**Assignment 1**

**Data Exploration and Regression Analysis**

**Introduction**

For the purpose of the Data Exploration and Regression Analysis, I have chosen to work with Air Quality in Italy Dataset the UCI Machine Learning Repository have provided the dataset. The quality of the air is assessed by using five different type of metal oxide chemical embedded in an Air Quality Chemical Multisensory Device. These devices have placed in areas where there is significant amount of pollution throughout the cites of Italy. The data that is stored in the spreadsheet has been collected from March 2004 to February 2005.

One of the main of problem of the harsh air quality is that, when we breathe in these toxic pollutants into the lungs it damages the air ways and causes breathing problems and chronic disease, which then cause hospitalisation. In Italy it was recorded by World Health Organisation that Italy is one the highest number of deaths because of bad air quality it mentioned that there are 8220 deaths per year because of the increased concertation of 03 gas in the atmosphere.

**Objective**

For the purpose of the Regression Analysis, I want to analyse the Absolute Humidity and the measurement of air pollutants. The reason I want to analyse these two variables is because I want to see if the any type of association between these two variables.

For the purpose of this analysis, I have to make some assumption which are, I have to assume that the time that the data is collected is a good represent of the air quality in Italy and I also have to assume that the readings from the five different sensors are accurate and that they sensor are not faculty.

**Data Description**

The dataset contains 9358 hourly averaged responses from a set of five metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was placed on a field in a heavily polluted area, at street level, within an Italian city. Data were collected from March 2004 to February 2005 (one year), representing the longest freely available recordings of responses from on-field deployed air quality chemical sensor devices. A co-located reference certified analyser provided hourly averaged concentrations of CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx), and Nitrogen Dioxide (NO2). Cross-sensitivities, as well as concept and sensor drifts, are present, as described in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required), eventually affecting sensors' concentration estimation capabilities. -200 is assigned to missing values. This dataset is only for academic use. Commercial purposes are categorically prohibited.

**Attribute Information**

|  |  |
| --- | --- |
| **Date** | (DD/MM/YYYY) |
| **Time** | (HH.MM.SS) |
| **CO(GT)** | True hourly averaged concentration CO in mg/m^3 (reference analyzer) |
| **PT08.S1(CO)** | (Tin oxide) hourly averaged sensor response (nominally CO targeted) |
| **NMHC(GT)** | True hourly averaged overall Non Metanic Hydrocarbons concentration in microg/m^3 (reference analyzer) |
| **C6H6(GT)** | True hourly averaged Benzene concentration in microg/m^3 (reference analyzer) |
| **PT08.S2(**NMHC) | (titania) hourly averaged sensor response (nominally NMHC targeted) |
| **NOx (GT)** | True hourly averaged NOx concentration in ppb (reference analyzer) |
| **PT08.S3(NOx)** | (Tungsten oxide) hourly averaged sensor response (nominally NOx targeted |
| **NO2(GT)** | True hourly averaged NO2 concentration in microg/m^3 (reference analyzer) |
| **PT08.S4(NO2)** | (Tungsten oxide) hourly averaged sensor response (nominally NO2 targeted |
| **PT08.S5(O3)** | (Indium oxide) hourly averaged sensor response (nominally O3 targeted) |
| **T** | Temperature in Â°C |
| **RH** | Relative Humidity (%) |
| **AH** | AH Absolute Humidity |

**Data Types**

From the data description the data description above, we can see that there only four distinct types of data, which consists of float point number, integers, objects and datetime.

<class 'pandas. core.frame.DataFrame'>

RangeIndex: 9357 entries, 0 to 9356

Data columns (total 15 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Date 9357 non-null datetime64[ns]

1 Time 9357 non-null object

2 CO(GT) 9357 non-null float64

3 PT08.S1(CO) 9357 non-null float64

4 NMHC(GT) 9357 non-null int64

5 C6H6(GT) 9357 non-null float64

6 PT08.S2(NMHC) 9357 non-null float64

7 NOx(GT) 9357 non-null float64

8 PT08.S3(NOx) 9357 non-null float64

9 NO2(GT) 9357 non-null float64

10 PT08.S4(NO2) 9357 non-null float64

11 PT08.S5(O3) 9357 non-null float64

12 T 9357 non-null float64

13 RH 9357 non-null float64

14 AH 9357 non-null float64

dtypes: datetime64[ns](1), float64(12), int64(1), object(1)

memory usage: 1.1+ MB

**Finding Missing Data**

When doing an analysis of a dataset its important to finding all Null Attributes reasoning being it can cause a major impact when analysing the data as it causes a miss representation of data.

<class 'pandas.core.frame.DataFrame'>

Int64Index: 827 entries, 0 to 826

Data columns (total 15 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Date 827 non-null datetime64[ns]

1 Time 827 non-null object

2 CO(GT) 827 non-null int64

3 PT08.S1(CO) 827 non-null float64

4 NMHC(GT) 827 non-null float64

5 C6H6(GT) 827 non-null float64

6 PT08.S2(NMHC) 827 non-null float64

7 NOx(GT) 827 non-null float64

8 PT08.S3(NOx) 827 non-null float64

9 NO2(GT) 827 non-null float64

10 PT08.S4(NO2) 827 non-null float64

11 PT08.S5(O3) 827 non-null float64

12 T 827 non-null float64

13 RH 827 non-null float64

14 AH 827 non-null float64

dtypes: datetime64[ns](1), float64(12), int64(1), object(1)

memory usage: 103.4+ KB

By removing the all the missing in the dataset, the dataset only contains only 827 rows of data. By removing all the missing values, it will reduce clutter in the dataset, therefore being able visualise the data better when examine the plot.

**Measurement Unit Error**

When looking through the description of the attributes. I noticed that one of the attributes “ CO(GT)” is measured in mg/m^3 and the rest of the independent variable were measured in microg/m^3. In order to fix this, I had to convert mg/m^3 to microg/m^3 which can be done by multiplying that entire column by hundred in the excel spreadsheet now that all the attributes are in the same unit of measurement we can move on the next phase of the analysis.

**Analysing Data**

By using the matrix plot is possible to see all the histograms and scatter plots that can be produced, from an analysis the matrix plots it can see that most the histograms are positively skewed, which mean that the mode value, which is the most common value that the variable produce is higher than the mean median. An example of the is when we look at the independent variable CO(GT) which is measure the concentration carbon dioxide in microg/m^3

A picture containing text, crossword puzzle, scoreboard, vector graphics

Description automatically generated

Now moving onto the scatter plot. The purpose of using a scatter plot is the determine if there is a positive or negative relationship between the independent variables and the dependent variables. By looking at the matrix plot we can see that most the variables have a positive relationship, what this means is that when the independent is increasing the dependent variable is also increase. An example for this Absolute Humidity vs the PT08.S5(O3), Absolute humidity is the measure of actual moisture in air and PT08.S5(O3) which is a sensor that detect gas that are released of the Ozone.

Chart, scatter chart

Description automatically generated

What the scatter plot actually tells us is that , when there is an increased presents of ozone gas in the air the amount of moisture in the air also increases. This can be said about other sensors as well reason being they follow the same trend, expect for the PT08.S3(NOx) , which measure the amount of Nitrogen oxide in the air.

Chart, scatter chart

Description automatically generated

**Corelation Analysis**

The purpose data correlation analysis is to evaluate the relationship between variables. Correlation of a sample is measured using the r value , which show the strength of the variable. The variables that are closer 1 have a positive association, correlation value that are closer have negative association.

From the correlation plot it can see that most of the variables have positive association where the correlation value is close to , keeping mind that there are few negative associations where the correlation value is close to -1

Calendar

Description automatically generated

From looking at the correlation matrix plot there is strong relationship between the references analyser that contains the concentration of gas in microg/m^3 and the sensor that are detecting the gases, the corelation of those variables range from 0.70 to 0.98 meaning that they have a strong positive association between the variables.

**Multiple Linear Regression - Ordinary Least Square Test**

|  |  |  |  |
| --- | --- | --- | --- |
| **OLS Regression Results** | | | |
| **Dep. Variable:** | **AH** | **R-squared:** | **0.944** |
| **Model:** | **OLS** | **Adj. R-squared:** | **0.943** |
| **Method:** | **Least Squares** | **F-statistic:** | **1143.** |
| **Date:** | **Fri, 15 Apr 2022** | **Prob (F-statistic):** | **0.00** |
| **Time:** | **00:28:47** | **Log-Likelihood:** | **1443.7** |
| **No. Observations:** | **827** | **AIC:** | **-2861.** |
| **Df Residuals:** | **814** | **BIC:** | **-2800.** |
| **Df Model:** | **12** |  |  |
| **Covariance Type:** | **nonrobust** |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | **57.853** | **Durbin-Watson:** | **0.472** |
| **Prob(Omnibus):** | **0.000** | **Jarque-Bera (JB):** | **245.041** |
| **Skew:** | **0.104** | **Prob(JB):** | **6.17e-54** |
| **Kurtosis:** | **5.659** | **Cond. No.** | **1.97e+05** |

Looking at the OLS test results, we can see that the Adjusted R squared value is 0.943, which means that the independent variable can explain 94 percent of the changes in Absolute Humidity. The Prob (F-Statistic) indicates the null hypothesis's accuracy (H0 Effect of all independent variable in regression model is zero). Our Prob(F-statistic) has an incredibly low value of 0.00 from the test above; because this value is less than 0.05, we reject the null hypothesis in support of the hypothesis that our model fits data better than the intercept only model. The p-value of this regression model is 0.00 which mean that we have to reject the null hypothesis (no statistically significant association between the air pollutants and Absolute Humidity

**Statistical Inference**

From the analysis I can that there is positive relationship between the air pollutants and the Absolute Humidity reason being the scatter plots have proven that there is positive trend, I can also prove this because the p-value is 0.00 which means that there some kind of association between the air pollutants reason being because 94% of the change in the water moisture content can be related to the air pollutant. What this means for the Air Quality in Italy is if more and more air pollutants are being released into the environment, which can cause the air pollutant to be absorbed into the moisture and if the citizen breath in that moisture and it cause harm to the lungs and cause harmful diseases.

Reference:

Saverio De Vito (saverio. devito **'@'** enea.it), ENEA - National Agency for New Technologies, Energy and Sustainable Economic Development