

Introduction

The Energy Consumption and Renewable Energy project included several files pertaining to both the production of wind and solar energy and the consumption of energy for a fictional town. The project included the following data files:

- solararray_production.csv*
- windfarm_production.csv*
- windfarm_windspeed.csv*
- solararray_solarangle.csv*
- powercity_consumption.csv*
- calendar_days_scenario.csv*
- powercity_population.csv*
- powercity_weather_consumption.csv*
- solararray_weather.csv*
- car_charging.csv*
- powercity_solarangle_consumption.csv*
- Powercity_weather_scenario.csv*

The first step in our preprocessing is to evaluate each of the source files and determine what transformations need to be applied to allow us to evaluate the completeness, accuracy and utility of the data. For the above files, we needed to create a unified datetime column/index for each one in order to join them in the future. *This was a very important step as each file had a different way to represent datetime.* We also removed unneeded columns and normalized column headers. We also had a problem with the *solararray_solarangle.csv* file having duplicated records which needed to be resolved. We then exported everything to the pickle format to be used for further processing.

Joining the Data

The next step was figuring out how all the files fit together. We determined that there are essentially three data groups, production, consumption, and car consumption (scenario year). We landed on the following:

Production_master has 5 files:

- solararray_production.csv*
- windfarm_production.csv*
- windfarm_windspeed.csv*
- solararray_solarangle.csv*
- solararray_weather.csv*

Consumption_master has 6 files:

- calendar_days_consumption.csv*
- powercity_solarangle_consumption.csv*
- powercity_weather_consumption.csv*
- powercity_consumption.csv*
- sector_Use_Matrix.xlsx*
- powercity_population.csv*

Car_Consumption_master has 3 files:

calendar_days_scenario.csv

car_charging.csv

Powercity_weather_scenario.csv

The third step is to figure out how to put all these datasets together. Joining the production_master datasets was straight forward. All we had to do was join on the datetime index. We end up with a single table with 42403 records of the following variables:

- **Time** (Index): Date-time stamp showing hour, day, month and year of the observation
- **Solar_KWH**, measuring solar electricity production in KWH
- **Wind_KWH**, measuring wind electricity production in KWH
- **Wind_Speed_AT_WINDFARM**, measuring the wind speed at the wind farm
- **Solar_Elevation**, measuring the inclination of the solar array
- **Cloud_Cover_Fraction**, measuring the proportion of cloud cover
- **Dew_Point**, measuring the dew point at the time of observation (the atmospheric temperature (varying according to pressure and humidity) below which water droplets begin to condense and dew can form - Wikipedia).
- **Humidity_Fraction**, how humid at the point of observation
- **Precipitation**: Amount of precipitation during the hour in millimeters
- **Pressure**: Pressure reading in millibars
- **Temperature**: Temperature reading in degrees Celsius
- **Visibility**: Visibility in kilometers
- **Wind_Speed_AT_SOLARRAY**, measuring wind speed at the solar array. Note: The solar array is not necessarily co-located with the wind farm, so differences are present. For our analysis, we will use wind speed at the wind farm as a relevant variable for wind power. The only impact wind speed would have on a solar array is if it was sufficiently forceful to actually bend, warp or break the solar array, or otherwise impact its function.

Joining the consumption datasets was a little trickier. The files *calendar_days_consumption.csv*, *powercity_solarangle_consumption.csv*, and *powercity_weather_consumption.csv* could all be joined on the time index. However, *powercity_consumption.csv*, *sector_Use_Matrix.xlsx*, *powercity_population.csv* needed some special processing. These, values needed to be joined on *sector* and *age_group* as well as date. We had to multiply our square_footage per person by our population number for each age_group. In the end we end up with a single dataset covering each hour of a single year for a hypothetical consumption year. The dataset contains most of the same columns as the production dataset except our production values (Solar_KWH, Wind_KWH) are replaced with consumption values. The full set of variables below

- **Time** (date-timestamp) - note, years are missing, so used '1900' as default.
- **FOOD_SERVICE** - Power consumption by this type of activity
- **GROCERY** - Power consumption by this type of activity
- **HEALTH_CARE** - Power consumption by this type of activity
- **K12_SCHOOLS** - Power consumption by this type of activity
- **LODGING** - Power consumption by this type of activity
- **OFFICE** - Power consumption by this type of activity
- **RESIDENTIAL** - Power consumption by this type of activity
- **STAND_ALONE_RETAIL** - Power consumption by this type of activity
- **Weekdays** - Day of the week.

- **HolidayName** - Name of the holiday (if applicable)
- **School_Day** - boolean, 1 if it is a school day, else 0.
- **Workday** - boolean, 1 if not a weekend or holiday, else 0.
- **Solar_Elevation**, measuring the inclination of the solar array
- **Cloud_Cover_Fraction**, measuring the proportion of cloud cover
- **Dew_Point**, measuring the dew point at the time of observation (the atmospheric temperature (varying according to pressure and humidity) below which water droplets begin to condense and dew can form - Wikipedia).
- **Humidity_Fraction**, how humid at the point of observation
- **Precipitable_Water**: The total precipitable water contained in a column of unit cross section extending from the earth's surface to the top of the atmosphere in millimeters
- **Pressure**: Pressure reading in millibars
- **Temperature**: Temperature reading in degrees Celsius
- **Visibility**: Visibility in kilometers

The car_consumption dataset was easily joined on the datetime index. It has the following values:

- **Time** (Index): Date-time stamp showing hour, day, month and year of the observation
- **Electricity_KW_SQFT**, measuring solar electricity production in KWH
- **Cloud_Cover_Fraction**: Amount of cloud cover (decimal from 0 being no clouds to 1 being fully cloudy)
- **Dew_Point**: Temperature of the dew point in degrees Celsius
- **Humidity_Fraction**: Fraction of humidity in the air (0 to 1)
- **Precipitation**: Amount of precipitation during the hour in milimeters
- **Pressure**: Pressure reading in millibars
- **Temperature**: Temperature reading in degrees Celsius
- **Visibility**: Visibility in kilometers
- **Wind_Speed**: Wind speed in meters per second
- **Weekdays** - Day of the week.
- **HolidayName** - Name of the holiday (if applicable)
- **School_Day** - boolean, 1 if it is a school day, else 0.

Notes for Modeling

Through the course of building these datasets a few important things were noticed. Our solar-production and wind-production data do not line up time wise perfectly. We will have to account for this in our models. We can look only at the timeframes for which we have data for both or impute the missing data.

When making predictions using car consumption, we will need to figure out how to handle data which overlaps with the standard consumption dataset.

Wind-Speed is missing from the consumption dataset, but present in the car_consumption dataset. We need to think about how to handle this for modeling. Additionally, solar-angle is present in the consumption but missing from car_consumption.

We filled missing values with zeros. In particular a sensor error was responsible for missing temperature and pressure readings. We will have to determine how we want to handle this, i.e. imputing, ignoring, etc.

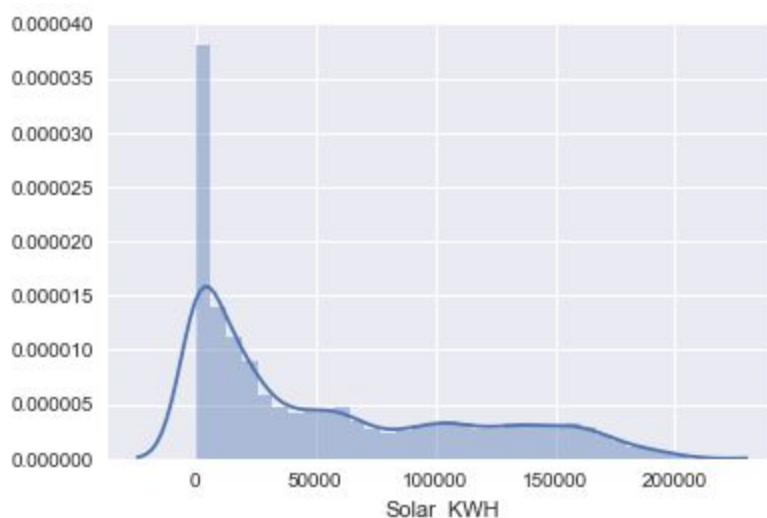
We will have to determine how we want to create a train-test split. Our production data covers multiple years while our consumption data is only a single ambiguous year.

We will be creating ARIMA models, and if time permits XGBoost or RNN models. Based on the description provided for the project task, we will be building models to predict (by hour): (1) solar energy production, (2) wind energy production, and (3) energy consumption within Power City. Additionally, we will need to score and analyze our consumption model when applied to the “Scenario Year”, which includes an additional source of energy consumption due to the adoption of electric vehicles.

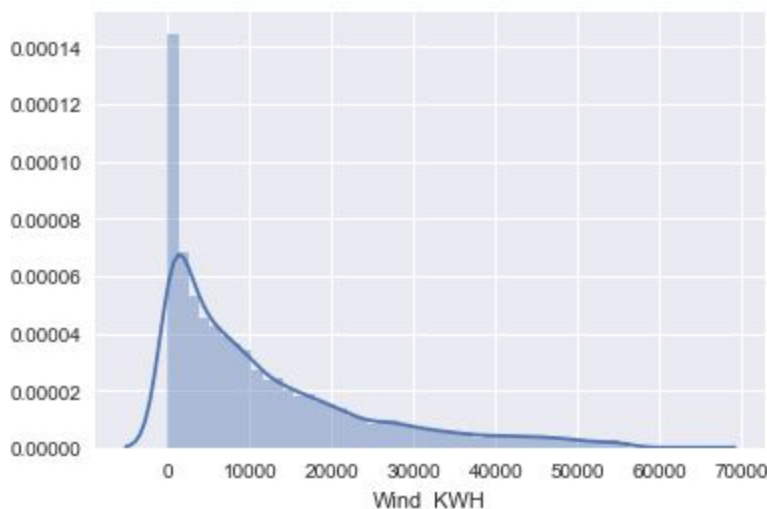
EDA

Production Data

Target Variables - Solar & Wind kWh

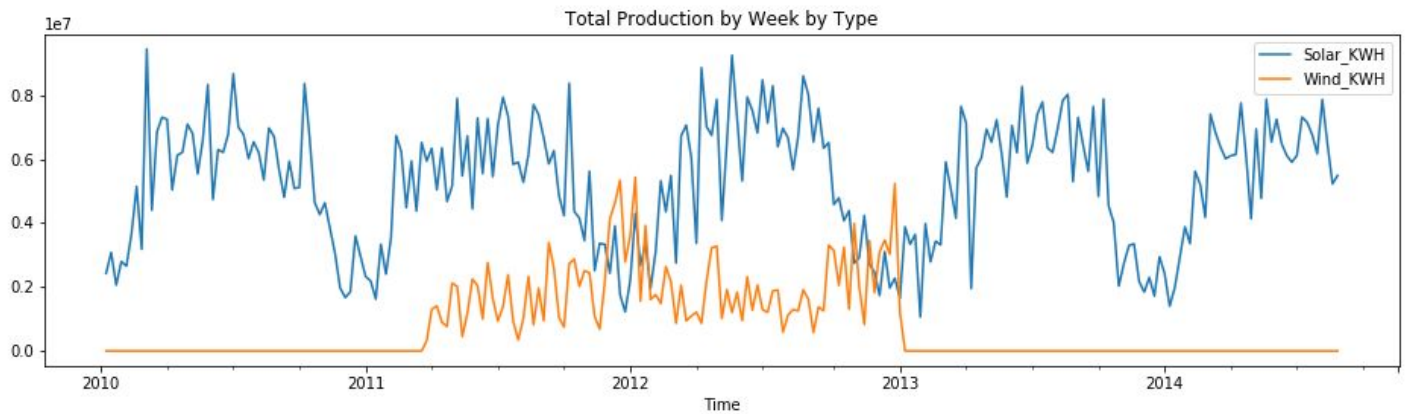


The distribution of solar energy production shows a right skewed, long-tail shape. There are a large frequency of zero output values, reflecting time outside of daylight hours, or during periods of heavy cloud cover.



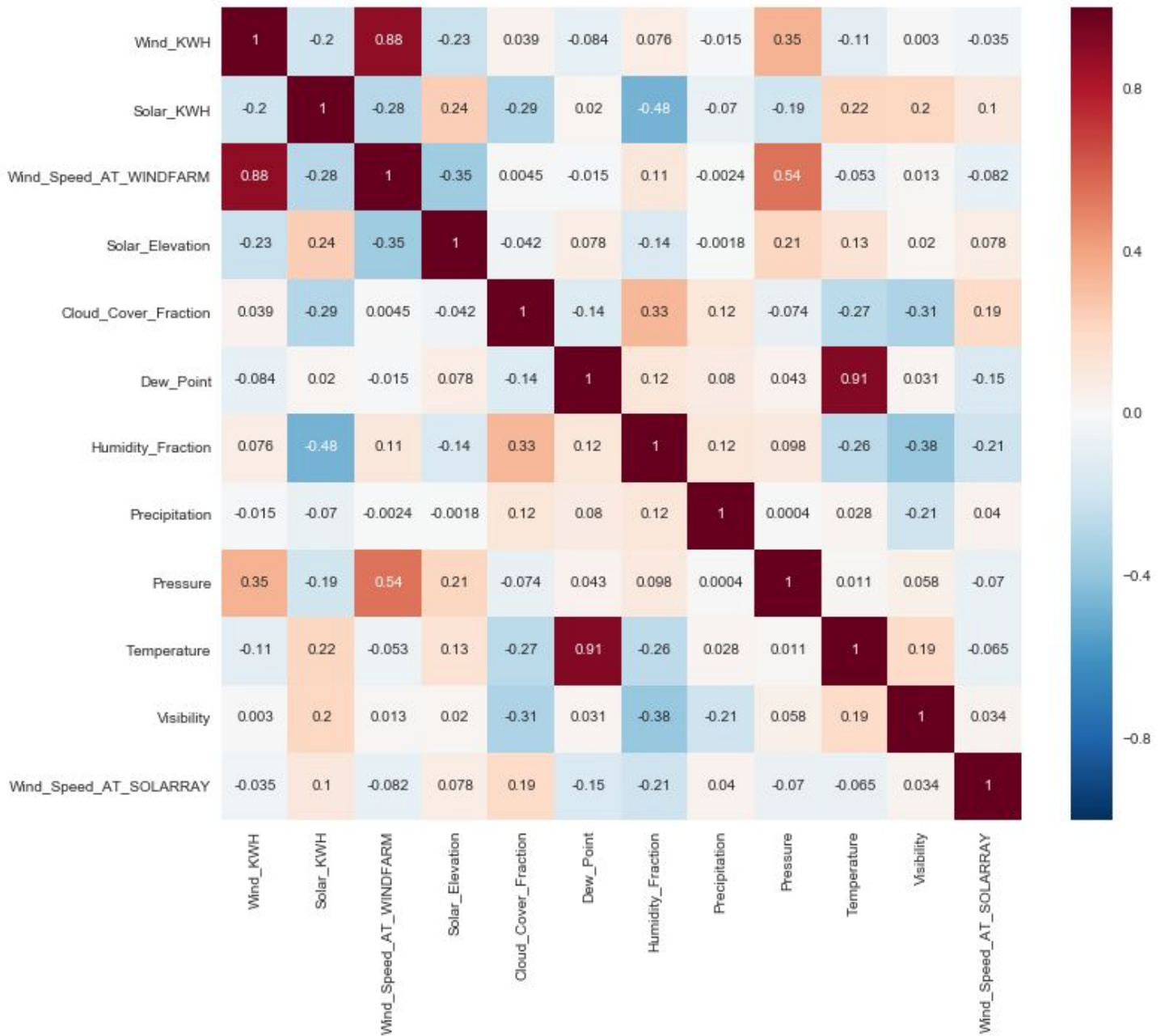
Wind output reflects a similar distribution, but with a more consistent diminishing frequency as energy output increases. This would suggest that wind is a less volatile aspect of weather than light

exposure. Similar to solar output, wind reflects a high frequency showing no hourly energy produced, which would represent times where wind speeds did not meet a minimum level to operate the turbines. Finally, one can see that the absolute level of output for solar is much higher relative to wind on average.



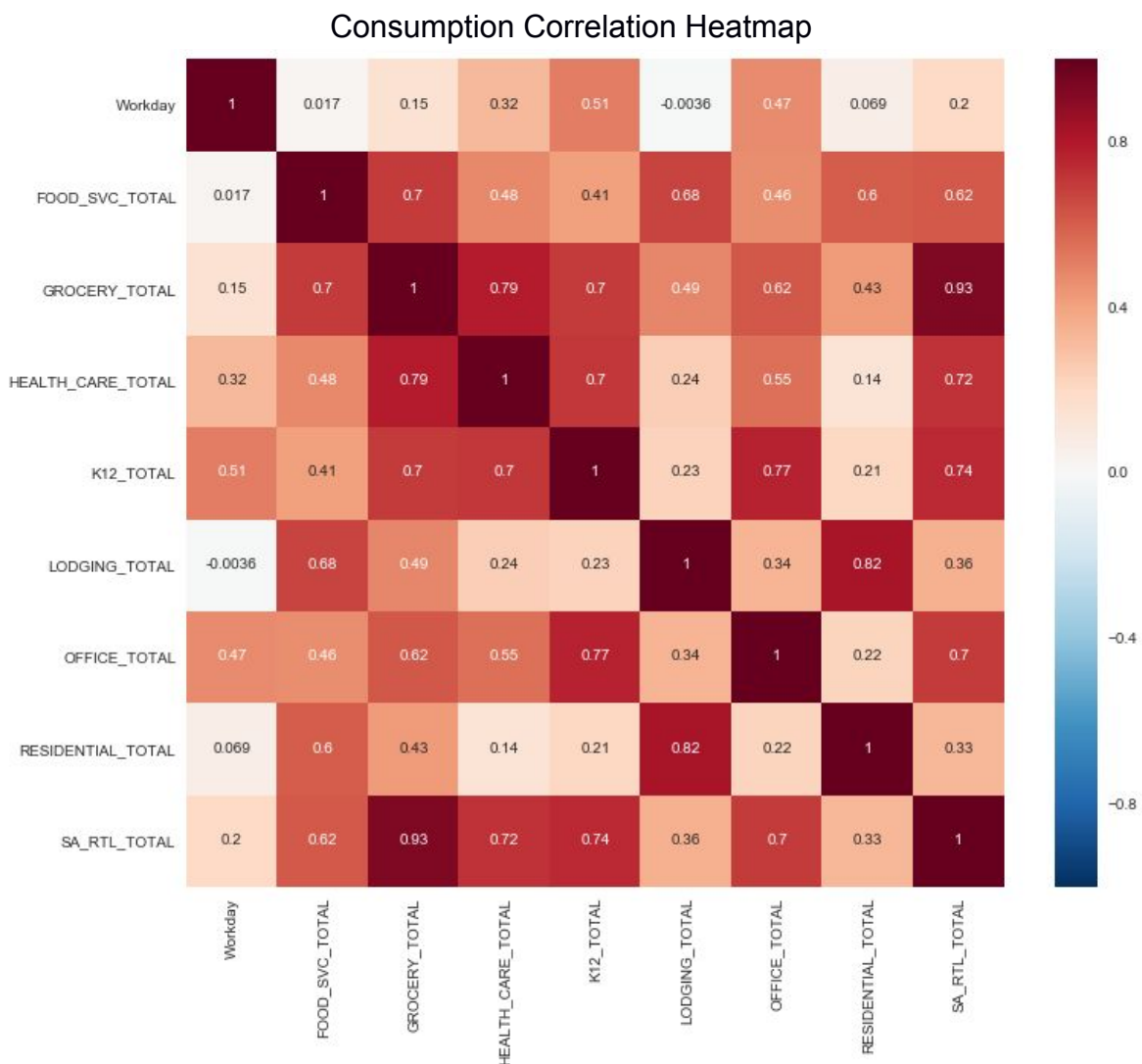
The time series above gives a clear overview of each energy output, as well as how they compare relative to one another. Solar generation shows an obvious seasonal trend on what appears to be an annual basis, where output peaks during the summer months. Wind production appears to potentially follow an inverse trend, but due to a lack of data it is more difficult to show compelling evidence. Both sources, however, do reflect a shorter term volatility, likely related to time of day as well as changing weather conditions. Once again, the larger overall energy output of solar relative to wind is made more clear by this visual.

Production Correlation Heatmap



The correlation heatmap above shows how individual predictors in the dataset are linearly related to each other and the target production variables. The two strongest relationships appear to be between wind production and wind speed, as well as temperature and dew point, both of which come at no surprise. Solar_kWh is weakly correlated with several features, namely Humidity, Cloud Cover, and Solar Angle.

Consumption Data



Similar to the figure for Production, the heatmap above describes how features in our Consumption data table linearly correlate with each other. The top relationships are fairly intuitive, as they include:

1. 0.93 - Stand Alone Retail / Grocery
2. 0.82 - Lodging / Residential
3. 0.79 - Healthcare / Grocery
4. 0.77 - K12 / Office

All of these relationships seem straightforward in that they are businesses or locations where people would be active during the same periods of time.

A summary of energy consumption by sector in a tabular format is shown below.

	mean	std	min	25%	50%	75%	max
FOOD_SVC_TOTAL	8389.24	2635.54	3653.72	7675.90	9000.07	10277.44	15115.61
GROCERY_TOTAL	2250.25	878.9	903.9	1436.71	2305.57	2911.75	4460.40
HEALTH_CARE_TOTAL	4139.1	1015.12	1804.62	3388.2	4016.86	4921.20	6182.32
K12_TOTAL	3179.47	1909.45	1311.44	1817.51	1868.66	4993.36	11272.87
LODGING_TOTAL	1432.78	455.73	688.03	1098.81	1393.73	1784.59	2804.22
OFFICE_TOTAL	16234.55	10146.95	3500.19	7260.22	13785.7	25107.90	60621.09
RESIDENTIAL_TOTAL	65907.64	24080.02	37010.14	46415.49	59572.37	81133.29	171016.6
SA_RTL_TOTAL	1410.24	882.19	144.2	489.43	1357.39	2251.02	3997.25

On average, the residential sector dominates in terms of consumption. While other sectors such as grocery and retail have very low minimums (likely outside business hours), residential consumption is consistently high even at its lowest levels. Building models that are able to represent these swings in use based on time of day and the type of day will be critical to making consistently accurate predictions.