Middlesex University London

MSc Data Science

4060 – Visual Data Analysis – Individual Course Work 2

Student:

- *Name*: Mykhailo
- *Surname*: Kaptyelov
- *Student number*: M00915847
- *Email*: MK2206@live.mdx.ac.uk

Visualisation of Boonsong Lekagul waterway contamination data

Introduction

In this project I am tasked to explore a given dataset in order to and visualise my findings via various graphs created in Python with the help of Altair library.

Due to default limitations of Altair of 5000 rows, I have decided to split the dataset in 28 parts for the utility of having a selection of even-sized samples instead of disabling it, and produce a highly detailed visualization of given data using loops.

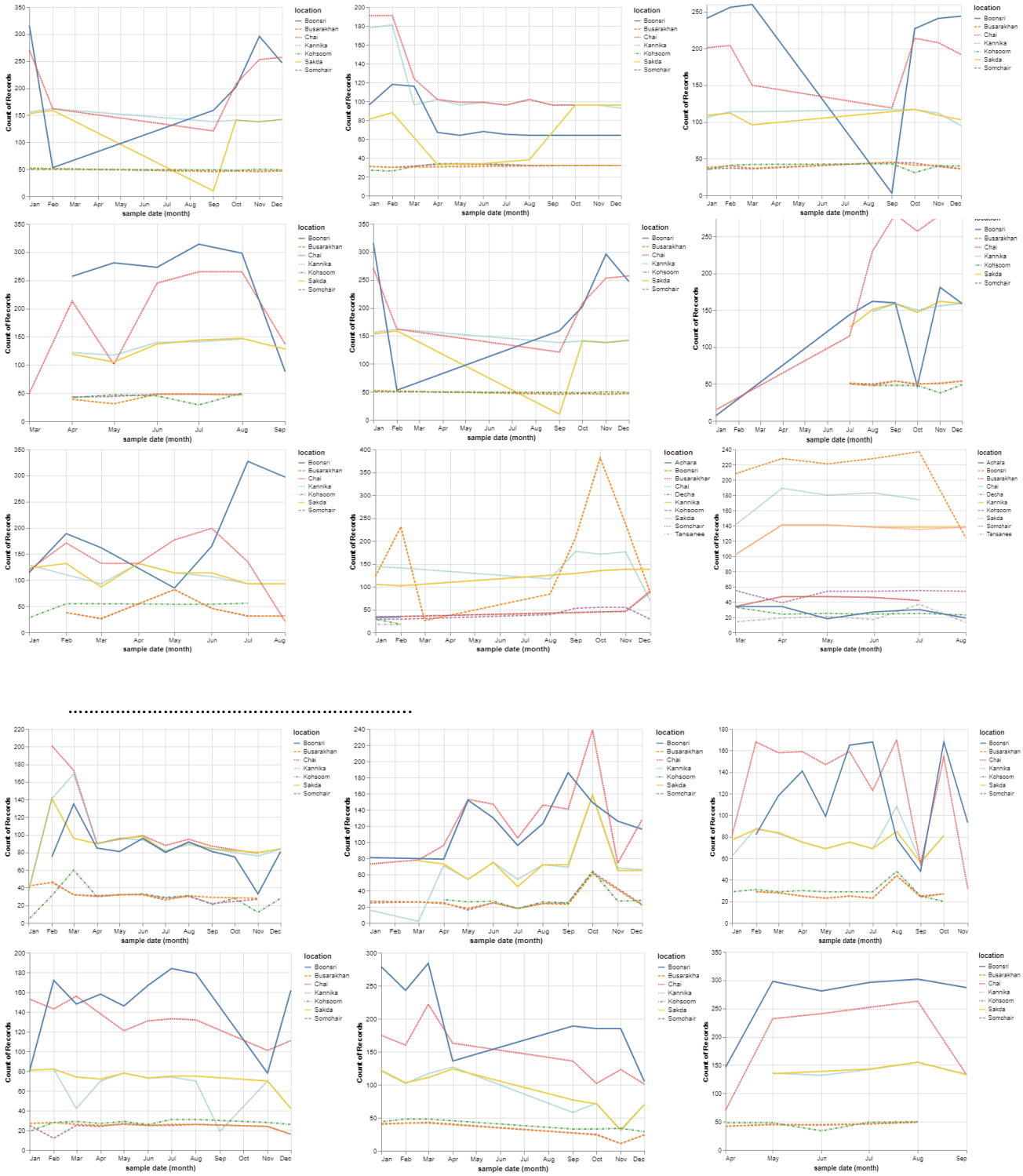
All the relevant code will be provided in the second(with video) submission, sorted by the number of analysis questions.

Analysis questions

- Describe trends and anomalies with respect to chemical contamination
  1. Trends: changes over time and/or sensor site;
  2. Anomalies: sudden change over time or one site significantly different from others.

- Describe any data quality and uncertain issues, such as:
  3. missing data,
  4. change in collection frequency,
  5. unrealistic values.

1. Trends: changes over time and/or sensor site:

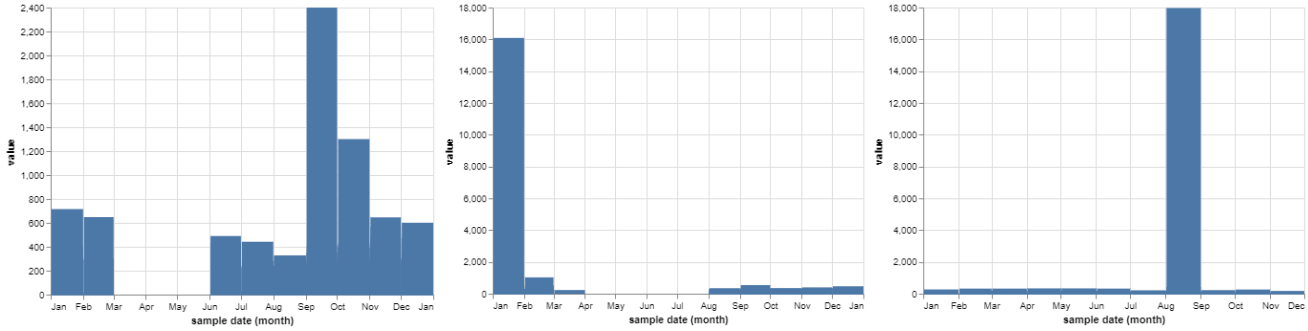


……

This series of line chart ordered in the recorded date, with entries from the beginning and the end of data portions being demonstrated, are showing several things, such as initially areas of Chai and Busarakhan leading in the amount of recorded contamination, but later being taken over by Boonsri, with chai consistently staying at the top, as well as the amount of recorded compounds being initially high, later significantly plummeting, and steadily creeping upwards toward the end of the recorded measurement period.

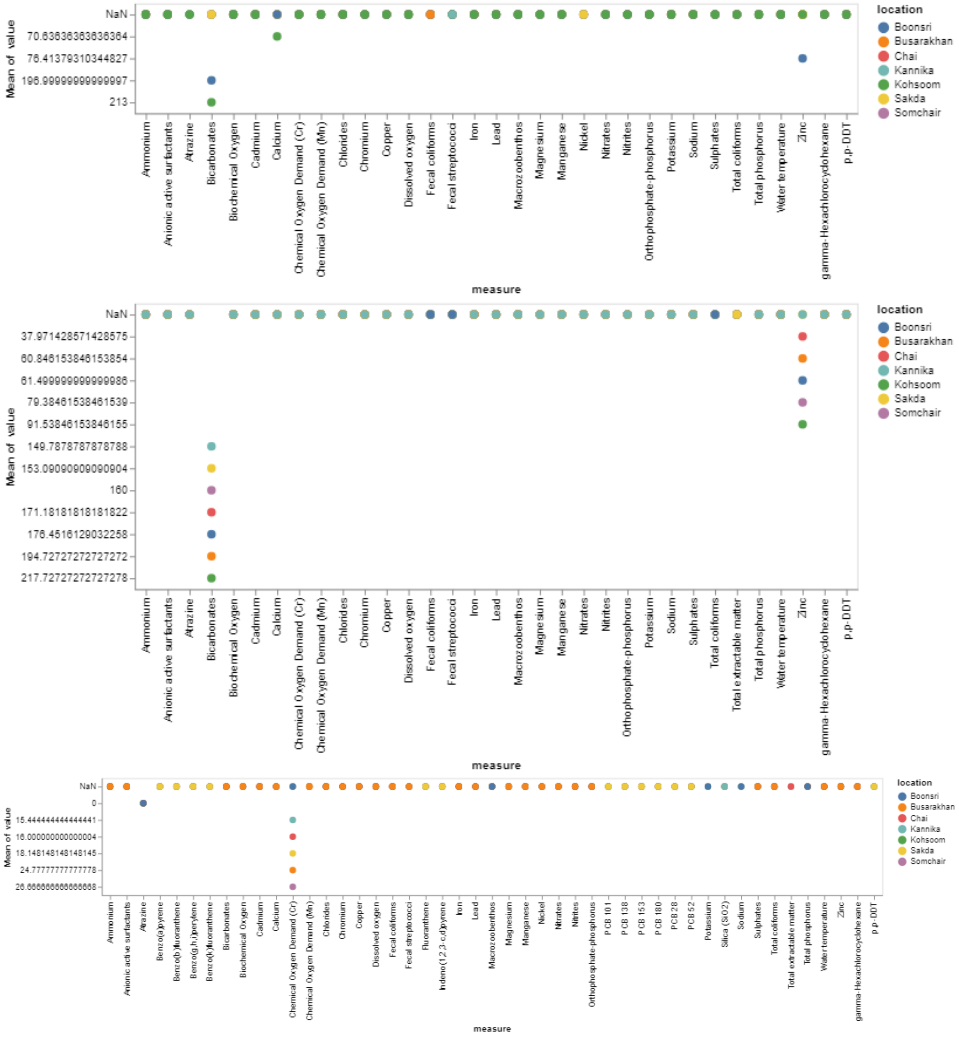
It may be rather accurately stated that, firstly, Chai is a rather contaminated area, secondly, the amount of contamination is on the rise, thirdly, Busarakhan making significant progress at decontamination, and lastly, area of Boonsri failing to do so.

2. Anomalies: sudden change over time or one site significantly different from others.



While regular amount in unites of measurement of various compounds averages at a value of around 2500, In the months of of January 2005, and August 2013, an incredible amount of volume of chemicals has been detected in the tracked areas, which does not make sense, unless a large-scale industrial accident has occurred in vicinity of at least one of the respective areas.

3. Missing data:

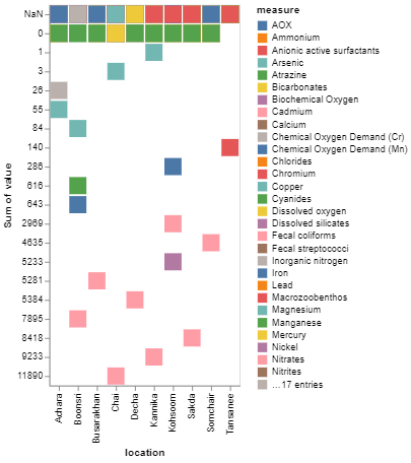


Throughout the dataset split parts, it is a rather common occurrence to see a period with almost no records of any detected compounds, with three examples being provided as evidence.

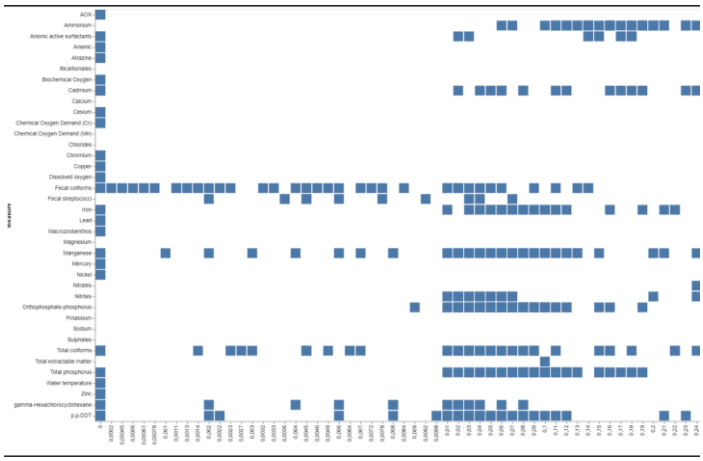
4. Changes in collection frequency:

Compared to the average level of detail of every part of the dataset split, the data around the years 2006, 2011 and 2012 appears to be almost absent.

Average-sized split, for comparison:



5. Unrealistic values:



Apart from the aforementioned at question 2 anomalies, that may also be classified as unrealistic values, throughout the data split the absence of most common compounds (such as trace minerals) contained in natural water reserves is frequently encountered, as shown by the example of excerpt from the very first entry, which I personally find unrealistic, and consider flawed measurement.