Comparison of Forecasting Models for US Seasonally-Adjusted Personal Consumption Expenditures

Introduction

The prediction of personal consumption expenditures (PCE) holds significant importance for understanding economic trends and making informed decisions. In this report, we undertake a comparative analysis of three distinct forecasting models to determine the most effective approach for predicting US seasonally-adjusted PCE. The models evaluated include a simple forecasting method, an Exponential smoothing model, and an ARIMA model.

Our primary objective is to identify the model that exhibits the highest predictive accuracy in forecasting PCE values. The analysis concludes the present and interpret of the selection criteria utilized to evaluate the performance of each model, methods used and their accuracies. We compare the predictions of each model with the actual PCE values in a single graph, facilitating a straightforward evaluation of their predictive capabilities. Our aim is to offer actionable insights into forecasting personal consumption expenditures by rigorously analyzing and interpreting the results.

Findings of the best-performing model to estimate the PCE for October 2024. Furthermore, we conduct a rolling forecasting comparison, evaluating the models' predictive performance using one-step ahead predictions without re-estimating parameters. This approach provides further insights into the robustness and stability of each model.

Data inspection

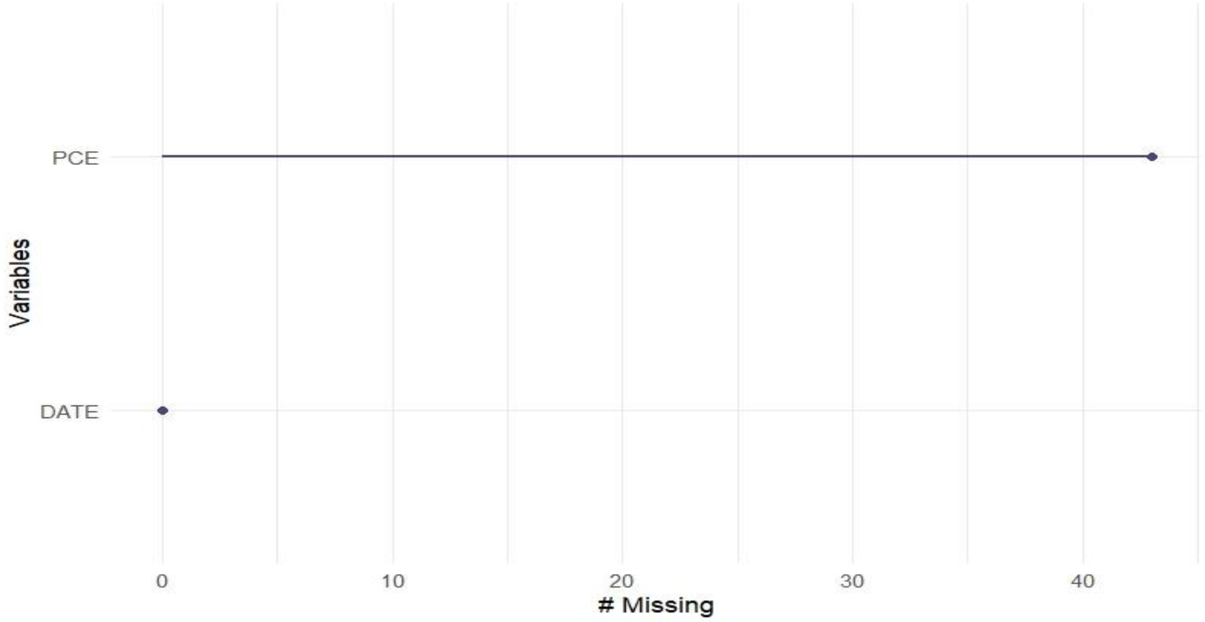
```
> skim(data)
── Data Summary ────────────────────────
                           Values
Name                       data
Number of rows             779
Number of columns          2
_______________________
Column type frequency:
  character                1
  numeric                  1
________________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 DATE                  0             1  10  10     0      779          0

── Variable type: numeric ────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate  mean    sd   p0   p25  p50   p75   p100 hist
1 PCE                  43         0.945 5792. 5067. 306. 1125. 4270 9897. 18859. █▂▂▂▁
```

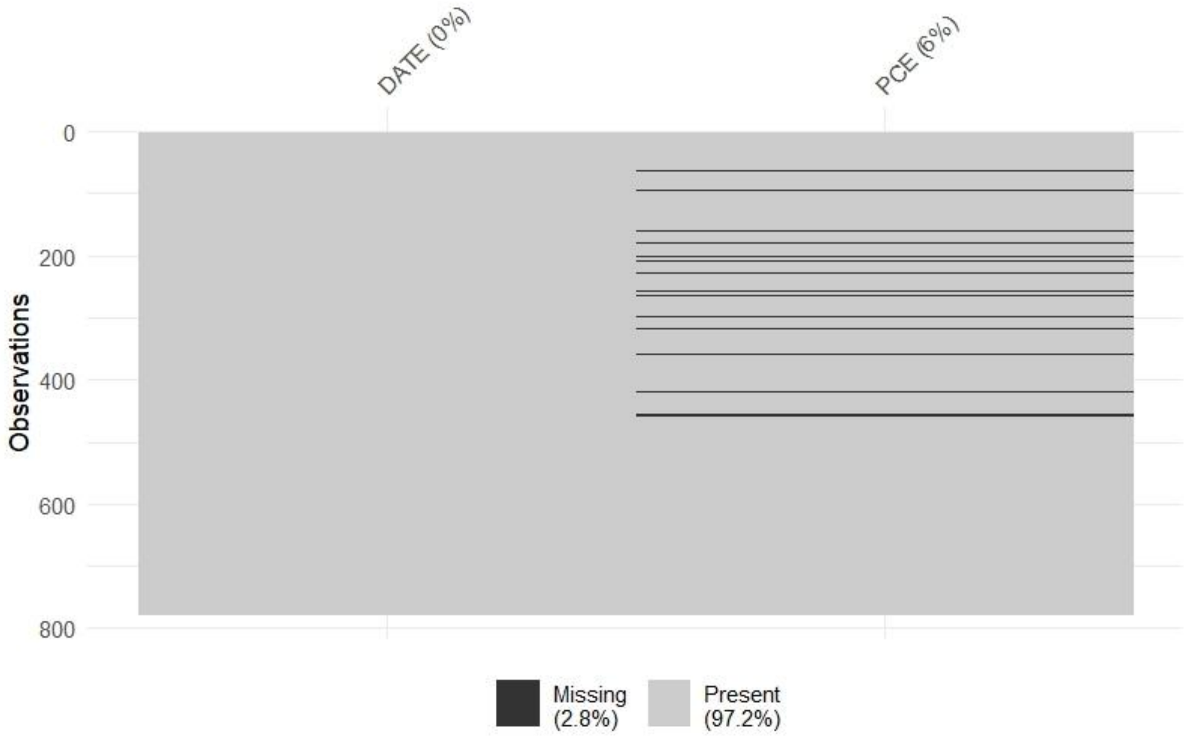
Data Preprocessing:

The initial phase of our analysis focuses on preparing the personal consumption expenditure (PCE) data for further examination. With the PCE data being seasonally adjusted, our aim in data preprocessing is to ensure its integrity and suitability for analysis. This involves implementing various techniques to address missing data, assess for white noise, and appropriately partition the dataset for subsequent analysis.

Missing Data Handling:

During the initial examination, it was identified that the PCE column contained a total of 43 missing values. To address this issue, we employed the moving average method from the imputeTS library. This method utilizes a moving window approach to compute the average of neighboring observations, thereby generating a smoothed estimate of the missing values. By incorporating information from adjacent data points, the moving average method effectively preserves the underlying trend in the data while reducing the impact of noise or fluctuations. Leveraging the moving average method ensures robustness and usability in anticipating future implications of the code.





Checking for white noise

White noise is characterized by random fluctuations with constant variance and no correlation between consecutive observations. Ljung box test was done to determine whether there is significant autocorrelation in a time series. The p-value here is almost zero, which is suggesting significant autocorrelation in the time series. Here we observed that the results of the Box-Ljung test suggest that there is significant autocorrelation in the time series. This implies that the time series data exhibit patterns or dependencies between consecutive observations, which should be considered when modelling or analysing the data.

```
> Box.test(myts, lag = 24, fitdf=0, type = "Lj")

	Box-Ljung test

data:  myts
X-squared = 16115, df = 24, p-value < 2.2e-16
```

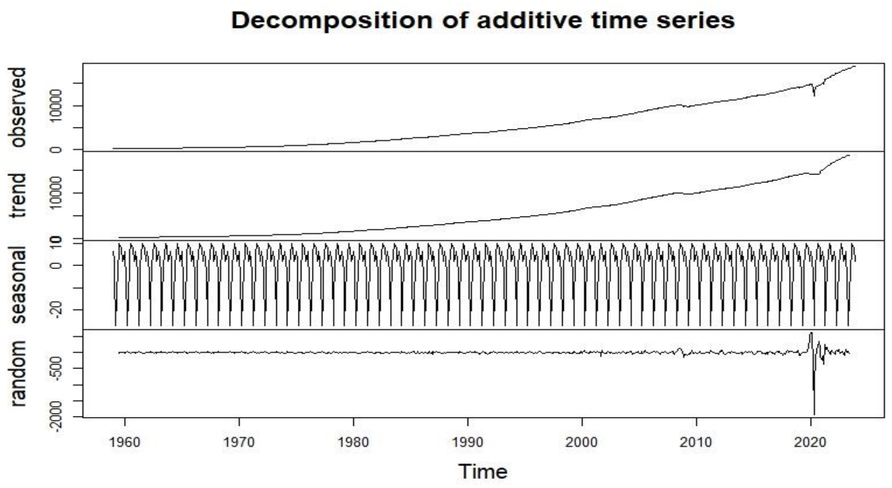
Dataset Splitting

The data is split into train and test for validation purposes, here the split that I have chosen is 80/20 split. Keeping in mind the data has a huge anomaly in it which can hinder the prediction and affect the overall accuracy of the forecast. Different splits are taken to test each splits ability to predict the best possible forecast such as 90/10, 85/15. This changes effect the RMSE, ME, MAE, MPE, MAPE, MASE, ACF1 while test there wasn't enough data to properly assess the training data.

Decomposing data

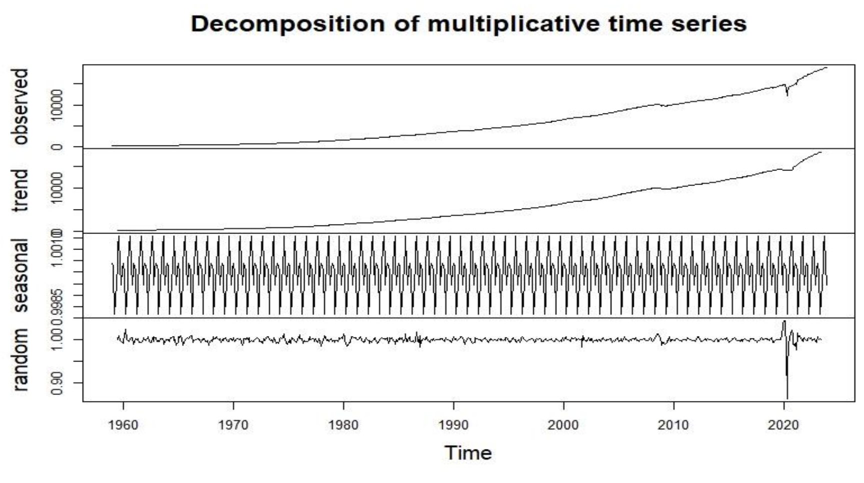
Additive decompose

Additive decomposition is a technique employed to break down a time series into its key constituents: trend, seasonality, and error. In this method, these components are presumed to be unrelated to the level of the series. Consequently, the magnitude of seasonal fluctuations and the trend's intensity are considered consistent across the series.



Multiplicative Decomposition

Multiplicative decomposition is another technique used to break down a time series into its constituent parts: trend, seasonality, and error. In contrast to additive decomposition, multiplicative decomposition assumes that the seasonal and trend components are proportional to the level of the series.



Simple Forecasting Methods

Here, we have employed four simple forecasting methods utilizing the forecast library. These methods utilize historical data to project future values. Here the ‘datasetcomplete’ is the complete data in a time series.

Naïve Method:

The Naïve Method, also known as the last observation method, predicts future values by simply taking the value of the last period as the forecast for the next period.

```
> accuracy(fcsnaive,datasetComplete)
                    ME      RMSE       MAE       MPE      MAPE     MASE      ACF1 Theil's U
Training set  194.5976  251.4858  202.7813  6.562192  6.645324  1.00000 0.9761424        NA
Test set     3412.3455 4141.8834 3412.3455 22.931063 22.931063 16.82772 0.9752355  17.50232
```

Mean Method:

The Mean Method, also called the Average Method, computes forecasts based on the average of past observations. It calculates the mean from historical data and uses this average to forecast future periods. The meanf function is used here

```
> accuracy(fcmean,datasetComplete)
                        ME      RMSE       MAE        MPE      MAPE     MASE      ACF1 Theil's U
Training set -1.421442e-13  3124.649  2658.459 -205.33471 239.61696 13.10998 0.9951973        NA
Test set      1.008371e+04 10354.918 10083.713   73.20916  73.20916 49.72704 0.9746013  49.13873
```

Drift Method:

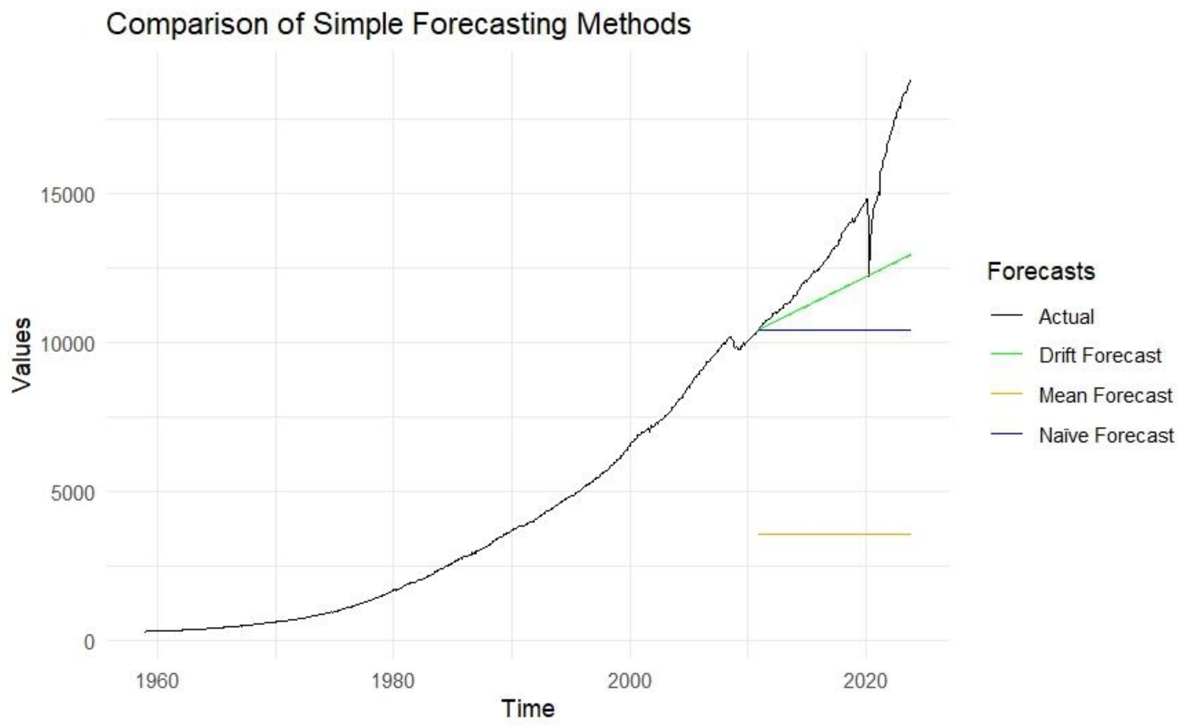
The Drift Method, also known as the Random Walk Method, incorporates the trend of the data into the forecasting process. It assumes that future values follow a linear trend, where each new observation drifts away from the previous value by a constant amount. The rwf function used to

```
> accuracy(fcdrift,datasetComplete)
                       ME       RMSE        MAE        MPE      MAPE       MASE      ACF1 Theil's U
Training set 2.564585e-14   24.83631   16.72042 -0.8283488  1.131257 0.08245545 0.1206498        NA
Test set     1.925996e+03 2545.27790 1926.57567 12.5872452 12.591982 9.50075796 0.9702413  10.38423
```

Evaluation of the models

The Drift Method outperforms the Naïve and Mean Methods, demonstrating lower RMSE and MAPE values on both the training and test sets. This indicates that the Drift Method successfully captures the underlying trend in the data, leading to more precise forecasts. The Drift Method appears to be the most effective for forecasting the personal consumption expenditures dataset.

Here is the plot which all the forecasts are plotted with respect to the whole dataset,



Exponential Smoothing Model

Exponential Smoothing Model to forecast personal consumption expenditures. Since the data is seasonally adjusted, we consider three variations of the Exponential Smoothing Model: Simple Exponential Smoothing Method, Holt's Method, and ETS (Error, Trend, and Seasonality) Auto Method.

Holt's Method

The Holt’s incorporates trend into the forecast as well, Holt’s method is also known as the double exponential smoothing. The Holt’s method account both the level and trend of the data which makes it perfect time series with linear trend.

```
> accuracy(fcholt,datasetComplete)
                    ME       RMSE       MAE        MPE      MAPE       MASE        ACF1 Theil's U
Training set   0.4355215   22.47443  12.37527 0.02493505 0.3979732 0.06102766 -0.01491604        NA
Test set     564.4029497 1145.65882 645.24152 3.25457736 3.9165706 3.18195835  0.95685202   4.39793
```

ETS Method

ETS stands for error, trend, seasonality, this method automatically selects the optimal model to get the best forecast by the selection of optimal combination of (Error, Trend, and Seasonality) for the time series. The model use automated algorithms to forecast the most suitable parameters.

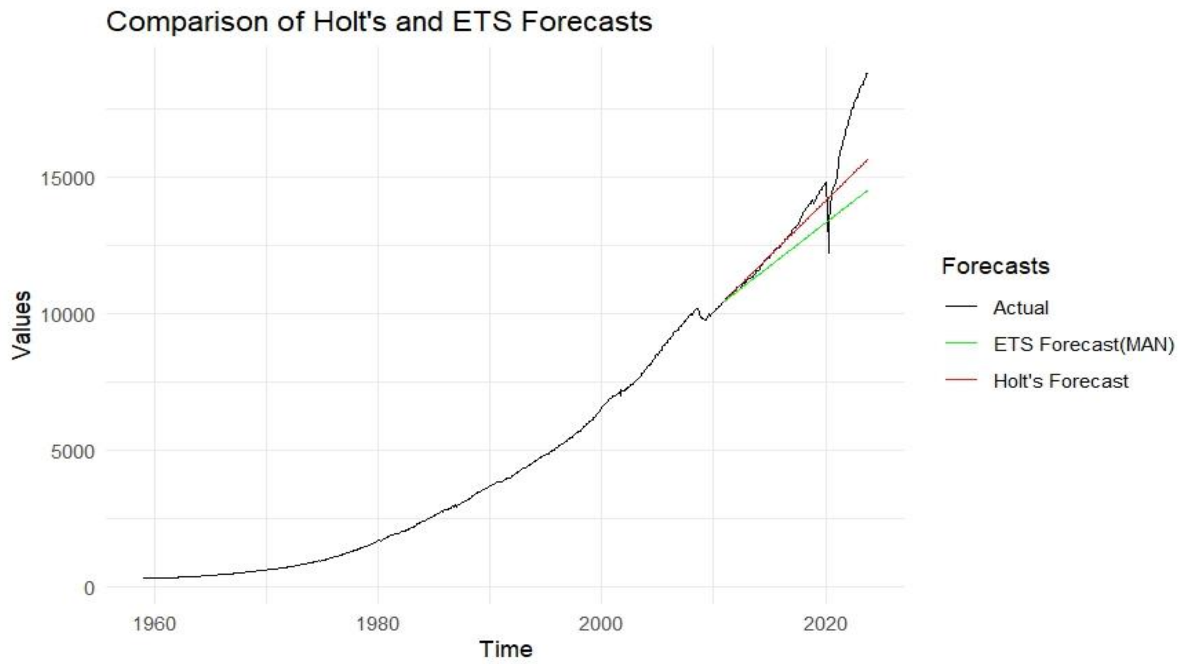
The model chosen for this data was,

```
Forecast method: ETS(M,A,N)

Model Information:
ETS(M,A,N)

Call:
 ets(y = train)

  Smoothing parameters:
    alpha = 0.8075
    beta  = 0.0589

  Initial states:
    l = 304.9825
    b = 1.5614

  sigma:  0.0054

     AIC     AICc      BIC
7043.600 7043.698 7065.773
```

```
> accuracy(fcets,datasetComplete)
                      ME       RMSE        MAE        MPE      MAPE      MASE      ACF1 Theil's U
Training set    0.6772146   22.67917   12.25876 0.04721351 0.3937112 0.0604531 0.1006397        NA
Test set     1137.6440849 1697.43018 1155.18692 7.18639219 7.3283490 5.6967144 0.9643726  6.696954
```

The Holt's Method performs better than both the Simple Exponential Smoothing Method and ETS Method, showing lower RMSE and MAPE values on both the training and test sets. This suggests that the Holt's Method effectively captures the trend components in the data, resulting in more accurate forecasts.

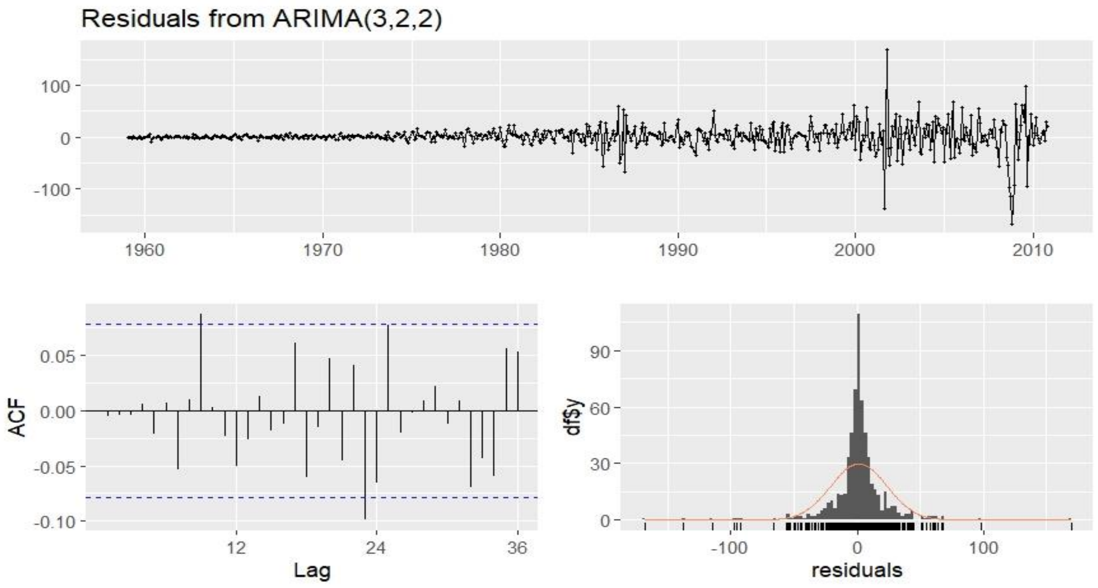
Holt's method appears to be the best model for this dataset as it has the lowest error rates on the test set, indicating better generalization to unseen data. The ETS model seems to be over fitting to the training data, and the SES model has the highest error rates on both the training and test sets.



ARIMA Model

ARIMA model and its components (autoregressive, differencing, moving average). ARIMA is a flexible and widely used model for time series forecasting, capable of capturing both linear and nonlinear relationships in the data. It's particularly useful for analysing and predicting data with trends and seasonal patterns. It is denoted as ARIMA(p, d, q), 'p' signifies the autoregressive component's order (AR), 'd' represents the degree of differencing necessary for achieving stationarity in the time series, and 'q' indicates the order of the moving average component (MA).

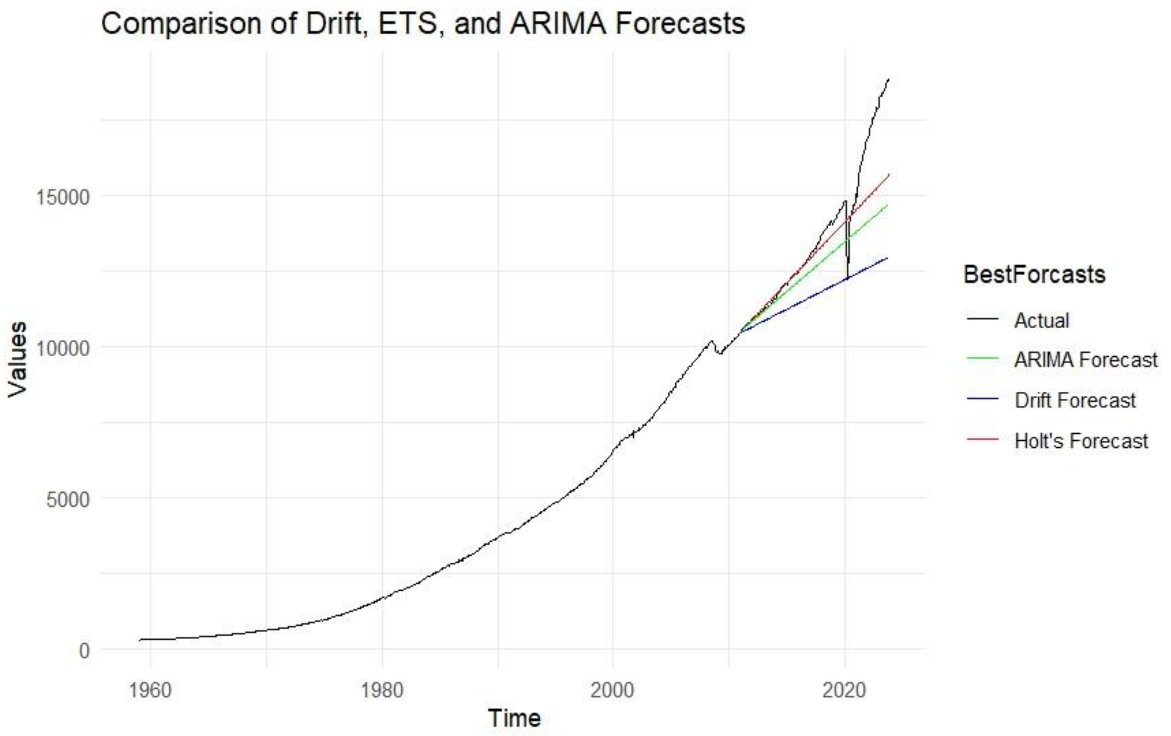
```
> summary(afit)
Series: train
ARIMA(3,2,2)

Coefficients:
         ar1     ar2     ar3      ma1     ma2
      0.4571  0.1957  0.0658  -1.5282  0.5374
s.e.  0.1650  0.0445  0.0579   0.1622  0.1579

sigma^2 = 494.1:  log likelihood = -2805.9
AIC=5623.79   AICc=5623.93   BIC=5650.38

Training set error measures:
                   ME     RMSE      MAE        MPE      MAPE       MASE          ACF1
Training set 1.256645 22.10412 12.36151 0.06713522 0.4029679 0.06095981 -0.005380483
```

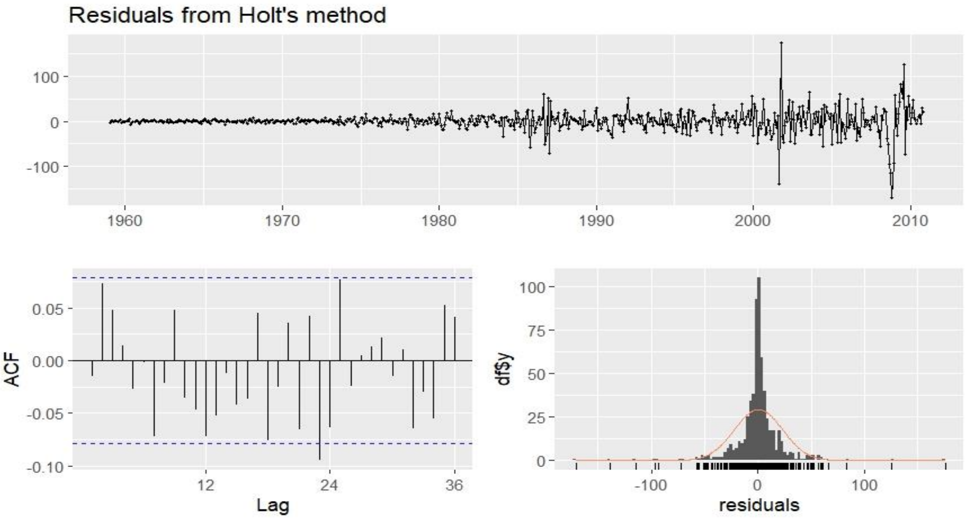


Other combination such as Arima(1,2,1), Arima(2,1,3), Arima(1,2,3) were tried but this had the best results.

Evaluation Criteria

Model	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
fcauto_arima	1.256645	22.10412	12.36151	0.06713522	0.4029679	0.06095981	-0.00538048	NA
fcholt	0.4355215	22.47443	12.37527	0.02493505	0.3979732	0.06102766	-0.01491604	NA
fcdrift	2.56E-14	24.83631	16.72042	-0.8283488	1.131257	0.08245545	0.1206498	NA
Test set								
fcauto_arima	1021.242576	1593.06514	1042.73242	6.36201339	6.5351073	5.14215378	0.963691956	6.240727
fcholt	564.4029497	1145.65882	645.24152	3.25457736	3.9165706	3.18195835	0.95685202	4.39793
fcdrift	1.93E+03	2545.2779	1926.57567	12.5872452	12.591982	9.50075796	0.9702413	10.38423





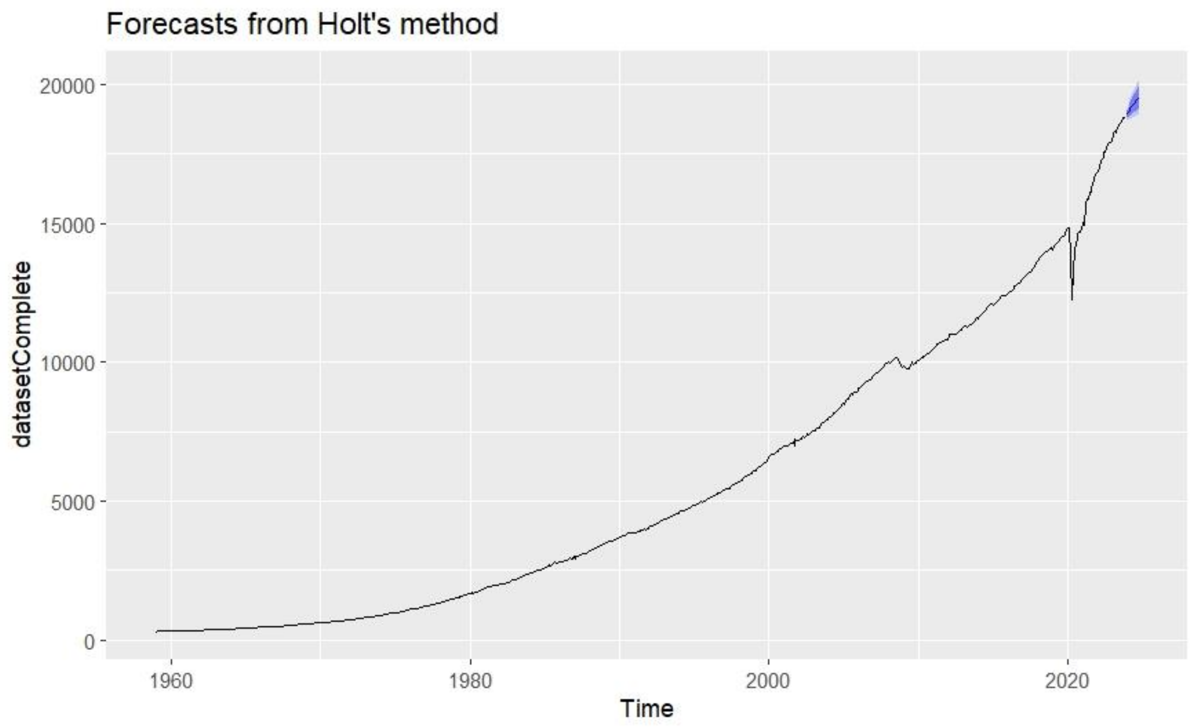
Prediction for October 2024

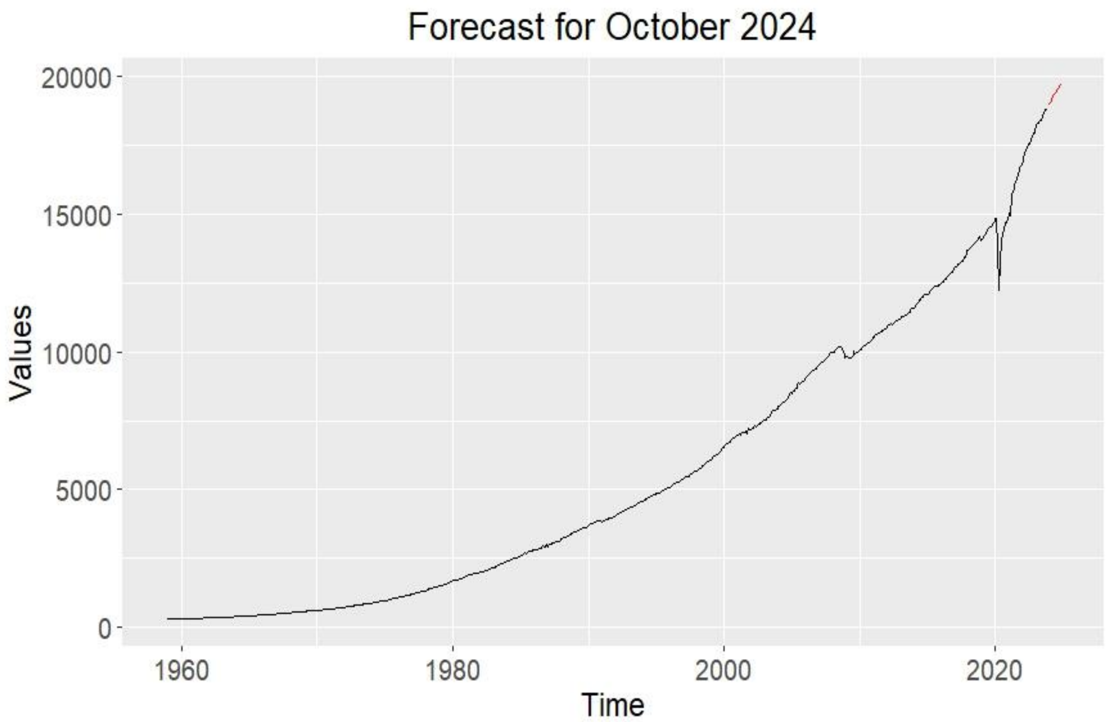
These are the prediction for 2024 October, this done using the best method (Holt's Linear method)

```
> fcholt_predict
         Point Forecast     Lo 80    Hi 80    Lo 95    Hi 95
Dec 2023       18923.27 18805.23 19041.31 18742.74 19103.80
Jan 2024       18987.63 18819.53 19155.74 18730.54 19244.73
Feb 2024       19052.00 18844.67 19259.33 18734.92 19369.08
Mar 2024       19116.36 18875.29 19357.44 18747.67 19485.06
Apr 2024       19180.73 18909.32 19452.13 18765.65 19595.81
May 2024       19245.09 18945.73 19544.46 18787.25 19702.94
Jun 2024       19309.46 18983.88 19635.04 18811.53 19807.39
Jul 2024       19373.83 19023.38 19724.27 18837.86 19909.79
Aug 2024       19438.19 19063.95 19812.43 18865.84 20010.54
Sep 2024       19502.56 19105.39 19899.72 18895.15 20109.96
Oct 2024       19566.92 19147.56 19986.28 18925.56 20208.28
```

The values donates the best fitting forecast, which is denoted by Forecast and by low 80 and High 80 the prediction could be inside there 80% of the time, but with 90 low and high the forecast is said to be 90% sure about the accuracy of the data.

In conclusion, based on the forecasted values for personal consumption expenditures (PCE) for the year 2024, we anticipate a gradual increase in PCE over the months, with October 2024 estimated to reach $19,566.92. These forecasts, derived from our chosen predictive model Holt's method, which provides valuable insights for economic planning and decision-making.



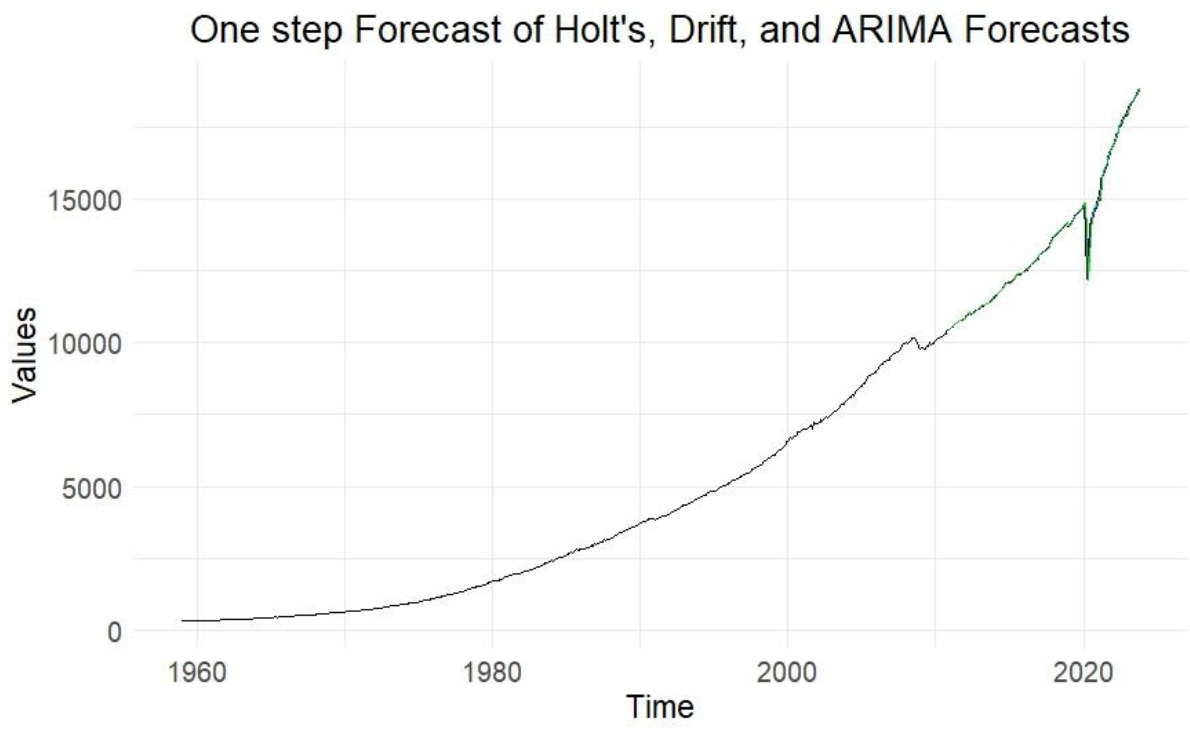


One-Step Ahead Rolling Forecasting

This involves predicting the value of a time series variable at the next time point based on the historical data available up to the current time point. This report aims to discuss the process of one-step ahead forecasting using various forecasting techniques and evaluate their performance based on accuracy metrics.

Model	Mean Error (ME)	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)	Mean Percentage Error (MPE)	Mean Absolute Percentage Error (MAPE)	Autocorrelation at Lag 1 (ACF1)	Theil’s U statistic
ARIMA	8.84	221.46	73.88	0.0496	0.536	0.26	1.085
Holt	16.99	200.12	69.41	0.101	0.501	0.174	0.974
Drift	30.16	201.36	75.69	0.188	0.541	0.183	0.976

While the ARIMA model has the lowest Mean Error (ME), it has the highest Root Mean Squared Error (RMSE) and Theil’s U statistic. On the other hand, the Holt model has the lowest Root Mean Squared Error (RMSE) and Theil’s U statistic, but its Mean Error (ME) is higher than that of the ARIMA model. The Drift model has the highest Mean Error (ME) and Mean Absolute Error (MAE), but its Root Mean Squared Error (RMSE) and Theil’s U statistic are comparable to those of the Holt model.



In the evaluation of our forecasting models, we observed varying levels of performance based on one-step ahead predictions. The drift model exhibited a Mean Error (ME) of 30.16, Root Mean Square Error (RMSE) of 201.36, and Mean Absolute Error (MAE) of 75.69. The Holt model showed improved metrics with ME of 16.99, RMSE of 200.12, and MAE of 69.41. Conversely, the ARIMA model demonstrated the lowest error metrics, with ME of 8.84, RMSE of 221.46, and MAE of 73.88.

Overall, these results indicate that the Holt model outperformed both the drift and ARIMA models in terms of accuracy, as evidenced by lower error metrics. However, it's essential to consider other factors such as computational complexity and model interpretability when selecting the most suitable

Part 2: Topic Modelling of Hotel Customer Reviews

Introduction

This report delves into the analysis of hotel reviews, aiming to unveil the underlying themes and sentiments expressed within them. Hotel reviews serve as invaluable resources for both hotel management and potential guests, providing insights into various aspects of the guest experience, including service quality, amenities, cleanliness, and overall satisfaction.

The main objective of this study is to conduct a thorough analysis of hotel reviews by leveraging advanced natural language processing (NLP) techniques and machine learning algorithms. Specifically, the goals are as follows:

- Sentiment Analysis: The first step involves performing sentiment analysis to categorize reviews as positive, negative, or neutral based on their associated ratings. This analysis aims to gauge the overall sentiment expressed in the reviews and identify trends in guest satisfaction.
- Topic Modeling: The study utilizes Latent Dirichlet Allocation (LDA), a probabilistic generative model, to uncover latent topics within the corpus of hotel reviews. By identifying recurring themes and topics, we gain insights into the key aspects of the guest experience that are most commonly discussed in the reviews.
- Visualization: To facilitate understanding and interpretation, the distribution of topics and their associated keywords are visualized using interactive visualization tools such as LDAvis. These visualizations provide a clear and intuitive representation of the topics discussed in the reviews, aiding in the identification of patterns and trends.

Data Sampling

Sampling Process

To obtain a manageable subset of the hotel review dataset, we employed the sample_n() function from the dplyr package in R. Prior to sampling, it was imperative to ensure reproducibility of results. Therefore, the set.seed(868) function was utilized, which is was specified in the instruction.
A sample of 2000 reviews was selected, The sample_n() function was chosen for taking the 2000 review

Selecting reviews that are English

The data consist of many languages, where in this analysis we are only considering the English reviews. The tables shows all the languages that are present in the data, using cld3 library we use language detection function to detect all the languages present in the data.

```
> table(language) # There are 7982 english reviews
language
  ar   cs   da   de   el   en   es   fi   fr   gl   id   it   iw   ja   ko   mi   nl   no 
   2    1   22  196    6 7982  348    3  421    1    2  595    3   63    5    1   51   26 
  pl   pt   ru   sv   th   tr   zh 
   8  153   55   44    1    2    5 
```

The function detects there are 7982 reviews which are English and the rest of the data consist of other languages. Here we only considering the English reviews for this analysis.

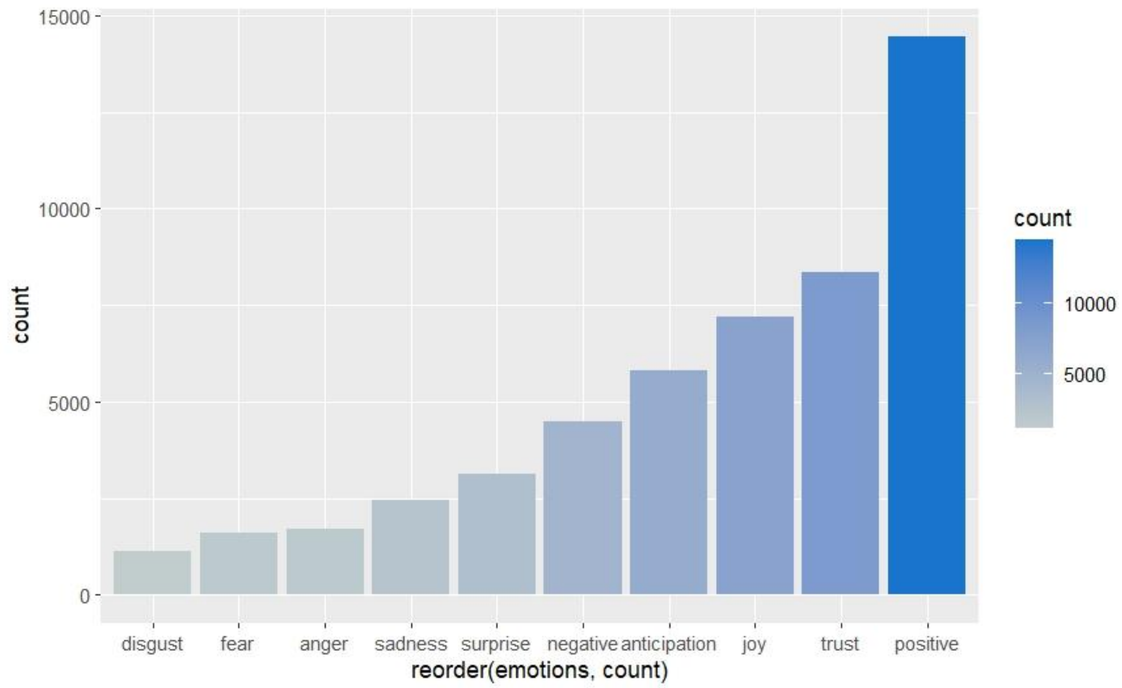
```
> skim(Hotel_data)
── Data Summary ────────────────────────
                           Values    
Name                       Hotel_data
Number of rows             2000      
Number of columns          2         
_______________________              
Column type frequency:               
  character                1         
  numeric                  1         
________________________             
Group variables            None      

── Variable type: character ──────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min  max empty n_unique whitespace
1 Text.1                0             1  50 5372     0     2000          0

── Variable type: numeric ────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate mean   sd p0 p25 p50 p75 p100 hist 
1 Review.score          0             1 4.05 1.14  1   4   4   5    5 ▁▁▁▃▇
```

Classification of Reviews

Classification of hotel reviews into positive, negative, or neutral categories is a fundamental step in s analysis to identify the reviews we are considering to sentiment analysis. In this section, we outline the criteria used to classify reviews and describe the methodology employed for sentiment analysis.

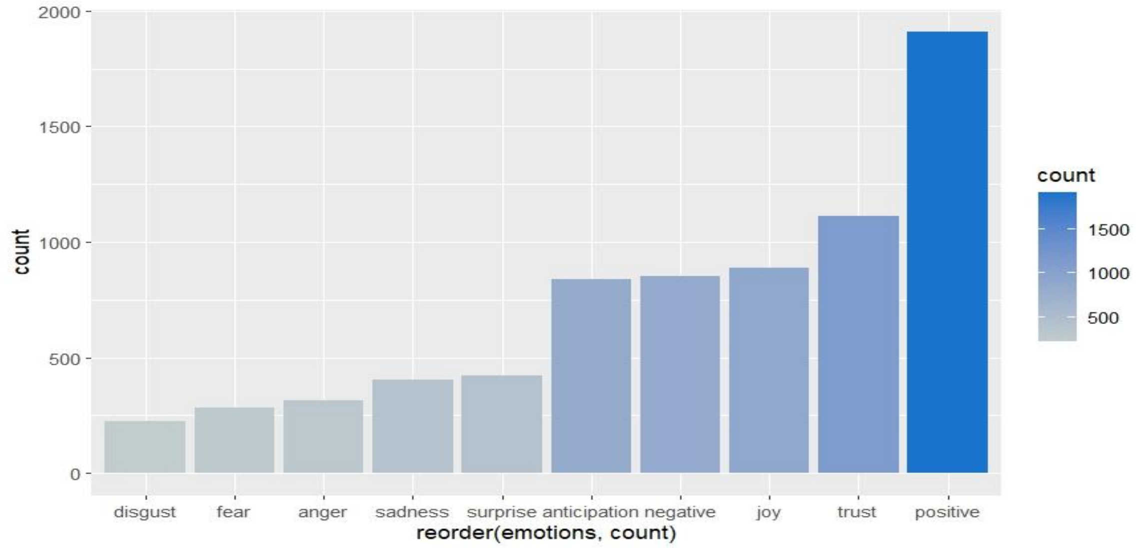
Criteria for Classification for Sentimental Analysis

The classification of reviews was based on the Likert scale ratings provided by customers. Reviews with higher ratings (4 or 5) were categorized as "positive," indicating high satisfaction levels. Conversely, reviews with lower ratings (1 or 2) were classified as "negative," signifying lower satisfaction levels. Reviews with a rating of 3 were considered "neutral," representing a moderate level of satisfaction.



Sentimental Analysis for “neutral” the objective was to identify underlying factors contributing to a neutral sentiment and to determine whether these factors aligned more closely with positive or negative sentiments.

Neutral reviews offer valuable insights into hotel experiences, highlighting aspects that may not strongly impact overall satisfaction but are still noteworthy. By examining neutral sentiments and incorporating pertinent factors into either positive or negative categories, a more thorough comprehension of guest sentiment and opportunities for enhancement can be attained.



The graph illustrates a tendency toward the positive reviews with the sentiment analysis results. So the ‘3’ neutral was also included in the positive reviews.

Data classification into Positive and Negative reviews

The data was split into 2 with,

Positive data	Negative data
1773	227

Text Pre-processing

Aimed at refining raw text into a clean, structured format suitable for further analysis. In the context of this study on hotel review.

Corpus Creation

The hotel reviews dataset is transformed into a corpus, where each review serves as a separate document. This allows for the organization and manipulation of text data at the document level.

Wordcloud Visualization

To gain a visual understanding of the most common terms in both positive and negative reviews, word clouds are generated. These visualizations depict the relative frequency of terms through varying font sizes, providing insights into the key themes and topics present in the reviews. Word cloud is used to identify he words each of the reviews have,

Positive word cloud

Here there are the top 20 words in the positive review and 100 words presented as tht world cloud below.

```
> frequency_pos[1:20]
    hotel      room     staff    london      good      stay breakfast     great  location 
     2798      2290      1361      1106      1087      1016       983       908       798 
    rooms     clean    stayed      nice       one  friendly      well      just   helpful 
      739       716       687       630       613       578       559       548       500 
  service    really 
      500       489 
```

Negative cloud

Here there are the top 20 words in the positive review and 100 words presented as tht world cloud below.

```
> frequency_neg[1:20] #Example of the output
     room      hotel breakfast      stay       one     staff    london     night       bed
      505        432       155       152       143       137       128       127       118
     good      rooms       get    stayed      even      just      time     small     didnt
      116        108       100        98        95        94        92        90        87
     like     shower
       87         81
```

Topic Modelling

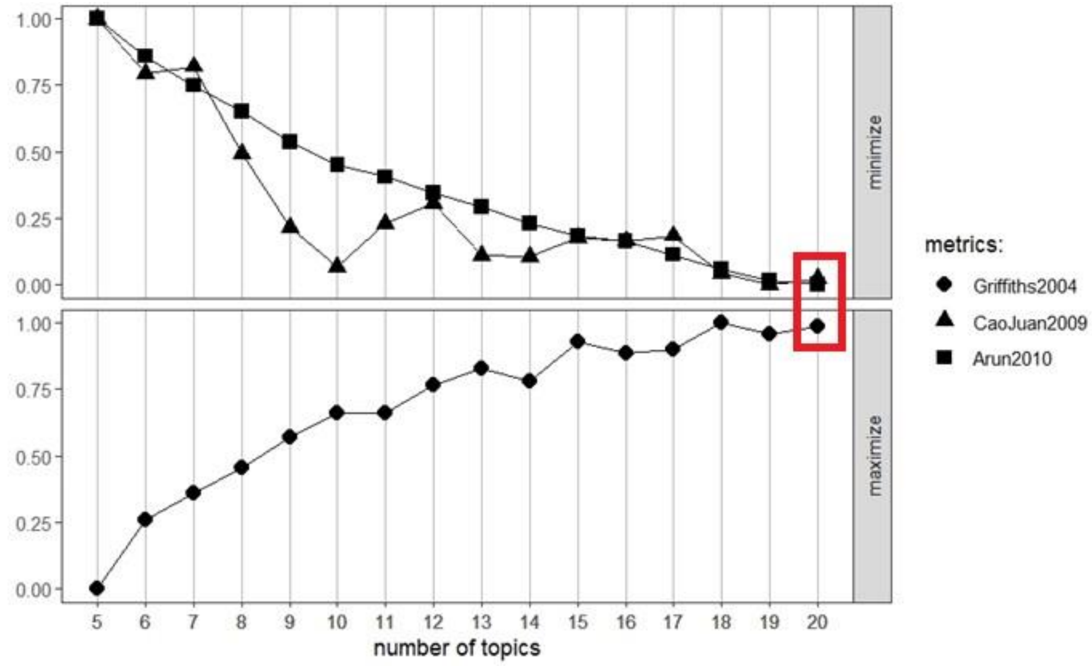
Latent Dirichlet Allocation (LDA) stands as a prevalent probabilistic model renowned for revealing latent themes embedded within a corpus of texts. Its fundamental assumption posits that every document comprises a blend of topics, each represented as a probability distribution over words.

Given its proven efficacy in discerning hidden patterns within textual data, LDA emerges as an apt selection for our analytical pursuits.

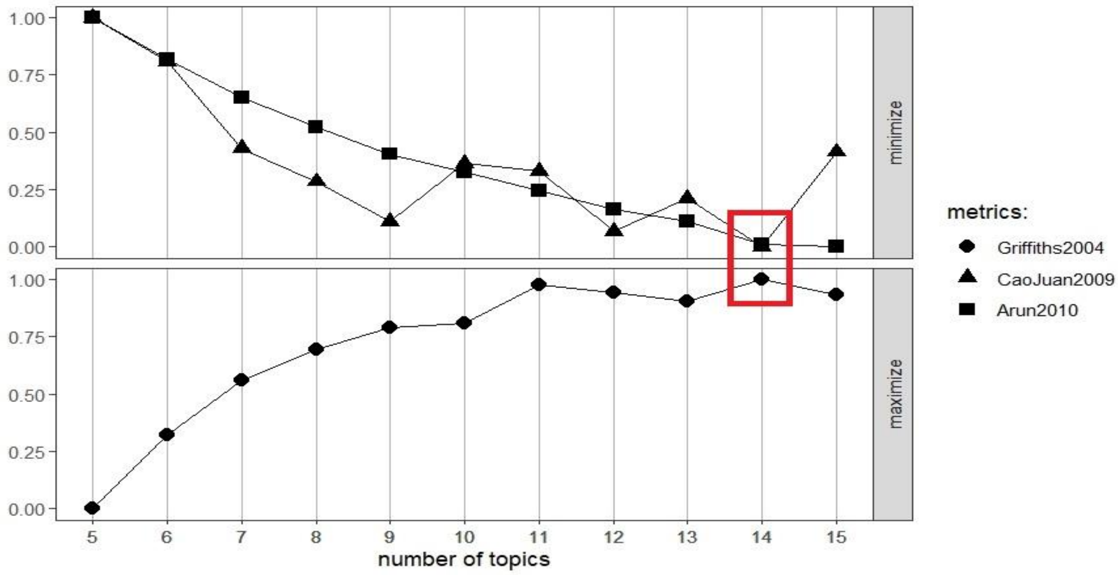
Number of Topics: Determining the appropriate number of topics is crucial for the interpretability of the results. We conducted a thorough evaluation using various metrics such as Griffiths2004, CaoJuan2009, and Arun2010 to select the optimal number of topics.

As the data is split into Positive and Negative reviews the topics has be found by,

Determining positive number of topics,



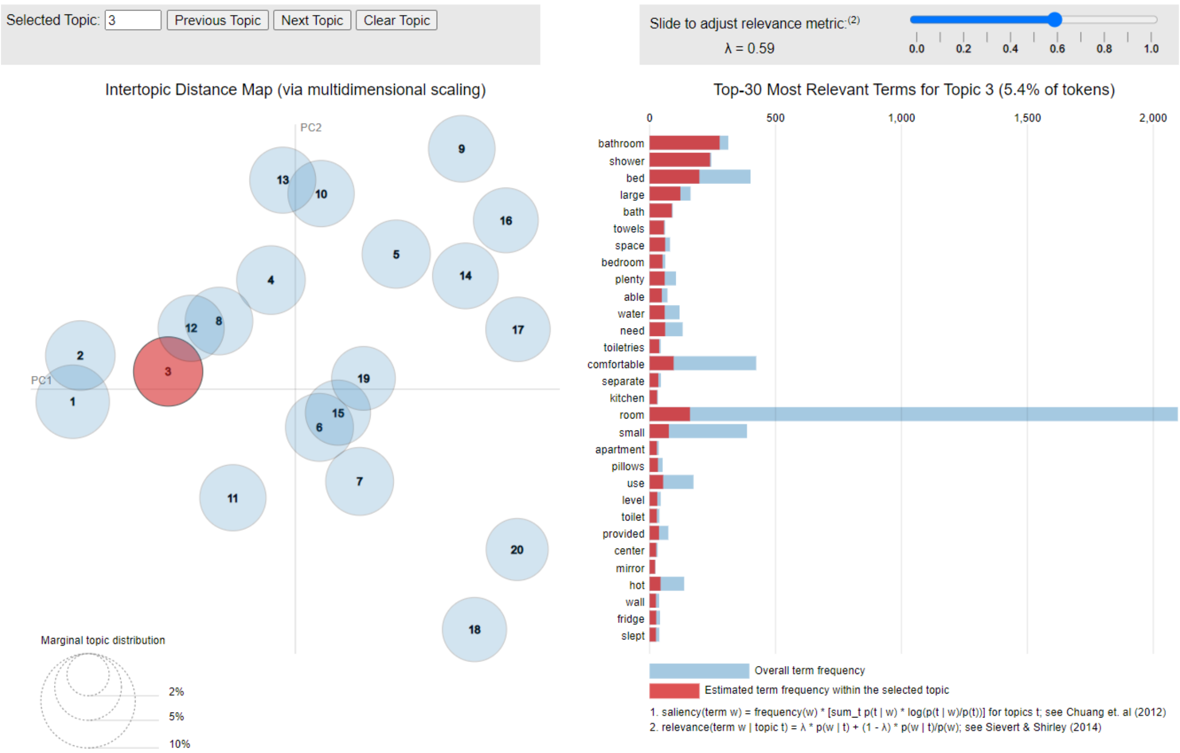
Determining negative number of topics,



```
> topicProbabilities[0:10,1:5]
           V1         V2         V3         V4         V5
1  0.02427184 0.05339806 0.10194175 0.07281553 0.02427184
2  0.05905512 0.02755906 0.03543307 0.02755906 0.01968504
3  0.03790614 0.04873646 0.01624549 0.03429603 0.04151625
4  0.02118644 0.02118644 0.02118644 0.07203390 0.05508475
5  0.02777778 0.29575163 0.03104575 0.03758170 0.03758170
6  0.03164557 0.03164557 0.03164557 0.05696203 0.03164557
7  0.06250000 0.08173077 0.04326923 0.02403846 0.06250000
8  0.07731959 0.05670103 0.02577320 0.04639175 0.04639175
9  0.03571429 0.06428571 0.07857143 0.03571429 0.05000000
10 0.08677686 0.07024793 0.06198347 0.05371901 0.02892562
```

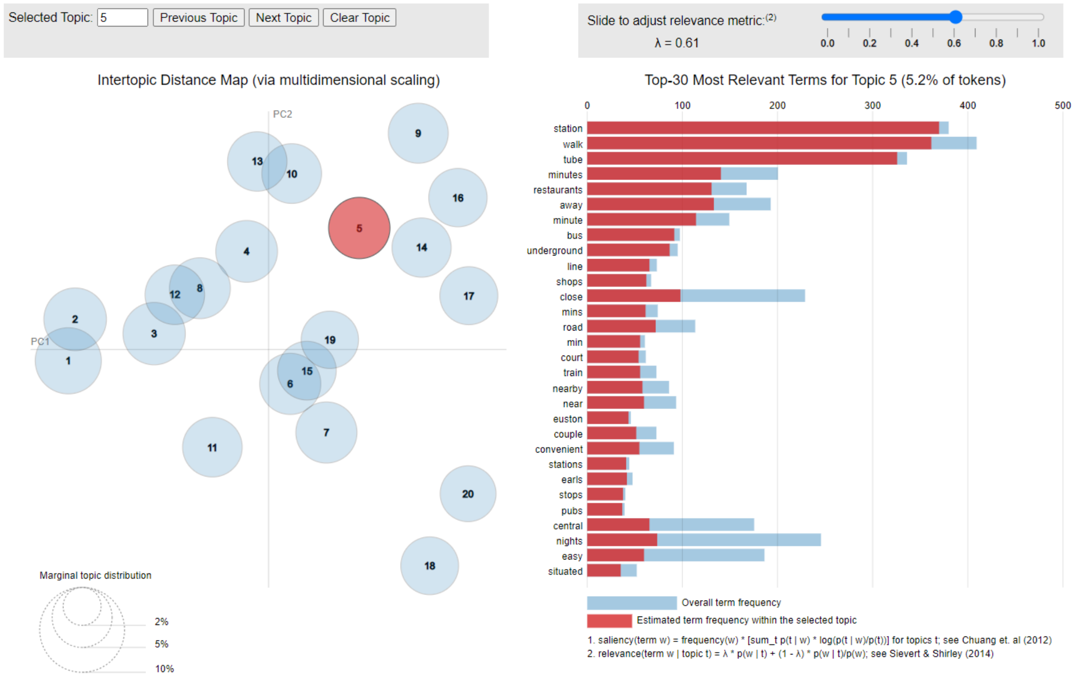
Positive Topic – 1

We could classify this as ‘Room’ or what the topics that could be associated with the rooms, bathrooms, facilities in the room. This could suggest the rooms are a factor satisfaction.



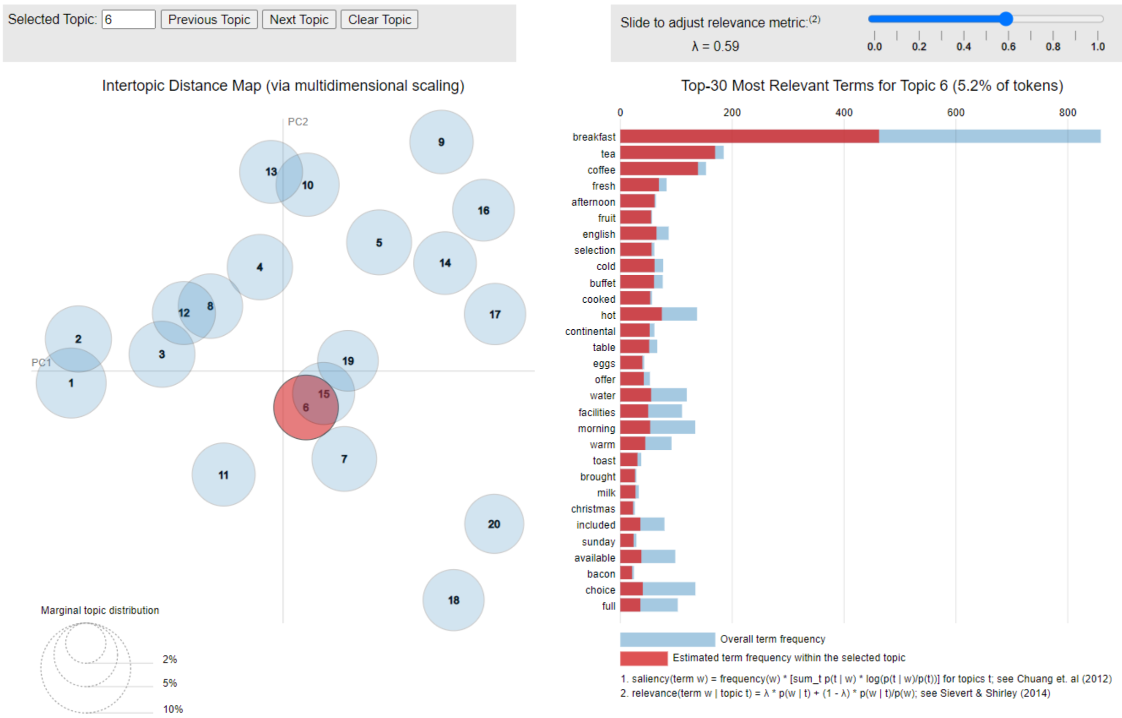
Positive Topic – 2

Transportation/Convenience, there are words such as station, tube, walk indicates that the hotel's placement in the city could be a factors that the customer's satisfaction.



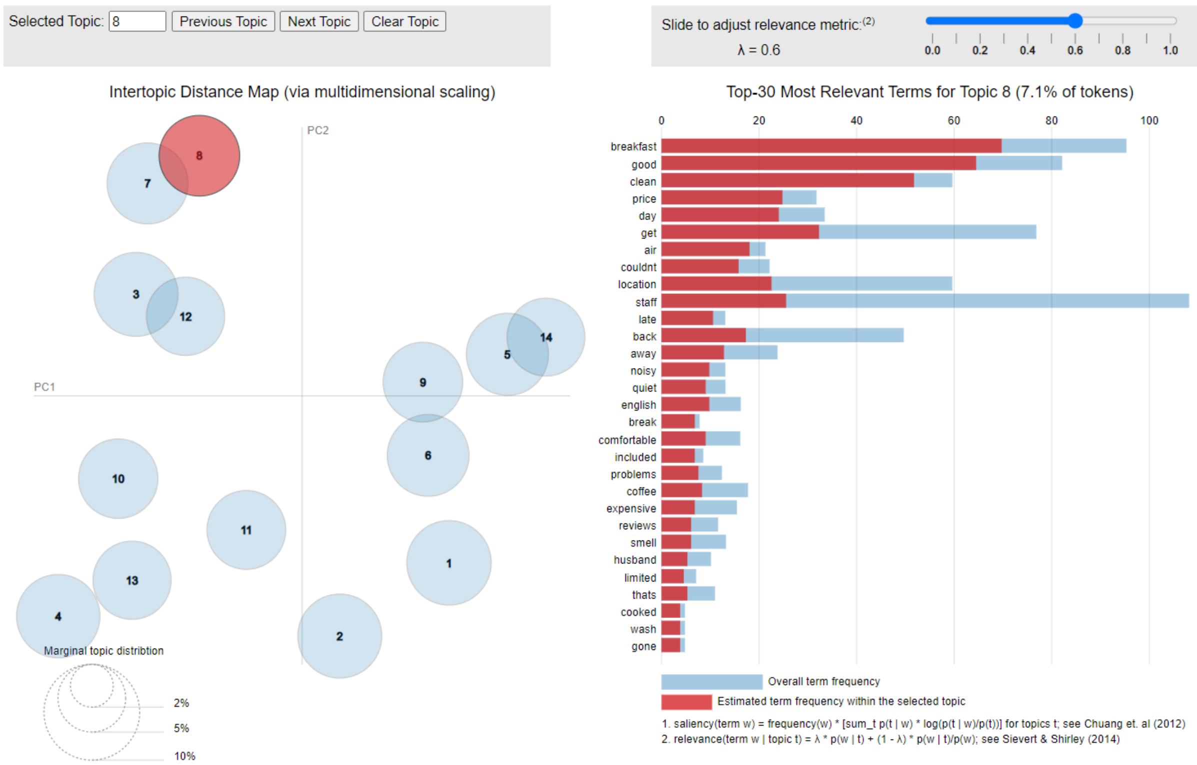
Positive Topic – 3

Food - Words like breakfast, tea, buffet, cooked could suggest the food in the hotel is a factor to the customer's satisfaction.



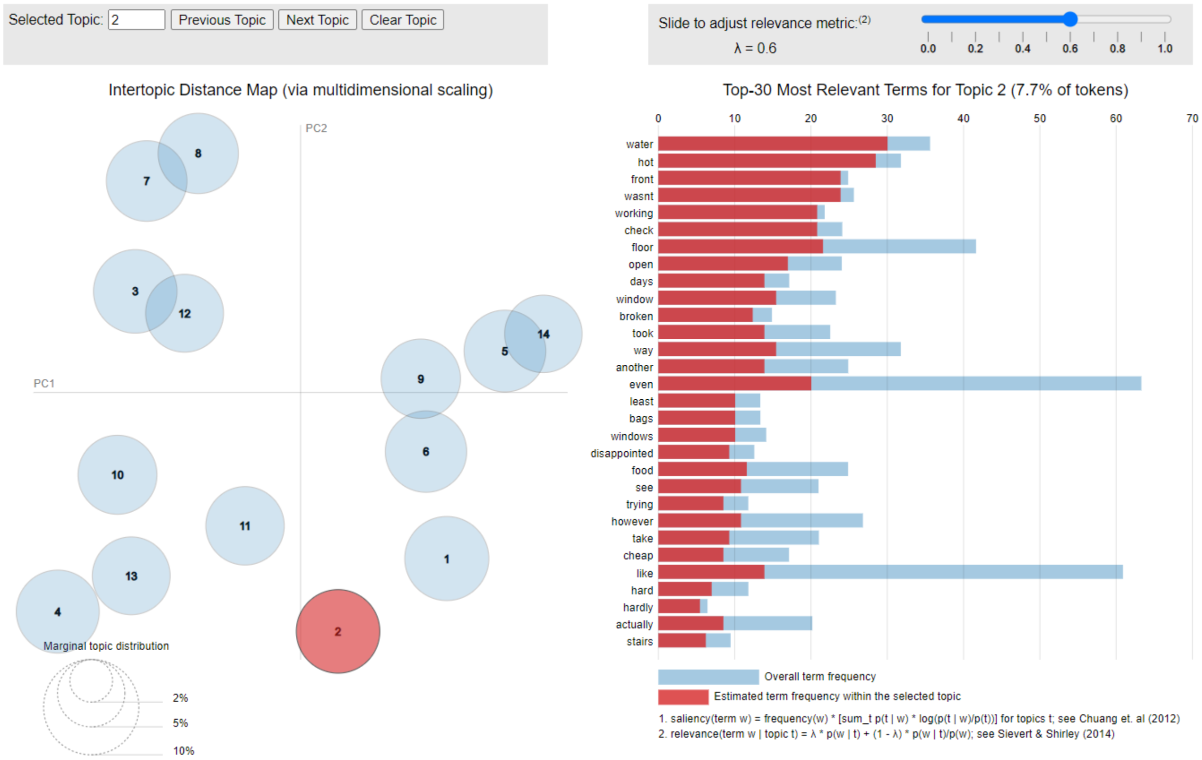
Negative Topic-1

Words like clean, expensive, coffee, smell all could be an indication about the restaurant's food or service.



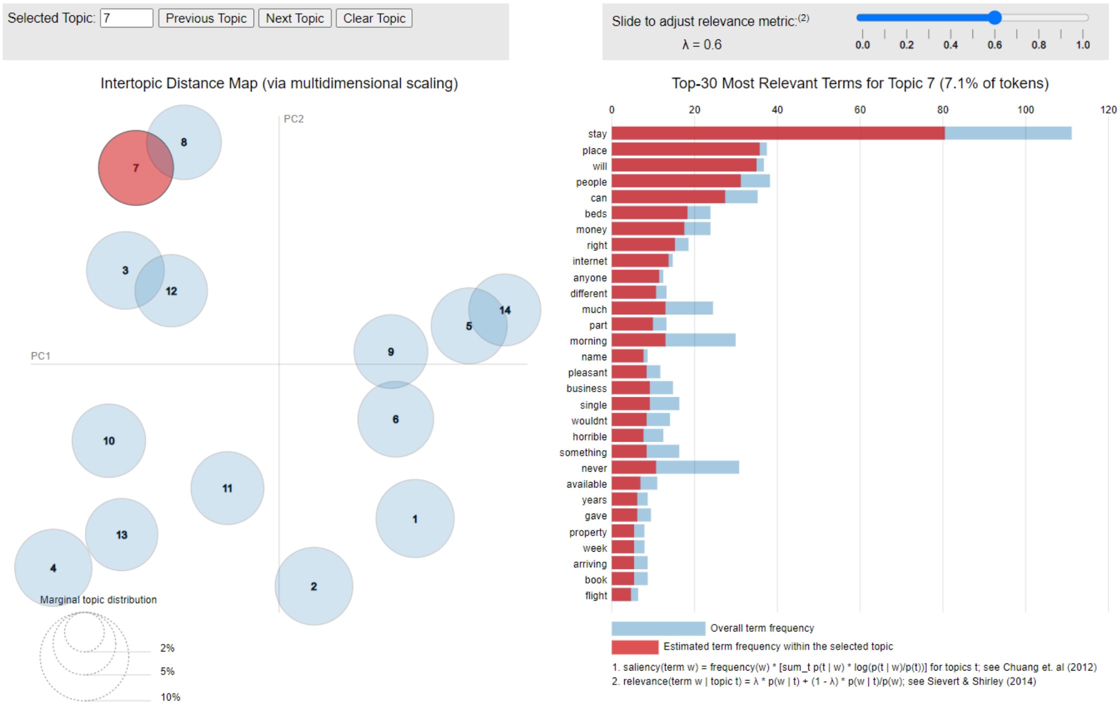
Negative Topic- 2

Disappointing food, stair inconvenience, broken window could indicate that the maintance of the hotel is not done properly which led to some dissatisfaction.



Negative Topic-3

General Topics – These are general topics with a high label count, this could be because of the association of to negative words there is clear topic for this. The words horrible, never



Selected Topic: 7
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0.6
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 7 (7.1% of tokens)
0 20 40 60 80 100 120
stay
place
will
people
can
beds
money
right
internet
anyone
different
much
part
morning
name
pleasant
business
single
wouldnt
horrible
something
never
available
years
gave
property
week
arriving
book
flight
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Analysis of Factors Affecting Satisfaction

```
Topic 1: Hotel Experience
 [1] "hotel"      "really"     "will"       "time"       "back"       "everything" "much"       "even"
 [9] "nothing"    "just"
Topic 2: Service Quality
 [1] "made"       "wonderful" "staff"      "like"       "amazing"    "way"        "feel"       "best"
 [9] "special"    "birthday"
Topic 3: Staff Interactions
 [1] "great"      "location"  "staff"      "rooms"      "stay"       "perfect"    "fantastic" "friendly"
 [9] "always"     "stayed"
Topic 4: Room Comfort
 [1] "bit"     "quite"    "didnt"    "however" "though"   "get"      "wasnt"    "enough"   "seemed"   "main"
Topic 5: Location Convenience
 [1] "get"     "can"      "much"     "like"     "just"     "better" "price"   "people" "think"    "hotels"
Topic 6: Dining Experience
 [1] "service"    "lovely"     "bar"        "food"       "restaurant" "weekend"    "dinner"     "well"
 [9] "time"       "view"
Topic 7: Check-in and Check-out Process
 [1] "room"       "check"     "back"       "day"        "arrived"    "went"       "reception" "asked"
 [9] "told"       "got"
Topic 8: Room Quality
 [1] "room"     "floor"    "door"     "noise"    "night"    "problem" "next"     "front"    "bed"      "outside"
Topic 9: Noise Concerns
 [1] "room"     "free"     "hotel"    "service" "wifi"     "access"   "use"      "also"     "lounge"   "desk"
Topic 10: Breakfast Quality
 [1] "breakfast" "tea"        "coffee"     "hot"        "fresh"      "english"    "afternoon" "cold"
 [9] "buffet"     "selection"
Topic 11: London Experience
 [1] "london"     "clean"      "value"      "family"     "central"    "inn"        "place"      "premier"
 [9] "money"      "breakfast"
Topic 12: Hotel Cleanliness
 [1] "london"    "hotel"     "many"      "hotels"    "business" "one"       "place"     "stayed"    "last"
[10] "best"
Topic 13: Area Surroundings
 [1] "nice"       "hotel"      "area"       "room"       "day"        "just"       "little"     "also"
 [9] "reception" "big"
Topic 14: Transportation
 [1] "station"     "walk"        "tube"        "minutes"     "away"        "restaurants" "minute"
 [8] "close"       "bus"         "underground"
```

```
Topic 16: Night Stay Experience
 [1] "one"      "night"    "stayed"   "room"     "stay"     "two"      "nights"   "booked"   "although"
[10] "going"
Topic 17: Bathroom Amenities
 [1] "bathroom"    "shower"      "bed"         "room"        "large"       "comfortable" "bath"
 [8] "small"       "clean"       "space"
Topic 18: Location Convenience
 [1] "hotel"    "location" "london"   "walking"  "within"   "distance" "modern"   "bridge"   "tower"
[10] "view"
Topic 19: Friendly Staff
 [1] "staff"       "friendly"    "helpful"     "clean"       "comfortable" "excellent"   "stay"
 [8] "definitely"  "recommend"   "extremely"
Topic 20: Hotel Comfort
 [1] "hotel"   "well"    "small"   "park"    "quiet"   "located" "rooms"   "street"  "london"  "close"
```

Here we can observe that the positive topics consisting of Hotel experience, Service quality and Staff's interaction towards the customers.

Analysis of Factors Affecting Dissatisfaction

```
Topic 1: Service Quality
 [1] "service"  "people"   "never"    "hotel"    "dont"     "open"     "internet" "someone"  "need"
[10] "windows"
Topic 2: Disliked Features
 [1] "like"    "place"  "really" "better" "can"    "stay"   "think"  "much"   "pay"    "first"
Topic 3: Room Conditions
 [1] "rooms"  "one"    "work"   "time"   "though" "front"  "even"   "best"   "get"    "days"
Topic 4: Staff Issues
 [1] "staff"     "water"    "cold"     "breakfast" "long"     "let"      "see"      "sink"
 [9] "reviews"   "evening"
Topic 5: Amenities Feedback
 [1] "tea"        "bar"        "get"        "given"      "said"       "also"       "minutes"    "bit"
 [9] "experience" "thought"
Topic 6: Bed and Bathroom Concerns
 [1] "bed"         "bathroom"    "shower"      "day"         "found"       "next"        "wall"
 [8] "quite"       "comfortable" "keep"
Topic 7: Night Stay Experience
 [1] "room"   "nights" "well"   "two"    "tiny"   "one"    "get"    "small"  "money"  "say"
Topic 8: Room Quality
 [1] "room"      "hotel"     "reception" "hot"       "will"      "rooms"     "time"      "poor"
 [9] "desk"      "booking"
Topic 9: Reception Concerns
 [1] "room"    "back"    "night"   "told"    "another" "asked"   "check"   "stay"    "went"    "away"
Topic 10: Hotel Experience
 [1] "hotel"     "staff"     "lobby"     "guests"    "night"     "room"      "paid"      "thing"
 [9] "available" "least"
Topic 11: Room Accessibility
 [1] "room"  "door"  "staff" "beds"  "night" "right" "still" "star"  "nice"  "also"
Topic 12: Breakfast Quality
 [1] "breakfast" "good"      "location"  "clean"     "small"     "stayed"    "morning"   "bad"
 [9] "left"      "give"
Topic 13: Location Feedback
 [1] "hotel"    "london"   "great"    "stayed"   "helpful"  "close"    "around"   "hotels"   "station"
[10] "friendly"
Topic 14: General Dissatisfaction
 [1] "hotel"  "just"   "booked" "stay"   "didnt"  "even"   "got"    "floor"  "felt"   "double"
```

Here we can observe that the positive topics consisting of Service quality, Disliked features, Staff Issues.

Limitations

- The dataset exhibited a significant skew towards positive reviews, resulting in an imbalance between positive and negative sentiments. This disparity may have impacted the model's ability to accurately discern negative feedback.
- During the data splitting process, positive and negative reviews were inadvertently combined, potentially introducing bias into the analysis. This mixing could have influenced the training and evaluation phases of the models.
- The scarcity of negative review data posed challenges in effectively evaluating the performance of sentiment analysis models.