Solar Energy Analysis

# Introduction

# The U.S. solar industry continued on its record-breaking trajectory in Q2 2015 with 1,393 megawatts (MW) of installed solar capacity, making this the largest Q2 in history. As has been the case over the last 18 months, the residential and utility-scale markets led the way, installed 463 and 729 MW, respectively.

Roughly 20,000 MW of solar capacity is forecasted to come online over the next two years, doubling the country's existing solar capacity. Growth is expected to be broad-based, with more than 16 states expected to top the 100 MW mark in 2016, up from 9 states in 2014.

The purpose of this analysis is to dig deeper into the acceptance (sentiment analysis) of consumers and their likelihood to have solar installations in their homes, also find the best places zip codes to have solar panels installed, concentrating in New York and using California as a benchmark

# Literature Review

Mainly used the literature and data sets found in <https://www.californiasolarstatistics.ca.gov/>, <http://gosolarcalifornia.org/> and <http://www.seia.org/>.

There are now over 22,700 MW of cumulative solar electric capacity operating in the U.S., enough to power more than 4.6 million average American homes.

With over 135,000 installations in the first half of 2015, nearly 784,000 U.S. homes and businesses have now gone solar and a new solar project was installed every 2 minutes

New York has a considerable amount of solar potential and has consistently been in the top 10 US solar markets. Pursuant to Governor Cuomo’s stated commitment to the NY-SUN initiative, the New York Public Service Commission (PSC) recently announced a 10-year $1 billion commitment to developing a self-sustaining solar market in New York State

I also did some research on Sentiment analysis using R from the following pages.

<http://thinktostart.com/sentiment-analysis-on-twitter/> and <http://thinktostart.com/wordcloud-syria-on-twitter/>

There are also data sets of current solar installation in California and New York that will be used as part of the analysis to determine the penetration of solar installation at the zip code level by cross referencing the US census data for population.

# Dataset

Downloaded twitter data for solar energy and green energy keywords, to which sentiment analysis will be performed.

<https://dev.twitter.com/overview/api/tweets>

My initial analysis has led me to use the following attributes from the tweets collected.

There were 8000 tweets collected using Keyword: SolarEnergy and 5000 tweets collected using GreenEnergy

* user.id
* created\_at
* time\_zone
* id
* text
* country
* location
* place
* coordinates

I am also using solar installation data for New York from

<https://data.ny.gov/Energy-Environment/Statewide-200kW-or-Less-Residential-Non-Residentia/3x8r-34rs>?

Fields extracted:

* City
* County
* State
* Zip Code
* Sector
* Electric Utility
* Purchase Type
* Date Application Received
* Date Install
* Project Cost
* $Incentive
* Total Nameplate KW
* Expected KWh Annual Production
* Location 1

I am also using solar installation data for California from

<https://www.californiasolarstatistics.ca.gov/data_downloads/>

Fields extracted:

* Program Administrator
* Program
* Total Cost
* Current Incentive Application Status
* Host Customer Sector
* Host Customer Physical Address City
* Host Customer Physical Address County
* Host Customer Physical Address State
* Host Customer Physical Zip Code
* Solar Contractor Company Name

I am also using the 2010 US Census Population By Zip Code from

<http://blog.splitwise.com/2013/09/18/the-2010-us-census-population-by-zip-code-totally-free/>

Fields extracted

* Zip Code ZCTA
* 2010 Census Population

# Approach

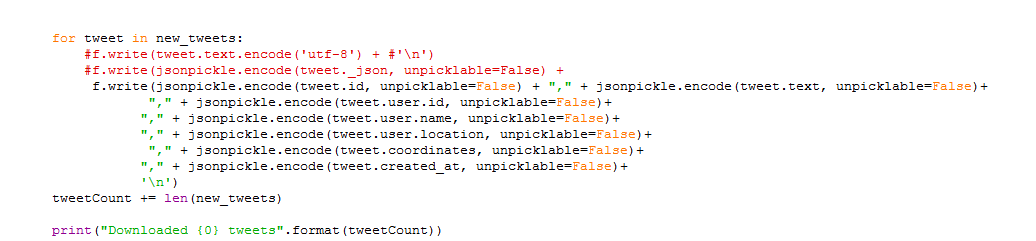
Once this is done, explain each of the steps in detail. What are you planning to do in each step or have already done? For example, in the above case you would create subheadings for each of the steps.

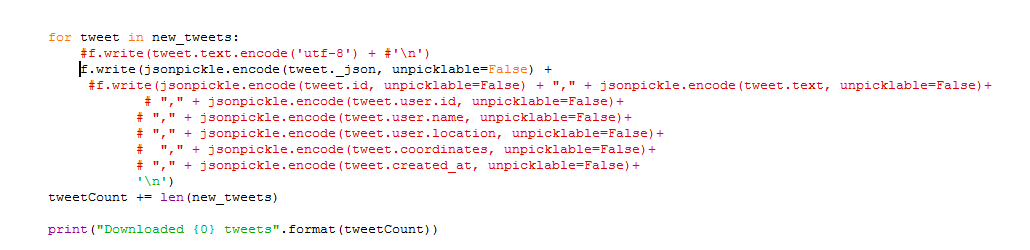
## Step 1: Data Research

## Scrapped Twitter: used two different approaches

<https://github.com/jcasallas/Capstone/blob/master/Python%20API%20Twitter>

1. Modified a python twitter API code (provided by a fellow student) to only extract desired fields.



1. Used unmodified python twitter API code (provided by a fellow student) to extract the entire JSON data from the API.
2. 

## Downloaded Installation Data for California and New York

## Zip Code population data

## Step 2: Data Clean up

Find useful data fields for the purpose of this analysis

## Step 3: Uploaded data to Hive

Uploaded data to Hive

<https://github.com/jcasallas/Capstone/blob/master/Hive%20table%20creation>

## Step 4: Zip Code Installation Analysis

This step is in progress – The goal is to complete analysis on how many installation are per zip code and correlate it to the population per zip. We will find out the penetration percentage for each zip and will focus our efforts on the zips with the highest penetrations

## Step 5: Tweeter Analysis

Tweets will also be grouped by location and coordinates (Where available) and sentiment analysis and world clouds preform to visualize the impact

## Step 4: Tweet data and Zip Code data Correlation

Find a correlation and overlay sentiment analysis by location on Solar installation to determine based on this what would be the best possible zip code to expand solar installation to.