Acea Smart Water Analytics

MATH 189 Group 2 Final Project

Kasen Teoh, Chung En Pan, Nathan Fallahi,
Parsa Ganjooi, and Eamon Jarrett-Mann

Winter 2021

1 Introduction

Every day the number of humans on the planet continues to rise and with it, the consumption of natural resources, such as water. As humans, we use tens, if not hundreds, of gallons of water a day, in the restroom, kitchen, garden, and simply drinking. While it may seem like there is an abundance of water, mainly the oceans, we require fresh water, water from aquifers, springs, lakes, and rivers to sustain our lives. With the increasing amount of life on the planet, we need to make strategic decisions regarding where water should be extracted from and how to ration water portions to allow the water source to last. For instance, if a particular waterbody was close to being drained and we knew it would not quickly replenish naturally, we could avoid using the last of the water. Additionally, in different seasons, such as spring and summer, where rainfall might be scarce, conservation of water is critical. In this project, we aim to analyze water availability in four different types of water bodies, aquifers, springs, rivers, and lakes, and attempt to predict the availability on a daily basis. To achieve this analysis, we break our goal into several parts:

1. Analyze the seasonal rainfall throughout the entirety of the study by using both graphical visualizations and interval estimates of seasonal rainfall

2. Analyze the connections between the variables in the dataset and the variables to predict (Depth to groundwater for aquifers, flow rate for springs, lake level and flow rate for lakes, and hydrometry for rivers). Here, we can also identify which variables affect the outcome variables significantly

3. Determine whether a linear regression model is suitable for the datasets in question

2 Data

Our data was sourced from a Kaggle Analytics Competition (*Acea Smart Water Analytics*, 2020). There are a total of nine datasets consisting of four aquifers, three springs, one lake, and one river which each contain various features. Not all days have the same measurements as some features were only recorded for certain dates. All datasets have measurements from a specific date until June 6, 2020. Each measurement includes the date it was measured on as a variable. For all the datasets, rainfall is measured in millimeters, depth to groundwater in meters, temperature in degrees Celsius, volume in cubic meters, hydrometry (groundwater level) in meters, flow rate in cubic meters per second, and lake level (river water level) in meters. All measurements/features are continuous numerical variables, except for the date which is discrete and numerical.

The Auser aquifer dataset has measurements for 8,154 days since 1998. There are measurements of rainfall at 10 locations, depth to rainfall at 5 locations, temperature at 4 locations, volume at 5 locations, and hydrometry at 2 locations. The Doganella aquifer dataset has measurements for 6,026 days since 2004. There are 2 rainfall, 9 depth to groundwater, 9 volume, and 2 temperature measurements. The Luco aquifer dataset has measurements for 7,487 days since 2000. There are 10 rainfall, 4 depth to groundwater, 4 temperature, and 4 volume measurements. The Petrignano aquifer dataset has measurements for 5,223 days since 2006. There are 1 rainfall, 2 depth to groundwater, 2 temperature, 1 volume, and 1 hydrometry measurements.

The Bilancino lake dataset has measurements for 6,603 days since 2002. There are 5 rainfall, 1 temperature, 1 lake level, and 1 flow rate measurements.

The Arno river dataset has measurements for 8,217 days since 1998. There are 14 rainfall, 1 temperature, and 1 hydrometry measurements.

The Amiata water spring dataset has measurements for 7,487 days since 2000. There are 5 rainfall, 3 depth to groundwater, 3 temperature, and 4 flow rate measurements. The Lupa water spring dataset has measurements for 4,199 days since 2009. There is 1 rainfall and 1 flow rate measurement. The Madonna di Canneto water spring dataset has measurements for 3,113 days since 2012. There is 1 rainfall, 1 temperature, and 1 flow rate measurement.

3 Background

Water availability is defined as the amount of water that humans can extract without harmful repercussions to the ecosystem and other organisms. This is basically saying how much water can we use without harming the environment or organisms that live in the environment.

An aquifer is a body of rock or sediment that holds groundwater. Groundwater is the precipitation that has combined with the soil below the surface and collected in empty spaces underground. Most groundwater, including a significant amount of our drinking water, comes from aquifers. In order to access the groundwater, a well must be created by drilling a hole that reaches the aquifer. While wells extract water from aquifers, aquifers also flow to other springs. If groundwater abstraction exceeds groundwater recharge for extensive areas and long time, overexploitation or persistent groundwater depletion can occur (Gleeson et al., 2010). Since the 1960s groundwater abstraction has more than doubled, resulting in an increase in groundwater depletion. (Wada et al., 2010). From the 1960's to 2010, the demand for groundwater had doubled, consequently the abstraction of groundwater doubled, leading to a depletion of groundwater. If this rate continues, we will have depleted our groundwater resources in the near future.

According to the US Geology Survey, a spring is a water resource formed from an aquifer being filled to the point that the water overflows onto the land surface. Because springs are derived from when aquifers' groundwater reaches the surface, the depletion of groundwater directly impacts springs. With the disappearance of aquifers and groundwater, there is no more water underground to flow to springs, and hence, the disappearance of springs follows. For instance, in the Nubian Aquifer system about 25% of the total withdrawals in 1998 resulted in reductions in natural flow to other water bodies, such as springs and oases (*Regional National Strategy for the Utilisation of The Nubian Sandstone Aquifer System*, 2001).

Lakes are bodies of water that are surrounded by land. If a lake flows into a river or ocean, it is considered an opened lake while if it does not flow into another water body, it is considered closed. Lakes' water supplies are replenished through groundwater seepage and/or natural means, i.e rain, snow, etc. Many organisms, not limited to just humans, depend on lakes as a cornerstone of life: fish, plants, birds, and etc. According to Inland water loss in the Aral Sea, Lake Urmia, the Great Salt Lake, Lake Abert, and Lake Poopo accounts for more than half of total inland water loss in terminal lakes on several of Earth's continents (Wine & Laronne, 2020).

A river is a natural flowing body of water that flows towards oceans, lakes, or other rivers. Every drop of water in rivers flowing across the land once fell to Earth from the atmosphere and given enough time, every drop will return to the atmosphere as water vapor (Karr & Chu, 1998). However, if humans continue down the current path and are not wise about the extraction of water, rivers could dry up permanently, further depleting the resource. Retaining the biological parts of riverine ecosystems, as well as the processes that nourish them is crucial to retaining water supplies and all the goods and services associated with water (Karr & Chu, 1998).

These four water bodies are essential to human life. While there are many more water bodies, such as oceans, we aim to highlight which season would be the smartest time to extract water from each of these four water bodies, allowing for the continuation for human growth, while also maintaining a hospitable ecosystem for other life and organisms as well.

4 Analysis

4.1 Aquifers

For Aquifers we have a total of four datasets: Auser, Petrigano, Doganella, and Luco.

4.1.1 Auser

The Auser aquifer consists of two subsystems, north and south. The north subsystem is a water table (or unconfined) aquifer while the south subsystem is an artesian (or confined) groundwater. The north subsystem partly influences the behavior of the south subsystem.

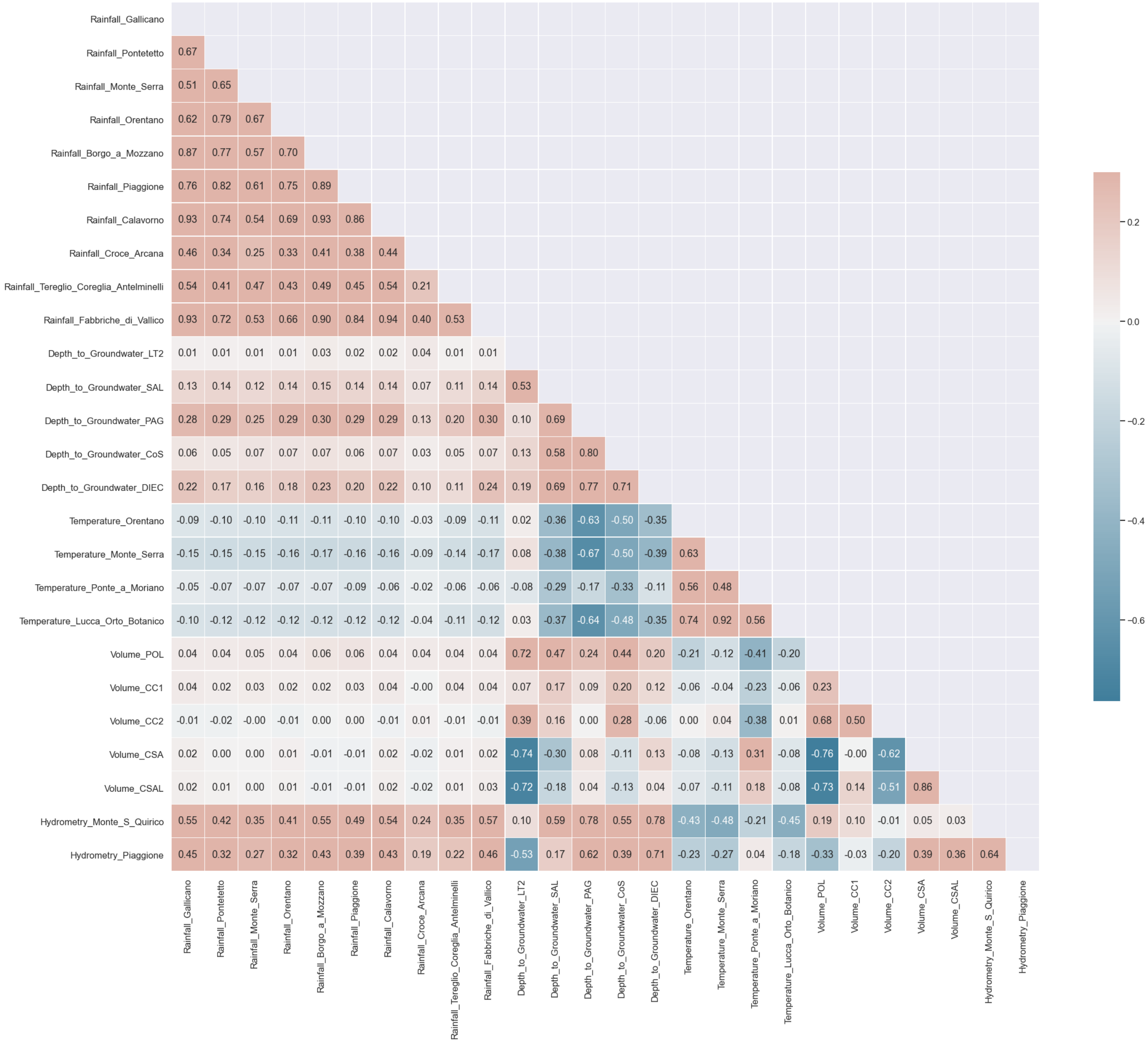

Figure 1: Auser Aquifer Correlation Heatmap

This heatmap shows a correlation matrix for all the variables in the Auser aquifer dataset. As you can see, many of the rainfalls are correlated across regions, which makes sense as they are likely close to each other. Additionally, it seems like the depth to groundwater (outcome) measures are not well correlated with other features other than the other depths to groundwater.

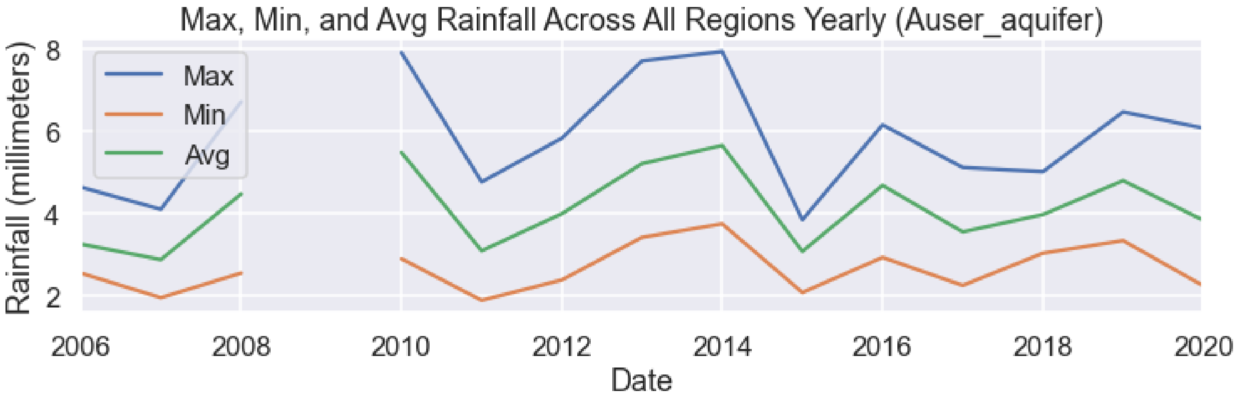

Figure 2: Auser Aquifer Rainfall

As you can see above, the Auser aquifer has a wide variety of rainfall meaures, with some years receiving as much as 5 mm on average and other years staying much drier with half the volume of rainfall. This graph gives a good idea of how the rainfall changes from year to year. Additionally, we see that there was no data collected in the year 2009.

Region	Winter	Spring	Summer	Autumn
Rainfall_Gallicano	[6.003, 7.76]	[1.501, 1.94]	[3.002, 3.88]	[4.502, 5.82]
Rainfall_Pontetetto	[4.248, 5.505]	[1.062, 1.376]	[2.124, 2.753]	[3.186, 4.129]
Rainfall_Monte_Serra	[4.797, 6.269]	[1.199, 1.567]	[2.399, 3.134]	[3.598, 4.702]
Rainfall_Orentano	[3.767, 4.943]	[0.942, 1.236]	[1.884, 2.472]	[2.826, 3.707]
Rainfall_Borgo_a_Mozzano	[5.689, 7.314]	[1.422, 1.828]	[2.845, 3.657]	[4.267, 5.485]
Rainfall_Piaggione	[4.835, 6.243]	[1.209, 1.561]	[2.418, 3.121]	[3.626, 4.682]
Rainfall_Calavorno	[5.539, 7.169]	[1.385, 1.792]	[2.769, 3.585]	[4.154, 5.377]
Rainfall_Croce_Arcana	[2.918, 4.192]	[0.729, 1.048]	[1.459, 2.096]	[2.188, 3.144]
Rainfall_Tereglio_Coreglia_Antelminelli	[4.983, 6.517]	[1.246, 1.629]	[2.491, 3.258]	[3.737, 4.888]
Rainfall_Fabbriche_di_Vallico	[7.273, 9.359]	[1.818, 2.34]	[3.636, 4.68]	[5.455, 7.02]
Temperature_Orentano	[13.311, 18.83]	[3.328, 4.708]	[6.655, 9.415]	[9.983, 14.123]
Temperature_Monte_Serra	[11.707, 15.876]	[2.927, 3.969]	[5.853, 7.938]	[8.78, 11.907]
Temperature_Ponte_a_Moriano	[10.191, 15.515]	[2.548, 3.879]	[5.096, 7.757]	[7.643, 11.636]
Temperature_Lucca_Orto_Botanico	[16.481, 21.822]	[4.12, 5.455]	[8.24, 10.911]	[12.361, 16.366]
Volume_POL	[-13712.149, -10578.761]	[-3428.037, -2644.69]	[-6856.074, -5289.38]	[-10284.112, -7934.071]
Volume_CC1	[-23675.845, -18787.258]	[-5918.961, -4696.814]	[-11837.923, -9393.629]	[-17756.884, -14090.443]
Volume_CC2	[-17184.123, -13582.672]	[-4296.031, -3395.668]	[-8592.062, -6791.336]	[-12888.093, -10187.004]
Volume_CSA	[-4384.341, -2222.9]	[-1096.085, -555.725]	[-2192.17, -1111.45]	[-3288.256, -1667.175]
Volume_CSAL	[-3558.796, -1701.984]	[-889.699, -425.496]	[-1779.398, -850.992]	[-2669.097, -1276.488]
Hydrometry_Monte_S_Quirico	[0.382, 0.476]	[0.096, 0.119]	[0.191, 0.238]	[0.287, 0.357]
Hydrometry_Piaggione	[-0.436, 0.034]	[-0.109, 0.009]	[-0.218, 0.017]	[-0.327, 0.026]

Table 1: Auser Aquifer Prediction Intervals

Above, we see the generated prediction intervals of the estimated seasonal rainfall, temperature, volume, and hydrometry. Using these prediction intervals along with the correlations and regression coefficients will allow us to determine what affects the depth to groundwater the most and what season will be optimal for water extraction.

```
Linear regression PCA models for Auser_aquifer
    RMSE when predicting the Depth_to_Groundwater_LT2 1.144
    RMSE when predicting the Depth_to_Groundwater_SAL 1.131
    RMSE when predicting the Depth_to_Groundwater_PAG 0.805
    RMSE when predicting the Depth_to_Groundwater_CoS 2.074
    RMSE when predicting the Depth_to_Groundwater_DIEC 0.777
```

When predicting depth to ground water using linear regression models, there is a bit of error for all depths, but CoS is significantly worse (about 2x as much error) than any of the other presented linear regression models.

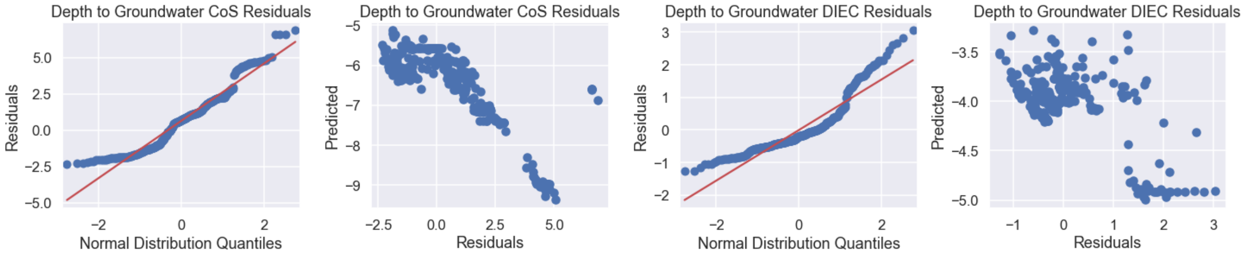

Figure 3: Auser Aquifer Residuals

Above, we see two of the five residual plots. In both cases, we see that the distribution of the residuals does not follow a normal distribution. The Cos residuals scatter plots shows a negative sloped scatter plot while the DIEC residuals scatter plots dot not should much of an obvious patter and is somewhat centered around 0.

Predictor	**Coefficient**	**Standard Error**	**P-Value**
Rainfall_Gallicano	3.046	0.202	0.0
Rainfall_Monte_Serra	-0.011	0.005	0.042
Rainfall_Orentano	0.021	0.005	0.0
Rainfall_Piaggione	-0.01	0.004	0.016
Rainfall_Calavorno	-0.012	0.004	0.002
Rainfall_Croce_Arcana	-0.006	0.003	0.016
Temperature_Ponte_a_Moriano	0.245	0.03	0.0
Temperature_Lucca_Orto_Botanico	-0.041	0.009	0.0
Volume_POL	-0.233	0.031	0.0
Volume_CC1	0.026	0.003	0.0
Volume_CC2	0.0	0.0	0.0
Volume_CSA	0.0	0.0	0.041
Volume_CSAL	0.0	0.0	0.0
Hydrometry_Piaggione	0.001	0.0	0.0

Table 2: Predicting Depth to Groundwater CoS

Predictor	Coefficient	Standard Error	P-Value
Rainfall_Gallicano	0.603	0.1	0.0
Rainfall_Pontetetto	0.871	0.041	0.0
Rainfall_Orentano	0.008	0.003	0.003
Rainfall_Piaggione	-0.01	0.002	0.0
Rainfall_Calavorno	-0.008	0.002	0.0
Rainfall_Croce_Arcana	-0.004	0.001	0.003
Temperature_Monte_Serra	-0.002	0.001	0.038
Temperature_Ponte_a_Moriano	0.046	0.015	0.002
Temperature_Lucca_Orto_Botanico	-0.011	0.004	0.014
Volume_POL	-0.039	0.015	0.009
Volume_CC1	0.018	0.001	0.0
Volume_CC2	0.0	0.0	0.0
Volume_CSA	0.0	0.0	0.0
Volume_CSAL	0.0	0.0	0.0
Hydrometry_Monte_S_Quirico	-0.0	0.0	0.002
Hydrometry_Piaggione	0.0	0.0	0.0

Table 3: Predicting Depth to Groundwater DIEC

As you can see from the tables above (Table 2 and 3), most of the features in the dataset are not strongly correlated with the varied depth outcomes. Above, we displayed two out of the five coefficients as they were relatively similar. In fact, even the other depths are not strongly correlated with the depth outcomes, and many are not statistically significant. The two constant and strong correlated throughout all five regression models was the rainfall in the regions of Gallicano and Pontetto. This was not too surprising as the these two were constantly held the highest Pearson correlation coefficient with the depth to ground water variables.

4.1.2 Petrigano

The Petrigano aquifer is water table groundwater and is also fed by the Chiascio river. The grounderwater levels are influences by rainfall, depth to groundwater, temperatures, drainage volumes, and level of the Chiascio river.

Variable	**Non-Null Count**
Rainfall_Bastia_Umbra	4199
Depth_to_Groundwater_P24	5168
Depth_to_Groundwater_P25	5184
Temperature_Bastia_Umbra	4199
Temperature_Petrignano	4199
Volume_C10_Petrignano	5025
Hydrometry_Fiume_Chiascio_Petrignano	4199

Table 4: Petrigano Aquifer Dataset

Above we see that this is one of the few complete datasets that we are given. For the predictor variables we have between 4000-5000 observations and for the outcome variables we have around 5000 observations. Hence, we do not need to drop any big pieces of data.

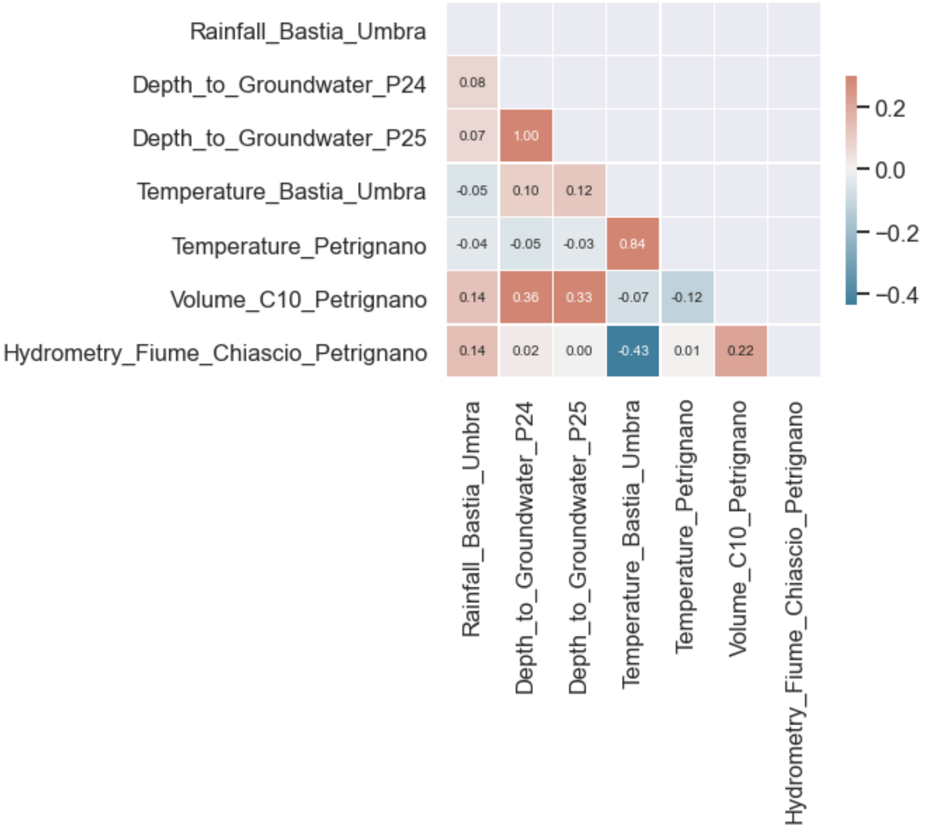

Figure 4: Petrigano Aquifer Correlation Heatmap

As you can see from the heatmap above, many of the rainfalls are correlated across regions, which makes sense as they are likely close to each other. Additionally, it seems like the depth to groundwater (outcome) measures are not well correlated with other features other than the other depths to groundwater.

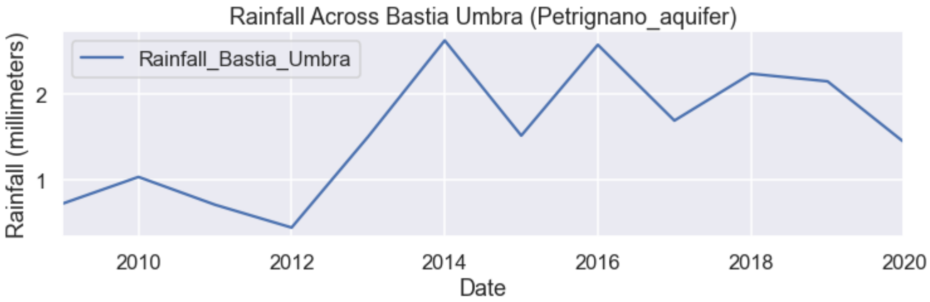

Figure 5: Petrigano Aquifer Rainfall

As you can see above, the Petrignano aquifer has a less variety of rainfall measures than most of the other aquifers, with values between .5mm and 2.5 mm. This graph gives a good idea of how the rainfall changes from year to year, and seems relatively consistent with the Auser number as 2014 is a relatively wet year in these regions as well.

Region	**Winter**	**Spring**	**Summer**	**Autumn**
Rainfall_Bastia_Umbra	[1.476, 2.403]	[0.369, 0.601]	[0.738, 1.202]	[1.107, 1.802]
Temperature_Bastia_Umbra	[14.467, 21.935]	[3.617, 5.484]	[7.234, 10.967]	[10.851, 16.451]
Temperature_Petrignano	[13.314, 20.347]	[3.328, 5.087]	[6.657, 10.174]	[9.985, 15.26]
Volume_C10_Petrignano	[-43118.272, -34054.395]	[-10779.568, -8513.599]	[-21559.136, -17027.197]	[-32338.704, -25540.796]
Hydrometry_Fiume_Chiascio_Petrignano	[2.764, 3.589]	[0.691, 0.897]	[1.382, 1.795]	[2.073, 2.692]

Table 5: Petrigano Aquifer Prediction Intervals

Above is the prediction intervals for the Petrignano aquifer. We see that in the Summer and Autumn months, the rainfall is fairly constant while in the Winter it could be double.

```
Linear regression PCA models for Petrignano_aquifer
    RMSE when predicting the Depth_to_Groundwater_P24 2.339
    RMSE when predicting the Depth_to_Groundwater_P25 2.234
```

The RMSE for these models is uniformly higher than the presented models for the Auser aquifer.

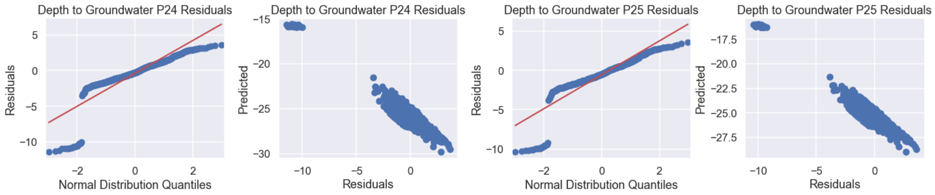

Figure 6: Petrigano Aquifer Residuals

Once again, we see that the residuals are not normally distributed and not randomly scatted. This leads us to believe that a linear model is not suitable for our data.

Predictor	Coefficient	Standard Error	P-Value
Rainfall_Bastia_Umbra	-3.395	0.081	0.0
Temperature_Petrignano	-0.258	0.014	0.0
Volume_C10_Petrignano	0.206	0.014	0.0
Hydrometry_Fiume_Chiascio_Petrignano	0.001	0.0	0.0

Table 6: Predicting Depth to Groundwater at P24

Predictor	Coefficient	Standard Error	P-Value
Rainfall_Bastia_Umbra	-3.504	0.08	0.0
Temperature_Petrignano	-0.266	0.014	0.0
Volume_C10_Petrignano	0.218	0.014	0.0
Hydrometry_Fiume_Chiascio_Petrignano	0.001	0.0	0.0

Table 7: Predicting Depth to Groundwater at P25

The Petrignano aquifer only has 2 depth measures, both of which have strong negative correlation with rainfall in Bastia Umbra and a small amount of negative correlation with the temperature in the Petrignano region. Further, there is a small amount of positive correlation with the Volume c10 Petrignano. As you can see in Table 6 and 7, there is little correlation with any of the other features present in the dataset.

4.1.3 Doganella

The wells field Doganella is fed by two underground aquifers not fed by rivers or lakes but fed by meteoric infiltration. The upper aquifer is a water table with a thickness of about 30m. The lower aquifer is a semi-confined artesian aquifer with a thickness of 50m and is located inside lavas and tufa products. These aquifers are accessed through nine wells. Approximately 80% of the drainage volumes come from the artesian aquifer. The aquifer levels are influenced by the following parameters: rainfall, humidity, subsoil, temperatures and drainage volumes.

Variable	Non-Null Count
Rainfall_Monteporzio	5399
Rainfall_Velletri	5374
Depth_to_Groundwater_Pozzo_1	2537
Depth_to_Groundwater_Pozzo_2	2736
Depth_to_Groundwater_Pozzo_3	2774
Depth_to_Groundwater_Pozzo_4	2374
Depth_to_Groundwater_Pozzo_5	2508
Depth_to_Groundwater_Pozzo_6	2428
Depth_to_Groundwater_Pozzo_7	2311
Depth_to_Groundwater_Pozzo_8	2551
Depth_to_Groundwater_Pozzo_9	2339
Volume_Pozzo_1	1356
Volume_Pozzo_2	1360
Volume_Pozzo_3	1360
Volume_Pozzo_4	1360
Volume_Pozzo_5+6	1360
Volume_Pozzo_7	1360
Volume_Pozzo_8	1360
Volume_Pozzo_9	1360
Temperature_Monteporzio	4564
Temperature_Velletri	4383

Table 8: Doganella Aquifer Dataset

In the above output, we see the Doganella aquifer dataset. This dataset is particularly large with 9 outcome variables and 12 predictor variables. Because the number of valid data is fairly similar throughout the dataset, around 1000-2000, we do not need to drop any variable entirely.

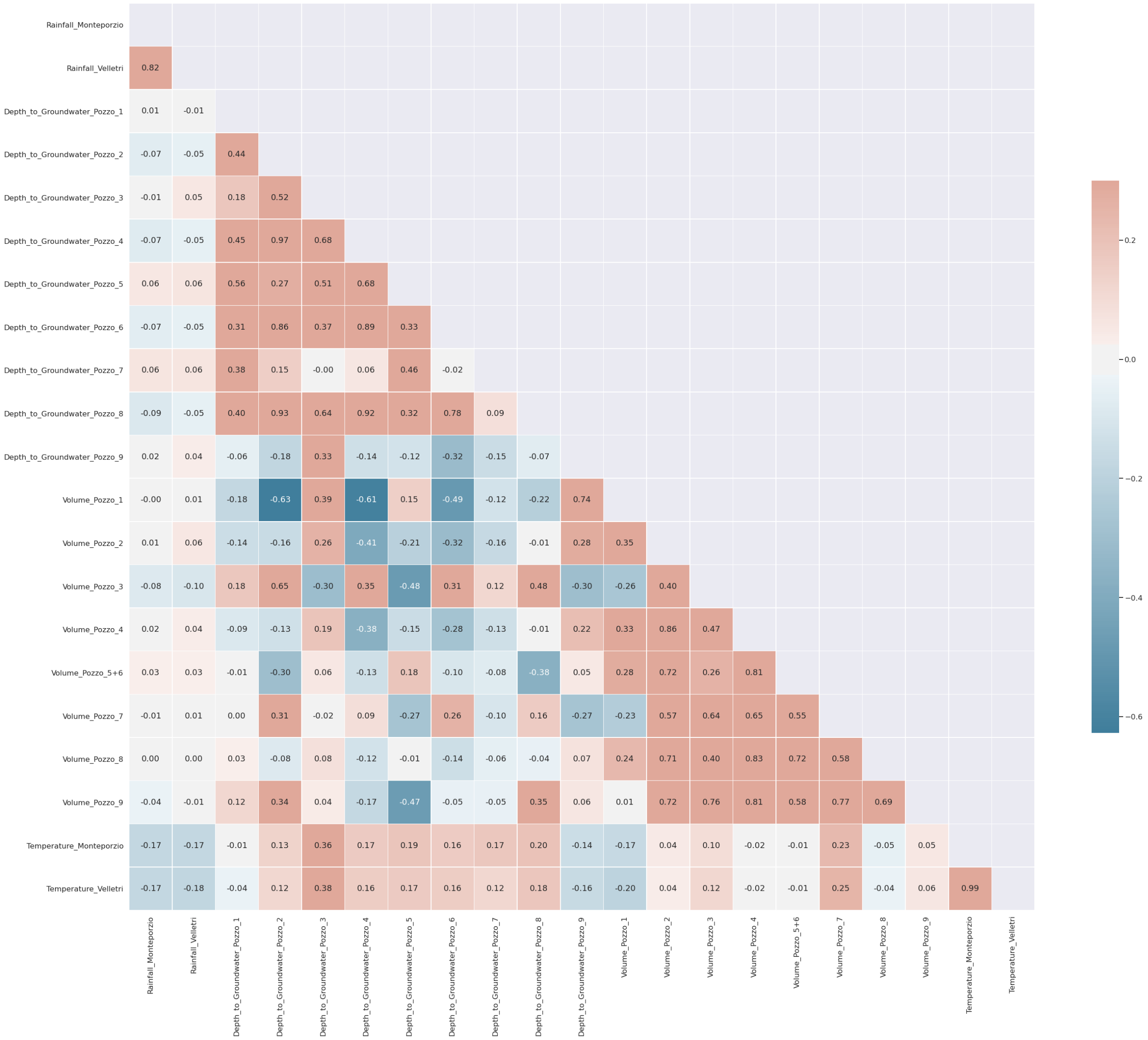

Figure 7: Doganella Aquifer Correlation Heatmap

There are only 2 rainfall features for this aquifer, with a variety of volume measures and a ton of (9 total!) depth meausures. As you can see from the heatmap above, some of the depths are strongly correlated with each other as are some of the volumes, but not some of the outcomes have little correlation with the other features.

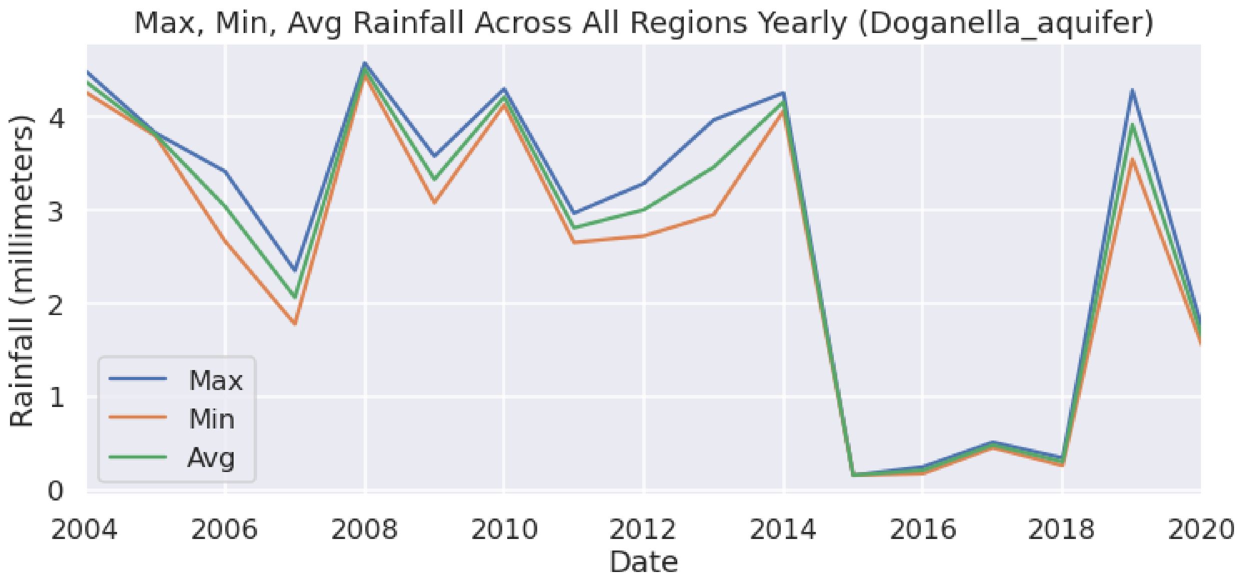

Figure 8: Doganella Aquifer Rainfall

The rainfall for the 2 regions present in this dataset roughly splits the difference between the Auser and Petrignano aquifers, with a good amount of volume in most years, but very little in 2015-2018.

Region	**Winter**	**Spring**	**Summer**	**Autumn**
Rainfall_Monteporzio	[3.044, 4.364]	[0.761, 1.091]	[1.522, 2.182]	[2.283, 3.273]
Rainfall_Velletri	[3.415, 4.816]	[0.854, 1.204]	[1.708, 2.408]	[2.562, 3.612]
Volume_Pozzo_1	[1180.588, 2678.48]	[295.147, 669.62]	[590.294, 1339.24]	[885.441, 2008.86]
Volume_Pozzo_2	[3582.7, 5506.575]	[895.675, 1376.644]	[1791.35, 2753.287]	[2687.025, 4129.931]
Volume_Pozzo_3	[3249.697, 5527.544]	[812.424, 1381.886]	[1624.849, 2763.772]	[2437.273, 4145.658]
Volume_Pozzo_4	[3695.168, 5615.466]	[923.792, 1403.866]	[1847.584, 2807.733]	[2771.376, 4211.599]
Volume_Pozzo_5+6	[6904.481, 10799.117]	[1726.12, 2699.779]	[3452.241, 5399.559]	[5178.361, 8099.338]
Volume_Pozzo_7	[2592.443, 4009.952]	[648.111, 1002.488]	[1296.222, 2004.976]	[1944.333, 3007.464]
Volume_Pozzo_8	[3830.988, 5793.899]	[957.747, 1448.475]	[1915.494, 2896.95]	[2873.241, 4345.424]
Volume_Pozzo_9	[3401.738, 5294.93]	[850.434, 1323.732]	[1700.869, 2647.465]	[2551.303, 3971.197]
Temperature_Monteporzio	[13.428, 18.996]	[3.357, 4.749]	[6.714, 9.498]	[10.071, 14.247]
Temperature_Velletri	[15.851, 21.945]	[3.963, 5.486]	[7.926, 10.972]	[11.889, 16.459]

Table 9: Doganella Aquifer Prediction Intervals

We see above that in the two rainfalls are similar for each season. However, when looking at the different volumes, we see that the volume 5 and 6 has the greater interval of all.

```
Linear regression PCA models for Doganella_aquifer
    RMSE when predicting the Depth_to_Groundwater_Pozzo_1 5.556
    RMSE when predicting the Depth_to_Groundwater_Pozzo_2 0.487
    RMSE when predicting the Depth_to_Groundwater_Pozzo_3 9.705
    RMSE when predicting the Depth_to_Groundwater_Pozzo_4 0.388
    RMSE when predicting the Depth_to_Groundwater_Pozzo_5 0.068
    RMSE when predicting the Depth_to_Groundwater_Pozzo_6 3.686
    RMSE when predicting the Depth_to_Groundwater_Pozzo_7 0.597
    RMSE when predicting the Depth_to_Groundwater_Pozzo_8 1.149
    RMSE when predicting the Depth_to_Groundwater_Pozzo_9 1.411
```

The RMSE for these models is varies quite a bit. For example, the 5th depth is two orders of magnitude less than the 3rd. This means that our linear model might be suitable for a few of the predictions, but not all.

Predictor	Coefficient	Standard Error	P-Value
Volume_Pozzo_1	6.648	1.309	0.0
Volume_Pozzo_2	-6.503	1.318	0.0
Volume_Pozzo_3	-0.005	0.001	0.0
Volume_Pozzo_5+6	-0.005	0.001	0.0
Temperature_Monteporzio	-0.017	0.005	0.001

Table 10: Predicting Depth to Groundwater Pozzo 3

Predictor	Coefficient	Standard Error	P-Value
Volume_Pozzo_1	5.02	1.077	0.0
Volume_Pozzo_2	-5.032	1.084	0.0
Volume_Pozzo_5+6	-0.002	0.001	0.04
Temperature_Monteporzio	-0.016	0.004	0.0

Table 11: Predicting Depth to Groundwater Pozzo 8

As you can see from the tables above (Table 10 and 11), the Doganella aquifer has the most depth measurements by far: a total of 9! For all of them the largest coefficients are on volume Pozzo 1 & volume Pozzo 2, and they have large positive coefficients with around 5 to 6 in most cases.

4.1.4 Luco

This is an underground aquifer which is not fed by rivers or lakes but fed by meteoric infiltration. It is accessed through wells called Pozzo1, Pozzo3, and Pozzo4.

Variable	**Non-Null Count**
Rainfall_Simignano	6822
Rainfall_Siena_Poggio_al_Vento	951
Rainfall_Mensano	1722
Rainfall_Montalcinello	6525
Rainfall_Monticiano_la_Pineta	2205
Rainfall_Sovicille	6657
Rainfall_Ponte_Orgia	1260
Rainfall_Scorgiano	3036
Rainfall_Pentolina	2116
Rainfall_Monteroni_Arbia_Biena	3104
Depth_to_Groundwater_Podere_Casetta	3346
Depth_to_Groundwater_Pozzo_1	1012
Depth_to_Groundwater_Pozzo_3	920
Depth_to_Groundwater_Pozzo_4	969
Temperature_Siena_Poggio_al_Vento	7487
Temperature_Mensano	7487
Temperature_Pentolina	7487
Temperature_Monteroni_Arbia_Biena	7487
Volume_Pozzo_1	2008
Volume_Pozzo_3	2008
Volume_Pozzo_4	2008

Table 12: Luco Aquifer Dataset

Above, we see the information for the dataset of aquifer Luco. The variables that we want to study are the four depth to groundwater variables against the other variables. However, we see that the depth to groundwater variables have around 1000-3000 observations while the other predictor variables have anywhere from 1000-7000 observations. Because with an unknown outcome value, we are unable to utilize the observation, we drop any observation day that is missing values in all four depth to groundwater measurements.

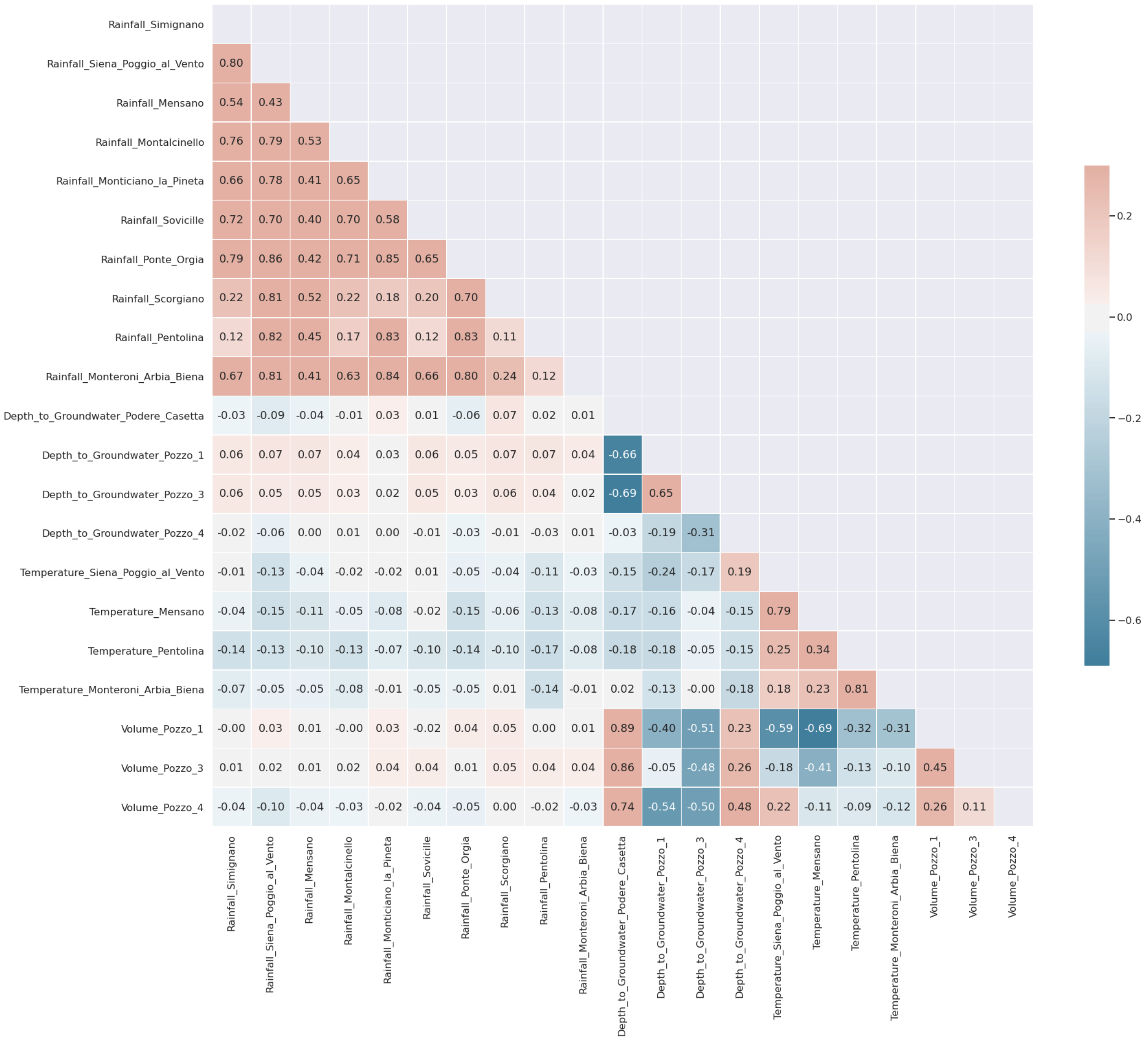

Figure 9: Luco Aquifer Correlation Heatmap

The heatmap above, most of the features in this dataset have small (absolute) correlation, with lots of correlation with the same variable (i.e the rainfall in one region is strongly correlated with rainfall in another region, but not with manyo of the other types of features).

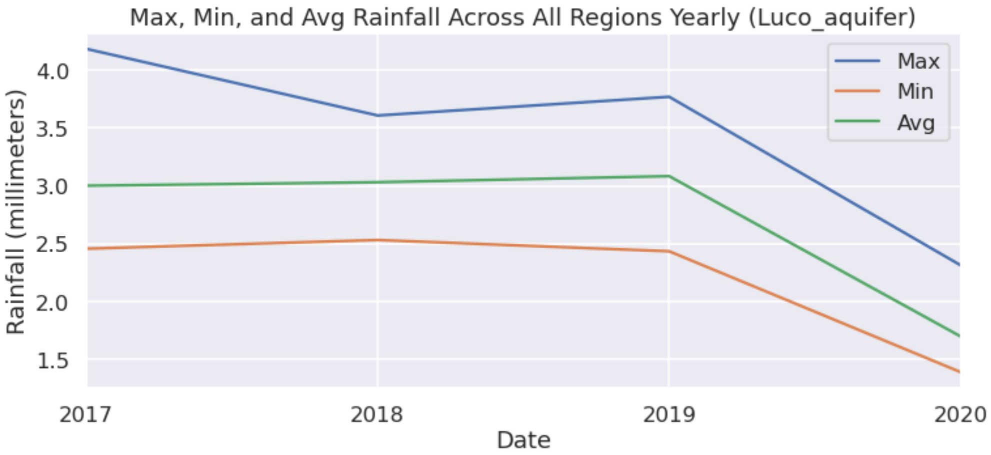

Figure 10: Luco Aquifer Rainfall

This dataset has by far the fewest number of observations, with only 3 years of data present. The rainfall numbers seem roughly consistent with the Doganella aquifer, although there are far more rainfall regions present in this dataset.

Region	**Winter**	**Spring**	**Summer**	**Autumn**
Rainfall_Simignano	[2.713, 4.317]	[0.678, 1.079]	[1.357, 2.159]	[2.035, 3.238]
Rainfall_Siena_Poggio_al_Vento	[1.167, 3.213]	[0.292, 0.803]	[0.583, 1.607]	[0.875, 2.41]
Rainfall_Mensano	[1.653, 4.654]	[0.413, 1.164]	[0.826, 2.327]	[1.24, 3.491]
Rainfall_Montalcinello	[2.742, 4.308]	[0.685, 1.077]	[1.371, 2.154]	[2.056, 3.231]
Rainfall_Monticiano_la_Pineta	[2.634, 5.231]	[0.658, 1.308]	[1.317, 2.616]	[1.975, 3.923]
Rainfall_Sovicille	[2.532, 4.146]	[0.633, 1.037]	[1.266, 2.073]	[1.899, 3.11]
Rainfall_Ponte_Orgia	[2.139, 4.666]	[0.535, 1.167]	[1.07, 2.333]	[1.604, 3.5]
Rainfall_Scorgiano	[3.055, 6.115]	[0.764, 1.529]	[1.528, 3.057]	[2.292, 4.586]
Rainfall_Pentolina	[3.743, 20.617]	[0.936, 5.154]	[1.872, 10.309]	[2.807, 15.463]
Rainfall_Monteroni_Arbia_Biena	[2.179, 3.839]	[0.545, 0.96]	[1.089, 1.92]	[1.634, 2.88]
Temperature_Siena_Poggio_al_Vento	[2.018, 7.508]	[0.504, 1.877]	[1.009, 3.754]	[1.513, 5.631]
Temperature_Mensano	[4.259, 10.596]	[1.065, 2.649]	[2.13, 5.298]	[3.195, 7.947]
Temperature_Pentolina	[13.544, 20.069]	[3.386, 5.017]	[6.772, 10.035]	[10.158, 15.052]
Temperature_Monteroni_Arbia_Biena	[13.995, 20.288]	[3.499, 5.072]	[6.997, 10.144]	[10.496, 15.216]
Volume_Pozzo_1	[-250.543, -155.802]	[-62.636, -38.95]	[-125.271, -77.901]	[-187.907, -116.851]
Volume_Pozzo_3	[-234.631, -143.639]	[-58.658, -35.91]	[-117.315, -71.82]	[-175.973, -107.73]
Volume_Pozzo_4	[-214.29, -128.153]	[-53.572, -32.038]	[-107.145, -64.077]	[-160.717, -96.115]

Table 13: Luco Aquifer Prediction Intervals

Above, we see the prediction intervals of the predictor variables. The region of Pentolina has an extremely large interval for the winter months with as much as 30.617 millimeters of rainfall. We believe that this large interval be my due to a few outliers.

```
Linear regression PCA models for Luco_aquifer
    RMSE when predicting the Depth_to_Groundwater_Podere_Casetta 0.211
    RMSE when predicting the Depth_to_Groundwater_Pozzo_1 0.637
    RMSE when predicting the Depth_to_Groundwater_Pozzo_3 0.525
    RMSE when predicting the Depth_to_Groundwater_Pozzo_4 1.063
```

The RMSE for these particular models is somewhat good. Meaning that they are close to the threshold of 0.5 with the exception of predicting the depth to groundwater at Pozzo 4. This may mean that our linear regression model is doing well for the variables given in predicting the first three depths to groundwater while for the last one, we may not have enough variables to accurately predict.

Predictor	**Coefficient**	**Standard Error**	**P-Value**
Temperature_Siena_Poggio_al_Vento	-0.308	0.049	0.0
Temperature_Mensano	0.113	0.028	0.0
Temperature_Pentolina	0.148	0.071	0.037
Volume_Pozzo_1	0.024	0.002	0.0
Volume_Pozzo_3	0.005	0.002	0.001
Volume_Pozzo_4	0.014	0.001	0.0

Table 14: Predicting Depth to Groundwater Podere Casetta

Predictor	**Coefficient**	**Standard Error**	**P-Value**
Temperature_Siena_Poggio_al_Vento	-0.503	0.11	0.0
Temperature_Pentolina	0.332	0.159	0.038
Volume_Pozzo_1	0.041	0.004	0.0
Volume_Pozzo_4	0.011	0.002	0.0

Table 15: Predicting Depth to Groundwater Pozzo 1

As you can see from the above tables (Tables 14 and 15) for the Luco aquifer, the largest coefficient (in absolute terms) is temperature, with Siena Poggio al Vento being largest for the first few depths. As you can see from the other coefficients, they do not seem to be very predictive, with much smaller relative coefficients.

4.2 Water Springs

For Springs we have a total of three datasets: Amiata, and Lupa and Madonna Di Canneto

4.2.1 Amiata

The Amiata waterbody is composed of a volcanic aquifer not fed by rivers or lakes but fed by meteoric infiltration. This aquifer is accessed through the Ermicciolo, Arbure, Bugnano and Galleria Alta water springs.

Variable	**Non-Null Count**
Rainfall_Castel_del_Piano	6291
Rainfall_Abbadia_S_Salvatore	3586
Rainfall_S_Fiora	2633
Rainfall_Laghetto_Verde	2865
Rainfall_Vetta_Amiata	2212
Depth_to_Groundwater_S_Fiora_8	3569
Depth_to_Groundwater_S_Fiora_11bis	3594
Depth_to_Groundwater_David_Lazzaretti	3242
Temperature_Abbadia_S_Salvatore	3583
Temperature_S_Fiora	7484
Temperature_Laghetto_Verde	3604
Flow_Rate_Bugnano	2008
Flow_Rate_Arbure	2008
Flow_Rate_Ermicciolo	2008
Flow_Rate_Galleria_Alta	2008

Table 16: Amiata Water Spring Dataset

Above, we see that there are a varying amount of missing data. With a total of 7487 observations, all of the variables are missing over half of the observations, i.e missing 3500 or more days during the study. Because we want to study the relationship between the variables and the flow rate, we drop any missing observations in the flow rate variables because a missing data in the flow rate variable is meaningless, as we are unable to study any relationship.

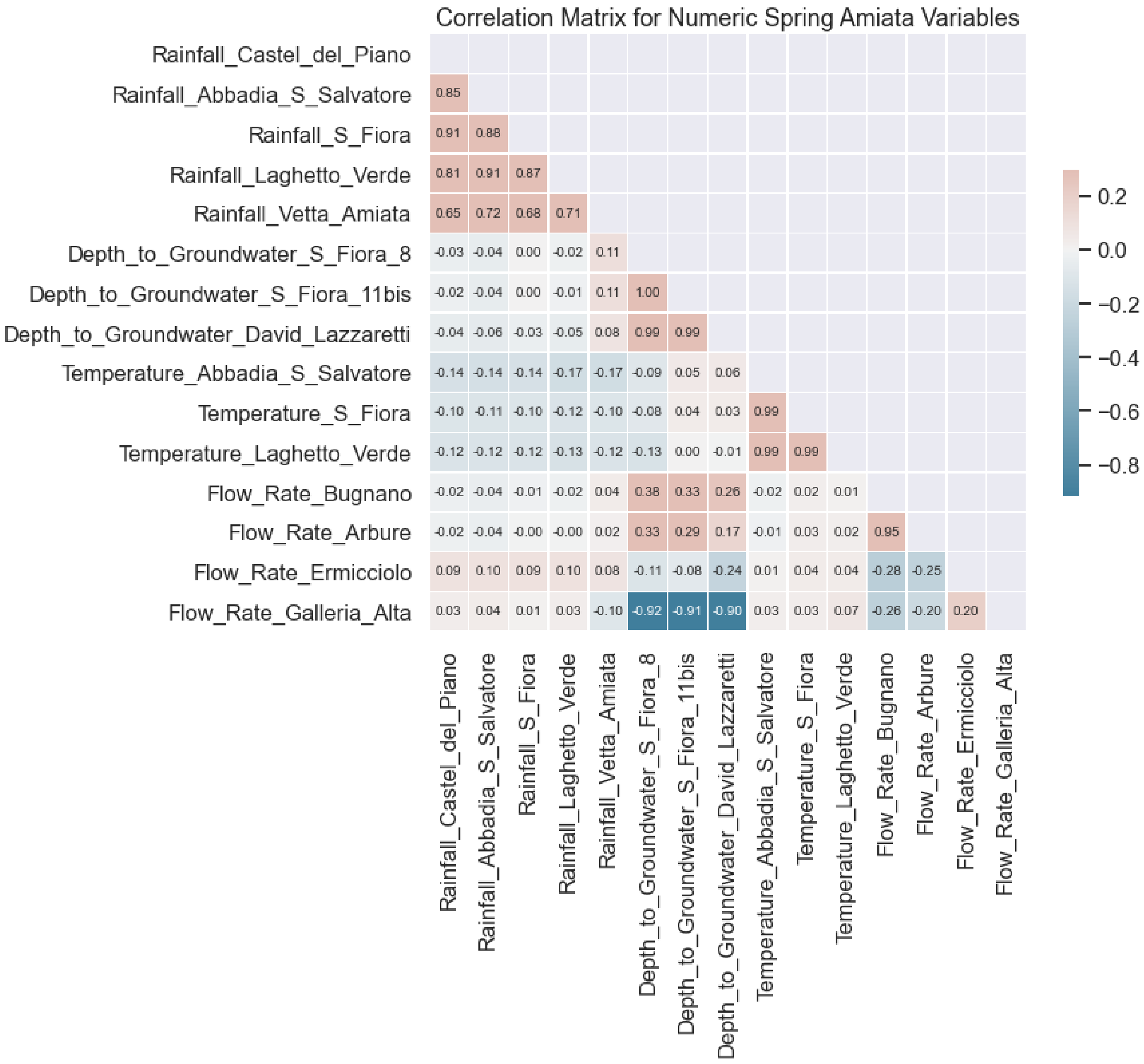

Figure 11: Amiata Spring Correlation Heatmap

In this dataset of analyzing water spring Amiata, we see that the three most correlated variables with the flow rate of Galleria Alta are the three depth to groundwater variables. Additionally, they are also the most correlated with the other three flow rates as well. Furthermore, we notice in the five rainfalls, they are fairly correlated with each other, i.e multicollinearity.

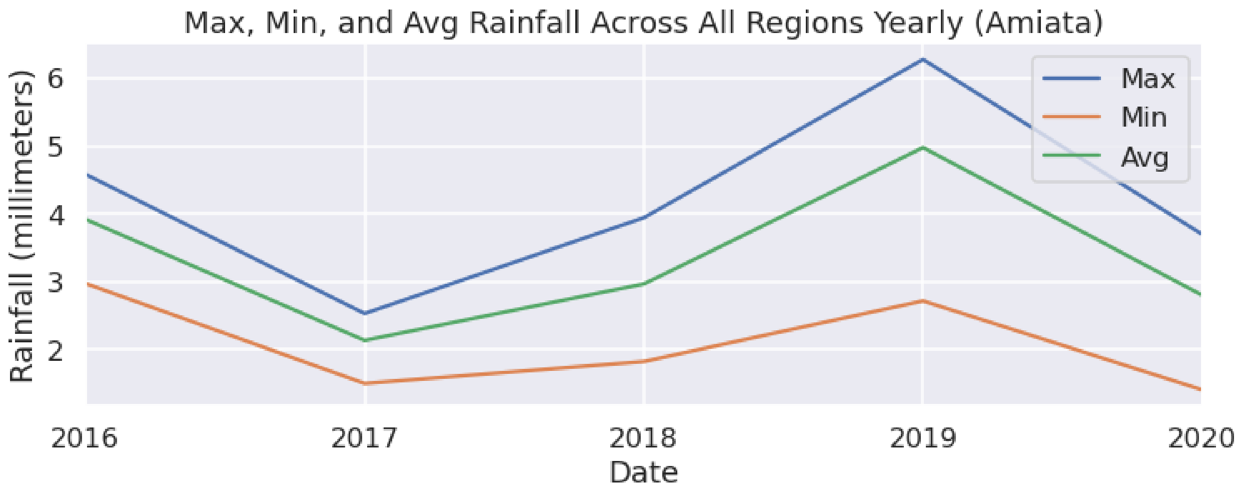

Figure 12: Amiata Spring Rainfall

Above, we see the maximum, minimum, and average yearly rainfall across all five regions. We note that there was a decrease in rainfall in year 2017 and a fairly large increase in the year 2019.

Region	**Winter**	**Spring**	**Summer**	**Autumn**
Rainfall_Castel_del_Piano	[2.541, 5.037]	[0.635, 1.259]	[1.27, 2.519]	[1.906, 3.778]
Rainfall_Abbadia_S_Salvatore	[3.633, 7.172]	[0.908, 1.793]	[1.817, 3.586]	[2.725, 5.379]
Rainfall_S_Fiora	[3.31, 7.065]	[0.827, 1.766]	[1.655, 3.533]	[2.482, 5.299]
Rainfall_Laghetto_Verde	[3.996, 7.972]	[0.999, 1.993]	[1.998, 3.986]	[2.997, 5.979]
Rainfall_Vetta_Amiata	[2.986, 5.0]	[0.747, 1.25]	[1.493, 2.5]	[2.24, 3.75]
Depth_to_Groundwater_S_Fiora_8	[-54.596, -29.134]	[-13.649, -7.284]	[-27.298, -14.567]	[-40.947, -21.851]
Depth_to_Groundwater_S_Fiora_11bis	[-74.04, -39.55]	[-18.51, -9.888]	[-37.02, -19.775]	[-55.53, -29.663]
Depth_to_Groundwater_David_Lazzaretti	[-488.26, -323.373]	[-122.065, -80.843]	[-244.13, -161.687]	[-366.195, -242.53]
Temperature_Abbadia_S_Salvatore	[7.389, 19.884]	[1.847, 4.971]	[3.694, 9.942]	[5.541, 14.913]
Temperature_S_Fiora	[7.728, 18.936]	[1.932, 4.734]	[3.864, 9.468]	[5.796, 14.202]
Temperature_Laghetto_Verde	[7.304, 18.657]	[1.826, 4.664]	[3.652, 9.329]	[5.478, 13.993]

Table 17: Amiata Spring Prediction Intervals

Above we see that the rainfall at Laghetoo Verde is the greatest in millimeters and that the depth to groundwater at David Lazzaretti is the deepest.

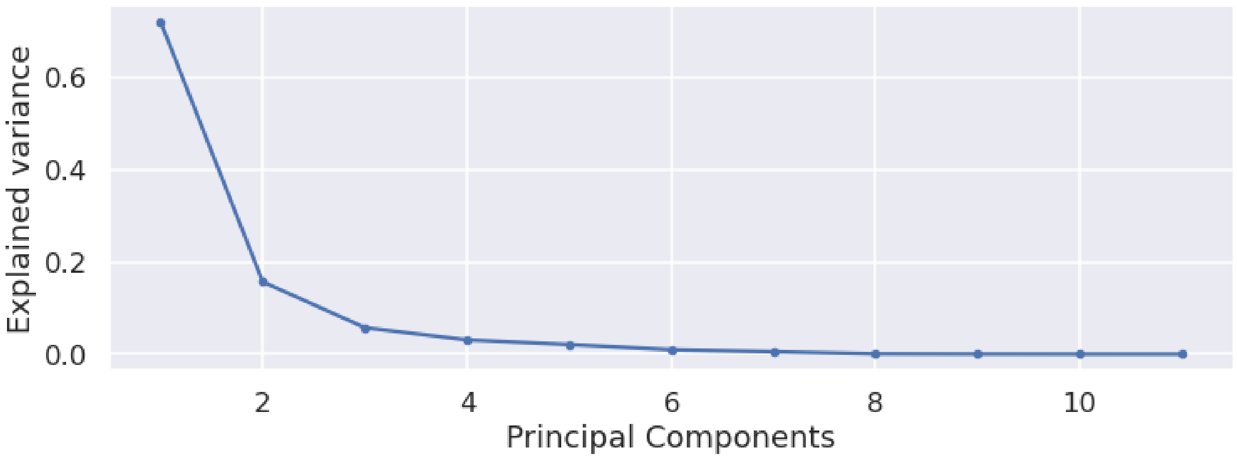

Figure 13: Amiata Spring Principal Component Analysis

To fix the problem of multicollinearity and to reduce the number of variables in our regression model, we utilize principal component analysis. Above we plot the different principal components and the corresponding variances. We see that just by using 5 principal components, we are able to explain most of the variance in our dataset.

```
Linear regression PCA models for Spring Amiata
    RMSE when predicting the Flow Rate of Bugnano: 0.056
    RMSE when predicting the Flow Rate of Arbure: 0.44
    RMSE when predicting the Flow Rate of Ermicciolo: 1.888
    RMSE when predicting the Flow Rate of Galleria Alta: 1.137
```

Above, we see the root mean squared errors when predicting the four flow rates for the water spring Amiata. When using a linear regression to predict the first two flow rates of Bugnano and Arbure, we get a low error, meaning that the linear regression model was pretty accurate in predicting the flow rate. As for the last two, flow rates of Ermicciolo and Galleria Alta, the root mean squared error is fairly high. Referring back to the correlation matrix above, we see that there are variables correlated with these two outcome variables; however, with a root mean squared error this high, it might suggest that a linear model does not fit the data as well.

Predictor	**Coefficient**	**Standard Error**	**P-Value**
Rainfall_Abbadia_S_Salvatore	-0.002	0.001	0.035
Rainfall_Laghetto_Verde	0.002	0.0	0.0
Rainfall_Vetta_Amiata	0.034	0.014	0.017
Depth_to_Groundwater_S_Fiora_8	0.109	0.006	0.0
Depth_to_Groundwater_S_Fiora_11bis	-0.022	0.002	0.0

Table 18: Amiata Spring Predicting Flow Rate in Bugnano

Predictor	**Coefficient**	**Standard Error**	**P-Value**
Rainfall_Laghetto_Verde	0.013	0.004	0.0
Depth_to_Groundwater_S_Fiora_8	0.965	0.045	0.0
Depth_to_Groundwater_S_Fiora_11bis	-0.176	0.017	0.0

Table 19: Amiata Spring Predicting Flow Rate in Arbure

Predictor	**Coefficient**	**Standard Error**	**P-Value**
Rainfall_Vetta_Amiata	5.044	0.299	0.0
Depth_to_Groundwater_S_Fiora_11bis	-0.603	0.047	0.0
Depth_to_Groundwater_David_Lazzaretti	-0.193	0.071	0.007
Temperature_S_Fiora	0.27	0.093	0.004

Table 20: Amiata Spring Predicting Flow Rate in Ermicciolo

Predictor	**Coefficient**	**Standard Error**	**P-Value**
Rainfall_Vetta_Amiata	-3.91	0.171	0.0
Depth_to_Groundwater_S_Fiora_8	3.682	0.069	0.0
Depth_to_Groundwater_S_Fiora_11bis	-0.066	0.027	0.014

Table 21: Amiata Spring Predicting Flow Rate in Alta

Above, we see the coefficients of all four regressions for each flow rate at difference locations (Tables 17-20). Common throughout all of the four is the depth to groundwater at South Fiora 11 bis with the depth to ground water at South Fiora 8 common throughout three of the four. This was not surprising because looking at the correlation matrix, we see that these two depth to groundwater locations are extremely correlated with each four of the flow rates. However, if looking at the correlation between these two depth to groundwater locations, we notice that there is an indication of a perfect correlation. Hence, we are unsure of which of these predictors is more significant because the presence of one affects the presence of the other.

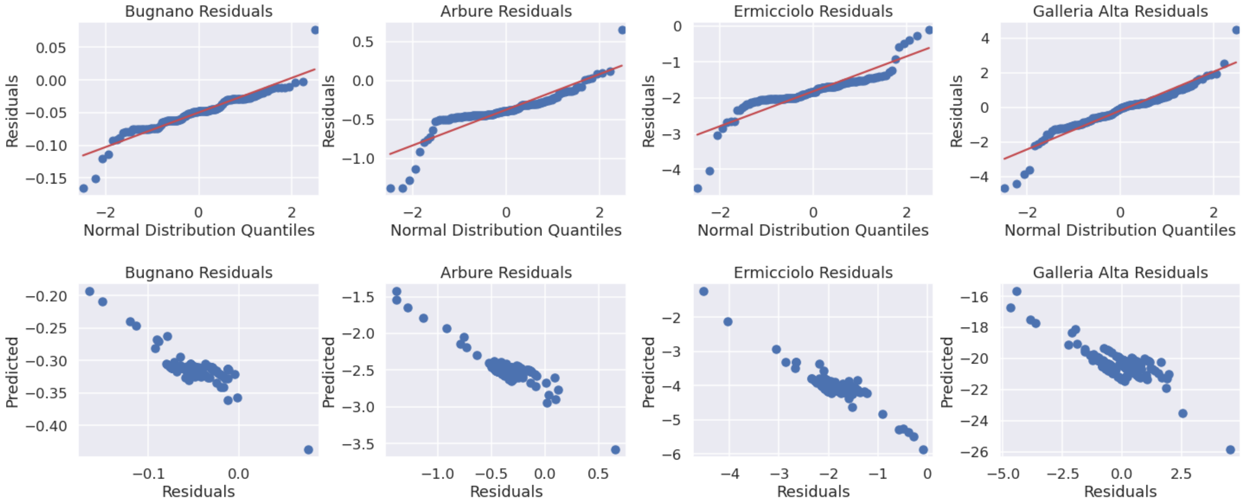

Figure 14: Amiata Spring Residuals

When looking at the residuals, to confirm that a linear regression is correct for the data, we see the first and the last residuals, the regions of Bugnano and Galleria Alta are fairly normal. As for Arbure and Ernicciolo, the distributions of the residuals are somewhat normal with many deviations from the normal distribution. Looking at the scatter plot of the residuals vs the predicted values, we see that there seems to be a strong negative correlation instead of a random scatter. This hints at the decision that our data is not right for a linear regression.

4.2.2 Lupa and Madonna di Canneto

The Lupa water spring is located in the Rosciano Valley, on the left side of the Nera river. It provides drinking water to the city of Terni and the towns around it.

The Madonna di Canneto spring is situated at an altitude of 1010m above sea level in the Canneto valley. It does not consist of an aquifer and its source is supplied by the water catchment area of the river Melfa.

Variable	**Non-Null Count**
Rainfall_Terni	4199
Flow_Rate_Lupa	3817

Table 22: Lupa Spring Dataset

Variable	Non-Null Count
Rainfall_Settefrati	2557
Temperature_Settefrati	2557
Flow_Rate_Madonna_di_Canneto	1387

Table 23: Madonna Spring Dataset

Above we see the information about null values in both the Lupa and Madonna di Canneto datasets. In both datasets, because we want to predict the flow rate from the other variables given, any null values that the flow rate contains is dropped. We had decided on this decision because with an unknown outcome value, we are unable to predict or make assumptions on. While this may decrease the accuracy of our model and increase the error, we did not want to impute the value, possibly feeding the model incorrect information.

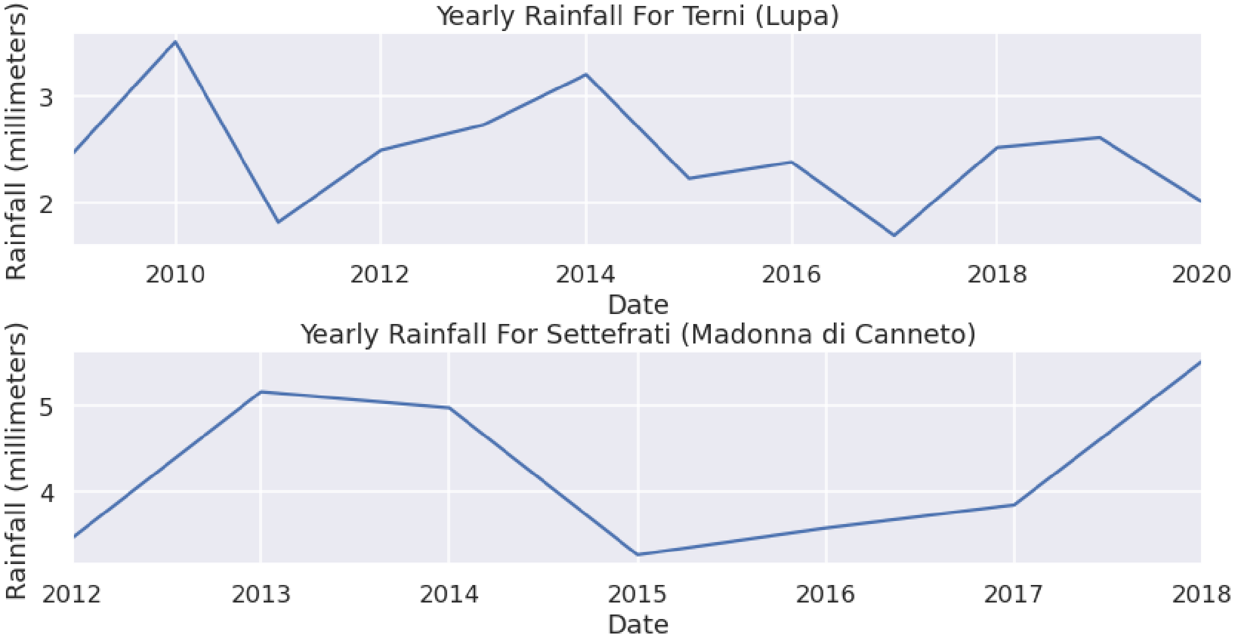

Figure 15: Lupa and Madonna di Canneto Yearly Rainfall

Above, we plot the yearly rainfall for the rainfall regions of the water spring Lupa (Top) and Madonna di Canneto (bottom). In both of these graphs, unlike the other datasets, we only plot a single region because there are very limited numbers of variables in both of these datasets, hence, we are unable to take the maximum, minimum, or average for the entire year. In the region of Terni for the water spring Lupa, we see an overall decrease in the rainfall from the year 2010 and 2020. As for the water spring Madonna di Canneto, we see almost a steady rainfall from year 2012 to 2018 with an all time low yearly rainfall in year 2015.

Season	Rainfall_Terni
Winter	[2.857, 3.942]
Spring	[0.714, 0.985]
Summer	[1.429, 1.971]
Autumn	[2.143, 2.956]

Table 24: Prediction Intervals for Lupa

Season	Rainfall_Settefrati	Temperature_Settefrati
Winter	[2.761, 6.741]	[7.861, 23.803]
Spring	[0.69, 1.685]	[1.965, 5.951]
Summer	[1.381, 3.37]	[3.93, 11.902]
Autumn	[2.071, 5.056]	[5.895, 17.853]

Table 25: Prediction Intervals for Madonna di Canneto

In the two tables above, we see the prediction intervals in the predictor variables of the water spring Lupa (Table 21) and Madonna di Canneto (Table 22). It is no surprise that in both water springs, the Winter season is the one with the most rainfall. What is surprising is that the summer months, there is approximately twice as much rainfall that in the Spring. We hypothesize that the autumn months in Italy may begin earlier than in the USA, i.e the ending months of Summer, leading into Autumn, experience high rainfalls.

```
RMSE when predicting the Flow Rate of Lupa: 15.792
RMSE when predicting the Flow Rate of Madonna di Canneto: 18.652
```

Above, we see the root mean squared error of predicting both flow rates is extremely high. While this might suggest that a linear regression is not suitable for these two datasets, we are unsure of whether this is true or not. When looking at the data for both Lupa and Madonna di Canneto, we see that we needed to drop about a few hundred observations for Lupa and around half for Madonna di Canneto. Therefore, this bias in our data may have led calculation of the error to be high, when in reality if the model was trained with more complete data, a linear regression may be suitable.

TODO: Coefficients

Predictor	Coefficient	Standard Error	P-Value
Rainfall_Settefrati	15.77	0.267	0.0

Table 26: Flow Rate Prediction of Mandonna Di Canneto

When predicting the flow rate for Lupa, we did not find any significant predictors. This is because the dataset for the Lupa spring was fairly small. For the Mandonna Di Canneto, we see that the rainfall for the region of Settefrati was pretty significant with a coefficient of 15.77; however, this may also be due to a small dataset, as the model has not seen much variation.

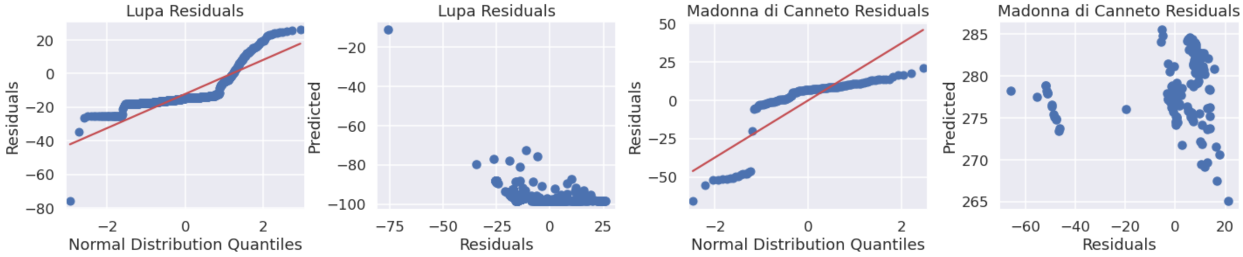

Figure 16: Lupa and Madonna di Canneto Residual Plots

Above, we see the residuals plotted for the remaining two springs. It is no surprise that the residuals are neither normal nor random. The extremely large root mean square error informs us that a linear regression is not suitable for our data. Hence, looking at the residuals further confirms this fact.

4.3 Lakes

The Bilancino lake is an artificial lake located in the municipality of Barberino di Mugello. It is used to refill the Arno river during the summer months. Indeed, 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.

Variable	**Non-Null Count**
Rainfall_S_Piero	6603
Rainfall_Mangona	6026
Rainfall_S_Agata	6026
Rainfall_Cavallina	6026
Rainfall_Le_Croci	6026
Temperature_Le_Croci	6025
Lake_Level	6603
Flow_Rate	6582

Table 27: Bilancino Lake Dataset

Before beginning our analysis, we want to identify any missing data in our observations. The above list highlights the number of null data in each variable present. We see that there are a total of 6603 entries for the lake data. however, 6 of these variables have the exact same number of missing data, i.e the study had not begun but the data for two other variables was available. Hence, we drop the 578 observations with missing data in these 6 columns.

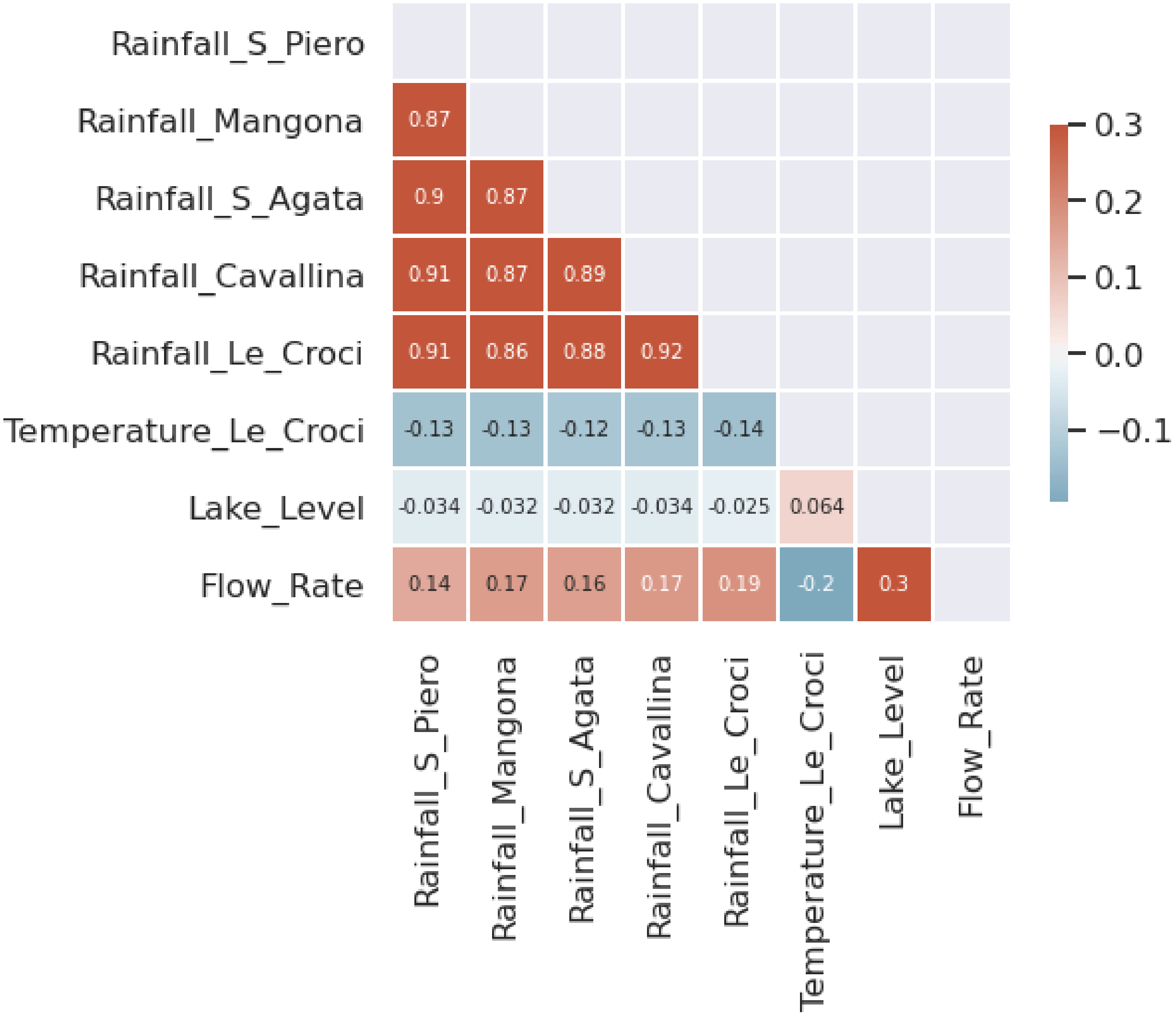

Figure 17: Lake Bilancino Correlation Matrix

From the correlation matrix, we see that the temperature is the most correlated with the flow rate and the lake level. We found this to be surprising because a positive correlation between the temperature and the lake level meant that as the temperature rose, so did the lake level. We expected at the very least the correlation be a negative value because as the temperature rises, the more water evaporates, and hence a lower lake level.

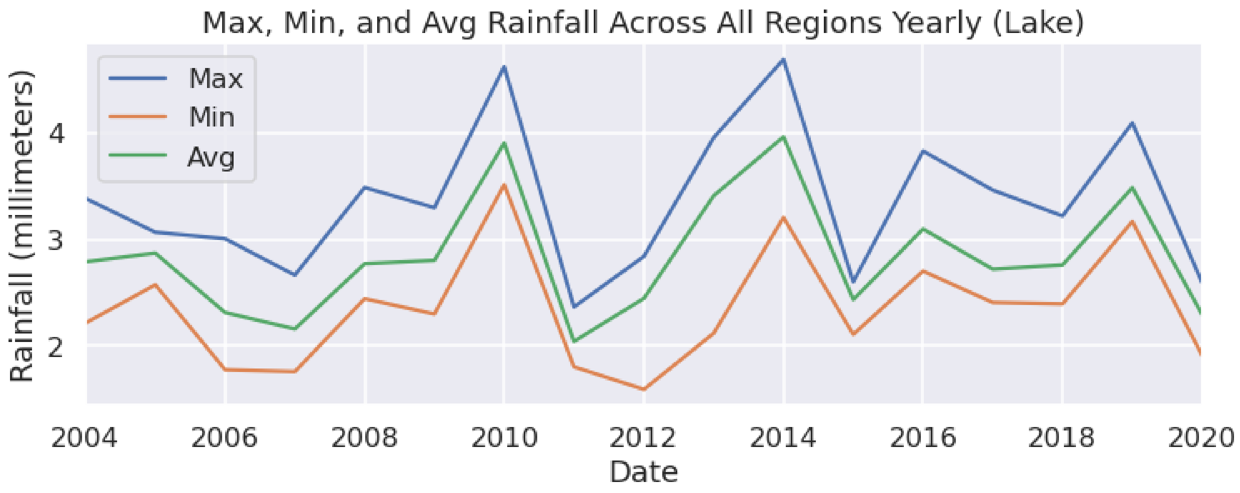

Figure 18: Lake Bilancino Yearly Rainfall Across all Regions

Above, we see in the winters of year 2010 and 2014, there was a dramatic increase in the rainfall. Additionally, we notice that the rainfall of year 2012 was far lower than other years in the study.

Season	**Rainfall_S_Piero**	**Rainfall_Mangona**	**Rainfall_S_Agata**	**Rainfall_Cavallina**	**Rainfall_Le_Croci**	**Temperature_Le_Croci**
Winter	[3.197, 4.095]	[4.363, 5.575]	[3.381, 4.366]	[3.48, 4.498]	[4.009, 5.137]	[13.783, 21.051]
Spring	[0.799, 1.024]	[1.091, 1.394]	[0.845, 1.092]	[0.87, 1.125]	[1.002, 1.284]	[3.446, 5.263]
Summer	[1.598, 2.047]	[2.181, 2.787]	[1.691, 2.183]	[1.74, 2.249]	[2.004, 2.569]	[6.892, 10.526]
Autumn	[2.398, 3.071]	[3.272, 4.181]	[2.536, 3.275]	[2.61, 3.374]	[3.007, 3.853]	[10.337, 15.789]

Table 28: Lake Bilancino Prediction Intervals

Because we want to identify what is the best way to extract water, we calculated the confidence intervals of each region in terms of rainfall or temperature for each season. For instance, we are 95% confidence that the South Piero region would have between 3.197 and 4.095 millimeters of rainfall during the Winter Season.

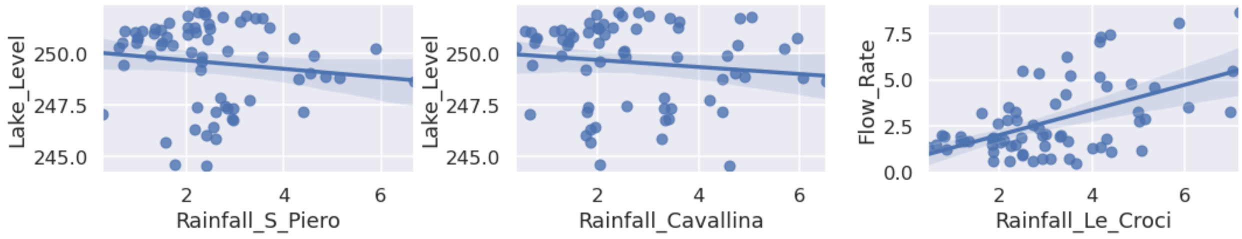

Figure 19: Lake Bilancino Regression Plot for Three Predictors

Above, we see the regression plots for some predictors and outcome variables separately. We see that the first two plots do not show a linear relationship while the one on the very right shows some linearly relationship. Other variables follow the same behaviors, with some showing no linear relationship at all and some hinting at the existence of this relationship.

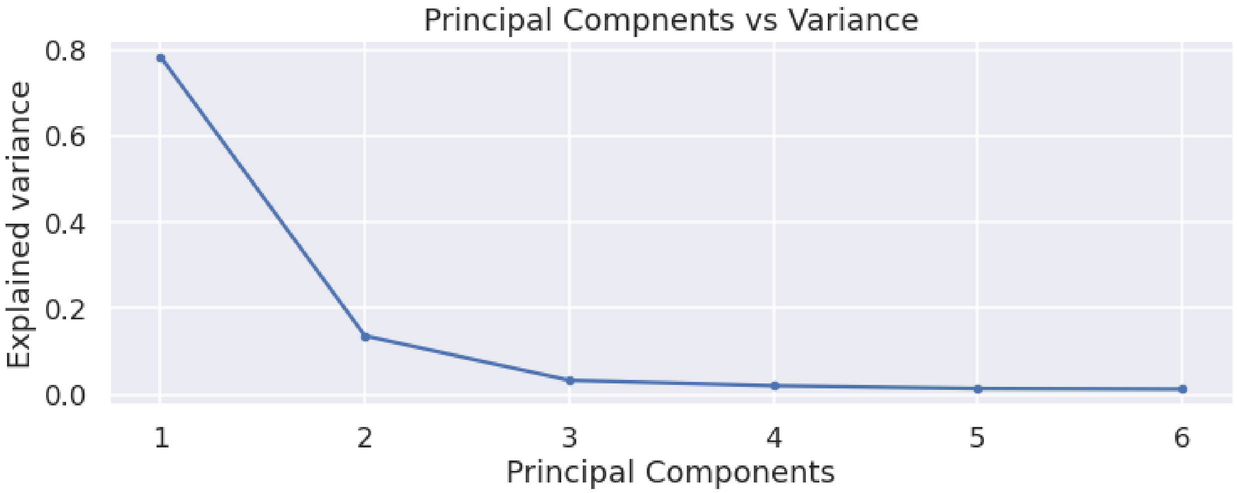

Figure 20: Lake Bilancino Principal Component Analysis Variance Plot

Because our dataset has multiple predictor variables, carrying out principal component analysis (PCA) can aid in decreasing the dimension space with a trade off of variance. In the plot above of the variance explained by each component, we see that most of the variance in the data can be explained by 3 principal components.

```
RMSE when predicting the Lake Level: 1.805
RMSE when predicting the Flow Rate: 3.963
```

Typically, when looking at the root mean squared error for predictions, we aim for a value that is less than 0.5. Above, we see that we resulted in an error value of 1.805 and 3.963 when predicting the lake level and flow rate, respectively. When referring back to the correlation matrix above, we see that all the five rainfalls and the temperature variable with the lake level do not have high correlations, less than 0.1. Hence, we believe that we are unable to use linear regression to predict the lake level.

Predictor	**Coefficient**	**Standard Error**	**P-Value**
Rainfall_S_Piero	1.47	0.344	0.0
Rainfall_S_Agata	-1.714	0.526	0.001
Rainfall_Cavallina	2.918	0.49	0.0
Rainfall_Le_Croci	13.414	0.086	0.0

Table 29: Lake Bilancino Predicting Lake Level Coefficients

Predictor	Coefficient	Standard Error	P-Value
Rainfall_S_Piero	0.048	0.014	0.001
Rainfall_Cavallina	0.173	0.021	0.0
Rainfall_Le_Croci	0.113	0.004	0.0

Table 30: Lake Bilancino Predicting Flow Rate Coefficients

Looking at the lake level (Table 16) first, we see that the rainfall of Le Croci is an important predictor when predicting the rainfall. This was interesting because when just looking at the correlation matrices, there all variables had an extremely low correlation with the lake level. Moving on to the flow rate (Table 17), we see that the rainfall at Cavallina and Le Croci are important predictors in our model.

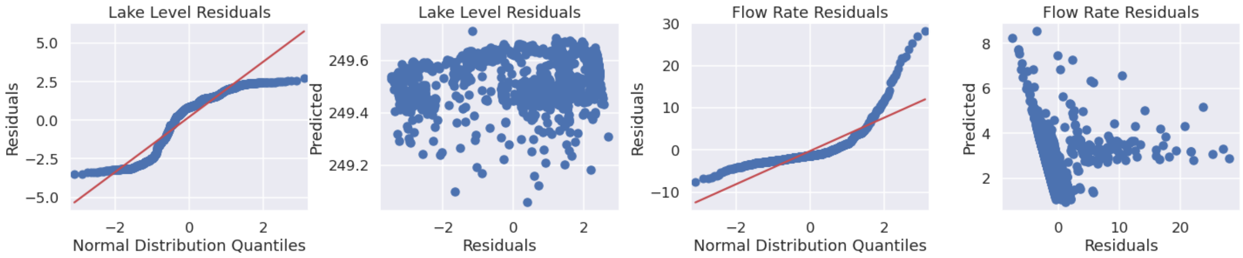

Figure 21: Lake Bilancino Residual Plot

When looking at the lake level residuals, the left two, we see that the residuals are neither normally distributed nor randomly scattered. In the scatter plot of lake level residuals, we see that the residuals are extremely high. Hence, our independent variables are not good predictors of the lake level. This is not surprising because when looking at the correlation matrix above, the correlation between all variables and the lake level was extremely low. Looking at the right two residual plots, we once again see that the flow rate residuals are also neither normally distributed nor random. Hence, once again, a linear regression may be incorrect.

4.4 Rivers

Arno is the second largest river in peninsular Italy and the main waterway in Tuscany. It has a relatively torrential regime, due to the nature of the surrounding soils (marl and impermeable clays). Arno results to be the main source of water supply of the metropolitan area of Florence-Prato-Pistoia. The availability of water for this waterbody is evaluated by checking the hydrometric level of the river at the section of Nave di Rosano.

Variable	**Non-Null Count**
Rainfall_Le_Croci	8217
Rainfall_Cavallina	6026
Rainfall_S_Agata	6026
Rainfall_Mangona	6026
Rainfall_S_Piero	6026
Rainfall_Vernio	4283
Rainfall_Stia	1283
Rainfall_Consuma	1283
Rainfall_Incisa	4568
Rainfall_Montevarchi	1647
Rainfall_S_Savino	1283
Rainfall_Laterina	1283
Rainfall_Bibbiena	2378
Rainfall_Camaldoli	1283
Temperature_Firenze	6192
Hydrometry_Nave_di_Rosano	8169

Table 31: River Arno Dataset

Analyzing the missing data in the river dataset, we see that the first 5 rainfalls are the same as that of the lakes dataset. However, the following 10 variables are missing at least 40% of the data. Because the observations were collected from year 2004 to 2020, we arbitrarily select a threshold of year 2010, that is if a particular variable does not have data for at least the year 2010, we will drop that particular variable.

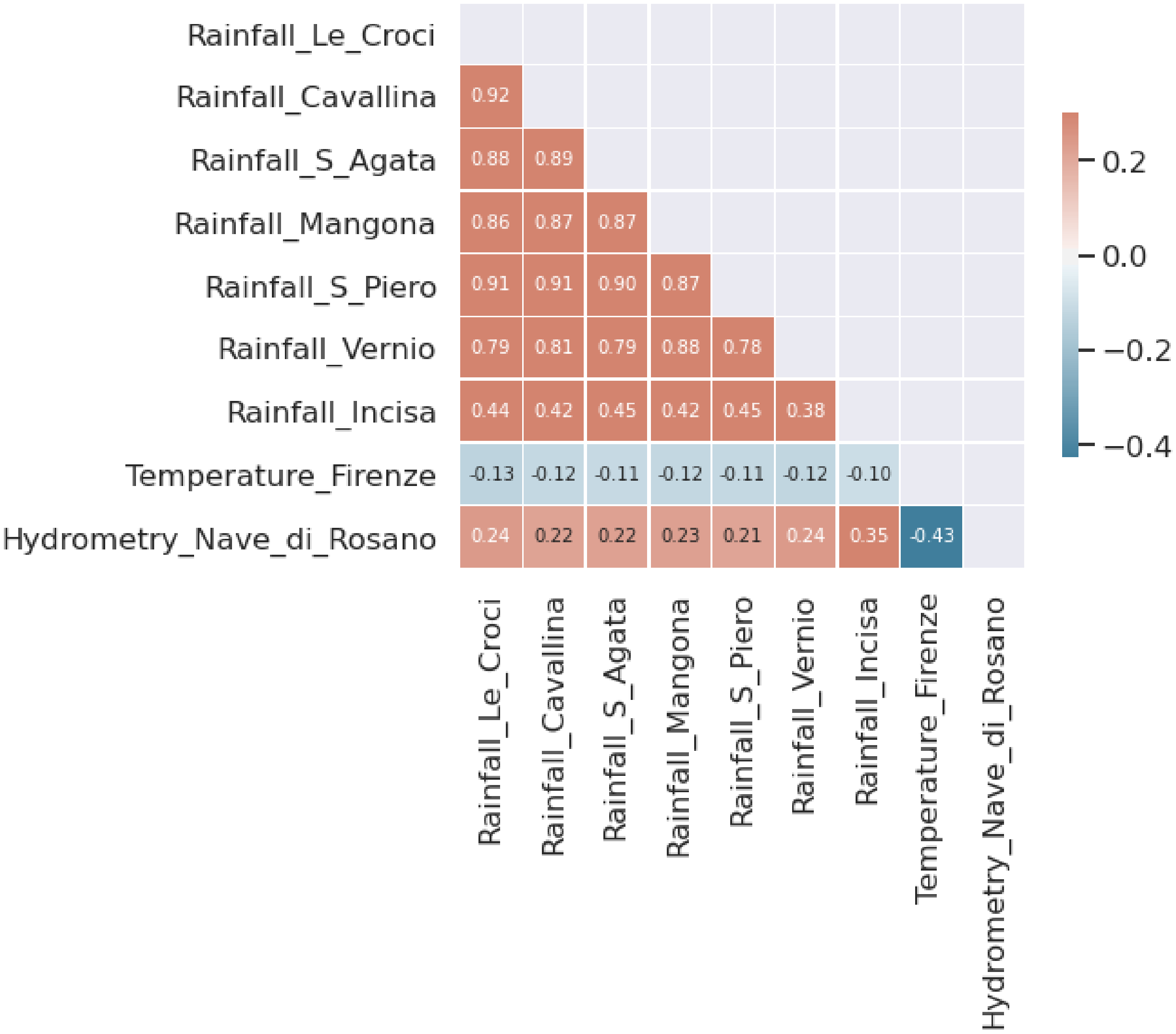

Figure 22: River Arno Correlation Matrix

Once again, as seen in the lake data set, temperature is the most correlated with hydrometry. However, this time it has a negative correlation, i.e as the temperature increases, the river level decreases. Following the temperature is the rainfall at Incisa and at Vernio and Le Croci.

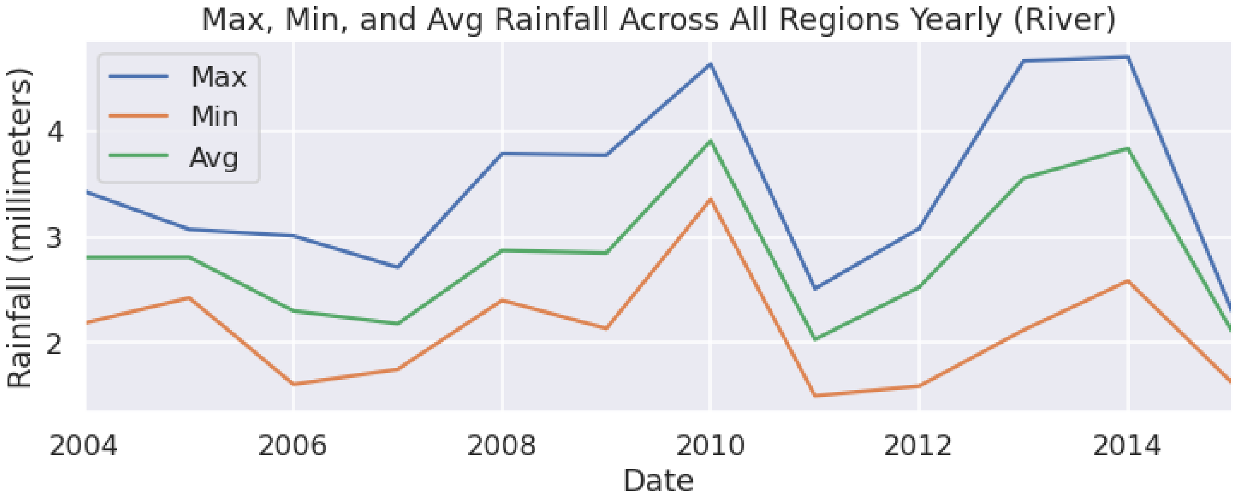

Figure 23: River Arno Yearly Rainfall Across all Regions

In the above plot, we see the maximum, minimum, and average rainfalls for all 7 regions between the years 2004 and 2014. We see in the years 2010, 2013, and 2014, there was an increase in rainfall and 2011 had the lowest rainfall throughout all years.

Region	Winter	Spring	Summer	Autumn
Rainfall_Le_Croci	[4.009, 5.137]	[1.002, 1.284]	[2.004, 2.569]	[3.007, 3.853]
Rainfall_Cavallina	[3.48, 4.498]	[0.87, 1.125]	[1.74, 2.249]	[2.61, 3.374]
Rainfall_S_Agata	[3.381, 4.366]	[0.845, 1.092]	[1.691, 2.183]	[2.536, 3.275]
Rainfall_Mangona	[4.363, 5.575]	[1.091, 1.394]	[2.181, 2.787]	[3.272, 4.181]
Rainfall_S_Piero	[3.197, 4.095]	[0.799, 1.024]	[1.598, 2.047]	[2.398, 3.071]
Rainfall_Vernio	[4.304, 5.816]	[1.076, 1.454]	[2.152, 2.908]	[3.228, 4.362]
Rainfall_Incisa	[2.871, 3.847]	[0.718, 0.962]	[1.435, 1.923]	[2.153, 2.885]
Temperature_Firenze	[15.306, 24.42]	[3.826, 6.105]	[7.653, 12.21]	[11.479, 18.315]

Table 32: River Arno Prediction Intervals

Above, we see the confidence intervals for each rainfall region. Throughout all 8 regions, region Vernio has the highest estimated rainfall with as much as 5.816 millimeters of rainfall in the winter and 2.908 in the Summer.

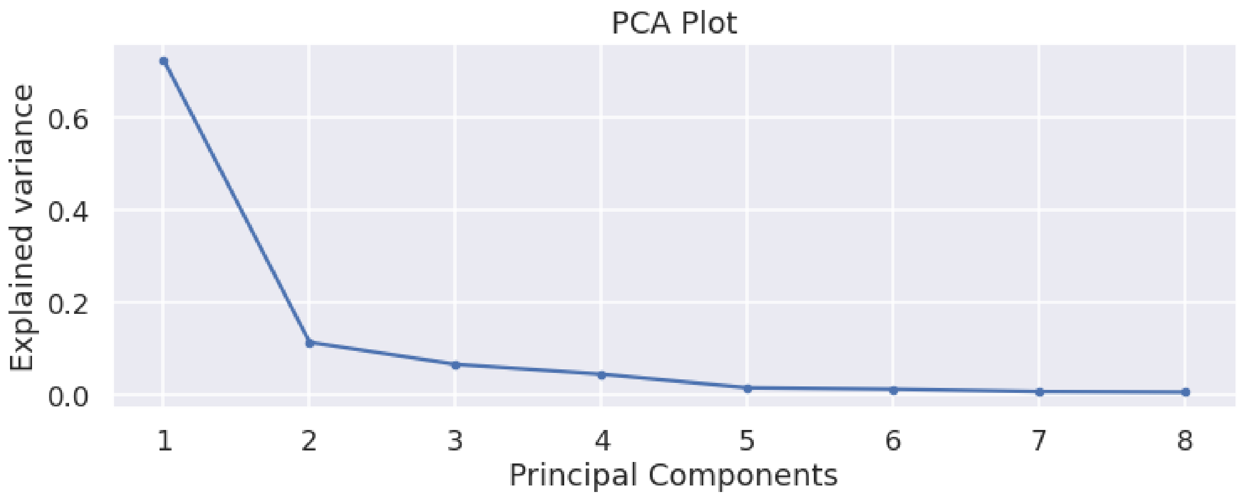

Figure 24: River Arno Principal Component Analysis Variance Plot

Above, we see for the river dataset, that using 5 principal components can explain most of the data's variance.

```
RMSE when predicting the Hydrometry: 0.785
```

The root mean square error of 0.785. While this is closer to 0.5, i.e using a linear regression fits this model better than lake, it is still considered high. We believe this may be due to the fact that the relationship between the rainfall and temperature and the hydrometry is not linear.

Predictor	**Coefficient**	**Standard Error**	**P-Value**
Rainfall_Le_Croci	-0.017	0.006	0.008
Rainfall_Cavallina	0.013	0.006	0.033
Rainfall_Mangona	-0.016	0.007	0.018
Rainfall_S_Piero	0.008	0.003	0.016
Rainfall_Vernio	0.045	0.003	0.0
Rainfall_Incisa	0.057	0.001	0.0

Table 33: River Arno Predicting Hydrometry Coefficients

We see that when predicting hydrometry, all of the variables coefficients are extremely close to 0. The highest of the 6 significant predictors is the rainfall at Incisa. Referring back to the correlation matrix, this is inline with our hypothesis because the Pearson correlation between Incisa rainfall and hydrometry is fairly high.

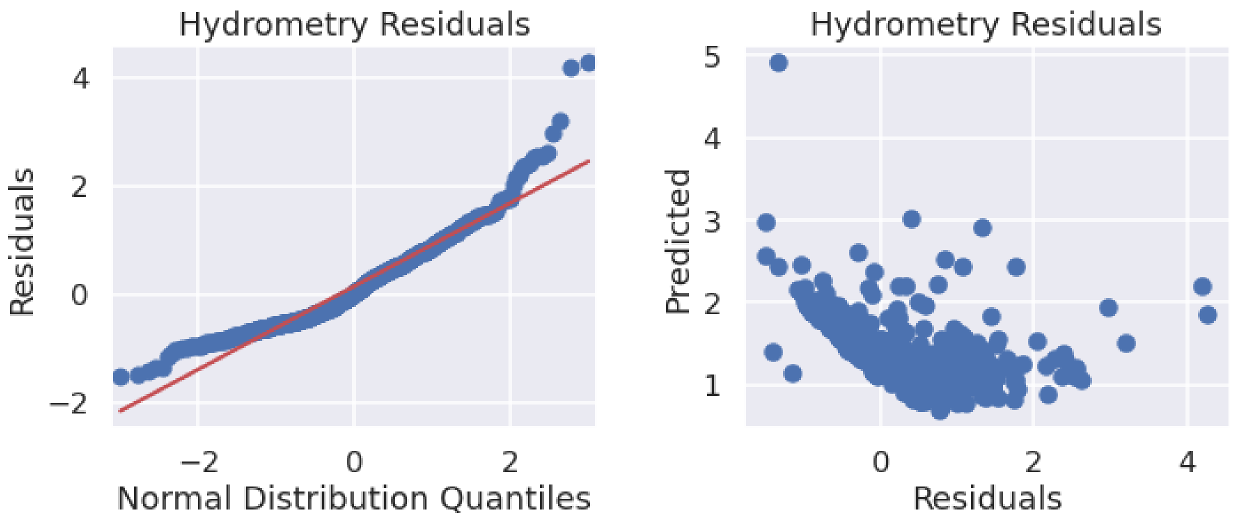

Figure 25: River Arno Residual Plot

Plotting the residuals, we see that the distribution of the residuals is fairly normal with the right end of the distribution becoming more sparse. On the right, we see the scatter plot of the predicted vs the residuals. This is fairly surprising because our root square mean error was close to the threshold at around 0.7, so we had expected the scatter to be fairly random centered around 0. Nevertheless, we see that the points are near zero; however, because the scatter plot is not random, we are unable to advocate for the use of a linear regression.

5 Limitations

In our analysis, there were some limitations. For all of the datasets, there were at minimum, a few hundred missing observations. In some cases, many variables were missing more than half of the years the study was performed. In these cases, we had decided to drop the variables entirely. When there were just a few missing observations within a variable, we had dropped that observation entirely because we had decided that imputing the missing data may lead to biased results. This would not represent an accurate interpretation of the data set.

Another factor that may have decreased our model's power is multicollinearity. In many of our datasets, there were multiple variables with strong correlation, and occasionally, perfect correlation among each other. Because multicollinearity affects the way we interpret our model, we had attempted to utilize principal component analysis to reduce the number of components used in our linear regression. In most cases, the particular dataset's variance was able to be explained in using half the number of the principal components.

One of the most important limitations in our analysis is the linear regression model. Our dataset is a time-series data, meaning that the observations within the data are not independent. While linear regression is usually not the best choice when predicting time-series data, a linear regression model may be suitable if the residuals of the linear model are normally distributed and do not show a pattern. When we trained and tested our model, a few of our residuals were normally distributed while the others were not and the spread of the residuals were often in groups. Because of this, we believe that a linear regression model is not suitable for this data and suggest exploring other models, such as vector autoregression (VAR) moving average which can take into account the time portion of our data. To effectively use VAR the data is assumed to be stationary, meaning that the shape of the distribution doesn't change over time. If we were to explore a VAR model, we would need to first make sure our data is stationary.

6 Conclusion

For the water types of aquifers, we had a total of four datasets. Each dataset had various types of predictors with many outcomes of depth to groundwater at different locations. The first aquifer, the Auser, had quite an amount of multicollinearity among a few of its independent variables. To fix this we had implemented PCA into our linear regression. The rainfall that we had found to be a constant predictor throughout all five depth to groundwaters was the rainfall at Gallicano. This was somewhat surprising because when looking at the correlation matrix, we saw that the Gallicano rainfall was not very correlated with the many of the depth to groundwater variables. The prediction intervals of Gallicano show that Winter is the best season for water extraction as there is an abundance of rainfall, at least double or triple the other seasons.

The second aquifer we analyzed was the Petrignano. Originally, we had hypothesized that the variable volume c1 at Petrignano would be the most significant predictor because it had the highest correlation. However, we were wrong because the rainfall at Bastia Umbra had the highest coefficient while the volume was the second least. When we look at the prediction intervals for the rainfall at Bastia Umbra, we notice that the rainfalls for all the seasons are fairly low. around 2 millimeters and less. Because of the low rainfall, we suggest that this aquifer may not be the best location to extract water from, as it is pretty easy to deplete the water resource.

The third aquifer was the Doganella. First visualizing the correlations among the given variables, we see that volume 1 at Pozzo has the highest correlation with all of the depth to groundwater variables. Hence, we hypothesized that this volume would be the most significant predictor. Following our regression model, we see that we were, in fact, correct with volume 2 at Pozzo not far behind. As for the best season for water extraction, because the two significant predictors are volume 1 and volume 2 at Pozzo, we suggest either Winter or Autumn for optimal water extraction.

The last aquifer was Luco. Upon first look, we see that once again, the volume 1 and volume 3 at Pozzo had the highest correlations with the depth to groundwater. For these two variables, we had checked for multicollinearity because they are roughly in the same region. However, we did not find high correlations between these two. Proceeding to the model coefficients the temperature of Siena Poggio al Vento was the most significant. The correlation between the temperature of Siena Poggio al Vento and the hydrometry levels were negative, i.e as the temperature goes down, the depth to groundwater increases. Hence, we recommend the Summer season for extraction because as the temperature increases, the depth to groundwater decreases and less digging is needed to reach the waters.

For the water springs, we had a total of three datasets. Each dataset analyzes a different water spring. The first water spring, Amiata, had the most amount of variables available. When just looking at the correlation matrix, we see that the variables most correlated with the flow rate, the variable to predict, are the depth to groundwater variables. The linear regression coefficients also confirms this fact, that the depth to groundwater variables are the most important. Looking at the prediction intervals, we suggest water extract during the Spring season because during Spring, the depth to groundwater is the least. The second spring, Lupa, we were given a single predictor variable, a single rainfall at a particular region. However, after performing linear regression, we saw that there were no significant predictors. Nevertheless, this may be due to the fact that there was too little data. If there was a strong correlation, the best time for water extraction would be in the Winter seasons and not in the Spring seasons as the Spring seasons have the lowest estimated rainfall. The third spring, Madonna di Canneto, was similar to Lupa in the sense that there were very few predictor variables. However, with our model, we were able to find that the singular rainfall at Settefrati was significant with a high coefficient. Looking at the confidence intervals, we once again suggest for Winter water extraction and avoid water extraction during the Spring months to allow for a restoration of water levels.

For the lakes, there was only one lake to be analyzed, but predicting two variables, the lake level and the flow rate. Looking at the lake level results first, we see that the rainfall at Le Croci is the most important when predicting the lake level and with the rainfall at Cavallina second. When looking at the flow rate, we see the opposite, Cavallina is first and Le Croci is second. When looking at the prediction intervals for these two rainfall regions, we see that the optimal time for water extraction is in the Winter or Autumn months as these two seasons have an estimated rainfall twice as much as the Spring and Summer months.

For the rivers, once again there was only one river to be analyzed with a single outcome variable, the hydrometry or river level. Originally, we had predicted that the temperature at Firenze would be the most significant predictor in our regression model because it had the highest Pearson correlation. When we look at the regression coefficients, we see that almost all of the predictor variables had a low coefficient, i.e did not have much of an effect on predicting the hydrometry. Additionally, Firenze's temperature was not one of the predictors that was significant in our model, i.e it had a p-value larger than the alpha value of 0.5. Nevertheless, the predictor variable that did have the highest coefficient was rainfall at Incisa. Referring back to the prediction intervals, we see that the seasons of Winter, Summer, or Autumn would be the best for water extraction and to avoid Spring as the average rainfall in the Spring was less than 1 millimeter.

From the nine datasets above, we do not have a definite answer of when is the best time to extract water without depleting the source, as seen above, it depends on the water body type. Answering the question of whether a linear regression is suitable for our data set, when we look at the residuals of our linear regression models, we see that most of the residuals were not normally distributed and not a random scatter. This leads us to be hesitant in our regression model, as more errors were introduced due to this violation of assumption. Further research had led us to the vector autoregression model which may be of better use in this analysis. Nevertheless, when looking at the correlation matrix and the prediction intervals, we are still able to get an idea of what factors are the most important in the forecasting depth to groundwater, flow rate, lake level, and hydrometry.

7 Theory

7.1 Pearson Correlation

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

where:

- x_i are the values of the independent variable
- $\bar{x}$ is the mean value of the independent variable
- y_i are the values of the dependent variable
- $\bar{y}$ is the mean value of the dependent variable

The Pearson Correlation Coefficient is a measure of the strength in a linear relationship between two variables. The correlation attempts to draw a line of best fit through the two variables and the coefficient indicates how far away the data points are from this line of best fit. The coefficient can take values between -1 and 1, inclusive with -1 and 1 having a strong linear relationship and 0 having no linear relationship.

7.2 Confidence Intervals/Prediction Intervals

In this analysis, we have implemented a prediction interval in all of our datasets. A prediction interval is a type of confidence interval that allows us to estimate a new observation in regression analysis based on an existing model. Confidence intervals usually estimate the population mean or proportion, but when it comes to estimating for an individual case, confidence intervals would not be appropriate because we need to take account of the uncertainty of sample deviation of each observation. The formulation for prediction intervals is similar to that of confidence intervals:

$$\hat{y}_h \pm t_{(1-\alpha/2, n-2)} \times \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$$

where $\hat{y}_h$ is the predicted-value.

Note in the critical value, $t_{(1-\alpha/2,n-2)}$, the degree of freedom is $n-2$ because we use MSE. Then the standard error for the prediction is:

$$\sqrt{MSE\left(1+\frac{1}{n}+\frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

Prediction intervals are typically wider than that of confidence intervals because they take individual observations as well as parameter estimates into consideration. In our analysis we had grouped our data by seasons, four for each year. We, then, fed the grouped data into an ordinary least squares regression and attempted to predict the average rainfall for that specific season.

7.3 Principal Component Analysis (PCA)

In many of our datasets, we have numerous variables and with it, multicollinearity. Multicollinearity is when many independent variables are highly correlated with one another. For instance, in the lake dataset, all of the rainfalls have a high Pearson Correlation of around 0.8-0.9. This may be because the regions of recorded rainfall may be near each other, and hence when a single region experiences, rain, other nearby regions follow as well. To remedy this problem of multicollinearity, we need to drop some of the independent variables, i.e predictor variables before performing a regression on our data otherwise, our regression results would be biased as the precision of the estimated coefficients lowers and thus the power, i.e ability to correctly predict, is lowered as well.

Principal Component Analysis (PCA) can aid in the dropping of variables in our regression model because PCA captures the amount of variance explained in the data by each principal component. Hence, we may be able to explain most of the variance in our data by a few of the independent variables.

To utilize PCA, we first create a covariance matrix of the data, say X and the covariance matrix is denoted by

$$C = X^T X$$

where ^{T} is the transpose of a matrix.

Then because $C = X^T X$ is symmetric, i.e $C^T = C$, it will have an eigenvalue decomposition of

$$C = Q\Lambda Q^T$$

where Q is orthogonal and Λ is diagonal.

$$Q = \begin{bmatrix} v_1 & v_2 & \cdots & v_r \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_r \end{bmatrix}$$

The entries of Λ are the eigenvalues of C while the columns of Q are the eigenvectors corresponding to the eigenvalues in Λ.Because the number of nonzero eigenvalues of C tells us how many independent columns X we have, the rank of C informs us of how many independent variables we have. The eigenvalue decomposition of C, then tells us how many eigenvectors, or principal components, correspond to nonzero eigenvalues. The eigenvector corresponding to the largest eigenvalue is the first principal component and the corresponding to the second largest eigenvalue is the second principal component and etc.

7.4 Regression Plots

In our study, we analyzed whether we would be able to predict a particular outcome variable(s) based on many predictor variables using a linear regression. In a linear regression, one of the obvious assumptions is that the relationship between the predictor and outcome (independent and dependent) variables is linear. A regression is beneficial in this case because it allows us to visually see whether a particular predictor has a linear relationship with a particular outcome variable.

7.5 Linear Regression

Linear Regression is used to predict the relationship between two variables. Typically this uses an independent variable to predict the dependent variable. This model contains an error term which is used to compute the variability of the dependent variable. Linear regression is used for the correlation between the two variables which is not to be confused with causation. The point of it is to see whether the independent and dependent variables have a positive relationship, negative relationship, or no relationship at all. For our analysis, we use multiple predictors to estimate one dependent variable, this process is so called multiple linear regression. To obtain a good model, there are five assumption we must follow:

1. **Linear relationship:** Linear regression requires all predictors to have a linear relationship to the dependent variables. This can be checked by scatter matrices. When looking at the scatter matrices, it is also crucial to check for outliers as linear regression is sensitive to outliers.

2. **Multivariate normality:** Linear regression requires the independent variables to be roughly normal, this can be checked by qq- plot for normality and from a goodness-of- fit test. if the variables are not normal, this can be resolved by transformations.

3. **No or little multicollinearity:** Linear regression requires correlation between independent variables to be as low as possible. This can be check by the correlation matrix.

4. **No auto- correlation:** Linear regression requires the data to have no auto- correlation, meaning the data should not be dependent on the previous data. This can be checked by auto- correlation plot.

5. **Homoscedasticity:** Linear regression requires the residual of the data to not be equal across the line. This can be check by the scatter plot of the residuals. There should be no clear pattern in the plot.

7.6 Root Mean Squared Error

Root mean square error can be calculated through the linear regression. It interprets the error from the residuals of the regression line and the actual value. i.e. it measures how well the regression line fits into the data point. RMSE can be obtained from this formula:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x_i})^2}{N}}$$

where x_i is the actual value and $\hat{x_i}$ is the estimated value.

The root mean square error takes any positive value. The desired value of the error is 0 with any value less than 0.5 acceptable. A large RMSE, greater than 0.5, suggests that our model is possibly incorrect and is failing to account for the important features in our datasets.

7.7 QQ-Plots

Quantile-quantile plots or qq-plots are used to compare whether two distributions are similar or identical. Typically, a qq-plot is used to test whether a piece of data or sample is normally distributed. If the two distributions are similar or identical, the qq-plot should represent a straight line with no curves in the center. In our analysis, we had utilized qq-plots as a way to test whether the residuals of our regression results were normally distributed.

8 Contributions

- Kasen Teoh: In this project, I completed the lakes and rivers analysis and helped in the water spring analysis. I also wrote the introduction and background sections as well as the PCA, plots, and correlation in the theory section. I also summarized our findings in the conclusion section.

- Chung En Pan: I completed the spring analysis. I also completed the data portion, RSME (root mean squared error), linear regression in the theory section.

- Nathan Fallahi: I helped edit the introduction, wrote the data section, helped with creating rainfall plots, and formatted most of the document.

- Parsa Ganjooi: I helped with the aquifer analysis, wrote some of the background, wrote some of the data section, and wrote the linear regression portion of the theory.

- Eamon Jarrett-Mann: Wrote code and analysis for the aquifer portion of the project.

References

Acea smart water analytics. (2020). Acea Group. Retrieved from `https://www.kaggle.com/c/acea-water-prediction/`

Gleeson, T., VanderSteen, J., Sophocleous, A., Taniguchi, M., Alley, W., Allen, D., & Zhou, Y. (2010). *Commentary: Groundwater sustainability strategies, nat.* Geosci.

Karr, J. R., & Chu, E. W. (1998). *Restoring life in running waters: better biological monitoring.* Island Press.

Regional national strategy for the utilisation of the nubian sandstone aquifer system (Vol. 3). (2001). Centre for Environment & Development for the Arab Region and Europe.

Wada, Y., Van Beek, L. P., Van Kempen, C. M., Reckman, J. W., Vasak, S., & Bierkens, M. F. (2010). Global depletion of groundwater resources. *Geophysical research letters*, *37*(20).

Wine, M. L., & Laronne, J. B. (2020). In water-limited landscapes, an anthropocene exchange: Trading lakes for irrigated agriculture. *Earth's Future*, *8*(4), e2019EF001274.