Global CO2 emissions: A comparison

MATH1324 Assignment 3

Anup Sakpal (s3801788), Vignesh Gopalakrishnan (s3795594), Kanishk Jain (s3810978)

Last updated: 22 October, 2019

RPubs link information

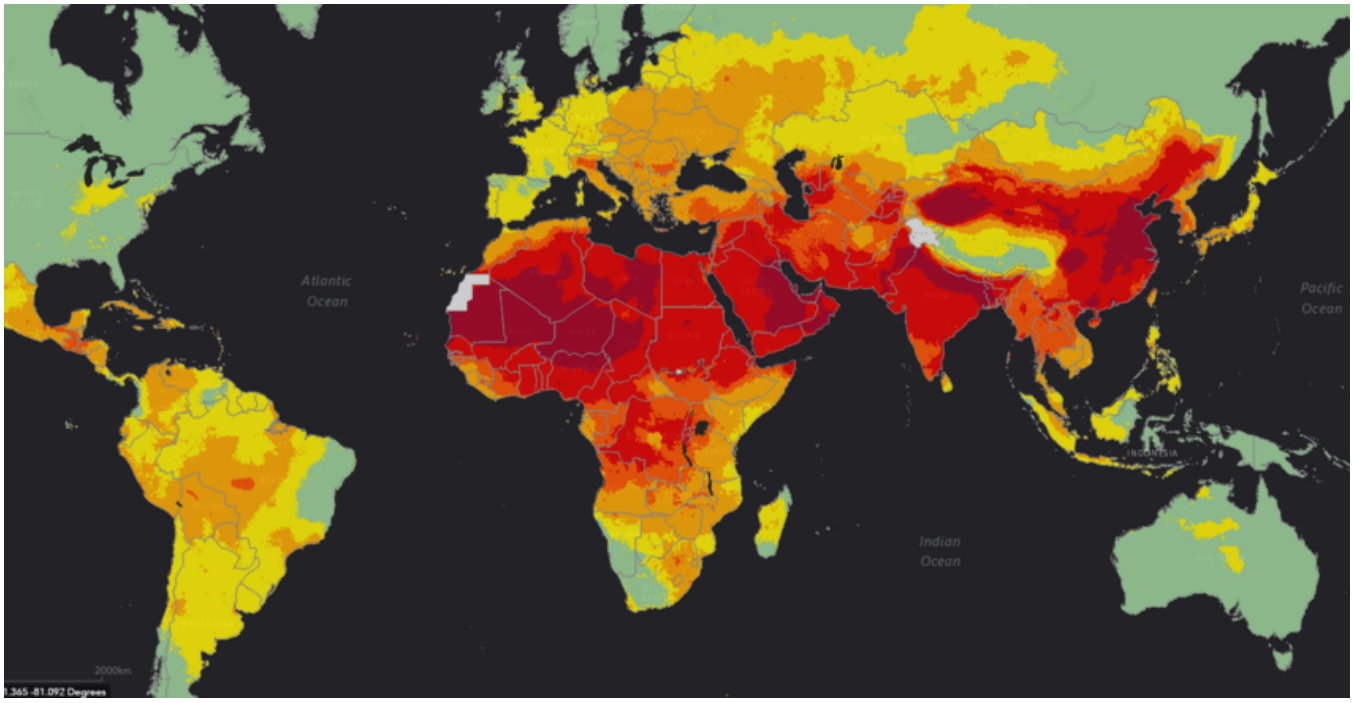
rpubs link : http://rpubs.com/gvignu/541660

Introduction

- Global pollution has been a topic of great concern over the past two decades
- Air pollution: mixing of unwanted pollutants and gases reaching to harmful concentrations
- Among the different pollutants CO2 has been the primary pollutant causing air pollution which is primarily emitted by burning of fossil fuel, forest fires etc
- excessive CO2 emissions increase the global temperature leading to global warming

Introduction Cont.

- A lot of discussion has been going on whether the pollution has increased or decreased over the past two decades.
- A lot of awareness campaigns and world treaties have been signed over the past decade by all the countries in the world to cut down CO2 emissions.



Problem Statement

- The investigation seeks to explore if there was any difference in the CO2 emission between two decades viz. 1994-2003 and 2004-2013
- We combined the CO2 emissions of years from 1994 to 2003 and 2004 to 2013 and performed paired t-test to find out if there was a significant difference in pollution over the two decades.

Data

- We have collected the dataset through Kaggle.
- This open data set study explores the impact of countries on global warming
- https://www.kaggle.com/catamount11/who-is-resposible-for-global-warming#Metadata_Country_API_EN.ATM.CO2E.PC_DS2_en_csv_v2_10576797.csv
- The method of data sampling is not known as this is not mentioned by the author of the dataset

Data Cont.

- The original dataset consists of Country Name, Country Code, Indicator Name, Indicator Code and CO2 emission values from 1960s
- We have simplified the dataset for our study.The simplified dataset consists of Country name, CO2 emission values from 1993 to 2014
- Country name is a factor with 246 levels, each representing one country
- CO2 emissions (metric tonnes per capita) from 1993 to 2014 spanning 20 years, is a numeric value

Descriptive Statistics and Visualisation

- We removed the countries which has missing values of emission for more than 11 years as imputing such values with mean/median of rest of the values will make the dataset baised
- The countries with missing values less than 10 were found out and their summary statistics were run to find out the normality of the data for such countires. We also ran Shapiro-Wilk normality test for confirm the normality
- The tables were gathered using tidyr function to find the mean and median for each country
- for countries which followed the normality, the missing values were replaced by mean and for countries which did not fit the normality, the missing values were replaced by median
- After replacing the missing values we combined the years and made them into two decades. The final dataset after preprocessing has three columns namely Country Name (factors with 241 levels), decade_1 (numeric), decade_2 (numeric), d (differnce between decade_1 and decade_2)
- Even though small values and large values of pollution by some countries may seem like outliers, they should not be removed as they can impact the global mean pollution. Hence, it is important in our investigation to keep all outliers. So, we haven't removed any outliers as they are important in determining an unbaised golbal mean

Continued

```
CO2_rev2[62,c(20,21,22)] <- sum_stats_country[62,'mean']
CO2_rev2[113,2:5] <- sum_stats_country[113,'median']
CO2_rev2[115,2] <- sum_stats_country[115,'mean']
CO2_rev2[151,2:12] <- sum_stats_country[151,'mean']
CO2_rev2[217,2:9] <- sum_stats_country[217,'median']
CO2_rev2[237,c(2:4,22)] <- sum_stats_country[237,'mean']

CO2_1994_2003 <- CO2_rev2[,2:11]
CO2_2004_2013 <- CO2_rev2[,12:21]
decade_1 <- rowSums(CO2_1994_2003)/10
decade_2 <- rowSums(CO2_2004_2013)/10
CO2_rev4 <- cbind(CO2_rev2[,1],decade_1,decade_2)
CO2_rev4 <- CO2_rev4 %>% mutate(d = decade_2 - decade_1)
CO2_rev4
```

Decsriptive Statistics Cont.

- we performed a descriptive summary of the two decades
- density plot was plotted which suggests a t-distribution for the given dataset
- qqplot was ploted to the test the normality along with the Shapiro-Wilk's test and a histogram with a normal curve overlay
- Even though the results do not confirm the normality since the sample size taken (241) is greater than the minimum sample (30) it is safe to assume normality as per Central Limit Theorem

```
knitr::kable(table1)
```

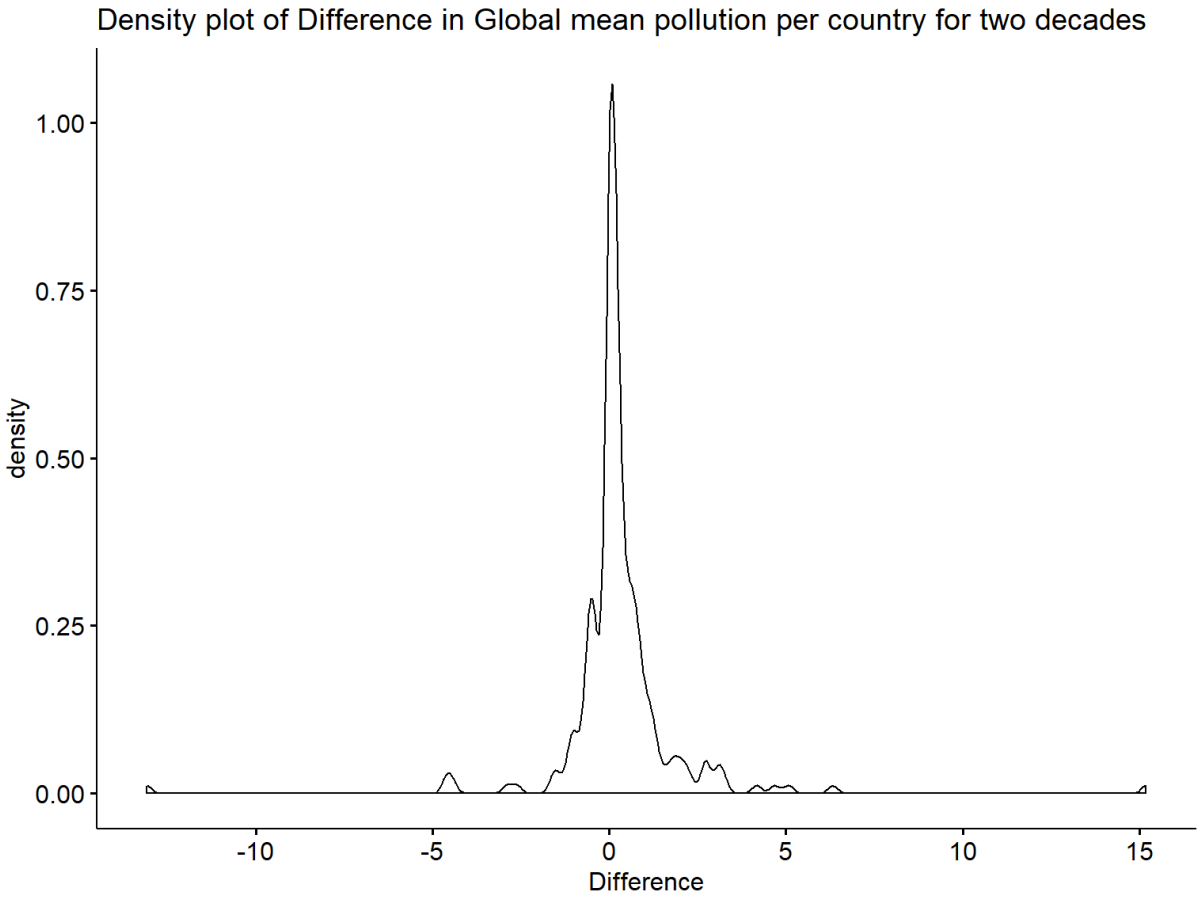
Min	Q1	Median	Q3	Max	Mean	SD	n	Missing
0.0207126	0.6250148	2.349967	6.22987	61.69089	4.512485	6.324376	241	0

```
knitr::kable(table2)
```

Min	Q1	Median	Q3	Max	Mean	SD	n	Missing
0.0253415	0.8056525	2.84751	6.57605	48.60564	4.799115	6.074046	241	0

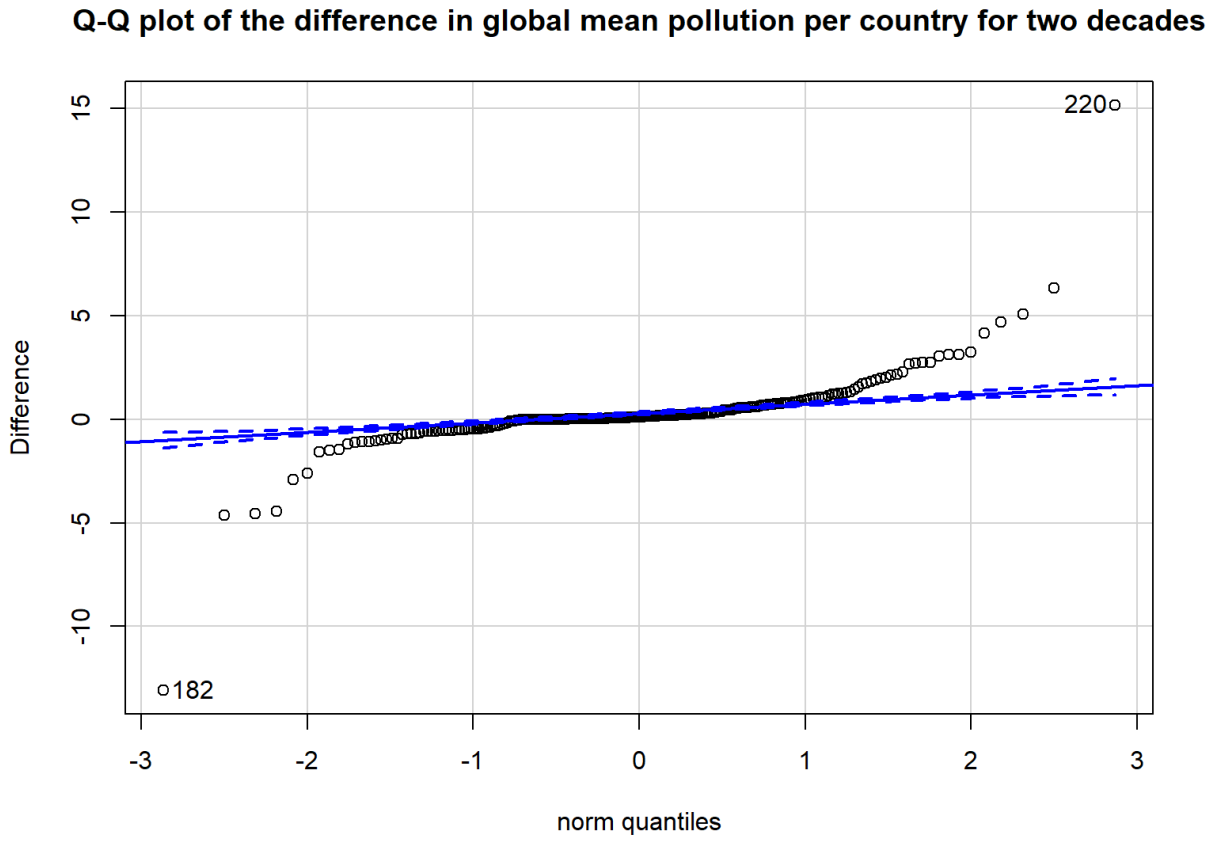
Continued

```
ggdensity(CO2_rev4$d,
          main = "Density plot of Difference in Global mean pollution per country for two decades",
          xlab = "Difference")
```



Continued

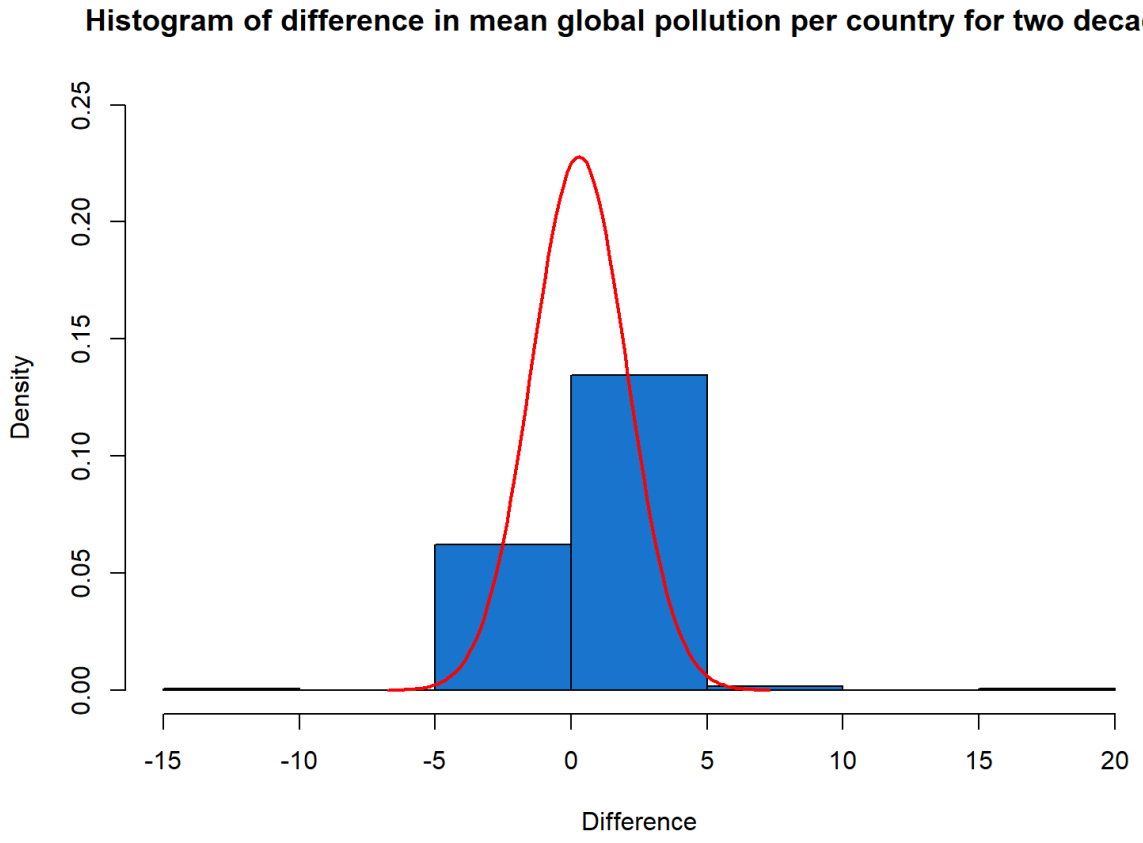
```
qqPlot(CO2_rev4$d, dist="norm", ylab = "Difference", main = "Q-Q plot of the difference in global mean pollution per country for
  two decades")
```




```
## [1] 220 182
```

Continued

```
x <- seq(min(CO2_rev4$d),max(CO2_rev4$d))
mu <- mean(CO2_rev4$d)
sd <- sd(CO2_rev4$d)
CO2_rev4$d %>% hist(xlab="Difference",
main="Histogram of difference in mean global pollution per country for two decades", prob=TRUE, ylim = c(0, 0.25), col =
  "dodgerblue3")
curve(dnorm(x,mu,sd),xlim = c(mu-sd*4, mu+sd*4),col="red", add=TRUE, lwd= 2)
```



Hypothesis Testing

- Since, we measure the same sample twice, the measurements are said to be "paired" or "dependent". The dataset measures the pollution of same countries in two different decades. Hence, the paired-samples t-test, also known as the dependent samples t-test, was used to check for a statistically significant mean change or difference in pollution in this situation.
- the paired sample t-test assumes the data are normally distributed. In our case normality can be assumed due to a large sample size as per Central limit theorem
- The statistical hypotheses for the paired-samples t-test are as follows:

$$H_0 : \mu\Delta = 0$$

$$H_A : \mu\Delta \neq 0$$

```
t.test(CO2_rev4$decade_2, CO2_rev4$decade_1,
       paired = TRUE,
       alternative = "two.sided")
```

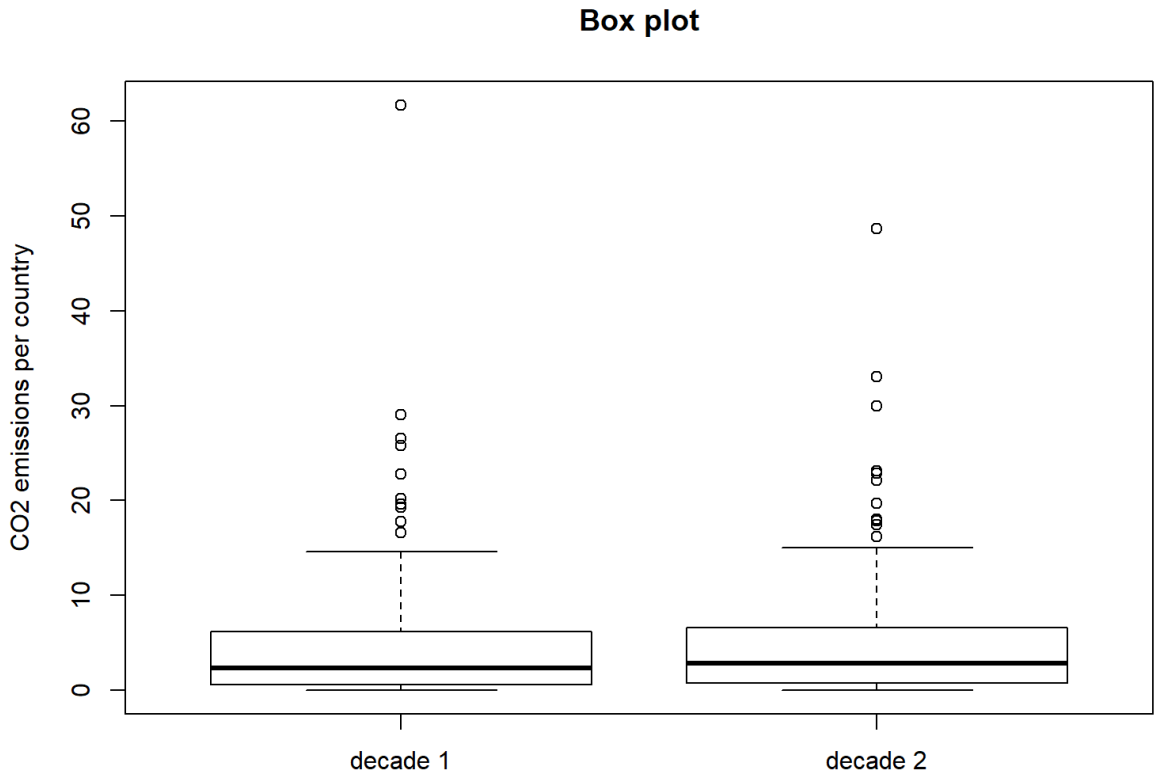
```
##
##  Paired t-test
##
## data:  CO2_rev4$decade_2 and CO2_rev4$decade_1
## t = 2.5419, df = 240, p-value = 0.01166
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06450068 0.50875958
## sample estimates:
## mean of the differences
##               0.2866301
```

Hypothesis Testing Cont.

- R reported t = 2.54, degrees of freedom = 240 at alpha = 0.05 (95% confidence interval)
- the p-value was reported to be 0.012 < 0.05, hence we reject the null hypothesis
- We find that there is a statistically significant difference in the mean global pollution between the two decades
- the mean difference was found to be 0.287 which means decade_2 had a higher pollution than decade_1
- A box plot and a dependent sample assessment plot were used to visualise the paired sample t-test
- The box plot plot revealed a higher average for pollution for decade_2 over decade_1
- the dependent sample assessment plot was not very clear due to large data in the dataset but the decade_2 clearly shows higher value than decade_1 but it is not clear if the 95% CI line (green) overlaps the identity line to say the significance of difference

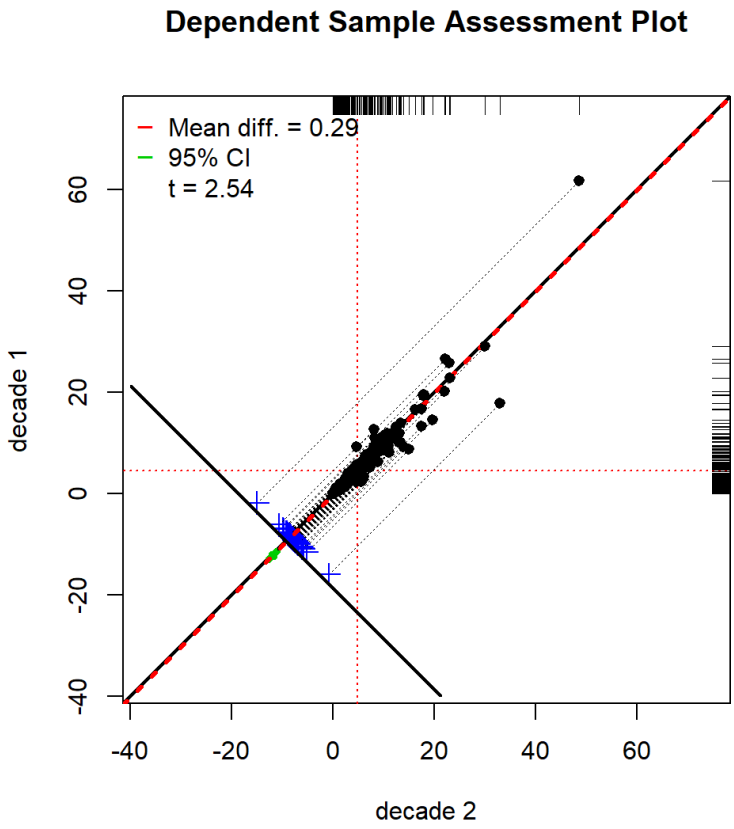
Continued

```
boxplot(
  CO2_rev4$decade_1,
  CO2_rev4$decade_2,
  main = "Box plot",
  ylab = "CO2 emissions per country",
  xlab = ""
)
axis(1, at = 1:2, labels = c("decade 1", "decade 2"))
```



Continued

```
granova.ds(
  data.frame(CO2_rev4$decade_2, CO2_rev4$decade_1),
  xlab = "decade 2",
  ylab = "decade 1"
)
```

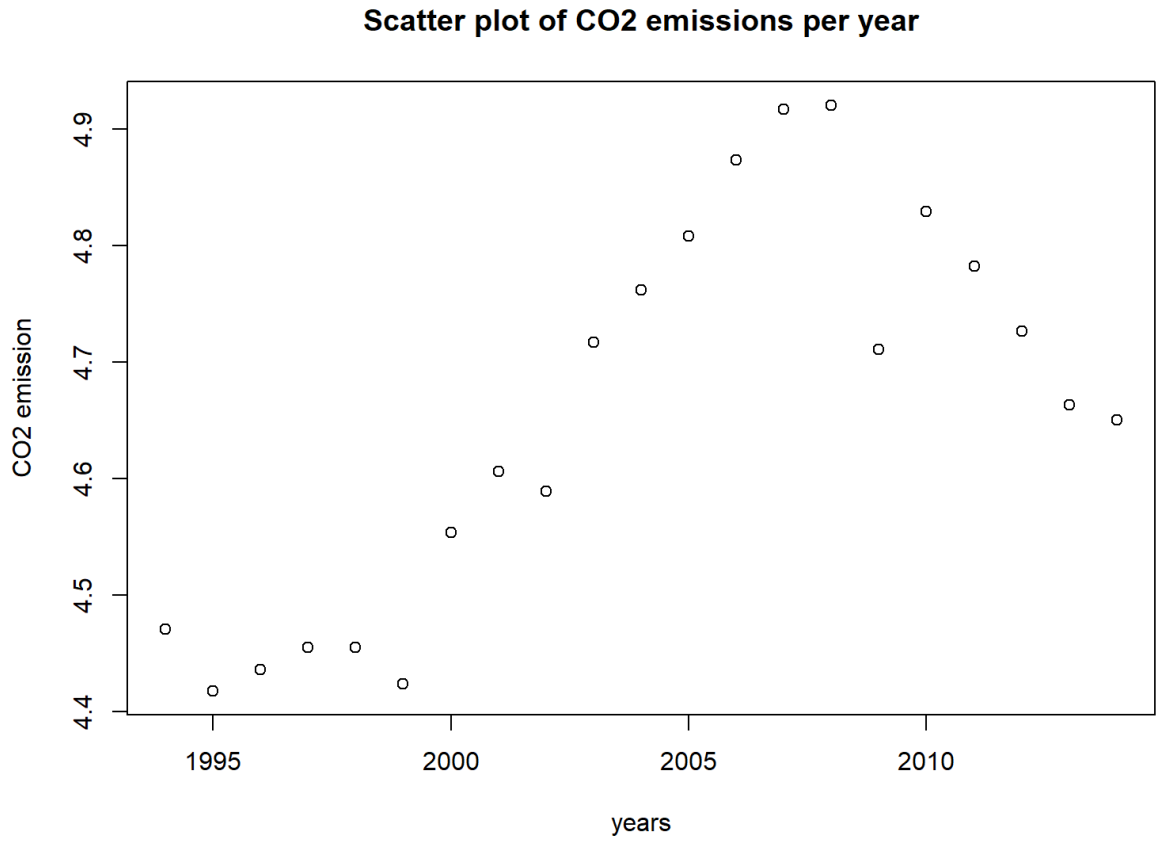



```
##             Summary Stats
## n                 241.000
## mean(x)             4.799
## mean(y)             4.512
## mean(D=x-y)         0.287
## SD(D)               1.751
## ES(D)               0.164
## r(x,y)              0.961
## r(x+y,d)           -0.144
## LL 95%CI            0.065
## UL 95%CI            0.509
## t(D-bar)            2.542
## df.t              240.000
## pval.t              0.012
```

Continued

- The year wise plot of global mean for all the 20 years was plotted to observe year wise change
- the second half of decade_2 shows a decrease in the pollution levels

```
mean_global <- colMeans(CO2_rev2[,-1])
years <- c(1994:2014)
CO2_mean_global <- as.data.frame(cbind(years,mean_global))
plot(mean_global ~ years, data = CO2_mean_global, main="Scatter plot of CO2 emissions per year", xlab = "years", ylab = "CO2
  emission")
```



Discussion

- There was a statistically significant difference in the pollution levels of decade_1 (1994 to 2003) and decade_2 (2004 to 2013) as seen from the paired sample t-test
- The visualisation of the test results through box plot and dependent sample assessment plot reveal that the decade_2 is more polluting than decade_1
- the data was tidied properly with the most suited missing values as per the result of normality testing for each country which improve the accuracy of the obtained results
- the results could only be obtained for predicting global mean pollution and not for each and every country
- Regression or time series analysis can be performed in this dataset for predicting the global pollution in the coming years
- From our investigation it can be clearly concluded that even after the awarness campaigns and global treaties signed by different countries over the past decade there was a statistically significant rise in global mean pollution over the decades 1994-2003 and 2004 to 2013

References

-[1]William Doane, Assessment & Evaluation, Statistics Dependent Sample Assessment, Plots using GRANOVA AND R, 2010 JULY 2014, https://drdoane.com/dependent-sample-assessment-plots/

-[2] Testing the Null: Data on Trial, James Baglin, https://astral-theory-157510.appspot.com/secured/MATH1324_Module_07.html

-[3] Sampling: Randomly Representative, James Baglin, https://astral-theory-157510.appspot.com/secured/MATH1324_Module_07.html

-[4] Tidy and Manipulate: Tidy Data Principles and Manipulating Data, Dr. Anil Dolgun, http://rare-phoenix-161610.appspot.com/secured/Module_04.html

-[5] Scan: Missing Values, Dr. Anil Dolgun, http://rare-phoenix-161610.appspot.com/secured/Module_05.html

-[6] Scan: Outliers, Dr. Anil Dolgun, http://rare-phoenix-161610.appspot.com/secured/Module_06.html

-[7] Get: Importing, Scraping and Exporting Data with R, Dr Anil Dolgun http://rare-phoenix-161610.appspot.com/secured/Module_02.html