****

**By**

**Sunil Raj Thota**

**Date: 04/28/2021**

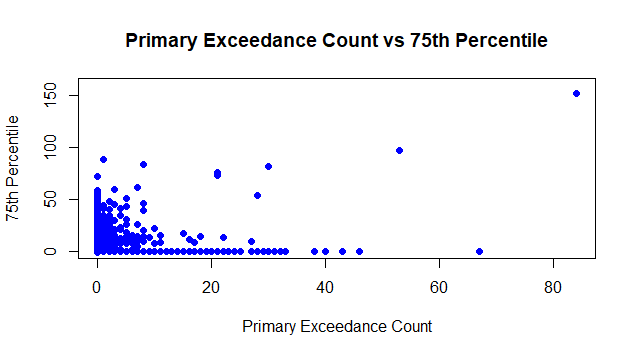
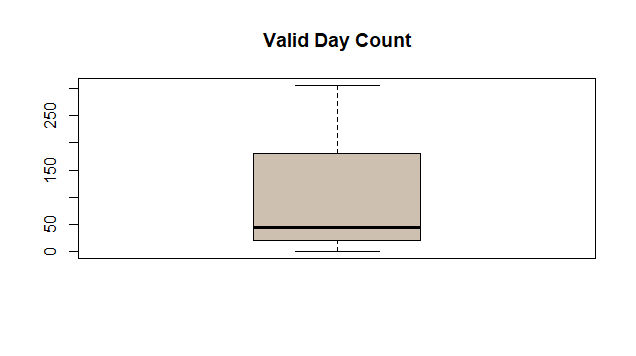
**Title: XN Project: Digging Deeper**

**ALY 6080 – Integrated Experiential Learning**

**Prof. Atherley, Valerie**

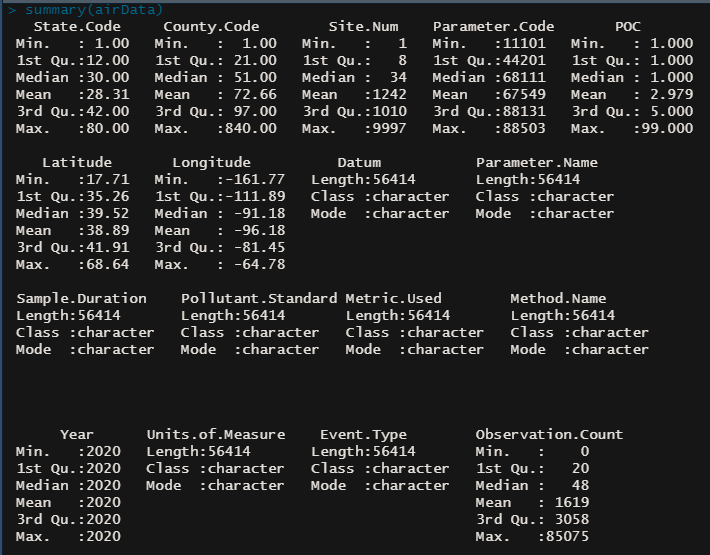
In this Assignment, I will be digging deeper into the Annual Conc. By monitor in 2020 Dataset from the EPA website where I took the monitor dataset which has 56514 rows with 55 parameters. In this, each parameter is measured at a site. It is used to keep track of the 24 hours activities and monitors accordingly. The country, state, latitude, longitude and FIPS codes, county, parameter name are a few of the attributes of this dataset. Engage analytical team members in the process when trying to understand the results. The better your team members understand what you want to get out of your results, the better they can advise you on the best data analysis strategy.

Before utilizing this dataset, we have to clean, structure, and transform it into useful data and do some exploratory data analysis on top of it. To perform this, we can use R/ Python and scrutinize the dataset. Exploratory Data Analysis (EDA) is needed to detect them, as it examines the data for unexpected findings that can have a significant effect on the results of our analysis and statistical modeling. Data planning guarantees data integrity, resulting in accurate observations. Insights would likely be not good if the data is not prepared well. EDA is, at its heart, a creative operation. And, as in most innovative processes, the trick to answering high-quality questions is to ask a lot of them. Since you don't know what insights are found in your dataset, it's difficult to ask revealing questions at the start of your study.



The above 2 plots show the basic EDA on the Valid Day Count and Primary Exceedance Count vs 75th Percentile. This is referred to as data wrangling/preparation, and it is an important aspect of the Project. The data that I got to work on won't be suitable for the research right away, it will need to be manipulated and organized. Many a time, we will be sorting one variable, but we can sort several variables by specifying one key parameter and then the others. Missing data is also known as a data that is not available for a column/ row in the given dataset. It’s encoded as NaNs. In coming days, I will be working on the Inferential Statistics, deep analysis, modelling the data, working on the data science algorithms like Decision tree, random forest, and others. I will completing these tasks by allocating time and segregate the work among my team.

The process of collecting data and expressing it in a summary form to conduct statistical analysis based on specific variables is known as data aggregation. It is difficult to determine when data transformations are required, and it is even more difficult to specify the type of transformation required. One of the most critical aspects of data preparation is data transformation, which includes the development of new record fields from existing values in the dataset. Give a brief overview of your characteristics and group them into categories. This will have a major impact on the visualizations and statistical methods you use. Visualize the distribution of the data to get a better understanding of it. This collaborative process is refined by gaining a better understanding of the data and its applications. In the below screenshot, we can see the basic summary of the dataset which gives us the Min, Max, Mean, Median, 1st Q, and 3rd Quartile of each column.



Visualize possible relationships between exposure and outcome Identify outliers and anomalies. Agile project management is a very useful method for process development teams who want to demonstrate value early and integrate continuous input from executives and clinicians as they move from iteration to iteration because it focuses on frequent contact, eliminating barriers and meeting commitments within defined timeframes.

We'll conduct exploratory data analysis in this manner. EDA (Exploratory Data Analysis) allows one to see beyond the numbers. The more we delve into the details, the more insights we gain. We will spend nearly 80% of our time using EDA to interpret data and solve different business problems. Going further, I will be analyzing a variety of columns. My preferred way of showcasing results are through the Visualization Dashboard.

**Reference:**

[1] Ronald K. Pearson, Exploratory Data Analysis using R was retrieved from www.ru.ac.bd/wp-content/uploads/sites/25/2019/03/102\_05\_02\_Pearson\_Exploratory-data-analysis-using-R-2018.pdf