Problem 4: Statistical analysis to support new data product

Problem Statement: In this problem we will further explore the RealDirect business. You should realize by now RealDirect has built a business around existing real estate (buying and selling) business by creatively repurposing the data and building a ***data product*** around it. They have created a web (and mobile) portal with tools to facilitate real estate related operations. Assume you have been hired by ReaDirect to extend the line of product offerings. You put on your thinking cap and realize that NY is a prime location for ***apartment rental*** since buying real estate (houses and apartments) is beyond their means for many. You also realize that many prospective clients take to twitter when they need something and want to express their sentiments and status. You plan to recommend to the executive team at RealDirect that they should offer apartment rental as a product. You want to arm yourself with data to prove your recommendation. You plan to collect twitter data about apartment rental and real estate (buying a house) for a week on a daily basis and show the feasibility of your recommendation with this statistical analysis. Also provide a pricing model (ex: subscription or one time registration etc.). Prepare the tar file as suggested in Problems 3 and 4 and submit the RDExP4UserName.tar

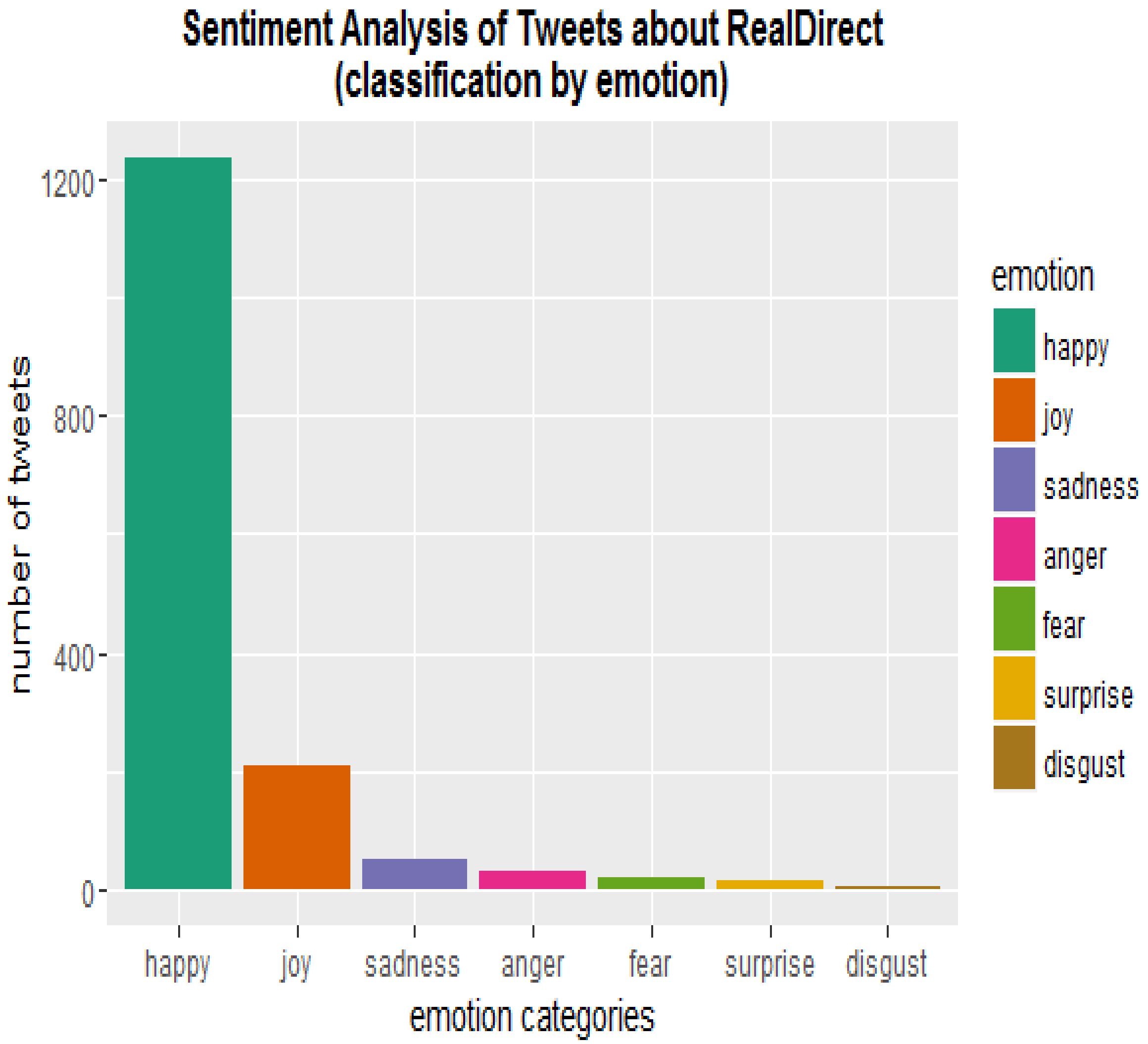
In this problem the tweets are collected in the same method as in the 1st problem. That is using the "**twitteR**" API package. This package has many features that can be easily applied on twitter data like sentimental analysis which was done for this problem. The twitter data collected was mainly on the topic **"apartment rentals in NYC**".

Since sentimental analysis was carried out on the twitter data, apart from the actual tweets, no other information was required. Hence all other unnecessary data was cleaned out. The sample of the cleaned out data is in the sample_tweet.json file.

On these tweets data, Sentimental analysis was carried out using the "Naive Bayes classifier algorithm". It is a machine learning approach for detection of sentiment and text classification. In our problem, the data was classified once based on its emotions and once based on its polarity.

The classification based on emotion segregates the tweets based on emotions like happy, joy, sadness, anger, fear, surprise and disgust. The graph which was plotted based on this emotional classification of tweets, tell us how many tweets are supporting the apartment rentals in NYC, if they are happy with the rates, location and other such stuffs.

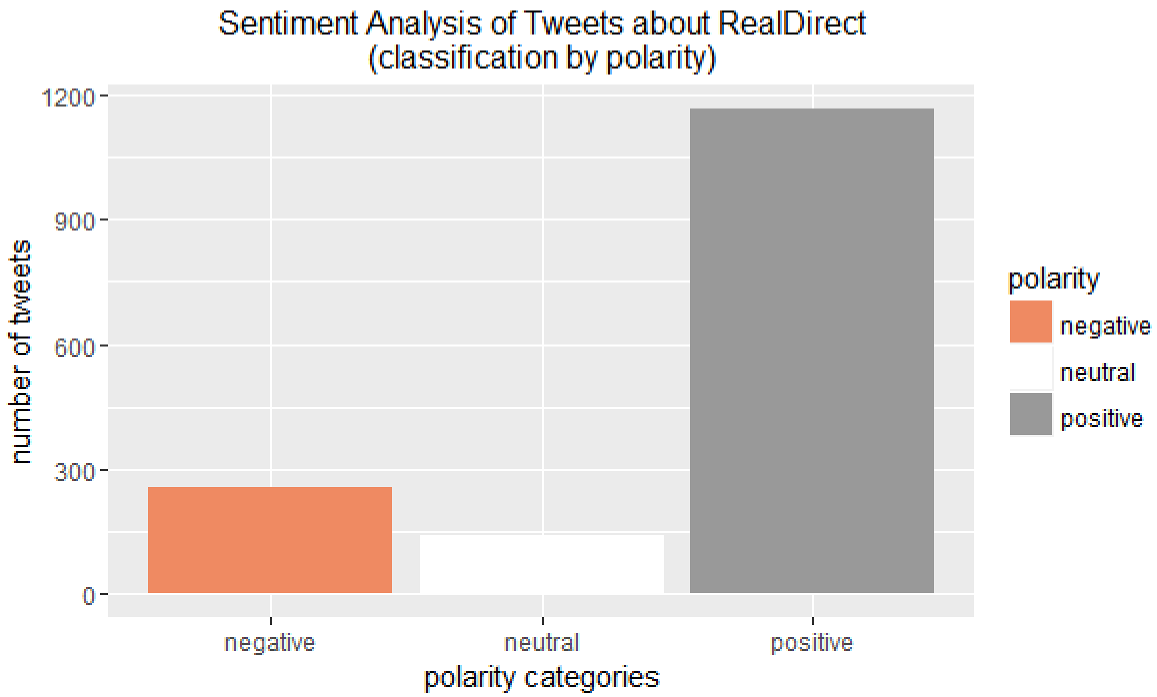
The following graph is based on the emotional classification of the tweets.



This graph shows the number of tweets that fall in various emotion categories. Based on this graph we can see that a majority of tweets are in the "happy" category. This tells us that user the month the real estate is on a rise. More people might be investing in buying house and selling them.

The classification based on the polarity segregates the tweets based in to either positive or negative or neutral. The polarity classification can help us identify the overall mindset of the people in NYC involved in some form of real estate business, either be it a broker or seller or a buyer.

The next graph is based on the polarity classification of the tweets.



In this graph the number of tweets are Classified into 3 categories- negative, neutral and positive. From the above graph, it can be infered that more number of tweets are in the positive category. This again confirms our inference from the previous graph that real estate is on a rise. There is a lot of activity and talk abut buying, selling and renting of houses.

After observing the data which was of only a month's time, if we collect data over a long period of time, end compare the maximum emotion category and high polarity category of all the months, then it would be more useful. From such a data, we can infer which month sees more activity and in which month the real estate is on a rise. This can help the real estate company in deciding the rates of the houses, the areas to invest in for future development or selling etc..