CISC 451 – Topics in Data Analytics

Final Report

Team Members: Will Murray (20054564), Emma Ritcey (20007918), Rylen Sampson (20051918)

Project Title: Analysis of Air Quality for Preventive Actions and Public Health

Problem Statement

Air quality, or a lack thereof, is emerging as a new and undoubtedly dangerous threat to public health. This threat is an even more prominent issue in developing countries where there is an access barrier to professional healthcare and in countries that are largely dependent on sectors with excessive industrial pollution. International health groups such as the World Health Organization have put tremendous work into performing broad-level analysis to estimate the effect a country's air quality may have on its citizens. WHO's analysis can provide insight into which countries require the help of data analytics the most in the fight against deteriorating air quality.

We obtained air quality data made public by OpenAQ, a non-profit organization, and implemented data science algorithms in hopes to provide actionable information for citizens living in countries with poor air quality on how to better protect themselves from this growing issue. To supplement this data we also gathered various country indicators from the World Bank, as we believe these may be useful in further explaining trends seen in the air quality data we collected.
Air quality, with the given circumstances, is a pressing issue as COVID-19 heavily affects the respiratory system. We used the most recent air quality data available as this could play a part in saving lives where COVID-19 is still out-of-control.

With the focus of data analytics heavily skewed towards developing complex and *fancy* models, there is a gap in developing real-world projects composed of explainable models with actionable results. While traditional statistical work has been historically used for decision-making, it lacks a certain degree of precision when it comes to predictive modelling and assessing future scenarios. We used a culmination of data analysis methods to advise government or policy groups on the risk associated with certain decisions and how it may influence the concentration of harmful air pollutants in their region.

Data Description

As mentioned, our main data source is OpenAQ. They have done the technical dirty work by collecting raw measurements of PM2.5 from numerous unformatted sources across the world. This data remains un-cleaned and un-analyzed, but available for those who want to look at issues surrounding air quality with a finer granularity.

OpenAQ provides air quality data for a total of 12667 locations worldwide. For the scope of this project and to keep our results actionable we focused on a vetted subset of 19 countries. To select which countries to focus on, we utilized data made available by the World Health Organization (WHO). The data hosted on the Global Health Observatory (GHO), a data repository run by WHO, contains information estimated joint effects of air pollution. This information was used to select various countries impacted in a widespread manner by air pollution. Similarly WHO, possesses PM2.5 records within the GHO, except these measurements are only given yearly and are specified for urban areas, a fact we find to be not representative which is why we opted for PM2.5 data from OpenAQ.

To elevate the problem scope, we hypothesized looking at the history of numerous country indicators would provide fruitful insights. Most of the data we were searching for was readily available in various databases hosted on the World Bank. Querying these databases for the selected countries and years was straightforward. As some of the data lies in different databases, we ended up with multiple csv files although these were small enough that it was easier to piece them together using Microsoft Excel rather

than a programming language such as Python. This obtained data was not perfect though, it included missing values for certain country, year and indicator combinations which needed to be filled. The process of how we did this is outlined in the proceeding *Methodology* section. The originally downloaded data is available in the world_bank_data folder.

Data we collected, but wasn't used, includes geospatial coordinates for each country. This was helpful in the exploratory analysis for understanding which countries were close enough together that their data may be tied together via geographical factors.

To provide a more complete understanding of the data we collected for this project, we did an extensive exploration of the data. Key observations are highlighted in the following figures.

Distance Between Countries

A piece of information we figured would be good to keep in mind throughout the project is the distance between each country. As nearby countries may be exposed to similar effects, this could help explain the difference in results for our supervised learning approaches.

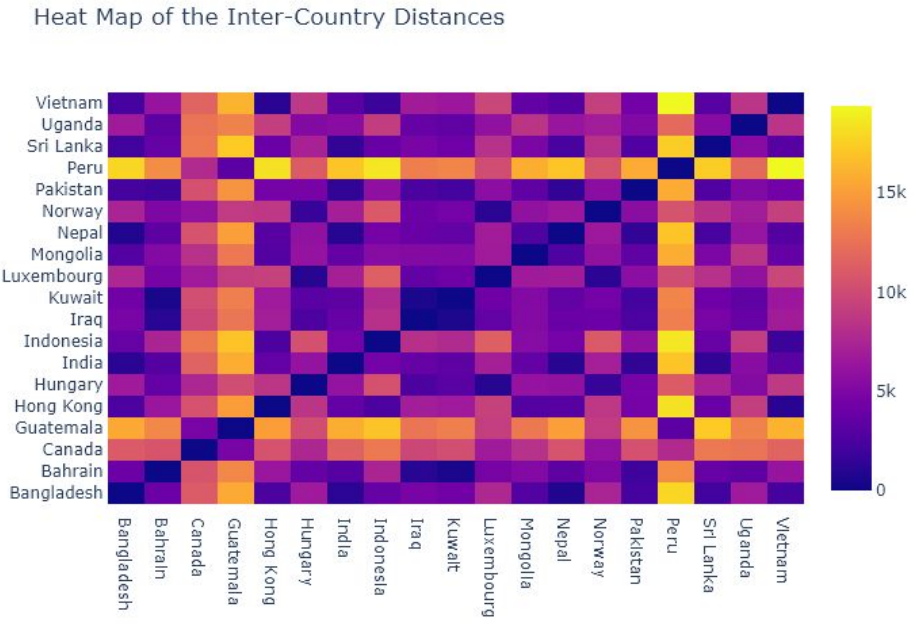

Figure 1: Heatmap of the distances between countries.

Feature Correlation Matrix

Given that we are using country indicators, it is clear that a number of our features could act as proxies for one another. A correlation matrix will help identify these features which we can then take into account for feature selection.

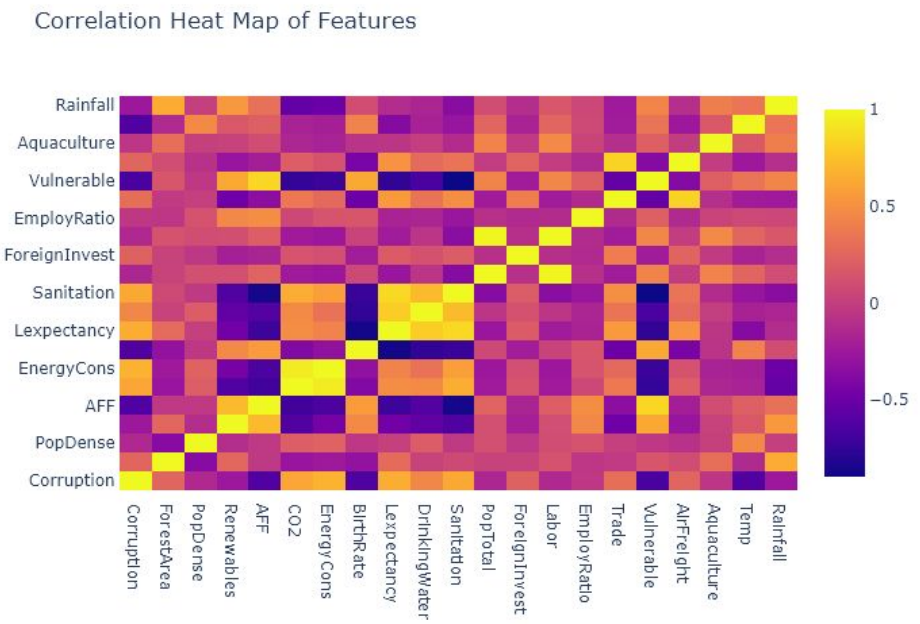

Figure 2: Correlation matrix of the collected features.

Feature Distribution

To provide more context to the correlation matrix, we looked at the distribution of all our features. Rather than share all figures we only share a couple to give context but, the remainder are in the figures folder of our submission.

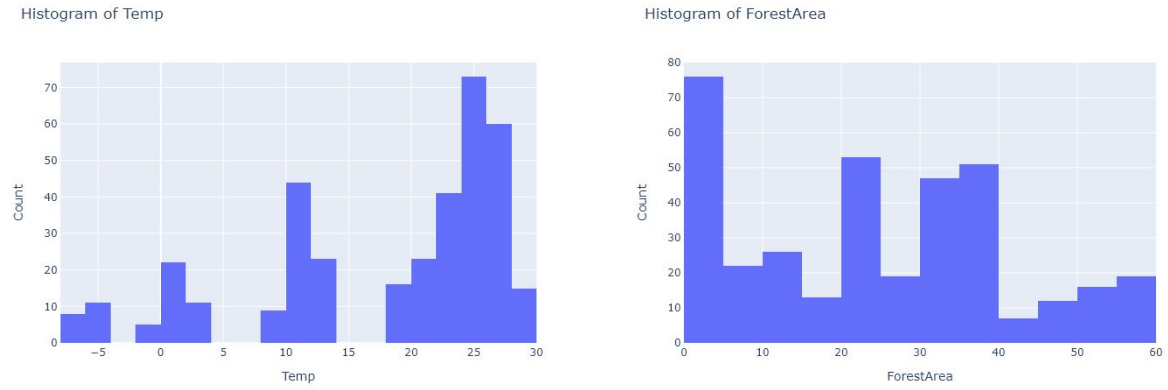

Figure 3 (A & B): Distributions of the Temperature and Forest Area features.

Feature Time Series Plots

As the data we collected is in fact time series data, it is a good idea to see how for each country a feature changes over time. Therefore, we plotted the country data for each feature as a line plot composed of multiple lines. Once again, we only show a couple here but the remainder are available in the figures folder.

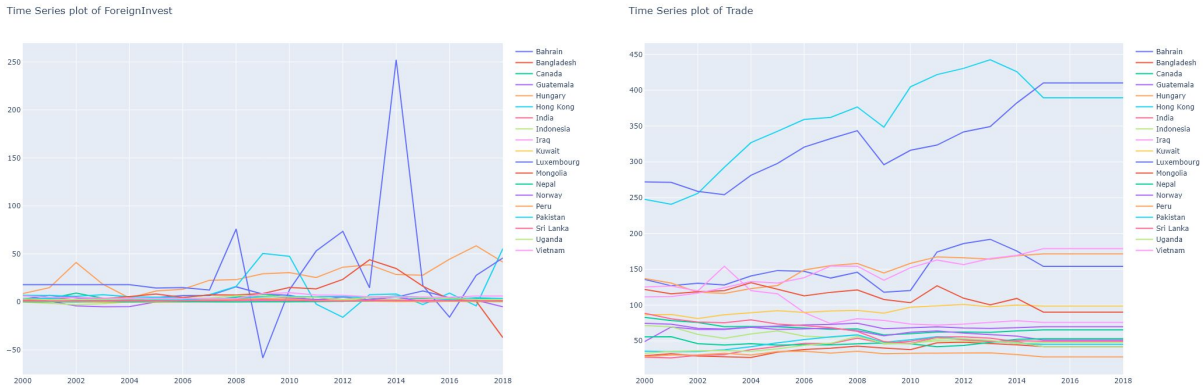

Figure 4 (A & B): Time series plots of the Foreign Investment and Trade features. These two features were selected as certain countries tend to deviate from a similar trend.

Time Series Stationarity

A fundamental component of the statistics and predictive analysis associated with time series data is the concept of stationarity. That is, as time moves forward the data remains constant in mean and variance. For each feature, we averaged the data across countries to see if this was in fact the case with our data. **Note**: An inconsistent average in one country would reflect an inconsistency of the total average.

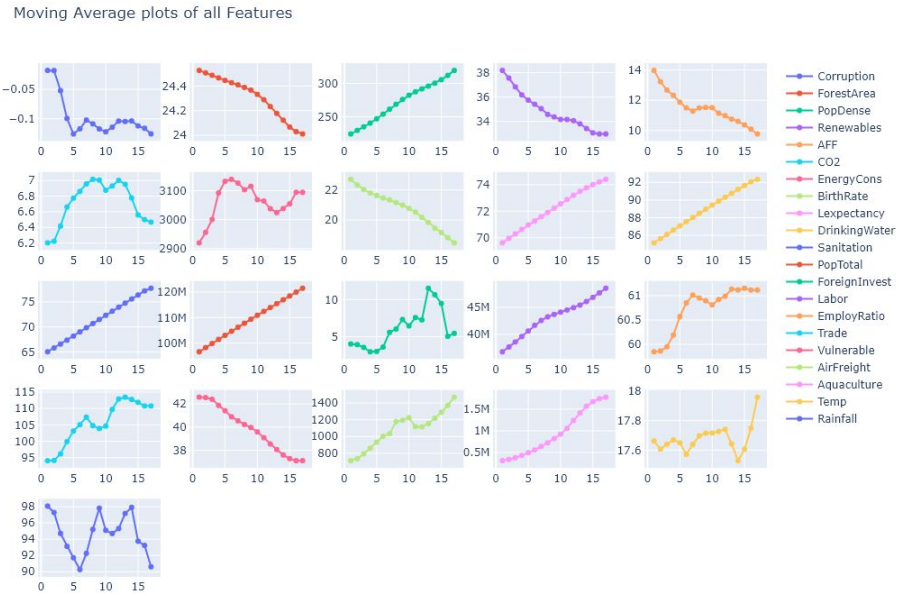

Figure 5: Moving averages to check for stationarity.

Evidently, part of our preprocessing will have to include asserting stationarity by using the rolling average of the data.

Air Quality Time Series Data

To exemplify our data cleaning motivations, we show two plots for the same time series plot in Figure 6; one before cleaning and one after.

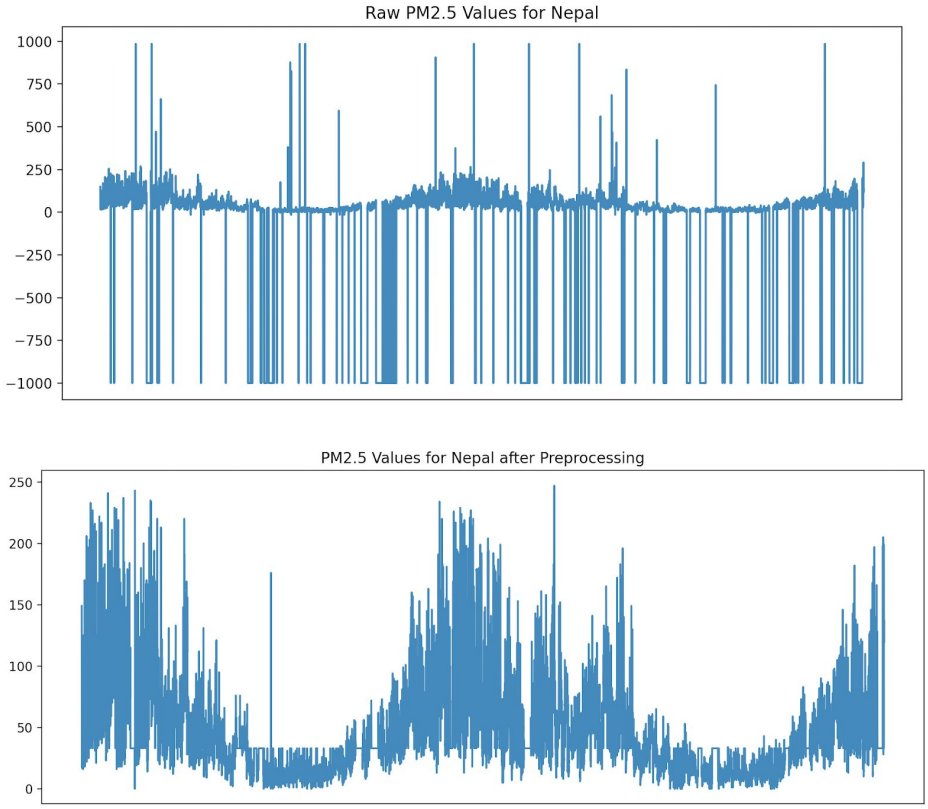

Figure 6 (A & B):Time series plots of Nepal before and after cleaning.

Exploratory Statistics

To conclude the overview of the collected data, we provide the following statistics on each feature prior to filling the missing values.

Table 1: Displays the summary statistics of each feature in the dataset.

Feature	Mean	Median	Min	Max	Variance	Missing
Corruption	-0.1000	-0.4440	-1.5006	2.2945	1.0648	19.0000
ForestArea	24.3424	23.1931	0.2722	59.4898	299.0790	38.0000
PopDense	272.4205	147.5719	1.5432	2017.2737	144191.4688	0.0000
Renewables	35.3178	37.1011	0.0000	94.3178	807.6454	57.0000
AFF	11.7856	11.0544	0.1600	38.2439	88.8503	8.0000
CO2	6.7047	1.8708	0.0608	31.5157	74.5387	38.0000
EnergyCons	3091.9612	909.1753	143.0538	11988.7754	13684713.0000	88.0000
BirthRate	20.7352	19.5010	6.9000	48.5840	81.7499	0.0000
Lexpectancy	72.1511	72.9350	46.2290	84.9341	47.1802	0.0000
DrinkingWater	88.6716	90.6993	26.7516	100.0000	213.4431	19.0000
Sanitation	71.0771	72.3631	15.1242	100.0000	732.1583	19.0000
PopTotal	109202720.0000	25167256.0000	436300.0000	1352617344.0000	72348449223213000.0000	0.0000
ForeignInvest	5.9173	2.1423	-58.3229	252.3081	310.8524	10.0000
Labor	43054396.0000	8341363.0000	176475.0000	512765184.0000	10374376465629100.0000	0.0000
EmployRatio	60.6325	60.7405	37.6860	84.2230	105.3807	57.0000
Trade	104.3098	69.3113	26.2748	442.6200	8062.4731	57.0000
Vulnerable	40.3147	44.3020	1.5890	87.4480	845.5478	57.0000
AirFreight	1226.3669	266.7955	0.0000	13293.1758	6248099.0000	45.0000
Aquaculture	858040.5000	34220.3008	0.0000	16600000.0000	4601696223232.0000	72.0000
Temp	17.6991	22.3552	-7.0771	28.9191	98.7697	0.0000
Rainfall	94.7139	85.2695	1.5379	298.3898	5200.4819	0.0000

Describe the Challenges

We had a few notable challenges throughout this project. The primary challenge we faced which caused plenty of initial frustration was data collection.
Because we obtained data from so many sources, each source had data from different ranges of years, and were collected in different intervals. This made it very challenging to find enough quality data that were sampled at the same rate, or at least could be processed to have the same rate, and which corresponded to the same years. Data collection was the most time consuming task, and there is still lots of room to supplement and improve our current data.
Another challenge we faced was being able to use the limited data in a practical and useful way.

More towards the methodology side, the lack of formal training with time series models and data required us to put in extensive time learning new libraries, concepts and ways of handling data in general. Although we call this a challenge, it was extremely rewarding gaining this new information we hope to apply in the future.

Methodology

All code for the methodology section can be found in the code folder. Code that was no longer relevant to the project vision got moved to the archive folder.

Data Cleaning

There existed two main spots where data cleaning needed to occur. The first being with the fine granularity air quality data as this is raw collected data and device measurement errors are inevitable. The second being within the collected data from the World Bank where certain indicators were missing for a country in a given year. This data cleaning proved to be the more difficult task of the two as the data we are dealing with is time series data meaning how we filled in these missing values required further thought.
In the air quality data, there were many erroneous data values such as -999 and values far too large to be a proper measure, which can be seen in the plot above in Figure X. Instances of -999 were merely replaced with the median of the data for that country. To determine unrealistically large data values, an upper threshold was found by plotting the values in a histogram and if there was a large gap in the histogram, the threshold was placed somewhere in the range where the gap occurred. Any values above this threshold were deemed erroneous and were replaced with the median of the data for that country.
For the latter case of filling in the missing indicator values, we came across two distinct cases that we accommodated for.

1) A country did not possess data for an entire feature, in this situation the average for each year of other countries was used. This gives a trend representative of the feature while we are unable to make any assertions regarding the scale of the data that this country would have had.
2) A country was only missing a few values. Here, we originally planned to use a linear imputer as found in scikit-learn but this would then treat the time scale as increasing or decreasing for whichever direction we wished to impute. Instead, we used the time interpolation method provided by pandas in Python.

After the data was cleaned, one more step needed to be performed for the air quality data as each country still had varying numbers of recordings over the two years, ranging from about 1500 to 17000. Therefore, 1500 recordings were uniformly sampled from each of the countries. Although there is information loss involved, downsampling made the algorithms in the next steps much more efficient and feasible to run.

Clustering

The countries were clustered based on their air quality over a two year span, which we analyzed based on the countries PM2.5 values. PM2.5 refers to atmospheric particulate matter; particles which are formed through burning fuel and chemical reactions in the atmosphere and suspend in the air for long periods of time. High PM2.5 values indicate decreased air quality. Initially, we hypothesized there would be three clusters representing poor, average, and high air quality. After further analysis, we determined that only two clusters existed in our dataset, which we deemed to be high air quality (low PM2.5 values) and low air quality (high PM2.5 values). Various clustering methods, such as k-means for time series data and hierarchical clustering, achieved very similar results and all pointed to the existence of 2 clusters. Therefore, hierarchical clustering was ultimately used due to its efficiency and visual explainability, as the dendrogram in Figure X clearly shows why two clusters is appropriate. The clustering of the countries based on their air quality is shown below, where cluster 1 and cluster 2 correspond to countries of high air quality and low air quality, respectively. Figure Y displays the mean PM2.5 values for each country in each cluster. This clustering gave us an idea of each country's air quality, as well as labels which made it possible to explore classification later on.

Cluster 1	**Cluster 2**
Canada (ca)	Bangladesh (bd)
Guatemala (gt)	Bahrain (bh)
Hong Kong (hk)	Indonesia (id)
Hungary (hu)	India (in)
Sri Lanka (lk)	Iraq (iq)
Luxembourg (lu)	Kuwait (kw)
Mongolia (mn)	Nepal (np)
	Norway (no)
	Peru (pe)
	Pakistan (pk
	Uganda (ug)
	Vietnam (vm)

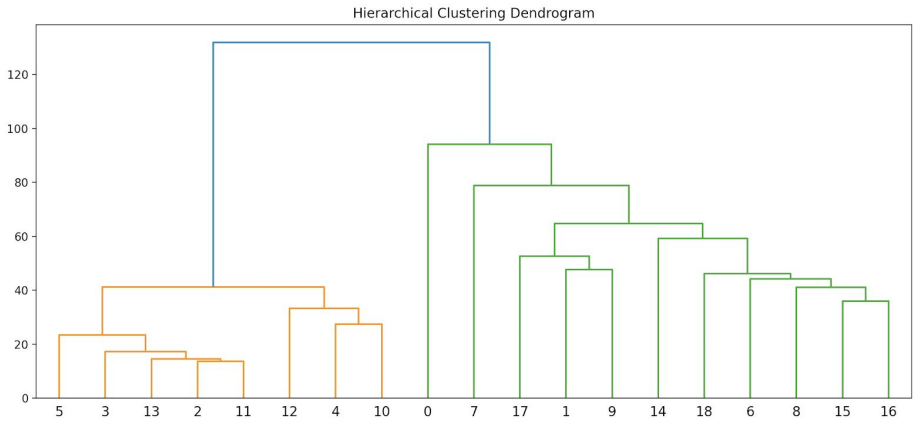

Figure 7: Dendrogram from hierarchical clustering which motivated the number of clusters to select.

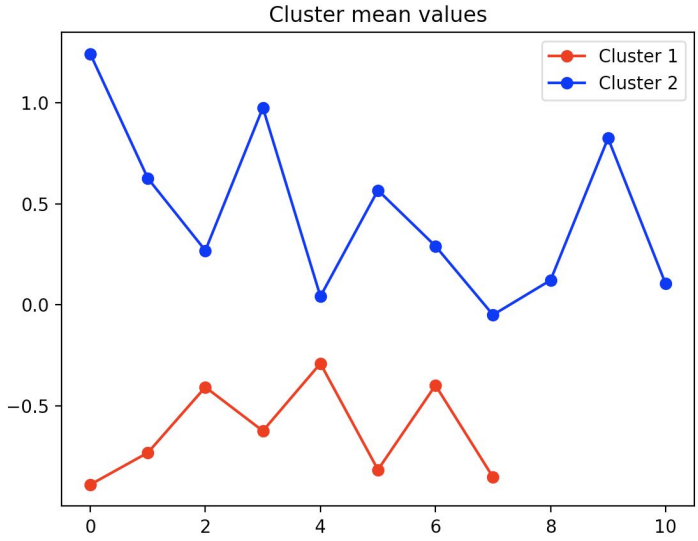

Figure 8: Mean PM2.5 concentration of the countries in each cluster.

Classification

A basic classification model was run on the dataset to determine whether or not we could accurately determine if a country would be in high or low air quality risk with some given data. Ideally this information could be given to citizens and governments and allow for improved living conditions and healthier lives. This was done using a combination of simple preprocessing tools and a K Nearest Neighbours classification model all from SciKit-Learn. The preprocessing tools were labelbinarizer, used to turn the “high” and “low” labels to binary 0 and 1 values and standardscaler, which simply scaled the data allowing for more robust models to be trained. The labels were gathered from the earlier clustering algorithms run during the exploration phase. The data was then split into training and testing sets, at 70% and 30% of the data respectively.

The first approach used took a random sampling of the data for each set. This approach gave a notably high accuracy of 100%. With this number being so high a little more investigation was done and the cause was determined. Because each country has time series data over ten years or more, there are multiple entries for each country with very similar data, all with the same air quality label. Because the split

shuffled the data, nearly every single time the model was run it would take ~70% of the data from each country and leave the last 30% from those same countries for the testing set. The remaining 30% could be modelled exceedingly accurately because it was so similar to the training data - a classic case of overfitting. The second approach aimed to minimize this problem entirely.
The second approach used only the first 13 countries to train and the remaining 6 countries to test. This approach was implemented by simply turning the traintestsplit function to have a shuffle value of "False". The results were much more sensical and this was the final model used.

Regression

We attempted to perform regression on the yearly air quality data, to determine how well the features in our dataset could predict a country's PM2.5 value. We used a basic linear regression model from the scikit-learn package in Python. The data was first standardized and 13 countries were randomly chosen to be a part of the training data, while 6 were randomly chosen for the testing data. Because there was data from many different countries, we decided to train a separate linear regression model for each of the training set countries; 13 models in total. Each of the test set countries were put through each of the 13 models, obtaining 13 different sets of predictions for each country. The average of the 13 predictions was taken for each country, which was used as the final predictions.

Forecasting

With all of the work we did between classifying countries as low or high with regard to PM2.5 concentration, and attempting to regress air quality values for previously seen dates, we were confident the collected country indicators and steps we took to clean and process the data would allow us to make a future forecast for the PM2.5 concentrations of each country. A barrier immediately became evident though with the lack of country indicator data of future years making it impossible to include the data we have for model training.
This work is unsupervised in the sense that we are attempting to predict PM2.5 concentrations for two years in the future, with daily predictions. I.e. 730 predictions. Here we break down the work that went into forecasting these values and attempting to keep the predictions as robust as possible.

To begin, we preprocessed the air quality data as outlined in the *Data Cleaning* subsection by removing -999 occurrences with the country's median. Then a country's data was split, 70%and 30%, for training and validation respectively. We then uniformly sample these sets, essentially splitting them into multiple copies to provide a more stable estimate of the hyperparameter. These sets were then used to tune two hyperparameters within our model, the window size for computing the rolling average and the order parameter used by the Autoregressive Integrated Moving Average (ARIMA) model. For each country, the code in *forecasting.py* provides the best hyperparameter pair. Once the hyperparameters have been tuned, we can then use those and the entire set of air quality data to fit an ARIMA model. The ARIMA max lag parameter is chosen using AIC criterion. This is done by the statsmodels library which we also use for the constructing the ARIMA.

The final models constructed using the outlined routine, are then used to forecast 730 observations into the future which covers two years into the future after 2020. The plots of our forecasting can be found in the figures folder.

Evaluation

Classification

The classification results were consistent and accurate to the countries and data presented. The model achieved 87% accuracy in determining a country's air quality risk level. Standardization improved the model from 74% to 87% on the same set of data.

Regression

The regression results were very inconsistent and sensitive to the set of countries chosen for the train and test sets. Figure X shows the actual air quality values and predicted values of Vietnam, which had some of the best results. The overall trend is very similar, however the magnitudes of the values are drastically different. We attempted to solve this problem by only standardizing the features, and not the PM2.5 values, however, this resulted in the same issue. Ultimately, we could not determine what caused this to occur. Root mean-squared error (RMSE) was used to determine the accuracy of the model. Because of the inconsistency of the model, these values change significantly each time the model is run. Overall, we saw RMSE values between 20 and 2000, which are very large and indicate that this model cannot accurately predict the air quality of a country.

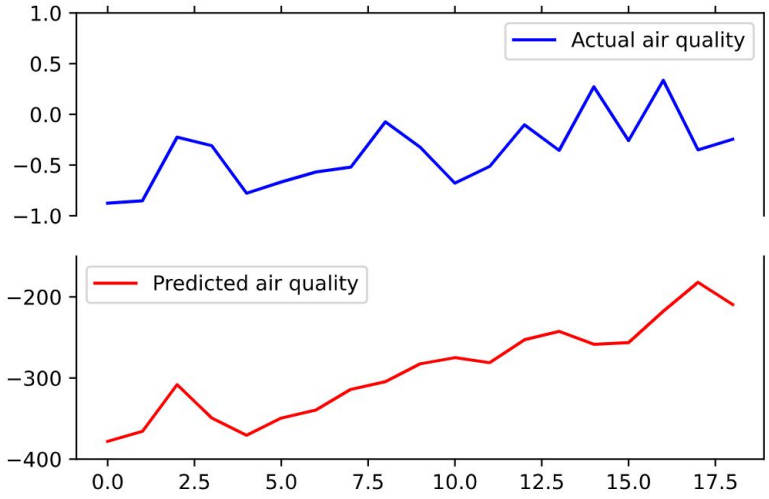

Figure 9: Standardized actual air quality values (blue) compared with the predicted air quality (red) for Uganda using the regression model described in the Methodology section

Conclusions

General

Overall, this analysis gave us a good insight into how analytics are currently applied for social good and in a practical setting. We now recognize the limitations involving pulling data from numerous sources with different purposes and how it can pose a constraint on a project. Data was our main roadblock from the get-go and we believe our work gave a creative and practical solution given the constraint with practicality always on our mind during this project. Between our various modeling approaches, and in-depth exploratory analysis the information garnered could be valuable to an organization working in the realm of public health.

Classification

The classification model was successful in predicting which risk class a country should be placed in. This information would be worthwhile to governments and citizens alike in determining their air quality without needing to rely on explicit readings. Having more data and more countries would give a more robust model; however, with the limited amount of data given, this model was still quite accurate which is very encouraging. With more data the robustness would increase and concerned parties can be even more certain in taking actions to optimize their health.

Regression

The regression model was unsuccessful in predicting a country's air quality given the set of features we obtained. Although it sometimes was able to pick up general trends in air quality, it was sufficient in predicting accurate magnitudes for the air quality measure. Different data would likely need to be collected in order to improve the predictions, as tuning the model likely won't be enough to make up for the lack of accuracy. Sheer lack of data is also a problem that we were challenged with, and increasing the amount of data we had would likely improve the model, although it is unknown whether that would be enough to construct an effective model. It is likely that the features we obtained are not appropriate indicators for air quality.

Forecasting

From the forecast plots for each country we see despite our best efforts these forecasts are not entirely robust. We can see there exists predictions which given the historical data, intuitively do not make sense. Otherwise, it appears the ARIMA model is capable of picking up on some aspects of the seasonality and trend with the data to make future predictions which fall within a plausible value range. Some noticeable results include India's tendency to higher concentrations of PM2.5 in the future, the variation in concentrations for Pakistan and despite Luxembourg's odd drastic spike the country's ability to keep PM2.5 concentrations extremely low.
It will be interesting for us to see how the air quality of these countries changes over these next few years to see whether the modeling we did has any merit.

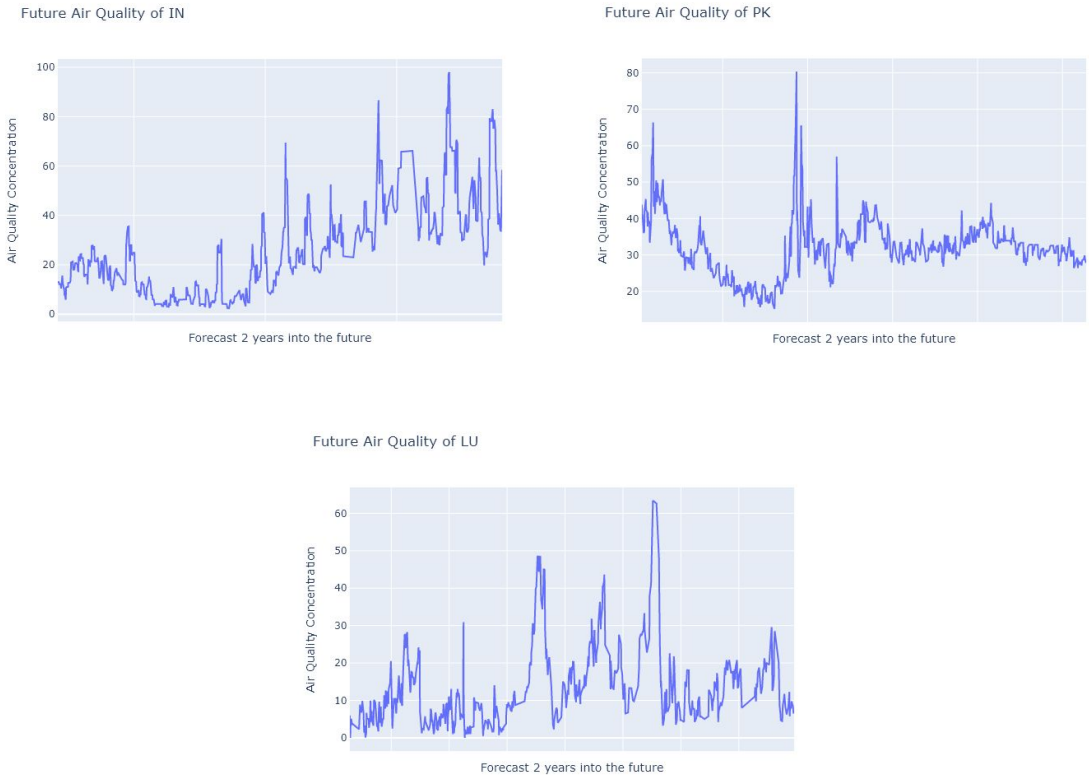

Figure 10 (A, B & C): Future forecasts of India, Pakistan and Luxembourg.

Future Work

Work that could continue to elevate this project can be attributed to either the data collection aspect or modelling.

For the former, we came to the conclusion that country indicators available on a fine granularity are very difficult to obtain for the fact it doesn't benefit countries to record data on a granularity past yearly aside for economic data (quarterly) and a few other indicators not related to air quality. Thus, to further this part of the project some serious digging would have to be done to uncover not publicly-known data whether that is scraping large amounts of websites or reaching out to those who head different government divisions.

Speaking on the subject of modeling for this project is difficult as the small dataset makes it difficult to highlight any blatant issues associated with our model. The impression we received though is that the original PM2.5 and country indicator data we collected is quality for modeling. With that knowledge, with more of it at a finer granularity we suspect that interpretable models like decision trees, and logistic regression would provide good results on top of a model useful to an organization who would like to investigate PM2.5 under the guise of these country indicators.