MA615 HW4

Jin Wen Lin

2024-09-27

Part (b)

Here is the process of converting missing or unknown values into NAs.

```
buoy_data <- read.csv("buoy_data.csv")
# Data with replaced NA
cleaned_buoy <- buoy_data
cleaned_buoy[] <- lapply(buoy_data, function(x){replace(x, x == 99.00 | x == 999.0, NA)})
```

Converting missing values into NAs is beneficial for us to clean the data and do calculations. R has specific functions that deals with NA values such as the na.omit(), and is.na() etc. Hence, converting missing or unknown values into NAs are efficient in data cleaning by making the data consistent and easy in dealing with calculations such as calculating means or averages etc. However, noticed that there are different kinds of placeholders in this data, such as 99, 999, 99.00, and 999.0 etc. They might have some other meanings behind and if we convert them into NAs, the information behind them seems to be lost so this is a reason that we might need to consider about. By looking at the data, we can see that two variables visibility and TIDE (water level in feet) are filled with NAs. Before converting them into NAs, they all filled with the placeholder of 99. In addition, some of the years did not record the air temperature and replacing them as 999. For variables like WVHT, DPD, APD, and MWD, they seems to record the values with specific period during the recent years. Sometimes record them in half an hour, and sometimes record them for every hour.

Part (c)

Some factors that might related to climate changes are wind speed, sea surface temperature, and sea level pressure. Calculations and visualizations of the listed factors will be presented in the following.

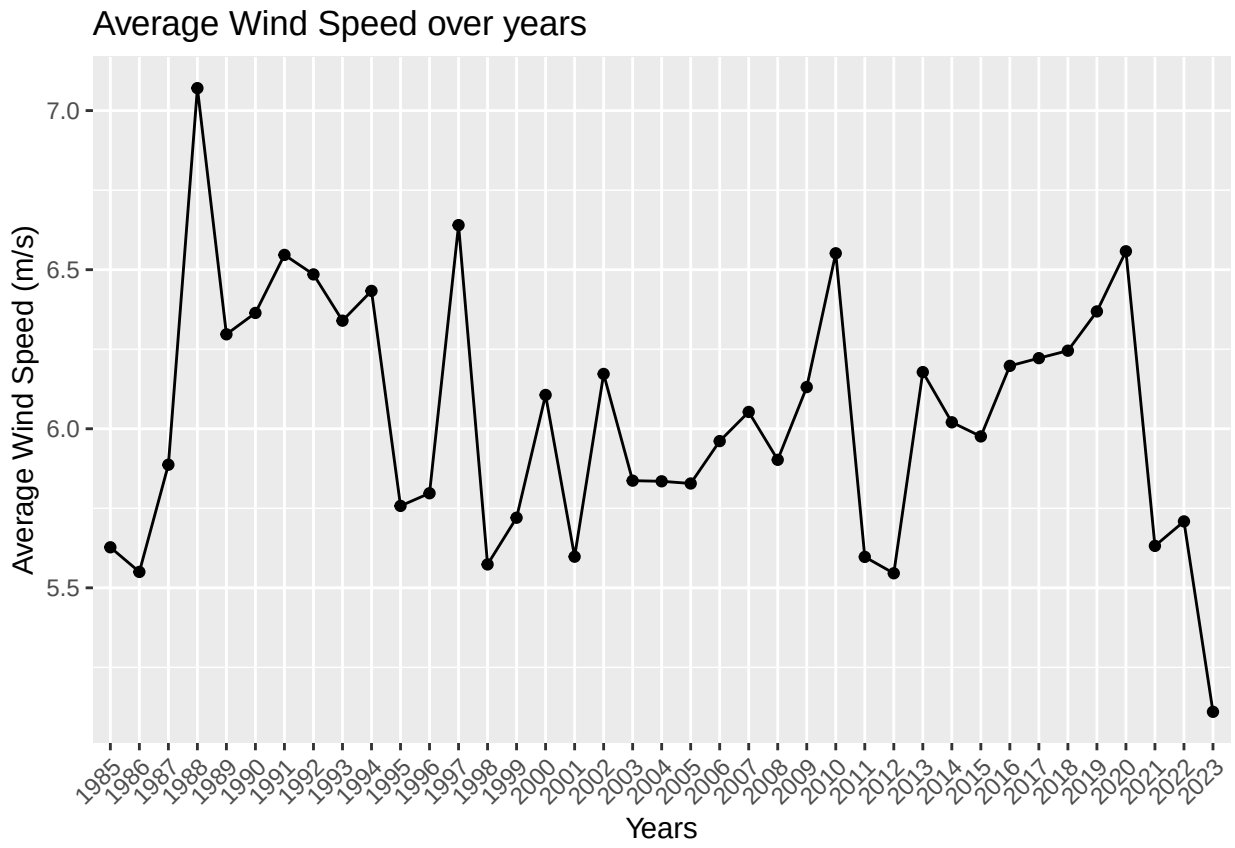
Wind Speed (m/s) First, calculate the average wind speed for each year. Here are the first few rows about the average of wind speed in each year.

```
# create a new data for wind speed average (AVWS)
buoy <- data.frame(Year = 1985:2023)
buoy <- cleaned_buoy %>% group_by(Year) %>% summarise(AVWS = mean(WSPD, na.rm = TRUE))
ws <- buoy %>% select(Year, AVWS)
head(ws)
```

```
## # A tibble: 6 x 2
##    Year  AVWS
##   <int> <dbl>
## 1  1985  5.63
```

```
## 2  1986  5.55
## 3  1987  5.89
## 4  1988  7.07
## 5  1989  6.30
## 6  1990  6.36
```

Below is the graph about the Average wind speed over years.

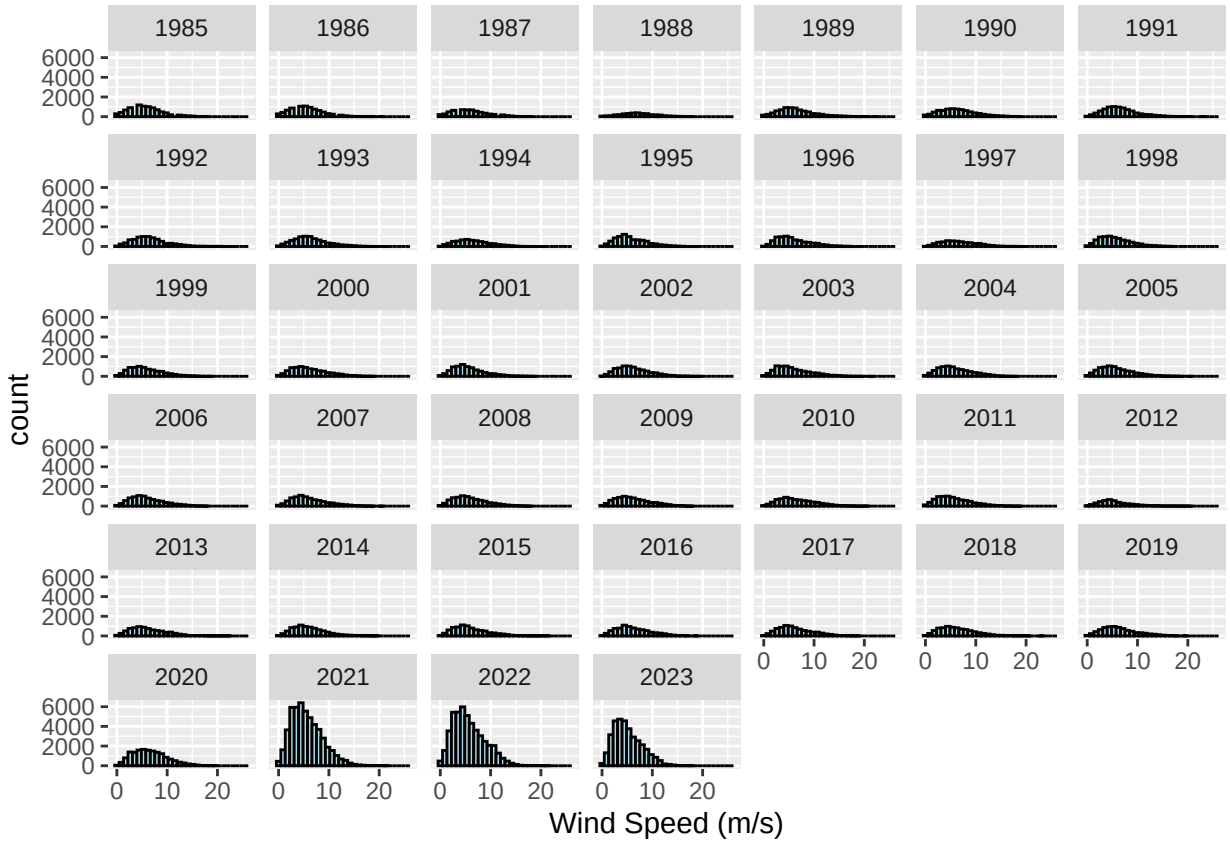


First by looking the plot of average wind speed for each year, it is hard to capture an overall trend since there are lots of peaks and troughs. However, we can see that there seems to exit a periodic pattern here. By looking at the peaks, we can see that they occurred at 1988, 1997, 2010, and 2020. Hence, there might exist some weather patterns for a period of 10 years.

```
# histogram
ggplot(cleaned_buoy, aes(WSPD)) +
  geom_histogram(fill = "lightblue", color = "black") +
  facet_wrap(~Year) + labs(x = "Wind Speed (m/s)")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 33193 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



Now, by looking at the histograms of wind speed for each year, we can see that the distributions are right skewed in the recent years of 2021 to 2023, which is indicating that the wind speeds are not very high. Compared to the years from 1985 to 2019, they are only little bit right skewed with less counts and looks more uniform instead of obvious right skewed pattern. Therefore, there might be an overall decrease in wind speed over years here.

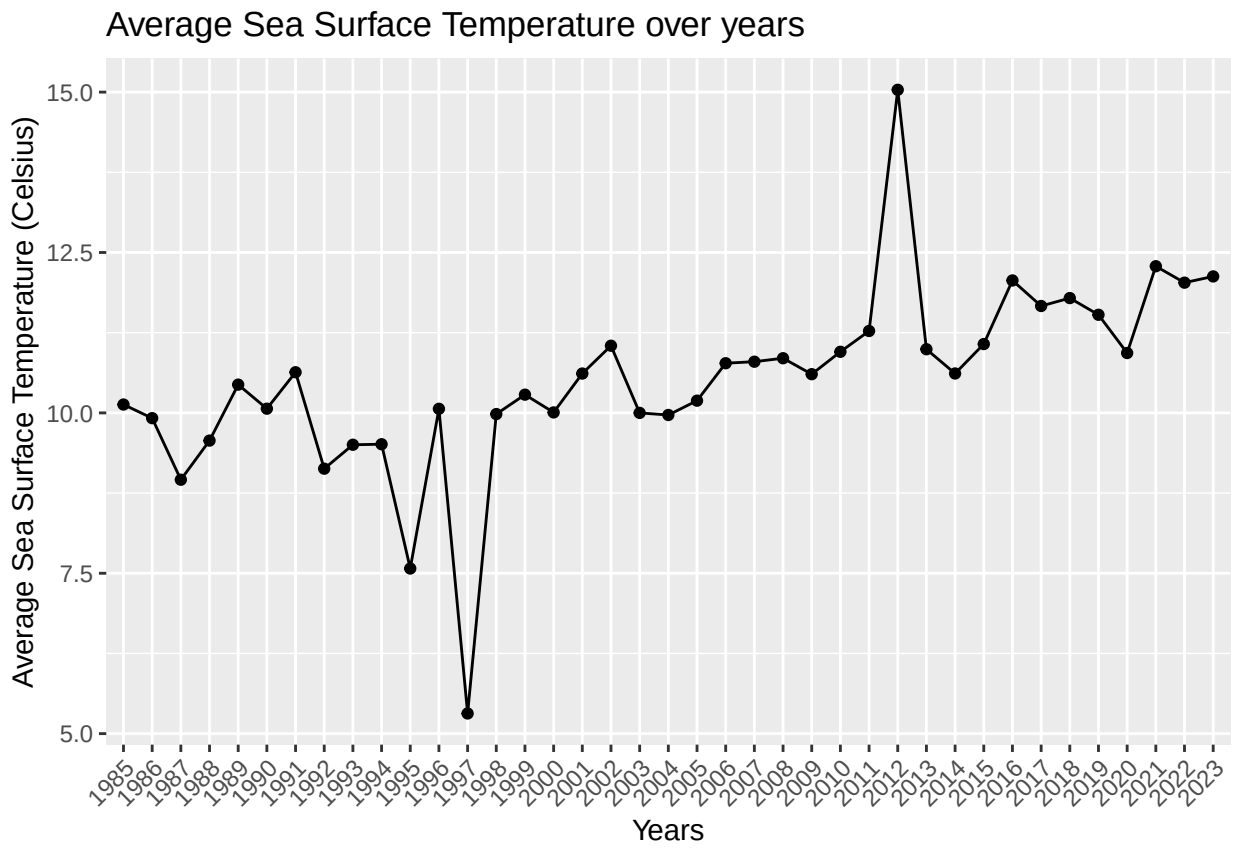
Sea Surface Temperature Next, let's investigate the factor sea surface temperature. Here are the first few rows about the average of sea temperature in each year. The following are the visualizations for sea temperature.

```
# calculate the average temperature for each year (AVTMP)
buoy <- buoy %>%
  mutate(cleaned_buoy %>% group_by(Year) %>%
           summarise(AVTMP = mean(WTMP, na.rm = TRUE)))
tmp <- buoy %>% select(Year, AVTMP)
head(tmp)
```

```
## # A tibble: 6 x 2
##    Year AVTMP
##   <int> <dbl>
## 1  1985 10.1
## 2  1986  9.92
## 3  1987  8.96
## 4  1988  9.57
## 5  1989 10.4
## 6  1990 10.1
```

Below is the graph about the Average Sea Temperature over years.

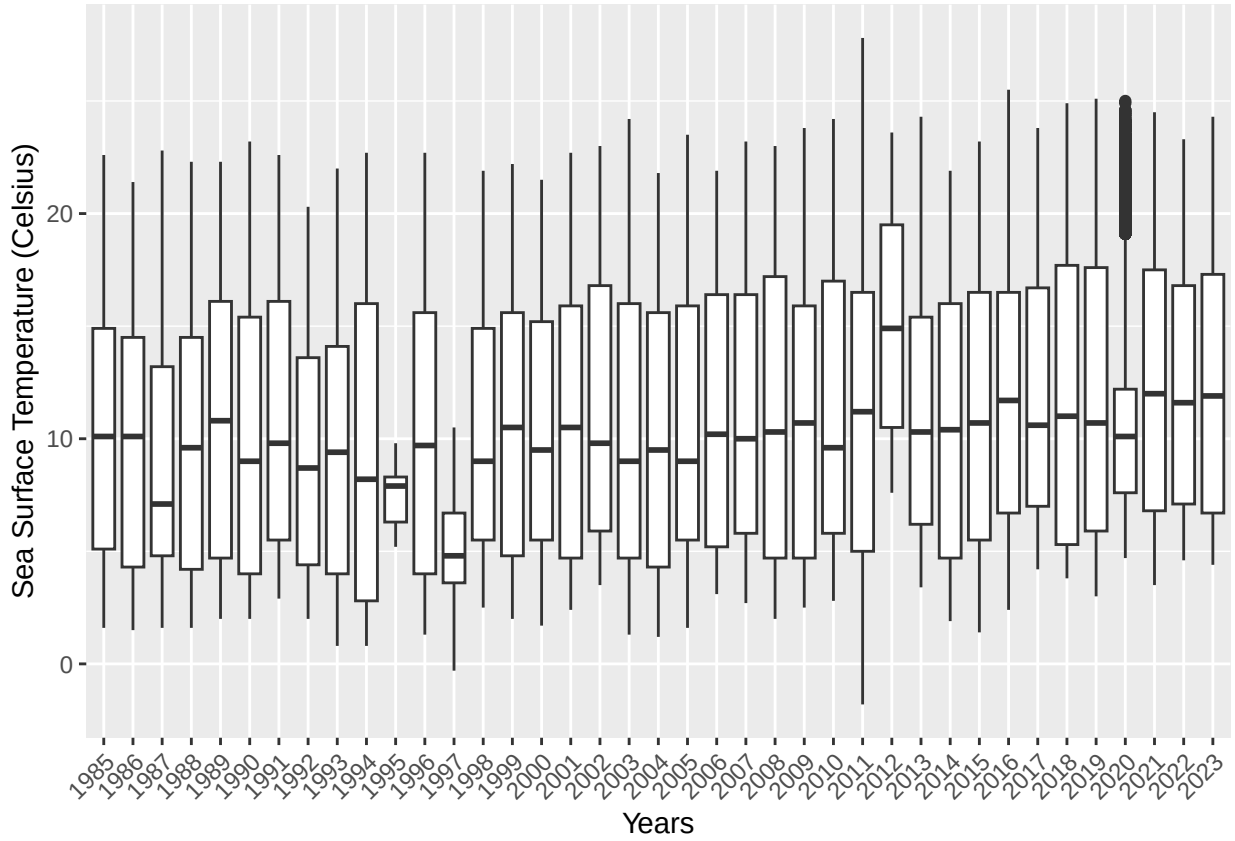
```
# Average sea temperature over years
ggplot(buoy, aes(as.character(Year), AVTMP)) +
  geom_point() + geom_line(group = 1) + theme(axis.text.x=element_text(angle=45,hjust=1)) +
  labs(x = "Years", y = "Average Sea Surface Temperature (Celsius)",
       title = "Average Sea Surface Temperature over years")
```



By looking at the above graph, we can see that there seems to have an overall trend of rising in average sea surface temperature over the years. As the sea temperature increase, there are several information can be interpreted such as the issue of global warming, and cause the glaciers to melt and therefore increase the sea level. However, by looking at the year of 1997, there was a sharp decrease in the sea surface temperature and a significant increase in the year of 2012. These are the outliers here and it would be better to do some further research to explore what was happening in these two years.

```
# boxplot of sea surface temperature for each year
ggplot(cleaned_buoy, aes(factor(Year), WTMP)) +
  geom_boxplot() + theme(axis.text.x=element_text(angle=45,hjust=1)) +
  labs(x = "Years", y = "Sea Surface Temperature (Celsius)")
```

```
## Warning: Removed 13197 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



The above is the box plot of sea surface temperature for each year. We can see that the most of the boxes have long height, which are indicating that the variability for sea surface temperature in each year was quite large. Some of the years like 1995, 1997 and 2020 seem to have short height, so the sea surface temperature of these years have less spread. By looking at the median of temperatures in each year, there seems to have an overall trend of a little increase in sea surface temperatures. Most of the median line for each year is located near the middle of the box, which the distribution could be symmetric. Some years like 1995, 1997 and 2011 might have some skewness here. In addition, there are many outliers in the year of 2020, and might needed further research.

Sea Level Pressure (hPa) Lastly, let's explore the factor sea level pressure.

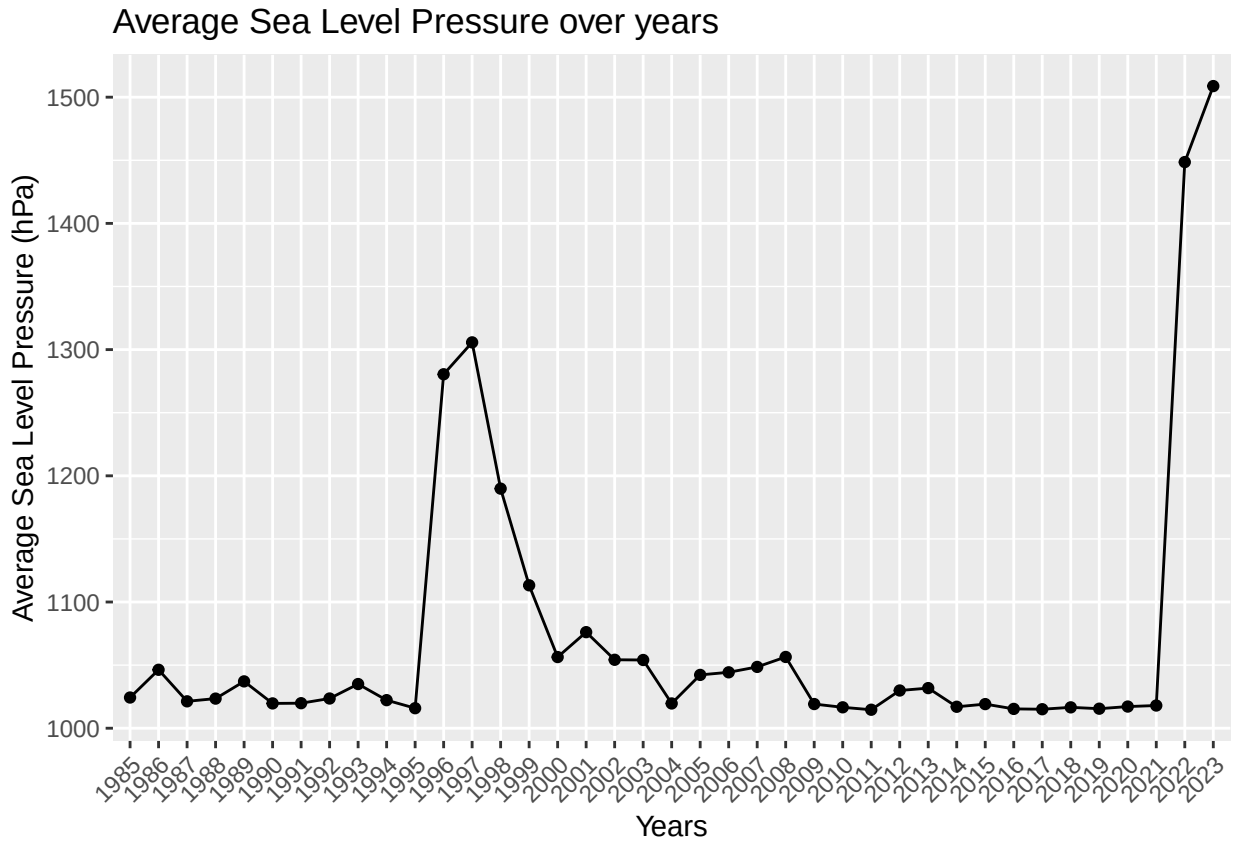
```
# Calculate the average sea level pressure for each year (AVPRES)
buoy <- buoy %>%
  mutate(cleaned_buoy %>% group_by(Year) %>%
           summarise(AVPRES = mean(Pressure, na.rm = TRUE)))
pres <- buoy %>% select(Year, AVPRES)
head(pres)
```

```
## # A tibble: 6 x 2
##    Year AVPRES
##   <int>  <dbl>
## 1  1985  1024.
## 2  1986  1046.
## 3  1987  1021.
## 4  1988  1023.
```

```
## 5  1989  1037.
## 6  1990  1020.
```

Above table shows the first few average sea level pressure of each year.

```
# Average sea level pressure over years
ggplot(buoy, aes(as.character(Year), AVPRES)) +
  geom_point() + geom_line(group = 1) + theme(axis.text.x=element_text(angle=45,hjust=1)) +
  labs(x = "Years", y = "Average Sea Level Pressure (hPa)",
       title = "Average Sea Level Pressure over years")
```



By looking at the graph above, we can see that most of the time, the average sea level pressure looks constant, between 1000hPa and 1050hPa. However, there are significant increase in the years of 1997 and 2023. There might exist some unusual things going on, maybe associated with some extreme weathers, which is needed to do more research.

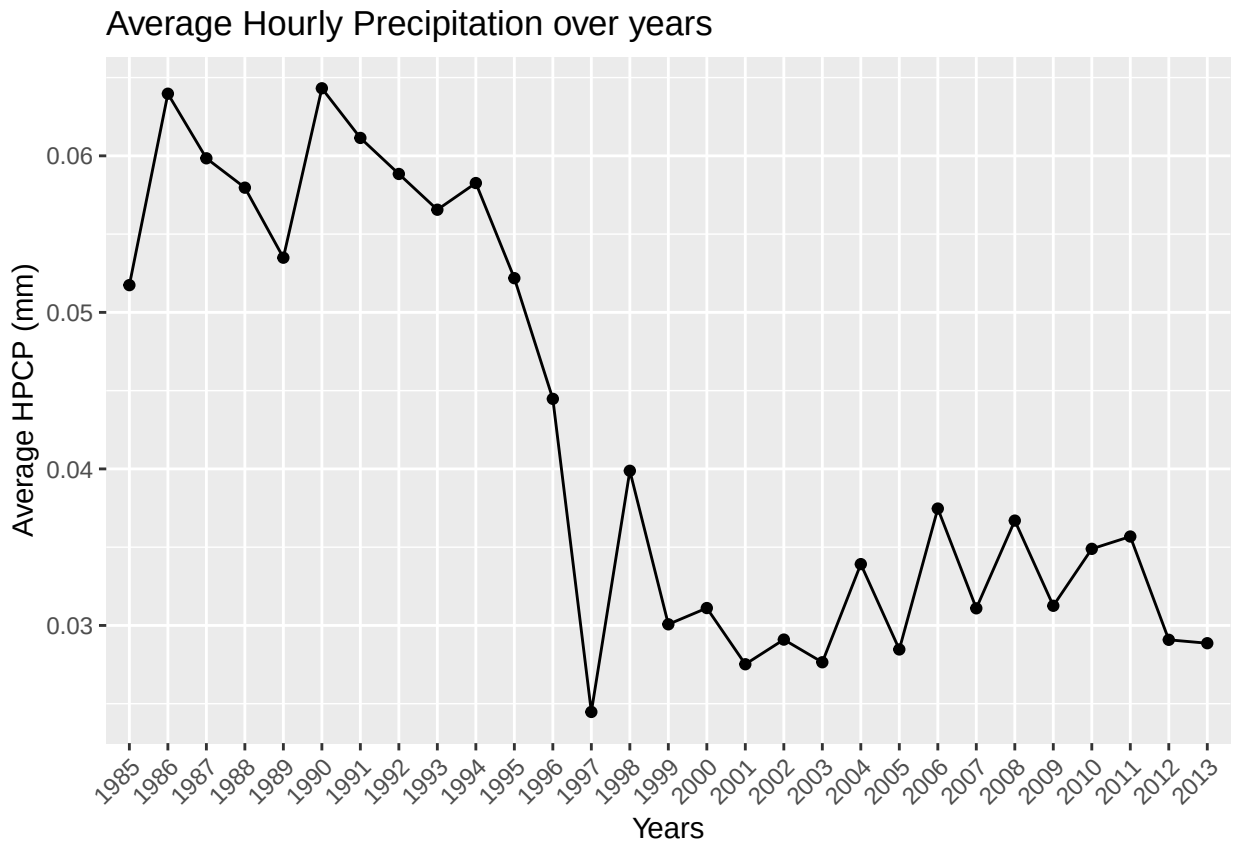
Overall, the factors wind speed, sea surface temperature, and sea level pressure do associated with climate change. The decrease in wind speed, increase in sea surface temperature and some extreme situations of sea level pressure might indicate an extreme weather and may also related to the issue of global warming as well.

Part (d)

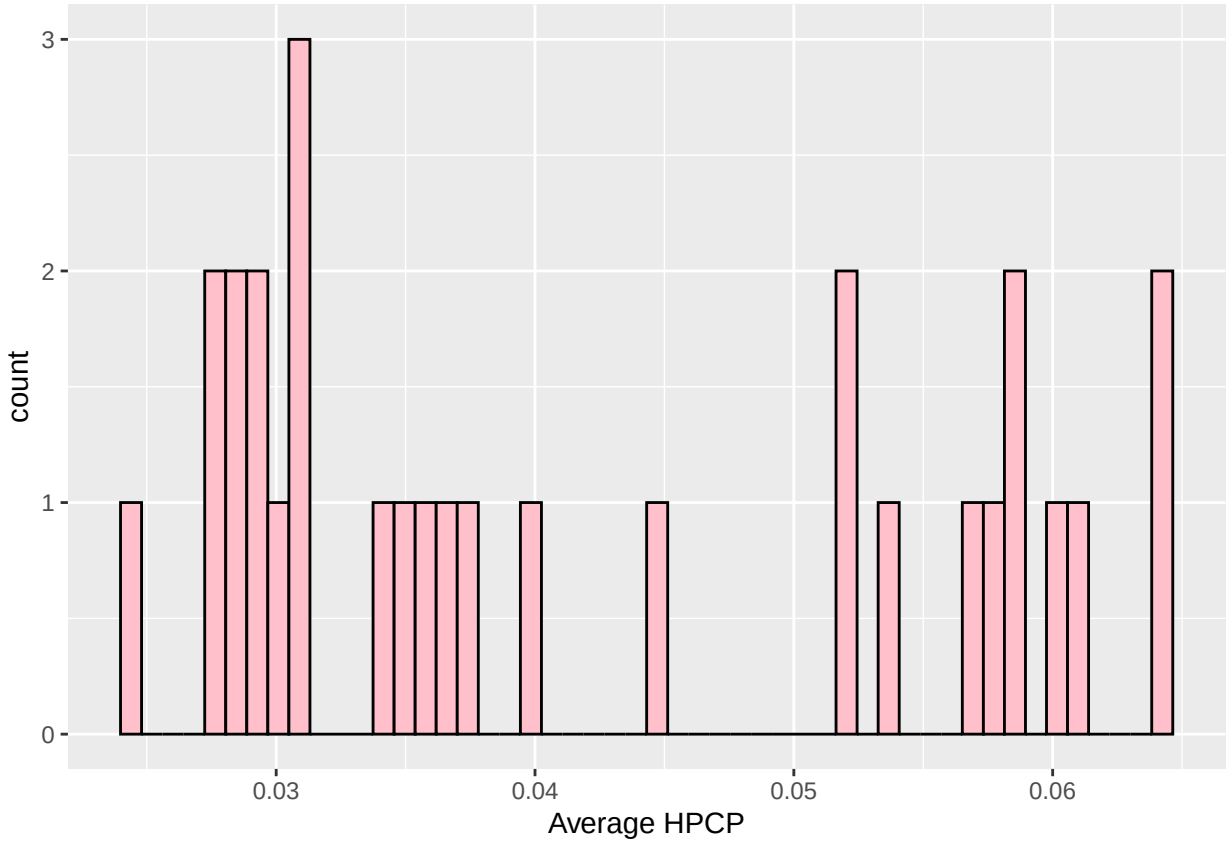
First, let's display few average hourly precipitation for each year through a table to get a sense of it.

```
##   Year    AVGRAIN
## 1 1985 0.05173975
## 2 1986 0.06396825
## 3 1987 0.05984211
## 4 1988 0.05796667
## 5 1989 0.05349306
## 6 1990 0.06431535
```

The following is a graph about the average hourly precipitation over the years.



From the above graph, we can see that there is a huge decrease in average hourly precipitation started from 1994 to 1997. After 1997, the precipitation seems to go back up and having small fluctuations over the years to 2013. Overall, there is a trend of decreasing in HPCP over the years started from 1985 to 2013.



From the histogram above, it is hard to see the distribution of the average hourly precipitation over years. Maybe a little bit right skewed since there highest bar here is located near 0.03.

Now, it is the time to build a simple linear regression model using the factors from part(C). The factors are wind speed, sea surface temperature, and sea level pressure.

```
##
## Call:
## lm(formula = AVGHCPC ~ AVWS + AVTMP + AVPRES, data = new_buoy)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.016601 -0.010697 -0.002733  0.009257  0.022777
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.221e-01  7.344e-02   1.663   0.1088
## AVWS         5.261e-03  6.597e-03   0.797   0.4327
## AVTMP       -3.038e-03  1.879e-03  -1.617   0.1184
## AVPRES      -7.646e-05  3.613e-05  -2.116   0.0444 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01252 on 25 degrees of freedom
## Multiple R-squared:  0.2371, Adjusted R-squared:  0.1456
## F-statistic:  2.59 on 3 and 25 DF,  p-value: 0.07528
```

By looking at the fitted linear regression model above, we can see that the overall result might not went

well. Since the coefficients for intercept, the average wind speed, and the average sea surface temperature are not significant due to the reason of their p-values, where are all greater than 0.05. Which we failed to reject the null hypothesis of coefficients are all equal to 0. There are many other factors to think about, such as considering making a necessary transformation so the distribution of variables would become symmetric and maybe try to consider interaction. It is also possible that the linear regression is not the best way to predict precipitation.

Therefore, there are too many things needed to be considered in order to predict the precipitation. Beside the things mentioned above, correlation between variables is also an very important factor to think about for building a linear regression, Therefore, it is needed to do lots of work before fitting the model. Hence, I feel definitely more sympathy for making wrong predictions in forecasting weather.