**CNCITY EDA – Insights and Takeaways**

**Intro:**

**What questions does this data analysis attempt to answer?**

* Predicting the pressure\_1 and pressure\_2 values
* Finding the correlation between different variables (pressure\_2 and is\_maintenance, outside\_temperature and pressure\_1, etc)
* Seasonality and trend

**Brainstorming:**

* **Preprocessing:** Grouped the values in the ‘time’ column by hour using ‘max’ as the aggregation method
* Sometimes ‘max’ can skew the distribution of the dataset
* Better to process the columns separately using different aggregation methods depending on the nature of each variable

1. Data from April (4/1/2024 ~ 4/29/2024)
   1. A graph with blue lines

      Description automatically generated🡪 tells us ABSOLUTELY nothing about ‘pressure\_2’
   2. A graph with blue lines

      Description automatically generated
      1. The vertical ‘traces’, or lines, indicate that the data has been compressed due to incorrect preprocessing
   3. Converted the values in the ‘hour’ column into datetime objects then concatenated ‘hour’ and ‘day’ to make a new column called ‘time’
      1. Values in the ‘time’ column are of the form: 2024-04-01 0:00:00, 2024-04-01 1:00:00,…
      2. A graph with blue lines

         Description automatically generated
2. January through June
   1. A graph with blue lines

      Description automatically generated
3. Autocorrelation plots (Jan, Feb, March, April, May, June)
   1. A graph with blue dots and numbers

      Description automatically generated
4. Conclusion/Bottom Line
   1. Building a linear regression model to predict the pressure\_2 values