AIR QUALITY DATA ANALYSIS

## Introduction:

Analyse which factors can influence the quality of the air and see what the possible consequences of air pollution are.

## Dataset Description:

The data set contains the responses of a gas multisensory device deployed on the field in an Italian city. Hourly responses averages are recorded along with gas concentrations references from a certified analyser. Dataset - Vito,Saverio. (2016). Air Quality. UCI Machine Learning Repository.

Dataset link: <https://doi.org/10.24432/C59K5F>

Spreadsheet:<https://docs.google.com/spreadsheets/d/15itWzBsLsLYR7s9oZloHtxHUYYOSraJB/edit?usp=sharing&ouid=115954894100227476468&rtpof=true&sd=true>

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | Date | Qualitative | Ordinal |
| **Time** | Time (in hours) | Quantitative | Discrete |
| **CO (GT)** | True hourly averaged concentration CO  in mg/m^3 | Quantitative | Continuous |
| **PT08.S1(CO)** | (tin oxide) hourly averaged sensor  response (nominally CO targeted) | Quantitative | Discrete |
| **NMHC(GT)** | True hourly averaged overall Non  Methanic Hydrocarbons concentration in microg/m^3 | Quantitative | Discrete |
| **C6H6(GT)** | True hourly averaged Benzene  concentration in microg/m^3 | Quantitative | Continuous |
| **PT08.S2(NMHC)** | (titania) hourly averaged sensor response  (nominally NMHC targeted) | Quantitative | Discrete |
| **NOx (GT)** | True hourly averaged NOx concentration  in ppb | Quantitative | Discrete |
| **PT08.S3(NOx)** | (tungsten oxide) hourly averaged sensor  response (nominally NOx targeted) | Quantitative | Discrete |
| **NO2(GT)** | True hourly averaged NO2 concentration  in microg/m^3 | Quantitative | Discrete |
| **PT08.S4 (NO2)** | (tungsten oxide) hourly averaged sensor  response (nominally NO2 targeted) | Quantitative | Discrete |
| **PT08.S5 (O3)** | (indium oxide) hourly averaged sensor  response (nominally O3 targeted) | Quantitative | Discrete |
| **T** | Temperature | Quantitative | Continuous |
| **RH** | Relative Humidity | Quantitative | Continuous |

## Concept application:

## Module 1:

The NO2 levels in the air quality of various states are assumed to be normally distributed. The mean NO2 level is 58 parts per billion (ppb), and the standard deviation is 12 ppb. What is the probability that the NO2 level in a randomly selected state falls between 51 ppb and 62 ppb?

Analysis:

The Normal random variable we are looking at is NO2 level.

Let the NO2 level be ‘x’.

Population Mean, µ = 58 ppb

Population Standard deviation, σ = 12 ppb

The probability that the NO2 level falls between 51 ppb and 62 ppb is **P (51 < x < 62)**

Standardizing the normal random variable

P ()

= P (-0.5833 < Z < 0.3333) (∵ )

= P (Z < 0.3333) – P (Z < -0.5833)

= 0.6306 - 0.2798 (use of the formula NORM.S.DIST(Z, Cumulative) in MS Excel)

= 0.3507

The probability that the NO2 level falls between 51 ppb and 62 ppb is 0.35 i.e., there is 35% chances that a randomly selected NO2 level falls in the range of 51 ppb and 62 ppb.

## Module 2:

A climatologist wants to determine how temperature affects air quality in a specific region. The mean temperature recorded is 9.8 degrees Celsius, and the known population standard deviation is 3.2 degrees Celsius. If a sample of 230 observations was collected for the study, construct a 95% confidence interval for the true mean temperature in this region. The climatologist is interested in understanding how temperature influences air quality.

Analysis:

The Normal random variable we are looking at is Temperature.

Let the Temperature be ‘x’.

Sample Mean, x̄ = 9.8 0C

Population Standard deviation, σ = 3.2 0C

Sample Size, n = 230

100(1- α) = 95%

α/2 = 0.025

Z α/2 = -1.96

Confidence Interval

= 9.8 ± -0.4136

9.3864 to 10.2136

Thus, we are 95% confident that the true mean number of Temperature recorded will be in the above interval.

## Module 3:

To ensure compliance with government regulations that limit atmospheric pollutant emissions through industrial smokestacks, Greentech Corp collects a random sample of 51 temperature readings. The sample's mean temperature is found to be 26.8 degrees Celsius, with a known population standard deviation of 3.3 degrees Celsius. The objective is to determine whether the data supports the claim that Greentech Corp's average temperature is below the mandated limit of 30 degrees Celsius at a significance level of 0.02. It is assumed that the population of temperature readings is approximately normally distributed.

1. What is the null and alternative hypothesis?
2. Calculate the value of the test statistic for this hypothesis test.
3. What is the P value associated with the z statistic?
4. State the decision.

Analysis:

A) :

*Alternative Hypothesis*, :

B) Given Sample Mean ( = 26.8,

Population Standard Deviation(s) =3.3.

Population Mean (μ)= 30,

Sample Size, n = 51

Population **Standard** Deviation () is Known, and Sample Size = 51, Hence, we will use a Normal Distribution, which means we need to use the Z-Test Statistic,

**Z = ,** Z = , Z = -6.96

C) With a p-value of approximately 0.000 and a given level of significance (α) of 0.02, we compare the p-value to the level of significance.

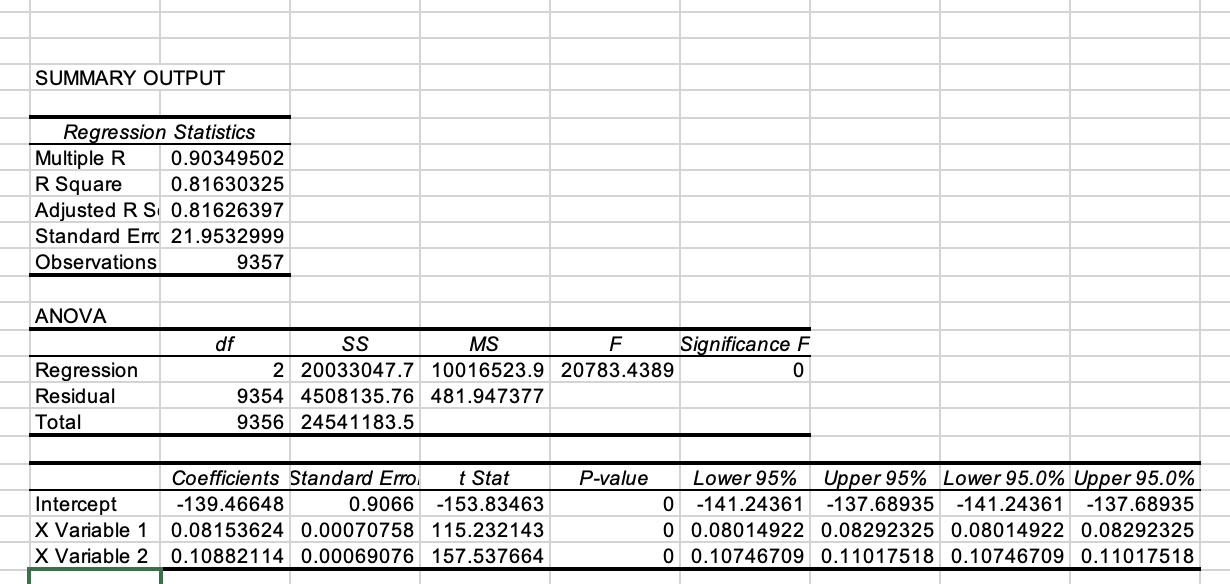
D) Since the p-value is much smaller than the significance level, we reject the null hypothesis.

Therefore, there is sufficient evidence at the 0.02 significance level to conclude that Greentech Corp’s average temperature is below the mandated limit of 30 degrees Celsius.

## Module 4:

Is there a significant statistical relationship between relative humidity (dependent variable) and the concentrations of tungsten oxide and tin oxide (independent variables) in predicting air quality? Using this dataset of air quality measurements, perform a multiple regression analysis to assess the impact of these two independent variables on relative humidity and, consequently, air quality. Report your findings to determine if tungsten oxide and tin oxide concentrations have a significant influence on relative humidity and air quality at 0.10 level of significance.

Analysis:



From the above output, as the p-value is less than the level of significance, we can state that there is a statistically significant relationship between these variables.