BIG DATA ANALYTICS

FINAL PROJECT REPORT

ANALYSIS OF GLOBAL TEMPERATURE DATA AND FORECASTING

IST-718_SUMMER 2022

Prof : JILL(Jillian Lando)

Contents

OVER VIEW

Climate change is the most widely speaking topic in this information age, while the technology is developing in a rapid phase, and similarly the industrial growth is occurring, however the cascading impacts are more droughts, raise in temperatures, increase in snow melt, sea level pressure rises…so and on and on.

There are lot of research are happening around the world to identify the causes, gauge the impacts, analyze the trends, and predict the future impacts.

We can’t recreate the natural resources however we can reduce the damage and restore and thus take measures to worsen the impact for future generations.

Basic idea is to identify the answers and results that were anticipated as part of research for this final project as part of IST 718 project submission.

QUESTIONS PLANNED TO EXPLORE

1) What is the trend in temperature rise?

2) is there any influence of the this rises over the seasons?

3) is it possible to visualize the top ten cities in the world that are impacting?

Over the years I wish to consider for 50 years data for this .

4) Are there any non-impacted countries due to the rise?

5) Topmost continent that got impacted due to this change in temperature

Approach

For the project purpose, I will be trying to incorporate the OSEMIN approach of data analysis and trying to fit the SOAR (Specify, Observe, Analyze, Recommend)

Here the main categorical variable is temperature and how the temperature is increasing over time, however both temperature and time are continuous variables, and the relationship (linear) has been observed across the geo locations (longitude and latitude) are continuous variable however the specific city, country, state has been considered as discrete.

The story is all about how best statistical algorithms, models, can be experimented to have the analysis of the temperature rise over years using the datasets are available.

Data Review

Data has been collected from Kaggle here is the link to access the same.

https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

Along with the data also collected the ISO continents to map the regions to continents

Data clean up

After extracting the Kaggle data and loading them to python environment and check the data details

For the ease of analysis to this specific assignment I will be mostly using “Land Average Temperature” as the key variable to analyze in relation to date(YYYY/MM/DD)

Removing the nulls and blanks as some of the 200 years old data doesn’t have all the columns with relevant data

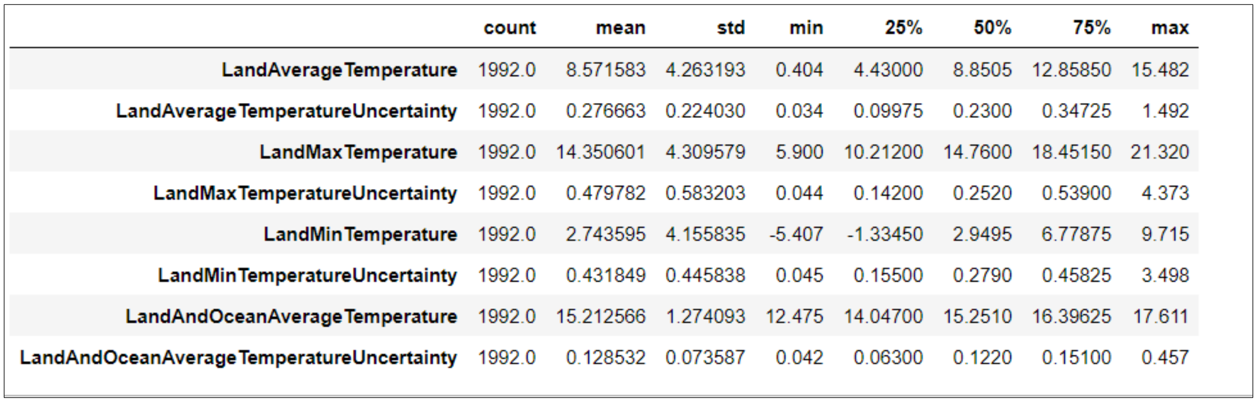
Also, for few geo locations that codes are not mapped

Descriptive statistics

Here is the summary for the key data set for the analysis global temperature by looking into the variables like

```
0   LandAverageTemperature                     1992 non-null   float64
 1   LandAverageTemperatureUncertainty          1992 non-null   float64
 2   LandMaxTemperature                         1992 non-null   float64
 3   LandMaxTemperatureUncertainty              1992 non-null   float64
 4   LandMinTemperature                         1992 non-null   float64
 5   LandMinTemperatureUncertainty              1992 non-null   float64
 6   LandAndOceanAverageTemperature             1992 non-null   float64
 7   LandAndOceanAverageTemperatureUncertainty  1992 non-null   float64
dtypes: float64(8)
memory usage: 140.1 KB
(1992, 8)
```

The Summary is



	count	mean	std	min	25%	50%	75%	max
LandAverageTemperature	1992.0	8.571583	4.263193	0.404	4.43000	8.8505	12.85850	15.482
LandAverageTemperatureUncertainty	1992.0	0.276663	0.224030	0.034	0.09975	0.2300	0.34725	1.492
LandMaxTemperature	1992.0	14.350601	4.309579	5.900	10.21200	14.7600	18.45150	21.320
LandMaxTemperatureUncertainty	1992.0	0.479782	0.583203	0.044	0.14200	0.2520	0.53900	4.373
LandMinTemperature	1992.0	2.743595	4.155835	-5.407	-1.33450	2.9495	6.77875	9.715
LandMinTemperatureUncertainty	1992.0	0.431849	0.445838	0.045	0.15500	0.2790	0.45825	3.498
LandAndOceanAverageTemperature	1992.0	15.212566	1.274093	12.475	14.04700	15.2510	16.39625	17.611
LandAndOceanAverageTemperatureUncertainty	1992.0	0.128532	0.073587	0.042	0.06300	0.1220	0.15100	0.457

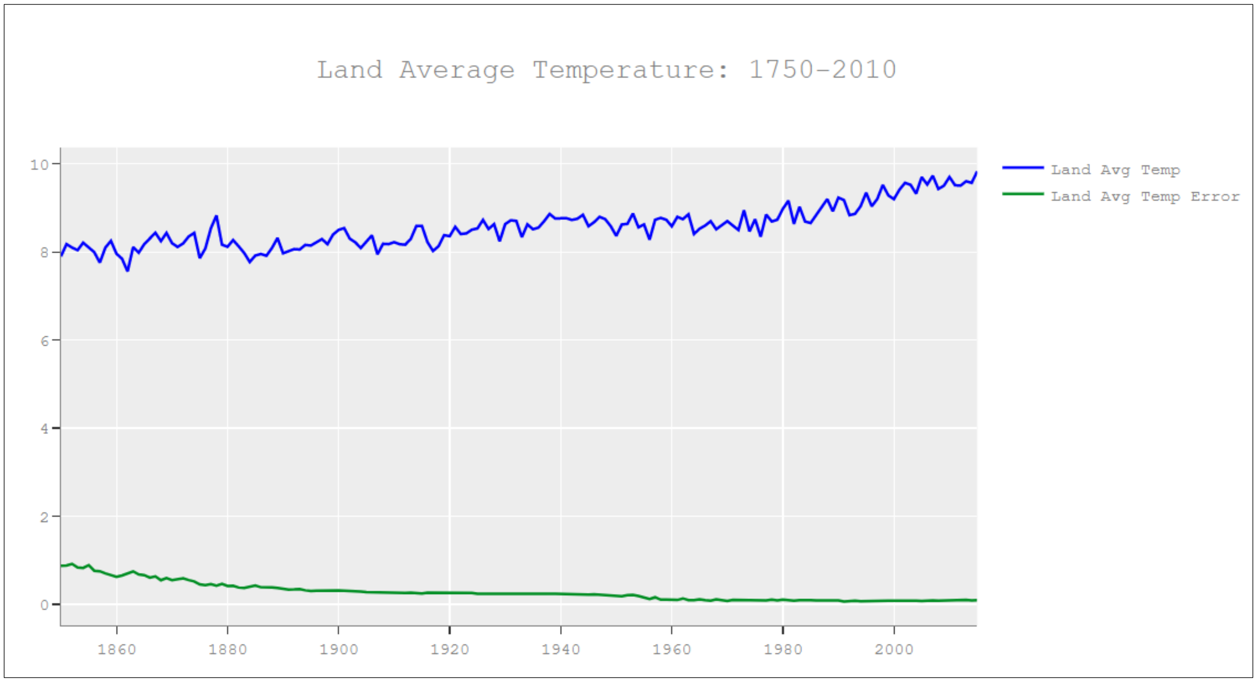
Like wise for the required data sets the descriptive statistics are measured and utilized wherever needed. Details are available in code note books that are submitted.

Data Analysis

Data Transformation

Import all the required packages to read the data as data frames and to plot the graphs also provide the results

After initial clean up and rearranging of the columns to be fit with analysis, the initial plot shows the increasing trend of Average Land Temperature



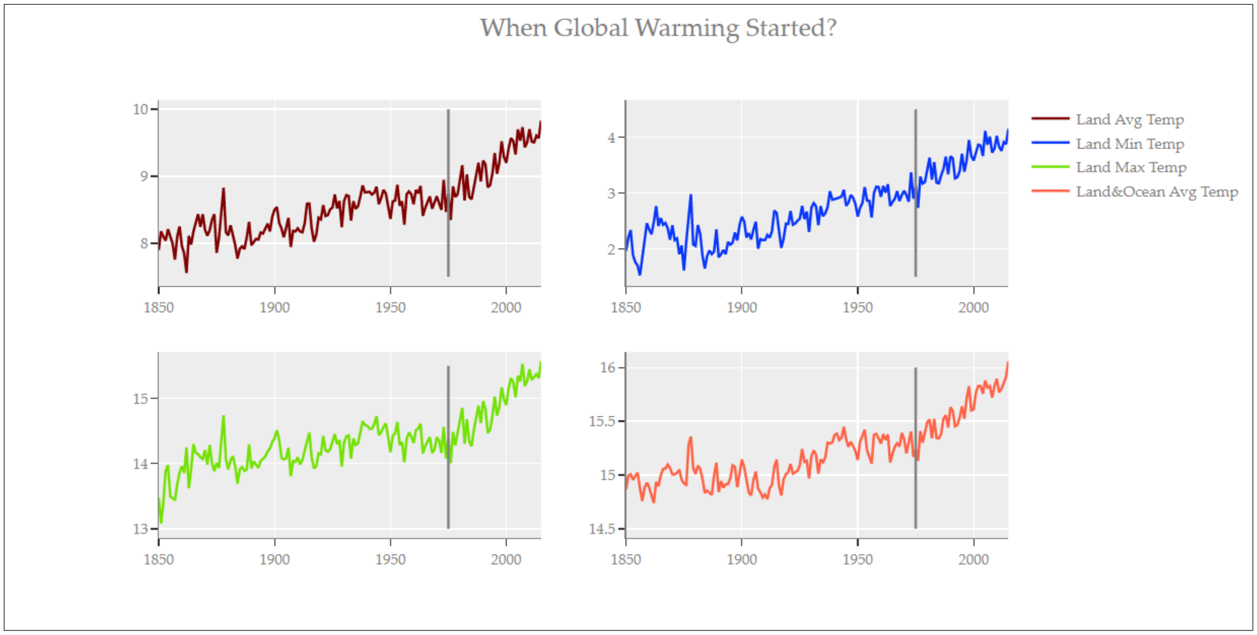
We can view there is not much change in 1700's to mid-1800's according to the geologists survey the global warming started noticed at the time of industrial revolution 1850 and reached to the level of concerning during the First World War however it did have wider impact noticed until the World War -II. However, the impacts are started to a greater observance across the globe is post 1975.

The data then recorded has attributes are little less than the comparisons that we draw in this modern information age. Hence considering data from 1950 to have better depiction of the when the global warming has started.

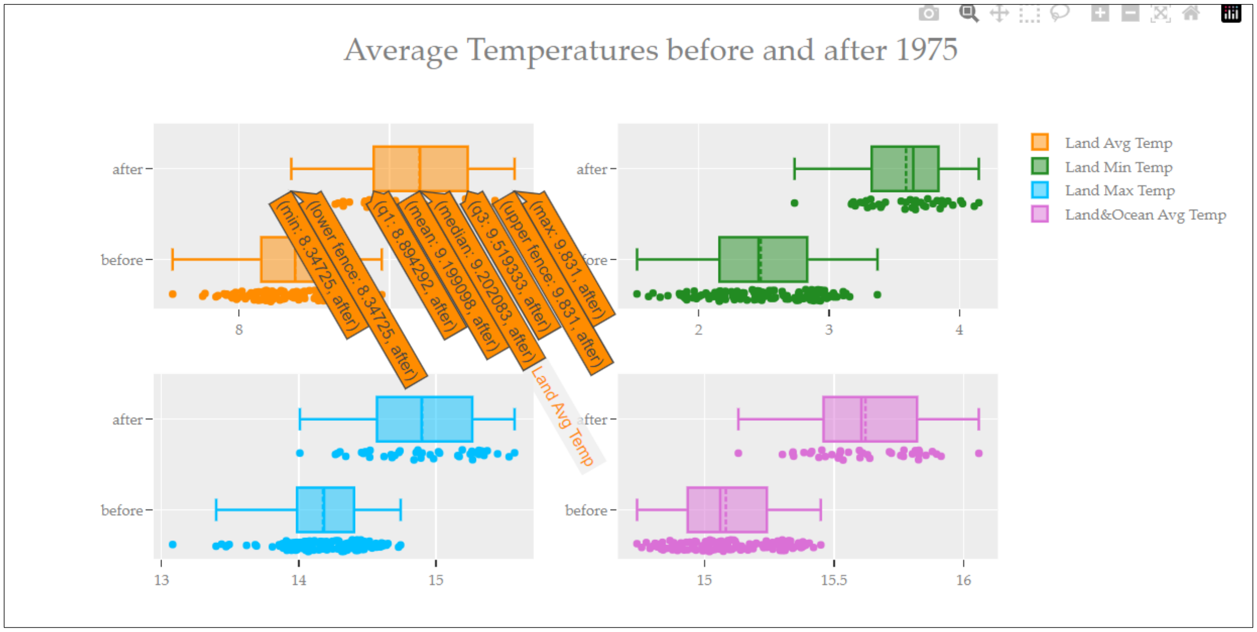
Analysis for the

Below plots representation starting from 1850's to the available data. How the averages of temperatures are raising and how the

For research question : Is there global warming and increase in trend?

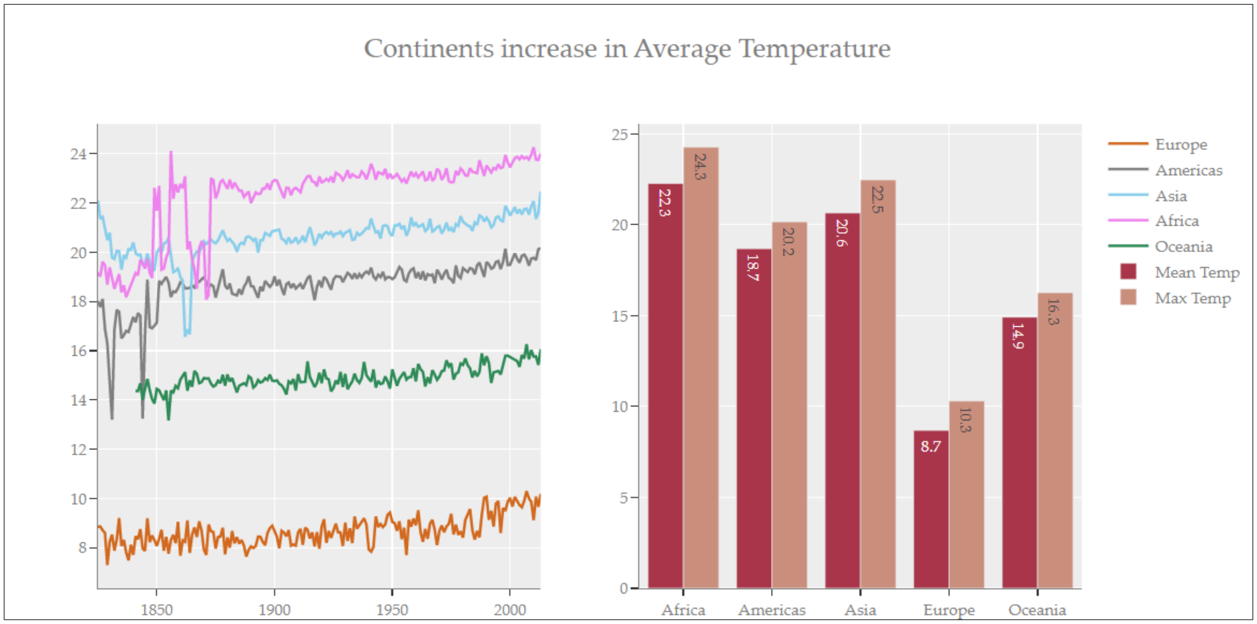


The below box plot provides the evidence that the impact of temperature is more predominant after 1975 in comparison with the 4 key variables temperature. (i.e, before 1975 and after 1975)



The above graphs and the analysis prove that there is global warming and the that is in increasing trend. Thus, provides result to research question

Trends with continents One of the research Question



This chart and scatter plot provides evidence the trend in rise in temperatures while mapping the geo locations to the ISO codes and tagging to the appropriate continent.

Is there any top continent increase in the temperature rise.

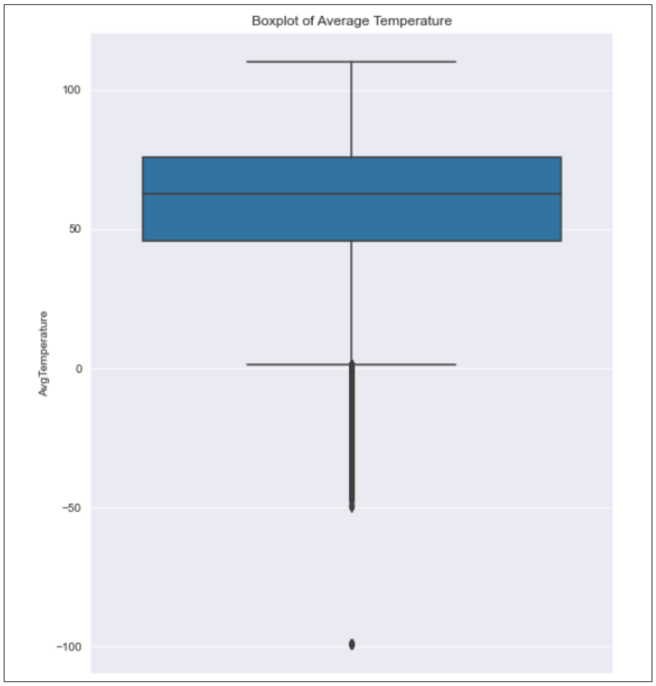
Nevertheless, it's the AFRICA's the top there are many reasons along the industrial revolution along with its geo location near to the equator and water evaporation. Etc.,

The above statement is from a research paper about Africa's hot temperature rise.

Checking for outliers

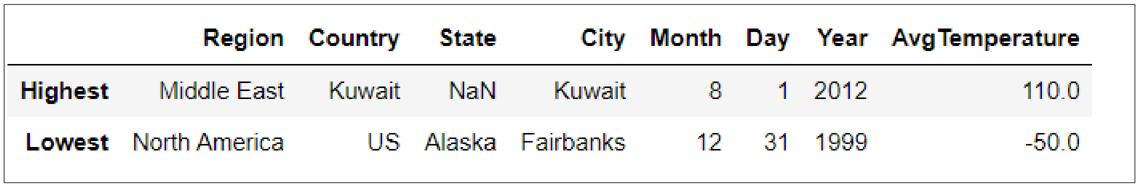
The daily temperature data set has been considered for the analysis however there are lot of missing data and for the certain time periods there has been highest records.

Outliers are present and can be seen in the figure below. Hence considering all the rows having '*AvgTemperature*' greater than -70°F for further analysis.



Another research questions the top hottest cities and coldest cities in the world

The below result from the analysis shows that Kuwait has the highest temperature and the city in Alaska “Fairbanks” recorded the lowest temperature



	Region	Country	State	City	Month	Day	Year	AvgTemperature
Highest	Middle East	Kuwait	NaN	Kuwait	8	1	2012	110.0
Lowest	North America	US	Alaska	Fairbanks	12	31	1999	-50.0

Note: One of the data correctness problem Kuwait is one of the country in middle east region however the geo location code is not having any appropriate match with the actual city it was shown as “none”

Those corrective methods as part of data scrubbing needs to be improved as part of my analysis.

Research question 4

Additionally, the results obtained are provided the top countries that are hot and cold

- The top five coldest countries in the world are: ['Mongolia', 'Iceland', 'Norway', 'Canada', 'Finland']
- The top five hottest countries in the world are: ['Guyana', 'Indonesia', 'Thailand', 'Nigeria', 'Haiti']

The above result is the from the datasets that are considered for the analysis of the current assignment.

However, the actuals are available

https://en.wikipedia.org/wiki/Highest_temperature_recorded_on_Earth

The notebook “IST 718_Final Project_Analysis2_TimeSeries_Forecasting”, provides different view of the decomposition analysis of the average temperature data for the further analysis,

Hypothesis of the data before starting the data modelling

Imported the required model packages for python

- Null Hypothesis (H0): If failed to be rejected, it suggests the time series has a unit root, meaning it is non-stationary. It has some time dependent structure.
- Alternate Hypothesis (H1): The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary. It does not have time-dependent structure.
- p-value > 0.05: Fail to reject the null hypothesis (H0), the data has a unit root and is non-stationary.
- p-value <= 0.05: Reject the null hypothesis (H0), the data does not have a unit root and is stationary.

Data Modeling

Experimenting to buuild a model that describes those relationships more mathematically right and proven with results.

As learnt in the course and labs ARIMA or SARIMA are best suited models for the time series analysis of such continuous research related data.

Autoregressive Integrated Moving Average Model(ARIMA)

An ARIMA model is a class of statistical models for analyzing and forecasting time series data.

It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.

AR: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations.

I: Integrated. The use of differencing of raw observations (e.g., subtracting an observation from an observation at the previous time step) to make the time series stationary.

MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Each of these components are explicitly specified in the model as a parameter. A standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.

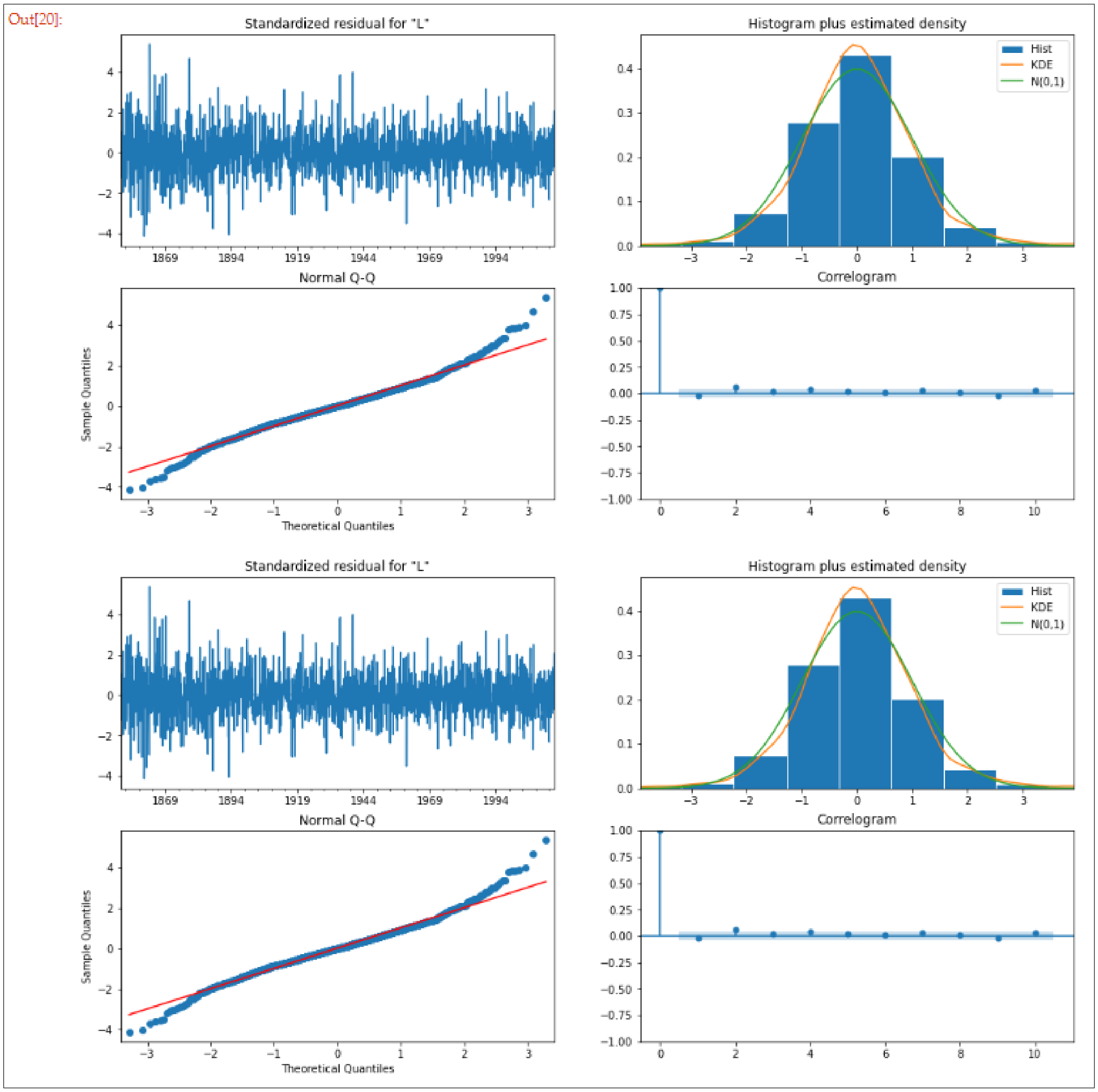
The parameters of the ARIMA model are defined as follows:

p: The number of lag observations included in the model, also called the lag order.

d: The number of times that the raw observations are differenced, also called the degree of differencing.

q: The size of the moving average window, also called the order of moving average.

A linear regression model is constructed including the specified number and type of terms, and the data is prepared by a degree of differencing to make it stationary, i.e., to remove trend and seasonal structures that negatively affect the regression model.



The results are pretty much evident the datasets that are considered for modelling are fit and provided the meaningful outcome of the prediction for temperature rise.

The train, test , split data sets are mentioned in the Jupyter notebook.

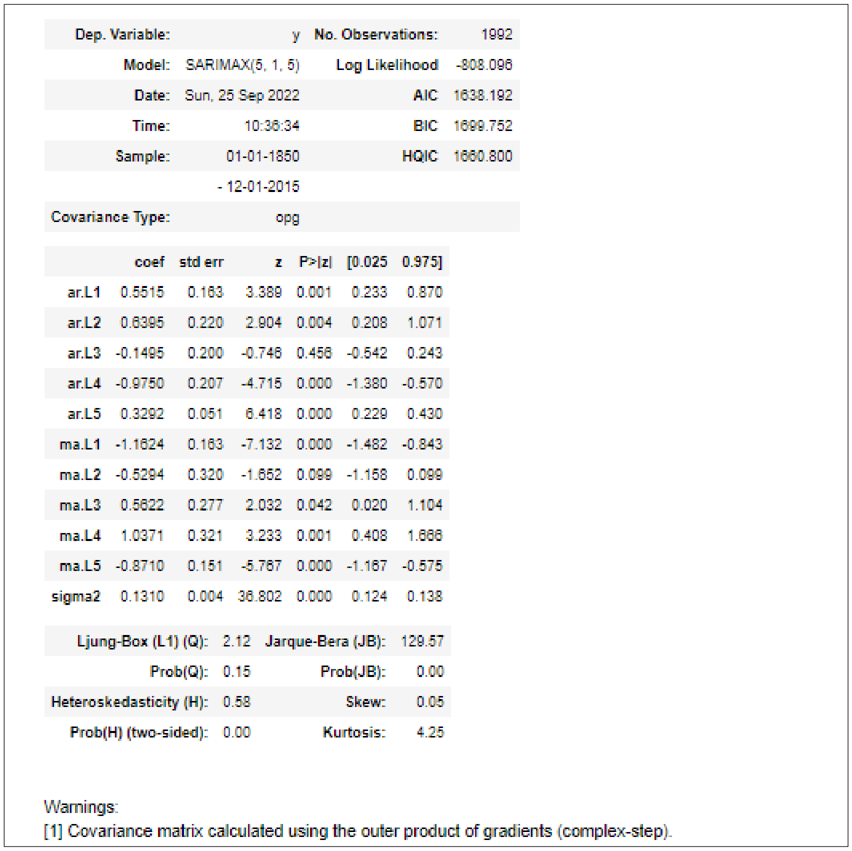
Request to refer the note book named

“IST 718_Final Project_Analysis2_TimeSeries_Forecasting”.ipynb.

Further analysis with Auto ARIMA

To be precise the lower the ‘p’ value and closest to “zero” means that the data is not that significant

However, with ARL1 coefficient shows the data used is valid to provide forecast analysis.



```
     Dep. Variable:                 y   No. Observations:     1992
            Model:    SARIMAX(5, 1, 5)  Log Likelihood     -808.096
             Date:    Sun, 25 Sep 2022             AIC     1638.192
             Time:            10:36:34             BIC     1699.752
           Sample:          01-01-1850            HQIC     1660.800
                          - 12-01-2015
 Covariance Type:                 opg

              coef   std err        z   P>|z|  [0.025   0.975]
     ar.L1   0.5515    0.163    3.389   0.001   0.233    0.870
     ar.L2   0.6395    0.220    2.904   0.004   0.208    1.071
     ar.L3  -0.1495    0.200   -0.746   0.456  -0.542    0.243
     ar.L4  -0.9750    0.207   -4.715   0.000  -1.380   -0.570
     ar.L5   0.3292    0.051    6.418   0.000   0.229    0.430
     ma.L1  -1.1624    0.163   -7.132   0.000  -1.482   -0.843
     ma.L2  -0.5294    0.320   -1.652   0.099  -1.158    0.099
     ma.L3   0.5622    0.277    2.032   0.042   0.020    1.104
     ma.L4   1.0371    0.321    3.233   0.001   0.408    1.666
     ma.L5  -0.8710    0.151   -5.767   0.000  -1.167   -0.575
    sigma2   0.1310    0.004   36.802   0.000   0.124    0.138

     Ljung-Box (L1) (Q):  2.12   Jarque-Bera (JB):  129.57
                Prob(Q):  0.15           Prob(JB):    0.00
Heteroskedasticity (H):  0.58               Skew:    0.05
   Prob(H) (two-sided):  0.00           Kurtosis:    4.25


Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Analysis with R

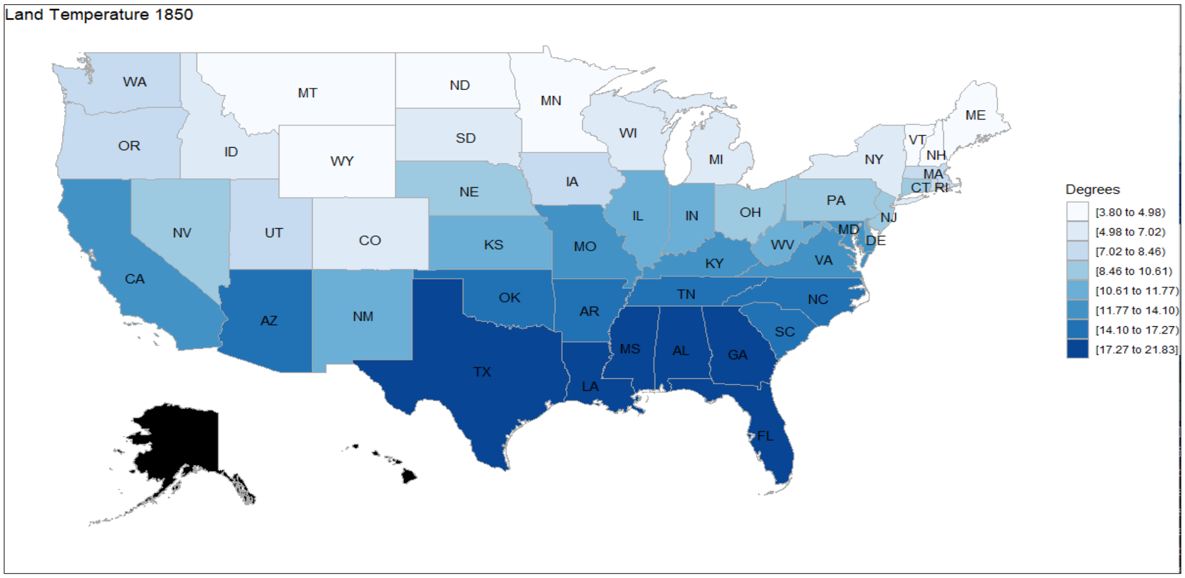
I tried to get the results of USA states temperature rise by using R,

There are few more graphical representations inline with analysis using python are provided as part of output those are available in ipynb notebooks that are uploaded.

This attempt to describe how the rise is and how the graph shows it for a better view.

Only few data corrections are done.

- ✓ At the time 1850 there are not many states in United Nations
- ✓ For Instant States like New Mexico, Colorado, Dakota Territory …etc are not included in the states
- ✓ Removed Alaska and Hawaii from analysis

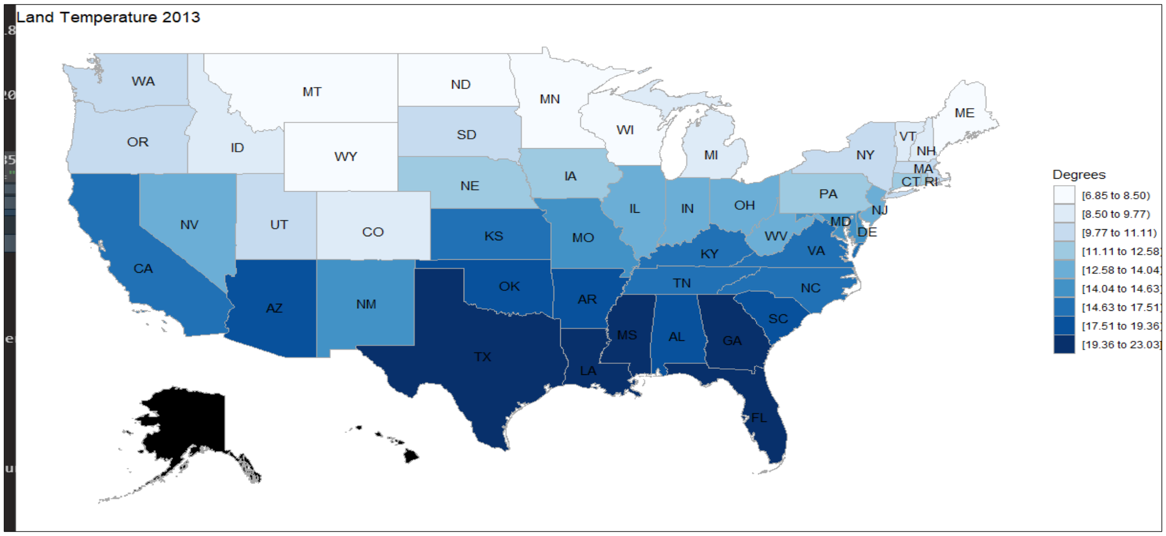


The above picture shows that the max temperature was set at 21.83 in degree Celsius

However, the modern states expanded and the likely hood of the temperature almost a decade (2013) has been used to depict the below map

The key observations are

- There is rise in max temperature to 23.03 degrees which is almost 2 degrees rise in 175 years in the states. But it has drastic impact on few states in the country
- The most affected are down south states like Texas, Florida resulting heavy storms along with few west coast states like California wide spread of wild fire.



Conclusion

It’s pretty evident the average land temperature is rising over years it will be having wide spread impacts on over all climate like water scarcity, CO2, air pollution, sea level rise, snow melt in Greenland, droughts and followed by other natural calamities.

Scientists are proposing lots of options for better living and restoring natural resources like choose alternate approaches.

This assignment (final project) a greater opportunity for me to experiment with large datasets and learn different methods of modelling the data .

Improvements Needed

Much improved way of data scrubbing

Need to pay much attention to evaluate the models by checking the accuracy, precision

I tried to get the confusion matrix to gain confidence on data that was used for analysis however hit with multiple roadblocks.

References

Along with Asynch course material also referred the available resources from internet.

https://medium.com/swlh/temperature-forecasting-with-arima-model-in-python-427b2d3bcb53

https://valueml.com/global-warming-prediction-using-machine-learning-in-python/

https://www.section.io/engineering-education/building-a-time-series-weather-forecasting-application-in-python/

https://scied.ucar.edu/learning-zone/climate-change-impacts/predictions-future-global-climate

https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

https://www.statsmodels.org/dev/generated/statsmodels.tsa.arima.model.ARIMA.fit.html

https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

https://towardsdatascience.com/single-and-multi-step-temperature-time-series-forecasting-for-vilnius-using-lstm-deep-learning-b9719a0009de

END OF THE DOCUMENT