**Project Report GitHub URL (insert URL here)**

<https://github.com/BMSinnott/UCDPA_BrendanSinnott.git>

**Abstract (Short overview of the entire project and features)**

This project looks to see if there is any correlation between rainfall events and the sale of pumps that can be used for flood control and general dewatering. The scope of this project will look at both rainfall characteristics in Germany such as daily rainfall levels, rainfall patterns during the years and months and extreme events over a 5-year period. This will then be compared to the sales patterns for watering pumps that are used for various application such as domestic flood management to large industrial scale dewatering pumps. Although the scope of this project is for Germany only, the intent will be using this model and extend the scope to a more global view that may aid long term global business strategy.

**Introduction (Explain why you chose this project use case)**

The author works for a multinational company that is a leader in fluid engineering that specialize in pumping, agitation, mixing, separation and purification technologies for fluids of all types. As part of the project selection process, the objective was to identify a data analytics project that could link freely available ‘big data’ to the overall business strategy of this company. Although the company covers a broad range of products and markets such as chemical processing, water, oil & gas, mining, construction, domestic and commercial wastewater management, it was decided that the focus should be on one of the key growth areas which is the water network.

After an extensive search for freely available data sets that might link to some products in this business segment it became apparent a direct link could be made between several topical subjects not just globally but for the company itself. This was namely linking the subjects of sustainability, the focus on global warming and products that impact one of our most valuable resources which is water. The company has a water division that has a range of products for the entire water network, spanning, desalination, clean water distribution, wastewater processing and flood management, called dewatering.

According to Forbes1, some two thirds of the world’s land area will experience wetter, more variable conditions as the Earth warms, making extreme rainfall and flooding more likely, a new study has found. Research by the Institute of Atmospheric Physics (IAP) of the Chinese Academy of Sciences (CAS) and the UK Meteorological (Met) Office, published recently in the journal Science Advances, indicates that most communities will experience even more extreme wet conditions with each degree of temperature rise. There are extensive datasets that are freely available to assess these trends in rainfall. Given Germany is one of the largest markets for the company, the author did some research on that region and discovered most of the data is publicly available and can be accessed via the Climate Data Center2 portal.

Using this data and the sales data from the company, this project will aim to assess if there is any linkage to these extreme weather events and the demand for dewatering products in the author’s company. It will look at rainfall data trends and assess if there is any correlation between the sale of products that can be used to manage these extreme events. It will assess if the company should be doing anything different, such as optimizing the supply chain to better align to these events and prepare for growth in these areas.

**Dataset (Provide a description of your dataset and source. Also justify why you chose this source)**

**German Weather Data:** This first dataset came from ‘The Deutscher Wetterdienst’ which is a higher federal authority under the Federal Ministry of Transport and Digital Infrastructure. This site has both a interactive mode which gives tabular and graphical previews of all the data from the German weather stations. In addition, they have and Open Data Server that can be directly accessed for download.

The site has data from 1,084 weather stations across Germany from 1900 to 2021. The overall dataset has more than 17 million weather observations. For the purpose of this project, we will only review the precipitation readings. I chose Germany and this dataset source as this is one of the largest customer bases for the company in review. The project demonstrates the download of these weather station raw files directly from the website.

Graphical user interface, application, map

Description automatically generated

Further research also identified other projects on GITHUB that had used similar datasets and as such was deemed an efficient use of some consolidated data. The main project referenced was [GitHub - janikvalentin/heavy\_rainfall\_events: An investigation of occurrence and intensity of heavy rainfall events in Germany](https://github.com/janikvalentin/heavy_rainfall_events) . This project contained a subset of data that will be used in this project. The datafile was a zipped CSV file on GITHUB and was downloaded to this project folder under the subfolder ‘Weather Data’.

**Sales Data:** This is a confidential data set which was made available by the company. It included line level sales data for all products sold by one legal entity globally. The data was pulled from a sale database application and output to an excel file. It contained data for the period Jan 2016 to Nov 2021. Each year of sales was in a separate sheet in the excel file and the file contained approximately 210,000 lines of data. For each line of data, the various details of region, product, quantity, date and twenty other parameters. This was an exceptionally clean data file and was deemed to contain all the critical data to conduct a robust analysis of sales in relation to the products in question.

A ’sample dataset’ was used to upload to GIT Hub to demonstrate the codes works. The project for this is called **UCDPA\_Brendan\_Sinnott\_SampleData.ipynb**. Although the full project is stored on GITHUB and shows output the actual dataset will not load for this **UCDPA\_Brendan\_Sinnott\_Ver1 – Master.ipynb.**

**Product Reference:** This dataset was a small table that cross references the product in the sales file to the product family. As the sales dataset contained all sales data for all products it was necessary to narrow down this data set to products family. As the product family did not exist in the raw sales date this CSV cross reference file was required to enable us to select ‘Dewatering’ products only. A merge of both datasets and a slice of Dewatering products was performed in the project.

A screenshot of a computer

Description automatically generated

1. **Implementation Process (Describe your entire process in detail)** 
   1. **Importing of Python Packages**

This section imported all the required packages and was extended a number of times as the project developed. The main packages used were ‘BeautifulSoup’ to retrieve the data from the German weather data website, ‘numpy’ and ‘pandas’ were used extensively for dataframe handling, various file and parameter handling packages such as ‘datetime’, ‘requests’, ‘zip’, ‘io’, ‘os’ and ‘glob’. ‘Seaborne’, ‘matplotlib’ and ‘Image’ were used for the graphics and saving figures. Other packages such as ‘time’ and ‘warnings’ were used for testing and debugging the code.

**1.2. Retrieve the Weather Data from Website**

This section retrieves the Weather Data for rainfall levels in Germany. It was acquired by scraping the data from the public source the German meteorological service <http://opendata.dwd.de>. BeautifulSoup was used to acquire all the files needed from the website. The first step was to get a list of the file that contained the daily weather readings ‘tageswerte’ translated as ‘daily values’. Although not a necessary step a list of these was then stored in the file 'Weather File List.txt', sample below.

A screenshot of a computer

Description automatically generated with medium confidence

Using the acquired list of files on the website then a loop was created to extract each of these files and store them in the ‘Weather Data’ folder of the project. These files were zipped on the website, and such must be extracted using the ZipFile function. Given the extensive dataset this step takes some time to complete and generates 15,200 files and required greater than 1.6GB of storage.

Graphical user interface

Description automatically generated with medium confidence

Although we could have then used some scripts to combine all these files and datasets into once datafile, it was deemed more efficient to use a previously consolidate dataset from another project. This project created a file ‘weather\_data.csv’ which was a zipped file and already stored on GITHUB.

### 1.2. Load & review the consolidated Weather Data

### Create Functions to define extreme rain conditions and new column to capture level

The German Weather Service speaks of extreme rainfall if the amount of precipitation exceeds 40 litres per m in 1 hour or 60 litres per m within 6 hours. Since the dataset gathered is aggregated over the daily amount of precipitation, we will use the threshold of 60 liters per m per day to classify a extreme rainfall event.

These next two sections of code created two new columns in the ‘weather\_data’ dataframe. The first one being [‘RATING’] was created using a custom function RATING which populated the column based on the variables for ‘Extreme’, ‘Normal’ or ‘Heavy’.

The second section again used a custom function to populate a new column [‘PEAK’] with the rainfall level on that day only if was considered an ‘Extreme’ level of rain or a PEAK event. If the recording was normal or just heavy on the day, then this column was just populated with zero.

### Load & review the consolidated Sales Data

This section loads the sales dataset, cleans it up, strips out the data for Germany and the Dewatering products only and finally ‘desensitises’ the data so it can be shared. Only the required columns are loaded and then cleaned up to add some additional columns as needed for [‘Month’], [‘YEAR’], [‘Product Group’] among others. The data set will exclude 2021 as at the time we did not have a full dataset for neither sales nor weather for Dec. Note, the ‘Capacity’ column is a parameter which measure the relative pumping capacity of the particular products, this is worth noting as it is a key variable used later in the analysis.

Table

Description automatically generated with medium confidence

1. **Results including charts and data:**
   1. Now that we have two data sets it is time to see what the data is telling us and then compare the two datasets together to see if we can draw any conclusions or correlations across the two data sets. Let’s first of all look at the weather data.
   2. Rainfall: Looking at the mean of rainfall at each station over the five year period would indicate that there is no major change in the overall peak trends, if anything recent years would indicate a reduction in overall annual rainfall. However, there is a higher frequency and higher level of rain events in recent years from the end of 2018 on.

Chart, line chart

Description automatically generated

* 1. Reducing the scope to the past 5 years and looking at seasonal patterns does reveal some interesting findings that do not match current assumptions in the business. Peak rainfall is in the summer months and extreme events also occur in the summer months.

Chart, line chart

Description automatically generated

Chart, bar chart

Description automatically generated

* 1. Sales: Looking at the sales data also revealed some interesting facts. The normal forecasting model for dewatering would show an increase as we approach year end with relatively flat results in the earlier month. However actual sales would indicate peaks in March and in the summer months. This needs to be considered in future forecasting models.

Chart, histogram

Description automatically generated

* 1. **Combined Rainfall & Sales Datasets:** combined we can see some clear correlation between the summer peak rain and the summer peak sales events. Taking into consideration the total rainfall these correlations are still present.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

* 1. If we also consider the natural business seasonality, i.e., December is a short month for production and sales, this carries into a typical slower Jan and a ramp up in Feb and Mar. If we remove this business impact and level these months, we can see even a higher correlation with the summer peak rainfall and sales.

Chart, histogram

Description automatically generated

1. **Insights (Point out at least 5 insights in bullet points)**
2. In Section 5.1 we looked at the weather data to assess any trends on total and extreme rain events over a 50 year period. The Global warming discussion covers a lot about rise in temperature, however, it typically does not talk much about rain. The analysis of the rain fall over the past 50 years would indicate although no significant trend jump out, there is a decline in mean rainfall, total from extreme events and the number of extreme events. Other than 2002, which had a well know peak year of floods, it can be said that rainfall overall if anything, is in decline.
3. As we look at the seasonality of the rainfall over the year, section 5.1. demonstrates that the summer months are typically the highest period of rainfall in both total volume, number of peak events and total rainfall from peak events. A significant number of extreme events fall in the summer months May-July and this is later seen to impact sales in this period.
4. Overall sales for this business look to be very stable, section 5.2. The apparent drop in Q1 2018 was associated with an ERP deployment. Over the 5 year period very little seasonality is apparent in sales with the exception of year end, this is related to factory and sales office closures during the Christmas holiday periods. When looking closer by mean and span by month, the main variations that should be considered investigating further for forecasting purposes would be March and July periods. Sales in these months is typically higher.
5. As the scope was narrowed to dewatering pumps overall and then to dewatering for Germany, section 5.2.4, some unexpected trends were identified. In both cases March was the peak month for sales followed by higher sales and variation in the summer months. This would not align with the typical forecasting for this product range. Although March may be impacted by 'and overall seasonally high production month' the variation in the summer months warrants better alignment to forecast modelling.
6. Finally, as we combine both rainfall and sales, section 5.3. a number of interesting facts emerged. Todays forecast from the sales team is typically very linear over the year with a slight increase in winter months. As this forecast process is very new to the organization it may be based on the thought that more dewatering is needed in the winter, however, in reality this does match actual sales history. Most interesting is the peak sales in the summer months appears to be heavily correlated to peak rainfalls in these months. As current forecast models do not reflect this, going forward we should take this into consideration in future forecasting models. This alignment could potentially significantly help improve supply chain alignment to these peak periods thereby improving customer satisfaction.
7. **References (Include any references if required)**
8. ‘The Deutscher Wetterdienst’ <https://www.dwd.de/DE/Home/home_node.html>
9. An investigation of occurrence and intensity of heavy rainfall events in Germany. GitHub - <https://github.com/janikvalentin/heavy_rainfall_events>
10. Are heavy rain events increasing in frequency and intensity in Germany? <https://towardsdatascience.com/are-heavy-rainfall-events-increasing-in-frequency-in-germany-2129b5d9d448>
11. Sales Data from company databases