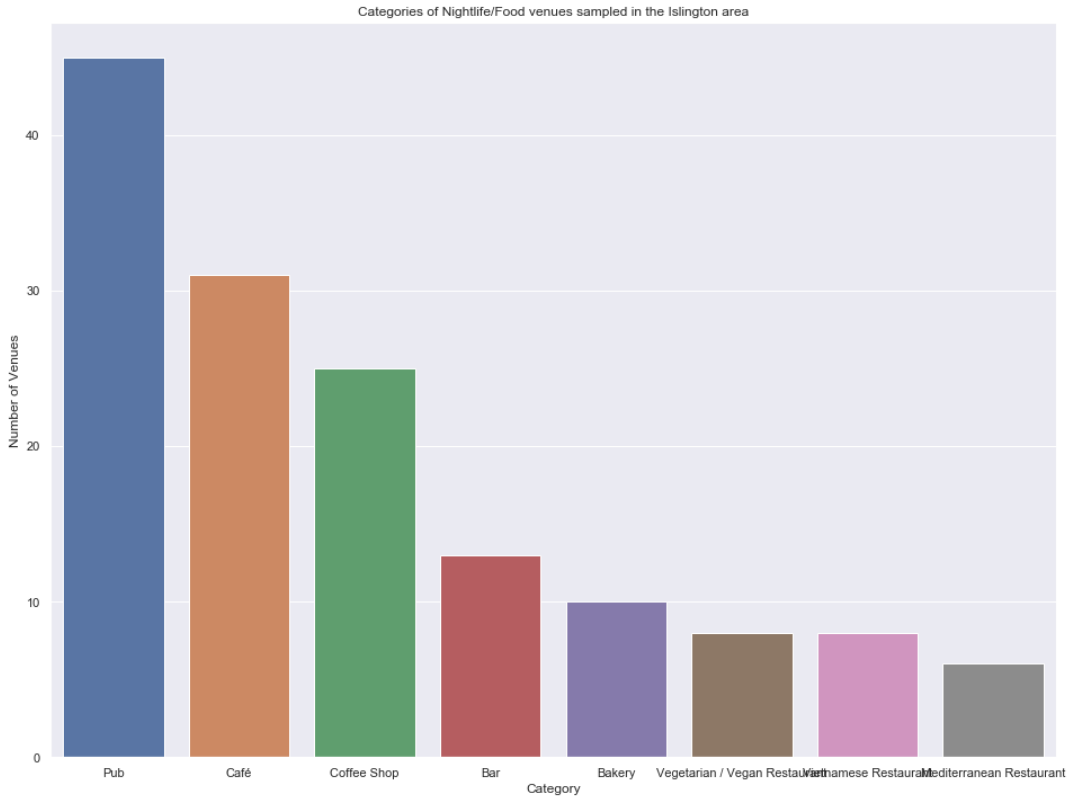
Coursera Data Science Capstone – Islington Pubs

Islington Pubs: The Problem

- My friends want to open a pub in the Islington area of London.
- They don’t yet know when it will be popular, so they don’t know when to staff it or run promotions.
  - Running promotions at the wrong time can hurt revenue; and overstaffing can hurt profitability; which is a big problem when London rents are so high.
- Can we use data science to help?

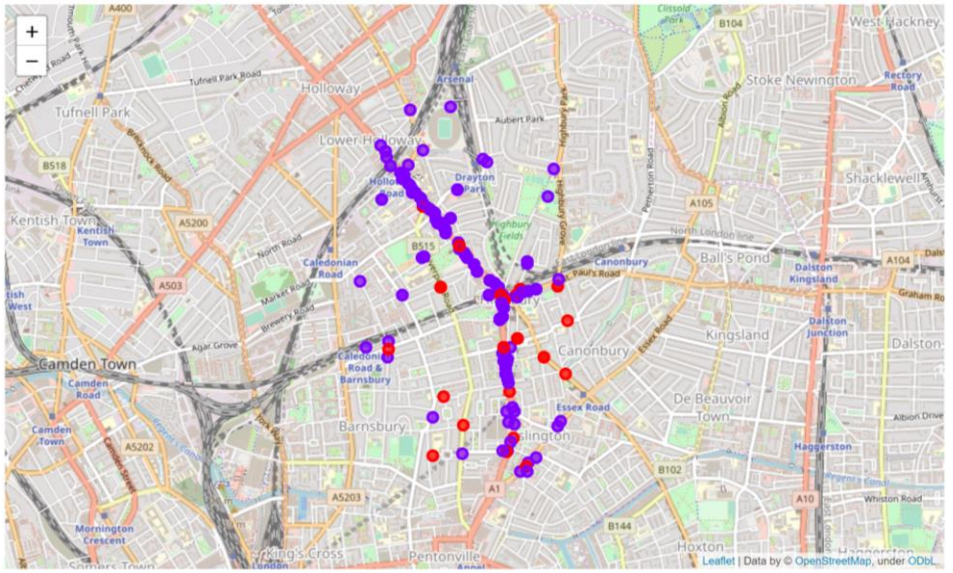
The Islington Area: a venue guide:

- I used the Foursquare API to obtain details on 250 food/nightlife venues in the local area
- The largest category of venue is pubs: there are more of them than coffee shops! This makes my investigation even more important.



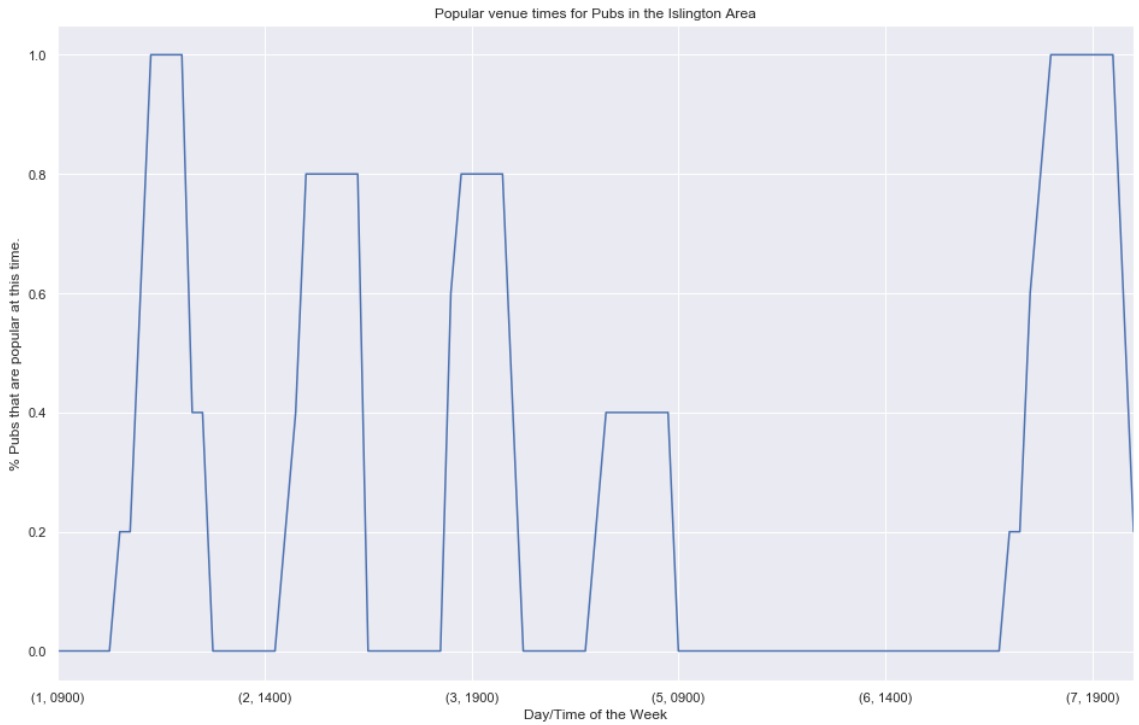
Islington Pubs: Where are they?

- The map on the right shows venues sampled in the area: there is an even distribution of pubs (red) among the other types of venues (purple)
- As a consequence it’s unlikely that our friends will find a location that there are no pubs in order to gain a competitive advantage. As a result staffing hours are very important.



Popular Times

- I then used the Foursquare API to gather data on when venues are open and when they are popular, for a sample of 31 venues (due to API limits).
- The graph on the right shows the proportion of the pubs in the area that were ‘popular’/not according to the API by time.
- Clearly there are peaks and troughs, so it’s worth exploring the times that are most and least popular for pubs.

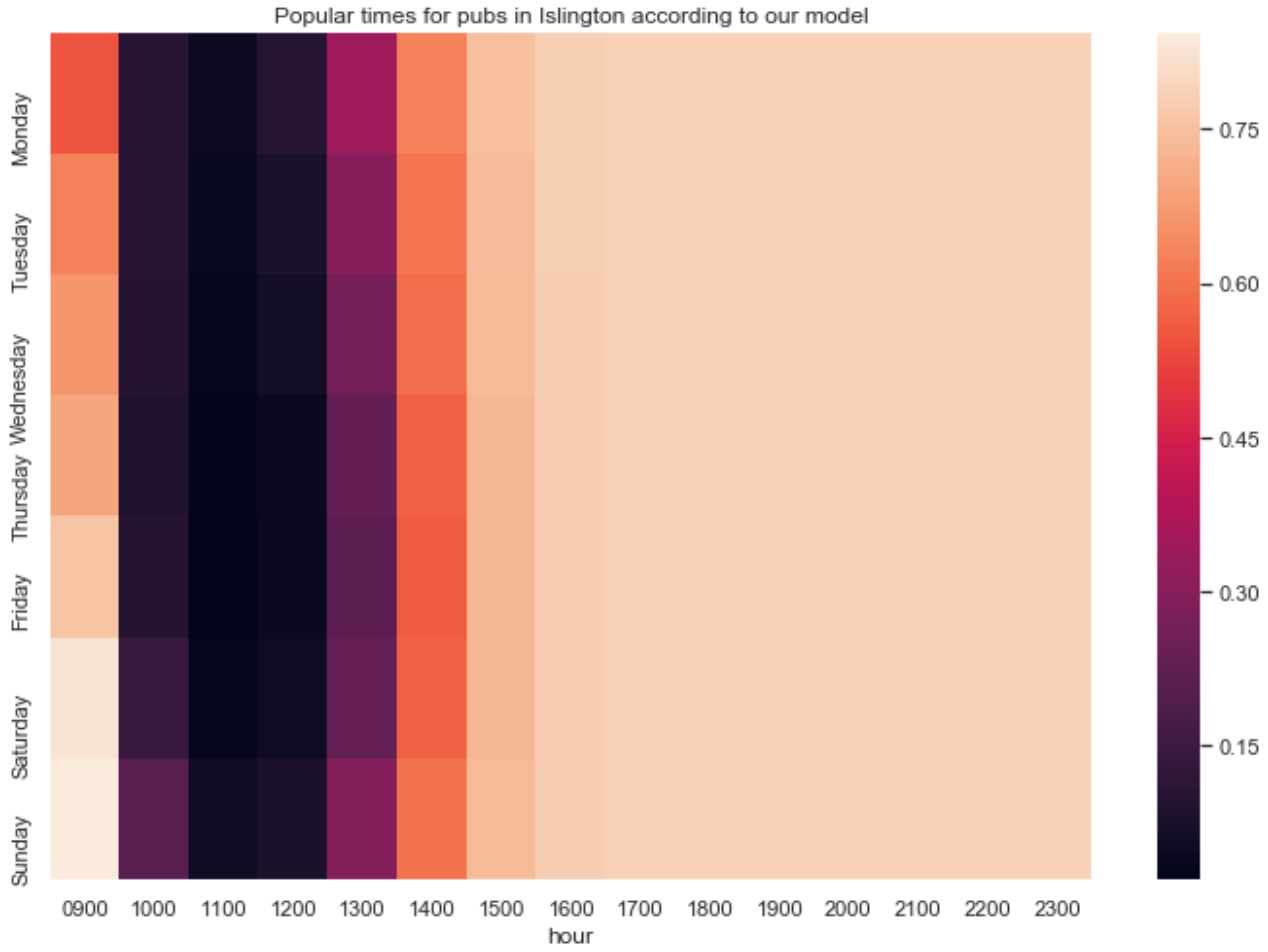


Model development:

- I built rows for every venue at every time from 9AM to 11PM marking the venue as popular or not, and extracted features for the type of venue.
- This resulted in a 2.2k row sample with 7 input features, modelling the binary “is_popular” feature. I used a training sample of 80% of the data, and reserved 20% for testing.
- I tried building models with Decision Trees, Logistic Regression, and Support Vector Classifiers to find the best model.
- The best model, a SVC with c=3.0, had an ***84.3% f1-score.***
- I used then used this to predict against dummy data for all days and hours of the week, to see when the model thought a pub would be most popular.

Model Predictions:

- The model predicts that pubs will be most popular in the evenings (17.00 onwards, increasing until 23.00) and from 15.00 on weekends.
- It also predicts that pubs will be popular at 09.00 on weekdays – though this may be because not many are open!



Conclusion:

- According to my model, my friends should look to fully staff their pub from 5PM onwards on weekdays, and from 3PM on weekends.
- While 9AM is also popular according to the model, it may not be practical to fully staff pubs during this time for the short period of popularity.
- If looking to extend the project, I would gather data on further venues in the area to increase the sample size. I would also want to gather data on the locations when pubs were popular for longest, in order to maximise the duration of their popularity.