Interactive Analysis and Comparison of Hotels in Singapore

Cheng Yi Xing, 2201626, 2201626@sit.singaporetech.edu.sg

Declan Fong Yi Ren, 2201461, 2201461@sit.singaporetech.edu.sg

Koh Qin Jiang Darren, 2202575, 2202575@sit.singaporetech.edu.sg

Lan Yiling Elizabeth, 2201982, 2201982@sit.singaporetech.edu.sg

Lee Ying Zhen, 2201785, 2201785@sit.singaporetech.edu.sg

***Abstract*—(A Brief History of the Hotel Industry, n.d.) Hotels have been around for a long time and can be traced back to the early eight century in Japan in the form of guesthouses, other establishments were also constructed around the globe in countries such as the UK, US, Hong Kong, and Malaysia. In the very beginning these establishments were mostly patronized by wealthy travelers however, with the start of the 20th century saw globalization which gave the average person the affordability to travel and with that, hotels saw a huge influx of guests and coupled with the rapid advancement of technology, led to the rise of travel sites where potential guests can make informed decisions on their choice of stay. This research project aims to take an analytical approach to determine the factors that affect the overall satisfaction of hotel guests, and to streamline the process of selecting a hotel based on an individual's preference with the assistance of an interactive dashboard. Data in this project will be based on all hotels in Singapore, sourced, and structured from Booking.com via data crawling algorithms and analyzed through text mining. A word cloud would be generated to provide a quick overview of critical attributes travelers seek for when it comes to patronizing a hotel. Sentiment analysis via Valence Aware Dictionary and sEntiment Reasoner (VADER) identifies positive and negative words that have been tokenized from guest reviews. Through this information we are able to assign scores to each word and attain a final compound score where higher scores indicate a more desirable guest experience. We then proceed to develop a desktop browser and mobile based web- page where guests are able to leverage on an interactive dashboard with the assistance of attribute filters such as, ’Staff’, ‘Facilities’, ‘Cleanliness’, ‘Comfort’, ‘Value’ and ‘Location’. These filters aid potential guests in narrowing down their booking options and to experience a hassle-free approach to accommodation. A few brief insights in our project has shown that guests generally have no issues with higher hotel rates in exchange for an enhanced experience, guests who visit with their families are more likely to patronize higher star hotels while solo travelers tend to do the opposite. Last but not least, text analysis has revealed that guests pay heed to the hotel’s breakfast, check-in process, staff service and location.**

**Keywords— Hotel guests, Travelers, Informed decision,**

**Experience, Interactive dashboard, Text mining, Data crawling,**

**Singapore, Sentiment analysis, Web-based**

# *Introduction*

Travelers around the world might not share a common destination however, one commonly shared aspect would be accommodation. Settling down on a hotel might seem like an easy task but with a sea of information displayed on travel sites such as Booking.com, making an informed decision would require painstaking effort to sieve through countless reviews of a single establishment, these sites could display hundreds of such hotels at any given time, making the process an arduous one. Despite all the hassle, research has shown that travel review sites ultimately still affect a traveler’s hotel selection. Ines (2020) travelers are prone to visiting review sites to seek for reliable and unbiased insights to a hotel, and reviews by other guests are deemed more valuable than marketing efforts from hotels. Zhao, et al. (2015) also states, consumers tend to rely more on peer reviews than information provided by business entities because peer customers are more independent and trustworthy, plus consumer reviews may include critical information that hotels are reluctant to reveal to the public. The research team leveraged on technology such as the implementation of data crawling which has greatly assisted us in attaining all relevant data located on hotel website Booking.com in the year 2022, the next procedure taken was data cleaning which enhanced the overall reliability and accuracy of our data. This instrumental data was further separated into ‘General' and ‘Guest Review’ data, with the first consisting of establishment attributes such as ’Staff’, ‘Facilities’, ‘Cleanliness’, ‘Comfort’, ‘Value’ and ‘Location’ and the latter providing guest sentiments via the VADER method on their experience with the establishment. The project ultimately seeks to offer accurate and trustworthy information which can be further customized to meet the demands of potential hotel guests, and hotel management on a desktop browser and mobile based web-page. With this information, hotel guests are empowered to make decisive and well-informed selections on accommodation with ease.

# *Related works*

One key difference between our project and the research done in (Zhang et al., 2022) article is the lack of a range of hotel attributes utilized for performing a recommendation. (Zhang et al., 2022) The results of review classification can be presented to users as reasons for hotel recommendation when only the travel types of users are known. Therefore, the model in this study can be applied to the hotel recommendation system to help users make better decisions. Meanwhile, it can also help hotel managers to improve the service quality of hotels. In future work, the author will continue to consider the influence of other variables such as hotel star, specific location, price, room type, distance, and user attributes on user preference. Looking at a separate research article (Hargreaves, 2015), it differs from our project in terms of the scope of local hotels observed and time period of data pulled from these establishments. (Hargreaves, 2015) There are many aspects to measure a hotels’ performance in terms of its customer experience and customer satisfaction. Customer feedback in terms of ratings and reviews was collected from 5 hotels located in Singapore from 2005 to 2014 using Trip Advisor online travel agency data. The fundamental issue for the survival and growth of hotels is measurement and management of services quality and customer experience. The objective of this study is to analyze the customer ratings and reviews using statistical analysis, and to identify for each of the 5 Singaporean hotels, the attributes ratings by customers on location, sleep quality, room maintenance, service quality, value for money, cleanliness and overall. Further, to identify attributes on which each of the hotel stands out, factors which can be improved and the factors which influence customer satisfaction. In (Akhtar et al., 2017), the study did share a common toolkit with our project when it came to sentiment analysis for a quick comparison between various hotels. However, it did not conduct an extended study on the ratings of establishments based solely on key attributes such as ’Staff’, ‘Facilities’, ‘Cleanliness’, ‘Comfort’, ‘Value’ and ‘Location’. (Akhtar et al., 2017) The study also analyses the sentiment scores of the hotels on the basis of their aspects, which gives a better understanding as to which of the aspects of the hotel under study are better than the others as per the user comments and on which of these aspects more improvement needs to be done. The hotel review summary this paper provides and sentiment score are separate from the ranking provided by the website and can be more realistic than the one provided by the website.

There are a variety of directions that this work can take. We can also rate the hotels as per the aspects crawled[14], i.e. how is the hotel in terms of the location for example. This can better help the customer in his decision making process of which hotel to choose as per his/her requirement. This can also be of help to the hotel management because they will now be aware of what areas they need improvement in and what their strong points are.

# *Methods/Approach*

*Scraping*

The *Scrapy* and *Twisted* (to run Scrapy without the command line) packages were used to scrape the links for all the hotels from the Booking.com search results. The *BeautifulSoup*, *Requests, CSV, Time, Itertools* and *Random* modules were used in the extraction part of the scraping. The following pseudo-code explains how the whole scraping process works:

1. Begin procedure

2. For each page from the search results:

a. Get the link for each hotel and append it to the list of links

3. Loop through the list of links to make sure the links are compatible with the extraction part:

a. Check if “hotel” is in the URL string

i. Replace “.html” with “.en-gb.html” and the aid with “304142”

4. To extract the general data of each hotel:

a. Cycle the user agent to prevent being blocked by booking.com

b. Extract the general data fields using the respective link from the list

c. Write the row to the general data CSV file

5. To extract the review data of each hotel:

a. Cycle the user agent to prevent being blocked by booking.com

b. Set a random delay to prevent being blocked by booking.com

c. To extract reviews from each page:

i. Form the page URL using the hotel id and page number

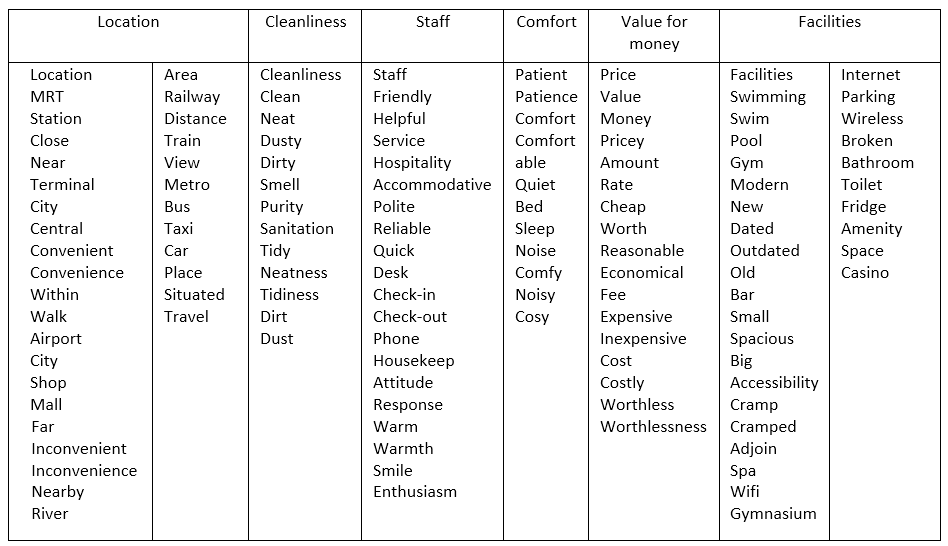
ii. Extract all the reviews

d. Write the hotel reviews to the review data CSV file

6. End procedure

*Classification of Reviews*

The first step for classifying the reviews was to define some associative words for each category, as shown in (Table 1). The associative words were stored in a CSV file.

Table 1: Manually Defined Associative Words Table

The Pandas package is used and the following pseudo-code explains how the review classification works:

1. Begin procedure

2. Read the associative words and review data CSV files using Pandas

3. Repeat the following for both positive and negative reviews for each row of the review data DataFrame:

a. Set the category as “no\_cat” if the review string is “no\_pos\_des” or “no\_neg\_des”

b. Set the score for each category to 0

c. Check if each word in the review contains an associative word

d. If the word is associated with a category the category score is incremented

e. The review is classified as the categories with a score of 1 or more

f. If the review is not classified, then it is classified as “General”

4. Add the classification of each review to the review data and save it as a CSV file

5. End procedure

*Sentiment Analysis*

The *Pandas, Numpy, NLTK* and *TQDM* packages are used. The *Vader* and *SentimentIntensityAnalyzer* modules from *NLTK* are used. The following pseudo-code explains how the sentimental analysis is done:

1. Begin procedure

2. Read the review data CSV file

3. Repeat the following for both positive and negative reviews:

a. Calculate the polarity score for each row

b. Merge the DataFrame into the review DataFrame

4. Save the final merged DataFrame as a CSV

*Word Clouds and Word Frequency for Positive and Negative Reviews*

A positive word cloud and a negative word cloud are generated

The *wordcloud* and *Pandas* packages are used and the following pseudo-code explains how the word clouds are generated:

1. Begin procedure

2. Read the review data CSV file using pandas

3. Filter the resulting DataFrame by hotel\_id, user\_country and type\_of\_travel

4. Convert the filtered DataFrame to a dictionary and then to text

5. Set the stop words

6. Add “no\_pos\_des” or “no\_neg\_des” depending on the word cloud and “hotel” to stop words

7. Generate the word cloud using the resulting text from step 4

8. Save the word cloud as a file

9. From step 4, separate words by spaces and store each word into list

10. Count the frequency of words in the list and store word frequency as dictionary

11. Sort dictionary

12. Retrieve 5 reviews containing the word

*Website*

A website was chosen as the GUI. The website consists of the frontend, backend and datasets acting as the database.

Nextjs, React and MUI Material React library was used for the frontend.

The backend uses the Flask, Flask\_Restful, Pandas and AST packages

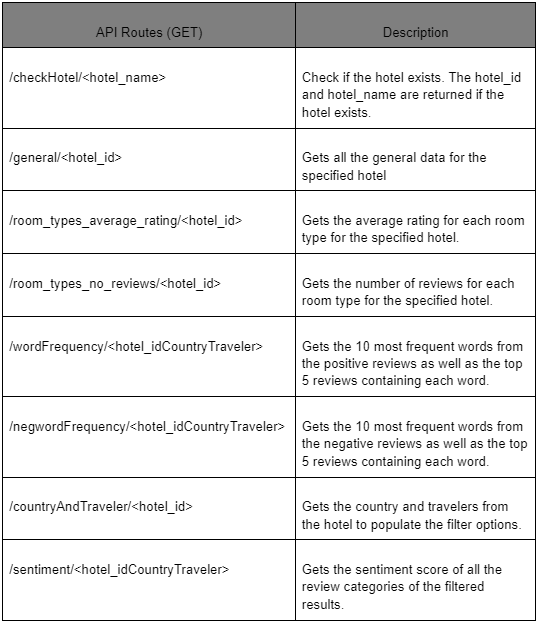


Table 2: Website Table

# *Dataset used*

We used *Scrapy, Requests,* and *BeautifulSoup* to extract the massive data located on travel review site Booking.com. The site consists of a variety of travel features worldwide, such as hotel accommodations, flights, car rentals, attractions, etc. To focus on our problem statement, we first filter the site to display only hotels in Singapore. These hotels range from 0 to 5 stars. Our team used *Scrapy* to extract the links to all the hotels in Singapore on Booking.com. With those links, *Requests* is used to request for the HTML of the link. *BeautifulSoup* is then used to extract the retrieved HTML elements and its contents. We proceed to save the data collected, into a csv format. After which, we have 2 finalized datasets, namely “General Data” and “Reviews Data”, both of which cover feedback and reviews from hotel guests, since 2019. “General Data” includes data from about 260 different hotels and “Reviews Data” consists of more than 170,000 worth of reviews left by actual hotel guests. The 2 datasets can also be merged using its primary key - hotel\_id, in order to draw further analysis and create potential relationships. Refer to (Figure 1) and (Figure 2) for the data structure of the dataset.

* General Data
* Broader dataset
* Overview of the hotel
* Covers common areas hotel guests look out for in their hotel decision making process
* Reviews Data
* More detailed dataset
* Analyses hotel guest’s actual review
* Analyses hotel guest’s hotel stay booking details
* Analyses profile of hotel guests

With these techniques we will be able to structure critical attributes of the review site - Booking.com and begin our logical approach to analyzing the data.

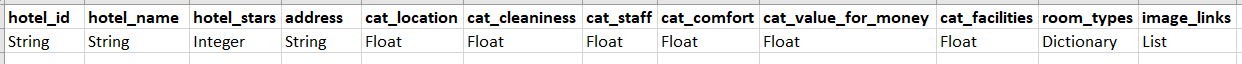


Figure 1: General Data



Figure 2: Reviews Data

# *System diagram*

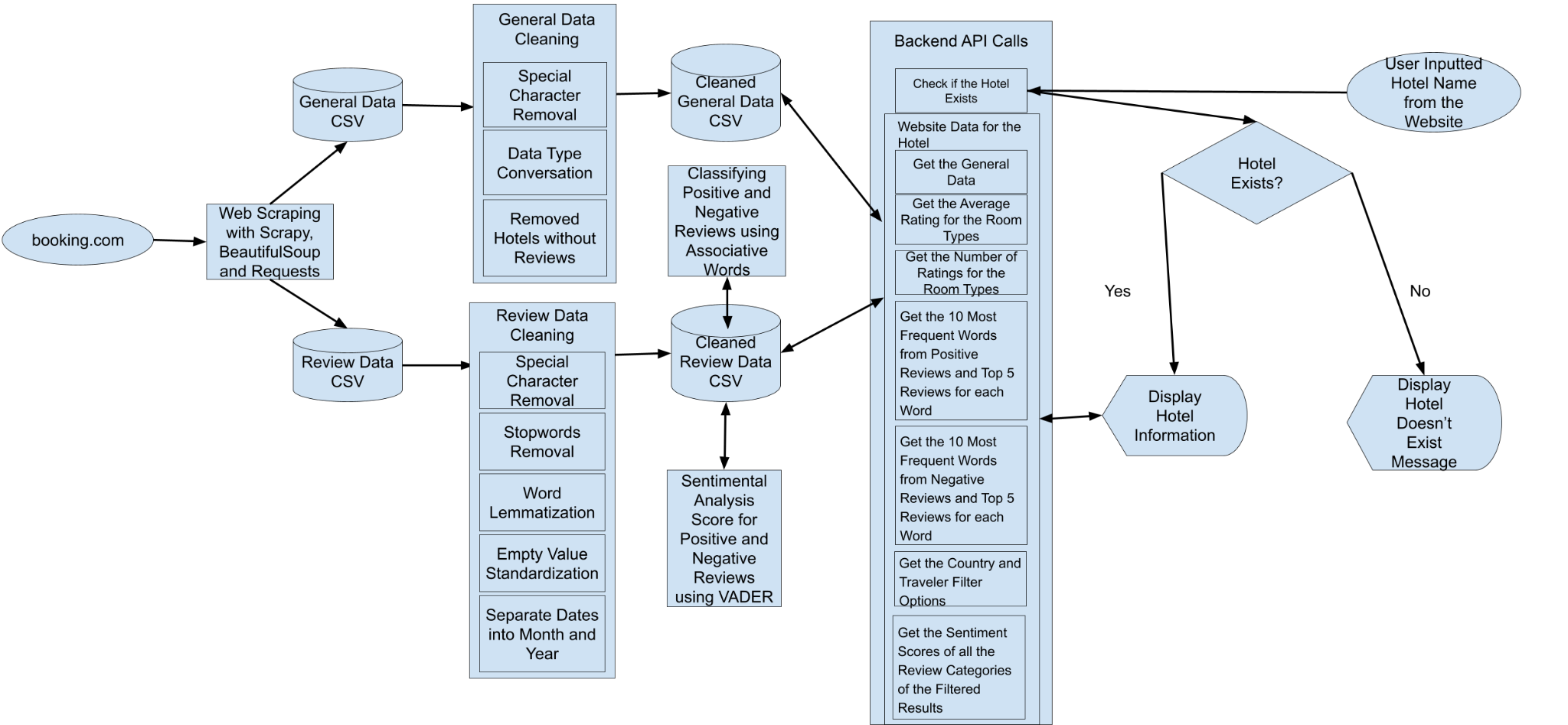


Figure 3: System Diagram

*Refer to Figure 3: System Diagram in Appendix For a Larger Chart*

*Data pre-processing*

General data preprocessing steps include:

1. Removing special characters for room\_types (Figure 4: room\_types)
2. Added 25, 50, 75 percentile prices columns (Figure 5: roomtypes\_clean, 25percentile, 50percentile, 75percentile)
3. Removed hotels with no reviews (Figure 6: Removed Hotels)

Review data preprocessing steps include:

1. Removed special characters, emojis, words in other languages (Figure 7 and 8: Removed Special Characters)
2. Removed Stop words, lemmatized reviews data (Figure 9 and 10: title\_cleaned, pos\_des\_cleaned)
3. Standardized Empty, NA, Nil Values
4. Separated date into month and year (Figure 11: Date columns)

* Classifier

We added a list of associative words under these 6 categories. For the positive and

negative reviews, we check if there are words in the reviews that matches the list of

associative words, it will add scores to the specific category(s). The category with the highest score will be the review category.

* Sentiment Analysis

There is an algorithm built in VADER which automatically cleans and gives each word in both positive\_des and negative\_des columns a score. From each word, they add up to get a final compound score. The higher the compound score, the better the overall sentiment of the guest who submitted that review.

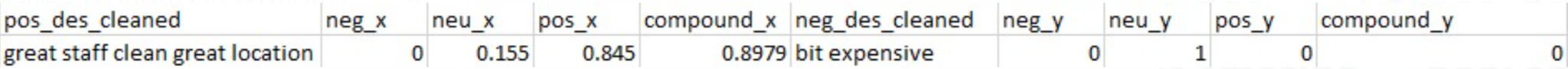


Figure 12: Sentiment Analysis

# *Data analysis*

Using both General Data and Reviews Data, we are able to draw insights and conclusions about the hotel industry. The analysis will be focused on looking at hotels ranging from 1 to 5 stars.

****

Figure 14: Average Rating of Hotels Across Different Stars

(Figure 14) displays the average overall hotel rating out of ten, across different hotel stars. Similarly, we can see that the average hotel rating increases, across the increasing number of hotel stars. The average rating of hotels across different stars increases as the number of stars increases, with the exception of 3 star hotels which have a lower average rating than 2 stars. This suggests that travelers on a budget may want to go for 2 stars hotels rather than 3 stars hotels.

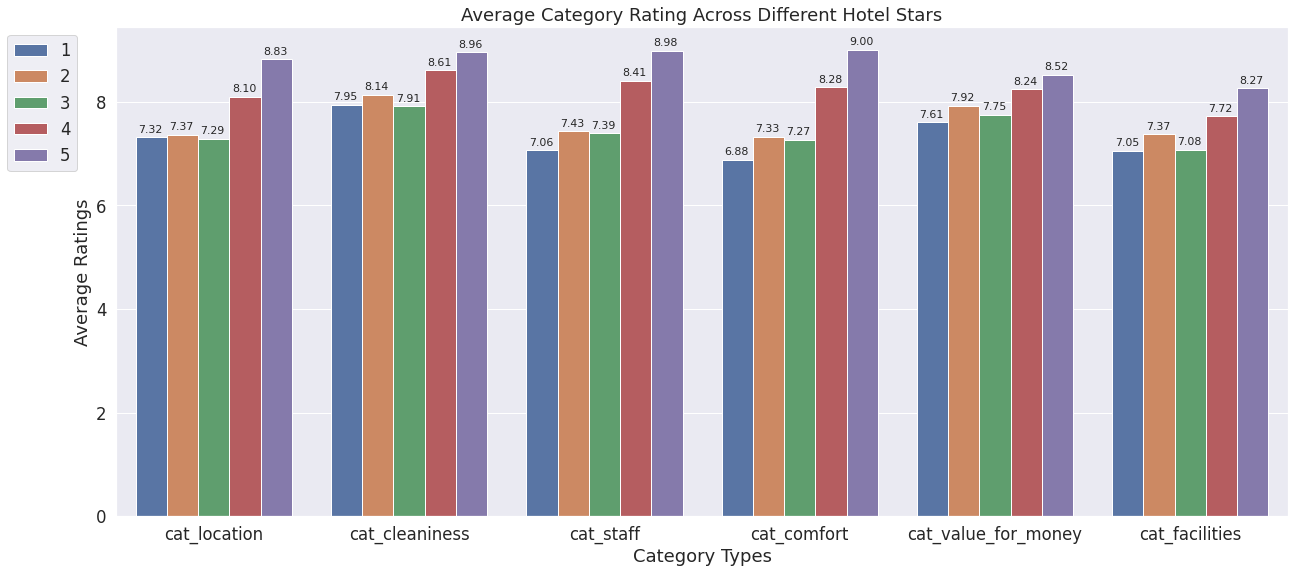


Figure 15: Average Rating on Different Categories Across Different Stars

The ratings can be further categorized into 6 different categories, Location, Cleanliness, Service Level of Staff, Comfort, Value for Money, and Hotel Facilities. Generally, the higher the hotel stars, the higher the ratings across all categories. It is observed that 4 and 5 star hotels stand out in terms of their staff and comfort, with ratings much higher than the other star hotels. Comparing 2 and 3 stars hotels, 2 star hotels have higher ratings across all categories. It suggests that 2 star hotels are able to deliver better results than guest expectations compared to 3 star hotels. Ratings on value on money for 1 star hotels are significantly higher than its other ratings. 5 Star hotels may have a much higher price range than others but guests still gave 5 star hotels a high rating for Value for money. Guests may think that 5 stars hotels are worth it as they are willing to spend more to get better location, cleanliness, comfort, service and facilities. This can be suggested by the high ratings guests gave for other categories.

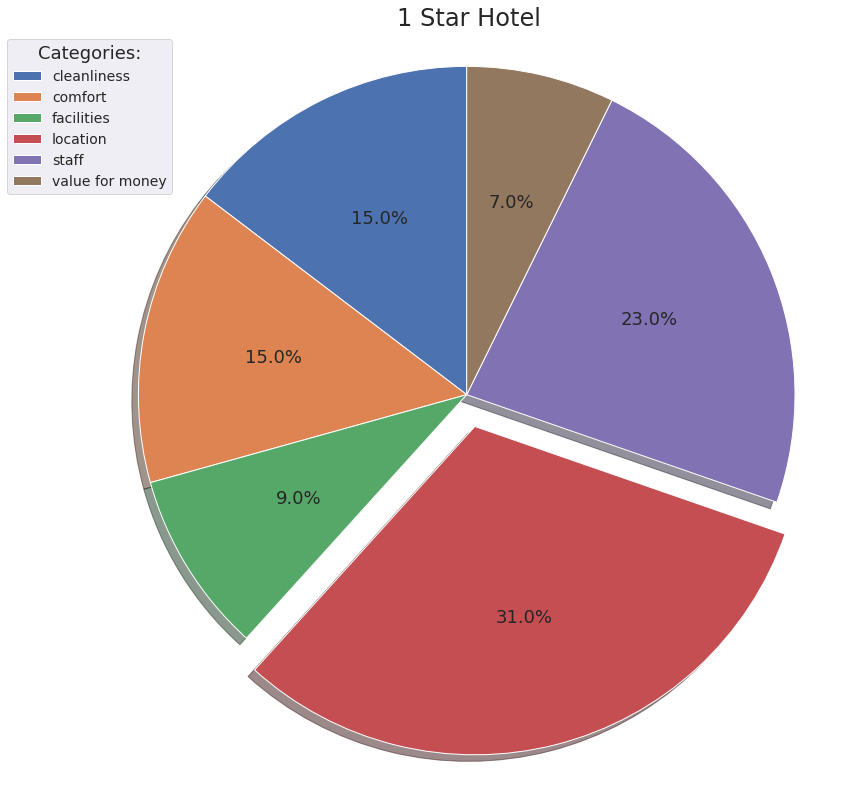
****

Figure 16: Percentages of Different Categories for Positive Reviews for 1 Star Hotels

*Refer to Figure 17, 18, 19, 20 for Percentages of Different Categories for Positive Reviews for 2-5 Star hotel*

In this segment, the percentage of each category (Location, Staff etc) that makes up the positive reviews of each hotel star is analyzed. This helps to identify which areas are more appreciated by guests, and identify any underlying trends across the different hotel stars. Generally, across all stars hotels, the location category takes up the highest percentage of positive reviews. This suggests that hotels in Singapore are often in locations which are seen as prime and convenient for the guests. The percentage for facilities category increases as the number of hotel stars increases. This implies that higher stars hotels have facilities that are much more appreciated by the guests. It is observed that the percentage of positive reviews for the “value for money”' category decreases as hotel stars increase. The staff and facilities percentage reviews for 5-star hotels stands out when compared to other star hotels. This indicates that 5-star hotels are excellent at ensuring their staff provides top-notch service and their facilities are highly regarded.

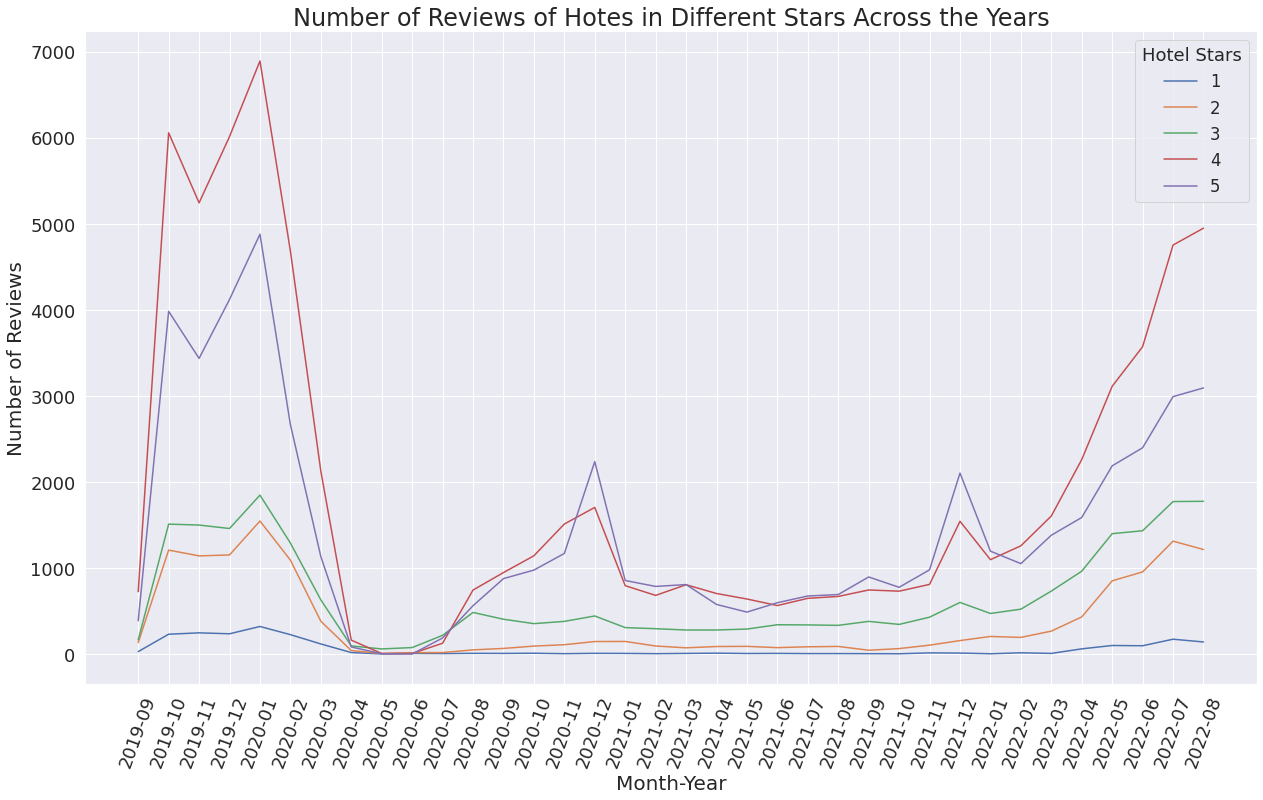


Figure 21: Total Reviews of Different Hotel Stars From Sept 2019 to Aug 2022

The number of reviews during a period of time can provide a picture of the demand of the hotels. After April 2020, demand for hotels dropped to almost 0 due to Singapore closing down her borders and entering into circuit breaker. In July 2020, the Singapore government announced the reopening of hotels which caused the demand to increase till december 2020 where it saw a steep drop in demand for higher stars hotels. Since december 2020, there has been a steady increase in demand, showing that the hotel industry is slowly getting back its pre-covid days.

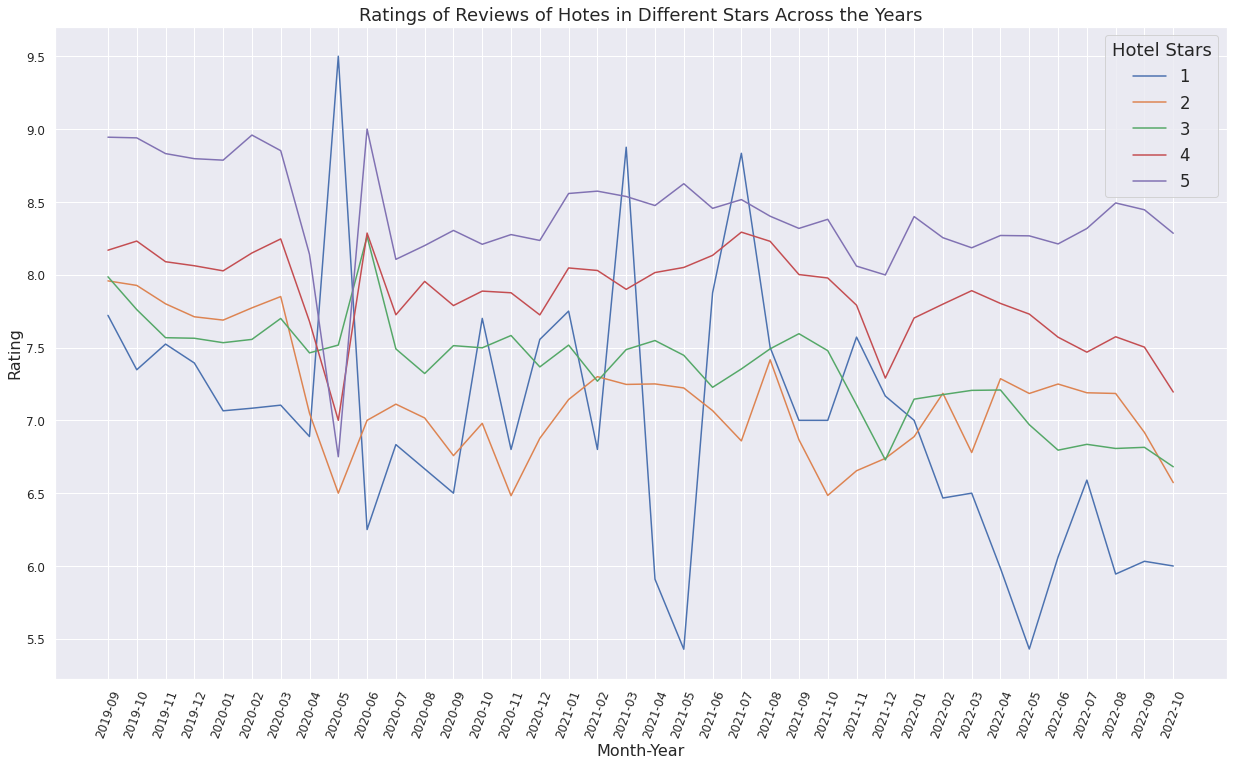


Figure 22: Average Ratings of Different Star Hotels From Sept2019 to Sept 2022

However, despite the increasing demand, the ratings have been on a downwards trend. It is observed that in may 2020, there is a steep drop in reviews due to circuit breaker and the ratings shot back up when Singapore announced the reopening of hotels in June 2020. Generally, the ratings of 2-5 star hotels follow a similar trend but ratings of 1 star hotels fluctuate a lot. Since may 2021, the ratings for 1-4 star hotels are on a downtrend while 5 star hotels are showing improvements to its rating steadily.

In this segment, 5 hotels with the highest number of reviews from each of the 5 different star categories have been chosen to do further analysis on. The hotels selected are All day in checkers inn (Cambell), Circular House (Circular), Aerotel Singapore (Aerotel), Hotel Boss (Boss) and Marina Bay Sands (Marina).

*Type of travelers (Solo, Family) and the percentage they make up of for the 5 hotels*

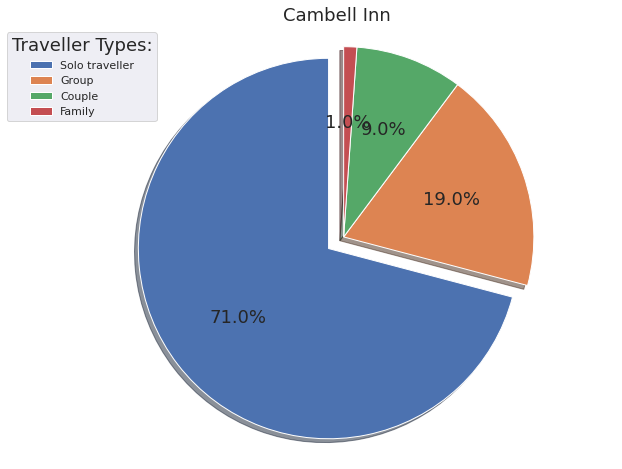


Figure 23: Percentages of Different Types of Travelers for All Day in Checkers Inn (Cambell)

*Refer to Figure 24, 25, 26, 27 for Percentages of Types of Travelers for 2-5 Star hotel*

Guests were classified into 4 segments, Family, Solo, Couple and Group. Generally, solo travelers make up the majority of the guests for Cambell, Circular and Aerotal, while Couples make up the majority in Boss and Marina. Percentage of guests traveling as a family increases as the number of stars increases, from 1.0% in Cambell to 33.0% in Marina. Inversely, the percentage of guests traveling solo decreases as the number of stars increases, from 71.0% to 8.0%.

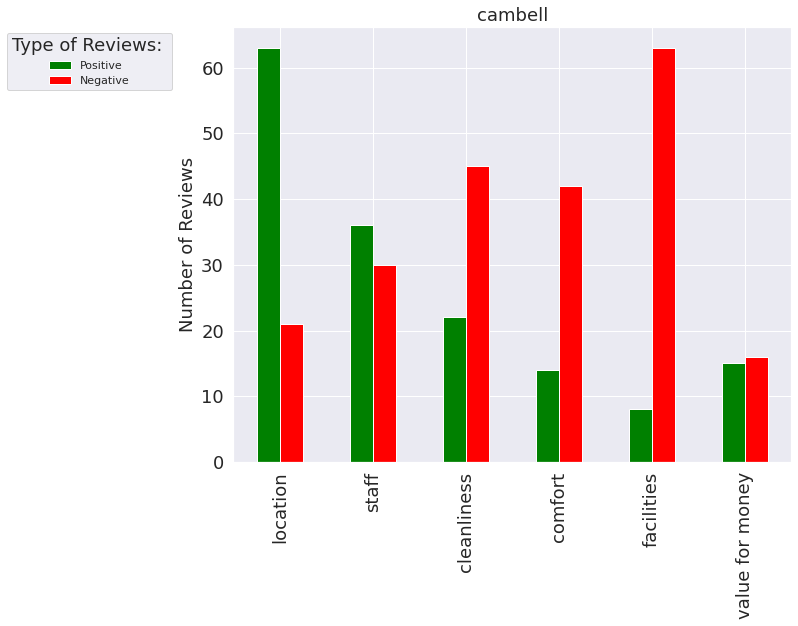


Figure 28: Percentages of Different Types of Travelers for All Day in Checkers Inn (Cambell)

*Refer to Figure 29, 30, 31, 32 for Percentages of Types of Travelers for 2-5 Star hotel*

Cambell has an overwhelming amount of negative reviews on Cleanliness, Comfort and Facilities. This suggests that due to the lacking comfort, cleanliness and facilities of the hotel, customers have given negative reviews on the value for money too despite the prices being low. Even though Boss is a 4 star hotel, there are more negative reviews on the facilities than positive. Marina has more positive reviews in all categories except for Value For Money, likely due to its high pricing. Although Aerotel has higher positive reviews on all categories, it has higher negative reviews for Value for money. The price of $445 which likely what many guests think would be too much for a 3 star hotel where the median price is. This could be due to the high price which many guests may think would be too expensive for a 3 star hotel where the median price is $350.45.

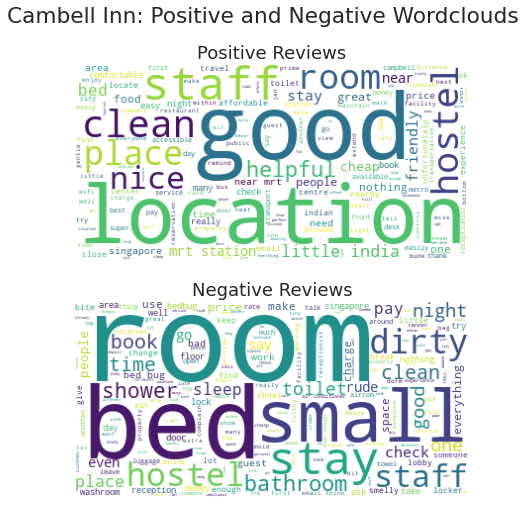


Figure 33: Text Analysis Word Cloud For Cambell Inn

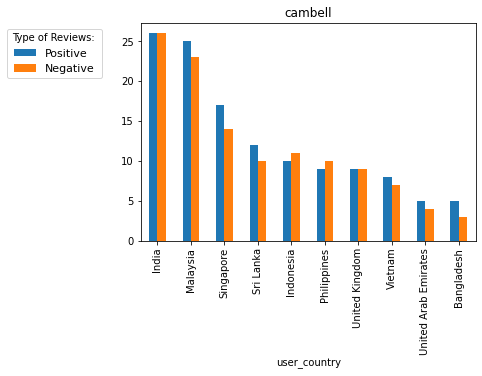


Figure 34: Top 10 Countries That Cambell Inn Guests Are From

*Text analysis on Cambell’s positive reviews suggests the following*

* Guests find the location at Little India to be great and service to be excellent. This is suggested by the words “location”, “staff” and “little india”. It is also backed by the high positive to negative reviews on Location and Staff.
* The hotel is popular with indians. This is implied by the word “indian” and also from (Figure 33) where we can see travelers from india gave the most amount of reviews.

*Text analysis on Cambell’s negative reviews suggests the following*

* Guests are not satisfied with the quality of their room in terms of cleanliness and size. This is hinted at by the words “bed”, “dirty”, “room” and “small”. There is also a high amount of negative reviews on Cleanliness and Comfort which further insinuate the statement.

*Text analysis on Circular House positive reviews suggests the following*

* Guests are satisfied with the overall experience of living in the unique pods of the hotel. This is suggested by the words “location”, “clean” and “staff” and also the high number of positive reviews on those areas.

Refer to Figure 35: Text Analysis Word Cloud For Circular House

*Text analysis on Circular House negative reviews suggests the following*

* Guests find the breakfast served at the hotel to be inadequate. This is implied by the word “breakfast”
* Some guests also find the room’s (pods) cleanliness to be lacking. This is suggested by the words “room”, (not)“clean” and “dirty”
* The words “clean” and “room” appeared in high frequency in both the positive and negative reviews. This may imply that the hotel has been inconsistent in ensuring the cleanliness of the rooms.

Refer to Figure 36: Text Analysis Word Cloud For Aerotel Hotel

*Text analysis on the positive reviews of Aerotel Hotel suggests the following*

* The hotel’s strongest areas are cleanliness and comfort of their rooms. This is suggested by the words “room”, “clean”, “comfortable” having high frequencies in positive reviews. It is further supported by an extremely high positive to negative ratio on Cleanliness and Comfort.

*Text analysis on the negative reviews of Aerotel Hotel suggests the following*

* Guests may find the booking and checking in processes to be not satisfactory. This is hinted at by the words “book”, “check”, “staff” and “hour”.
* Guests may find that the prices of the hotels are too expensive for a 3 star hotel. This is suggested by the word “expensive” and the high number of negative reviews for Value of Money.

Refer Figure 37: Text Analysis Word Cloud For Hotel Boss

*Text analysis on positive reviews of Hotel Boss suggests the following*

* Guests find the location of the hotel to be prime and convenient. This is suggested by the words “location”, “good location”, “close”, “near”. Hotel boss is located in central and has 4 MRT within 1km away and many of Singapore's top attractions are near the hotel too

*Text analysis on negative reviews of Hotel Boss suggests the following*

* Guests may find the rooms to be too small. It is implied by the words “small” and “room”. The high negative reviews on Comfort also suggests that.
* Guests find the breakfast served at the hotel to be inadequate. This is implied by the word “breakfast”

Refer Figure 38: Text Analysis Word Cloud For Marina Bay Sands

*Text analysis on positive reviews of Marina Bay Sands suggests the following*

* Guests are impressed by the world-renowned hotel facility, infinity pool, which offers a great view. This is suggested by the high frequency of words like “Infinity pool” and “View”. The positive to negative review on facilities are also high

*Text analysis on negative reviews of Marina Bay Sands suggests the following*

* Guests may be dissatisfied with the check in process and the price of Marina. This is suggested by the words “Room”, “Check”, “Price” and “Expensive”.

# *Result and insights*

In summary, the average overall rating for 3-star hotels is lower than 2-star hotels as well as across the 6 categories. This suggests that 2-star hotels are exceeding expectations compared to 3-star hotels. Despite the high prices, guests from 5 star hotels find their stay there to be value for their money. For guest segments, families tend to visit higher star hotels while solo travelers visit lower stars hotels. In light of Singapore opening her borders, demand for hotels has been steadily increasing since 2022, however, the ratings have been on a downward trend. Based on the text analysis conducted, hotel guests tend to review the hotel’s breakfast, checking in process, service level of staff and convenience of location.

Data is known to be exclusive to every hotel, as no 2 hotels will have the same datasets. The project aims to tackle this by creating a website allowing users to search for a hotel and get analytics on more specific information, such as room types, room prices, and rating for each category.

The website will be analyzing

* Room types
* Room prices and it’s average ratings
* Room type demands over the years
* Country and Type of traveler filters for hotel review analysis
* Top 10 frequent words for positive and negative reviews and 5 reviews that have the word
* Overall sentiment score and number of positive and negative on the 6 categories of review (Location, Comfort, Cleanliness, Facilities, Staff and Value for Money)

Refer Figure 39: Website Overview

# *Conclusion*

Through secondary research, we realized that travelers often find it difficult to decide on which hotel to choose due to information overload. Therefore, there is a need for a user-friendly platform which allows users to draw insights and filter out information that is relevant to them.

*Refer to Table 3 For Project Limitation and Future Work*

# *References*

[1] A Brief History of the Hotel Industry. (n.d.). Customer Alliance.<https://www.customer-alliance.com/en/resources/article/hotel->

[2] Akhtar, N., Zubair, N., Kumar, A., & Ahmad, T. (2017). Aspect based Sentiment Oriented Summarization of Hotel Reviews. Procedia Computer Science, 115, 563–571.<https://doi.org/10.1016/j.procs.2017.09.115>

[3] Barreiros, I. (2020, February 26). How Hotels Can Leverage “Social Proof” to Increase Direct Bookings. GuestCentric.<https://blog.guestcentric.com/how-hotels-can-leverage-social-proof-to-increase-direct-bookings/?utm_source=HotelBookingsVsGuestDecisionMakingProcess&utm_medium=September2021&utm_campaign=HospitalityNet>

[4] Barreiros, I. (2021, September 13). Hotel Bookings - What is the Guest’s Decision-Making Process? | By Ines Barreiros. Hospitality Net.<https://www.hospitalitynet.org/opinion/4106390.html>

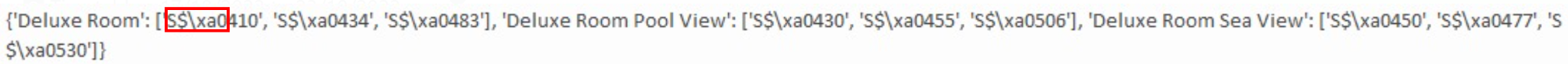
[5] Hargreaves, C. (2015, September 17). (PDF) Analysis of Hotel Guest Satisfaction Ratings and Reviews: An Application in Singapore. ResearchGate. <https://www.researchgate.net/publication/291832800_Analysis_of_Hotel_Guest_Satisfaction_Ratings_and_Reviews_An_Application_in_Singapore>

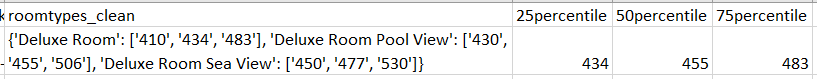
[6] Zhang, L., Guo, J., Kang, R., Zhao, B., Zhang, C., & Li, J. (2022). Hotel Review Classification Based on the Text Pretraining Heterogeneous Graph Neural Network Model. Computational Intelligence and Neuroscience, 2022, 5259305.<https://doi.org/10.1155/2022/5259305>

# *Appendix*

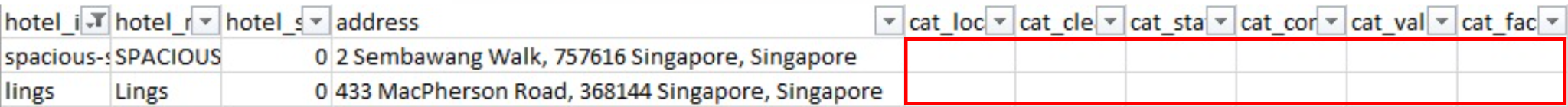
|  |
| --- |
| *Project Limitations* |
| 1. General data scraped from booking.com are not the most updated. The prices and room available for each hotel changes each day 2. Did not do in depth testing on GUI, so there are bugs around that is yet to be resolved 3. Data scraper took 16 hours to scrape all the reviews of all the hotels. 4. Classifier algorithm confidence score can be improved |
| *Future Implementations (Process and Improvements)* |
| *Data Scraper*   1. Automate scraping process in the website’s backend in order to extract live data from booking.com 2. Use threading to concurrently run the scraper to scrape more efficiently and quicker 3. Use rotating proxies in order to prevent website server from blocking IP address   *Data Analytics*   1. Improve the classifier by manually adding more associative words and write a program that uses NLTK to add synonyms for each word to the associative words CSV file. The minimum score to categorize the reviews can also be increased. 2. Replacing the VADER model with the RoBERTa model which is significantly more accurate and consistent when it comes to recognizing relationships between words, and the context of the sentence. As a deep learning model RoBERTa is more inclined to recognize text based sarcasm and also understand memes naturally like a human. All in all this model would churn out far more reliable results and provide quality insights to guest reviews.   *Website*   1. Under the positive and negative reviews segment, includes 5 reviews that has either the highest and lowest sentiment score, instead of the first 5 reviews that contains the word 2. Add a “Download Report of Hotel” feature so that hotel management is able to download a copy of more detailed analysis of their hotel. A separate dashboard catered to hotel management would be advantageous, as the positive and negative insights would serve to identify opportunities of advancement in providing guests with the best experiences. 3. Handle errors and bugs that may arise from the website 4. Comparison feature where users are able to compare hotels side by side 5. Conduct primary research on potential users to better understand their needs so as to add features to the website |

*Table 3: Project Limitations and Future Implementations (Process and Improvements)*

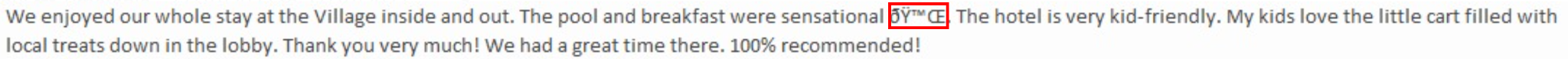
*Figure 4: room\_types*



*Figure 5: roomtypes\_clean, 25percentile, 50percentile, 75percentile*



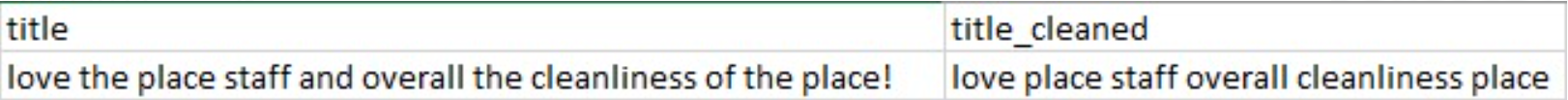
*Figure 6: Removed Hotels*



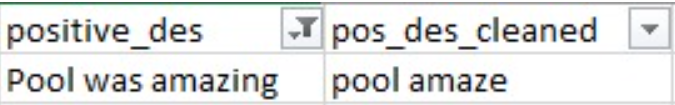
*Figure 7: Removed Special Characters*



*Figure 8: Removed Special Characters*



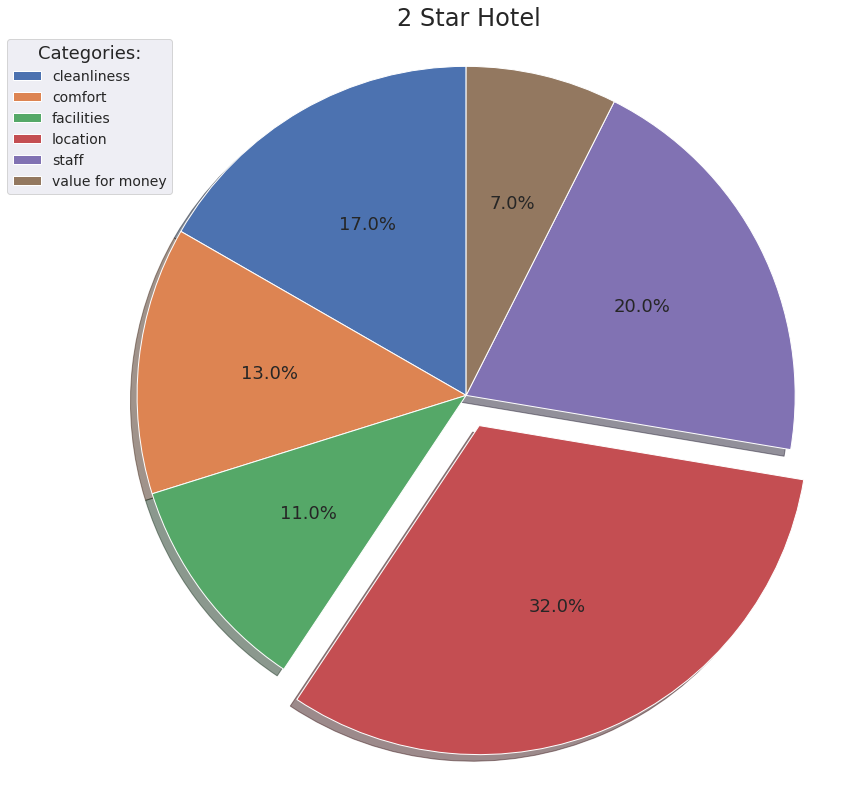
*Figure 9: title\_cleaned*



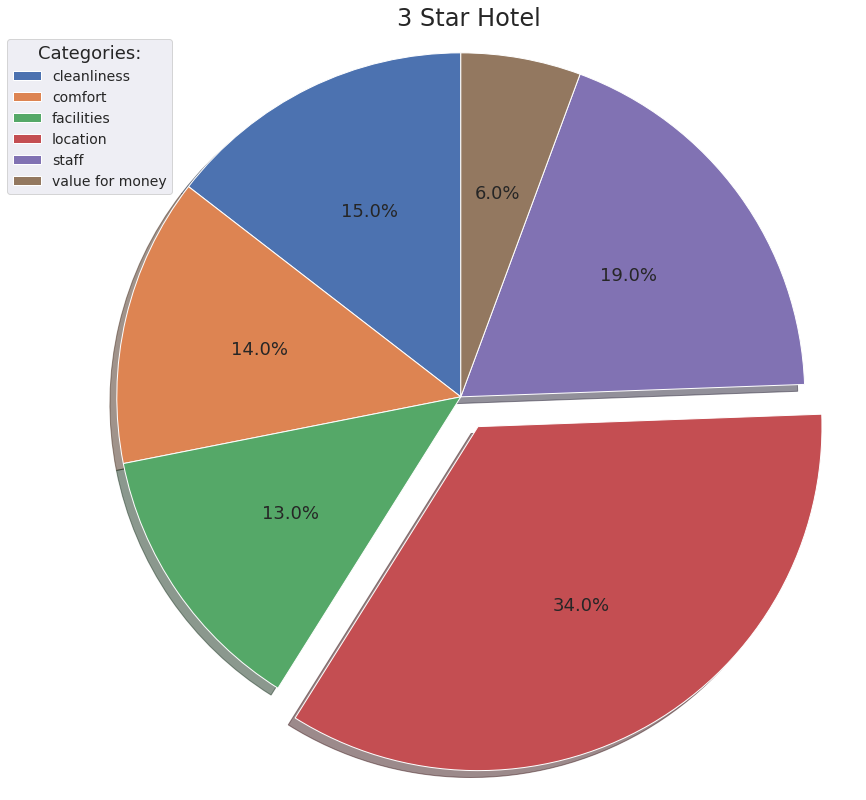
*Figure 10: pos\_des\_cleaned*



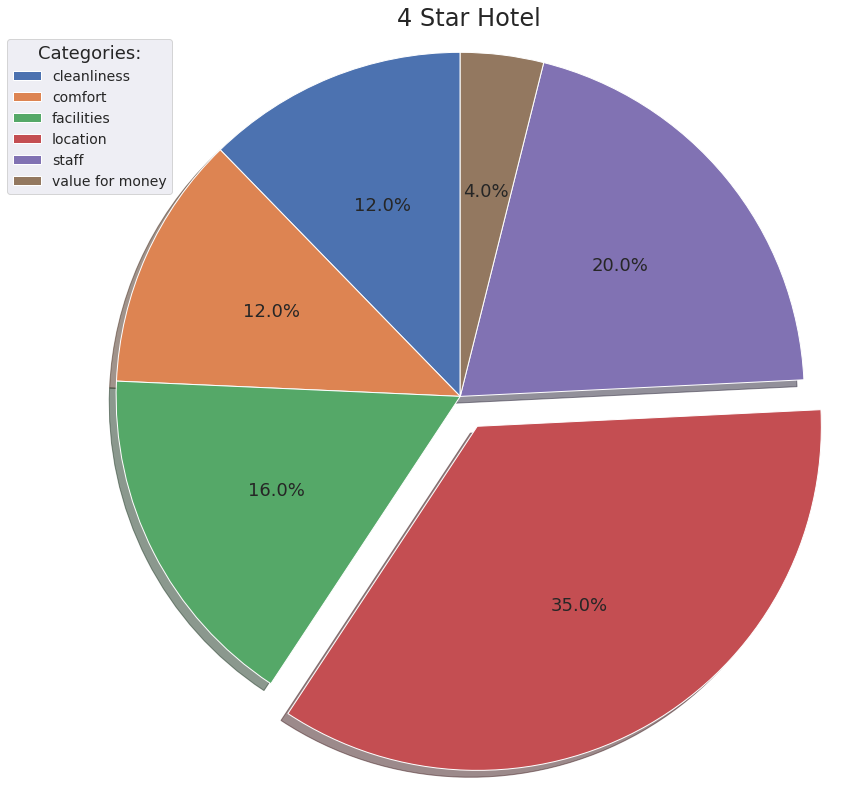
*Figure 11: Date Columns*



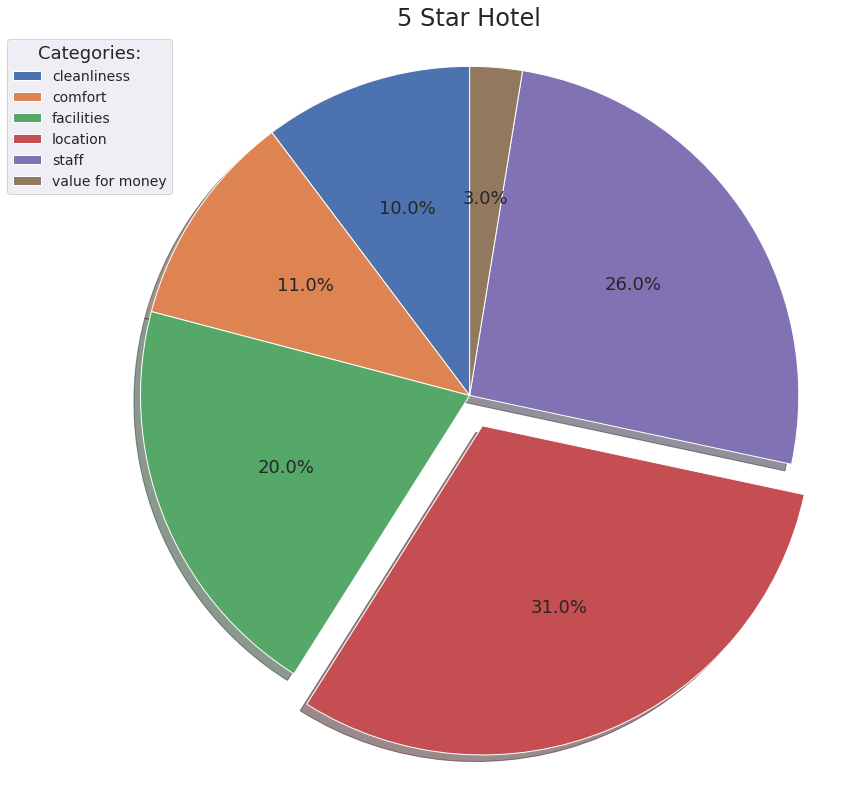
*Figure 17: Percentages of Different Categories for Positive Reviews for 2 Star Hotels*



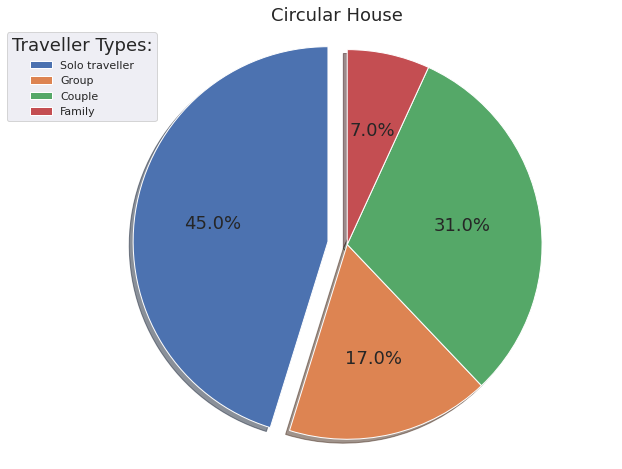
*Figure 18: Percentages of Different Categories for Positive Reviews for 3 Star Hotels*



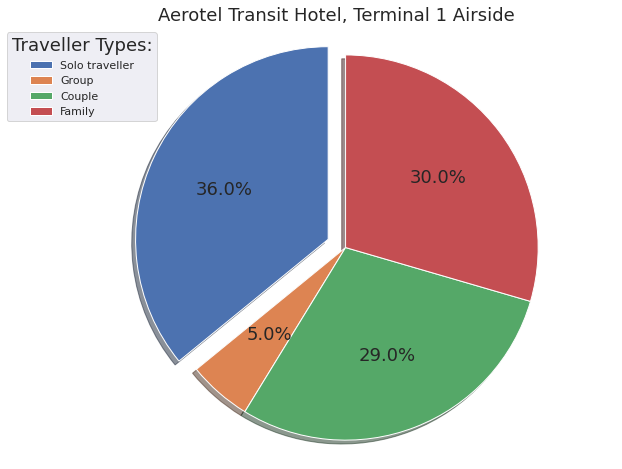
*Figure 19: Percentages of Different Categories for Positive Reviews for 4 Star Hotels*



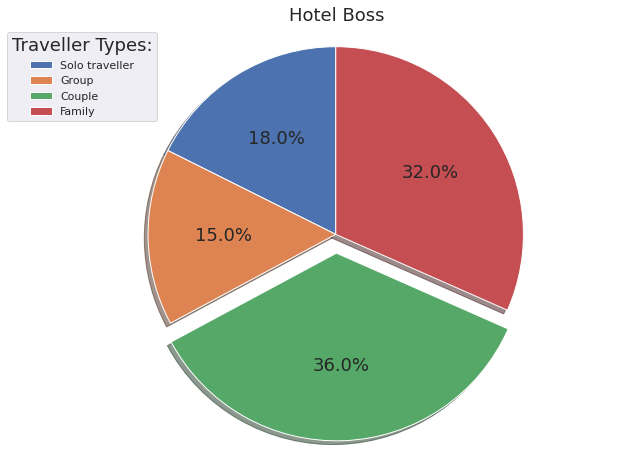
*Figure 20: Percentages of Different Categories for Positive Reviews for 5 Star Hotels*



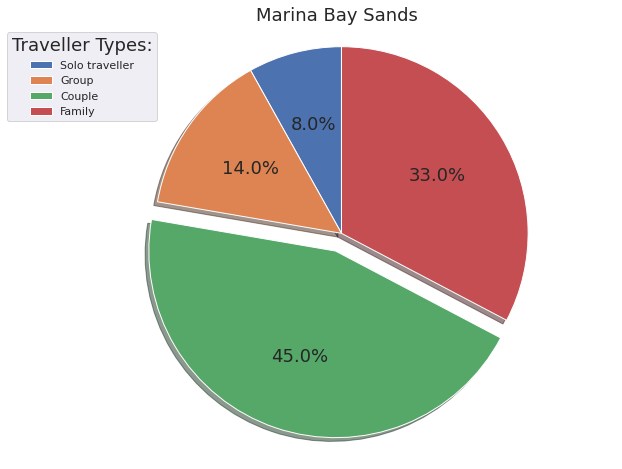
*Figure 24: Percentages of Different Types of Travelers for Circular House (Circular)*



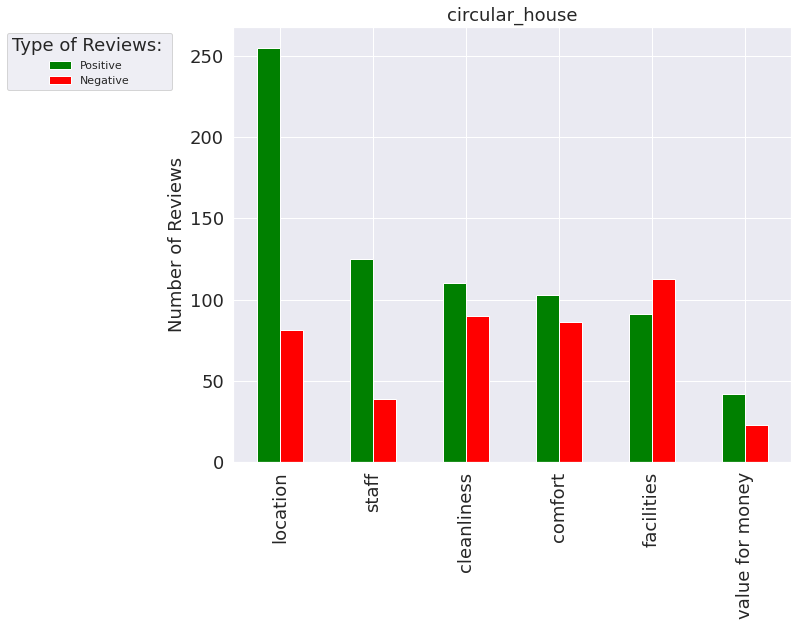
*Figure 25: Percentages of Different Types of Travelers for Aerotel Singapore (Aerotel)*



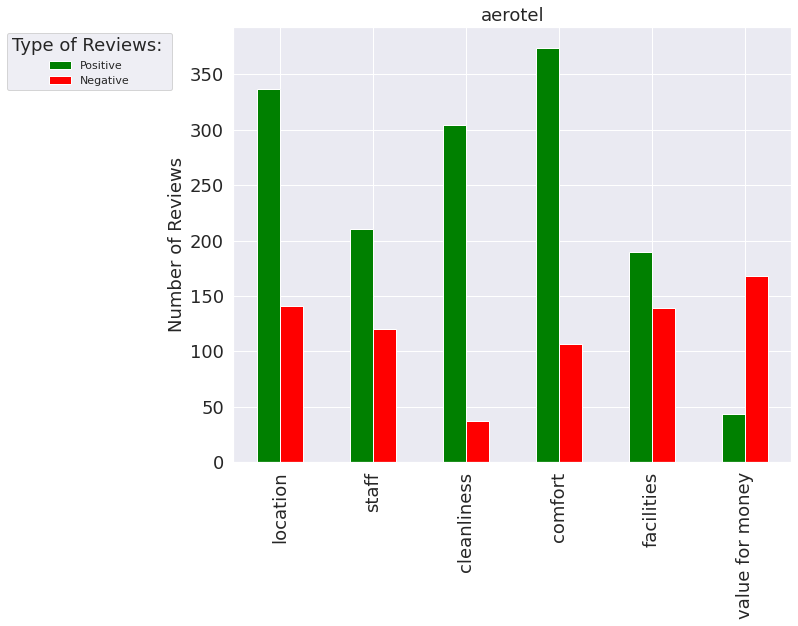
*Figure 26: Percentages of Different Types of Travelers for Hotel Boss (Boss)*



*Figure 27: Percentages of Different Types of Travelers for Marina Bay Sands (Marina)*



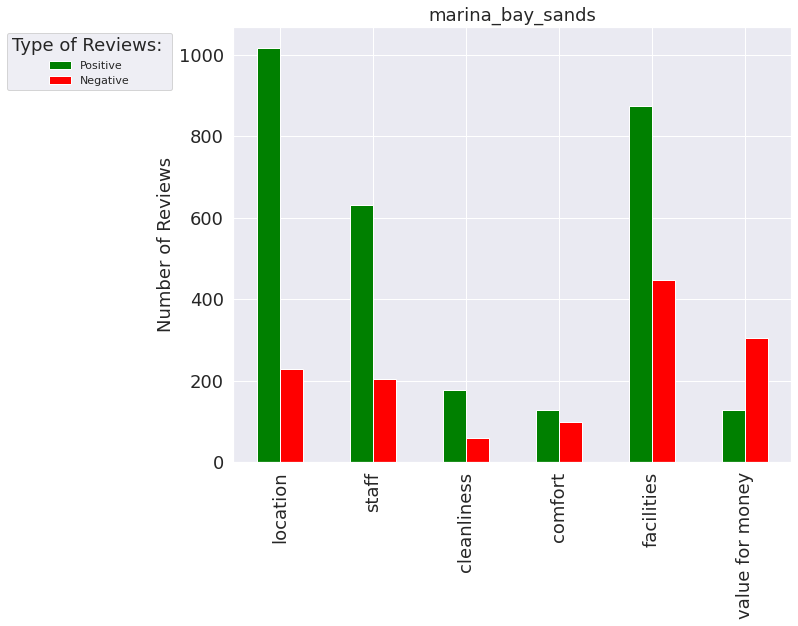
*Figure 29: Percentages of Different Types of Travelers for Circular House (Circular)*



*Figure 30: Percentages of Different Types of Travelers for Aerotel Singapore (Aerotel)*



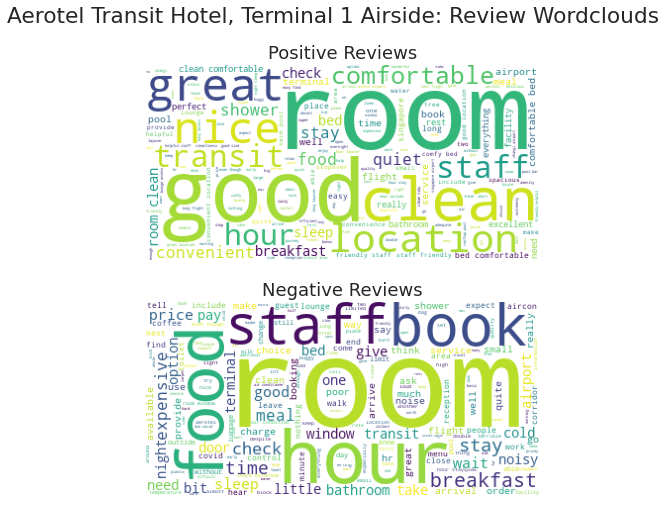
*Figure 31: Percentages of Different Types of Travelers for Hotel Boss (Boss)*



*Figure 3: Percentages of Different Types of Travelers Marina Bays Sands (Marina)*



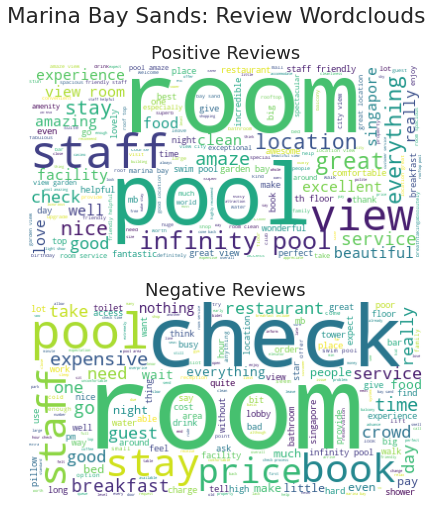
*Figure 35: Text Analysis Word Cloud For Circular House*



*Figure 36: Text Analysis Word Cloud For Aerotel Hotel*

​​

*Figure 37: Text Analysis Word Cloud For Hotel Boss*



*Figure 38: Text Analysis Word Cloud For Marina Bay Sands*



*Figure 39: Website Overview*

# *Reflection*

Ying Zhen Reflection:

In this project, I help to do reviews data cleaning and create a word cloud on negative reviews. I feel grateful for my group members. They helped me whenever I faced any difficulties with the tasks. I think I still have a lot of areas to improve on in my programming. When I face an issue, I will try to solve it on my own. I think that trying is good, but I wasted too much time trying. I should raise the issue when I am still not able to solve it after 1 hour, so that it will not affect the group’s progress. Through this project, I learned that communication is key in a group project. In future group projects, I will ask immediately if I face any issues.

Elizabeth Reflection:

Though I learned Python before in Poly, many of such libraries used in this project weren't covered then. At the start of the Python project, I was mainly worried about doing a project using such libraries, within such a short timeframe. However, after the past 8 weeks, I not only picked up new technical skills such as data crawling and scraping, Pandas, Seaborn, API calls etc. but also forged new friendships with my groupmates. I overcame obstacles by searching online, and reaching out to the team mates when in doubt, we often helped each other and compromised, which is what I believe made our project a success. Furthermore, it was also interesting to see how some of our insights drawn from the data analytics done could be very useful and valuable to actual hotel managers, in specific areas to look out for when overseeing their hotels.

Yi Xing:

The project has been very exciting as I was able to work with my wonderful group mates and learn much about them. This project has also provided me with insights of what an industry project is like and how it is like working in a team for a tech project. I was able to upgrade myself as a programmer through the experience and also become a better team member

Darren:

I was in an engineering course in poly and wasn’t exposed to Python back then. This proved to be a challenge in university as I had to take a lot of initiative in my self learning in order to do the work that was needed from me for the project. There was steep learning curve at the start but I am glad that through countless hours of practice, I was able to add much value to the project and I feel that I was a much better programmer compared to when I first started university

Declan:

The project was an eye-opening one getting to not only make new friends in my group, but also apply technical skills that I learned in Poly to elevate our project. I also learned the importance of teamwork, helping one another, and the importance of constant group meetings and updating one another. It was enriching to see how our project pieced together with everyone’s contribution.