

EDA

Youlan Shen

2023-04-26

Data Exploratory Analysis

hurrican703.csv collected the track data of 703 hurricanes in the North Atlantic area since 1950. For all the storms, their location (longitude & latitude) and maximum wind speed were recorded every 6 hours. The data includes the following variables

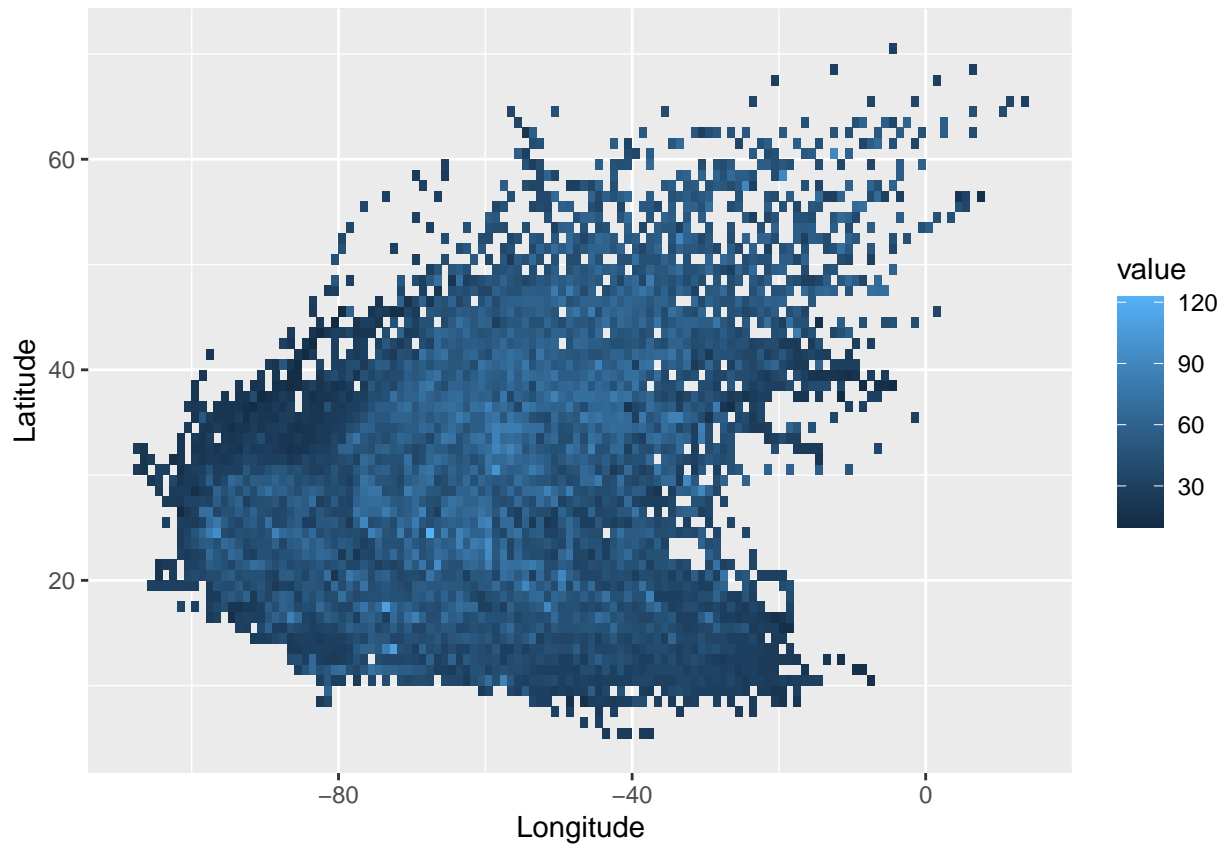
1. **ID**: ID of the hurricanes
2. **Season**: In which the hurricane occurred
3. **Month**: In which the hurricane occurred
4. **Nature**: Nature of the hurricane
 - ET: Extra Tropical
 - DS: Disturbance
 - NR: Not Rated
 - SS: Sub Tropical
 - TS: Tropical Storm
5. **time**: dates and time of the record
6. **Latitude** and **Longitude**: The location of a hurricane check point
7. **Wind.kt** Maximum wind speed (in Knot) at each check point

```
# library all packages that we need at the beginning
library(tidyverse)
library(dplyr)
library(readxl)
library(car)
library(gtsummary)
library(corrplot)
library(caret)
```

Summary table and Plot for hurricane data

```
library(ggplot2)
dt = read.csv("hurrican703.csv")
ggplot(data=dt, aes(x = Longitude, y = Latitude)) +
  stat_summary_2d(data = dt, aes(x = Longitude, y = Latitude, z = dt$Wind.kt),
    fun = median, binwidth = c(1, 1), show.legend = TRUE)
```

```
## Warning: Use of 'dt$Wind.kt' is discouraged.
## i Use 'Wind.kt' instead.
```



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

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##   between, first, last

## The following object is masked from 'package:purrr':
##
##   transpose
```

```
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
```

Hurricane data on World Map

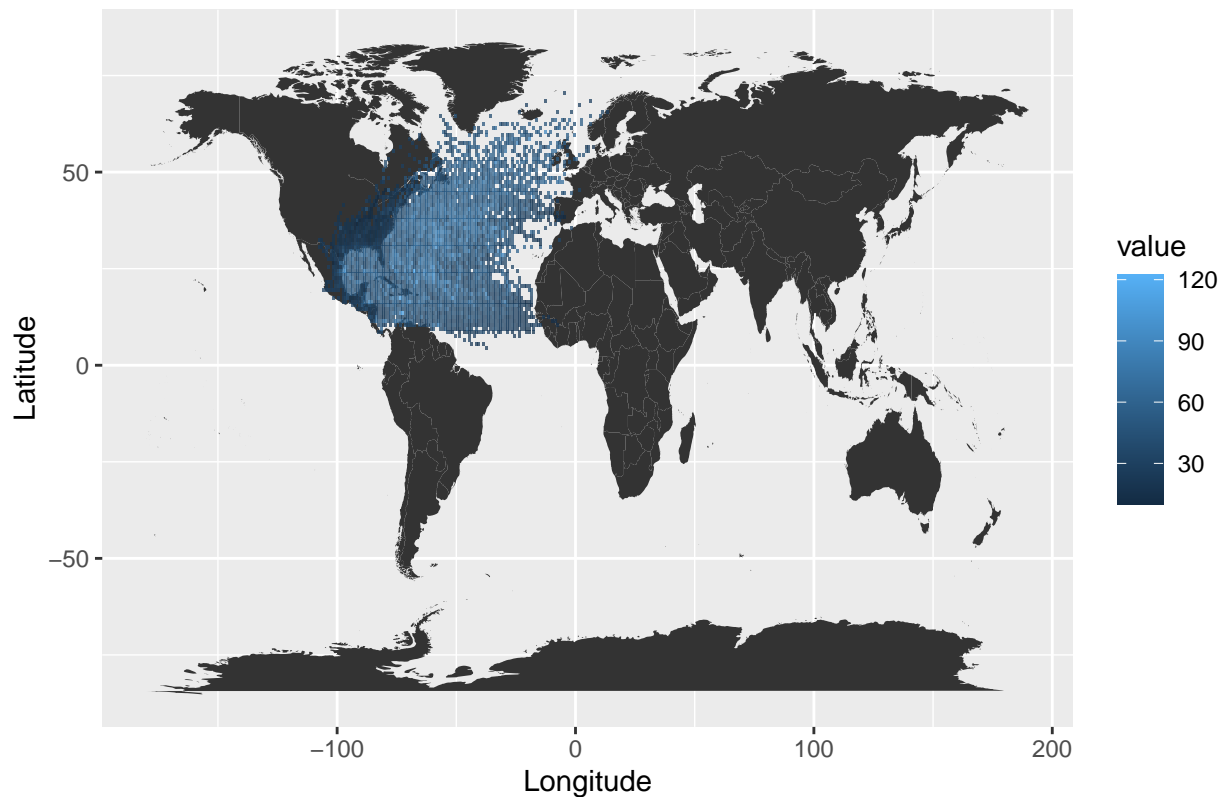
```
library(maps)
```

```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##      map
```

```
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 = c(1, 1),
    show.legend = TRUE, alpha = 0.75) +
  ggtitle(paste0("Atlantic Windstorm mean knot"))
```

Atlantic Windstorm mean knot



Track of Each Hurricane on Map

```
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())
```

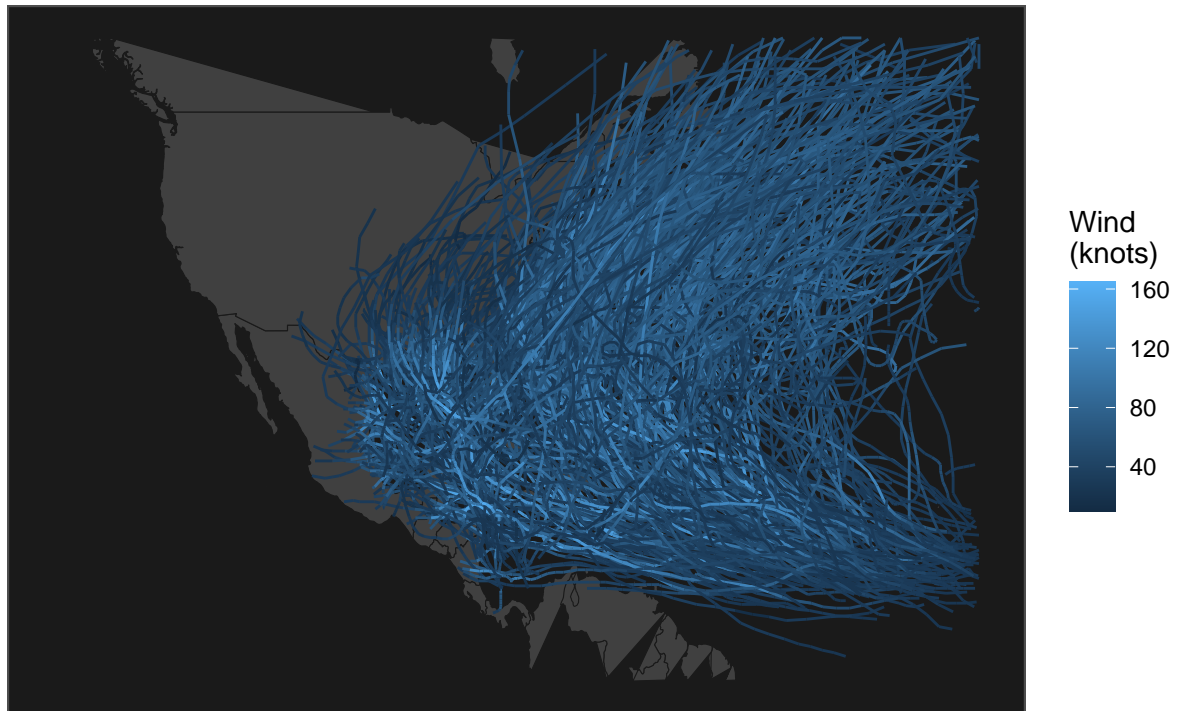
```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
```

```
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)

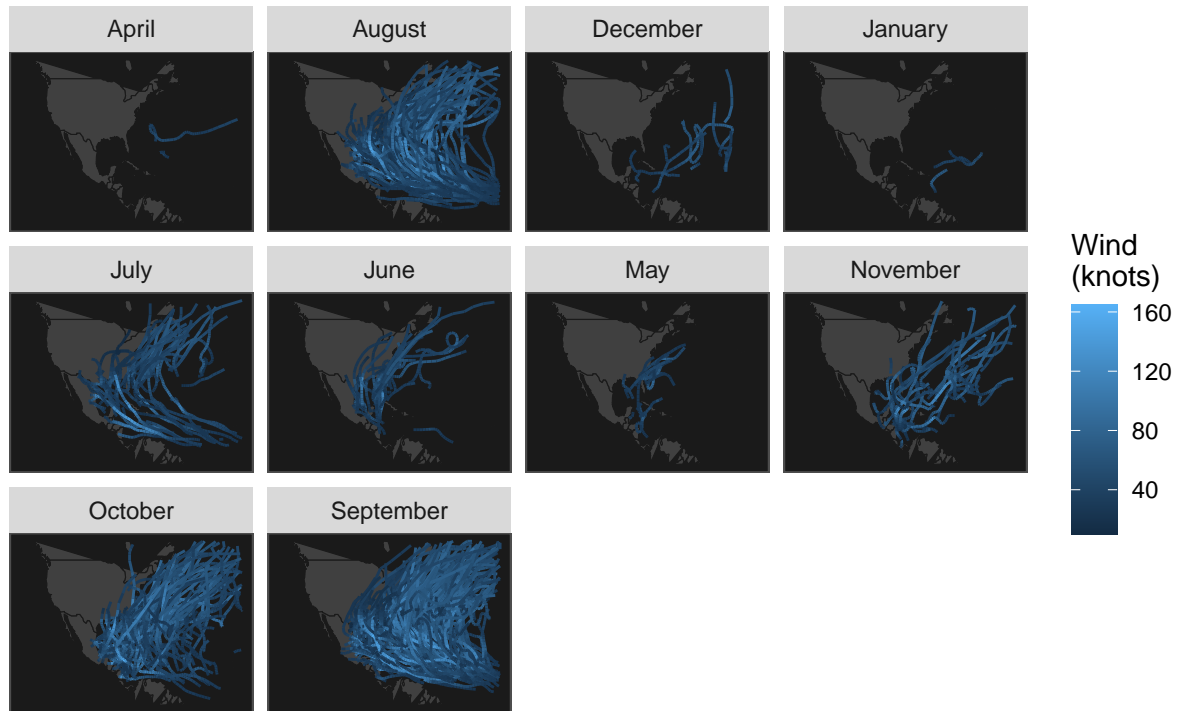


Track of Each Hurricane by Month on Map

```
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()’).
```

Atlantic named Windstorm Trajectories by Month (1950 – 2013)



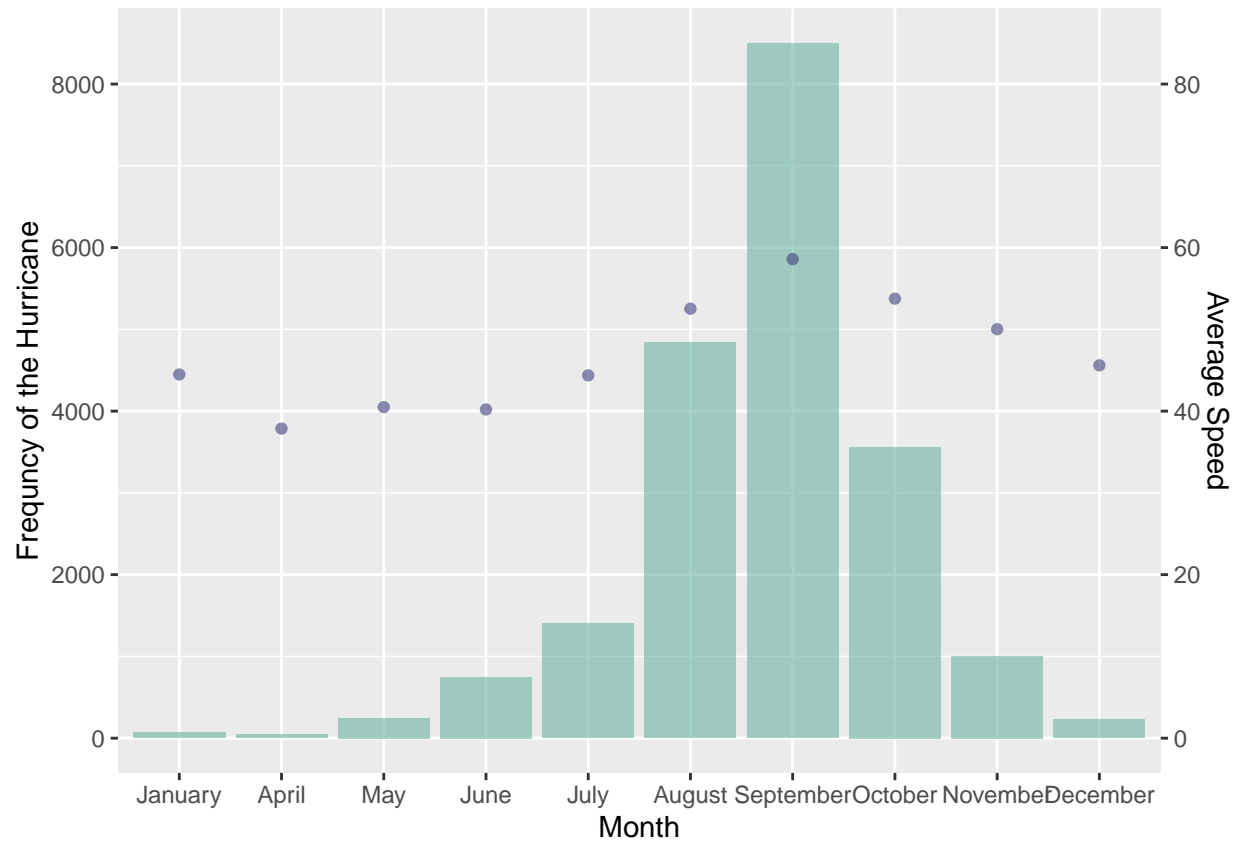
Others

```
# read in data from CSV file
hurricane <- read.csv("hurrican703.csv")
# tidy data on date
hurricane <- as_tibble(hurricane) %>%
  separate(time, into = c("Date", "Hour"), sep = " ") %>%
  mutate(Hour = ifelse(Hour == "00:00:00", 0,
                      ifelse(Hour == "06:00:00", 6,
                             ifelse(Hour == "12:00:00", 12, 18))),
         Date = str_remove(Date, "\\("),
         Date = yday(Date),
         Month = factor(Month, levels = month.name))
# tidy data on latitude longitude wind_kt
hurricane <- hurricane %>%
  group_by(ID) %>%
  mutate(Lat_change = Latitude - lag(Latitude, 1),
         Long_change = Longitude - lag(Longitude, 1),
         Wind_change = lag(Wind.kt, 1) - lag(Wind.kt, 2),
         Wind_prev = lag(Wind.kt, 1)) %>%
  na.omit()
hurricane
```

```
## # A tibble: 20,636 x 13
```

```
## # Groups:   ID [699]
##   ID      Season Month Nature  Date  Hour Latit~1 Longi~2 Wind.kt Lat_c~3 Long_~4
##   <chr>   <int> <fct> <chr>  <int> <dbl>   <dbl>   <dbl>   <int>   <dbl>   <dbl>
##  1 ABLE~   1950 Augu~ TS      224    12    18.2   -57.4     45    0.5    -1.10
##  2 ABLE~   1950 Augu~ TS      224    18    19     -58.6     50    0.800  -1.20
##  3 ABLE~   1950 Augu~ TS      225     0    20     -60      50    1      -1.40
##  4 ABLE~   1950 Augu~ TS      225     6    20.7   -61.1     50    0.700  -1.10
##  5 ABLE~   1950 Augu~ TS      225    12    21.3   -62.2     55    0.600  -1.10
##  6 ABLE~   1950 Augu~ TS      225    18    22     -63.2     55    0.700   -1
##  7 ABLE~   1950 Augu~ TS      226     0    22.7   -63.8     60    0.700 -0.600
##  8 ABLE~   1950 Augu~ TS      226     6    23.1   -64.6     60    0.400 -0.800
##  9 ABLE~   1950 Augu~ TS      226    12    23.4   -65.4     60    0.300 -0.800
## 10 ABLE~   1950 Augu~ TS      226    18    23.9   -66      65    0.5    -0.600
## # ... with 20,626 more rows, 2 more variables: Wind_change <int>,
## #   Wind_prev <int>, and abbreviated variable names 1: Latitude, 2: Longitude,
## #   3: Lat_change, 4: Long_change
```

```
hurricane %>%
  group_by(Month) %>%
  summarise(count = n(),
            Ave.Speed = mean(Wind.kt)) %>%
  ggplot(aes(x = Month)) +
  geom_col(aes(y = count), fill = "#69b3a2", alpha = 0.6) +
  geom_point(aes(y = Ave.Speed*100), color = "#404080", alpha = 0.6) +
  scale_y_continuous(
    name = "Frequency of the Hurricane",
    sec.axis = sec_axis(~.*0.01, name = "Average Speed"))
```



```
save(hurricane, file = "hurricane.RData")
```