Weather Analysis

This report explores historical weather patterns in the North-West Florida and Southern Alabama geographic areas.

The data comes from NOAA (https://www.ncdc.noaa.gov/). Specifically, it was downloaded from this FTP site.

I focused on six measurements:

- **TMIN, TMAX:** the daily minimum and maximum temperature *(in tenths of degrees C)*
- **TOBS:** The average temperature for each day *(in tenths of degrees C)*
- **PRCP:** Daily Precipitation *(in mm)*
- **SNOW:** Daily snowfall *(in mm)*
- **SNWD:** The depth of accumulated snow *(in mm)*

In my geographic location (*fig 1.*), there are 177 unique weather stations and $12,249$ data observations between all stations. I have data from year $1890 - 2012$. Specifically my data is distributed between the following 6 measurements:

Measurement	Rows of data
TMIN	2021
TOBS	1344
TMAX	2020
SNOW	2107
SNWD	1973
PRCP	2784
TOTAL	12249

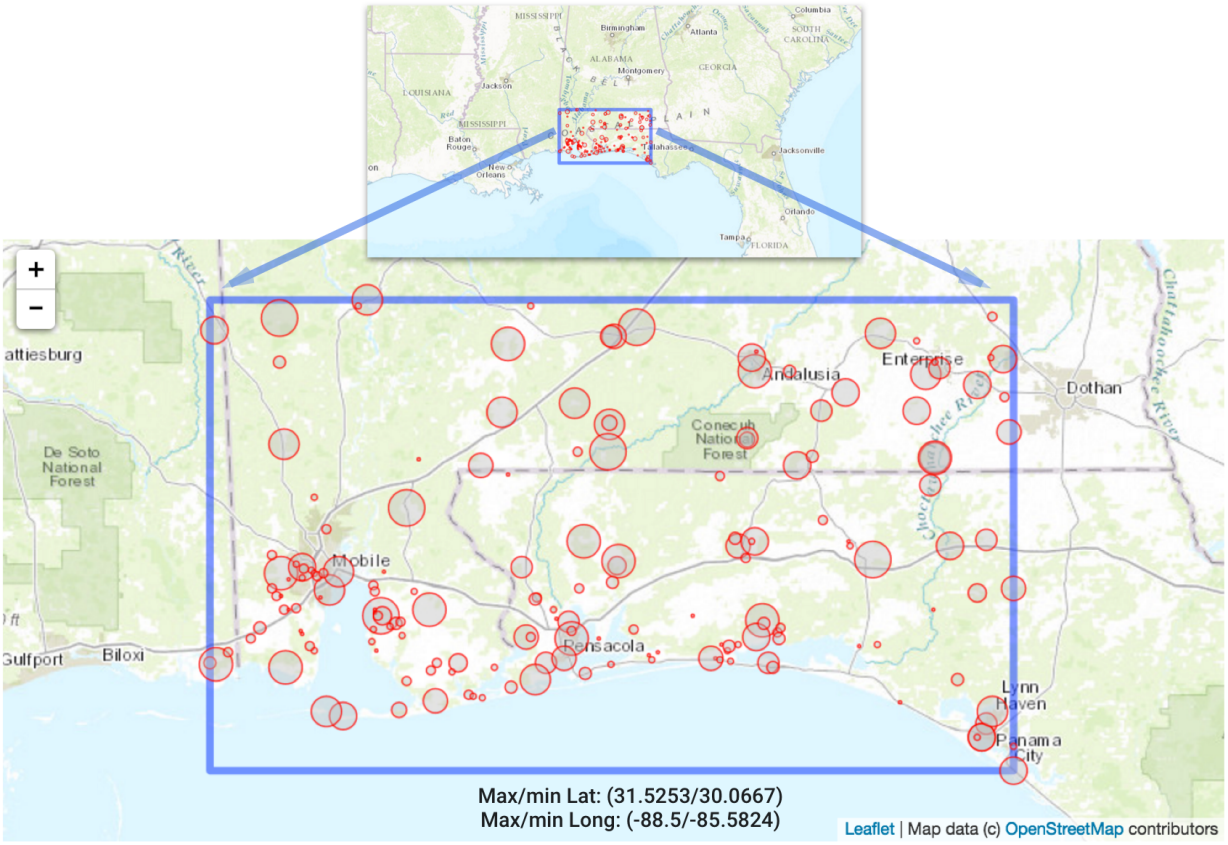


Figure 1: Geographic Location, Radius of circle corresponds to the amount of data generated by that station.

Each observation includes the **station, its elevation, lat/long**, as well as the **measurement type, year** and a **vector** of length 365. These vectors are measurements from this station for each day of that year. I will be diving deeper into these vectors for each type of measurement.

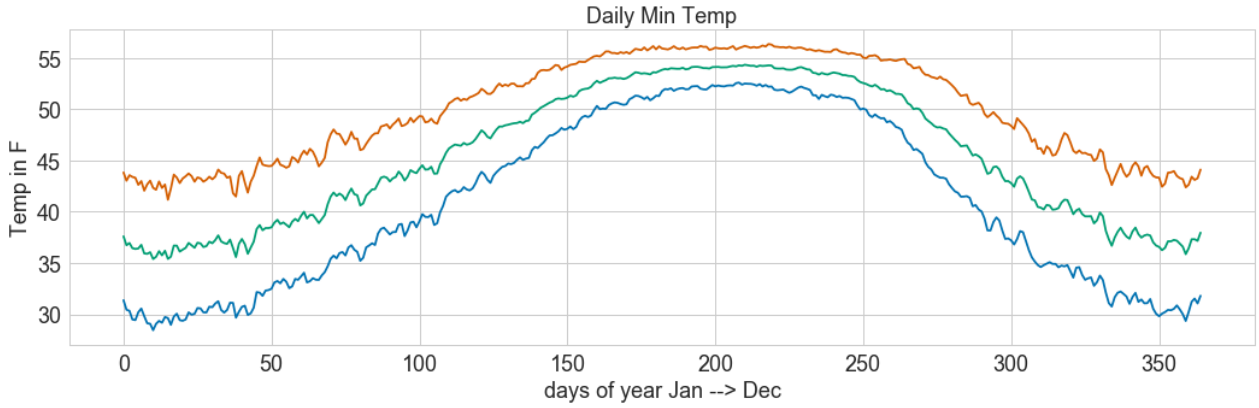
	elevation	latitude	longitude	measurement	station	undefs	vector	year
0	14.9	30.4132	-86.6635	PRCP	US1FLOK0014	38	[0, 0, 0, 0, 176, 91, 0, 66, 0, 126, 96, 86, 0...	2009.0
1	6.4	30.2119	-85.6828	TMAX	USW00003882	5	[64, 90, 240, 90, 128, 88, 128, 81, 224, 80, 8...	1999.0
2	6.4	30.2119	-85.6828	TMAX	USW00003882	3	[32, 91, 120, 91, 72, 91, 152, 90, 0, 88, 184,...	2000.0

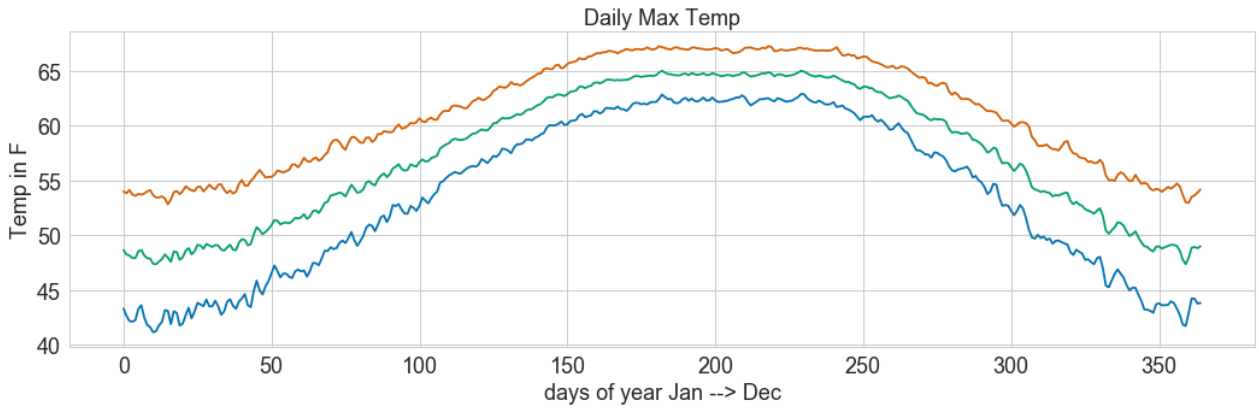
	elevation	latitude	longitude	measurement	station	undefs	vector	year
12246	74.7	30.7244	-86.0939	SNWD	USC00082220	0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	2007.0
12247	74.7	30.7244	-86.0939	SNWD	USC00082220	0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	2008.0
12248	74.7	30.7244	-86.0939	SNWD	USC00082220	0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	2009.0

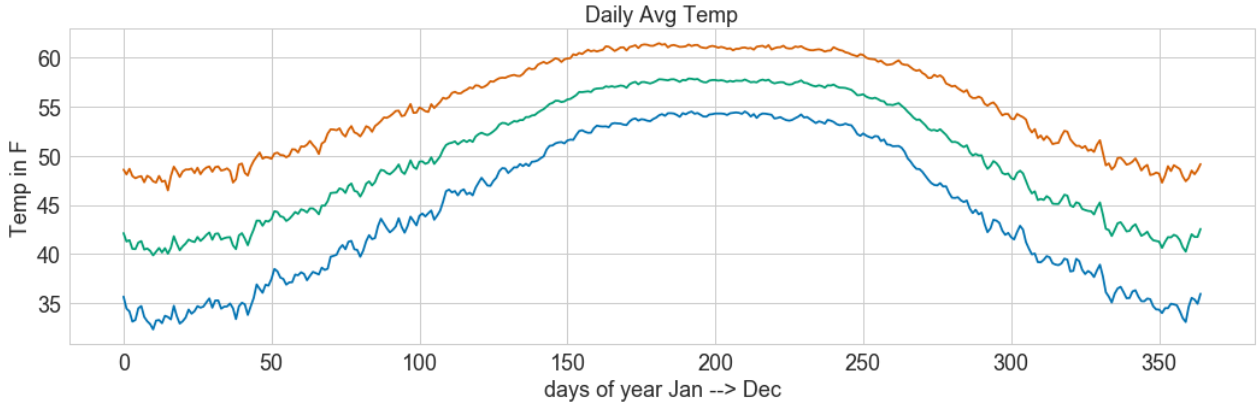
Univariate Analysis

Plots of all individual measurements. These plots provide us with a sense of how the data looks, if there are weird data points and a general trend. E.g. the temperatures through out the year follow a trend that we would expect to observe (colder in the winter months, and a slow change in the weather throughtout the year. The mean data analyzed does match up approximatly with data reported at US Climate Data (http://www.usclimatedata.com/climate/pensacola/florida/united-states/usfl0715).

Plot of mean temperatures (*middle line*) and a +/− one standard deviation (*top and bottom line*) :







Interestingly, the temperature during summer time have less variance, winter months have greater variability in their temperatures. Perhaps in the geographic region summers are mostly hot, whereas winter months have both cold and hot days, or perhaps bigger ranges of temperatures are observed.

Because of my region (northern Florida, southern Alabama) and the fact that the snow data is so sparse, I will not be performing much more analysis:

```
Percentage of Zeroes in the Snow (SNOW) Data = 99.08%
Percentage of Zeroes in the Snow (SNWD) Data = 99.05%
```

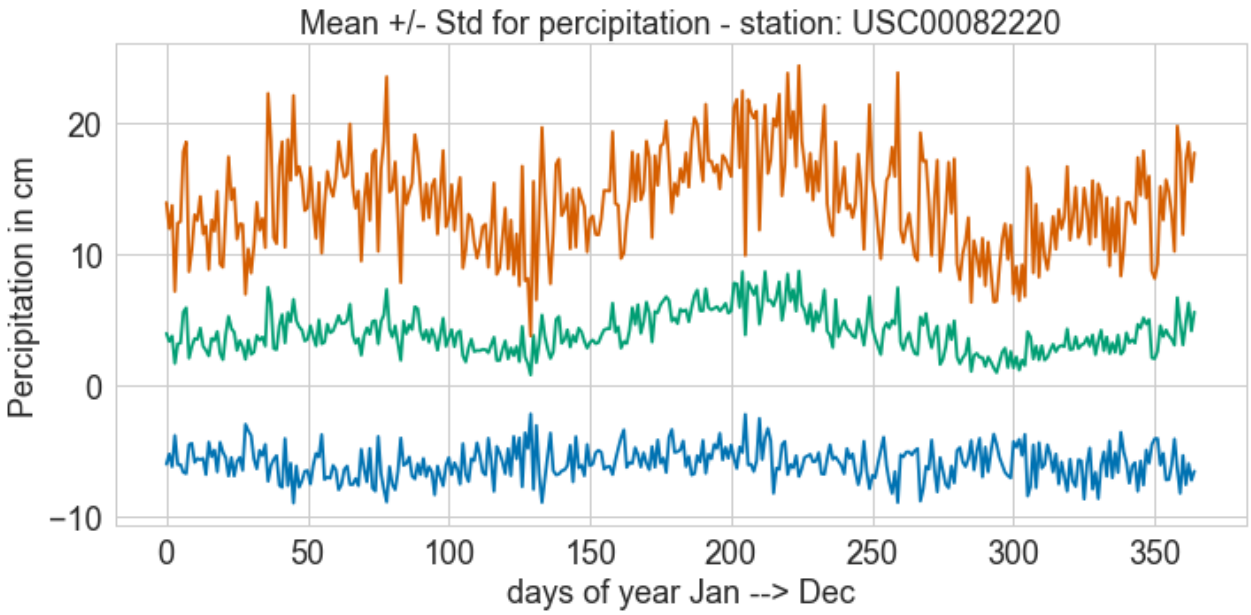
Instead I will delve into precipitation and temperature.

My Precipitation data still has a good amount of zeroes, approximately:

```
Percentage of Zeroes in the Precipitation Data = 70.18%
```

This means that of all the days of data I have, over this entire region, at each station for all my years of data it rains approx 30% of the time. Below is a plot for one particular weather station. This station had the most amount of precipitation weather data, 97 years worth to be exact.

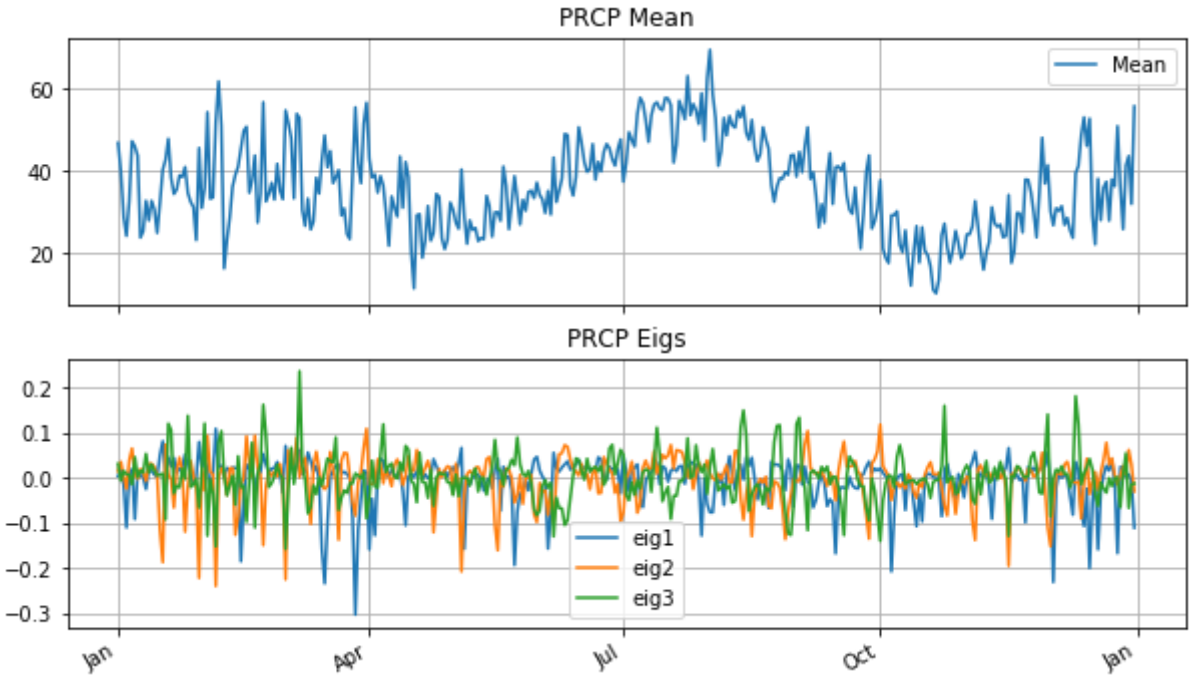
The three lines include: Mean (*middle line*) and a +/− one standard deviation (*top and bottom line*) :



We can tell that rain at this station on average occurs more frequently during the winter and especially the late summer months. This logic seems correct for the northern Gulf of Mexico as that is Hurricane season and the air temperature mixed with the hot water causes more storms.

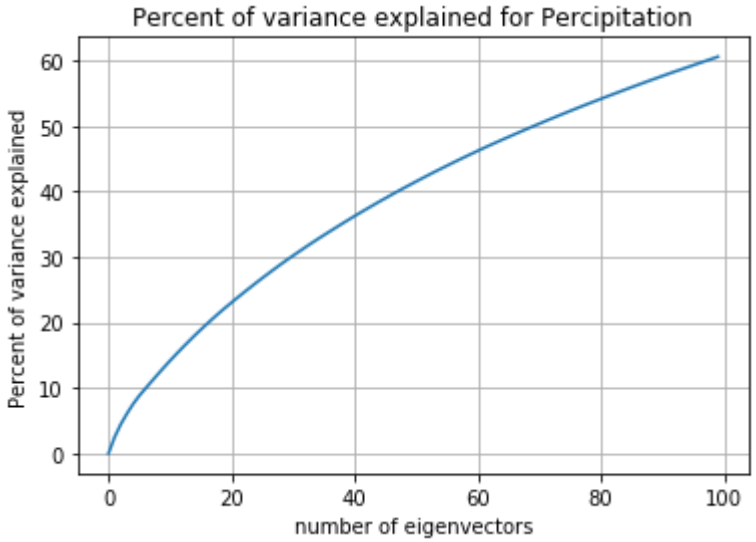
PCA Analysis for Percipitation

One issue with precipitation in my region, is that rain appears to occur at a reasonable rate throughout the year, with some spikes during hurricane season. Here is a plot of the Precipitation mean and the top three eigenvectors for all precipitation weather stations:



The mean line shows the regional storm patterns nicely; however, the 4 top eigenvectors look like noise and do not really seem to offer much information. Let's look at the explained ratio of variance for the eigenvectors to gain more insight.

The main issue is that the top 100 (about 33% of the dimensions) eigenvectors only explain about 60% of the variance, and the top 5 eigenvectors explain a paltry few percentage points.

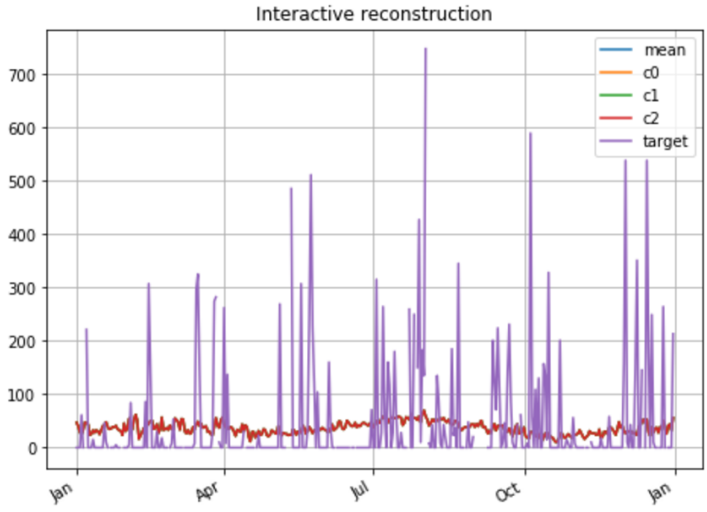


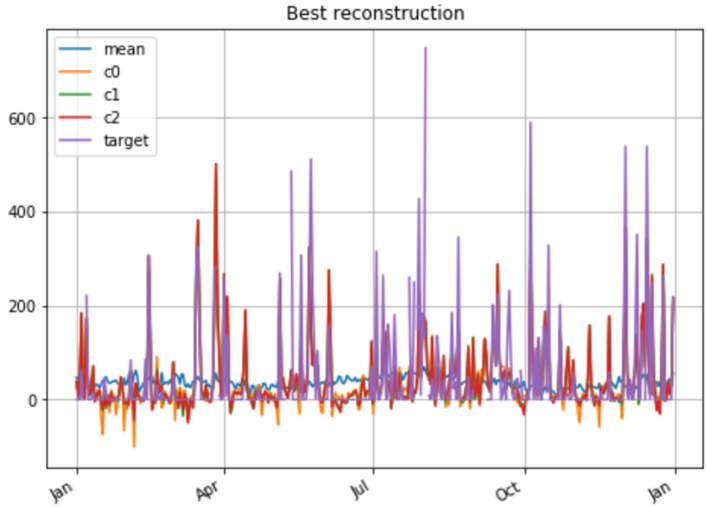
Only being able to explain a small amount of the variance with the top several eigenvectors is not very good. Here are the top 5 eigenvalues:

```
0.0250605 ,  0.01959968,  0.01621266,  0.01465037,  0.01255439
```

These eigenvalues tell us how much of the variance can be explained by their associative eigenvectors. These values are very low. We will not be able to represent the true signal of precipitation with any merit. But lets next look at the amount of reconstruction we can do with 3 eigenvectors.

The top plot shows the reconstruction with all the eigenvectors zeroed out, thus the mean is essentially shown (hidden) behind the red line. The purple is the true signal. The best reconstruction, with optimal eigenvalues, at best only reconstructs a small amount if the original signal. Nothing to be excited about. You can clearly see the eman in the background as the blue line, but much of the variability in the target signal is not being captured by this reconstruction.





Thoughts on PCA Analysis

I checked each covariance matrix and wrote a function to determine if the Cov was a positive semi definite matrix. Surprisingly each Cov matrix failed this test. I checked this function by trying to compute the Cholesky decomposition, again each matrix failed.

```
TMIN's cov is positive semi def: False
TOBS's cov is positive semi def: False
TMAX's cov is positive semi def: False
SNOW's cov is positive semi def: False
SNWD's cov is positive semi def: False
PRCP's cov is positive semi def: False
```

I assume that these are false for perhaps one of the following reasons:

- (i.) For each Cov at least one, or more, variables can be a linear combination of other variables.
- (ii.) Perhaps there is some really strong (almost uniform) colinearity between variables, and inexactness of floating point numerical computations there is a propagation of linearity and thus several of the eigenvalues become negative. This might be less likely the case because the negative eigenvalues are very large e.g. -6.5293e+02, whereas a very tiny negative eigenvalue might be contributed to floating point error.
- (iii.) The amount of missing data is causing the sample Cov to become unstable. This is probably because the way NaNs are being computed is causing errors. Even with using np.nanmean() vs np.mean() there appears to be awkward results.

Perhaps imputing data to remove the NaNs and then performing an ablation study to find highly correlated data and remove them would provide a better sample Cov. I assume that weather data from close proximity stations within the same year are strongly correlated. This type of data adds no new information about the covariance and taking them out during the preprocessing stage might provide a better set of data to work with.

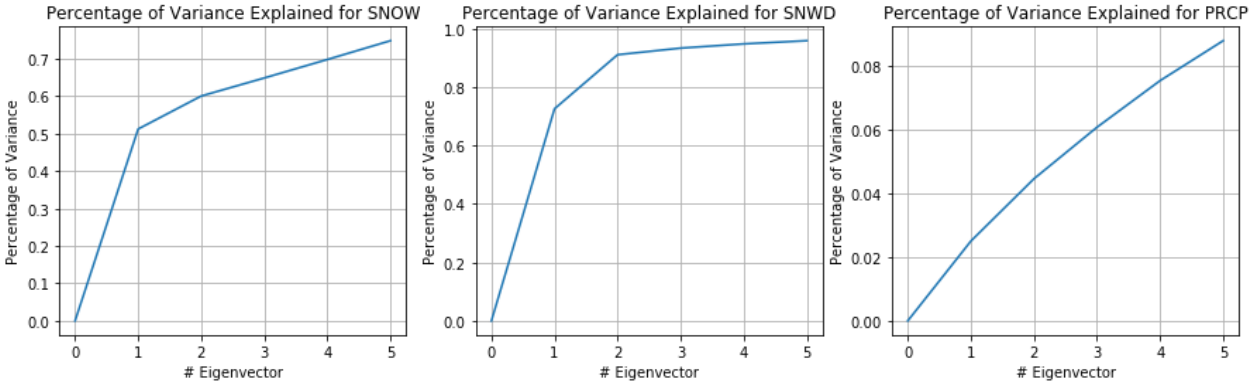
We can also look at the max magnitude of each eigenvalue:

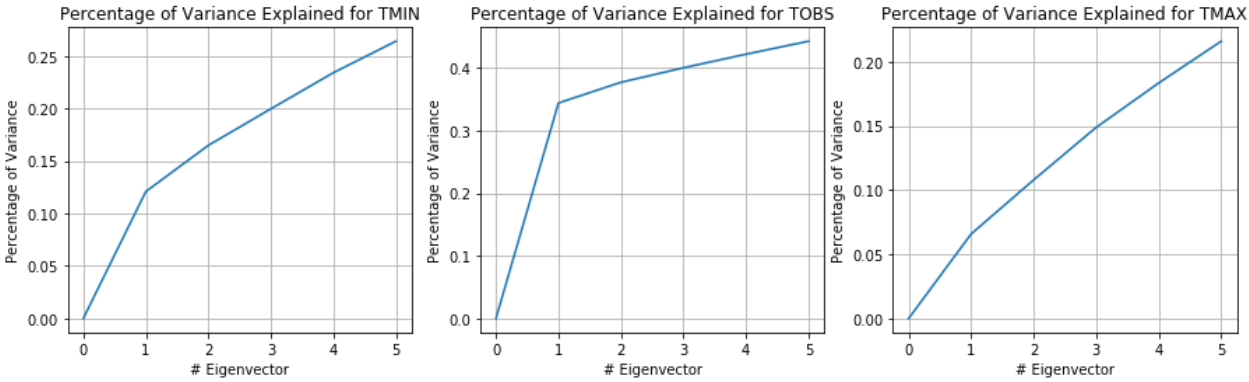
```
TMIN 100485.309663
TOBS 376528.613346
TMAX 36345.0352096
SNOW 312.258081267
SNWD 1416.53530402
PRCP 85913.3689075
```

Looking at these values it appears that from the data we have TOBS Temperature is most important for explaining the variance in the weather for this particular region then SNOW/SNWD. Precipitation is also fairly important, perhaps because of the spikes for hurricane season and winter. These attributes make it much more important that the 99% 0.0 value SNOW data!

Percentage of Variance Explained by Top Eigenvectors

Let's look at the Percent of Variance Explained for each of the Measurements:





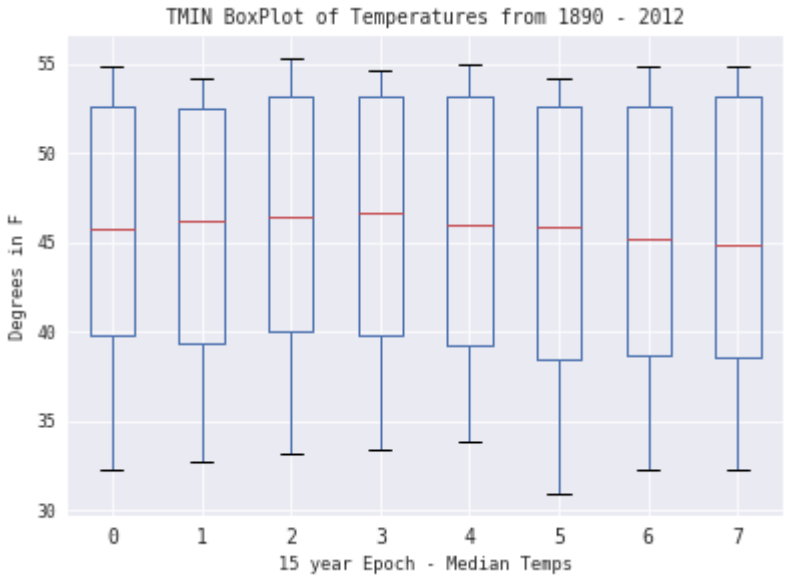
- **SNOW/SNWD:** The percent of explained variance is very good here for the first few Eigenvectors, however this is misleading, because over 99% of the data is missing.
- **PRCP:** going back to the previous analysis on precipitation, this shows that there is no real observable patterns to discern, because it is raining fairly consistently throughout the year, with elevated rain during hurricane season.
- **TMIN/TOBS/TMAX:** The top 5 eigenvectors explain 25% of variance for TMIN, 43% for TOBS and 25% for TMAX. Of the three, TOBS is best explained by the top 5 eigenvectors, especially the first vector which accounts for approx 35% of the variance.

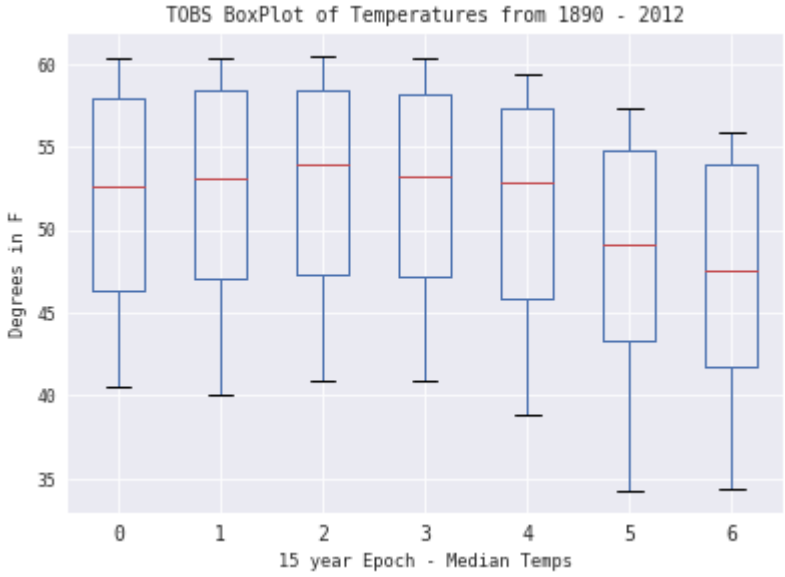
15 Year Moving Epoch for TMIN, TOBS, TMAX

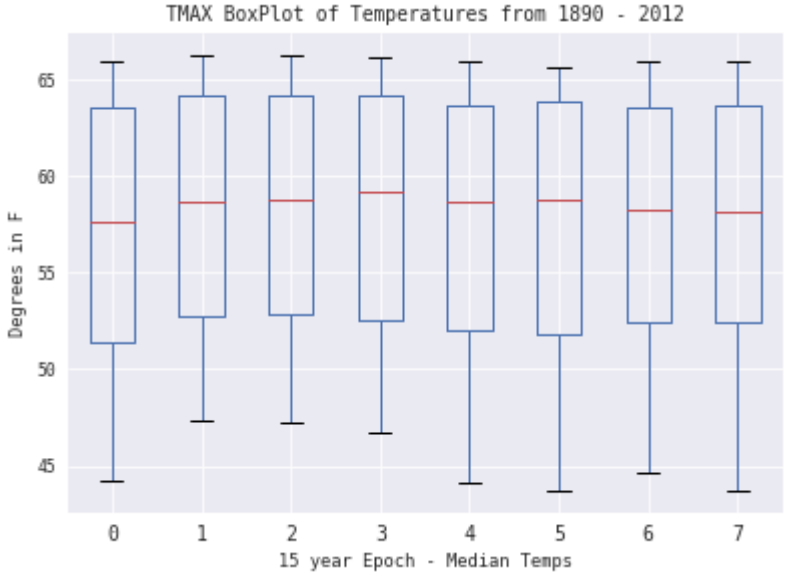
In the below plots, I combined temperatures within 15 year epochs. So boxplot 1 is from year 1890 - 1905, boxplot two is 1906 - 1920 and so on. I wanted to show a general temperature trend over the data for my specific geographical region. All of my station's elevations are within a small range so they should not factor in too much. One issue with these measurements that I would like to explore more into is what time of day these readings are taken. Because the temperature for TOBS may change if it was taken during the morning, noon, or evening. Especially if these stations are not taking it during the same time within the same year across all stations, or at the same time of day across all of a station's years of data collection.

The data below seems to suggest a slight uptick in the temperature followed by a dip in the avg temperature. Obviously this is for a small geographic region and as stated above there are some lurking variables which most likely remain unaccounted. Perhaps a different measurement technique, or standard protocol.

One quick note, TOBS had some missing years and thus only has 7 epochs made, vs TMIN/TMAX. There is also a drastic decrease in the TOBS temp measurement in the late 1900s. The cause for this is probably due to the time TOBS (which is temperature during time of observation) was recorded. Along the lines of what I said in the previous paragraph. I believe that we should also see a similar downward curve in the TMIN/TMAX plots if the temperature was on average decreasing such a drastic amount in the past 50 years.



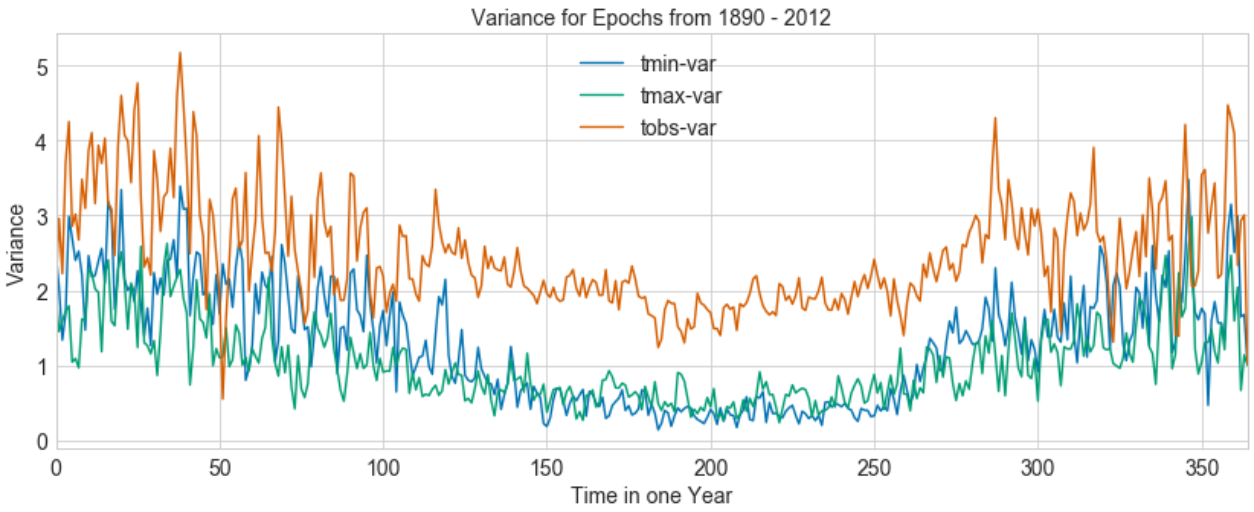




A plot of the variance for the epoch data above reveals that winter and fall months have a high

variability in temperature. Whereas summer months have less variance. This is true for all three temperature measurements. Overall TOBS has a higher variance and this could be accounted for if TOBS is taken at different times during the day between stations. Or perhaps rain during the stormy months and winter is far more dependent on the severity of storms, though in summer it rains only a little bit here and there. But a large storm could dump a tremendous amount of rain one week and then none the following week.

Another hypothesis is that during the summer the weather is uniformly hotter throughout the day and perhaps during the colder months the temperature range during the day and night are larger. If this is true then it makes sense to observe a larger variance of data during the colder months.



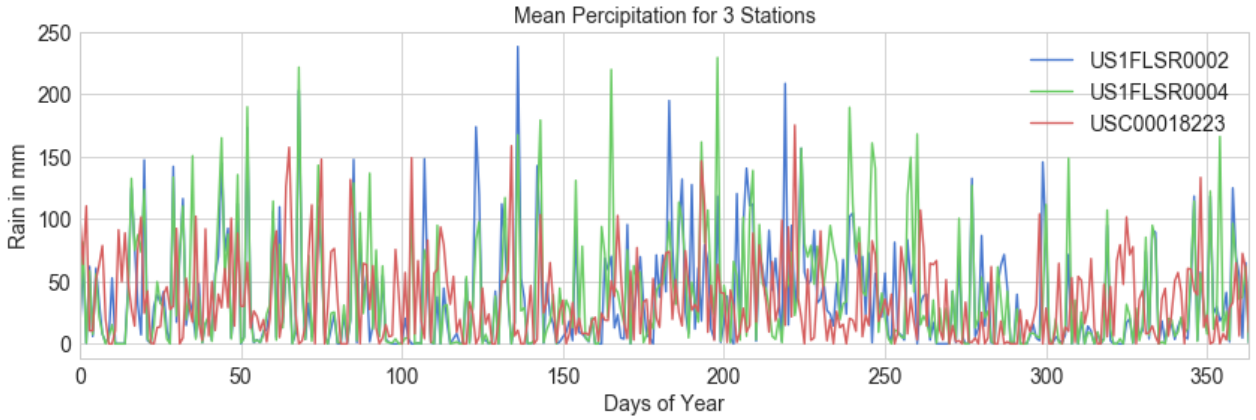
In [ ]:

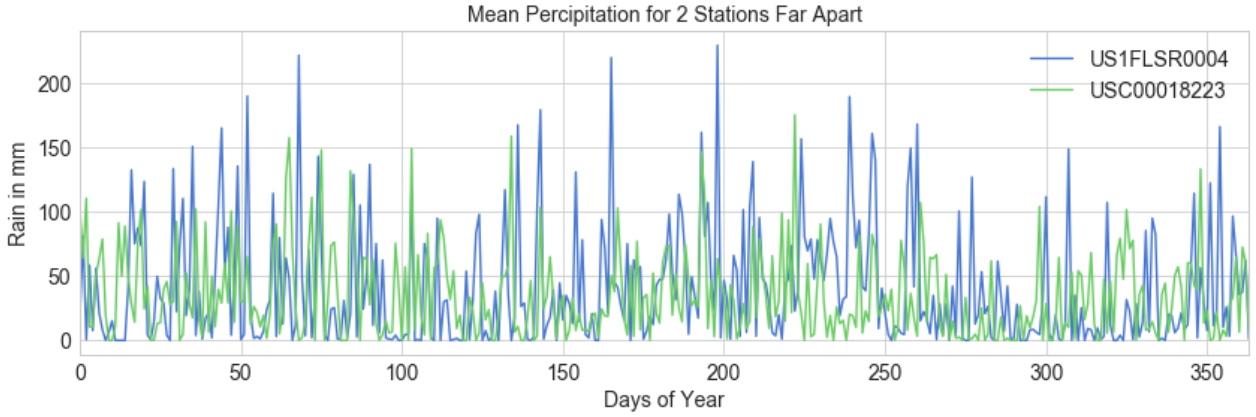
Analyzing Precipitation between 3 Stations

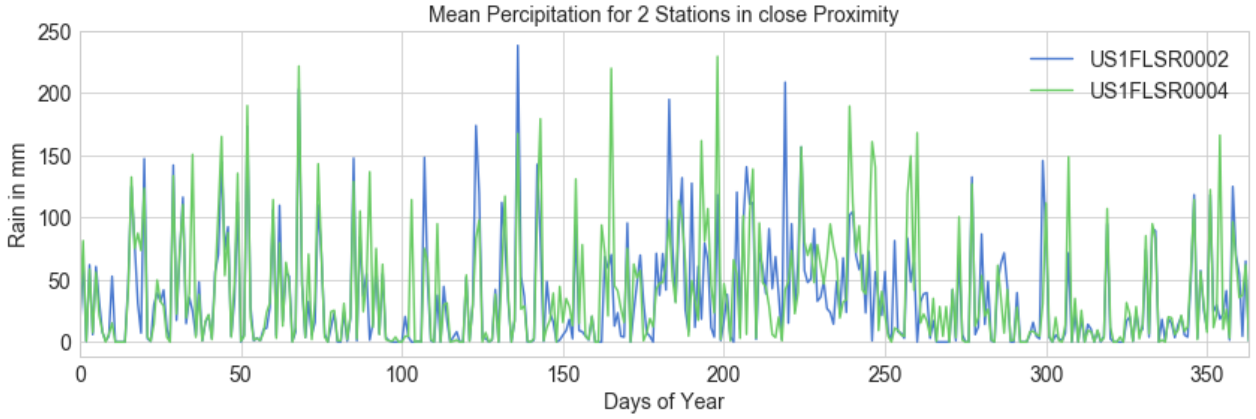
I pulled all the precipitation data for three weather stations. Two stations are very close to each other and the third station is very far apart.

```
distance from US1FLSR0002 --> US1FLSR0004 = 11.5 km
distance from US1FLSR0002 --> USC00018223 = 138 km
```

Below are three plots overlapping all three stations, two close stations and 2 far away stations. There appears to be more variance in the plot of two stations far apart, whereas the stations that are in close proximity seem to line up closer. This makes sense and is an expected result. Weather stations close together tracking the same measurement in the same years should have similar data. This is a helpful result one the less, because we can determine that there will be data that is highly collinear. This can be problematic in an analysis. Because highly collinear data does not explain any more of the variance.







I wanted to explore weather stations that are far apart and really close together. These stations had roughly the same amount of data ~7 years worth and had mostly overlapping years.

Conclusion

There are a lot of interesting patterns to tease out of this data set. I would like to spend more time running some statistical test to compare inter weather station data, especially between weather station on the coast and those inland.

There are certainly some internal issues with this data set that need to be taken care of before a meaningful analysis can occur. E.g. why are the eigenvalues negative and messed up, why are the Covs not semi positive definite? Also how are collinear weather station data being taken care of and

what is the best approach for dealing with NaN data (impute via regression model, mean imputation, interpolation, or simply remove the data)? Additionally, the snow measurements are really useless and should be removed.

Further analysis could look at Hurricane data and the years when there has been heavy hurricanes. These data points could be removed and I would like to see what happens to the precipitation plots. I bet we would not see that sharp rise in precipitation during September/October/November.

Thanks for reading through my report!

Cheers, Kyle