A Data Research on Climate Change

Ning Shangguan

Climate change is a hot research topic widely discussed in various media. It also becomes an important political topic discussed by politicians. Is climate change real? How does the climate change look like? If the climate change is happening and the earth becomes warmer, why my place is still so cold in winter and has so much snow every year? Why disastrous snowstorms hit Texas this year and caused so much damage if the climate change is real? Is the CO_2 emission from fossil fuel causing the climate change?

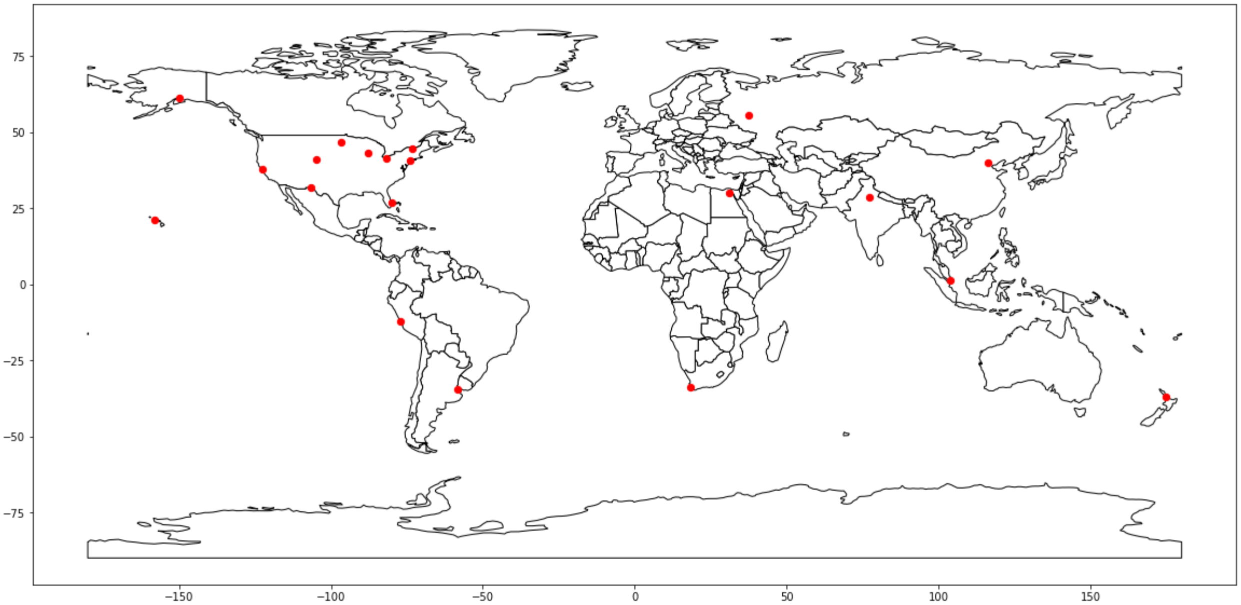
To answer those questions, one way is to do some historical data research. Data can tell us if the climate change is real and what the climate change look like.

1. Data

University of Dayton provides a collection of daily mean temperature of 157 US cities and 167 international cities from 1995 to 2015 at https://academic.udayton.edu/kissock/http/weather/

The daily mean temperature data of each city can be read merged into a Pandas dataframe. Totally 19 cities (9 cities in contiguous US, 2 cities in Alaska and Hawaii, plus 9 international cities) were selected (**Figure 1**） and their temperature data were read and combined into a pandas dataframe table.

Figure 1. 20 cities selected for the temperature data research



Additionally, a 150 years of monthly mean temperature of New York city data (www.weather.gov/media/okx/Climate/CentralPark/monthlyannualtemp.pdf) was found and added to this temperature research. The historical data of CO2 concentration can be found at https://ourworldindata.org/grapher/global-co-concentration-ppm.

2. Data Cleaning

University of Dayton provided a temperature dataset in text format for each city. The daily mean temperature data of each city could be read into Pandas dataframe by using pandas. read_table function. I converted the columns of year, month and day to one column of datetime data and used it as an index. The temperature of 19 cities could be merged into a whole Pandas dataframe by the datetime column. Finally, I had a 7627×19 Pandas table.

```
<class 'pandas.core.frame.DataFrame'>
Index: 7627 entries, 1995-01-01 to 2015-11-18
Data columns (total 20 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   AK_Anchorage_Temp        7614 non-null   float64
 1   ND_Fargo_Temp            7602 non-null   float64
 2   FL_West_Palm_Beach_Temp  7602 non-null   float64
 3   Vermont_Burlington_Temp  7605 non-null   float64
 4   TX_El_Paso_Temp          7608 non-null   float64
 5   Wyoming_Cheyenne_Temp    7611 non-null   float64
 6   San_Francisco_Temp       7591 non-null   float64
 7   Cleveland_Temp           7613 non-null   float64
 8   Milwaukee_Temp           7572 non-null   float64
 9   New_York_City            7607 non-null   float64
 10  Honolulu_Temp            7609 non-null   float64
 11  SA_Cape_Town             7610 non-null   float64
 12  Egypt_Cairo              7607 non-null   float64
 13  India_Delhi              7591 non-null   float64
 14  China_Beijing            7614 non-null   float64
 15  Singapore_Temp           7613 non-null   float64
 16  New_Zealand Auckland     7584 non-null   float64
 17  Russia_Moscow            7613 non-null   float64
 18  Argentina Buenos_Aires   7608 non-null   float64
 19  Peru_Lima                7606 non-null   float64
dtypes: float64(20)
memory usage: 1.2+ MB
```

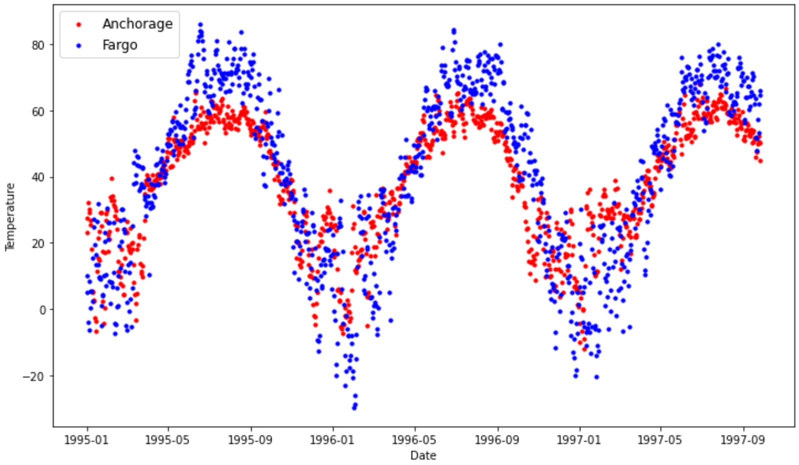
The missing values could not be found by using the isnull() function. Instead, the missing values were shown as -99 °F, an unlikely low temperature. The -99 value was replaced with null. Each city had 13-55 missing values, compared with the 7627 total values, the missing values did not account for a significant part. I used the linear interpolate method to fill the null values.

For the 150 years of history of New York City (Weather.gov data) data, the data file was in PDF form, so I converted it from PDF to Excel. The excel file was read into the Pandas dataframe. There is no missing value in the dataset and it only contains monthly mean temperature data. The dataframe was melted and converted to the same format as the first dataframe. It is an 1812×2 rows.

3. Data Wrangling

Looking at the visualization of the daily mean temperature of cities, I only found using scatter plot generates too many dots and makes it difficult to see clearly. Figure 2 only shows the daily mean temperature of two and half year of two cities

Figure 2. Daily Mean Temperature of Anchorage and Fargo Between 1995 and 1997



By taking the temperature of monthly mean, it is much easier to see the temperature changes over the years. **Figure 3** shows the monthly mean temperature of 3 cities (Anchorage, Fargo and Cleveland) between 1995 and 2000.

Figure 3. Monthly Mean Temperature of Anchorage, Fargo and Cleveland between 1995 and 2000

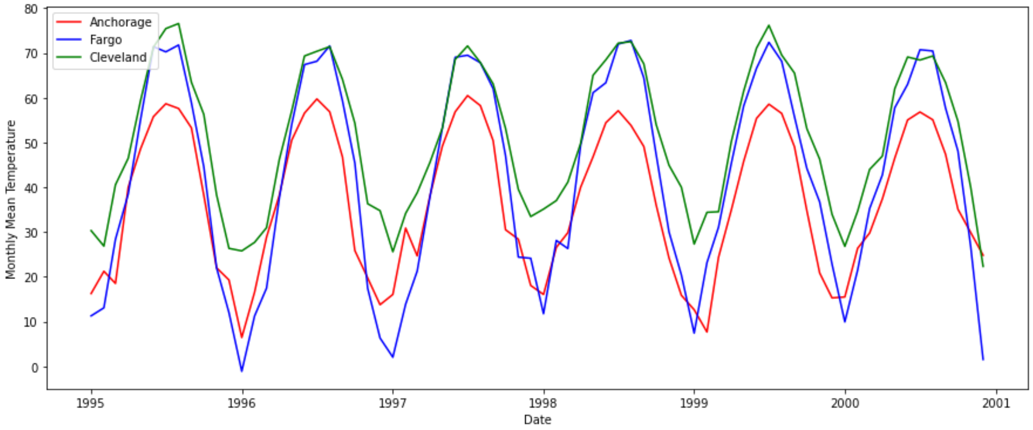


Figure 4. Yearly Mean Temperature of 4 US cities between 1995 and 2014

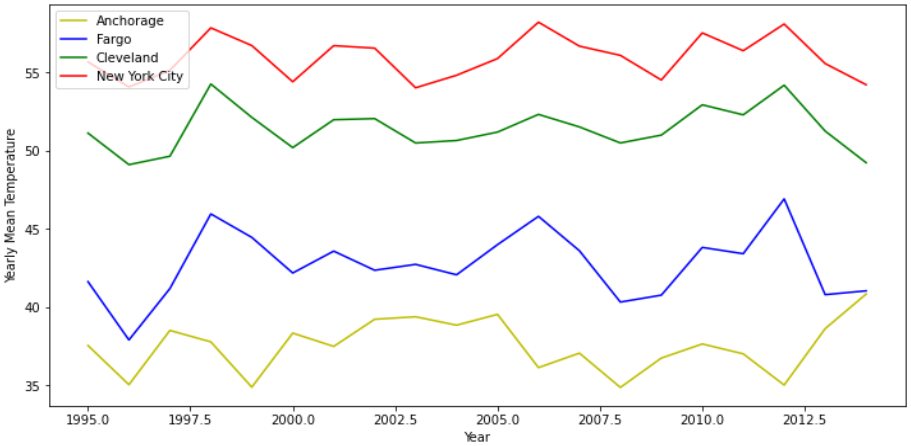
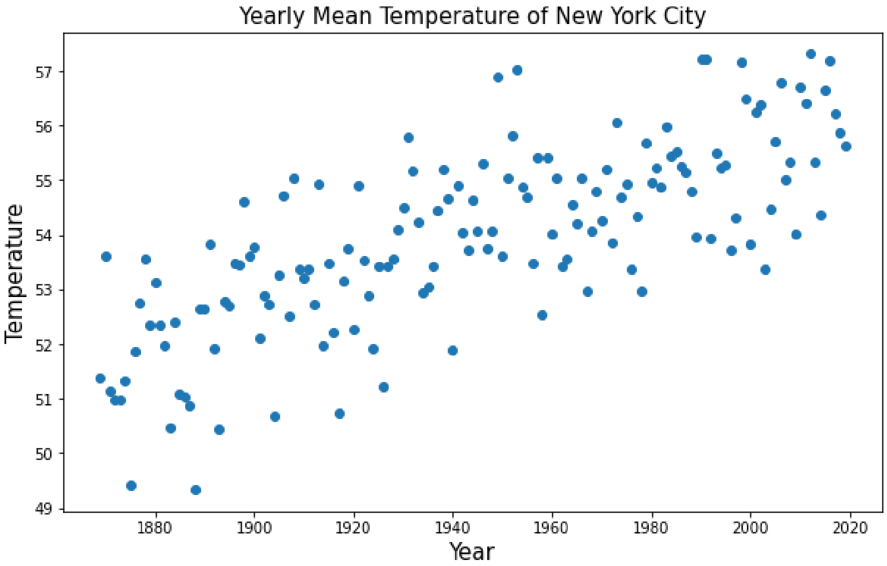


Figure 4 shows the yearly mean temperature of 4 US cities (New York City, Fargo, Cleveland and Anchorage from 1995 to 2014) over the 20 years. Just looking at the graph, the 20 years of the temperature data still cannot give an obvious answer to the climate change.

Finally, the 150 years of the temperature data of New York City gave us a clear idea of how climate change looks like (**Figure 5**).

Figure 5. Yearly Mean Temperature of New York City between 1869 and 2019



4. Exploratory Data Analysis and Initial Findings

From the total 19 cities, I took average of 1995-1999 mean temperature (Mean_1) and 2010-2014 mean temperature (Mean_2).

Mean_2 – Mean_1:

```
AK_Anchorage_Temp           1.131106
ND_Fargo_Temp               1.083954
```

```
FL_West_Palm_Beach_Temp     0.657804
Vermont_Burlington_Temp     1.061802
TX_El_Paso_Temp             1.430969
Wyoming_Cheyenne_Temp       0.768072
San_Francisco_Temp          0.513417
Cleveland_Temp              0.820345
Milwaukee_Temp              0.385734
New_York_City               0.533899
Honolulu_Temp              -0.102903
SA_Cape_Town                1.023740
Egypt_Cairo                 1.613335
India_Delhi                 1.005422
China_Beijing              -0.821714
Singapore_Temp             -0.119140
New_Zealand Auckland        0.080367
Russia_Moscow               1.955066
Argentina Buenos_Aires      1.046030
Peru_Lima                  -0.839321
```

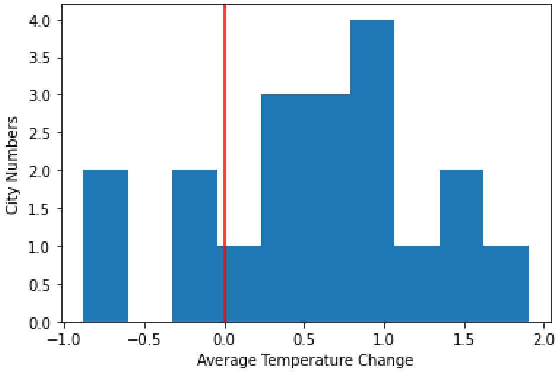


Figure 6. Mean Temperature Change of Cities

Mean of the total: 0.661

From the 19 cities, 15 cities showed increased temperature and 4 cities had decreased temperature from the 5 years of average temperature (1995-1999 vs. 2010-2014). The total average is +0.661 °F.

Permutation hypothesis test:

I assume there is no temperature change during the 20 years. The yearly mean temperatures of each city follows the standard distribution.

So I used the mean of the 20 years temperature and the standard deviation of the yearly mean temperature of each city to generate 10000 data (as the sample pool of the yearly mean temperature) for each city.

Then I randomly picked 5 data (5a) from the pool as the yearly mean temperature of 1995-1999, and other 5 data (5b) as the those of 2010-2014.

c=Average of 20 cities of Mean(5a)-Mean(5b)

I repeated 1000 times and obtained a list of c.

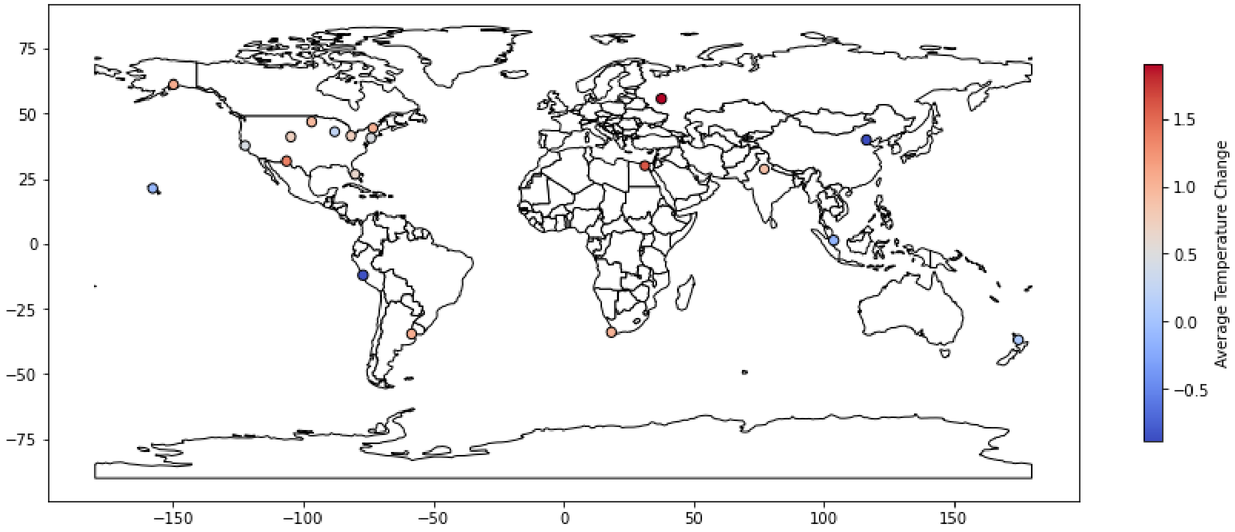
5% -95% of c is between -0.27 and 0.27 (°F)

1% -99% of c is between -0.41 and 0.41 (°F)

Since I got a 0.66 (°F) from the 20 cities. It is well above 99% of the possible data. **It means the original hypothesis that there was no climate change is wrong. The overall temperature of 20 cities increased during the years of 1995-2014.**

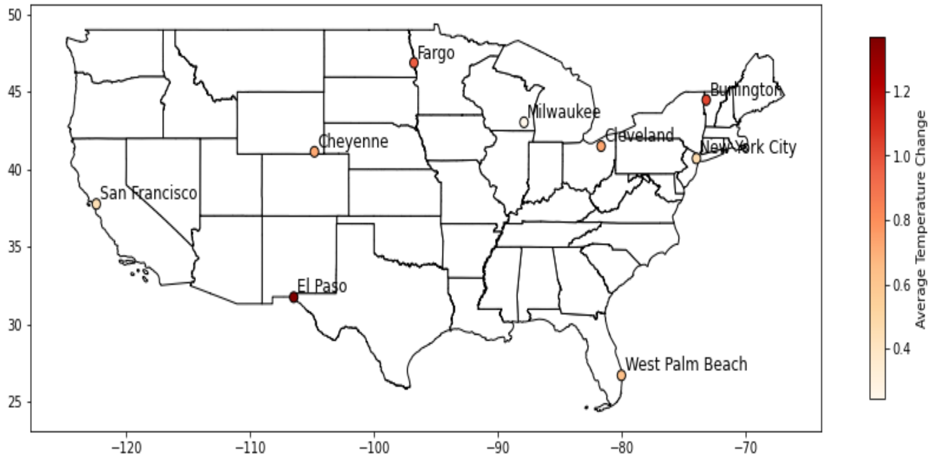
The figure below shows the cities on the world map and their change of the mean temperature of 5 years between 1995-1999 and 2010-2014.

Figure 7. World Map of the Mean Temperature Changes of the Selected Cities



All of the 8 cities in contiguous US showed elevated temperatures during this 15 years' span temperature comparison.

Figure 8. Mean Temperature Change of Selected Cities in Contiguous US



The daily mean temperature relationships between US cities have an almost linear relationship. For example, Fargo and Milwaukee are 2 US cities about 600 miles away. Their daily mean temperatures of two cities have a close to linear relationship.

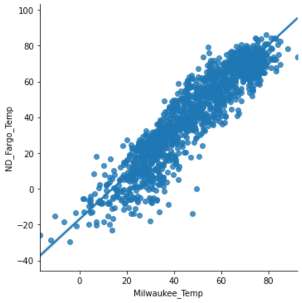
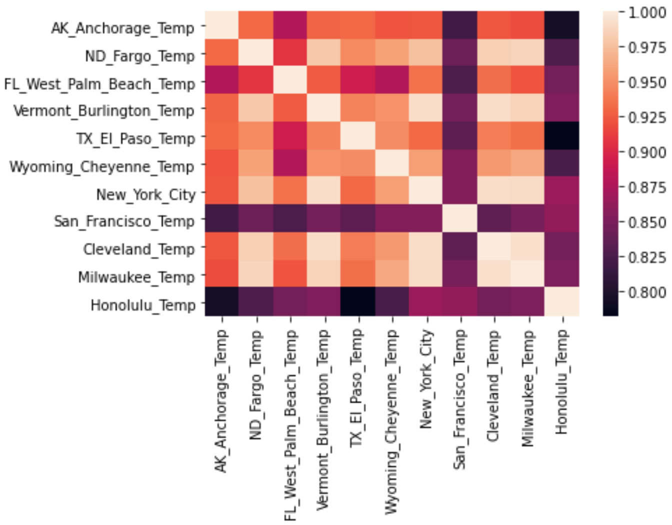


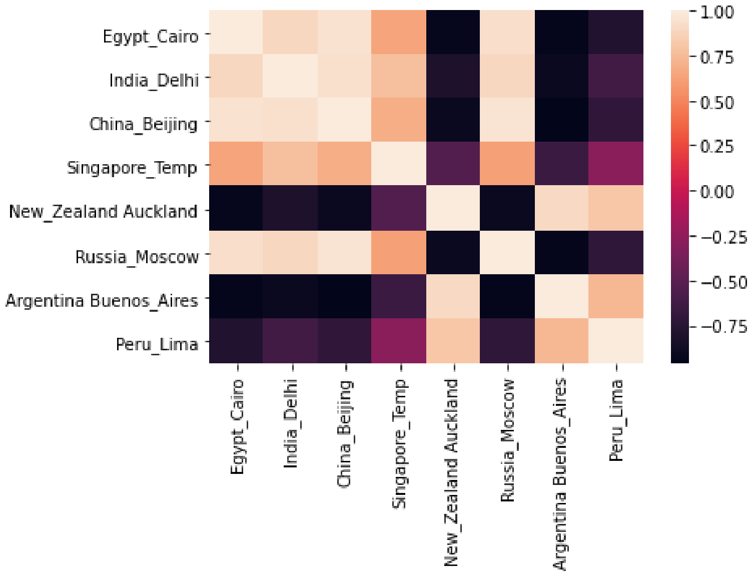
Figure 9. Heat Map of the Correlation of Monthly Mean Temperatures of US Cities



Most of US cities have very high degree of correlation (>0.9) of monthly mean temperatures to each other (**Figure 9**). The only exceptions are Honolulu and San Francisco, these 2 cities have mild temperature and low changes around the year, so they have relatively lower degree of correlation to other cities (around 0.8).

The correlation of the monthly temperatures between international cities also show interesting trend (**Figure 10**): the cities of middle to high latitudes in both northern or southern hemisphere have high degree of positive correlation coefficient (Moscow and Beijing , >0.9), cities between northern and southern hemisphere show high degree of negative relationship (for example: Beijing and Auckland, <0.9), but for the cases of between equatorial cities and mid to high latitude cities, there is low degree of correlation (for example: Singapore and Cairo).

Figure 10. Heat Map of the Correlation of Monthly Mean Temperatures of International Cities



5. Machine Learning and Predictions

Long term temperature prediction is a difficult task for meteorologists. Based on the current data, I tried to make some temperature prediction using various machine learning methods.

Prediction I:

If I know some other cities (Burlington and Cleveland) of monthly mean temperature, can I predict the monthly mean temperature of New York city?

KNN model: K =2, Train Score= 0.993, Test Score=0.981

Linear Regression: Train Score= 0.994, Test Score= 0.990, MAE=1.28°F

Gradient Boosting: Train Score= 0.998, Test Score= 0.988, MAE=1.36 °F

Conclusion: It is easy to predict the NYC temperature if you know the temperature of other US cities at the same time.

Prediction II:

Based on the 150 years of historical data of the monthly mean temperature of New York city, can we predict the future temperatures of NYC?

KNN model: K=1, Test Score =0.892; K=2, Test Score =-0.272.

Linear Regression: Train Score = 0.965, Test Score= 0.959, MAE= 2.45°F

Gradient Boosting: Train Score = 0.969, Test Score= 0.930, MAE= 2.89°F

Random Forest: Test Score= 0.928, MAE= 2.95 °F

Check the long term prediction of the Linear Regression and the Gradient Boosting Models:

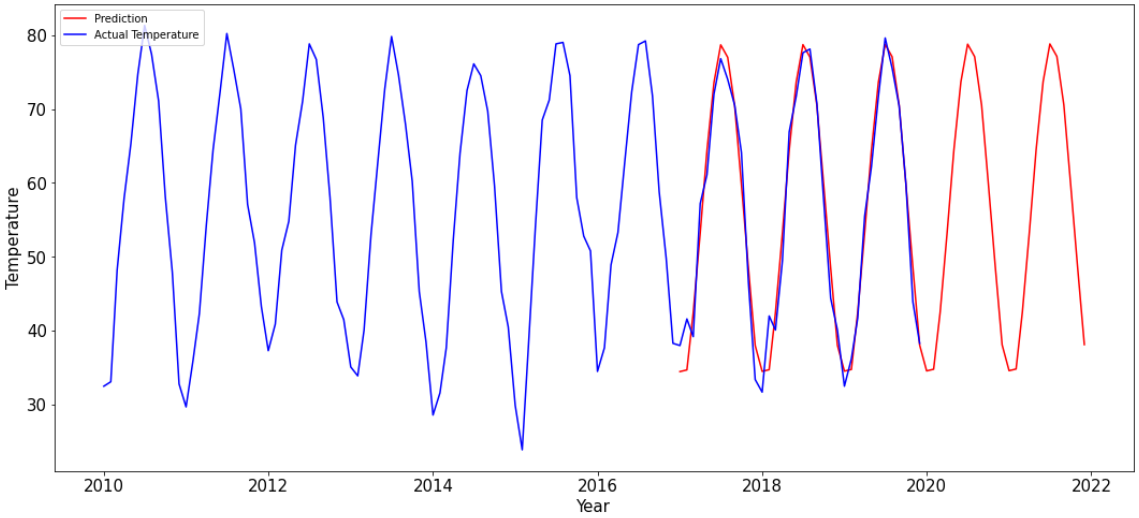
Year 2019, the monthly mean temperature of NYC (From January to December):
Actual Temperature: [32.5, 36.2 ,41.7, 55.5, 62.2, 71.7, 79.6, 75.5,70.4, 59.9, 43.9,38.3]

Year 2100, the monthly mean temperature of NYC:

```
Linear Regression: [36.7, 37.3, 45.0, 55.7, 66.8, 75.9,
 81.1, 79.4, 72.8, 62.1, 50.9, 40.4]
Gradient Boosting: [33.8, 34.7, 42.3, 52.8, 62.8, 71.3,
 76.3, 74.7, 67.9, 57.2, 47.9, 29.7]
```

It is obvious that Gradient Boosting does not give a good prediction; the values predicted by Linear Regression is reasonable. **Figure 11** shows the predicted monthly mean temperature by the linear model and the actual temperature of New York City (NYC).

Figure 11. The Actual Monthly Mean Temperature and the Predicted Temperature of NYC



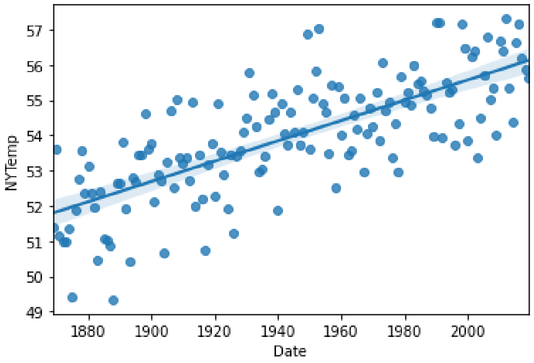
Conclusion: Predicting future temperature based on historical data is challenging. We can still get a rough idea of the future temperatures if the climate change trend remains same.

Prediction III:
If I know the monthly mean temperature of NYC, Milwaukee, Cleveland and Burlington, can I predict the monthly mean temperature of NYC next year?

KNN model: K =4, Train Score= 0.951, Test Score=0.916
Linear Regression: Train Score= 0.969, Test Score= 0.969
Gradient Boosting: Train Score= 0.992, Test Score= 0.914

Conclusion: Adding data of other US cities did not help predicting the NYC temperatures next year.

6. Conclusion and the Future Work



The Climate Change is a complicated research project. Just looking at 20 years of climate data is not enough to give us a deep understanding of how it looks like. Fortunately, we have 150 years of weather data of New York City (Figure on the right), we do see a linear like relationship between time and the yearly mean temperature. From the case of New York City, we can clearly see the climate change over the 150 years of history.

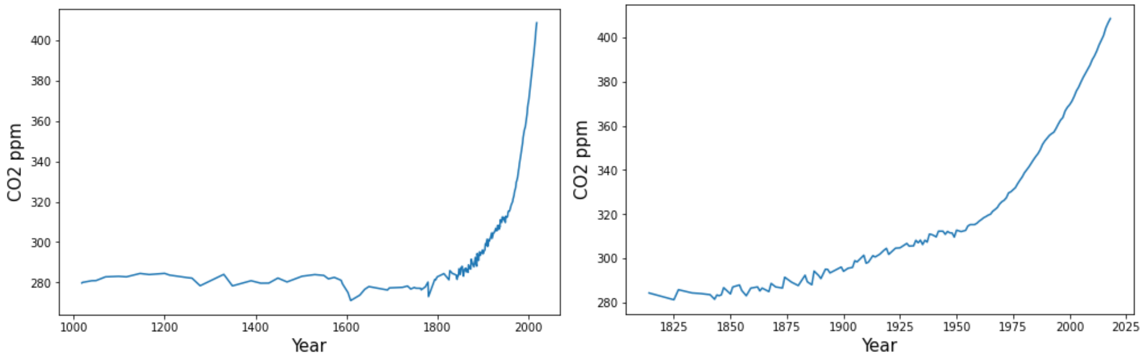
We need to have more data (longer history) to do a better climate change research. One important part of the research needs to be done in the future is the relationship of the CO_2 concentration in the atmosphere and the temperature increase.

Figure 12 shows the CO_2 concentration before 1800 remains almost unchanged for a long time. After 1800 and before 1950, the increase of CO_2 level was still modest. Recent 50 years saw the most rapid increase of the CO_2 concentration, but we have not seen the same temperature increase from the New York City yearly mean temperature. The temperature increase between 1869 and 2019 seems steady. Is it because there is a time lag between CO_2 concentration climbing and the corresponding greenhouse gas effect? Any other factors affect the climate but we did not take into consideration?

From this data research, we can see the climate change is real and the rate of warming is slow. You cannot see an obvious change within 10 or 20 years, at some places you can even see some temperature decease within 10 -20 years. From the 150 years of weather record of New York City, the mean temperature increase per decade is about 0.3°F. But during the recent 50 years we have seen a much faster increase of CO_2 concentration than any other time in the earth history, will the temperature increase trend remain same as last 150 years is a big question we cannot answer now.

Figure 12. CO_2 concentration in atmosphere

(data from https://ourworldindata.org/grapher/global-co-concentration-ppm)



The climate change of earth is a long term and slow-going process and at the same time it is extremely important since it is about the future of the earth and human beings. I hope more research can be done on climate change and more importantly more work will be done to stop the climate change.