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) is 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 then 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.

Next Steps

- Look for stronger correlations/more obvious relationships amongst all variables in the dataset vs. only park size, number of amenities and number of people living within a 10-minute walk. There could likely be a relationship I'm not expecting that could improve the predictive power of my model.
- Build predictive model that predicts the number of tweets inside a park boundary based on chosen variables

Initial Conclusions/Limitations

- Need more data about the areas surrounding a park. Knowing what's inside isn't enough.
- Are tweets a good indicator of park visitation? Could there be any bias in choosing tweets as the indicator?