Exploratory Data Analysis:

Using Data Analysis to Examine the Effectiveness of Seattle's Solar Station

Alex Dantinne, Yabo Gao

Northwestern University

Introduction

The project aims to look at how well the four solar panels at Seattle's North Transfer Station works to generate solar data. The dataset is hosted by the City of Seattle. This project is very important to us because we believe that global warming is an immediate crisis and want to help alleviate it by switching to renewable energy sources. We picked this data because we want to learn more about the solar industry. We think this data will help us learn more about how well the City of Seattle generates solar energy. There are 871 attempted collections from each inverter and no sampling was used. However, not all attempts were successful. An inverter takes variable DC, direct current, output from solar panels and converts it to AC, alternating current, which homes and businesses are wired to use. We will explain the data prepping and cleaning in the next section but for now, we will list a few hypotheses:

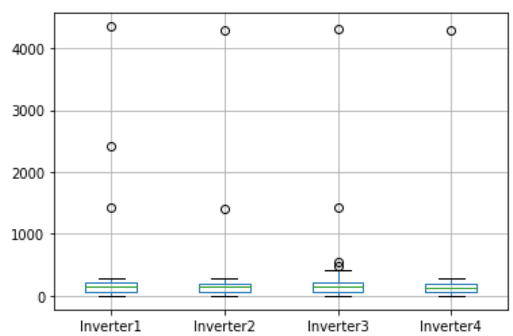
- The solar panels are not that different from each other in terms of overall and seasonal effectiveness. The seasons are categorized as follows: Fall: September-November, Winter: December-February, Spring: March-May, and Summer: June-August
- The average amount of energy generated by each panel during different seasons will be ranked as follows: Summer > Spring >= Fall > Winter. This will also be true for all four inverters combined.

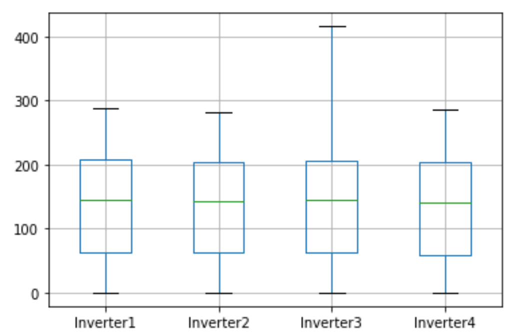
The results of our hypothesis will help us answer the following questions:

- Can the City of Seattle rely on solar energy sources all year round or should they seek alternative sources of energy to meet their needs?
- Should all solar panels continue to function as is or should the City consider funding for a replacement/some replacements?

Data Preparation and Analysis

Before we proceed, we'd like to first provide an overview of the data and make it more accessible to you, the reader. We read the solar CSV file in our Jupyter notebook to examine the data. Once examined, we wrote code to remove the missing values, all entries that do not have the values of all four inverters recorded, and both mild and extreme outliers from the data. We started off with 871 rows. 154 rows were being removed for having missing or inconsistent values, bringing the total to 717. Five rows were subsequently removed for being outliers, bringing the total to 712. The side by side boxplots show the solar data as-is from the CSV file and after preparing the data for analysis.





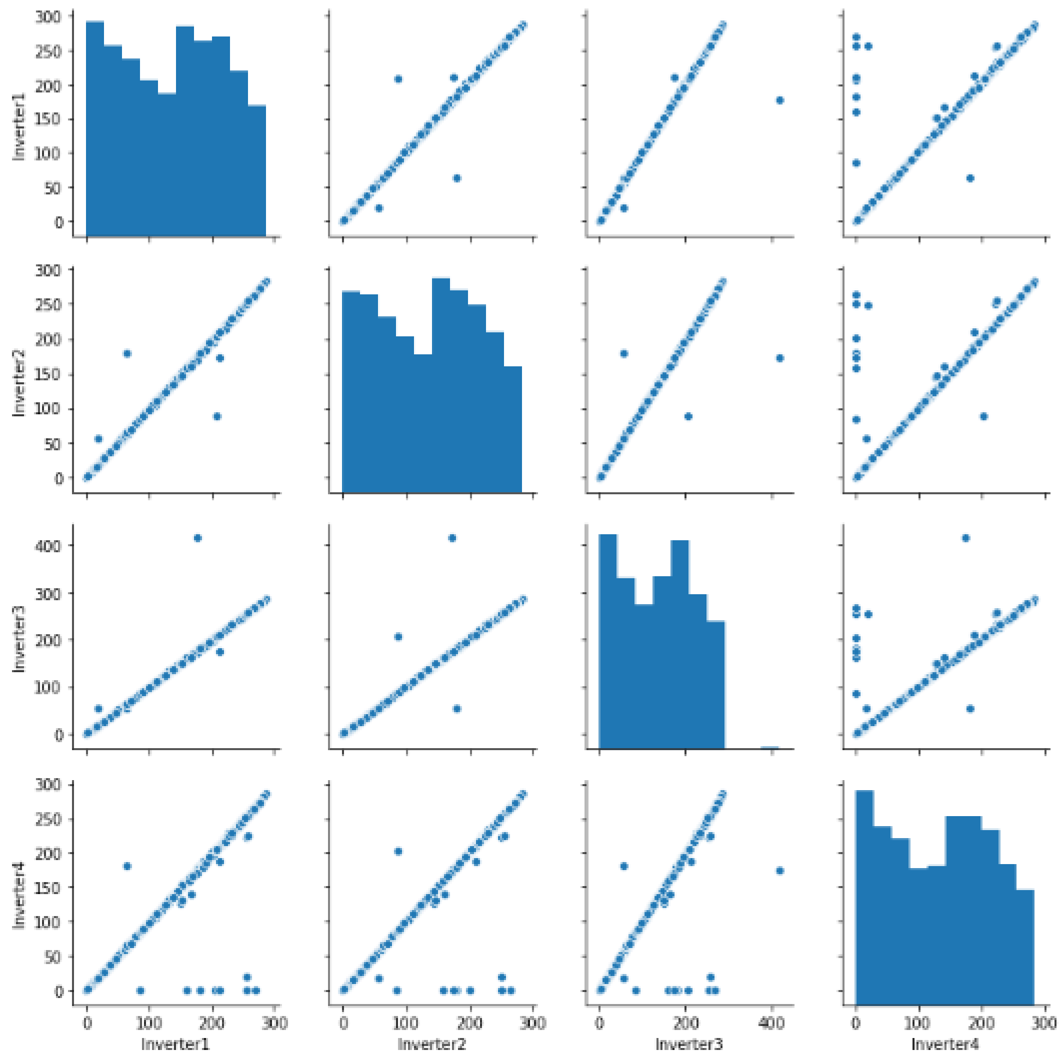
Now that we have a more reasonable dataset to analyze, we will now generate a dictionary from our data. Since Date is associated with all the inverters, we will orient the dictionary by index to minimize repetition. Below is a screenshot of our dictionary:

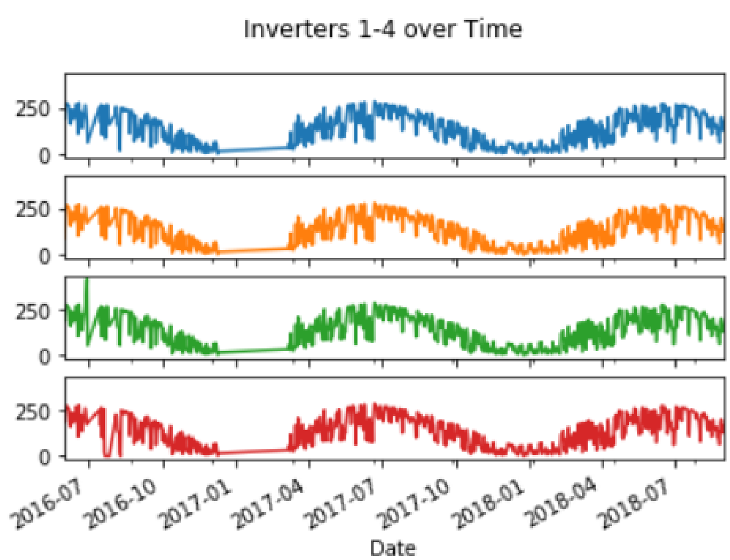
```
847: {'Date': '2016-Aug-06',
 'Inverter1': 166.6,
 'Inverter2': 159.9,
 'Inverter3': 163.8,
 'Inverter4': 140.3},
848: {'Date': '2016-Aug-05',
 'Inverter1': 255.4,
 'Inverter2': 250.7,
 'Inverter3': 255.1,
 'Inverter4': 221.8},
849: {'Date': '2016-Aug-04',
 'Inverter1': 257.8,
 'Inverter2': 254.3,
 'Inverter3': 257.3,
 'Inverter4': 225.4},
```


For the remainder of this section, we will be highlighting each of our four hypotheses in bullet form and testing out these hypotheses immediately after the bullet points. We will also answer questions that corresponds to our hypothesis along in this area.

- The solar panels are not that different from each other in terms of overall effectiveness
  - The correlation matrix below indicates that there's a very high correlation or similarity between each and every one of the solar panel inverters. The formula for each individual entry can be found in Saha's article. The correlation value for the diagonals is 1 because every column is identical to itself. The pair plot below is a visual representation of the correlation matrix. The line plots below are designed to provide a graphical presentation of the similarities among the four inverters and seasonal variability.

	Inverter1	Inverter2	Inverter3	Inverter4
Inverter1	1.000000	0.996608	0.993630	0.961893
Inverter2	0.996608	1.000000	0.990665	0.963460
Inverter3	0.993630	0.990665	1.000000	0.957213
Inverter4	0.961893	0.963460	0.957213	1.000000


Inverter1
Inverter2
Inverter3
Inverter4
Inverter1
Inverter2
Inverter3
Inverter4


Inverters 1-4 over Time
Date

The value of this correlation matrix and the similarity of line plots supports our hypothesis and therefore we can conclude that all solar panels continue to function as is.

- The solar panels are not that different from each other in terms of seasonal effectiveness
    - The data frame used to generate the previous correlation matrix has been broken down into 16 individual arrays, with each inverter of the four inverters being broken down by season as follows: Fall: September-November, Winter: December-February, Spring: March-May, and Summer: June-August
    - Four one-way anova tests has been performed on each season for all four inverters. How this test is performed can be found in Black's book chapter 11. The results and reference F value for alpha = 0.95 are shown below:

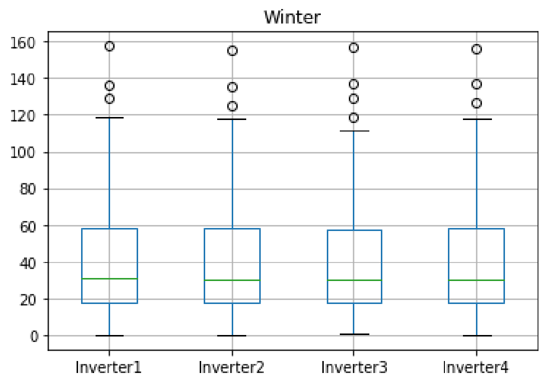
Season	Computed F Value	Reference F value[1]
Spring	0.04639969626702745	2.656532
Summer	1.474615087786745	2.640422
Fall	0.02562335891891728	2.654237
Winter	0.005055204124114	2.695534

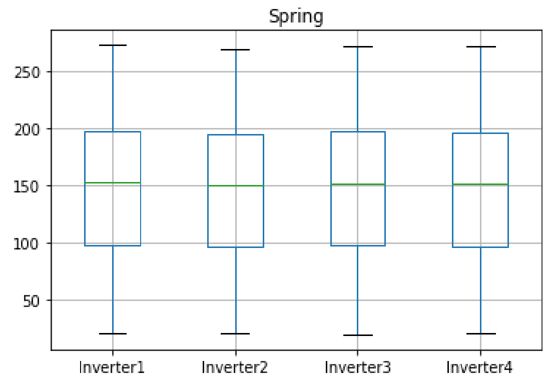
Because all of the Computed F Values are less than the Reference F Values, we can say that the solar panels are not different from each other in terms of seasonal effectiveness. Below are boxplots from the seasonal data, in yearly order from winter to fall.

[1] These values have been computed using R Console

	Inverter1	Inverter2	Inverter3	Inverter4
count	100.000000	100.00000	100.000000	100.000000
mean	41.064000	40.60300	40.596000	40.690000
std	31.279911	30.88821	31.153085	31.198386
min	0.300000	0.00000	0.600000	0.000000
25%	17.750000	17.42500	17.525000	17.275000
50%	30.850000	30.30000	30.400000	30.150000
75%	58.350000	57.87500	57.700000	58.250000
max	157.700000	155.40000	157.000000	156.300000

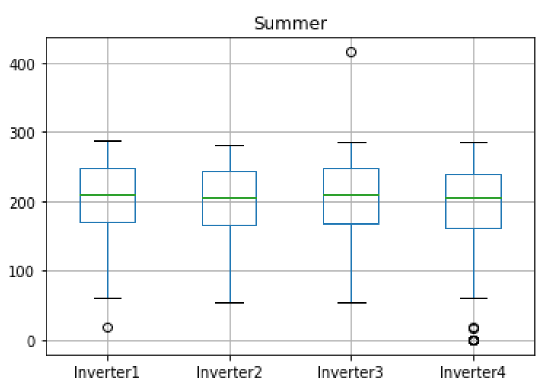
	Inverter1	Inverter2	Inverter3	Inverter4
count	178.000000	178.000000	178.000000	178.000000
mean	150.251685	147.732584	149.409551	148.691573
std	66.590281	65.635348	66.482823	66.355726
min	20.200000	19.900000	19.300000	20.100000
25%	97.900000	96.175000	97.150000	96.200000
50%	152.400000	149.950000	151.900000	151.450000
75%	197.850000	194.325000	197.100000	195.775000
max	272.800000	269.700000	272.000000	271.700000

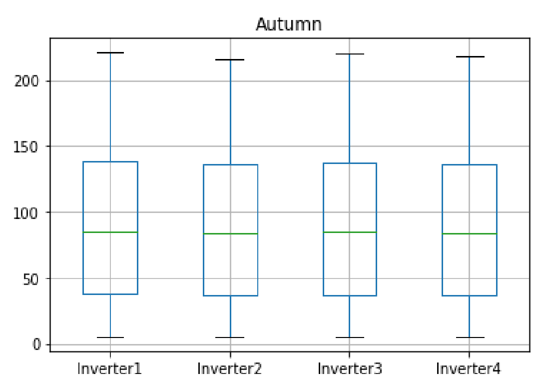




	Inverter1	Inverter2	Inverter3	Inverter4
count	252.000000	252.000000	252.000000	252.000000
mean	202.861111	198.643651	202.660317	193.596825
std	54.720009	53.170085	55.969230	63.245736
min	19.100000	55.400000	55.500000	0.000000
25%	169.800000	165.200000	168.475000	162.150000
50%	210.650000	207.150000	209.750000	206.000000
75%	248.725000	243.825000	247.375000	240.500000
max	287.200000	282.800000	416.100000	285.400000

	Inverter1	Inverter2	Inverter3	Inverter4
count	182.000000	182.000000	182.000000	182.000000
mean	90.959890	89.534066	90.350549	89.873077
std	60.318105	59.373238	60.143023	59.889218
min	5.100000	5.400000	5.300000	5.100000
25%	37.525000	36.550000	36.900000	36.600000
50%	85.300000	83.550000	84.600000	84.000000
75%	138.050000	135.650000	137.250000	136.550000
max	221.200000	216.400000	219.900000	218.400000





- The average amount of energy generated by each panel during different seasons will be ranked as follows: Summer > Spring >= Fall > Winter
    - To test this, we have computed the mean of all 16 arrays and grouped them by inverters as follows. Mean is defined as the sum of all items in the array divided by the total number of items in the array

	Inverter1	Inverter2	Inverter3	Inverter4
Spring Summer Fall Winter	150.25168539325838 202.86111111111114 90.95989010989005 41.06399999999999	147.73258426966294 198.6436507936506 89.53406593406596 40.602999999999994	149.40955056179777 202.66031746031754 90.95989010989005 40.596	148.69157303370787 193.5968253968253 90.95989010989005 40.69

Examining the means above, the order is actually Summer>Spring>Fall>Winter for all the inverters. We find this to be pretty consistent with what we suspected since we thought there would be no sunlight during the winter season for the inverter to absorb power from. While we can dive more into this during further research, we have sufficient evidence to say that the City of Seattle can rely on this energy source all year round. All they need to do is store half of the energy generated during summer and load it back during winter.

- The average amount of energy generated by all panels combined during different seasons will be ranked as follows: Summer > Spring >= Fall > Winter
    - From the clear results during our last finding, we can safely say that the average amount of energy generated by all panels combined during different seasons will be actually ranked as follows: Summer > Spring >Fall > Winter

Conclusion

In this study we are able to study the four solar panels at Seattle's North Transfer Station. Due to the way this dataset is organized, we had some issues coming up with a dictionary for this dataset, but everything else was able to run rather smoothly. We were right in that all four inverters have similar overall functionalities and in our hypothesis of the order of mean energy production. We are able to determine that all solar panels can continue to function as is and the City can rely on this all year round if they are able to store half of the energy generated during Summer and transfer it into Winter. A research topic that can spring from this study is: “How can the extra energy converted by the inverters in Summer be stored until the Winter months ?” This result is true to our original hypothesis but because of our limited knowledge in the science behind solar panels, we are unable to provide further explanation.

References

Black, K. (2016, September 23). Business Statistics: For Contemporary Decision Making, 9th Edition - Wiley. Retrieved February 29, 2020, from https://www.wiley.com/en-us/Business Statistics: For Contemporary Decision Making, 9th Edition-p-9781119320890

Bock, T. (2018, August 16). What is a Correlation Matrix? Retrieved February 29, 2020, from https://www.displayr.com/what-is-a-correlation-matrix/

Saha, S. (2018, October 5). Let us understand the correlation matrix and covariance matrix. Retrieved February 29, 2020, from https://towardsdatascience.com/let-us-understand-the-correlation-matrix-and-covariance-matrix-d42e6b643c22

scipy.stats.f_oneway¶. (2019, December 19). Retrieved February 29, 2020, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html

Seltman, H. J., & Seltman, H. J. (2018). Exploratory Data Analysis. In *Experimental Design and Analysis* (pp. 61–98).

What are outliers in the data? (n.d.). Retrieved February 29, 2020, from https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm