Final Project

Group F

2021-11-06

Introduction

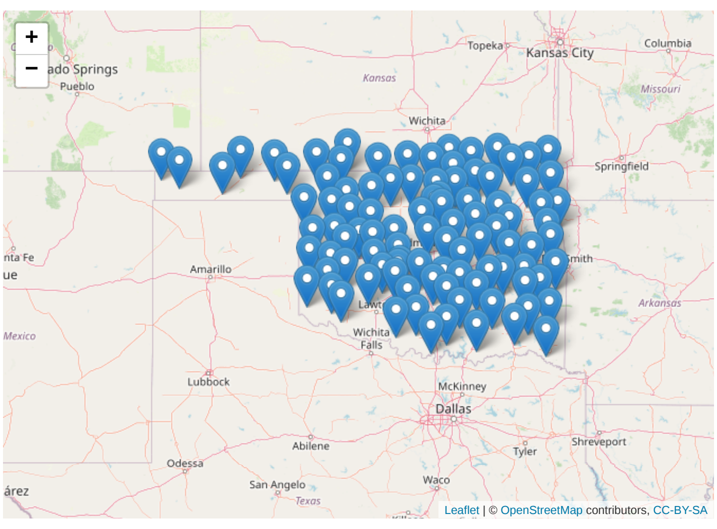
In this project we will work with a partly preprocessed dataset created from the original data given in the Kaggle platform corresponding to the “AMS 2013-2014 Solar Energy Prediction Contest”. For more information on the dataset it can be found here https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest/overview/evaluation.

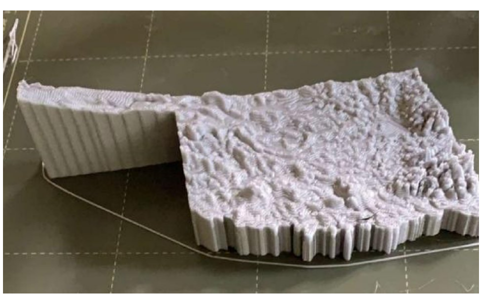
Dataset

- A total dimension of 6909 rows and 456 columns.
- Each row corresponds to information of a particular day, ranging from 1994-01-01 to 2012-11-30. The first column, ‘Date’, informs you of which day corresponds to each row.
- The next 98 columns (from 2nd to 99th position) gives the real values of solar production recorded in 98 different weather stations. These columns are only informed until 2007-12-31 (row 5113); after this date these 98 columns contain NA or missing values. These missing values we will attempt to **predict** to achieve the final goal of the project.
- The remaining columns are variables created from different weather predictors given in the Kaggle competition. They are the result of performing Principal Component Analysis, PCA, over the original data.

WEATHER STATIONS LOCATIONS

We thought it was important to visualize how spread out these 98 different whether stations are across the state of Oklahoma. We took advantage of the leaflet package in order to do this

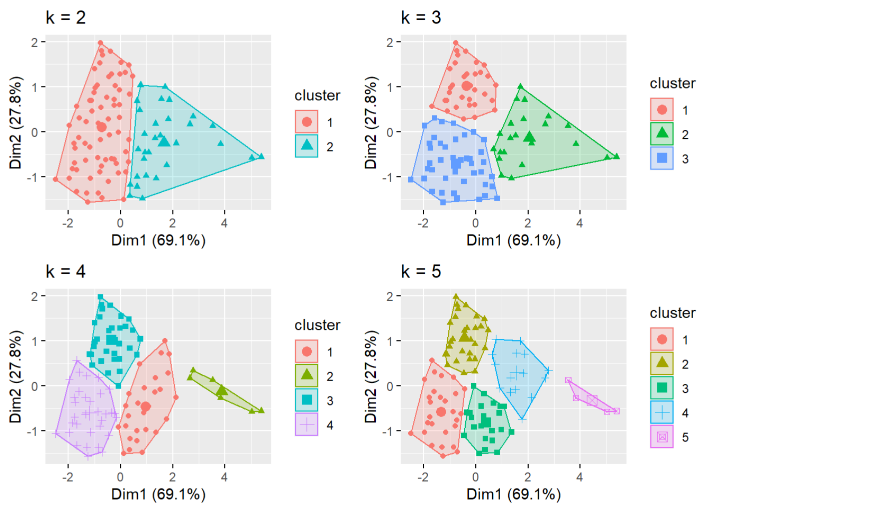




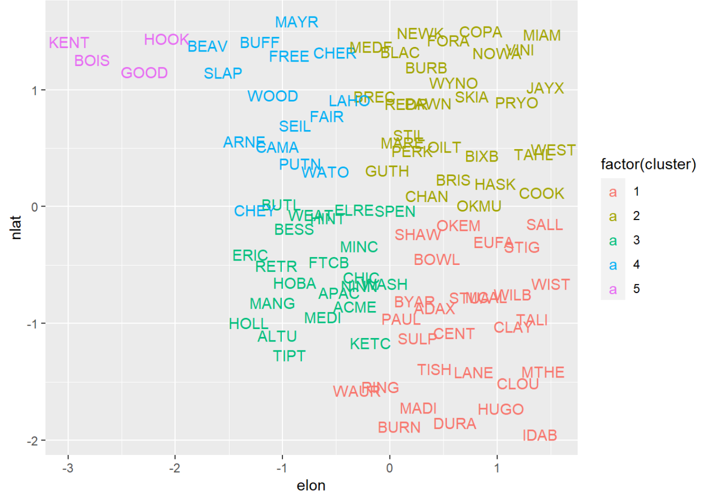
Elevation increases significantly in the northwest

EDA

Cluster Analysis to group weather stations beased on geographic coordinates

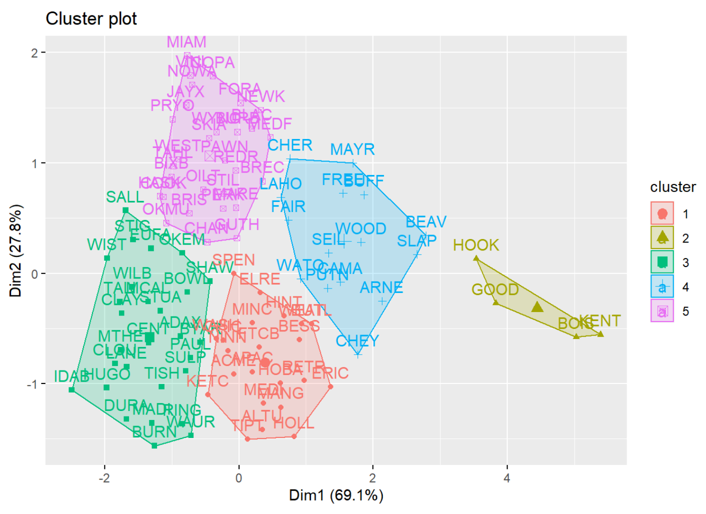


Clustering with k= 5



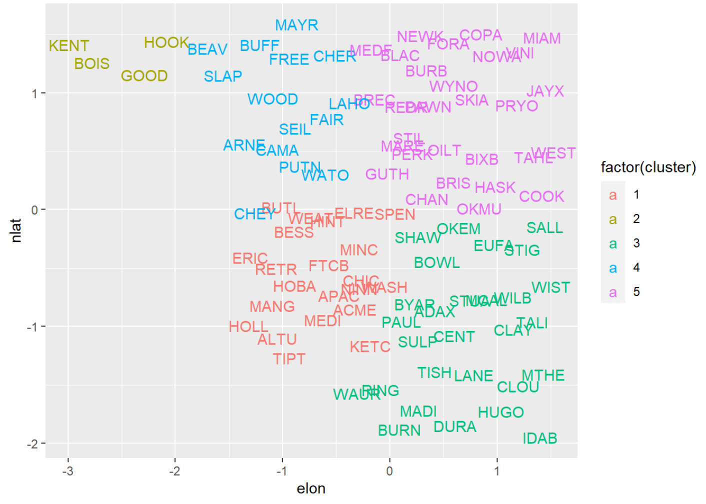
Based on these clusters, we decided to go wth 5 clusters since it is the one that best represents the Oklahoma map

```
## K-means clustering with 5 clusters of sizes 22, 4, 28, 15, 29
## 
## Cluster means:
##         nlat       elon       elev
## 1 -0.5728740 -0.6445489  0.2900088
## 2  1.3141019 -2.5323351  3.4817824
## 3 -1.0495895  0.7234280 -0.6998768
## 4  0.8787289 -0.9868012  0.9000444
## 5  0.8122205  0.6501879 -0.4900496
## 
## Clustering vector:
## ACME ADAX ALTU APAC ARNE BEAV BESS BIXB BLAC BOIS BOWL BREC BRIS BUFF BURB BURN 
##    1    3    1    1    4    4    1    5    5    2    3    5    5    4    5    3 
## BUTL BYAR CAMA CENT CHAN CHER CHEY CHIC CLAY CLOU COOK COPA DURA ELRE ERIC EUFA 
##    1    3    4    3    5    4    4    1    3    3    5    5    3    1    1    3 
## FAIR FORA FREE FTCB GOOD GUTH HASK HINT HOBA HOLL HOOK HUGO IDAB JAYX KENT KETC 
##    4    5    4    1    2    5    5    1    1    1    2    3    3    5    2    1 
## LAHO LANE MADI MANG MARE MAYR MCAL MEDF MEDI MIAM MINC MTHE NEWK NINN NOWA OILT 
##    4    3    3    1    5    4    3    5    1    5    1    3    5    1    5    5 
## OKEM OKMU PAUL PAWN PERK PRYO PUTN REDR RETR RING SALL SEIL SHAW SKIA SLAP SPEN 
##    3    5    3    5    5    5    4    5    1    3    3    4    3    5    4    1 
## STIG STIL STUA SULP TAHL TALI TIPT TISH VINI WASH WATO WAUR WEAT WEST WILB WIST 
##    3    5    3    3    5    3    1    3    5    1    4    3    1    5    3    3 
## WOOD WYNO 
##    4    5 
## 
## Within cluster sum of squares by cluster:
## [1]  9.299395  3.244412 17.614317 10.179914 15.491898
##  (between_SS / total_SS =  80.8 %)
## 
## Available components:
## 
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```




```
## # A tibble: 5 x 5
##   cluster   nlat   elon   elev stations
##     <int>  <dbl>  <dbl>  <dbl>    <dbl>
## 1       1 -0.573 -0.645  0.290       NA
## 2       2  1.31  -2.53   3.48        NA
## 3       3 -1.05   0.723 -0.700       NA
## 4       4  0.879 -0.987  0.900       NA
## 5       5  0.812  0.650 -0.490       NA
```

Final clusters




```
## ACME ADAX ALTU APAC ARNE BEAV BESS BIXB BLAC BOIS BOWL BREC BRIS BUFF BURB BURN 
##    1    3    1    1    4    4    1    5    5    2    3    5    5    4    5    3 
## BUTL BYAR CAMA CENT CHAN CHER CHEY CHIC CLAY CLOU COOK COPA DURA ELRE ERIC EUFA 
##    1    3    4    3    5    4    4    1    3    3    5    5    3    1    1    3 
## FAIR FORA FREE FTCB GOOD GUTH HASK HINT HOBA HOLL HOOK HUGO IDAB JAYX KENT KETC 
##    4    5    4    1    2    5    5    1    1    1    2    3    3    5    2    1 
## LAHO LANE MADI MANG MARE MAYR MCAL MEDF MEDI MIAM MINC MTHE NEWK NINN NOWA OILT 
##    4    3    3    1    5    4    3    5    1    5    1    3    5    1    5    5 
## OKEM OKMU PAUL PAWN PERK PRYO PUTN REDR RETR RING SALL SEIL SHAW SKIA SLAP SPEN 
##    3    5    3    5    5    5    4    5    1    3    3    4    3    5    4    1 
## STIG STIL STUA SULP TAHL TALI TIPT TISH VINI WASH WATO WAUR WEAT WEST WILB WIST 
##    3    5    3    3    5    3    1    3    5    1    4    3    1    5    3    3 
## WOOD WYNO 
##    4    5
```

Summary

```
##      Date                ACME               ADAX               ALTU         
##  Length:6909        Min.   :   12000   Min.   :  510000   Min.   :     900  
##  Class :character   1st Qu.:11404200   1st Qu.:10611000   1st Qu.:11674500  
##  Mode  :character   Median :16946400   Median :16299300   Median :17073600  
##                     Mean   :16877462   Mean   :16237534   Mean   :17119189  
##                     3rd Qu.:23734800   3rd Qu.:23027400   3rd Qu.:23903700  
##                     Max.   :31347900   Max.   :31227000   Max.   :31411500  
##                     NA's   :1796       NA's   :1796       NA's   :1796      
##       APAC               ARNE               BEAV               BESS         
##  Min.   :    3300   Min.   :  477300   Min.   :     300   Min.   :  510600  
##  1st Qu.:11637000   1st Qu.:11666400   1st Qu.:11493600   1st Qu.:11712600  
##  Median :17062500   Median :17578500   Median :17520900   Median :17176500  
##  Mean   :17010565   Mean   :17560173   Mean   :17612143   Mean   :17304074  
##  3rd Qu.:23909400   3rd Qu.:24503700   3rd Qu.:24683100   3rd Qu.:24241200  
##  Max.   :31616100   Max.   :32645700   Max.   :32884800   Max.   :31887900  
##  NA's   :1796       NA's   :1796       NA's   :1796       NA's   :1796      
```

```
##
##      PC353               PC354               PC355
##  Min.   :-23.66467   Min.   :-21.90907   Min.   :-29.65249
##  1st Qu.: -1.58636   1st Qu.: -1.66117   1st Qu.: -1.60810
##  Median : -0.07761   Median :  0.01887   Median :  0.07844
##  Mean   : -0.02131   Mean   : -0.04450   Mean   :  0.04149
##  3rd Qu.:  1.52298   3rd Qu.:  1.56665   3rd Qu.:  1.74447
##  Max.   : 25.13188   Max.   : 22.87146   Max.   : 17.98370
##
##      PC356               PC357
##  Min.   :-20.50932   Min.   :-22.30097
##  1st Qu.: -1.56128   1st Qu.: -1.52905
##  Median : -0.03687   Median :  0.04224
##  Mean   : -0.02142   Mean   :  0.06599
##  3rd Qu.:  1.51265   3rd Qu.:  1.61865
##  Max.   : 18.59593   Max.   : 24.75692
##
```

Counting missing values

```
count_nas <- function(x){
  ret <- sum(is.na(x));
  return(ret);
}

sapply(dat, count_nas);
```

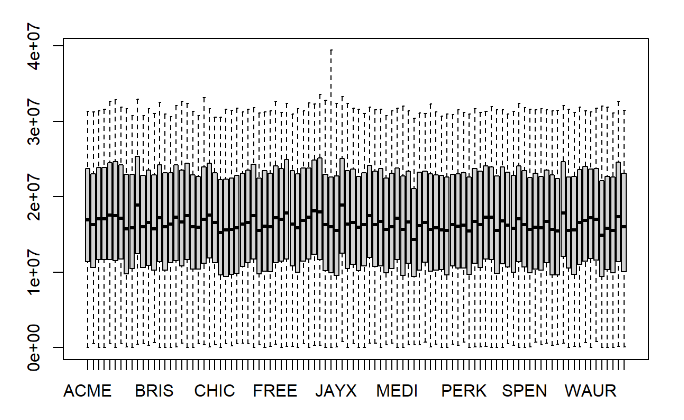
```
##  Date  ACME  ADAX  ALTU  APAC  ARNE  BEAV  BESS  BIXB  BLAC  BOIS  BOWL  BREC
##     0  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796
##  BRIS  BUFF  BURB  BURN  BUTL  BYAR  CAMA  CENT  CHAN  CHER  CHEY  CHIC  CLAY
##  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796
##  CLOU  COOK  COPA  DURA  ELRE  ERIC  EUFA  FAIR  FORA  FREE  FTCB  GOOD  GUTH
##  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796
##  HASK  HINT  HOBA  HOLL  HOOK  HUGO  IDAB  JAYX  KENT  KETC  LAHO  LANE  MADI
##  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796
##  MANG  MARE  MAYR  MCAL  MEDF  MEDI  MIAM  MINC  MTHE  NEWK  NINN  NOWA  OILT
##  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796
##  OKEM  OKMU  PAUL  PAWN  PERK  PRYO  PUTN  REDR  RETR  RING  SALL  SEIL  SHAW
##  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796
##  SKIA  SLAP  SPEN  STIG  STIL  STUA  SULP  TAHL  TALI  TIPT  TISH  VINI  WASH
##  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796  1796
##  WATO  WAUR  WEAT  WEST  WILB  WIST  WOOD  WYNO   PC1   PC2   PC3   PC4   PC5
##  1796  1796  1796  1796  1796  1796  1796  1796     0     0     0     0     0
##   PC6   PC7   PC8   PC9  PC10  PC11  PC12  PC13  PC14  PC15  PC16  PC17  PC18
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  PC19  PC20  PC21  PC22  PC23  PC24  PC25  PC26  PC27  PC28  PC29  PC30  PC31
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  PC32  PC33  PC34  PC35  PC36  PC37  PC38  PC39  PC40  PC41  PC42  PC43  PC44
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  PC45  PC46  PC47  PC48  PC49  PC50  PC51  PC52  PC53  PC54  PC55  PC56  PC57
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  PC58  PC59  PC60  PC61  PC62  PC63  PC64  PC65  PC66  PC67  PC68  PC69  PC70
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  PC71  PC72  PC73  PC74  PC75  PC76  PC77  PC78  PC79  PC80  PC81  PC82  PC83
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  PC84  PC85  PC86  PC87  PC88  PC89  PC90  PC91  PC92  PC93  PC94  PC95  PC96
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  PC97  PC98  PC99 PC100 PC101 PC102 PC103 PC104 PC105 PC106 PC107 PC108 PC109
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC110 PC111 PC112 PC113 PC114 PC115 PC116 PC117 PC118 PC119 PC120 PC121 PC122
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC123 PC124 PC125 PC126 PC127 PC128 PC129 PC130 PC131 PC132 PC133 PC134 PC135
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC136 PC137 PC138 PC139 PC140 PC141 PC142 PC143 PC144 PC145 PC146 PC147 PC148
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC149 PC150 PC151 PC152 PC153 PC154 PC155 PC156 PC157 PC158 PC159 PC160 PC161
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC162 PC163 PC164 PC165 PC166 PC167 PC168 PC169 PC170 PC171 PC172 PC173 PC174
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC175 PC176 PC177 PC178 PC179 PC180 PC181 PC182 PC183 PC184 PC185 PC186 PC187
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC188 PC189 PC190 PC191 PC192 PC193 PC194 PC195 PC196 PC197 PC198 PC199 PC200
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC201 PC202 PC203 PC204 PC205 PC206 PC207 PC208 PC209 PC210 PC211 PC212 PC213
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC214 PC215 PC216 PC217 PC218 PC219 PC220 PC221 PC222 PC223 PC224 PC225 PC226
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC227 PC228 PC229 PC230 PC231 PC232 PC233 PC234 PC235 PC236 PC237 PC238 PC239
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC240 PC241 PC242 PC243 PC244 PC245 PC246 PC247 PC248 PC249 PC250 PC251 PC252
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC253 PC254 PC255 PC256 PC257 PC258 PC259 PC260 PC261 PC262 PC263 PC264 PC265
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC266 PC267 PC268 PC269 PC270 PC271 PC272 PC273 PC274 PC275 PC276 PC277 PC278
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC279 PC280 PC281 PC282 PC283 PC284 PC285 PC286 PC287 PC288 PC289 PC290 PC291
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC292 PC293 PC294 PC295 PC296 PC297 PC298 PC299 PC300 PC301 PC302 PC303 PC304
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC305 PC306 PC307 PC308 PC309 PC310 PC311 PC312 PC313 PC314 PC315 PC316 PC317
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC318 PC319 PC320 PC321 PC322 PC323 PC324 PC325 PC326 PC327 PC328 PC329 PC330
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC331 PC332 PC333 PC334 PC335 PC336 PC337 PC338 PC339 PC340 PC341 PC342 PC343
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC344 PC345 PC346 PC347 PC348 PC349 PC350 PC351 PC352 PC353 PC354 PC355 PC356
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC357
##     0
```

removing rows with missing DATA

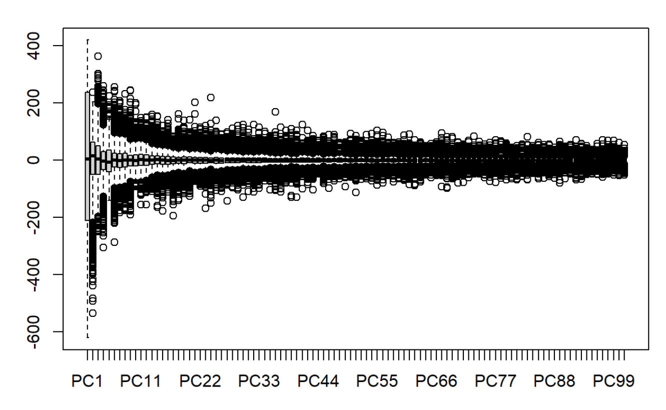
```
df_red <- df[1:5113, ]
sapply(df_red, function(x){sum(is.na(x))});
```

```
##  Date  ACME  ADAX  ALTU  APAC  ARNE  BEAV  BESS  BIXB  BLAC  BOIS  BOWL  BREC
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  BRIS  BUFF  BURB  BURN  BUTL  BYAR  CAMA  CENT  CHAN  CHER  CHEY  CHIC  CLAY
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  CLOU  COOK  COPA  DURA  ELRE  ERIC  EUFA  FAIR  FORA  FREE  FTCB  GOOD  GUTH
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  HASK  HINT  HOBA  HOLL  HOOK  HUGO  IDAB  JAYX  KENT  KETC  LAHO  LANE  MADI
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  MANG  MARE  MAYR  MCAL  MEDF  MEDI  MIAM  MINC  MTHE  NEWK  NINN  NOWA  OILT
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  OKEM  OKMU  PAUL  PAWN  PERK  PRYO  PUTN  REDR  RETR  RING  SALL  SEIL  SHAW
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  SKIA  SLAP  SPEN  STIG  STIL  STUA  SULP  TAHL  TALI  TIPT  TISH  VINI  WASH
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  WATO  WAUR  WEAT  WEST  WILB  WIST  WOOD  WYNO   PC1   PC2   PC3   PC4   PC5
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##   PC6   PC7   PC8   PC9  PC10  PC11  PC12  PC13  PC14  PC15  PC16  PC17  PC18
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  PC19  PC20  PC21  PC22  PC23  PC24  PC25  PC26  PC27  PC28  PC29  PC30  PC31
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  PC32  PC33  PC34  PC35  PC36  PC37  PC38  PC39  PC40  PC41  PC42  PC43  PC44
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  PC45  PC46  PC47  PC48  PC49  PC50  PC51  PC52  PC53  PC54  PC55  PC56  PC57
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  PC58  PC59  PC60  PC61  PC62  PC63  PC64  PC65  PC66  PC67  PC68  PC69  PC70
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  PC71  PC72  PC73  PC74  PC75  PC76  PC77  PC78  PC79  PC80  PC81  PC82  PC83
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  PC84  PC85  PC86  PC87  PC88  PC89  PC90  PC91  PC92  PC93  PC94  PC95  PC96
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##  PC97  PC98  PC99 PC100 PC101 PC102 PC103 PC104 PC105 PC106 PC107 PC108 PC109
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC110 PC111 PC112 PC113 PC114 PC115 PC116 PC117 PC118 PC119 PC120 PC121 PC122
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC123 PC124 PC125 PC126 PC127 PC128 PC129 PC130 PC131 PC132 PC133 PC134 PC135
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC136 PC137 PC138 PC139 PC140 PC141 PC142 PC143 PC144 PC145 PC146 PC147 PC148
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC149 PC150 PC151 PC152 PC153 PC154 PC155 PC156 PC157 PC158 PC159 PC160 PC161
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC162 PC163 PC164 PC165 PC166 PC167 PC168 PC169 PC170 PC171 PC172 PC173 PC174
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC175 PC176 PC177 PC178 PC179 PC180 PC181 PC182 PC183 PC184 PC185 PC186 PC187
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC188 PC189 PC190 PC191 PC192 PC193 PC194 PC195 PC196 PC197 PC198 PC199 PC200
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC201 PC202 PC203 PC204 PC205 PC206 PC207 PC208 PC209 PC210 PC211 PC212 PC213
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC214 PC215 PC216 PC217 PC218 PC219 PC220 PC221 PC222 PC223 PC224 PC225 PC226
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC227 PC228 PC229 PC230 PC231 PC232 PC233 PC234 PC235 PC236 PC237 PC238 PC239
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC240 PC241 PC242 PC243 PC244 PC245 PC246 PC247 PC248 PC249 PC250 PC251 PC252
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC253 PC254 PC255 PC256 PC257 PC258 PC259 PC260 PC261 PC262 PC263 PC264 PC265
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC266 PC267 PC268 PC269 PC270 PC271 PC272 PC273 PC274 PC275 PC276 PC277 PC278
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC279 PC280 PC281 PC282 PC283 PC284 PC285 PC286 PC287 PC288 PC289 PC290 PC291
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC292 PC293 PC294 PC295 PC296 PC297 PC298 PC299 PC300 PC301 PC302 PC303 PC304
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC305 PC306 PC307 PC308 PC309 PC310 PC311 PC312 PC313 PC314 PC315 PC316 PC317
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC318 PC319 PC320 PC321 PC322 PC323 PC324 PC325 PC326 PC327 PC328 PC329 PC330
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC331 PC332 PC333 PC334 PC335 PC336 PC337 PC338 PC339 PC340 PC341 PC342 PC343
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC344 PC345 PC346 PC347 PC348 PC349 PC350 PC351 PC352 PC353 PC354 PC355 PC356
##     0     0     0     0     0     0     0     0     0     0     0     0     0
## PC357
##     0
```

Visualize data distribution of weather stations



Visualize data distribution of top 100 Principal Components



It is evidenced in this plot that PC2 & above have plenty of outliers

Finding Outliers

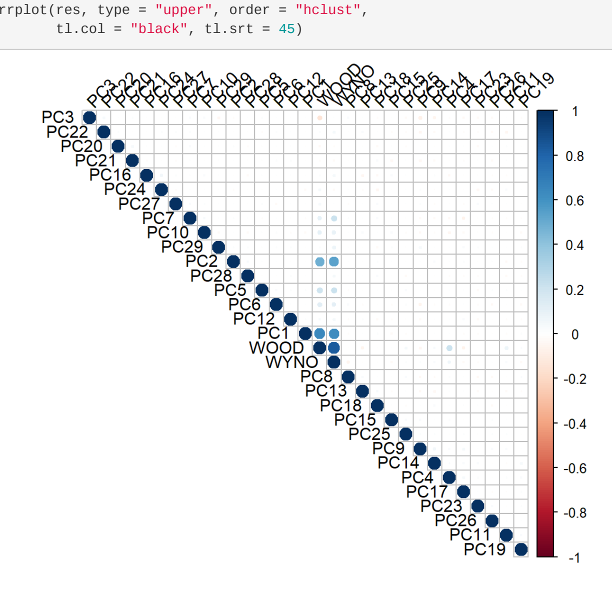
```
##          ACME          ADAX          ALTU          APAC          ARNE
##  1.200000e+04  5.100000e+05  9.000000e+02  3.300000e+03  4.773000e+05
##          BEAV          BESS          BIXB          BLAC          BOIS
##  3.000000e+02  5.106000e+05  7.470000e+04  3.600000e+04  4.671000e+05
##          BOWL          BREC          BRIS          BUFF          BURB
##  4.977000e+05  2.688000e+05  6.426000e+05  1.200000e+03  3.210000e+04
##          BURN          BUTL          BYAR          CAMA          CENT
##  3.000000e+02  1.017000e+05  5.457000e+05  6.300000e+03  1.500000e+03
##          CHAN          CHER          CHEY          CHIC          CLAY
##  5.190000e+05  3.993000e+05  5.580000e+04  3.945000e+05  1.230000e+04
##          CLOU          COOK          COPA          DURA          ELRE
##  3.160350e+07  3.145350e+07  3.171360e+07  5.427000e+05  4.794000e+05
##          ERIC          EUFA          FAIR          FORA          FREE
##  4.200000e+03  3.113400e+07  3.690000e+04  1.239000e+05  4.146000e+05
##          FTCB          GOOD          GUTH          HASK          HINT
##  4.380000e+04  1.281000e+05  1.281000e+05  1.500000e+03  5.052000e+05
##          HOBA          HOLL          HOOK          HUGO          IDAB
##  3.600000e+03  3.291000e+05  3.201000e+05  3.282300e+07  3.944280e+07
##          JAYX          KENT          KETC          LAHO          LANE
##  3.237450e+07  7.938000e+05  2.730000e+04  5.202000e+05  3.000000e+02
##          MADI          MANG          MARE          MAYR          MCAL
##  6.600000e+03  5.511000e+05  5.640000e+05  1.500000e+03  3.774000e+05
##          MEDF          MEDI          MIAM          MINC          MTHE
##  5.802000e+05  1.500000e+03  3.200880e+07  3.864000e+05  3.041520e+07
##          NEWK          NINN          NOWA          OILT          OKEM
##  3.840000e+05  6.861000e+05  3.229740e+07  4.227000e+05  1.200000e+03
##          OKMU          PAUL          PAWN          PERK          PRYO
##  3.000000e+02  4.437000e+05  2.949000e+05  6.987000e+05  3.000000e+02
##          PUTN          REDR          RETR          RING          SALL
##  6.600000e+04  7.830000e+04  1.365000e+05  1.800000e+03  3.155520e+07
##          SEIL          SHAW          SKIA          SLAP          SPEN
##  2.400000e+04  4.794000e+05  3.069000e+05  1.350000e+04  1.560000e+04
##          STIG          STIL          STUA          SULP          TAHL
##  3.163350e+07  3.157290e+07  3.169590e+07  5.691000e+05  3.138360e+07
##          TALI          TIPT          TISH          VINI          WASH
##  3.143400e+07  5.760000e+05  9.000000e+02  1.746000e+05  7.800000e+04
##          WATO          WAUR          WEAT          WEST          WILB
##  6.162000e+05  3.300000e+03  7.593000e+05  3.203010e+07  6.300000e+03
##          WIST          WOOD          WYNO           PC1           PC2
##  1.680000e+04  1.884000e+05  8.700000e+04 -6.158103e+02 -5.355114e+02
##           PC3           PC4           PC5           PC6           PC7
##  3.642267e+02 -3.050638e+02  2.250383e+02 -2.865765e+02  2.074084e+02
##           PC8           PC9          PC10          PC11          PC12
##  2.317553e+02  2.453236e+02  2.012154e+02 -1.553629e+02 -1.551643e+02
##          PC13          PC14          PC15          PC16          PC17
##  1.600801e+02 -1.640057e+02 -1.764960e+02  1.417673e+02 -1.937855e+02
##          PC18          PC19          PC20          PC21          PC22
##  1.616515e+02  1.388195e+02  1.410499e+02  2.027049e+02  1.161806e+02
##          PC23          PC24          PC25          PC26          PC27
## -1.683955e+02  2.199148e+02  1.395319e+02 -9.702311e+01 -1.417869e+02
##          PC28          PC29          PC30          PC31          PC32
##  1.120534e+02  1.200663e+02 -1.290919e+02 -1.087599e+02  1.243084e+02
##          PC33          PC34          PC35          PC36          PC37
##  1.162189e+02  1.012680e+02 -1.027788e+02  1.692115e+02  1.147545e+02
##          PC38          PC39          PC40          PC41          PC42
##  8.228080e+01 -1.291575e+02 -1.290914e+02  1.249909e+02  1.035214e+02
##          PC43          PC44          PC45          PC46          PC47
## -1.042286e+02  6.959360e+01  8.471449e+01 -9.715078e+01  8.903736e+01
##          PC48          PC49          PC50          PC51          PC52
## -6.455406e+01 -9.266506e+01  9.256876e+01 -1.128904e+02  6.967469e+01
##          PC53          PC54          PC55          PC56          PC57
##  8.994106e+01 -7.266175e+01  1.033283e+02  9.668562e+01  8.041210e+01
##          PC58          PC59          PC60          PC61          PC62
## -7.561729e+01 -6.448290e+01  8.791393e+01  9.188996e+01 -8.758565e+01
##          PC63          PC64          PC65          PC66          PC67
## -7.563266e+01  5.065436e+01 -6.041020e+01  6.339131e+01  9.556350e+01
##          PC68          PC69          PC70          PC71          PC72
## -9.707587e+01  8.272304e+01 -7.896915e+01 -7.826387e+01  6.975846e+01
##          PC73          PC74          PC75          PC76          PC77
##  8.428817e+01  6.518741e+01 -6.044956e+01 -7.308455e+01 -7.695525e+01
##          PC78          PC79          PC80          PC81          PC82
##  6.674916e+01 -6.843903e+01 -6.448964e+01 -6.826343e+01 -5.616706e+01
##          PC83          PC84          PC85          PC86          PC87
##  8.152141e+01  5.064356e+01  6.237217e+01 -6.317895e+01 -5.674073e+01
##          PC88          PC89          PC90          PC91          PC92
## -6.322324e+01 -7.966507e+01 -6.470725e+01  5.431693e+01  7.076370e+01
##          PC93          PC94          PC95          PC96          PC97
## -6.574163e+01  5.110237e+01 -4.663167e+01  5.728569e+01  7.087758e+01
##          PC98          PC99         PC100
##  6.867961e+01  6.868623e+01  6.178891e+01
```

Create a function to tipify the dataset (substract the mean & divide by standard deviation)

Correlation Analysis

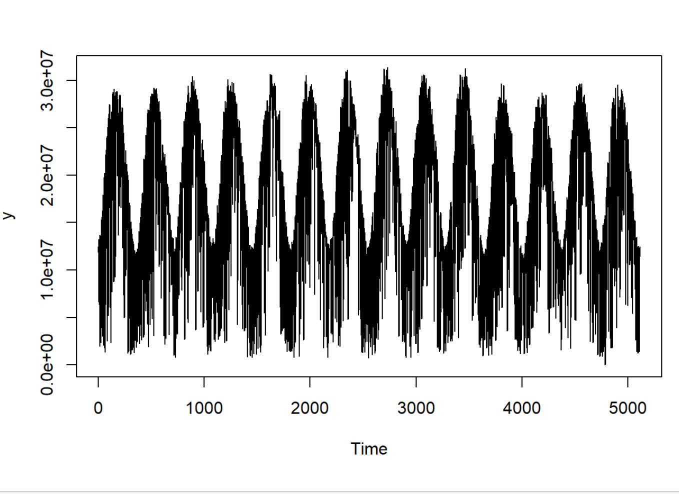
Correlation plot

```
library(corrplot)
corrplot(res, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```

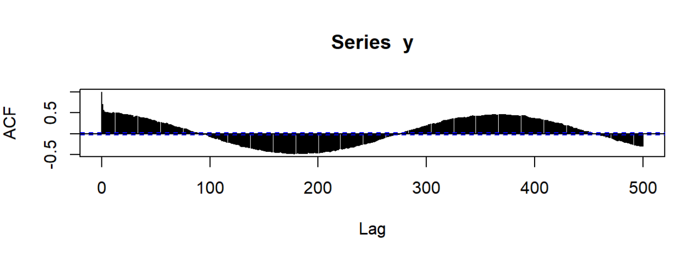


Time Series Model

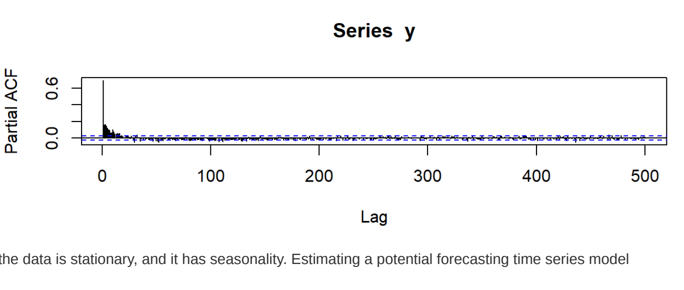
```
y<-df_red[,2][1:5113]
nlags = 500
ts.plot(y)
```




```
par(mfrow=c(2,1))
acf(y, nlags)
pacf(y, nlags)
```

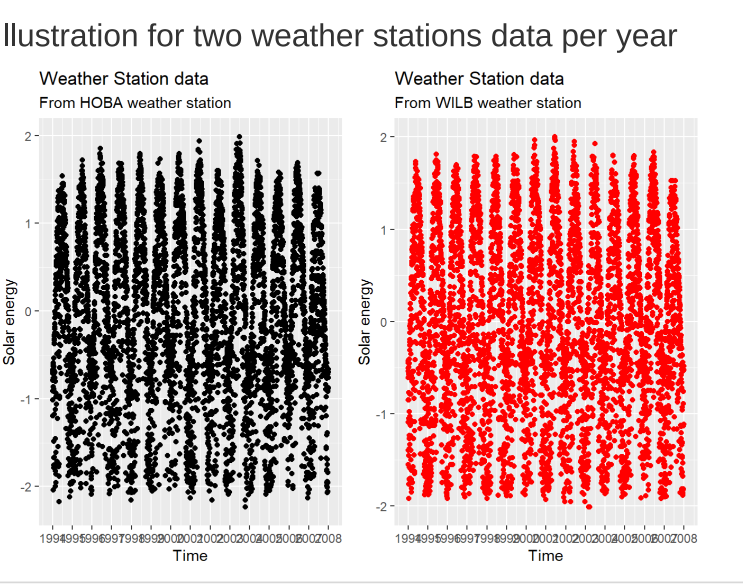


We can clearly notice



the data is stationary, and it has seasonality. Estimating a potential forecasting time series model

Illustration for two weather stations data per year




```
s= 365
fit<-arima(y,order=c(0,1,3),seasonal=list(order=c(0,1,0),period=s))
fit


par(mfrow=c(3,1))
ts.plot(fit$residuals)
acf(fit$residuals)
pacf(fit$residuals)
```

Forecasting

```
y.pred<-predict(fit,n.ahead=1796)
y.pred$pred # point predictions
y.pred$se  # standard error for the point predictions to compute confidence intervals
```

plotting predictions

```
new <- c(y,y.pred$pred) # real data + predicted values

plot.ts(new,main="Predictions",
        ylab="Solar Energy",col=3,lwd=2) # time series plot
lines(y,col=4,lwd=2) # for the second series
legend("topleft",legend=c("Predictions","Historical"),col=c(3,4),
       bty="n",lwd=2)
```

Machine Learning

Regression Model

error analysis

```
##          1          2          3          4          5          6
##   331619.1  1439548.7 -1467137.8 -1916356.1  2421672.1  4428116.6
```

Compute MAE

```
## [1] "MAE = NA"
```

Compute MSE

```
## [1] "MSE = NA"
```

Train & test split

```
## [1] 5113  456
```

```
## [1] 3579  456
```

```
## [1] 1534  456
```

Train model

```
# build linear regression model to predict HOBA using as predictor variable PC1
model_l1_train <- lm(HOBA ~ PC1, data = train);

# Get model predictions for train
predictions_train <- predict(model_l1_train, newdata = train);

# Get model predictions for test
predictions_test <- predict(model_l1_train, newdata = test);

# Get errors
errors_train <- predictions_train - train$HOBA;
errors_test <- predictions_test - test$HOBA;

# Compute Metrics
mse_train <- round(mean(errors_train^2), 2);
mae_train <- round(mean(abs(errors_train)), 2);

mse_test <- round(mean(errors_test^2), 2);
mae_test <- round(mean(abs(errors_test)), 2);

# Build comparison table
comp <- data.table(model = c("lm_1var"),
                   mse_train = mse_train, mae_train = mae_train,
                   mse_test = mse_test, mae_test = mae_test);
library(e1071); # LIBSVM

### train standard SVM model
model <- svm(HOBA ~ PC1, data = train)  # . all columns except the target

# Check model info
print(model);
```

```
##
## Call:
## svm(formula = HOBA ~ PC1, data = train)
##
##
## Parameters:
##    SVM-Type:  eps-regression
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  1
##     epsilon:  0.1
##
##
## Number of Support Vectors:  3050
```

```
# Get model predictions
predictions_train <- predict(model, newdata = train); # There is no formula here! Less straightforward
 to explain
# (Accuracy-'explainability' tradeoff)
predictions_test <- predict(model, newdata = test);

# Get errors
errors_train <- predictions_train - train$HOBA;
errors_test <- predictions_test - test$HOBA;

# Compute Metrics
mse_train <- round(mean(errors_train^2), 2);
mae_train <- round(mean(abs(errors_train)), 2);

mse_test <- round(mean(errors_test^2), 2);
mae_test <- round(mean(abs(errors_test)), 2);

# Build comparison table
comp <- rbind(comp,
              data.table(model = c("standard svm"),
                         mse_train = mse_train, mae_train = mae_train,
                         mse_test = mse_test, mae_test = mae_test));
comp; # Worse than linear regression. What is wrong
```

```
##           model    mse_train mae_train     mse_test mae_test
## 1:      lm_1var 3.479681e+13   4651964 3.362257e+13  4554945
## 2: standard svm 3.423940e+13   4481607 3.354866e+13  4413707
```