Part 1. EDA Highlights

Correlation

For solar production, the sun intensity, measured by time and sunlight strength, is the main factor that determines the solar production. The sun intensity is influenced by both geographic characteristics described in *station_info.csv*, and weather factors described in all the PCA weather predictors.

Statistical analysis on the solar production (measured in mean) and elevation (meters), solar production and latitude, are performed. Especially on latitudes, the interest is to investigate into whether latitude (degree changes within Oklahoma State) determines the hours of sunlight due to the Earth revolution around sun, a factor that directly influence the time under sunlight for each station.

From the statistical analysis result (figure 1.), **solar energy production and the elevation of the station are significantly correlated** with a correlation coefficient of 0.85 and p-value of 2.47607910^{-28} (which is less than the significance level alpha = 0.05).However, there is no significant correlation between solar energy production and the latitude as figure 2 shown below. **Therefore, elevation is selected to be a key variable to be considered in our model.**

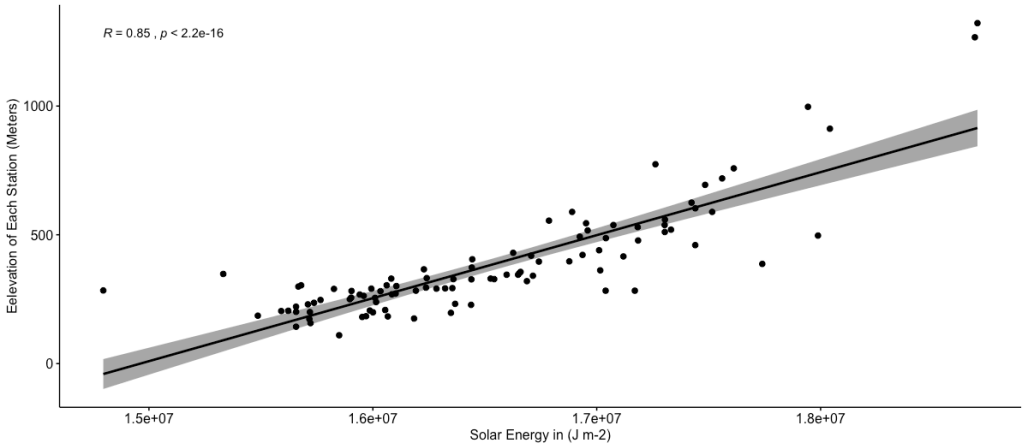


Figure.1 Correlation between solar energy production and elevation

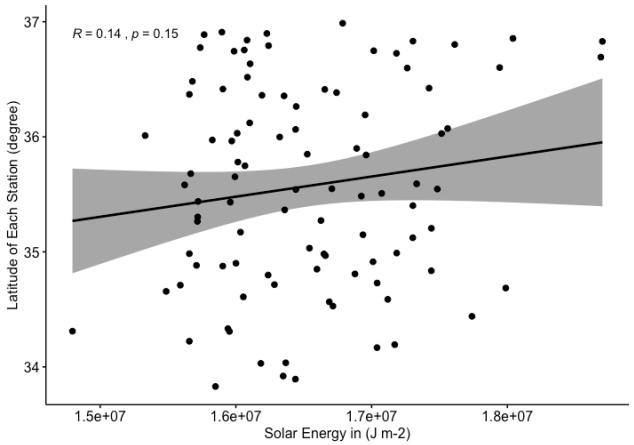


Figure.2 Correlation between solar energy production and latitude

Descriptive Statistics of PCA variables

For the weather predicators described by the PCA dataset (357 columns), main question to answer is whether all or some PCAs are included as the train dataset. A criterion of PCA selection needs to be set.

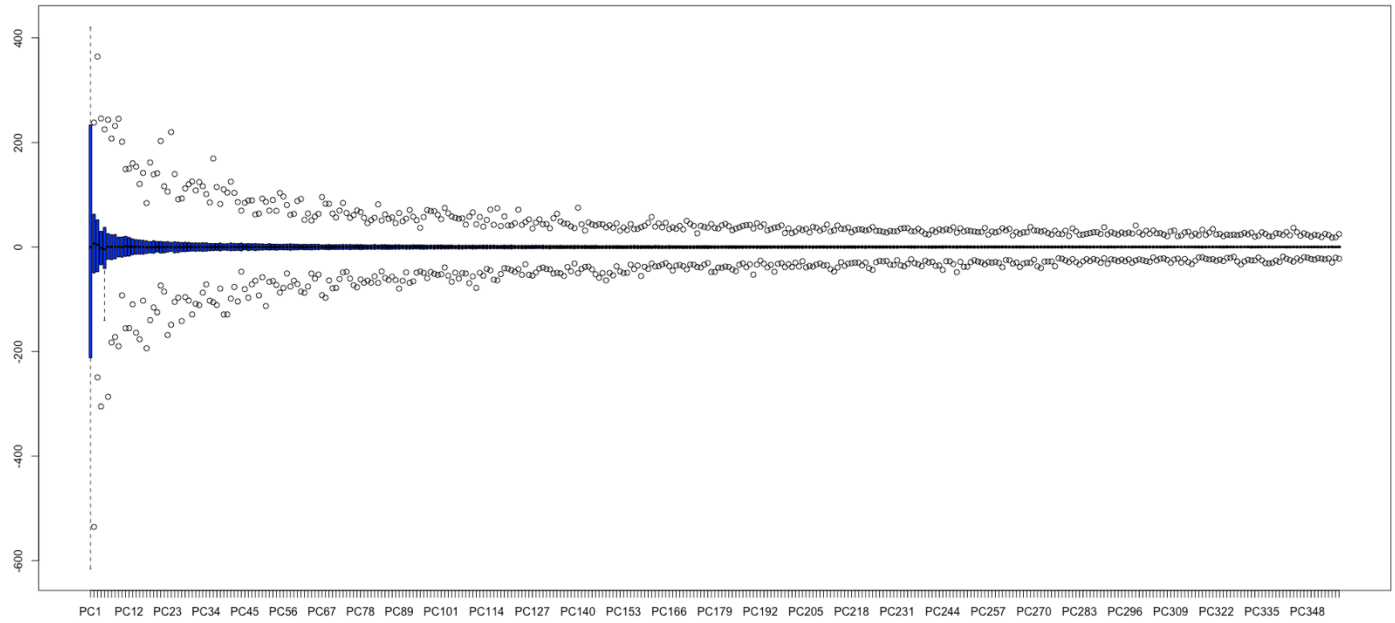


Figure.3 Descriptive analysis of 357 columns of PCAs

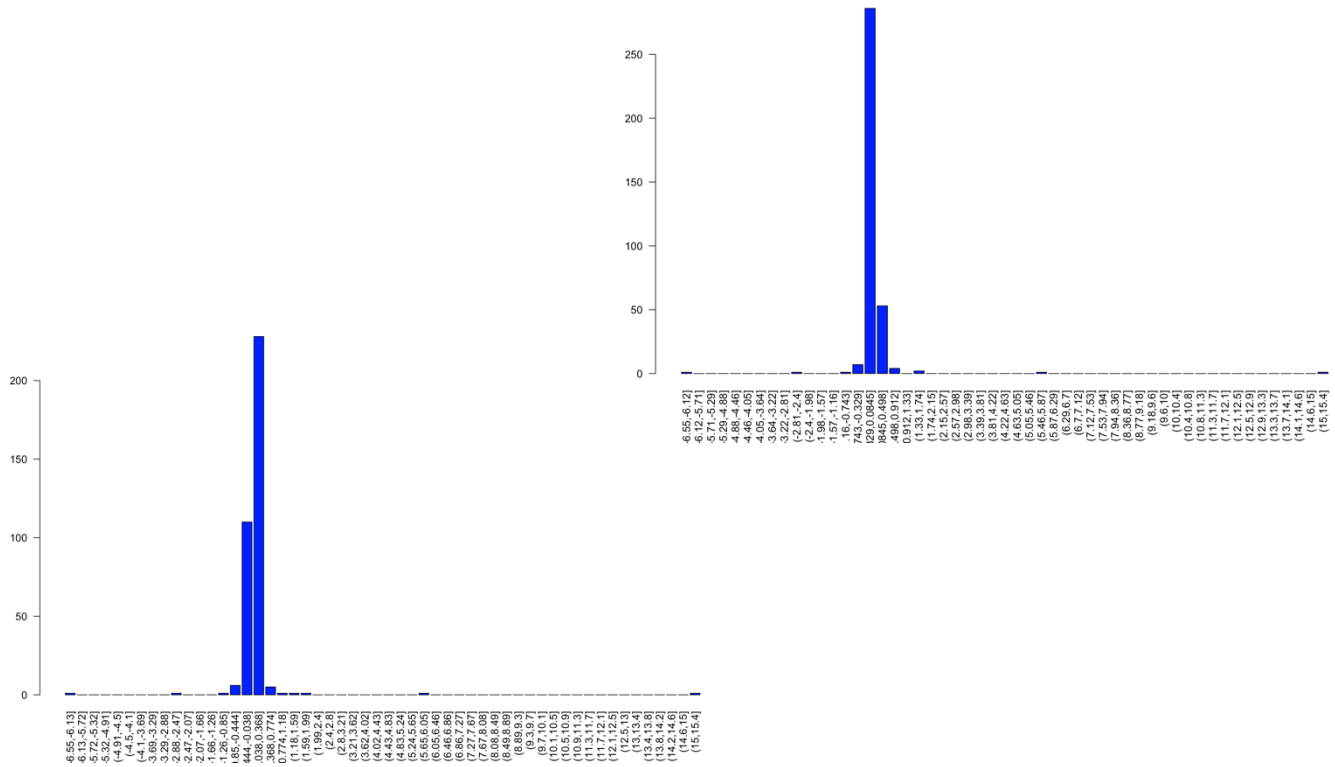


Figure.4 Analysis of statistical summary of all PCAs(top-right: Bin=53),(bottom-left: Bin=54)

The summary for each PCA column is done and plotted in figure 3.

Based on the PCA method, the first principal components show greater variance. The goal is to exclude as many near-0 PCAs as possible [*section [1.1.3] Reduce PCAs in code*]. Further steps are taken to **identify first 18 PCs to be the most "variant" PCs "away from the 0".**

Descriptive Statistics of Solar Production of stations

From 1994 to 2007, ***the Station IDAB*** has the highest solar production overtime. The station is chosen to demonstrate a 14-year span of solar production. The overall pattern indicates the change of seasons.

The criteria for outliers has further been identified based both on **graphical pattern** (circled in orange in figure.5) and the **numerical result from *outlier(),*** with a value of 39442800.

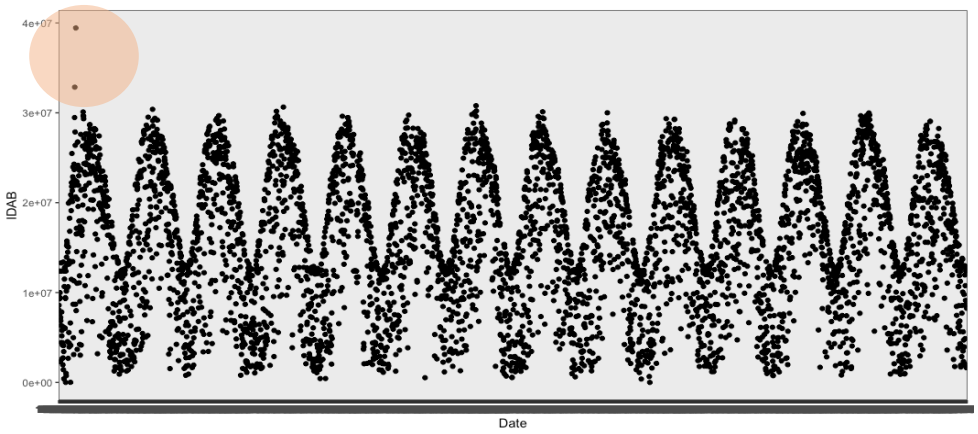


Figure.5 A 14-year span of solar production and its seasonal pattern at IDAB Station.

To future investigate into the smaller ranged outliers, we graphed Station ALTU with an outlier of 900. Many low-production points (marked by a straight line in figure 6) around the production level of global minimum are identified graphically, therefore we do not consider smallest numerical results from ***outlier()*** as outliers.

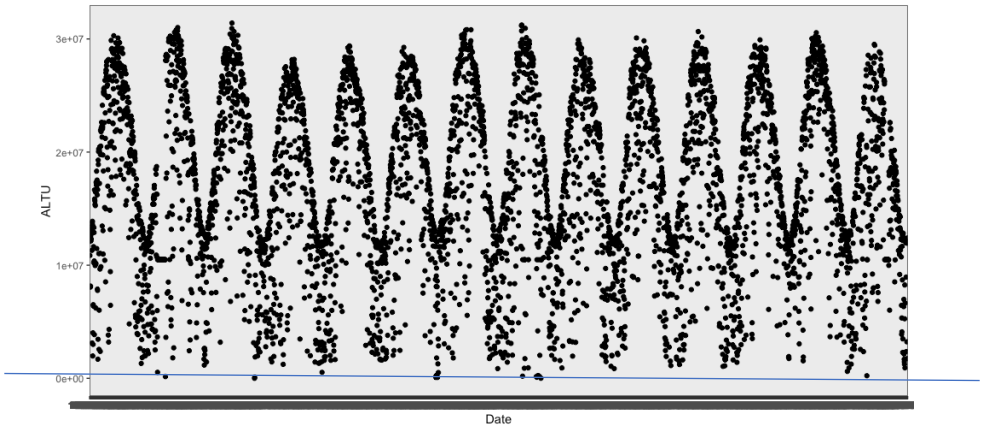


Figure.6 A 14-year span of solar production pattern at ALTU Station.

After plotting max of the result from ***outlier()*** (figure 7) that fall into the high production level, we further conclude that **ONLY two outlier points** identified in ***Station IDAB*** (marked in orange circle in figure 7) are outliers. We correct them with the local maximum of IDAB Station.

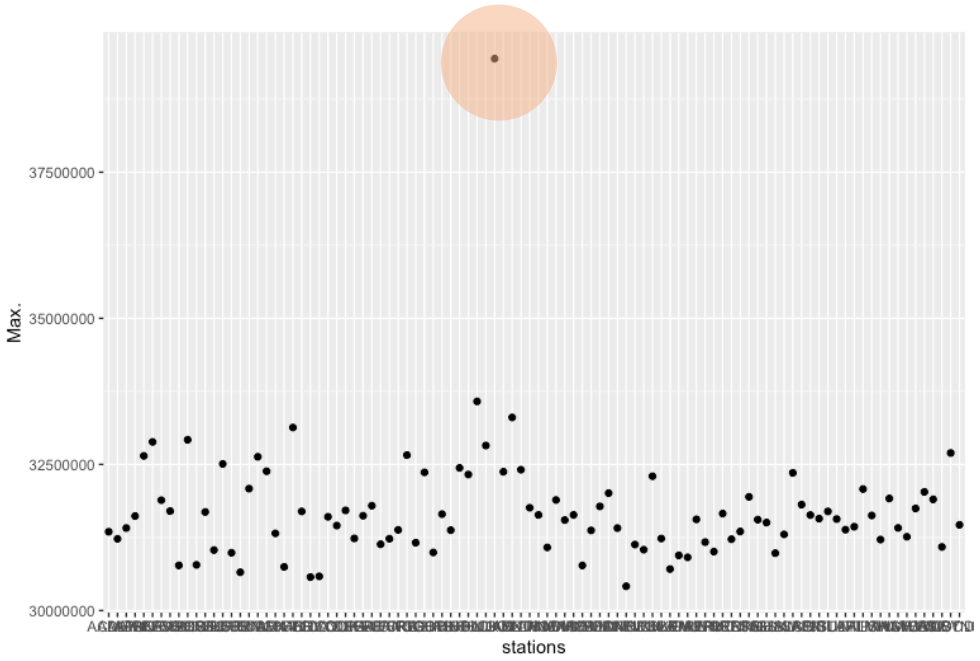
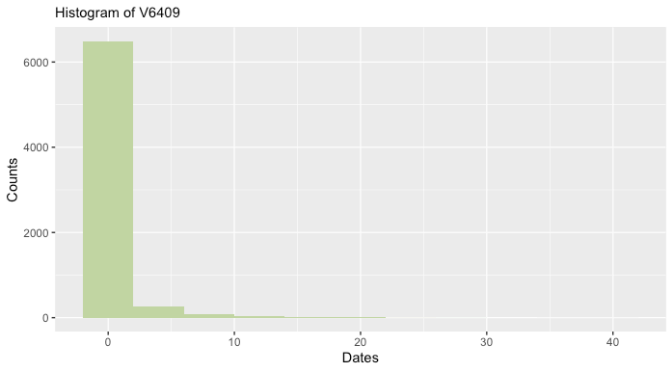
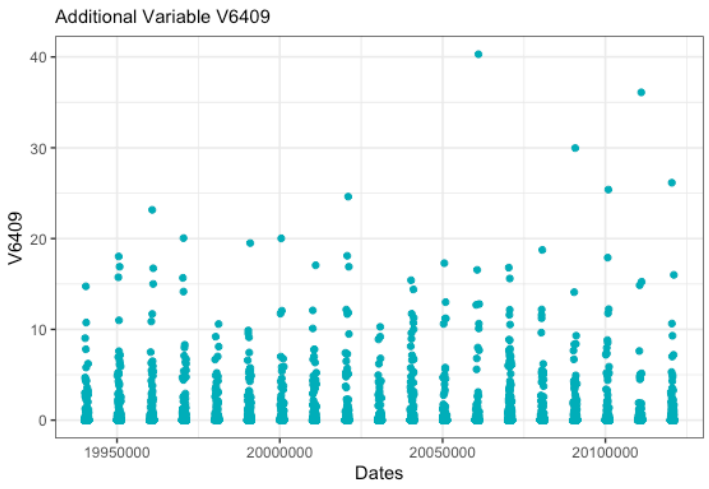


Figure.7 Plotting of outlier() results (max of the max).

Descriptive Statistics of Additional Variables





Mapping Visualization and basic GIS

Figure 7. is an interactive mapping of all stations, with the clustering option of freezeAtZoom = 6, a five clusters of stations are identified. This is used in the third model as a spatial clustering criterion.

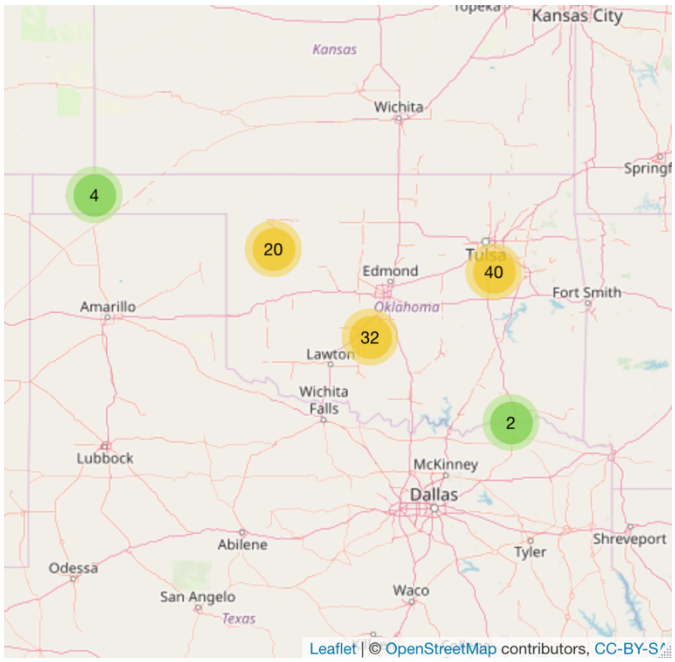


Figure.7 Basic mapping of all solar stations. For interactive map: http://rpubs.com/n_men/map

Though all stations have relatively equal circle sizes, spatially, **three geo-clusters of high-production solar farms are identified** (represented by overlapping of circles in Figure.8): Northeast of Lawton, Stillwater and Northwest of Lawton.

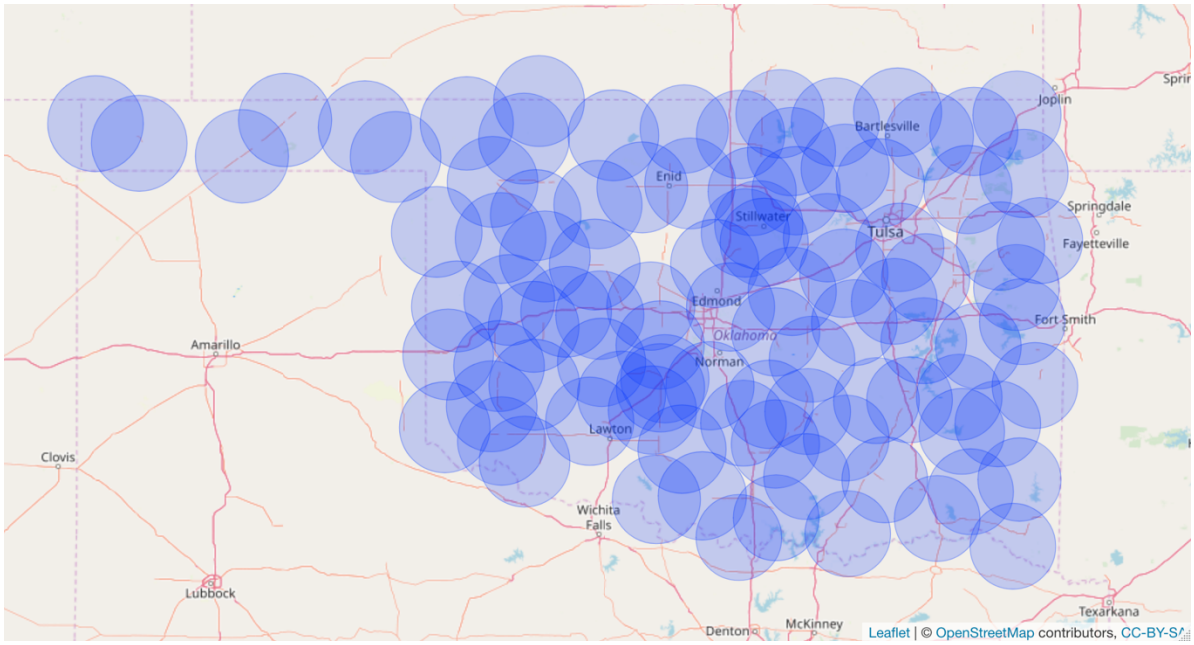


Figure. 8 A mapping on all stations with circle size representing sum of solar production.