**Final Report**

**Data Description:**

The dataset that’s been selected for this project is titled “South Korean Pollution- Pollutant levels in South Korea”. The data was sourced from Kaggle (<https://www.kaggle.com/datasets/calebreigada/south-korean-pollution?resource=download>). It consists of sample observations of pollution level measurements in various cities and their districts in the country of South Korea from the year 2013 to 2022. The data fields/columns consist of the following:

* date - date of measurement
* pm25 - fine particulate matter (PM2.5) (µg/m3)
* pm10 - fine particulate matter (PM10) (µg/m3)
* o3 - Ozone (O3) (µg/m3)
* no2 - Nitrogen Dioxide (NO2) (ppm)
* so2 - Sulfur Dioxide (SO2) (ppm)
* co - Carbon Monoxide (CO) (ppm)
* Lat - Latitude where measurement was taken
* Long - Longitude where measurement was taken
* City - City where measurement was taken
* District - District where measurement was taken
* Country - Country where measurement was taken

**Goal of Data research Project**: Above are six major pollutants measured to estimate pollution levels. The goal of this project is to compare the average levels of pollution between cities and districts as well as check for any upward or downward trends over the years. Attention will also be paid to the pollution level comparisons between metropolitan and more rural areas. The goal is to find the difference between pollution levels in the different districts of South Korea and decipher the safest in terms of health for tourists or citizens to visit while calculating the likely pollution levels they’ll experience.

**Research Hypothesis and Methodology**: Despite their hazardous power plants, with their moves towards cleaner air, it is hypothesized that South Korea has reduced its overall levels of pollution.

It is also hypothesized that province districts experience lower average pollution than metropolitan districts.

This project thesis will make use of more than 34,000 sample observations to find the mean, standard deviation and yearly trends of the various pollutants in South Korea. Comparisons will be made, and confidence intervals will be constructed.

**Module 1 research question:**

In a sample of 34530 daily pollution measurements with a mean of 53.225 µg/m3 and a standard deviation of 39.952 µg/m3, what is the probability that a pollution measurement taken on a random day for pm25 in South Korea falls in the healthy range for the average individual, between the good (12 µg/m3) and Unhealthy levels of concern (55.4 µg/m3)? (According to the EPA’s Air Health quality index)

**Result:**

μ=53.225 µg/m3

σ=39.952 µg/m3

P(12 µg/m3 < X < 55.4 µg/m3)

P(X>12 µg/m3) = P(X> (12-53.225)/39.952) = P(X> -1.03)

P(X<55.4 µg/m3) = P(X< (55.4-53.225)/39.952) = P(X< 0.05)

P(X> -1.03) = 0.8485

P(X< 0.05) = 0.5199

P(12 µg/m3 < X < 55.4 µg/m3) = 0.8485-0.5199= 0.3286 or 32.86 %

The probability that a pollution measurement taken on a random day for pm25 in South Korea falls in the healthy range for the average individual is **32.86%**

**Module 2 research question:**

Given the sample of 34530 daily pollution measurements, what is the 95% confidence interval for the average pm 2.5 measurement of the population of South Korea?

**Result:**

μ=53.225

σ=39.952

n=34530

At 95% confidence interval, the critical value that should be used in constructing the confidence interval is 1.960

 A picture containing text, clock, watch, gauge

Description automatically generated

E = 1.960 \* = 0.4214

Lower Endpoint: 53.225 - 0.4214 = 52.8036 52.8

Upper Endpoint: 53.225 + 0.4214 = 53.6464 53.6

I am 95% confident that the true mean PM 2.5 fine particulate matter pollutant level found withing south Korea is between 52.8 µg/m3 and 53.6 µg/m3

52.8 < µ < 53.6

**Module 3 research question:**

As a result of the pandemic lockdown, it is believed that South Korea reduced its concentration level in fine particulate matter (PM2.5) in the year of 2021 compared to 2020. With the mean and standard deviation of the fine particulate matter (PM2.5) set as 62.53 and 30.16 respectively, is there sufficient evidence at the 0.01 level that this claim is true considering a sample of 6555 pollutant measurements around the country in the year of 2021 produced a sample mean of 62.135 µg/m3

**Result:**

H0: µ ≥ 63 µg/m3

HA: µ < 63 µg/m3

Graphical user interface

Description automatically generated with medium confidence z = -1.06

As a one-tailed test we will reject H0 if z is less than the critical value of the z test statistic at 0.01 = -2.33

Since z = -1.06 is not less than -2.33, **we will fail to reject the null hypothesis**

**Thus, we cannot be 99% sure that South Korea reduced its average PM 2.5 concentration levels from the year 2020 in the year of 2021.**

**Module 4 research question:**

Using the dataset, we are trying to compare the average pm2.5 concentration levels in the districts of Seoul and Gangwon. Using a sample of 5771 measurements, it was found that conc. levels in the Seoul district have an average of 66.14 µg/m3 and a standard deviation of 35.21 µg/m3. On the other hand, using a sample of 3473 measurements, it was found that conc. levels in the Gangwon district have a mean of 27.44 µg/m3 and a standard deviation of 33.95 µg/m3. Can it be concluded at the 0.01 level of significance that the average pm 2.5 measurement for Gangwon is significantly lower than the average PM2.5 measurement for the district of Seoul since it’s a less populated and metropolitan district region? Seoul will be µ1 while Gangwon will be µ2.

**Result:**

H0: µ1 ≤ µ2

HA: µ1 > µ2

A picture containing text, clock

Description automatically generated

Z = 52.34

We will reject H0 if z is greater than the critical value of the z test statistic at 0.01 = 2.33

Since z = 52.34 is greater than than 2.33, **we will reject the null hypothesis**

**Therefore, we can be 99% sure that the average PM2.5 measurement for the province district of Gangwon is significantly lower than the average PM2.5 measurement for the more populated metropolitan district of Seoul.**

**Module 5 research question:**

Using a regression analysis of the data with pm25 as the dependent variable and o3, no2, co, so2 as the independent variables, is there a significant relationship between these pollutants and the fine particulate matter pm 2.5?

**Result:**



Analysis of the results displayed show that the linear relationship between the dependent variable pm2.5 and the independent variables o3, no2, so2 & co is not strong at all. Compounding that with what is presented by the Adjusted R-Square, it must be very weak. Only 0.082 of the dependent variation is explained by the regression, the coefficient of determination. The p-value for the regression equation is 0 meaning that a statistically significant linear relationship exists between the independent and dependent variables at any level of significance.

The regression results in the following equation:

Ŷ = 34.678 + 0.063(o3) + 0.938(no2) – 0.444(so2) + 0.891(co)

The p values of the variables are low enough to be statistically significant at any level. Therefore, though the regression seems to be overfitted, there is a significant relationship between pm25 and the other four pollutants.