**Solar Relationship to Socioeconomic and Demographic Factors in the United States**

Atharva Fulay, Niki Tavakoli, Aamirah Dastgir

INF 550 Final Project Report

## 1. Introduction

In this project, our group focused on a sustainability issue, specifically within the realm of solar energy. We looked into whether there is a relationship between socioeconomic or demographic factors in the US population and the progress/incentivization of solar installations and production. Since California is home to liberal government policies, wealthy individuals and technologically advanced companies, we hypothesized that the solar production and installations in California would be higher than other parts of the nation. Additionally, we extended the hypothesis to Southwestern states (California, Arizona, Nevada, New Mexico, Utah and Colorado) because these states' governments also offer incentives for solar installations and are continuously pushing for further development.

## 2. Background

Solar powered electricity has been a popular method for generating green energy while allowing individuals and corporations to save exceptional amounts of money. California and the Southwestern states, in general, have been proactive about offering incentives to individual homeowners, apartment buildings, and commercial buildings to encourage solar installations and similar technologies. This is due to the sunny climate in these states which allows solar energy to be readily available and harnessed. In a similar vein, California is home to some of the most technologically advanced and socially aware companies. Furthermore, citizens of the San Francisco Bay Area and Los Angeles Metropolitan area, on average, are paid better compared to other areas in the nation.

However, one caveat to solar power is the high cost of the initial installation and fees. The aim of our analysis is to determine if a connection exists between socioeconomic or demographic factors and current production of solar as well as potential energy that could be

generated. We looked at multiple sources that provide insight into the association of income, demographics, and solar production.

## 3. Datasets

As part of our analysis, we collected and analyzed three data sources. These data sources are joined on United States Postal codes (ZIP codes). These data sources were downloaded as CSVs from their respective websites.

1. Google's Project Sunroof dataset contains data about how many buildings have solar installed and, more importantly, the potential of how much solar power can be generated in each ZIP code. Google calculated this amount based on an analysis using Google Maps and research about how much energy the average solar panel generates. This data source is the most limiting, containing about 11,000 usable rows.
2. The U.S. 2010 Census dataset provided data containing each ZIP code's demographics including age, sex, race, heritage, relationship, and housing information.
3. The Internal Revenue Service dataset provided data on adjusted gross income as well as totals earned by wages for every ZIP code.

## 4. Process and Methods

To begin our analysis, we cleaned the data and identified interesting attributes from each data source. We loaded the data from Project Sunroof into a MySQL database and looked at the values within each column. To reduce noise from small states or invalid amounts, we filtered out any ZIP codes that had a `count_qualified` of less than 25. This means that we analyzed only the ZIP codes which had 26 or more buildings that are suitable for solar, according to Google. We also visualized these in Tableau to see if there were any outliers.

Since the Census data had over 400 columns of data per ZIP Code, we split up the CSV into four tables, each were confined to household, race/heritage, relationships, and sex/age data, respectively. Lastly, the IRS data was added as a table to the database. Since the amount of data was so massive, it was difficult to immediately rule out any outliers. However, there were a handful of NULL records due to data that could potentially be too specific to ZIP codes. These

were discarded from our analysis. Finally, we created some proxy variables that could be used to determine current production from ZIP codes, percentages, and potential gain. Our fully joined tables contained 9,927 ZIP codes with up to 430 attributes for each.

Once we had the database set up, we used Jupyter notebooks to run a handful of preliminary analyses and Tableau for some basic visualizations. One of the most valuable visualizations was to figure out where the ZIP codes were on a map (Figure 1).

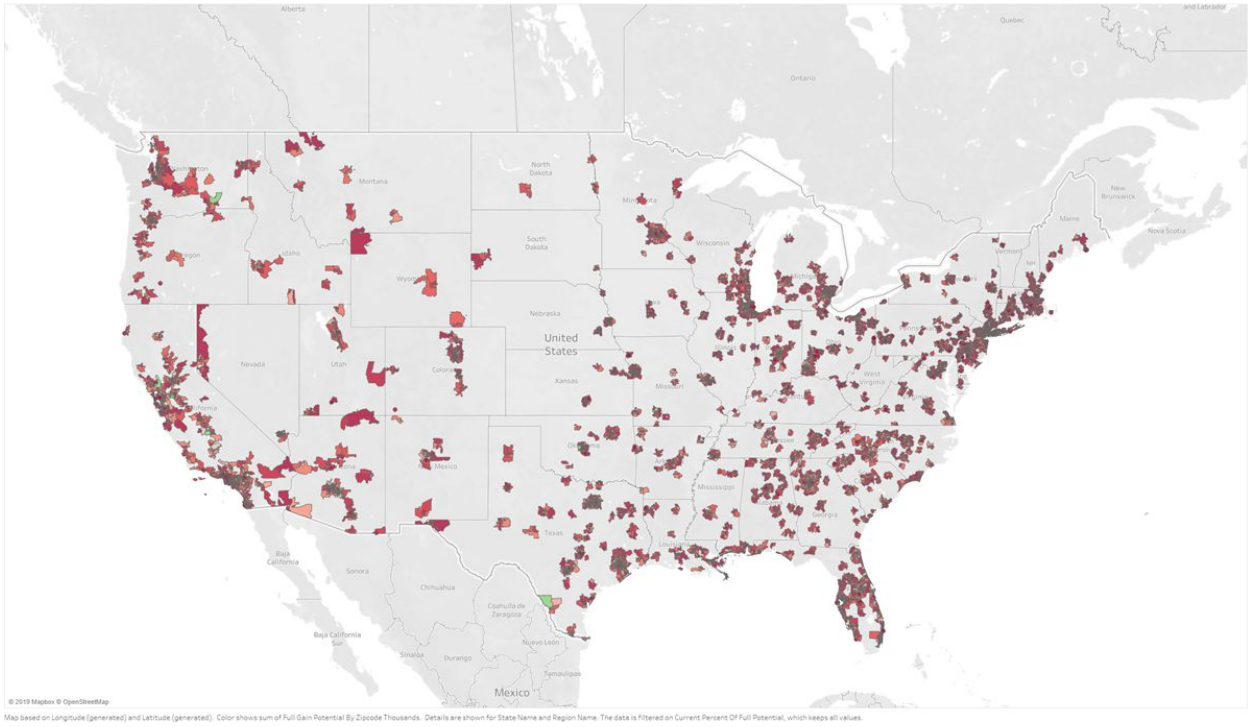



**Figure 1. ZIP codes in our analysis. Red to green (low to high) indicates potential solar energy that can be generated in that ZIP code.**

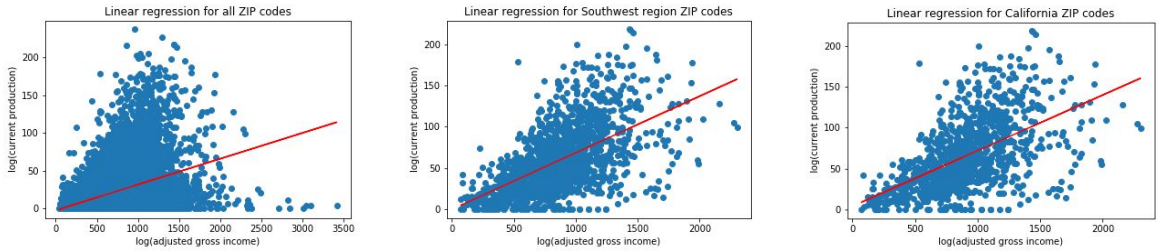
According to Figure 1, many of the ZIP codes are near urban areas rather than rural areas of the United States. The reason for why some ZIP codes are scattered across rural areas is not clear as it likely pertains to Google's study specifications.

Our next step was to see if there is a linear relationship between the average gross income and the production amount of solar in a certain ZIP code. Ideally, we would have liked to find a line that could represent the data well with a high $R^2$ value. We also used a correlation matrix to see if there were any non-trivial correlations that we could not see immediately. We used Pearson's Correlation method and chose the (standardized) strength of association as follows: 0.1 - 0.3 is weak, 0.3 - 0.7 is moderate, and 0.7 to 1 is strong. If the value is positive, there is an

uphill association and if the value is negative then there is a downhill association. Lastly, we utilized K-means clustering to determine if certain states' current percentage of full potential to actual amount of production in kilowatts could be similar to one another.

## 5. Analysis Results

Results from the linear regression analysis showed that there is little to no relation between the average adjusted gross income and the production of solar power in a ZIP code. Even with a logarithm transformation on the earnings and production, the result of `y = 0.034x - 2.227` had an $R^2$ value of 0.152, indicating this line cannot be trusted as a predictor with much confidence (Figure 2). This procedure was repeated for ZIP codes in Southwestern states as well as for California to see if it would perform better with a smaller sample (Figures 3 and 4). It resulted in equations `y = 0.069x - 0.058` and `y = 0.068x + 4.076` for $R^2$ values of 0.328 and .308, respectively. Although better, it still means that these linear regression predictions will not be useful in determining the production values based on that ZIP code's average adjusted gross income.




**Figures 2, 3, and 4 from left to right. Linear regression plots for all, Southwest, and California ZIP codes. They have $R^2$ values of 0.152, 0.328, and 0.308, respectively.**

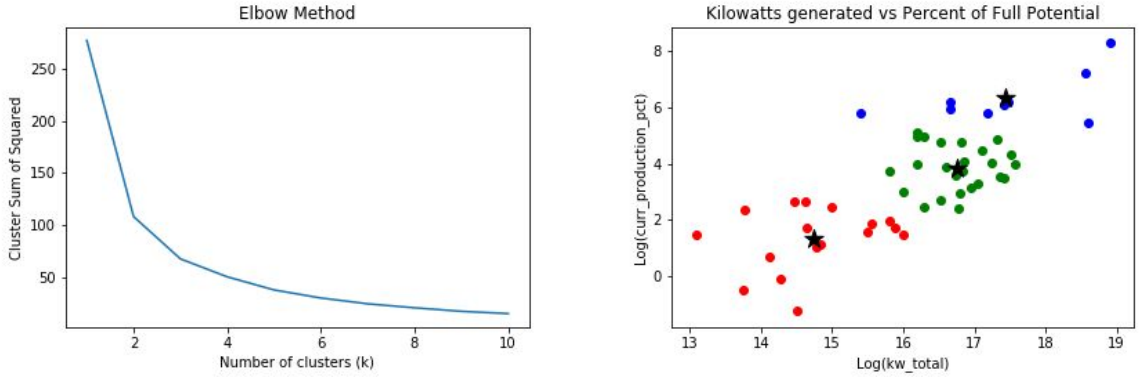
One noteworthy point is that the `log(current production)` averages for California, the Southwest region, and the nation were 65.29, 59.03, and 22.66, respectively. Even though the relationship may not be linear or fit well on a line, California and the Southwest region are producing more solar than the rest of the nation on average. In fact, all ZIP codes outside of the Southwest region produce a `log(current production)` of 15.68.

The correlation matrix and analysis was the most informative of where relationships in the data between the sets exist. One field we created as a proxy was the current production, set this to the formula: `curr_production = kw_total * existing_installs /`

count qualified. This can be used as an estimation for how much solar energy the ZIP code was producing at the time of Google's study. We looked at the top and bottom 25 fields for Pearson's correlation from each data source for any correlation to current production. Obviously existing_installs topped the list with 0.97, but the 2nd one was the most interesting: Number; RACE - Total population - Two or More Races - White; Asian [3] came in at 0.555. Although this is not enough to be a strong association, it is the 2nd strongest one that the correlation matrix produced, and is well within the bounds of a moderate positive association. A few further down the list lies Number; HISPANIC OR LATINO AND RACE - Total population - Hispanic or Latino - Asian alone, at 0.453 furthering the idea that the Asian (either mixed race with White or only Asian) population is living in ZIP codes that generate the most solar power.

To take it one step further, we looked into Number; RACE - Total population - Two or More Races - White; Asian [3] and any attributes it was associated with. After the expected Asian race-related fields, existing_installs_count popped up at a Pearson correlation value of 0.595. This confirms the notion that the Asian population lives in areas where solar production is producing the most power.

Lastly, we used K-means clustering on states to see if certain states were acting in similar ways. For this analysis, we created another field: curr_production_pct = existing_installs / count qualified. This gives us the percentage of homes and buildings that had solar installation at the time of Google's study. We compared that to the kw_total field to see if the states that are already producing solar power are producing a lot of power, or if it is minimal in terms of kilowatts generated. To determine the number of K's, we used the Elbow method shown below.




**Figures 5 and 6 from left to right. Figure 5 shows the sum of squared distance with relation to a growing K value. Figure 6 shows the clusters when k=3.**

We ended up choosing k=3 with the result above (Figure 6). What the k-means clustering shows is that, more or less, the states that are already producing more solar are indeed the states that have the most potential in terms of total kilowattage. Although it does not inform us of anything related to the population, it shows that the states offering incentives and encouraging their citizens to invest in solar are the states that should be doing so. Since they have the most to gain from solar energy (in terms of kilowatts), they are right to take the most advantage. Among the states in blue (top right) are California, Texas, Florida, Arizona, Colorado and Hawaii.

```
kmeans = KMeans(n_clusters=3, init='k-means++', max_iter=300, n_init=10, random_state=0)
pred_y = kmeans.fit_predict(X)

for i in range(len(X)):
    if pred_y[i] == 0:
        col = "blue"
    elif pred_y[i] == 1:
        col = "red"
    elif pred_y[i] == 2:
        col = "green"
    plt.scatter(X[i][0], X[i][1], c=col)

plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], marker="*", s=200, c='black')
```

**A snippet of code to generate and color the K-means plot (Figure 6). X is an array of 2-D tuples containing log(kw_total) and log(curr_production_pct) as x, y coordinates.**

## 6. Observation and conclusion

From the linear regression analysis, there is no clear linear relationship between solar production and the average adjusted gross income of the residents of that ZIP code. This was the case for ZIP codes nationwide, within the region of the Southwest states, and only in California. However, the `log(current production)` average values reveal that California and the Southwest region are currently producing more solar energy than the rest of the nation.

From the correlation matrix analysis, an interesting moderately-positive correlation appears. People of Asian descent (either mix of White/Asian or Asian alone) live in ZIP codes that produce more solar power. There is no evidence that Asian households or building owners are creating these installations but it could solely be due to the location where Asians are residing: high cost of living areas like the San Francisco Bay Area or the Los Angeles metropolitan area.

Lastly, the analysis from the clustering provided evidence that the states which are providing incentives and encouraging its residents to invest in solar are the states that have the most potential kilowattage to gain from solar power. States like California, Texas, Florida, Arizona, Colorado and Hawaii are leading in current percent of full solar potential while having the highest kilowattage potential.

## 7. References


1. Zaric, Drazen. "Better Heatmaps and Correlation Matrix Plots in Python." Medium, Towards Data Science, 15 Apr. 2019, towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-41445d0f2bec.
2. Keen, Ben Alex. "K-Means Clustering in Python." *Ben Alex Keen*, 10 May 2017, benalexkeen.com/k-means-clustering-in-python/.
3. Chauhan, Nagesh Singh. "A Beginner's Guide to Linear Regression in Python with Scikit-Learn." *Medium*, Towards Data Science, 7 Sept. 2019, towardsdatascience.com/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-83a8f7ae2b4f.
4. Bronshtein, Adi. "Simple and Multiple Linear Regression in Python." *Medium*, Towards Data Science, 2 Aug. 2018, towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9.
5. Blais, Henry. "K-Means Practical." *Medium*, Towards Data Science, 7 Mar. 2019, towardsdatascience.com/k-means-practical-1ab126e52f58.

6. Maklin, Cory. “K-Means Clustering Python Example.” *Medium*, Towards Data Science, 21 July 2019, towardsdatascience.com/machine-learning-algorithms-part-9-k-means-example-in-python-f2ad05ed5203.