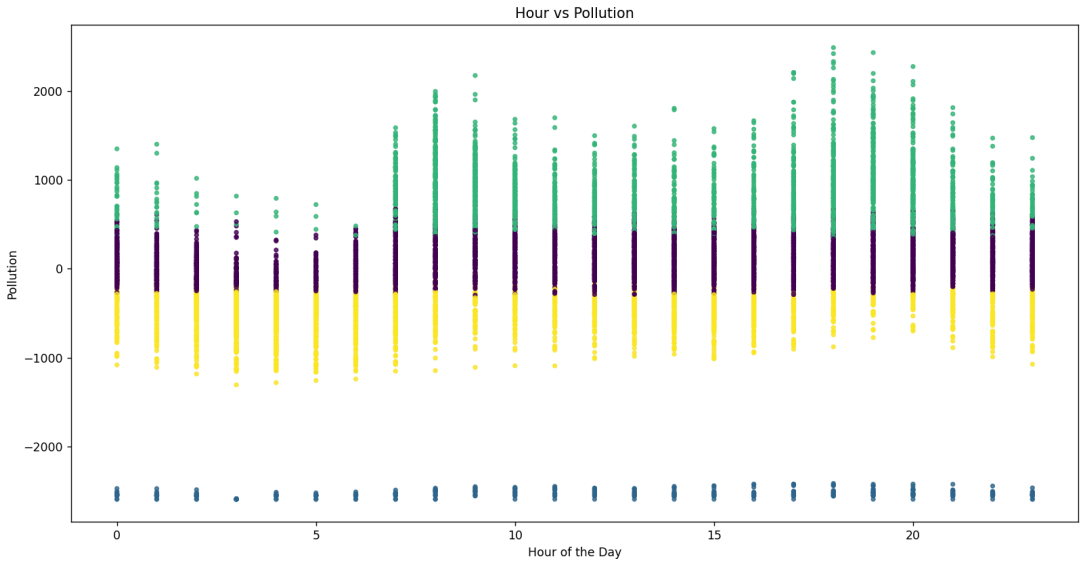
The data used is Air Quality, from https://archive.ics.uci.edu/dataset/360/air+quality. I obtained it from the UCI Machine Learning Repository outlined in the assignment tools. I downloaded the AirQualityUCI.csv file from the attached link and read it as a pandas dataframe in my Python script. It is interesting because it shows hourly air quality levels in an Italian city based on the presence of various harmful chemicals in the air. This is very useful training information for machine learning as it can be used to feed models predicting air quality levels. In this case, I chose to cluster groups by air quality by the hour. This helps determine what time has better, and therefore safer, air quality which is important information for human health awareness.

After importing necessary libraries, I read the CSV file using comma as a delimiter, as all features were stuck together into 1 feature due to being separated by commas. After separating them properly, I had to drop the last 2 features, as they were unnamed and had no values stored but NaN, and according to the description of the dataset it only had 15 features while these obsolete features made the total add up to 17, so I simply removed them as they are not supposed to be there. I made a new feature, Hour, by removing the minutes from the Time feature. I then used this to measure pollution levels by the hour. Another step of preprocessing was replacing commas with periods in decimal values. For certain features related to specific chemicals, a comma was used in place of a period. To do PCA analysis all data must be numerical, so I changed this to fit that requirement. I then performed PCA analysis and split the data into 5 components, as there are many types of chemicals present and I did not want to have the data skewed by unbalanced levels among them. I determined that the first component was the most accurate representation of overall pollution level across all chemical features. I consolidated these again with the Hour feature, then applied K-means clustering to the data with 4 clusters and

plotted the data using the Hour feature as the X-axis and the principal component, representing pollution level, as the Y-axis.



This was the resulting plot. Due to the creation of the 'Hour' feature being discrete, all values are present at clear hourly intervals on the X-axis, while the pollution level is continuous. The 'Hour' feature is based on 24 hour time (Hour 8 as 8AM and Hour 20 as 8PM). It is clear that certain hours, such as Hour 6 (6AM) have less instances in the set as there are less points at that time. The 4 clusters are clearly labeled by colors, blue for minimal pollution levels, yellow for relatively low pollution, purple for average levels, and green for high levels.

The conclusion that can be drawn from the results is an understanding of which hours have higher and lower average pollution. Hours 7-11 experience a spike in air pollution, especially at 8 and 9. After that point, the average pollution remains quite high until there is another spike between hours 17-19, the peak pollution levels. Average pollution levels drop significantly after hour 20, and are at their lowest points between hours 3-6. One can infer the higher pollution levels may be from the peak traffic hours; the times where many people go to work in the morning and return from work in the evening, and the lowest pollution levels are due

to the low traffic. People are less likely to be driving, smoking, or doing any kind of activity outside in the middle of the night or very early in the morning.

References

Vito, S. (2008). Air Quality [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C59K5F.

GeeksforGeeks. (2024, August 29). *K means clustering - introduction*. GeeksforGeeks. https://www.geeksforgeeks.org/k-means-clustering-introduction/