

**1 INTRODUCTION**

In this report, we will be analyzing the El Nino dataset - containing oceanographic and surface meteorological readings taken from a series of buoys positioned throughout the equatorial Pacific - in order to understand and visualize the data. But most importantly, we want to be able to understand the story that this data is telling us. That is why visualizations are going to be imperative in this report. Through data visualizations, we are able to formulate and put together a story that can help extrapolate information from the data. The dataset can be found in the following URL’s:

1. <https://archive.ics.uci.edu/ml/datasets/El+Nino>
2. <https://www.kaggle.com/uciml/el-nino-dataset>

The UCI Machine Learning Repository cites the following as the dataset’s source:

1. *Original Owner: Pacific Marine Environmental Laboratory, National Oceanic and Atmospheric Administration, US Department of Commerce,* [*http://www.pmel.noaa.gov/*](http://www.pmel.noaa.gov/)
2. *Donor: Dr Di Cook, Department of Statistics, Iowa State University, dicook* ***'@'*** *iastate.edu,* [*http://www.public.iastate.edu/~dicook/*](http://www.public.iastate.edu/~dicook/)

El Nino is a climate pattern that explains the unusual warming of surface waters in the eastern tropical Pacific Ocean. The dataset pertinent to this pattern contains 12 variables that will help with our analysis. In order to develop the crucial visualizations needed for this report, we will use the R Programming Language in R Studio. As stated previously, the main objective of this report is to understand the dataset and be able to create a story through data visualizations. No more than four data visualizations will be created to help meet our objectives. Methodologies and tools that will help us achieve our objective include pre-downloaded basic R packages, data visualization techniques found in R, inferential reasoning based on findings, tidyverse package, naniar package, and lattice package.

Let us learn the brief approach behind the data collecting. The Tropical Atmosphere Ocean (TAO) array - developed by the international Tropical Ocean Global Atmosphere (TOGA) program - was used to collect the data. This TAO array is made up of 70 moored buoys that span the equatorial Pacific. These moorings were created by the National Oceanic and Atmospheric Administration’s (NOAA) Pacific Marine Environmental Laboratory (PMEL). Each single mooring measures a multitude of characteristics including air temperature, relative humidity, surface winds, sea surface temperatures, and subsurface temperatures 500 meters down. Some of the buoys measure currents, rainfall, and solar radiation. As stated, there are a multitude of variables in the dataset that correspond with the characteristics that the moorings collect (these will be explored later in the report). It is also important to note that all readings were taken at the same time of the day as stated by the UCI Machine Learning Repository.

**2 ANALYSIS**

**2.1 Importing Data**

I imported the data into R Studio by setting a working directory based on my Data Folder located in my QTM2623 Folder that can be found in Windows (C:) drive. After importing the data, I named the dataset “ElNino”.

**2.2 Data Dictionary**

I created a Data Dictionary because it is helpful to catalog and communicate the structure and content of the data to the viewer. It provides descriptive details about each variable found in the dataset. The table below is the Data Dictionary:

| **Variable Name** | **Variable Data-Type\*** | **Units** | **Description** |
| --- | --- | --- | --- |
| Observation | Integer | Number | The observation that the data comes from. |
| Year | Integer | Unit of Time | The year data was recorded |
| Month | Integer | Unit of Time | The month data was recorded |
| Day | Integer | Unit of Time | The day data was recorded |
| Date | Integer | Number | The full date data was recorded |
| Latitude | Numeric | Degrees and in Minutes | The angular distance of a place north or south of the earth's equator, or of a celestial object north or south of the celestial equator, usually expressed in degrees and minutes. |
| Longitude | Numeric | Degrees and in Minutes | The angular distance of a place east or west of the meridian at Greenwich, England, or west of the standard meridian of a celestial object, usually expressed in degrees and minutes. |
| Zonal Winds | Character | Meters per second | Winds that circulate at the same latitude, parallel to the equator (thermalize the atmosphere longitudinally). |
| Meridional Winds | Character | Meters per second | Winds along the local meridian. In a horizontal coordinate system fixed locally with the x axis directed eastward and the y axis northward, the meridional wind is positive if from the south, and negative if from the north. |
| Humidity | Character | Percentage | Relative humidity measures the amount of water the air can hold over a temperature range. Moreover, it measures how much moisture is currently in the air. |
| Air Temp | Character | Degree Celsius | The temperature that describes the kinetic energy, or energy of motion, of the gasses that make up air. |
| Sea Surface Temp | Character | Degree Celsius | The water temperature that is close to the ocean's surface. |

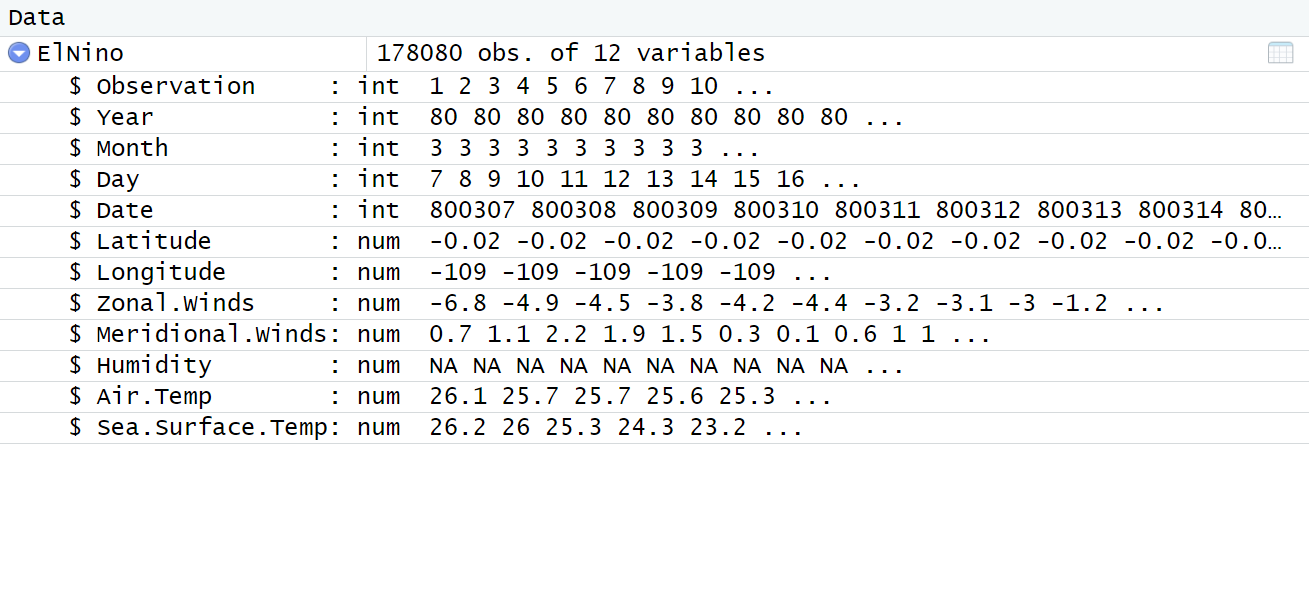
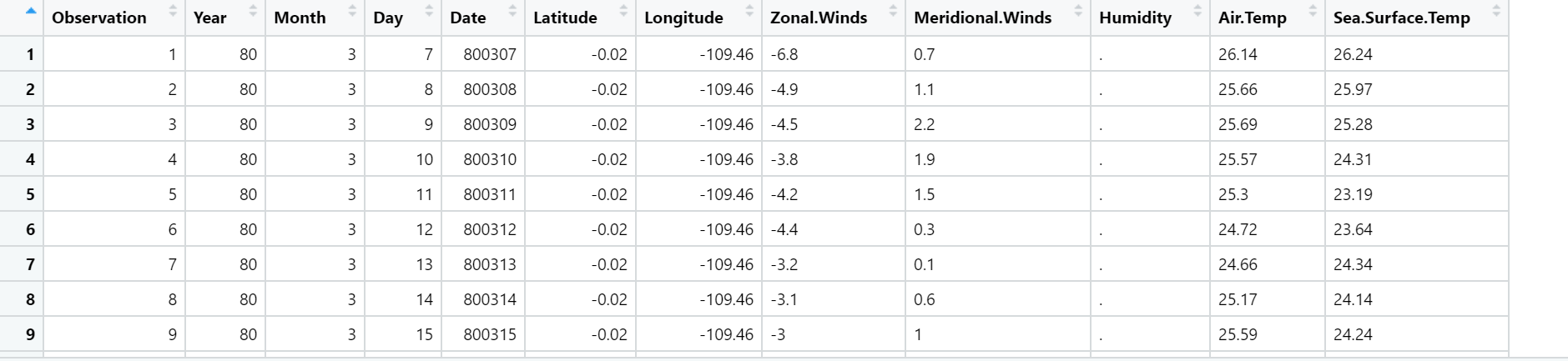
\*= data-type seen as how R reads the variable

**2.3 Data Structure**

Once the data was imported, I analyzed the dataset and its structure. Using the is.data.frame(df) function, I checked whether the dataset was indeed a dataframe. R Studio responded with “TRUE”, confirming that this dataset is indeed a dataframe. We can confirm that it is a dataframe because the column names are non-empty, row names are unique, data is stored as a data-type, and each column contains the same number of data items.

Using the str(df) function, I was able to get an overarching view of the data frames structure. As stated previously, there are 12 variables in the dataset, and a total of 178080 observations (this is the dimension of the dataframe). 5 of the variables are listed as an integer data-type, 2 are listed as a numeric data-type, and 5 are listed as a character data-type. There are missing values in the dataframe which are mentioned in the next section. Refer to Exhibit A for the first few lines of the dataframe and a view of the structure.

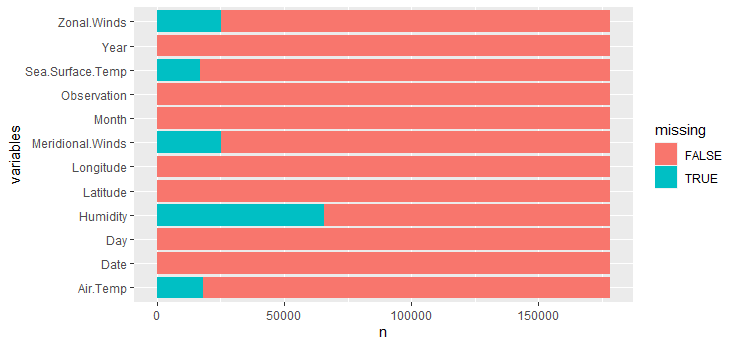
*Exhibit A*



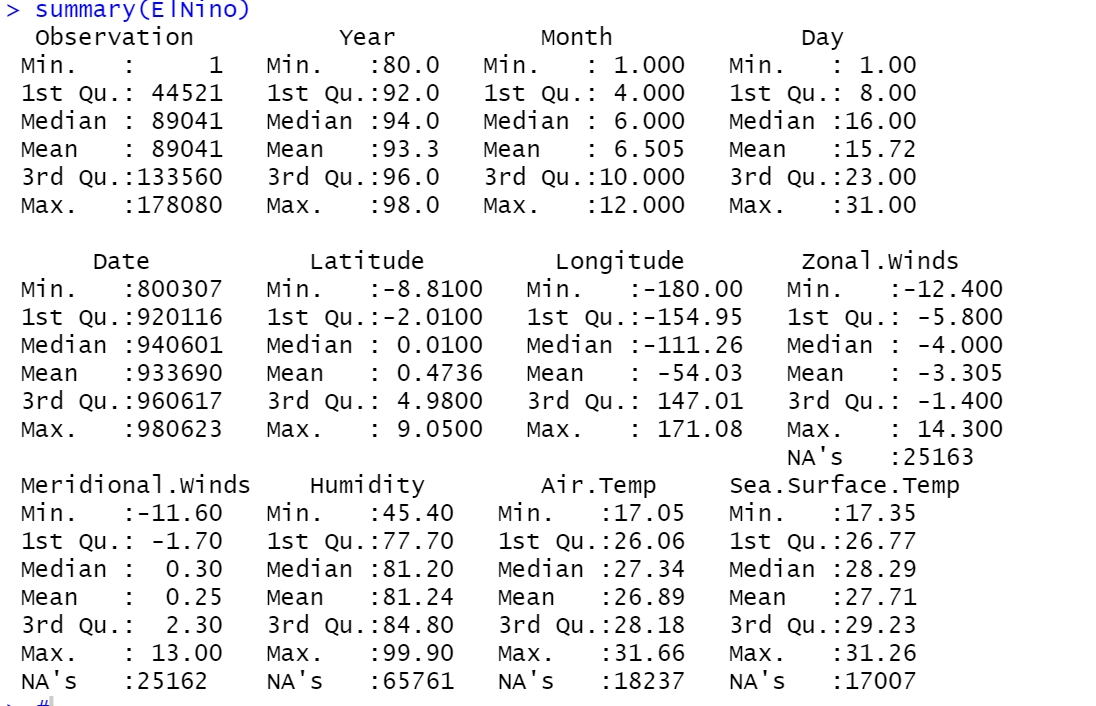
**2**.**4 Missing Values and Data Structure Manipulation**

There are many missing values in the dataframe. Using the summary(df) function I was able to see specifically which columns had missing values. The following variables had missing values: Zonal.Winds (with 25,163 missing values), Meridional.Winds (with 25,162 missing values), Humidity (with 65,761 missing values), Air.Temp (with 18,237 missing values), and Sea.Surface.Temp (with 17,007 missing values). There are a total of 151,330 missing values in the entire dataframe. I created a barchart using a naniar package as well in Exhibit B that shows the distribution of missing values in each of the variables. Humidity ranked highest with the most missing values. This is possibly due to the fact that capturing relative humidity can be very difficult because moisture sensors on the moors require a suitable environment in order to take readings. Exhibit C shows the statistical summary of the dataframe and includes the counts of missing values for each variable. Many of the variables also needed to be converted into a different data-type. I converted Zonal.Winds, Meridional.Winds, Humidity, Air.Temp, and Sea.Surface.Temp into the numeric data-type. The missing data was not required to be removed as the packages I installed removed them for me.

*Exhibit B*

**

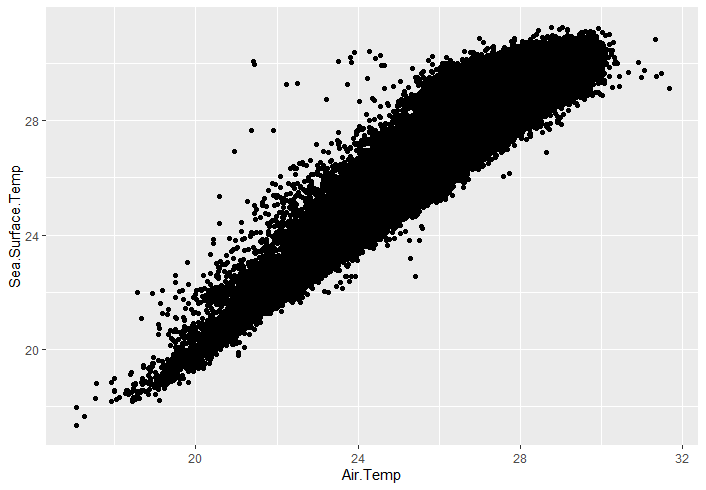
*Exhibit C*

**

**2.5 Sea.Surface.Temp and Air.Temp**

The first plot I decided to create was a scatterplot using ggplot found in the tidyverse package that showed the relationship between Sea.Surface.Temp and Air.Temp! As one can see there is a positive relatively-linear relationship between the two variables. What exactly can that tell us? Well, we know that Air Temperature is the kinetic energy of the gas molecules that make up the air, and that Sea Surface Temperature is the temperature that is close to the ocean surface. I wanted to know how similar air temperature trends with sea surface temperature as it increases. My initial hypothesis was that if the air temperature is hot, then the kinetic energy of the air that is close to the surface of the ocean will transfer to the liquid at the top of the ocean. Thus, the sea-surface will capture this moving kinetic energy and thus see an increase in temperature. My hypothesis was fairly correct based on the scatterplot seen in Exhibit D. We see that when air temperature increases, so does sea surface temperature. This tells us that the kinetic energy in the gas molecules that compose the air are easily able to transfer their energy to the sea surface temperature (as learned in Physics energy does not disappear, rather it is transferred). Though the liquid has a greater mass and can slow down the movement of the kinetic energy (thus accounting for discrepancies which are causing the scatterplot to not have a perfect positive linear relationship), we can see that there is an overall positive relationship between the two variables. When creating this function, it is important to note that ggplot removed 29124 rows of missing data. Though that is a large amount of data being missed out, it should not affect the scatterplot as there are almost 140,000 more rows of data that is being used to generate the scatterplot.

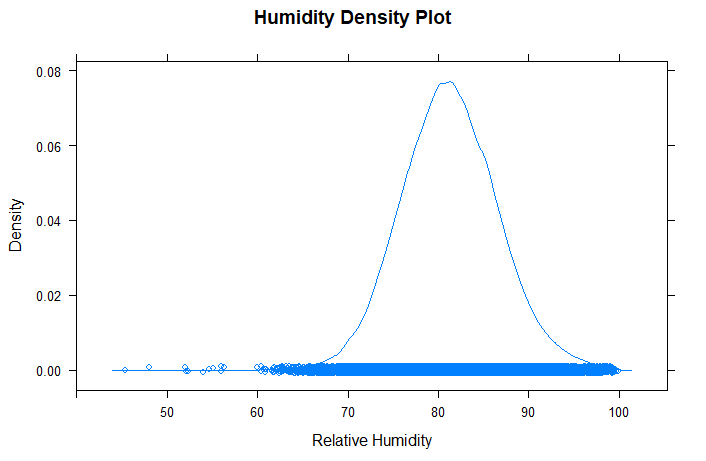
*Exhibit D*

****

**2.6 Humidity**

The second plot I created using the lattice package is a density plot of humidity. As you can see in Exhibit E, there is a slight negative skewness in this density plot. This density plot tells us that there are many outliers in the data on the left of the x-axis. Moreover, we can infer that the mode is probably greater than the median which is greater than the mean. Because the mean is less than the median and there is evidence of outliers, the plot tells us that there were difficulties with readings. As previously stated, the moorings on the buoys need to be in specific conditions in order for moisture content to be accurately captured (which is how relative humidity is captured). Because there are discrepancies in the data, we can infer that some of these moorings were not in proper conditions so they could not get accurate readings, causing a possible mishap of data missed on the left side of the x axis in the plot below. Moreover, the density plot, though slightly negatively skewed, has a relatively normal distribution where it is peaking, showing us that the majority of the readings had relative humidity above 80%. This variable had the most missing values as stated previously out of any other variable in the dataframe. In order to clean the data, and make the plot more reliable, I removed all missing values from the column.

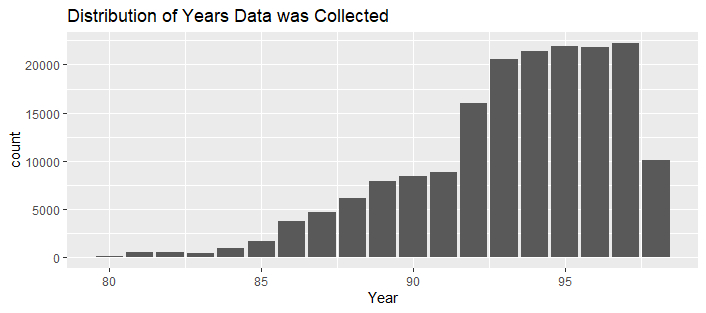
*Exhibit E*

****

**2.7 Year**

Using the ggplot package, I was able to generate a bar chart showing the count of data collected by the Year. Though the chart may seem simple at first glance, there is a lot of information to take away from it. It is evident that data started to get collected in larger quantities after the year 1990. 1997 seems to have the most collections of observations. What was happening before 1990? Why is there more data being collected after 1990? Well, in 1992 the United Nations Framework Convention on Climate Change was created to combat human involvement in climate change and to make the world a safer place. This was the basis for the Kyoto and Paris accords that set the basis for greenhouse gas emissions. Before then, it was very common for companies to engage in corruptive behavior and bypass certain regulations and engage in creating an abundance of greenhouse gas emissions. After this framework was created, it became clear that climate change and environmental monitoring would become a new norm. Thus, this can explain the possibility of more data being collected in the 1990s. Moreover, it makes sense that less data was being collected especially in the beginning of the 1980s as the moorings had just been set, and it probably took time for more buoys to be set over the years. The barchart in Exhibit F tells us a story of not only manipulating the law, but implementing global laws concerning climate change. As the El Nino experiences irregular and unusual warming, it was probably a suitable place for scientists to conduct readings and make observations about climate change.

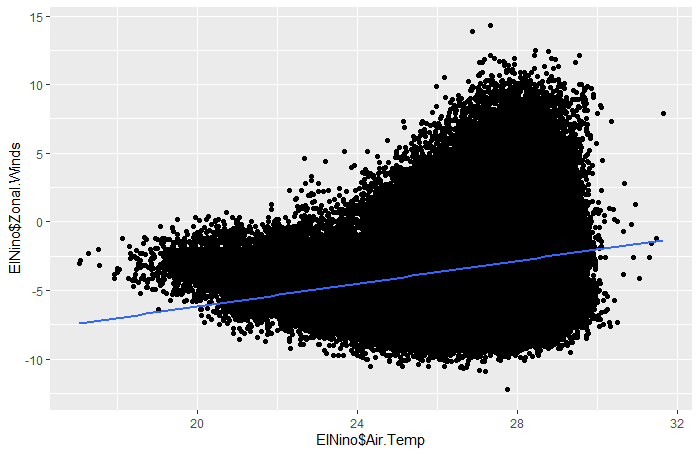
*Exhibit F*

****

**2.8 Air Temperature and Zonal Winds**

The last graphic I created was a scatterplot with a line of best fit between Air Temperature and Zonal Winds. I wanted to see the effect that the rising global air temperatures (due to climate change) would have on zonal winds along the equator. As predicted, I believed that because air temperature (defined as the kinetic energy of the gas molecules that compose the air) when increased, would transfer its energy into solar energy thus creating stronger winds. Though the line of best fit is not an accurate measure of the scatterplot, we can see a relatively exponential trend in the plot. This shows us that when air temperature increases, the energy is transferred into solar energy (which is what powers winds). I wanted to look at this plot because I wanted to see what attributes to rising winds. Though the winds could also become stronger when negative (-10 m/s has the same speed as 10 m/s), they become stronger in a positive direction (implying north of the equator). I found this very fascinating as north of the equator brings us areas such as Florida and Louisiana. The zonal winds along the equator have a tendency to travel northwards when air temperatures increase. Could this possibly explain the devastating hurricane that wrecked New Orleans? Maybe global warming may have propelled these devastating storms that plague Florida and Louisiana. Though I am not a scientist, this plot shows signs that zonal winds will increase when air temperature increases and will move northwards as well. It is also important to note that the ggplot function removed 38,015 rows containing missing values. This could possibly explain the slight discrepancies that skew a picture-perfect exponential relationship between air temperature and zonal winds.

*Exhibit G*



**3 CONCLUSION**

After analyzing and manipulating the data structure, I was able to generate four plots that told me a lot about the data being collected of El Nino. The big takeaways from my plots were that air temperature can affect zonal winds and sea surface temperatures due to transfer of energies, humidity can be hard to capture as it requires pristine conditions to measure relative humidity, and data collection was far more prevalent in the 90s compared to the 80s.

It is imperative to understand that the use of packages such as ggplot helped me remove missing values from columns that were included in my data visualization, thus some of the manipulation of the data was not needed. Moreover, it is also important to realize that there were many missing values in the dataset. However, they did not have an effect on my plots except for possibly my last graphic.

Using the methodologies and tools listed in my introduction, I was able to complete the objective of being able to tell a story through data visualizations and found some fascinating stories that coincided with true historical events. Though the El Nino data is aimed to find the unusual warming of the equatorial Pacific, there are many variables that have a story behind them that can explain global trends as seen in my third graphic.