Nonhlanhla Mazibuko

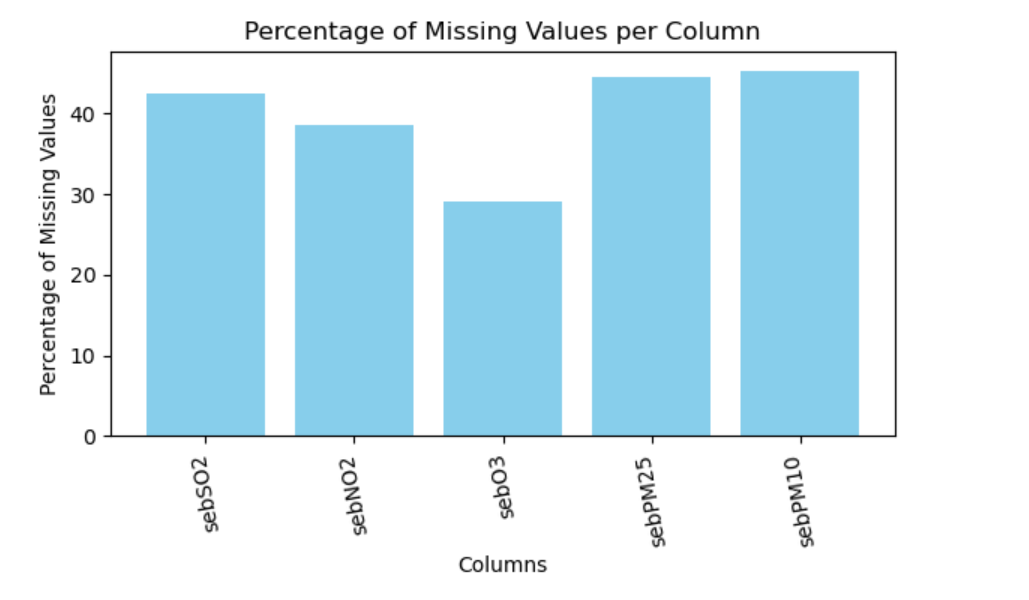
1889514

Health Analytics Practical report 1

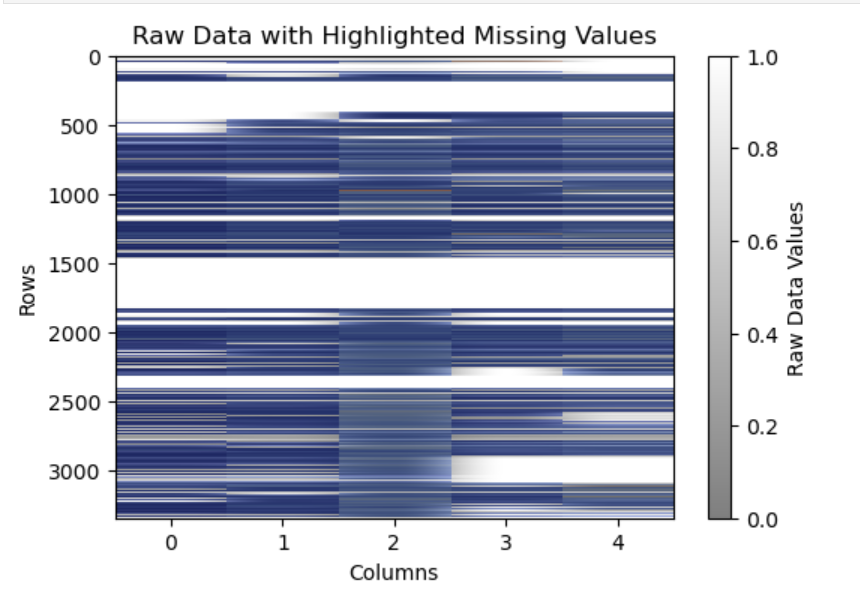
In health analytics, data is used as a tool for extracting valuable insights from different healthcare data sources, and facilitating informed decision-making in the healthcare sector. Healthcare professionals use these insights to accurately diagnose conditions and implement effective treatment plans and interventions for their patients. As healthcare systems increasingly use analytics to improve their healthcare outcomes and operations, the need for high quality data that is complete and accurate has become very important. Missing data poses a significant challenge within clinical settings, leading to potentially dire consequences, such as misdiagnoses, inappropriate treatments, and compromised patient safety. There are multiple ways in which missing data can be handled, one way is to remove the missing data and another would be to impute the missing data. This report will focus on using a multivariate imputation on missing values, and analyzing how the imputation affected the analysis of the results.

Multivariate imputation is

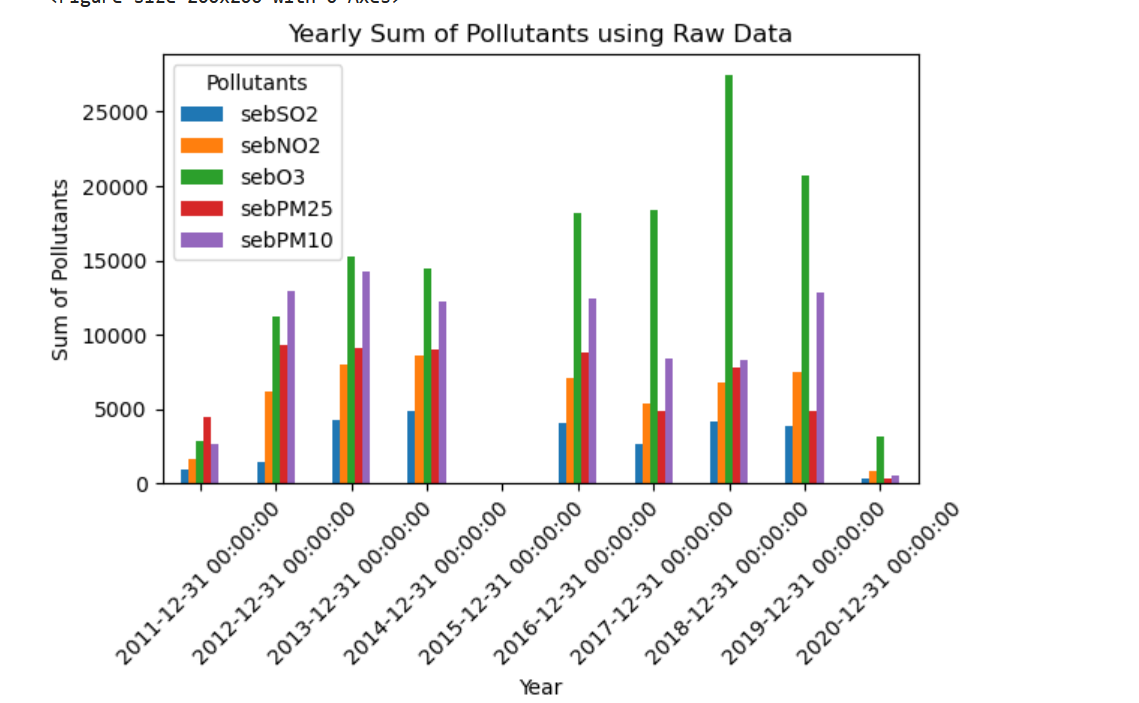
The assumption made was that the missing data is *Missing at Random (MAR)*, meaning the probability of missingness is related to the observed data but not the unobserved data. To address this, we employed the K-Nearest Neighbors Imputer (KNNImputer) from the sk-learn library to fill in the missing values, ensuring the dataset could be used for further analysis without significant gaps that might produce inaccurate results.



The dataset includes pollutant measurements, which are sebSO2, sebNO2, sebO3, sebPM25 and sebPM10. A summary of the data was obtained using basic descriptive statistics. These provided insights into the distribution of values, with means, medians, and standard deviations calculated for each pollutant. The data indicated significant levels of missing values across all pollutants, with the highest missing percentage for sebPM10 (45.38%) and the lowest for sebO3 (29.12%).

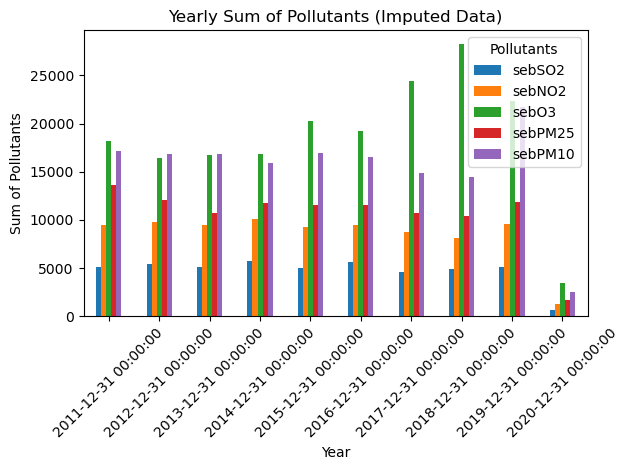


A heatmap was created to visualize the missing data across the dataset. This helped in visually identifying patterns of missingness in the dataset to visualize how much of the dataset was incomplete. The heatmap showed too much missing data, indicated by the white spaces in the heatmap, particularly between the fourth year and the sixth year for all pollutants. This irregularity suggested that an imputation technique would be required for this analysis in order to avoid based results. A date column from January 2011 to February 2020, spanning 10 years was added in the dataset. This column was to enable the creation of a time series for the data.



Yearly sums of pollutants were calculated and plotted from the raw dataset. This was done in order to compare the visualizations of data before and after imputation. The resulting plot revealed irregular patterns in the sum of pollutants throughout the years. In the fifth year, the yearly sums bars are empty because of the missing data for that whole year. The pollutants across the years varied significantly and for most pollutants the yearly sums were smaller. These results could have been the reflection of the data gaps rather than the true trends and patterns of the pollutants.

The missing data was then imputed using the KNNImputer, which works by finding the nearest neighbors (with n\_neighbors = 5) based on feature similarity and using their values to estimate the missing values. The advantage of using KNN is that it retains variability in the data by considering multiple features for imputation, making it well-suited for datasets with spatial or temporal correlations.



The yearly sums were then recalculated using the imputed data and a new time series was then plotted. This plot shows that after imputation there is a general trend of the yearly sums being constant for each pollutant across the years, and all the pollutants have almost the same yearly sums. This suggests that KNNImputation successfully mitigated the gaps, providing more reliable estimates of pollutant levels.

while the KNNImputer provides a more complete dataset, it does not reconstruct the actual missing data. Instead, it gives an approximation based on existing data patterns. Therefore, conclusions drawn from this imputed data should still be interpreted with caution mostly for datasets that do not follow trends captured by nearest neighbors.