Part 1:- Comparing Time Series Models for Personal Consumption Expenditures Forecasting

Introduction:

The study assesses the forecasting performance of three different models using the seasonally-adjusted personal consumption expenditures (PCE) dataset. We evaluate the performance of three models:
Autoregressive Integrated Moving Average (ARIMA).
Exponential smoothing (ETS)
Simple Forecasting (The Drift Method)

Each model uses a unique strategy to discover trends in the PCE data. Whereas ARIMA use iterative procedures to identify autoregressive and moving average components, the Drift Method assumes a constant rate of change, whereas ETS dynamically modifies weights in accordance with history. The goal is to identify the most effective forecasting model for this unique dataset by measuring performance using appropriate error metrics and visual evaluations."

Data Preprocessing:

Several procedures in preparation were done to get the PCE dataset ready for analysis. These stages included resolving missing values, dividing the data into training and testing sets, and ensuring that the date column was properly formatted.

The PCE column was discovered to be missing observations on an initial inspection. To overcome this issue, we used the na_interpolation function from the imputeTS package in R. Using several interpolation approaches, this function estimates and fills in missing values based on surrounding observations. The spline interpolation technique was selected for this investigation since it is well known for managing time series data. .

```
21
22  #missing_values handling
23  missing_values <- sum(is.na(PCE_data$PCE))=
24  if (missing_values > 0) {
25
26    PCE_data$PCE <- na_interpolation(PCE_data$PCE, option = "spline")
27
28  } else {
29    # No missing values, proceed with your existing code
30  }
```

Fig 1: Missing Values Handling code snippet

Following the preprocessing processes, the date data was moved to a new column called Date_POSIX, making it easier to use for time series analysis. By changing the date column's original character format (DATE) to POSIXct format, the representation of dates and times was standardised. This translation guaranteed that dates and timings were represented consistently, allowing for simple incorporation into the analytical framework.

```
# Create a new vector column Date_POSIX and arrange the date
PCE_data$Date_POSIX <- as.POSIXct(PCE_data$DATE, format = "%m/%d/%Y")
```

Fig 2: Creation of standard date column

The dataset was divided into training and testing subsets in order to assess the forecasting models. In accordance with time series forecasting best practices, an 80/20 split was used, meaning that 80 percent of the data were used for model training and the remaining 20 percent were set aside for testing. This partitioning technique allows an unbiased evaluation of model performance while also delivering a large volume of data for robust model training.

Additionally, a line plot depicting the PCE data over time was created using the R ggplot2 tool to help with the visualisation of trends and patterns. The x- and y-axes on this diagram represented PCE levels and corresponding years, respectively. These visual representations improve understanding of time series properties and can affect model selection decisions. Overall, the PCE dataset was methodically prepared for additional analysis by exacting preparation, which included proper date formatting, efficient handling of missing values, and appropriate data division into training and testing groups.

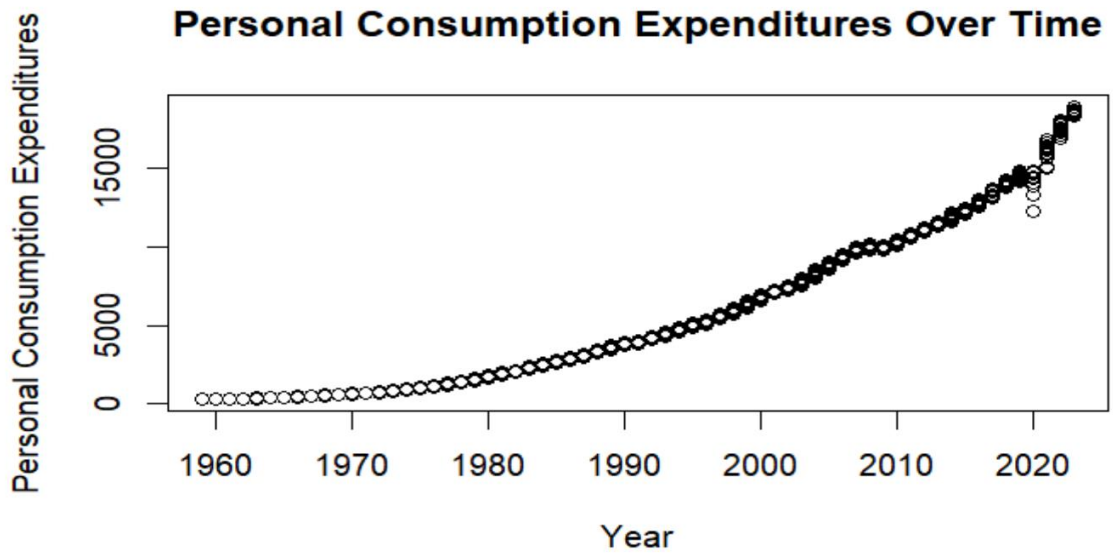

Fig 3

Model Selection and Evaluation:

Because of their wide range of applications and appropriateness for time series forecasting tasks, Simple Forecasting (Drift technique), Exponential Smoothing (ETS), and Autoregressive Integrated Moving Average (ARIMA) were the three forecasting models selected for comparison.

Under the assumption that the time series will change at a constant rate, the Simple Forecasting (Drift) technique predicts future values by averaging the difference between successive observations in the training data. It is easy to use and functions best with datasets that have a linear trend.

By weighting recent data more heavily, the Exponential Smoothing (ETS) model accurately captures trends and seasonal patterns by combining weighted averages of past records. This

model automatically updates the smoothing settings and component components in response to the data.
The Autoregressive Integrated Moving Average (ARIMA) model is a robust and widely used method for forecasting time series data. It combines the autoregressive and moving average components to mimic data autocorrelation. In this analysis, the auto.arima function selects the best ARIMA model based on data properties automatically.

RMSE, or root mean square error, calculates the average squared difference between the expected and actual numbers to provide a general estimate of the error magnitude. Conversely, the average of the absolute discrepancies between the projected and known values, or the Mean Absolute error, or MAE, indicates the normal degree of inaccuracy. In time series forecasting, both metrics are commonly employed to assess model performance. Higher RMSE values imply more inaccuracies, whereas MAE identifies common errors.
The RMSE and MAE values for each of the three models are displayed in the table below:

Description: df [3 × 3]

Model <chr>	RMSE <dbl>	MAE <dbl>
ETS	1725.324	1181.331
ARIMA	1595.695	1045.538
Drift	2545.278	1926.576

3 rows

Fig: 4

According to the results, the ARIMA model outperformed the other two models in forecasting the PCE dataset, with the lowest RMSE and MAE values. The ARIMA model achieved an RMSE of 1595.695 and an MAE of 1045.538, demonstrating its ability to reduce both the average squared difference and the average absolute difference between actual and projected values.
The ETS model likewise did rather well, with an RMSE of 1725.324 and an MAE of 1181.331. These results, while slightly higher than those of the ARIMA model, indicate that the ETS model accurately captured data trends and patterns.
However, as compared to the other models, the Simple Forecasting (Drift) technique performed comparatively worse, as evidenced by its highest RMSE and MAE values. The Drift approach, which assumed a constant rate of change, may not have sufficiently taken into consideration the complexity of the PCE time series, as evidenced by its RMSE of 2545.278 and MAE of 1926.576.
Overall, the ARIMA model emerges as the most suitable choice for forecasting the PCE dataset based on the results.
Model Forecasting and Visualisation:
The plot below visually compares the actual PCE values with the projections generated by each forecasting model, allowing for a qualitative assessment of their performance:

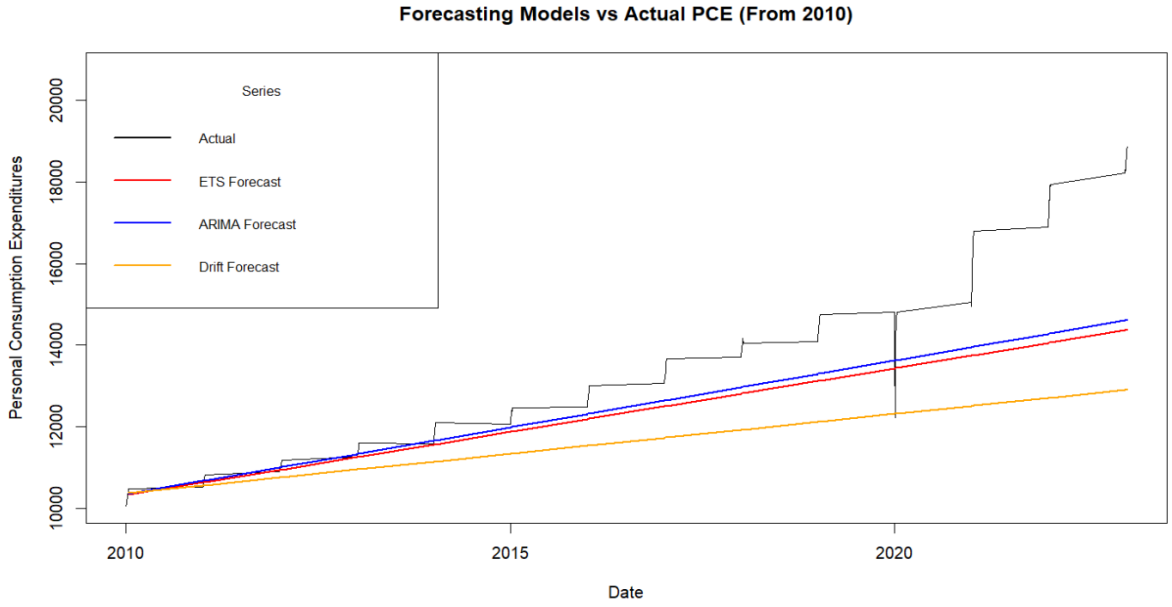

Fig 5 Forecasting Models VS Actual PCE

The graph compares the ARIMA model (green), ETS model (red), and Drift technique (blue) predictions to the actual PCE values (black). Notably, the ARIMA projections closely track PCE trends, particularly in later years, supporting the lower RMSE and MAE values. While the ETS model has commendable performance, the ARIMA model stands out both visually and numerically.

The limits of presuming a constant rate of change are shown by the difference between the Drift method's output and the actual values, which implies the Drift approach performs poorer than ARIMA. The greater RMSE and MAE values for Drift highlight its poor performance, with considerable disparities seen between projected and actual values, particularly in latter parts of the time series. This mismatch implies that the Drift approach does not sufficiently reflect the PCE dataset's complicated dynamics.

On the other hand, the ARIMA model performs better, as seen by its strong numerical and graphic correlation with the real data. This shows that ARIMA is better at capturing underlying patterns and dynamics in the PCE dataset than the other forecasting models.

Forecast for October 2024:

The ARIMA model, which was chosen as the most effective, predicted a PCE value of 19665.85 for October 2024. This forecast was created by using the forecast function with an anticipation of 11 steps forward, which corresponds to the time leading up to October 2024 from the previous recorded data point.

Static Forecasting Comparison:

When using the static approach to forecasting, assumptions are established for a predetermined goal period without the model parameters being updated or re-estimated throughout time. This static approach was used to assess the performance of each forecasting model in this study. The predicted PCE values were compared to the actual data over the target time, and the root mean squared error (RMSE) was calculated to determine the accuracy of each model's predictions. Based on the RMSE values collected, ARIMA was shown to be the highest performing model for forecasting PCE, with an RMSE of

1595.69520964861. The ETS and Drift models had RMSE values of 1725.32430933037 and 2545.27789857663, respectively.

```
Static Forecasting Comparison:
  Best Performing Model:ARIMA
  RMSE (ETS):1725.32430933037
  RMSE (ARIMA):1595.69520964861
  RMSE (Drift):2545.27789857663
```

Conclusion:

To sum up, we assessed three different forecasting models using the PCE dataset: Drift, ETS, and ARIMA. Initially, while using the Static Forecasting technique, the ARIMA model performed better in terms of RMSE and MAE. Long-term trends and complex patterns are particularly well-captured by ARIMA.

Part 2:- Topic Modelling Analysis on Hotel Reviews Data

Introduction:

Reviewers' perceptions and decision-making processes are greatly influenced by customer reviews in the hotel sector. They provide essential insights into the elements influencing guest happiness and discontent. Using advanced text mining techniques, such as topic modelling, is a great way to find underlying trends in unstructured text data, such online reviews. Our research will use topic modelling on the "HotelData.csv" dataset, which includes internet reviews and ratings, to identify the major issues driving visitor happiness and discontent. In order to provide useful insights for hotel management and service improvement activities, we aim to uncover major themes found in both good and negative evaluations by analysing a sample of 2,000 reviews.

Data Preprocessing and Sampling:

In order to provide a representative and consistent sample, the study used the sample_n() function from the dplyr package in R to randomly select 2,000 reviews from the "HotelData.csv" dataset. To assure consistency, a seed value of 205, calculated from the final three digits of the student ID, was used. This method ensures consistent and repeatable outcomes by employing the same set of reviews for each analysis.

Before topic modelling, the text data was cleaned and prepared for review. To begin, the reviews were tokenized to separate the information into distinct terms. Punctuation marks were then removed using the gsub() function because they had no semantic relevance. Furthermore, common terms like "the," "a," and "is," which don't add anything to the subject material, were filtered out using a bespoke stop words list.

To improve the analysis, stemming and lemmatization techniques were used. Lemmatization converts words into their base or dictionary forms, whereas stemming lowers words to their root form by deleting affixes. These algorithms were implemented with the SnowballC and textstem packages, which enable the combining of relevant phrases to improve topic modelling accuracy.

Document-term matrices (DTMs) representing the frequency of occurrence of each term throughout matching sets of reviews for both positive and negative feelings were created using the preprocessed and cleaned text. These DTMs provided input for the ensuing topic modelling analysis.

Classification of Positive and Negative Reviews:

Reviews were divided into good and negative categories using customer ratings as the foundation. Reviews that received a score of four or five were regarded as positive, suggesting high levels of satisfaction, whereas those that received a score of one or two were categorised as negative, indicating low levels of satisfaction. After preprocessing, 1486 of the initial 2,000 reviews were categorised as good, with 245 classed as unfavourable. This segmentation allowed for separate topic modelling assessments for each group, generating useful insights for improving customer happiness.

Topic Modelling for Positive Reviews:

Positive reviews were analysed using the Latent Dirichlet Allocation (LDA) technique, with the goal of revealing the elements impacting customer happiness. To measure the frequency of terms in the review corpus, a document-term matrix (DTM) was created using pre-processed data. Determining the right number of subjects is critical for assuring model quality and interpretability. Model effectiveness was evaluated using measures such as Griffiths2004, CaoJuan2009, and Arun2010, which helped determine the optimal amount of topics. The ldatuning package's FindTopicsNumber_plot method was used to visualise metric values across different topic numbers, allowing for more informed subject selection decisions.

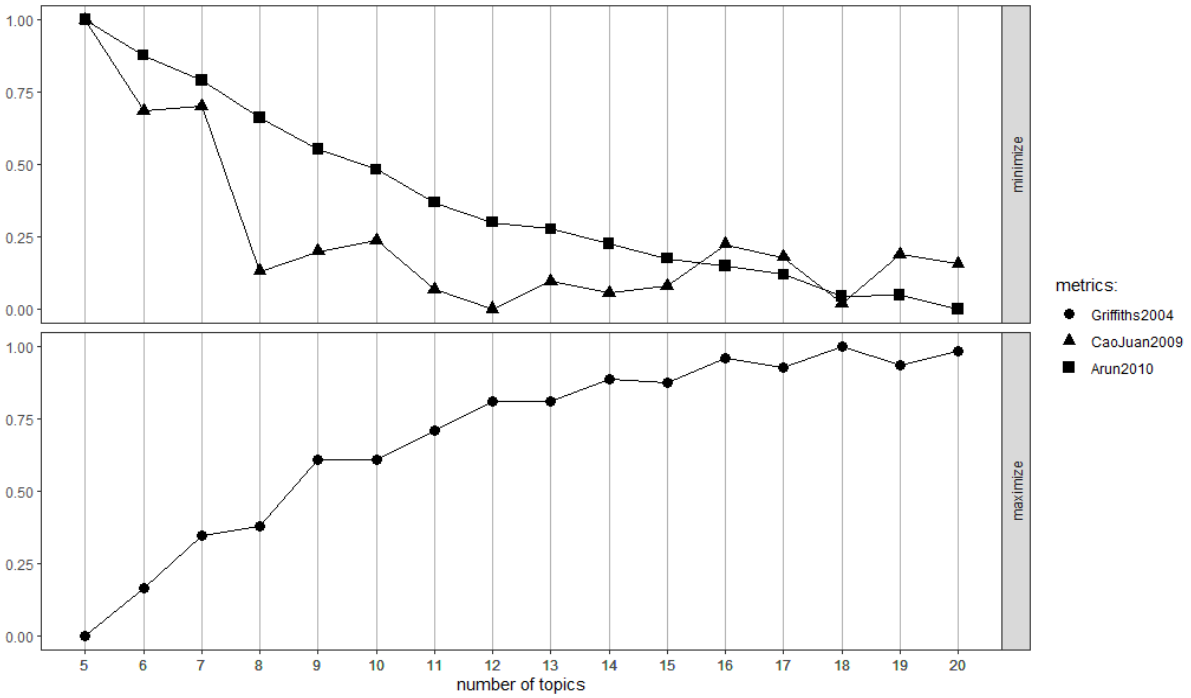

Fig 1: Topic Number Identification plot for Positive Reviews

After a thorough examination of the metrics and visualisations, the best number of themes for the positive review corpus was determined to be 18, as seen in the graph displaying

metric values. The topic-term distributions and document-topic distributions were then obtained by training the LDA model with the document-term matrix (DTM) and the desired number of topics.
Following analysis, it was discovered that the three most talked-about themes in good reviews were themes 7 (134 reviews), 1 (112 reviews), and 3 (109 reviews), highlighting their importance in affecting consumer happiness.
The analysis revealed three primary elements that significantly impacted customer satisfaction:

Room Quality (Topic 1): Positive reviews have often emphasised the value of tidy, comfortable mattresses and well-appointed bathrooms, underscoring the role that room quality plays in guaranteeing patron happiness.

Location and Accessibility (Topic 7): The hotel's well-positioned position in relation to public transportation and its easy access to neighbouring attractions were frequently cited as reasons for positive feedback. This shows that a well-chosen site is critical in improving consumer satisfaction.

Overall Hotel Experience (Topic 3): Excellent feedback was constantly focused on the outstanding overall hotel experience, which included things like the hotel's ambience, events, and the reservation or arrival procedure. This comprehensive experience substantially improves client happiness and loyalty.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
[1,]	"bed"	"breakfast"	"hotel"	"est"	"staff"	"stay"	"walk"
[2,]	"room"	"good"	"love"	"très"	"friend"	"night"	"minut"
[3,]	"bathroom"	"room"	"just"	"pour"	"help"	"good"	"station"
[4,]	"shower"	"free"	"return"	"chambr"	"excel"	"clean"	"street"
[5,]	"floor"	"small"	"view"	"les"	"veri"	"one"	"locat"
[6,]	"need"	"includ"	"wonder"	"une"	"perfect"	"two"	"park"
[7,]	"use"	"hot"	"book"	"petit"	"comfort"	"famili"	"tube"
[8,]	"get"	"coffe"	"special"	"dan"	"everyth"	"travel"	"close"
[9,]	"larg"	"buffet"	"arriv"	"nous"	"definit"	"king"	"just"
[10,]	"size"	"avail"	"recept"	"bien"	"big"	"day"	"restaur"
[11,]	"quit"	"eat"	"suit"	"avec"	"weekend"	"premier"	"away"
[12,]	"nois"	"english"	"birthday"	"light"	"great"	"inn"	"shop"
[13,]	"doubl"	"choic"	"everyth"	"plus"	"locat"	"comfort"	"distanc"
[14,]	"sleep"	"bite"	"husband"	"hôtel"	"breakfast"	"will"	"within"
[15,]	"littl"	"full"	"offer"	"mai"	"high"	"cross"	"underground"
[16,]	"just"	"continent"	"even"	"des"	"earli"	"meal"	"easi"
[17,]	"small"	"bar"	"upgrad"	"qui"	"superb"	"next"	"garden"
[18,]	"onli"	"area"	"beauti"	"tout"	"recommend"	"etc"	"oxford"
[19,]	"window"	"modern"	"surpris"	"londr"	"och"	"bedroom"	"squar"
[20,]	"busi"	"deal"	"extra"	"métro"	"break"	"also"	"high"

	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14
[1,]	"per"	"hotel"	"room"	"hotel"	"veri"	"hotel"	"stay"
[2,]	"non"	"time"	"check"	"que"	"good"	"und"	"much"
[3,]	"con"	"servic"	"get"	"muy"	"nice"	"person"	"realli"
[4,]	"che"	"look"	"can"	"para"	"hotel"	"die"	"make"
[5,]	"una"	"good"	"find"	"con"	"quiet"	"meet"	"hotel"
[6,]	"molto"	"visit"	"day"	"una"	"clean"	"das"	"like"
[7,]	"colazion"	"staff"	"ask"	"las"	"locat"	"sehr"	"can"
[8,]	"sono"	"room"	"say"	"metro"	"price"	"ist"	"feel"
[9,]	"due"	"new"	"book"	"por"	"excel"	"der"	"enjoy"
[10,]	"camera"	"also"	"come"	"del"	"room"	"bad"	"place"
[11,]	"londra"	"last"	"back"	"person"	"comfort"	"zimmer"	"need"
[12,]	"personal"	"review"	"first"	"los"	"restaur"	"war"	"will"
[13,]	"della"	"year"	"arriv"	"londr"	"pleasant"	"taxi"	"mani"
[14,]	"era"	"money"	"desk"	"todo"	"overal"	"super"	"want"
[15,]	"come"	"stay"	"becaus"	"desayuno"	"near"	"man"	"littl"
[16,]	"posizion"	"disappoint"	"work"	"muito"	"friend"	"ein"	"noth"
[17,]	"del"	"london"	"befor"	"habitacion"	"spacious"	"mit"	"high"
[18,]	"questo"	"spend"	"give"	"son"	"close"	"nicht"	"home"
[19,]	"anch"	"dure"	"morn"	"hay"	"also"	"negat"	"sure"
[20,]	"camer"	"expect"	"tell"	"com"	"euston"	"на"	"thing"

	Topic 15	Topic 16	Topic 17	Topic 18
[1,]	"room"	"london"	"great"	"staff"
[2,]	"one"	"hotel"	"london"	"bar"
[3,]	"onli"	"station"	"stay"	"great"
[4,]	"much"	"can"	"locat"	"love"
[5,]	"didnt"	"area"	"hotel"	"room"
[6,]	"stay"	"place"	"recommend"	"food"
[7,]	"pay"	"also"	"central"	"servic"
[8,]	"night"	"tube"	"staff"	"fantast"
[9,]	"problem"	"get"	"good"	"back"
[10,]	"recept"	"reason"	"breakfast"	"tea"
[11,]	"lot"	"line"	"visit"	"restaur"
[12,]	"nice"	"use"	"clean"	"drink"
[13,]	"think"	"stop"	"valu"	"attent"
[14,]	"seem"	"conveni"	"trip"	"will"
[15,]	"howev"	"min"	"welcom"	"comfort"
[16,]	"like"	"access"	"help"	"extrem"
[17,]	"bite"	"bus"	"choic"	"amaz"
[18,]	"small"	"easi"	"spotless"	"lobbi"
[19,]	"tri"	"price"	"busi"	"definit"
[20,]	"enough"	"citi"	"friend"	"modern"

Fig 2: Topics and terms with number of topic matches to respective positive reviews.

Topic Modelling for Negative Reviews:

Topic modelling was employed to analyse negative reviews, replicating the approach used to analyse good reviews, with the goal of discovering characteristics that contribute to consumer discontent. A different document-term matrix (DTM) showing the likelihood of each term occurring in this corpus was created using pre-processed text from the unfavourable reviews. Using the same criteria and the FindTopicsNumber_plot function, the optimal number of topics for the negative review corpus was identified. After careful consideration, 11 subjects were chosen as the best fit for the study of unfavourable reviews.

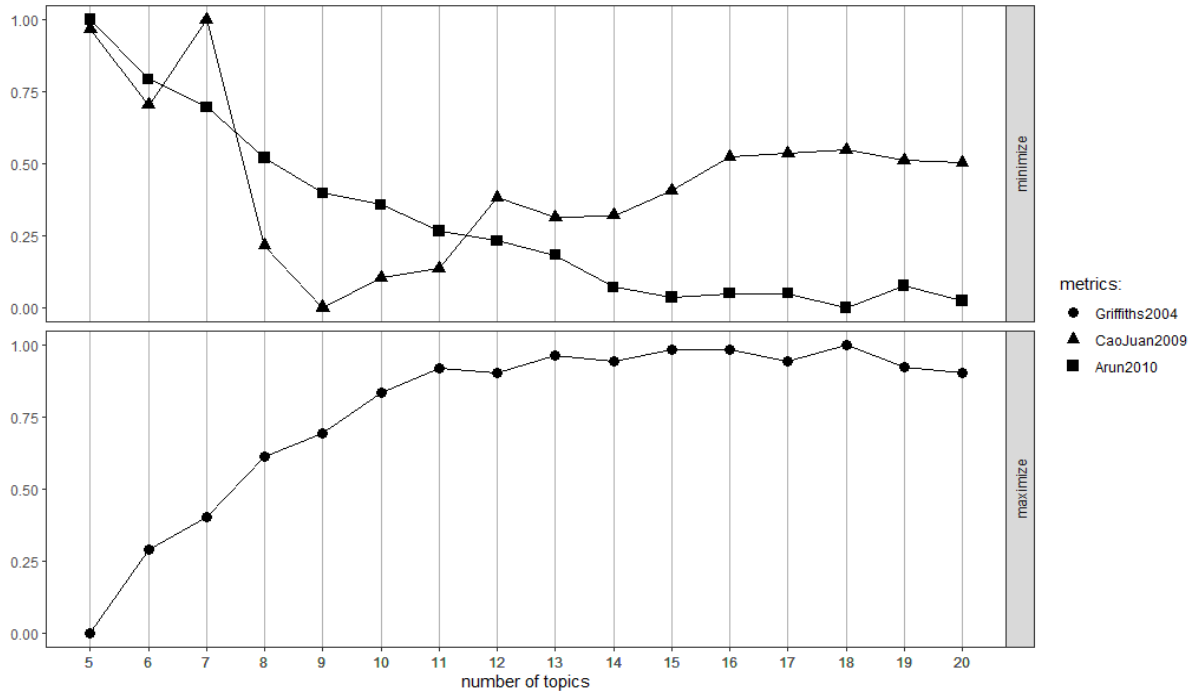

Fig 3: Topic Number Identification plot for Negative Reviews

The specified number of topics (11), which were obtained from the graph's metrics and the DTM of the bad reviews, were then used to train the LDA model, producing distinct document-topic and topic-term distributions that were particular to the negative reviews. The most common themes in unfavourable reviews were determined by analysing the probability and frequency of reviews that covered each subject. Notably, Topics 4 (32 reviews), 2 (30 reviews), and 5 (27 reviews) were the top three most often discussed topics. The primary causes contributing to customer dissatisfaction were as follows:

Room Size and Facility Issues (Topic 4): Negative reviews frequently cited guests' dissatisfaction with the rooms' amenities and spaciousness. Specific comments raised concerns about undersized rooms, cramped bathrooms, and insufficient facilities. Poor reviews highlighted complaints about room size, inadequate showers, and out-of-date décor, compounding general discontent with the hotel experience.

Check-in and Payment Issues (Topic 2) This issue focused on unpleasant experiences with the check-in process, payment procedures, and contacts with hotel workers at reception. Guests were dissatisfied with inefficiencies or delays at check-in, payment processing issues, and communication breakdowns with reception staff.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
[1,]	"one"	"hotel"	"est"	"veri"	"bed"	"que"	"die"
[2,]	"staff"	"check"	"pour"	"good"	"night"	"las"	"man"
[3,]	"say"	"get"	"chambr"	"room"	"breakfast"	"por"	"und"
[4,]	"day"	"recept"	"nous"	"small"	"stay"	"para"	"der"
[5,]	"guest"	"tell"	"très"	"bathroom"	"much"	"con"	"das"
[6,]	"give"	"ask"	"les"	"area"	"room"	"park"	"zimmer"
[7,]	"bar"	"arriv"	"petit"	"shower"	"just"	"com"	"een"
[8,]	"get"	"night"	"une"	"window"	"price"	"porqu"	"person"
[9,]	"back"	"pay"	"des"	"place"	"friend"	"estaba"	"wir"
[10,]	"food"	"tri"	"light"	"floor"	"first"	"muy"	"hilton"
[11,]	"tea"	"can"	"bus"	"door"	"feel"	"на"	"ein"
[12,]	"comfort"	"wait"	"car"	"bad"	"sleep"	"cama"	"ist"
[13,]	"lobbi"	"back"	"dan"	"open"	"didnt"	"pero"	"war"
[14,]	"rude"	"anus"	"avon"	"old"	"think"	"что"	"modern"
[15,]	"becaus"	"even"	"que"	"around"	"look"	"los"	"ich"

	Topic 8	Topic 9	Topic 10	Topic 11
[1,]	"hotel"	"room"	"che"	"hotel"
[2,]	"london"	"can"	"per"	"stay"
[3,]	"veri"	"work"	"non"	"good"
[4,]	"locat"	"onli"	"con"	"like"
[5,]	"also"	"time"	"due"	"will"
[6,]	"nice"	"book"	"una"	"make"
[7,]	"clean"	"want"	"era"	"never"
[8,]	"station"	"even"	"camera"	"room"
[9,]	"great"	"one"	"abbiamo"	"servic"
[10,]	"staff"	"two"	"sono"	"time"
[11,]	"help"	"wasnt"	"del"	"experi"
[12,]	"close"	"take"	"molto"	"need"
[13,]	"recommend"	"thing"	"siamo"	"pay"
[14,]	"walk"	"turn"	"solo"	"use"
[15,]	"find"	"next"	"metro"	"seem"

Fig 4: Topics and terms with number of topic matches to respective negative reviews.

Conclusion:

The topic modelling analysis of 2,000 hotel reviews provided an intuitive understanding of the elements that influence consumer pleasure and discontent. According to the research, significant factors influencing consumer satisfaction include location and accessibility, staff friendliness and service, and room quality. Negative aspects that contributed to consumer discontent, on the other hand, were primarily related to check-in procedures, payment processes, and room quality. These findings help hotel managers discover areas for development in order to increase overall visitor satisfaction and experiences.