**Data Mining for PM2.5 of Chengdu**

**Name: Lam Kwan Hon**

**Student ID: 6247872**

**GitHub:** [**https://github.com/hon611/iteration4-assignment**](https://github.com/hon611/iteration4-assignment)

1. **Situational Understanding**

In China, there is a serious air pollution issue. This report is about one of the China places which are the Cheng Du air problem. PM 2.5 relates to atmospheric particulate matter (PM) having a diameter of less than 2.5 micrometers, about 3 percent of a human hair's diameter.

Commonly written as PM2.5, electrons are so tiny in this category that only an electron microscope can detect them. The fine particles are so tiny and light, they tend to remain in the air longer than heavier particles. The likelihood of being inhaled into the bodies by humans and animals will be increased. Particles less than 2.5 micro meters can bypass the nose and neck and penetrate deep into the lungs due to their minute size, and some may even enter the circulatory system.

**1.1 Identify the objectives of the situation**

PM 2.5 did not take account of the Chinese air quality standard. After 2011, the information was disclosed. PM 2.5 became an air quality factor in China from that moment on. PM 2.5 air pollution has always been one of the country's worst content, in China (Cheng Du). The local government of China has launched a vast amount of funds to develop Cheng Du. So, air pollution in Cheng Du has been a serious issue in China.

Because of the development of Cheng Du, China government is not easy to get the balance between healthy air and development. In this PM 2.5 research, which helps the state and individuals prepare in advance. It can also demonstrate the trend of PM2.5 in Cheng Du, it provides efficient assistance for future governance, according to the information.

**1.2 Assessing the Situation**

**Resources**: Air pollution is a worldwide issue, and lately it is more severe. The reasons or variables that could affect the problem can be discovered after several types of research and data analysis. Besides, there are five cities, and a city will be used as a study for saving time.

**Requirements**: No deployment requirements are available now. Furthermore, the requirements would be further discussed if it is implemented.

**Assumption**: The PM 2.5 concentration and humility connections and the humility factor can be used to forecast PM 2.5 concentration.

**Data**: Chengdu’s PM 2.5 data is selected for this study, and environmental data of this type are more reliable.

**Limits**: Since this dataset has been downloaded from Kaggle, which offers free access to customers and the data needed to download, but there is some missing value in the dataset, this is the limitation.

**Risk**: Because of citizens ' livelihoods, factory emission, the relative government policies, and other reasons, certain reasons cannot be analysed for modifying PM 2.5.

**Contingency**: We could browse several relative website information in an attempt to collect suitable information or evidence to resolve the issue.

**1.3 Determine data mining of objectives**

The organization has been able to efficiently evaluate the information and to detect the variables affecting PM2.5 concentration through information analysts.

• The research utilizes historical PM2,5 and environment information and logs for several years to predict the connection of PM2,5 concentration to environmental variables.

• This information is to detect the motion towards the concentration of PM2.5 in Chengdu.

* 1. **Produce a project plan**

The overview of the planning analysis report for the study is as shown in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| Phase | Time | Resources | Risks |
| Situation  understanding | 1 week | Website, books, news | May be changed in topic |
| Data understanding |  | Technique issues |
| Data preparation | 1 week | SPSS modeler, database analyst | Data problems, wrong critical thinking |
| Modeling | SPSS modeler, plot the graph | Technology issues |
| Evaluation | Evaluate the graph, make the prediction. | Plotting the wrong graph using the wrong way to analysis |
| Interpretation | 1 week | Data mining | If the value is wrong, we should change and make the right interpretation. |

**2 Data Understanding**

**2.1 Collect initial data**

Data Type: Kaggle Data These data have been uploaded to Kaggle's website, and "PM2.5 Five Chinese City Data— Measurement for Shenyang, Chengdu, Beijing, Gangzhou and Shanghái."

The dataset covers five Chinese cities. then I would like to choose Chengdu as my research subject because Chengdu has a large number of citizens and it is a modern city. As a result. I extracted it to be my dataset. Therefore, in two places in Chengdu. I removed two PM 2.5 concentration columns and retained the PM 2.5 concentration provided by the US embassy to make information more accurate and unprejudiced.

The historical information of PM 2.5 (district-shared) are included in the dataset as we need to predict the trend of PM 2.5 concentration over time in the city of Chengdu. The missing values are shown in the N / A dataset.

• Chengdu Historical PM information: Historical data include the previous daily hourly concentration of PM in the Chengdu area. Since it is important to research the trend of PM 2.5 concentration in the city over time, historical PM 2.5 information should be included. Since the most recent data from the 2015-2019 PM 2.5 concentration in China are not available, I include the historical review data from 2010-2015.

• City database: the data includes the air quality of the main city in China with PM 2.5 concentration. These data provide insights into the air quality of the Chinese city and help to reduce the target of this study.

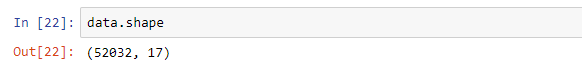
• Climate data: Climate changes over time. These data are intended to assist in assessing the impact of climate change on PM 2.5 concentration.

However, it should be noted that there is a potential problem with the missing N / A information. This can influence the reliability of the findings of the study.

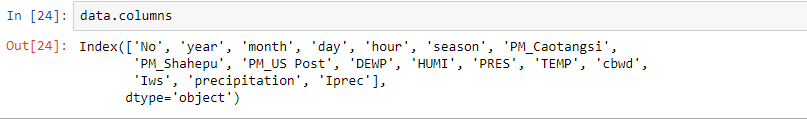
2.2 **Describing Data**

* Amount of Data

In the PM 2.5 Chengdu data, there are 52032 data rows and 17 columns. In the column sets, the information is as follow:



By using data. columns, we can see there are 17 attributes.



No: row number

year: year of data in this row

month: month of data in this row

day: day of data in this row

hour: the hour of data in this row

season: season of data in this row

PM: PM2.5 concentration (ug/m^3)

DEWP: Dew Point (Celsius Degree)

TEMP: Temperature (Celsius Degree)

HUMI: Humidity (%)

PRES: Pressure (hPa)

cbwd: Combined wind direction

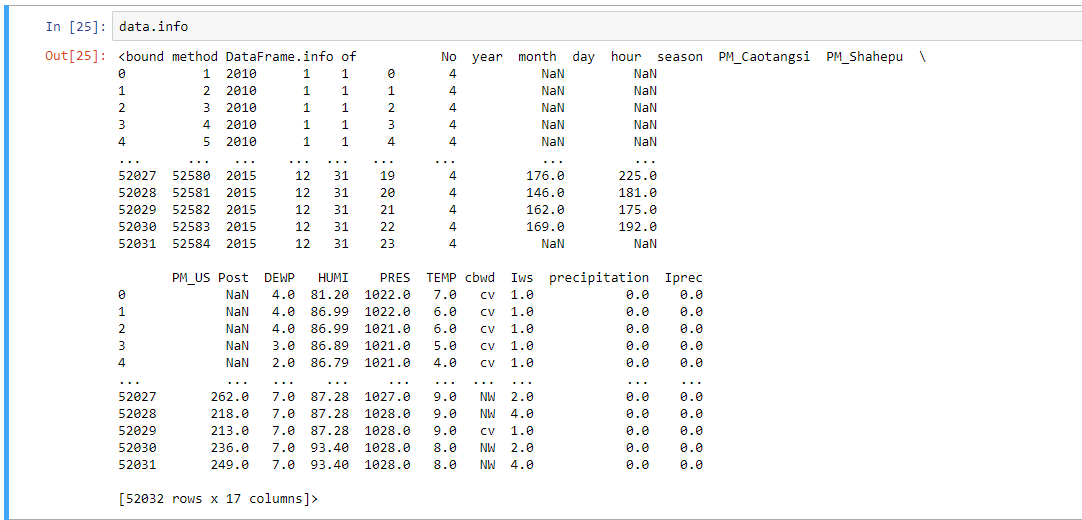
Iws: Cumulated wind speed (m/s)

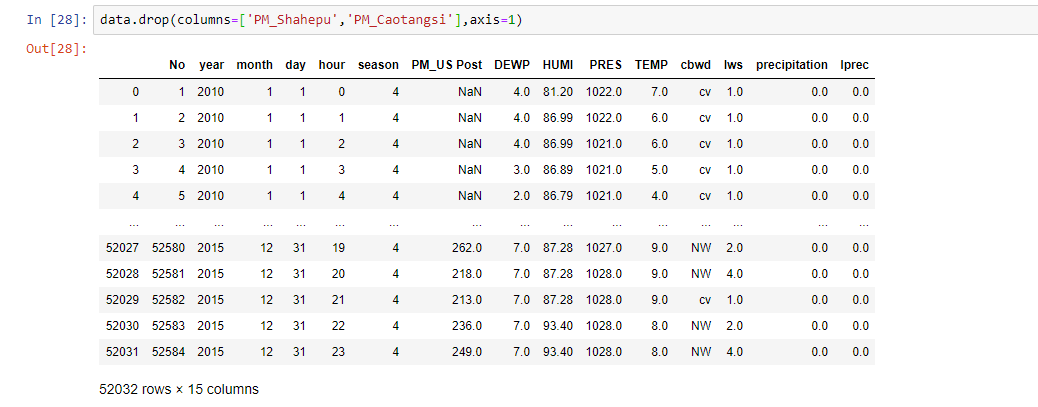
precipitation: hourly precipitation (mm)

Iprec: Cumulated precipitation (mm)

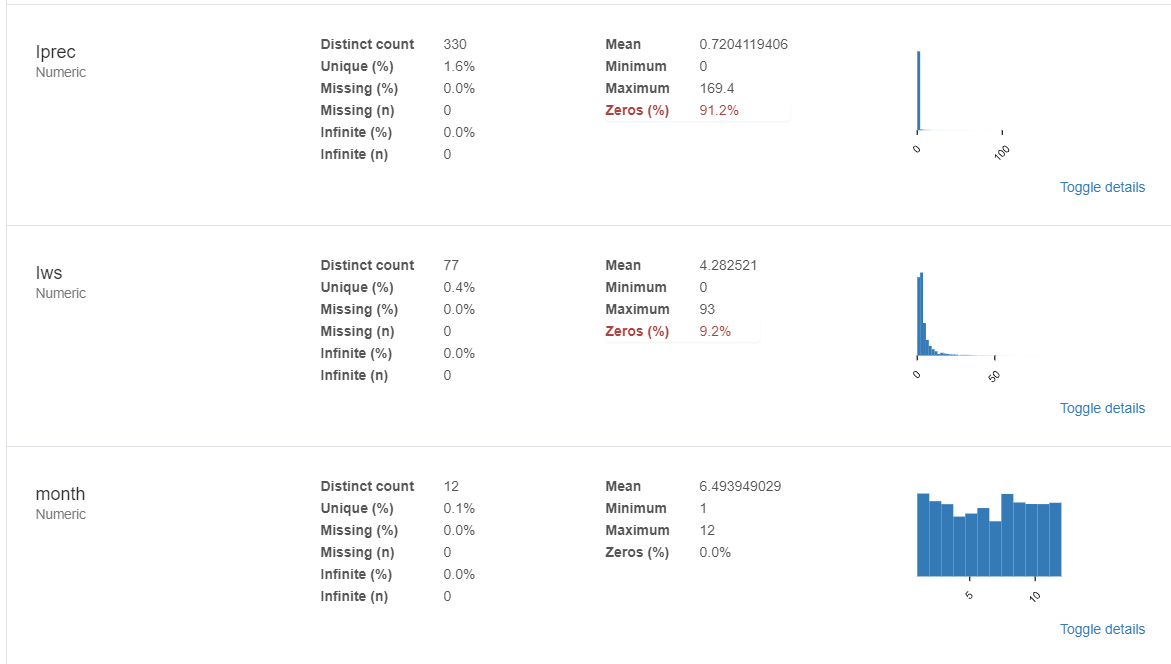
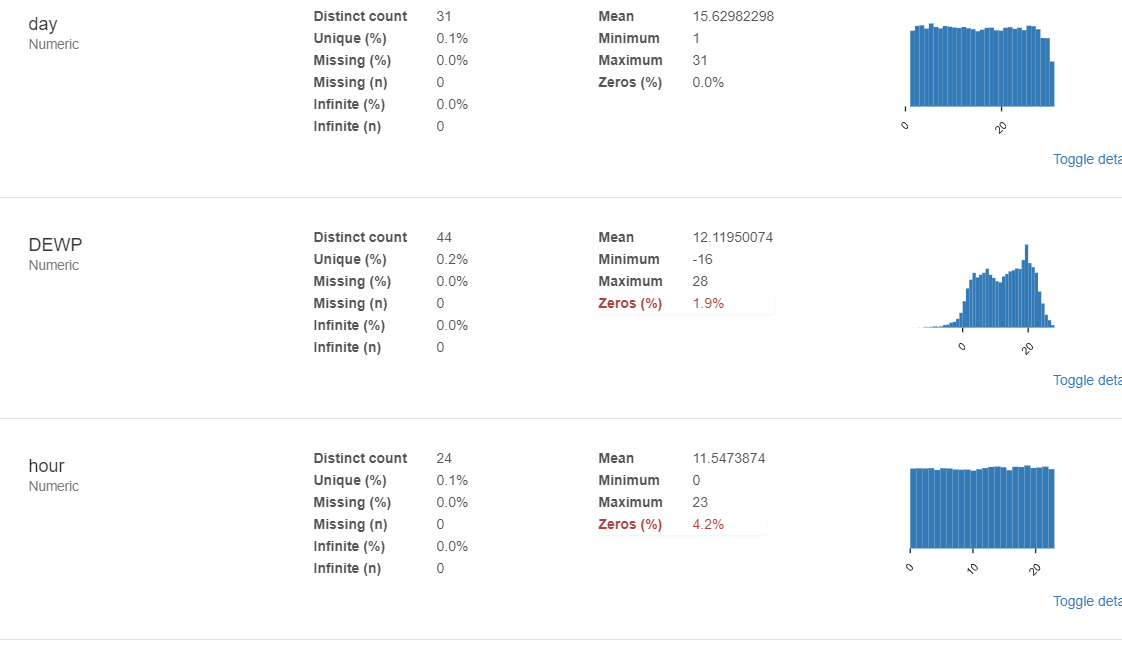
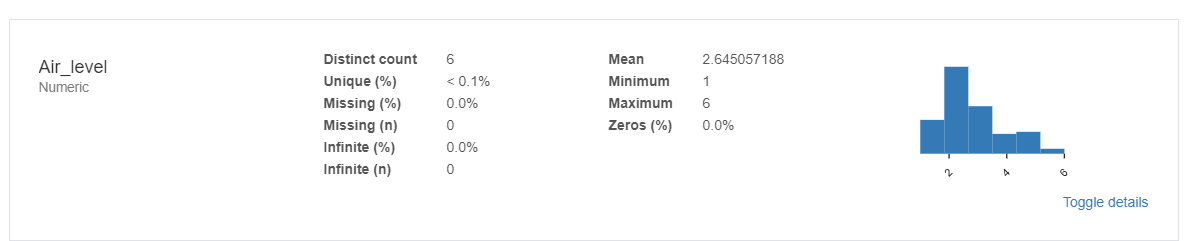
**2.3.1 Exploring the data**

Because the PM 2.5 concentration standard in China is a little different from other countries. So, I will filter the PM\_Caotangsi and China PM\_Shahepu and leave the American standard PM\_US Post into my research. I will set the PM\_US Post to target in the variable factor. I need to drop the PM\_Caotangsi and China PM\_Shahepu by using data.drop(columns). After that, I can get them from 17 columns to 15 columns data.

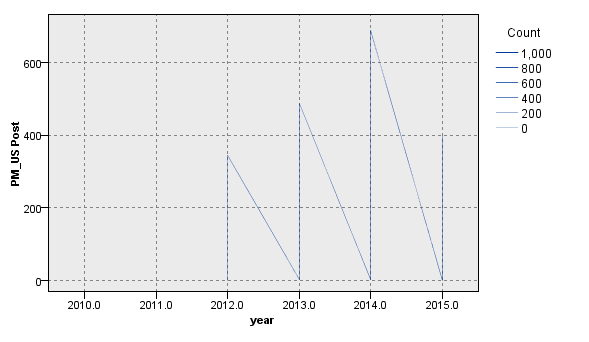




**2.3 Visualization**

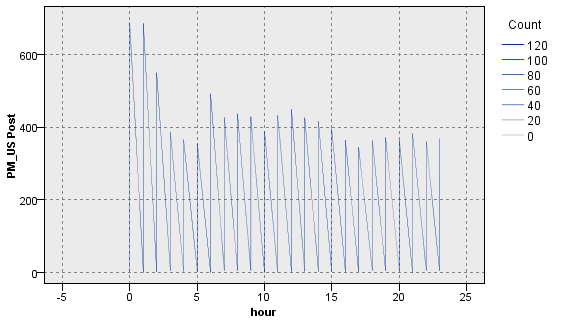
****By using pandas profile, we can get the overview report of variables graph.

**Year compares with PM2.5**

****

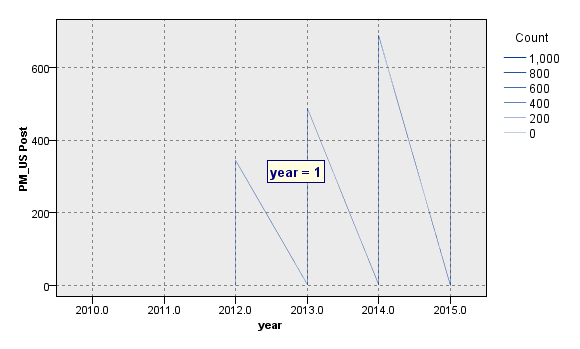
The chart indicates the trend of PM 2.5 concentration combined with the year. There was a significant decrease in 2014. In 2013, PM 2.5 concentration drew the Chinese government’s attention, and the government decided to take strategy to reduce the PM 2.5 concentration. The figure for 2014 shows that the Chinese government achieved a successful result in PM 2.5 governance in 2014.

**PM2.5 with Hour**

****

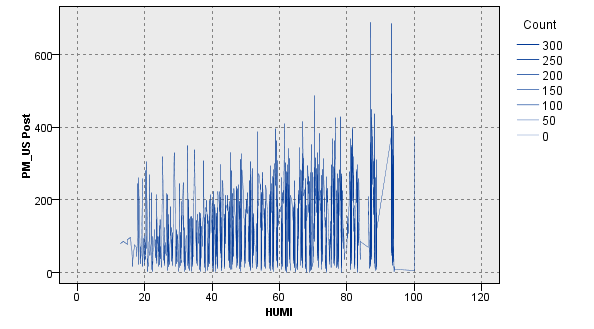
The graph shows the trend of PM 2.5 in combination with hours per day. Increases in the concentration of PM 2.5 in the morning peak (12:00 am), meaning that the diesel emissions may play a significant role in increasing the concentration of PM 2.5. At increased altitude, the temperature reduces, the bottom air is warmer, the top air is cooler and the cold air drops strongly. Querol, Xavier, et al (2001) stated that the warm water floats and forms a convection. At some stage, however, the temperature of the reduced atmosphere close to the floor is very small because the soil temperature decreases significantly, and the air of the upper side is not so rapidly cooled off.

**PM 2.5 with year**

****

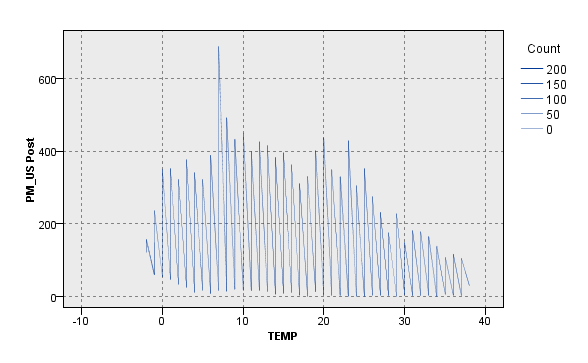
The graph shows the trend of PM 2.5 in combination with the year. In 2015, there was a substantial reduction. As the above graph, the air pollution was increasing from 2011 to 2014 and 2014 of the air pollution issue was serious. The concentration of PM 2.5 attracted China's attention in 2014, and the government decided to adopt a strategy to reduce the concentration of PM 2.5. The 2014 figure demonstrates that the Chinese local government accomplished a good outcome in PM 2.5 in 2015.

**PM2.5 with HUMI**

****

The graph shows the trend in concentration of PM 2.5 in combination with humility. With the rise in humidity, the reduced concentration rate of PM 2.5 would have more points. This shows that the greater the humidity of the atmosphere, the reduced the PM2.5 concentration, the smaller the humidity of the atmosphere, the greater the PM2.5 concentration and the greater the PM2.5 concentration. (Cheng et al.,2015) mentioned that in some Chengdu seasons, for instance in winter and early spring, crops were not found covering the surface of the soil, and a powerful wind could lead to more dust particles in the atmosphere. In dry seasons such as fall, the growing interaction of green space and other components of urban regions with greater EDs and LPIs would lead to quicker near-earth wind speeds and increased atmospheric moisture.

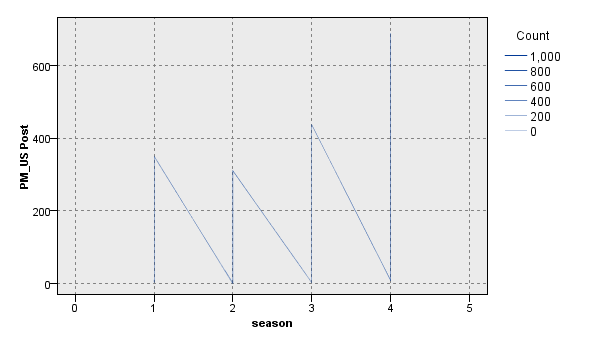
**PM 2.5 with TEMP**

****

The graph shows the PM 2.5 trend combined with temperature. When the temperature is low, the concentration PM 2.5 is always higher. It might be less if the temperature were higher, so the temperature may affect the concentration of PM 2.5. Around 7 to 9 Celsius provides the highest PM concentration. Li, Y., Ma, Z., Zheng, C., & Shang, Y. (2015) said that the cold temperatures and stagnant air are able to produce an accumulation of these substances close to the floor, especially when the weather is called temperature reversals.

Warm air lies close the floor in other seasons and circumstances, and the air can readily rise and carry pollutants. In a temperature reversal, cold air is caught by a layer of warm air close the floor. The warm air functions like a deck that holds these substances down. Smoke cannot increase during the reversal of the temperature and carbon monoxide can achieve unhealthy concentrations.

**Season & PM2.5**

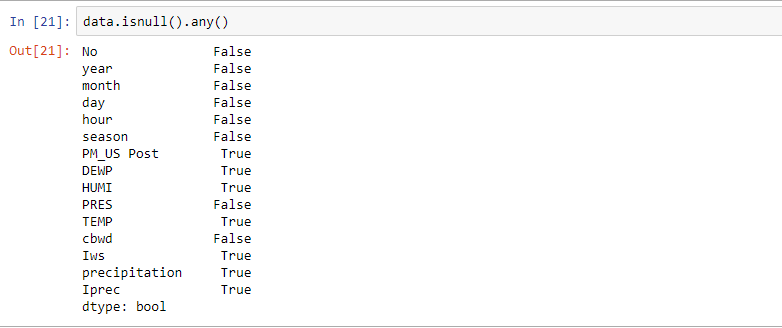
****

The graph shows the trend of PM 2.5 in combination with various seasons. The largest concentration of PM 2.5 is in winter, followed by autumn and spring. The lowest is the summer season. Furthermore, the humidity and the temperature would be greater and the concentration of PM 2.5 would reduce.

The reason for the adverse correlation may be that elevated temperatures encourage air convection and thus dilute and disperse air pollutants (Luo et al. 2017). Because of the temperature of the winter is low the low temperature can result in enhanced national heating emission levels and electricity manufacturing.

## **2.4 Verify the data quality**

By using the data.info(), we can get the data of type. There are eight variables which are a float. Six Integers variable and one object.



**Missing data:** PM US Post has 17340missing documents. Seasonal variables DEWP(Dew Point (Celsius Degree), "HUMI(Humidity (Percent), PRES(HPa), TEMP(Temperature (Celsiuus Degree)), precipitation(Hourly Rainfall (mm))" This failure may occur due to technical problems in the data collection system that could disintegrate the detector or human errors in data processing.

**Measurement errors:** Because of the deviations have appeared, also, there is one observation point as well. There are measurement errors.

1. **Data preparation**

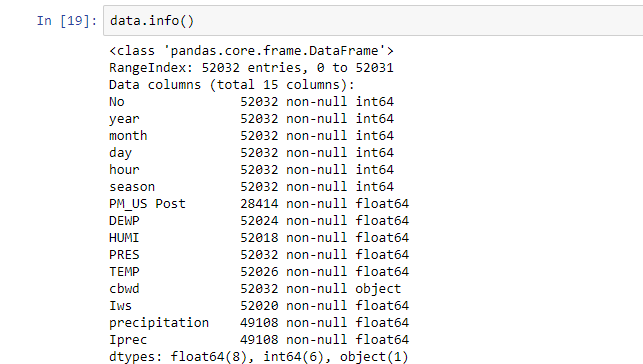
**3. 1 Select the data**

Data of PM2.5 in Chengdu are available from the Kaggle website. The PM2.5 concentration in Chengdu, timeline and relative weather conditions are taken into consideration in this dataset. Furthermore, we should, therefore, analyse the PM 2.5 concentration in this research so that the PM US post is set to the target.

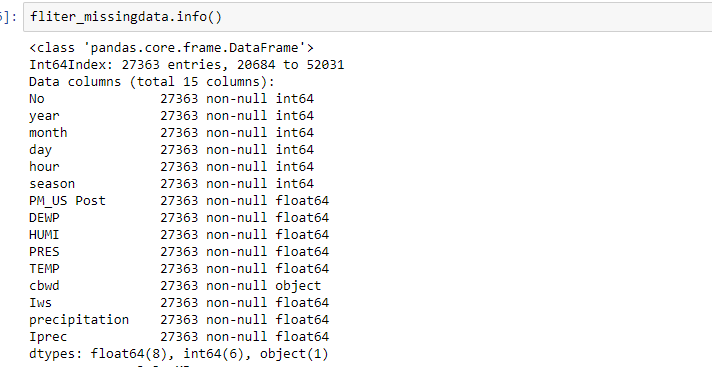
**3.2 Clean the data**

****

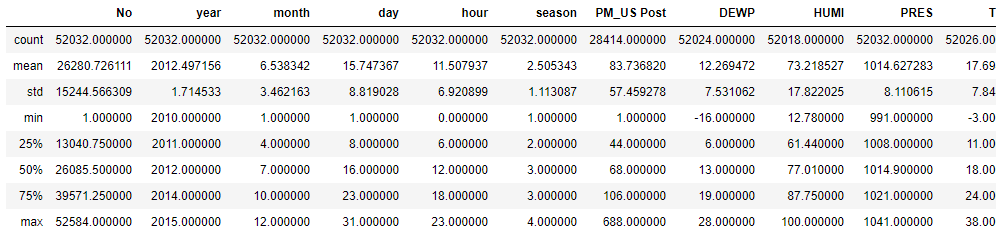
By using data.dropna() to delete these and output fill\_Missing to a CSV file.

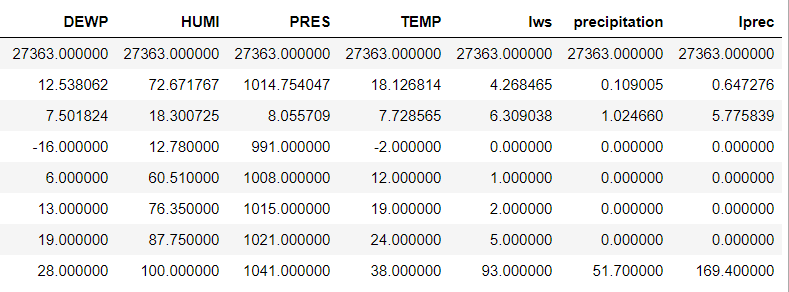


After, by using fliter\_missingdata.info(), I can exactly filter the missing data. There is 27363 missing values in the PM\_US Post.



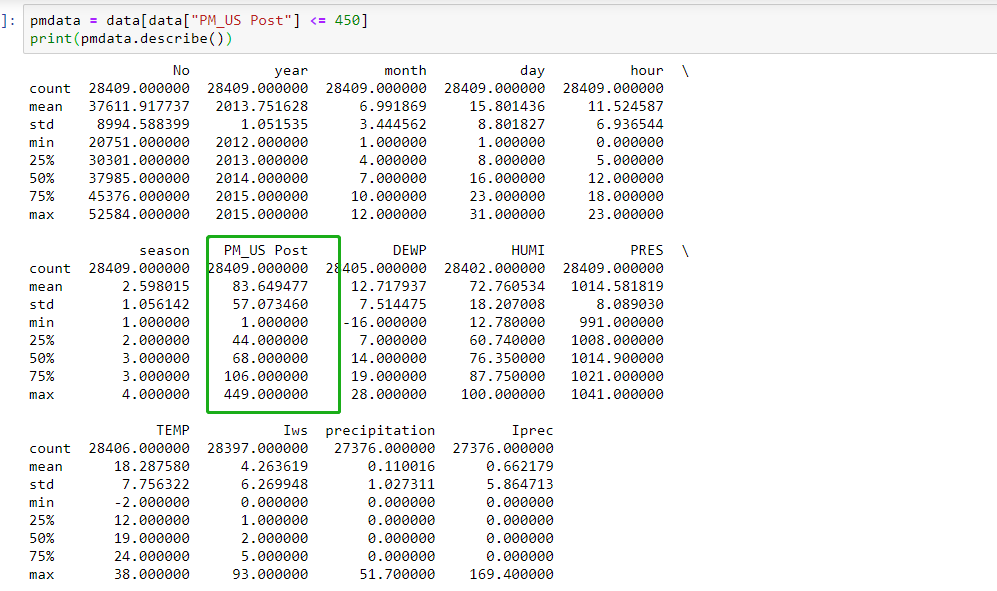
After considering the outliers and extremes. We would use data.describe() function.



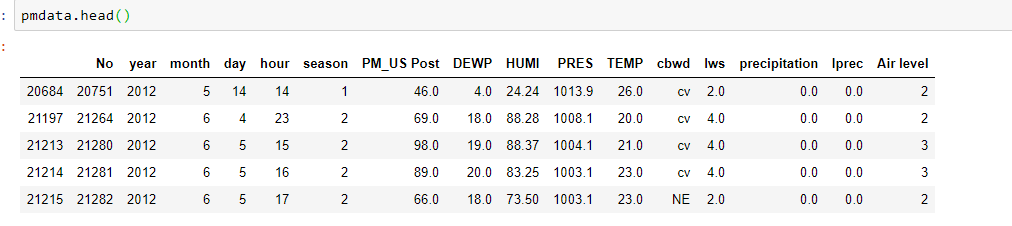


Through the Tukey statistic method, we can compute the quartile Q1 which is 25% and quartile Q3 which is 75% for the interquartile range is Q3–Q1, which will then be lower than Q1–1.5IQR or information volumes larger than Q3 + 1.5 IQR. For instance, for the column PM\_US Post:

Q1 is 44, Q3 is 106 in PM\_US Post, the IQR is 62. So, the interquartile range inside in -49 to 199. The minimum value is 1 and the maximum value is 688. (Schwartz, J., Laden, F., & Zanobetti, A. (2002) stated that the worst level of PM 2.5 is around 450 to 550. I will set the worst level which is 450. All of the data from PM\_US Post which is over 450 should be filtered.



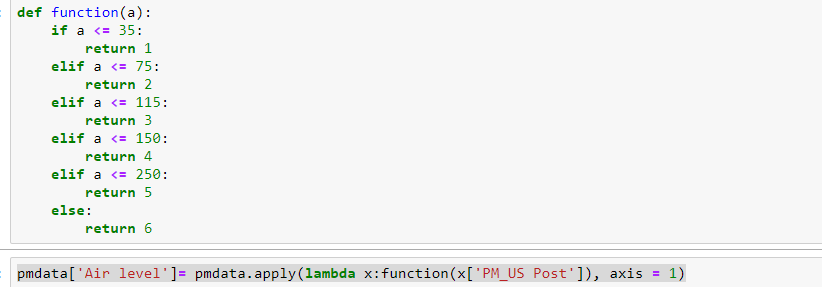
As we can see, there are no significant outliers for all the attributes. On the temporary, it can say that the data set is accurate.

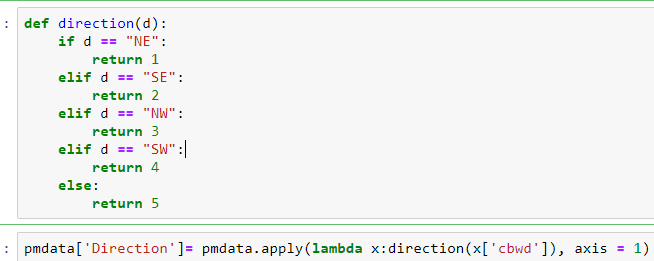


By using pmdata.head(), we can see that cbwd which is object type. The prime thing I need to do is changing the object type to Integer.

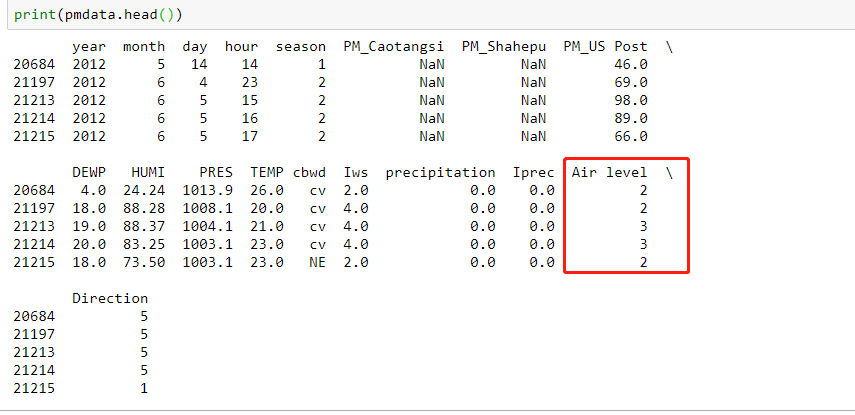
**3.3 Data Construction**

To predict the PM 2.5 concentration, the six quality levels should be built, the new variable name is "Air level".



****

The one Object type ‘cdwd'should be changed to Integer. Five attributes, 'NE', 'SE', 'NW', 'SW' are built using if..else . Then, we would use the same method to transfer these into numbers.



## 

## **3.4 Integrate various data sources**

By using the pmdata.drop to drop down ‘cbwd’ and ‘PM\_US Post’, because the attributes from right now are irrelevant.

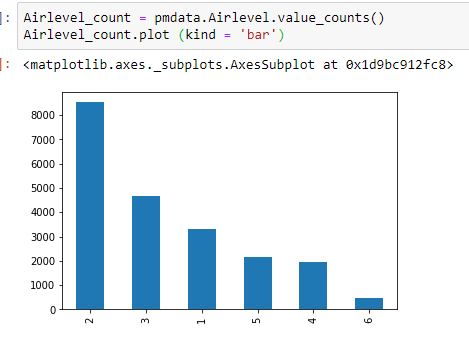




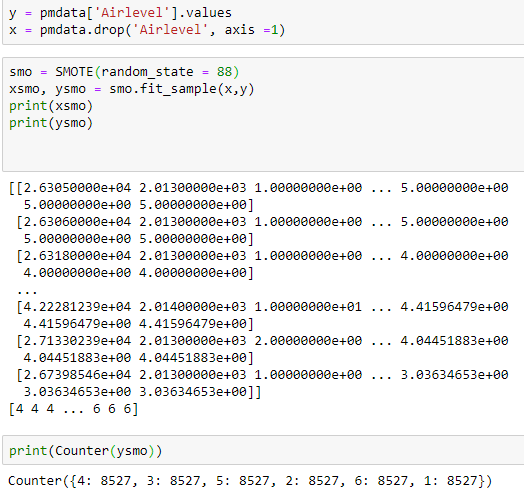
## 

## **3.5 Format the data**

After using the Counter() and plotting the boxplot, it shows that the data is imbalanced. From the picture below, we can comprehend that the allocation of data for 'Airlevel ' is unbalanced. SMOTE Algorithm is an excellent option for balancing information. Where the data is severely imbalanced, the anticipated results are often partial, i.e. the classification results are oriented towards more observations. How can I deal with this sort of problem? The easiest and hardest way is to create data to either cut off a part of the more (for example, the sub-sample) or sample a bootstrap (for example, over-sampling). However, there is an issue with that. For the first method, the data cut off will lead to the loss of some hidden information. The second method produces the model with a simple copy of the returned samples. Fitting. Fitting. Fitting. Fitting.



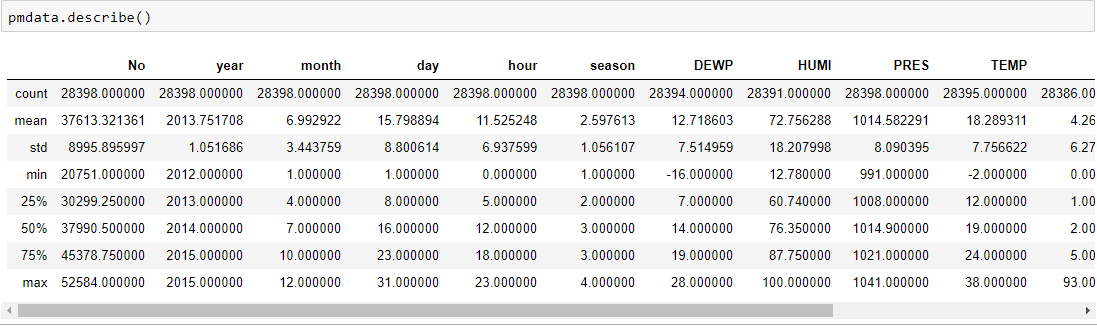
After using SMOTE, the result is balanced right now.

****

**4.0 Data Transformation**

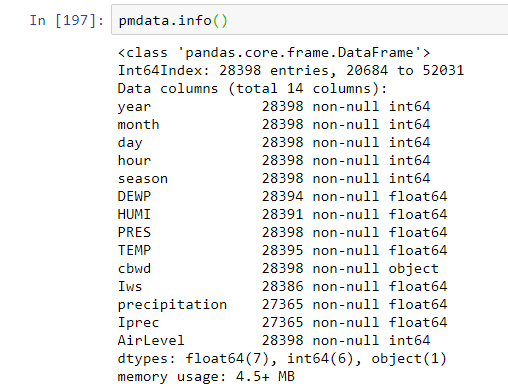
Some information transformations are necessary to make the evaluation easy and logical for Pm 2.5 concentration data set. This study enables both the decrease of data and projection of information.

**4.1 Data reduced**

****

When the “No" factor is shown on the graph, it is the meaningless factor in this study. We should delete it.

****



By using the drop(columns), the not-related variable “No” are deleted. Clearly, contrary to other rows, the ' No ' row would not impact the PM rates at all. Therefore, we will delete this column to will the variables we will examine in the future.

**4.2 Project the data**

I don't need to project the response data in this study because it's a classification issue. Because I use decision trees or random forest algorithms, there is no need to transform the parameters as they are not statistically based.

## **Project the data**

After the elimination of the useless data, Decision tree, Random Forest and XGBoost tree which will be used to manage that data.

**5.0 Data-mining methods selection**

**5. 1 Match and discuss the objectives of data mining to data mining methods**

This research focuses on data mining objectives: 1. Explanation: Essential characteristics related to Chengdu pm 2.5 concentration. Explain the relationship between features and variable response.

2. Prediction: In Chengdu, forecast the pm 2.5 intensity, despite detailed weather conditions.

Different approaches will be used in this analysis. About the following strategies.

• **Supervised Learning-Regression** is the monitoring learning task to model the constant numerical variables and predict them. Types include real-estate price predictions, shifting stock levels, and test scores.

Regression tasks are defined as numerical target variable by marked datasets. For each observation you can use to monitor your algorithm, you have a "ground truth" value.

• **Classification**: Classification Models use one or more input field values to estimate one or more outputs or target areas. The following approaches include: option trees (C&R tree, QUEST, algorithms CHAID andC5.0), return (linear, spatial, generalised linear, and Cox regression algorithms).

Classification models help organisations to predict a known outcome, for example whether a customer buys or leaves or if a transaction is a recognised fraud pattern. Modelling techniques include machine learning, rule induction, subgroup identification, statistical methodology and different models generation.

Classification is checked for simulation and foresight of categorical variables. Types include churn forecasts, e-mail spam, financial manipulation, and students ' letter ratings.

Most regression algorithms have their equivalents, as you will see. Algorithms are modified instead of real numbers to estimate rank (or class probabilities).

• Unsupervised Learning–The clustering is an unintended learning method to identify natural statistical groups based on the intrinsic structure of your dataset. Types include consumer segmentation, e-commerce aggregation and review of social networks.

Since the clustering is not monitored (no "right response") the visualisation of data is usually used to evaluate results. If a "correct response" exists (i.e. you have pre-labeled clusters in your training set), classification algorithms are usually more suitable.

**Association**

Models of associations will find patterns in your data that associate one or more entities with one or more other entities (for example occurrences, purchases, and features). Models construct sets of rules that define these relations. In this case, data regions can work as inputs and goals. These associations can be discovered manually, but algorithms for association rules are much quicker and can detect more complex patterns. Examples of the use of such algorithms are Apriori and Carma. Another assembly model is a sequence detection model that detects sequential models in time-structured data.

Association models are most useful in anticipating different outcomes— for example, customers who bought X also bought Y and Z. Association models associate a certain conclusion with several conditions (e.g. purchasing anything). The advantage of algorithms of association rules compared to more standard algorithms of a decision tree (C5.0 and C&RT) is that associations can happen. A decision tree algorithm constructs rules with only one conclusion, while combined algorithms attempt to find many rules which can each be determined.

**Classification**

Classification models use one or more field values to predict one or more output or target regions. Examples of such techniques include regression (linear, logistic, generalized linear and Cox regression algorithms), selection trees (C&R Tree, QUEST, CHAID, and C5.0 algorithms), neural networks, Bayesian networks, and vector support machines.

Classification models support organizations to predict a known result, such as whether a client buys or leaves, or whether a company matches a recognized fraud pattern. Modeling techniques include machine learning, rule induction, subgroup identification, statistical methods, and model formation.

**Segmentation**

Divide data into segments or clusters of papers with a similar pattern input field. Since they're only interested in input fields, segmentation models have no concept of output or target field. Examples of segmentation models are Korhonen networks, K-Means clustering, error detection, and two-step clustering.

Segmentation models (also known as "clustering models") are useful when the result is unknown (e.g. identifying new fraud patterns or identifying interest groups in your customer base). Clustering models focus on acknowledging groups of similar papers and group labeling records. This is done without the benefit of prior knowledge of organizations and their characteristics, and it distinguishes clustering models from other modeling methods in that the model has no predefined output or target field to predict. These models have no right or inaccurate answers. Their value is determined by their ability to capture interesting data groups and provide useful descriptions of them. Clustering models are often used to produce clusters or segments that are subsequently used as inputs (e.g. by segmenting prospective customers into homogeneous subgroups).

As for my study, I don’t need to use regression and new factors or group factors.

**5.2 Select the appropriate data-mining methods based on the discussion**

This research predicts the trend of PM 2.5 and therefore employs the value of one or more areas of input to estimate the values of one or more areas of production or targets, so classification needs to be used. Moreover, the research will identify the connection between weather conditions and the concentration of PM 2.5, which implies that the regression model may be used in the classification to evaluate key indicators of concentration of PM 2.5. Nevertheless, the precision may be poor as too many variables affect the level of PM 2.5 and there are many kinds of results.

**6. Data mining algorithm selection**

There are many different data mining algorithms in the IBM SPSS modeler. They have distinct characteristics and methods for data assessment. For these studies, appropriate algorithms should be chosen to manage the data-set.

**6.1 Conduct exploratory analysis and discuss**

Until data mining, the data performance should be reviewed carefully to ensure that this is already progressing. Data was cleaned correctly in the previous step when the irrelevant data were dropped and the corresponding data remained in order. There are no missing data because all the missing data have been allocated the target mean. In addition, various graphs and tables must represent the dataset. Visualization facilitates data collection comprehension and study. Charts and tables should be used as a table and graphs during data analysis as they are easier to understand and intuitive. A map contains a lot of data that reflects thousands of words. In the visualization section histograms and dispersion plots are conducted to detect data quality, such as outliers and extremes, and the trend towards PM 2.5 in Chengdu is identified. At the same time, the relations between the factors and the concentration in PM 2.5 would also be found to see how they affect PM 2.5 concentration.

Therefore, since we want machine learning to be used in our education, it is important to develop a more convenient and accurate model. More decision-making insights will then be gained, based on the results provided by these models. For data mining a range of machine learning approaches can be used. In general, supervised methods and unattended methods are available.

This study should identify the relationship between PM 2.5 and the weather and then predict the trend of PM 2.5 in Chengdu. From the previous step, we already known that the "supervised" classification is a great choice. As a consequence, methods should be carefully chosen in classification. Classification models are used to predict the target value with one or more feature fields. A variety of types are offered: decision trees (C&R Tree, C5.0 al-gorithms), neural networks and support systems for vectors and bayesian networks. Each has its own advantages and disadvantages.

**•Classification**: **• Decision tree**: Decision tree learning uses an informatics decision book (as a predictive model) to shift from point observations to conclusions about the target value of the item (represented in the paper). It is one of the predictive models of statistics, data extraction and machine analysis. The tree models in which the target variable can contain a separate set of values are called classification tree; in these structures, leaves represent class labels, and branches represent combinations of characteristics that lead to such class labels. Decision trees where discrete values (typically real numbers) can be considered to be called regression trees by the objective variable.

A decision tree can be used to simply and visually depict decisions and decision-making. A decision tree reflects data in data mining (but it could be determined by the corresponding classification tree). This page deals with decision trees for data mining.

Advantages: Comprehensible and understandable. After a short explanation, people can understand the mechanisms of the decision tree. Shrubs can also be depicted graphically in a way which nonexperts can easily understand. You can process both numerical and categorical data. For the analysis of data sets, other approaches are usually used which have only one type of parameter. (For example, relationship rules can only be implemented with nominal values, while neural networks can be used with numerical variables or conversions of class values to 0-1 values). Because trees can handle performance predictors, stupid variables are not required. If one situation in a system is observed, boolean logic can easily explain the case. By addition, in a blackbox model, for example, an artificial neural network, the explanation for the result is generally hard to understand.

A prototype can be numerically evaluated. This takes into account the reliability of the model.

Non-statistic methodology without implication of learning knowledge or residual evaluation, e.g. no distributive, independent or continuous variance assumptions Well achieved for large data sets. Large amounts of data can be processed with standard computer tools in a reasonable time.

**Benefit**: Trees can be very unsolid. In various areas of optimality and even simple terms, the problem of learning a suitable choice tree is called NP-complete. Realist decision-making algorithms are therefore based on heuristic concepts like the greedy algorithm, where local choices are feasible. These algorithms can not guarantee the return of the globally optimal decision tree. Some methods such as the Dual Info Distance (DID) tree are suggested to reduce local optimalism's greedy effect.

**CART Tree**: Clasification and Regression tree (CART) is non-parametric tree learning software that generates a categorical or numerical classification or regression tree depending on whether the dependent variable is a categorical or numeric tree.

Decisive books consist of a set of rules based on models in the data set: rules based on the variables ' values are divided into the best fraction, to make observations based on the dependent variable After a rule is selected and the node is divided into two, a single method (i.e. a recursive procedure) is applied in each "kinder" node. (Otherwise the data is separated as far as possible and then the tree is split later.) Each branch of the tree ends in a terminal node. This statement falls into one terminal node and a set of rules are unique to each terminal node.

**Random forests** of Leo Breiman are a very popular method of predictive analysis.

Advantages: C&R Tree models are quite stable in the event of problems such as missing data and large numbers of fields. We typically don't have to forecast long periods of learning. C&R Tree models, however, are often easier to understand than some other models–the laws extracted from the model have a simple explanation.

Random forests are a supervised algorithm for learning. It can be used for both regression and classification. The algorithm is also the most flexible and easiest to use. There are trees in a wood. The more trees it has, the better the wood. Random forests generate decision-making trees on randomly selected data samples, obtain prediction from each tree and select by voting the best solution. It also provides a very good indicator of the value of the function.

Random forests have a wide range of applications, including suggestion motors, object identification and selection of features. It can be used to categorize faithful lenders, classify fraudulent activities and predict disease. At the base of the Boruta algorithm, essential features are selected in a dataset.

**C5.0 Tree**: A templateC5.0 operates by dividing the sample by the region with the largest data gain. The first sub-sample is then split again, usually based on a different region, and the process repeats until it is impossible to further split the sub-samples. Finally, the lowest-level divisions are reviewed and those that do not contribute significantly to the quality of the template are removed or removed.

Two versions can be generated by C5.0. A decision tree is a simple description of the splits of the algorithm. Terminal node (or "leaf") is a special part and a terminal node in a tree, in each case. In other words, one data record raised to a decision tree can be predicted exactly.

**Advantages**: C5.0 models are quite robust to problems such as missing data and large quantities of input fields. Long training times are typically not needed to estimate. Nevertheless, C5.0 models are more easily understood because the laws derived from the system were very easily interpreted.

**Return**: • Linear Return: linear models forecast a continuous target based on linear relations between an objective and a predictor or more.

**Advantages**: Linear models are relatively simple and provide a readily understood numerical measurement method. The characteristics of such models are well known and can typically be developed very fast in the same set of data as other models.

**Logistic regression**: logistic regression is a statistical method of classified data based on input field values, also known as marginal regression. • Logistic regression. This is similar to linear regression; it takes a categorical target area instead of numerical. Both binomial (two category targets) and multinomial (two category target) models are supported.

Logistic regression works by building a set of equations that relate the input field's value with the probability of each class of the output field. The probabilities of new data can be determined once the model is created. For each record, the likelihood of membership is determined for each potential output group. The highest probability target group is assigned to this record as the estimated output value.

**Benefits**: logistic regression models are often accurate. Symbolic and numerical data fields can be treated. In all target categories, we can offer estimated probabilities to easily identify a second-best guess. When group membership is based on constant field values (e.g. high IQ compared to low IQ), the rich information provided in a wide range of values can be viewed using a linear regression. Logistic models are most efficient when group membership is a categorical domain. Logistic models can also allow automatic field selection, although by other approaches, such as tree models and functional selection, these can be more efficiently done in large data sets.

**6.2 Select data-mining algorithms based on discussion**

The goal of the present study is to determine the future atmospheric value of PM 2,5 concentrations in Chengdu and to define the relationship between PM 2,5 concentration and environment. The Decision Tree algorithm can create multiple branches with less resources, as Decision Tree algorithm.

The reason why Random Forest is chosen is because it is an ensemble algorithm based on decision tree and works better than a single decision tree algorithm. XGBoost Tree is chosen as the base model because it is an advanced gradient algorithm that promotes tree structure. Upgrade algorithms to test poor graders and apply them to a strong final grade. XGBoost Tree is more useful and the flexibility is higher.

**6.3 Choose suitable models and parameters**

**Random Forest**

****

Random Forest is a sort of monitored, ensemble-based learning machine algorithm. Learning ensemble is a kind of learning in which you multiply distinct kinds of algorithms or the same algorithm to create a stronger forecast model. The algorithm of the random forest combines several algorithms of the same nature: several choice trees leading to a forest of trees, therefore the name "Random Forest." For both regression and classification assignments, the random forest algorithm is used.

I choose the random forest model because There are various trees and then each tree is trained on a subset of data since the random forest algorithm is not partial. The random forest algorithm depends basically on the strength of the' crowds' and therefore reduces the general bias of the algorithm. This is a very stable algorithm. Even when the latest dataset data point is entered, the general algorithm is not influenced because fresh data can affect a tree, but it is very difficult for it to affect all trees. The random forest algorithm operates well if both categorical and numerical characteristics are present. Also, the random forest algorithm operates well when information has missing values or has not been well-scaled. There is default value with no max depth.

**Decision Tree**

****

Decision Tree is an ML algorithm type white box. It shares the inner decision-making logic that is not accessible in algorithms like the Neural Network in the black box form. Its training time is quicker than the algorithm of the neural network. The time complexity of decision-making bodies depends on the number of documents and the number of characteristics in the information. The decision tree is a non-parametric method that does not rely on the assumptions of probability. Decision trees can manage with excellent precision high dimensional information.

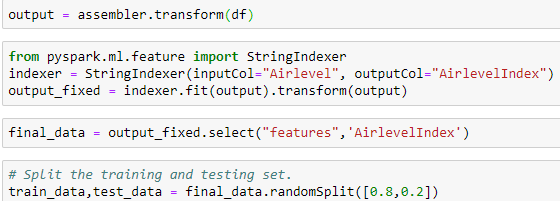
ID3 is an algorithm created by Ross Quinlan to generate a decision tree. ID3 is the precursor of the C4.5 algorithm, typically used in machine education and natural language treatment.

Decision tree is widely used in data mining. The goal is to create a model based on multiple input variables to forecast a target variable. -inner node corresponds to one of the input variables; children have edges for each of the potential input variable values. The leaf is a target variable value given the root-to-leaf input variable values.

For example, a decision tree is a simple representation. Suppose all output features have infinite distinct domains, and this section has a single target feature called "classification." Every classification domain element is called a class. A decision tree or a classification tree is a tree with an internal (non-leaf) level output variable. The node arcs labeled with an input function are labeled with each possible target or output value, or the arc leads to a different input decision node. Each tree is labeled as a class or probability distribution over the classes, meaning a data set is classified by the tree into either a certain class or a given probability distribution (which, if well-constructed, is skewed to certain class sub-sets).

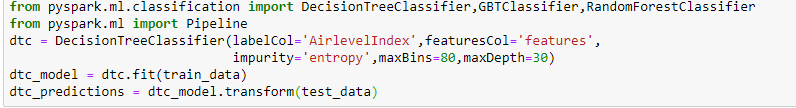
A tree is formed by dividing the source set into subsets comprising the tree's root node-the successor children. The distinction was based on a series of controversial, category-based rules. This process is repeated recursively on each derived subset, called recursive partitioning. This recursion is achieved if the node subset has the same values as the target parameter, or if splitting no longer adds value to forecasts.

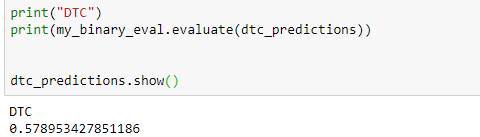
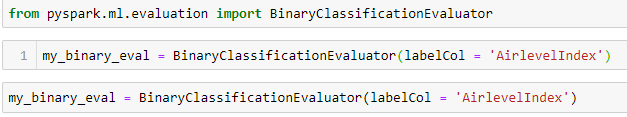
**7. 1 Create and justify test modelling**

****

Turrini, E., Carnevale, C., Finzi, G., & Volta, M. (2018) stated that minimizing the mistakes caused by data discrepancies and then understand the model better. A training subset that accounts for 80 percent, and a 20 percent test subset, would be an excellent choice for allocating the percentage for these two subsets. By using the partition, our data is separated into two subsets, a training subset, and a test subset. The data in the training subset would usually be larger and the test subset would contain fewer data.

**7.2 Data Ming Model (Accuracy Testing)**

****

****

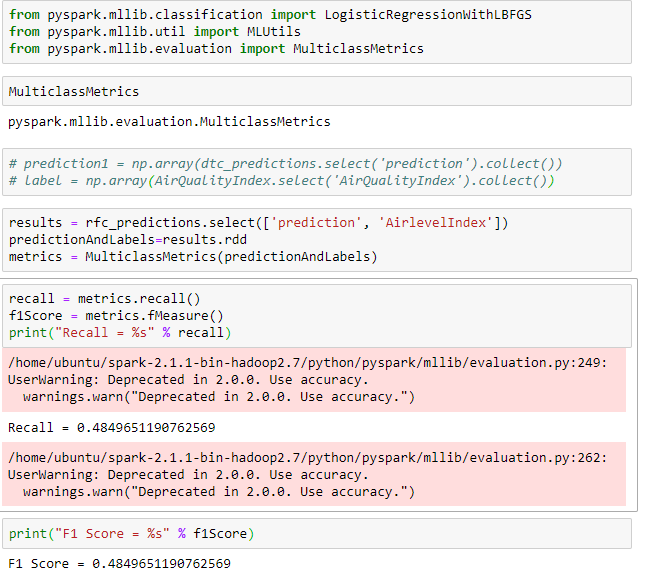
**The accuracy for decision tree model is around 58%.**

Classification can be viewed as two separate issues–binary and multi-class classification. In the binary grouping, only two classes are involved in a better understood function, while multiple classes include the assigning of an object to one of several categories. Since many classification methods have especially been developed for binary classification, multi-class classification often requires multiple binary classifiers to be used together.

Precision (also known as positive predictive value) is the fraction of the appropriate cases among the instances retrieved in pattern recognition and binary classification. Recall (also known as awareness) is the fraction of appropriate cases that have been obtained over the complete number of the appropriate cases. Accuracy and reminder are therefore based on an understanding and significance measurement.

After that, by using the evaluation model, the command code (recall) is as followed:

**The Logistic Regression Classifier:**



**As we can see, the accuracy by using the logistic model is lower than before, which is 48.5%.**

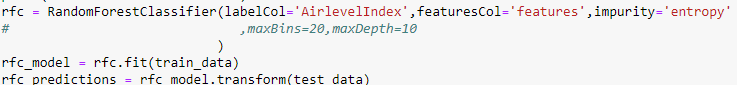
For the classification function, the consistency is the number of truly positive elements (i.e., the number of good items that belong to the positive category) divided by the total number of elements which belong to the positive class (i.e. the sum of true positive and false positive elements incorrectly identified as being class members).

Reminders are described in this context as the number of true positive elements divided by the total number of elements which actually belong to the positive group, that is, the total number of true positive and false negative elements which were not classified as belonging to but should have been.

When the precision and the retrieval rate are small, the cost of F1 value is small (0.485). The F1 is the best value at 1 (perfect accuracy and retrieval rate) while the worst value is 0. The F1 value is a measure of test accuracy in the binary classification. Therefore, the record for this graph is 0.485 and the score for F1 is also 0.485.

**Random Forest**

**The random forest classifier:**

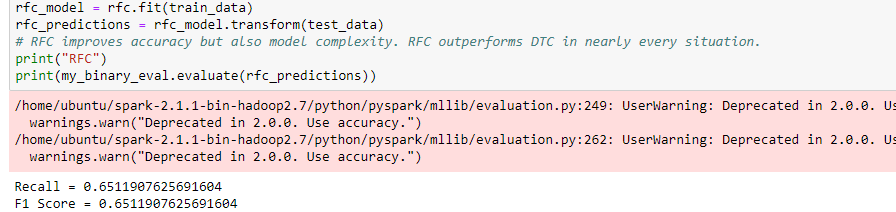
****





**As we can see, the accuracy of RFC is 57.3%.**

**The accuracy of the random forest model:**

****

By using the Random Forest model, the accuracy is higher than using the Decision Tree mode.

For the classification function, the consistency is the number of truly positive elements (i.e., the number of good items that belong to the positive category) divided by the total number of elements which belong to the positive class (i.e. the sum of true positive and false positive elements incorrectly identified as being class members).

Reminders are described in this context as the number of true positive elements divided by the total number of elements which actually belong to the positive group, that is, the total number of true positive and false negative elements which were not classified as belonging to but should have been.

When the precision and the retrieval rate are small, the cost of F1 value is small (0.485). The F1 is the best value at 1 (perfect accuracy and retrieval rate) while the worst value is 0. The F1 value is a measure of test accuracy in the binary classification. Therefore, the record for this graph is 0.651 and the score for F1 is also 0.651.

## **7.3 Search for patterns**

The accuracy of the pattern:

|  |  |
| --- | --- |
| Modeling | Accuracy |
| Random Forest | 57.3% |
| Decision Tree | 48.5% |

# The accuracy of Random Forest is higher than Decision Tree.

# **8. Interpretation**

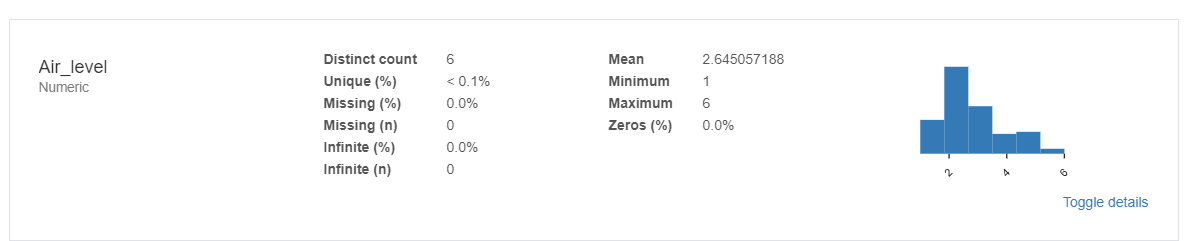
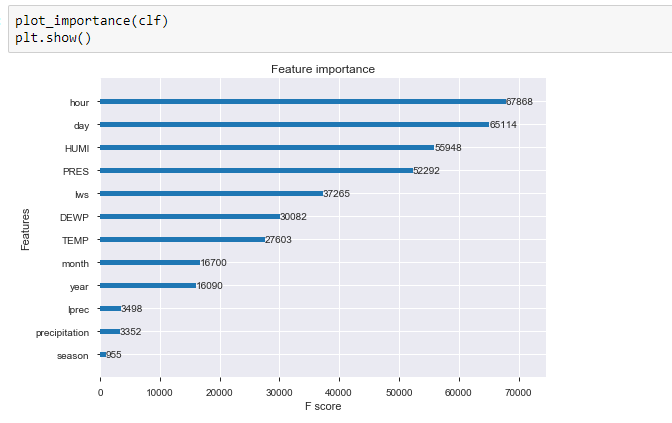
## 

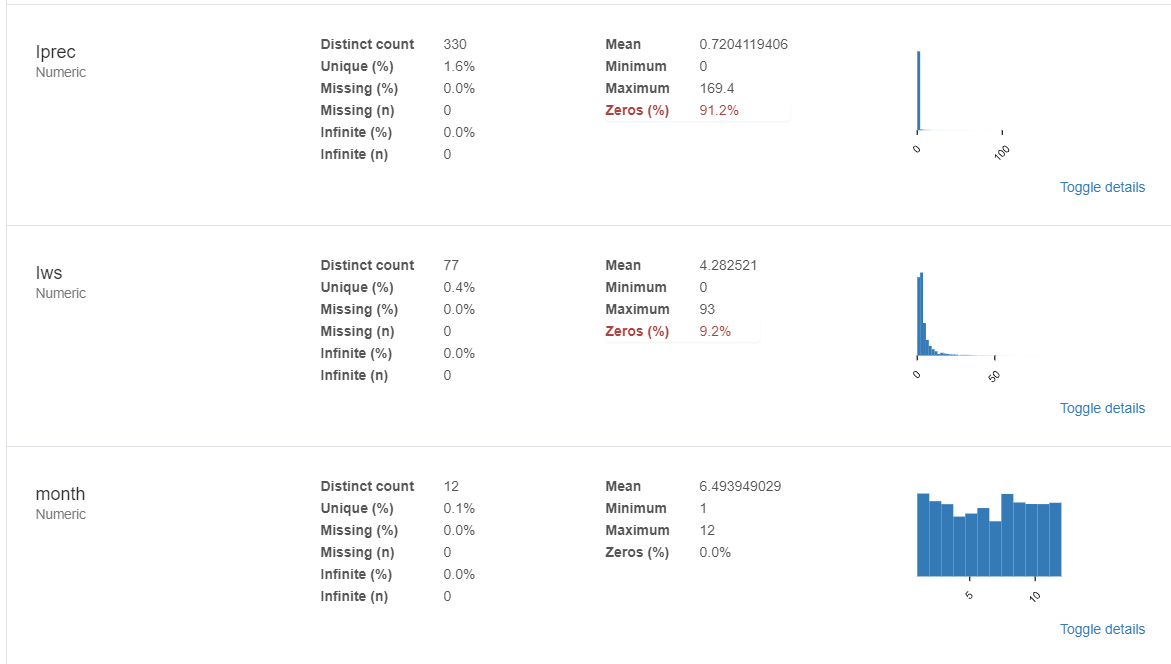
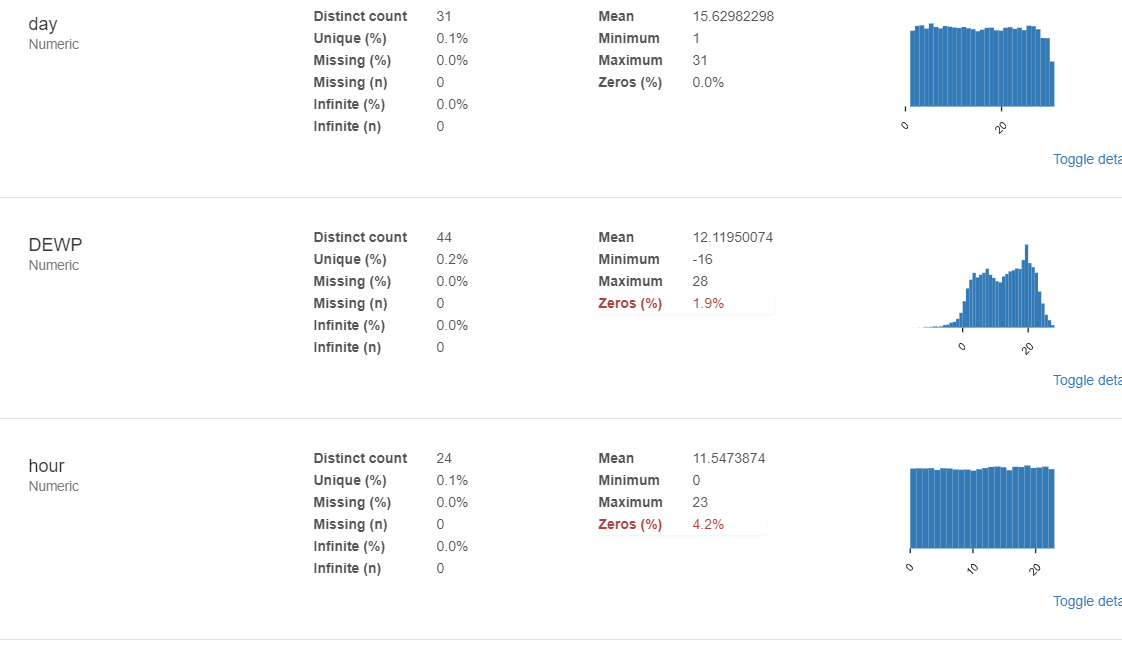
## **8.1 Study and discuss the patterns**

Modeling is a spiraling, continuing optimization technique. At the end of each model, it is necessary to determine whether analysis results are meaningful, the characteristics are clear and acceptable. If the outcome is not optimal, we must modify the model and optimize it. We can do this by altering the input value, variables, route, and mining algorithm coefficient.

The above therapy acquires a sequence of analytical results and patterns. They define the multi-sided goal problem. They must be inspected and evaluated to acquire sensitive and complete decision-making data. Comparison, accuracy, and assist must be used to check model results to determine model value. We need to introduce more layers and background users to test and verify the final optimization model through comprehensive multi-model comparison.

**8.2 Pattern Visualization**

****

****

Base on the different variable, we want to get which variables is the most important factor to affect Chengdu Pm 2.5 factors. As we can see that hour and day are the highest factors. On the temporary, we need to select the most accurate model in phase 6-7.

**8.3 Interpret the results, models, and patterns**

From the above figures, air quality at the dew point, higher pressures and decreased accumulated rain times in autumn and winter can be discovered to be worse.

As for the primary seasonal factor, people usually use coal and gas for heating in autumn and winter. The mixture of the climate and public outcomes in a higher concentration of Pm2,5 and in autumn and winter a lower quality of water.

Human activities could also affect the concentration of PM 2.5 not only in environmental conditions. In the morning and evening, a large quantity of government or personal transportation would generate PM 2,5.

Random Tree performance is better after comparison than other classification algorithms. Somehow the Random Tree can predict the air quality index. Following the prediction, people can prepare to prevent air pollution before it happens.

**8.4 Assess and evaluate results, models, and patterns**

The Random Forest algorithm can solve the problem of classification. The Decision Tree development method was based on the principle of the highest level of knowledge gain for choosing nodes and separating points. This analysis thus uses specificity to demonstrate the effect of this algorithm and the data gain of Random Forest algorithms. The analytical picture shows that the lower limit of accuracy is approximately 57 percent higher than the Decision Tree; thus, this result is satisfactory and reliable because the study uses training data set and testing information to analyse the model.

Furthermore, after analysing Decision Tree, the development of a time graph contrasts the air quality pattern between 2010 and 2015 through training and testing. It shows that training and testing trends are similar, except in some extreme areas. Ultimately, the findings of the study can be agreed based upon the above proof, as the predicted target is greater than 57 percent. It can help people to predict air quality to some extent after using this model.

**8.5 Iterate prior (step one to seven)**

**In phase 1**, this research recognizes and sets the objectives of the pm2.5 concentration in Chengdu. After reading more, we fully understand the subject and changed the goal several times to meet the requirement.

**In phase two,** this research originally investigates the data and discovered some initial data from the dataset. Because of the incorrect dataset, three datasets were discovered to be analyzed. Some of them, however, are unrelated to this subject. Finally, we have chosen the most suitable.

**In phase 3**, data mining including, cleaning, selection, and construction was prepared.

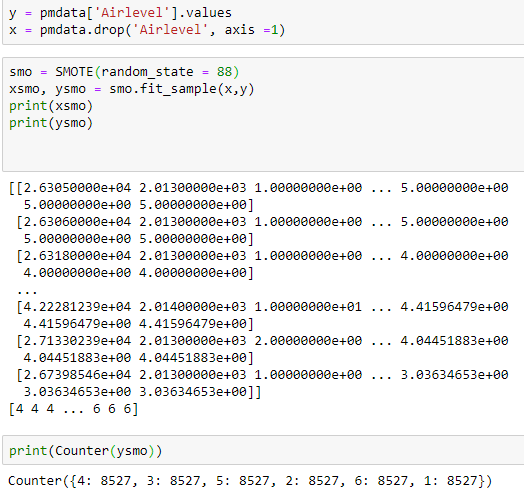
In this research, we used various ways of cleaning information, including the clean missing information feature for dealing with missing values and replacing the function with outer and extreme values in pandas. We might choose to remove them after comparing them.

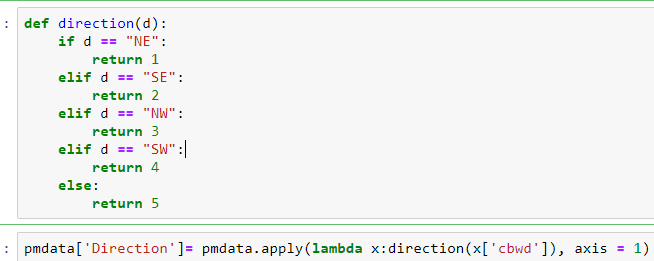
****





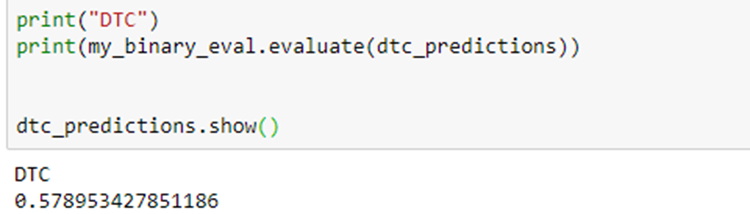
**In Phase 4,** I start to clean and delete meaningless data and variables. Also, to make the data balance.

****

****

**In phase five**, this research confirmed the use of regression techniques for mining information after a debate of the distinct techniques. However, once the precision has been found is small and the air quality has to be predicted, we add a different process, a classification, to predict.

**In phase six**, this research discovers different appropriate models and algorithms by discussion. To selected three models in order to get the highest accuracy. I chose the Decision tree and Random Forest Model to compare their accuracy. The outcome is that in my dataset, Random Forest is more accurate.





**In this phase seven**, this study defines the template test method, performs data mining and searches for correct patterns. In this step we iterate at least the tree model, after data mining with various parameters, to look for the right pattern. Following the combination of the tests, we can achieve the best model in this analysis. Finally, this study is published in Github and the URL is https://github.com/hon611/iteration4-assignment.git.

**Reference:**

Schwartz, J., Laden, F., & Zanobetti, A. (2002). The concentration-response relation between PM (2.5) and daily deaths. *Environmental health perspectives*, *110*(10), 1025-1029.

Turrini, E., Carnevale, C., Finzi, G., & Volta, M. (2018). A non-linear optimization programming model for air quality planning including co-benefits for GHG emissions. *Science of the Total Environment*, *621*, 980-989.

Zhang, Y. L., & Cao, F. (2015). Fine particulate matter (PM 2.5) in China at a city level. *Scientific reports*, *5*, 14884.

**Disclaimer**

"I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic integrity and copyright. (See: <https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.html>).

I also acknowledge that I have appropriate permission to use the data that I have utilised in this project. (For example, if the data belongs to an organisation and the data has not been published in the public domain then the data must be approved by the rights holder.) This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student's misuse of data."

Name: Kwan Hon Lam