EL NINO (UCI/ML)

```
library(ggplot2) # Data Visualisation
library(moments) # Moments: skewness, kurtosis
```

Importing table and assigning column

```
taoall <- data.table::fread('elnino/tao-all2.dat.gz')

colnames(taoall) <- c("obs", "year", "month","day", "date", "lat", "lng",
                      "zon_wind", # Zonal wind - for west, + for east
                      "mer_wind", # Meridional wind - for south, + for north
                      "humidity", "air_temp", "sub_temp"
                      )
```

Column conversion and Basic stats

```
## Conversion of the variables into numeric (right format for columns)
taoall$zon_wind <- as.numeric(taoall$zon_wind)
taoall$mer_wind <- as.numeric(taoall$mer_wind)
taoall$humidity <- as.numeric(taoall$humidity)
taoall$air_temp <- as.numeric(taoall$air_temp)
taoall$sub_temp <- as.numeric(taoall$sub_temp)

taoall$date <- lubridate::ymd(taoall$date)  # Formatting data column

### nrow(taoall)
### Summary for data set
summary(taoall)
```

```
##       obs              year          month             day
##  Min.   :     1   Min.   :80.0   Min.   : 1.000   Min.   : 1.00
##  1st Qu.: 44521   1st Qu.:92.0   1st Qu.: 4.000   1st Qu.: 8.00
##  Median : 89041   Median :94.0   Median : 6.000   Median :16.00
##  Mean   : 89041   Mean   :93.3   Mean   : 6.505   Mean   :15.72
##  3rd Qu.:133560   3rd Qu.:96.0   3rd Qu.:10.000   3rd Qu.:23.00
##  Max.   :178080   Max.   :98.0   Max.   :12.000   Max.   :31.00
##
##       date                 lat               lng            zon_wind
##  Min.   :1980-03-07   Min.   :-8.8100   Min.   :-180.00   Min.   :-12.400
##  1st Qu.:1992-01-16   1st Qu.:-2.0100   1st Qu.:-154.95   1st Qu.: -5.800
##  Median :1994-06-01   Median : 0.0100   Median :-111.26   Median : -4.000
##  Mean   :1993-10-19   Mean   : 0.4736   Mean   : -54.03   Mean   : -3.305
```

```
##  3rd Qu.:1996-06-17   3rd Qu.: 4.9800   3rd Qu.: 147.01   3rd Qu.: -1.400
##  Max.   :1998-06-23   Max.   : 9.0500   Max.   : 171.08   Max.   : 14.300
##                                                           NA's   :25163
##     mer_wind         humidity        air_temp        sub_temp
##  Min.   :-11.60   Min.   :45.40   Min.   :17.05   Min.   :17.35
##  1st Qu.: -1.70   1st Qu.:77.70   1st Qu.:26.06   1st Qu.:26.77
##  Median :  0.30   Median :81.20   Median :27.34   Median :28.29
##  Mean   :  0.25   Mean   :81.24   Mean   :26.89   Mean   :27.71
##  3rd Qu.:  2.30   3rd Qu.:84.80   3rd Qu.:28.18   3rd Qu.:29.23
##  Max.   : 13.00   Max.   :99.90   Max.   :31.66   Max.   :31.26
##  NA's   :25162    NA's   :65761   NA's   :18237   NA's   :17007
```

Some more basic Stats

```
###  Standard deviation
apply(na.omit(taoall[, 8:12]), 2, sd)
```

```
## zon_wind mer_wind humidity air_temp sub_temp
## 3.423210 3.021228 5.275265 1.674481 1.871993
```

```
### Inter Quartile Range
apply(na.omit(taoall[, 8:12]), 2, IQR)
```

```
## zon_wind mer_wind humidity air_temp sub_temp
##     4.40     4.10     7.10     1.86     2.17
```

```
### Skew
apply(na.omit(taoall[, 8:12]),2, skewness)
```

```
##    zon_wind    mer_wind    humidity    air_temp    sub_temp
##  0.97559003  0.02541101  0.10411561 -1.47285706 -1.45919519
```

```
### kurtosis
apply(na.omit(taoall[, 8:12]),2, kurtosis)
```

```
## zon_wind mer_wind humidity air_temp sub_temp
## 3.702570 2.758855 3.113213 5.775632 5.441958
```

Na's Exploration

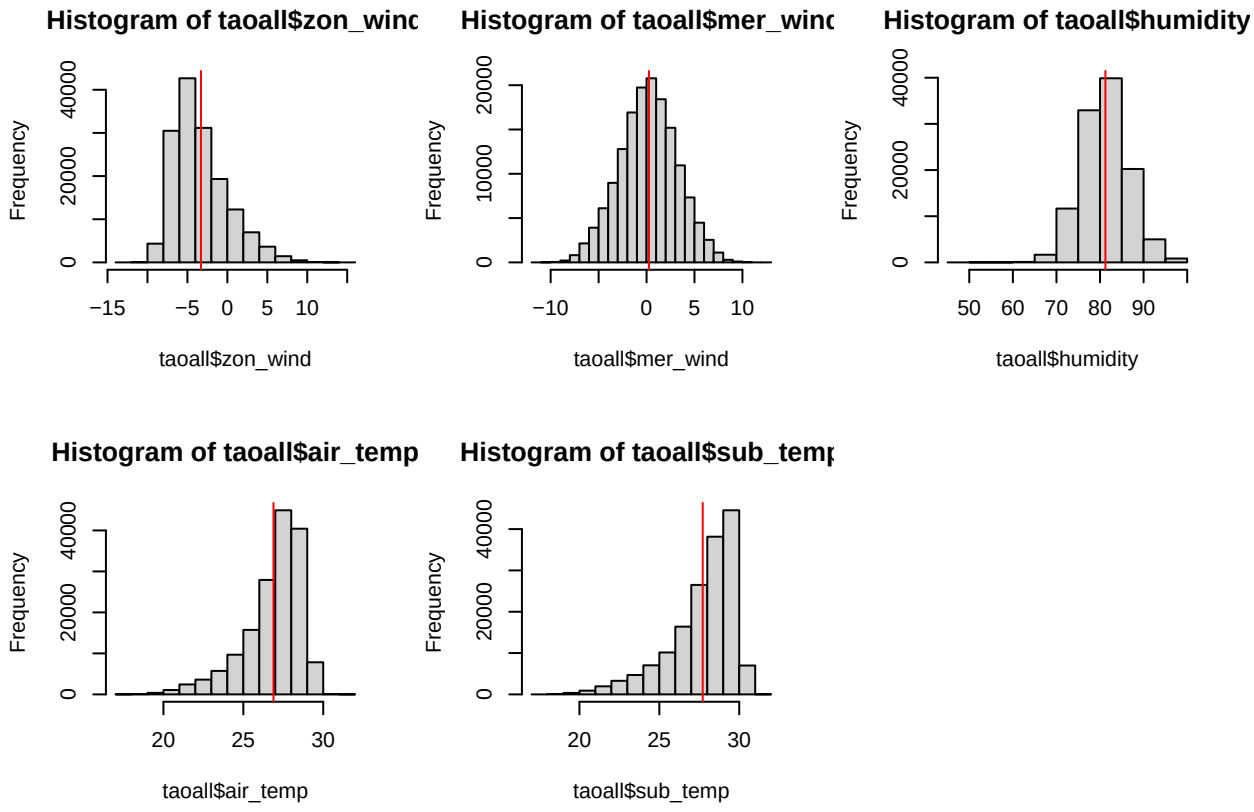
```
## Na's Exploration
a <- sum(is.na(taoall$zon_wind))/nrow(taoall) # NA's in percent for zonal wind
b <- sum(is.na(taoall$mer_wind))/nrow(taoall) # NA's in percent for meridional wind
c <- sum(is.na(taoall$humidity))/nrow(taoall) # NA's in percent for humidity
d <- sum(is.na(taoall$air_temp))/nrow(taoall) # NA's in percent for air temprature
e <- sum(is.na(taoall$sub_temp))/nrow(taoall) # NA's in percent for subsurface temprature
```

```
### Printing Data Frame

(data.frame(Variable = c("Zonal Wind", "Meridional Wind", "Humidity", "Air Temprature", "Subsurafe Tempr
            Sum = c(sum(is.na(taoall$zon_wind)),sum(is.na(taoall$mer_wind)),sum(is.na(taoall$humidity))
                    sum(is.na(taoall$sub_temp))),
            Percentage = c(a, b, c, d ,e)
              ))
```

```
##               Variable   Sum Percentage
## 1           Zonal Wind 25163 0.14130166
## 2      Meridional Wind 25162 0.14129605
## 3             Humidity 65761 0.36927785
## 4       Air Temprature 18237 0.10240903
## 5 Subsurafe Temprature 17007 0.09550202
```

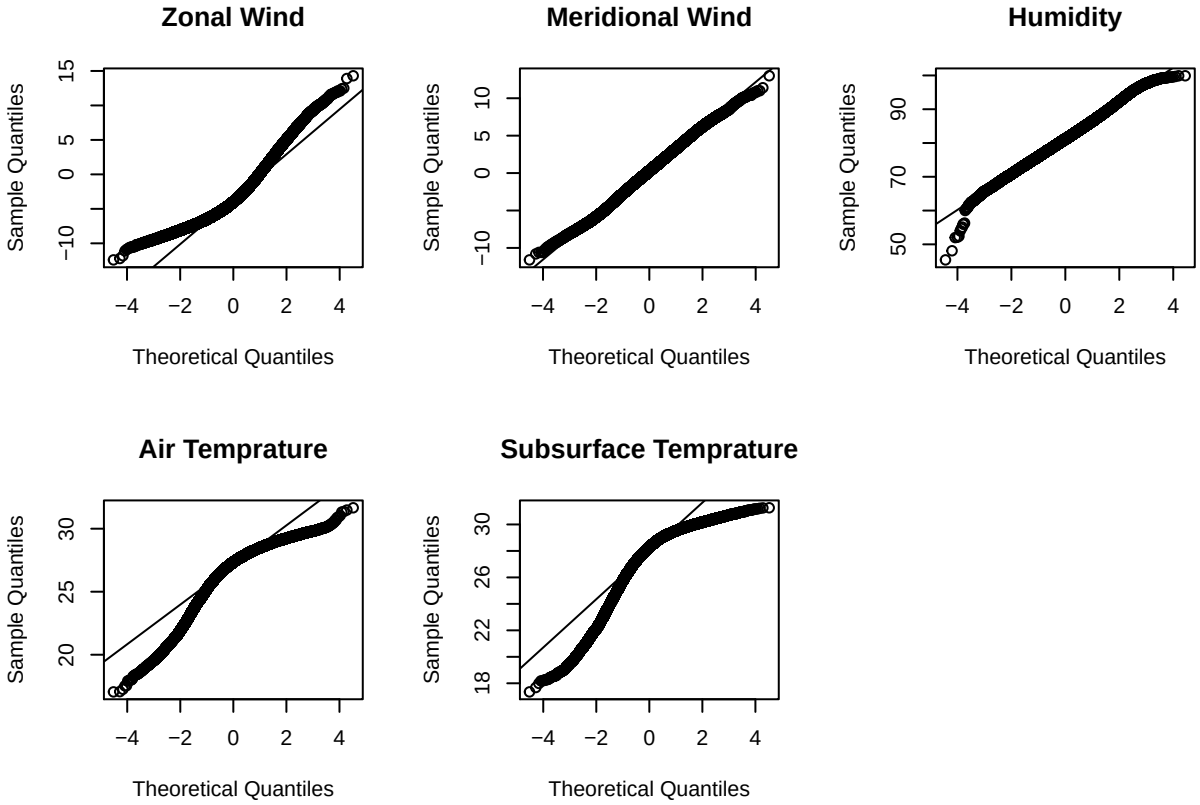
Univariate Analysis

```
################ Univariate Exploration ########################
par(mfrow = c(2,3))
hist(taoall$zon_wind);abline(v = mean(na.omit(taoall$zon_wind)), col = 'red')
hist(taoall$mer_wind);abline(v = mean(na.omit(taoall$mer_wind)), col = 'red')
hist(taoall$humidity);abline(v = mean(na.omit(taoall$humidity)), col = 'red')
hist(taoall$air_temp);abline(v = mean(na.omit(taoall$air_temp)), col = 'red')
hist(taoall$sub_temp);abline(v = mean(na.omit(taoall$sub_temp)), col = 'red')
```



Normality Visualisation

```
### Normality check
par(mfrow = c(2,3))
qqnorm(taoall$zon_wind, main = "Zonal Wind"); qqline(taoall$zon_wind)
qqnorm(taoall$mer_wind, main = "Meridional Wind"); qqline(taoall$mer_wind)
qqnorm(taoall$humidity, main = "Humidity"); qqline(taoall$humidity)
qqnorm(taoall$air_temp, main = "Air Temprature"); qqline(taoall$air_temp)
qqnorm(taoall$sub_temp, main = "Subsurface Temprature"); qqline(taoall$sub_temp)
```

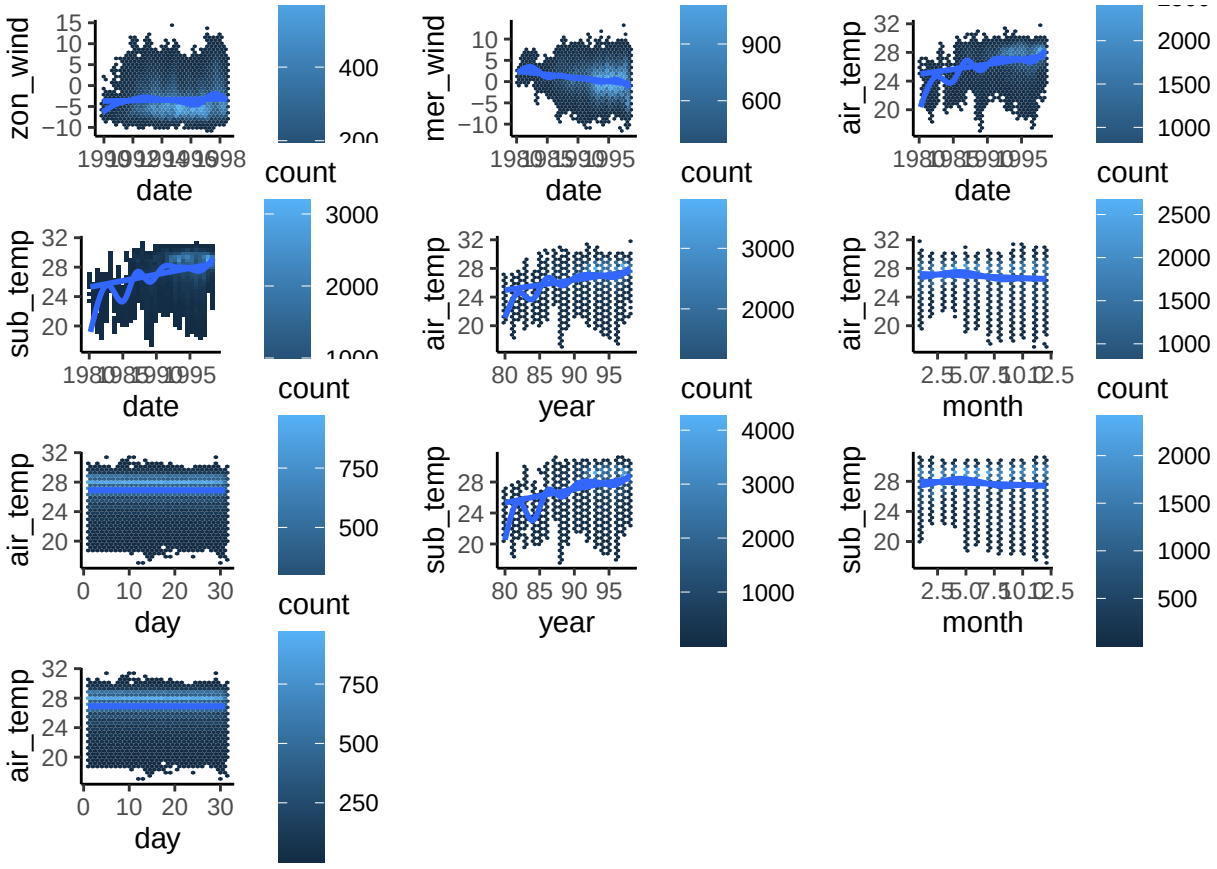


Advance Plot

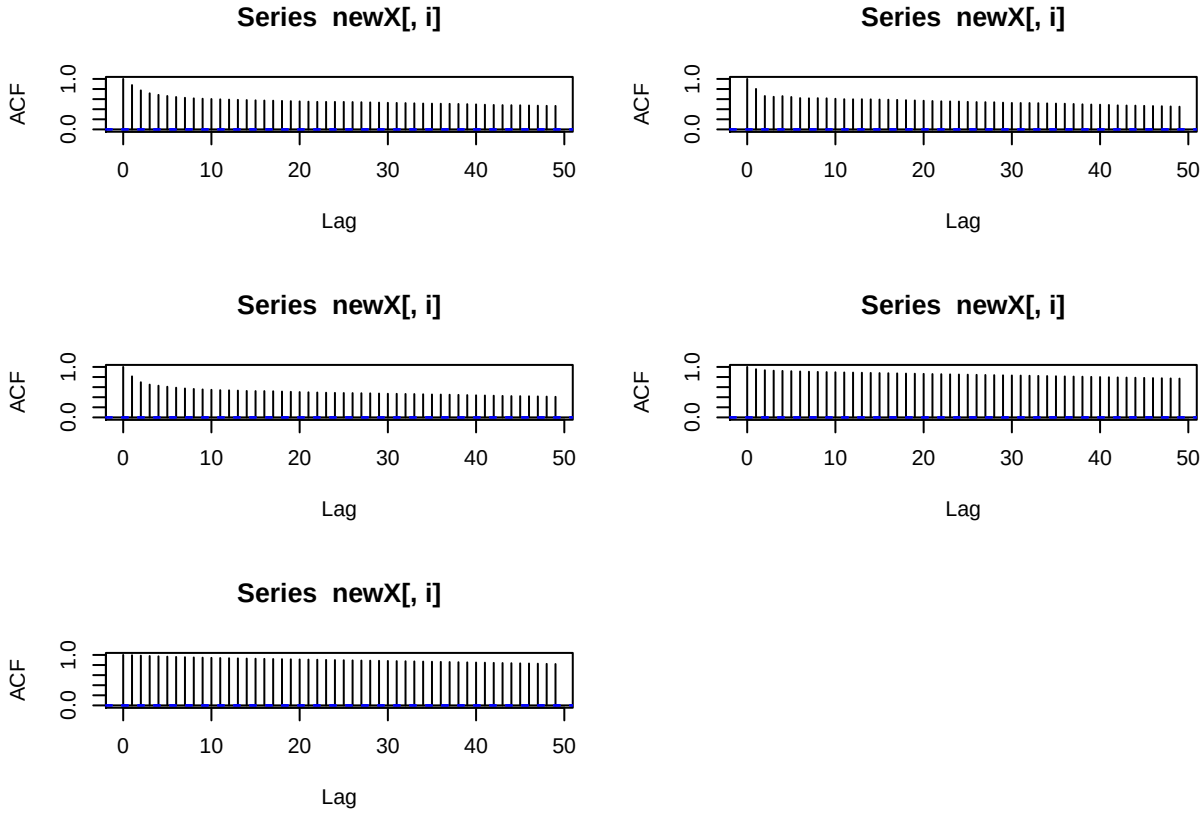
```
################ Visualization ################################


p1 <- ggplot(na.omit(taoall), aes(date, zon_wind))+
  geom_hex()+
  geom_smooth(method = "lm")+
  geom_smooth()+
  theme_classic()

p2 <- ggplot(taoall, aes(date, mer_wind))+
  geom_hex()+
```

```
  geom_smooth(method = "lm")+
  geom_smooth()+
  theme_classic()

p3 <- ggplot(taoall, aes(date, air_temp))+
  geom_hex()+
  geom_smooth(method = "lm")+
  geom_smooth()+
  theme_classic()


p4 <- ggplot(taoall, aes(date, sub_temp))+
  geom_bin2d( )+
  geom_smooth(method = "lm")+
  geom_smooth()+
  theme_classic()
##################### Air #####################################
p5 <-  ggplot(taoall, aes(year, air_temp))+
  geom_hex()+
  geom_smooth(method = "lm")+
  geom_smooth()+
  theme_classic()
p6 <-  ggplot(taoall, aes(month, air_temp))+
  geom_hex()+
  geom_smooth(method = "lm")+
  geom_smooth()+
  theme_classic()

p7 <-  ggplot(taoall, aes(day, air_temp))+
  geom_hex()+
  geom_smooth(method = "lm")+
  geom_smooth()+
  theme_classic()

##################### Sub Surface Air ##########################


p8 <-  ggplot(taoall, aes(year, sub_temp))+
  geom_hex()+
  geom_smooth(method = "lm")+
  geom_smooth()+
  theme_classic()



p9 <-  ggplot(taoall, aes(month, sub_temp))+
  geom_hex()+
  geom_smooth(method = "lm")+
  geom_smooth()+
  theme_classic()

p10 <-  ggplot(taoall, aes(day, air_temp))+
  geom_hex()+
```

```
  geom_smooth(method = "lm")+
  geom_smooth()+
  theme_classic()

gridExtra::grid.arrange(p1, p2, p3, p4, p5, p6, p7, p8 , p9 , p10)
```




```
par(mfrow = c(3,2))
ls <- apply(na.omit(taoall[,8:12]), 2, acf)
#par(mfrow = c(2,2))
#ls[1]; ls[2]; ls[3]; ls[4]
```



Discussion

This data set is downloaded from http://archive.ics.uci.edu/ml/datasets/El+Nino machine learning repositories.Primarily used by Dr. Di Cook, Department of Statistics, Lowa State University, data set collected by the international Tropical Ocean Global Atmosphere (TOGA) program using Tropical Atmosphere Ocean (TAO) array. Data set consist of 11 variables and 178080 observations, of which year, month and day are just by part of date column. Technically variable of interest in this case are zonal wind, meridional wind, humidity, air temperature and sub surface temperature. And obviously time phenomena makes it more interesting and complex to study. Though the variables of interest have unknown values in the data set, there is no technical explanation of which why this has happened.

Data has observation from 1980,Aug 03 to 1998,June 23 for daily interval. Zonal wind has minimum -12.4 and maximum 14.3, where (+) for east and (-) for west. Then meridional winds has minimum of -11.60 and maximum of 13.0, where (-) stands for south and (+) for north. On the other hand, temperature for air and sub surface seem to be quit identical in basic exploration have (17.05, 17.35) minimum, (31.66, 31.26) maximum and mean (26.89, 27.71) respectively. Even kurtosis for temperature has quite large and identical value of 5.44 and 5.77, which is excess kurtosis cases(>3). This data set is suffering from unknown data cells, which is quite high for humidity (36 percent). Clearly omitting these observation from the data set will cause lost of many information from the data. but a systematic analysis can give valuable insights. Meridional wind and humidity are seem to be normal in initial graphical analysis. when zonal wind is left skewed and air temperature and sub suraface temprature are again characterizing identical right skewed. Results for normality is being anticipated by normality graphs, Meridional wind and Humidity is still seem to be normal while other are seem following normal for cluster of observation around mean while tails are thick (non normal). From the advance analysis graph, trend for zonal wind seem to be stagnant over the period of time. Meridional wind is showing downward trend over the years but stagnant between months and days. In contrast to this, air and sub surface temperature are identical here too, increasing trend for years and constant for months and days. This is proved temperature for air and sub suraface air are highly correlated.

```
corrplot::corrplot(cor(na.omit(taoall[,8:12])),type = "lower")
```