THE CURIOSITY CUP 2023
A Global SAS® Student Competition

Team Name: Data Voyagers

Unveiling Traveler Preferences and Experiences: An Analysis of Hotel Reviews from TripAdvisor.com using SAS

ABSTRACT

This paper describes a systematic approach for analyzing a dataset of 20,490 hotel reviews collected from TripAdvisor.com using SAS. The goal of this study is to answer several business questions related to customer satisfaction in the travel industry, including identifying the top 10 most frequently used words in hotel reviews, the most common words across all reviews, and the most commonly referenced entities in the reviews. The results of the analysis are significant in that they provide valuable insights into the preferences and experiences of travelers, which can be used to inform decisions related to hotels and rooms. Additionally, this study highlights the usefulness of opinion mining and term frequency analysis as powerful techniques for extracting insights from online text, and provides a valuable resource for researchers interested in analyzing hotel reviews.

INTRODUCTION

One effective strategy for making informed decisions about accommodations is to consult reviews from individuals who have previously stayed at the hotel. Through this approach, individuals can obtain valuable insights into various aspects of the hotel, such as its atmosphere, layout, and other relevant factors. Additionally, reviews can help individuals identify specific services that may be of interest to them during their stay. Opinion mining, a form of aspect-based sentiment analysis, is a powerful technique that can be used to extract insights from online text, where the text is represented as a bag of words and each word is assigned a score based on its relationship to the topic, as well as the contextual relationships between words. This process is known as word sense disambiguation. Recent advancements in technology have facilitated the use of term frequency to classify user feedback, making it an increasingly useful tool for identifying and analyzing relevant information from online reviews.

The purpose of this paper is to conduct an analysis of a hotel reviews dataset that were collected from the website TripAdvisor.com. The dataset consists of 20,490 reviews, with each review containing an average of 12.03 sentences and 104.36 words. The vocabulary size of the dataset is 25,448, indicating that the reviews contain a wide variety of unique words. The reviews are likely to be diverse in terms of the hotels that are being reviewed, as well as the experiences and opinions of the reviewers. The reviews could potentially be used for tasks such as sentiment analysis, topic modeling, or information extraction, depending on the specific research question being addressed. It is important to note that the dataset contains inconsistencies data, because it is based on user-generated content. Additionally, the dataset contain errors or inaccuracies due to the use of automatic text extraction methods or the presence of spam or fake reviews. This dataset provides a valuable resource for researchers interested in analyzing hotel reviews and understanding the opinions and experiences of travelers on TripAdvisor.com.

The present study proposes a systematic approach for analyzing data from TripAdvisor using SAS, consisting of a three-step process. Firstly, the data is automatically imported and verified for accuracy, followed by an assessment of discrepancies in the data. Subsequently, data preparation is performed, as the output from this stage forms the basis of the analysis phase. During the data analysis stage, an appropriate algorithm is selected to generate summaries and visualizations that align with the hypothesis. Finally, the results are transformed into a suitable format that can be shared with relevant stakeholders, facilitating effective communication and dissemination of findings.

The goal of this study is to answer several business questions related to customer satisfaction in the travel industry. Specifically, we seek to determine the top 10 most frequently used words in hotel reviews, identify the most common words across all reviews, and pinpoint the most commonly referenced entities, such as people, places, and organizations, mentioned in the reviews.

METHODOLOGY

The methodology of this study involves a four-step process for analyzing a dataset of hotel reviews collected from TripAdvisor.com using SAS. The following sections describes these three stages in details.

IMPORTING AND ACCESSING DATASET

The given text describes a code that uses the SAS programming language to import a data file in CSV format located at a specific file path. The imported data is saved in an output file named "comments" while the "replace" option overwrites any existing dataset with the same name. The "guessingrows=max" option automatically determines the data types for input variables based on the maximum number of rows in the data file, reducing the likelihood of errors and saving time. This code provides a reproducible and efficient method of importing large datasets into the SAS environment, making it a valuable tool for data analysis and research.

```
PROC IMPORT DATAFILE="/home/u14644163/Hotel-reviwer/tripadvisor.csv"
    OUT=comments
    DBMS=CSV
    replace;
    guessingrows=max;
RUN;
```

EXPLORATION AND PREPARING DATA

In this stage, the data need to be explored in order to gain a better understanding of its structure, distribution, and relationships between variables. This involves conducting various descriptive statistics and visualizations to summarize and present the data in a meaningful way.

The initial stage of data exploration involves data cleaning to remove any unwanted characters from the comments data and converting them to lowercase. The code provided processes a dataset called "comments_cleaned" using SAS. The dataset is created by reading in a dataset called "comments". The "set" statement specifies that the contents of the "comments" dataset should be used as input to create the new dataset. The code then proceeds to clean the comments data by performing a series of manipulations on the "review" variable. The "compress" function is used to remove certain characters from the text, including asterisks, pipes, commas, semicolons, exclamation marks, question marks, double quotes, and apostrophes. The resulting text is stored in a new variable called "comment". Next, the "tranwrd" function is used to replace any double spaces in the text with single spaces. This is done to ensure that the text is properly formatted and does not contain any unnecessary white space. Finally, the "lowcase" function is used to convert all

characters in the "comment" variable to lowercase. This step is important to ensure consistency in the data, as uppercase and lowercase characters may be treated as separate entities by some analysis techniques. The code for this stage is presented below:

```
data comments_cleaned;
  set comments;
  comment = compress(review,'*|.,;:!?"''');
  comment = tranwrd(comment,' ',' ');
  comment = lowcase(comment);
run;
```

The second stage of data exploration involves processing the cleaned comments data by extracting relevant words for analysis. To accomplish this, the researcher may choose to exclude common words such as "the," "and," "of," "to," "in," "is," "it," "for," "as," "with," and "was" using a stopword list defined by the variable %let STOPWORDS = the and of to in is it for as with was;. This list is then used in the if statement to exclude any words that match the stopwords from being output. The code defines a new dataset called comments_words using the set statement to specify the input data source as comments_cleaned. The do loop is used to iterate through each word in the comment using the scan function. If the word is not included in the defined stopword list, it is output to the word variable, which is defined with a length of 500 characters. Finally, the resulting output is saved in the comments_words dataset for further analysis. By removing stopwords and extracting relevant words, researchers can gain insights into the most common themes and topics present in the comments data. This process enables researchers to identify patterns and trends in the data that can be used to inform decision-making, such as identifying areas for improvement in a hotel or product based on customer feedback. The code for this stage is presented below:

```
data comments_words;
  set comments_cleaned;
  length word $500;
  do i=1 to countw(comment,' ');
    word = scan(comment,i,' ');
    if word not in ("&STOPWORDS") then output;
  end;
run;
```

The text describes the two stages of data exploration, which involve cleaning and processing a dataset to gain a better understanding of its structure, distribution, and relationships between variables. The first stage involves cleaning the data by removing unwanted characters and converting text to lowercase using SAS. The second stage involves extracting relevant words for analysis by excluding common words using a stopword list and iterating through each word in the comments using the scan function. The resulting output is saved in a new dataset called "comments_words" for further analysis, which can help identify patterns and trends in the data that inform decision-making.

ANALYZING DATA

The data analysis section is a crucial component of any research project, particularly when it comes to investigating the hospitality industry. In this section, we will use various analytical techniques to explore and uncover valuable insights into different aspects of the hotel, including customer feedback, service quality, and overall performance. Through data analysis, we can gain a deeper understanding of the challenges and opportunities faced by the hotel and identify strategies for improving customer satisfaction and driving business growth

What are the top 10 words used most frequently in the hotel reviews, which is the first insight that we want to be extracted from the hotel review dataset?. The first part of the code uses the PROC FREQ procedure to count the frequency of each unique word in the

comments_words dataset. The NOPRINT option is used to suppress the output in the SAS log. The resulting frequency table is saved to a new dataset called words_freq using the OUT= statement. The second part of the code is a data step that reads in the words_freq dataset created in the previous step. The IF statement is used to only keep the words with a count greater than 1, meaning only words that occur more than once are retained. The DROP statement is used to remove the i variable, which was created during the counting process in the previous PROC FREQ step. The resulting dataset is still named words_freq. The third part of the code sorts the words_freq dataset by descending frequency count using the PROC SORT procedure. The sorted dataset is still named words_freq. The fourth and final part of the code creates a new dataset called words_freq_10 that contains only the top 10 most frequent words from the words_freq dataset. This is achieved by using the SET statement with the OBS= option to read in only the first 10 observations from the words_freq dataset, which are the 10 words with the highest frequency counts. The resulting dataset is named words_freq_10. Appendix A1 and Appendix 2 illustrate the frequency values of the top words in the hotel reviews dataset.

The next insight that can be extracted using this code is to identify the most common entities (such as people, places, and organizations) mentioned in the hotel reviews. The following code is used for text analysis on the "comments_cleaned" dataset. The "proc TGPARSE" step uses the Text Analytics engine to identify and extract key information from the comments text. The options "entities=yes" and "tagging=yes" turn on the recognition of named entities and parts of speech, respectively, while "stemming=yes" enables word stemming for analyzing word forms. The "key=Key4" option specifies that the output should be written to a dataset called "Key4". The "var comment" statement indicates that the "comment" variable in the "comments_cleaned" dataset should be analyzed. After running the "proc TGPARSE" step, the code sorts the "Key4" dataset by frequency using "proc sort" with the "by descending freq" option. This sorts the dataset in descending order based on the frequency of the key phrases extracted in the previous step. Overall, this code is used to identify and extract key phrases and concepts from the hotel review comments, which can provide valuable insights into common themes and issues mentioned in the reviews.

```
proc TGPARSE data=comments_cleaned
    /* turn the entity finder on */
    entities=yes stemming=yes
    tagging=yes key=Key4 out=Out4;
    var comment;
run;
proc sort data=Key4;
  by descending freq;
run;
```

To find the most common words in the documents, we used the obtained results from Key4 and Out4. The given code performs text mining on the "comments_cleaned" dataset to extract important terms and phrases that can be used for further analysis. The proc TGPARSE is used for entity recognition and to extract key phrases from the comments. The output is stored in the dataset named "Out4". Next, proc freq is used to compute the frequency of each term and store the result in the dataset named "Frequencies". The frequency of each term is weighted by its count. Then, proc sort is used to sort the terms by their frequency in descending order. In the next step, the dataset "TopWords" is created to store the top 10 most frequent terms. The set statement is used to retrieve the top 10 terms from the "Frequencies" dataset, and the keep statement is used to keep only the _TERMNUM_ and count variables. Then, proc sort is used to sort the "Key4" dataset in descending order by frequency. Next, a left join operation is performed between "Topwords" and "Key4" datasets using the _TERMNUM_ variable to link the two datasets. The result of the join is stored in the dataset "final_data". Finally, proc sort is used to sort the "final_data" dataset by the count variable in ascending order, and nodupkey option is used to remove duplicate observations based on the count variable. The resulting dataset,

"sorted_final_data" is then printed using the proc print statement to display the top 10 most frequent terms along with their counts.The code od doing this is presented below:
proc TGPARSE data=comments_cleaned.

```
proc freq data=Out4;
  tables _TERMNUM_ / missing out=Frequencies;
  weight  _COUNT_;
run;
proc sort data=Frequencies;
  by descending count;
run;
data TopWords;
  set Frequencies (obs=10);
  keep _TERMNUM_ count;
run;
proc sort data=Key4;
  by descending freq;
run;
proc sql;
    create table final_data as
    select Topwords.*, Key4.key, Key4.Term
    from Topwords
    left join Key4
    on Topwords._TERMNUM_ = Key4.key;
quit;
proc sort data=final_data out=sorted_final_data nodupkey;
  by count ;
run;
proc print data=sorted_final_data;
  var Term count;
run;
```

The fact that "hotel" and "room" are the most common words in the reviews (see Appendix 3) could mean that these two aspects are the most important and frequently mentioned by customers in their reviews of the hotel. This suggests that customers are likely to focus on the quality of the hotel rooms and their overall experience with the hotel during their stay. It may also indicate that these are the areas where the hotel could focus on improving their services and amenities to better meet customer needs and expectations.

CONCLUSION

This paper outlines a systematic approach for analyzing a dataset of 20,490 hotel reviews using SAS, with the aim of answering business questions related to customer satisfaction in the travel industry. The study identified the top 10 most frequently used words in hotel reviews, the most common words across all reviews, and the most commonly referenced entities in the reviews. The results provide valuable insights into the preferences and experiences of travelers, which can be used to inform decisions related to hotels and rooms. The study also highlights the usefulness of opinion mining and term frequency analysis as powerful techniques for extracting insights from online text. This dataset and methodology provide a valuable resource for researchers interested in analyzing hotel reviews. Further work could include sentiment analysis, topic modeling, and information extraction to deepen our understanding of the data.

REFERENCES

RefeLutz, C. (2017). SAS Programming by Example: A Practical Guide for Learners. New York, NY: Springer.

Kowalski, T. (2017). SAS for Data Analysis: Intermediate Statistical Methods. Boston, MA: Springer.
Barr, K. (2016). Mastering SAS: Advanced Techniques Made Easy. San Francisco, CA: SAS Institute Inc.
Curry, R. (2015). SAS Programming: A Gentle Introduction. London, UK: Springer.
Ott, R. (2014). SAS Macro Programming Made Easy. New York, NY: SAS Institute Inc.rence
SAS Institute Inc. Base SAS® Procedures Guide. Cary, NC: SAS Institute Inc.
Curtis, J. (2007). SAS® For Dummies®. Hoboken, NJ: John Wiley & Sons, Inc.

ACKNOWLEDGMENTS

We would like to acknowledge Dr. Rasha Shakir AbdulWahhab for her supervision and support throughout our work

CONTACT INFORMATION

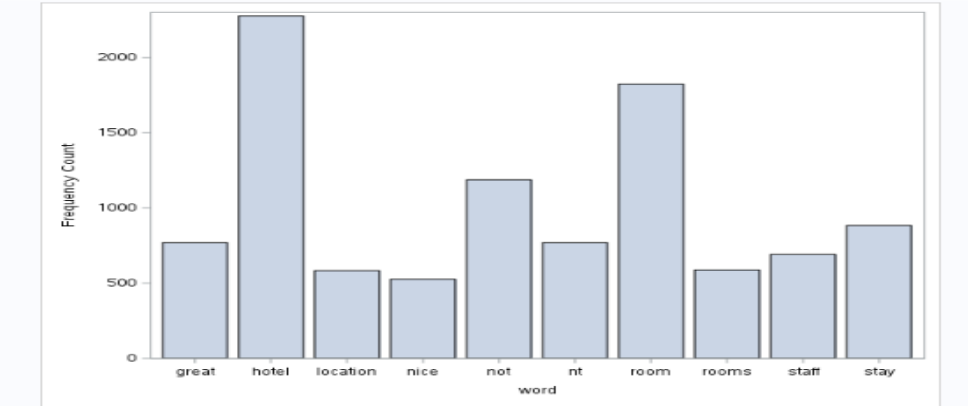
Your comments and questions are valued and encouraged. Contact the author at:

APPENDIX A

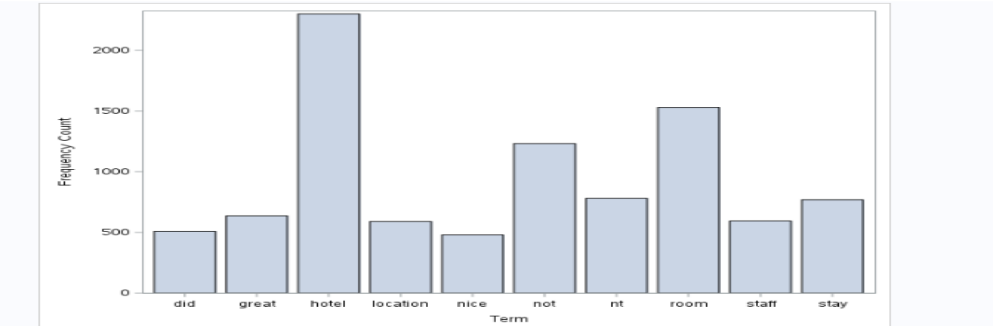
Top 10 Words in Visitor Comments

Obs	word	COUNT	PERCENT
1	hotel	2277	2.72940
2	room	1824	2.18639
3	not	1187	1.42283
4	stay	883	1.05844
5	great	769	0.92179
6	nt	769	0.92179
7	staff	691	0.82829
8	rooms	588	0.70482
9	location	583	0.69883
10	nice	526	0.63051

A1. The top Words in Visitor Comments.



A2. The Frequency count of the top Words in Visitor Comments.



A3. The most comment words among document.