CHAPTER 4

INITIAL RESULT

4.1 Exploratory Data Analysis (EDA)

4.1.1 Primary Data

1. Visualizations: The visualization of the primary data in order to get a feel and an idea of the underlying data of the research. Through the Microsoft Power BI, line charts, bar charts, and scatter plots are created out of data sets. Line graphs are especially ideal in demonstrating trends in air pollution over time and bar graphs can be used to compare concentration values of various pollutants across different cities. Categorically, scatter plots are used to determine the correlation between air quality indices and individual health risks and costs, particularly the prevalence of respiratory diseases.
2. Descriptive Statistics: Descriptive statistics brings out the characteristics of the data in terms of numerical value hence giving an insight on the variations within the data from the mean. Here, by utilizing Power BI, measures include Mean, Median, Mode, Standard deviation and Range for both the air quality and health data. For instance, the average concentration of PM2.5 and PM10 pollutants, and the average number of respiratory-related hospital admission, averaged values are calculated and compared.
3. Initial Insights: The visualizations and descriptive statistics give the first understanding of the results obtained in the dataset. For example, an increase in the value of PM2.5 during dry season may be noted, which has been associated with an increase in admission of patient's especially with asthma and other respiratory diseases. Furthermore, it should be noted that regions with a higher level of industrialization could reveal increased levels of pollution and the corresponding levels of cardiovascular diseases.

4. Feature Engineering: In feature engineering, new features are derived from the original data that can affect the models' performance positively. Therefore, in this study we create new features from the available data including moving averages of the pollutants and lagged variables in Python. For example, using the moving average of PM2.5 in the past week can also be a strong predictor of current health status. Lagged variables can be used when prior values of a variable in question have an impact on the present state of health.

4.1.2 Secondary Data

1. Visualizations: Like in the case of primary data, the secondary data is also analysed and represented by Power BI and Python. The secondary data sources comprise the air quality data obtained from the IQAir database beside the record of health of the population available from the Ministry of Health. Tabulations are used to display trends, distributions, and geographical variations in the data and line charts, histograms, and heat maps are used to present them. For instance, there is a heat map, which presents the geographical distribution of pollutants in various cities and clearly shows zones with consistently elevated levels of different pollutants.
2. Descriptive Statistics: Descriptive statistics for secondary data are normally calculated in order to get broad results. Thus, averages of pollutants, standard deviations and interquartile ranges are also calculated. The health data statistics also comprise the average hospital admission rate for respiratory and cardiovascular diseases, stratified by age and gender.
3. Initial Insights: First findings of the secondary research demonstrate relationships and patterns which are consistent or opposite with those acquired from the primary research. For example, the changes of the climate in certain months might be associated with deterioration in the quality of air, and higher incidence of respiratory diseases. These assist to verify the findings arising out of the primary data analysis process and put them into a wider perspective.

4. Feature Engineering: In secondary data analysis, other features are created specifically to provide an improvement to the model. Therefore, using Python, new features like the abnormal levels of the pollutant or the products between some pollutants are built. Anomalies can help in uncovering the trends of pollutant concentrations while the interaction terms detect the overall impact of several pollutants on health consequences.

4.2 Machine Learning

The first machine learning models are trained through Python's Scikit-learn to estimate the air pollution level and potential health impact. Some of the techniques used include; decision tree, predictive random forests and support vector machines (SVM). For example, one type of model, such as the random forest model, may analyse data to estimate the chance of days with high pollution, and another, such as an SVM model, may estimate the possibility of respiratory diseases' hospitalization. By evaluating models on a certain set of data the performance is laid down as a baseline in order to optimize it with further solutions. The obtained feature importance scores help in determining the main predictors while hyperparameters are adjusted to optimize the model. All possible measures are taken to avoid overfitting and to obtain greater generalizability of the models: cross-validation techniques.