Leeds University Business School

UNIVERSITY OF LEEDS

Assessed Coursework Coversheet

For use with *individual* assessed work

Student ID Number:	2	0	1	8	0	0	7	5	5
Module Code:	LUBS5309M								
Module Title:	Forecasting and Advanced Business Analytics								
Module Leader:	Panagiotis Stamolampros								
Declared Word Count:	2985								

Please Note:

Your declared word count must be accurate, and should not mislead. Making a fraudulent statement concerning the work submitted for assessment could be considered academic malpractice and investigated as such. If the amount of work submitted is higher than that specified by the word limit or that declared on your word count, this may be reflected in the mark awarded and noted through individual feedback given to you.

It is not acceptable to present matters of substance, which should be included in the main body of the text, in the appendices ("appendix abuse"). It is not acceptable to attempt to hide words in graphs and diagrams; only text which is strictly necessary should be included in graphs and diagrams.

By submitting an assignment you confirm you have read and understood the University of Leeds **Declaration of Academic Integrity** (http://www.leeds.ac.uk/secretariat/documents/academic_integrity.pdf).

Time Series Modelling of US Personal Consumption Expenditure Index

Introduction

Personal consumption expenditure (PCE) is a measure of economic activity, reflecting the total value of consumer spending on goods and services. Accurately predicting PCE is important for policymakers, businesses, and investors to make informed decisions and anticipate economic trends. In this analysis, predictive ability of three forecasting models is compared, using United States PCE data.

- A simple forecasting method
- An exponential smoothing model
- An ARIMA model

The objective is to identify the best-performing model among the three for which, a comprehensive analysis will be conducted, consisting of multiple phases.
First step is data analysis and exploration, including data decomposition, data pre-processing, handling missing data, and dataset splitting.
Next, performance of each model is evaluated and best one is chosen based on multiple accuracy measures and the residual analysis.
Lastly, all models' performance is tested using a one-step ahead rolling forecasting approach, without re-estimation of the model parameters. This will assess the robustness and efficiency of the chosen forecast model.

Data description

This time series data provided contains PCE data spanning from January 1959 to November 2023, with a monthly frequency.

	DATE	PCE
1	01/01/1959	306.1
2	01/02/1959	309.6
3	01/03/1959	312.7
4	01/04/1959	312.2
5	01/05/1959	316.1
6	01/06/1959	318.2
7	01/07/1959	317.8
8	01/08/1959	320.2

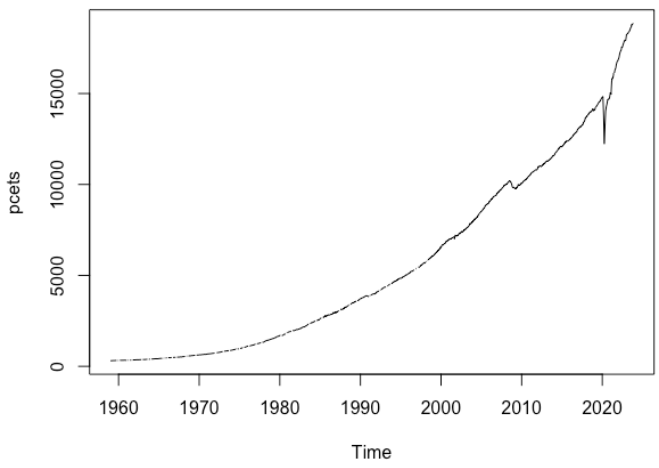


Figure 1. Visual representation of time series

It can be observed that there is a clear increasing trend over time with multiple missing observations up till year 2000, with no visible outliers.

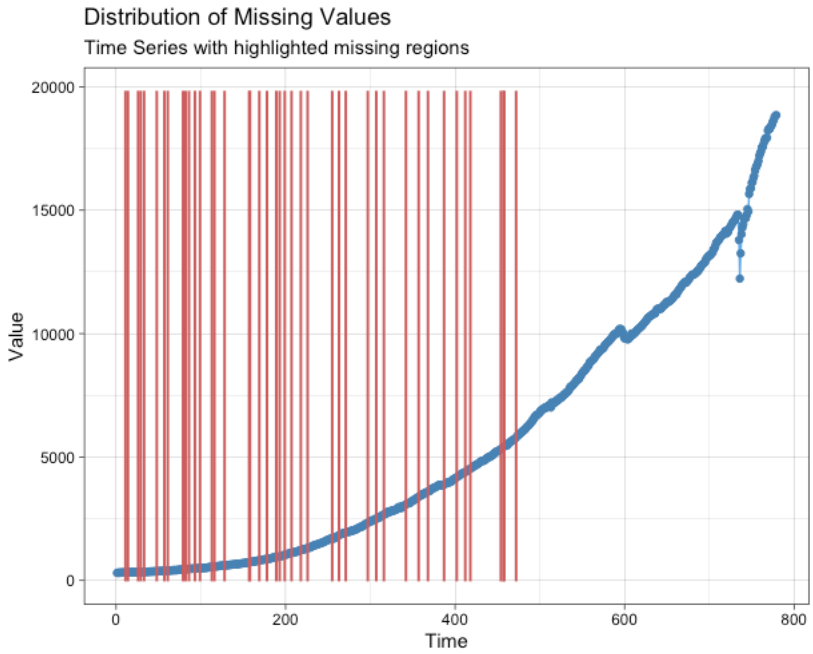


Figure 2. Missing values plot distribution

Data Analysis

Additive decomposition is used since the time series is seasonally adjusted.

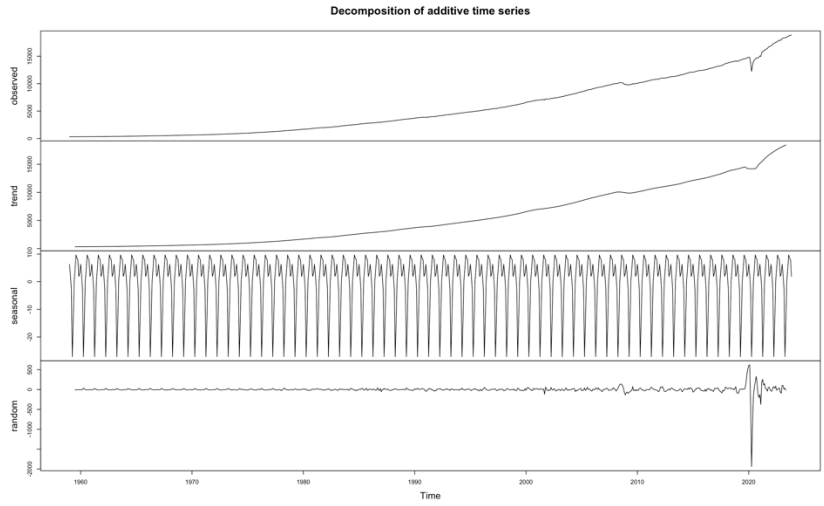


Figure 3. Additive decomposition of time series

Figure 3 illustrates that there is a gradual increasing trend in the time series. Seasonal component is present. The random component resembles white noise prior to 2000, after which variance increases with time. This pattern suggest that residual time series is non-stationary.

To impute the missing observations, linear interpolation method is most suitable. It is a straightforward and widely used method for estimating missing values in data. Essentially, it

involves connecting two known points with a straight line and assuming that the points along this line represent the values that are not observed.
Since the data values in time series follow a clear linear trend without much fluctuation, so the missing values can be estimated using the adjacent data points. It will help preserve the underlying trend and maintain the overall smoothness of the data.

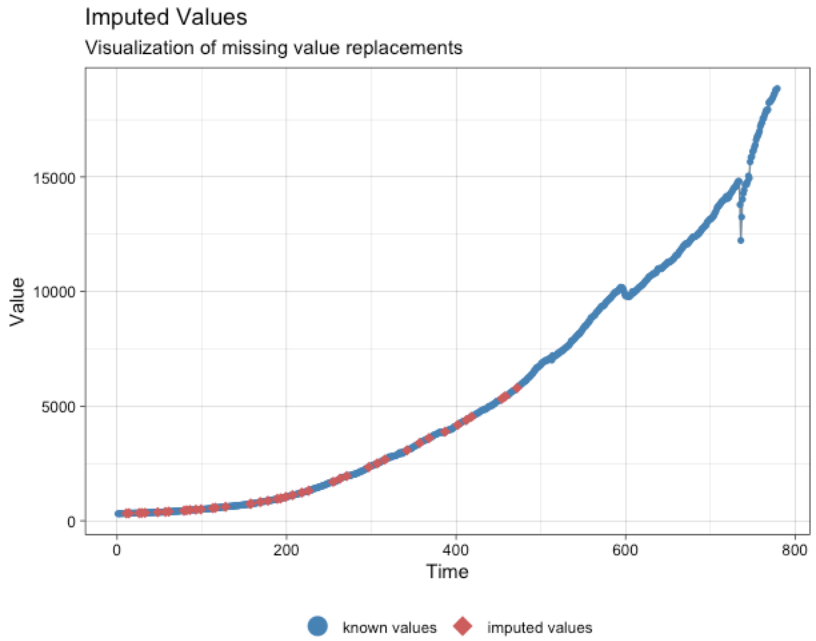


Figure 4. Visual representation of time series after imputation

A seasonal plot is constructed for the underlying seasonal pattern to be seen more clearly. It is especially useful in identifying years in which the pattern changes.

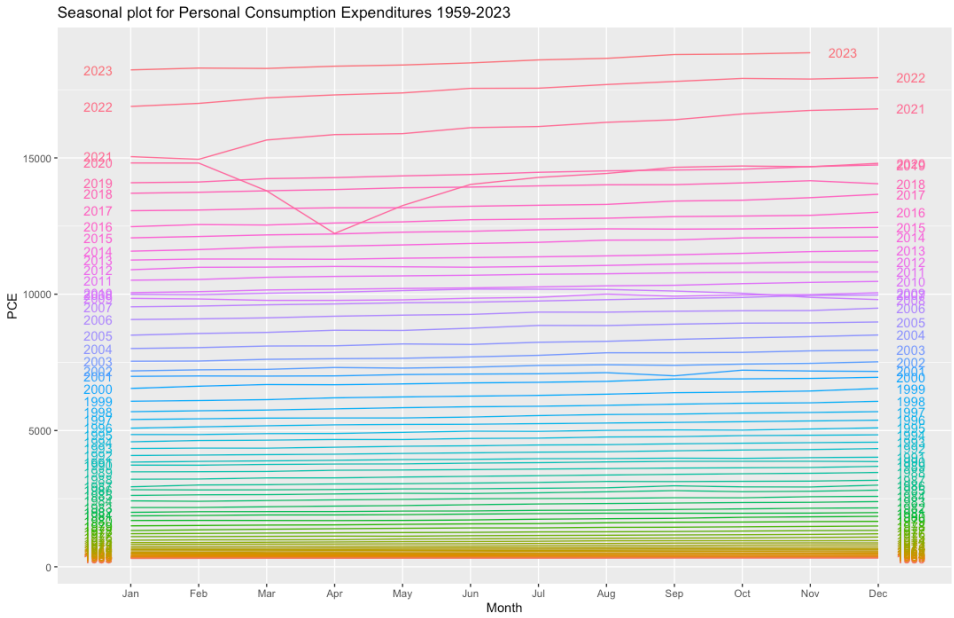
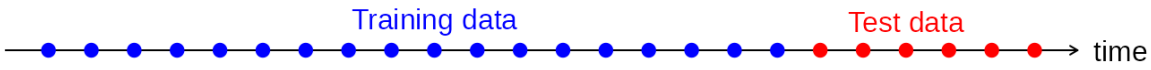


Figure 5. Seasonal distribution of time series

From above plot, it is evident that there is a significant dip in PCE index in April 2020. This was probably because WHO declared COVID-19 pandemic in March 2020. Widespread lockdowns, social distancing measures, and economic uncertainty led to declined consumer spending patterns, impact on specific industries such as retail, hospitality, and transportation.

Forecasting methodology

To compare the forecast accuracy of the models, the time series data is separated into training and testing data in the ratio of 80:20. The "in-sample" or training data is used to estimate the parameters of the forecast model and the "out-of-sample" or testing data is used to predict its accuracy.



The 3 models used for time series modelling are as follows:

1. The "drift method" also called "random walk with drift" is utilised. It incorporates a trend by allowing the forecast to change gradually over time. Specifically, it assumes that future values will be equal to the most recent observed value plus a constant "drift" or trend factor.
2. "Holt's linear trend method" also known as "double exponential smoothing method" is utilised. It is an extension of simple exponential smoothing that incorporates a linear trend component in addition to the level component. In our case, both are suitable because our time series exhibits an increasing linear trend.
3. A non-seasonal ARIMA model is utilised, which captures the autocorrelation in the time series by combining autoregression (AR), differencing (I), and moving average (MA) components.

While exponential smoothing models are based on a description of the trend and seasonality in the data, ARIMA models aim to describe the autocorrelations in the data. The models discussed above are trained on the training data, then future values are predicted using the trained model on the testing dataset. Firstly, Fixed origin fixed holdout sample approach is utilised, and the best model is chosen based on 3 criteria:

- Model having smallest errors metrics for test set.
- Visually which model which best captures the trend of time series.
- Model having smallest forecast residuals.

Lastly, the robustness of the model is assessed by cross validation. A rolling origin forecast approach, without re-estimation of the model parameters, is utilised to compare the models' forecasting accuracy and check whether the outcome is different.

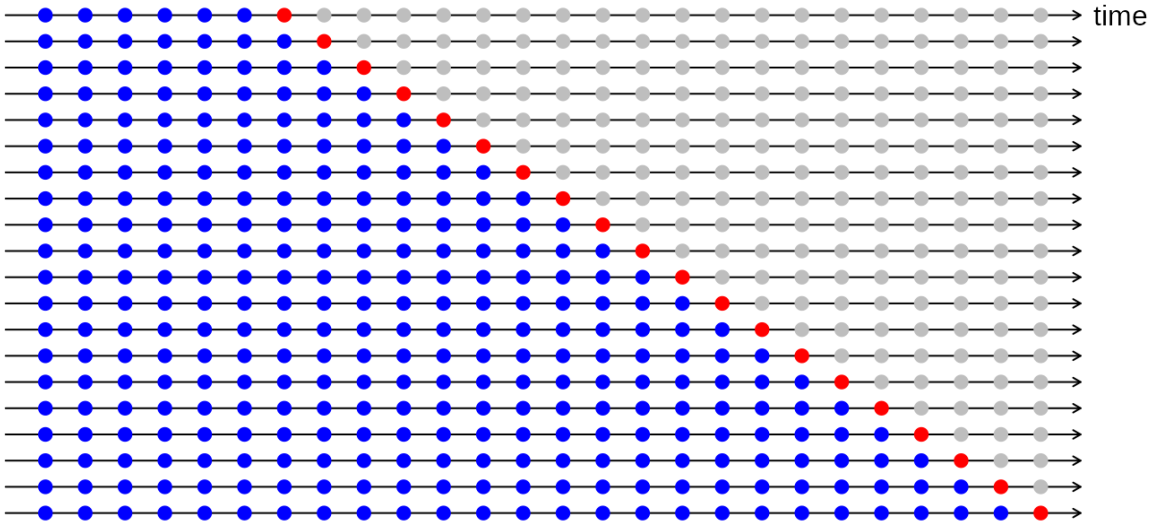


Figure 6. One step ahead rolling origin forecast with variable test sample size

Results and Discussion

On observing from the graph below, it is obvious that the holt linear method is best for these data as it closely tracks the pattern of the actual data values.

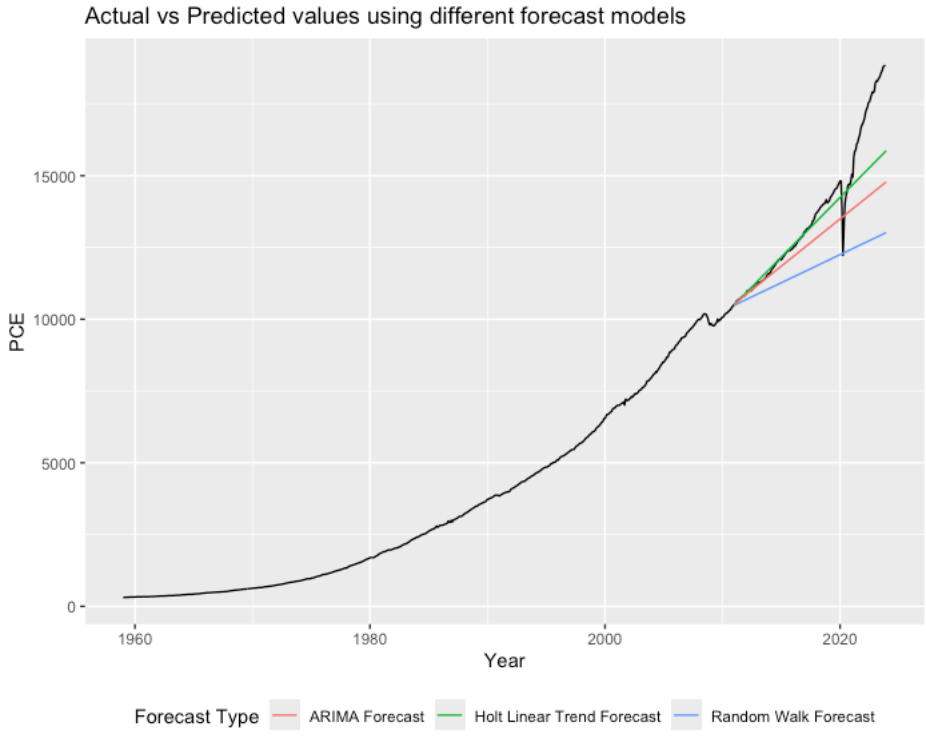


Figure 7. Comparison of actual and forecasted values using different models

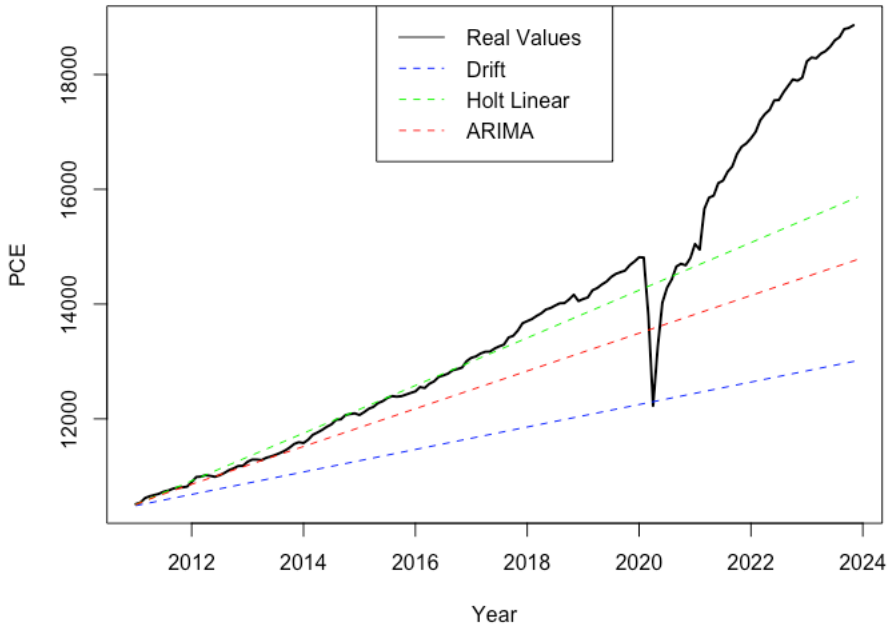


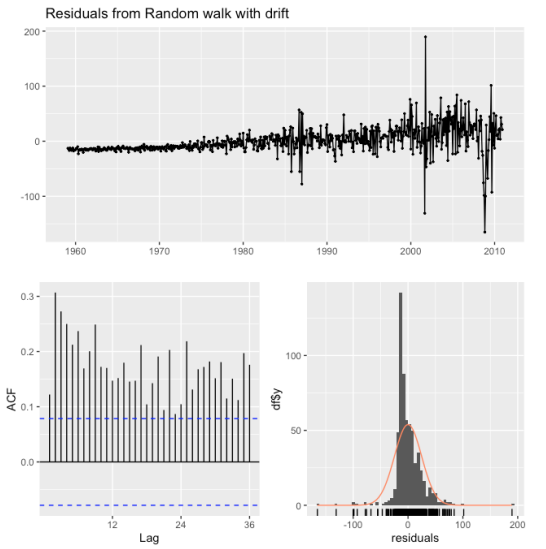
Figure 8. A close-up view of Figure 6

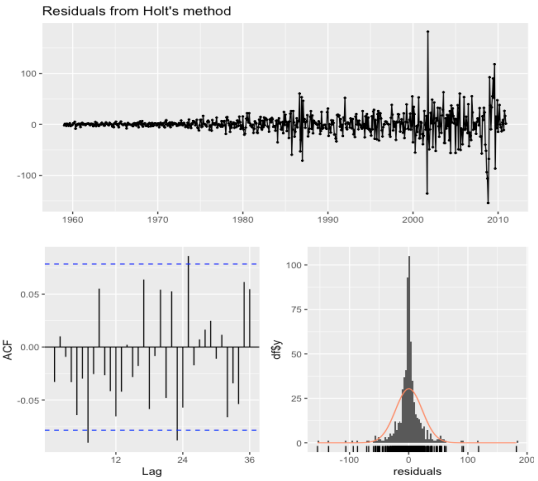
While visual inspection can provide valuable insights, it is often necessary to support it with quantitative measures of forecast accuracy, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or Mean Absolute Percentage Error (MAPE). These metrics provide numerical assessments of forecast accuracy and can help objectively evaluate the performance of different forecasting models.
Forecast accuracy of models is measured by summarising the forecast errors in different ways. A forecast "error" is the difference between an observed value and its forecast. Here "error" does not mean a mistake, it means the unpredictable part of an observation.

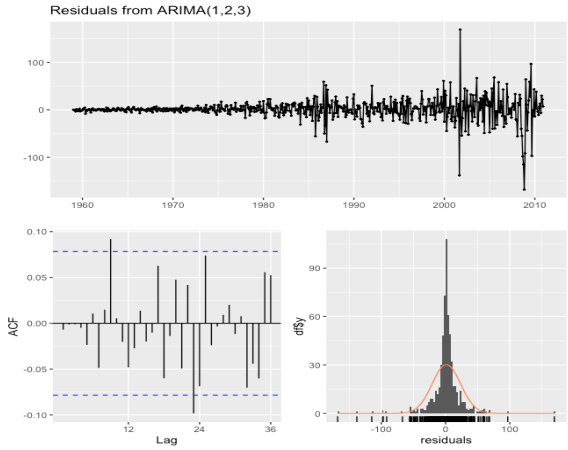
	RMSE	MAE	MAPE	MASE
Drift method	2534.5738	1915.4286	12.497530	9.429404
Holt Linear method	830.1576	536.2572	3.502422	2.639924
Arima method	1589.2877	1038.6794	6.499549	5.113283

Figure 9. Error metrics for all models for fixed origin approach

Sometimes, different accuracy measures will lead to different results as to which forecast method is best. However, on observing fig. 9, it can be concluded that holt linear method has the smallest error across all categories. Note that forecast errors are different from residuals in two ways. Firstly, residuals are computed using the training dataset, whereas forecast errors are determined using the test dataset. Secondly, residuals are derived from one-step forecasts, whereas forecast errors can encompass multi-step forecasts.


Residuals from Random walk with drift
200
100
0
-100
1960
1970
1980
1990
2000
2010
0.3
0.2
0.1
0.0
ACF
12
24
36
Lag
100
50
0
dfSy
-100
0
100
200
residuals


Residuals from Holt's method
100
0
-100
1960
1970
1980
1990
2000
2010
0.05
0.00
-0.05
ACF
12
24
36
Lag
100
75
50
25
0
dfSy
-100
0
100
200
residuals


Residuals from ARIMA(1,2,3)
100
0
-100
1960
1970
1980
1990
2000
2010
0.10
0.05
0.00
-0.05
-0.10
ACF
12
24
36
Lag
90
60
30
0
dfSy
-100
0
100
residuals

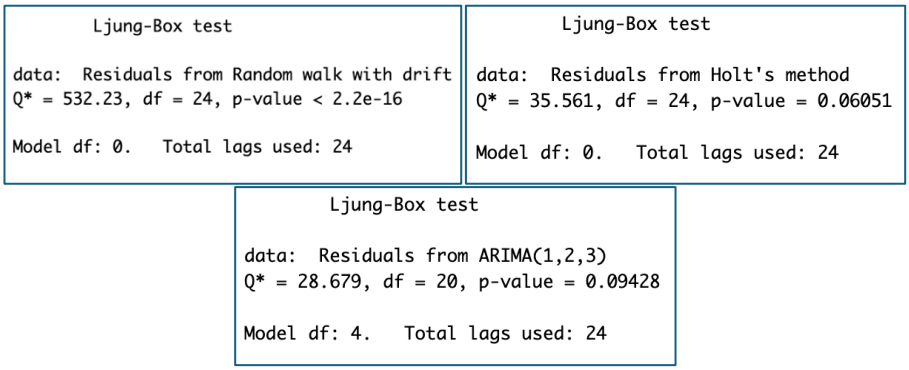


Figure 10. Residual analysis of all models

The null hypothesis for the Ljung-Box test is that there is no autocorrelation in the residuals. In this case, p-value is > 5% significant level, which means that the null hypothesis is favoured and there is a 95% chance that residuals are independent and are mostly white noise.

On observing fig. 10, drift model can be rejected because the residuals are not white noise and all spikes in ACF plot go beyond confidence interval. For drift and arima models, p-value > 5%, and the residuals for both have a normal distribution with majority having zero mean. Despite arima exhibiting a greater p-value which indicates stronger likelihood of residuals being white noise, the difference in errors between arima and holt is significant, with arima having almost twice the errors compared to holt. Hence, holt linear is chosen as the best mode, prioritising accuracy.

The prediction for October 2024 PCE index using holt linear model is as follows:

Oct 2024	17026.20	9631.024	24421.38	5716.254	28336.15

Where 17026 is the point forecast, 9631 and 24421 are the low and high values at 85% confidence interval, and 5716 & 28336 are low and high value at 95% confidence interval.

Next, time series cross-validation is implemented across all models and errors are compared.

	RMSE	MAE	MAPE	MASE
Drift method	2534.574	1915.4286	12.497530	20.649353
Holt Linear method	1085.053	605.4722	3.682584	6.527317
Arima method	1589.288	1038.6794	6.499549	11.197524

Figure 11. Error metrics for all models for rolling origin approach

From fig. 11, it can be observed that for Arima and Drift models, there wasn't significant difference in RMSE, MAE and MAPE. However, MASE has increased across all models. This is because rolling origin has higher error rate due to the evolving nature of test data with time. Although the Holt model has slight increase in all metrics, they remain lower than those of the other models. Hence, Holt linear remains the preferred model among the three.

Limitations:

The aforementioned ARIMA model may not be the best because the *auto.arima()* function takes only a specific set of models into consideration, and not the entire range of potential

combinations, to identify the most optimal model. Consequently, there might be alternative arima models that haven't been evaluated yet and could potentially yield better results.
The forecasting accuracy of models could be increased if train them on shorter time series because it would capture most recent trends in data.

Topic Modelling analysis of hotel reviews

Introduction

Customer reviews play an important role in shaping perceptions and decisions in the hospitality industry. In today's digital era, online platforms provide a large repository of such reviews, offering valuable insights into customer satisfaction and dissatisfaction. In this part, a dataset of hotel reviews will be analysed.
The primary objective is to apply topic modelling to find out the underlying patterns and factors discussed in both positive and negative reviews. The aim is to gain a comprehensive understanding of the key factors influencing customer satisfaction and dissatisfaction in the hotel industry.
To achieve this, first, the reviews will be categorised as positive or negative by sentiment analysis, followed by a detailed description of the methodology used. This includes pre-processing steps such as data cleaning and stop words elimination, as well as the selection of the number of topics using appropriate criteria.

Furthermore, the top three factors affecting both satisfaction and dissatisfaction of customers will be highlighted. This will provide informative insights to hotel management which will help them to enhance guest experience & overall service quality.

Data description

A dataset containing 10,000 hotel reviews has been provided It contains both positive and negative feedback expressed in various languages including English, French, and others. There are 25 distinct languages in the entire dataset. Each review is accompanied by a rating on a Likert scale ranging from 1 to 5, with 1 indicating the lowest level of satisfaction.

4	I last stayed at this hotel some time ago, and am plea…
3	Booking in was a nightmare. I had a reservation but t…
4	questo è il 5 anno di fila che vado in questo hotel in q…
4	The best hotel! Very clean, good profetional staff
2	距離Euston車站或是Euston地鐵站都很便利，為台服務人…

Data Preprocessing

Firstly, a sample of 2000 reviews is randomly chosen from the dataset to minimize selection bias and ensure creation of a representative sample. Secondly, all non-English reviews are translated into English, using G-sheets, to prevent the loss of information. Then, the following pre-processing steps are applied:

- Reviews are converted to lowercase to ensure consistency in the dataset.
- Punctuation marks are replaced with spaces to ensure that a review does not get divided into sentences while running sentiment analysis later.
- A space is inserted between numeric characters followed by alphabetic characters to ensure that the numbers can be easily removed later without creating unnecessary custom stop words and avoid complexity.

Methodology

a) **Sentiment analysis**

First, the pre-processed reviews are divided into positive and negative category based on the *sentiment score*. Sentiments of the review are quantified by assigning sentiment scores to individual words and aggregating them to compute the overall sentiment of a review.

A lexicon-based approach for sentiment analysis is utilised, where each word in the text is assigned a polarity score based on its sentiment orientation (positive or negative). These polarity scores are then aggregated to compute an overall sentiment score for the entire text. Reviews having sentiment score > 0 are classified as positive otherwise negative.

b) **Corpus creation and cleaning**

The reviews (both positive and negative) are first converted to corpus, which is nothing but a collection of text documents. Then the corpus is converted to lowercase to avoid duplicate results. After that, all the numeric characters are removed.

The documents in the corpus are then reduced to their base or dictionary form, known as *lemma* in a process called lemmatisation. It is useful for standardising words and typically results in a more meaningful base word because it considers the context of the word within the sentence and the grammatical structure of the language e.g. "*running*" to "*run*", "*better*" to "*good*" etc.
Next, *stop words* are filtered and removed. Stop words are common words that are often filtered out during text analysis because they are insignificant e.g. articles, prepositions, conjunctions etc. It is important to remove them because they have no semantic value, and their removal can enhance the interpretability of the analysis results and provide clearer insights.

c) Generating Document Term Matrix (DTM)

The corpus is then used to create the DTM. It is a matrix that represents the frequency of words that occur in a collection of documents. Each row of the matrix corresponds to a document, and each column corresponds to a term. In our case, each entry of the matrix represents the frequency of occurrence of each term in each document.
For example, consider a set of documents:
Document 1: "The quick brown fox"
Document 2: "Jumped over the lazy dog"
So, DTM will look like:

	the	**quick**	**brown**	**fox**	**jumped**	**over**	**lazy**	**dog**
Doc 1	1	1	1	1	0	0	0	0
Doc 2	1	0	0	0	1	1	1	1

d) Topic modelling

Topic modelling is a statistical modelling approach that identifies groups of similar words within a corpus of text. This method uses semantic structures in text to understand unstructured data without training data. It analyses documents to identify common themes and provide a suitable cluster.
In our case, Latent Dirichlet Allocation is used. This topic modelling method reveals the hidden structure in a set of observations by looking at the relationships between words in a document and grouping them into topics.
It assumes documents as a mixture of topics and topics as a mixture of words e.g. Document A may have "10% Topic 1" and "90% Topic 2", while Document B has "40% Topic 1" and "60% Topic 2".
To determine the number of topics, 3 metrics are used: Griffiths2004, CaoJuan2009, Arun2010.

Results and Discussion

Upon visualising the density plot of sentiment scores of the reviews, it's evident that most reviews express positive feedback rather than negative ones. On average, the sentiment score sits at 0.627083, ranging from a low of -0.983464 to a high of 3.189068. It is possible

that the graph is left skewed because some negative reviews might include positive words in a sarcastic context, resulting in an overall positive sentiment score of the sentence.

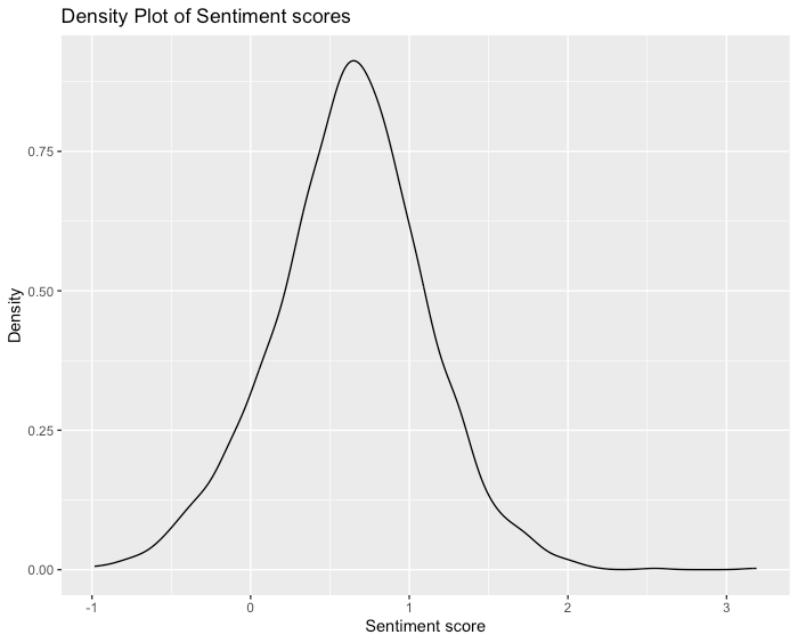
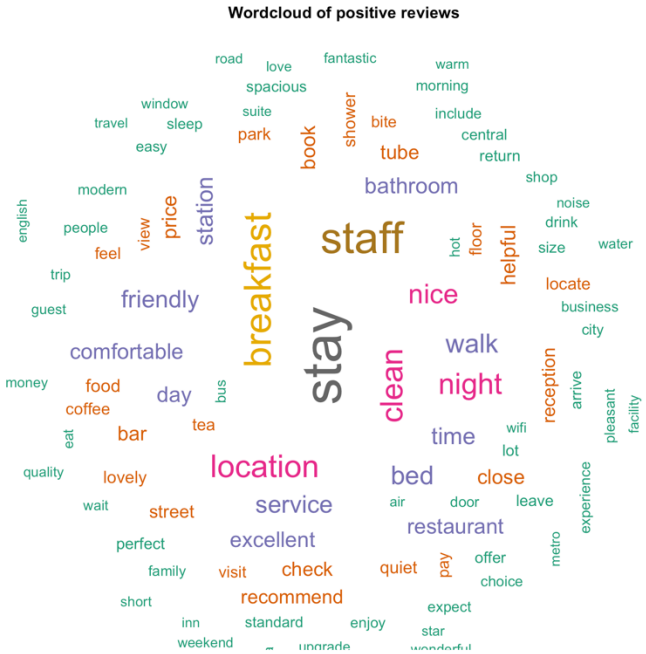


Figure 12. Density plot distribution of sentiment scores

let's visualise the most frequently used terms in both category of reviews in the form of a wordcloud.



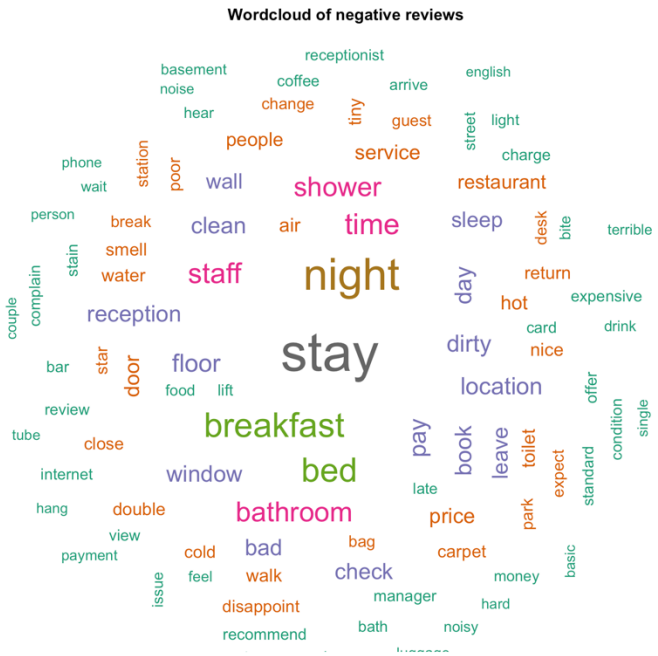
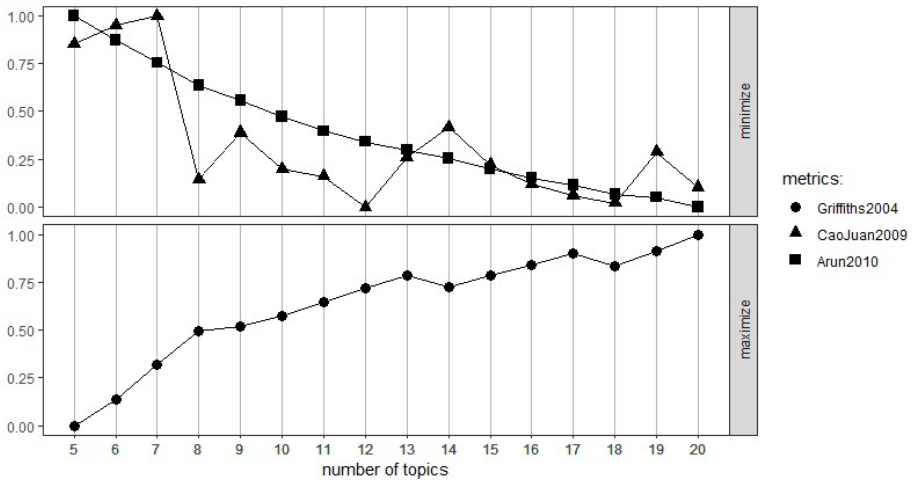


Figure 13. Word cloud for positive and negative reviews

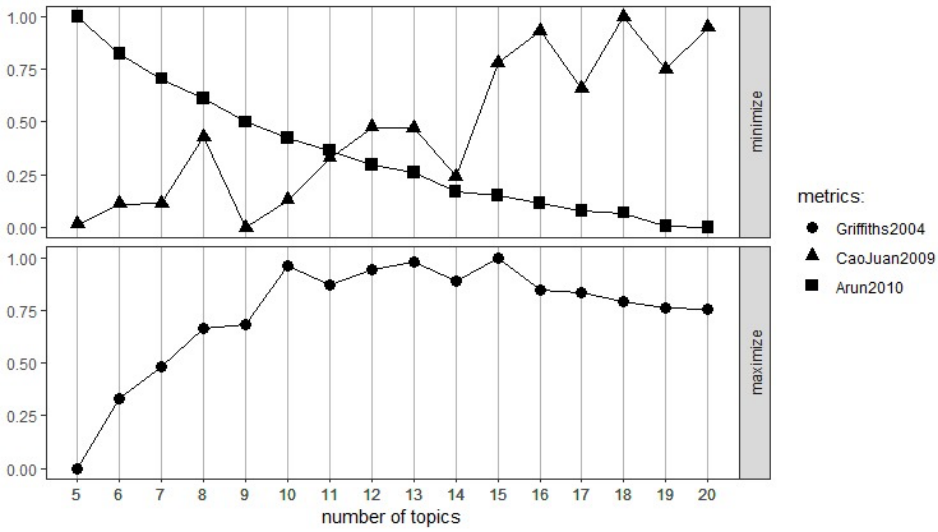
As we can see, there are some common words appearing in both plots, such as *'stay'*, *'breakfast'*, *'bed'* etc. Both plots contain insignificant words as well like *'bus'*, *'metro'*, *'air'* etc which might make sense after topic modelling is done.
Both plots depict their respective scenarios: the negative word cloud reveals a negative pattern, while the positive one reveals a positive pattern. However, they differ in magnitude, as the second plot contains fewer negative words compared to the abundance of positive words in the first plot.

To determine an optimal number of topics, metrics are evaluated across a range of topic numbers, starting from 5 and extending up to 20. We want to minimize the Arun2010 and CaoJuan2009 and maximize the Griffiths2004 criteria.

Metric plot for positive DTM:



Metric plot for negative DTM:



From the plots above, it is evident that the most suitable number of topics which captures the underlying pattern of both positive and negative DTMs are 20, and 14.

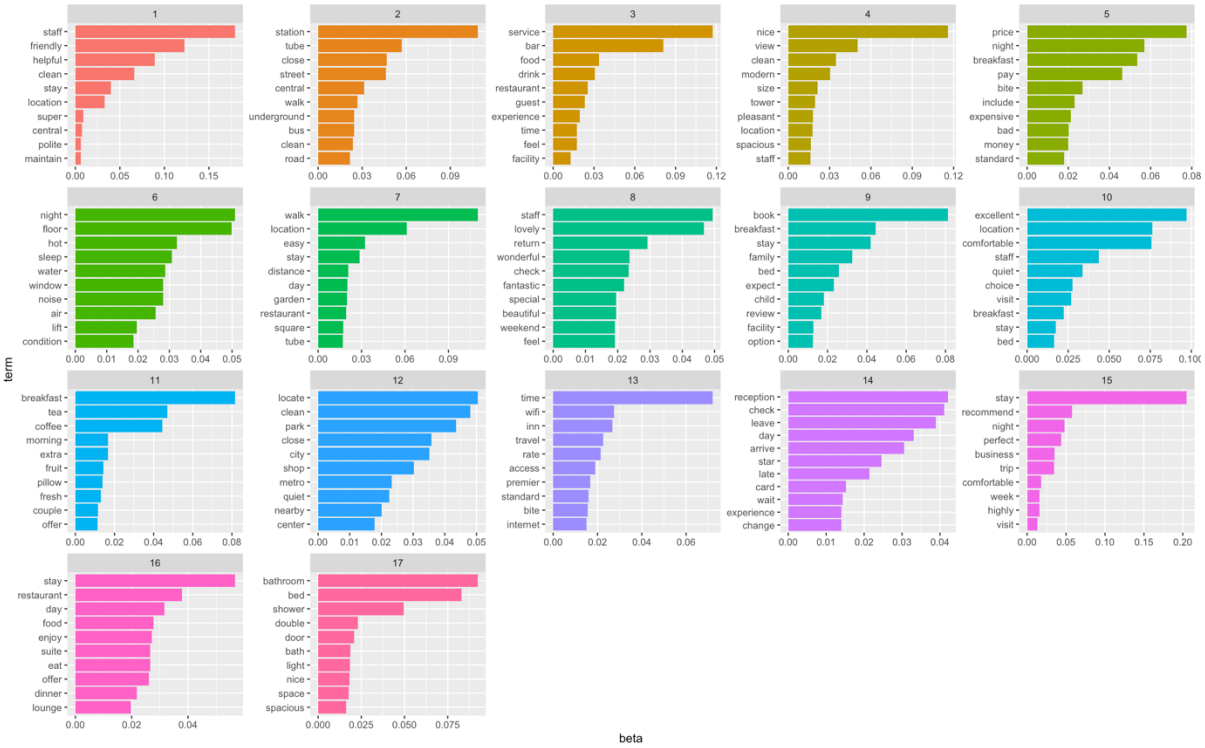


Figure 14. Top 10 words in each positive topic

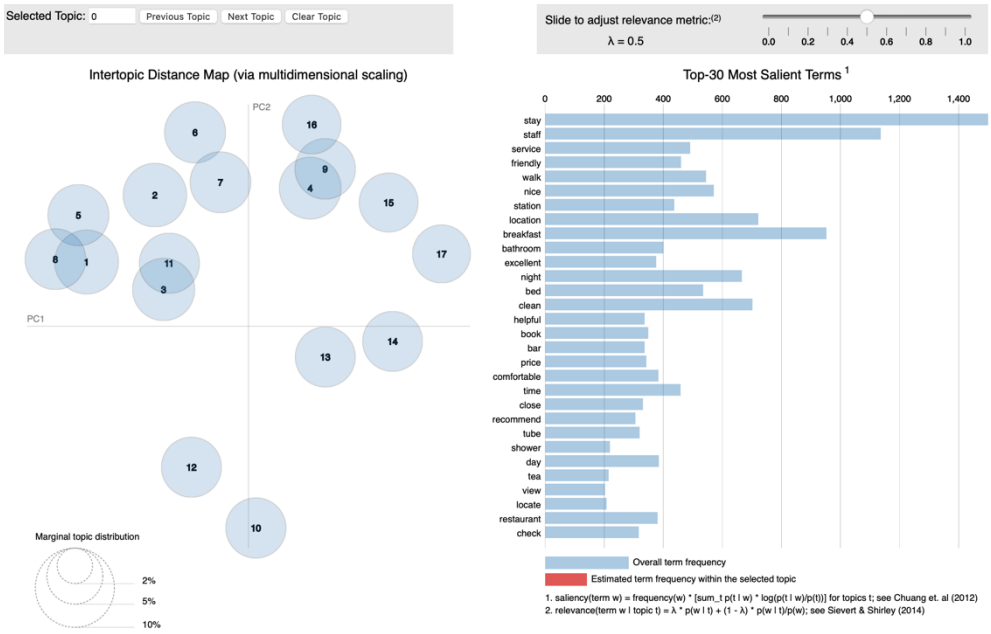


Figure 15. Positive topics visualisation

In figure 15, each bubble represents a topic, and its size represents the significance of that topic in the corpus. In our case, since the size is same for all bubbles, so we will choose the topics based on the distance criteria. The greater the distance between the bubbles, the more different they are e.g. Topics 1 & 8 share similar theme while Topics 17 & 12 would share different theme. Hence, the major 3 topics chosen which affect customer satisfaction are 1, 10, 17.

- **Topic 1: Staff behaviour**
  words like "*staff*" "*friendly*" and "*helpful*", revolves around the quality of staff mangement. Customers value positive interactions with staff members and expect helpful and friendly service throughout their stay. Their satisfaction is influenced by the professionalism, responsiveness, and helpfulness of hotel staff.
- **Topic 10: Hotel location**
  words like "*excellent*", "*central*" and "*location*" suggest discussions about the hotel's location and convenience. Customers prefer hotels that are centrally located, close to public transportation, and offer easy access to nearby attractions.
- **Topic 17: Restaurant quality**
  words like “service”, “restaurant”, “experience” shows that the hotel has exceptional service, offers guests a pleasant dining experience with delicious food and drinks at the bar and restaurant. Every customer feels welcomed and enjoy their time spent utilising the hotel's facilities.

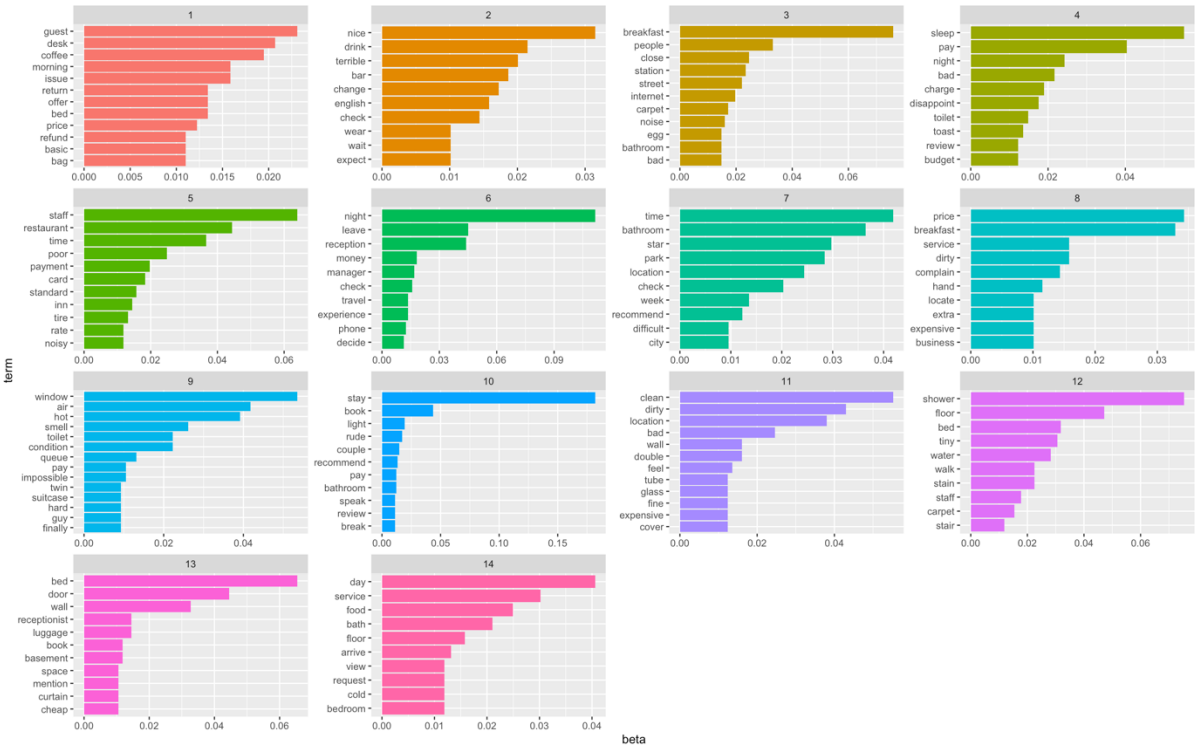


Figure 16. Top 10 words in each negative topic

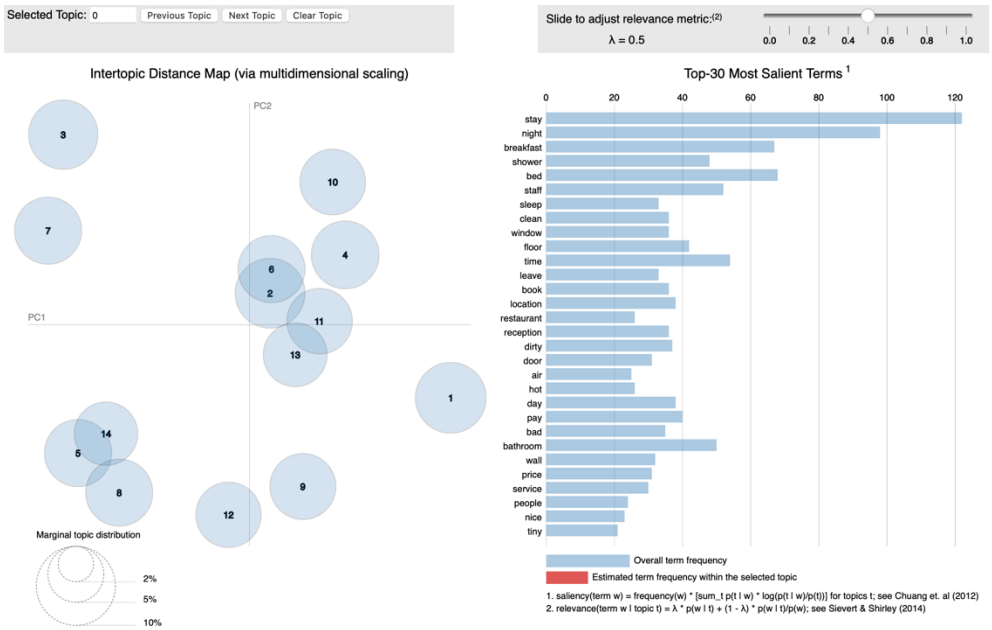


Figure 17. Negative topics visualisation

Like positive topics, the major 3 topics chosen which affect customer dissatisfaction are topics 1, 3 and 12.

- **Topic 1: Amenities**
  words like "*coffee*", "*issue*" shows that many guests were disappointed with the lack of basic amenities like decent morning coffee. Issues with refunds and unhelpful staff regarding lost luggage stirred the frustration of some guests who felt the price wasn't worth the experience.

- **Topic 3: Food quality**
  words like "*egg*", "*breakfast*", "*bad*" shows that guests encountered breakfast issues throughout their stay. Complaints included bad breakfast and unclean bathrooms. The lack of a good internet connection, noisy street further contributed to the negative experience.
- **Topic 12: Room/Facilities**
  Words like "tiny", "stain", "water" shows that guests mentioned malfunctioning showers, wet floors, and uncomfortable beds in tiny rooms. Water leaks, stained carpets, and dirty staircases added to the negative impression.

Limitations

Lemmatization would be computationally intensive if the dataset is large. Bigram tokenisation may be more suitable for this case but requires more resources and is computationally intensive. There are other metrics such as semantic coherence, perplexity score which might be more efficient in determining the number of topics.

References

Hyndman, R. (2014). *Forecasting: Principles & practice*. [online] Available at: https://robjhyndman.com/uwafiles/fpp-notes.pdf.

Anon, (n.d.). *Interpolation methods for time series data – d3VIEW*. [online] Available at: https://www.d3view.com/interpolation-methods-for-time-series-data/.

Oracle.com. (2022). *Introduction to forecating with ARIMA in r*. [online] Available at: https://blogs.oracle.com/ai-and-datascience/post/introduction-to-forecasting-with-arima-in-r.

Donato_TH (2022). *Basic Time Series Forecasting with R (Part 1)*. [online] Medium. Available at: https://medium.com/@designbynattapong/basic-time-series-forecasting-part-1-28238764ed7.

Shabou, S. (n.d.). *Chapter 4 Time Series Forecasting | Time Series with R*. [online] *s-ai-f.github.io*. Available at: https://s-ai-f.github.io/Time-Series/time-series-forecasting.html.

Kabacoff, R.I. (2015). *R in Action*. Simon and Schuster.

SVETUNKOV, I. (2023). *Forecasting and Analytics with the Augmented Dynamic Adaptive Model (Adam)*.

Robinson, J.S. and D. (n.d.). *6 Topic modeling | Text Mining with R*. [online] *www.tidytextmining.com*. Available at: https://www.tidytextmining.com/topicmodeling.

ladal.edu.au. (n.d.). *Topic Modeling with R*. [online] Available at: https://ladal.edu.au/topicmodels.html.