PREDICTING HURRICANE LANDFALLS FOR THE US GULF COAST

Springboard Capstone Project
Seth Hill
May 1st 2018

OVERVIEW

- Problem and Motivation
- Cleaning and Wrangling Datasets
- Initial Observations
- Statistical Analysis Results
- Machine Learning Predictive Performance
- Future Steps

PROBLEM AND MOTIVATION

- Every year the Gulf Coast is impacted by numerous tropical weather systems including hurricanes causing billions in damage and loss of life.
  - Currently, Gulf Coast states could use a simple way to assess or forecast potential landfalls for upcoming hurricane seasons.
- The federal, state, and local governments, especially those along the Gulf Coast, would be interested in a way to predict the impact of upcoming hurricane seasons.
  - This could help with budgeting and resource allocation allowing for a more timely response to potential hurricane damage.

CLEANING AND WRANGLING DATASETS

NOAA WORLD OCEAN DATABASE (WOD)

- Massive database storing info on many aspects of the world's oceans.
- Grouped by time, location, and/or scientific instrument utilized.
- Data is stored in proprietary WOD profiles:
  - Depth profiles of various ocean properties.
  - Time and location stored as attributes.

WIKI TABLE HISTORICAL HURRICANE LANDFALLS

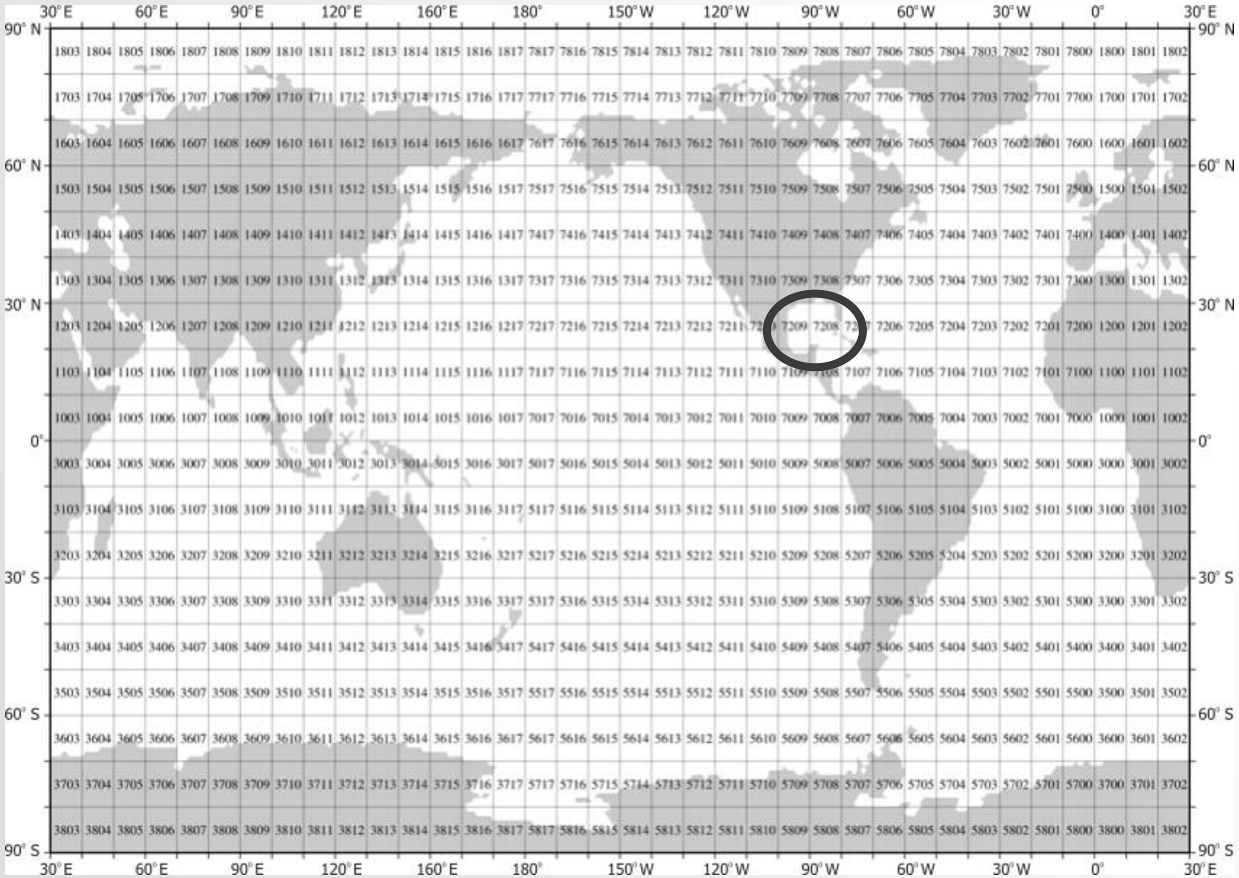
- Wiki tables separated by state.
- Contain info on date, name, and category of hurricane landfalls.
- Footnotes and special comments embedded in the tables.

CLEANING AND WRANGLING DATASETS
WORLD OCEAN DATABASE
DATA SELECTION

- Data is grouped by instruments utilized:
  - OSD utilized due to bigger variety of data available in one dataset.

Dataset	Source
OSD	Bottle, low-resolution Conductivity-Temperature-Depth (CTD), low-resolution XCTD data, and plankton data
CTD	High-resolution Conductivity-Temperature-Depth (CTD) data and high-resolution XCTD data
MBT	Mechanical Bathythermograph (MBT) data, DBT, micro-BT
XBT	Expendable (XBT) data
SUR	Surface only data (bucket, thermosalinograph)
APB	Autonomous Pinniped Bathythermograph - Time-Temperature-Depth recorders attached to elephant seals
MRB	Moored buoy data from TAO (Tropical Atmosphere-Ocean), PIRATA (moored array in the tropical Atlantic), MARNET, and TRITON (Japan-JAMSTEC)
PFL	Profiling float data
DRB	Drifting buoy data from surface drifting buoys with thermistor chains
UOR	Undulating Oceanographic Recorder data from a Conductivity/Temperature/Depth probe mounted on a towed undulating vehicle
GLD	Glider data

CLEANING AND DATA WRANGLING
WORLD OCEAN DATABASE
DATA SELECTION



- After choosing measurements from OSD instrument:
  - Data can be selected by location or time.
  - Data was chosen by location to focus on the Gulf of Mexico.
  - This corresponded to map sections 7208 and 7209.

CLEANING AND WRANGLING DATASETS
WORLD OCEAN DATABASE
DATA WRANGLING AND CLEANING

- File Download and Incorporation into Pandas data frames:
  - Files need to be unzipped and transformed from a proprietary ascii format organized in profiles by location and time.
    - Utilized wodpy python package to store profiles in data frames.
  - Compiled and filtered out empty profiles with pandas into one data frame.

CLEANING AND WRANGLING DATASETS
WORLD OCEAN DATABASE
DATA WRANGLING AND CLEANING

- Exploring, Manipulating, and Cleaning
  - Parsed individual day, month, year columns into one datetime column.
  - Zipped individual latitude and longitude columns into one location column of tuples rounded to the nearest degree.
  - Dropped all data predating 1960 due to lack of data reliability.
  - Only Oxygen, Temperature, Phosphate, Silicate, Salinity data utilized.
    - Other measurements sparse in data points.
  - Data resampled by location and month.
    - Missing values filled with location average first.
    - Remaining missing values back filled.

CLEANING AND WRANGLING DATASETS
WORLD OCEAN DATABASE
DATA WRANGLING AND CLEANING

- Finished Product

		oxygen	phosphate	salinity	silicate	temperature
location	**date**					
(20.0, -80.0)	**1965-05-31**	4.595000	0.00	36.150000	5.0	27.620000
(20.0, -81.0)	**1965-05-31**	4.640000	0.23	36.260000	6.0	26.930000
	1972-04-30	4.640000	0.23	35.870000	6.0	26.340000
(20.0, -82.0)	**1965-05-31**	4.590000	0.23	36.120000	6.0	27.050000
	1970-02-28	4.633333	0.23	35.820000	6.0	25.400000
	1972-04-30	4.750000	0.23	35.887500	6.0	26.350000
	1989-12-31	4.560000	0.23	35.917000	6.0	27.530000
(20.0, -83.0)	**1968-04-30**	4.580000	0.23	36.196000	6.0	27.075885
	1968-11-30	4.532500	0.23	36.049000	6.0	27.840000
	1970-05-31	4.532500	0.23	36.064875	6.0	28.300000

CLEANING AND WRANGLING DATASETS
HISTORICAL HURRICANE LANDFALLS
DATA SELECTION

- Data was scraped from List of United States Hurricanes wiki page into pandas data frames using the python wikipedia package.
  - Data tables were separated by state and each table was stored in a split table format.

Out[48]:

	Name	Saffir-SimpsonCategory	Date of closest approach	Year	Unnamed: 4	Name.1	Saffir-SimpsonCategory.1	Date of closest approach.1	Year.1
0	Unnamed	3	August 26	1852.0	NaN	Unnamed	2	October 18	1916.0
1	Unnamed	1	September 29	1917.0	NaN	Unnamed	3	August 21	1926.0
2	Unnamed	1[notes 1]	August 31	1856.0	NaN	Unnamed	1	September 1	1932.0
3	Unnamed	1	September 16	1859.0	NaN	Baker	1	August 31	1950.0
4	Unnamed	2	August 12	1860.0	NaN	Camille	1	August 18	1969.0
5	Unnamed	1	September 16	1860.0	NaN	Eloise	1[notes 1]	September 23	1975.0
6	Unnamed	1	July 30	1870.0	NaN	Frederic	3	September 13	1979.0
7	Unnamed	1[notes 1]	September 10	1882.0	NaN	Elena	3	September 2	1985.0
8	Unnamed	2	October 3	1893.0	NaN	Opal	1[notes 1]	October 4	1995.0
9	Unnamed	1	August 15	1901.0	NaN	Danny	1	July 19	1997.0
10	Unnamed	2	September 27	1906.0	NaN	Ivan	3	September 16	2004.0
11	Unnamed	1	September 14	1912.0	NaN	Dennis	1 [notes 1]	July 10	2005.0
12	Unnamed	2	July 5	1916.0	NaN	Katrina	1	August 29	2005.0

CLEANING AND WRANGLING DATASETS
HISTORICAL HURRICANE LANDFALLS
DATA WRANGLING AND CLEANING

- Fixed the side by side data by making one longer data frame for each state.
- Parsed times from separate columns into one column.
- Deleted blank rows.
- Removed special [notes] in the category columns.
- Renamed the remaining columns.
- Replaced special entries in quotations in the named column with "Unnamed".
- Added a corresponding state label for each dataframe entry.
- Removed any entries pre-dating 1960 from the dataframe.

CLEANING AND WRANGLING DATASETS
HISTORICAL HURRICANE LANDFALLS
DATA WRANGLING AND CLEANING

- Finished Product

Out[46]:

	name	category	state
1960-09-10	Donna	4	Florida
1960-09-15	Ethel	1	Mississippi
1961-09-11	Carla	4	Texas
1963-09-17	Cindy	1	Texas
1964-08-27	Cleo	2	Florida
1964-09-10	Dora	2	Florida
1964-09-15	Ethel	1	Louisiana
1964-10-03	Hilda	3	Louisiana
1964-10-14	Isbell	2	Florida
1965-09-08	Betsy	3	Florida
1965-09-10	Betsy	3	Louisiana

INITIAL OBSERVATIONS

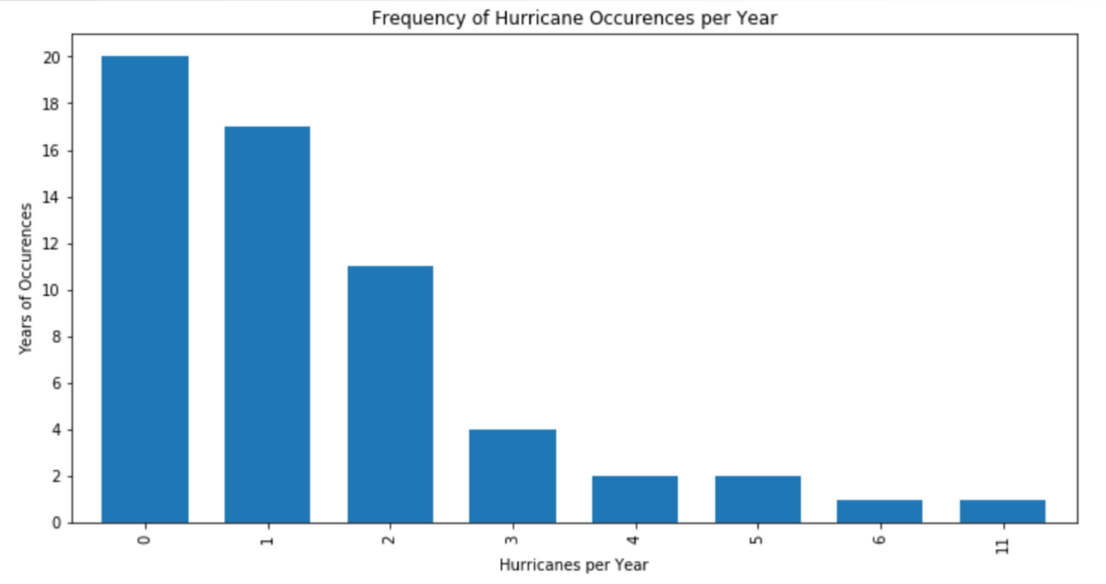
- Answer three main questions:

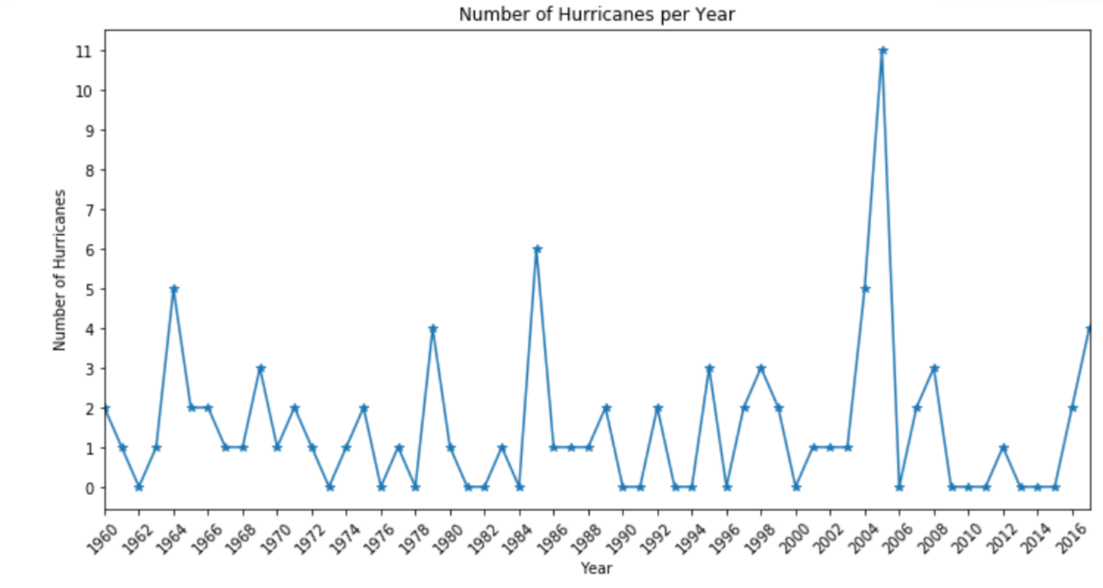
1. What are the historical hurricane landfall summary stats?
2. Do any WOD ocean properties correlate with hurricane landfall frequency?
3. Are there any correlations when data is broken down by state?

- For one and two, the data was averaged yearly for the gulf as a whole.

INITIAL OBSERVATIONS
HISTORICAL HURRICANE LANDFALL STATS

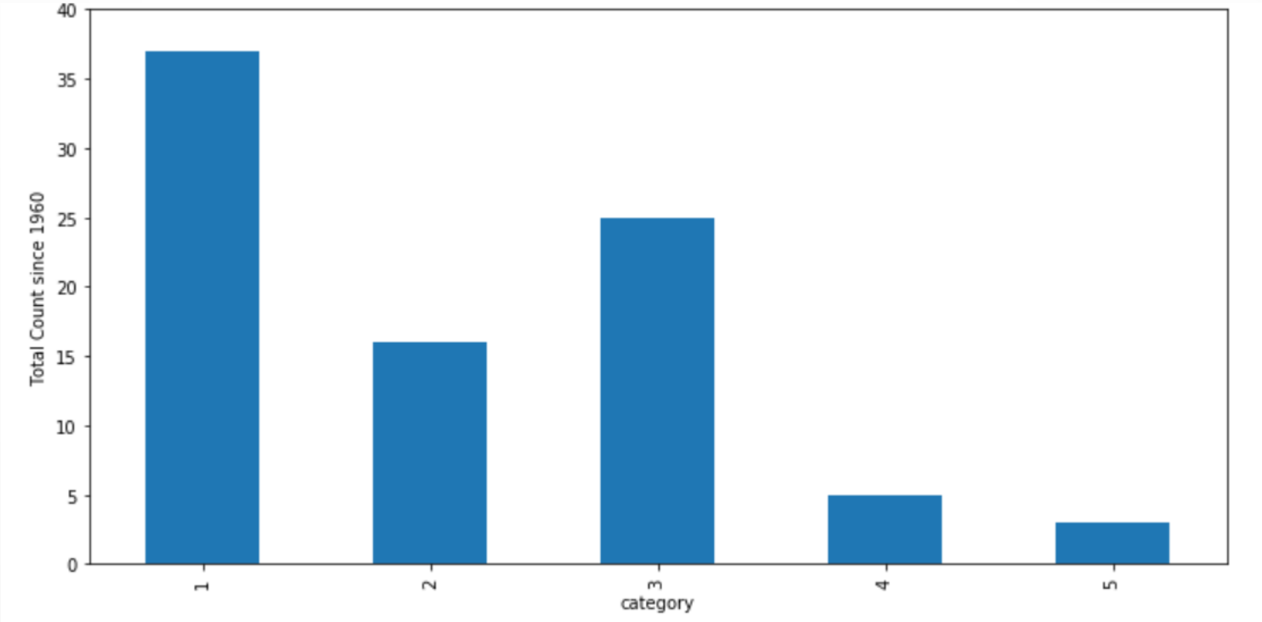
- 90% of years show US gulf coast had 3 landfalls or less.
- Most years have 0 or 1 landfall per year.
- 2005 is an unusually active year.





INITIAL OBSERVATIONS
HISTORICAL HURRICANE LANDFALL STATS

- Hurricane strength is notated by categories one through five.
  - One is the weakest and five is the strongest.
- The most frequent occurrence is category one with category five being least frequent.



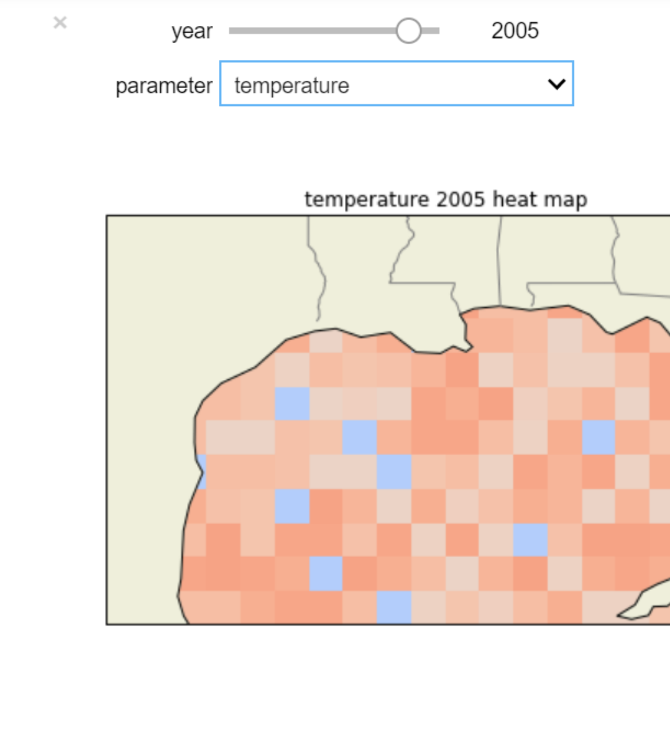
INITIAL OBSERVATIONS
WOD OCEAN PROPERTIES AND LANDFALL FREQUENCY

- Pearson Correlation Coefficients were calculated for each Ocean Property vs. Landfall Frequency combination.
- Initially only temperature appears to be relevant.

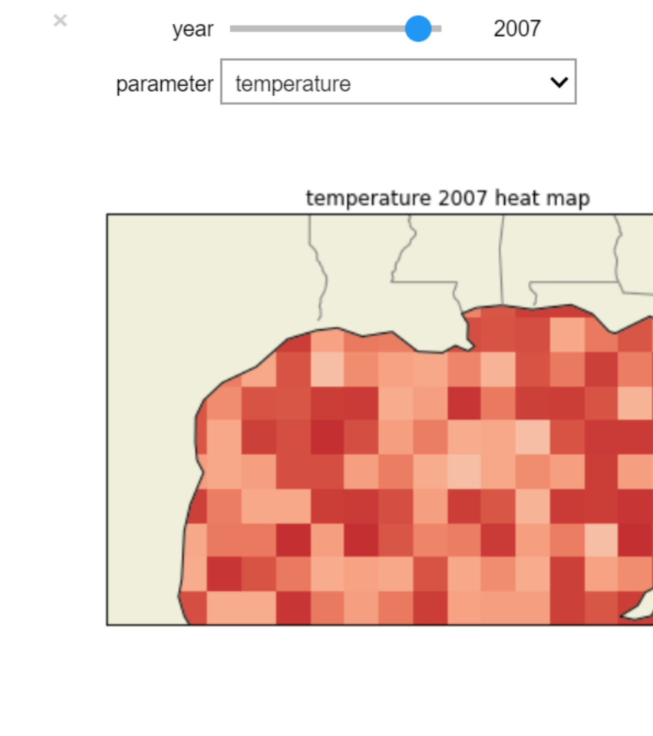
Ocean Parameters	Correlations with Number of Hurricanes per Year
Oxygen	-0.139101
Phosphate	0.161184
Salinity	-0.214846
Silicate	0.122182
Temperature	-0.362684

INITIAL OBSERVATIONS
WOD OCEAN PROPERTIES AND LANDFALL FREQUENCY

- Generated heatmaps also illustrate the apparent negative correlation between temperature and landfall frequency.



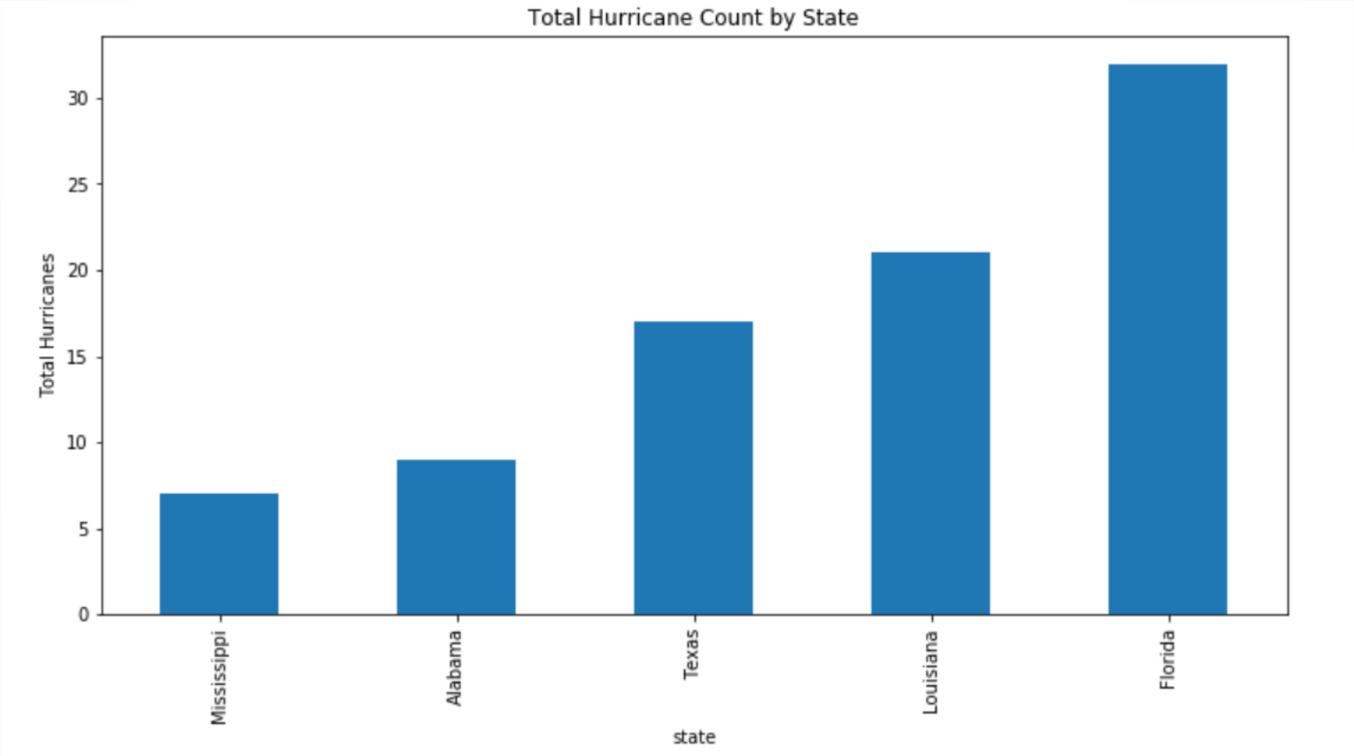
11 landfalls



2 landfalls

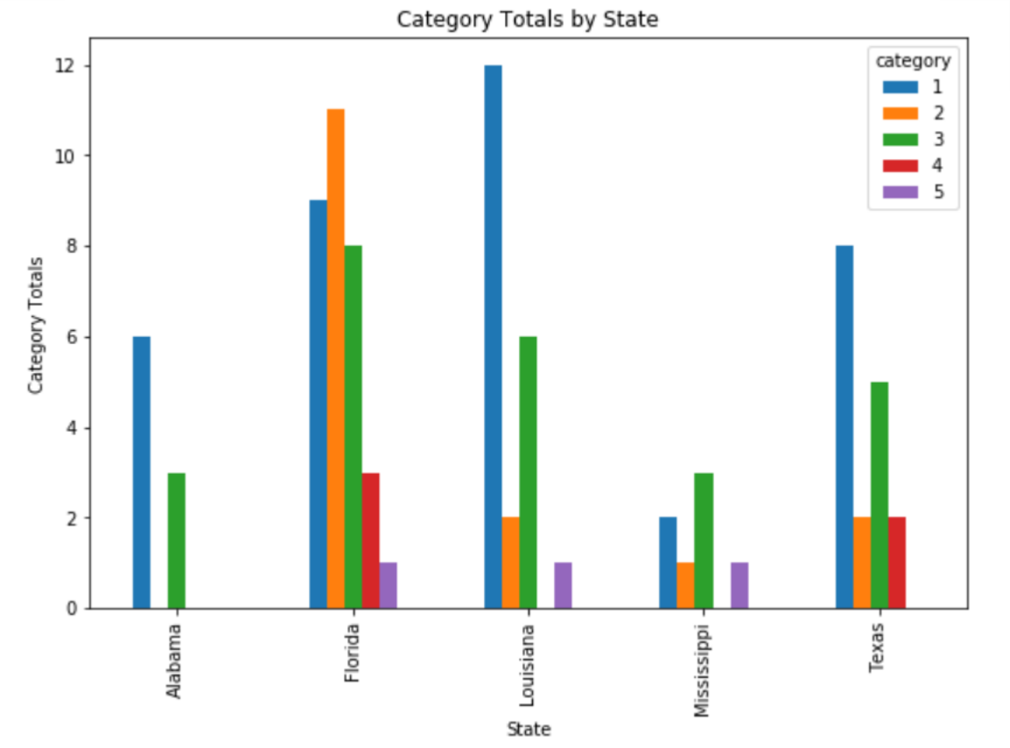
INITIAL OBSERVATIONS
DATA ANALYZED BY STATE

- Florida and Louisiana have the most landfalls historically.
- Mississippi and Alabama have the least landfalls historically.



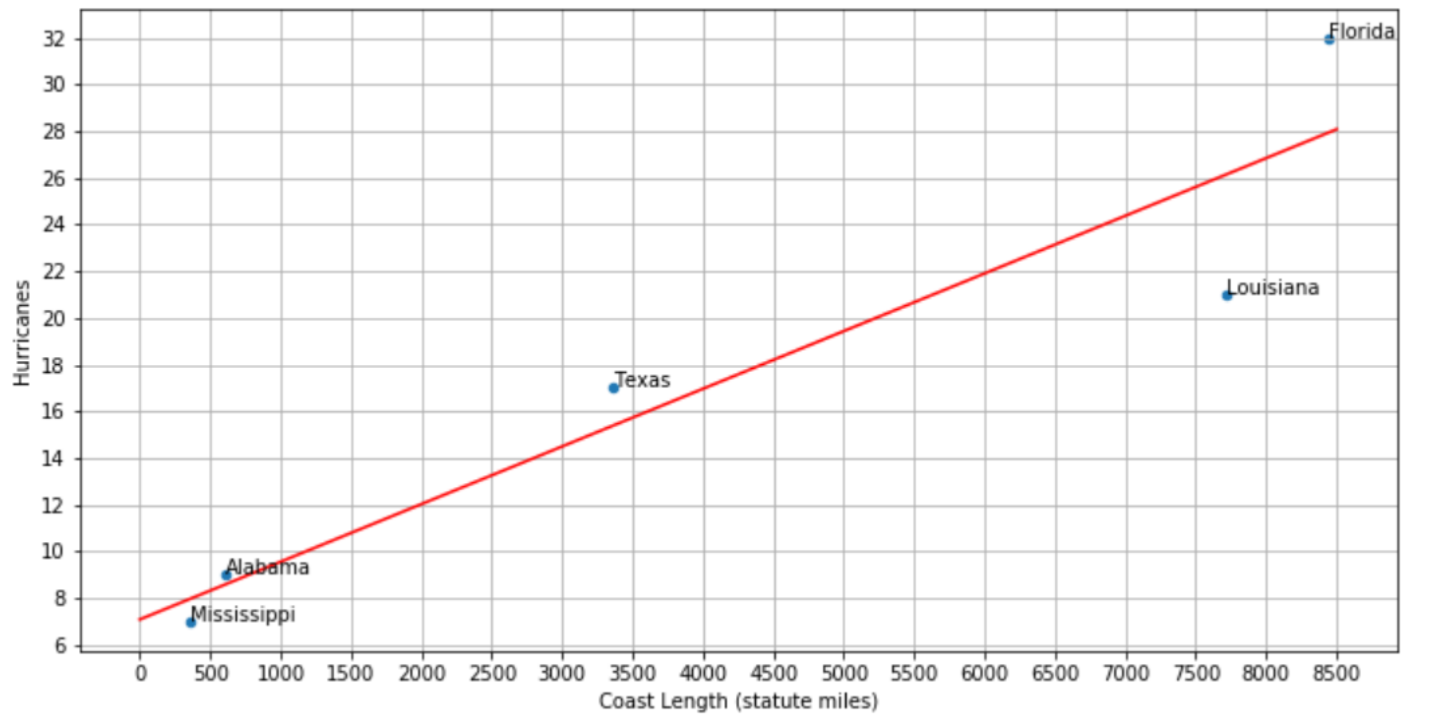
INITIAL OBSERVATIONS
DATA ANALYZED BY STATE

- Florida is impacted by higher strength storms on average.
- Other states mostly impacted by category 1 landfalls.



INITIAL OBSERVATIONS
DATA ANALYZED BY STATE

- There appears to be a strong correlation between coast length and historical landfall totals.




```
The correlation between coast length and number of hurricanes: 0.9402222593630062
```

STATISTICAL ANALYSIS RESULTS

- Hypotheses:
  - Correlation of landfall frequency to any of the World Ocean Database features.
  - Correlation of average yearly strength of landfalls to any of the World Ocean Database features.
  - Coast length correlates with overall landfalls per state.

STATISTICAL ANALYSIS RESULTS
LANDFALL FREQUENCY AND AVERAGE STRENGTH

- Test setup utilizing hacker statistics:
  - Null hypothesis is that the data aren't correlated.
  - Create permutation replicates of the wod feature data while holding the hurricane data constant.
  - Test statistic is the correlation coefficient.
  - p-value is the sum of the correlations at least as extreme as the empirical correlation.
  - alpha = 0.05 significance level.

STATISTICAL ANALYSIS RESULTS
LANDFALL FREQUENCY AND AVERAGE STRENGTH

- Test results:
  - The only ocean property significantly correlated to landfall frequency is temperature.
  - The only ocean property significantly correlated to average yearly strength is oxygen content.

Landfall Frequency			
WOD Feature	**Empirical Correlation**	**p-value**	**Significant Result**
Oxygen	-0.139	0.156	No
Phosphate	0.161	0.106	No
Salinity	-0.215	0.075	No
Silicate	0.122	0.156	No
Temperatu re	-0.363	0.006	Yes

Average Yearly Strength			
WOD Feature	**Empirical Correlation**	**p-value**	**Significant Result**
Oxygen	-0.243	0.038	Yes
Phosphate	0.033	0.382	No
Salinity	0.052	0.355	No
Silicate	0.133	0.174	No
Temperatu re	-0.204	0.070	No

STATISTICAL ANALYSIS RESULTS
COAST LENGTH AND OVERALL LANDFALLS BY STATE

- Test setup:
  - Null hypothesis is that the data aren't correlated.
  - Create permutation replicates of the coast length data while holding the number of hurricanes data constant.
  - Test statistic is the correlation coefficient.
  - p-value is the sum of the correlations at least as extreme as the empirical correlation.
  - alpha = 0.05 significance level.
- Test results:
  - The coast length was significantly correlated to the overall landfalls by state.

Empirical Correlation	p-value	Significant Result
0.940	0.009	Yes

MACHINE LEARNING PREDICTIONS

- Goal:
  - Predict categorical yearly landfall frequency.
- Categories:
  - No Impacts (0)
    - 0 hurricanes
  - Moderate (1)
    - 1-2 hurricanes
  - Severe (2)
    - 3+ hurricanes

MACHINE LEARNING PREDICTIONS

- Tested 4 different feature sets and 3 different algorithms:
  - Feature Sets:
    - Temperature only.
    - All WOD features.
    - Above with last years WOD features.
    - Above with last years hurricane frequency and average yearly strength.
  - Algorithms:
    - Logistic Regression
    - SVM Classification
    - kNN
- All features were standardized and key hyperparameters tuned.
- Performance was based on the f1 scores of the models predictions of a set aside test set.

MACHINE LEARNING PREDICTIONS

- Model performance results and tuned hyperparameter(s):

	Logistic Regression	**SVM Classification**	**kNN**
Temperature Only	f1 = 0.60 C = 373	f1 = 0.66 C = 22425 gamma = 0.04	f1 = 0.71 k = 9
All WOD Features	f1 = 0.50 C = 0.001	f1 = 0.60 C = 88 gamma = 0.008	f1 = 0.60 k = 11
+ Last Year's WOD Features	f1 = 0.33 C = 0.001	f1 = 0.49 C = 243 gamma = 0.0004	f1 = 0.38 k = 7
+ Last Year's Frequency and Strength	f1 = 0.47 C = 2.74	f1 = 0.27 C = 203 gamma = 0.00004	f1 = 0.37 k = 13

MACHINE LEARNING PREDICTIONS

- The best performing combination is kNN with temperature as the only feature.
  - k = 9
  - f1 score = 0.71
    - Precision = 0.8
    - Recall = 0.75
- More data and research into new relevant features needs to be done in order to improve model performance.

FUTURE STEPS

- Research new features.
  - El Nino
  - Wind Patterns coming off of Africa
  - and more…
- Add in more data.
  - Global ocean properties
  - Global landfalls
- Test more algorithms.
  - Random Forests
  - Decision Trees
  - Neural Networks
  - Etc…