**Data imputation health Analytics assignment**

**Introduction**

This report outlines the analysis of air pollutant data for the Sebokeng area from January 2011 to February 2020.The datasets were downloaded from the South African Air Quality Information System (Saaqis (environment.gov.za)) under the Department of Forestry and Fisheries Environment, (DFFE).

Missing values were present across all variables and thus the analysis will be done by addressing the missing data through imputation. The datasets were incomplete and required cleaning, we will be using a multivariate method of imputation on the ‘Sebokeng\_data’ dataset to handle the missing values.

The dataset and Pollutants will be plotted as a time series before and after imputation. This will allow for a better understanding of the environment and health conditions in Sebokeng.

**Data variables**

The dataset contains five variables which are important for evaluating the air quality and thus the environmental and health conditions in Sebokeng.

The five variables mainly being, Sulphur Dioxide(SO2), Nitrogen Dioxide(NO2) ,Ozone(O3), Particulate Matter <2.5(PM2.5 ) and Particulate Matter (PM10) .

**Data cleaning**

The dataset was incomplete and contained several missing values. The dataset was cleaned by removing or dropping rows with all values missing. Missing values could interfere with the identification of trends and the assessment of health risks in Sebokeng

**Descriptive statistics**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **VARIABLE** | **sebSO2** | **sebNO2** | **sebO3** | **sebPM25** | **sebPM10** |
| **count** | 1925.000000 | 2055.000000 | 2371.000000 | 1853.000000 | 1827.000000 |
| **mean** | 13.877048 | 25.350595 | 55.590021 | 31.673353 | 46.370742 |
| **std** | 11.669311 | 10.510577 | 24.103645 | 18.867618 | 24.009341 |
| **min** | 0.190387 | 0.000000 | 0.000000 | 3.634714 | 2.460333 |
| **25%** | 6.187676 | 18.360628 | 37.442500 | 19.661917 | 30.136500 |
| **50%** | 10.360277 | 23.416105 | 53.699667 | 27.742417 | 41.075083 |
| **75%** | 17.229557 | 30.367969 | 71.014333 | 38.294458 | 56.197313 |
| **max** | 115.727693 | 115.438737 | 199.584250 | 288.647042 | 194.949870 |

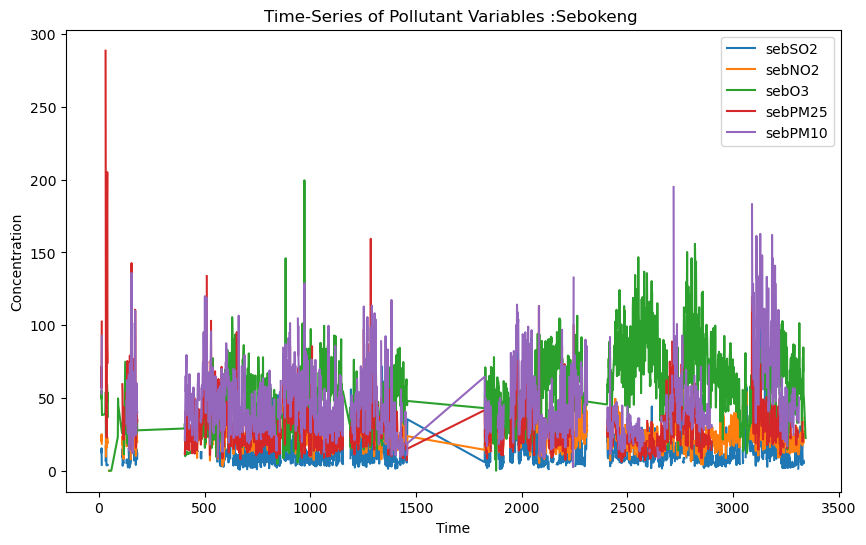
**State the percentage of missing data**

The percentages show the completeness of the dataset for each variable, this is important for determining the quality and reliability of the findings derived from the data

|  |  |  |  |
| --- | --- | --- | --- |
| |  |  | | --- | --- | | **Variable** |  | | **Percentage Missing** |
| **sebSO2** | 20.091324 |
| **sebNO2** | 14.694894 |
| **sebO3** | 1.577418 |
| **sebPM25** | 23.080116 |
| **sebPM10** | 24.159402 |

**Time-series of dataset (Visualisation)**

After cleaning a time series was done for each pollutant.



The time series shows gaps in the time series highlighting the importance to imputate , This was further substantiated by the high percentage of missing data for the variables sebSO2- 20.091324 , sebPM25 -23.080116 and sebPM10 -24.159402 respectfully

**Multivariate imputation**

Imputing the missing values was important because the large gaps in the time series data could distort the modelling or analysis

Multivariate Imputation was done using the IterativeImputer from sklearn. MICE (Multiple Imputation by Chained Equations) is a method for handling missing data in which each missing value is modelled by using the other variables in the dataset. This is a simple yet effective method, it maintains the overall trend and continuity over time

**Descriptive statistics (Imputed)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **sebSO2** | **sebNO2** | **sebO3** | **sebPM25** | **sebPM10** |
| **count** | 2409.000000 | 2409.000000 | 2409.000000 | 2409.000000 | 2409.000000 |
| **mean** | 13.997828 | 25.547520 | 55.631861 | 31.826846 | 46.181893 |
| **std** | 10.658141 | 9.895108 | 23.941968 | 17.256416 | 22.531673 |
| **min** | -1.639424 | 0.000000 | 0.000000 | 3.634714 | 2.460333 |
| **25%** | 6.838309 | 19.173572 | 37.734417 | 21.276133 | 32.062847 |
| **50%** | 11.683453 | 24.299078 | 53.783500 | 29.186167 | 42.902042 |
| **75%** | 17.304524 | 29.960855 | 70.785583 | 37.111542 | 52.795652 |
| **max** | 115.727693 | 115.438737 | 199.584250 | 288.647042 | 273.435889 |

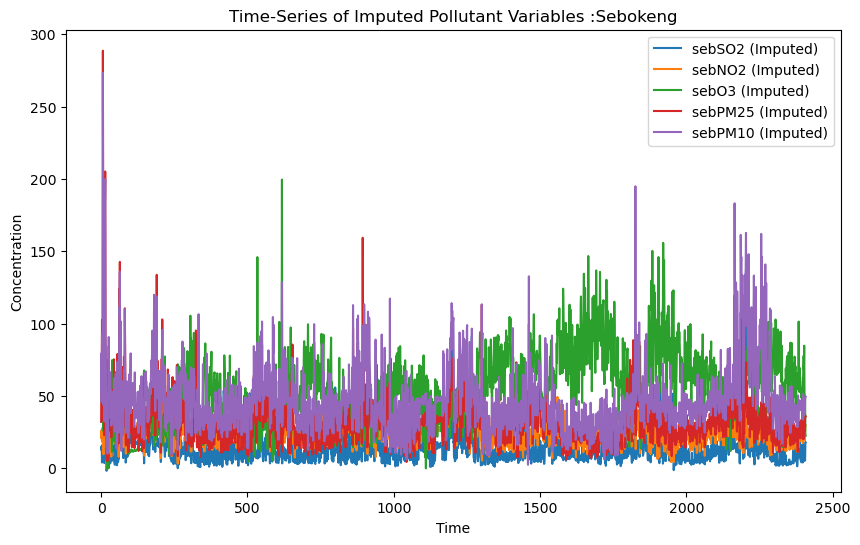
The imputed Descriptive statistics after imputation show minimal difference, this indicates that the imputation filled in the missing data while maintaining the integrity of the dataset.

The max values for the variables are significantly higher than their mean and median which suggests high pollution environments in the area.

The mean and median Concentrations of Ozone- O3 and Particulate Matter PM10- are high indicating low air Quality in Sebokeng

**Time-series of dataset (Visualisation)**

Time series plots were generated before and after imputation.



The time gaps that were evident in between time (0-500 and 1000-1500 ) in the time series prior to imputation are filled

The visualisation indicate that the imputation process effectively filled in the missing data while maintaining the integrity and continuity of the original time series patterns.

**Conclusion**

The analysis of the Sebokeng Dataset included the handling of missing data through imputation to ensure accurate time series analysis. MICE capture more complex relationships in the data. The resulting imputed time series was complete and continuous to the dataset. The dataset show and high pollution and low air Quality environments in Sebokeng. This negatively impacts the health outcomes of the Area. The increased pollutant environments can be attributed to by the increased industrialization and firms found in the area