Faculty of Engineering and Technology

Assessment 2 Exploratory Data Analysis Report

**Acknowledgment**

The datasets were acquired from the UC Irvine Machine Learning Repository, and all parts of the report and data analysis were developed using the accessible meta data and information given in UC repository. GitHub and open-source code repositories are used to gain assistance with R programming. Open AI is only used to fix the errors to speed up the work. The link option datasets are as follows.

https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength

https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

**Student ID:** 1075535  
**Student Name:** Zainab Mahmood

**Module Name:** Introduction to Data Analytics  
**Module Code:** 7010DATSCI  
**Programme:** MSc Data Science

Contents

[1 Exploratory Data Analysis for EI Nino Dataset 1](#_Toc156057646)

[1.1 Description of Data 1](#_Toc156057647)

[1.2 Verifying the Different Aspects of Data 1](#_Toc156057648)

[1.2.1 Distribution and Relationship between the Variables 1](#_Toc156057649)

[1.2.2 Statistical Measure 1](#_Toc156057650)

[1.3 Visualization of Multivariate 2](#_Toc156057651)

[1.4 Plot’s Trends and Discussion 5](#_Toc156057652)

[1.5 Snip of Code 6](#_Toc156057653)

[2 Exploratory Data Analysis for Glass Identification Dataset 7](#_Toc156057654)

[2.1 Description of Dataset 7](#_Toc156057655)

[2.2 Verifying the Aspects of Data 8](#_Toc156057656)

[2.2.1 Identification of Outliers 8](#_Toc156057657)

[2.2.2 Distribution and Relationship between Variables 8](#_Toc156057658)

[2.2.3 Linear regression Model 8](#_Toc156057659)

[2.3 Visualization of Multi-Variate and Trends 9](#_Toc156057660)

[2.3.1 Oxides that Predict the Glass Type 9](#_Toc156057661)

[2.3.2 Oxides that Predict the Reflective Index (RI) of Glass 10](#_Toc156057662)

[2.4 Discussion 12](#_Toc156057663)

[2.5 Code Snip 13](#_Toc156057664)

[3 Exploratory Data Analysis for Diabetes Health Predication Dataset 15](#_Toc156057665)

[3.1 Description of Data 15](#_Toc156057666)

[3.2 Verifying the Aspects of Data 15](#_Toc156057667)

[3.2.1 Data Duplication 15](#_Toc156057668)

[3.2.2 Identification of Outliers 15](#_Toc156057669)

[3.2.3 Distribution and Relationship between the Variables 15](#_Toc156057670)

[3.3 Visualization of Multivariate 15](#_Toc156057671)

[3.4 Plot’s Trends and Discussion 19](#_Toc156057672)

[3.5 Code Snip 21](#_Toc156057673)

[4 Data Analysis Report Concrete Compressive Strength 22](#_Toc156057674)

[4.1 Description of Data 22](#_Toc156057675)

[4.2 Verifying the Aspects of Data 22](#_Toc156057676)

[4.2.1 Distribution and Relationship between the Variables 22](#_Toc156057677)

[4.2.2 Statistical Measure 22](#_Toc156057678)

[4.3 Visualization of Multivariate 23](#_Toc156057679)

[4.4 Plot’s Trends and Discussion 26](#_Toc156057680)

[4.5 Code Snip 28](#_Toc156057681)

# List of Figures

[Figure 1: Occurrence of El Nino Event with the Change in Buoys Position in Northern and Southern Hemispheres 2](#_Toc156057695)

[Figure 2: Impact of Buoys Position Change on Sea Surface and Air Temperature 3](#_Toc156057696)

[Figure 3: Timeline for variation in Sea Surface Temperatures from 1980 to 1998 in North and south Hemispheres 3](#_Toc156057697)

[Figure 4: Monthly Timeline of Sea Surface and Air Temperature for Year 1990 and 1997 in North and south Hemispheres 4](#_Toc156057698)

[Figure 5: Variations in Humidity, Zonal Wind, and Meridional Wind in relation to Sea Surface Temperature Over the 18 Years 4](#_Toc156057699)

[Figure 6 : Sodium (Na) and Aluminium (Al) Oxides as Predicator for Glass Type( Windows and Non-Windows) 9](#_Toc156057700)

[Figure 7: Oxides (chemicals) Average Amount Present in Glass Type(Windows and non-Windows) 10](#_Toc156057701)

[Figure 8: Correlation Matrix Heatmap to explain the correlation (Positive & Negative) of Refractive Index with other Oxides. 10](#_Toc156057702)

[Figure 9: Scatter Plots of Oxides to Show the Spread for Refracive Index and presence of Outliers 11](#_Toc156057703)

[Figure 10: Most Significant Oxides to Predict the Refractive Index Based on Linear Regression Model Results 12](#_Toc156057704)

[Figure 11: Correlation Heatmap to Show How Diabetes is correlated significantly with Other Exploratory Variables 16](#_Toc156057705)

[Figure 12: Diabetic People and their Body Mass Index Relationship with Disease and Habits 17](#_Toc156057706)

[Figure 13: Diabetic Distribution among Different Age Group of People in America 17](#_Toc156057707)

[Figure 14: Effect of Blood Pressure, Smoke, Cholesterol, Eating, Health and Walk on Diabetes 18](#_Toc156057708)

[Figure 15: Gender Distribution for Diabetes and different Health Factors 19](#_Toc156057709)

[Figure 16: Impact of Age and Cement Content on Concrete Comprehensive Strength (Mpa) 23](#_Toc156057710)

[Figure 17: Water and Cement Content Compsition Impact of Concrete Strength 24](#_Toc156057711)

[Figure 18: Cement and Fly Ash Content Composition and Their effect on Concrete Strength 24](#_Toc156057712)

[Figure 19: Comparison of Five content of Concrete Composition to Achieve the Concrete Strength (MPa) 25](#_Toc156057713)

[Figure 20: Impact of Superplasticizer on Water and Concrete Strength 25](#_Toc156057714)

# Exploratory Data Analysis for EI Nino Dataset

## Description of Data

The warm phase of the El Nino Southern Oscillation (ENSO) is known as El-Nino. Every three to seven years, there is a shift of the ocean and climate that lasts for nine to twelve months. El Nino, which is located in the Equatorial Pacific, affects weather patterns worldwide and is responsible for the unusually warm Pacific Ocean temperatures. There are twelve variables in the data. This report analyses the variations in "Air\_Temp," "Sea\_Surface\_Temp," and the impact of the position of the buoys on temperature. For certain places, data from buoys taken from 1980. Rainfall, sun exposure, current levels, and subsurface temperatures are among the other data that were collected at various sites. The data contains missing values. Since certain buoys are not equipped to measure solar radiation, rainfall, or currents, these numbers vary depending on the specific buoy. Because some buoys were put into service earlier than others, the quantity of data that is accessible also varies depending on the buoy.

## Verifying the Different Aspects of Data

### Distribution and Relationship between the Variables

Initially, the data is looked for non-numeric values in the dataset and converted them in so that analysis would continue. There are 15133 missing values in the data due to buoys as defined in first section, but they are not imputed because they are not necessary for this report. Datetime format was used for the "Date" column, and all relevant columns were renamed. A data summary was calculated to determine how the variables were distributed and box plots and descriptive statistics are used to find outliers. There are no outliers observed in the box plots of latitude, longitude, zonal winds, meridional winds, humidity, air temperature, and sea surface temperature. The position of buoys was change to 180 degree and a new variable hemisphere is created based on latitude, longitude calculation.

### Statistical Measure

To determine which factors are statistically significant for sea surface temperature, the linear regression model and correlation matrix are used in statistical testing. The p-value of the latitude, longitude, humidity, air temperature, zonal winds, and meridional winds is less than the significant value of 0.5, indicating their significance. According to the correlation matrix results, there is an positive correlation between latitude, longitude, zonal winds, and meridional winds. Latitude exhibits the maximum correlation value of 1, followed by longitude and zonal winds. Conversely, there is a negative correlation between sea surface temperature, air temperature, and humidity.

## Visualization of Multivariate

The section explains the reasons of the plotting of these variables in this part, along.

with annotations of the plot's trends. and section 4 provides more discussion on trends. The figure shows how frequently El Nino events occur in the dataset; a threshold of +0.5°C is used to identify anomalies related to El Nino. The El Nino event frequency in the northern and southern hemispheres is displayed using a map and bar chart.

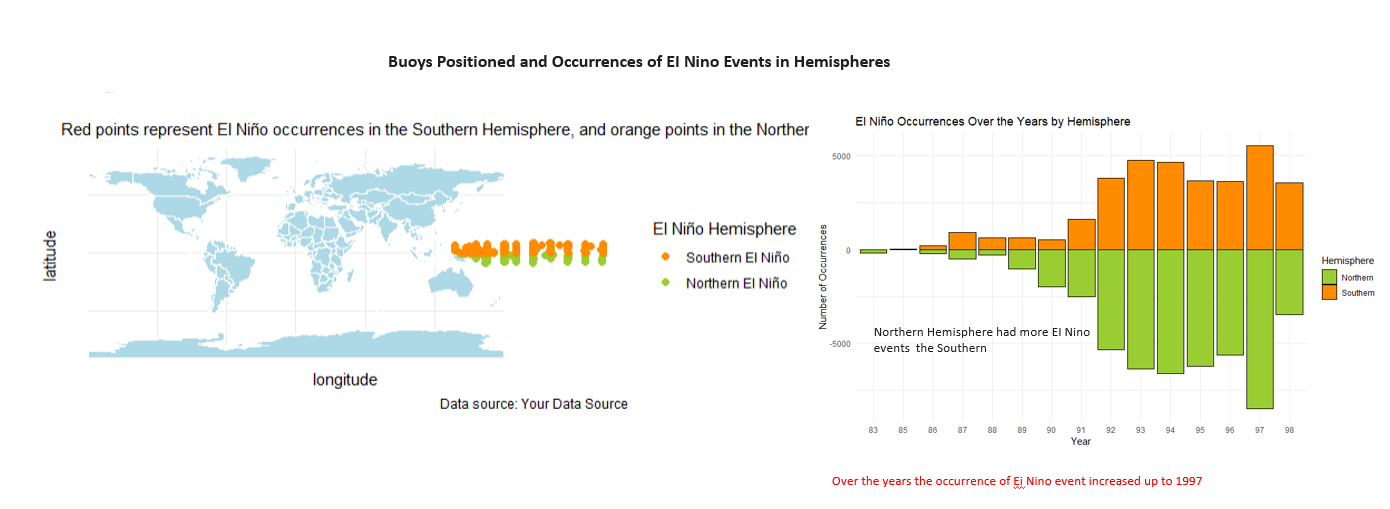


Figure 1: Occurrence of El Nino Event with the Change in Buoys Position in Northern and Southern Hemispheres

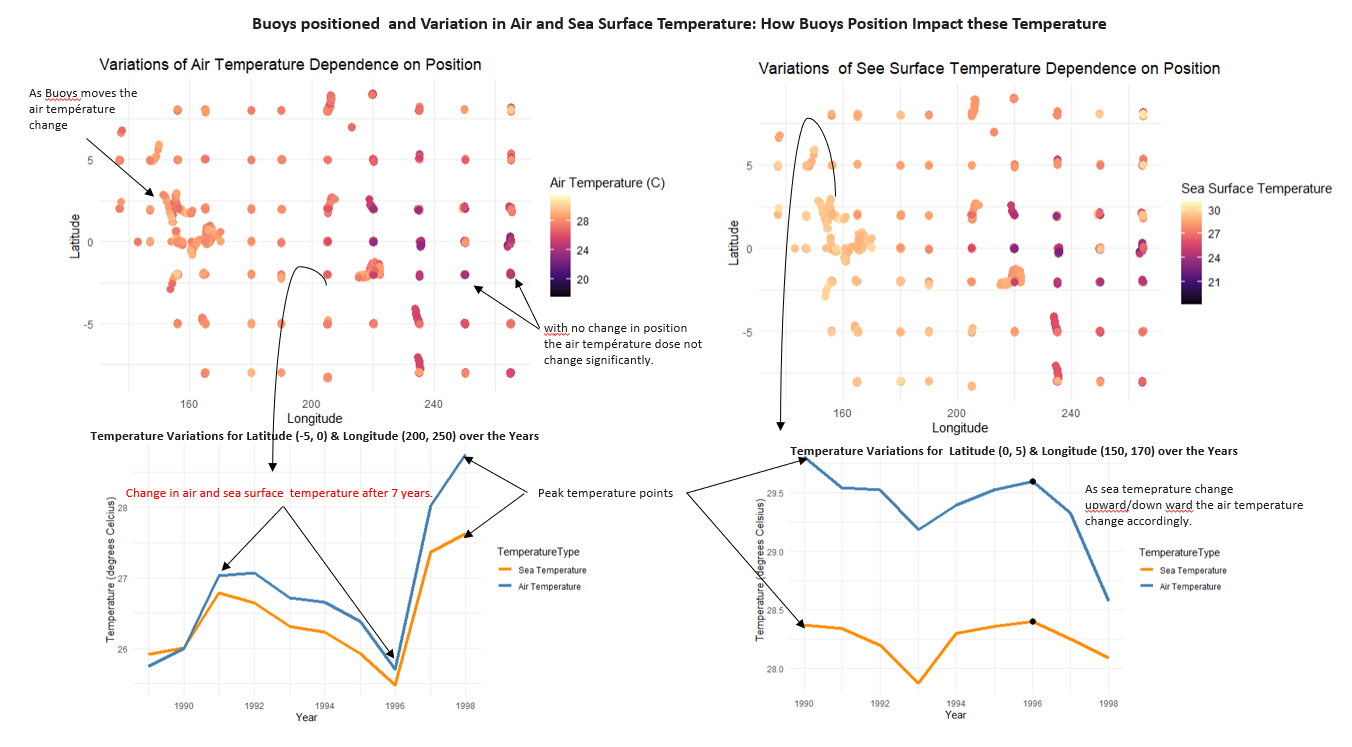
The air and sea surface temperatures are highly dependent on the buoy's position, which can vary dramatically. The impact of fluctuation on temperature is depicted in the image below, which has a plot grid. The last row grid provides more information about how the air and water temperatures vary at a given location.

Figure 2: Impact of Buoys Position Change on Sea Surface and Air Temperature

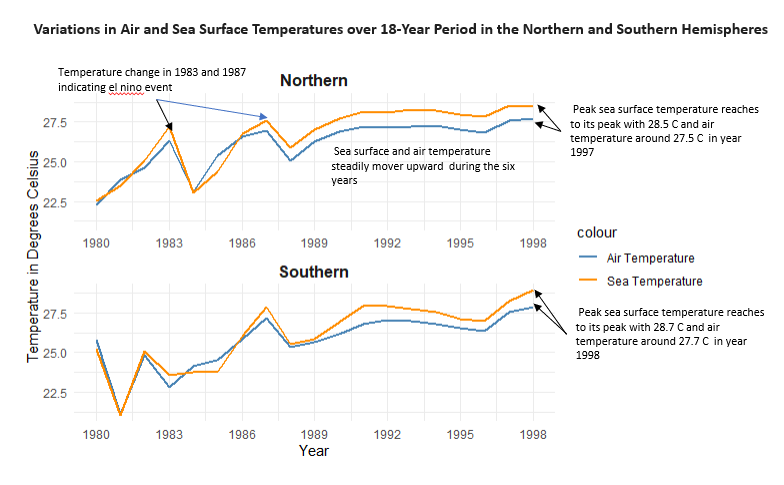
The chart displays the time series of variations in air and sea surface temperatures during a period of eighteen years, from 1980 to 1998. The figures show trends and sudden shifts. 

Figure 3: Timeline for variation in Sea Surface Temperatures from 1980 to 1998 in North and south Hemispheres

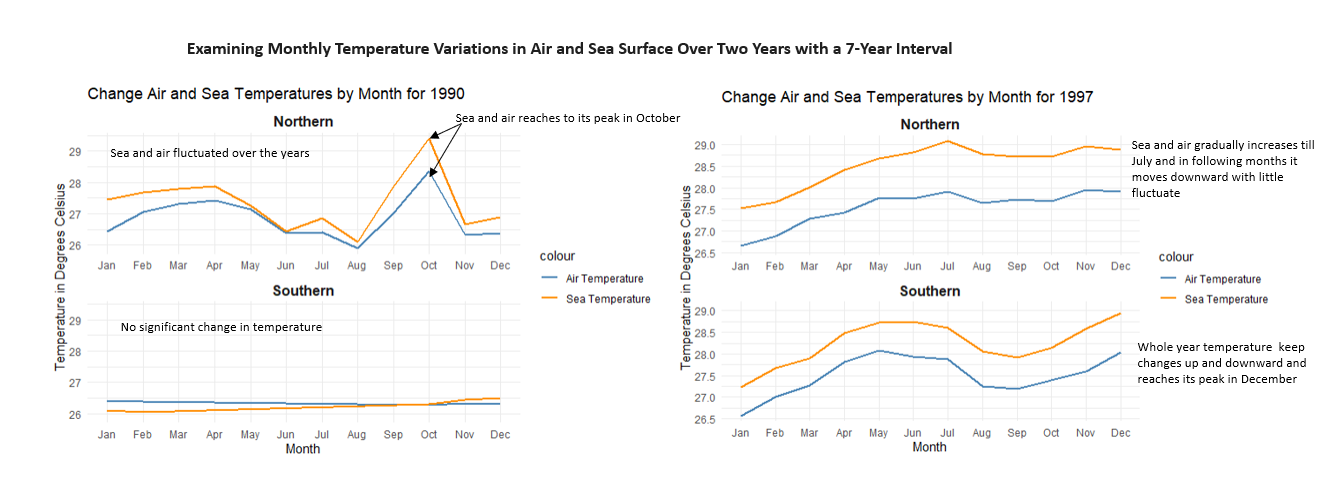
The monthly interval of temperature variation for the years 1990 and 1997 is illustrated. It is evident that the temperature fluctuates somewhat steadily in the southern hemisphere and quite a bit in the northern hemisphere. 

Figure 4: Monthly Timeline of Sea Surface and Air Temperature for Year 1990 and 1997 in North and south Hemispheres

The air temperature has a linear relationship with sea temperature, and the position of the buoys affects the sea surface temperature. The winds and air humidity are affected by these temperature changes. and is useful in weather and rainfall forecasting. The figure below illustrates how variations in sea surface temperature affect humidity and the zonal and meridional winds.

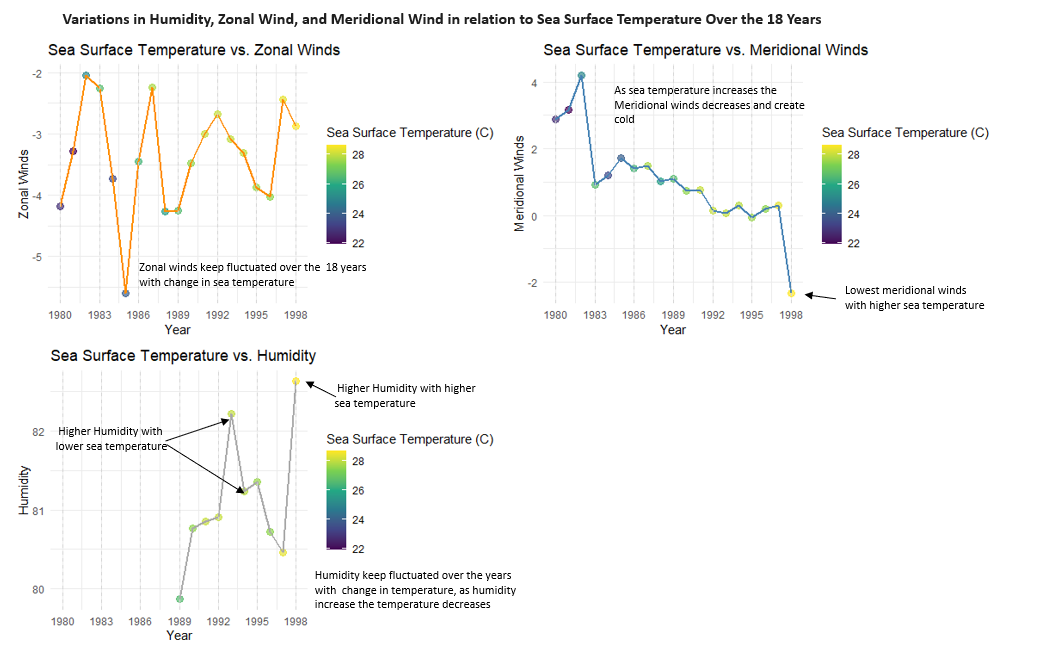


Figure 5: Variations in Humidity, Zonal Wind, and Meridional Wind in relation to Sea Surface Temperature Over the 18 Years

## Plot’s Trends and Discussion

From these plots following trends are observed on the relationships between these variables and buoys position.

* Sea surface and air temperature has strong positive correlation relationship, as sea surface temperature arises the air temperature arises.
* As humidity increases, air temperature tends to decrease, and vice versa.
* The temperature changes with latitude as the position of buoys changes
* The change in temperature and El-Nino events happened more in northern hemisphere then southern hemisphere.
* The negative correlation indicates that high wind speed, associated with cool sea surface temperature.

Plots of the relationships between the variables and buoy positions show different interesting patterns. First, there is a strong positive correlation between air and sea surface temperatures, meaning that air temperature tends to rise together with sea surface temperature. Furthermore, humidity and air temperature have an inverse relationship, which means that lower air temperatures are linked to higher humidity levels and vice versa. Temperature differences are largely dependent on geographic latitude; as buoys move along different latitudes, variances are evident. Compared to the southern hemisphere, the northern hemisphere sees more extreme variations in temperature as well as El Nino events. The sea surface temperature and wind speed are found to be negatively correlated, indicating that lower sea surface temperatures are related to higher wind speeds.

In both hemispheres, the air and sea surface temperatures have risen over time. In both hemispheres, the sea surface temperature peaked in 1997. Peak sea surface temperatures in the Northern and Southern Hemispheres, respectively, were around 28.5°C and 28.7°C. In 1998, the air temperature reached its highest point. Peak air temperature was 27.5°C in the North Hemisphere and 27.7°C in the South Hemisphere. The air temperature and the sea surface temperature are highly correlated. This is due to the fact that the ocean is a significant heat reservoir that affects the air's temperature above it.

## Snip of Code

df\_EI=read.table('tao-all2.txt', na.strings = '.', header= FALSE) #Load data file  
df\_EI$V5 = as.character(df\_EI$V5) # Change column into character   
df\_EI$V5 <- as.Date(df\_EI$V5, '%y%m%d') # changing date in date format  
colnames(df\_EI) = c("obs", "year", "month", "day", "date", "latitude", "longitude", "zon\_winds", "mer\_winds", "humidity", "air\_temp", "s\_s\_temp.")  
summary(df\_EI) # printing data statistics na\_count <- colSums(is.na(df\_EI))  
total\_na\_count <- sum(na\_count) # Na valuse in data set   
print(total\_na\_count) # Display the resultpar(mfrow = c(3, 3)) # setting plot area  
boxplot(df\_EI$obs)  
boxplot(df\_EI$latitude)  
boxplot(df\_EI$longitude)  
boxplot(df\_EI$zon\_winds)  
boxplot(df\_EI$mer\_winds)  
boxplot(df\_EI$humidity)  
boxplot(df\_EI$air\_temp)  
boxplot(df\_EI$s\_s\_temp.)

numeric\_columns <- sapply(df\_EI[, 6:ncol(df\_EI)], is.numeric) # Calculate correlation matrix for numeric columns only (excluding the first 5 columns)  
cor\_matrix <- cor(df\_EI[, 6:ncol(df\_EI)][, numeric\_columns], use = "complete.obs")  
library(corrplot)

corrplot(cor\_matrix, method = "color")

# Function to assign hemisphere based on latitude  
assign\_hemisphere <- function(latitude) {  
 if (latitude >= 0.0) {  
 return("Northern")  
 } else {  
 return("Southern")  
 }  
}  
# Create 'Hemisphere' column based on 'Latitude'  
df\_EI$Hemisphere <- sapply(df\_EI$latitude, assign\_hemisphere)

# Assuming 'df' is your data frame  
model <- lm(s\_s\_temp. ~ latitude+longitude+zon\_winds + mer\_winds + humidity + air\_temp, data =df\_EI)  
# Display the summary of the linear model  
summary(model)library(ggplot2)  
plot\_title <- "World Distribution of Buoys"  
ggplot(data = df\_EI, aes(x = longitude, y = latitude, color = Hemisphere)) +  
 geom\_point() +  
 labs(title = plot\_title) +  
 theme\_minimal()