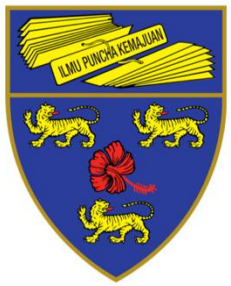


UNIVERSITY OF MALAYA

MASTER OF DATA SCIENCE (SEMESTER 1 – 2023/2024)

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

WQD7005 DATA MINING

GROUP PROJECT

PREDICTION FOR AIR QUALITY INDEX

Group Member	Matric Number	Contribution
Wong Jia Hui (Leader)	S2192852	• Model parts • Assess parts • Report Formating
Low Boon Kiat	17138399	• Sample parts • Explore parts
Jie HongSheng	22064728	• Model parts
Zhao Zihui	S2187551	• Modify parts

Project Presentation Video Link:

https://www.youtube.com/watch?v=q6T8J79DVbo&t=30s

Contents

1 Introduction

In this project, we harness a dataset from the Global Meteorological Data Repository, encompassing a wide array of environmental and meteorological data. This includes temperature, wind speed, air quality index (AQI), and concentrations of various pollutants across a range of geographical regions. Utilizing this dataset is vital for deriving insights through data mining, applying the SEMMA (Sample, Explore, Modify, Model, and Assess) methodology of SAS Enterprise Miner. The sampling phase is particularly critical as it provides a succinct overview of the challenges we aim to address and a comprehensive description of the data. This phase helps in selecting the most appropriate data for our specific project requirements.

During the exploration phase, we conduct an exploratory data analysis (EDA) to explore correlations between variables such as temperature, ozone (O3), carbon dioxide (CO2), nitrogen dioxide (NO2), sulfur dioxide (SO2), PM2.5, PM10, and AQI. By gaining a good understanding of these interrelations, it helps to develop a robust model that can accurately predict AQI based on the selected features. In the data modification phase, we meticulously detail the processes of data cleaning and feature engineering implemented to refine the data for effective modeling. For the modeling phase, we have utilized a variety of machine learning models, including Decision Tree, Neural Network, and Ensemble techniques, to predict the Air Quality Index (AQI). Subsequently, in the assessment phase, we evaluate the performance of these models using various performance metrics, such as the misclassification rate, average squared error, ROC index, and confusion matrix.

The development of an accurate predictive model for the Air Quality Index (AQI) is immensely beneficial. It plays a pivotal role in safeguarding public health, guiding policy formulation, promoting environmental sustainability, and ensuring the overall well-being of communities. Accurate AQI predictions empower stakeholders to take proactive measures in mitigating pollution-related risks and enhancing the quality of life.

2 Business Understanding

2.1 Analysis Goal

The main objective of this project is to perform data sampling and data exploration on the selected dataset to understand the correlations between weather parameters, pollutant's concentration, and Air Quality Index (AQI). By understanding the correlations between those parameters, we can identify the types of data modifications required and select the right features for constructing our model to accurately predict the AQI during the Modify and Modeling phase in SEMMA methodology. In addition, we also aim to explore the distribution and range of variables to identify if there are any outliers through the exploration.

Moreover, a predictive model for AQI offers several significant benefits such assisting government to make better-informed decisions in policy planning and resource allocation, better public health protection from timely alerts, and environmental quality improvement.

2.2 Analysis Data

This dataset contains meteorological and environmental data from various geographical locations, covering the latest 3 months of data records (29/08/2023 – 28/11/2023), inclusive of AQI, variety of weather parameters and pollutants' concentration. Additionally, we reclassify the roles and measurement levels of variables to ensure data accuracy.

During the exploration phase, we will generate comprehensive statistical information of the data, including minimum, maximum, average values, missing values, and standard deviation. We will employ both bivariate and multivariate analysis methods. For example, we have observed the relationship between AQI and various environmental factors such as temperature, cloud cover, precipitation, humidity, etc. We also studied the relationship between pollutants' concentration and AQI. Furthermore, we also plotted a world map to showcase the AQI in different geographical locations, using variables like longitude, latitude, AQI and country name. For identifying the best features affecting the AQI, we created a correlation matrix chart to showcase the correlations between those variables.

In this study, we conduct a thorough analysis of the selected dataset, identifying suitable features for building a predictive AQI model. Ultimately, our goal is to construct a model that predicts AQI, aiding environmental scientists and policymakers in making more informed decisions for environmental management and public health protection.

3 Dataset Description

3.1 Data introduction

This study employed a dataset sourced from Kaggle's Global Meteorological Data Repository which provides daily weather information for capital cities around the world. It contains a comprehensive set of features such as meteorological, weather data, pollutants' concentrations, and Air Quality Index (AQI) from various geographical locations around the world. The dataset initially provided over 40+ features, but we conducted a simple data preprocessing to remove duplicated features due to using different units and measurement systems (Metric System and Imperial System). Therefore, the dataset used in this project after preprocessing comprises 27 features with a total of 17,149 entries, spanning from 28/08/2023 to 29/11/2023.

The complete metadata of the dataset is shown in Table 3.1. A quick overview on the details of the dataset is shown below:

- **Basic Information:**
  - Data is generated from 185 countries with 197 unique locations
  - Number of features: 27 features
  - Period: 28/08/2023 to 29/11/2023 (Updated daily)
- **Meteorological Parameters:**
  - Temperature records range from a minimum of -23.3°C to a maximum of 45.4°C
  - Wind speeds from 0 to 141.1 km/h
  - Atmospheric pressure from 955 to 1053 millibars
  - Precipitation amounts from 0 to 31 mm
  - The dataset also includes other meteorological parameters such as wind degree, cloud cover, 'feels like' temperature, visibility, and UV index.
- **Environmental Pollutants Concentrations:**
  - Pollutant's concentrations for Carbon Monoxide (CO), Ozone (O_3), Nitrogen Dioxide (NO_2) and Sulphur Dioxide (SO_2)
  - Concentrations of PM2.5 and PM10. The concentration ranges for PM2.5 and PM10 are 0.5 to 1558.8 μg/m³ and 0.5 to 2504.3 μg/m³ respectively
  - Air Quality Index (AQI)
- **Additional Information:**
  - Includes information on the moon phase (encompassing 8 different types of moon phases) and moon illumination

The distinct feature of this dataset is its coverage of a diverse array of meteorological and environmental parameters. Therefore, the dataset is applicable to various types of in-depth analysis such as climate analysis, weather prediction, environmental impact, tourism planning, and geographical patterns. However, in this project, we are focusing on the environmental impact where the dataset is used for creating a model to predict AQI. Through this dataset, we will conduct a detailed analysis of how various parameters influence the AQI.

Overall, this dataset provides a rich platform for applying skills in feature engineering, exploratory data analysis, and data organization to predict the Air Quality Index in different regions.

Table 3.1 Variables and descriptions of the dataset.

Variable Name	Data Type	SI units	Description
country	object	-	Country where the weather data is recorded
location_name	object	-	Name of the specific location （city） within the country
latitude	float64	-	Geographical latitude coordinate of the location
longitude	float64	-	Geographical longitude coordinate of the location
timezone	object	-	Timezone pertaining to the location
last_updated	datetime64 [ns]	-	Last update time of the data, recorded as a datetime
temperature_celsius	float64	°C	Temperature measured in degrees Celsius
condition_text	object	-	Description of the weather conditions
wind_kph	float64	km/h	Wind speed measured in kilometers per hour
wind_degree	int64	°	Direction of the wind in degrees
pressure_mb	int64	mbar	Atmospheric pressure measured in millibars
precip_mm	float64	mm	Precipitation amounts in millimeters
humidity	int64	%	Relative humidity percentage
cloud	int64	%	Percentage of cloud cover
feels_like_celsius	float64	°C	'Feels like' temperature in degrees Celsius
visibility_km	float64	km	Visibility distance in kilometers
uv_index	int64	-	Ultraviolet index
gust_kph	float64	km/h	Speed of wind gusts in kilometers per hour
air_quality_ Carbon_Monoxide	float64	μg/m³	Concentration of Carbon Monoxide
air_quality_Ozone	float64	μg/m³	Concentration of Ozone
air_quality_ Nitrogen_dioxide	float64	μg/m³	Concentration of Nitrogen Dioxide

Table 3.1: Variables and Descriptions of the dataset, Cont.

Variable Name	Data Type	SI units	Description
air_quality_Sulphur_dioxide	float64	µg/m³	Concentration of Sulphur Dioxide
air_quality_PM2.5	float64	µg/m³	Concentration of PM2.5 particles
air_quality_PM10	float64	µg/m³	Concentration of PM10 particles
air_quality_gb-defra-index	int64	-	Air quality index as per DEFRA standards
moon_phase	object	-	Current phase of the moon
moon_illumination	int64	-	Percentage of moon illumination

3.2 Air Quality Index (AQI) – DEFRA Standard

Within this project, our focus lies in understanding the associations among different variables and the Air Quality Index (AQI) as per DEFRA standards. Therefore, we would like to present a concise introduction to the AQI, enhancing the reader's understanding of this concept. AQI as per Department for Environment, Food & Rural Affairs (DEFRA) standards is defined as the air quality standards set by DEFRA to monitor air pollution levels, and thus protecting the public health and the environment. According to DEFRA's official page, the AQI provides information on the levels of air pollution and offers suggested measures and health guidance. The AQI has a range from 1 to 10, which is categorized into four bands, starting from low (1) to very high (10) as shown in Table 3.2.

Table 3.2 An Overview and descriptions of AQI as per DEFRA standards.

Air Pollution Banding	AQI	Accompanying health messages for at-risk individuals	Accompanying health messages for the general population
Low	1 – 3	Enjoy usual outdoor activities	Enjoy usual outdoor activities
Moderate	4 – 6	Individuals with lung/heart issues should consider reducing outdoor physical activities, especially if they experience symptoms.	Enjoy usual outdoor activities
High	7 – 9	Individuals with lung/heart issues or elderly should reduce their outdoor physical activities, especially if they experience symptoms. Those with asthma may find they need to use their reliever inhaler more frequently.	Individuals who experiencing discomfort should consider reducing outdoor activities.
Very High	10	Individuals with lung/heart issues or elderly should avoid outdoor physical activities. Those with asthma may find they need to use their reliever inhaler more frequently.	Reduce physical outdoor activities, especially if experience symptoms such as cough or sore throat.

4 Methodology

4.1 SEMMA Description

SEMMA is a data mining framework proposed by the SAS Institute to provide a structured and intuitive pathway for data analysis. This framework enables users to effectively extract valuable insights from large volumes of data. It mainly consists of five key steps: Sampling, Exploration, Modification, Modeling, and Evaluation. Data scientists can follow a clear workflow under the SEMMA framework, which not only enhances the efficiency of data analysis but also ensures the reliability and accuracy of the results.

The SEMMA framework is widely used across various industries and fields, such as market analysis, customer behavior prediction, and product recommendation systems. It serves as a powerful tool for addressing complex business and research challenges. By implementing SEMMA in SAS Enterprise Miner, data scientists can systematically manage the entire data mining process, thereby better uncovering the latent value of the data.

4.2 SEMMA Flow Chart

The full SEMMA process flow is shown in Figure 4.1.

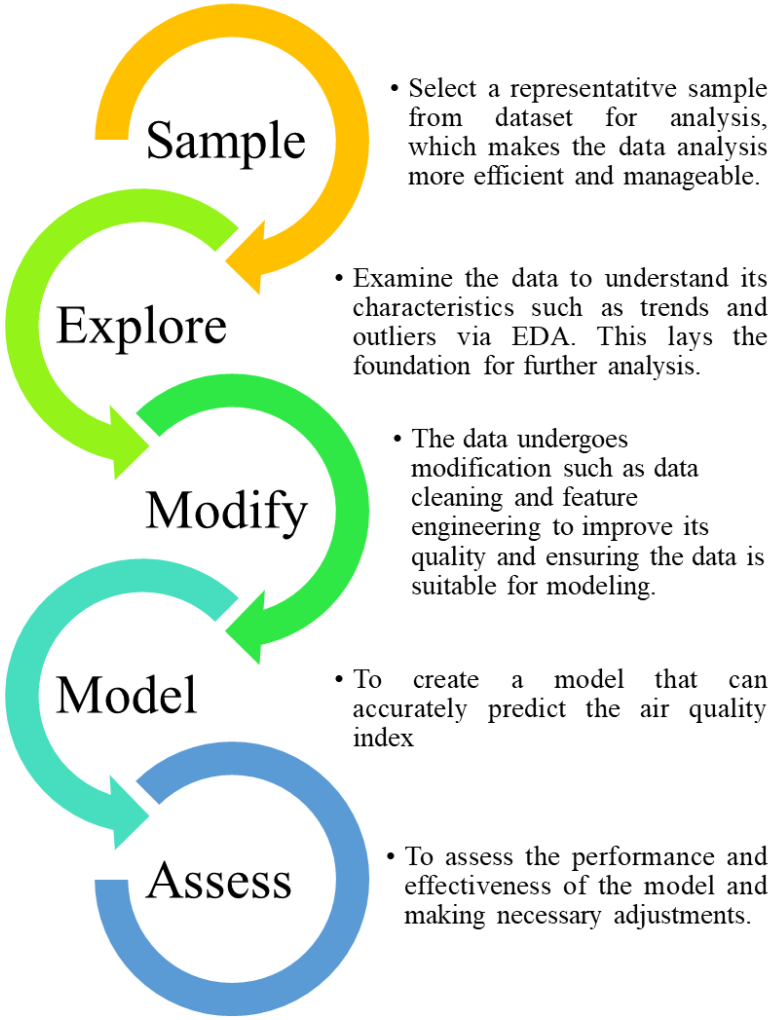


Figure 4.1 SEMMA process flow chart.

5 Sample

5.1 Basic Data Cleaning

The dataset was downloaded from Kaggle as CSV file and imported into Talend Data Preparation for basic data cleaning. The original dataset contains 41 columns. Among them, 14 columns were dropped because some of them are duplicated columns with different units of measurement and do not provide additional information. Additionally, some columns are not important for air quality prediction such as sunrise, sunset, moonrise and moonset.

Table 5.1 List of columns dropped from original dataset.

	Column Name	Description	Dropped?
1	country	Country of the weather data	
2	location_name	Name of the location (city)	
3	latitude	Latitude coordinate of the location	
4	longitude	Longitude coordinate of the location	
5	timezone	Timezone of the location	
6	last_updated_epoch	Unix timestamp of the last data update	/
7	last_updated	Local time of the last data update	
8	temperature_celsius	Temperature in degrees Celsius	
9	temperature_fahrenheit	Temperature in degrees Fahrenheit	/
10	condition_text	Weather condition description	
11	wind_mph	Wind speed in miles per hour	/
12	wind_kph	Wind speed in kilometers per hour	
13	wind_degree	Wind direction in degrees	
14	wind_direction	Wind direction as a 16-point compass	/
15	pressure_mb	Pressure in millibars	
16	pressure_in	Pressure in inches	/
17	precip_mm	Precipitation amount in millimeters	
18	precip_in	Precipitation amount in inches	/
19	humidity	Humidity as a percentage	
20	cloud	Cloud cover as a percentage	
21	feels_like_celsius	Feels-like temperature in Celsius	
22	feels_like_fahrenheit	Feels-like temperature in Fahrenheit	/
23	visibility_km	Visibility in kilometers	

Table 5.1: List of columns dropped from original dataset, Cont.

	Column Name	Description	Dropped?
24	visibility_miles	Visibility in miles	/
25	uv_index	UV Index	
26	gust_mph	Wind gust in miles per hour	/
27	gust_kph	Wind gust in kilometers per hour	
28	air_quality_Carbon_Monoxide	Air quality measurement: Carbon Monoxide	
29	air_quality_Ozone	Air quality measurement: Ozone	
30	air_quality_Nitrogen_dioxide	Air quality measurement: Nitrogen Dioxide	
31	air_quality_Sulphur_dioxide	Air quality measurement: Sulphur Dioxide	
32	air_quality_PM2.5	Air quality measurement: PM2.5	
33	air_quality_PM10	Air quality measurement: PM10	
34	air_quality_us-epa-index	Air quality measurement: US EPA Index	/
35	air_quality_gb-defra-index	Air quality measurement: GB DEFRA Index	
36	sunrise	Local time of sunrise	/
37	sunset	Local time of sunset	/
38	moonrise	Local time of moonrise	/
39	moonset	Local time of moonset	/
40	moon_phase	Current moon phase	
41	moon_illumination	Moon illumination percentage	

We re-categorized the values in the “condition_text” column. The purpose is to group similar values together and make it easier to analyze data in the Explore phase.

Table 5.2: Comparison between new and old categories in "condition_text" column.

New Category	Old Category
Blizzard	Blizzard
Clear	Clear
Cloudy	Cloudy
Fog	Fog, Freezing fog
Heavy rain	Heavy rain, Heavy rain at times, Moderate or heavy rain shower, Moderate or heavy rain with thunder, Torrential rain shower
Drizzling	Light drizzle, Patchy light drizzling

Table 5.2 Comparison between new and old categories in "condition_text" column, Cont.

New Category	Old Category
Snow	• Heavy snow • Light snow • Light snow showers • Moderate or heavy snow • Moderate or heavy snow with thunders • Moderate snow • Patchy light snow • Patchy moderate snow
Rain	• Light freezing rain • Light rain • Light rain shower • Moderate rain • Moderate rain at times • Patchy light rain • Patchy light rain with thunder • Patchy rain possible
Sleet	• Light sleet • Moderate or heavy sleet • Possible sleet possible
Mist	Mist
Overcast	Overcast
Partly cloudy	Partly cloudy
Sunny	Sunny
Thundery outbreaks possible	Thundery outbreaks possible

We corrected the spelling error on Cameroon's country name because 5 values in the "country" column were misspelled as "Cameron" in the dataset (noisy data). Additionally, we changed the timezone format for Malaysia from "Kuala Lumpur" to "Asia/Kuala Lumpur" because the timezone format for most countries in the dataset are region/city (inconsistent data). Subsequently, the dataset was exported from Talend Data Preparation as CSV file. The CSV file was imported to SAS Enterprise Miner and saved as a SAS file for data exploration. After basic data cleaning, the dataset contains a total of 27 variables and 17149 observations.

Table 5.3 List of variables in dataset after basic data cleaning in Talend Data Preparation.

	Variable Name	Description
1	country	Country of the weather data
2	location_name	Name of the location (city)
3	latitude	Latitude coordinate of the location
4	longitude	Longitude coordinate of the location

Table 5.3 List of variables in dataset after data cleaning in Talend Data Preparation, Cont.

	Variable Name	Description
5	timezone	Timezone of the location
6	last_updated	Local time of the last data update
7	temperature_celsius	Temperature in degrees Celsius
8	condition_text	Weather condition description
9	wind_kph	Wind speed in kilometers per hour
10	wind_degree	Wind direction in degrees
11	pressure_mb	Pressure in millibars
12	precip_mm	Precipitation amount in millimeters
13	humidity	Humidity as a percentage
14	cloud	Cloud cover as a percentage
15	feels_like_celsius	Feels-like temperature in Celsius
16	visibility_km	Visibility in kilometers
17	uv_index	UV Index
18	gust_kph	Wind gust in kilometers per hour
19	air_quality_Carbon_Monoxide	Air quality measurement: Carbon Monoxide
20	air_quality_Ozone	Air quality measurement: Ozone
21	air_quality_Nitrogen_dioxide	Air quality measurement: Nitrogen Dioxide
22	air_quality_Sulphur_dioxide	Air quality measurement: Sulphur Dioxide
23	air_quality_PM2.5	Air quality measurement: PM2.5
24	air_quality_PM10	Air quality measurement: PM10
25	air_quality_gb-defra-index	Air quality measurement: GB DEFRA Index
26	moon_phase	Current moon phase
27	moon_illumination	Moon illumination percentage

5.2 Sampling Strategy

Generally, the Sample phase involves selecting a representative, right-volume subset of data from a large dataset, making it manageable and efficient to analyze without losing the essential characteristics of the original dataset. In this project, we will use the entire dataset for Explore phase followed by using samples of varying sizes for Model phase due to the following reasons.

- Our dataset contains a total of 17149 observations. The size of our dataset is relatively small when compared to the one depicted in the example report with 1048575 observations. Additionally, the example report used the entire dataset. Given the size of our dataset is small, it should be manageable to use the entire dataset.

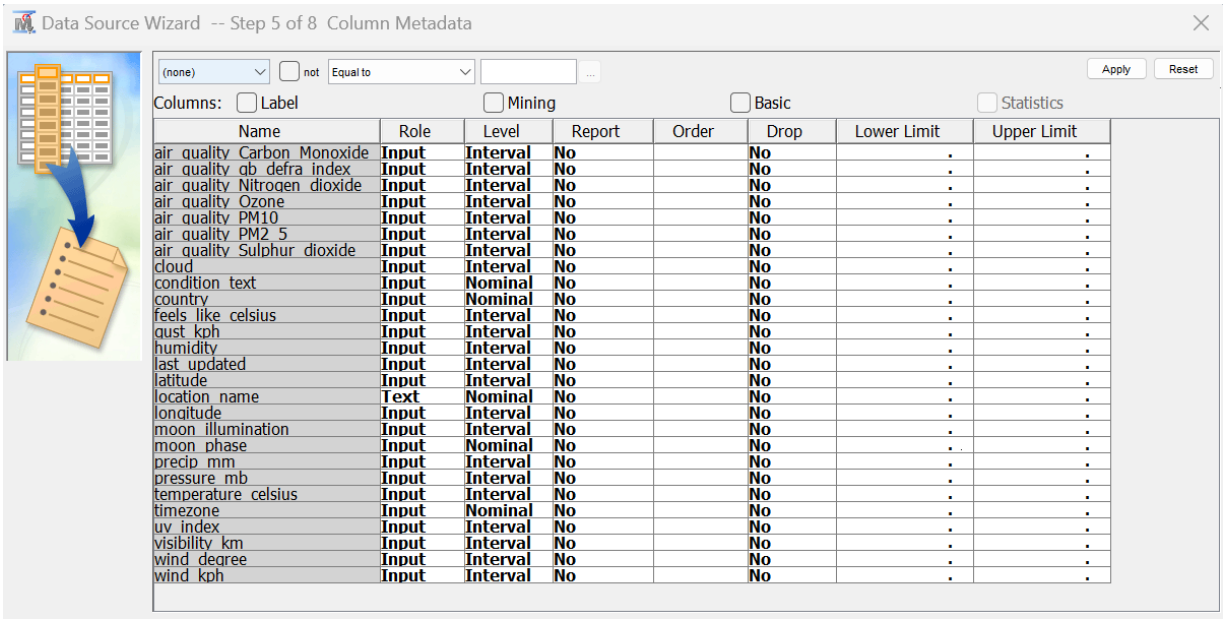
- The Explore phase involves conducting univariate and multivariate analysis to study interconnected relationships between data elements and to identify gaps in the data. According to Applied Analytics Using SAS Enterprise Miner E-book, it is wise to examine the entire dataset if our goal is to examine the data for potential problems. This phase can help determine if the data quality and sample size of our dataset is good enough for model building. This justifies why we selected the entire dataset for the Explore phase.
- In the Model phase, a very large sample size will increase the model complexity and the time required to build it. On the other hand, a very small sample size can lead to a less accurate model although it can be built more quickly and with less complexity.

Our sampling strategy for Model phase is to utilize the entire dataset for the initial modeling phase. This allows us to maximize the amount of data available for training the model. If the modeling outcome is unsatisfactory, we plan to return to the Sample phase for potential resampling. Here is our consideration for the sampling plan. In terms of sampling method, we will perform proportional stratified sampling, whereby the proportion of records in each stratum is the same in the sample as it is in the population. The target variable "air_quality_gb-defra-index" is a categorical variable with 10 levels: 1 indicating good air quality and 10 indicating bad air quality. By stratifying on the target variable, we maintain the same proportion of records in each level that are present in the entire dataset within the samples. Otherwise, we might obtain a sample with very few or even no records of any Air Quality Index levels. For example, there might be no records of level 1 Air Quality Index in the sample if simple random sampling was used. Proportional stratified sampling is appropriate in this case because the levels of categorical data could be easily under- or over-represented if simple random sampling was used. In SAS Enterprise Miner, the Sample node enables us to easily extract a sample from the input data source.

To sum up, the decision to use the entire dataset for the Explore phase was driven by the small size of our dataset, coupled with our intention to thoroughly examine the data for any potential data quality issues by leveraging the full dataset for a comprehensive exploration. For Model phase, we ensure that our model is initially trained on the full dataset to maximize the use of available data, and if necessary, subsequent sampling with proportional stratification will be conducted in a way that maintains the representativeness of the target variable across different levels. This approach enhances the robustness and reliability of our modeling process, particularly for handling a categorical target variable with multiple levels.

5.3 Metadata

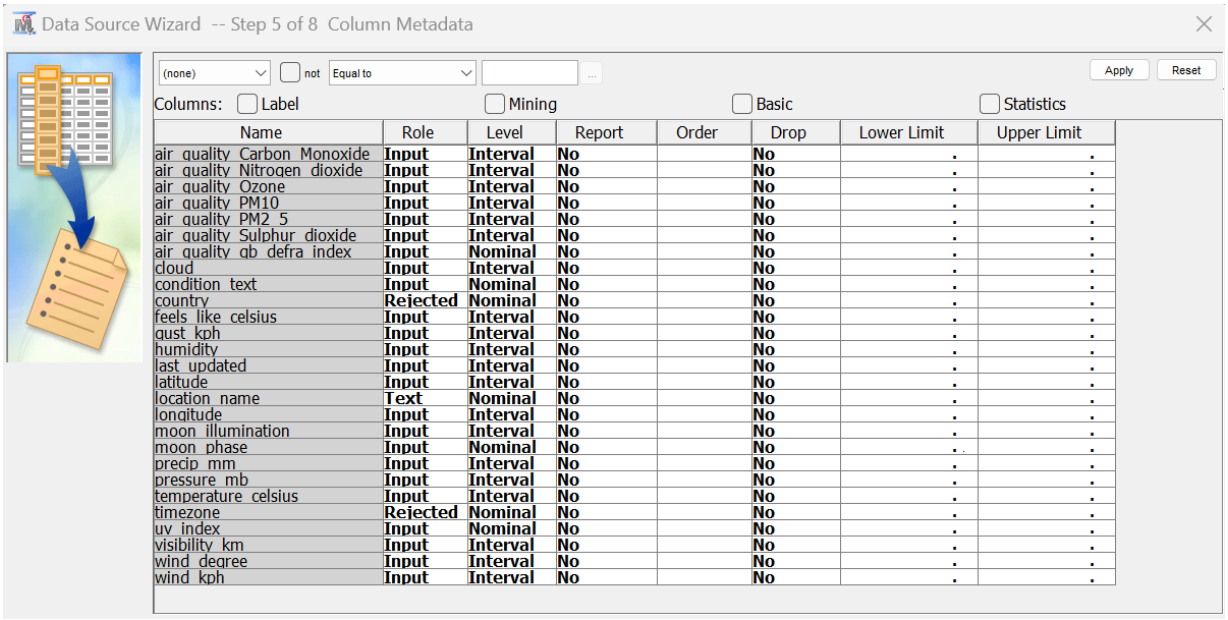
Metadata provides information about the variables such as their roles and measurement levels. In SAS Enterprise Miner, column metadata can be defined using either the Basic or Advanced settings. The Basic setting sets the initial measurement levels and roles based on variable attributes such as variable name, data type, and assigned SAS format. The Advanced setting assigns the initial measurement levels and roles based on variable attributes as well as distributions of the variables.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
air_quality_Carbon_Monoxide	Input	Interval	No		No	.	.
air_quality_gb_defra_index	Input	Interval	No		No	.	.
air_quality_Nitrogen_dioxide	Input	Interval	No		No	.	.
air_quality_Ozone	Input	Interval	No		No	.	.
air_quality_PM10	Input	Interval	No		No	.	.
air_quality_PM2_5	Input	Interval	No		No	.	.
air_quality_Sulphur_dioxide	Input	Interval	No		No	.	.
cloud	Input	Interval	No		No	.	.
condition_text	Input	Nominal	No		No	.	.
country	Input	Nominal	No		No	.	.
feels_like_celsius	Input	Interval	No		No	.	.
gust_kph	Input	Interval	No		No	.	.
humidity	Input	Interval	No		No	.	.
last_updated	Input	Interval	No		No	.	.
latitude	Input	Interval	No		No	.	.
location_name	Text	Nominal	No		No	.	.
longitude	Input	Interval	No		No	.	.
moon_illumination	Input	Interval	No		No	.	.
moon_phase	Input	Nominal	No		No	.	.
precip_mm	Input	Interval	No		No	.	.
pressure_mb	Input	Interval	No		No	.	.
temperature_celsius	Input	Interval	No		No	.	.
timezone	Input	Nominal	No		No	.	.
uv_index	Input	Interval	No		No	.	.
visibility_km	Input	Interval	No		No	.	.
wind_degree	Input	Interval	No		No	.	.
wind_kph	Input	Interval	No		No	.	.

Figure 5.1 Column metadata in Basic setting.

With the Basic setting, all the variables except location name are assigned the input role. This is not true because there should be a target variable and an ID variable. Additionally, we should only include variables that we intend to use in the modeling process. Therefore, the column metadata defined by the Basic setting cannot be accepted. On the other hand, with the Advanced setting, two variables namely country and timezone were rejected for having too many distinct values (default threshold is 20). This is true given country and timezone are unlikely significant for air quality prediction. However, target variable and ID variable are still not defined by the Advanced setting. Therefore, the column metadata defined by the Advanced setting cannot be accepted. Manual adjustment is required to correctly define the column metadata.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
air_quality_Carbon_Monoxide	Input	Interval	No		No	.	.
air_quality_Nitrogen_dioxide	Input	Interval	No		No	.	.
air_quality_Ozone	Input	Interval	No		No	.	.
air_quality_PM10	Input	Interval	No		No	.	.
air_quality_PM2_5	Input	Interval	No		No	.	.
air_quality_Sulphur_dioxide	Input	Interval	No		No	.	.
air_quality_gb_defra_index	Input	Nominal	No		No	.	.
cloud	Input	Interval	No		No	.	.
condition_text	Input	Nominal	No		No	.	.
country	Rejected	Nominal	No		No	.	.
feels_like_celsius	Input	Interval	No		No	.	.
gust_kph	Input	Interval	No		No	.	.
humidity	Input	Interval	No		No	.	.
last_updated	Input	Interval	No		No	.	.
latitude	Input	Interval	No		No	.	.
location_name	Text	Nominal	No		No	.	.
longitude	Input	Interval	No		No	.	.
moon_illumination	Input	Interval	No		No	.	.
moon_phase	Input	Nominal	No		No	.	.
precip_mm	Input	Interval	No		No	.	.
pressure_mb	Input	Interval	No		No	.	.
temperature_celsius	Input	Interval	No		No	.	.
timezone	Rejected	Nominal	No		No	.	.
uv_index	Input	Nominal	No		No	.	.
visibility_km	Input	Interval	No		No	.	.
wind_degree	Input	Interval	No		No	.	.
wind_kph	Input	Interval	No		No	.	.

Figure 5.2 Column metadata in Advanced setting.

5.4 Reclassification of Role and Level of Variables

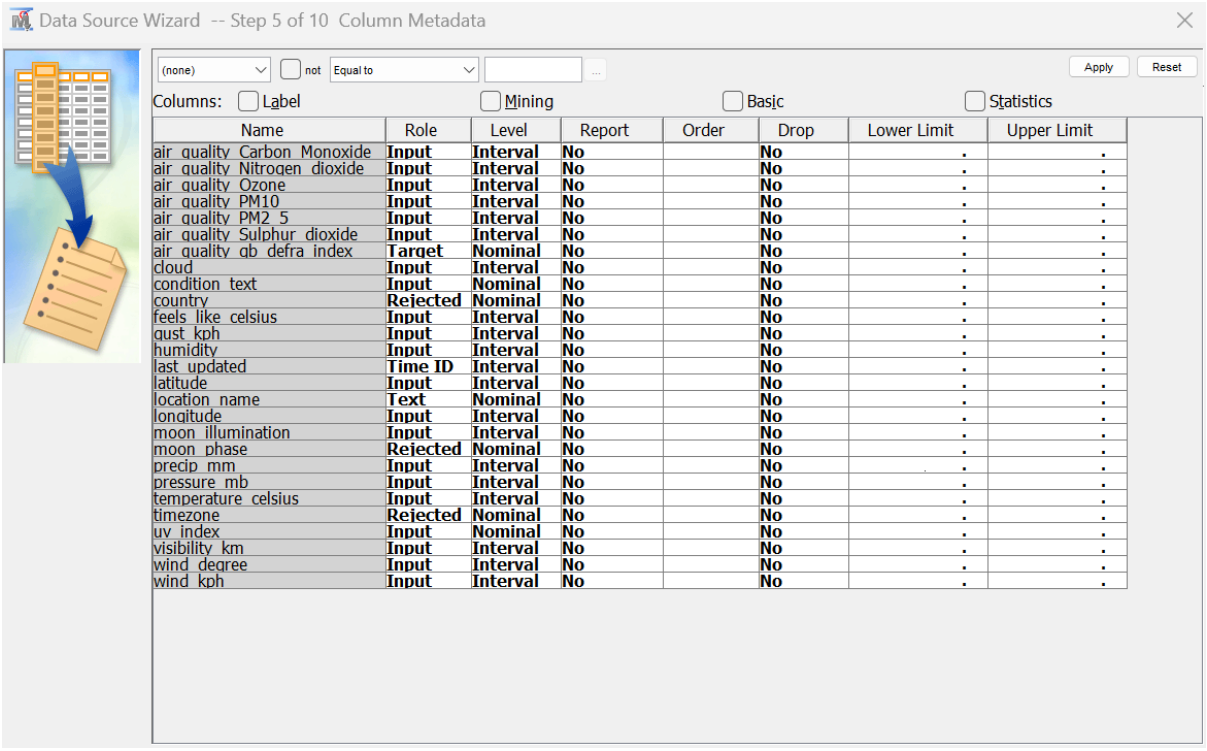
The roles and measurement levels of some variables were re-classified to correctly define the column metadata. For instance, air_quality_gb-defra-index variable should be assigned the role of target given that the main goal of our analysis is air quality prediction. The last_updated variable should be assigned the role of Time ID because it contains the timestamp of data update. The moon_phase variable should be rejected because moon phase is unlikely important for air quality prediction.

Table 5.4 Comparison between column metadata in Advanced setting and column metadata after manual re-classification.

Variable	Advanced Setting		Manual Re-classification	
	Role	**Level**	**Role**	**Level**
air_quality_Carbon_Monoxide	Input	Interval	Input	Interval
air_quality_Nitrogen_dioxide	Input	Interval	Input	Interval
air_quality_Ozone	Input	Interval	Input	Interval
air_quality_PM10	Input	Interval	Input	Interval
air_quality_PM2.5	Input	Interval	Input	Interval
air_quality_Sulphur_dioxide	Input	Interval	Input	Interval
air_quality_gb-defra-index	Input	Nominal	Target	Nominal
cloud	Input	Interval	Input	Interval

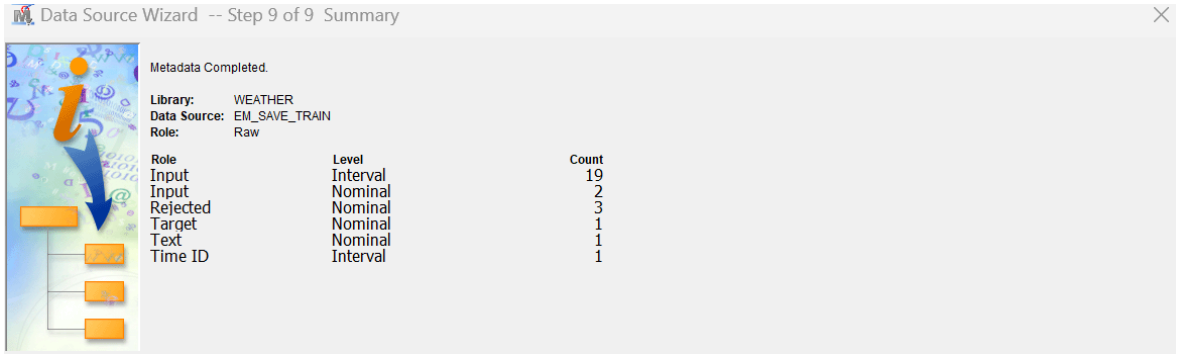
Table 5.4 *Comparison between column metadata in Advanced setting and column metadata after manual re-classification, Cont*

Variable	Advanced Setting		Manual Re-classification	
	Role	**Level**	**Role**	**Level**
condition_text	Input	Nominal	Input	Nominal
country	Rejected	Nominal	Rejected	Nominal
feels_like_celsius	Input	Interval	Input	Interval
gust_kph	Input	Interval	Input	Interval
humidity	Input	Interval	Input	Interval
last_updated	Input	Interval	Time ID	Interval
latitude	Input	Interval	Input	Interval
location_name	Text	Nominal	Text	Nominal
longitude	Input	Interval	Input	Interval
moon_illumination	Input	Interval	Input	Interval
moon_phase	Input	Nominal	Rejected	Nominal
precip_mm	Input	Interval	Input	Interval
pressure_mb	Input	Interval	Input	Interval
temperature_celsius	Input	Interval	Input	Interval
timezone	Rejected	Nominal	Rejected	Nominal
uv_index	Input	Nominal	Input	Nominal
visibility_km	Input	Interval	Input	Interval
wind_degree	Input	Interval	Input	Interval
wind_kph	Input	Interval	Input	Interval



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
air_quality_Carbon_Monoxide	Input	Interval	No		No	.	.
air_quality_Nitrogen_dioxide	Input	Interval	No		No	.	.
air_quality_Ozone	Input	Interval	No		No	.	.
air_quality_PM10	Input	Interval	No		No	.	.
air_quality_PM2_5	Input	Interval	No		No	.	.
air_quality_Sulphur_dioxide	Input	Interval	No		No	.	.
air_quality_gb_defra_index	Target	Nominal	No		No	.	.
cloud	Input	Interval	No		No	.	.
condition_text	Input	Nominal	No		No	.	.
country	Rejected	Nominal	No		No	.	.
feels_like_celsius	Input	Interval	No		No	.	.
gust_kph	Input	Interval	No		No	.	.
humidity	Input	Interval	No		No	.	.
last_updated	Time ID	Interval	No		No	.	.
latitude	Input	Interval	No		No	.	.
location_name	Text	Nominal	No		No	.	.
longitude	Input	Interval	No		No	.	.
moon_illumination	Input	Interval	No		No	.	.
moon_phase	Rejected	Nominal	No		No	.	.
precip_mm	Input	Interval	No		No	.	.
pressure_mb	Input	Interval	No		No	.	.
temperature_celsius	Input	Interval	No		No	.	.
timezone	Rejected	Nominal	No		No	.	.
uv_index	Input	Nominal	No		No	.	.
visibility_km	Input	Interval	No		No	.	.
wind_degree	Input	Interval	No		No	.	.
wind_kph	Input	Interval	No		No	.	.

Figure 5.3 Column metadata after manual re-classification.



Role	Level	Count
Input	Interval	19
Input	Nominal	2
Rejected	Nominal	3
Target	Nominal	1
Text	Nominal	1
Time ID	Interval	1

Figure 5.4 Summary details about column metadata after manual re-classification.

6 Explore

6.1 Summary Statistics

Upon accessing and examining the dataset, a set of summary statistics is produced. The purpose of these summary statistics is to provide an overview of the data pattern, encompassing details like minimum and maximum values, mean, presence of missing values, and standard deviation. Figure 6.1 and Figure 6.2 display the summary statistics for interval and class variables. It shows that our dataset has some missing values. Pressure_mb variable has 9 missing values; condition_text variable has 22 missing values; country variable has 7 missing values.

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
air_quality_Carbon_Monoxide	INPUT	568.364	1436.985	17149	0	96.8	283.7	36315.9	12.73704	208.8085
air_quality_Nitrogen_dioxide	INPUT	12.7675	22.65928	17149	0	0	4.4	337.2	4.175877	26.51848
air_quality_Ozone	INPUT	40.75872	32.20778	17149	0	0	37.2	555	1.763954	10.44748
air_quality_PM10	INPUT	41.42908	99.53617	17149	0	0.5	12.9	2504.3	8.240322	103.9945
air_quality_PM2_5	INPUT	24.1824	67.3488	17149	0	0.5	7.6	1558.8	9.536901	122.8243
air_quality_Sulphur_dioxide	INPUT	7.091737	16.55664	17149	0	0	1.7	335.7	6.153057	60.47203
cloud	INPUT	35.35588	33.08028	17149	0	0	25	100	0.425445	-1.19751
feels_like_celsius	INPUT	22.19765	10.88277	17149	0	-36.8	24.6	73.6	-0.47493	0.641658
gust_kph	INPUT	17.53299	11.28636	17149	0	0	15.3	110.5	1.226396	2.286825
humidity	INPUT	72.48195	20.29754	17149	0	4	77	100	-0.93965	0.289414
latitude	INPUT	19.30084	24.58315	17149	0	-41.3	17.25	63.83	-0.30613	-0.7684
longitude	INPUT	21.90824	65.69636	17149	0	-175.2	23.24	179.22	0.00864	0.334742
moon_illumination	INPUT	50.89335	35.13772	17149	0	0	49	100	-0.01611	-1.48828
precip_mm	INPUT	0.161076	0.748877	17149	0	0	0	31	17.24163	480.749
pressure_mb	INPUT	1013.218	6.689865	17140	9	964	1013	1053	-0.60629	3.484119
temperature_celsius	INPUT	20.84551	8.491842	17149	0	-31	22.5	45.4	-0.80762	0.890103
visibility_km	INPUT	9.702548	2.548837	17149	0	0	10	32	2.0093	17.32982
wind_degree	INPUT	162.189	105.521	17149	0	1	150	360	0.230127	-1.12867
wind_kph	INPUT	11.06719	7.945606	17149	0	3.6	9	141.1	2.079137	10.39417

Figure 6.1 Interval variable summary statistics.

Class Variable Summary Statistics
(maximum 500 observations printed)

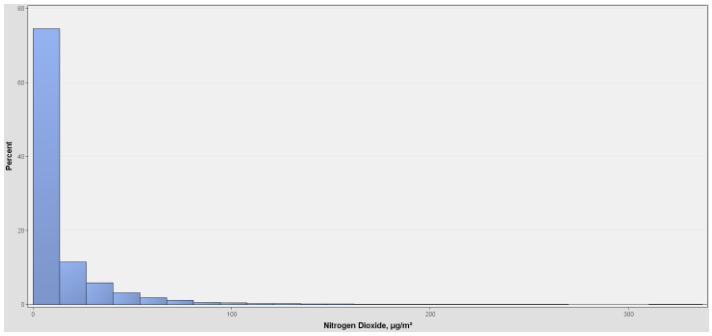
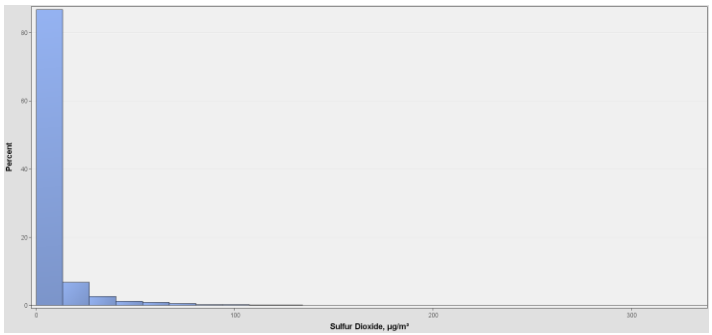
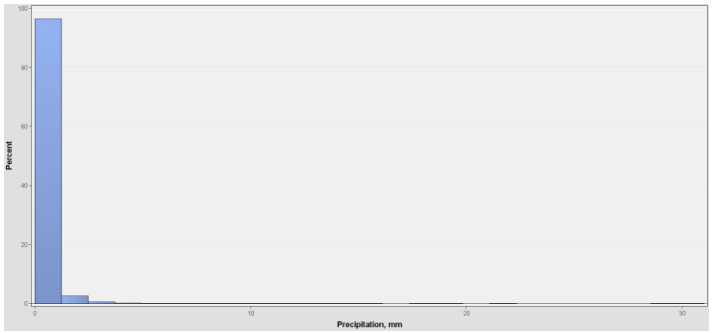
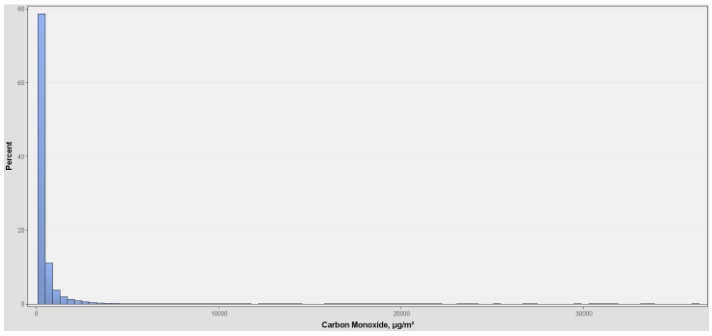
Data Role=TRAIN

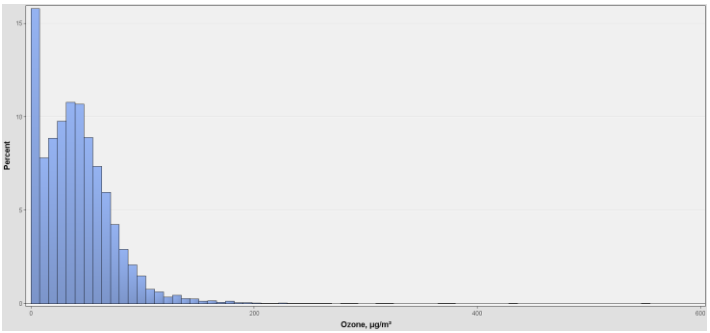
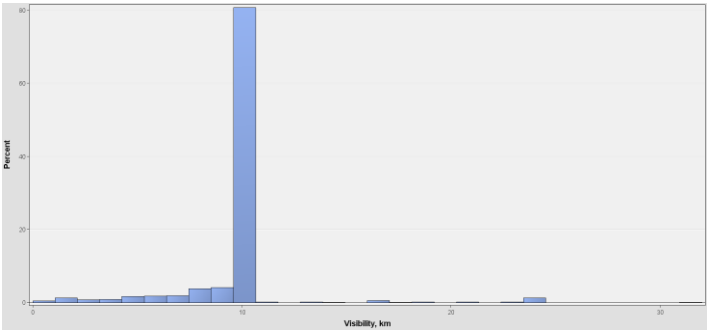
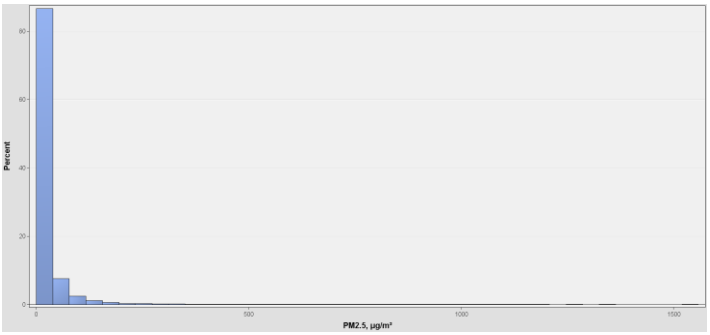
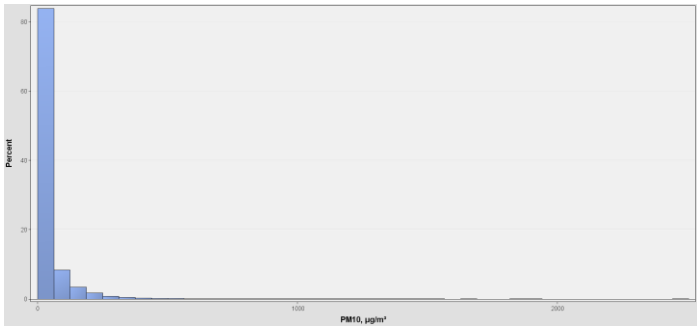
Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	condition_text	INPUT	15	22	Partly cloudy	39.34	Clear	34.08
TRAIN	uv_index	INPUT	11	0	1	75.59	6	9.21
TRAIN	country	REJECTED	186	7	Bulgaria	1.54	Indonesia	1.09
TRAIN	moon_phase	REJECTED	8	0	Waxing Crescen	22.72	Waning Crescen	20.47
TRAIN	timezone	REJECTED	183	0	Europe/Rome	1.54	Asia/Bangkok	1.53
TRAIN	air_quality_gb_defra_index	TARGET	10	0	1	60.97	2	16.62

Figure 6.2 Class variable summary statistics

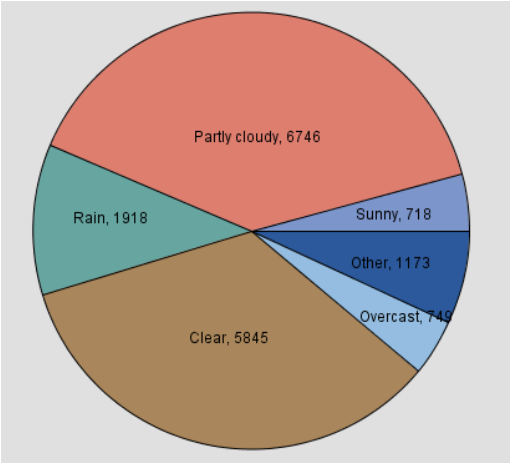
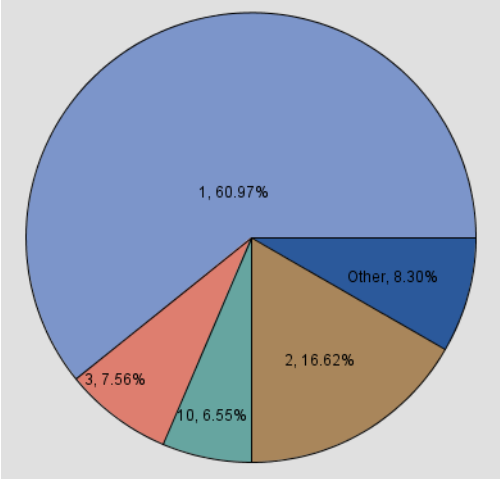
6.2 Univariate Analysis

During the explore phase, we initiate our data discovery with univariate analysis. During this stage, each parameter within the dataset is scrutinized independently to assess both the distribution and central tendency of the data. In addition to providing individual descriptions for each parameter, univariate analysis offers a visual representation of the overall trend in response to each parameter. Histogram is the most commonly used graph for univariate analysis. This graphical representation not only aids in understanding the skewness and curvature of the data but also serves as a valuable tool for visualizing patterns in both discrete and continuous data types.

Histogram		
No	Variable	Findings
1	Nitrogen Dioxide (NO_2) 	• As can be seen from the figure, most of the NO_2 concentration is below 100 µg/m³, and this part accounts for more than 90%.
2	Sulphur Dioxide (SO_2) 	• As shown in the picture, the SO_2 has a similar trend to NO_2. However, SO_2 has higher percentage (~80%) in having concentration less than 12.5µg/m³ as compared to the NO_2 (~70%).
4	precip_mm 	• Generally speaking, the precipitation is low, and most of the precipitation even tends to 0mm.
5	Carbon Monoxide (CO) 	• Overall, the carbon monoxide content is not too high. But there are also some extreme values

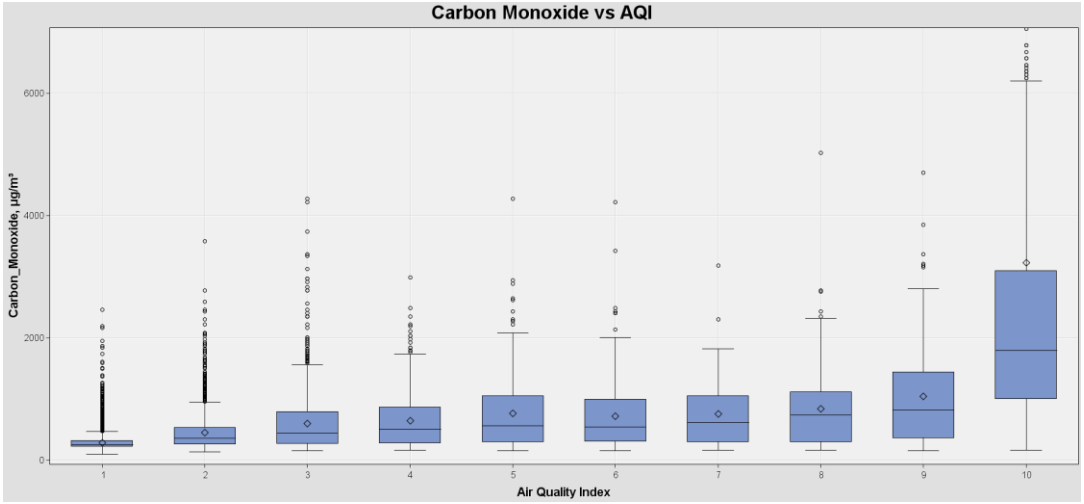
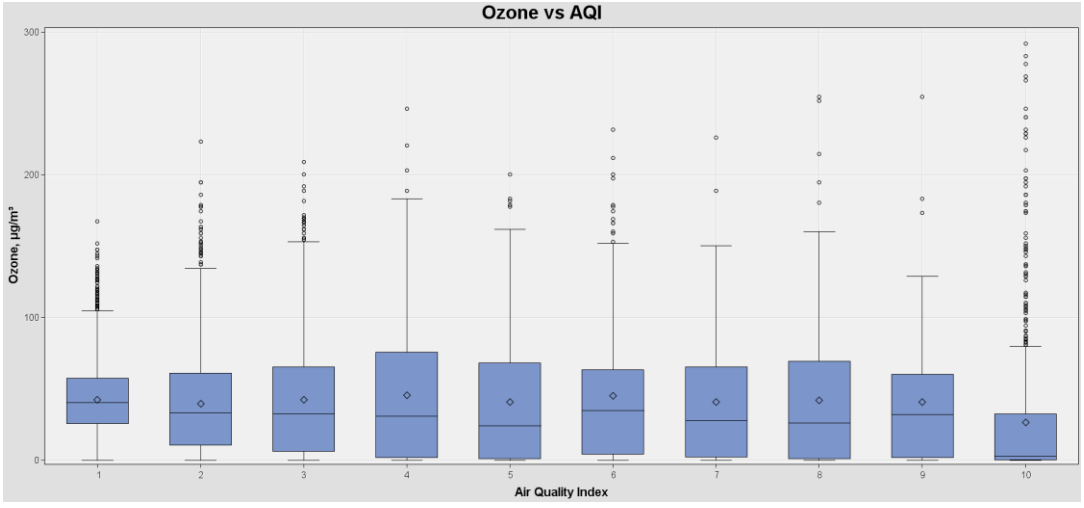
6	Ozone(O_3) 	• The vast majority of ozone levels are below 200 µg/m³, but there are some values that are too high
7	Visibility 	• The visibility is around 10km in most cases, and 10km of visibility is considered to be quite good visibility. In most cases,10 kilometers of visibility are considered a long sight distance in open space. This visibility usually indicates clear weather and good atmospheric conditions
8	PM2.5 	• Most PM2.5 concentrations are less than 100 µg/m³, which falls between "good" and "lightly polluted" levels. This is relatively healthy air quality,
9	PM10 	• Most PM10 concentrations are less than 50 µg/m³, which usually fall into the "excellent" range, indicating relatively good air quality.

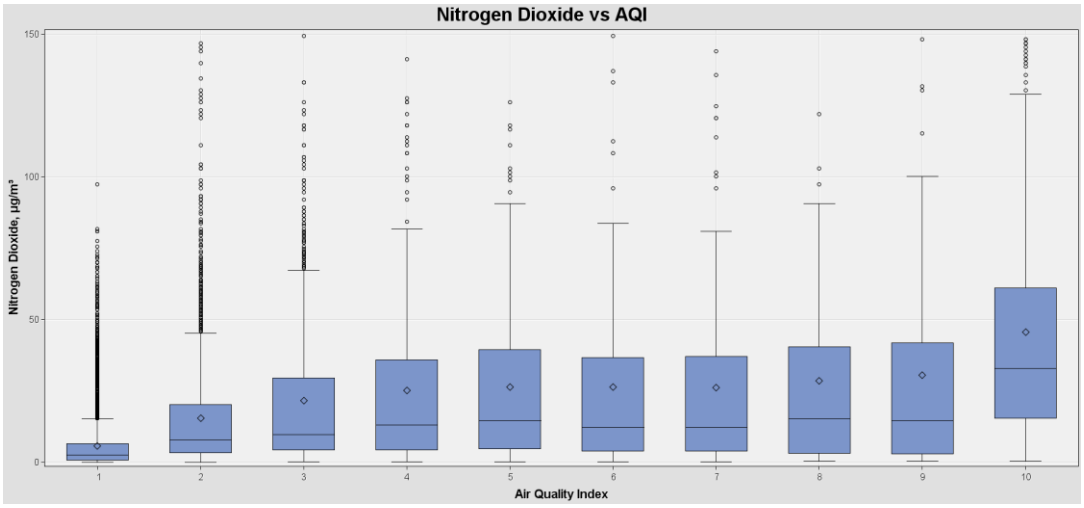
A characteristic that lacks quantifiable measures is termed a categorical variable, commonly known as a qualitative variable. Examples of categorical variables include nominal and ordinal variables. A nominal variable describes an arbitrarily assigned name, tag, or group, while an ordinal variable involves numbers assigned based on the hierarchy of groups. For illustrating the relative contribution of each component to the whole, the SAS Pie Chart generates straightforward, category, or layered diagrams that represent information as segments of a pie. Each segment corresponds to a specific data category, and the size of each segment reflects the extent to which that category has contributed to the overall graphical statistic.

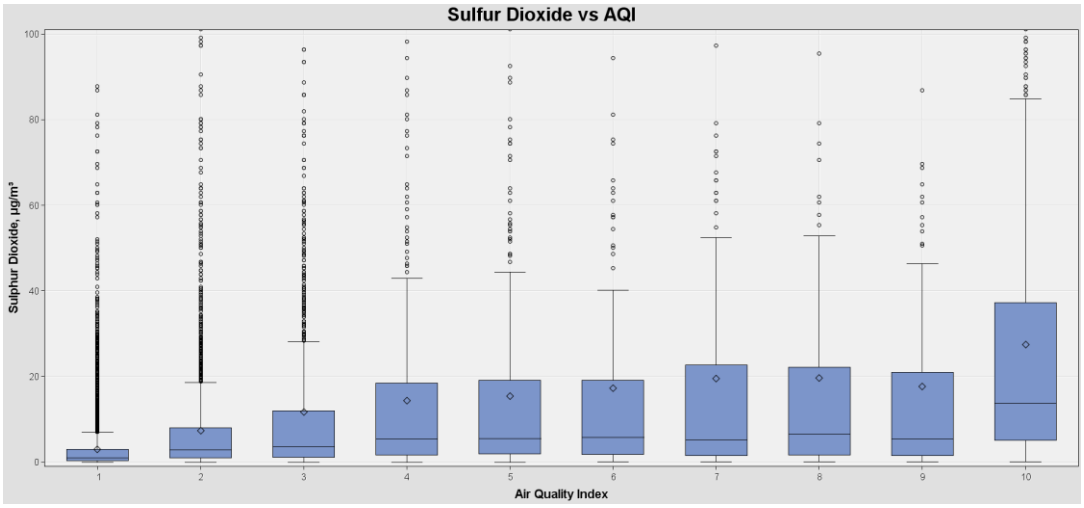
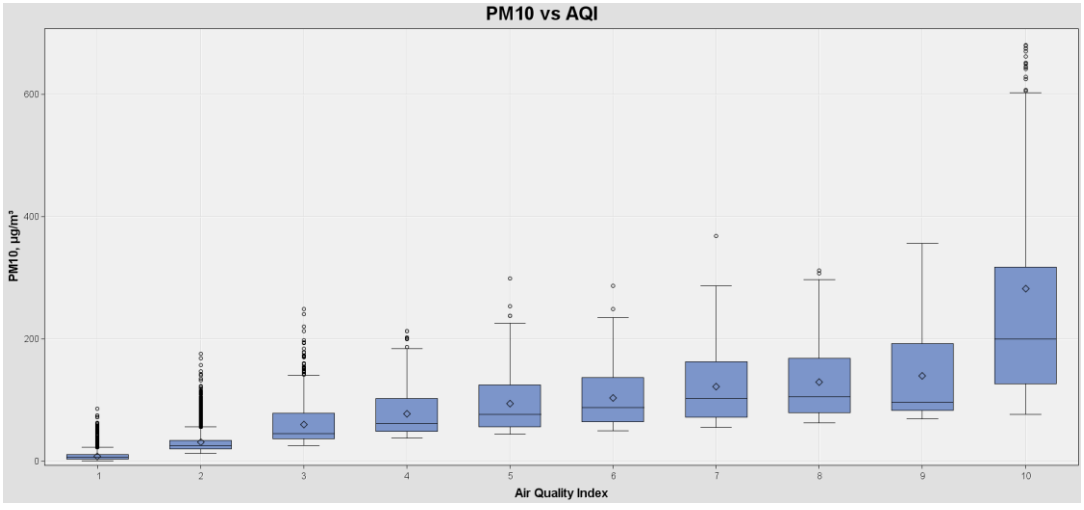
Pie charts		
No.	Variable	Findings
1	Weather Conditions 	• The weather conditions are mostly partly cloudy and clear, and these two types account for most of the conditions.
2	AQI 	• In the DEFRA Air Quality Index (AQI), lower values correspond to better air quality, while higher values indicate worse air quality. Specifically, DEFRA AQI values are graded and range from 1 to 10. Most of the AQIs in the picture are 1-3, indicating that the air quality is good and there is almost no risk to health.

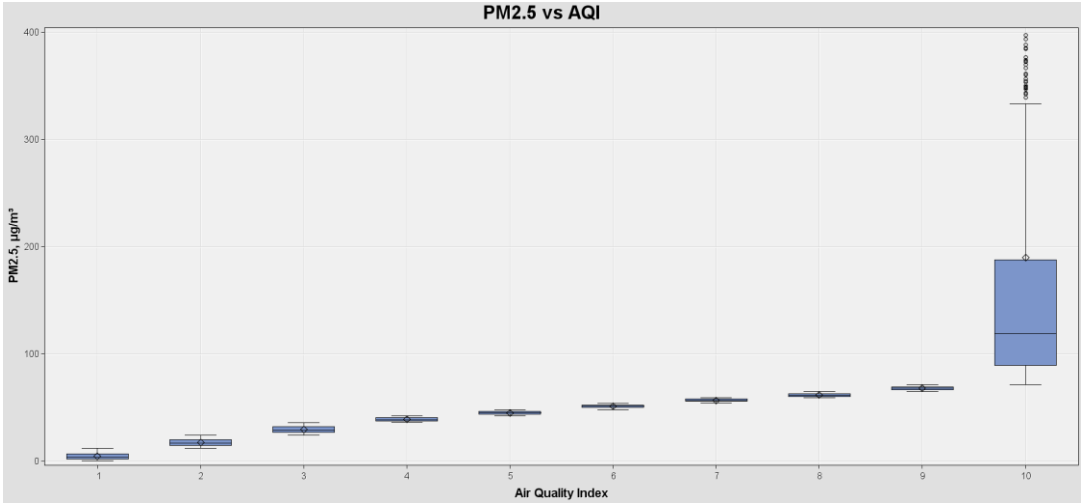
6.3 Bivariate Analysis

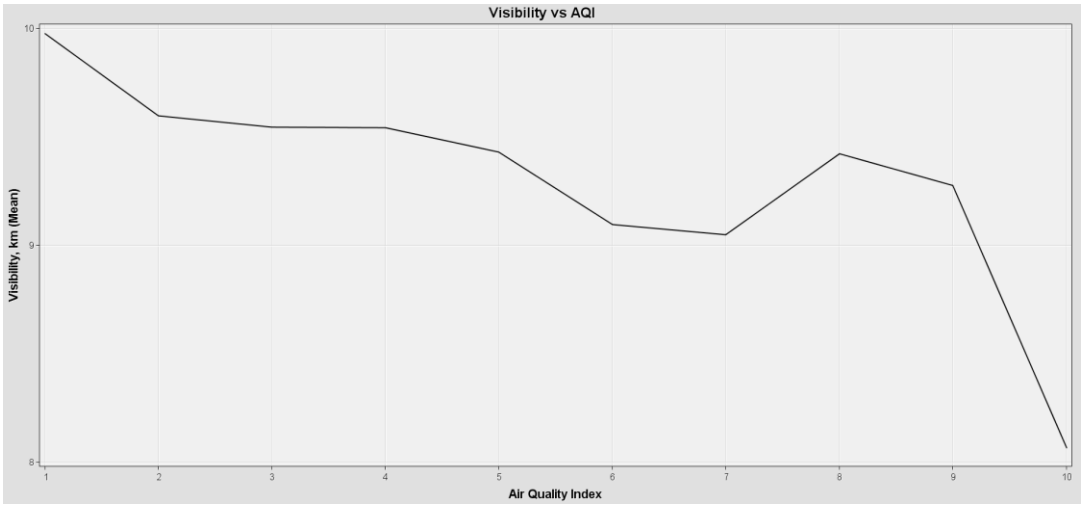
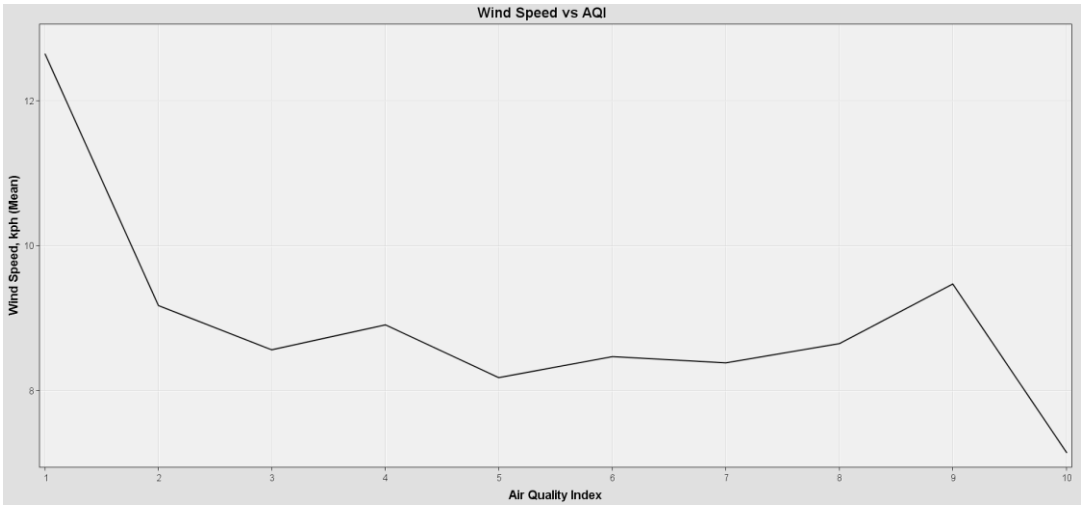
Bivariate analysis involves examining two variables simultaneously to investigate their relationship. This analysis aims to determine whether there is an association between the two variables, assess the strength of this association, identify differences between the variables, and evaluate the significance of these differences. Various techniques, such as line charts and box plots, are employed in bivariate analysis. The goal is to uncover patterns or trends in the data, offering valuable insights into the interdependence of the two variables under consideration.

No	**Bivariate Analysis Chart**
1.	Variables Used: Carbon Monoxide (CO), Air Quality Index (AQI)  Observations: • The overall CO concentration is skewed to the right, indicating that most outliers of CO concentration are greater than the median. • The overall median of the CO concentration in each AQI category varies from 250 µg/m³ to 750 µg/m³, except for AQI = 10, where its value falls around 1750 µg/m³. • The mean of CO concentration shows an increase from AQI=1 to AQI = 2 and then fluctuates in a small range of values from AQI = 2 to AQI = 9 prior to a sharp increase when AQI = 10. • Outliers are found across AQI, but it may not necessarily warrant their removal. CO concentration can be influenced by both natural and anthropogenic sources. The presence of outliers might be reflective of specific events or conditions contributing to elevated CO concentration. Removing outliers without a justified reason could lead to a loss of important information.
2.	Variables Used: Ozone (O_3), AQI 

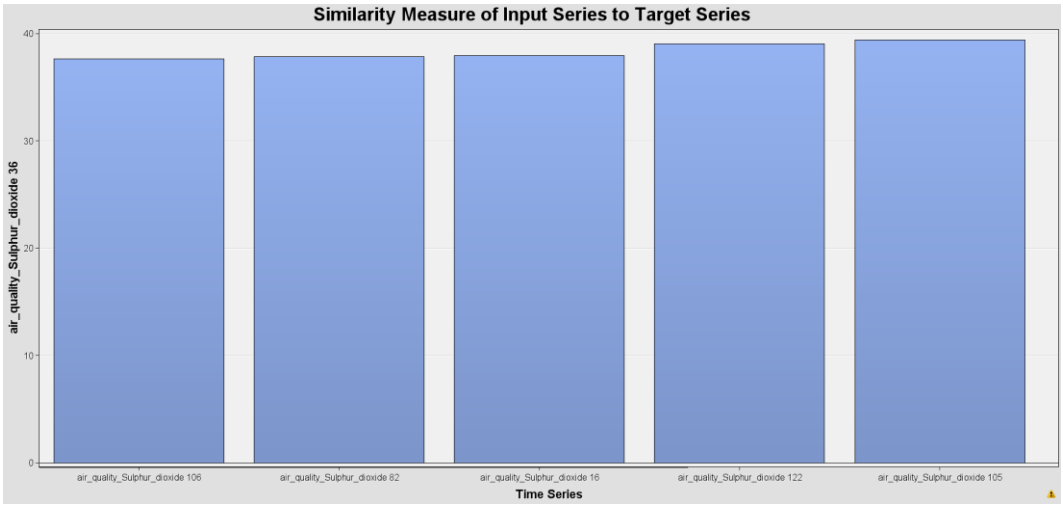
	Observations: • The overall O_3 concentration is skewed to the right, indicating that most outliers of O_3 concentration are greater than the median. • The overall median of the O_3 concentration in each AQI category varies from 20 µg/m³ to 40 µg/m³, except for AQI = 10, where its value falls around 2 µg/m³. • The mean of O_3 concentration fluctuates in a small range of values from AQI=1 to AQI = 9 prior to a sharp dip when AQI = 10. • Outliers are found across AQI, but their removal may not be necessary. O_3 is a variable pollutant, and the presence of outliers may have environmental significance. Extreme O_3 levels can be influenced by specific atmospheric conditions and sources, and removing outliers might mask important insights.
3.	Variables Used: Nitrogen Dioxide (NO_2), AQI  Observations: • The overall NO_2 concentration is skewed to the right, indicating that most outliers of NO_2 concentration are greater than the median. • The overall median of the NO_2 concentration in each AQI category varies from 2 µg/m³ to 15 µg/m³, except for AQI = 10, where its value falls around 40 µg/m³. • The mean of NO_2 concentration shows a significant increase from AQI=1 to AQI= 2 and then a gradual increase in concentration from AQI = 2 to AQI = 9 before another significant increase when AQI = 10. • Outliers are found across AQI, but it may not necessarily warrant their removal. Outliers in NO_2 concentration may represent genuine environmental variations. NO_2 is a common air pollutant, and concentration can vary due to factors such as industrial activities, and meteorological conditions. Retaining outliers allows for a more accurate representation of the environmental conditions.

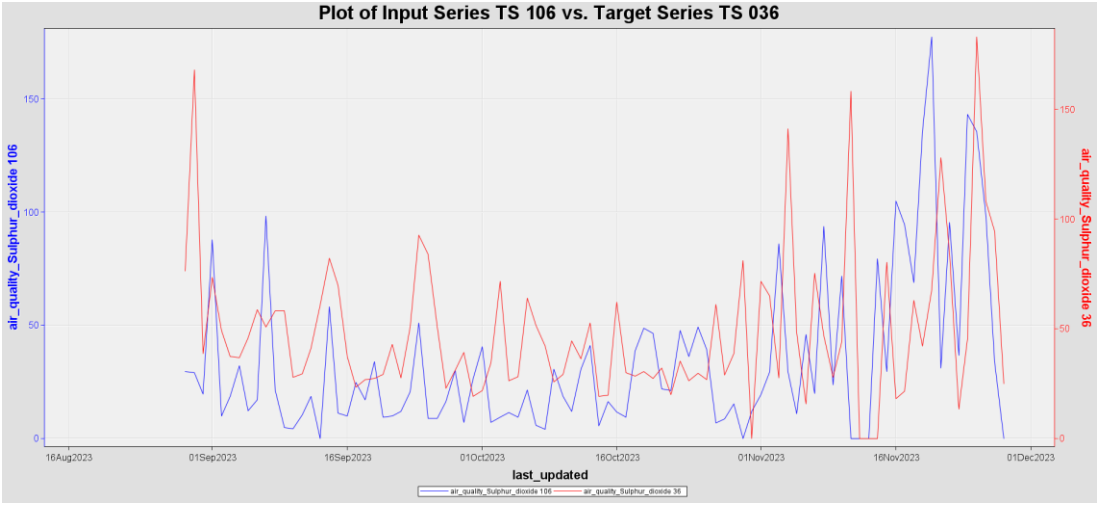
4.	Variables Used: Sulphur Dioxide (SO_2), AQI  Observations: • The overall SO_2 concentration is skewed to the right, indicating that most outliers of SO_2 concentration are greater than the median. • The overall median of the SO_2 concentration in each AQI category varies from 1 µg/m³ to 7 µg/m³, except for AQI = 10, where its value falls around 14 µg/m³. • The mean of SO_2 concentration shows a significant increase from AQI=1 to AQI=4 and then fluctuates in a small range of values from AQI=4 to AQI=9 prior to a sharp rise when AQI = 10. • Outliers are found across AQI, but their removal may not be necessary. Outliers in SO_2 may reflect genuine environmental variations. Removing outliers without justification might lead to an oversimplified view that does not capture the true complexity of the data.
5.	Variables Used: PM10, AQI  Observations: • The overall PM10 concentration is skewed to the right, suggesting that most outliers in SO2 concentration are greater than the median. • The overall trend in the median for PM10 concentration in each AQI category indicates an incremental increase from AQI=1 to AQI=8, followed by a slight drop at AQI=9, and a sharp increase at AQI=10. • The mean of PM10 concentration also shows a continuous increase in its

	concentration across the AQI with a sudden spike at AQI=10. • Outliers are found across AQI, but their removal may not be necessary. Outliers in air quality data may represent genuine environmental variations. Extreme levels of PM10 occurs due to specific air quality incidents. Removing such outliers could result in a loss of information about these extreme events.
6.	Variables Used: PM2.5, AQI  Observations: • The overall PM2.5 concentration has approximately equal proportions around the median across the AQI, except for AQI=10 where it is skewed to the right, suggesting that most outliers are greater than the median. • The overall trend in the median for PM2.5 concentration in each AQI category indicates an incremental increase from AQI=1 to AQI=9, followed by a steep climb at AQI=10. • The mean of PM2.5 concentration also shows a continuous increase in its concentration across the AQI with a sudden spike at AQI=10. • Outliers are found at AQI=10 which is a threshold for concern. Removing data points at this level might lead to an incomplete understanding of the frequency and impact of extremely poor air quality incidents.

7.	Variables Used: Visibility, AQI  Observations: • The visibility and AQI are inversely correlated. Visibility decreases when AQI increases. This is because more air pollutants scatter in the air and absorb light, making it difficult to see distant objects. • The relationship between AQI and visibility is nonlinear. This means that the rate of change in visibility is not constant as AQI increases. For example, the increase of AQI from 9 to 10 results in a significant decrease in visibility.
8.	Variables Used: Wind Speed, AQI  Observations: • The wind speed and AQI are inversely correlated. AQI increases when wind speed decreases. This is because higher wind speeds disperse contaminants more effectively, resulting in lower concentrations. • The relationship between AQI and wind speed is nonlinear. This means that the rate of change in wind speed is not constant as AQI increases. For example, the increase of AQI from 9 to 10 results in a significant decrease in wind speed.

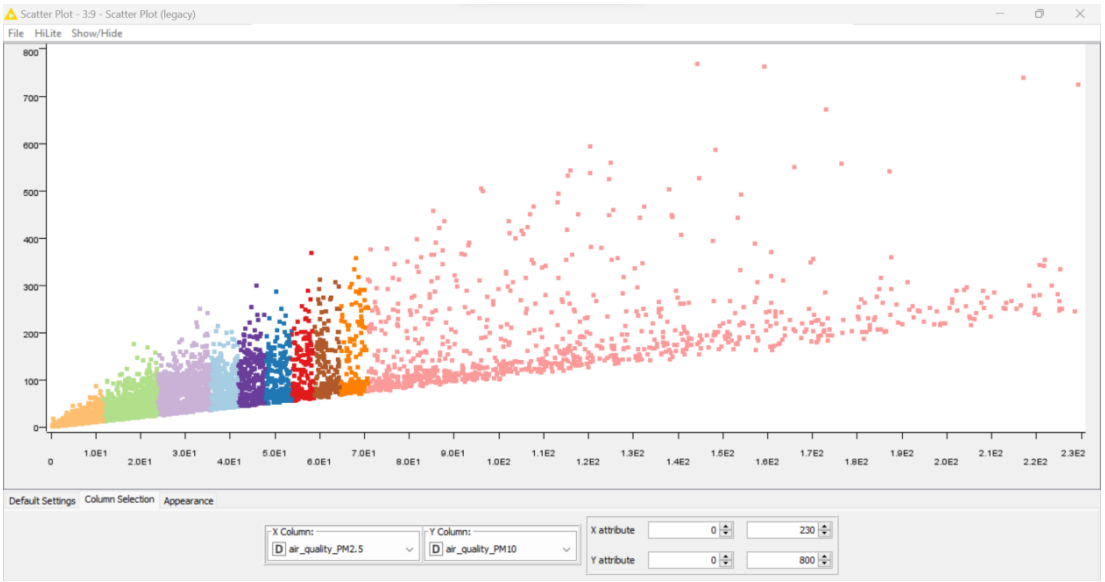
9. Time Series Similarity Plot





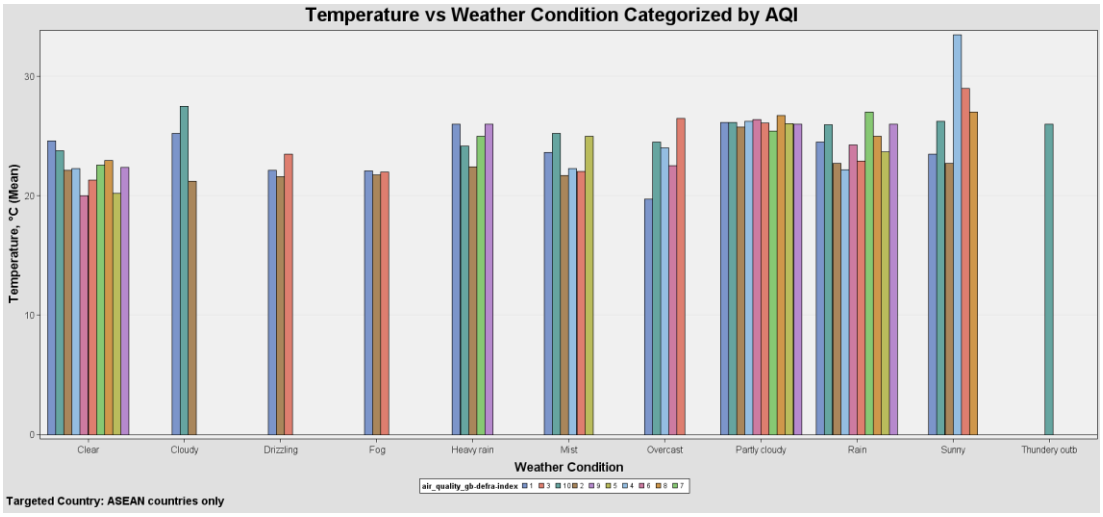
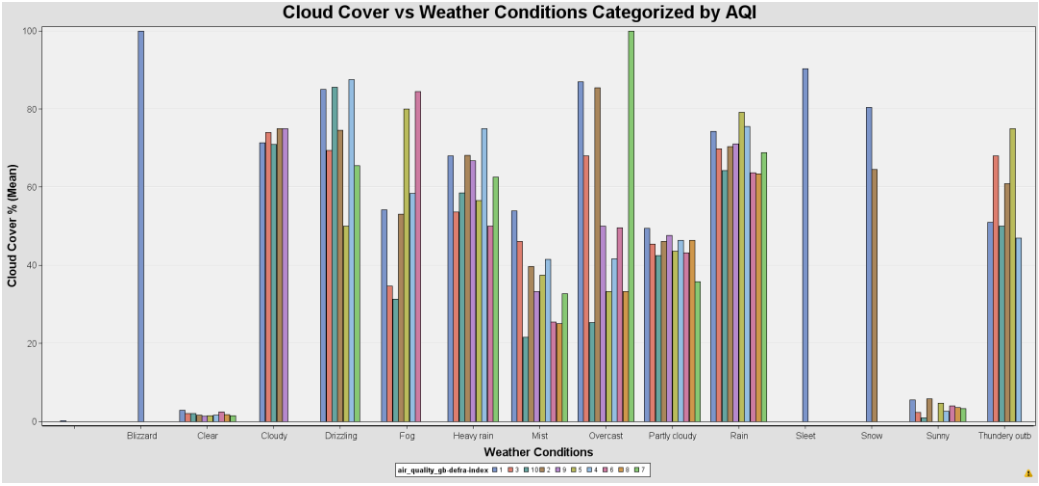
Observations:

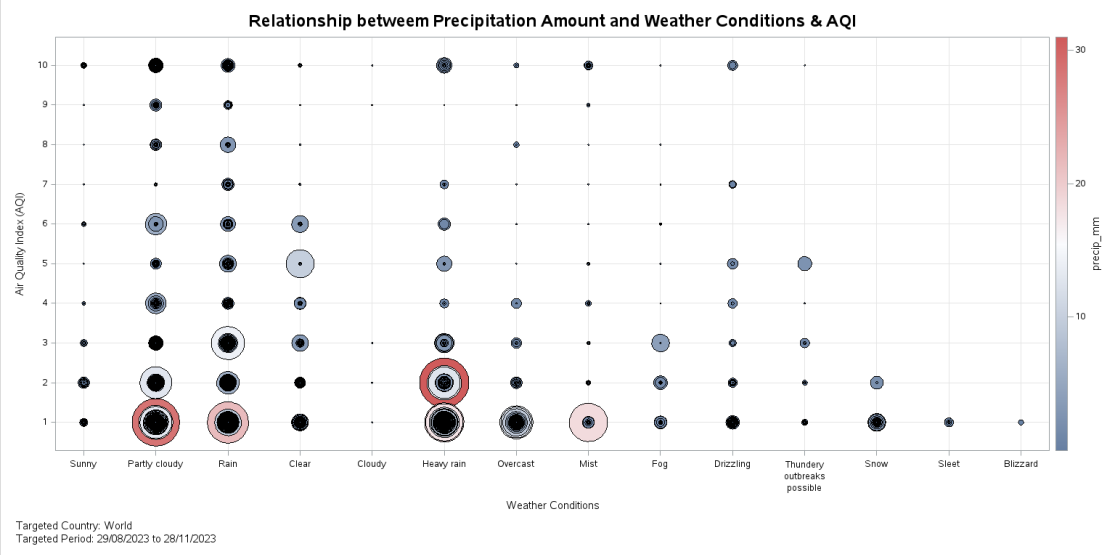
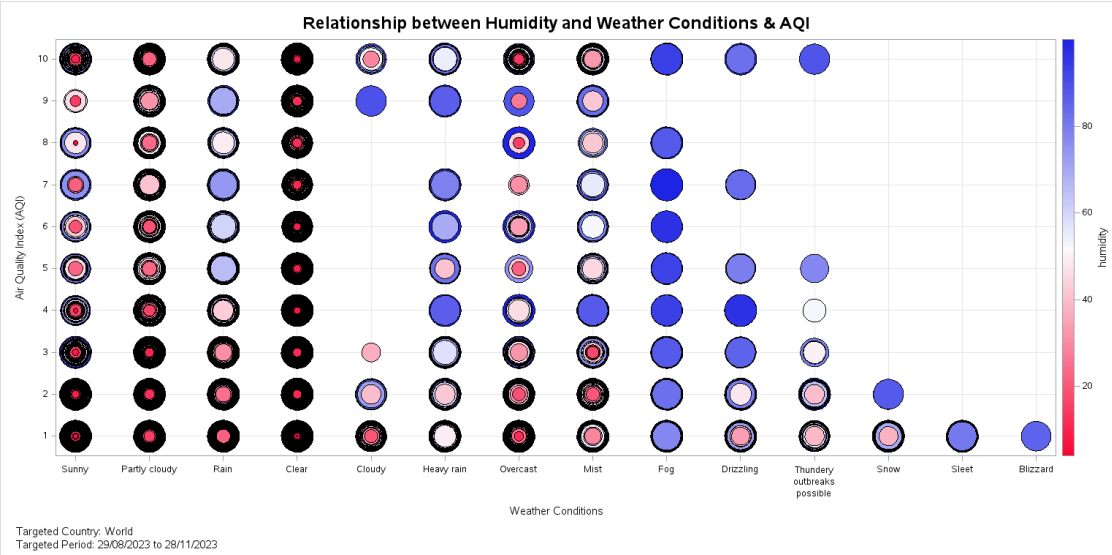
- The analysis is focused on identifying time series with similar SO_2 concentration patterns to China. SO_2 concentration patterns in China (TS 036) is identified as the most similar to Mexico (TS 106), Japan (TS 082), Belarus (TS 016), Norway (TS 122) and Mauritius (TS 105).
- One likely justification could be a combination of geographical proximity and shared industrial activities. Similar latitudinal positions can contribute to shared meteorological conditions and potentially similar patterns in environmental factors, including SO_2 concentrations. Additionally, similarities in industrial activities among these countries could lead to comparable emissions of SO_2 and result in similar concentration patterns over time.

10	KNIME DBSCAN Plot  Observations: • The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) plot with 10 clusters for PM2.5 and PM10 concentrations indicates that the data points have been grouped into 10 distinct clusters based on their density in the feature space defined by PM2.5 and PM10 concentrations. • The identification of 10 clusters suggests that there are inherent patterns or subgroups in the data related to PM2.5 and PM10 concentrations. Each of the 10 clusters may correspond to specific environmental conditions, sources of pollution, or other factors influencing PM2.5 and PM10 concentrations. • Given that the data points were categorized into 10 AQI classes, it provides a direct link between the clustering results and air quality levels. Each cluster identified by DBSCAN may correspond to a specific range or pattern of AQI classes. For example, light orange cluster might represent data points with consistently good air quality (low AQI), while light red cluster might represent points with consistently poor air quality (high AQI).

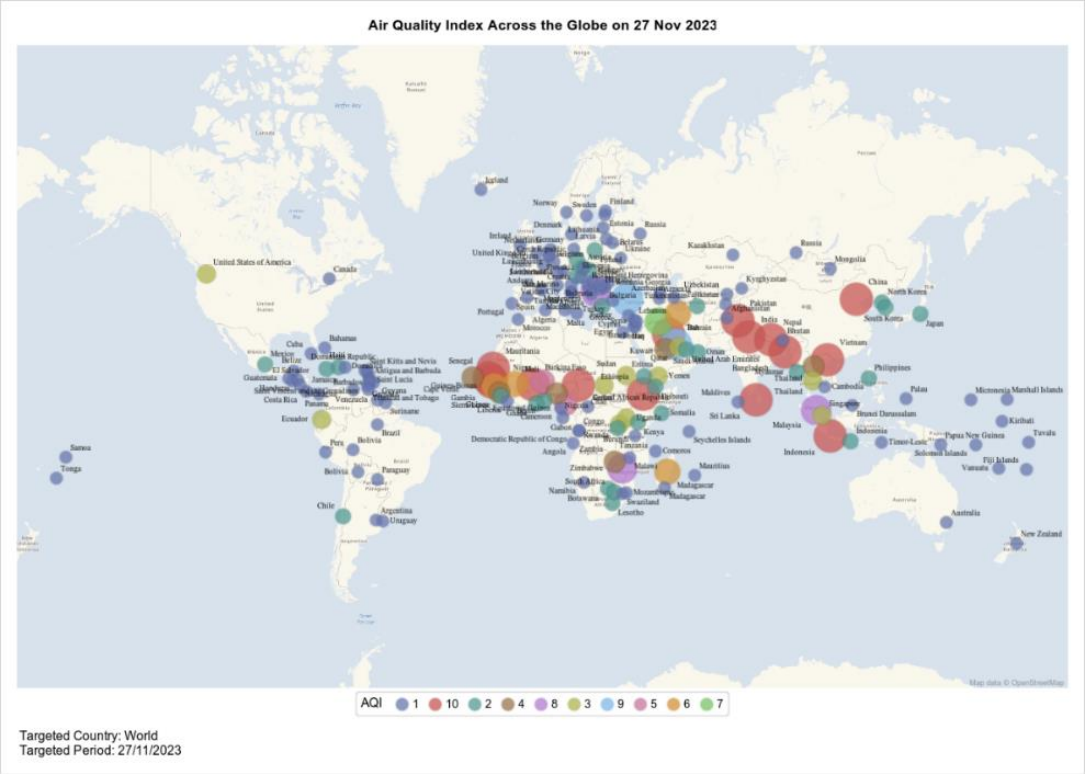
6.4 Multivariate Analysis

Multivariate analysis refers to statistical techniques used to analyze data that involves multiple variables. Unlike univariate analysis (which focuses on a single variable) and bivariate analysis (which involves the relationship between two variables), multivariate analysis considers the simultaneous analysis of three or more variables. This analysis aims to identify underlying patterns and relationships within the data, such as identifying groups of related variables or identifying important variables that explain most of the variation in the data. Various graphical visualization techniques have been used for the multivariate analysis bar charts, bubble plots, bar-line charts, etc. In addition, we also utilize the Variable Clustering node in the SAS Enterprise Miner to plot out the correlation matrix between the variables. By conducting multivariate analysis, we can understand the complex relationships between multiple variables and uncover patterns, trends, or structures within the data. This can help inform decisions for modeling and predicting outcomes in the next stage of the SEMMA process, the Model stage.

Three Variables	
1.	Variables Used: Temperature, Weather Conditions, Air Quality Index (AQI)  Observations: • The above graph focuses on the mean temperature recorded in ASEAN countries as countries with four seasons will affect the representative of the data. • The mean temperature is affected by the type of weather conditions. As example, the mean temperature is the highest during Sunny as compared to drizzling and heavy rain. • The mean temperature is not highly affected by the AQI as there are marginal differences in the mean temperature across the AQI for each weather condition. Therefore, temperature variable has no effect on the AQI.
2.	Variables Used: Cloud Cover, Weather Conditions, Air Quality Index (AQI)  Observations: • The mean cloud cover % is mainly affected by the type of weather conditions. As example, the mean cloud cover % is the highest during overcast as compared to clear or sunny day. No significant influence of cloud cover % on AQI is observed as the data shows inconsistency.

3.	Variables Used: Precipitation Amount, Weather Conditions, Air Quality Index (AQI)  Observations: • The type of weather conditions has impacts on the precipitation amount. As examples, the precipitation amount is higher (larger bubble size) during heavy rain as compared to sunny day or cloudy day. • The precipitation amount has a negative correlation with the AQI as the amount of precipitation increases, the air quality of the city improves. As examples, AQI=1 has greater precipitation amount as compared to other AQI.
4.	Variables Used: Humidity, Weather Conditions, Air Quality Index (AQI)  Observations: • The above graph shows that humidity does not have significant impacts on the weather conditions and AQI as the bubble size is almost identical across the weather conditions and AQI.

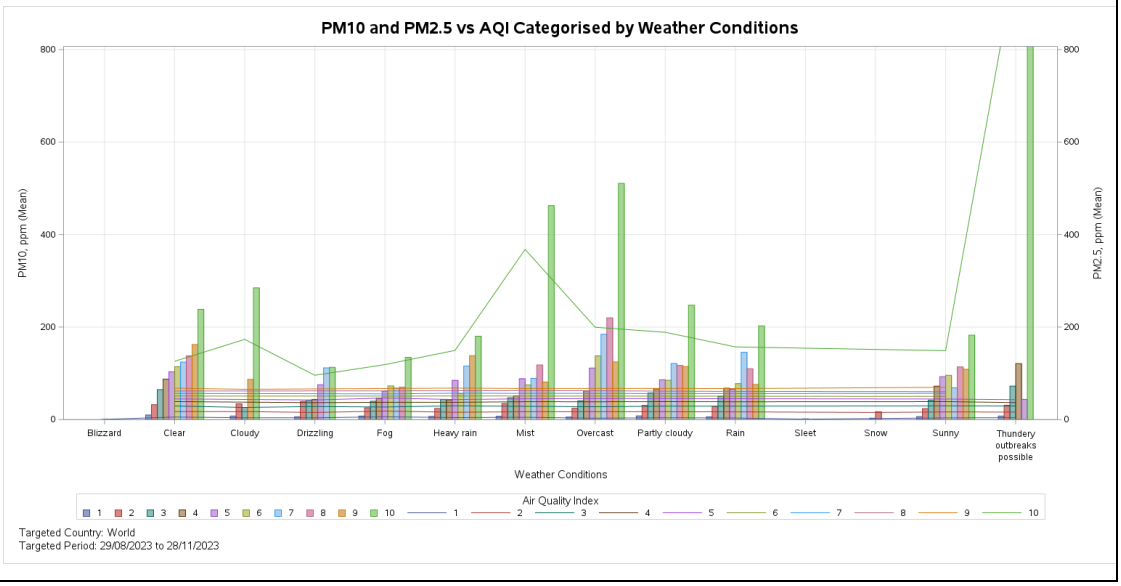
Four Variables	
5.	Variable Used: Longitude, Latitude, Air Quality Index, Country

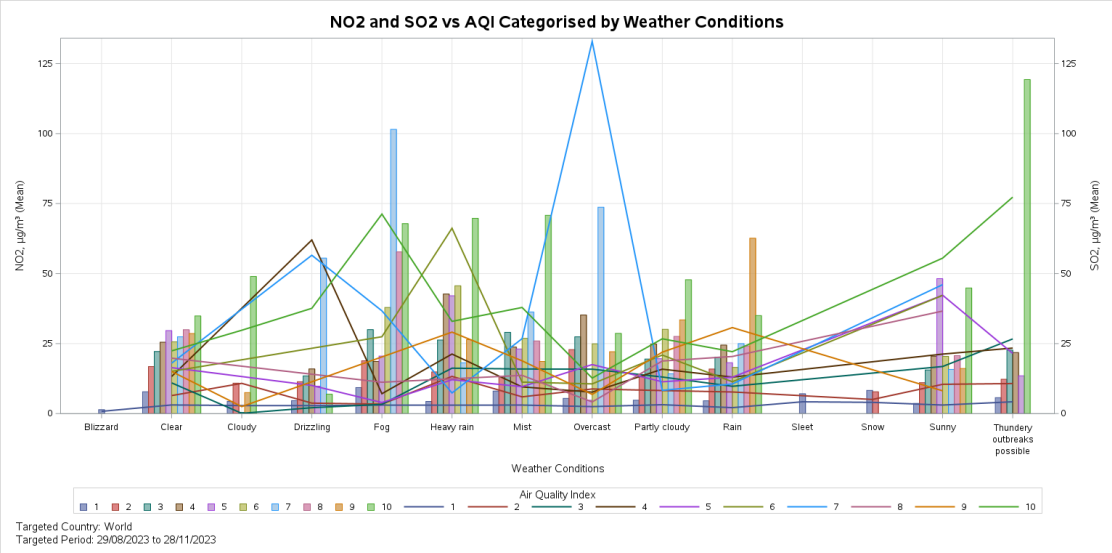


Observations:

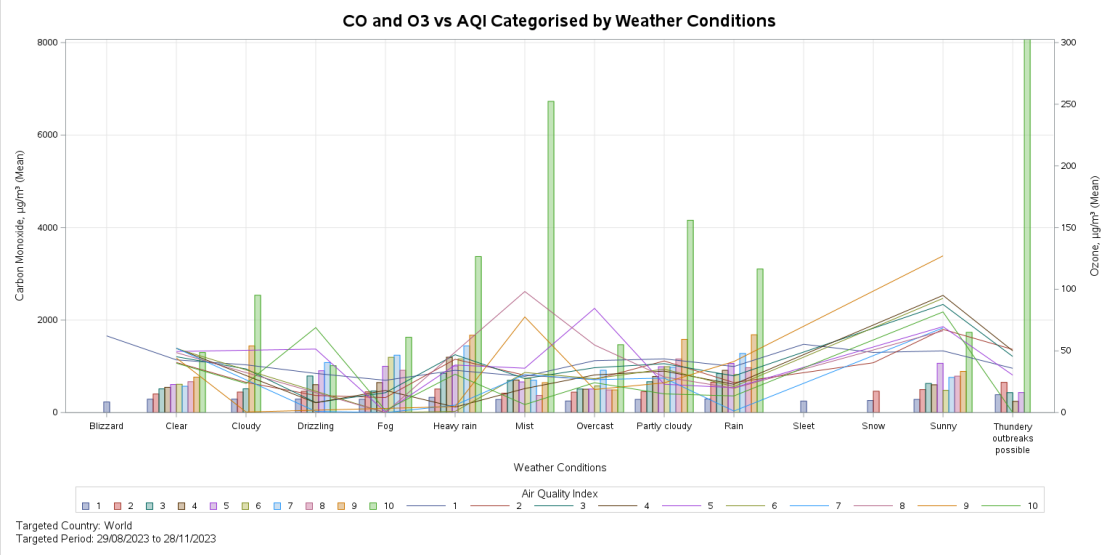
- The above figure uses longitude and latitude to pinpoint the AQI on 27 Nov 2023 of that location and then labelled with the country name. Based on the figure, both Asia and Africa continents have the highest AQI on 27 Nov 2023 as compared to North America, South America, Europe and Australia.
- For Asia continent, high AQI value are found at Western Asia, Eastern Asia and South-East Asia. For Africa continent, high AQI values are mainly found at West Africa and East Africa.

6.	Variable Used: PM10, PM2.5, Air Quality Index, Weather Conditions



	Observations: • The mean of PM10 concentration is positively correlated with the AQI value. Besides, the mean of PM10 concentration is also partially affected by the weather conditions. As examples, the mean of PM10 concentration is lower during drizzling, fog, heavy rain as compared to other weathers like clear, overcast and cloudy. • The mean of PM2.5 concentration is positively correlated with the AQI value. Besides, the mean of PM2.5 concentration is also partially affected by the weather conditions. As examples, the mean of PM2.5 concentration is lower during drizzling, fog, heavy rain as compared to other weathers like clear, overcast and cloudy. • Both mean for PM10 and PM2.5 concentration is found to be maximum when AQI=10 and the weather is thundery outbreaks possible.
7.	Variable Used: Nitrogen Dioxide, Sulphur Dioxide, Air Quality Index, Weather Conditions  Observations: • The mean of NO_2 concentration is generally positively correlated with the AQI value. Besides, the mean of NO_2 concentration is also partially affected by the weather conditions. As examples, the mean of NO_2 concentration is lower during clear, sunny, snow as compared to other weathers like cloudy, overcast, fog etc. • The mean of SO_2 concentration is generally positively correlated with the AQI value. Besides, there is not a significant impact on the mean of SO_2 concentration by the weather conditions due to the inconsistency of the data. • The peak for the mean of NO_2 concentration is found at AQI=10 and the weather condition are thundery outbreaks possible. • The peak for the mean of SO_2 concentration is found at AQI=7 and the weather conditions are overcast.

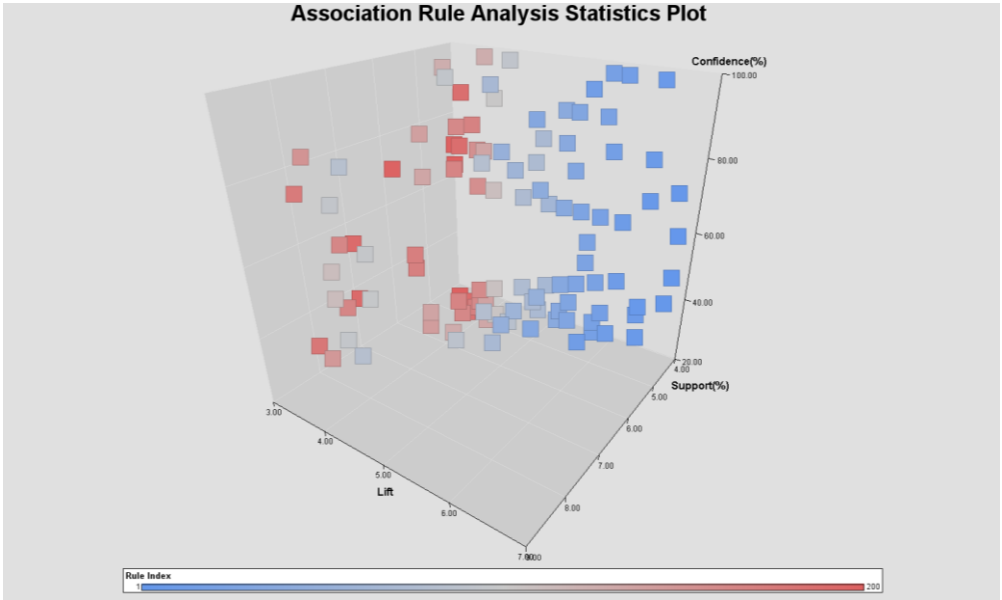
8.	Variable Used: Carbon Monoxide, Ozone, Air Quality Index, Weather Conditions



Observations:

- The mean of CO concentration is positively correlated with the AQI value. Besides, the mean of CO concentration is also partially affected by the weather conditions. As examples, the mean of CO concentration is lower during clear, sunny, drizzling as compared to other weathers like cloudy, heavy rain, mist, etc.
- The mean of O_3 concentration does not show any significant correlation with AQI as the trend is inconsistent. Besides, weather conditions have slight influence on the mean of O_3 concentration as higher O_3 concentration is found during sunny day as compared to the others.
- The peak for the mean of CO concentration is found at AQI=10 and the weather conditions are thundery outbreaks possible.
- The peak for the mean of O_3 concentration is found at AQI=9 and the weather condition are sunny.

9.	Association Rule Analysis Statistics Plot



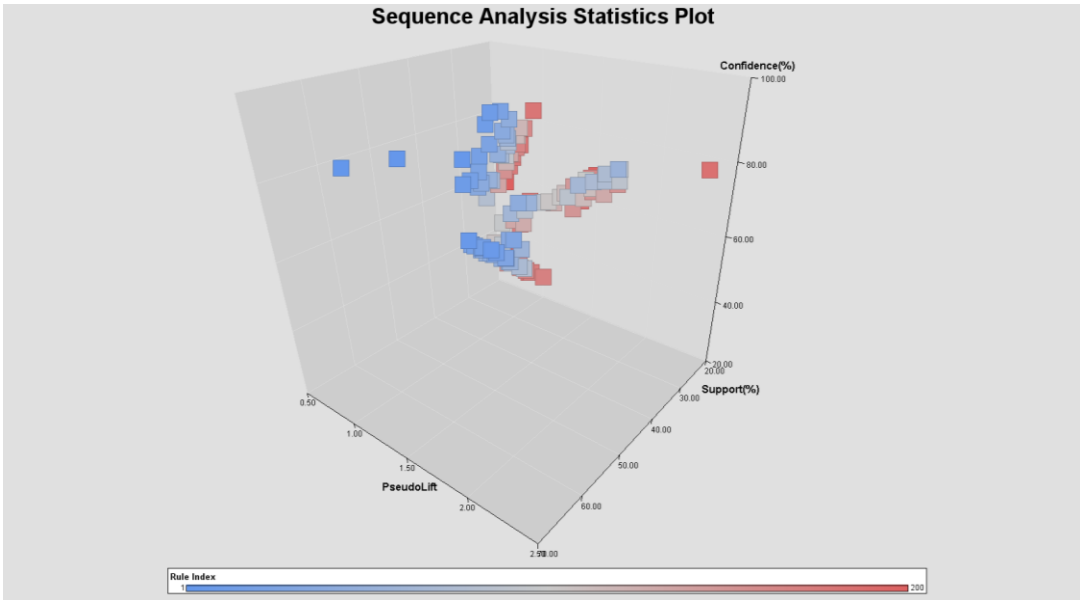
Rules Table

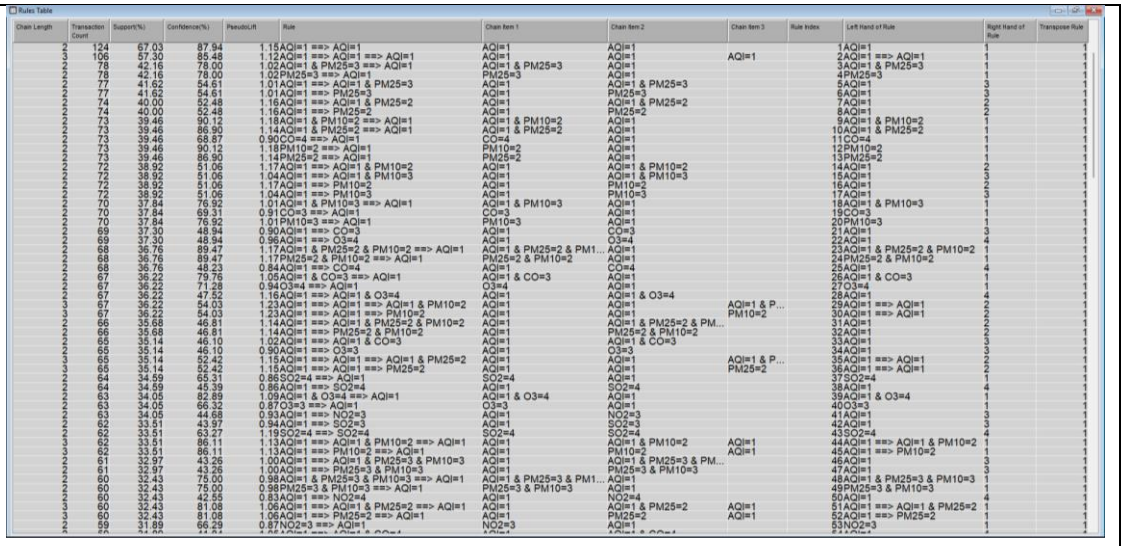
Relations	Expected Confidence (%)	Confidence (%)	Support (%)	Lift	Transaction Count	Rule	Left Hand of Rule	Right Hand of Rule	Rule Item 1	Rule Item 2	Rule Item 3	Rule Item 4	Rule Item 5	Rule Index	Transpose Rule
4	7.57	52.94	4.86	7.00	9.00	SO2=5 & PM10=5 & CO=5 ==> AQI=10	SO2=5 & PM10=5 & CO=5	AQI=10	SO2=5	PM10=5	CO=5	=====>	AQI=10	1	1
4	9.19	64.29	4.86	7.00	9.00	AQI=10 ==> SO2=5 & PM10=5 & CO=5	AQI=10	SO2=5 & PM10=5 & CO=5	AQI=10	=====>	SO2=5	PM10=5	CO=5	2	1
4	6.49	45.00	4.86	6.94	9.00	SO2=2 & PM10=2 ==> O3=3 & NO2=2	SO2=2 & PM10=2	O3=3 & NO2=2	SO2=2	PM10=2	=====>	O3=3	NO2=2	3	1
4	10.81	75.00	4.86	6.94	9.00	O3=3 & NO2=2 ==> SO2=2 & PM10=2	O3=3 & NO2=2	SO2=2 & PM10=2	O3=3	NO2=2	=====>	SO2=2	PM10=2	4	1
4	7.57	52.38	5.95	6.92	11.00	PM10=5 & NO2=5 & CO=5 ==> AQI=10	PM10=5 & NO2=5 & CO=5	AQI=10	PM10=5	NO2=5	CO=5	=====>	AQI=10	5	1
4	11.35	78.57	5.95	6.92	11.00	AQI=10 ==> PM10=5 & NO2=5 & CO=5	AQI=10	PM10=5 & NO2=5 & CO=5	AQI=10	=====>	PM10=5	NO2=5	CO=5	6	1
4	4.86	32.14	4.86	6.61	9.00	NO2=5 & CO=5 ==> SO2=5 & AQI=10	NO2=5 & CO=5	SO2=5 & AQI=10	NO2=5	CO=5	=====>	SO2=5	AQI=10	7	1
4	15.14	100.00	4.86	6.61	9.00	SO2=5 & AQI=10 ==> NO2=5 & CO=5	SO2=5 & AQI=10	NO2=5 & CO=5	SO2=5	AQI=10	=====>	NO2=5	CO=5	8	1
4	12.43	81.82	4.86	6.58	9.00	NO2=5 & AQI=10 ==> SO2=5 & CO=5	NO2=5 & AQI=10	SO2=5 & CO=5	NO2=5	AQI=10	=====>	SO2=5	CO=5	9	1
4	5.95	39.13	4.86	6.58	9.00	SO2=5 & CO=5 ==> NO2=5 & AQI=10	SO2=5 & CO=5	NO2=5 & AQI=10	SO2=5	CO=5	=====>	NO2=5	AQI=10	10	1
4	10.27	64.29	4.86	6.26	9.00	AQI=10 ==> PM10=5 & O3=2 & NO2=5	AQI=10	PM10=5 & O3=2 & NO2=5	AQI=10	=====>	PM10=5	O3=2	NO2=5	11	1
4	7.57	47.37	4.86	6.26	9.00	PM10=5 & O3=2 & NO2=5 ==> AQI=10	PM10=5 & O3=2 & NO2=5	AQI=10	PM10=5	O3=2	NO2=5	=====>	AQI=10	12	1
4	4.86	30.00	4.86	6.17	9.00	PM10=5 & CO=5 ==> SO2=5 & AQI=10	PM10=5 & CO=5	SO2=5 & AQI=10	PM10=5	CO=5	=====>	SO2=5	AQI=10	13	1
4	5.95	36.67	5.95	6.17	11.00	PM10=5 & CO=5 ==> NO2=5 & AQI=10	PM10=5 & CO=5	NO2=5 & AQI=10	PM10=5	CO=5	=====>	NO2=5	AQI=10	14	1
4	16.22	100.00	5.95	6.17	11.00	NO2=5 & AQI=10 ==> PM10=5 & CO=5	NO2=5 & AQI=10	PM10=5 & CO=5	NO2=5	AQI=10	=====>	PM10=5	CO=5	15	1
4	16.22	100.00	4.86	6.17	9.00	SO2=5 & AQI=10 ==> PM10=5 & CO=5	SO2=5 & AQI=10	PM10=5 & CO=5	SO2=5	AQI=10	=====>	PM10=5	CO=5	16	1
4	13.51	81.82	4.86	6.05	9.00	NO2=5 & AQI=10 ==> PM10=5 & O3=2	NO2=5 & AQI=10	PM10=5 & O3=2	NO2=5	AQI=10	=====>	PM10=5	O3=2	17	1
4	5.95	36.00	4.86	6.05	9.00	PM10=5 & O3=2 ==> NO2=5 & AQI=10	PM10=5 & O3=2	NO2=5 & AQI=10	PM10=5	O3=2	=====>	NO2=5	AQI=10	18	1
4	4.86	29.03	4.86	5.97	9.00	PM10=5 & NO2=5 ==> SO2=5 & AQI=10	PM10=5 & NO2=5	SO2=5 & AQI=10	PM10=5	NO2=5	=====>	SO2=5	AQI=10	19	1
4	16.76	100.00	4.86	5.97	9.00	SO2=5 & AQI=10 ==> PM10=5 & NO2=5	SO2=5 & AQI=10	PM10=5 & NO2=5	SO2=5	AQI=10	=====>	PM10=5	NO2=5	20	1
4	5.41	32.14	4.86	5.95	9.00	NO2=5 & CO=5 ==> O3=2 & AQI=10	NO2=5 & CO=5	O3=2 & AQI=10	NO2=5	CO=5	=====>	O3=2	AQI=10	21	1
4	10.81	64.29	4.86	5.95	9.00	AQI=10 ==> SO2=5 & NO2=5 & CO=5	AQI=10	SO2=5 & NO2=5 & CO=5	AQI=10	=====>	SO2=5	NO2=5	CO=5	22	1
4	7.57	45.00	4.86	5.95	9.00	SO2=5 & NO2=5 & CO=5 ==> AQI=10	SO2=5 & NO2=5 & CO=5	AQI=10	SO2=5	NO2=5	CO=5	=====>	AQI=10	23	1
4	15.14	90.00	4.86	5.95	9.00	O3=2 & AQI=10 ==> NO2=5 & CO=5	O3=2 & AQI=10	NO2=5 & CO=5	O3=2	AQI=10	=====>	NO2=5	CO=5	24	1
4	9.73	56.25	4.86	5.78	9.00	SO2=2 & O3=3 ==> PM10=2 & NO2=2	SO2=2 & O3=3	PM10=2 & NO2=2	SO2=2	O3=3	=====>	PM10=2	NO2=2	25	1
4	8.65	50.00	4.86	5.78	9.00	PM10=2 & NO2=2 ==> SO2=2 & O3=3	PM10=2 & NO2=2	SO2=2 & O3=3	PM10=2	NO2=2	=====>	SO2=2	O3=3	26	1
4	11.35	64.29	4.86	5.66	9.00	AQI=10 ==> PM10=5 & O3=2 & CO=5	AQI=10	PM10=5 & O3=2 & CO=5	AQI=10	=====>	PM10=5	O3=2	CO=5	27	1
4	7.57	42.86	4.86	5.66	9.00	PM10=5 & O3=2 & CO=5 ==> AQI=10	PM10=5 & O3=2 & CO=5	AQI=10	PM10=5	O3=2	CO=5	=====>	AQI=10	28	1
4	6.49	36.00	4.86	5.55	9.00	PM10=5 & O3=2 ==> CO=5 & AQI=10	PM10=5 & O3=2	CO=5 & AQI=10	PM10=5	O3=2	=====>	CO=5	AQI=10	29	1
4	5.41	30.00	4.86	5.55	9.00	PM10=5 & CO=5 ==> O3=2 & AQI=10	PM10=5 & CO=5	O3=2 & AQI=10	PM10=5	CO=5	=====>	O3=2	AQI=10	30	1
4	16.22	90.00	4.86	5.55	9.00	O3=2 & AQI=10 ==> PM10=5 & CO=5	O3=2 & AQI=10	PM10=5 & CO=5	O3=2	AQI=10	=====>	PM10=5	CO=5	31	1
4	13.51	75.00	4.86	5.55	9.00	CO=5 & AQI=10 ==> PM10=5 & O3=2	CO=5 & AQI=10	PM10=5 & O3=2	CO=5	AQI=10	=====>	PM10=5	O3=2	32	1
4	6.49	35.48	5.95	5.47	11.00	PM10=5 & NO2=5 ==> CO=5 & AQI=10	PM10=5 & NO2=5	CO=5 & AQI=10	PM10=5	NO2=5	=====>	CO=5	AQI=10	33	1
4	16.76	91.67	5.95	5.47	11.00	CO=5 & AQI=10 ==> PM10=5 & NO2=5	CO=5 & AQI=10	PM10=5 & NO2=5	CO=5	AQI=10	=====>	PM10=5	NO2=5	34	1
4	7.57	40.91	4.86	5.41	9.00	O3=2 & NO2=5 & CO=5 ==> AQI=10	O3=2 & NO2=5 & CO=5	AQI=10	O3=2	NO2=5	CO=5	=====>	AQI=10	35	1
4	5.95	32.14	4.86	5.41	9.00	SO2=5 & PM10=5 ==> NO2=5 & AQI=10	SO2=5 & PM10=5	NO2=5 & AQI=10	SO2=5	PM10=5	=====>	NO2=5	AQI=10	36	1
4	11.89	64.29	4.86	5.41	9.00	AQI=10 ==> O3=2 & NO2=5 & CO=5	AQI=10	O3=2 & NO2=5 & CO=5	AQI=10	=====>	O3=2	NO2=5	CO=5	37	1
4	15.14	81.82	4.86	5.41	9.00	NO2=5 & AQI=10 ==> SO2=5 & PM10=5	NO2=5 & AQI=10	SO2=5 & PM10=5	NO2=5	AQI=10	=====>	SO2=5	PM10=5	38	1
4	5.41	29.03	4.86	5.37	9.00	PM10=5 & NO2=5 ==> O3=2 & AQI=10	PM10=5 & NO2=5	O3=2 & AQI=10	PM10=5	NO2=5	=====>	O3=2	AQI=10	39	1
4	16.76	90.00	4.86	5.37	9.00	O3=2 & AQI=10 ==> PM10=5 & NO2=5	O3=2 & AQI=10	PM10=5 & NO2=5	O3=2	AQI=10	=====>	PM10=5	NO2=5	40	1
3	13.51	71.43	5.41	5.29	10.00	AQI=10 ==> PM10=5 & O3=2	AQI=10	PM10=5 & O3=2	AQI=10	=====>	PM10=5	O3=2		41	1
3	13.51	71.43	5.41	5.29	10.00	PM2_5= & AQI=10 ==> PM10=5 & O3=2	PM2_5= & AQI=10	PM10=5 & O3=2	PM2_5=	AQI=10	=====>	PM10=5	O3=2	42	1
3	13.51	71.43	5.41	5.29	10.00	AQI=10 ==> PM2_5= & PM10=5 & O3=2	AQI=10	PM2_5= & PM10=5 & O3=2	AQI=10	=====>	PM2_5=	PM10=5	O3=2	43	1
3	7.57	40.00	6.49	5.29	12.00	PM10=5 & CO=5 ==> AQI=10	PM10=5 & CO=5	AQI=10	PM10=5	CO=5	=====>	AQI=10		44	1
3	7.57	40.00	5.41	5.29	10.00	PM10=5 & O3=2 ==> AQI=10	PM10=5 & O3=2	AQI=10	PM10=5	O3=2	=====>	AQI=10		45	1
4	7.57	40.00	6.49	5.29	12.00	PM2_5= & PM10=5 & CO=5 ==> AQI=10	PM2_5= & PM10=5 & CO=5	AQI=10	PM2_5=	PM10=5	CO=5	=====>	AQI=10	46	1
4	7.57	40.00	5.41	5.29	10.00	PM2_5= & PM10=5 & O3=2 ==> AQI=10	PM2_5= & PM10=5 & O3=2	AQI=10	PM2_5=	PM10=5	O3=2	=====>	AQI=10	47	1
3	16.22	85.71	6.49	5.29	12.00	AQI=10 ==> PM10=5 & CO=5	AQI=10	PM10=5 & CO=5	AQI=10	=====>	PM10=5	CO=5		48	1
4	16.22	85.71	6.49	5.29	12.00	PM2_5= & AQI=10 ==> PM10=5 & CO=5	PM2_5= & AQI=10	PM10=5 & CO=5	PM2_5=	AQI=10	=====>	PM10=5	CO=5	49	1
4	7.57	40.00	6.49	5.29	12.00	PM10=5 & CO=5 ==> PM2_5= & AQI=10	PM10=5 & CO=5	PM2_5= & AQI=10	PM10=5	CO=5	=====>	PM2_5=	AQI=10	50	1

Observations:

- The association rules provide insights into how specific combinations of air quality parameters tend to co-occur. These associations can be valuable for understanding patterns and dependencies in the air quality data.
- Many rules have relatively high confidence and support, indicating that the associations are frequent, and the rules are well-supported by the data.
- The lift values, particularly in the range of 6.00 to 7.00, suggest strong positive correlations between the antecedent and consequent in the rules. This indicates that when certain conditions are present, there is a higher likelihood of the associated outcomes.
- The highest lift rule is Rule 1. When SO_2 is in category 5, PM10 is in category 5, and CO is in category 5, there is a 52.94% confidence that the AQI will be in category 10. The lift of 7.00 suggests a strong positive correlation. Conversely, when the AQI is in category 10, there is a 64.29% confidence that SO_2, PM10, and CO will all be in category 5. This shows that lift is symmetrical.

10. Sequence Analysis Statistics Plot





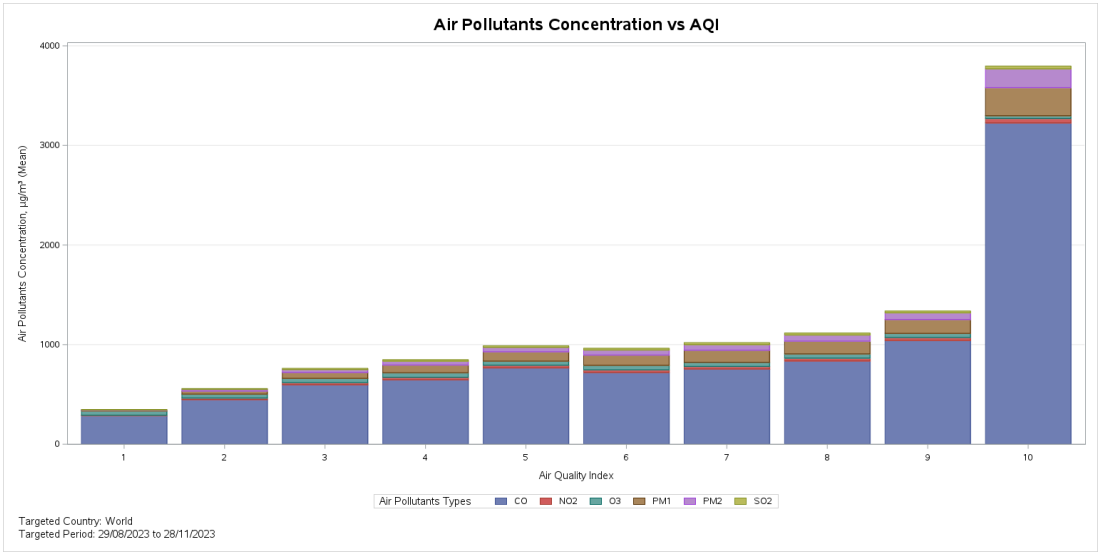
Chain Length	Transaction Count	Support(%)	Confidence(%)	PseudoLift	Rule	Chain Item 1	Chain Item 2	Chain Item 3	Rule Index	Left Hand of Rule	Right Hand of Rule	Transpose Rule
2	124	67.03	87.94	1.15	AQI=1 ==> AQI=1	AQI=1	AQI=1			1 AQI=1	1	1
3	106	57.30	85.48	1.12	AQI=1 ==> AQI=1 ==> AQI=1	AQI=1	AQI=1	AQI=1		2 AQI=1 ==> AQI=1	1	1
2	78	42.16	78.00	1.02	AQI=1 & PM25=3 ==> AQI=1	AQI=1 & PM25=3	AQI=1			3 AQI=1 & PM25=3	1	1
2	78	42.16	78.00	1.02	PM25=3 ==> AQI=1	PM25=3	AQI=1			4 PM25=3	1	1
2	77	41.62	54.61	1.01	AQI=1 ==> AQI=1 & PM25=3	AQI=1	AQI=1 & PM25=3			5 AQI=1	3	1
2	77	41.62	54.61	1.01	AQI=1 ==> PM25=3	AQI=1	PM25=3			6 AQI=1	3	1
2	74	40.00	52.48	1.16	AQI=1 ==> AQI=1 & PM25=2	AQI=1	AQI=1 & PM25=2			7 AQI=1	2	1
2	74	40.00	52.48	1.16	AQI=1 ==> PM25=2	AQI=1	PM25=2			8 AQI=1	2	1
2	73	39.46	90.12	1.18	AQI=1 & PM10=2 ==> AQI=1	AQI=1 & PM10=2	AQI=1			9 AQI=1 & PM10=2	1	1
2	73	39.46	86.90	1.14	AQI=1 & PM25=2 ==> AQI=1	AQI=1 & PM25=2	AQI=1			10 AQI=1 & PM25=2	1	1
2	73	39.46	68.87	0.90	CO=4 ==> AQI=1	CO=4	AQI=1			11 CO=4	1	1
2	73	39.46	90.12	1.18	PM10=2 ==> AQI=1	PM10=2	AQI=1			12 PM10=2	1	1
2	73	39.46	86.90	1.14	PM25=2 ==> AQI=1	PM25=2	AQI=1			13 PM25=2	1	1
2	72	38.92	51.06	1.17	AQI=1 ==> AQI=1 & PM10=2	AQI=1	AQI=1 & PM10=2			14 AQI=1	2	1
2	72	38.92	51.06	1.04	AQI=1 ==> AQI=1 & PM10=3	AQI=1	AQI=1 & PM10=3			15 AQI=1	3	1
2	72	38.92	51.06	1.17	AQI=1 ==> PM10=2	AQI=1	PM10=2			16 AQI=1	2	1
2	72	38.92	51.06	1.04	AQI=1 ==> PM10=3	AQI=1	PM10=3			17 AQI=1	3	1
2	70	37.84	76.92	1.01	AQI=1 & PM10=3 ==> AQI=1	AQI=1 & PM10=3	AQI=1			18 AQI=1 & PM10=3	1	1
2	70	37.84	69.31	0.91	CO=3 ==> AQI=1	CO=3	AQI=1			19 CO=3	1	1
2	70	37.84	76.92	1.01	PM10=3 ==> AQI=1	PM10=3	AQI=1			20 PM10=3	1	1
2	69	37.30	48.94	0.90	AQI=1 ==> CO=3	AQI=1	CO=3			21 AQI=1	3	1
2	69	37.30	48.94	0.96	AQI=1 ==> O3=4	AQI=1	O3=4			22 AQI=1	4	1
2	68	36.76	89.47	1.17	AQI=1 & PM25=2 & PM10=2 ==> AQI=1	AQI=1 & PM25=2 & PM1...	AQI=1			23 AQI=1 & PM25=2 & PM10=2	1	1
2	68	36.76	89.47	1.17	PM25=2 & PM10=2 ==> AQI=1	PM25=2 & PM10=2	AQI=1			24 PM25=2 & PM10=2	1	1
2	68	36.76	48.23	0.84	AQI=1 ==> CO=4	AQI=1	CO=4			25 AQI=1	4	1
2	67	36.22	79.76	1.05	AQI=1 & CO=3 ==> AQI=1	AQI=1 & CO=3	AQI=1			26 AQI=1 & CO=3	1	1
2	67	36.22	71.28	0.94	CO=3 ==> AQI=1	CO=3	AQI=1			27 CO=4	1	1
2	67	36.22	47.52	1.16	AQI=1 ==> AQI=1 & O3=4	AQI=1	AQI=1 & O3=4			28 AQI=1	4	1
3	67	36.22	54.03	1.23	AQI=1 ==> AQI=1 ==> AQI=1 & PM10=2	AQI=1	AQI=1	AQI=1 & P...		29 AQI=1 ==> AQI=1	2	1
3	67	36.22	54.03	1.23	AQI=1 ==> AQI=1 ==> PM10=2	AQI=1	AQI=1	PM10=2		30 AQI=1 ==> AQI=1	2	1
2	66	35.68	46.81	1.14	AQI=1 ==> AQI=1 & PM25=2 & PM10=2	AQI=1	AQI=1 & PM25=2 & PM...			31 AQI=1	2	1
2	66	35.68	46.81	1.14	AQI=1 ==> PM25=2 & PM10=2	AQI=1	PM25=2 & PM10=2			32 AQI=1	2	1
2	65	35.14	46.10	1.02	AQI=1 ==> AQI=1 & CO=3	AQI=1	AQI=1 & CO=3			33 AQI=1	3	1
2	65	35.14	46.10	0.90	AQI=1 ==> O3=3	AQI=1	O3=3			34 AQI=1	3	1
3	65	35.14	52.42	1.15	AQI=1 ==> AQI=1 ==> AQI=1 & PM25=2	AQI=1	AQI=1	AQI=1 & P...		35 AQI=1 ==> AQI=1	2	1
3	65	35.14	52.42	1.15	AQI=1 ==> AQI=1 ==> PM25=2	AQI=1	AQI=1	PM25=2		36 AQI=1 ==> AQI=1	2	1
2	64	34.59	65.31	0.86	SO2=4 ==> AQI=1	SO2=4	AQI=1			37 SO2=4	1	1
2	64	34.59	45.39	0.86	AQI=1 ==> SO2=4	AQI=1	SO2=4			38 AQI=1	4	1
2	63	34.05	82.89	1.09	AQI=1 & O3=4 ==> AQI=1	AQI=1 & O3=4	AQI=1			39 AQI=1 & O3=4	1	1
2	63	34.05	66.32	0.87	O3=3 ==> AQI=1	O3=3	AQI=1			40 O3=3	1	1
2	63	34.05	44.68	0.93	AQI=1 ==> NO2=3	AQI=1	NO2=3			41 AQI=1	3	1
2	62	33.51	43.97	0.94	AQI=1 ==> SO2=3	AQI=1	SO2=3			42 AQI=1	3	1
2	62	33.51	63.27	1.19	SO2=4 ==> SO2=4	SO2=4	SO2=4			43 SO2=4	4	1
3	62	33.51	86.11	1.13	AQI=1 ==> AQI=1 & PM10=2 ==> AQI=1	AQI=1	AQI=1 & PM10=2	AQI=1		44 AQI=1 ==> AQI=1 & PM10=2	1	1
3	62	33.51	86.11	1.13	AQI=1 ==> PM10=2 ==> AQI=1	AQI=1	PM10=2	AQI=1		45 AQI=1 ==> PM10=2	1	1
2	61	32.97	43.26	1.00	AQI=1 ==> AQI=1 & PM25=3 & PM10=3	AQI=1	AQI=1 & PM25=3 & PM...			46 AQI=1	3	1
2	61	32.97	43.26	1.00	AQI=1 ==> PM25=3 & PM10=3	AQI=1	PM25=3 & PM10=3			47 AQI=1	3	1
2	60	32.43	75.00	0.98	AQI=1 & PM25=3 & PM10=3 ==> AQI=1	AQI=1 & PM25=3 & PM1...	AQI=1			48 AQI=1 & PM25=3 & PM10=3	1	1
2	60	32.43	75.00	0.98	PM25=3 & PM10=3 ==> AQI=1	PM25=3 & PM10=3	AQI=1			49 PM25=3 & PM10=3	1	1
2	60	32.43	42.55	0.83	AQI=1 ==> NO2=4	AQI=1	NO2=4			50 AQI=1	4	1
3	60	32.43	81.08	1.06	AQI=1 ==> AQI=1 & PM25=2 ==> AQI=1	AQI=1	AQI=1 & PM25=2	AQI=1		51 AQI=1 ==> AQI=1 & PM25=2	1	1
3	60	32.43	81.08	1.06	AQI=1 ==> PM25=2 ==> AQI=1	AQI=1	PM25=2	AQI=1		52 AQI=1 ==> PM25=2	1	1
2	59	31.89	66.29	0.87	NO2=3 ==> AQI=1	NO2=3	AQI=1			53 NO2=3	1	1

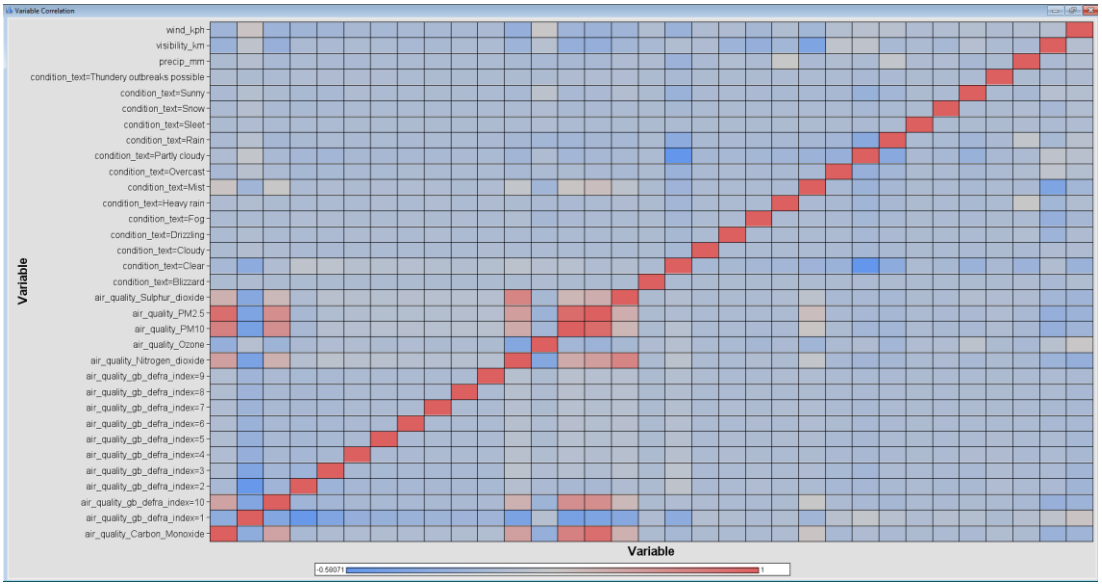
Observation:

- The sequence analysis helps uncover patterns in the temporal ordering of air quality parameter occurrences.
- The confidence, support and lift for many of the rules changes after time ID is considered. In Sequence 1 ("AQI=1 ==> AQI=1"), when the AQI is in category 1, there is an 87.94% confidence that the next occurrence will also have the AQI in category 1. The pseudo lift of 1.15 in Sequence 1 suggests that the occurrence of AQI=1 in the first position increases the likelihood of AQI=1 in the subsequent position by approximately 15% compared to its overall likelihood. The support of 67.03% for Sequence 1 indicates that this sequence is present in approximately 67.03% of the transactions (sequences) in the dataset.

More than Four Variables

11. Variables Used: AQI, Carbon Monoxide, Ozone, Nitrogen Dioxide, Sulphur Dioxide, PM2.5 and PM10.



	Observations: • The total air pollutants concentration contributed by the CO, NO_2, O_3, PM10, PM2.5 and SO_2 is positively correlated with the AQI. This is shown by the upward trend in air pollutants concentration when the AQI increases. • The major contributions to the total air pollutants concentration are CO and followed by PM10 and PM2.5 across the AQI. • The concentration of NO_2, SO_2 and $O_3$is roughly the same or only marginal increase in the concentration across the AQI in comparison to the concentration of CO, PM10 and PM2.5. • There is a spike increase in the total air pollutants concentration when AQI is up by 1 from AQI=9 to AQI=10.
12.	Variables Correlation Matrix  The above figure shows the correlation matrix between the interested variables. A table which shows the correlation value of all variables is attached in the appendix.

6.5 Summary of Key Findings from Exploration

- From univariate analysis (histogram) and bivariate analysis (boxplots), we found that all pollutants' concentration variables (CO, NO_2, SO_2, O_3, PM2.5, and PM10) contain outliers but it may not necessarily warrant their removal. Outliers in air quality data may represent genuine environmental variations.
- From the pie chart analysis, weather condition with the most occurrence in the period of 28/08/2023 to 29/11/2023 is partly cloudy, followed by clear and rain.
- From bivariate analysis – Chart No.7 and No.8, wind speed and visibility show negative correlation with the AQI.
- From multivariate analysis – Chart No.1, No.2 and No.4, variables like temperature, cloud cover and humidity do not show any correlation with AQI.
- From multivariate analysis – Chart No.3, it shows that precipitation amount has negative correlation with the AQI.
- From multivariate analysis – Chart No.6, No.7 and No.8, variables like CO, NO_2, SO_2, PM2.5, and PM10 have positive correlation with AQI with the exception for O_3. Besides, they are also partially affected by weather conditions.

7 Modify

7.1 Data Cleaning

7.1.1 Finding Missing values

In SAS Enterprise, we have utilized the "StatExplore" component to check if there are any missing values in the variables as shown in Figure 7.1. Referring to Figure 7.1, there are a total of three variables with missing values as shown below: -

- condition_text: 22 missing values
- pressure_mb: 9 missing values.
- Visibility_km: 48 missing values.

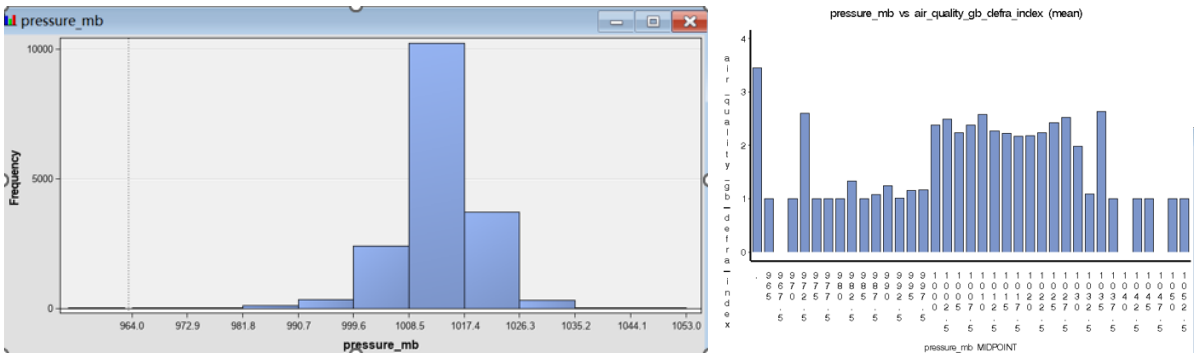
Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage		
TRAIN	condition_text	INPUT	15	22	Partly cloudy	39.34	Clear	34.08		
	pressure_mb	INPUT	1013.218	6.689865	17140	9	964	1013	1053	-0.60629	3.484118
	temperature_celsius	INPUT	20.84551	8.491842	17149	0	-31	22.5	45.4	-0.80762	0.890103
	uv_index	INPUT	2.223395	2.247711	17149	0	1	1	11	1.451918	0.465081
	visibility_km	INPUT	9.729285	2.501822	17101	48	0.1	10	32	2.25408	18.27215

Figure 7.1 The above figures show that there are missing values found in condition_text, pressure_mb, and visibility_km

7.1.2 Handing missing values

In this section, we will describe the techniques employed for the imputation of missing values of each variable.

7.1.2.1 pressure_mb



Firstly, we analyze the distribution of the 'pressure_mb' variable. This distribution appears to be relatively uniform and central, indicating that the missing values occur at random. Next, we investigated its correlation with the target variable and found that 'pressure_mb' does not exhibit a direct relationship with either the target variable or any other variables in the dataset. This lack of correlation is crucial for determining the appropriate method for imputing its missing values.

- Missing value type

The missing values in "pressure_mb" satisfies the Missing Completely at Random (MCAR) – the missingness of a data point is completely unrelated to any observed or unobserved variable. And the reason for the missing value is purely random.

- Imputation Method

Since the distribution of the data closely resembles a normal distribution, we have the option to use either the median or the mean for imputation. For “pressure_mb” variable, we have chosen to use the mean imputation method as shown in Figure 7.2 and the result is shown in Figure 7.3.

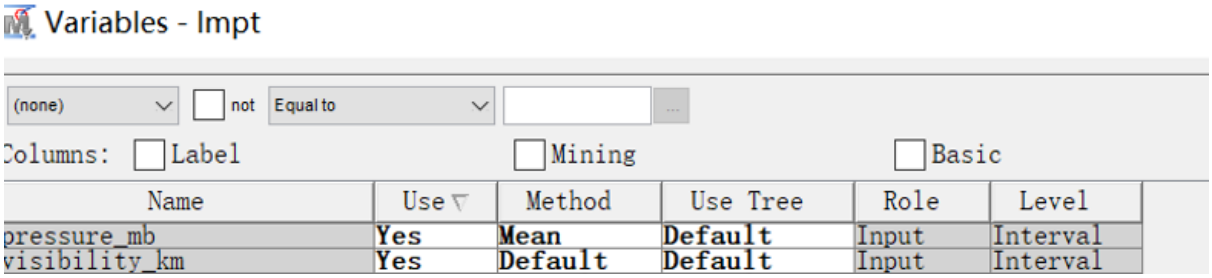


Figure 7.2 Mean imputation for "pressure_mb" variable.

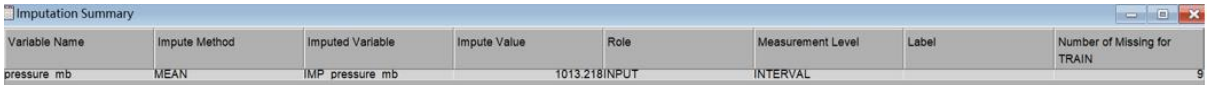
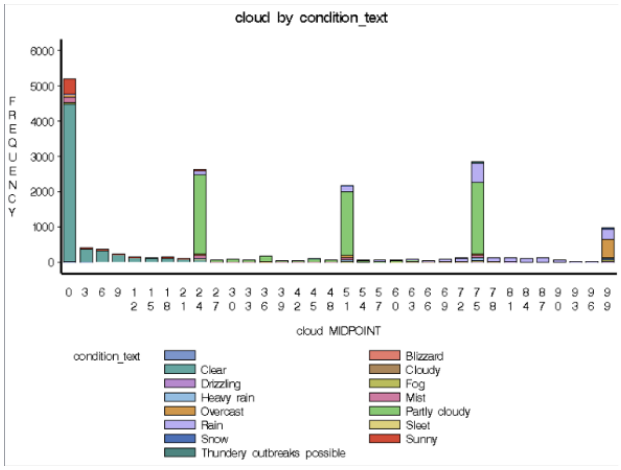
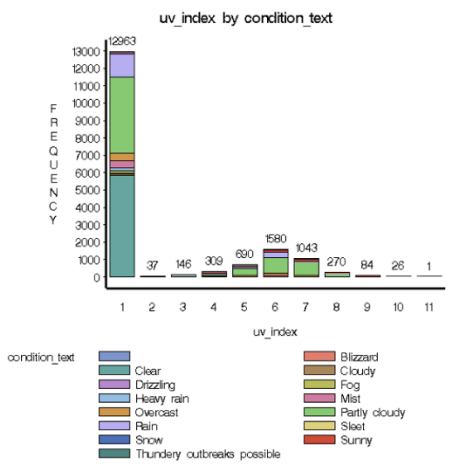


Figure 7.3 Results for "pressure_mb" variable after mean imputation.

7.1.2.2 condition_text





Upon analyzing the “condition_text” variable correlation with cloud cover and the UV index, it was observed that the missing value shows a significant relation to both cloud cover and the UV index. However, it does not show any relationship with the missing value's own characteristics.

- Missing Value Type

It meets the definition of Missing at Random (MAR) – the missingness of a data point is related to some observed variables but not the missing value itself.

- Imputation Method

Since the missing values align with the attributes typically associated with "sunny" conditions, we have opted to use values indicative of "sunny" weather for the imputation process.

Replacement Editor-WORK.OUTCLASS

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric
condition_text	Partly cloudy		6746	C	Partly cloudy	.
condition_text	Clear		5845	C	Clear	.
condition_text	Rain		1918	C	Rain	.
condition_text	Overcast		749	C	Overcast	.
condition_text	Sunny		718	C	Sunny	.
condition_text	Mist		478	C	Mist	.
condition_text	Heavy rain		242	C	Heavy rain	.
condition_text	Fog		135	C	Fog	.
condition_text	Cloudy		98	C	Cloudy	.
condition_text	Snow		78	C	Snow	.
condition_text	Drizzling		74	C	Drizzling	.
condition_text	Thundery outbreaks possible		37	C	Thundery outbreaks possible	.
condition_text		Sunny	22	C		.
condition_text	Sleet		8	C	Sleet	.

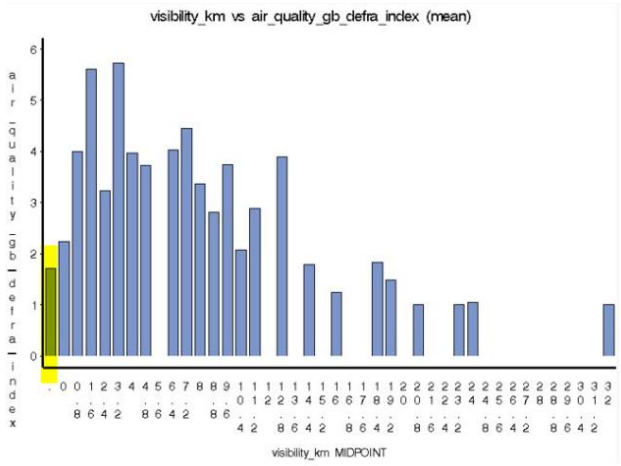
Figure 7.4 Replaceing missing value with "Sunny" for "condition_text" variable.

Total Replacement Counts

Variable	Role	Label	Train
condition_text	INPUT		22

Figure 7.5 Results for "condition_text" variable after imputation.

7.1.2.3 visibility_km



Based on the relationship between the visibility displayed in the picture and the target attribute, as well as the comprehensive consideration of the missing values corresponding to other attribute values.We discovered that the air quality values corresponding to the missing data are exceptionally low. This suggests that the missing variable instances occur under conditions of extremely low visibility, likely below the currently recorded minimum values. This could be attributed to limitations in equipment or other factors that hinder detection at certain low visibility levels.

- Missing Value Type

It meets the definition of Missing Not at Random (MNAR) – the missingness of a data point is related to the value that's missing itself, possibly in combination with other observed variables. In the project, it is in combination with other observed variables.

- Imputation Method

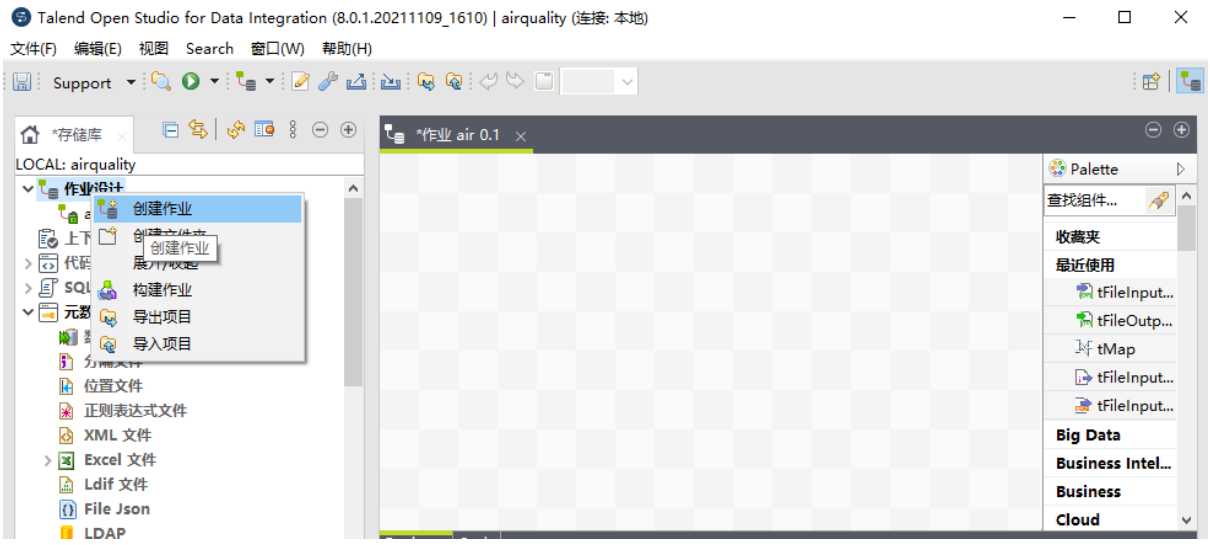
Considering that the minimum value of visibility is 0.1, and the missing value is less than the minimum value, we use "0" for imputing.

7.1.3 Checking Data Quality using Talend Data Integration

In this section, we will use Talend Data Integration to detect the format of all data to ensure that all data conforms to our predefined pattern.

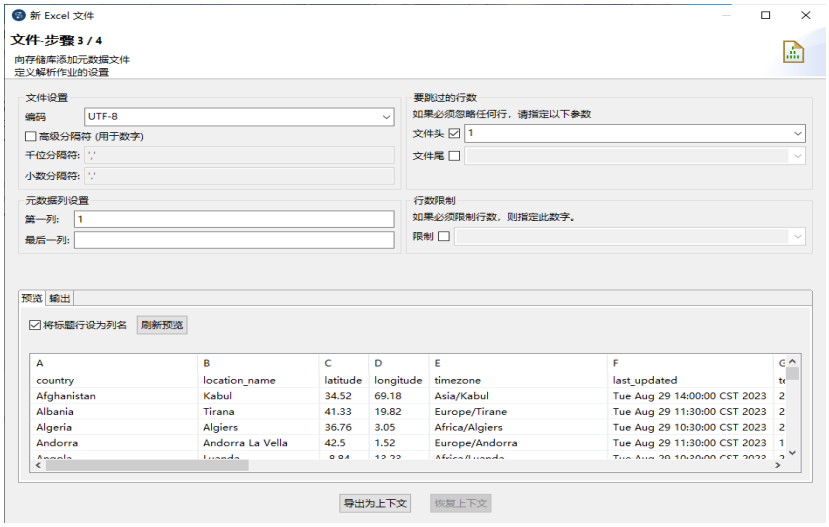
- Data input

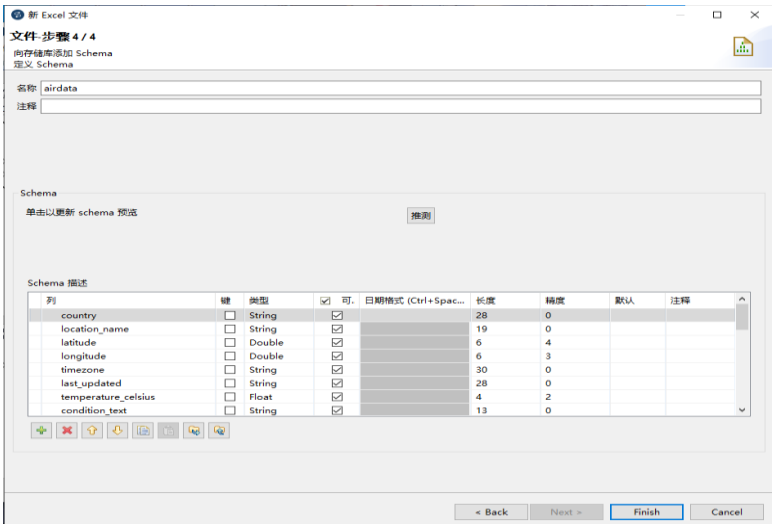
First, we create a new job.



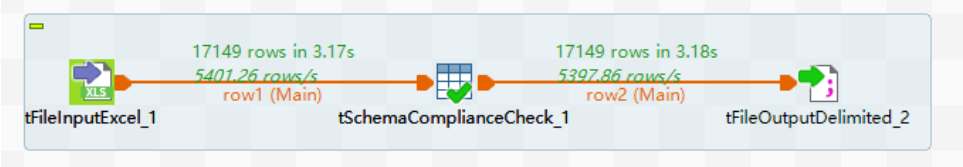
Second, we select a type of document according to our dataset.

Third, we input our dataset.

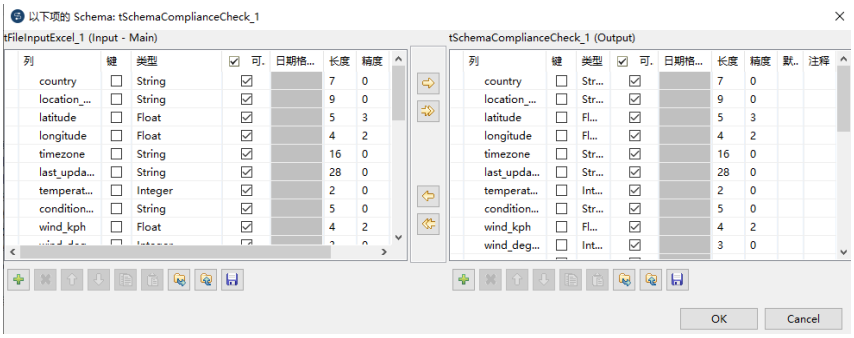




- Workflow



- Edit Attributes



- Result

All attributes meet the modeling standards and can be used for the feature engineering.

7.2 Feature Engineering

According to the previous sampling phase in SAS Enterprise Miner, the role and level of the attributes from the selected dataset are as shown in the figure. To perform feature engineering on this dataset, we have decided to use two different tools – SAS Enterprise Miner and KNIME.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
air_quality_Carbon_Monoxide	Input	Interval	No		No	.	.
air_quality_Nitrogen_dioxide	Input	Interval	No		No	.	.
air_quality_Ozone	Input	Interval	No		No	.	.
air_quality_PM10	Input	Interval	No		No	.	.
air_quality_PM2_5	Input	Interval	No		No	.	.
air_quality_Sulphur_dioxide	Input	Interval	No		No	.	.
air_quality_gb_defra_index	Target	Nominal	No		No	.	.
cloud	Input	Interval	No		No	.	.
condition_text	Input	Nominal	No		No	.	.
country	Rejected	Nominal	No		No	.	.
feels_like_celsius	Input	Interval	No		No	.	.
gust_kph	Input	Interval	No		No	.	.
humidity	Input	Interval	No		No	.	.
last_updated	Time ID	Interval	No		No	.	.
latitude	Input	Interval	No		No	.	.
location_name	Text	Nominal	No		No	.	.
longitude	Input	Interval	No		No	.	.
moon_illumination	Input	Interval	No		No	.	.
moon_phase	Rejected	Nominal	No		No	.	.
precip_mm	Input	Interval	No		No	.	.
pressure_mb	Input	Interval	No		No	.	.
temperature_celsius	Input	Interval	No		No	.	.
timezone	Rejected	Nominal	No		No	.	.
uv_index	Input	Nominal	No		No	.	.
visibility_km	Input	Interval	No		No	.	.
wind_degree	Input	Interval	No		No	.	.
wind_kph	Input	Interval	No		No	.	.

Figure 7.6 The metadata of the attributes from the used dataset.

7.2.1 SAS Enterprise Miner for feature engineering

7.2.1.1 Data integrity confirmation

- Integrity confirmation

By using the "StatExplore" component, we can check the data integrity via the summary statistics. Referring to Figure 7.7 and Figure 7.8, all the attributes from the dataset are free from missing values.

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
air_quality_Carbon_Monoxide	INPUT	568.364	1436.985	17149	0	96.8	283.7	36315.9	12.73704	208.8085
air_quality_Nitrogen_dioxide	INPUT	12.7675	22.65928	17149	0	0	4.4	337.2	4.175877	26.51848
air_quality_Ozone	INPUT	40.75872	32.20778	17149	0	0	37.2	555	1.763954	10.44748
air_quality_PM10	INPUT	41.42908	99.53617	17149	0	0.5	12.9	2504.3	8.240322	103.9945
air_quality_PM2_5	INPUT	24.1824	67.3488	17149	0	0.5	7.6	1558.8	9.536901	122.8243
air_quality_Sulphur_dioxide	INPUT	7.091737	16.55664	17149	0	0	1.7	335.7	6.153057	60.47203
cloud	INPUT	35.35588	33.08028	17149	0	0	25	100	0.425445	-1.19751
feels_like_celsius	INPUT	22.19765	10.88277	17149	0	-36.8	24.6	73.6	-0.47493	0.641658
gust_kph	INPUT	17.53299	11.28636	17149	0	0	15.3	110.5	1.226396	2.286825
humidity	INPUT	72.48195	20.29754	17149	0	4	77	100	-0.93965	0.289414
latitude	INPUT	19.30084	24.58315	17149	0	-41.3	17.25	63.83	-0.30613	-0.7684
longitude	INPUT	21.90824	65.69636	17149	0	-175.2	23.24	179.22	0.00864	0.334742
moon_illumination	INPUT	50.89335	35.13772	17149	0	0	49	100	-0.01611	-1.48828
precip_mm	INPUT	0.161076	0.748877	17149	0	0	0	31	17.24163	480.749
pressure_mb	INPUT	1013.218	6.688681	17149	0	964	1013	1053	-0.60618	3.485423
temperature_celsius	INPUT	20.84551	8.491842	17149	0	-31	22.5	45.4	-0.80762	0.890103
visibility_km	INPUT	9.702548	2.548837	17149	0	0	10	32	2.0093	17.32982
wind_degree	INPUT	162.189	105.521	17149	0	1	150	360	0.230127	-1.12867
wind_kph	INPUT	11.06719	7.945606	17149	0	3.6	9	141.1	2.079137	10.39417

Figure 7.7 Summary statistics for interval variable.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	condition_text	INPUT	14	0	Partly cloudy	39.34	Clear	34.21
TRAIN	uv_index	INPUT	11	0	1	75.59	6	9.21
TRAIN	air_quality_gb_defra_index	TARGET	10	0	1	60.97	2	16.62

Figure 7.8 Summary statistics for class variable.

7.2.1.2 Features selection

To perform the feature selection, we are using the "Variable Selection" component and set up the connection as shown in Figure 7.9.

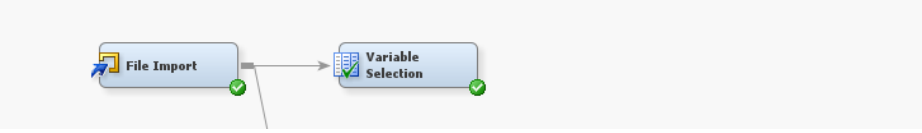


Figure 7.9 Setting up the connection between component for features selection.

- The relationship between each variable and the target variable.

(1) Sequential R-Squared graph: it is used to display the incremental contribution of variables when they are included in the model. The R-Squared value is a statistical metric that measures the goodness of fit of the model; it ranges from 0 to 1, with higher values indicating a better fit of the model to the data.

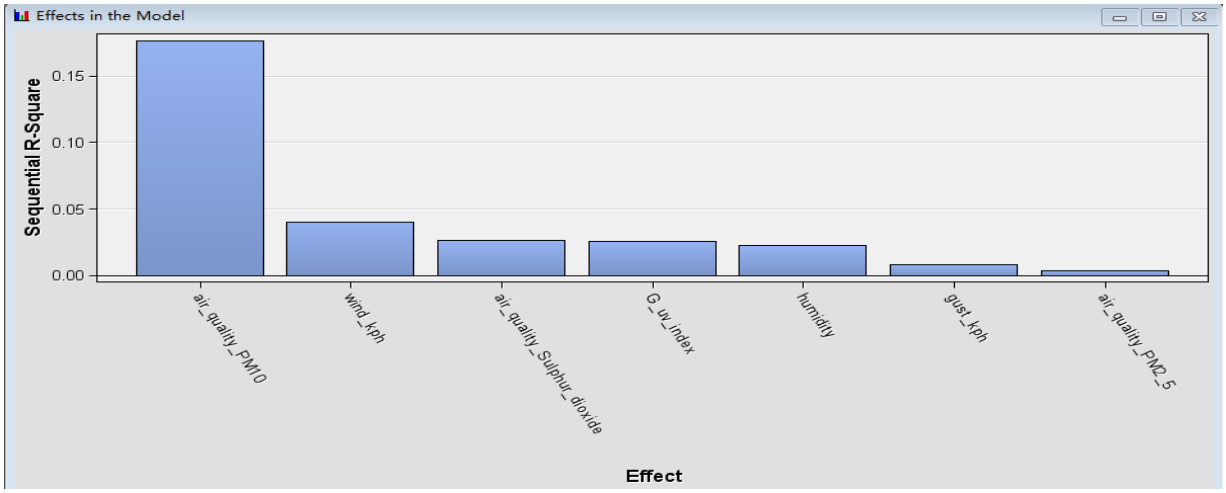


Figure 7.10 Sequential R-squared graph to show the contribution of variables to the AQI.

Findings from Figure 7.10:

- air_quality_PM10: this variable, once added to the model, resulted in the largest increase in R-Squared value, indicating its significant importance in predicting the target variable.
- wind_kph: following closely is wind speed （measured in kilometers per hour）, which also makes a relatively large contribution to the model.
- Other variables: such as air_quality_Sulphur_dioxide, uv_index, humidity, gust_kph, and air_quality_PM2.5, contribute less significantly, but their contributions are still positive, indicating that they also enhance the model's predictive capability to a certain extent.

(2) R-Squared values graph: it is a statistical tool that visually represents the explanatory power of independent variables in a regression model, guiding the model-building process by quantifying the contribution of each variable to the model's predictive capability in relation to the dependent variable.

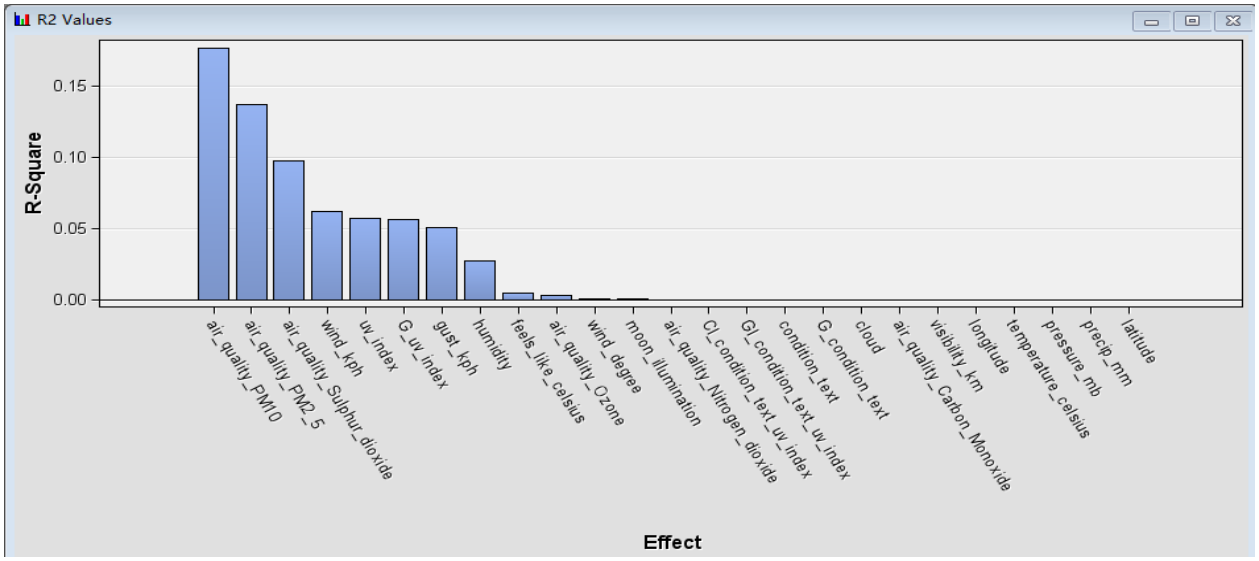


Figure 7.11 The R2 scores between all the variables and AQI.

Findings from Figure 7.11:

- **air_quality_PM10:** shows the highest R-Squared value among the variables, suggesting that PM10 levels have the strongest linear relationship with the target variable, which may be an air quality index. It significantly contributes to the model, explaining a good portion of the variance in the target.
- **air_quality_PM2.5 and wind_kph:** indicating that the particulate matter with a diameter of 2.5 micrometers or less and the wind speed are also important predictors for the target variable. They have a substantial but lesser contribution compared to PM10.
- **The following variables:** like uv_index, gust_kph, air_quality_Sulphur_dioxide, and others contribute incrementally less to the R-Squared value. However, their presence in the model still accounts for some of the variability in the target variable, indicating that while they are less important than PM10 or PM2.5, they do have a role in predicting air quality.
- **Variables toward the end of the chart:** such as longitude, latitude, visibility_km, and air_quality_Carbon_Monoxide, have the smallest bars, which means their contribution to the explained variance in the target variable is minimal in this particular model.

(3) The variable selection process: it is a crucial step in model building, where the goal is to identify the most relevant predictors out of possibly many variables. This process helps in simplifying the model, improving its interpretability, and potentially enhancing its predictive performance by avoiding overfitting.

Variable Selection

Variable Name	Role	Measurement Level	Type	Label	Reasons for Rejection
G_uv_index	Input	Nominal	Numeric	Grouped Levels for uv_ind...	
air_quality_Carbon_Monox...	Rejected	Interval	Numeric	air_quality_Carbon_Monox...	Varsel:Small R-square val...
air_quality_Nitrogen_dioxide	Rejected	Interval	Numeric	air_quality_Nitrogen_dioxide	Varsel:Small R-square val...
air_quality_Ozone	Rejected	Interval	Numeric	air_quality_Ozone	Varsel:Small R-square val...
air_quality_PM10	Input	Interval	Numeric	air_quality_PM10	
air_quality_PM2_5	Input	Interval	Numeric	air_quality_PM2.5	
air_quality_Sulphur_dioxide	Input	Interval	Numeric	air_quality_Sulphur_dioxide	
cloud	Rejected	Interval	Numeric	cloud	Varsel:Small R-square val...
condition_text	Rejected	Nominal	Character	condition_text	Varsel:Small R-square val...
feels_like_celsius	Rejected	Interval	Numeric	feels_like_celsius	Varsel:Small R-square val...
gust_kph	Input	Interval	Numeric	gust_kph	
humidity	Input	Interval	Numeric	humidity	
latitude	Rejected	Interval	Numeric	latitude	Varsel:Small R-square val...
longitude	Rejected	Interval	Numeric	longitude	Varsel:Small R-square val...
moon_illumination	Rejected	Interval	Numeric	moon_illumination	Varsel:Small R-square val...
precip_mm	Rejected	Interval	Numeric	precip_mm	Varsel:Small R-square val...
pressure_mb	Rejected	Interval	Numeric	pressure_mb	Varsel:Small R-square val...
temperature_celsius	Rejected	Interval	Numeric	temperature_celsius	Varsel:Small R-square val...
uv_index	Rejected	Nominal	Numeric	uv_index	Varsel:Small R-square val...
visibility_km	Rejected	Interval	Numeric	visibility_km	Varsel:Small R-square val...
wind_degree	Rejected	Interval	Numeric	wind_degree	Varsel:Small R-square val...
wind_kph	Input	Interval	Numeric	wind_kph	

Figure 7.12 Results from variable selection.

Findings from Figure 7.12:

- **Variables such as air_quality_PM10, air_quality_PM2.5, and wind_kph** were included as inputs in the model because they were likely found to have a significant contribution to the model's ability to predict the dependent variable, which, in this case, could be related to air quality measures.
- **Variables like uv_index, temperature_celsius, and visibility_km** were rejected due to their small contribution to the model's R-Squared value, indicating that they do not add much predictive value in the context of the other variables included.

(4) Effect table: starting with the most significant variable at the top. This particular table is sorted by the R-Squared values in descending order, meaning variables at the top of the list have a greater impact on the model than those at the bottom. The variables and classes with "R2 < MINR2" next to them indicate that these effects have an R-Squared value lower than a minimum threshold set within the model （MINR2）, suggesting that they have a relatively low explanatory power and may not be useful in the final model.

	Effect	DF	R-Square				
72				96	AOV16: feels_like_celsius	15	0.051829	
73				97	AOV16: gust_kph	12	0.051559	
74	Var: air_quality_PM10	1	0.176859	98	Var: gust_kph	1	0.050226	
75	AOV16: air_quality_Nitrogen_dioxide	14	0.161902	99	AOV16: air_quality_Carbon_Monoxide	15	0.042839	
76	Var: air_quality_Nitrogen_dioxide	1	0.157689	100	AOV16: humidity	15	0.037253	
77	AOV16: latitude	15	0.139778	101	AOV16: temperature_celsius	15	0.036606	
78	Var: air_quality_PM2_5	1	0.136582	102	Var: humidity	1	0.027352	
79	AOV16: longitude	14	0.122149	103	AOV16: visibility_km	13	0.024105	
80	Class: condition_text*uv_index	84	0.118665	104	Var: visibility_km	1	0.017810	
81	Group: condition_text*uv_index	5	0.116674	105	Var: longitude	1	0.017045	
82	Var: air_quality_Sulphur_dioxide	1	0.096968	106	AOV16: wind_degree	15	0.016886	
83	AOV16: air_quality_PM10	13	0.092310	107	AOV16: pressure_mb	15	0.013387	
84	Class: condition_text	13	0.088302	108	Var: temperature_celsius	1	0.008782	
85	Group: condition_text	3	0.086952	109	Var: pressure_mb	1	0.007200	
86	AOV16: air_quality_Sulphur_dioxide	13	0.086458	110	Var: feels_like_celsius	1	0.004165	R2 < MINR2
87	AOV16: cloud	15	0.083491	111	Var: air_quality_Ozone	1	0.003202	R2 < MINR2
88	AOV16: air_quality_Ozone	12	0.074836	112	Var: precip_mm	1	0.002812	R2 < MINR2
89	Var: cloud	1	0.071037	113	Var: latitude	1	0.002218	R2 < MINR2
90	AOV16: air_quality_PM2_5	14	0.070932	114	AOV16: precip_mm	12	0.001534	R2 < MINR2
91	Var: wind_kph	1	0.061782	115	AOV16: moon_illumination	15	0.001168	R2 < MINR2
92	Var: air_quality_Carbon_Monoxide	1	0.060258	116	Var: wind_degree	1	0.000133	R2 < MINR2
93	AOV16: wind_kph	10	0.059797	117	Var: moon_illumination	1	0.000103	R2 < MINR2
94	Class: uv_index	10	0.056940					
95	Group: uv_index	2	0.056001					

Figure 7.13 Results of Effect Table checking whether the variables are "R2 < MINR2".

Findings from Figure 7.13:

- **High R-Squared Variables:** variables such as "air_quality_PM10" and "air_quality_Nitrogen_dioxide" have relatively high R-Squared values, suggesting they are strong predictors for the model and explain a significant portion of the variance in the dependent variable.
- **Variables with Degrees of Freedom Greater Than One:** for categorical variables like "condition_text" with a DF of 84, the R-Squared value is quite high, indicating that this categorical variable significantly influences the dependent variable's variance across its many categories.
- **Low R-Squared Variables:** towards the bottom of the list, variables such as "precip_mm" and "moon_illumination" have very low R-Squared values, which means they explain a very small amount of variance in the dependent variable and might not be valuable in the predictive model.

(5) Effects Chosen for Target: this directly guides which variables should be considered in the modeling process. Within the context of modeling, variables that possess both a high R-Squared value and significant F values and p-values are typically selected, as they provide the most information and are most likely to enhance the predictive accuracy of the model.

The DMINE Procedure

Effects Chosen for Target: _DUMMY_TARGET_

Effect		DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Var:	air_quality_PM10	1	0.176859	3684.178450	<.0001	721.768233	0.195910
Var:	wind_kph	1	0.040291	882.458052	<.0001	164.429880	0.186332
Var:	air_quality_Sulphur_dioxide	1	0.026225	594.246219	<.0001	107.023896	0.180100
Group:	uv_index	2	0.025553	299.600362	<.0001	104.283754	0.174038
Var:	humidity	1	0.022304	539.438305	<.0001	91.023880	0.168738
Var:	gust_kph	1	0.007727	188.919099	<.0001	31.532208	0.166909
Var:	air_quality_PM2_5	1	0.003796	93.305102	<.0001	15.490003	0.166015

Figure 7.14 Results from the Effects Chosen for Target.

Findings from Figure 7.14**:**

- **R-Squared （R^2）:** Indicates the proportion of the variance in the target variable that is explained by each variable independently. In model selection, a higher R^2 value typically means the variable is more useful in explaining the target variable.
- **F Value:** Represents the significance of the variable within the model. The higher the F value, the more significant the contribution of the variable to the model.
- **p-Value:** Provides the statistical significance of the F test. A p-value less than 0.05 （commonly used as the threshold for significance testing） means that the improvement to the model by the variable is statistically significant, hence these variables are meaningful and should be included in the final predictive model.

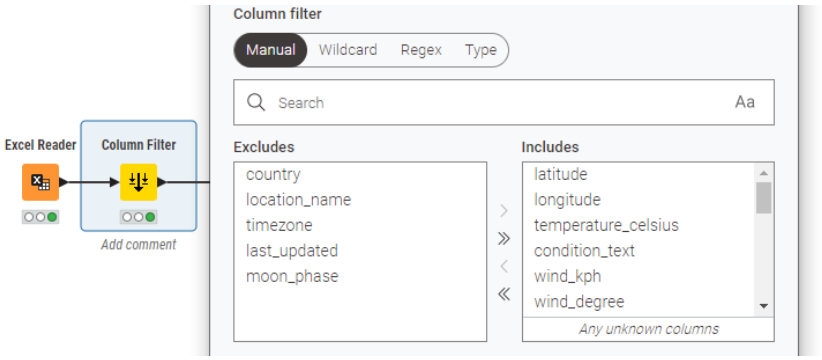
7.2.1.3 Outcomes from SAS EM Features Selection

The suggested final variables for the modeling process include air_quality_PM10, wind_kph, air_quality_Sulphur_dioxide, uv_index, humidity, gust_kph, and air_quality_PM2.5.

7.2.2 “DBSCAN” after feature engineering

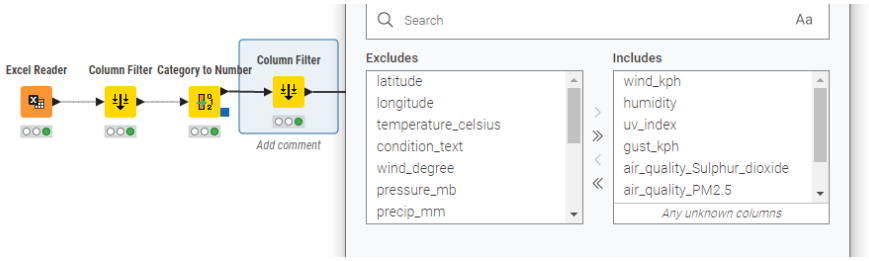
7.2.2.1 Initial data input

The initial input data will be the same as what we have used for the SAS Enterprise Miner.



7.2.2.2 DBSCAN

Based on the outputs from feature engineering in SAS Enterprise Miner, the initial data input will be filtered out by selecting these variables – air_quality_PM10, wind_kph, air_quality_Sulphur_dioxide, uv_index, humidity, gust_kph and air_quality_PM2.5 as an input for the DBSCAN.



- Workflow

The workflow for DBSCAN is arranged as shown in Figure 7.15

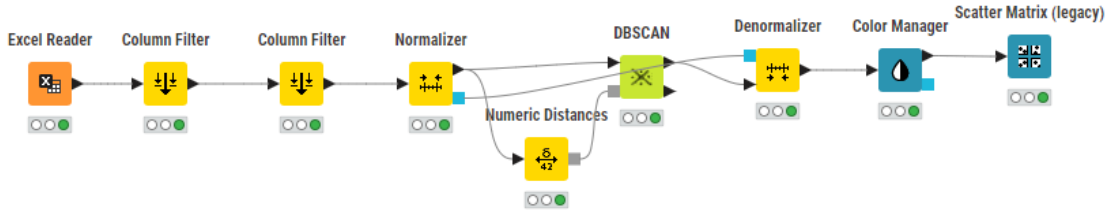


Figure 7.15 The arrangement of components in KNIME for DBSCAN.

- Identify noise data (the output of DBSCAN)

#	RowID	wind_kph Number (double)	humidity Number (double)	uv_index Number (double)	gust_kph Number (double)	air_quality_Sulphu... Number (double)	air_quality_PM2.5 Number (double)	air_quality_PM10 Number (double)	air_quality_gb-def... Number (double)	Cluster String
1	Row0	11.5	19	7	13.3	0.4	7.9	11.1	1	Cluster_74
2	Row1	6.1	54	6	11.9	1.8	28.2	29.6	3	Cluster_47
3	Row2	13	30	7	5.4	12.6	6.4	7.9	1	Cluster_74
4	Row3	9.7	51	4	11.9	0.2	0.5	0.8	1	Cluster_58
5	Row4	3.6	69	6	5.8	26.9	139.6	203.3	10	Noise
6	Row5	15.1	79	1	37.4	0.5	0.8	1.9	1	Cluster_73
7	Row6	11.2	71	1	13.7	1.3	2.1	3.5	1	Cluster_73
8	Row7	9	26	8	8.3	1.1	5	6.2	1	Cluster_63
9	Row8	15.1	62	1	15.1	0.5	4	5.8	1	Cluster_73
10	Row9	19.1	82	4	25.9	13	13.1	14.9	2	Cluster_0
11	Row10	22	36	7	15.8	0.9	5.6	6.8	1	Cluster_74
12	Row11	14.4	80	1	19.8	1.2	1.4	2.1	1	Cluster_73
13	Row12	13	54	9	16.2	9.7	69.3	152.9	9	Noise
14	Row13	5.4	52	7	7.6	11.1	84.8	95.6	10	Noise

- Relationships and distributions within data

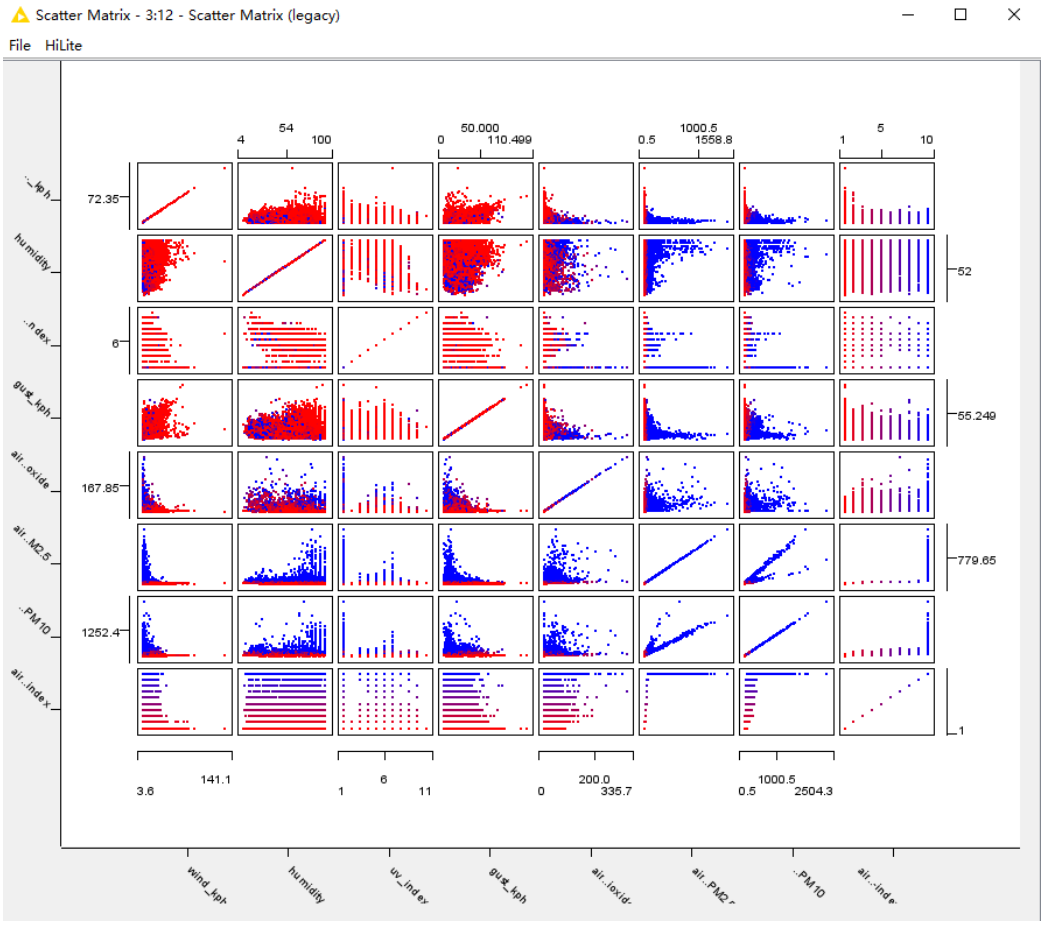


Figure 7.16 Results of Scatter Matrix.

Findings from Figure 7.16:

- **Histogram Distribution:** The histograms on the diagonal display the distribution of each variable. Most histograms exhibit distinct separate peaks, indicating that different clusters are well-differentiated across their respective variables.
- **Color Separation:** Different colors (representing different clusters) are well separated in the scatter plots; this indicates good clustering results. In our chart, for the most part, there are clear boundaries between colors, suggesting that the different clusters are well separated across these variable pairs.
- **Cluster Consistency:** Consistency in color clustering can be observed across multiple pairs of variables in the scatter plots. Most of the plots show data points of the same color grouped together, demonstrating consistency in the clustering.
- **Outliers:** Most scatter plots demonstrate clear and distinct color clustering, indicating that the clustering results for these pairs of variables are quite distinct.

7.2.2.3 Outcomes from DBSCANS

Overall, the clustering results of these variables are quite good. Although there is some degree of overlap for a minority of variable pairs, in most cases, there is a clear delineation between clusters, with consistency maintained across different pairs of variables. Therefore, we can consider the clustering results to be effective, and the clusters are distinctly separable for the majority of variable pairs.

8 Model

8.1 Data Partition Ratio

Before moving on to the modeling phase to predict the air quality index, it is important to determine the data splitting ratio for the dataset. Since identifying the optimal data splitting ratio for achieving the best model performance is challenging, we have opted to conduct experiments with three different ratios: 70:30, 80:20, and 90:10 for our modeling process as shown in Figure 8.1. Subsequently, we will evaluate the performance of each model based on all three data splitting ratios.

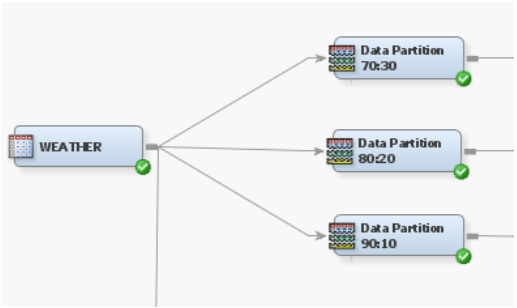


Figure 8.1 Setting up three different data splitting ratio using the "Data Partition" nodes.

8.2 Decision Tree Modeling Results

A decision tree model is a visual representation of a flowchart-like tree structure used in machine learning and decision-making processes. A typical decision tree model consists of these key components, namely root node, decision node, leaf/ terminal node, splitting, branch/ sub-tree, parent and child node.

Figure 8.2 shows the arrangement of the nodes in SAS Enterprise Miner to perform the modeling using decision tree. The result output of the decision tree models in predicting air quality index is presented in Figure 8.3. In addition, the confusion matrix from the output is also show in Figure 8.4. Table 8.1 shows a summary of the performance metrics that are essential for the assessment of the performance metrics.

Table 8.1 A summary of the important performance metrics for decision tree models to assess their models' performance.

	Decision Tree					
	70:30		80:20		90:10	
	Train	Validate	Train	Validate	Train	Validate
Misclassification Rate	0.159	0.152	0.209	0.207	0.210	0.207
Average Squared Error	0.021	0.021	0.029	0.029	0.029	0.029
ROC index	0.98	0.99	0.98	0.98	0.97	0.97
False Negative	120	34	14	3	0	0
True Negative	10795	4657	12254	3063	13625	1523
False Positive	418	156	560	149	793	85
True Positive	665	304	883	223	1010	113

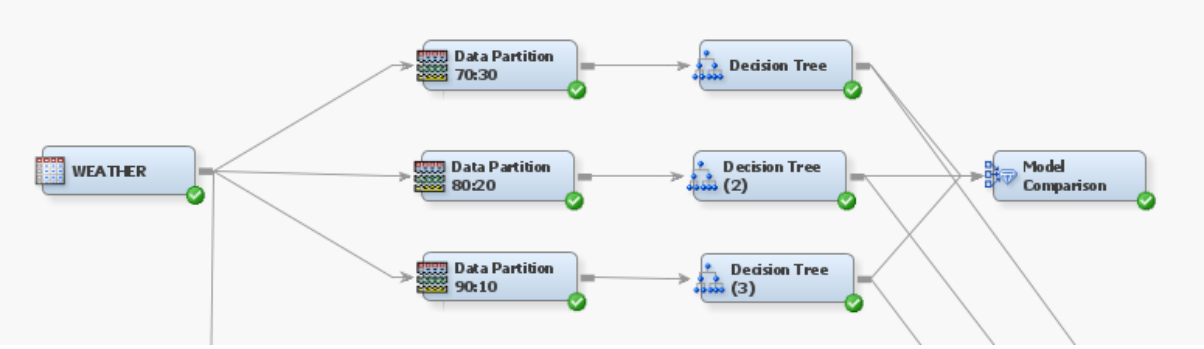


Figure 8.2 Setting up decision tree modeling using 3 different data partitions.

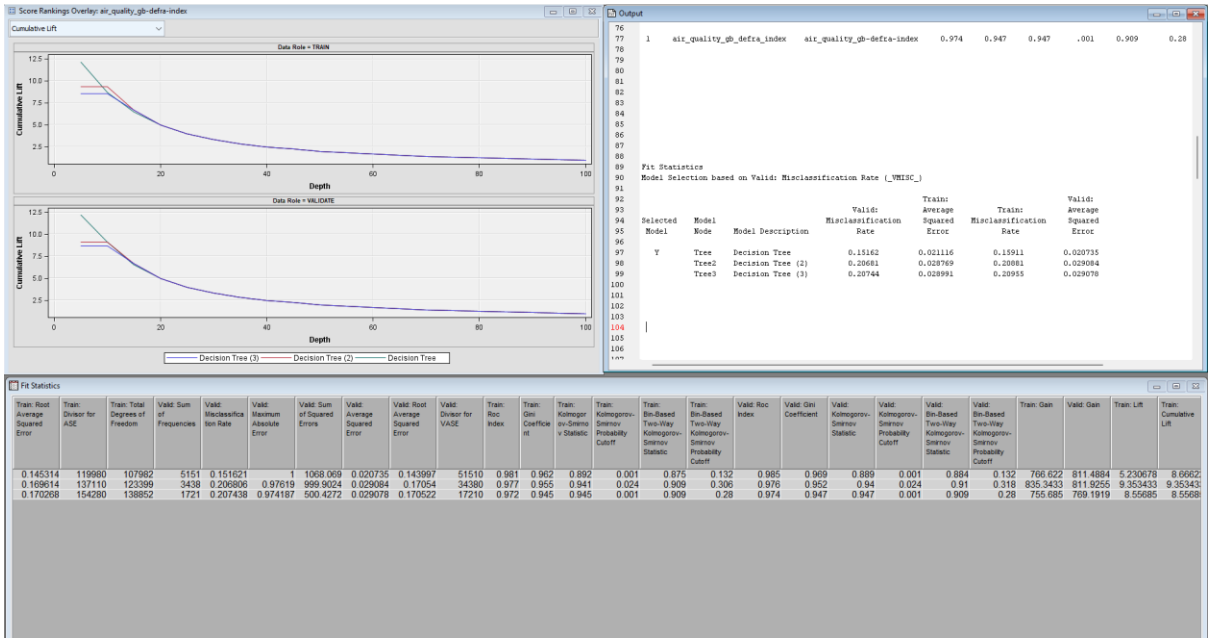


Figure 8.3 Results output from the model comparisons of 3 different decision tree models.

```
Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model                          Data                                                                    False      True       False      True
Node      Model Description    Role       Target                  Target Label                         Negative   Negative   Positive   Positive

Tree      Decision Tree        TRAIN      air_quality_gb_defra_index  air_quality_gb-defra-index       120        10795      418        665
Tree      Decision Tree        VALIDATE   air_quality_gb_defra_index  air_quality_gb-defra-index       34         4657       156        304
Tree2     Decision Tree (2)    TRAIN      air_quality_gb_defra_index  air_quality_gb-defra-index       14         12254      560        883
Tree2     Decision Tree (2)    VALIDATE   air_quality_gb_defra_index  air_quality_gb-defra-index       3          3063       149        223
Tree3     Decision Tree (3)    TRAIN      air_quality_gb_defra_index  air_quality_gb-defra-index       0          13625      793        1010
Tree3     Decision Tree (3)    VALIDATE   air_quality_gb_defra_index  air_quality_gb-defra-index       0          1523       85         113


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*
```

Figure 8.4 Confusion matrix output from the model comparisons of 3 different decision tree models.

8.3 Neural Network Modeling Results

A neural network model is a computational framework inspired by the human brain's structure and functioning. It comprises of interconnected nodes (neurons) organized into layers, which learn and make predictions by adjusting the strengths of connections based on the input train data.

In SAS Enterprise Miner, the neural network model is based on the Multilayer Perceptron (MLP). The arrangement of the nodes to perform the modeling using neural network in SAS Enterprise Miner is shown in Figure 8.5. According to the SAS Enterprise Miner E-book for applied analytics, the "Neural Network" node lacks the input selection capabilities. Therefore. The node "Variable Selection" is added to the arrangement and connects to the data source – WEATHER to filter out variables that are significantly important to the neural network modeling. The results output from the "Variable Selection" is shown in Figure 8.6.

Referring to Figure 8.7, it presented the result output of the neural network models for predicting air quality index. In addition, the confusion matrix from the output is also show in Figure 8.8. Table 8.2 shows a summary of the performance metrics that are essential for the assessment of the performance metrics.

Table 8.2 A summary of the important performance metrics for neural network models to assess their models' performance.

	Neural Network					
	70:30		80:20		90:10	
	Train	Validate	Train	Validate	Train	Validate
Misclassification Rate	0.214	0.212	0.211	0.212	0.213	0.219
Average Squared Error	0.028	0.028	0.027	0.027	0.027	0.028
ROC index	0.99	0.99	0.99	0.99	0.99	0.99
False Negative	41	20	47	14	55	5
True Negative	10853	4653	12361	3088	13955	1548
False Positive	360	160	453	124	463	60
True Positive	744	318	850	212	955	108

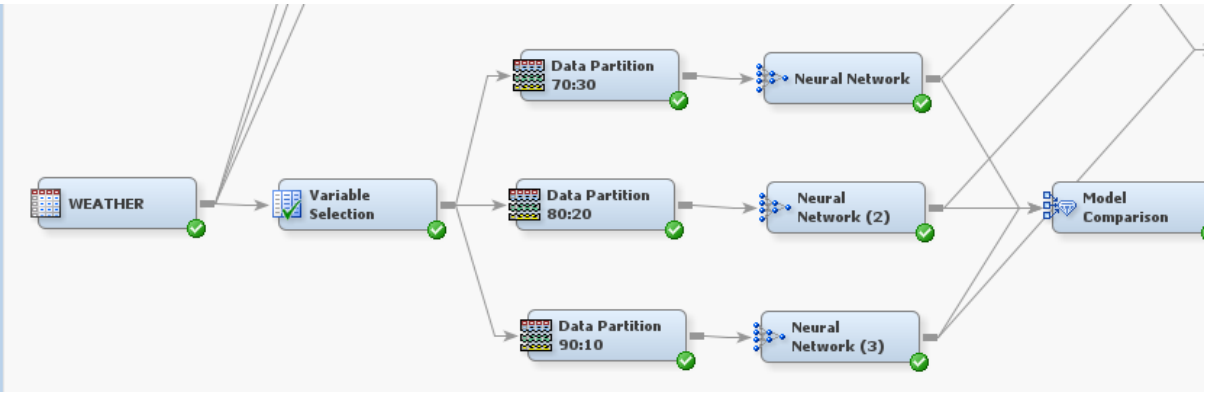
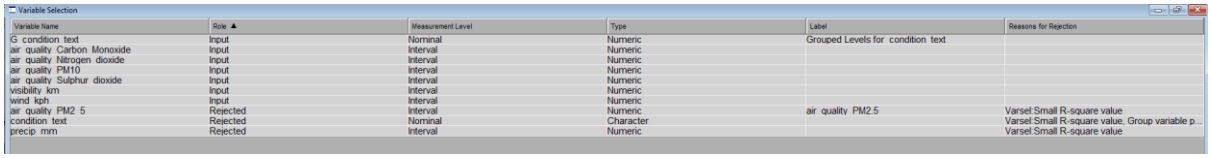


Figure 8.5 Setting up neural network modeling using 3 different data partitions.



Variable Name	Role	Measurement Level	Type	Label	Reasons for Rejection
G condition text	Input	Nominal	Numeric	Grouped Levels for condition text	
air quality Carbon Monoxide	Input	Interval	Numeric		
air quality Nitrogen dioxide	Input	Interval	Numeric		
air quality PM10	Input	Interval	Numeric		
air quality Sulphur dioxide	Input	Interval	Numeric		
visibility km	Input	Interval	Numeric		
wind kph	Input	Interval	Numeric		
air quality PM2 5	Rejected	Interval	Numeric	air quality PM2.5	Varsel:Small R-square value
condition text	Rejected	Nominal	Character		Varsel:Small R-square value, Group variable p...
precip mm	Rejected	Interval	Numeric		Varsel:Small R-square value

Figure 8.6 Results output from the "Variable Selection" node.

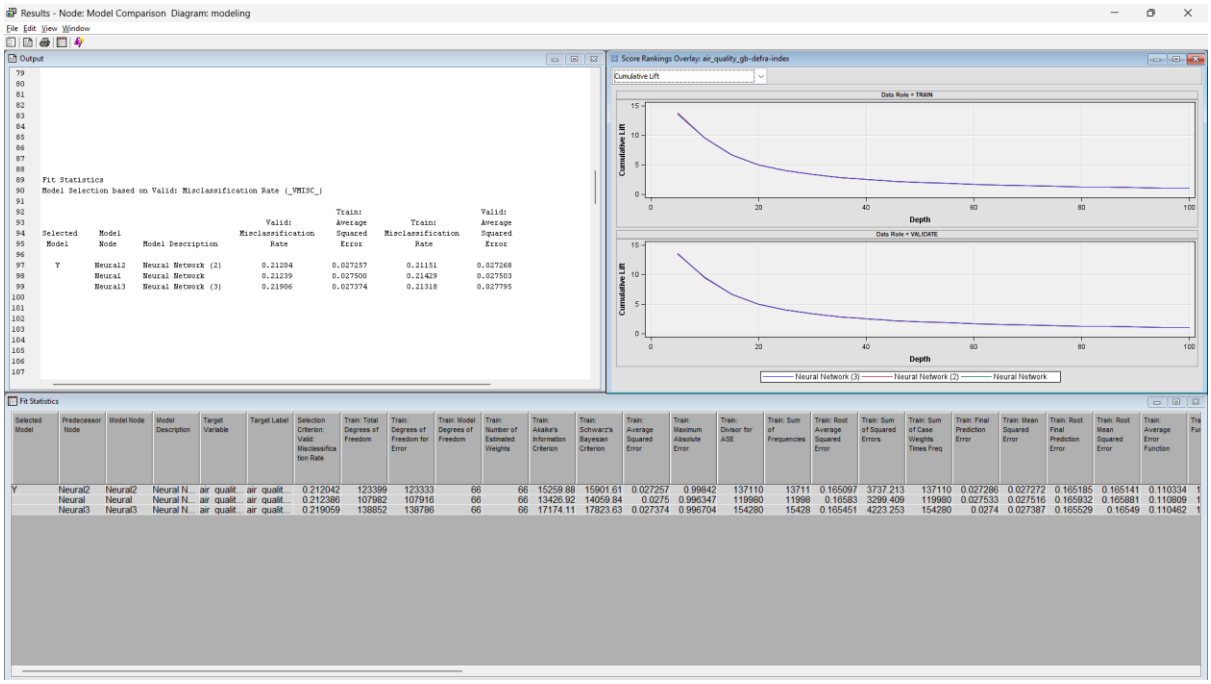


Figure 8.7 Results output from the model comparisons of 3 different neural network models.

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Neural	Neural Network	TRAIN	air_quality_gb_defra_index	air_quality_gb-defra-index	41	10853	360	744
Neural	Neural Network	VALIDATE	air_quality_gb_defra_index	air_quality_gb-defra-index	20	4653	160	318
Neural2	Neural Network (2)	TRAIN	air_quality_gb_defra_index	air_quality_gb-defra-index	47	12361	453	850
Neural2	Neural Network (2)	VALIDATE	air_quality_gb_defra_index	air_quality_gb-defra-index	14	3088	124	212
Neural3	Neural Network (3)	TRAIN	air_quality_gb_defra_index	air_quality_gb-defra-index	55	13955	463	955
Neural3	Neural Network (3)	VALIDATE	air_quality_gb_defra_index	air_quality_gb-defra-index	5	1548	60	108

----
* Score Output

----
* Report Output

Figure 8.8 Confusion matrix output from the model comparisons of 3 different neural network models.

8.4 Ensemble Modeling Results

Referring to the SAS Enterprise E-book, the Ensemble models can be defined as a model which is built on the foundation of predictions from multiple models to create a single consensus prediction. The advantage of ensemble models is that the combined model might produce better performance than the individual models that compose it.

The arrangement of the nodes to perform the modeling using neural network in SAS Enterprise

Miner is shown in Figure 8.9. As mentioned earlier, the ensemble model is built on top of the foundations from other models' predictions. Therefore, the "Control Point" node is used and connected with the predictions made by the Decision Tree model and Neural Network model. Then, the control point is passed to the Ensemble node for modeling purposes.

Referring to Figure 8.10, it presented the result output of the neural network models in predicting the air quality index. In addition, the confusion matrix from the output is also show in Figure 8.11. Table 8.3 shows a summary of the performance metrics that are essential for the assessment of the performance metrics.

Table 8.3 A summary of the important performance metrics for ensemble models to assess their models' performance.

	Ensemble Models (DT & NN)					
	70:30		80:20		90:10	
	Train	Validate	Train	Validate	Train	Validate
Misclassification Rate	0.149	0.148	0.197	0.195	0.195	0.203
Average Squared Error	0.022	0.021	0.026	0.026	0.026	0.026
ROC index	0.99	0.99	1.00	0.99	1.00	1.00
False Negative	41	20	11	2	0	0
True Negative	10811	4638	12213	3055	13583	1517
False Positive	402	175	601	157	835	91
True Positive	744	318	886	224	1010	113

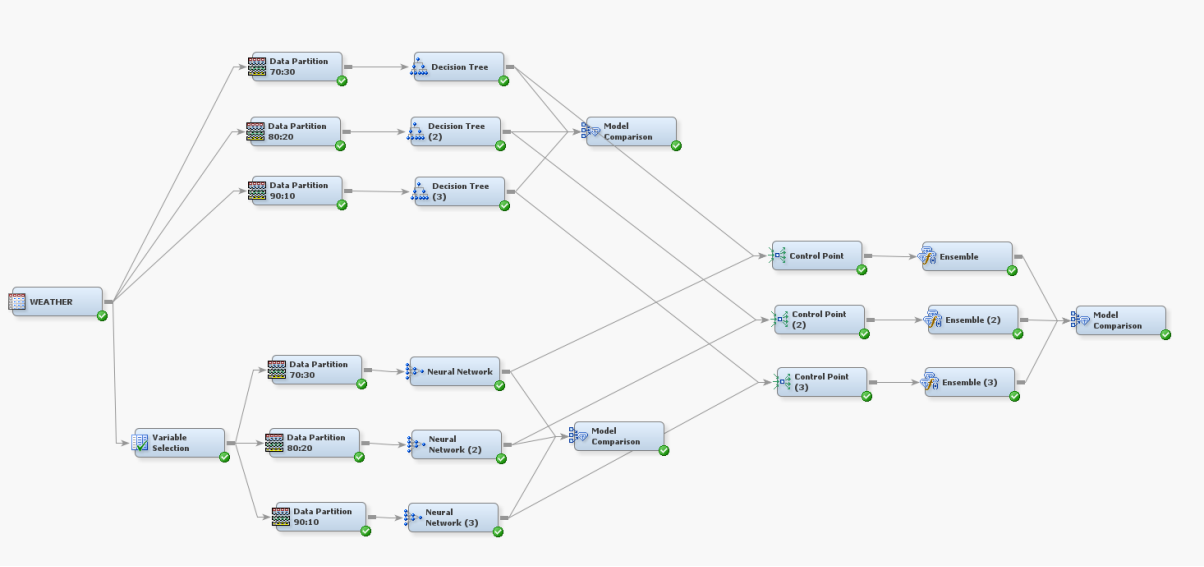


Figure 8.9 Setting up ensemble modeling using 3 different data partitions.

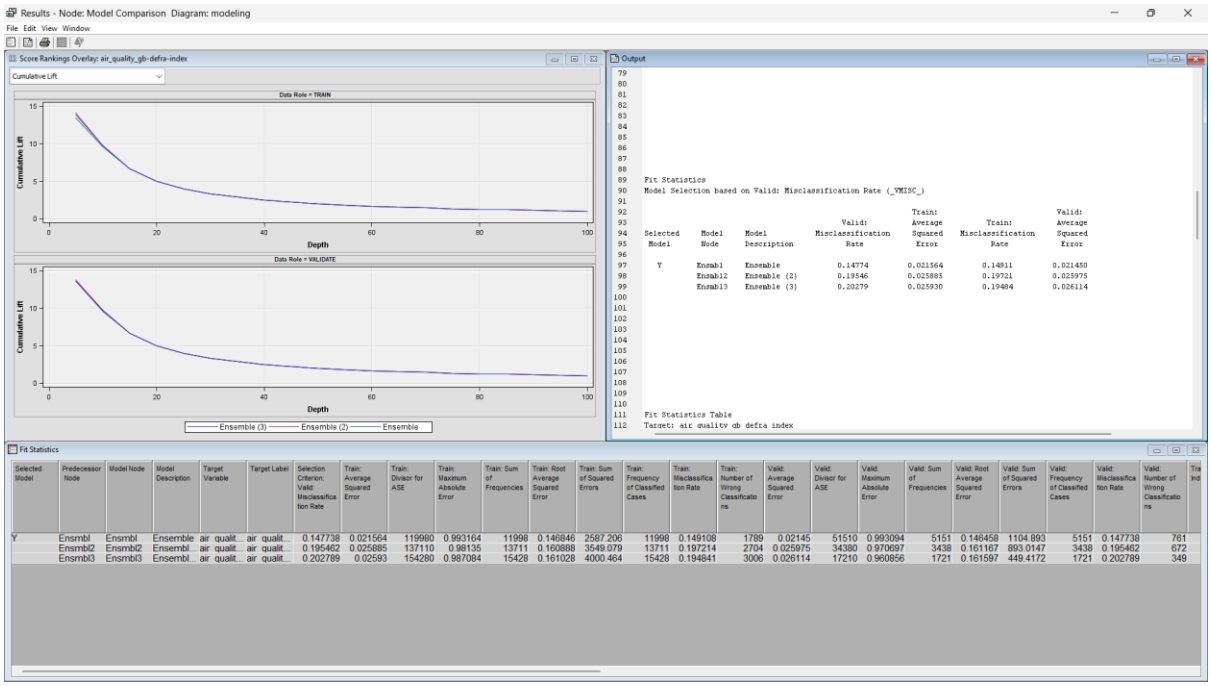


Figure 8.10 Results output from the model comparisons of 3 different ensemble models.

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Ensmbl	Ensemble	TRAIN	air_quality_gb_defra_index	air_quality_gb-defra-index	41	10811	402	744
Ensmbl	Ensemble	VALIDATE	air_quality_gb_defra_index	air_quality_gb-defra-index	20	4638	175	318
Ensmbl2	Ensemble (2)	TRAIN	air_quality_gb_defra_index	air_quality_gb-defra-index	11	12213	601	886
Ensmbl2	Ensemble (2)	VALIDATE	air_quality_gb_defra_index	air_quality_gb-defra-index	2	3055	157	224
Ensmbl3	Ensemble (3)	TRAIN	air_quality_gb_defra_index	air_quality_gb-defra-index	0	13583	835	1010
Ensmbl3	Ensemble (3)	VALIDATE	air_quality_gb_defra_index	air_quality_gb-defra-index	0	1517	91	113

--
* Score Output
--

--
* Report Output
--

Figure 8.11 Confusion matrix output from the model comparisons of 3 different ensemble models.

8.5 Result Comparison between the Selected Models

In this segment, we have condensed the comparisons of results among the chosen models in predicting the air quality index and categorized them according to the data splitting ratio, as illustrated Table *8.4*, Table *8.5*, and Table *8.6*. Result comparison is a crucial step in assessing the performance of the models conducted in SAS Enterprise Miner, as it provides valuable insights into its effectiveness and reliability of the models. By systematically comparing and contrasting these results, we can gain a comprehensive understanding of how well a model performs under various conditions.

Table 8.4 Results comparison between the selected model for Data Splitting Ratio 70:30

	Data Splitting Ratio 70:30					
	Decision Tree		Neural Network		Ensemble	
	Train	Validate	Train	Validate	Train	Validate
Misclassification Rate	0.159	0.152	0.214	0.212	0.149	0.148
Average Squared Error	0.021	0.021	0.028	0.028	0.022	0.021
ROC index	0.98	0.99	0.99	0.99	0.99	0.99
False Negative	120	34	41	20	41	20
True Negative	10795	4657	10853	4653	10811	4638
False Positive	418	156	360	160	402	175
True Positive	665	304	744	318	744	318

Table 8.5 Results comparison between the selected model for Data Splitting Ratio 80:30

	Data Splitting Ratio 80:20					
	Decision Tree		Neural Network		Ensemble	
	Train	Validate	Train	Validate	Train	Validate
Misclassification Rate	0.209	0.207	0.211	0.212	0.197	0.195
Average Squared Error	0.029	0.029	0.027	0.027	0.026	0.026
ROC index	0.98	0.98	0.99	0.99	1.00	0.99
False Negative	14	3	47	14	11	2
True Negative	12254	3063	12361	3088	12213	3055
False Positive	560	149	453	124	601	157
True Positive	883	223	850	212	886	224

Table 8.6 Results comparison between the selected model for Data Splitting Ratio 70:30

	Data Splitting Ratio 90:10					
	Decision Tree		Neural Network		Ensemble	
	Train	Validate	Train	Validate	Train	Validate
Misclassification Rate	0.210	0.207	0.213	0.219	0.195	0.203
Average Squared Error	0.029	0.029	0.027	0.028	0.026	0.026
ROC index	0.97	0.97	0.99	0.99	1.00	1.00
False Negative	0	0	55	5	0	0
True Negative	13625	1523	13955	1548	13583	1517
False Positive	793	85	463	60	835	91
True Positive	1010	113	955	108	1010	113

9 Assess

Based on the results comparisons stated in Section 8.5, we are then able to assess the performance of each model from the perspective of misclassification rate, average squared error, ROC index and confusion matrix.

9.1 Misclassification Rate, Average Squared Error and ROC Index

Misclassification rate is a measure of the proportion of incorrectly classified instances, where lower values indicate better performance. Referring to the results tabulated in Section 8.5, the misclassification rate of models trained by data splitting ratio of 70:30 outperformed the other data splitting ratios – 80:20 and 90:10. Moreover, the Ensemble model seems to have the lowest misclassification rates on both the training and validation sets as compared to Decision Tree and Neural Network across the data splitting ratio.

For average squared error, it is a measure of the average of the square differences between the predicted and actual values, where lower values indicate better performance. However, there is no significant difference in the average squared error value between models across the data splitting ratio. All three models have relatively low values in the average squared error.

Lastly, for ROC index, according to the SAS Enterprise Miner E-book, a strong model will have a ROC index greater than 0.7. Viewing from the tabulated results, all three models are strong in predicting the air quality index with the ROC index ranging from 0.97 to 1.00.

9.2 Confusion Matrix

Referring to the confusion matrix recorded in Section 8.5, we are then able to calculate the precision, recall, F1-score, accuracy and specificity for all models. All the calculated performance metrics are then tabulated in Table *9.1*, Table *9.2*, and Table *9.3* The formulas use to calculate those performance metrics are listed below: -

- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $Specificity = \frac{TN}{TN+FP}$
- $Accuracy = \frac{TP+TN}{TP+FN+TN+FP}$
- $F1 = \frac{2 \times precision \times recall}{precision + recall}$

Table 9.1 Performance metrics calculated from confusion metrics of all models for data splitting ratio 70:30.

	Data Splitting Ratio 70:30					
	Decision Tree		Neural Network		Ensemble	
	Train	Validate	Train	Validate	Train	Validate
Precision	0.614	0.661	0.674	0.665	0.649	0.645
Recall	0.847	0.899	0.948	0.941	0.948	0.941
F1-score	0.712	0.762	0.788	0.779	0.771	0.765
Accuracy	0.955	0.963	0.967	0.865	0.963	0.962
Specificity	0.963	0.967	0.968	0.967	0.964	0.964

Table 9.2 Performance metrics calculated from confusion metrics of all models for data splitting ratio 80:20.

	Data Splitting Ratio 80:20					
	Decision Tree		Neural Network		Ensemble	
	Train	Validate	Train	Validate	Train	Validate
Precision	0.608	0.599	0.652	0.631	0.596	0.588
Recall	0.984	0.986	0.948	0.938	0.988	0.991
F1-score	0.752	0.746	0.773	0.754	0.743	0.738
Accuracy	0.958	0.956	0.964	0.960	0.955	0.954
Specificity	0.956	0.954	0.965	0.961	0.953	0.951

Table 9.3 Performance metrics calculated from confusion metrics of all models for data splitting ratio 90:10.

	Data Splitting Ratio 90:10					
	Decision Tree		Neural Network		Ensemble	
	Train	Validate	Train	Validate	Train	Validate
Precision	0.560	0.571	0.673	0.643	0.547	0.54
Recall	1.00	1.00	0.946	0.956	1.00	1.00
F1-score	0.718	0.727	0.787	0.769	0.708	0.713
Accuracy	0.949	0.951	0.966	0.962	0.946	0.947
Specificity	0.945	0.947	0.968	0.963	0.942	0.943

Based on the above tabulated results, those performance metrics – precision, recall, F1-score, accuracy and specificity provide valuable insights into its performance. They offer a comprehensive evaluation of the model's ability to make accurate predictions.

The Decision Tree model exhibits a notable decrease in precision as the data splitting ratio increases, suggesting a potential susceptibility to overfitting with larger training datasets. However, its recall remains consistently high, and reaching a perfect score in the 90:10 split, indicating robust performance in identifying positive instances or due to no false negative is recorded. The F1-score experiences a slight decline with higher data splitting ratios, reflecting a balance between precision and recall. The accuracy, while generally high, shows a modest decrease with expanded training datasets. Specificity decreases as well, pointing to a potential trade-off with sensitivity.

In contrast, the Neural Network model demonstrates insignificant changes in the performance metrics across different splitting ratios. In summary, the neural network model has the capabilities to capture both positive and negative instances as indicated by the F1-score which reflect a balance between precision and recall.

Lastly, for the Ensemble model, a combination of individual models from Decision Tree and Neural Network, shows a decreasing trend in precision with higher data splitting ratios, suggesting a susceptibility to overfitting. The performance of the ensemble model shows similar characteristics to the Decision Tree, where recall remains consistently high, reaching a perfect score in the 90:10 split, showcasing the ensemble's ability to capture positive instances effectively. The F1-score exhibits a declining trend with higher data splitting ratios, hinting at a potential trade-off between precision and recall. In terms of accuracy, it shows a high value but experiences a slight decline with larger training datasets. The specificity decreases as well,

indicating a potential compromise with sensitivity too.

9.3 Summary of Model Assessment

After assessing all the model's performance with different data splitting ratios, models that trained with data splitting ratio 70:30 show the best performance with Ensemble model as the best model compared to Decision Tree and Neural Network. In general, while the Ensemble model showcases high accuracy and specificity, there is a notable need to focus on enhancing its capability to identify positive instances due to low precision and recall. This is due to the fact that the dataset used for training the model is highly imbalanced and resulted in overfitting of the majority class in the dataset. Therefore, it is essential to address the overfitting and data imbalances issues, so that the model's overall performance can be improved.

9.4 Solutions to Data Imbalances and Overfitting

To resolve the data imbalances and overfitting issues, there are two common approaches which are oversampling and undersampling.

Oversampling focuses on increasing the number of instances in the minority class through methods such as duplicating values or generating synthetic examples, like using the Synthetic Minority Over-sampling Technique (SMOTE). This process aims to create a more balanced class distribution between the minority and majority classes. Achieving balance in the dataset enables the model to effectively capture patterns within the minority class during training, thereby mitigating any bias towards the majority class. As a result, the overall performance of the model experiences improvement.

On the other hand, undersampling tackles data imbalance by reducing the number of instances in the majority class, typically done randomly to bring the distribution closer to that of the minority class. However, the use of undersampling comes with a trade-off, as there is a potential risk of discarding valuable information present in the majority class. Therefore, careful consideration of dataset characteristics becomes crucial when opting for undersampling.

Both oversampling and undersampling contribute to resolving the overfitting issue in a model by preventing the model from excessively memorizing the training data. By training the model on a balanced dataset, it allows for better generalization to unseen instances, resulting in improved prediction performance.

10 Conclusions

Initially, the dataset contains 41 variables and among them, 14 variables were dropped because some of them are duplicated columns with different unit measurement systems and do not provide additional information. Additionally, some variables like sunrise, sunset, moonrise and moonset are dropped too as it is not important for air quality prediction. After the simple data cleaning and data quality check using Talend Data Preparation and Talend Data Integration, the dataset used for our project contains a total of 27 variables. After importing the dataset to SAS Enterprise Miner, the tools labelled 20 of the variables as interval variables and the remaining as nominal variables. However, there are some errors in the labelling of the type of variable. Therefore, we manually revise the metadata, and the output is shown in Table 10.1.

Table 10.1 Revised metadata.

Role	Type of Variable	Count
Input	Interval	19
Input	Nominal	2
Target	Nominal	1
Rejected	Nominal	3
Time ID	ID	1
Text	Nominal	1

For data sampling, we have decided to use the full dataset entries as the amount of data is still manageable as compared to the store sales data. For exploration parts, as stated in section 6.5, most of the interested variables contain outliers which needs to be treated during the modification process. In addition, from the exploratory data analysis, the variables which influence the Air Quality Index (AQI) are pollutants' concentration (CO, NO_2, SO_2, PM2.5, and PM10), wind speed, visibility, precipitation amount and weather conditions.

From the modify phase, it focusses on improving the quality and relevance of the data through data imputation, feature selection, and data transformation techniques. These modifications enhance the dataset's suitability for modelling part. Sven significant features, namely air_quality_PM10, wind_kph, air_quality_Sulphur_dioxide, uv_index, humidity, gust_kph, and air_quality_PM2.5, were selected based on their significant higher R-square values to the target variable compared to others. From the key findings obtained through the exploration and modification of our dataset, we can use that information to construct an accurate model for predicting the Air Quality Index (AQI) by incorporating the identified features that exhibit correlations with AQI.

In terms of the findings from Modeling and Assessment, models that trained with data splitting ratio 70:30 show the best performance with Ensemble model as the best model compared to Decision Tree and Neural Network. Although, Ensemble model showcases high accuracy and specificity, there is a notable need to focus on enhancing its capability to identify positive instances due to low precision and recall due to data imbalanced and resulted in overfitting of the majority class in the dataset. Therefore, in future work, we suggest adopting the solutions presented at Section 9.4 to resolve the data imbalance and overfitting issue.

In conclusion, this project taught us well in understanding the availability of various tools like Talend Data Integration, Talend Data Preparation, SAS Enterprise Miner and KNIME can help us to streamline the data mining process without going through the hassle to code from scratch. In addition, the good graphic user interface and documentations also help new users like us to pick up the tools quickly.

11 Teamwork and Collaboration

In this cooperation project, our team demonstrated efficient teamwork and collaboration. In order to confirm the progress of our peers at any time and increase efficient communication between teams, we collaborate by sharing documents in real time on Teams as shown in Figure 11.1.

Figure 11.1 Shared documents on Microsoft Teams

We held a total of 8 team meetings throughout the project work, mainly to update the project progress and discuss any problems encountered. For major discussion meetings, the meetings were recorded for references as shown in Figure 11.2.

Figure 11.2 Important meeting records

For each meeting, we will each summarize the content of what we have done, new discoveries, and difficulties encountered. In addition, we will also coordinate the overall project and discuss what changes personal make will cause changes to others, so that everyone can accurately grasp the progress and changes of the group, keep up with the overall progress of the group, and not fall behind. Finally, we work out what to do next and when the next meeting will be, based on everyone's time and workload.

Besides, using the Microsoft teams, we also created a Whatsapp work group for collaboration as it is much easier to reach out to the group mates.

Our project has not been smooth sailing. When we were about to start our project, our team leader believed that it was necessary for us to generate new hypotheses based on the actual

model results while modeling, and it is necessary to explore the Sample and Explore sections based on the new hypotheses. Because we have completed the Sample and Explore parts as well as possible in the assignment part, and going back to the Sample and Explore parts means that we need to explore more deeply, and we may not be able to explore the information we want, so we There was a heated discussion on this point at that meeting. In the end, we decided to divide the four-person team into two groups, with Low Boon Kiat and ZhaoZihui to work together to improve all our pre-modeling work, and Low Boon Kiat to re-explore our Sample and Explore parts and adopt new technologies are explored in many aspects and in depth. ZhaoZihui adjusts and decides on Modify parts based on the results of Sample and Explore, striving to provide correct and accurate information for the modeling part. In this way, we can ensure that our model selection and attribute input and other important modeling directions are correct. The Model and Assessment parts are conducted by the team Wong Jia Hui and Jie Hongsheng, so that the two persons can discuss and analyze together, ultimately determine the selection and improvement of the model, and strive to obtain the best evaluation results.

It is worth mentioning that the team leader will organize and guide us to learn the use of each software in depth during the project, and record videos to ensure that everyone can learn the use of all software. This allows our team to complete group projects efficiently and within the specified timeframe.

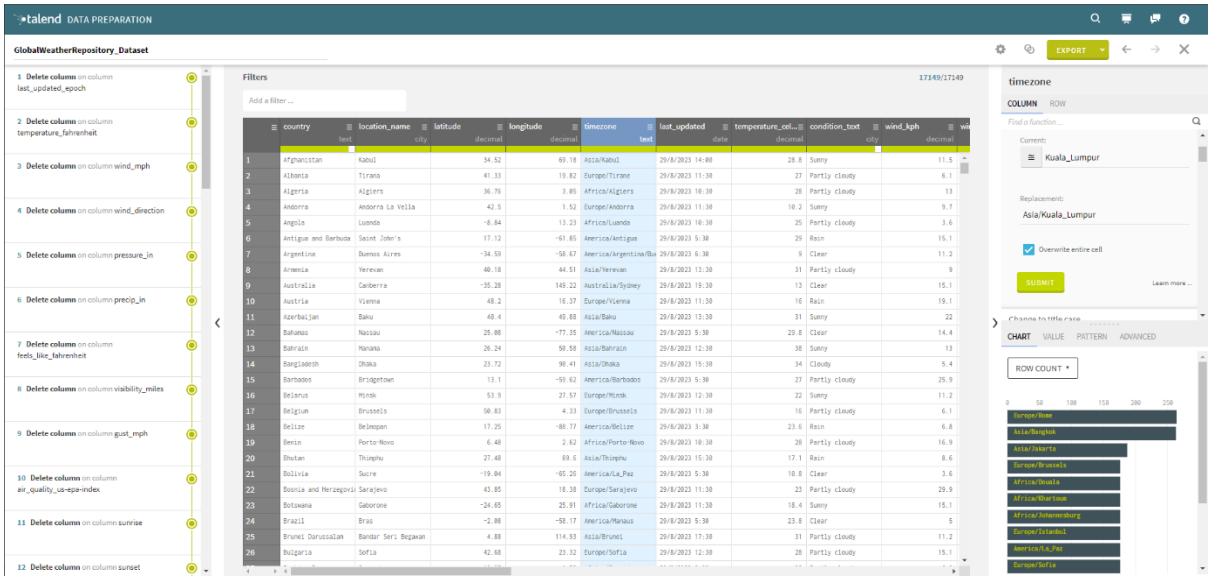
Thanks to DR. TEH YING WAH, we learned the use of a variety of software in this class, which provided the necessary conditions for our group's project development. Through this group project, we realized that in order to achieve good results in actual data mining and analysis projects, the SEMMA process is often iterative and dynamic rather than strictly linear. The results of the model may indicate the need for data processing. Further cleaning or transformation, such as handling outliers or applying different data normalization methods. During the evaluation phase, we often need to go back to the model phase or even the data preprocessing phase to optimize the model's performance. This challenged our team’s proficiency in using various software technologies and timely and effective communication. In the end, through collaborative cooperation and repeated confirmations, we successfully completed our team project.

Appendix

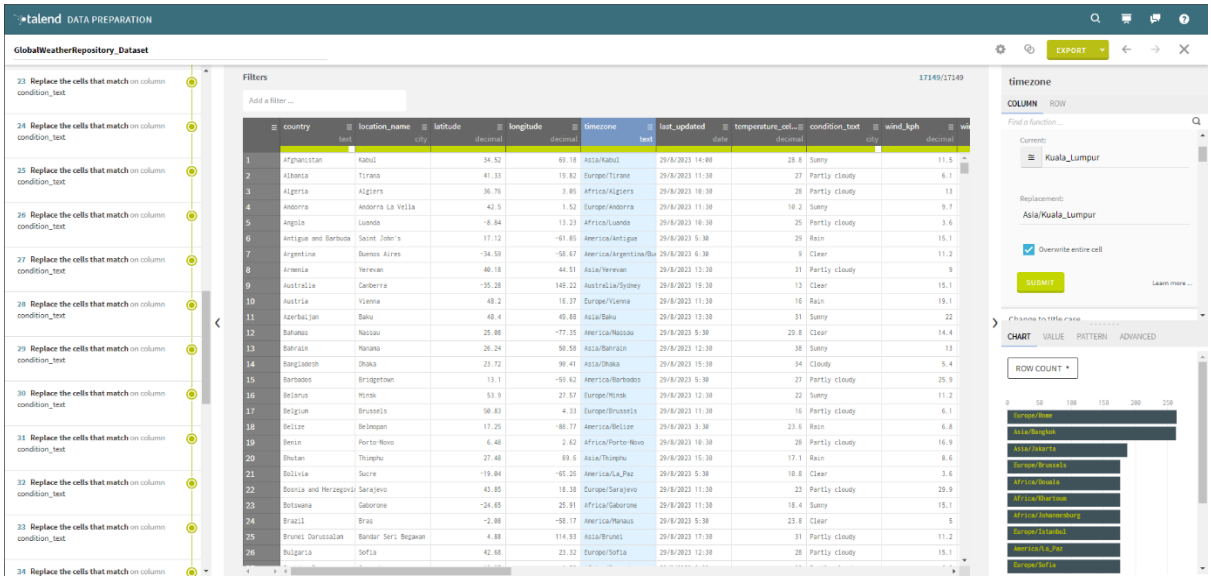
Appendix A – Procedure for Basic Data Cleaning

Talend Data Preparation is used to perform basic data cleaning.

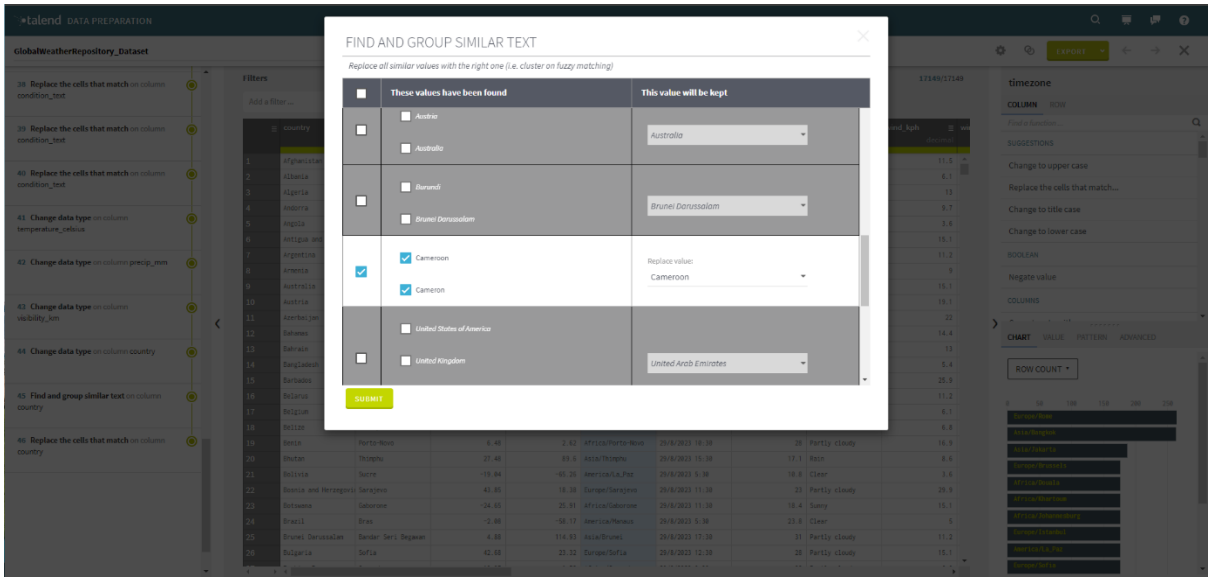
i. Drop 14 columns because these are duplicated columns with different units of measurement and do not provide additional information.



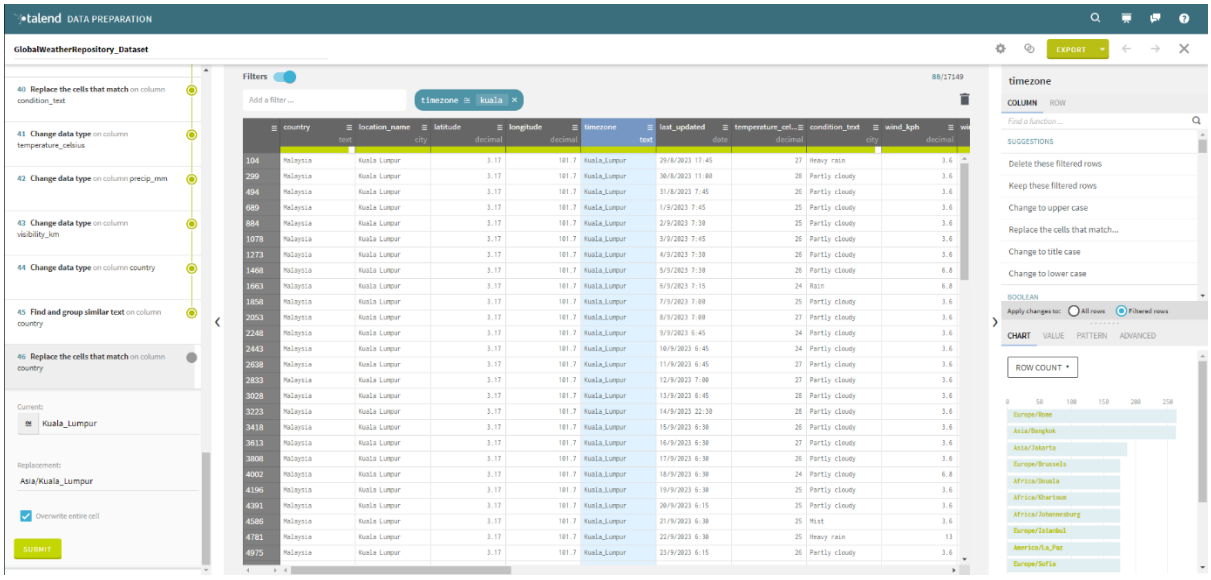
ii. Re-categorize values in the “condition_text” column.



iii. Correct the spelling error on Cameroon’s country name (noisy data).



iv. Change the timezone format for Malaysia from “Kuala Lumpur” to “Asia/Kuala Lumpur” because the timezone format for most countries are region/city (inconsistent data).

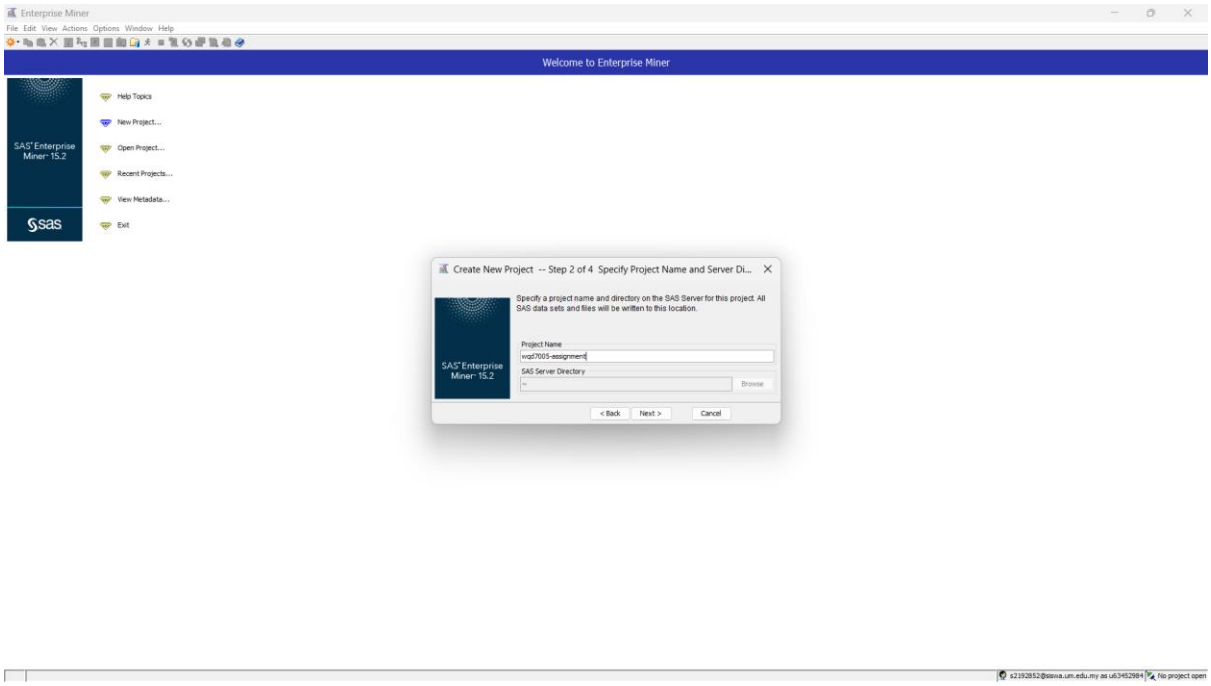


Appendix B – Procedure for Conducting Sample and Explore

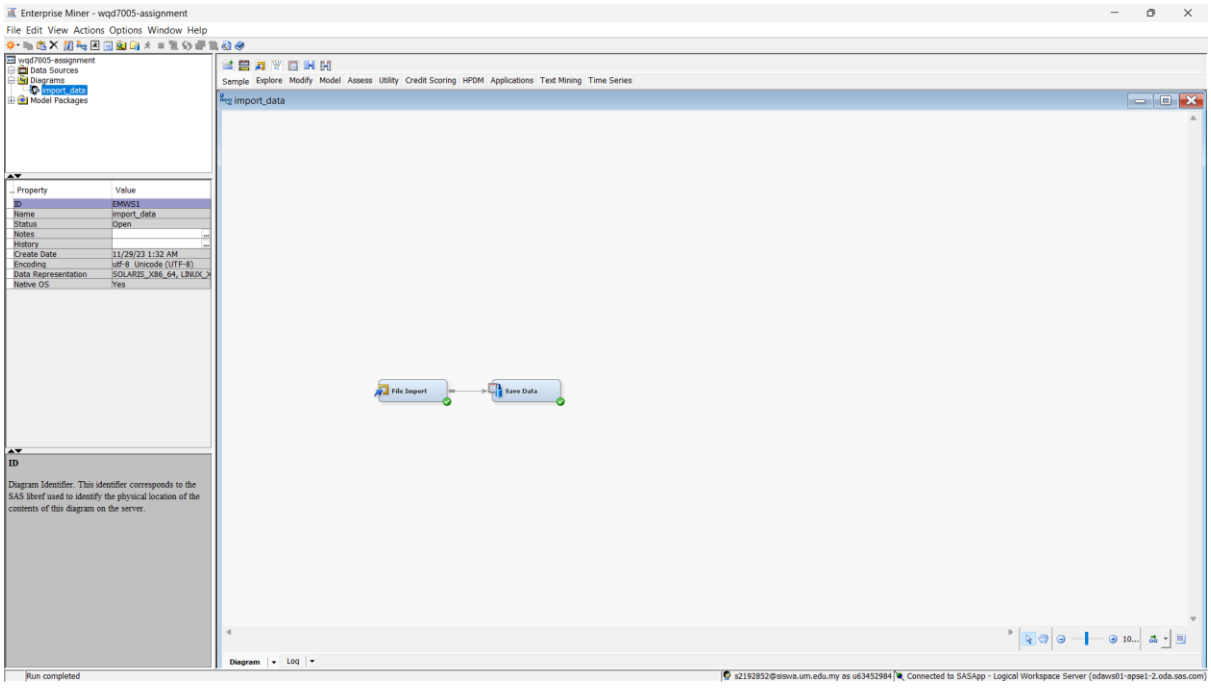
SAS Enterprise Miner is used to carry out the first two steps of SEMMA – Sample and Explore.

Project Setup

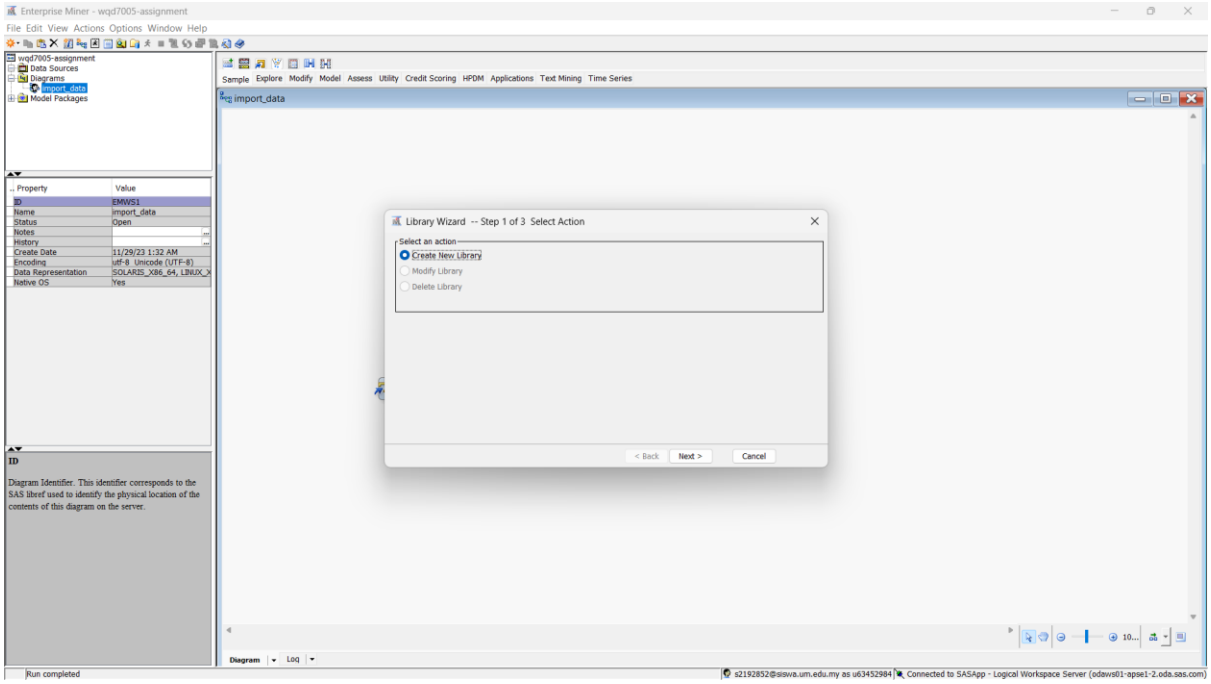
i. Create a new Enterprise Miner Project.

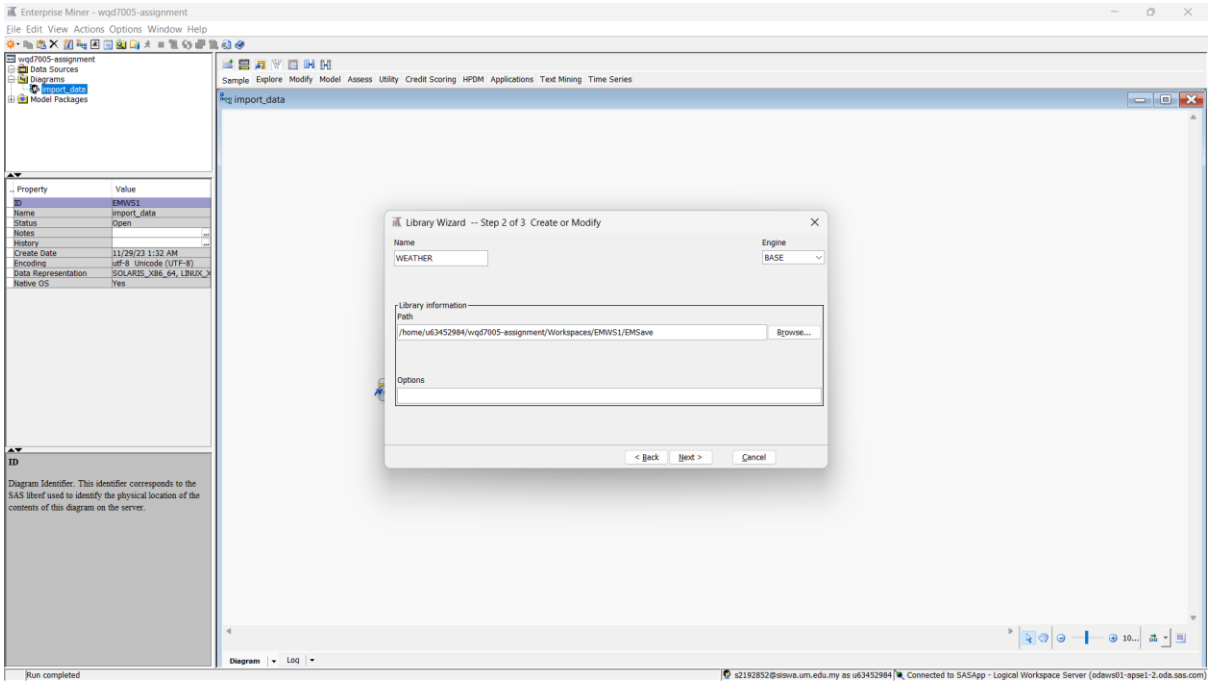


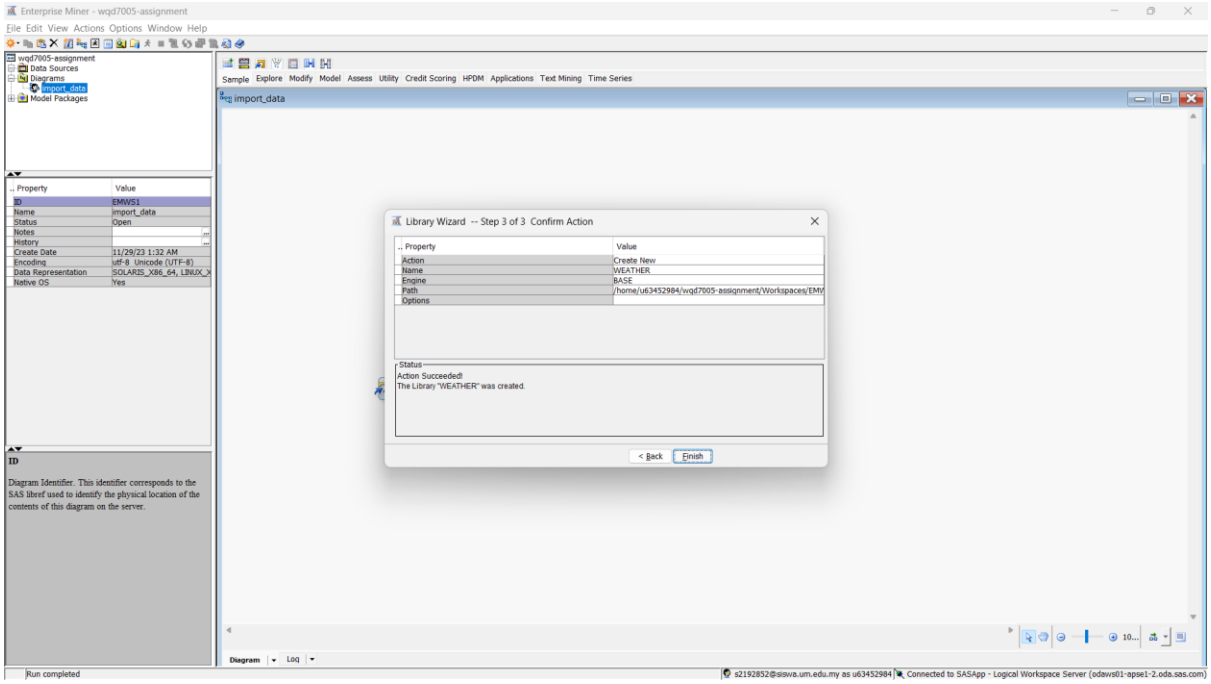
ii. Create a diagram with the nodes for file import and save data.



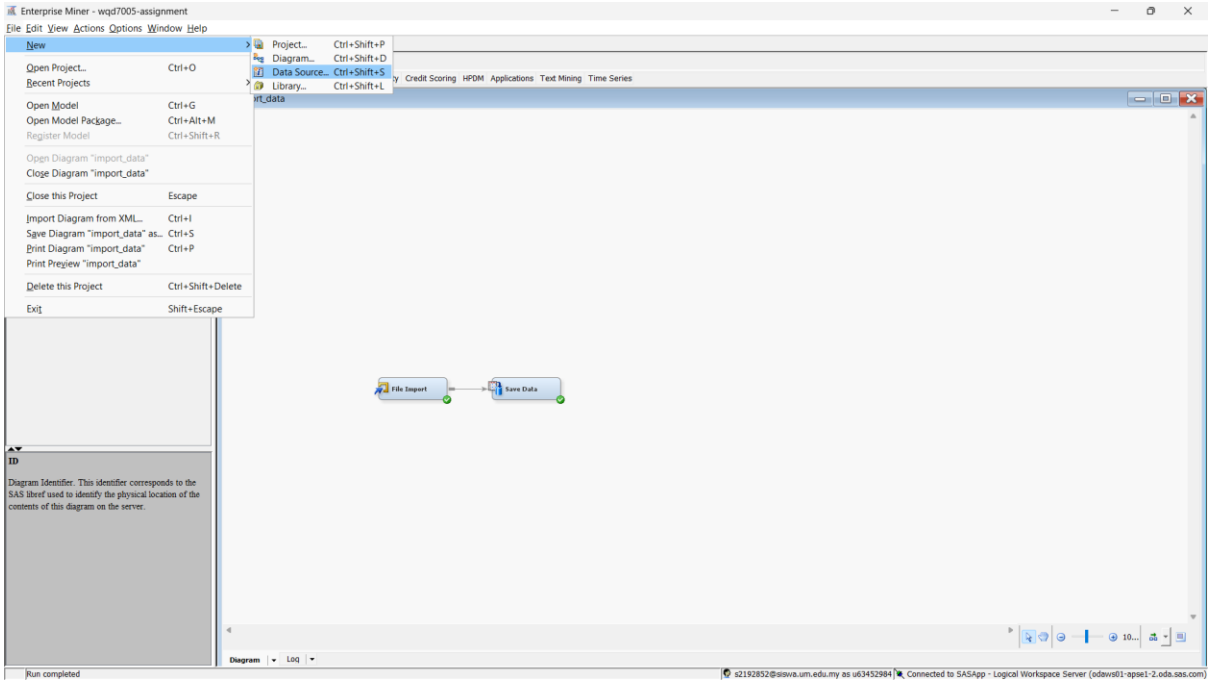
iii. Create a new library.

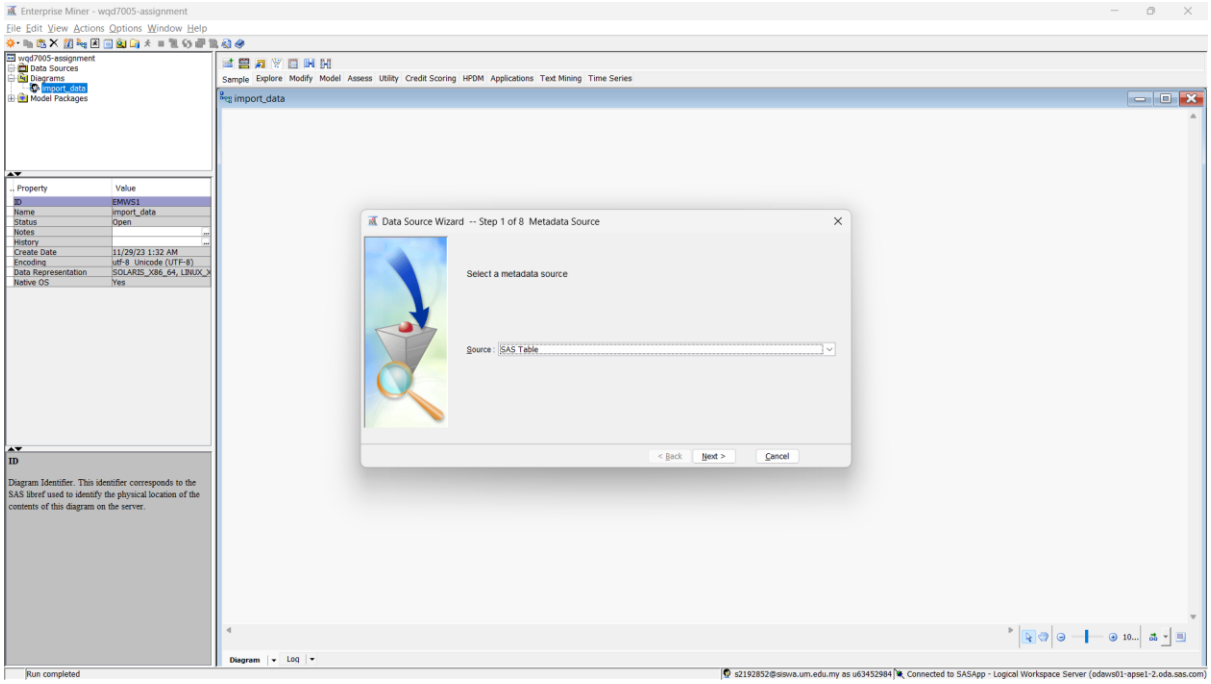

Enterprise Miner - wqd7005-assignment
File Edit View Actions Options Window Help
wqd7005-assignment
Data Sources
Diagrams
import_data
Model Packages
Sample Explore Modify Model Assess Utility Credit Scoring HPDM Applications Text Mining Time Series
import_data
Property
Value
ID
EMWS1
Name
import_data
Status
Open
Notes
History
Create Date
11/29/23 1:32 AM
Encoding
utf-8 Unicode (UTF-8)
Data Representation
SOLARIS_X86_64, LINUX_X
Native OS
Yes
ID
Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.
Library Wizard -- Step 1 of 3 Select Action
Select an action
Create New Library
Modify Library
Delete Library
< Back
Next >
Cancel
Diagram
Log
Run completed
s2192852@siswa.um.edu.my as u63452984
Connected to SASApp - Logical Workspace Server (odaws01-apse1-2.oda.sas.com)

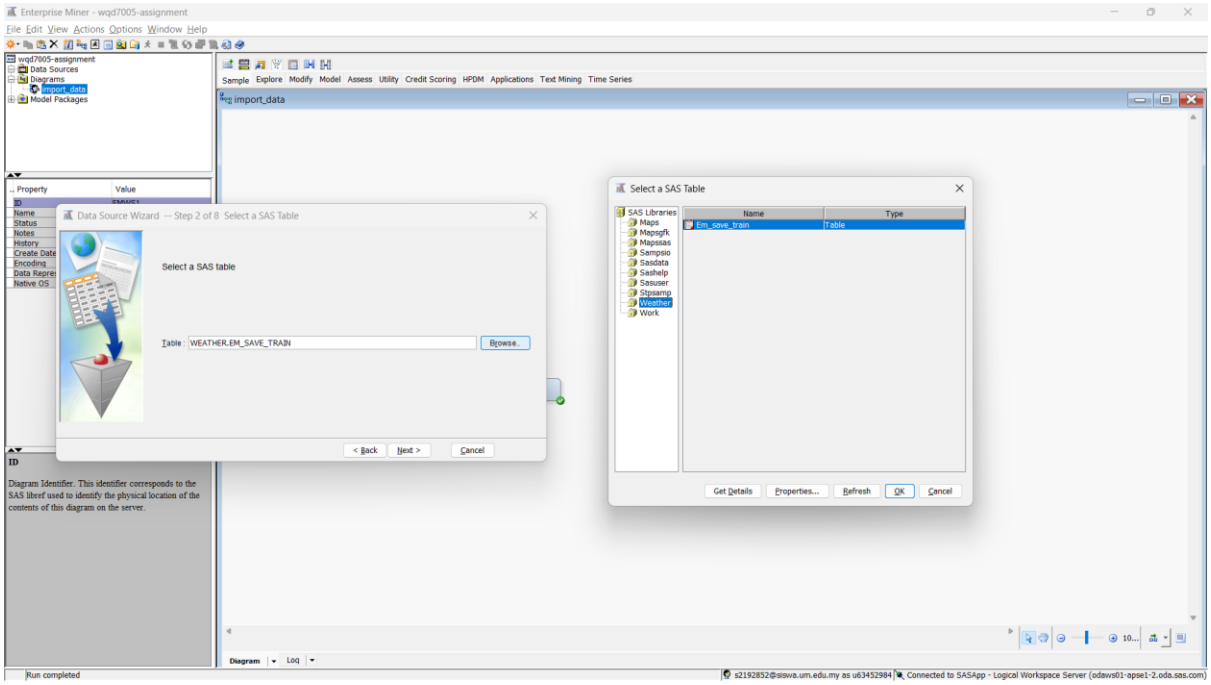

Enterprise Miner - wqd7005-assignment
File Edit View Actions Options Window Help
wqd7005-assignment
Data Sources
Diagrams
import_data
Model Packages
Sample Explore Modify Model Assess Utility Credit Scoring HPDM Applications Text Mining Time Series
import_data
Property
Value
ID
EMWS1
Name
import_data
Status
Open
Notes
History
Create Date
11/29/23 1:32 AM
Encoding
utf-8 Unicode (UTF-8)
Data Representation
SOLARIS_X86_64, LINUX_X
Native OS
Yes
ID
Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.
Library Wizard -- Step 2 of 3 Create or Modify
Name
WEATHER
Engine
BASE
Library information
Path
/home/u63452984/wqd7005-assignment/Workspaces/EMWS1/EMSave
Browse...
Options
< Back
Next >
Cancel
Diagram
Log
Run completed
s2192852@siswa.um.edu.my as u63452984
Connected to SASApp - Logical Workspace Server (odaws01-apse1-2.oda.sas.com)



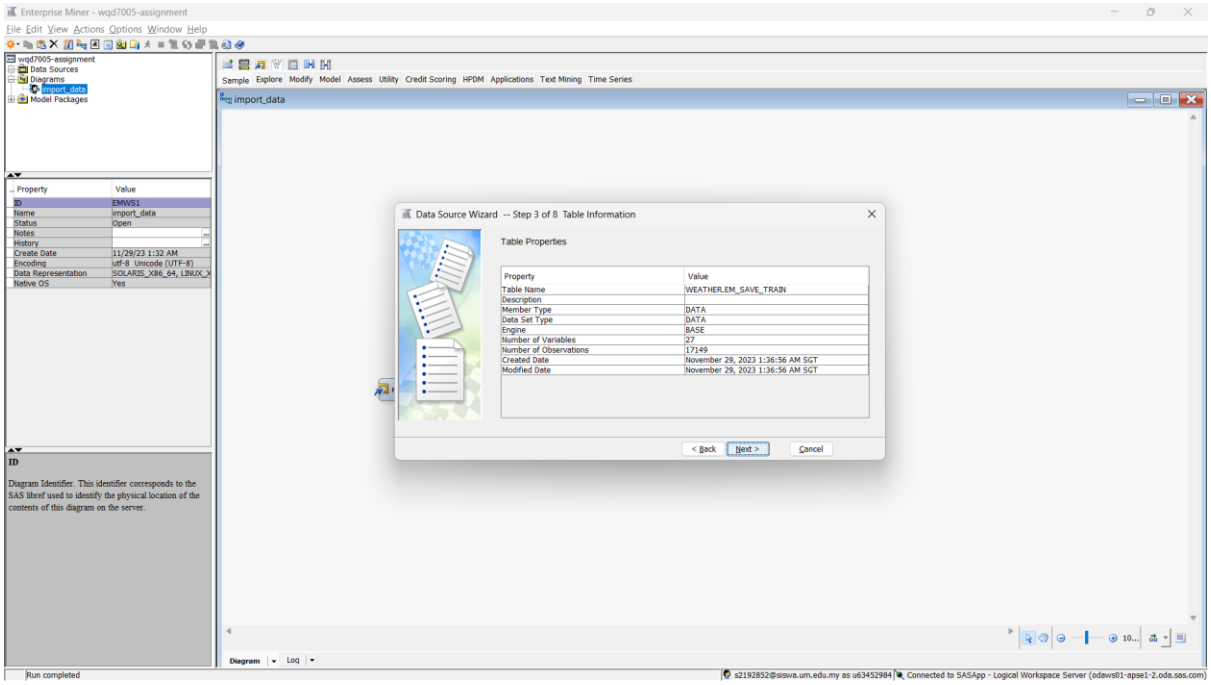
iv. Specifying the data source.



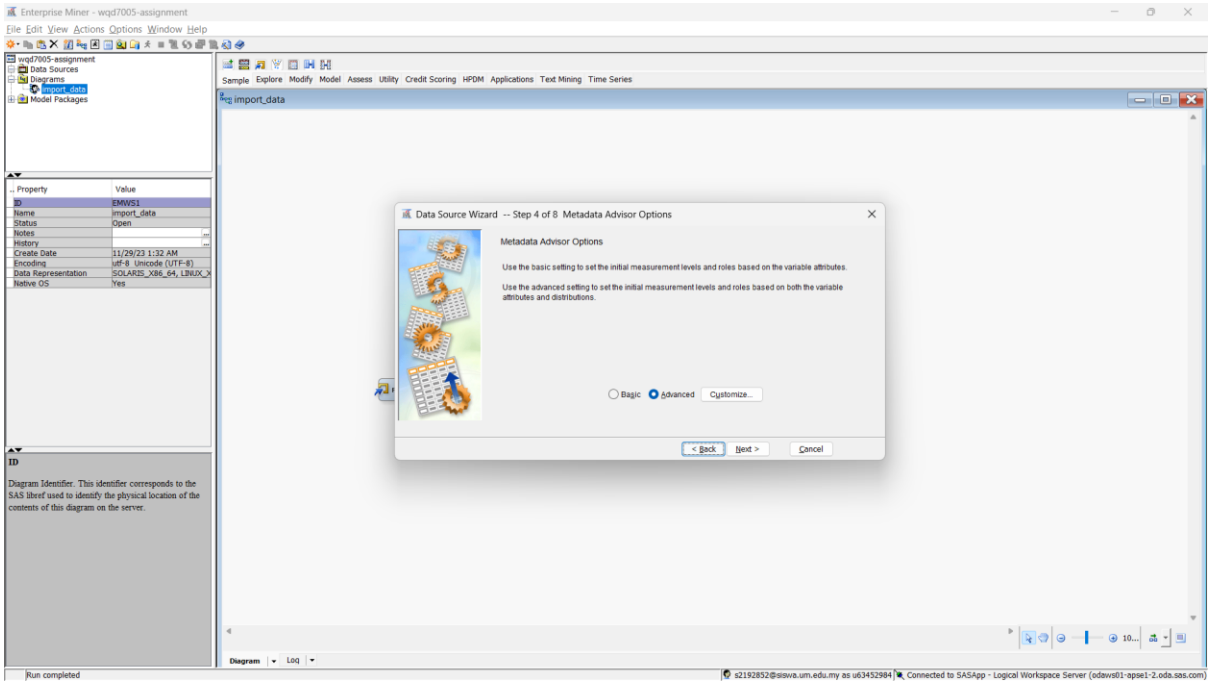

Enterprise Miner - wqd7005-assignment
File Edit View Actions Options Window Help
wqd7005-assignment
Data Sources
Diagrams
import_data
Model Packages
Sample Explore Modify Model Assess Utility Credit Scoring HPDM Applications Text Mining Time Series
import_data
Property
Value
ID
EMWS1
Name
import_data
Status
Open
Notes
History
Create Date
11/29/23 1:32 AM
Encoding
utf-8 Unicode (UTF-8)
Data Representation
SOLARIS_X86_64, LINUX_X
Native OS
Yes
ID
Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.
Data Source Wizard -- Step 1 of 8 Metadata Source
Select a metadata source
Source : SAS Table
< Back
Next >
Cancel
Diagram
Log
Run completed
s2192852@siswa.um.edu.my as u63452984
Connected to SASApp - Logical Workspace Server (odaws01-apse1-2.oda.sas.com)

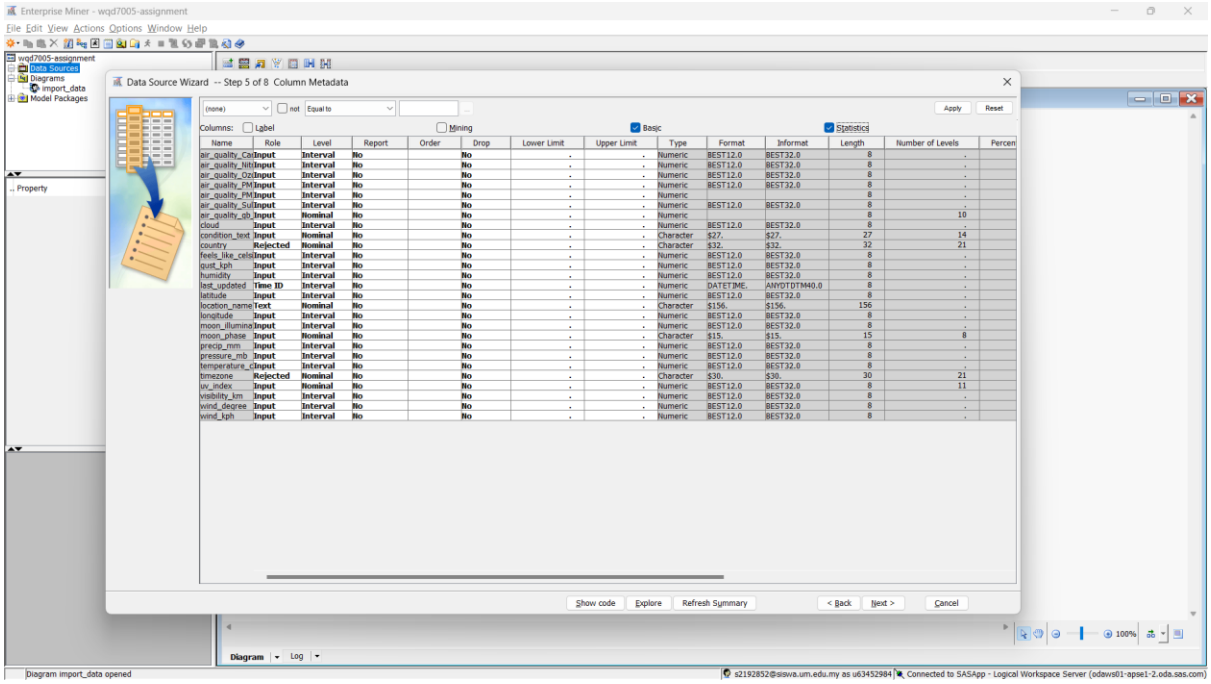

Enterprise Miner - wqd7005-assignment
File Edit View Actions Options Window Help
wqd7005-assignment
Data Sources
Diagrams
import_data
Model Packages
Sample Explore Modify Model Assess Utility Credit Scoring HPDM Applications Text Mining Time Series
import_data
Property
Value
ID
Name
Status
Notes
History
Create Date
Encoding
Data Repres
Native OS
ID
Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.
Data Source Wizard -- Step 2 of 8 Select a SAS Table
Select a SAS table
Table : WEATHER.EM_SAVE_TRAIN
Browse...
< Back
Next >
Cancel
Select a SAS Table
SAS Libraries
Maps
Mapsgfk
Mapssas
Sampsio
Sasdata
Sashelp
Sasuser
Stpsamp
Weather
Work
Name
Type
Em_save_train
Table
Get Details
Properties...
Refresh
OK
Cancel
Diagram
Log
Run completed
s2192852@siswa.um.edu.my as u63452984
Connected to SASApp - Logical Workspace Server (odaws01-apse1-2.oda.sas.com)

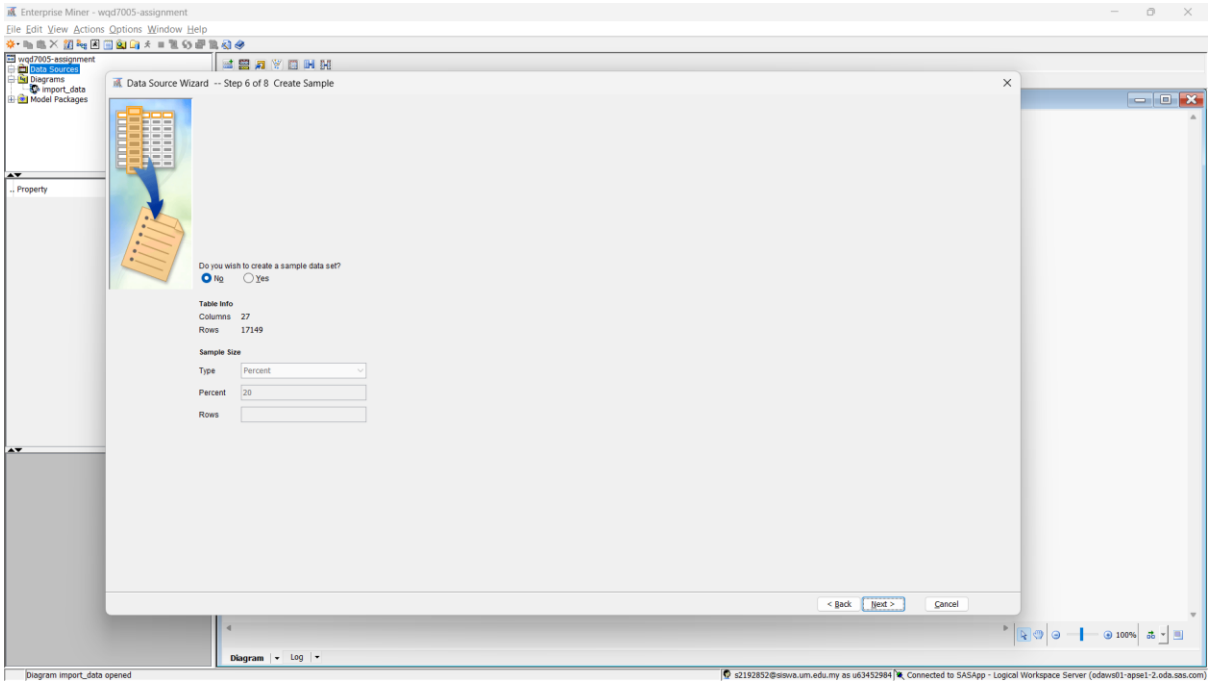
v. Defining column metadata.

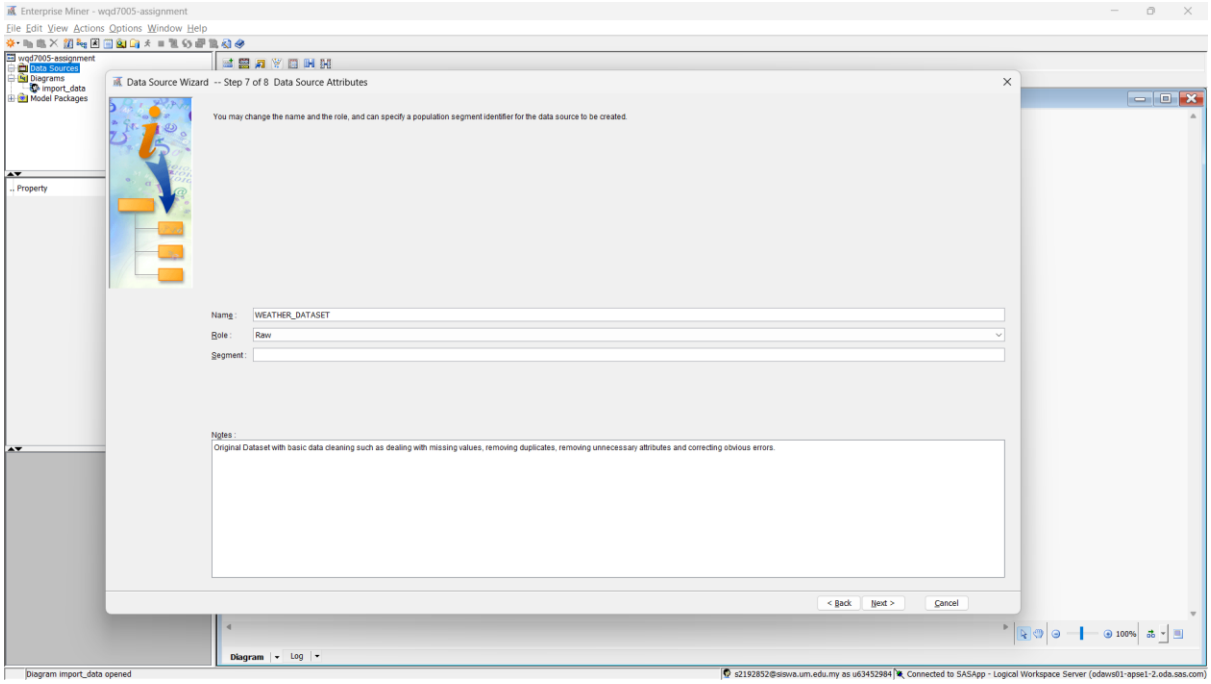


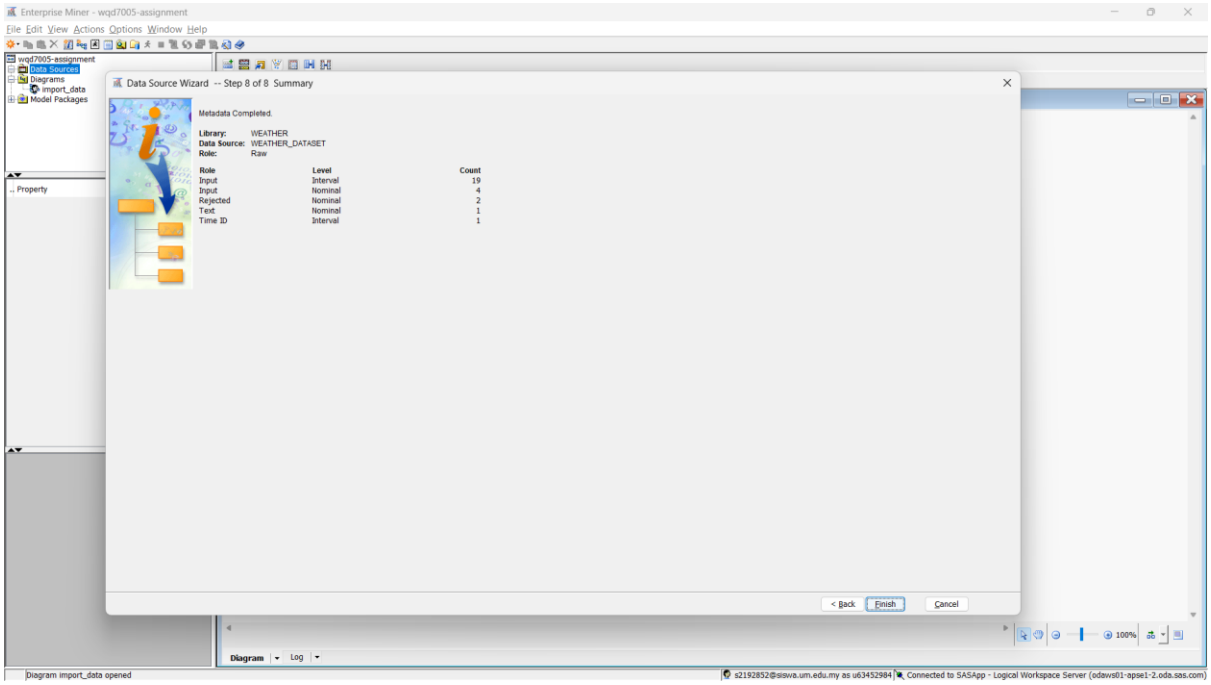
vi. Perform advanced setting.




Enterprise Miner - wqd7005-assignment
File Edit View Actions Options Window Help
wqd7005-assignment
Data Sources
Diagrams
import_data
Model Packages
Property
Data Source Wizard -- Step 5 of 8 Column Metadata
Columns: Label Mining Basic Statistics
Name Role Level Report Order Drop Lower Limit Upper Limit Type Format Informat Length Number of Levels
air_quality_Ca Input Interval No No Numeric BEST12.0 BEST32.0 8 .
air_quality_Nit Input Interval No No Numeric BEST12.0 BEST32.0 8 .
air_quality_O2 Input Interval No No Numeric BEST12.0 BEST32.0 8 .
air_quality_PM Input Interval No No Numeric BEST12.0 BEST32.0 8 .
air_quality_PM Input Interval No No Numeric 8 .
air_quality_Su Input Interval No No Numeric BEST12.0 BEST32.0 8 .
air_quality_gb Input Nominal No No Numeric 8 10
cloud Input Interval No No Numeric BEST12.0 BEST32.0 8 .
condition_text Input Nominal No No Character $27. $27. 27 14
country Rejected Nominal No No Character $32. $32. 32 21
feels_like_cels Input Interval No No Numeric BEST12.0 BEST32.0 8 .
gust_kph Input Interval No No Numeric BEST12.0 BEST32.0 8 .
humidity Input Interval No No Numeric BEST12.0 BEST32.0 8 .
last_updated Time ID Interval No No Numeric DATETIME. ANYDTDTM40.0 8 .
latitude Input Interval No No Numeric BEST12.0 BEST32.0 8 .
location_name Text Nominal No No Character $156. $156. 156 .
longitude Input Interval No No Numeric BEST12.0 BEST32.0 8 .
moon_illumina Input Interval No No Numeric BEST12.0 BEST32.0 8 .
moon_phase Input Nominal No No Character $15. $15. 15 8
precip_mm Input Interval No No Numeric BEST12.0 BEST32.0 8 .
pressure_mb Input Interval No No Numeric BEST12.0 BEST32.0 8 .
temperature_c Input Interval No No Numeric BEST12.0 BEST32.0 8 .
timezone Rejected Nominal No No Character $30. $30. 30 21
uv_index Input Nominal No No Numeric BEST12.0 BEST32.0 8 11
visibility_km Input Interval No No Numeric BEST12.0 BEST32.0 8 .
wind_degree Input Interval No No Numeric BEST12.0 BEST32.0 8 .
wind_kph Input Interval No No Numeric BEST12.0 BEST32.0 8 .
Apply Reset
Show code Explore Refresh Summary < Back Next > Cancel
Diagram Log
Diagram import_data opened
s2192852@siswa.um.edu.my as u63452984 Connected to SASApp - Logical Workspace Server (odaws01-apse1-2.oda.sas.com)


Enterprise Miner - wqd7005-assignment
File Edit View Actions Options Window Help
wqd7005-assignment
Data Sources
Diagrams
import_data
Model Packages
Property
Data Source Wizard -- Step 6 of 8 Create Sample
Do you wish to create a sample data set?
No Yes
Table Info
Columns 27
Rows 17149
Sample Size
Type Percent
Percent 20
Rows
< Back Next > Cancel
Diagram Log
Diagram import_data opened
s2192852@siswa.um.edu.my as u63452984 Connected to SASApp - Logical Workspace Server (odaws01-apse1-2.oda.sas.com)

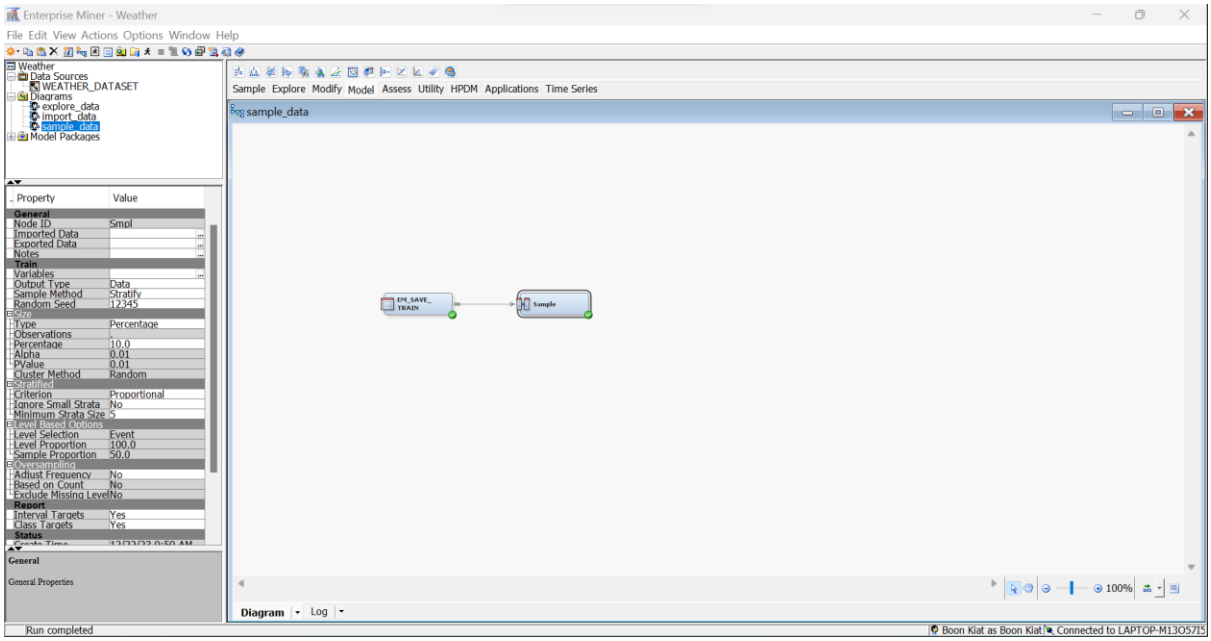


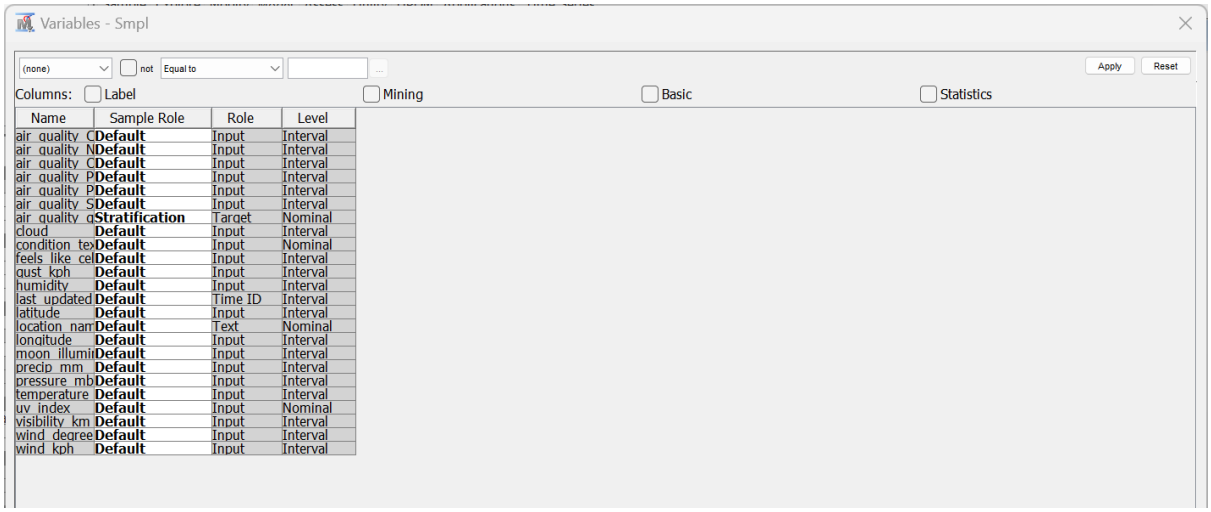


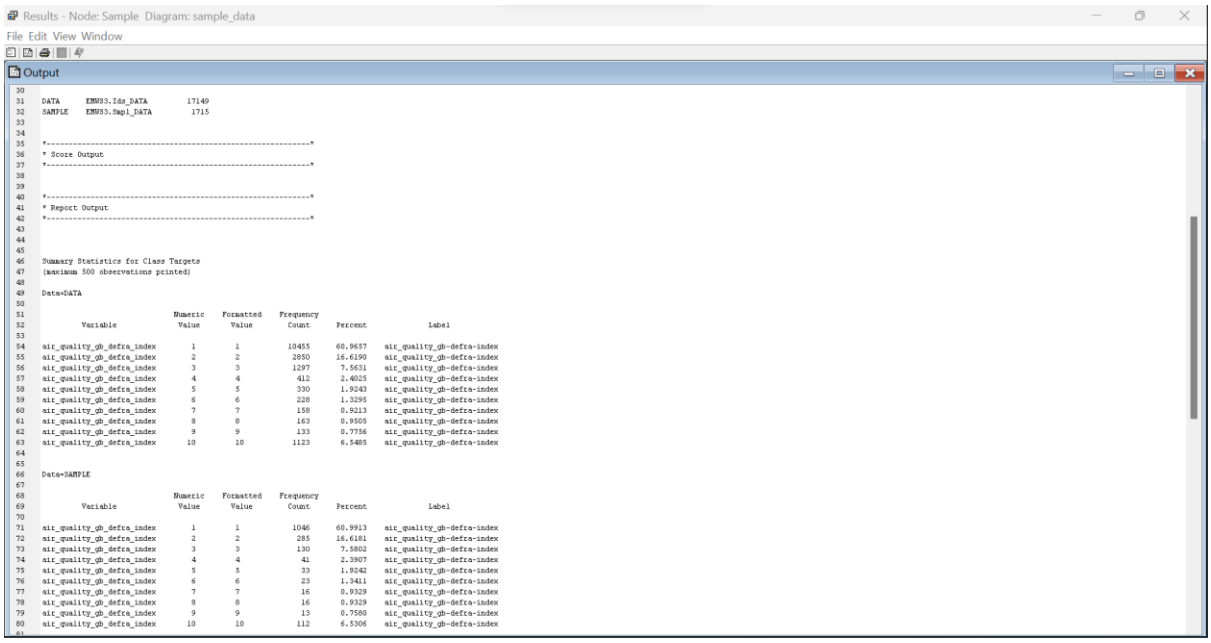
Sample Part

i. Generate samples to be used in Model phase.

Create a new diagram with the data source node and add another node named Sample. Specify the following: Sample Method = Stratify; Size Type = Percentage; Size Percentage = 10/50/100, Stratified Criterion = Proportional; Sample Role for Variables “air_quality_gb-defra-index” = Stratification.

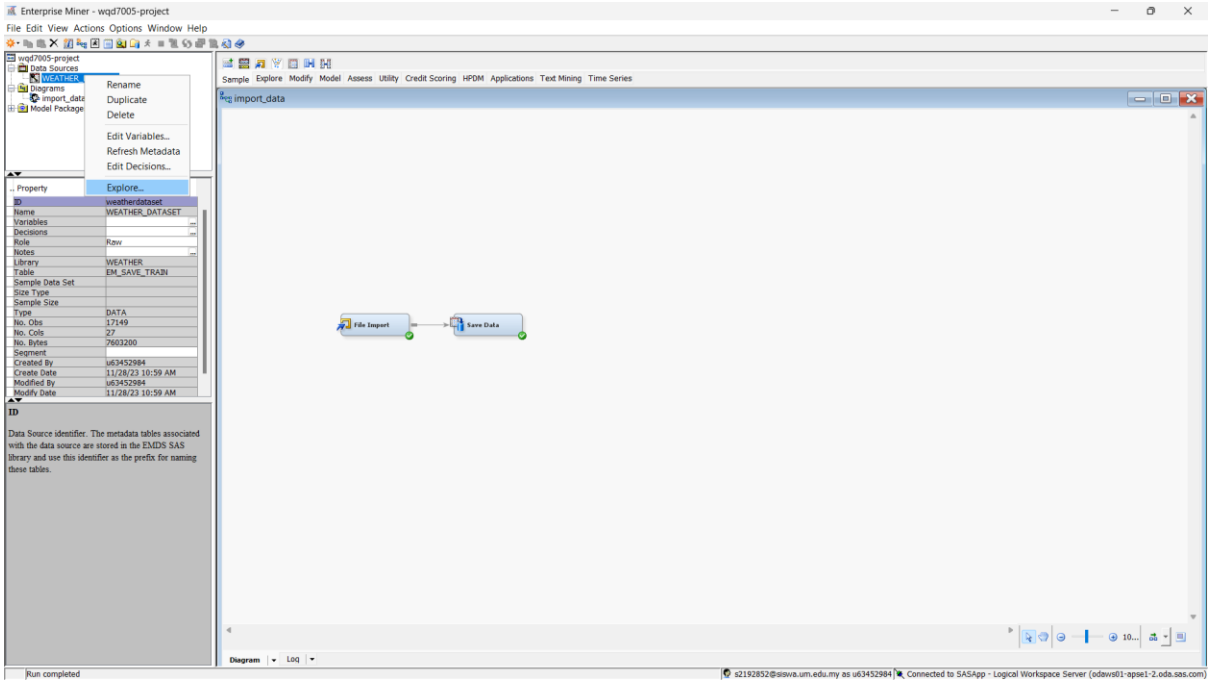

Enterprise Miner - Weather
File Edit View Actions Options Window Help
Weather
Data Sources
WEATHER_DATASET
Diagrams
explore_data
import_data
sample_data
Model Packages
Sample Explore Modify Model Assess Utility HPDM Applications Time Series
sample_data
EM_SAVE_TRAIN
Sample
Diagram
Log
Run completed
Boon Kiat as Boon Kiat
Connected to LAPTOP-M13O5715


Variables - Smpl
Apply
Reset
Columns:
Label
Mining
Basic
Statistics
Name
Sample Role
Role
Level
air_quality_C Default Input Interval
air_quality_N Default Input Interval
air_quality_C Default Input Interval
air_quality_P Default Input Interval
air_quality_P Default Input Interval
air_quality_S Default Input Interval
air_quality_g Stratification Target Nominal
cloud Default Input Interval
condition_tex Default Input Nominal
feels_like_cel Default Input Interval
gust_kph Default Input Interval
humidity Default Input Interval
last_updated Default Time ID Interval
latitude Default Input Interval
location_nam Default Text Nominal
longitude Default Input Interval
moon_illumi Default Input Interval
precip_mm Default Input Interval
pressure_mb Default Input Interval
temperature Default Input Interval
uv_index Default Input Nominal
visibility_km Default Input Interval
wind_degree Default Input Interval
wind_kph Default Input Interval

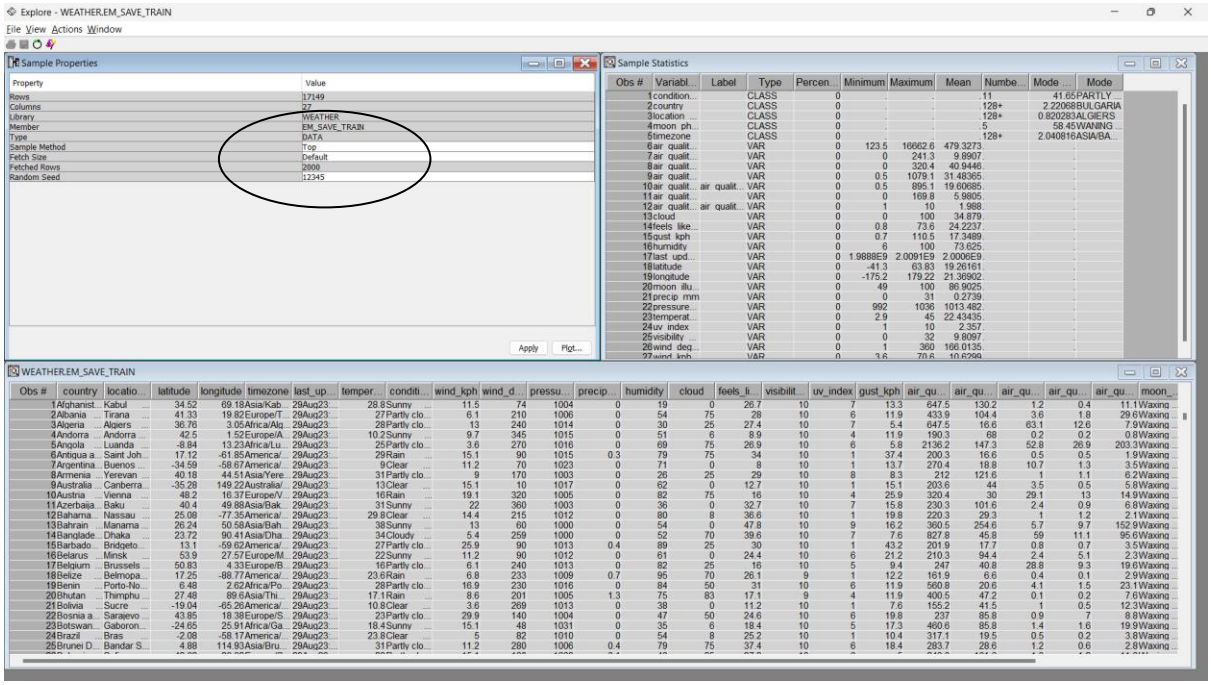

Results - Node: Sample Diagram: sample_data
File Edit View Window
Output
DATA EMWS3.Ids_DATA 17149
SAMPLE EMWS3.Smpl_DATA 1715
* Score Output
* Report Output
Summary Statistics for Class Targets
(maximum 500 observations printed)
Data=DATA
Variable Numeric Value Formatted Value Frequency Count Percent Label
air_quality_gb_defra_index 1 1 10455 60.9657 air_quality_gb_defra_index
air_quality_gb_defra_index 2 2 2850 16.6190 air_quality_gb_defra_index
air_quality_gb_defra_index 3 3 1297 7.5631 air_quality_gb_defra_index
air_quality_gb_defra_index 4 4 412 2.4025 air_quality_gb_defra_index
air_quality_gb_defra_index 5 5 330 1.9243 air_quality_gb_defra_index
air_quality_gb_defra_index 6 6 228 1.3295 air_quality_gb_defra_index
air_quality_gb_defra_index 7 7 158 0.9213 air_quality_gb_defra_index
air_quality_gb_defra_index 8 8 163 0.9505 air_quality_gb_defra_index
air_quality_gb_defra_index 9 9 133 0.7756 air_quality_gb_defra_index
air_quality_gb_defra_index 10 10 1123 6.5485 air_quality_gb_defra_index
Data=SAMPLE
Variable Numeric Value Formatted Value Frequency Count Percent Label
air_quality_gb_defra_index 1 1 1046 60.9913 air_quality_gb_defra_index
air_quality_gb_defra_index 2 2 285 16.6181 air_quality_gb_defra_index
air_quality_gb_defra_index 3 3 130 7.5802 air_quality_gb_defra_index
air_quality_gb_defra_index 4 4 41 2.3907 air_quality_gb_defra_index
air_quality_gb_defra_index 5 5 33 1.9242 air_quality_gb_defra_index
air_quality_gb_defra_index 6 6 23 1.3411 air_quality_gb_defra_index
air_quality_gb_defra_index 7 7 16 0.9329 air_quality_gb_defra_index
air_quality_gb_defra_index 8 8 16 0.9329 air_quality_gb_defra_index
air_quality_gb_defra_index 9 9 13 0.7580 air_quality_gb_defra_index
air_quality_gb_defra_index 10 10 112 6.5306 air_quality_gb_defra_index

Explore Part

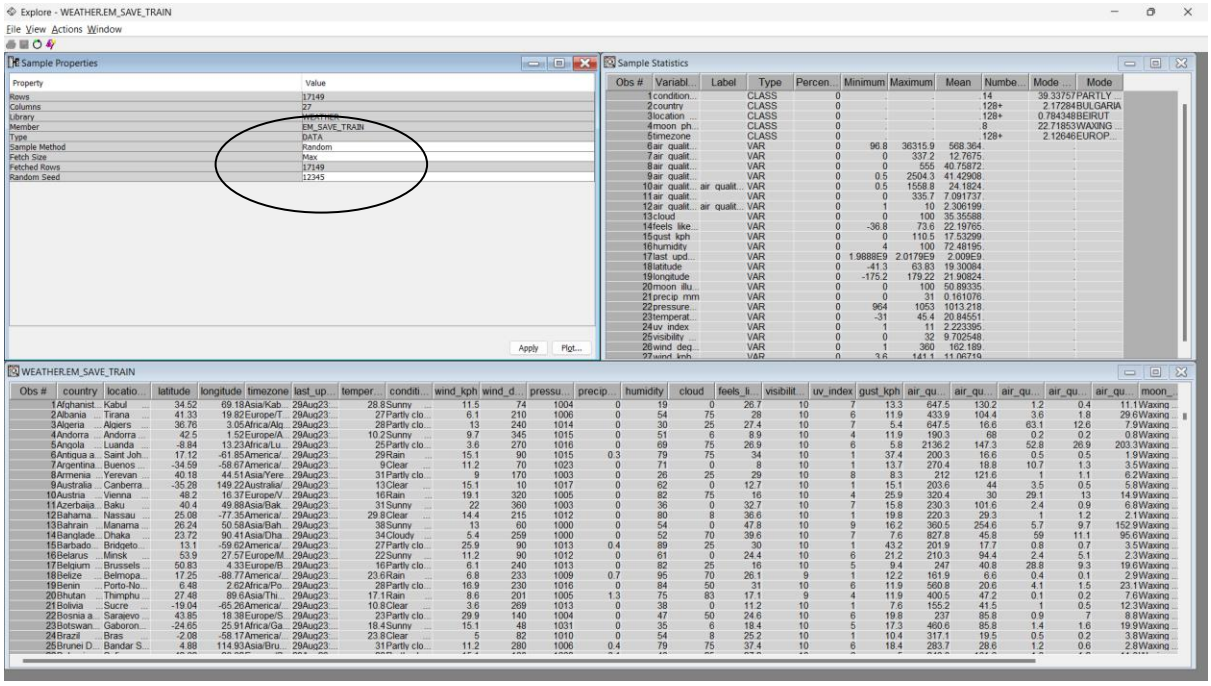
i. Changing the Explore Sample Size.



The initial explore only fetch 2000-observation sample.

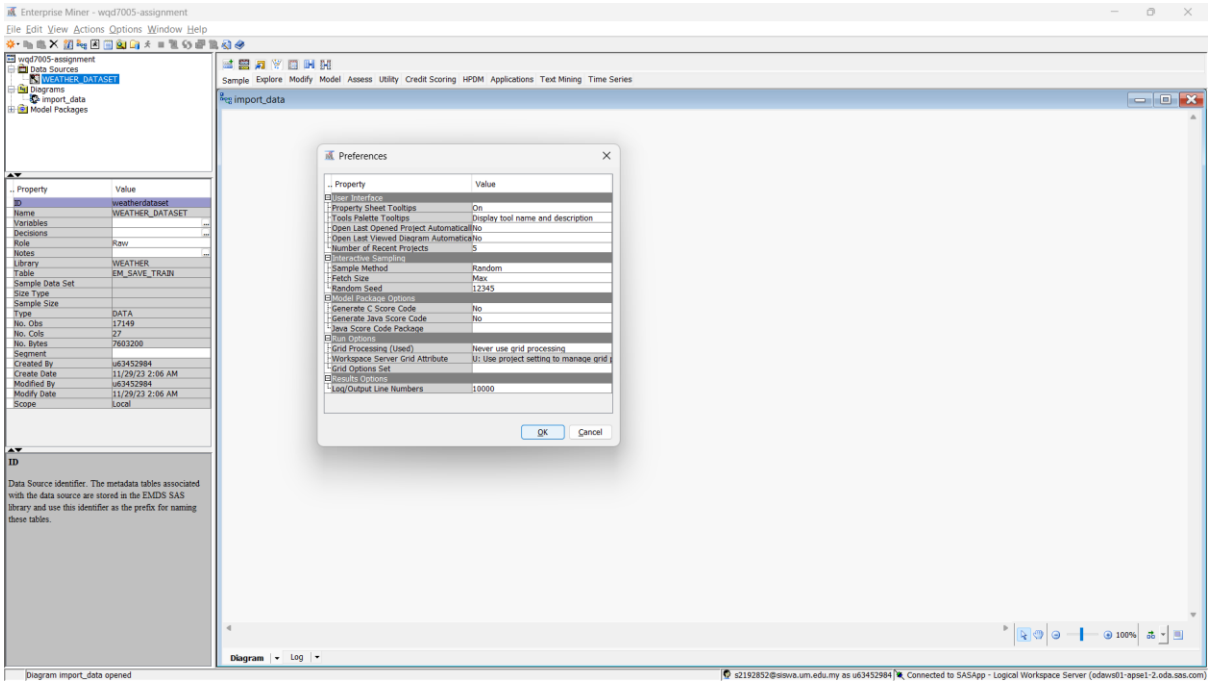


Change the setting to random and set max for the fetching size.



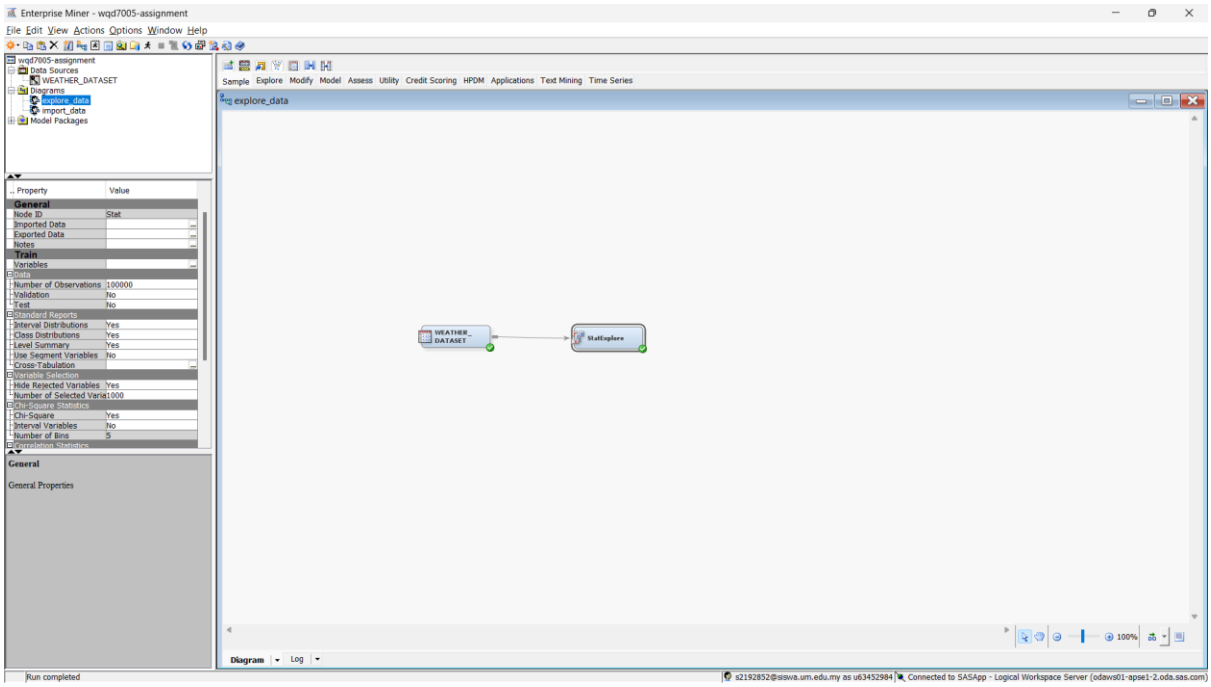
ii. Changing the explore window sampling default.

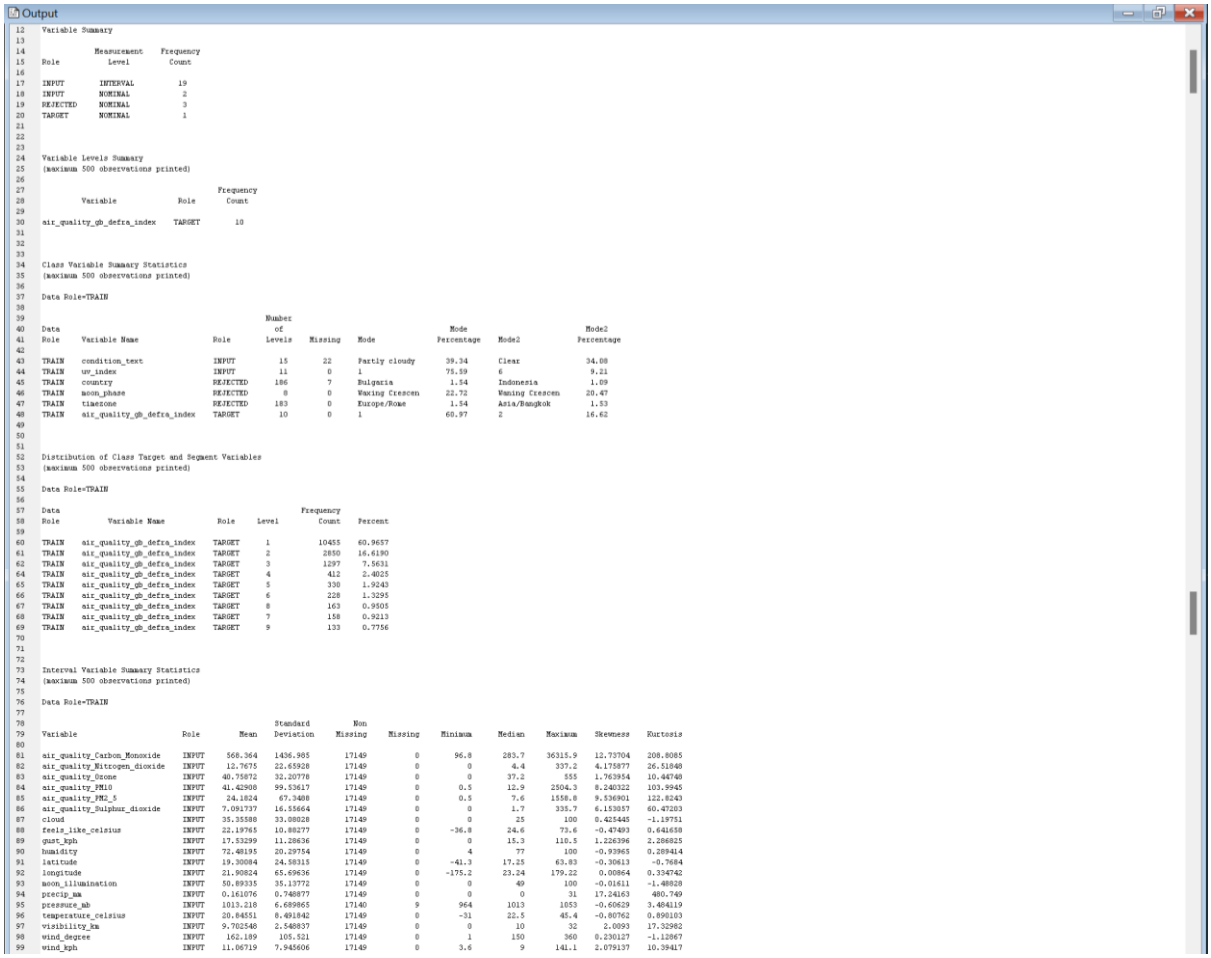
Select Options -> Preferences from the main menu.



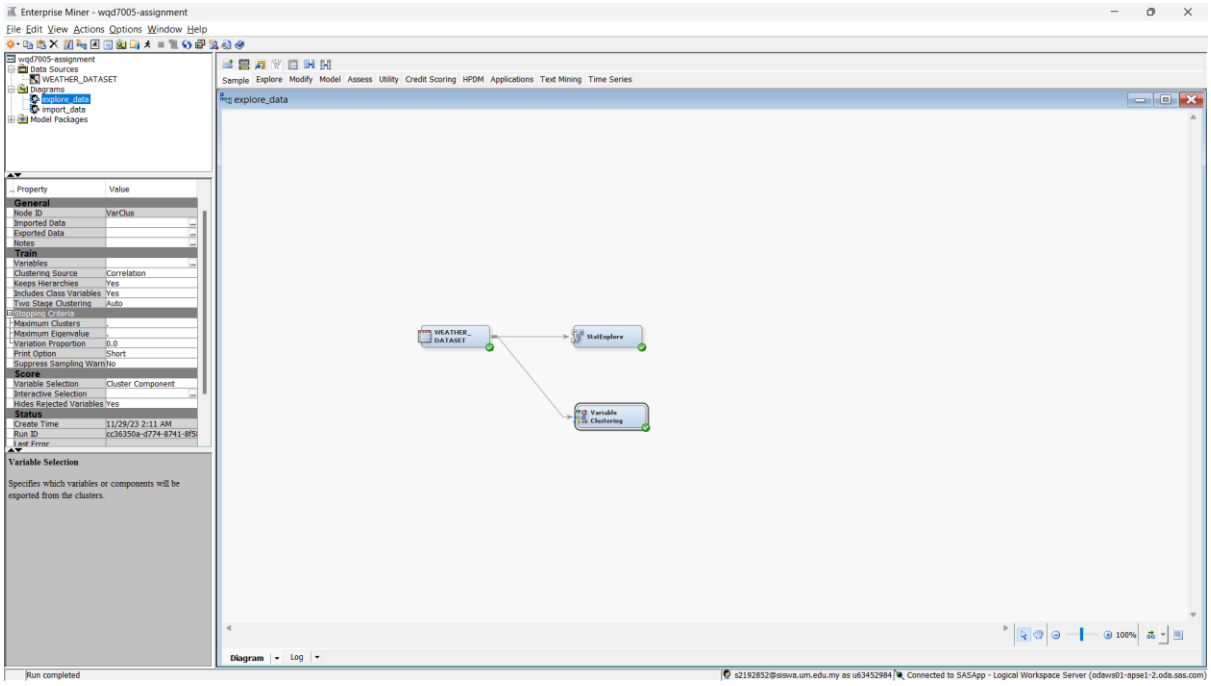
iii. Generate a Statistics Table.

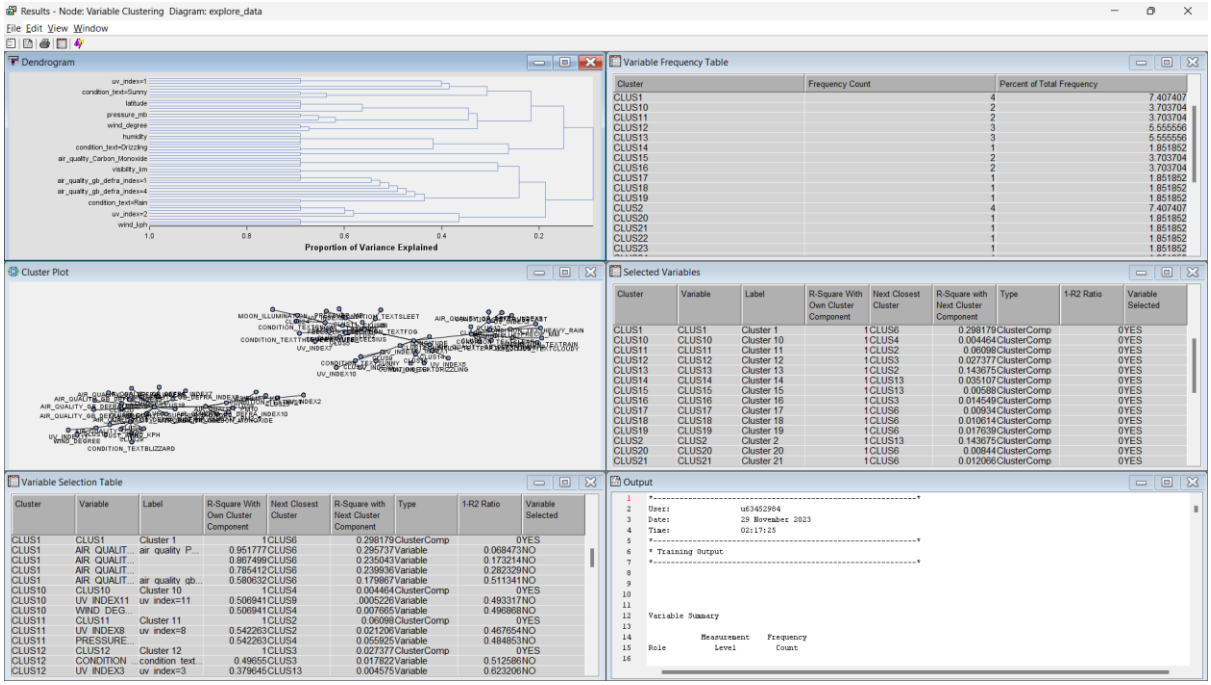
Create a new diagram with the data sources node and add another node called StatExplore.



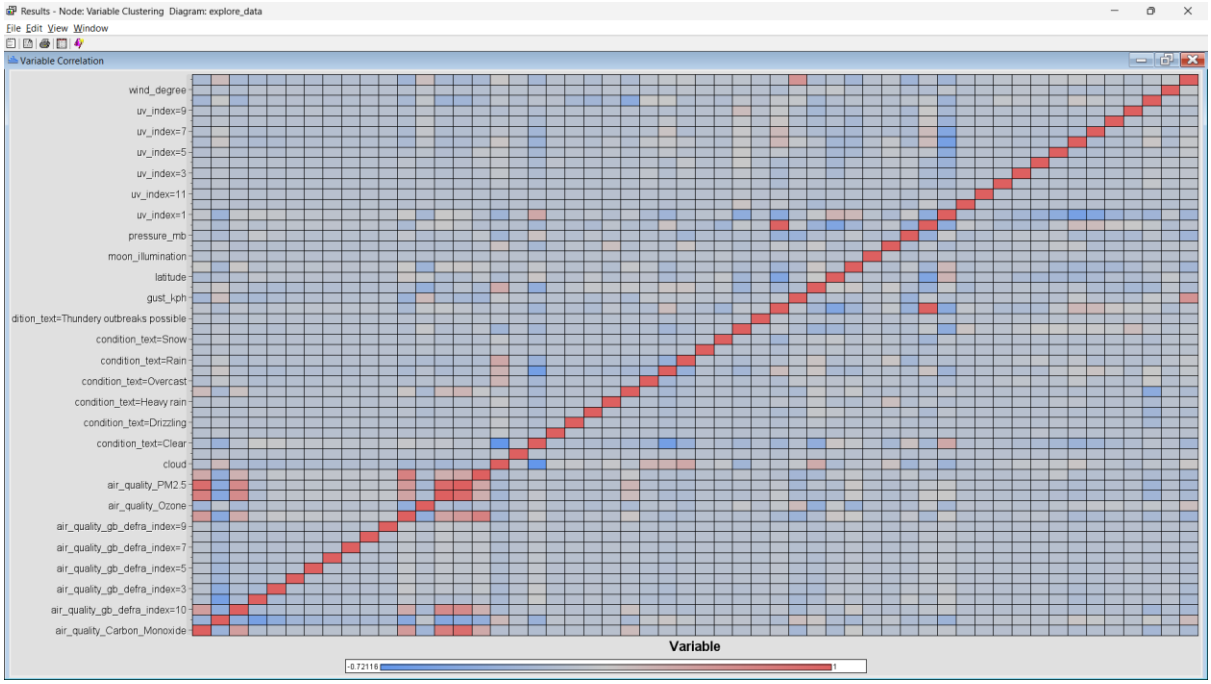


iv. To show the correlation matrix of variables using variable clustering.





The correlation matrix can be found from the view tab (located at top left) > model > variable correlation. The result is shown below.



As we have tons of variables including nominal used for the correlation analysis. Therefore, it is hard to analyze the correlation value from the chart. Therefore, table below shows the complete correlation values between the variables.

Variable	Variable	Correlation
air_quality_gb_defra_index=1	air_quality_Carbon_Monoxide	-0.2454752835669658
air_quality_gb_defra_index=1	air_quality_Nitrogen_dioxide	-0.3971005581960198
air_quality_gb_defra_index=1	air_quality_Ozone	0.056588090302376506
air_quality_gb_defra_index=1	air_quality_PM10	-0.420545905824364
air_quality_gb_defra_index=1	air_quality_PM2.5	-0.36957048053859254
air_quality_gb_defra_index=1	air_quality_Sulphur_dioxide	-0.31139610684276486
air_quality_gb_defra_index=1	air_quality_gb_defra_index=1	1.0
air_quality_gb_defra_index=1	air_quality_gb_defra_index=10	-0.330823592530848
air_quality_gb_defra_index=1	air_quality_gb_defra_index=2	-0.5579416230914609
air_quality_gb_defra_index=1	air_quality_gb_defra_index=3	-0.35747619018354176
air_quality_gb_defra_index=1	air_quality_gb_defra_index=4	-0.19607804337174767
air_quality_gb_defra_index=1	air_quality_gb_defra_index=5	-0.17505561648072965
air_quality_gb_defra_index=1	air_quality_gb_defra_index=6	-0.14506865193167776
air_quality_gb_defra_index=1	air_quality_gb_defra_index=7	-0.12051420442130119
air_quality_gb_defra_index=1	air_quality_gb_defra_index=8	-0.12242423725457974
air_quality_gb_defra_index=1	air_quality_gb_defra_index=9	-0.11048828858653424
air_quality_gb_defra_index=1	condition_text=Blizzard	0.006110462708841677
air_quality_gb_defra_index=1	condition_text=Clear	-0.2569633767844627
air_quality_gb_defra_index=1	condition_text=Cloudy	0.0321180713609366
air_quality_gb_defra_index=1	condition_text=Drizzling	0.01802738212012803
air_quality_gb_defra_index=1	condition_text=Fog	-0.007173602075546932
air_quality_gb_defra_index=1	condition_text=Heavy rain	0.01972447126595972
air_quality_gb_defra_index=1	condition_text=Mist	-0.10123978865036912
air_quality_gb_defra_index=1	condition_text=Overcast	0.08853255965172815
air_quality_gb_defra_index=1	condition_text=Partly cloudy	0.16058515588894573

air_quality_gb_defra_index=1	condition_text=Rain	0.0783864971814342
air_quality_gb_defra_index=1	condition_text=Sleet	0.01728652710471633
air_quality_gb_defra_index=1	condition_text=Snow	0.05053478361643161
air_quality_gb_defra_index=1	condition_text=Sunny	0.04133934644194007
air_quality_gb_defra_index=1	condition_text=Thundery outbreaks possible	-0.0014357058355244415
air_quality_gb_defra_index=1	precip_mm	0.05302470521240489
air_quality_gb_defra_index=1	visibility_km	0.13345447489192822
air_quality_gb_defra_index=1	wind_kph	0.2485593296718882
air_quality_gb_defra_index=10	air_quality_Carbon_Monoxide	0.48967504364509384
air_quality_gb_defra_index=10	air_quality_Nitrogen_dioxide	0.38172826010878824
air_quality_gb_defra_index=10	air_quality_Ozone	-0.11915886077506546
air_quality_gb_defra_index=10	air_quality_PM10	0.6400809647117754
air_quality_gb_defra_index=10	air_quality_PM2.5	0.650950452812413
air_quality_gb_defra_index=10	air_quality_Sulphur_dioxide	0.3262025712152931
air_quality_gb_defra_index=10	air_quality_gb_defra_index=1	-0.330823592530848
air_quality_gb_defra_index=10	air_quality_gb_defra_index=10	1.0
air_quality_gb_defra_index=10	air_quality_gb_defra_index=2	-0.11818079464850707
air_quality_gb_defra_index=10	air_quality_gb_defra_index=3	-0.07571906894081368
air_quality_gb_defra_index=10	air_quality_gb_defra_index=4	-0.04153240772825258
air_quality_gb_defra_index=10	air_quality_gb_defra_index=5	-0.037079527690992146
air_quality_gb_defra_index=10	air_quality_gb_defra_index=6	-0.03072781784746496
air_quality_gb_defra_index=10	air_quality_gb_defra_index=7	-0.025526800395401398
air_quality_gb_defra_index=10	air_quality_gb_defra_index=8	-0.025931375334246914
air_quality_gb_defra_index=10	air_quality_gb_defra_index=9	-0.02340315402919799
air_quality_gb_defra_index=10	condition_text=Blizzard	-0.0020214852253647806
air_quality_gb_defra_index=10	condition_text=Clear	0.017290915889993812
air_quality_gb_defra_index=10	condition_text=Cloudy	0.004948657827258274
air_quality_gb_defra_index=10	condition_text=Drizzling	-0.010234267284199535
air_quality_gb_defra_index=10	condition_text=Fog	-0.012910784619195126
air_quality_gb_defra_index=10	condition_text=Heavy rain	-0.0036918306093320016
air_quality_gb_defra_index=10	condition_text=Mist	0.19861521174235736
air_quality_gb_defra_index=10	condition_text=Overcast	-0.0069758515161999785
air_quality_gb_defra_index=10	condition_text=Partly cloudy	-0.0331799370577337
air_quality_gb_defra_index=10	condition_text=Rain	-0.05953341327589536
air_quality_gb_defra_index=10	condition_text=Sleet	-0.005718790999164133
air_quality_gb_defra_index=10	condition_text=Snow	-0.017893493056049276
air_quality_gb_defra_index=10	condition_text=Sunny	-0.004728971266163042
air_quality_gb_defra_index=10	condition_text=Thundery outbreaks possible	-0.0072288918539409 35
air_quality_gb_defra_index=10	precip_mm	-0.028162523087915326
air_quality_gb_defra_index=10	visibility_km	-0.16978842569059358
air_quality_gb_defra_index=10	wind_kph	-0.13060764153102558
air_quality_gb_defra_index=2	air_quality_Carbon_Monoxide	-0.03724097972843525
air_quality_gb_defra_index=2	air_quality_Nitrogen_dioxide	0.05248383729720807
air_quality_gb_defra_index=2	air_quality_Ozone	-0.01794431706022523
air_quality_gb_defra_index=2	air_quality_PM10	-0.04717726380068615
air_quality_gb_defra_index=2	air_quality_PM2.5	-0.0470176327664564
air_quality_gb_defra_index=2	air_quality_Sulphur_dioxide	0.0066311644743843115
air_quality_gb_defra_index=2	air_quality_gb_defra_index=1	-0.5579416230914609
air_quality_gb_defra_index=2	air_quality_gb_defra_index=10	-0.11818079464850707
air_quality_gb_defra_index=2	air_quality_gb_defra_index=2	1.0
air_quality_gb_defra_index=2	air_quality_gb_defra_index=3	-0.12770195710837176
air_quality_gb_defra_index=2	air_quality_gb_defra_index=4	-0.07004536406101936
air_quality_gb_defra_index=2	air_quality_gb_defra_index=5	-0.062535479120788
air_quality_gb_defra_index=2	air_quality_gb_defra_index=6	-0.0518231738937262
air_quality_gb_defra_index=2	air_quality_gb_defra_index=7	-0.0430515379389514
air_quality_gb_defra_index=2	air_quality_gb_defra_index=8	-0.043733862909533856
air_quality_gb_defra_index=2	air_quality_gb_defra_index=9	-0.03946995933578293
air_quality_gb_defra_index=2	condition_text=Blizzard	-0.0034092814816109695

air_quality_gb_defra_index=2	condition_text=Clear	0.13338331170360201
air_quality_gb_defra_index=2	condition_text=Cloudy	-0.017220878363325823
air_quality_gb_defra_index=2	condition_text=Drizzling	-0.010271719425897469
air_quality_gb_defra_index=2	condition_text=Fog	0.015180454564498817
air_quality_gb_defra_index=2	condition_text=Heavy rain	-0.013570387172458278
air_quality_gb_defra_index=2	condition_text=Mist	-0.008982481461819675
air_quality_gb_defra_index=2	condition_text=Overcast	-0.0547854040383378
air_quality_gb_defra_index=2	condition_text=Partly cloudy	-0.06834186655883193
air_quality_gb_defra_index=2	condition_text=Rain	-0.02721354806040737
air_quality_gb_defra_index=2	condition_text=Sleet	-0.009644872990419958
air_quality_gb_defra_index=2	condition_text=Snow	-0.025521748448515794
air_quality_gb_defra_index=2	condition_text=Sunny	-0.024499477347472657
air_quality_gb_defra_index=2	condition_text=Thundery outbreaks possible	0.009625037037565823
air_quality_gb_defra_index=2	precip_mm	-0.013588098937610354
air_quality_gb_defra_index=2	visibility_km	-0.018693947047743013
air_quality_gb_defra_index=2	wind_kph	-0.10661222880544897
air_quality_gb_defra_index=3	air_quality_Carbon_Monoxide	0.005848666825041234
air_quality_gb_defra_index=3	air_quality_Nitrogen_dioxide	0.11029193913998916
air_quality_gb_defra_index=3	air_quality_Ozone	0.012117139257853668
air_quality_gb_defra_index=3	air_quality_PM10	0.051903116193978464
air_quality_gb_defra_index=3	air_quality_PM2.5	0.02133570117258179
air_quality_gb_defra_index=3	air_quality_Sulphur_dioxide	0.08030166447537847
air_quality_gb_defra_index=3	air_quality_gb_defra_index=1	-0.35747619018354176
air_quality_gb_defra_index=3	air_quality_gb_defra_index=10	-0.07571906894081368
air_quality_gb_defra_index=3	air_quality_gb_defra_index=2	-0.12770195710837176
air_quality_gb_defra_index=3	air_quality_gb_defra_index=3	1.0
air_quality_gb_defra_index=3	air_quality_gb_defra_index=4	-0.044878440410687476
air_quality_gb_defra_index=3	air_quality_gb_defra_index=5	-0.040066816853592535
air_quality_gb_defra_index=3	air_quality_gb_defra_index=6	-0.03320338544398509
air_quality_gb_defra_index=3	air_quality_gb_defra_index=7	-0.027583351245038325
air_quality_gb_defra_index=3	air_quality_gb_defra_index=8	-0.028020520513033477
air_quality_gb_defra_index=3	air_quality_gb_defra_index=9	-0.02528861462580303
air_quality_gb_defra_index=3	condition_text=Blizzard	-0.0021843449294153268
air_quality_gb_defra_index=3	condition_text=Clear	0.12192054344805732
air_quality_gb_defra_index=3	condition_text=Cloudy	-0.018759570776205638
air_quality_gb_defra_index=3	condition_text=Drizzling	-0.008736836570400162
air_quality_gb_defra_index=3	condition_text=Fog	0.006961890823211415
air_quality_gb_defra_index=3	condition_text=Heavy rain	-0.008045124326130401
air_quality_gb_defra_index=3	condition_text=Mist	0.0038160435291695136
air_quality_gb_defra_index=3	condition_text=Overcast	-0.03954683863591354
air_quality_gb_defra_index=3	condition_text=Partly cloudy	-0.0750372655567171
air_quality_gb_defra_index=3	condition_text=Rain	-0.01263784215754659
air_quality_gb_defra_index=3	condition_text=Sleet	-0.006179521850898523
air_quality_gb_defra_index=3	condition_text=Snow	-0.019335071231824874
air_quality_gb_defra_index=3	condition_text=Sunny	-0.03006398204836583
air_quality_gb_defra_index=3	condition_text=Thundery outbreaks possible	9.58432917608479E-4
air_quality_gb_defra_index=3	precip_mm	-0.019930848619778424
air_quality_gb_defra_index=3	visibility_km	-0.017626270061795575
air_quality_gb_defra_index=3	wind_kph	-0.09003699558085437
air_quality_gb_defra_index=4	air_quality_Carbon_Monoxide	0.00864623072951625
air_quality_gb_defra_index=4	air_quality_Nitrogen_dioxide	0.08556710373886106
air_quality_gb_defra_index=4	air_quality_Ozone	0.02200573328600698
air_quality_gb_defra_index=4	air_quality_PM10	0.0572916588113832
air_quality_gb_defra_index=4	air_quality_PM2.5	0.03401050139084467
air_quality_gb_defra_index=4	air_quality_Sulphur_dioxide	0.06919271972931089
air_quality_gb_defra_index=4	air_quality_gb_defra_index=1	-0.19607804337174767
air_quality_gb_defra_index=4	air_quality_gb_defra_index=10	-0.04153240772825258
air_quality_gb_defra_index=4	air_quality_gb_defra_index=2	-0.07004536406101936

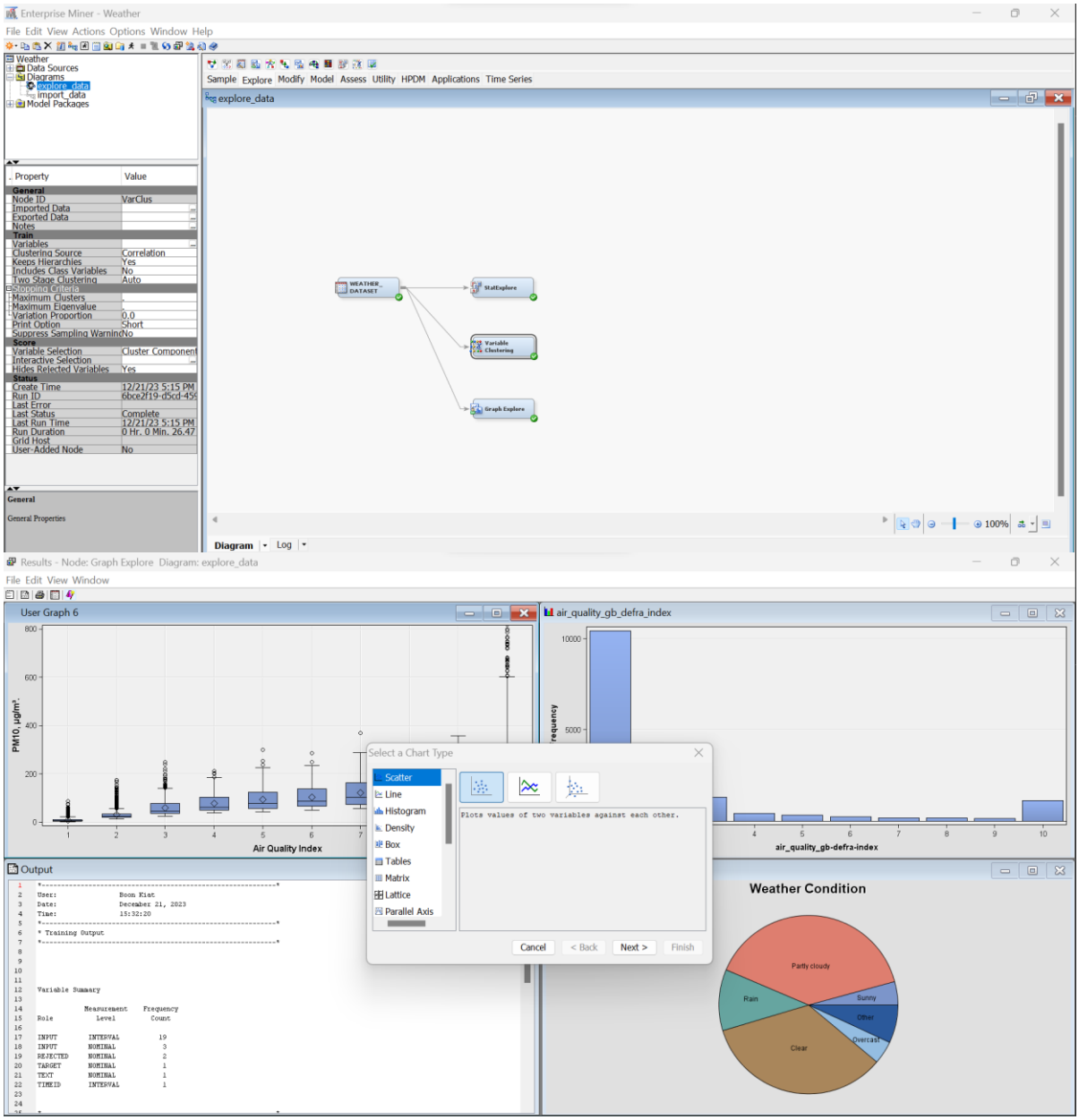
air_quality_gb_defra_index=4	air_quality_gb_defra_index=3	-0.044878440410687476
air_quality_gb_defra_index=4	air_quality_gb_defra_index=4	1.0
air_quality_gb_defra_index=4	air_quality_gb_defra_index=5	-0.02197691278054884
air_quality_gb_defra_index=4	air_quality_gb_defra_index=6	-0.01821227547443607
air_quality_gb_defra_index=4	air_quality_gb_defra_index=7	-0.015129649722925189
air_quality_gb_defra_index=4	air_quality_gb_defra_index=8	-0.01536943994404939
air_quality_gb_defra_index=4	air_quality_gb_defra_index=9	-0.013870971582369487
air_quality_gb_defra_index=4	condition_text=Blizzard	-0.0011981275720457049
air_quality_gb_defra_index=4	condition_text=Clear	0.07067479654470482
air_quality_gb_defra_index=4	condition_text=Cloudy	-0.011894547106156478
air_quality_gb_defra_index=4	condition_text=Drizzling	0.0012907476515755737
air_quality_gb_defra_index=4	condition_text=Fog	-0.0010485541661700351
air_quality_gb_defra_index=4	condition_text=Heavy rain	-0.009085146695954163
air_quality_gb_defra_index=4	condition_text=Mist	-0.003432711719262415
air_quality_gb_defra_index=4	condition_text=Overcast	-0.022349683366431145
air_quality_gb_defra_index=4	condition_text=Partly cloudy	-0.044490041492267096
air_quality_gb_defra_index=4	condition_text=Rain	-0.015803431953691587
air_quality_gb_defra_index=4	condition_text=Sleet	-0.0033895084113854594
air_quality_gb_defra_index=4	condition_text=Snow	-0.010605413842088482
air_quality_gb_defra_index=4	condition_text=Sunny	0.005229124651909916
air_quality_gb_defra_index=4	condition_text=Thundery outbreaks possible	9.117091556103372E-4
air_quality_gb_defra_index=4	precip_mm	-0.017779911301276912
air_quality_gb_defra_index=4	visibility_km	-0.009958238618799482
air_quality_gb_defra_index=4	wind_kph	-0.042708737689665134
air_quality_gb_defra_index=5	air_quality_Carbon_Monoxide	0.0194080232400591
air_quality_gb_defra_index=5	air_quality_Nitrogen_dioxide	0.08399963349457511
air_quality_gb_defra_index=5	air_quality_Ozone	-7.562036673205341E-4
air_quality_gb_defra_index=5	air_quality_PM10	0.07410620388645688
air_quality_gb_defra_index=5	air_quality_PM2.5	0.04309656352189026
air_quality_gb_defra_index=5	air_quality_Sulphur_dioxide	0.0702640738568029
air_quality_gb_defra_index=5	air_quality_gb_defra_index=1	-0.17505561648072965
air_quality_gb_defra_index=5	air_quality_gb_defra_index=10	-0.037079527690992146
air_quality_gb_defra_index=5	air_quality_gb_defra_index=2	-0.062535479120788
air_quality_gb_defra_index=5	air_quality_gb_defra_index=3	-0.040066816853592535
air_quality_gb_defra_index=5	air_quality_gb_defra_index=4	-0.02197691278054884
air_quality_gb_defra_index=5	air_quality_gb_defra_index=5	1.0
air_quality_gb_defra_index=5	air_quality_gb_defra_index=6	-0.016259653839210297
air_quality_gb_defra_index=5	air_quality_gb_defra_index=7	-0.013507530541615907
air_quality_gb_defra_index=5	air_quality_gb_defra_index=8	-0.013721611752663912
air_quality_gb_defra_index=5	air_quality_gb_defra_index=9	-0.012383801060961852
air_quality_gb_defra_index=5	condition_text=Blizzard	-0.001069670816478789
air_quality_gb_defra_index=5	condition_text=Clear	0.0662983922246283
air_quality_gb_defra_index=5	condition_text=Cloudy	-0.010619278123250187
air_quality_gb_defra_index=5	condition_text=Drizzling	0.0037300594231459427
air_quality_gb_defra_index=5	condition_text=Fog	-0.0028713189937041557
air_quality_gb_defra_index=5	condition_text=Heavy rain	0.0012349473787044572
air_quality_gb_defra_index=5	condition_text=Mist	0.0020675254724752097
air_quality_gb_defra_index=5	condition_text=Overcast	-0.023704018412263276
air_quality_gb_defra_index=5	condition_text=Partly cloudy	-0.04502215288847045
air_quality_gb_defra_index=5	condition_text=Rain	-0.007957083248558117
air_quality_gb_defra_index=5	condition_text=Sleet	-0.0030261036591269596
air_quality_gb_defra_index=5	condition_text=Snow	-0.009468358752643317
air_quality_gb_defra_index=5	condition_text=Sunny	-0.005969044918754591
air_quality_gb_defra_index=5	condition_text=Thundery outbreaks possible	0.0026346972802170974
air_quality_gb_defra_index=5	precip_mm	-0.008697980544865867
air_quality_gb_defra_index=5	visibility_km	-0.015045202630158091
air_quality_gb_defra_index=5	wind_kph	-0.050884047970437586
air_quality_gb_defra_index=6	air_quality_Carbon_Monoxide	0.01232398104177151

air_quality_gb_defra_index=6	air_quality_Nitrogen_dioxide	0.06921633622217754
air_quality_gb_defra_index=6	air_quality_Ozone	0.015561384151104116
air_quality_gb_defra_index=6	air_quality_PM10	0.07225220776313088
air_quality_gb_defra_index=6	air_quality_PM2.5	0.04599761471080189
air_quality_gb_defra_index=6	air_quality_Sulphur_dioxide	0.0714495877306561
air_quality_gb_defra_index=6	air_quality_gb_defra_index=1	-0.14506865193167776
air_quality_gb_defra_index=6	air_quality_gb_defra_index=10	-0.03072781784746496
air_quality_gb_defra_index=6	air_quality_gb_defra_index=2	-0.0518231738937262
air_quality_gb_defra_index=6	air_quality_gb_defra_index=3	-0.03320338544398509
air_quality_gb_defra_index=6	air_quality_gb_defra_index=4	-0.01821227547443607
air_quality_gb_defra_index=6	air_quality_gb_defra_index=5	-0.016259653839210297
air_quality_gb_defra_index=6	air_quality_gb_defra_index=6	1.0
air_quality_gb_defra_index=6	air_quality_gb_defra_index=7	-0.011193695386596697
air_quality_gb_defra_index=6	air_quality_gb_defra_index=8	-0.011371104562691622
air_quality_gb_defra_index=6	air_quality_gb_defra_index=9	-0.010262460364426969
air_quality_gb_defra_index=6	condition_text=Blizzard	-8.8643658785045E-4
air_quality_gb_defra_index=6	condition_text=Clear	0.06116572607330196
air_quality_gb_defra_index=6	condition_text=Cloudy	-0.008800199575413475
air_quality_gb_defra_index=6	condition_text=Drizzling	-0.007641698775742941
air_quality_gb_defra_index=6	condition_text=Fog	0.0011818035777294655
air_quality_gb_defra_index=6	condition_text=Heavy rain	-0.0052549368056581
air_quality_gb_defra_index=6	condition_text=Mist	0.008180297747476124
air_quality_gb_defra_index=6	condition_text=Overcast	-0.004877950879488054
air_quality_gb_defra_index=6	condition_text=Partly cloudy	-0.03928060154782483
air_quality_gb_defra_index=6	condition_text=Rain	-0.020192448479479565
air_quality_gb_defra_index=6	condition_text=Sleet	-0.002507733183661606
air_quality_gb_defra_index=6	condition_text=Snow	-0.007846432281723855
air_quality_gb_defra_index=6	condition_text=Sunny	-0.006471709614775215
air_quality_gb_defra_index=6	condition_text=Thundery outbreaks possible	-0.005397652073785876
air_quality_gb_defra_index=6	precip_mm	-0.008828319678017602
air_quality_gb_defra_index=6	visibility_km	-0.027701743514633094
air_quality_gb_defra_index=6	wind_kph	-0.03801843908870026
air_quality_gb_defra_index=7	air_quality_Carbon_Monoxide	0.01256523053379502
air_quality_gb_defra_index=7	air_quality_Nitrogen_dioxide	0.05655515663767222
air_quality_gb_defra_index=7	air_quality_Ozone	-3.406840844870809E-4
air_quality_gb_defra_index=7	air_quality_PM10	0.07780685811273326
air_quality_gb_defra_index=7	air_quality_PM2.5	0.04626626329576181
air_quality_gb_defra_index=7	air_quality_Sulphur_dioxide	0.07261860071180926
air_quality_gb_defra_index=7	air_quality_gb_defra_index=1	-0.12051420442130119
air_quality_gb_defra_index=7	air_quality_gb_defra_index=10	-0.025526800395401398
air_quality_gb_defra_index=7	air_quality_gb_defra_index=2	-0.0430515379389514
air_quality_gb_defra_index=7	air_quality_gb_defra_index=3	-0.027583351245038325
air_quality_gb_defra_index=7	air_quality_gb_defra_index=4	-0.015129649722925189
air_quality_gb_defra_index=7	air_quality_gb_defra_index=5	-0.013507530541615907
air_quality_gb_defra_index=7	air_quality_gb_defra_index=6	-0.011193695386596697
air_quality_gb_defra_index=7	air_quality_gb_defra_index=7	1.0
air_quality_gb_defra_index=7	air_quality_gb_defra_index=8	-0.009446421411633503
air_quality_gb_defra_index=7	air_quality_gb_defra_index=9	-0.008525427304629046
air_quality_gb_defra_index=7	condition_text=Blizzard	-7.363975520020837E-4
air_quality_gb_defra_index=7	condition_text=Clear	0.04752887733683366
air_quality_gb_defra_index=7	condition_text=Cloudy	-0.0073106700617795025
air_quality_gb_defra_index=7	condition_text=Drizzling	0.012274095380249478
air_quality_gb_defra_index=7	condition_text=Fog	-0.001683728470957122
air_quality_gb_defra_index=7	condition_text=Heavy rain	0.009160576260008312
air_quality_gb_defra_index=7	condition_text=Mist	0.013332975352011526
air_quality_gb_defra_index=7	condition_text=Overcast	-0.011649111758142646
air_quality_gb_defra_index=7	condition_text=Partly cloudy	-0.03767333406533579
air_quality_gb_defra_index=7	condition_text=Rain	-0.012918690137426634

air_quality_gb_defra_index=7	condition_text=Sleet	-0.0020832720612321476
air_quality_gb_defra_index=7	condition_text=Snow	-0.0065183382583779235
air_quality_gb_defra_index=7	condition_text=Sunny	-0.0018746501298169617
air_quality_gb_defra_index=7	condition_text=Thundery outbreaks possible	-0.004484040740391323
air_quality_gb_defra_index=7	precip_mm	-0.01440948073671548
air_quality_gb_defra_index=7	visibility_km	-0.024784714290299797
air_quality_gb_defra_index=7	wind_kph	-0.03255558970052788
air_quality_gb_defra_index=8	air_quality_Carbon_Monoxide	0.0182563550337343
air_quality_gb_defra_index=8	air_quality_Nitrogen_dioxide	0.06786237283921534
air_quality_gb_defra_index=8	air_quality_Ozone	0.003818376372523266
air_quality_gb_defra_index=8	air_quality_PM10	0.08615583407611443
air_quality_gb_defra_index=8	air_quality_PM2.5	0.05445373352410509
air_quality_gb_defra_index=8	air_quality_Sulphur_dioxide	0.07402009357411239
air_quality_gb_defra_index=8	air_quality_gb_defra_index=1	-0.12242423725457974
air_quality_gb_defra_index=8	air_quality_gb_defra_index=10	-0.025931375334246914
air_quality_gb_defra_index=8	air_quality_gb_defra_index=2	-0.043733862909533856
air_quality_gb_defra_index=8	air_quality_gb_defra_index=3	-0.028020520513033477
air_quality_gb_defra_index=8	air_quality_gb_defra_index=4	-0.01536943994404939
air_quality_gb_defra_index=8	air_quality_gb_defra_index=5	-0.013721611752663912
air_quality_gb_defra_index=8	air_quality_gb_defra_index=6	-0.011371104562691622
air_quality_gb_defra_index=8	air_quality_gb_defra_index=7	-0.009446421411633503
air_quality_gb_defra_index=8	air_quality_gb_defra_index=8	1.0
air_quality_gb_defra_index=8	air_quality_gb_defra_index=9	-0.008660547028878687
air_quality_gb_defra_index=8	condition_text=Blizzard	-7.480687364024953E-4
air_quality_gb_defra_index=8	condition_text=Clear	0.03956709530663763
air_quality_gb_defra_index=8	condition_text=Cloudy	-0.007426537066157263
air_quality_gb_defra_index=8	condition_text=Drizzling	-0.006448871837522738
air_quality_gb_defra_index=8	condition_text=Fog	0.004874705922736248
air_quality_gb_defra_index=8	condition_text=Heavy rain	-0.01171986607605016
air_quality_gb_defra_index=8	condition_text=Mist	-0.0019837574464308216
air_quality_gb_defra_index=8	condition_text=Overcast	-0.012112910145944228
air_quality_gb_defra_index=8	condition_text=Partly cloudy	-0.017371569437603027
air_quality_gb_defra_index=8	condition_text=Rain	-0.013787207135836347
air_quality_gb_defra_index=8	condition_text=Sleet	-0.002116289895575515
air_quality_gb_defra_index=8	condition_text=Snow	-0.006621647574916189
air_quality_gb_defra_index=8	condition_text=Sunny	-0.0024740919000177533
air_quality_gb_defra_index=8	condition_text=Thundery outbreaks possible	-0.004555108421425545
air_quality_gb_defra_index=8	precip_mm	-0.012331286304640095
air_quality_gb_defra_index=8	visibility_km	-0.010779370577006493
air_quality_gb_defra_index=8	wind_kph	-0.02985872195405907
air_quality_gb_defra_index=9	air_quality_Carbon_Monoxide	0.02914742618193628
air_quality_gb_defra_index=9	air_quality_Nitrogen_dioxide	0.06871024332218743
air_quality_gb_defra_index=9	air_quality_Ozone	-4.315668824454108E-4
air_quality_gb_defra_index=9	air_quality_PM10	0.08694057777854024
air_quality_gb_defra_index=9	air_quality_PM2.5	0.05716890348375914
air_quality_gb_defra_index=9	air_quality_Sulphur_dioxide	0.056531207905927076
air_quality_gb_defra_index=9	air_quality_gb_defra_index=1	-0.11048828858653424
air_quality_gb_defra_index=9	air_quality_gb_defra_index=10	-0.02340315402919799
air_quality_gb_defra_index=9	air_quality_gb_defra_index=2	-0.03946995933578293
air_quality_gb_defra_index=9	air_quality_gb_defra_index=3	-0.02528861462580303
air_quality_gb_defra_index=9	air_quality_gb_defra_index=4	-0.013870971582369487
air_quality_gb_defra_index=9	air_quality_gb_defra_index=5	-0.012383801060961852
air_quality_gb_defra_index=9	air_quality_gb_defra_index=6	-0.010262460364426969
air_quality_gb_defra_index=9	air_quality_gb_defra_index=7	-0.008525427304629046
air_quality_gb_defra_index=9	air_quality_gb_defra_index=8	-0.008660547028878687
air_quality_gb_defra_index=9	air_quality_gb_defra_index=9	1.0
air_quality_gb_defra_index=9	condition_text=Blizzard	-6.751345671717549E-4
air_quality_gb_defra_index=9	condition_text=Clear	0.044133500754906844

air_quality_gb_defra_index=9	condition_text=Cloudy	0.002116056773469006
air_quality_gb_defra_index=9	condition_text=Drizzling	-0.005820128665862944
air_quality_gb_defra_index=9	condition_text=Fog	-0.007875184768509118
air_quality_gb_defra_index=9	condition_text=Heavy rain	0.006329702712913576
air_quality_gb_defra_index=9	condition_text=Mist	0.009259001962969993
air_quality_gb_defra_index=9	condition_text=Overcast	-0.012388597281699528
air_quality_gb_defra_index=9	condition_text=Partly cloudy	-0.02356697328500463
air_quality_gb_defra_index=9	condition_text=Rain	-0.018718523242356972
air_quality_gb_defra_index=9	condition_text=Sleet	-0.0019099587954048438
air_quality_gb_defra_index=9	condition_text=Snow	-0.0059760593537885875
air_quality_gb_defra_index=9	condition_text=Sunny	-0.008524473313040652
air_quality_gb_defra_index=9	condition_text=Thundery outbreaks possible	-0.004111000771545266
air_quality_gb_defra_index=9	precip_mm	-0.012199286518943074
air_quality_gb_defra_index=9	visibility_km	-0.014849557779036105
air_quality_gb_defra_index=9	wind_kph	-0.01777299816102356

v. Plot charts by adding a node named Graph Explore.

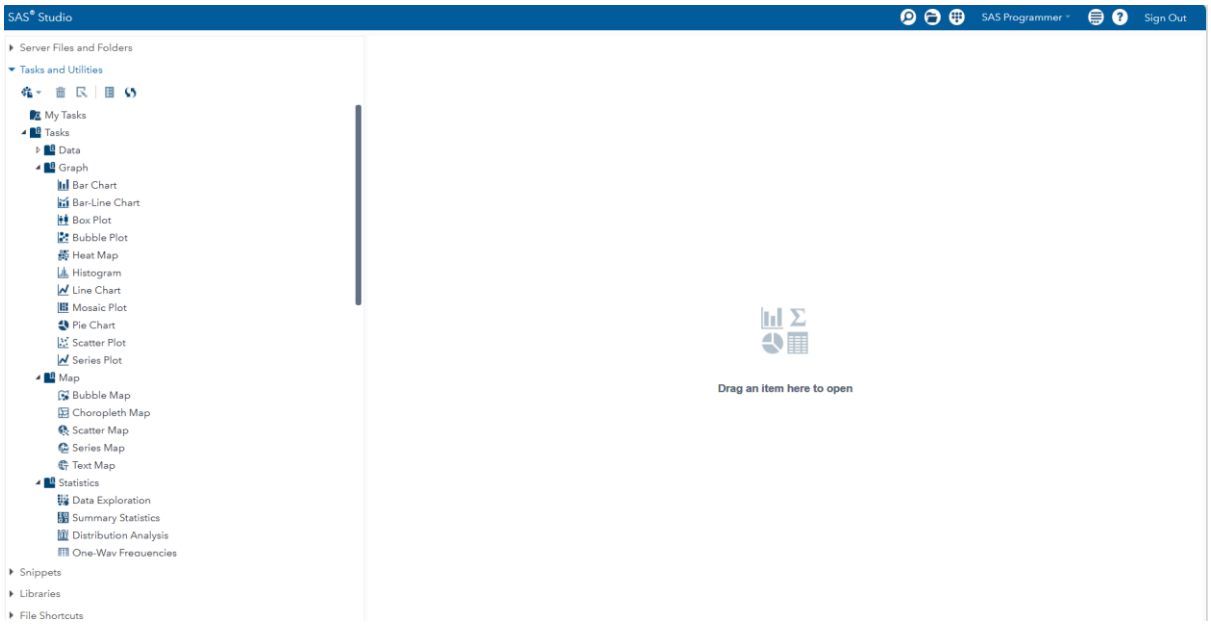


Appendix C – Plotting Charts Using SAS Studio

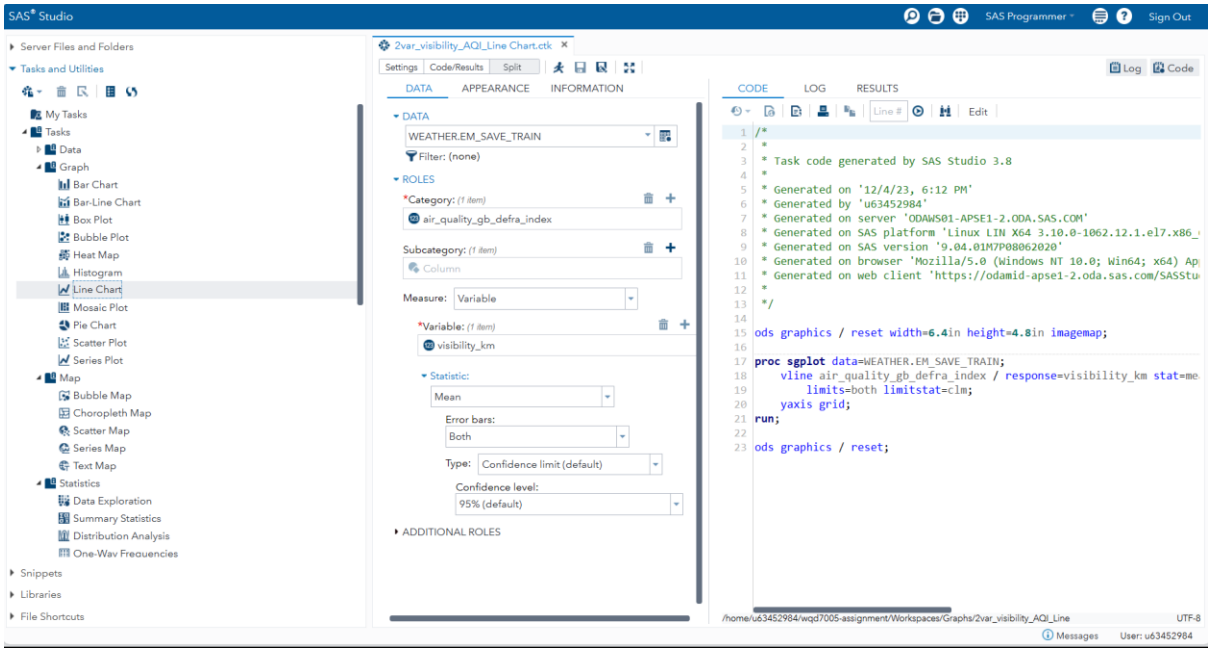
All the visualizations graphs are plotted through SAS Studio as it offers the flexibility to customize the graphs through editing the SAS code.

i. Launch the SAS Studio.

ii. Navigate to the tasks and utilities, and you can find the different types of charts at the graph section.

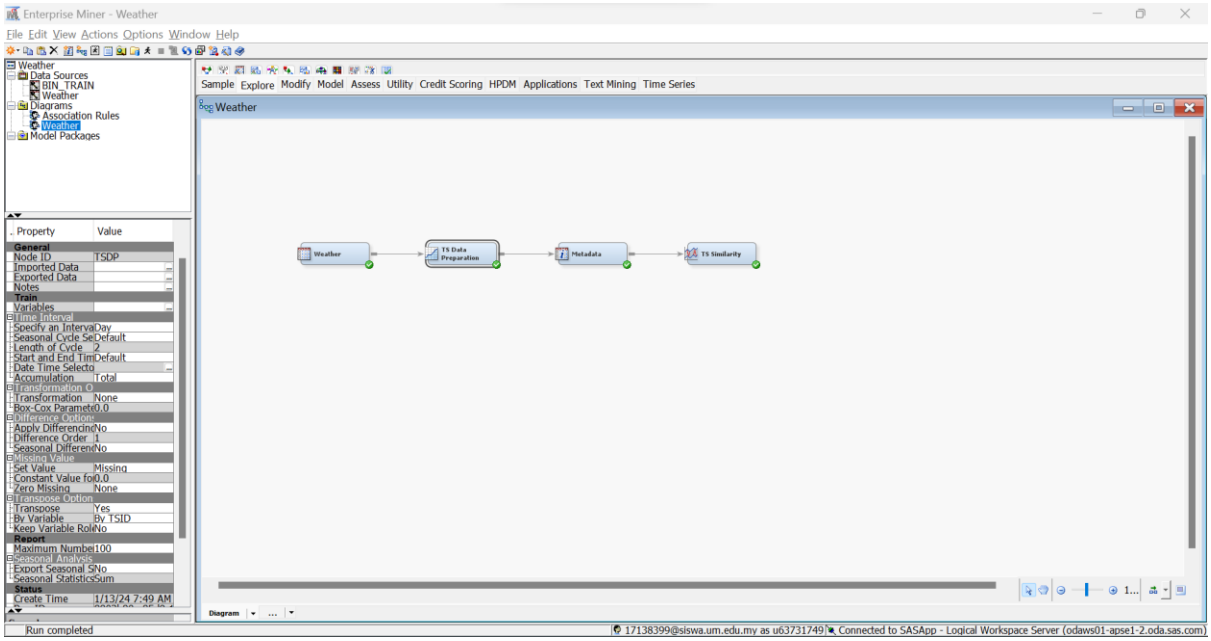


iii. Select a graph method. Then, you can start plotting by selecting the wanted variables.



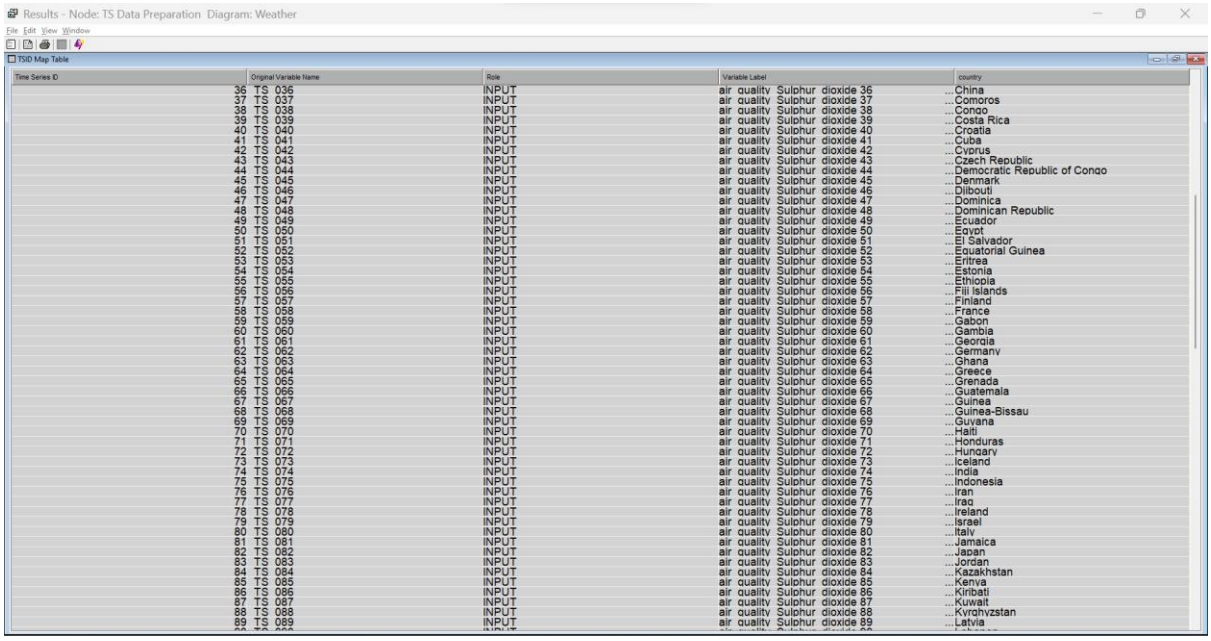
Appendix D – Time Series Similarity Analysis using SAS Enterprise Miner

(i) Connect data source to TS Data Preparation node, followed by Metadata node and TS Similarity node.

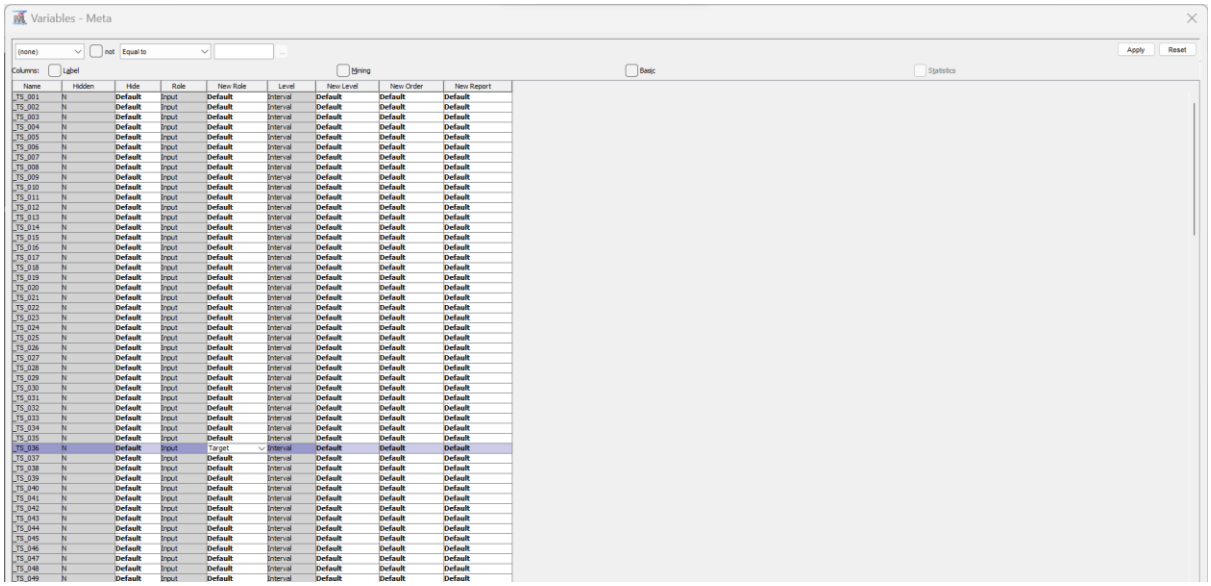


(ii) Right click on TS Data Preparation node and click Edit Variables on the menu. Specify cross ID, time ID and target.

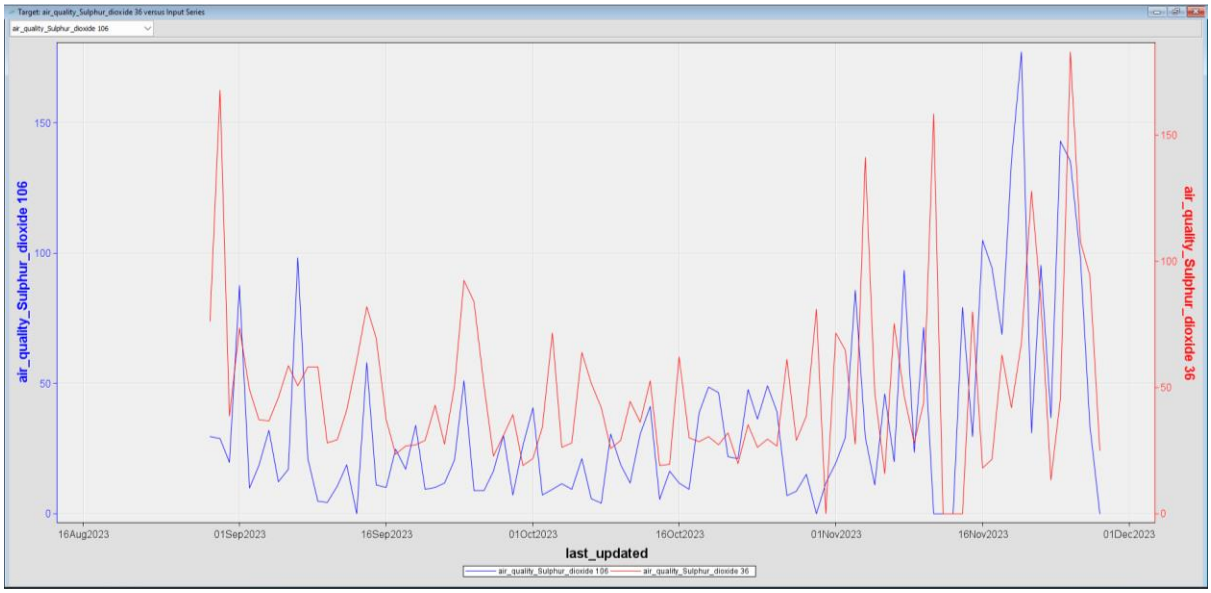
(iii) Run TS Data Preparation node. Examine the series labels.

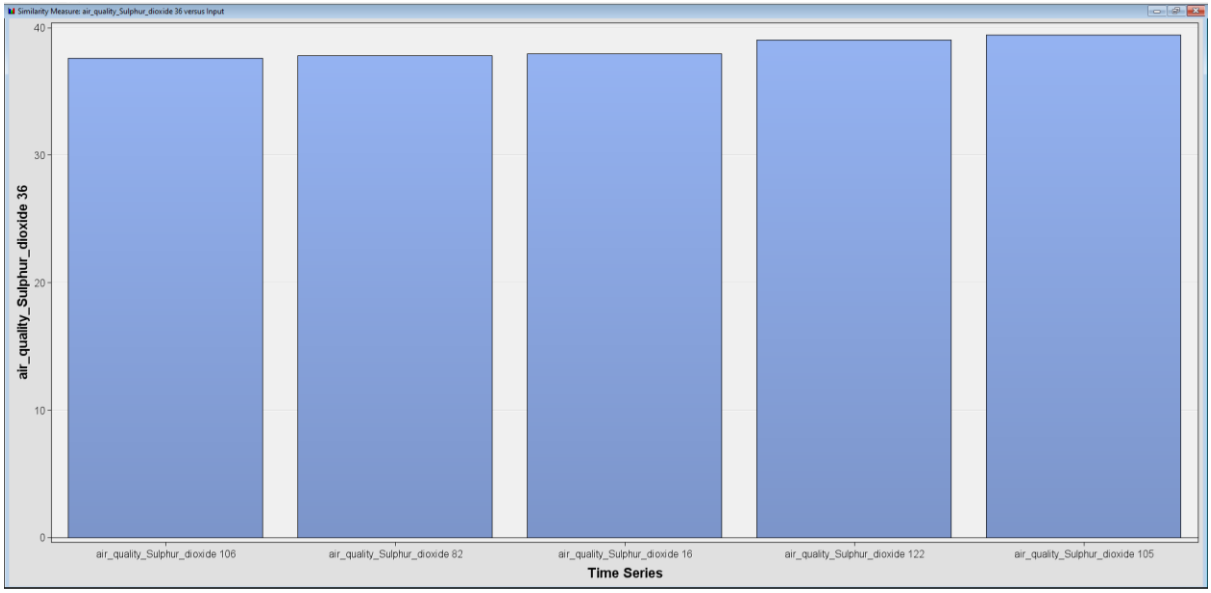


(iv) On Metadata node, specify TS 036 (China) as the target. The goal is to identify countries that exhibit similar SO_2 concentration patterns to China.



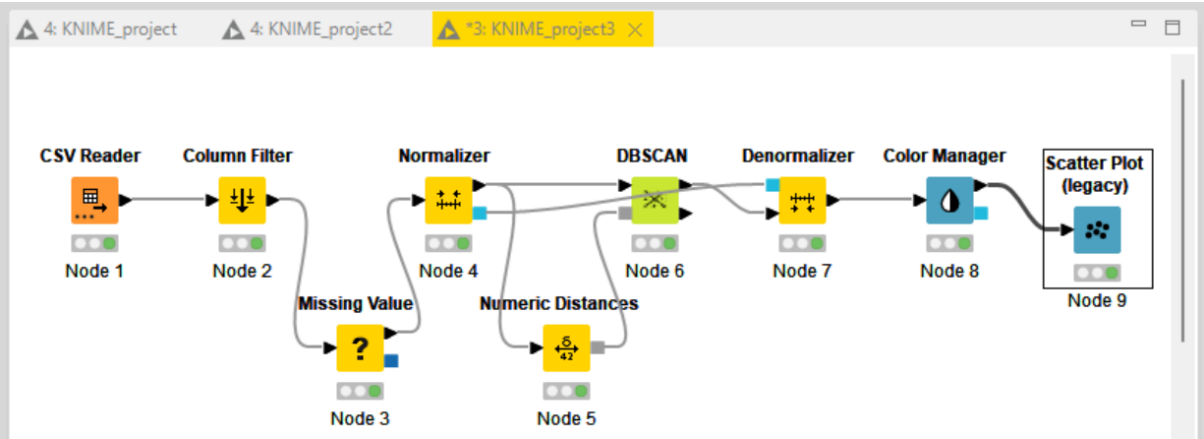
(v) On TS Similarity node, obtain the plots of target series versus input series (TS 106 in figure below), as well as a bar chart comparing the similarity measure of each input series to the target series.



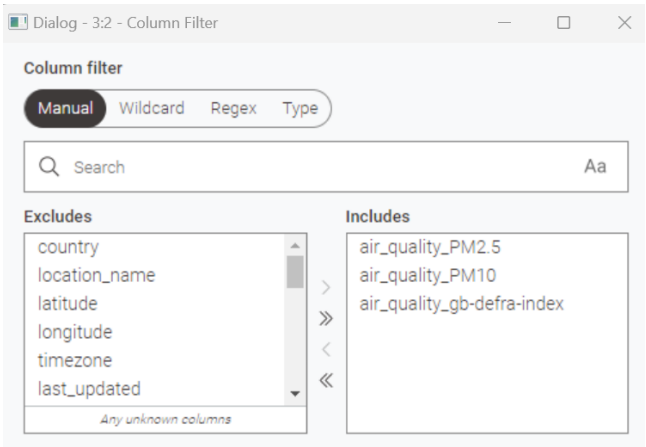


Appendix E – DBSCAN Plot using KNIME

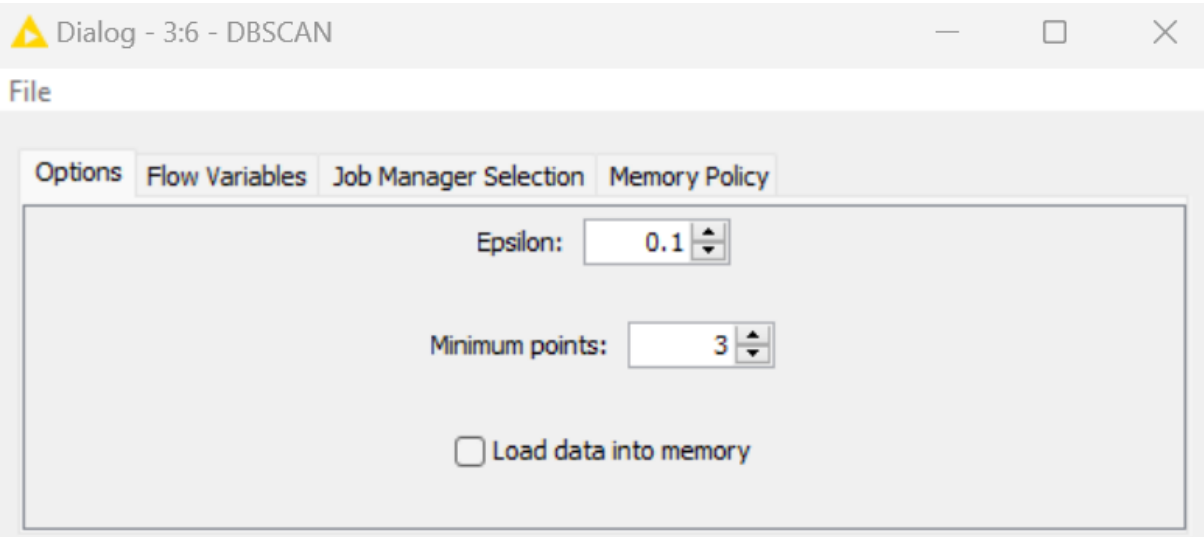
(i) Connect all the nodes.



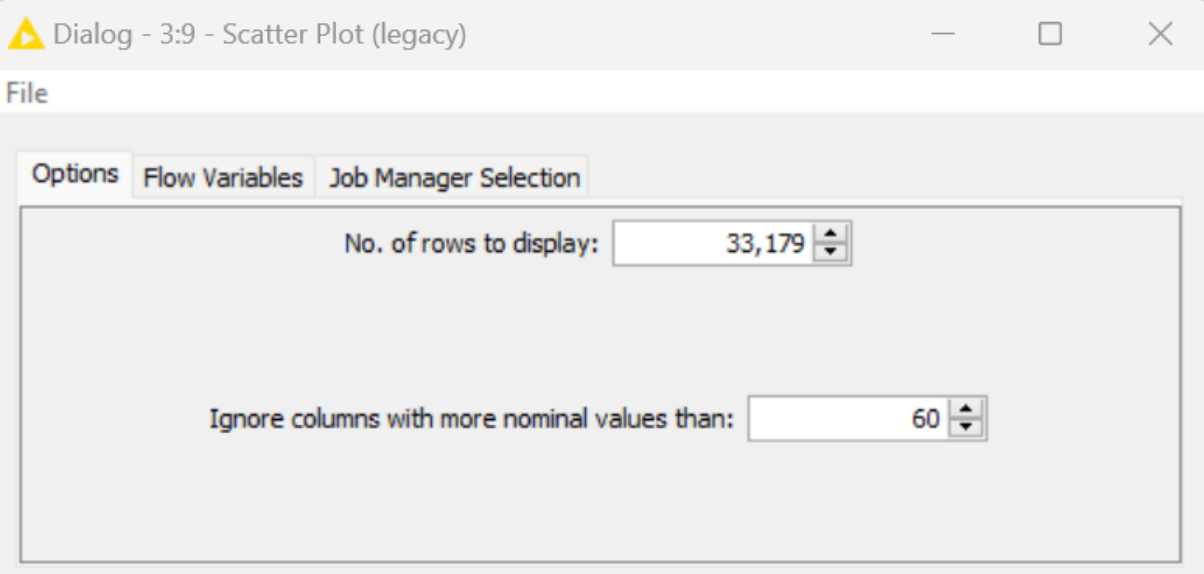
(ii) Configure the Column Filter node as below.



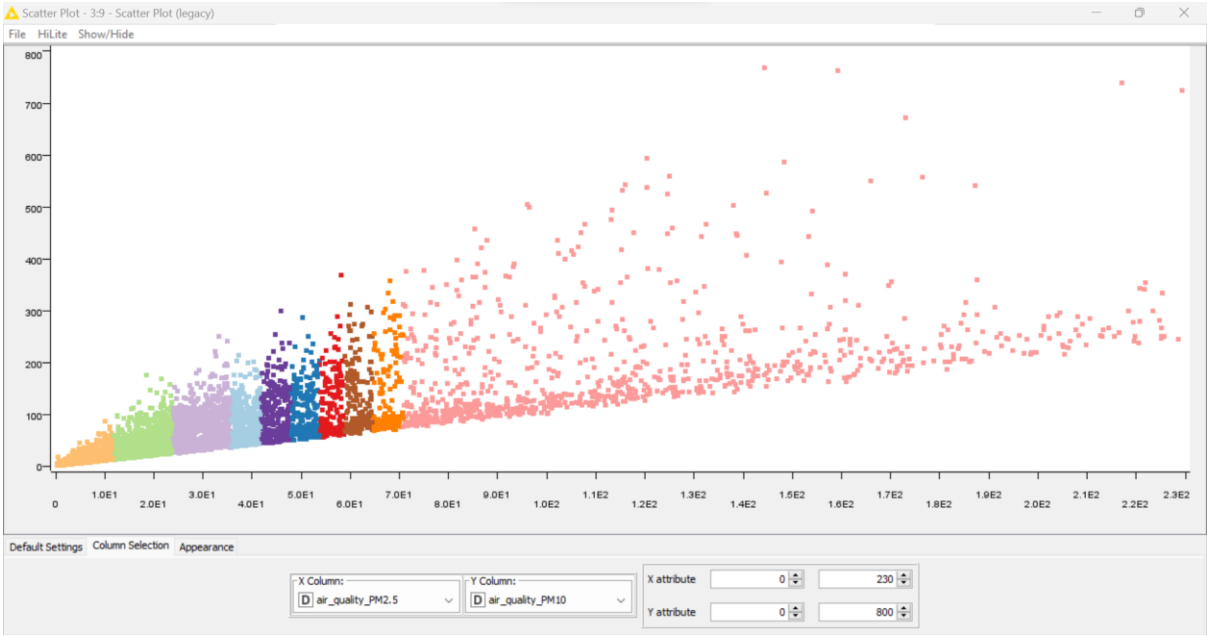
(iii) Configure the DBSCAN node as below.



(iv) Configure the Scatter Plot node as below. Include all rows (33179 rows).

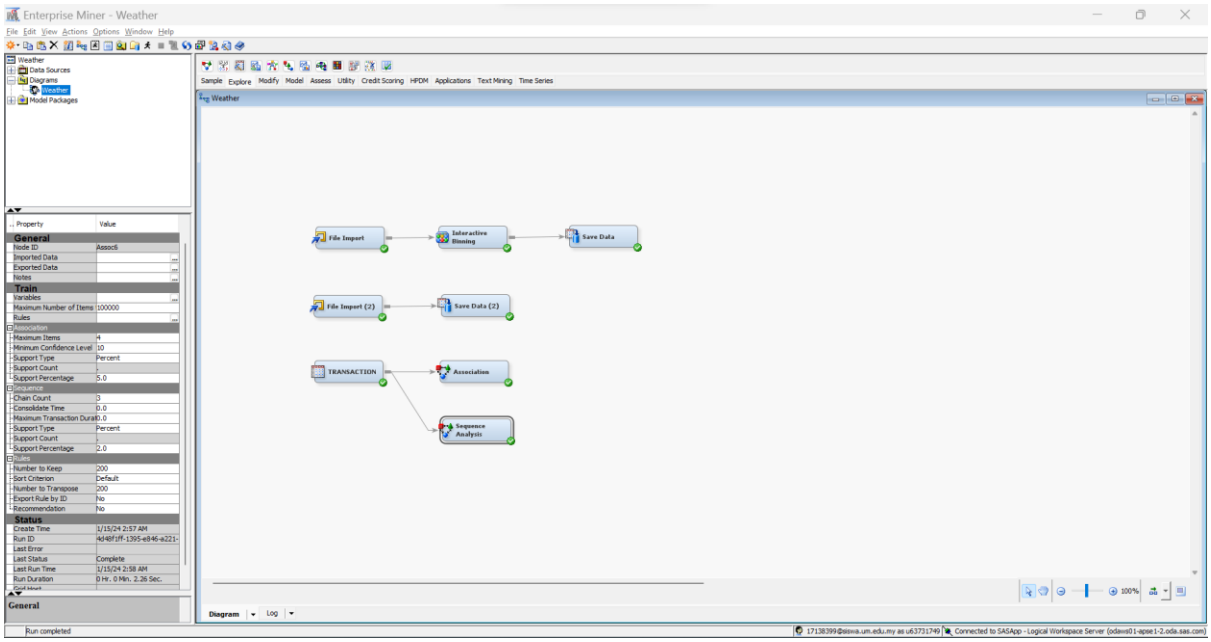


(v) View scatter plot. It displays 10 clusters.

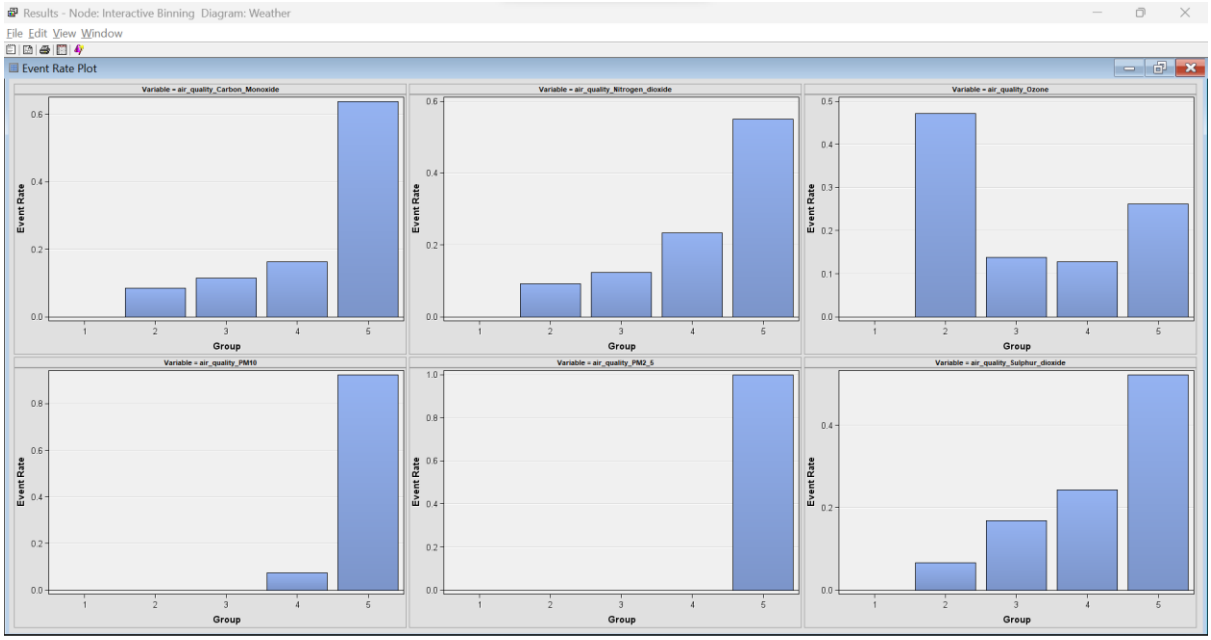


Appendix F – Association Rule and Sequence Analysis using SAS Enterprise Miner

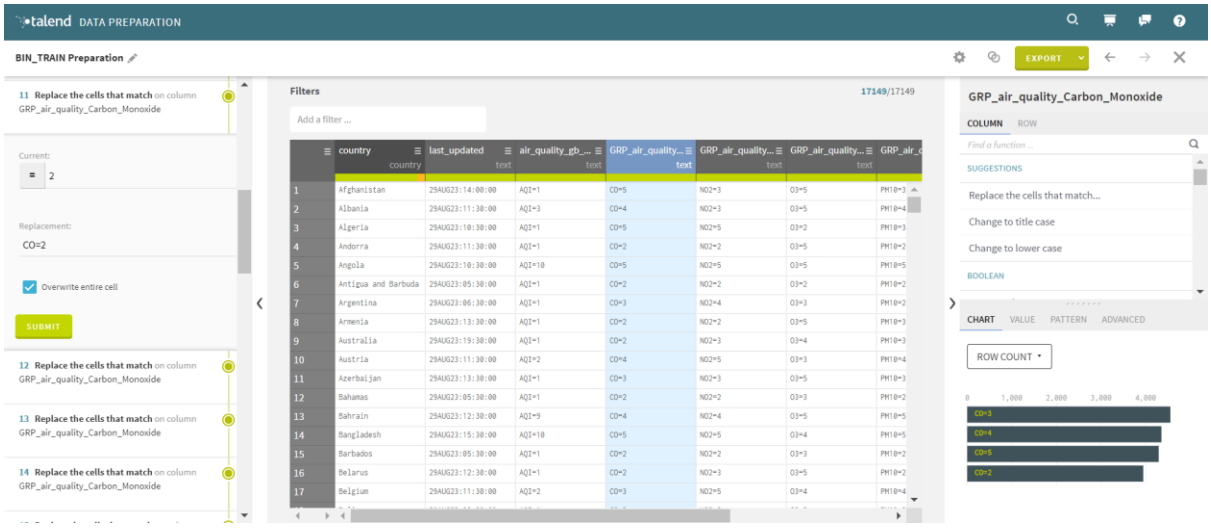
(i) Connect all the nodes as shown in the diagram.



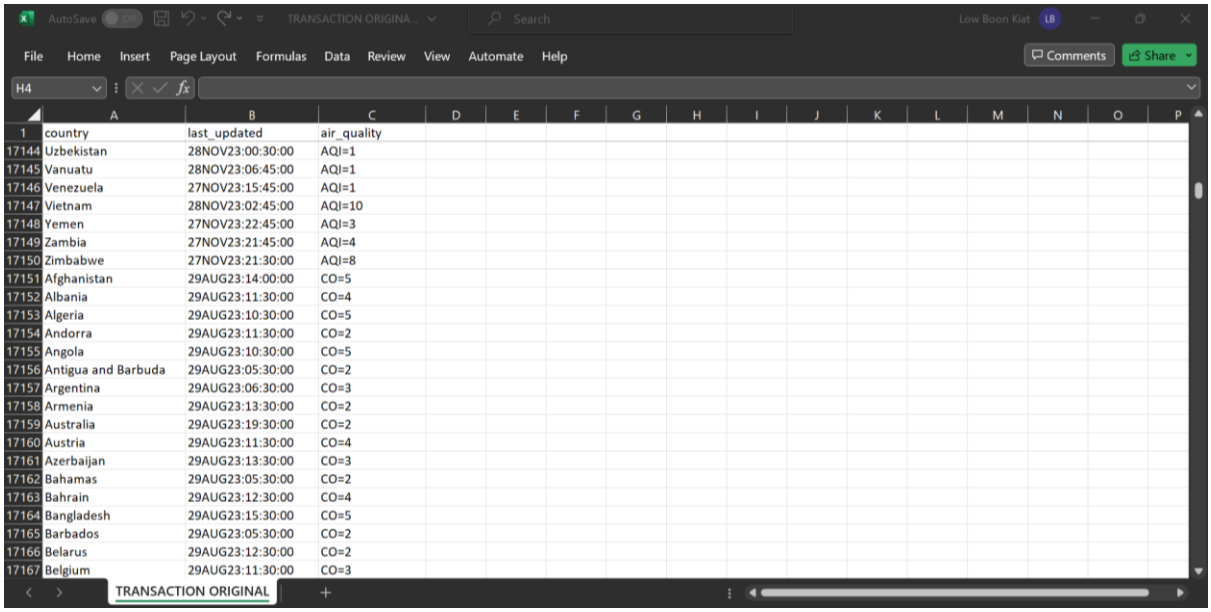
(ii) Use Interactive Binning node to discretize the concentrations of pollutants by putting them into bins.



(iii) Export the binned data from SAS Enterprise Miner as a CSV. Notice that the concentration of pollutants was grouped into "1" until "5" after binning. By using Talend Data Preparation, rename each cell content by specifying the pollutant name. For example, rename the group "2" of CO concentration into "CO=2".



(iv) The creation of transactional data involves structuring it into a set of items per transaction. By using Talend Data Preparation, merge all the columns of air pollutant concentration into single column. Eventually, the transactional dataset should consist of only 3 columns i.e. country (ID), last_updated (sequence), and air_quality (target). Export the data from Talend Data Preparation as a CSV.

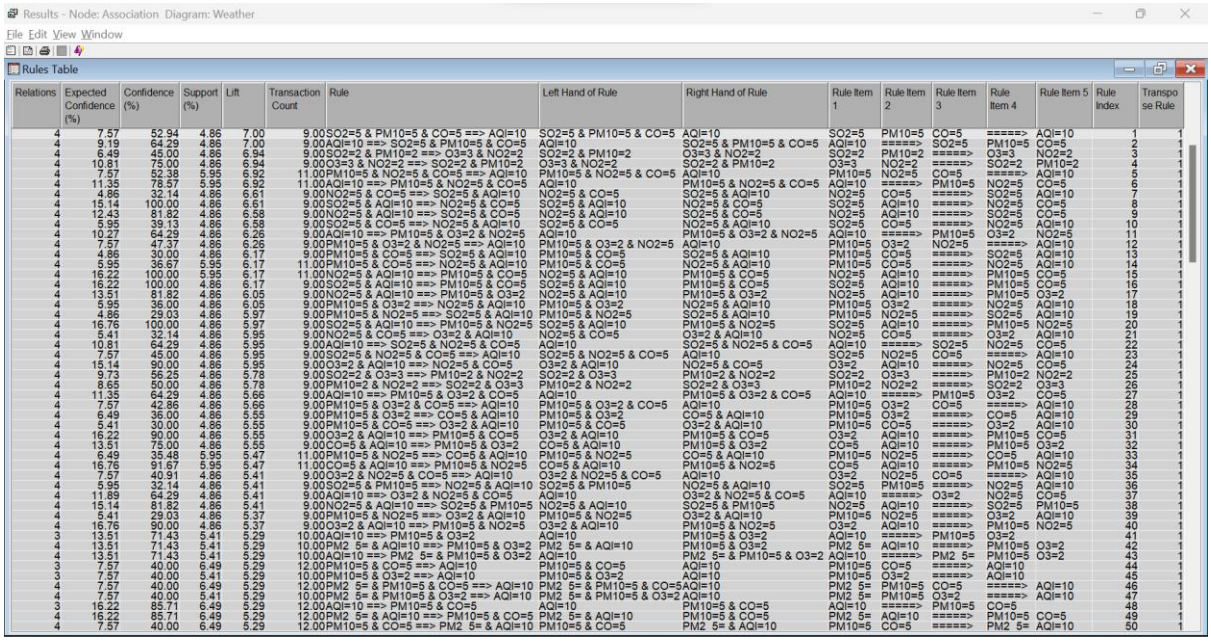


	country	last_updated	air_quality
17144	Uzbekistan	28NOV23:00:30:00	AQI=1
17145	Vanuatu	28NOV23:06:45:00	AQI=1
17146	Venezuela	27NOV23:15:45:00	AQI=1
17147	Vietnam	28NOV23:02:45:00	AQI=10
17148	Yemen	27NOV23:22:45:00	AQI=3
17149	Zambia	27NOV23:21:45:00	AQI=4
17150	Zimbabwe	27NOV23:21:30:00	AQI=8
17151	Afghanistan	29AUG23:14:00:00	CO=5
17152	Albania	29AUG23:11:30:00	CO=4
17153	Algeria	29AUG23:10:30:00	CO=5
17154	Andorra	29AUG23:11:30:00	CO=2
17155	Angola	29AUG23:10:30:00	CO=5
17156	Antigua and Barbuda	29AUG23:05:30:00	CO=2
17157	Argentina	29AUG23:06:30:00	CO=3
17158	Armenia	29AUG23:13:30:00	CO=2
17159	Australia	29AUG23:19:30:00	CO=2
17160	Austria	29AUG23:11:30:00	CO=4
17161	Azerbaijan	29AUG23:13:30:00	CO=3
17162	Bahamas	29AUG23:05:30:00	CO=2
17163	Bahrain	29AUG23:12:30:00	CO=4
17164	Bangladesh	29AUG23:15:30:00	CO=5
17165	Barbados	29AUG23:05:30:00	CO=2
17166	Belarus	29AUG23:12:30:00	CO=2
17167	Belgium	29AUG23:11:30:00	CO=3

(v) Import the transactional CSV into SAS Enterprise Miner and save it as a SAS table named Transaction. Connect Transaction data source to Association nodes. In Association Analysis, select Use -> No for Sequence variable. In Sequence Analysis, select Use -> Yes for Sequence variable.

Name	Use	Role	Level
air_quality	Yes	Target	Nominal
country	Yes	ID	Nominal
last_updated	No	Sequence	Interval

(vi) Examine the Rule Table and 3D plot containing Lift, Support and Confidence.



Relations	Expected Confidence (%)	Confidence (%)	Support (%)	Lift	Transaction Count	Rule	Left Hand of Rule	Right Hand of Rule	Rule Item 1	Rule Item 2	Rule Item 3	Rule Item 4	Rule Item 5	Rule Index	Transpose Rule
4	7.57	52.94	4.86	7.00	9.00	SO2=5 & PM10=5 & CO=5 ==> AQI=10	SO2=5 & PM10=5 & CO=5	AQI=10	SO2=5	PM10=5	CO=5	=====>	AQI=10	1	1
4	9.19	64.29	4.86	7.00	9.00	AQI=10 ==> SO2=5 & PM10=5 & CO=5	AQI=10	SO2=5 & PM10=5 & CO=5	AQI=10	=====>	SO2=5	PM10=5	CO=5	2	1
4	6.49	45.00	4.86	6.94	9.00	SO2=2 & PM10=2 ==> O3=3 & NO2=2	SO2=2 & PM10=2	O3=3 & NO2=2	SO2=2	PM10=2	=====>	O3=3	NO2=2	3	1
4	10.81	75.00	4.86	6.94	9.00	O3=3 & NO2=2 ==> SO2=2 & PM10=2	O3=3 & NO2=2	SO2=2 & PM10=2	O3=3	NO2=2	=====>	SO2=2	PM10=2	4	1
4	7.57	52.38	5.95	6.92	11.00	PM10=5 & NO2=5 & CO=5 ==> AQI=10	PM10=5 & NO2=5 & CO=5	AQI=10	PM10=5	NO2=5	CO=5	=====>	AQI=10	5	1
4	11.35	78.57	5.95	6.92	11.00	AQI=10 ==> PM10=5 & NO2=5 & CO=5	AQI=10	PM10=5 & NO2=5 & CO=5	AQI=10	=====>	PM10=5	NO2=5	CO=5	6	1
4	4.86	32.14	4.86	6.61	9.00	NO2=5 & CO=5 ==> SO2=5 & AQI=10	NO2=5 & CO=5	SO2=5 & AQI=10	NO2=5	CO=5	=====>	SO2=5	AQI=10	7	1
4	15.14	100.00	4.86	6.61	9.00	SO2=5 & AQI=10 ==> NO2=5 & CO=5	SO2=5 & AQI=10	NO2=5 & CO=5	SO2=5	AQI=10	=====>	NO2=5	CO=5	8	1
4	12.43	81.82	4.86	6.58	9.00	NO2=5 & AQI=10 ==> SO2=5 & CO=5	NO2=5 & AQI=10	SO2=5 & CO=5	NO2=5	AQI=10	=====>	SO2=5	CO=5	9	1
4	5.95	39.13	4.86	6.58	9.00	SO2=5 & CO=5 ==> NO2=5 & AQI=10	SO2=5 & CO=5	NO2=5 & AQI=10	SO2=5	CO=5	=====>	NO2=5	AQI=10	10	1
4	10.27	64.29	4.86	6.26	9.00	AQI=10 ==> PM10=5 & O3=2 & NO2=5	AQI=10	PM10=5 & O3=2 & NO2=5	AQI=10	=====>	PM10=5	O3=2	NO2=5	11	1
4	7.57	47.37	4.86	6.26	9.00	PM10=5 & O3=2 & NO2=5 ==> AQI=10	PM10=5 & O3=2 & NO2=5	AQI=10	PM10=5	O3=2	NO2=5	=====>	AQI=10	12	1
4	4.86	30.00	4.86	6.17	9.00	PM10=5 & CO=5 ==> SO2=5 & AQI=10	PM10=5 & CO=5	SO2=5 & AQI=10	PM10=5	CO=5	=====>	SO2=5	AQI=10	13	1
4	5.95	36.67	5.95	6.17	11.00	PM10=5 & CO=5 ==> NO2=5 & AQI=10	PM10=5 & CO=5	NO2=5 & AQI=10	PM10=5	CO=5	=====>	NO2=5	AQI=10	14	1
4	16.22	100.00	5.95	6.17	11.00	NO2=5 & AQI=10 ==> PM10=5 & CO=5	NO2=5 & AQI=10	PM10=5 & CO=5	NO2=5	AQI=10	=====>	PM10=5	CO=5	15	1
4	16.22	100.00	4.86	6.17	9.00	SO2=5 & AQI=10 ==> PM10=5 & CO=5	SO2=5 & AQI=10	PM10=5 & CO=5	SO2=5	AQI=10	=====>	PM10=5	CO=5	16	1
4	13.51	81.82	4.86	6.05	9.00	NO2=5 & AQI=10 ==> PM10=5 & O3=2	NO2=5 & AQI=10	PM10=5 & O3=2	NO2=5	AQI=10	=====>	PM10=5	O3=2	17	1
4	5.95	36.00	4.86	6.05	9.00	PM10=5 & O3=2 ==> NO2=5 & AQI=10	PM10=5 & O3=2	NO2=5 & AQI=10	PM10=5	O3=2	=====>	NO2=5	AQI=10	18	1
4	4.86	29.03	4.86	5.97	9.00	PM10=5 & NO2=5 ==> SO2=5 & AQI=10	PM10=5 & NO2=5	SO2=5 & AQI=10	PM10=5	NO2=5	=====>	SO2=5	AQI=10	19	1
4	16.76	100.00	4.86	5.97	9.00	SO2=5 & AQI=10 ==> PM10=5 & NO2=5	SO2=5 & AQI=10	PM10=5 & NO2=5	SO2=5	AQI=10	=====>	PM10=5	NO2=5	20	1
4	5.41	32.14	4.86	5.95	9.00	NO2=5 & CO=5 ==> O3=2 & AQI=10	NO2=5 & CO=5	O3=2 & AQI=10	NO2=5	CO=5	=====>	O3=2	AQI=10	21	1
4	10.81	64.29	4.86	5.95	9.00	AQI=10 ==> SO2=5 & NO2=5 & CO=5	AQI=10	SO2=5 & NO2=5 & CO=5	AQI=10	=====>	SO2=5	NO2=5	CO=5	22	1
4	7.57	45.00	4.86	5.95	9.00	SO2=5 & NO2=5 & CO=5 ==> AQI=10	SO2=5 & NO2=5 & CO=5	AQI=10	SO2=5	NO2=5	CO=5	=====>	AQI=10	23	1
4	15.14	90.00	4.86	5.95	9.00	O3=2 & AQI=10 ==> NO2=5 & CO=5	O3=2 & AQI=10	NO2=5 & CO=5	O3=2	AQI=10	=====>	NO2=5	CO=5	24	1
4	9.73	56.25	4.86	5.78	9.00	SO2=2 & O3=3 ==> PM10=2 & NO2=2	SO2=2 & O3=3	PM10=2 & NO2=2	SO2=2	O3=3	=====>	PM10=2	NO2=2	25	1
4	8.65	50.00	4.86	5.78	9.00	PM10=2 & NO2=2 ==> SO2=2 & O3=3	PM10=2 & NO2=2	SO2=2 & O3=3	PM10=2	NO2=2	=====>	SO2=2	O3=3	26	1
4	11.35	64.29	4.86	5.66	9.00	AQI=10 ==> PM10=5 & O3=2 & CO=5	AQI=10	PM10=5 & O3=2 & CO=5	AQI=10	=====>	PM10=5	O3=2	CO=5	27	1
4	7.57	42.86	4.86	5.66	9.00	PM10=5 & O3=2 & CO=5 ==> AQI=10	PM10=5 & O3=2 & CO=5	AQI=10	PM10=5	O3=2	CO=5	=====>	AQI=10	28	1
4	6.49	36.00	4.86	5.55	9.00	PM10=5 & O3=2 ==> CO=5 & AQI=10	PM10=5 & O3=2	CO=5 & AQI=10	PM10=5	O3=2	=====>	CO=5	AQI=10	29	1
4	5.41	30.00	4.86	5.55	9.00	PM10=5 & CO=5 ==> O3=2 & AQI=10	PM10=5 & CO=5	O3=2 & AQI=10	PM10=5	CO=5	=====>	O3=2	AQI=10	30	1
4	16.22	90.00	4.86	5.55	9.00	O3=2 & AQI=10 ==> PM10=5 & CO=5	O3=2 & AQI=10	PM10=5 & CO=5	O3=2	AQI=10	=====>	PM10=5	CO=5	31	1
4	13.51	75.00	4.86	5.55	9.00	CO=5 & AQI=10 ==> PM10=5 & O3=2	CO=5 & AQI=10	PM10=5 & O3=2	CO=5	AQI=10	=====>	PM10=5	O3=2	32	1
4	6.49	35.48	5.95	5.47	11.00	PM10=5 & NO2=5 ==> CO=5 & AQI=10	PM10=5 & NO2=5	CO=5 & AQI=10	PM10=5	NO2=5	=====>	CO=5	AQI=10	33	1
4	16.76	91.67	5.95	5.47	11.00	CO=5 & AQI=10 ==> PM10=5 & NO2=5	CO=5 & AQI=10	PM10=5 & NO2=5	CO=5	AQI=10	=====>	PM10=5	NO2=5	34	1
4	7.57	40.91	4.86	5.41	9.00	O3=2 & NO2=5 & CO=5 ==> AQI=10	O3=2 & NO2=5 & CO=5	AQI=10	O3=2	NO2=5	CO=5	=====>	AQI=10	35	1
4	5.95	32.14	4.86	5.41	9.00	SO2=5 & PM10=5 ==> NO2=5 & AQI=10	SO2=5 & PM10=5	NO2=5 & AQI=10	SO2=5	PM10=5	=====>	NO2=5	AQI=10	36	1
4	11.89	64.29	4.86	5.41	9.00	AQI=10 ==> O3=2 & NO2=5 & CO=5	AQI=10	O3=2 & NO2=5 & CO=5	AQI=10	=====>	O3=2	NO2=5	CO=5	37	1
4	15.14	81.82	4.86	5.41	9.00	NO2=5 & AQI=10 ==> SO2=5 & PM10=5	NO2=5 & AQI=10	SO2=5 & PM10=5	NO2=5	AQI=10	=====>	SO2=5	PM10=5	38	1
4	5.41	29.03	4.86	5.37	9.00	PM10=5 & NO2=5 ==> O3=2 & AQI=10	PM10=5 & NO2=5	O3=2 & AQI=10	PM10=5	NO2=5	=====>	O3=2	AQI=10	39	1
4	16.76	90.00	4.86	5.37	9.00	O3=2 & AQI=10 ==> PM10=5 & NO2=5	O3=2 & AQI=10	PM10=5 & NO2=5	O3=2	AQI=10	=====>	PM10=5	NO2=5	40	1
3	13.51	71.43	5.41	5.29	10.00	AQI=10 ==> PM10=5 & O3=2	AQI=10	PM10=5 & O3=2	AQI=10	=====>	PM10=5	O3=2		41	1
4	13.51	71.43	5.41	5.29	10.00	PM2_5= & AQI=10 ==> PM10=5 & O3=2	PM2_5= & AQI=10	PM10=5 & O3=2	PM2_5=	AQI=10	=====>	PM10=5	O3=2	42	1
4	13.51	71.43	5.41	5.29	10.00	AQI=10 ==> PM2_5= & PM10=5 & O3=2	AQI=10	PM2_5= & PM10=5 & O3=2	AQI=10	=====>	PM2_5=	PM10=5	O3=2	43	1
3	7.57	40.00	6.49	5.29	12.00	PM10=5 & CO=5 ==> AQI=10	PM10=5 & CO=5	AQI=10	PM10=5	CO=5	=====>	AQI=10		44	1
3	7.57	40.00	5.41	5.29	10.00	PM10=5 & O3=2 ==> AQI=10	PM10=5 & O3=2	AQI=10	PM10=5	O3=2	=====>	AQI=10		45	1
4	7.57	40.00	6.49	5.29	12.00	PM2_5= & PM10=5 & CO=5 ==> AQI=10	PM2_5= & PM10=5 & CO=5	AQI=10	PM2_5=	PM10=5	CO=5	=====>	AQI=10	46	1
4	7.57	40.00	5.41	5.29	10.00	PM2_5= & PM10=5 & O3=2 ==> AQI=10	PM2_5= & PM10=5 & O3=2	AQI=10	PM2_5=	PM10=5	O3=2	=====>	AQI=10	47	1
3	16.22	85.71	6.49	5.29	12.00	AQI=10 ==> PM10=5 & CO=5	AQI=10	PM10=5 & CO=5	AQI=10	=====>	PM10=5	CO=5		48	1
4	16.22	85.71	6.49	5.29	12.00	PM2_5= & AQI=10 ==> PM10=5 & CO=5	PM2_5= & AQI=10	PM10=5 & CO=5	PM2_5=	AQI=10	=====>	PM10=5	CO=5	49	1
4	7.57	40.00	6.49	5.29	12.00	PM10=5 & CO=5 ==> PM2_5= & AQI=10	PM10=5 & CO=5	PM2_5= & AQI=10	PM10=5	CO=5	=====>	PM2_5=	AQI=10	50	1