lst[[index]]
  next_sheet <- water_df_lst[[index + 1]]

  # check for column name and row name differences using Sets
  non_matching_cols <- setdiff(colnames(current_sheet),
   ↪  colnames(next_sheet))
  non_matching_rows <- setdiff(rownames(current_sheet),
   ↪  rownames(next_sheet))
  if((length(non_matching_cols) != 0) || (length(non_matching_rows !=
   ↪  0))){
    # log the current sheet
    sheets_with_changes[[names(water_df_lst)[index+1]]] <- list(
      non_matching_cols = non_matching_cols,
      non_matching_rows = non_matching_rows
    )
  }
}
View(sheets_with_changes)
```

```
# confirm all columns are numeric
sheets_not_numeric = c()
for(sheet in names(water_df_lst)) {
  if (any(sapply(water_df_lst[[sheet]], function(df) !is.numeric(df))))
   ↪  {
    sheets_not_numeric <- c(sheets_not_numeric, sheet)
  }
}
```

B.2 Data Analysis

```
#'
#' Investigating statistical questions
#'

#'
#' Setup
#'

# imports
require("dplyr")
require("ggplot2")
require("DBI") # database communication
require("BSDA") # for z-test
require("zoo") # contains manipulation tools for time-series and NA
require("reshape2") # for dataframe manipulation
require("forecast") # ARIMA
# require("devtools") # for downloading external packages (e.g., from
 ↪  GitHub)

# download external ClickHouse package by patzaw - if not done so
 ↪  already
# devtools::install_github("patzaw/ClickHouseHTTP")

# data download from ClickHouse
clickhouse_conn <- dbConnect(ClickHouseHTTP::ClickHouseHTTP(),
                             host = "localhost",
                             port = 8123,
                             user = "alroberts",
                             db = "uk_water")

tables_to_download <- dbListTables(clickhouse_conn)
water_df_lst <- list()
for (table in tables_to_download) {
  table_to_add <- dbReadTable(clickhouse_conn, table)
  rownames(table_to_add) <- table_to_add$Company
  table_to_add$Company <- NULL
  water_df_lst[[table]] <- table_to_add
}

# close the connection
dbDisconnect(clickhouse_conn)

# confirm all the row names are consistent
check_row_names_consistency <- function(df_list) {
  row_names_list <- lapply(df_list, row.names)
  first_row_names <- row_names_list[[1]]
  all_identical <- all(sapply(row_names_list, function(x) identical(x,
   ↪  first_row_names)))
}

if (check_row_names_consistency(water_df_lst)) {
  "Row names are consistent."
} else {
  "Row names inconsistent."
}

#'
#' QID 1: Water Demand
#'

# get and add water demand data into one dataframe
water_consumption_df <- list()
for (sheet in names(water_df_lst)) {
  data <- water_df_lst[[sheet]]
  column_to_add <- data %>% mutate(Consumption =
   ↪  measured.household...consumption +
        measured.non.household...consumption +
        unmeasured.household...consumption +
        unmeasured.non.household...consumption) %>% select(
        Consumption
        )
  colnames(column_to_add) <- sheet
  if(length(water_consumption_df) == 0) {
    water_consumption_df <- column_to_add
  } else {
    water_consumption_df <- bind_cols(water_consumption_df,
     ↪  column_to_add)
  }
```

```
}

consumption_means_df <- data.frame(Column =
 ↪  names(water_consumption_df),
                                   Mean=colMeans(water_consumption_df,
                                    ↪  na.rm = TRUE))
ggplot(consumption_means_df, aes(x=Column, y=Mean, group=1)) +
  geom_line() +
  geom_point() +
  labs(title =
   ↪  "Estimated, averaged daily water demand for England and Wales",
       x = "Time Period",
       y = "Mean Consumption (Ml/d)") +
  theme_minimal()

# perform K-S Test on '19-20' and '20-21' to confirm both are normally
 ↪  distributed
ks_test_19_20 <- ks.test(water_consumption_df$`19-20`, "pnorm",
                          mean(water_consumption_df$`19-20`),
                          sd(water_consumption_df$`19-20`))

ks_test_20_21 <- ks.test(water_consumption_df$`20-21`, "pnorm",
                        mean(water_consumption_df$`20-21`, na.rm =
                         ↪  TRUE),
                        sd(water_consumption_df$`20-21`, na.rm = TRUE))
print(ks_test_19_20)
print(ks_test_20_21)

# performing Z-test for hypothesis testing; to see if mean difference
 ↪  is significant
# cannot do directly do paired z-test, have to do one sample z-test of
 ↪  differences instead
mean_water_demand_diffs <- na.omit(water_consumption_df$`20-21`)
 ↪  -water_consumption_df[rownames(water_consumption_df) !=
 ↪  "Bournemouth Water",]$`19-20`
mean_diff <- mean(mean_water_demand_diffs)
sd_diff <- sd(mean_water_demand_diffs)
n <- length(mean_water_demand_diffs)

z_test_19_20_20_21 <- z.test(x=mean_water_demand_diffs,
                             alternative = "two.sided", mu = 0,
                             sigma.x = sd_diff/sqrt(n),
                             conf.level = 0.95)
print(z_test_19_20_20_21)
```

```
#'
#' QID 2: Leakages Over Time
#'

# get leakage data over time
leakage_df <- list()
population_df <- list()
for (sheet in names(water_df_lst)) {
  data <- water_df_lst[[sheet]]
  column_to_add_leakage <- data.frame(sheet =
  ↪   data['total.leakage..ml.d.'])
  column_to_add_population <- data.frame(sheet =
  ↪   data['total.population'])
  colnames(column_to_add_leakage) <- sheet
  colnames(column_to_add_population) <- sheet
  if(length(leakage_df) == 0) {
    leakage_df <- column_to_add_leakage
    population_df <- column_to_add_population
  } else {
    leakage_df <- bind_cols(leakage_df, column_to_add_leakage)
    population_df <- bind_cols(population_df, column_to_add_population)
  }
}

leakage_means_df <- data.frame(Column = names(leakage_df),
                                   Mean=colMeans(leakage_df, na.rm =
                                   ↪   TRUE))

leakage_plot <- ggplot(leakage_means_df, aes(x=Column, y=Mean,
↪   group=1)) +
  geom_line() +
  geom_smooth(method="lm", se=FALSE, color="blue", linetype = 'dashed')
  ↪   +
  geom_point() +
  labs(title =
  ↪   "Annual averaged leakage of water across water companies for England and Wales",
       x = "Time Period",
       y = "Mean Leakage (Ml/d)") +
  theme_minimal()

leakage_subset <- leakage_means_df %>% filter(Column %in% c('11-12',
↪   '12-13', '13-14', '14-15', '15-16',
                                                           '16-17',
                                                           ↪   '17-18',
                                                           ↪   '18-19',
                                                           ↪   '19-20',
                                                           ↪   '20-21',
                                                           '21-22',
                                                           ↪   '22-23',
                                                           ↪   '23-24'))

leakage_plot + geom_smooth(data = leakage_subset, aes(Column, Mean,
↪   group=1), method="lm",
                           se=FALSE, color="red", linetype="dashed")

leakage_percentages_df <- leakage_df[row.names(leakage_df) !=
↪   "Bournemouth Water",] %>%
  mutate(Percentage = `23-24` / sum(`23-24`) * 100)

combined_df <- na.omit(data.frame(Leakage=leakage_df$`23-24`,
                          Population=population_df$`23-24`,
                          Company=row.names(leakage_df)))

ggplot(combined_df, aes(x=Leakage, y=Population, color=Company)) +
  geom_point(size=3) +
  geom_smooth(method="lm", se=FALSE, color="black", linetype="dashed")
  ↪   +
  labs(title =
  ↪   "The relationship between 'leakage' and the 'population' for water companies, '23-
       x = "Leakage (Ml/d)",
       y = "Population (000s)")
  theme_minimal()

summary(leakage_df$`23-24`)

#'
#' QID 3: Water Availability by Region
#'

water_availability <- list()
population_df <- list()
for (sheet in names(water_df_lst)) {
  data <- water_df_lst[[sheet]]
  column_to_add_water_availability <- data.frame(sheet =
  ↪   data[colnames(data) %in% c('total.water.available.for.use',
                                                                                  ↪
  column_to_add_population <- data.frame(sheet =
    ↪  data['total.population'])
  colnames(column_to_add_water_availability) <- sheet
  colnames(column_to_add_population) <- sheet
  if(length(water_availability) == 0) {
    water_availability <- column_to_add_water_availability
    population_df <- column_to_add_population
  } else {
    water_availability <- bind_cols(water_availability,
      ↪  column_to_add_water_availability)
    population_df <- bind_cols(population_df, column_to_add_population)
  }
}

combined_df <-
 ↪  na.omit(data.frame(Availability=water_availability$`23-24`,
                                  Population=population_df$`23-24`,

                                   ↪  Company=row.names(water_availability)))

ggplot(combined_df, aes(x=Availability, y=Population, color=Company)) +
  geom_point(size=3) +
  geom_smooth(method="lm", se=FALSE, color="black", linetype="dashed")
    ↪  +
  labs(title =
    ↪  "The relationship between 'water availability' and the 'population' for water comp
       x = "Total Water Availability (Ml/d)",
       y = "Population (000s)")
theme_minimal()

linear_availability_population <- lm(Availability ~ Population,
 ↪  data=combined_df)
print(summary(linear_availability_population))

#'
#' QID 4: Individual Water Consumption
#'

individual_consumption_df <- list()
thames_water_df <- list()
united_utilities_df <- list()
for (sheet in names(water_df_lst)) {
  data <- water_df_lst[[sheet]]
  column_to_add <- data %>% mutate(Individual =
    ↪  average.household...pcc) %>% select(
                                       Individual
                                     )
  column_to_add_thames <- data[row.names(data) == "Thames Water",] %>%
  ↪  mutate(Individual = average.household...pcc) %>% select(
    Individual
  )
  column_to_add_utilities <- data[row.names(data) ==
  ↪  "United Utilities",] %>% mutate(Individual =
  ↪  average.household...pcc) %>% select(
    Individual
  )
  colnames(column_to_add) <- sheet
  colnames(column_to_add_thames) <- sheet
  colnames(column_to_add_utilities) <- sheet
  if(length(individual_consumption_df) == 0) {
    individual_consumption_df <- column_to_add
    thames_water_df <- column_to_add_thames
    united_utilities_df <- column_to_add_utilities
  } else {
    individual_consumption_df <- bind_cols(individual_consumption_df,
    ↪  column_to_add)
    thames_water_df <- bind_cols(thames_water_df, column_to_add_thames)
    united_utilities_df <- bind_cols(united_utilities_df,
    ↪  column_to_add_utilities)
  }
}

individual_consumption_means_df <- data.frame(Time =
↪  names(water_outage),
                              Mean=colMeans(individual_consumption_df,
                               ↪  na.rm = TRUE))
thames_consumption = data.frame(Time = names(thames_water_df),
                                Individual=colMeans(thames_water_df))
united_util_consumption = data.frame(Time = names(united_utilities_df),

                                 ↪  Individual=colMeans(united_utilities_df))

ggplot(individual_consumption_means_df, aes(x=Time, y=Mean, group=1)) +
  geom_line(aes(color="England&Wales")) +
  geom_line(data = thames_consumption, aes(x=Time, y=Individual,
  ↪  color="Thames Water")) +
  geom_line(data = united_util_consumption, aes(x=Time, y=Individual,
  ↪  color="United Utilities")) +
  labs(title =
  ↪  "Averaged daily water consumption PCC recorded across England and Wales",
       x = "Time Period",
       y = "Water Consumed (l/h/d)") +
  theme_minimal()

j#'
#' QID 5: Raw Water Imports
#'

# get all raw water import data
raw_water_df <- list()
for (sheet in names(water_df_lst)) {
  data <- water_df_lst[[sheet]]
  column_to_add <- data.frame(sheet = data['raw.water.imported'])
  colnames(column_to_add) <- sheet
  if(length(raw_water_df) == 0) {
    raw_water_df <- column_to_add
  } else {
    raw_water_df <- bind_cols(raw_water_df, column_to_add)
  }
}

# for getting only x-axis labels for companies who have imported raw
↪   water within the time period
raw_water_13_14_no_zeros <- raw_water_df %>% filter(`13-14` != 0)
ggplot(raw_water_13_14_no_zeros,
↪   aes(x=row.names(raw_water_13_14_no_zeros), y=`13-14`)) +
  geom_bar(stat="identity") +
  labs(title =
  ↪   "Bar chart of daily imported raw water by each company for 2013-2014",
       x="Company",
       y="Raw Water Imported (Ml/d)") +
  theme_minimal()

raw_water_23_24_no_zeros <- raw_water_df %>% filter(`23-24` != 0)
ggplot(raw_water_23_24_no_zeros,
↪   aes(x=row.names(raw_water_23_24_no_zeros), y=`23-24`)) +
  geom_bar(stat="identity") +
  labs(title =
  ↪   "Bar chart of daily imported raw water by each company for 2023-2024",
       x="Company",
       y="Raw Water Imported (Ml/d)") +
  theme_minimal()

# hypothesis testing using Wilcoxon Signed-Rank Test
```

```
wilcox_results <- wilcox.test(raw_water_df$`13-14`,
↪  raw_water_df$`23-24`, paired=TRUE, exact=FALSE)
print(wilcox_results)

#'
#' QID 6: Water Treatment Efficiency
#'
summary(water_df_lst[['13-14']]$treatment.works.losses.and.operational.use)
summary(water_df_lst[['23-24']]$treatment.works.losses.and.operational.use)

water_treatment <- list()
for (sheet in names(water_df_lst)) {
  data <- water_df_lst[[sheet]]
  column_to_add <- data %>% mutate(Treatment =
  ↪  data$treatment.works.losses.and.operational.use) %>% select(
                                       Treatment
                                     )
  colnames(column_to_add) <- sheet
  if(length(water_treatment) == 0) {
    water_treatment <- column_to_add
  } else {
    water_treatment <- bind_cols(water_treatment, column_to_add)
  }
}

treatment_means_df <- data.frame(Column = names(water_treatment),
                                 Mean=colMeans(water_treatment, na.rm
                                 ↪  = TRUE))
ggplot(treatment_means_df, aes(x=Column, y=Mean, group=1)) +
  geom_line() +
  geom_point() +
  labs(title =
  ↪  "Annual averaged daily treatment works losses and operational use for England and
       x = "Time Period",
       y = "Water loss (Ml/d)") +
  theme_minimal()

ks_test_treatment_10_11 <- ks.test(water_treatment$`10-11`, "pnorm",
                         mean(water_treatment$`10-11`),
                         sd(water_treatment$`10-11`))

ks_test_treatment_11_12 <- ks.test(water_treatment$`11-12`, "pnorm",
                         mean(water_treatment$`11-12`),
                         sd(water_treatment$`11-12`))
print(ks_test_treatment_10_11)
print(ks_test_treatment_11_12)

#'
#' QID 7: Water Outages
#'

water_outage <- list()
for (sheet in names(water_df_lst)) {
  data <- water_df_lst[[sheet]]
  column_to_add <- data %>% mutate(Outage =  data$outage.experienced)
  ↪  %>% select(
    Outage
  )
  colnames(column_to_add) <- sheet
  if(length(water_outage) == 0) {
    water_outage <- column_to_add
  } else {
    water_outage <- bind_cols(water_outage, column_to_add)
  }
}

outage_means_df <- data.frame(Time = names(water_outage),
                                   Mean=colMeans(water_outage, na.rm =
                                    ↪  TRUE))
ggplot(outage_means_df, aes(x=Time, y=Mean, group=1)) +
  geom_line() +
  geom_point() +
  labs(title =
  ↪  "Averaged daily water outage recorded across England and Wales",
       x = "Time Period",
       y = "Outage Experienced (Ml/d)") +
  theme_minimal()

summary(water_df_lst$`18-19`['outage.experienced'])
summary(water_df_lst$`22-23`['outage.experienced'])

fit <- auto.arima(outage_means_df$Mean)
forecastedValues <- forecast(fit, 1)
forecastedValues <- data.frame(forecastedValues)

forecast_df <- data.frame(
  Time = '24-25',
  Mean = forecastedValues$Point.Forecast,
  Lo.95 = forecastedValues$Lo.95,
  Hi.95 = forecastedValues$Hi.95
)
previous_outage_row <- data.frame(
  Time = '23-24',
  Mean = outage_means_df['23-24',]$Mean,
  Lo.95 = outage_means_df['23-24',]$Mean,
  Hi.95 = outage_means_df['23-24',]$Mean
)

forecast_df <- rbind(previous_outage_row, forecast_df)

ggplot(outage_means_df, aes(x=Time, y=Mean, group=1)) +
  geom_line() +
  geom_line(data=forecast_df, aes(x=Time, y=Mean, color="Forecast")) +
  geom_line(data=forecast_df, aes(x=Time, y=Lo.95, color="Forecast"),
   ↪  linetype="dashed") +
  geom_line(data=forecast_df, aes(x=Time, y=Hi.95, color="Forecast"),
   ↪  linetype = "dashed") +
  labs(title =
   ↪  "Averaged daily water outage recorded across England and Wales",
       x = "Time Period",
       y = "Outage Experienced (Ml/d)") +
  theme_minimal()
```

Appendix C Docker Compose

```
name: clickhouse-water-data
services:
  clickhouse-server:
    image: clickhouse/clickhouse-server:latest
    container_name: clickhouse-server
    ulimits:
      nofile:
        soft: 242144
        hard: 242144
    ports:
      - 9000:9000
      - 8123:8123
    environment:
      CLICKHOUSE_USER: "alroberts"
      CLICKHOUSE_PASS: "WaterInvestigation"
      CLICKHOUSE_DB: "uk_water"
      CLICKHOUSE_DEFAULT_ACCESS_MANAGEMENT: 1
    healthcheck:
      test: ["CMD", "wget", "--no-verbose", "--tries=1", "--spider",
       ↪  "http://localhost:8123/ping"]
      interval: 10s
      retries: 2
      start_period: 20s

  ch-ui:
    image: ghcr.io/caioricciuti/ch-ui:latest
    container_name: clickhouse-gui-manager
    depends_on:
      clickhouse-server:
        condition: service_healthy
    restart: always
    ports:
      - 5521:5521
    environment:
      VITE_CLICKHOUSE_URL: "http://localhost:8123"
      VITE_CLICKHOUSE_USER: "alroberts"
```