DSC520 Final Project

Carla Harrell

11/13/2019

## R Markdown

#Introduction

#Tornadoes are formed when hot air and cold air meet in a powerful storm, and the warm air begins spiraling upward, creating a funnel cloud. This funnel cloud forces objects on the ground and in the air around and upward, and can be strong enough to uproot trees, move cars and tear apart buildings. This extremely energetic air current is part of a special, very tall storm called a supercell.

#The supercells that make tornadoes form when a large mass of cold air moves over a large mass of warm air. This situation is highly unstable, because the warm air is lighter than the cold air. The greater the temperature difference, the more unstable the situation. Winds are generated as the warm air moves up and through the cold air and the cold air moves down and around the warm air. If the warm air contains a large amount of water vapor, it tends to condense into clouds as it moves up and cools.

#When this movement of warm air upward is energetic enough, the moving air starts to twist on its way up, creating the potential for a funnel cloud. These funnel clouds are the effect of rapidly moving and spinning warm air and condensing water vapor going upward and spiraling.

#Research questions #1. What is the frequency of occurrence of tornadoes in Dallas? #2. What category of Tornadoes? #3. How many occurrences per year? #4. What time of the season does the storms occur? #5. What is the average length of miles of the Tornadoes?

#Approach #1. Get the number of observations (rows) and variables, and a head of the first cases. Review the data to look for the data types, zeros, infinite numbers and missing values. Removes all variables with NA, zeros or high variables. #2. Analyze numerical and categorical at the same to see what contributes the occurrence of tornadoes.  
#3. Use head and structure to determine the type of variables and review the data to see if any data need to be removed. #4. Plot the date looking for distributions of the day. Form a histogram to find the distribution of the variables. #5. Use summary statistics to explore the distribution of the variables. #6. Find out the correlation between the variables by using correlation matrix and try to find out which variables #7. Build a regression model to predict the outcome variable.

#Because the orignal dataset was limited with Dallas, Texas, I used a dataset for all the counties in Texas and then narrowed it down to Dallas.

FILENAME <- "c:/users/carla/OneDrive/Documents/R/Assignments/tornado\_new.csv"  
  
#FILENAME <- "tornado\_new.csv"  
  
tornado\_df <- read.csv(FILENAME, stringsAsFactors = FALSE)  
  
### explore the data   
  
head(tornado\_df)

## Including Plots

library(tidyverse)

## -- Attaching packages -------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1 v purrr 0.3.2  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts ----------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

# taking a quick look  
glimpse(tornado\_df)

## Observations: 1,997  
## Variables: 11  
## $ County <chr> "Fannin", "Hunt", "Bell", "Grayson", "Hill", "Limes...  
## $ Year <int> 1880, 1883, 1883, 1885, 1885, 1885, 1886, 1888, 188...  
## $ Date <chr> "5/28/1880", "6/2/1883", "4/27/1883", "4/21/1885", ...  
## $ Time\_CST <chr> "22:15:00", "19:15:00", "15:45:00", "17:35:00", "10...  
## $ Rating <int> 4, 2, 4, 2, 2, 2, 2, 2, 1, 2, 4, 3, 3, 4, 4, 4, 4, ...  
## $ Length <dbl> -999, -999, -999, -999, -999, -999, -999, -999, -99...  
## $ Width <int> -999, -999, -999, -999, -999, -999, -999, -999, -99...  
## $ Start.Lat <dbl> -999, -999, -999, -999, -999, -999, -999, -999, -99...  
## $ Start.Long <dbl> -999, -999, -999, -999, -999, -999, -999, -999, -99...  
## $ End.Lat <dbl> -999, -999, -999, -999, -999, -999, -999, -999, -99...  
## $ End.Long <dbl> -999, -999, -999, -999, -999, -999, -999, -999, -99...

sapply(tornado\_df, function(x) sum(is.na(x))) # no missing values in the data

## County Year Date Time\_CST Rating Length   
## 0 0 0 0 0 0   
## Width Start.Lat Start.Long End.Lat End.Long   
## 0 0 0 0 0

summary(tornado\_df)

## County Year Date Time\_CST   
## Length:1997 Min. :1880 Length:1997 Length:1997   
## Class :character 1st Qu.:1970 Class :character Class :character   
## Mode :character Median :1985 Mode :character Mode :character   
## Mean :1982   
## 3rd Qu.:2001   
## Max. :2019   
## Rating Length Width Start.Lat   
## Min. :0.000 Min. :-999.00 Min. :-999.00 Min. :-999.00   
## 1st Qu.:0.000 1st Qu.: 0.10 1st Qu.: 10.00 1st Qu.: 31.80   
## Median :1.000 Median : 0.50 Median : 30.00 Median : 32.52   
## Mean :1.025 Mean : -67.01 Mean : 18.07 Mean : -38.75   
## 3rd Qu.:2.000 3rd Qu.: 2.00 3rd Qu.: 80.00 3rd Qu.: 33.07   
## Max. :5.000 Max. : 36.37 Max. :1760.00 Max. : 33.94   
## Start.Long End.Lat End.Long   
## Min. :-999.00 Min. :-999.00 Min. :-999.00   
## 1st Qu.: -97.92 1st Qu.: 31.83 1st Qu.: -97.92   
## Median : -97.27 Median : 32.54 Median : -97.23   
## Mean :-159.49 Mean : -38.74 Mean :-159.47   
## 3rd Qu.: -96.60 3rd Qu.: 33.08 3rd Qu.: -96.57   
## Max. : -95.32 Max. : 33.95 Max. : -95.31

########## dropping rows where values are -999  
tornado\_df = tornado\_df[!tornado\_df$Start.Lat ==-999,]   
str(tornado\_df)

## 'data.frame': 1859 obs. of 11 variables:  
## $ County : chr "Comanche" "Erath" "Denton" "Limestone" ...  
## $ Year : int 1950 1950 1950 1950 1951 1951 1952 1952 1952 1952 ...  
## $ Date : chr "4/29/1950" "2/12/1950" "5/29/1950" "2/12/1950" ...  
## $ Time\_CST : chr "15:30:00" "1:15:00" "17:00:00" "6:10:00" ...  
## $ Rating : int 1 1 1 2 4 2 1 0 4 2 ...  
## $ Length : num 11.5 2.3 3.4 3.4 2 14.5 13.8 0.2 2.3 4.9 ...  
## $ Width : int 200 233 267 100 200 20 33 10 200 200 ...  
## $ Start.Lat : num 31.9 32.1 33.4 31.5 33.4 ...  
## $ Start.Long: num -98.6 -98.3 -96.9 -96.5 -98.8 ...  
## $ End.Lat : num 31.7 32.1 33.3 31.6 33.4 ...  
## $ End.Long : num -98.6 -98.3 -96.9 -96.5 -98.7 ...

summary(tornado\_df)

## County Year Date Time\_CST   
## Length:1859 Min. :1950 Length:1859 Length:1859   
## Class :character 1st Qu.:1973 Class :character Class :character   
## Mode :character Median :1987 Mode :character Mode :character   
## Mean :1987   
## 3rd Qu.:2002   
## Max. :2019   
## Rating Length Width Start.Lat   
## Min. :0.0000 Min. : 0.010 Min. : 0.00 Min. :30.57   
## 1st Qu.:0.0000 1st Qu.: 0.200 1st Qu.: 10.00 1st Qu.:32.03   
## Median :1.0000 Median : 0.500 Median : 33.00 Median :32.60   
## Mean :0.8887 Mean : 2.171 Mean : 93.57 Mean :32.53   
## 3rd Qu.:1.0000 3rd Qu.: 2.100 3rd Qu.: 100.00 3rd Qu.:33.10   
## Max. :5.0000 Max. :36.370 Max. :1760.00 Max. :33.94   
## Start.Long End.Lat End.Long   
## Min. :-99.12 Min. :30.57 Min. :-99.09   
## 1st Qu.:-97.75 1st Qu.:32.05 1st Qu.:-97.72   
## Median :-97.20 Median :32.62 Median :-97.17   
## Mean :-97.17 Mean :32.54 Mean :-97.14   
## 3rd Qu.:-96.55 3rd Qu.:33.13 3rd Qu.:-96.52   
## Max. :-95.32 Max. :33.95 Max. :-95.31

head(tornado\_df)

###Exploring the time of the year tornadoes most likely to occur.

library(plotly)

## Warning: package 'plotly' was built under R version 3.6.3

##   
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':  
##   
## last\_plot

## The following object is masked from 'package:stats':  
##   
## filter

## The following object is masked from 'package:graphics':  
##   
## layout

library(magrittr)

## Warning: package 'magrittr' was built under R version 3.6.3

##   
## Attaching package: 'magrittr'

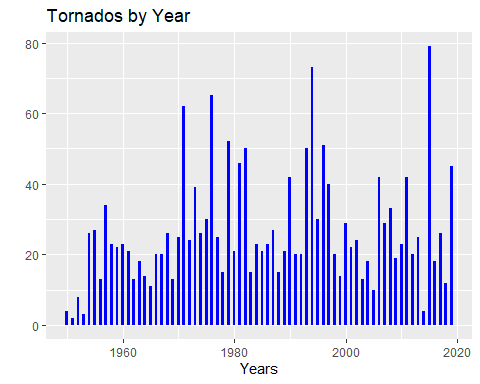
## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

# Tornadoes by the year from 1880 to 2019 in the Texas counties.  
summary(tornado\_df$Year)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1950 1973 1987 1987 2002 2019

# Median number of tornadoes occured around 1985  
# Mean of tornadoes occured around 1982  
#x=year y = number of tornadoes  
qplot(tornado\_df$Year,  
 geom="histogram",  
 binwidth = 0.5,   
 main = "Tornados by Year",   
 xlab = "Years",   
 fill=I("blue"))



# There is a negative skew from around 1880 to 1950. It is possible they not had proper technology to handle the data.  
# In 2015 there is a peak of tornadoes occuring within Texas.

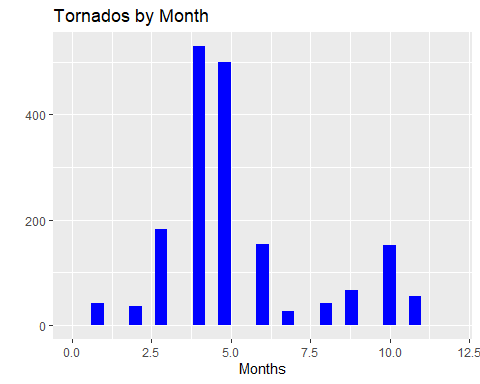
# Exploring the time of the year tornadoes most likely to occur.

#Modifying months  
months <- as.integer(format(as.Date(tornado\_df$Date, format="%m/%d/%Y"),"%m"))  
summary(months)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 4.000 5.000 5.557 6.000 12.000

# From this data storms will most likely occur between April and June.  
  
#x=months y = number of tornadoes  
qplot(months,  
 geom="histogram",  
 main = "Tornados by Month",   
 xlab = "Months",   
 fill=I("blue"),  
 xlim = c(0,12),  
 binwidth=0.4)

## Warning: Removed 2 rows containing missing values (geom\_bar).



#Exploring the time of day tornadoes most likely to occur.

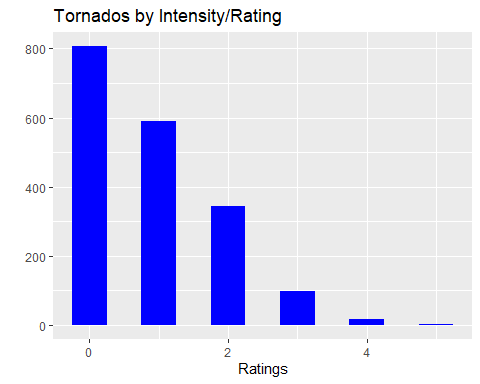
#Correcting time  
tornado\_df$time <- as.POSIXct(strptime(tornado\_df$Time\_CST, format="%H:%M:%S"))  
R <- paste('1970-01-01')  
R <- strptime(tornado\_df$time, format="%Y-%m-%d %H:%M:%S")  
  
#x=Time of Day y = number of tornadoes  
plot\_ly(x = (as.numeric(R) \* 1000), type = "histogram") %>%   
 layout(xaxis=list(type="date", tickformat="%H:%M:%S"))

## PhantomJS not found. You can install it with webshot::install\_phantomjs(). If it is installed, please make sure the phantomjs executable can be found via the PATH variable.

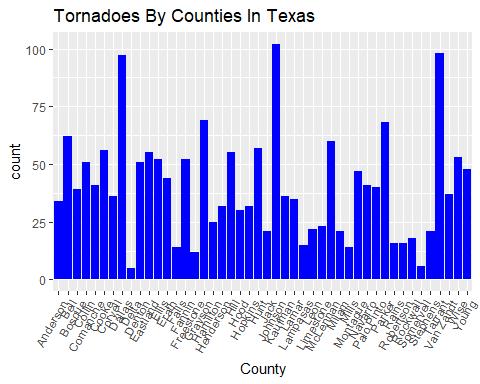
# Looking at the histogram, most likely to occur between 13:30 and 22:00.   
# The data is negatively skewed and this could account for the tracking of the data between 1800 to 1950.

# Exploring tornado intensity

# Most tornadoes are category 0  
#x=rating y = number of tornadoes  
  
qplot(tornado\_df$Rating,  
 geom="histogram",  
 main = "Tornados by Intensity/Rating",   
 xlab = "Ratings",   
 fill=I("blue"),  
 binwidth=0.5)

 # Exploring tornado by counties in Texas

# More tornadoes have occured in Johnson county followed by Dallas and Tarrant.   
#x=Counties In Texas y = number of tornadoes  
ggplot(data = tornado\_df,aes(x=County))+  
 geom\_bar(fill=I("blue"))+ theme(axis.text.x = element\_text(angle = 60, hjust = 1))+ ggtitle("Tornadoes By Counties In Texas")



# Analyzing data through correlation.

library(lubridate)

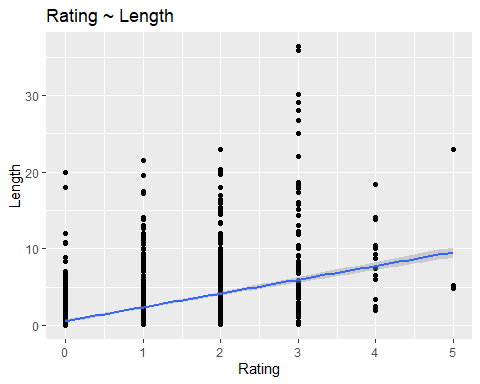
##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

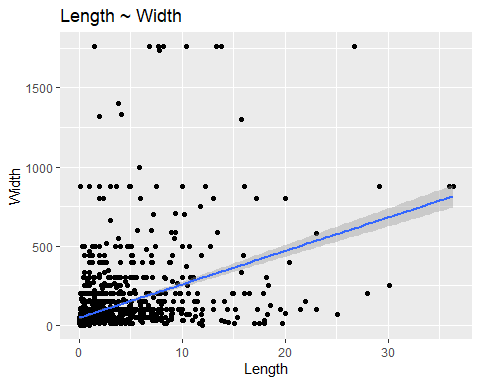
library(dplyr)  
  
  
new\_data = tornado\_df %>% mutate(Month = month(mdy(tornado\_df$Date)), Day = day(mdy(tornado\_df$Date))) %>% select(-County, -Date, -Time\_CST, -time)  
  
cor(new\_data)

## Year Rating Length Width  
## Year 1.0000000000 -0.38621347 -0.0794522255 0.064581256  
## Rating -0.3862134680 1.00000000 0.4540818951 0.359020579  
## Length -0.0794522255 0.45408190 1.0000000000 0.426074547  
## Width 0.0645812560 0.35902058 0.4260745467 1.000000000  
## Start.Lat -0.0334916913 -0.03296075 -0.0367749658 -0.001254270  
## Start.Long 0.0001732233 0.09482503 0.0539201014 0.054438316  
## End.Lat -0.0329776689 -0.01611386 0.0009315138 0.016558111  
## End.Long -0.0065056930 0.11560481 0.1032968206 0.074498965  
## Month -0.0655015636 0.02806085 0.0187894815 -0.001425843  
## Day 0.0747064255 -0.06889164 -0.0306166242 0.059943993  
## Start.Lat Start.Long End.Lat End.Long  
## Year -0.03349169 0.0001732233 -0.0329776689 -0.006505693  
## Rating -0.03296075 0.0948250338 -0.0161138562 0.115604814  
## Length -0.03677497 0.0539201014 0.0009315138 0.103296821  
## Width -0.00125427 0.0544383163 0.0165581112 0.074498965  
## Start.Lat 1.00000000 0.1458395938 0.9985657601 0.144651889  
## Start.Long 0.14583959 1.0000000000 0.1499647937 0.998118291  
## End.Lat 0.99856576 0.1499647937 1.0000000000 0.150242801  
## End.Long 0.14465189 0.9981182906 0.1502428014 1.000000000  
## Month -0.01066690 0.0512583277 -0.0069211401 0.049803563  
## Day -0.03506764 -0.0128626059 -0.0356120944 -0.015038086  
## Month Day  
## Year -0.065501564 0.07470643  
## Rating 0.028060852 -0.06889164  
## Length 0.018789482 -0.03061662  
## Width -0.001425843 0.05994399  
## Start.Lat -0.010666899 -0.03506764  
## Start.Long 0.051258328 -0.01286261  
## End.Lat -0.006921140 -0.03561209  
## End.Long 0.049803563 -0.01503809  
## Month 1.000000000 -0.03628878  
## Day -0.036288780 1.00000000

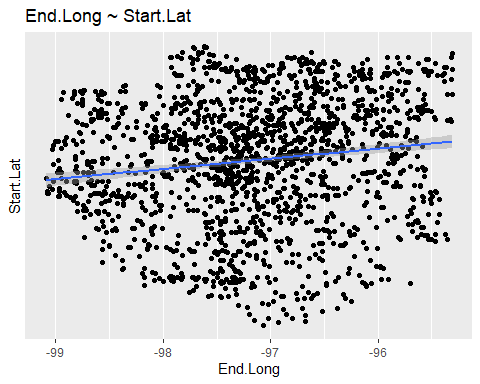
#Model to explore linear correlations  
library(ggplot2)  
#Rating ~ Length  
ggplot(data = new\_data, aes(x = Rating, y = Length)) + ggtitle("Rating ~ Length") +  
 geom\_point() + geom\_smooth(method = "lm")



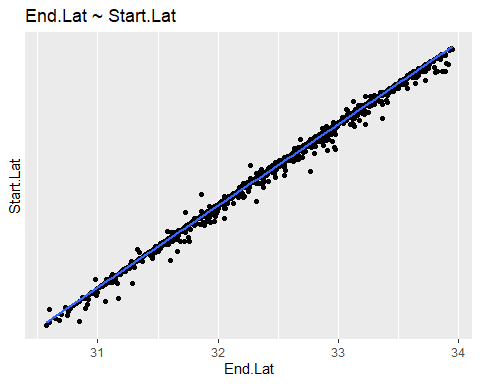
#"Length ~ Width"  
ggplot(data = new\_data, aes(x = Length, y = Width)) + ggtitle("Length ~ Width") +  
 geom\_point() + geom\_smooth(method = "lm")



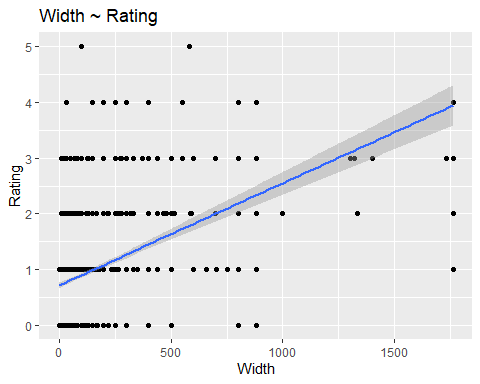
#"End.Long ~ Start.Lat"  
ggplot(data = new\_data, aes(x = End.Long, y = Start.Lat)) + ggtitle("End.Long ~ Start.Lat") +  
 geom\_point() + scale\_y\_log10() + geom\_smooth(method = "lm")



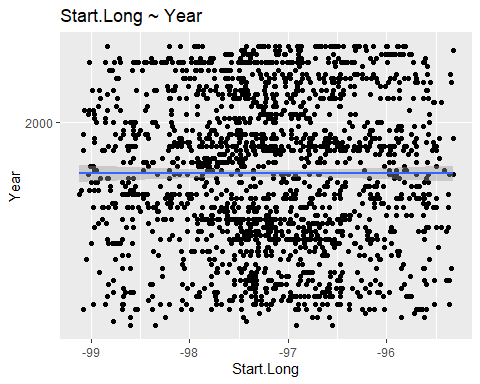
#"End.Lat ~ Start.Lat"  
ggplot(data = new\_data, aes(x = End.Lat, y = Start.Lat)) + ggtitle("End.Lat ~ Start.Lat") +  
 geom\_point() + scale\_y\_log10() + geom\_smooth(method = "lm")



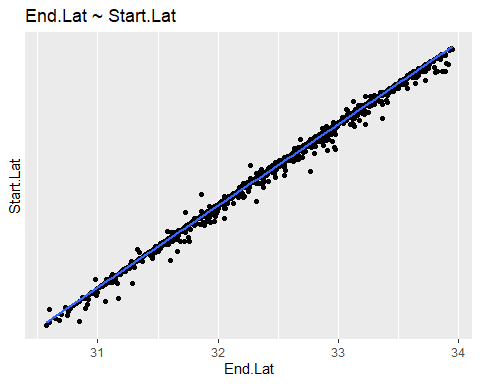
#"Width ~ Rating"  
ggplot(data = new\_data, aes(x = Width, y = Rating)) + ggtitle("Width ~ Rating") +  
 geom\_point() + geom\_smooth(method = "lm")



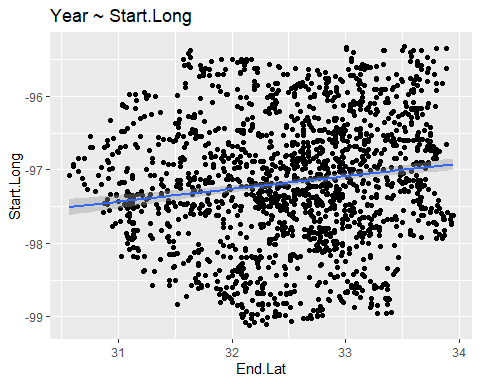
#"Start.Long ~ Year"  
ggplot(data = new\_data, aes(x = Start.Long, y = Year)) + ggtitle("Start.Long ~ Year") +  
 geom\_point() + scale\_y\_log10() + geom\_smooth(method = "lm")



#"End.Lat ~ Start.Lat"  
ggplot(data = new\_data, aes(x = End.Lat, y = Start.Lat)) + ggtitle("End.Lat ~ Start.Lat") +  
 geom\_point() + scale\_y\_log10() + geom\_smooth(method = "lm")



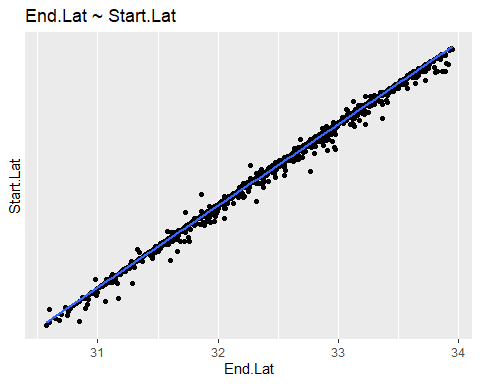
#"Year ~ Start.Long"  
ggplot(data = new\_data, aes(x = End.Lat, y = Start.Long)) + ggtitle("Year ~ Start.Long") +  
 geom\_point() + geom\_smooth(method = "lm")

 ###Linear Models

library(ggplot2)  
reg\_lm1 <- lm(Start.Lat~ End.Lat,data=tornado\_df)  
summary(reg\_lm1)

##   
## Call:  
## lm(formula = Start.Lat ~ End.Lat, data = tornado\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.35883 0.00280 0.01191 0.01407 0.27125   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.066271 0.040402 1.64 0.101   
## End.Lat 0.997567 0.001241 803.73 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04073 on 1857 degrees of freedom  
## Multiple R-squared: 0.9971, Adjusted R-squared: 0.9971   
## F-statistic: 6.46e+05 on 1 and 1857 DF, p-value: < 2.2e-16

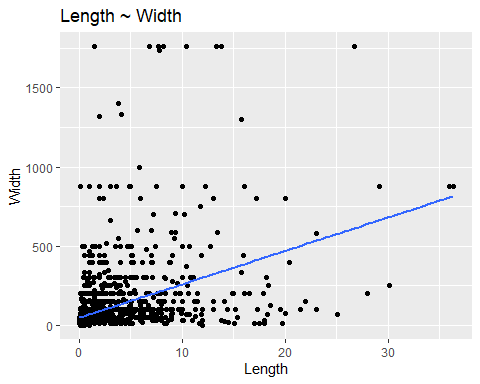
ggplot(data = new\_data, aes(x = End.Lat, y = Start.Lat)) + ggtitle("End.Lat ~ Start.Lat") +  
 geom\_point() + scale\_y\_log10() + geom\_smooth(method = "lm")



reg\_lm3 <- lm(Length~ Width,data=tornado\_df)  
  
summary(reg\_lm3)

##   
## Call:  
## lm(formula = Length ~ Width, data = tornado\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.9976 -1.3527 -1.0129 -0.0495 27.4378   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.3667613 0.0889600 15.36 <2e-16 \*\*\*  
## Width 0.0085970 0.0004236 20.30 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.434 on 1857 degrees of freedom  
## Multiple R-squared: 0.1815, Adjusted R-squared: 0.1811   
## F-statistic: 411.9 on 1 and 1857 DF, p-value: < 2.2e-16

# Linear Model  
library(ggplot2)  
ggplot(data = tornado\_df, aes(x = Length, y = Width)) + ggtitle("Length ~ Width") +  
 geom\_point() +   
 geom\_smooth(method = "lm", se = FALSE)



# The original question was to demonostrate the frequency of storms that occured in Dallas, Texas.

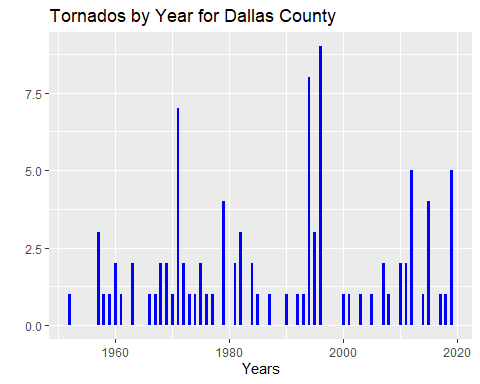
dallas = tornado\_df[tornado\_df$County =='Dallas',]   
str(dallas)

## 'data.frame': 97 obs. of 12 variables:  
## $ County : chr "Dallas" "Dallas" "Dallas" "Dallas" ...  
## $ Year : int 1952 1957 1957 1957 1958 1959 1960 1960 1961 1963 ...  
## $ Date : chr "3/2/1952" "6/12/1957" "4/3/1957" "4/2/1957" ...  
## $ Time\_CST : chr "22:00:00" "8:10:00" "10:50:00" "16:15:00" ...  
## $ Rating : int 0 3 0 3 2 3 0 0 2 0 ...  
## $ Length : num 0.2 0.5 0.5 17.3 3 7.3 0.1 0.1 0.3 0.5 ...  
## $ Width : int 10 10 10 100 300 67 10 13 33 17 ...  
## $ Start.Lat : num 33 32.7 32.6 32.7 32.9 ...  
## $ Start.Long: num -96.6 -96.9 -96.9 -96.8 -96.6 ...  
## $ End.Lat : num 33 32.7 32.6 32.9 32.9 ...  
## $ End.Long : num -96.6 -96.9 -96.9 -96.9 -96.6 ...  
## $ time : POSIXct, format: "2020-07-28 22:00:00" "2020-07-28 08:10:00" ...

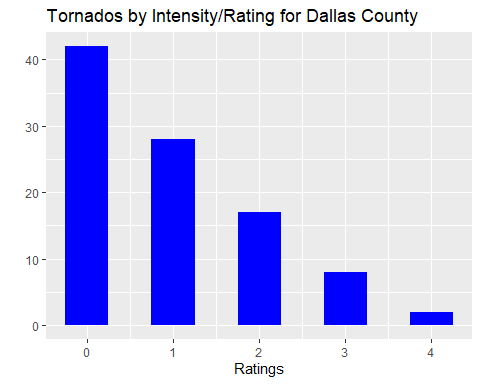
summary(dallas)

## County Year Date Time\_CST   
## Length:97 Min. :1952 Length:97 Length:97   
## Class :character 1st Qu.:1971 Class :character Class :character   
## Mode :character Median :1993 Mode :character Mode :character   
## Mean :1989   
## 3rd Qu.:2005   
## Max. :2019   
## Rating Length Width Start.Lat   
## Min. :0.0000 Min. : 0.010 Min. : 10.0 Min. :32.55   
## 1st Qu.:0.0000 1st Qu.: 0.500 1st Qu.: 13.0 1st Qu.:32.62   
## Median :1.0000 Median : 1.000 Median : 50.0 Median :32.76   
## Mean :0.9691 Mean : 2.334 Mean : 130.5 Mean :32.74   
## 3rd Qu.:2.0000 3rd Qu.: 2.010 3rd Qu.: 100.0 3rd Qu.:32.86   
## Max. :4.0000 Max. :28.000 Max. :1760.0 Max. :32.98   
## Start.Long End.Lat End.Long   
## Min. :-97.03 Min. :32.55 Min. :-97.03   
## 1st Qu.:-96.90 1st Qu.:32.63 1st Qu.:-96.87   
## Median :-96.78 Median :32.77 Median :-96.77   
## Mean :-96.79 Mean :32.76 Mean :-96.77   
## 3rd Qu.:-96.69 3rd Qu.:32.89 3rd Qu.:-96.67   
## Max. :-96.53 Max. :32.99 Max. :-96.51   
## time   
## Min. :2020-07-28 00:55:00   
## 1st Qu.:2020-07-28 12:08:00   
## Median :2020-07-28 16:25:00   
## Mean :2020-07-28 15:12:50   
## 3rd Qu.:2020-07-28 20:00:00   
## Max. :2020-07-28 22:50:00

## histograms for Dallas   
  
#On an average it appears around 9 tornadoes occurred in Dallas yearly from 1952 to 2019 with 40 percent being category 0.  
qplot(dallas$Year,  
 geom="histogram",  
 binwidth = 0.5,   
 main = "Tornados by Year for Dallas County",   
 xlab = "Years",   
 fill=I("blue"))



qplot(dallas$Rating,  
 geom="histogram",  
 main = "Tornados by Intensity/Rating for Dallas County",   
 xlab = "Ratings",   
 fill=I("blue"),  
 binwidth=0.5)



library(ggplot2)

## Summary

#1. Summarize the problem statement you addressed.

#On October 20, 2019 around 10 tornadoes hit Dallas, Texas. It left numerous individuals with homes. There were no notice signs until the tornadoes started to cause a lot of harm. Many began to say Dallas is not a decent place to live because it is in Tornado Alley which is where a high frequency of storms occurs. Because the storms occurred within a 2-mile radius of my home, it clearly is the motivation of this dataset. The quest to search for the best dataset began for the frequency of tornadoes within the Dallas, Texas area. There was not a lot of information in the first dataset I found. However, I found another that gave more information and listed the whole state of Texas. From those datasets, I can provide the frequency of tornadoes that occur in Dallas, Texas.

#2. Summarize how you addressed this problem statement (the data used and the methodology employed).

#In the original data set, I could not find any correlation of the data. After the 3rd week, I was able to find a dataset with counties in Texas at <https://data.world/dhs/historical-tornado-tracks/workspace/file?filename=Historical_Tornado_Tracks.csv>

#In my approach, I cleaned the informational indexes to expel all NA’s and - 999 to help comprehend the information. At that point, I plotted the information utilizing ggplots, and qqplots. As I was endeavoring to plot dated related factors, I needed to isolate the dates with the administrator work so as to plot the information. The date variable was an integer. The lubridate package was used to subset the dates to make date formatting simpler.

#In this clean dataset, I could not find out everything I wanted to know about the data. The histograms were reviewed with ggplots and qqplots histogram Then, correlations were performed on the variables. From there two linear regression models on data that correlated. However, the information related more to composition of a tornado like width, length, longitude, latitude. Obviously, those would relate. I added loess function in a few of the plots to see if there was any correlation also. Because this dataset did not provide the right information relating to weather conditions, I decided to narrow my research to the city of Dallas with yearly frequency of tornadoes and the category of tornadoes.

#3.Summarize the interesting insights that your analysis provided.

#In the state of Texas, the median number of tornadoes occurred around 1985 with the mean of tornadoes occurring around 1982. From this data, tornadoes will most likely occur between April and June. Looking at the histogram, most likely to occur between 13:30 and 22:00 central time and most tornadoes are EF0. The counties where most of tornadoes have occurred were Johnson county followed by Dallas and Tarrant.

#On an average, it appears 9 tornadoes occur in Dallas yearly from 1952 to 2019 with a 40% chance of it being an EF3 category. The average length within Dallas is 2 miles. The tornadoes more likely to occur between 1200 and 20:00 central time.

#4.Summarize the implications to the consumer (target audience) of your analysis.

#Tornadoes can hit within a matter of seconds. Whenever there are it is best to try to get to a safe area. Some are saying don’t purchase homes in Dallas. When looking at the graphs, they can strike anywhere in the state of Texas. It is likely they will occur in Dallas and Dallas is ranked 3 on the list of counties.

#Tornadoes can hit inside only seconds. At whatever point there are it is ideal to attempt to get to a protected territory. Some are stating don’t buy homes in Dallas. When taking a gander at the diagrams, they can strike anyplace in the province of Texas. It is likely they will happen in Dallas and Dallas is positioned 3 on the rundown of areas.

#5.Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

#Based upon my dataset, the categories limit the use to predict tornadoes in Dallas, From the data, start and end shows positive correlation. In order to predict tornadoes, other weather conditions datasets should be available. From there I could predict patterns. I plan on in the future bringing other datasets that could provide more warnings. I came close to being in the Tornado the day that it occurred. I was two hours away from being in the path of that storm. There was no warning until the tornado struck from the news.

#In view of my dataset, the categories limited my ability to predict tornadoes in Dallas forming. Some dataset does correlate, but only with the composition of a tornado. To anticipate tornadoes, other climate conditions datasets ought to be accessible. From that point I could foresee designs. I plan on later bringing different datasets that could give more alerts. I came close to being in the Tornado path the day it occurred. I There was no warning until the tornado struck from the news. In the future, I hope someone can give phone alerts to let everyone know they are coming. It is miracle that no one perished from the tornado in Dallas.