Gulf Coast Area Rainfall Prediction

By Jimmy Smart

Background

Our livelihood is so greatly affected by the weather. Depending on the region someone lives in usually determines the type of weather patterns that they'll see on a yearly, monthly or even daily basis. A large majority of industries greatly depend on their region's weather and climate. For instance, farmers need a good mixture of rain and sunlight so that their crops may grow. Too much sun but not enough rain, may cause issues to their crop yields. Just as too much rain and not enough sun can affect their crop yields as well. Seasonal temperatures factor in as well, as a majority of crops grow based on the daily temperatures. Some may grow better in months that are typically known to have cooler temperatures, while other crops prefer to grow in the warmer months.

Commercial fishing also keeps a keen eye on weather patterns as life found in the Gulf usually have their own growing seasons as well as seasons for farming.

For farmers, commercial fishing and various other industries who depend on their regional climates, their best approach in co-existing with weather patterns is by using past trends to create future predictions.

The Problem

History is known to repeat itself at times but predicting if/when history will repeat itself is an ongoing issue. The best approach is look for patterns of the past while comparing it to patterns of the present. The goal of this project is: Can we predict rainfall for the Gulf Coast area?

Weather has shown signs of yearly and seasonal patterns but it's not always set in stone. Just like other factors in our world, weather patterns are changing as well. Global warming has become a hot debate in our culture today as more and more people feel that the certain factors in our environments are raising temperatures around the world which in turn are also raising the temperatures of our oceans, seas and gulfs. This project isn't necessarily a global warming, project, but as we go through our historical weather data and try to predict future weather patterns, we will be able to see if there has been a general uptick in temps and see if there is a correlation with any effects it's having on our data over the years.

Data

This project will build a model that focuses on a dataset consisting of decades of daily weather from various locations within our Gulf Coast area. Where we're looking at consists of an area in the southern part of the United States that shares a coastline on the Gulf of Mexico. Our data is from https://ww.ncdc.noaa.gov/ and consists of 5 different datasets from each of our 5 chosen cities along the Gulf Coast. Each dataset consists of daily weather data and is saved as a CSV file.

Cities used for our Dataset

Predictor	Description
New_Orleans.csv	dataset of daily New Orleans, LA weather 1/1/2010-12/31/2019
Houston.csv	dataset of daily Houston, TX weather 1/1/2010-12/31/2019
Pascagoula.csv	dataset of daily Pascagoula, MS weather 1/1/2010-12/31/2019
Mobile.csv	dataset of daily Mobile, AL weather 1/1/2010-12/31/2019
Tampa.csv	dataset of daily Tampa, FL weather 1/1/2010-12/31/2019

Terms and definitions

Dewpoint Average Temperature- temperature where water vapor starts to condense out of the air (the temperature at which air becomes completely saturated). Above this temperature the moisture stays in the air. If the dew-point temperature is close to the dry air temperature - the relative humidity is high If the dew point is well below the dry air temperature - the relative humidity is low

Humidity(%)- concentration of water vapor present in the air. Water vapor, the gaseous state of water. Humidity indicates the likelihood for precipitation, dew, or fog to be present. The amount of water vapor needed to achieve saturation increases as the temperature increases.

Sealevel_pressure(Hg)- atmospheric pressure at sea level at a given location. When observed at a reporting station that is not at sea level (nearly all stations), it is a correction of the station pressure to sea level. This correction takes into account the standard variation of pressure with height and the influence of temperature variations with height on the pressure. The temperature used in the sea level correction is a twelve hour mean, eliminating diurnal effects.

Data Wrangling and Cleaning

Our 5 datasets were given a quick cleaning within their original Excel format before being saved as CSV files and read into our notebook as one through glob. Concat was used to create one dataframe.

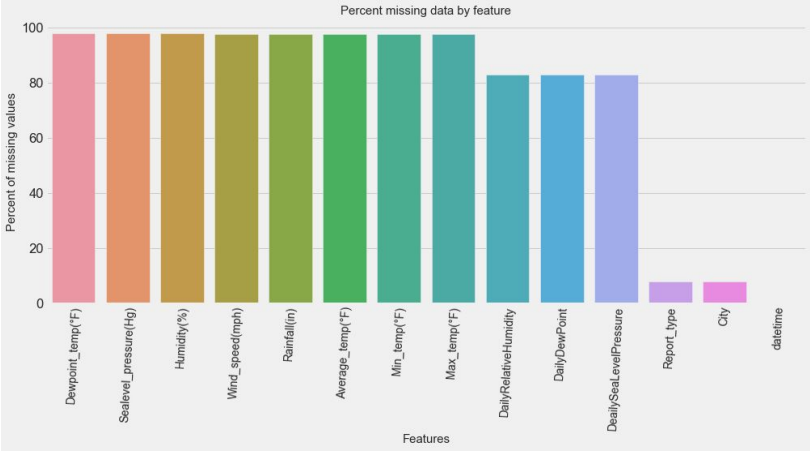
Our initial exploratory data analysis (EDA) consisted of looking at the shape of our dataframe (699929, 141). Our datetime value was converted into a readable datetime object through to_datetime.

Since our dataset was long and consisted of values that we were not going to be using we focused on the following values:

```
df = df[[
'STATION',
        'datetime',
        'REPORT_TYPE',
        'DailyDewPoint',
        'DailyRelativeHumidity',
        'DeailySeaLevelPressure',
        'DailyAverageDewPointTemperature',
        'DailyAverageRelativeHumidity',
        'DailyAverageSeaLevelPressure',
        'DailyMaximumDryBulbTemperature',
        'DailyMinimumDryBulbTemperature',
        'DailyAverageDryBulbTemperature',
        'DailyPeakWindSpeed',
        'DailyPrecipitation'
       ]]
```

We then renamed the values to more column friendly names.

Missing Data



A values had a large percentage of null values due to the structure of the dataset having hourly, daily, monthly data so we chose to focus on only the daily data figures by using .loc to pull out rows that only consisted of daily data.

From here we check again for an updated percentage of missing data before properly filling in any missing data.

We replaced any mislabeled data before converting our values to float. We checked for any outstanding outliers and made the appropriate changes.

Creating New Timeseries Columns

Time plays an important role in our project. Our dataset consists of data compiled from over 10 years with each year being able to be broken down into months, weeks, days and years. We created a new dataframe that represented these timestamps.

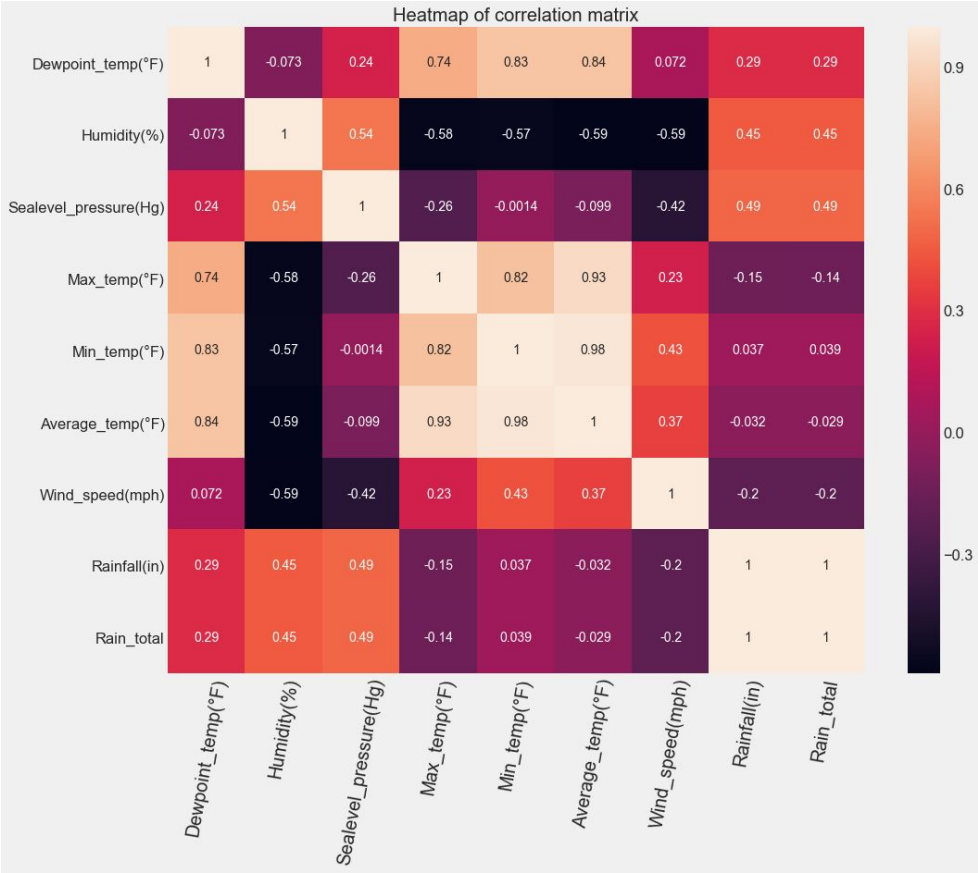
Data Visualization

With our EDA completed we were then able to start visualizing our data better. Starting with various groupby dataframes that represented various aspects such as focusing on individual cities and being able to see the yearly and or monthly means/sums of our values.

Various groupby dataframes showed us trends by years or months or seasonally. We could see if certain months had more less rainfall occurrences as well as seeing if certain years had more rainfall or hotter/colder temperatures than other years or other cities.

Heatmap Correlation

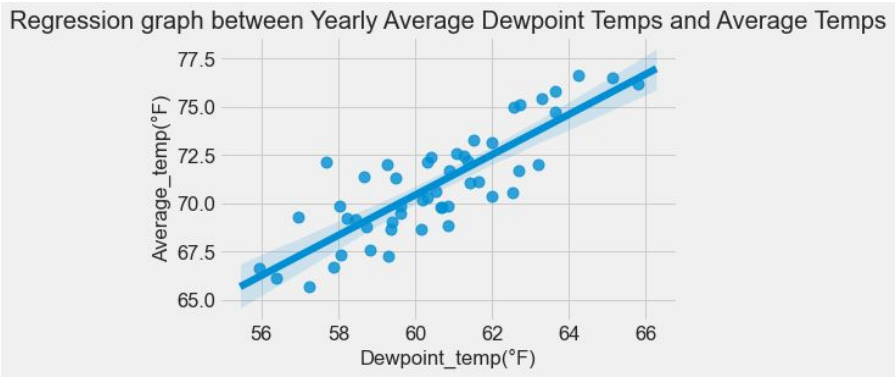
Using one of the dataframe that we created with our groupby. we can take a look at the overall correlations within our dataset.



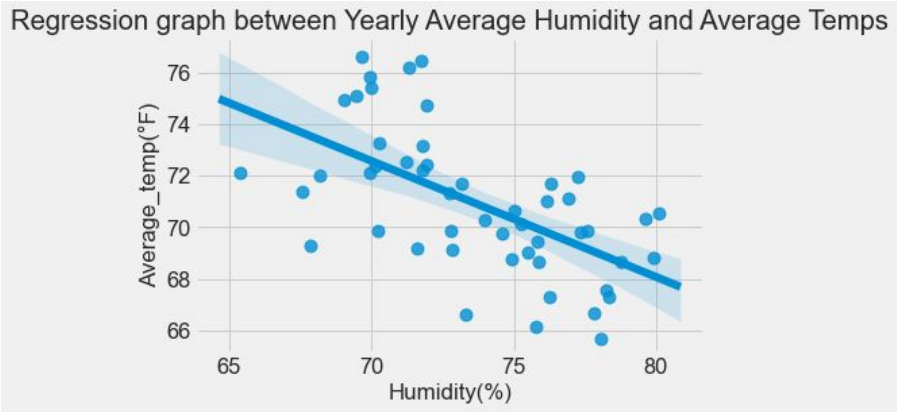
From this heatmap we can see that Dewpoint and our max, min, average daily temperatures have a positive correlation. While Humidity has a negative correlation with our temperature variables.

Rainfall has decent correlations with Humidity(0.31) and Sealevel Pressure (0.49) which makes sense because as a storm rolls in humidity usually rises as does sealevel pressures.

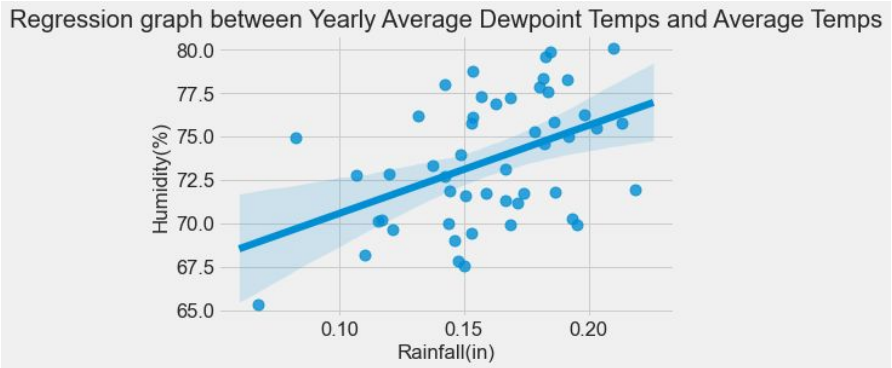
Below we will use regression plots to show these variables.



As our Dewpoint temperatures rose, so did our average temperatures.

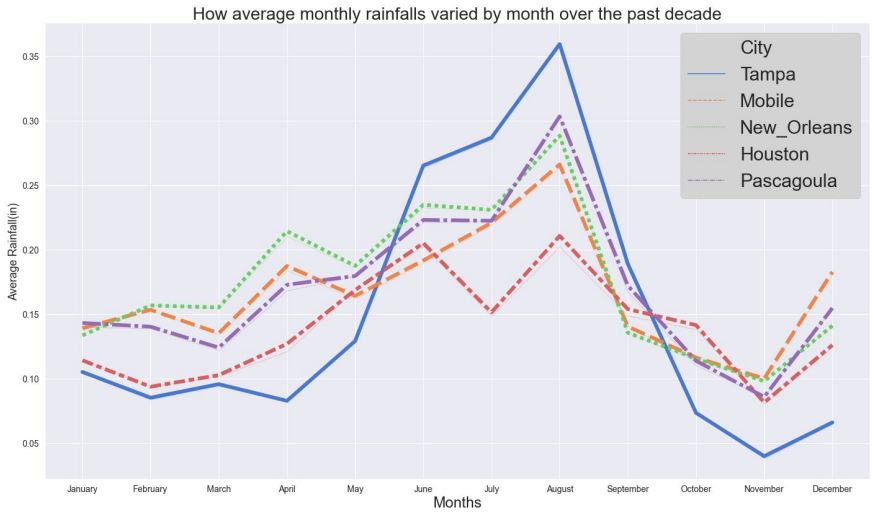


As our Humidity percentages rose our average temperatures decreased.

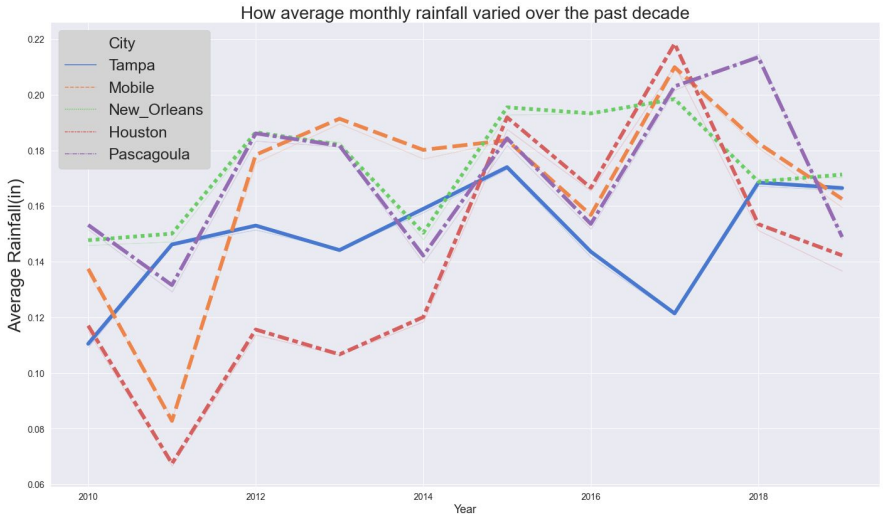


As our rainfall increased so did our humidity percentages.

Rainfall

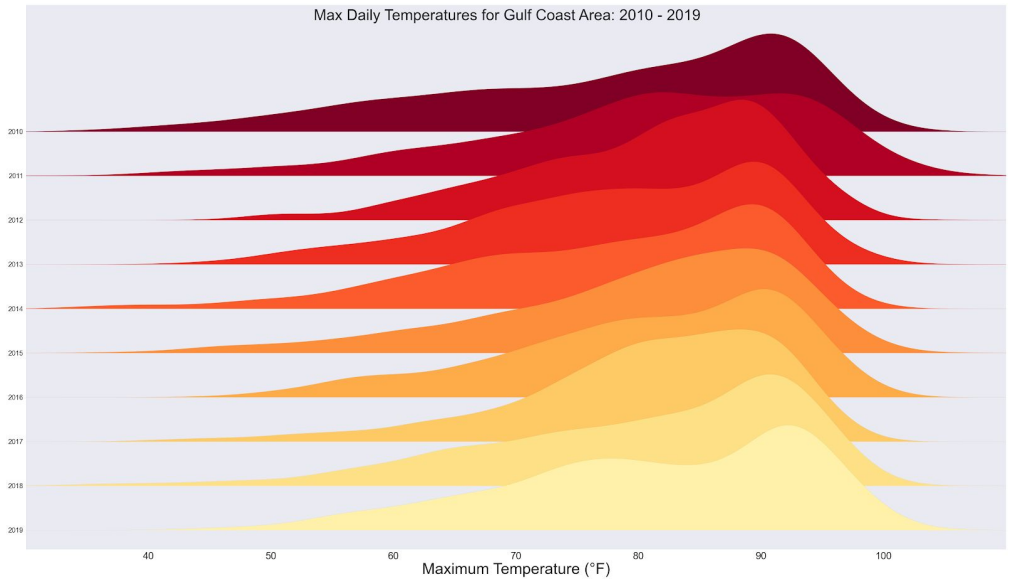


A look at how average monthly rainfalls varied by months for each of our cities. We can see that Tampa had the highest and lowest totals amongst our 5 cities. We can also see that the summer months have the highest average rainfalls of the year. Considering hurricanes and tropical storms are known to affect this area of the country and that hurricane season is known to be between early June and late November, this may also be a factor.



Our yearly averages for rainfall amongst our cities tells a little bit of a different story than the monthly averages. Tampa has stayed someone consistent throughout the decade. They had some highs and lows but Houston and Mobile show more extreme averages. Lows in 2011 and highs in 2017.

Temperatures



Here we can see how our average temperatures have faired over the months of our dataset. Summer months are typically known to be hot in this area with a bit of a break from the heat in the other months of the year.

Year	Max_temp(°F):
2010	78.081644
2011	80.559648
2012	80.828415
2013	78.713029
2014	78.003297
2015	80.202192
2016	80.704372
2017	80.885479
2018	80.456438
2019	81.356164

As we can see, there has been an uptick in average max temperatures in the Gulf Coast Area. Gaining almost 3 degrees hotter over the past decade. At least in the Gulf Coast area there is a warming trend occurring.

Statistical Analysis

Now that we've spent some time on data wrangling and creating some visuals to better see what our dataset has, we can dive deeper into the statistical information.

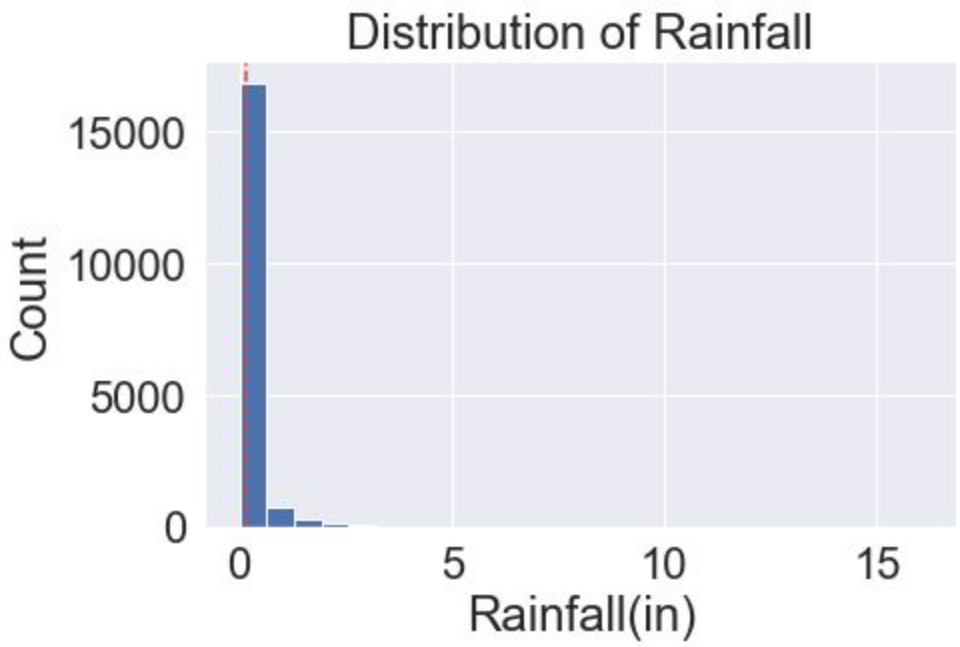
Null Hypothesis: Can we predict rain patterns for years to come based on past data from the Gulf Coast Area.

Alternate Hypothesis: We will not be able to predict future rain patterns based off of past weather data for the Gulf Coast Area.

Frequentism

We will focus on the Rainfall variable within our dataset.

mean: 0.15989694677410646
Std: 0.15989694677410646



We have a probability value of (1) and a critical value of (**1.96**) and confidence intervals of (**0.1434** and **0.1763**)

We separated our rainfall by determining if the day had rain or no rain.

```
df_rain = df_time['Rainfall(in)'].loc[df_time['Rain or Not'] >= 1]
df_no_rain = df_time['Rainfall(in)'].loc[df_time['Rain or Not'] == 0]
```

Results:

Number of Rainy Days: 5695
Number of non Rainy Days: 12548
Mean of Rainy Days: 0.5122036874451316
Standard Deviation of Rainy Days: 0.834636440830481

To calculate the **t-test**, 1st we went with a manual approach by calculating the value of the test statistic and then its probability (the p-value). We also used ttest_ind then verified that we got the same results from both.

T-test: **68.74592164509636**
Ttest_ind Result (statistic= **68.73988574169559**, pvalue= **0.0**)

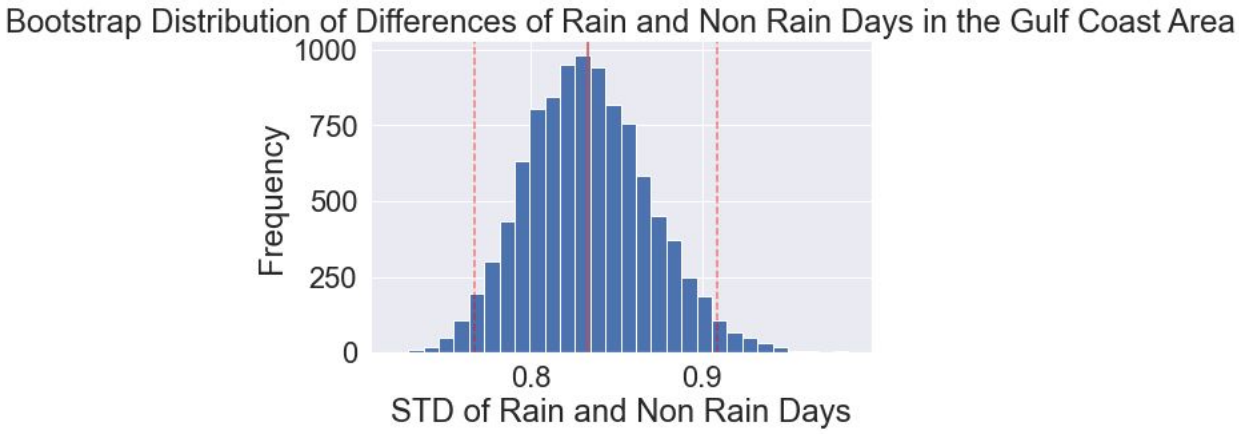
When we set the equal_var to False we got the result:

Ttest_indResult (statistic= **46.3077936849106**, pvalue= **0.0**)

We were able to obtain the same t-test results and noticed a slight change when set to False.

Bootstrapping

We'll create a bootstrap sampling to estimate a 95% confidence interval lower limit.



The observed difference of standard deviations falls within the 95% Confidence Interval.

Difference of STD for rainfall and non rainfall days: **0.834636440830481**
Difference of STD for bootstrap samples: **0.8332447469738007**
The 95% confidence interval for the difference between the standard deviations of rainfall and non rainfall days is: [**0.76690786 0.90884574**]

Confidence interval and p-value

We'll perform a bootstrapped hypothesis test at the 5% significance level (α =0.05) to calculate the p-value of the observed difference between rainfall days and non rainfall days.

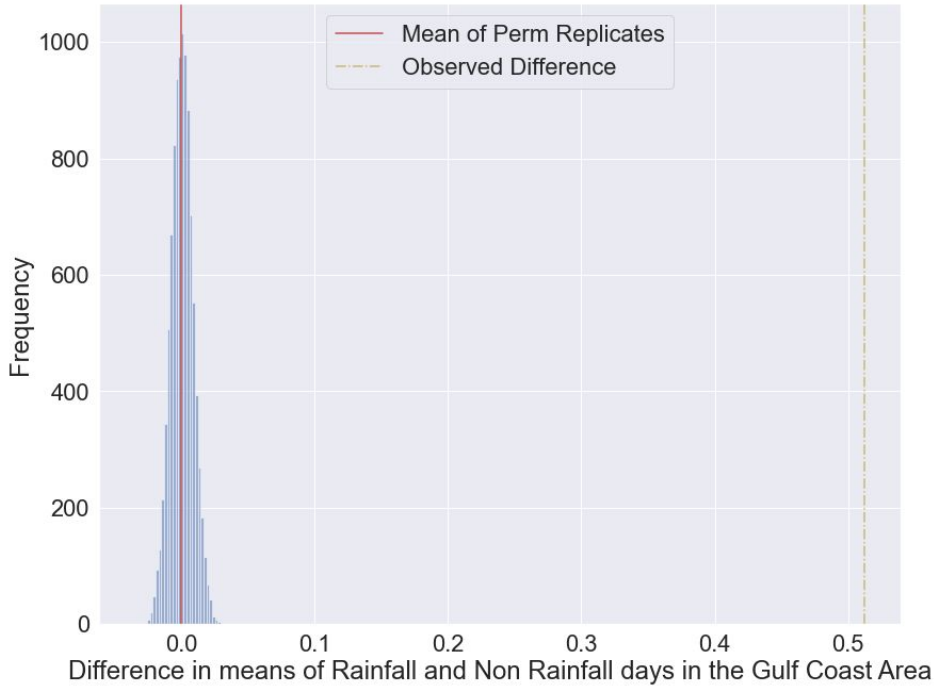
```
def diff_of_means(data_1, data_2):
    diff = np.mean(data_1) - np.mean(data_2)
    return diff

# Compute difference of mean rainfall and non rainfall: empirical_diff_means
empirical_diff_means = diff_of_means(rainfall, no_rainfall)

# Draw 10,000 permutation replicates: perm_replicates
perm_replicates = draw_perm_reps(rainfall, no_rainfall, diff_of_means,
size=N_rep)

# Compute permutation p-value: perm_p
perm_p = np.sum(perm_replicates >= empirical_diff_means) /
len(perm_replicates)

#Compute p-value: p
p = np.sum(bs_perm_rep >= diff_means)/len(bs_perm_rep)
```

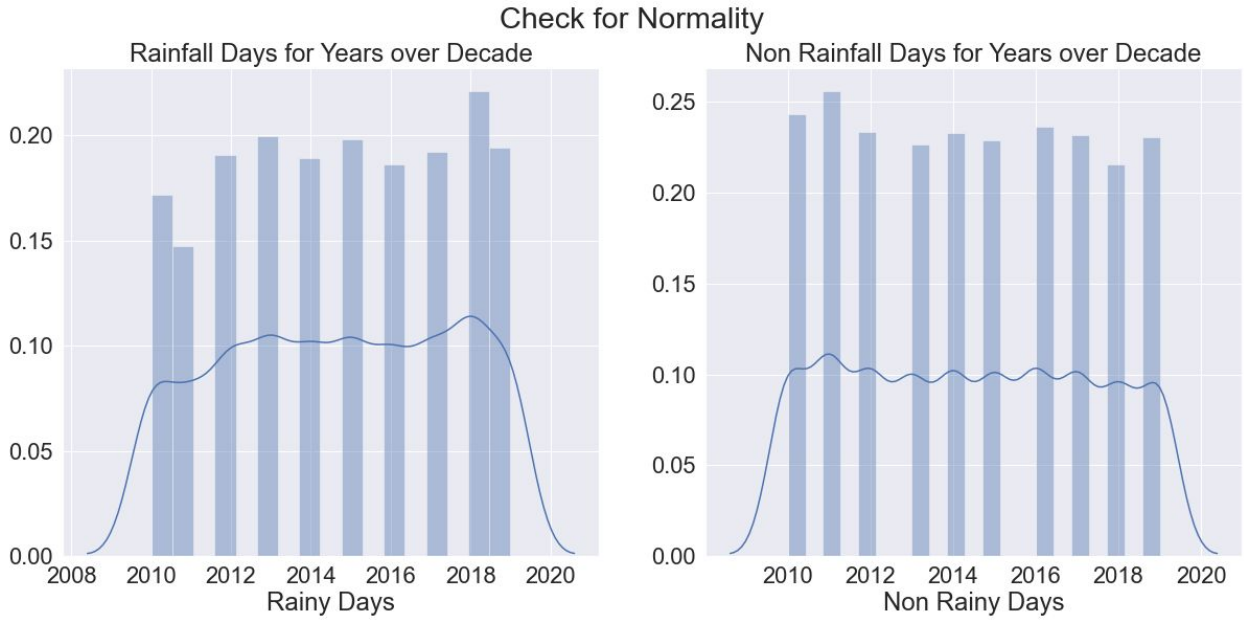
Perm p-value = 0.0
P-value= 0.7383

Null Hypothesis: Can we predict rain patterns for years to come based on past data from the Gulf Coast Area.

Alternate Hypothesis: We will not be able to predict future rain patterns based off of past weather data for the Gulf Coast Area.

With the p-value from the bootstrap being 0.7383, I can not reject the null hypothesis.





Summary

Through our data wrangling efforts and visualizations we were able to see weather trends than can be usual to serve any clients looking to see how weather has been and may be in the future.

When looking at temperatures over the past decade we were able to see that there has been an upward trend towards average temperatures rising each year. A degree hotter might not seem like much but the overall story can show devastating results if proper precautions aren't met. Farmers may need to adjust their watering and feeding habits to make sure their crops and/or animals can adapt to rising temperatures.

As for rainfall, we were able to see trends over the past decade. For the most part, the area hasn't had to deal with any long lasting droughts. Being able to have a visual of upcoming trends can help various industries to adjust their habits based on the time of year and expected rainfall or non rainfall.

Statistically, we were able to come up with our hypothesis and ultimately go with our alternate hypothesis. We knew going into this project that predicting weather isn't an easy task. People have been trying to crack it for centuries and will continue doing so. There are multiple variables in play and looking farther away to get an idea of what weather may be be coming towards the area. Our inferential statistics did give us inside and we may need to add more features in the future to better our results.

As we continue on with our project, we'll use machine learning to try and predict what the future holds for the Gulf Coast area in regards to rainfall. Knowing if the upcoming months or years will be rainy or mostly dry will help them better plan now for what's to come.