SEMESTER 2, 2021/2022

CSCI 4341

Section 01

BIG DATA ANALYTICS

Project Proposal on Daily Temperature of Major Cities

PREPARED BY:

NAME	MATRIC NO.	Email
Mohamed Moubarak Mohamed Misbahou Mkouboi	1820705	mkouboimoubarak18@gmail.com
Umayma Ahhmed	1810704	umaymaahhmed@gmail.com
Malaika Pandit	1917494	malaikaapandit@gmail.com
Mohammad Azwaan	1839075	mohammadazwaan@gmail.com

LECTURER: Dr. SHARYAR WANI

Table of content

Abstract

Our research aims to study and discover the daily temperatures of major cities around the globe. There is a need to analyze how and what factors are contributing to global warming, thus influencing the daily temperatures. In our research, we analyzed an open source dataset and extracted information from the dataset to answer our research questions. Statistical analysis and machine learning are our main methods to conduct this research.

Introduction

The term "global warming" refers to the rise in global average temperatures brought on by the greenhouse effect. When sunlight heats the earth's surface, certain gases in the atmosphere function like glass in a greenhouse, trapping the heat as it radiates back into space. Earth's temperature rises as a result of the atmospheric buildup of greenhouse gases. Climate change, another name for this process, is changing the climate rapidly. It is also acknowledged that those of us who reside in the West or other affluent nations contribute significantly more to this issue than do those who reside in poor nations. The ordinary European contributes almost 2.5 times as much carbon to the atmosphere as a Latin American. Greenhouse gases, variations in the sun's intensity, industrial & agricultural activities, and deforestation are some of the causes of this alarming issue.

Background

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a method of analyzing and manipulating data to extract information, and hidden trends and enhance understanding of the data at hand. EDA helps analyze data to answer research questions, and problems and helps to achieve business goals by making educated decisions based on validated information. Not only that, EDA can also help to correlate variables or attributes to find underlying relationships that lead to new discoveries. In relation to our research, through EDA, we are able to perform statistical analysis on the dataset to find useful information regarding the daily temperature of major cities, thus helping to answer our research questions and problems.

Linear Regression

Linear Regression is an algorithm that helps us to better understand and visualize the relationships between dependent and independent variables. By visualizing how an independent variable affects a dependent one, we are not only able to determine the strength of the relationships between these variables but also possibly predict the value of the dependent variable. As described above, linear regression can be used for gaining predictive insights, which is why we will be using this algorithm to help us predict the daily average temperature amongst different major cities.

Support Vector Machine (SVM)

A recently created classification method called support vector machine (SVM) tries to maximize the margin between two opposing classes in order to solve the classification problem. The addition of kernels is one of the key SVM methods that give it the capacity to handle infinite or nonlinear features in a high-dimensional feature space. In addition to SVM, the kernel trick has been extensively used in a variety of other machine learning algorithms, including principal component analysis (PCA), linear discriminant analysis (LDA), and marginal Fisher analysis (MFA), leading to the development of numerous potent kernel-based learning machines. This will help to forecast the average temperature of major cities in the future.

Long Short Term Memory Networks (LSTMs)

In deep learning, long short-term memory networks are used. Many recurrent neural networks (RNNs) are able to learn long-term dependencies, particularly in tasks involving sequence prediction. Aside from singular data points like photos, LSTM has feedback connections, making it capable of processing the complete sequence of data. This has uses in machine translation and speech recognition, among others. A unique version of RNN called LSTM exhibits exceptional performance on a wide range of issues. Consequently, this will help to obtain accurate and exceptional results in the prediction of the daily temperature of countries around the globe.

Logistic Regression

Predictive analytics and categorization frequently make use of this kind of statistical model, also referred to as a logit model. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. Given that the result is a probability, the dependent variable's range is 0 to 1. In logistic regression, the odds—that is, the probability of success divided by the probability of failure—are transformed using the logit formula.

Research Questions/Hypothesis

- How can we predict the average temperature in the forthcoming years?
- Using Logistic Regression, we can analyze which will be the most affected countries that will experience the highest temperature in the future.

Data Science Questions

➔ Descriptive:
  - What is the total number of regions, countries, and cities that are experiencing average temperature?
  - Which number of months has the most probability of experiencing the lowest average temperature?

➔ Diagnostic:
  - What is the highest and the lowest average temperature Asia experienced in 2020?
  - What are the different countries from 2010 to 2015 in the African region?

➔ Predictive:
  - What are the future results/ analysis of the average temperature?
  - What are the future countries that will have the highest temperature?

➔ Prescriptive:

- Which are the top cities experiencing average temperature in 2005?
- Which are the top regions experiencing the lowest temperature in 2015?
- Which countries are mostly affected by high temperatures during month 12?

➔ Inferential:
  - What is the temperature range of Europe from 2017 to 2019?

➔ Exploratory:
  - How are average temperatures classified in the different regions in May and June?

Research Objectives

- To find a formula for calculating the total number of regions, countries, and cities experiencing average temperature.
- To figure out the different countries in the Asian region from 2010 to 2015
- To provide estimations about the future results of the average temperature.
- To make comparisons between the top countries and regions that have the highest and lowest temperatures.
- To identify and select the temperature range of Europe from 2017 to 2019.
- To arrange the average temperature in terms of different regions in May and June.

Research Significance

The benefit of conducting this research is that first of all, it will help to control the temperature in the future since we can make predictions based on the datasets we have which include the past analysis of the temperature. Moreover, people can take preventative measures so that we can save our world from being destroyed by this alarming and dangerous problem. For instance, we can develop strategies to reduce the contributing factors such as greenhouse gases, deforestation, and so on. The original research problem is concerned with the daily average temperature of the different cities, countries, and regions across the globe. This dataset has been explored in more than

70 notebooks on Kaggle and therefore, it is existing and modified. As for the novelty criteria of my project, the elements leading to the new exploration may be the warmth and expansion of ocean water in volume as a result of rising temperature, causing ice melting and more water flowing to the seas from glaciers and ice caps. This is also an impact of climate change and global warming. Human activities are the root of this phenomenon. Specifically, since the industrial revolution, carbon dioxide and other greenhouse gas emissions have raised temperature, even in the higher poles significantly. Consequently, the research problem is the same as one of the existing ones but we are going to use different Machine Learning algorithms and Deep Learning as it is a heavily explored dataset. The contribution will be substantial and we will get more accurate results about this problem.

Literature Review/Relevant Work

No	Year	Author	Research Problem/ Application	Technique Used	Results	Future Works (if any
1	2018	Harvey Zheng	Analysis of Global Warming Using Machine Learning	Random Forest, Support Vector Regression, Lasso	Four machine learning algorithms that have frequently been shown to deliver satisfactory performance were compared.	Future research can also look into other machine learning techniques, particularly ensemble-bas ed techniques like xgboost and neural networks, to find better models.
2	2018	Paul A. O'Gorman & John G. Dwyer	Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of	RF, Convection Scheme, and Idealized GCM Simulations	Our findings indicate that the application of ML is potential for both the creation of novel	It is still being investigated whether an ML parameterizati on should be nonlocal in space and

			Climate, Climate Change and Extreme Events		parameterizations and for new diagnostics of the interplay between the large-scale subgrid processes and	time. How well convective-momentum trends can be predicted, in addition to the boundary layer, radiation, and large-scale cloud schemes should be used, and whether these should all be replaced
3	2019	Sultan Noman Qasem Saeed Samadianfard, Hamed Sadri Nahand, Amir Mosavi, Shahaboddin Shamshirband, and Kwok-wing Chau	Estimating Daily Dew Point Temperature Using Machine Learning Algorithms	Support Vector Regression (SVR), Gene Expression Programming, M5 Model Tree, and Evaluation Criteria	The results of this study revealed that the SVR-6, using two input parameters of T and RH, and GEP-10 using three parameters of T, RH, and S, had appropriate performance in the estimation of DPT values.	M5-15 is proposed as the most accurate method for the estimation of DPT values at the Tabriz synoptic station, Iran.
4	2020	Zinabu Assefa Alemu & Michael O. Dioha	Climate change and trend analysis of temperature: the case of Addis Ababa, Ethiopia	Mann-Kendall (MK) trend test and Sen's slope estimate	The study showed that both the Mann–Kendall trend test and Sen’s Slope estimator reveals that there is a tendency of temperature	

					increase in the study area. Thus, the increasing trend of temperature due to climate change and other factors can lead to weather extremes in the capital city.	
5.	2020	A V Dergunov, Oleg Yakubailik	Comparative analysis of data on air temperature based on current weather data sets for 2007-2019	Global Forecast System Model,	A detailed analysis revealed that there may be months when the air temperature from different data sets can vary significantly over a period of time. It was found that the value of the air temperature with which the analyzed data sets differ from each other depending on the time of year. In winter, this figure is higher than in summer.	The ERA5 -Land dataset provides hourly or monthly average data with a spatial resolution of about 9 km, but it has a much more limited set of meteorologic al information compared to the GFS model. Modern ERA5 -Land and GFS datasets are not inferior to CRU TS data, and their advantages, in the form of high spatial and

						temporal resolution, can provide better results for solving tasks
6	2020	Himanshu Vishwakarma	Climate Change Analysis Using Machine Learning	Linear regression, support vector machine, lasso regression, Elastic Net	In this study, more than satisfactory accuracy is achieved for the temperature prediction as well as the greenhouse gasses prediction. After the successful prediction, forecasted data for the next 10 years, and also a graphical representation of those predicted and forecasted data with the help of the matplot library is obtained.	The data set is used for this model for 100 – 150 years. But the prediction will be better if the data available is for more than 200 years. In the future, the prediction can be better by using a greater number of data. The other physical entities which are responsible for temperature increase can be also predicted. The data can be predicted for each week as well as each country, which will make this more attractive and reasonable for all people.

7	2021	Marwah Sattar Hanoon Ali NajahAhmed , Nur'atiah Zaini, Arif Razzaq, Pavitra Kumar, Mohsen Sherif, Ahmed Sefelnasr& Ahmed El-Shafe,	Developing machine learning algorithms for meteorological temperature and humidity forecasting at Terengganu state in Malaysia	Gradient boosting tree (G.B.T.), Random forest (R.F.), Artificial neural network (ANN), and Statistical evaluation	The results showed that MLP-NN is a leading algorithm compared with other models in predicting monthly relative humidity with a correlation coefficient equal to 0.8462, followed by RBF-NN with a correlation coefficient equal to 0.7113	future works can be explored by hybridizing the MLP-NN model with optimizers for more accuracies.
8	2021	Olufemi P. Abimbola, George E. Meyer Aaron R. Mittelstet Daran R. Rudnick Trenton E. Franz	Knowledge-guided machine learning for improving daily soil temperature prediction across the United States	ANNs and fuzzy logic, ANFIS, and ANFIS model development	The findings of this study showed that ANFIS-based NT-ACMA models were able to reduce prediction RMSE with increasing soil depth for both Lancaster and Polk weather stations.	Future studies could focus on using new ML algorithms with NT-ACMA feature transformation as a method for predicting soil temperature in the United States and other climates and soils around the world
9	2021	M. Purushotham Reddy, A. Aneesh, K. Praneetha, S. Vijay	Global Warming Analysis, and Prediction Using Data Science	Linear Regression	Linear regression model have been successfully built to predict the average temperatures of	For future work though the linear regression model has been used in the current project to

					Hyderabad. Various visualizations have been done to understand the current situation of global warming. So we can say that global warming is increasing every year and it surely needs to be tackled. Carbon dioxide emissions have also been analyzed in this project	predict the average temperatures, and various other machine learning models can also be used. Root mean squared error can be reduced.
10	2021	Shanza Zia	Climate Change Forecasting Using Machine Learning SARIMA Model	Sarima Model	A straightforward machine learning technique was proposed for the analysis to determine early prediction before climate changes. Multiple linear regressions and the SARIMA Model are used to examine various factors to improve and develop the system. The SARIMA parameters appear to be	

					well-fitted With the test set (baseline vs. extrapolation) in place, RMSE measures the model's standard deviation when it emanates to	

Methodology

1) Datasets used: city_temperature.csv (link: [daily temperature of major cities datasets]())

2) Data preprocessing:

a) Understanding the Data.

Before we begin manipulating the data, we must first assess it to determine what it is about. Each attribute is examined.

b) Data Cleaning

This dataset contains rows with blank values. As a result, the rows are removed to address this issue. The same is true for duplicate rows. This step is critical before proceeding to analyze each research question to avoid bias or error in the analysis.

c) Attribute Selection

There are some characteristics that are unique to each question. As a result, only important attributes relevant to the question at hand should be chosen.

3) Tools used:

a) Google Collab: Platform for Coding
b) Python: Data Pre-processing, Data Analysis, Machine Learning

4) Language used: Python

5) Libraries used:

Libraries	Programming Language	Purpose
numpy	Python	Mathematical Operations
pandas	Python	Data Analysis, Data Pre-processing
matplotlib	Python	Data Analysis, Data Visualization
seaborn	Python	Data Analysis, Data Visualization
sklearn	Python	Machine Learning
Spark	Python	Data Pre-processing, Data Analysis, Data Visualization

6) Algorithm used:

Support Vector Machines (SVM)

Long Short Term Memory Networks (LSTMs)

Linear Regression

Logistic Regression

Results

➔ Descriptive:

- What is the total number of regions, countries, and cities that are experiencing average temperature?

```python
fig = go.Figure()

fig.add_trace(go.Indicator(
    mode = "number",
    value = dftemp['Region'].nunique(),
    domain = {'row': 0, 'column': 0},
    title = {'text': "Total number of Regions"}))

fig.add_trace(go.Indicator(
    mode = "number",
    value = dftemp['Country'].nunique(),
    domain = {'row': 0, 'column': 1},
    title = {'text': "Total number of Countries"}))

fig.add_trace(go.Indicator(
    mode = "number",
    value = dftemp['City'].nunique(),
    domain = {'row': 1, 'column': 2},
    title = {'text': "Total number of Cities"}))

fig.update_layout(
    grid = {'rows': 3, 'columns': 2, 'pattern': "independent"},
    template = {'data' : {'indicator': [{
        'title': {'text': "Speed"},
        'mode' : "number"}]
                    }})

fig.show()
```

- Which number of months has the most probability of experiencing the lowest average temperature?

```
df_temp.groupBy("Month").min("AvgTemperature")
```

Month	min(AvgTemperature)
12	-99.0
1	-99.0
6	-99.0
3	-99.0
5	-99.0
9	-99.0
4	-99.0
8	-99.0
7	-99.0
10	-99.0
11	-99.0
2	-99.0

➔ Diagnostic:
- What is the highest and the lowest average temperature Asia experienced in 2020?

```python
#highest
df_high = df_temp[['Region', 'Year', 'AvgTemperature']]
df_high.filter((df_high.Region == 'Asia') & (
    (df_high.Year == '2020') |
  (df_high.AvgTemperature == 110.0))).show()
```

```
+------+----+--------------+
|Region|Year|AvgTemperature|
+------+----+--------------+
|  Asia|2020|          20.3|
|  Asia|2020|          27.1|
|  Asia|2020|          29.5|
|  Asia|2020|          32.4|
|  Asia|2020|          32.0|
|  Asia|2020|          32.9|
|  Asia|2020|          32.0|
|  Asia|2020|          28.4|
|  Asia|2020|          26.6|
|  Asia|2020|          32.5|
|  Asia|2020|          28.7|
|  Asia|2020|          25.6|
|  Asia|2020|          26.0|
|  Asia|2020|          21.3|
|  Asia|2020|          26.3|
|  Asia|2020|          25.0|
|  Asia|2020|          26.3|
|  Asia|2020|          26.2|
|  Asia|2020|          35.1|
|  Asia|2020|          28.7|
+------+----+--------------+
only showing top 20 rows
```

```
#lowest
df_low = df_temp[['Region', 'Year', 'AvgTemperature']]
df_low.filter((df_low.Region == 'Asia') & (
    (df_low.Year == '2020') &
  (df_low.AvgTemperature == -99.0))).show()
```

```
+------+----+--------------+
|Region|Year|AvgTemperature|
+------+----+--------------+
|   Asia|2020|         -99.0|
|   Asia|2020|         -99.0|
|   Asia|2020|         -99.0|
|   Asia|2020|         -99.0|
|   Asia|2020|         -99.0|
+------+----+--------------+
```

- What are the different countries from 2010 to 2015 in the African region?

```
df_q3 = dftemp[['Region', 'Country', 'Year']]
df_dif = df_temp[['Region', 'Country', 'Year']]
df_dif.filter((df_dif.Region == 'Africa') & (df_dif.Year.between(2010, 2015))).show(2000)
```

```
+------+-------+----+
|Region|Country|Year|
+------+-------+----+
|Africa|Algeria|2010|
|Africa|Algeria|2010|
|Africa|Algeria|2010|
|Africa|Algeria|2010|
|Africa|Algeria|2010|
|Africa|Algeria|2010|
|Africa|Algeria|2010|
|Africa|Algeria|2010|
|Africa|Algeria|2010|
|Africa|Algeria|2010|
```

```
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
|Africa|Algeria|2011|
```

```
|Africa|Algeria|2012|
|Africa|Algeria|2012|
|Africa|Algeria|2012|
|Africa|Algeria|2012|
|Africa|Algeria|2012|
|Africa|Algeria|2012|
|Africa|Algeria|2012|
|Africa|Algeria|2012|
|Africa|Algeria|2012|
|Africa|Algeria|2012|
|Africa|Algeria|2012|
|Africa|Algeria|2012|
|Africa|Algeria|2012|
|Africa|Algeria|2012|
|Africa|Algeria|2012|
|Africa|Algeria|2012|
|Africa|Algeria|2012|
```

```
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
|Africa|Algeria|2013|
```

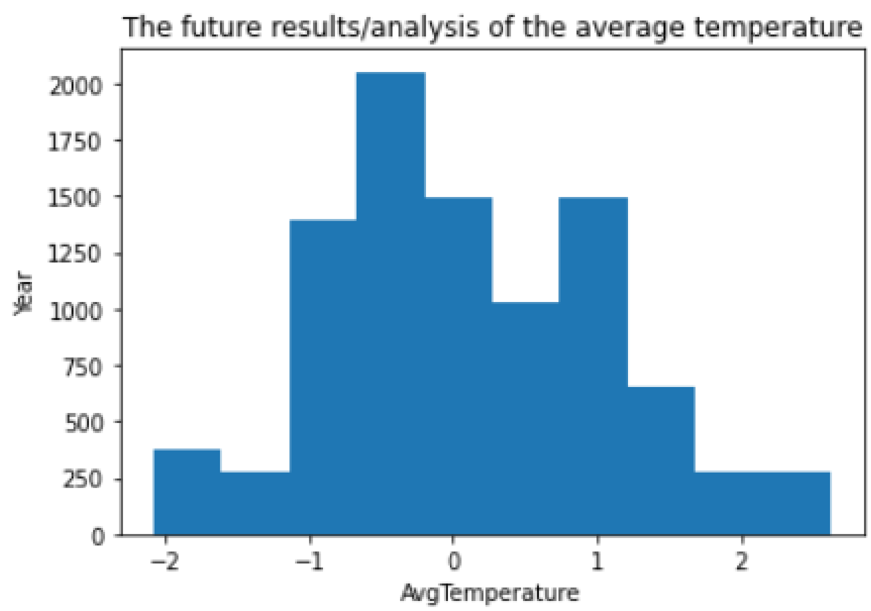
```
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
|Africa|Algeria|2014|
```

```
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
|Africa|Algeria|2015|
```

- Predictive:

- What are the future results/ analysis of the average temperature?

```
from matplotlib.pyplot import hist
dftemp = pd.DataFrame({
    '_id': np.random.randn(100),
    'total': 100 * np.random.rand()
})
hist(dftemp._id, weights=dftemp.total)
# Labels and Titles
plt.title('The future results/analysis of the average temperature')
plt.xlabel('AvgTemperature')
plt.ylabel('Year')
plt.show()
```



- What are the future countries that will have the highest and lowest temperature?

```
df_q5 = df_temp[['Country', 'AvgTemperature']]
df_q5.filter( (
  (df_q5.AvgTemperature == 110.0))).show()
```

```
+-------+--------------+
|Country|AvgTemperature|
+-------+--------------+
| Kuwait|         110.0|
+-------+--------------+
```

```
df_q6 = df_temp[['Country', 'AvgTemperature']]
df_q6.filter( (
  (df_q6.AvgTemperature == -99.0))).show()
```

```
+-------+--------------+
|Country|AvgTemperature|
+-------+--------------+
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
|Algeria|         -99.0|
```

➔ Prescriptive:

- Which are the top cities experiencing average temperature in 2005?

```
df_q7 = df_temp[['City', 'Year', 'AvgTemperature']]
df_q7.filter( (
    (df_q7.Year == '2005')
  )).show()
```

```
+-------+----+--------------+
|   City|Year|AvgTemperature|
+-------+----+--------------+
|Algiers|2005|          50.3|
|Algiers|2005|          46.7|
|Algiers|2005|          47.4|
|Algiers|2005|          48.1|
|Algiers|2005|          45.4|
|Algiers|2005|          43.1|
|Algiers|2005|          43.4|
|Algiers|2005|          44.2|
|Algiers|2005|          42.7|
|Algiers|2005|          45.0|
|Algiers|2005|          44.2|
|Algiers|2005|          43.7|
|Algiers|2005|          49.0|
|Algiers|2005|          51.0|
|Algiers|2005|          44.5|
|Algiers|2005|          43.4|
|Algiers|2005|          43.7|
|Algiers|2005|          46.3|
|Algiers|2005|          56.9|
|Algiers|2005|          56.9|
+-------+----+--------------+
only showing top 20 rows
```

- Which are the top regions experiencing the lowest temperature in 2015?

```
df_q8 = df_temp[['Region', 'Year', 'AvgTemperature']]
df_q8.filter( (
    (df_q8.Year == '2015') &
  (df_q8.AvgTemperature == -99.0))).show()
```

```
+------+----+--------------+
|Region|Year|AvgTemperature|
+------+----+--------------+
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
|Africa|2015|         -99.0|
+------+----+--------------+
only showing top 20 rows
```

- Which countries are mostly affected by high temperatures during month 12?

```
df_q9 = df_temp[['Country', 'Month', 'AvgTemperature']]
df_q9.filter( (
    (df_q9.Month == '12') |
  (df_q9.AvgTemperature == 110.0))).show()
```

```
+-------+-----+--------------+
|Country|Month|AvgTemperature|
+-------+-----+--------------+
|Algeria|   12|          57.3|
|Algeria|   12|          55.9|
|Algeria|   12|          55.1|
|Algeria|   12|          57.7|
|Algeria|   12|          53.8|
|Algeria|   12|          51.8|
|Algeria|   12|          59.3|
|Algeria|   12|          58.2|
|Algeria|   12|          53.2|
|Algeria|   12|          52.6|
|Algeria|   12|          52.9|
|Algeria|   12|          52.1|
|Algeria|   12|          54.4|
|Algeria|   12|          52.3|
|Algeria|   12|          62.7|
|Algeria|   12|          62.2|
|Algeria|   12|          55.6|
|Algeria|   12|          54.1|
|Algeria|   12|          52.1|
|Algeria|   12|          58.1|
+-------+-----+--------------+
only showing top 20 rows
```

➔ Inferential:

- What is the temperature range of Europe from 2017 to 2019?

```
df_q10 = df_temp[['Region', 'Year', 'AvgTemperature']]
df_q10.filter (
    (df_q10.Region == 'Europe') &
  (df_q10.Year.between(2017, 2019))).show()
```

```
+------+----+--------------+
|Region|Year|AvgTemperature|
+------+----+--------------+
|Europe|2017|          35.9|
|Europe|2017|          44.4|
|Europe|2017|          45.3|
|Europe|2017|          46.5|
|Europe|2017|          44.6|
|Europe|2017|          33.6|
|Europe|2017|          26.4|
|Europe|2017|          24.9|
|Europe|2017|          26.9|
|Europe|2017|          25.8|
|Europe|2017|          29.4|
|Europe|2017|          32.4|
|Europe|2017|          43.9|
|Europe|2017|          43.7|
|Europe|2017|          40.6|
|Europe|2017|          42.8|
|Europe|2017|          44.3|
|Europe|2017|          45.4|
|Europe|2017|          44.6|
|Europe|2017|          40.8|
+------+----+--------------+
only showing top 20 rows
```

➔ Exploratory:

- How are average temperatures classified in the different regions in May and June?

```
df_q11 = df_temp[['Region', 'Month', 'AvgTemperature']]
df_q11.filter (
    (df_q11.Month.between(5, 6))).show()
```

```
+------+-----+--------------+
|Region|Month|AvgTemperature|
+------+-----+--------------+
|Africa|    5|          60.4|
|Africa|    5|          59.7|
|Africa|    5|          61.5|
|Africa|    5|          63.9|
|Africa|    5|          63.7|
|Africa|    5|          66.7|
|Africa|    5|          75.1|
|Africa|    5|          77.2|
|Africa|    5|          74.6|
|Africa|    5|          69.4|
|Africa|    5|          67.4|
|Africa|    5|          68.2|
|Africa|    5|          63.3|
|Africa|    5|          60.4|
|Africa|    5|          65.6|
|Africa|    5|          68.9|
|Africa|    5|          68.4|
|Africa|    5|          70.1|
|Africa|    5|          70.3|
|Africa|    5|          68.0|
+------+-----+--------------+
only showing top 20 rows
```

Discussion

The region in the Middle East with the highest temperature from the dataset is Kuwait and the lowest is from the African region which is Tanzania. From our analysis, we found that the main factors that affect continents with the highest and lowest temperature are caused by global warming excluding the testing factors. For that citizens and also the govt needs to take actions against cutting down trees and forest. Therefore, the government of each country needs to prioritize the population's living conditions and focus more on planting trees and stopping deforestation unnecessary.

However, if we consider the average temperature it still comes from the middle east region. So from the last 10 years, it has only increased. The situation is getting only worse. If it continues in the upcoming decades it will be hard for the people to live there. At the same time, the other regions have also increased due to population or deforestation and lack of awareness about global warming.

From our analysis, what we can notice is that the temperature of the regions has shown a significant decrease too for the following regions: Europe, Africa, and North America. However, on the other hand, the highest temperature has almost reached around 70 which is actually a lot than tolerable.

Plantation and taking necessary steps against global warming and deforestation help reduce temperature and it will also cause rainfall in humid areas which will calm down the temperature in certain countries. Regarding the analysis of the temperature rates, what we can observe from the dataset is that continents such as Africa, Europe, and North America had lower temperatures compared to other continents. In the coldest countries with rising temperatures, there is ice melting which is increasing the sea levels. In the previous works, there is research done that proves rising temperature will melt the icebergs and increase the sea level. so in the near future, some countries might get underwater like Maldives and Iceland.

The relationship between global warming, temperature rising, and deforestation is possible because people have become careless. Modernization is taking place. Lots of empty spaces have become developed with the high-rise buildings and industries, factories. A few countries are taking caution and getting used to healthy, greener lifestyles but the developing countries are still struggling and modernization, it's really affecting the temperature and humidity levels. Therefore, countries such as Japan and Brazil, Malaysia, and Canada are adapting to a more eco-friendly, greener environment.

Future Work

One of our possible future implementations to our research project may be the addition of more data that seems to be missing from the dataset we currently have. We will be adding this data into our project whenever it is updated on Our World In Data, or possibly manually too by merging our current dataset with other datasets (from credible sources).

References

[1] Y. P. Trust For the Environment, “Global Warming,” *Young People’s Trust For the Environment*. https://ypte.org.uk/factsheets/global-warming/introduction.

[2] “5 Causes of Global Warming,” *Sciencing*, Nov. 22, 2019. https://sciencing.com/5-causes-global-warming-8232444.html.

[3] Zheng, H., 2018. Analysis of Global Warming Using Machine Learning. https://www.scirp.org/html/5-2570177_86337.htm

[4] P. A. O'Gorman and J. G. Dwyer, “Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events,” *Journal of Advances in Modeling Earth Systems*, vol. 10, no. 10, pp. 2548–2563, 2018.

[5] S. N. Qasem, S. Samadianfard, H. Sadri Nahand, A. Mosavi, S. Shamshirband, and K.-wing Chau, “Estimating daily dew point temperature using machine learning algorithms,” *MDPI*, 20-Mar-2019. https://www.mdpi.com/2073-4441/11/3/582.

[6] M. S. Hanoon, A. N. Ahmed, N. Zaini, A. Razzaq, P. Kumar, M. Sherif, A. Sefelnasr, and A. El-Shafie, “Developing machine learning algorithms for meteorological temperature and humidity forecasting at Terengganu state in Malaysia,” *Nature News*, 23-Sep-2021. https://www.nature.com/articles/s41598-021-96872-w.

[7] “Knowledge‐guided machine learning for ... - Wiley Online Library.” https://acsess.onlinelibrary.wiley.com/doi/epdf/10.1002/vzj2.20151.

[8] “National Geographic,” *National Geographic*. https://www.nationalgeographic.com/.

[9] M. Purushotham Reddy, A. Aneesh, K. Praneetha and S. Vijay, "Global Warming Analysis and Prediction Using Data Science," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud) (I-SMAC), 2021, pp. 1055-1059, doi: 10.1109/I-SMAC52330.2021.9640944.

[10] a V Dergunov, Oleg Yakubailik, A. V. D. O. Y. (2020). Comparative analysis of data on air temperature based on current weather data sets for 2007–2019. *Research Gate*. https://doi.org/10.1088/1755-1315/548/3/032034

[11] Alemu, Z.A., Dioha, M.O. Climate change and trend analysis of temperature: the case of Addis Ababa, Ethiopia. *Environ Syst Res* 9, 27 (2020). https://doi.org/10.1186/s40068-020-00190-5

[12] H.V. (2021). Climate Change Analysis Using Machine Learning. *International Journal of Science and Research (IJSR)*. https://www.ijsr.net/archive/v9i8/SR20722101621.pdf

[13] Shanza Zia, S. Z. (2021). Climate Change Forecasting Using Machine Learning SARIMA Model. *iRASD Journal of Computer Science and Information Technology*. https://journals.internationalrasd.org/index.php/jcsit/article/download/645/414