HW4 Qb-Qd

Suheng Yao

September 24, 2024

Question b

```
combined_data[combined_data==99 | combined_data==999] <- NA
print(head(combined_data))
```

```
##               DateTime  WDIR  WSPD   GST  WVHT   DPD   APD   MWD  PRES  ATMP
##                 <POSc> <int> <num> <num> <num> <num> <num> <int> <num> <num>
## 1: 1985-01-01 00:00:00    NA     4     5    NA    NA    NA    NA    NA   4.7
## 2: 1985-01-01 01:00:00    NA     4     5    NA    NA    NA    NA    NA   5.1
## 3: 1985-01-01 02:00:00    NA     4     5    NA    NA    NA    NA    NA   5.6
## 4: 1985-01-01 03:00:00    NA     4     5    NA    NA    NA    NA    NA   5.8
## 5: 1985-01-01 04:00:00    NA     4     5    NA    NA    NA    NA    NA   5.8
## 6: 1985-01-01 05:00:00    NA     4     5    NA    NA    NA    NA    NA   5.3
##     WTMP  DEWP   VIS  TIDE    WD    BAR
##    <num> <num> <num> <num> <int>  <num>
## 1:   6.7    NA    NA    NA    60 1030.3
## 2:   6.7    NA    NA    NA    80 1030.0
## 3:   6.6    NA    NA    NA   100 1030.1
## 4:   6.7    NA    NA    NA   100 1029.4
## 5:   6.7    NA    NA    NA   110 1028.6
## 6:   6.7    NA    NA    NA    90 1027.8
```

It may not be always appropriate to convert 99 or 999 to NA because sometimes 99 or 999 could be meaningful data points, and converting them to NA will cause data loss and affect the model's performance. Also, some missing values are in the columns with only character or factor values, converting them directly with NA may not be helpful with further data cleaning steps. Further analysis on the pattern of NA:

```
total_missing <- sum(is.na(combined_data)) # find the total number of missing value in data
print(total_missing)
```

```
## [1] 3092771
```

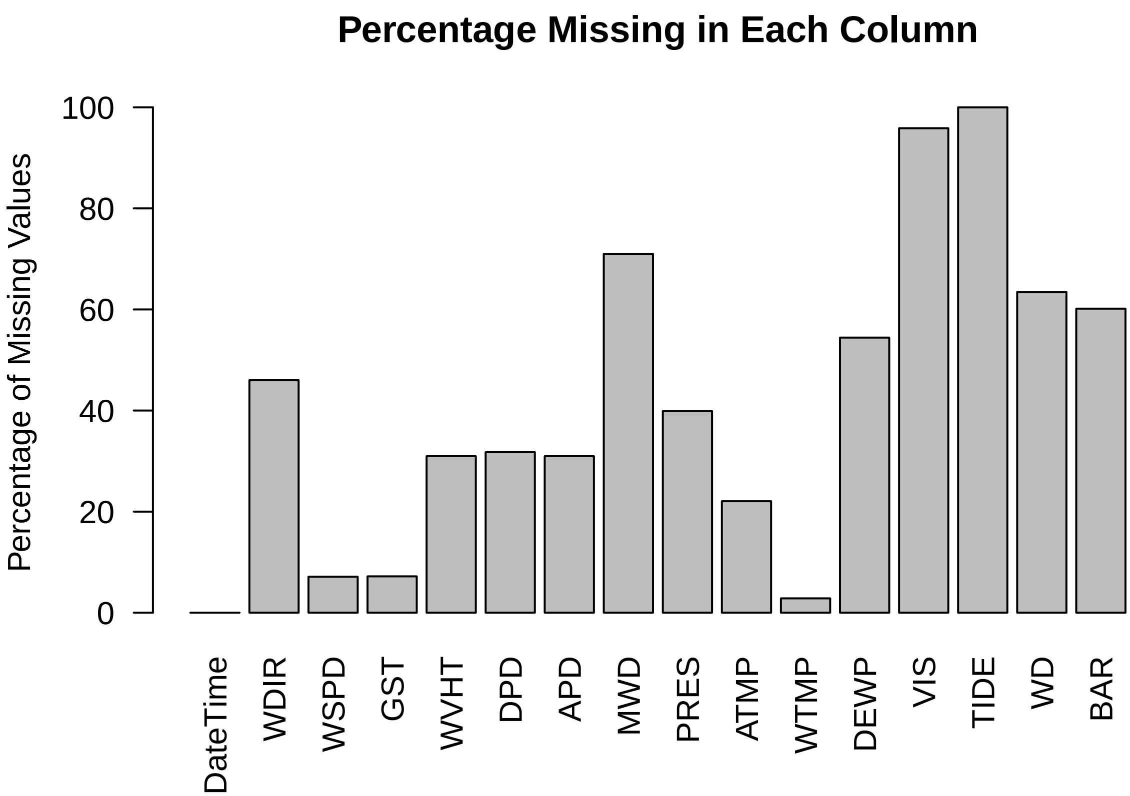
```
missing_col <- colSums(is.na(combined_data)) # find the number of missing values in each column
print(missing_col)
```

```
## DateTime     WDIR     WSPD      GST     WVHT      DPD      APD      MWD
##        0   214406    33193    33495   144283   147975   144283   330839
##     PRES     ATMP     WTMP     DEWP      VIS     TIDE       WD      BAR
##   185927   102771    13197   253630   446734   465973   295757   280308
```

```r
# find the percentage of missing values in each column
missing_percent <- (colSums(is.na(combined_data))/nrow(combined_data)) * 100
print(missing_percent)
```

```
##   DateTime       WDIR       WSPD        GST       WVHT        DPD        APD
##   0.000000  46.012537   7.123374   7.188185  30.963811  31.756132  30.963811
##        MWD       PRES       ATMP       WTMP       DEWP        VIS       TIDE
##  70.999607  39.900810  22.055141   2.832138  54.430192  95.871220 100.000000
##         WD        BAR
##  63.470845  60.155417
```

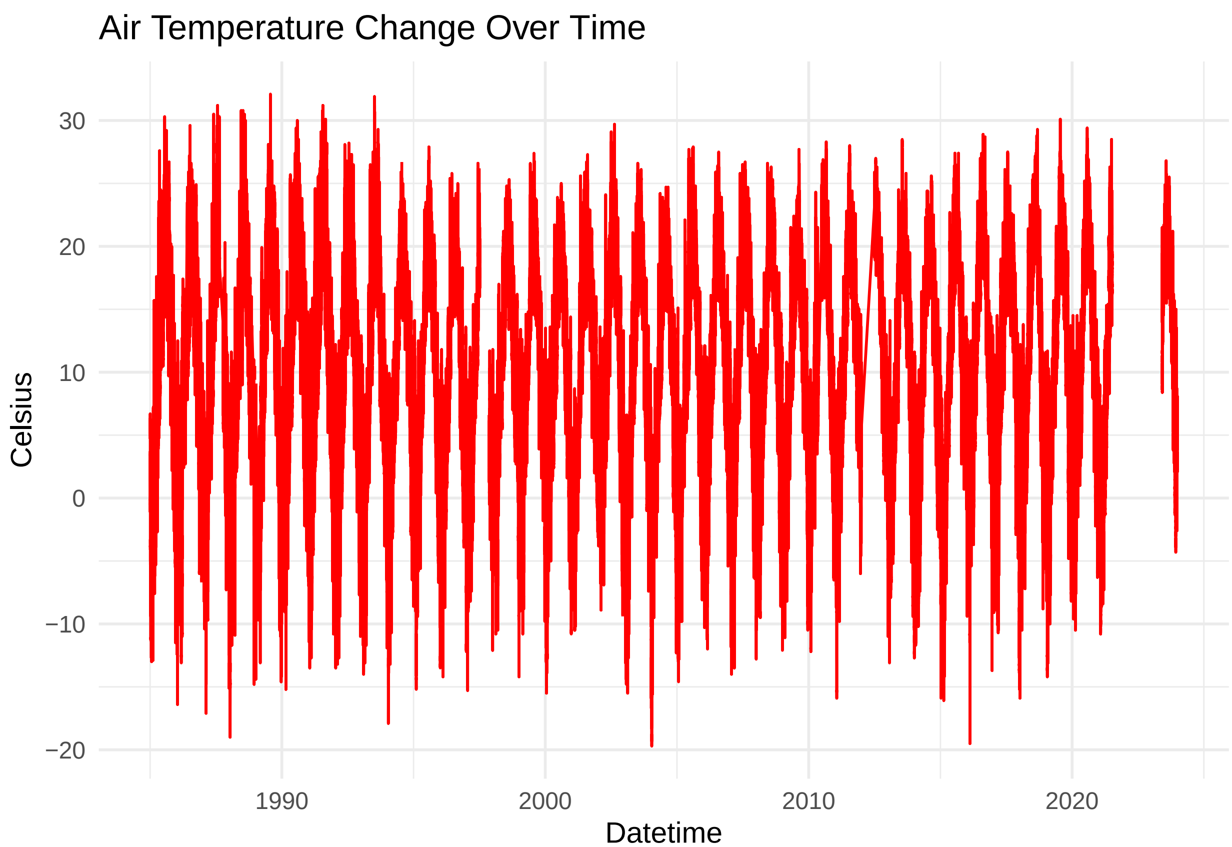
```r
barplot(missing_percent, main = "Percentage Missing in Each Column",
        ylab = "Percentage of Missing Values", las=2)
```



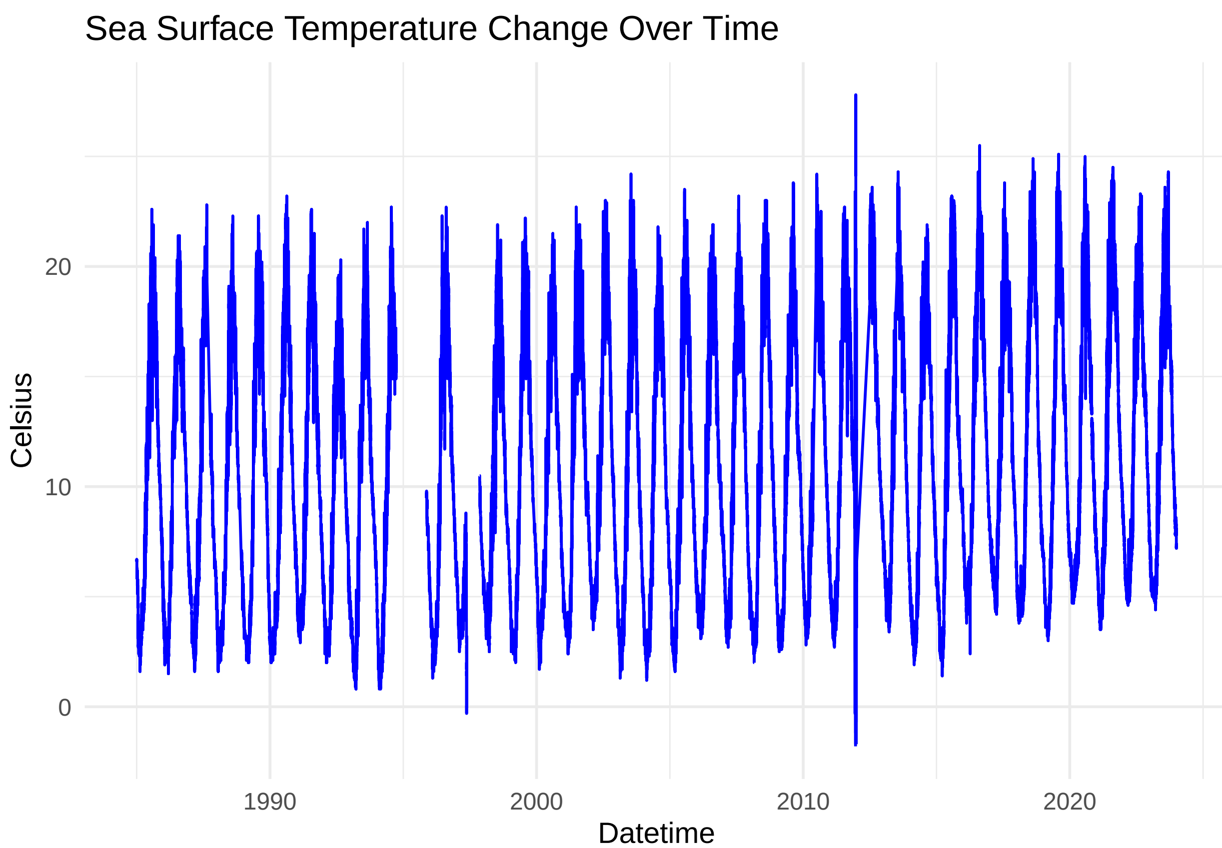
The graph shows that there are a lot of missing values in VIS and TIDE variables. In most of the datasets from 1985 to 2023, the station visibility data(VIS) is filled with 99, and only starting from 2000, the TIDE variable had been added to the data. So VIS and TIDE tend to have a lot of missing values. Also, MWD(Wave Measurements), WD(Wind Directions) and BAR(Pressure) get a lot of missing values because WD and BAR are all old metrics, and the station stop using those labels long time ago, adn for MWD, the reason could be related to that the wave measurements are not directly measured by sensors on board the buoys.

Question c

```r
library(ggplot2)
library(dplyr)
ggplot(combined_data, aes(x = DateTime, y = ATMP)) +
  geom_line(colour = "red") +
  labs(x = "Datetime", y = "Celsius", title = "Air Temperature Change Over Time") +
  theme_minimal()
```

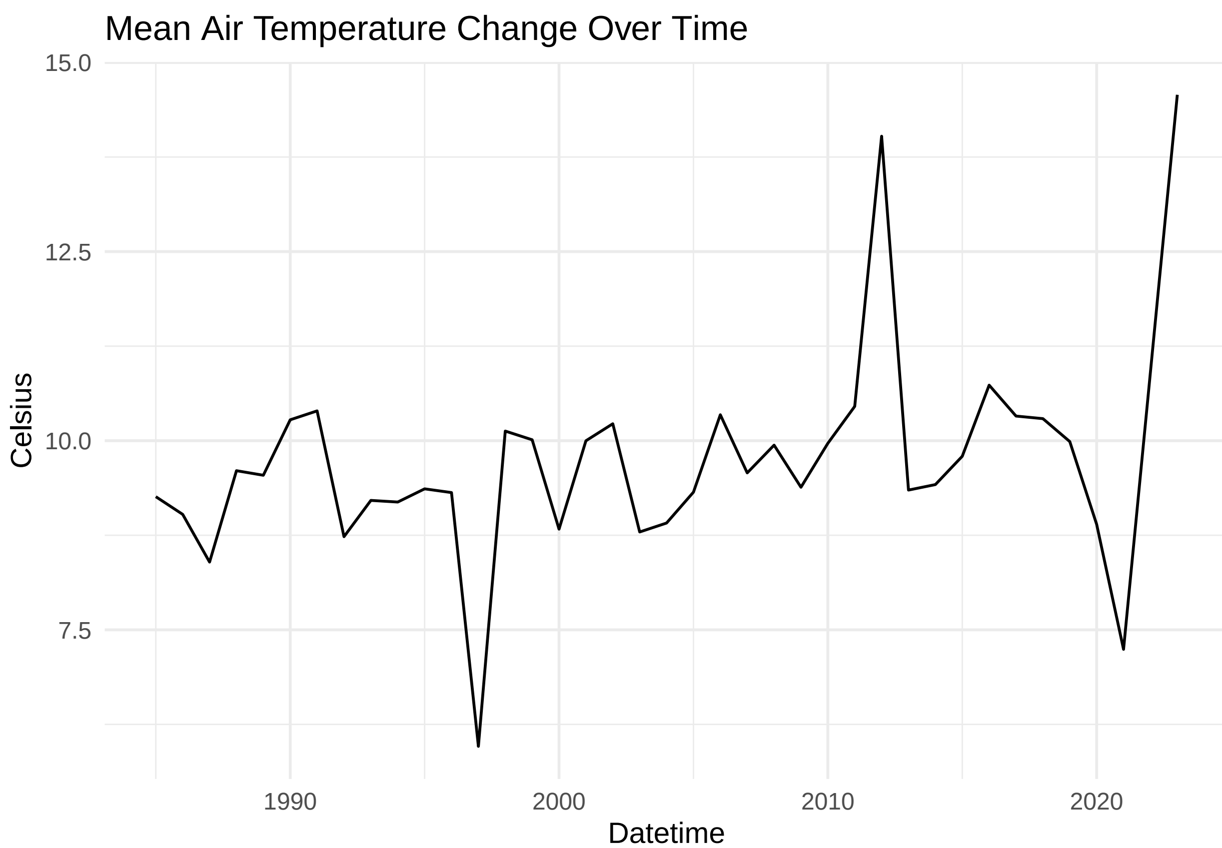



```r
ggplot(combined_data, aes(x = DateTime, y = WTMP)) +
  geom_line(colour = "blue") +
  labs(x = "Datetime", y = "Celsius", title = "Sea Surface Temperature Change Over Time") +
  theme_minimal()
```



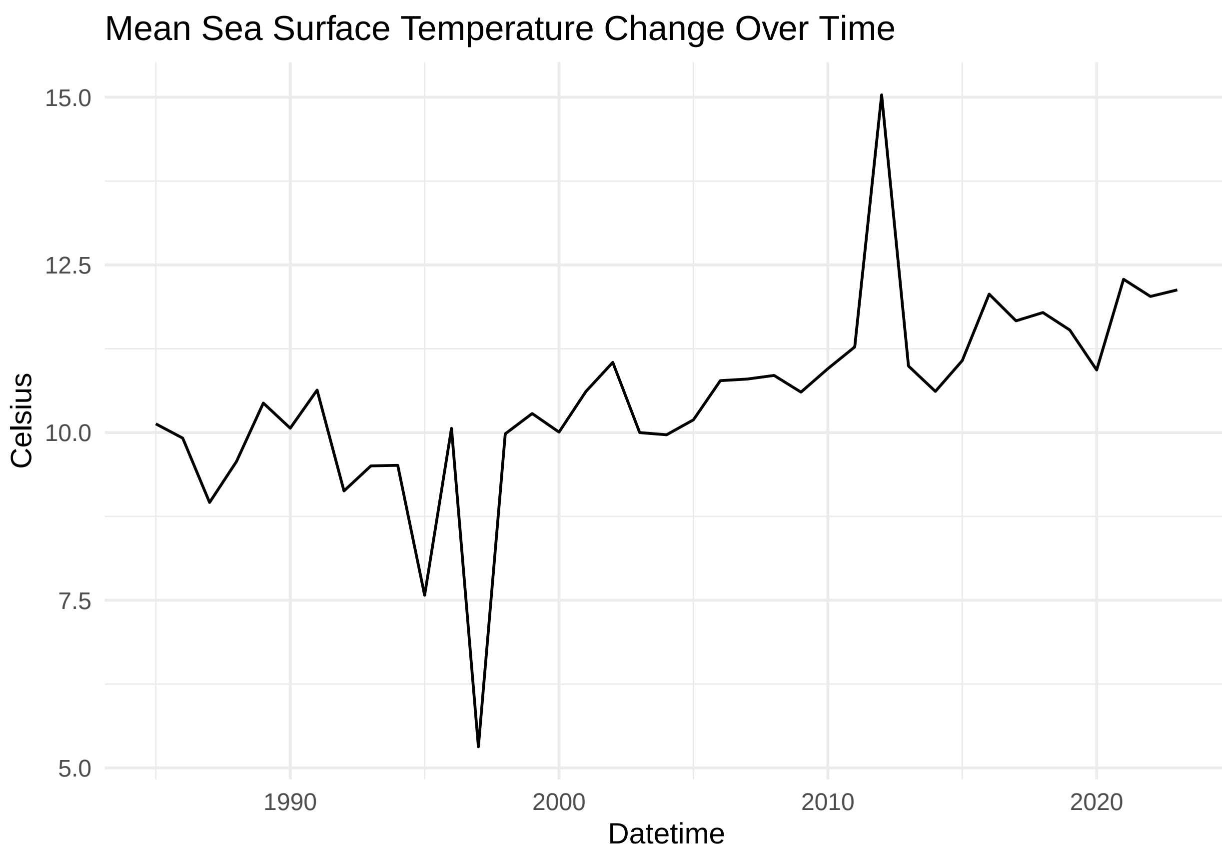

```
mean_year_air <- combined_data %>%
  mutate(Year = as.integer(format(DateTime, "%Y"))) %>%
  group_by(Year) %>%
  summarise(mean_atmp = mean(ATMP, na.rm = TRUE)) %>%
  filter(!is.na(mean_atmp))

ggplot(mean_year_air, aes(x=Year, y=mean_atmp))+
  geom_line()+
  labs(x = "Datetime", y = "Celsius", title = "Mean Air Temperature Change Over Time")+
  theme_minimal()
```




```
mean_year_sea <- combined_data %>%
  mutate(Year = as.integer(format(DateTime, "%Y"))) %>%
  group_by(Year) %>%
  summarise(mean_wtmp = mean(WTMP, na.rm = TRUE))

ggplot(mean_year_sea, aes(x=Year, y=mean_wtmp))+
  geom_line()+
  labs(x = "Datetime", y = "Celsius", title = "Mean Sea Surface Temperature Change Over Time")+
  theme_minimal()
```



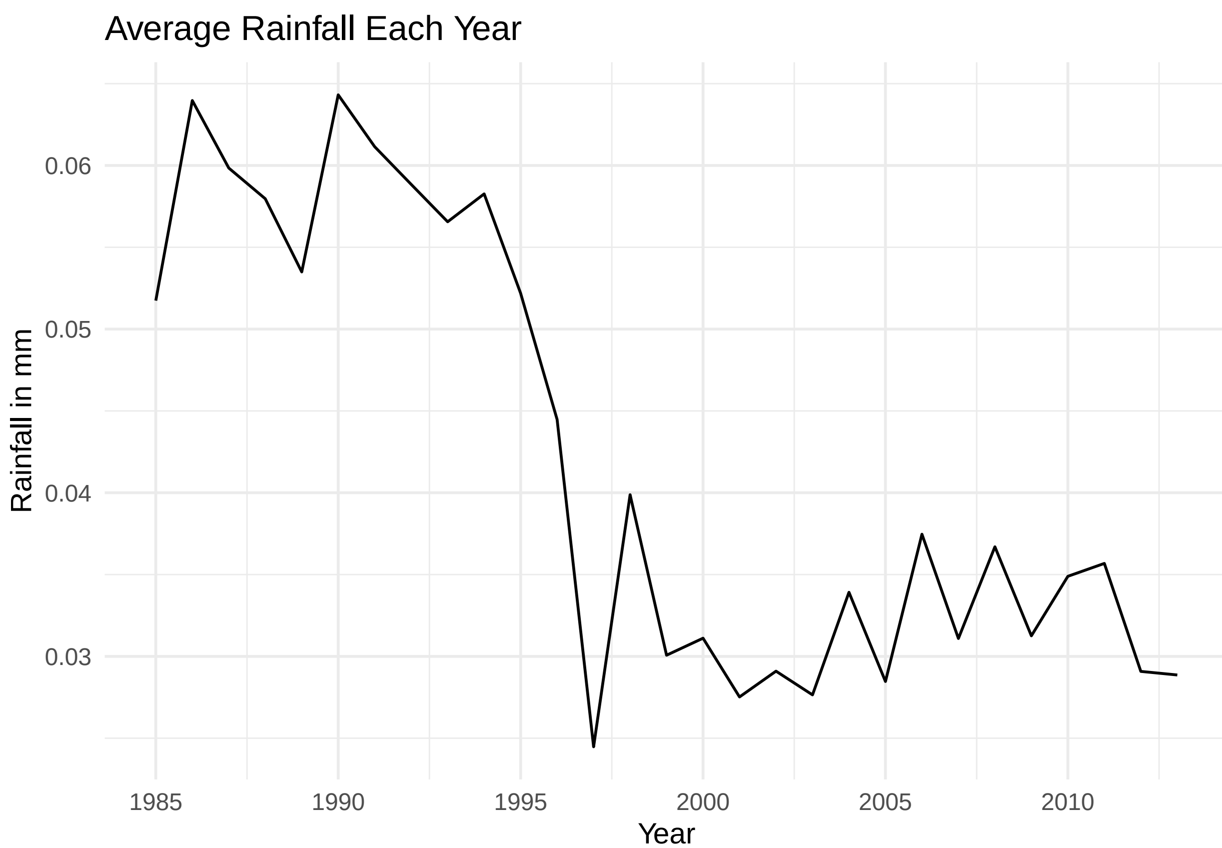
After drawing the ATMP(Air Temperature) and WTMP(Sea Surface Temperature) change over time, I can clearly see that temperature in 2023 is much higher than temperature in 1985, indicating that the climate change is a real issue. Also, from the mean air and sea surface temperature line graph, the average temperature is clearly increasing from 1985 to 2023, especially for the sea surface temperature.

Question d

```r
library(lubridate)
library(dplyr)
df <- read.csv("Rainfall.csv")
df$DATE <- ymd_hm(df$DATE) # change the column to a date-time object

# Plot the average rainfall changing over time
mean_year_rain <- df %>%
  mutate(Year = as.integer(format(DATE, "%Y"))) %>%
  group_by(Year) %>%
  summarise(mean_rain = mean(HPCP, na.rm = TRUE))

ggplot(mean_year_rain, aes(x=Year, y=mean_rain))+
  geom_line()+
  labs(x = "Year", y = "Rainfall in mm", title = "Average Rainfall Each Year")+
  theme_minimal()
```

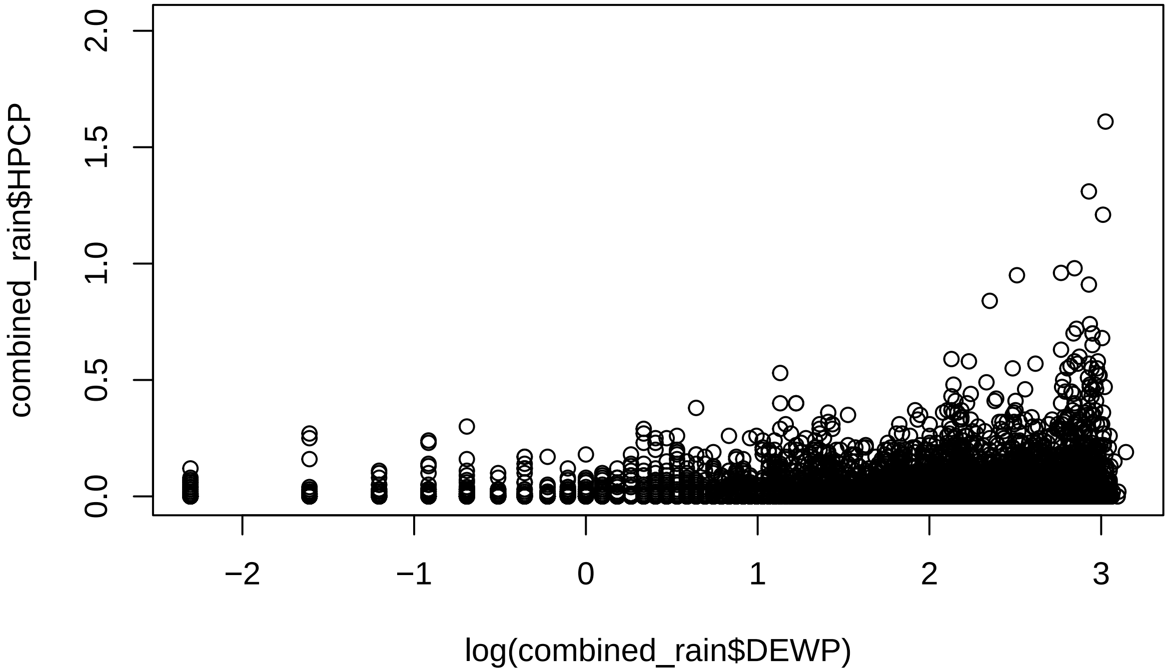



```r
# summary of the distribution of the mean rainfall
dist_stats <- data.frame(
  Standard_Deviation = sd(mean_year_rain$mean_rain, na.rm = TRUE),
  Variance = var(mean_year_rain$mean_rain, na.rm = TRUE),
  IQR = IQR(mean_year_rain$mean_rain, na.rm = TRUE),
  Min = min(mean_year_rain$mean_rain, na.rm = TRUE),
  Max = max(mean_year_rain$mean_rain, na.rm = TRUE),
  Median = median(mean_year_rain$mean_rain, na.rm = TRUE)
)
print(dist_stats)
```

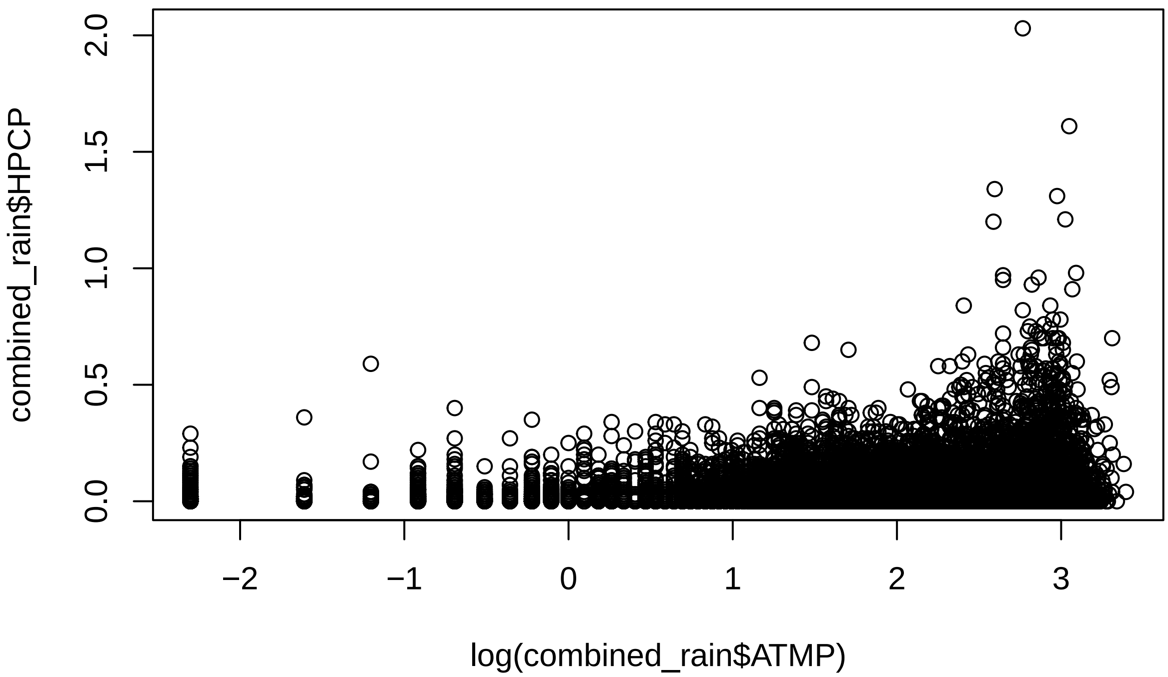
```
##   Standard_Deviation     Variance        IQR        Min        Max     Median
## 1         0.01353984 0.0001833274 0.02648594 0.02447411 0.06431535 0.03669661
```

From the plot of the average rainfall, it is clear that there is a significant decrease in rainfall from 1990 to 1998. Also, from the summary table shown, the standard deviation is 0.013, and the IQR is 0.026, indicating that the data is quite spread out. Additionally, there is a 0.04 difference between min value and max value, showing that the variability in rainfall data is intense.

```r
df <- df %>% rename(DateTime=DATE)
combined_rain <- inner_join(combined_data, df, by="DateTime")
plot(log(combined_rain$DEWP), combined_rain$HPCP)
```




```
plot(log(combined_rain$ATMP), combined_rain$HPCP)
```



Since the air temperature and dew point temperature are related to the amount of rainfall, and from the two graphs above, there are some patterns within the scatterplot, it is a good idea to fit a simple linear regression to try to predict the amount of rainfall:

```
combined_rain <- combined_rain %>%
  filter(!is.na(DEWP) & !is.na(ATMP))

modeld <- lm(HPCP~ATMP+DEWP,
             data = combined_rain)
summary(modeld)
```

```
##
## Call:
## lm(formula = HPCP ~ ATMP + DEWP, data = combined_rain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05132 -0.03199 -0.02044  0.00149  1.56064
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0231890  0.0014940  15.521  < 2e-16 ***
## ATMP        -0.0001362  0.0003289  -0.414    0.679
## DEWP         0.0014099  0.0002821   4.999 5.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.07049 on 10293 degrees of freedom
## Multiple R-squared:  0.02168,    Adjusted R-squared:  0.02149
## F-statistic:   114 on 2 and 10293 DF,  p-value: < 2.2e-16
```

From the output of the model, both variables are statistically significant, with p-values less than 0.05, but the R^2 is only 0.01046, indicating that only 1% variability is explained by the model, so this is not a good model to fit. To improve, since the rainfall data may be dependent on time, in the future exploration, I could try time series analysis model to capture the pattern.