**Description**

The data collection comprises oceanographic and meteorological observations of the surface taken from a series of buoys distributed in the equatorial Pacific. In order to grasp and forecast El Nino/Southern Oscillation (ENSO) cycles, the data is supposed to help.The dataset has following feature columns:

* Date
* Latitude
* Longitude
* Zonal winds
* Meridional winds
* Relative Humidity
* Air Temperature
* Sea Surface Temperature
* Subsurface Temperature

For certain sites, data from the buoys dates back to 1980. Rainfall, sun radiation, current levels, and subsurface temperatures were all measured in various areas. The data indicated that the bouys traveled around to different sites based on their latitude and longitude. The approximate position was within a degree of the latitude values. Despite this, the longitude measurements were frequently five degrees off from the approximate position. The data contains missing values. Because not all buoys can monitor currents, rainfall, or sun radiation, these numbers may be absent depending on the buoy. Because some buoys were commissioned sooner than others, the amount of data provided varies each buoy.

Initially, the data was downloaded from the uci ml repo website and loaded using R. Apart from data analysis of other dataset using Google Colab Notebooks, data analysis for this dataset was done using R studio.The .dat extension file downloaded was read using read.table function in R.

dataset<-read.table("tao-all2.dat", na.strings=".")

The first 10 instances of the data are shown below:

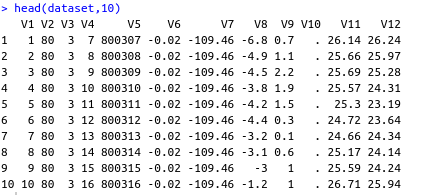


Fig: Screenshot showing first 10 instances of El Nino dataset

The columns of the dataset were renamed to appropriate names since column names were not appropriate in the dataset.



After renaming the columns of the dataset, the head of the dataset is given below with 10 as parameter.

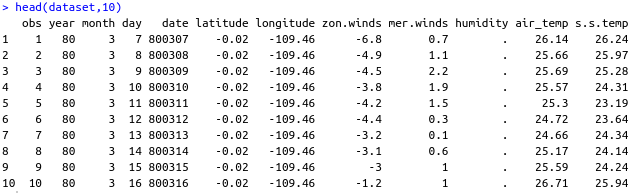


Fig: Screenshot showing first 10 instances of El Nino dataset

**Data Validation**

Data Validation was done by checking the number of rows and columns present in the dataset and comparing with the information of the dataset present in the UCI ml repo website. Data validation was done using stopifnot function in R as shown below:



**Data Cleaning**

Initially, we have to check whether the data contains any null values or not to lead to whether data cleaning is required or not.

row.has.na <- apply(dataset, 1, function(x){any(is.na(x))})

sum(row.has.na)

The result obtained was 84145 indicating there are manyy NAl values in the dataset. There are several missing values in the dataset. Thus, we apply further data cleaning steps in the dataset as shown below:

str(dataset)

dataset$longitude <- dataset$longitude + ifelse(dataset$longitude<0, 360, 0)

dataset\_without\_na <- na.omit(dataset)

After the data cleaning steps for the missing values, we convert non numeric columns into numeric using following apply function.



**Data Analysis**

After the data cleaning process, we perform analysis on the data. Initially, the summary of the dataset is generated using the summary function in R. We obtain the following results as a summary of the dataset.

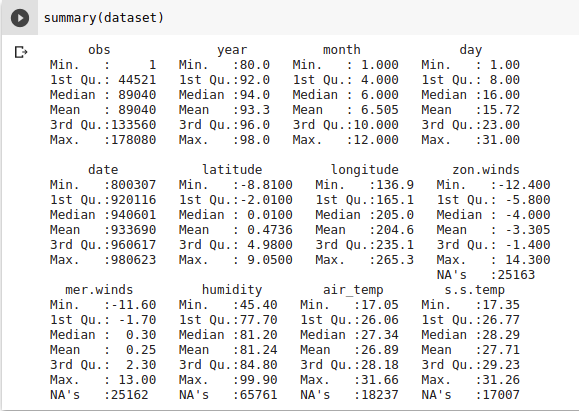
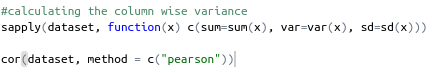
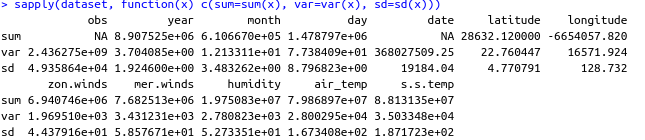


Fig: Screenshot showing the summary generated of the dataset

After generating such summary of the dataset, we calculate the column wise variance of the dataset.





We also calculate the pearson correlation coefficient of different features of the dataset using cor() function. The correlation result obtained are shown below:

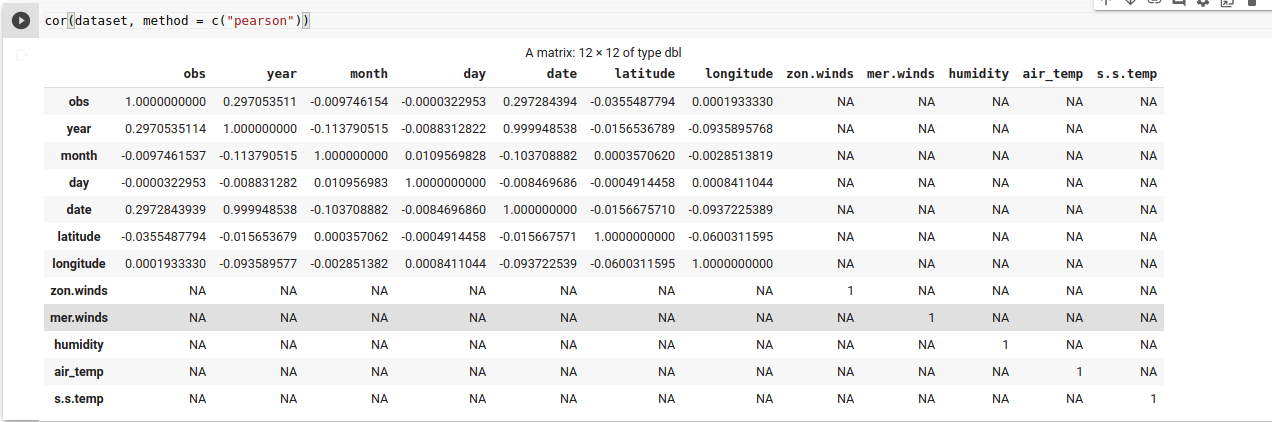
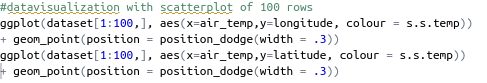


Fig: Screenshot of the correlation result obtained from the dataset

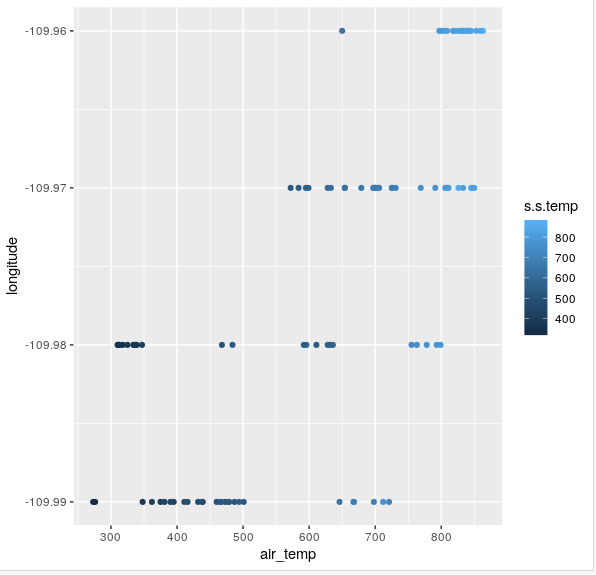
The correlation coefficient indicates the level of linear relationship between the two variables. The correlation coefficient close to -1 indicates strong negative linear relationship whereas close to +1 indicates strong positive linear relationship. From the above correlation plot, we can deduce that, “air\_temp” and “s\_s. Temp” are highly correlated with a strong positive linear relationship. In the wind data, both zonal and southern winds fluctuated from -10 m/s to 10 m/s. No linear relation was seen in the plot of the two wind variables. Also, against the other three meteorological results, the plots of and wind variable showed no linear relationships. In the tropical Pacific, relative humidity levels usually vary from 70 percent to 90 percent.

Both the temperature of the air and the temperature of the water surface ranged from 20 to 30 degrees Celsius. A positive linear relationship is seen in the plot of the two temperature variables. There are also equivalent plot designs for the two temperatures as one is plotted against time. There was no linear relation between the graphs of the other meteorological variables and the temperature variables.

After correlation plot, we performed data visualization using scatter plot considering only the first 100 instances of the data of the air temperature against longitude and lattitude.



Following scatter plot was obtained



Scatter plot of first 100 instances of data of longitude and air temperature

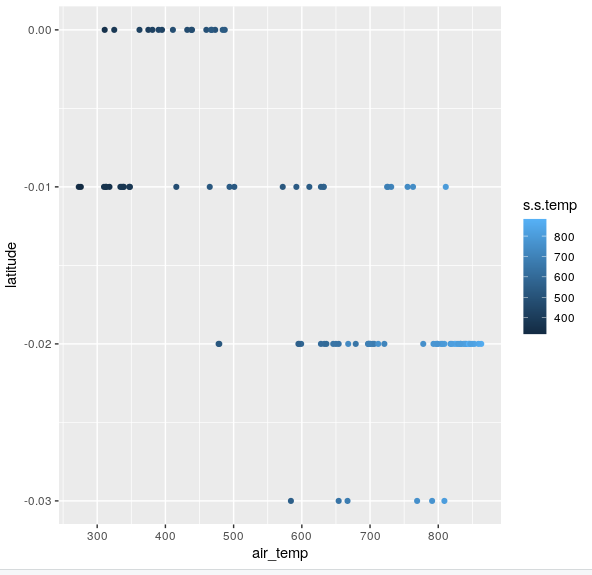


Fig: Screenshot showing the scatter plot of first 100 instances of data of latitude and air temperature

Variation of temperature with different longitudinal ranges

The study of variation of air temperature wit longitudinal ranges is perfomed with the help of scatter plot of air temperature with longitude in the data. The box plot of the air temperature is also developed to visualize the distribution of the air temerature data.

Box plot

Following code was used to develop the box plot of air temperature

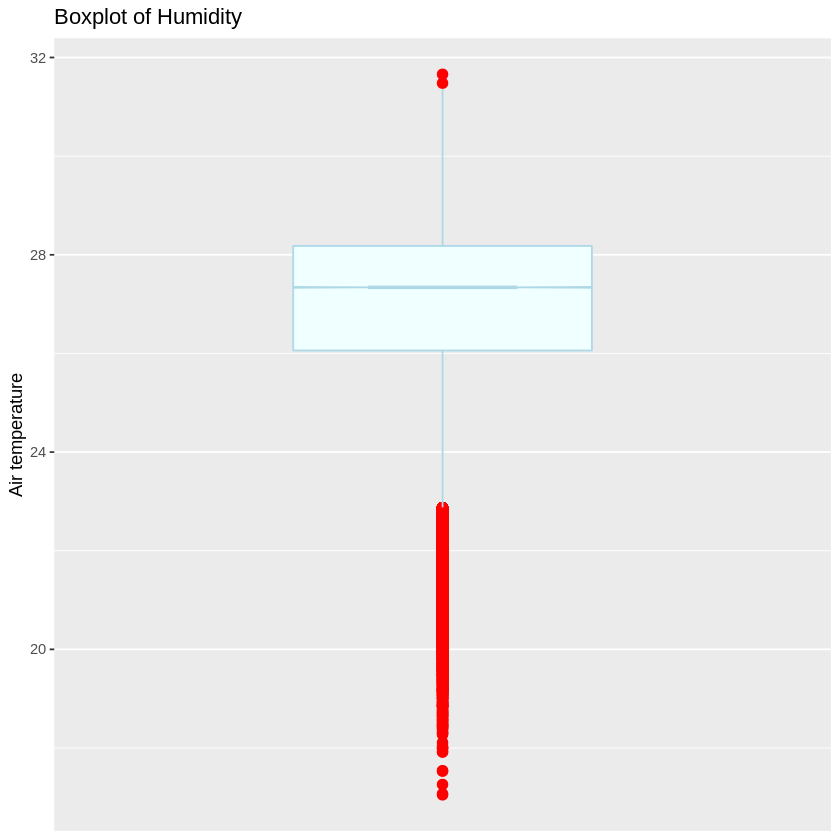
ggplot(dataset,aes(y=dataset$air\_temp)) +

geom\_boxplot(color="lightblue", notch=TRUE, fill="azure", outlier.color="red", outlier.shape=20,outlier.size=4) +

ylab("Air temperature") +

scale\_x\_discrete() +

ggtitle("Boxplot of Air Temperature")



The box plot pof the air temperature shows that there are many outliers in the dataset and the data distribution is highy skewed.

Scatter plot

Following code was used to develop the scatter plot of air temperature with respect to longitude

ggplot(dataset,aes(x=air\_temp,y=longitude)) +

geom\_point(color = "blue") +

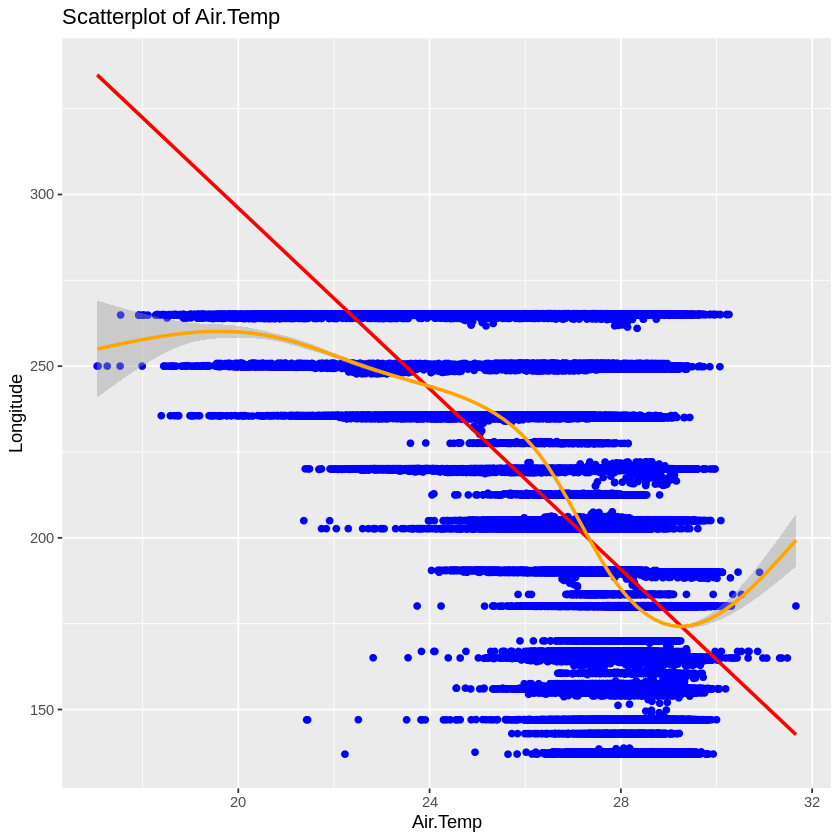
xlab("Air.Temp") +

ylab("Longitude") +

ggtitle("Scatterplot of Air.Temp") +

geom\_smooth(method = "lm",color = "red") +

geom\_smooth(color = "orange")



The scatterplot of the air temperature with the longitude shows that air temperature variation is very high in larger longitudinal ranges whereas the temperature variation is low in lower longitudinal ranges.