EXPLORATORY DATA ANALYSIS

ON

CLIMATE CHANGE: EARTH SURFACE TEMPERATURE DATA FOR ASIAN COUNTRIES

Ravi Teja Buddabathuni | 09-12-2016

Executive Summary

This report summarizes the data analysis results associated with the Climate Change: Earth surface temperature changes in Asian countries. The purpose of this report is to document the analysis results of the year wise climatic changes or earth's surface temperature changes for Asian countries.

Data set is taken from the kaggle website from this link. Later the dataset is being cleaned by using MS Excel, Here the priority is taken for 2011, 2012, 2013 years and only for Asian countries. The considered rows were as below:

- Date
  Date refers to the date of observation where the temperature was measured.
- Avg_Temp
  This is the average land temperature that is recorded at several points in the country.
- Avg_Temp_Uncertain
  This is the 95% confidence interval around the minimum land temperature.
- Country
  It specifies the name of the country where we measured the temperature.
- Year
  It is the year when the temperature was measured.

Methods used for analysis

The following methods were used to perform the analysis on the dataset. Based on these methods few interpretations were drawn.

- Boxplot
- Comparison Boxplot
- Plotting
- Resistant Line
- Straightening
- Smoothening
- Median Polish – Additive Fit
- Extended Fit

Introduction

Meet the Data:

Data: Climate Change: earth's surface temperature data for Asian Countries for years 2011, 2012, 2013

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

One of the major threat of our age is the temperature change. The organization Berkeley Earth recorded the changes of temperatures across all the countries from 1850 – 2013. Here in the analysis part we are considering the Asian Countries for the years 2010 – 2013 to observe the notable changes in the temperature of the earth's surface.

The table below shows the sample data from the data set.

date	Avg_Temp	Avg_Temp_Uncertain	Country	year
8/1/2013	19.971	0.223	China	2013
8/1/2013	26.555	0.242	India	2013
8/1/2013	25.669	0.303	Japan	2013
8/1/2013	26.661	0.223	Malaysia	2013
8/1/2013	28.487	0.375	Pakistan	2013
8/1/2013	27.068	0.41	Philippines	2013
8/1/2013	13.819	0.328	Russia	2013
8/1/2013	34.683	1.272	Saudi Arabia	2013
8/1/2013	27.372	0.417	Singapore	2013
8/1/2013	27.548	0.289	Thailand	2013
8/1/2013	35.617	1.961	United Arab Emirates	2013
7/1/2013	20.482	0.153	China	2013
7/1/2013	27.012	0.197	India	2013
7/1/2013	24.286	0.369	Japan	2013
7/1/2013	26.717	0.223	Malaysia	2013
7/1/2013	30.502	0.263	Pakistan	2013
7/1/2013	27.372	0.441	Philippines	2013

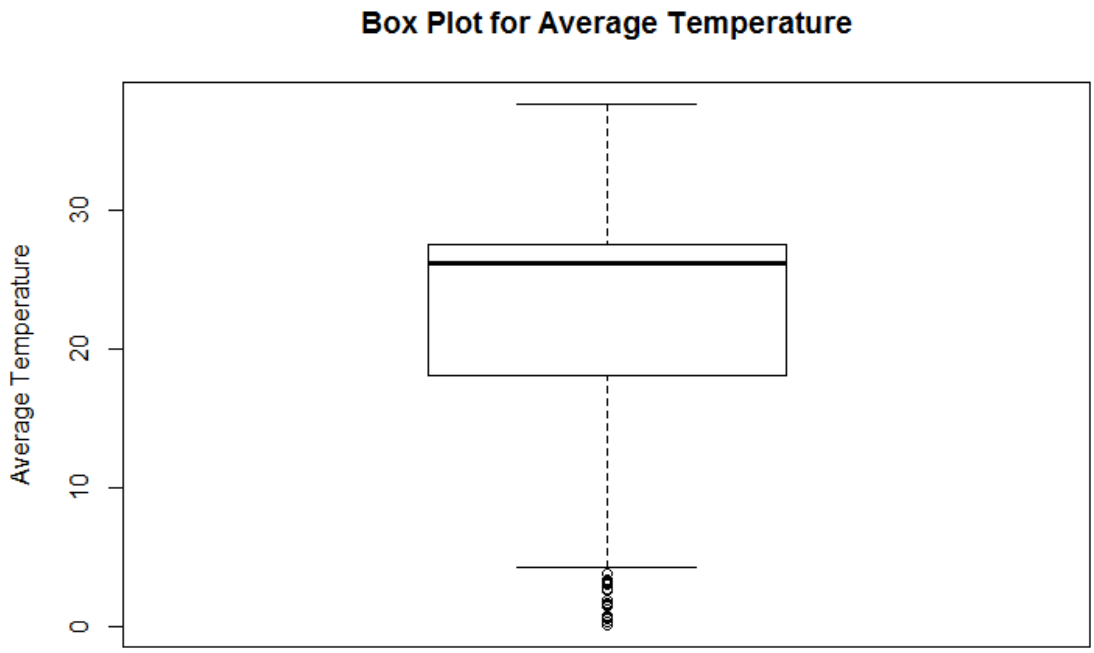
Stem Plot:

Here the dataset contains 352 rows, so considering the stem plot in this case is not possible, as the stem plot would be better to consider a maximum of 50 values.

Box Plot:

Considering the boxplot for the Average Temperature, we got the graph as below.

Looking at the boxplot we can say that the data is left skewed and having few outliers.



Basic interpretations from the boxplot:

Here the data is left skewed. It is having few outliers and those are mild outliers as these outliers are less than 3/2 times of the lower quartile.

Summary Values:

- Min = 0.047
- 1st quartile = 18.150
- Median = 26.240
- Mean = 22.580
- 3rd quartile = 27.570
- Max = 37.710
- Fourth Spread = 9.449
- Step = 14.1735

Here the maximum temperature was recorded at United Arab Emirates in the year of 2012 as 37.710 and the minimum temperature as 0.047 at Japan in 2011.

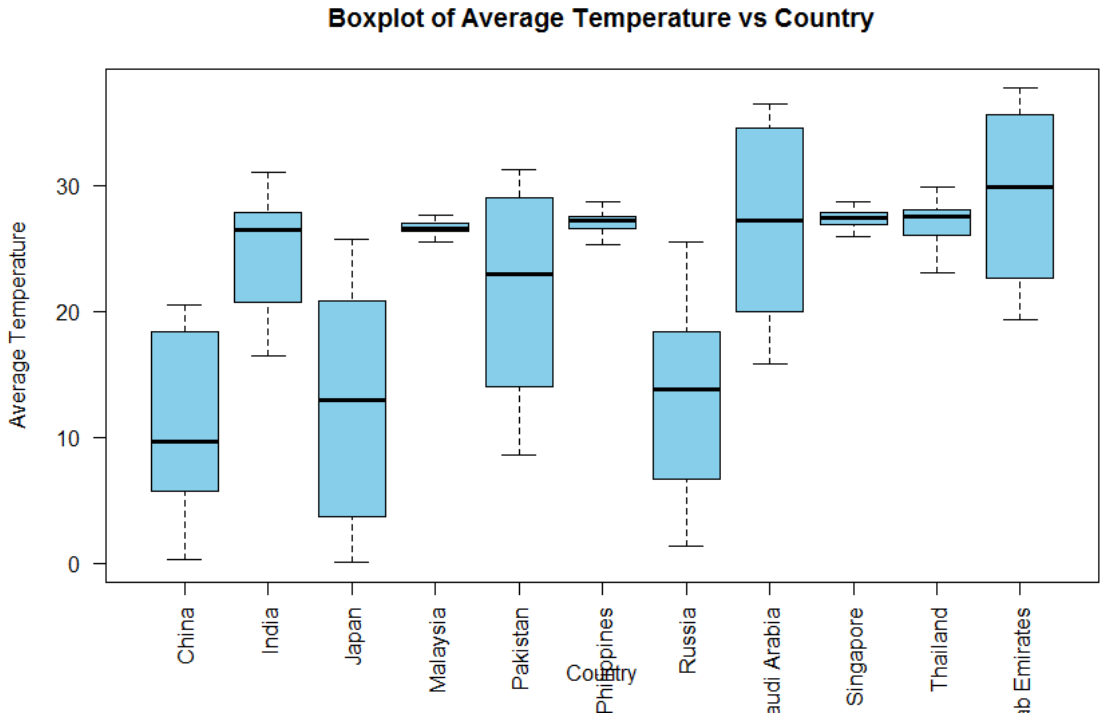
Comparison Box Plot:

Here in this method, we'll compare the batches of the average temperature by country.

Below is the boxplot for the comparison. Here from the below display it is easy to interpret the distribution of the Average Temperature for each Country. Here the sorting order is alphabetical.

The interpretation of the box plot will be easy as it represents the distribution of spreads for various countries,

- China, Japan, Russia are coldest countries in the Asian continent, compared to other countries.
- United Arab Emirates is the hottest country.
- Countries like India, Malaysia, Philippines, Arabia, Singapore, Thailand are having similar temperatures.
- Here China, Japan, Pakistan, Russia, Arabia, and UAE are having similar amount of spreads. This indicates they are having similar fourth spreads.



Country	Observations	Minimum Temperature	Year Observed	Maximum Temperature	Year Observed
China	32	0.254	2102	20.482	2013
India	32	16.478	2011	31.014	2013
Japan	32	0.047	2011	25.703	2012
Malaysia	32	25.531	2011	27.614	2013
Pakistan	32	8.554	2012	31.228	2011
Philippines	32	25.303	2011	28.673	2013
Russia	32	1.366	2011	25.53	2013
Saudi Arabia	32	15.795	2011	36.495	2012
Singapore	32	25.957	2011	28.662	2013
Thailand	32	23.116	2011	29.885	2013
United Arab Emirates	32	19.35	2012	37.713	2012

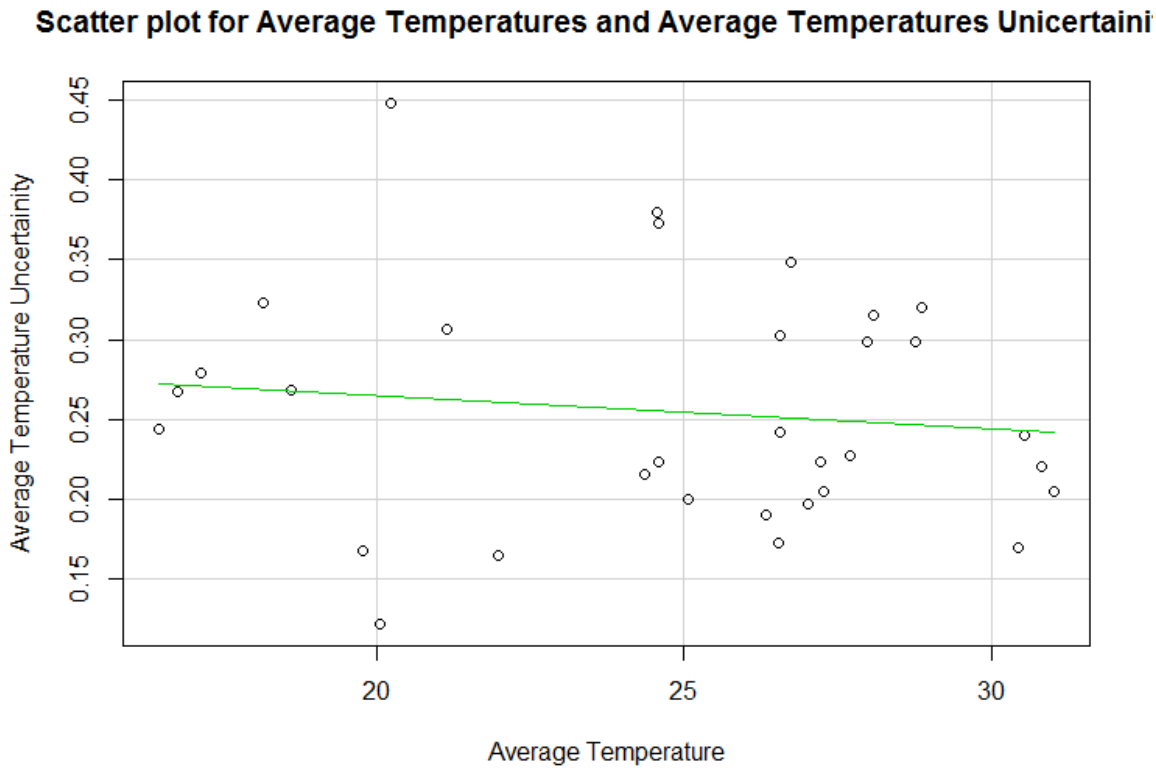
Plotting:

For this method, a subset of the full data set which is having the average temperatures of India has been considered. Here we are finding out the relation between the Average Temperature and Average Temperature Uncertainty. That is based on the average temperature how average temperature uncertainty of India for various years had changed.

The below plot represents the scatterplot between the average temperature and average temperature uncertainty on x and y axis respectively. The slope of the line that fits the data is -0.002048. That is for every 1-unit increase in the average temperature there will be a steady decrease of 0.002048 average temperature uncertainty.

The fit line equation for the line passing through the points is as below,

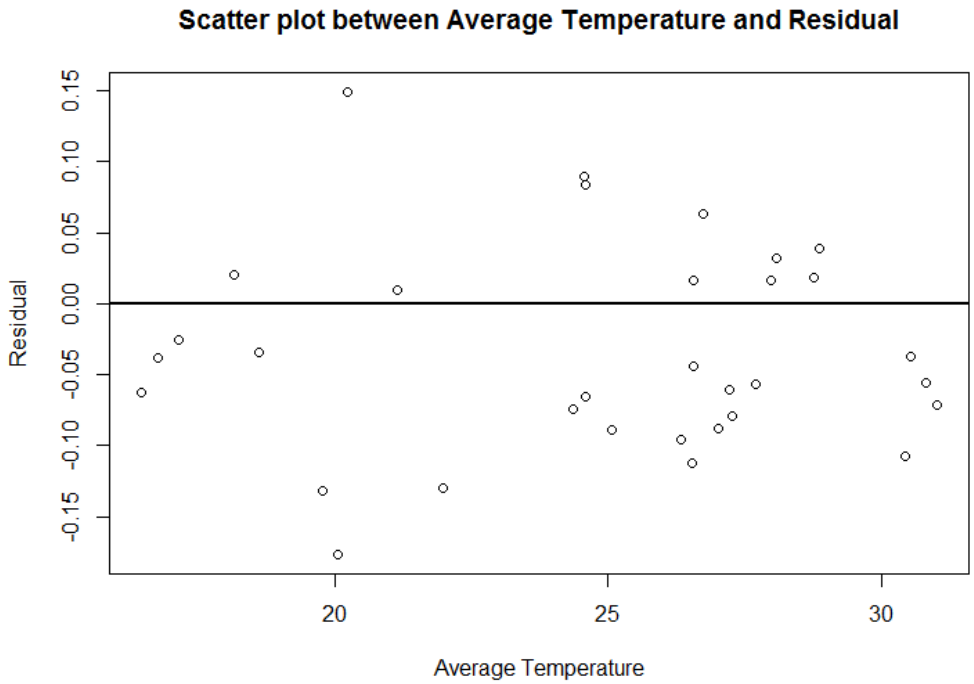
FIT = -0.002048*(Avg_Temp - 16.778) + 0.30566



- Looking at the above graph, the uncertainty temperature increases when average temperature increases and sometimes decreases. And pretty much constant all the time.

- Uncertainty temperatures decreases by 0.0020 over the increase of average temperatures

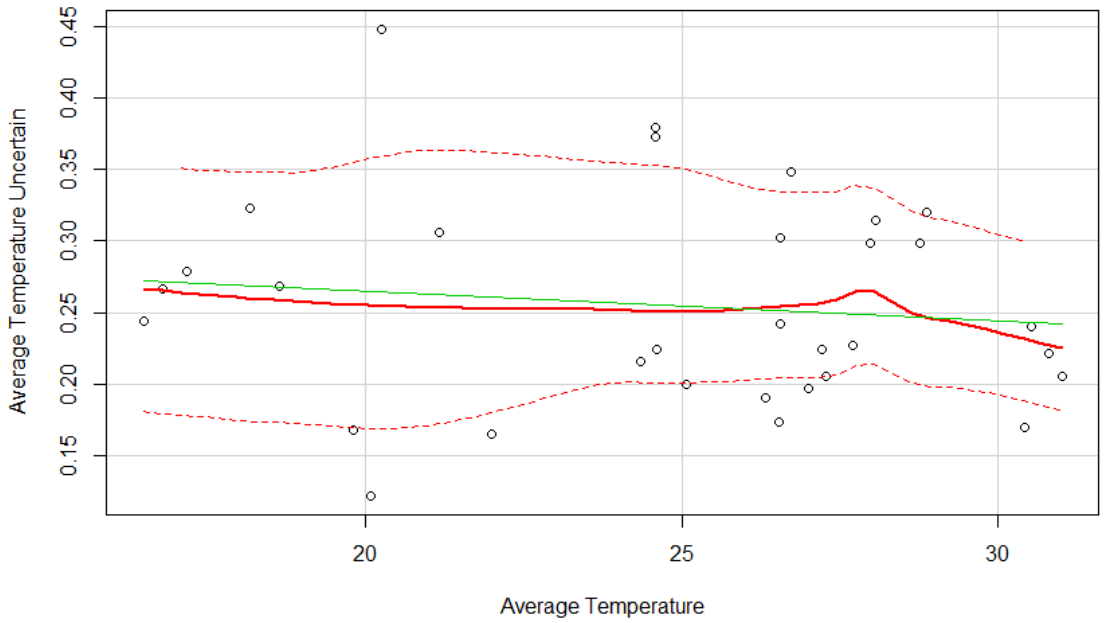
From the above fit line after finding out the residuals, the scatterplot between the IMDB Score and the Residual is



From the above plot we can see there is no general trend in the residuals. Here there is an increase in the average temperature uncertainty after 25.5 and after 29 it got decreased gradually. There is an unusual residual at 20 degrees centigrade.

Alternative Fit:

Here the aim is to find out the fit for the data, i.e., the common pattern in the scatterplot between the average temperature and average temperature uncertainty. The below is the predicted fit for the scatter plot.



- There is a general decrease and increase in the temperatures.
- There is a dip at 20 C which is observed in Feb of 2013.
- For rest uncertainty remained pretty much constant.

Plotting Resistant Line:

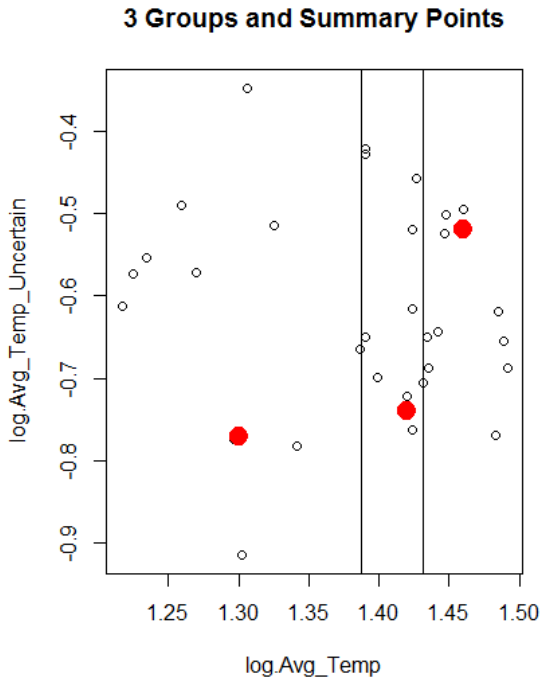
For the above plotted data, we'll transform it in to the log values. Then the data of the average temperature is grouped to in to 3 data subsets and median values of those data sets were noted down. The resultant plot is as below,

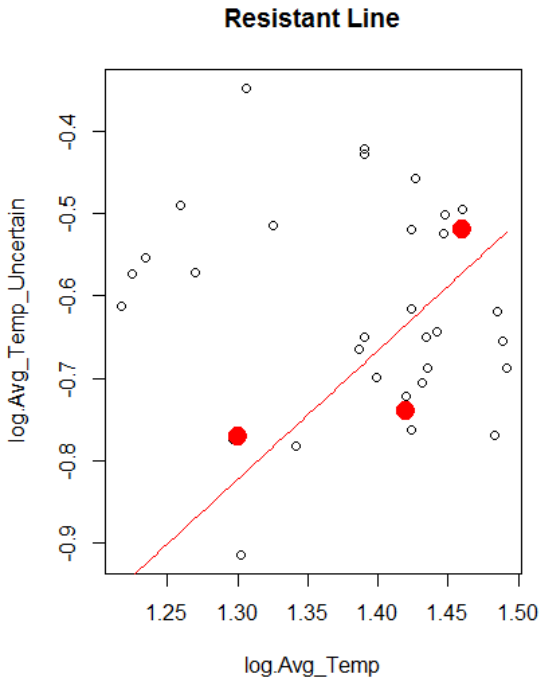
In the later graph, a line which fits the 3 summary points was drawn. The slope of the line that fits the 3 summary points 1.5625

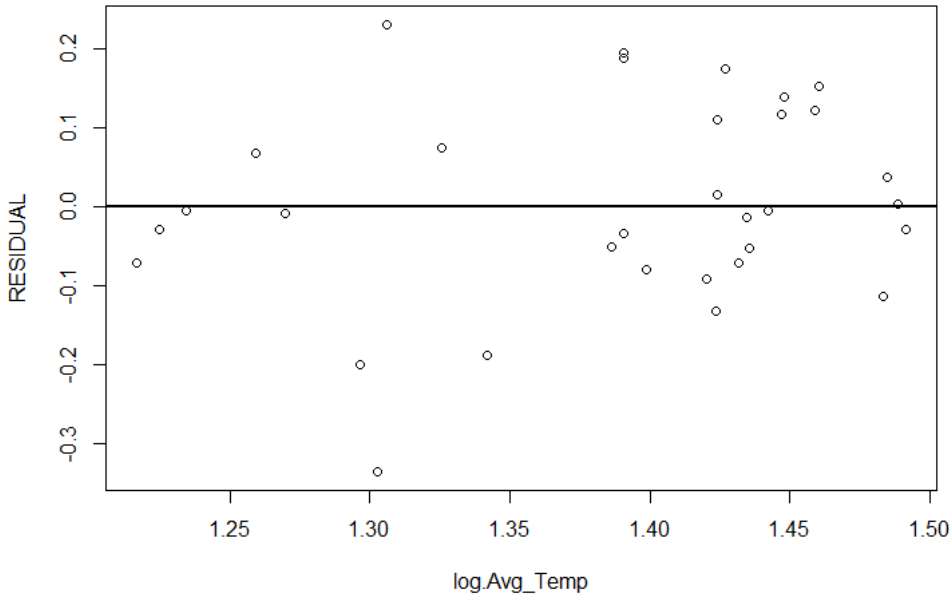
The equation of the line passing through these 3 points is,

1.5625 * (x - 1.42) - 0.635

Later we try to improve the line by fitting them to the residuals, by finding out the residuals from Actual – Fit, we can plot the line between the log of average temperature and residuals. The resulting graph with the zero line is as below,







Here from the above plot, we didn't see any pattern between the average temperature and residuals.

Straightening:

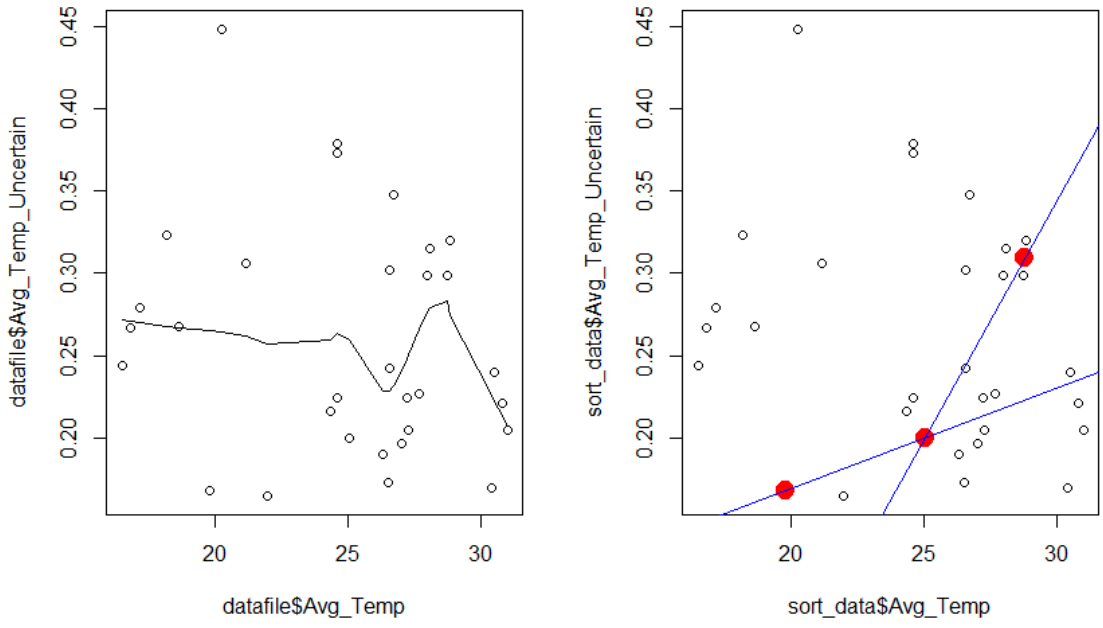
Here the consideration for straightening plots are average temperature and average temperature uncertainty, here plotting the symmetric points after considering the 3 equal groups of data on average temperature and the lines joining 1st group, 2nd group and 2nd group, 3rd group were drawn.

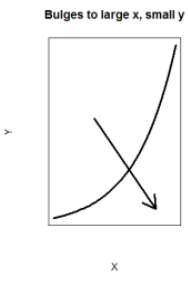
The slopes for these two lines are considered to find out the half slope ratio to make it a bit straight.

Ml = 0.006072106

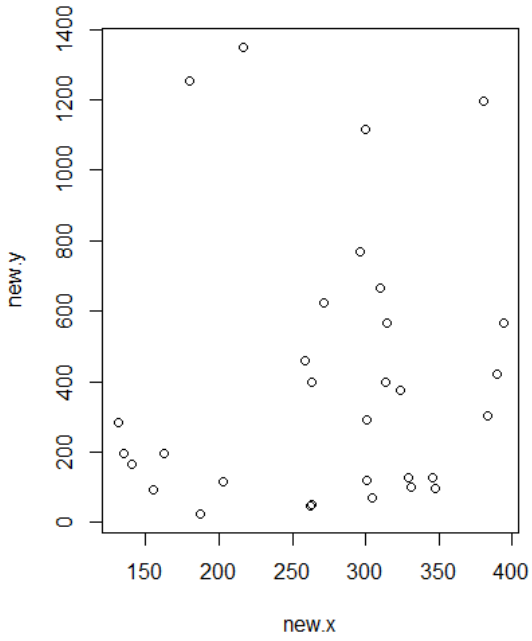
Mr = 0.02901251

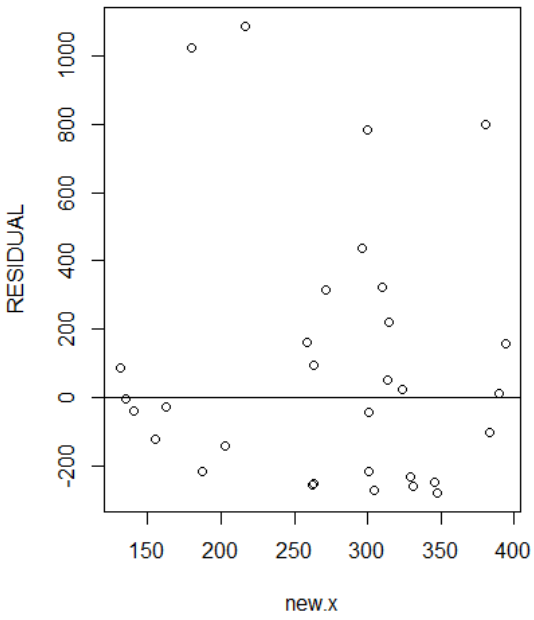
Here the curve of the two lines joining these above 3 points is bulges to large values of x and small values of y





We can achieve the half slope ratio as 1 by the transformation of x to 1.74 and y to -4





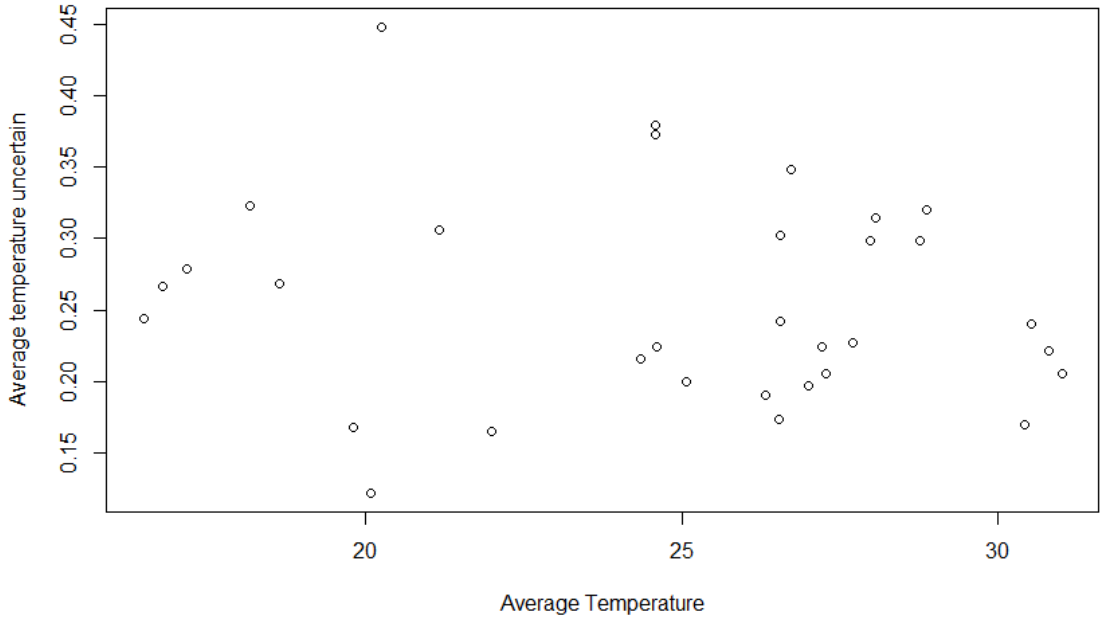
Here from the above plot 1 we can conclude that the data is scattered randomly and it seems with the increase in average temperature there is a notable increase in the average temperature uncertainty and there is no pattern in the residuals vs x plot which is mentioned in plot 2.

From the above residual plot, there are two unusual large residuals in the graph, they belong to 2013 temperatures. In that year there us some kind of high temperatures recorded in India.

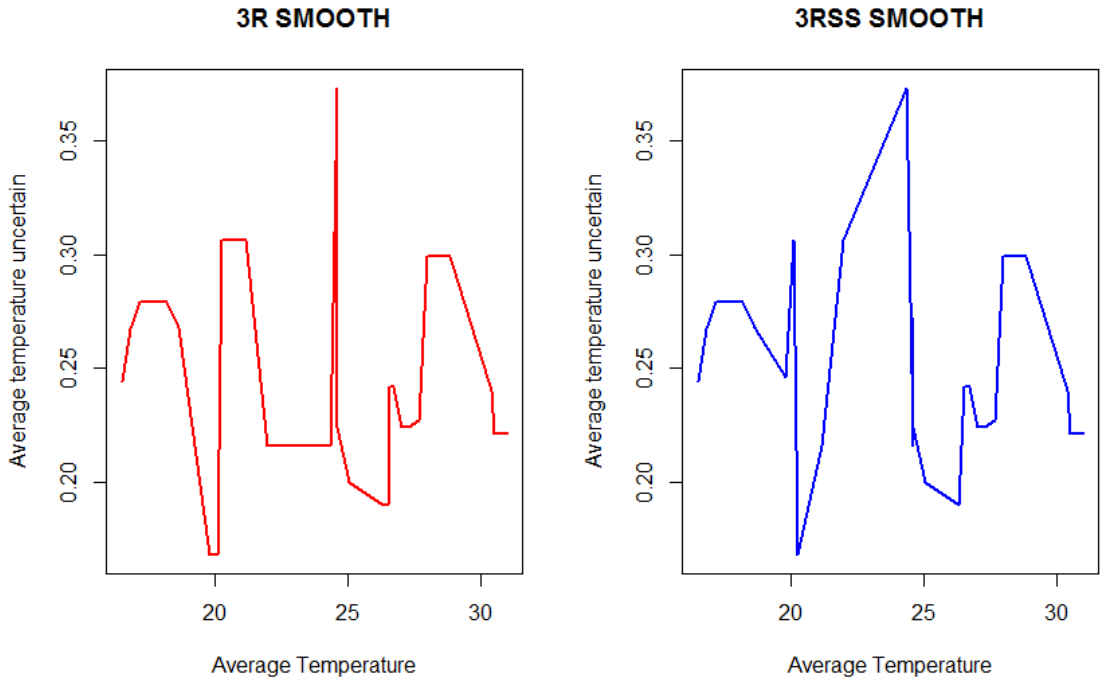
Smoothening:

Here smoothing method is being applied to average temperature and average temperature uncertainty along x and y axis respectively.

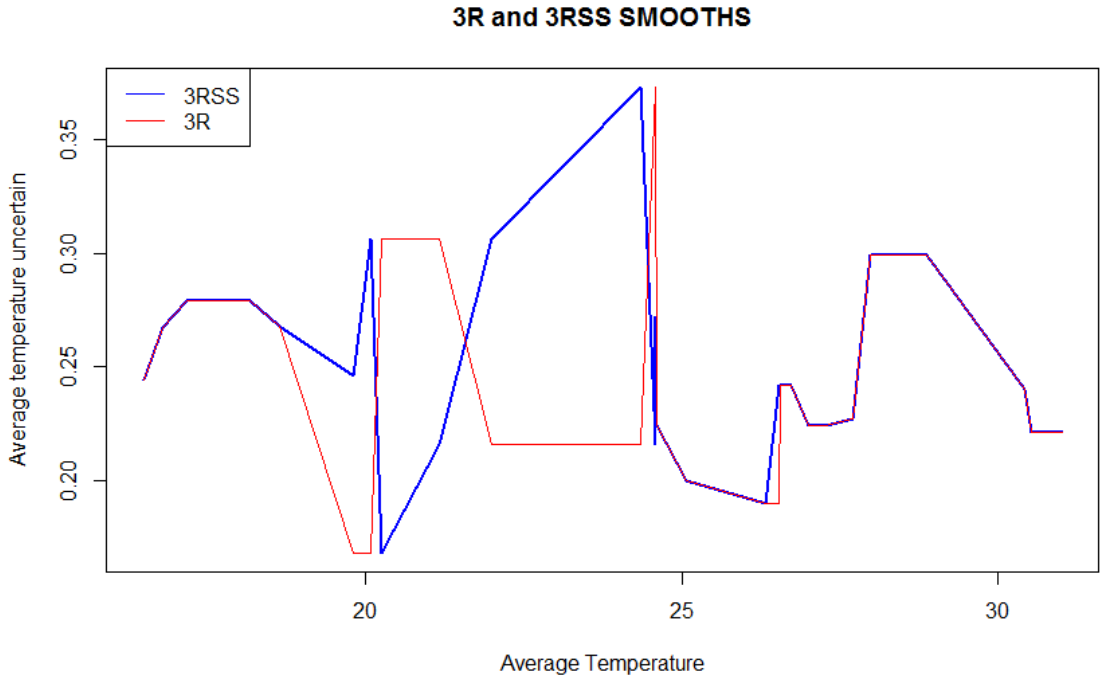
Below is the scatterplot between average temperature and average temperature uncertainty,



From the above plot we took the y axis data and smoothened the data of medians and applied with the in-build smoothing methods of 3R and 3RSS methods.



From the below graph, we can observe the amount of smoothening in the data.



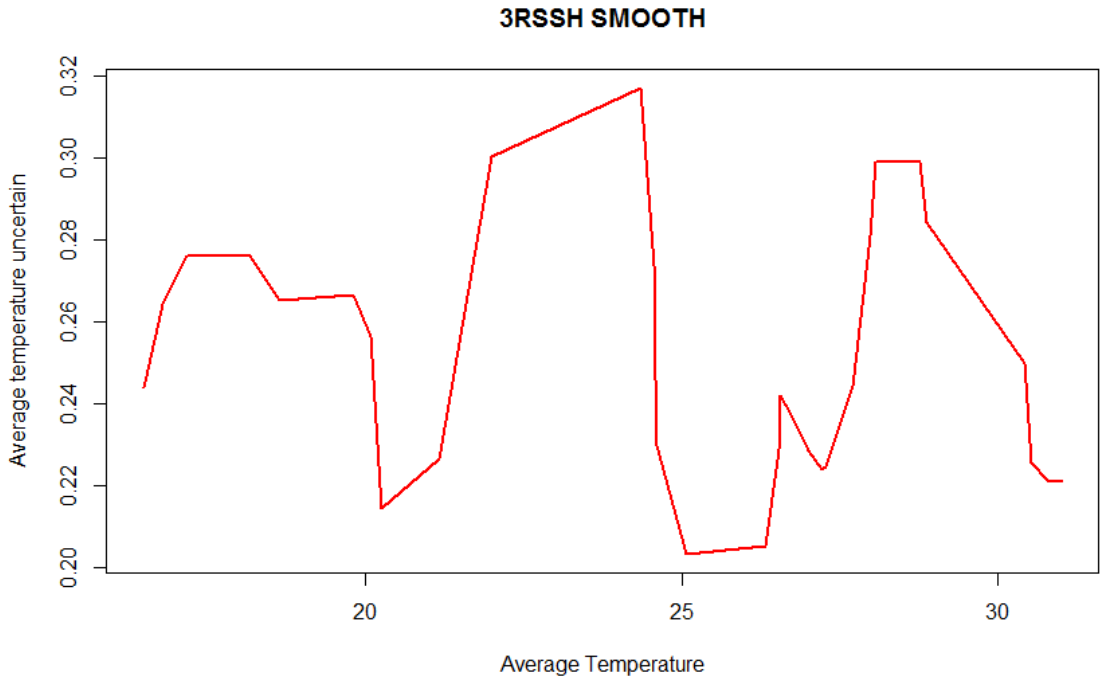
From the above plot we achieved removing few peaks and deep valleys in the data,

For smoothening the data more, we apply hanning method,

The below graph is the resultant graph after hanning the data,

From the hanning method we applied the smoothening of end values of the data by applying 3RSSH and 3RS3R method respectively.

Here the temperatures were gradually increasing and decreasing, there are times where temperatures went to 16.478 and increased over the period of time to 31.014 in the year 2011 and 2013 respectively.



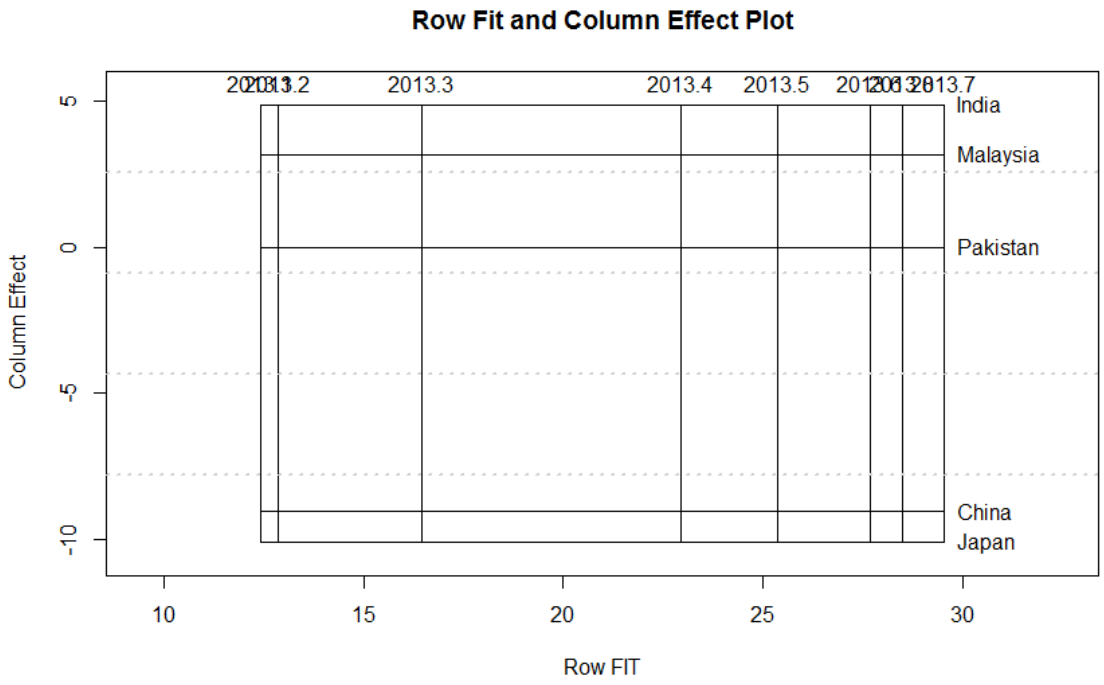
Summary for Smoothening:

- Here 3RSS looks bit rough, for further smoothening Hanning is applied.
- Bumpy monotone sequences are removed from the data.
- Removed the high peaks and low valleys of the data.

Median Polish - Additive Fit

Here temperatures of 2013 for the countries China, India, Japan, Malaysia, Pakistan were considered.

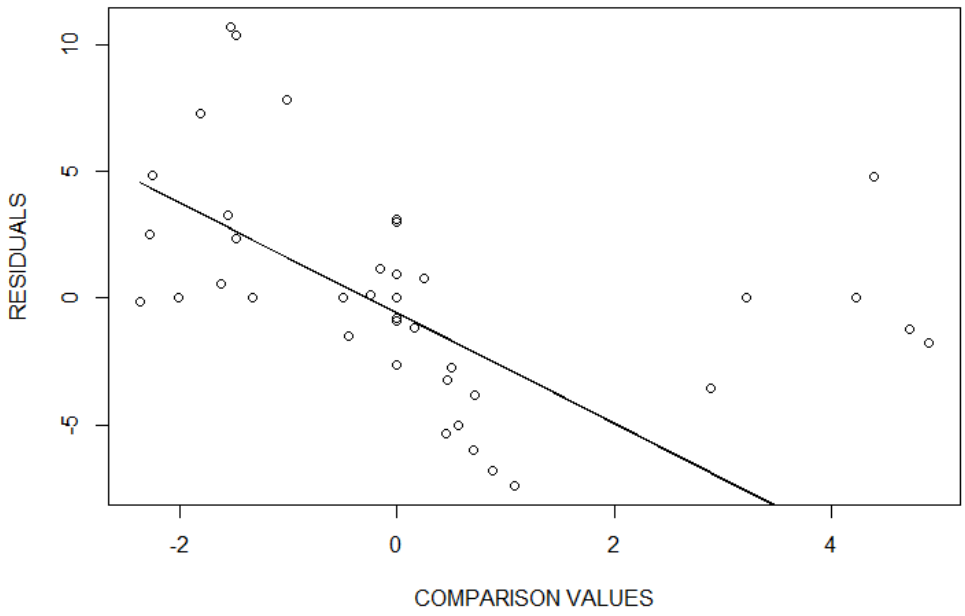
Median polish is applied to the two way data, and the resultant plot is as below,



- From Two-way table, temperature is classified in to country wise and month wise.
- Applying median polish and additive fit
- China and Japan are comparatively colder than India and Malaysia.
- Few residuals are large in the month of January. Mostly China and Malaysia
- Here plotting with Row fits and Column effects.
- According to graph, one should stay in India in August month if we need warm weather.
- And to stay in china or Japan if we need cold weather.

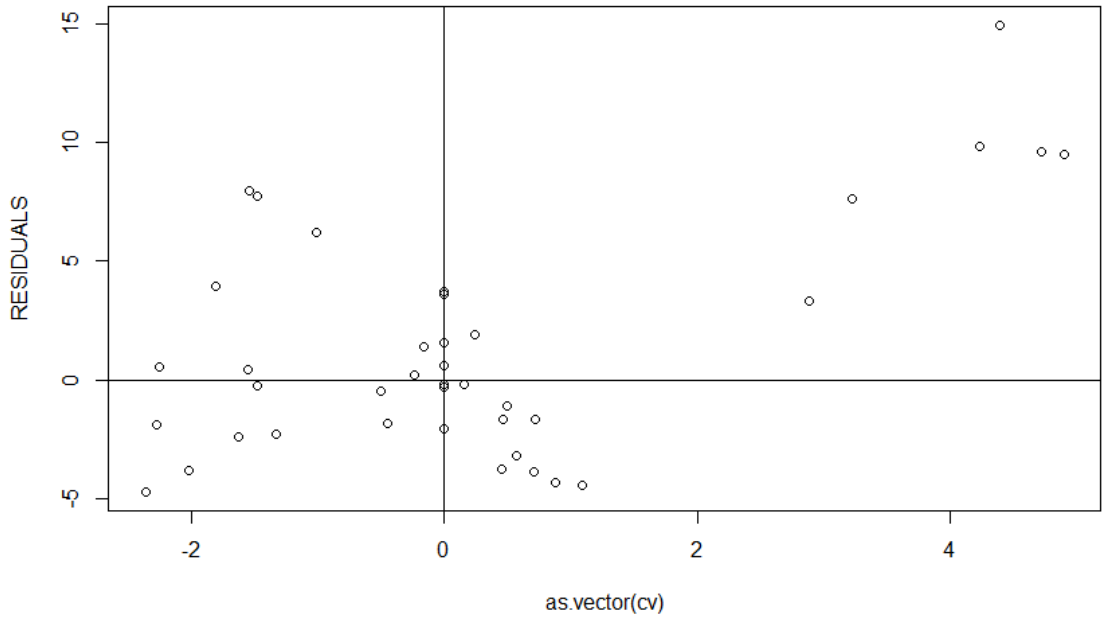
Extended Fit:

For the above model, while using the extended fit, the graph is as below,



The above graph is the plot between the comparison values and the residuals. The median-median line passing through the residuals is having a slope of 2.177

After plotting the residuals and the comparison values and the horizontal and vertical zero lines, the residual plot is as below,

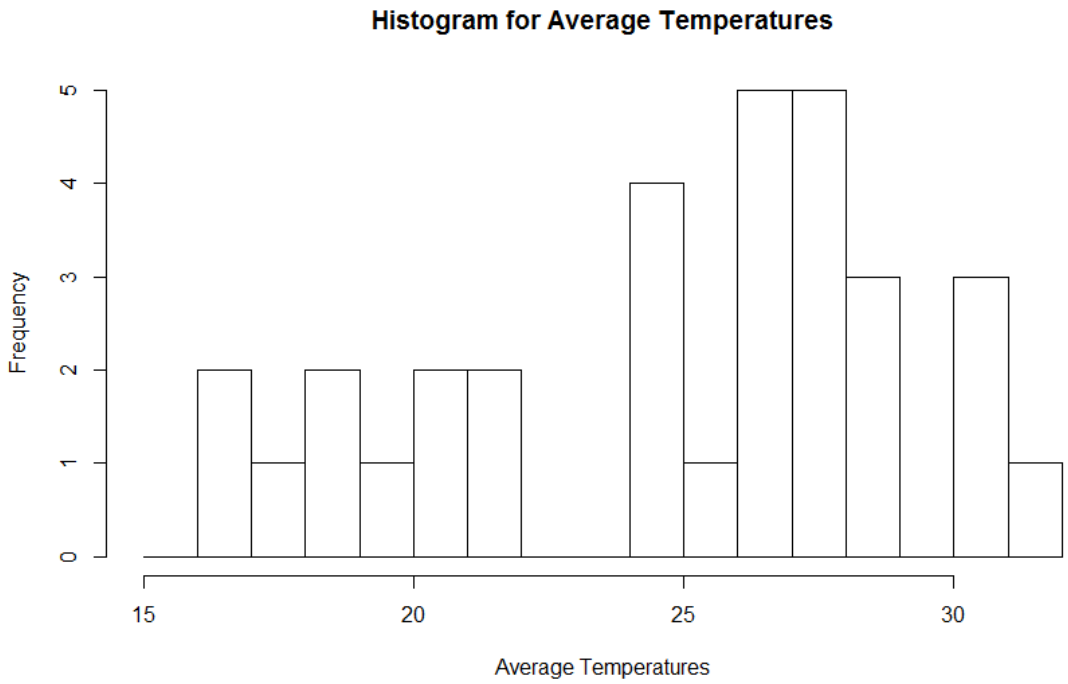


Here the residuals are not following any pattern, and there are few residuals those are unusual., they are for the months of June and July. In which high temperatures were recorded in almost all the countries.

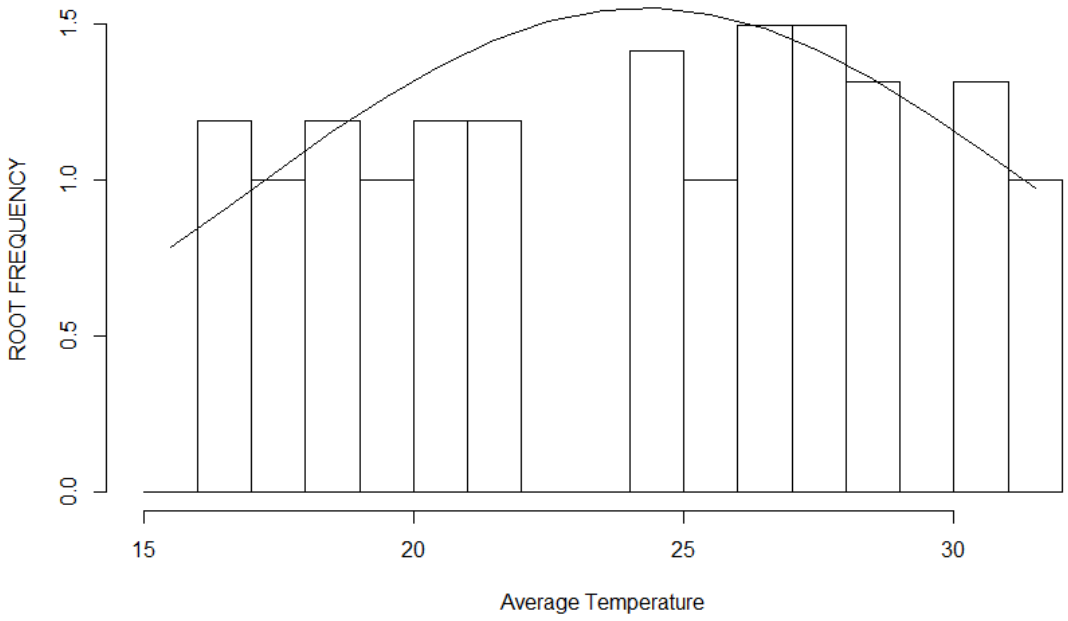
Rootgram:

For the rootgram method the data for country India is being considered.

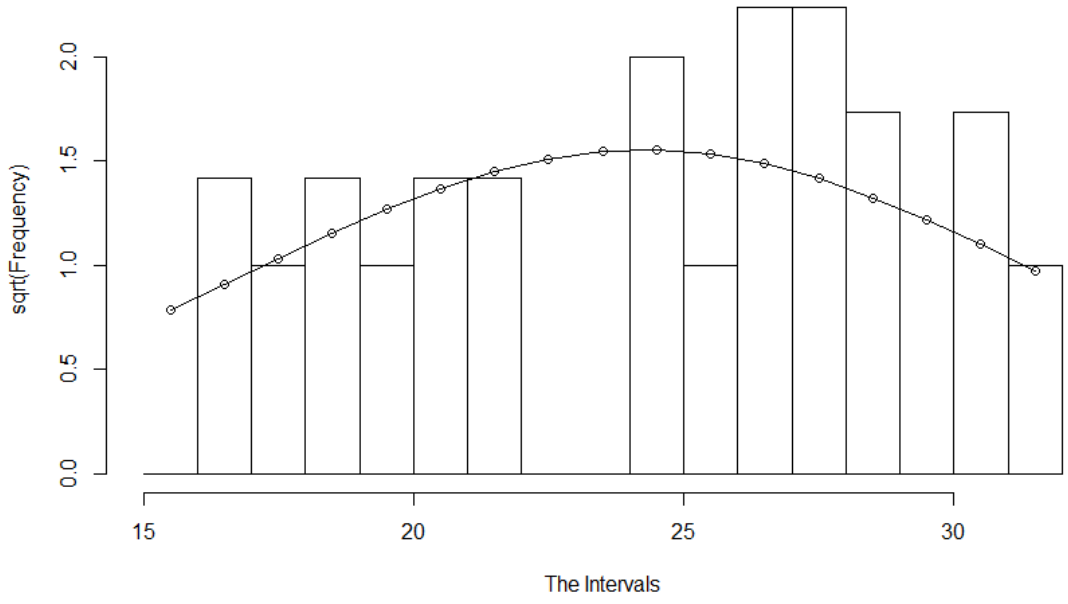
The below plot is the histogram for average temperatures of India.



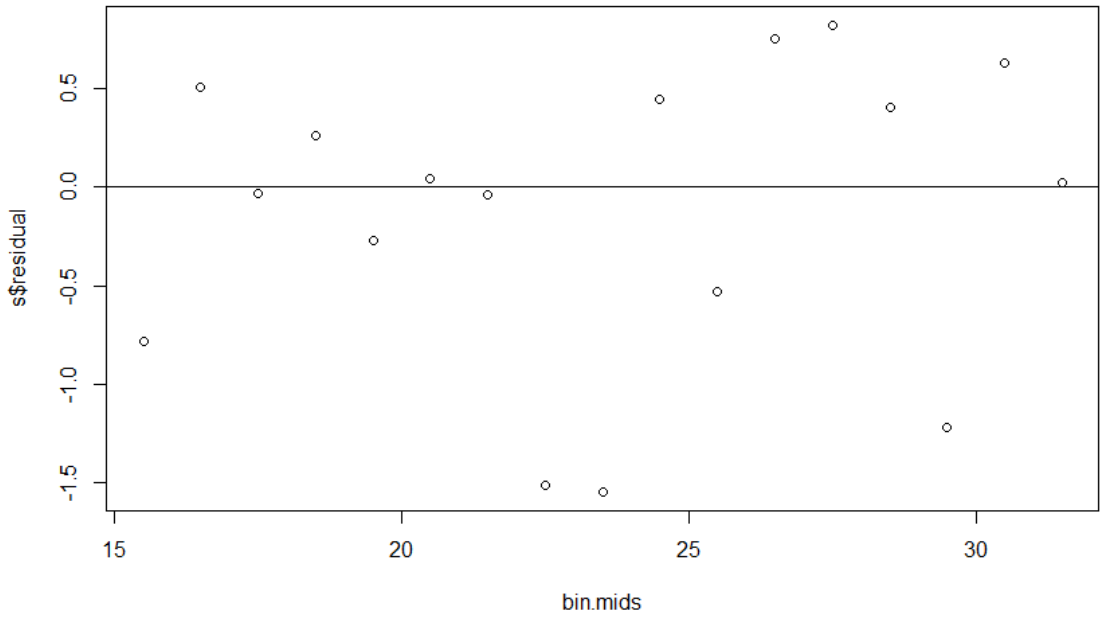
The below graph is for the square root of the data



The rootgram after converting the data in to square roots is as below,



By considering the bins mid values, a plot is drawn between the bins mids and the residuals.



From the above graph, we can infer that there is no pattern in the residuals. There are few unusual residuals between 20-25.