Data Analytics Career Track

Southern Water Corp Case Study Part III - Descriptive Statistics

Previously, in the financial and economics section, you've helped Southern Water Corp improve their understanding of the financial and economic performance. In Particular, you identified that Surjek is the most expensive of the three desalination plants that Southern Water Corp operates. This is largely due to Surjek's large cost-base, however, this has also been driven by **higher maintenance costs due to desalination pumps failing**. The Engineering Team doesn't have time to continuously look at data to proactively point out when the pump is failing; they've also got Kootha and Jutik to look at. They're stretched thin in terms of resources and may benefit from a data analyst **creating an alarm for them** that is able to **detect failure on their behalf.**

As a Data Analyst, you've got to get used to using statistics to tease out insight from your provided data. Most commonly, the data you will be analysing will be **time series data.** Time series data is essentially data which is recorded at a **specific time.** For example, if we were to check the temperature **every hour for one day,** we'd have 24 unique **time series data points.**

In this case study, you will be dealing with minute by minute **time-series data.** The data will be recorded every minute for a number of variables like temperature. In particular, you will use each of the descriptive techniques you learned and reinforce this with analyzing the behavior of a desalination plant pump **before failure and during failure.**

The engineers at Southern Water Corp look at the variables in the table on the next page to identify and track pump performance.

You needn't understand these terms in any great detail for your analysis. You'll be looking at the visual data trends. We have included the explanation as to what each of these variables track to make it easier to interpret the trends you will be plotting out later.

Please note that the data has been split into **three (3) tabs**, for ease of your analysis.

Data Repository Table – Raw: This includes all the **raw** data that you will use for your initial analysis.
DRT – Rolling Mean: This includes the raw data transformed into 30 Minute Rolling Mean.
DRT – Rolling Stdev – This includes the raw data transformed into 30 Minute Standard Deviation.

Variable	Variable Explanation
Volumetric Flow Meter 1 / 2	This is a flow meter reading which tracks the amount of desalinated water pumping through the pipe on a minutely basis.
Pump Speed (RPM)	Pump Speed (Rotations Per Minute) indicates how fast the pump impeller is rotating. Essentially, the faster the Pump is running, the more energy the pump is consuming. (E.g. for periods of high demand, the pump speed may ramp up).
Pump Torque	Pump Torque is a meter output indicating how much force is being applied to the pump at any given point in time. For example, if the Pump Torque is very high – this indicates the Pump is undergoing significant stress due to heavy workloads.
Ambient Temperature	Ambient Temperature is read off a temperature-sensor which informs us how hot/cool the pump is. For periods of **high demand** the pump temperature may increase to reflect the increased workload.
Horse Power	Horse Power references the pump's motor and motor speed. As the Pump works harder (or eases off), the horse power will either increase or decrease based off the energy required.
Pump Efficiency	Pump Efficiency is a measure of how effectively the pump operates. Essentially, each pump has a theoretical amount of water it can 'pump.' A measurement is taken every minute to check the flow rate versus this theoretical maximum and a pump efficiency is derived. For example, if a pump is producing 40 m3/d of water and the pump theoretical maximum is 40 m3/d, the pump efficiency would be 100%.

Now that we've covered the variables that will be in the dataset you will be analyzing, it's time to briefly cover **why** this statistical analysis matters.

Data Analysis is **only valuable** when it can be tied back to some form of **business insight with tangible/non-tangible outcomes**. If Southern Water Corp can **proactively identify pumps performing abnormally** they will likely be able to reduce their maintenance costs. More importantly, having an **alert** in place will save the engineers time. This means they can let the system analyze failures on their behalf as opposed to them always looking at the signal(s) themselves. If you are able to do this reliably, SWC can reduce the overall production costs and improve the overall **market economics** which will become a business advantage.

This is where you come in; With the knowledge and foundations you've established across the Statistics Unit, take **one final dive** into Southern Water Corp's Data and help management better understand the insights they can capture from the rich data assets they have.

You're going to create your first **Statistical Data Alarm** that will help inform the Asset Engineers when they should take a look at Surjek's Pump Performance.

With management very much aware that **Surjek's costs are rapidly dwarfing the other plants**, they have requested the following Statistical Data Analysis for you to complete:

1) **Descriptive Statistical Analysis** – Analyzing the desalination plant data using statistical means, are you able to identify any trends or behaviors that will enable them to better understand pump failure?

2) **Inferential Statistical Analysis** – Analyzing the desalination plant using inferential statistical means, are you able to identify any **correlations** between the variables which may indicate when the pump may fail?

Ultimately, your challenge as a data analyst is to tell a meaningful story that drives **meaningful** insights. For your statistical analysis, you'll be measuring your analytical success by ensuring you can clearly tell the Executives at Southern Water Corp a meaningful story. This story should unpack which variables need to be monitored closely with respect to pump failure, as well as creation of a multivariate linear equation that will become the statistical alarm the engineers can use to automatically pick up failure.

This assignment is broken up into two (2) overall components. In each section, you analyze one of the two question(s) listed above and tell the overall story you've uncovered.

In Part I of the assignment (8.4) you're going to tackle Descriptive Statistics.
In Part II of the assignment (8.6) you're going to tackle the Inferential Statistics.

Lastly, in Part II of the assignment (8.6) you're going to conclude our overall analysis with a presentation making use of the story you've uncovered in both Part I and Part II of the Statistical Data Analysis.

Time to get started!