Global Climate Change Visual Analysis Tool

by

Dylan Beaufort Dias

A Project Proposal Submitted
in
Partial Fulfillment of the
Requirements for the Degree of
Master of Science
Supervised by

Prof. Tae Oh

School of Information

B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York

November 2021

The project proposal "Global Climate Change Visual Analysis Tool" by Dylan Beaufort Dias has been examined and approved by the following Examination Committee:

Prof. Tae Oh
Project Committee Chair

Prof. Michael McQuaid
Committee Member

Acknowledgments

I'm grateful to Prof. Tae Oh for his weekly feedback and suggestions that helped me to get this capstone to move smoothly. A special thanks to Prof. Michael McQuaid for his great understanding of the concepts applied in this project and feedback that helped me complete this project at ease. I would like to thank the IST department for giving me this opportunity to learn and grow during my time at RIT. Finally, I would like to thank everyone who has helped me and been with me throughout my masters. Thank you all!.

Abstract

Global Climate Change
Visual Analysis Tool

Dylan Beaufort Dias

Supervising Professor: Prof. Tae Oh

Climate change can be defined as the change in the usual weather conditions of a certain region. We can also define climate change as the change in Earths climate like rise in temperature, melting of glaciers, and more frequent hurricanes, floods, storms, etc. The main reason for the current climate change is due to human activities like burning of fossil fuels like natural gas, oil, and coal which in turn trap heat inside the atmosphere causing the Earths average temperature to rise. The main goal of this climate change interface is to analyze the Earths climate over the years and predict future climate change based on human activities. Through this analysis, we can understand what changes are taking place in the Earths atmosphere and over the world. The ARIMA-based statistical forecasting model is used to predict future temperature values, and k-means clustering is used to group countries based on their carbon dioxide emission levels. Climate change is a serious global environmental problem, and this paper provides an overview on how the rate of climate change has escalated over these years leading to global warming.

Contents

List of Tables

List of Figures

1 Introduction

Climate can be described as the change in the average weather conditions (like rainfall, snow, temperature) of a region over a long period of time. Global climate change refers to the average weather changes taking place over the entire Earth. Such effects can be seen by rising sea levels, shrinking glaciers, melting of ice, changes in plants blooming times, etc. According to a study conducted by NASA (National Aeronautics and Space Administration) over the past 150 years Earth's average temperature has been increasing at a much quicker rate. Some parts of Earth are getting much warmer than other. The past five years have been recorded as the warmest years. In facts, the Earths surface temperature has increased by 2° C in the past 100 years which is growing concern among many scientists. Earths climate is constantly changing but due to human interaction, this effect has been accelerated. Certain gases are preventing the heat from escaping the atmosphere which is making the Earth hotter. This is called the greenhouse effect. Human activities such as burning fuel to power factories, cars, and buses are changing the natural greenhouse, leading to a warmer Earth.

Climate change is a global environmental problem that affects not only us but even plants, animals, air, land mostly everything. If efforts are not taken now to reduce CO2 emission it can lead to disastrous effects in the near future. We need to analyze which parts of the Earth are getting hotter so that initiatives can be taken to reduce the carbon footprint in those parts of Earth.

In this project, we will develop a global climate change interface that will help us to give a brief overview on how Earths climate has been changing over all these years. We will develop this interface using Dash which is a web microframework of python. We will generate all the necessary visualizations based on the data which will help to analyze the situation at hand. Also, we will be implementing a time series model to predict future temperature change values and cluster countries into groups based on their emission rate. All of this will help us understand the overall changes taking place all over the Earth so that the necessary precautions will be taken to reduce it as much as possible.

2 Problem Statement

Ideal:
Climate change is a global problem and efforts need to be made to reduce it as much as possible. Carbon emissions from fossils fuels have been linked to global warming and climate change. Developing this interface can help analyze how the burning of fossil fuel over all these years has caused the Earths temperature to rise. Implementing a time series model to predict future climate change can also help. In turn, necessary steps can be taken to reduce carbon emission levels.
Reality:
According to research done by National Aeronautics and Space Administration "Climate change: Vital signs of the planet," n.d. carbon dioxide levels in the air are at their highest

in 65000 years and the main source of this rise is due to human activities such as deforestation and burning fossil fuels, as well as natural processes such as respiration and volcanic eruptions. The temperature of the earth is also changing in fact nineteen of the hottest years have occurred since 2000. If measures are not taken to reduce the level of carbon dioxide in the air it will keep increasing at an immeasurable rate and which in turn will affect the global temperature of the earth.

Consequences:

To solve this problem we need to identify countries who are the main sources of carbon dioxide emission by clustering them into various groups based on their emission levels. By knowing which countries have been emitting more carbon dioxide than the other restrictions can be put on such countries to prevent future emissions. In order to check how the climate will change in the future, forecasting models can be implemented to give us an overview on future climate change. But for these models to give accurate predictions, we need to supply them with as much historical data as possible.

3 Goals and Objectives

Develop a visualization tool to analyze fossil fuel and temperature data, predict future temperature change values and group countries based on their carbon dioxide emission level.

Goal 1: Implement the visualization tool by storing the data on the cloud server using MongoDB Atlas and deploying the application on AWS Elastic Beanstalk.

Objective 1: Collect the data for fossil fuel types such as fuel, coal and gas and also for temperature change for various countries.

Objective 2: Perform exploratory analysis on the data collected.

Objective 3: Implement a user-friendly dashboard.

Objective 4: Create a web-interface which consists 4 sections that are introduction, exploratory analysis, dashboard and statistical models.

Goal 2: Implementing ARIMA forecasting model and K-means clustering algorithms.

Objective 1: Determine the number of clusters for K-means using Elbow method.

Objective 2: Implement the K-means algorithm to cluster countries into similar groups based on the carbon dioxide emmission level.

Objective 3: Divide the dataset as training dataset and test dataset and test dataset in the ratio of 70:30.

Objective 4: Find the optimum values of p, d and q using grid search method.

Objective 5: Validate the model accuracy using the RMSE score with the test dataset.

Objective 6: Implement the new ARIMA with the new order score to predict future temperature change values.

4 Related Work

Webster et al. (2003) wanted to improve the climate change policies, and came to a conclusion that accurate quantitative descriptions of the uncertainty in climate outcomes under various possible policies are needed. Here they applied two different policies on earths system. The first policy talks about the absence of greenhouse gases on earths ecosystem and the second policy talks about the presence of greenhouse gases on earths ecosystem. Based on the first policy there is a one in forty chance that global mean surface temperature change will exceed 4.9 °C by the year 2100 and based on the second policy there is a lower chance that temperature will exceed 3.2 °C, thus reducing but not eliminating the chance of substantial warming. In reality it will take time to implement such policies, but it is a great effort overall.

Scholze et al. (2006) analyzed climate change risk on world ecosystem. Here they considered distribution of outcomes based on three set of models. First model talks about global temperature less than 2°C, second model talks about global temperature between 2-3°C, and third model talks about temperature more than 3°C. This method helps to give an overview of climate change for three different scenarios.

El-Mallah and Elsharkawy (2016) developed an ARIMA time series model to predict the annual warming trend. They found out that the non-seasonal linear trend model ARIMA (3-1-2) and quadratic trend model ARIMA (3-2-3) are most optimum models obtained for the data.

Girvetz et al. (2009) developed a web-based tool called Climate Wizard. The main purpose for developing this tool is to provide non-climate specialist with simple analyses and innovative graphical depictions for conveying how climate has and is projected to change within specific geographic areas throughout the world.

Lai and Dzombak (2020) implemented an ARIMA-based forecasting model to provide estimates of confidence intervals for temperature and precipitation extremes in different return periods, and to provide future daily temperature and precipitation simulations. It said that ARIMA model is an efficient, interpretable, and reliable method for obtaining near-term (220 years) regional temperature and precipitation forecasts for use in various engineering applications.

Kodinariya and Makwana (2013) compared various methods available to determine the value of K which is the most controversial issues in clustering algorithms.

A thorough review of the literature revealed that the ARIMA model is best suited to tackle the on-going climate change issue. The correct model can accurately predict the next 5-10 years of climate change values. Also by using k-means clustering, countries can be grouped based on their emission level in order to identify highly emitting countries.

5 Methodology

5.1 Dataset Selection and Quality

The source of this dataset is (Andrew & Peters, 2021). The fossil fuels dataset consists of 3 csv files for coal, oil, and gas fossil fuel type. Each csv file contains 223 countries and carbon dioxide emission per million tonnes between the years 2000 to 2019 for all those countries.

The source of this dataset is (SY, 2020). The temperature change dataset consists of a single csv file containing temperature change data for 274 countries, major cities, and continents. It contains temperature change data between the years 2000 to 2019 for all those countries, cities, and continents.

While checking for the completeness of the fossil fuels dataset it consisted of missing values which were indicated by "None" which meant that the fossil fuel emission for that particular country and year was 0. By making use of Excel's "Replace By" function those "None" values were replaced with the value 0.

5.2 User Interface Design

5.2.1 Interface Flowchart

Figure 1 show the flowchart that demonstrates how the web interface will be deloyed.

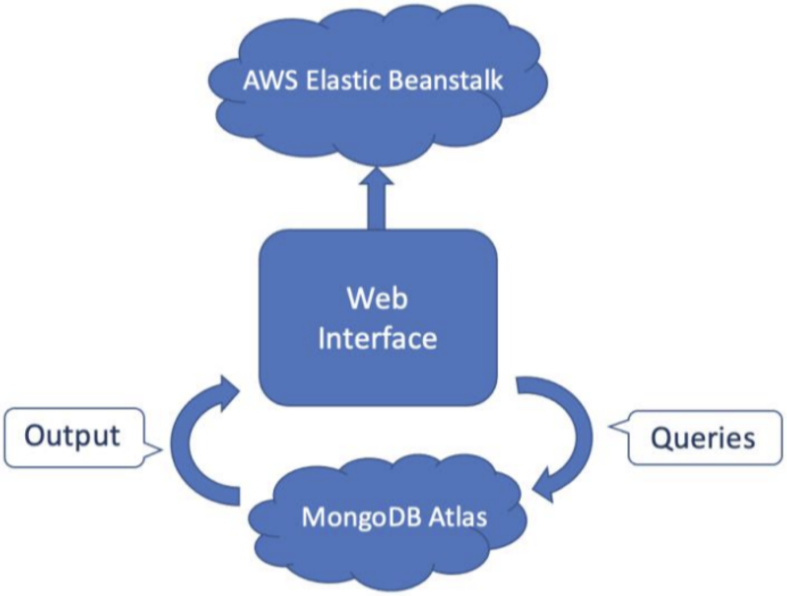
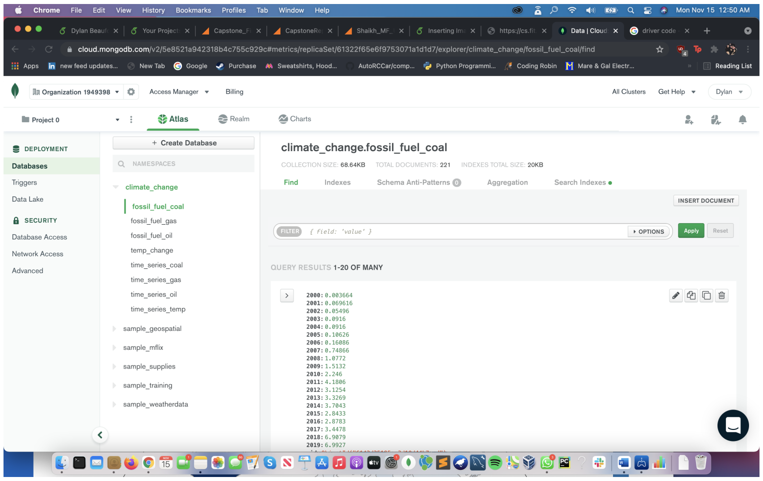


Figure 1. Web interface flowchart

Firstly, the datasets were stored on a cloud server which can be seen in figure 2. The cloud server used for this project is MongoDB Atlas. MongoDB Atlas is a fully managed cloud database that can be used to store and manage databases on the cloud server of your choice which in this case is AWS. The datasets were then stored in the cloud server under the database name "climate_change" and the IP address was set to 0.0.0.0/0 indicating the datasets can be accessed from any network. Using the driver code the datasets were then accessed by the python code created.



climate_change
fossil_fuel_coal
fossil_fuel_gas
fossil_fuel_oil
temp_change
time_series_coal
time_series_gas
time_series_oil
time_series_temp

sample_geospatial
sample_mflix
sample_supplies
sample_training
sample_weatherdata

climate_change.fossil_fuel_coal

COLLECTION SIZE: 68.64KB TOTAL DOCUMENTS: 221 INDEXES TOTAL SIZE: 20KB

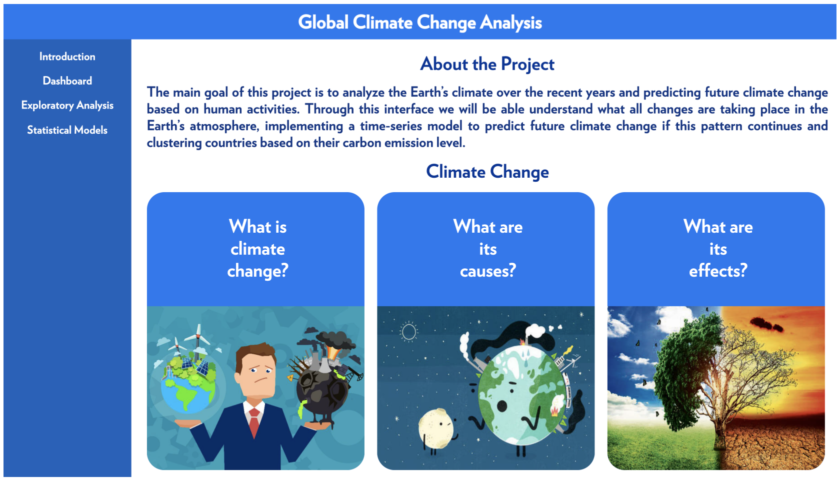
`{ field: 'value' }`

QUERY RESULTS 1-20 OF MANY

```
2000: 0.003664
2001: 0.069616
2002: 0.05496
2003: 0.0916
2004: 0.0916
2005: 0.10626
2006: 0.16086
2007: 0.74866
2008: 1.0772
2009: 1.5132
2010: 2.246
2011: 4.1806
2012: 3.1254
2013: 3.3269
2014: 3.7043
2015: 2.8433
2016: 2.8783
2017: 3.4478
2018: 6.9079
2019: 6.9927
```

Figure 2. MongoDB Atlas

Figure 3 shows the web interface that was created using Dash which is a python framework created by Plotly for creating interactive web applications. Dash makes use of a callback function which is used to update the visualizations in real-time. The main plus point of plotly is its interactive nature and of course visual quality.



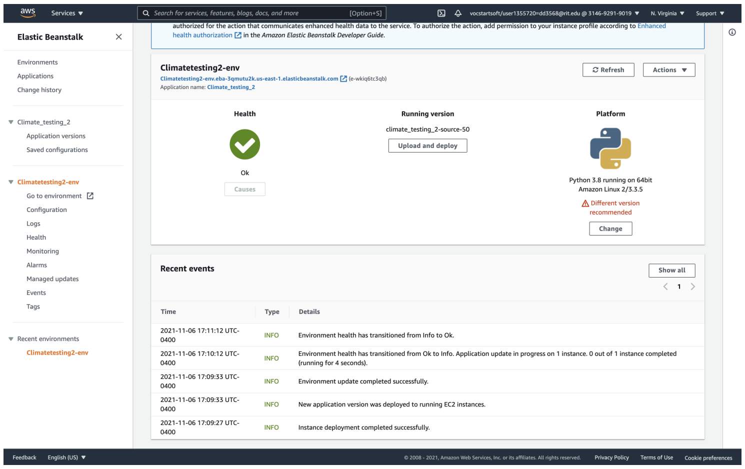
About the Project

The main goal of this project is to analyze the Earth's climate over the recent years and predicting future climate change based on human activities. Through this interface we will be able understand what all changes are taking place in the Earth's atmosphere, implementing a time-series model to predict future climate change if this pattern continues and clustering countries based on their carbon emission level.

Climate Change

Figure 3. Web interface

Finally, the web interface was deployed using AWS Elastic Beanstalk which is seen in figure 4 which is a compute service used for quickly deploying high-scaled applications. This is done by creating a requirements.txt file which contains a list of python libraries used along with their version numbers. The python file, text file, assets folder containing .css file, and image are then all zipped together and uploaded on Elastic Beanstalk. AWS Elastic Beanstalk automatically handles the details of capacity provisioning, load balancing, scaling, and application health monitoring.



authorized for the action that communicates enhanced health data to the service. To authorize the action, add permission to your instance profile according to Enhanced health authorization in the *Amazon Elastic Beanstalk Developer Guide*.

Climatetesting2-env

Climatetesting2-env.eba-3qmutu2k.us-east-1.elasticbeanstalk.com (e-wkiq6tc3qb)

Application name: Climate_testing_2

Health: Ok

Running version: climate_testing_2-source-50

Platform: Python 3.8 running on 64bit Amazon Linux 2/3.3.5

Different version recommended

Recent events

Time	Type	Details
2021-11-06 17:11:12 UTC-0400	INFO	Environment health has transitioned from Info to Ok.
2021-11-06 17:10:12 UTC-0400	INFO	Environment health has transitioned from Ok to Info. Application update in progress on 1 instance. 0 out of 1 instance completed (running for 4 seconds).
2021-11-06 17:09:33 UTC-0400	INFO	Environment update completed successfully.
2021-11-06 17:09:33 UTC-0400	INFO	New application version was deployed to running EC2 instances.
2021-11-06 17:09:27 UTC-0400	INFO	Instance deployment completed successfully.

© 2008 - 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Figure 4. AWS Elastic Beanstalk

5.3 Exploratory Data Analysis

Fossil Fuels Dataset: Exploratory data analysis is an approach to analyze the data which in turn can help to summarize the main characteristics of the dataset which can be used to detect various patterns and get a better understanding of the data.

In order to get a summary of the dataset various statistical parameters have been calculated which can be seen in figure 5.

Count: Count of the number of data points.

Mean: Average fossil fuel value for the country.

std: Standard deviation is a measure of how spread the data is.

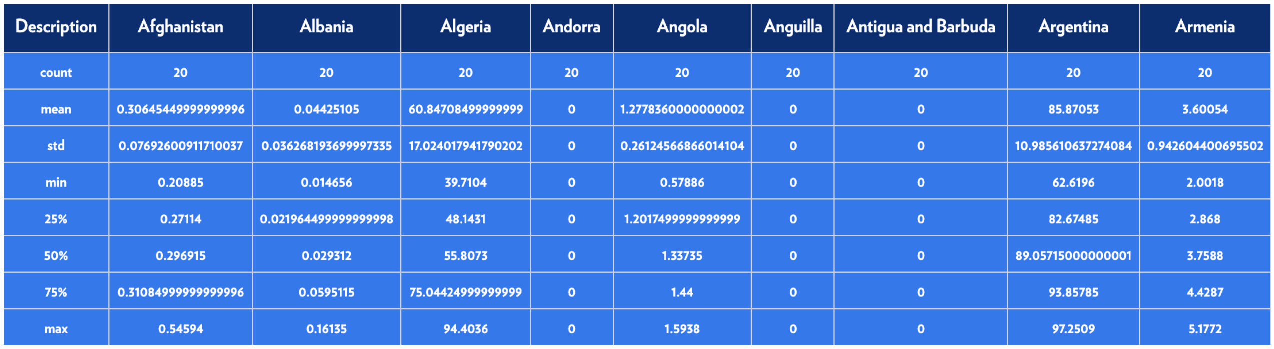
Min: Minimum fossil fuel value for the country.

25%: 25th percentile is the value at which 25% of the data lie below that value, and 75% of the data lie above that value.

50%: 50th percentile is the median of data.

75%: 75th percentile is the value at which 75% of data will be less than 75th percentile; 25% of data will be more than 75th percentile.

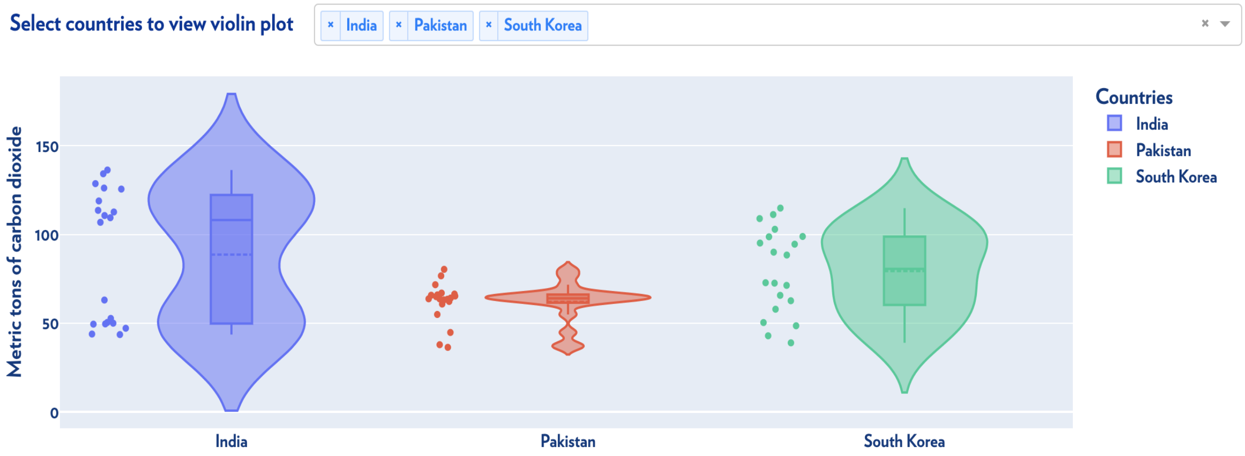
Max: Maximum fossil fuel value for the country.



Description	Afghanistan	Albania	Algeria	Andorra	Angola	Anguilla	Antigua and Barbuda	Argentina	Armenia
count	20	20	20	20	20	20	20	20	20
mean	0.30645449999999996	0.04425105	60.84708499999999	0	1.2778360000000002	0	0	85.87053	3.60054
std	0.07692600911710037	0.036268193699997335	17.024017941790202	0	0.26124566866014104	0	0	10.985610637274084	0.942604400695502
min	0.20885	0.014656	39.7104	0	0.57886	0	0	62.6196	2.0018
25%	0.27114	0.021964499999999998	48.1431	0	1.2017499999999999	0	0	82.67485	2.868
50%	0.296915	0.029312	55.8073	0	1.33735	0	0	89.05715000000001	3.7588
75%	0.31084999999999996	0.0595115	75.04424999999999	0	1.44	0	0	93.85785	4.4287
max	0.54594	0.16135	94.4036	0	1.5938	0	0	97.2509	5.1772

Figure 5. Carbon dioxide data description

There is also a violin plot generated which can be seen in figure 6 to visualize the distribution of the data and its probability density. This chart is a combination of a Box Plot and a Density Plot that is rotated and placed on each side, to show the distribution shape of the data. The user can add the country and the fossil fuel type they wish to visualize which is used to compare the violin plots for various countries using this visualization.



Select countries to view violin plot

× India × Pakistan × South Korea

Figure 6. Violin plot

5.4 Temperature Dataset:

In order to get a summary of the dataset, various statistical parameters have been calculated which can be seen in figure 7.

Count: Count of the number of data points.

Mean: Average temperature value for the country.

std: Standard deviation is a measure of how spread out data is.

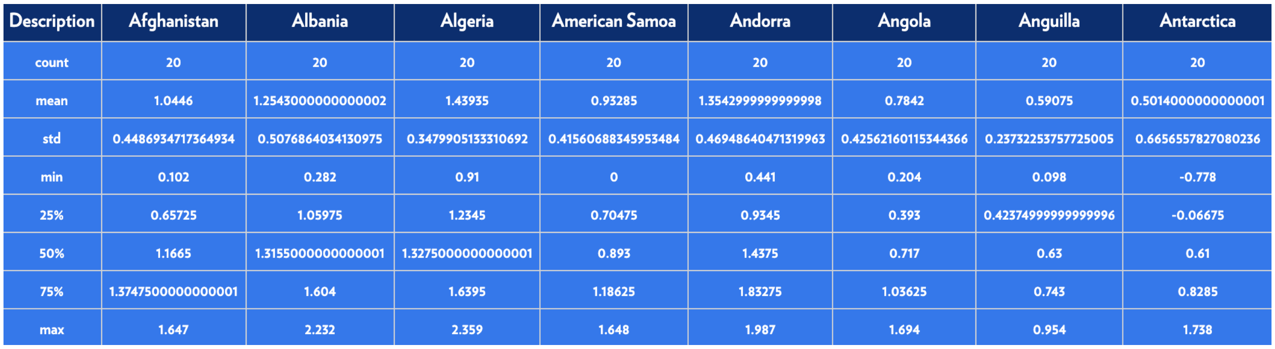
Min: Minimum temperature value for the country.

25%: 25th percentile is the value at which 25% of the data lie below that value, and 75% of the data lie above that value.

50%: 50th percentile is the median of data.

75%: 75th percentile is the value at which 75% of data will be less than 75th percentile; 25% of data will be more than 75th percentile.

Max: Maximum temperate value for the country.



Description	Afghanistan	Albania	Algeria	American Samoa	Andorra	Angola	Anguilla	Antarctica
count	20	20	20	20	20	20	20	20
mean	1.0446	1.2543000000000002	1.43935	0.93285	1.3542999999999998	0.7842	0.59075	0.5014000000000001
std	0.4486934717364934	0.5076864034130975	0.3479905133310692	0.41560688345953484	0.46948640471319963	0.42562160115344366	0.23732253757725005	0.6656557827080236
min	0.102	0.282	0.91	0	0.441	0.204	0.098	-0.778
25%	0.65725	1.05975	1.2345	0.70475	0.9345	0.393	0.42374999999999996	-0.06675
50%	1.1665	1.3155000000000001	1.3275000000000001	0.893	1.4375	0.717	0.63	0.61
75%	1.3747500000000001	1.604	1.6395	1.18625	1.83275	1.03625	0.743	0.8285
max	1.647	2.232	2.359	1.648	1.987	1.694	0.954	1.738

Figure 7. Temperature data description

In figure 8 we have created a violin plot which is used to see the distribution of temperature change data for all 7 seven continents. Figure 9 shows a time series line plot showing the temperature change taking place in these continents from 2000 to 2019. It can be seen from the line plot that the temperature is mostly stationary in those regions. Figure 10 is a bar plot for the 10 ten hottest countries based on the year selected. The user can select the year they desire to view and based on the selected year a bar plot will be generated ranking the top 10 countries.

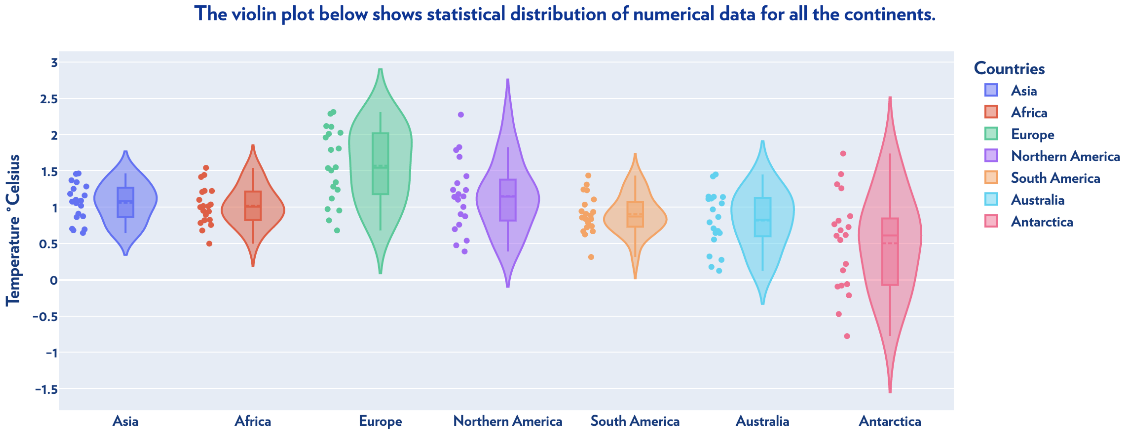


Figure 8. Continents violine plot

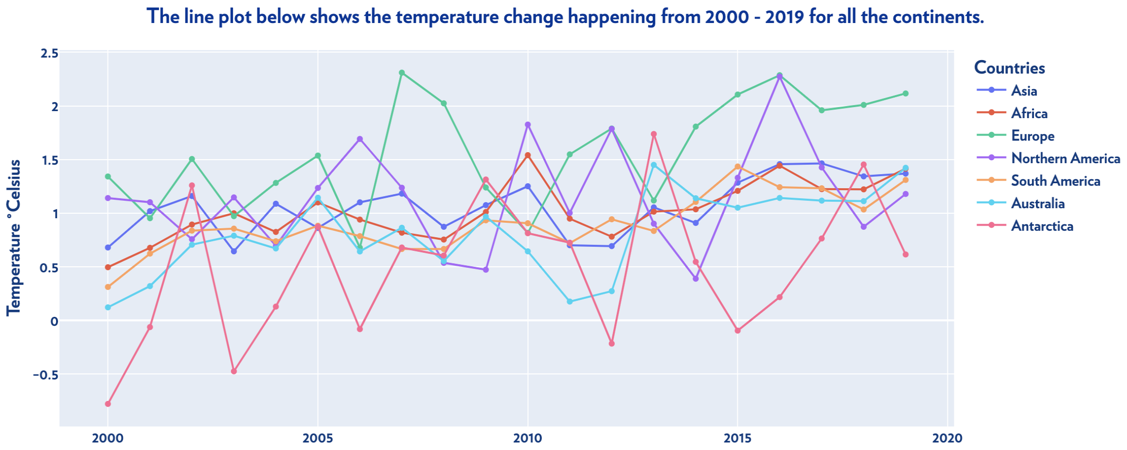


Figure 9. Continents time series line plot

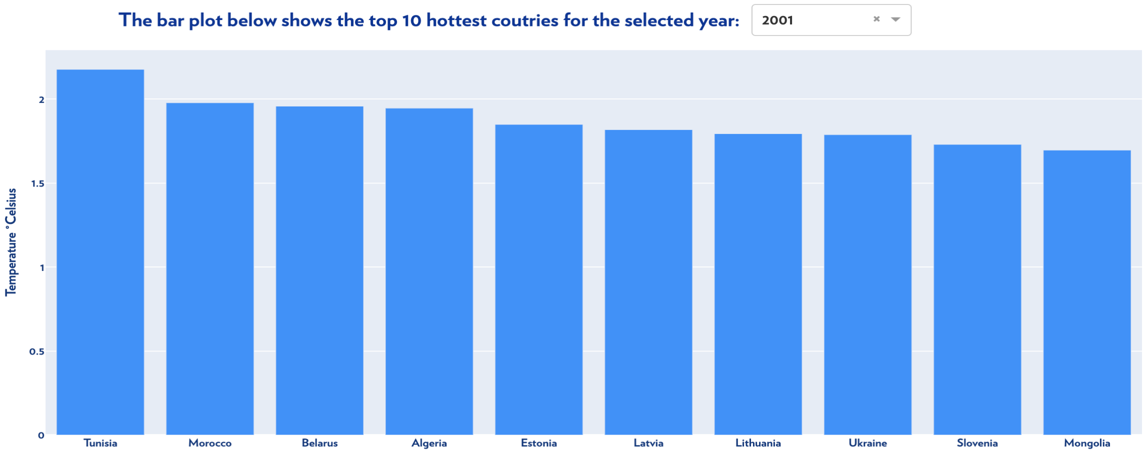


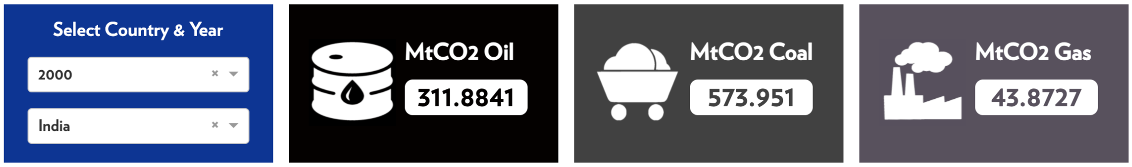
Figure 10. Top 10 hottest countries

5.5 Dashboard

The dashboard created helps in giving useful insights of the data at hand and also helps to identify trends. Using this dashboard the user will be able to interact with the data and manipulate the visualizations that are created.

5.5.1 Fossil fuels data:

The fossil fuels dashboard consists of 5 major visualizations. Visualization 1 can be seen in figure 11 where the user will be able to view the million tonnes of coal, oil, and gas emitted based on the year and country selected. In figure 12 we can see the second visualization which consists of a geo chart which the user can hover over to view the emission level of that country. The user can change the fossil fuel type of the geo chart by using the radio buttons and also change the year using the slider. Figure 13, shows visualization 3 and 4 which consists of a radar plot and time-series area plot. The radar plot consists of three variables total coal, total oil, and total gas emission. Using the dropdown box the user can add n number of countries to compare the total emission levels for each fossil fuel type. The time-series area plot shows the total emission level for each fossil fuel type between the years 2000 to 2019. It also consists of a range slider to select a particular range of years. Finally, visualization 5, consists of a time series line plot. The user can select the fossil fuel type using the radio buttons, select n number of countries using the dropdown box and also select a range of years using the range slider.



Select Country & Year

2000

India

MtCO2 Oil 311.8841

MtCO2 Coal 573.951

MtCO2 Gas 43.8727

Figure 11. Fossil fuel values

5.5.2 Temperatue data:

The temperature change dashboard consists of 4 major visualizations. Visualization 1 can be in figure 15 which is a time series line plot for the temperature change of the entire world. In figure 16 we can see the 2nd visualization which consists of time series line chart for countries. The user can add n number of countries from the dropdown list. In figure 17 we can see the 3rd visualization which consists of a geo chart which the user can hover over to view the temperature of that country. The user can also change the year by selecting a year from the slider. And finally, from figure 18 we can see the 4th visualization which consists of a violin plot to show the distribution of temperature change data. The user can add n number of countries from the dropdown list to this visualization.

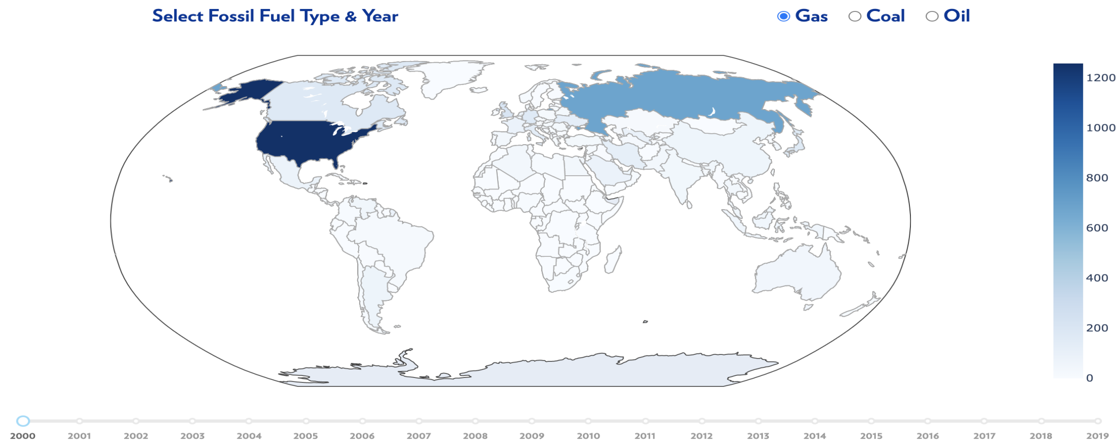


Figure 12. Fossil fuels geochart

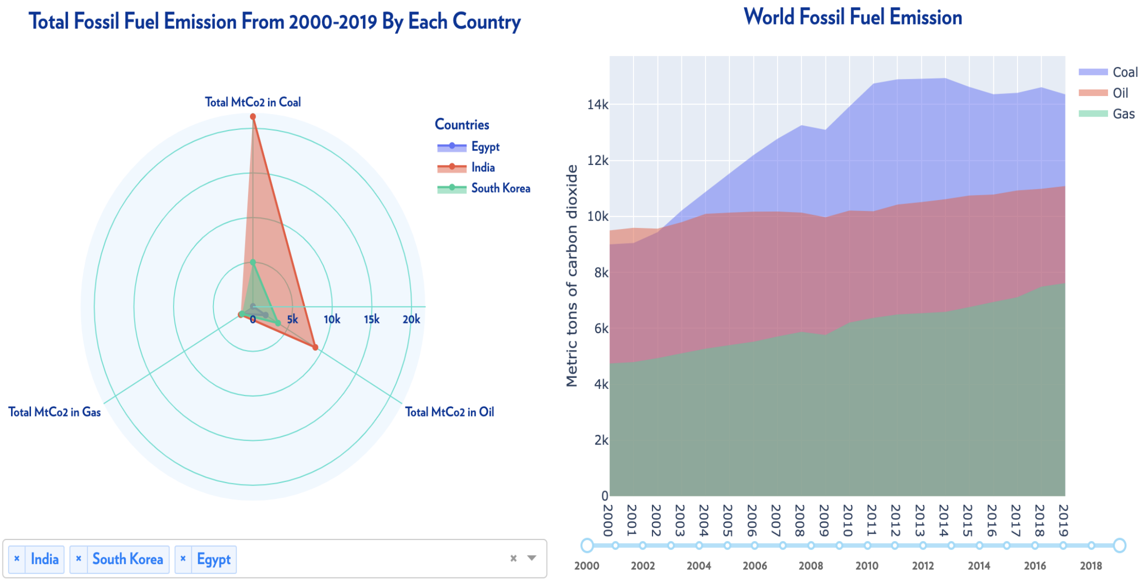
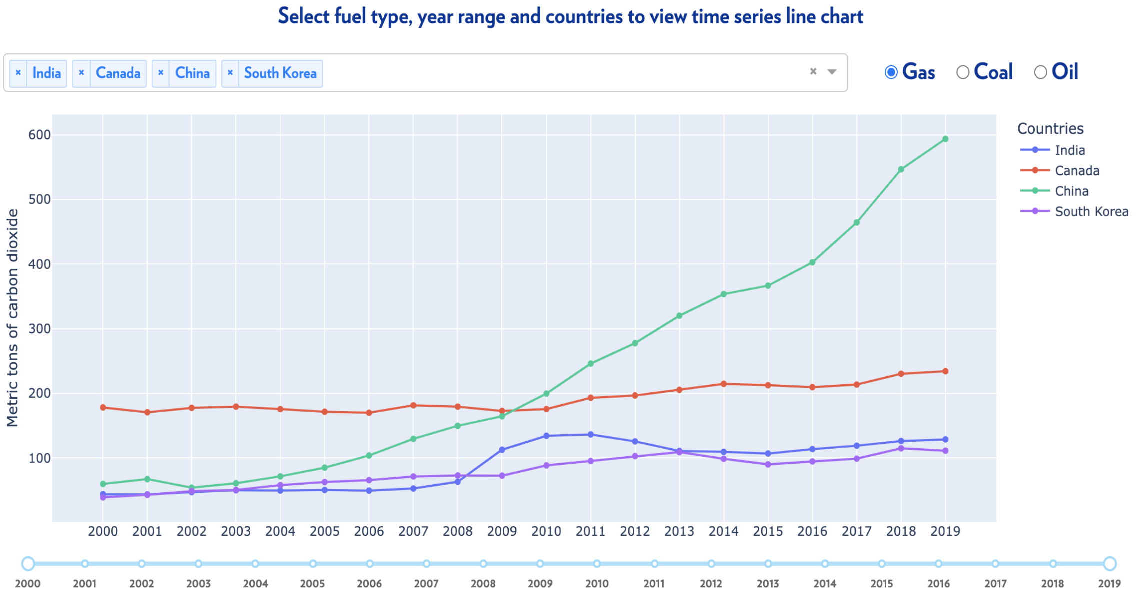


Figure 13. Fossil fuels radar and area plot



Select fuel type, year range and countries to view time series line chart

× India × Canada × China × South Korea

● Gas ○ Coal ○ Oil

Figure 14. Fossil fuels line plot

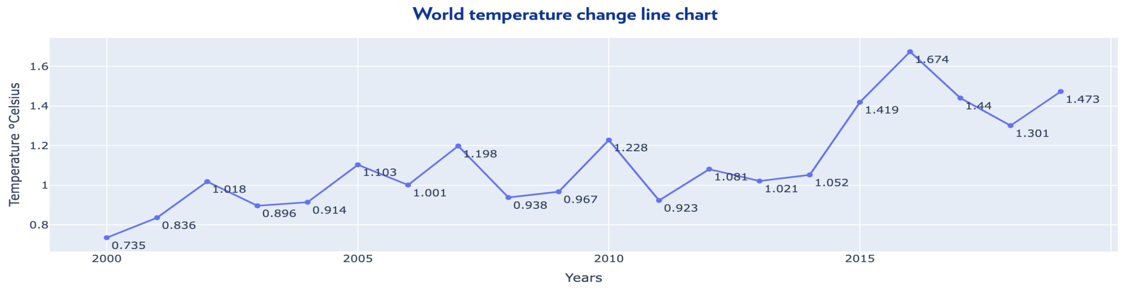


Figure 15. World temperature change line plot

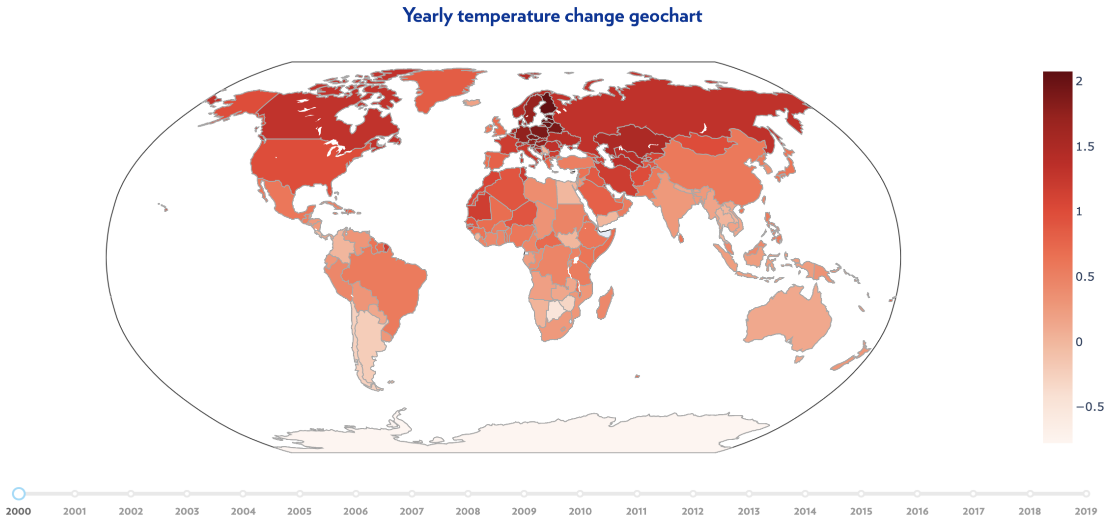


Figure 16. Temperature change geochart

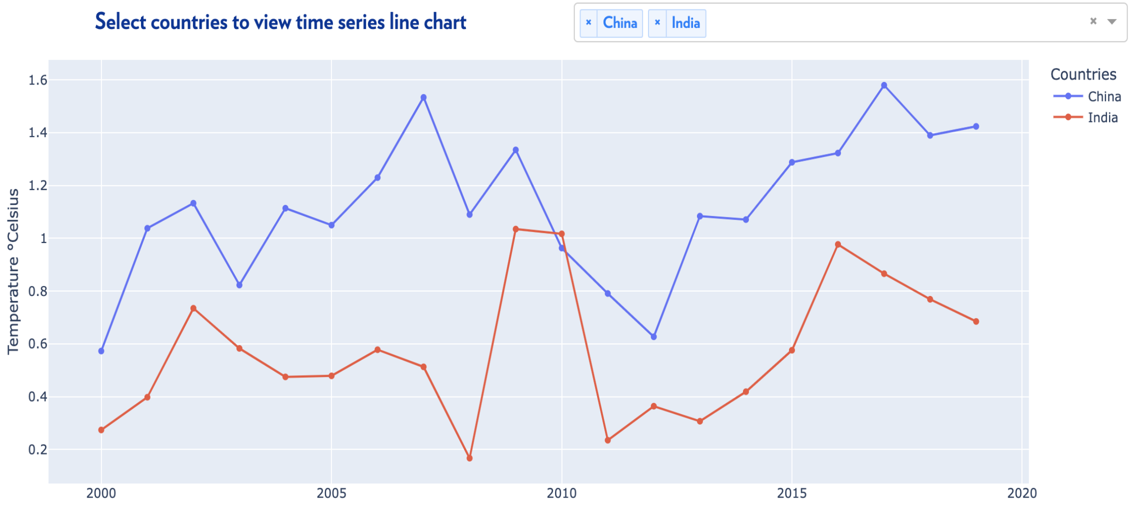


Figure 17. Temperature change line plot

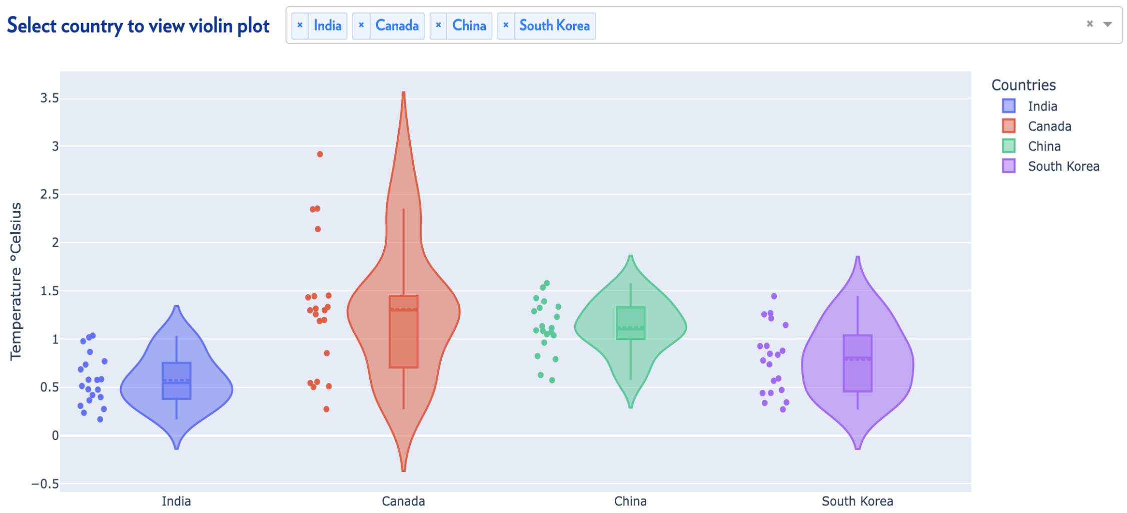


Figure 18. Temperature change violin plot

5.6 Core Algorithms

5.6.1 K-means Clustering

K-means Algorithm: K-means clustering algorithm is an unsupervised learning algorithm that is used to cluster items into K groups. In k-means clustering, we need to first define k. Based on the k value it will randomly select k distinct centroids. After the k centroids have been selected it will calculate the Euclidean distance between each point and the centroid. Each point will be assigned to the nearest cluster based on the smallest euclidian distance. We then calculate the mean of this cluster and based on the mean the new centroids are calculated. K-means algorithm aims to select the centroids that minimize sums of the squared error function. K-means keeps track of the total variance of each cluster and chooses the cluster with the lowest sum of variance.

K means clustering metrics: In order to get the best out of the k-means clustering algorithm, we need to determine the value of k. In this case, we will make use of the Elbow method to give us an idea on what a good k number would be based on the sum of squared distance (SSE) between data points and their assigned clusters centroids. In the Elbow method, you start with k=1 and keep increasing the value of k by 1, calculating your clusters and the cost. At some value for K, the cost drops dramatically, and after that, it decreases at a very slow rate. That is the K value needs to be selected.

5.6.2 ARIMA

ARIMA Model The autoregression integrated moving average(ARIMA) model is a combination of the autoregressive model(AR) and moving average model(MA). ARIMA model

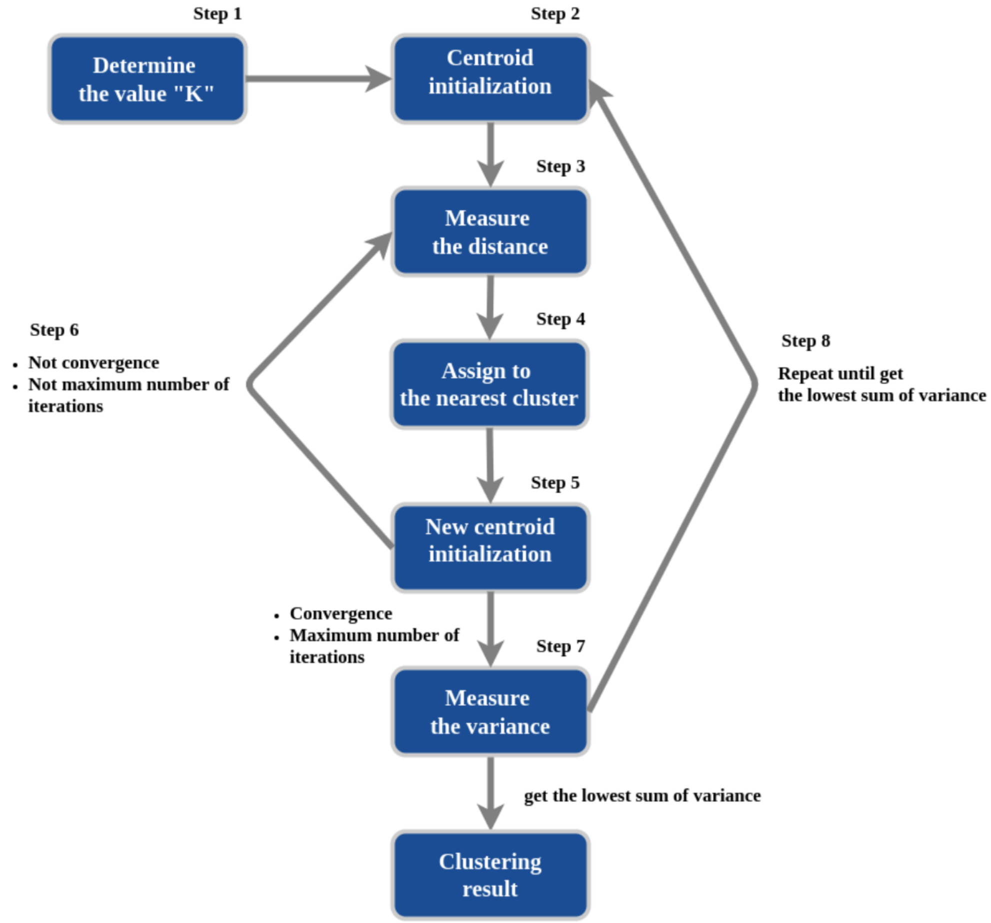

Step 1
Determine the value "K"
Step 2
Centroid initialization
Step 3
Measure the distance
Step 4
Assign to the nearest cluster
Step 5
New centroid initialization
Step 6
Not convergence
Not maximum number of iterations
Convergence
Maximum number of iterations
Step 7
Measure the variance
Step 8
Repeat until get the lowest sum of variance
get the lowest sum of variance
Clustering result

Figure 19. K-means clustering steps

makes use of past data to predict future values as it believes that the future reflects the past. The 'AR' part of ARIMA shows that the time series is regressed on its own past data. The 'MA' part of ARIMA indicates that the forecast error is a linear combination of past errors. The 'I' part of ARIMA shows that the data values have been replaced with differenced values of d order to obtain stationary data, which is the requirement of the ARIMA model approach. Each part is included in the ARIMA model as parameters p, q, and d. p indicates the number of lags, q is the number of forecast errors in the model, and d is known as the degree of differencing indicating the number of times the lagged indicators have been subtracted to make the data stationary.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \beta_1 Y_{t-1} \epsilon_t + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q} \quad (1)$$

ARIMA model metrics: In order to obtain the best value of p, q, and d the grid search method was used. Grid search is a tuning technique that attempts to compute the optimum values of hyperparameters. It is an exhaustive search that is performed on the specific parameter of the model. The value of p was set to a range of 0 to 10 while the value of d and q was set to a range of 0 to 5. The models were evaluated using their root mean squared error(RSME) score and the model with the lowest RMSE score is selected.

6 Experiments and Results

6.1 Exploratory data analysis queries:

Three queries were created in order to get a better understanding of the fossil fuels data. While browsing through the dataset for each fossil fuel type it was noticed that some countries did not emit any carbon dioxide for the past 20 years. So a count query was generated to return the number of countries that have 0 carbon dioxide emissions for the past 20 years. Table 1 shows the results of the count of countries with 0 carbon dioxide emission and with some carbon dioxide emission in the past 20 years.

Fossil Fuel Type	Countries with 0 CO2 emission	Countries with some CO2 emission
Coal	91	130
Gas	94	127
Oil	8	213

Table 1. Count of countries with 0 carbon dioxide emission and with some carbon dioxide emission.

A count of countries for various ranges of carbon dioxide emission was also generated which included a count of countries with less than 500 million tonnes of carbon dioxide emission, count of countries that have between 500 to 5000 million tonnes of carbon dioxide emission, and a count of countries with more than 5000 million tonnes of carbon dioxide emission. Table 2 shows the results of the count of countries for various ranges of carbon dioxide emission.

Count			
Fossil Fuel Type	Less than 500 MtCo2	Between 500 to 5000 MtCo2	More than 5000 MtCo2
Coal	187	27	7
Gas	179	39	3
Oil	168	43	10

Table 2. Count of countries for various ranges of carbon dioxide emmission.

Table 3 ranks the top 10 countries based on their emission level for each fossil fuel type.

6.2 K-means clustering:

6.2.1 Elbow plot:

In order to determine the value of k, an elbow plot was created which can be in figure 21. From the plot, we can see that at 5 the value of SSE begins to decrease indicating that 5 is the ideal number of clusters for this dataset.

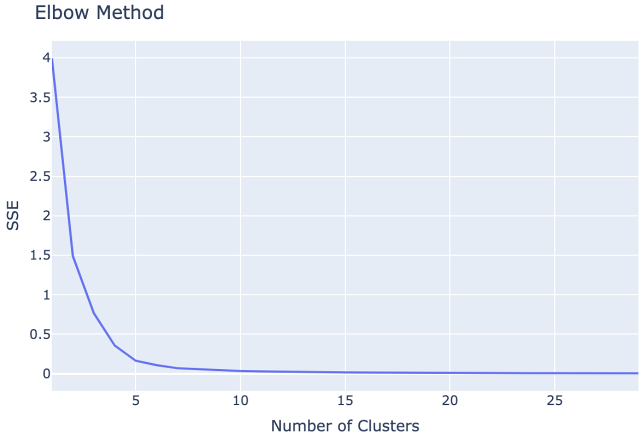


Figure 20. Elbow plot

6.2.2 Clusters Formed:

By setting the value of k to 5 the k-means clustering algorithm was implemented. The scatter plot below shows the cluster each country was assigned to.

Rank	Country	Coal Emission
1	China	113.2969k
2	USA	37.09321k
3	India	21.32379k
4	Japan	8658.594
5	Russia	8326.758
6	South Africa	7736.58
7	Germany	6921.295
8	South Korea	4994.114
9	Poland	4300.983
10	Australia	3774.14

Rank	Country	Gas Emission
1	USA	27.05318k
2	Russia	15.22427k
3	Iran	5380.543
4	China	4716.607
5	Japan	4283.579
6	Canada	3841.138
7	UK	3648.633
8	Germany	3418.595
9	Italy	2933.77
10	Saudi Arabia	2932.787

Rank	Country	Oil Emission
1	USA	47.59442k
2	China	21.06458k
3	Japan	11.12963k
4	India	9091.686
5	Russia	6995.889
6	Saudi Arabia	6097.1
7	Germany	5828.727
8	Brazil	5476.419
9	Mexico	5367.562
10	Canada	5191.643

Table 3. Top 10 countries ranked based on their total million tonnes of carbon dioxide emission for each fossil fuel type.

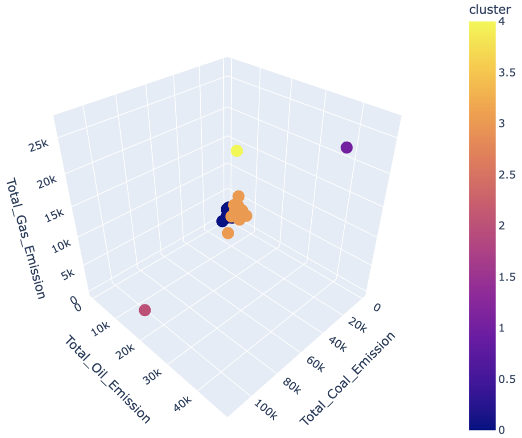


Figure 21. K-means clusters

In order to get a better understanding of the clusters. A words cloud was of country names was created for each cluster. It can be seen in figure 23 that cluster 0 consists of 205 countries, clusters 1, 2, and 4 consist of 1 country each, and cluster 3 consists of 13 countries.

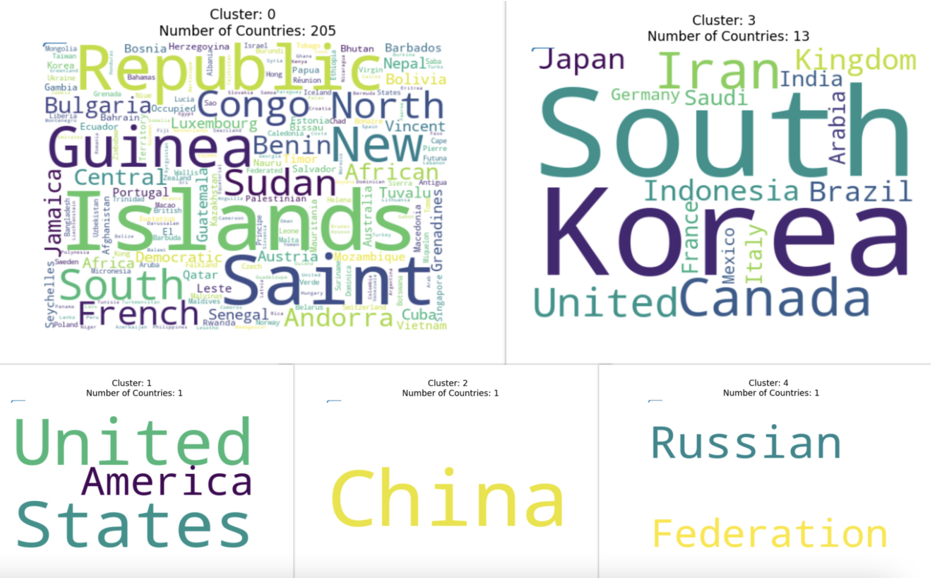
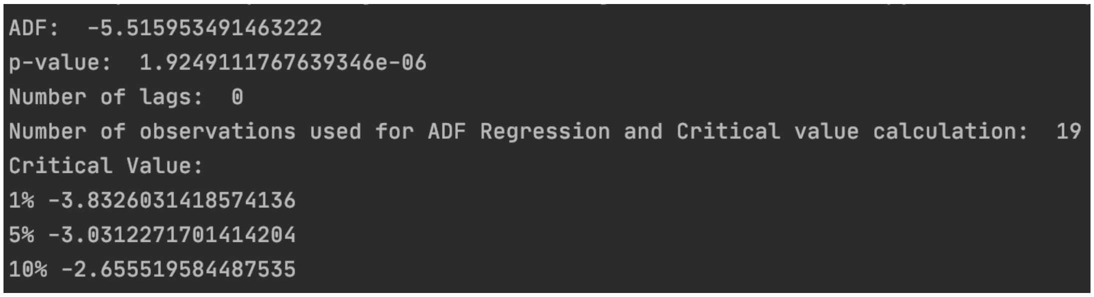


Figure 22. Countries word cloud

6.3 ARIMA Model:

6.3.1 Augmented Dickey-Fuller Test:

In order to implement the ARIMA model we need to check if the data is stationary or not for that, we will be using the ADF (Augmented Dickey-Fuller) test. From the ADF test, we get a p-value of 1.92e-06 which means that we can reject the null hypothesis. Therefore, inferring that the data is stationary



```
ADF:  -5.515953491463222
p-value:  1.9249111767639346e-06
Number of lags:  0
Number of observations used for ADF Regression and Critical value calculation:  19
Critical Value:
1% -3.8326031418574136
5% -3.0312271701414204
10% -2.655519584487535
```

Figure 23. Augmented Dickey-Fuller Test

6.3.2 Forecasting Results:

We will be dividing the dataset into training and testing sets where the last 9 instances will be to test the model. We will be using the RMSE score to validate the model which basically tells us the difference between the predicted and observed values. We got an RMSE score of 0.187 for the model with values (9, 1, 0). From figure 24 we can see the ARIMA model predictions on the test dataset.

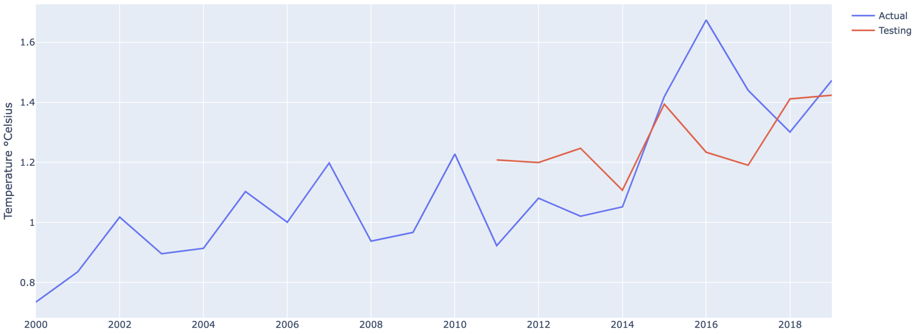


Figure 24. Test set predictions

Using this model we were able to predict the next 10 years of temperature change values which can be seen in figure 25 and table 4.

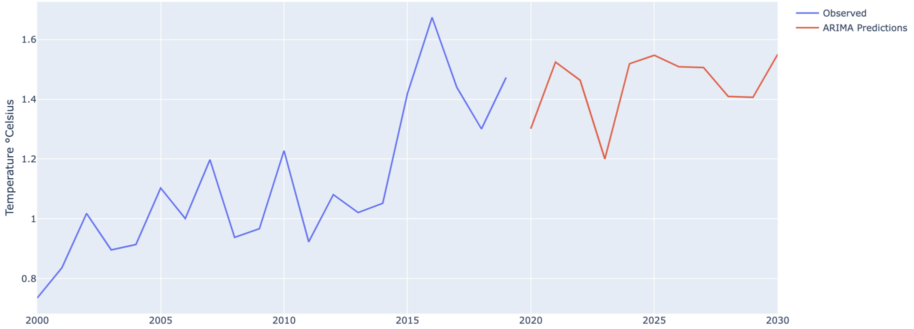


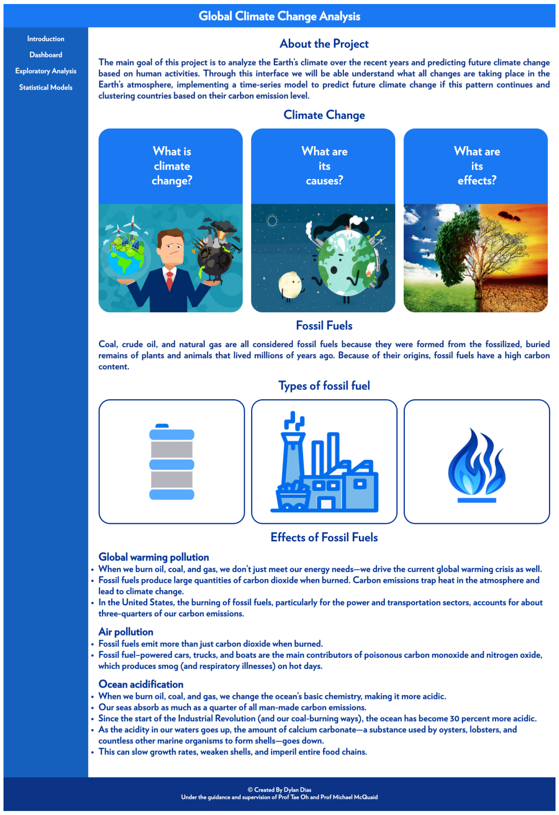
Figure 25. 10 years of temperature change values prediction

Year	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030
Temperature °C	1.30	1.52	1.46	1.20	1.52	1.54	1.51	1.51	1.41	1.406	1.55

Table 4. ARIMA model predictions.

6.4 Visualization Tool:

The final visualization tool consists of 4 main sections introduction section, dashboard section, exploratory analysis section, and statistical models sections. The introduction section as seen in figure 26 consists of a small introduction about the project, what is climate change and its effects, various types of fossil fuels, and the effects of fossil fuels. Figure 27 shows the dashboard that has been implemented. Using this dashboard the user will be able to choose between 2 sets of dashboards i.e. fossil fuels and temperature by using the radio buttons. Each dashboard consists of a set of visualizations that the user can manipulate to compare countries. In figure 28 we can see the exploratory analysis section. In the exploratory analysis section, the user will be able to choose between 2 exploratory analyses i.e. fossil fuels and temperature by using the radio buttons. This section consists of a description of the data along with various queries that have been created to give a better overview of the data. Finally, figure 28 and 29 is the statistical model's section which consists of 2 models that have been created for each dataset. For the fossil fuels, the k-means clustering model has been implemented to group countries based on their emission levels and for the temperature dataset, an ARIMA model was implemented to predict temperature values till 2030.



Global Climate Change Analysis

About the Project

The main goal of this project is to analyze the Earth's climate over the recent years and predicting future climate change based on human activities. Through this interface we will be able understand what all changes are taking place in the Earth's atmosphere, implementing a time-series model to predict future climate change if this pattern continues and clustering countries based on their carbon emission level.

Climate Change

Fossil Fuels

Coal, crude oil, and natural gas are all considered fossil fuels because they were formed from the fossilized, buried remains of plants and animals that lived millions of years ago. Because of their origins, fossil fuels have a high carbon content.

Types of fossil fuel

Effects of Fossil Fuels

Global warming pollution

- When we burn oil, coal, and gas, we don't just meet our energy needs—we drive the current global warming crisis as well.
- Fossil fuels produce large quantities of carbon dioxide when burned. Carbon emissions trap heat in the atmosphere and lead to climate change.
- In the United States, the burning of fossil fuels, particularly for the power and transportation sectors, accounts for about three-quarters of our carbon emissions.

Air pollution

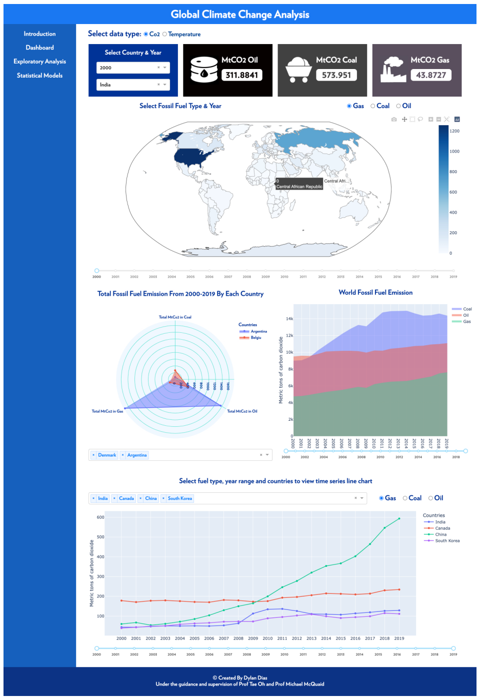
- Fossil fuels emit more than just carbon dioxide when burned.
- Fossil fuel–powered cars, trucks, and boats are the main contributors of poisonous carbon monoxide and nitrogen oxide, which produces smog (and respiratory illnesses) on hot days.

Ocean acidification

- When we burn oil, coal, and gas, we change the ocean's basic chemistry, making it more acidic.
- Our seas absorb as much as a quarter of all man-made carbon emissions.
- Since the start of the Industrial Revolution (and our coal-burning ways), the ocean has become 30 percent more acidic.
- As the acidity in our waters goes up, the amount of calcium carbonate—a substance used by oysters, lobsters, and countless other marine organisms to form shells—goes down.
- This can slow growth rates, weaken shells, and imperil entire food chains.

© Created By Dylan Dias
Under the guidance and supervision of Prof Tae Oh and Prof Michael McQuaid

Figure 26. Introductory section



Global Climate Change Analysis

Introduction
Dashboard
Exploratory Analysis
Statistical Models

Select data type: ● Co2 ○ Temperature

Select Country & Year: 2000, India

MtCO2 Oil: 311.8841

MtCO2 Coal: 573.951

MtCO2 Gas: 43.8727

Select Fossil Fuel Type & Year ● Gas ○ Coal ○ Oil

Total Fossil Fuel Emission From 2000-2019 By Each Country

Denmark, Argentina

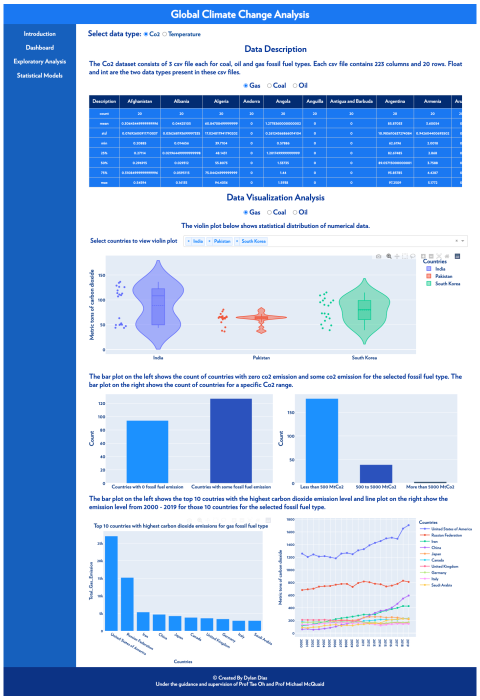
World Fossil Fuel Emission

Select fuel type, year range and countries to view time series line chart

India, Canada, China, South Korea ● Gas ○ Coal ○ Oil

© Created By Dylan Dias
Under the guidance and supervision of Prof Tae Oh and Prof Michael McQuaid

Figure 27. Dashboard section



Global Climate Change Analysis

Introduction
Dashboard
Exploratory Analysis
Statistical Models

Select data type: ● Co2 ○ Temperature

Data Description

The Co2 dataset consists of 3 csv file each for coal, oil and gas fossil fuel types. Each csv file contains 223 columns and 20 rows. Float and int are the two data types present in these csv files.

● Gas ○ Coal ○ Oil

Description	Afghanistan	Albania	Algeria	Andorra	Angola	Anguilla	Antigua and Barbuda	Argentina	Armenia	Aru
count	20	20	20	20	20	20	20	20	20	2
mean	0.30645449999999996	0.04425105	60.84708499999999	0	1.2778360000000002	0	0	85.87053	3.60054	0
std	0.07692600911710037	0.036268193699997335	17.024017941790202	0	0.26124566866014104	0	0	10.985610637274084	0.942604400695502	0
min	0.20885	0.014656	39.7104	0	0.57886	0	0	62.6196	2.0018	0
25%	0.27114	0.021964499999999998	48.1431	0	1.2017499999999999	0	0	82.67485	2.868	0
50%	0.296915	0.029312	55.8073	0	1.33735	0	0	89.05715000000001	3.7588	0
75%	0.31084999999999996	0.0595115	75.04424999999999	0	1.44	0	0	93.85785	4.4287	0
max	0.54594	0.16135	94.4036	0	1.5938	0	0	97.2509	5.1772	0

Data Visualization Analysis

● Gas ○ Coal ○ Oil

The violin plot below shows statistical distribution of numerical data.

Select countries to view violin plot: India, Pakistan, South Korea

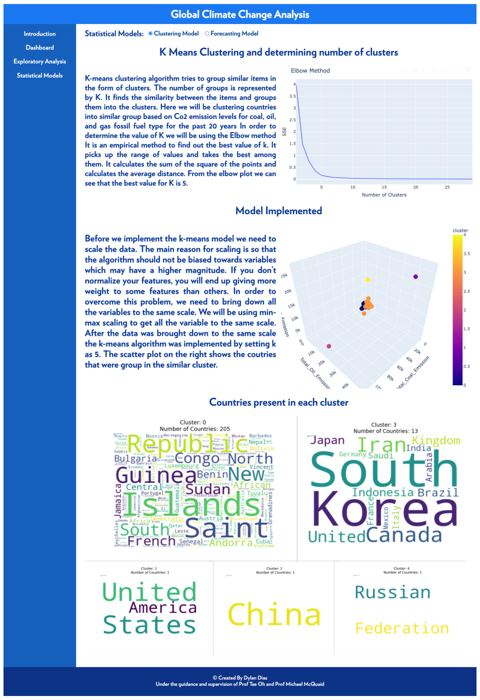
The bar plot on the left shows the count of countries with zero co2 emission and some co2 emission for the selected fossil fuel type. The bar plot on the right shows the count of countries for a specific Co2 range.

The bar plot on the left shows the top 10 coutries with the highest carbon dioxide emission level and line plot on the right show the emission level from 2000 - 2019 for those 10 countries for the selected fossil fuel type.

Top 10 countries with highest carbon dioxide emissions for gas fossil fuel type

© Created By Dylan Dias
Under the guidance and supervision of Prof Tae Oh and Prof Michael McQuaid

Figure 28. Exploratory analysis section



Global Climate Change Analysis
Introduction
Dashboard
Exploratory Analysis
Statistical Models
Statistical Models: ● Clustering Model ○ Forecasting Model

K Means Clustering and determining number of clusters

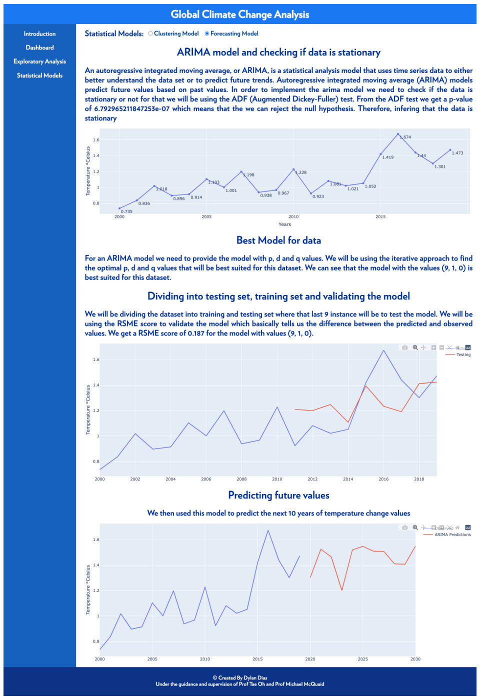
K-means clustering algorithm tries to group similar items in the form of clusters. The number of groups is represented by K. It finds the similarity between the items and groups them into the clusters. Here we will be clustering countries into similar group based on Co2 emission levels for coal, oil, and gas fossil fuel type for the past 20 years In order to determine the value of K we will be using the Elbow method It is an empirical method to find out the best value of k. It picks up the range of values and takes the best among them. It calculates the sum of the square of the points and calculates the average distance. From the elbow plot we can see that the best value for K is 5.

Model Implemented

Before we implement the k-means model we need to scale the data. The main reason for scaling is so that the algorithm should not be biased towards variables which may have a higher magnitude. If you don't normalize your features, you will end up giving more weight to some features than others. In order to overcome this problem, we need to bring down all the variables to the same scale. We will be using min-max scaling to get all the variable to the same scale. After the data was brought down to the same scale the k-means algorithm was implemented by setting k as 5. The scatter plot on the right shows the coutries that were group in the similar cluster.

Countries present in each cluster

Figure 29. K-means clustering section



Global Climate Change Analysis
Introduction
Dashboard
Exploratory Analysis
Statistical Models
Statistical Models: ○ Clustering Model ● Forecasting Model

ARIMA model and checking if data is stationary

An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. Autoregressive integrated moving average (ARIMA) models predict future values based on past values. In order to implement the arima model we need to check if the data is stationary or not for that we will be using the ADF (Augmented Dickey-Fuller) test. From the ADF test we get a p-value of 6.792965211847253e-07 which means that the we can reject the null hypothesis. Therefore, infering that the data is stationary

Best Model for data

For an ARIMA model we need to provide the model with p, d and q values. We will be using the iterative approach to find the optimal p, d and q values that will be best suited for this dataset. We can see that the model with the values (9, 1, 0) is best suited for this dataset.

Dividing into testing set, training set and validating the model

We will be dividing the dataset into training and testing set where that last 9 instance will be to test the model. We will be using the RSME score to validate the model which basically tells us the difference between the predicted and observed values. We get a RSME score of 0.187 for the model with values (9, 1, 0).

Predicting future values

We then used this model to predict the next 10 years of temperature change values

Figure 30. ARIMA section

7 Future Work

ARIMA model makes its predictions using past values. In order to obtain more accurate predictions from the models we can supply it with more historical data. The interface can be made more responsive as in order to make it render on various other devices and windows.

8 Conclusion

The final web interface consists of 4 major sections an introductory section, dashboard, exploratory analysis section, and statistical models section. The introductory section gives a brief overview on what is climate change, causes of climate change, the effects of climate change, fossil fuel type, and its impacts on the environment. The dashboard can be changed between fossil fuel and temperature change data. Each dashboard consists of a number of visualizations with which the user can interact by adding countries of their choice and even changing the year they want to view. The exploratory analysis section gives a brief overview of the fossil fuels and temperature change datasets. This is done with the help of queries generated, visualizations, and descriptions of the data. Finally, the statistical model's section consists of the two major models implemented which are the clustering model and the ARIMA forecasting model. For the clustering model, the k-means model was implemented where we used to elbow plot to determine the value of k as 5, divided the countries into 5 clusters, and created a word cloud of countries for each cluster. Finally for the ARIMA model we used the augmented dickey-fuller test to check if the dataset is stationary or not. A p-value of 1.92e-06 indicated that the dataset is stationary we then used the grid search tuning method to find the most optimum order for the model where the last 9 instances were used to test the model. The model with the order (9, 1, 0) was selected which had the lowest RMSE score of 0.187 and this model was used to predict the next ten years of future climate change values.

References

Al Sayah, M. J., Abdallah, C., Khouri, M., Nedjai, R., & Darwich, T. (2021). A framework for climate change assessment in mediterranean data-sparse watersheds using remote sensing and arima modeling. *Theoretical and Applied Climatology*, *143*(1), 639–658.

Andrew, R., & Peters, G. (2021). The global carbon projects fossil co2 emissions dataset. https://doi.org/10.6084/m9.figshare.16729204.v1

Chalikias, M. S., & Ntanos, S. (2015). Countries clustering with respect to carbon dioxide emissions by using the iea database. *HAICTA*, 347–351.

Climate change: Vital signs of the planet. (n.d.). https://climate.nasa.gov/

El-Mallah, E., & Elsharkawy, S. (2016). Time-series modeling and short term prediction of annual temperature trend on coast libya using the box-jenkins arima model. *Advances in Research*, 1–11.

Girvetz, E. H., Zganjar, C., Raber, G. T., Maurer, E. P., Kareiva, P., & Lawler, J. J. (2009). Applied climate-change analysis: The climate wizard tool. *PLoS One*, *4*(12), e8320.

Goswami, K., Hazarika, J., & Patowary, A. (2017). Monthly temperature prediction based on arima model: A case study in dibrugarh station of assam, india. *International Journal of Advanced Research in Computer Science*, *8*(8).

Kigerl, A. (2016). Cyber crime nation typologies: K-means clustering of countries based on cyber crime rates. *International Journal of Cyber Criminology*, *10*(2).

Kijewska, A., & Bluszcz, A. (2016). Research of varying levels of greenhouse gas emissions in european countries using the k-means method. *Atmospheric Pollution Research*, *7*(5), 935–944.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering. *International Journal*, *1*(6), 90–95.

Lai, Y., & Dzombak, D. A. (2020). Use of the autoregressive integrated moving average (arima) model to forecast near-term regional temperature and precipitation. *Weather and Forecasting*, *35*(3), 959–976.

Nury, A., Koch, M., & Alam, M. (2013). Time series analysis and forecasting of temperatures in the sylhet division of bangladesh. *4th International Conference on Environmental Aspects of Bangladesh (ICEAB), August*, 24–26.

Perkel, J. M. (2018). Data visualization tools drive interactivity and reproducibility in online publishing. *Nature*, *554*(7690), 133–134.

Rahman, A., & Hasan, M. M. (2017). Modeling and forecasting of carbon dioxide emissions in bangladesh using autoregressive integrated moving average (arima) models. *Open Journal of Statistics*, *7*(4), 560–566.

Scholze, M., Knorr, W., Arnell, N. W., & Prentice, I. C. (2006). A climate-change risk analysis for world ecosystems. *Proceedings of the National Academy of Sciences*, *103*(35), 13116–13120.

Swain, S., Nandi, S., & Patel, P. (2018). Development of an arima model for monthly rainfall forecasting over khordha district, odisha, india. *Recent findings in intelligent computing techniques* (pp. 325–331). Springer.

SY, S. (2020). Temperature change. https://www.kaggle.com/sevgisarac/temperature-change/activity

Wang, H., Huang, J., Zhou, H., Zhao, L., & Yuan, Y. (2019). An integrated variational mode decomposition and arima model to forecast air temperature. *Sustainability*, *11*(15), 4018.

Webster, M., Forest, C., Reilly, J., Babiker, M., Kicklighter, D., Mayer, M., Prinn, R., Sarofim, M., Sokolov, A., Stone, P., et al. (2003). Uncertainty analysis of climate change and policy response. *Climatic change*, *61*(3), 295–320.

Zhang, Y., Bilheux, J.-C., Bilheux, H. Z., & Lin, J. Y. (2019). An interactive web-based tool to guide the preparation of neutron imaging experiments at oak ridge national laboratory. *Journal of Physics Communications*, *3*(10), 103003.