CMP3749M Big Data Assessment Report

A logo of a university

Description automatically generated

Mathews Joy

[25186202@students.lincoln.ac.uk](mailto:25186202@students.lincoln.ac.uk)

JOY20779936

University of Lincoln

School of Computer Science

When finished when ever I mention a specific function e.g. “toPandas()” include a reference next to it of the link to that functions documentation.

**Task 1 – Analysis of Nuclear Plants dataset**

**Part 1**

In the dataset we analysed for task 1 ”nuclear\_plants\_small\_dataset.csv” we used the pyspark “.isNull()” function to confirm there was no missing data points in the dataset. This function “is used to check if the current expression is NULL/None or column contains a NULL/None value”1. In our scenario we identified that there was no missing values in our dataset, as a result no further action was required. In the case a column contained 1 or more empty values, it is important to identify it and take necessary steps to fix the missing values, this is to avoid skewing the results from analysis you might perform. First it is important to identify what type of missing data you have; “Missing completely at random (MCAR) data is just randomly missing, missing at random (MAR) meaning missing conditionally at random based on another observation or missing not at random (MNAR) which is missing as part of how it is collected (deliberately missing)”2. Based on the previous statement by identifying these types helps in choosing the correct method to handle the missing data. For instance, if the data is classified as MCAR, it might be best to simply exclude those cases from your dataset. Although simply deleting data is not recommend, but as the missing data is MCAR, removing the data would not bias the analysis performed. This is because the missingness is not dependent on observed or unobserved data. But if it’s MAR or MNAR, more sophisticated techniques might be needed to avoid bias in your analysis. These could include imputation, which involves filling in missing values with substituted values such as mean of the column with missing data. Other options include acceptance/ignore them, “this is the most conservative option involves accepting your missing data: you simply leave these cells blank. It’s best to do this when you believe you’re dealing with MCAR or MAR values”3. Although the data seemed quite clean and structured, upon inspection with the “.printSchema()” function (which displays the data frame columns and there datatypes), we can see the column names are inconsistent. For example, in the pressure sensor column names, they have a random space value in the column names, which the other columns do no have. Due to this inconsistent it is best to keep the column names in uniform so we used the “.replace()” function on each column name to check for random empty spaces in the column names and get remove them. This means in future analysis no unexpected errors will be through when trying to access specific columns.

**Part 2**

For this task we are asked to find the summary statistics (minimum, maximum, mean, and median) for each group of subjects (normal and abnormal), and then for each group plot the boxplots for each feature. We can first approach this task by converting the pyspark data frame to “pandas” data frame; “pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool” 4. The reason we convert to pandas is it allows cleaner syntax for performing aggregate functions in general and integrates with graphical visualisations easier, however these calculations can also be performed directly on the pyspark data frame. We can convert our current data frame to pandas via the “.toPandas()” function. Next,

**Task 2 - MapReduce for Margie Travel dataset**

**Task 3 - Big Data Tools and Technology Appraisal**

**References**

1. [PySpark isNull() & isNotNull() - Spark By {Examples} (sparkbyexamples.com)](https://sparkbyexamples.com/pyspark/pyspark-isnull/)
2. Week 3 big data workshop
3. [Missing Data | Types, Explanation, & Imputation (scribbr.co.uk)](https://www.scribbr.co.uk/stats/missing-values/)
4. https://pandas.pydata.org/