Climate Change Analysis

Problem Statement

Problem

Climate change has become a huge problem that affects the very existence of life on the planet and a majority of climate change in the recent years in our industrial age is caused due to human activity. This caused a dramatic increase of temperatures on both land and water surfaces, meltdown of ice caps, rise in sea levels, increase in the magnitude and frequency of hurricanes, and extinction of many species.

This adverse effects can be slowed down / stopped / reversed with changes in human activity and proper education. Even in this science era, many people - often educated, and in powerful influential positions - disregard the climate change phenomenon as a hoax and refuse to take any action.

This project is to show the correlation between the rise in temperatures due to rise in atmospheric greenhouse gases (CO2) caused by human activity and to conclude that controlling the CO2 emissions can regulate climate change phenomenon.

Client Use Case

This solution is applicable to the general public as well as many companies in various industries like automobile, fossil fuel extraction, food processing, chemicals, etc. that cause an increase in atmospheric greenhouse gases.

The solution provides the client an education of how rise in CO2 levels is causing a rise in average temperatures, which is the reason for all the adverse effects on various life forms including humans on Earth.

This study would encourage companies to design their processes to be more efficient and build clean energy facilities, and encourage people to be cautious about their actions to reduce the carbon footprint.

Data Source

1. *Temperature* - All the temperature data necessary for the study is downloaded from National Centers for Environmental Information data repository of NOAA.
    a. The dataset is a global daily temperature readings from Jan 1975 to Oct 2018. The data set includes daily data for - Average temperature, Minimum temperature, Maximum temperature, Precipitation, details regarding thunderstorms, hail, and tornado occurrence for the day

2. *CO2 Levels* - All the CO2 levels data is downloaded from Earth System Research Laboratory - Global Monitoring Division of NOAA
    a. The dataset includes monthly data of concentration of CO2 measured at various locations in the US (different states). Even though the data is a global measure, data collected in various states is assumed to be state level CO2 data for our analysis
3. *Land Cover* - The land cover data is downloaded from the National Land Cover Database of USGS.

Methodology

1. Problem approach
    a. The aim of the project is to show the trends in temperature increase and the prime factor is the increase in CO2 levels due to human activity in the industrial age.
    b. Since the cause has many factors and the data is available to show the correlation between factors and effects, this study will be a **supervised** learning.
    c. This supervised learning model is going to be **regression and time series** forecast model to show the correlation, recent trends and future trends
    d. The variable this project is aiming to predict is temperature - average, min, and max by state in USA
    e. This prediction models are based on variables - CO2 levels, and land cover data variables (tree cover data specifically, which contributes to the atmospheric CO2 levels)
    f. The training data will be the raise of CO2 levels. The idea is to build a forecast models with varying levels of CO2 data once the relationship is established and show the sensitivity analysis by changing the trend in factor variables - increasing trend, prediction trend, decreasing trend in CO2, and also show how CO2 is changed in the atmosphere by varying the tree cover predictors

Output

The deliverable will be a presentation with associated visualizations of sensitivity analysis, and an explanation of the approach.

Data Processing

Overview

I chose to analyze climate change for my Capstone Project 1 at Springboard. The objective of my project is to analyze the climate change trend in various states of United States and deduce how greenhouse gas (esp. CO2) emissions, and human activity (deforestation) is contributing to the climate change. I used US Government's public data for my analysis. The datasets I used are -

1. Climate information
    a. Source: NOAA
    b. Data attributes:
        i. Climate dataset
            1. Station name
            2. Substation number
            3. Date of observation
            4. Average temperature
            5. Dewpoint
            6. Sea level pressure
            7. Station pressure
            8. Visibility
            9. Windspeed
            10. Max windspeed
            11. Windgust
            12. Max temperature
            13. Min temperature
            14. Precipitation
            15. Snow depth
            16. FRSHTT (Fog, Rain, Snow, Hail, Thunder, Tornado)
        ii. Station dataset
            1. Station number
            2. Substation number
            3. State
            4. Country
            5. Begin
            6. End
            7. Latitude
            8. Longitude
            9. ICAO
            10. Elevation
            11. Station Name

c. Frequency: Daily (1975-2018)
d. States: All US

2. CO2 Concentration
   a. Source: https://www.esrl.noaa.gov/gmd/dv/data/
   b. Data attributes:
      i. Year
      ii. Month
      iii. CO2 Concentration
   c. Frequency: Monthly (1958-2017 : *varies by state*)
   d. States: AK, CA, CO, FL, HI, OK, OR, UT, VI, WA, WI
      i. CO2 concentration data couldn’t be found for all the US states because not all US states have CO2 assessment stations.
3. Fuel Consumption:
   a. Source: EPA
   b. Data attributes:
      i. State
      ii. Year
      iii. Petroleum product category (petrol/coal)
      iv. Consumption (Petrol: barrels, Coal: tons)
   c. Frequency: Annual (1980-2016)
   d. States: All US
4. Forest Area:
   a. Source: USDA
   b. Data Attributes:
      i. Region
      ii. Region or State
      iii. Year
      iv. Total land
      v. Total cropland
      vi. Cripland used for crops
      vii. Cropland used for pasture
      viii. Cropland idled
      ix. Grassland pasture and range
      x. Forest-use land (all)
      xi. Forest-use land grazed
      xii. Forest-use land not grazed
      xiii. All special uses of land
      xiv. Land in rural transportation facilities
      xv. Land in rural parks and wildlife areas
      xvi. Land in defense and industrial areas
      xvii. Farmsteads, roads, and miscellaneous farmland
      xviii. Land in urban areas
      xix. Other land

   c. Frequency: 5 years (1945-2012)
   d. States: All US

Data Wrangling

Cleaning Data

Climate Data

Packages: pandas, os, numpy
Methodology:

1. Stations List dataset
   a. Load Station list dataset using pandas read_csv method
   b. The analysis needs only state and country information from this dataset to map station number with the main climate dataset
   c. So delete all the irrelevant columns except Station information, State, and Country
   d. Extract all the US states and save the information in a pandas DataFrame.
2. Climate dataset
   a. Climate dataset is downloaded in a series of zipped directories - by Station Number by Year
   b. Using Package os, loop through the all the zipped folders load the data into pandas DataFrame using read_csv method
   c. Missing values are listed as ‘9999.9’ or ‘999.9’. Specify the na_values parameter in read_csv method to identify the missing values
   d. For the sake of simplicity, the project only analyzes temperature and frequency of other events like hail, tornado, thunder, snow, and rain. So drop all the columns in the dataset except what’s needed before proceeding to next steps
   e. Merge Station and Substation information to match that in the station list DataFrame. This station information will be used to extract the State from station list dataset
   f. As stated above, the project also analyzes the frequency of national events like rain, snow, hail, thunder, and tornado. All this information is listed under one column (FRSHTT) as 1/0 for respective event in the dataset
   g. Separate the FRSHTT into different columns with corresponding header and convert the data into number
   h. Due to limitations in the computing power, and the availability of rest of the datasets, I chose to analyze monthly data. So the next step is to upsample the data into Monthly frequency
   i. Create a dictionary with aggregate functions for each column

j. Use the dictionary to aggregate the daily data into monthly using the aggregate methods defined in the dictionary and pandas native resample method grouped by station number
k. Join the stations_list and climate_data Data_Frames to get the state information for all the data. Now that we have State information, drop the Station_Num column from the climate dataset
l. Run the above logic for each year and process all the data into one pandas DataFrame
m. Save the cleaned climate data into a CSV file using pandas to_csv method for analysis. This step eliminates processing raw data every time.

CO2 Concentration

Packages: pandas, os
Methodology:

1. CO2 concentration data is downloaded in different files by State
2. Using os package, loop through all the files and load each file into a pandas DataFrame using read_csv method
3. Set the columns for the DataFrame and extract the State code from the file name.
4. Convert the Month and Year columns into month end dates
5. Append all the individual State DataFrames into one with all the States
6. Save the cleaned CO2 concentration data into a CSV file using pandas to_csv method for analysis. This step eliminates processing raw data every time.

Fuel Consumption

Packages: pandas
Methodology:

1. Load the fuel consumption data into Pandas DataFrame using read_csv method
2. The consumption data is broken out into various products, of which we'll need Total Petroleum and Total Coal consumptions - identified as 'PATCP' and 'CLTCP'
3. Extract only these two codes data and discard the rest
4. The data is yearly, and we'll need to downsample it to monthly. So convert the Year column into a year end date
5. Using pandas' resample method, downsample the annual data to monthly using a user defined function to interpolate the consumption numbers into all the months based on the trend for the year
    a. ***Assumption***: I assumed a linear relation to interpolate total yearly consumption to monthly
6. Transpose the row data into columns - Petrol_Consumption, and Coal_Consumption
7. Save the cleaned fuel consumption data into a CSV file using pandas to_csv method for analysis. This step eliminates processing raw data every time.

Forest & Urban Land Data

Packages: pandas
Methodology:

1. Forest and urban land use data comes in a CSV file with various other categories of land use.
2. The data also has State names instead of State codes. Since State codes are used in the rest of the datasets, I used another file with State names and corresponding codes
3. Load land use data into DataFrame using pandas read_csv method. Missing values are coded as “N.A.” in the dataset. Load only the required columns - State, Year, Forest land, and Urban land.
4. Using pandas merge method, inner join State codes and land use DataFrames. Inner join will also eliminate the region data in the land use dataset and store only state information
5. The data is on a 5 year frequency. This needs to be downsampled to monthly. Convert the Year column into datetime with year end date
6. Drop any NA values
    a. ***Assumption***: Since the data is being downsampled to monthly, missing values are filled using linear interpolation method of pandas resample method
7. Using pandas resample method, downsample the data to monthly frequency using interpolate method
8. Rename the columns accordingly and save the DataFrame into CSV using pandas to_csv method.

Missing values

All the datasets involved are resampled to monthly frequency from daily (GSOD), annual (Fuel Consumption), and 5 years (Land use), missing values have been filled with appropriate interpolation, except for CO2 Concentration. CO2 Concentration is monthly data with no missing values, but the range of data varies widely by state.
The CO2 missing data for trailing years is forecasted until 2018 using ARIMA forecasting model. The best p, d, and q parameters for the ARIMA model are manually chosen. Python statistical package - statsmodels - offers basic forecasting ability that doesn’t necessarily offer the best fit. I chose the p, d, and q parameters manually by observing the fit of all the forecasted plots iteratively generated by the model for different parameters for each state and chose the parameters that offer best fit for each state. I saved these parameters in a dictionary and arrived at forecasted CO2 levels based on defined parameters for each state where CO2 data is available.

Outliers

GSOD data has outliers especially for minimum temperatures, which are taking the minimum down and are affecting the average temperatures. GSOD computed the TAVG reading using these outliers, and they had to be corrected as well. I confirmed the presence of outliers using some observations -

1. Wild swings in minimum temperatures - FL state has minimum temperatures averaging around 60F, but it would suddenly drop to -13F one day.
2. Fact searching revealed that the the least temperatures ever recorded in FL state is -2F in 1880

Possible solutions:

1. Deeper look at the data revealed that such outliers usually happened when Dewpoint data in the dataset is recorded as missing.
    a. Elimination of all the daily records with missing dewpoint information would eliminate some such outliers in minimum temperature, as well as wrongly computed average temperature in the dataset.
    b. This process might remove correct entries as well, but because I'm downsampling daily data to monthly on means, we don't effectively lose information.
    c. This solution might not remove all the outliers because not all outliers are missing dewpoint information. Such remaining outliers can be adjusted by smoothing the Min Temperature and Max Temperature curves by taking averages for resampling instead of taking Min and Max for the month respectively.
    d. Pro: This solution is easy to implement
    e. Con: This might not eliminate all the outliers, and taking average of Min and Max temperatures, we may lose legitimate min / max temperatures occurred in the month.
2. Identify outliers using percent change method
    a. Derive the percent daily change and identify the outliers if they swing by more than a certain threshold and remove them
    b. Pro: With correct threshold, this might identify outliers effectively and remove them
    c. Con: This solution is hard to implement. It might also remove legitimate swings in temperature.

Upon careful review of the data, I chose to eliminate bad data and outliers using solution #1. Taking averages of Min and Max temperatures would also help us to effectively compute the trend of raise in temperatures.

Statistical Analysis

In this document, I'm going to present the observations of temperature trend all over USA and some interesting insights in Alaska. For a detailed report on various states/regions of USA, please refer to

Climate Change Analysis

Overview

Factor Analysis

To establish the correlation between factors and temperatures, I used basic regression plots provided by Python plotting library Seaborn. *At this time, factor analysis is limited to the visual interpretations and the notebook doesn't present any metrics or regression models. These models shall be presented in future iterations of the analysis.*

Temperature Change Analysis

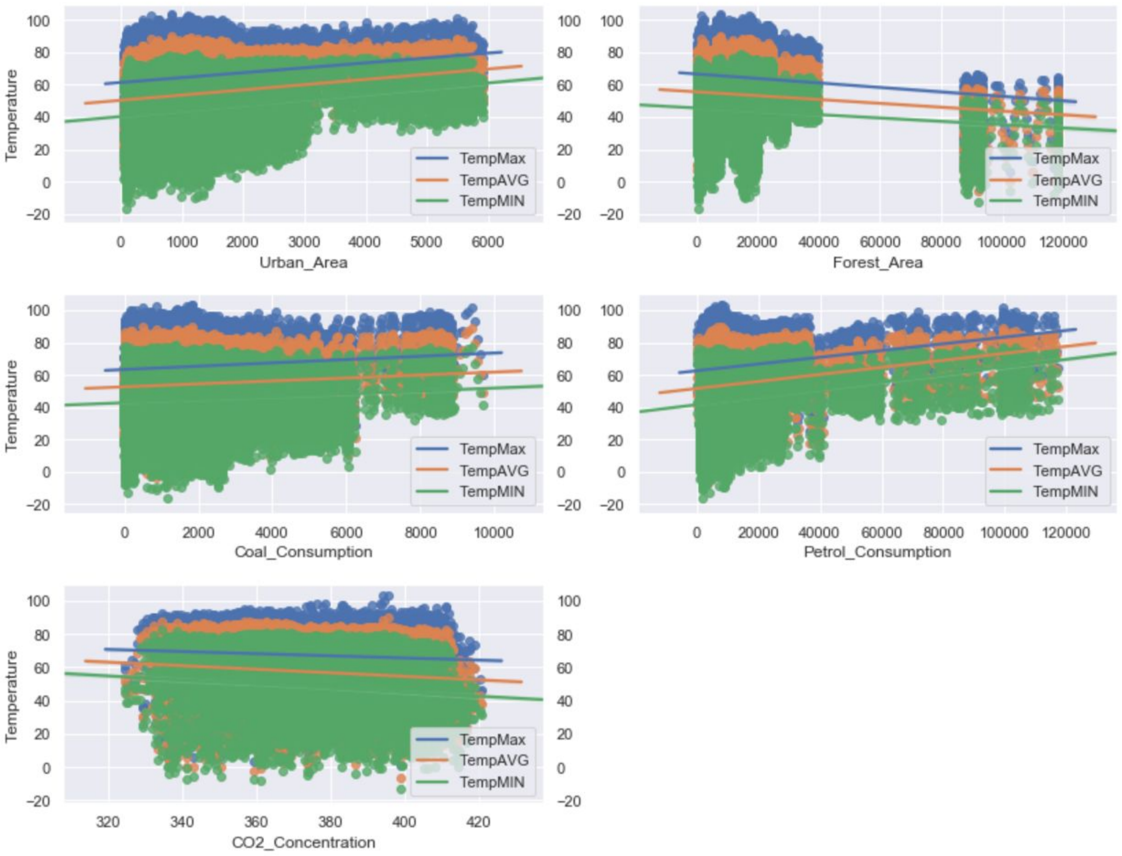
To analyze the temperature difference over the last 43 years (1975 - 2018), I used statistical methods to compare means hypothesis testing on bootstrapped samples:

1. The first set of data is first 5 year monthly temperatures - 1975-1979;
2. The second set of data is recent 5 year monthly temperatures - 2014-2018.
3. Using these samples of data, I used bootstrapping methods to arrive at a sample means of 10,000 so the data is normally distributed according to Central Limit Theorem.
4. I used two ways to analyze the temperature means between to data samples using Hypothesis testing
    a. H_0 : Mean temperature in US DID NOT change between 1975-79 and 2014-18
    b. H_a : Mean temperature in US DID change between 1975-79 and 2014-18
5. I used two was of statistical analysis to deduce at the conclusion of hypothesis testing
    a. Shifted means method
        i. Compute the difference in means between both the samples. This difference in means will be the empirical mean difference, which serves as a benchmark statistic for further analysis
        ii. The means of both the samples are shifted in a way that there is no difference between the means of both the samples - essentially modify the data points of both the samples that they have the same mean
        iii. Now compare the empirical mean difference with all the means in the shifted sample.
        iv. The probability of having the sample means at least as extreme as that of the empirical mean difference is the p-value that concludes the hypothesis test. A $p < 0.05$ rejects H_0 and statistically suggests that temperatures changed.

b. Compare means method
   i. Since the bootstrapped samples are means of randomly selected samples, the Standard Deviation of the samples becomes Standard Error of means.
   ii. Using the means of the samples, and standard error of means of 1975 sample (since we are comparing 2018 temperature against 1975 temperature), compute the z value.
   iii. Z value statistically suggests the distance between the means in terms of standard errors on a bell curve.
   iv. The z value can be used to arrive at the p-value using Python library Scipy's CDF function. A $p < 0.05$ rejects H_0 and statistically suggests that temperatures changed.

USA

Factor Correlation



The regression plot above shows effect of all the factors over Max, Avg, and Min temperatures over the years.

Urban Area: Increase in Urban Area corresponds to increase in temperatures

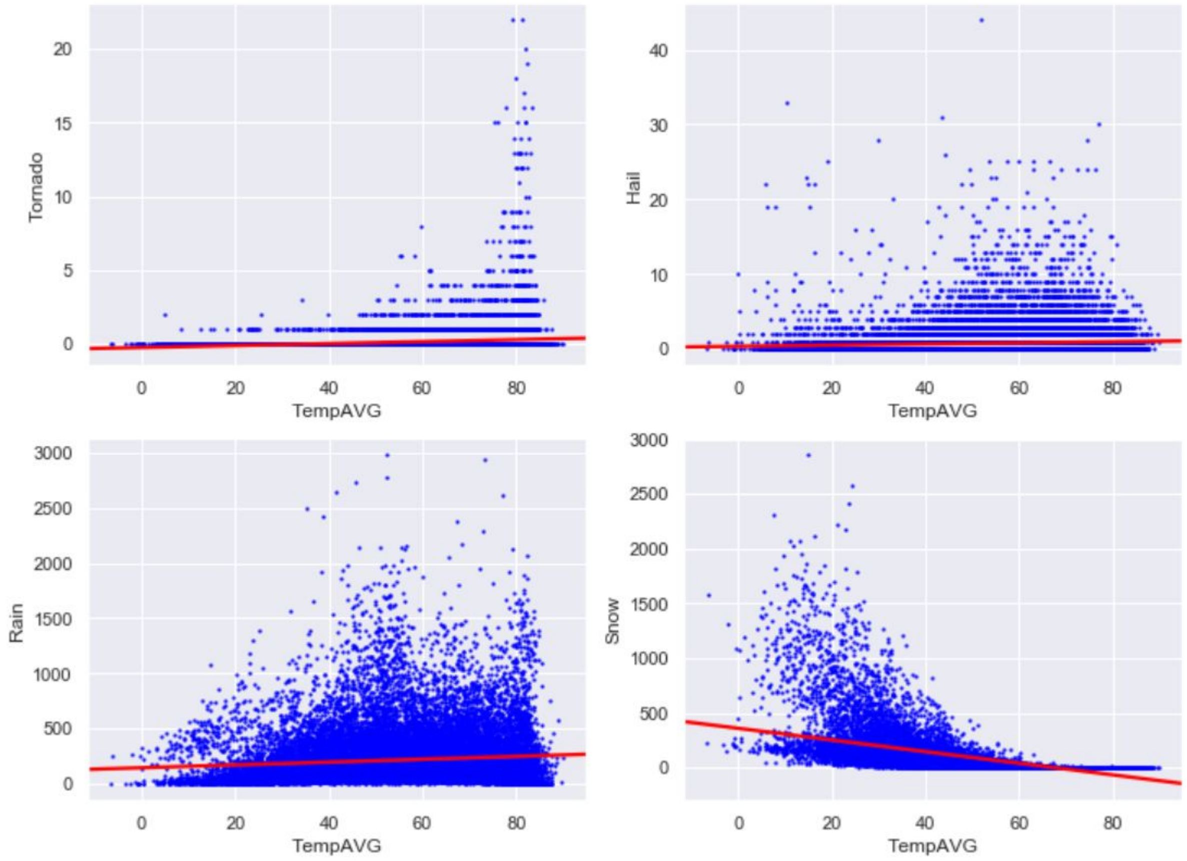
Forest Area: Increase in Forest Area has a profound impact in decreasing the temperatures. Scientific reason we should plant more trees!

Coal Consumption: Increase in consumption of coal corresponds to increase in temperatures

Petrol Consumption: Increase in consumption of petrol corresponds to increase in temperatures

CO2 Concentration: The chart shows that increase in atmospheric CO2 concentration correspond to decrease in temperatures, contrary to the scientific explanation. This could be due to low correlation values of CO2 data with available temperature data. Further analysis is required to better present this relationship.

Natural Event Correlation



Raise in temperatures have adverse effects on Earth's natural phenomena. The chart below shows the effect of temperature changes in natural event occurrences.

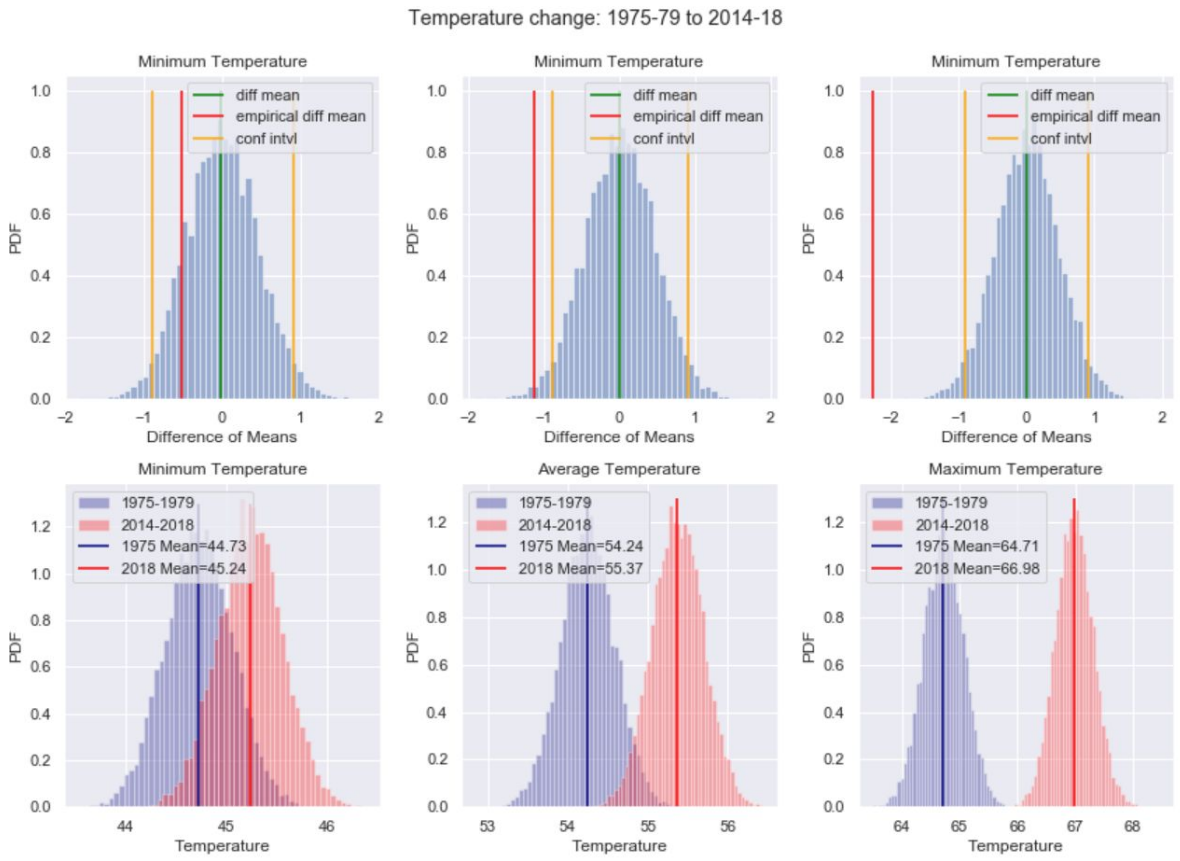
Tornado: Raise in temperature corresponds to increase in tornado occurrences.

Hail: Raise in temperature corresponds to increase in hail occurrences.

Rain: Raise in temperature corresponds to increase in rainfall.

Snow: Raise in temperature corresponds to decrease in snowfall.

Temperature Change Analysis



```
Hypothesis t-test:

Minimum Temperature:
Shifted Means p-value = 0.1358
Compare Means p-value = 0.0638
    Fail to reject H0: The minimum temperature data between 1975-79 and
2014-18 is NOT STATISTICALLY SIGNIFICANTLY DIFFERENT; Data suggests mean
minimum temperature is not different between 1975 and 2018

Average Temperature:
Shifted Means p-value = 0.0072
Compare Means p-value = 0.0004
    Reject H0: The average temperature data between 1975-79 and 2014-18 IS
STATISTICALLY SIGNIFICANTLY DIFFERENT; Data suggests that mean temperature
is different between 1975 and 2018.
```

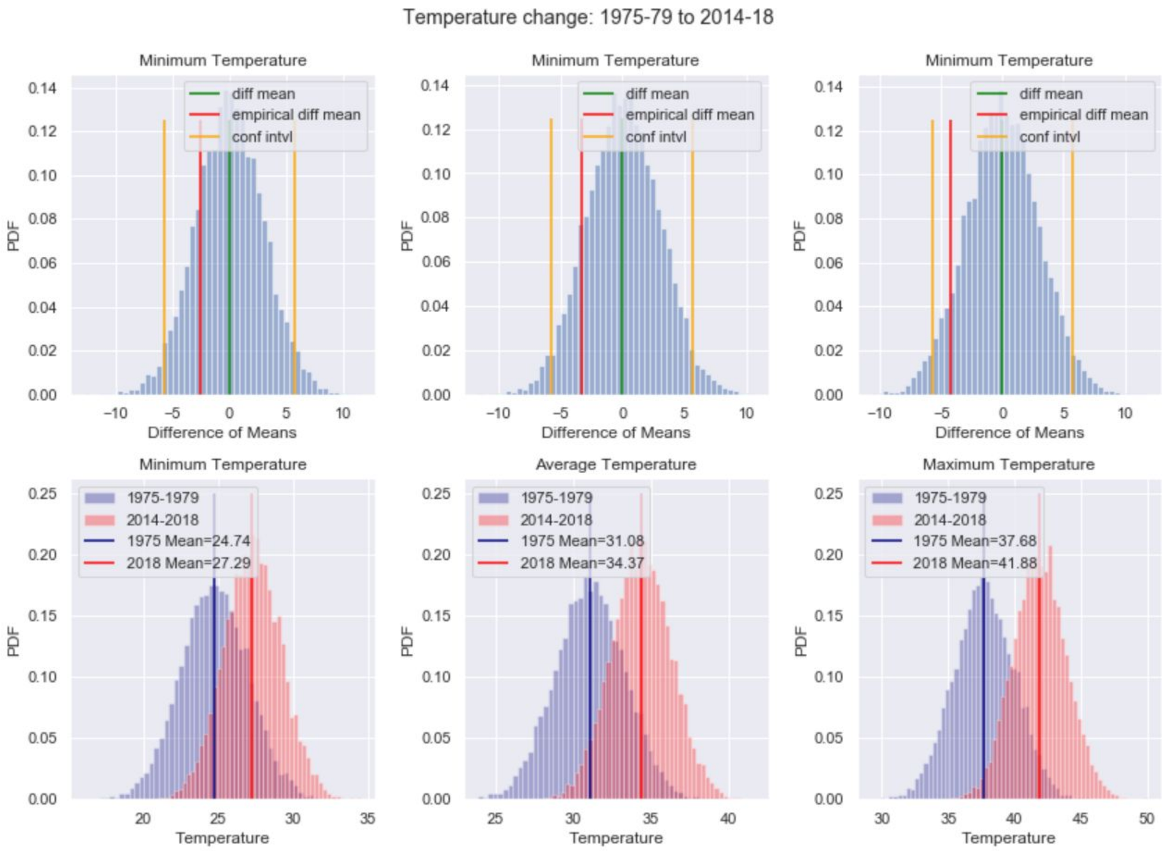
```
On average, 5 year mean temperature in USA between 1975-79 and 2014-18
changed around 1.13F

Maximum Temperature:
Shifted Means p-value = 0.0
Compare Means p-value = 0.0
    Reject H0: The temperature data between 1975-79 and 2014-18 IS
STATISTICALLY SIGNIFICANTLY DIFFERENT; Data suggests that mean max
temperature is different between 1975 and 2018.

On average, 5 year mean MAX temperature in USA between 1975-79 and 2014-18
changed around 2.27F
```

The visual data analysis and statistical data analysis of temperatures in USA from 1975 show that the climate change has an impact in temperatures.
Although min temperature changes cannot be confidently rejected, mean and max temperature levels significantly increased - by 1.1℉ and 2.2℉ respectively.
The statistical analysis also shows that min temperatures increased, although not to a 95% confidence level.

Insights into Alaska



```
Hypothesis t-test:

Minimum Temperature:
Shifted Means p-value = 0.1942
Compare Means p-value = 0.128
    Fail to reject H0: The minimum temperature data between 1975-79 and
2014-18 is NOT STATISTICALLY SIGNIFICANTLY DIFFERENT; Data suggests mean
minimum temperature is not different between 1975 and 2018

Average Temperature:
Shifted Means p-value = 0.1342
Compare Means p-value = 0.0712
    Fail to reject H0: The average temperature data between 1975-79 and
2014-18 is NOT STATISTICALLY SIGNIFICANTLY DIFFERENT; Data suggests mean
temperature is not different between 1975 and 2018

Maximum Temperature:
Shifted Means p-value = 0.0823
```

```
Compare Means p-value = 0.0327
    Reject H0: The temperature data between 1975-79 and 2014-18 IS
STATISTICALLY SIGNIFICANTLY DIFFERENT; Data suggests that mean max
temperature is different between 1975 and 2018.

On average, 5 year mean MAX temperature in AK between 1975-79 and 2014-18
changed around 4.2F
```

The visual data analysis and statistical data analysis of temperatures in Alaska from 1975 show that the climate change has an impact in temperatures in Alaska.

Although min and mean temperature changes cannot be confidently rejected, max temperature levels significantly increased - by 4.1℉

The statistical analysis also shows that min and mean temperatures increased, although not to a 95% confidence level.

Temperature raise in Alaska is particularly dangerous because melting ice caps in Alaska directly contribute to the rising sea levels. Temperature raise in this state is not just dangerous for humans, but also for the wildlife that habitat in Alaska. Temperature increase is affecting the beauty of Alaska, which is tourism dependent state, leading to an economic downturn in the state.

Conclusion

Through the visual and statistical exploration of data, we can conclude that human activity present as factors for climate change. Data from all the states show a strong correlation between raise in temperature and the amount of fossil fuels we burn, trees we cut down, and cities we expand.

The statistical analysis shows that the climate change has a profound impact on the extreme states in the heat spectrum - Temperatures significantly rose in colder states like Alaska and hotter states like Florida, and Texas. While the temperature change is not statistically significant is the medium temperature states, it prevails.

Temperature change is a long term phenomenon. The analysis of last 40 years shows United States had a significant impact due to climate change. It will only be worse in the coming decades if no preventive action is taken.

One interesting insight is that strangely, increase in CO2 concentration seem to inversely affect the temperature. All the charts show that increase in CO2 concentration seem to decrease temperatures. Further investigation with statistical evidence is required to dive into this insight and present a valid conclusion.

Conclusively, an immediate action is absolutely necessary to counter climate change or life's future on Earth will be grim.