Thought process & methodology behind data identifying data abnormalities.  
  
I began by looking at the flow\_1 file and plotting the data

At this juncture I decided not to transform the unix time to human readable date time.

On zooming in, we can see that there exists cyclic trend within the data.   
Observations with value 0 could be treated as erroneous data points but will be handled later.

Empty observations can be dealt with in 2 ways. We can ignore and remove these data points, or we can impute the values based on either overall mean values, median values or using KNN algorithm.

My preference would be to use KNN looking at closest neighbours (k=4).

I have chosen 4 as I would want imputation based on the most recent/closest variables. There are, on average 3 readings taken every hour (1 observation every 16 mins 40 seconds). Taking 4 would try to accommodate the most recent hours readings into consideration.

Using the average over time would not conform with the cyclic trend of our values or give preference to most recently values.

Another method would be to build a regression or time series model. However, that is a complex process that would require much more effort.

The easiest and possibly best method would be to discard those observations. I felt the imputation algorithm would provide a decent fix and hence used it.

I used KNN as the algorithm provides realistic suggestions for missing values but not for the missing time. The algorithm calculates the Euclidean distance between points and therefore will not be a good fit for date values.

Looking at the unix timestamp fields, there is no pattern to which timestamp gets 3 or 4 readings. The majority have 3 readings. A pivot table of count of timestamps shows this:

Table

Description automatically generated

There are two periods where we have missing data. Each of these instances have 15 missing timestamp slots between 4 possible timeslot values. I have decided against using any method to filling this in and have decided to remove these values. There is no fixed logic I could justify applying here and therefore feel the best course of action would be to remove it rather than add a personal bias. The data showed a pattern of 3-3-3-4-3-4 being applied however, there wasn’t sufficient consistency of this across the data.

I then converted the dates to human readable format and dropped the original columns from the file keeping the human readable date format and imputed values.

I plotted the daily average and daily total flow.

Outliers tend to skew the average value. From the graph we can see that there was an outlier at the beginning of our data that skews our average high even though the total flow is low.

We can also see overall low flow average low flow in the middle of march. This is due to the presence of 0 values and low flow readings. I decided against doing anything for these figures as they seemed to follow a trend.

Outlier detection:

For a normal distribution, we would consider any values over 3 standard deviation above mean and 3 below mean as outliers. However, our data does not follow a normal distribution.

Plotting just the flow values in descending order   
  
  
There seems to be just the one value that is much higher even thought there exists some values above the 3 SD mark. I decided to keep those that fall within 3.5 SD to accommodate this and drop the others. 3.5 was chosen arbitrarily but works to accommodate the few high values and removes the outliers.

I have also added a filter to only have positive value as flow should not be negative. The direction of flow can change which could be indicated by another flag (not present in the data) but the absolute value of it will always be positive.

This method worked well to fix the data issues in flow 1.

I would be repeating the same steps for the other 3 files in the folder flow.

For flow 2, I only had to impute values. Everything else was the same as within flow 1.

Flow 3, also had missing values for ‘value’ but no missing values on time. Using the same logic as above we imputed values using KNN k=4.

Here, we did not drop values > 3 standard deviations from the mean.

A plot of the graph shows us that even though there were values that went above threshold to be outliers, on visual inspection they seem to follow a trend/pattern and don’t stick out as outliers.

For this reason, I have not discarded them.   
  
  
  
  
Flow 4 has missing time fields but not missing values. We look at the distribution of measures over timestamps as we did for flow 1.

Looking at the pivot data, we can see a clear and defined pattern of 3-3-3-4-3-4.

I shall use this pattern to back fill missing dates. **Time permitting,** **I will come back to do as it is large and complex in scope and implementation. For now, I will remove these values!**

After removing the values, I have plotted the average and total flow by day

I also plotted the flow values over the time period

And plotted these values in descending order:

None of these graphs suggest there are any outliers to our data.

While calculating for outliers, we do get a small sample that goes over the 3 standard deviation threshold, but none of them go over the 3.6 standard deviation threshold. We can just call these high value observations.

My next step would be to append these 4 files together. I decided against appending them at the beginning and then running the data manipulation. This was done to account for any bias within the individual readings and to keep it to those files and not spill over to the entire dataset. This would include seasonal values as we are looking at time series data.

Reading rain files:

Rain\_1 has no missing data. It does have negative values that will be replaced by 0.

Plotting the values against time:

From the graph we can see that most days have value of rain as 0.2. We can either say that the instrument could possibly need calibration or that the instrument gets that much rain daily.

Looking at it from a daily day perspective:

We observe that on average each day receives between 0.2 to 0.3 measure of rain. Mid April, we can see that the average is much higher even though the total rain collected was less. This usually happens due to high data values. There can be a case made to remove these values as outlier. I have chosen not to remove them as it seems like a baseline of 0.2 exists for the timestamp and any value on top of it would be rainfall for that day which could vary. The assumption is that some days it could only drizzle and provide smaller measures while other days it would be a heavy down pour and provide much more rain.

Rain\_2 has no missing data. Plotting the values

we observe that at the end there is a very big increase in values. These very much seem to be outliers and we would have to remove them from consideration.

Applying a standard deviation of 3 removes these outliers.

Looking at the average and total rain plotted against dates after removing the outliers:

We can see that the average is between 0.2 & 0.3 which is similar to the data we observed in rain 1. To reinforce this decision, we can see that the average holds steady even when the totals vary.

I then proceeded to join both the rain files into a single file.

Once I completed processing my data, I began to view the relationship between rain and flow. We have multiple timestamp across both rain and flow. However, many of these do not overlap or match. I decided to look at the data from a day to day perspective. This would give easy to understand figures that are easier for making decisions.

I grouped the data within flow and rain by calendar dates and computed the mean and total values. Using both mean and total would account for any dates having either more or less measurements.

Dates where either rain or flow data were absent were assigned 0 values. The assumptions used here were that there was no rain or flow on that day which is why we see no data coming through.

After this transformation, I subtracted the rain values from the flow measurements. This was done as rain values are already accounted for with flow.

I further decided to plot the relationship between flow and rain.

The above plot can be interpreted into 2 parts i.e. 1st part of LHS and 2nr part or RHS. The beginning shows a relationship between the peaks on both flow and rain. The peaks for both correspond showing a relationship. The second half of the graph shows no rain but a continuation of flow. We know that the absolute values measured for rain is much smaller in magnitude as that of low. Assuming that they are both reporting the same unit of measurement. I also looked at the correlation between these fields which is 0.46. This is not a strong correlation. The graph suggests that this would be spurious correlation.

The above lot shows the relationship between total rain and flow. It provides the same conclusions from the graph depicting the mean values. The correlation between Total flow and total rain is 0.42.

On removing the null values after joining the data rather than replacing with 0 (to look at those dates that have both rain and flow measures), we see that the correlation for mean values is 0.19 and for total values is 0.38 which would reinforce that rain has no major impact on flow.

s