Climate Change - Analysis Report

Anil Kumar Pallekonda
CISC 520-90 Late Fall 2016
Deprtment: Analytics
E-mail: apallekonda@my.harrisburgu.edu
Student ID:157959
University: Harrisburg University

Vinaya Rajappa
CISC 520-90 Late Fall 2016
Deprtment: ISEM
E-mail: VRajappa@my.harrisburgu.edu
Student ID:171494
University: Harrisburg University

Avneesh Sharma
CISC 520-90 Late Fall 2016
Deprtment: Analytics
E-mail: AvnSharma@my.harrisburgu.edu
Student ID: 186585
University: Harrisburg University

Abstract: Global temperatures and its trends. How global temperatures are correlated with the USA temperatures? What will be the USA temperatures in the next 10 years? Clustering countries based on their average temperatures.

INTRODUCTION

Global Warming: According to the life science explanation[3], Global warming is the term used to describe a gradual increase in the average temperature of the Earth's atmosphere and its oceans, a change that is believed to be permanently changing the Earth's climate. In general people say[1],*"Climate change is the biggest threat of our age while others say its a myth based on dodgy science."*

As a team, we would like to explore the global temperatures to accept or reject the above mentioned statement. Apart from this, as a part of CISC 520-90 Late Fall 2016 course project, we (project team) have selected *Global Temperature* data set from kaggle (open data source)to perform the below tasks.

- Data Preparation
- Descriptive analysis of the data
- Hypothesis testing
- Forecasting analysis
- Clustering

This data has been collected from various weather stations across the world. The data set has data from the year 1875 to the year 2015 and has been categorized into 5 different categories. Out of the five categories, we have chosen *"average global temperatures"* and *"average temperatures by country"* (subsets of main data set) for our analysis. We have considered data from the year 1900 to the year 2012 for analysis purpose. Till the year 1900, all the countries do not have complete data and both data sets are having common data till the year 2012, this is the reason we have restricted our data from the year 1900 to 2012. We have incorporated results, interpretations and their corresponding visuals in this report. We have used statistical software *"R"* and *"R-Studio"*(IDE) to perform all our analysis.

DATA PREPARATION

Data preparation is a crucial part in any data analysis portfolio. In this step, we have performed several tasks to load, understand and clean the data.

We have used below *"R"* commands to load the data into *"RStudio"*.

```
global_climate ← fread("GlobalTemperatures.csv",header = TRUE)
global_temp_country ← fread("GlobalLandTemperaturesByCountry.csv", header = TRUE)
```

When we loaded the data into R, we have observed that the global average temperature data has $3,192$ observations with 11 variables. Global temperatures based on each country has $577,462$ observations with 6 variables. When we observed structure of the data, the dt column had been loaded in the format of character. We have changed dt column into date format using *as.date()* function. From this, we have extracted *year* and *month* columns which will be used in further analysis.

As a part of data preparation, we have created new columns and named them as *"Year"* and *"Month"*. In this analysis, we have not utilized few columns in the data set as they are not useful in our scope of analysis. These columns are *"MinLandAverageTemperature", "MaxLandAverageTemperature", "LandOceanAverageTemperature", "LandOceanAverageTemperatureUncertainty"* and *"LandAverageTemperatureUncertainty"* in global temperature data set. We have omitted column *"AverageTemperatureUncertainty"* from country data set in this analysis. To know the impact of uncertainty in the temperatures, we have used *"Average Temperature Uncertainty"* of the major cities data set.

Before checking for the missing values, we have restricted the data from the year 1900 to the year 2015. The number of observations got reduced from $3,192$ to $1,392$ and from $577,462$ to $329,868$. After that, we have checked for the missing values using the below *"R"* code:

```
sapply(unique(global_climate_1900), function(x)any(is.na(x)))
sapply(unique(global_temp_country_1900), function(x)any(is.na(x)))
```

The results show that there are missing values in *"AverageTemperature"* and *"AverageTemperatureUncertainty"* variables. When we check for the total number of missing values, there are 0.32% to the total number of observations. This value is negligible amount when compared with total number of observations, so we have ignored/removed those observations.

Exploratory Analysis

To understand the data in a better way, we have plotted different graphs/charts for each data segment. Apart from these visualizations, we are going to compute few basic statistics on several temperatures i.e. global temperatures, country wide temperatures and year wise temperatures. Finally, under exploratory analysis, we are going to perform hypothesis test which will confirm the presence of global warming across the world and in the USA. With these visualizations, we can conclude if the existence of global warming is true or not.

When we plot the global average temperatures on the graph, the graph looks like in Fig.1.

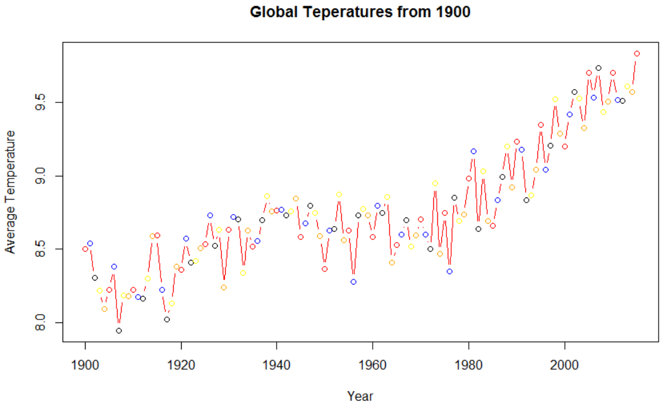


Fig. 1. Global Temperatures

On observing the above graph, we can say that there is a significant amount of rise in the average temperatures since the year 1900 to the year 2015. To plot this, we have computed the average temperatures based on year (12 months average temperatures set to the respective year). Though there are few places where the temperatures dropped across the world, after the year 1980 there is a significant growth in the global temperatures.

When we plot the histogram of global average temperatures and normal density curve to that. The histogram depicts that the data has been distributed normally. The basic statistics of computed yearly average land temperatures are as follows. The minimum temperature is 7.947^oC and maximum is 9.837^oC. The 25% or 1st quartile value is 8.514^oC, 50% or median value is 8.699^oC, 75% or 3rd quartile value is 8.956^oC. We have used box plot and box plot matrix to identify outliers in the land average temperatures. There are no outliers in this global average temperatures.

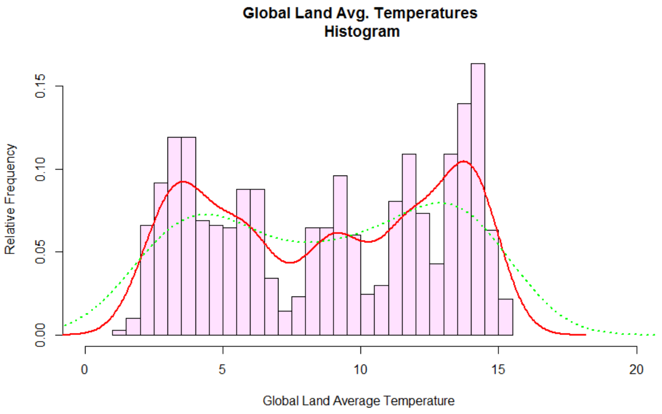


Fig. 2. Global Temperatures Wide dispersion

When we plot the global land average temperatures(monthly) with 50 bins from the year 1900 to the year 2015, the histogram looks like in Fig.2. By looking at this graph, we can say that it has bi-model distribution.

When we look at the country wise data, there are few countries which are having low average temperatures from the year 1900 to the year 2015 and few countries which are having very high average temperatures in the same time period. When we look at the average temperatures across the countries, the average temperatures is rising year by year. We can confirm this using the graph in Fig.3.

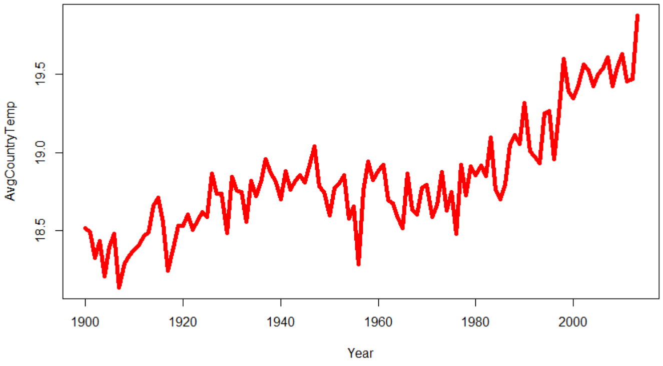


Fig. 3. Average Temperatures by Country

We have plotted 20 countries which are having the least average surface temperatures and 20 countries which are having the highest average surface temperatures in the Fig.4 & Fig.5. When we look at the below graphs, we have found that *"Djibouti"*, *"Mali"* countries are having the highest average temperatures on the earth's land surface. Countries *"Greenland"*, *"Denmark"* has the lowest temperatures on the earth's land surfaces.

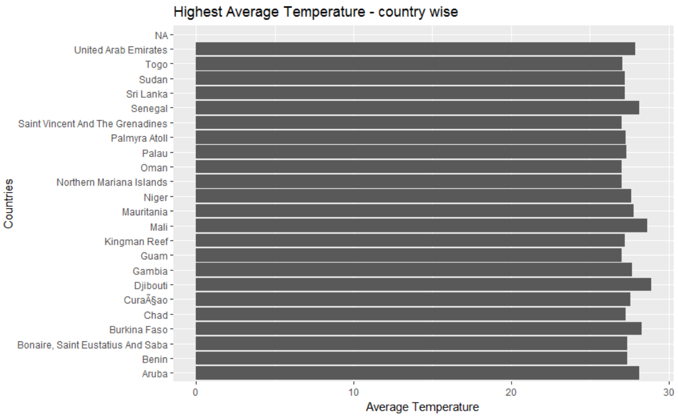


Fig. 4. Countries With Highest Temperatures

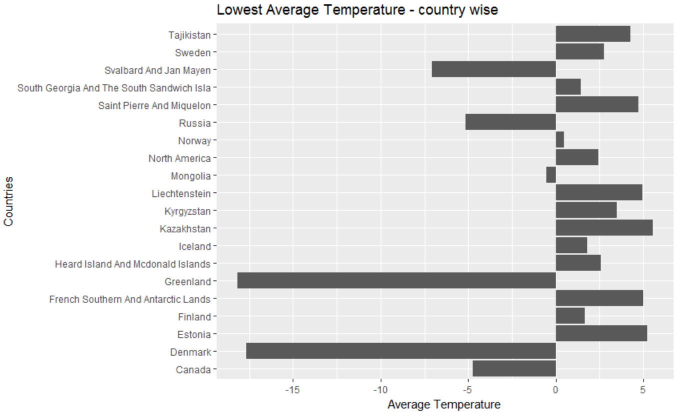


Fig. 5. Countries With Lowest Temperatures

We have used *order* function to sort the temperatures from low to high. In the graph, countries are arranged in alphabetical order on the y-axis.

Now let us perform some exploratory analysis on the major cities dataset. Using the *glimpse()* function we find that the dataset has 239,177 observations and 7 variables. The variables are

- dt, type = factor
- Average Temperature, type = double
- Average Temperature Uncertainty, type = double
- City, type = factor
- Country, type = factor
- Latitude, type = factor
- Logitude, type = factor

It seems as if the variable *dt* or date has been captured in a factor format. Using the function *separate()*, we split the date into *Year, Month and Date*. Now that we have the necessary data structure, let us look at the quality and completeness of the dataset. The function *is.na()* shows that the entire dataset has 22,004 observations that have missing values. Since the percentage of missing values is relatively very small, we have chosen to omit them from the analysis using *na.omit()*.After omitting, we are left with 228,175 observations.

Now that the dataset has a clean structure and has no missing values, we performed some exploratory analysis to get a complete idea about the dataset. Using the *unique()* function, we found that the dataset has temperatures recorded for 100 cities from 49 countries. The data ranges from 1849 - 2013, but since the temperature uncertainty is high, we will omit 1849. We then wanted to find out how the average temperature across all the years trended, which we've plotted in Figure 6. It shows the average temperature for all the cities for each individual year. It can be seen that it took almost 79 years for the average temperature recorded to change from 18 °C to 19 °C, but its only taken about 55 years for the average temperature to change from 19 °C to 20 °C. This indicates that there are hints of climate change across major cities of the world as the last 13 years of recorded data have temperatures higher than 19.5 °C.

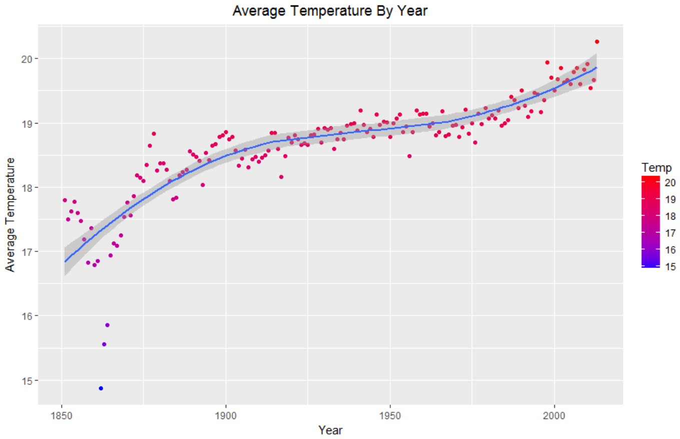


Fig. 6. Average Temperature by Year

We also wanted to understand what impact, if any, did the variable *Average Temperature Uncertainty* have or if it offers any insight. Hence, to fully understand its impact, we have plotted the variable but also factored in the *Country* and *City* as can be seen from Figure 7.

As could have been predicted, in general the uncertainty is high in the last 2 centuries, which can be attributed to the techonological reasons. However the countries with the highest variability in uncertainty are Italy, Russia, United Kingdom, United States, Canada, Bangladesh and Morocco. It will be interesting to take this variable into account and maybe do the same analysis on a subset of the data for the recent years in order to get a better estimate of variablity. We've taken a look at how the temperature trend looks as a single average value

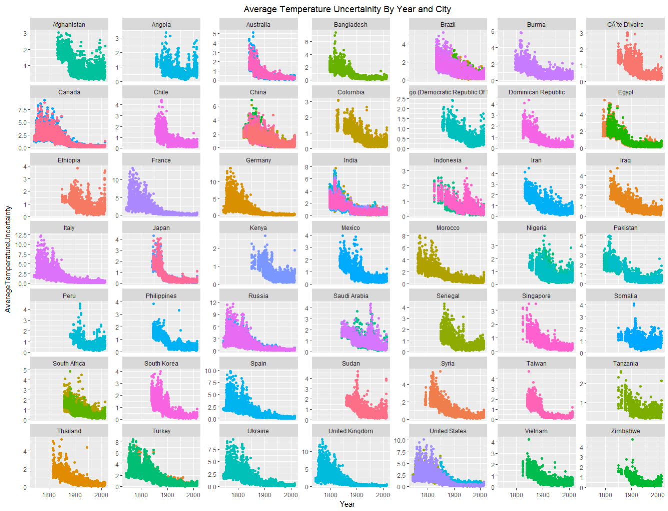


Fig. 7. Average Temperature Uncertainty by Year and City

for the entire year. It might be interesting to analyze the trend of all the available data points plotted together, which is what we can see in Figure 8. We think it shows that there are no distinct outliers that seem to be present in the dataset from this chart.

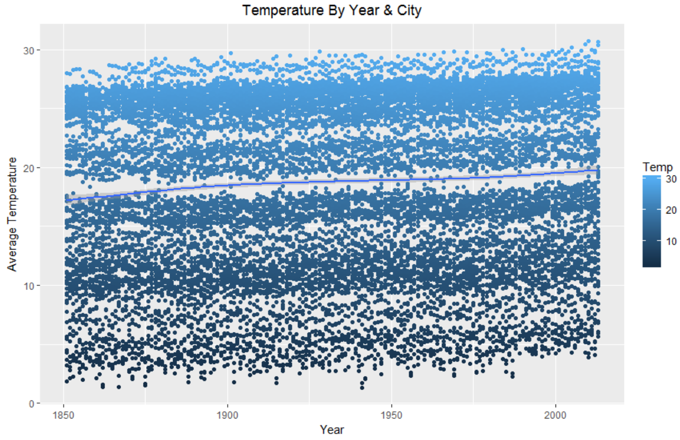


Fig. 8. Average Temperature by Year and City

However, Figure 8 doesn't help us understand much of anything apart from that. Hence to get a better understanding, we plotted Figure 9, which should help give an insight about the same dataset with the added visual aids of *City* and *Country*. Each *City* has a different color plotted within the *Country* facet.

It definitely has some interesting insights. It helps to understand the Average temperature trend for each City, how the different cities compare to each other, which cities have the highest average temperature vs the ones that have the lowest.

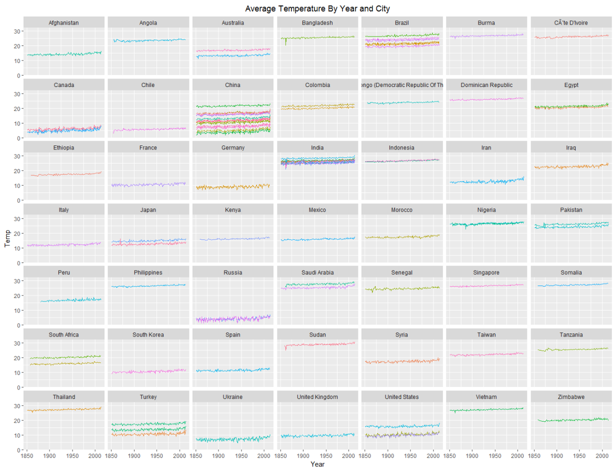


Fig. 9. Average Temperature by Year and City and Country

As it could've been guessed, the countries located near the equator have a higher average temperature compared to others. Russia and Canada have the lowest average temperatures and Saudi Arabia and India have the highest. If you look at the trend for United States, you will see 2 distinct zones of temperature trend. Let us investigate this further.For a better context, let us plot this on a map to understand the comparison of temperatures of the major cities:

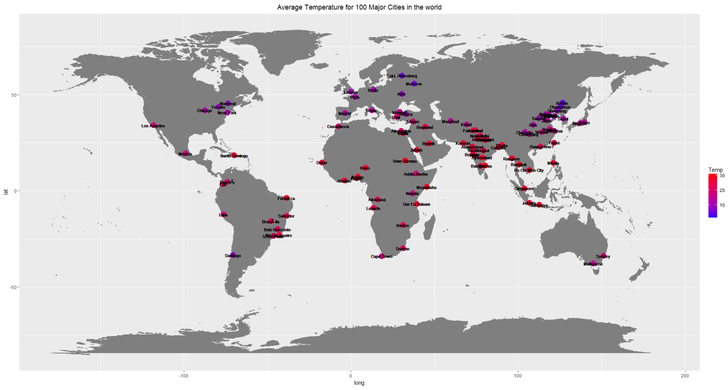


Fig. 10. Average Temperature by City

This figure shows the average temperature across the years for all the 100 major cities that are recorded in the dataset. It shows a nice comparison amongst the temperature levels in the various cities. It shows which cities have a higher temperature than others.This chart shows that data has been collected for 3 USA cities. Let us try and get a better granular understanding of those 3 cities.

Analysis on the USA

Let's dig into more granular level. Let's analyze the United States of America temperature patterns. Our intention is to analyze the USA temperature patterns over the years and we would like to compare the USA temperatures with global temperatures. We would like to compare temperature growth rate between the last 50 years and the last 10 years. When we look at the basic statistics, for the USA, average temperature from the year 1900 to the year 2013 is 8.95^oC across all the years with the standard deviation of 8.9^oC. The minimum temperature is -6.74^oC and maximum temperature is 23.01^oC. The skewness of the USA temperatures from the year 1900 to the year 2013 is -0.01^oC and kurtosis is -1.44^oC.

First, lets compare the USA surface temperatures with global surface temperatures. When we look at the graph in Fig.11, we can say that most of the time, the USA temperatures are equal or bit higher to the global temperatures(though it has few up and downs between the years) but after the year 2008, there are significant changes in the USA average temperatures. This is the most worrying part in terms of climate change. With this information we can say that there is global warming across the world including the USA. Earth temperatures are increasing drastically in the recent years.

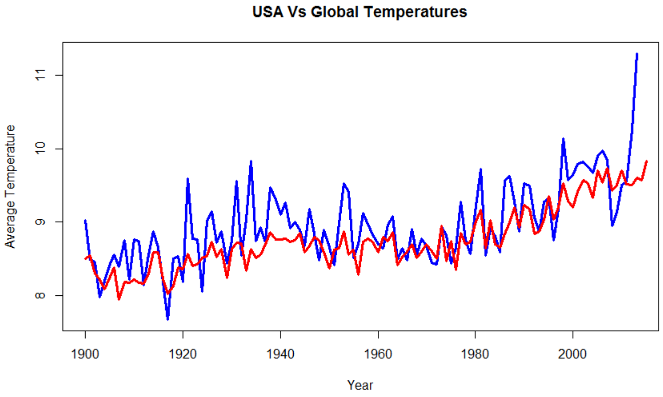


Fig. 11. Global Temperatures Vs. USA Temperatures

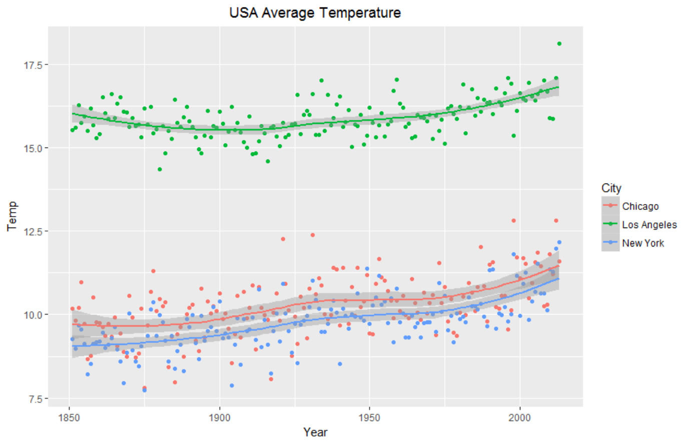


Fig. 12. USA City Temperatures by Year

Plotting an *Average Temperature* chart for just United States gives us Figure 12. The 3 cities that have data collected for are *Chicago, New York and Los Angeles*. Upon closer inspection, there seems to be a lot of variability in the trends. It also feels that there might be some outliers present when we zoom in to such a closer level of detail. However, quite interestingly, the temperature gain levels off from about 1935 to 1970 before rising rapidly for *Chicago*. All 3 cities record a rapid rise in the average temperature after 1975. It also seems that the variability in *Chicago's* temperature is higher compared to *New York* despite having almost similar average temperature trends. Hence, we can say with a certain degree of confidence, that global warming is a true phenomenon.

Hypothesis Testing

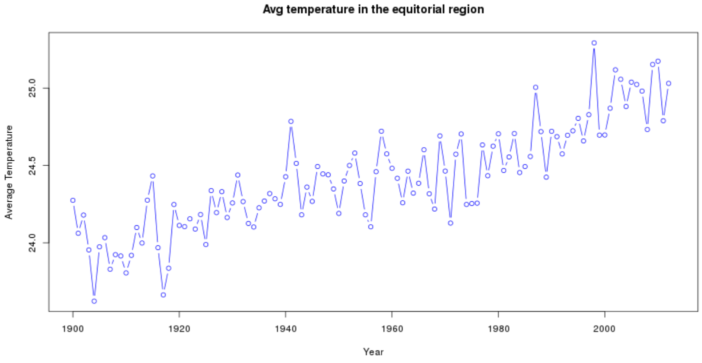


Fig. 13. Equatorial Region Temperatures Vs. USA Temperatures

We can see from the Fig.14 that the mean temperature of the polar region is steadily increasing. The alarming fact is that that average temperature is slightly above freezing point. This correlates to the recent incident of largest iceberg poised to break away from the North Antarctic[4]. This would cause the global sea level to rise by 10cm which can submerge small islands nations. To show that global warming is a recent phenomenon driven by industrial age, we can look at the history of any major city in a developed world.

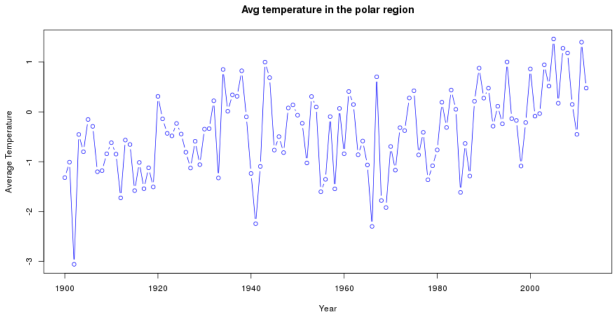


Fig. 14. Polar Region Temperatures Vs. USA Temperatures

When we relate these plots to the real incidents in the world, we can say that global warming will be a biggest threat in future.

Correlation

When we calculate the correlation between Global year wise temperatures and the USA year wise temperatures from the year 1900 to the year 2012, the *"Pearson"* correlation coefficient is 0.79. Correlation coefficients for *"Spearman"* and *"Kendall"* methods are 0.76 and 0.57 respectively. When we consider the Pearson correlation, there is a strong correlation between global average temperatures and the USA average temperatures. We can say that, 79% of global temperatures can be explained using the USA temperatures and vice versa.

Forecast Analysis

As this is time series data, we have converted the data set into time series object in order to perform further analysis. We have used *ts()* function to convert the data into "ts" object and this is useful to perform forecast analysis on time series data. To know the monthly patterns across all the years, we have used *monthsplot* graph. When we look at the monthly plot, temperatures are slightly skewed to the left. When we observe monthly patterns, July and August months have the highest temperatures across the years and January and December months have the lowest temperatures. When we observed Fig.15, there are more fluctuations in January, February and December months when compared with June, July and August months so there are more fluctuations in the lowest temperatures of recent years.

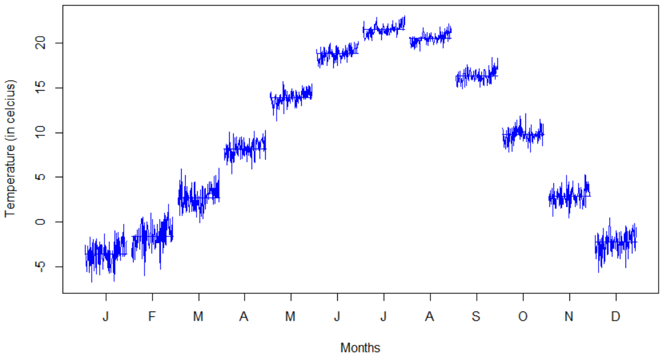


Fig. 15. Month Plot on USA Temperatures

We have plotted Seasonality chart to check the seasonality & trend of the ts object which is shown in Fig.16. We have used *"stl()"* method to calculate the trend and seasonality. When we look at below seasonality and trend chart, we can say that there is a strong seasonality in the chosen temperature data set. When we look at the trend line, overall trend is upwards from the year 1900 to the recent years. When we look at the reminder graph in *Seasonality & Trend* plot, the reminder graph waves are smaller at the end after the year 2000.

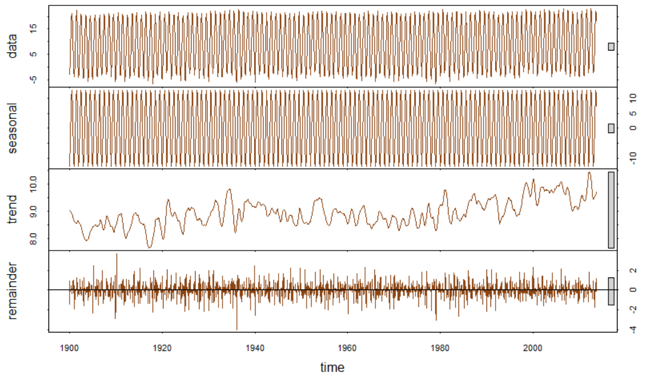


Fig. 16. Seasonality & Trend

There are several methods to forecast time series data i.e. 'Moving average', 'ARIMA', 'trend models', 'simple', 'linear', 'quadratic', 'random walks', 'Seasonal exponential smoothing', etc., etc. We have chosen, moving average to forecast the temperatures in the USA. When we look at the graph in Fig.17, the forecasted temperatures are increasing(with 95% confidence intervals). In the graph, blue color lines represents forecasted values and black lines represents historical values. When we look at grayed portion(this represents upper level and lower level of confidence interval) in the graph, the average temperatures are too high when compared with historic temperatures.

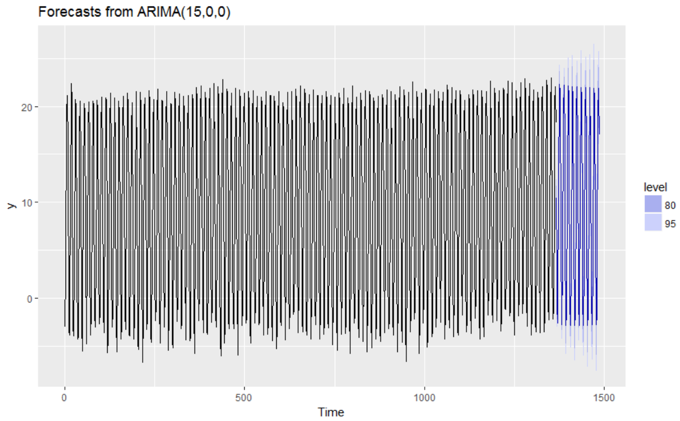


Fig. 17. 10 Years USA Forecasted Temperatures with Historic Values

We have forecasted temperatures for the next 10 years for the country USA. We can see the forecasted values represented in a different color in Fig.17 & Fig.18. We have plotted 10 years forecasted value with 1 year historic values to identify the patterns in a better way. When we look at Fig.18, the purple and blue shaded areas represents the 95% and 85% confidence intervals respectively. Pinnacle of the confidence interval is raising year by year which represents average temperatures will raise more in the future.

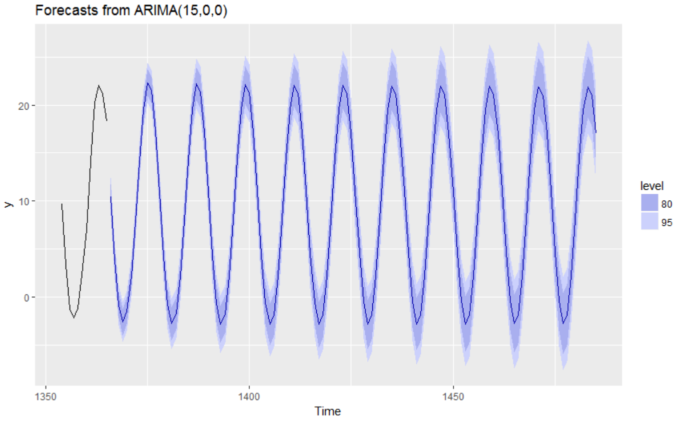


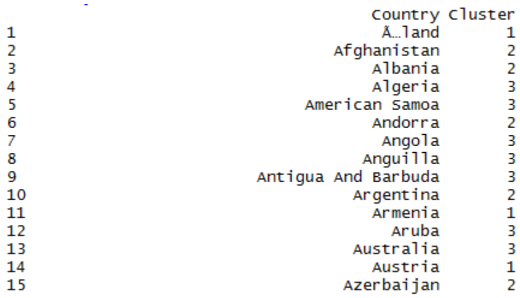
Fig. 18. 10 Years USA Forecasted Temperatures

To find out the best fit and to compute the forecasted values, we have used *forecast*(a library in 'R') package. The best fit was computed using *ARIMA()* function and *forecast()* function is used to calculate forecast values. Finally, we have used *plot()* function to plot the forecasted values with historic values.

Clustering

Clustering is one of the most commonly used technique in data mining area. As a part of this project analysis, we would like to bifurcate the countries into 5 clusters. This can be done by calculating the distance between the points when we plotted the countries average temperatures on temperatures space. We have chosen "Euclidean distance" method to calculate the distance between the points. The distance calculation goes like this. After plotting the country points in the temperatures space, when we have decided to divide the countries into 5 groups, randomly 5 points will be selected. From those points to every other point in the space, distance will be calculated. The points which have least distance from the selected points are formed as one group. Later the centeriod will be calculated for each group and these centeriods are treated as cluster points in that arranged group. Again, the distance calculation and grouping process will be performed as mentioned above. For every cycle, centeriod and cluster points will change. At certain point, the points in cluster will not change. At that point, the algorithm stops calculating distance process and produces final groups which we call as clusters. In our scenario, countries are grouped into 5 segments. In our scenario, algorithm run for 3 iterations to complete the process.

We have added a column to the country wise data set and named it as *cluster*. We have added cluster values(1 to 5) to the new column(cluster) of country data set. Fig.19 represents few rows of data with countries with their respective cluster values. The cluster value 1 represents that 1st country i.e. "land" will come under cluster 1; the country "Argentina" comes under cluster 2 and the country "Angola" comes under Cluster 3. We have showed first 15 rows of data with cluster values from country temperatures data set.



```
                                   Country Cluster
1                                    Ã…land       1
2                              Afghanistan       2
3                                  Albania       2
4                                  Algeria       3
5                           American Samoa       3
6                                  Andorra       2
7                                   Angola       3
8                                 Anguilla       3
9                      Antigua And Barbuda       3
10                               Argentina       2
11                                 Armenia       1
12                                   Aruba       3
13                               Australia       3
14                                 Austria       1
15                              Azerbaijan       2
```

Fig. 19. Clustered Countries

When we plot the clusters in the temperatures space with different color segmentation, the plot looks like in Fig.20. This graph is unable to tell a clear story of the picture.

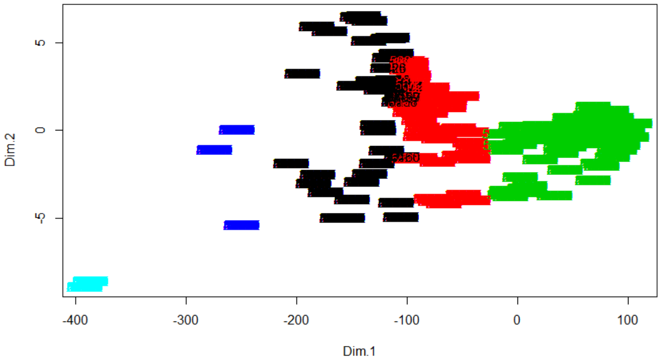


Fig. 20. Clustered Countries

To represent a better way and to understand with clear picture, we have plotted all the countries with different color gradient on the world map. The world map looks like in Fig.21. When we look at Fig.21, we can identify that several countries are represented with same color that means those countries come under single cluster. The lowest temperature cluster is represented with color "Azure" and the highest temperature cluster is represented with color "Red". We have used color gradients to represent all 5 clusters(Lowest - Azure, Highest - Red).

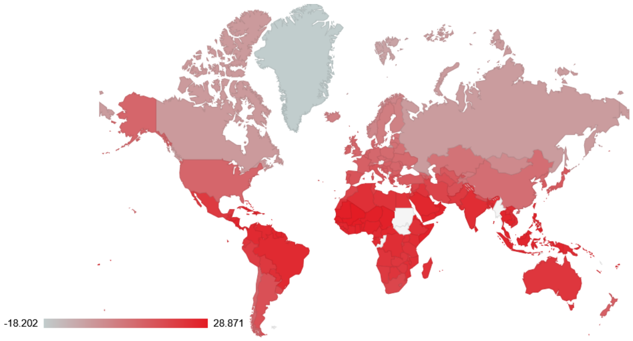


Fig. 21. Clustered Countries

Conclusions

As mentioned in the hypothesis, global warming exists on Earth's surface. This might be converted into severe threat to the future generations. The graphs under exploratory analysis will support this statement. The graphs represented under ”USA Analysis” section supports that there is a temperature rise in the USA since year 1900. With this information, we can confirm that global warming exists in the USA due to increase in temperatures year by year. Forecast analysis results are confirming that in future there is more uncertainty in climate temperatures. When we divided countries into groups, few countries in South Asia and few countries in Africa region have the highest temperatures.

Assumptions

This data has been downloaded from the internet source with an assumption of everything being presented in the data to be correct. We have not considered the uncertainty while forecasting the future temperatures. There is more scope to perform in-depth analysis on global temperatures data set(on a whole). Due to time constraint, we have not focused on in-depth analysis of global temperatures data set.

References

[1] http://www.kaggle.com
[2] http://berkeleyearth.org
[3] http://www.livescience.com
[4] http://www.bbc.co.uk/news/science-environment-38522954