Emmalee Dolfi
Capstone Project - Final Report

What brings people to parks? Using social media, amenity availability and park accessibility to predict park visitation rates

Why do people visit parks? How do these visitors choose one park over the other? For this project, I intend to identify what amenities (e.g. basketball courts, walking trails), facilities (e.g. restrooms), and/or general park attributes (e.g. park size, park accessibility) lead to higher visitation rates (i.e., higher number of tweets) in parks. Answering this questions could help answer the difficult question of -- what makes a park a “high quality” park?

For this specific study, I collected social media and park amenity/facility data in New York, NY, and thus, the anticipated client in this case would be the NYC Parks Department. However, the client could be any agency responsible for park acquisition, park design, park creation and/or park maintenance. These agencies could use the conclusions from this study in their future park planning efforts, to

Data Wrangling

https://github.com/edolfi/springboard_datascience/blob/master/Capstone/codes/ParkFeatures_DataWrangling.ipynb

How did I wrangle the data?

1. The parks dataset I collected from the NYC parks dept had thousands of park polygons in it, each with a unique identifier. This same identifier was found in all other datasets I collected from the NYC open data portal: (add link!). I chose to include the following amenities in my analysis:
    - Handball
    - Tennis
    - Basketball courts
    - Softball fields
    - Running tracks
    - Play areas
    - Preserves
    - Spray parks/spray areas

    I first got a count of each of these amenities in each park, as well as a combined total for each park across all of the amenities. I then merged these amenity counts with the park polygons using the common identifier into a single dataset.

I chose to do a left join for these merge, so I could keep all parks, whether or not they had amenities inside of them.

2. The 10-minute walk statistics dataset also had this common identifier. I merged this data with the park polygons + amenities data, but used an inner join instead so that all park would have 10-minute walk statistics. This left me with 1,705 park polygons.
3. I then took a random sample of 100 park polygons for which to collect tweets. The geolocated tweets were provided (anonymously) in JSON format. I performed a spatial join on the geolocated tweets with the park polygons to tag each tweet with the park it was geolocated within. The code for the random sampling can be found here:

   https://github.com/edolfi/springboard_datascience/blob/master/Capstone/codes/Dolfi_Capstone_RandomSampling.ipynb

   The park tagging was done in ESRI's ArcMap software, and thus not written in python (although it could've been!).

What other data sources could I have used?

- Public transportation - bus stop locations, subway stations, etc.
- More detailed set of amenities
- Neighborhood crime statistics or occurrences
- Recreation rental facilities / programming within parks

Initial Findings

https://github.com/edolfi/springboard_datascience/blob/master/Capstone/codes/ParkFeatures_DataStorytelling.ipynb

https://github.com/edolfi/springboard_datascience/blob/master/Capstone/codes/NYC_ParkFeatures_Tweets_InferentialStats.ipynb

Summary of patterns within park data (no tweets):

- Amenity count does not appear to be strongly correlated with park size. There are many small parks with a high amenity count.
- Manhattan has the lowest median # of park amenities and Queens has the highest.
- Staten Island tends to have larger parks on average than the other boroughs. Manhattan parks are relatively small.
- There does not seems to be a strong correlation between # of people served and park size, or # of people served and amenity count.

- Handball courts are the most popular amenities in NYC parks. Brooklyn has the highest number of handball courts and play areas/playgrounds. Staten Island has the highest number of preserves in their park system.

Summary of patterns within park data and tweets:

- The number of tweets within a park boundary is not strongly linearly correlated with park size, total number of park amenities, or number of people living within a 10-minute walk of a park.
- I did not find any significant difference in means for small vs. large parks and # of tweets, parks with many amenities vs parks with few amenities and # of tweets, or parks with many people living within a 10-minute walk vs. parks with few people living within a 10-minute walk and # of tweets.

In-depth analysis using Machine Learning

https://github.com/edolfi/springboard_datascience/blob/master/Capstone/codes/NYC_ParkVisitationPrediction.ipynb

Choice of ML algorithm

I chose to use a Random Forest regressor and a Random Forest classifier as my algorithms for a few reasons. The first reason is because I was interested in building a regression model as well as a classification model. Also, the types and ranges of my variables are all over the place, and scaling is not necessary for Random Forest. Additionally, Random Forests provide an easy way to evaluate feature importance, and since I am using a relatively large number of variables, this was something I wanted to easily explore. Lastly, with the number of samples in my dataset, the processing speed (i.e. a major limitation of Random Forest algorithms) would not be a factor.

Summary of steps

- Random Forest regression model: I ran a random forest regression model using all variables in my first run, and then performed a grid search to determine the best-performing parameters. The latter results in the highest r-squared values on both the training and test data:

  ```
  Train R-squared ::  0.8653115296513108
  Test R-squared  ::  0.25302542372881365
  ```

  - Although these were the highest r-squared values, the model performs well on our training data and not on the test data, which means it is overfit. These results led me to try out a random forest classifier…

- Random forest classification model: I ran several random forest classification models on the data. The first run included all variables, the second run removed the dummy/categorical variables along with the total # of amenities (since all the other amenities were included), and the third used only park size, total number of amenities, and the total population living within a 10-minute walk a that park.
    - Results from 1st run: the hyper-parameterized version of this variable set resulted in the highest classification accuracy.

      ```
      Train Accuracy ::  0.9871794871794872
      Test Accuracy  ::  0.65
      ```

      Park size (acres) and white population living within a 10-minute walk of a park werethe two most important features driving our model results, though neither of them are exceptionally great.

    - Results from 2nd run: the hyper-parameterized version of this variable set also resulted in the highest classification accuracy on the test data, yet performed slightly better on the training data than the previous model. However, that doesn’t mean much in this case, as the model is overfit in both cases.

      ```
      Train Accuracy ::  1.0
      Test Accuracy  ::  0.65
      ```

      Park size (acres) was once again the most important feature in the model, and the white population within a 10-minute walk of the park and the # of high income households living within a 10-minute walk of the park is were #2 and #3. Again, none of these variables had too strong on an importance.

    - Results from 3rd run: the hyper-parameterized version of this small set of variables (park size, total number of amenities, and total population living within a 10-minute walk) was the worst-performing of the three models.

      ```
      Train Accuracy ::  0.9871794871794872
      Test Accuracy  ::  0.45
      ```

Conclusions/Limitations

We can conclude from the above model runs that park size, total number of amenities, and the total population living within a 10-minute walk are not great predictors of the number of tweets (and possibly park visitation) for NYC parks when they are the only variables used in the prediction model. Park size, however, was the “most important” variable compared to all others in each model run above, yet the relative importance was not very high in the grand scheme (~.10).

The highest testing accuracy of the above models was 65%. Although better than random chance, this result is not dependable enough to be a true predictor of the number (or class) or tweets within a park boundary in NYC. Each of the variables used in the datasets above were descriptive of what is inside the park, about the physical characteristics park itself, or the populations that reside immediately surrounding the park boundary. These results are consistent with what was found in Hamstead et al. (2018; https://doi.org/10.1016/j.compenvurbsys.2018.01.007). These factors alone are not reliable indicators of park usage, and thus more variables should be considered in future analyses. Some examples of additional variables that could be considered are:

- Public transportation near the park - bus stop locations, subway stations, etc.
- More detailed set of amenities
- Neighborhood crime statistics or occurrences
- Recreation rental facilities / programming within parks

An assumption of this analysis that could be called into question is: is the number of tweets within a park truly representative of park visitation? To answer this question, one would need to collect park visitation data in the same parks for which he or she is gathering twitter data, and ensure the two trends are highly correlated. One must also consider whether or not counting tweets is representative the population using the park. For example, low-income populations may be less likely to carry a device that allows them to tweet outside of their home, if at all. Similarly, children are much less likely to carry this type of device, or use social media, yet are often heavy users of a park.

In future analyses, it would be interesting to perform some qualitative analysis to measure local perception of park quality - how do the users of the park feel about it? Is it clean? Is it safe? Is it too crowded? One could then look into whether or not a “high quality” park is correlated with high number of tweets? This could help clear up the question of whether or not tweets are a good indicator of park visitation. Lastly, it would be interesting to flip the model around and see if physical characteristics of a park, such as the park type (e.g., community, neighborhood, pocket, regional) or park size, can be predicted using the number of tweets. This is an interesting follow-up research question to investigate!

Online resources:

https://medium.com/rants-on-machine-learning/the-unreasonable-effectiveness-of-random-forests-f33c3ce28883

https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

Data sources:

The Trust for Public Land

Twitter

NYC Parks and Recreation Department