```
library(data.table)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)

file_root <- "https://www.ndbc.noaa.gov/view_text_file.php?filename=44013h"
tail <- ".txt.gz&dir=data/historical/stdmet/"

years <- 1985:2023

buoy_data_list <- list()

for (year in years) {
  path <- paste0(file_root, year, tail)
  header <- scan(path, what = 'character', nlines = 1)
  skip_lines <- ifelse(year >= 2007, 2, 1)
  buoy <- fread(path, header = FALSE, skip = skip_lines)
  num_cols <- ncol(buoy)
  if (length(header) > num_cols) {
    header <- header[1:num_cols]
  } else if (length(header) < num_cols) {
    header <- c(header, paste0("V", (length(header) + 1):num_cols))
  }

  colnames(buoy) <- header

  # Fix the year column to always be four digits
  if ("YY" %in% colnames(buoy)) {
    buoy$YY <- ifelse(buoy$YY < 100, ifelse(buoy$YY > 20, 1900 + buoy$YY, 2000 + buoy$YY), buoy$Y
Y)
  }

  if ("YY" %in% colnames(buoy) & "MM" %in% colnames(buoy) & "DD" %in% colnames(buoy) & "hh" %in% c
olnames(buoy) & "mm" %in% colnames(buoy)) {
    buoy$Date <- ymd_hms(paste(buoy$YY, buoy$MM, buoy$DD, buoy$hh, buoy$mm))
  }

  buoy_data_list[[as.character(year)]] <- buoy
}
```

```
## Warning in fread(path, header = FALSE, skip = skip_lines): Stopped early on
## line 5114. Expected 16 fields but found 17. Consider fill=TRUE and
## comment.char=. First discarded non-empty line: <<2000 08 01 00 78 4.3 5.1 0.58
## 8.33 5.36 999 1022.9 17.3 17.5 15.0 99.0 99.00>>
```

```
buoy_data_list <- rbindlist(buoy_data_list, fill = TRUE)

buoy_data_list <- buoy_data_list %>%
  mutate(Year = coalesce(as.numeric(YY), as.numeric(YYYY), as.numeric(`#YY`))) %>%
  select(-YY, -YYYY, -`#YY`) %>%
  select(Year, everything())

buoy_data_list <- buoy_data_list %>%
  mutate(Wind_Direction = coalesce(WD, WDIR)) %>%
  select(-WD, -WDIR)

buoy_data_list <- buoy_data_list %>%
  mutate(Pressure = coalesce(BAR, PRES)) %>%
  select(-BAR, -PRES)


if (all(c("Year", "MM", "DD", "hh", "mm") %in% colnames(buoy_data_list))) {
  buoy_data_list[, date := ymd_hms(paste(Year, MM, DD, hh, mm))]
}
```

```
## Warning: 411397 failed to parse.
```

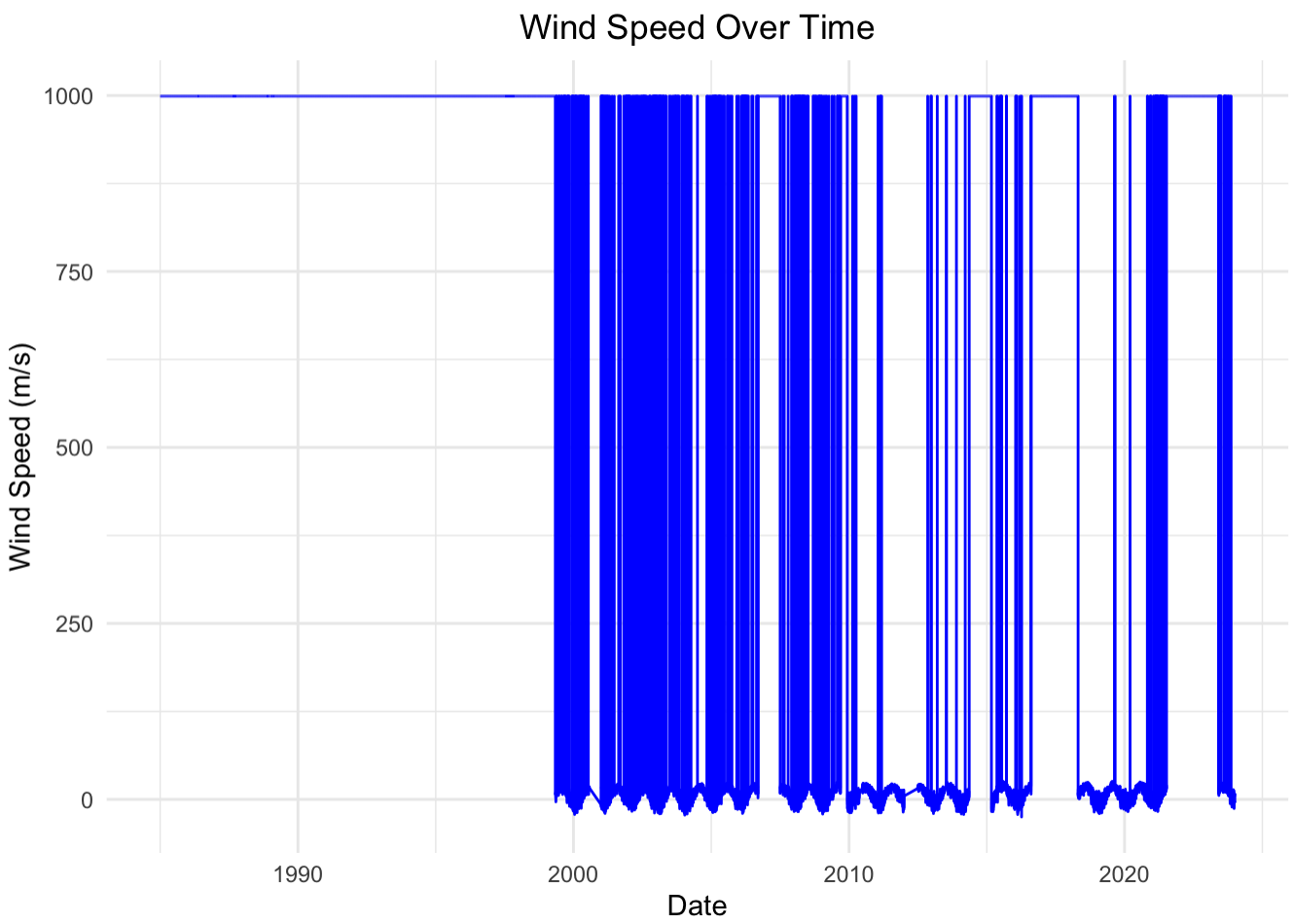
```
buoy_data_list = buoy_data_list %>% mutate(date=0)

summary(buoy_data_list[, .(Year, MM, DD, hh, mm)])
```

```
##       Year            MM               DD              hh
##  Min.   :1985   Min.   : 1.000   Min.   : 1.00   Min.   : 0.0
##  1st Qu.:1998   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 5.0
##  Median :2013   Median : 7.000   Median :16.00   Median :11.0
##  Mean   :2009   Mean   : 6.593   Mean   :15.73   Mean   :11.5
##  3rd Qu.:2021   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:17.0
##  Max.   :2023   Max.   :12.000   Max.   :31.00   Max.   :23.0
##
##        mm
##  Min.   : 0.00
##  1st Qu.:10.00
##  Median :40.00
##  Mean   :31.58
##  3rd Qu.:50.00
##  Max.   :50.00
##  NA's   :164650
```

```
#unique(buoy_data_list$Year)
#unique(buoy_data_list$MM)
#unique(buoy_data_list$DD)
#unique(buoy_data_list$hh)
#unique(buoy_data_list$mm)


buoy_data_list <- buoy_data_list %>%
  mutate(date = ifelse(complete.cases(Year, MM, DD, hh, mm),
                       make_datetime(year = Year, month = MM, day = DD, hour = hh, min = mm),
                       as.POSIXct(NA)))


buoy_data_list$date <- make_datetime(
  year = ifelse(is.na(buoy_data_list$Year), 2000, buoy_data_list$Year),
  month = ifelse(is.na(buoy_data_list$MM), 1, buoy_data_list$MM),
  day = ifelse(is.na(buoy_data_list$DD), 1, buoy_data_list$DD),
  hour = ifelse(is.na(buoy_data_list$hh), 0, buoy_data_list$hh),
  min = ifelse(is.na(buoy_data_list$mm), 0, buoy_data_list$mm)
)

str(buoy_data_list)
```

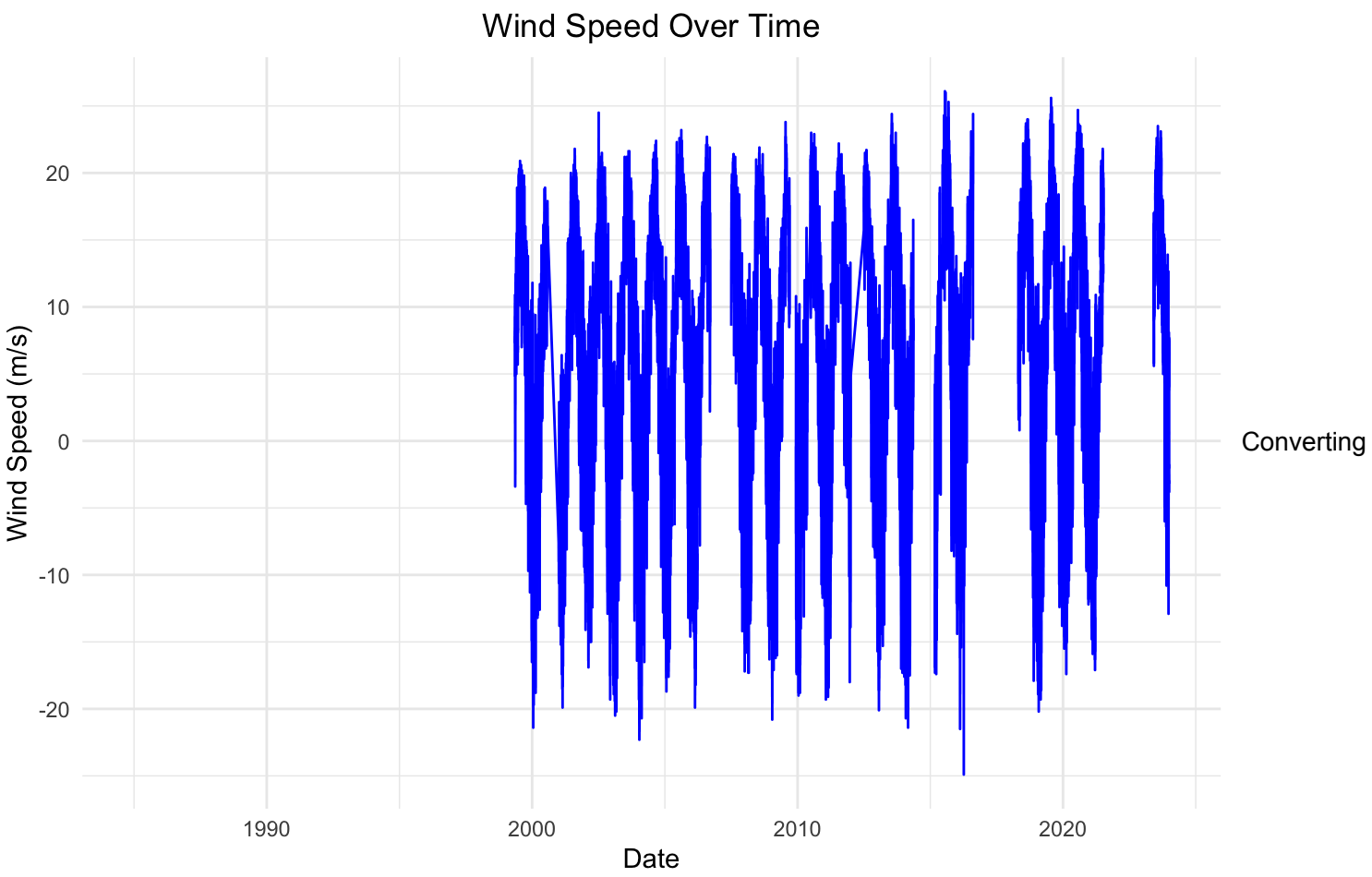
```
## Classes 'data.table' and 'data.frame':   462301 obs. of  19 variables:
##  $ Year          : num  1985 1985 1985 1985 1985 ...
##  $ MM            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ DD            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ hh            : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ WSPD          : num  4 4 4 4 4 4 4 4 6 7 ...
##  $ GST           : num  5 5 5 5 5 5 6 5 6 8 ...
##  $ WVHT          : num  99 99 99 99 99 99 99 99 99 99 ...
##  $ DPD           : num  99 99 99 99 99 99 99 99 99 99 ...
##  $ APD           : num  99 99 99 99 99 99 99 99 99 99 ...
##  $ MWD           : int  999 999 999 999 999 999 999 999 999 999 ...
##  $ ATMP          : num  4.7 5.1 5.6 5.8 5.8 5.3 5.5 5.8 5.9 6.2 ...
##  $ WTMP          : num  6.7 6.7 6.6 6.7 6.7 6.7 6.7 6.7 6.7 6.7 ...
##  $ DEWP          : num  999 999 999 999 999 999 999 999 999 999 ...
##  $ VIS           : num  99 99 99 99 99 99 99 99 99 99 ...
##  $ TIDE          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ mm            : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Wind_Direction: int  60 80 100 100 110 90 60 30 40 40 ...
##  $ Pressure      : num  1030 1030 1030 1029 1029 ...
##  $ date          : POSIXct, format: "1985-01-01 00:00:00" "1985-01-01 01:00:00" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
#b

buoy_data_list <- buoy_data_list %>% filter(!is.na(date))
ggplot(buoy_data_list, aes(x = date, y = DEWP)) +
  geom_line(color = "blue") +
  labs(title = "Wind Speed Over Time", x = "Date", y = "Wind Speed (m/s)") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



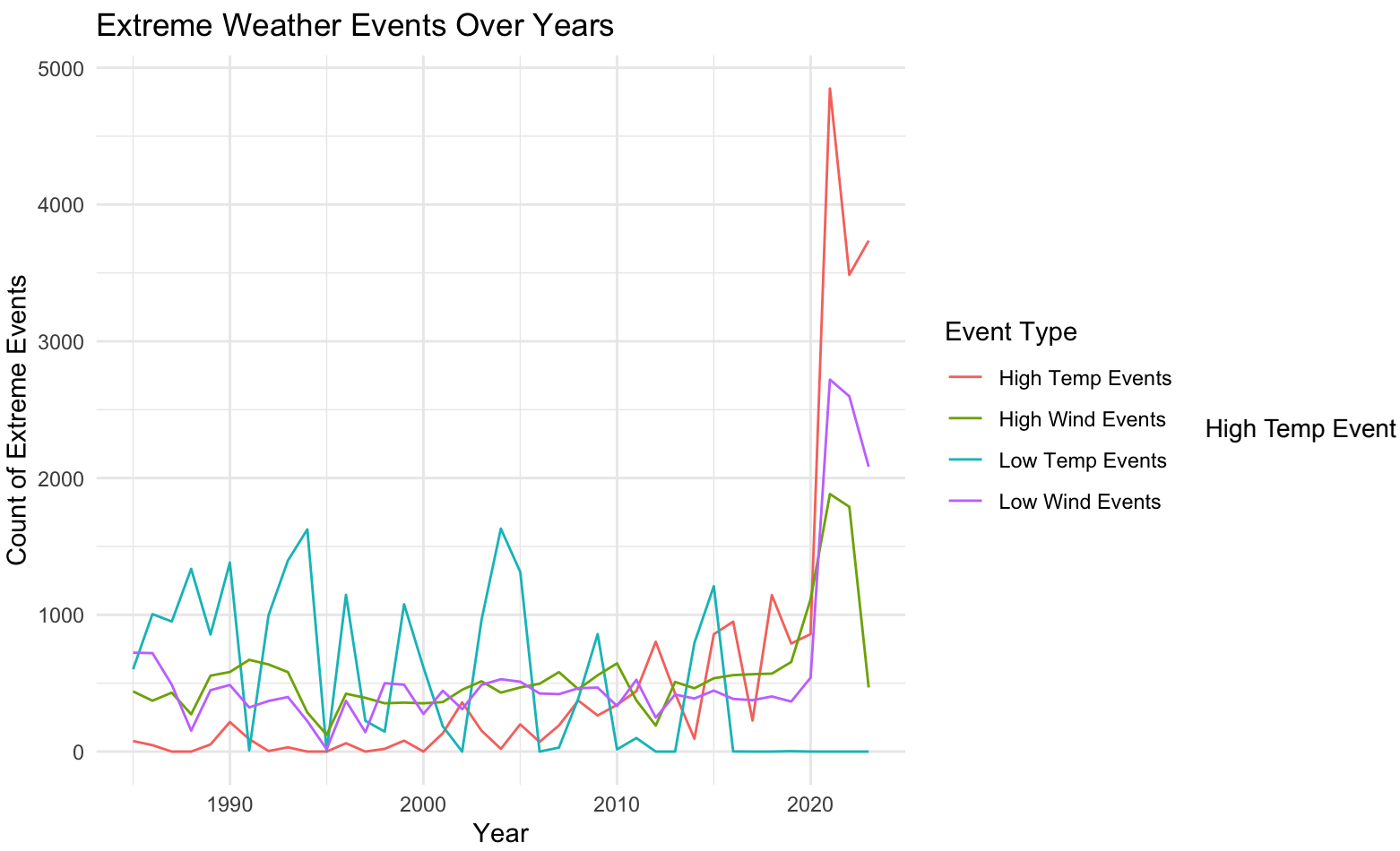

```
buoy_data_list[buoy_data_list == 999] <- NA
ggplot(buoy_data_list, aes(x = date, y = DEWP)) +
  geom_line(color = "blue") +
  labs(title = "Wind Speed Over Time", x = "Date", y = "Wind Speed (m/s)") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: Removed 119140 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



missing/empty data to NA is not always appropriate. When you need to analyze a situation where the data had 999 before (like DEWP), 999 into NA is not appropriate. From the two graphs I plotted, it's hard to see the change in wind speed when there's 999. When changing 999 to NA, the change in wind speed can become more obvious.

```
#C
library(tidyr)
library(ggplot2)
data=buoy_data_list
data[data==99]= NA
data <- data %>%
  mutate(WTMP = if_else(WTMP == 999, NA_real_, WTMP),
         WSPD = if_else(WSPD == 999, NA_real_, WSPD),
         date = as.Date(date, format = "%Y-%m-%dT%H:%M:%S"),
         Year = format(date, "%Y")) %>%
  drop_na(WTMP, WSPD)

extreme_thresholds <- data %>%
  summarise(
    High_WTMP = quantile(WTMP, 0.95, na.rm = TRUE),
    Low_WTMP = quantile(WTMP, 0.05, na.rm = TRUE),
    High_WSPD = quantile(WSPD, 0.95, na.rm = TRUE),
    Low_WSPD = quantile(WSPD, 0.05, na.rm = TRUE)
  )

extreme_events <- data %>%
  mutate(
    Extreme_High_WTMP = WTMP >= extreme_thresholds$High_WTMP,
    Extreme_Low_WTMP = WTMP <= extreme_thresholds$Low_WTMP,
    Extreme_High_WSPD = WSPD >= extreme_thresholds$High_WSPD,
    Extreme_Low_WSPD = WSPD <= extreme_thresholds$Low_WSPD
  ) %>%
  group_by(Year) %>%
  summarise(
    High_WTMP_Count = sum(Extreme_High_WTMP, na.rm = TRUE),
    Low_WTMP_Count = sum(Extreme_Low_WTMP, na.rm = TRUE),
    High_WSPD_Count = sum(Extreme_High_WSPD, na.rm = TRUE),
    Low_WSPD_Count = sum(Extreme_Low_WSPD, na.rm = TRUE)
  )

ggplot(extreme_events, aes(x = as.integer(Year))) +
  geom_line(aes(y = High_WTMP_Count, color = "High Temp Events")) +
  geom_line(aes(y = Low_WTMP_Count, color = "Low Temp Events")) +
  geom_line(aes(y = High_WSPD_Count, color = "High Wind Events")) +
  geom_line(aes(y = Low_WSPD_Count, color = "Low Wind Events")) +
  labs(title = "Extreme Weather Events Over Years",
       x = "Year",
       y = "Count of Extreme Events",
       color = "Event Type") +
  theme_minimal()
```



gradually increases with time. After 2020 High Temp Event went up extremely fast.