Hyperiondev

TASK

Exploratory Data Analysis on the “Dam Levels Individual Nov2015-Mar2016” Data Set

Visit our website

Introduction

Summary of the data set

The data set was downloaded from the "South African Open Data" site and is titled:

"Dam Levels Individual Nov2015-Mar2016"

The data set is in CSV format and is supplied by the Department of Water and Sanitaion. It provides the names and coordinates of 211 dams in South Africa with historic water levels from November 2015 until March 2016.

Additionally, the province, catchment area and dam capacity are provided creating a great set of numbers to perform an Exploratory Data Analysis.

DATA CLEANING

The data set is made up of a combination of continuous and categorical data.

In order to pick up outliers and incorrect data points, a series of visuals were generated.

The first visual was a scatter plot of the coordinate points for each dam, with the intention of locating any positions that are in a questionable location. The plot shows an expected grouping of data points and the info is assumed to be correct. A formal geographic plot will be generated during the exploration part of the task.

The second method used was to generate line plots of the continuous data types, making it easy to spot outliers. Two instances were found and the data points were further investigated. Both instances proved to be true as both belonged to either South Africa's largest dam (the Gariep Dam) or longest river (the Orange River).

The format of each column was printed out with the column name to verify all dtypes are correct. The results showed all columns are formatted as the correct types.

MISSING DATA

Methods - Initial exploration by way of producing a ”shape”, “isnull()” and “missingno matrix” to get a feel for the size of the dataset and determine the degree of missingness. Columns and rows with a large amount of missing data were dropped.

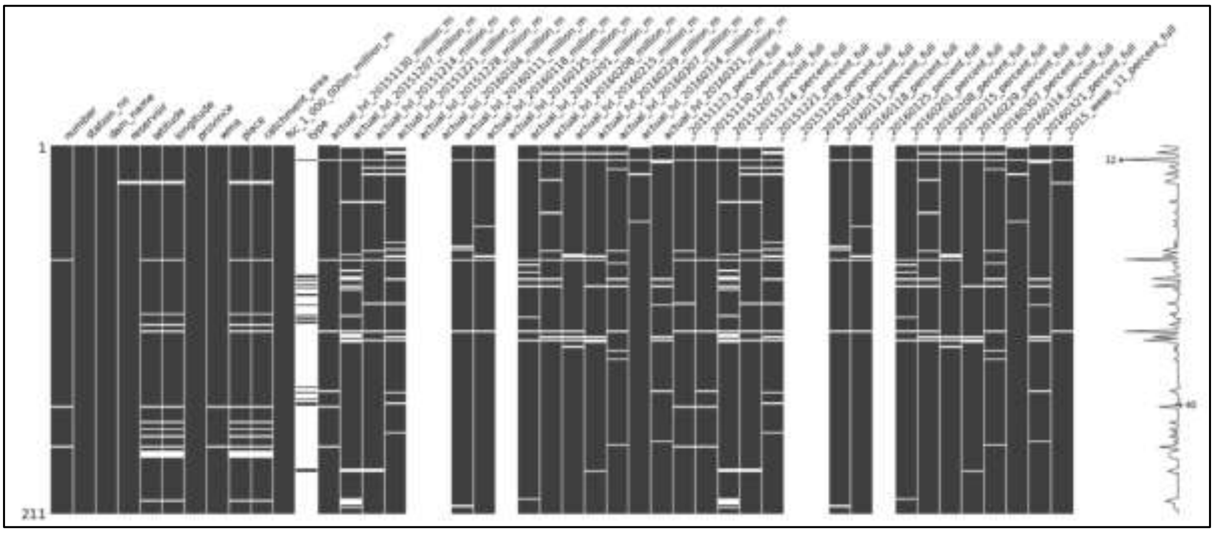


Figure 1: A missingno.matrix of the df prior to cleaning

The data set in raw form has 46 columns, some of which have a large portion of data missing (more than 15%). The following columns have been dropped from the set to create a more efficient dataframe as they are mostly blank:

- Type
- actual_lvl_20151228_million_m
- actual_lvl_20160104_million_m
- actual_lvl_20160125_million_m
- _20151228_percent_full
- _20150104_percent_full
- _20160125_percent_full

The remaining columns show missingness of 10% or less. For these values, the data is weekly records of dam levels and is therefore not possible to impute the blank inputs, but is sufficient to perform an analysis.

Furthermore, an isnull().sum” analysis was performed over axis=1 and rows with more than 15 data points missing out of the 46 total columns were removed.

For the categorical and geographical type columns with missing data points, there is a dam name available for 100% of the records, we are therefore able to find missing longitude/latitude/place/catchment area details by conducting further research for certain records where blanks occur if necessary. For the purpose of this task, this has not been undertaken.

Within the continuous type columns representing dam levels, these are classified as MNAR (Missing Not At Random) and therefore it is not possible to impute and replace the values.

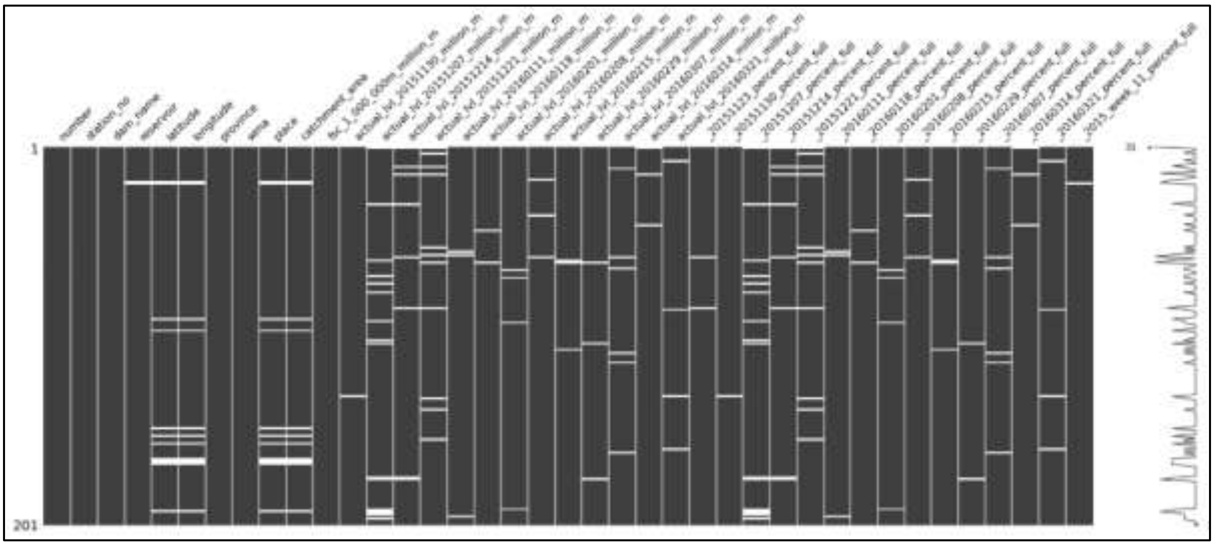
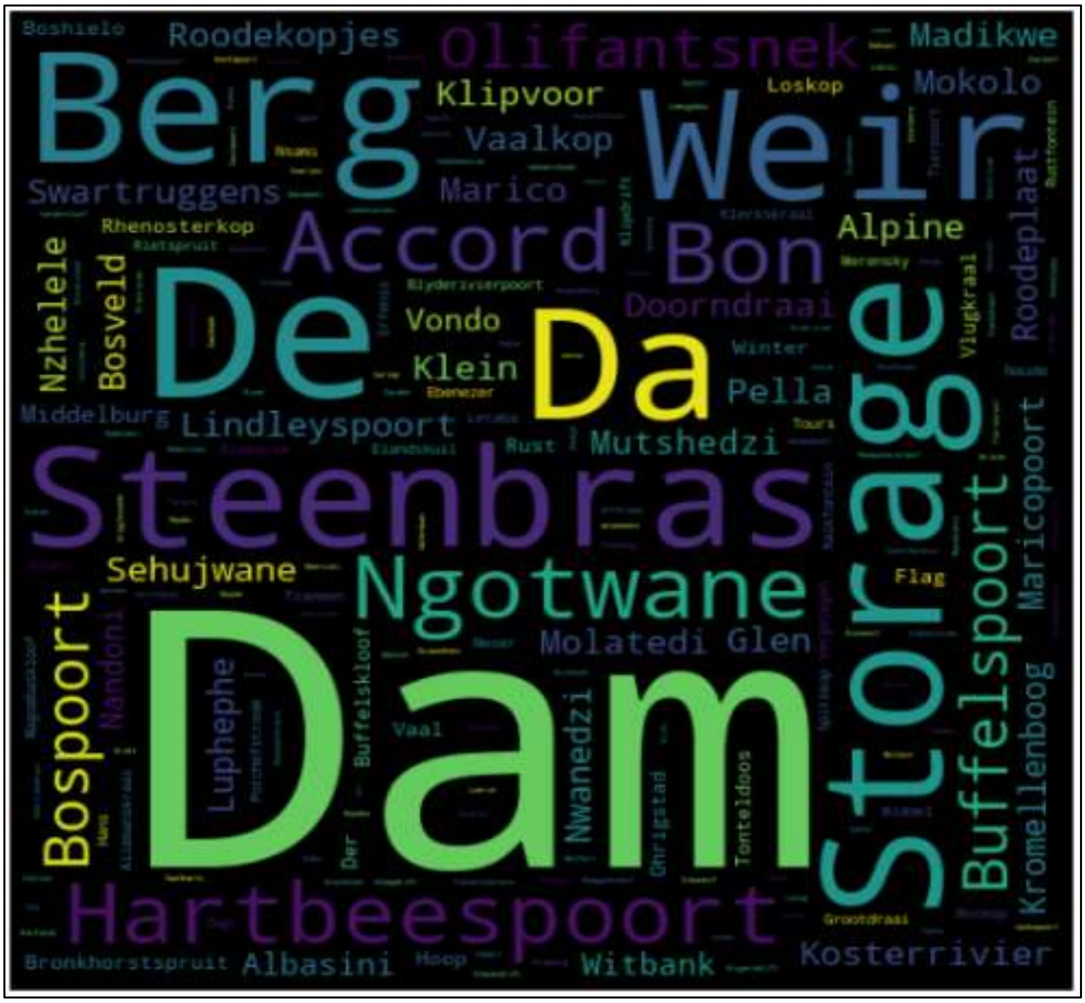


Figure 2: A missingno.matrix of the df post cleaning cleaning

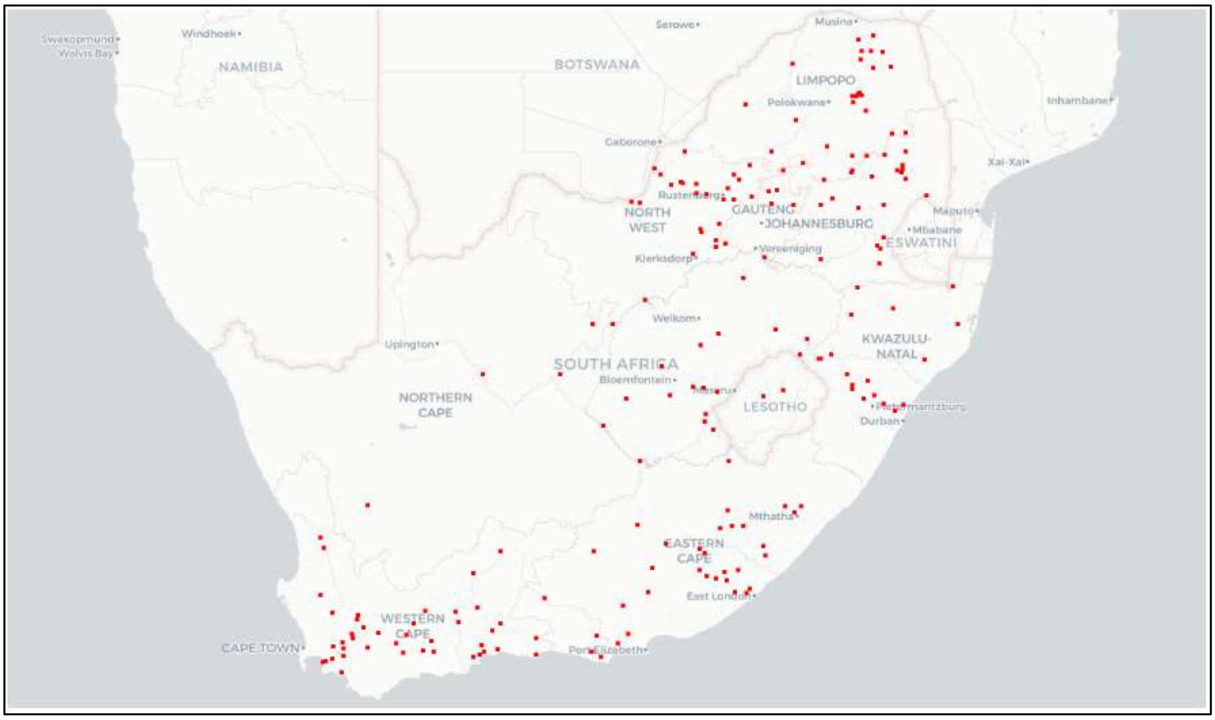
DATA STORIES AND VISUALIZATIONS

The EDA conducted on the historic dam levels data provided insight into the state of South African dams during the record period as well as level variations over time.

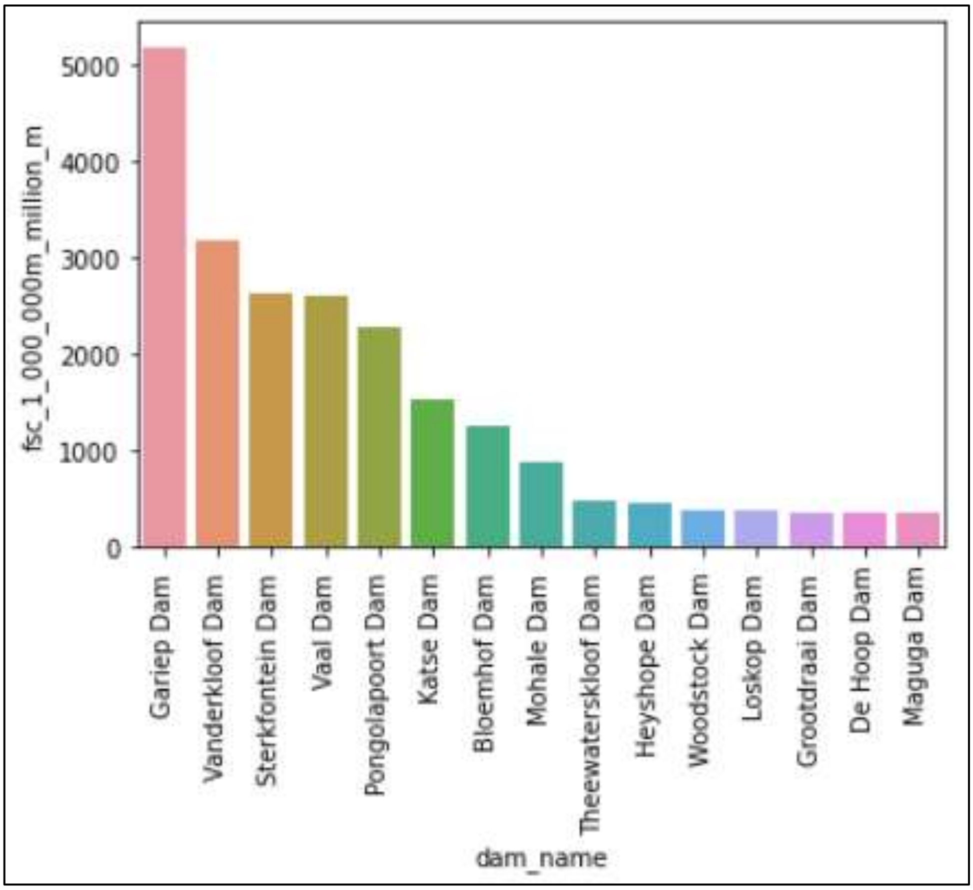
The data represents dams across the entire country, with familiar names shown in the below WordCloud.



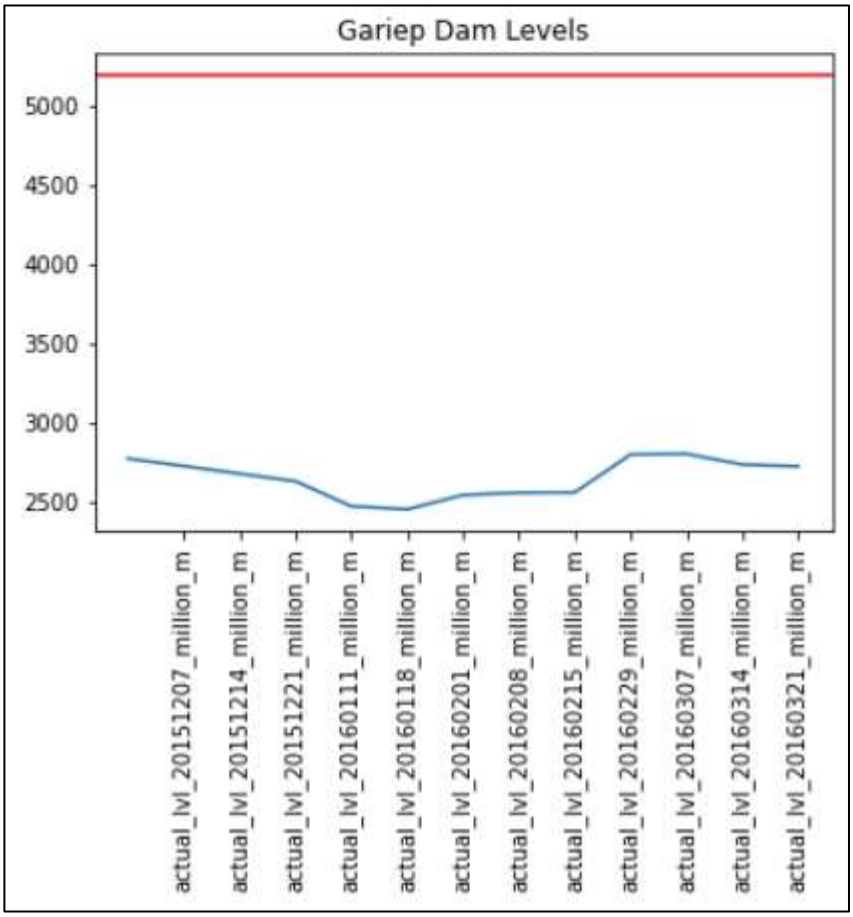
A geographical visual created with Geoplotlib provides an overview of the dam locations, each dam shown with a red square and based on the coordinates stored within the CSV file.



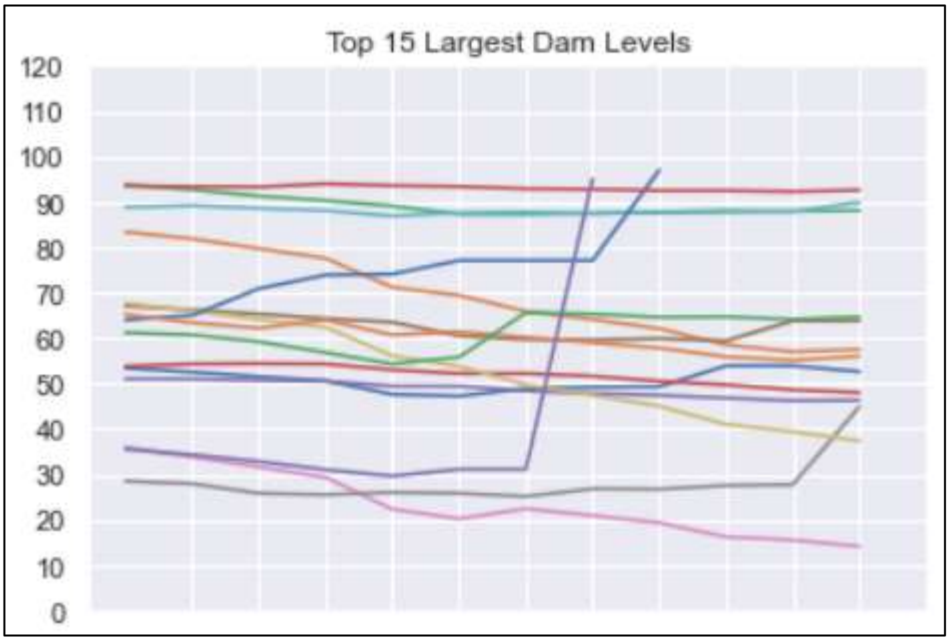
As a starting point, a bar chart was created to show the names and comparative sizes of the fifteen largest dams within the country. The biggest by full capacity being the Gariep Dam, it is well known that this is the largest dam in the country.



Further analysis was conducted on Gariep Dam by plotting the levels as a time series compared to the total dam capacity.

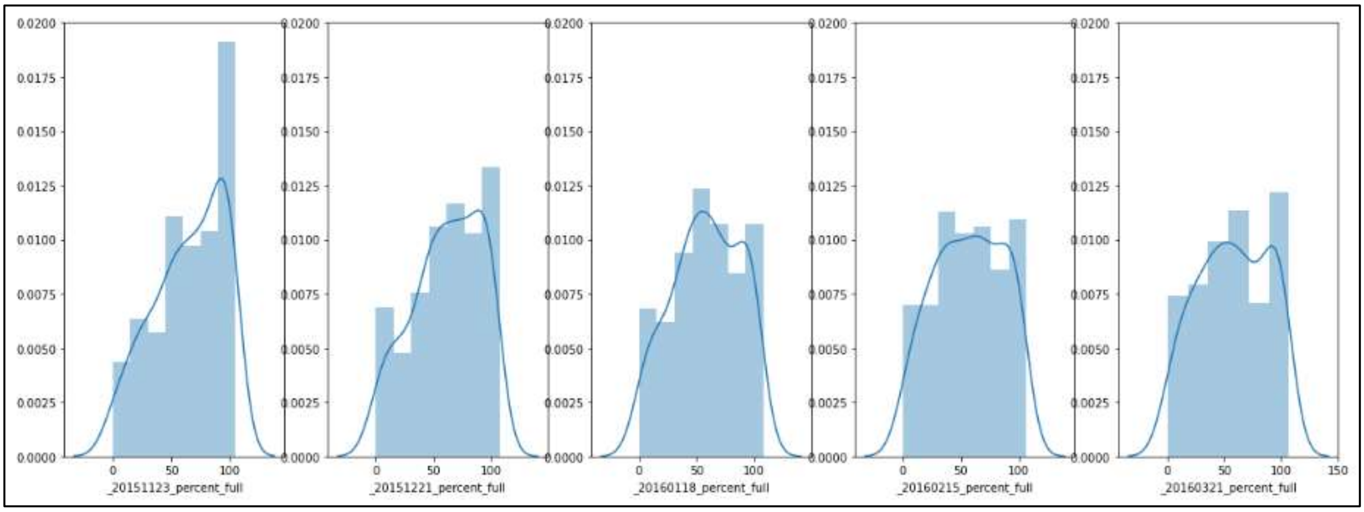


The total capacity is shown by the red line and actual measurements by the blue. It is clear from this graphic that the dam was hovering just above 50% full during the period.

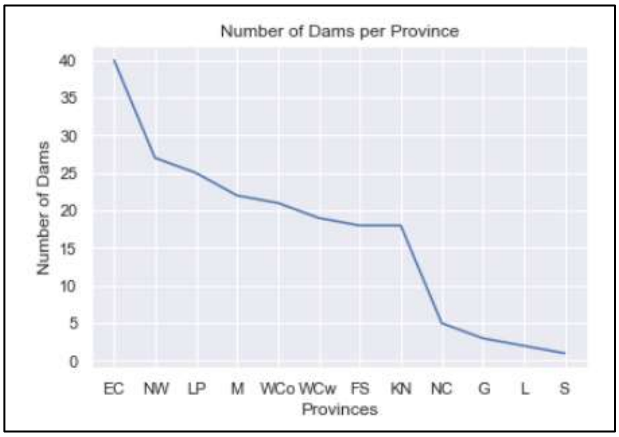


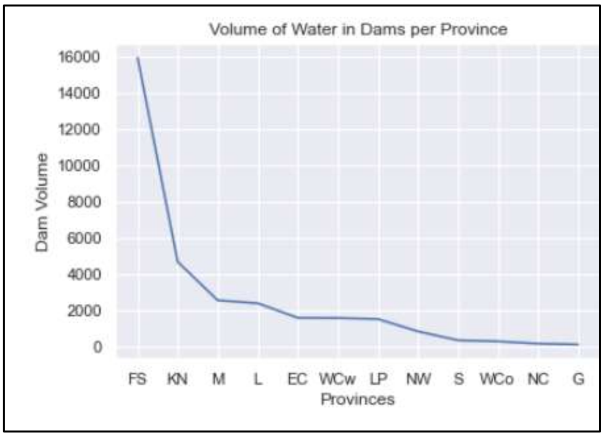
The second graph provides information on the top fifteen largest dams with the levels shown as percentage of full capacity. From the data, we see that there is a wide spread of levels ranging from 15% up to 95%.

Furthermore, we can see how the levels vary throughout the period. The following graph looks into a relationship between the levels of the entire population vs the months during the period. The graphs show a move from a large amount showing a 100% level in Nov 2015 to a shift towards lower level throughout the months to March 2016.



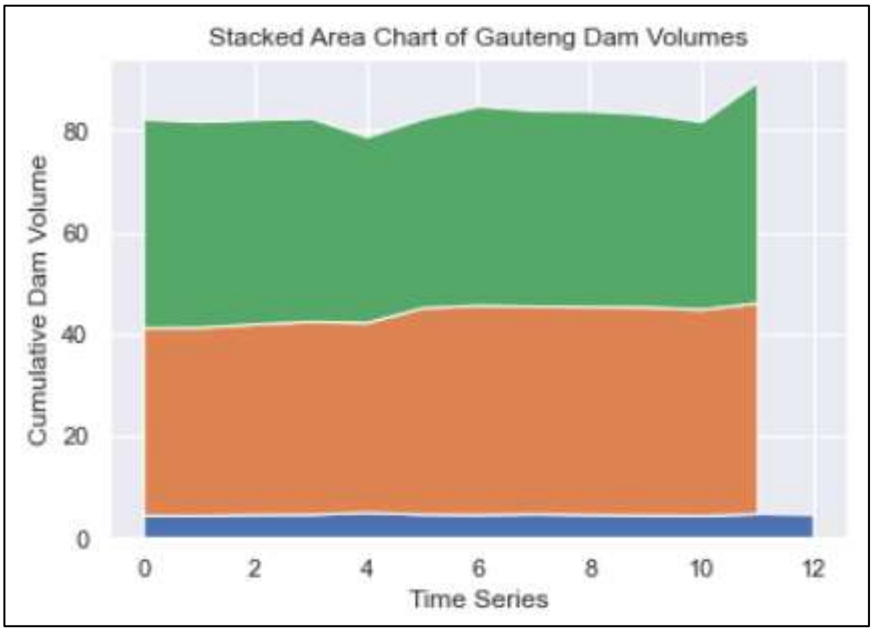
The Geoplotlib graphic shows a good spread of the dams across the country. The following two graphs show a comparison of dams and total capacity per province. It is noted that there is no strong link between the number of dams and total capacity per province.



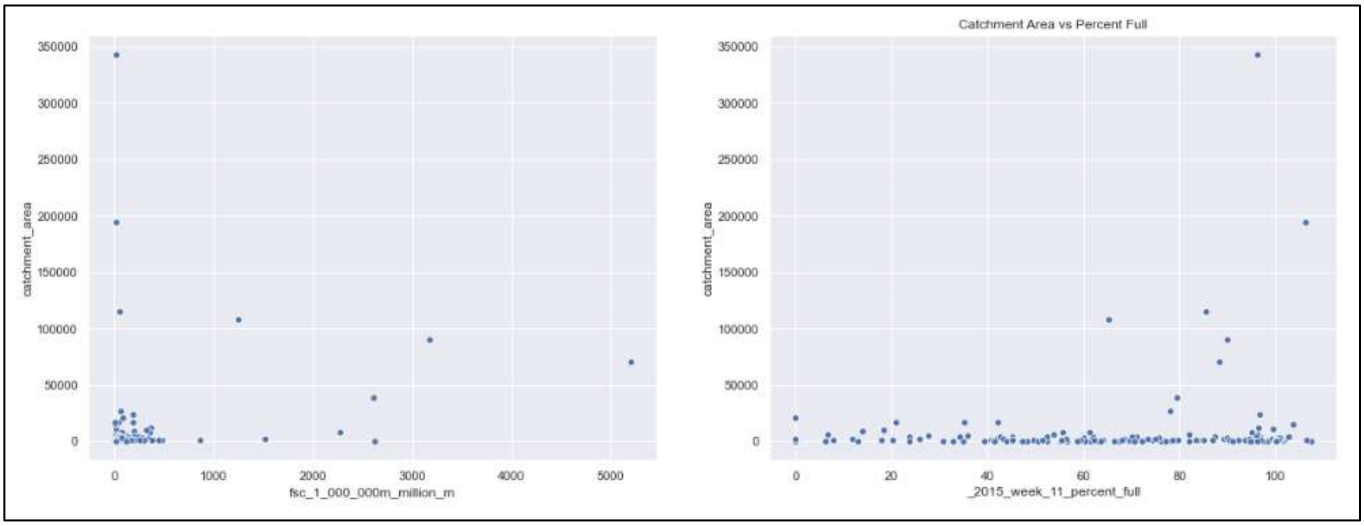


Further to the above analysis, a more detailed investigation was conducted on the dams within Gauteng, with the intention of generating a graphical presentation of the portion allocated per dam over a period of time.

The stacked area chart shows that the Roodeplaat and Bronkhorstspruit Dams hold the large majority of water compared to the Bon Accord Dam. Note, the Rietvlei Dam did appear in the initial CSV data set but was removed due to lack of sufficient data.



A final analysis was completed to determine if there is a relationship between the catchment areas per dam compared to the capacity and recorded dam levels.



The scatter plot shows a fairly strong direct relationship between the size of the catchment area the water levels, but not a very strong trend when looking at the catchment area size vs capacity. These results are as expected.

ENSURE THIS DOCUMENT IS NEAT AND CAN BE ADDED IN YOUR PORTFOLIO

THIS REPORT WAS WRITTEN BY : JUSTIN BAINBRIDGE