Gulf Coast Area Rainfall Prediction

Jimmy Smart

Project Outline

- Introduction
- Data Wrangling
  - Dropping redundant variables
  - Identify feature values
- Data Visualization
  - Rainfall
  - Temperatures
  - Summary
- Statistical Analysis
  - Frequentism
  - Bootstrapping
- Model Optimization
  - FB Prophet
  - Linear Regression
  - XGBoost
  - Random Forest
- Comparing models
- Conclusion

Introduction

Q:

Can we predict rainfall for the Gulf Coast Area?

- Dataset contains historical weather data of five cities along the Gulf Coast shores.

- Our data is from https://ww.ncdc.noaa.gov/

- The original datasets for all 5 cities *New_Orleans.csv*, *Houston.csv, Pascagoula.csv, Mobile.csv* and *Tampa.csv* were loaded in and saved as 1 dataset into a Jupyter Notebook.

- The goal of this project is to use the features, from the datasets listed above, to predict daily rainfall amounts based on past rainfall
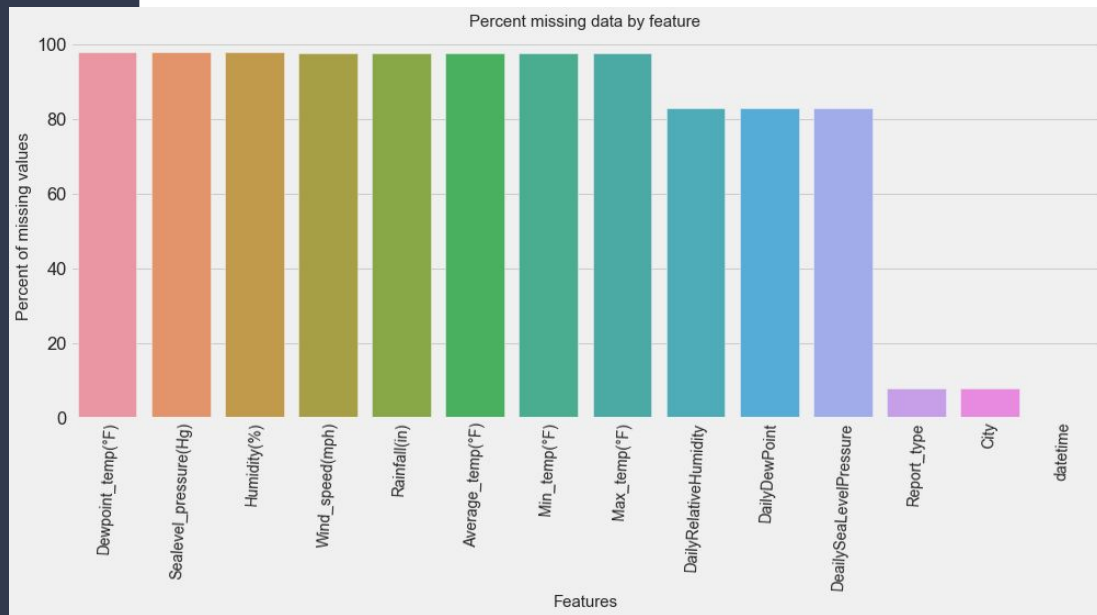
Featured values from dataset

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.

Featured values:

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

Identifying missing data


Percent missing data by feature

- Our featured values had a large percentage of null values due to the structure of the dataset having hourly, daily, monthly data.
- Focus only on the daily data figures by using .loc to pull out rows that only consisted of daily data.

	null_sum	null_pct	dtypes	count	mean	median	std	min	max
Average_temp	0	0.0	float64	18243	70.944581	74.000000	12.746689	26	95
City	0	0.0	object	18243	NaN	NaN	NaN	Houston	Tampa
Dewpoint	0	0.0	float64	18243	60.499926	64.500000	13.934829	4.70833	80
Humidity	0	0.0	float64	18243	73.628929	75.000000	11.397478	21	100
Max_temp	0	0.0	float64	18243	79.980266	82.000000	12.198700	31	109
Min_temp	0	0.0	float64	18243	61.407170	65.000000	14.031359	16	85
Rainfall(in)	0	0.0	object	18243	0.159897	0.000000	0.523273	0	16.07
Sealevel_press	0	0.0	float64	18243	30.049150	30.035833	0.140398	29.44	30.6954
Wind_speed	0	0.0	object	18243	23.381571	22.000000	37.303786	0	2237
datetime	0	0.0	datetime64	18243	NaN	NaN	NaN	2010-01-01	2019-12-31

EDA Analysis

Heatmap of correlation matrix
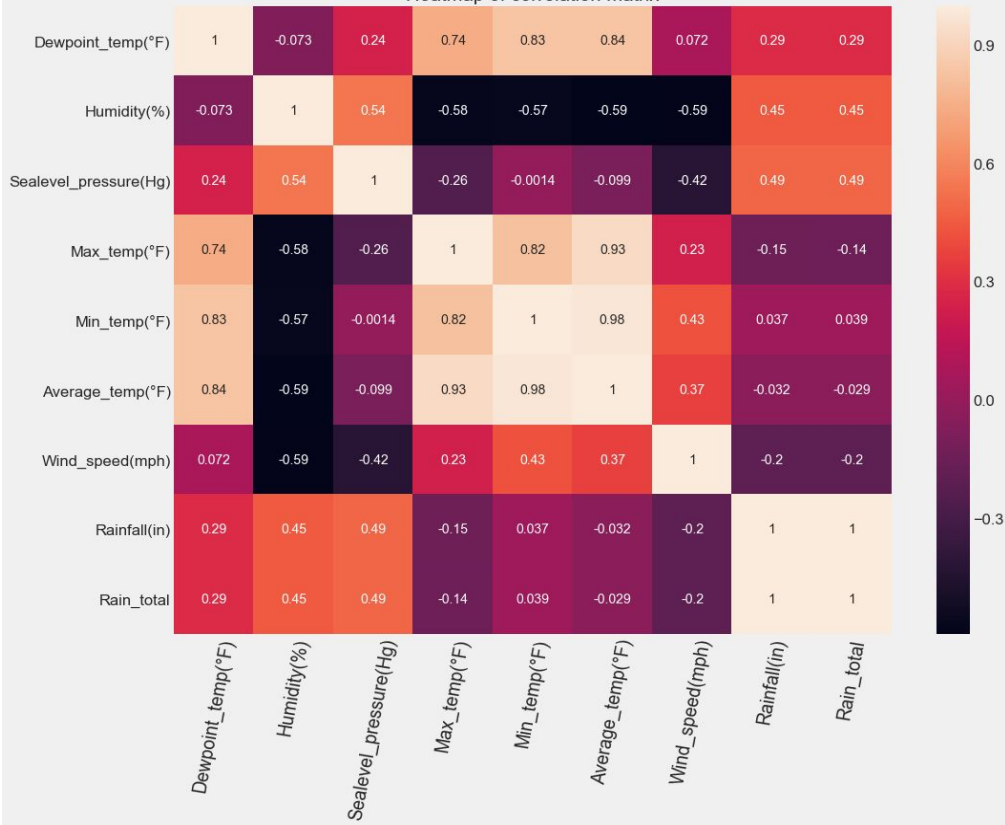
- From this heatmap we can see that Dewpoint and our max, min, average daily temperatures have a positive correlation.

- 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.
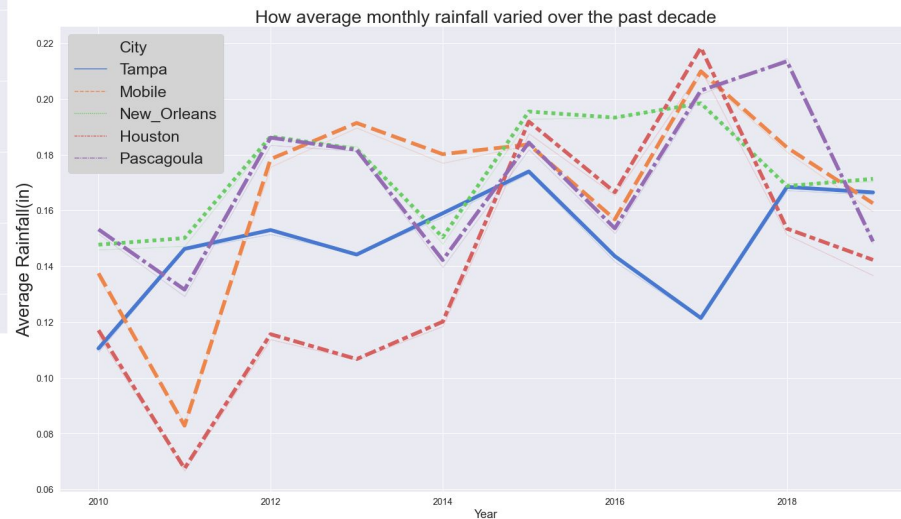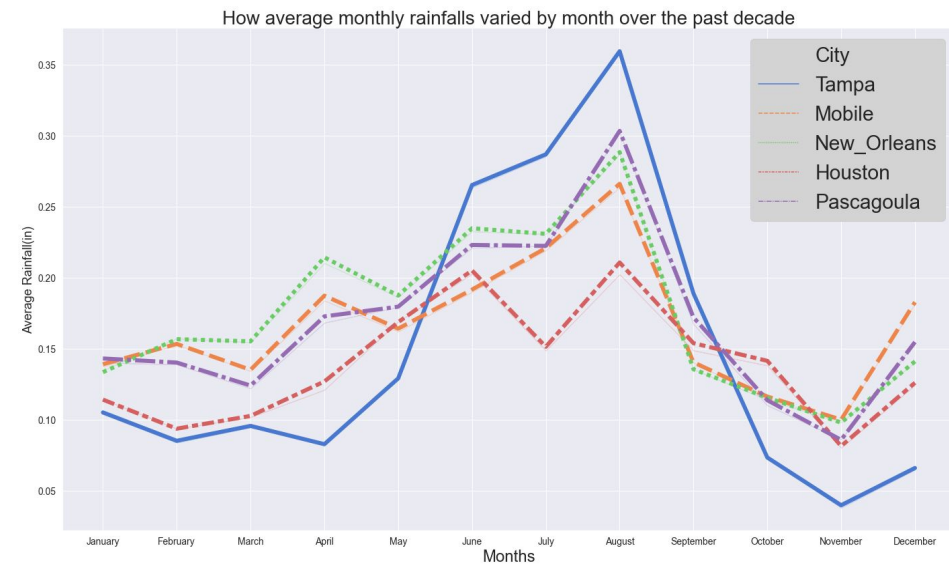
Data Visualization

City	Dewpoint(°F)	Humidity(%)	Sealevel_pressure(Hg)	Max_temp(°F)	Min_temp(°F)	Avg_temp(°F)	Wind_speed(mph)	Rainfall(in)	Rain_total
Houston	59.784873	72.047958	30.021814	80.700192	60.899699	71.047136	24.181694	0.139984	510.80
Mobile	59.931493	77.123816	30.064963	78.147244	59.645736	69.145873	22.132712	0.166597	607.58
New Orleans	60.473006	71.836942	30.054149	79.387229	62.785969	71.346396	24.034804	0.174412	636.43
Pascagoula	59.076459	76.829816	30.054634	78.810907	56.954782	68.133461	20.773089	0.169819	619.67
Tampa	63.233489	70.308030	30.050197	82.854755	66.748698	75.049055	22.719649	0.148676	542.52

- Yearly averages of our features separated by our 5 cities
- Total rainfall amounts for the past decade

Rainfall

(left) 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.
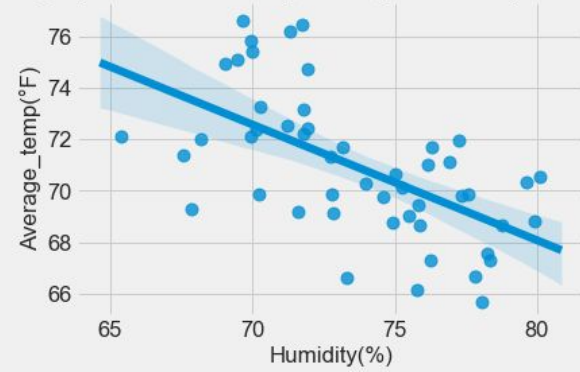
(right) yearly averages for rainfall amongst our cities tells a little bit of a different story than the monthly averages.
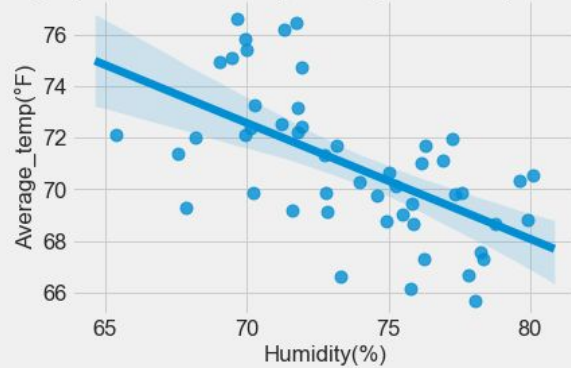

How average monthly rainfalls varied by month over the past decade

How average monthly rainfall varied over the past decade
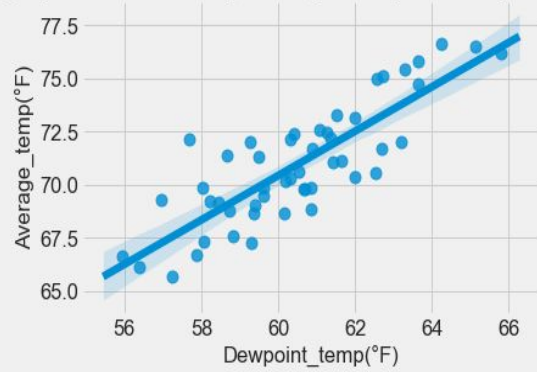
(top left) As our Dewpoint temperatures rose, so did our average temperatures. (top right) As our Humidity percentages rose our average temperatures decreased. (bottom center) As our rainfall increased so did our humidity percentages.

Rain or No Rain based on Dewpoint Temperatures
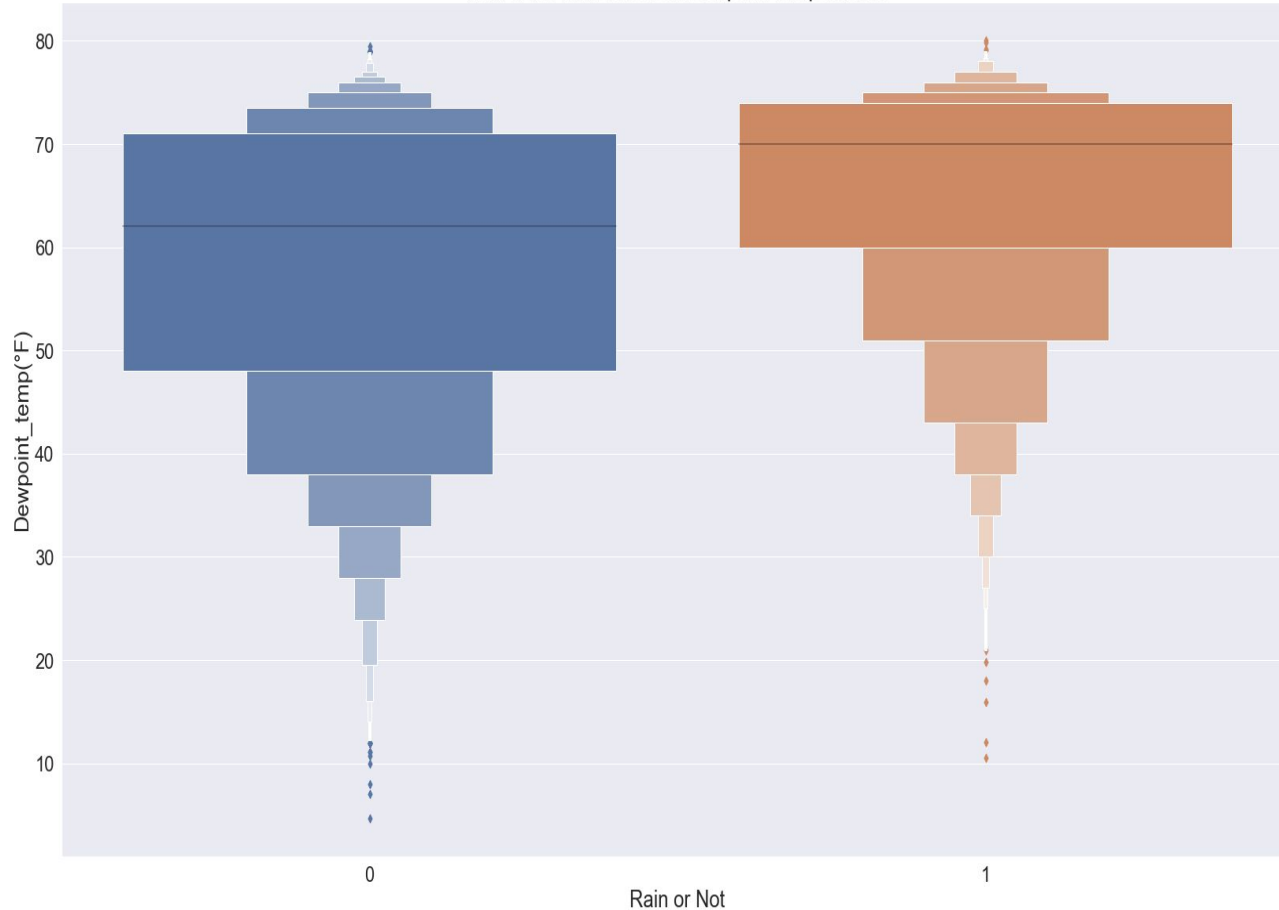
- We created a new feature (Rain-1 or Not-0)

- As we can see from the figure to the left, on days that we had rain, our average Dewpoint temperatures were higher than days without rain.

- Our non rainy days also showed lower falling outliers than on rainy days

Max Daily Temperatures for Gulf Coast Area: 2010 - 2019
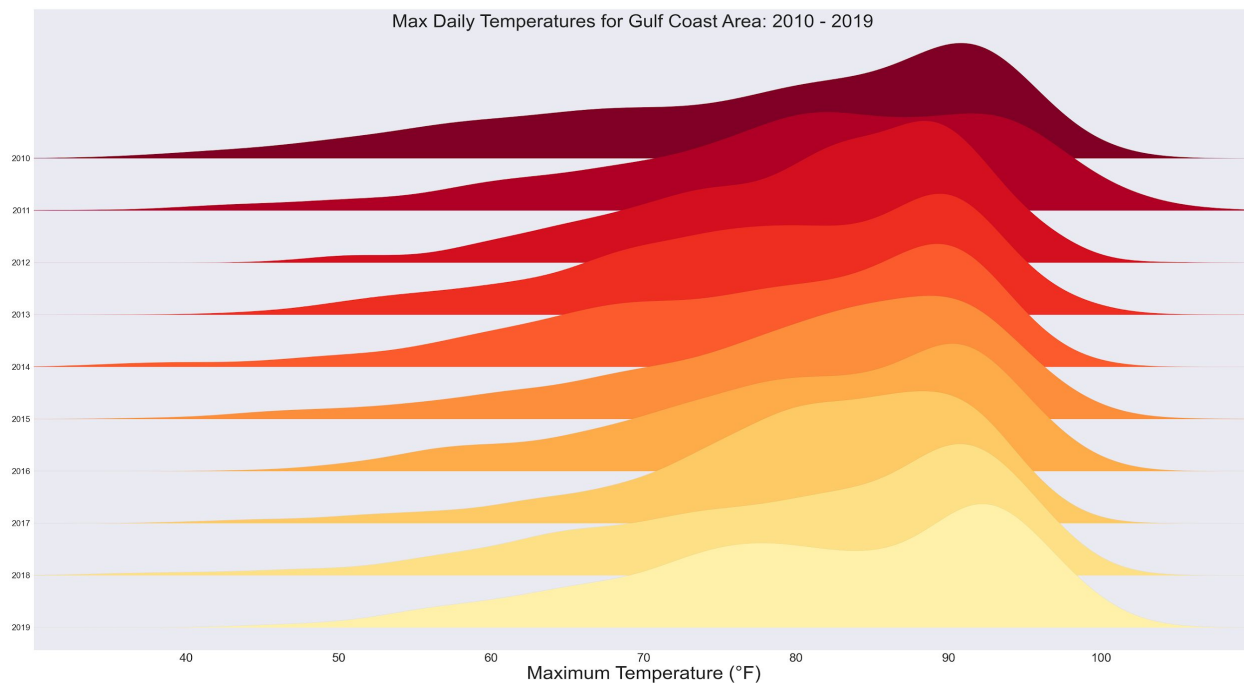
Max_temp(°F):	
Year	
2010	78.081644
2011	80.559648
2012	80.828415
2013	78.713029
2014	78.003297
2015	80.202192
2016	80.704372
2017	80.885479
2018	80.456438
2019	81.356164

Here we can see how our average temperatures have fared 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.

Data Visualization Summary

- We were able to see weather trends that can be useful to any clients looking to see how weather has been and may be in the future.

- Rainfall amounts fluctuate during parts of the season but look to be holding steady overall

- 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.

Bootstrapping

Bootstrap sampling to estimate a 95% confidence interval lower limit.

**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.

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 day is: *[0.76690786 & 0.90884574]*

Bootstrap Distribution of Differences of Rain and Non Rain Days in the Gulf Coast Area
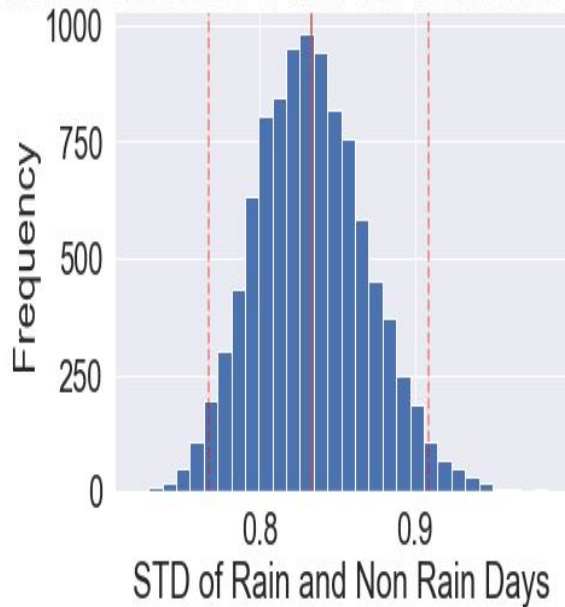

Performed our test at 5% significance level to calculate p-value between rainfall days and non rainfall days

```
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

With a p-value of 0.7383, I can reject the alternate hypothesis and accept the null hypothesis

Rainfall(in) in Gulf Coast Area

Normal dist. (mu = 0.10 and sigma = 0.26)

Probability Plot
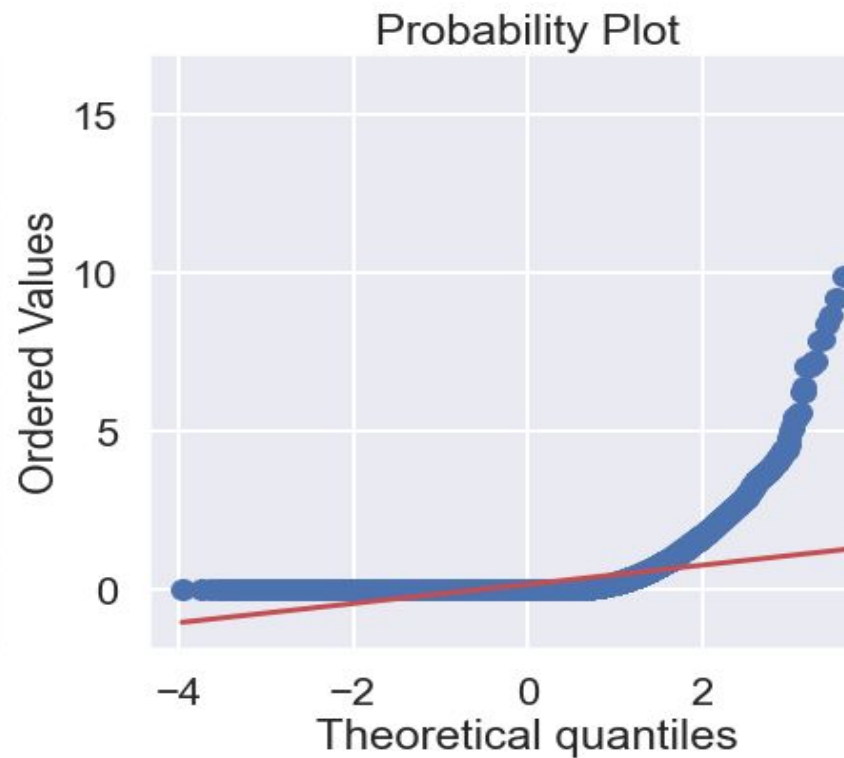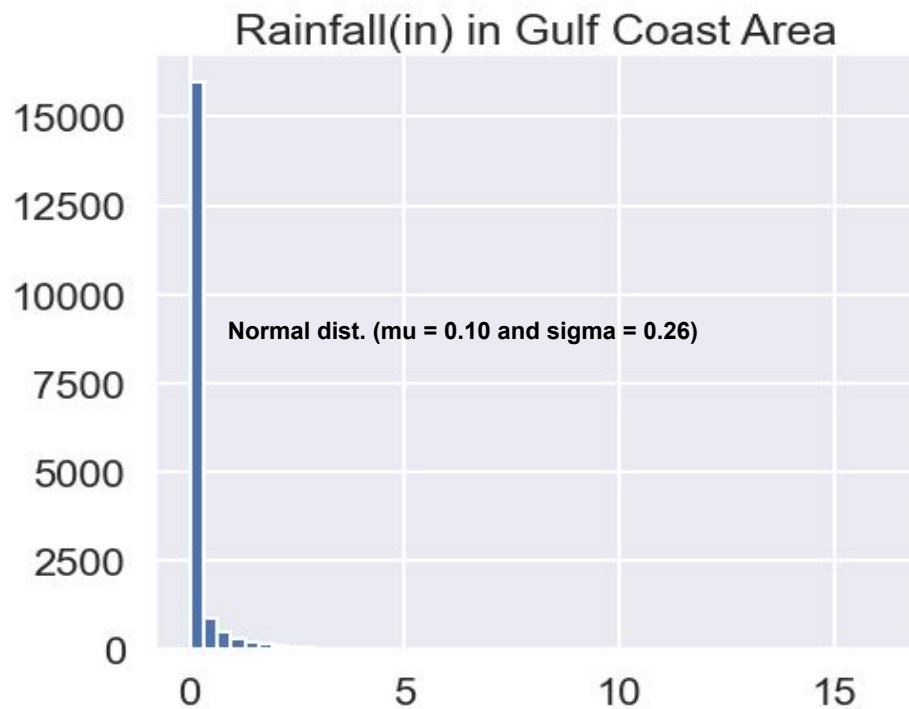
	count	mean	std	min	25%	50%	75%	max
Rainfall(in)	18243.0	0.159897	0.5233	0.0	0.0	0.0	0.03	16.07

Model Optimization

- FB Prophet
- Linear Regression
- XGBoost
- Random Forest

2 datasets:
- without Rainfall outliers
- with Rainfall outliers

Comparing Models
without outliers

- Our r2 percentages fluctuate and showed an issue of overfitting with our Training set especially for XGBoost and RF

- Testing r2 relatively showed 20-30% relationship between your testing model and our dependent variable (Rainfall)

- MAE results are good as the values aren't too high which means we weren't "punished" too much for huge errors.

- RMSE results are good too considering they stayed fairly low as well

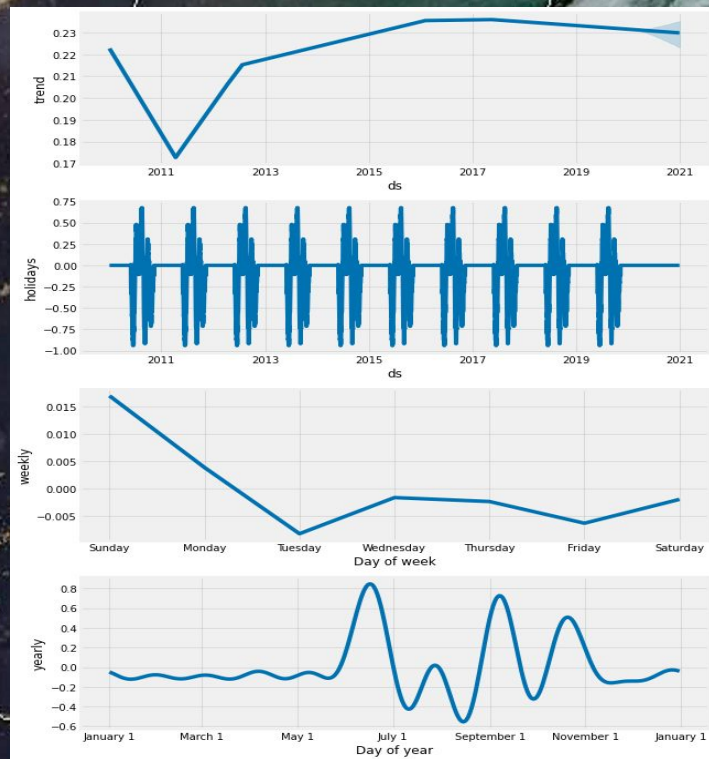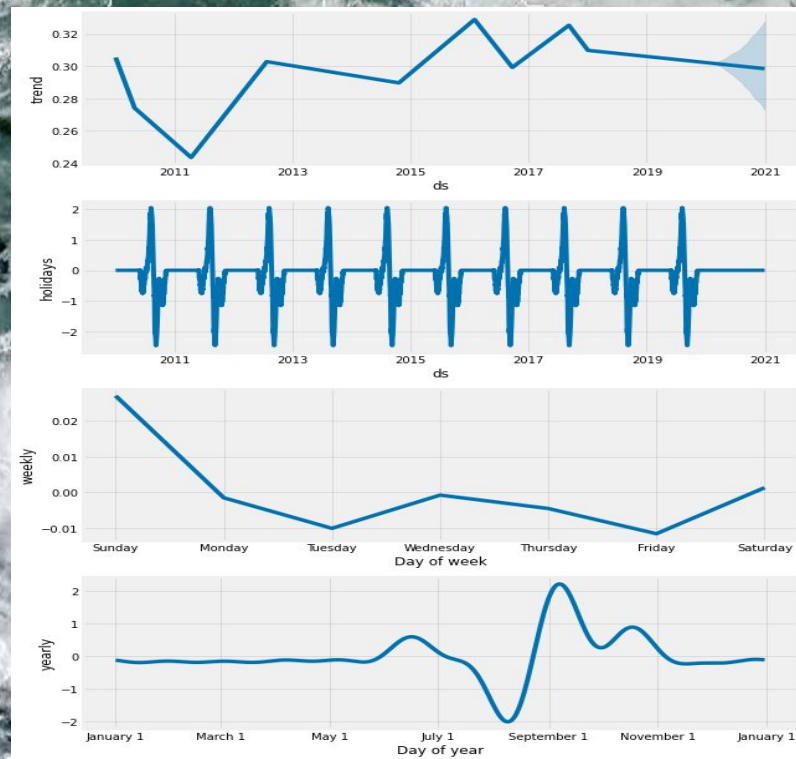| name	R2 - Test	R2 - Train	MAE - Test	MAE - Train	RMSE - Test	RSME - Train
FB Prophet	-0.001	NaN	0.248	NaN	0.468	NaN
Linear Regression(multiple)	0.223	-2.600	0.214	0.220	0.396	0.406
XGBoost	0.255	0.650	0.161	0.220	0.373	0.217
Random Forest	0.292	0.847	0.167	0.065	0.382	0.142
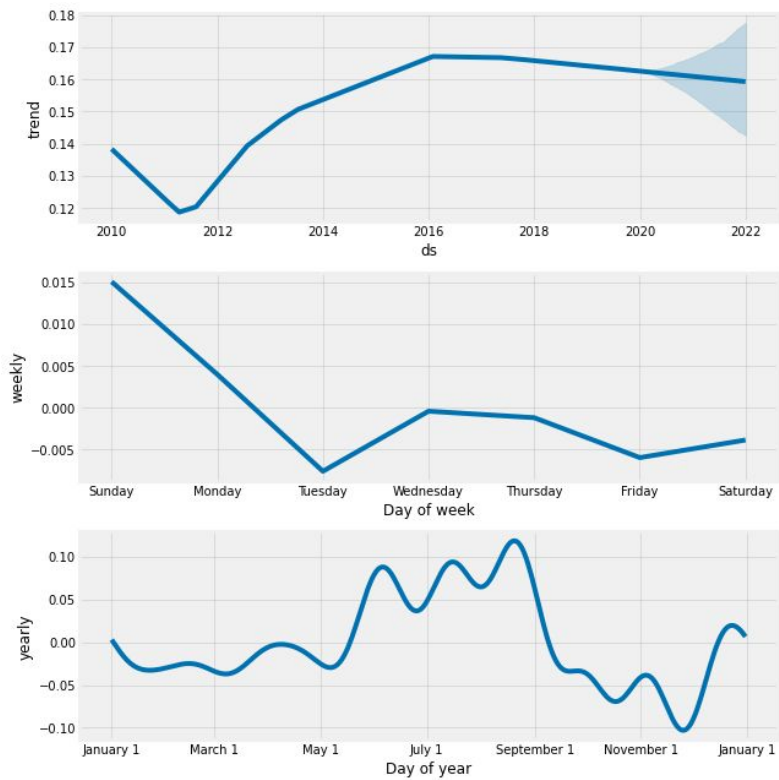
Model without outliers
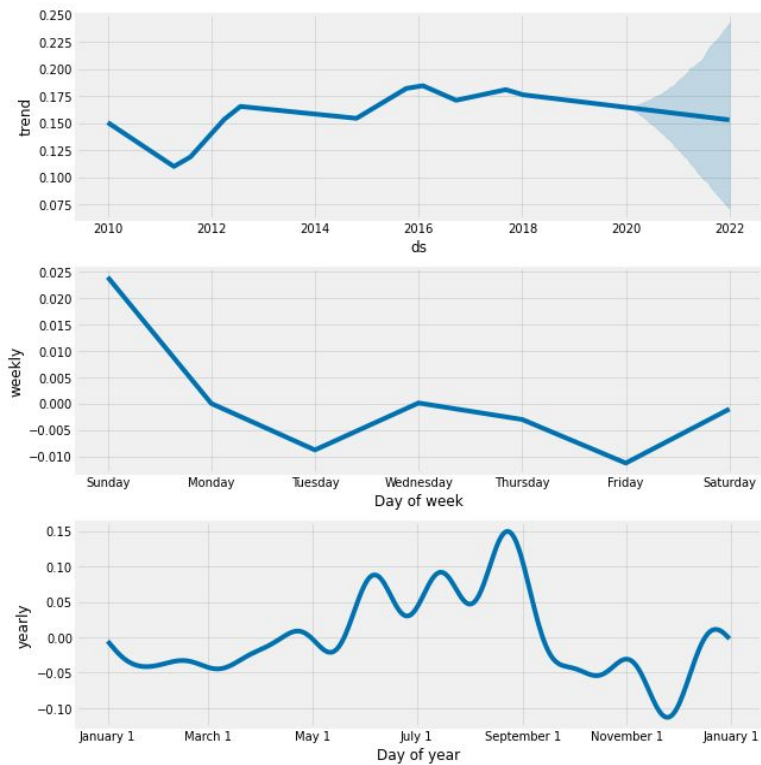
Model with outliers

FB Prophet - trends

Hurricane Season

Model without outliers

Model with outliers

Importance Features
(without outliers)

Features	FB Prophet	LR (multiple)	XGBoost	Random Forest
Sealevel_pressure(Hg)	-0.194383	-0.212812	0.104949	0.149999
Max_temp(°F)	0.007984	-0.002402	0.102125	0.096114
Average_temp(°F)	0.064095	0.009596	0.102010	0.062910
Min_temp(°F)	0.109440	0.007863	0.138275	0.088046
Dewpoint_temp(°F)	0.159427	-0.017660	0.071421	0.103797
Wind_speed(mph)	0.326622	0.019038	0.241180	0.212961
Humidity(%)	0.331100	0.018864	0.240040	0.286172

Importance Features
(with outliers)

Features	FB Prophet	LR (multiple)	XGBoost	Random Forest
Sealevel_pressure(Hg)	-0.194946	-0.228957	0.106688	0.150871
Max_temp(°F)	0.005526	0.006074	0.130203	0.098977
Average_temp(°F)	0.060142	-0.006568	0.135132	0.061892
Min_temp(°F)	0.104485	0.017008	0.122700	0.099537
Dewpoint_temp(°F)	0.150118	-0.019476	0.096356	0.104089
Wind_speed(mph)	0.311427	0.022094	0.192597	0.212199
Humidity(%)	0.321254	0.020152	0.216324	0.272435

Linear Regression Prediction
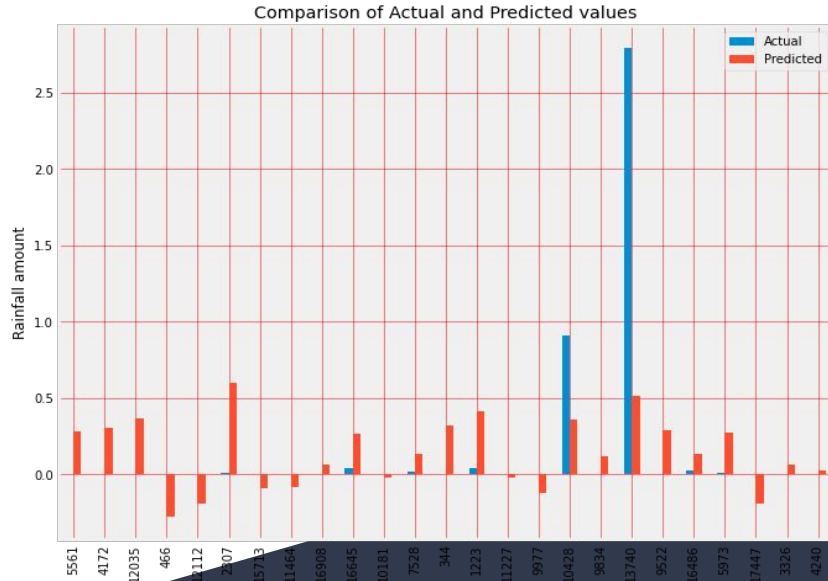
Without outliers
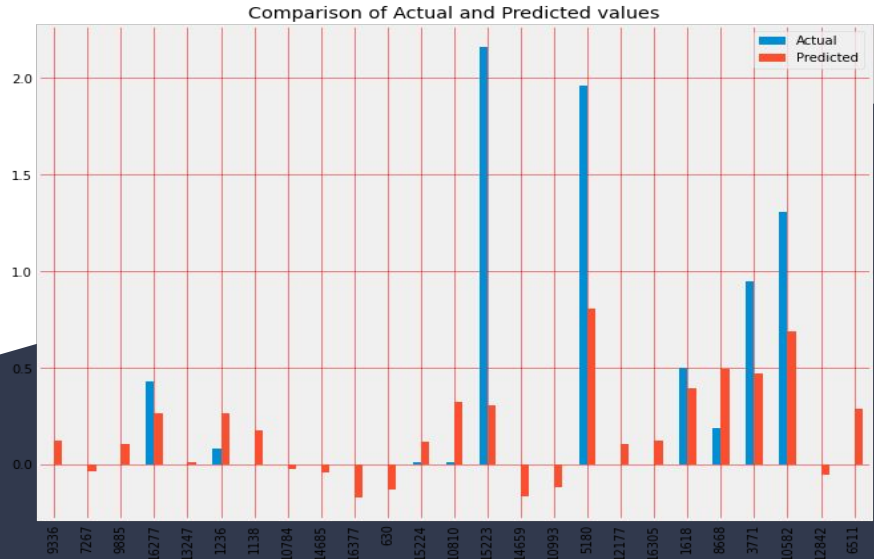
With outliers

Comparing Models
with outliers

- We didn't see anything too drastic between our 2 sets of models

- R2 scores stayed relatively near our models without outliers. XGBoost's overfitting is still high but was able to come down some.

- Again r2 relatively showed 20-30% relationship between your testing models and our dependent variable (Rainfall)

- MAE results are good as the values aren't too high which means we weren't "punished" too much for huge errors.

- RF RMSE results are good considering they are fairly low

name	R2 - Test	R2 - Train	MAE - Test	MAE - Train	RMSE - Test	RSME - Train
FB Prophet	-0.003	NaN	0.261	NaN	0.544	NaN
Linear Regression (multiple)	0.174	-2.726	0.235	0.233	0.511	0.456
XGBoost	0.346	0.277	0.167	0.233	0.409	0.318
Random Forest	0.311	0.851	0.170	0.065	0.196	0.026

Conclusion Results

Results:

- XGBoost and RF model with outliers included looks to be the models worth focusing on and trying to achieve better results with. Their respective MAE and RMSE metrics were in good ranges

- Our models did not score as high as we would have liked in regards to r2 but we were able to see benefits from them

- We were able to find which features offered more importance. Coupled with more features could help increase our predictions

- Prophet and LR models offered useful historical and future trends as well as offer insight of important features

Next Steps:

- Predicting weather can be tricky as we've seen. To build a better model we might need to look at historical weather data from other parts of the country and world.
- The features we used didn't have as big of an impact as we would have like. For future modeling we could bring in more features and find a better mixture of features that show more importance to our Rainfall variable