Final Project – Lessa Fleming

Topic: Weather Analysis

Data Source #1: Kaggle US Weather Events (2016 - 2022)
Data format: CSV file

Data Description and Project Scope:

The data runs from 2016-2022 displaying the weather data for the entire US. This CSV file has fourteen columns that display where the weather was registered and what the weather was, as well as the severity of the weather. This analysis will focus on the state of SC and work to predict what the weather will be for 2023 and 2024.

Data Preparation and Analysis Plan:

Part I:

1. Locate the weather file that will allow for proper forecasting.
2. Select the file and load into Jupyter using PySpark and store it into Mongo.

Part II

1. Install library and applicable modules (example: pandas, PySpark, RandomForest, Plt)
2. Scrub and Explore the Data
3. Transform Data
    a. Create graphs.
4. Analysis
    a. Forecast graphs.
    b. Chart of weather type

Part III:

1. Conclusion

After downloading the data from Kaggle, the data set is loaded into Jupyter. The file size was large and took time to load. For further analysis, the data was then cleaned to remove data and create a data frame to only view data for South Carolina for forecasting later. This data contains all states and daily weather events starting in 2016. The data also contained Null data in certain columns.

Analysis -

Loading the data Longitude and latitude are not needed for analysis or the airport code column. These were removed and the data was viewed. The data was then filtered, and a new data frame was created only to show SC data. The file was so large this will aid in analysis.

EventId	Type	Severity	StartTime(UTC)	EndTime(UTC)	Precipitation(in)	TimeZone	AirportCode	LocationLat	LocationLng	City	County	State	ZipCode
W-59597	Rain	Light	2016-01-08 09:55:00	2016-01-08 10:35:00	0.0	US/Eastern	KSPA	34.9157	-81.9565	Roebuck	Spartanburg	SC	29376
W-59598	Rain	Light	2016-01-08 10:55:00	2016-01-08 11:35:00	0.0	US/Eastern	KSPA	34.9157	-81.9565	Roebuck	Spartanburg	SC	29376
W-59599	Rain	Light	2016-01-08 12:15:00	2016-01-08 12:35:00	0.0	US/Eastern	KSPA	34.9157	-81.9565	Roebuck	Spartanburg	SC	29376
W-59600	Rain	Light	2016-01-08 12:55:00	2016-01-08 16:35:00	0.13	US/Eastern	KSPA	34.9157	-81.9565	Roebuck	Spartanburg	SC	29376
W-59601	Rain	Light	2016-01-08 16:55:00	2016-01-08 17:15:00	0.02	US/Eastern	KSPA	34.9157	-81.9565	Roebuck	Spartanburg	SC	29376
W-59602	Fog	Severe	2016-01-08 22:15:00	2016-01-08 22:35:00	0.0	US/Eastern	KSPA	34.9157	-81.9565	Roebuck	Spartanburg	SC	29376
W-59603	Rain	Light	2016-01-09 19:55:00	2016-01-09 20:55:00	0.01	US/Eastern	KSPA	34.9157	-81.9565	Roebuck	Spartanburg	SC	29376

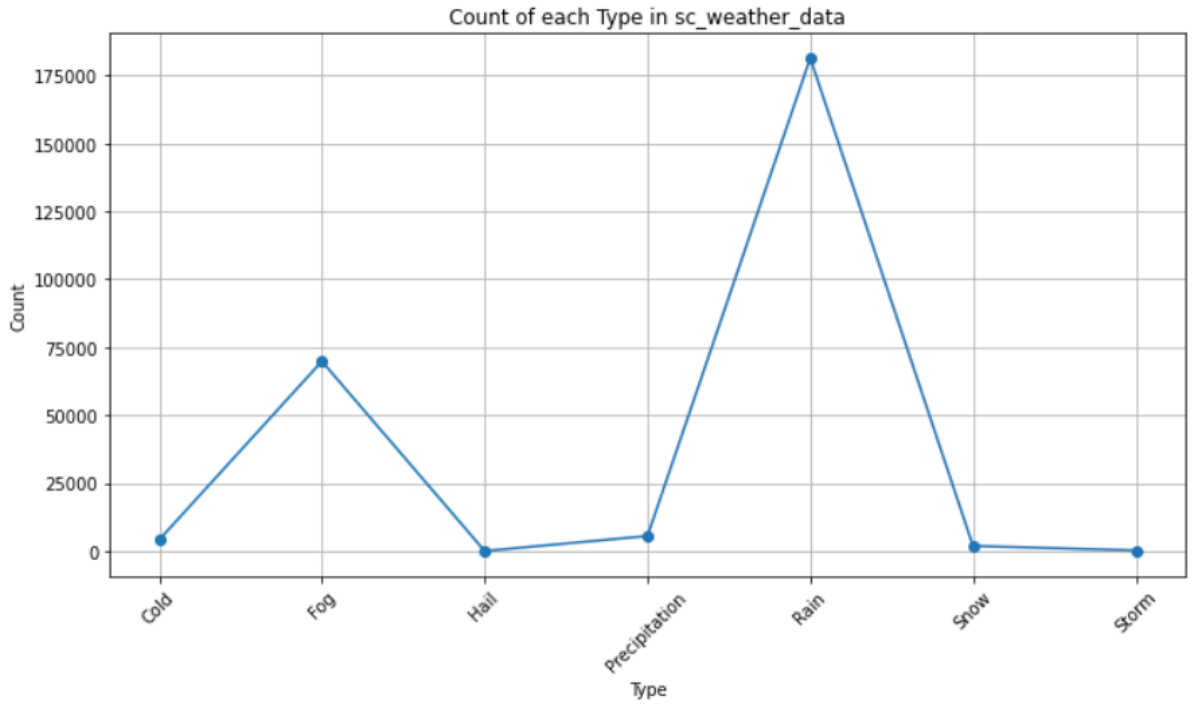
The data was then viewed to see the summary statistics for Type, Precipitation and ZipCode: This showed the average precipitation was 9.3 with a standard deviation of 3.8. The max precipitation is 9.99 with the lowest of six. This will aid when viewing the forecasting.

summary	Type	Precipitation(in)	ZipCode
count	262938	262938	262938
mean	null	0.09266907027512254	29484.194977523217
stddev	null	0.3796076585291063	237.19485055474027
min	Cold	0.0	29020
25%	null	0.0	29360.0
50%	null	0.0	29527.0
75%	null	0.06	29649.0
max	Storm	9.99	29926

Charts and graphs

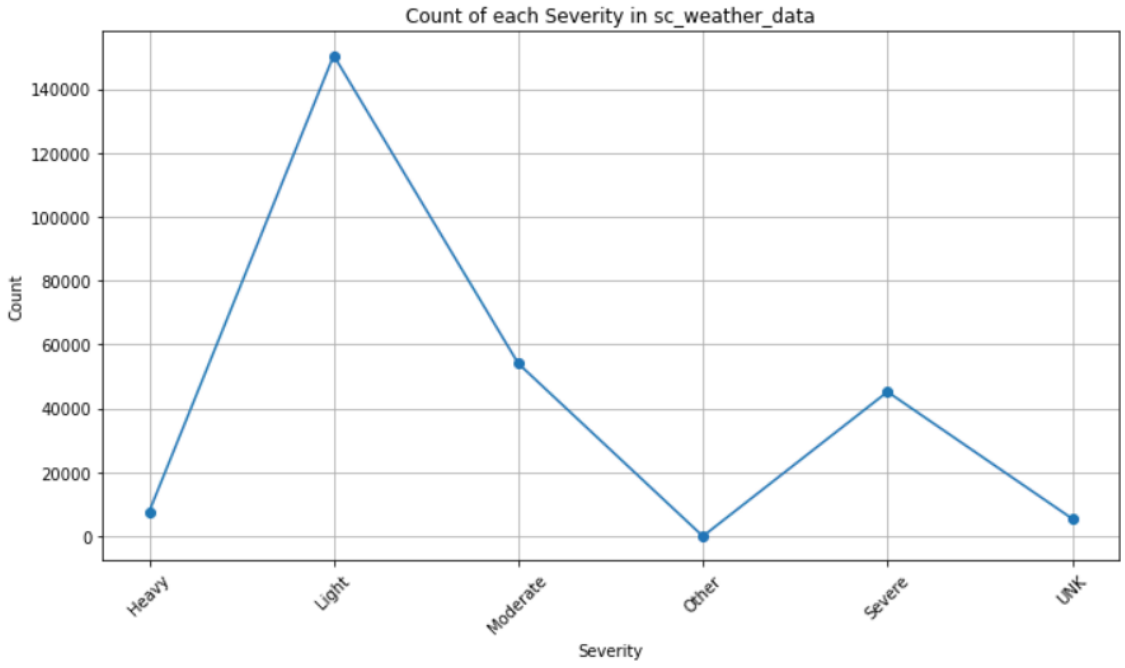
Line Chart of weather type

This first shows that fog and rain are the most weather types in SC. With Storms, hail and snow the lowest.

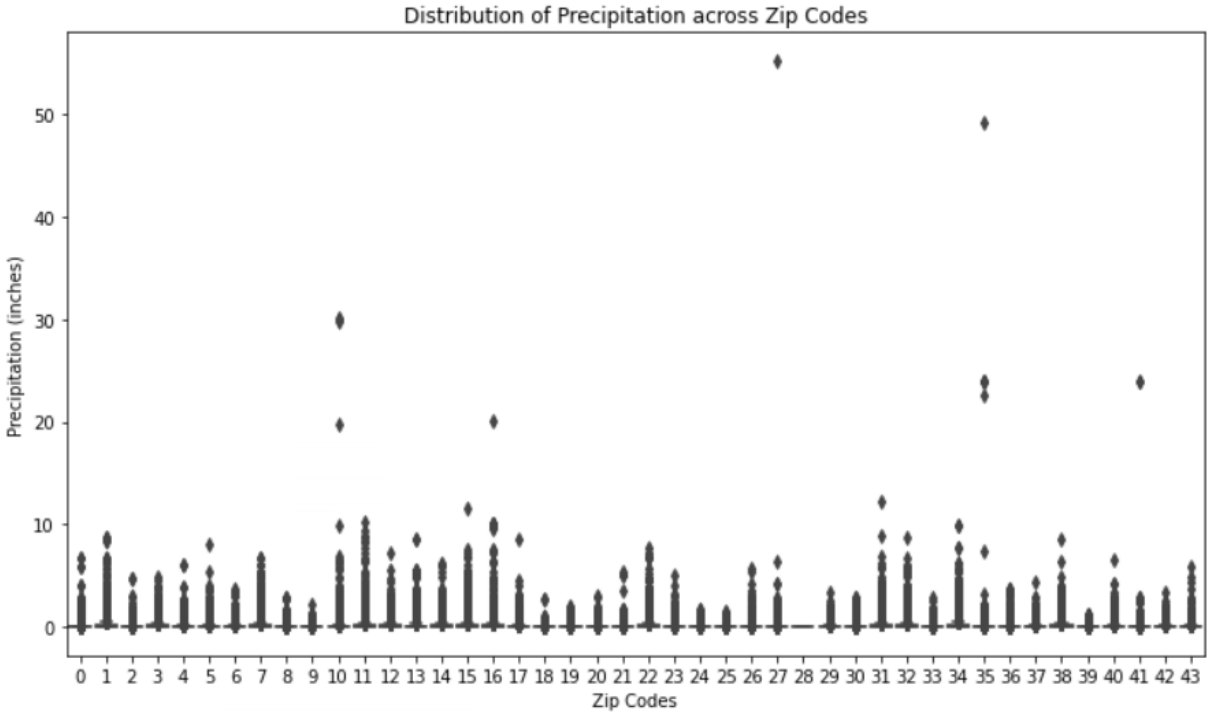


Severity Plot

This plot shows the severity of the weather types with the highest as light followed by moderate and severe.



The next plot shows the distribution of precipitation across zip codes. Due to there being so many different zip codes they are represented by numbers. In this, you can see that most zip codes are even in their precipitation with a few outliers with higher precipitations. These could be the coastal areas or have gotten hurricane activity over the years. There are a few with lower precipitation as well.



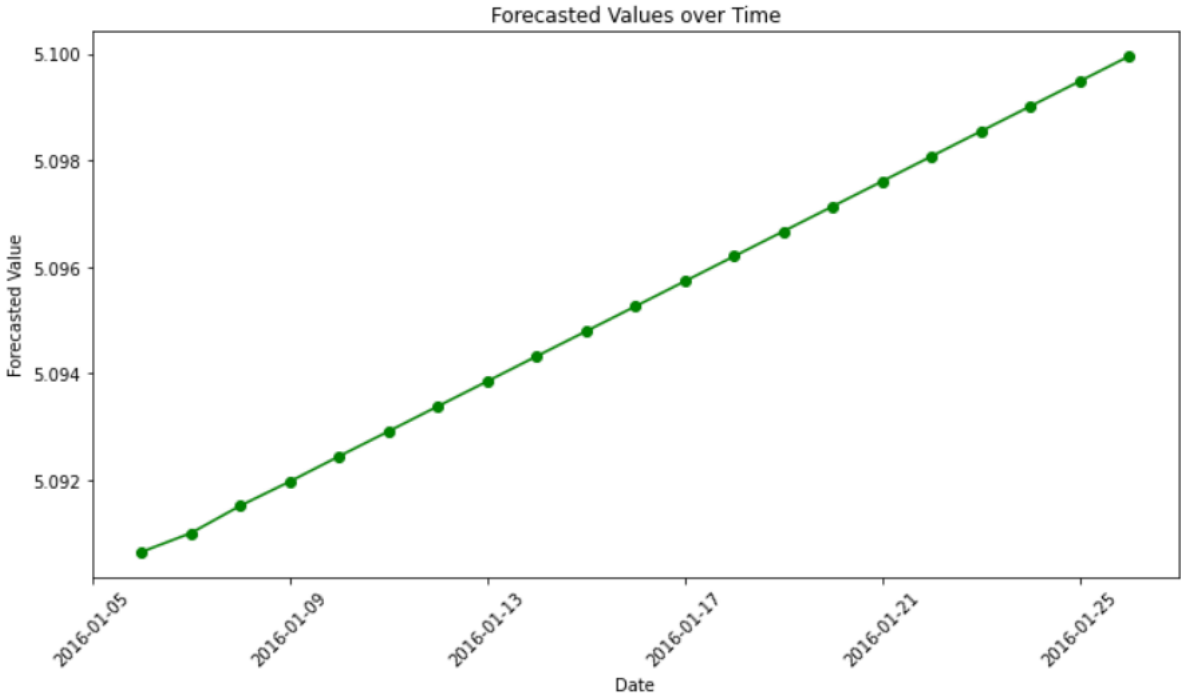
Arima Forecast

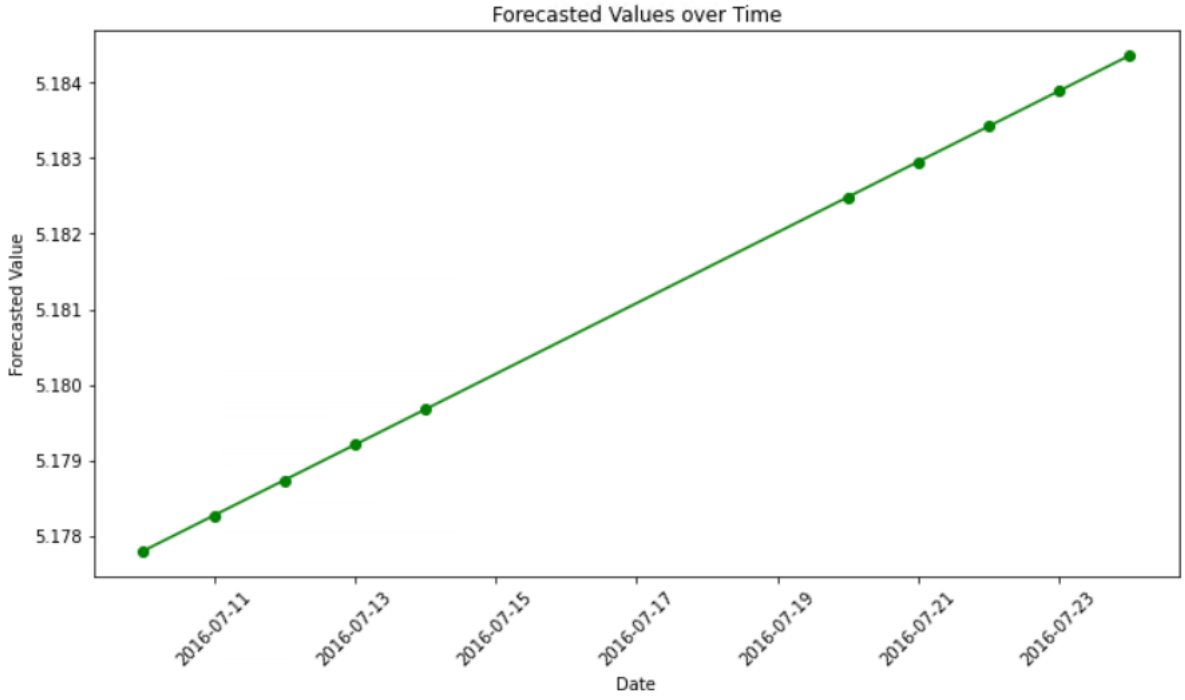
The Arima forecasted_steps to represent dates for the future, these are represented as numbers, each shows the predicted precipitation for the date. This assumes that the start date of the data is 01/06/2016. This data set ends in 2022 and the forecast predicts for 2023.

```
2542    4.549203
2543    5.289536
2544    5.010973
2545    5.116749
2546    5.077549
2547    5.093035
2548    5.087892
2549    5.090531
2550    5.090235
2551    5.091046
2552    5.091439
2553    5.091990
2554    5.092481
2555    5.092995
2556    5.093500
2557    5.094009
2558    5.094516
2559    5.095024
2560    5.095531
2561    5.096039
2562    5.096547
2563    5.097054
2564    5.097562
2565    5.098070
2566    5.098577
2567    5.099085
2568    5.099593
2569    5.100100
2570    5.100608
           ...
2747    5.182485
2748    5.182954
2749    5.183423
2750    5.183892
2751    5.184361
Name: predicted_mean, Length: 210, dtype: float64
```

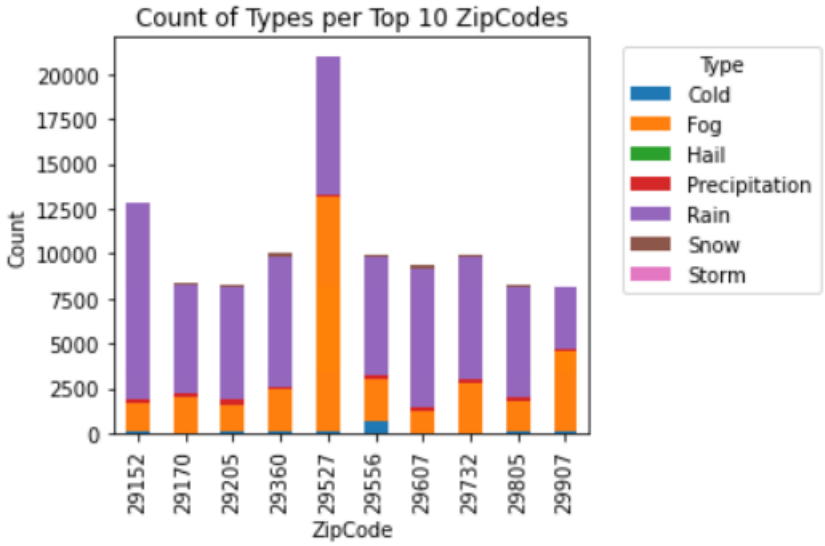
This shows the forecasting step dates to correlate with the precipitation estimates above and will be used below to plot the results.

```
Date for forecast step 2551: 2022-12-31 00:00:00
Date for forecast step 2552: 2023-01-01 00:00:00
Date for forecast step 2553: 2023-01-02 00:00:00
Date for forecast step 2554: 2023-01-03 00:00:00
Date for forecast step 2555: 2023-01-04 00:00:00
Date for forecast step 2556: 2023-01-05 00:00:00
Date for forecast step 2557: 2023-01-06 00:00:00
Date for forecast step 2558: 2023-01-07 00:00:00
Date for forecast step 2559: 2023-01-08 00:00:00
Date for forecast step 2560: 2023-01-09 00:00:00
Date for forecast step 2561: 2023-01-10 00:00:00
Date for forecast step 2562: 2023-01-11 00:00:00
Date for forecast step 2563: 2023-01-12 00:00:00
```


Forecasted Values over Time
Forecasted Value
5.100
5.098
5.096
5.094
5.092
2016-01-05
2016-01-09
2016-01-13
2016-01-17
2016-01-21
2016-01-25
Date


Forecasted Values over Time
Forecasted Value
5.184
5.183
5.182
5.181
5.180
5.179
5.178
2016-07-11
2016-07-13
2016-07-15
2016-07-17
2016-07-19
2016-07-21
2016-07-23
Date

After plotting the results of the forecasting, a chart was created to see the types of precipitation for the top ten zip codes. As you can see Rain is the most type of precipitation.



Conclusion

After analysis and forecasting for 2023 Rain Type will be the most frequent type of weather based on historical weather data. Fog will be the next most frequent type followed by cold which is still one of the lowest weather types as well. In the ARIMA forecasting the most comment estimated precipitation is lower than the summary precipitation numbers of around 9. In the forecast, during winter months precipitation is higher than the forecast for July and summer months as would be expected.