Problem Statement

The Energy Information Administration (EIA) has a plethora of hourly electricity demand data for different regions in the lower 48 states. Energy companies rely on accurate forecasts to predict upcoming demand. The goal of this project is to train a machine learning model on weather/electricity data to generate a more nuanced energy-demand forecast. The client for this type of project would be any energy company looking to better understand how the demand for their electricity varies hour to hour. An improved forecast helps energy companies take next steps to fully capitalize on and perhaps better brace for large spikes in electricity demand. Additionally an improved predictive model will shed insight on the features that most heavily affect electricity demand.

Data Cleaning and Wrangling

My project consists of primarily two data sets: (1) electricity data from the EIA, i.e., electricity demand, and (2) hourly, local climatological data (LCD) downloaded from the National Oceanic and Atmospheric Administration (NOAA) website. To limit the scope of the project, I retrieved both data sets from the city of Los Angeles and the city of Seattle (specifically, Los Angeles Department of Water and Power data from EIA and the Los Angeles Downtown LCD data and Seattle City and Light from EIA and Seattle Tacoma International Airport LCD data). Electricity data were retrieved using EIA's API and then unpacked into a pandas `DataFrame`. Since LCD data are not available via NOAA's API, I manually downloaded LCD data from the NOAA website as a CSV file which I imported to a pandas `DataFrame`. Electricity data contain hourly entries from July 2015 to present. To align the data correctly, I downloaded the LCD data from July 2015 to September 2018 and later cut the electricity data to fit the date range.

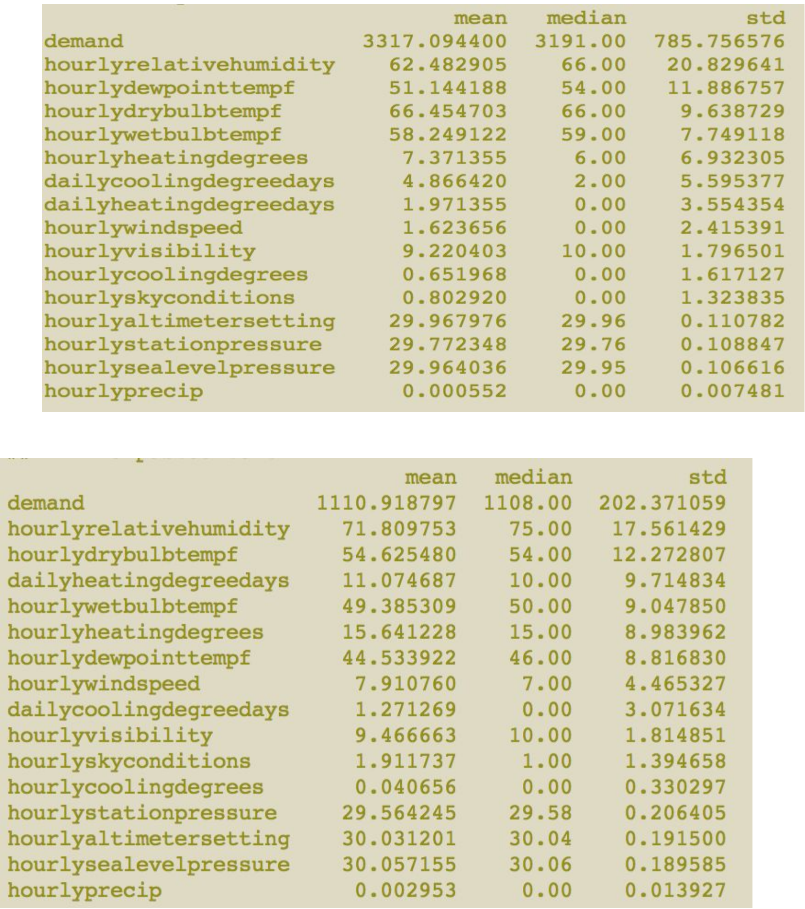
The electricity data required some cleaning--there were very few null values in the set, however there were a very small number of outliers where data was either entered in as zero or a large negative number. Removing outliers cut only ~.01% of the data. LCD data on the other hand required more cleansing. Throughout the cleaning of the LCD data, I frequently referred to the LCD documentation for guidance on how to treat the features. Firstly, all features (except daily cooling/heating degrees) that did not record hourly entries, i.e., monthly and daily records, were dropped. I kept hourly heating and cooling degrees because they are very relevant to energy consumption. I used `bfill` to fill the the hourly entries of a specific day with that day's heating/cooling degrees. Additionally I added hourly heating and cooling degrees to get higher granularity on a metric which is relevant to energy consumption. The hourly heating/cooling degree days are the number of degrees below/above 65 degrees Fahrenheit (wet bulb temperature) (negative values are set to zero) respectively and represents the amounting heating/cooling necessary for that temperature. I dropped a number of hourly columns with excessive null entries compared to the size of the data set and those with low predictive power. I dropped hourly pressure tendency and change because we of excess null values and because measures of pressure were already present, hourly wind direction and gust speed for the same reasons, and hourly present weather type because of excess null values and the dispersity of possible values, indicating low predictive power.

One of the main challenges in cleaning the LCD data was that there were in some cases multiple entries for the same hour. I wanted to have just one entry per hour such that I could eventually align LCD data with the hourly entries in the electricity data. I wrote a function that goes through the weather data and for each instance where there are multiple hourly entries, I keep the hourly entry closest to the hour and drop the remaining rows. For instance, if there are entries at 11:05, 11:36, and 11:56, I kept the entry corresponding to 11:05 and dropped the other rows. I performed the cleaning this way because the data closest to the hour is most representative for that hour. I then renamed the date such that the minute column is set to 00 to once again align with the electricity data. Either way, the values for multiple, per-hour entries are very similar, so the choice of which entry to keep doesn't make a huge difference.

The hourly sky condition feature required additional cleaning. The hourly sky condition feature contains information about each cloud layer (up to three) using six different categories to describe the cloud coverage. As stated in the documentation, the full state of the sky can be best determined the the category of the last (i.e., highest) layer. With this in mind, I replaced the sky condition feature with the condition of the upper most layer to make the feature consistent. This feature also included some information about the cloud layer height which I excluded because these data were only present for a small number of entries. There were also a number of features of mixed float and string types. After looking at their unique values, I saw that some values were string and floats arbitrarily. I simply converted strings to floats to make the features consistent. Finally, I convert our categorical value to numeric using Pandas' `factorize` function to be used for analyzing correlations among other tasks.

Next we deal with null values and outliers. To determine whether to fill null values with median of `ffill`, I plotted histograms, violin plots, and printed stats for each feature. These plots can be found here (for LA) and here (for Seattle). Features where the median was close the mean suggests few outliers and that the median is an appropriate statistic to fill null entries. Conversely, null values in features with outliers present were filled using `ffill`, assuming that previous values have some predictive value for subsequent values since we are dealing with time-stream data. For outliers, because there was not a single feature where the presence of outliers significantly shifted the median from the mean and because the data is such high granularity, outliers were kept without further cleaning. Lastly, we cut both data sets to fit the same date range as mentioned above, and combine the electricity and weather data into a single `DataFrame` on date index. Tables of summary statistics for both Seattle and LA can be seen in **Figure 1**.

To obtain more training data, I performed the same analysis described above with weather and electricity data from Seattle. With both data sets (LA and Seattle), we are left with ~54K data points from July 2015 to September 2018 to perform statistical analysis and predictive modeling. The code for this process can be found here.



```
                                mean   median         std
demand                   3317.094400  3191.00  785.756576
hourlyrelativehumidity     62.482905    66.00   20.829641
hourlydewpointtempf        51.144188    54.00   11.886757
hourlydrybulbtempf         66.454703    66.00    9.638729
hourlywetbulbtempf         58.249122    59.00    7.749118
hourlyheatingdegrees        7.371355     6.00    6.932305
dailycoolingdegreedays      4.866420     2.00    5.595377
dailyheatingdegreedays      1.971355     0.00    3.554354
hourlywindspeed             1.623656     0.00    2.415391
hourlyvisibility            9.220403    10.00    1.796501
hourlycoolingdegrees        0.651968     0.00    1.617127
hourlyskyconditions         0.802920     0.00    1.323835
hourlyaltimetersetting     29.967976    29.96    0.110782
hourlystationpressure      29.772348    29.76    0.108847
hourlysealevelpressure     29.964036    29.95    0.106616
hourlyprecip                0.000552     0.00    0.007481
```

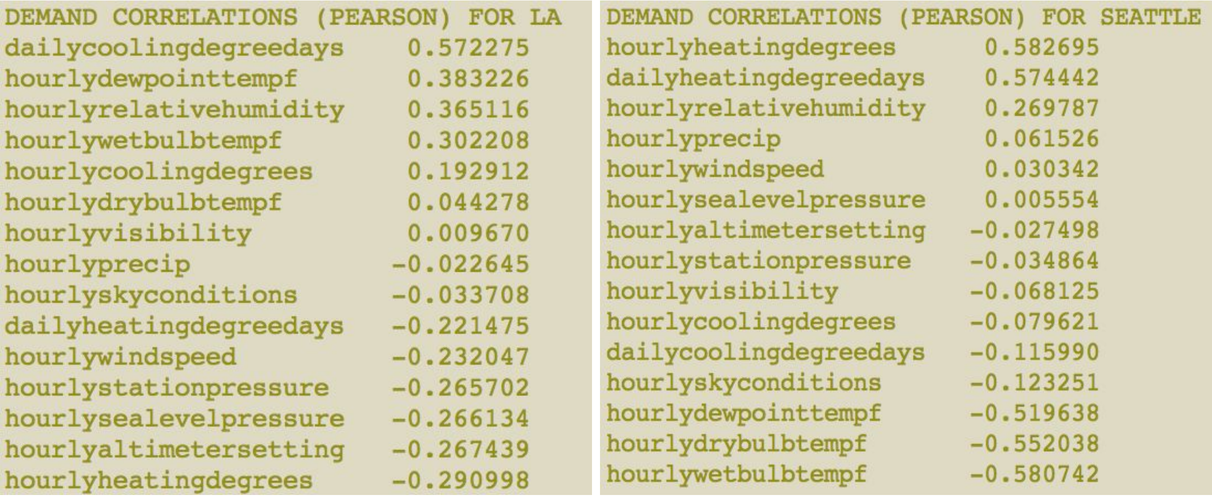
```
                                mean   median         std
demand                   1110.918797  1108.00  202.371059
hourlyrelativehumidity     71.809753    75.00   17.561429
hourlydrybulbtempf         54.625480    54.00   12.272807
dailyheatingdegreedays     11.074687    10.00    9.714834
hourlywetbulbtempf         49.385309    50.00    9.047850
hourlyheatingdegrees       15.641228    15.00    8.983962
hourlydewpointtempf        44.533922    46.00    8.816830
hourlywindspeed             7.910760     7.00    4.465327
dailycoolingdegreedays      1.271269     0.00    3.071634
hourlyvisibility            9.466663    10.00    1.814851
hourlyskyconditions         1.911737     1.00    1.394658
hourlycoolingdegrees        0.040656     0.00    0.330297
hourlystationpressure      29.564245    29.58    0.206405
hourlyaltimetersetting     30.031201    30.04    0.191500
hourlysealevelpressure     30.057155    30.06    0.189585
hourlyprecip                0.002953     0.00    0.013927
```

Figure 1. Summary statistics for all the features in the data set for both LA (top) and Seattle (bottom). As discussed, since the median is not too far away from the mean for almost

all of these features, filling null values with either the median of `ffill` is an appropriate procedure.

Exploratory Data Analysis

Here we perform some initial exploratory data analysis (EDA) for both the Seattle and Los Angeles weather/electricity data sets. One of the first questions we ask is what is the relationship between the independent variables (the weather features) and the dependent variable (electricity demand measured in MWh). Pandas has a nice method (the .corr() method) for finding the Pearson correlation coefficients between all the different variables. **Figure 2** shows the sorted Pearson correlation coefficients for both Seattle and Los Angeles. The results from the Pearson correlation coefficients between Seattle and Los Angeles are interesting. Namely that cooling degree days (CDD) are a significant, positive correlation and heating degree days (HDD) are a significant, negative correlation for LA, whereas the exact opposite is true for Seattle. **Figure 3** shows how the electricity demand varies in time for both LA and Seattle and sheds light on why we observe this behavior. Seattle's electricity demand spikes in the winter months, indicating that the majority of power is used for heating, whereas LA's electricity demand spikes in the summer months, indicating that the majority of power is used for cooling. This is consistent with the computed Pearson correlation coefficients, but a more thorough multivariate regression would be informative.



```
DEMAND CORRELATIONS (PEARSON) FOR LA
dailycoolingdegreedays      0.572275
hourlydewpointtempf         0.383226
hourlyrelativehumidity      0.365116
hourlywetbulbtempf          0.302208
hourlycoolingdegrees        0.192912
hourlydrybulbtempf          0.044278
hourlyvisibility            0.009670
hourlyprecip               -0.022645
hourlyskyconditions        -0.033708
dailyheatingdegreedays     -0.221475
hourlywindspeed            -0.232047
hourlystationpressure      -0.265702
hourlysealevelpressure     -0.266134
hourlyaltimetersetting     -0.267439
hourlyheatingdegrees       -0.290998
```

```
DEMAND CORRELATIONS (PEARSON) FOR SEATTLE
hourlyheatingdegrees        0.582695
dailyheatingdegreedays      0.574442
hourlyrelativehumidity      0.269787
hourlyprecip                0.061526
hourlywindspeed             0.030342
hourlysealevelpressure      0.005554
hourlyaltimetersetting     -0.027498
hourlystationpressure      -0.034864
hourlyvisibility           -0.068125
hourlycoolingdegrees       -0.079621
dailycoolingdegreedays     -0.115990
hourlyskyconditions        -0.123251
hourlydewpointtempf        -0.519638
hourlydrybulbtempf         -0.552038
hourlywetbulbtempf         -0.580742
```

Figure 2. Pearson correlation coefficients for both LA (left) and Seattle (right). As discussed, CDD are positive correlations for LA, while they are negative in Seattle, and vice versa for HDD.

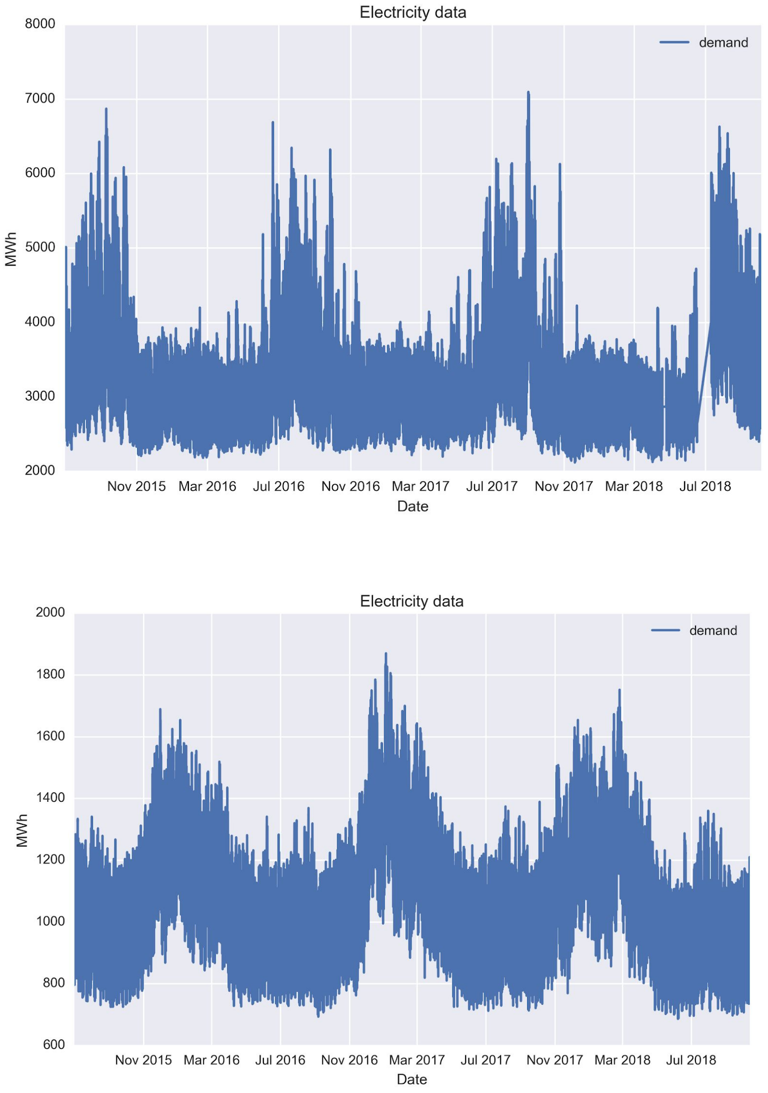
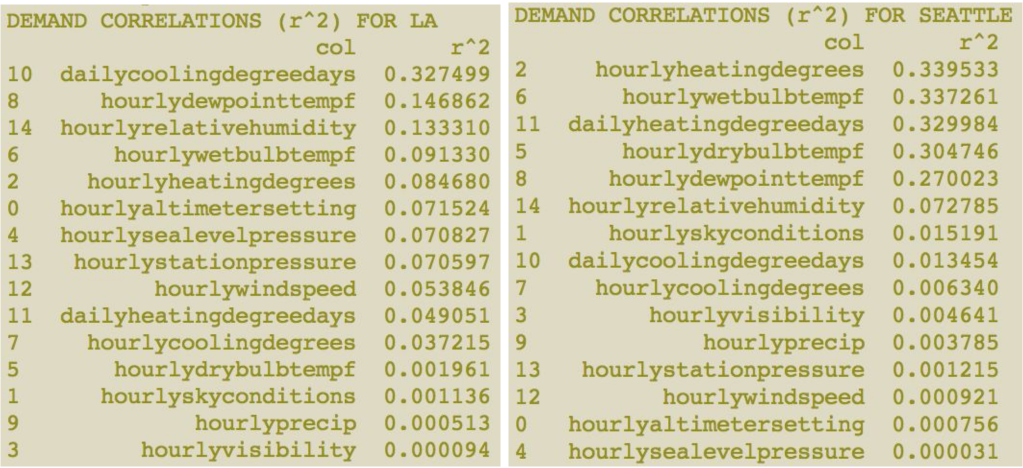


Figure 3. Electricity demand versus time for both LA (top) and Seattle (bottom). Demand peaks in the winter in Seattle, whereas it peaks in the summer in LA.

Additionally, **Figure 4** shows computed r^2 values for both the Seattle and Los Angeles data sets. Here we can see the relative power of each of the features for both data sets. Once again, HDD is a significant feature for Seattle while CDD is a significant feature for LA.



```
DEMAND CORRELATIONS (r^2) FOR LA
                         col       r^2
10  dailycoolingdegreedays  0.327499
8       hourlydewpointtempf  0.146862
14  hourlyrelativehumidity  0.133310
6        hourlywetbulbtempf  0.091330
2      hourlyheatingdegrees  0.084680
0   hourlyaltimetersetting  0.071524
4   hourlysealevelpressure  0.070827
13   hourlystationpressure  0.070597
12         hourlywindspeed  0.053846
11  dailyheatingdegreedays  0.049051
7      hourlycoolingdegrees  0.037215
5        hourlydrybulbtempf  0.001961
1       hourlyskyconditions  0.001136
9             hourlyprecip  0.000513
3          hourlyvisibility  0.000094
```

```
DEMAND CORRELATIONS (r^2) FOR SEATTLE
                         col       r^2
2      hourlyheatingdegrees  0.339533
6        hourlywetbulbtempf  0.337261
11  dailyheatingdegreedays  0.329984
5        hourlydrybulbtempf  0.304746
8       hourlydewpointtempf  0.270023
14  hourlyrelativehumidity  0.072785
1       hourlyskyconditions  0.015191
10  dailycoolingdegreedays  0.013454
7      hourlycoolingdegrees  0.006340
3          hourlyvisibility  0.004641
9             hourlyprecip  0.003785
13   hourlystationpressure  0.001215
12         hourlywindspeed  0.000921
0   hourlyaltimetersetting  0.000756
4   hourlysealevelpressure  0.000031
```

Figure 4. r^2 values for both LA (left) and Seattle (right).

In our data set, we have three features for pressure (hourly sea level, station pressure, and hourly altimeter setting) and three for temperature (hourly dew point, wet, and dry bulb temperature). Having multiple features for essentially the same quantity raises concerns over collinearity. **Figure 5** shows a scatter matrix for the different pressure features for both LA and Seattle. As predicted, these features show high collinearity and as a result we keep the feature with the highest r^2 as referenced in **Figure 4**. For LA it's altimeter setting, whereas for Seattle it's station pressure. Since the all the pressure features have similar values in LA, we choose the highest for Seattle, station pressure, and drop the rest of the pressure columns.

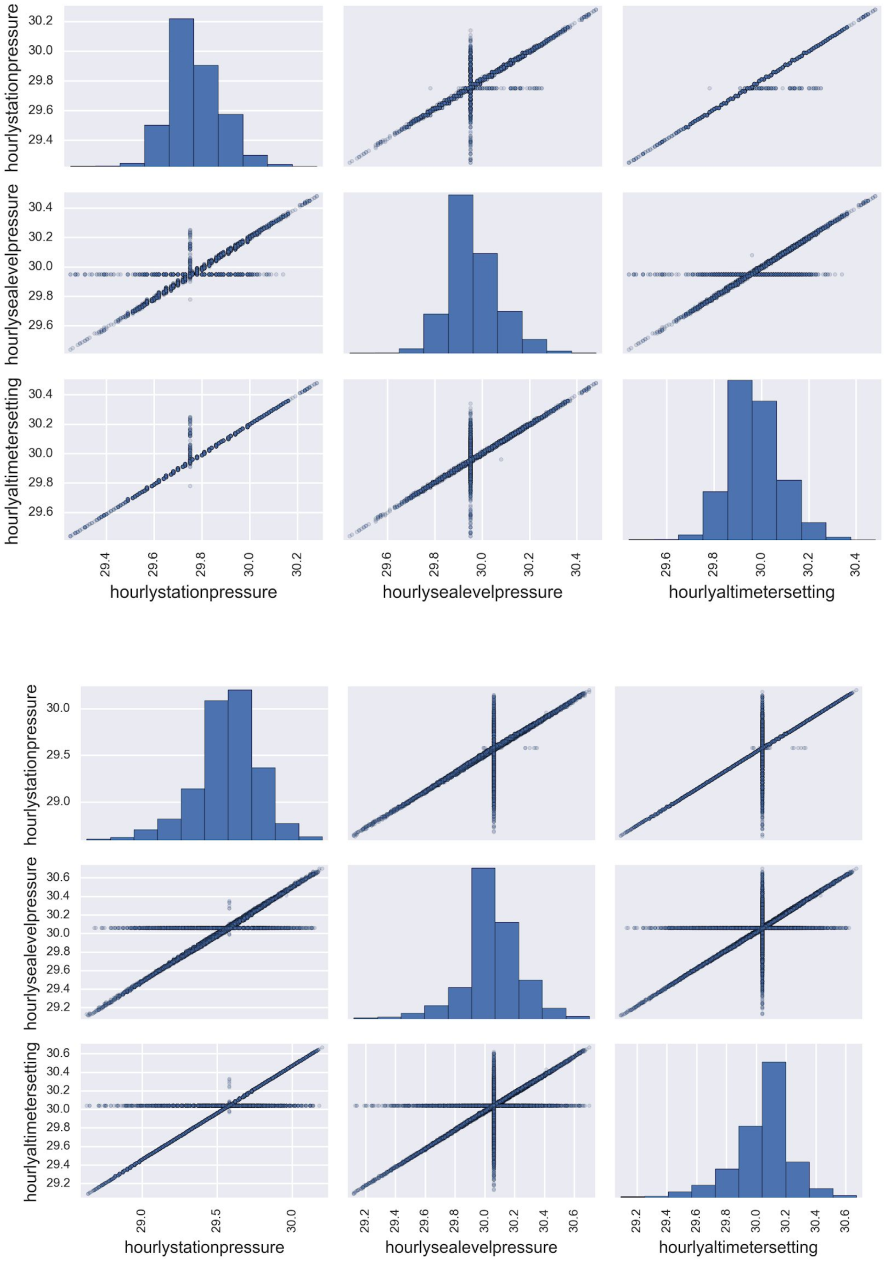


Figure 5. Scatter matrices for pressure features in both LA (top) and Seattle (bottom). Collinearity is obvious in the scatter plots and the histograms also show very similar values.

Figure 6 repeats the above analysis but with the three temperature features. Collinearity is not as obvious but still strong. Since we derive hourly CDD and HDD from wet bulb temperature, we drop that column. Dry bulb temp has a low r^2 value for LA, so we also drop that column, leaving us with dew point temperature as the main temperature feature.

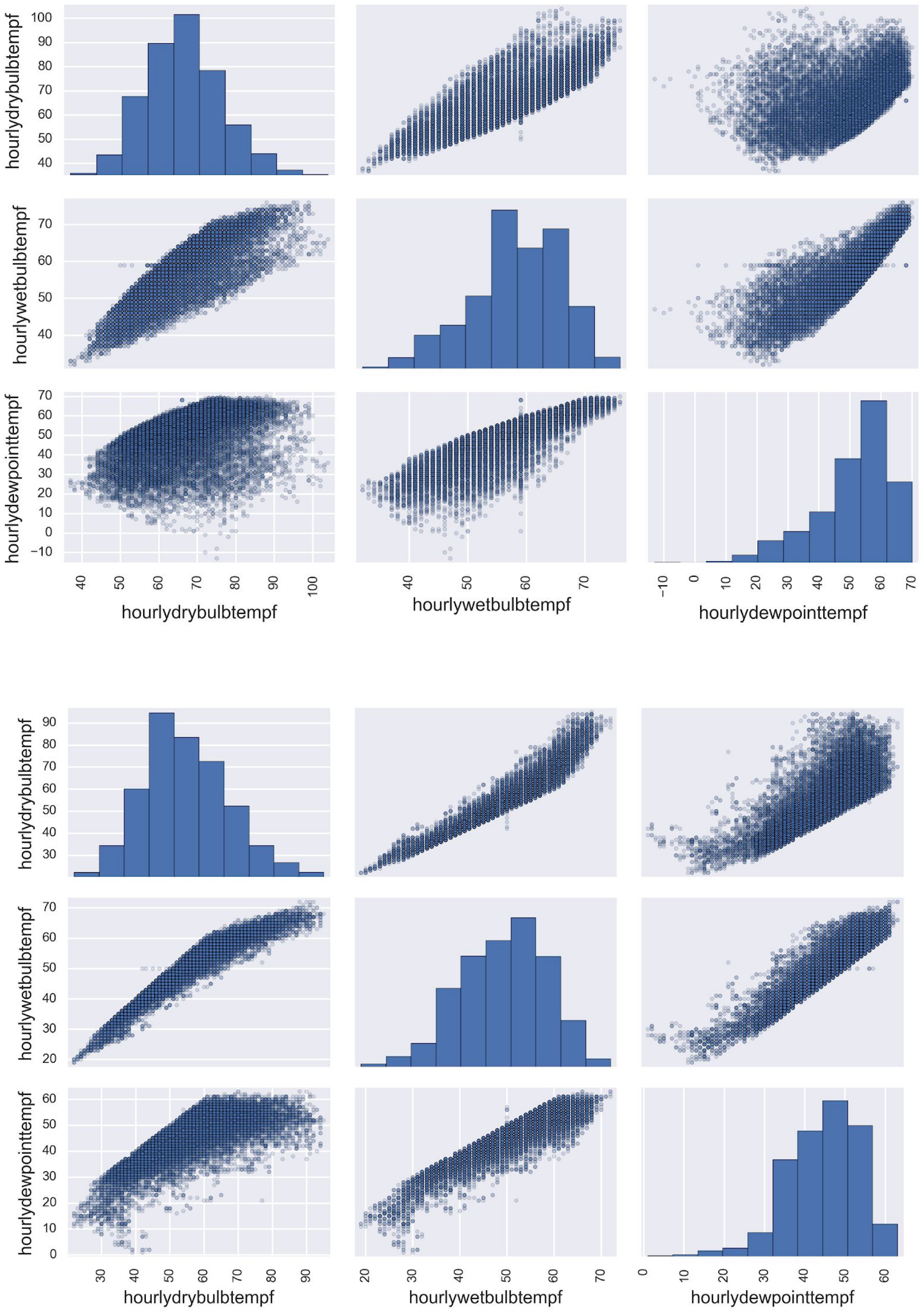
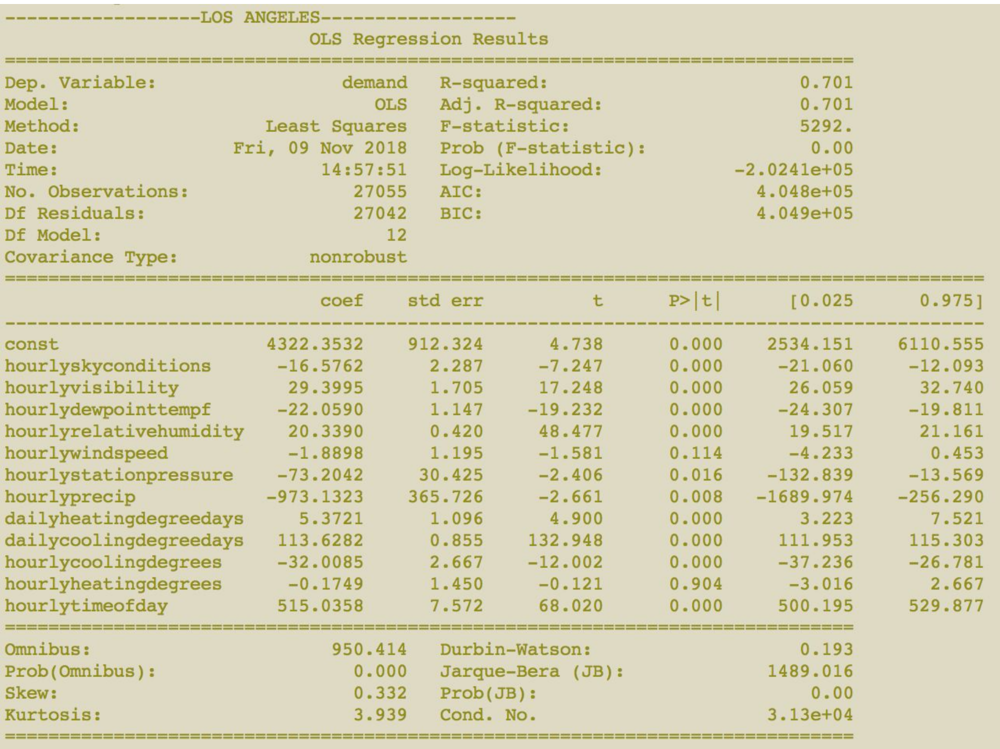


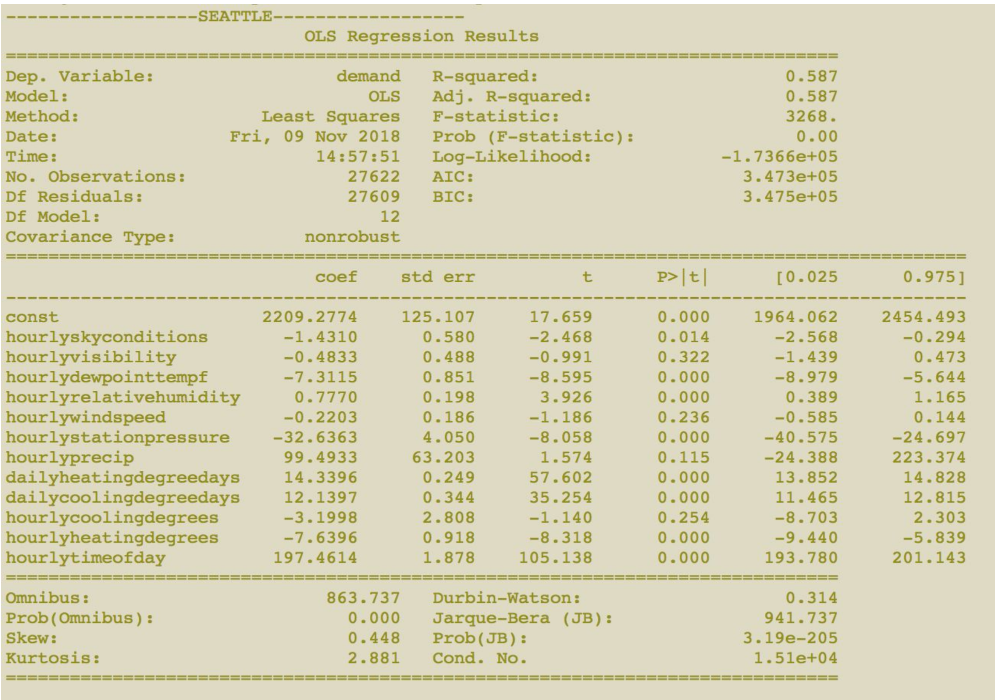
Figure 6. Scatter matrices for temperature features in both LA (top) and Seattle (bottom). Collinearity is obvious in the scatter plots and the histograms also show very similar values.

Since people behave different in the daytime versus the nighttime regardless of how hot or cold it is that day, an ‘hourlytimeofday’ column was added, representing if the current time is day or night (zero between 6AM and 6PM, one otherwise). The temperature data is important for identifying heating and cooling peaks, but the regression can start leaning on the temperature as a proxy for daytime. This feature increased our multivariate regression r^2 considerably (from 0.686 to 0.701 for LA and from 0.419 to 0.587 for Seattle), signifying its importance.

Finally we perform a multiple regression on all the features versus electricity demand. **Figure 6** shows the results for both Seattle and LA.



```
------------------LOS ANGELES------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                 demand   R-squared:                       0.701
Model:                            OLS   Adj. R-squared:                  0.701
Method:                 Least Squares   F-statistic:                     5292.
Date:                Fri, 09 Nov 2018   Prob (F-statistic):               0.00
Time:                        14:57:51   Log-Likelihood:            -2.0241e+05
No. Observations:               27055   AIC:                         4.048e+05
Df Residuals:                   27042   BIC:                         4.049e+05
Df Model:                          12
Covariance Type:            nonrobust
==========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                    4322.3532    912.324      4.738      0.000    2534.151    6110.555
hourlyskyconditions       -16.5762      2.287     -7.247      0.000     -21.060     -12.093
hourlyvisibility           29.3995      1.705     17.248      0.000      26.059      32.740
hourlydewpointtempf       -22.0590      1.147    -19.232      0.000     -24.307     -19.811
hourlyrelativehumidity     20.3390      0.420     48.477      0.000      19.517      21.161
hourlywindspeed            -1.8898      1.195     -1.581      0.114      -4.233       0.453
hourlystationpressure     -73.2042     30.425     -2.406      0.016    -132.839     -13.569
hourlyprecip             -973.1323    365.726     -2.661      0.008   -1689.974    -256.290
dailyheatingdegreedays      5.3721      1.096      4.900      0.000       3.223       7.521
dailycoolingdegreedays    113.6282      0.855    132.948      0.000     111.953     115.303
hourlycoolingdegrees      -32.0085      2.667    -12.002      0.000     -37.236     -26.781
hourlyheatingdegrees       -0.1749      1.450     -0.121      0.904      -3.016       2.667
hourlytimeofday           515.0358      7.572     68.020      0.000     500.195     529.877
==============================================================================
Omnibus:                      950.414   Durbin-Watson:                   0.193
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1489.016
Skew:                           0.332   Prob(JB):                         0.00
Kurtosis:                       3.939   Cond. No.                     3.13e+04
==============================================================================
```



```
------------------SEATTLE------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                 demand   R-squared:                       0.587
Model:                            OLS   Adj. R-squared:                  0.587
Method:                 Least Squares   F-statistic:                     3268.
Date:                Fri, 09 Nov 2018   Prob (F-statistic):               0.00
Time:                        14:57:51   Log-Likelihood:            -1.7366e+05
No. Observations:               27622   AIC:                         3.473e+05
Df Residuals:                   27609   BIC:                         3.475e+05
Df Model:                          12
Covariance Type:            nonrobust
==========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                    2209.2774    125.107     17.659      0.000    1964.062    2454.493
hourlyskyconditions        -1.4310      0.580     -2.468      0.014      -2.568      -0.294
hourlyvisibility           -0.4833      0.488     -0.991      0.322      -1.439       0.473
hourlydewpointtempf        -7.3115      0.851     -8.595      0.000      -8.979      -5.644
hourlyrelativehumidity      0.7770      0.198      3.926      0.000       0.389       1.165
hourlywindspeed            -0.2203      0.186     -1.186      0.236      -0.585       0.144
hourlystationpressure     -32.6363      4.050     -8.058      0.000     -40.575     -24.697
hourlyprecip               99.4933     63.203      1.574      0.115     -24.388     223.374
dailyheatingdegreedays     14.3396      0.249     57.602      0.000      13.852      14.828
dailycoolingdegreedays     12.1397      0.344     35.254      0.000      11.465      12.815
hourlycoolingdegrees       -3.1998      2.808     -1.140      0.254      -8.703       2.303
hourlyheatingdegrees       -7.6396      0.918     -8.318      0.000      -9.440      -5.839
hourlytimeofday           197.4614      1.878    105.138      0.000     193.780     201.143
==============================================================================
Omnibus:                      863.737   Durbin-Watson:                   0.314
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              941.737
Skew:                           0.448   Prob(JB):                    3.19e-205
Kurtosis:                       2.881   Cond. No.                     1.51e+04
==============================================================================
```

Figure 7. Multivariate OLS regression for all the features + constant. Shown are a list of coefficients with errors, t-statistics, confidence intervals.