Data Analysis on Historical Weather Data

Group Zeus

CHAN, Chun Hin (20853893, chchanec@connect.ust.hk)
CHAN, Long Ki (20774516, lkchanar@connect.ust.hk)
CHUNG, Lok Wang (20852241, lwchung@connect.ust.hk)

Project Video Link:

The Youtube video link of our 5-minute project demonstration video is shown here:
https://youtu.be/FonI2BFH_Xw

Background & Aim:

In the current century, climate change is one of the most crucial problems we need to face. The global average temperature has become higher, leading to more frequent natural disasters. Hence, historical weather data would be helpful for us to analyze the change in our climate in recent years. The best way to obtain them would be those collected in different cities in different countries around the world. They are useful for analyzing the climate patterns in meteorology.

Hence, the aim of this project is to investigate the weather data in recent years to observe the weather patterns around the world, and build a machine learning pipeline to do prediction on the weather data.

Tools & Setting:

Google Colaboratory:

Google Colaboratory provides an online platform to develop and run Jupyter Notebook. The default configuration of Google Colaboratory is equipped with 12.7 GB RAM, 107.7 GB ROM and Intel(R) Xeon(R) CPU @ 2.20GHz.

PySpark:

PySpark is a tool that includes the APIs for managing the RDDs, analyzing the dataset, and creating machine learning pipelines for various regression tasks. We will use version 3.5.1 of PySpark.

Introduction:

The dataset is retrieved from Kaggle from this website:
https://www.kaggle.com/datasets/balabaskar/historical-weather-data-of-all-country-capitals

Inside this dataset, there are 11 columns in the column. In summary, they are the information of the date of the record, the country, the city, the location, temperature, wind speed, wind direction and air pressure.

There are 324647 rows of weather data. Notice that the columns **tavg**, **tmin**, **tmax**, **wdir**, **wspd** and **pres** have less than 324647 number of data, indicating that there are some missing values in these columns.

For the number of distinct dates, there are 1745 of them. By considering there are 365 days in a year, this is equivalent to around 4.78 years.

For the number of distinct countries, there are 193 of them. Given that we have 195 countries in total in the world, we can see the dataset covers the weather data of the whole world.

For the number of distinct cities, there are 191 of them.

Data Preprocessing:

We have changed the date column to date data type, wind direction column to integer type, and rest of the data column (except the country and city name) to double data type.

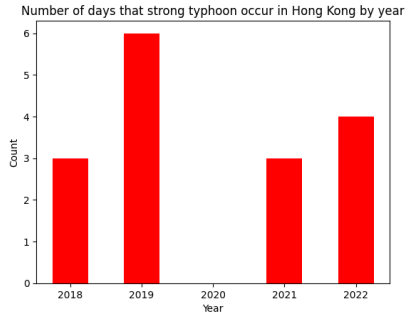
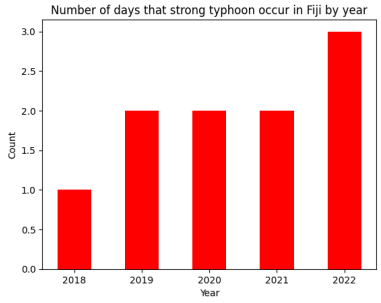
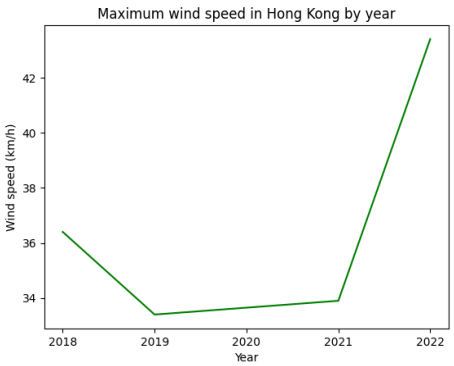
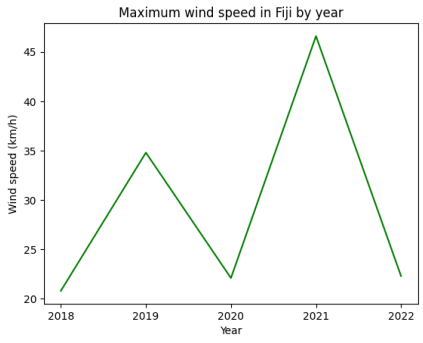
Also, we have removed the rows of data that contain any NULL values.

Analysis on the frequency of typhoons in tropical regions:

Typhoons are one kind of extreme meteorological phenomenon that usually happens in the tropical countries or cities. The characteristic of a typhoon is extremely high wind speed and very low sea-level air pressure. To count the number of possible typhoon days occurring in different regions or countries, we can use the mean-standard deviation method to identify those dates where the typhoons possibly happen.

As there are many tropical countries or cities regarded as tropical areas, we choose Hong Kong and Fiji since they are usually striked by typhoons during summer.

After the analysis, we get the following graphs:

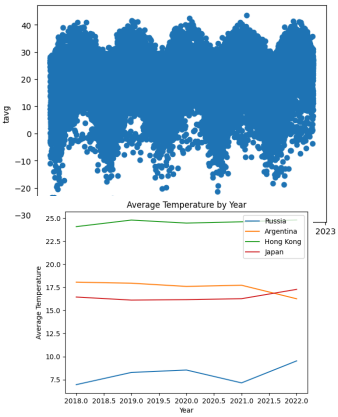
Hong Kong	Fiji
	
	

From the analysis on Hong Kong and Fiji, we observed several things:

(a) For Hong Kong, the number of days that strong typhoons that occur is mostly at least 3 days per year, but there is an overall increase in the maximum wind speed throughout 2018 to 2022, meaning a stronger typhoon in recent years.
(b) For Fiji, the number of days that strong typhoons that occur is no more than 3 days per year, but there is an overall increase in the number of days the strong typhoons occur. However, even though the maximum wind speed in each year alternate each year, this still shows an overall increase in the maximum wind speed, meaning a stronger typhoon in recent years.

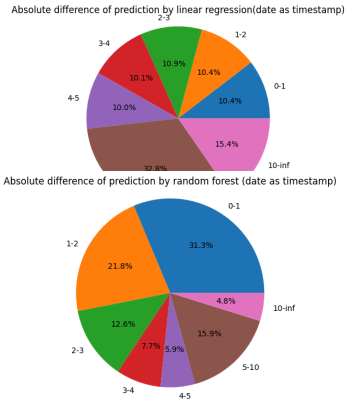
Thus, there is a possible belief that the intensity of the typhoon is becoming stronger and stronger.

Analysis of temperature global warming:



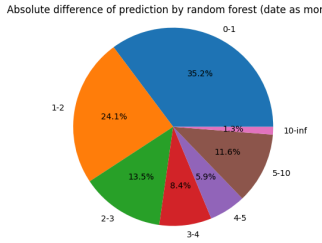
Global warming is always an important issue, therefore we may try to use the data from the dataset to visualize the problem. Therefore we may create various graphs to generate some figures about it and try to understand the issue. To start with, we first look at the scatter graph of tavg (average temperature) to the date. We can see that the graph is not intuitive and difficult to understand. Therefore, we try to make a line graph for three countries (Russia, Japan, and Hong Kong). We can see that all three places have a slight increase in average temperature compared with 2018.
However, with further investigation, it is still questionable whether the increase in temperature is caused by global warming. La Niña happened in 2017-2018 and 2020-2022 and may cause a colder eastern Pacific and a warmer western Pacific. The opposite event of La Niña, El Niño happened during 2018-2019.[1] By looking at the data of Eastern Pacific countries, for example Argentina, we can see that they are having a drop in temperature from 2021.
Then, we use longitude to get the data of the Western Pacific region and the Southern Pacific region. We can see that the eastern Pacific region follows the data of El Niño and La Niña which increased temperature in 2018 and decreased in 2020. But looking at the Western Pacific part, we can see that they always have an increase in temperature which can be an effect of global warming.
Finally, we may also look at the line graph of the average temperature of all the places together by year, we can see that the temperature is increasing from 2018 to 2022. Therefore, we may conclude that even though La Niña and El Niño did affect the temperature, global warming is still pushing the temperature upward.

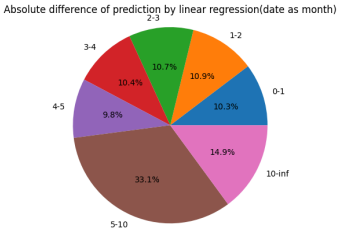
Using linear regression and random forest to predict temperature:



In this part, we have tried to do a regression prediction on the temperature, with the column provided by the dataset, we will use ["Latitude", "Longitude", "wdir", "wspd", "pres", "date"] as our feature column. The column we want to predict will be the tavg. In this task, we will not use column tmax tmin as they are already the temperature and we should not use them to predict tavg. First, we need to vectorize each column into a numerical value, which we first convert the date into a timestamp and other columns to float. We then split our dataset into the train dataset and test dataset.
For the linear regression, we get a Root Mean Squared Error of 8.03 and 4.46 in random forest. The above graph shows the absolute difference between the predicted value and the actual value of the test dataset.

However, in the exploring of data, we find that the month actually has a bigger impact than the year, therefore, in the second trial, we convert the date column into a month (1-12) only and predict again. Then we get a Root mean squared error of 7.99 for linear regression and 3.35 for random forest which is lower than using timestamp. The percentage of predicted value's error of <=3 has also increased for both linear regression(+0.2%) and random forest(+7.1%)





Analysis of Extreme Temperature in Polar and Tropical Countries:

Extreme weather is severe weather or climate conditions that are unusual and have great impacts on our society. It is one of the features of climate change. Heat waves and cold waves are examples of extreme weather. Heat waves are periods of abnormally high temperatures while cold waves are in low temperatures.

We analyse the weather data with the mean-standard deviation method on cities that are located in the Polar Region and Tropical Region to identify the number of days that are in extremely high and low temperatures and investigate the trend of the temperature.

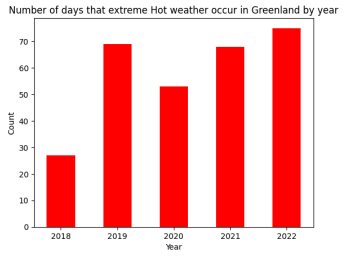
In our analysis, we define the occurrence of extremely high and low temperatures to be:

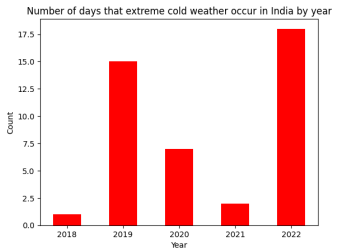
- The value of tmin on a certain date is <= 1.5 times the standard deviation lower than the population mean of the tavg from 2018 to 2022 in the Polar Region Countries as extremely low temperature,
- The value of tmax on a certain date is >= 1.5 times the standard deviation higher than the population mean of the tavg from 2018 to 2022 in the Polar Region Countries as extremely high temperature,
- The value of tmin on a certain date is <= 1.5 times the standard deviation lower than the population mean of the tmin from 2018 to 2022 in the Tropical Region Countries as extremely low temperature, and
- The value of tmax on a certain date is >= 1.5 times the standard deviation higher than the population mean of the tmax from 2018 to 2022 in the Tropical Region Countries as extremely high temperature

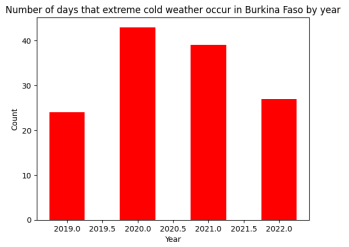
We have picked Greenland and Russia for analysing extreme weather in Polar Region Countries and India and Burkina Faso for analysing extreme weather in Tropical Region Countries.

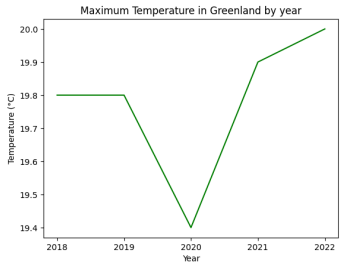
The results are as follows:

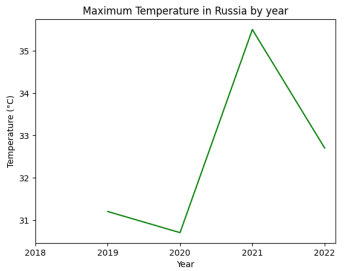
Extremely hot weather			
Greenland	Russia	India	Burkina Faso

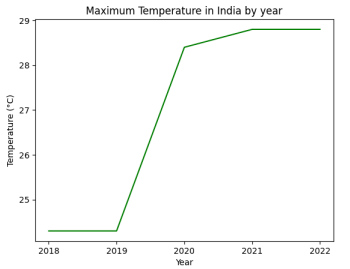


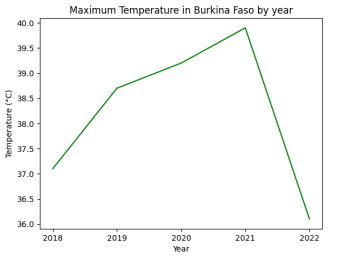






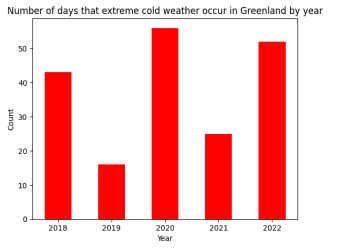


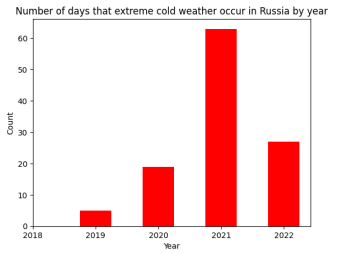


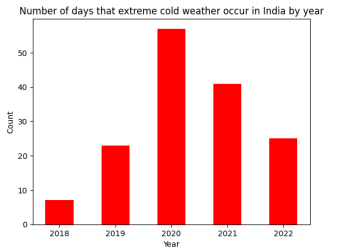


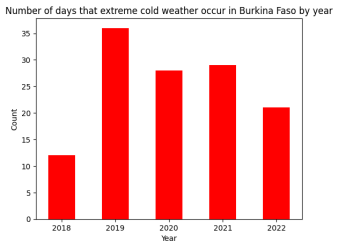
Extremely cold weather

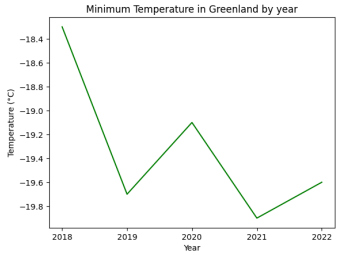
Greenland	Russia	India	Burkina Faso

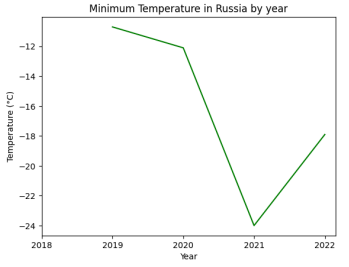


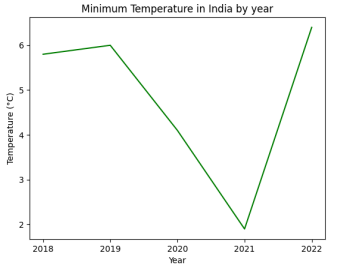


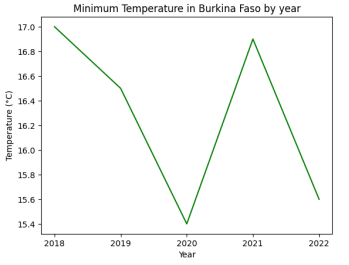












From the results, we have the following observations:

- The increase in maximum temperatures in both Polar Countries and Tropical Countries has shown that extreme heat weather has higher and higher temperatures.
- The decreasing minimum temperatures in both Polar Countries and Tropical Countries have also shown that we are having much colder temperatures in extreme cold weather.
- The increase in the number of days that extreme hot/cold weather has also suggested that heat waves and cold waves are more frequent and occur for a longer period.

Reference:

[1]http://www.bom.gov.au/climate/enso/#tabs=Pacific-Ocean&pacific=History&enso-impacts=La-Ni%C3%B1a-impacts