Hurricanes and Emergencies:
1st Capstone Project Report

1. Background, problems and questions to be answered

It is needless to say that tropical storms and hurricanes belong to the most extreme natural phenomena on Earth. Even if they are located above the ocean far from the continents, they disrupt air and marine traffic, pose threat to living species. Once a hurricane moves inland, the consequences of the landfall may be devastating. Massive storm surges, excessive rainfalls and powerful winds bring flooding, destruction of property and sometimes to loss of life. Very often the damage is so extensive, that the state and local government officials have to declare a state of emergency in order to mobilize technical and human resources and attract additional federal funding to help alleviate the aftermath. The scale of disaster suggests that comprehensive investigation of everything related to storms and hurricanes is of paramount importance to develop measures of protection and minimize damage.

The techniques predicting the trajectories of motion of storms and hurricanes, the territories they are expected to affect and for how long, are well developed now. The predictions give enough time to the population of coastal areas to prepare to embrace disastrous storms, albeit even now it is often not clear how these storms will behave once they move deeper inland. Such an incomplete predictability may make it difficult to estimate in advance the total monetary loss, number of people affected, as well as the number of counties subject to emergency declarations. In light of this, it may be useful to undertake a statistical analysis of historical characteristics of storms and hurricanes and try to analyze how the above-mentioned quantities depend on various characteristics. For example, data permitting, one can build predictive models and investigate how the total monetary damage depends on the initial place of birth of a hurricane, its maximal strength, average strength, as well as the total time and distance travelled.

In general, in order to answer as many questions as possible, one may need to analyze large number of datasets related to both the hurricane characteristics and the data specifying the consequences. In this project, however, we will try to answer just one major question: which characteristics of hurricanes are most typical for dangerous storms that, once hitting land, are likely to lead to emergency declarations? Other questions we would like to explore are:

a) What do the Atlantic and Pacific hurricanes have in common and what are their differences?
b) How does the maximum strength of both Atlantic and Pacific hurricanes depend on distance travelled, time travelled, first latitude and longitude as well as the strength averaged over the time of existence?
c) How many storms and hurricanes resulted in emergency declarations? What is specific to these hurricanes?
d) Is the distance travelled by hurricanes considerably correlated with the time travelled? How random the velocity of motion is?
e) Is there any dependence between the maximum strength of a hurricane and its place of origin (latitude and longitude)?

2. Potential Clients

There are two groups of clients that could be interested in the findings from this project. The first type of clients is the insurance companies that estimate the costs associated with hurricanes and tropical storms. The second group of clients consists of the government agencies that are in charge of declaring emergencies and distributing relief funds.

3. Datasets used, data wrangling and exploration

3.1 Datasets used

We will import three datasets. The first two data files are presented by the Department of the Interior of US Geological Survey. They describe the tracks of historic hurricanes originated in Pacific and Atlantic oceans. The datasets can be 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

The data are contained in .dbf files and we use the DBF5 reader to obtain the Pacific and Atlantic hurricane tracks data-frames.

The file containing information about the emergency declarations is 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. The file can be downloaded from:

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

3.2 Data wrangling for the ‘hurricanes’ datasets

a) Checking the information about the hurricane tracks datasets indicates that these files contain no missing values. Since the hurricanes started receiving names only after 1949, and only after that year the data are more or less accurate, we will take into account the 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. Namely, we will consider only the depressions, storms (both subtropical and tropical) and hurricanes of all five categories. The data-frames about characteristics of hurricanes contain many columns, but we will keep only those describing year, month, day, name, longitude of the center, latitude of the center, sustained wind and category of each hurricane or storm.

b) 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'.

c) 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. The effective strength and wind strength are averaged over the time when a storm is considered dangerous. 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 group-by 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 df_Atl_new contains 527 rows, while the data frame for Pacific hurricanes 679 rows.

d) 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. And finally, we create the combined list of the names of the hurricanes (Atlantic and Pacific) to be used later.

3.3 Data wrangling for the ‘emergencies’ dataset

a) We first select the rows where the column 'Disaster Type' is equal to 'Hurricane' and 'Typhoon'. The data-frame that previously had 46185 rows now has only 8883 rows. Then 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' treating each such entry as a distinct county. We then recast the 'Declaration Date' column in the standard datetime format and create three separate columns, corresponding to year, month and day. There are no missing or incorrectly entered data in these columns.

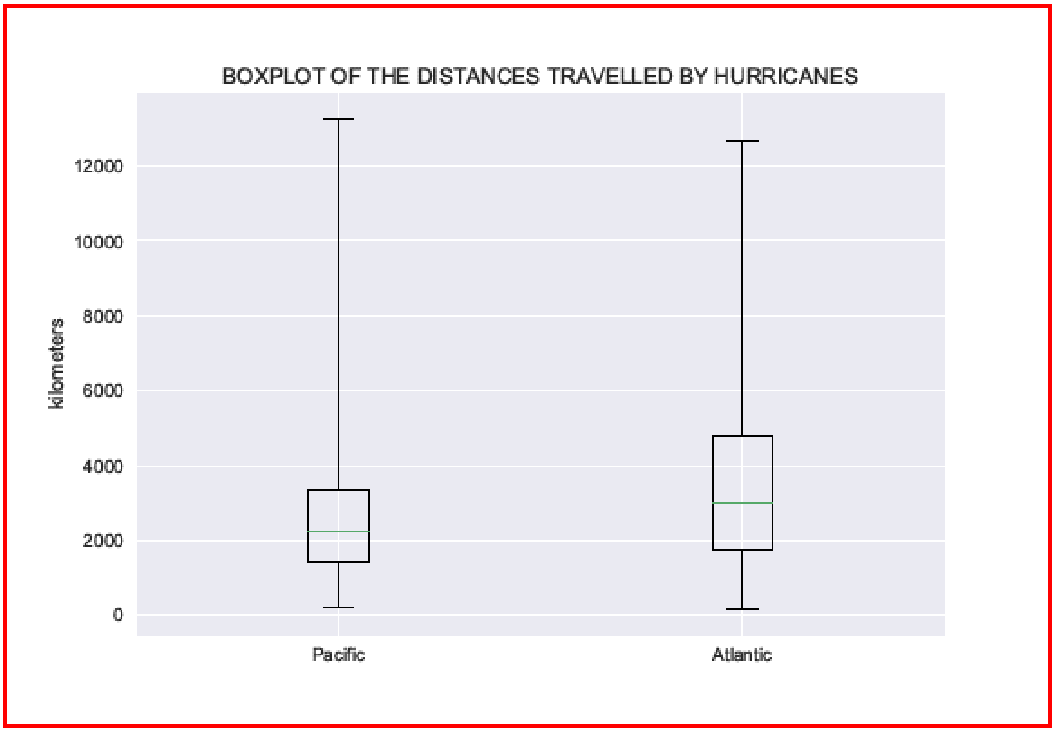
b) The majority of values in the column 'Disaster Title' contain the name of a hurricane. Thus, our next task will be to single out the name of a hurricane in each entry of the column. To do this, we capitalize the entry and remove everything that does not contain a name from the combined list of names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

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

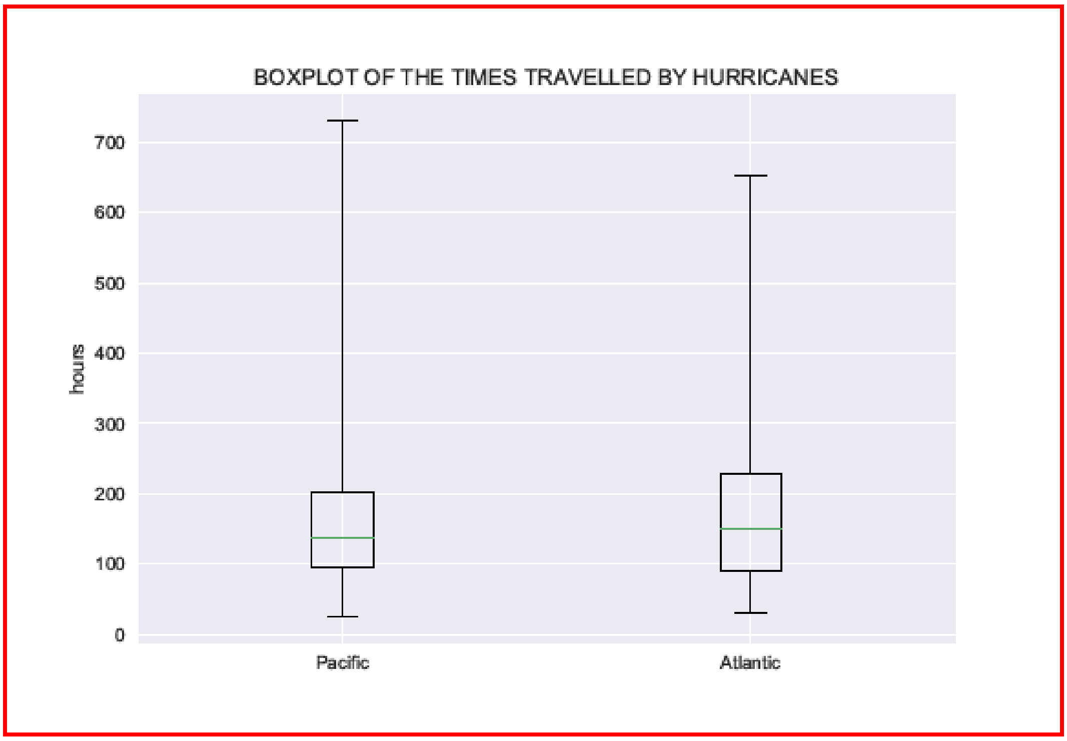
3.4 Atlantic and Pacific hurricanes: comparison of some characteristics

Let's compare some characteristics of Pacific and Atlantic hurricanes.

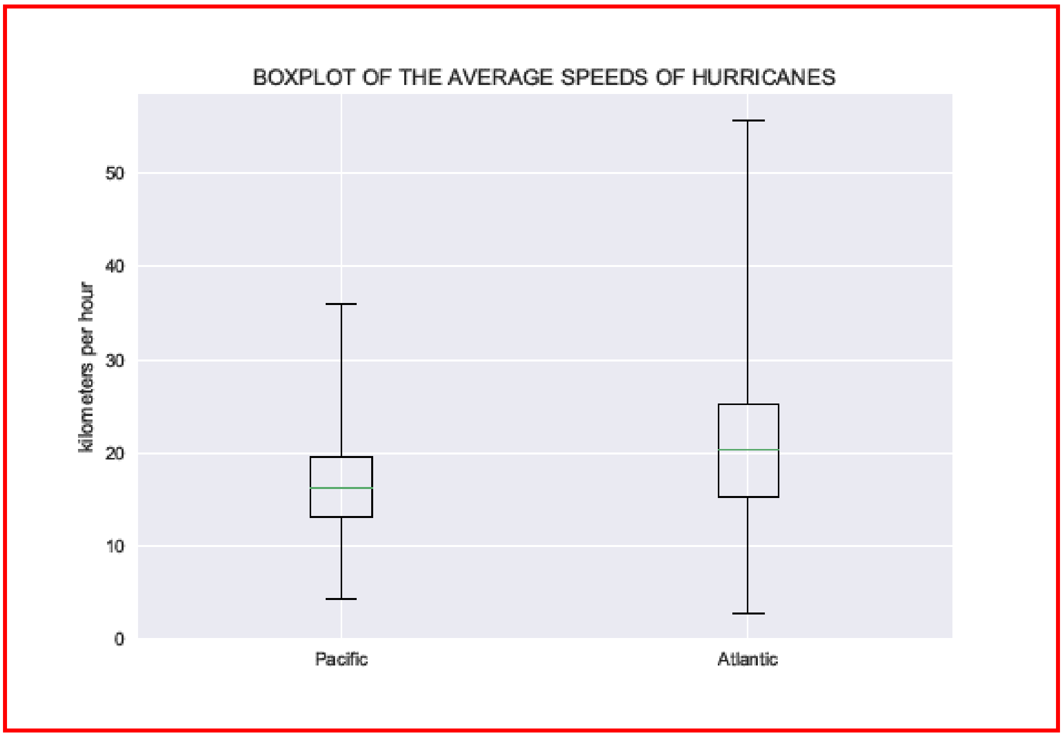
a) First, we look at the column that gives the distance travelled by each hurricane while it is considered dangerous. The descriptions as well as boxplot reveal that the travel distances vary dramatically. The range of variation is especially pronounced for Pacific hurricanes. However, the distribution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



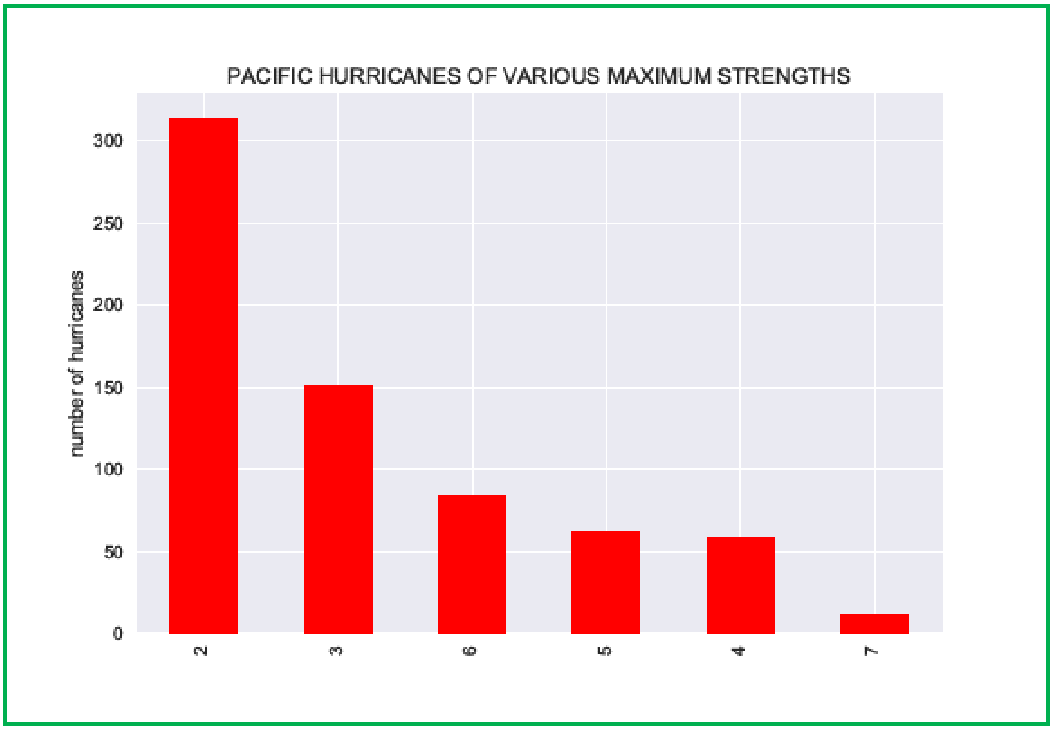
b) Next, we look at the column that gives the total time travelled by each hurricane while it is considered dangerous. The descriptions as well as boxplot reveal that the travel times vary a lot as well, with the variation range being more pronounced for Pacific hurricanes than for Atlantic ones. We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

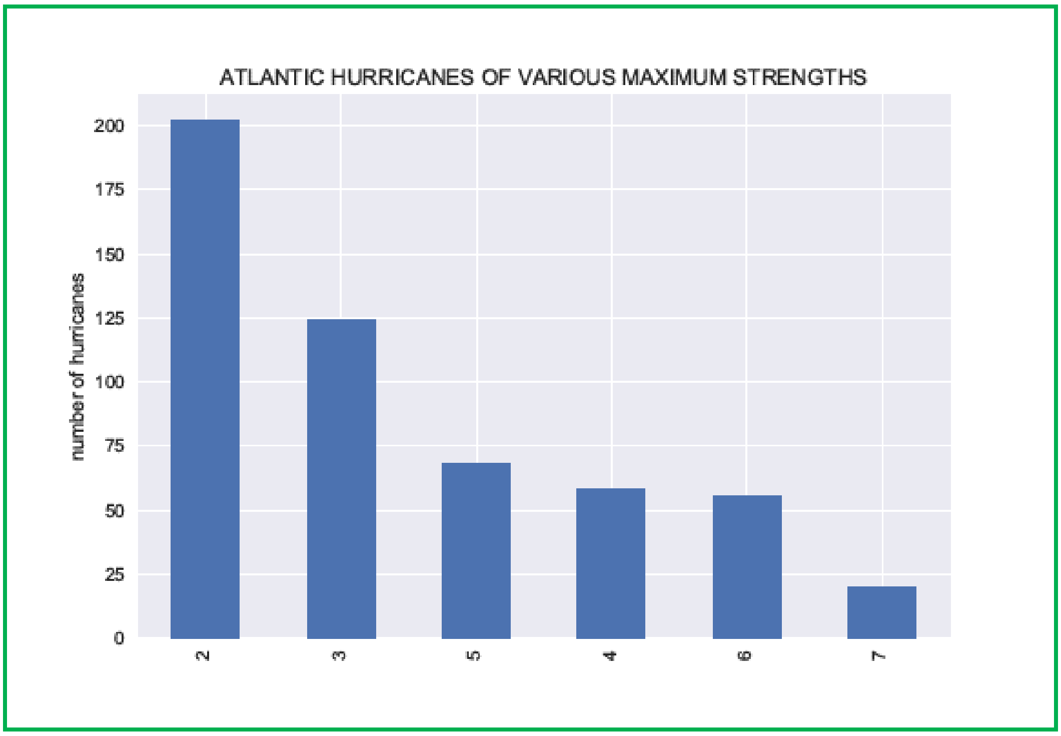


c) We then make a box plot of the average speeds. We see that the spread of values of speeds for Atlantic hurricanes is much broader than for Pacific ones.

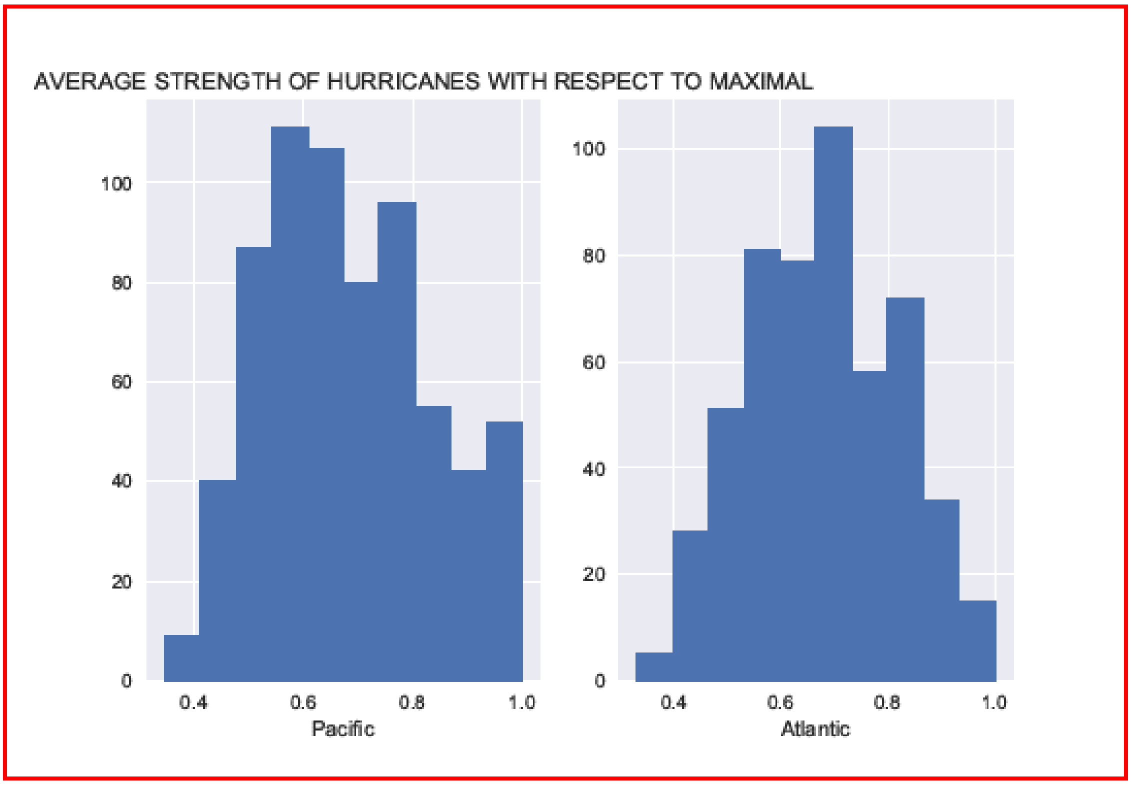


d) It is then a good idea to look at the distribution of the total number of hurricanes and storms with respect to the value of maximum strength. For Pacific hurricanes and storms, the number of lows that never reached the level of a hurricane is the largest compared to the number of those that reached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

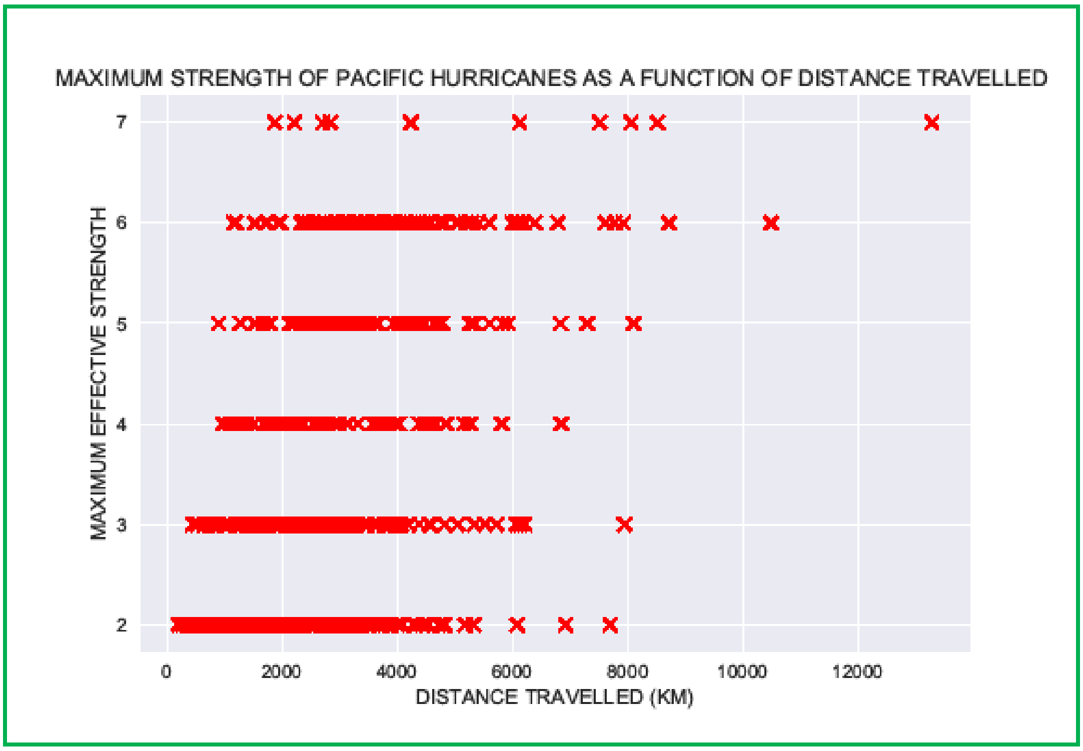


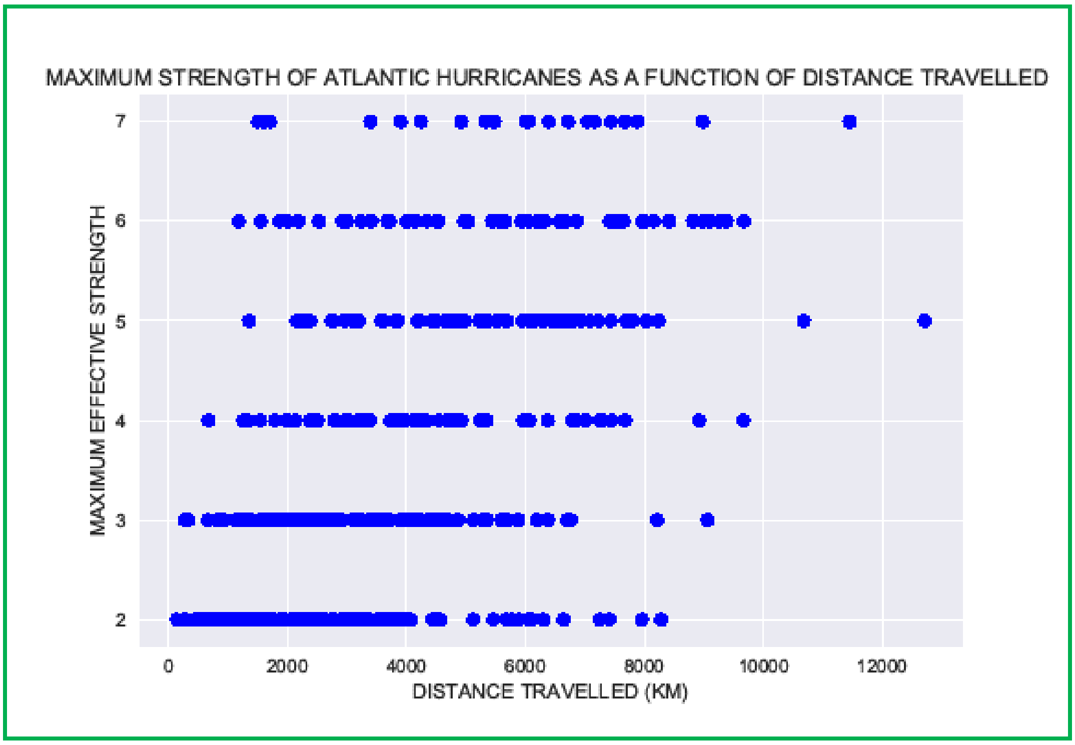


e) Let's make the histograms of the average strengths of hurricanes with respect to the maximum strength. We see that the average strengths of Pacific hurricanes have broader distribution than those of Atlantic ones. While the average strength of Atlantic hurricanes is peaked at around 0.7, the same parameter for Pacific ones is maximal at values of around 0.6.

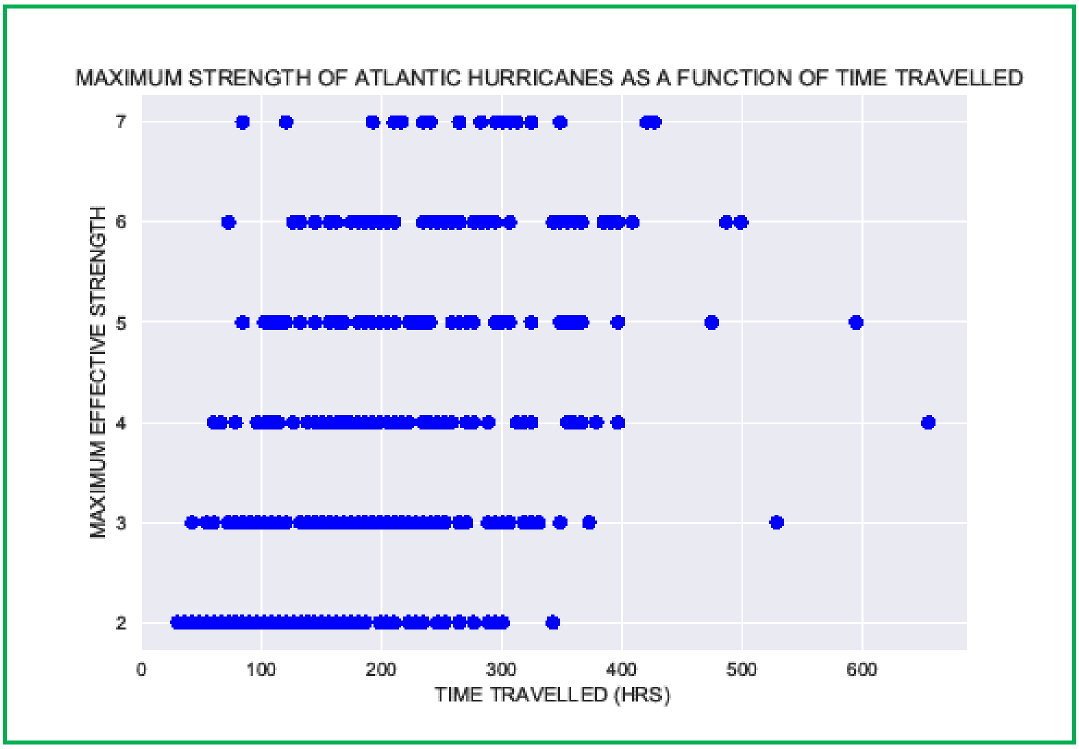


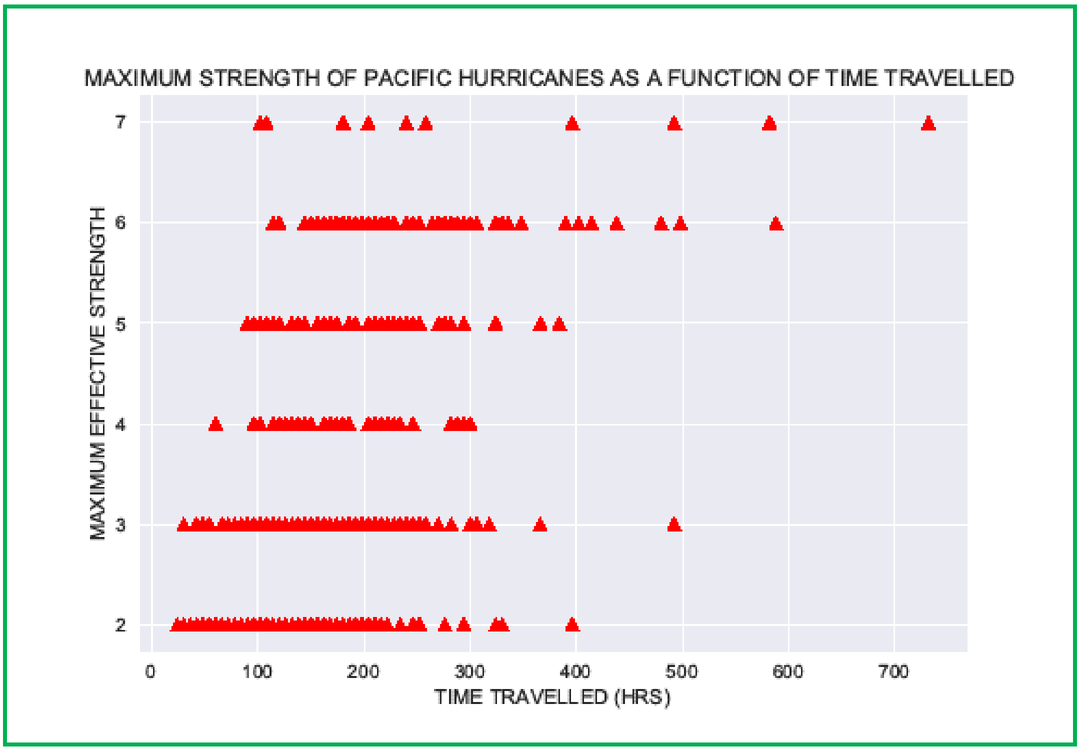
e) At the next step, we look at the dependence of maximum effective strength of hurricanes on the distance travelled while being dangerous. We see that the average distance travelled, as well as the spread of distances, has the tendency to increase with increasing maximum strength. This is almost true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



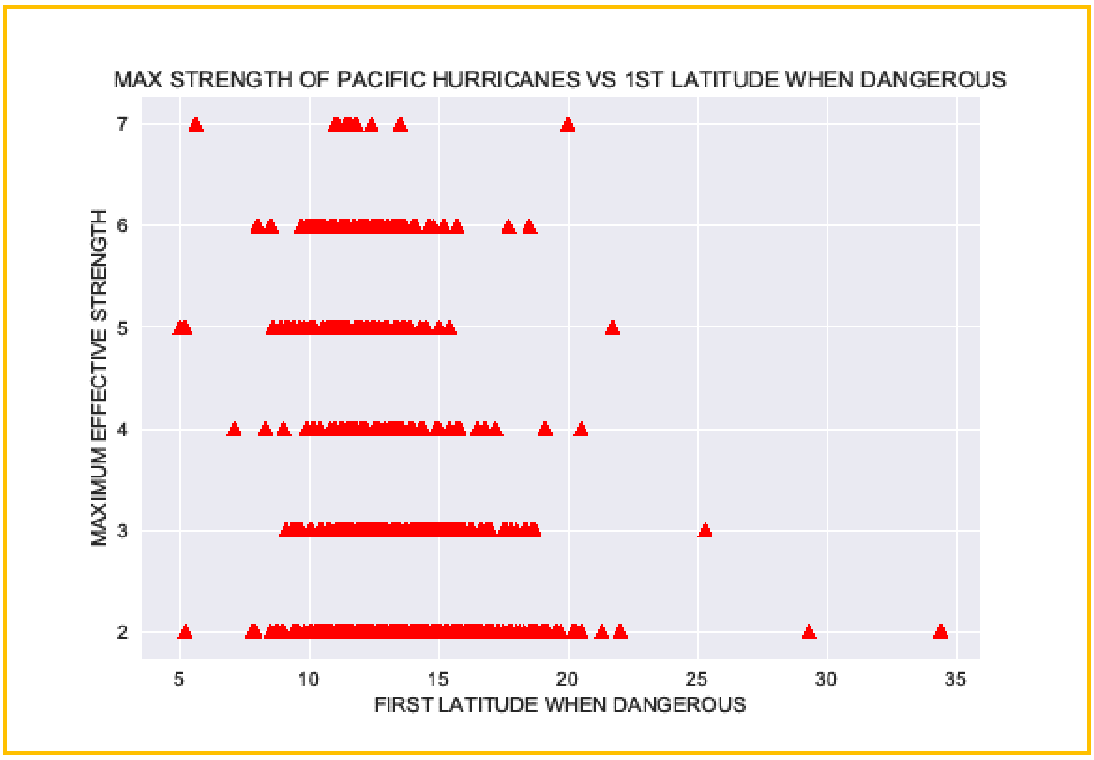


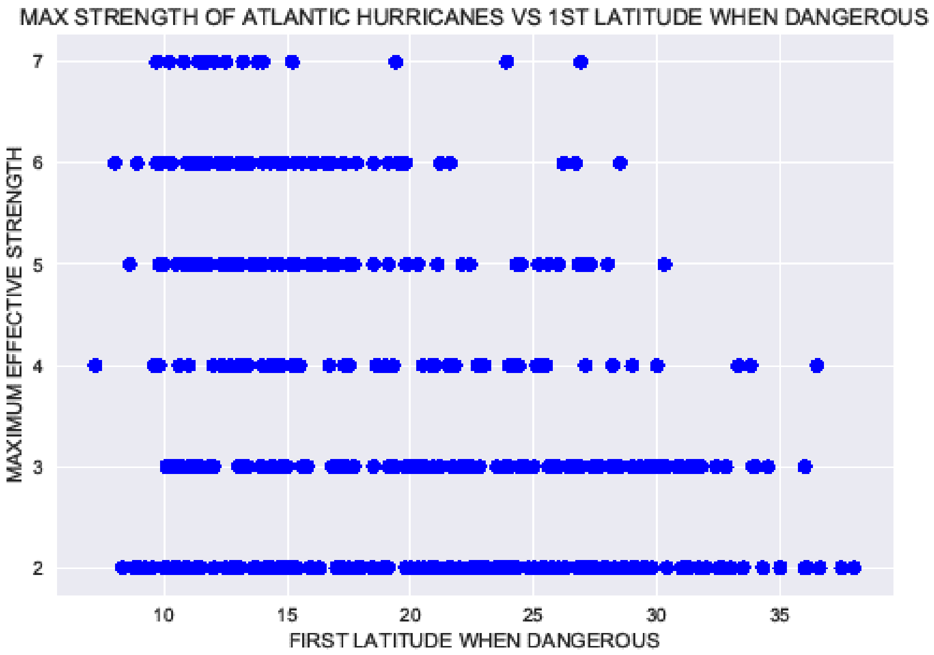
f) We then look at the dependence of the maximum effective strength of hurricanes on the time travelled while being dangerous. The dependence here is similar to that for the maximum strength vs distance travelled. The minimal and average times travelled by hurricanes have the tendency to increase with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

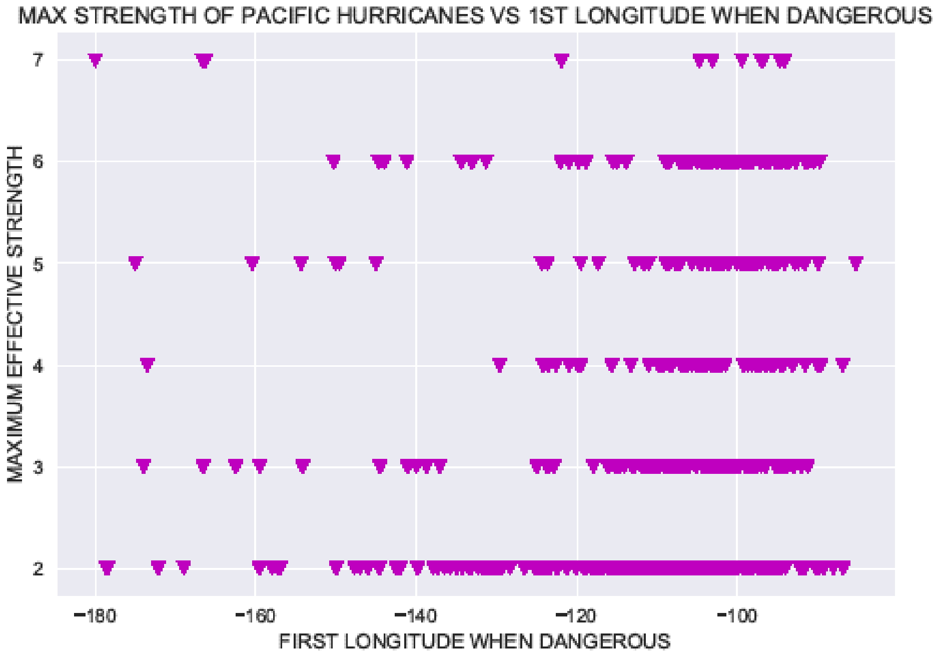


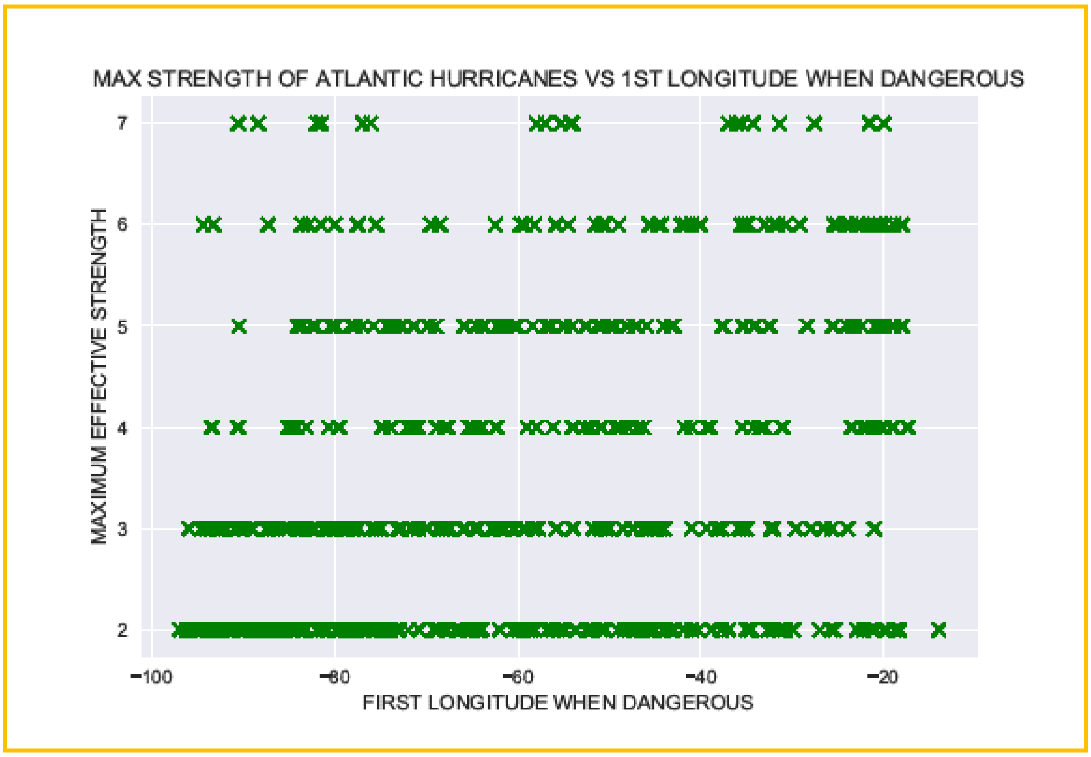


g) It behooves also to look at the dependence of the maximum strength of hurricanes on the latitude and longitude when they first became dangerous. The tendency that we observe for both Pacific and Atlantic hurricanes is very interesting. The larger the maximal strength of a hurricane is, the smaller the range of latitudes and longitudes where the hurricane is more likely to form.

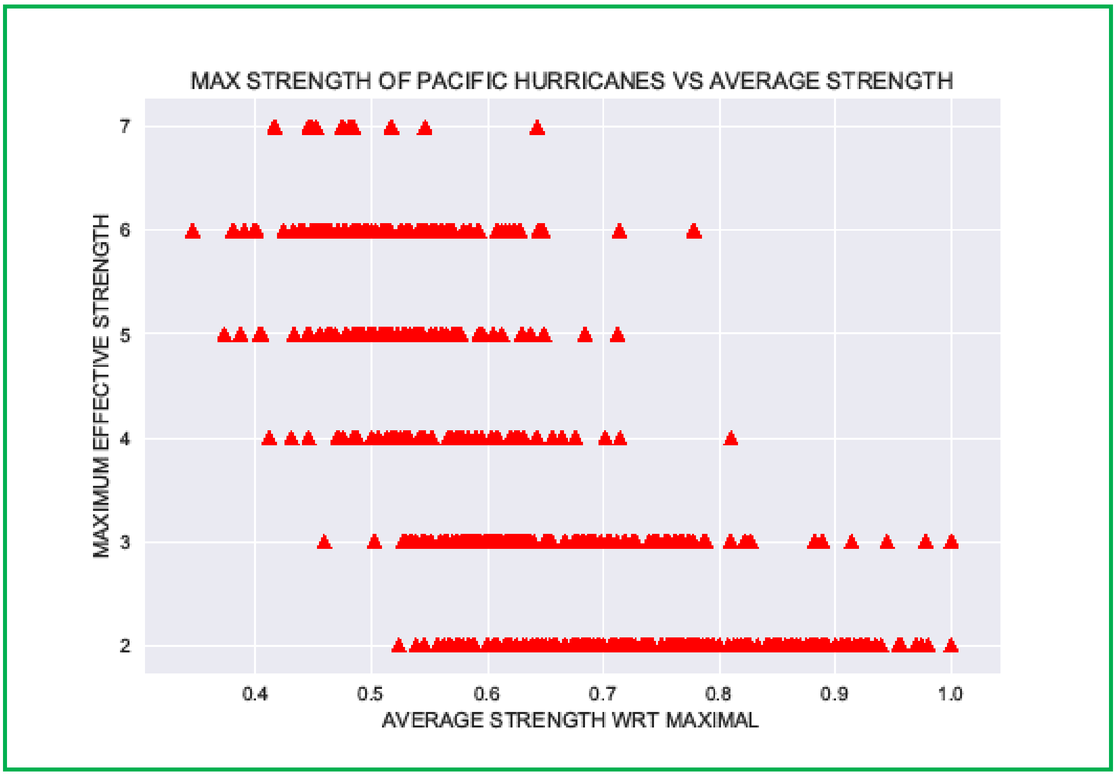


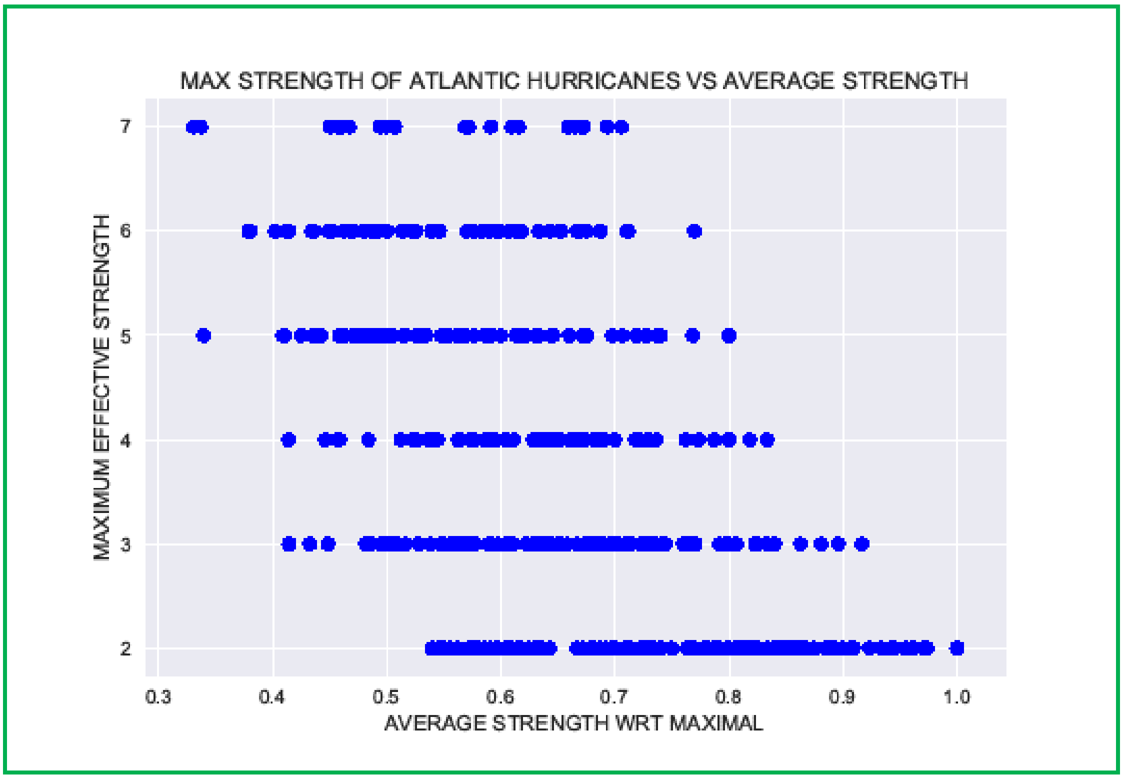

MAX STRENGTH OF ATLANTIC HURRICANES VS 1ST LATITUDE WHEN DANGEROUS
MAXIMUM EFFECTIVE STRENGTH
7
6
5
4
3
2
10
15
20
25
30
35
FIRST LATITUDE WHEN DANGEROUS


MAX STRENGTH OF PACIFIC HURRICANES VS 1ST LONGITUDE WHEN DANGEROUS
MAXIMUM EFFECTIVE STRENGTH
7
6
5
4
3
2
−180
−160
−140
−120
−100
FIRST LONGITUDE WHEN DANGEROUS



h) 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.





3.4 Emergencies and Atlantic hurricanes merged data: Exploratory Data Analysis

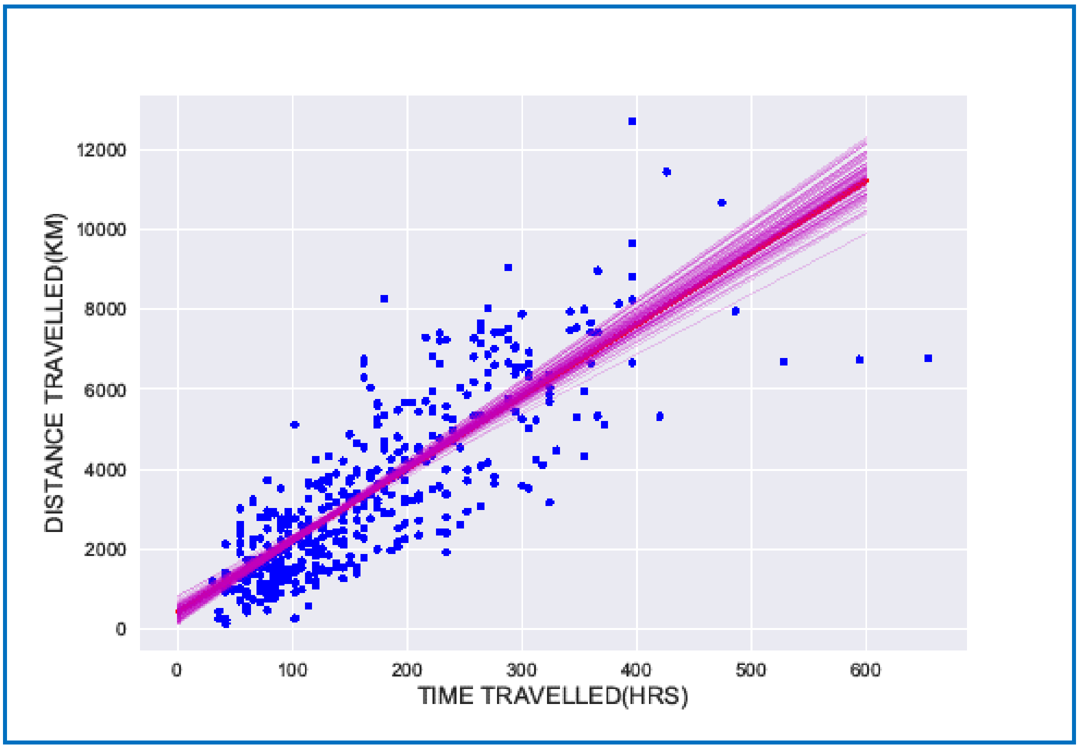
We will try to merge the hurricane and emergency data-frames, based on the 'NAME' and 'YEAR' columns. The resultant data-frames become too small, however. Merging the emergency and Pacific hurricanes data-frame gives only 3 rows, while merging the emergency with Atlantic hurricanes leads to a combined data-frame of only 47 rows. Thus, these merged data-frames will be hardly useful for making predictions.

As a result, we will follow another approach. We create a new data-frame from df_merge, containing 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 to years greater than 1964, in which case our data-frame will contain 402 observations with 47 rows having 1 in 'LED TO EMERGENCIES' columns. This new and major data-frame is called df_merge_new. 'LED TO EMERGENCIES' column contains the binary dependent variable, while variables in other columns will be regarded as independent.

a) The first question we would like to ask is how random the average speed of Atlantic hurricanes is? The column 'AVERAGE SPEED(KMPH)' is obtained by taking the relation of the 'DISTANCE TRAVELLED(KM)' value to the 'TIME TRAVELLED(HRS)' value, so that the values in the

'AVERAGE SPEED(KMPH)' column are determined by the values of the two other columns. But how independent is the time travelled by a hurricane from the distance travelled? The mean value of the average speed of Atlantic hurricanes is 21.132572425720934 km per hours, while the standard deviation is 8.538384619924566 km per hour and is quite large compared to the mean.

The distance and time travelled columns are strongly correlated; the Pearson coefficient is 0.813291065794. The coefficient of determination which is equal to the square of Pearson coefficient is 0.6614423577. We then make the scatter plot of the distance travelled vs the time travelled and make the linear fit. We see that the slope of the fit, that is in fact the fitted velocity of the hurricane motion, is 17.9969045062 km per hour and is slightly less than the mean value of the average speed 21.132 km per hour.

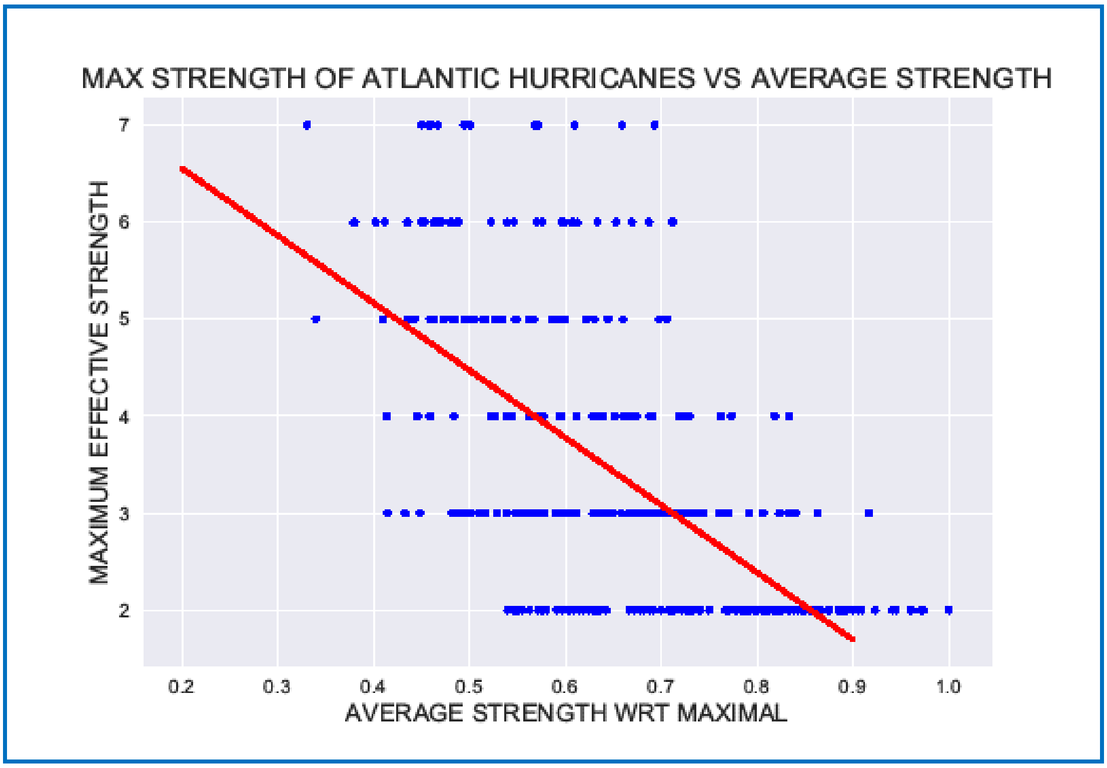


It is instructive also look at the 95% confidence interval for the slope using pairs bootstrap method and visualize the variability of the slope. The confidence interval for the slope (fitted velocity of a hurricane) is [16.01, 20.29] km per hour. We see that the mean value of the average speed being equal to 21.13 km per hour is slightly higher than the upper bound of the confidence interval.

b) What is the percentage of hurricanes of a given maximum strength that led to emergency declarations? The answer to this question 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 hurricanes) resulted in that outcome from 1965 to 2004 years.

c) We can then look at the Pearson coefficients between some columns describing independent variables.

- The maximum effective strength of a hurricane is strongly correlated with the average strength with respect to maximal; the Pearson coefficient is -0.657 and negative. The coefficient of determination is 0.432670262484. This means that the larger the maximum strength achieved by the hurricane, the smaller amount of time it spent being that strong. This is in complete agreement with the visualization data presented in the data storytelling section. It is useful to visualize it once again here and to attempt a linear fit.



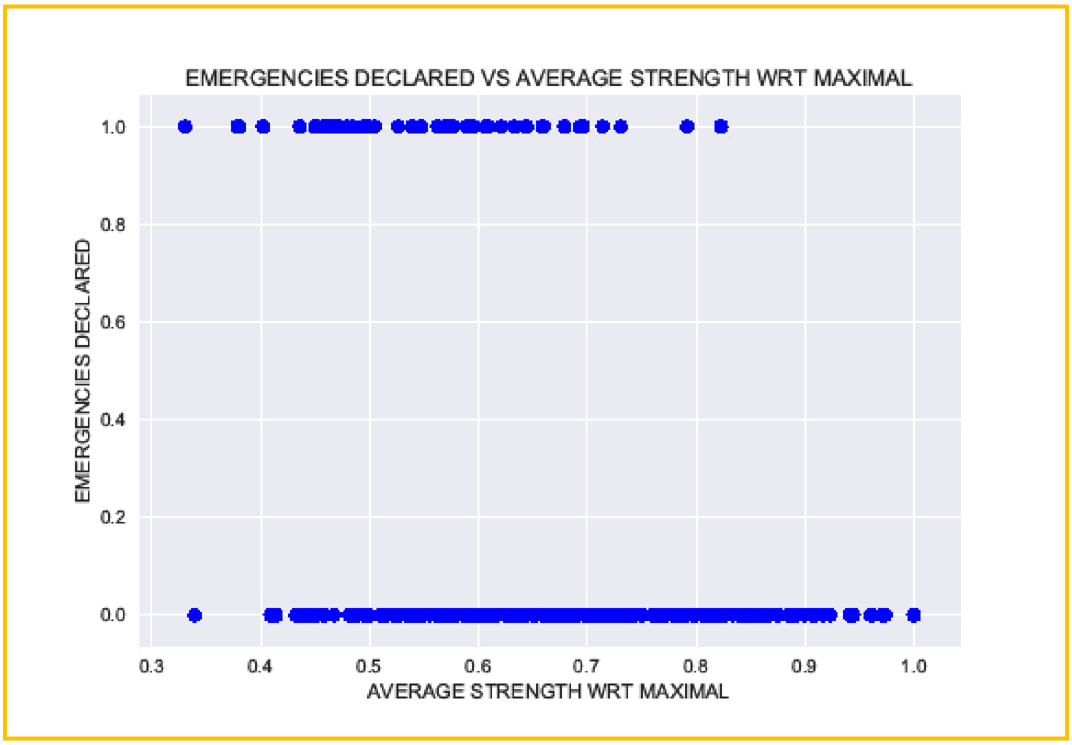
- It is completely natural that the maximum effective strength is strongly positively correlated with the average wind strength inside a hurricane (Pearson coefficient is 0.858). This is because the strength of a hurricane at any moment of time is primarily based on the speed of the wind.

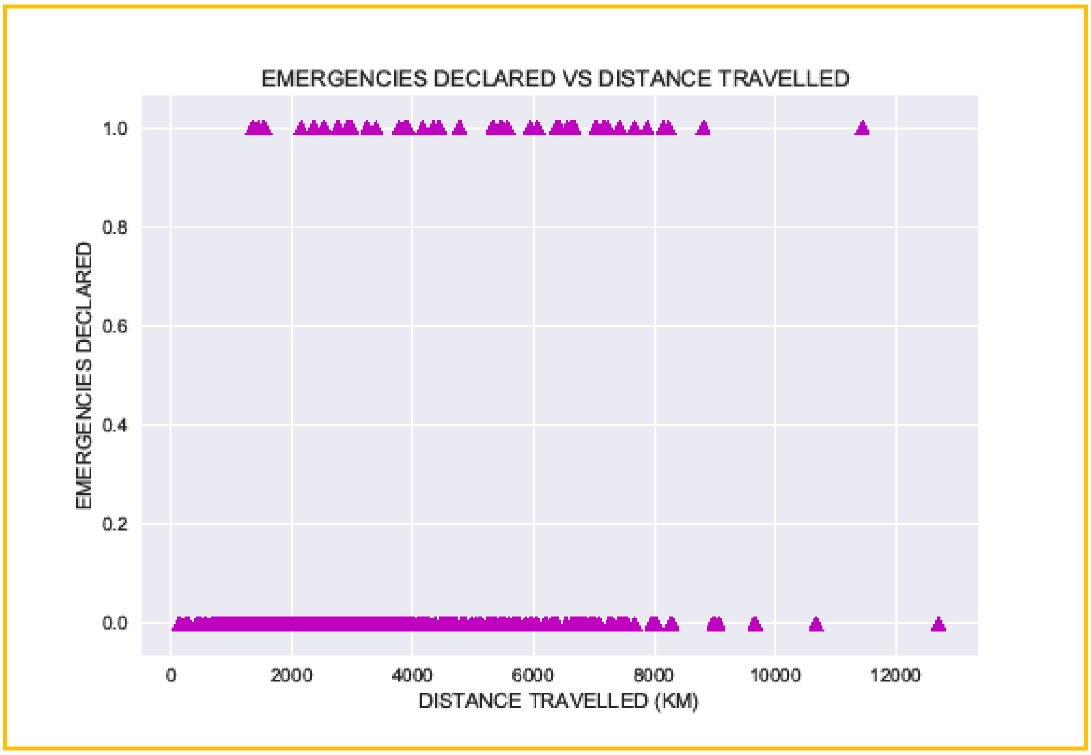
- It may seem first a bit surprising that the first latitude is rather strongly correlated with the first longitude when a hurricane first becomes dangerous (Pearson coefficient is -0.485). This correlation may be explained by the presence of land (islands and continents) that affects the place of formation of hurricanes.

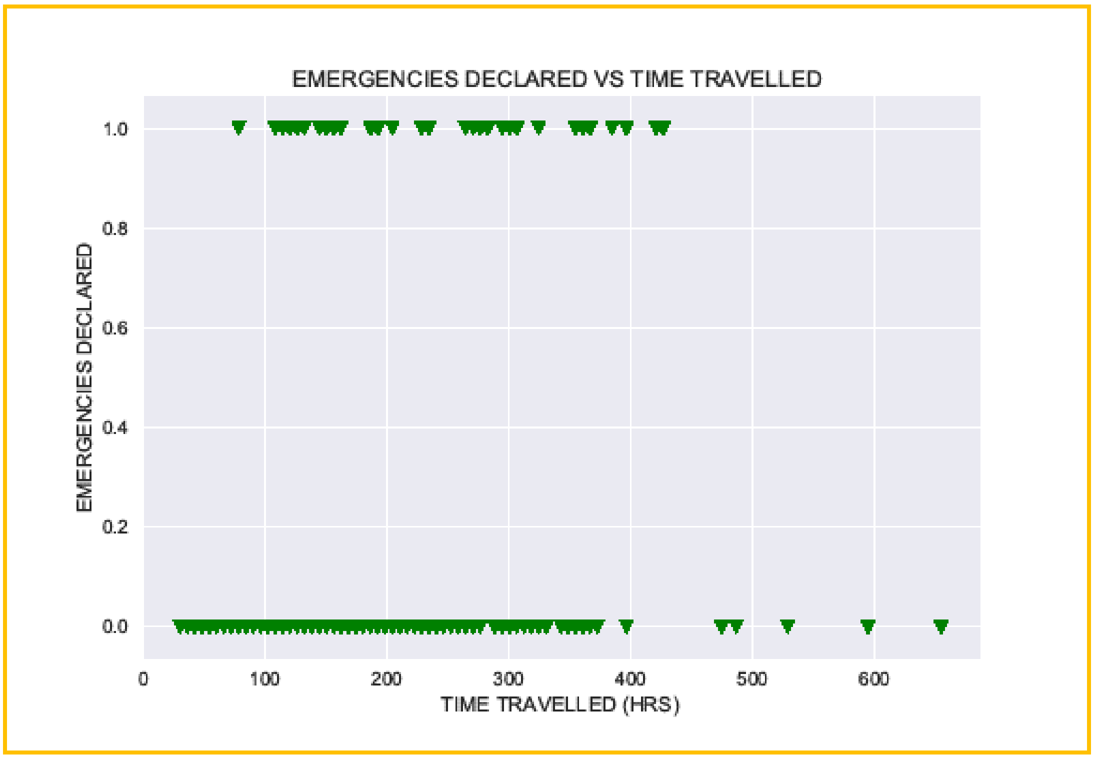
- Finally, it's worth looking at the correlation between the maximum effective strength and first latitude and longitude when a hurricane first became dangerous. The correlation between the maximum strength and longitude is rather weak (Pearson coefficient is 0.268), while the correlation involving latitude is stronger (Pearson coefficient is -0.324) and negative. The latter correlation can be explained by the fact that the strongest hurricanes form closer to the equator and gain power during their motion above the ocean.

d) As a final part of EDA, we will consider the correlation between the dependent binary variable of the 'LED TO EMERGENCIES' column and some of the independent variables.

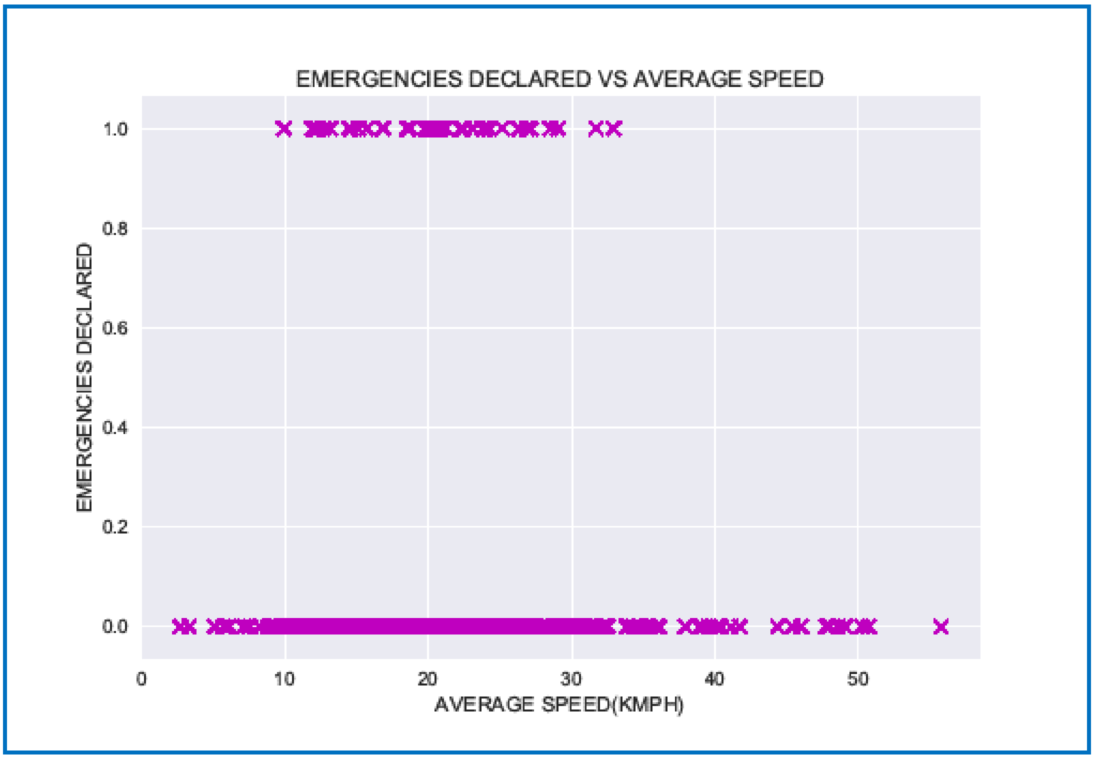
- How does the value of 'LED TO EMERGENCIES' column depends on the average strength of a hurricane, distance travelled and time travelled? Investigation shows that the average strengths, distances and times are rather broadly distributed for both, hurricanes that led to emergencies and those that did not. Our dependent variable is not very strongly correlated with distance (Pearson coefficient is 0.268) and time (Pearson coefficient is 0.280) travelled by hurricanes. We can formulate the null hypothesis that these variables are correlated and calculate the corresponding p-values using permutation method. They are equal to 1 meaning that both null-hypothesis are true.

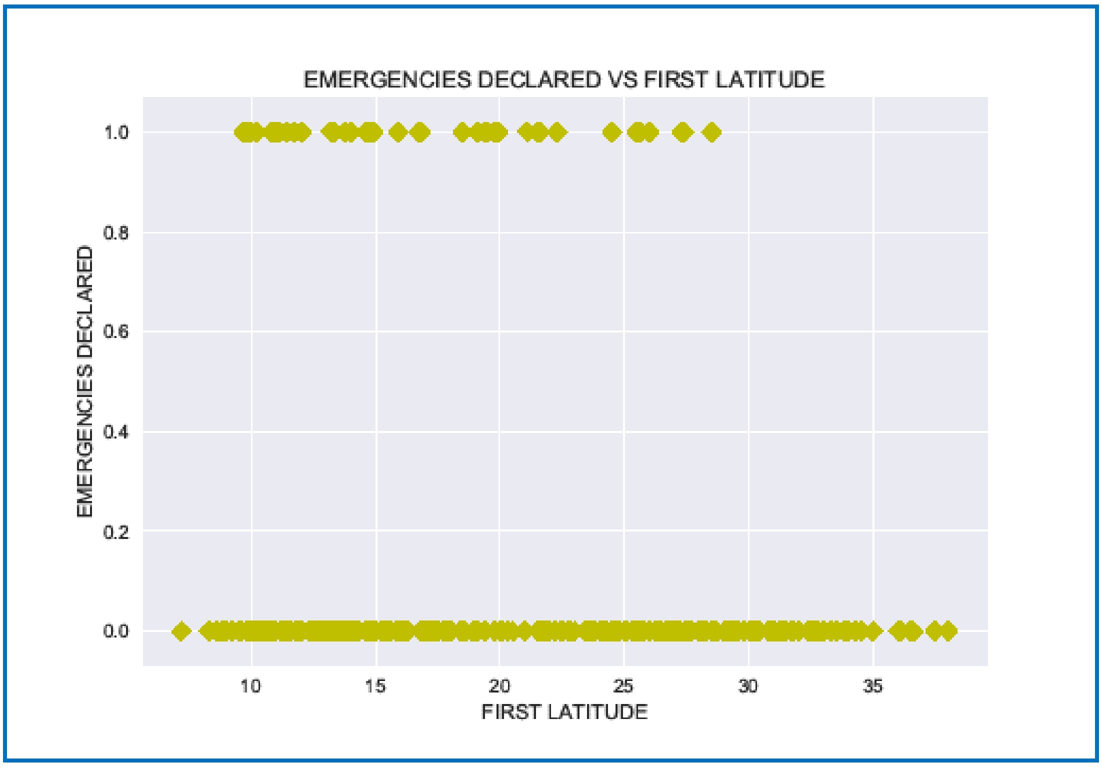


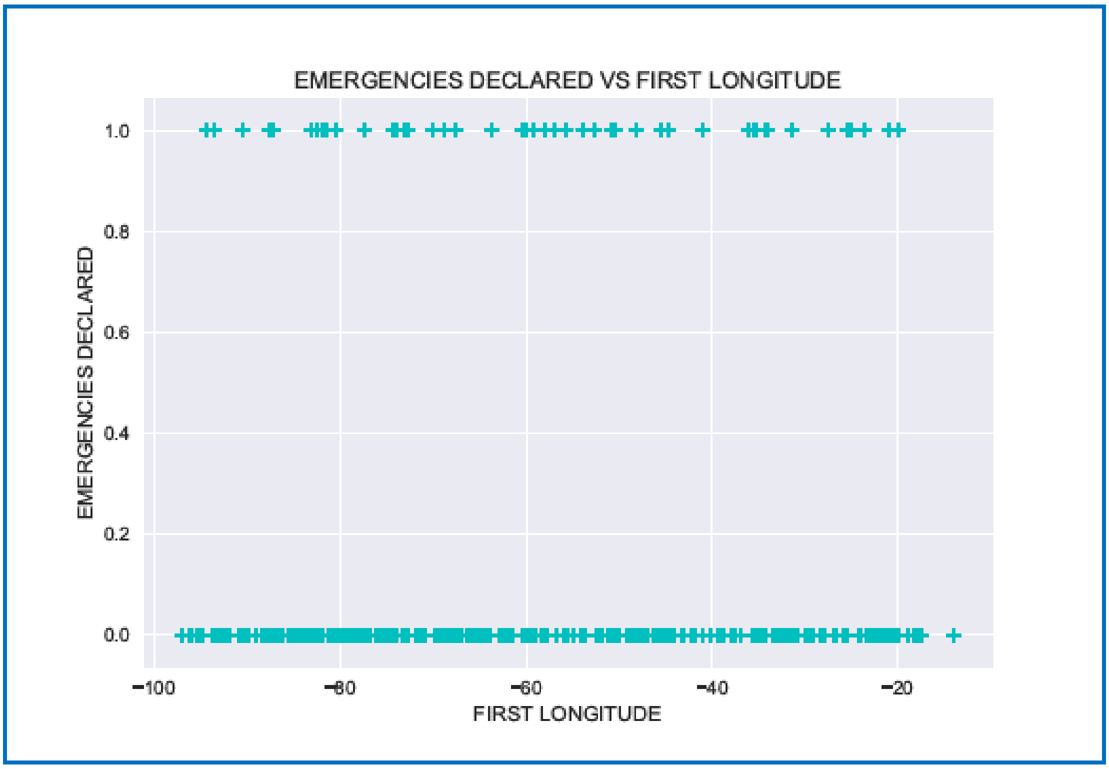




- Considering how the value of 'LED TO EMERGENCIES' column depends on the average speed of a hurricane, we see that average speeds of the hurricanes that led to emergencies are relatively narrowly distributed. This is not true for the dependences of the value of 'LED TO EMERGENCIES' on the first latitudes 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. Looking at the correlation between the dependent ‘emergencies’ variable and first latitude and longitude variables, we obtain that the target variable is weakly but correlated with the first latitude variable (Pearson coefficient is -0.178) and almost uncorrelated with the first longitude (Pearson coefficient is 0.022). Let's test two null-hypothesis: the presence of correlation in the case of latitude and the lack of correlation in the case of longitude. The computations show that in both cases we obtain large p-values, meaning that both null-hypothesis should be kept. We thus come to a remarkable conclusion that the value of the first latitude affects the probability of emergency declarations, but the value of the first longitude does not.







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

4. Predictive Modelling

Let's ask the following question. Will a hurricane lead to an emergency declaration in any state with 50 percent probability given the following features: first longitude, first latitude, distance travelled, time travelled, average speed, average wind strength, maximum effective strength and average strength with respect to maximal? To answer this question, we create the X and y sets each containing 402 observations. X has eight aforementioned features relevant to making predictions. y contains 47 values equal to 1.0 corresponding to the hurricanes that led to emergency declarations, and 355 values equal to 0.0 describing the hurricanes that didn't. We see that the dataset is very unbalanced because only about 10 percent of all hurricanes resulted in emergencies declarations.

If a hurricane resulted in emergency declarations, the relevant observation is regarded positive. Negative observations mean then the absence of emergencies. Under these assumptions, false positives (type I errors) correspond to the wrongful prediction of emergencies for the hurricanes that in reality did not lead to any emergencies. In its turn, false negatives (type II errors) incorrectly predict the absence of emergencies for the hurricanes that in fact resulted in them. I think that in this problem the large number of false negatives is more detrimental than the excess of false positives: missed prediction of an emergency for a dangerous hurricane may lead to deaths, while overprediction of an emergency potentially leads just to monetary losses. While it is clearly desirable to minimize the errors of both types, we will mainly be interested in those classifiers that lead to the smallest possible numbers of false negatives.

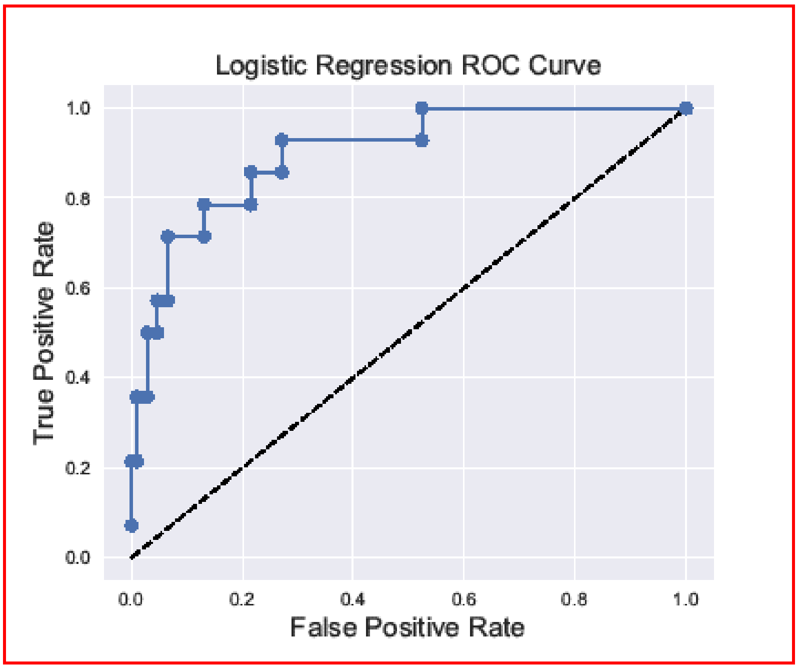
We will use five classifiers: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier and Gaussian Naive Bayes. It is worth mentioning now that all five methods give approximately the same accuracy when tested on the test set, but very different relative weights of false positives (fp) and false negatives (fn); we clearly observe the method dependent fp-fn tradeoff. In this context, the quantities of particular interest to us will be precision defined as tp/(tp+fp), and recall given by tp/(tp+fn) (tp and tn denote the numbers of true positives and negatives respectively).

To proceed, we first standardize features by moving the mean and scaling to unit variance. Then, for all methods we use 70 percent of data for training and 30 percent for testing. Also for all methods, we use the 5-fold cross-validation and the same value for the random_state parameter used for random train-test splitting. To achieve the same distribution of labeled data in the train and test sets, we use the stratified splitting. Each method uses its own reasonable set of parameters to be tuned during the grid search cross-validation. Whenever possible, we plot the table that reflects the relative importance of features determining the level of accuracy, as well as the receiver operating characteristic (ROC) curve. For each method we print the best values of parameters obtained as a result of cross-validation, training scores, test scores, confusion matrix and classification report.

4.1 Classification Models: Training and Testing

a) Logistic Regression

Logistic Regression returns the coefficients for each feature, which might be positive or negative, giving an indication of whether a particular feature is weakening or strengthening the probability that a hurricane may lead to emergencies. The treatment shows that the most important feature in determining this probability is the 'MAXIMUM EFFECTIVE STRENGTH'. It is followed by the 'DISTANCE TRAVELLED' and the 'AVERAGE WIND STRENGTH WRT MAXIMAL', which is not surprising given rather substantial degree of correlation among all these three features. The overall accuracy on the test set is about 90 percent, and we note further that running Logistic Regression gives the recall and precision equal to 0.71 and 0.45 respectively. The recall is well higher than precision, and this is what we are looking for following our approach. To increase the accuracy, we used the 'roc_auc' scoring function to choose parameters during cross-validation. A small difference in the test and training scores is also a good sign.

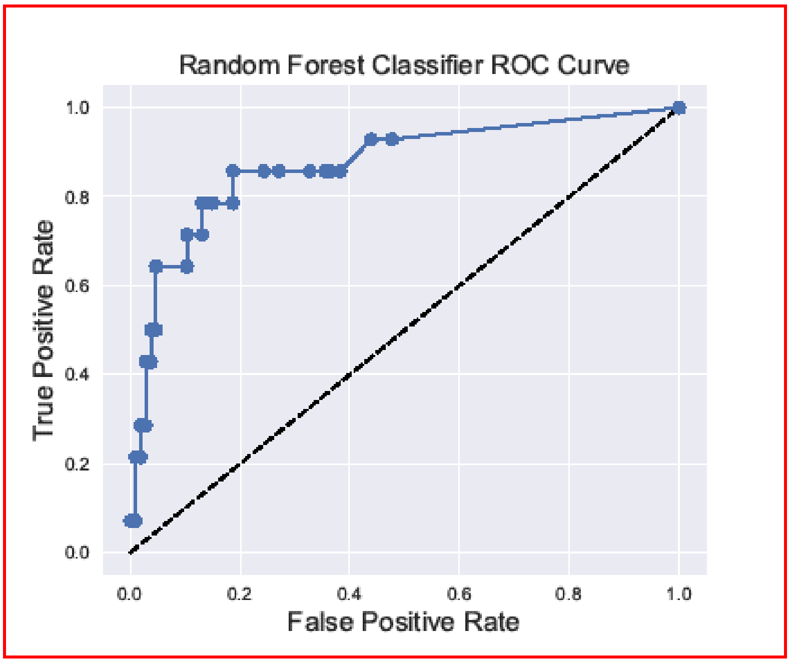


b) Decision Tree Classifier

Decision Tree Classifier gives the test score accuracy of 88 percent and precision of 0.5. The recall is 0.47 which is lower compared to Logistic Regression. The most important feature seems to be the 'AVERAGE WIND STRENGTH', followed by 'AVERAGE STRENGTH WRT MAXIMAL'. We the see that overall Decision Tree Classifier performs worse than Logistic Regression.

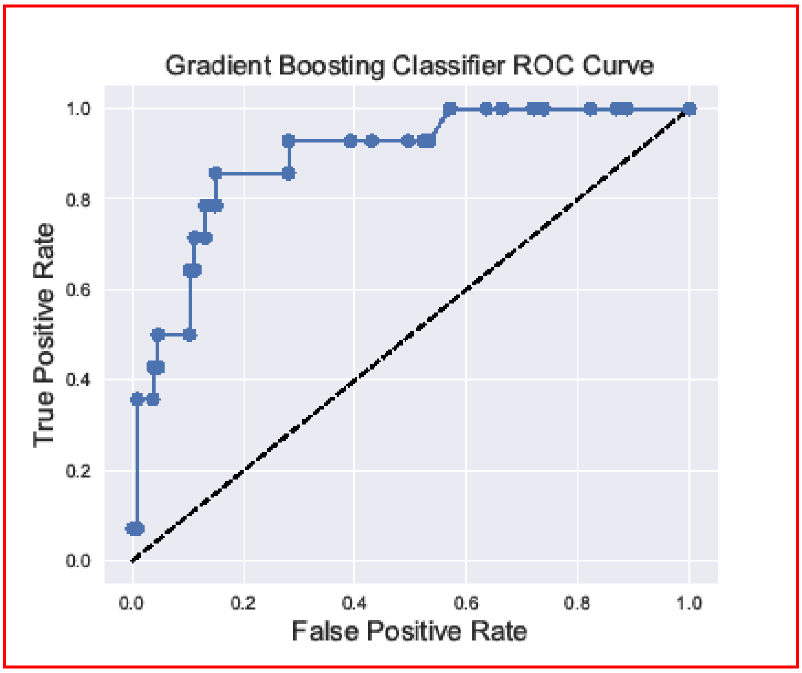
c) Random Forest Classifier

Random Forest Classifier creates random trees and evaluates them, so we expect it to perform better than Decision Tree Classifier. Indeed, the test score is around 87 percent, while precision and recall are 0.50 and 0.64 respectively. We again used the 'roc_auc' scoring function to choose parameters during cross-validation. The most important features are the same as for Decision Tree Classifier.



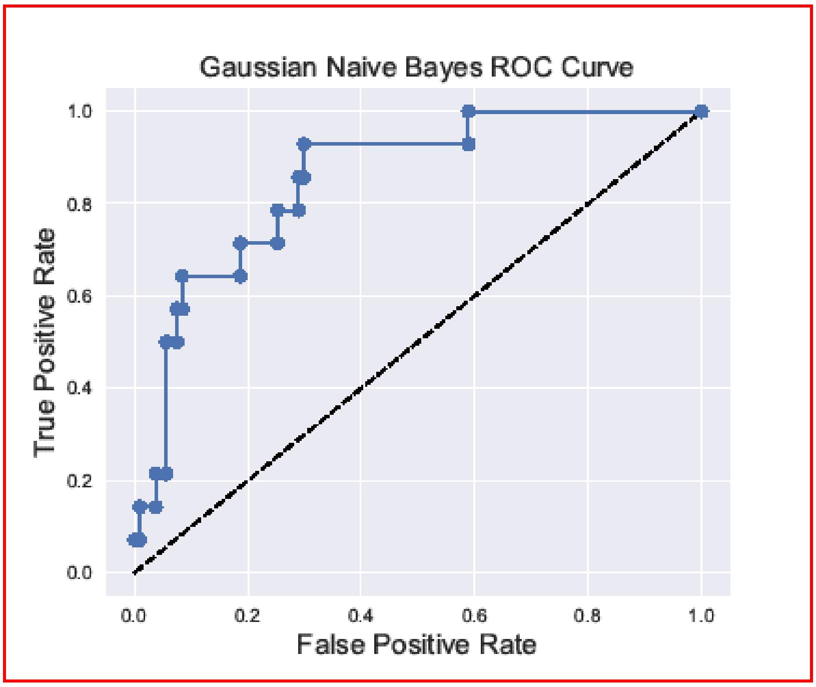
d) Gradient Boosting Classifier

Gradient Boosting Classifier is an ensemble method that builds a strong learner out of many weak learners, typically decision trees. It leads to a high overall accuracy (about 89 percent) but much lower recall compared to precision. Though the 'AVERAGE WIND STRENGTH' seems to be the most important feature, the difference between the least and most importance features is not very large; the relative importance of features looks not very pronounced contrary to other methods. We tend to conclude that the Gradient Boosting Classifier is not very useful for our strategy to minimize the number of false negatives.

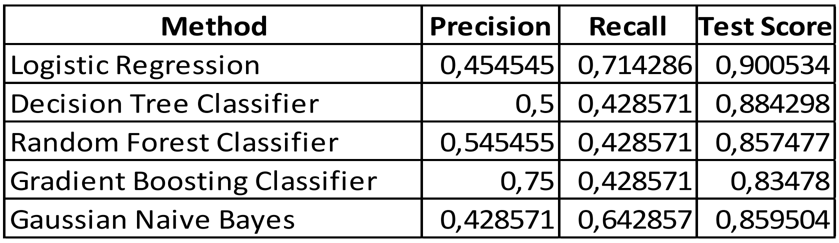


e) Gaussian Naive Bayes

Gaussian Naive Bayes belongs to a family of simple probabilistic classifiers based on applying Bayes' theorem, and is the fastest method among those considered here. It gives a decent overall accuracy (around 86 percent), and a higher value of recall (0.64) compared to precision (0.43). Unfortunately, it is not possible to obtain the relative importance of features in this method. Again, it is worth noting that the training and test scores do not vary significantly.



Let's create the comparison table that contains test scores, precisions and recalls pertaining to each method. The summary suggests that Logistic Regression is the best algorithm from the point of view of high test score accuracy and high value of recall. At the same time, Decision Tree Classifier displays the poorest performance. Gaussian Naive Bayes seems to be the best second method as seen from the table below.



Method	Precision	Recall	Test Score
Logistic Regression	0,454545	0,714286	0,900534
Decision Tree Classifier	0,5	0,428571	0,884298
Random Forest Classifier	0,545455	0,428571	0,857477
Gradient Boosting Classifier	0,75	0,428571	0,83478
Gaussian Naive Bayes	0,428571	0,642857	0,859504

4.2 Classification Models: Statistical Significance

The results in the table above are obtained after running five classification models given the particular sets of tunable parameters to choose from for each model, the particular value of cross-validation parameter, and, what is probably most important, the particular way of splitting the entire data set onto the training and test sets. One can ask then the question of how the results may change if we use other sets and values for the above-mentioned parameters. The fact that the results are likely to be affected significantly is bolstered by the small size of our dataset. Will the preferences regarding the methods change or remain the same? Trying to answer this question at least approximately, we run the algorithms for 100 different ways of the train-test splitting by assigning 100 different values to the random_state parameter. The test scores, recalls and precisions will be stored in arrays for each of five models. We then create another table that contains means, standard deviations and 95 percentiles of the test score, recall and precision for all five classification models used to make predictions.

	DTC	GNB	GBC	LOG REG	RFC
mean value of the test score	0,857933884	0,826198347	0,869165554	0,896455274	0,864889853
standard deviation of the test score	0,029445041	0,031371339	0,048908156	0,029467757	0,042342659
95 percentile of the test score	[0.80557851 0.91735537]	[0.7642562 0.88863636]	[0.78214286 0.93826769]	[0.8370494 0.94587784]	[0.79216455 0.94047897]
mean value of precision	0,403297136	0,370987004	0,503673863	0,371504983	0,458637884
standard deviation of precision	0,11935387	0,063708735	0,135852881	0,049284029	0,132304952
95 percentile of precision	[0.22374582 0.65535714]	[0.25758065 0.51141304]	[0.25307487 0.75]	[0.27796875 0.47619048]	[0.21055556 0.77625]
mean value of recall	0,416428571	0,691428571	0,351428571	0,827142857	0,487142857
standard deviation of recall	0,1245625	0,118183531	0,108175972	0,093775677	0,192825394
95 percentile of recall	[0.21428571 0.64285714]	[0.4625 0.89464286]	[0.14285714 0.57142857]	[0.64285714 1.]	[0.14285714 0.82321429]

The results of the table above show that overall the test scores for all methods don’t vary significantly. Logistic Regression remains the best method in terms of the values for the test score and recall. The standard deviations are also quite small for these parameters for the Logistic Regression. Gaussian Naive Bayes is only somewhat worse than Logistic Regression in terms of the accuracy of the test score and recall. Gradient Boosting Classifier seems to give the highest average value of precision, but the lowest average value of recall. One should notice also very large 95 percentile intervals for the values of precision and recall for Decision Tree and Random Forest, which is indicative of the high dependence of results on the train-test splitting for these models.

5. Future Research and Recommendations

To conclude, Logistic Regression is the best method to make predictions trying to avoid the presence of false negatives as much as possible. The most important features that play the role in determining whether this or that hurricane will lead to emergencies are the 'MAXIMUM EFFECTIVE STRENGTH', 'DISTANCE TRAVELLED' and the 'AVERAGE WIND WRT MAXIMAL'. It is a bit surprising that Logistic Regression contrary to the tree based methods does not list the 'FIRST LAT' and 'FIRST LONG' features as those that strengthen the probability of emergencies. It looks like the more powerful the hurricane is, the more likely it is to make a landfall and bring damage; weaker storms are more likely to dissolve while still in the ocean or veer simply off the coast. Unfortunately, this particular observation is of limited help in making early enough predictions whether a hurricane is capable of inflicting the level of damage that may trigger emergency declarations. The above-mentioned features are rather well correlated but are nothing more than the documented properties of each hurricane. For the future investigation, it is much more important to look at datasets with features that potentially may shed light on the issues like why some hurricanes gain extraordinary power, spend a lot of time above the ocean surface and travel large distances. As a result of the current study, one can only recommend to pay extra attention to the storms that reach high levels of strength, especially those that do this travelling long distances.

6. Acknowledgements

I would like to thank the teams working on building and improving Scikit-Learn and Pandas, as well as opensource communities.

Special thanks to my mentor Hassan Waqar who provided me with many invaluable advices and guidance.