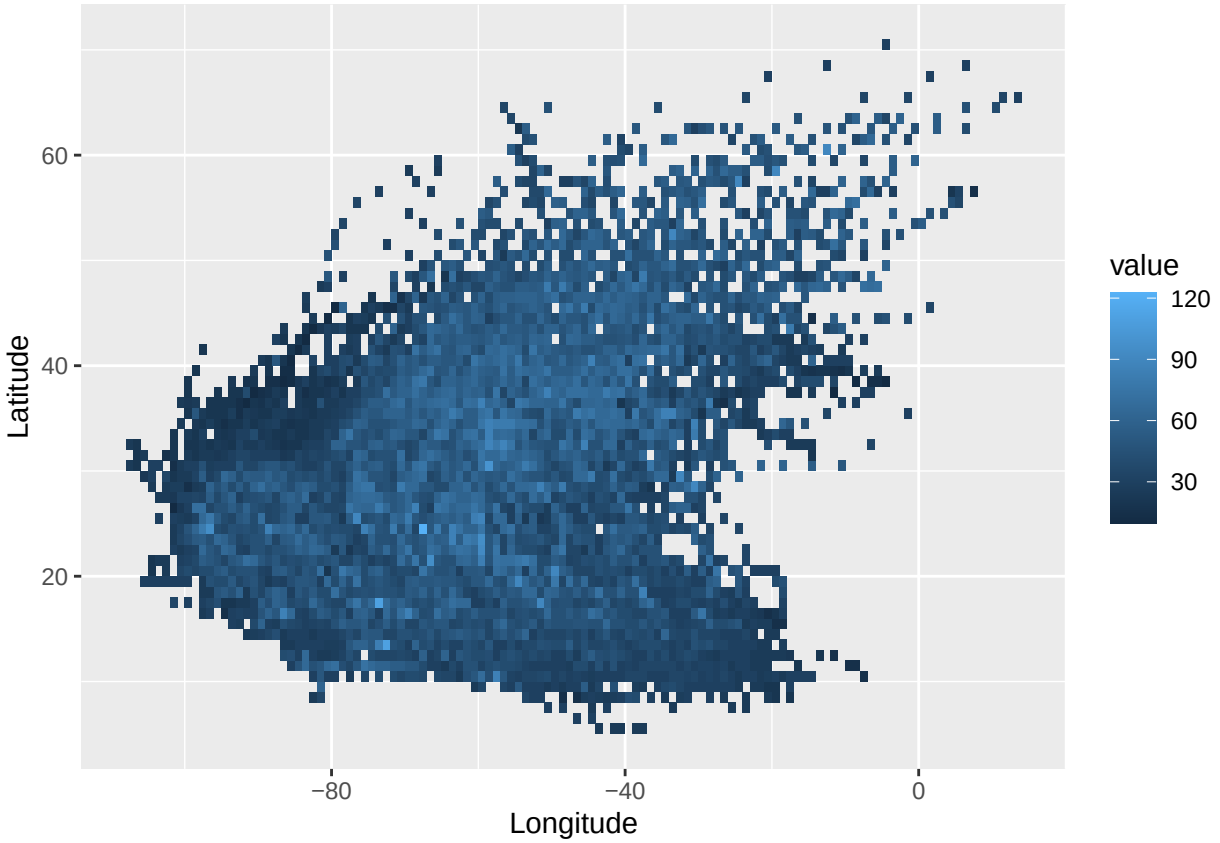
# EDA

```
origin_df <- read.csv("hurrican703.csv")

# identify a problem in the data wrangling part far below:
# different hurricanes with same names.
# for visualization, we need to fix the problem first.
dt <-
  origin_df %>%
  mutate(
    # 2 hurricanes with the name ALICE.1954
    ID = ifelse(ID == "ALICE.1954" & Month == "June", "ALICE.1954(1)", ID),
    ID = ifelse(ID == "ALICE.1954", "ALICE.1954(2)", ID),
    # 4 hurricanes with the name SUBTROP:UNNAMED.1974
    ID = ifelse(ID == "SUBTROP:UNNAMED.1974" & Month == "June", "SUBTROP:UNNAMED.1974(1)", ID),
    ID = ifelse(ID == "SUBTROP:UNNAMED.1974" & Month == "July", "SUBTROP:UNNAMED.1974(2)", ID),
    ID = ifelse(ID == "SUBTROP:UNNAMED.1974" & Month == "August", "SUBTROP:UNNAMED.1974(3)", ID),
    ID = ifelse(ID == "SUBTROP:UNNAMED.1974", "SUBTROP:UNNAMED.1974(4)", ID),
    # 2 hurricanes with the name SUBTROP:UNNAMED.1976
    ID = ifelse(ID == "SUBTROP:UNNAMED.1976" & Month == "May", "SUBTROP:UNNAMED.1976(1)", ID),
    ID = ifelse(ID == "SUBTROP:UNNAMED.1976", "SUBTROP:UNNAMED.1976(2)", ID)
  )
```

**Given code for visualization**

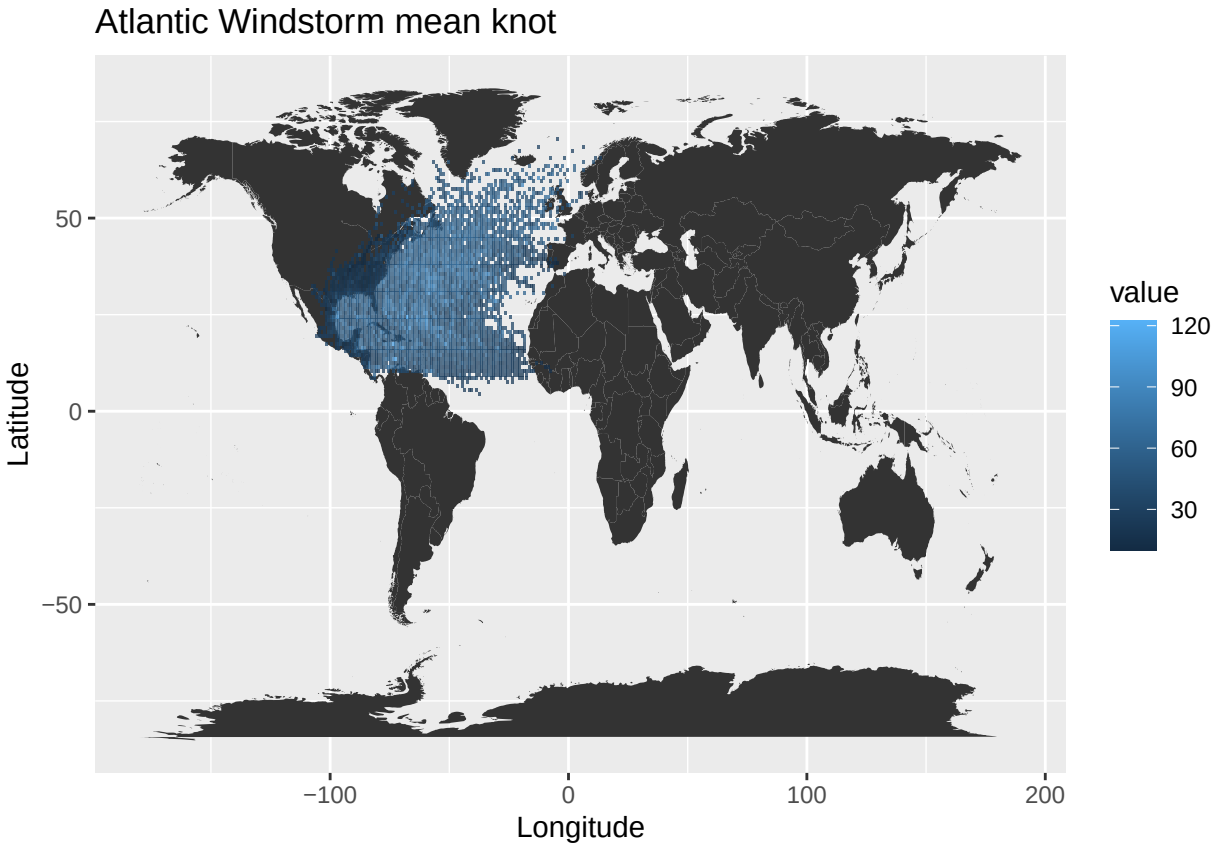
```
ggplot(data=dt, aes(x = Longitude, y = Latitude)) +
  stat_summary_2d(data = dt, aes(x = Longitude, y = Latitude, z = Wind.kt), fun = median, binwidth = c(
```




```
dt <- as.data.table(dt)
summary(dt)
```

```
##       ID                 Season        Month              Nature
##  Length:22038       Min.   :1950   Length:22038       Length:22038
##  Class :character   1st Qu.:1969   Class :character   Class :character
##  Mode  :character   Median :1989   Mode  :character   Mode  :character
##                     Mean   :1986
##                     3rd Qu.:2003
##                     Max.   :2013
##      time               Latitude       Longitude          Wind.kt
##  Length:22038       Min.   : 5.00   Min.   :-107.70   Min.   : 10.00
##  Class :character   1st Qu.:18.70   1st Qu.: -78.70   1st Qu.: 30.00
##  Mode  :character   Median :26.50   Median : -64.05   Median : 45.00
##                     Mean   :26.99   Mean   : -62.91   Mean   : 52.28
##                     3rd Qu.:33.60   3rd Qu.: -48.60   3rd Qu.: 65.00
##                     Max.   :70.70   Max.   :  13.50   Max.   :165.00
```

```
map <- ggplot(data = dt, aes(x = Longitude, y = Latitude)) +
  geom_polygon(data = map_data(map = 'world'), aes(x = long, y = lat, group = group))
map +
  stat_summary_2d(data = dt, aes(x = Longitude, y = Latitude, z = dt$Wind.kt), fun = median, binwidth =
  ggtitle(paste0("Atlantic Windstorm mean knot"))
```




```
map <- ggplot(dt, aes(x = Longitude, y = Latitude, group = ID)) +
  geom_polygon(data = map_data("world"),
               aes(x = long, y = lat, group = group),
               fill = "gray25", colour = "gray10", size = 0.2) +
  geom_path(data = dt, aes(group = ID, colour = Wind.kt), size = 0.5) +
  xlim(-138, -20) + ylim(3, 55) +
  labs(x = "", y = "", colour = "Wind \n(knots)") +
  theme(panel.background = element_rect(fill = "gray10", colour = "gray30"),
        axis.text.x = element_blank(), axis.text.y = element_blank(),
        axis.ticks = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
```

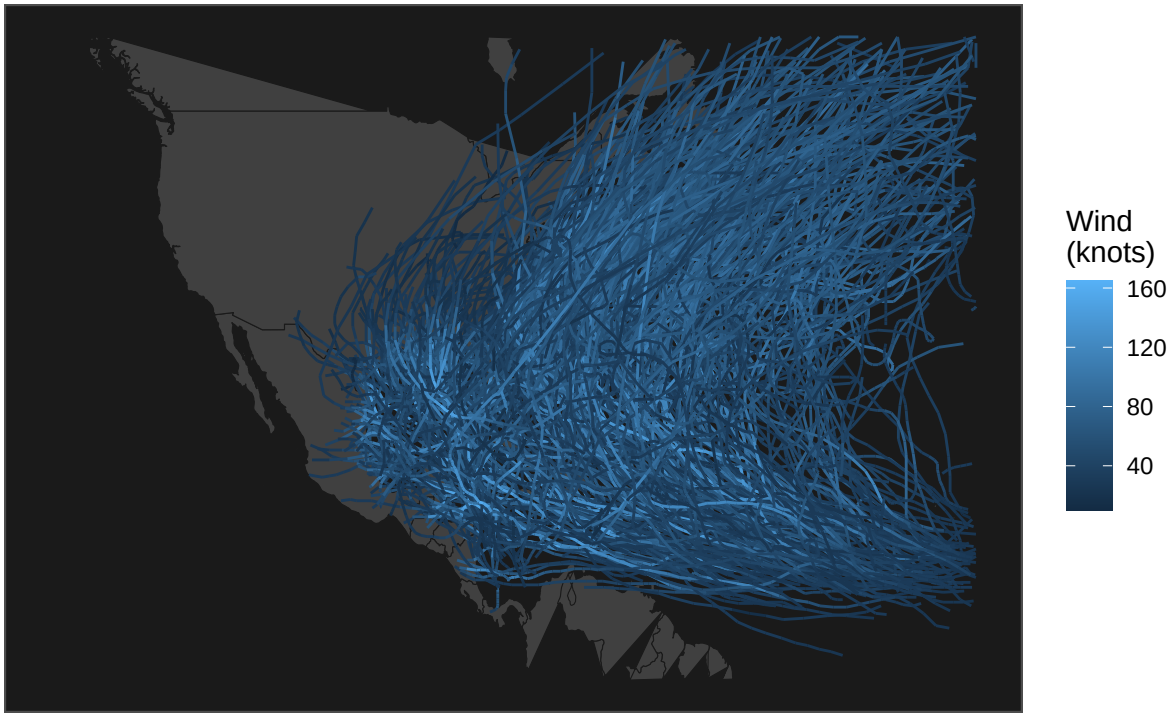
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
seasonrange <- paste(range(dt[, Season]), collapse=" - ")

map + ggtitle(paste("Atlantic named Windstorm Trajectories (",
                    seasonrange, ")\n"))
```

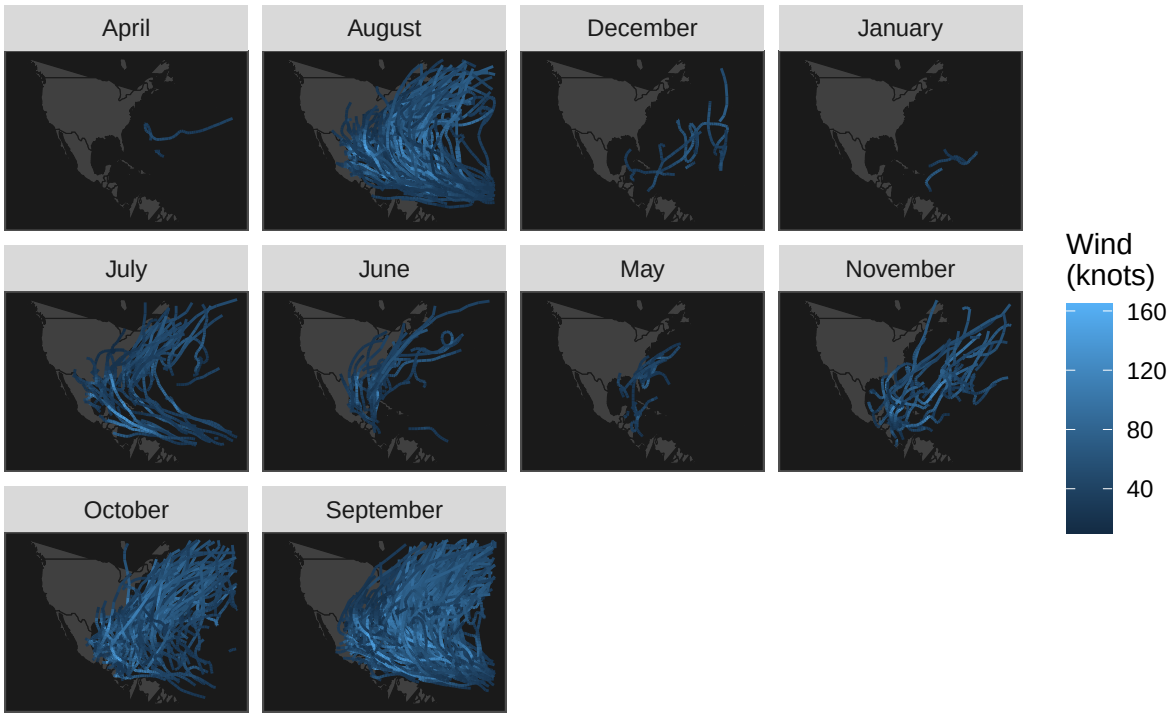
```
## Warning: Removed 522 rows containing missing values (‘geom_path()‘).
```

Atlantic named Windstorm Trajectories ( 1950 − 2013 )




```
mapMonth <- map + facet_wrap(~ Month) +
  ggtitle(paste("Atlantic named Windstorm Trajectories by Month (",
                seasonrange, ")\n"))
mapMonth
```

```
## Warning: Removed 522 rows containing missing values (‘geom_path()‘).
```




**Data wrangling**

```
hurricane_df <- origin_df %>%
  mutate(
    Month = as.factor(Month), # April-January
    Nature = as.factor(Nature), # TS,ET,DS,SS,NR
    # note: one hurricane can have multiple natures throughout its life
    time = gsub("[()]", "", time),
    time = paste0(ifelse(substr(time, 1, 2) > 23, "19", "20"), time),
    time = as.POSIXct(time, format = "%Y-%m-%d %H:%M:%S"),
    hour = substr(time, 12, 19)
  ) %>%
  # remove data not at six-hour time intervals. (613 observations)
  filter(hour %in% c("00:00:00", "06:00:00", "12:00:00", "18:00:00")) %>%
  select(-hour)

# remove hurricanes that has only 2 (<3) observations (change the threshold if you wish)
few_id <- hurricane_df %>%
  group_by(ID) %>%
  summarize(obs = n()) %>%
  filter(obs < 3) %>%
  .$ID
few_id
```

```
## [1] "BARBARA.2013"     "SIMONE.1961"      "TEN:UNNAMED.1988"
```

```
hurricane_df <- hurricane_df %>% filter(!(ID %in% few_id)) # remove 3 hurricanes

# check if any missing hours (e.g., 0 directly to 12, missing 6)
issue_id <- hurricane_df %>%
  group_by(ID) %>%
  summarize(
    obs = n(),
    obs2 = as.integer(difftime(max(time), min(time), units = "hours"))/6 + 1,
    diff = obs - obs2
  ) %>%
  filter(abs(diff) > 0.167) %>% # some = 1/6. may due to DST time-shift
  .$ID
issue_id
```

```
## [1] "ALICE.1954"           "SUBTROP:UNNAMED.1974" "SUBTROP:UNNAMED.1976"
```

```
# no such issue. but identify a new issue:
# several different hurricanes with the same name. (3 such hurricane names)
# e.g.,
# https://en.wikipedia.org/wiki/Hurricane_Alice_(December_1954)
# https://en.wikipedia.org/wiki/Hurricane_Alice_(June_1954)
origin_df %>%
  filter(ID %in% issue_id) %>%
  select(ID, Month, time)
```

```
##                       ID     Month                time
## 1             ALICE.1954      June (54-06-24 12:00:00)
## 2             ALICE.1954      June (54-06-24 18:00:00)
## 3             ALICE.1954      June (54-06-25 00:00:00)
## 4             ALICE.1954      June (54-06-25 06:00:00)
## 5             ALICE.1954      June (54-06-25 12:00:00)
## 6             ALICE.1954      June (54-06-25 18:00:00)
## 7             ALICE.1954      June (54-06-26 00:00:00)
## 8             ALICE.1954      June (54-06-26 06:00:00)
## 9             ALICE.1954      June (54-06-26 12:00:00)
## 10            ALICE.1954      June (54-06-26 18:00:00)
## 11            ALICE.1954  December (54-12-30 06:00:00)
## 12            ALICE.1954  December (54-12-30 12:00:00)
## 13            ALICE.1954  December (54-12-30 18:00:00)
## 14            ALICE.1954  December (54-12-31 00:00:00)
## 15            ALICE.1954  December (54-12-31 06:00:00)
## 16            ALICE.1954  December (54-12-31 12:00:00)
## 17            ALICE.1954  December (54-12-31 18:00:00)
## 18            ALICE.1954   January (55-01-01 00:00:00)
## 19            ALICE.1954   January (55-01-01 06:00:00)
## 20            ALICE.1954   January (55-01-01 12:00:00)
## 21            ALICE.1954   January (55-01-01 18:00:00)
## 22            ALICE.1954   January (55-01-02 00:00:00)
## 23            ALICE.1954   January (55-01-02 06:00:00)
## 24            ALICE.1954   January (55-01-02 12:00:00)
## 25            ALICE.1954   January (55-01-02 18:00:00)
```

```
## 26            ALICE.1954   January (55-01-03 00:00:00)
## 27            ALICE.1954   January (55-01-03 06:00:00)
## 28            ALICE.1954   January (55-01-03 12:00:00)
## 29            ALICE.1954   January (55-01-03 18:00:00)
## 30            ALICE.1954   January (55-01-04 00:00:00)
## 31            ALICE.1954   January (55-01-04 06:00:00)
## 32            ALICE.1954   January (55-01-04 12:00:00)
## 33            ALICE.1954   January (55-01-04 18:00:00)
## 34            ALICE.1954   January (55-01-05 00:00:00)
## 35            ALICE.1954   January (55-01-05 06:00:00)
## 36            ALICE.1954   January (55-01-05 12:00:00)
## 37            ALICE.1954   January (55-01-05 18:00:00)
## 38            ALICE.1954   January (55-01-06 00:00:00)
## 39            ALICE.1954   January (55-01-06 06:00:00)
## 40  SUBTROP:UNNAMED.1974      June (74-06-24 18:00:00)
## 41  SUBTROP:UNNAMED.1974      June (74-06-25 00:00:00)
## 42  SUBTROP:UNNAMED.1974      June (74-06-25 06:00:00)
## 43  SUBTROP:UNNAMED.1974      June (74-06-25 12:00:00)
## 44  SUBTROP:UNNAMED.1974      June (74-06-25 18:00:00)
## 45  SUBTROP:UNNAMED.1974      June (74-06-26 00:00:00)
## 46  SUBTROP:UNNAMED.1974      July (74-07-16 00:00:00)
## 47  SUBTROP:UNNAMED.1974      July (74-07-16 06:00:00)
## 48  SUBTROP:UNNAMED.1974      July (74-07-16 12:00:00)
## 49  SUBTROP:UNNAMED.1974      July (74-07-16 18:00:00)
## 50  SUBTROP:UNNAMED.1974      July (74-07-17 00:00:00)
## 51  SUBTROP:UNNAMED.1974      July (74-07-17 06:00:00)
## 52  SUBTROP:UNNAMED.1974      July (74-07-17 12:00:00)
## 53  SUBTROP:UNNAMED.1974      July (74-07-17 18:00:00)
## 54  SUBTROP:UNNAMED.1974      July (74-07-18 00:00:00)
## 55  SUBTROP:UNNAMED.1974      July (74-07-18 06:00:00)
## 56  SUBTROP:UNNAMED.1974      July (74-07-18 12:00:00)
## 57  SUBTROP:UNNAMED.1974      July (74-07-18 18:00:00)
## 58  SUBTROP:UNNAMED.1974      July (74-07-19 00:00:00)
## 59  SUBTROP:UNNAMED.1974      July (74-07-19 06:00:00)
## 60  SUBTROP:UNNAMED.1974      July (74-07-19 12:00:00)
## 61  SUBTROP:UNNAMED.1974      July (74-07-19 18:00:00)
## 62  SUBTROP:UNNAMED.1974      July (74-07-20 00:00:00)
## 63  SUBTROP:UNNAMED.1974      July (74-07-20 06:00:00)
## 64  SUBTROP:UNNAMED.1974      July (74-07-20 12:00:00)
## 65  SUBTROP:UNNAMED.1974    August (74-08-10 12:00:00)
## 66  SUBTROP:UNNAMED.1974    August (74-08-10 18:00:00)
## 67  SUBTROP:UNNAMED.1974    August (74-08-11 00:00:00)
## 68  SUBTROP:UNNAMED.1974    August (74-08-11 06:00:00)
## 69  SUBTROP:UNNAMED.1974    August (74-08-11 12:00:00)
## 70  SUBTROP:UNNAMED.1974    August (74-08-11 18:00:00)
## 71  SUBTROP:UNNAMED.1974    August (74-08-12 00:00:00)
## 72  SUBTROP:UNNAMED.1974    August (74-08-12 06:00:00)
## 73  SUBTROP:UNNAMED.1974    August (74-08-12 12:00:00)
## 74  SUBTROP:UNNAMED.1974    August (74-08-12 18:00:00)
## 75  SUBTROP:UNNAMED.1974    August (74-08-13 00:00:00)
## 76  SUBTROP:UNNAMED.1974    August (74-08-13 06:00:00)
## 77  SUBTROP:UNNAMED.1974    August (74-08-13 12:00:00)
## 78  SUBTROP:UNNAMED.1974    August (74-08-13 18:00:00)
## 79  SUBTROP:UNNAMED.1974    August (74-08-14 00:00:00)
```

```
## 80  SUBTROP:UNNAMED.1974    August (74-08-14 06:00:00)
## 81  SUBTROP:UNNAMED.1974    August (74-08-14 12:00:00)
## 82  SUBTROP:UNNAMED.1974    August (74-08-14 18:00:00)
## 83  SUBTROP:UNNAMED.1974    August (74-08-15 00:00:00)
## 84  SUBTROP:UNNAMED.1974   October (74-10-04 00:00:00)
## 85  SUBTROP:UNNAMED.1974   October (74-10-04 06:00:00)
## 86  SUBTROP:UNNAMED.1974   October (74-10-04 12:00:00)
## 87  SUBTROP:UNNAMED.1974   October (74-10-04 18:00:00)
## 88  SUBTROP:UNNAMED.1974   October (74-10-05 00:00:00)
## 89  SUBTROP:UNNAMED.1974   October (74-10-05 06:00:00)
## 90  SUBTROP:UNNAMED.1974   October (74-10-05 12:00:00)
## 91  SUBTROP:UNNAMED.1974   October (74-10-05 18:00:00)
## 92  SUBTROP:UNNAMED.1974   October (74-10-06 00:00:00)
## 93  SUBTROP:UNNAMED.1974   October (74-10-06 06:00:00)
## 94  SUBTROP:UNNAMED.1974   October (74-10-06 12:00:00)
## 95  SUBTROP:UNNAMED.1974   October (74-10-06 18:00:00)
## 96  SUBTROP:UNNAMED.1974   October (74-10-07 00:00:00)
## 97  SUBTROP:UNNAMED.1974   October (74-10-07 06:00:00)
## 98  SUBTROP:UNNAMED.1974   October (74-10-07 12:00:00)
## 99  SUBTROP:UNNAMED.1974   October (74-10-07 18:00:00)
## 100 SUBTROP:UNNAMED.1974   October (74-10-08 00:00:00)
## 101 SUBTROP:UNNAMED.1974   October (74-10-08 06:00:00)
## 102 SUBTROP:UNNAMED.1974   October (74-10-08 12:00:00)
## 103 SUBTROP:UNNAMED.1974   October (74-10-08 18:00:00)
## 104 SUBTROP:UNNAMED.1974   October (74-10-09 00:00:00)
## 105 SUBTROP:UNNAMED.1976       May (76-05-21 12:00:00)
## 106 SUBTROP:UNNAMED.1976       May (76-05-21 18:00:00)
## 107 SUBTROP:UNNAMED.1976       May (76-05-22 00:00:00)
## 108 SUBTROP:UNNAMED.1976       May (76-05-22 06:00:00)
## 109 SUBTROP:UNNAMED.1976       May (76-05-22 12:00:00)
## 110 SUBTROP:UNNAMED.1976       May (76-05-22 18:00:00)
## 111 SUBTROP:UNNAMED.1976       May (76-05-23 00:00:00)
## 112 SUBTROP:UNNAMED.1976       May (76-05-23 06:00:00)
## 113 SUBTROP:UNNAMED.1976       May (76-05-23 12:00:00)
## 114 SUBTROP:UNNAMED.1976       May (76-05-23 18:00:00)
## 115 SUBTROP:UNNAMED.1976       May (76-05-24 00:00:00)
## 116 SUBTROP:UNNAMED.1976       May (76-05-24 06:00:00)
## 117 SUBTROP:UNNAMED.1976       May (76-05-24 12:00:00)
## 118 SUBTROP:UNNAMED.1976       May (76-05-24 18:00:00)
## 119 SUBTROP:UNNAMED.1976       May (76-05-25 00:00:00)
## 120 SUBTROP:UNNAMED.1976       May (76-05-25 06:00:00)
## 121 SUBTROP:UNNAMED.1976       May (76-05-25 12:00:00)
## 122 SUBTROP:UNNAMED.1976       May (76-05-25 18:00:00)
## 123 SUBTROP:UNNAMED.1976 September (76-09-13 12:00:00)
## 124 SUBTROP:UNNAMED.1976 September (76-09-13 18:00:00)
## 125 SUBTROP:UNNAMED.1976 September (76-09-14 00:00:00)
## 126 SUBTROP:UNNAMED.1976 September (76-09-14 06:00:00)
## 127 SUBTROP:UNNAMED.1976 September (76-09-14 12:00:00)
## 128 SUBTROP:UNNAMED.1976 September (76-09-14 18:00:00)
## 129 SUBTROP:UNNAMED.1976 September (76-09-15 00:00:00)
## 130 SUBTROP:UNNAMED.1976 September (76-09-15 06:00:00)
## 131 SUBTROP:UNNAMED.1976 September (76-09-15 12:00:00)
## 132 SUBTROP:UNNAMED.1976 September (76-09-15 18:00:00)
## 133 SUBTROP:UNNAMED.1976 September (76-09-16 00:00:00)
```

```
## 134 SUBTROP:UNNAMED.1976 September (76-09-16 06:00:00)
## 135 SUBTROP:UNNAMED.1976 September (76-09-16 12:00:00)
## 136 SUBTROP:UNNAMED.1976 September (76-09-16 18:00:00)
## 137 SUBTROP:UNNAMED.1976 September (76-09-17 00:00:00)
```

```
# manually correct those data
hurricane_df <-
  hurricane_df %>%
  mutate(
    # 2 hurricanes with the name ALICE.1954
    ID = ifelse(ID == "ALICE.1954" & Month == "June", "ALICE.1954(1)", ID),
    ID = ifelse(ID == "ALICE.1954", "ALICE.1954(2)", ID),
    # 4 hurricanes with the name SUBTROP:UNNAMED.1974
    ID = ifelse(ID == "SUBTROP:UNNAMED.1974" & Month == "June", "SUBTROP:UNNAMED.1974(1)", ID),
    ID = ifelse(ID == "SUBTROP:UNNAMED.1974" & Month == "July", "SUBTROP:UNNAMED.1974(2)", ID),
    ID = ifelse(ID == "SUBTROP:UNNAMED.1974" & Month == "August", "SUBTROP:UNNAMED.1974(3)", ID),
    ID = ifelse(ID == "SUBTROP:UNNAMED.1974", "SUBTROP:UNNAMED.1974(4)", ID),
    # 2 hurricanes with the name SUBTROP:UNNAMED.1976
    ID = ifelse(ID == "SUBTROP:UNNAMED.1976" & Month == "May", "SUBTROP:UNNAMED.1976(1)", ID),
    ID = ifelse(ID == "SUBTROP:UNNAMED.1976", "SUBTROP:UNNAMED.1976(2)", ID)
  )

# check again
issue_id <- hurricane_df %>%
  group_by(ID) %>%
  summarize(
    obs = n(),
    obs2 = as.integer(difftime(max(time), min(time), units = "hours"))/6 + 1,
    diff = obs - obs2
  ) %>%
  filter(abs(diff) > 0.167) %>%
  .$ID
issue_id # no such problem
```

```
## character(0)
```

```
summary(hurricane_df) # 21691 observations, 704 hurricanes
```

```
##       ID                Season           Month      Nature
##  Length:21691       Min.   :1950   September:8866   DS:  969
##  Class :character   1st Qu.:1969   August   :5127   ET: 2149
##  Mode  :character   Median :1989   October  :3730   NR:   96
##                     Mean   :1986   July     :1490   SS:  751
##                     3rd Qu.:2003   November :1047   TS:17726
##                     Max.   :2013   June     : 801
##                                    (Other)  : 630
##       time                          Latitude       Longitude
##  Min.   :1950-08-12 00:00:00.000   Min.   : 5.00   Min.   :-107.70
##  1st Qu.:1969-09-06 15:00:00.000   1st Qu.:18.60   1st Qu.: -78.30
##  Median :1989-09-02 12:00:00.000   Median :26.50   Median : -63.70
##  Mean   :1986-05-13 19:44:45.021   Mean   :27.01   Mean   : -62.63
##  3rd Qu.:2003-09-26 18:00:00.000   3rd Qu.:33.60   3rd Qu.: -48.40
##  Max.   :2013-11-23 06:00:00.000   Max.   :70.70   Max.   :  13.50
```

```
##
##     Wind.kt
##  Min.   : 10.00
##  1st Qu.: 30.00
##  Median : 45.00
##  Mean   : 51.99
##  3rd Qu.: 65.00
##  Max.   :165.00
##
```

```
diff_df <-
  hurricane_df %>%
  group_by(ID) %>%
  mutate(
    # diff between t & t-6
    lat_diff = Latitude - lag(Latitude),
    long_diff = Longitude - lag(Longitude),
    wind_diff = Wind.kt - lag(Wind.kt),
    # time: (t - t0)/6, where t0 is the first observation time.
    # you can change this value (e.g. +1, or -1) as you wish
    time = round(difftime(time, min(time), units = "hours")/6) %>% as.integer
  ) %>%
  drop_na %>%
  select(ID, lat_diff, long_diff, wind_diff, time)

summary(diff_df)
```

```
##       ID              lat_diff         long_diff         wind_diff
##  Length:20987       Min.   :-2.9000   Min.   :-5.6000   Min.   :-65.0000
##  Class :character   1st Qu.: 0.1000   1st Qu.:-0.9000   1st Qu.:  0.0000
##  Mode  :character   Median : 0.5000   Median :-0.3000   Median :  0.0000
##                     Mean   : 0.5622   Mean   : 0.0275   Mean   :  0.1584
##                     3rd Qu.: 0.9000   3rd Qu.: 0.7000   3rd Qu.:  5.0000
##                     Max.   : 5.5000   Max.   :11.0000   Max.   : 55.0000
##       time
##  Min.   :  1.00
##  1st Qu.:  8.00
##  Median : 17.00
##  Mean   : 20.73
##  3rd Qu.: 30.00
##  Max.   :117.00
```