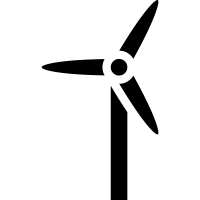
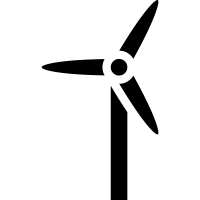
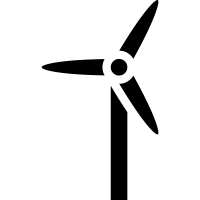
Alexsandra Guerra

Analysis of solar PV installations by utility and installer in Califoronia

Capstone Project Final Report, Springboard Data Science Intensive



Abstract

California’s progressive environmental policies have led to a great spike in residential & small-scale commercial PV installations over the recent decade. While distributed solar power is expected to continue to grow, the recent year of 2016 has seen a slowdown in the rate of adoption due to change in net-energy-metering rules limitations. State regulations and incentive play a large part in the adoption of solar. California’s three investor owned utilities (IOUs) are aiming to decrease PV system application to completion time, which has proven to also be a hindrance to PV adoption. This project provides the client, an IOU, with an interactive data set to be viewed and used by engineers and managers who work on new technology demonstration projects that focus on increasing control and use of distributed PV systems.

Contents

[Introduction 2](#_Toc478808215)

[Client & Business Need 2](#_Toc478808216)

[Deliverables 2](#_Toc478808217)

[Data Source 2](#_Toc478808218)

[Limitations of Data 3](#_Toc478808219)

[Approach 3](#_Toc478808220)

[Process Overview 3](#_Toc478808221)

[Outline of Process: 3](#_Toc478808222)

[Part 1: Clean Up 4](#_Toc478808223)

[Installer: 4](#_Toc478808224)

[App Status: 4](#_Toc478808225)

[Duplicates & Nulls 4](#_Toc478808226)

[Erroneous installation data: 4](#_Toc478808227)

[Other changes: 5](#_Toc478808228)

[Part 2: Analysis 5](#_Toc478808229)

[Part 3.1: Visualization (Python & HTML) 5](#_Toc478808230)

[Part 3.2: PowerBI Visualization 5](#_Toc478808231)

[Data Validation 8](#_Toc478808232)

[Findings 8](#_Toc478808233)

[Client Feedback 9](#_Toc478808234)

[Next Steps 9](#_Toc478808235)

[Appendix 10](#_Toc478808236)

[Screenshots of Webpage output 10](#_Toc478808237)

# Introduction

Renewable energy in increasing at a rapid pace in California. California is progressive on its environmental policies. For example, California is known globally in the energy industry for its Renewable Portfolio Standard which aims to achieve 33% of its energy from renewable energy sources by 2020, 50% by 2030 (SB 350). Regulators at California Public Utility Commission (CPUC) actively set requirements for electric utilities like Southern California Edison (SCE) to comply with. Much of the work that utilities like SCE are faced with is updating a legacy, outdated electric system with new intelligent and communication devices.

This effort is called Grid Modernization, which was born from the fact that distributed energy resources like solar PV panels and energy storage batteries are increasing at a rapid and exponential rate. These devices are changing the technical and physical requirements of a grid that was built for a different purpose, namely, to deliver energy in only ONE direction, from plant to customer. With increase of solar panels, the grid will need to change and serve as a bi-directional system for energy flow to and from customers and the utility. Being able to predict and understand the adoption of solar PV among California residents and businesses will help utilities like SCE to better meet the needs of their customers and requirements set for by the CPUC.

# Client & Business Need

Southern California Edison’s (SCE) mission is to provide clean, reliable, and safe power to its residents. In order to meet the clean energy demands of its customers, SCE has many teams and projects aimed to meet the demands for increased solar PV systems by its customers. It is also conducting studies to better understand the adoption of solar PV systems by commercial customers. SCE research and development engineers will use this analysis on a quarterly basis to keep track of the increase of solar on the grid, and to compare and track performances against other CA IOUs.

## Deliverables

This analysis will help SCE engineers to:

* Identify potential demonstration locations for technology projects that are dependent on high penetration of distributed energy resources (DERs) such as residential solar
* Provide visuals and data for presentations in support of these projects and project proposals
* Identify potential partners (solar installers) for demonstration projects

While there is available data sets and visualizations on the California Gov site, the visualizations do not separate and aggregate the data by utility and installers. The client needs this open data set to explore client-specific questions. The solutions must be:

* Refreshed on a quarterly, or more frequent basis (in fact this can be updated monthly)
* Available to internal stakeholders via common and accessible platform (SharePoint)

# Data Source

California Solar Statistics publishes all investor owned utility (IOU) solar PV net energy metering (NEM) interconnection data per CPUC Decision (D.)14-11-001, and all IOU data from the California Solar Initiative incentive program per California Senate Bill 1 (SB-1). Data available at:  <https://www.californiasolarstatistics.ca.gov/data_downloads/>. With this data accessible, California residents and vendors have more information available to them. Vendors and utilities are able to keep track of their performance of PV interconnections, thus allowing for higher adoption rates of solar PV panels. **The specific dataset examined was the Net Energy Metering (NEM)** Current Interconnection Dataset**.** A **NEM qualified system** is one that meets the following criteria:

* Total capacity under 1000 kW (1MW)
* System sized to offset part or all of past 12-month billing history
* Rule 21 Compliant (an electrical and power systems set of requirements)

### Limitations of Data

The data set is limited to the record keeping of the individual utilities that provide the data to CSI. For example, only data that was required after a date was included, so the complete history and total installations is not represented. Additionally the data required significant cleaning and matching of installer names. Data also included erroneous data, such as large installations (over 1MW) which do not technically qualify as NEM and therefore should not be in the dataset.

# Approach

## Process Overview

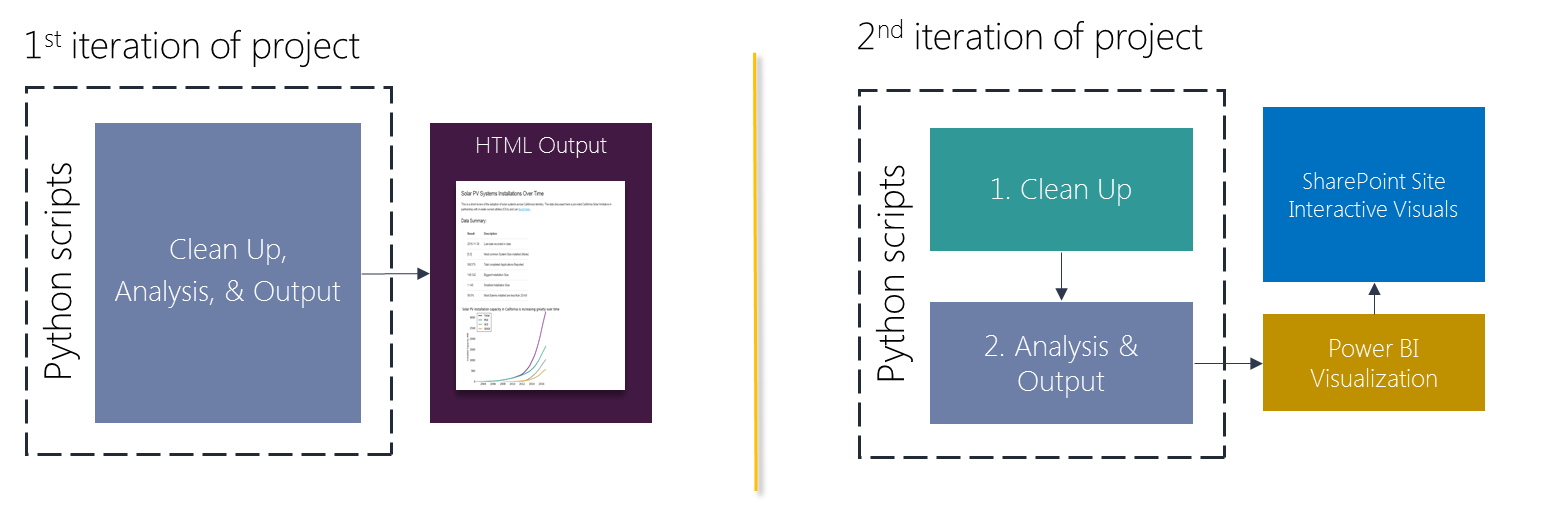
After discussing with the client how the data should be available to internal engineers, it was determined that SharePoint would be a suitable platform to share the data analysis and report. The project was to provide a script to clean up data and visualize installations for IOUs and vendors as described above in an HTML output page that would be embedded on a SharePoint team site, and therefore also accessible easily to other teams within the clients’ organization.

However, the project deliverables shifted, or pivoted, after a few meetings with the client and after discovering a technology alignment. **Figure 1: Project Evolution & Iterations** illustrates this. After the milestone report and after showing the client the current work to output to HTML (static), it was requested to provide visualizations that engineers could interact with. It was also discussed that internal teams were looking at various data analysis and visualization packages, including PowerBI, a business analytics service provided by Microsoft. It provides interactive visualizations with self-service business intelligence capabilities, where end users can create reports and dashboards by themselves, without having to depend on any information technology staff or database administrator. Aligning with this new adoption of PowerBI by the client’s team, PowerBI was chosen as the platform to visualize the data. The scripts were then updated and organized to output pandas DataFrames to cvs that could easily be queried and refreshed automatically with PowerBI, and an interactive report instantly generated and embedded to a SharePoint Site.

Below is a list of steps taken throughout the process. These steps are referred to throughout the report.

### Outline of Process:

1. **Data exploration** – looked at the data using python summary tools, histograms, and plots. Plotted many different representations and grouping of the data.
2. Presented to client – feedback from client identified which plots and information was most helpful and met client needs
3. **First iteration of project:**
   1. Created data cleaning scripts – this process was the longest part of the project. (see [Part 1: Clean Up](#_Part_1:_Clean))
   2. Generated HTML reporting scripts
   3. Presented HTML report and visualization to client – feedback from the client identified data accuracy concerns
   4. Validated data and investigated data insights
   5. Met with client to discuss findings
4. **Second iteration of project**
   1. Client asks for new visualization tool – PowerBI in SharePoint
   2. Split up scripts for faster processing and organization – outputs specific dataframes for PowerBI
   3. Created PowerBI queries of dataframe outputs
   4. Created visualizations with PowerBI and new interactive filtering of data set
5. Presented to client
6. Identified next steps



**Figure 1: Project Evolution & Iterations**

## Part 1: Clean Up

Needed to reformat string values to be consistent (all uppercase) The data required a lot of clean up. The final code submitted is just a condensed fraction of the code that was used to explore and analyze and determine the clean-up steps. The process of discovering these issues was long and iterative.

### Installer:

The clean-up process first started with the installer names which was noticed in the first stage, 1. Data Exploration. I analyzed the list of unique values for each installer in order to understand the variability of installer spelling. I noticed that the significant majority of issues came from variations in capitalization of the vendors, such as SunRun vs Sunrun or SolarCity vs Solarcity.

### App Status:

Only applications with status of “installed” were included. The client was interested only in the current capacity of solar PV systems connected to the grid in California.

### Duplicates & Nulls

Although no duplicates were found in the dataset currently available, it is good practice to include this especially since this script is to be used in the future on updated data sets that may or may not include duplicate entries. Thus this step was mostly **preventative.** Dropped duplicates of the following subset of columns was performed: 'System Size DC', 'Application Id', 'Utility', 'App Complete Date', 'Service Zip'.

In the most recent dataset, December 2016, a total of 12,105 rows with null/NA ‘System Size DC’ values were removed. System Size DC is the key metric looked at in this dataset.

### Erroneous installation data:

The 0.5% quantiles were removed from the dataset as a precaution and as a way to remove erroneous data. Additionally, any system above 1000 kW in System Size DC was removed, since these installations violate the criteria for this data set where NEM installations are less than 1MW (1000 kW). After step 3 of this process, it was discovered that many applications were recorded to have system size above 1MW, as the installed capacity was determined to be different than other recorded values. See [Data Validation](#_Data_Validation) section.

### Other changes:

The dataset was used with time-index and was shrunk by 60 irrelevant and empty columns that the client was not interested in. This made the data set easier to work with as it was smaller and required less memory.

## Part 2: Analysis

The client wanted to understand the adoption rates of NEM systems. Thus, the installations added over time was aggregated over yearly, quarterly, and monthly frequency.

Dataframes that were created and saved for input into PowerBI:

* Cumulative sum of PV installations - A cumulative sum of installed PV capacity calculated over time.
* Installations by city – grouping the installations by Service City and summing System Size for each city, but preserving the time values. This includes number of installations, total installed capacity, and average installation size
* Top PV installers, by utility – the number of installations and the total capacity for each utility was determined by looping and filtering through each Installer name. Each row was an instance of # of installs, total install capacity, and utility
* Top PV installers – this is similar to the above dataframe except it does not include utility. It looks at the installer performance over all territories. It is used to give context to their regional performance
* Note: some of the data from these dataframes were also used in the HTML visualizations and output

## Part 3.1: Visualization (Python & HTML)

The folder directory uploaded to GitHub (Capstone Project Files) is the directory for this webpage output. A template html file was used and then updated in the python script #2, with the summary details of the results, as well as all of the plots that were identified by the customer in step 2 of this process – Present exploration visuals to client and discuss.

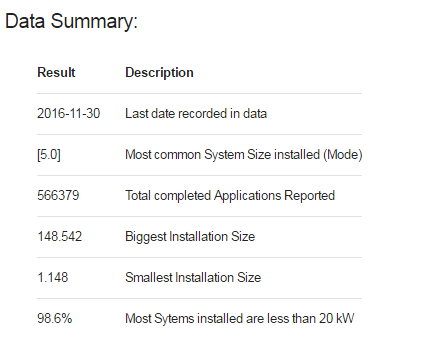
The **index.html** in the github directory is the output HTML webpage with the plots and description of the data. Figure 2: Summary table for HTML Outputis a screenshot of the summary table that was created from the scripts and output to this page that summarized key characteristics of the data and analysis. For convenience, instead of going to the github files, screenshots of the webpage are included in the [Appendix](#_Screenshots_of_Webpage) of this report.

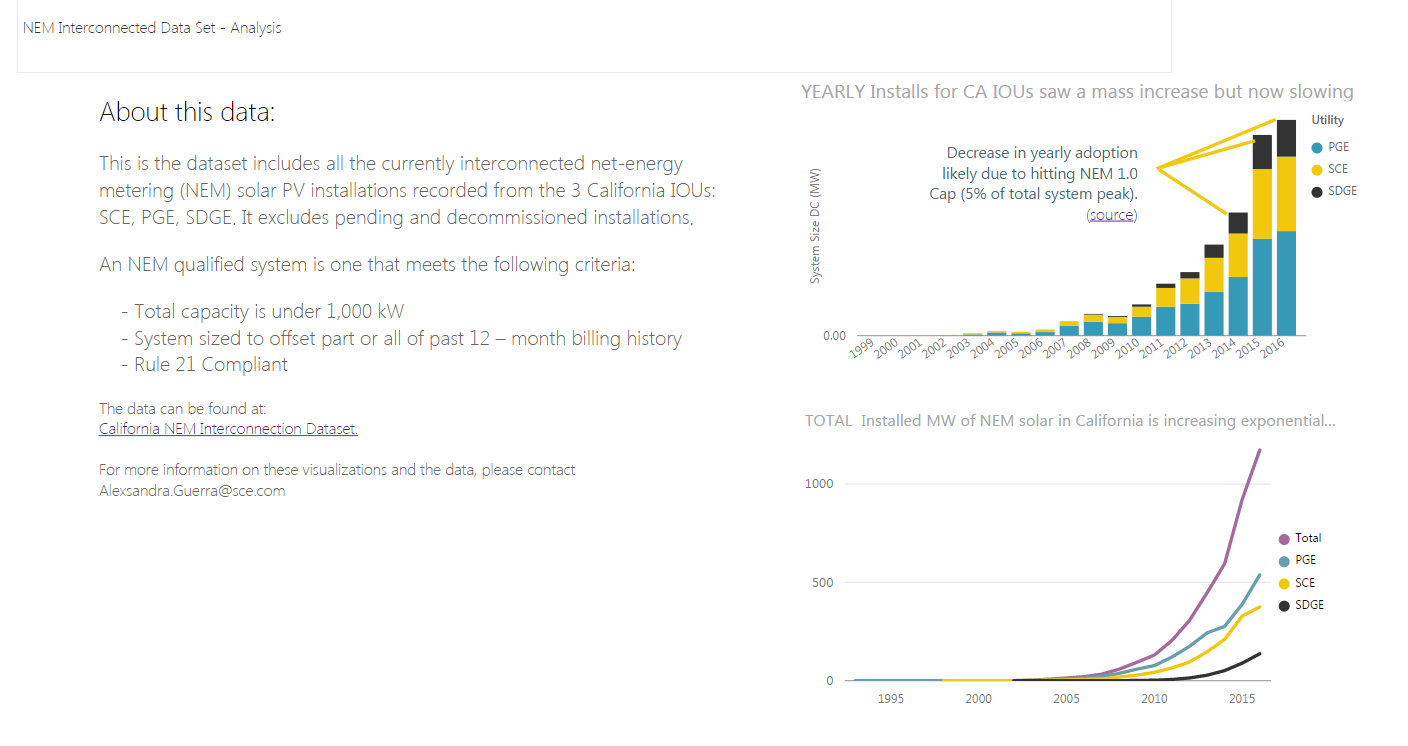
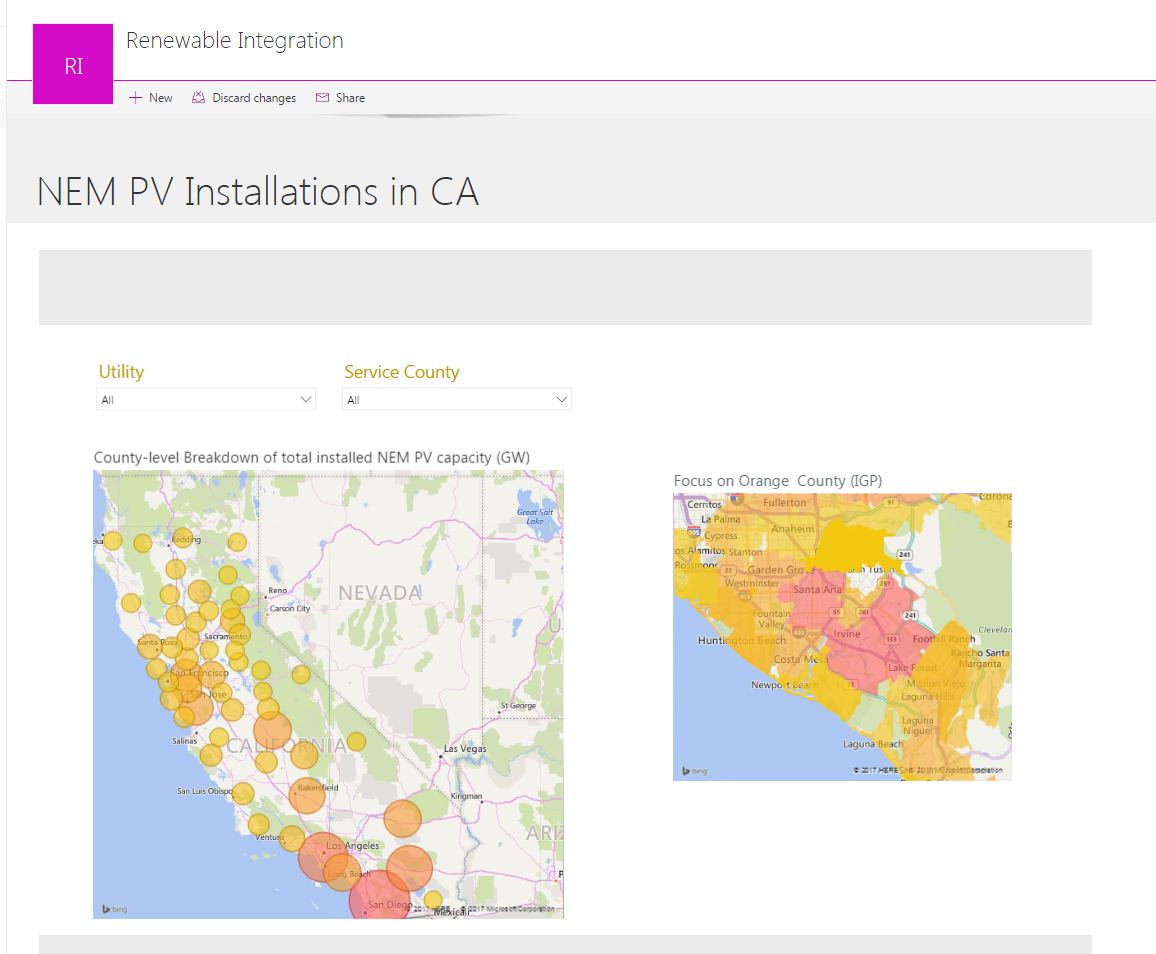
Figure : Summary table for HTML Output

## Part 3.2: PowerBI Visualization

PowerBI is a useful tool for data analysis and visualization. However, python is a more agile and easier to use tool for data clean up (in my opinion) and so the work that was already done for this project was used to output several summary DataFrames that were then input into the PowerBI software tool. This allowed for quick upload and automatic visualization of the updated data from the python scripts. PowerBI also provides and link (public) for users to embed into web pages, including SharePoint. Below are screenshots of the three pages of visualizations included in the interactive PowerBi report. It includes slicers and filters that allow users to filter the data.

*Figure 3: NEM yearly install and cumulative installs* shows the first page (of 3) from the interactive web output on SharePoint. It gives information and context to the data, as well as pointing out some finding about slowed adoption (see next section). It provides two plots: yearly added capacity for each IOU and cumulative installed PV capacity for each IOU over time as well as the combined of all 3 IOUs.

*Figure 4: Installations Grouped By County* shows a breakdown of installations per county. This was down using the Installations by city dataframe mentioned in the previous section, linked to a lookup table of the filtered dataset that matches county to city. Each bubble represents a count and the color and size of each bubble represents the total installed NEM in that county (left plot). The image to the right is a zoomed in heat map of the Orange County area, where the client is demonstrating many projects, and is thus of interest to the client. This heat map breaks the installation down by city and is a good lens for looking at potential partner cities. *Figure 5: Example of slicer filter and data tooltips* illustrates the interactiveness of these plots. The selection boxes allow filtering of the graphic and plotted bubbles. Additionally, in all of the plots in this tool, hovering the mouse over the data point can reveal more information (which can be added by linking common fields in separate datasets (dataframes)).



1 of 3 >

Figure : NEM yearly install and cumulative installs

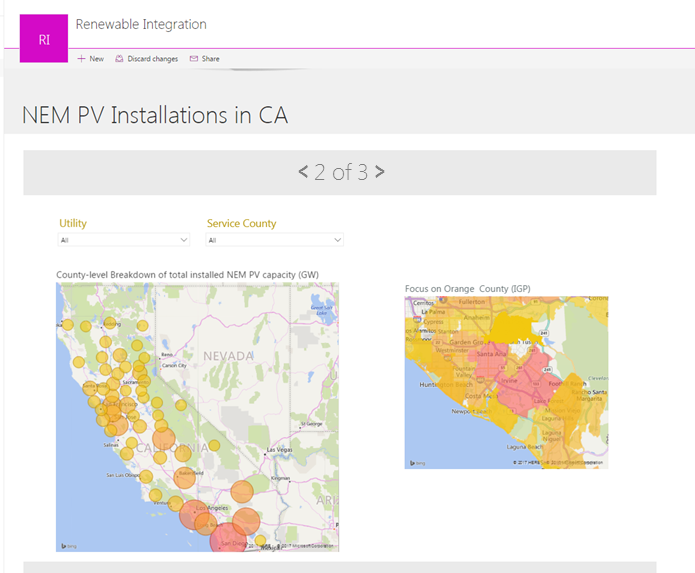


Figure : Installations Grouped By County

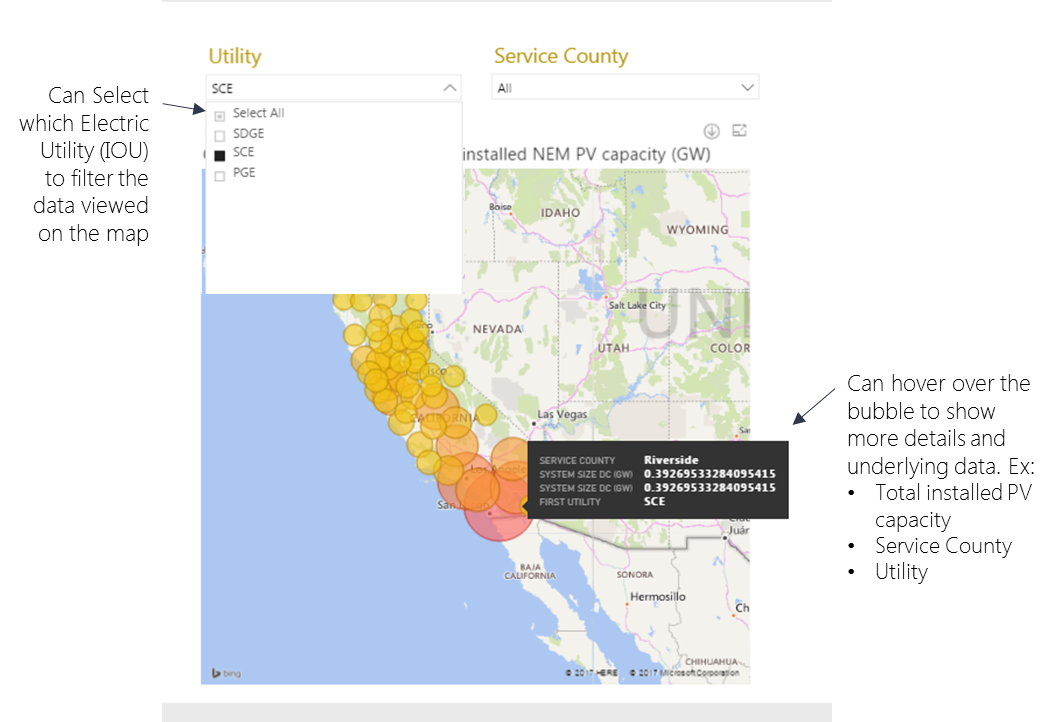


Figure : Example of slicer filter and data tooltips

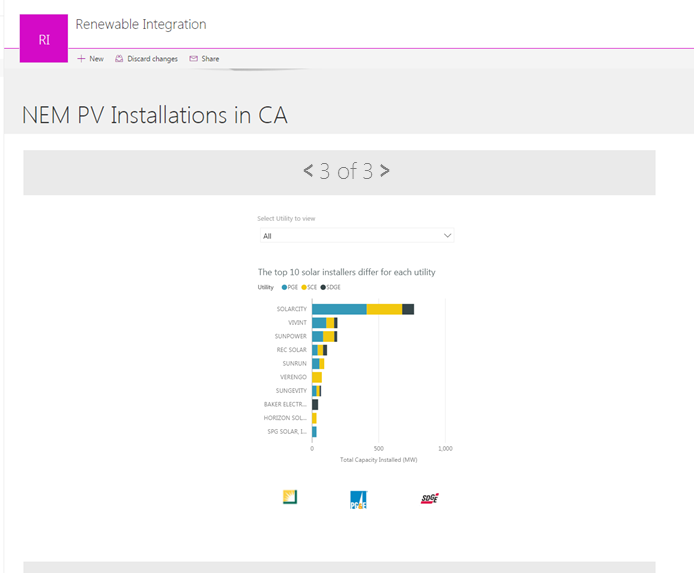


Figure : Top Installers for the three IOUs

## Data Validation

In order to validate the data and visuals, internal data from the client that recorded solar PV installation was used. It is likely that the internal dataset is the source of the publically available dataset. However, validation was conducted not just to verify the data, but to also verify that the scripting and visualizatiosn were correct. Thus, the working data set (NEM from ca.gov) was validated against the internal client data using other visualizations tools,SAP Lumira and PowerBI. Infact, this process contributed to the pivot that resulted in the 2nd iteration with PowerBI webpage and visualizations as the deliverable. The plots created in the HTML output were validated against the plots in these tools. The cumulative sum of installations for the client, SCE, was verified to be within .1 MW of the plot and working dataset, which was acceptable to the client.

The data in the final PowerBi page was also validated against the public visualization[[1]](#footnote-1) and data that also included publically owned utilities (POUs) such as Los Angeles Department of Water and Power (LADWP). The data from Figure 4: Installations Grouped By County was seen to be of the same magnitude, where in the NEM PowerBI set, Sand Diego County has 606 MW installed NEM, and in the entire set of PV installations there is a recorded 636 MW.

# Findings

* While the data set here of IOU installed NEM solar PV systems is only a subset of PV distributed installations in California, it makes up approximately and thus is a reasonable representation of the total California system, as the difference is less than 3%.
* There is a slowdown in adoption of PV, where the rate of change in added solar capacity yearly is decreasing. This is likely due to:
  + NEM 1.0 quotas being met. Regulators set the NEM 1.0 (first version) to cap incentives at 5% of the peak capacity of each utility’s system. NEM 2.0 is now coming into effect. This is likely true for PGE[[2]](#endnote-1), however SCE has not yet met its NEM limit.[[3]](#endnote-2)
  + Saturation: The cities and customers who are most likely to adopt PV systems have already adopted and a natural saturation is being met.
  + Nonetheless, there is still a large growth in NEM installations that can be expected.
* San Diego County has the most solar NEM installations than any other county, Los Angeles being second. However, the dataset does not include LADWP (Los Angeles Department of Water and Power), which is publically owned and does not have to report these installations like the IOUs, and therefore not included in the dataset. Thus, the recorded difference here in Solar PV between these two counties is larger than the actual difference.
* Each utility has its own top installers. In San Diego (San Diego Gas and Electric), the top installers include smaller installers who are not prominent in the larger territories of PGE and SCE, such as: Baker Electric, Sullivan, and Stellar Solar, which are not even in the top 10 installers of SCE or PGE.
* Solar City is the strongest player in installing solar systems in California’s large IOUs. It has over 122,000 installations and 760 MW of installed NEM solar, out-performing the runner-up, Vivint, by over 300% which has approx. 33,000 installations and 190 MW of added capacity.

# Client Feedback

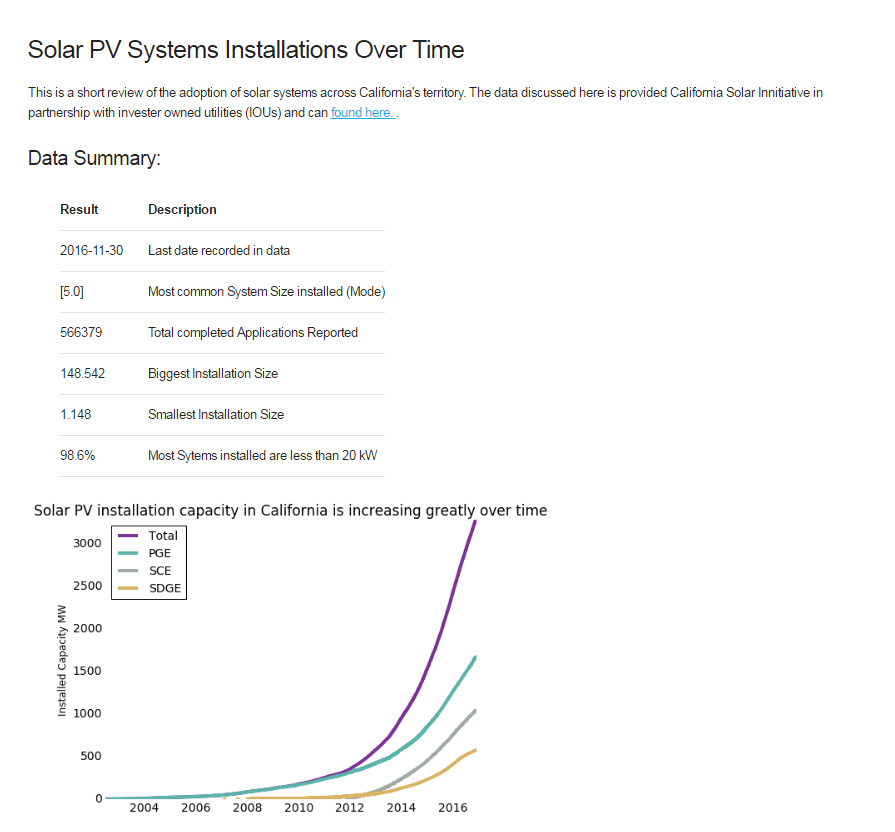
The client had positive feedback on the final project shared on the client’s team’s SharePoint site. The client particularly likes the county and city bubble & heat maps. The client asked that the link be send to an internal stakeholder who is responsible for keeping the internal records of installations, and who would have interest in using the site as a tool for identifying cities in SCE’s territory that could serve as potential partners on projects. The client also asked this to be send to other principle project managers who could use the tool and the graphics in presentation and in analysis. The client also liked the first page that showed a high level overview of installations. This provided the client with a utility-level insight into PV adoption, which is not currently publically available. The client also identified possible next steps (see below).

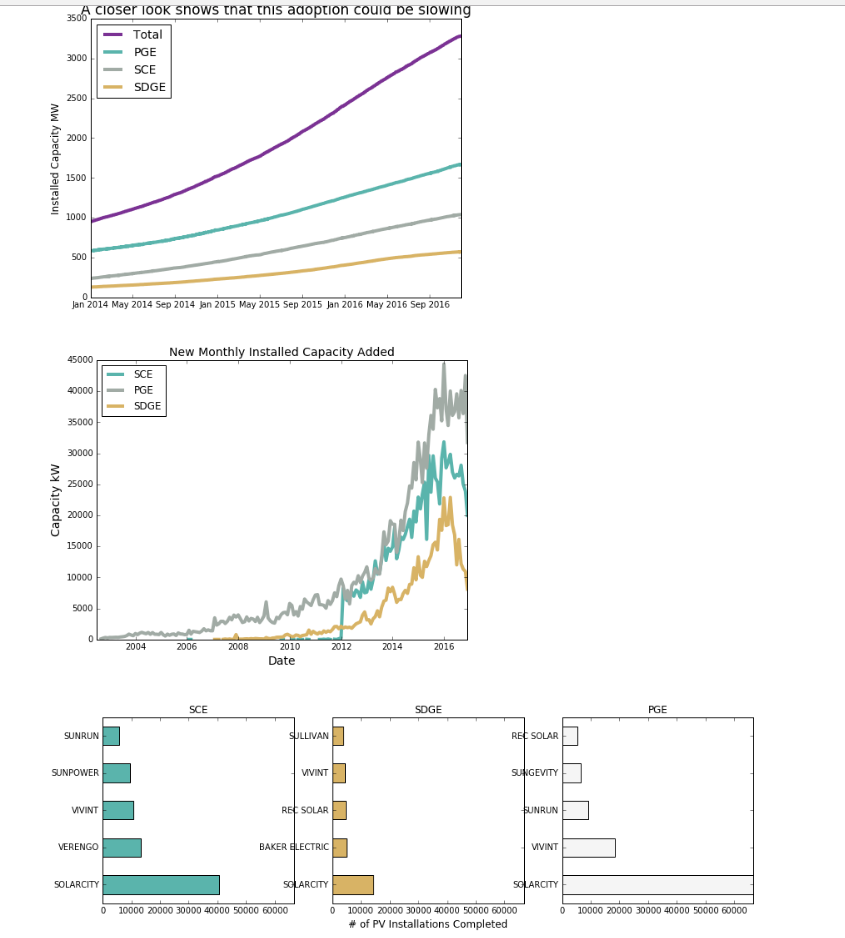
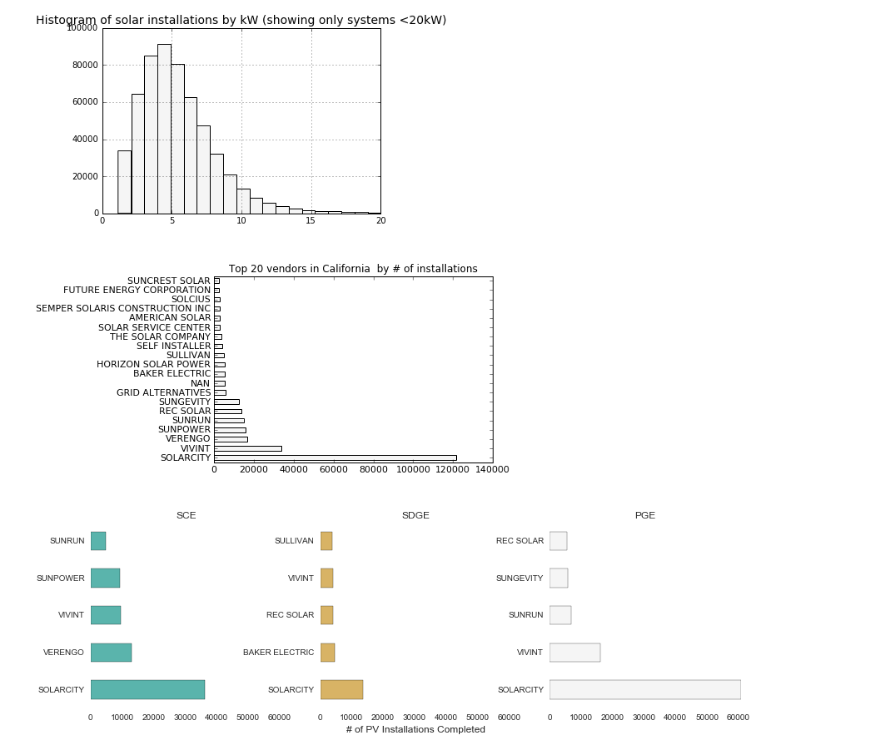
# Next Steps

* The client suggested that it would be good to see the performance of utilities in terms of customer base size and territory size. This would give context to the penetration of PV systems per utility.
* The client also suggested that adding a page to keep track of installation application turn over time. SCE and the client are constantly working to decrease application-to-interconnection time, i.e. the time it take for permitting and approval before the customer can actually connect their system to the grid. This has effected PV adoption in the past. Reducing these barriers is an aim of California’s citizens, regulators, and utilities.

# Appendix

## Screenshots of Webpage output



1. http://www.californiadgstats.ca.gov/charts/ [↑](#footnote-ref-1)
2. http://bit.ly/2c4ZYVL [↑](#endnote-ref-1)
3. http://on.sce.com/2onqGPO [↑](#endnote-ref-2)