An Analysis of Tropical Cyclones in the Atlantic Basin: 1950-2015

I. Introduction

Team Phoenix was interested in tropical cyclones and how often this rare phenomenon affects the United States coastlines. These events are often extremely dangerous to life and property and can have immense economic impact on society.

For our project, we decided to create a visualization dashboard for meteorologists to view historical data on hurricanes. Our dashboard allows meteorologists to monitor tropical cyclone activity and analyze hurricane metrics from the past.

II. Data

https://www.kaggle.com/noaa/hurricane-database

Using the National Hurricane Center (NHC) tropical cyclone historical database, known as HURDAT (HURricane DATabase), we analyzed named storms of at least tropical storm strength from the years 1950 to 2015. The Atlantic HURDAT database contains six-hourly information on the location (latitude/longitude coordinates), maximum winds (knots), and central pressure (millibars) of each storm. This data set was chosen because of the integrity and reliability of data compiled by the NHC and the large body of works that have been presented in past literature (Landsea (2004) and Landsea (2007)).

a. Data Cleaning/Data Processing

The National Hurricane Center (NHC) tropical cyclone historical database consists of data from both the Atlantic basin and the Pacific Ocean. For our analysis, our group decided to focus on tropical cyclones in the Atlantic ocean. The Atlantic ocean dataset that we are using has 22 columns and 49,105 rows. The following are the original columns in our dataset: ID, Name, Date, Event, Status, Latitude, Longitude, Maximum Wind, Minimum Pressure, Low Wind NE, Low Wind SE, Low Wind SW, Low Wind NW, Moderate Wind NE, Moderate Wind SE, Moderate Wind SW, Moderate Wind NW, High Wind NE, High Wind SE, High Wind SW, High Wind NW.

We first chose to condense our data. This involved eliminating all entries with a date before 1950, the year the United States started naming storms. We also removed all Low, Moderate, and High Wind columns and only kept the Maximum Wind column in our dataset. This left us with the following columns: ID, Name, Date, Event, Status, Latitude, Longitude, Maximum Wind, Minimum Pressure. To clean our data further, we converted all entries in the Date column to the Python datetime format, cleaned the Latitude and Longitude columns by removing “N”, “S”, “E”, and “W”, from its entries, and used the str.strip() function to eliminate extra whitespace from all the columns labeled ‘object.’ These changes make our data easier to sort and query. Finally, we added a Category column that accounted for the strength of the tropical storm (category 0 indicates a storm less than hurricane strength). The Category column was populated based on the Maximum Wind column from the Saffir-Simpson hurricane wind scale.

Upon further examination, we see that -999 indicates missing values in our new *atlantic_df* Pandas dataframe. We replaced these values with ‘NA’ to allow for aggregation functions to be applied to our data set.

After completing the above data cleaning and processing, our final unaggregated data set consists of columns: ‘ID’, ‘Name’, ‘Date’, ‘Time’, ‘Event’, ‘Status’, ‘Latitude’, ’Longitude’, ‘Maximum Wind’, ‘Minimum Pressure’, ‘Datetime’, and ‘Category’.

To perform aggregation functions on our data set, we further processed the data to account for storm track lengths and landfall data. Our final aggregated data set consists of columns: 'ID', 'Name', 'initialDate', 'endDate', 'duration', 'netDistanceKm', 'totalDistanceKm', 'maxLandSpeed', 'minLandSpeed', 'meanLandSpeed', 'pressureMean', 'pressureStDev', 'pressureMin', 'pressure25Pct', 'pressureMedian', 'pressure75Pct', 'pressureMax', 'pressureDelta', 'windMean', 'windStDev', 'windMin', 'wind25Pct', 'windMedian', 'wind50Pct', 'windMax', 'windDelta', 'maxCategory', 'landfallBool', 'landfallTimeDelta', 'landfallDatetime', 'landfallCategory', 'landfallLong', 'landfallLat'.

III. Experimental Design

As stated above, we are using the National Hurricane Center (NHC) tropical cyclone historical database, known as HURDAT (HURricane DATabase) in our analysis. Below is an outline of our experimental design process which details each step we took to analyze our data.

a. Process:

i. Obtained data from kaggle hurricane database

ii. Cleaned and processed data to correct the formatting of variables, and to create new variables

1. Turned date and time variables into a proper datetime format

2. Converted longitude and latitude strings into proper decimal form, to allow distance to be calculated from coordinates

3. Created a variable representing hurricane category, based upon existing wind speed

4. Cleaned various string variables

iii. Created a new, aggregated dataset with new variables per-storm, including:

1. Distance traveled, based upon the distance between each observation's longitude and latitude values

2. Duration, based upon the delta between the initial datetime value for each storm, and the final datetime value

3. Landspeed, based upon the distance traveled and the duration of the storm

4. Variables related to landfall, whether or not the storm reached land; includes variables such as:
    a. Datetime for when the storm arrived on land
    b. A boolean variable denoting whether or not the particular storm achieved landfall
    c. What Category the storm was when it achieved landfall
    d. Coordinates at which the storm achieved landfall
5. Various summary statistics for windspeed and pressure

iv. Saved the cleaned and aggregated datasets into separate csv files

v. Read in our new datasets from csv files

vi. Ran several queries on the data that revealed the following interesting metrics from our data:

1. 132 hurricanes made landfall
2. 277 reached high magnitude, but did not make landfall
3. 165 Category 1 storms
4. 22 Category 5 storms

vii. Produced Visualizations based on our experimental queries.

Our experimental design process ended with the production of several graphs and plots that highlight the results of our experimental queries.

IV. Results

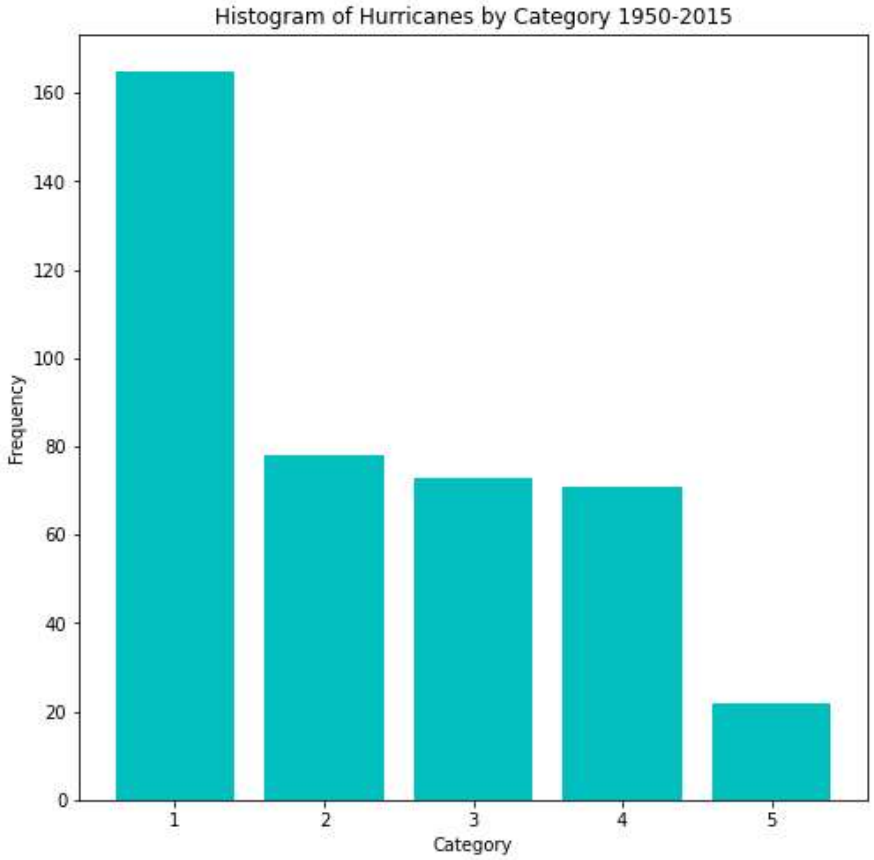

Figure 1. Histogram of hurricane strength storms by Category (1950-2015).

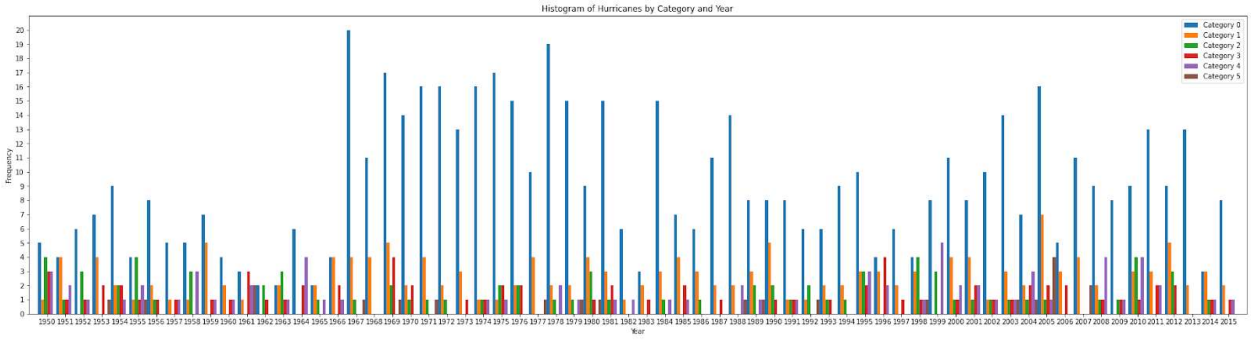

Figure 2. Histogram of hurricane-strength storms by category for each year (1950-2015).

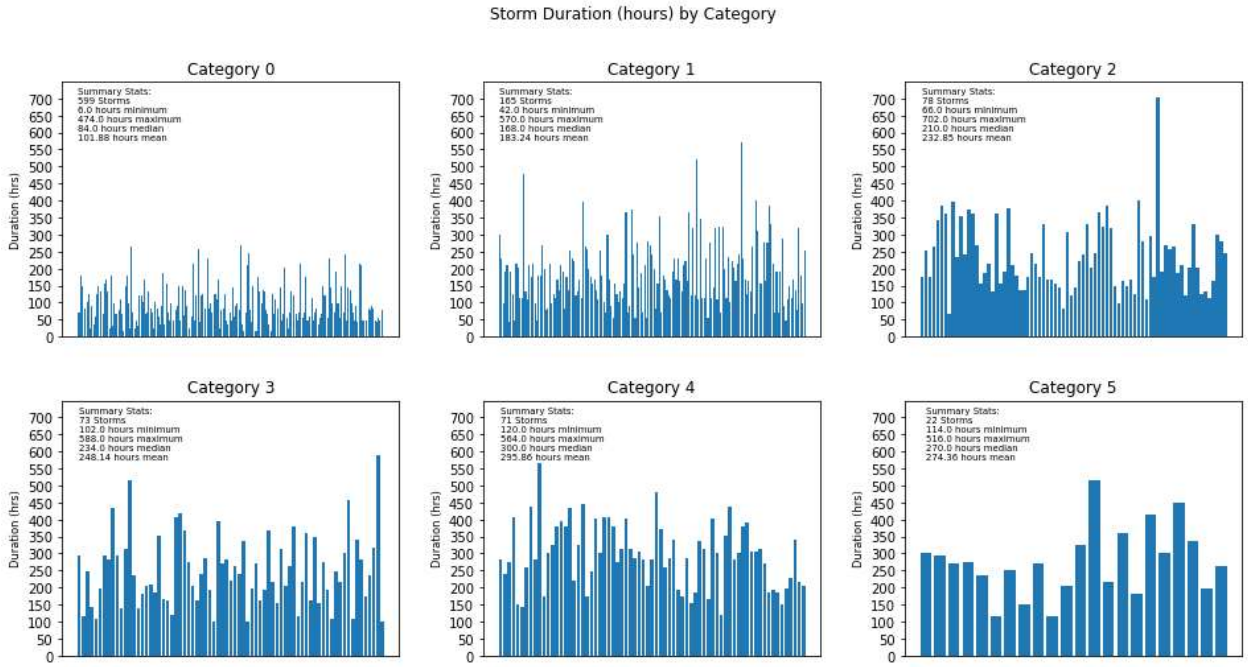

Figure 3. Histograms of Storm Duration (in hours) by Category, with summary statistics (1950-2015).

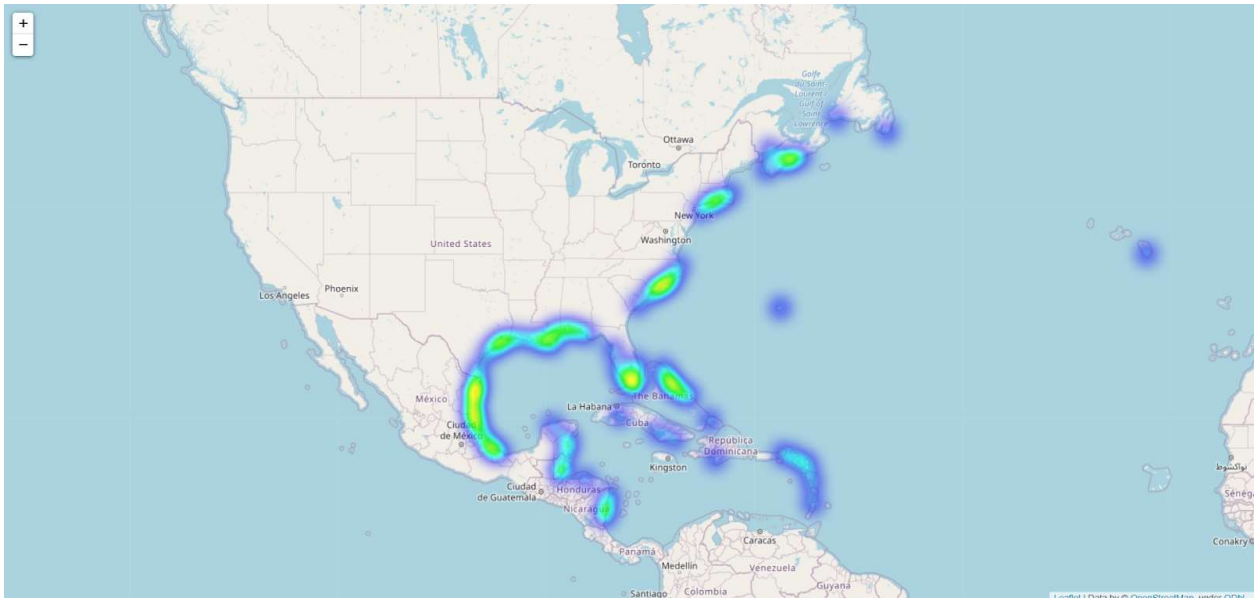

Figure 4. Heat map of storms that made landfall (1950-2015).

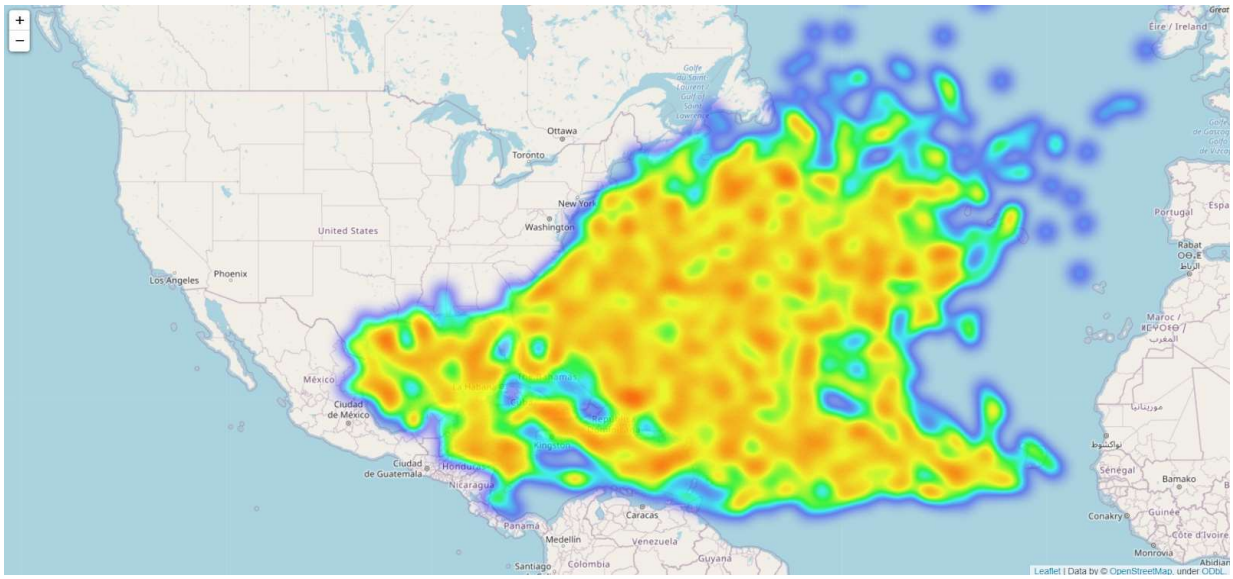

Figure 5. Heat map of storms that did not make landfall (1950-2015).

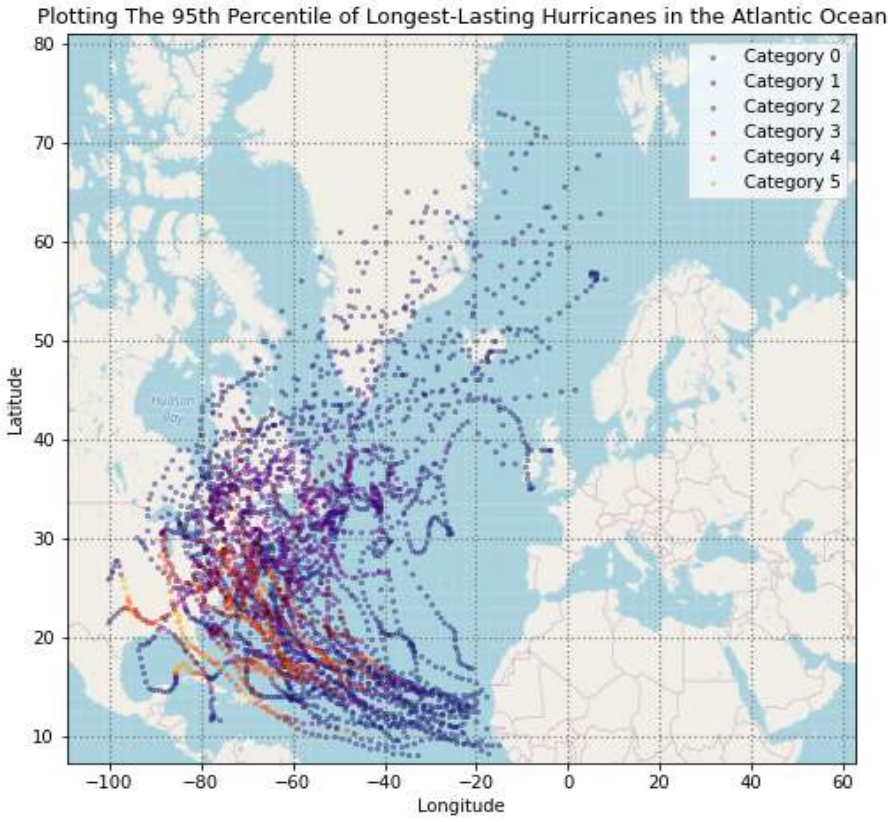

Figure 6. Plot of the longest duration hurricane tracks (1950-2015).

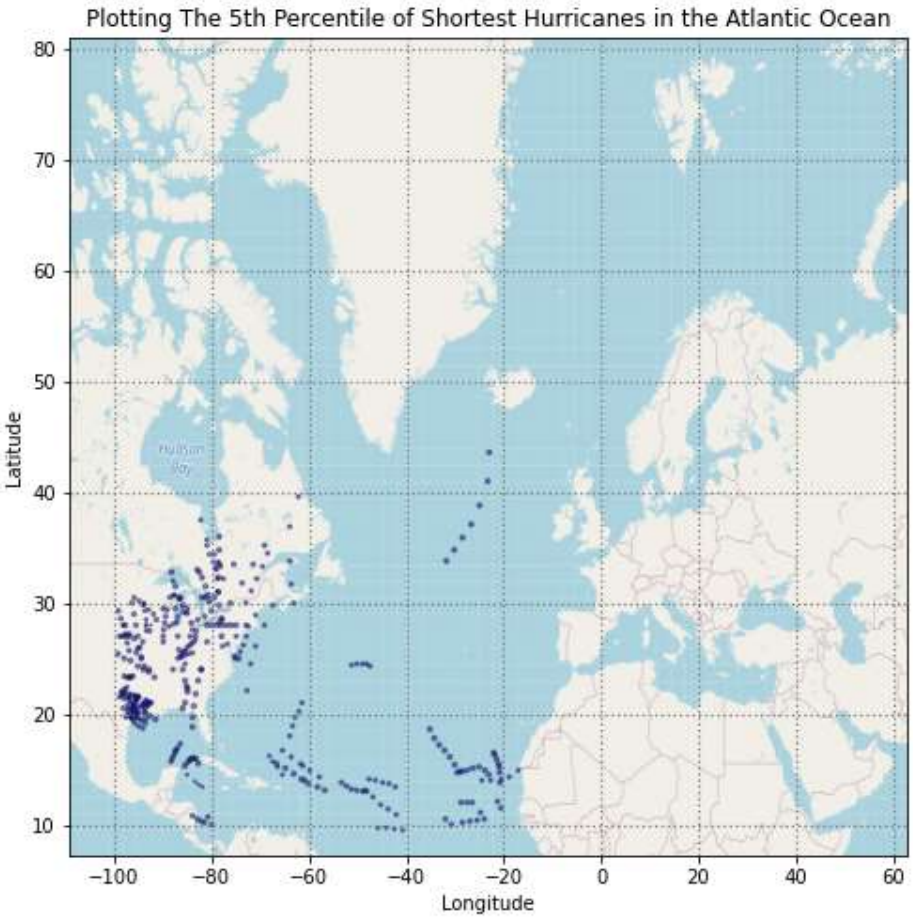

Figure 7. Plot of the shortest duration hurricane tracks (1950-2015).

V. Testing

We divided our code into three sections: Data Cleaning, Data Aggregation, and Data Visualization. We focused our testing on the data cleaning section of our project to ensure that our data was processed correctly before running our analysis.

We created a CleanData_test.py that contained all of our tests for CleanData.py. Our CleanData.py file contains six functions that each handle a unique step of our data cleaning process. We intentionally structured these functions to all have a return value of the processed data to better test its functionality.

Our CleanData_test.py contains three separate units tests for the data cleaning section of our project. We test the readData() function which reads from our kaggle dataset and stores its contents into a pandas dataframe. This test checks whether the correct dataset was downloaded by verifying the column names are present in the returned dataframe.

```
# test read_data function
def test_readData(self):
    data = readData("atlantic.csv")

    # assert that the read data has the correct columns
    assert list(data.columns) == ['ID', 'Name', 'Date', 'Time', 'Event',
                                  'Status', 'Latitude','Longitude',
                                  'Maximum Wind', 'Minimum Pressure',
                                  'Low Wind NE','Low Wind SE',
                                  'Low Wind SW', 'Low Wind NW',
                                  'Moderate Wind NE', 'Moderate Wind SE',
                                  'Moderate Wind SW', 'Moderate Wind NW',
                                  'High Wind NE', 'High Wind SE',
                                  'High Wind SW', 'High Wind NW'
                                  ]
```

Our next unit test evaluates the condenseData() function. This function removes hurricanes before 1950 from our dataset and subsets relevant variables and columns in our dataset. Our unit test for this function checks whether the correct columns were removed from our dataframe.

```python
# test condense_data function
def test_condenseData(self):
    data = readData("atlantic.csv")

    condensed_data = condenseData(data)

    # assertions for all remaining columns in df
    assert list(condensed_data.columns) == ["ID", "Name", "Date", "Time",
                                            "Event", "Status", "Latitude",
                                            "Longitude", "Maximum Wind",
                                            "Minimum Pressure"
                                            ]

    # assertions for columns that should have been removed from df
    assert "Low Wind SE" not in condensed_data.columns
    assert "Low Wind SW" not in condensed_data.columns
```

Finally, the last unit test for this section tests our createAdditionalColumns() function. This test confirms that the correct column headers were added to our dataframe.

```python
# test create_additional_columns function
def test_createAdditionalColumns(self):
    data = readData("atlantic.csv")
    data = condenseData(data)
    data = removeWhitespace(data)
    data = processMaxWind(data)
    data = createAdditionalColumns(data)


    # assertions for all added columns
    assert "Datetime" in data.columns
    assert "Category" in data.columns
```

We also added a few unit tests to test the validity of the data used to plot our HeatMaps:

```
import unittest
import HeatMap as hm


class HeatMapTest(unittest.TestCase):


    #Test to determine that all values in the landfall dataframe 'Event' column have a value of 'L'
    def testLandFallCols(self):
        df = hm.loadData()
        df_landfall = hm.hurricaneLandFall(df)
        found = df_landfall[df_landfall['Event'].str.contains('L')]
        land_count = len(found)

        df_unique_id = hm.unique(df_landfall['ID'].tolist())
        id_count = len(df_unique_id)
        self.assertEqual(land_count, id_count)


    #Test to determine that all values in the no_landfall dataframe 'Event' column do NOT have a value of 'L'
    def testNoLandFallCols(self):
        df = hm.loadData()
        df_no_landfall = hm.hurricaneNoLandFall(df)
        found = df_no_landfall[df_no_landfall['Event'].str.contains('L', na=False)]
        land_count = len(found)

        self.assertEqual(land_count, 0)

if __name__ == '__main__':
    unittest.main()
```

VI. Beyond Specifications

An aspect of our project that we experimented with was the visualization of our data. We incorporated several additional libraries to incorporate more complex graphs and plots. For example, we used the colorspacious, folium, and seaborn packages to enhance our visualizations.

The translation of our initial data into something that can be visualized on maps involved more intensive cleaning and processing; as the use of coordinate positions required well-cleaned and precise data. Calculation of distance and landspeed-related variables also required further processing of our initial data to transform positional variables into distances.

VII. Conclusion

In summary, there were 165 Category 1 hurricanes in the Atlantic basin from 1950-2015 and 22 Category 5 hurricanes (Figure 1). Figure 2 reveals a period of time from the mid-1960s through the mid-1980s where we saw more tropical cyclone generation over previous years (or at least they were better detected due to advances in technology). Tropical cyclones that did not reach hurricane strength tended to have a shorter life-span than storms that reached hurricane strength (Figure 3). Heat maps of land fallen tropical cyclones showed an area of greatest activity on the eastern coast of Mexico, southern Florida, Bahamas and North Carolina (Figure 4). Storms that did not make landfall had a less distinguishable pattern (Figure 5). In comparing the longest duration storms (Figure 6) to the shortest duration storms (Figure 7), clearly the strongest storms lasted the longest.

Furthermore, our data set revealed that Hurricane Wilma (2005) had the lowest pressure on record for the Atlantic Basin from 1950-2015 at 882.0 mb. Even more interesting is that the year 2005 contained three storms that were in the top 10 most intense storms during our data period (Wilma, Rita, and Katrina).

Storm ID/Year	***Name***	***Min Pressure(mb)***
AL252005	WILMA	882.0
AL081988	GILBERT	888.0
AL182005	RITA	895.0
AL041980	ALLEN	899.0
AL091969	CAMILLE	900.0
AL122005	KATRINA	902.0
AL042007	DEAN	905.0
AL131998	MITCH	905.0
AL092004	IVAN	910.0
AL101955	JANET	914.0

Similarly for maximum wind (kts) data, Hurricane Wilma (2005) is second only to Hurricane Allen (1980), with Wilma, Rita, and Katrina all in the top 10 cases.

Storm ID/Year	***Name***	***Max Wind(kts)***
AL041980	ALLEN	165.0
AL252005	WILMA	160.0
AL081988	GILBERT	160.0
AL182005	RITA	155.0
AL131998	MITCH	155.0
AL091969	CAMILLE	150.0
AL062007	FELIX	150.0
AL091979	DAVID	150.0
AL122005	KATRINA	150.0
AL051977	ANITA	150.0

Citations:

Landsea, C. W., 2007: Counting Atlantic tropical cyclones back in time. Eos, Trans. Amer. Geophys. Union, 88 (18), 197–203.

——, and Coauthors, 2004: The Atlantic hurricane database reanalysis project: Documentation for the 1851–1910 alterations and additions to the HURDAT database. Hurricanes and Typhoons: Past, Present and Future, R. J. Murname and K.-B. Liu, Eds., Columbia University Press, 177–221.