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 report is to analyze several variables that may lead into predicting what areas or consumers have a higher probability to have solar installations in their homes, also find the best places zip codes to have solar panels installed, concentrating in New York, once the best predictive model is achieved the model can be applied to other areas.

# Literature Review

There are currently no research papers on what areas, zip codes or customers are more favorable to or have a higher probability to enter into the solar industry by having solar installations on their roofs based on demographic data.

However, there were are few articles/papers showing how Solar power will continue its growth in the years to come and the areas that areas that are more favorable for large buildings (commercial) applications.

The following is a quote by eminent inventor and futurist Ray Kurzweil during the latest interview

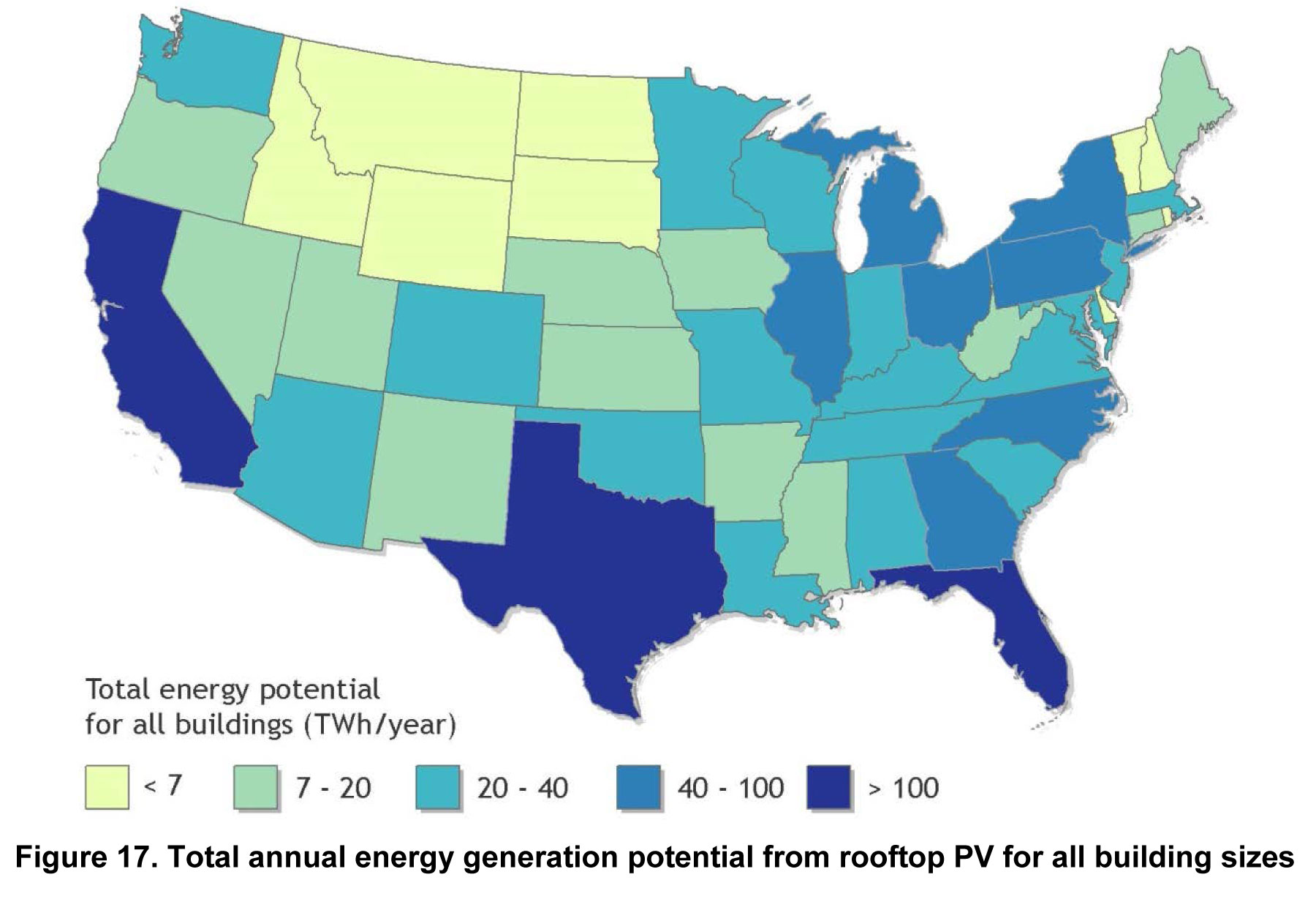
"Solar panels are coming down dramatically in cost per watt. And as a result of that, the total amount of solar energy is growing, not linearly, but exponentially. It’s doubling every 2 years and has been for 20 years. And again, it’s a very smooth curve. There’s all these arguments, subsidies and political battles and companies going bankrupt, they’re raising billions of dollars, but behind all that chaos is this very smooth progression."

Solar cell efficiency is improving rapidly with technological progress and as solar costs decline, residential system sales increase exponentially. Therefore creating a model that can be used by solar companies to better understand and target customers should be in the top priority.

If residential solar adoption is like air conditioning, where 50 years later they had 80 percent saturation, solar on buildings could follow a similar path.

Some states with below-average solar resource (such as Minnesota, Maine, New York, and South Dakota) have similar or even greater potential to offset total sales than states with higher-quality resource (such as Arizona and Texas). This highlights the observation that solar resource is only one of several factors that determine the offset potential.

Many New England states, despite their lower solar resources, could generate nearly half of electricity used from rooftop solar.



Source: GTM Research

# Dataset

There are 2 different data sets used in the analysis as described below.

## Dataset 1: Tweets

Tweeter Data with 27000 individual tweets, data was collected randomly over a period of 60 days, with search words: GreenEnergy, SolarEnergy and SolarPanels.

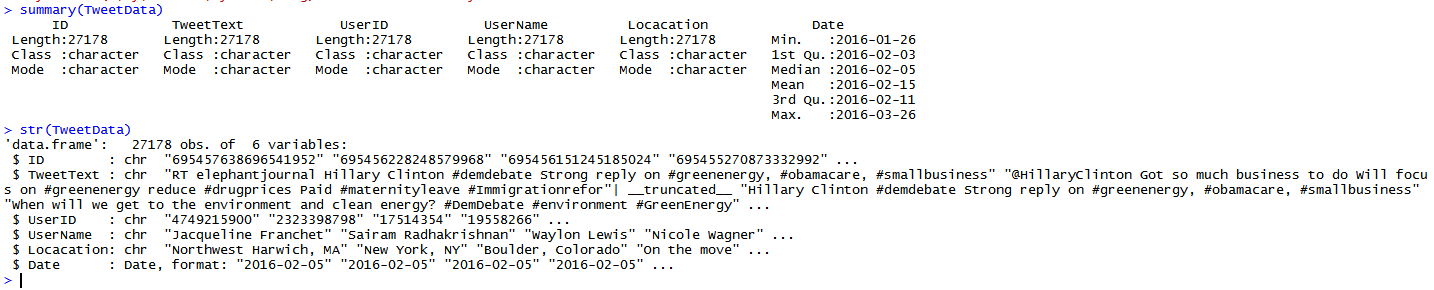
This data set will be used to provide some insight into what people are saying and “feeling” about Solar Power.

Python Script used to download the dataset can be found here:

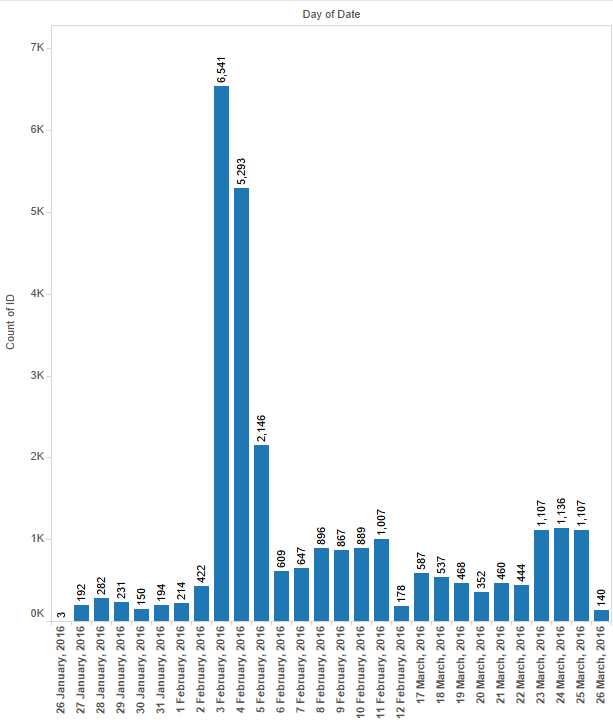
<https://github.com/jcasallas/Capstone/tree/master/Code>

File: Twitter\_SB Github

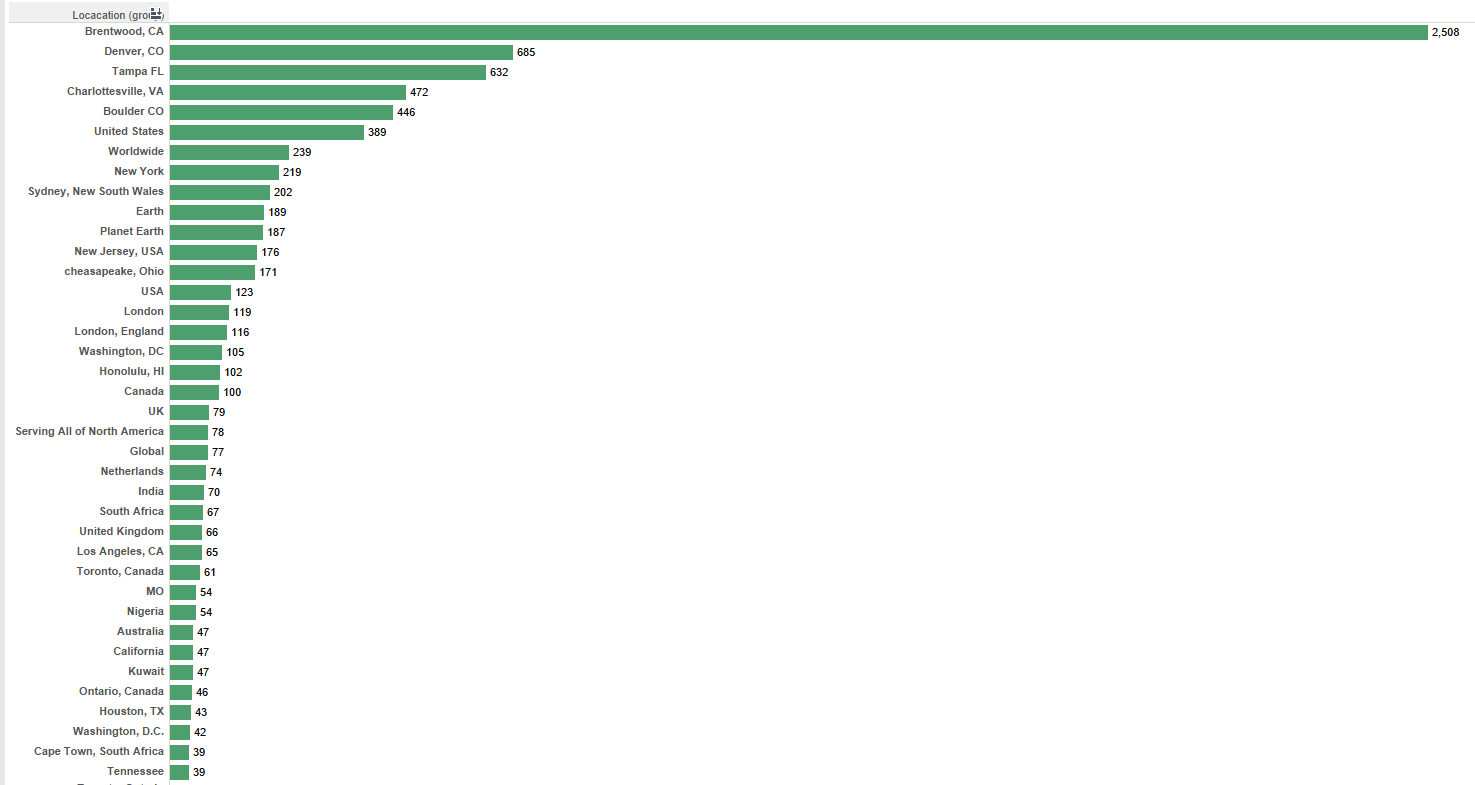
This data contains 7 attributes



Number of tweets recorded per day



Top Locations



Lists with the positive and the negative words.

We can find them here:

<https://github.com/mjhea0/twitter-sentiment-analysis/tree/master/wordbanks>

## Dataset 2: Census Data and Demographics

There are several data sets used for the location and demographic analysis, below is a breakdown of the data sets:

New York Solar Installation Zip Level

About the data: Statewide 200kW or Less Residential/Non-Residential Solar Photovoltaic Incentive Program dataset includes the following data points for projects completed in the incentive Program beginning December 2000: Project number, location, sector, application received date, installation date, electric utility, purchase type, inverter manufacturer, inverter quantity, PV module manufacturer, PV quantity, project cost, incentives, kilowatt capacity, and expected annual kilowatt hour production.

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

Income Data

Zip code file of two of the most commonly requested characteristics: median household income and mean household income:

Source: <http://www.psc.isr.umich.edu/dis/census/Features/tract2zip/>

Population

The 2010 US Census Population By Zip Code

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

Household Size

AVERAGE HOUSEHOLD SIZE OF OCCUPIED HOUSING UNITS BY TENURE - Universe: Occupied housing units

Source: <http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_14_5YR_B25010&prodType=table>

Search and dataset found here:

<https://github.com/jcasallas/Capstone/tree/master/DataSets/Census%20Data%20Sets/American%20Fact%20Finder%20Saved%20Searches>

<https://github.com/jcasallas/Capstone/blob/master/DataSets/Census%20Data%20Sets/ACS_14_5YR_B25010_Average%20HoseholdSize.xls>

Structure Type and Ownership

TENURE BY UNITS IN STRUCTURE - Universe: Occupied housing units

Source: <http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_14_5YR_B25032&prodType=table>

Search and dataset found here:

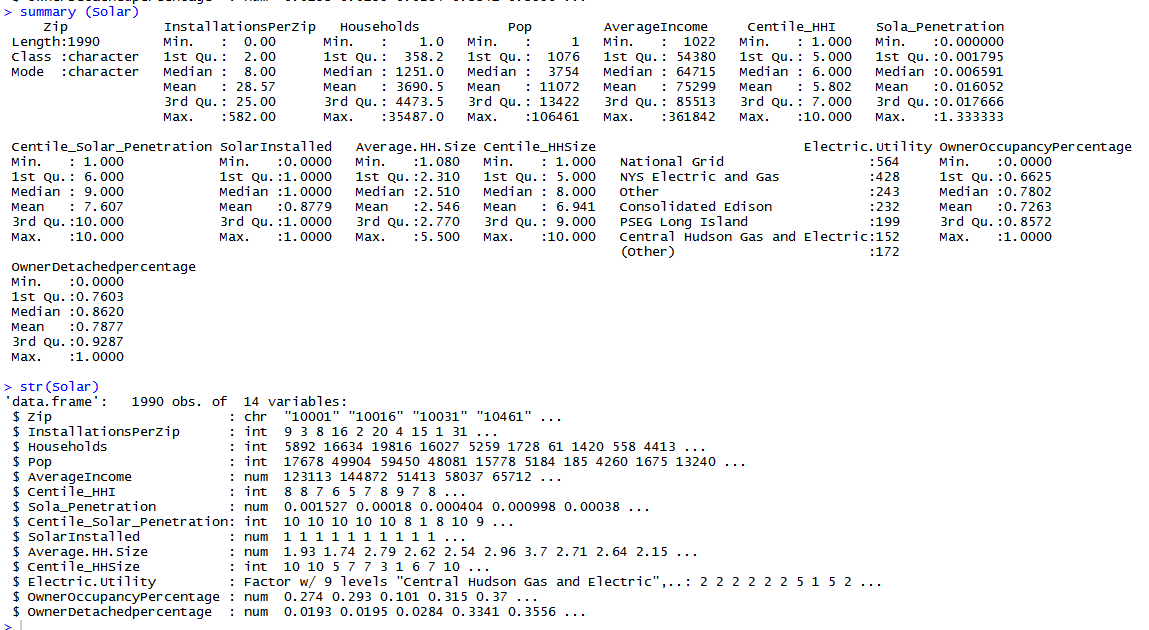
<https://github.com/jcasallas/Capstone/tree/master/DataSets/Census%20Data%20Sets/American%20Fact%20Finder%20Saved%20Searches>

<https://github.com/jcasallas/Capstone/blob/master/DataSets/Census%20Data%20Sets/ACS_14_5YR_B25032_Structure%20Type%20and%20Ownership.xls>

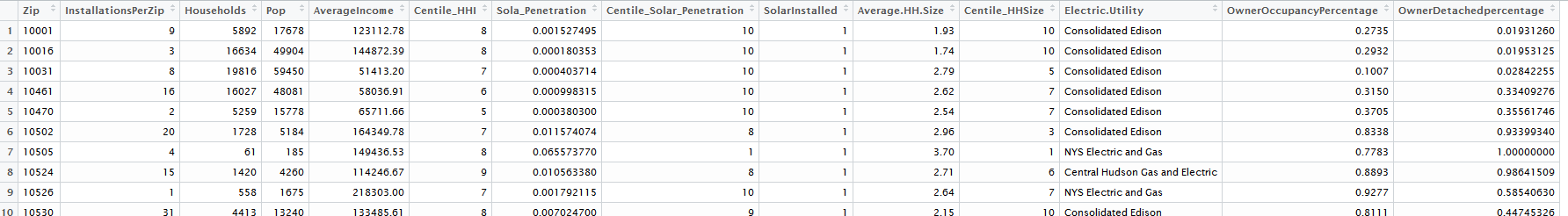
All the data sets above where modified to provide Zip level demographic information. Although there were margin of errors in the American Finder data sets, they were not included in the final results.

The final table compiled can be found here

<https://github.com/jcasallas/Capstone/blob/master/DataSets/CapstoneFileSolar_Final_20160329_V2.csv>



*Snapshot of the data set*



# Approach

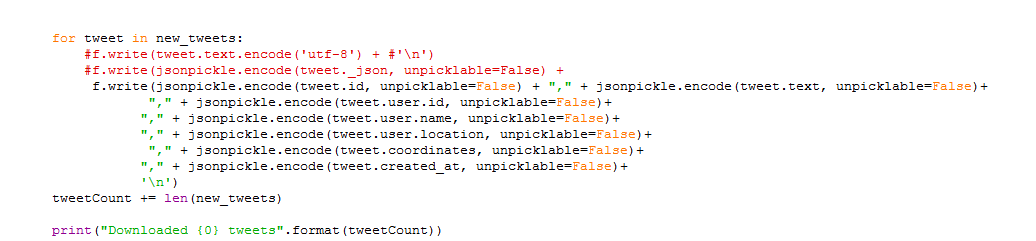
This research was done in two part which would be combined to gather a better understanding of “acceptance” of solar power and which areas have a higher probability to purchase solar panel installations.

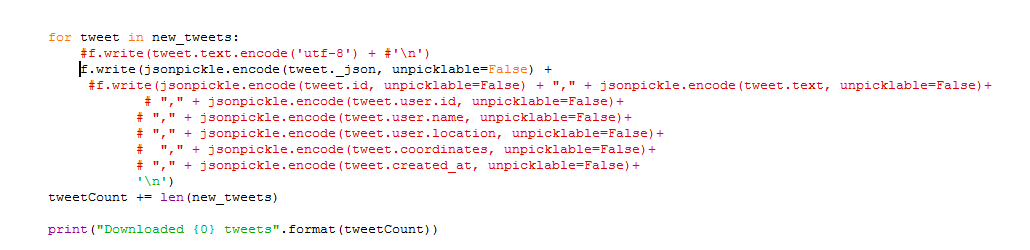
## Part 1 – Acceptance research

## Step 1: Data Research

## Scrapped Twitter: Used two different approaches

<https://github.com/jcasallas/Capstone/blob/master/Code/Twitter_SB%20for%20Git.py>

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



After experimenting with data loading and manipulation it was decided that option (a) was a more suitable option for the goal in mind, downloading and saving only relevant fields.

## Step 2: Data Clean up

After a few runs of the Twitter API, there were 5 files with data results for individual search words.

The files were compiled into one using excel.

The final file can be found here:

<https://github.com/jcasallas/Capstone/blob/master/DataSets/Tweet%20DataSets/JCDataGreenEnergy_All.csv>

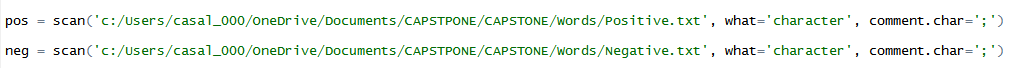
## Step 3: Positive and Negative Keywords

Lists with the positive and the negative words. We can find them here:

<https://github.com/mjhea0/twitter-sentiment-analysis/tree/master/wordbanks>

After downloading the ZIP files they were saved in a local folder

We now have to load the words in variables to use them, with the following code:



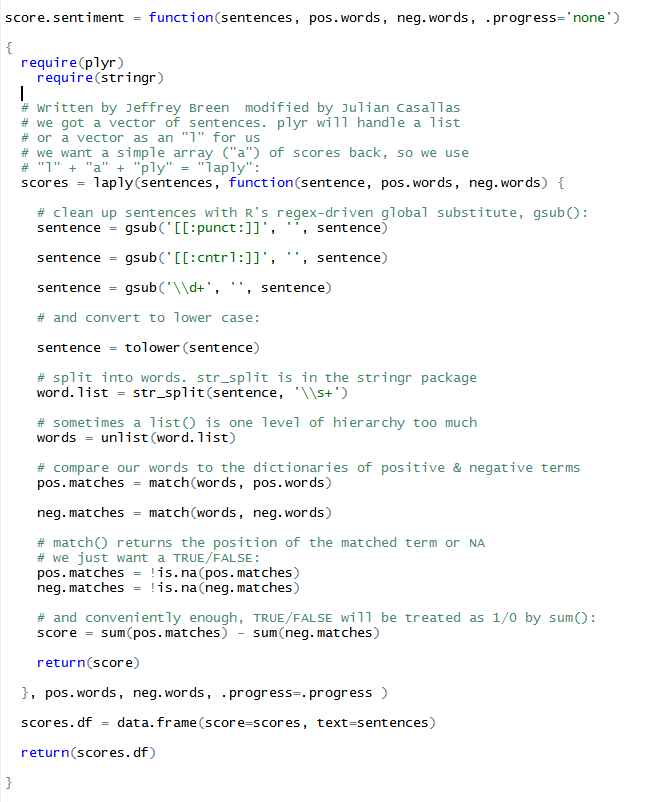
## Step 4: Performed extra clean up on tweets by removing special characters

Sometimes the tweeter text has invalid characters in, so we have to remove them.



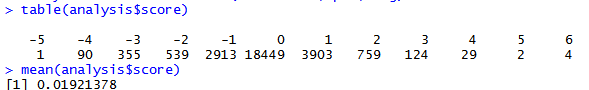
## Step 5: Tweet Analysis by words classification into positive and negative

Used an algorithm for analyzing our words



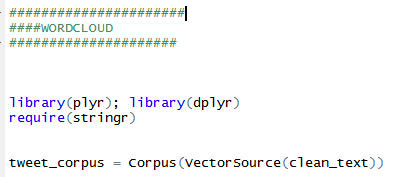
## Step 6: Sentiment Analysis completed

The positive values stand for positive tweets and the negative values for negative tweets. The mean tells you about the overall mood of your sample.



## Step 7: Added Clean text to a corpus

Add this clean text to a so called Corpus, this is the main structure in the tool tm to save collections of text documents.



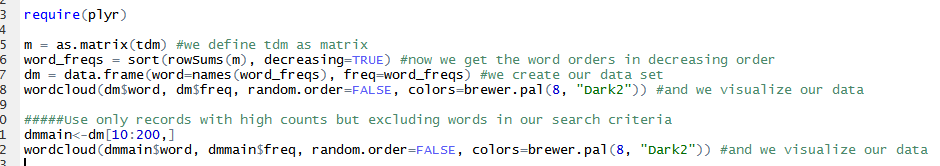
## Step 8: Removed Stop word and number and convert to lowercase

Transform this Corpus in a Term-document Matrix. This matrix describes the frequency of terms that occur in a collection of documents.



## Step 9: Arrange words by frequencies

Ok now we have our tdm. We have to do now is arrange our words by frequencies and put them in the word cloud.



## Part 2 – Predictive Models Research

## Step 1: Data Research

Data downloaded from several open data sites, including Factfinder and Open Data NY. Data was saved in Excel and CSV formats

## Step 2: Data Clean Up

Files cleaned up to only contain relevant data at the zip level.

Code can be found here:

<https://github.com/jcasallas/Capstone/blob/master/Code/CapstoneSQLCode.sql>

Data reduced to the following fields

* Zip
* InstallationsPerZip
* State
* Population /3 to be an estimated of HouseHolds
* Population
* AverageIncome
* Average House Hold Size
* OwnerOccupancyPercentage
* OwnerDetachedpercentage

## Step 3: SQL

SQL was used to perform data analysis and combining all cleaned up files into one major file

Analyzing and compiling number of installation per zip

Creating Installation counts by Zip

SELECT COUNT([Zip Code]) AS InstallationsPerZip, [Zip Code] INTO #InstallationsByZip FROM [JC\_NY\_SolarInstall\_Capstone] group by [Zip Code]

Data holding table with zip data and from Census data available at Factfinder.com

SELECT DISTINCT

CAP.Zip

, InstallationsPerZip

, US.State

, Cap.pop/3 AS HouseHolds

, Cap.pop

, CAp.Mean AS AverageIncome

, CASE WHEN [Average HH Size] <>'-' then [Average HH Size]

ELSE (SELECT AVG (CAST([Average HH Size] AS float)) FROM [JC\_Zip\_CapPopIncome] where [Average HH Size] <>'-')

END AS [Average HH Size]

, [OwnerOccupancyPercentage]

, [OwnerDetachedpercentage]

INTO JC\_NY\_InstallsPerZip\_CP

FROM [dbo].[JC\_Zip\_CapPopIncome] AS Cap

INNER JOIN [dbo].[JC\_CAP\_OwnerRenter\_Units] AS Own

ON Own.Zip = Cap.Zip

INNER JOIN [dbo].[JC\_ZipDataUS] AS US

ON US.Zip= CAP.Zip

LEFT JOIN #InstallationsByZip AS InstallsZ

ON InstallsZ.[Zip Code] =CAp.Zip

LEFT JOIN [dbo].[JC\_NY\_SolarInstall\_Capstone] AS Installs

ON Installs.[Zip Code]= CAp.Zip

WHERE US.state='NY'

Adding Solar Installation Data from NY open data site

SELECT

SolarIns.Zip

, CAP.City

, CAP.County

, CAP.State

, CAP.Sector

, CAP.[Electric Utility]

, CAP.[Expected KWh Annual Production]

, CAST(SolarIns.InstallationsPerZip as int) AS InstallationsPerZip

, CAST (SolarIns.HouseHolds as int) AS HouseHolds

, SolarIns.pop

, SolarIns.AverageIncome

, CAST (SolarIns.[Average HH Size] as float) AS [Average HH Size]

, [OwnerOccupancyPercentage]

, [OwnerDetachedpercentage]

INTO #JC\_NY\_SolarCap\_2016

FROM JC\_NY\_InstallsPerZip\_CP AS SolarIns

LEFT JOIN [dbo].[JC\_NY\_SolarInstall\_Capstone]AS CAP

ON Cap.[Zip Code]= SolarIns.Zip

SELECT \* FROM #JC\_NY\_SolarCap\_2016

Dealing with low households and Missing data

UPDATE #JC\_NY\_SolarCap\_2016

SET HouseHolds = pop

WHERE HouseHolds <1

UPDATE #JC\_NY\_SolarCap\_2016

SET [Electric Utility]='Other'

WHERE [Electric Utility] IS NULL

UPDATE #JC\_NY\_SolarCap\_2016

SET InstallationsPerZip = 0

WHERE InstallationsPerZip IS NULL

Generating Final table to be used in Predictive Models

SELECT

Zip

, City

, County

, State

, Sector

, [Electric Utility]

, [Expected KWh Annual Production]

, CAST(InstallationsPerZip as int) AS InstallationsPerZip

, CAST (HouseHolds as int) AS HouseHolds

, CAST(CAST(InstallationsPerZip as float)/CAST (HouseHolds as float) AS FLOAT )as SolarPenetration

, pop

, AverageIncome

, CAST ([Average HH Size] as float) AS [Average HH Size]

, [OwnerOccupancyPercentage]

, [OwnerDetachedpercentage]

INTO JC\_NY\_SolarCap\_2016

FROM #JC\_NY\_SolarCap\_2016

/\*\*\*\*\*\*FinalTable\*\*\*\*\*\*\*/

SELECT DISTINCT

Zip

, InstallationsPerZip

, [Households]

, [Pop]

, [AverageIncome]

, NTILE(10) OVER(ORDER BY [AverageIncome] DESC) AS Centile\_HHI

, SolarPenetration AS Sola\_Penetration

, NTILE(10) OVER(ORDER BY SolarPenetration DESC) AS Centile\_Solar\_Penetration

, CASE WHEN InstallationsPerZip >0 THEN '1' ELSE '0' END AS SolarInstalled

, [Average HH Size]

, NTILE(10) OVER(ORDER BY [Average HH Size] DESC) AS Centile\_HHSize

, [Electric Utility]

, [OwnerOccupancyPercentage]

, [OwnerDetachedpercentage]

FROM JC\_NY\_SolarCap\_2016

## Step 4: Analysis the data sets and relationships

Code can be found here:

<https://github.com/jcasallas/Capstone/blob/master/Code/NY%20Solar%20R%20code_V2.R>

Here we examine a few of the most important fields believed to be highly correlated to Solar Installations

Figure 1: shows a higher number of installations are present when the percentage of detached homes is higher, this makes sense since majority of residential installation are for detached homes.

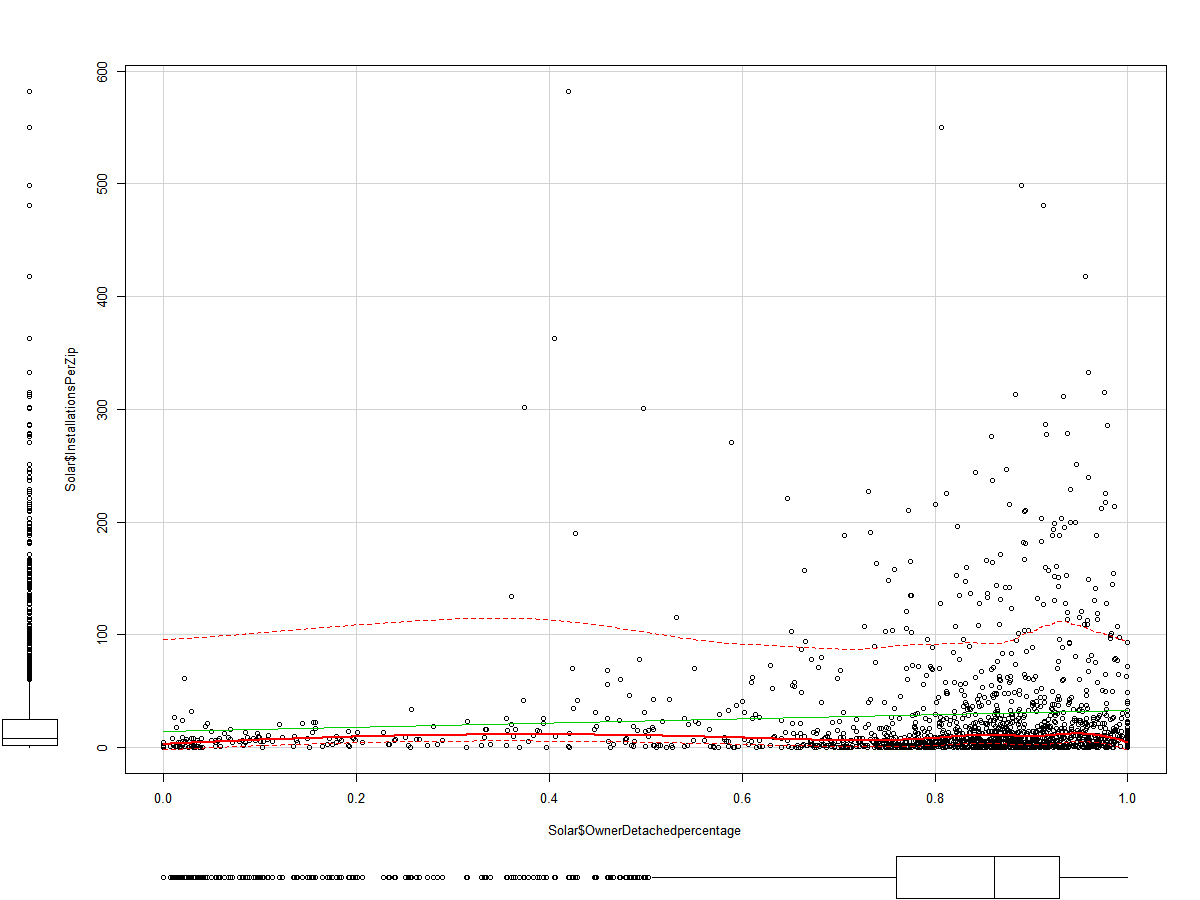


Figure 1 Installations vs DetachedPercentages

Figure 2: shows the relation between installations and Income, here we can see there are a high number of installations for customer with incomes around the $100,000, this shows that solar installations although have come down in price in recent years, customers still need a significant income, it also shows how customers with an income of $100,000 are more likely to purchase solar panels.

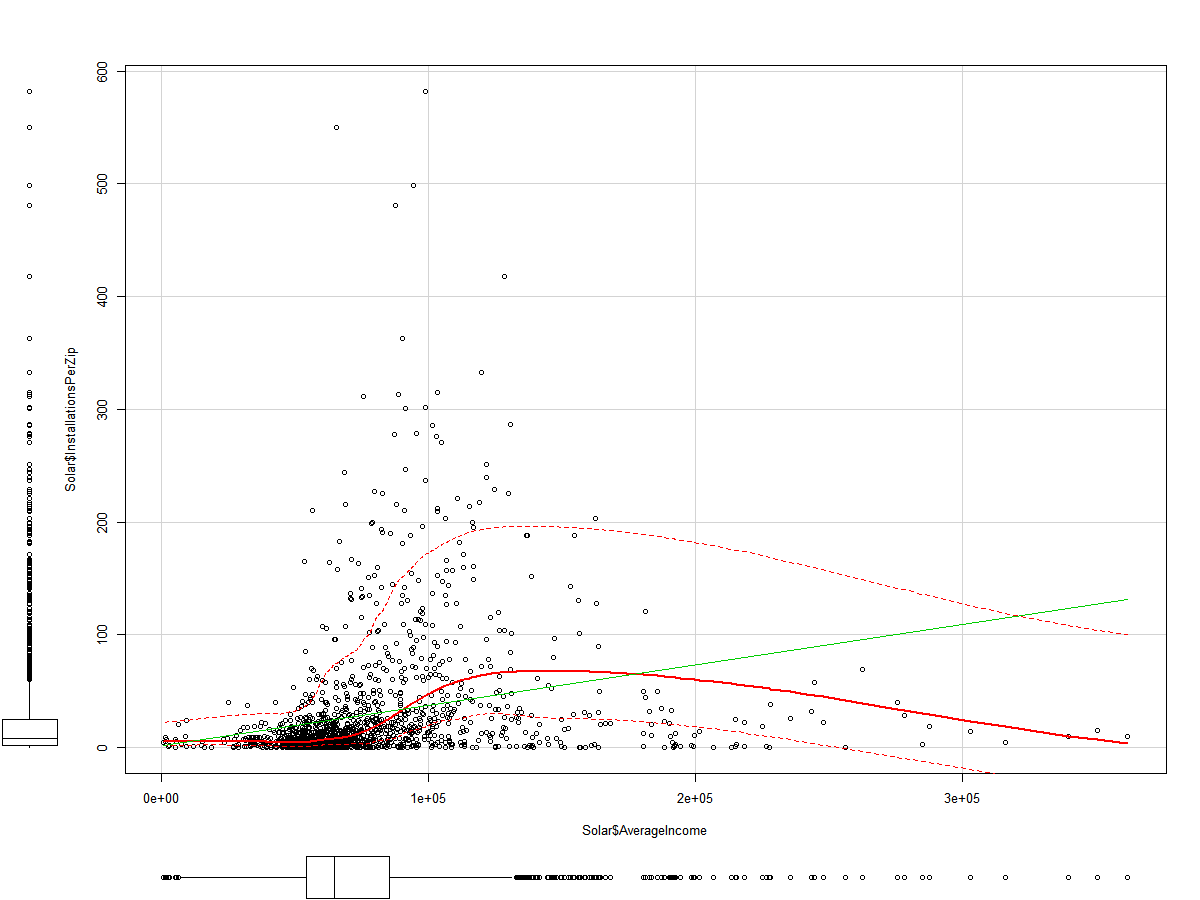
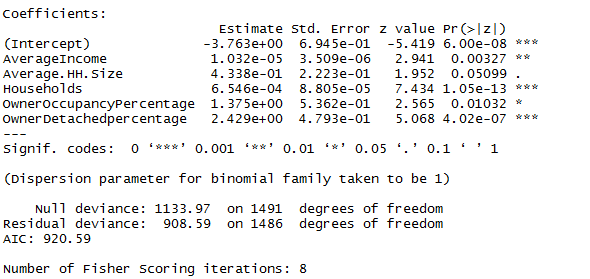


Figure 2 Installations vs Income

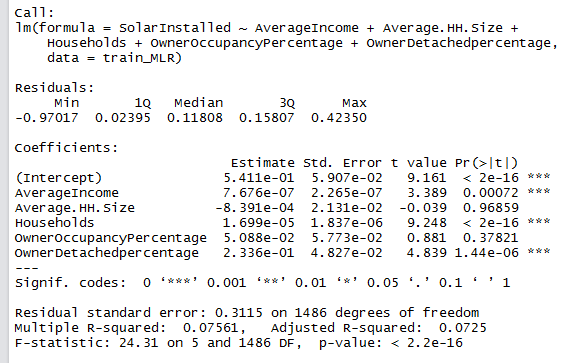
## Step 5: Predictive Models

Model 1 - logistic regression Model



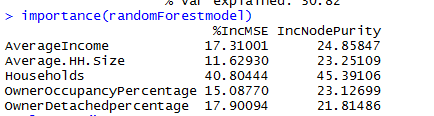
Now we can analyze the fitting and interpret what the model is telling us.  
First of all, we can see that Average.HH.Size,  and OwnerOccupancyPercentage are not statistically significant. As for the statistically significant variables, OwnerDetachedPercentage has the lowest p-value suggesting a strong association of the OwnerDetachedPercentage with the probability of having solar installation. The positive coefficient for this predictor suggests that all other variables being equal, owner of a detached home is more likely to have solar panels installations.

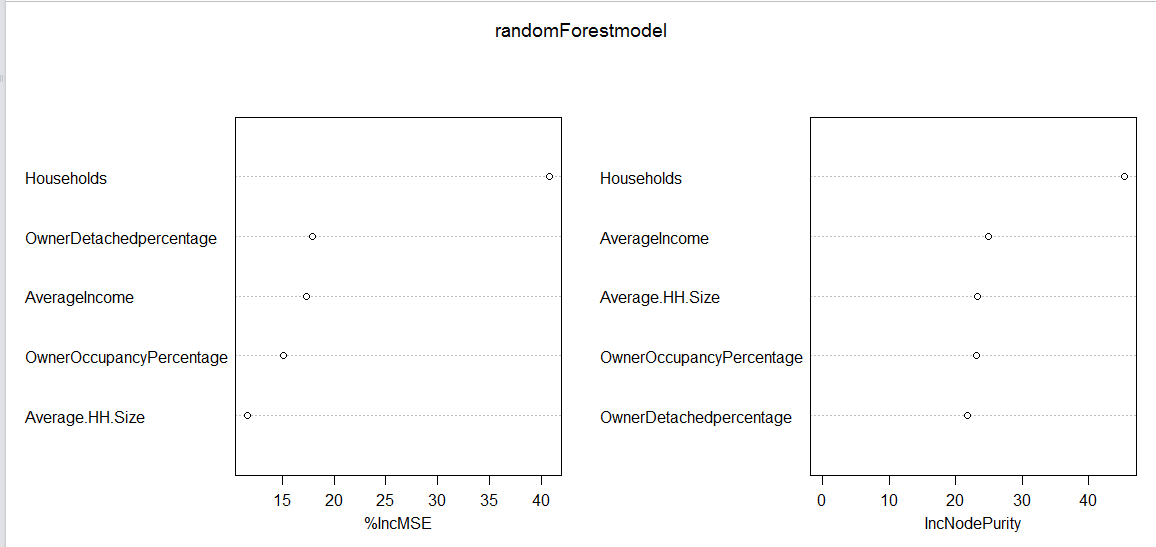
Model 2 Multivariate



OwnerDetachedPercentage has the lowest p-value suggesting a strong association of the OwnerDetachedPercentage with the probability of having solar installation.

Model 3 Random Forest Model

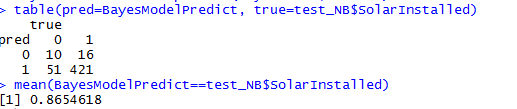




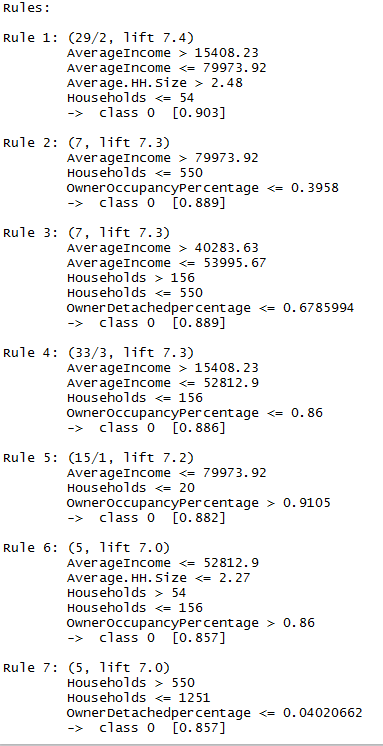
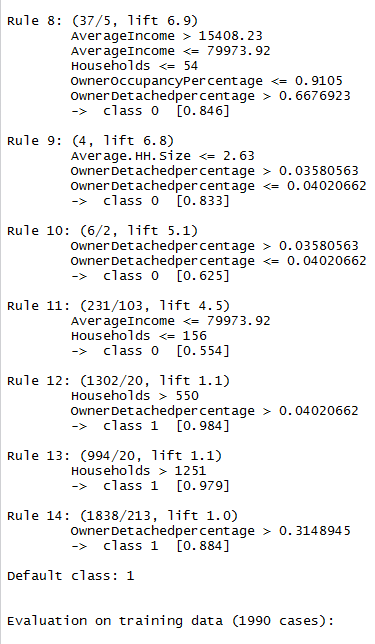
The first graph shows that if a variable is assigned values by random permutation by how much will the MSE increase. In this case if we randomly permute the households (i.e. an observation which had households =100 but you randomly assign the households = 500 the MSE will increase by 100% on an average. Higher the value, higher the variable importance.

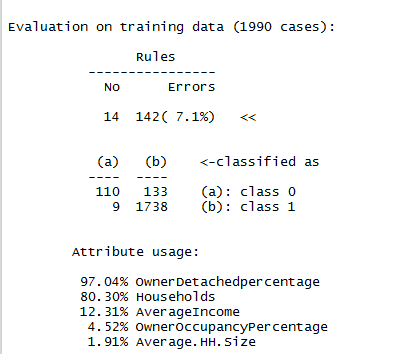
On the other hand, Node purity is measured by Gini Index which is the difference between RSS before and after the split on that variable.

Model 3 Bayes Classifier



## Step 6: Knowledge Discovery



## Step 6: Measure Performance

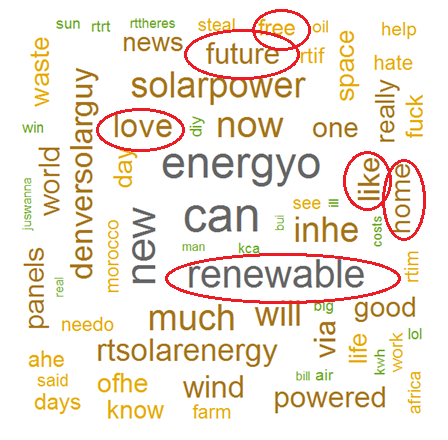
The performance of each model was measure and plotted in an ROC Curve

An ROC curve is the most commonly used way to visualize the performance of a binary classifier.

# Results

First we look at the first part of the analysis, Solar acceptance.

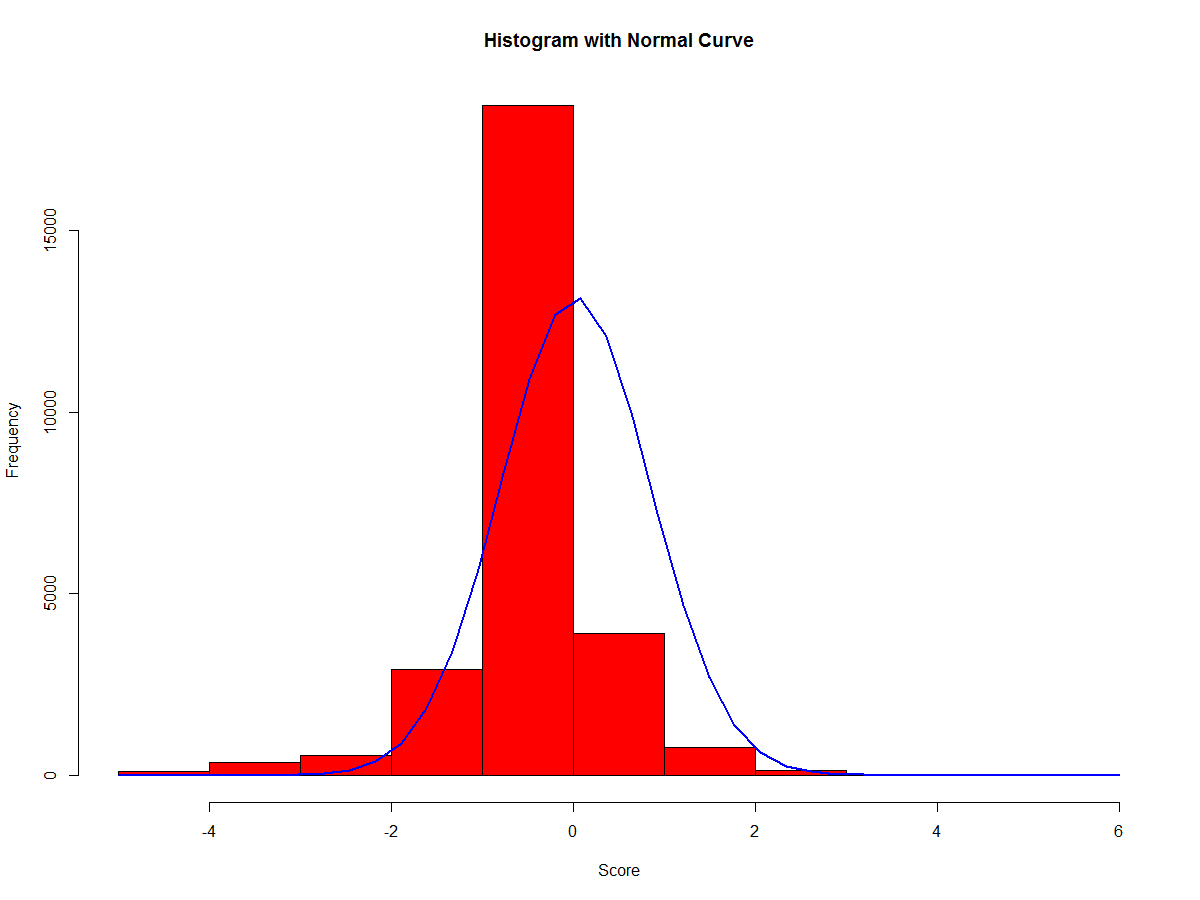
Below is a word cloud of tweet gather over a period of time containing search results for solar keywords



As we can see here there are several positive words, associated with the search which lead us to believe

There is a positive “feeling” associated with solar energy.

Now we analyze the positive and negative words



Here we can see most of the works scored neutral score = 0, however, we can also see there is a slight lift on the positive side of the graph, which supports the previous word cloud



Based on this we would assume the acceptance and sentiment for Solar Power is positive, next we look at how to predict if someone is more likely to purchase solar installations.

During the analysis we looked at four main models:

1. Logistic
2. Multivariate
3. Random Forest
4. Naïve Baynes.

##### Using 5 variables to predict SolarInstalled

* Households
* AverageIncome
* Average.HH.Size
* OwnerOccupancyPercentage
* OwnerDetachedpercentage

That is a great benefit of using an ROC curve to evaluate a classifier instead of a simpler metric such as misclassification rate, in that an ROC curve visualizes all possible classification thresholds, whereas misclassification rate only represents error rate for a single threshold.

## 

#### Logistic



#### Multivariate



#### Random Forest



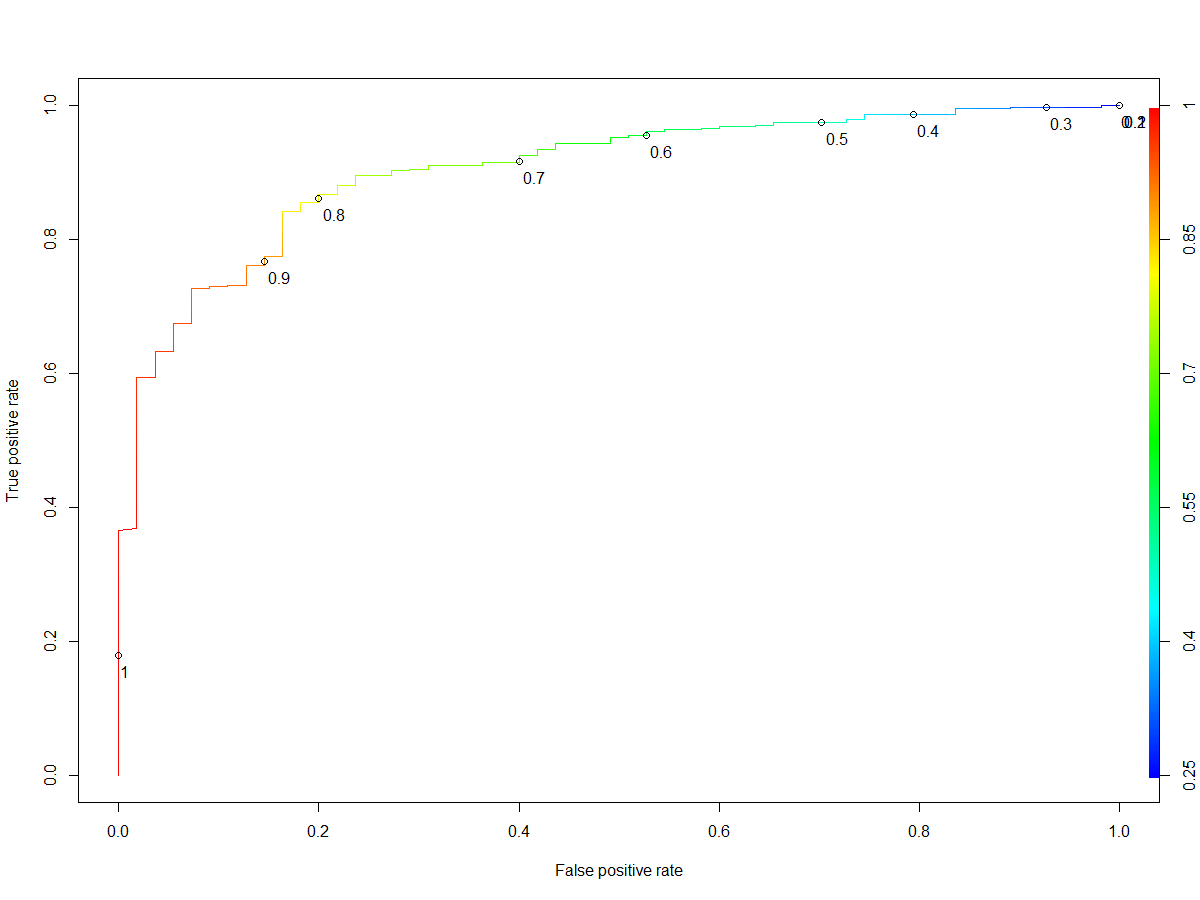
#### Naïve Baynes



## 

#### Random Forest

Random Forest performed better than the Logistic, Multivariate and Naïve Baynes models. This means the Random forest classifier has done a better job at separating the classes. Therefore we would use Random Forest to predict whether or not a person is more or less likely based on the variables

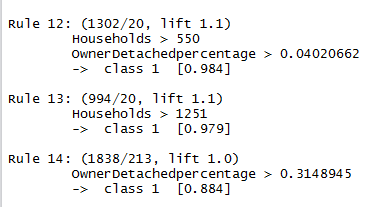


Based on the performance of each model, it was determined that the Random Forest Model was the best performing model with a 90.4% accuracy and did very good job at separating the classes.

# Conclusions

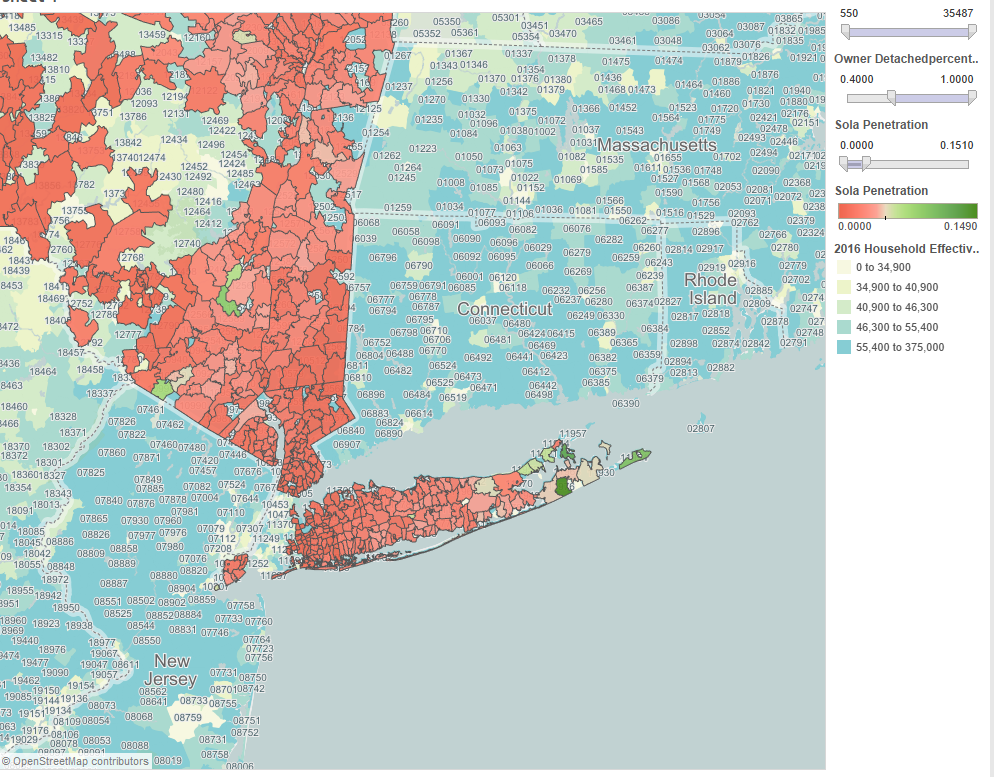
In this report, I have considered several demographic variables for predicting and selecting suitable areas with a higher probability within New York for potential residents to have solar installations on their residential rooftops. I did analysis on some of o the most available and usable demographic data from several census and open data sources,

Using Statistical models I made zip level estimates/predictions of the probability of solar installations, and generated some simple rules that can be implemented to target specific set of customers



During the analysis I was able to determine that areas where there is a higher penetration of detached homes, where there is a higher number of households and higher income are more likely to have installations

The green areas in the map below show the targeted Zips for solar installations



## Future Work

There are many possible directions for future analyses, and it is my hope that the data there will be more data available from municipalities, utility providers, and solar energy researchers.

There is also a potential to incorporate Lidar data to estimate rooftop area suitable for Solar development/installations. The Lidar data correlates to the elevation of the first object detected and creates a digital surface model for each city.