Hurricanes and Emergencies: Data Wrangling Steps

Part1: Data wrangling for files describing the Pacific and Atlantic hurricane tracks

We will be using the data files presented by the Department of the Interior of US Geological Survey, that describe the tracks of historic hurricanes originated in Pacific and Atlantic oceans. The datasets were downloaded from the webpages:

https://catalog.data.gov/dataset/historical-north-atlantic-tropical-cyclone-tracks-1851-2004-direct-download

https://catalog.data.gov/dataset/historical-eastern-north-pacific-tropical-cyclone-tracks-1949-2004-direct-download

1) We first import all necessary modules including the .dbf file reader DBF5 to obtain the Pacific and Atlantic hurricane tracks dataframes. We then read the tracks files into the data-frames df_Atl and df_Pac. Running .info() indicates that these files contain no missing values.

2) Since the hurricanes started receiving names only after 1949, and only after that year the data are more or less accurate, we will limit ourselves to hurricanes that occurred in 1950 or later. If a hurricane after 1949 is still not named, we will remove it from the database as well. Then, we will limit ourselves only with the storms that can be considered dangerous. We remove from the database the low-pressure systems, named 'SUBTROP', 'SUBTROP 2', 'SUBTROP 3', 'SUBTROP 4', that could not reach the dangerous limit, and consider only the depressions, storms (both subtropical and tropical) and hurricanes of all five categories.

3) The data-frames df_Atl and df_Pac contain many columns, but we will keep only the columns describing year, month, day, name, longitude of the center, latitude of the center and category of each hurricane. We will omit the columns describing internal pressure and sustained wind since they themselves determine the category.

4) It is convenient to assign an integer number ranging from 1 to 7 to each category of a hurricane based on its strength. The weakest low-pressure system, depression, is given 1, while the strongest one, category five

hurricane, is assigned 7. The new column is called 'EFFECTIVE STRENGTH'.

5) For the future analysis, we will create a table that contains: 1) the first longitude and latitude when hurricane is dangerous, 2) the last longitude or latitude when hurricane is still dangerous (or before making a landfall), 3) the total time travelled in hours, 4) the total distance travelled (in km), 5) the average wind strength, 6) the maximum effective strength and 7) the average effective strength when it is still dangerous. We first define the auxiliary functions that help to compute us 1) through 4). The total distance is calculated using the Haversine formula from the imported gpxy module, while the total number of hours is computed by taking the number of entries for a given hurricane and multiplying by 6 hrs. After that we use the combination of groupby and merging technique to create the new table containing the aforementioned information for both Pacific and Atlantic hurricanes. The new data frame for Atlantic hurricanes contains 527 rows, while the data frame for Pacific hurricanes 679 rows.

6) At the next step, we will create a new column that gives the average speed of each hurricane during the time when it is still dangerous.

7) Finally, we create the combined list of the names of the hurricanes (Atlantic and Pacific) to be used in Part 2.

Part 2: Data wrangling for 'Emergencies_database.csv'

Here we will do the data preparation for the csv-file 'Emergencies_database.csv' taken from the KAGGLE open source database. Contrary to another interesting data file 'DisasterDeclarationsSummaries.csv', this file contains information about all counties in all states where the state of emergency because of a hurricane was declared. Counting the counties affected by disastrous hurricanes is much more informative than counting just the states. The file can be downloaded from;

https://www.kaggle.com/fema/federal-disasters

1) We first select the rows with the column 'Disaster Type' equal to 'Hurricane' and 'Typhoon' from 'Emergencies_database.csv' that initially had 46185 rows. Now the data-frame has only 8883 rows.

2) Now we choose the columns 'Declaration Date', 'State', 'County', 'Disaster Title' that are only interesting to us, and look at the missing values. Some values are missing from the 'County' column, and we fill them with the value 'Some name' regarding each such entry as a distinct county.

3) We then recast the 'Declaration Date' column in the standard datetime format and create three separate columns, corresponding to year, month and day with the names to match the hurricane tracks data-frames. There are no missing or incorrectly entered data in these columns. The earliest year is 1954 and the latest is 2016.

4) The majority of values in the column 'Disaster Title' contains the name of a hurricane. Thus, our next task will be to single out the name of a hurricane in each entry of the column. To do this, we capitalize the entry and remove everything that does not contain a name from the combined list of names for Atlantic and Pacific hurricanes created from the hurricane tracks data-frames.

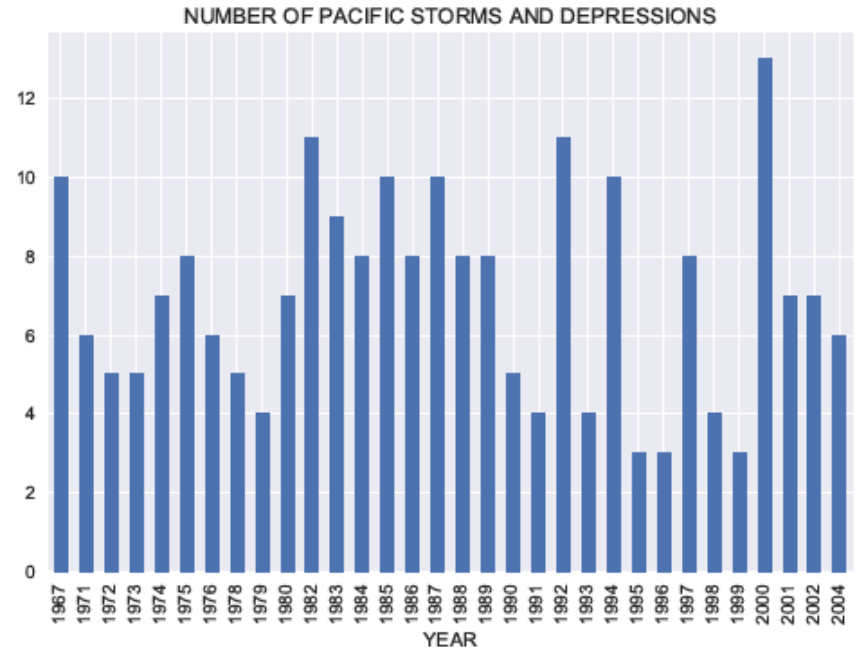
5) The new data-frame contains 8883 rows but 1664 rows are empty and do not contain the name of a hurricane. We will drop those rows and rename the name of the column 'Disaster Title' to name. The new data-frame contains 7219 rows, but examining it further we see that the sensible data about the number of counties affected is for hurricanes dated by 1965 and later. Thus we slightly trim our data-frame to arrive at the new data-frame containing 7207 rows.

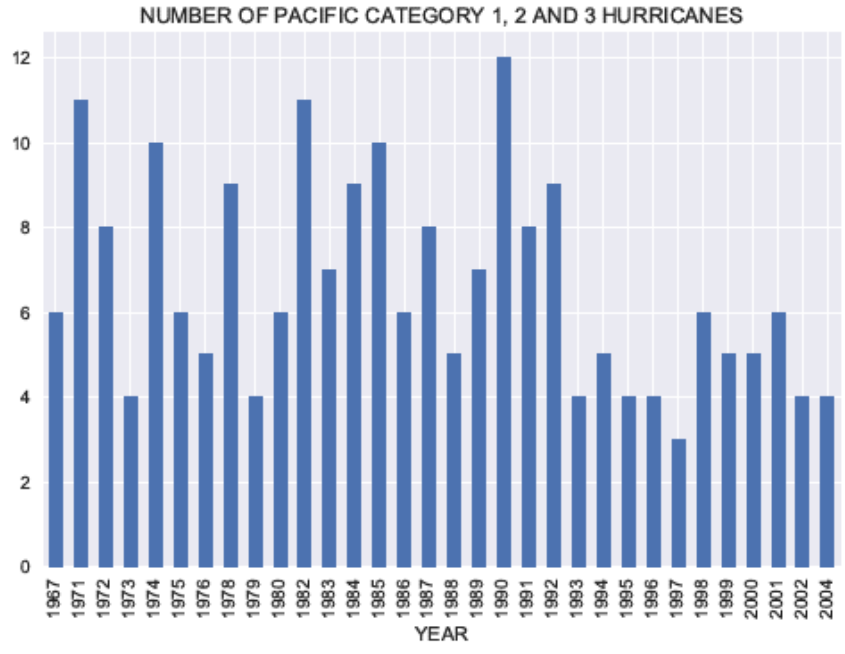
6) Finally, we will count the number of affected counties grouping them by year and name. The new data-frame is called df_emerg_new.

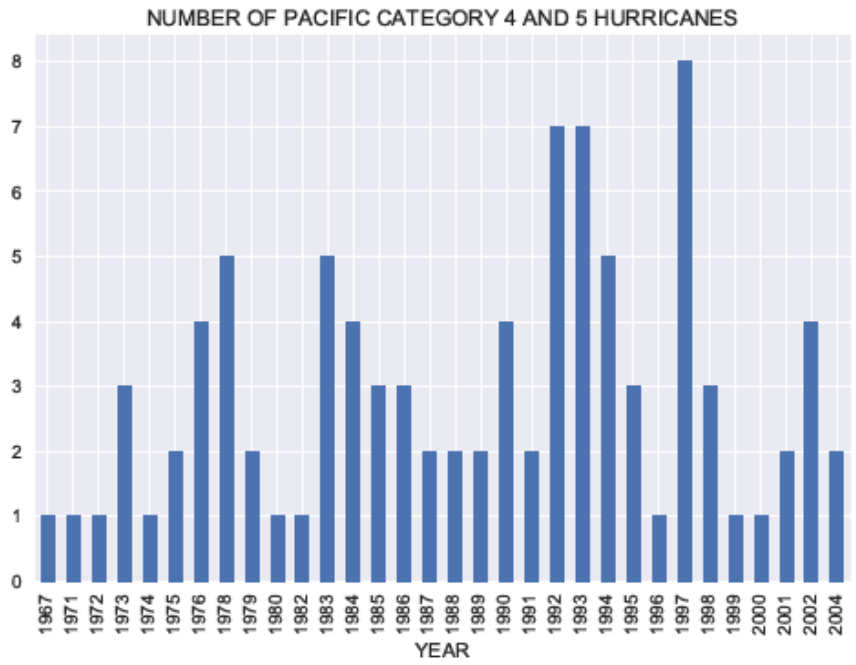
Part 3: Data Storytelling - Hurricane tracks datasets

Here we will make some plots to gain insights about tendencies that may require further investigation. To do this, we first import the plotting tools.

1) We first look at how the number of Pacific storms of different categories depends on year, and try to determine the tendencies.

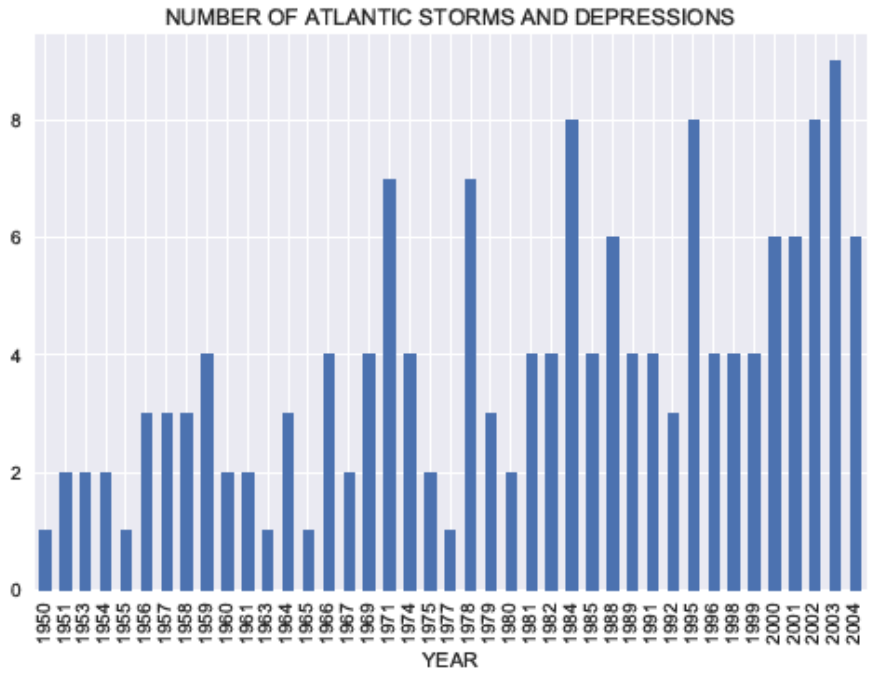


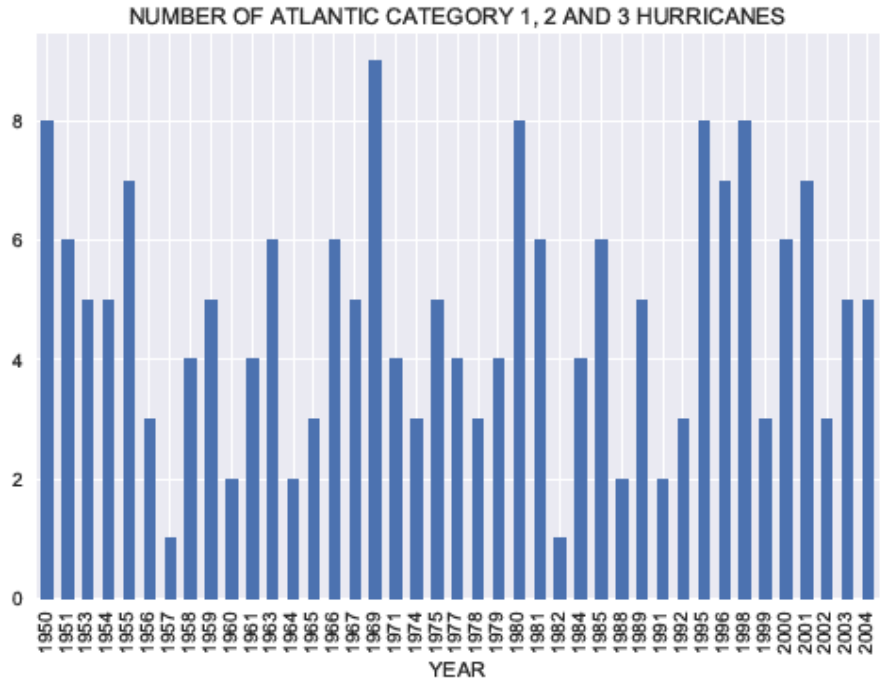


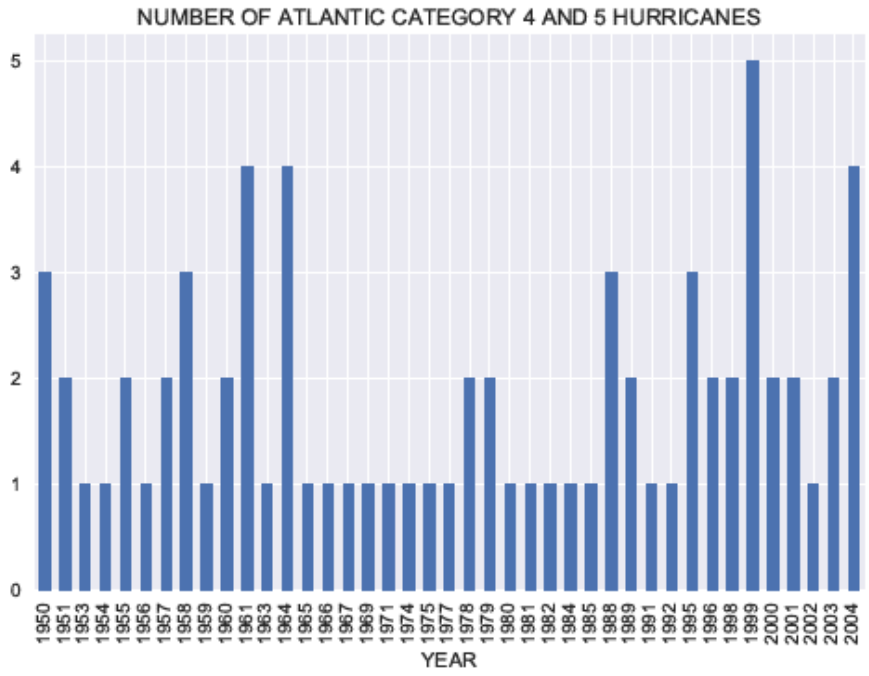


Looking at the plots for the numbers of Pacific low-pressure systems, we come to the conclusion that the numbers of hurricanes of all strengths display large fluctuations from year to year. One can distinguish between the high-frequency hurricane seasons and the low-frequency ones. One can see that the number of category 1,2 or 3 hurricanes was relatively small for all years between 1994 and 2004.

2) Next, we look at how the number of Atlantic storms of different categories depends on year, and try to determine similar tendencies.


NUMBER OF ATLANTIC STORMS AND DEPRESSIONS
0
2
4
6
8
1950 1951 1953 1954 1955 1956 1957 1958 1959 1960 1961 1963 1964 1965 1966 1967 1969 1971 1974 1975 1977 1978 1979 1980 1981 1982 1984 1985 1988 1989 1991 1992 1995 1996 1998 1999 2000 2001 2002 2003 2004
YEAR

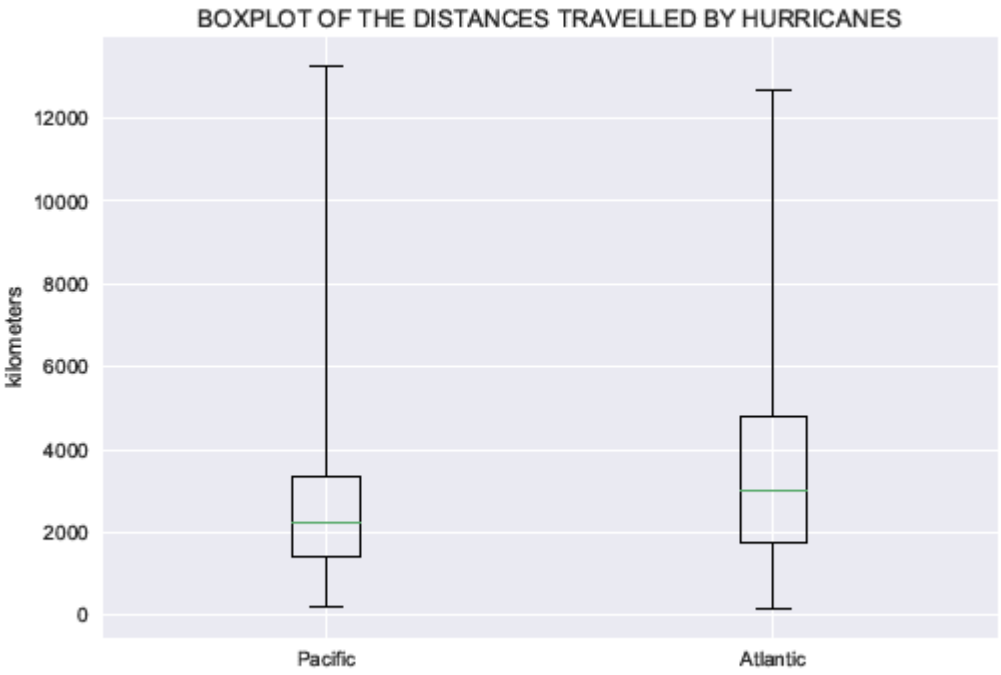

NUMBER OF ATLANTIC CATEGORY 1, 2 AND 3 HURRICANES
0
2
4
6
8
1950 1951 1953 1954 1955 1956 1957 1958 1959 1960 1961 1963 1964 1965 1966 1967 1969 1971 1974 1975 1977 1978 1979 1980 1981 1982 1984 1985 1988 1989 1991 1992 1995 1996 1998 1999 2000 2001 2002 2003 2004
YEAR

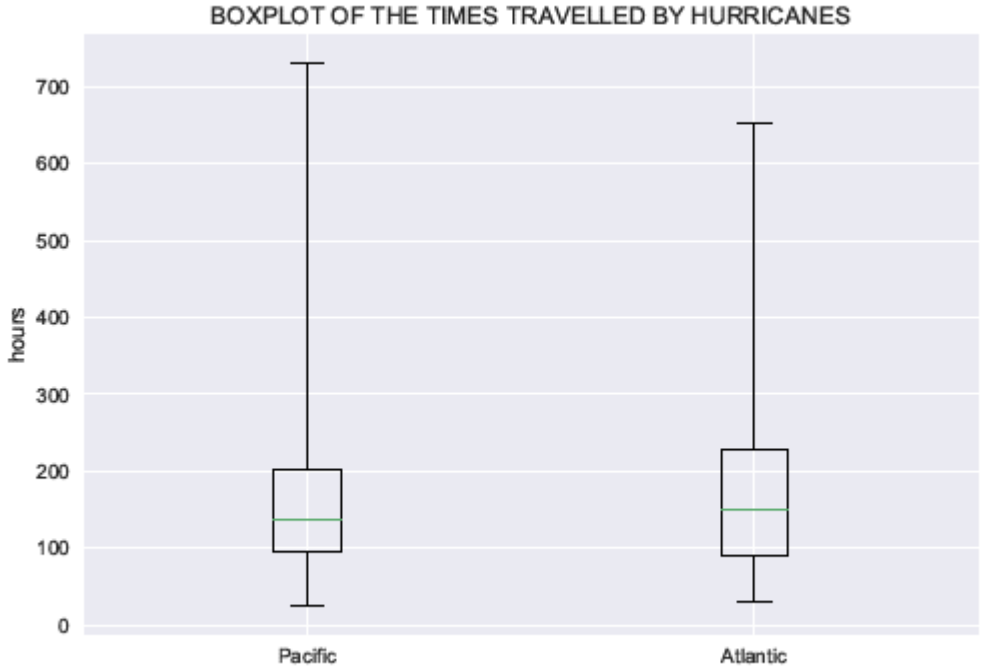


The analysis of plots reveals that the number of Atlantic low-pressure systems, that reached the level of a depression or storm, has a tendency to increase on average during the time interval between 1950 and 2004 years. At the same time, the number of hurricanes that reached the level of category 1, 2, or 3 is approximately the same. The average number of the most dangerous hurricanes of 4 and 5 categories is maximal during the last decade preceding 2004.

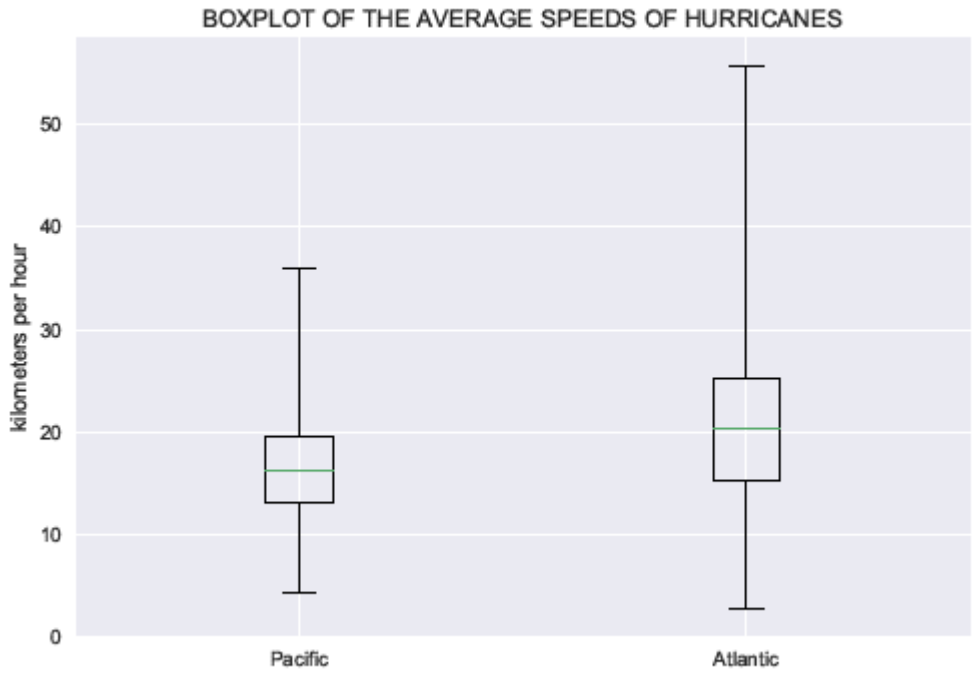
3) Then we look at the column that gives the distance travelled by each hurricane while it is considered dangerous. The descriptions as well as boxplot reveal that the travel distances vary dramatically. The range of variation is especially pronounced for Pacific hurricanes. However, the distribution of the distances for Atlantic hurricanes is broader than for the Pacific ones.

4) Next, we look at the column that gives the total time travelled by each hurricane while it is considered dangerous. The descriptions as well as boxplot reveal that the travel times vary a lot as well, with the variation range being more pronounced for Pacific hurricanes than for Atlantic ones. We should notice also that the range of top 25 percent times and distances is much larger than the range of bottom 75 percent.

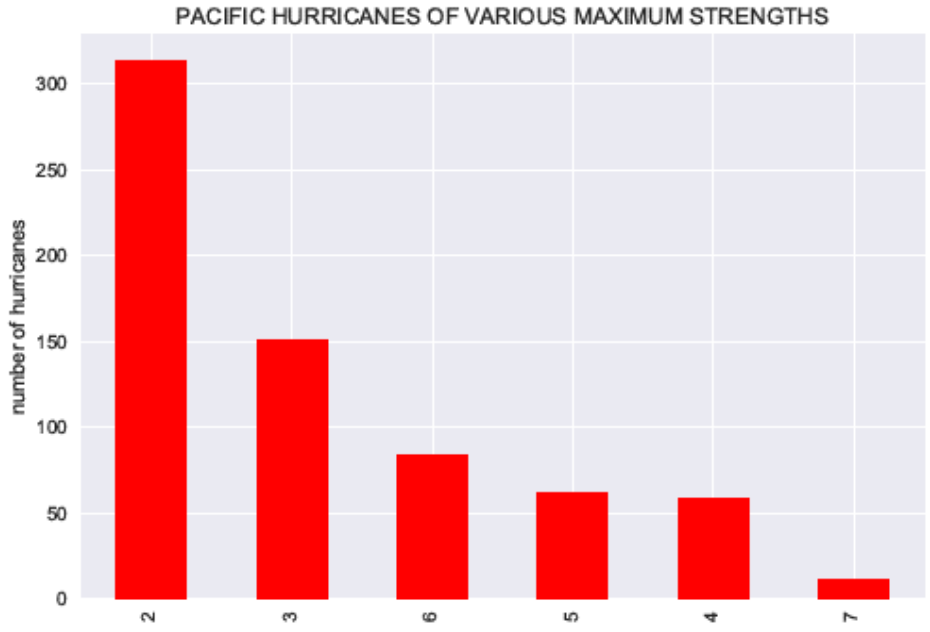


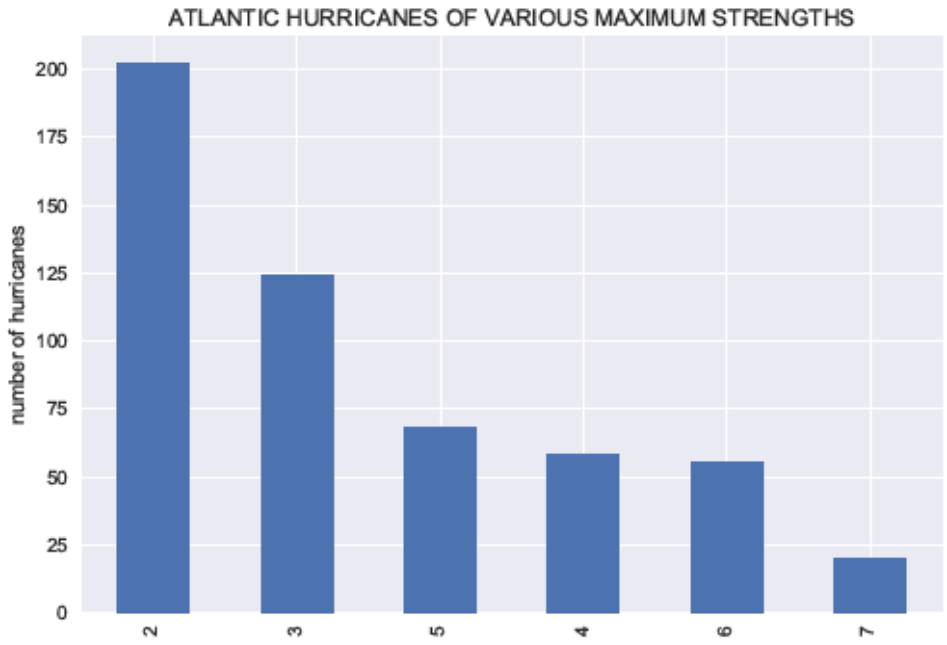


5) We then make a box plot of average speeds. We see that the spread of values of speeds for Atlantic hurricanes is much broader than for Pacific ones.

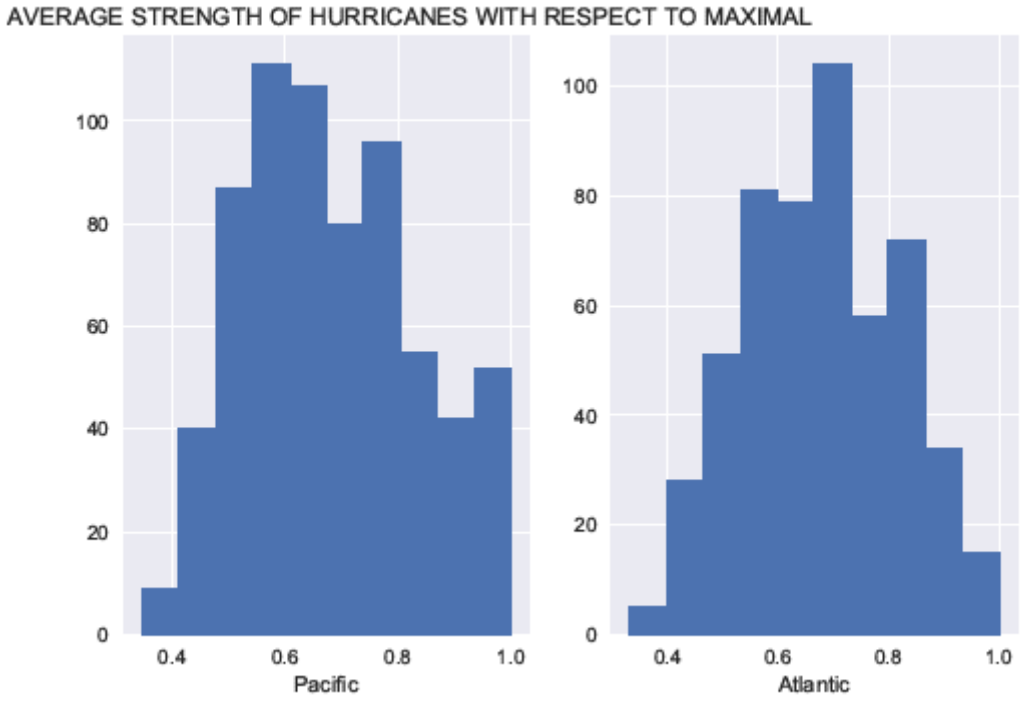


6) It is then a good idea to look at the distribution of the total number of hurricanes and storms with respect to the value of maximum strength. For Pacific hurricanes and storms, the number of lows that never reached the level of a hurricane is the largest compared to the number of those that reached that level becoming a hurricane of a certain maximum strength. The same is true for Atlantic lows, albeit the difference is not that pronounced.

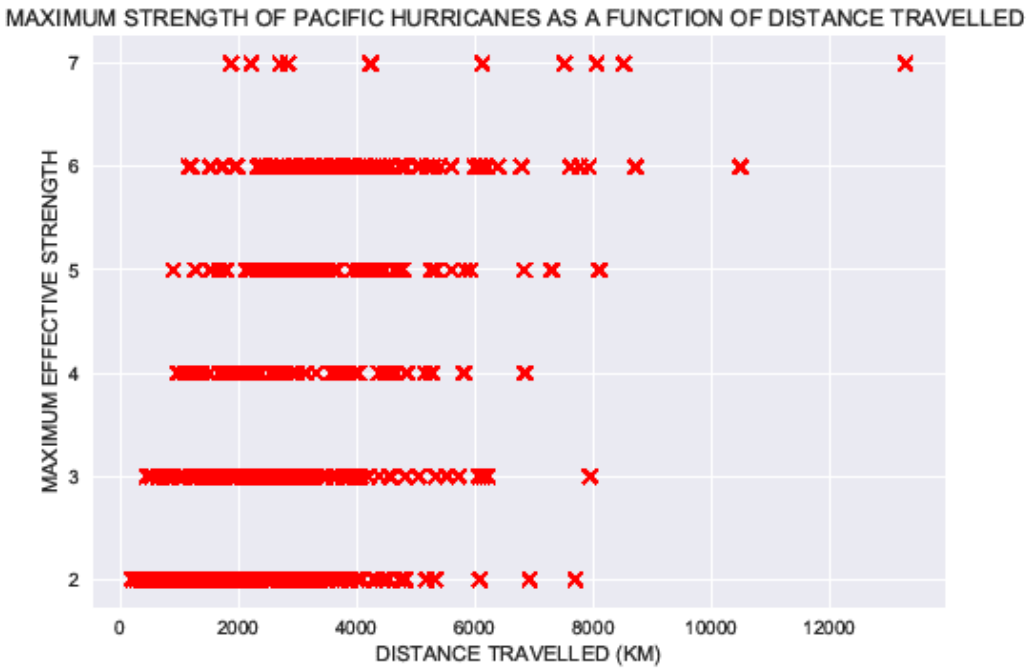


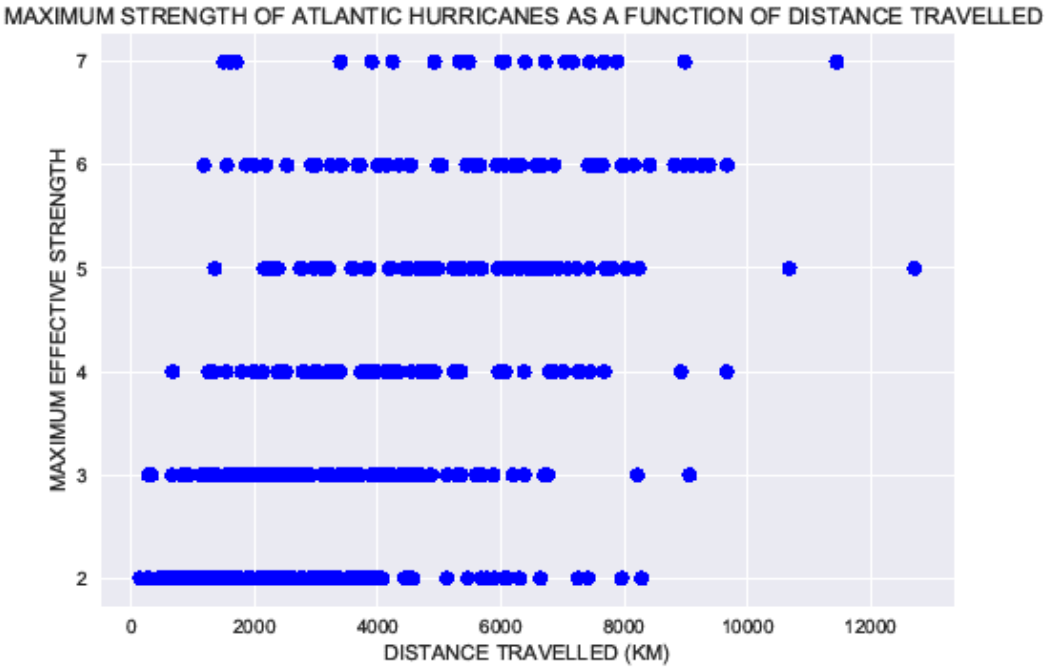


7) Let's make the histograms of the average strengths of hurricanes with respect to the maximum strength. We see that the average strengths of Pacific hurricanes have broader distribution than those of Atlantic ones. While the average strength of Atlantic hurricanes is peaked at around 0.7, the same parameter for Pacific ones is maximal at values of around 0.55.

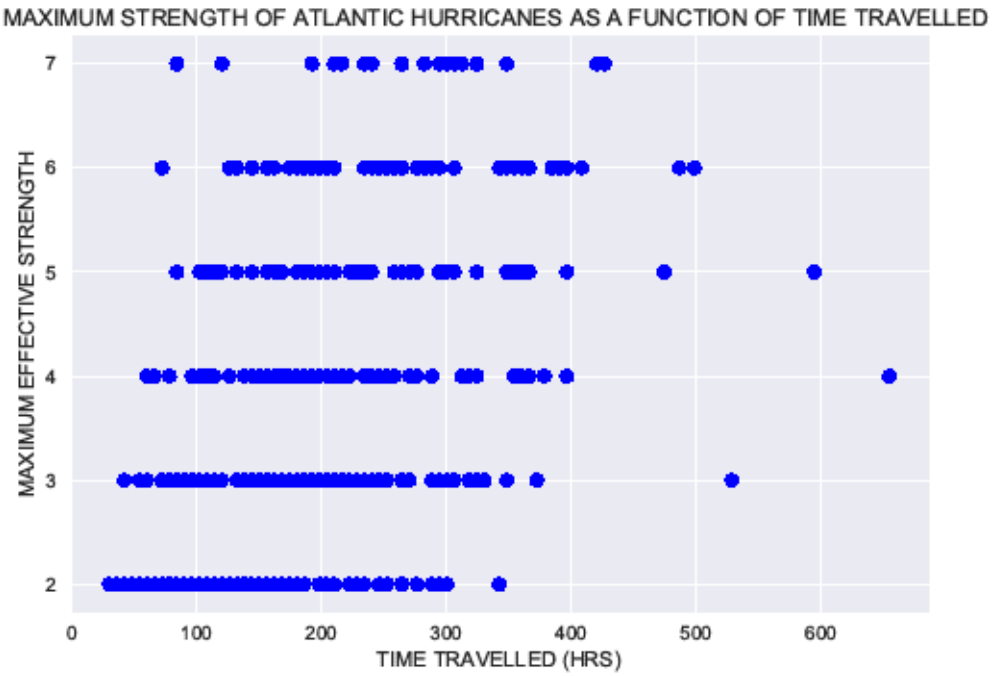


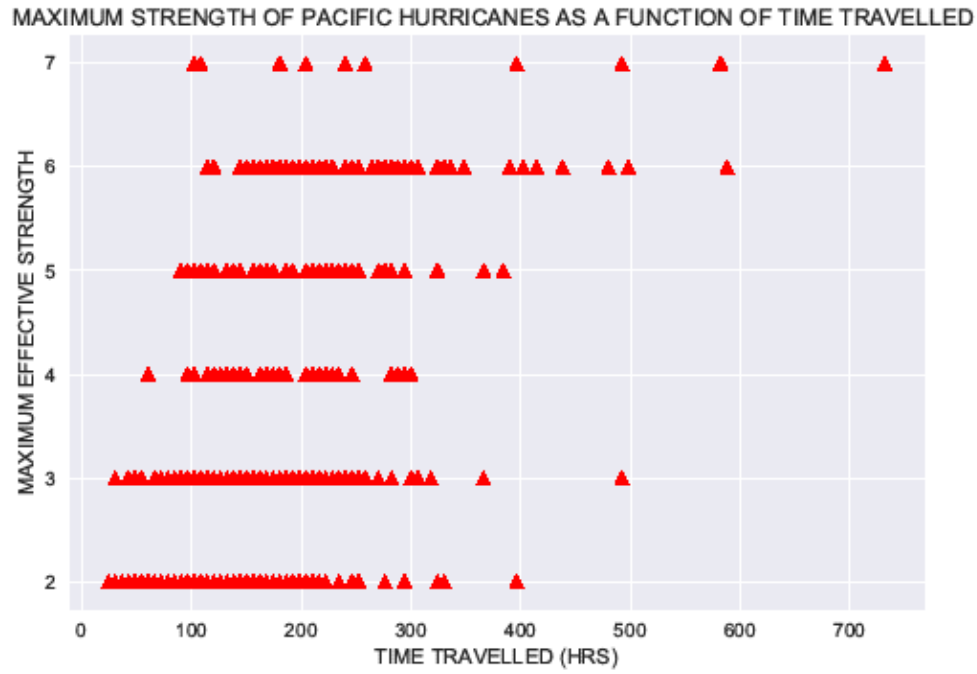
8) At the next step, we look at the dependence of maximum effective strength of hurricanes on the distance travelled while being dangerous. We see that the average distance travelled, as well as the spread of distances, has the tendency to increase with increasing maximum strength. This is almost true for both Pacific and Atlantic hurricanes. The minimal distance corresponding to hurricanes of various strengths increases as well for hurricanes of larger maximum strength.





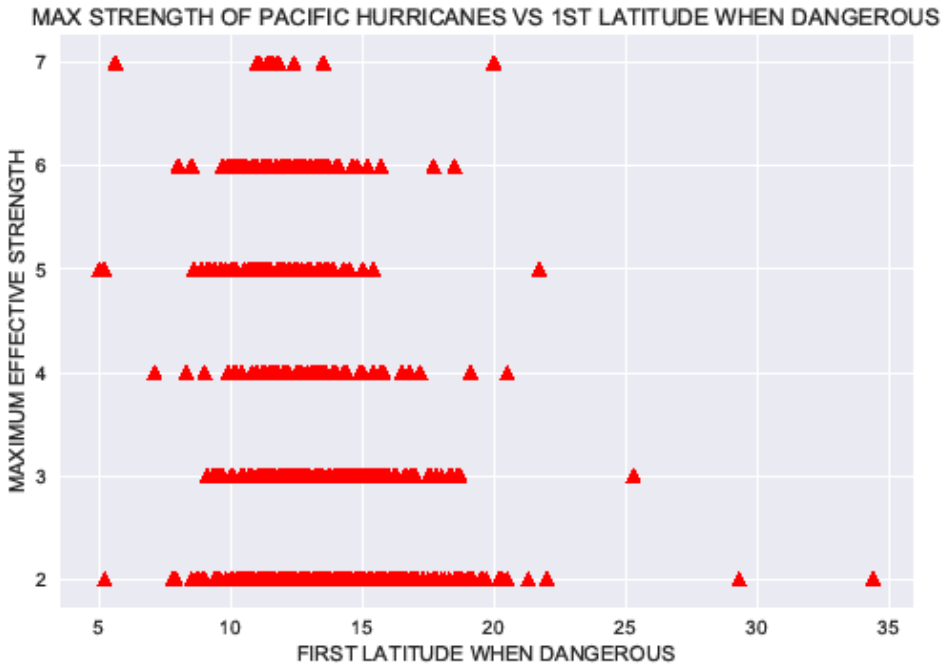
9) We then look at the dependence of the maximum effective strength of hurricanes on the time travelled while being dangerous. The dependence here is similar to that for the maximum strength vs distance travelled. The minimal and average times travelled by hurricanes have the tendency to increase with increasing value of the maximum strength.

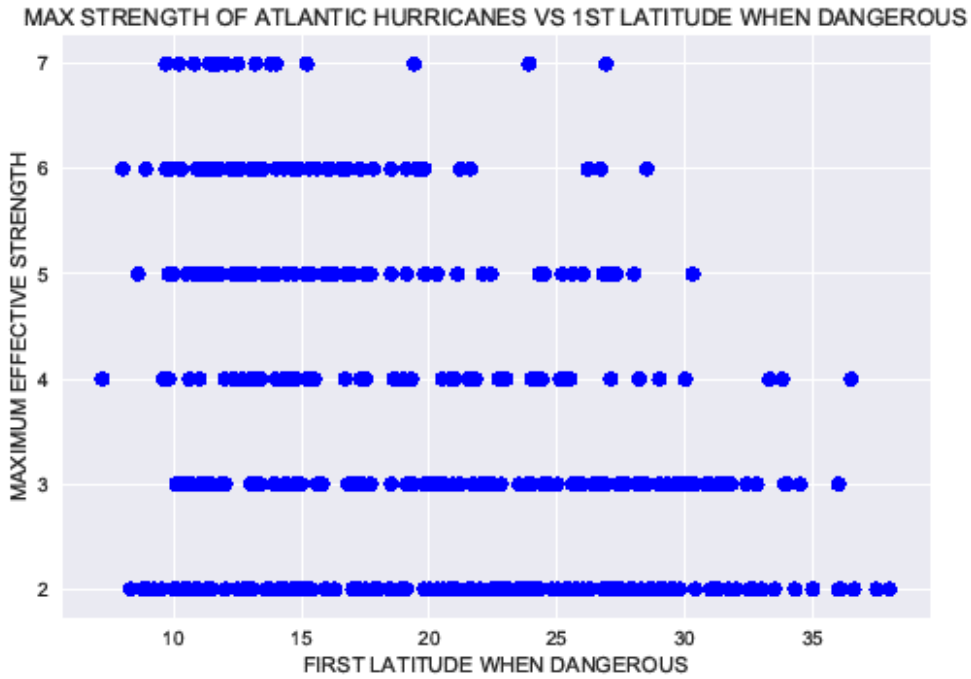


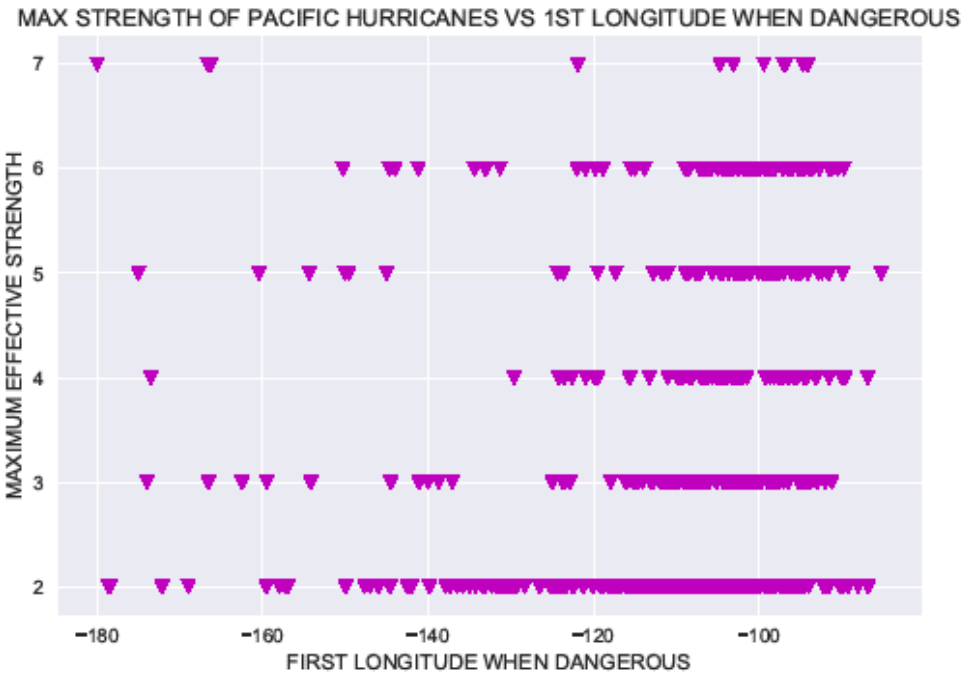


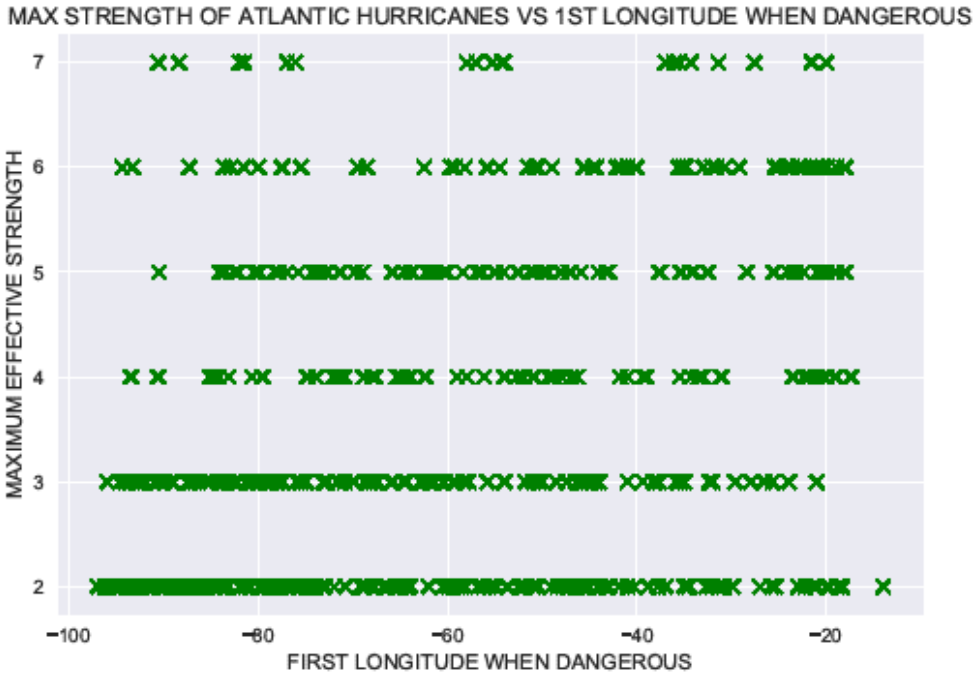
10) It behooves also to look at the dependence of the maximum strength of hurricanes on the latitude and longitude when they first became dangerous. The tendency that we observe for both Pacific and Atlantic

hurricanes is very interesting. The larger the maximal strength of a hurricane is, the smaller the range of latitudes and longitudes where the hurricane is more likely to form.

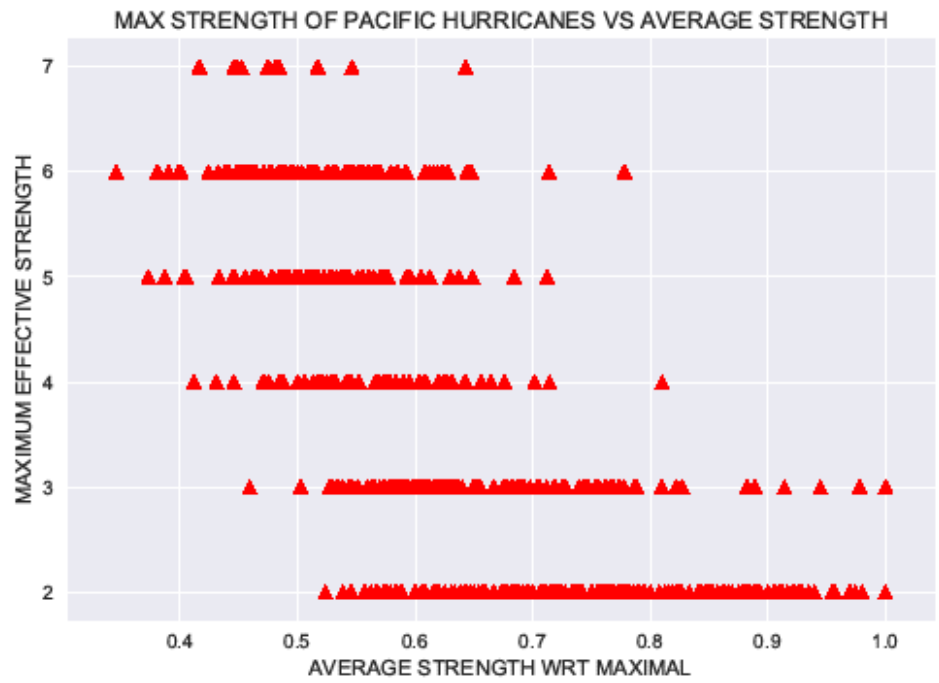


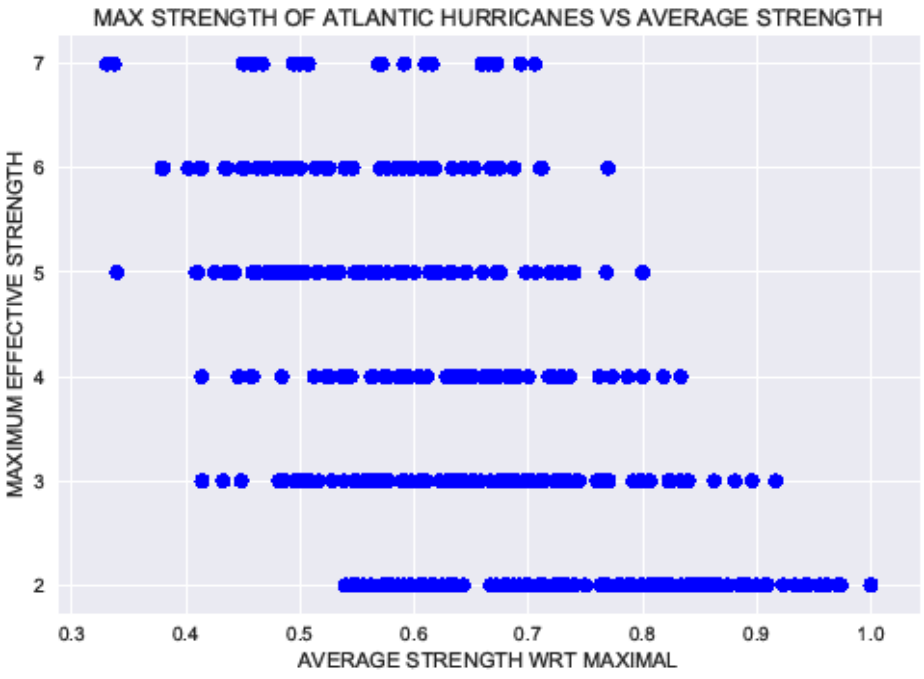






11) Next, we plot the maximum effective strength of a hurricane vs its average strength while that hurricane is considered dangerous. As one can see from the plots, the result is not very unexpected. For both Atlantic and Pacific hurricanes, the larger the maximum strength is, the smaller amount of time a hurricane is in the state with the strength close to maximal.





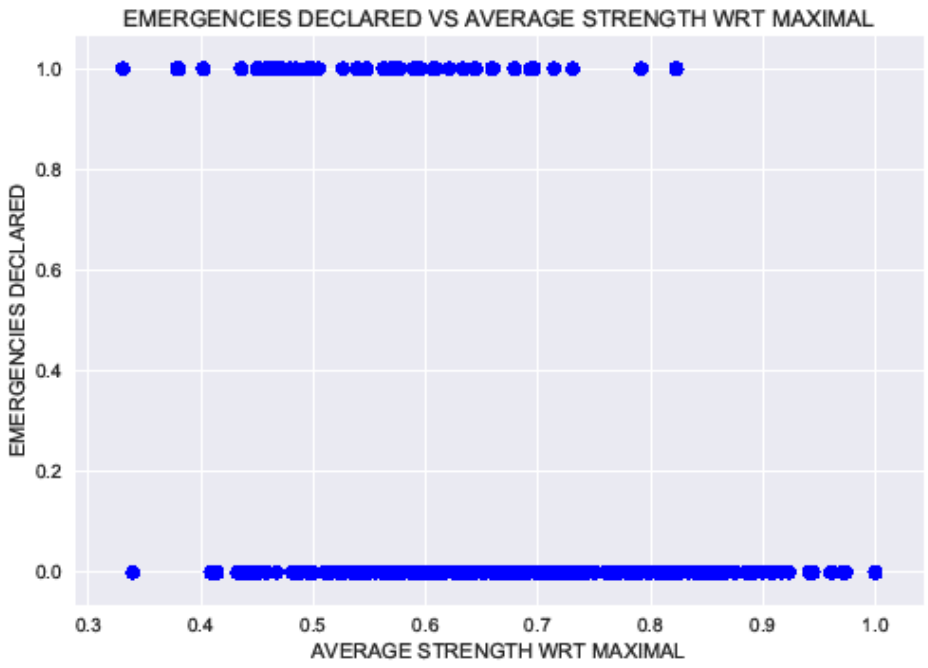
Part 4: Data Storytelling – Hurricanes and Emergencies

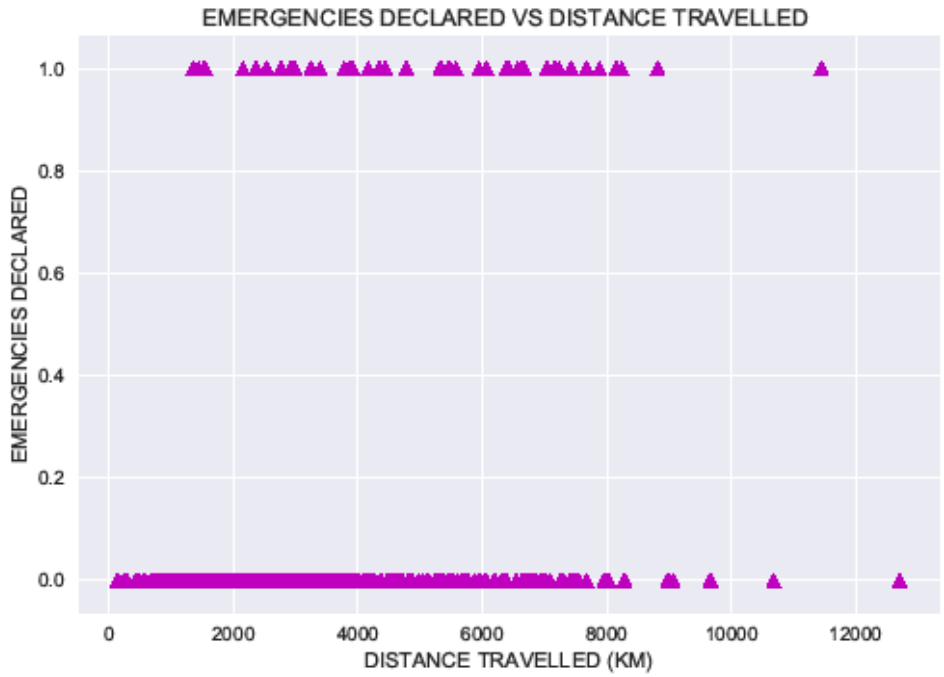
We will try to merge the hurricane and emergency data- frames, based on the 'NAME' and 'YEAR' columns. The resultant data-frames become too small, however. Merging with Pacific hurricanes data-frame gives only 3 rows, while merging with Atlantic hurricanes one leads to a combined

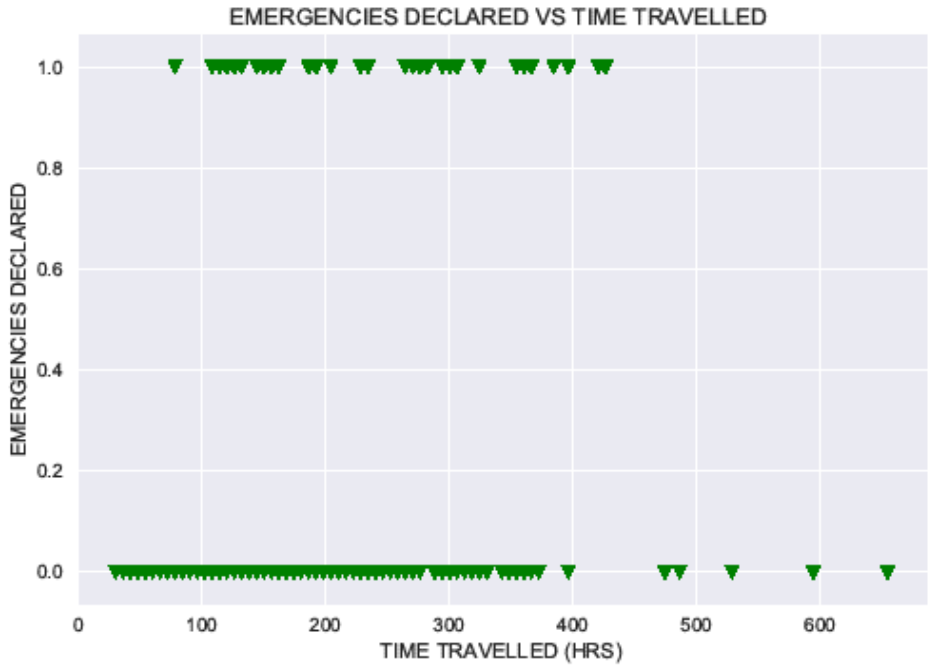
data-frame of only 47 rows. Thus, this merged data-frame will be hardly useful for making predictions.

Thus, we will try to do the following. We create from df_merge a new data-frame that contains only 'YEAR', 'NAME' and one more column 'LED TO EMERGENCIES' with the value 1. Then we merge this data-frame with df_Atl_new, and if no emergencies occurred for a given hurricane the value in the 'LED TO EMERGENCIES' column will be just 0. We will further limit ourselves with years greater than 1964, in which case our data-frame will contain 402 observations with 47 rows having 1 in 'LED TO EMERGENCIES' columns.

1) Let's try to do some data storytelling on this new data-frame, and first ask a simple question: what is the percentage of hurricanes of a given maximum strength that led to emergency declarations. The result appears to be very interesting. Only 1.1 percent of Atlantic hurricanes that reached the maximum strength 1 (tropical/subtropical storm) led to emergency declaration, while 83 percent of maximum strength 7 (category 5 hurricane) resulted in that outcome from 1965 to 2004 years.
2) One can then consider how the value of 'LED TO EMERGENCIES' column depends on the average strength of a hurricane, its distance and time travelled? We see that the average strengths, distances and times are rather broadly distributed for both, hurricanes that led to emergencies and those that did not.

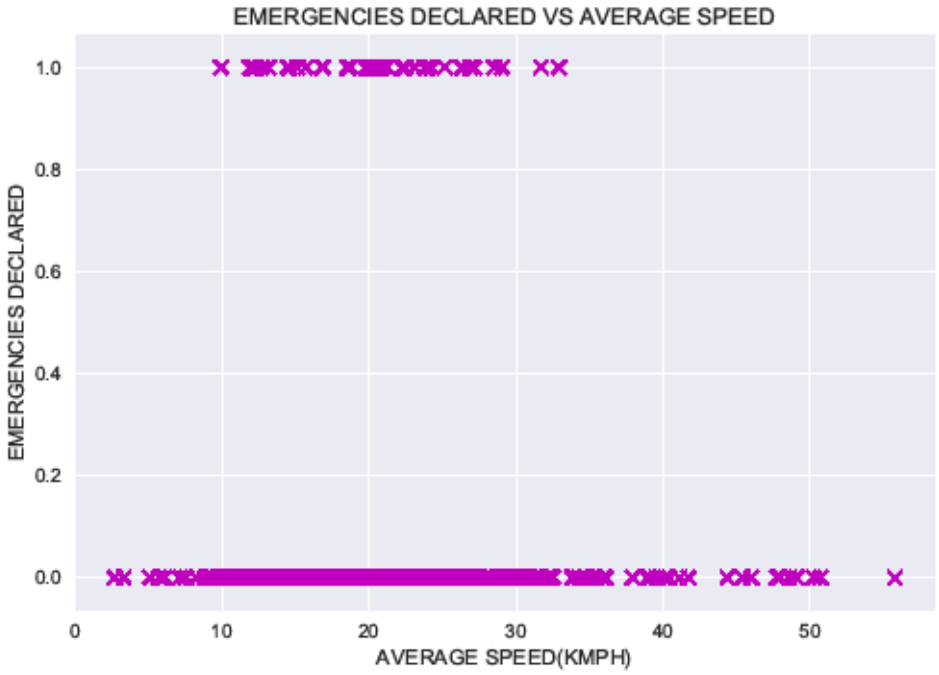


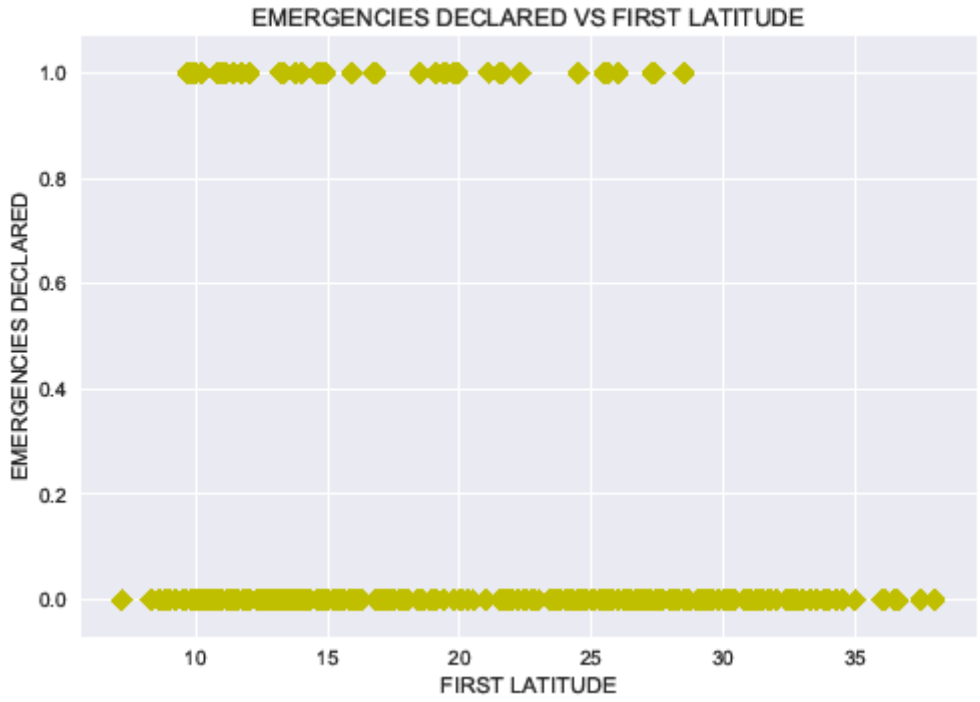


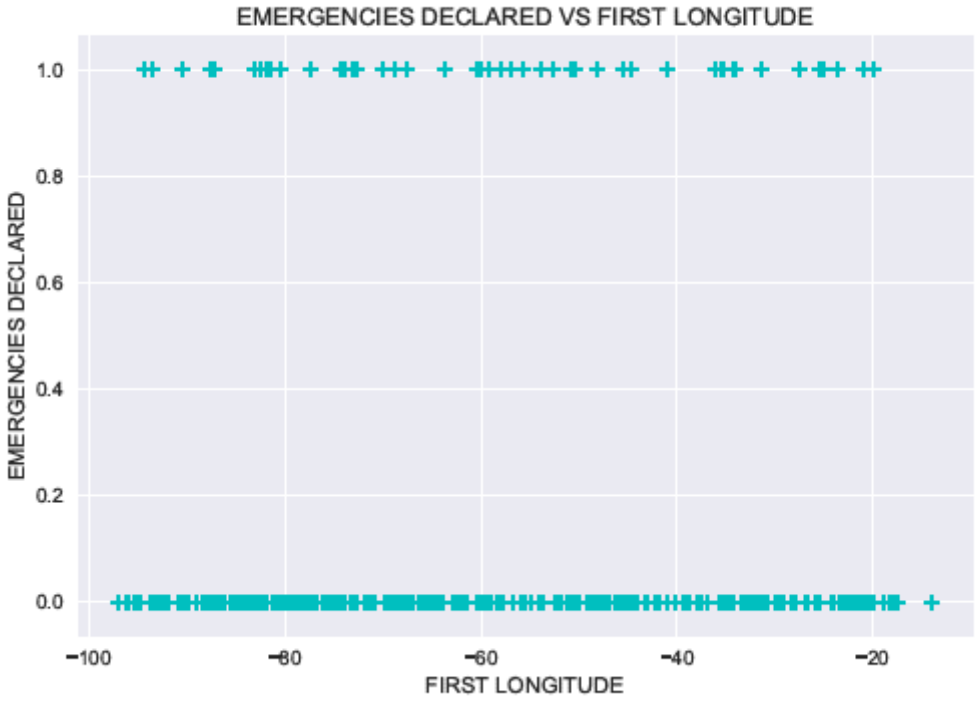


3) If we consider how the value of 'LED TO EMERGENCIES' column depends on the average speed of a hurricane, we will see that average speeds of the hurricanes that led to emergencies are reltively narrowly distributed. This is not true for the dependences of the value of 'LED TO EMERGENCIES' on the first latitudes and longitudes. The hurricanes that led to emergencies originated from a broad segment of latitudes. The range of longitudes where hurricanes that caused troubles

originated is almost equal to that for hurricanes that did not lead to any emergencies.







As a result of EDA of the merged data frame, one can conclude that one can try to build a model that will allow us to predict whether a hurricane will lead to an emergency declaration or not.