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 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.

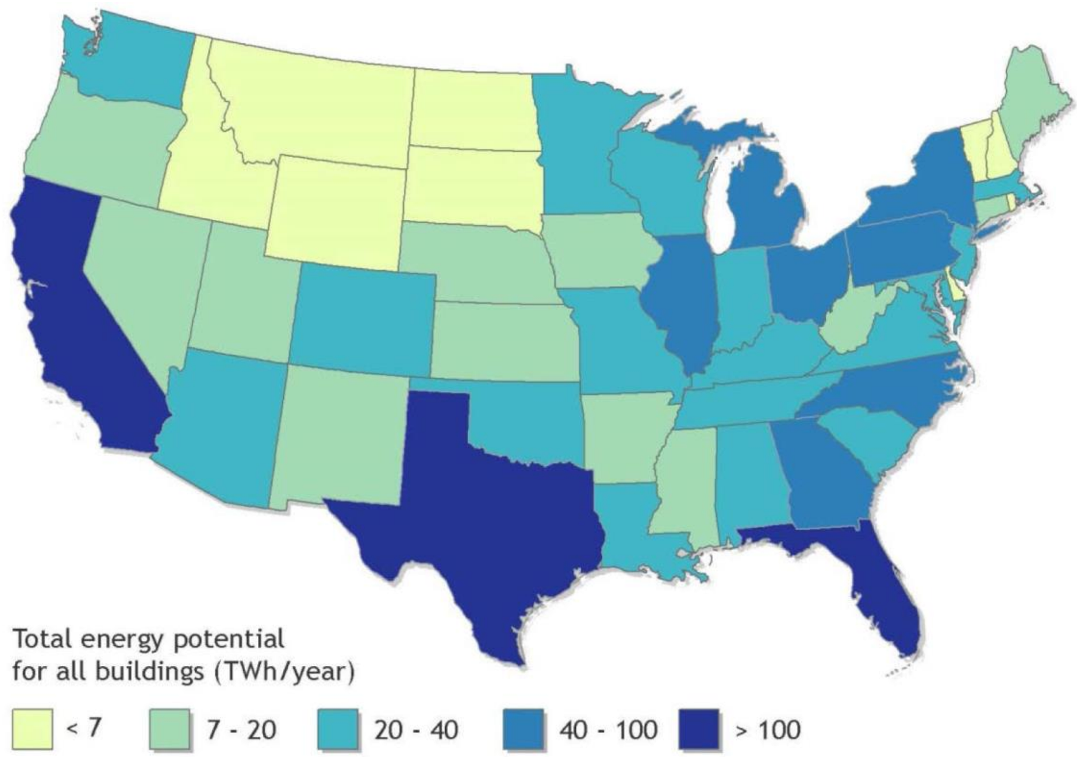


Figure 17. Total annual energy generation potential from rooftop PV for all building sizes

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:

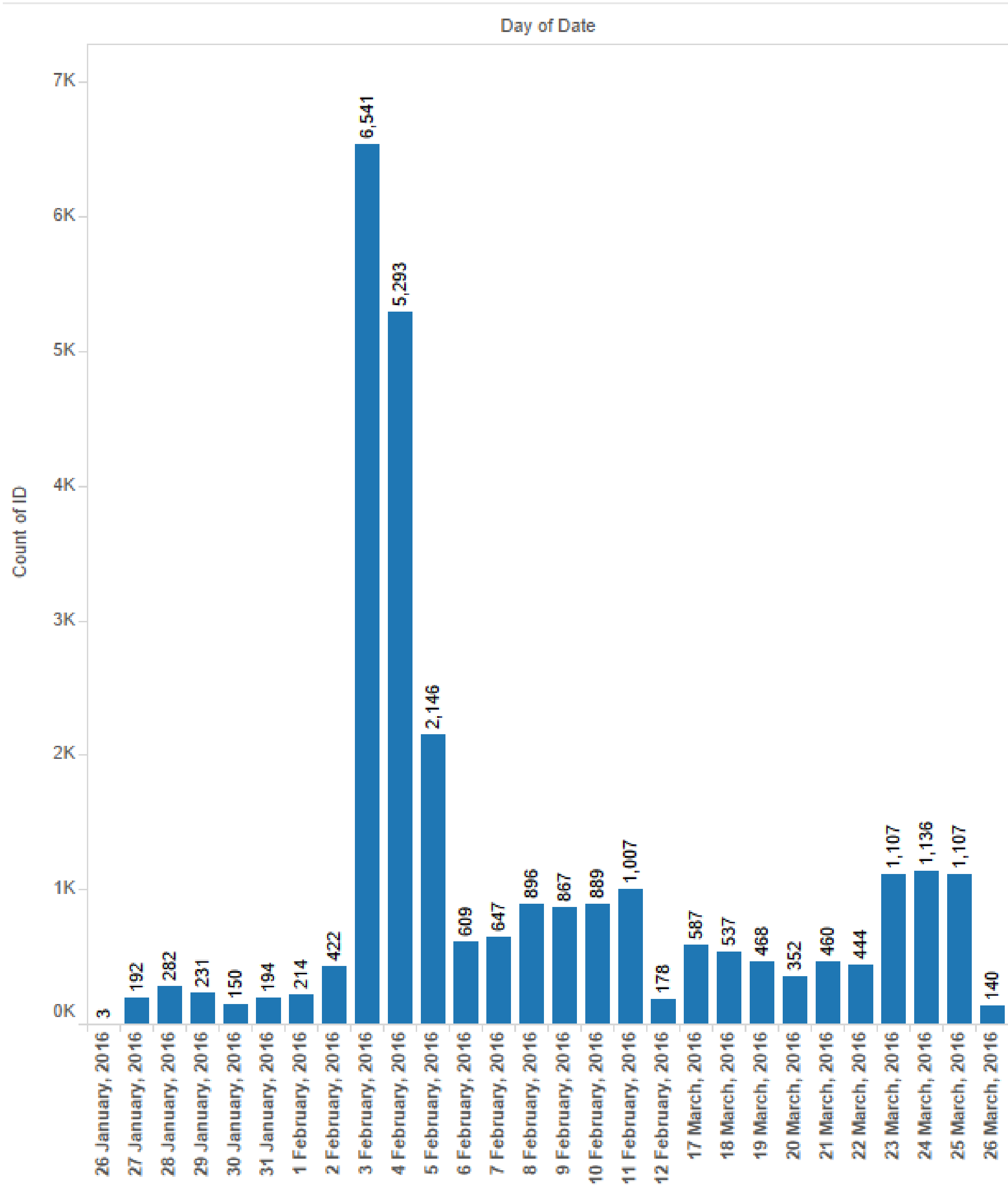
https://github.com/jcasallas/Capstone/tree/master/Code

File: Twitter_SB Github

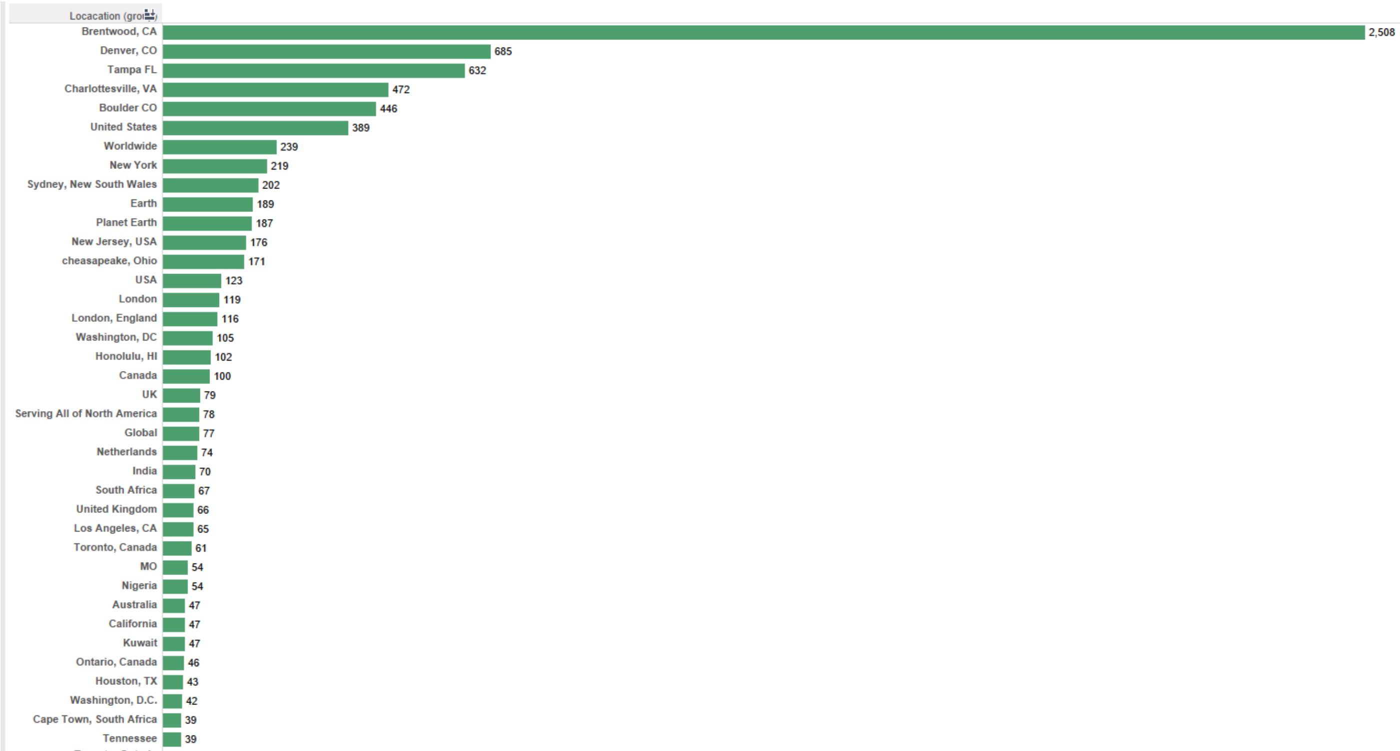
This data contains 7 attributes

```
> summary(TweetData)
      ID              TweetText            UserID             UserName           Locacation             Date           
 Length:27178       Length:27178       Length:27178       Length:27178       Length:27178       Min.   :2016-01-26  
 Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:2016-02-03  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :2016-02-05  
                                                                                                 Mean   :2016-02-15  
                                                                                                 3rd Qu.:2016-02-11  
                                                                                                 Max.   :2016-03-26  
> str(TweetData)
'data.frame':	27178 obs. of  6 variables:
 $ ID        : chr  "695457638696541952" "695456228248579968" "695456151245185024" "695455270873332992" ...
 $ TweetText : chr  "RT elephantjournal Hillary Clinton #demdebate Strong reply on #greenenergy, #obamacare, #smallbusiness" "@HillaryClinton Got so much business to do Will focu
s on #greenenergy reduce #drugprices Paid #maternityleave #Immigrationrefor"| __truncated__ "Hillary Clinton #demdebate Strong reply on #greenenergy, #obamacare, #smallbusiness" 
"When will we get to the environment and clean energy? #DemDebate #environment #GreenEnergy" ...
 $ UserID    : chr  "4749215900" "2323398798" "17514354" "19558266" ...
 $ UserName  : chr  "Jacqueline Franchet" "Sairam Radhakrishnan" "Waylon Lewis" "Nicole Wagner" ...
 $ Locacation: chr  "Northwest Harwich, MA" "New York, NY" "Boulder, Colorado" "On the move" ...
 $ Date      : Date, format: "2016-02-05" "2016-02-05" "2016-02-05" "2016-02-05" ...
> 
```

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

```
> summary (Solar)
     Zip            InstallationsPerZip   Households          Pop          AverageIncome     Centile_HHI     Sola_Penetration
 Length:1990        Min.   :  0.00      Min.   :    1.0   Min.   :     1   Min.   :  1022   Min.   : 1.000   Min.   :0.000000
 Class :character   1st Qu.:  2.00      1st Qu.:  358.2   1st Qu.:  1076   1st Qu.: 54380   1st Qu.: 5.000   1st Qu.:0.001795
 Mode  :character   Median :  8.00      Median : 1251.0   Median :  3754   Median : 64715   Median : 6.000   Median :0.006591
                    Mean   : 28.57      Mean   : 3690.5   Mean   : 11072   Mean   : 75299   Mean   : 5.802   Mean   :0.016052
                    3rd Qu.: 25.00      3rd Qu.: 4473.5   3rd Qu.: 13422   3rd Qu.: 85513   3rd Qu.: 7.000   3rd Qu.:0.017666
                    Max.   :582.00      Max.   :35487.0   Max.   :106461   Max.   :361842   Max.   :10.000   Max.   :1.333333

 Centile_Solar_Penetration SolarInstalled    Average.HH.Size Centile_HHSize                          Electric.Utility OwnerOccupancyPercentage
 Min.   : 1.000            Min.   :0.0000   Min.   :1.080   Min.   : 1.000   National Grid                 :564     Min.   :0.0000
 1st Qu.: 6.000            1st Qu.:1.0000   1st Qu.:2.310   1st Qu.: 5.000   NYS Electric and Gas          :428     1st Qu.:0.6625
 Median : 9.000            Median :1.0000   Median :2.510   Median : 8.000   Other                         :243     Median :0.7802
 Mean   : 7.607            Mean   :0.8779   Mean   :2.546   Mean   : 6.941   Consolidated Edison           :232     Mean   :0.7263
 3rd Qu.:10.000            3rd Qu.:1.0000   3rd Qu.:2.770   3rd Qu.: 9.000   PSEG Long Island              :199     3rd Qu.:0.8572
 Max.   :10.000            Max.   :1.0000   Max.   :5.500   Max.   :10.000   Central Hudson Gas and Electric:152     Max.   :1.0000
                                                                              (Other)                       :172

 OwnerDetachedpercentage
 Min.   :0.0000
 1st Qu.:0.7603
 Median :0.8620
 Mean   :0.7877
 3rd Qu.:0.9287
 Max.   :1.0000

> str(Solar)
'data.frame':	1990 obs. of  14 variables:
 $ Zip                      : chr  "10001" "10016" "10031" "10461" ...
 $ InstallationsPerZip      : int  9 3 8 16 2 20 4 15 1 31 ...
 $ Households               : int  5892 16634 19816 16027 5259 1728 61 1420 558 4413 ...
 $ Pop                      : int  17678 49904 59450 48081 15778 5184 185 4260 1675 13240 ...
 $ AverageIncome            : num  123113 144872 51413 58037 65712 ...
 $ Centile_HHI              : int  8 8 7 6 5 7 8 9 7 8 ...
 $ Sola_Penetration         : num  0.001527 0.00018 0.000404 0.000998 0.00038 ...
 $ Centile_Solar_Penetration: int  10 10 10 10 10 8 1 8 10 9 ...
 $ SolarInstalled           : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Average.HH.Size          : num  1.93 1.74 2.79 2.62 2.54 2.96 3.7 2.71 2.64 2.15 ...
 $ Centile_HHSize           : int  10 10 5 7 7 3 1 6 7 10 ...
 $ Electric.Utility         : Factor w/ 9 levels "Central Hudson Gas and Electric",..: 2 2 2 2 2 2 5 1 5 2 ...
 $ OwnerOccupancyPercentage : num  0.274 0.293 0.101 0.315 0.37 ...
 $ OwnerDetachedpercentage  : num  0.0193 0.0195 0.0284 0.3341 0.3556 ...
> 
```

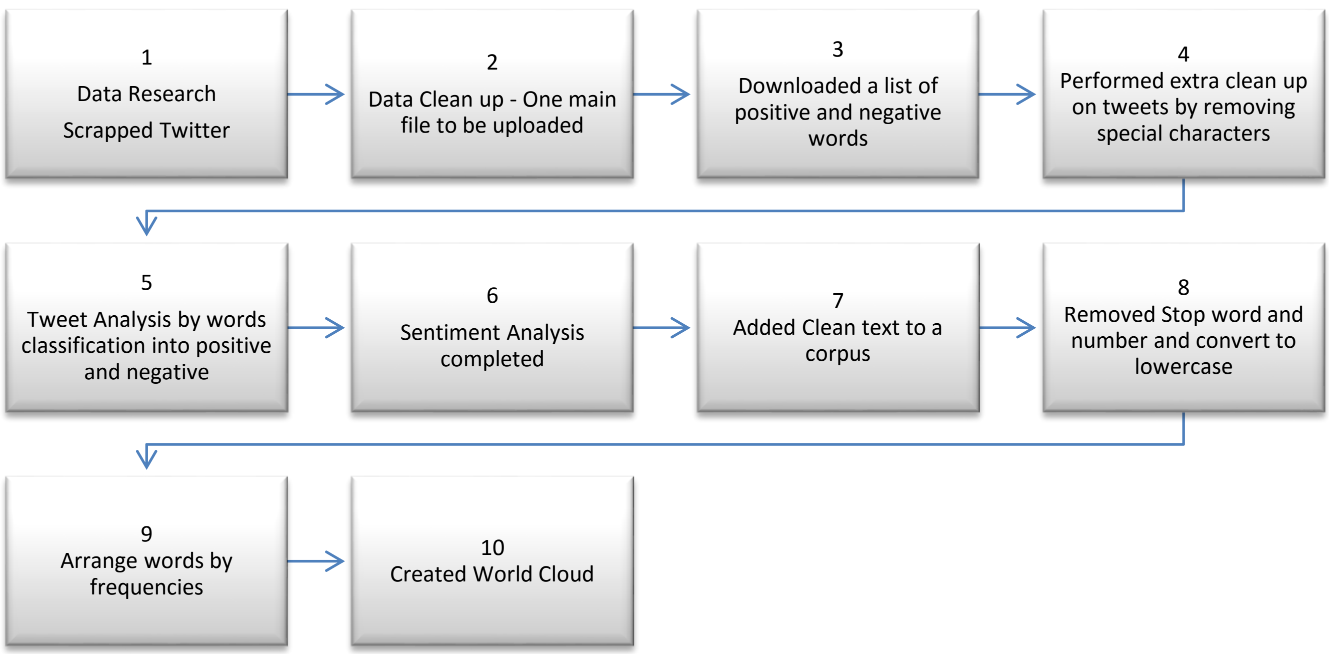
Snapshot of the data set

	Zip	InstallationsPerZip	Households	Pop	AverageIncome	Centile_HHI	Sola_Penetration	Centile_Solar_Penetration	SolarInstalled	Average.HH.Size	Centile_HHSize	Electric.Utility	OwnerOccupancyPercentage	OwnerDetachedpercentage
1	10001	9	5892	17678	123112.78	8	0.001527495	10	1	1.93	10	Consolidated Edison	0.2735	0.01931260
2	10016	3	16634	49904	144872.39	8	0.000180353	10	1	1.74	10	Consolidated Edison	0.2932	0.01953125
3	10031	8	19816	59450	51413.20	7	0.000403714	10	1	2.79	5	Consolidated Edison	0.1007	0.02842255
4	10461	16	16027	48081	58036.91	6	0.000998315	10	1	2.62	7	Consolidated Edison	0.3150	0.33409276
5	10470	2	5259	15778	65711.66	5	0.000380300	10	1	2.54	7	Consolidated Edison	0.3705	0.35561746
6	10502	20	1728	5184	164349.78	7	0.011574074	8	1	2.96	3	Consolidated Edison	0.8338	0.93399340
7	10505	4	61	185	149436.53	8	0.065573770	1	1	3.70	1	NYS Electric and Gas	0.7783	1.00000000
8	10524	15	1420	4260	114246.67	9	0.010563380	8	1	2.71	6	Central Hudson Gas and Electric	0.8893	0.98641509
9	10526	1	558	1675	218303.00	7	0.001792115	10	1	2.64	7	NYS Electric and Gas	0.9277	0.58540630
10	10530	31	4413	13240	133485.61	8	0.007024700	9	1	2.15	10	Consolidated Edison	0.8111	0.44745326

Approach

This research was done in two sections 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

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

```
for tweet in new_tweets:
    #f.write(tweet.text.encode('utf-8') + #'\n')
    #f.write(jsonpickle.encode(tweet._json, unpicklable=False) +
     f.write(jsonpickle.encode(tweet.id, unpicklable=False) + "," + jsonpickle.encode(tweet.text, unpicklable=False)+
            "," + jsonpickle.encode(tweet.user.id, unpicklable=False)+
           "," + jsonpickle.encode(tweet.user.name, unpicklable=False)+
           "," + jsonpickle.encode(tweet.user.location, unpicklable=False)+
            "," + jsonpickle.encode(tweet.coordinates, unpicklable=False)+
           "," + jsonpickle.encode(tweet.created_at, unpicklable=False)+
           '\n')
tweetCount += len(new_tweets)

print("Downloaded {0} tweets".format(tweetCount))
```

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

```
for tweet in new_tweets:
    #f.write(tweet.text.encode('utf-8') + #'\n')
    f.write(jsonpickle.encode(tweet._json, unpicklable=False) +
     #f.write(jsonpickle.encode(tweet.id, unpicklable=False) + "," + jsonpickle.encode(tweet.text, unpicklable=False)+
            # "," + jsonpickle.encode(tweet.user.id, unpicklable=False)+
           # "," + jsonpickle.encode(tweet.user.name, unpicklable=False)+
           # "," + jsonpickle.encode(tweet.user.location, unpicklable=False)+
           #  "," + jsonpickle.encode(tweet.coordinates, unpicklable=False)+
           # "," + jsonpickle.encode(tweet.created_at, unpicklable=False)+
           '\n')
tweetCount += len(new_tweets)

print("Downloaded {0} tweets".format(tweetCount))
```

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:

```
pos = scan('c:/Users/casal_000/OneDrive/Documents/CAPSTPONE/CAPSTONE/Words/Positive.txt', what='character', comment.char=';')
neg = scan('c:/Users/casal_000/OneDrive/Documents/CAPSTPONE/CAPSTONE/Words/Negative.txt', what='character', comment.char=';')
```

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.

```r
clean.text <- function(some_txt)
{
  some_txt = gsub("&amp", "", some_txt)

  some_txt = gsub("(RT|via)((?:\b\\W*@\\w+)+)", "", some_txt)

  some_txt = gsub("@\\w+", "", some_txt)

  some_txt = gsub("[[:punct:]]", "", some_txt)

  some_txt = gsub("[[:digit:]]", "", some_txt)

  some_txt = gsub("http\\w+", "", some_txt)

  some_txt = gsub("[ t]{2,}", "", some_txt)

  some_txt = gsub("^\\s+|\\s+$", "", some_txt)

  # define "tolower error handling" function

  try.tolower = function(x)

  {

    y = NA

    try_error = tryCatch(tolower(x), error=function(e) e)

    if (!inherits(try_error, "error"))

      y = tolower(x)

    return(y)

  }

  some_txt = sapply(some_txt, try.tolower)

  some_txt = some_txt[some_txt != ""]

  names(some_txt) = NULL

  return(some_txt)

}
```

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

Used an algorithm for analyzing our words

```
score.sentiment = function(sentences, pos.words, neg.words, .progress='none')
{
  require(plyr)
    require(stringr)

  # written by Jeffrey Breen  modified by Julian Casallas
  # we got a vector of sentences. plyr will handle a list
  # or a vector as an "l" for us
  # we want a simple array ("a") of scores back, so we use
  # "l" + "a" + "ply" = "laply":
  scores = laply(sentences, function(sentence, pos.words, neg.words) {

    # clean up sentences with R's regex-driven global substitute, gsub():
    sentence = gsub('[[:punct:]]', '', sentence)

    sentence = gsub('[[:cntrl:]]', '', sentence)

    sentence = gsub('\\d+', '', sentence)

    # and convert to lower case:

    sentence = tolower(sentence)

    # split into words. str_split is in the stringr package
    word.list = str_split(sentence, '\\s+')

    # sometimes a list() is one level of hierarchy too much
    words = unlist(word.list)

    # compare our words to the dictionaries of positive & negative terms
    pos.matches = match(words, pos.words)

    neg.matches = match(words, neg.words)

    # match() returns the position of the matched term or NA
    # we just want a TRUE/FALSE:
    pos.matches = !is.na(pos.matches)
    neg.matches = !is.na(neg.matches)

    # and conveniently enough, TRUE/FALSE will be treated as 1/0 by sum():
    score = sum(pos.matches) - sum(neg.matches)

    return(score)

  }, pos.words, neg.words, .progress=.progress )

  scores.df = data.frame(score=scores, text=sentences)

  return(scores.df)

}
```

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.

```
> table(analysis$score)

   -5    -4    -3    -2    -1     0     1     2     3     4     5     6 
    1    90   355   539  2913 18449  3903   759   124    29     2     4 
> mean(analysis$score)
[1] 0.01921378
```

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.

```
######################
####WORDCLOUD
#####################



library(plyr); library(dplyr)
require(stringr)


tweet_corpus = Corpus(VectorSource(clean_text))
```

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.

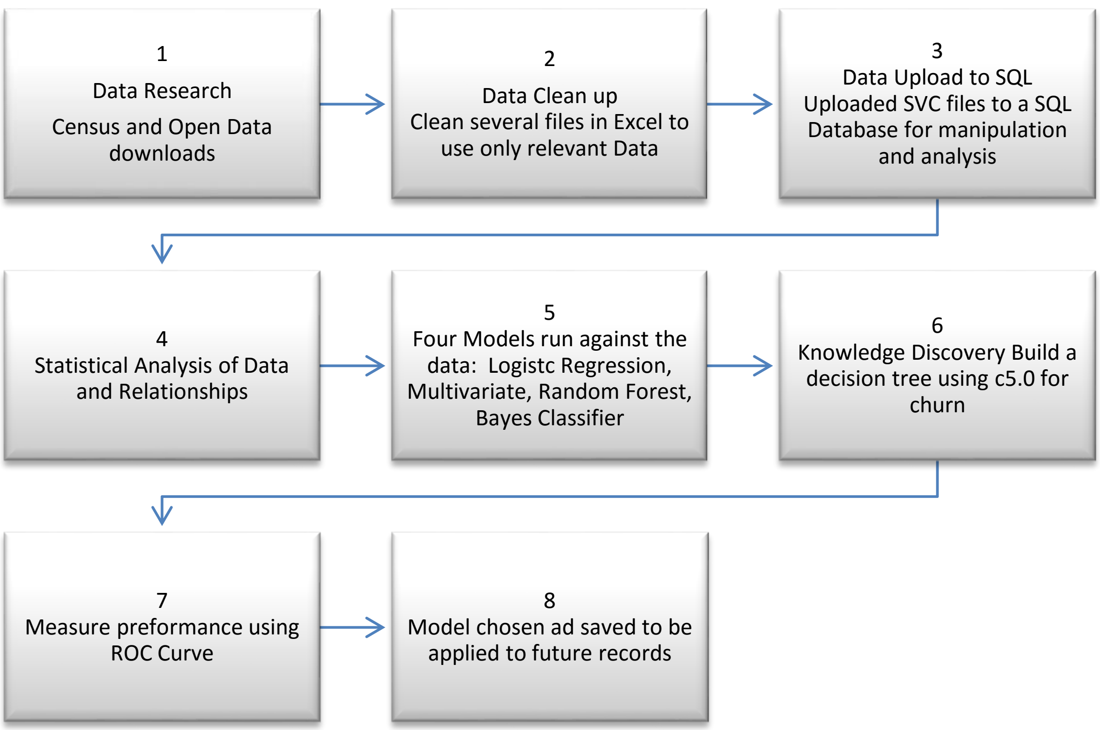
```
tdm = TermDocumentMatrix(tweet_corpus, control = list(removePunctuation = TRUE,stopwords = c("machine", "learning", stopwords("english")), removeNumbers = TRUE, tolower = TRUE))
```

Step 9: Arrange words by frequencies

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

```
require(plyr)

m = as.matrix(tdm) #we define tdm as matrix
word_freqs = sort(rowSums(m), decreasing=TRUE) #now we get the word orders in decreasing order
dm = data.frame(word=names(word_freqs), freq=word_freqs) #we create our data set
wordcloud(dm$word, dm$freq, random.order=FALSE, colors=brewer.pal(8, "Dark2")) #and we visualize our data

#####Use only records with high counts but excluding words in our search criteria
dmmain<-dm[10:200,]
wordcloud(dmmain$word, dmmain$freq, random.order=FALSE, colors=brewer.pal(8, "Dark2")) #and we visualize our data
```

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(InstallationsPerZipas 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.

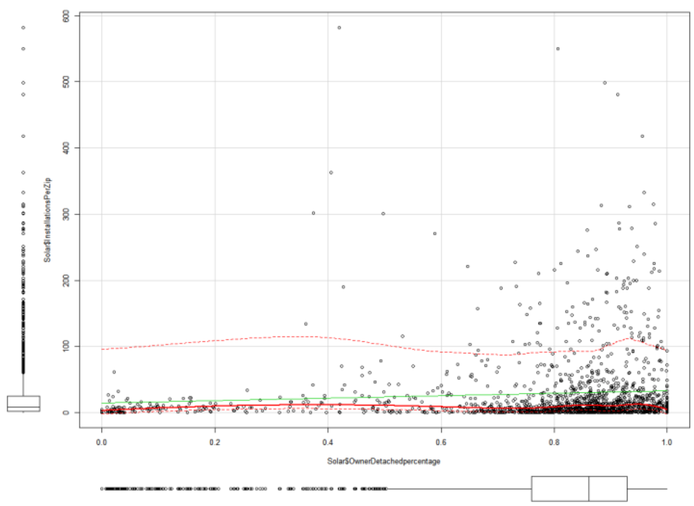


Figure 1 Installations vs DetachedPercentages

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.

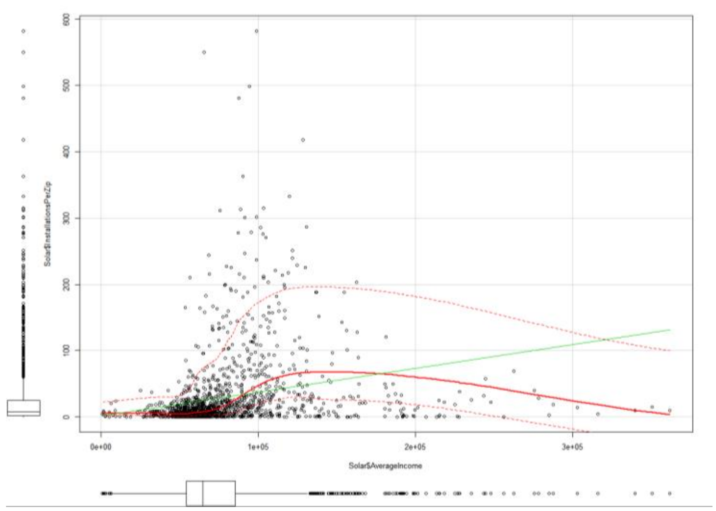


Figure 2 Installations vs Income

Step 5: Predictive Models

Model 1 - logistic regression Model

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -3.763e+00  6.945e-01  -5.419 6.00e-08 ***
AverageIncome             1.032e-05  3.509e-06   2.941  0.00327 ** 
Average.HH.Size           4.338e-01  2.223e-01   1.952  0.05099 .  
Households                6.546e-04  8.805e-05   7.434 1.05e-13 ***
OwnerOccupancyPercentage  1.375e+00  5.362e-01   2.565  0.01032 *  
OwnerDetachedpercentage   2.429e+00  4.793e-01   5.068 4.02e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1133.97  on 1491  degrees of freedom
Residual deviance:  908.59  on 1486  degrees of freedom
AIC: 920.59

Number of Fisher Scoring iterations: 8
```

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.

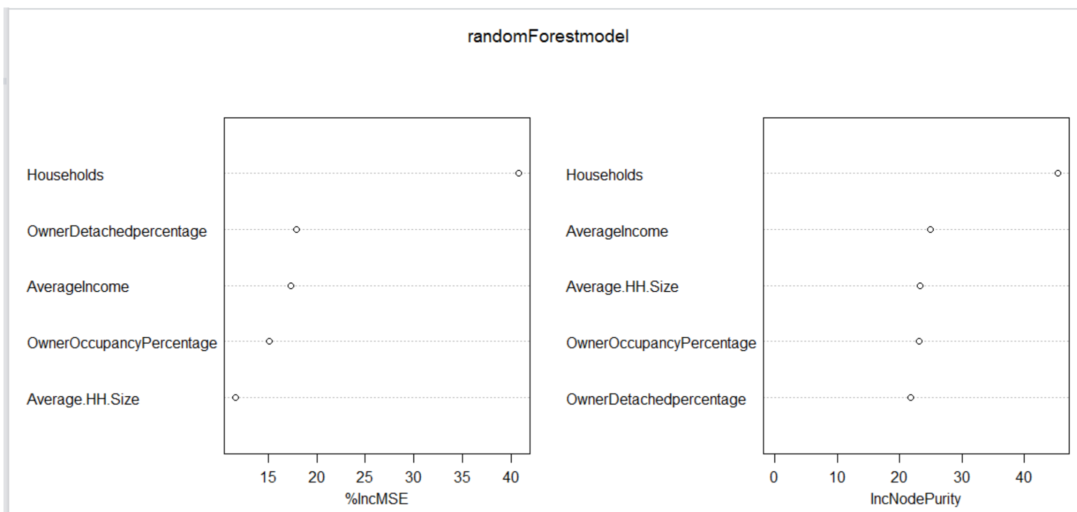
Model 2 Multivariate

```
Call:
lm(formula = SolarInstalled ~ AverageIncome + Average.HH.Size + 
    Households + OwnerOccupancyPercentage + OwnerDetachedpercentage, 
    data = train_MLR)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.97017  0.02395  0.11808  0.15807  0.42350 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               5.411e-01  5.907e-02   9.161  < 2e-16 ***
AverageIncome             7.676e-07  2.265e-07   3.389  0.00072 ***
Average.HH.Size          -8.391e-04  2.131e-02  -0.039  0.96859    
Households                1.699e-05  1.837e-06   9.248  < 2e-16 ***
OwnerOccupancyPercentage  5.088e-02  5.773e-02   0.881  0.37821    
OwnerDetachedpercentage   2.336e-01  4.827e-02   4.839 1.44e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3115 on 1486 degrees of freedom
Multiple R-squared:  0.07561,	Adjusted R-squared:  0.0725 
F-statistic: 24.31 on 5 and 1486 DF,  p-value: < 2.2e-16
```

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

```
> importance(randomForestmodel)
                          %IncMSE IncNodePurity
AverageIncome            17.31001      24.85847
Average.HH.Size          11.62930      23.25109
Households               40.80444      45.39106
OwnerOccupancyPercentage 15.08770      23.12699
OwnerDetachedpercentage  17.90094      21.81486
```



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

```
> table(pred=BayesModelPredict, true=test_NB$SolarInstalled)
    true
pred   0   1
   0  10  16
   1  51 421
> mean(BayesModelPredict==test_NB$SolarInstalled)
[1] 0.8654618
```

Step 6: Knowledge Discovery

```
Rules:

Rule 1: (29/2, lift 7.4)
	AverageIncome > 15408.23
	AverageIncome <= 79973.92
	Average.HH.Size > 2.48
	Households <= 54
	->  class 0  [0.903]

Rule 2: (7, lift 7.3)
	AverageIncome > 79973.92
	Households <= 550
	OwnerOccupancyPercentage <= 0.3958
	->  class 0  [0.889]

Rule 3: (7, lift 7.3)
	AverageIncome > 40283.63
	AverageIncome <= 53995.67
	Households > 156
	Households <= 550
	OwnerDetachedpercentage <= 0.6785994
	->  class 0  [0.889]

Rule 4: (33/3, lift 7.3)
	AverageIncome > 15408.23
	AverageIncome <= 52812.9
	Households <= 156
	OwnerOccupancyPercentage <= 0.86
	->  class 0  [0.886]

Rule 5: (15/1, lift 7.2)
	AverageIncome <= 79973.92
	Households <= 20
	OwnerOccupancyPercentage > 0.9105
	->  class 0  [0.882]

Rule 6: (5, lift 7.0)
	AverageIncome <= 52812.9
	Average.HH.Size <= 2.27
	Households > 54
	Households <= 156
	OwnerOccupancyPercentage > 0.86
	->  class 0  [0.857]

Rule 7: (5, lift 7.0)
	Households > 550
	Households <= 1251
	OwnerDetachedpercentage <= 0.04020662
	->  class 0  [0.857]

Rule 8: (37/5, lift 6.9)
	AverageIncome > 15408.23
	AverageIncome <= 79973.92
	Households <= 54
	OwnerOccupancyPercentage <= 0.9105
	OwnerDetachedpercentage > 0.6676923
	->  class 0  [0.846]

Rule 9: (4, lift 6.8)
	Average.HH.Size <= 2.63
	OwnerDetachedpercentage > 0.03580563
	OwnerDetachedpercentage <= 0.04020662
	->  class 0  [0.833]

Rule 10: (6/2, lift 5.1)
	OwnerDetachedpercentage > 0.03580563
	OwnerDetachedpercentage <= 0.04020662
	->  class 0  [0.625]

Rule 11: (231/103, lift 4.5)
	AverageIncome <= 79973.92
	Households <= 156
	->  class 0  [0.554]

Rule 12: (1302/20, lift 1.1)
	Households > 550
	OwnerDetachedpercentage > 0.04020662
	->  class 1  [0.984]

Rule 13: (994/20, lift 1.1)
	Households > 1251
	->  class 1  [0.979]

Rule 14: (1838/213, lift 1.0)
	OwnerDetachedpercentage > 0.3148945
	->  class 1  [0.884]

Default class: 1


Evaluation on training data (1990 cases):
```

```
Evaluation on training data (1990 cases):

             Rules
      ----------------
        No      Errors

        14  142( 7.1%)   <<


       (a)   (b)    <-classified as
      ----  ----
       110   133    (a): class 0
         9  1738    (b): class 1


	Attribute usage:

	 97.04% OwnerDetachedpercentage
	 80.30% Households
	 12.31% AverageIncome
	  4.52% OwnerOccupancyPercentage
	  1.91% Average.HH.Size
```

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

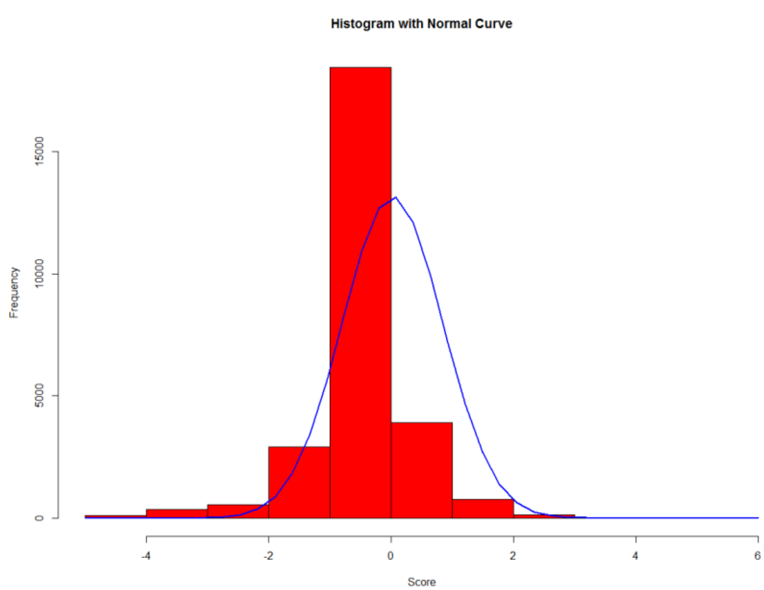
First we look at the first part of the analysis, Solar acceptance.

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

```
   -5    -4    -3    -2    -1     0     1     2     3     4     5     6
    1    90   355   539  2913 18449  3903   759   124    29     2     4
> 
```

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.

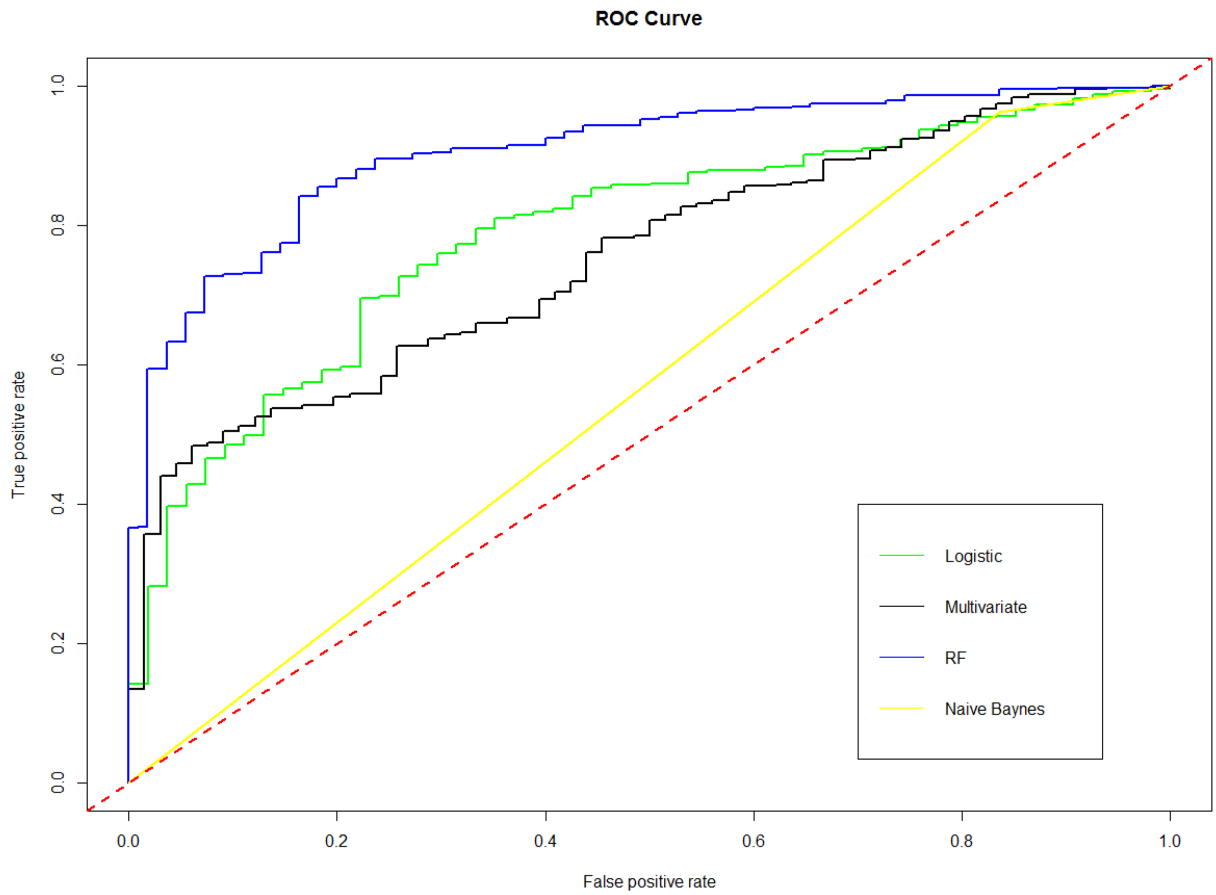
During the analysis we looked at four main models:

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

Using 5 variables to predict Solar Installed

- 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
[1] 0.7873707

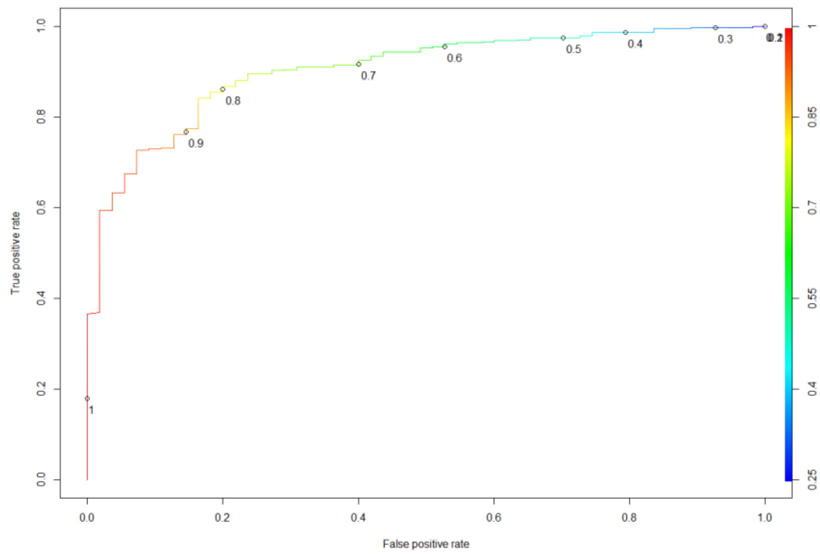
Multivariate
[1] 0.7530163

Random Forest
[1] 0.9043915

Naïve Baynes
[1] 0.5636606

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

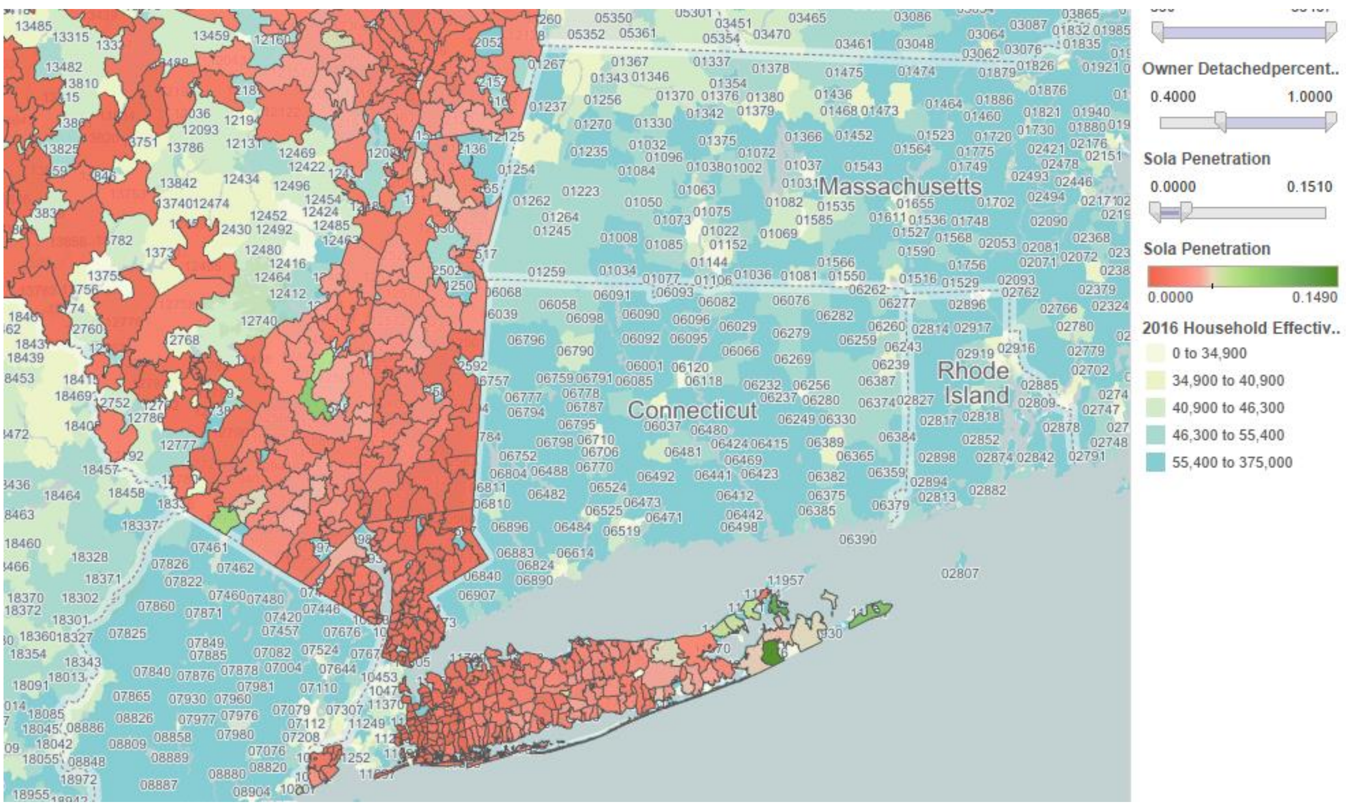
In this report, I have considered several demographic variables for predicting and selecting suitable areas with a higher probability within New York for potential residents to have solar installations on their residential rooftops. I did analysis on some of o the most available and usable demographic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

```
Rule 12: (1302/20, lift 1.1)
        Households > 550
        OwnerDetachedpercentage > 0.04020662
        ->  class 1  [0.984]

Rule 13: (994/20, lift 1.1)
        Households > 1251
        ->  class 1  [0.979]

Rule 14: (1838/213, lift 1.0)
        OwnerDetachedpercentage > 0.3148945
        ->  class 1  [0.884]
```

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.