CAPSTONE PROJECT

Applied Data Science Capstone

JOSHUA MONTGOMERY

Table of Contents

PARKS AND RECREATION

Section 1: Introduction

Often when people think about data science, they imagine modern companies building profiles and recommendation engines using ‘big data’. These people certainly aren’t wrong, but the concept of using practical data to derive models which predict future demand has been around for a long time before big data. A great example of this demand planning would be aircraft companies, where it’s not uncommon for the design of a new aircraft to take over 6 years. This means that they must design these aircraft for a future market and hence must be able to predict the future trend in the aircraft industry, and then design an aircraft that will have a niche in the future market. Almost all companies will have some level of demand planning, whether it is a baker predicting demand variation of the week or a trans-national corporation predicting how many delivery drivers need to be trained to meet future demand.

Therefore, as my capstone project I decided to analyse parks in London, and then build a model which could predict visitor numbers for a planned park. To do this, the foursquare API will be used to gather location data of parks in London and then combined with data on size of a park and visitor numbers. This data will then be used to train a multiple linear regression model for the expected number of visitors per annum. Finally, the predictions will be analysed to determine the accuracy of the model

Section 2: Method

The first part of this project was to gather data for the location of parks in London, using the Foursquare API. To do this a query was submitted to find any place with park in the name or category within a certain radius of a centrally chosen location. In this case Buckingham Palace was assumed to be a reasonable approximation for the centre of London and the search radius was set to 10km. Due to the size of the search radius the centre wasn’t massively significant, as shown in figure 3.2, and the centre could have been moved and still suggest the same number of parks. The data then had to be cleaned to eliminate rows that contained irrelevant data, and any columns that weren’t going to be used in making the model. Using the folium software package these data points were plotted over a map of London as figure 3.3.

Next the data for size of the parks and number of visitors had to be gathered, unfortunately there were no tables online that contained the relevant data, hence it had to be manually collected. This was relatively easy for the Royal Parks and those still run by the local council as they had readily available reports on each parks performances, as they are legally required to do, However a few of the parks are now privately operated and hence are under no obligation to release such figures, hence greatly complicating finding accurate and current data. Further discussion of this is outside the scope of a methodology and will be analysed further in the discussion.

Once the manually collected data had been uploaded to the notebook, both data frames were indexed by the name of the parks and then a join was performed on the two tables. This data was then further split into test and train data. A multilinear regression fit was then on training data and used to make predictions on the test data. The success of these predictions was then evaluated and finally predictions were made for three hypothetical parks

Section 3: Results

After the foursquare API had been run the data displayed here as figure 3.1 had been generated. Figure 3.1 has several irrelevent columns cropped out, to make iteasier to read.

	name	categories	address	cc	city	country	crossStreet	distance	formattedAddress	labeledLatLngs	lat	lng	neighborhood	postalCode
0	Hyde Park	Park	Serpentine Rd	GB	London	United Kingdom	NaN	1587	[Serpentine Rd, London, Greater London, W2 2TP...	[{'label': 'display', 'lat': 51.50778087767913...	51.507781	-0.162392	NaN	W2 2TP
1	St James's Park	Park	The Mall	GB	London	United Kingdom	Horse Guards Rd	650	[The Mall (Horse Guards Rd), London, Greater L...	[{'label': 'display', 'lat': 51.50325316049429...	51.503253	-0.132995	NaN	SW1A 2BJ
2	Green Park	Park	Piccadilly	GB	London	United Kingdom	Constitution Hill	385	[Piccadilly (Constitution Hill), London, Great...	[{'label': 'display', 'lat': 51.50465559886703...	51.504656	-0.143788	NaN	SW1A 1BW
3	Battersea Park	Park	Albert Bridge Rd	GB	Battersea	United Kingdom	NaN	2651	[Albert Bridge Rd, Battersea, Greater London, ...	[{'label': 'display', 'lat': 51.47951201381755...	51.479512	-0.156984	NaN	SW11 4NJ
4	Regent's Park	Park	Chester Rd	GB	London	United Kingdom	NaN	3339	[Chester Rd, London, Greater London, NW1 4NR, ...	[{'label': 'display', 'lat': 51.53047945949403...	51.530479	-0.153766	NaN	NW1 4NR
5	Green Park London Underground Station	Metro Station	Piccadilly	GB	London	United Kingdom	at Stratton St	595	[Piccadilly (at Stratton St), London, Greater ...	[{'label': 'display', 'lat': 51.5067341345345,...	51.506734	-0.142630	NaN	W1J 9DZ
6	St James's Park Lake	Lake	Horse Guards Rd	GB	London	United Kingdom	NaN	631	[Horse Guards Rd, London, SW1A 2BJ, United Kin...	[{'label': 'display', 'lat': 51.50270552998373...	51.502706	-0.133038	NaN	SW1A 2BJ
7	St. James's Park London Underground Station	Metro Station	Petty France	GB	London	United Kingdom	NaN	566	[Petty France, London, Greater London, SW1H 0B...	[{'label': 'display', 'lat': 51.4997101149314,...	51.499710	-0.134187	NaN	SW1H 0BD
8	Victoria Park	Park	Grove Rd	GB	London	United Kingdom	NaN	8460	[Grove Rd, London, Greater London, E3 5TB, Uni...	[{'label': 'display', 'lat': 51.53849910020006...	51.538499	-0.035290	Old Ford	E3 5TB
9	Hyde Park Corner Bus Stop F	Bus Stop	NaN	GB	London	United Kingdom	NaN	615	[London, Greater London, SW1W 0QH, United King...	[{'label': 'display', 'lat': 51.502087, 'lng':...	51.502087	-0.150721	Green Park	SW1W 0QH

Figure 3.1- The data output from the foursquare API displayed as pandas dataframe.

Upon reading the name column it is clear that the search query also returned data on any locations in the search area with 'park' in the title. The location of Underground stations and bus stops is not relevent to this report, hence the table was further filtered to only keep data that has the categories label of 'Park'. Most of the columns are not also unneeded, hence a new data frame was formed with only the 'name', 'distance', 'lat' , 'lng' and 'postalCode' columns, and was then displayed as figure 3.2.

	distance	lat	lng	postalCode
name				
Hyde Park	1587	51.507781	-0.162392	W2 2TP
St James Park	650	51.503253	-0.132995	SW1A 2BJ
Green Park	385	51.504656	-0.143788	SW1A 1BW
Battersea Park	2651	51.479512	-0.156984	SW11 4NJ
Regent Park	3339	51.530479	-0.153766	NW1 4NR
Victoria Park	8460	51.538499	-0.035290	E3 5TB
Holland Park	4318	51.503148	-0.204153	W14
Clissold Park	7639	51.561438	-0.088457	N16 9HJ
Finsbury Park	8179	51.570321	-0.100937	N 4 2
Greenwich Park	10245	51.477521	0.000858	SE10 9NF
Queen Elizabeth Olympic Park	9926	51.540296	-0.012938	E20 2ST
Richmond Park	11542	51.438905	-0.274728	TW10 5HS
Brockwell Park	6142	51.450931	-0.106065	SE24 0PA
Archbishop Park	1792	51.497800	-0.116692	SE1 7LE

Figure 3.2- The refined data for London parks, that was origionally output by the foursquare API.

Trying to plot the parks data using folium, proved to be difficult, as the folium plugin couldn't handle data that had apostrophes in, as they cause a HTML parsing error. Hence all apostrophes were removed form the data, an example being St James's park going to St James Park. After this change the folium software generated the map of London seen in figure 3.3

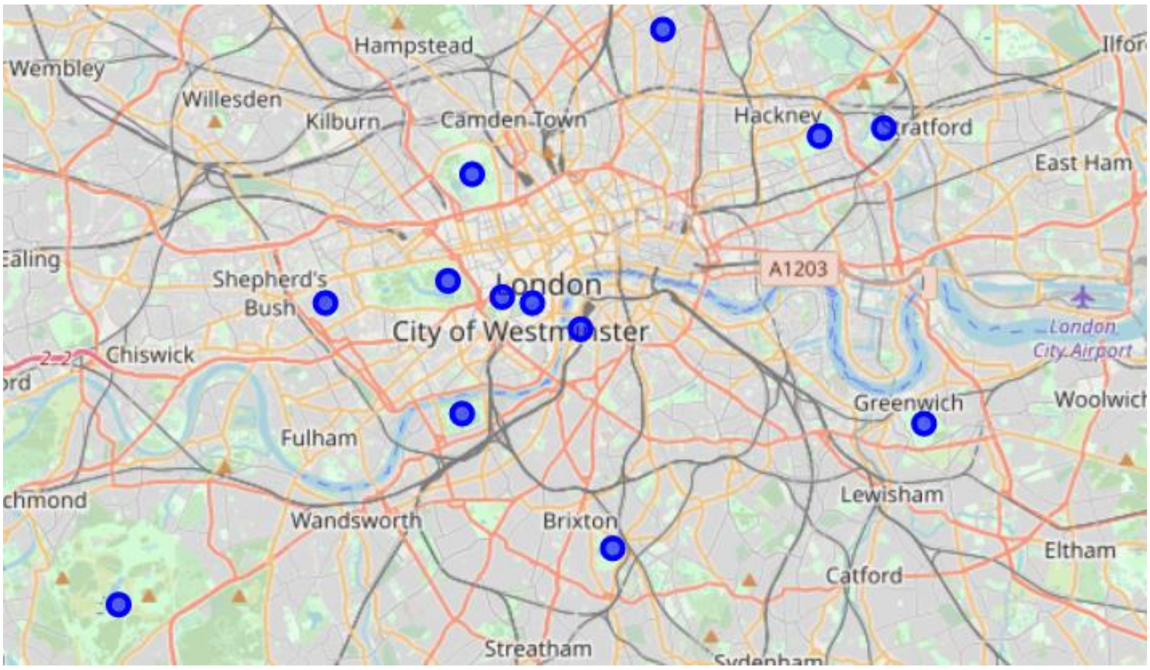

Figure 3.3- Map of London with all parks within 10km of Buckingham Palace labeled.

The data for visitors and area was then joined with the foursquare data, and the reuslting data frame is displayed here as figure 3.4.

name	distance	lat	lng	postalCode	visitors	size
Hyde Park	1587	51.507781	-0.162392	W2 2TP	10.30	350.0
St James Park	650	51.503253	-0.132995	SW1A 2BJ	13.00	58.0
Green Park	385	51.504656	-0.143788	SW1A 1BW	10.90	47.0
Battersea Park	2651	51.479512	-0.156984	SW11 4NJ	3.00	200.0
Regent Park	3339	51.530479	-0.153766	NW1 4NR	6.70	410.0
Victoria Park	8460	51.538499	-0.035290	E3 5TB	9.00	213.0
Holland Park	4318	51.503148	-0.204153	W14	5.26	54.0
Clissold Park	7639	51.561438	-0.088457	N16 9HJ	3.00	55.8
Finsbury Park	8179	51.570321	-0.100937	N 4 2	1.50	110.0
Greenwich Park	10245	51.477521	0.000858	SE10 9NF	3.90	180.0
Queen Elizabeth Olympic Park	9926	51.540296	-0.012938	E20 2ST	6.00	560.0
Richmond Park	11542	51.438905	-0.274728	TW10 5HS	4.40	2500.0
Archbishop Park	1792	51.497800	-0.116692	SE1 7LE	0.30	9.7

Figure 3.4- The data used to train and test the model. Visitors is million per year and size is in acres.

The model was trained using 80% of the data in figure 3.4, with the rest resereved for testing. Next, the fit was then evaluated with the test data and found to have a residual sum of squares of 6.8 and a variance score of 0.03

Finally, the model was used for what it was intended for, and generated data for three hypothetical parks. The parameters input and the output visitor predications is displayed in figure 3.5. Again, distance is distance from Buckingham Palace.

Name	Size (Acres)	Distance (m)	Predicted Visitors Per Year (million)
Buckingham Palace	39	0	8,81
Small Park	10	100	8.72
Large Park	800	10000	4.34

Figure 3.5- Proposed parks and projected visitor numbers.

Section 4: Discussion

The variance score of only 0.03 is excellent, when considering the small scope of this project, and suggests that the average is very close to predicting the actual value. However, the residual sum of squares of 6.8 suggests a substantial lack of accuracy on more extreme values. This model hence proves to be useful in suggesting trends, but the user of the script shouldn't exact the predicted values to exactly align with future data.

From the model as distance from the centre of London increases, the visitor count decrease, hence they can be said to have a negatively proportional relationship. There is a clear positive relationship between size of the park and number of visitors. However, of the three proposed parks, the two smallest (but most centrally located) parks were predicted the visitors compared to the large park. Suggesting that the most important feature in the success of the park is its location.

The largest source of error is most likely to be the values for visitors per year, as there were substantial problems with the collection of that data. As mentioned earlier, in section 2, there were very few published figures for the privately run parks, such as Battersea park, meaning that some of the visitor counts were out of date. All of the data for the royal parks and the council owned parks was for the economic year of 2018-2019, however this was not the case for the privately run parks, with a notable outlier being Battersea park were the only data available was for 2013, hence introducing a substantial amount of error into the model.

If the experiment were to be repeated, more independent variables would produce a more accurate (and more precise) model. For example, if data could be collected for yearly monetary investment or crime rates that would lead to a much more complete picture of parks in London. Unfortunately, this data is not publicly available for all the parks and retrieving it would likely require several freedom of information requests. As this model is purely to asses' technical abilities, that level of data collection would be inappropriate.

Section 5: Conclusion

The model clearly shows that parks that are closer to the city centre get more visitors, and so do larger parks. However, as it is not feasible to achieve both of these things on a finite budget, proximity to the city centre should be prioritised. If greater resources were available, it would benefit the analysis to include more independent variables. The model is limited in that all the parks analysed were to free enter, and it not clear if the parks were considered profitable by the owners. If a similar report were to be created but with profit as the dependent variable it would likely be more useful commercially.