Climate Change: Earth Surface Temperature Data KDD Project

Sunday, April 30, 2017

Team: Analysis Paralysis –
Vyom Sharma , Piyusha Biswas ,
Tong Zhang , Raka Choudhury

Dataset :

Climate Change: Earth Surface Temperature Data from Kaggle
caught our attention the most while we were looking at multiple available datasets.
In the current scenario of the world,Climate Change is a major concern for all countries.
Over the years, the surface temperature of earth has been steadily rising.

- Climate Change indicates change in overall climate pattern of the Earth. It includes change in average temperature of Earth's surface, changes in precipitation levels, rise in sea level , more frequent occurrences of weather-related disasters etc.Climate change occurs due to multiple reasons.As part of our project, we will consider few of these parameters.

We considered the various factors known to directly or indirectly cause climate change.Finally, we chose few parameters to be considered as part of our project.Population of the country , CO_2 emission, Forest area and Sea level were the shortlisted parameters. We found all these related data online (mentioned in References).

We decided to analyze climate change based on above mentioned indicators only for few selected countries of the world. The temperature data can be found in a different dataset and has to be merged with the supporting datasets.

Problem Statement

- Analysis of patterns in climate change based on parameters like :
  - Time
  - Countries
  - CO_2 emission levels
  - Forest Area
  - Population
  - Rise in sea level

Dataset Collection

Based on the problem statement,we'll need data regarding some of the causes of climate change and global warming

- **Global Climate Change data -** This has been taken from Kaggle.It has multiple types of data like average land temperature, country- wise temperature, major city-wise temperature etc.

First few records from Climate Change dataset

	dt	AverageTemperature	AverageTemperatureUncertainty	City	Country	Latitude	Longitude
1	1901-01-01	25.915	0.786	Abidjan	Côte D'Ivoire	5.63N	3.23W
2	1901-02-01	27.913	0.772	Abidjan	Côte D'Ivoire	5.63N	3.23W
3	1901-03-01	27.512	1.178	Abidjan	Côte D'Ivoire	5.63N	3.23W
4	1901-04-01	26.816	1.270	Abidjan	Côte D'Ivoire	5.63N	3.23W
5	1901-05-01	25.837	0.770	Abidjan	Côte D'Ivoire	5.63N	3.23W
6	1901-06-01	25.195	0.536	Abidjan	Côte D'Ivoire	5.63N	3.23W
7	1901-07-01	24.303	0.723	Abidjan	Côte D'Ivoire	5.63N	3.23W
8	1901-08-01	24.152	0.883	Abidjan	Côte D'Ivoire	5.63N	3.23W
9	1901-09-01	25.039	0.678	Abidjan	Côte D'Ivoire	5.63N	3.23W
10	1901-10-01	25.678	1.185	Abidjan	Côte D'Ivoire	5.63N	3.23W
11	1901-11-01	26.087	0.939	Abidjan	Côte D'Ivoire	5.63N	3.23W

- **Greenhouse Gas Emissions data :**

Data includes year wise total greenhouse gas emission values for different countries.It also has readings for individual greenhouse gases like CO_2, N_2O etc.

- **Forest Area dataset :**

Based on our study about the domain, we concluded that climate change and forest area are connected concepts.Depletion of forest area might be a cause of climate change as well.Hence, we plan to use this data along with the available climate change data.

- **Population Data :**

Data includes year wise total population values for different countries.

First few records from Population dataset(for few years)

Country N	Country C	Indicator I	Indicator (	1960	1961	1962	1963	1964	1965	1966	1967
Aruba	ABW	Populatio	SP.POP.TC	54208	55435	56226	56697	57029	57360	57712	58049
Afghanist	AFG	Populatio	SP.POP.TC	8994793	9164945	9343772	9531555	9728645	9935358	10148841	10368600
Angola	AGO	Populatio	SP.POP.TC	5270844	5367287	5465905	5565808	5665701	5765025	5863568	5962831
Albania	ALB	Populatio	SP.POP.TC	1608800	1659800	1711319	1762621	1814135	1864791	1914573	1965598
Andorra	AND	Populatio	SP.POP.TC	13414	14376	15376	16410	17470	18551	19646	20755
Arab Worl	ARB	Populatio	SP.POP.TC	92540534	95077992	97711191	1E+08	1.03E+08	1.06E+08	1.09E+08	1.12E+08
United Ar	ARE	Populatio	SP.POP.TC	92612	100985	112240	125216	138220	150318	161077	171781
Argentina	ARG	Populatio	SP.POP.TC	20619075	20953079	21287682	21621845	21953926	22283389	22608747	22932201
Armenia	ARM	Populatio	SP.POP.TC	1867396	1934239	2002170	2070427	2138133	2204650	2269475	2332624
American	ASM	Populatio	SP.POP.TC	20012	20478	21118	21883	22701	23518	24320	25116
Antigua a	ATG	Populatio	SP.POP.TC	54681	55403	56311	57368	58500	59653	60818	62002
Australia	AUS	Populatio	SP.POP.TC	10276477	10483000	10742000	10950000	11167000	11388000	11651000	11799000
Austria	AUT	Populatio	SP.POP.TC	7047539	7086299	7129864	7175811	7223801	7270889	7322066	7376998

❖ **Sea level rise data :**

Data includes year wise adjusted sea level values(in mm)

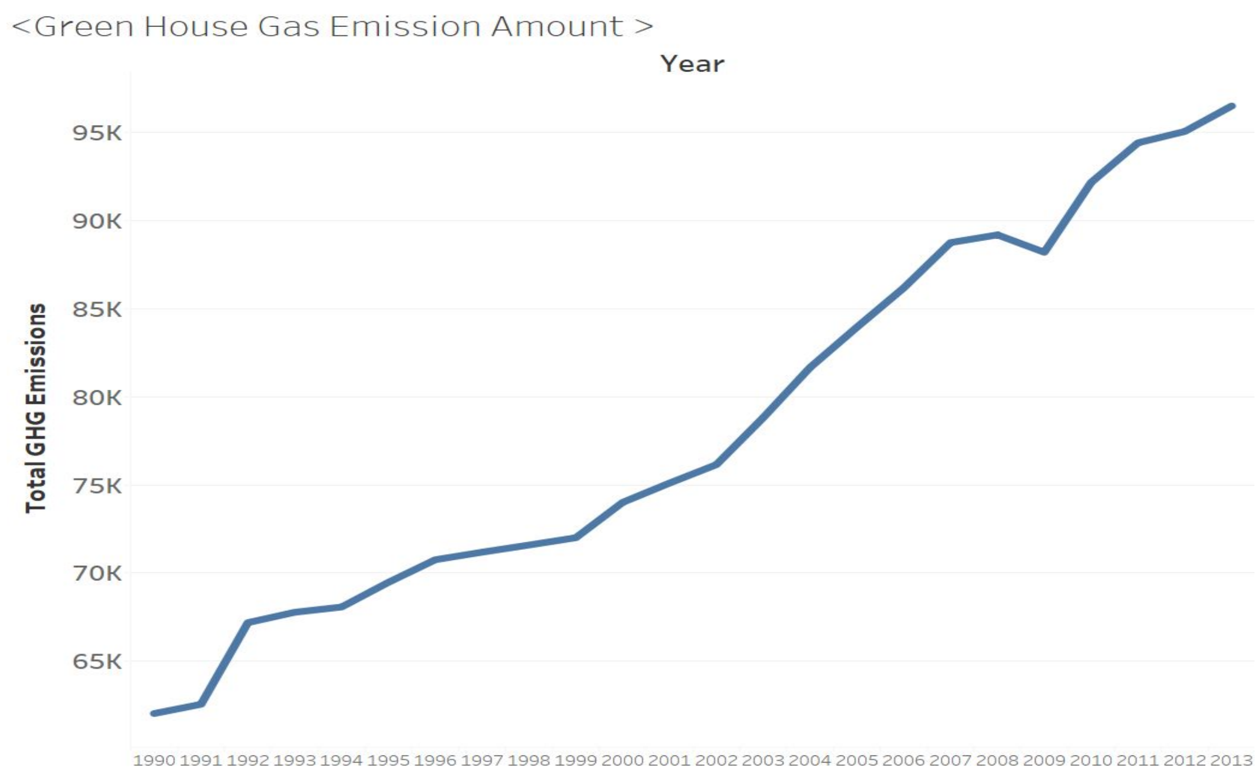
First few records from Rise in sea-level dataset(for few years)

```
> sealevel
   Year CSIRO.Adjusted.Sea.Level Lower.Error.Bound Upper.Error.Bound
1  1880                0.0000000      -0.95275590         0.9527559
2  1881                0.2204724      -0.73228346         1.1732283
3  1882               -0.4409449      -1.34645669         0.4645669
4  1883               -0.2322835      -1.12992126         0.6653543
5  1884                0.5905512      -0.28346457         1.4645669
6  1885                0.5314961      -0.33070866         1.3937008
7  1886                0.4370079      -0.38188976         1.2559055
8  1887                0.2165354      -0.60236220         1.0354331
9  1888                0.2992126      -0.51968504         1.1181102
10 1889                0.3622047      -0.45669291         1.1811024
```

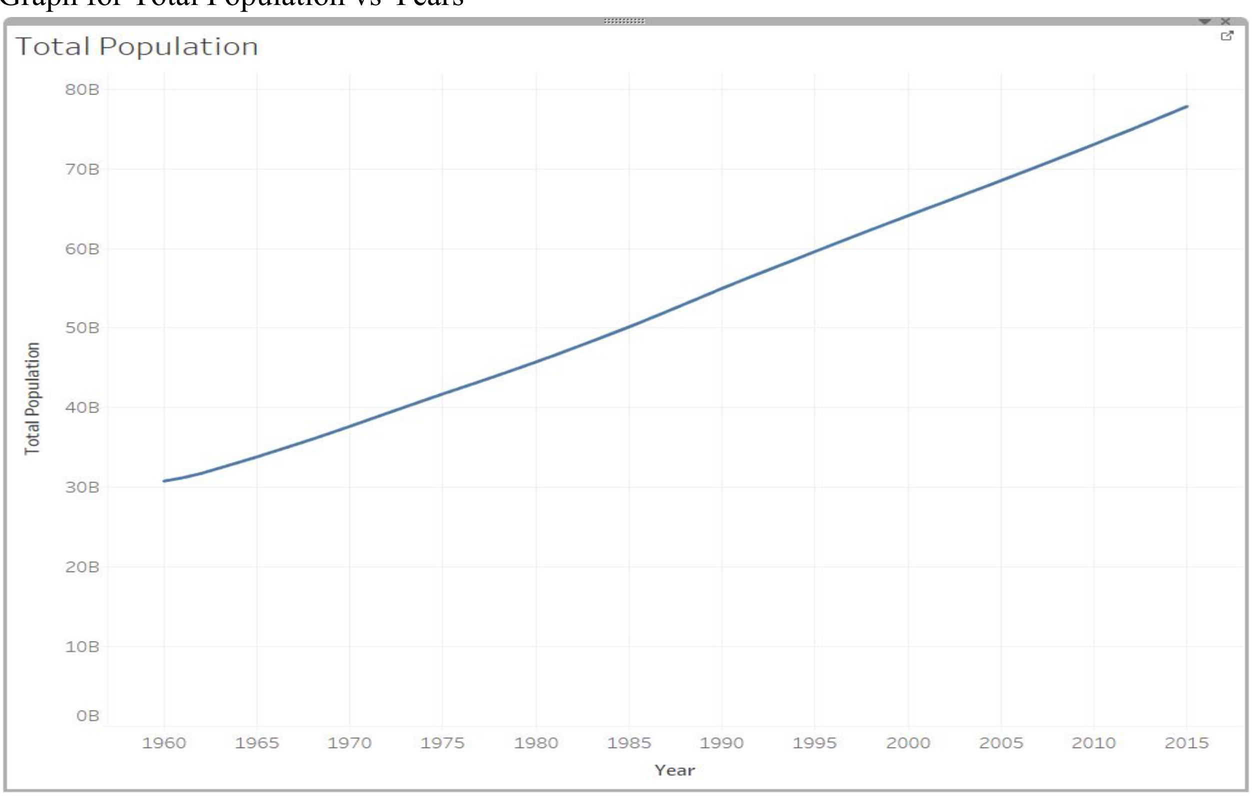
Data Understanding and Data Preparation Phase

Exploratory Data Analysis

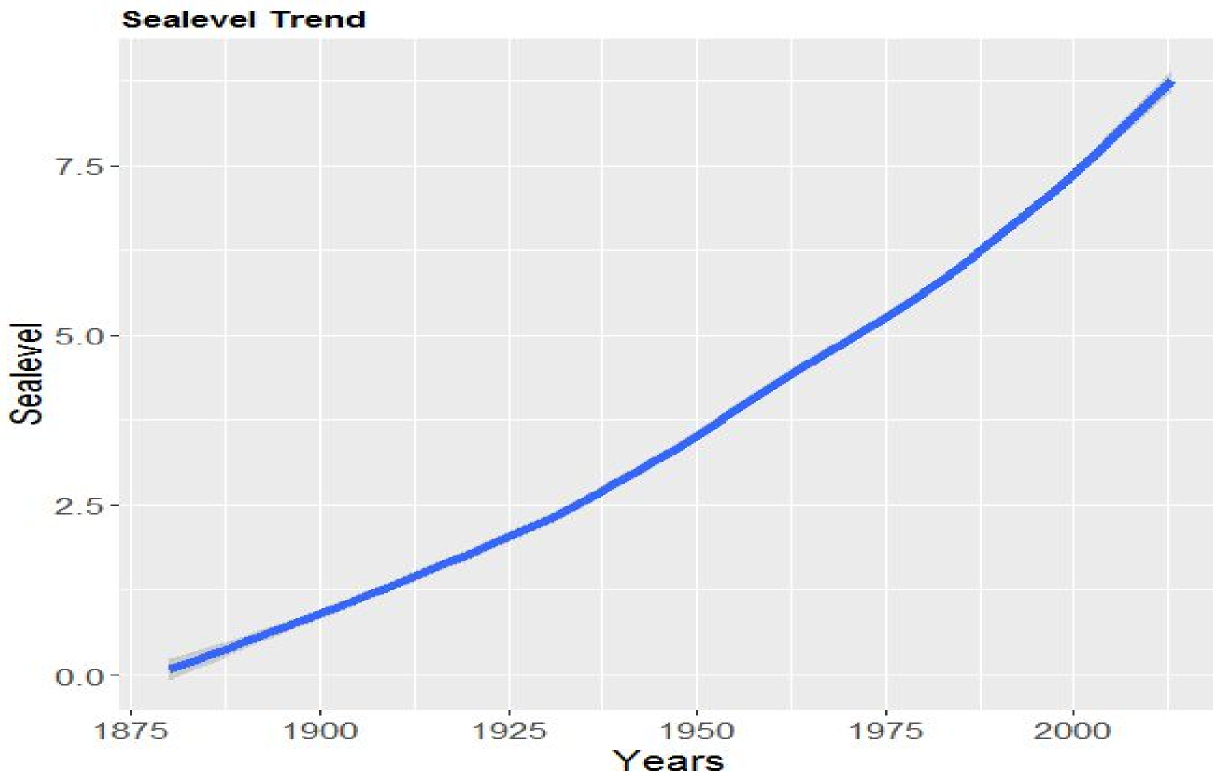
Graph for Total GHG emission vs Years



Graph for Total Population vs Years



Graph for Sea level rise(in mm) vs Years plotted using R

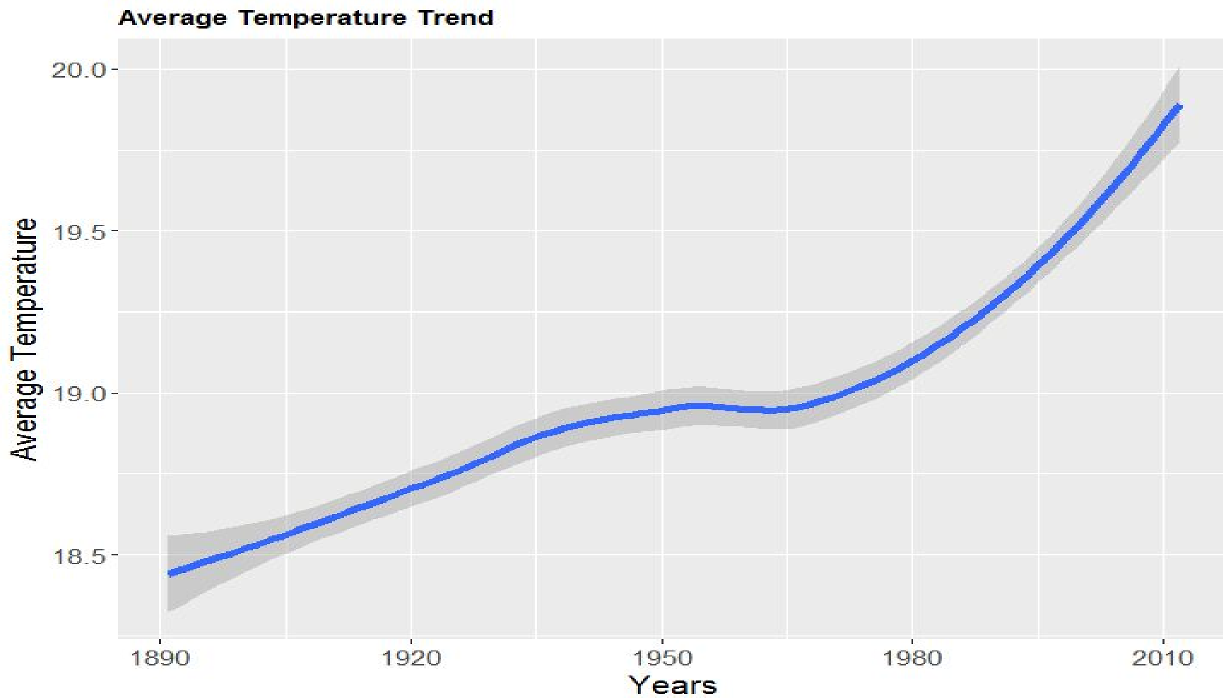


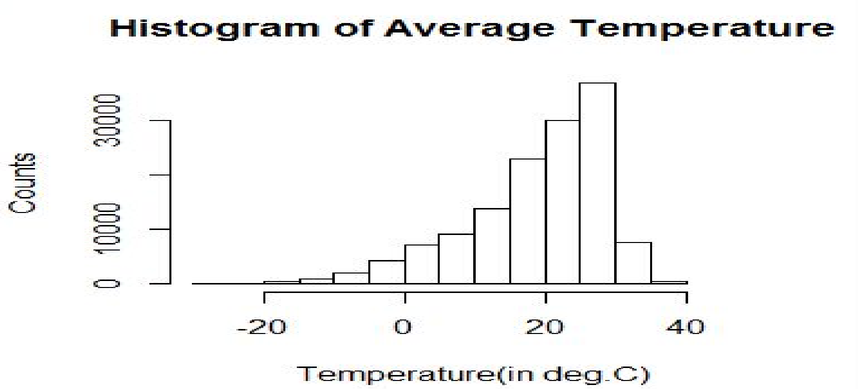
Summary for Rise in sea-level dataset

Year	CSIRO.Adjusted.Sea.Level	Lower.Error.Bound	Upper.Error.Bound	NOAA.Adjusted.Sea.Level
Min. :1880	Min. :-0.4409	Min. :-1.346	Min. :0.4646	Min. :6.297
1st Qu.:1914	1st Qu.: 1.6329	1st Qu.: 1.079	1st Qu.:2.2402	1st Qu.:6.853
Median :1947	Median : 3.3130	Median : 2.915	Median :3.7106	Median :7.498
Mean :1947	Mean : 3.6503	Mean : 3.205	Mean :4.0960	Mean :7.423
3rd Qu.:1980	3rd Qu.: 5.5876	3rd Qu.: 5.330	3rd Qu.:5.8455	3rd Qu.:8.012
Max. :2014	Max. : 9.3268	Max. : 8.992	Max. :9.6614	Max. :8.664
	NA's :1	NA's :1	NA's :1	NA's :113

EDA on Global Climate Change Dataset

- **Distribution: Graph for Average Temperature (in Celsius) vs Year plotted using R**





Correlations

Correlations between Global climate change and parameters like population,forest area etc. were observed.
 E.g : Increase in population might be related to the increase in Average temperature over the years.

```
> cor(total$avg_Temp,total$TotalPopulation)
[1] 0.8568113
```

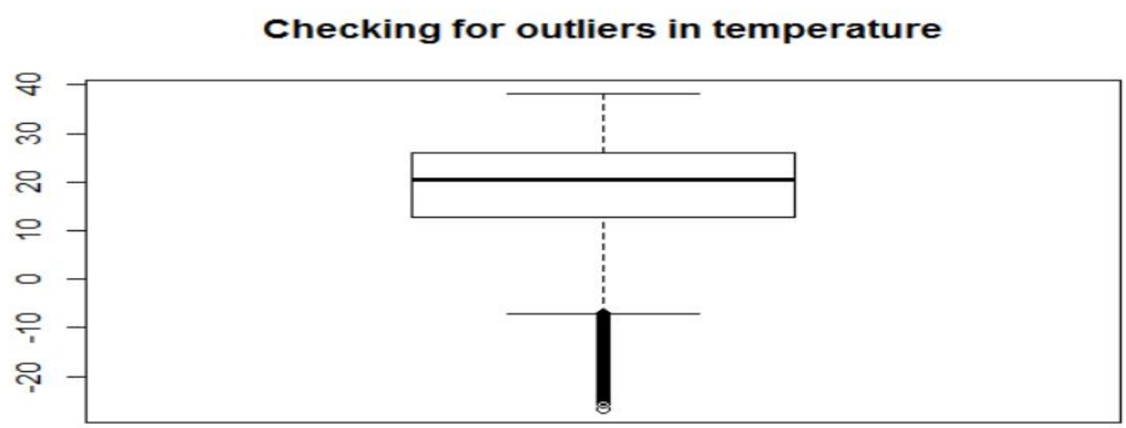
Interpretation of Temperature increase
Based on our study about the domain of global warming, we found that any amount of temperature increase (even 1 deg. C) indicates the occurrence of global warming.

- Increase in CO_2 emissions is a major cause of global warming.
- Increase in population might also impact climate change.
- And Slightest increase in global temperature can lead to rise in sea level.

Hence, we can plan to use the collected extra datasets along with the Kaggle data for climate change.

Outliers :

- Mostly negative temperatures.
- But, there are areas in the world which experience such low temperatures.
- Hence, those values can not be ignored.



Kurtosis

- Kurtosis is a measure of whether the data is heavy-tailed or light-tailed relative to a normal distribution.
- Data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case
- **Here,kurtosis value is 0.7805135.**
- It shows that the distribution of the data is *platykurtic*, since the computed value is less than 3. Also,this indicates absence of outliers.

Skewness

- It is a measure of symmetry of data.

- If skewness is less than -1 or greater than 1, the distribution is highly skewed.
- **For average temperature, skewness value is -1.035394.**
- Hence,we can say that the distribution is highly skewed.
- A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Missing values :

- There are 11002 rows with missing values for average temperature.
- Pattern in missing values : Data before 1890 has maximum missing values.We used mice package to check patterns in missing values.
- Analyzing the dataset and reading information about climate change we reached a conclusion that methods like substituting with constant or mean, imputation won't be appropriate.Hence, removing the missing values will be most suitable.Even after removal,leftover data shows temperature variation over years.

Checking Number of rows with NA
(in Global climate change dataset)

```
           dt            AverageTemperature AverageTemperatureUncertainty
            0                         11002                         11002
         City                       Country                      Latitude
            0                             0                             0
    Longitude
            0
```

Percentage of missing data in each column

```
           dt            AverageTemperature AverageTemperatureUncertainty
     0.000000                      4.599941                      4.599941
         City                       Country                      Latitude
     0.000000                      0.000000                      0.000000
    Longitude
     0.000000
```

Checking pattern in missing data

```
       dt AverageTemperature AverageTemperatureUncertainty Longitude   City Country Latitude
176846  1                  1                             1         1      0       0        0      3
 51329  1                  1                             1         0      0       0        0      4
  8571  1                  0                             0         1      0       0        0      5
  2431  1                  0                             0         0      0       0        0      6
        0              11002                         11002     53760 239177  239177   239177 793295
```

Summary of climate change dataset before handling missing values

```
       dt              AverageTemperature AverageTemperatureUncertainty     City              Country         
 Min.   :1743-11-01   Min.   :-26.77     Min.   : 0.040               Length:239177      Length:239177     
 1st Qu.:1864-02-01   1st Qu.: 12.71     1st Qu.: 0.340               Class :character   Class :character  
 Median :1914-02-01   Median : 20.43     Median : 0.592               Mode  :character   Mode  :character  
 Mean   :1910-11-08   Mean   : 18.13     Mean   : 0.969                                                    
 3rd Qu.:1963-12-01   3rd Qu.: 25.92     3rd Qu.: 1.320                                                    
 Max.   :2013-09-01   Max.   : 38.28     Max.   :14.037                                                    
                      NA's   :11002      NA's   :11002                                                     
   Latitude          Longitude        
 Length:239177      Length:239177     
 Class :character   Class :character  
 Mode  :character   Mode  :character  
```

Summary of climate change dataset after handling missing values

```
       dt              AverageTemperature AverageTemperatureUncertainty     City              Country         
 Min.   :1901-01-01   Min.   :-26.77     Min.   :0.040                Length:135200      Length:135200     
 1st Qu.:1929-02-22   1st Qu.: 14.12     1st Qu.:0.272                Class :character   Class :character  
 Median :1957-04-16   Median : 21.35     Median :0.379                Mode  :character   Mode  :character  
 Mean   :1957-04-16   Mean   : 19.01     Mean   :0.457                                                     
 3rd Qu.:1985-06-08   3rd Qu.: 26.37     3rd Qu.:0.550                                                     
 Max.   :2013-08-01   Max.   : 38.28     Max.   :4.756                                                     
   Latitude          Longitude        
 Length:135200      Length:135200     
 Class :character   Class :character  
 Mode  :character   Mode  :character  
```

The initial collected data for datasets like forest area,CO2 emission,population etc. contain a lot of missing values for every attribute and for every country. Regression Modeling was used to handle these missing values.To reduce the influence of conditions in one country on another country, missing values for each country were estimated separately.Finally, all the parts have been merged to get the final dataset for analysis.This has been done to take care of different situation in different countries.For instance, populations of China and India are much higher compared to many other countries. Likewise, areas of Malaysia and Pakistan are relatively smaller when compared to countries like USA and China. This might in turn affect the forest area in these countries.

```
> View(data_train)
> usa.initial <- read.csv("usa.csv",header = TRUE,sep = ",")
> names(usa.initial)<-c("year","forestarea","co2emission","poptot")
> usa <- usa.initial
> str(usa)
'data.frame':   56 obs. of  4 variables:
 $ year       : int  1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 ...
 $ forestarea : num  NA NA NA NA NA NA NA NA NA NA ...
 $ co2emission: num  16 15.7 16 16.5 17 17.5 18.1 18.6 19.1 19.9 ...
 $ poptot     : num  1.81e+08 1.84e+08 1.87e+08 1.89e+08 1.92e+08 1.94e+08 1.97e+08 1.99e+08 2.01e+08 2.03e+08 ...
> names(usa)
[1] "year"        "forestarea"  "co2emission" "poptot"
> summary(usa)
      year        forestarea     co2emission        poptot         
 Min.   :1960   Min.   :33.00   Min.   :15.70   Min.   :181000000  
 1st Qu.:1974   1st Qu.:33.10   1st Qu.:18.60   1st Qu.:213500000  
 Median :1988   Median :33.20   Median :19.40   Median :243000000  
 Mean   :1988   Mean   :33.34   Mean   :19.31   Mean   :248714286  
 3rd Qu.:2001   3rd Qu.:33.67   3rd Qu.:20.02   3rd Qu.:285750000  
 Max.   :2015   Max.   :33.90   Max.   :22.50   Max.   :321000000  
                NA's   :30      NA's   :4                          
> 
```

Similar estimation is done for multiple countries like South Africa, India, China, Brazil, Canada etc.Ultimately, these country-wise data are merged together.

```
> usa.mreg.out2 <- lm(usa$forestarea ~usa$co2emission+usa$poptot)
> summary(usa.mreg.out2)

Call:
lm(formula = usa$forestarea ~ usa$co2emission + usa$poptot)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.068918 -0.030081  0.003016  0.022192  0.093839 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      3.406e+01  3.486e-01   97.71  < 2e-16 ***
usa$co2emission -1.536e-01  1.251e-02  -12.28 1.76e-10 ***
usa$poptot       7.517e-09  5.532e-10   13.59 3.09e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04103 on 19 degrees of freedom
  (34 observations deleted due to missingness)
Multiple R-squared:  0.9741,	Adjusted R-squared:  0.9714 
F-statistic: 357.2 on 2 and 19 DF,  p-value: 8.455e-16
```

```
#===============================================
#combine all countries data in to one data frame
#===============================================
climateData <- rbind(usaTraining,canTraining,braTraining)
climateData <- rbind(climateData,chinaTraining,indiaTraining)
climateData <- rbind(climateData,egypttraining,southafricatraining)

#=======================================================
#Normalize all the variables by using min-max normalization
#=======================================================

climateData.norm<-climateData

mmnorm.forestarea <- (climateData$forestarea - min(climateData$forestarea))/(max(climateData$forestarea) - min(climateData$forestarea))
climateData.norm$forestarea <- mmnorm.forestarea


mmnorm.co2emission <-(climateData$co2emission - min(climateData$co2emission))/(max(climateData$co2emission) - min(climateData$co2emission))
climateData.norm$co2emission <- mmnorm.co2emission

mmnorm.poptot <-(climateData$poptot - min(climateData$poptot))/(max(climateData$poptot) - min(climateData$poptot))
climateData.norm$poptot <- mmnorm.poptot



#====================================================================
```

Transformation to get global total population per year

Year	TotalPopulation
1960	30786322073
1961	31198688844
1962	31749131664
1963	32432215473
1964	33122167297
1965	33828078352
1966	34571322687
1967	35312745454
1968	36067742980
1969	36860134487
1970	37663599353
1971	38488325433
1972	39307925119
1973	40121815934
1974	40937765743
1975	41738696090
1976	42525873683
1977	43309655854
1978	44107431889
1979	44922822213
1980	45747291012

Data Preparation: Normalization:

Due to different scale of measurements for each variable data had to be normalized.
E.g. Min-Max normalization

<u>Before Normalization</u>

Filter

	year	forestarea	co2emission	poptot	temperature	country
1	1960	32.965506	16.000000	1.81e+08	6.720098	USA
2	1961	33.034128	15.700000	1.84e+08	6.492202	USA
3	1962	33.010608	16.000000	1.87e+08	6.910322	USA
4	1963	32.948857	16.500000	1.89e+08	7.016024	USA
5	1964	32.894622	17.000000	1.92e+08	6.274974	USA
6	1965	32.832871	17.500000	1.94e+08	6.465711	USA
7	1966	32.763280	18.100000	1.97e+08	6.281553	USA
8	1967	32.701528	18.600000	1.99e+08	6.866044	USA
9	1968	32.639777	19.100000	2.01e+08	6.480260	USA
10	1969	32.531955	19.900000	2.03e+08	6.763688	USA
11	1970	32.362704	21.100000	2.05e+08	6.599193	USA
12	1971	32.400612	21.000000	2.08e+08	6.234191	USA
13	1972	32.308147	21.700000	2.10e+08	6.243403	USA
14	1973	32.200324	22.500000	2.12e+08	6.839272	USA
15	1974	32.368928	21.500000	2.14e+08	6.600926	USA
16	1975	32.552889	20.400000	2.16e+08	6.209032	USA
17	1976	32.445067	21.200000	2.18e+08	6.497176	USA
18	1977	32.414030	21.500000	2.20e+08	7.208443	USA
19	1978	32.359796	22.000000	2.23e+08	6.895796	USA
20	1979	32.405543	21.800000	2.25e+08	6.694786	USA
21	1980	32.574147	20.800000	2.27e+08	7.066892	USA
22	1981	32.742752	19.800000	2.29e+08	7.767236	USA
23	1982	32.949587	18.600000	2.32e+08	6.465607	USA

<u>After Using Min - Max Normalization</u>

	year	forestarea	co2emission	poptot	temperature	country
1	1960	0.43850882	0.7076286434	0.126895212	6.720098	USA
2	1961	0.43943073	0.6941345808	0.129135858	6.492202	USA
3	1962	0.43911474	0.7076286434	0.131376503	6.910322	USA
4	1963	0.43828514	0.7301187478	0.132870267	7.016024	USA
5	1964	0.43755652	0.7526088521	0.135110912	6.274974	USA
6	1965	0.43672692	0.7750989565	0.136604675	6.465711	USA
7	1966	0.43579199	0.8020870817	0.138845321	6.281553	USA
8	1967	0.43496239	0.8245771860	0.140339084	6.866044	USA
9	1968	0.43413279	0.8470672904	0.141832848	6.480260	USA
10	1969	0.43268424	0.8830514574	0.143326611	6.763688	USA
11	1970	0.43041043	0.9370277078	0.144820375	6.599193	USA
12	1971	0.43091971	0.9325296869	0.147061020	6.234191	USA
13	1972	0.42967747	0.9640158330	0.148554784	6.243403	USA
14	1973	0.42822893	1.0000000000	0.150048547	6.839272	USA
15	1974	0.43049405	0.9550197913	0.151542311	6.600926	USA
16	1975	0.43296549	0.9055415617	0.153036074	6.209032	USA
17	1976	0.43151694	0.9415257287	0.154529838	6.497176	USA
18	1977	0.43109997	0.9550197913	0.156023601	7.208443	USA
19	1978	0.43037135	0.9775098956	0.158264247	6.895796	USA
20	1979	0.43098596	0.9685138539	0.159758010	6.694786	USA
21	1980	0.43325108	0.9235336452	0.161251774	7.066892	USA
22	1981	0.43551620	0.8785534365	0.162745537	7.767236	USA
23	1982	0.43829495	0.8245771860	0.164986183	6.465607	USA
24	1983	0.43849692	0.8245771860	0.166479946	6.950866	USA

Modeling

ARIMA time series model (Autoregressive integrated moving average)

Time series model are used when the data has a dependency on time. ARIMA models are, in theory, the most general class of models for forecasting a time series which can be made to be “stationary”. The time series can be represent as ARIMA(p,d,q)(P,D,Q)m.

Data Preparation Changes

- Data had to be transformed into time series format to be able to apply ARIMA time series model.
- For applying the time series model ,we had to also use the major city wise climate change data.
- Total Population dataset , CO2 emission dataset and Forest area dataset had to be integrated into our previously integrated dataset.

ts() function - Convert a numeric vector into an R time series object.

The format is **ts(***vector*, **start=, end=, frequency=)** where :

- start and end are the times of the first and last observation
- frequency is the number of observations per unit time (1=annual, 4=quarterly, 12=monthly, etc.).

<u>Transform the data into time series format</u>

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1990	12.96821	14.57313	17.36101	19.98670	22.40623	23.91269	24.38834	24.32777	22.53726	20.13708	17.54136	13.97481
1991	12.68308	14.09189	17.01887	19.93335	22.43805	24.07881	24.54578	24.20086	22.65751	19.80891	16.48127	13.31980
1992	12.50978	13.79592	16.82521	20.01001	22.48707	23.79935	24.22999	24.00869	22.30302	19.39159	16.21639	13.51563
1993	12.46158	14.10242	16.67400	19.88770	22.66416	23.90910	24.20531	24.06681	22.33786	19.82967	16.22784	13.77235
1994	13.03094	13.73309	16.96686	20.39195	22.89067	24.08682	24.70636	24.33059	22.68588	20.00924	17.17630	13.66613
1995	12.93887	14.62651	17.03312	19.70149	22.71972	24.43867	24.64747	24.54573	22.61229	20.35351	16.43924	13.22204
1996	12.58024	13.82294	16.62634	19.55174	22.71805	23.96375	24.31320	24.06357	22.51903	19.74019	16.37508	13.78803
1997	12.44596	14.20684	17.10379	19.33254	22.44274	24.13969	24.76553	24.52155	22.57798	19.82104	16.99427	13.85002
1998	13.10717	15.25846	17.24993	21.19343	23.36349	24.48269	24.90708	24.79855	23.28794	20.58570	16.93898	14.17895
1999	13.22298	15.00630	17.24232	20.60812	22.80969	24.28449	24.90748	24.38565	23.18035	20.05863	16.77982	13.90168
2000	12.32430	13.90121	17.29798	20.67422	23.03674	24.15630	24.67627	24.48956	22.80447	20.18565	16.68947	13.83875
2001	12.66861	14.38087	17.58976	20.43839	23.23670	24.10982	24.93453	24.62795	23.00110	20.59572	16.99794	13.57083
2002	13.31622	15.46412	18.24215	20.60843	23.06637	24.34122	25.35836	24.53164	22.93419	20.13988	16.72986	13.53578
2003	12.79954	14.43474	17.07866	20.38075	23.22125	24.43278	24.68177	24.58769	22.95315	20.05940	17.05252	13.88011
2004	12.69015	14.76807	17.92678	20.72800	22.68653	23.94269	24.57806	24.33165	23.14777	20.13639	17.25484	13.87552
2005	12.75448	13.78233	17.02063	20.76632	22.78080	24.77572	24.86599	24.60775	23.30576	20.33532	17.05777	13.23400
2006	12.99794	14.71250	17.19218	20.23390	22.95222	24.38026	25.03133	24.71884	23.01361	20.96664	17.28912	14.03938
2007	13.43582	15.07378	17.61691	20.48394	23.48876	24.49131	24.67724	24.70171	23.09679	20.37684	16.86562	13.94252
2008	12.36067	13.73491	18.13619	20.66613	22.71193	23.94896	24.73896	24.45372	22.97694	20.54477	17.05691	13.97525
2009	12.87896	15.18893	17.51858	20.55247	23.18394	24.58638	24.82574	24.73854	23.26629	20.51236	16.89762	13.85521
2010	12.89556	15.00602	17.99422	20.60926	23.43986	24.58293	25.32517	24.92402	23.11360	20.24914	17.36539	13.43564
2011	12.01797	14.30767	16.96784	20.28739	22.99368	24.32631	24.91556	24.57709	22.97664	20.23834	17.21146	13.77041
2012	12.44496	13.51301	17.26119	20.62158	23.45831	24.66302	25.21899	24.77204	23.05836	20.48724	16.94874	13.57143
2013	12.75366	14.61521	17.33937	19.98301	23.40596	24.34176	24.95132	24.77023				

Split the dataset into training and test

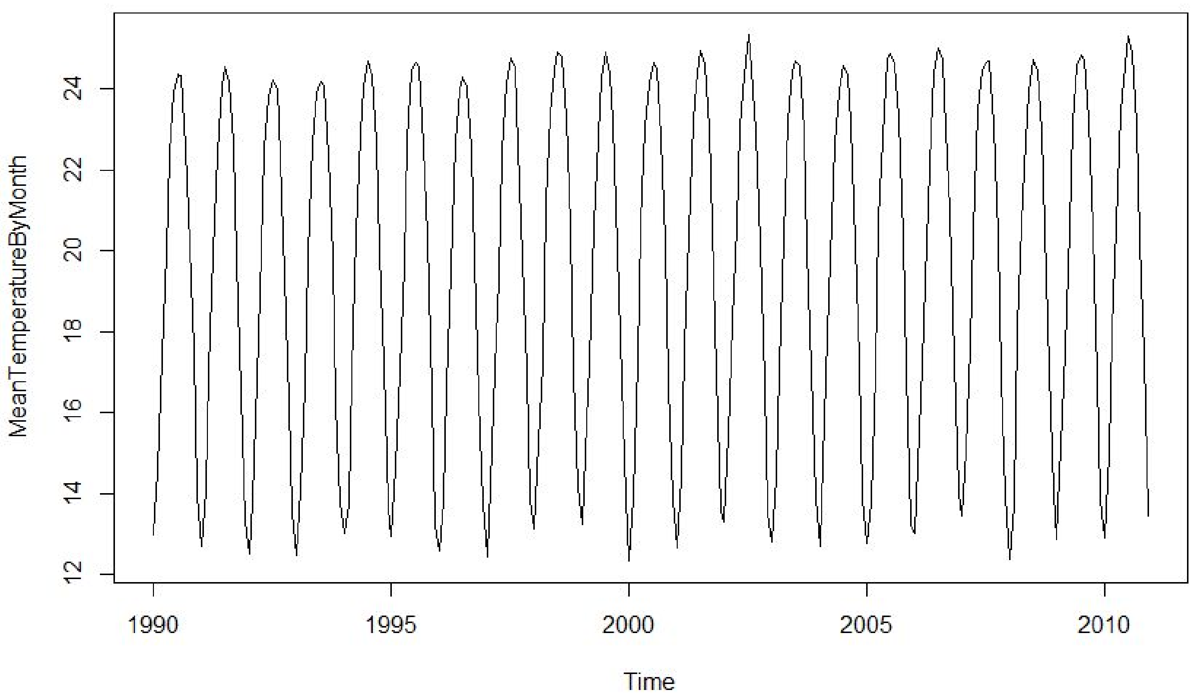
training dataset: from 1990-01 to 2010-12

```
training<-ts(MeanTemperatureByMonth,start=c(1990,1),end=c(2010,12), fre=12)
training
```

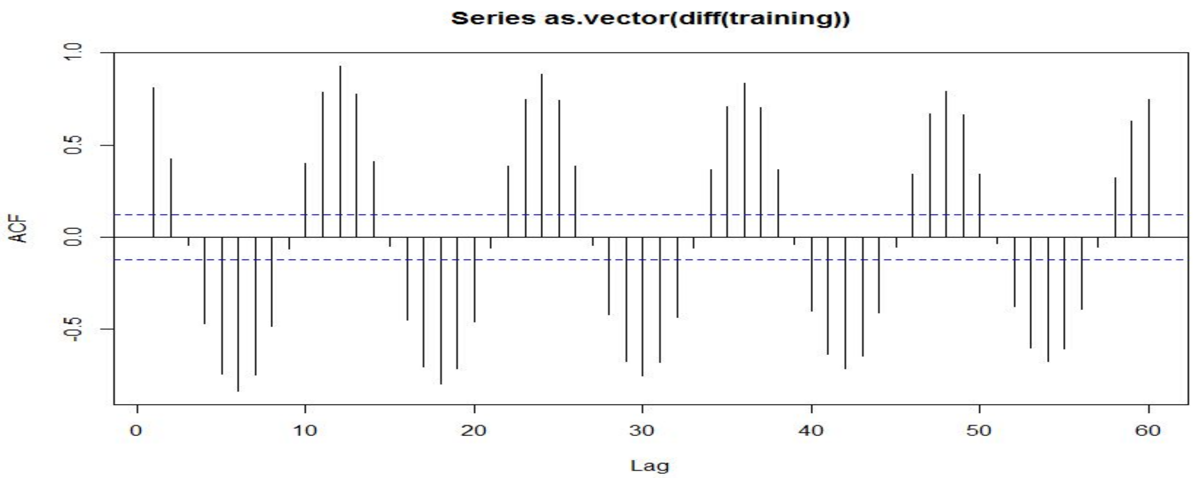
test dataset from 2011-01 to 2013-08

```
test<-ts(MeanTemperatureByMonth,start=c(2011,1),end=c(2013,8), fre=12)
test
```

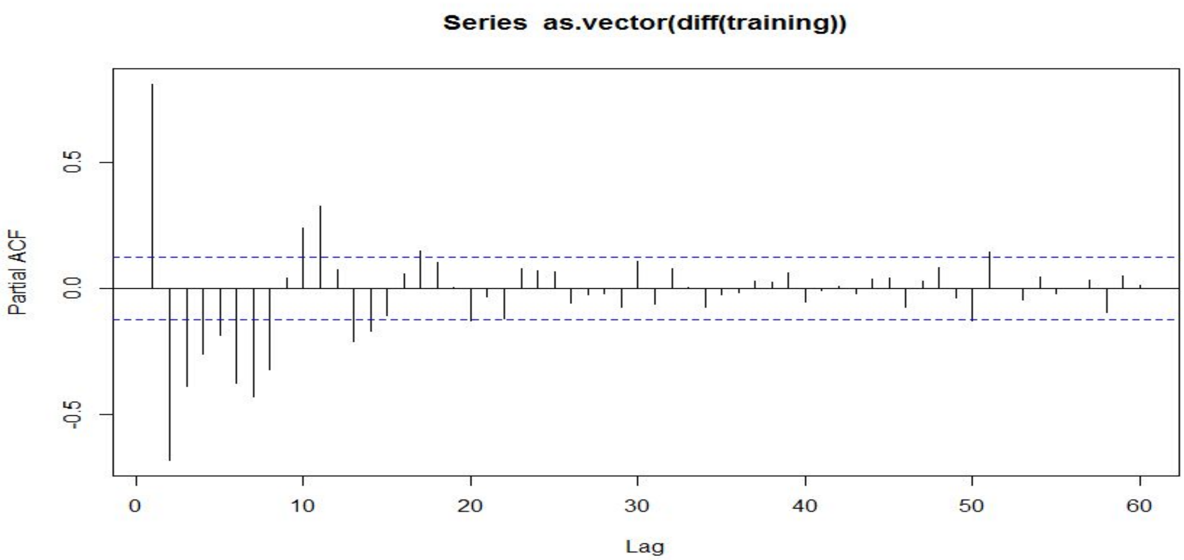
The plot of the training data



ACF Plot



PACF Plot



Choose the parameter for the ARIMA

Candidate one: ARIMA(1,1,1)(0,1,1)12

```
arima(x = training, order = c(1, 1, 1), seasonal = list(order = c(0, 1, 1),
    period = 12))

Coefficients:
         ar1      ma1     sma1
      0.2188  -0.8388  -0.9630
s.e.  0.0937   0.0599   0.1112

sigma^2 estimated as 0.08739:  log likelihood = -62.76,  aic = 131.52
```

Candidate two: ARIMA(3,1,1)(0,1,1)12

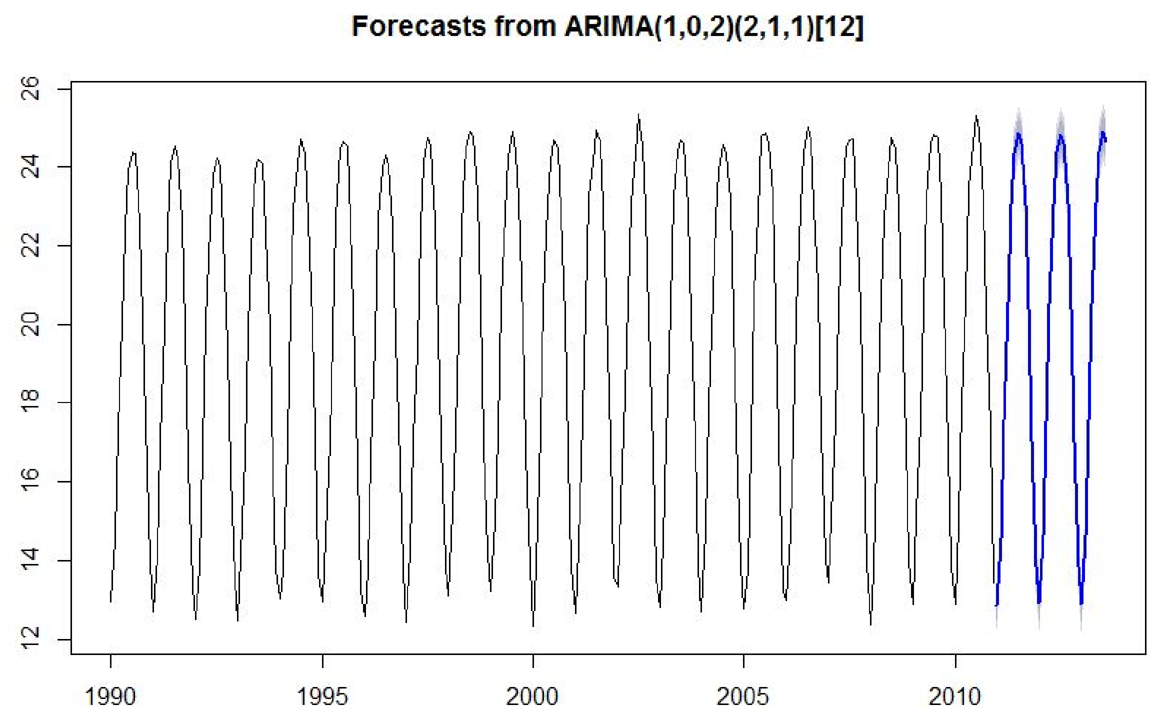
```
Call:
arima(x = training, order = c(3, 1, 1), seasonal = list(order = c(0, 1, 1),
    period = 12))

Coefficients:
         ar1      ar2     ar3     ma1     sma1
      0.3786  -0.0072  0.1765  -1.000  -0.9237
s.e.  0.0649   0.0694  0.0661   0.081   0.0645

sigma^2 estimated as 0.0861:  log likelihood = -61.22,  aic = 132.45
```

Candidate three: ARIMA(1,0,2)(2,1,1)12

```
Series: training
ARIMA(1,0,2)(2,1,1)[12]

Coefficients:
         ar1      ma1      ma2     sar1     sar2     sma1
      0.9638  -0.5964  -0.1704  -0.1137  -0.1256  -0.8635
s.e.  0.0292   0.0736   0.0736   0.0800   0.0772   0.0633

sigma^2 estimated as 0.08877:  log likelihood=-56.97
AIC=127.95   AICc=128.43   BIC=152.31
```

Choose the model which has the lowest AIC value. The candidate three is chosen.

Use the model to make prediction



Model evaluation

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	0.01993893	0.2870993	0.2134004	0.04628141	1.202766	0.637102	-0.01127898	NA
Test set	-0.40853755	0.4936912	0.4560682	-2.10688109	2.405546	1.361581	0.56433841	0.2287747

K-Means Clustering and Decision Modeling

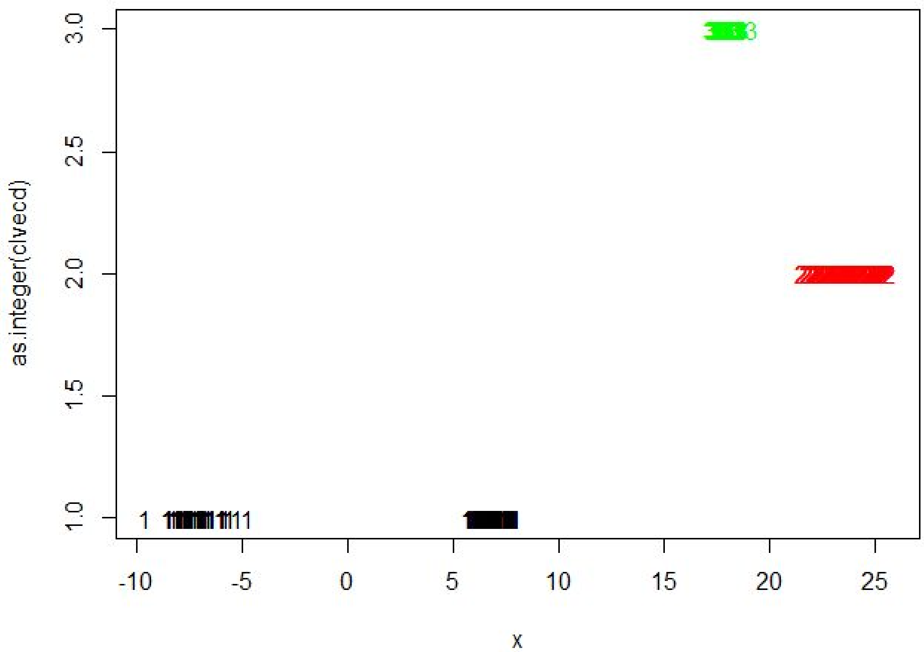
Parameters Used:-

- Year – year of observed values
- forestarea - Proportions of forest area in the country
- co2emission – CO2 emissions (metric ton per captia)
- poptot – Population in total
- country – country's respective code

The records in the dataset were binned into three categories using "k means clustering". 3 clusters are created using the algorithm. From cluster analysis, we can find which country falls into which bin.

```
K-means clustering with 3 clusters of sizes 159, 159, 53

Cluster means:
       [,1]
1  2.190298
2 23.879397
3 18.047000

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [71] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[141] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[211] 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[281] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[351] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3

Within cluster sum of squares by cluster:
[1] 6941.98364  190.79951   10.23514
 (between_SS / total_SS =  84.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
```

3 bins are created : 1 - low ,2 - medium and 3-high range temperatures. Countries in a particular region came under the same category. It can be visualized by following plot :



We use CART and C5.0 algorithm to build a decision tree.We classify the countries into categories of high, middle and low temperatures.The target variable for this decision tree will be tempCat and the predictor variables are other variables.

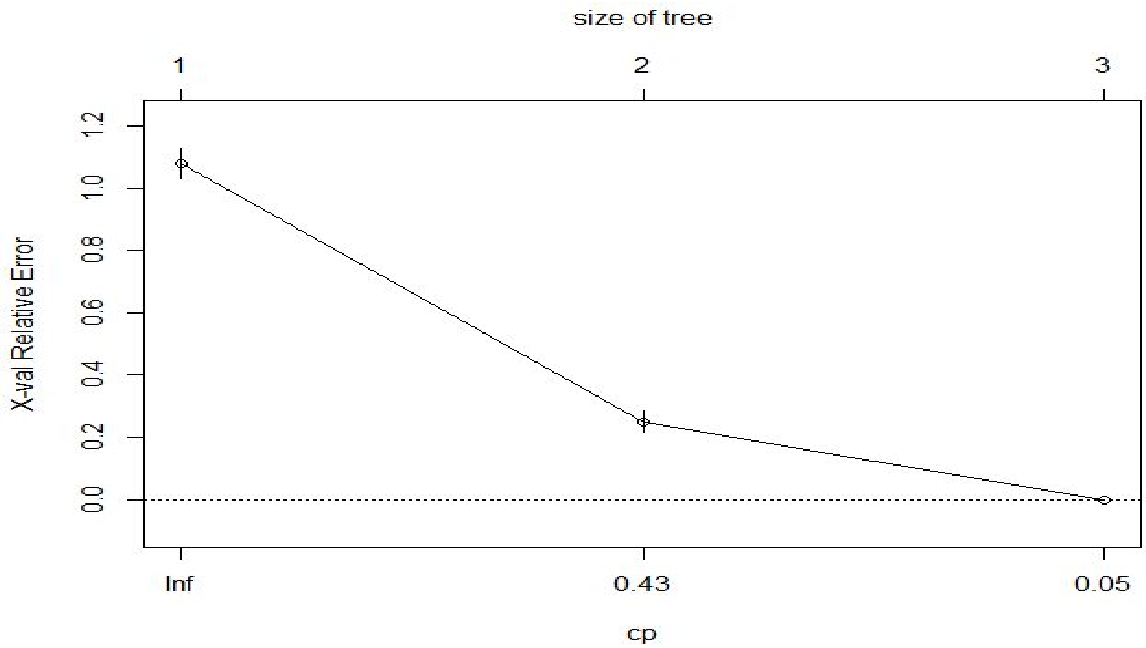
For developing this model,we need to split the records in our dataset into training and test datasets.For this we will create a new dataframe where rows are copies of the original data frame but selected based on random generation.We should check that we get same data in both the data frames by using summary and also head functions to check that records are randomized.

```
> plotcluster(climateData.norm$temperature,climateclusters$cluster)
> library(rpart)
> library(rpart.plot)
> data_rand <- climateData.norm[order(runif(371)), ] #creating  a new dataframe
> summary(climateData.norm$temperature)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -9.525   6.565  17.980  13.750  23.740  25.700
> summary(data_rand$temperature) # we check that we get the same data in both dataframes...
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -9.525   6.565  17.980  13.750  23.740  25.700
> head(climateData.norm$temperature)
[1] 6.720098 6.492202 6.910322 7.016024 6.274974 6.465711
> head(data_rand$temperature) # we check the order of both dataframes are different !
[1] 24.369567 -8.513794 18.524628 22.963850  5.689573 -5.186193
> prop.table(table(data_train$tempCat))

     HIGH       LOW    MEDIUM
0.4234234 0.4354354 0.1411411
> prop.table(table(data_test$tempCat))

     HIGH       LOW    MEDIUM
0.4736842 0.3684211 0.1578947
>
```

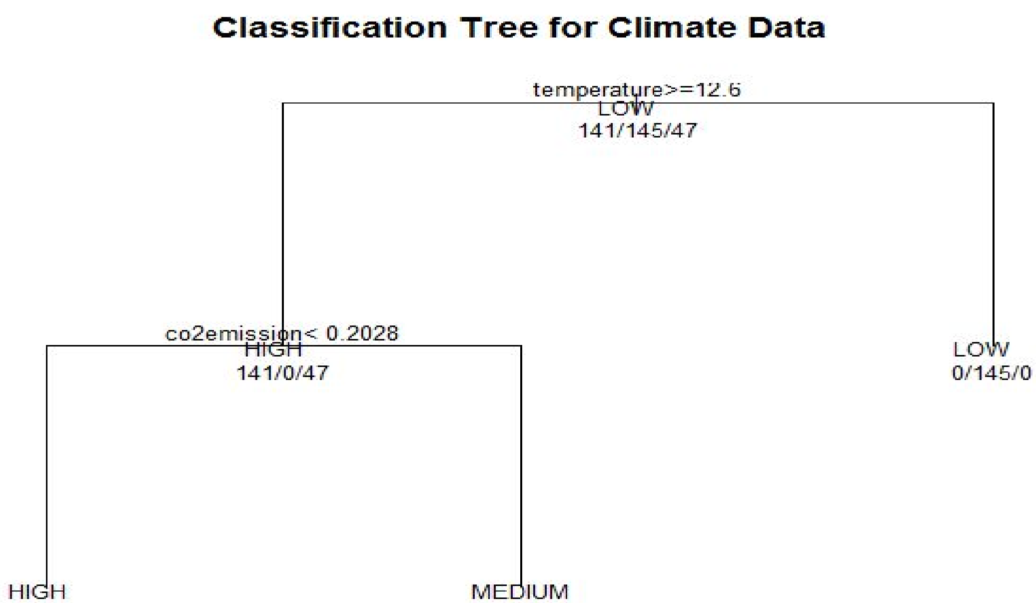
These proportions are nearly equal and so our partitions are not biased.We use CART analysis.

```
> summary(data_model_cart) # detailed summary of splits
Call:
rpart(formula = tempCat ~ forestarea + co2emission + poptot +
    temperature, data = data_train, method = "class")
  n= 333

    CP nsplit rel error   xerror       xstd
1 0.75      0      1.00 1.079787 0.04735215
2 0.25      1      0.25 0.250000 0.03379496
3 0.01      2      0.00 0.000000 0.00000000

Variable importance
temperature co2emission  forestarea      poptot
         45          36          10           8

Node number 1: 333 observations,    complexity param=0.75
  predicted class=LOW     expected loss=0.5645646  P(node) =1
    class counts:   141   145    47
   probabilities: 0.423 0.435 0.141
  left son=2 (188 obs) right son=3 (145 obs)
  Primary splits:
      temperature < 12.59678   to the right, improve=133.02550, (0 missing)
      co2emission < 0.1679556  to the left,  improve= 77.30134, (0 missing)
      forestarea  < 0.09856714 to the right, improve= 39.04606, (0 missing)
      poptot      < 0.1310031  to the right, improve= 17.23784, (0 missing)
  Surrogate splits:
      co2emission < 0.4579885  to the left,  agree=0.859, adj=0.676, (0 split)
      forestarea  < 0.369968   to the left,  agree=0.715, adj=0.345, (0 split)
      poptot      < 0.1310031  to the left,  agree=0.688, adj=0.283, (0 split)

Node number 2: 188 observations,    complexity param=0.25
  predicted class=HIGH    expected loss=0.25  P(node) =0.5645646
    class counts:   141     0    47
   probabilities: 0.750 0.000 0.250
  left son=4 (141 obs) right son=5 (47 obs)
  Primary splits:
      co2emission < 0.2028153  to the left,  improve=70.5, (0 missing)
      temperature < 20.34192   to the right, improve=70.5, (0 missing)
      forestarea  < 0.1773834  to the right, improve=23.5, (0 missing)
      poptot      < 0.03831503 to the right, improve=23.5, (0 missing)
  Surrogate splits:
      temperature < 20.34192   to the right, agree=1, adj=1, (0 split)

Node number 3: 145 observations
  predicted class=LOW     expected loss=0  P(node) =0.4354354
    class counts:     0   145     0
   probabilities: 0.000 1.000 0.000

Node number 4: 141 observations
  predicted class=HIGH    expected loss=0  P(node) =0.4234234
    class counts:   141     0     0
   probabilities: 1.000 0.000 0.000

Node number 5: 47 observations
  predicted class=MEDIUM  expected loss=0  P(node) =0.1411411
    class counts:     0     0    47
   probabilities: 0.000 0.000 1.000
```

The cross-validation of the results can also be visually represented as follows:



Classification tree is made.CART analysis is used.

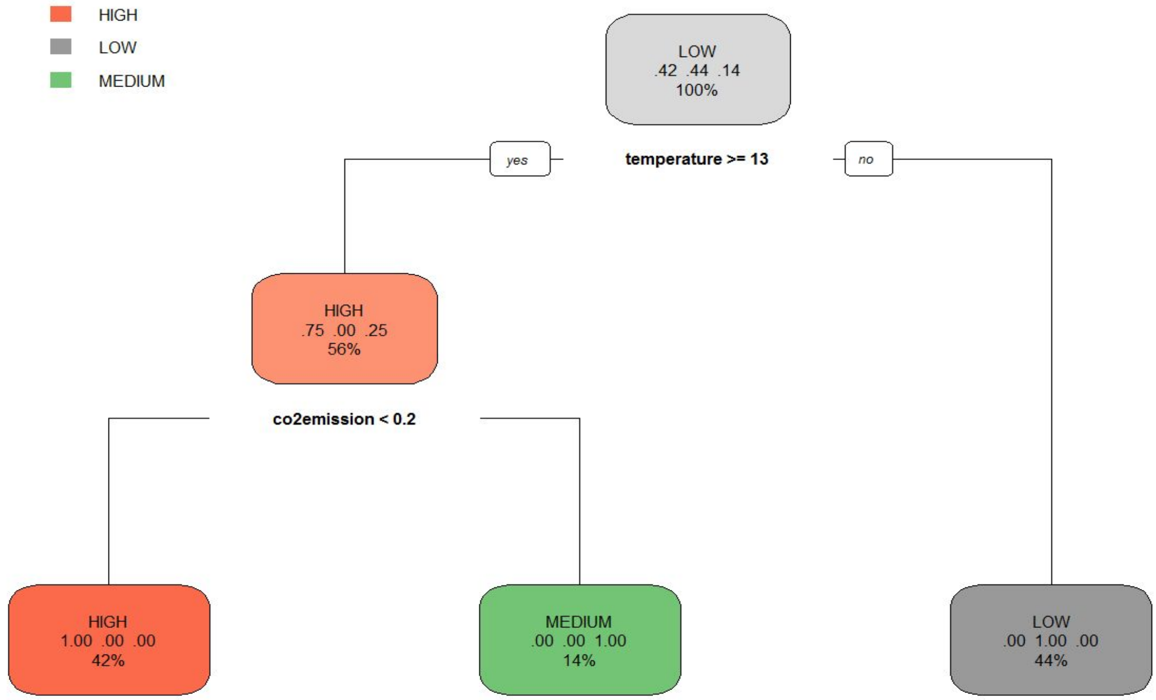


Then we develop the model using C5.0 algorithm.

```
Call:
C5.0.default(x = x, y = y)

Classification Tree
Number of samples: 333
Number of predictors: 4

Tree size: 3

Non-standard options: attempt to group attributes

> summary(data_model_c50)

Call:
C5.0.default(x = x, y = y)


C5.0 [Release 2.07 GPL Edition]        Thu Apr 27 06:07:16 2017
-------------------------------

Class specified by attribute `outcome'

Read 333 cases (5 attributes) from undefined.data

Decision tree:

temperature <= 8.045993: LOW (145)
temperature > 8.045993:
:...temperature <= 19.14038: MEDIUM (47)
    temperature > 19.14038: HIGH (141)


Evaluation on training data (333 cases):

            Decision Tree
          ----------------
          Size      Errors

             3    0( 0.0%)   <<


           (a)   (b)   (c)    <-classified as
          ----  ----  ----
           141                (a): class HIGH
                 145          (b): class LOW
                        47    (c): class MEDIUM


        Attribute usage:

        100.00% temperature
```

Based on summary of C5.0 model ,we can see that temperature does not have any contribution to classification of records. So another model is used to see how other variables influence the prediction.

These results can be visualized using rpart function.



```
Call:
C5.0.default(x = data_train[, c(2, 3, 4)], y = as.factor(data_train$tempCat))


C5.0 [Release 2.07 GPL Edition]         Thu Apr 27 06:11:11 2017
-------------------------------

Class specified by attribute `outcome'

Read 333 cases (4 attributes) from undefined.data

Decision tree:

co2emission > 0.1644467:
:...forestarea <= 0.179574: MEDIUM (47)
:   forestarea > 0.179574: LOW (106)
co2emission <= 0.1644467:
:...poptot <= 0.4779297: HIGH (111)
    poptot > 0.4779297:
    :...forestarea <= 0.265372: LOW (39)
        forestarea > 0.265372: HIGH (30)


Evaluation on training data (333 cases):

	    Decision Tree
	  ----------------
	  Size      Errors

	     5    0( 0.0%)   <<


	   (a)   (b)   (c)    <-classified as
	  ----  ----  ----
	   141                (a): class HIGH
	         145          (b): class LOW
	                47    (c): class MEDIUM


	Attribute usage:

	100.00% co2emission
	 66.67% forestarea
	 54.05% poptot


Time: 0.0 secs
```

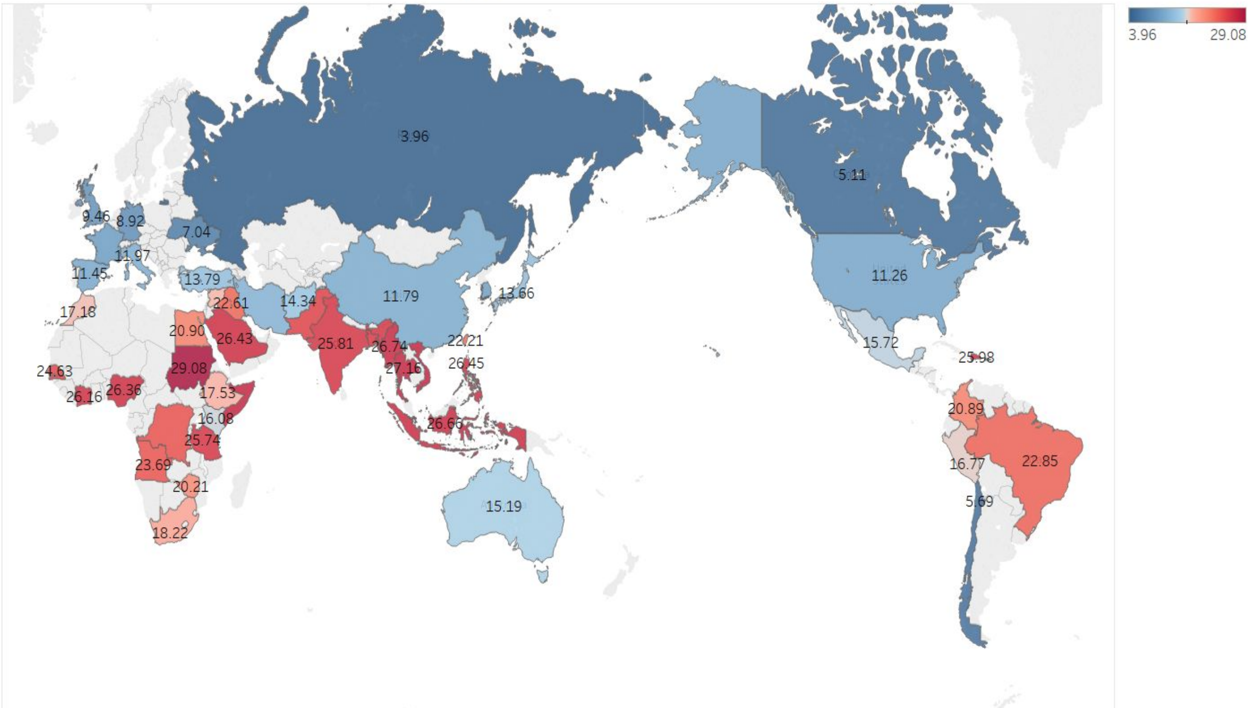
We see that there are other attributes that contribute to classification. We test these results by using confusion matrix.Accuracy has been computed.

```
> CrossTable(data_test$tempCat, data_predict_c50,
+            prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
+            dnn = c('actual ', 'predicted '))


   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  38


             | predicted
      actual |      HIGH |       LOW |    MEDIUM | Row Total |
-------------|-----------|-----------|-----------|-----------|
        HIGH |        18 |         0 |         0 |        18 |
             |     0.474 |     0.000 |     0.000 |           |
-------------|-----------|-----------|-----------|-----------|
         LOW |         0 |        14 |         0 |        14 |
             |     0.000 |     0.368 |     0.000 |           |
-------------|-----------|-----------|-----------|-----------|
      MEDIUM |         0 |         0 |         6 |         6 |
             |     0.000 |     0.000 |     0.158 |           |
-------------|-----------|-----------|-----------|-----------|
Column Total |        18 |        14 |         6 |        38 |
-------------|-----------|-----------|-----------|-----------|
```

```
Total Observations in Table:  38


             | predicted
      actual |      HIGH |       LOW |    MEDIUM | Row Total |
-------------|-----------|-----------|-----------|-----------|
        HIGH |        18 |         0 |         0 |        18 |
             |     0.474 |     0.000 |     0.000 |           |
-------------|-----------|-----------|-----------|-----------|
         LOW |         0 |        14 |         0 |        14 |
             |     0.000 |     0.368 |     0.000 |           |
-------------|-----------|-----------|-----------|-----------|
      MEDIUM |         0 |         0 |         6 |         6 |
             |     0.000 |     0.000 |     0.158 |           |
-------------|-----------|-----------|-----------|-----------|
Column Total |        18 |        14 |         6 |        38 |
-------------|-----------|-----------|-----------|-----------|

> data_predict_c50_01<- predict(data_model_c50_01,data_test)
> CrossTable(data_test$tempCat, data_predict_c50_01,
+            prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
+            dnn = c('actual ', 'predicted '))

   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  38


             | predicted
      actual |      HIGH |       LOW |    MEDIUM | Row Total |
-------------|-----------|-----------|-----------|-----------|
        HIGH |        18 |         0 |         0 |        18 |
             |     0.474 |     0.000 |     0.000 |           |
-------------|-----------|-----------|-----------|-----------|
         LOW |         0 |        14 |         0 |        14 |
             |     0.000 |     0.368 |     0.000 |           |
-------------|-----------|-----------|-----------|-----------|
      MEDIUM |         0 |         0 |         6 |         6 |
             |     0.000 |     0.000 |     0.158 |           |
-------------|-----------|-----------|-----------|-----------|
Column Total |        18 |        14 |         6 |        38 |
-------------|-----------|-----------|-----------|-----------|
```

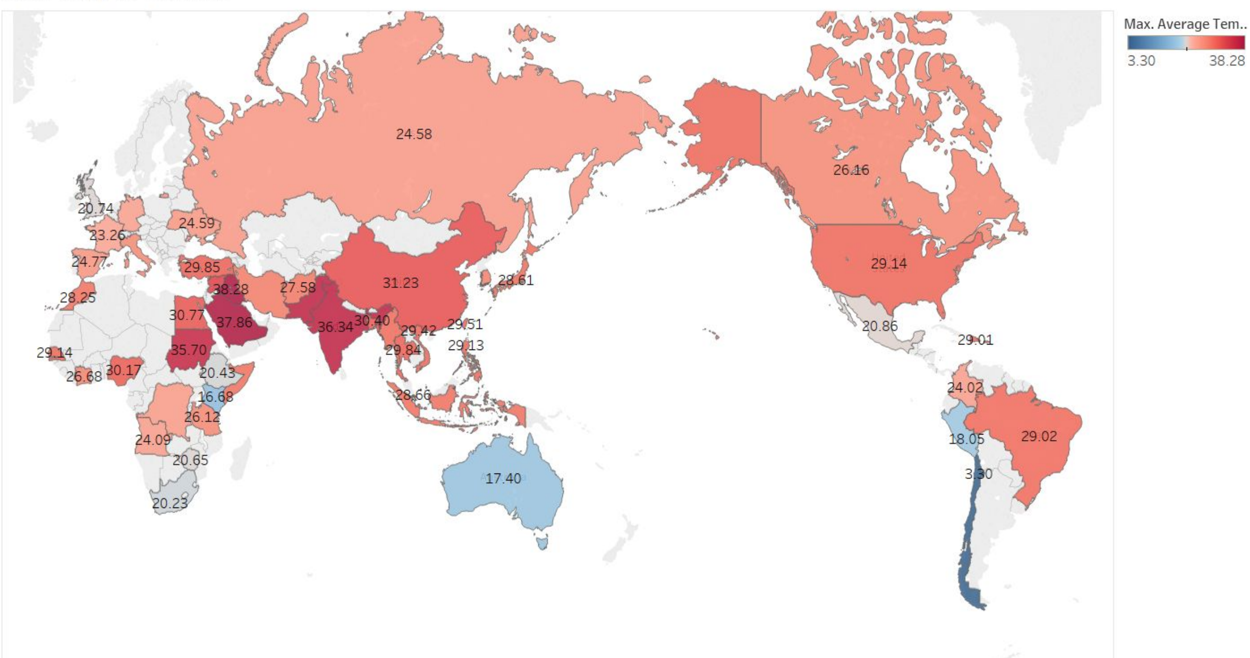
We used Tableau for some of our visualizations.

Average Temp for all the Years from 1750-2012



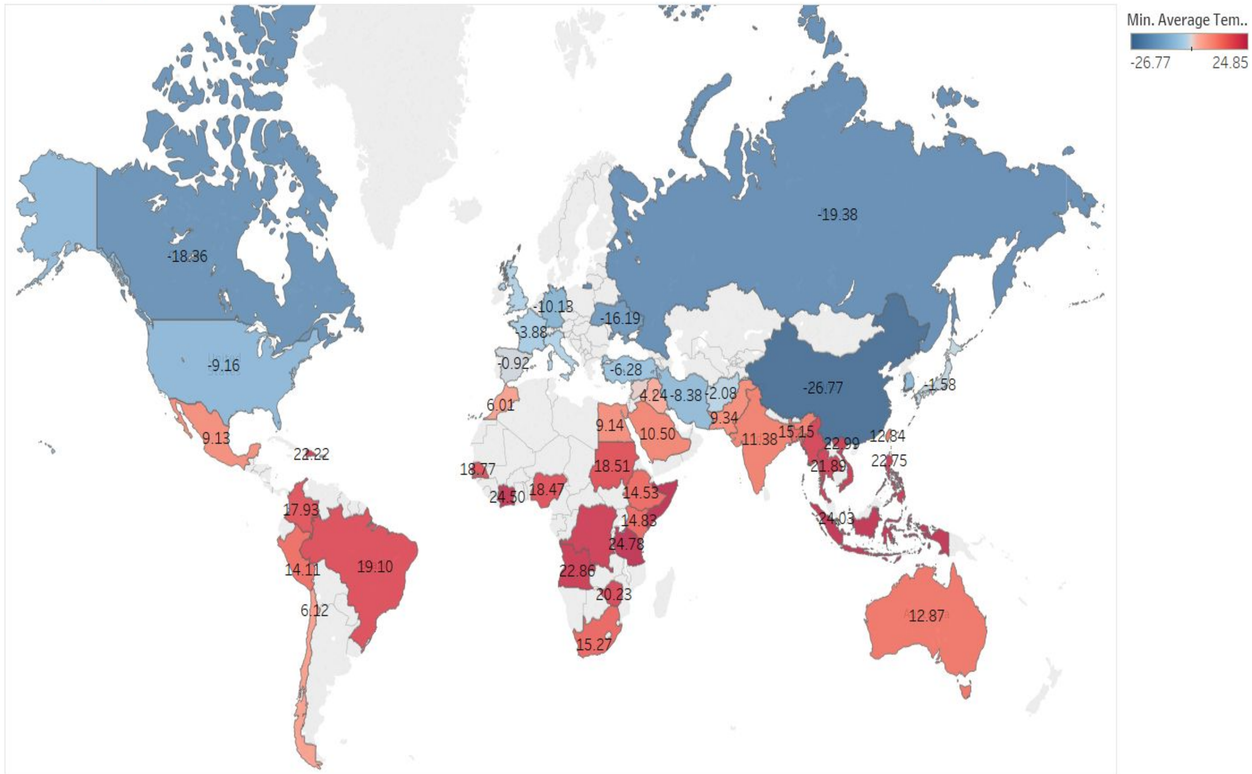
Map based on Longitude (generated) and Latitude (generated). Color shows average of Average Temperature. Details are shown for Country1.

Max Temp in Summer



Map based on Longitude (generated) and Latitude (generated). Color shows maximum of Average Temperature. Details are shown for Country1. The data is filtered on Dt Month, which keeps June, July and August.

Min Temp in Winters



- Map based on Longitude (generated) and Latitude (generated). Color shows minimum of Average Temperature. Details are shown for Country1. The data is filtered on Dt Month, which keeps January, February, November and December.

These 2 visualizations show that winters are colder and summer are warmer in some countries.

Observations

Over the years it is observed that as the population of the country increases,the CO2 emission increases leading to the rise in the temperatures.This can be justified from the plots as well as from the correlation values.The gradual increase in temperature in turn leads to increase in sea level .ARIMA Time series model also helped us in creating a model to predict future temperatures. From k-means clustering, we categorized the records in the dataset into three categories-High, Middle and Low. Models designed using CART and C5.0 models predict the results well.Hence , we can say that the predictive capabilities are very much high for these chosen models.

Limitations

- Although CO2 emissions were expected to have a very high impact on climate change.We could not see any such strong dependence.On doing some background research we understood that the given temperatures in our dataset are with respect to land areas only.We don't have any data for sea or ocean temperature.But a major geographical area on Earth is covered with water.So, analyzing the water temperature as well might help us get more accurate knowledge about climate change.

- We limited the range of predictor variables for classifying the temperature category.There are many more causes and effects of global warming and climate change.This makes the model more or less biased to certain to predictor variables.

References

- https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data
- http://www.wri.org/sites/default/files/CAIT_Country_GHG_Emissions_-_csv_02022017.zip
- http://data.worldbank.org/indicator/SP.POP.TOTL
- http://data.okfn.org/data/core/sea-level-rise
- http://www.ucsusa.org/global_warming/science_and_impacts/science/each-countrys-share-of-co2.html#.WNRNFjsrI2w
- http://www.statmethods.net/advstats/timeseries.html
- http://people.duke.edu/~rnau/411arim3.htm
- https://link.springer.com/article/10.1007/s00376-012-1252-3