جامعة مصر للمعلوماتية
EGYPT UNIVERSITY
OF INFORMATICS

Egypt University of Informatics
Faculty of Computing & Information Sciences
Data Analysis Course

Analysis of Nitrogen Dioxide Levels during the COVID-19 Lockdown

Cairo, Egypt vs Lisbon, Portugal

Submitted by:

Ahmed Yasser Hassanein (21-101006)

Omar Mohamed Elabasery (21-101173)

Roba Mahmoud Zenhoum (21-101074)

Abdulrahman Khaled Saleh (21-101079)

Supervised by:

Dr. Mohamed Taher ElRefaei PhD

Eng. Nadine ElSaeed

Eng. Mira Raheem

20/05/2023

Introduction

Knowing whether the Coronavirus disease, known as Covid-19, has an effect on the Nitrogen Dioxide levels in Cairo, Egypt does not seem as an interesting topic at all, but it really is! Nitrogen Dioxide is a chemical compound with the formula NO_2. Elevated levels of Nitrogen Dioxide can cause damage to the human respiratory tract and increase a person's vulnerability to, and the severity of, respiratory infections and asthma. Long-term exposure to high levels of Nitrogen Dioxide can cause chronic lung disease. So, how is the Covid-19 pandemic related to the Nitrogen Dioxide levels. The dataset, thoroughly explained in the coming sections, suggests that the Covid-19 lockdown is responsible for decreasing the Nitrogen Dioxide levels which makes the air quality much better which, in return, relieves you from the damage mentioned above. How? Unfortunately, the reasons for the better air quality are not mentioned, but according to sources, mentioned in the “References” section, human daily life events like driving your vehicle are one of the main reasons of us living in an environment with poor air quality.

Project 2 Enhancement

Since I failed to prove my Null Hypothesis (H_0) in Project 2, I decided that merging the Cairo dataset with another dataset may enhance data and analysis, and if I fail to reject the (H_0) again, then the Covid-19 lockdown is unlikely to have an effect on the air quality and the $N0_2$ levels. I researched the countries that were strict with the Covid-19 lockdown protocol. How can strictness be measured? “To measure the stringency for instance, they (authors of the tracker) took into account nine indices: from school and workplace closings, transportation, restrictions on travelling and gatherings to stay at home requirements and public information campaigns.” "If people are allowed to gather in groups of 50, that gets one rating, if they can only gather in groups of 5, that is a more intense rating," explained co-author of the tracker, Toby Phillips.

How Can Cairo’s Dataset be Enhanced?

No, Cairo’s dataset will not be enhanced, but it will be merged with another dataset of country that is part of the European Union (EU). Surprisingly, Portugal took the lead over the EU with a 66.2/100 score of applying a strict response policy. Fortunately, I found a data pool that had datasets of most of the countries in Europe with their cities in a separate dataset. I chose Lisbon, Portugal as my dataset because it is the capital of Portugal, and Cairo is the capital of Egypt, so it made sense to compare the two capitals. I hypothesized that if Lisbon, Portugal had lower levels of NO_2 after the Covid-19 lockdown, then it is possible that the strict response policy was one of the reasons of better air quality, and that Cairo may have implemented a less strict response policy which may have been one of the reasons of my failing to reject H_0.

Research Question

Does the Covid-19 lockdown have an effect on the air quality by reducing Nitrogen Dioxide levels?

Hypothesis

$\boldsymbol{\mu_1}$: maximum NO_2 concentration in April 2019

$\boldsymbol{\mu_2}$: maximum NO_2 concentration in April 2020

Null hypothesis (H_0): $\mu_1 = \mu_2$ (the two population means are equal)

There is **<u>no difference</u>** between the maximum NO2 concentration, before and after the Covid-19 Pandemic, in Lisbon, Portugal. The Covid-19 Pandemic has no effect on the maximum NO2 concentration in Lisbon, Portugal.

Alternative hypothesis (H_a): $\boldsymbol{\mu_1 \neq \mu_2}$ (the two population means are not equal)

There is a **<u>right-tailed difference</u>**, $\mu_1 > \mu_2$, between the maximum NO2 concentration, before and after the Covid-19 Pandemic, in Lisbon, Portugal. The maximum NO2 concentration is higher in April 2019 than in 2020 due to the Covid-19 pandemic which led to lockdown/quarantine all over Portugal.

Sampling Method:

Unfortunately, there is few information on the dataset, so we had to take a smart guess on the Sampling Method. Data may have been collected using the Convenience Sampling Method because the data is focused only on the month of April in 2019 and 2020 in Cairo, Egypt. The Covid-19 lockdown took time during the middle days of the month of March, 2019 which gives us more reason to believe why April was the chosen month because most people were scared of the pandemic, much caring about their health and loved ones, and strictly following the Covid-19 Protocol. So, this gives us the more reason to test if the Covid-19 has an effect on the air quality as there were less people on the streets.

Regarding the Lisbon, Portugal dataset, it was rich in data, and there were many options to customize the dataset from. To elaborate, the data pool gave me the flexibility to choose the country, city, air pollutant (CO_2, NO_2, CO, O_3, etc.), year from, year to, time coverage, and data source. I had several ways to best utilize the data, and I did!

Number of Samples Collected: 30 (Cairo, Egypt)

Number of Samples Collected: About 4800 (Lisbon, Portugal)

Bias Identification:

- Choosing only the month of April in 2019 and 2020
- Analysing the air quality based on the NO_2 levels only (Carbon Monoxide, Nitrogen Monoxide, Ozone, and many other gases contribute to the poor air quality levels we live in)
- Very small dataset
- Targeting air quality in Cairo, Egypt (It can be affected by other Egyptian governorates)

Collected Data/Dataset:

Cairo:

The Spatio-Temporal, belonging to both space and time, dataset of NO_2 levels during the Covid-19 lockdown in Cairo, Egypt. This dataset was created to monitor the changes in NO_2 concentration due to the lockdown measures of the Covid-19 pandemic in Cairo, Egypt. The daily data was collected from the Sentinel 5P, an Earth observation satellite developed by the European Space Agency, remote sensing platform for April 2019 and 2020. The daily data were processed using MATLAB and ARC/GIS, then the monthly means of April 2019 and 2020 were created and used to calculate the changes in NO_2 concentration.

Dataset Columns:

- Date: The date of the day the data was processed.
- Min: The minimum computed NO_2 concentration value of the corresponding date.
- Max: The maximum computed NO_2 concentration value of the corresponding date.
- Mean: The mean of all computed NO_2 concentration values during the day of the corresponding date.
- SD: The standard deviation of all computed NO_2 concentration values during the day of the corresponding date.

Lisbon:

The data pool has data that come from two dataflows: "E1a and E2a. The E1a data are reported to EEA by member states every September and covers the year before the delivery. This means that data delivered in September 2017 covers 2016. EEA also receives up-to-date (E2a) data on hourly basis from most of its member states. Because E1a data are validated and considered an official delivery, all E2a data are deleted before E1a data are imported. This is to ensure that no E2a data are mixed with E1a data."
"E2a (UTD) files for the current year are updated once a day. The update job starts at 01:00 AM and normally finishes around 05:30 AM).
The most recent year of E1a (historic) files are updated on monthly basis, whereas previous years of E1a only updates on request. However, at the end of each yearly reporting cycle (end of December) we run the export job for the most recent years. This is to ensure that if some

countries have redelivered historic data, e.g. for 2013, this will also be reflected in the download service."

Dataset Columns:

CountryCode: Country iso code

Namespace: Unique namespace as provided by the country

AirQualityNetwork: Network identifier

AirQualityStation: Local ID of the station

AirQualityStationEoICode: Unique station identifier as used in the past AirBase system

Samplingpoint: Local ID of the samplingpoint

SamplingProcess: Local ID of the samplingprocess

Sample: Local ID of the sample (also known as the feature of interest)

AirPollutant: Short name of pollutant.

AirPollutantCode: Reference (URL) to the definition of the pollutant in data dictionary

AveragingTime: Defines the time for which the measure has been taken (hour, day, etc)

Concentration: The measured value/concentration

UnitOfMeasurement: Defines the unit of the concentration

DateTimeBegin: Defines the start time (yyyy-mm-dd hh:mm:ss Z) of the measurement

DateTimeEnd: Defines the end time (yyyy-mm-dd hh:mm:ss Z) of the measurement

Validity: The validity flag for the measurement

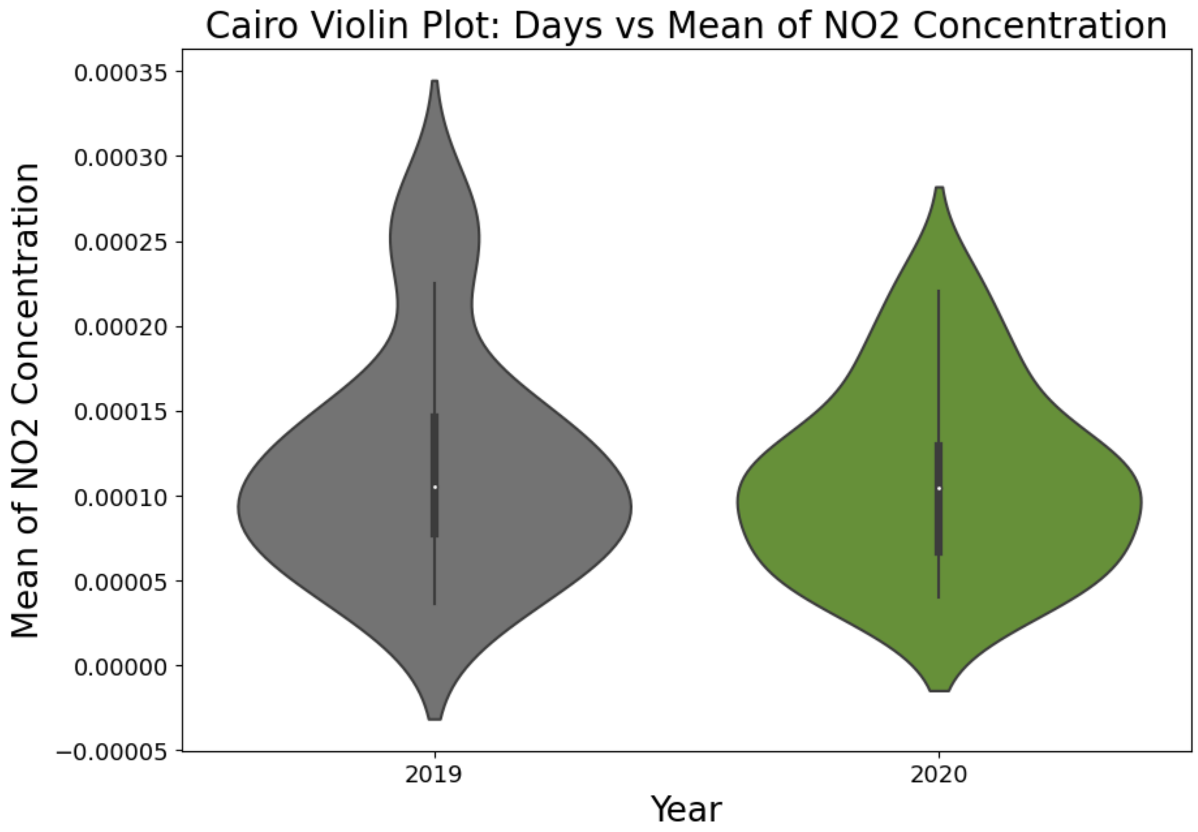
Verification: The verification flag for the measurement

Analysis:

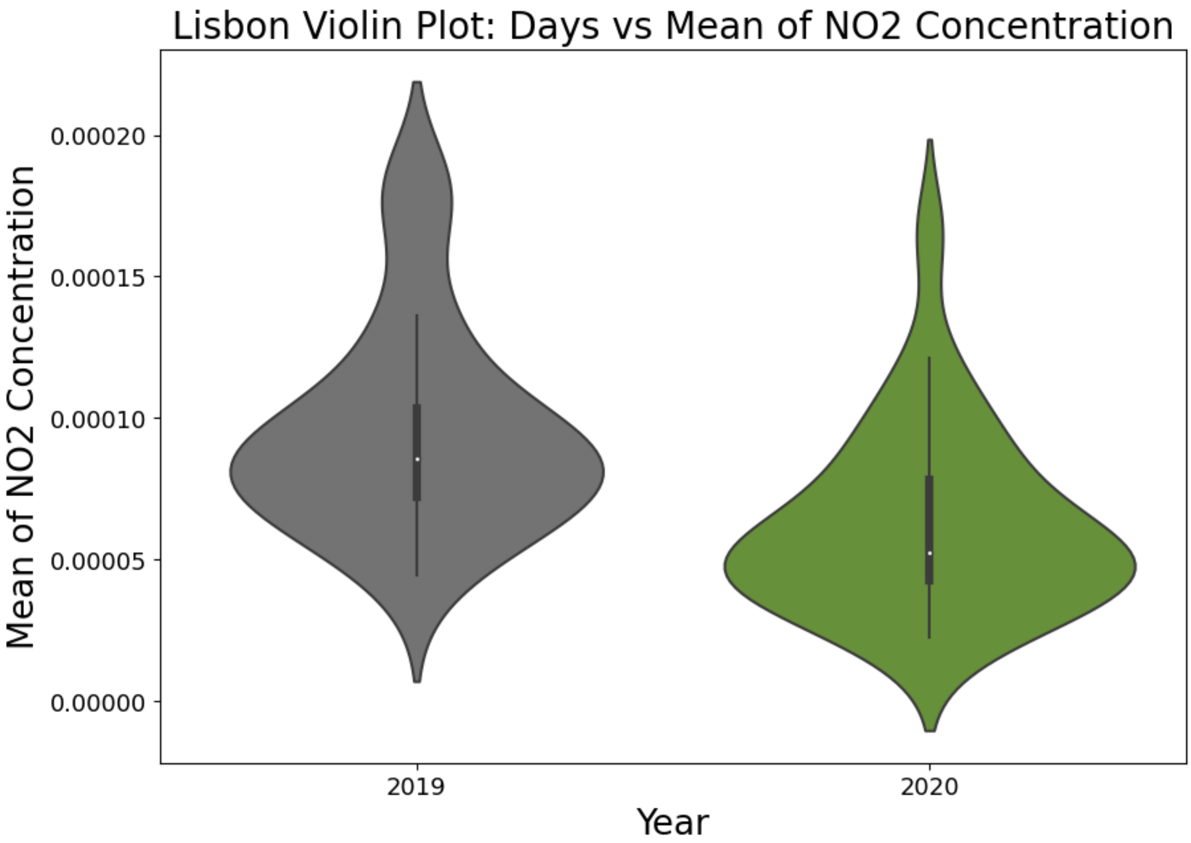
The cleaning stage, found in the notebook file, took a lot of effort because I had to transform the original Lisbon dataset to match the month, days, and years of the Cairo dataset to be able to conduct my analysis and test my hypothesis.

After cleaning the dataset, the next step is both visualizing and analysing the dataset to find patterns to support the hypothesis we came up with at the beginning of the report.

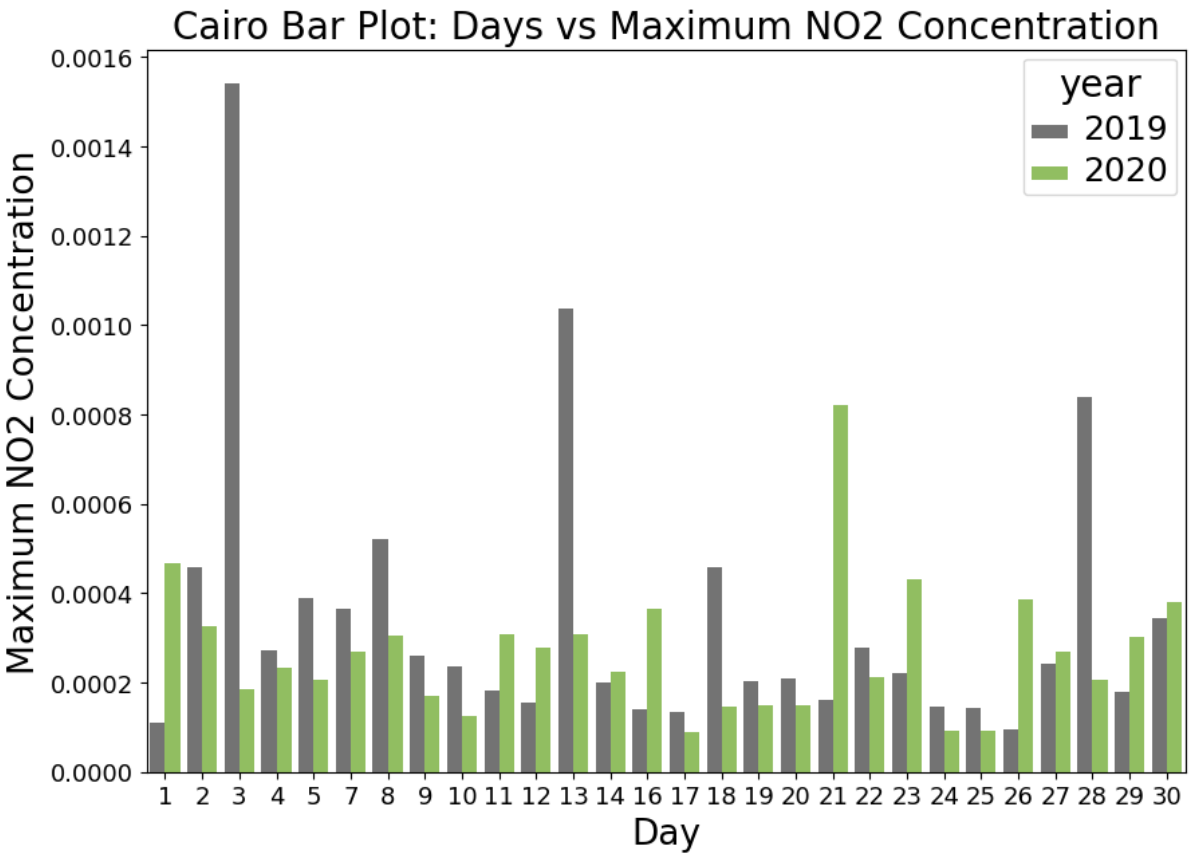
This section will cover all visualizations used in the Jupyter Notebook, platform used to analyse the dataset, that will help us answer the research question and conduct the hypothesis testing to either reject or fail to reject H_0.



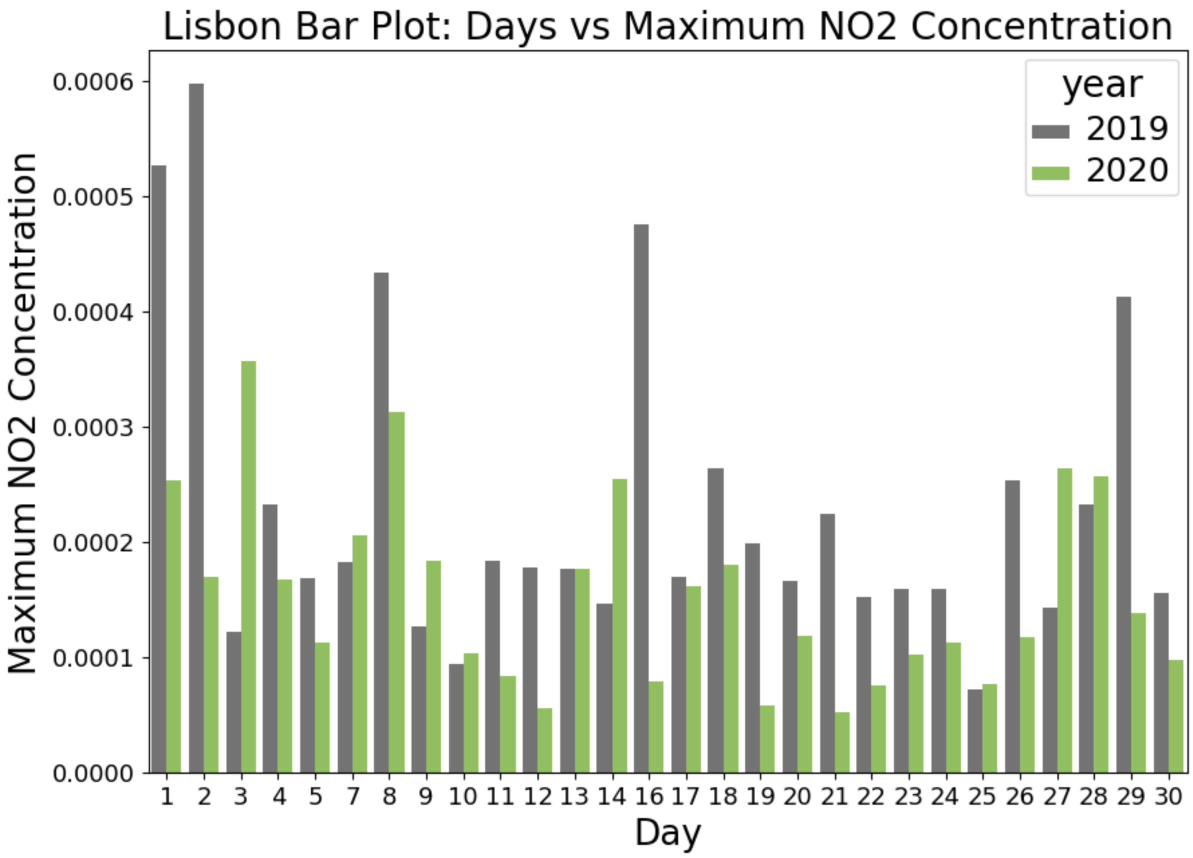
Violin Plot showing the mean of Nitrogen Dioxide for every day in the month of April in Cairo in 2019 and 2020. The Violin Plot shows most of the calculated data is between the 0.00005-0.00015 range for both years. It is shown that 2019 has quite higher means, but they will be later visualized by other plots to be classified as outliers or not. The Violin Plot shows that the "Mean" may not be the best variable used to formulate the hypothesis upon.



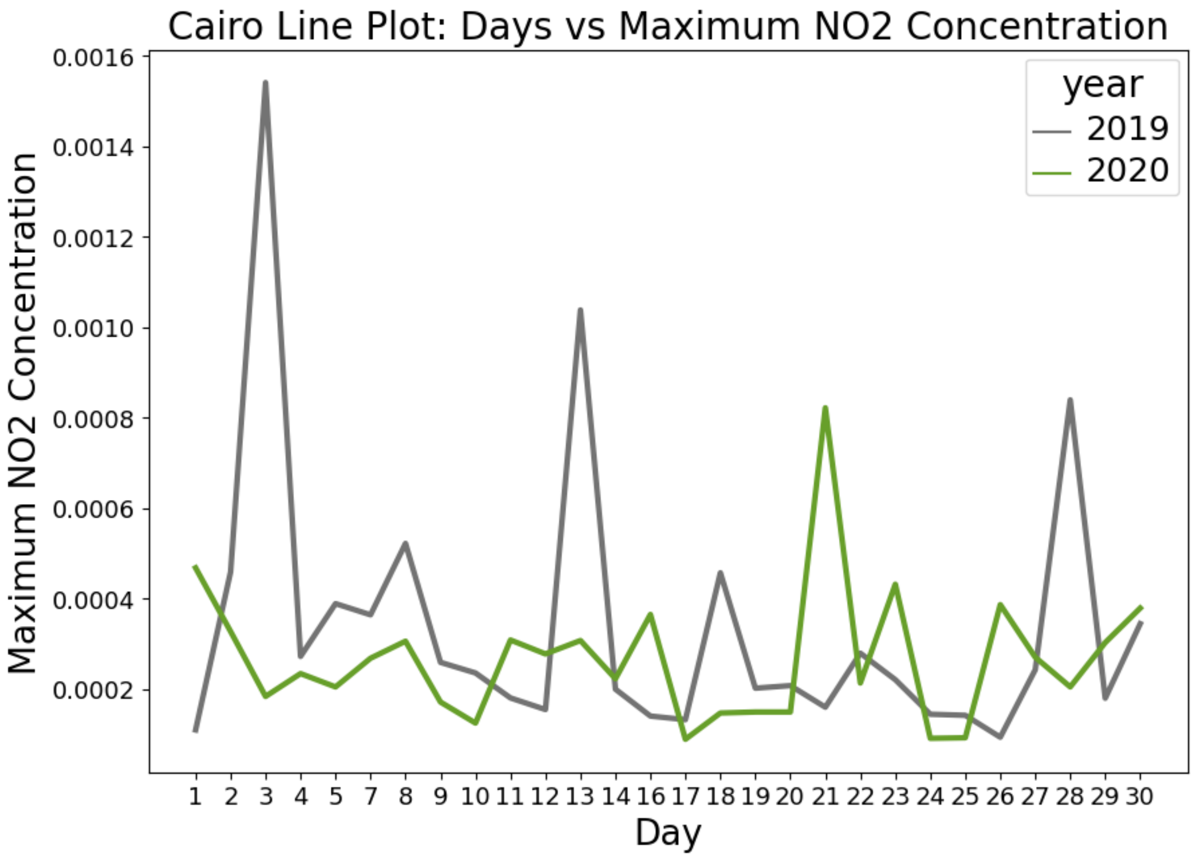
Violin Plot showing the mean of Nitrogen Dioxide for every day in the month of April in Lisbon in 2019 and 2020. The Violin Plot shows most of the calculated data is between the 0.00005-0.00010 range for both years. It is shown that 2019 has quite higher means, but they will be later visualized by other plots to be classified as outliers or not. The Violin Plot shows that the "Mean" may not be the best variable used to formulate the hypothesis upon.



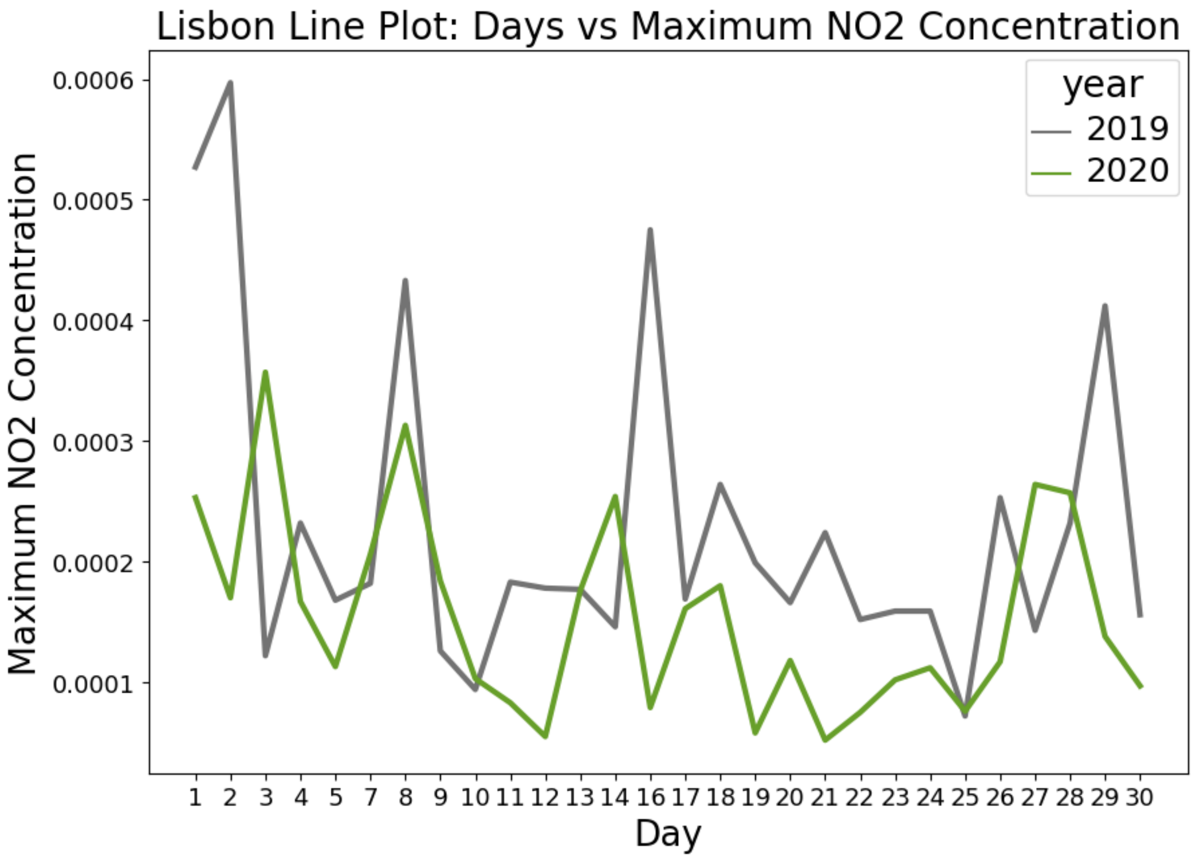
Bar Plot showing the maximum computed value of Nitrogen Dioxide for each day in the month of April in Cairo in 2019 and 2020. The bar plot shows that the values in both years are close for most of the day with the 2019 data having a higher edge on 2020's. On notable few days of April, "3", "13", and "28", 2019 shows that the level of Nitrogen Dioxide is much higher and concentrated.



Bar Plot showing the maximum computed value of Nitrogen Dioxide for each day in the month of April in Lisbon in 2019 and 2020. The bar plot shows that the values in 2019 are considerably higher than 2020's. On notable few days of April, "1", "2", "16", and "29", 2019 shows that the level of Nitrogen Dioxide is much higher and concentrated.



Line Plot showing the maximum computed value of Nitrogen Dioxide for each day in the month of April in Cairo in 2019 and 2020. The Line Plot shows that the maximum NO2 concentration in 2019 is much higher than the value in 2020 which is a motive to construct a hypothesis and conduct hypothesis testing on the two following categories, **maximum NO2 concentration values in 2019** and **maximum NO2 concentration values in 2020.**



Line Plot showing the maximum computed value of Nitrogen Dioxide for each day in the month of April in Lisbon in 2019 and 2020. The Line Plot shows that the maximum NO2 concentration in 2019 is much higher than the value in 2020 which is a motive to construct a hypothesis and conduct hypothesis testing on the two following categories, **maximum NO2 concentration values in 2019** and **maximum NO2 concentration values in 2020.**

Hypothesis Testing Steps for Cairo

- Step 1: Define null and alternative hypothesis

  μ_1: maximum NO_2 concentration in April 2019

  μ_2: maximum NO_2 concentration in April 2020

  Null hypothesis (H_0): $\mu_1 = \mu_2$ (the two population means are equal)

  There is no difference between the maximum NO2 concentration, before and after the Covid-19 Pandemic, in Cairo, Egypt. The Covid-19 Pandemic has no effect on the maximum NO2 concentration in Cairo, Egypt.

  Alternative hypothesis (H_a): $\mu_1 \neq \mu_2$ (the two population means are not equal)

  There is a right-tailed difference, $\mu_1 > \mu_2$, between the maximum NO_2 concentration, before and after the Covid-19 Pandemic, in Cairo, Egypt. The maximum NO_2 concentration is higher in April 2019 than in 2020 due to the Covid-19 Pandemic which led to lockdown all over Egypt.

- Step 2: Choose the appropriate test

  The appropriate test to conduct the hypothesis testing is the paired samples T-test. A paired samples T-test is used to compare the means of two samples when each observation in one sample can be paired with an observation in the other sample.

  ### Paired Samples T-test: Motivation

  A paired samples T-test is commonly used in two scenarios:

  - A measurement is taken on a subject before and after some treatment – e.g., the maximum NO2 concentration is measured in April 2019 and April 2020, and the treatment, between both periods, is the Covid-19 pandemic.
  - A measurement is taken under two different conditions – e.g., the maximum NO2 concentration is measured in April 2019 and April 2020, and the condition, between both periods, is the Covid-19 pandemic.

  In both cases we are interested in comparing the mean measurement between two groups in which each observation in one sample can be paired with an observation in the other sample.

- Step 3: Calculate the p-value

  p-value: [0.28502196]

  All calculations are found in the Jupyter Notebook that is attached to this report.

- Step 4: Determine the statistical significance

  The commonly used significance level threshold is 0.05. Since p-value here (0.285) is greater than 0.05, we can say that it is statistically insignificant based on the collected sample. Thus, we fail to reject the null hypothesis (H_0). In Laymen's terms, this usually means that we do not have statistical evidence that the difference in groups is not due to chance. Reasons for failing to reject the null hypothesis (H_0) are small sample size, bias in data, weak effect size between the two sample sets.

Hypothesis Testing Steps for Lisbon

- Step 1: Define null and alternative hypothesis

  μ_1: maximum NO_2 concentration in April 2019

  μ_2: maximum NO_2 concentration in April 2020

  Null hypothesis (H_0): $\mu_1 = \mu_2$ (the two population means are equal)

  There is **no difference** between the maximum NO2 concentration, before and after the Covid-19 Pandemic, in Lisbon, Portugal. The Covid-19 Pandemic has no effect on the maximum NO2 concentration in Lisbon, Portugal.

  Alternative hypothesis (H_a): $\mu_1 \neq \mu_2$ (the two population means are not equal)

  There is a **right-tailed difference**, μ1 > μ2, between the maximum NO2 concentration, before and after the Covid-19 Pandemic, in Lisbon, Portugal. The maximum NO2 concentration is higher in April 2019 than in 2020 due to the Covid-19 pandemic which led to lockdown/quarantine all over Portugal.

- Step 2: Choose the appropriate test

  The appropriate test to conduct the hypothesis testing is the paired samples T-test. A paired samples T-test is used to compare the means of two samples when each observation in one sample can be paired with an observation in the other sample.

  ### Paired Samples T-test: Motivation

  A paired samples T-test is commonly used in two scenarios:

  - A measurement is taken on a subject before and after some treatment – e.g., the maximum NO2 concentration is measured in April 2019 and April 2020, and the treatment, between both periods, is the Covid-19 pandemic.
  - A measurement is taken under two different conditions – e.g., the maximum NO2 concentration is measured in April 2019 and April 2020, and the condition, between both periods, is the Covid-19 pandemic.

  In both cases we are interested in comparing the mean measurement between two groups in which each observation in one sample can be paired with an observation in the other sample.

- Step 3: Calculate the p-value

  p-value: [0.01064432]

  All calculations are found in the Jupyter Notebook that is attached to this report.

- Step 4: Determine the statistical significance

  The commonly used significance level threshold is 0.05. Since the p-value here (0.010) is less than 0.05, I can reject the Null Hypothesis (H0). Thus, there is evidence that the Alternative Hypothesis (Ha) holds. In conclusion, it is safe to say that the maximum NO2 concentration is higher in April 2019 than in April 2020 which means that the Covid-19 Pandemic's Lockdown has had a positive effect on the NO2 levels in Lisbon, Portugal where the NO2 levels have decreased in 2020 which was about one year since the Covid-19 Pandemic's Lockdown took place.

Conclusion

At the 5% significance level, there is sufficient evidence to support the claim that there is a right-tailed difference, $\mu_1 > \mu_2$, between the maximum NO_2 concentration, before and after the Covid-19 Pandemic, in Lisbon, Portugal. Since the p-value here (0.010) is less than 0.05, I can reject the Null Hypothesis (H_0). In conclusion, it is safe to say that the maximum NO_2 concentration is higher in April 2019 than in April 2020 which means that the Covid-19 Pandemic's Lockdown has had a positive effect on the NO_2 levels in Lisbon, Portugal where the NO_2 levels have decreased in 2020 which was about one year since the Covid-19 Pandemic's Lockdown took place.

Cairo, Egypt vs Lisbon, Portugal

The main question is:

Why is the same H_0 chosen was rejected in Lisbon, Portugal and failed to get rejected in Cairo, Egypt?

Please note that both cities' datasets contained the same sample numbers and columns, and each produced a different result.

There are many reasons for failing to reject the H_0 in Cairo, Egypt.

- Cairo did not apply a strict response policy to the Covid-19 Pandemic like Lisbon did.
- The NO_2 levels in Cairo in 2019 were higher than the levels in Lisbon in 2019.
- Lisbon, Portugal belongs to the European Union which shows that there are severe consequences if certain measures were not taken. On the other hand, Cairo, Egypt does not belong to any entity or alliance and takes decisions that it sees fit.
- The reported number of individuals affected by the Covid-19 Pandemic in Portugal was higher than in Egypt which may be a reason for applying a strict response policy in the first place.

Any potential issues

- The Lisbon dataset contained a surplus of data, and about 90% of it had to be dropped for it to match Cairo's dataset which was very small in size.
- The Cairo dataset limited what I could have analysed with the dataset as a whole, and there were other countries that calculated other air pollutants which could have been utilised a whole lot better.
- The weak effect size between the data obtained in April 2019 and April 2020 is a contributing factor to failing to reject the null hypothesis. A large effect size means that research finding has practical significance, while a small effect size indicates limited practical applications. In other words, the larger the effect size, the more important the effect.
- There are other potential issues faced that are mentioned in the "Bias Identification" section.

References

A. AbdelSattar, S. Swidar, and D. Abdelgawad, “Spatio-temporal dataset of nitrogen dioxide levels during COVID-19 lockdown in Cairo, Egypt,” *data.mendeley.com*, vol. 1, Aug. 2020, doi: https://doi.org/10.17632/8cp52kynh6.1.

G. M. Sullivan and R. Feinn, “Using effect size—or why the P value is not enough,” *Journal of Graduate Medical Education*, vol. 4, no. 3, pp. 279–282, Sep. 2012, Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/

“COVID-19: which European country has the strictest response policy?,” *euronews*, Aug. 31, 2020. https://www.euronews.com/2020/08/31/covid-19-which-european-country-has-the-strictest-response-policy#:~:text=Kosovo%20(69.44%2F100)%20appears (accessed May 20, 2023).

“Failing to Reject the Null Hypothesis,” *Statistics By Jim*, Mar. 03, 2020. https://statisticsbyjim.com/hypothesis-testing/failing-reject-null-hypothesis/#:~:text=Failing%20to%20reject%20the%20null%20indicates%20that%20our%20sample%20did

D. Wright, “How to State the Conclusion about a Hypothesis Test,” *Dawn Wright, Ph.D.*, Nov. 21, 2018. https://www.drdawnwright.com/how-to-state-the-conclusion-about-a-hypothesis-test/

Europa.eu, 2019. https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm (accessed May 20, 2023).

I. Chiang, “Understanding Null Hypothesis Testing – Research Methods in Psychology,” *Opentextbc.ca*, Oct. 13, 2015. https://opentextbc.ca/researchmethods/chapter/understanding-null-hypothesis-testing/