

INSY 5378-002: Data Science: A Programming Approach

Spring 2017

Group Project 3 – Correlating number of tweets with weather and day of the week

Members:

|  |  |  |
| --- | --- | --- |
| **Name** | **Major** | **Email** |
| Anish Grandhi | Information Systems | anish.grandhi@mavs.uta.edu |
| Meghna Goswami | Information Systems | meghna.goswami@mavs.uta.edu |
| Vinothkumaar Udhayakumar | Industrial Engineering | vinothkumaar.udhayakumar@mavs.uta.edu |

Table of Contents

[I. Problem Statement 3](#_Toc482118135)

[II. Data Collection 3](#_Toc482118136)

[1. People’s actions and moods 3](#_Toc482118137)

[2. Weather 3](#_Toc482118138)

[III. Data Cleaning 4](#_Toc482118139)

[IV. Data exploration 4](#_Toc482118140)

[V. Preparation of Data 11](#_Toc482118141)

[VI. Machine Learning Models 12](#_Toc482118142)

[VII. Changing the Target variable type (Extra Work) 15](#_Toc482118143)

[VIII. Conclusion 18](#_Toc482118144)

[IX. References 18](#_Toc482118145)

# Problem Statement

It’s a general notion that we are happy when the weather is pleasant and we are not happy when the weather is gloomy. Through our data science project, we wanted to analyze people’s moods’, actions and find out if they are correlated to weather. We used twitter to collect our data. We plan to use number of tweets as a measurement of success. If the number of tweets are more compared to other days, we can tell that people are happy and doing the actions mentioned below or at least tweet about it.

The topics are –

1. Happy – The happiness state (or mood) of a person
2. Ride/Going out – The event (or action) of a person going out of his/her home
3. Food – The event (or action) of a person cooking, eating out or talking about food
4. Workout – The event (or action) of a person working out
5. Spending money – The event (or action) of a person buying something.

We collected data on 2 states in the US – California, New Jersey as they are densely populated. Based on the data collected on the above topics and weather data, we try to build a model using various algorithms. We want to predict the number of tweets based on the weather conditions or any other external factors such as day of the week.

# Data Collection

## People’s actions and moods

We used Twython module in Python and collected tweets on the topics based on keywords and locations.

The keywords used for each topic are –

1. Happy - 'happy', 'excited', 'joy'
2. Ride - 'beach', 'ride', 'outing', 'sunny', 'going out'
3. Food - 'food', 'restaurant', 'eating out', 'burger', 'pizza', 'cheese'
4. Workout - 'workout', 'cardio', 'gym', 'exercise', 'jogging'
5. Spending money – ‘bought’

The data was collected over a period -

* New Jersey - 7-Apr-2017 to 29-Apr-2017
* California – 9-Apr-2017 to 29-Apr-2017

## Weather

For weather information, we used the historical weather information found at wunderground’s website – www.wunderground.com. Selecting a place and date will give weather information. We copied the information as a table in excel file.

# Data Cleaning

Data cleaning process involved in removing the tweets that do not fit our problem statement. For the topic “happy”, we removed all the tweets tweeted by people because of events, special occasions (Earth day, Penguin Day), advertisements. Special occasions include birthdays, anniversary, Easter etc. We also removed tweets containing words such as ‘unhappy’, ‘not happy’.

For other topics ‘ride’, ‘food’, ‘workout’ – we removed tweets which contained advertisements.

# Data exploration

1. Tweet Count

After cleaning the data (removing tweets that are irrelevant), the tweets count from each of the state are –

|  |  |  |
| --- | --- | --- |
| **Topics/State** | **California** | **New Jersey** |
| Happy | 29109 | 22893 |
| Workouts | 4254 | 2748 |
| Going out | 26421 | 7083 |
| Food | 19131 | 14656 |
| Spending | 4975 | 3357 |

From the above tweets count, we can see there is not much difference between the states based on the topics except for Workouts and “Going out”.

California has approximately 35% more tweets in terms of workouts which could mean that California is a much more active state than New Jersey.

California has a staggering 26421 tweets compared to 7083 tweets from New Jersey when it comes to going out. This is probably because California has more number of beaches (Long Beach, Laguna Beach etc.) from where a lot of people tweet, post pictures.

Overall, California looks more active on Twitter than New Jersey as more number of tweets were collected in California compared to New Jersey.

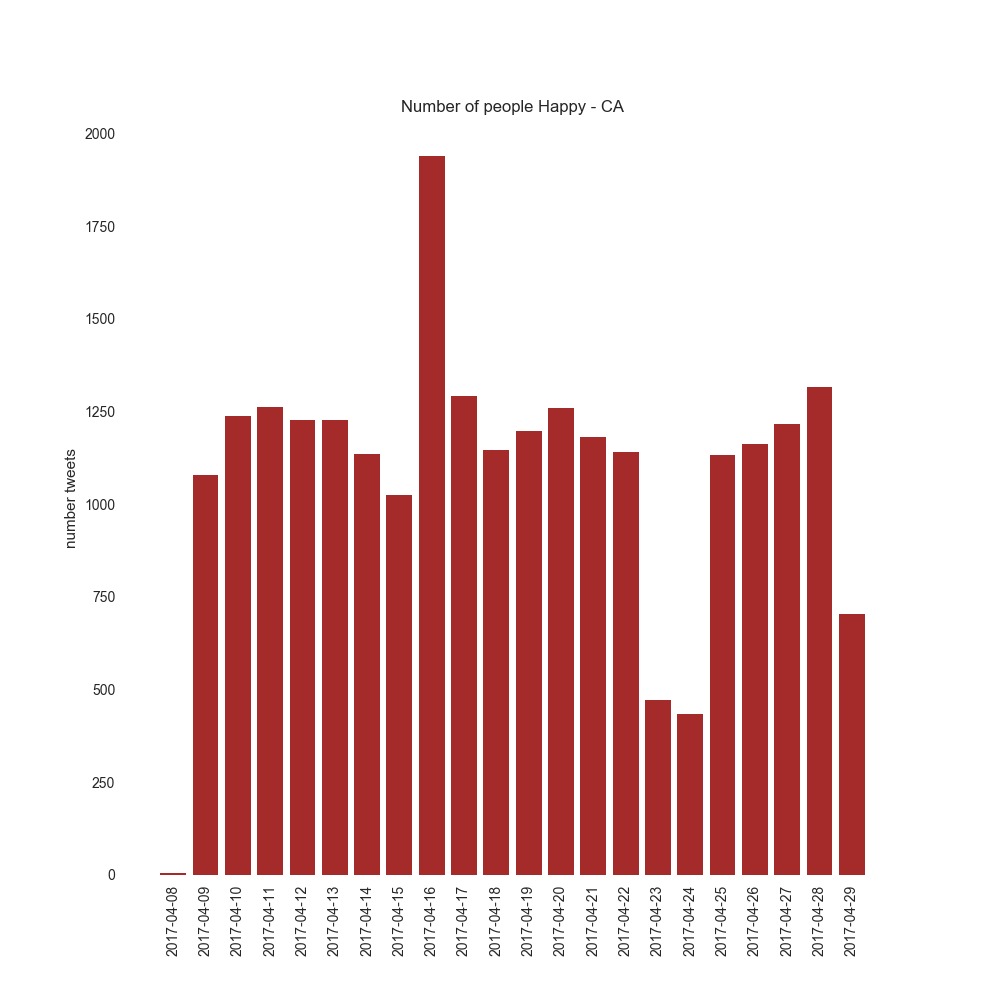
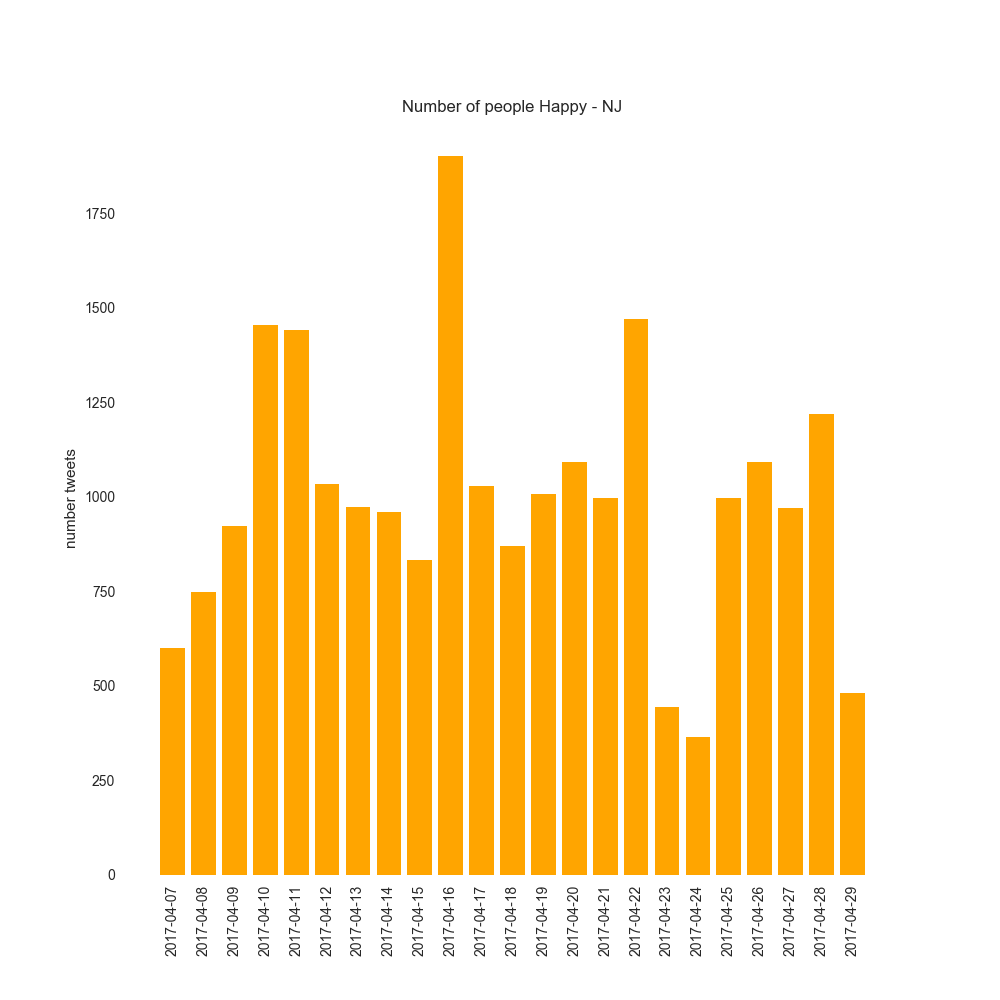
1. Weather in New Jersey and California

Weather in New Jersey

Weather in California

From the weather plots, we can see that New Jersey has temperatures spiking up and down. It has temperatures as high as 79F and as low as 48F. On the other hand, California has a moderate and pleasant temperature conditions. The humidity in both the regions are erratic.

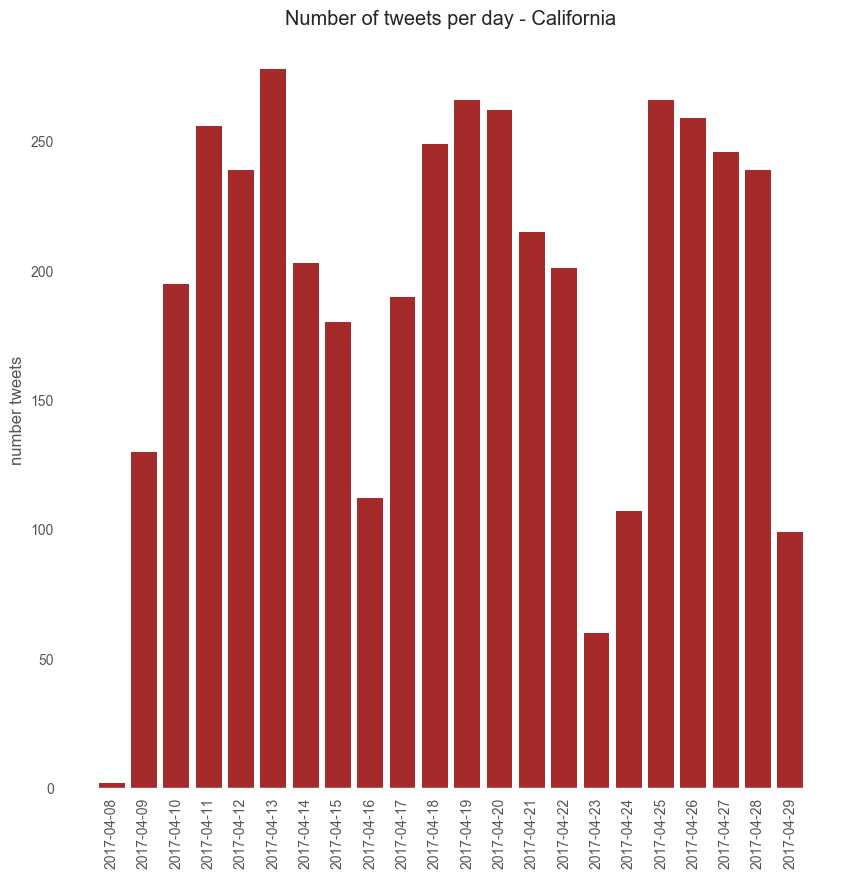
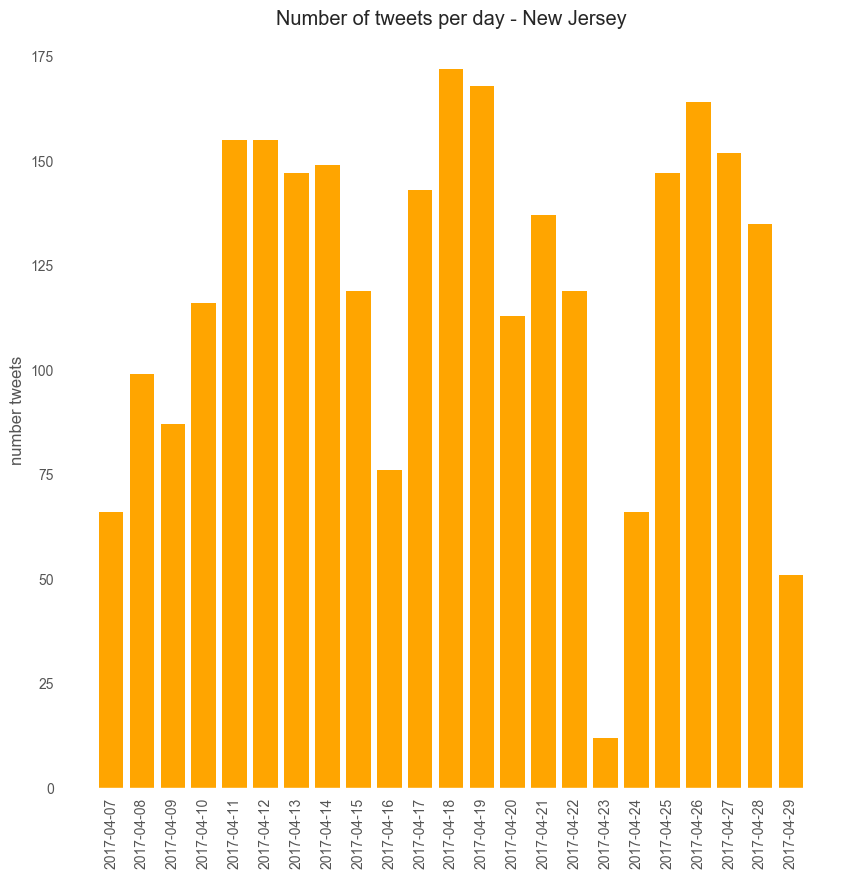
We plotted topic-wise histograms – number of tweets collected per day in California and New Jersey.

* Topic – Happy

Number of people happy in CA

Number of people happy in CA

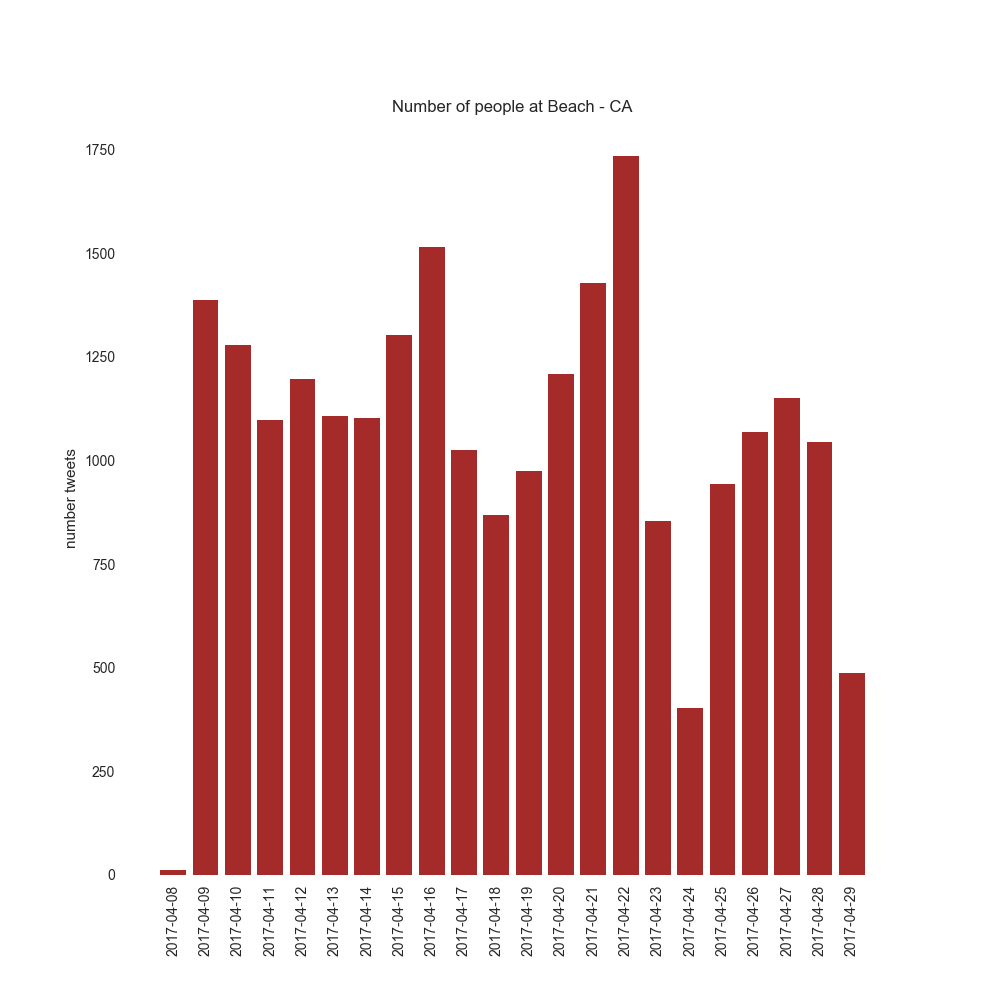
Looking at the histograms on topic Happy, the weekdays show no much difference in the number of tweets. But on weekends, one weekend (23-Apr) shows around 500 tweets while another weekend (16-Apr) shows approximately 2000 tweets. 16-Apr happens to be Easter because of which there seems to be a surge in the number of tweets on that day owing to celebrations.

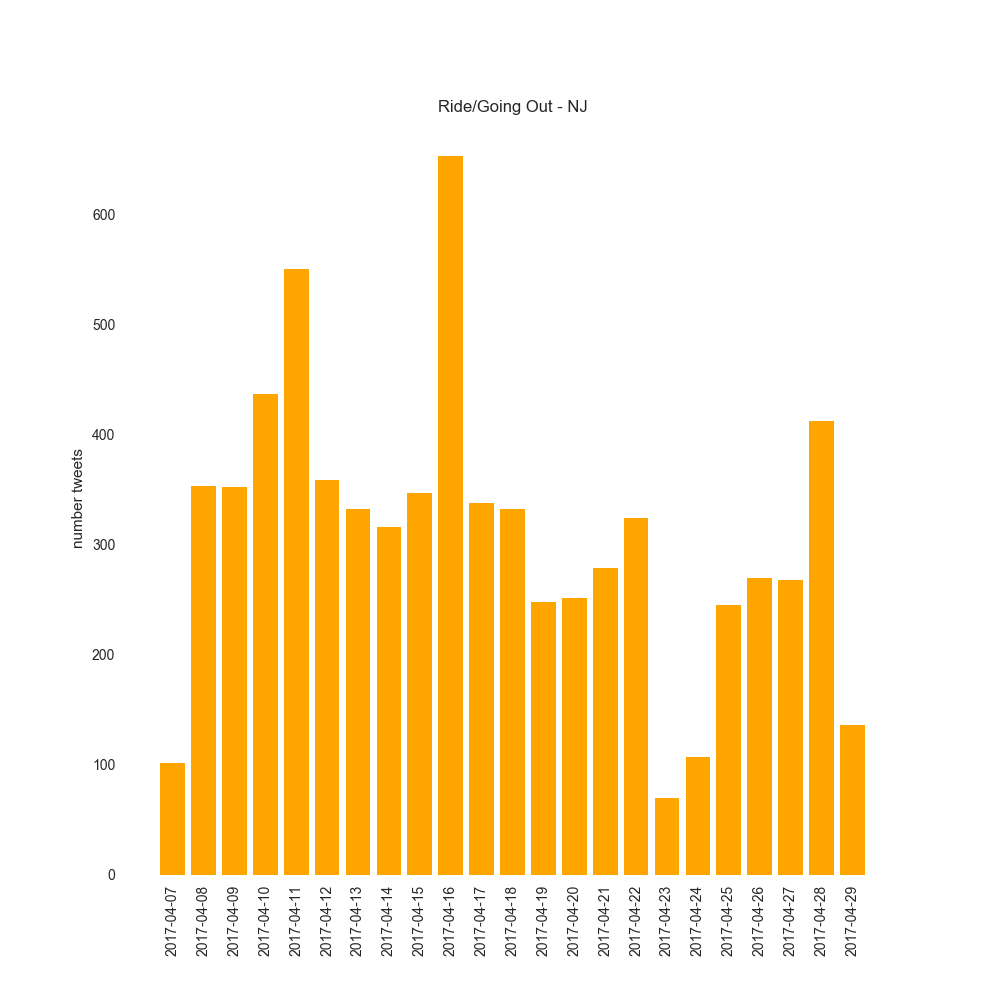
* Topic – Workout

Number of tweets by people on Workout in New Jersey

Number of tweets by people on Workout in California

Looking at the histograms on topic Workouts, the weekdays see higher number of tweets compared to Sundays (16-Apr, 23-Apr).

* Topic – Going out/Ride

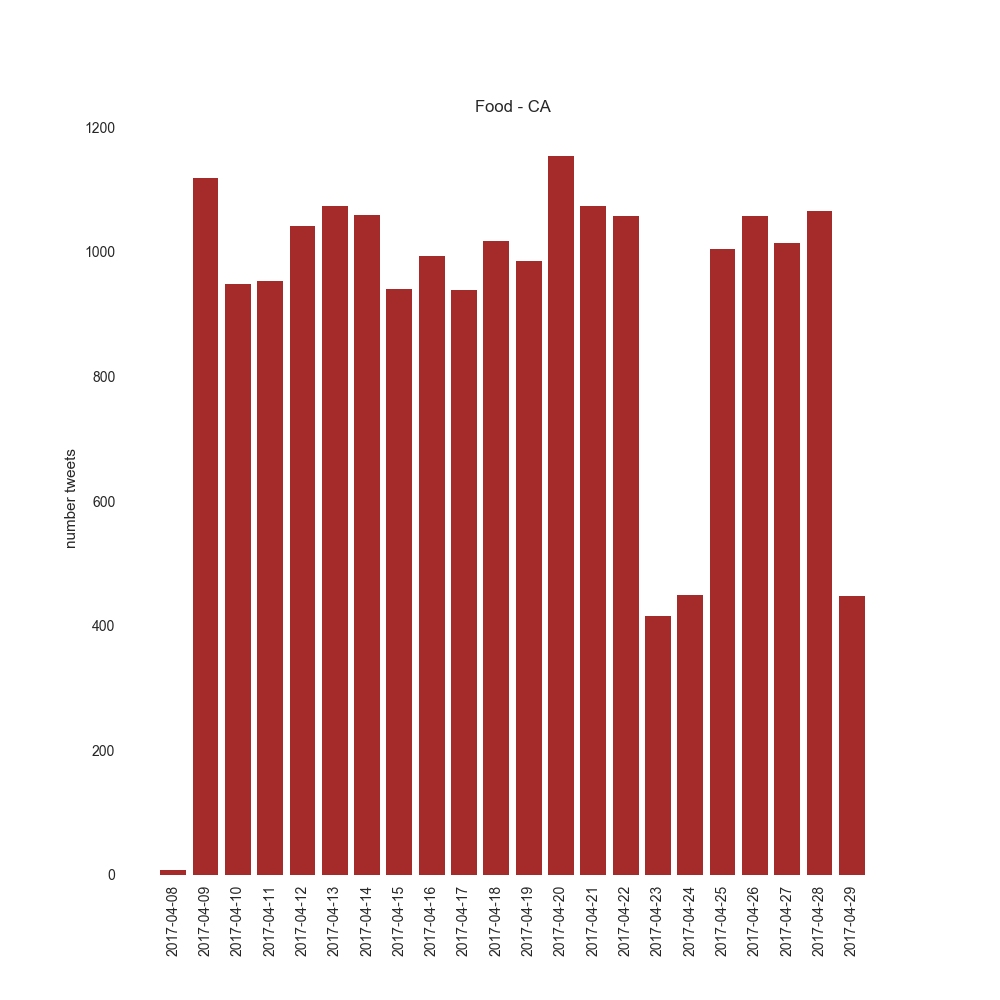


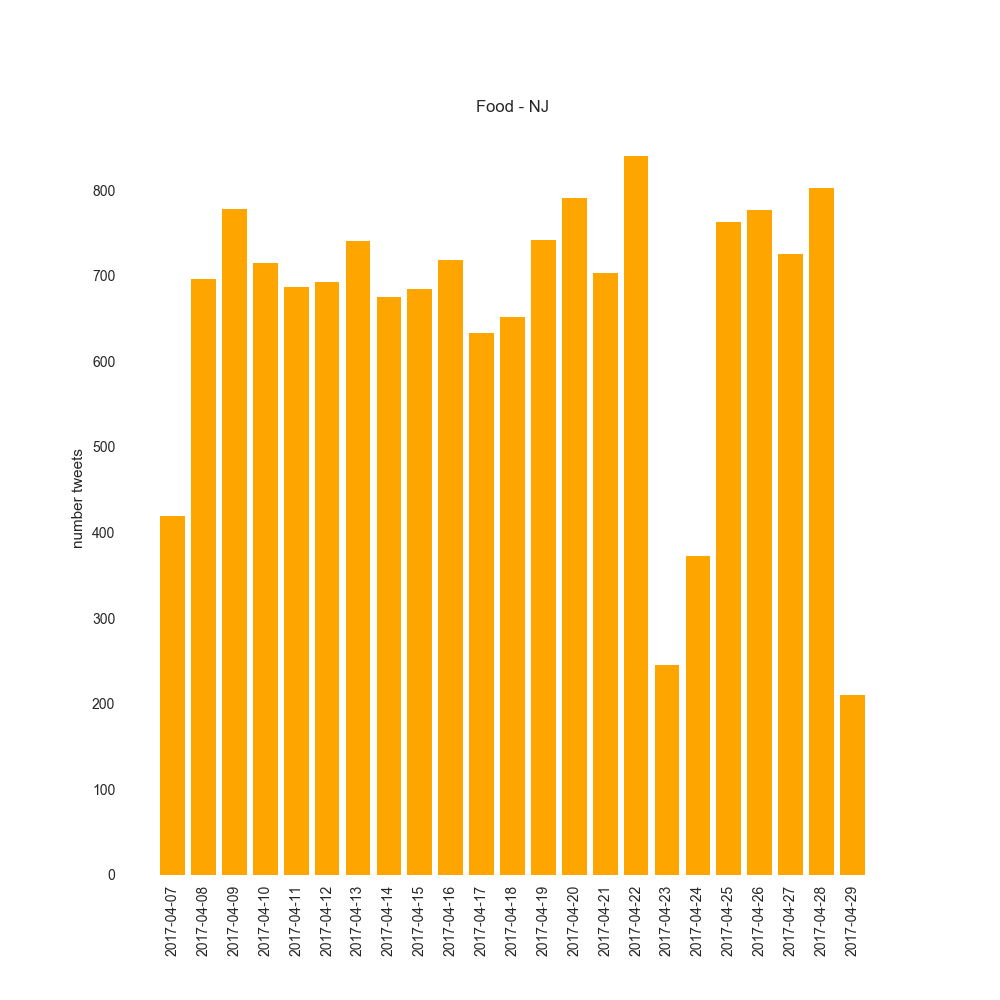
Number of tweets in New Jersey

Number of tweets in New Jersey

Number of tweets in California

Looking at the histogram on topic ride/going out, the number of people who go out of their homes on any given day is lesser in New Jersey compared to California. It is primarily because California has more number of beaches that attract people and all the cities in California are located closer to the beaches.

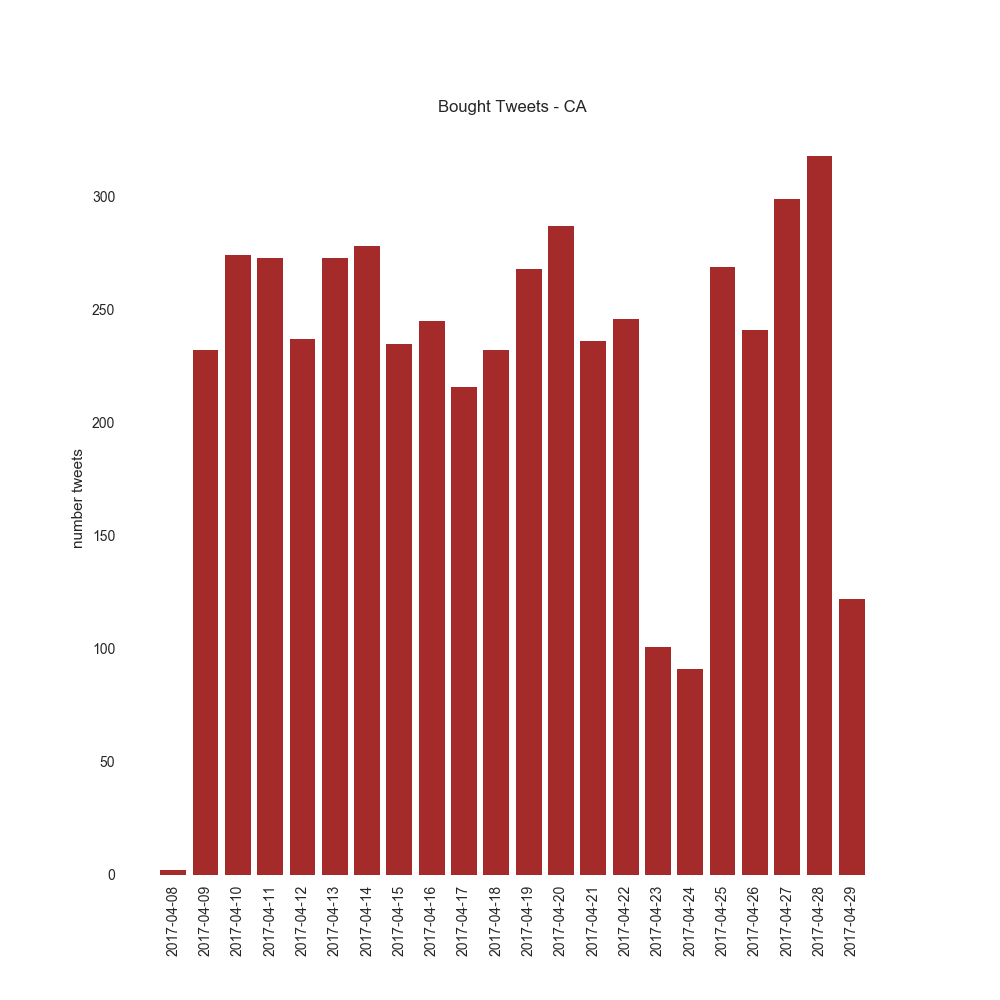
* Topic – Food

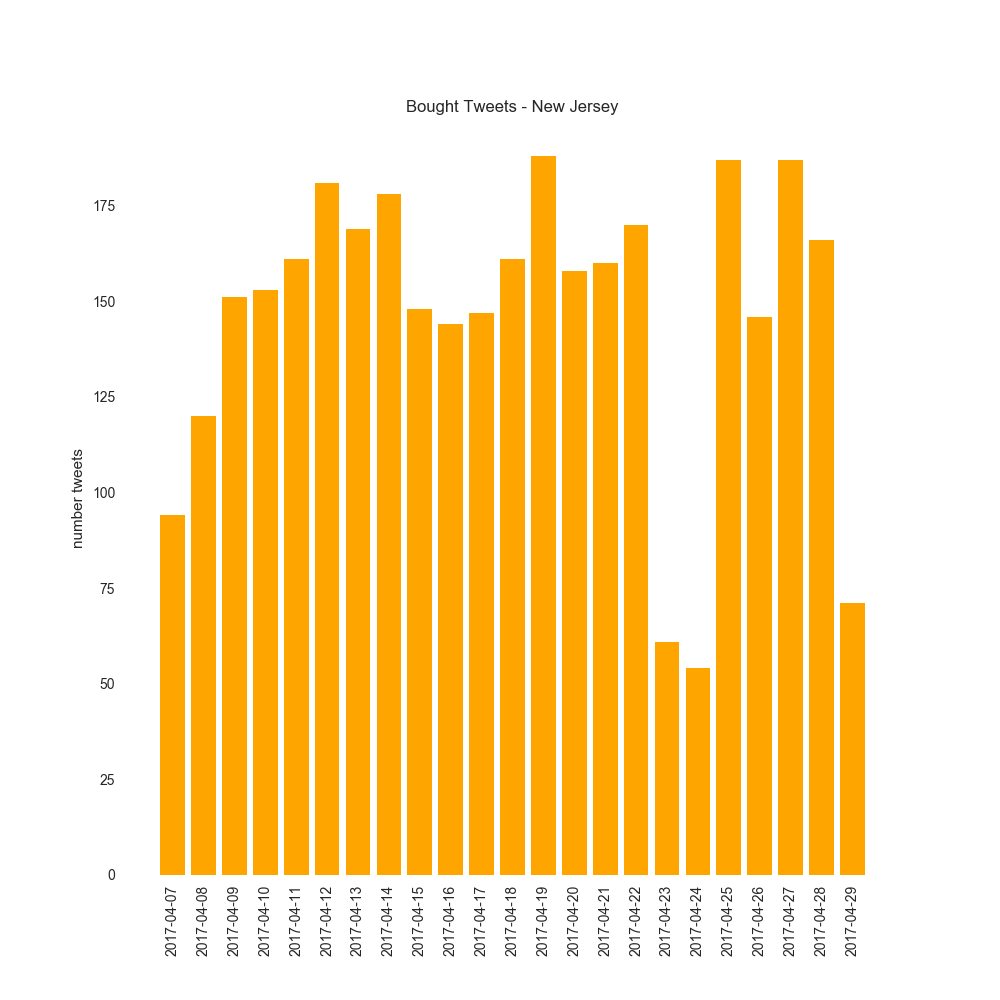


Number of people who tweeted about Food – New Jersey

Number of people who tweeted about Food - California

The histograms on topic Food show that only on one period – 23rd,24th April (Sunday, Monday) the tweets are less. On all other days, number of tweets are approximately same number.

* Topic – Spending money



Number of people who bought in New Jersey

Number of people who bought in California

The above two histograms show the trend in the number of people who tweeted that they have bought something. Here, the number of tweets are approximately same in both California and New Jersey.

# Preparation of Data

To prepare the data for machine learning models, we used python to collect tweets. We used pandas to create a data frame that contains dates, number of tweets each day.

From wunderground’s website, we collected historical weather data for the duration for which we collected tweets. The weather data includes average, maximum, minimum temperature, humidity, and events such as thunderstorm, rain. Events in weather are considered as bad weather. We assigned the value ‘1’ on the day any weather event occurred and ‘0’ if there is no event.

We combined the data of number of tweets and weather data in one excel file for faster access whenever we run the data. Along with this data, we also added another column called ‘Weekend’ since we believed the day of the week could be affecting the number of tweets. For this purpose, all weekdays have been assigned ‘0’ and weekend’s are assigned the number ‘1’.

Following are the features that we will use for training our models –

1. Temp – Average temperature value in a day
2. Humidity – Average humidity value in a day
3. Tmax – Maximum temperature recorded in a day
4. Tmin - Minimum temperature recorded in a day
5. Hmax - Maximum temperature recorded in a day
6. Hmin - Minimum temperature recorded in a day
7. Weather Events – Events such as Thunderstorms, Rain, Fog if occurred, value is ‘1’ or else ‘0’
8. Weekend – If weekend, value is ‘1’ or else ‘0’
9. Number of Tweets – The number of tweets recorded in a day after removing irrelevant tweets. This is our **target** variable.

# Machine Learning Models

Due to less time on our hands, we could not prepare models on ‘bought’ topic.

**Types of models used** -

We used Regression models from sklearn to develop models and test the model’s accuracy.

**Train and Test data –**

As the data that we have is limited we decided to split the data for training and testing purposes.

Training data – 2/3rd of the data set

Testing data – 1/3rd of the data set

**Model Validation –**

We used sklearn’s cross validation kfold method (10 splits) for validating the model i.e., testing the accuracy of the model.

**Selecting Features** –

For feature selection, we must select those features which are highly correlated with the response variable while having low inter correlation with other control variables. To quickly find out which features to be selected, we used Weka software attribute selection feature. We chose ‘CfsSubsetEval’ Attribute evaluator which selects the best features.

**Details of models** –

We used three different regression models – Linear Regression, Ridge (alpha = 2.0), ElasticNet (alpha = 1.0).

Following are the mean accuracy values after cross validation –

1. Topic – Happy, Location California

The feature columns used - 'Temp' (average temperature), 'Weather Events', 'Weekend'



The mean accuracy of the models is in negative. This means that the models are not able to fit the data properly.

Let’s check if adding tweets about birthday would make any difference:



The feature columns used are 'Temp', 'Weather Events'. The accuracy remains same or worsens.

1. Topic – Happy, Location New Jersey

The feature columns used are - 'tmax', 'Weather Events'



The mean accuracy of the models is in negative. This means that the models are not able to fit the data properly.

Let’s check if adding tweets about birthday would make any difference:



As we can see, although the models’ accuracy improves, the mean of the accuracy remains negative.

1. Topic – Workout, Location – California

The features used are – ‘Weekend’



We see that Linear Regression, Ridge models have a mean accuracy of 23.64% and 19.6% respectively. On adding other features such as Average Temperature, Weather Effects then the mean accuracies go down. This means that the number of tweets depend on weekend or weekday than other features which contain weather information.

1. Topic – Workout, Location – New Jersey

The features used are – ‘Weekend’



Again, only when ‘Weekend’ feature is used, the Linear Regression, Ridge models see a positive accuracy. Otherwise, the accuracies are in negative.

1. Topic – Food, Location – California

The features used are – 'Weekend', 'Temp',' Weather Events'



The combination of these three features gives the least negative values of accuracy.

1. Topic – Food, Location – New Jersey

The features used are – 'Weekend'



Only providing ‘Weekend’ gives the least negative values of accuracy.

1. Topic – Ride/Going Out, Location – California

The features used are – 'Weekend', 'Temp',' Weather Events'



The combination of these three features gives the least negative values of accuracy.

1. Topic – Ride/Going Out, Location – New Jersey

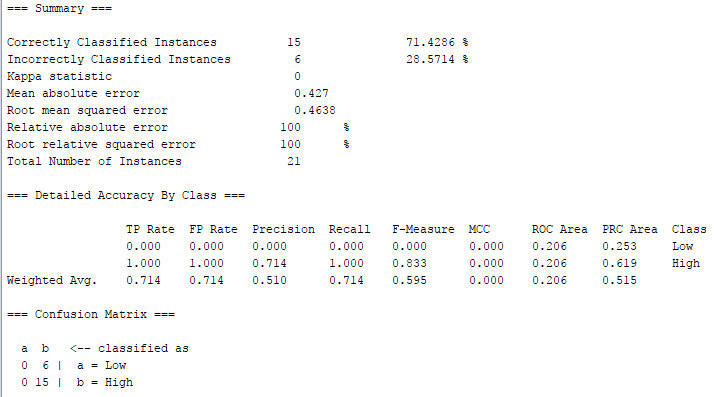
The features used are – 'Weather Events'

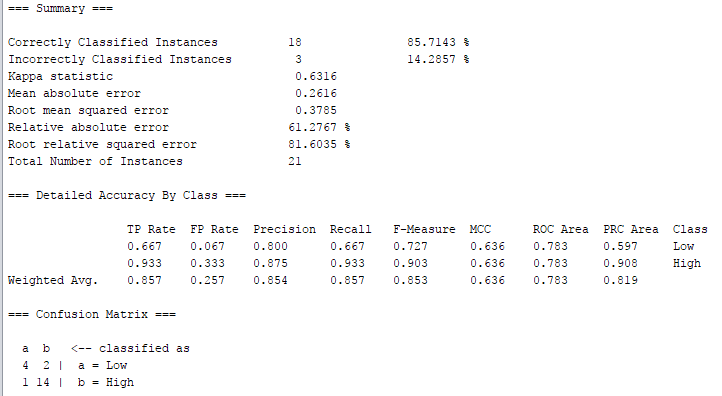


Only providing ‘Weather Events’ gives the least negative values of accuracy.

# Changing the Target variable type (Extra Work)

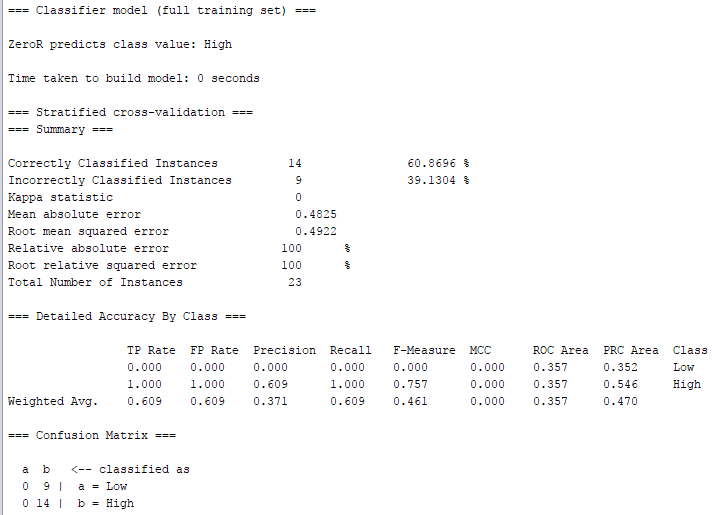
If we change the type of target variable “Number of Tweets” to only high and low, it becomes a classification problem and one can use many other machine learning algorithms available such as ZeroR, OneR, NaiveBayes.

On curiosity, I took the average number of tweets for the topic “**Happy**”, location **California** and changed the values of number of tweets to “High” if the value is higher than average value or else “Low”. Through Weka software, I applied ZeroR algorithm with 10-fold cross validation test options to check how classification accuracy.

The ZeroR could correctly classify 71% of the instances. Using this as baseline, I tried other algorithms and NaiveBayes algorithm came back with the best accuracy results.

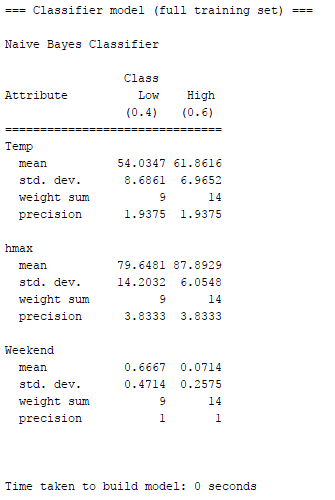
Naïve Bayes correctly classified 85% of the instances. The features used for the above 2 algorithms are - Weather Events, Weekend, number\_tweets

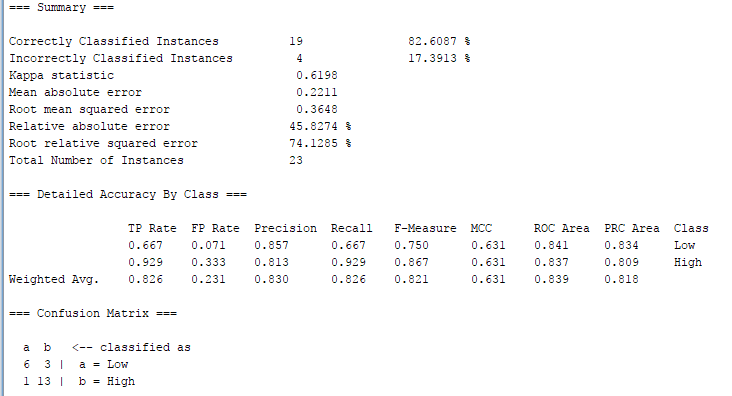
For the topic **Happy**, location **New Jersey**, applying the same algorithms ZeroR, NaïveBayes resulted in similar accuracy levels. The features used are ‘Temp”, ‘hmax’,’weekend’.

ZeroR along with confusion matrix–

ZeroR algorithm for **Happy** topic and location **New Jersey** gives 60% accuracy in classifying the models.

NaiveBayes –

The model built by Naïve Bayes is –

The actual summary and results of NaiveBayes model along with confusion matrix–

As we can see, the accuracy of NaïveBayes is 82% in classifying the target variable.

# Conclusion

For the data collected from Twitter using specific keywords, we can conclude that there is no correlation with the weather data. The data could have been better if not Easter holiday in between because of which a surge of tweets are seen in case of ‘Happy’ topic. Since there is no good correlation between the weather data and the tweet data, the notion that we are happy and do many actions during pleasant weather cannot be concluded.

Only the topic Workout showed positive mean accuracy values, when “Weekend” feature was used. This could mean that the number of people going out to gym or working out depends on the day of the week. They probably must be resting on weekends.

When the target variable is changed from numerical to nominal, it becomes easier for algorithms to build a good model that can be used for prediction,

There is a lot of future scope on the project. With more data, one can find correlations with moods and actions of people. Since, the data we had was fairly limited to two weeks, the change in temperature or humidity is not so much as when you compare between various seasons.

# References

Code References -

* <https://chrisalbon.com/machine-learning/linear_regression_scikit-learn.html>
* <http://blog.coderscrowd.com/twitter-hashtag-data-analysis-with-python/>
* <http://machinelearningmastery.com/evaluate-performance-machine-learning-algorithms-python-using-resampling/>