DSE230
Spring 2017
HW5
A53226915

Wyoming Weather Analysis

This is a report on the historical analysis of weather patterns in an area that covers approximately ¼ the area of the state of Wyoming.

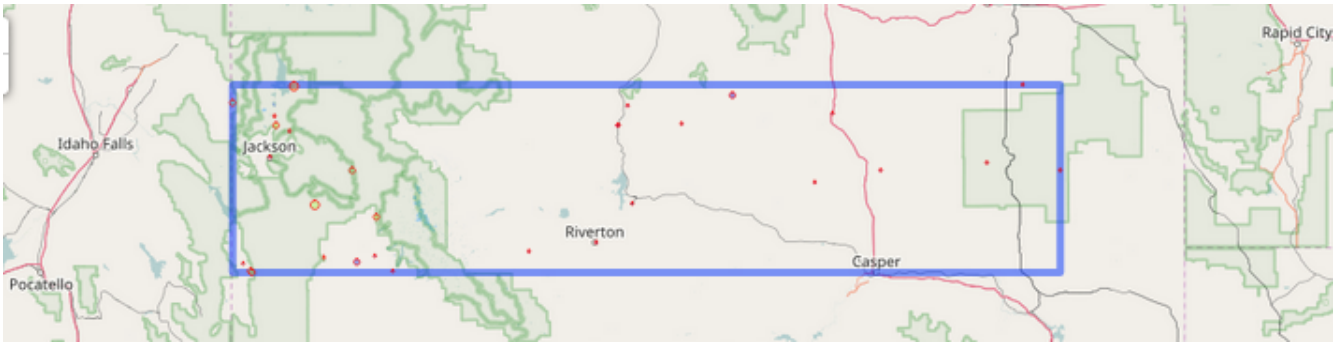
The data used here comes from NOAA. Specifically, it was downloaded from

ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/

I focused on six measurements:

- **TMIN, TMAX:** the daily minimum and maximum temperature.
- **TOBS:** The temperature at observation time each day.
- **PRCP:** Daily Precipitation (in mm)
- **SNOW:** Daily snowfall (in mm)
- **SNWD:** The depth of accumulated snow.

Study Area

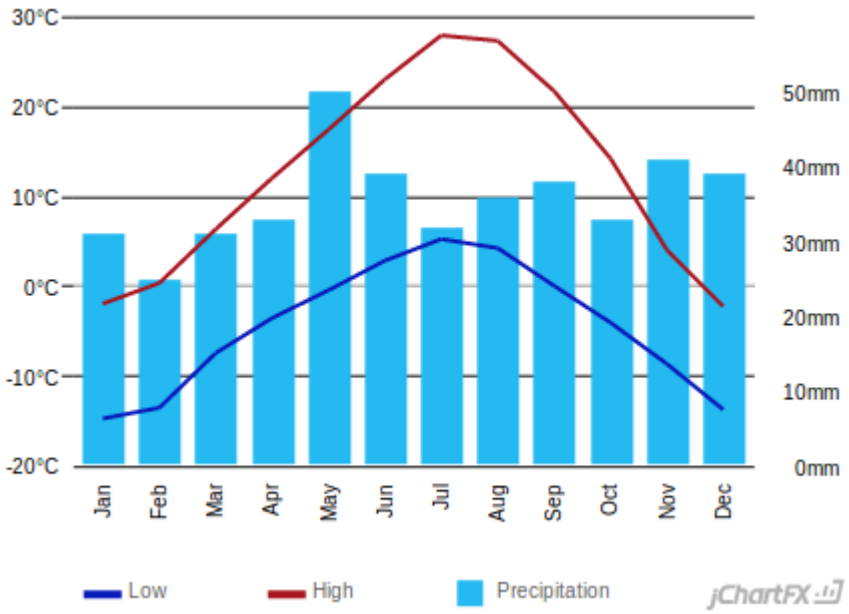


Data Details

The data was collected by stations associated with NOAA's Daily Global Historical Climatology Network. The data was filtered to only include stations within the United States. Stations were partitioned into 256 geographic rectangles and randomly assigned to analysts for analysis. Only years that had 50 or less missing entries (NaN) were included for each station. This data set was index SBBSSBSB and contained 12246 station/year pairs from statons with Wyoming. The weather data for each day for each station was stored as a 365 member vectors (i.e. numpy arrays) for efficiency and convenience of analysis.

Sanity-check: comparison with outside sources¶

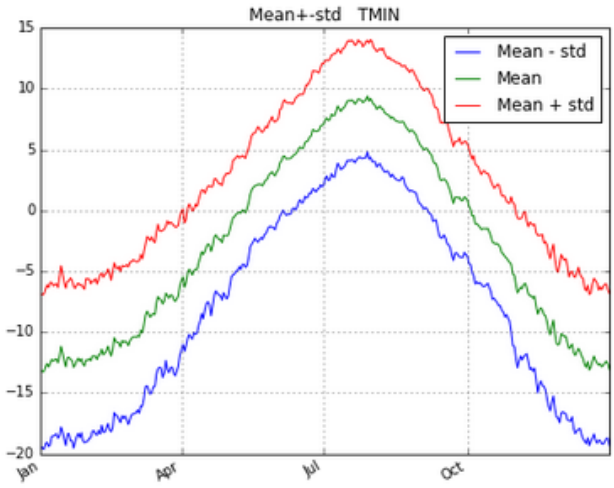
I start by comparing some of the general statistics with graphs that I obtained from a site called US Climate Data The graph below shows the daily minimum and maximum temperatures for each month, as well as the total precipitation for each month.

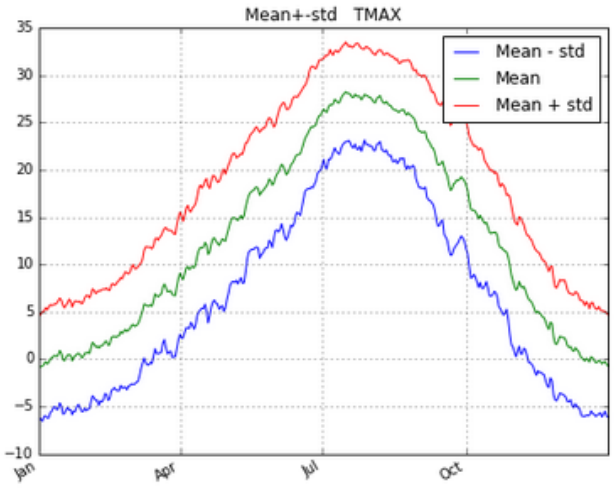


	Jan	Feb	Mar	Apr	May	Jun
Average high in °C:	-1.9	0.4	6.4	12.1	17.5	23.1
Average low in °C:	-14.7	-13.5	-7.4	-3.5	-0.4	2.9
Av. precipitation in mm:	31	25	31	33	50	39
Days with precipitation:	-	-	-	-	-	-
Hours of sunshine:	-	-	-	-	-	-
Average snowfall in cm:	40.6	27.9	17.8	5.1	2.5	0

	Jul	Aug	Sep	Oct	Nov	Dec
Average high in °C:	28	27.4	21.8	14.3	4.1	-2.2
Average low in °C:	5.3	4.3	0.1	-4	-8.6	-13.7
Av. precipitation in mm:	32	36	38	33	41	39
Days with precipitation:	-	-	-	-	-	-
Hours of sunshine:	-	-	-	-	-	-
Average snowfall in cm:	0	0	0	2.5	27.9	45.7

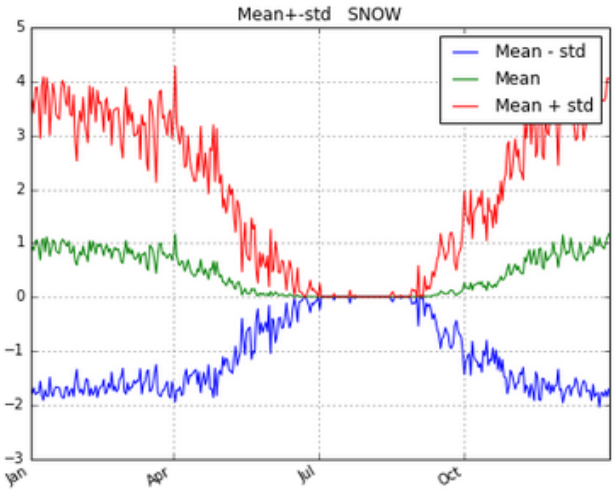
The study data agrees with the above web site for Temperature

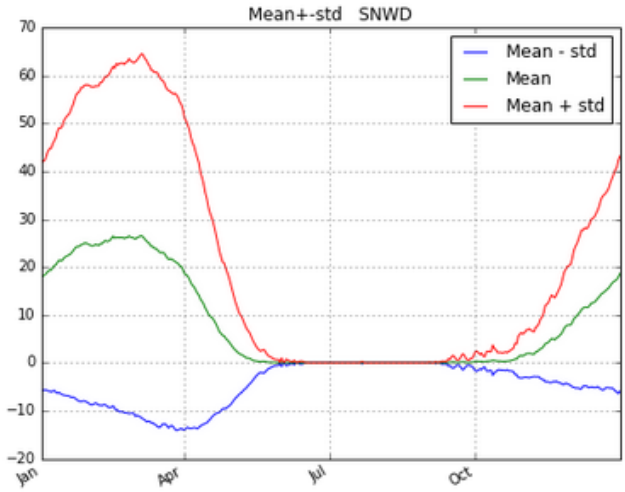




.

and for daily snowfall and average snow depth (both in cm)



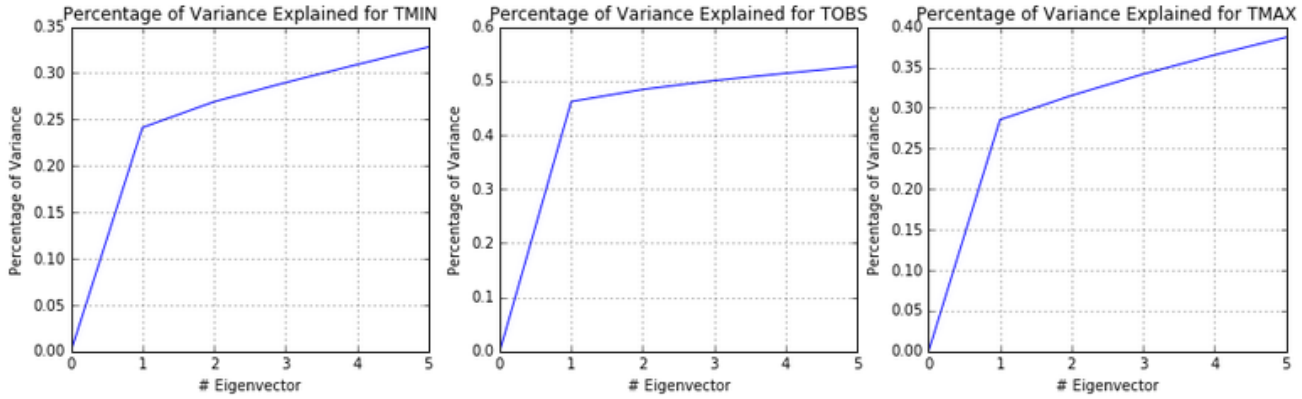


PCA analysis

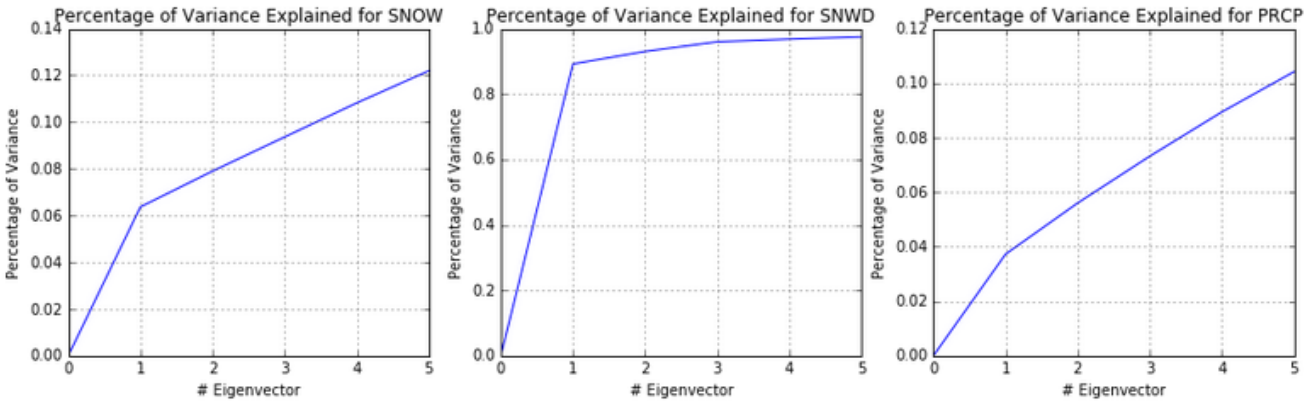
To reduce the determine the components that most predict each variable, Principal Components Analysis (PCA) was performed on each of the six variables. PCA transforms the data onto a new coordinate system in order to find components that explain the most variance. It is used to reduce dimensionality of large data sets to make analysis more manageable. Eigen vectors are the axes of the transformed data space.

For each of the six measurement, I computed the percentage of the variance explained as a function of the number of eigen vectors used.

Percentage of variance explained.



As shown above, the top 5 eigen vectors explain just over 30%, 50% and 30% for TMIN, TOBS and TMAX, respectively. The explained variance is still rising for all 3 variables and it was found that the percentage kept rising at a consistent rate until over 80% of the variance was explained for all 3 variables with 100 eigen vectors. While this is a significant reduction from 365, typically only the top several eigen vectors are used for analysis so the percentages above are somewhat lacking to continue with further analysis.



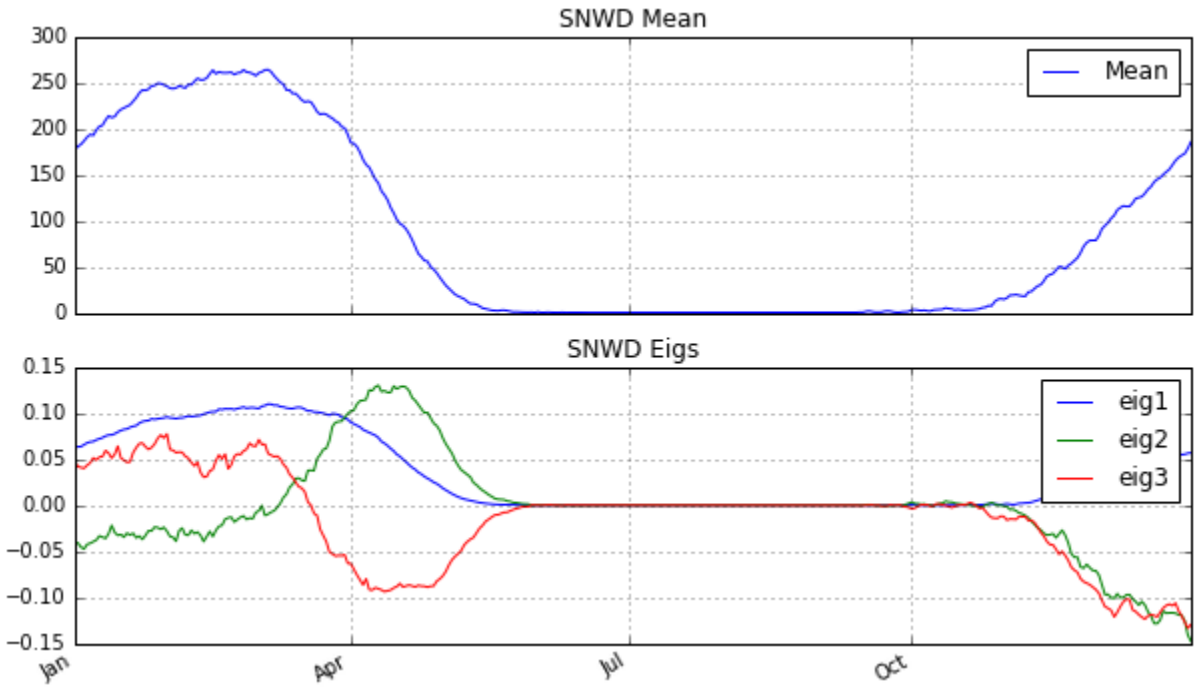
Of the second 3 variables, snow depth has the most variance explained with the fewest eigen vectors. The top 3 eigen vectors explain well over 90% of the total variance. It makes sense that snow depth would be more predictable because it varies less between days and between the same date in different years.

Analysis of snow depth

I chose to analyze the eigen-decomposition for snow-depth because the first 3 eigen-vectors explain over 90% of the variance.

Below is a graph of the mean and the top 3 eigen-vectors.

One can see that the snow season is from early November through early May, where the middle of February marks the peak of the snow-depth.

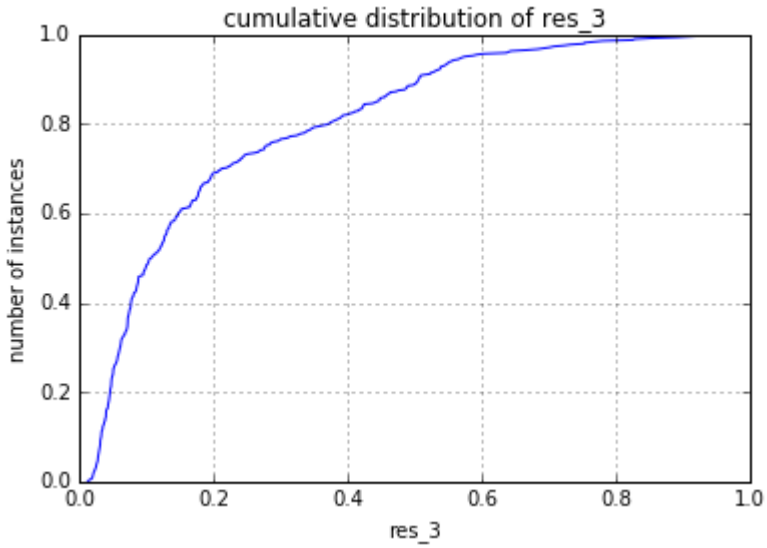


Interpreting the eigen vectors, it is clear that the first eigen vector closely follows the mean from January through May while slightly lagging behind the mean at the start of the season from November and December. The interpretation of this shape is that the first eigen vector represents the overall amount of snow above/below the mean, but without changing the distribution over time.

Eigen vectors 2 and 3 move between positive and negative values so they correspond to the changing distribution of the snow depth, but don't contribute as much to the snow depth total. Eigen vector 2 corresponds to more late season snow while eigen vector 3 corresponds to more snow in mid-winter. Both eigen vectors 2 and 3 correspond with less snow in early winter.

Residual Analysis

By looking at the cumulative distribution of residual 3 (the residual variance after the Mean and the first 2 eigen vectors have been subtracted out), one can see that over 90% of the instances have a residual 3 value of 0.6 or less.



Estimating the effect of the year vs the effect of the station

To estimate the effect of time vs. location on the first eigenvector coefficient for snow I computed:

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

Then computed the RMS (Root Mean Squared Error) before and after subtracting either the row or the column vector.

```
total RMS                   =  1122.69047274
RMS removing mean-by-station=  1006.7957443
RMS removing mean-by-year   =  702.159177023
```

This shows that the RMS decreased more when mean-by-year was removed which says that on average over all stations, the snow depth varied more by year than by station.