NATIONAL UNIVERSITY OF SINGAPORE

FACULTY OF SCIENCE

AY 2021/2022, SEMESTER 1
DSA3101 : Data Science in Practice

Assignment 1 Report

Group 3:
FRANCELINE BENETTA (A0200662U),
GORDON THAM QI HAO(A0204749Y),
KUEI YU FEI (A0205517L),
LEE MIN LI (A0205279B),
LOW HON ZHENG (A0204803R),
LOW JIA LIN (A0204877W),
SAW LIN MIN (A0205450U),
USMAN HASAN SIDDIQUI (A0201488E)

Specific Contributions:

Data Cleaning & EDA:

KUEI YU FEI, LOW JIA LIN

Cluster Model Codes:

K-PROTOTYPE: LEE MIN LI,
K-MEANS: SAW LIN MIN,
DBSCAN: USMAN HASAN SIDDIQUI

Cluster Analysis:

LEE MIN LI, SAW LIN MIN, USMAN HASAN SIDDIQUI

Sentiment Analysis:

FRANCELINE BENETTA, LOW HON ZHENG

LDA Topic Modelling:

GORDON THAM QI HAO

Presentation Slides:

FRANCELINE BENETTA, GORDON THAM QI HAO, KUEI YU FEI,
LEE MIN LI, LOW HON ZHENG, LOW JIA LIN, SAW LIN MIN,
USMAN HASAN SIDDIQUI

Table of contents

1. Introduction

1.1 Business Idea and Objective

We are a group of investors looking to set up our first luxury hotel in Brussels, Belgium. Our definition of luxury hotels comprises all 4* and 5* hotels in the city. In order to gain sufficient market insight to make informed and strategic business decisions, we will utilise the following methods on a dataset of customer reviews on local hotels in Brussels.

1. Customer segmentation using DBSCAN, K-Means and K-Prototype
2. Sentiment analysis on reviews in each cluster

From the cluster analysis, we hope to come up with specific marketing and pricing strategies to target each customer group (cluster). This will of course be done in the interest of maximising sales. From the aspect-based sentiment analysis on hotel reviews for each cluster found, we hope to discover the positive and negative aspects from customers reviews, and thereafter include the most desirable and optimal aspects, and take caution of any negative facets that might surface in our hotel. Through the aforementioned Unsupervised Machine Learning techniques, we aim to cater to a wider customer base by having various room types catered to each cluster and the proportion of room types to be allocated according to the proportion of each cluster. This will ensure customers have a desirable room type according to their needs in our hotel. By implementing these strategies successfully, we aim to garner more positive reviews and increase the attraction of the hotel.

We hypothesise that most hotels cater to a niche target audience depending on their purpose of travel and budget etc. This is done by hotels aligning their marketing and pricing efforts to target those customers. We aim to disrupt this trend and cater to a variety of target audiences and ensure all their preferences are considered in line with their budget.

For our marketing strategies and pricing recommendation, we aim to gauge it using the clusters formed. We will adjust our pricing based on demands and take seasonality into account to ensure revenue is indeed maximised. With the price set, we can market our services accordingly.

1.2 Data Wrangling

The dataset has 26,386 rows, and 15 columns and is from Data World (Link). This dataset shows reviews on Bookings.com for accommodations in Brussels, Belgium.
The columns are `review_title, reviewed_at, reviewed_by, images, crawled_at, url, hotel_name, hotel_url, avg_rating, nationality, rating, review_text, raw_review_text, tags, meta.`

We dropped the following columns that are redundant for our analysis, namely `images`, `crawled_at, url, hotel_url, meta`, and renamed some of our columns for ease of usage when conducting our analysis.

After checking for missing values, we found 1 missing value in `review_title` and 16 missing values in `nationality`. Rows with missing values and duplicates were removed.

A new column named `hotel_type` was added, where reviews of accommodation are classified as 'Hotel', 'Hostel', 'Motel', 'Bed & Breakfast', 'Guest House', 'Apartment', 'Chalet' or 'Others', based on `hotel_name`.

The 'tags' column contains useful information and was split into 4 different columns, `nights_stayed`, `room_type`, `traveller_type`, `trip_type`.

We then selected 20 hotels that are at least 4 stars from Bookings.com as our goal is to target luxury hotels. These 20 hotels consist of 10 hotels that were highest rated by reviewers and 10 hotels that were rated the lowest by reviewers. A new column named `price_per_night` was added [Assumption: we retrieved prices from the same website, Bookings.com], which is the price per night of a room type at each hotel. A new column named `total_spending` was added, which is the amount spent by a reviewer, calculated by multiplying `price_per_night` and `nights_stayed`.

In `review_text`, we noticed that there was the presence of escape characters (i.e. '\n') and carriage returns (i.e. '\r').

```
'none,\n\nRoom, location, breakfast - everything according to prior expectations.'
```

Figure 1. Display of the presence of the "\n" character

Such an example can be seen from Figure 1 above. A reason for this is possibly because the raw text data wasn't converted properly when the dataset was first created. To solve this issue, we replaced such instances with a blank space.

1.3 Feature Engineering

Column Added	Methodology	Assumption
`price_per_night`	The average price of each hotel from booking.com[1] for each quarter of the year.	• Source used was accurate. • Price does not vary significantly in each quarter. • Due to lack of historical data, we assume variance across each season in 2018-2021 is the same as 2021
`total_spending`	`Price_per_night * nights_stayed`	-
`booking_freq`	Groupby nationality and reviewed_by to get the frequency of booking of each reviewer	• People with the same name and nationality are the same person
`hotel_type`	Simple regex from `hotel_names` to extract the type of accommodation	• Most hotel names contained the type of accommodation and were accurate
`quarter`	Split the `reviewed_at` consisting of the date of a review into 4 quarters of the year	• Customer reviews were written soon after they have stayed in their respective accommodation
`nights_stayed`	The original dataset has a `tag` column that contained this information [2]	-
`room_type`		-
`traveller_type`		-
`trip_type`		-

Table 1. Description of new features engineered

[1] Same as the source of the original dataset
[2] Used regex to split tag into these predictors/columns

In conclusion, our cleaned dataset consists of the following columns, with 1,980 rows and 16 columns:

- `review_title`
- `review_date`
- `reviewer`
- `hotel_name`
- `hotel_avg_rating`
- `nationality`
- `rating`
- `review_text`
- `raw_review_text`
- `hotel_type`
- `nights_stayed`
- `room_type`
- `traveller_type`
- `price_per_night`
- `total_spending`
- `booking_freq`
- `quarter`

1.4 Exploratory Data Analysis

The data types of all columns are as shown below (Figure 2). There are 10 categorical variables, 6 numerical variables, and 1 DateTime variable.

```
review_title                 object
review_date          datetime64[ns]
reviewer                     object
hotel_name                   object
hotel_avg_rating            float64
nationality                  object
rating                      float64
review_text                  object
raw_review_text              object
quarter                    category
hotel_type                   object
nights_stayed                 int64
room_type                    object
traveller_type               object
price_per_night               int64
total_spending                int32
booking_freq                float64
dtype: object
```

Figure 2. Data Types of columns

	hotel_avg_rating	rating	nights_stayed	price_per_night	total_spending	booking_freq
count	1980.000000	1980.000000	1980.000000	1980.000000	1980.000000	1980.00000
mean	8.592980	8.677879	1.892929	156.174242	296.438889	1.49596
std	0.393461	1.482277	1.185925	62.524032	237.237301	1.45628
min	7.800000	1.000000	1.000000	84.000000	84.000000	1.00000
25%	8.400000	8.000000	1.000000	108.000000	143.000000	1.00000
50%	8.500000	9.000000	2.000000	143.000000	222.000000	1.00000
75%	8.900000	10.000000	2.000000	178.500000	356.000000	1.00000
max	9.500000	10.000000	16.000000	1008.000000	3024.000000	11.00000

Figure 3. Descriptive statistics of the numerical variables.

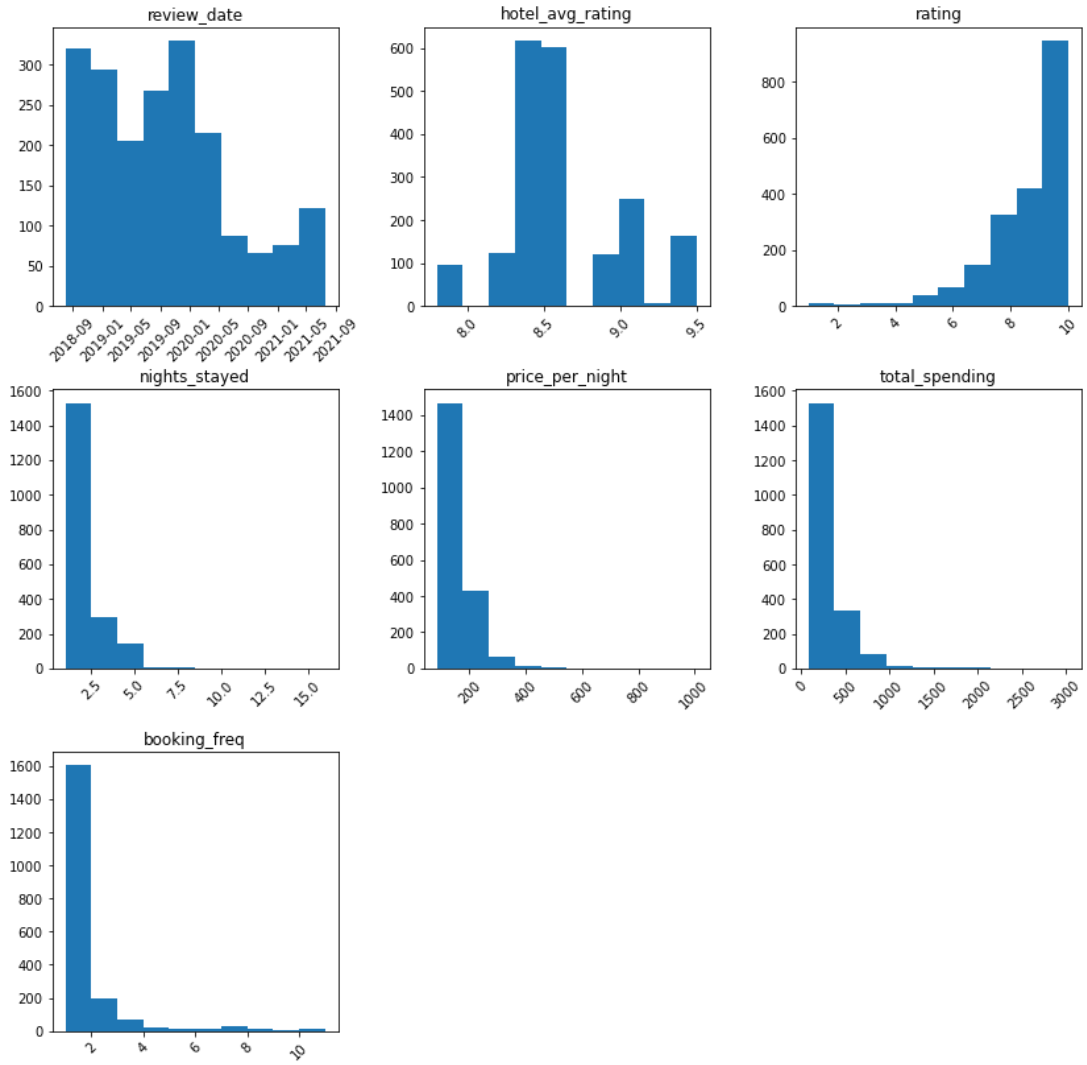


Figure 4. Histograms of the numerical variables, where numbers are grouped into bins and the height of a bar shows how many numbers are in each bin

The review dates span from July 2018 to July 2021, hence the data collected is recent and up-to-date. Looking at the histogram of the various hotels' average ratings, it is ensured that the hotels that we have selected are all above 7.5, hence we are able to derive credible insights from established and renown hotels. In addition, from the histogram of reviewers' rating, it can be deduced that most of the reviews are positive, given that a large majority of reviews have a rating of 8 and above. This is a reasonable occurrence as it is proven that most people tend to feel happier when travelling abroad, hence they might feel satisfied more easily when overseas. Hence, we might have to focus more on the positive aspects of our competitors, and be sure to include them in order to attract more customers. Furthermore, from the histogram of nights stayed, we can conclude that most reviewers in our dataset were short-term visitors, and were not there for extended stays. Thus, this group of customers fits our target audience.

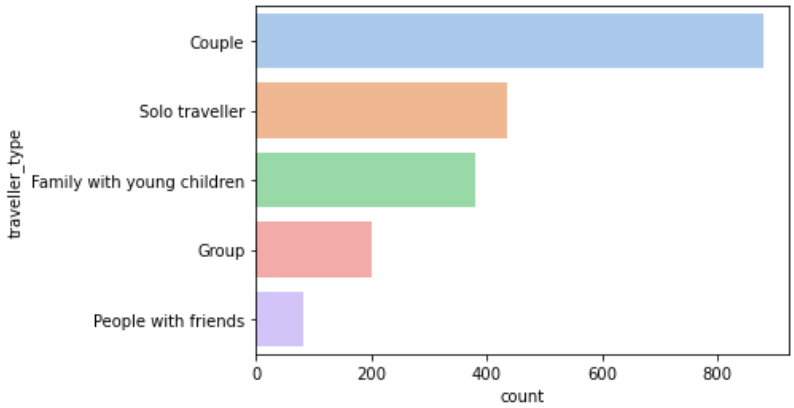


Figure 5. Bar Plot of 'traveller_type'. Each bar represents a unique class, and the x-axis is the count of occurrences of a class.

From Figure 5, we can tell that the most common traveller type among the reviewers is couples, followed by solo traveller, and then family with young children. A possible strategy we can derive from this data is that we can provide exclusive couple deals at our amenities, such as a special promotion for a massage for two at our spa or parlour, or provide extra security for solo travellers, especially for young women.

1.4.1 Perceptual Map

A perceptual map is a visual representation of the perceptions of customers or potential customers about specific attributes of a brand. In this particular case, we decided to create a competitor perceptual map which depicts where our competitors stand in terms of average hotel prices and average hotel ratings.

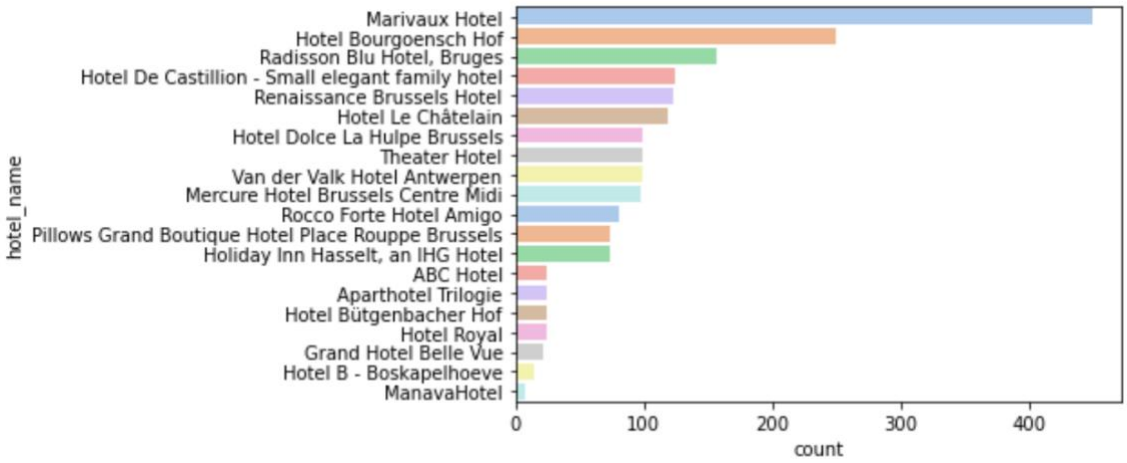


Figure 6. Plot of categorical variable 'hotel_name'. Each bar represents a unique class, and the x-axis is the count of occurrences of a class.

From Figure 6, we are able to determine the sizes of each competitor on the map. The higher the number of reviews for a certain hotel, the larger the size of the hotel will be on the perceptual map. The hotel variables are arranged in descending order.

	hotel_name	avg_price_per_night	hotel_avg_rating
0	Marivaux Hotel	110.250000	8.5
128	Hotel Bourgoensch Hof	172.800000	8.4
309	Hotel Royal	202.000000	9.0
449	ManavaHotel	109.500000	9.2
452	Hotel Dolce La Hulpe Brussels	162.250000	8.4
536	Pillows Grand Boutique Hotel Place Rouppe Brus...	155.000000	9.0
876	Grand Hotel Belle Vue	241.333333	9.0
923	Hotel De Castillion - Small elegant family hotel	260.571429	9.5
997	Mercure Hotel Brussels Centre Midi	128.000000	7.8
1100	ABC Hotel	228.000000	9.5
1168	Hotel Le Châtelain	258.000000	8.9
1221	Theater Hotel	123.500000	8.4
1386	Van der Valk Hotel Antwerpen	172.400000	8.4
1435	Hotel B - Boskapelhoeve	212.500000	9.4
1500	Holiday Inn Hasselt, an IHG Hotel	209.166667	8.4
1573	Aparthotel Trilogie	201.666667	9.1
1597	Radisson Blu Hotel, Bruges	200.000000	8.5
1650	Renaissance Brussels Hotel	177.000000	8.2
1876	Hotel Bütgenbacher Hof	248.800000	9.1
1900	Rocco Forte Hotel Amigo	550.600000	9.1

Figure 7. Table of average price per night and average ratings of 20 shortlisted hotels

From Figure 7, we then extracted each hotel's overall average rating, and the average price each hotel charges by taking the mean of the prices of the different room types for each hotel.

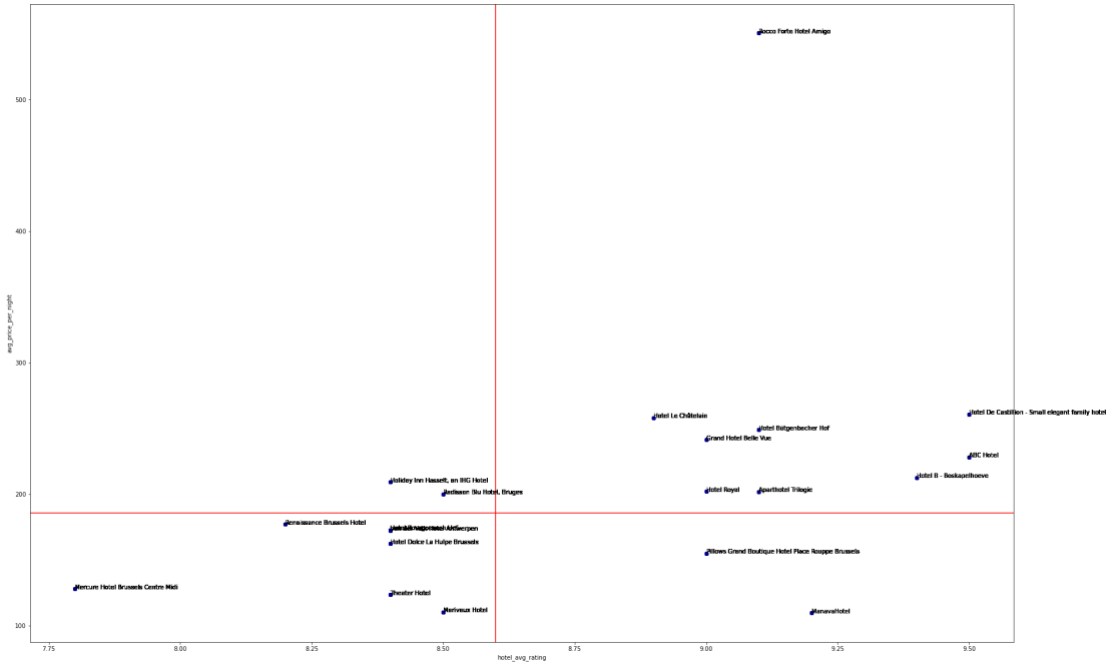


Figure 8. Perceptual map of 20 hotels

A plot of `hotel_avg_rating` and `avg_price_per_night` was created. The means of `hotel_avg_rating` and `avg_price_per_night` were used to determine the axis for the perceptual map. In order to create a more precise version of our map (Figure 9), the sizes of each hotel were incorporated, where the size represents the number of review counts for each hotel.

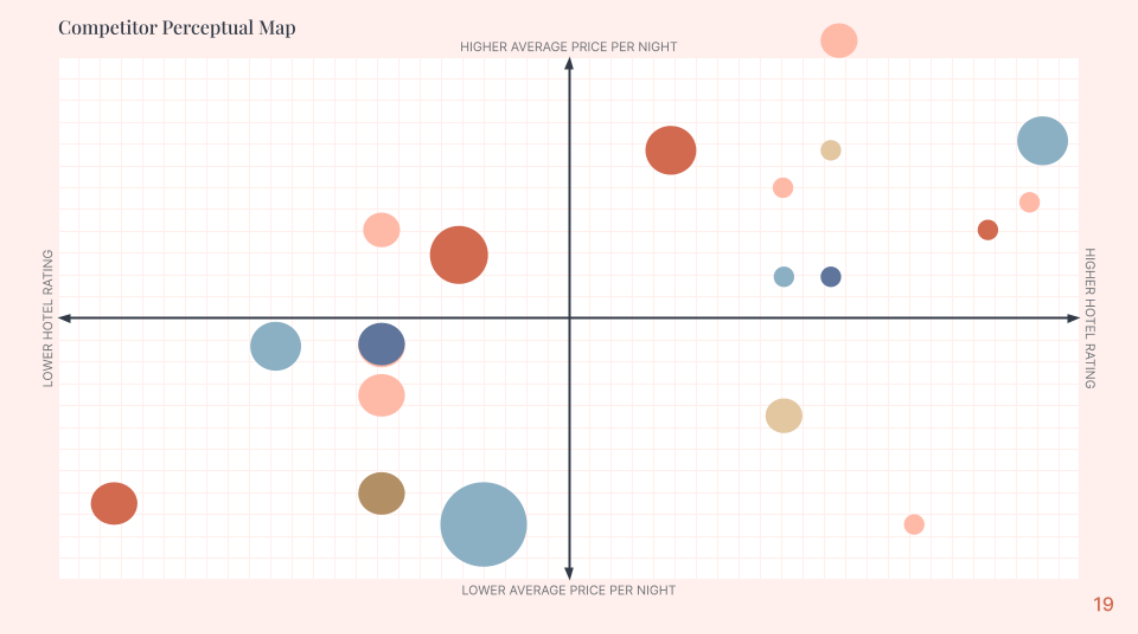


Figure 9. Final perceptual map of 20 hotels

2. Cluster Models

2.1 Further cleaning of the dataset

We realised that our dataset seems to be unbalanced with regards to the `traveller_type` variable. As such, we decided to recategorize them into the following:

Before Recategorizing		After Recategorizing	
`traveller_type`	Count	`traveller_type`	Count
Solo	435	Solo	435
Couple	880	Couple	880
Family with young children	381	Group	665
Travellers with friends	83		
Group Travellers	201		

Table 2. Recategorizing of traveller_type column

Afterwhich, we took 400 random samples each from the 3 distinct `traveller_type` to get a balanced dataset for clustering.

2.2 Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is a density-based clustering that identifies clusters in a large spatial dataset by looking at the local density of the data points. We conducted DBSCAN on customers' `nationality`, `rating` of the hotel, a `quarter` of the year, `nights_stayed`, `room_type`, `traveller_type`, `booking_freq` and `price_per_night`. As the algorithm only works on continuous columns, we applied a label encoder on categorical columns to represent them in numbers. We tried a density-based clustering algorithm to make the clusters more accurate and meaningful.

By aggregation depending on the density of points near the chosen centroids. This model is robust for numerical variables, doesn't need the number of clusters to be specified prior and can deal with arbitrary shaped clusters

After encoding categorical variables and standardising the continuous variables. We used the below k distance graph from sklearn to find the optimal value of eps, which was found

to be 0.035 (also from model iterations) and chose min_samples to be 9 according to the criteria that min_samples >= D+1 where D is the dimensions of the data.

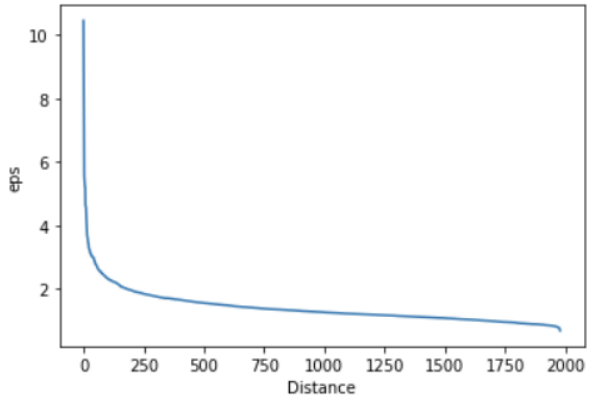


Figure 10. DBScan - Selecting eps value

2.2.1 Results

Finally, from the above clusters plot here, we find that there are two clusters formed (-1,0) but they are not clear in distinction, but the majority of the points are in one cluster (as per the cluster_id column in the plot below). The variable `price_per_night` is the one with the most distinction between the clusters.

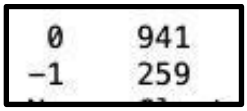


Figure 11. Cluster Segmentation

Next, using the eps and min_samples parameters, the DBScan model was run *(code: (Ren, 2019))* on a standardised sample of 400 observations randomly selected using the 'sample' function of the 'random' package.

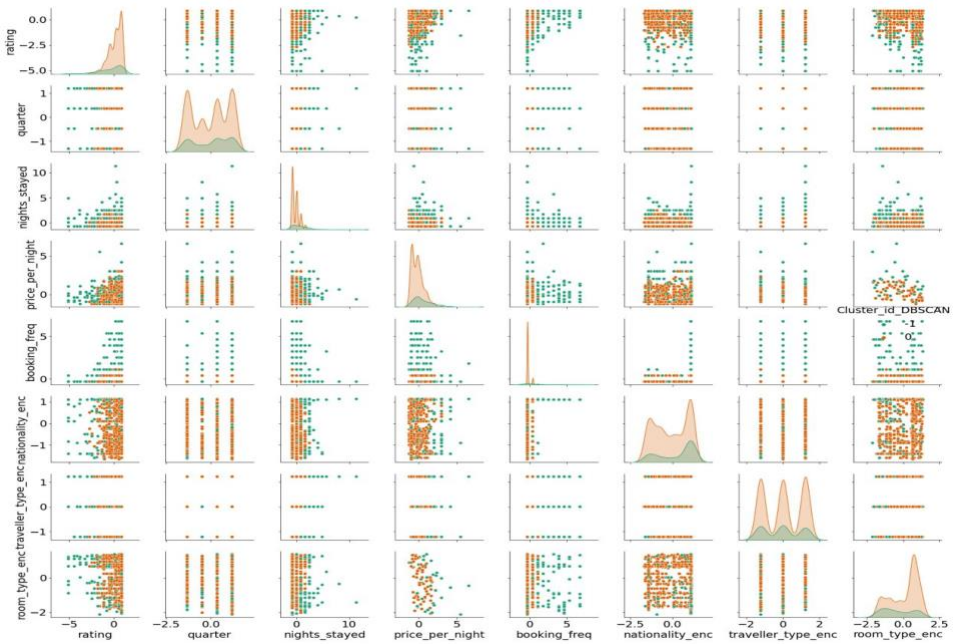


Figure 12. DBScan results

2.2.2 Conclusion

From Figure 12, there were no apparent or distinct shapes that could be identified. Also, since the above clustering algorithm can only handle numerical variables, it proves to be insufficient since a categorical variable needs to be included in order to effectively explore the research questions. Hence, we proceed to explore with the k-means model and k-prototype clustering model.

2.3 K-Means

K-Means is a simple and straightforward method to define clusters using numerical attributes.

Before fitting the model with our dataset, we begin by finding out if variables are correlated with one another. If the variables are highly correlated, we should only choose one of them to represent the attribute.

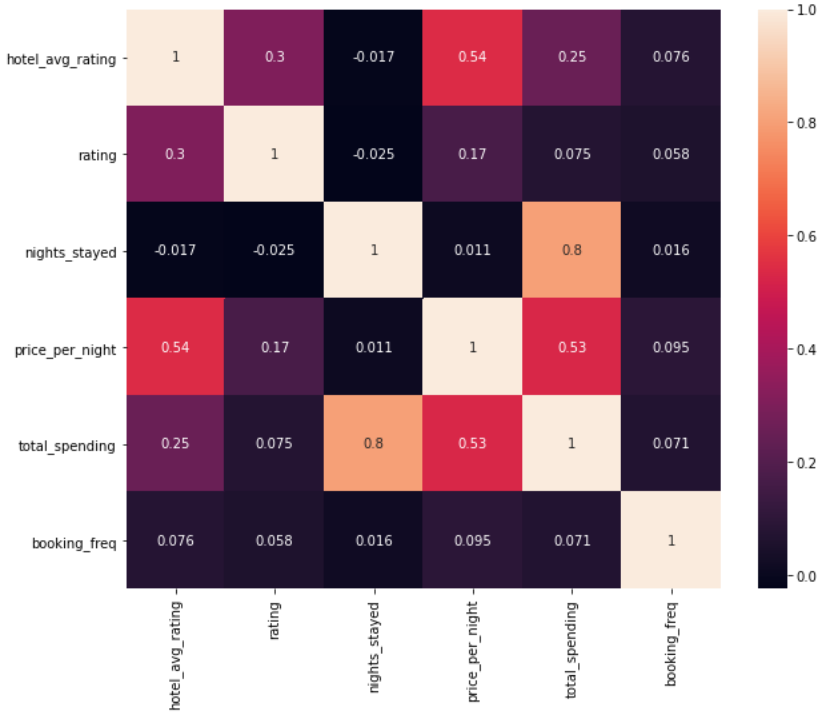


Figure 13. Heatmap of correlation of variables

From the heatmap (Figure 13), we can see that correlation values between `price_per_night` and `hotel_avg_rating`, `nights_stayed` and `total_spending` is quite high so we have to drop columns that represent similar attributes. Since `total_spending` is derived from `price_per_night` and `nights_stayed`, we decided to use `total_spending` to represent it. `hotel_avg_rating` is the mean of ratings of all

reviews for the hotel and hotels with the same `hotel_name` will have the same value. `rating` is given by each reviewer to the hotel based on their satisfaction with the hotel. Since both features are similar, we decided to use `rating` as it has more variation.

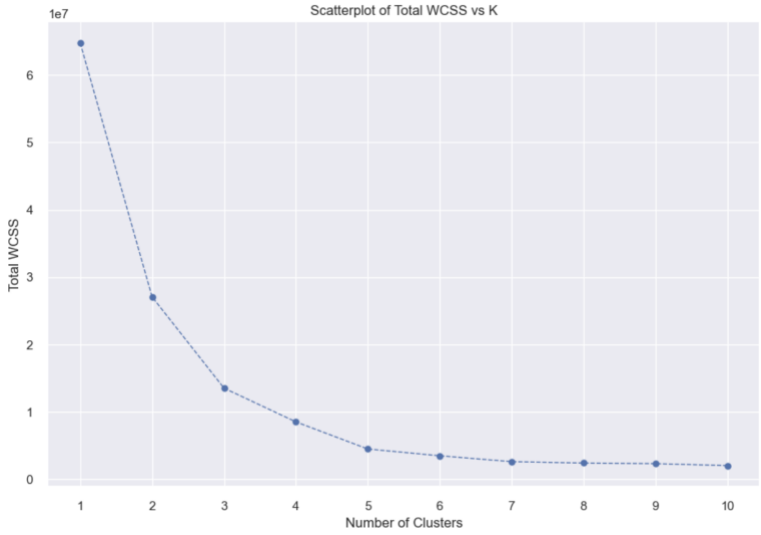


Figure 14. WCSS Graph for K-Means

Within Cluster Sum of Squares (WCSS) measures the variability of observations within each cluster. To determine the appropriate number of clusters to form, we plot Total WCSS vs K (Figure 14) and locate the elbow point of the graph to get k = (2, 3) where total WCSS is reasonably minimized.

We conducted K-means clustering to cluster the dataset based on the 3 distinct continuous columns which are `total_spending` by each customer, `booking_freq` of each customer, `rating` by reviewers to get 2 and 3 clusters. To decide on the optimum number of clusters, we compare the silhouette score and their segmentation of customers.

2.3.1 Results

K = 2:

Clusters	Total Spending	Booking Frequency	Rating	Total
0	227.193732	1.457740	8.648433	1053
1	767.605442	1.394558	8.844218	147

Figure 15. Cluster results for K = 2

Silhouette score: 0.7130897076164787

We then inferred the following customer profiles:
Cluster 0: customers with low spending at lower-rated hotels
Cluster 1: customers with high spending at higher-rated hotels

K = 3:

Clusters	Total Spending	Booking Frequency	Rating	Total
0	190.040276	1.418872	8.626122	869
1	481.795222	1.515358	8.746758	293
2	1204.263158	1.657895	9.157895	38

Figure 16. Cluster result for K = 3

Silhouette score: 0.6321047936501327

We then inferred the following customer profiles:
Cluster 0: customers who have the least spending at lowest-rated hotels
Cluster 1: customers who have medium spending
Cluster 2: customers who have the highest spending at highest rated hotels

2.3.2 Conclusion

Even though the silhouette score for K = 3 is lower than K = 2, we are able to identify 3 types of customers who have different levels of spending from low, medium to high which is more insightful than having 2 clusters. Hence, we decided to proceed with the model with 3 clusters.

For K = 3, the majority of the customers belong to the lower spending tier with only 3.17% of customers having the highest total spending. It is evident that the customer segmentation is unbalanced and not distinct, and the `booking_freq` is not helpful at all since it has similar values for all 3 clusters. The only distinctive variable is the `total_spending` which is not enough for us to make further analysis. The ANOVA test results revealed that only the `total_spending` variable is significant with a p-value of 0.0. Therefore, this model is not ideal in helping us to achieve our business objective.

2.4 K-Prototype

In order to make full use of our dataset that consists of categorical and numerical data, we employed the K-Prototype model. It combines K-Modes and K-means to handle discrete variables and continuous variables simultaneously.
The K-Prototype algorithm divides the dataset into different subclusters to minimize the value of the Cost Function. The Cost Function is shown in the following formula:

$$F(U,Q) = \sum_{l=1}^{k}\sum_{i=1}^{n} u_{il} d(x_i, q_l)$$

Figure 17. Cost function formula for K-Prototype

The K-prototype algorithm combines the “means” of the numerical part and the “modes” of the categorical part to build a new hybrid Cluster Center “prototype”. On the basis of “prototype,” it builds a Dissimilarity Coefficient formula and the Cost Function applicable to the mixed-type data.

The parameter γ is introduced to control the influence of the Categorical Feature and the Numerical Feature on the clustering process. It is assumed that the mixed-type dataset has Numerical Feature and Categorical Feature. For any , the definition of the Dissimilarity Coefficient of k-prototypes is shown in the following formula:

$$d(x_i, q_l) = \gamma \sum_{s=1}^{p} \delta\left(x_{i,s}^{C} - q_{l,s}^{C}\right) + \sum_{s=p+1}^{m} \sqrt{\left(x_{i,s}^{N} - q_{l,s}^{N}\right)^2}$$

$$\text{where } \delta(x_{i,s}, q_{l,s}) = \begin{cases} 0, x_{i,s} = q_{l,s} \\ 1, x_{i,s} \neq q_{l,s} \end{cases}$$

Figure 18. Dissimilarity coefficient for K-Prototype

When deciding on what variables to use, we omitted `booking_freq` as most of the values are 0. Also, since `total_spending` = `price_per_night * nights_stayed` and `price_per_night` are linearly dependent, so we decided to use `price_per_night` as it would be useful for our pricing strategy. `Trip_type` was not used because it was not very informative and most rows did not have a `trip_type`. `room_type` was also not used because every hotel has different names for similar types of rooms eg. 'Deluxe Twin Room' vs 'Deluxe Two Beds'. Therefore, we selected `traveller_type`, `quarter`, `price_per_night` and `rating` as the variables for the model.

Therefore, to get the appropriate number of clusters to form, we plot the Cost vs K (Figure 19) and locate the elbow point of the graph at k = (3,4) which is also the cost-minimizing point.

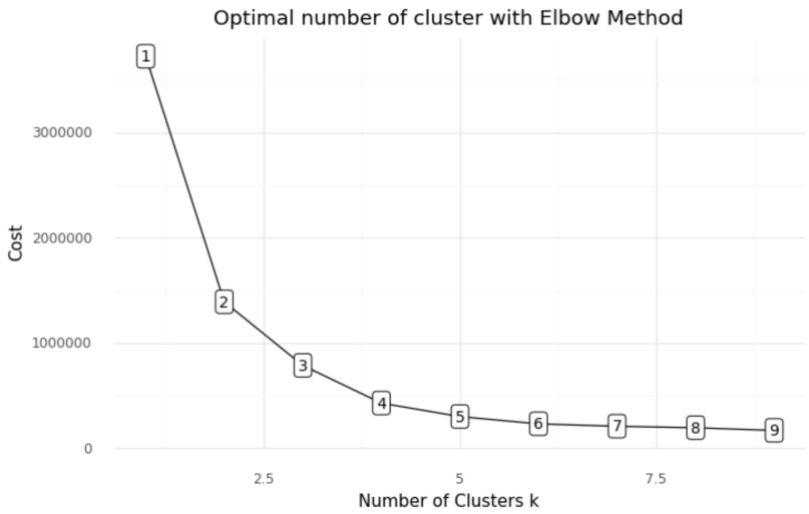


Figure 19. K-Prototype: Elbow Plot

We conducted K-Prototype clustering on `traveller_type, quarter, price_per_night and rating` to get 3 and 4 clusters. To decide on the optimal number of clusters, we ran both models to compare their silhouette scores and their segmentation of customers.

2.4.1 Results

K = 3:

	Cluster	Total	traveller_type	quarter	price_per_night	rating
0	0	102	Group	4	283.667	9.296
1	1	382	Group	1	182.144	8.790
2	2	716	Solo traveller	4	118.251	8.521

Figure 20. K-Prototype: Cluster results for K=3

Silhouette score: 0.1520792725244602

We then inferred the following customer profiles:
Cluster 0: Group travellers who travel in the 4th quarter of the year, staying in high tier rooms
Cluster 1: Group travellers who travel in the 1st quarter of the year, staying in mid-tier rooms.
Cluster 2: Solo travellers who travel in the 4th quarter of the year, staying in low-tier rooms.

K = 4:

	Cluster	Total	traveller_type	quarter	price_per_night	rating
0	0	208	Group	3	211.125	8.951
1	1	435	Solo traveller	1	103.814	8.431
2	2	492	Couple	4	150.807	8.662
3	3	65	Couple	4	306.323	9.474

Figure 21. K-Prototype - Cluster results for K=4

Silhouette score: 0.18921947456225274

We then inferred the following customer profiles:

Cluster 0: Group travellers who travel in the 3rd quarter of the year, staying in medium-high tier rooms
Cluster 1: Solo travellers who travel in 1st quarter of the year and stays in the cheap room
Cluster 2: Couple who travels in 4th quarter of the year and stay in low tier rooms
Cluster 3: Couple who travels in 4th quarter of the year and stay in expensive rooms

2.4.2 Conclusion

The silhouette score for 4 clusters was higher than that of 3 clusters and the customer segmentation appears better, as there was further segregation between high and low spending ‘Couples’. As such, we decided to proceed with 4 clusters.

From the customer profile, we are able to see the distinct segmentation of customers from the quarter of the year, type of traveller and the price of the room they choose to stay in. The model with 4 clusters is ideal for us to make further analysis and do customer segmentation.

2.5 Most appropriate model

Out of all 3 models, K-Prototype performs best and forms the most distinct and meaningful clusters for us to do our customer segmentation analysis. Hence, we chose it as our model of choice for further analysis.

3. Analysis of cluster results

3.1 2D Analysis between variable and clusters

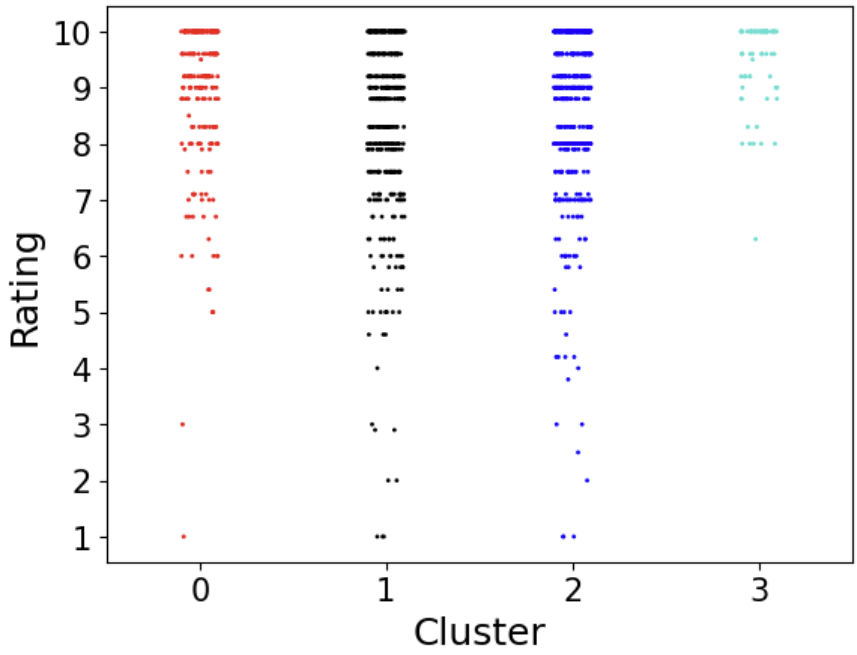


Figure 22. Plot of Rating vs Cluster

As seen from Figure 22, Cluster 3 consists of customers who go only for the high rated hotels and this information is also consistent with the customer profile. We believe that since there are couples on a high budget, they are willing to spend more money for enjoyment and luxurious accommodation. It is also evident that most of the customers with a medium budget in Cluster 0 do not stay in hotels that are lowly rated.

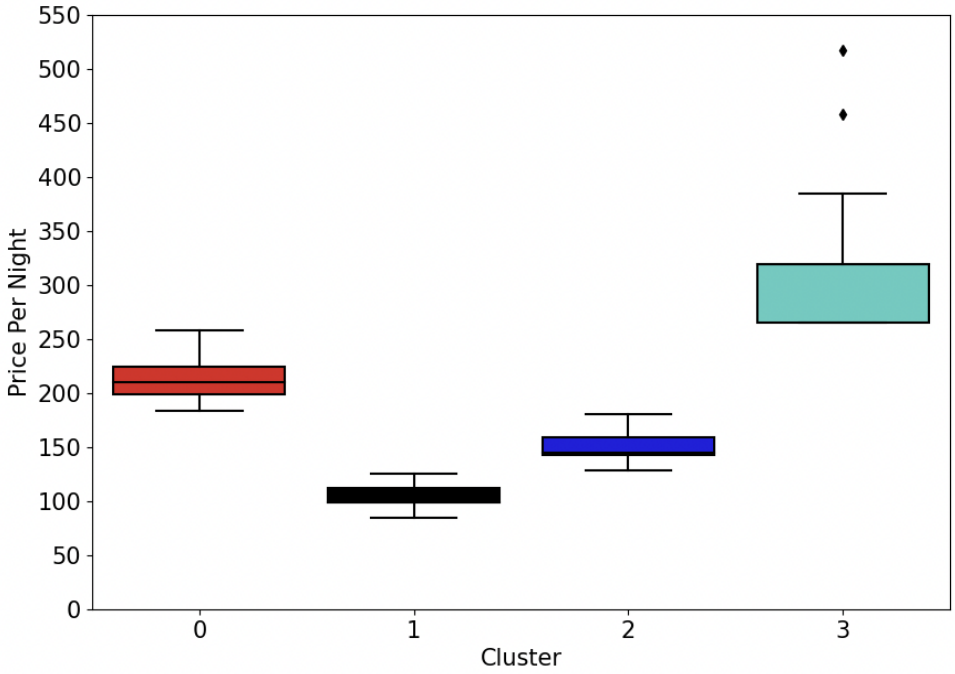


Figure 23. Boxplot of Price Per Night against Cluster

In Figure 23, we can see that the clusters are well-separated by the `price_per_night`, suggesting that it plays a significant role in the clustering model. Furthermore, we can see that most of the customers belong to Cluster 1 and Cluster 2 so it will be appropriate to target them to garner more business and revenue by adjusting the pricing policies of our hotel's rooms. In addition, there seems to be an overlap between Cluster 1 and Cluster 3. We can attempt to convert customers in Cluster 1 to stay in more expensive rooms for a better experience.

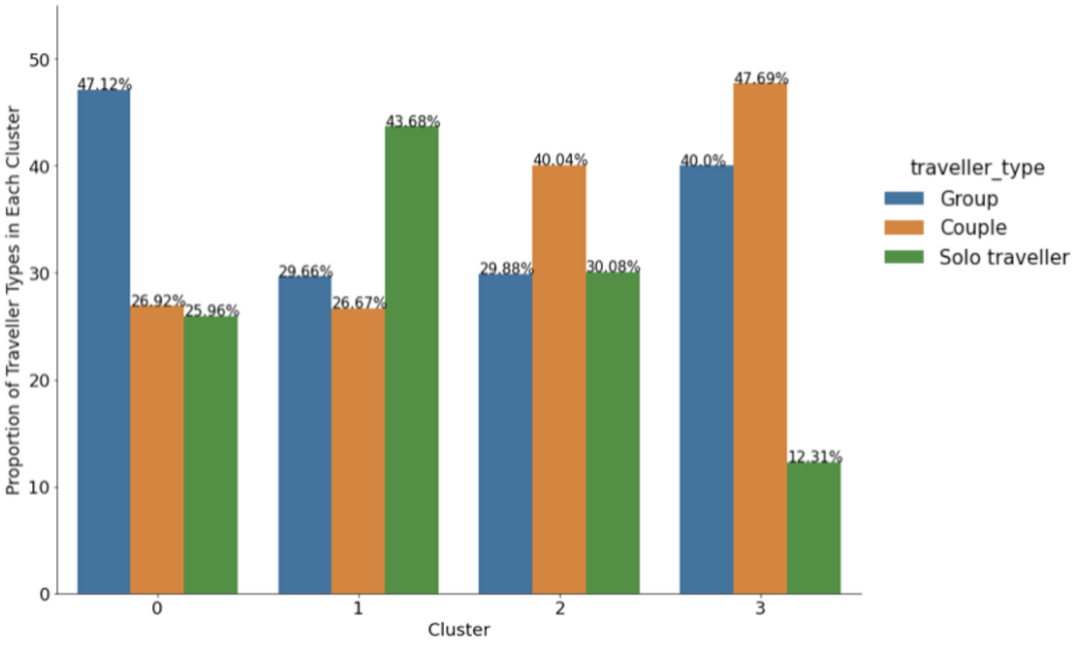


Figure 24. Bar Plot of Proportion of Traveller Types in Each Cluster

From the bar graph above (Figure 24), we can tell the proportion of the majority `traveller_type` in each cluster. This is useful for us in our hotel planning as on top of catering for the majority `traveller_type` in the cluster, this diagram also informs us on the proportion of the other types of traveller that we have to keep in mind and serve as well. cluster 3 and 4 are the 2 clusters we need to keep in mind as the proportions of all Solo and Group travellers are relatively high within its own cluster.

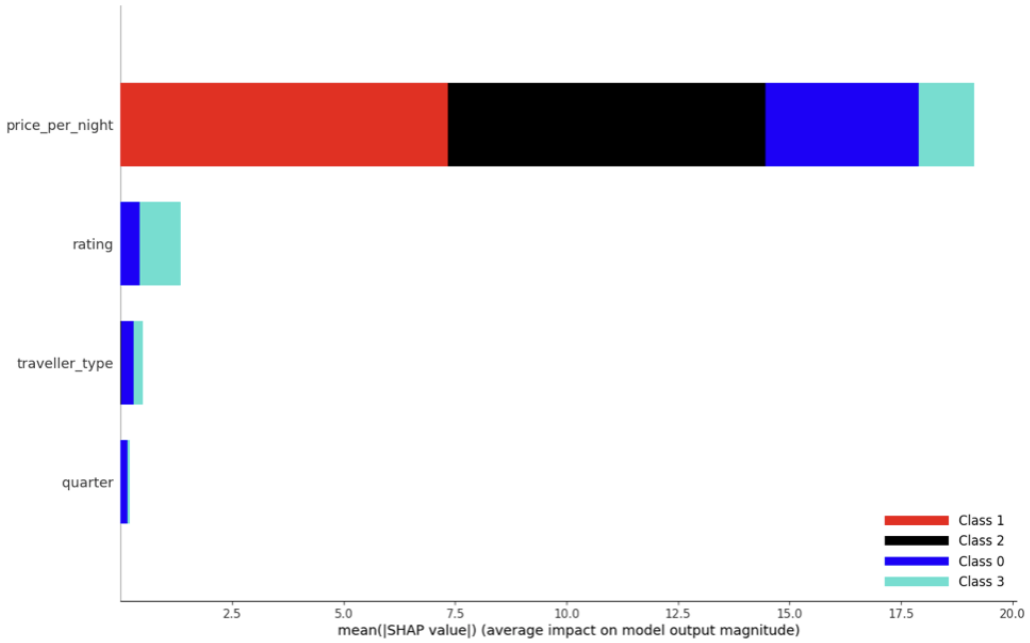


Figure 25. Feature importance of clusters using SHAP

SHAP (SHapley Additive exPlanations) explain the prediction of an instance x by computing the contribution of each feature to the prediction. The SHAP explanation method computes Shapley values from coalitional game theory.

This graph shows the contribution of each variable to the cluster formed and the variables are ordered according to their significance in descending order. As we can see from Figure 23 and Figure 25, the `price_per_night` variable is the most distinct variable in generating well-divided clusters and contributes the most to the formation of clusters. Hence, it is reasonable that it is the most significant variable.

3.2 Niche Hypothesis

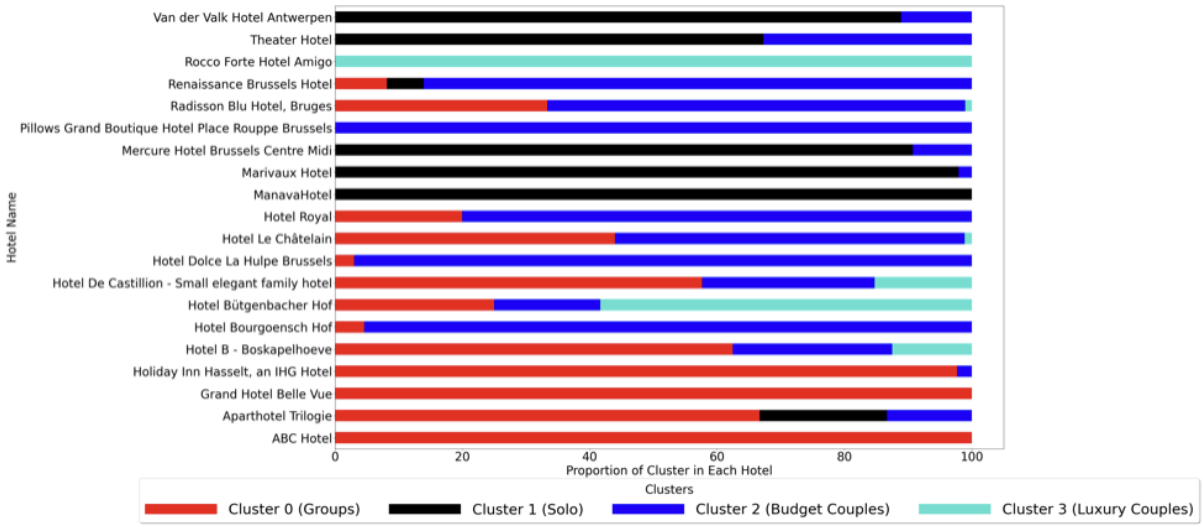


Figure 26. Proportion of Cluster in each hotel

To support our hypothesis of every hotel targeting a niche target audience, we conducted some analysis on our clusters. As shown in Figure 26, each hotel tends to target a specific group of travelers (i.e. one colour dominating the entire bar) such as ABC Hotel catering to only Group travellers while Pillow Grand Boutique Hotel Place Rouppe Brussels caters to Couples on a high budget. Hence, we feel that there is a gap in the market and the goal of our hotel is to fill this gap by catering to all audiences, while maintaining their preferences and budget

3.3 3D Visualisation between variables and clusters

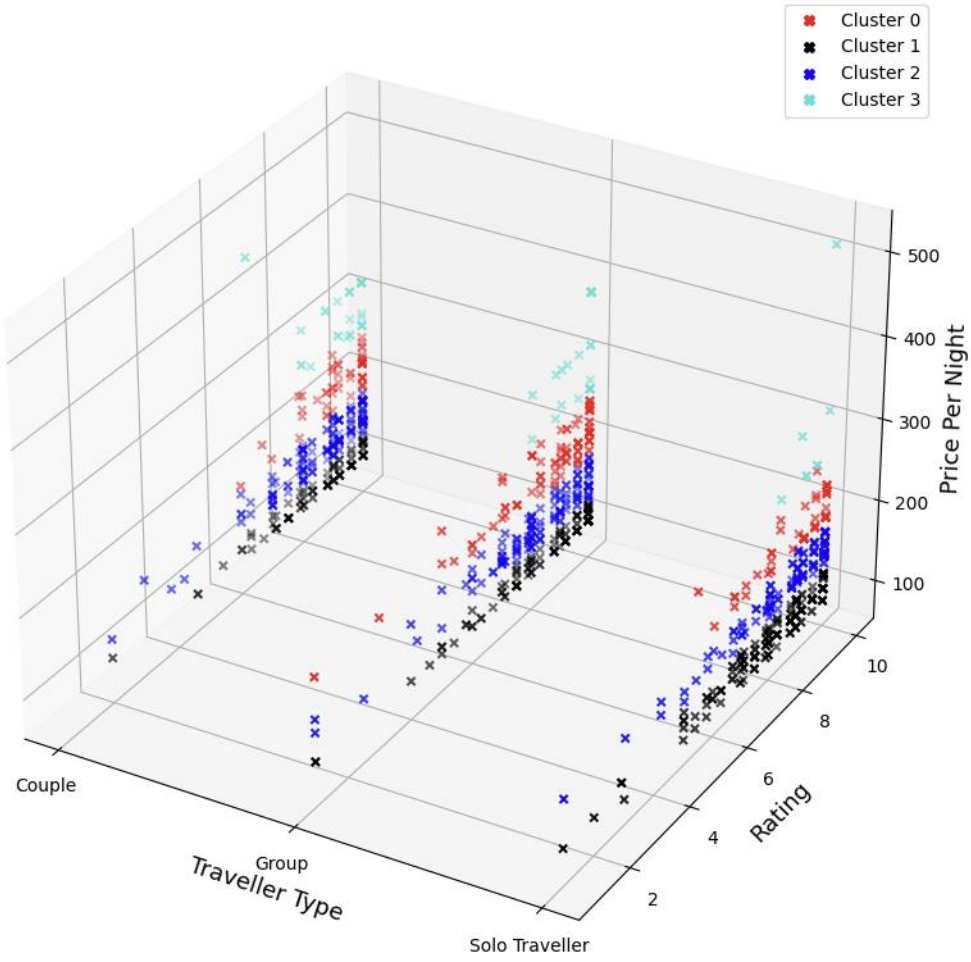


Figure 27. 3D Visualisation of Price Per Night, Traveller Type and Rating

From the 3D visualisation (Figure 27), the clusters were very apparent especially for the variable price_per_night, highlighted in the above 2D plot (Figure 23), where we can see distinct clusters according to how much customers are willing to spend in their respective clusters. We can also infer that price_per_night seems to be proportional to rating and we can tell that high spending couples (Light green) had great satisfaction from their stay since ratings were generally high (≅10).

From the above visualisations, we are able to generalise the clusters into the following customers profiles:

Cluster 0: Group travellers with medium budget

Cluster 1: Solo travellers with low budget
Cluster 2: Couple with low budget
Cluster 3: Couple with high budget

4. Sentiment Analysis

4.1 Why sentiment analysis?

Sentiment analysis allows us to unlock more business insights into the preferences/dislikes of hotel guests. This would provide us with valuable information in terms of what features and facilities we should or should not have to cater to the mass public.

4.2 Preliminary cleaning of review_text

As text data is in natural human format, in sentences or in paragraphs with the occasional spelling error or mistake it is important for us to break the text down into a format the computer can easily understand. To do this, we have to carry out cleaning on the review_text.

Apart from the removal of "\n" and "\r" shared in Chapter 1.2, further exploration highlighted that there were certain rows of reviews with the entry of "There are no comments available for this review". Reviews with this value were removed so as to only retain meaningful and factual reviews for our analysis. Furthermore, to reduce possible errors that might occur in our analysis, we converted the reviews to lower case.

4.2.1 Spelling errors

Another noticeable issue from our exploration is the presence of spelling mistakes in some of the reviews. This issue was relatively easy to resolve and after trying out different packages available in Python such as *pyspellchecker* and the autocorrect library, we concluded that the autocorrect library performed better. The autocorrect library corrects words by using it's Speller class. This Speller class assigns a score based on how close the input word/sentence is from another word in a list of possible words. The highest scoring word would then be used to correct any potential spelling errors in the text. A demonstration of this can be seen in the following figure.

```
docx = ['calandar', 'lightening', 'misspel', 'neccessary', 'bussiness', 'recieve', 'adress']
for word in docx:
    print(f'{word}:{check.get_candidates(word)}')
```

```
calandar:[(126400, 'calendar')]
lightening:[(31395, 'lightning'), (3850, 'tightening')]
misspel:[(3878, 'mussel'), (18785, 'misses'), (115214, 'missed'), (2614, 'dispel'), (100557, 'missile')]
neccessary:[(355501, 'necessary')]
bussiness:[(1048453, 'business')]
recieve:[(228363, 'receive'), (16438, 'relieve'), (3845, 'recieved')]
adress:[(518789, 'address'), (80605, 'dress')]
```

Figure 28. Demonstration of the use of the Speller class

As seen from Figure 28, the variable docx consists of words with spelling errors. For example, using the get_candidate attribute in the Speller class, *"lightening"* would be deemed to be close to the words *"lightning"* and *"tightening"*. Since the word *"lightning"* has a higher score due to its closer similarity to *"lightening"*, the original word *"lightening"* would be corrected to *"lightning"*. This is an important step to ensure that our analysis would be representative of the data.

4.2.2 Splitting of reviews

Taking into consideration that there might be different sentiments or topics mentioned in the same review, we decided to split up the reviews into a list of sentences for each review. Here our assumption is that the information within the same sentence would all relate to the same topic. From our research, we noticed that the sent_tokenize function from the NLTK library will be able to split the reviews up into sentences. It is able to provide a regular expression based tokenizer that splits reviews into sentences using an unsupervised algorithm to build a model for abbreviation words, collocations, and words that start sentences and then use that model to find sentence boundaries.

Next, we moved on to the second phase of cleaning these sentences to make them optimal for sentiment analysis, with the following steps:

1. The sentences are split into individual words. We did this by first splitting the sentences by any punctuations and then further splitting each of these phrases by whitespace.
2. We removed unhelpful parts of data, or noise, by removing words that contained numbers (i.e. 0-9). These steps help us to clean up the data by removing unwanted punctuations as well as words that may hinder the analysis of the sentiments.
3. Stopwords were also removed since keywords are more important than general terms in sentiment analysis. Apart from that, we decided to lemmatize the words to convert them to their base form. This helped us reduce the number of overall terms to certain "root" terms. We executed this using the Wordnet Lemmatizer package available in python. Attaching the "part-of-speech" tags for each token, the Wordnet Lemmatizer from the NLTK library was able to accurately convert the words to their base form.

4. Words of length less than or equal to one were also removed to improve the quality of data before we merged the words back into sentences.

```
1 clean_text_for_sentiment("the chefs are cooking an1 extremely delicious meal!")
```

```
'chef cook extremely delicious meal'
```

Figure 29. Demonstration of phase 2 of cleaning

Taking a look at Figure 29, an example of the second phase of cleaning can be seen. Feeding the sentence "the chefs are cooking an1 extremely delicious meal!" into the cleaning function, the sentence is first split into words *["the", "chefs", "are", "cooking", "an1", "extremely", "delicious", "meal", ""]*. Next, those words containing numbers would be removed to reduce noise in the data. This would give us an updated list of *["the", "chefs", "are", "cooking", "extremely", "delicious", "meal", ""]*. Following that, stopwords in the list like *"the"* and *"are"* would be removed while words like *"chefs"* and *"cooking"* would be reduced to their root terms *"chef"* and *"cook"* by the lemmatizing step. Lastly, all strings with length less than or equal to 1 would be removed and as such the empty string "" would be removed. Merging these remaining words in the list ultimately returns us the sentence *"chef cook extremely delicious meal"* which is a form more suitable for carrying out sentiment analysis.

Now that the sentences have been cleaned up, we are ready to analyse their sentiment. Using the NLTK library, we were able to use the SentimentIntensityAnalyzer package to identify the sentiments of these sentences post-processing. This is performed by relying on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a sentence can then be obtained by summing up the intensity of each word in it. As the sentiment score returned is a range between [-1, 1], those with a sentiment score of >= 0 will be assigned to positive while those < 0 will be assigned to form the negative class.

4.3 Latent Dirichlet Allocation (LDA)

4.3.1 Definition and Explanation

In order for us to group the sentences based on the general topics, we decided to use the Latent Dirichlet Allocation (LDA) approach. LDA is a topic modelling approach that is used to extract topics from a given set of text documents. It is a generative probabilistic model, meaning that it estimates the probabilities of instances as well as probabilities of topic labels. We view our text data as a collection of composites made up of parts, where the collection is the dataset, composites are the reviews from each customer and parts are the individual words in each review. Hence we are able to pre-process our text data to obtain a Document-Term matrix, giving us the frequency of all words in a given sentence that we can use to run the algorithm.

The LDA approach makes 2 major assumptions: Documents are made up of a mixture of topics and that each topic is described by a mixture of words. There are 2 hyperparameters when running the LDA approach that influences the Dirichlet distribution that the LDA approach is based upon. They are:

1. Alpha, that controls the per-document topic distribution, affecting how likely a document contains a particular topic. A higher value means that a document is made up of more topics.
2. Beta, that controls the per-topic word distribution, affecting how likely a topic contains a particular word. A higher value means that each topic is represented by more words.

4.3.2 Application

With the hotel review data that we have, we can break down the text reviews into individual sentences and we assume that each sentence is written with a specific topic in mind from the customer. As such, we would use low Alpha and Beta values in order to get a sentence to be labelled under a more specific topic.

We are also required to choose a K value, which is the number of topics that the reviews are supposedly made up of. We chose our K value by checking 4 metrics (Griffiths2004, CaoJuan2009, Arun2010, Deveaud2014) to evaluate how well clustered the topics are on a range of K values from 2 to 10. The 4 metrics use different probabilistic models to estimate the accuracy of topic allocation within a given corpus. This was done with the help of the ldatuning package in R. From the figure below, we concluded that setting K = 6 would obtain the best clustering. To find the best K value, we would have to find the point where the 2 metrics at the top are minimized and have a small difference between them and the 2 metrics at the bottom are maximized and have a small difference between them. At K = 6, all 4 metrics agreed with each other in terms of minimizing the differences between themselves while also maximising/minimizing their values based on each metric. This means that all the reviews in the dataset could generally be classified under one of 6 topics.

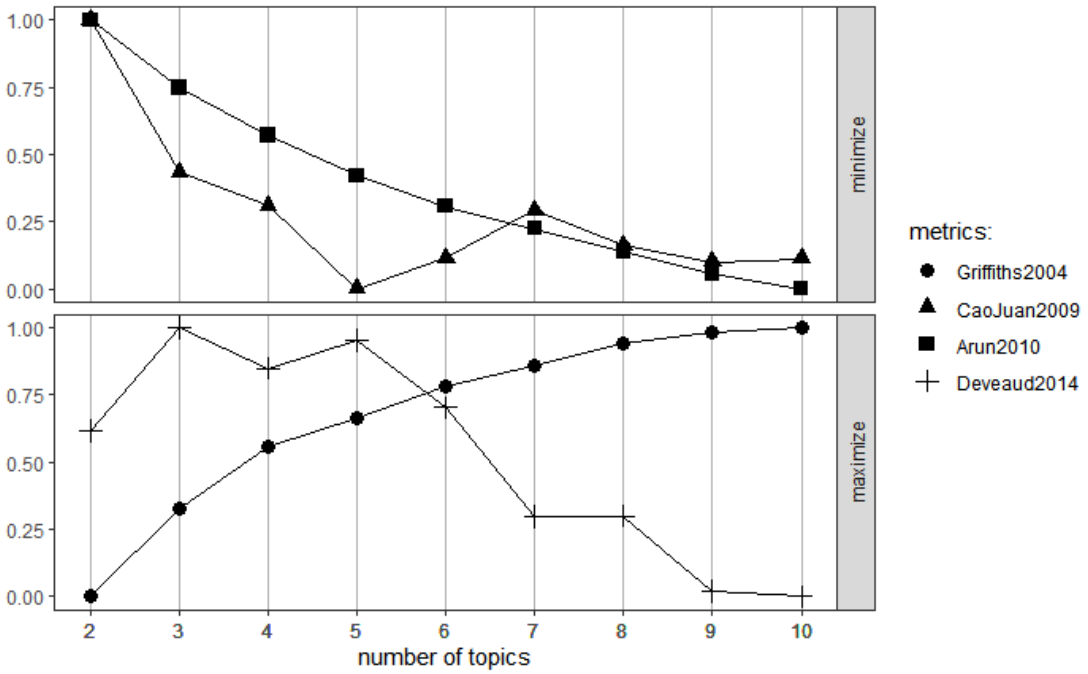


Figure 30. Lda tuning plot - conceptual coherence of topic allocation with different K values

Now that we have decided on the number of topics to have among the reviews, we will share more about the assigning process. LDA works as an iterative process by first randomly assigning topics to each word in the reviews. After which, it works towards optimizing the assignments of words and topics by adjusting and correcting the assignments over each iteration. LDA does this by computing 2 different probabilities that help it make its decisions.

1. P1 = The proportion of words in each review (D) that are currently assigned to a particular topic (k), ie: proportion (topic k / document D) (Θ)
2. P2 = The proportion of reviews where a particular word (w) is also assigned to a particular topic (k) ie: proportion(word w / topic k) (Φ)

Using P1 and P2, the LDA algorithm is able to estimate a probability of which is the most relevant topic for the current word (Z) and to replace the previous topic assignment if needed.

4.3.3 Collapsed Gibbs Sampling

Apart from using the LDA approach to assign topics to the reviews, we also used Collapsed Gibbs Sampling, which uses the underlying technique of Markov Chain Monte Carlo (MCMC) to sample from our dataset and get the conditional probabilities Θ, Φ and Z. MCMC works by simulating the sampling of words over the underlying distribution of words in the text (Monte Carlo) and is dependent on estimated probabilities obtained from previous samples (Markov Chain) We then collapse the latent variables (Θ, Φ) and marginalize them out analytically using integration methods. This gives us the conditional probabilities of topic allocation for a particular word in a certain sentence, W, given all

other topic allocations of the other words and of the same word in a different sentence in a computationally efficient way.

$$P(Z(i) \mid Z(-i), W)$$

Figure 31. Conditional probability obtained

4.4 Use of LDA to group reviews of the same topic

From Chapter 2, we have split our text data into 4 clusters of hotel guests. Now, we use the LDA algorithm on the text data of each cluster separately to analyse the text reviews and find out what aspects of the hotel guests generally review on and how important and closely related each aspect is.

Before running the algorithm, we have to pre-process our text data to obtain a Document-Term Matrix that shows the frequency of words in a particular sentence. We do this by first cleaning the data of stop-words, punctuation and any other unwanted characters/ symbols. We also check the total counts of a particular word in the whole cluster and filter out words that occur less than 5 times. This step is necessary because we want to be able to only retain the words that are commonly mentioned by hotel guests in their reviews. These words would then be used to split the review data into the 6 main topics. We chose to remove low-frequency words because they may be used to describe very specific and uncommon parts of hotels which can negatively affect our topic modelling. After preprocessing, we run the LDA with a collapsed Gibbs sampling algorithm on the cleaned data to obtain the 6 topic groupings for each of the 4 clusters.

By exploring our visualisations for the 4 different clusters, we were able to identify the 6 common topics as *“Food-related”, “Hotel amenities/vicinity”, “General comments”, “Room amenities”, “Staff/service”, “Location”.*

We reference the LDA visualisation for cluster 2 below as an example of how we interpret our results:

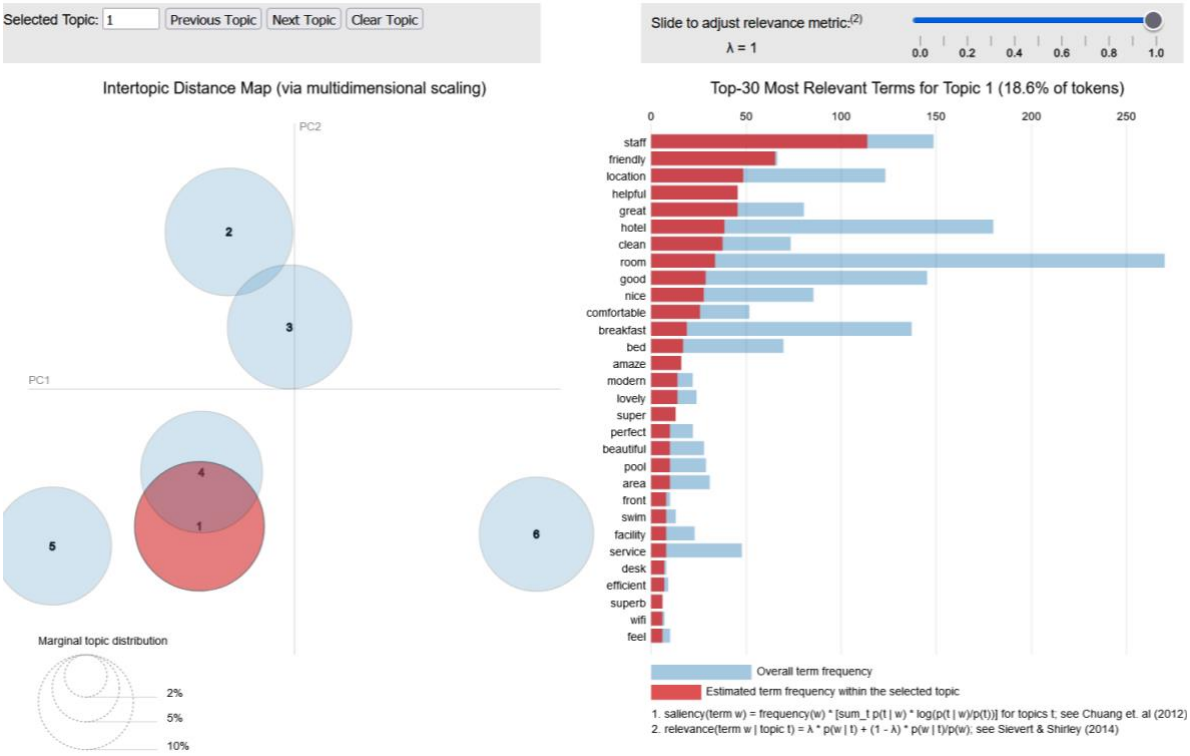


Figure 32. Cluster 2 LDA Visualisation for topic 1

From the figure above, we can see the topic clusterings on the left and the most relevant words for the topics on the right for topic 1 in this example. Blue bars show the total number of words in the corpus and the red bars show the number of times that word appears under this topic. We see that in the top 30 words, words such as *"staff", "friendly", "helpful"* are present and are also highly specific to this topic. Hence this allows us to infer that this topic is referring to the topic of Staff/service of the hotel.

Referring to figure 32 again, the distances between the topic bubbles on the left are an approximation of the semantic relationship between the topics. Looking at the example of topics 1 and 4, the topic bubbles are close together and partially overlapping. This could mean that the content of different topics could be closely related in the reviews of this cluster.

On top of that, sizes of the topic bubbles tell us how frequently talked about the topics are amongst all the reviews. In the case of cluster 2, we see that all the topics were rather evenly discussed.

4.4.1 Observations

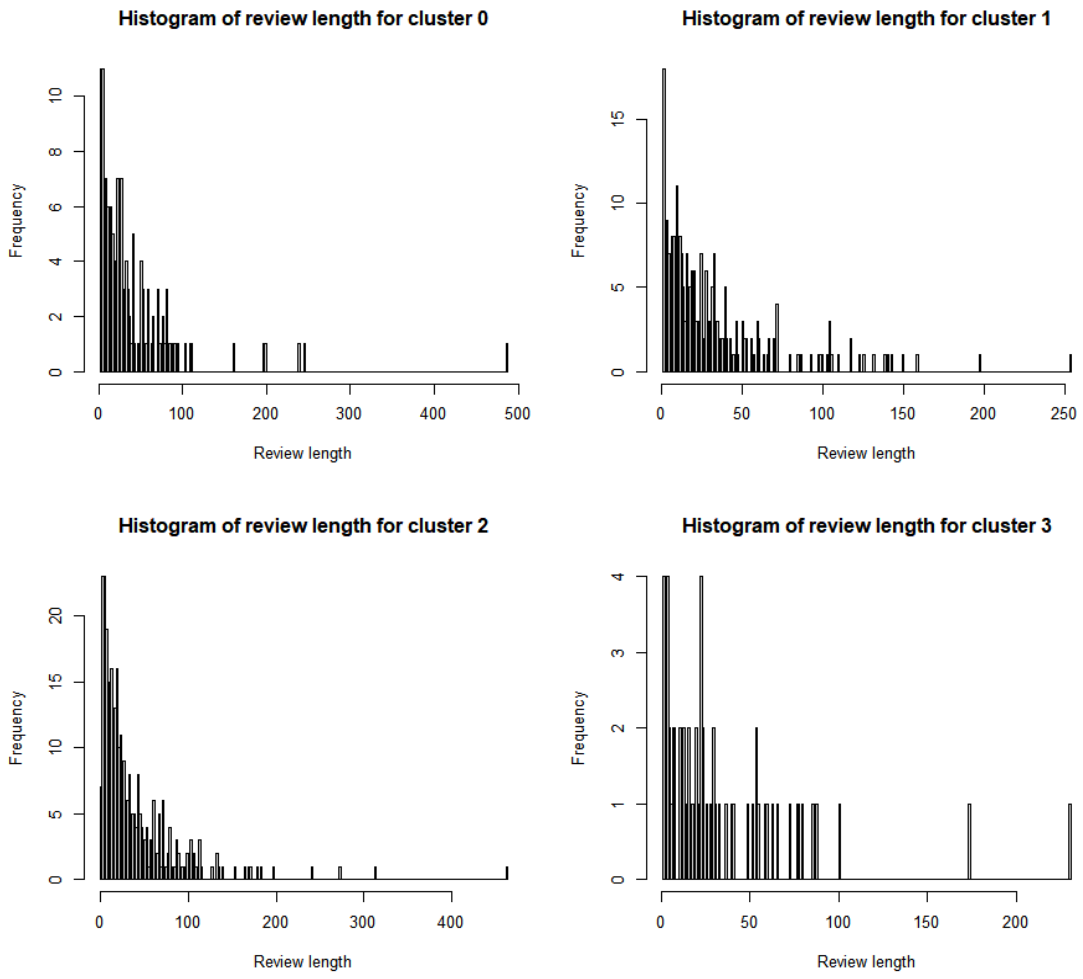


Figure 33. Histogram of review lengths from each cluster

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Number of reviews	141.0	266.00	329.00	62.00
Unique terms used for LDA	115.0	178.00	258.00	41.00
Average review Length	42.5	33.91	41.06	36.84

Figure 34. Summary of Text data from each cluster

	Segment	Total	traveller_type	quarter	price_per_night	rating
0	0	208	Group	3	211.125	8.951
1	1	435	Solo traveller	1	103.814	8.431
2	2	492	Couple	4	150.807	8.662
3	3	65	Couple	4	306.323	9.474

Figure 35. Recap of characteristics of each cluster

Difference in the "Total" column in figure 35 and "Number of reviews" row in figure 34 is due to deletion of rows which have no available reviews.

From figure 33, we are able to see the distributions for length of reviews under each of the 4 clusters. For all clusters, the distribution appears similar, except for cluster 3 where the gradient of descent is gentler. Referencing from figure 34, we see that cluster 3 has only 62 reviews available for our topic modelling. The gentler gradient therefore tells us that we have more well spread out review lengths in cluster 3 as compared to the other 3 clusters where the majority of review lengths are generally on the shorter side.

We can also notice from figure 34 that cluster 1 has the shortest review length. Referencing from figure 33, we confirm that the length of the majority of the reviews from this cluster are rather short. Looking at figure 35, we notice that cluster 1 is made up of Solo Travellers and are also the lowest spenders. This could represent people such as businessmen, people that are transiting in the city and backpackers. We also see that there were 435 customers in the cluster but only 266 reviews that had non-empty reviews, showing that 38% of customers did not leave reviews. This together with the short reviews allows us to infer that a large percentage of these customers may have put lesser effort into writing a review and possibly stayed in the hotel just for the shelter over their heads for that period and did not require much luxuries during their stay. Hence, we could consider a lower budget in room design for this cluster of customers if any furnishing cost recommendations are needed.

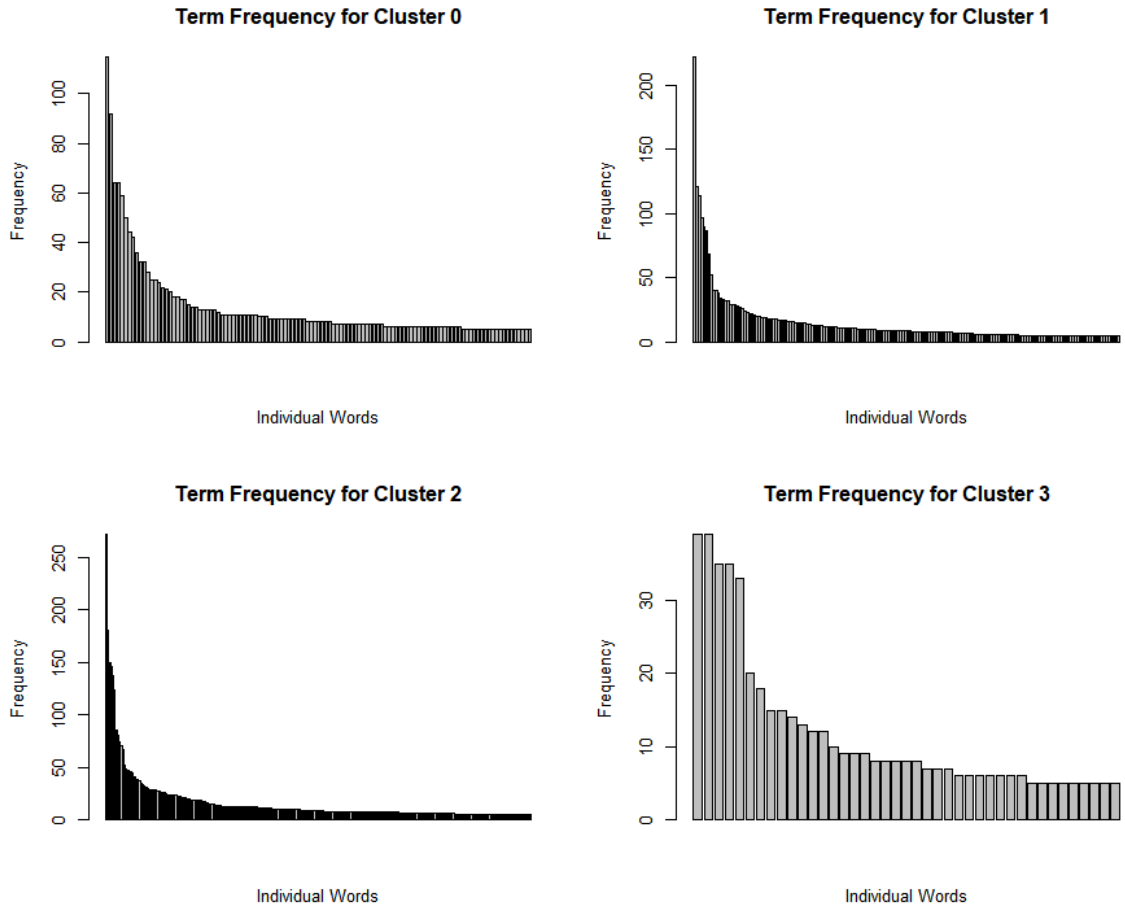


Figure 36. Individual Word Frequency in each cluster

From figure 36, we see that in all 4 clusters, there exists a rather steep gradient of descent. This tells us that there are a small group of words in every cluster that are very commonly mentioned in the reviews. These could be words very relevant to the hotel such as: “rooms”, “breakfast”, “staff”. Or it could also mean that there are some aspects of the hotels that are very commonly talked about in a particular cluster of customers. This gives us the opportunity to figure out the difference in what customers from different clusters are paying attention to or looking out for when they stay in the hotel. By identifying these words through further analysis, we can see the exact reviews on those words and allow us to learn about what we should improve or change to a hotel to attract a particular cluster of customers to the hotel.

However once again, we notice that there is a gentler gradient in the plot for cluster 3. Referencing from figure 34, we see that there were only 41 unique terms that were used in the LDA topic modelling for topic 3. This could therefore affect the interpretability of our topic modelling.

The LDA visualisation for the other 3 clusters namely cluster 0, cluster 1 and cluster 3 have been included in the following lines.

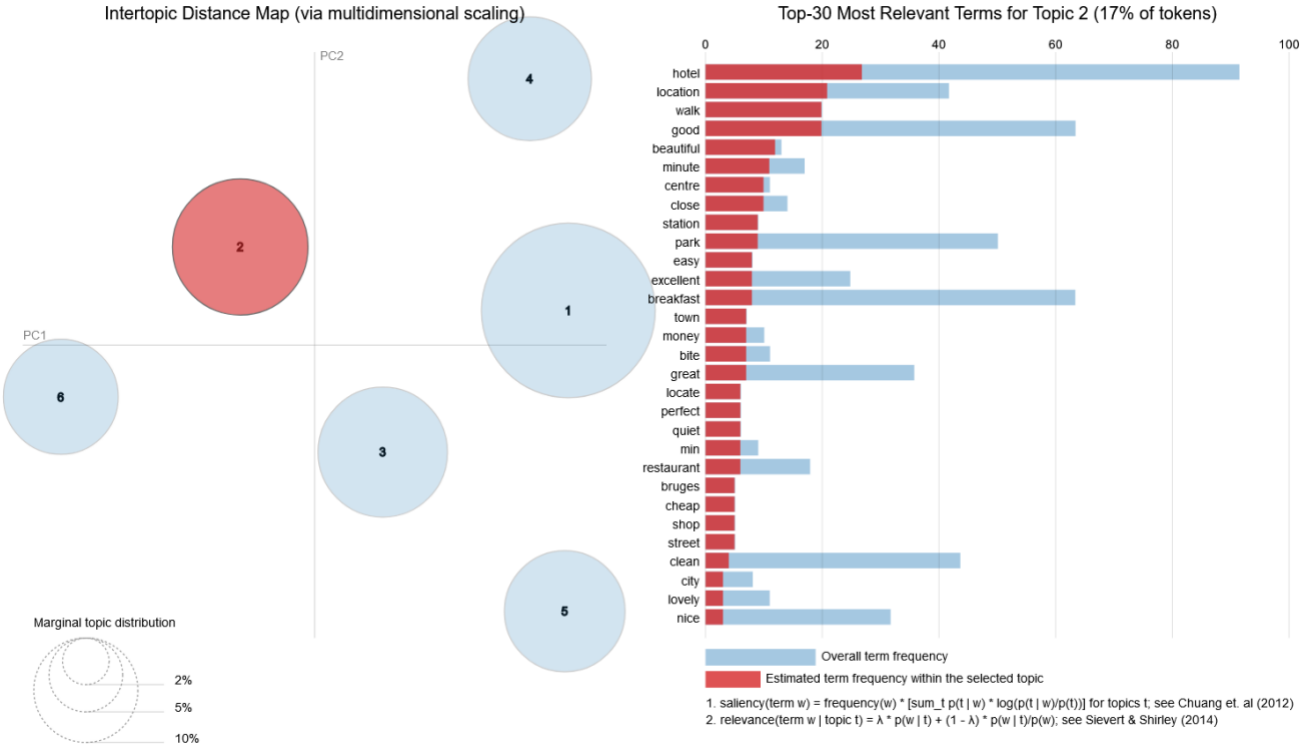


Figure 37. Cluster 0 LDA Visualisation

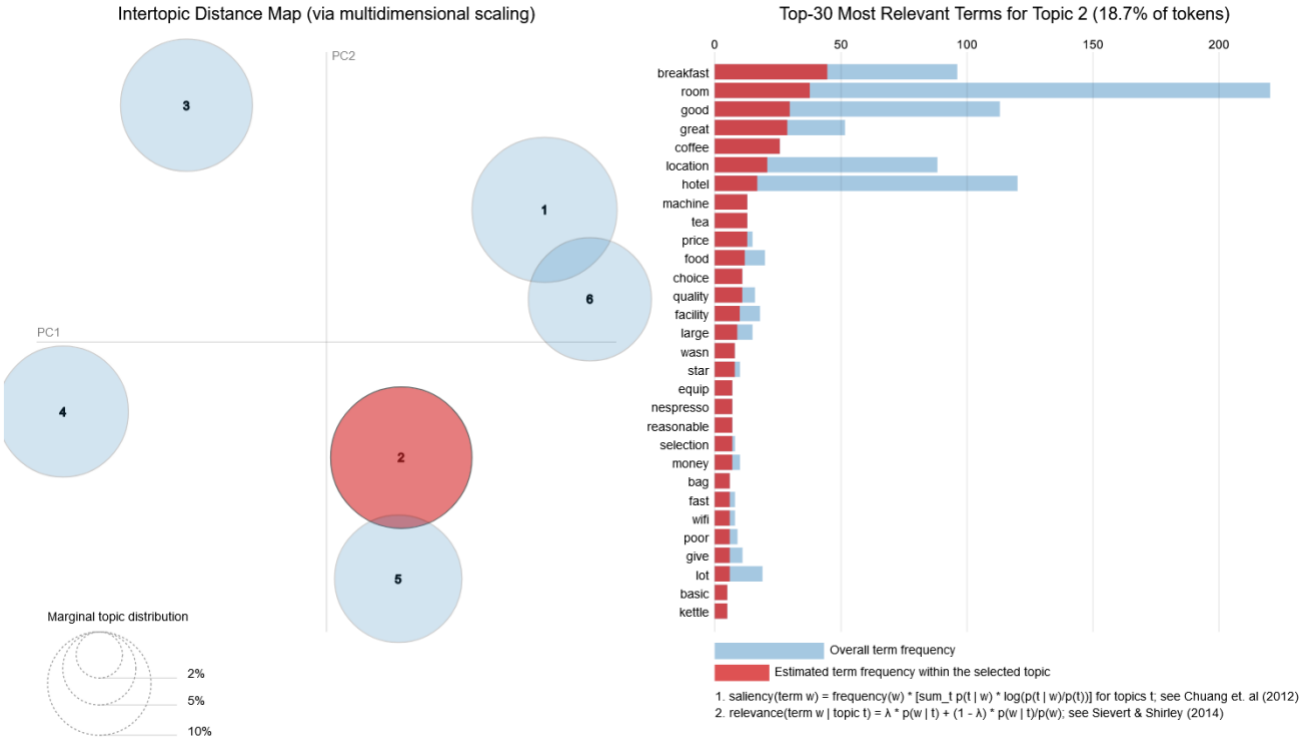


Figure 38. Cluster 1 LDA Visualisation

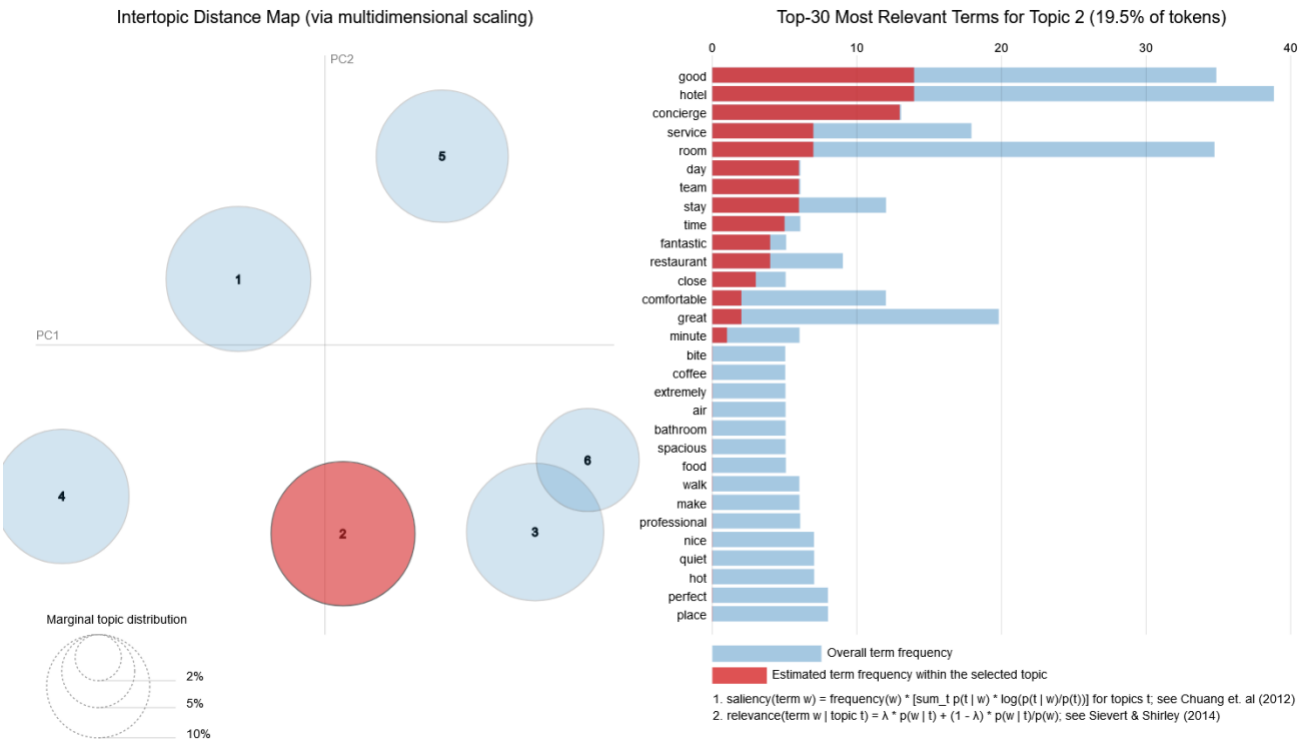
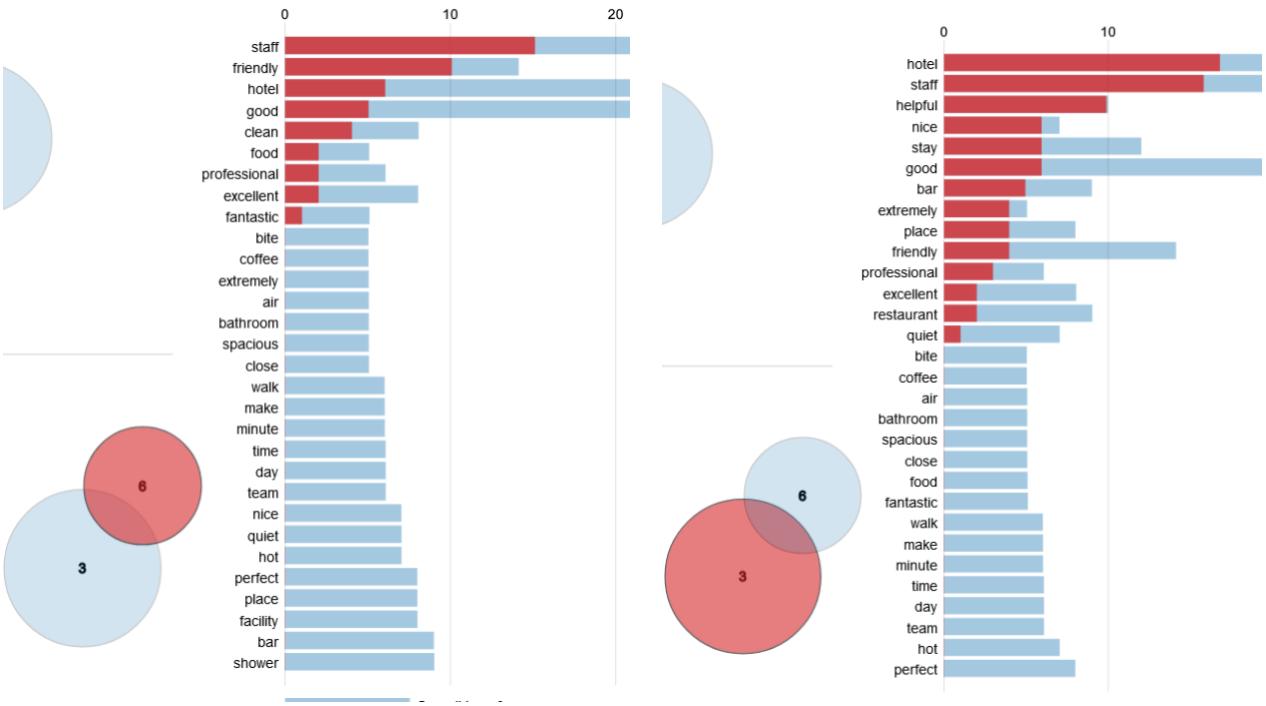


Figure 39. Cluster 3 LDA visualisation

From our analysis of the LDA visualisations, we were able to obtain interpretable topics from clusters 0, 1 and 2, where labellings were easily assigned and there was a large number of words allocated to each topic to provide good evidence to justify the topic namings.

However, for cluster 3, due to the smaller number of reviews available along with the comparatively small number of unique terms used in topic modelling, it reduces the ability for us to differentiate the topics. The reason why the number of unique terms in this cluster is significantly lower is because there are a lot of terms that have frequencies less than 5 and hence did not pass the word count checking mentioned in paragraph 2 of Chapter 4.4. This meant that there were fewer words classified under a certain topic and hence reduced the ability for us to differentiate the topics. We use topics 3 and 6 to illustrate this.



Cluster 6 General comments **Cluster 3 Staff/service**

Figure 41. Cluster 3 LDA Visualisation topics 6 and 3

As seen above, after assigning the topic namings for all the topic clusters to our best ability, we gave cluster 6 the naming of “General comments” and cluster 3 the naming of “Staff/service”. By the number of red bars shown in both figures, we can see that there are few words to justify with strong confidence why we gave these labels to each of the clusters. As the topics are closely related in this cluster, we also see words like “staff” present in both clusters which further introduces ambiguity during the naming process.

As mentioned in paragraph 2 of Chapter 4.4, the filtering out of terms with frequency below 5 was so that low frequency words that may describe a very specific and uncommon part of hotels will not affect our topic modelling much, with the intention to gather insights that may allow us to attract larger groups of people from one cluster to another.

In the case for cluster 3, referencing from figure 35, we see that this cluster has the highest spending on rooms and gave the highest ratings, telling us that this group may have stayed in the most luxurious rooms available.

Hence, unlike the other 3 clusters, we would want to deepdive into the reviews from this cluster of customers by further identifying specific aspects that customers in this cluster are looking out for, which made their stay and experience great.

5. Insights

Now that the clusters and the topics from each cluster have been formed. Let us carry out some further analysis of the reviews from hotel guests in each cluster.

5.1 Overall distribution

Firstly let's take a look at the overall distribution of sentiments across all the reviews from our dataset from all 4 clusters. To allow us to analyse the sentiment of the reviews, we decided to utilise the same SentimentAnalyzer package as mentioned above in Chapter 4.2.

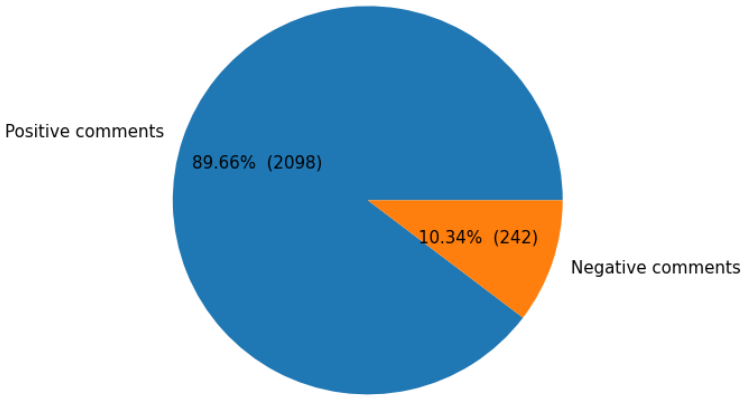


Figure 42. Pie chart displaying the proportion of sentiments within the dataset

As seen from the pie chart above, looking at the reviews in its totality, this dataset is largely skewed towards having more positive sentiments. This would mean that most of the insights derived from our individual clusters will be understanding things that reviewers liked about their stay rather than what they disliked about it.

5.2 Common Topics in Cluster 0, 1, and 2

Using the LDA visualisations, for each of the 3 clusters (Cluster 0, 1 and 2) that provided good topic clustering, we were able to obtain the top 20 words that were associated with each of the 6 topics (“Location”, “Food related”, “Hotel amenities/vicinity”, “Room amenities", “General comments”, “Staff/service”). Analysing the top 20 words for each of the topics in the 3 clusters we will be working on, we noticed that there exists words in it that were either very general or did not suit the topic it was assigned to. Hence, to minimise the overlap between the different topics, words like “good”, “extremely” as well as words that did not suit the topic of its own cluster were being removed.

Following that, let us now take a look at cluster 0. Treating all of the words in the 6 topics in cluster 0 as a bag of words, we create a dataframe in which each row represents a

certain word. On top of that, we take note of all sentences with this specific word in it along with the polarity of them. This results in a dataframe as shown below.

	word	pos_count	pos_desc	neg_count	neg_desc	total_count	topics
0	room	104	[enjoyed the option to have breakfast and even...	8	[my partner was with me next to the phone in t...	112	[Room amenities, Hotel amenities/vicinity, Sta...
1	hotel	81	[very comfortable beds and the best bath in an...	10	[we stayed on a side of the hotel with a very ...	91	[Room amenities, Staff/service, General commen...
2	breakfast	59	[enjoyed the option to have breakfast and even...	3	[we didn't have the hotel's breakfast preferri...	62	[Hotel amenities/vicinity, Food related]
3	min	40	[just a real minor thing: really small selecti...	5	[we didn't work this out and spent several min...	45	[Location]
4	clean	42	[everything was great, but the breakfast quali...	2	[nothing ,clean with good facilities., nothing...	44	[Room amenities, Staff/service, Food related]

Figure 43. Table showing us information of the reviews in each cluster

Referring to the figure above, the dataframe has 6 columns. The first column named “word” refers to the word from the bag of words that is associated with this row. Next, the pos_count and pos_desc columns show us the number of positive sentiment sentence tokens that have the “word” present in it along with the actual sentence tokens. This is repeated for sentence tokens that have a negative sentiment. The next row shows us the total count of both positive sentiment sentence tokens and negative sentiment sentence tokens for the associated “word”. Lastly, the topics column indicates which of the 6 topics the description of the “word” in the first column belongs to.

Using this data frame above, we were able to obtain the number of positive and negative sentence tokens associated with each of the 6 topics. This can be seen in the figure below.

	index	positive	negative
0	Room amenities	362	40
1	Hotel amenities/vicinity	231	27
2	Staff/service	290	28
3	General comments	159	36
4	Food related	372	30
5	Location	226	27

Figure 44. Table showing the number of sentence tokens for each of the topics in cluster 0

This figure above shows us all 6 topics along with the number of positive as well as negative sentence tokens associated with each of the topics for cluster 0.

Using this data we were able to analyse the sentiment of the sentence tokens from cluster 0. To find out which is the most frequently occurring word present in the bag of words, we chose to adopt a barplot to allow the distribution of the sentences to be clearly depicted for further analysis.

5.2.1 Cluster 0

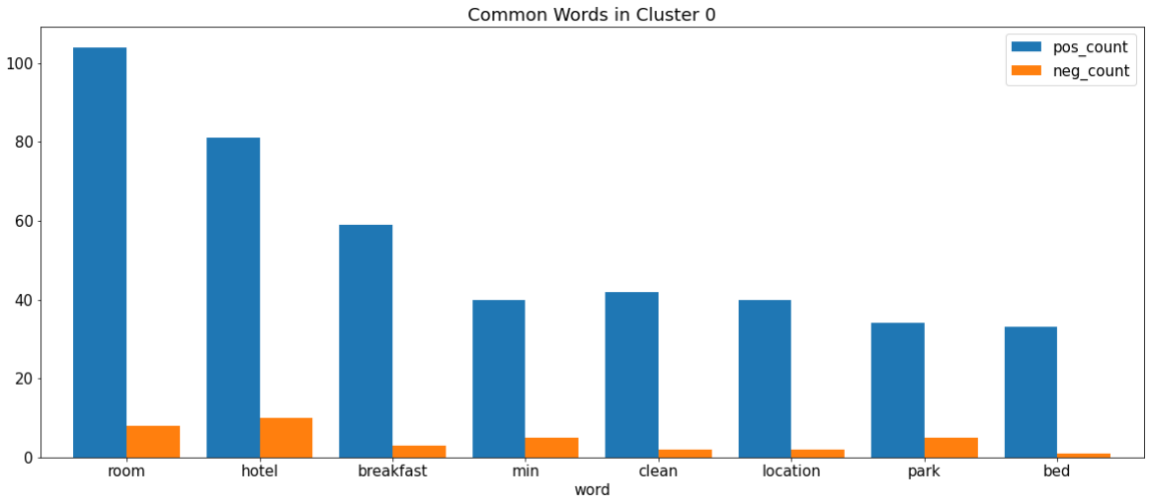


Figure 45. Common Words in Cluster 0

Looking at the figure above, it shows us the most frequently occurring words amongst the bag of words from the sentence tokens. Words like “room” as well as hotel and breakfast are highly mentioned among the reviews from cluster 0. Comparing the blue bars and the orange bars we can tell that the distribution of positive sentence tokens is very much larger than the negative ones amongst the reviews in cluster 0.

Another possible visualisation we can make with the information we have is a barplot showing the distribution of the sentence tokens among the 6 topics for each cluster.

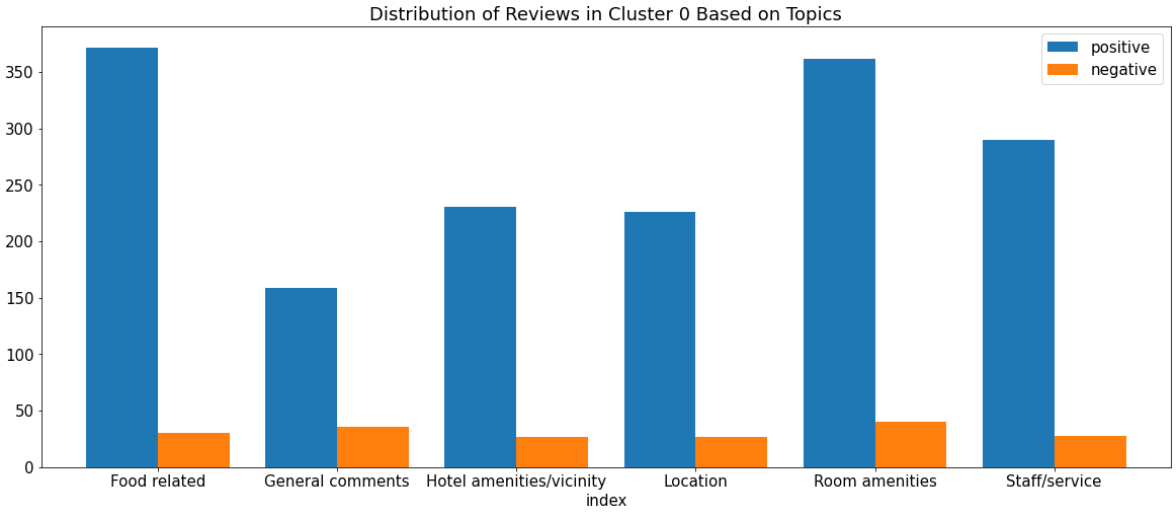


Figure 46. Distribution of Reviews in Cluster 0 Based on Topics

The figure above shows us the distribution of the sentence tokens amongst the 6 topics identified by LDA. From this, we can tell that for the hotel guests in cluster 0, most of the

reviews made were about food-related matters or room amenities matter. Further interpretation can then be made on these reviews to find out exactly what this group of hotel guests like or dislike about their stay in competitor hotels and from there allows us to work on it to stand out from them.

5.2.2 Cluster 1 and 2

The following steps can then be repeated for the other 2 clusters, cluster 1 and cluster 2 and the results can be seen below.

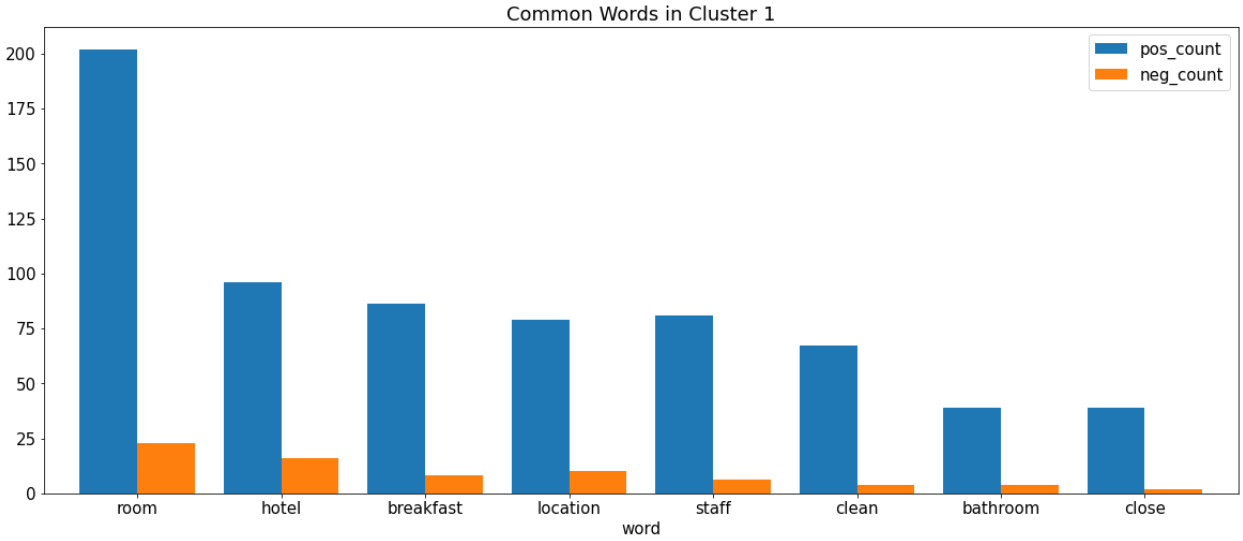


Figure 47. Common Words in Cluster 1

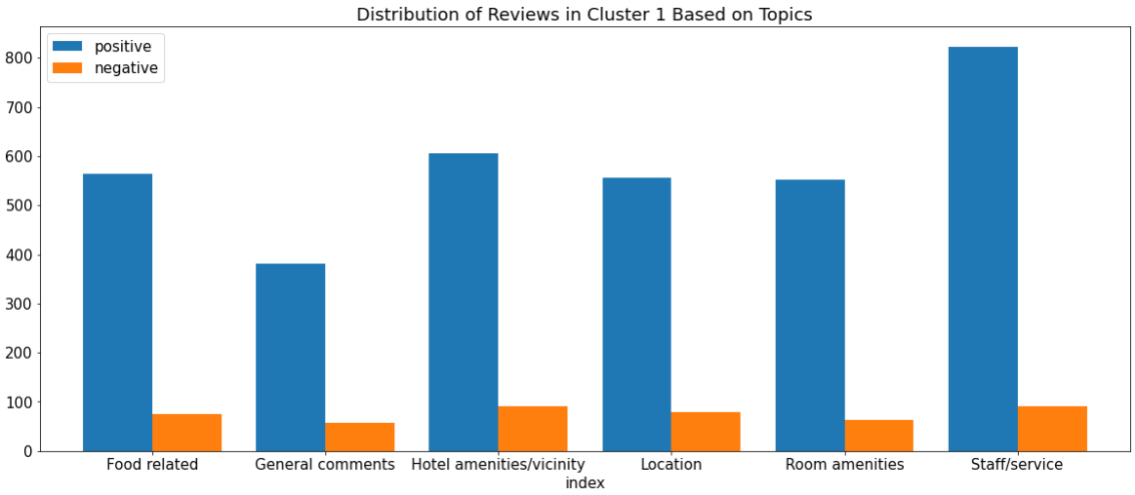


Figure 48. Distribution of Reviews in Cluster 1 Based on Topics

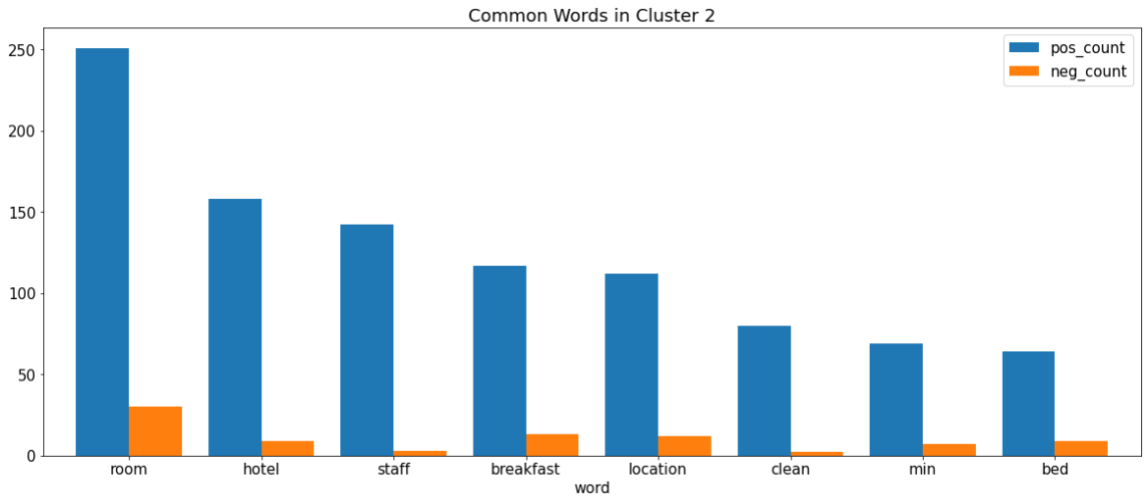


Figure 49. Common Words in Cluster 2

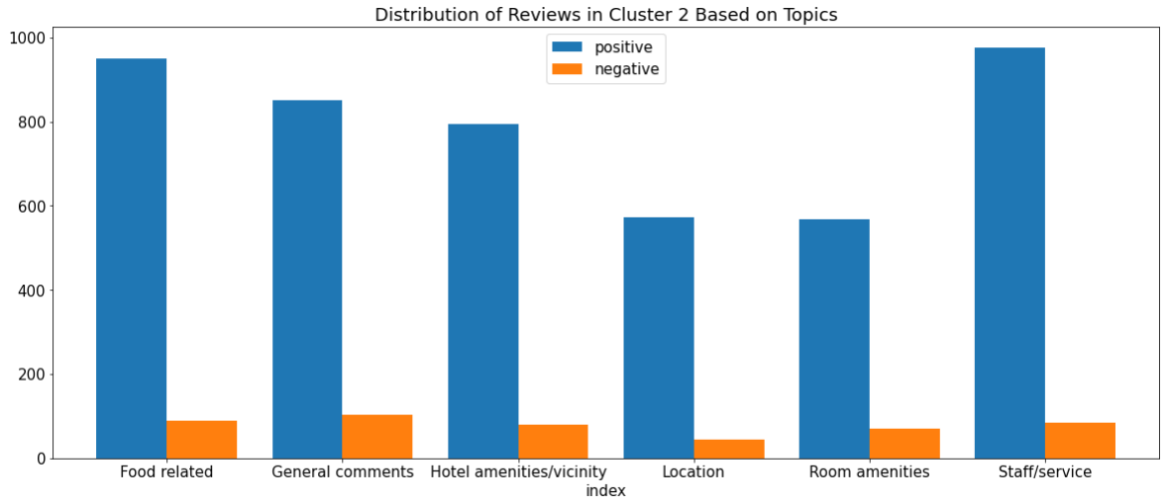


Figure 50. Distribution of Reviews in Cluster 2 Based on Topics

5.3 Most Prominent Features across Clusters based on each Topic

After having a clearer understanding of the aspects that different clusters of hotel guests look out for on their hotel stays, we went a step further to make further analysis on each of the topics from each of the clusters. By finding the most frequently occurring noun from the reviews that suited the topic for each of the topics, we were able to sieve through the reviews in a topic to only keep those that contain that specific “noun”. Next by using the noun_chunks attribute from the spaCy package available in python, we were able to extract the base noun phrases present within each sentence token. Essentially, noun chunks are made up of a noun and the words describing the noun. An example of the noun chunks extracted from a sentence can be seen below.

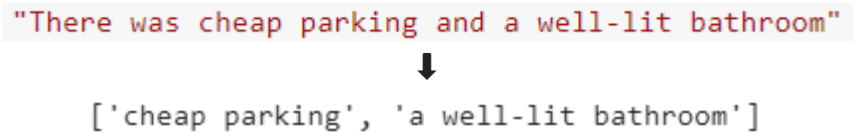


Figure 51. Example of noun_chunks attribute from spaCy package

By keeping only the noun chunks that contained the specific “noun”, we were able to review essential information regarding the noun in a more concise and simple manner. We would reference the positive reviews to improve and propose it for inclusion in the hotel while ensuring that the negative items are absent or fixed in our hotel.

5.3.1 Positive Feedback

Now that we know how to get the descriptions of the words, let’s first see what people from each cluster like from their respective hotel.

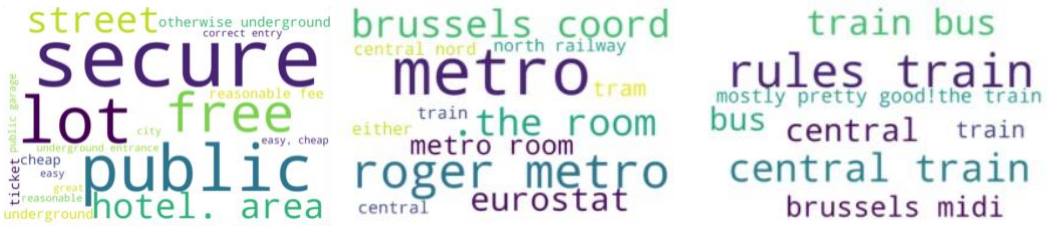


Figure 52. Positive description of the most prominent noun on ‘Location’ in cluster 0 (left), cluster 1 (middle), and cluster 2 (right) respectively

As previously mentioned, we were focusing on ‘parking space’ for cluster 0 and ‘station’ for cluster 1 and 2. This is fairly reasonable as people in cluster 0 travelled in groups and hence might prefer their personal vehicle as means of transportation. While travellers in both cluster 1 and 2, who are travelling individually, prefer public transportation such as train/metro station.

Based on the figure above, we see that people from cluster 0 highly value secure, cheap, and easy to access parking space. While most people in cluster 1 appreciate the close proximity of metro stations from the hotels. Similarly, positive reviews about ‘station’ in cluster 2 show that people were pleased with the convenience of being close to train stations from their hotels.

Figure 53. Positive description of the most prominent noun on ‘Food’

in cluster 0 (left), cluster 1 (middle), and cluster 2 (right) respectively

Across all three clusters, we observed a significantly large proportion of people mentioning breakfast in their reviews on topic 'Food'. Hence, by using the word cloud function, we then observed what people enjoy about their breakfast during their stay. Although most reviews simply mentioned that they had a pleasant breakfast experience, we could see that venues could be one of the things contributing to the positive breakfast experience for people in cluster 0. This was indicated by words such as, seating, seaside, or area. Meanwhile, words like buffet, great selection, and arrangement could be seen in cluster 1. This meant that people in cluster 1 generally view the choice of meals as a significant contributor to their well received breakfast experience. On the other hand, positive reviews on cluster 2 were generally more distributed, with some notable mentions about buffet, free complementaries, and quality of breakfast.

Figure 54. Positive description of the most prominent noun on 'Room Amenities' in cluster 0 (left), cluster 1 (middle), and cluster 2 (right) respectively

Upon exploring reviews on the topic 'Room Amenities', we decided on 'bed' for cluster 0 and bathroom for both cluster 1 and 2 as the most prominent features mentioned. Observing the word cloud figures above, we concluded that people in cluster 0 appreciate large and comfortable bed(s) and room(s). On the other hand, people in cluster 1 especially liked the aesthetic of the bathroom and the quality of the bathroom products that come with it. This can be seen with the mention of words such as huge and modern, well equipped, and complimentary. Meanwhile people in cluster 2 made some more general comments such as nice or fine with mentions of ventilation.

Figure 55. Positive description of the most prominent noun on 'Staff/Service' in cluster 0 (left), cluster 1 (middle), and cluster 2 respectively

The consensus determining feature on the topic ‘Staff/Service’ across all three clusters is the behaviour of the staff. On all three clusters, we observed that people are most grateful for how friendly, supportive, and helpful the staff are.

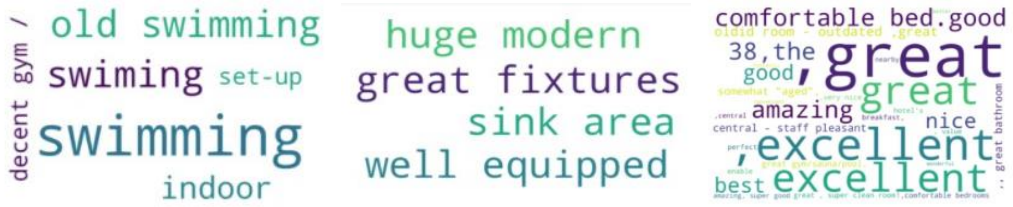


Figure 56. Positive description of the most prominent noun on ‘Hotel Amenities/Vicinity’ in cluster 0 (left), cluster 1 (middle), and cluster 2 (right) respectively

The most prominent features on the topic ‘Hotel Amenities/Vicinity’ generally differ across clusters. After further inspection, we decided on ‘pool’, ‘bathroom’, and ‘location’ as the features we will further investigate for cluster 0, 1, and 2 respectively. As seen from the word cloud figures above, people generally like having a sports facility such as an indoor pool/gym with nice sets up in their hotel. For cluster 1, we could see similar preferences of general bathrooms in the hotel as their en suite bathroom, such as the aesthetics and the equipment of the bathrooms. Lastly, we observed comments on how ideal the location of the hotels are on reviews from cluster 2. This indicated that people in cluster 2 highly value strategic location, which is in line with their comments on train stations in ‘Location’ topic.

Figure 57. Positive description of the most prominent noun on ‘General Comments’ in cluster 0 (left), cluster 1 (middle), and cluster 2 (right) respectively

General comments give us an overview of features that people most valued during their stay in the hotels. Hence, we decided to get possible descriptions of the hotel for all clusters using the noun chunk to find descriptions of the word ‘hotel’ for cluster 0 and 1 and ‘room’ for cluster 2. Whilst the word cloud on general comments have a wider range of descriptions, including some mentioned in previous topics, we could conclude that freshly renovated hotels with a parking area in a nice location (e.g. high rise) is positively received by people in cluster 0. On the other hand, people in cluster 1 generally made remarks on the aesthetic or the service of the hotels. Similarly, we could see that people in cluster 2 also favoured the quality of service in addition to the hygienic aspect of the hotels.

5.3.2 Negative Feedback

As seen from figure 42 and the corresponding explanation in Chapter 5.1, our dataset is largely composed of positive reviews. Hence, we were only able to get a few negative descriptions for the features on each topic as listed below.

Figure 58. Negative description of the most prominent noun on 'Location' in cluster 0

As mentioned in Chapter 5.3.1, people in cluster 0 have a positive impression on hotels with parking space. However, they also respond negatively when these parking spaces do not come with a reserve parking option.

Figure 59. Negative description of the most prominent noun on 'Room Amenities' and 'Hotel Amenities' in cluster 1

As observed from the figure above, people in cluster 1 had bad experiences with hygienic issues during their stay. Although we might not be able to generalize this to a problem with all of the hotels, it was an important reminder that hygiene is something that people correlate with their impression of the hotels.

Figure 60. Negative description of the most prominent noun on 'Food' (top left), 'General Comments' (top right), and 'Hotel Amenities' (bottom) in cluster 2

Finally, people from cluster 2 had issues with poor quality of breakfast or no cooked meal options. Additionally, unaccommodating facilities had also caused them inconvenience.

In conclusion, their overall bad experience could be attributed to the quality of the hotel in general that did not quite meet their expectations.

5.4 Observations on Cluster 3

As what was mentioned in Chapter 4.4.1, in order to better understand the needs and wants from hotel guests in cluster 3, we would have to take a closer look due to the unique properties of this cluster. First, we use the word cloud to display the most frequently mentioned words from this cluster.

Figure 61. Word cloud of words present from reviews in cluster 3

Looking at the above word cloud, it gives us an insight as to what most of the reviews are mentioning. Referencing the exact frequency of the words from cluster 3, we see that word “staff” actually has the highest word frequency, only tied with “hotel”. As the customers in this cluster are the highest paying and also the ones who gave the best ratings out of all 4 clusters, we use this as an example to illustrate how we can use the reviews to add value to our future customers to justify them paying a premium for our hotel.

Upon inspecting the reviews with the word “staff” in them, we see items such as “very well run hotel with exceptionally trained staff” and “the staff were outstanding, warm and kind and everyone knew my name from when I checked in”. One other review that stood out was “The staff were more than great and helpful we were there during la tour de france 2019, the roads leading to the hotel were closed the hotel sent us one of the employees to guide us through the crazy traffic he walked more than half an hour to reach us and another 1 hour in the car trying to figure out the quicker and less buster road to the hotel ..... I can easily say that I have never experienced a more professional converge service like the one I have received from the converge team”.

From the above review, we observe something interesting. By reading through the reviews in this cluster, we observed that a large part of the reviews in this cluster talked about

potential improvements. However, the average rating from this cluster is still 9.474/ 10 which is the highest out of all 4 clusters. Looking at the 3 sentences with the word “Staff”, a common theme amongst them is that apart from the elusive praise for the staff, the reviewers used words that describe emotions and explained how the staff’s gestures made their experience better. Hence we can infer that emotions play a big part in a customer’s overall view of their hotel experience and that the staff’s service played a big part of it.

One insight that we can take from this is that, as a hotel, customers who receive good service will be happy and willing to pay more for their experience. Since infrastructure and amenities improvements require fundings and may not always be an available option, to attract customers to stay at our hotel even with higher prices, we can differentiate our hotel by investing in our staff through more welfare and service training. A Staff force that is well taken care of indirectly translates to happier customers as they would be more likely to be motivated to go the extra mile to make the customers feel special as shown in the examples above. This is a form of emotional selling and is also beneficial for long term business prospects.

6. Recommendations

For our recommendations, we assume the following:

1. Customer's budget is assumed from the historical data of price of the room stayed
2. Travellers will book a hotel of a similar type and budget to one they have booked historically
3. We assume that there is no significant change in preference of hotel guests and hence, will approve of the recommendations made

6.1 Pricing & Packages

Cluster	A quarter of the year	Description	Price per night (rounded off to nearest ten)
0	3	Group travellers on a medium budget	210
1	1	Solo travellers on a low budget	100
2	4	Couples on a low budget	150
3	4	Couples on a high budget	310

Table 3. Summary of the customer profiles formed using K-Prototype

From table 3, we came up with a pricing strategy to maximise profit and allow us to cater to the demands of the customers based on the quarter of the year, customer's budget, price of the rooms per night and also the type of traveller they are.

In the first quarter of the year, we should offer discount packages for single rooms to cater to solo travellers with a low budget.

In the second quarter of the year, which seems to be the non-peak travel period, we should offer all rooms at a discounted price to attract travellers to stay. We should also market our low prices online and on social media to compete with local hotels and improve brand recognition.

In the third quarter of the year, we can release attractive bundled room packages to cater to the demand of group travellers. On top of that, we should also collaborate with local private hire and tourist attractions to offer transport and bundled tickets to go along with the room package, offering them a wholesome experience.

The fourth quarter of the year is usually the peak travel period and the price of the room will surge because of the high demand. We can come up with creative themed double rooms to attract couples who are willing to spend for better rooms and higher enjoyment.

As seen from Figure 24, there would also be solo and group travellers to account for as well. Since we have limited hotel rooms, during the busy period, we have to forgo some customers with a low budget at our hotels and prioritise customers who we expect to have high spending. Therefore, for solos and couples with a low budget, we are unable to accommodate them if they are unwilling to pay.

6.2 Features of the Hotel

Topic	Cluster 0 (Groups)	Cluster 1 (Solo)	Cluster 2 (Couples)
Location	Cheap & secure parking Reserve Parking	Near train station	Near train station
General	Prefer newly renovated hotel with parking area	Prefer aesthetically pleasant-looking hotels with quality service	Valued hygienic (i.e. smell, cleanliness) and quality of service
Staff/ Service	Friendly, helpful and polite		
Room	Comfortable and large beds	Clean and big bathrooms with quality bathroom products. Not good to have worn out items and dirty bathroom	Big bathrooms
Food	Good quality breakfast in a comfortable dining hall	Wide selection for breakfast	Breakfast buffet Complimentary breakfast Not good to have no

			cooked breakfast
Hotel Amenities	Indoor pool / gym with nice set up	Bathrooms aesthetic and equipments	Strategic location

Table 4. Summary of insights from sentiment analysis

Consolidating all feedback in table 4, we aim to be a holistic hotel by fulfilling the features listed and filling the identified market gap. By catering to the wants of each cluster individually, we can meet the expectation of each cluster in order to achieve our goal. With cluster 2 (Couples on low budget) being the largest cluster, there are opportunities in generating larger revenues if we can offer luxury packages to the couples in Cluster 2, hence shifting this group to Cluster 3 (couples on high budget). These couples can spend more on their rooms while still leaving with a better overall experience. This can be done by introducing certain fine-dining and special couple suite rooms. This can be achieved by liaising with fine-dining restaurants to set up their restaurants in the hotel itself, since convenience is something that customers seek, as seen from their reviews of the hotel being near train stations.

With the current recovery of the hotel industry due to Covid-19, Belgium is currently low on infectious rates and over 72% of the population is fully vaccinated. With people feeling more safe and a large proportion of the population feeling deprived of travel, locals and tourists alike are more inclined to go for vacations and business trips. Due to the aforementioned reasons, we are planning to set up our hotel during this time period of a recovery from a pandemic.

7. Conclusion

1. We were able to expand our dataset with additional columns containing information about customers that were proven helpful for our analysis and results.

2. From the K-Prototype model, we identified 4 clusters which provide meaningful insights that aided us in catering to different types of travellers with various budgets during specific periods of the year as seen in 6.2.1.

3. We met our objective of using the clusters to come up with appropriate pricing and marketing strategies with revenue maximising goals in mind.

4. Our niche hypothesis was also evident in 3.2 when we plot the Hotel against Proportion of Cluster in each hotel.

5. From the sentiment analysis and LDA method, we have identified 6 main topics that the reviews can be categorised under. From there, we analysed the distribution of the reviews

in each topic and cluster and deep dived into the individual reviews to gain insights from the experience hotel goers had with other hotels.

8. References

For Chapter 2 and 3:

Ruberts, A. (2020, May 16). *K-Prototypes - Customer clustering with mixed data types*. Well Enough. Retrieved September 14, 2021, from https://antonsruberts.github.io/kproto-audience/.

Bagavathy, P. (2020, May 26). *DBSCAN clustering*. Kaggle. Retrieved September 14, 2021, from https://www.kaggle.com/bagavathypriya/dbscan-clustering.

Kemmer, R. (2021, January 29). *Clustering on mixed data types in Python*. Medium. Retrieved September 14, 2021, from https://medium.com/analytics-vidhya/clustering-on-mixed-data-types-in-python-7c22b3898086.

Zazueta, Z. (2020, November 2). *Telecom_churn_project/notebook 2 - cleaning and kprototypes2.ipynb at master · zachzazueta/telecom_churn_project*. GitHub. Retrieved September 14, 2021, from https://github.com/zachzazueta/telecom_churn_project/blob/master/Notebook%202%20-%20Cleaning%20and%20KPrototypes2.ipynb.

Reza. (2021, January 20). *Slow at running the calculation of silhouette score in K prototypes clustering algorithm for mixed categorical and Numerical Data*. Stack Overflow. Retrieved September 14, 2021, from https://stackoverflow.com/questions/65814395/slow-at-running-the-calculataion-of-silhouette-score-in-k-prototypes-clustering.

Yadav, M. (2020, April 3). *Customer-segmentation-using-K-prototype-clustering/rfm.py at master · milindyadav-97/customer-segmentation-using-K-prototype-clustering*. GitHub. Retrieved September 14, 2021, from https://github.com/MilindYadav-97/Customer-Segmentation-using-K-prototype-clustering/blob/master/rfm.py.

Jia, Z., & Song, L. (2020). Weighted K-prototypes clustering algorithm based on the hybrid dissimilarity coefficient. *Mathematical Problems in Engineering*, *2020*, 1–13. https://doi.org/10.1155/2020/5143797

Molnar, C. (n.d.). *Interpretable machine learning*. 8.6 SHAP (SHapley Additive exPlanations). Retrieved September 15, 2021, from https://christophm.github.io/interpretable-ml-book/shap.html#definition.

Chapter 4:

Nltk.Tokenize package¶. nltk.tokenize package - NLTK 3.6.2 documentation. (n.d.). Retrieved September 14, 2021, from https://www.nltk.org/api/nltk.tokenize.html#nltk.tokenize.punkt.PunktSentenceTokenizer.

Hamza, A. (2019, January 30). *Effectively pre-processing the text data Part 1: Text Cleaning*. Medium. Retrieved September 14, 2021, from https://towardsdatascience.com/effectively-pre-processing-the-text-data-part-1-text-cleaning-9ecae119cb3e.

Prabhakaran, S. (2018, October 2). *Lemmatization approaches with examples in python*. Machine Learning Plus. Retrieved September 14, 2021, from https://www.machinelearningplus.com/nlp/lemmatization-examples-python/.

Nltk.Sentiment package¶. nltk.sentiment package - NLTK 3.6.2 documentation. (n.d.). Retrieved September 15, 2021, from https://www.nltk.org/api/nltk.sentiment.html.

Beri, A. (2020, May 27). *Sentimental analysis using vader*. Medium. Retrieved September 15, 2021, from https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664.

Neha Seth. (2021, June 28). *Topic modeling and Latent Dirichlet Allocation (lda) using Gensim*. Analytics Vidhya. Retrieved September 14, 2021, from https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/.

Arun R., Suresh V., Veni Madhavan C.E., Narasimha Murthy M.N. (2010) On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In: Zaki M.J., Yu J.X., Ravindran B., Pudi V. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2010. Lecture Notes in Computer Science, vol 6118. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-13657-3_43

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2008). A density-based method for adaptive LDA model selection. Neurocomputing, 72(7-9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. Document Numérique, 17(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101(Supplement 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

MCMC for Lda - Markov chain Monte Carlo. Coursera. (n.d.). Retrieved September 14, 2021, from https://www.coursera.org/lecture/bayesian-methods-in-machine-learning/mcmc-for-lda-RnMwB.

Liu, S. (2019, January 6). *Dirichlet distribution*. Medium. Retrieved September 14, 2021, from https://towardsdatascience.com/dirichlet-distribution-a82ab942a879.

William M. Darling. (2011, December 1). A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling. Retrieved September 14, 2021, from https://www.ccs.neu.edu/home/vip/teach/DMcourse/5_topicmodel_summ/notes_slides/sampling/darling-lda.pdf.

Chapter 5:

Linguistic features · spacy usage documentation. Linguistic Features. (n.d.). Retrieved September 14, 2021, from https://spacy.io/usage/linguistic-features#noun-chunks.

Chapter 6:

Cushman & Wakefield. (2021). *Hospitality market in Belgium: Belgium*. Cushman & Wakefield. Retrieved September 14, 2021, from https://www.cushmanwakefield.com/en/belgium/insights/hospitality-market-in-belgium.