Mufei.Xu 12992297
Mu Tian 13009535

Introduction

The dataset we choose is the air pollution data in United States from year 2000 to 2016, the dataset includes the name of the city, the state it belongs to, and the data of four major air pollutions including NO2, O3, SO2 and CO. With these data, we plan to find out the relationship between four major air pollution indicators, then predict the possible future trend of the air pollution.

This project is meaningful because against the background of industrial society, the air pollution issues i becoming the focus of attention in various countries. By doing this machine learning project, we can find how the major air pollutants are related to each other, thus helping the governments find out the most efficient way in reducing the air pollution. Also, we will be able to predict the air pollution situation in near future using machine learning, thus understanding the effect of current methods on air pollution control.

In this project, we will firstly use data visualization tools in Seaborn(eg. Box plot and Correlation thermogram) to visualize the whole dataset, so that we can understand gross changes in air pollution ov the period from 2000 to 2016, and with the correlation thermogram, we will be able to find out the relationship between different air pollution indicators.

Then, since there are too many air pollution indicators in too many cities, in the prediction part we only chose the pollution that most known to the public, which is CO, in the first city that was shown in the dataset which is Phoenix. Predicting the trend of CO in Phoenix is clearly a time series issue, so we firstly use the different method to improve the stability of the CO data, then build a ARIMA model that fits our dataset. Use the Q-Q plot and DW test to create a white noise test to ensure all the useful information in the data has been extracted. Using the ARIMA model we can predict the future trend of the CO data.

Finally, we use self-predicting test to test the reliability and creditability of the predicting model we have built.

Exploration

The major challenges in this project is to learn how to deal with the time series issues, including how to do the order determination and choose the most suitable ARIMA model to deal with the data, also how to make sure all the useful information in the dataset has been extracted in to the model.

The most useful data in the whole dataset includes the name of the city(As a unique label), data local(Used to track the time flow) and the indicators of four major pollution(Mean, 1st Max Value, 1st Ma Hour). In the visualizing part, all the indicators of the pollution are involved since we want to discover as many relationships as possible. In the predicting part, we only choose the CO Mean data in Phoenix, sinc there are too many air pollution indicators in too many cities and it will be hard to list the prediction of all of them. Besides, the CO Mean data is the most appropriate way to reflect the CO emissions of a day, so the 1st Max Value and 1st Max Hour is removed in order to improve the modeling efficiency.

Since we are dealing with a typical time series issue, we choose a widely used model named ARIMA, in order to find the model that suits our project the best, we will need to use the different method improving the stability of the data, then choose the best model according to the ACF(autocorrelation function) grap and PACF(partial autocorrelation function) graph.

To test the completeness of data extraction, we will adopt Q-Q plot and DW test as a white noise test. In order to test the reliability and creditability of the predicting model, we will let the model firstly predict the

trend of CO Mean in Phoenix from 2014 to 2016 and compare it to the real trend in that period of time, se
if the predicting result is reasonably convincing.

Methodology

By choosing the CO data in the predicting part, what we should do first is sample the monthly average of CO data which would be used to refilled the blank data in order to draw a time series chart of CO value (Figure 1).

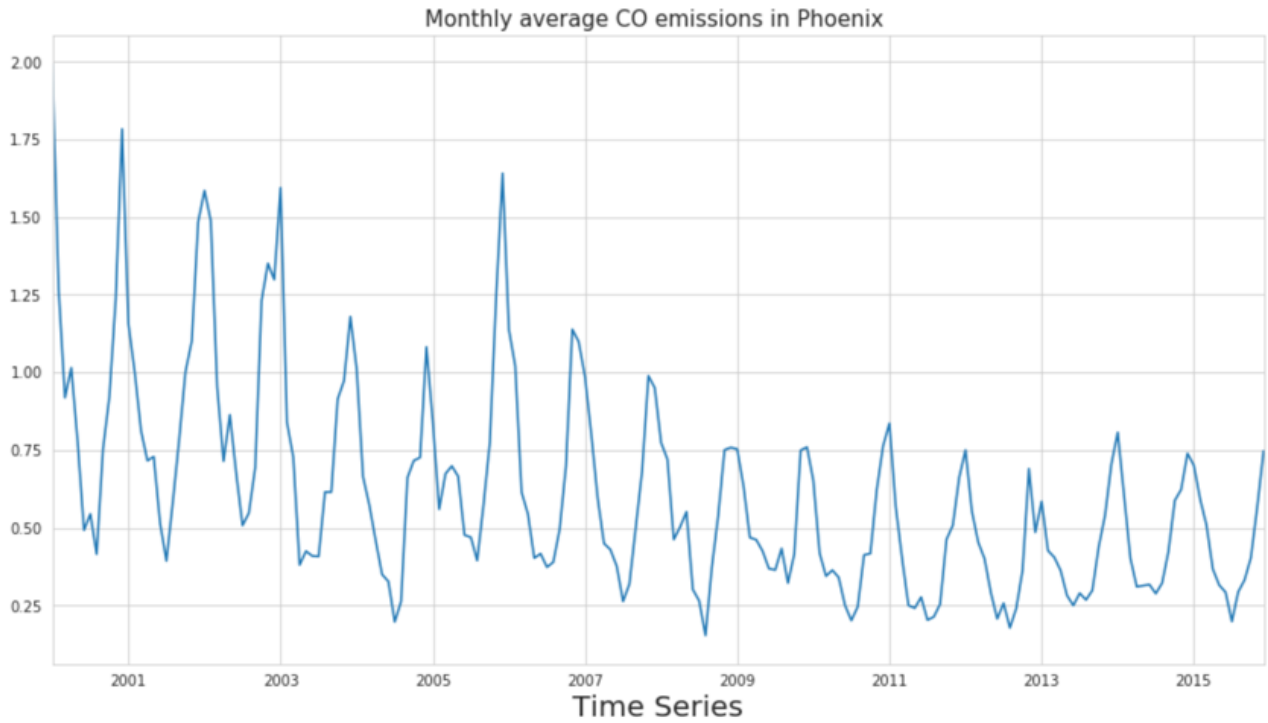


Figure 1. The monthly average value of CO in Phoenix

According to the chart, the line in the chart is too steep for further analysis. After that, we cointegration analysis the data by first-order difference in order to make the dataset to be a stationary series (Figure 2) although this step would lose some information in the dataset.

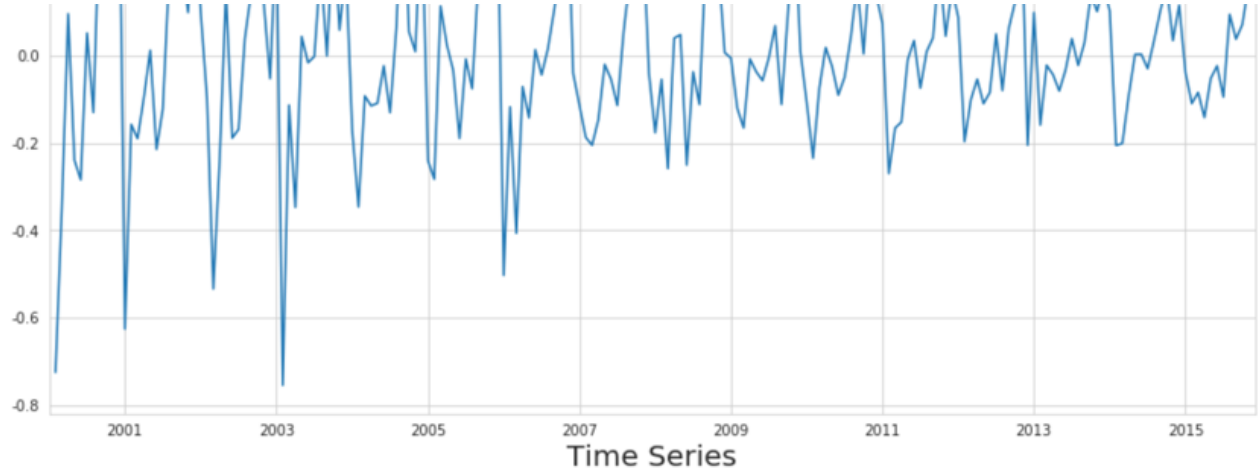


Figure 2. First difference of monthly average value of CO in Phoenix

When this chart looks stable, we use the ADF Test (augmented dickey-fuller test) to check if this is a stationary series (Figure 3).

```
(-4.2445102664096614,
 0.0005534600249798 51,
 15,
 175,
 {'1%': -3.468280364174926 7,
  '10%': -2.575652579591836 6,
  '5%': -2.878201724081632 7},
 -217.43874686904547)
```

Figure 3. ADF Test result

According to the result, the value of the ADF equals -4.245 which is remarkable lower than the three confidence coefficients related to 1%, 5% and 10%. Besides, the P-value is also lower than 0.001. According to those evidence, we consider the Test has passed, which means the series is stable.

When we get the stable series, the next stop should be getting the suitable ARIMA model. Before that, th autocorrelogram and partial autocorrelogram of stationary time series should be checked first in an attempt to confirm two parameters, p and q, in the formula:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

Figure 4. The formula of ARMA

By drawing the Autocorrelation coefficient and partial autocorrelation coefficient, the values of p and q ar confirmed to be 8 and 0.

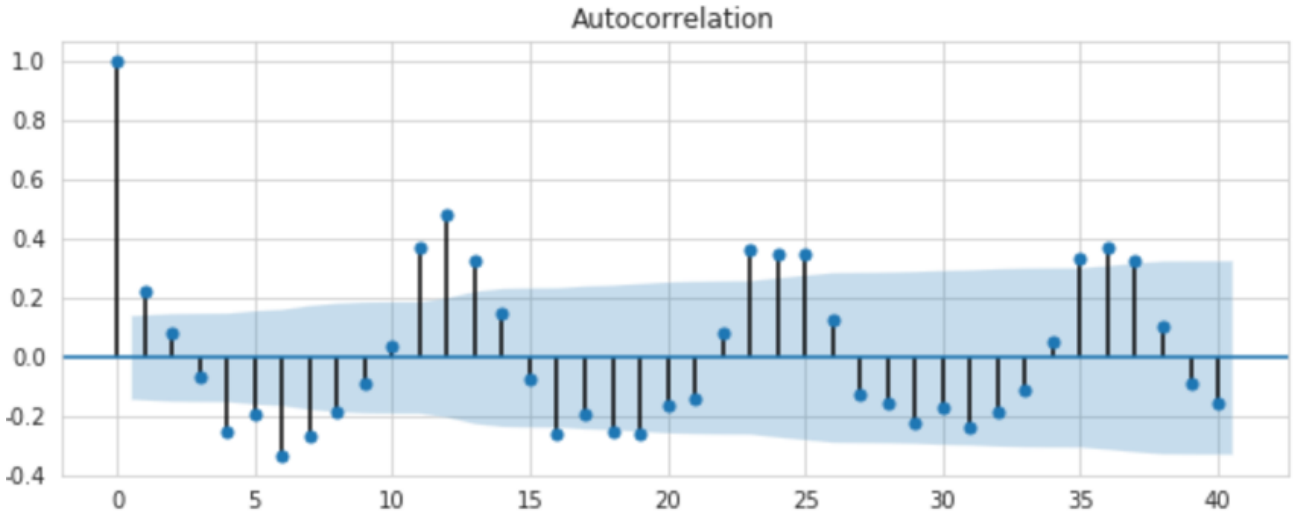


Figure 5. The chart of Autocorrelation

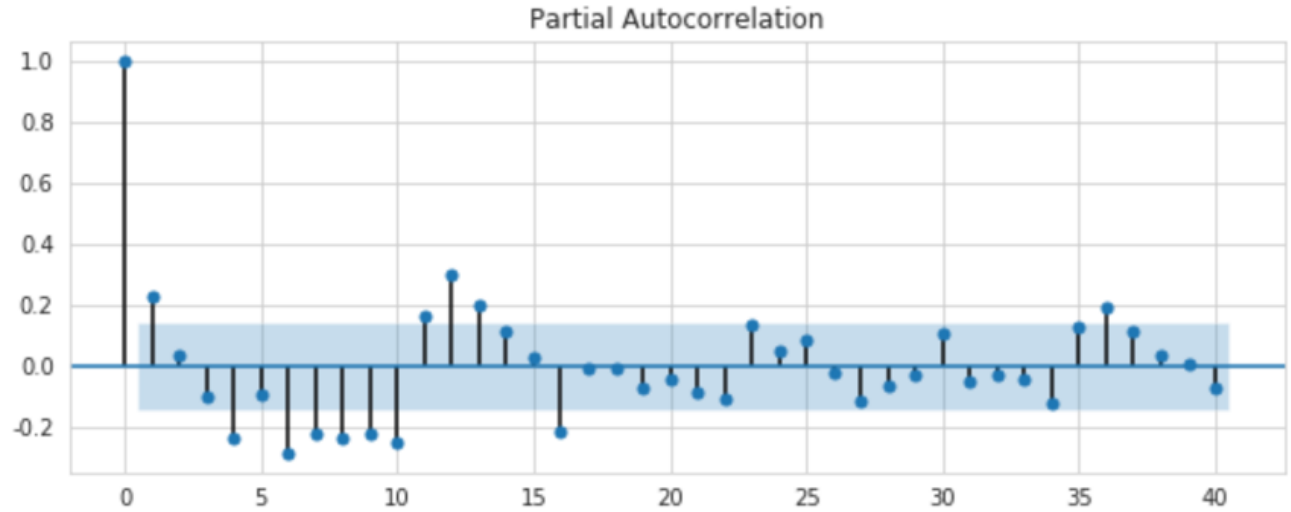
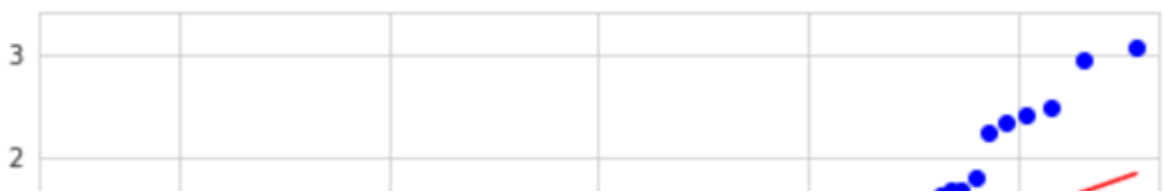


Figure 6. The chart of Partial Autocorrelation

From now on, we successfully create a model ARMA(8,0) to predict the future trend of the CO data.

Evaluation

After that, Q-Q Plot Test and DW Test are used to test if the model is suitable.



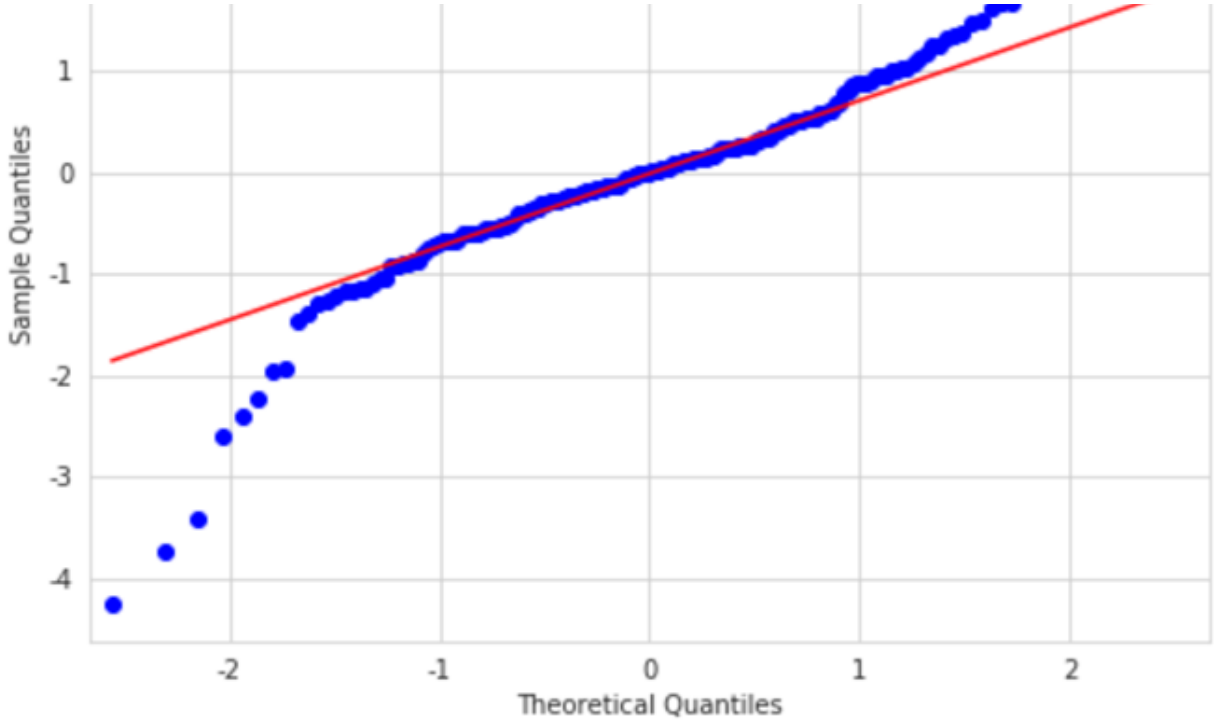


Figure 7. Q-Q Plot Test

According to the Figure 7, the residual series of the model could be considered a normal distribution, which means the residual series is a white nose series and the information of the residual has all been extracted. In such situation, the model does not need to be changed.

Another test is the DW Test, which directly use the residual value to check the autocorrelation of the tota random error terms.

```
# DW Test
print(sm.stats.durbin_watson(arma_mod20.resid.values))
```

```
2.0311876041782453
```

Figure 8. DW Test

As the relation of value of DW with parameter p showing below:

$$\begin{cases} 0, & \rho = 1 \end{cases}$$

$$DW = \begin{cases} 2, & \rho = 0 \\ 4, & \rho = -1 \end{cases}$$

Figure 9. Relationship between DW value and parameter p

According to the Figure 9, the test result is very close to 2, indicating that there is no autocorrelation. Afte testing the model, we use the model to predict trend of CO Mean in Phoenix from 2014 to 2015.

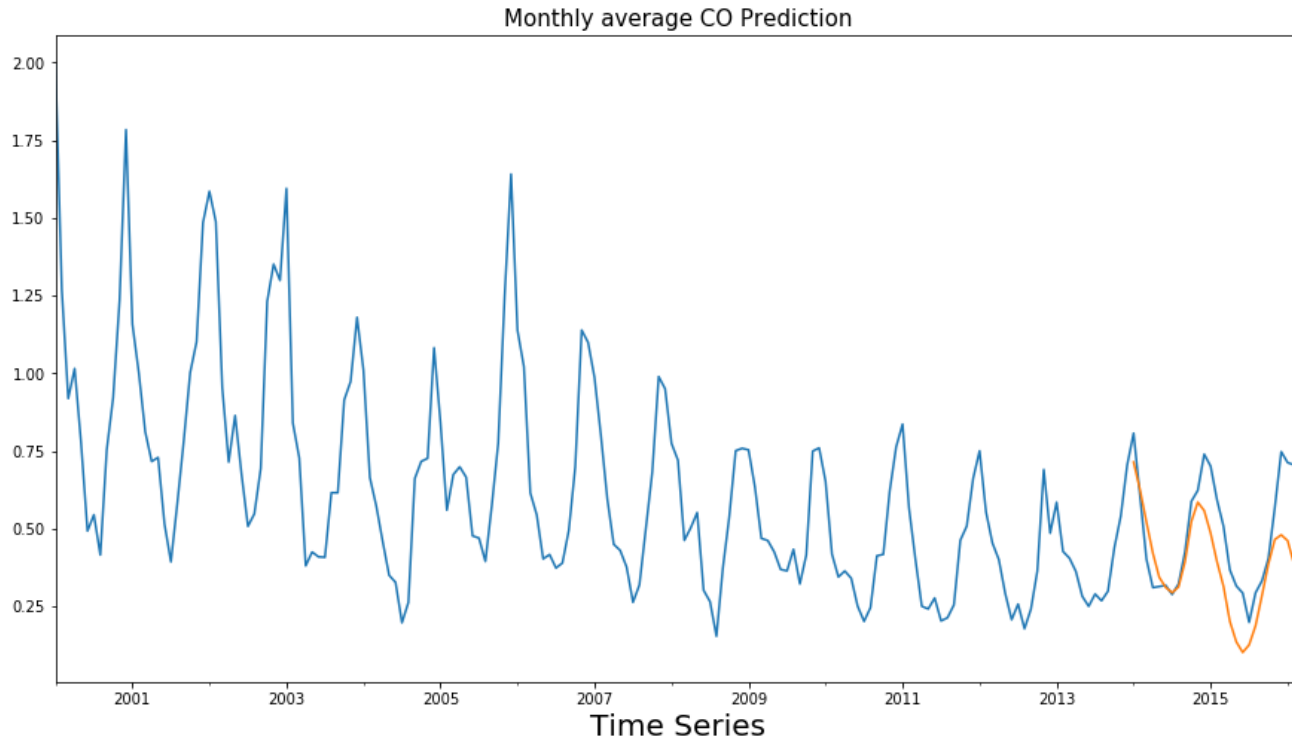
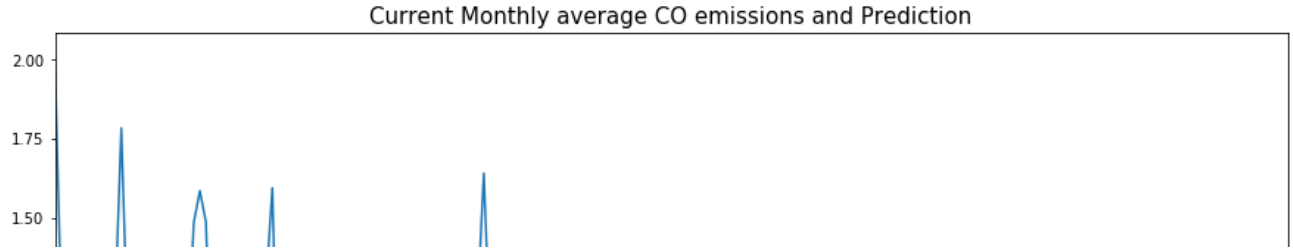


Figure 10. The prediction of CO emission

From the chart, we can clearly see that the prediction is quite reasonable.

Conclusion

We use the ARIMA model to predict the CO emission data which is related to the air pollution and the relationship between each type of air pollutants, and such information is extraordinary useful for future study.



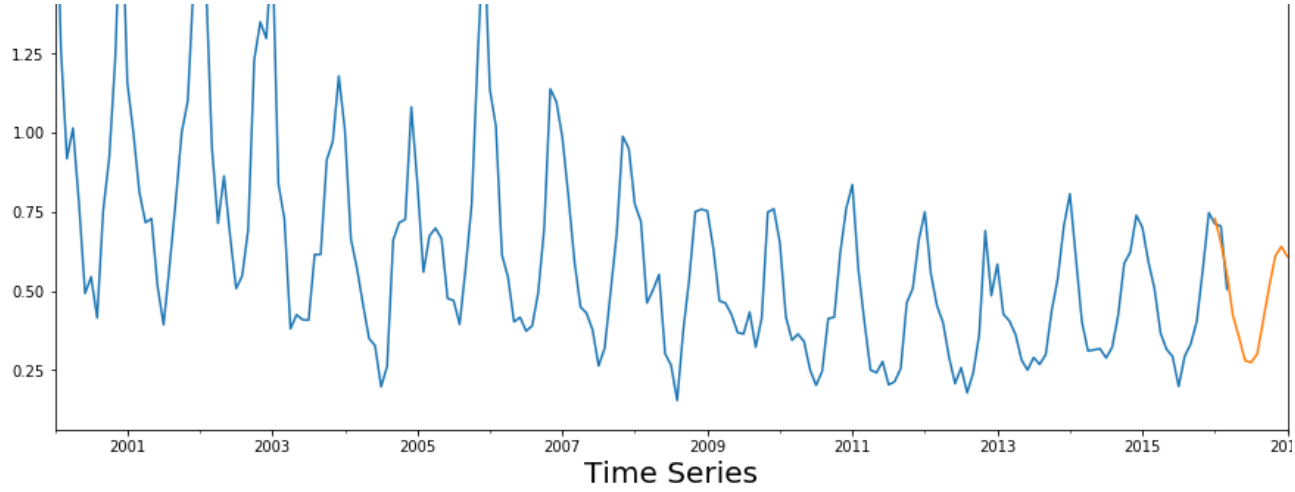


Figure 11. Prediction of Monthly average CO emission

According to Adebiyi at. El. (2014), while the pattern of ARIMA forecasting model is directional, the ANN model is toward value forecasting. Although these two models are all performing well with no significant difference, the performance of ANN model is better than ARIMA model in terms of forecasting accuracy on many occasions. For possible improvements, we can use the ANN model to analysis the data and do the prediction.

Ethical

This project has addressed a practical issue on air pollution, this project has provided progressive insigh including the relationships between each type of air pollutants, as well as the future trend of air pollution This could help the governments understand the key relationships in air environments controlling, also th prediction in this project can help relevant departments to know whether the current air pollution controlling strategy is effective.

To sum up, this project is aiming on a positive and progressive issue that can help the society to improve in air pollution controlling, thus helping us to achieve a better society. According to utilitarian approach, this project can provide valuable insights in air pollution control, and help us to improve our current air pollution controlling strategy which is definitely providing the greatest volume of benefits over harms for the majority of people, and also morally correct. According to Kantian Duty Based Ethics, the starting point is absolutely good will, regardless of the outcome of the project and its value, this project is done ir order to make us better understand the current situation of air pollution controlling, so that we can furthe optimizing the current air pollution controlling strategy.

So that this project is ethical and have positive effects to the whole society, which means there can be and should be more in-dept study on such air pollution issues.

Reference

Ayodele, A.A., Adewumi, A.O. & Charles, K.A. 2014, "Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction", *Journal of Applied Mathematics*, vol. 2014.

Link

https://colab.research.google.com/drive/1cPTeyIJhrb5CCMYHScIG7vX0L3bgj2rE