Project1-Kaylei-Nilson

Kaylei Nilson-Pierce

10/1/2020

Video link: https://youtu.be/UAF352c1VIU

Import data: only first 16 columns

```
col_names <- c("SID", "SEASON", "NUMBER", "BASIN", "SUBBASIN", "NAME", "ISO_TIME", "NATURE", "LAT", "LO

col_types=c('character',"integer","integer", "character", "character", 'character', 'character','charac

dat <- read.csv(file='ibtracs.NA.list.v04r00.csv', skip=86272,
         colClass=col_types, stringsAsFactors = FALSE, na.strings = "MM")

colnames(dat) <- col_names

head(dat, 5)
```

```
##             SID SEASON NUMBER BASIN SUBBASIN      NAME            ISO_TIME
## 1 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 00:00:00
## 2 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 03:00:00
## 3 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 06:00:00
## 4 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 09:00:00
## 5 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 12:00:00
##   NATURE     LAT      LON WMO_WIND WMO_PRES WMO_AGENCY TRACK_TYPE DIST2LAND
## 1     TS 30.5000 -76.5000       20          hurdat_atl       main       390
## 2     TS 30.3428 -76.8528       NA                           main       382
## 3     TS 30.2000 -77.2000       25          hurdat_atl       main       371
## 4     TS 30.0845 -77.5549       NA                           main       332
## 5     TS 30.0000 -78.0000       25          hurdat_atl       main       294
##   LANDFALL
## 1      379
## 2      371
## 3      341
## 4      294
## 5      239
```

```
dat$MONTH <-as.numeric(substr(dat$ISO_TIME, 6, 7))
str(dat, vec.len = 1)
```

Add month column

```
## 'data.frame':    36069 obs. of  17 variables:
##  $ SID       : chr  "1980199N31284" ...
##  $ SEASON    : int  1980 1980 ...
##  $ NUMBER    : int  49 49 ...
##  $ BASIN     : chr  "NA" ...
##  $ SUBBASIN  : chr  "NA" ...
##  $ NAME      : chr  "NOT_NAMED" ...
##  $ ISO_TIME  : chr  "1980-07-17 00:00:00" ...
##  $ NATURE    : chr  "TS" ...
##  $ LAT       : num  30.5 ...
##  $ LON       : num  -76.5 ...
##  $ WMO_WIND  : int  20 NA ...
##  $ WMO_PRES  : chr  " " ...
##  $ WMO_AGENCY: chr  "hurdat_atl" ...
##  $ TRACK_TYPE: chr  "main" ...
##  $ DIST2LAND : int  390 382 ...
##  $ LANDFALL  : int  379 371 ...
##  $ MONTH     : num  7 7 ...
```

Manipulating data frames that will be used later on.

```
dat2 <- filter(dat, SEASON %in% 1980:2019)
head(dat2, 5)
```

We will only be exploring data in 1980:2019.

```
##             SID SEASON NUMBER BASIN SUBBASIN      NAME            ISO_TIME
## 1 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 00:00:00
## 2 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 03:00:00
## 3 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 06:00:00
## 4 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 09:00:00
## 5 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 12:00:00
##   NATURE     LAT      LON WMO_WIND WMO_PRES WMO_AGENCY TRACK_TYPE DIST2LAND
## 1     TS 30.5000 -76.5000       20          hurdat_atl       main       390
## 2     TS 30.3428 -76.8528       NA                           main       382
## 3     TS 30.2000 -77.2000       25          hurdat_atl       main       371
## 4     TS 30.0845 -77.5549       NA                           main       332
## 5     TS 30.0000 -78.0000       25          hurdat_atl       main       294
##   LANDFALL MONTH
## 1      379     7
## 2      371     7
## 3      341     7
## 4      294     7
## 5      239     7
```

Create column "HURRICANE" The WMO_WIND column provides its numbers in knots. According to the textbook and websites provided, when a storm's sustained wind speed reaches 74 mph or 64 kt it is considered a hurricane. Based off this knowledge, I created a column that returned T/F values if the recorded storm's WMO_WIND was recorded to be = or > than 64 kt.

```
hurricane <- mutate(dat2, HURRICANE = WMO_WIND >= 64)
head(hurricane,5)
```

```
##             SID SEASON NUMBER BASIN SUBBASIN      NAME            ISO_TIME
## 1 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 00:00:00
## 2 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 03:00:00
## 3 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 06:00:00
## 4 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 09:00:00
## 5 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 12:00:00
##   NATURE     LAT      LON WMO_WIND WMO_PRES WMO_AGENCY TRACK_TYPE DIST2LAND
## 1     TS 30.5000 -76.5000       20          hurdat_atl       main       390
## 2     TS 30.3428 -76.8528       NA                           main       382
## 3     TS 30.2000 -77.2000       25          hurdat_atl       main       371
## 4     TS 30.0845 -77.5549       NA                           main       332
## 5     TS 30.0000 -78.0000       25          hurdat_atl       main       294
##   LANDFALL MONTH HURRICANE
## 1      379     7     FALSE
## 2      371     7        NA
## 3      341     7     FALSE
## 4      294     7        NA
## 5      239     7     FALSE
```

Create column to identify the category of storms I used the function cut() to assign category values 0:5 to corresponding WMO_WIND values.

I refrenced the textbook, chapter 8, for the WMO_WIND cutoffs that I used to distinguish the different categories.

- category 1: 64-82 kt
- category 2: 83-95 kt
- category 3: 96-112 kt
- category 4: 113-136 kt
- category 5: 137 kt or higher

```
wind <- hurricane$WMO_WIND
category <- cut(wind, breaks= c(0, 63, 82, 95, 112, 136, Inf), labels= c("0","1","2","3","4","5"))
```

```
dat3 <- mutate(hurricane, CATEGORY= category)
head(dat3, 5)
```

```
##             SID SEASON NUMBER BASIN SUBBASIN      NAME            ISO_TIME
## 1 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 00:00:00
## 2 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 03:00:00
## 3 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 06:00:00
## 4 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 09:00:00
## 5 1980199N31284   1980     49    NA       NA NOT_NAMED 1980-07-17 12:00:00
##   NATURE     LAT      LON WMO_WIND WMO_PRES WMO_AGENCY TRACK_TYPE DIST2LAND
## 1     TS 30.5000 -76.5000       20          hurdat_atl       main       390
## 2     TS 30.3428 -76.8528       NA                           main       382
## 3     TS 30.2000 -77.2000       25          hurdat_atl       main       371
## 4     TS 30.0845 -77.5549       NA                           main       332
## 5     TS 30.0000 -78.0000       25          hurdat_atl       main       294
```

```
##   LANDFALL MONTH HURRICANE CATEGORY
## 1      379     7     FALSE        0
## 2      371     7        NA     <NA>
## 3      341     7     FALSE        0
## 4      294     7        NA     <NA>
## 5      239     7     FALSE        0
```

Exploratory analysis

BASIN and SUBBASIN

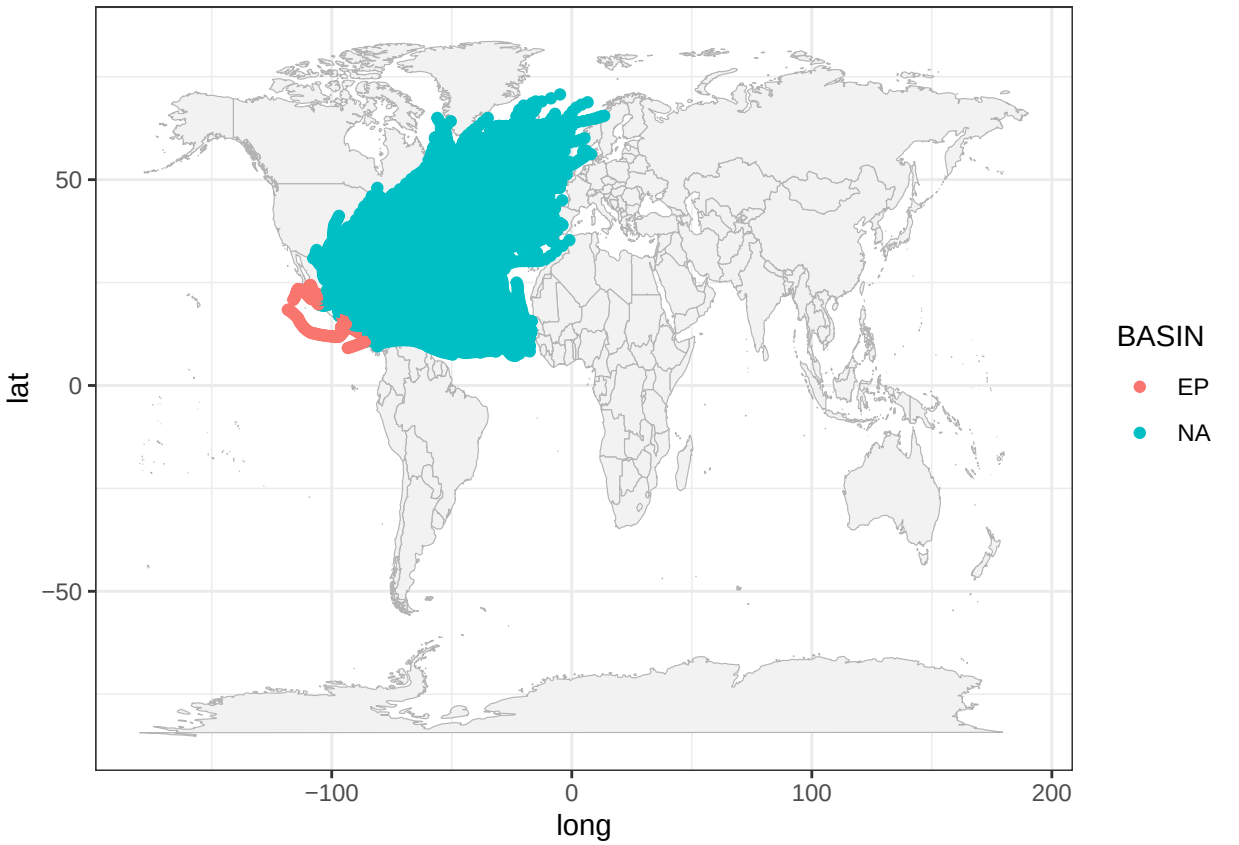
I have been asked to analyze hurricane data in the North Atlantic. When looking through the values in BASIN both NA and EP (Eastern Pacific) are logged. This lead me to have two theories:

1) the EP values were misrecorded since there's only a small minority of them and they are actually NA
2) the EP values do correspond to the Eastern Pacific and the data I am working with is not limited to the North Alantic as I had initially assumed.

In order to answer my question, I plotted the data set by latitude and longitude and colored by BASIN (NA and EP values) to see where they show up.

```
world_map <- map_data("world")

gg_world <- ggplot() +
  geom_polygon(data = world_map,
               aes(x = long, y = lat, group = group),
               fill = "gray95", colour = "gray70", size = 0.2) +
  theme_bw()
```

```
gg_world +
  geom_point(data = dat3, aes(x = LON, y = LAT, color = BASIN))
```



When I plotted the NA and EP values, they showed up in two differing areas and the EP values did in fact correspond to the Eastern Pacific.

After consulting my professor, I was told it was best to use all the data and not exclude EP values. Thus, the majority of the data points are from the Noth Atlantic, but some of the data points are from the Eastern Pacific.

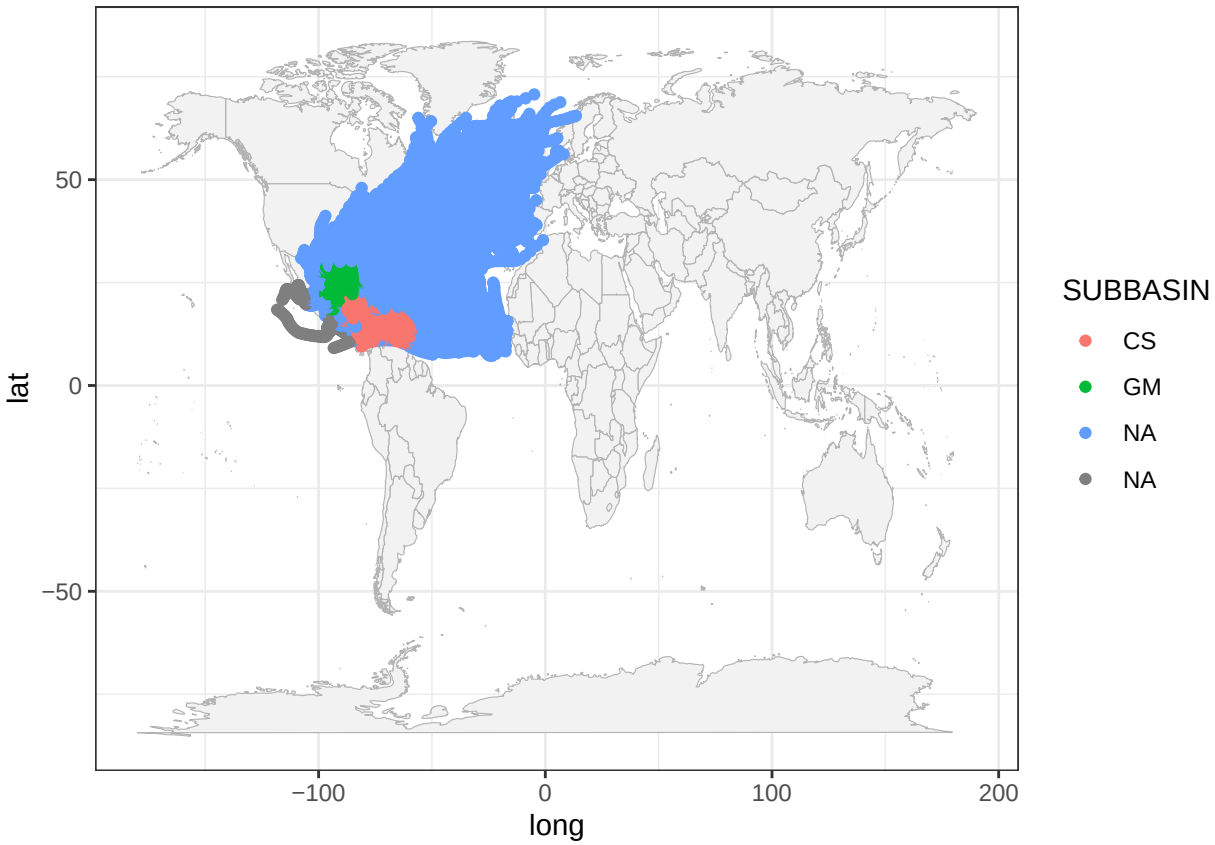
SUBBASIN

Another discrepency I notice is that SUBBASIN includes NA values, but the data dictionary says it should only include the following labels: MM, CS, GM, CP, BB, AS, WA, and EA. I have two theories:

1) the first is that the NA values in SUBBASIN mean it is located in the North Atlantic and NA was accidentally used to describe SUBBASIN column as it is used to describe in the BASIN column
2) the NA values stand for values that are "not applicable." In other words, the NA values do not represent the North Atlantic.

Again, I plotted the latitude and longitude and colored by SUBBASIN to see the results.

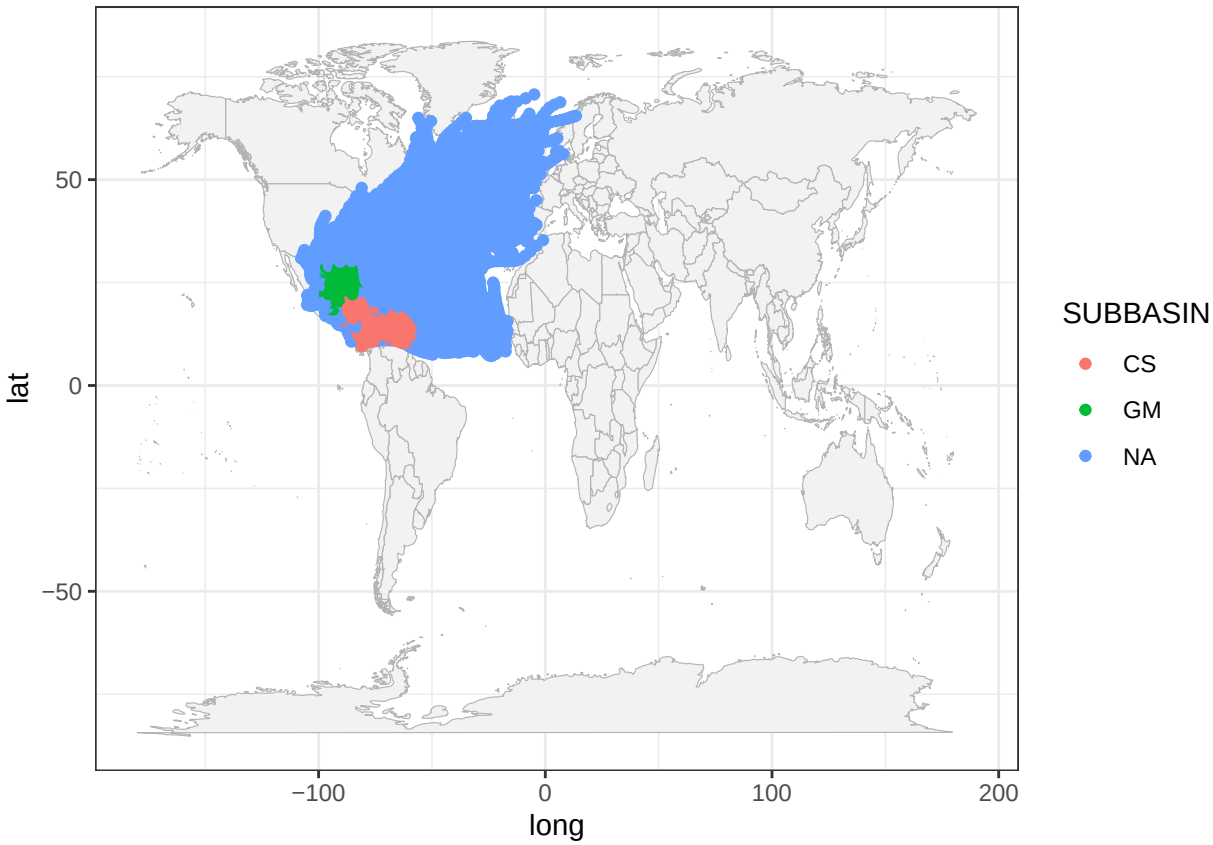
```
gg_world +
  geom_point(data = dat3, aes(x = LON, y = LAT, color = SUBBASIN))
```



The results of this map were supprising to me because they actually confirm that both of my theories are correct. The majority of the NA values really do correspond to the North Atlantic and were probably mistankingly used in the SUBBASIN column instead of the BASIN column. Notice that another group of NA values are gray and do not correspond to the North Atlantic. Interestingly, notice that the gray NA values correspond to the Eastern Pacific values (EP).

It would be difficult to filter out the gray NA values from the legitamite blue NA values in the SUBBASIN category since you can't really say exclude "NA." However, since the gray NA values are located in the EP, we can filter BASIN to exclude EP values then map again according to SUBBASIN and see if this got rid of the gray NA values for SUBBASIN.

```
NAONLY <- filter(dat3, BASIN != "EP")

gg_world +
  geom_point(data = NAONLY, aes(x = LON, y = LAT, color = SUBBASIN))
```

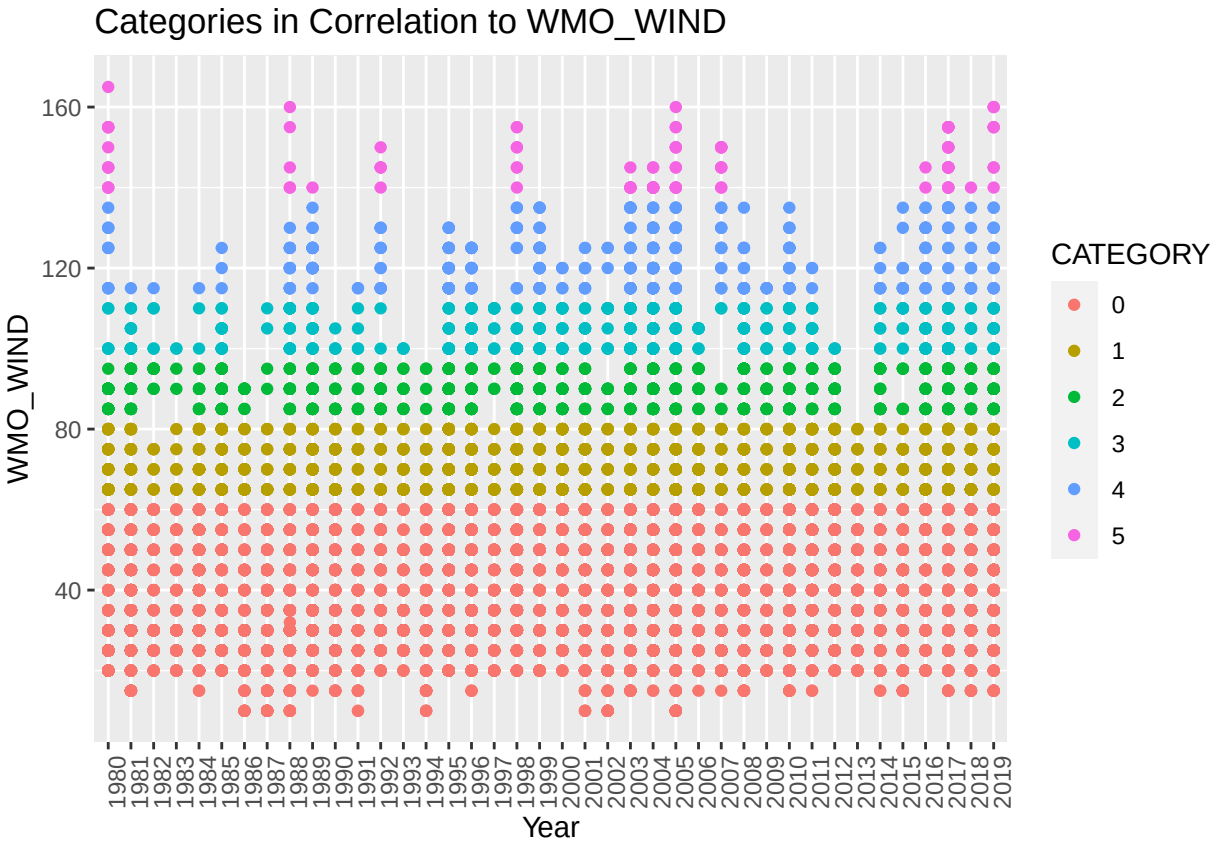


This map is accurate to the locations it is named as. If this were a legitimate report that depended on absolute accuracy, we would probably use the more filtered data set like this without the EP values.

SEASON, WMO_WIND, and CATEGORY

I am curious to see the distribution of WMO_WIND and CATEGORY over time to see if there is any pattern.

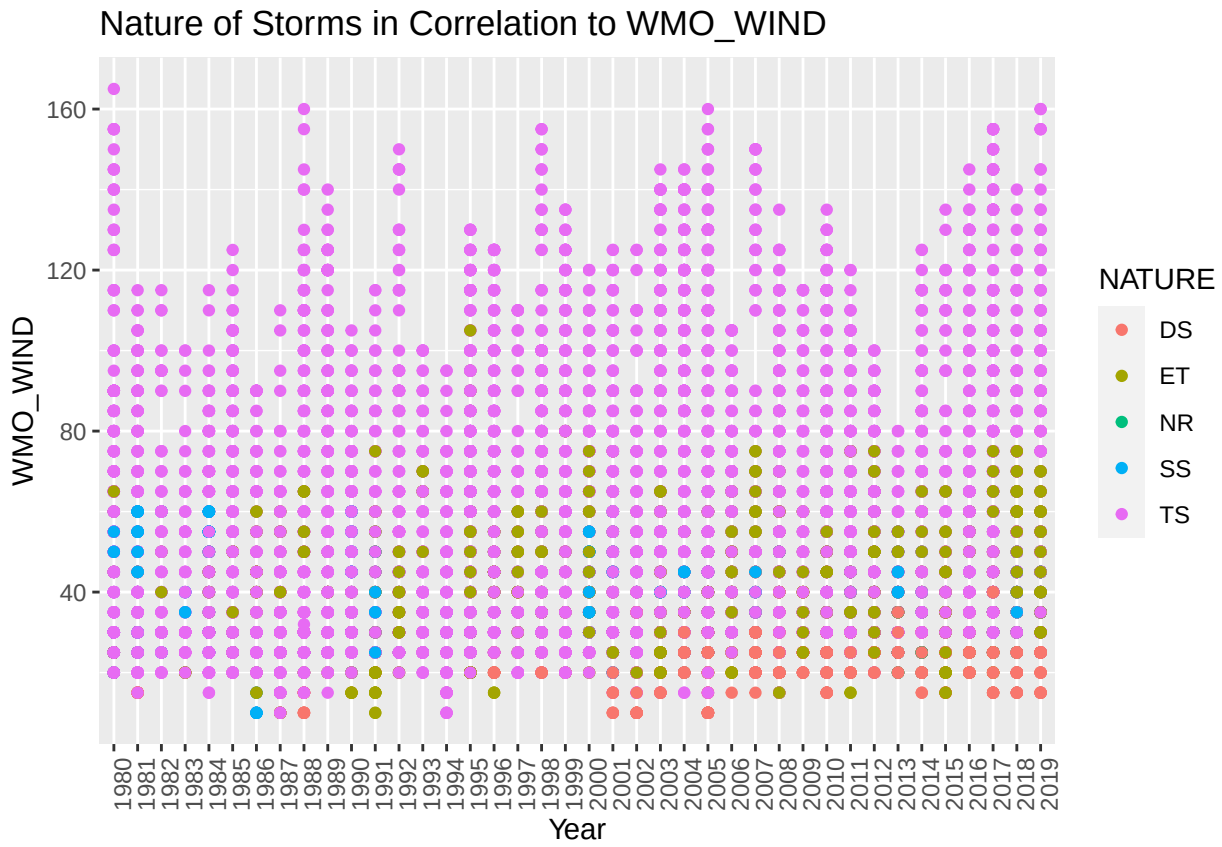
```
WindVals <- filter(dat3, WMO_WIND != "is.na")

ggplot(WindVals, aes(x=factor(SEASON), y=WMO_WIND)) +
  geom_point(aes(color=CATEGORY))+
  labs(x= "Year", title= "Categories in Correlation to WMO_WIND") +
  theme(axis.text.x = element_text(angle = 90))
```



SEASON, WMO_WIND and NATURE

After visualizing the correlation between WMO_WIND and CATEGORY, I was also curious to explore the NATURE of the storms and any correlation to WMO_WIND.

```
ggplot(WindVals, aes(x=factor(SEASON), y=WMO_WIND)) +
  geom_point(aes(color=NATURE))+ labs(x= "Year", title= "Nature of Storms in Correlation to WMO_WIND") +
  theme(axis.text.x = element_text(angle = 90))
```



REPORT

Claim A: Partial

Claim A) A typical hurricane season (during a calendar year) runs from June through November, but occasionally storms form outside those months.

This claim is a bit ambiguous in that it could be interpreted in two ways:

1) We look at hurricanes to answer the claim since the beginning of the statement says “a typical hurricane season”
2) We could look at storms because the second part of the claim states “occasionally storms form outside those months”

This contradiction of using a hurricane season, but also showing storms for the outside months led me to come up with my own solution. I decided I will plot storms and hurricanes to show the difference in distribution between the two so we can answer both parts of the claim.

First I will show a general visualization. I will seperate the storms and hurricanes, group by month, and calculate the mean number of occurances for each month over the time period 1980:2019.

```r
hurricane.na <- filter(dat3, HURRICANE != "is.na")

storms_hurricanes <- summarise(
  group_by(hurricane.na, HURRICANE, SEASON, MONTH),
  number = n()
)
```

```
## `summarise()` regrouping output by 'HURRICANE', 'SEASON' (override with `.groups` argument)
```

```r
head(storms_hurricanes,5)
```

```
## # A tibble: 5 x 4
## # Groups:   HURRICANE, SEASON [1]
##   HURRICANE SEASON MONTH number
##   <lgl>      <int> <dbl>  <int>
## 1 FALSE       1980     7     51
## 2 FALSE       1980     8     72
## 3 FALSE       1980     9     93
## 4 FALSE       1980    10     30
## 5 FALSE       1980    11     57
```

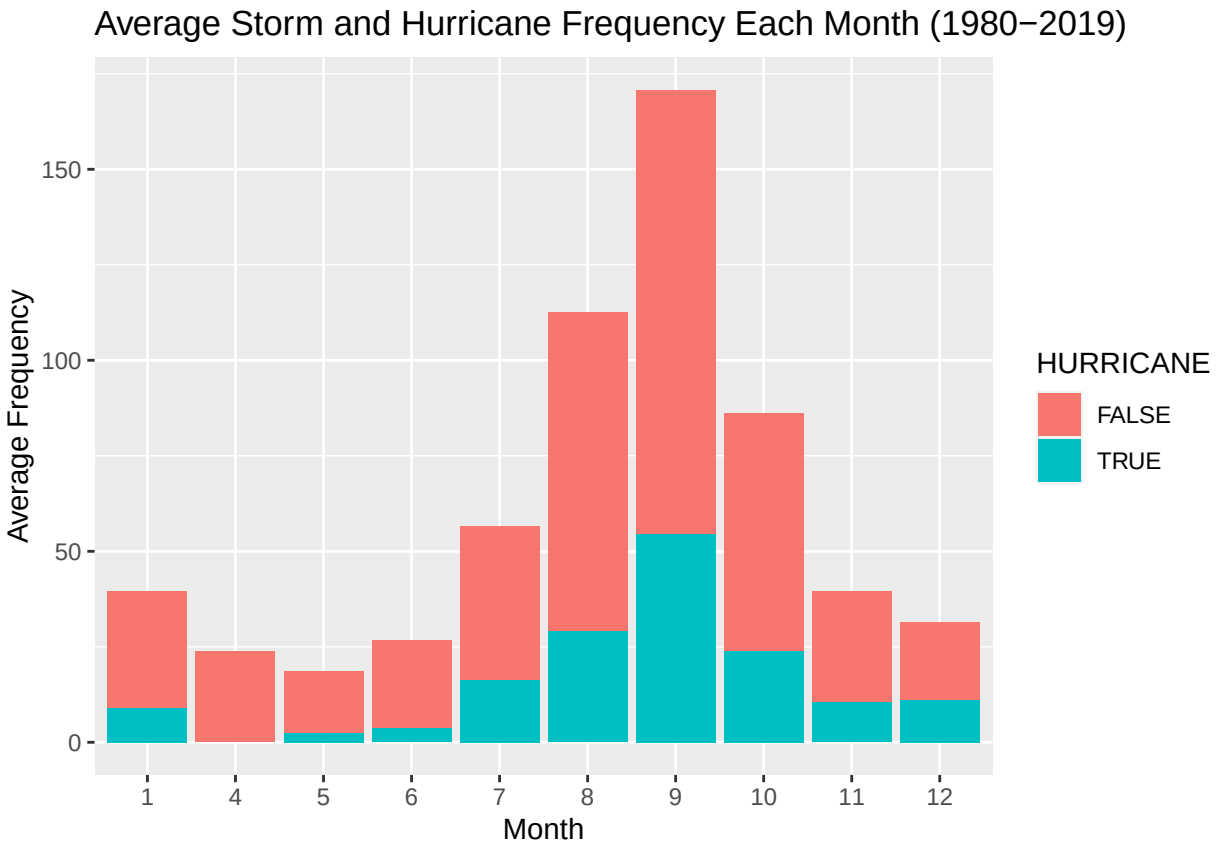
```r
mean_all <- summarise(
  group_by(storms_hurricanes, HURRICANE, MONTH),
  mean= mean(number))
```

```
## `summarise()` regrouping output by 'HURRICANE' (override with `.groups` argument)
```

```r
head(mean_all,10)
```

```
## # A tibble: 10 x 3
## # Groups:   HURRICANE [1]
##    HURRICANE MONTH  mean
##    <lgl>     <dbl> <dbl>
##  1 FALSE         1  30.5
##  2 FALSE         4  23.8
##  3 FALSE         5  16.2
##  4 FALSE         6  23.2
##  5 FALSE         7  40.1
##  6 FALSE         8  83.6
##  7 FALSE         9 116.
##  8 FALSE        10  62.1
##  9 FALSE        11  29.1
## 10 FALSE        12  20.5
```

```r
ggplot(mean_all, aes(x= factor(MONTH), y=mean, fill= HURRICANE)) +
  geom_bar(stat = "identity")+
  labs(x="Month", y= "Average Frequency", title= "Average Storm and Hurricane Frequency Each Month (198
```



This visualization shows the average (mean) distribution of hurricanes (blue) and storms(orange) for each month over the time period 1980:2019. This shows that there is a high frequency of hurricanes between months 7-12. This visualization also confirms that storms do occur on "outside" months.

From this visualization, we see that hurricane frequency is highest in the months 7-10. Interestingly, we don't see very high values for month 6 or 11 which are supposed to be high according to claim A's typical hurricane season assessment. It is interesting to note month 1 has a higher frequency average than month 11 which is unexpected since we are considering month 1 an outside month.

This graph gives us a good general overview of the difference in distribution for storms and hurricanes. This graph also provides us with questions to look farther into for instance does it appear to be a high number of occurances for storms and hurricanes in month 1 because it is skewed?

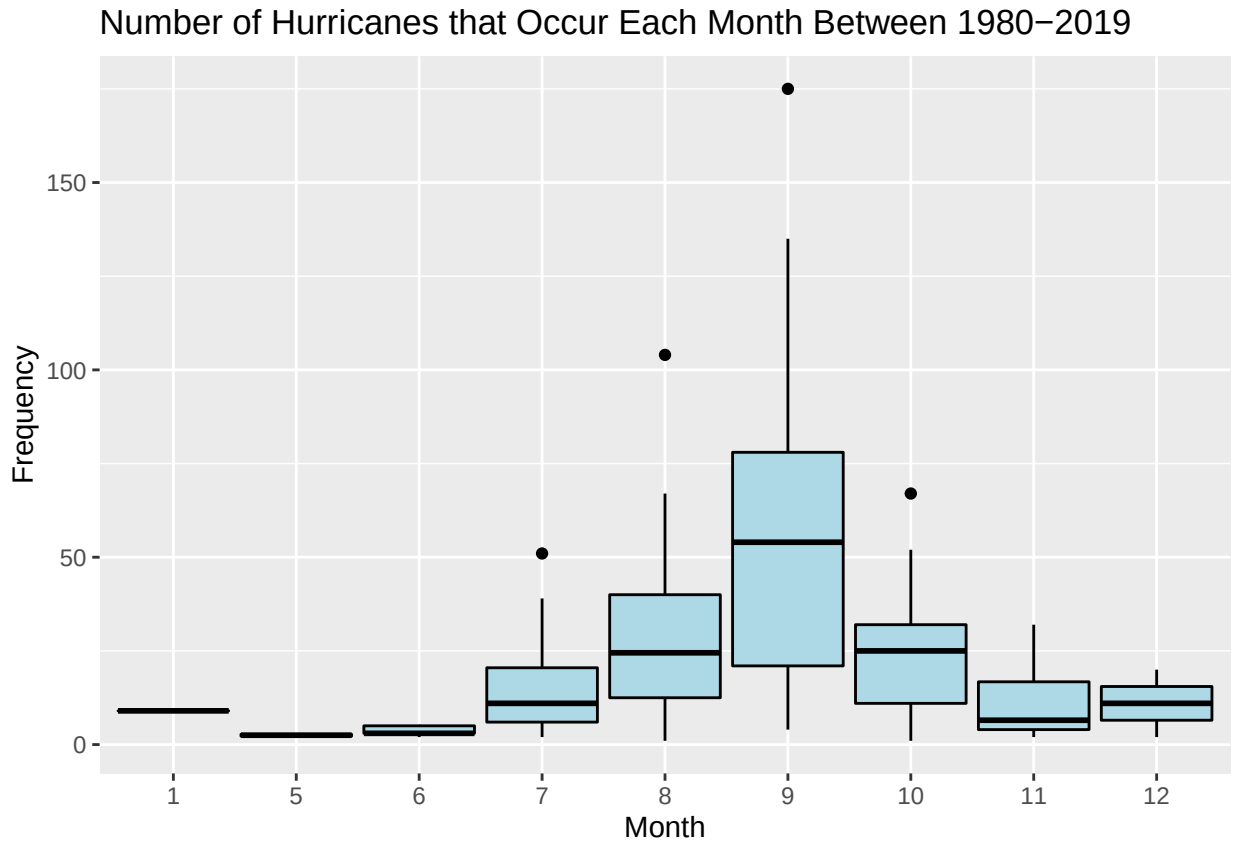
To look a little more into this, I'm going to plot a boxplot of the hurricanes and storms so we can see their outliers.

Hurricane Box Plot

```
hurricanedat <- filter(dat3, HURRICANE == "TRUE")

hurricanes_per_month <- summarise(
  group_by(hurricanedat, SEASON, MONTH),
  number = n()
)
```

```
## `summarise()` regrouping output by 'SEASON' (override with `.groups` argument)
```

```
ggplot(hurricanes_per_month, aes(x= factor(MONTH), y=number)) +
  geom_boxplot(color= "black", fill= "lightblue", width= 0.9) +
  labs(x= "Month", y="Frequency", title= "Number of Hurricanes that Occur Each Month Between 1980-2019")
```



Analysis

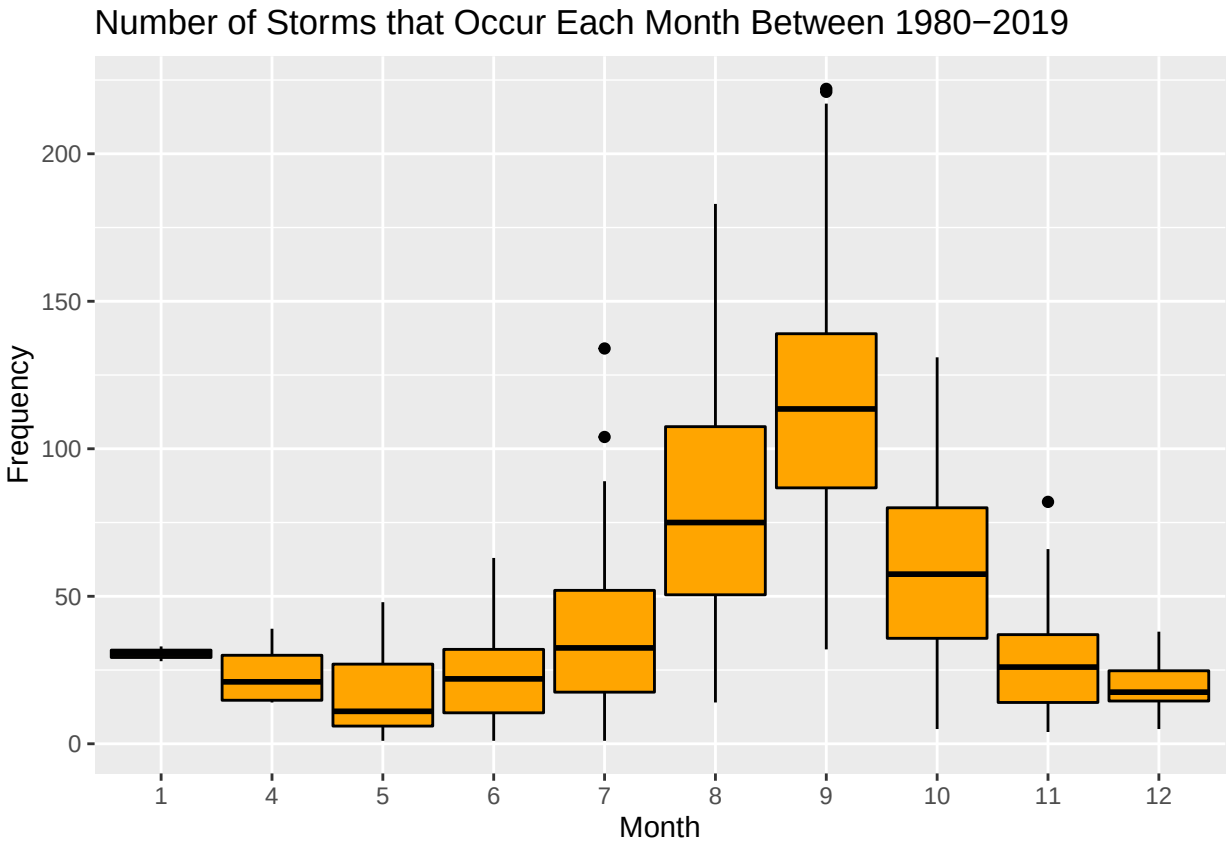
This boxplot shows that month 7, 8, 9, and 10 have outliers that make them right-skewed resulting in a higher mean. Despite this, months 7-12 do have higher mean occurances than months 1-6. This would support a "typical hurricane season," to be months 7-12 and "outside months," to be 1-6 in contrast to claim A's specified months 6-11 and 12-5 (respectively).

Storm Box Plot

```
stormdat <- filter(dat3, HURRICANE == "FALSE")

storms_per_month <- summarise(
  group_by(stormdat, SEASON, MONTH),
  number = n()
)
```

```
## `summarise()` regrouping output by 'SEASON' (override with `.groups` argument)
```

```
ggplot(storms_per_month, aes(x= factor(MONTH), y=number)) +
  geom_boxplot(color= "black", fill= "orange", width= 0.9) +
  labs(x= "Month", y="Frequency", title= "Number of Storms that Occur Each Month Between 1980-2019")
```

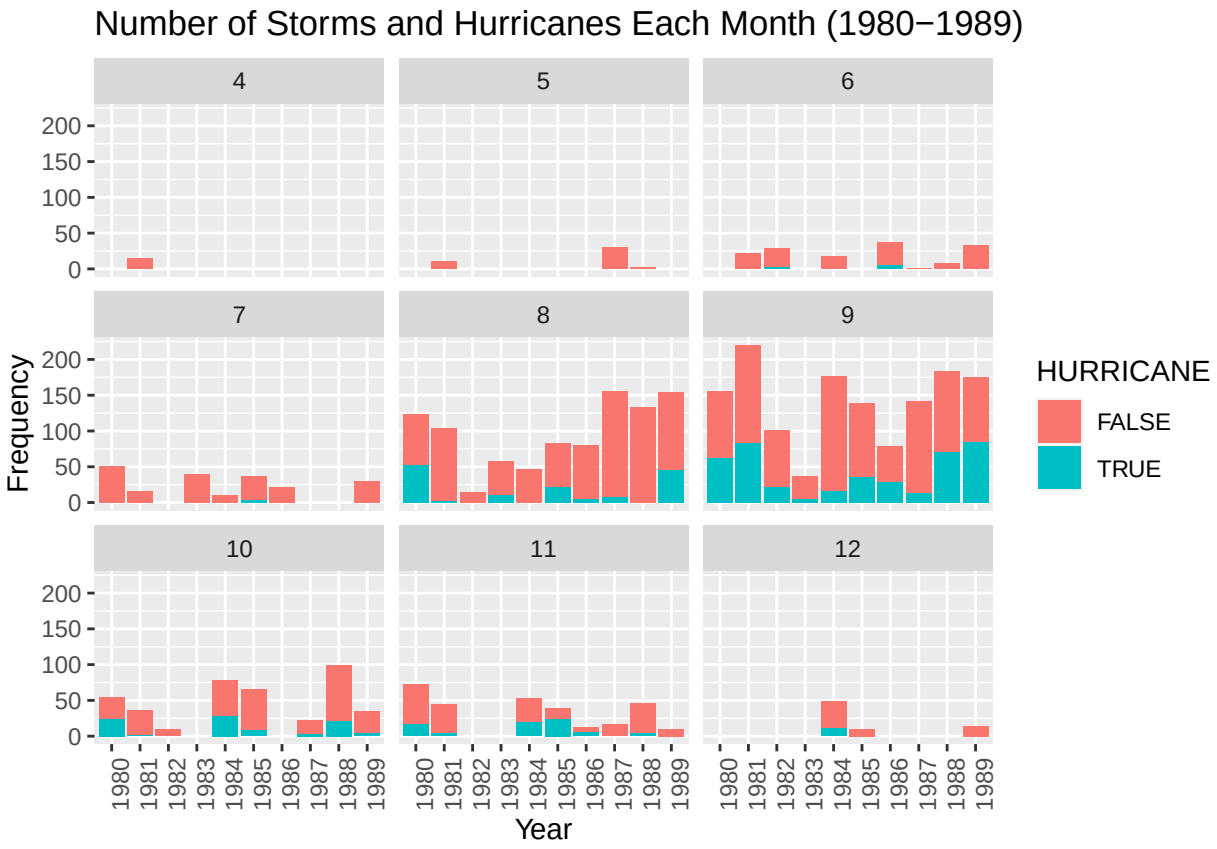


Analysis

Contrary to the idea I previously mentioned, month 1 for storms does not appear to have an outlier. On the other hand, month 11 does have an outier which makes it right-skewed, having a higher mean than is actually accurate of the average. Looking at the boxplot, one can see that month 11's mean without outliers is below that of month 1.

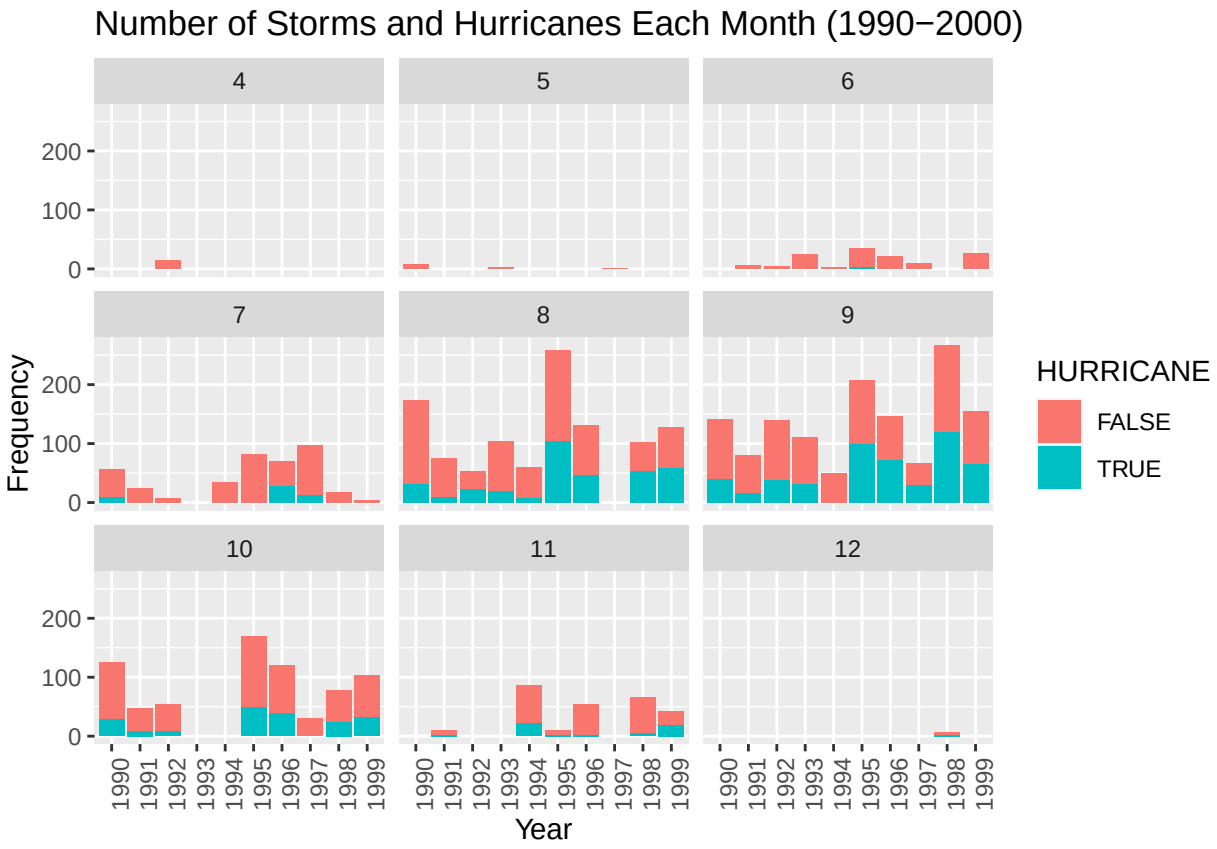
This discrepency leads me to think that "occasionally storms form in outside these months," could be wrong because the mean of month 1 is higher than month 6 and 11. However, I want to look deeper into this before I jump to conclusions, so I will visualize the distribution of months for each year.

```
sh_decade1 <- filter(storms_hurricanes, SEASON %in% 1980:1989)

ggplot(sh_decade1, aes(x= factor(SEASON), y=number, fill=HURRICANE)) +
  geom_bar(stat= "identity") +
  labs(x= "Year", y="Frequency", title= "Number of Storms and Hurricanes Each Month (1980-1989)")+
  facet_wrap(~MONTH)+
  theme(axis.text.x = element_text(angle = 90))
```



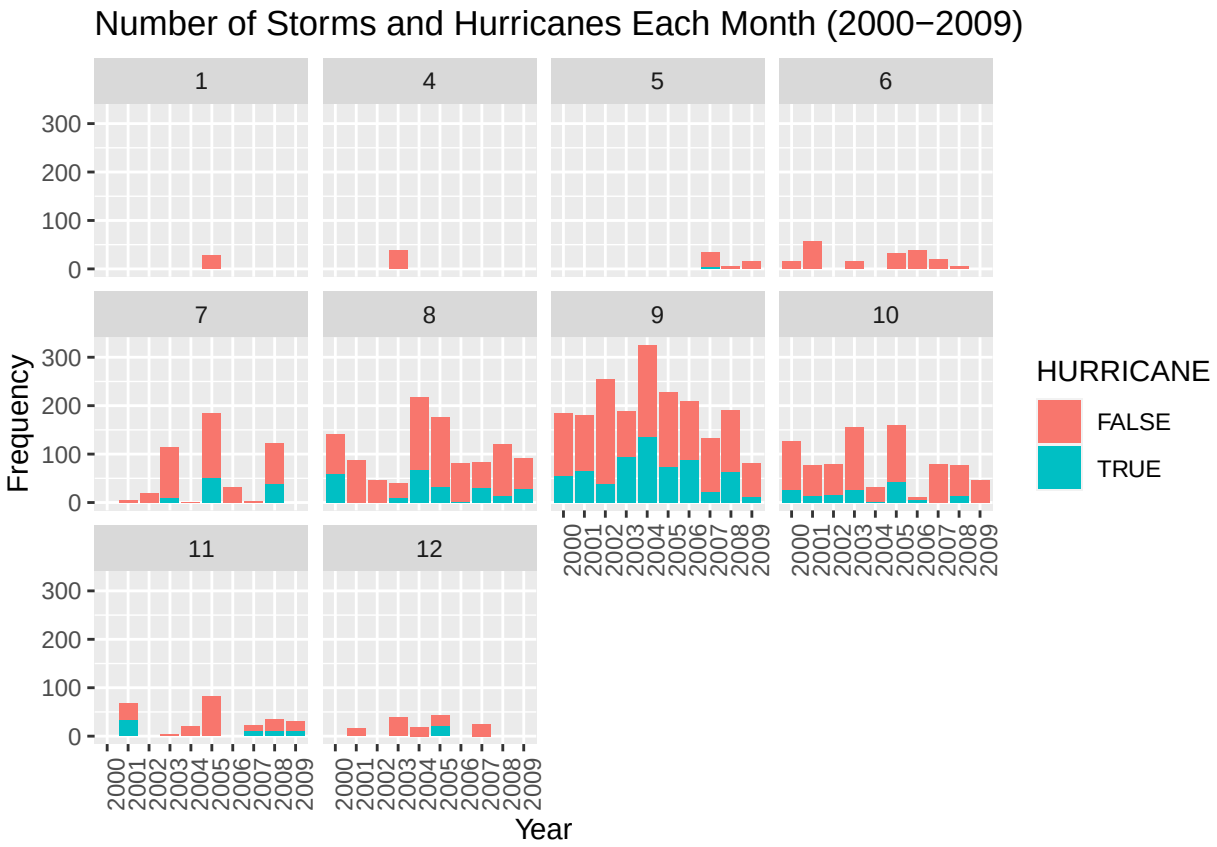
Summary: Month 1 does not appear. Month 4, 5, and 12 have low frequency. Month 7 seems slightly ahead of month 6. Month 11 seems slightly ahead of month 7. Month 8,9,10 have the highest occurances.

```
sh_decade2 <- filter(storms_hurricanes, SEASON %in% 1990:1999)

ggplot(sh_decade2, aes(x= factor(SEASON), y=number, fill=HURRICANE)) +
  geom_bar(stat= "identity") +
  labs(x= "Year", y="Frequency", title= "Number of Storms and Hurricanes Each Month (1990-2000)")+
  facet_wrap(~MONTH)+
  theme(axis.text.x = element_text(angle = 90))
```



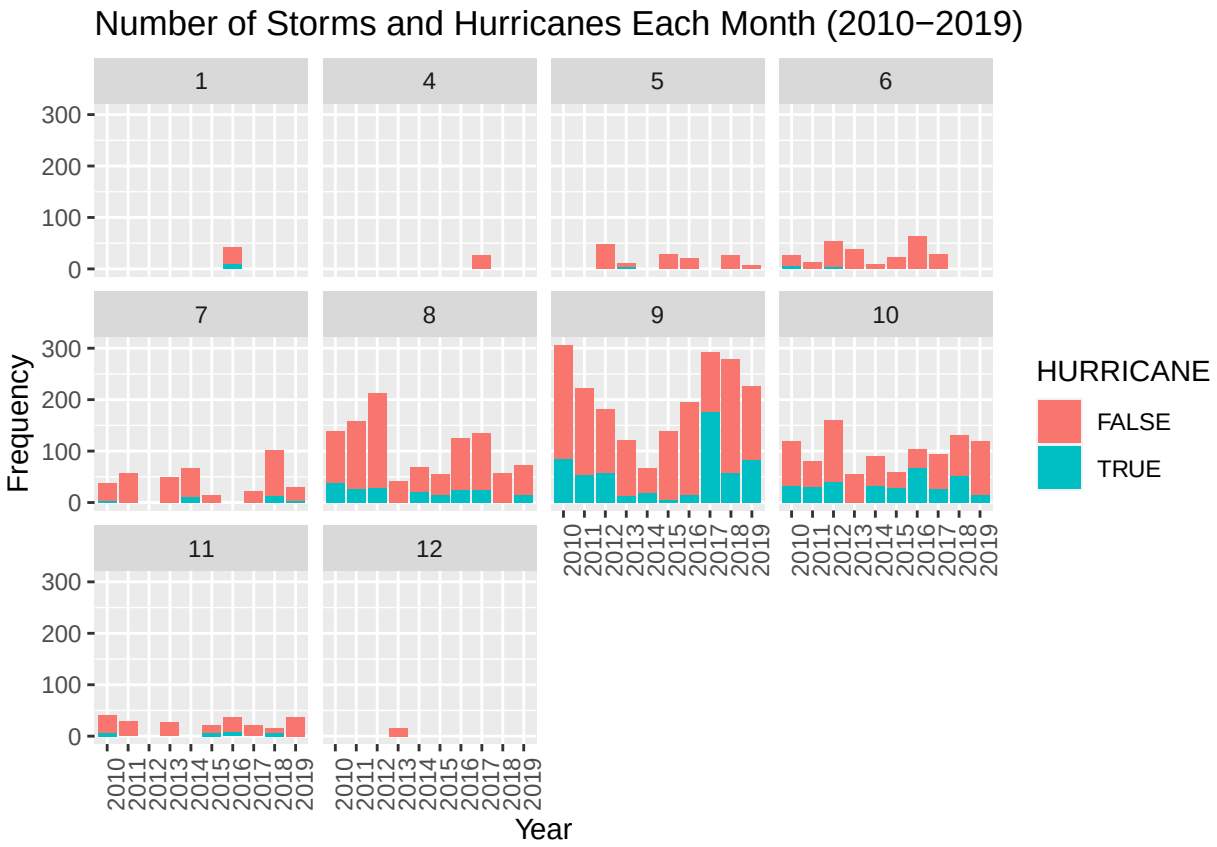
Summary: Month 1 does not appear. Month 4, 5, and 12 have low frequency. Month 7 has higher frequency than month 6. Month 11 seems slightly less than month 7. Month 8,9,10 have the highest occurances.

```
sh_decade3 <- filter(storms_hurricanes, SEASON %in% 2000:2009)

ggplot(sh_decade3, aes(x= factor(SEASON), y=number, fill=HURRICANE)) +
  geom_bar(stat= "identity") +
  labs(x= "Year", y="Frequency", title= "Number of Storms and Hurricanes Each Month (2000-2009)")+
  facet_wrap(~MONTH)+
  theme(axis.text.x = element_text(angle = 90))
```



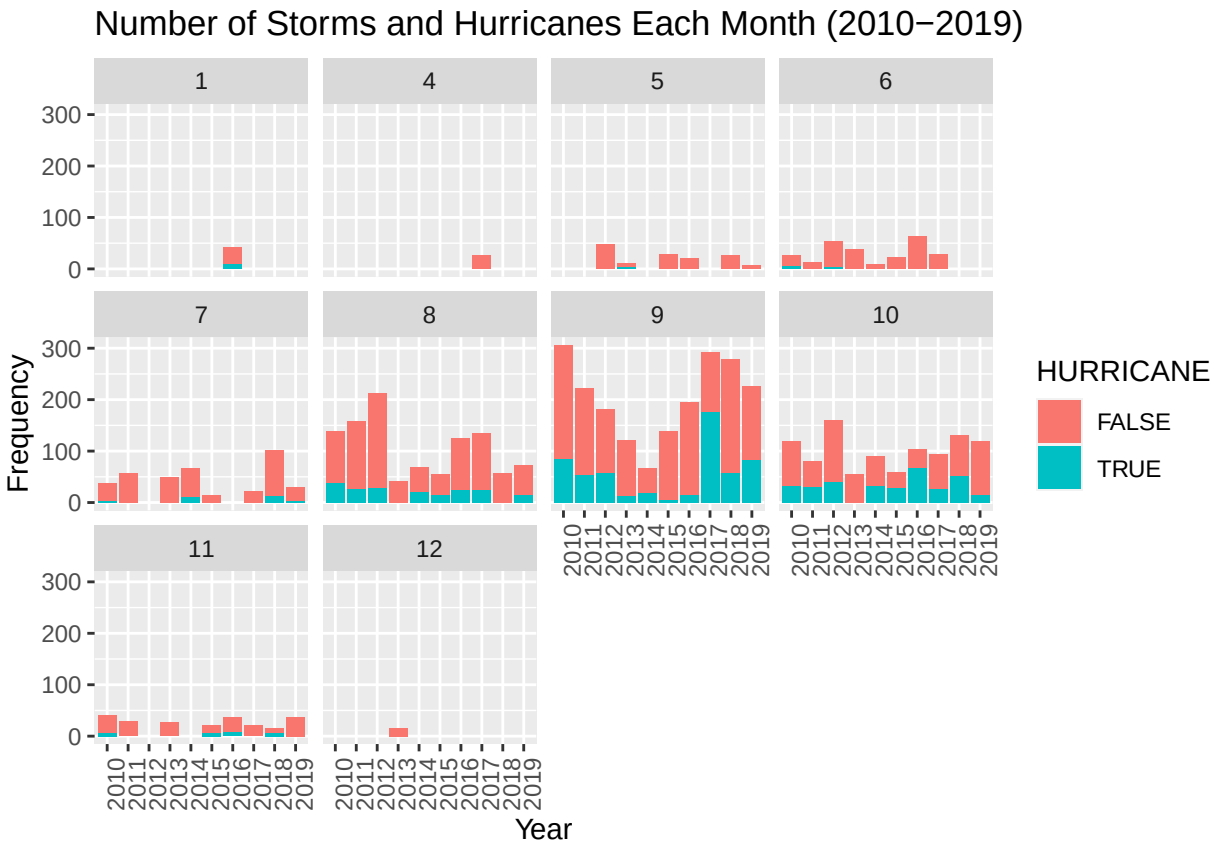
Summary: Month 1 appears. Month 4, 5, and 12 have low frequency. Month 7 has higher frequency than month 6. Month 11 appears the same amount of times as month 7 but has less frequency. Month 8,9,10 have the highest occurances.

```
sh_decade4 <- filter(storms_hurricanes, SEASON %in% 2010:2019)

ggplot(sh_decade4, aes(x= factor(SEASON), y=number, fill=HURRICANE)) +
  geom_bar(stat= "identity") +
  labs(x= "Year", y="Frequency", title= "Number of Storms and Hurricanes Each Month (2010-2019)")+
  facet_wrap(~MONTH)+
  theme(axis.text.x = element_text(angle = 90))
```



Summary: Month 1 appears. Month 4, 5, and 12 have low frequency. Month 7 has higher frequency than month 6. Month 11 appears about the same as month 6 but has less frequency. Month 8,9,10 have the highest occurances.

```
sh_decade4 <- filter(storms_hurricanes, SEASON %in% 2010:2019)

ggplot(sh_decade4, aes(x= factor(SEASON), y=number, fill=HURRICANE)) +
  geom_bar(stat= "identity") +
  labs(x= "Year", y="Frequency", title= "Number of Storms and Hurricanes Each Month (2010-2019)")+
  facet_wrap(~MONTH)+
  theme(axis.text.x = element_text(angle = 90))
```



Summary: Month 1 appears. Month 4, 5, and 12 have low frequency. Month 7 has higher frequency than month 6. Month 11 appears about the same as month 6 but has less frequency. Month 8,9,10 have the highest occurances.

Final Conclusion on Claim A:

I believe that the highest occurance of hurricanes occurs in the months 7-11. I think that hurricanes show up in month 6 at about the same frequency as outside months. I would deem months 12-5 outside months.

Storms do occur outside of the typical hurricane season. After looking at the decade distributions, I would describe them as "ocasionally," occuring.

Claim B: Partial

B) A typical year has 12 named storms, including six hurricanes of which three become major hurricanes (category 3, 4, and 5).

First I will isolate all named storms from dat3 and exclude any that are listed as "Not_Named."

```
named_storms <- filter(dat3, NAME != "NOT_NAMED")
head(named_storms,5)
```

```
##             SID SEASON NUMBER BASIN SUBBASIN  NAME            ISO_TIME NATURE
## 1 1980214N11330   1980     57    NA       NA ALLEN 1980-07-31 12:00:00     NR
## 2 1980214N11330   1980     57    NA       NA ALLEN 1980-07-31 15:00:00     NR
## 3 1980214N11330   1980     57    NA       NA ALLEN 1980-07-31 18:00:00     NR
## 4 1980214N11330   1980     57    NA       NA ALLEN 1980-07-31 21:00:00     NR
## 5 1980214N11330   1980     57    NA       NA ALLEN 1980-08-01 00:00:00     TS
##       LAT      LON WMO_WIND WMO_PRES WMO_AGENCY TRACK_TYPE DIST2LAND LANDFALL
## 1 11.0000 -30.0000       25          hurdat_atl       main      1417     1417
## 2 10.9509 -31.1101       NA                           main      1531     1531
## 3 10.9000 -32.2000       25          hurdat_atl       main      1650     1650
## 4 10.8496 -33.2574       NA                           main      1695     1655
## 5 10.8000 -34.3000       30     1010 hurdat_atl       main      1651     1603
##   MONTH HURRICANE CATEGORY
## 1     7     FALSE        0
## 2     7        NA     <NA>
## 3     7     FALSE        0
## 4     7        NA     <NA>
## 5     8     FALSE        0
```

I will filter out hurricanes by saying hurricane = false in order to make sure I am only working with storms.

```
onlystorms <- filter(named_storms, HURRICANE == "FALSE")
distinct_storms <- distinct(select(onlystorms, SEASON, NAME))
head(distinct_storms, 5)
```

```
##   SEASON    NAME
## 1   1980   ALLEN
## 2   1980  BONNIE
## 3   1980 CHARLEY
## 4   1980 GEORGES
## 5   1980    EARL
```

Using the distinct named storm data (distinct_storms), I will try to verify the first part of the claim "a typical year has 12 named storms." I will create a table that counts the number of named storms that occur for each season (year).

Based off of the table storms_count, which provides the number of storms that occured in each year, I will calculate the average using mean.

```
storm_count <- count(distinct_storms, SEASON)
storm_count %>% summarise(mean_storms_season = mean(n, na.rm = TRUE))
```

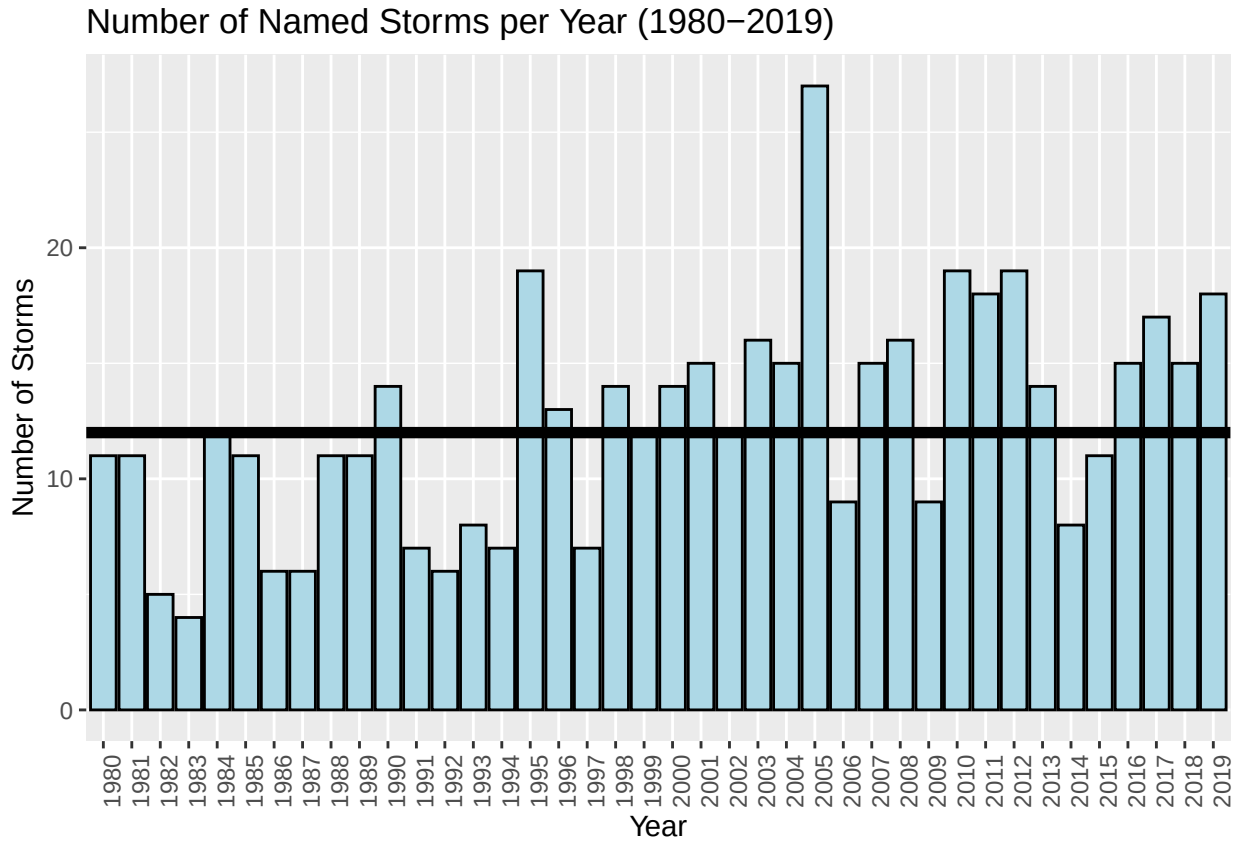
```
##   mean_storms_season
## 1             12.425
```

This mean provides us an average of the amount of storms that occured for each year. This average verifies that about 12 named storms did occur each season. I will also visualize the data to see the trend.

```
storms_per_season <- summarise(
  group_by(distinct_storms, SEASON),
  NAMED_STORMS = n()
)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
ggplot(data=storms_per_season, aes(x=factor(SEASON), y=NAMED_STORMS)) +
geom_bar(stat="identity", color= "black", fill= "lightblue") +
  scale_x_discrete("Year", labels = 1980:2019, limits= factor(1980:2019)) +
  labs(x= "Year", y= "Number of Storms", title= "Number of Named Storms per Year (1980-2019)")+
  geom_abline(slope=0, intercept=12,  col = "black", lwd=2) + theme(axis.text.x = element_text(angle = 9
```



Next I will move to answering the next part of the claim: a typical year includes 6 hurricanes. For this, I will filter hurricane = true.

```
named_hurricanes <- filter(named_storms, HURRICANE == "TRUE")
distinct_hurricanes <- distinct(select(named_hurricanes, SEASON, NAME))
head(distinct_hurricanes,5)
```

```
##   SEASON    NAME
## 1   1980   ALLEN
## 2   1980  BONNIE
## 3   1980 CHARLEY
## 4   1980 GEORGES
## 5   1980    EARL
```

```
hurricanes_per_season <- summarise(
  group_by(distinct_hurricanes, SEASON),
  NAMED_HURRICANES = n()
)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
head(hurricanes_per_season, 5)
```

```
## # A tibble: 5 x 2
##   SEASON NAMED_HURRICANES
##    <int>            <int>
## 1   1980                9
## 2   1981                7
## 3   1982                2
## 4   1983                3
## 5   1984                5
```

```
hurricanes_per_season <- count(distinct_hurricanes, SEASON)
hurricanes_per_season %>% summarise(mean_hurricane_season= mean(n, na.rm=TRUE))
```

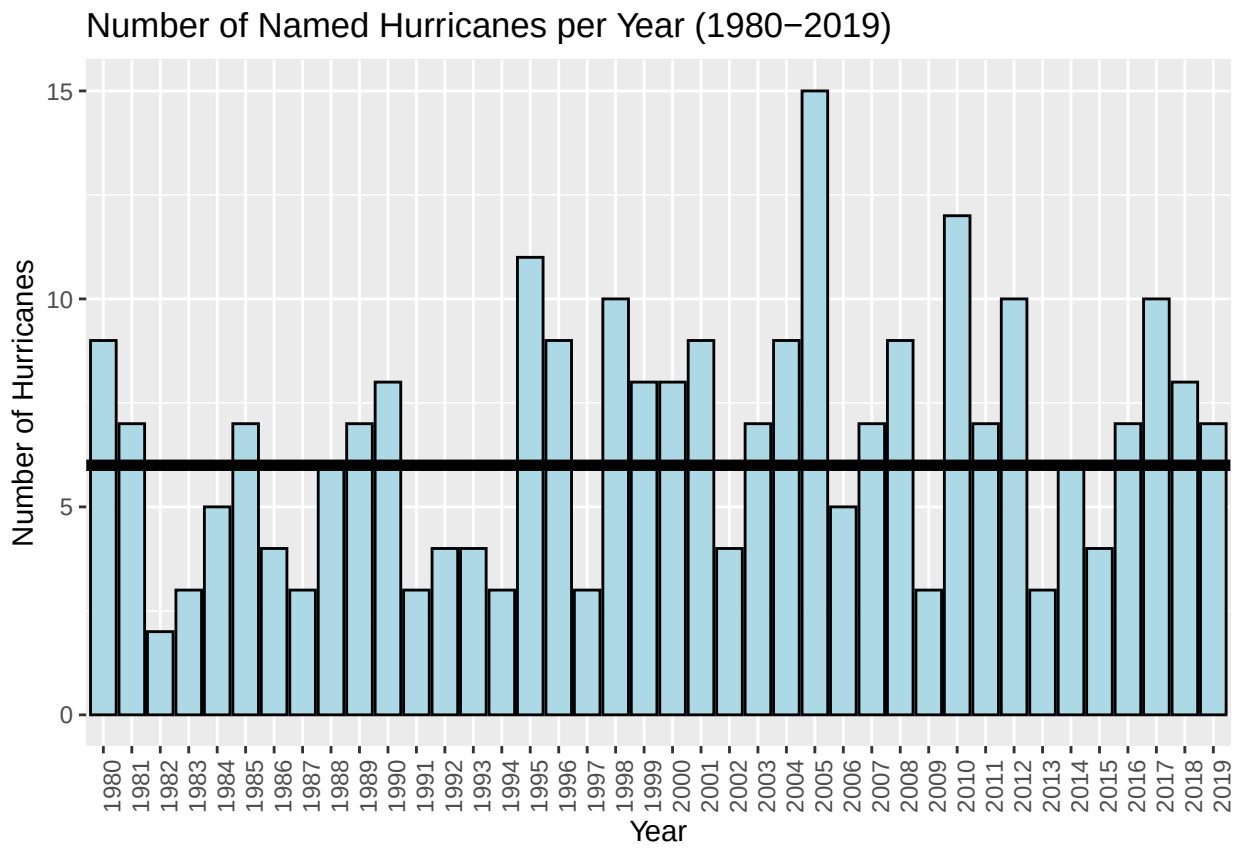
```
##   mean_hurricane_season
## 1                  6.65
```

This value is slightly above what I would consider to still be about 6. I am going to round up on this value and say it is closer to 7. Therefore, I will have to state that I believe the second part of the claim is false.

```
hurricanes_per_season <- summarise(
  group_by(distinct_hurricanes, SEASON),
  NAMED_HURRICANES = n()
)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
ggplot(data=hurricanes_per_season, aes(x=factor(SEASON), y=NAMED_HURRICANES)) +
geom_bar(stat="identity", color= "black", fill= "lightblue") +
  scale_x_discrete("Year", labels = 1980:2019, limits= factor(1980:2019)) +
  labs(x= "Year", y= "Number of Hurricanes", title= "Number of Named Hurricanes per Year (1980-2019)")+
theme(axis.text.x = element_text(angle = 90))
```



Finally, we will answer the third part of the claim which is three of the hurricanes become major hurricanes per year. Major hurricanes are classified by category 3, 4, and 5 which I have filtered for.

```
named_major <- filter(named_storms, CATEGORY %in% 3:5)
distinct_major <- distinct(select(named_major, SEASON, NAME))
head(distinct_major,5)
```

```
##   SEASON    NAME
## 1   1980   ALLEN
## 2   1980 FRANCES
## 3   1981   FLOYD
## 4   1981  HARVEY
## 5   1981   IRENE
```

```
major_per_season <- summarise(
  group_by(distinct_major, SEASON),
  NAMED_MAJOR_HURRICANES = n()
)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
head(major_per_season, 5)
```

```
## # A tibble: 5 x 2
```

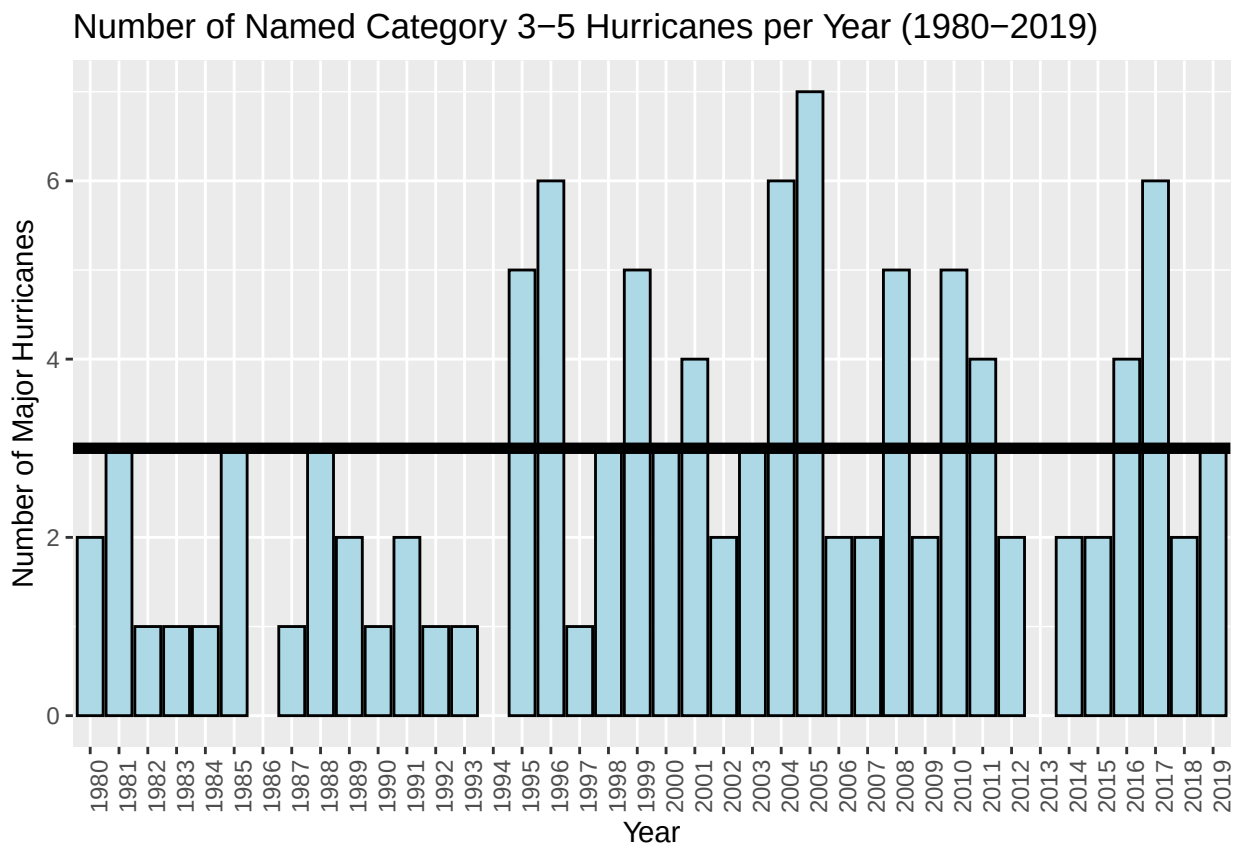
```
##   SEASON NAMED_MAJOR_HURRICANES
##    <int>                  <int>
## 1   1980                      2
## 2   1981                      3
## 3   1982                      1
## 4   1983                      1
## 5   1984                      1
```

```
major_per_season %>% summarise(mean_major_hurricane= mean(NAMED_MAJOR_HURRICANES, na.rm=TRUE))
```

```
## # A tibble: 1 x 1
##   mean_major_hurricane
##                  <dbl>
## 1                 2.92
```

This value is very close to 3 and I would consider the claim of 3 major hurricanes to be true.

```
ggplot(data=major_per_season, aes(x=factor(SEASON), y=NAMED_MAJOR_HURRICANES)) +
geom_bar(stat="identity", color= "black", fill= "lightblue") +
  scale_x_discrete("Year", labels = 1980:2019, limits= factor(1980:2019)) +
  labs(x= "Year", y= "Number of Major Hurricanes", title= "Number of Named Category 3-5 Hurricanes per
  geom_abline(slope=0, intercept=3,  col = "black", lwd=2) +
  theme(axis.text.x = element_text(angle = 90))
```



Final Conclusion on Claim B

I believe there are about 12 storms per year. I do not believe there are 6 hurrianes, I believe there are closer to 7 hurricanes per year. I do believe that there are about 3 major hurricanes per year.

Claim C: True

C) September is the most active month (where most of the hurricanes occur), followed by August, and October.

This question is similar to how I solved A; however, it specifically asks for hurricanes so I will filter for only hurricane = true.

```
hurricanedat <- filter(dat3, HURRICANE == "TRUE")
head(hurricanedat,5)
```

```
##             SID SEASON NUMBER BASIN SUBBASIN  NAME            ISO_TIME NATURE
## 1 1980214N11330   1980     57    NA       NA ALLEN 1980-08-03 00:00:00     TS
## 2 1980214N11330   1980     57    NA       NA ALLEN 1980-08-03 06:00:00     TS
## 3 1980214N11330   1980     57    NA       NA ALLEN 1980-08-03 12:00:00     TS
## 4 1980214N11330   1980     57    NA       NA ALLEN 1980-08-03 18:00:00     TS
## 5 1980214N11330   1980     57    NA       NA ALLEN 1980-08-04 00:00:00     TS
##    LAT   LON WMO_WIND WMO_PRES WMO_AGENCY TRACK_TYPE DIST2LAND LANDFALL MONTH
## 1 12.4 -51.4       65      985 hurdat_atl       main       790      762     8
## 2 12.6 -53.6       70      980 hurdat_atl       main       753      713     8
## 3 12.8 -55.6       80      975 hurdat_atl       main       633      543     8
## 4 12.9 -57.5       95      965 hurdat_atl       main       453      398     8
## 5 13.3 -59.1      110      950 hurdat_atl       main       357      321     8
##   HURRICANE CATEGORY
## 1      TRUE        1
## 2      TRUE        1
## 3      TRUE        1
## 4      TRUE        2
## 5      TRUE        3
```

```
hurricane_month <- distinct(select(hurricanedat, SID, SEASON, MONTH))
head(hurricane_month,5)
```

```
##             SID SEASON MONTH
## 1 1980214N11330   1980     8
## 2 1980227N13325   1980     8
## 3 1980234N36287   1980     8
## 4 1980245N16322   1980     9
## 5 1980249N18336   1980     9
```

Next I will filter for just the months in question: August, September, and October

```
fall_months <- filter(hurricane_month, MONTH %in% 8:10)
head(fall_months,5)
```

```
##             SID SEASON MONTH
## 1 1980214N11330   1980     8
## 2 1980227N13325   1980     8
## 3 1980234N36287   1980     8
## 4 1980245N16322   1980     9
## 5 1980249N18336   1980     9
```

Note: some of the same SID appear more than once because some of the same storms occur in more than one month. For instance, SID: 1981265N14328 occurs in month 9 and 10 of 1981. I am going to count this as a hurricane occuring in the month of 9 and a hurricane occuring in the month of 10.

```
fall_hurricanes <- summarise(
  group_by(fall_months, SEASON, MONTH),
  number = n()
)
```

```
## `summarise()` regrouping output by 'SEASON' (override with `.groups` argument)
```

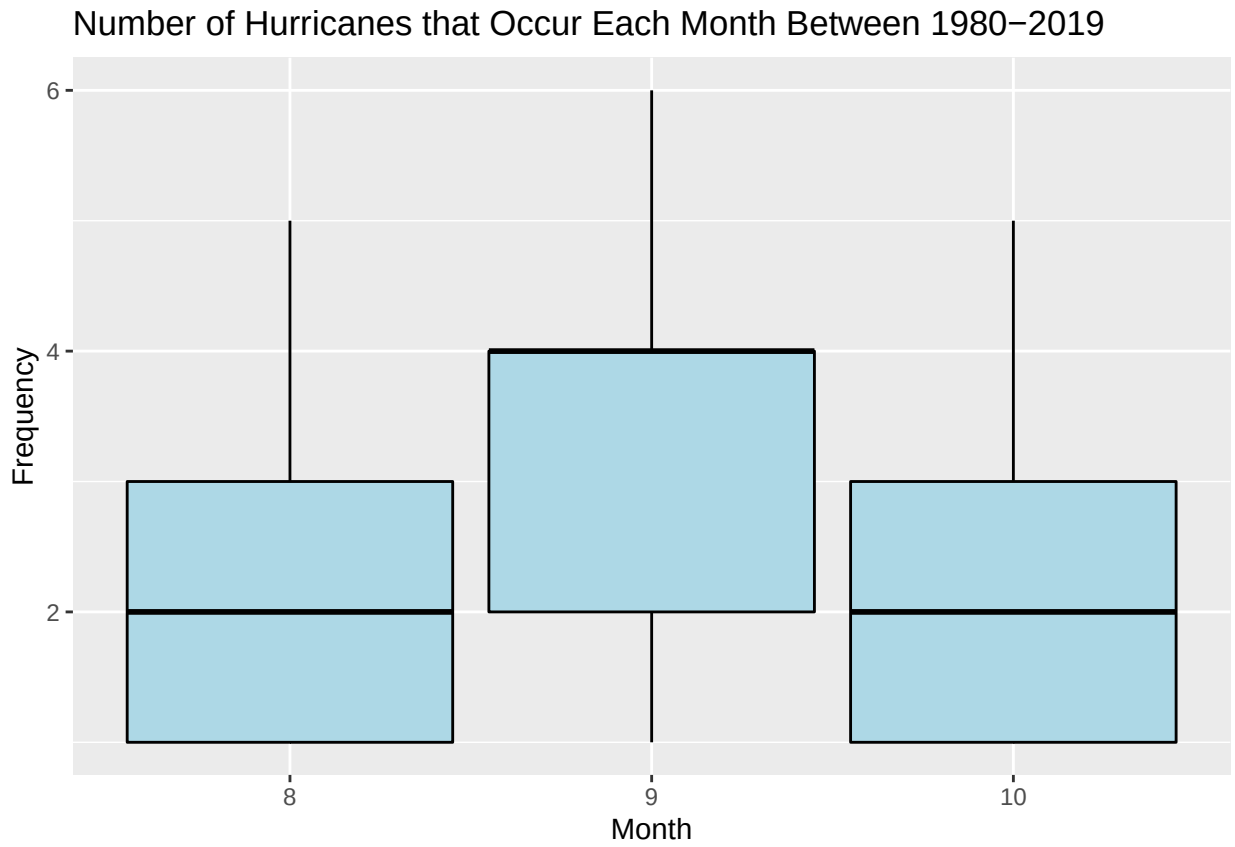
```
head(fall_hurricanes,5)
```

```
## # A tibble: 5 x 3
## # Groups:   SEASON [2]
##   SEASON MONTH number
##    <int> <dbl>  <int>
## 1   1980     8      3
## 2   1980     9      3
## 3   1980    10      1
## 4   1981     8      1
## 5   1981     9      5
```

Boxplot of Fall Months

I used a box plot to see the fall months' means and if there are any outliers skewing the means.

```
ggplot(fall_hurricanes, aes(x= factor(MONTH), y=number)) +
  geom_boxplot(color= "black", fill= "lightblue", width= 0.9) + labs(x= "Month", y="Frequency", title=
```



This boxplot shows that September definitely has the highest mean and therefore the most amount of hurricane occurances. September is a clear winner for most active month. From this boxplot it is difficult to tell a difference between August and October.

I calculated the means to give me a better idea since I can't differentiate the means on the boxplot.

```
fall_avg <- summarise(
  group_by(fall_hurricanes, MONTH),
  month_avg= mean(number, na.rm=TRUE))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
fall_avg
```

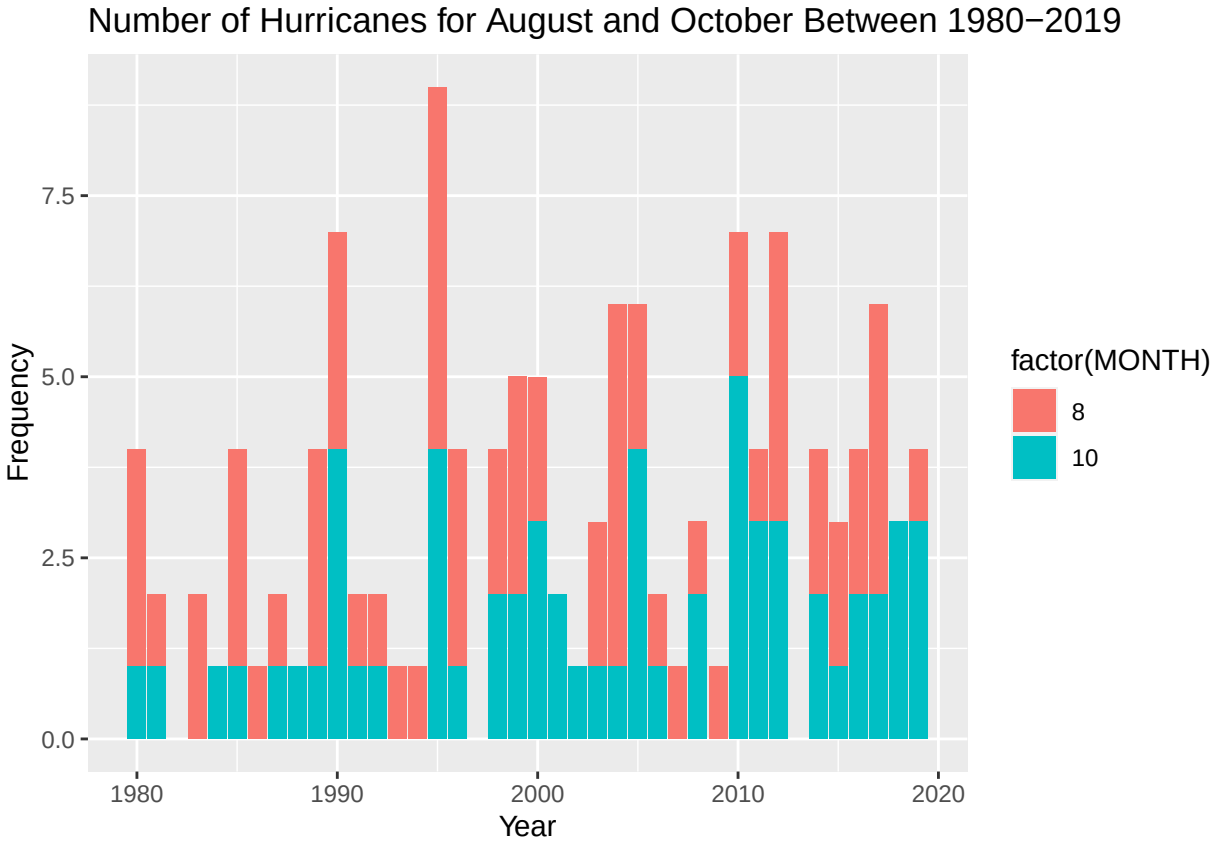
```
## # A tibble: 3 x 2
##   MONTH month_avg
##   <dbl>     <dbl>
## 1     8      2.09
## 2     9      3.15
## 3    10      1.97
```

The means confirm that September is the most active month, followed by August, then October as the claim states. I also want to visualize the differences between August and October to get a closer look.

```r
AugOct <- filter(fall_hurricanes, MONTH %in% 8 | MONTH %in% 10)
head(AugOct, 5)
```

```
## # A tibble: 5 x 3
## # Groups:   SEASON [3]
##   SEASON MONTH number
##    <int> <dbl>  <int>
## 1   1980     8      3
## 2   1980    10      1
## 3   1981     8      1
## 4   1981    10      1
## 5   1983     8      2
```

```r
ggplot(AugOct, aes(x= SEASON, y=number, fill= factor(MONTH))) +
  geom_bar(stat = "identity")+
  labs(x= "Year", y="Frequency", title= "Number of Hurricanes for August and October Between 1980-2019")
```



From looking at this graph, it is clear to see that August surpasses October every year.

Final Conclusion on Claim C

Given all the evidence, I definitely agree Claim C is true.

Claim D: True

D) During the analyzed period (1980-2019), no hurricanes made U.S. landfall before June and after November.

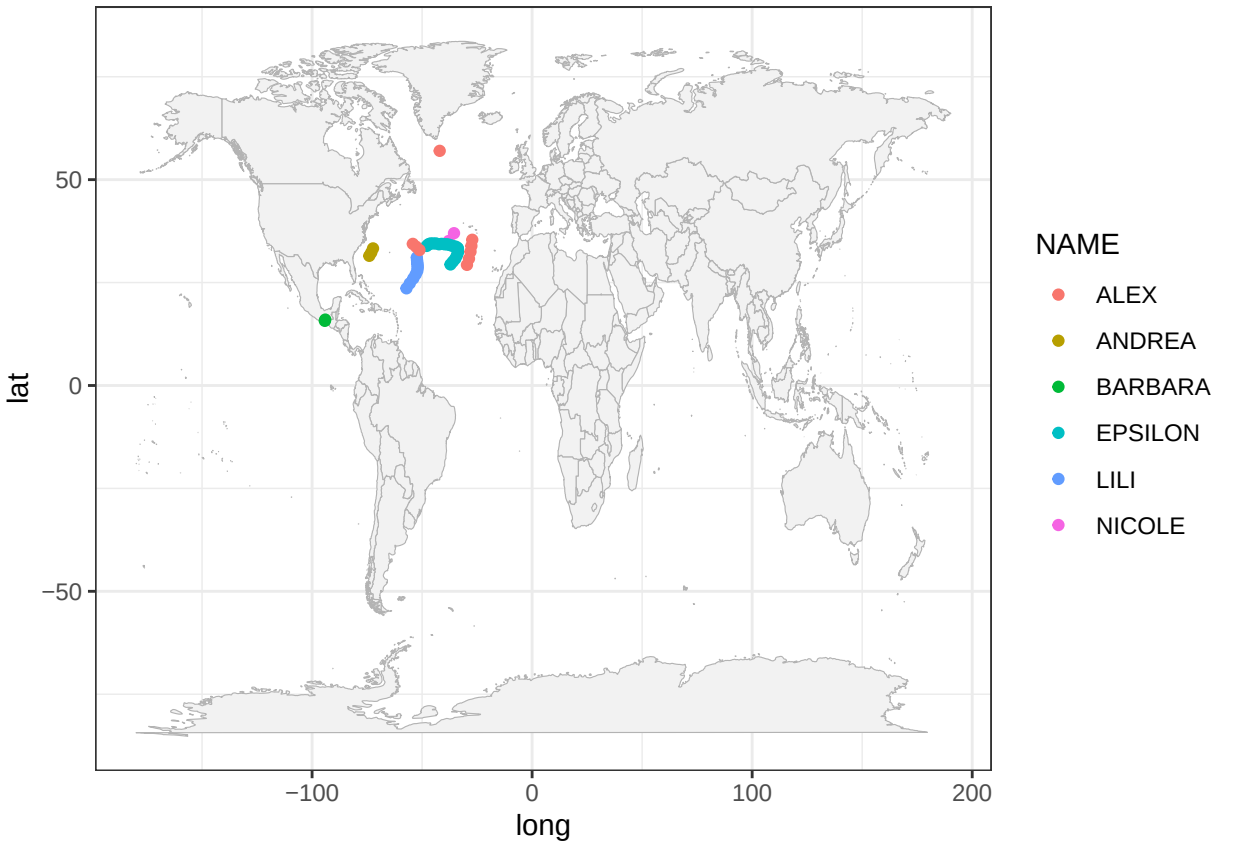
To assess this claim, I first filtered the data frame to only include hurricanes and only the time frame specified.

```
hurricanetime <- filter(dat3, HURRICANE == "TRUE", MONTH <6 | MONTH >11)
head(hurricanetime,5)
```

```
##             SID SEASON NUMBER BASIN SUBBASIN NAME            ISO_TIME NATURE
## 1 1984348N35300   1984    121    NA       NA LILI 1984-12-20 12:00:00     TS
## 2 1984348N35300   1984    121    NA       NA LILI 1984-12-20 18:00:00     TS
## 3 1984348N35300   1984    121    NA       NA LILI 1984-12-21 00:00:00     TS
## 4 1984348N35300   1984    121    NA       NA LILI 1984-12-21 06:00:00     TS
## 5 1984348N35300   1984    121    NA       NA LILI 1984-12-21 12:00:00     TS
##    LAT   LON WMO_WIND WMO_PRES WMO_AGENCY TRACK_TYPE DIST2LAND LANDFALL MONTH
## 1 31.1 -52.4       70      980 hurdat_atl       main      1726     1726    12
## 2 30.5 -52.3       70      980 hurdat_atl       main      1794     1794    12
## 3 30.0 -52.2       70      980 hurdat_atl       main      1850     1850    12
## 4 29.5 -52.1       70      980 hurdat_atl       main      1846     1831    12
## 5 29.0 -52.0       70      980 hurdat_atl       main      1818     1805    12
##   HURRICANE CATEGORY
## 1      TRUE        1
## 2      TRUE        1
## 3      TRUE        1
## 4      TRUE        1
## 5      TRUE        1
```

I then was curious just to see what these hurricanes looked like on a map. (This doesn't answer the question, just provides a good overall look)

```
gg_world +
  geom_point(data = hurricanetime, aes(x = LON, y = LAT, color = NAME))
```



This map shows me that only one hurricane, Andrea, seems relatively close to the U.S. This map also shows me that one hurricane, Barbra, appears to have made landfall.

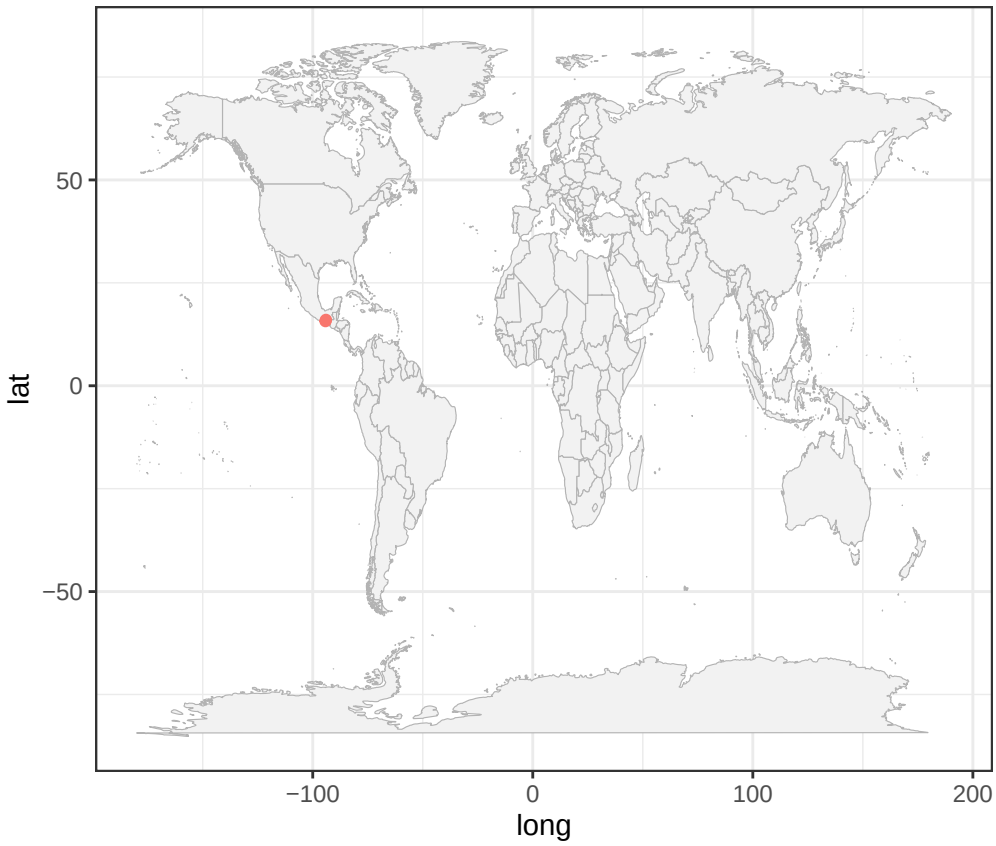
Filtering landfall The data dictionary specifies a hurricane to make landfall when it is =0. The data dictionary specifies that values less than "60 nmile" are likely to impact the land. I am going to assume that they meant to say 60 miles. The landfall category is recorded in km so I will convert 60 miles to km which is 96.5. I am going to filter for hurricanes that have a landfall below 97 because I want to see which hurricanes have an impact on the land.

```
Barb <- filter(hurricanetime, LANDFALL < 97)
Barb
```

```
##             SID SEASON NUMBER BASIN SUBBASIN    NAME            ISO_TIME NATURE
## 1 2013149N14264   2013     22    EP     <NA> BARBARA 2013-05-29 18:00:00     TS
## 2 2013149N14264   2013     22    EP     <NA> BARBARA 2013-05-29 19:50:00     TS
##    LAT   LON WMO_WIND WMO_PRES WMO_AGENCY TRACK_TYPE DIST2LAND LANDFALL MONTH
## 1 15.7 -94.2       65      986 hurdat_epa       main        39        0     5
## 2 16.0 -94.0       70      983 hurdat_epa       main         0        0     5
##   HURRICANE CATEGORY
## 1      TRUE        1
## 2      TRUE        1
```

Only one hurricane resulted in this filter and it was Barbara who's landfall is = 0.

```
gg_world +
  geom_point(data = Barb, aes(x = LON, y = LAT, color = NAME))
```



Visualization of Barbara

Final Conclusion on Claim D

Barbara is not located in the U.S. so we can safely say that no hurricanes between the time period 1980-2019 before the months June and after the months November made landfall.

Thank for reading my report !