Lake Bilancino Acea Water Group

1. Introduction

1.1 Motivation

Most people tend to take the availability of drinking water for granted, and don't really spend much time thinking about the process behind its collection and refinery. Natural freshwater water bodies such as rivers, lakes, water springs, and aquifers are often impacted by not only the impact that humans have made on the world, but also various natural factors which can vary dramatically from season to season. We are observing Lake Bilancino in Tuscany, Italy in order to find out more about its hydrometry.

1.2 Importance

As the earth becomes more densely populated and resources become more scarce, the integrity and reliance on water supply has become increasingly important. To reflect this, workers are using modern tools to observe, measure, and record different factors that may affect the quality of different water bodies. With multiple factors affecting the availability of water for daily consumption by the people of Tuscany, we aim to create predictions for water level and flow rate at different times of the year. We hope to deliver an outcome that will aid the Acea Group who are in charge of the management of water networks in many places across Italy, to have a better understanding of the availability and hydrometry of the water in Lake Bilancino.

1.3 Initial and Investigation Questions

Our central focus is to find out whether there is any significant relation between weather conditions in the different parts of Lake Bilancino and its water quality between different seasons. We will be exploring the relationships between different variables within our Lake Bilancino dataset through various methods of statistical analysis. We will expand and build our results based on the following investigation questions:

1) What areas and weather conditions are most affecting water quality in lakes?
2) Interval all the flow rates and categorize 0-25%,25%-50%,50%-75%,75%-100%, and group all the data annually to see if there is shortage or surplus in certain years.
3) Time series model of each year, to see the general behavior of the lake flow rate, to if if there is any unusual behavior (ex: drought)
4) Which parameter has the most significant effect in relation to groundwater levels in each of the lakers?
5) Analyze flow rate and water level between seasons
6) Number of rainfall affects the flow rate or water level of the water body

We hope to use this information to potentially predict the water level and flow rate of Lake Bilancino at different times of the year.

2. Data

The Acea Group deals with four different types of waterbodies: water spring, lake, river and aquifers.This project initially uses nine different datasets, completely independent and not linked to each other. Each dataset represents a different body of water. Although we understood the significance of each of these bodies of water, we wanted to focus specifically on lakes. We realized that lakes data could become the most insightful because it contains multiple potentially influential parameters such as rainfall and temperature that could have a significant impact on water level and flow rate.

The data is specifically from Lake Bilancino, an artificial lake located in the municipality of Barberino di Mugello (about 50 km from Florence). This lake is important because it is used to refill the Arno river during the summer months. During the winter months, the lake is filled up and then, during the summer months, the water of the lake is poured into the Arno river. The dataset starts on January 01, 2004 and ends on June 30, 2020. It tracks the rainfall of 5 of the areas adjacent to the lake. These include: San Piero a Sieve (10km east of the lake), Mangona (12km northwest of the lake), Sant'Agata sui Due Golfi (7km northeast of the lake), Cavallina (2km west of the lake), and Le Croci (4km north of the lake). We are able to see how the rainfall from these different cities could possibly affect the water levels and flow rate. The dataset also tracks the daily temperature in Le Croci.

2.1 Data Variables

Date (Numerical, Discrete) : The date of the recorded data points
Rainfall_S_Piero (Numerical, Continuous): The amount of rainfall in S. Piero (millimeters)
Rainfall_Mangona (Numerical, Continuous): The amount of rainfall in Mangona (millimeters)
Rainfall_S_Agata (Numerical, Continuous): The amount of rainfall in S.Agata (millimeters)
Rainfall_Cavallina (Numerical, Continuous): The amount of rainfall in Cavallina (millimeters)
Rainfall_Le_Croci (Numerical, Continuous): The amount of rainfall in Le Croci (millimeters)
Temperature_Le_Croci (Numerical, Discrete): The temperature in Le Croci (celsius)
Lake_Level (Numerical, Continuous): The level of the lake (meters)
Flow_Rate (Numerical, Continuous): The flow of the lake (m^3/sec)

3. Background

The Acea Group is an Italian multiutility group. It provides water, electricity and environmental services to millions of people spanning across areas such as Lazio, Tuscany, Umbria, Molise, Campania. In our report, we will be focusing solely on the Acea Group's water sector. With regards to their water services, the Acea Group has served over 9 million customers in four regions, managed 58,000 km of drinking water networks and has made over 460,000 analyses of drinking water [2]. Prior to entering homes and towns, the water has been inspected and deemed to be drinkable. The company invests in the future using technological innovation and raising awareness to responsible consumption. In other words, the Acea Group is constantly finding ways to encourage their customers to limit water waste and reduce environmental pollution but stay healthy at the same time using the data that dates back more than a decade ago.

Originally, we were given 9 different datasets, each pertaining to a different type of waterbody/ waterbody characteristics. We will be focusing on one type: lakes. Like the other bodies of water, have their own hydrologic mannerisms. These include but are not limited to water utilization, dike (barriers used to hold back water) construction, and water restriction. While it is important to take notice of emergency factors of lakes such as drought control, flooding evacuation procedures and diversions in case of a flood, we will be focusing on the water quality of the lake itself [1]. The water quality of lakes encompasses the water temperature, water levels and flowrate of the water.

Unlike rivers and streams, the end location of these types of water bodies are not addressed in the water analyses by the Acea Group. However, it is important to note that natural lakes are the direct result of a downward river. If the river has a high flow rate, it will result in a high volume of water in the downstream lake [2]. Since the lake we are observing is an artificial lake, we would have to take into concern the water infiltration. Water can seep in through the floor creating minor fluctuations in the water levels, similar to the artificial Lake Nasser, Egypt which also happens to be the largest artificial lake in the world [3]. Dry weather can also take a part in water levels, as during the summertime, the extreme heat can evaporate more waters than that of the non-summer seasons. Examining the water levels is quite important since it can also affect the environment. For example, low water levels can create clusters of *Dreissena Polymorpha,* a freshwater mussel species. Because of their high-body fat content and ability to filter water, these mussels create a nuisance to water basins since they end up clogging water intakes [4]. But the main reason the water level is so important is quite anti-climactic. The less water a lake intakes for the year, the less water consumers will be able to drink. Droughts and severe weather have hurt and helped the water quality and level Lake Bilancino, respectively. We hope to find correlation and see how different times of the year and their corresponding weathers affect these features.

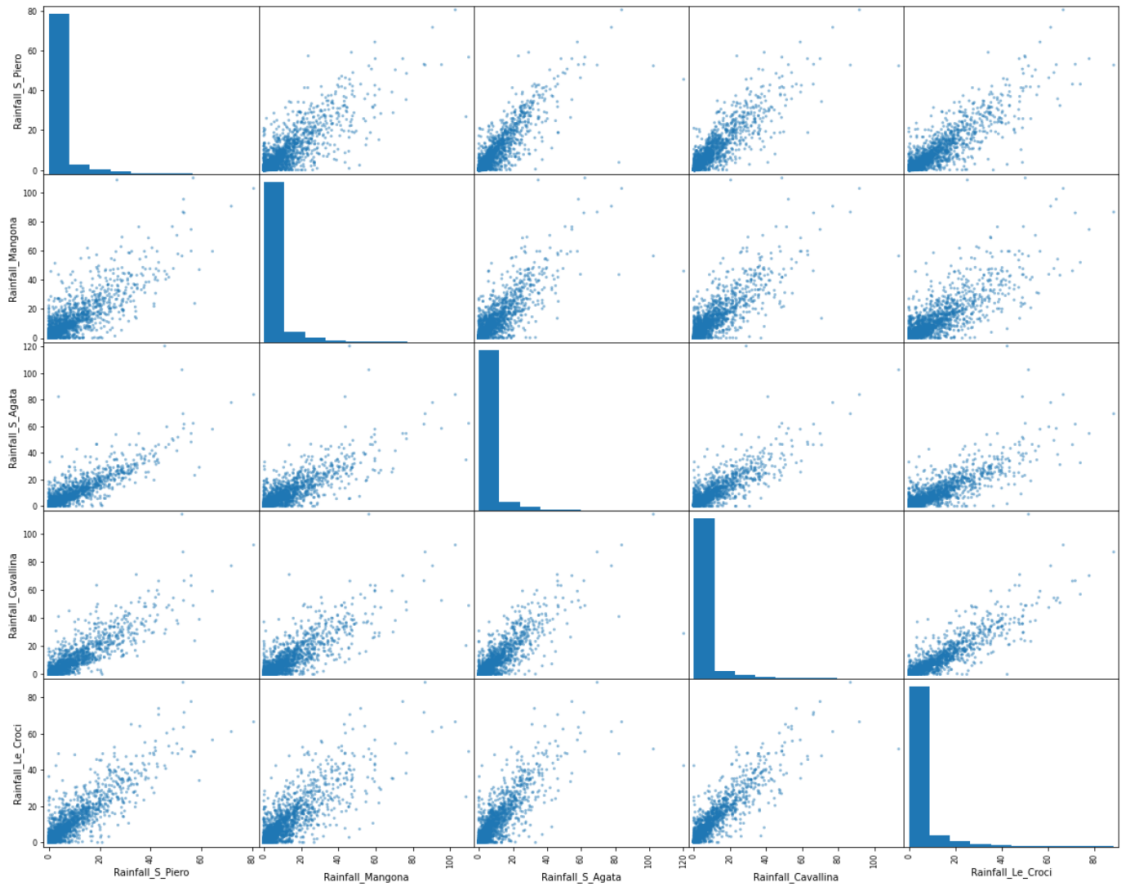
Due to global warming, a great number of lakes are having declining water levels. Normally, lakes go through cycle over the years, where the decline and incline balance out during the cycle(1). In addition, unusual or dramatic decline in lake level occurs during warm and dry years, moderate decline in the cold and dry years, slight decline in the warm and wet years(2).

4. Investigation/Analysis

4.1 What areas and weather conditions are most affecting water quality in lakes?

To answer this question, we want to analyze how the amount of rainfall in each of the five regions of the lake (S Piero, Mangona, Agata, Cavallina, and Le Croci) relate to both lake level and flow rate. Due to the fact that the surface area of Lake Bilancino is only about 1.9 square miles, we can infer that when it rains in one area, it is very likely that it will rain in the other areas as well. By creating a correlation matrix between the different rainfall areas, we can see this relationship in more detail.

Rainfall Correlation Matrix



Rainfall Correlation Coefficients

	Rainfall_S_Piero	Rainfall_Mangona	Rainfall_S_Agata	Rainfall_Cavallina	Rainfall_Le_Croci
Rainfall_S_Piero	1.000000	0.867117	0.897863	0.908543	0.910068
Rainfall_Mangona	0.867117	1.000000	0.865833	0.873150	0.863780
Rainfall_S_Agata	0.897863	0.865833	1.000000	0.887618	0.880652
Rainfall_Cavallina	0.908543	0.873150	0.887618	1.000000	0.922999
Rainfall_Le_Croci	0.910068	0.863780	0.880652	0.922999	1.000000

As we can see the rainfall areas are closely correlated with correlation coefficients ranging between ~0.86 - ~0.92. With this in mind, it is important to recognize that any difference in the way each area's rainfall may affect water level and flow rate may not be massive, but even the slightest relations found can be beneficial in aiding the management of water for Lake Bilancino. To proceed, we will examine the relationships between rainfall amounts and flow rate/lake level for each of the different locations. By running linear mixed models regression separated by each month in the dataset we can study the strength of the relationships across the different months that rainfall occurs. This is advantageous because of the dramatic changes in rainfall amounts between months/seasons. If we do this for both flow rate and lake level we get the following correlation coefficients:

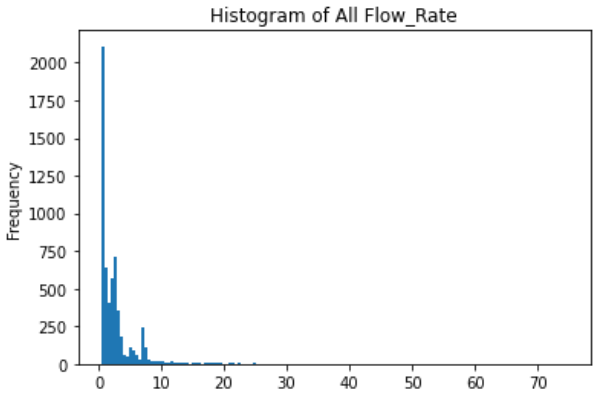
	Area	Flow Rate Corr	Lake Level Corr
0	Piero	0.079	0.010
1	Mangona	0.070	0.008
2	S Agata	0.085	0.010
3	Cavallina	0.084	0.009
4	Le Croci	0.087	0.009

As expected the values are very near each other for the same reasons as discussed earlier, but of these rainfall areas we can see that Le Croci held the most prominent relationship with the flow rate of the lake. This makes sense because rain can have a heavy impact on flow rate in a water body depending on amount and density of rainfall. For lake level, the regression does not hold much significance as for the flow rate because regardless of where the rain falls, it will spread across the lake and increase its level.

4.2 Interval all the flow rates and categorize 0-25%,25%-50%,50%-75%,75%-100%, and group all the data annually to see if there is shortage or surplus in certain years.

The interpretation of Flow rates will be explained by the data between 2004 to 2020, since prior year data does not consist of rainfalls measurement.

First, we construct a Histogram of Flow Rate of everyday



We see that the histogram is heavily right skewed with a tremendous amount of data stack on the left side. Following is the data statistics with 25,50,75,100 quantiles, median and etc.

```
descriptive_stats(LakeData,'Flow_Rate')
```

```
Mean: 2.7782041493776783
Median: 1.5
Q1: 0.6
Q2: 1.5
Q3: 3.0
Q4: 74.65
Inter-Quartile Range: 2.4
Standard Deviation: 4.130833602567077
Variance: 17.063786252097298
Range: 74.2
----------------------------------------
```

However, we still perform the Kurtosis test, we get the values of 60.3859, such big values tells us that the data are heavy-tail relative to a normal distribution. Therefore, we see that there are some significant variables that affect the flow rate of the lake.

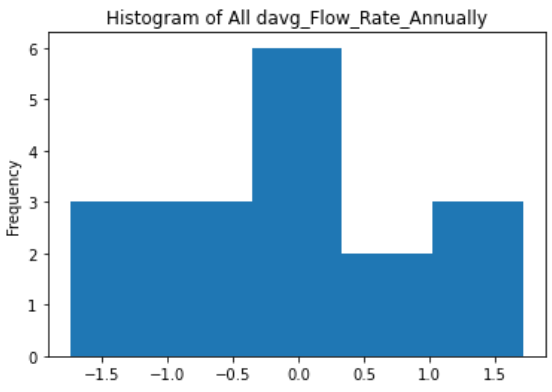
Given data lastly contains a variable, which is the rainfall amount with area that attribute to the change in lake flow rate.

In order to understand if the current water volume is at a shortage or a surplus, we calculate the average water flow of each year, and determine the differences in water flow rate with the overall average flow rate of the lake.

The Average Flow Rate of Bilancino Lake=2.778204

	avg_Flow_Rate_Annually	davg_Flow_Rate_Annually	sum_RF_Anually	davg_Lake_Level	time	avg_LakeL
0	3.764932	0.986727	5108.0	-0.277919	2004.0	250.292630
1	2.833151	0.054947	5243.8	0.162081	2005.0	250.732630
2	3.693973	0.915768	4222.0	-0.440549	2006.0	250.130000
3	1.489205	-1.288999	3939.6	-2.095618	2007.0	248.474932
4	1.536120	-1.242084	5074.8	-1.578637	2008.0	248.991913
5	3.073726	0.295522	5119.2	-1.019837	2009.0	249.550712
6	4.492877	1.714673	7134.0	0.396108	2010.0	250.966658
7	2.668219	-0.109985	3724.8	-1.418330	2011.0	249.152219
8	1.036339	-1.741865	4477.8	-4.641588	2012.0	245.928962
9	4.222658	1.444453	6228.6	-0.329152	2013.0	250.241397
10	4.096932	1.318727	7237.8	-0.592412	2014.0	249.978137
11	2.073288	-0.704916	4444.6	-0.963700	2015.0	249.606849
12	2.485710	-0.292494	5671.8	-0.994183	2016.0	249.576366
13	1.821068	-0.957136	4968.2	-1.449234	2017.0	249.121315
14	2.885014	0.106810	5041.0	-0.755700	2018.0	249.814849
15	2.548740	-0.229464	6366.8	-0.770303	2019.0	249.800247
16	2.253352	-0.524853	2087.6	0.000000	2020.0	250.570549

Following the histogram of the variable davg_Flow_Rate_Annually



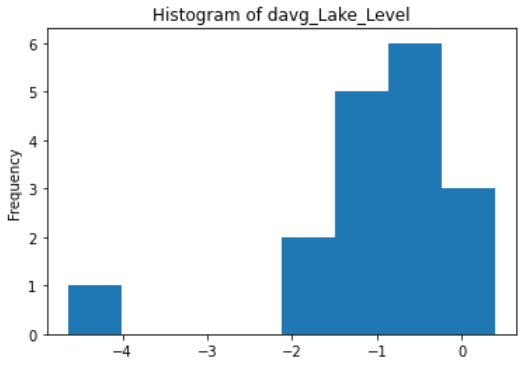
We see that the graph has a relative normal graph. However, to ensure the normality of the chart, we resample the data 500 times, using kurtosis to test if the distribution of the differences annual average flow rate is normal.

Kurtosis=-0.982815, which is roughly less than -3, we conclude that the distribution is light tails and has features of a platykurtic distribution.

4.3 Time series model of each year, to see the general behavior of the lake levels, to see if if there is any unusual behavior (ex: possible drought),

We group each set of data by year, based on each year's average values to determine any unusual behavior of the lake.

Base on the histogram of the Average Lake Level of each year



The majority of the data is located close to the right side of the graph.

```
descriptive_stats(Q1data,'davg_Lake_Level')

Mean: -0.9864103193627363
Median: -0.7703028752069656
Q1: -1.4183302724670739
Q2: -0.7703028752069656
Q3: -0.3291521902755221
Q4: 0.39610808369704387
Inter-Quartile Range: 1.0891780821915518
Standard Deviation: 1.1123456891517194
Variance: 1.2373129321744134
Range: 5.037695785612755
----------------------------------------
```

In addition, even though the data are left skewed, the overall values are allocated below zero. The majority values for average change in lake level is negative. We see from 2004 to 2020 (table in 4.2) only 2005, 2010, and 2020 have positive numbers for change in average water level.

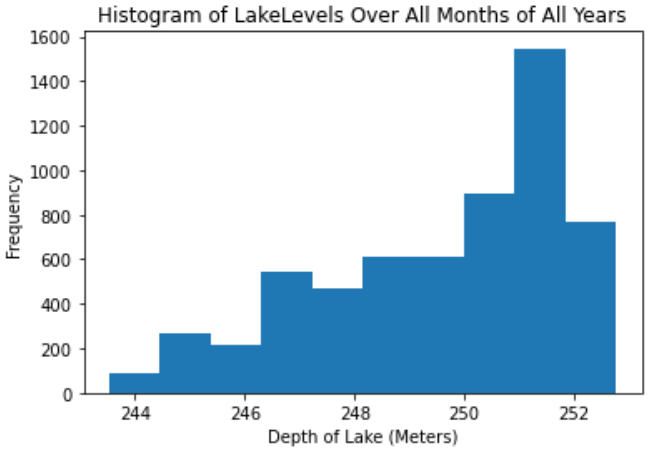
We can state such unusual behavior, which water level is consistently decreasing in water level, might be caused by drought or any other natural event.

With a deeper look at the change in Lake Level in time series

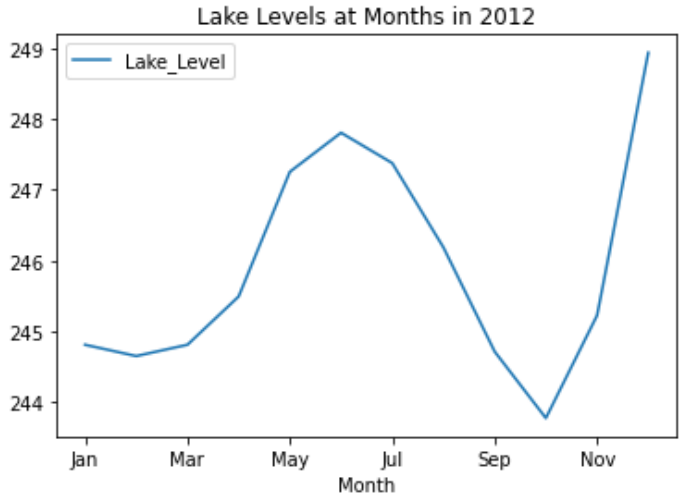
There is a clear indication where 2012 has a huge drop in water level, which can possibly represent there was a drought during 2012.

4.4 Investigation into the 2012 lake levels

While the rise and fall of the water levels may seem insignificant on paper, it has an actual impact on the amount of water lost or gained. One meter of water in the lake level lost is equivalent to around 5000L of water. To take this into perspective, we will take a look at an 'outlier' in the data. As mentioned earlier, 2012 was the year that the lake level took a noticeable dip. It decreased from the average of 249.58 meters (between all the years) to a 245.93 meters average (within 2012). That is a decrease in 3.66 meters of water depth, equal to around an 18276.57 L water decrease.



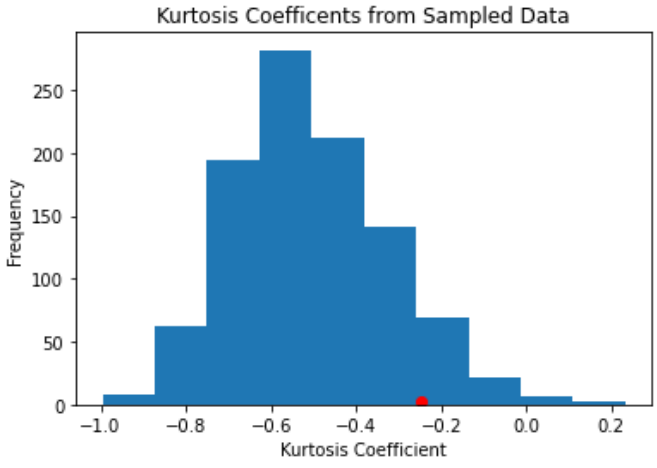
When plotting a preliminary histogram of the distribution of lake levels, we see a left skewed graph with a somewhat unimodal distribution shape. As stated earlier, we see a clear indication that one of the years had a drop in lake levels. We found out that it was during 2012 that this outlier occurred.



	Lake_Level
count	12.000000
mean	245.920833
std	1.577242
min	243.777742
25%	244.788844
50%	245.360667
75%	247.284274
max	248.934516

Here we can observe that the lowest water level at the lake during 2012 was in October. It had a level of 243.78 Meters, 2.143 Meters below the average for 2012. That is around 10715.46 liters of water below the 2012 average water level, which in our case, is already the lowest lake-level year of the dataset.

To see if this irregularity is normal or not, we will generate pseudo random lake levels from the given distribution and use that to decide whether a departure from normal distribution is big or small. We will also check the similarity of the simulated lake levels with the observed lake levels between the months of the different year. Then the kurtosis coefficients will be repeated on the same exact sample size. This will be able to give us a better picture of 2012's decline in lake levels. On a side note, in each iteration we collected 365 water levels since 2012 was a special case with 366 observations (it was a leap year).



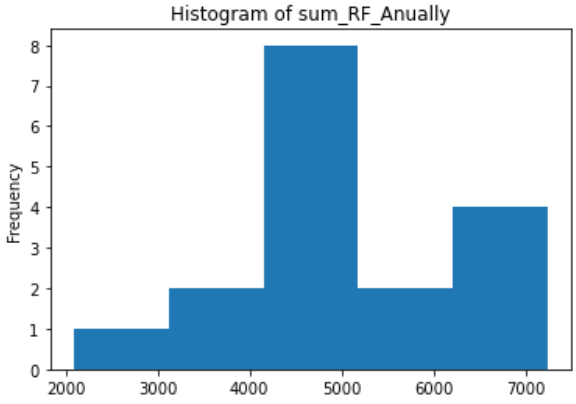
	Descriptive Stats
count	1000.000000
mean	-0.503742
std	0.183322
min	-0.959135
25%	-0.638738
50%	-0.516225
75%	-0.393962
max	0.110237

After a kurtosis test, we affirmed that the 2012 data rejects the null hypothesis and favors the alternative that it is not normally distributed. The Kurtosis Coefficient of the 2012 lake data is

-0.246, which falls under the skew of the graph, thus deeming the data is not normally distributed. This is presumably since the 2012 lake levels were left skewed to begin with. In the following sections, we will see if rainfall has an affect on this interesting year, as well as the other years.

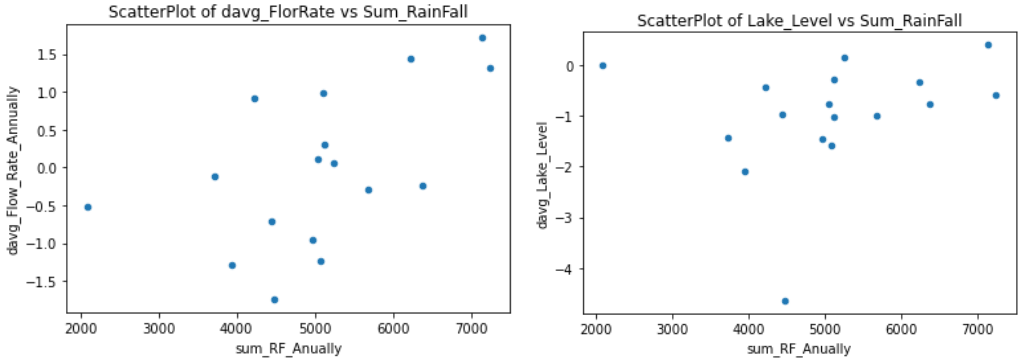
4.5 Number of rainfall affects the flow rate or water level of the water body

We first have to see if the distribution of rainfalls every year is relative normal



With kurtosis values= 0.3017652, which is small enough for us to conclude that the distribution of rainfall every year is relatively normal.

Then we want to see if the number of rainfalls can possibly predicts the lake water flow rate or lake water level



Both scatter plots show a positive correlation . where the regression line does helps us conclude the more rainfalls in the area the increase in water flow rate and lake water flow rate.

4.6 Analyze flow rate and water level between seasons

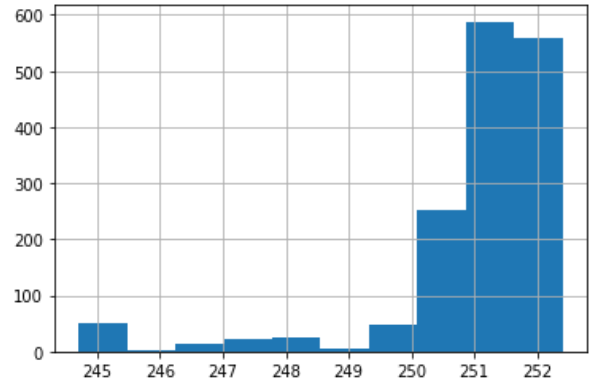
We wanted to evaluate if the different seasons had a significant effect on both the lake level and the flow rate of the river. We took the month given by the data parameter and parsed it into months. After that, we grouped the months into 3-month intervals and define those particular months as seasons. Spring was March to May. Summer was June to August. Autumn was September to November. Winter was December to February. After labeling specific observations, we were left with this count of observations for each season:

Spring	1564
Winter	1504
Summer	1502
Autumn	1456

We were glad to observe that the seasons were uniformly represented. Because of this, we assumed that we safely could run tests without concern that there would be bias due to significant differences in the size of data. Before we conducted any type of statistical testing, we needed to check the distribution of the data. In order to do that, we plotted the histograms of the lake levels by seasons.

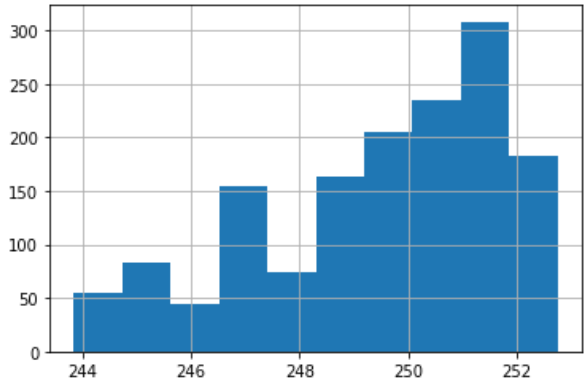
<u>Lake Level</u>

Lake Level Spring



Kurtosis coefficient: 7.908177509668862
Skewness: -2.746259672934576

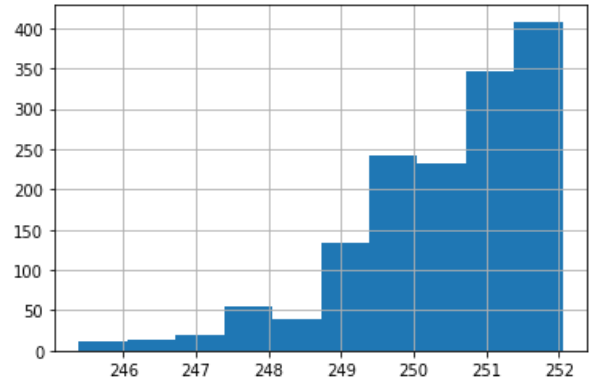
Lake Level Winter



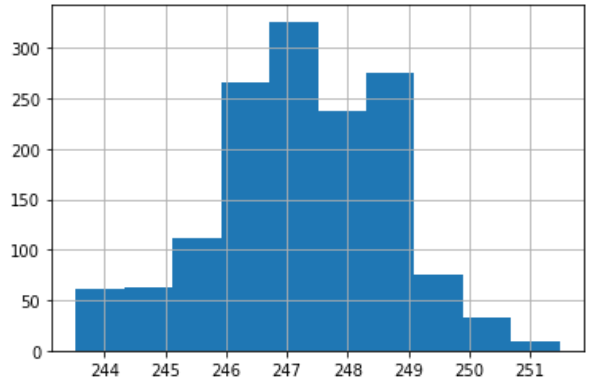
Kurtosis coefficient: -0.49416624536716514
Skewness: -0.7522149298567467

Lake Level Summer

Lake Level Autumn



Kurtosis coefficient: 1.3479750284887322
Skewness: -1.162173606084193



Kurtosis coefficient: -0.1288667126969134
Skewness: -0.23664741215027785

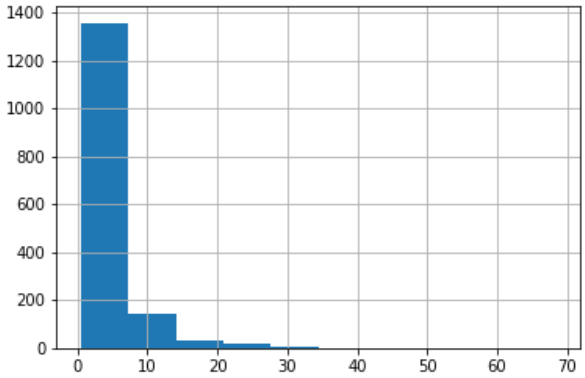
We realized that the distributions were not bell-shaped. Although the Lake Level during Autumn seems to like it is close to the normal distribution, the other seasons show a strong left skew. In order to verify this, we calculated the Kurtosis coefficient and skew which confirmed our initial assumption that all the seasons except for Autumn did not have a normal distribution. Because of the nature of the distributions, we knew that we weren't able to use standard student's t-tests to determine any significant difference between these groups. Instead, we decided that we should use the Kruskal test to evaluate the difference between the medians of each season to see if there could be any similarities or differences inferred.

Seasons Compared	**Statistic**	**P-Value**	**Reject/Accept**
Spring, Autumn	1815.3667	0.0	reject
Spring, Winter	439.7879	1.2039e-97	reject
Spring, Summer	316.4720	8.4993e-71	reject
Autumn, Winter	747.0003	1.8016e-164	reject
Autumn, Summer	1736.9111	0.0	reject
Winter, Summer	81.7568	1.5391e-19	reject

Based on the result of our Kruskal Wallis test, we are able to say that there is a statistically significant difference between each of the seasons to each other. This means that we can reject the notion that the medians of these seasonally based distributions are equal. We can also infer that the seasons have a significant effect on the lake level.

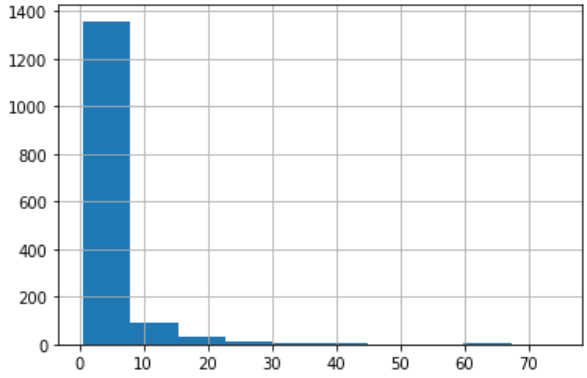
<u>Flow Rate</u>

Flow Rate Spring



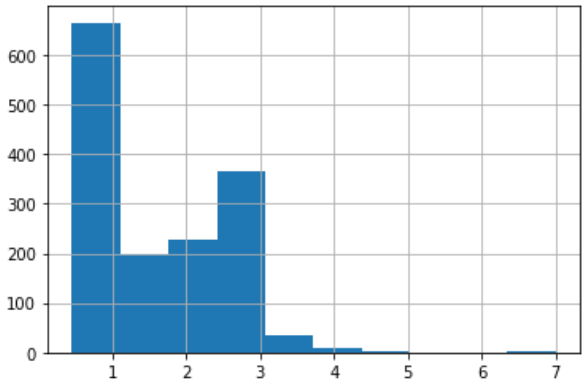
Kurtosis coefficient: 27.558922256853283
Skewness: 3.9065287852494053

Flow Rate Winter



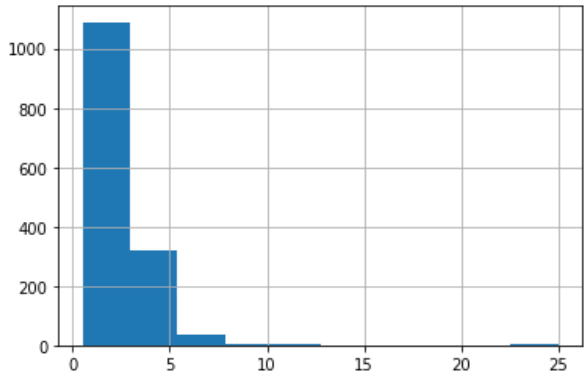
Kurtosis coefficient: 36.52927105368438
Skewness: 4.767458306303771

Flow Rate Summer



Kurtosis coefficient: 0.4743729364028657
Skewness: 0.6754771720981214

Flow Rate Autumn



Kurtosis coefficient: 7.908177509668862
Skewness: -2.746259672934576

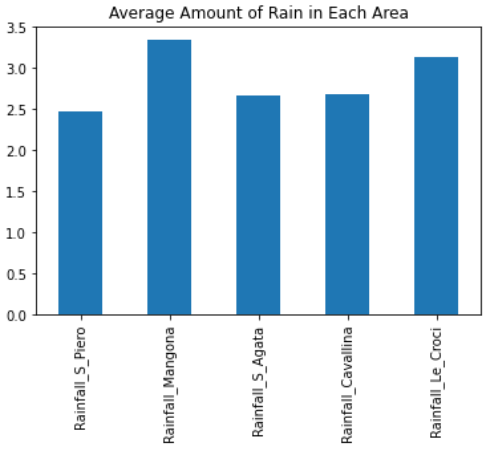
Similar to the Lake Level, the Flow Rate had an even stronger skew and we understood that the only way to evaluate the difference between these distributions was to run a statistical test that doesn't assume normality. Because of this, we also ran the Kruskal Wallis test on the Flow Rate.

Seasons Compared	**Statistic**	**P-Value**	**Reject/Accept**
Spring, Autumn	12.6810	0.0004	reject
Spring, Winter	11.7928	0.0006	reject
Spring, Summer	7.6998	0.0055	reject
Autumn, Winter	0.0013	0.9711	**accept**
Autumn, Summer	117.4666	2.2689e-27	reject
Winter, Summer	42.2913	7.8640e-11	reject

Based on the result of our Kruskal Wallis test, we are able to say that there is a statistically significant difference between each of the seasons to each other except for Autumn and Winter. This means that we can reject the notion that the medians of these seasonally based distributions are equal. However, the only two seasons that we can't reject would be Autumn and Winter. We can infer that the seasons have a significant effect on the lake level except for the differences between Winter and Autumn. This intuitively makes sense because Autumn and Winter are next to each other during the time, the lake is used to store water in preparation for Summer. Because of this, the H-statistic when comparing summer to Autumn and Summer is the highest which implies that they are the most different. This would make sense because the role of the lake is to fill up the Arno during the summers so I would expect the flow rates during the summer to be significantly different from all other seasons.

4.7 Investigate the amount of rainfall in each area

The amount of rainfall may have some effect on the water quality. The greater the rainfall of one area is, the water quality could become better or worse than another area. Here, we will analyze the amount of rainfall in each area and see which region has the amount of rainfall above or below average.



Rainfall_S_Piero	2.471635
Rainfall_Mangona	3.341212
Rainfall_S_Agata	2.670440
Rainfall_Cavallina	2.675187
Rainfall_Le_Croci	3.130390

Above is the graph and data of daily average rainfall for each area in Lake Bilancino, the average of the daily average rainfall is 2.8577728cm. We construct hypothesis testing for mean to see if some areas have higher or lower amounts of rainfall than other areas.

$$H_0 : \mu = 2.8577728$$
$$H_1 : \mu \neq 2.8577728$$

Area	z	P-value	Decision
S_Piero	-4.51	0.00001	Reject
Mangona	4.21	0.00001	Reject

S_Agata	-2.03	0.02118	Reject
Cavallina	-1.85	0.03216	Fail to Reject
Le_Croci	2.64	0.00415	Reject

At 95% confidence level, we reject the null hypothesis for all areas except Cavallina. We have sufficient evidence to claim that Mangona and Le Croci average daily rainfall is above average, while S Piero and S Agata average daily rainfall is below average. Since most of the sample mean are far differ from the population mean, we may also conclude that each area has a different amount of rainfall, which may result in different quality of water in different areas.

5. Theory

5.1 Normality Checks (Visual Methods)

The primary normality check in our study is graphical assessment of normality. Where histogram of data will be drawn. Visual inspection is usually unreliable and does not guarantee the normality of the distribution if the skewness of the data is extremely close to normal. However, when data is presented visually. We and the readers can judge the distribution assumption directly and immediately.

In our study we use the frequency distribution that plots the observed values against their frequencies to provide both a visual judgment of the distribution and insights about the gaps in the data and outliers.

5.2 Hypothesis Testing - Kruskal Wallis Test

The Kruskal Wallis test is the nonparametric equivalent to the One Way ANOVA. This means that the test doesn't assume your data comes from a normal distribution. This hypothesis test is used when the assumptions for ANOVA. The test is used to determine whether the medians of two or more groups are different. Similar to most statistical tests, it calculates the test statistic and compares it to a distribution point.

The test statistic used in this test is called the H statistic.

- $\mathbf{H_0}$: population medians are equal.
- $\mathbf{H_1}$: population medians are not equal.

$$H = (N - 1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{\bar{r}})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{\bar{r}})^2}$$

- $N$is the total number of observations across all groups

- g is the number of groups
- n_i is the number of observations in group i
- r_{ij} is the rank (among all observations) of observation j from group i
- $\overline{r}_i = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ is the average rank of all observations in group i
- $\overline{r} = \frac{1}{2}(N + 1)$ is the average of all the r_{ij}.

The Kruskal Wallis test will tell us if there is a significant difference between groups.

- **P-value ≤ α**: The differences between some of the medians are statistically significant
- **P-value > α**: The differences between the medians are not statistically significant

5.3 Regression Methods, Residual Plots

The scatter plot is a set of data points that are observed while the regression line is the prediction. When performing a simple linear regression, we use the best fit line to estimate or predict the future values of the dependent variable.

As residuals are the difference between any data point and the best fit regression line., we can express the residual with an equation:

Residual=Observed value- predicted value

In the case of our data, we construct the residual plots base on the data of average hydrometer, where overall data are grouped by year, and them calculate the average lake level and water flow of each year, for water flow, we will subtract the average water of each year by the overall water flow from 2004 to 2020 to obtain the change in rate.

5.4 Kurtosis test

The Kurtosis coefficient is used for checking the data for Normality. It is the same as the skewness coefficient, but is raised to the fourth power instead of the third. In symmetric distributions the skewness is zero, but with regards to the kurtosis coefficient, it's interpreted as the peak of distribution. For a normal distribution, a kurtosis coefficient is close to zero when calculating 1 sample.

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{x}}{s}\right)^4$$

Where n is the number of observations in the data, X is the data, x bar is the average of the data and S is the standardized data. When we run this hypothesis test we set the null hypothesis to: the data is normally distributed. The alternative hypothesis would be: the data is not normally distributed. We can deem the sample to not have a normal distribution if the original kurtosis coefficient (red dot) falls near the skewed end of the histogram of the simulated kurtosis coefficients. The original data would have had a normal distribution if the red dot was near the highest peak of the kurtosis coefficient distribution histogram.

6. Conclusion

From our investigation, it is possible, from 2004 to 2020, these years are generally cold and dry except 2012, where a hot and dry year with a drought might occur in the area. Rainfall does predict the changes in lake level and lake flow rate in a positive manner, however, there is a consistent drop in lake level over the years due to some environmental factors. These environmental factors are most likely due to changing weather conditions between seasons and due to climate change over the course of the past few decades. Through the use of the Kruskal Test we were able to see the significant effects that changing seasons can have on lake level, especially in the spring and summer months. For autumn and winter the change was found to be less prominent. Further analysis of the water levels throughout the past 20 years shows us that there was a large drop in water level in 2012, potentially indicative of a drought or perhaps a change in the water management system for Lake Bilancino.

Analysis of the rainfalls in Lake Bilancino led us to find that rainfalls in Le Croci seemed to have the most prominent relationship with the lake's measured flow rate. From between season analysis it could be seen that flow rate was the highest in the spring and summer months, most likely due to the lake's role in filling up the river Arno.

Our investigations about Lake Bilancino have shown that weather conditions play an active role in both the flow rate and lake level, both of which are important factors in the proper management of its water. Le Croci's rainfall area affects the flow rate the most, so we would suggest to have special attention and more resources dedicated to that area of the lake. Acea Water Group managers should also be prepared to have large changes in the spring and summer months and prepare accordingly. We hope that the variables that we have analyzed can play an important role in the potential development of any sophisticated machine learning prediction algorithms for the lake's water conditions in the future.

7. References

(1) B. Oskin, Scientific American (2014)
(2) X.-Y. Li, H.-Y. Xu, Y.-L. Sun, D.-S. Zhang, and Z.-P. Yang, Water Resources Management 21, 1505 (2006).
(3) https://www.britannica.com/science/lake/The-hydrologic-balance-of-the-lakes
(4) Managing water - Acea Group
(5) https://www.sciencedirect.com/science/article/pii/S1110982310000153?via%3Dihub
(6) http://www.aquaticinvasions.net/2006/AI_2006_1_4_Lori_Cianfanelli.pdf
(7) https://www.worldweatheronline.com/florence-weather-history/toscana/it.aspx
(8) https://www.theflorentine.net/2012/02/02/bilancino-running-dry/

8. Credits

Person	**Topic**
Benjamin Becze	Introduction, Investigation 4.1, Conclusion
Andrew Chin	Background, Investigation 4.4, Theory 5.4
Samuel Huang	Data Description, 4.6 Analyze flow rate and water level between seasons.
Benjamin Hu	Investigation 4.2,4.3,and 4.5, theory 5.1 and 5.3, last paragraph of background
Thomas Chen	Investigation 4.7