Analysis of California Solar Customer Demographics and Market Potential

EME 198: Statistical Methods in Design and Manufacturing
Prepared For: Cristina Davis and Stephanie Fung

Team Members:
Katherine Wikler
Ian de Vries
Trent Storm

Summary

This project aims to draw conclusions by comparing large data sets on solar panel installation and demographics from the National Renewable Energy Laboratory's OpenPV project and United States Census data. The scope of our collected data is narrowed down to residential solar installations only and is based on zip codes within California. By comparing data and highlighting trends using Python programming, we aimed to pinpoint which demographics of Californians are most likely to install solar panels, which areas would benefit the most from widespread solar installation, and which areas are the most marketable to sell solar panels. Then, based on our conclusions, we can perform financial analysis to see how much energy and money a single household could save over time.

To answer who would purchase solar panels, we looked at PV data that details all CA solar installations from the age of the buyer to their income level. For age and income level, we generated histograms and computed averages using Python to conclude that middle aged, upper-middle class homeowners are most likely to purchase solar panels.

To answer what areas would benefit the most from solar installation, we looked at weather data and observed areas of high average annual direct normal irradiance (DNI). The zip codes with the highest intensities were deemed to be the most beneficial areas to install.

To answer what areas are the most marketable, we looked at areas where there was very little to no solar installation already. Then, we compared these zip codes with the zip codes that showed high annual sun intensity. Zip codes that have high annual sun intensity and don't have solar already installed were deemed as the most marketable areas.

Our results showed that the typical solar panel customer is middle-aged (around 40 years old) with an income of $66,954 per year. The most beneficial areas to install included 24 of California's 58 counties (744 zip codes) which had higher than average annual DNI from the sun between 5-6 kW/m^2 per day. The most marketable areas to install solar included 22 of the 24 beneficial counties but only 478 zip codes. The zip codes were restricted by a median income level between $50-120k per year. The counties that were marketable had 266 zip codes that did not meet the minimum income level cutoff. The two counties that were deemed not marketable were Sierra and Glen.

Introduction

Many people wonder if solar panel installation is a worthwhile investment or just tree-hugger propaganda. To put it simply, it's no hoax. The benefits of installing solar heavily outweigh the upfront cost.

Historically, energy rates in California have risen at a rate of 6.7% each year. Installing solar significantly reduces energy bill costs and adds instant value to any home. According to the Appraisal Institute, a household's value will increase $20 for every $1 in annual utility bill savings due to solar energy equipment. For example, a $26,000 home photovoltaic (PV) system saves around $3,000 each year in energy expenses, which equates to a $60,000 increase in the home's value (1). Therefore, after 9 years of having solar installed, the PV system will pay for itself in addition to adding $60k in appraisal value to the house. The return on investment of solar installation proves that it is a good long-term investment and provides sustainability despite rising energy costs.

Not only do solar panels save you money, they also protect the environment. Renewable energy systems like solar panels reduce the amount of fossil fuels burned, thus reducing carbon dioxide emissions. Solar energy systems help decrease air pollution, global warming, acid rain, and other dangerous environmental impacts. Over 25 years, an average 4.5 kW home PV system will offset 278,000 lbs of carbon dioxide (1).

The purpose of this project is to draw conclusions by comparing PV data with demographics of residential areas in California. Through data analysis, we can determine the most marketable areas for solar installation and who is the most likely to purchase solar panels. Furthermore, we can perform financial analysis and see how much energy and money a single household can earn over time.

Methods

Python was the programming language used to analyze data from NREL and the U.S. Census Bureau, primarily because of its ability to handle large datasets and perform efficient statistical calculations and analysis using packages like pandas, seaborn, and numpy.

Data first needed to be pulled from both NREL and U.S. Census Bureau websites. Given more time, a Python-based API would have been developed to query and download data however for this project, data was pulled manually from the websites and saved in .csv (comma-separated value) format. NREL's OpenPV website allowed filtering by state and the resulting file was extremely large with almost 630,000 data entries. The U.S. Census Bureau's download website accommodate a broad range of filtering options; for this project, data was filtered by state and zipcode and chosen out of the American Community Survey (ACS) database for median age, household mean income, and household median income per zip code.

The OpenPV file was uploaded into a Python Jupyter notebook via Anaconda and parsed to include only relevant columns including zip code, PV system size (kW), and installation type (residential, commercial, utility, etc.). This project only focuses on residential PV systems, so data was filtered to only include residential entries capped at a maximum system size of 50 kW. Census data was also uploaded into Jupyter and parsed to include only median age per zip

code, household mean income per zip code, and household median income per zip code. Next, datasets were merged together to create a master dataset from which statistical parameters could be analyzed.

All data (household mean income, household median income, median age, and PV system size) was plotted as a histogram into bins to show individual parameter distributions (n = 50). A simple summary was also performed for each parameter to display mean, median, standard deviation, minimum/maximum values, and 25th, 50th, and 75th percentile values for the data. Data was also plotted in box plot format to show distributions in a different format and see potential outliers.

Next, scatter plots were created to compare household mean income, household median income, and median age against corresponding PV system size. The plots appear quite dense because there are almost 1500 data points being plotted. A quick linear regression was run to check correlation between each of the two data sets, however all R^2 values were extremely small, indicating that data would need to be further manipulated to run a satisfactory analysis. Household income data was rounded to the nearest $10,000 to create 32 income brackets ranging from $20,000 to $430,000 and the corresponding PV system size was averaged within each income bracket. Data was again plotted as a scatter plot and a linear regression was performed. Python's statsmodel package was used to perform an Ordinary Least Squares (OLS) regression analysis that provided values for regression line coefficients, goodness of fit, and confidence intervals. This method was applied to household mean income data, household median income data, and median age data.

After observing adequate correlation between PV system size and household income, it was necessary to identify zip codes that have not installed any PV capacity, i.e. the zip codes that were present on U.S. Census dataset but not present on the OpenPV dataset. The distribution of household median income for homes without any PV was plotted as a histogram using n = 100 to obtain more granularity. Since the relationship between PV system size and household median income showed the strongest correlation, that regression equation was used to predict PV system size for homes without any PV, based on their household median income. Assuming families within a certain income range would be more likely to install PV than others (between $50,000 and $120,000 based on the household median income residuals plot), a list of zip codes and counties that should install PV, and how much they are likely to install, was generated.

Next, GIS tools were utilized to map results directly onto a map of California (2). A dataset that correlates latitude and longitude values with zip code and county was imported and merged with the master dataset. It was also necessary to import the Basemaps package from mpl_toolkits into Anaconda. A map was created showing the distribution of installed residential PV systems across California, with another map showing commercial PV system distribution for reference. Maps were also created showing the areas where PV could be installed based on the estimation

analysis, restricted to the $50,000 - $120,000 income range and not restricted to any income range.

Finally, data for average annual direct normal irradiance (DNI) per county was imported and parsed. This data was merged with the master dataset and compared to the locations where PV installation was recommended, in order to analyze the economic payback of the system and see if PV installation would actually make sense in that area based on the amount of solar radiation the area received.

Discussion and Results

After comparing PV and demographic datasets, we determined that the average age that someone is most likely to purchase solar is 40 years old. We also found that the median income level of a single household that installs solar is around $66,954 per year, which is slightly above the median household income in CA of $63,783. The breakdown of household income by age is shown in the table below.

Median Household Income in California by Age of Householder

	California
Householder under 25 years	$30,785
Householder 25 to 44 years	$66,979
Householder 45 to 64 years	$76,370
Householder 65 years and over	$46,749

Figure 1:Median household income in California by age

Our demographic results make sense because middle-aged households make more money than others on average and are also young enough to see the long-term benefits of solar energy. Also, the median household income with solar installed is only slightly above the overall median income in California. We expected it to be higher because we thought that higher income households would be able to afford the initial cost of installation. Perhaps more and more people are becoming knowledgeable of the return on investment of installing solar, despite economic status.

A summary of demographic results from U.S. Census data are tabulated below, including averages, standard deviations, and 25th - 75th percentiles.

Table 1: Demographic Results, Summary

	Household Mean Income (USD)	Household Median Income (USD)	Median Age (Years)
Mean	72076.43	56172.15	44.54
Standard Deviation	39504.37	30776.07	11.32
Maximum	252925.00	202563.00	68.90
Minimum	24256.00	14758.00	18.80

The California plots show each county's commercial versus residential system size, average solar energy system's kW power, and their total direct normal irradiance (DNI), which is a measure of the sun's intensity in that area. By comparing the system size and the DNI plots, we can see which areas have solar installed already and which areas receive the most radiation from the sun. The areas with the highest DNI are the areas that would benefit the most from solar installation. The areas with low amounts of solar installation and high DNI are the most marketable counties (see fig. 2).

After performing linear regression on the datasets, we found that the strongest correlation exists between average PV system size (kW) and household median income, with an R^2 value of 0.668. OLS regression results gave an equation for the linear regression line which was used to make predictions about PV system size installations for counties and zip codes that did not currently have PV installed (according to the OpenPV dataset).

By filtering the master dataset to include only counties and zip codes that had no PV installed, a list of recommended areas for PV installation was generated by zip code and county. Restricting the dataset to include only household median incomes between $50,000 - $120,000 and average annual DNI between 5 - 6 kW/m^2/day resulted in 478 zip codes among 22 counties where it would be beneficial to install a PV system. A list of these 22 counties can be found in the appendix, along with the 727 zip codes and 24 counties that are just restricted to average annual DNI production.

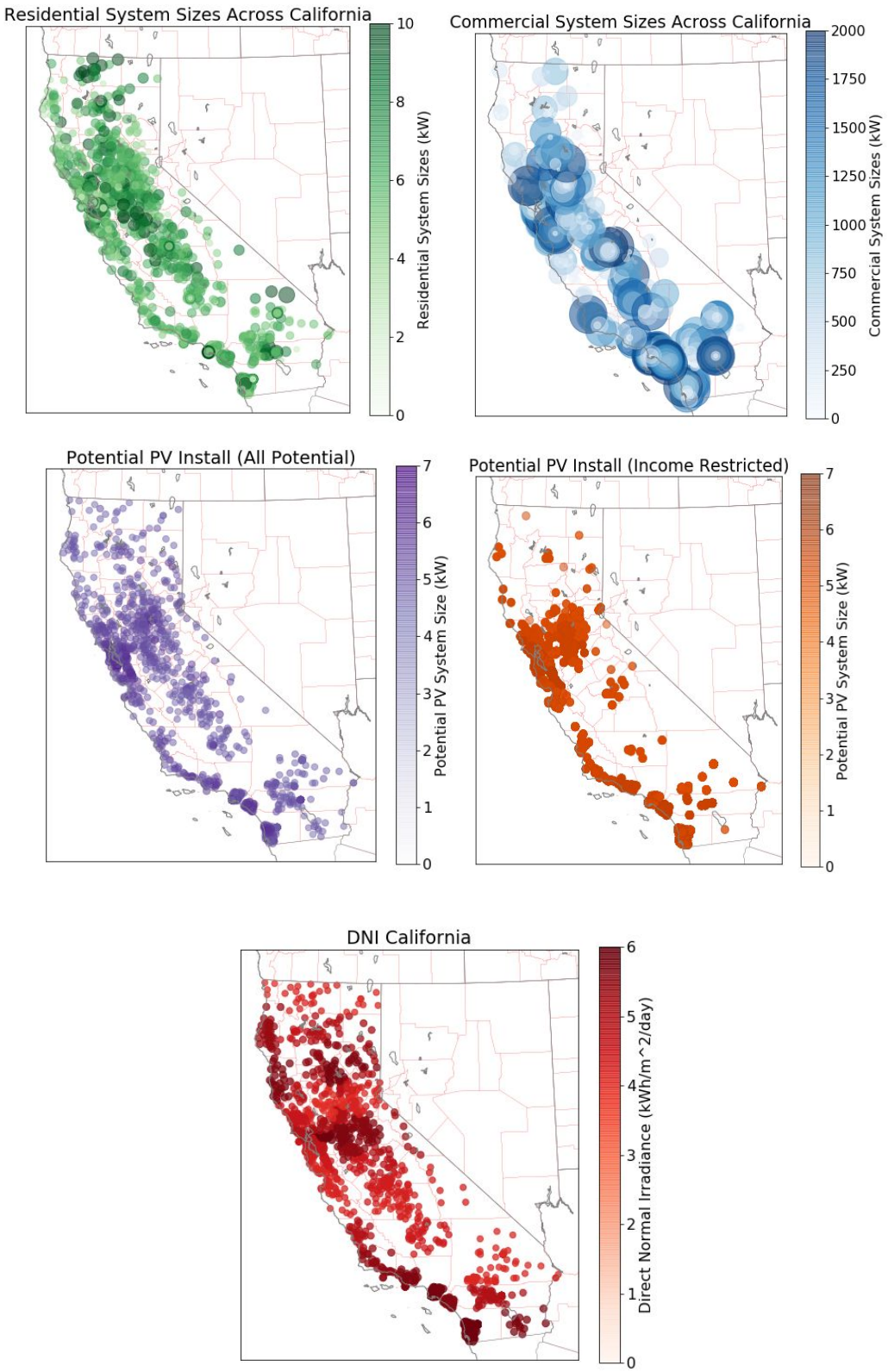


Figure 2: Data maps of system sizes, potential size, and DNI values across California

Conclusion

Python was utilized to analyze linear regression which showed which showed us which demographics correlated with solar panel installation. The data collected has shown that residential solar installation is not correlated to income, as it seems that household incomes across the spectrum are installing solar panels, with the median annual salary of solar customers being $66,954. It was also found that the median income was much more representative than the mean, due to high income outliers as shown in Figures A11 and A14. The median age of solar installations was 40 years old, which shows that middle aged households have a higher chance for solar installation, since this is around the age most people reach the median annual salary of solar customers. The areas with the most DNI, around 5-6 kW/m^2 per day included 24 counties (744 zip codes). These areas are the most beneficial for solar panels to be installed. The most marketable areas to install included 22 of the 24 counties (478 zip codes) which was restricted by income between $50-120k per year as shown in Table A1. Many of the 22 counties have certain zip codes that have low income, but only the entirety of Sierra County and Glenn County have income that is too low to be marketable.

In future projects it would be beneficial to conduct an in-depth analysis of financial gains from these marketable counties, as well as reduction in emissions if these gains are met. This information is extremely useful to companies related to solar energy as they can help improve the solar capacity across all of California. It would also be beneficial to examine more granular census data, including a breakdown of income range by age. This data would help us investigate if people in their 40s (who, on average, have the highest incomes) are truly the largest installers of PV systems.

This project was challenging because the team chose to code in Python, a new and somewhat unfamiliar coding language outside the typical Mechanical Engineering curriculum. However, Python proved easier to use, had better tools for statistical analysis, and could provide detailed graphics that best represented the data. It was simple to import all necessary datasets and parse/clean the data into appropriate datasets that could be used for analysis. Additionally, the team found simple linear regression straightforward, thanks to tools like OLS regression and packages like seaborn and pandas. However, it proved challenging to create California distribution maps because the team had to learn how to import and use GIS packages in Python. Luckily, Python is well-documented online and various forums like Stack Exchange provided helpful tips to guide our analysis.

Sources

[1] https://www.incomebyzipcode.com/california

[2] https://jakevdp.github.io/PythonDataScienceHandbook/04.13-geographic-data-with-basemap.html&sa=D&ust=1527897959177000&usg=AFQjCNEpgmODeU4mJy4MAhCDjjlY2xBmuA

Appendix

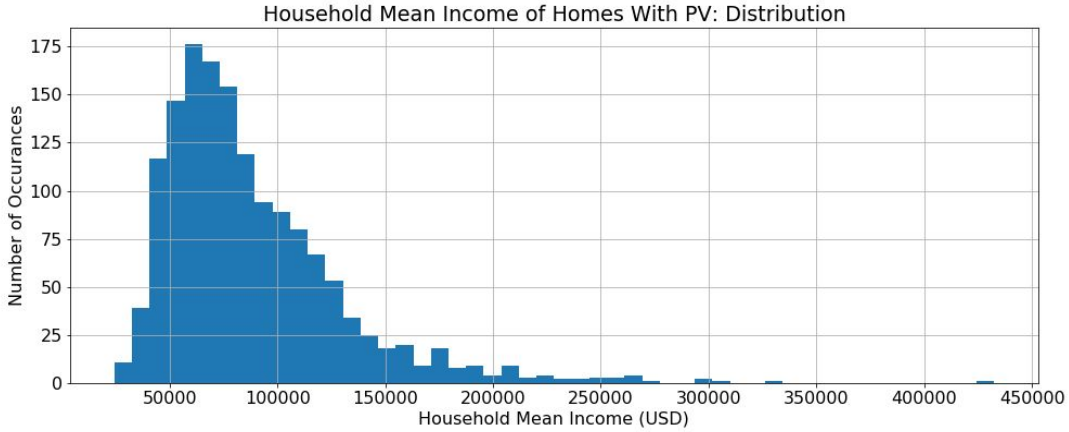


Figure A1: Mean income of homes in certain zip codes with solar installation

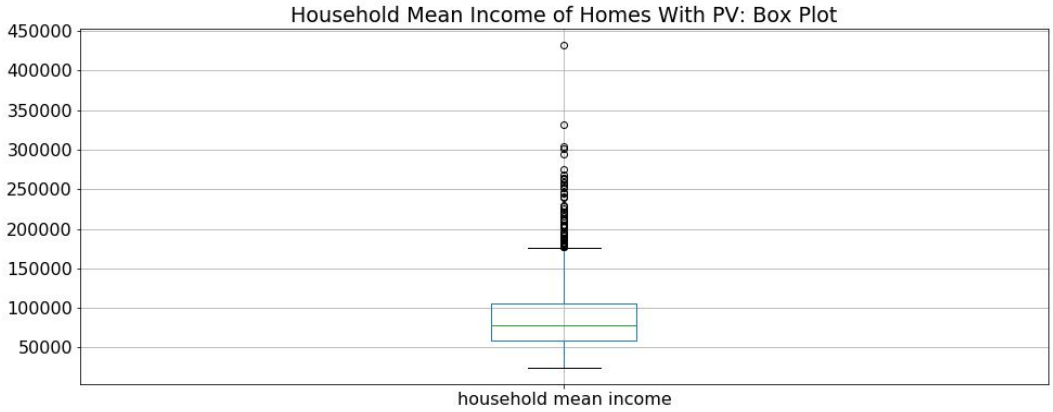


Figure A2: Mean income of homes with solar installation box plot shows high income outliers

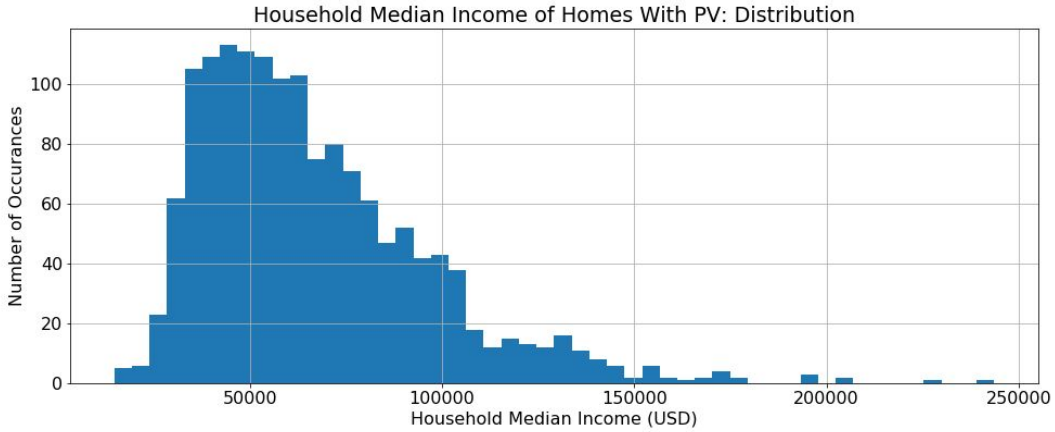


Figure A3: Median income of homes in certain zip codes that have installed solar.

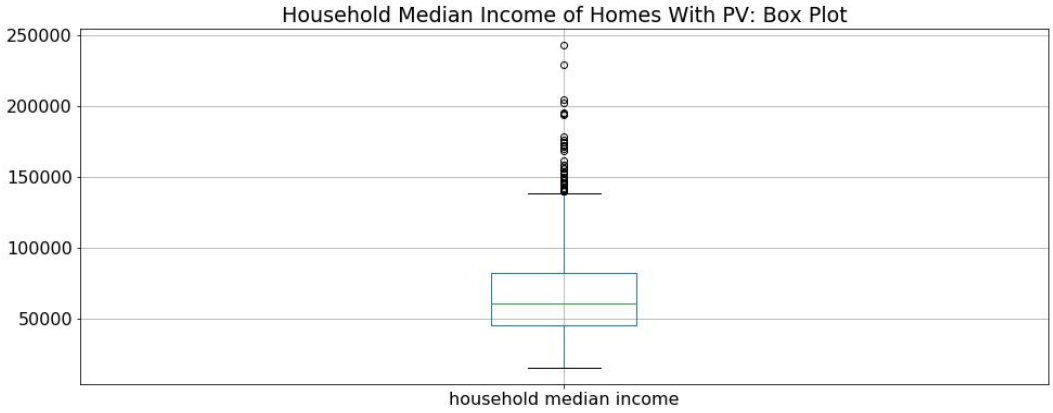


Figure A4: Box plot of median income of homes with solar installation.

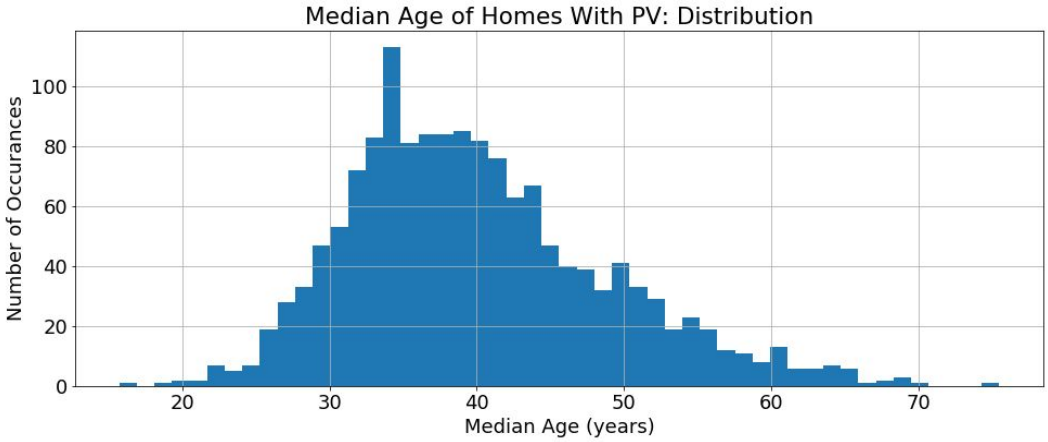


Figure A5: Median age of homes in certain zip codes that have installed solar.

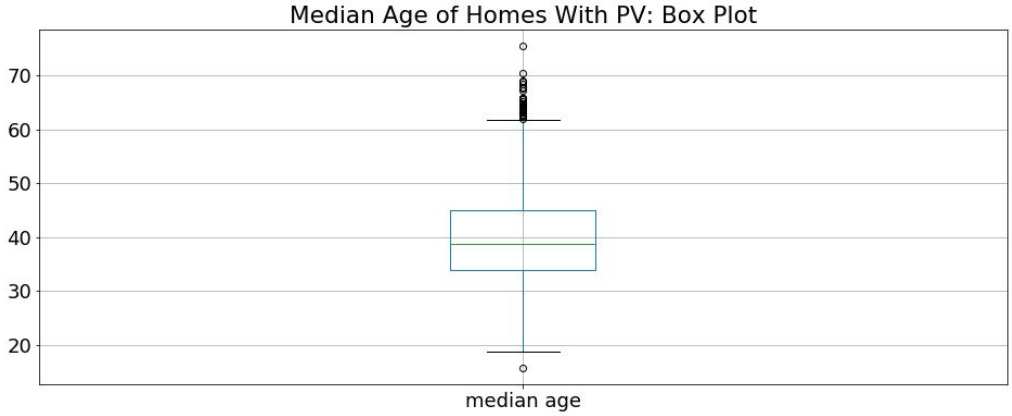


Figure A6: Box plot of median age of homes in certain zip codes that have installed solar.

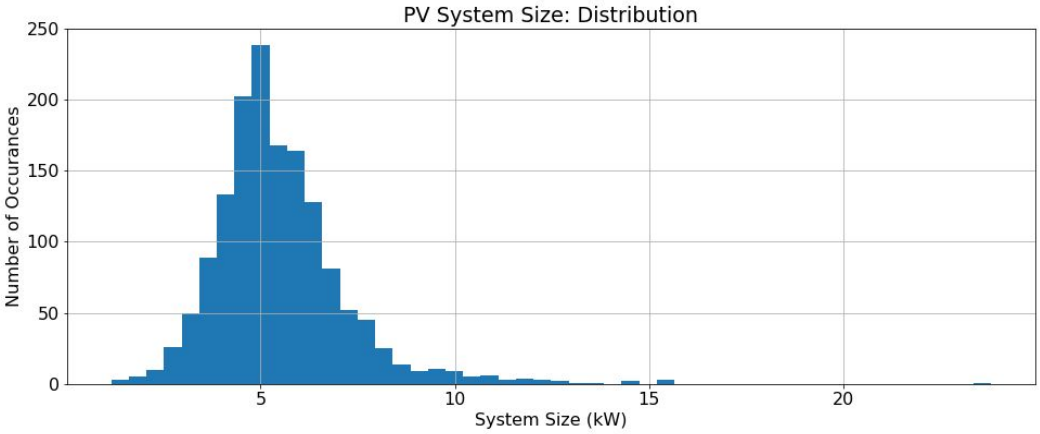


Figure A7: Plot of system sizes shows average around 4.6 kW.

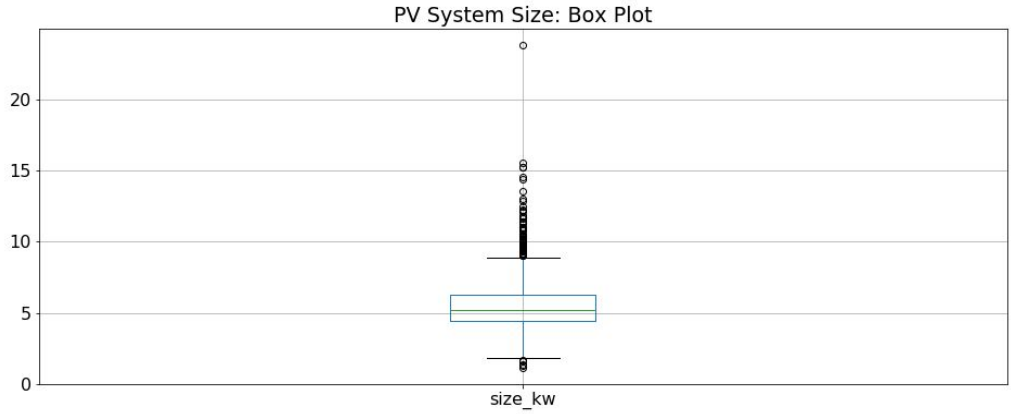


Figure A8: Box plot of PV system size.

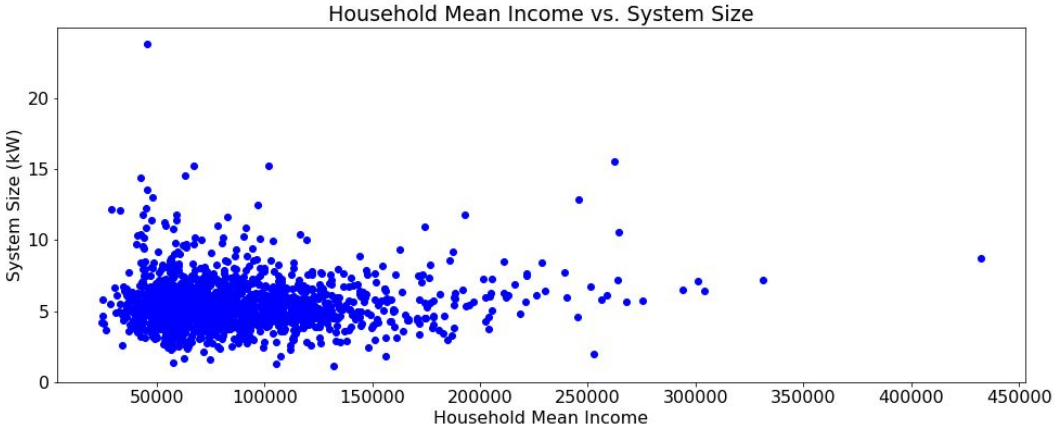


Figure A9: Mean income vs system size scatter plot.

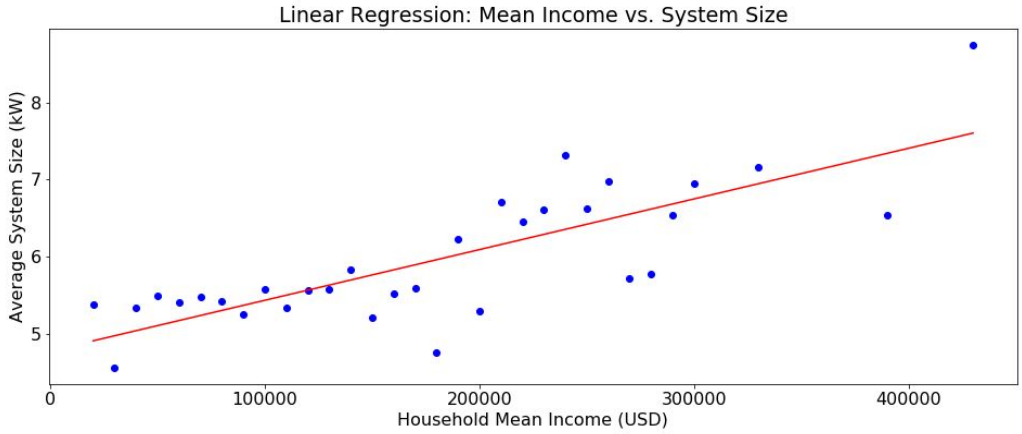
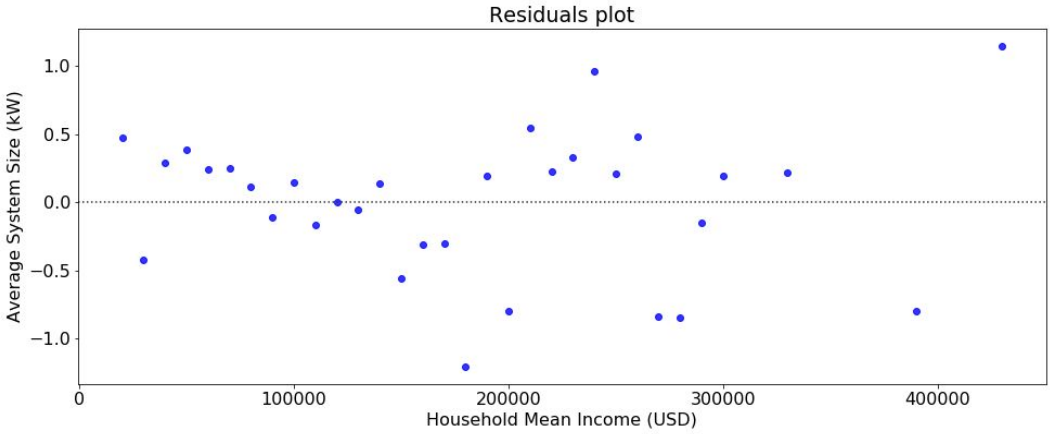


Figure A10: Mean income vs system size linear regression.



OLS Regression Results

Dep. Variable:	y	R-squared:	0.633
Model:	OLS	Adj. R-squared:	0.621
Method:	Least Squares	F-statistic:	51.81
Date:	Fri, 01 Jun 2018	Prob (F-statistic):	5.20e-08
Time:	09:06:21	Log-Likelihood:	-24.441
No. Observations:	32	AIC:	52.88
Df Residuals:	30	BIC:	55.81
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.7742	0.191	25.023	0.000	4.385	5.164
x	6.586e-06	9.15e-07	7.198	0.000	4.72e-06	8.45e-06

Omnibus:	0.810	Durbin-Watson:	1.867
Prob(Omnibus):	0.667	Jarque-Bera (JB):	0.439
Skew:	-0.287	Prob(JB):	0.803
Kurtosis:	2.984	Cond. No.	4.20e+05

Figure A11: Residual plot and results of average system size and household mean income.

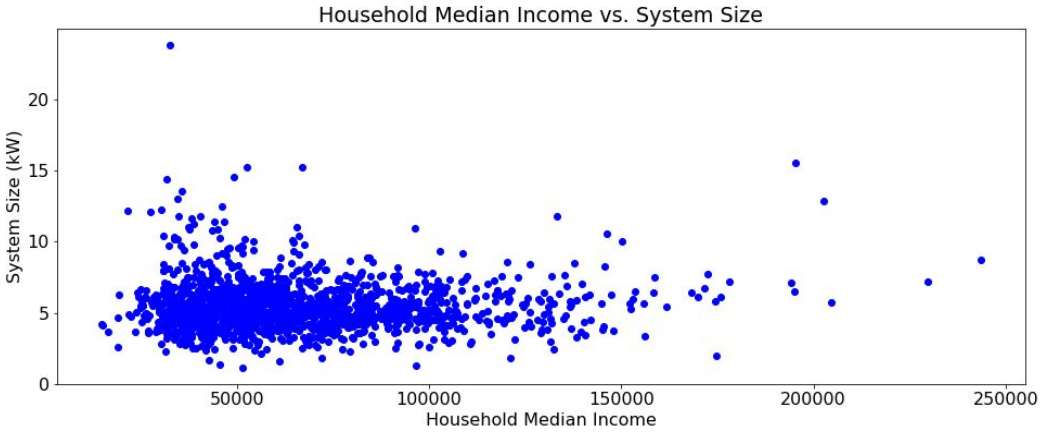


Figure A12: Scatter plot of median income vs system size.

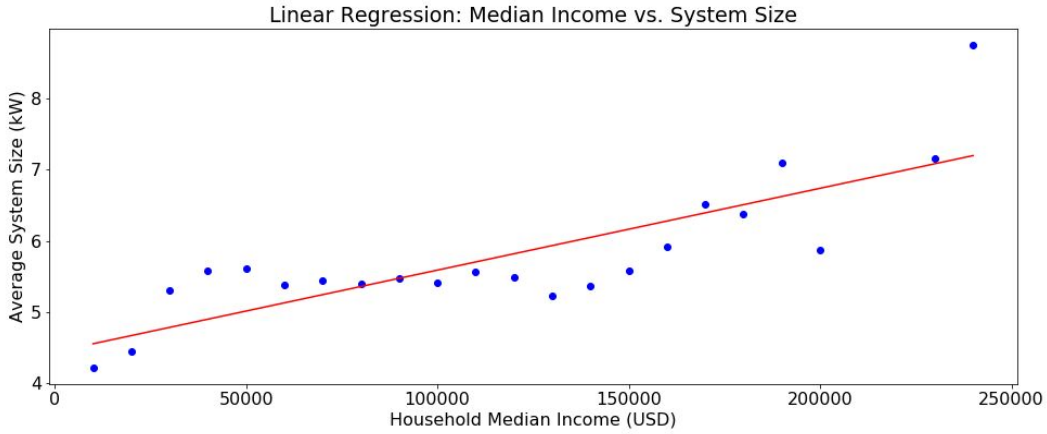
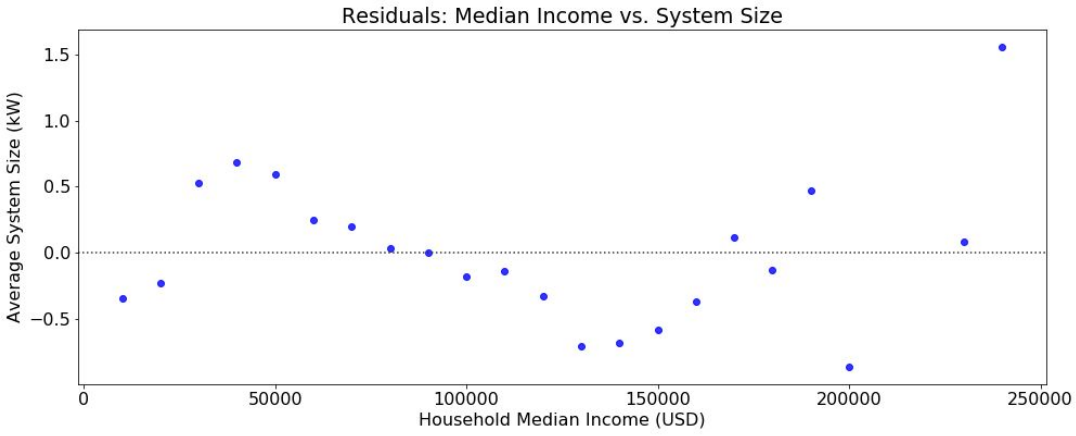


Figure A13: Linear regression of median income vs system size.



OLS Regression Results

Dep. Variable:	g	R-squared:	0.668
Model:	OLS	Adj. R-squared:	0.652
Method:	Least Squares	F-statistic:	40.27
Date:	Fri, 01 Jun 2018	Prob (F-statistic):	3.42e-06
Time:	09:06:47	Log-Likelihood:	-17.577
No. Observations:	22	AIC:	39.15
Df Residuals:	20	BIC:	41.34
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.4434	0.243	18.273	0.000	3.936	4.951
f	1.148e-05	1.81e-06	6.346	0.000	7.71e-06	1.53e-05

Omnibus:	6.019	Durbin-Watson:	1.023
Prob(Omnibus):	0.049	Jarque-Bera (JB):	3.744
Skew:	0.858	Prob(JB):	0.154
Kurtosis:	4.069	Cond. No.	2.72e+05

Figure A14: Residual plot and results of average system size and household median income.

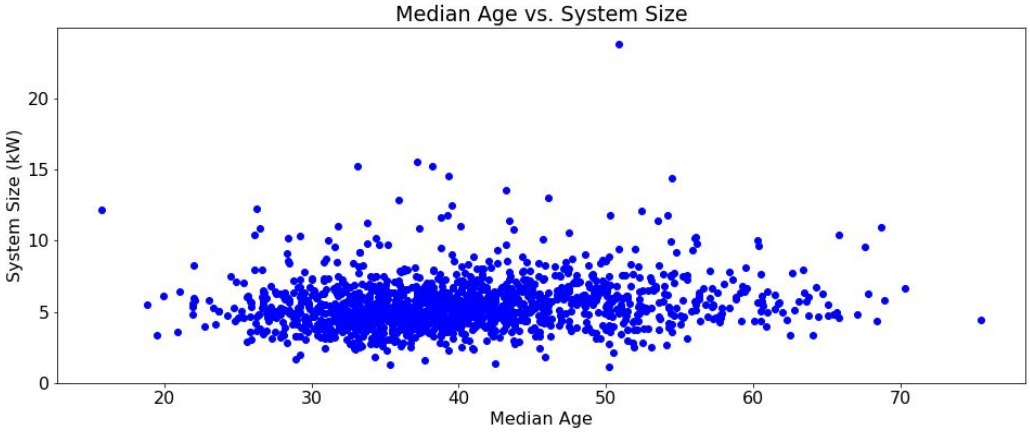


Figure A15: Median age vs system size scatter plot

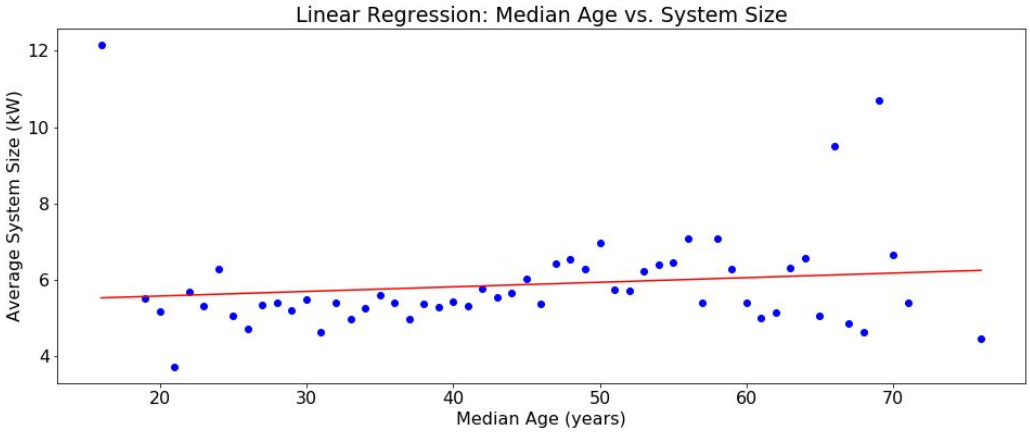


Figure A16: Median age vs system size linear regression line.

Table A1: Most beneficial counties and most marketable counties.

Most Beneficial (DNI only)	**Most Marketable (DNI and income)**
Amador	Amador
Butte	Butte
Calaveras	Calaveras
Contra Costa	Contra Costa
Glenn	
Humboldt	Humboldt
Imperial	Imperial
Lassen	Lassen
Los Angeles	Los Angeles
Mariposa	Mariposa
Mendocino	Mendocino
Mono	Mono
Orange	Orange
Plumas	Plumas
Sacramento	Sacramento
San Diego	San Diego
San Joaquin	San Joaquin
Santa Barbara	Santa Barbara
Sierra	
Solano	Solano
Stanislaus	Stanislaus
Tehama	Tehama
Tuolumne	Tuolumne
Ventura	Ventura