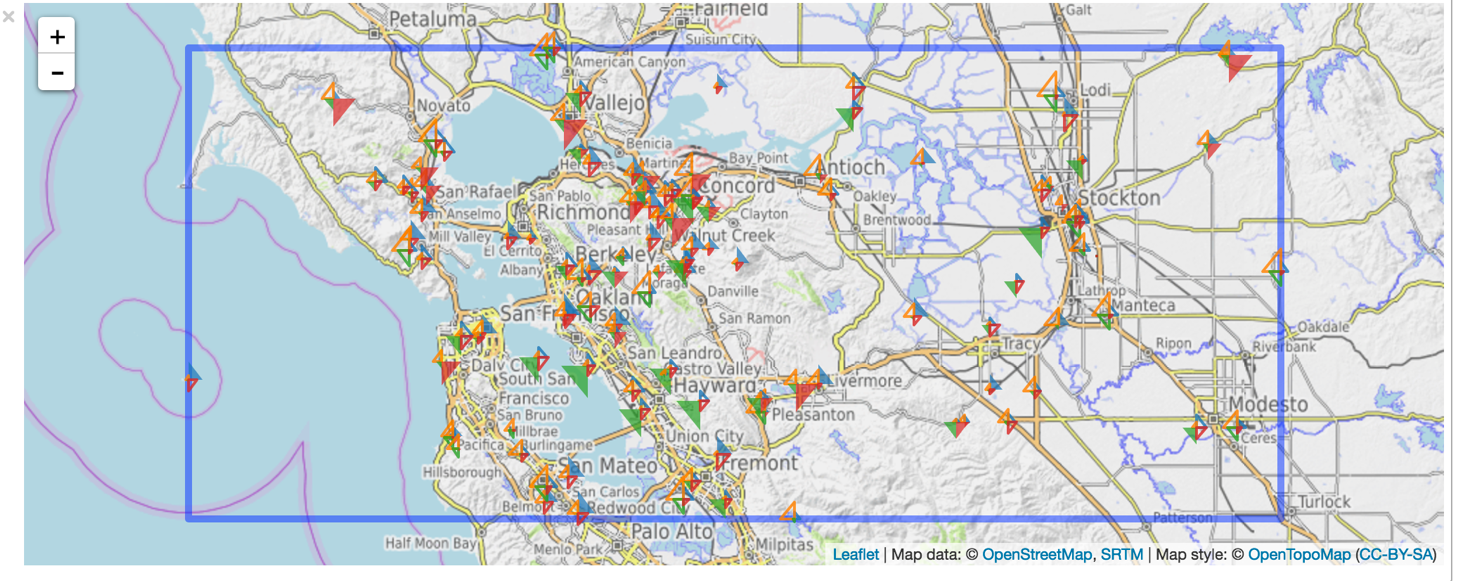
Northern California Weather Analysis

Chris Chen



# Fig 1: All Weather Stations in the My Data Set

# Introduction

This is a report on the historical analysis of weather patterns in an area that approximately overlaps the area of the state of Northern California.

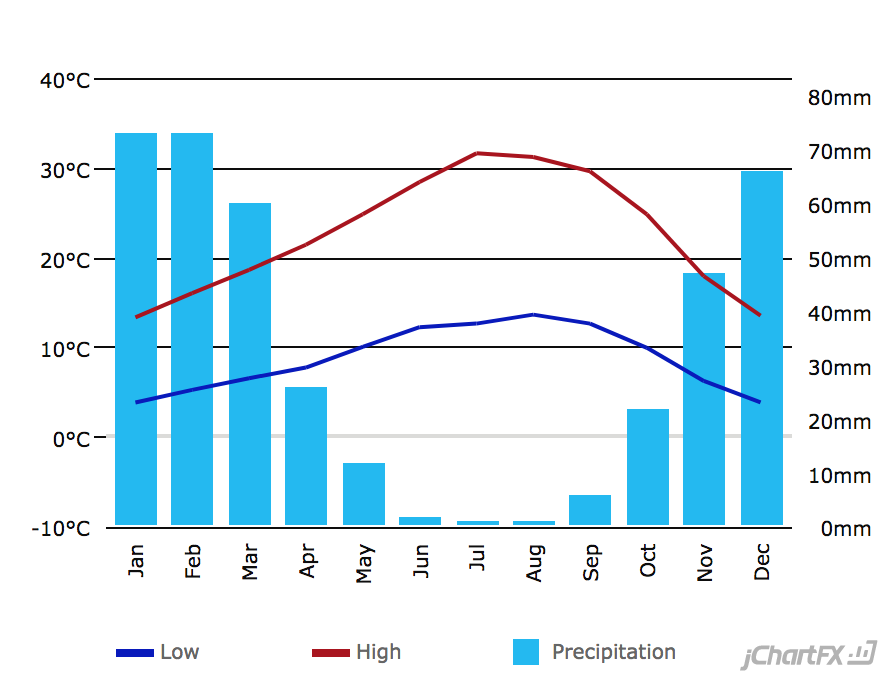
The data we will use here comes from [NOAA](https://www.ncdc.noaa.gov/). Specifically, it was downloaded from this [FTP site](ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/).

We focused on six measurements:

* **TMIN, TMAX**: The daily minimum and maximum tempature
* **TOBS:** The average temperature for each day.
* **PRCP:** Daily precipitation (in mm)
* **SNOW:** Daily snowfall (in mm)
* **SNWD:** The depth of accumulated snow.

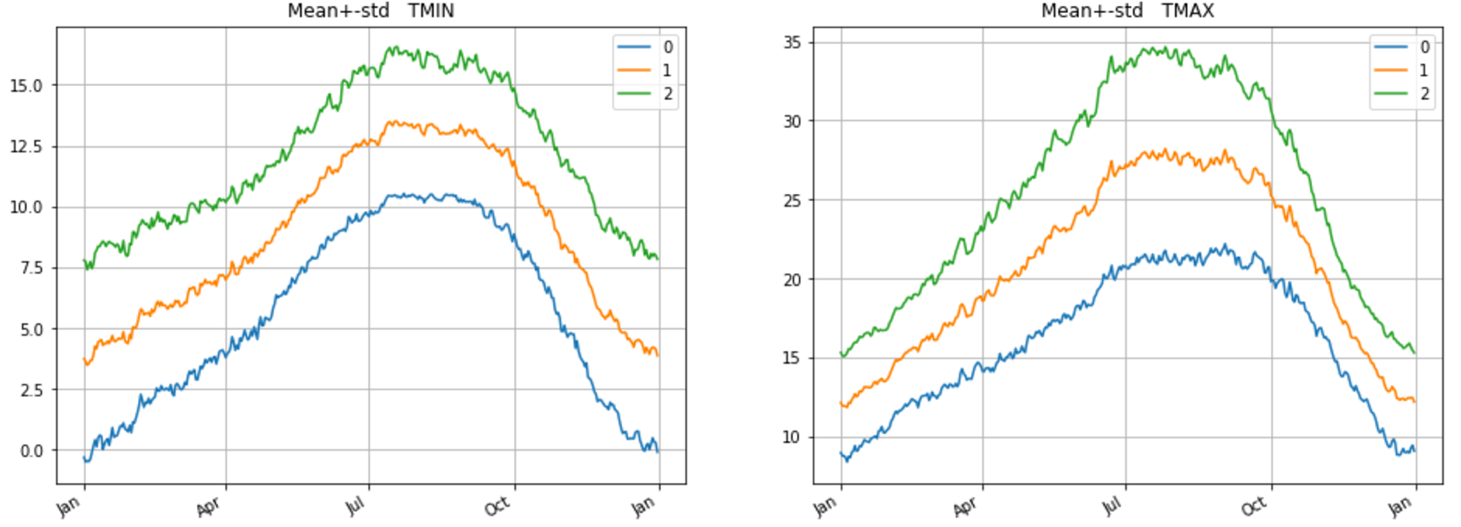
## Sanity-Check: comparison with outside sources

We start by comparing some of the general statistics with graphs that we obtained from a site call [US Climate Data](http://www.usclimatedata.com/climate/california/united-states/3174). The graph below shows the daily minimum and maximum temperatures for each month, as well as the total precipitation for each month.

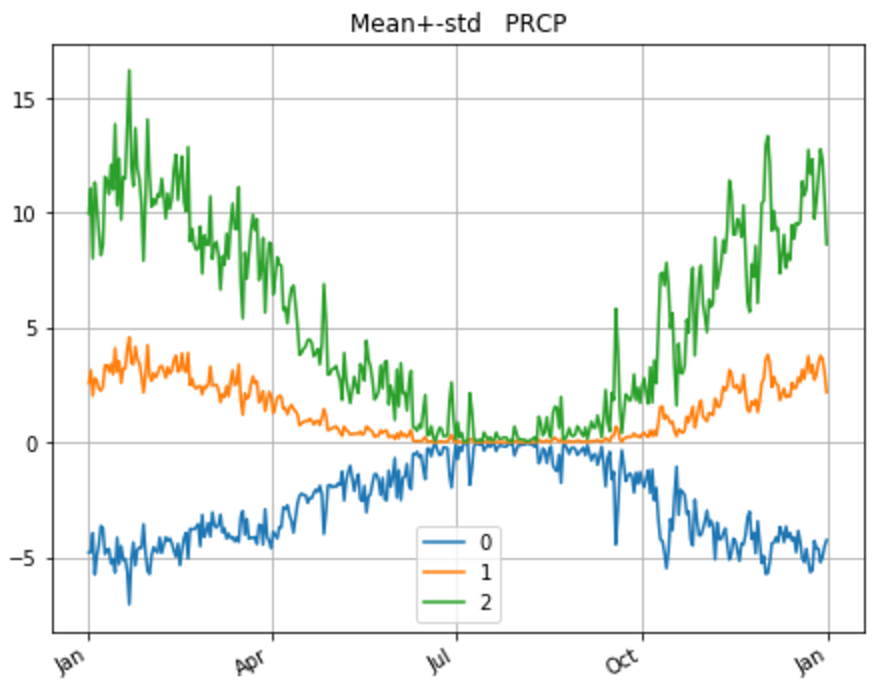


Three charts below (precipitation, Tmin, Tmax) are generated to validate with data from outside sources. We see that the min and max daily temperature agree with the ones we got from our data. For the rainfall chart, our data confirm with outside source that there is not much rain in summer and rain the most after January.





For precipitation, our analysis show the average rainfall is ~2.5 mm/day which translate to about 75 mm per month in January. The results are comparable what we are seeing in US-Climate-Data. However; there is clear agreement that raining season starts around October and end in April. Please note “1” in legend is equivalent to the mean.



### PCA Analysis

For each of the six measurement, we compute the percent of the variance explained as the number of eigenvectors used. Many of pre-computed statistics variables are provided for analysis. Below is a dictionary that contain the name of statistics variables, descriptions and shape of the variables.

[('SortedVals',

'Sample of values',

'vector whose length varies between measurements'),

('UnDef',

'sample of number of undefs per row',

'vector whose length varies between measurements'),

('mean', 'mean value', ()),

('std', 'std', ()),

('low100', 'bottom 1%', ()),

('high100', 'top 1%', ()),

('low1000', 'bottom 0.1%', ()),

('high1000', 'top 0.1%', ()),

('E', 'Sum of values per day', (365,)),

('NE', 'count of values per day', (365,)),

('Mean', 'E/NE', (365,)),

('O', 'Sum of outer products', (365, 365)),

('NO', 'counts for outer products', (365, 365)),

('Cov', 'O/NO', (365, 365)),

('Var', 'The variance per day = diagonal of Cov', (365,)),

('eigval', 'PCA eigen-values', (365,)),

('eigvec', 'PCA eigen-vectors', (365, 365))]

To do a sanity check, I want to verify that if all eigenvalues are positive for each of six measurements. Except for ‘PRCP’ measurement, every other measurements have eigenvalues that are negative. This suggest that there might be issues with the preprocessing data.

('TMIN', 3)

('TOBS', 43)

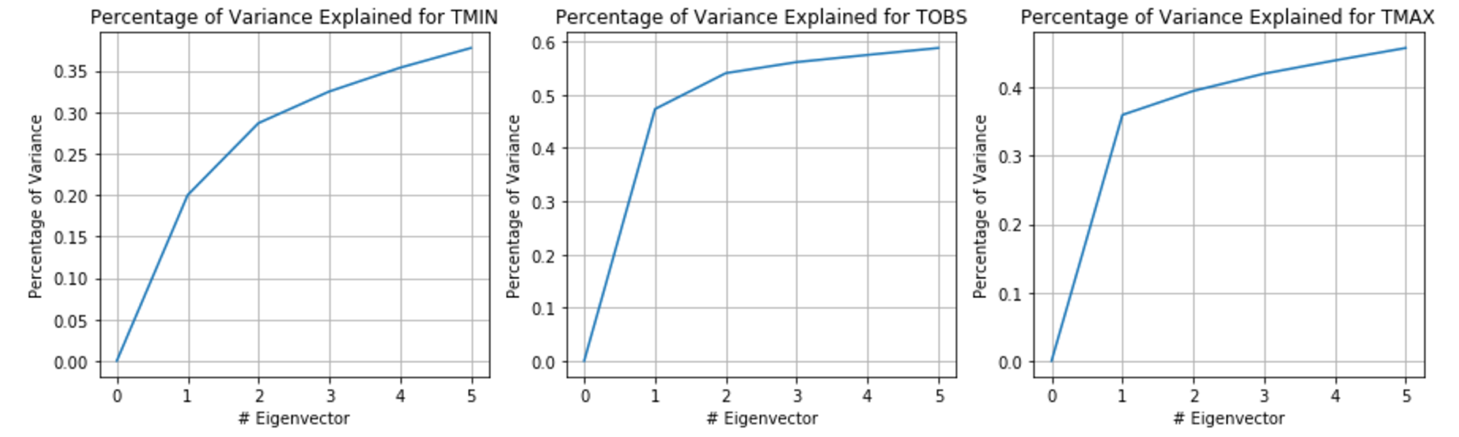
('TMAX', 27)

('SNOW', 12)

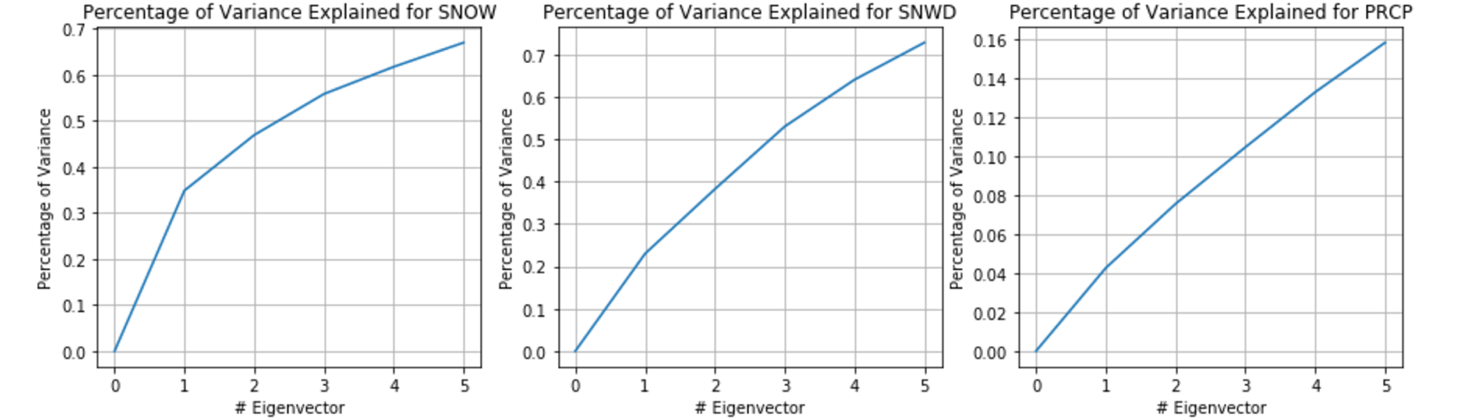
('SNWD', 4)

('PRCP', 0)

#### Percentage of Variance Explained



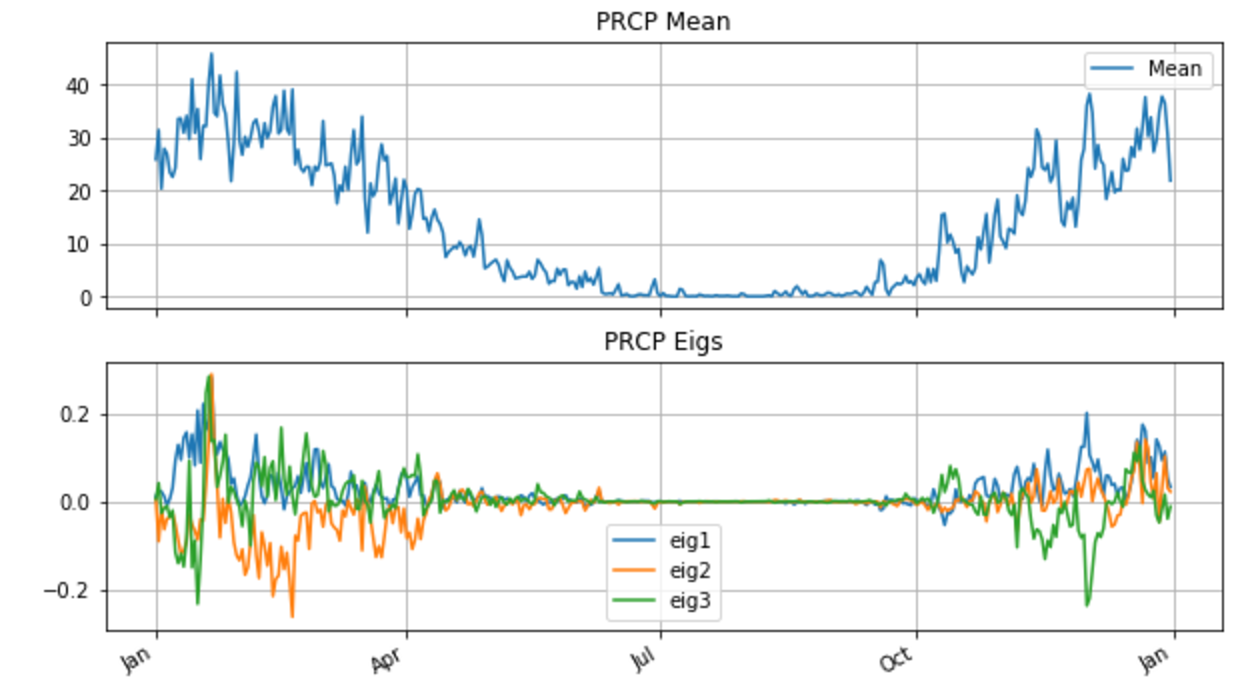
We see that the top 5 eigenvectors explained 37% of variance for TMIN, 58% for TOBS and 47% for TMAX. We conclude that of the three, TOBS is best explained by the top 5 eigenvectors. This is especially true for the first eigenvector which, by itself, explain 47% of the variance.



The top 5 eigenvectors explain 16% of variance for PRCP. However; the top 5 eigenvectors explained ~70% of the variance for SNOW and SNWD. This means that the top 5 eigenvectors capture most of the variation in the snow signals. It makes sense that in Northern California, both SNOW and SNWD varies less between days and between the same date on different years. The reasons are it hardly ever snow in Northern California.

## Analysis of Precipitation

I choose to analyze the eigen-decomposition for PRCP because there exist negative values for all other measurements. Recall, the eigenvalues and eigenvectors derived from the PCA need to satisfy positive semi-definite requirements.



The eigen-functions explain very little about the means. This is hardly surprising because the top eigenvalues explain a tiny percentage of overall variances.

### How well-explained are the vectors in this collection?

To answer this question, we extract all of the values of:

res\_1: residual variance after mean has been subtracted out

res\_2: residual variance after mean and the first eigen-vector have been subtracted out

res\_3: residual variance after mean and the first two eigen-vectors have been subtracted out.



### Analysis of Correlation of Precipitation of Weather Stations



The matrix above shows, for each pair of stations, the normalized log probability that the overlap in rain days is random.

We see immediately the first 30 stations are highly correlated with each other.

To find more correlations we use SVD (the term PCA is reserved for decomposition of the covariance matrix). As we shall see on the graph below that the top 10 eigenvectors explain about 90% of the square magnitude of the matrix.



## Analysis of correlation between precipitation across locations and/or Time

To estimate the effect of time vs. location on the first eigenvector coefficient we compute:

* The average row: mean-by-station
* The average column: mean-by-year

We then compute the RMS before and after subtracting either the row or the column vector.

total RMS = 180.793723228

RMS removing mean-by-station= 172.563345122

RMS removing mean-by-year = 80.9796786501

initial RMS= 180.793723228

0 after removing mean by year = 80.9796786501

0 after removing mean by stations= 67.4716166167

1 after removing mean by year = 66.5943913789

1 after removing mean by stations= 66.4794628951

2 after removing mean by year = 66.4526691357

2 after removing mean by stations= 66.4430363512

3 after removing mean by year = 66.4387621774

3 after removing mean by stations= 66.4366676894

4 after removing mean by year = 66.4355781453

4 after removing mean by stations= 66.4349863562

By removing mean by year, it will decrease the RMS by 100. This means that a lot of variations

can be explained by mean by years rather than mean by stations. **This means there is more correlation for precipitation among years than across stations.**