MLA
ASSIGNEMENT 1

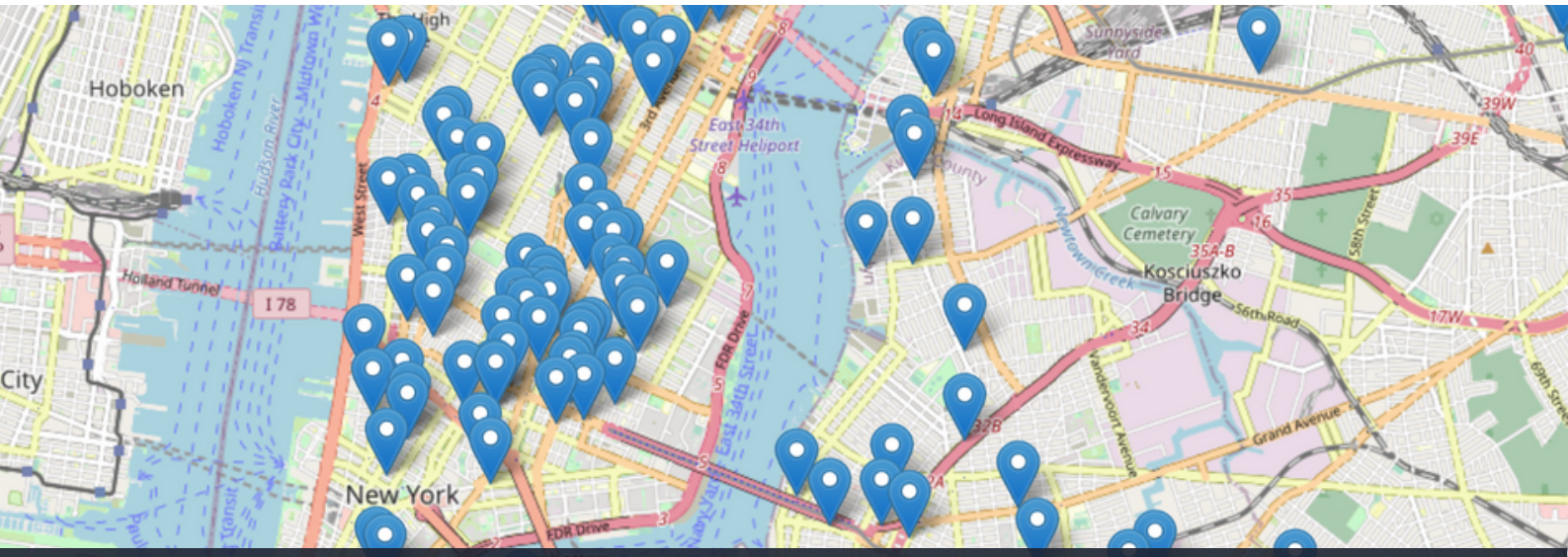
MSISS
21363700

MICHELIN ANALYSIS

An analysis into Michelin star Restaurants

ODHRAN RUSSELL

INTRODUCTION



We were tasked with the analysis of the dataset encapsulating various variables associated with Michelin and non-Michelin restaurants. The Michelin Guide is a prestigious accomplishment for any restaurant and classifies superior quality on all levels as rated by surveys on factors such as food, decor, price and service. We were tasked with two objectives:

- Can you reduce the dimension of the data so that its structure becomes easier to understand?
- Can you use clustering methods to identify meaningful structure in the data? What relationship, if any, does this structure have with a restaurant's inclusion in the Michelin guide?

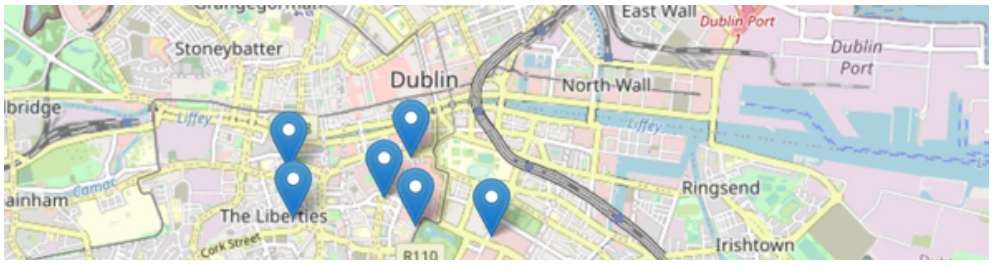
These sparked another question in my mind - What does it take for a restaurant to earn the coveted Michelin star?

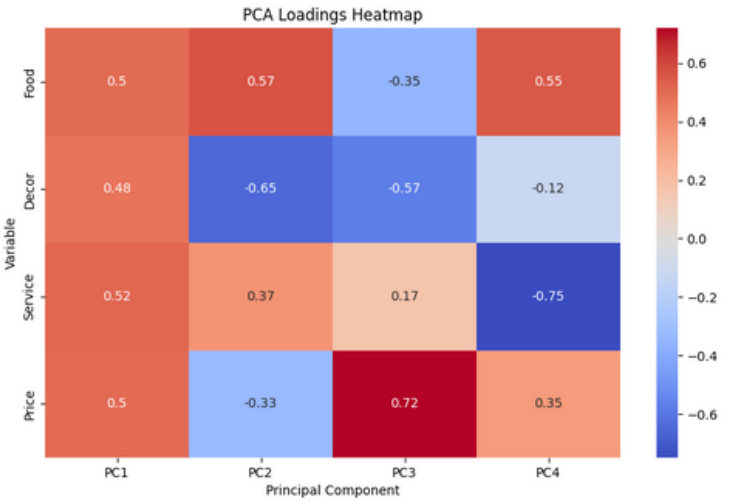
LET'S EXAMINE THE DATA

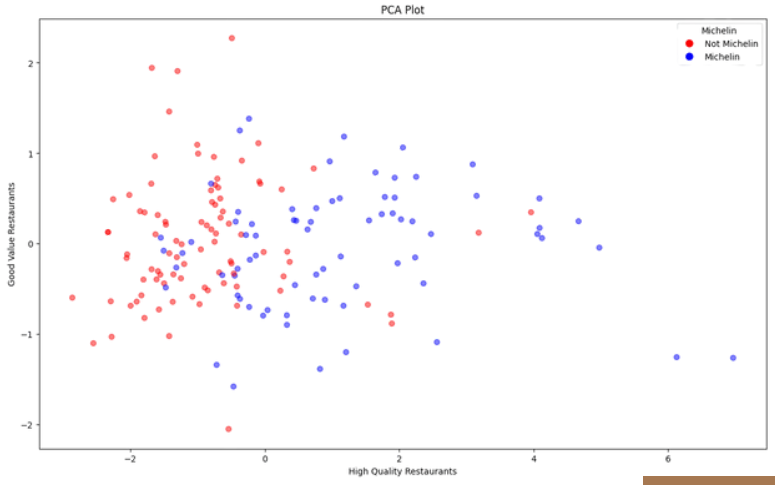
First things first - What exactly is a Michelin star ? A Michelin star is awarded to the very best of restaurants - with many receiving an increase in business after obtaining the star. According to the official guide it takes in 5 variables ~ " *the quality of the ingredients, the harmony of flavours, the mastery of techniques, the personality of the chef as expressed through their cuisine and, just as importantly, consistency both across the entire menu and over time.*" The dataset given had various columns including whether or not the given restaurant had a Michelin star, the name, food, decor, price, service ratings from Zagat surveys. They attempted to pseudo-represent the Michelin Guides' standards. Below is a snapshot of the data.

Statistic	Michelin Restaurants	Non-Michelin Restaurants	Total
Number of Restaurants	74	90	164
Average Food Rating	22.82	19.94	21.24
Average Decor Rating	21.42	17.31	19.16
Average Service Rating	21.34	18.36	19.7
Average Price	61.49	40.73	50.1

The dataset was fairly represented and contained no missing values and important factor before we begin our analysis. As expected, there were on average, higher values for Michelin star restaurants which would be a fair assumption to make. My goal was simple, I wanted to uncover the underling factors that influence Michelin star recipients. So lets address the lack of Michelin stars in Dublin vs New York City (pictured above) by examining the data in great detail.







Part 1

REDUCING THE DIMENSION OF THE DATA

One problem with the data is that there is plenty of dimensions (food, decor, price, service). How can we, maybe, group some of these features together to make this easier to visualise ? A technique called principal component analysis can help us achieve this. This works by extracting the main points from the data. Chances are the variables are not independent or to put it simply, a Michelin star restaurant won't score highly in just one dimension and non-Michelin star restaurants won't score poorly in just one dimension. The reality is - that a poor restaurant will have a lower price if the food is poor meaning we might be able to group these features together which is what we have done. Examining the heat map on the left - we can see what variables influence the 'grouped features PC1, PC2, PC3 etc. To preface, PC1 and PC2 capture a significant part of the variation in the data (88%!) meaning that PC3 and PC4 can be disregarded. But what is PC1 and PC2?

PC1

PC1 represents restaurants that have very high price, service, decor & food. These can be captured in one word - 'Good Quality Restaurants'

PC2

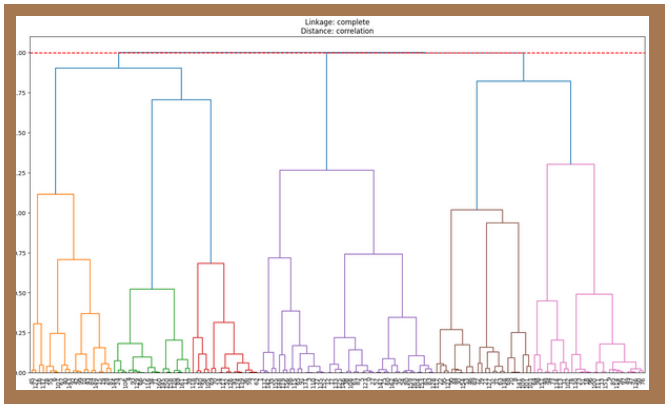
PC2 represents restaurants that have lower prices and poorer quality decor but still are great in service and food - "Good value restaurants

From the graph (above and right) we can see that there seems to be much more Michelin star restaurants in the 'Good quality restaurants' bracket - implying that exceptional all-round restaurants deserve the Michelin star

Part 2

CLUSTERING ANALYSIS

It is a bit of a let down to do all of that analysis to realise that good quality restaurants deserve a Michelin star. So we can turn to more advanced techniques which attempt to cluster the data by similar aspects. One technique we can use is called hierarchical clustering. This is a technique that groups datapoints based on different techniques such as furthest away, the most similar which can be defined by various metrics such as Manhattan (think of a city), euclidean (straight line) and many more. In my analysis, I wrote an algorithm that rewarded low entropy (so groups that had high proportions of Michelin vs non-Michelin) and even distribution (so we didn't end up with a group with one or two). The most optimal dendrogram is displayed below with distance: correlation and linkage: complete.

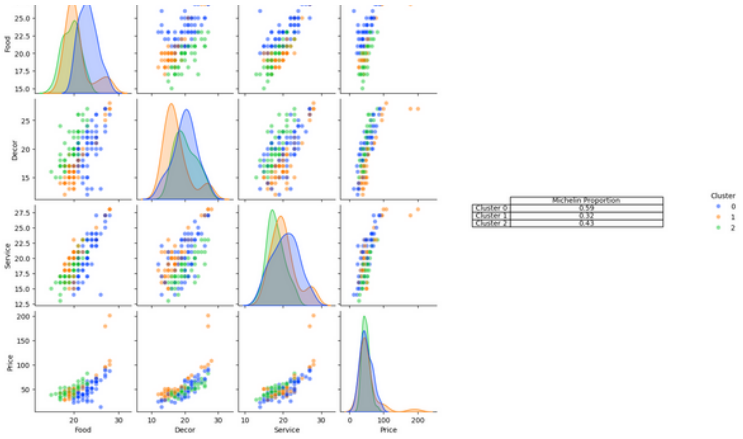


Complete & Correlation Dendrogram

Complete linkage links the furthest away by the distance metric - which in our case is correlation. Correlation measures distance by how similar the clusters features are (are price and service correlated?). This intuitively makes some sense as the Michelin stars can be divided based on consistency (correlation) and the complete linkage serves as testament to the 'superior' nature of receiving the Michelin star. We can now analyse the final clusters in a bit more detail. This should isolate some contrasting features between Michelin and non-Michelin star restaurants.

Analysing the clusters

To the right we have the clusters separated by the low-entropy and even distribution algorithm we can see data that reaffirms our discovery through reducing the dimensions. We have created a pair-plot for each of the variables and divided them into clusters based on the above. What we find is that the cluster with the most Michelin restaurants (cluster 0 - blue) is superior on all counts and the cluster with the lowest is inferior for all variable levels. It seems as if - Michelin starred restaurants are simply good quality restaurants. With this unsatisfactory discovery I turned to different methods to figure out what it truly takes to obtain the Michelin star.

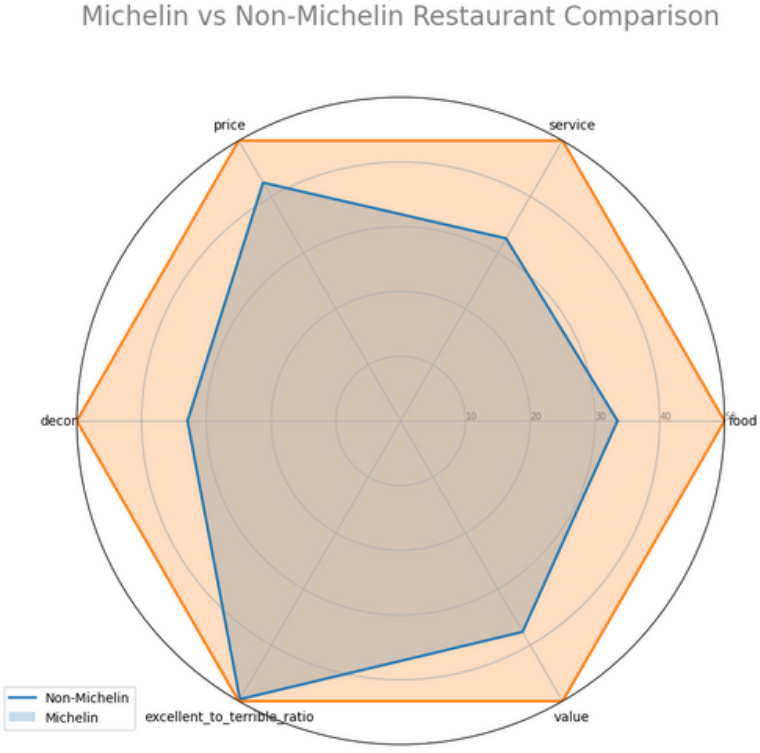


Part 3

OBTAINING MORE DATA

tripadvisor

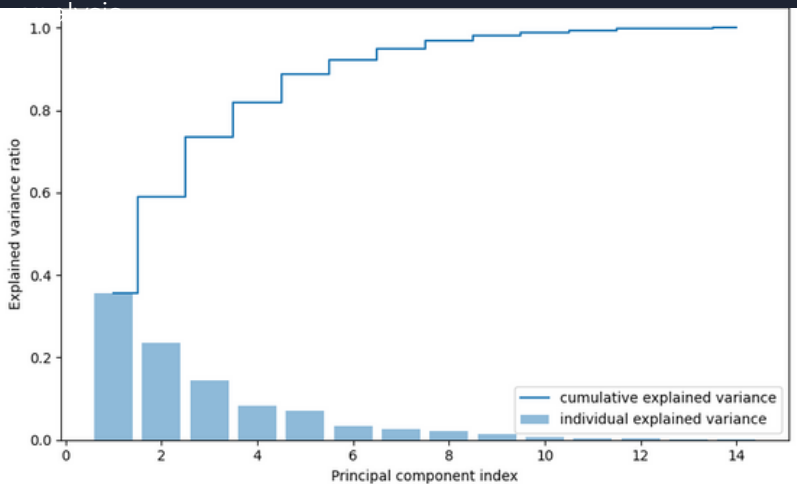
Since my question still lingered, I knew the obvious solution to determine the answer to my question was to gather more data. The obvious source was - TripAdvisor, with reviews and ratings that would cover a wide range of restaurants and that I could trust. I built a Selenium bot in Python that scraped the existing dataset for additional data and gathered a whole lot more. Now, we were working with a massive dataset that had been developed over 13 hours of scraping. It contained new variables such as latitude, longitude, the original variables, number of reviews, the number of excellent, very good, average, poor, terrible, I conducted sentiment analysis on three pages of reviews adding sentiment polarity (how negative/positive the review was), sentiment subjectivity (was it based on opinions or facts?), Tripadvisor URL, restaurant URL, cuisine types. Now I felt I was adequately equipped to answer my question. In addition, I also feature engineered two more columns that would later help in my advanced clustering analysis - value (food / price) and excellent to terrible (excellent / terrible) maintaining the linear nature of the variables. We are now dealing with real world data and hence unknown data. In the case of missing data, my strategy involved using the median. Below is a snapshot of the data (I ended up cutting some columns).

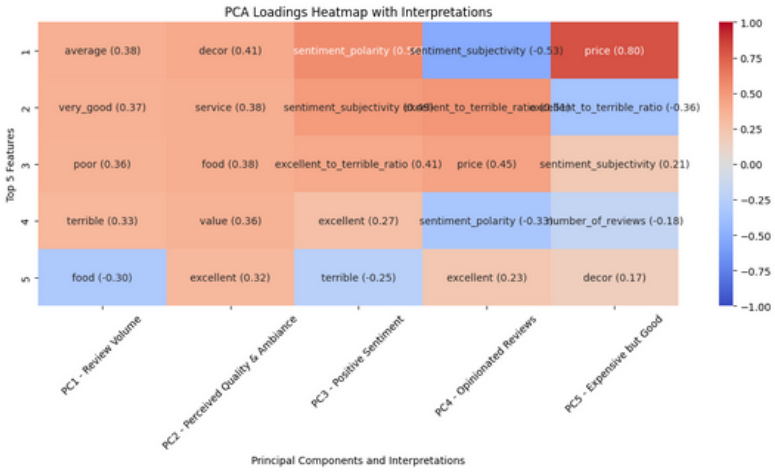


Part 4

PCA AGAIN

Now we conduct PCA again. I perhaps should mention now that we are getting a bit more advanced, about certain nuances involved with reducing the dimension of the data. We first ensure there is no empty values or else we have to use an imputer to handle those empty values. In addition we must normalise the data to a mean of 0 and a standard deviation of 1 as data can be represented across various different scales. We also must choose the number of PCA variables to consider. This is mainly done using a scree plot as seen below. The additive variation is represented by the bars and the cumulative denoted by the stair like line. I am going to go with five variables as it represents upwards of 85% of the variation. We can then see the interpretation from the PCA





PC1 - General popularity of a restaurant
PC2 - Quality
PC3 - Subjective reviews
PC4 - Good reputation (good sentiment & excellent to terrible)
PC5 - Expensive but not overpriced

Part 4

EVEN MORE ADVANCED CLUSTERING

K

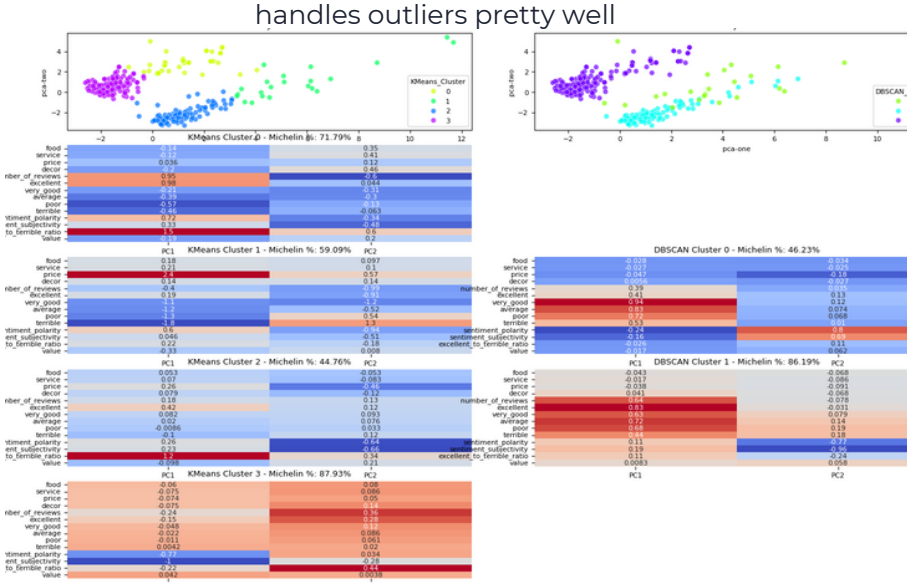
K Means

We use an advanced clustering technique that attempts to group data around a predetermined number of points by distance (euclidean or other) and calculating the mean of all those points and updating that point to the new location and iterating until the points no longer move (and the algorithm 'converges')

DB

DBSCAN (Density Based Spatial Clustering of Applications with noise)

This is another advanced technique that does not require to initialise a predetermined number of points. It divides the data into three clusters core, border and noise. A core point has x (input variable) within a certain circular distance of it, those points that lie within the radius are border points - and those outside are 'noise'. Hence, this algorithm can find non-circular clusters and handles outliers pretty well



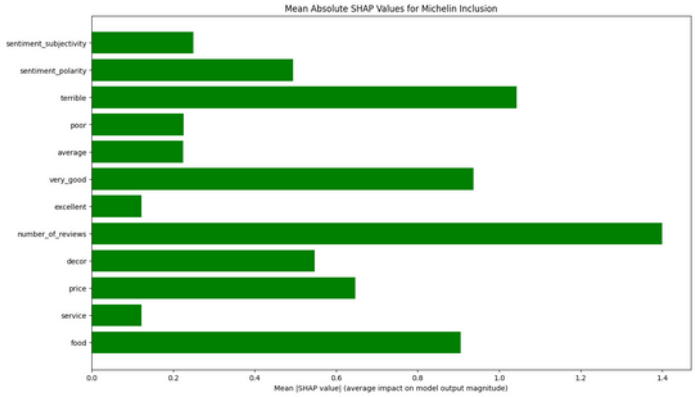
Finally

We finally uncover some insight into the data. I know the graphs to are right look complicated but all they are, are the same dimension reduction done for each of the clusters as defined above. It seems that in both clustering techniques that popularity is emerging as a common theme among Michelin star restaurants (see cluster 3 K-means and cluster 1 DBSCAN)

Part 5

PUTTING THIS ALL TOGETHER

I applied a variety of machine learning models to forecast which restaurants will be awarded a Michelin star. In this sense, a model is a tool for data analysis that finds patterns in data that may be used to forecast results. Our first model, logistic regression, had an accuracy rate of 79%. It was too simple to capture anything of value. Our accuracy was increased to 81% by using ensemble approaches, which integrate predictions from various models and better capture the interaction between variables like food and service. XGBoost achieved the highest accuracy of 85% by efficiently learning from more complex patterns in the data. The success of the XGBoost at finding the complex data suggests that the patterns in the data are much more complex than is capable of linear analysis. Nevertheless, we shall analyze two decisions made by the model to understand how beneficial linear analysis can be on data.



Analysing the Michelin Cluster

We have performed a SHAP analysis on the XGBoost model to identify how it is splitting up the data. Remember, XGBoost is like a decision classifier (split the data by number of reviews, then by food etc.). So what have we uncovered from the data? We can see that popularity (number of reviews), food, price, terrible reviews and very good (surprising?) triumph the decision classifier. This has, to a slightly higher degree, uncovered the same insights as our original linear analysis. This successfully predicts Michelin (or potential Michelin inclusion based on those parameters.

The outcome

The outcome (apart from the analysis) was a tool that given a TripAdvisor restaurant name, could access the various aspects of that restaurant and with the help of a decision classifier model called XGBoost could predict Michelin inclusion to an accuracy of 85% on test data which was further tested on a small sample of unseen data. So could restaurants in Dublin use this tool to see where they could improve on? Likely not, the model has many limitations and as such would require a huge amount of extra data and further optimisation before that is a reality. But, the conclusion from the new data show that getting your restaurant out there, so it could be worth the effort in hiring marketing people, social media experts and customer success teams? In addition to the above, I quickly came to the conclusion that this data was not linear with some of the new features having to be removed as they worsened the model outcome, since the new features were linearly derived. It is clear the above model (assuming data is of decent quality) captures a much more complex relationship than was first imagined when I analysed the PCA components of the first dataset. So what would I do differently if I was to start again?

1 I would understand the Michelin star process in a lot more detail. For example, on the website they highlight the chef and the food as being very important factors which are dataset captures only partly. In addition, there seem to be a huge amount of Michelin star restaurants in Europe as opposed to any other country which would have made a more determinable conclusion

2 I would also have left the scraping algorithm run for a bit longer. As the quality of the data could be trusted (TripAdvisor) it was safe to say there were very little more trusted data sources available (for free). This would have made conclusions a bit more rigid

Finally

To sum up the analysis in a couple of sentences: I realised that the overall quality of a restaurant is of significant weight when deciding on whether or not a restaurant should be included in the Michelin Guide. With the correlation hierarchical cluster affirming this via the correlation distance metric (insinuating Michelin restaurants features are all correlated price and price are similar). The reviews told a story of an “exceptional experience” and “impeccable customer service” whereas the normal restaurants used less elegant terminology such as “excellent”. In our more advanced analysis whereby we used k-means and DBSCAN to analyse the clusters considering the effect (and there were some) of outliers in our dataset in which we discovered the publicity of a restaurant as a potential factor in receiving a Michelin star. We then assumed non-linearity and attempted to find complex relationships using complicated decision tree model which validated our findings about having an established presence.

https://github.com/Odhran7/Data-Analysis-Restaurants