

FIT5147 Data exploration and visualisation – S2 2021

Programming Exercise 1: Tableau Public

Jiaming Ren, ID:217218863

Table of Contents

1. Introduction.....	2
2. Data checking and cleaning.....	2
2.1 load and check Data	2
2.2 Visualize Errors	2
3. Data exploration.....	4
3.1 Compare and contrast the average hourly (over 24 hours) solar energy generation trends of each location.	4
3.2.1 How does the amount of energy generated vary across the entirety of 2020 for each individual location.....	5
3.2.2 How does the amount of energy generated vary across the entirety of 2020 for all locations aggregated together.....	6
4. Summarize	6

1. Introduction

This data analysis report is for Programming Exercise 1: Tableau Public of the unit FIT5147 in Monash University. The dataset I am using is the records of the amount of energy generated by solar panels on top of the buildings in Monash University's Clayton Campus of the year 2020. I will use Tableau Public to read, explore and visualize the dataset. There are five problems in the dataset, and two issues are known, but it is irrelevant to this assignment.

Two Known issues:

- there are multiple time periods with missing (NULL) values. missing values are inherently different from zero (0), as its value is simply unknown rather than specifically zero.
- a solar meter had reported small amounts of energy being generated at night.

The other three unknown issues are:

- negative value
- naming error
- duplicate rows

I will explain how to use tools (python and Tableau Public) to explore and visualize and correct these issues in my data analysis report.

2. Data checking and cleaning

2.1 load and check Data

PE1_Solar_Data_Generation_2... Timestamp	Abc Mon Solar Meter	# Real Energy Into t...
12/7/2020 7:30:00 AM	LV_N1_Building_Carp...	132.909
12/7/2020 7:30:00 AM	LV_Engineering_72.S...	15.850
12/7/2020 7:30:00 AM	LV_Biomedical_Scien...	4.040
12/7/2020 7:30:00 AM	LV_Sport_01.PVDB-D...	22.217
12/7/2020 8:00:00 AM	LV_Sport_01.PVDB-D...	30.531
12/7/2020 8:00:00 AM	LV_N1_Building_Carp...	181.903

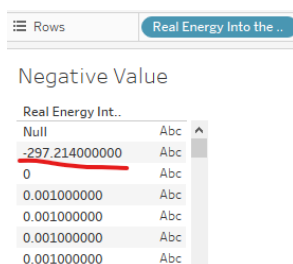
Figure 1: PE1_Solar_Data_Generation_2020.csv dataset

The Figure 1 is what the dataset looks like after loading the PE1_Solar_Data_Generation_2020.csv in Tableau. The dataset has three columns, and the data type for each column are Date & Time (qualitative), Text (qualitative) and Numerical (quantitative).

2.2 Visualize Errors

I will use Tableau to explore and visualize all three issues that has not been found.

Negative Value:



Real Energy Int..	Abc
Null	Abc
-297.214000000	Abc
0	Abc
0.001000000	Abc
0.001000000	Abc
0.001000000	Abc
0.001000000	Abc

Figure 2: Explore Negative value

From Figure 2, I explored the range of all energy value and found a negative value. There should be no negative value in the dataset because solar panels generate electricity, not consume it. Before I decide whether remove it or keep it, I will check whether the value is an outlier.

Solution:

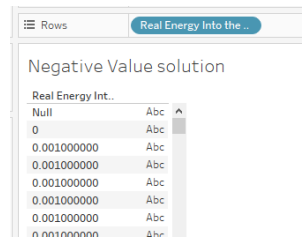


Figure 3: Corrected Negative Value

From Figure 3, the negative value has been kept and change to positive. I make the value positive and explore and compare it with the adjacent records or recent days at the same time. I find there is not too much difference. The records of value before and after the error are 285.6 and 275.0. Also, on February 17, 12 pm, the value is 271.1. Therefore, there is not much difference between adjacent data. I decide this data should keep and change to positive data.

Naming Error:

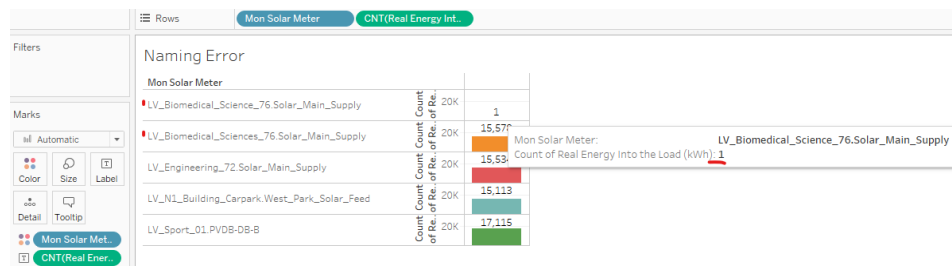


Figure 4: Explore Naming Error

Figure 4 shows there are five unique building names. However, two-building names science(LV_Biomedical_Science_76.Solar_Main_Supply) and sciences(LV_Biomedical_Sciences_76.Solar_Main_Supply) are similar. Also, we can see that there is only one data counted in the science building, whereas other buildings have much more records. Therefore, we can confirm that the science building has the wrong name.

Solution:

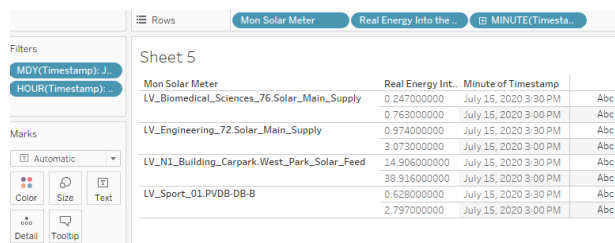


Figure 5: Corrected Naming Error

From Figure 5, there is only four buildings and the row from the science building is filled into the sciences building. I explored whether the sciences building has a record in that time that the single record that the science building has, and I find the sciences building missed the record on July 15, 3 pm. Furthermore, I compare it with the adjacent records or recent days at the same time, and I find the energy values are similar. Therefore, I correct the name of the science building and keep the value.

Duplicate rows:

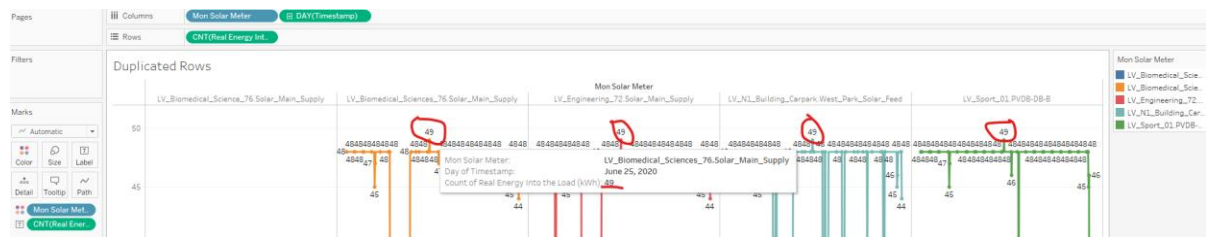


Figure 6: Explore Duplicate Rows

From Figure 6 we can learn that each building records real energy into the load every half an hour, which make 48 records a day. Also, for each building on June 25 has 49 records, which should be 48. Thus, I believe the dataset has duplicate rows. Therefore, I will explore these four records. If these four records are duplicate rows, I will drop them.

Solution:

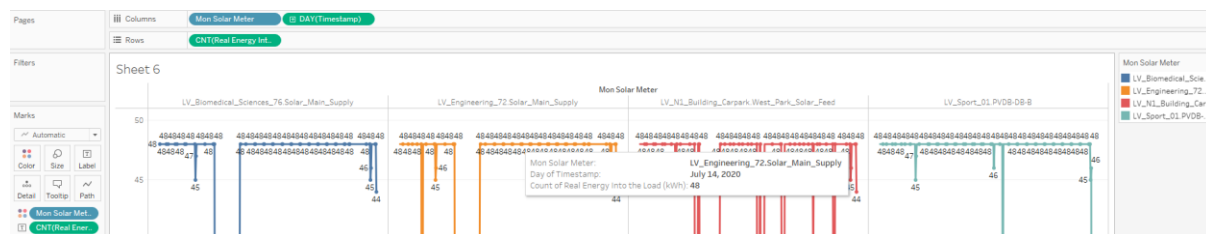


Figure 7: Correct Duplicate Rows

From Figure 7, I use python find and confirm there is one duplicate row for each building on June 25. I dropped all duplicated rows. Now, for each building on June 25 has 48 records.

3. Data exploration

3.1 Compare and contrast the average hourly (over 24 hours) solar energy generation trends of each location.

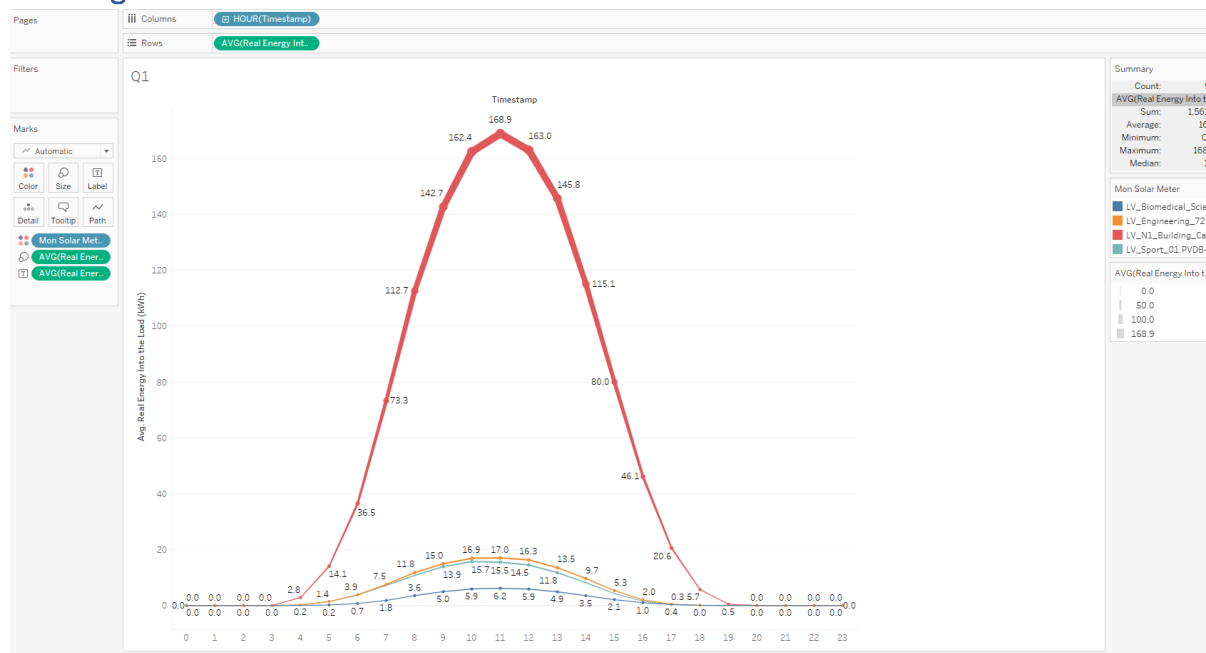


Figure 8: Compare and contrast the average hourly (over 24 hours) solar energy generation trends of each location

From Figure 8, we can observe that the peak is at 11 o'clock and 10 to 12 o'clock generated the most energy for each building. As we all know that 10 to 12 o'clock is at noon which is the hottest time in the day, which make sense why the peak time is from 10 to 12 o'clock.

As we all know, the sunrise is around 7 o'clock and the sunset around 18 o'clock. From Figure 8, we can observe some energy generated before 7 am and after 6 pm. It should be the false data (generating energy during the night) mentioned in the assignment introduction. We can see for each building that the energy generating start climb from 7 am until 11 am, then start decline until 6 pm, which follows the pattern of sunrise and sunset.

We also can observe that the building LV_N1_Building_Carpark.West_Park_Solar_Feed has a much higher average energy generated per hour. I can't find exact data, but I assume car park has flatter surface to deploy solar panels, which make car park can have more energy into the load.

3.2.1 How does the amount of energy generated vary across the entirety of 2020 for each individual location.

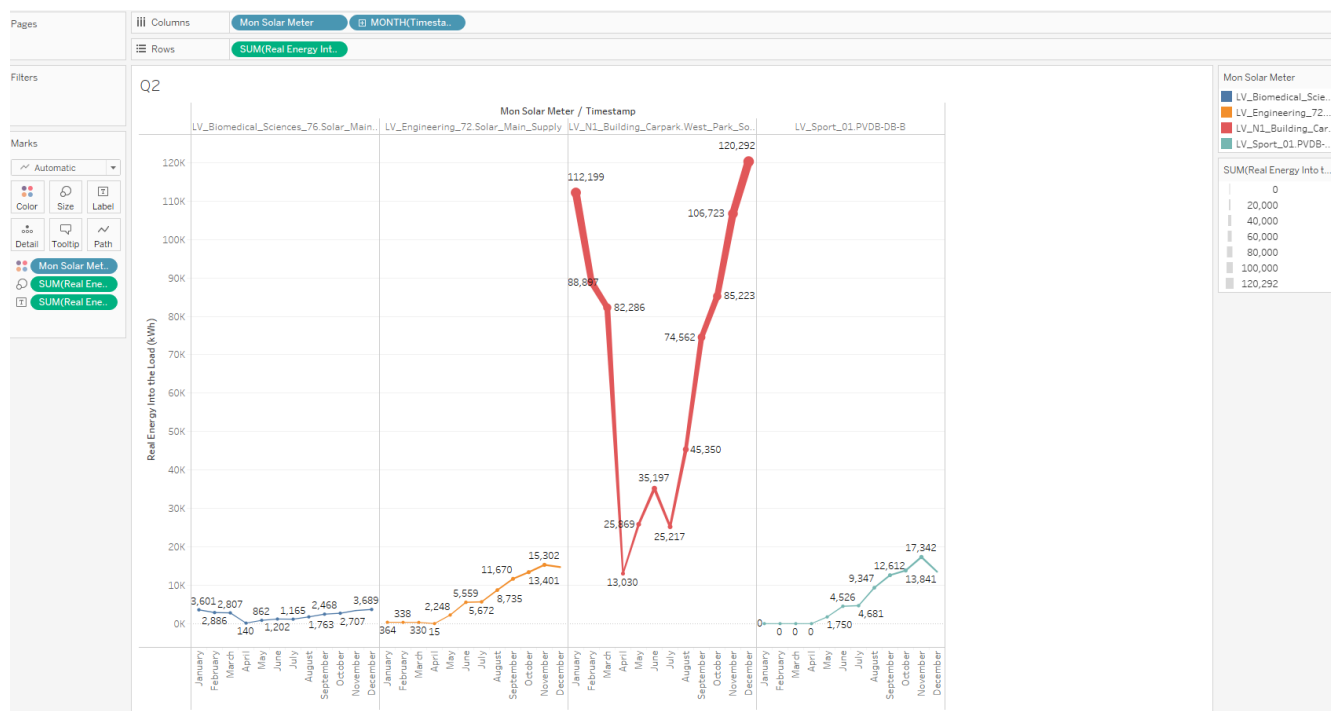


Figure 9: the amount of energy generated vary across the entirety of 2020 for each individual location

For all buildings, the months of the most energy generated into the load are from November to February. As we all know that November is the end of the spring season and summer start from December to February. These four months are the hottest months in Melbourne. Thus, it makes sense these months generated most energy into the load. From April to August generated the lowest energy into the load because it's during the fall and winter in Melbourne.

3.2.2 How does the amount of energy generated vary across the entirety of 2020 for all locations aggregated together.

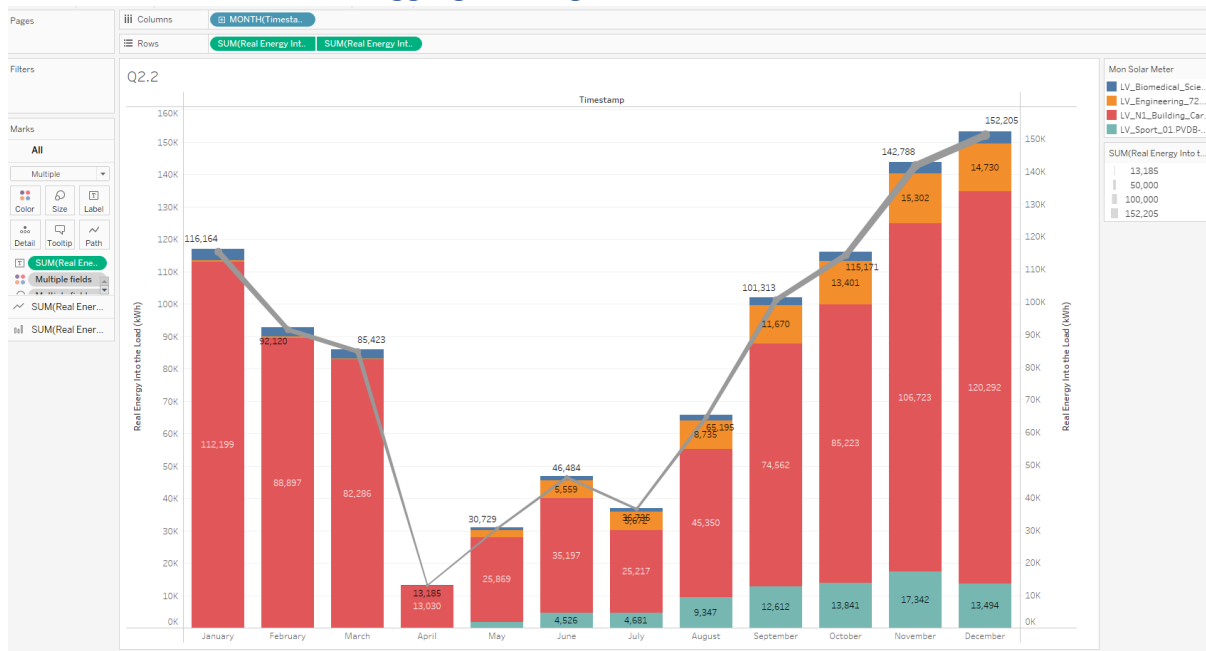


Figure 10: the amount of energy generated vary across the entirety of 2020 for all locations aggregated together

From Figure 10, we can observe the same pattern from Figure 9. From November to February generated the highest energy into the load because it's during the end of spring and summer in Melbourne. From April to August generated the lowest energy into the load because it's during the fall and winter in Melbourne. Thus, we can conclude that how much the energy generated into the load is related to the seasons.

4. Summarize

Overall, we use visualization found all issues and use python to correct them. Through visualization, we find that the LV_N1_Building_Carpark.West_Park_Solar_Feed has generated the most energy into the load, and the average of energy into the load follows a similar pattern how a daily sunrise and sunset. Furthermore, we find that how much the energy generated into the load is related to the seasons in 2020.