Leeds University Business School

UNIVERSITY OF LEEDS

Assessed Coursework Coversheet

For use with *individual* assessed work

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).

Part 1

The given data contains information about the monthly personal consumption expenditures (PCE) in the US from January 1959 to November 2023. This section of the report seeks to identify the best performing forecasting model for the given data as it helps in providing insights on the economic trend and anticipated fluctuations. The report will tackle following objectives:

- Compare the forecasting ability of simple forecasting model, exponential smoothing model and ARIMA model and suggest the best model based on the results.
- Using the best model from the first task to predict the PCE for October 2024.
- Using one step ahead rolling forecasting without re-estimation to compare the three models.

Before proceeding to test the forecasting abilities of the models it is essential to perform the data preparation steps. This is started by installing and loading necessary libraries, and the dataset into R. To test the forecasting models the input needs to be a structured sequence where each data point is associated with particular time. Hence, data is converted into time series with frequency equal to 12 since the data has monthly records. Once the data is converted to time series, a check for missing data is conducted as it is important to handle the missing data to have a reliable forecasting model. It is found that there are 779 data points, 43 out of which are missing values in the dataset. Imputation is carried out using the interpolation function to address the missing data. Interpolation function performs a linear interpolation by replacing the missing value with the mean of the data values before and after it.

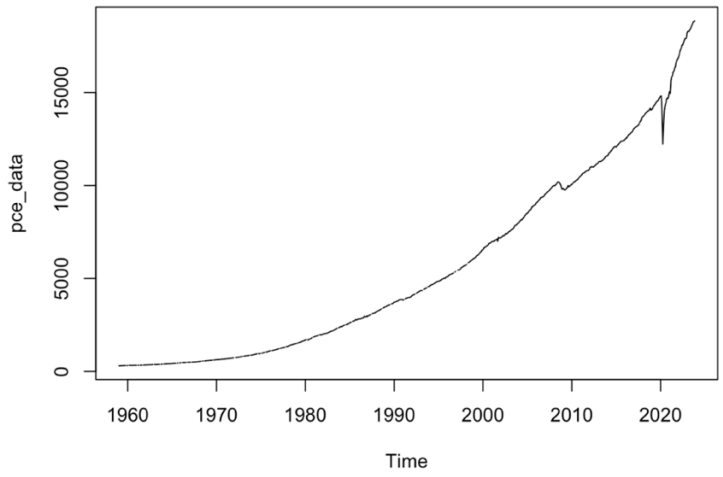

Figure 1.1: Plot of data time series

Although it is already known that the data is seasonally adjusted and from Figure1.1 it can be observed that the data has a clear trend, a check for seasonality is conducted to ensure the same. From the Figure 1.2 shown below it is confirmed that there is no seasonality in the dataset.

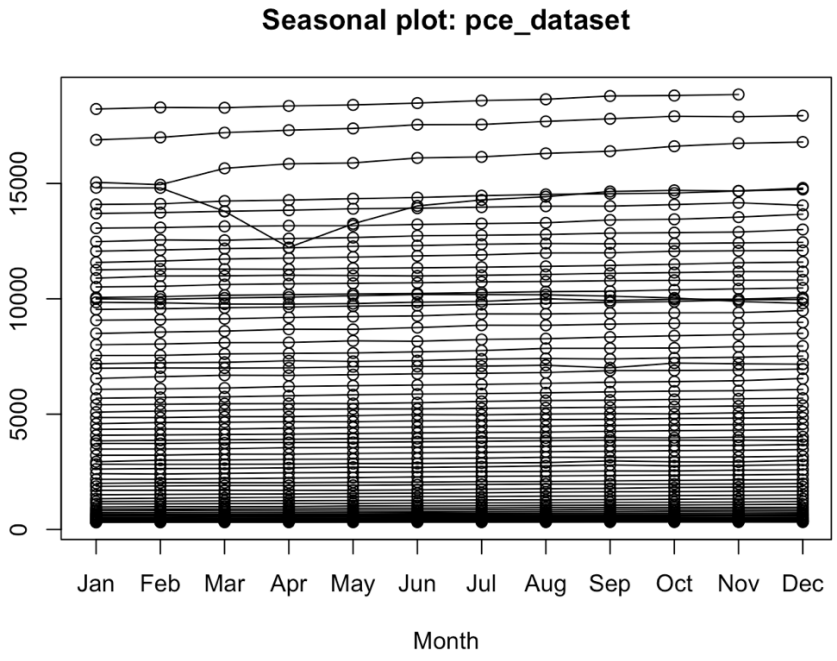

Figure 1.2: Seasonal plot of the data

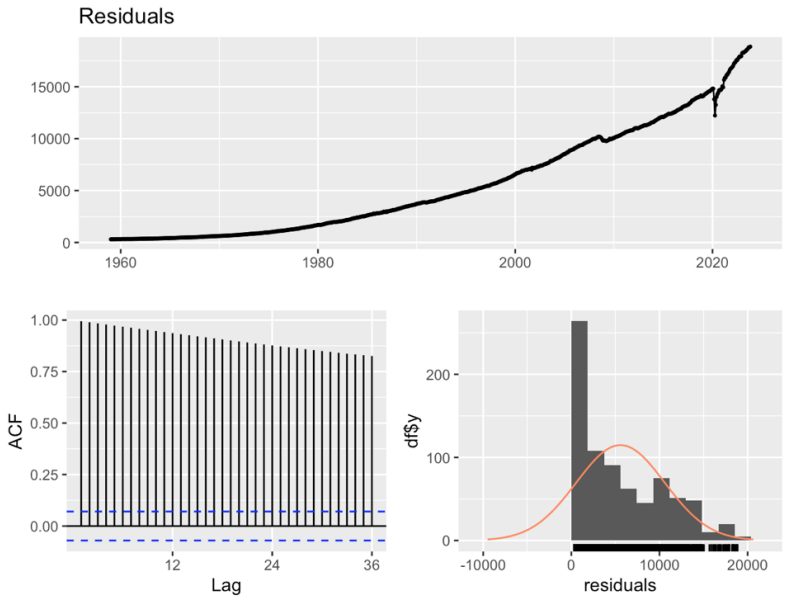

Figure 1.3: Residuals plot

From the residuals plot above we can see that the data exhibits a clear trend as the ACF plot decreases exponentially. Additive decomposition is performed to analyse if there is any anomaly or irregularity in the data to ensure better forecasting. From the figure below it is observed that there are no irregularities in the data and the seasonal component is constant.

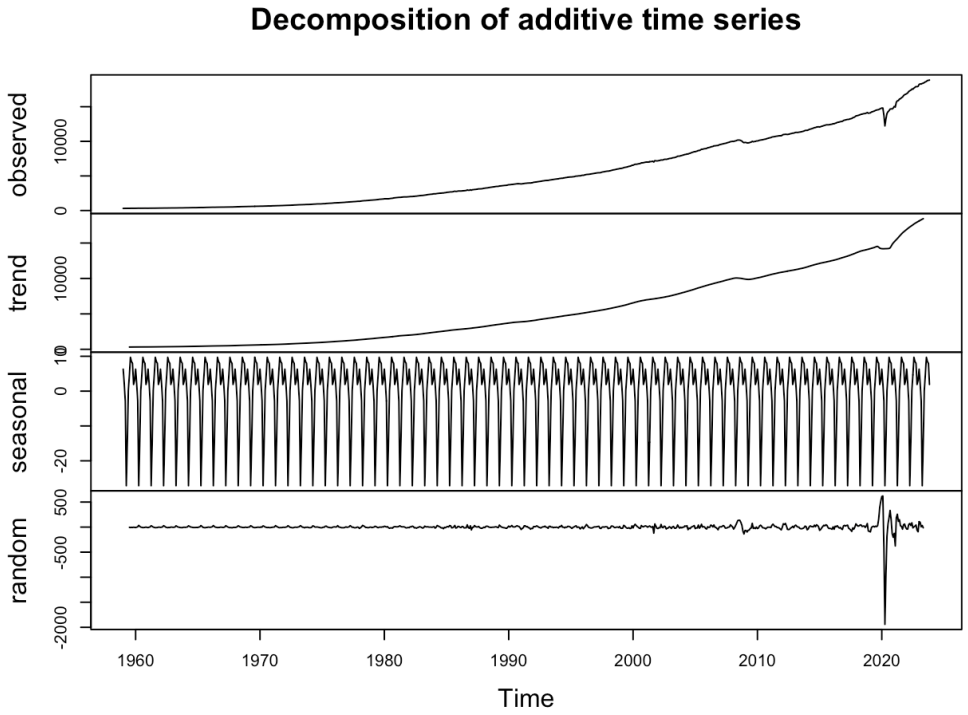

Figure 1.4: Additive decomposition plot

Now that it is clearly established that there is no seasonality, irregularity or missing values, the data can be split into train set and test set. This split is essential because it assists in evaluating the predictive ability of the model. The model is trained on the data using the train set and the test set is used to assess the performance of the model on the unseen data. This assures that the model is analysing the patterns rather than just memorizing and it can now be used to forecast the new data. For the purpose of this analysis the time series of the data is split as 80% (i.e 623 data points) for the train set and 20% (i.e 156 data points) for the test set.

Once the test and the train split is established the next step is to test the predictive ability of the models. Naive model, is the simplest among the four simple forecasting models. This model predicts the future values entirely on the basis of most recently observed value of the time series completely ignoring the other data points. Naive forecast method basically assumes that the projected value will be same as the recently observed value of the time series. After training

the naive model on the train set, summary as shown in Figure 1.4 suggests that mean error (ME) is a positive value indicating that the forecasted values are slightly greater than the actual value. Root mean square error (RMSE) is observed to be 29.69 and mean absolute error (MAE) which corresponds to the deviation of the predicted values from the actual value, is observed to have a value of 19.89. Figure 1.6 shows the plot for the forecast using the train set in naive method.

```
> summary(predict_naive)

Forecast method: Naive method

Model Information:
Call: naive(y = trainset_pce, h = 156)

Residual sd: 29.6977

Error measures:
                    ME     RMSE      MAE       MPE      MAPE       MASE       ACF1
Training set 16.28215 29.69766 19.89598 0.5641837 0.6559345 0.09811548 0.1206498
```

Figure 1.5: Summary of naive model

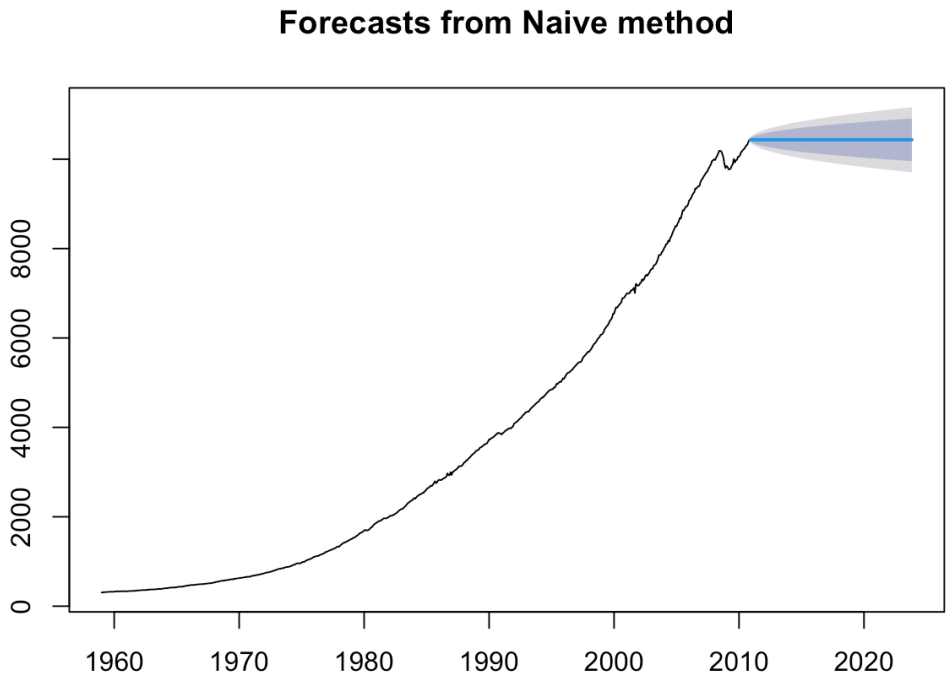

Figure 1.6: Naive forecast plot

In exponential smoothing, the model allocates exponentially reducing weights to past values in which the most recent value gets the highest weightage implying that the recent observation will have a greater impact on the forecast. Exponential smoothing does not account for trend and seasonality in the data and since PCE has the trend component but no seasonality, Holt's method is most appropriate type of exponential smoothing that can be used. Figure 1.7 shows the plot for the forecast using the train set in Holt's method.

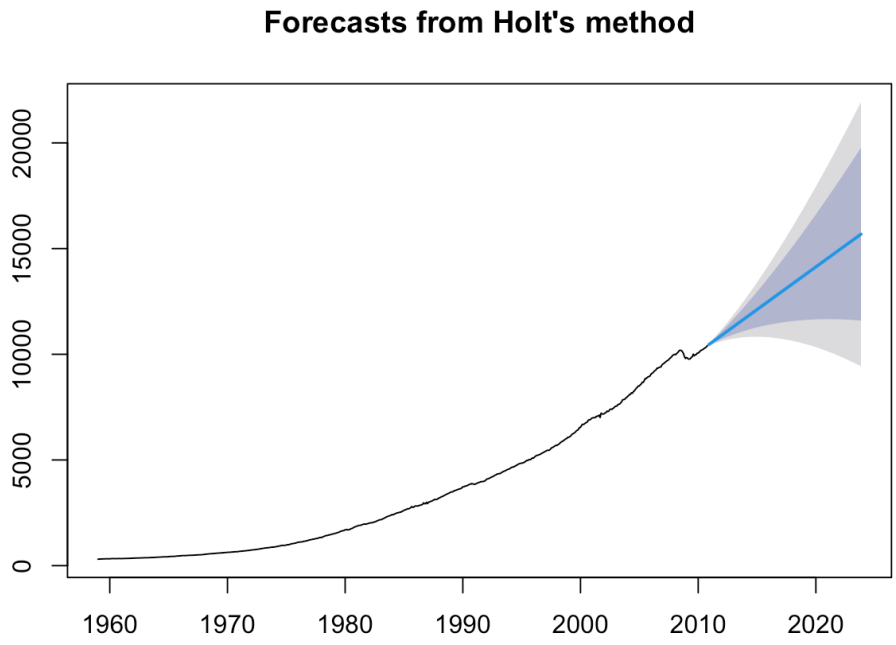

Figure 1.7: Holt's method forecast plot

Autoregressive integrated moving average (ARIMA) model is a forecasting model that predicts future values of a time series using the previously observed values by combining autoregression, difference and moving average components. The autoregressive component analyses the link between the past values and the current value. The moving average component checks for the impact of previous forecast errors and the differencing component removes trends or seasonality and helps in achieving stationarity.

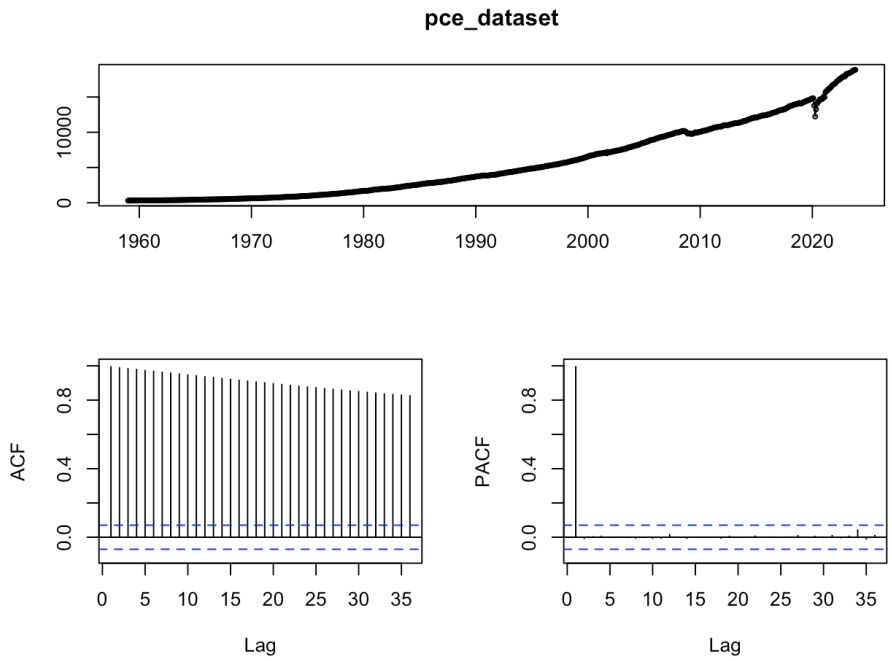

Figure 1.8: ACF and PACF plots

From the figure above it can be noticed that in the ACF plot, values beyond lag 35 are above the significance boundaries which implies that the present value can be predicted from the previous values until lag 35 and it also suggest non-stationarity around the mean. The PACF plot has a strong spike at lag 1 which implies a strong partial autocorrelation between the current observation and just preceding observation suggesting that taking the first difference can assist in removing the trend component.

```
> checkresiduals(arfit_pce)

        Ljung-Box test

data:  Residuals from ARIMA(3,2,2)
Q* = 27.895, df = 19, p-value = 0.08546

Model df: 5.   Total lags used: 24
```

Figure 1.9: Result of ARIMA model residuals check

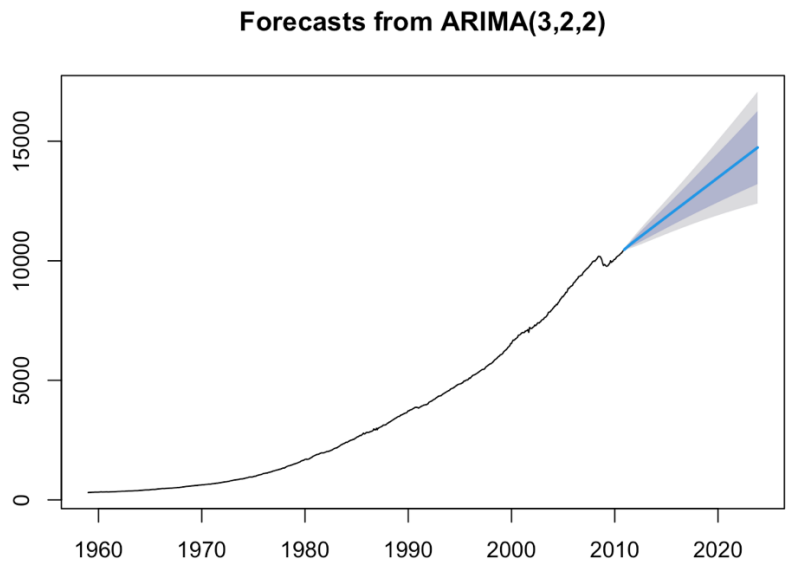

Figure 1.10: ARIMA model forecast plot

The ARIMA model chosen by the auto arima function has 3 autoregressive, 2 differencing and 2 moving average components. From Figure 1.8 it can be said that there is not enough evidence to reject the null hypothesis as the p-value is greater that 0.05. This implies that the residuals of ARIMA(3,2,2) do not have significant autocorrelation. Figure 1.10 shows the plot for the forecast using auto ARIMA function that employs the train set in the model.

Checking the accuracy of all three models using the test and train set, the predictive ability of each model can be evaluated. For the model to be a good fit, error measures have to be considerably less.

```
> #Check accuracy of all models
> accuracy(predict_naive,pce_dataset)
                    ME       RMSE        MAE         MPE       MAPE        MASE      ACF1 Theil's U
Training set   16.28215   29.69766   19.89598  0.5641837  0.6559345  0.09811548 0.1206498        NA
Test set     3204.14551 3976.13173 3204.14551 21.3499291 21.3499291 15.80099418 0.9746013  16.67327
> accuracy(fce_holt,pce_dataset)
                    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
> accuracy(fc_arima,pce_dataset)
                    ME       RMSE        MAE         MPE      MAPE       MASE         ACF1
Training set    1.256645   22.10412   12.36151 0.06713522 0.4029679 0.06095981 -0.005380479
Test set     1021.242570 1593.06514 1042.73242 6.36201334 6.5351073 5.14215375  0.963691956
             Theil's U
Training set        NA
Test set      6.240727
```

Figure 1.11: Accuracy of all three models

From the Figure 1.11 it can be observed that the naive model has the highest value of ME, RMSE and MAE on the test set suggesting that performance of the naive model is not the best. Holt's method and ARIMA model have lower values of MAE, this implies that these models have better accuracy than naive model. Although there is a possibility that all the models are overfitted since the predictive ability is not the best for test set as demonstrated by higher values of error measures. This is because the count of values given for training the model is insufficient and the predictive ability can be improved by providing a larger train set. However, with the current split Holt's model has a lower value of RMSE and ME on the test set in contrast with the ARIMA model thereby suggesting better performance. From accuracy checks and from the plot as shown in Figure 1.12 it can be concluded that Holt's model is the best out of the three models to forecast the personal consumption expenditures.

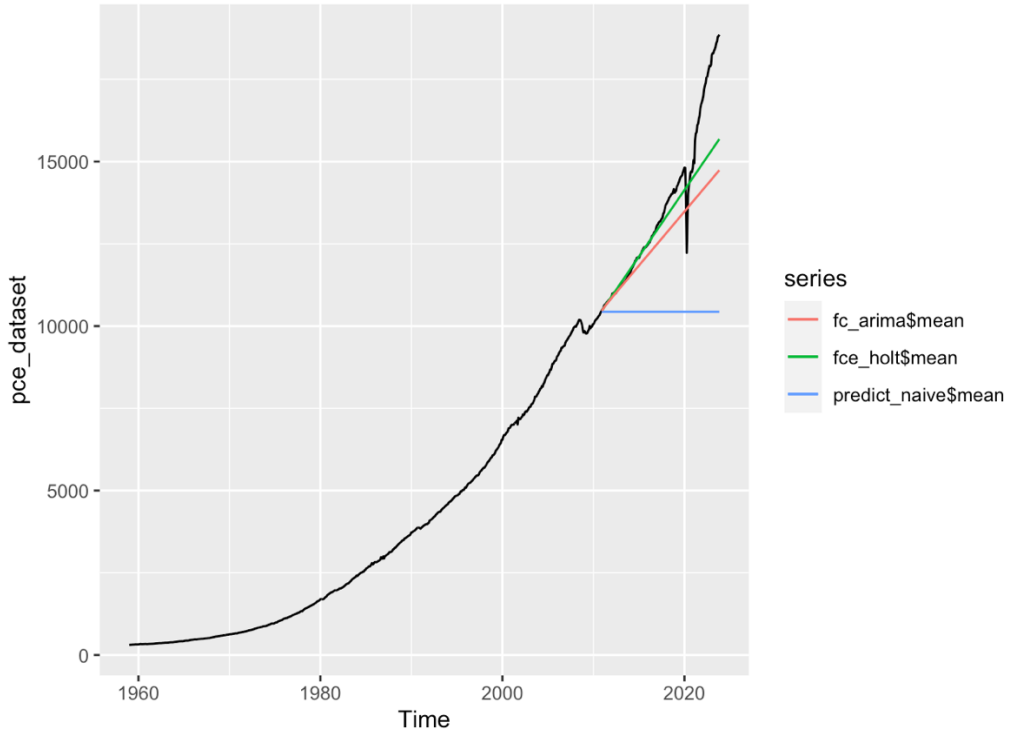

Figure 1.12: Plot of all three models and the main dataset

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
May 2024	15884.48	11569.33	20199.63	9285.035	22483.93
Jun 2024	15918.15	11564.58	20271.72	9259.934	22576.36
Jul 2024	15951.81	11559.70	20343.92	9234.661	22668.96
Aug 2024	15985.48	11554.72	20416.24	9209.215	22761.74
Sep 2024	16019.15	11549.62	20488.67	9183.596	22854.69
Oct 2024	16052.81	11544.41	20561.21	9157.806	22947.82

Figure 1.13: Forecasted value for October 2024

Using the Holt's model to predict the value of personal consumption expenditure for October 2024, the predicted value is 16052.81.

The next task in this analysis is to use one step ahead rolling forecast without re-estimation to compare the naive, Holt's and ARIMA models and determine which has a better performance. Rolling forecasting is a technique in which the model is first trained on a sample data set and then it is used to predict one step ahead data points for the test set which means the forecasting horizon is one (h=1). After the prediction is made for that time point, the available value is now updated into the model to predict the next value. This iterative process keeps repeating until end of the test set. For the purpose of this analysis the window size for the train set is 623 data points which occurs in the year 2010 in the data set and the test set is from 2011 to 2023.

Checking the accuracy of all three models after training the model it is observed that, ARIMA model has the lowest ME value but Holt's and naive model have the lowest RMSE and ACF1 values closer to zero. Although all three models performed well, Holt's model has lesser value of RMSE, ME, MAE and ACF1 thereby making it the best model in the rolling forecast as well.

```
> accuracy(fc_ar,pce_dataset)
                ME     RMSE      MAE         MPE      MAPE      ACF1 Theil's U
Test set 9.404508 221.2915 73.12321 0.05330363 0.5308014 0.2578749  1.082109
> train_holt <- window(pce_dataset,end=2010.99)
> fit_holt <- holt(train_holt)
> refit_holt <- holt(pce_dataset, model=fit_holt)
> fc_holt <- window(fitted(refit_holt), start=2011)
> accuracy(fc_holt,pce_dataset)
                ME     RMSE      MAE       MPE      MAPE      ACF1 Theil's U
Test set 17.03811 200.7651 69.78977 0.1012893 0.5038305 0.1739243 0.9744258
> train_naive <- window(pce_dataset,end=2010.99)
> fit_naive <- naive(train_naive)
> refit_naive <- naive(pce_dataset, model=fit_naive)
> fc_naive <- window(fitted(refit_naive), start=2011)
> accuracy(fc_naive,pce_dataset)
                ME     RMSE      MAE      MPE      MAPE      ACF1 Theil's U
Test set 54.11548 206.9241 92.44065 0.367856 0.6643489 0.1828965         1
```

Figure 1.14: Rolling forecast accuracy for three models

To conclude, in this analysis the forecasting abilities of the three models namely naive, Holt's and ARIMA model were compared to find the best model to forecast the personal consumption expenditure for October 2024. The results showed that Holt's model has the best accuracy and the forecast of PCE for October 2024 is 16052.81. Further, one step ahead rolling forecast confirmed that the Holt's model has the best performance.

Part 2

The data for this section of the report contains 10,000 customer reviews given online for hotels and their corresponding ratings on likert scale ranging from 1(low satisfaction) to 5(high satisfaction). The task outlined is to identify the top 3 key factors that lead to satisfactory or unsatisfactory reviews from the customer. Text analysis with topic modelling will be conducted to analyse the objectives.

After installing and loading all the necessary packages and libraries in R, the data is loaded. Since the data contains reviews in multiple languages it is important to filter the reviews that are only given in English for convenience of this analysis. The filtering is done using textcat function of textcat package. Once the filtering is successful, a test sample of reviews is split into positive and negative reviews. This split is on the basis of ratings given by the customers

and is categorized into 3 or less as negative reviews and 4 or above as positive reviews. To ensure that all the text, special characters and emojis present in the reviews is processed appropriately UTF-8 encoding is performed. Now the text that is converted into UTF is converted into a corpus. Corpus is a collection of documents with text and creating a corpus is usually the first step in topic modelling. Converting the document into corpus helps in cleaning the data and prepares it for the text mining tasks. The next step is text processing which involves operations like tokenization, removing punctuation, converting to lowercase, removing stop words and lemmatization. Tokenization converts text into words or meaningful units called tokens facilitating easier understanding and manipulation of data. Converting text to lowercase prevents considering the words with same spelling but have a dissimilar combination of upper and lowercase letters as different tokens. Removal of stop words and punctuation marks is also essential as they do not carry a lot of meaning. The next step is lemmatization in which the tokens are converted into their root form and similar tokens are grouped together. Consequently, due to the cleaning step it is very likely that all the tokens from certain documents might have been cleared which is why it is important to remove the empty documents from the corpus.

To validate the rating-based classification of reviews and gain additional insights into the emotional content of the reviews, a sentiment analysis was performed using the 'bing' lexicon. The sentiment analysis assigned positive and negative scores to words in each review, allowing to calculate an overall sentiment score for each review. The analysis revealed that reviews with higher ratings (4-5) generally corresponded with positive sentiment scores, while lower ratings (1-3) aligned with negative sentiment scores, confirming the reliability of our classification approach.

For example, in the image below, Review #2 with a rating of 5 had a sentiment score of 10, indicating strong positive sentiment in the text. Similarly, Review #3 with a rating of 4 had a sentiment score of 6, showing moderate positive sentiment. This alignment between numerical ratings and text sentiment provides additional confidence in our classification of positive and negative reviews for topic modeling.

```
> print(head(test))
  Review.score
1            4
2            5
3            4
4            4
5            5
6            4

Text.1
1 I have recently stopped here twice in a month, first time in a twin room and the second in a double room. The double room was tiny but nicely decorated and in in good condition, the twin was bi
gger. The only complaint i would have is the TV's have not been set up correctly and you get weird aspect proportions The position next to Euston Station is tucked away but it is in a relatively
quiet spot. The food and service in Brasserie43 is excellent and the staff are very pleasant and friendly especially Weronike, who remembered me and was very helpful and charming. It is not right
in the heart of the action and probably from the size of the the rooms does not justify its 4 star status, but this is London I guess. I would definitely stop here again.
2
I walked from Kings Cross. 25 Minutes at a steady pace. Right by the British Museum. Doorman was immediately friendly and polite as were the receptionists. Straight up to my. Room which was immac
ulate. I was there for business and I will use it again. Very happy stay and very comfortable. The door man was even more helpful when we running late for dinner. He went off and came back with a
taxi. very impressive. I hope the Bloomsbury stays like this.
3
Arrived at the hotel in the early hours to find no where for us to park. Staff done their best to help but to no avail. . As well as the parking problems we had, there is a tube line that runs by
this hotel which is nosey and causes vibrations in the room so if you are a light sleeper or trying to get some sleep after a long days work I suggest going to another Travelodge. Hotel is lovely
and clean and very easy to find. Staff where pleasant and friendly
4
If you happen to be in London over the weekend it is good value. I stayed in a large and well furnished club room and had cocktails followed dinner and wine for two. The total bill came to about
£170. For London that really is good value. The room had a nice practical layout which was helpful as I needed to get some work done. Internet signal was good. Very nice quality furniture. Clearl
y quite newly done rooms.
5                                                                                                                                                    We had a fantastic wedding reception at the Crown
Moran on 14th April. Our wedding was the best day of our lives and the Crown Moran played a key role in that! As we were planning our wedding, the staff were always friendly and professional, and
the hotel offers great value for money in relation to many other London hotels. On the day the food was fresh and delicious, the room and decor was perfect and the staff were professional yet rel
axed. Our guests commented on how nice the hotel is, the rooms are confortable and welcoming and in the morning the kids went swimming in the pool! Cate and Simon Jolley
6
The room was spacious and the pillows very comfortable. The shower still uses a curtain...I hate shower curtains! Very friendly and helpful staff that helps provide a relaxing atmosphere whether
at breakfast, dining or just a drink at the bar.
     lang review_id sentiment_score
1 english         1               8
2 english         2              10
3 english         3               6
4 english         4               9
5 english         5               9
6 english         6               3
>
```

Figure 2.2: Sentiment Analysis Results

Further, wordcloud can be used to derive insights on frequently encountered words in the corpus. Figure 2.1 shows the wordcloud for positive reviews and it can be observed that words like hotel, room, London, breakfast are prominent which implies that these words appear more frequently in the corpus.

Figure 2.2: Wordcloud for positive reviews

Figure 2.2 shows the wordcloud for negative reviews and it can be noticed that hotel, room, staff, stay are prominent.

Figure 2.3: Wordcloud for negative reviews

The results of sentiment analysis complement the wordcloud visualisation by providing quantitative evidence of the emotional tone in reviews. Reviews with high positive sentiment scores (like Review 2 with a score of 10) frequently contained words that appear prominently in the positive wordcloud, demonstrating the consistency between different analytical approaches.

To find the topics that are discussed in the reviews and to have an understanding of the factors that lead to positive or negative reviews, topic modelling using Latent Dirichlet Allocation is performed. Topic modelling is a technique used to generate a set of topics in the collection of documents. Initially is important to determine the number of topics (k). Different metrics like Griffiths2004, CaoJuan2009 and Arun2010 are used to determine the optimal number of topics that have to be extracted from the corpus. These metrics calculate the probability, frequency and divergence of each word within a topic facilitating easy interpretation of the topics.

From Figure 2.3 and Figure 2.4 it can be said that the optimal number of topics for both positive and negative reviews is between 15-20 topics as the curves seem to gain some stability in that duration. For positive reviews, 18 topics and for negative reviews 19 topics can be a good choice as it can be observed that there is a good compromise between the three metrics at these points.

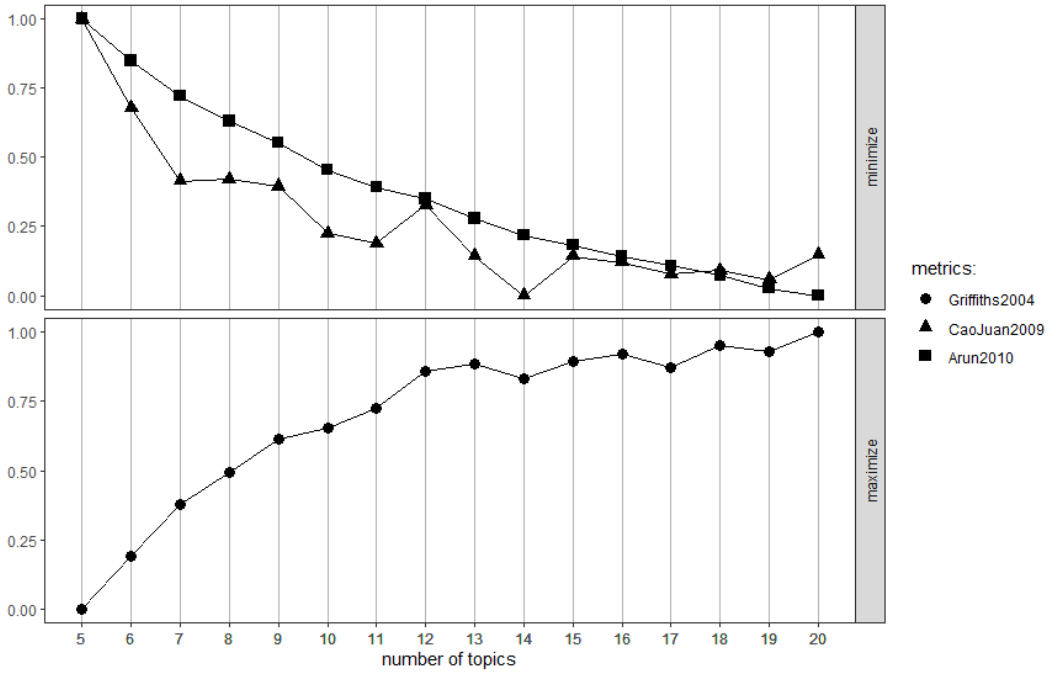

Figure 2.4: Plot of the metrics to find the number of topics for positive reviews

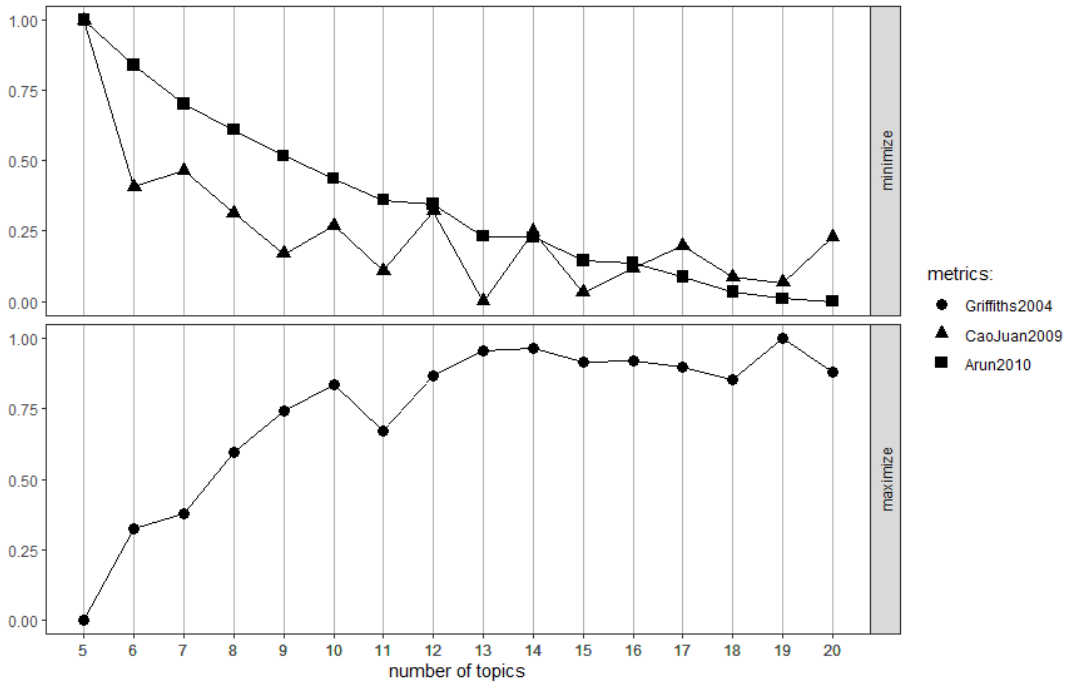

Figure 2.5: Plot of the metrics to find the number of topics for negative reviews

```
> ldaOut.positive
      Topic 1    Topic 2       Topic 3      Topic 4    Topic 5    Topic 6      Topic 7     Topic 8      Topic 9
 [1,] "room"     "well"        "lovely"     "room"     "one"      "staff"      "get"       "time"       "breakfast"
 [2,] "bed"      "rooms"       "will"       "shower"   "stayed"   "made"       "can"       "everything" "hotel"
 [3,] "night"    "small"       "staff"      "water"    "hotels"   "service"    "free"      "stay"       "good"
 [4,] "floor"    "clean"       "definitely" "bathroom" "best"     "experience" "bit"       "really"     "choice"
 [5,] "noise"    "stay"        "fantastic"  "one"      "time"     "stayed"     "price"     "place"      "room"
 [6,] "bathroom" "area"        "wonderful"  "desk"     "standard" "feel"       "wifi"      "much"       "fresh"
 [7,] "didnt"    "good"        "perfect"    "front"    "night"    "visit"      "dont"      "every"      "coffee"
 [8,] "double"   "comfortable" "really"     "bath"     "will"     "even"       "reception" "always"     "couple"
 [9,] "large"    "modern"      "tea"        "day"      "staying"  "without"    "better"    "staff"      "english"
[10,] "quite"    "two"         "amazing"    "found"    "ive"      "welcome"    "want"      "needed"     "fruit"
      Topic 10   Topic 11 Topic 12  Topic 13      Topic 14   Topic 15    Topic 16     Topic 17    Topic 18
 [1,] "hotel"    "hotel"  "street"  "staff"       "room"     "station"   "service"    "hotel"     "great"
 [2,] "london"   "nice"   "walk"    "friendly"    "back"     "tube"      "bar"        "location"  "location"
 [3,] "recommend" "also"  "just"    "helpful"     "went"     "walk"      "excellent"  "close"     "good"
 [4,] "stay"     "room"   "park"    "stay"        "booked"   "quiet"     "food"       "many"      "clean"
 [5,] "trip"     "little" "hotel"   "comfortable" "check"    "road"      "restaurant" "london"    "nights"
 [6,] "big"      "people" "minutes" "clean"       "morning"  "easy"      "quality"    "restaurants" "rooms"
 [7,] "view"     "like"   "around"  "excellent"   "arrived"  "also"      "suite"      "walking"   "value"
 [8,] "central"  "area"   "away"    "inn"         "night"    "london"    "drinks"     "within"    "money"
 [9,] "eye"      "really" "find"    "premier"     "given"    "convenient" "also"      "distance"  "always"
[10,] "highly"   "lots"   "tube"    "especially"  "reception" "minute"    "dinner"     "city"      "family"
```

Figure 2.6: Topics for positive reviews

Figure 2.5 shows the 18 topics and 10 most frequently occurring words under each topic for positive reviews. The themes for these topics are identified and the topics are labeled accordingly. The labels for 18 topics from the positive reviews are given in the table below.

Topic Number	Label
1	Description of the room facilities
2,11	Overall experience in the hotel
3,6,13	Staff and service
4	Room, bathroom amenities
5	Accommodation experience
7	Quality of additional facilities offered
8,18	Overall satisfaction
9	Breakfast options
10,17	Hotel location and recommendation
12,15	Convenience of transportation
14	Check in experience
16	Dining Experience

Table 1: Topic labels for positive reviews

```
> ldaOut.negative
      Topic 1 Topic 2   Topic 3    Topic 4   Topic 5      Topic 6     Topic 7  Topic 8   Topic 9     Topic 10 Topic 11
 [1,] "hotel" "room"    "bathroom" "said"    "stay"       "breakfast" "bar"    "hotel"   "room"      "can"    "one"
 [2,] "even"  "booked"  "really"   "never"   "like"       "small"     "area"   "around"  "hotel"     "wifi"   "night"
 [3,] "time"  "double"  "quite"    "asked"   "london"     "good"      "quiet"  "away"    "reception" "hotels" "room"
 [4,] "day"   "beds"    "helpful"  "morning" "will"       "nights"    "floor"  "stayed"  "paid"      "bed"    "noise"
 [5,] "well"  "two"     "much"     "told"    "just"       "although"  "main"   "staying" "good"      "better" "next"
 [6,] "first" "view"    "though"   "manager" "front"      "even"      "table"  "without" "big"       "also"   "thing"
 [7,] "found" "bed"     "looking"  "wanted"  "things"     "room"      "tired"  "years"   "desk"      "use"    "going"
 [8,] "park"  "window"  "stay"     "another" "experience" "long"      "modern" "enough"  "given"     "free"   "bit"
 [9,] "star"  "wait"    "near"     "service" "bathroom"   "clean"     "walk"   "part"    "door"      "make"   "thought"
[10,] "say"   "outside" "rate"     "check"   "people"     "toast"     "bath"   "work"    "back"      "tiny"   "two"
      Topic 12    Topic 13  Topic 14  Topic 15      Topic 16  Topic 17 Topic 18   Topic 19
 [1,] "good"      "however" "get"     "stay"        "rooms"   "room"   "hotel"    "just"
 [2,] "clean"     "place"   "didnt"   "service"     "staff"   "shower" "station"  "dont"
 [3,] "great"     "many"    "went"    "staff"       "nice"    "hot"    "tube"     "price"
 [4,] "friendly"  "bit"     "back"    "food"        "hotel"   "also"   "london"   "right"
 [5,] "location"  "walls"   "got"     "rooms"       "well"    "water"  "location" "much"
 [6,] "staff"     "hear"    "arrived" "etc"         "minutes" "bed"    "small"    "made"
 [7,] "stayed"    "three"   "sleep"   "comfortable" "access"  "one"    "close"    "pay"
 [8,] "london"    "sure"    "left"    "restaurant"  "free"    "tea"    "walk"     "want"
 [9,] "value"     "tiny"    "night"   "location"    "get"     "door"   "noisy"    "convenient"
[10,] "recommend" "need"    "problem" "weekend"     "floor"   "coffee" "road"     "didnt"
```

Figure 2.7: Topics for negative reviews

Figure 2.6 shows the 19 topics and 10 most frequently occurring words under each topic for negative reviews. The labels for 19 topics from the negative reviews are given in the table below.

Topic Number	Label
1	Hotel amenities
2	Room reservation issues
3	Bathroom cleanliness
4	Interaction with the staff
5	Complete stay experience
6	Breakfast Experience
7	Bar area and atmosphere
8	Surroundings of the hotel
9	Reception desk service
10	Quality of additional facilities offered
11	Disturbance during the night
12,15	Overall Experience

13	Room size
14	Arrival experience
16	Room amenities
17	Bathroom amenities
18	Accessibility of the hotel
19	Value for money paid

Table 2: Topic labels for negative reviews

On observing the topics, some of the positive terms may be found in the topics related to negative reviews. This could be because the customer started the review by pointing out positive aspects and then proceed to highlight the negative aspects. The reviews with rating 3 might have a combination of positive and negative feedbacks throughout the corpus.

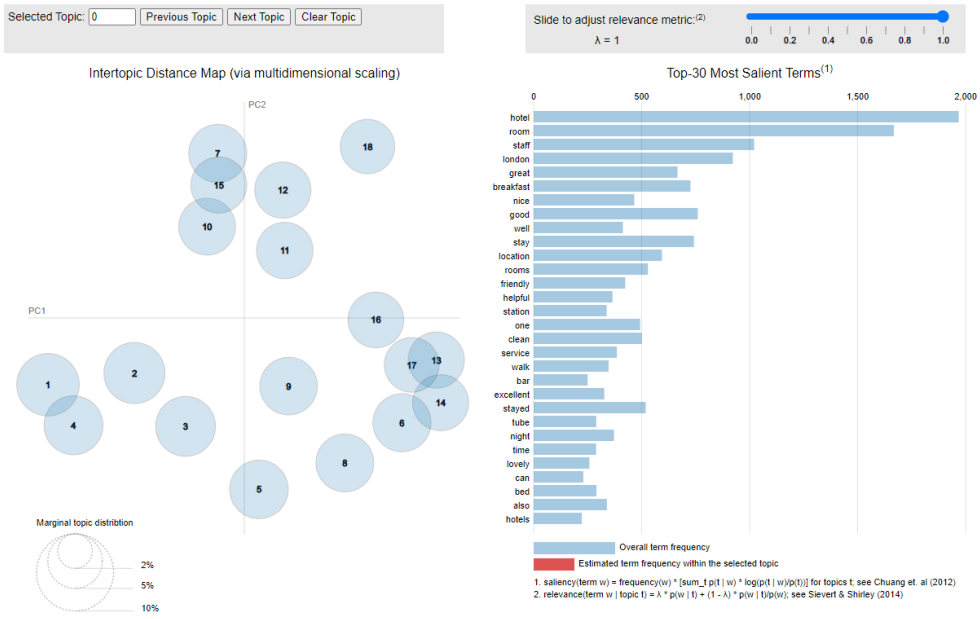

Figure 2.8: Visualisation of topic modelling for positive reviews

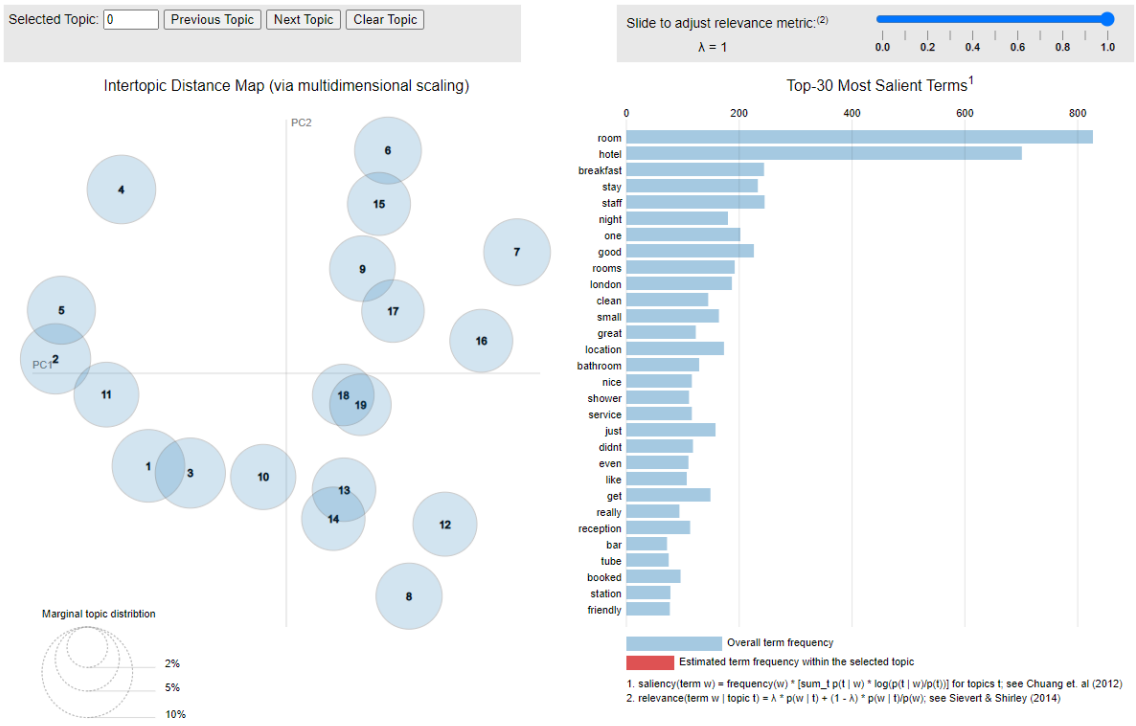

Figure 2.9: Visualisation of topic modelling for negative reviews

The intertopic distance map shows the relationship between the topics. The topics that are clustered demonstrate similarity in the topics and the topics that are far apart are mostly dissimilar to each other. From Figure 2.7 it can be said that hotel, room and staff are three most frequently occurring terms in all the topics for positive reviews. From Figure 2.8 it can be said that room, hotel and breakfast are three most frequently occurring terms in all the topics for negative reviews.

After establishing a clear understanding of topics, to identify the top three factors that affect the satisfaction and dissatisfaction of the customer, probability of occurrence or topic prevalence of each topic is checked. It can be measured by computing the number of documents assigned to each topic with respect to the total number of documents and sorting it in decreasing order. This provides insights on the distribution of topics in the data.

```
> posistive_prevalence
       14         3        13        18        15         1         9         7        17         2        10        11        12         8
0.05754663 0.05724532 0.05704297 0.05650678 0.05644602 0.05626596 0.05590125 0.05580637 0.05559430 0.05555454 0.05547038 0.05517823 0.05510862 0.05472702
       16         5         4         6
0.05470507 0.05395369 0.05353839 0.05340847
```

Figure 2.10: Probability of occurrence of topics for positive reviews

From the figure above it can be observed that the top three topics that lead to customer satisfaction are topic 14, 3 and 13 as they have a higher prevalence score. From Table 1 it can be inferred that topics 14, 3 and 13 highlight check-in experience and staff service are the factors that lead to satisfactory reviews.

```
> negative_prevalence
       18        14        12         2        17         4         9         1         6        15        11         5         3        16
0.05847278 0.05601938 0.05554474 0.05545676 0.05469747 0.05396193 0.05340343 0.05330593 0.05311348 0.05285474 0.05262162 0.05179279 0.05128480 0.05102702
        8        10        19         7        13
0.05068420 0.05043153 0.04985631 0.04830001 0.04717108
```

Figure 2.11: Probability of occurrence of topics for negative reviews

From the figure above it can be observed that the top three topics that lead to customer dissatisfaction are topic 18, 14 and 12. From Table 2 it can be inferred that dissatisfaction stems from accessibility of the hotel, customer's arrival experience at the hotel and overall experience with the hotel offers respectively.

By the means of text analysis and using topic modelling on the data containing online reviews provided by the customers, insights can be drawn on the factors influencing the satisfaction or dissatisfaction. Moreover, the sentiment analysis provides additional validation of our approach to categorizing reviews as positive or negative. The alignment between review ratings and sentiment scores suggests that customers' numerical ratings generally reflect the emotional content of their written reviews, strengthening the reliability of our topic modeling results. By working on factors like customer arrival experience and improving other services offered, the dissatisfaction of the customers can be reduced. Maintaining good staff service can further enhance the overall experience of the customer.

.