Project Assignment 1: Data Wrangling

Business Problem

We retrieved a time series dataset that contains oceanographic and surface meteorological readings taken from a series of buoys positioned throughout the equatorial Pacific. Each buoy recorded important data points such as date, latitude, longitude, zonal winds, meridional winds, relative humidity, air temperature, sea surface temperature, and subsurface temperatures. Data from some buoys start from as early as 1980. This data can be used to predict weather conditions throughout the world, such as droughts and floods. In particular, we can use this data to understand and predict climate change patterns in the equatorial Pacific.

Data Quality

We encountered a lot of missing data in this dataset mostly with a **Missing not at Random (MNAR)** pattern. Firstly, the buoys were commissioned at varying dates; hence the amount of data collected depends on the year. The range of the year at which the buoys were commissioned is 18 years with 1980 being the earliest and 1998 being the latest. In addition, the amount of data available is also dependent on the buoys' reliability. Weather conditions, such as currents, rainfall, and solar radiation, as well as the machine's mechanical issues hindered the buoys' ability to collect high-quality data consistently. Many buoys were unable to record variables because of these external reasons, especially for measuring humidity. Therefore, we choose to drop the humidity variable because of the large amount of missing data.

The buoys were not clearly labeled in the dataset, though data collected by each buoy was clustered and can be identified using the date column. Usually, data collected by a buoy followed a time order from the earlier date to the latest. Once the date column jumps from a late date to an early date, we know it is the start of data collected by a new buoy.

This dataset was originally obtained from NOAA Pacific Marine Environmental Laboratory, distributed by the UCI Machine Learning repository on Kaggle. Therefore, we strongly believe that the data source is of high quality and that this dataset is trustworthy.

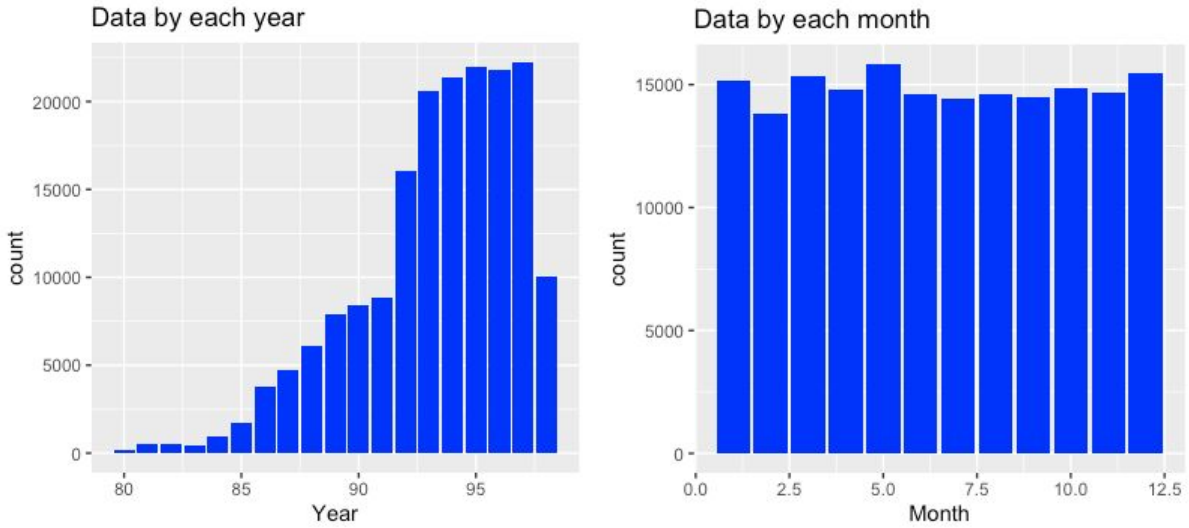
Use of Trifacta Wrangler

Trifacta detected that 14% of observations in "Zonal Winds" and "Meridional Winds ", 37% of observations in "Humidity", as well as 10% of observations in "Air Temp" and "Sea Surface Temp" was missing. The dataset uses "." to represent these missing values. We applied the "Replace" function to replace them with "null".

Final Data Quality

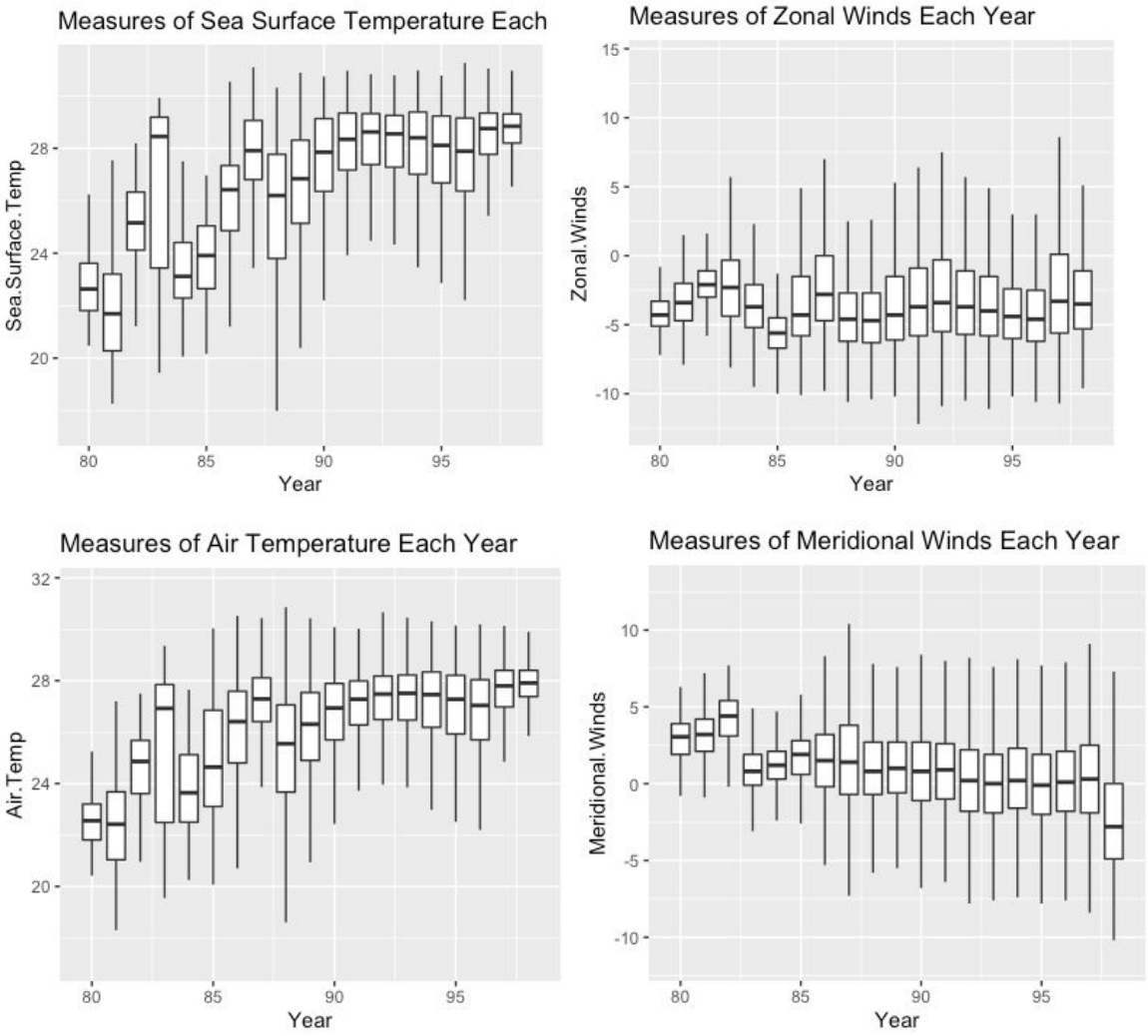
After we wrangled our data, we generated graphs below by R, showcasing our final data quality. There are also our reflections on the final data quality for future usage.

We observe that the data were not collected evenly throughout the years. A significant increase in data collection occurred from 1980 to 1991. The measures from 1992 to1997 are most abundant and evenly distributed. Therefore, in future data analysis, we may want to focus on the 1992 to 1997 data, for better standardization.



However, the months measured are distributed evenly. Therefore, in future analysis, we can use this data for analyzing year-long changes.

There are four key measures we will want to focus on for the later data analysis grouped by years after we have outliers removed. We may need to exclude “Humidity” because 37% of the “Humidity” data were missing.



Descriptive Statistics

It's challenging to create descriptive graphs for the key variables because the buoys were not stationary. The data provider indicated that these buoys traveled near the equator, which means that their north-south locations were relatively static. We programmed with R to show the data's geographic characteristics:

With the information of the north-south static location, we selected a latitude range that data were frequently collected, for example, the latitude = [160, 180]. Within this range, over 70% of data were collected between latitude 164.5 and 165.5. Thereby, we used Trifacta to further filter the numbers to the range and summarized the numbers in the graph:

Disclaimer: This graph only shows the frequency of each value, but not the actual distribution of the value. For the actual distribution of values, please refer to the Final Data Quality section.

Summary: Team Reflections on Trifacta Wrangler

Analytical modeling process

Trifacta Wrangler assisted our team in understanding the multiple dimensions of one buoy as a data collecting point: variations of locations and date. The further steps may include labeling buoys and developing research on individual buoys. Our team could track individual buoys' traveling path and the data that were collected at different locations. With a fixed time and the same path, we could compare the data collected by the different buoys.

After the data cleaning, we could apply the time series analysis in learning patterns and making weather predictions using this data

Combination with other software tools

Since it's hard to label buoys using the date column in Trifacta, we wrote a small program in Python to do that. Trifacta is easy to use; the interface is user-friendly, and non-tech people can master the tool easily. However, sometimes data cleaning requires slightly more complicated algorithms, and we can use more sophisticated tools to accomplish that. In general, Trifacta serves the first step cleaning purpose well and prepares data for further cleaning and analysis.

In addition, once we are done with data cleaning in Trifacta, we can export the cleaned dataset easily and have them ready for next step work.

Collaborative workflow within or across organizational boundaries

Trifacta can be smoothly embedded in a collaborative workflow for the following reasons.

The tool is easy to master. The interface is user-friendly and does not require much technical knowledge in advance. However, it also enables the feature of building up users' own data cleaning functions, so advanced data technicians can build more complicated data cleaning methods in Trifacta. In general, both business and technical people can use this tool to clean data.

Moreover, the tool records all cleaning steps and supports commenting functionality, so it is easy to comment on cleaning steps and make edits. Users can also retrieve the steps during a presentation to justify their cleaning efforts. Lastly, the tool supports future replication of cleaning procedures on a similar dataset. This functionality saves time for business if the business has similar datasets coming periodically.

Finally, collaborators on the workflow can access the tool simultaneously. If several technicians representing different organizations would like to work on the dataset at the same time, they can easily do so by logging into the platform. Moreover, in the case while there are several organizations working on the data analysis pipeline, after one party wraps up data cleaning in Trifacta, it can easily pass the task onto the next team, since both the procedure and results are easy to interpret. In general, using a simple, standard, and unified tool to do data cleaning makes it more smooth for different organizations to collaborate on the data project either simultaneously or vertically.