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)

Demo Video Link

Demo Video Link

- The YouTube video link of our 5-minute video is as shown below:
- https://youtu.be/FonI2BFH_Xw

Background & Aim

Background

- Climate change is a serious problem in current century
- Global average temperature increasing
- More frequency natural disasters

- Historical weather data is useful for analyzing climate change
- Obtain the weather data from different locations around the world
- Analyze the climate patterns

Aim

- Investigate the weather data in the recent years

- Analyze the weather patterns around the world
- Build a machine learning pipeline to do prediction on the weather data

Tools & Setting

Google Colaboratory

- an online platform to develop and run Jupyter Notebook

- Setting:
- RAM: 12.7 GB
- ROM: 107.7 GB
- CPU: Intel(R) Xeon(R) CPU @ 2.20GHz

PySpark

- Provides different APIs for:
- managing the RDDs
- analyzing the dataset
- creating machine learning pipelines for various regression tasks

- Setting:
- Version: 3.5.1

Introduction

Dataset

- Source: https://www.kaggle.com/datasets/balabaskar/historical-weather-data-of-all-country-capitals
- 324647 rows of weather data
- 11 columns:

date	The date where the weather data is recorded
country	The name of the country
city	The name of the city
Latitude	The latitude value of the city location
Longitude	The longitude value of the city location
tavg	The average air temperature in °C
tmin	The minimum air temperature in °C
tmax	The maximum air temperature in °C
wdir	The average wind direction in degrees (°)
wspd	The average wind speed in km/h
pres	The average sea-level air pressure in hPa

Dataset

- 1745 distinct dates
- 193 distinct countries
- 191 distinct cities

```
Number of distinct dates:
+--------------------+
|count(DISTINCT date)|
+--------------------+
|1745                |
+--------------------+
```

```
Number of distinct countries:
+-----------------------+
|count(DISTINCT country)|
+-----------------------+
|193                    |
+-----------------------+
```

```
Number of distinct cities:
+---------------------+
|count(DISTINCT city)|
+---------------------+
|191                  |
+---------------------+
```

Data Cleaning/Preprocessing

Data Preprocessing

- Convert the data type of some of the columns in the dataset

- string -> date: “date”
- string -> integer: “wdir”
- string -> double: “Latitude”, “Longitude”, “tavg”, “tmin”, “tmax”, “wspd”, “pres”

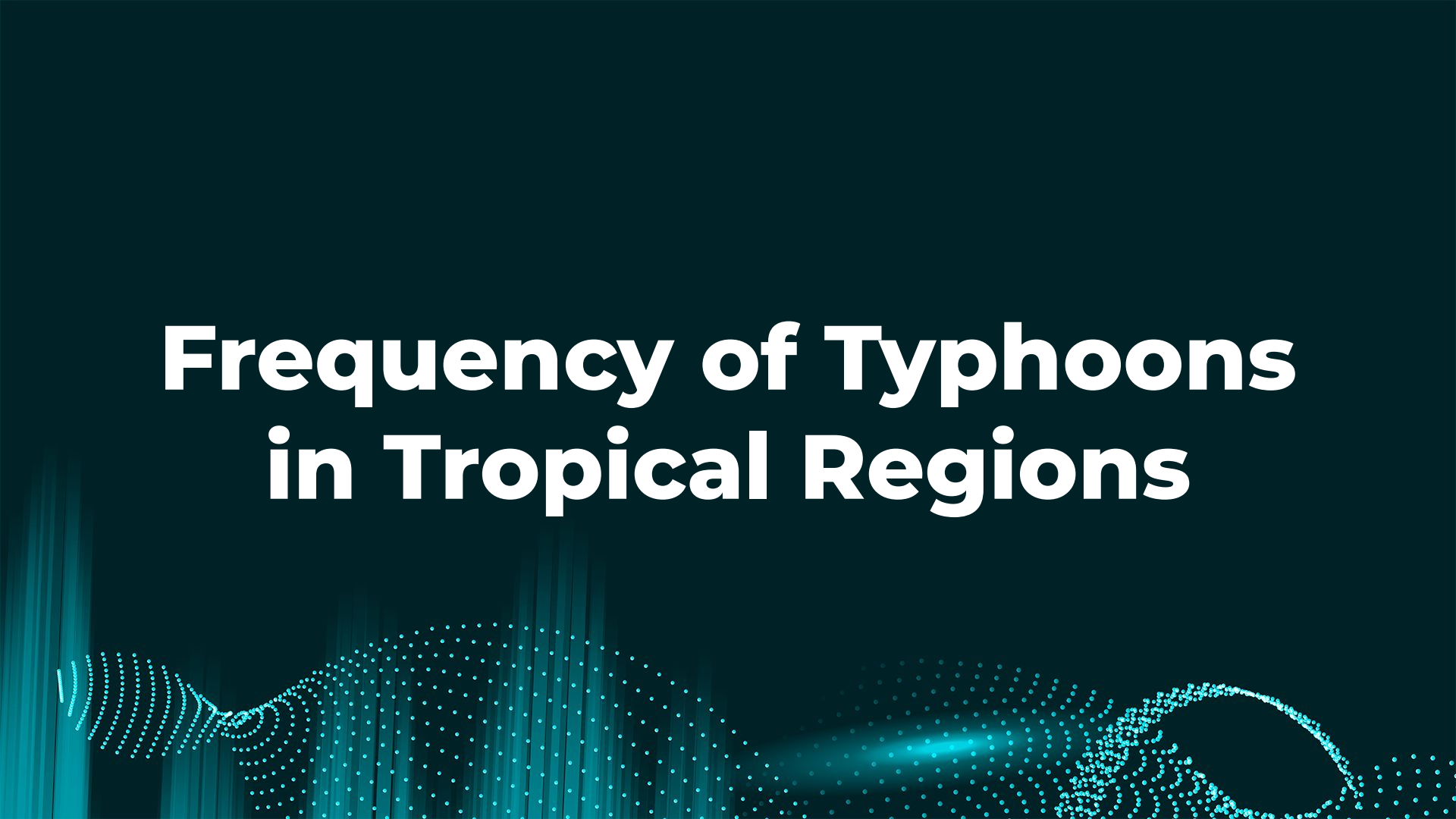
```
root
 |-- date: date (nullable = true)
 |-- country: string (nullable = true)
 |-- city: string (nullable = true)
 |-- Latitude: double (nullable = true)
 |-- Longitude: double (nullable = true)
 |-- tavg: double (nullable = true)
 |-- tmin: double (nullable = true)
 |-- tmax: double (nullable = true)
 |-- wdir: integer (nullable = true)
 |-- wspd: double (nullable = true)
 |-- pres: double (nullable = true)
```

Data Cleaning

- Remove the rows of data that contains any NULL values in any column
- 277551 rows of data remained after data cleaning

summary	country	city	Latitude	Longitude	tavg	tmin	tmax	wdir	wspd	pres
count	277551	277551	277551	277551	277551	277551	277551	277551	277551	277551
mean	NULL	NULL	20.756738623535117	12.14380587077041	20.817601089529468	17.223447222312295	24.816747192407835	164.4588454013857	13.446140348980911	1013.2937553818939
stddev	NULL	NULL	25.993528595821584	72.82123478430485	9.350438694717448	9.403817706778403	9.85308045415029	102.06904494374895	7.332545983445811	7.194732279250438
min	Abkhazia	Abu Dhabi	-54.43	-176.176447	-29.6	-31.7	-26.0	0	0.0	922.8
max	Western Sahara	Zagreb	78.062	179.198128	44.1	38.5	89.6	360	74.1	1058.0

Frequency of Typhoons in Tropical Regions



Frequency of Typhoons in Tropical Regions

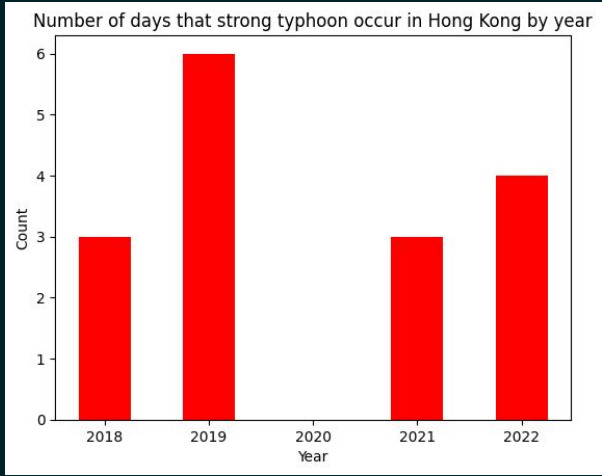
- Characteristics of typhoon:
  - uncommonly high wind speed on a certain date
  - uncommonly low sea-level air pressure on a certain date

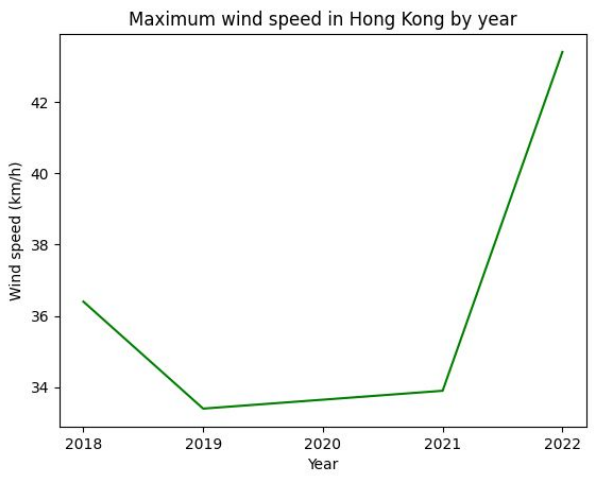
- Selected Hong Kong and Fiji as the tropical countries/cities

- Definition of typhoon(s) possibly occurred on a certain date:
  - wspd: >= 1.5 times the standard deviation greater than the mean of wspd
  - pres: <=1.5 times the standard deviation smaller than the mean of pres

Frequency of Typhoons in Tropical Regions

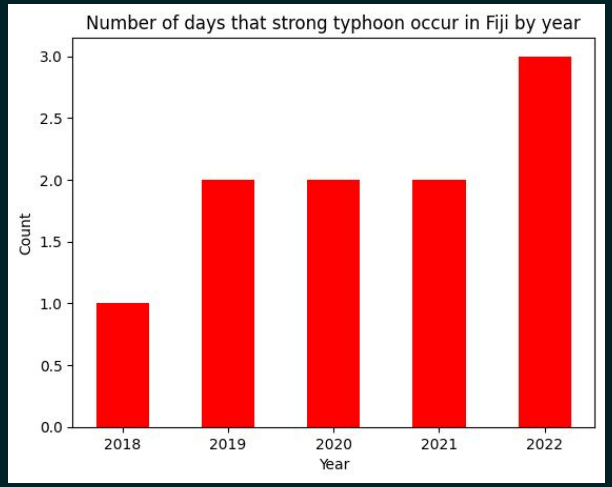
- Hong Kong:

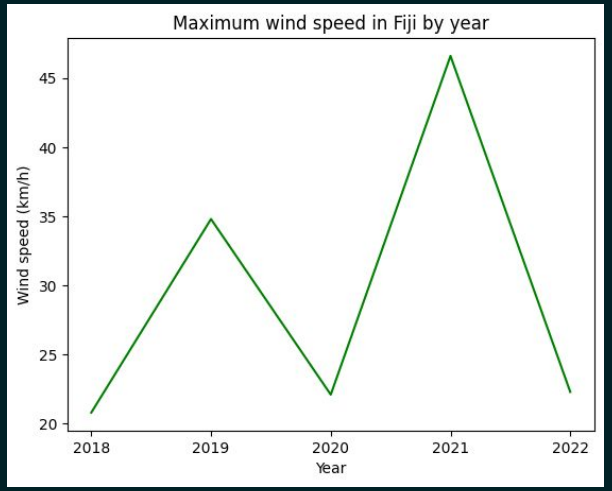




Frequency of Typhoons in Tropical Regions

- Fiji:





Frequency of Typhoons in Tropical Regions

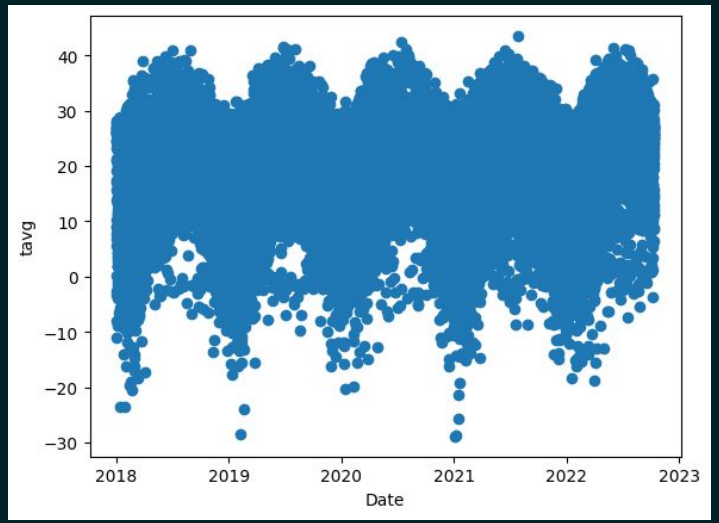
- Key observations:
  - Hong Kong: number of days that strong typhoons occur is mostly at least 3 days per year
  - Hong Kong: an overall increase in the maximum wind speed throughout 2018 to 2022, meaning a stronger typhoon in recent years
  - Fiji: number of days that strong typhoons that occur is no more than 3 days per year
  - Fiji: an overall increase in the number of days the strong typhoon occur
  - Fiji: 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

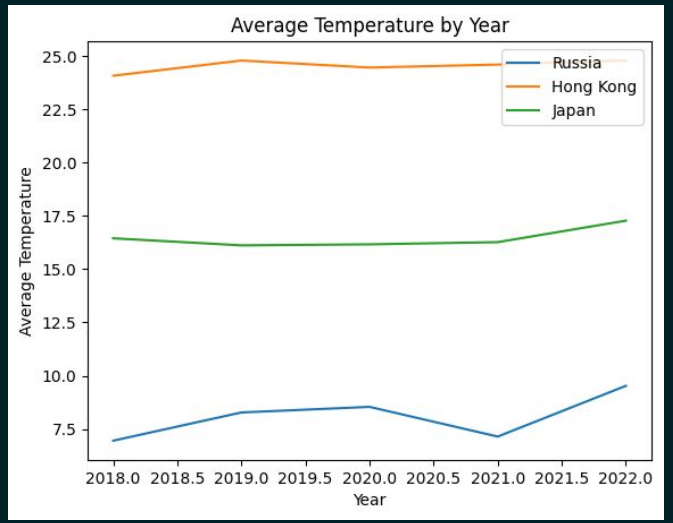
- Short summary: intensity of the typhoon is becoming stronger and stronger.

Global Warming

Global warming

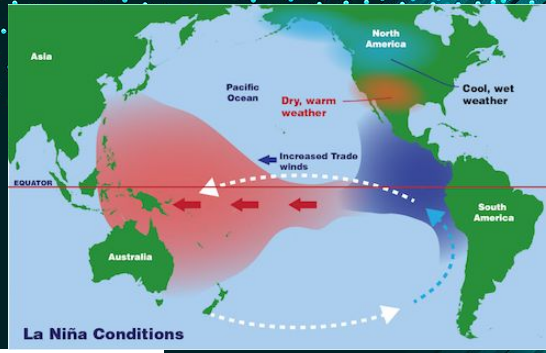
- Scatter graph of tavg of different time, not intuitive
- Changes to line chart of some selected places, did show increase in temperature, but..

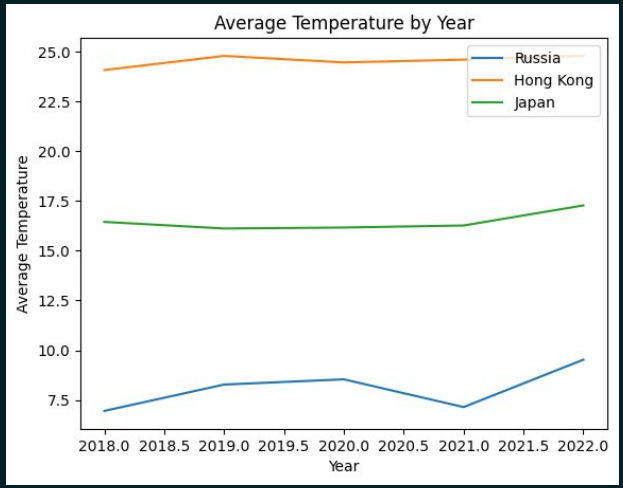


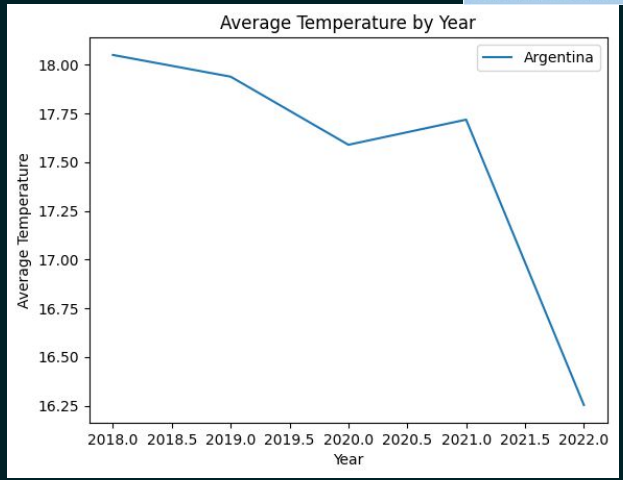


Global warming may not be the major reason for the previous countries

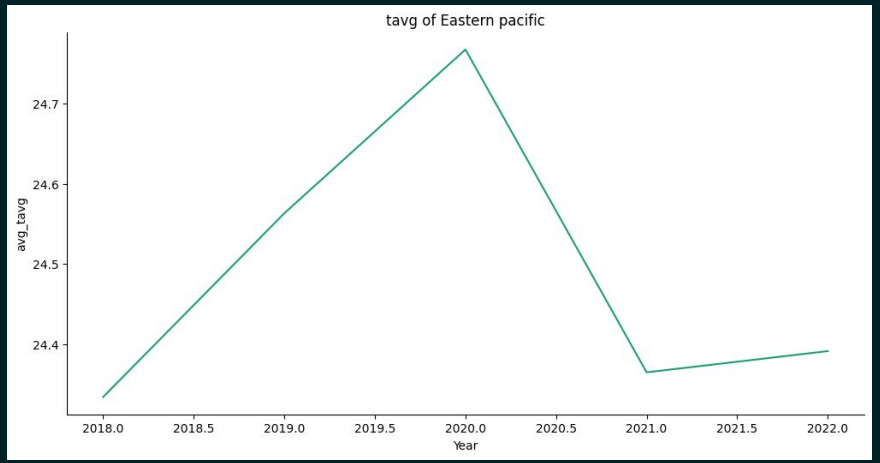
- La Niña cause a colder eastern Pacific and a warmer western Pacific

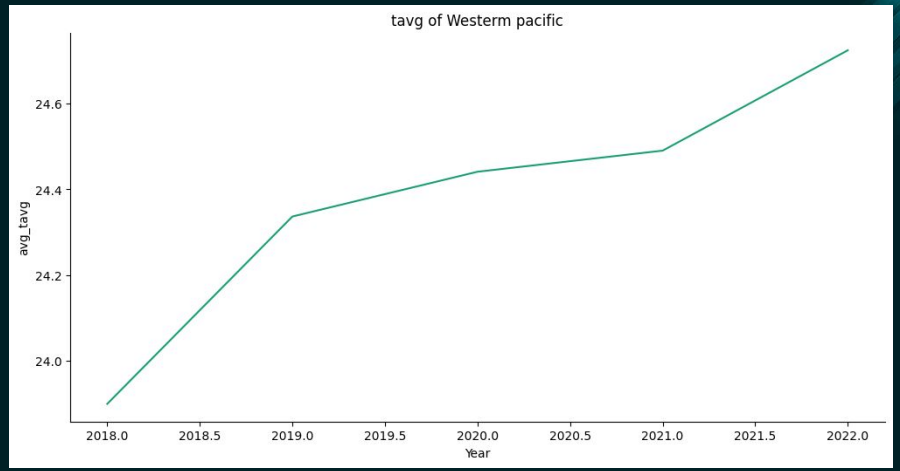




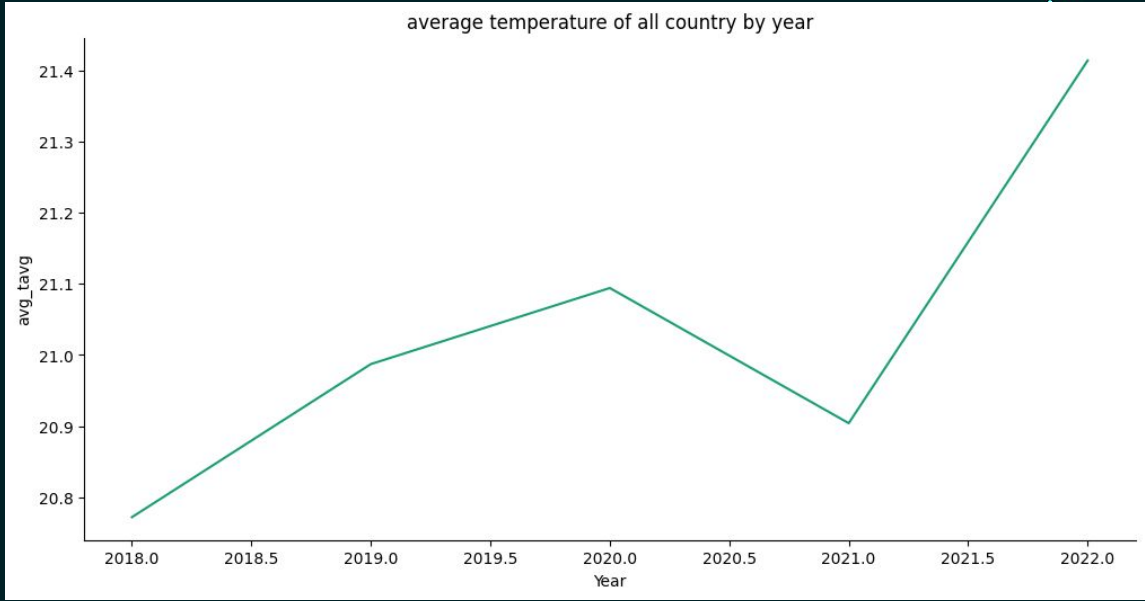


- La Niña cause a colder eastern Pacific and a warmer western Pacific
- happened in 2020-2022
- The opposite event, El Niño happened during 2018-2019
- Eastern pacific follow the La Niña and El Niño
- Western pacific area temperature keep increase





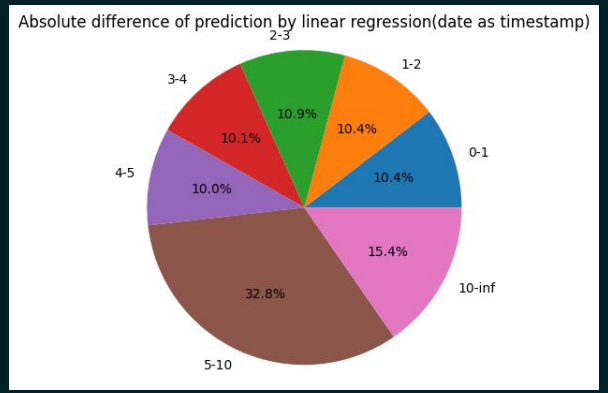
Overall, the temperature keep increase

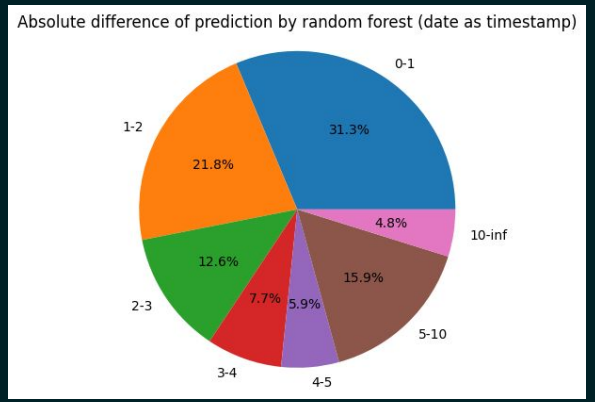


Linear Regression & Random Forest for Temperature Prediction

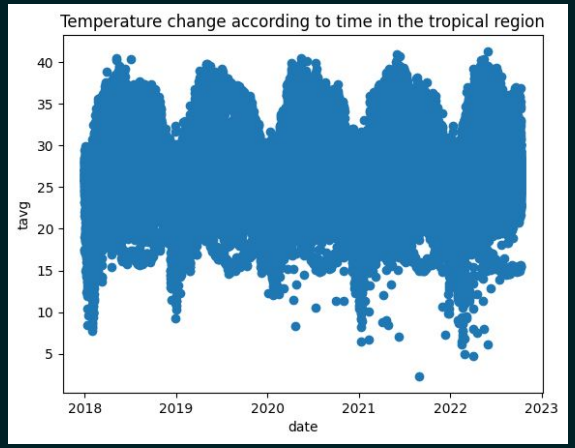
Regression problem on predicting temperature

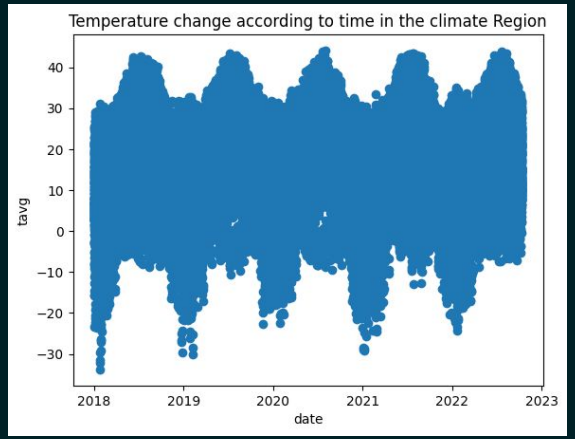
- Feature column: ["Latitude", "Longitude", "wdir", "wspd", "pres", "date"]
- Using timestamp to convert date column
- User float to convert other columns
- RMSE:8.03(linear regression)4.46(random forest)

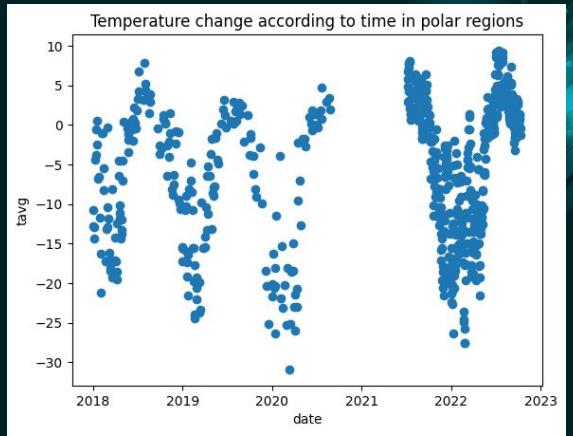




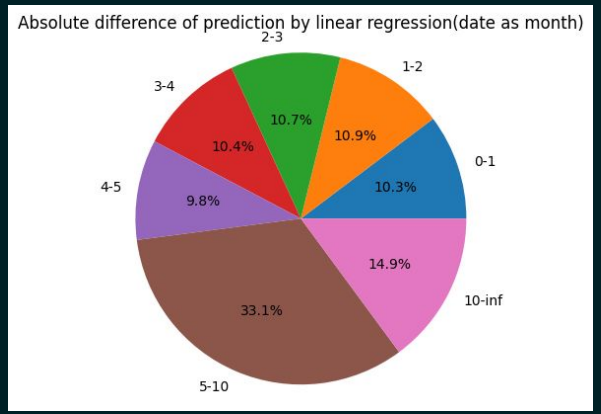
The temperature not heavily dependent on year but on month

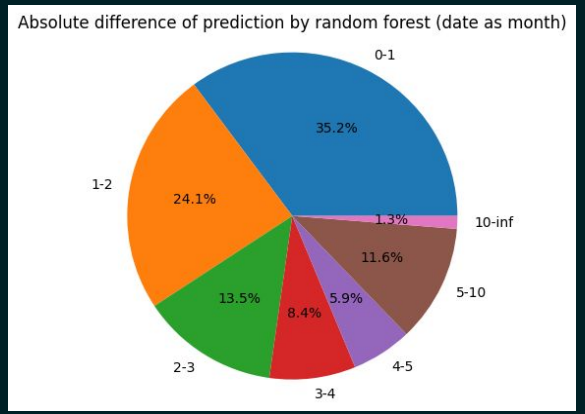




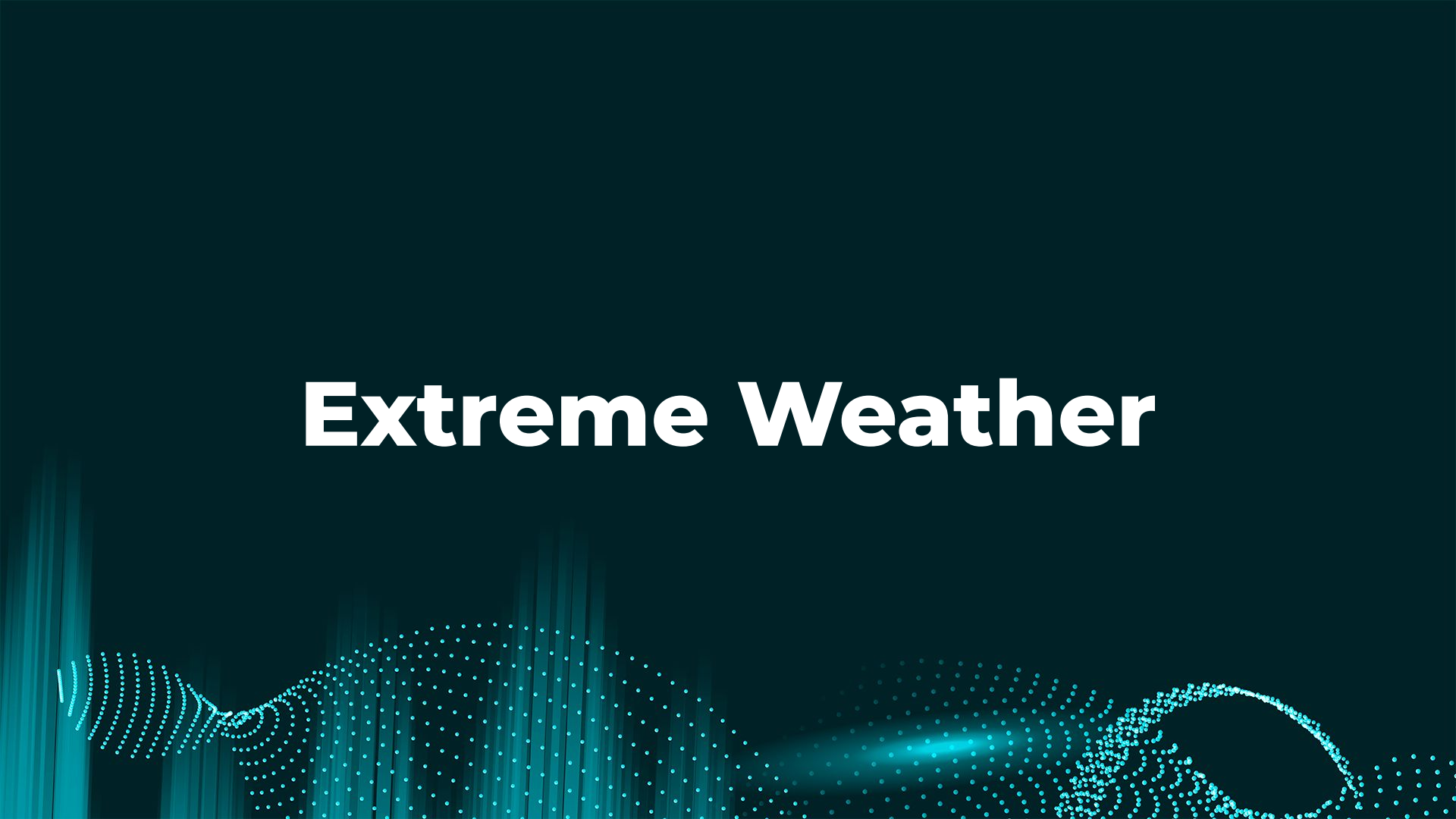


- Use month of the date to encode the date column
- RMSE:7.99(linear regression)3.35(random forest)
- Getting lower RMSE score
- The change of percentage of absolute difference of (0-3)
    - 0.2% increase in linear regression
    - 7.1% increase in random forest





Extreme Weather



Extreme Weather

Features:

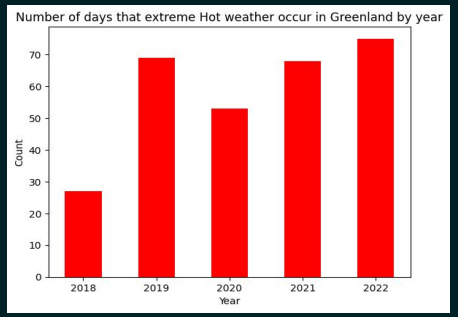
- Severe weather or climate conditions
- Unusual
- Devastating impact on society
- Caused by climate change

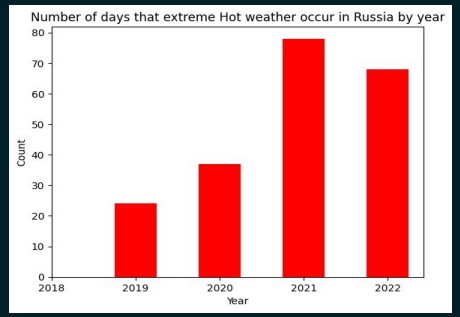
Our Analysis:

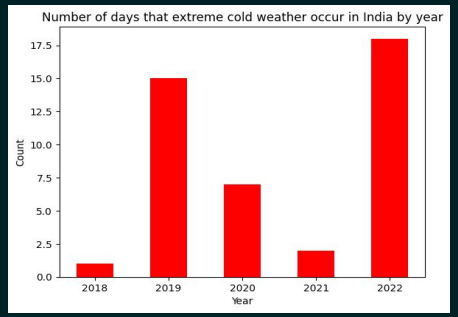
- Extremely hot weather (Heat Waves)
- Extremely cold weather (Cold Waves)
- Studied Greenland and Russia in Polar Region
- Studied India and Burkina Faso in Tropical Region

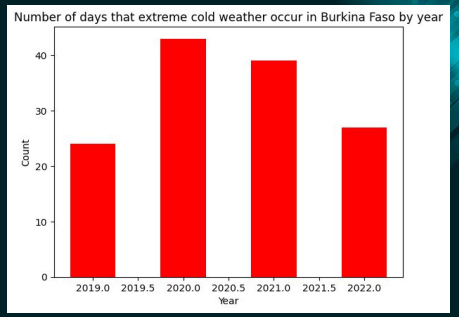
Occurrence of Extreme Weather

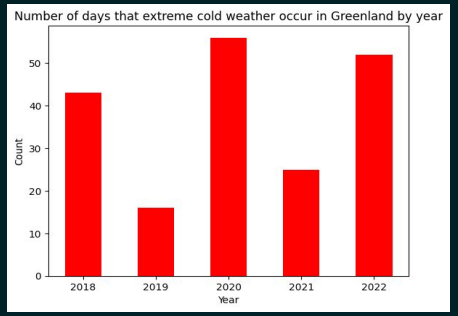
The increase in the number of days that extreme hot/cold weather has suggested that heat waves and cold waves are more frequent and occur for a longer period

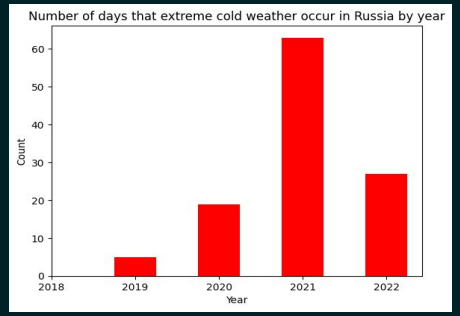


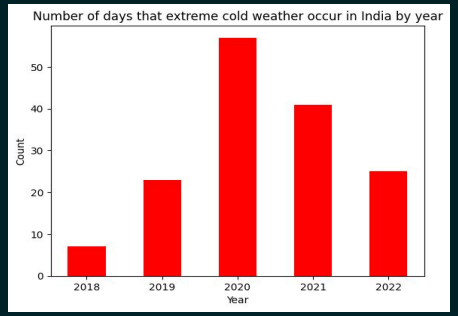


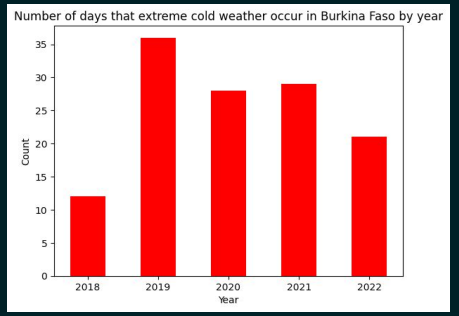








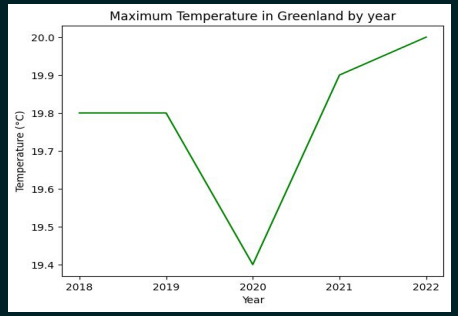


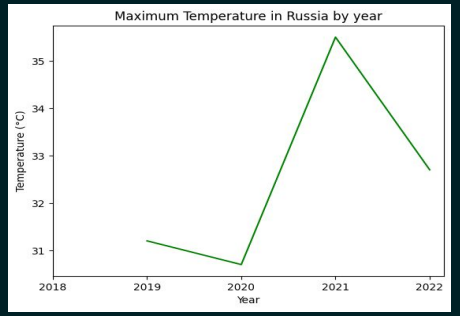


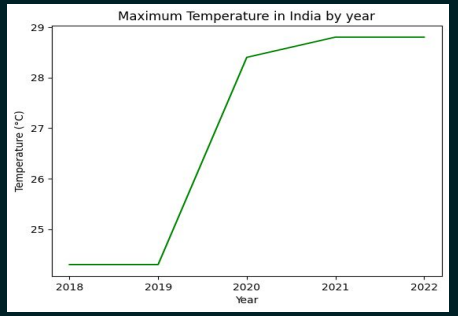
Trend of Extreme Weather Temperature

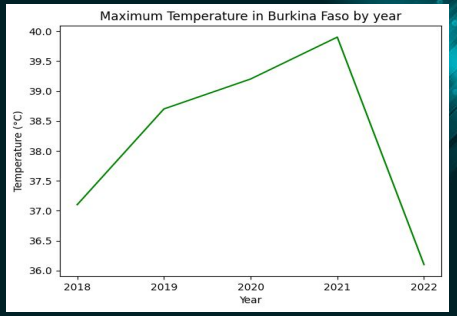
The increase in maximum temperatures in both Polar Countries and Tropical Countries has shown that extreme heat weather has higher and higher temperatures.

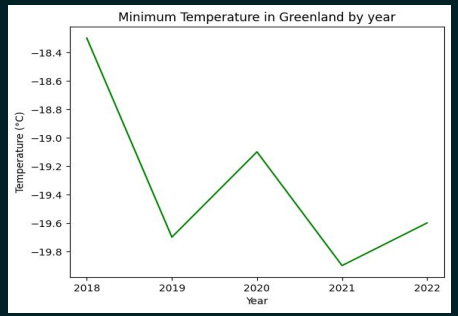
The decreasing minimum temperatures in both Polar Countries and Tropical Countries have also shown that we are having much colder temperatures in extreme cold weather.

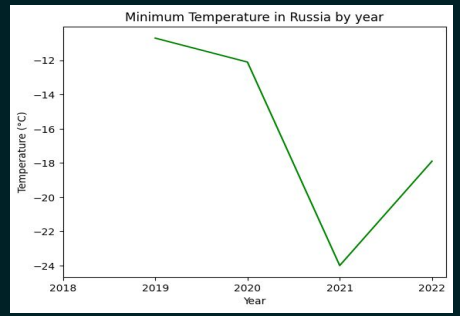


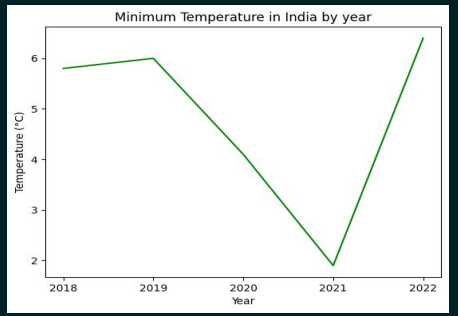


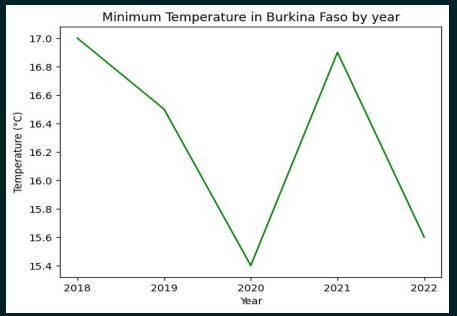












Reference

Reference

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

The End