Abstract

In this project, an interactive dashboard is designed for the objective of comparing London boroughs and regions over various metrics such as the rating and the total number of visitors for each place, an NLP model is also developed to predict the rating of each review and the emotional sentiment behind each review. In theory, the challenge of predicting the review rating or the sentiment of the review based on textual data only revolves around understanding and analysing sentiments from the texts provided. On the other hand, successful implementation of the NLP model along with a comparative analysis can be considered an essential asset for decision makers in the business world; generating insights on potential business opportunities and aiding the process of developing marketing strategies for already existing businesses, aiming to enhance the understanding of business decisions from a data analytics perspective, as well as utilizing business intelligence tools to provide useful visualisations to back the business' decision-making process. The project implementation successfully achieved satisfactory results in terms of accuracy and model evaluation for the NLP model, with the optimum accuracy reaching around 88.6% for the sentiment prediction. As for the geo-tagged data, PowerBI and ArcGIS were used to visualise the data and provide useful insights for the data collected. The data was collected from Google Maps using Google Maps' API platform, which is available for the public, thus removing any ethical issues, all personal information retrieved where excluded from the data retrieved, information such as the user's name, location, and other user-specific information. The outcome of this project is completely unbiased and is only focused on areas and businesses for business-focused intentions, therefore, the project completely satisfies all the ethical and legal principles. The results achieved in this project satisfied the design requirements and specifications, providing accurate and satisfactory results.

Keywords: Natural Language Processing; Sentiment Analysis; Geo-Spatial; Python; PowerBI; ArcGIS; Google Maps; London

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Methodology

1.1.1 Data Retrieval

The data retrieval stage is an essential stage for this project, the data required to complete this project needs specific information about the places in London along with the coordinates for the place, this can be achieved by accessing the Google Maps API. Using an API to retrieve the data provides the most recent and consistent data available which is an advantage to using online resources of data, it also supports automation for future work which is out of scope for this project but nonetheless an important factor.

The data retrieval process aims to collect raw data from the Google Maps Places API which provides location information along with personal reviews from people regarding the place in the query, the process uses 2 main functions in the Google Maps Places API which are the “Nearby search” function and the “Details” function.

The “Nearby Search” function takes the coordinates, radius of search in meters, search keyword or place type, and the API access key, the output is an XML file that shows the details of up to 60 places per search, to obtain a sufficient amount of data, plenty of search queries needed to be done, the final output produced around 4851 rows of data without any cleaning.

As for the “Details” function, the input required for this query is the “place_id” column from the prior search along with the API access key, the output is in the form of a JSON document that includes several information about the place along with 5 reviews, given the total number of reviews for the place exceeds 5 reviews, otherwise it would result in returning all the reviews. The challenge in this application is that it requires the “place_id” to be entered separately for each place, thus, making the review retrieval process quite exhaustive. The most suitable approach for this was to automate the process for the queries.

1.1.2 Data Cleaning

The data cleaning process is an important part of every data project, it can either make or break the project. Data cleaning is basically removing or fixing any inaccurate, erroneous, and duplicate data in the dataset. The project relies heavily on the integrity and the validity of the dataset, mainly because all the insights included are based on the retrieved and cleaned datasets.

The data retrieved is relatively clean in terms of context, but there were several duplicates and nulls in both data-frames. After dropping the duplicates in the Places data-frame, the number of rows in the data-frame was 3640 rows.

The data cleaning process for both data-frames consisted of selecting relative columns, dropping duplicates, dropping specific rows based on crucial null values, fixing column types, and merging and rearranging the columns. The Places data-frame eventually reached 3639 places, and the Reviews data-frame yielded 13383 reviews.

The Places data-frame columns are (name, area, latitude, longitude, rating, no_of_reviews, place_id, price_level, and category), as for the Reviews data-frame, it consists of (place name, address, phone_number, latitude, longitude, average rating, no_of_reviews, place_id, review_text, review_rate).

1.1.3 Natural Language Processing

It is important to define the process of Natural Language Processing (NLP), according to IBM, “Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.” (IBM Cloud Education, 2020), which means that NLP is a combination of computational algorithms that are based on specific rules regarding any human language with statistical models, which is known as machine learning models.

NLP is used in this project to connect sentiments from the texts provided and feed it into a machine learning model to predict future reviews and sentiments, there are several methods to using NLP, where a few of these models are experimented in this project.

Design & Strategy

3.1 Data Collection

The data collection process consists of several steps, firstly, in order to obtain data from Google's Places API, a set of coordinates was needed to adequately complete the queries, the places query takes a number of inputs which are:

- Location; a coordinates pair of longitude and latitude representing the centre point of the search.
- Radius; a value in meters that represents the radius of the search area.
- Keyword/Type; the keyword and place type of the places to look for.
- Key; API key generated by the Google API Platform.

The total queries for each coordinate pair revolved around 21 Keyword/Type places, which are: Café, bar, point of interest, restaurant, museum, movie theatre, art gallery, bookstore, bowling alley, casino, library, night club, park, tourist attraction, zoo, health, lodging, clothing store, gym, stadium, and Iconic Places, with a fixed radius set at 100,000 meters.

This resulted in far more types in the output of the data collected. A sheet was constructed containing the API queries as links for 7 main areas in London which are: Wembley, Greenwich, Soho, Westminster, Kingston Upon Thames, Chelsea, and the centre point of Greater London, the sheet was used in an automation tool that was developed for the purpose of data collection using Python. The tool relies on Python libraries to open a chrome window, insert the link, save the output as an XML, and convert the output into a CSV using the online tool (https://www.convertcsv.com) that converts XML into CSV, the automation tool was run for all the regions mentioned.

Given the fact that Google's API provides up to 60 places for each query in up to 3 pages, each page of queries was saved as a separate CSV file, where each region would have around 60 CSV files, and then the files would be appended using Python to a general CSV file for each region, with the duplicates being dropped along each step, mainly because the data had a lot of overlaps, the CSV files representing each region were then appended. The automation tool for the places retrieval process was run for each region separately, the final source code for the automation tool can be found in Appendix A.1.

The data collection process from the outputs of each automated procedure was a simple process of cleaning the collected data in terms of dropping nulls and removing duplicated places, the code for the data collection can be found in Appendix A.2.

As for the reviews, the process was a little different, the query for the review retrieval takes only the place ID and the API key, with the place ID being one of the columns in the data collected for the places. The automation tool was adjusted to retrieve the data from each query by feeding the place ID from the places data-frame that was constructed to the query, save the output in a list of JSON files as one large text file, the file was then converted into CSV using the same online tool (https://www.convertcsv.com). The adjusted automation tool for the reviews can be found in Appendix A.3. The general distribution of the data collected from Google Maps can be seen in Fig. 1 below:

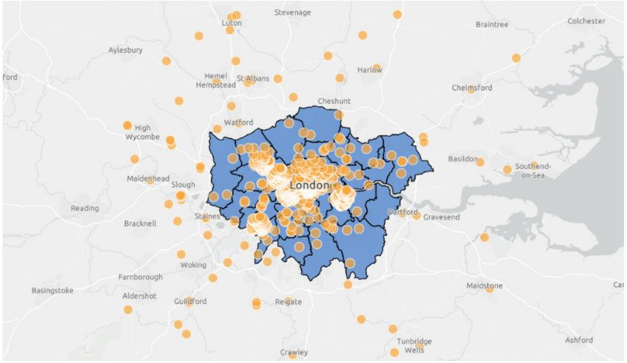


Figure 1: Collected data visualized on the London Map

In Fig. 1, the places are represented in orange circles on the map, it can be seen that there are several places that lie outside the London region, which is highlighted in blue on the map, these places can be considered as getaway destinations for London residents and visitors. The popularity of these destination is explored throughout the report.

3.2 Data Preparation and Cleaning

The data preparation method started with reading the two CSV files that were created from the data retrieval process, one for the places and the other for the reviews, with each data-frame being prepared separately.

As for the places data-frame, the first step was to drop all unnecessary columns from the raw data retrieved, which amounts to 25 columns in total. Afterwards, a new data-frame was constructed based on the original data-frame, fixing the column names, and setting the columns in adequate order. The data-frame required no further drops as the duplicates were dropped in each step of data collecting, having a total of 3639 rows and 21 columns in the cleaned data-frame.

The structure of the places data-frame has 21 columns, 13 of which represent the place’s main category and sub-categories, with the rest being the name, area, latitude, longitude, rating, no_of_reviews, price_level, and Place_ID. Having 13 types for each place proved to do more harm than good for the purpose of this project, thus, the first 4 categories for each place were only considered in the data as 1 main category and 3 sub-categories.

For the reviews data-frame the process was similar, starting off with selecting the necessary columns instead of dropping the unnecessary ones as the number of columns needed was much less than the ones not needed, and then dropping the places that had no reviews in total, and then creating a new data-frame with proper column names and order. The reviews data-frame required further manipulation as the reviews were column-based rather than row-based, the manipulation process resulted in having 5 similar rows for each place with the only difference being in the review text and review rate columns, as the data-frame had 3050 places in it, the total number of rows for the data-frame yielded 13383 rows and 10 columns after dropping rows with no review text or rate as well.

The structure of the reviews data-frame consists of 10 columns which are: Name, Address, Phone_Number, Latitude, Longitude, Rating, No_of_reviews, Place_ID, Review_text, and Review_rate.

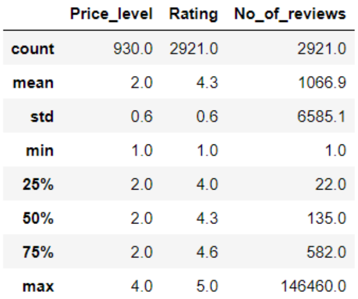
It was noticed that the places data-frame had 549 rows where the values of the average rating and total number of reviews were nulls, as well as a slight difference in some rows in terms of total number of reviews, this was mainly because the data collection for the places occurred before the reviews collection and there was a short amount of time between the two steps where the total number of reviews has slightly increased. Hence, the two data-frames were merged to achieve the best possible outcome for the places in the data-frames. The merging process was an inner join process based on the place ID column, and after dropping the rows that had duplicated Place ID, the merged data-frame had 2921 rows and 14 columns.

The structure of the merged and final data-frame has 14 columns which are: Name, Area, Address, Latitude, Longitude, Price_level, Rating, No_of_reviews, Type_0, Type_1, Type_2, Type_3, Phone_Number, and Place_ID. The Price_level and Phone_Number columns were only added as a reference to the place and did not prove to be essential for the project. The python code for the data preparation stage for all data-frames can be found in Appendix A.4.

3.3 Data Analysis

The initial statistical analysis of the final data-frame had some interesting insights, the data-frame had 930 places out of 2921 places where the price level of the places is defined, having an average price level of 2 out of 4 across the 930 places, as for the average rating, the lowest average rate in the data-frame is 1 out of 5 and the highest being 5 out of 5, with an average total reviews of 1067 per place. The place with the highest total reviews had 146460 total reviews, which can be assumed to be the value for total visitors, and the lowest was 1 review, with the price level indicating the average prices in the place, with 1 being the cheapest and 4 being the most expensive criteria. The insights for the initial statistical analysis of the data-frame can be seen in Table 1 below.

Table 1: Initial Statistical Insights



	Price_level	Rating	No_of_reviews
count	930.0	2921.0	2921.0
mean	2.0	4.3	1066.9
std	0.6	0.6	6585.1
min	1.0	1.0	1.0
25%	2.0	4.0	22.0
50%	2.0	4.3	135.0
75%	2.0	4.6	582.0
max	4.0	5.0	146460.0

In order to achieve a much deeper analysis of the data, the decision to create a dashboard to represent the data was made, the dashboard would be created using Microsoft’s PowerBI tool, providing interactive visualisations and giving more freedom in taking analytical insights based on various metrics. The dashboard creation will be further discussed in the Visualisations & Analysis section of this chapter, with the results being discussed in the next chapter.

3.4 NLP

The NLP part of the project aims to predict the sentiment behind every review, this can be used in several ways to enhance the customer review experience and support the objective of the project. The sentiment prediction tool can be used to analyse offensive and inappropriate language to automatically reject the review from being posted online to the general public, which enhances the customer experience, as for the objective of the project, the tool will be able to predict the sentiment behind a certain review based on textual data only and predict the review rate as well, this can be developed as a user interface where the review can be fed into the tool and it will automatically predict the output in terms of sentiment and rate.

3.4.1 Data Preparation

The data preparation stage for NLP takes the review text and review rate columns into a new data-frame used specifically for the NLP part of the project, since the main data-frames were already cleaned and prepared for implementation, no further data cleaning was done on these columns in terms of data validity and data type. There are 47 duplicated reviews in the text column only, as this is considered an extremely low percentage out of the 13383 rows in the data-frame and having 47 similar reviews for different places is a valid reason depending on the length of the review and other similar factors as well, the decision was made to proceed with the data with the 47 duplicated reviews.

3.4.2 Text Cleaning

The text cleaning process consists of 5 main stages: removing URLs, removing HTML, removing punctuations, removing stop-words, and lemmatizing the texts. Each stage in the text cleaning process is defined into a separate function and all functions were applied to the review text column.

The first 4 stages focus mainly on cleaning the text from any noise that may affect the NLP prediction, with the first 2 stages ensuring that all URL and HTML texts are removed to safeguard that the text is relevant to the English language only, the third and fourth stages guarantee that the words in the text have a higher influence on the decision making process for the models by removing all punctuations and stop-words. The text was then lemmatized and at a later stage it was also stemmed, the purpose of the lemmatization and stemming methods is to reduce the inflectional form of the word, by using the base form of the word it would allow the models to better predict the target by assessing base words instead of having various versions of the same word that affect the NLP analysis process.

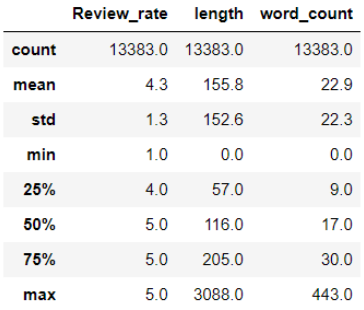
3.4.3 Feature Engineering

The feature engineering step of the NLP stage focuses on adding relevant columns for exploratory and NLP enhancement reasons, the main features that were added to the dataset are the word count and the text length columns, with the first representing the number of words in the cleaned text pre-stemming and the latter representing the length of the text in terms of characters. The critical feature that was added to the dataset is the sentiment column, which is a rule-based column based on the review rate for each review; the reviews that have 4- and 5-star ratings are classified as ‘Positive’ reviews, the reviews that have 1- and 2-star ratings are classified as ‘Negative’ reviews, and the 3-star rated reviews being classified as ‘Neutral’ reviews. The sentiment column will be used to train the ML algorithm and use NLP to predict the sentiment of the review based on supervised learning machine learning models that are pre-labelled.

As there are various ways to predict a sentiment from texts, in this project, a relationship between the words in the review and the review rate was established to connect the abstract sentiments of ‘Positive’, ‘Neutral’, and ‘Negative’, with the words used in the texts. The decision for this approach was mainly dependent on the data, the data retrieved is all classified into 5 ratings and the consistency of this rating was used to cluster highly rated reviews as positive sentiments, neutral reviews as neutral sentiments, and negative reviews as negative sentiments.

The average length for each text review is around 156 characters long, having around 23 words in each text, Table 2 below shows the overall statistics for the reviews based on review rate, review length, and word count.

Table 2: Textual Data Insights



	Review_rate	length	word_count
count	13383.0	13383.0	13383.0
mean	4.3	155.8	22.9
std	1.3	152.6	22.3
min	1.0	0.0	0.0
25%	4.0	57.0	9.0
50%	5.0	116.0	17.0
75%	5.0	205.0	30.0
max	5.0	3088.0	443.0

3.4.4 NLP Preparation

Machine learning models rely on numerical values to predict a certain value, therefore the text and the sentiment columns need to be properly adjusted to meet the NLP requirements, each column will be altered in a separate way depending on how the column is used, as the text column will be used as an input feature for the model and the sentiment column will be used as an output.

The sentiment column has been encoded to represent the sentiment in a numerical format, with the value 2 representing ‘Positive’, 1 representing ‘Neutral’, and 0 representing ‘Negative’. After the encoding procedure was completed, the texts in the text column were stemmed to increase the effect of the base form of the words as mentioned before to improve the accuracy for the target prediction in the algorithm.

The data was split into training and testing datasets with a ratio of 80-20% respectively. The text column was then further adjusted to enhance the NLP analysis, each review was tokenized, fit to the sequencing algorithm that converts the words into numbers for ML modelling, and pad the sequences to maintain a fixed sequence length throughout the dataset. The textual analysis for NLP has showed that there are 15926 words in the texts used for the project, with 13954 words out of the 15926 words being unique words.

3.4.5 Machine Learning Models

Since the aim of the NLP is to predict the review rate and the sentiment for each review based on textual data only, the most suitable ML approach for this project would be the Logistic Regression algorithm. Logistic Regression is able to handle multi-class outputs for the prediction, and in this project, 2 main outputs would be predicted separately, with both outputs having more than a binary output; the rate prediction have 5 separate classes, and the sentiment prediction have 3 separate classes, meaning that both outputs would require a multi-class logistic regression model.

The logistic regression models in this project are all built on TF-IDF, which is used to convert the texts into useful vectors, the class weight for all models is set to balanced in order to minimise the data imbalance effect on the model. TF-IDF was chosen because it showed better results in terms of accuracy and F1 score.

The systematic experiment for the models was designed to adopt the ’liblinear’ library, which is better suited for multi-class models, The first model uses an ’lbfgs’ solver which required the maximum iterations to exceed 100 iterations, while the models that use the ’liblinear’ solver, with the same penalty choice ’L2’, achieved better results in terms of accuracy and F1 score, the results achieved in the ‘liblinear’ model were comparatively better than the results achieved in the traditional first model.

The complete code for implementing the NLP stage of the project from the data preparation to the ML development and results can be found in Appendix B.

3.5 Visualisations & Analysis

The visualisation and analysis section aims to provide statistical visualisations that give insights on the data at hand, as well as providing a comparative analysis between the various areas in London and various categories for the places in the dataset. As there are several perspectives that can shape how the comparison would be in terms of metrics and user preference, it was found that ideally, a dashboard report for all the places would be developed to ensure all perspectives are met and each user can derive useful and relevant insights based on the users’ own interests and preference.

Microsoft PowerBI was used to develop the dashboard report, using ArcGIS Maps for Power BI to provide an improved geo-spatial analysis and visualisations for the data. The ArcGIS features integrated into Microsoft PowerBI add depth to the spatial analysis of the data as it helps analyse the patterns and trends that are difficult to represent using statistical and standard PowerBI features, the ArcGIS integration was primarily used to add regional depth into the visualisations by showing the specific London boroughs within the map and also using geo-spatial focused features such as heatmaps and map-clustering that clusters the places according to a certain viewpoint.

The dashboard design consists of 6 pages, each page aims to classify and satisfy a different comparison perspective, the first page in the report is a general overview of the dashboard along with a categories filter that can be used to classify the places. The other pages provide more detailed insights whether it was geo-spatial or statistical, with 2 pages focusing on statistical analysis of the dataset and 3 more pages providing geo-spatial insights in addition to a variety of statistical insights.

2. Results & Discussion

4.1 NLP

4.1.1 Exploratory Data Analysis

The data was explored to provide statistical insights prior to the NLP development stage, it was found that the data was more clustered around 5 star reviews, which is an acceptable situation given the fact that positive reviews in the dataset are much more popular than negative and neutral reviews, Fig. 2 shows the distribution of the data based on the rating for each review, while Fig. 3 shows the distribution of the data across the 3 sentimental categories, both figures are shown below:

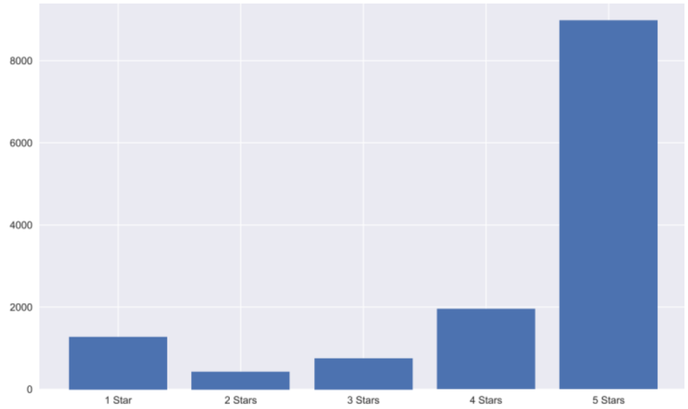


Figure 2: Review Rate Distribution

It can be seen that the 5-star ratings reach almost 9000 reviews, while the second most popular rating is the 4-star rating which has around 2000 reviews. On the other hand, the 1- 2- and 3-star ratings have an accumalted total of around 2500 reviews across all three categories, where the 1-star rating had the highest rank, followed by the 3-star rating and then the 2-star rating.

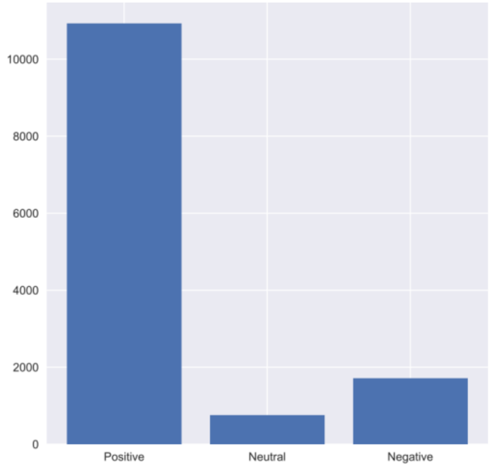


Figure 3: Sentiment Distribution

The sentiment distribution confirms the distribution of the ratings discussed above, with the positive reviews stacking almost 11000 reviews, while the negative reviews almost reach 2000 reviews, and the neutral reviews being slightly below 1000 reviews approximately.
The word-cloud is used to show the most popular words in the textual data, it is an illustration that shows the most common words in the dataset in a larger size than the less common ones, as can be seen in the word-cloud in Fig. 4, the most popular words in the dataset are Great, Good, Place, Nice, London, Amazing, and several others, the word-cloud can be seen below:

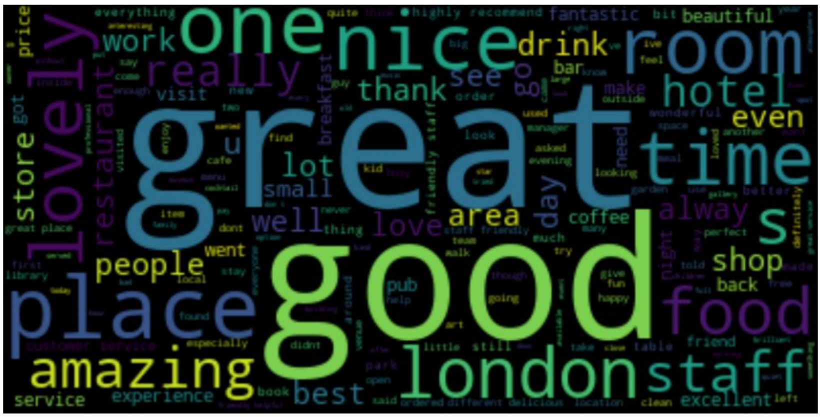


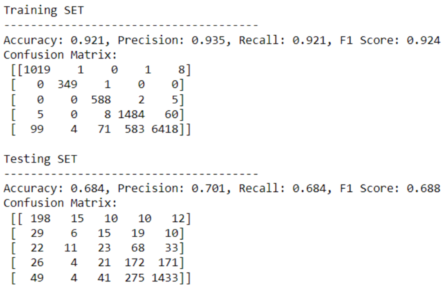
Figure 4: Word-cloud

As for the models used in the project, the design experimented 4 models, with 2 models aiming to predict the rating of the review, and the other 2 models aiming to predict the sentiment behind the review. The 2 models for each target are fairly similar, and both are based on Multi-Class Logistic Regression algorithms.

All models were trained with a TF-IDF vectorizer, the TF-IDF vectors were implemented with a (1, 2) n-grams, which means that the vectorizer would set the vectors to range between uni-grams and bi-grams. The decision to use uni-grams and bi-grams was an experimental decision based on the accuracy and F1 scores that were achieved across various configurations, the aforementioned configuration of (1, 2) n-grams achieved the best possible results for all models.

4.1.2 Model 1: Logistic Regression for Rating

The first model used a balanced class wright with maximum iterations set at 1000 iterations, the model was fit and trained accordingly with the TF-IDF vectors split across the training and testing data, the model had an accuracy score of 68.4% and an F1 score of 68.8%, the training results achieved much higher scores when compared to the testing results, this is due to several reasons, the main ones being that the data has an imbalance towards 4- and 5-star ratings, the imbalance effect on a multi-class model could lead to this kind of variation between the training and testing sets, as well as having several classes which widens the prediction criteria, nonetheless, the testing scores can be considered satisfactory in terms of ML algorithms. The full results of the model for the training and testing datasets are shown in Fig. 5 below along with the confusion matrix for each dataset:



```
Training SET
--------------------------------------
Accuracy: 0.921, Precision: 0.935, Recall: 0.921, F1 Score: 0.924
Confusion Matrix:
 [[1019    1    0    1    8]
 [   0  349    1    0    0]
 [   0    0  588    2    5]
 [   5    0    8 1484   60]
 [  99    4   71  583 6418]]

Testing SET
--------------------------------------
Accuracy: 0.684, Precision: 0.701, Recall: 0.684, F1 Score: 0.688
Confusion Matrix:
 [[ 198   15   10   10   12]
 [  29    6   15   19   10]
 [  22   11   23   68   33]
 [  26    4   21  172  171]
 [  49    4   41  275 1433]]
```

Figure 5: Model 1 Results

The confusion matrix shows the true/false results for each class, with the main diagonal being the true and accurate result for each class, in the training set, the 1-star results that were predicted accurately are 1019 reviews, with the same statistic being 349, 588, 1484, and 6418 for the 2-star, 3-star, 4-star, and 5-star classes, respectively. It can be seen that the numbers in the main diagonal dropped drastically in the testing set due to the reasons mentioned before and due to the fact that the accuracy has dropped in the testing set.

4.1.3 Model 2: Logistic Regression using ‘LibLinear’ solver for Rating

This model is an adaptation of the first model that uses a ‘Liblinear’ solver instead of the traditional L2 solver, the ‘Liblinear’ solver is one of the additions that enhance the results from the ‘liblinear’ Python library, the model was fit and trained using a balanced class weight, which is similar to model 1, with the solver being specified to be ‘Liblinear’ and asserting the penalty for the model to be assessed using the L2, the results achieved in this model were relatively better than the results achieved in model 1 for the same classification.

The liblinear model achieved an accuracy of 73.3% with an F1 score of 69.3%, the results increased slightly from the 68.4% and 68.8% results achieved in the basic model. The model achieved better results in the training set scores but it is more significant in the testing set score which are the scores that reflect the actual model performance.

The confusion matrix for this model clearly shows that the numbers have increased from the previous model, as the main diagonal numbers in both the training set and the testing set are higher than the numbers achieved in model 1, hence the higher accuracy.

4.1.4 Model 3: Logistic Regression for Sentiment

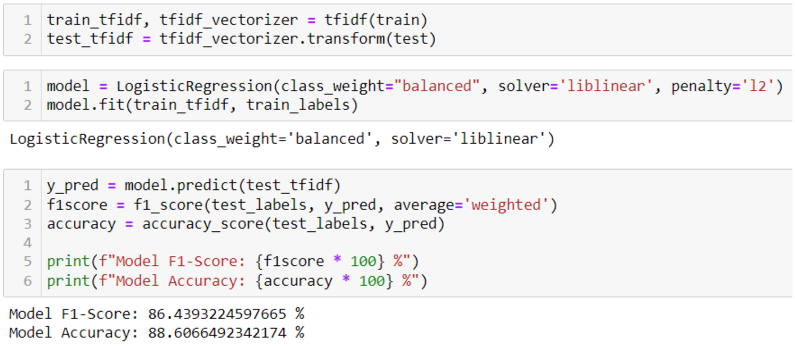
This model is an exact implementation of model 1, with the main difference being the target of the model; in model 1 it was the rating score and in this model it is the sentiment, the difference in the classification showed much more accurate results, and these results can be considered more relevant to the purpose of this project as it aims to show the emotion and sentiment behind each review rather than the actual rate.

The model was fit and trained using the same set of configurations for model 1, but the splitting process had the sentiment column in the target labels rather than the rating column, the splitting process also split the data into 80% training data and 20% testing data.

The model achieved significantly better results when compared to the rating-based models, the model produced an accuracy of 87.2% and an F1 score of 87.1%, which are ideal results for the testing set, the model achieved higher than expected in the training set, with the accuracy and F1 scores reaching almost 97.7%, this could be a sign of overfitting in the model due to the imbalance between the classes.

4.1.5 Model 4: Logistic Regression using ‘LibLinear’ solver for Sentiment

This model is an implementation of model 2 with the same set of configurations, with the target column being the sentiment column instead of the rating column. The model uses the ‘liblinear’ solver that achieved better results in the rating models. Unlike the rating models, the ‘liblinear’ solver for this model achieved a better overall accuracy when compared to model 3, but the F1 score decreased slightly, but this is a sign of better imbalance handling as the accuracy is considerably higher than the F1 score, and it is evident.



```
train_tfidf, tfidf_vectorizer = tfidf(train)
test_tfidf = tfidf_vectorizer.transform(test)
```

```
model = LogisticRegression(class_weight="balanced", solver='liblinear', penalty='l2')
model.fit(train_tfidf, train_labels)
```

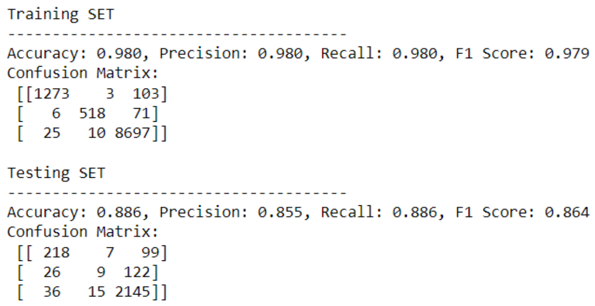
```
LogisticRegression(class_weight='balanced', solver='liblinear')
```

```
y_pred = model.predict(test_tfidf)
f1score = f1_score(test_labels, y_pred, average='weighted')
accuracy = accuracy_score(test_labels, y_pred)

print(f"Model F1-Score: {f1score * 100} %")
print(f"Model Accuracy: {accuracy * 100} %")
```

```
Model F1-Score: 86.4393224597665 %
Model Accuracy: 88.6066492342174 %
```

Figure 6: Model 4 Implementation

As can be seen in Fig. 6 above, the data was prepared for TF-IDF implementation, the model was trained using the ‘liblinear’ solver and the class_weight parameter is set to balanced to have better imbalance handling of the data. The model achieved an accuracy of 88.6% with an F1 score of 86.4%, and the training set accuracy and F1 score almost reach 98%. This model can be considered the best possible model out of the 4 models to satisfy the needs of this project, as the model shows better imbalance handling and achieved the best accuracy. The scores of this model along with the confusion matrices are shown in Fig. 7 below:



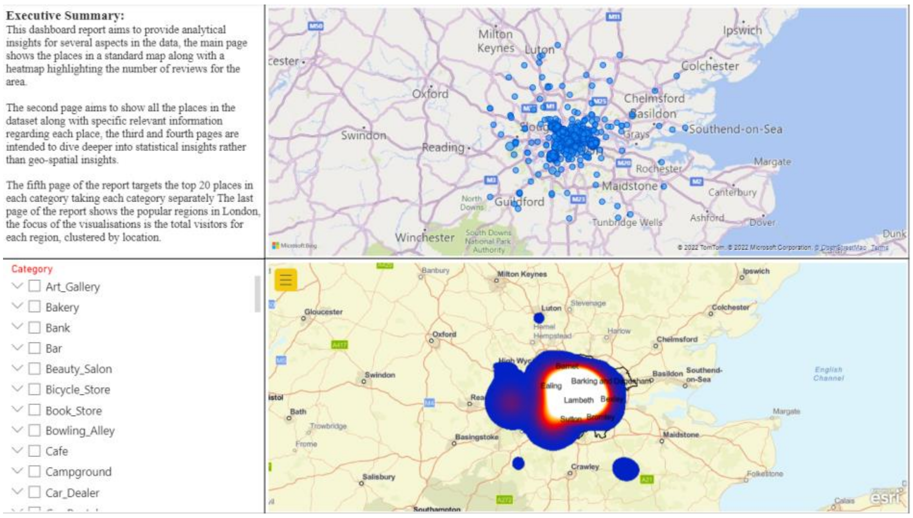
```
Training SET
--------------------------------------
Accuracy: 0.980, Precision: 0.980, Recall: 0.980, F1 Score: 0.979
Confusion Matrix:
 [[1273    3  103]
 [   6  518   71]
 [  25   10 8697]]

Testing SET
--------------------------------------
Accuracy: 0.886, Precision: 0.855, Recall: 0.886, F1 Score: 0.864
Confusion Matrix:
 [[ 218    7   99]
 [  26    9  122]
 [  36   15 2145]]
```

Figure 7: Model 4 Results

The confusion matrices for this model have 3 columns instead of 5, with each column representing the sentiment, the first column on the left represent the negative sentiment, the middle one represents the neutral sentiment, and the right column represents the positive sentiment. The confusion matrices for this model have higher numbers when compared to the confusion matrices of model 3, the main diagonal numbers increased, and it demonstrates that the model has in fact better results compared to all the other models.

4.2 Analysis

The dashboard created provides analytical insights for several aspects in the data used in this project, the first and main page of the report is the general overview of the dashboard, which shows the places in a basic map using the PowerBI maps, and a heatmap highlighting the number of reviews for the area using the ArcGIS maps for PowerBI. The general overview page can be filtered to select one or several categories for the places using the bottom left panel of the page, the map visualisations would auto-update accordingly to the selected categories. The first page of the report can be seen in Fig. 8 below:



Executive Summary:
This dashboard report aims to provide analytical insights for several aspects in the data, the main page shows the places in a standard map along with a heatmap highlighting the number of reviews for the area.

The second page aims to show all the places in the dataset along with specific relevant information regarding each place, the third and fourth pages are intended to dive deeper into statistical insights rather than geo-spatial insights.

The fifth page of the report targets the top 20 places in each category taking each category separately The last page of the report shows the popular regions in London, the focus of the visualisations is the total visitors for each region, clustered by location.

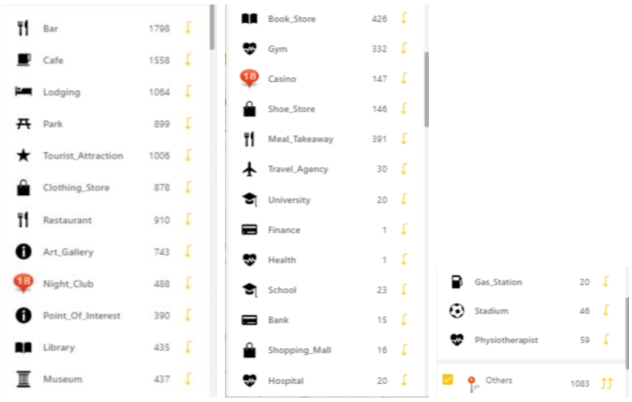
Category
- □ Art_Gallery
- □ Bakery
- □ Bank
- □ Bar
- □ Beauty_Salon
- □ Bicycle_Store
- □ Book_Store
- □ Bowling_Alley
- □ Cafe
- □ Campground
- □ Car_Dealer

Figure 8: Dashboard Overview

The heatmap shown in the bottom right panel in Fig. 8 above measures the density of total visitors in each area, where the bright yellow colour highlights the high density in total visitors, and the dark blue colour highlights lower intensity in such figures, the gradual distinction from bright yellow to dark blue represents a gradual decrease in the intensity of the number of total visitors on the map.

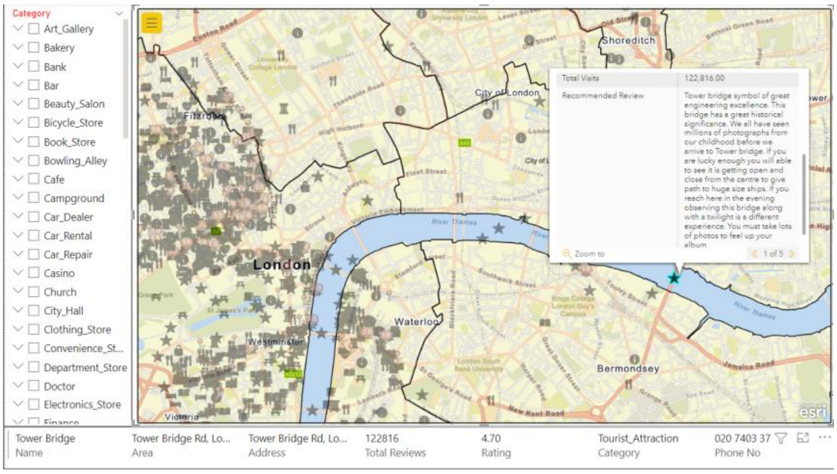
The second page aims to show all the places in the dataset along with specific relevant information for each place which are the main category, 3 sub-categories, place information, a recommended review, and the recommended review’s rate, each symbol in the map represents a single category (some categories have the same symbol), the symbols were chosen according to the main category of the place as shown in Table 3 below:

Table 3: Places Symbology



Bar	1798
Cafe	1558
Lodging	1064
Park	899
Tourist_Attraction	1006
Clothing_Store	878
Restaurant	910
Art_Gallery	743
Night_Club	488
Point_Of_Interest	390
Library	435
Museum	437
Book_Store	426
Gym	332
Casino	147
Shoe_Store	146
Meal_Takeaway	391
Travel_Agency	30
University	20
Finance	1
Health	1
School	23
Bank	15
Shopping_Mall	16
Hospital	20
Gas_Station	20
Stadium	46
Physiotherapist	59
Others	1083

The page also includes the categories filter that can be used according to the user's preference to show specific categories and get useful insights for a specific category, the bottom line of the page shows the place and relevant details for the place; the bottom line panel is bidirectionally interactive with the map, which means if a specific place is clicked on in the map the details of the place will instantly appear in the bottom figure, and vice-versa. The second page of the dashboard can be seen in Fig. 9 below:



Category
- ☐ Art_Gallery
- ☐ Bakery
- ☐ Bank
- ☐ Bar
- ☐ Beauty_Salon
- ☐ Bicycle_Store
- ☐ Book_Store
- ☐ Bowling_Alley
- ☐ Cafe
- ☐ Campground
- ☐ Car_Dealer
- ☐ Car_Rental
- ☐ Car_Repair
- ☐ Casino
- ☐ Church
- ☐ City_Hall
- ☐ Clothing_Store
- ☐ Convenience_St...
- ☐ Department_Store
- ☐ Doctor
- ☐ Electronics_Store
- ☐ Finance

Total Visits	122,816.00
Recommended Review	Tower bridge symbol of great engineering excellence. This bridge has a great historical significance. We all have seen millions of photographs from our childhood before we arrive to Tower bridge. If you are lucky enough you will able to see it is getting open and close from the centre to give path to huge size ships. If you reach here in the evening observing this bridge along with a twilight is a different experience. You must take lots of photos to feel up your album.

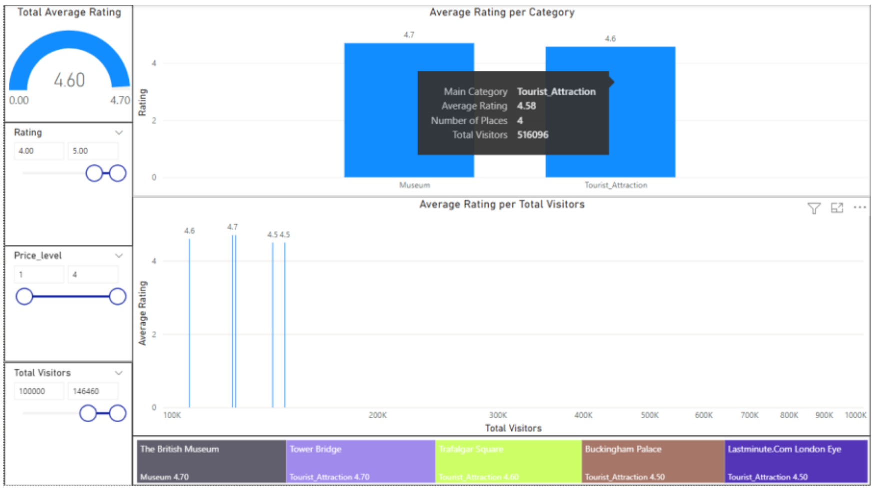
Zoom to 1 of 5

Name	Area	Address	Total Reviews	Rating	Category	Phone No
Tower Bridge	Tower Bridge Rd, Lo...	Tower Bridge Rd, Lo...	122816	4.70	Tourist_Attraction	020 7403 37

Figure 9: Places Map

As can be seen in the figure, the place clicked on is Tower Bridge in London, the relevant information is mainly shown in the bottom line of the page such as the address of the place, area, total reviews which represents the total number of visitors, average rating, main category, and the place's phone number if available. The page also utilises an interactive map tooltip that can switch between up to 5 reviews for each place, the tooltip integrated into the map displayed shows the general information of the place in addition to a specific recommended review along with that review’s rate, the recommended review tab can be scrolled through to view other recommended reviews out of the reviews for each place.

The third and fourth pages of the project are intended to show more statistical insights rather than geo-spatial insights, the third page of the dashboard report shows the total average rating for the places, the average rating per category, the average rating per total visitors, and the names of the places if the filters showed a relatively small number of places. The filters in the third page are the rating, price level, and total visitors, which can be seen on the left hand side of the page, a combination of these filters can be applied to show more detailed insights. As shown in Fig. 10 below, the filters applied focus on places that have an average rating higher than 4.0 and total visitors higher than 100k visitors.


Rating 4.00 5.00

Price_level 1 4

Total Visitors 100000 146460

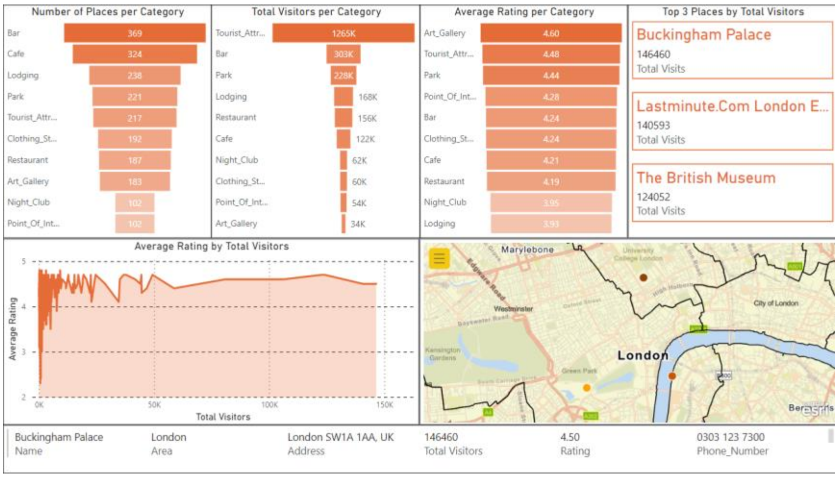
Main Category Tourist_Attraction
Average Rating 4.58
Number of Places 4
Total Visitors 516096

Figure 10: Average Rating Statistics

The figure shows that the average rating for the filtered places is 4.6 out of 5, distributed across 2 categories which are Museum and Tourist Attraction, the tooltip shows that Tourist Attraction has an average rating of 4.58, averaged over 4 places, with around 516k total visitors across all 4 places. It is also shown the average rating per total visitors for each place is represented in a vertical line in the graph, with the value being the average rate for the particular place, this graph aims to show whether there is any correlation between the number of total visitors and the average rating for each place. The bottom panel of the page shows the names of these places along with the main category and the average rate as can be seen.

This page can be utilised by policy makers in analysing the most and least visited places and assess the average rating for criteria, which can help in attracting more people for the least visited places or even help develop marketing strategies for places in close proximity to the most visited places, there are numerous use cases for this page out of which an improved service can be provided to the people and an improved business model can be adopted by businesses.

The fourth page adopts a similar statistical approach to the third page, the focus of this page is the top 10 categories in terms of number of places in each category in the dataset, the page shows the number of places in each category, the total visitors per category, the average rating per category, and the top 3 places in each category ordered by the number of total visitors. The page also shows the average rating for each category by the total visitors in an area graph to show the correlation for these metrics, and it also contains a small map that shows the location of these places on the London map. The bottom panel of the page is like the bottom panel in Page 2, it mainly shows the relevant information of the place clicked on.



Buckingham Palace
146460
Total Visits

Lastminute.Com London E...
140593
Total Visits

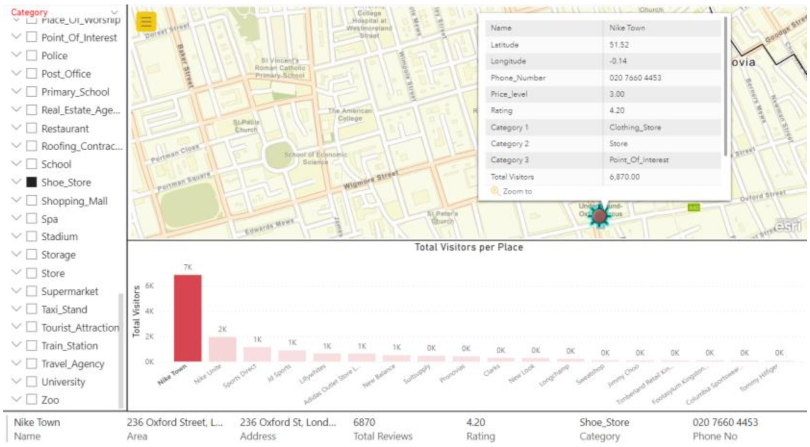
The British Museum
124052
Total Visits

Buckingham Palace	London	London SW1A 1AA, UK	146460	4.50	0303 123 7300
Name	Area	Address	Total Visitors	Rating	Phone_Number

Figure 11: Top 10 Categories Statistics

Fig. 11 above shows the most popular categories along with important metrics for each category such as the total visitors and the average rating for each category, it can also be seen that the most popular places in London are the Buckingham Palace with 146460 total visitors, followed by the London Eye which has 140593 visitors, and then The British Museum which has 124052 visitors in total, the locations of these three places can be seen in the small map as intended. This page can be utilised to analyse the most common categories for the places in London and assess the people’s engagement with such categories, it can be used by entrepreneurs to pursue ventures for popular categories that are not available in certain regions in London.

The fifth page of the report targets the top 20 places in each category based on a filter that takes each category separately, which can provide valuable insights for retailers looking to expand their business in London, the page comprises of a big map that shows the top places from the selected category in different sizes depending on the number of total visitors for each place, below the map there is a bar chart that shows the total number of visitors for the top 20 places in that particular category, the panel at the bottom of the page shows the most relevant information about the selected place as in the previous pages. As can be shown in Fig. 12 below:



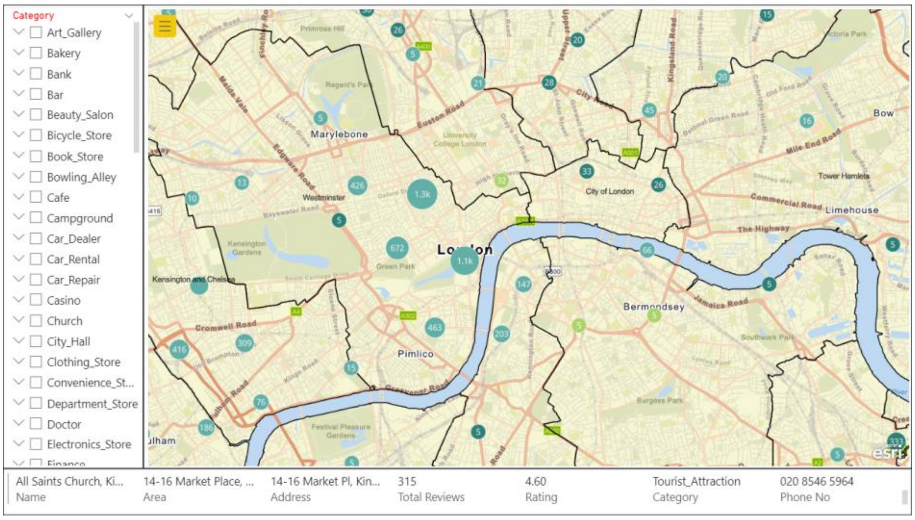
Name	Nike Town
Latitude	51.52
Longitude	-0.14
Phone_Number	020 7660 4453
Price_level	3.00
Rating	4.20
Category 1	Clothing_Store
Category 2	Store
Category 3	Point_Of_Interest
Total Visitors	6,870.00

Nike Town	236 Oxford Street, L...	236 Oxford St, Lond...	6870	4.20	Shoe_Store	020 7660 4453
Name	Area	Address	Total Reviews	Rating	Category	Phone No

Figure 12: Top 20 Places Map

The selected category in this figure is the shoe stores in London, it can be seen that Nike Town is the most visited shoe store in London with around 7000 visitors in total, the location of the store is shown in the map along with additional information such as the category, price level, phone number, and other relevant information. The bar chart compares the selected store with other stores in the same category and shows how each store compares in terms of total visitors.

The last page of the report shows the popular regions in London, the focus of the visualisations is the total visitors for each region, clustered by location, the category filter is also placed in this page for user interaction, and the bottom panel shows the relevant information of the place clicked on. Fig. 13 below shows the sixth and final page of the dashboard report:



Category

- □ Art_Gallery
- □ Bakery
- □ Bank
- □ Bar
- □ Beauty_Salon
- □ Bicycle_Store
- □ Book_Store
- □ Bowling_Alley
- □ Cafe
- □ Campground
- □ Car_Dealer
- □ Car_Rental
- □ Car_Repair
- □ Casino
- □ Church
- □ City_Hall
- □ Clothing_Store
- □ Convenience_St...
- □ Department_Store
- □ Doctor
- □ Electronics_Store
- □ Finance

Name	Area	Address	Total Reviews	Rating	Category	Phone No
All Saints Church, Ki...	14-16 Market Place, ...	14-16 Market Pl, Kin...	315	4.60	Tourist_Attraction	020 8546 5964

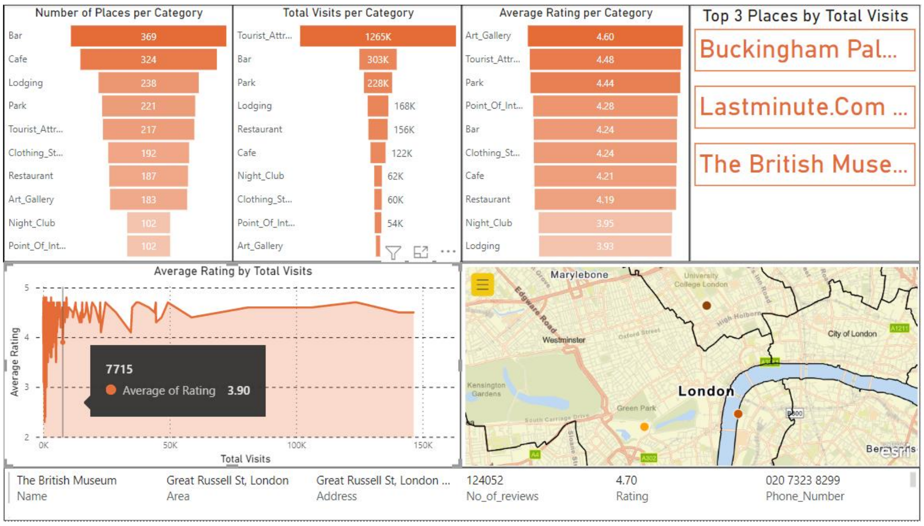
Figure 13: Reviews Clustered Map

As can be seen in the map, the total number of visitors is clustered by location, Green Park for example had around 672 total visitors distributed across several places, the details of these places can be further explored by clicking on the bubble and interacting with the visualisation.

The purpose of this visualisation is to show the most popular regions and boroughs in London represented by a bubble, the size of the bubble depends on the total number of visitors for each area and the colour of the bubble varies with the average rating of the clustered region, the dark green colour represents the cluster that has an average rating of 4.4 and above, and the bright green colour representing the cluster that has an average rating below 2.1, varying in three colours between 2.1 and 3.3, 3.3 and 4.0, and 4.0 and 4.4, with the colours being light green, green, and turquoise, respectively. The colour classification was selected based on the distribution of ratings in the data and the natural breaks between each rating.

The clustered visualisation shown in Fig. 13, can be extremely beneficial to highlight the popular areas in London, which is the main objective for this project, the valuable insights shown in the map represented by the bubble size and the bubble colour make the visualisation more comprehensive and simultaneously simpler to understand, highlighting specific areas in London can be used by location planners and transportation planners to ensure that the popular areas are well serviced and also work on improving the underperforming areas by developing the services provided in the area.

One example of the many interesting insights that can be obtained from the dashboard created can be seen in Fig. 14, it can be observed that the average rating for the places became more stable when the total number of visitors crossed 7715 visitors in total, with the average rating being more stable in the range of 3.9-5 stars, the average rating for the places with less than 7715 total visitors is evidently more arbitrary when compared to the places that have more than 7715 total visitors as highlighted in Fig. 14 below:



The British Museum
Name

Great Russell St, London
Area

Great Russell St, London ...
Address

124052
No_of_reviews

4.70
Rating

020 7323 8299
Phone_Number

Figure 14: Rating Insight