Advanced Forecasting and Text Analytics for Business Strategy: PCE Prediction and Hotel Review Insights

Part 1: PCE Prediction

1. Introduction

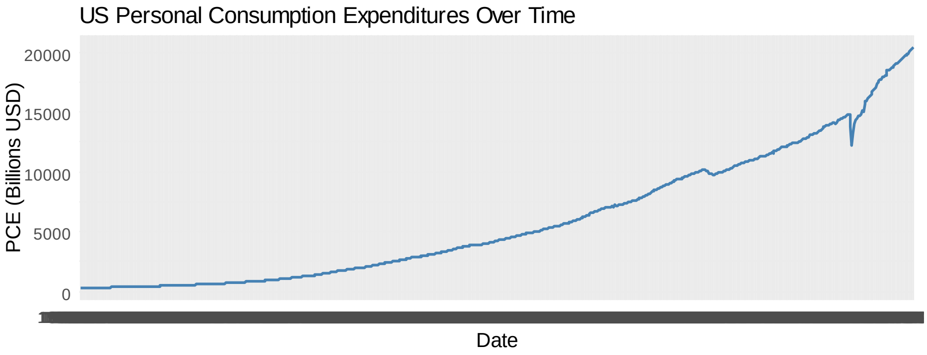
The objective of this report is to assess the performance of multiple forecasting techniques on historical data provided, identify the best model, and use it to predict US Personal Consumption Expenditures (PCE) for the upcoming year to inform business strategy. PCE is a measure of consumer spending on goods and services among households in the US. Forecasting such key indicators is important for insights into economic trends and consumer behavior.

2. Data Understanding and Exploration

The data comes from PCE.csv and contains seasonally adjusted US Personal Consumption Expenditures (PCE) in billions of USD.
The data has 792 observations and 2 variables.
The time ranges from January 1959 to December 2024 with monthly frequency.
There are two key columns: the observation date(character) and the PCE values(numeric).
Below are the summary statistics of PCE values:

Statistic	Value
Min	306.1
Max	20387.2
Mean	5908.7
Median	4212.9
1st Quartile	1116.8
3rd Quartile	9917.4

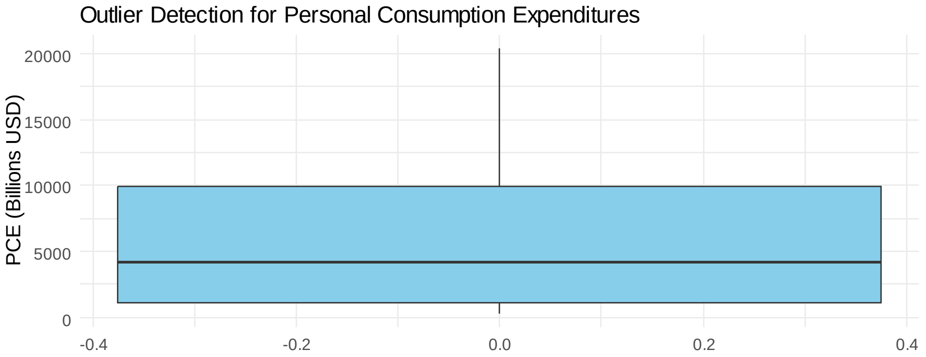
There are 54 missing personal consumption expenditures in the data.



The line plot shows a clear upward trend over time, indicating that the personal consumption expenditure steadily increased, eventually reaching $20 trillion. There are noticeable short-term fluctuations – during 2008 and 2020. This is followed by a sharp increase, indicating that the drop is a temporary decline and could be due to financial crisis and COVID-19.

This series shows trends and subtle fluctuations that indicate seasonality.

The Augmented Dickey-Fuller Test is a statistical test used to check for stationarity in time series. For the given data, the p-value is 0.99> 0.05. Thus, the null hypothesis is accepted and the time series is non-stationary. This suggests that transformations is required for applying certain time-series models.

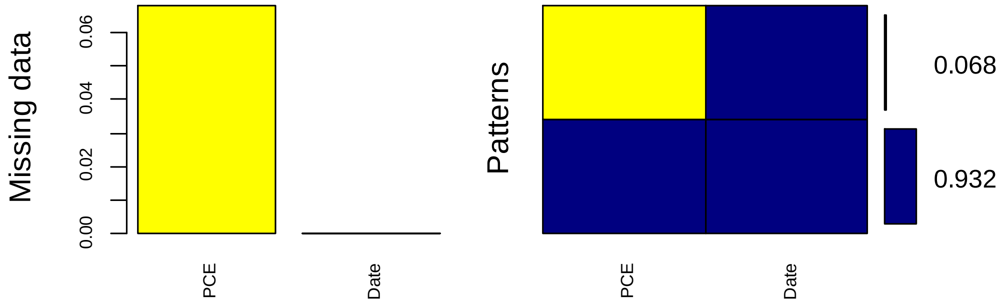


The boxplot shows that the data is a positively skewed distribution. Also, there are no extreme outliers beyond the whiskers. There are no outliers, but the wide interquartile range suggests significant variability in the data.

3. Data Cleaning and Preparation

The *date* column is converted to proper date format and is also chronologically arranged, which is crucial for accurate time series forecasting.

The dataset contains 54 missing values under the PCE column. Since it is a time series, deleting the rows would disrupt the time continuity.



The missing data constitutes 6.8% of the total data.

Only *PCE* values are missing and does not affect the *date* variable.

Linear Interpolation is used to impute the missing data. The method estimates the missing values by the trend of surrounding data points. It is the best choice for economic data, especially for gaps within the time series. It is effective for continuous numeric data and the time structure is also maintained, thus preserving the trend and smoothness of the dataset.

4. Modelling

4.1 Simple Forecasting

Choosing appropriate simple forecasting model depends on a lot of factors like:

- The characteristics of data
- Forecasting horizon
- Simplicity VS Complexity
- Availability of historical data

<u>Characteristics of the data:</u>

Characteristic	
Type	Time series
Variables	Dates and corresponding PCE values
Frequency	Monthly
Trend/Seasonality	Long-term upward trend with seasonality
Stationarity	Non-stationary

Forecasting Horizon:

The forecasting horizon affects the model choice. The objective is to determine PCE values for the next one year(medium-term).

Simplicity VS Complexity

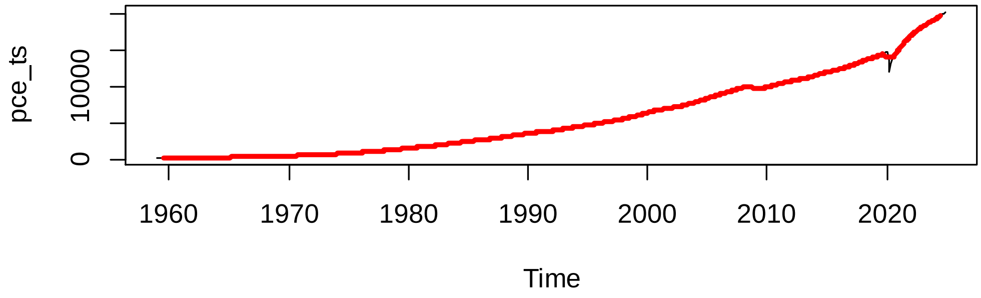
Simple models are easier to understand and apply, and the data shows that the patterns are consistent. Complex models like ARIMA capture intricate patterns.

Availability of Historical Data

There is enough historical data for around six decades which is helpful for training models, better insights and capturing long-term patterns.

Based on the above, there is a need for simple yet effecting forecasting method for non-stationary, monthly time series with consistent patterns and long historical data to forecast for the next one year.

For a baseline model, simple moving average is the best option as the data shows long-term upward trend and seasonality. It smoothens short-term fluctuations and is better for seasonal and slightly trending data. This method is also better suited for monthly economic indicators like PCE. However, it doesn't model seasonality explicitly and lags behind changes in trends.

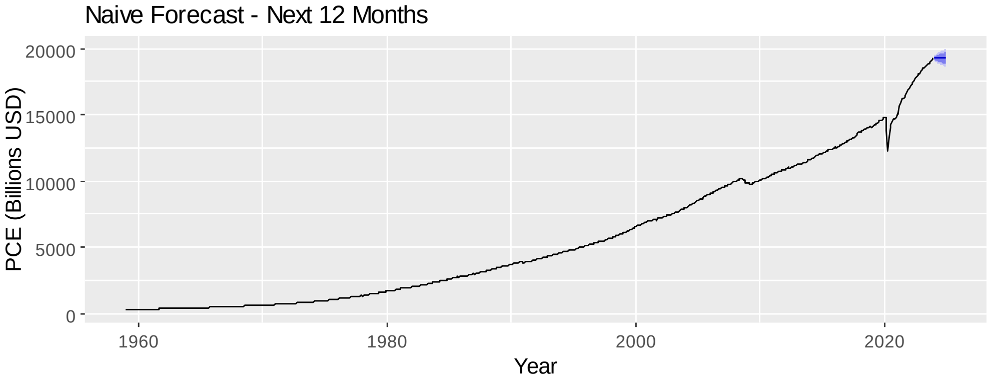


The moving average has smoothened out the short-term fluctuations, making the overall trend clear.

There is a visible sharp drop in 2020 followed by rapid increase, which indicates the economic impact of COVID-19.

The steepest part of the curve in the recent years (2021-2024), suggests post-pandemic recovery and inflation.

Then, naive model is implemented. It serves as a baseline to compare and evaluate against other complex models.



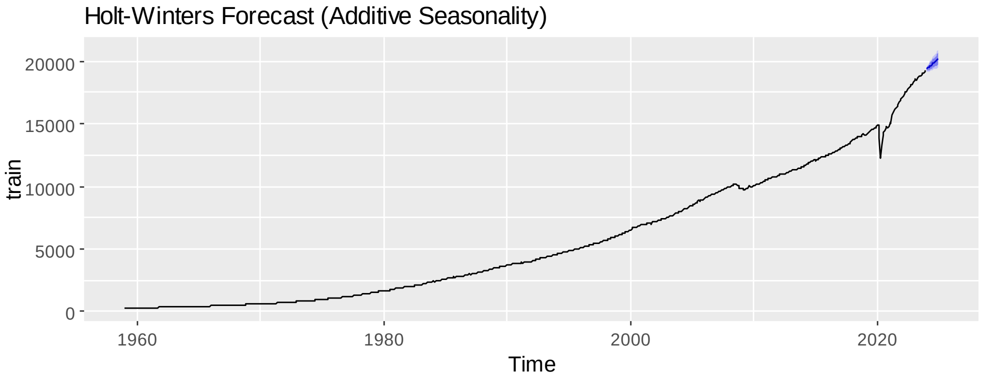
Error Measures:

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	23.44003	96.83551	33.94563	0.5283482	0.6594877	0.1155295	0.2074649
Test set	538.16667	596.93235	538.16667	2.8408973	2.8408973	1.8315788	0.7370880

The blue section shows the forecast using naïve model for next one year. The forecasted value is flat as the model simply projects last known PCE values into the future and continues the constant value.
Overall, the model fits well in the training set. However, it performs poorly in the test set. MASE>1. RMSE and MAE values are very large and also the residual autocorrelation is very high.

4.2 Exponential Smoothing

The Holt-Winters Triple Exponential Smoothing is used as it is suitable for monthly data and handles trend and seasonality very well. It gives more weight to recent years and is more efficient for medium-term forecasts like one year. It remains as a simple yet powerful forecasting tool, especially for the given data.



Error Measures:

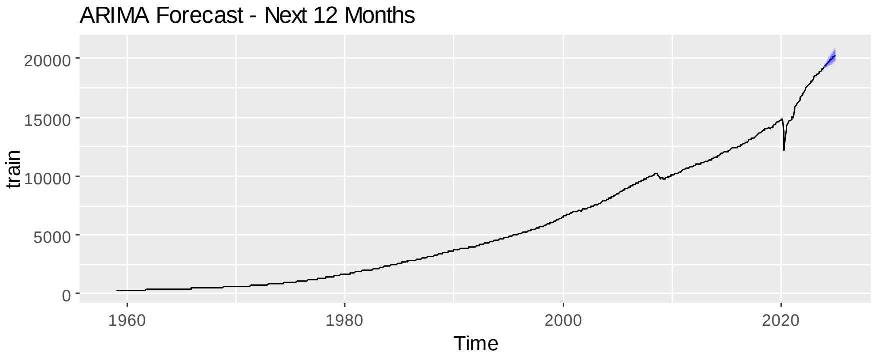
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	5.843335	92.47667	25.31986	0.1385624	0.6772991	0.08617279	0.1735411
Test set	110.318740	118.41070	110.31874	0.5842930	0.5842930	0.37545518	0.1984023

Forecasts increase gradually over year and the seasonal pattern is preserved. The graph aligns with the output, trend is continued upward, unlike the naive model.

Overall, the model performance is better on the training set, but not effectively in the test set. The model seems to be overfit, as it performs better on the training data, but struggles with the test data.

4.3 ARIMA Modelling

ARIMA modelling is powerful for forecasting economic indicators like PCE.



Error Measures:

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	4.941421	87.14500	24.94386	0.1096820	0.4498561	0.08489313	-0.008714626
Test set	20.691135	36.67561	26.14734	0.1104881	0.1393596	0.08898900	0.054887208

The ARIMA model is mathematically sound and there is no danger of unstable or unreliable forecasts. It achieves MAPE of 0.45%, suggesting highly accurate forecasts. The residuals show no significant autocorrelation and can be used for reliable forecasting.

Overall, the model performs better on the test set rather than training set, which is a good sign for generalizability. However, there is still room for improvement.

4.4 Choosing the best model

Model	RMSE	MAE
Naïve	596.93235	538.16667
Holt-Winters	118.41070	110.31874
ARIMA	36.67561	26.14734

5. Forecast Results

The naive model has very high RMSE and MAE. The Holt-Winters model is better, but the ARIMA MODEL performs the best with low RMSE and MAE. The lower ACF1 suggests that it has captured the time series patterns better than the other models. Hence, it will be used for forecasting the PCE values for 2023.

Month	Point Forecast
Jan 2025	20480.48
Feb 2025	20550.44
Mar 2025	20628.58
Apr 2025	20720.72
May 2025	20808.06
Jun 2025	20896.74

Jul 2025	20979.01
Aug 2025	21069.40
Sep 2025	21150.50
Oct 2025	21236.39
Nov 2025	21319.40
Dec 2025	21400.52

6. Conclusion

The report presents analysis of historical data using Naïve Model, Holt-Winters Triple Exponential Smoothing, and Arima Model. Arima Model is identified to make the most accurate predictions and is used to forecast PCE values for the next 12 months.

Part 2: Hotel Review Insights

1. Introduction

The objective of the report is to analyse customer reviews collected by Hotel Insight to identify factors that influence customer satisfaction and dissatisfaction. A random sample of 2000 reviews are selected from the dataset. Word frequency analysis and Latent Dirichlet Allocation (LDA) topic modelling is conducted to provide detailed interpretation of the factors and improve overall customer experience and business performance.

2. Data Understanding and Exploration

The dataset comes from HotelsData.csv and contains 10,000 observations and 2 variables.
The column ‘Text’ is of character and contains qualitative feedback, whereas the column ‘Review Score’ is integer and contains reviews rated on a Likert scale from 1(low satisfaction) to 5(high satisfaction).
Preliminary assessment showed that there are no missing values found in the data and confirmed that the review scores are ranked from 1 to 5.

3. Data Cleaning and Preparation

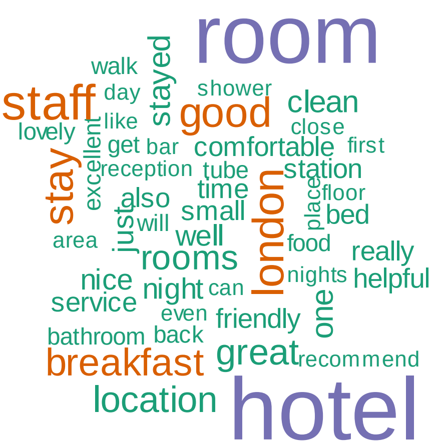
A random sample of 2000 reviews are selected and random seed of 504 is set for reproducibility.
A comprehensive data cleaning is conducted to ensure that raw text is converted to suitable format for analysis and deriving meaningful insights.
Initial exploration showed that the reviews are from 25 different languages:
Therefore, the dataset is filtered for English reviews for streamlined analysis and ensure consistency.
A corpus(collection of text documents) is created for natural language processing.
A custom function *cleandocs()* applied several text pre-processing techniques to the text corpus. The following is done to clean text and prepare them for further analysis:

- Converting text to lowercase – to ensure to get unique terms.
- Removing punctuation and numbers – to simplify text preprocessing and reduce noise.

- Removing stopwords – words that occur frequently but carry less semantic value are eliminated.
- Lemmatizing words – reducing different forms of words to single form.
- Removing extra whitespace.

With this, a clean version of data is obtained to create a document term matrix. This contains frequency of terms that occur in the collection of reviews. Rare words that appear in less than 1% of reviews is removed to focus on important words.

4. WORD FREQUENCY ANALYSIS



Sentiment analysis (also known as opinion mining or emotion AI) refers to natural

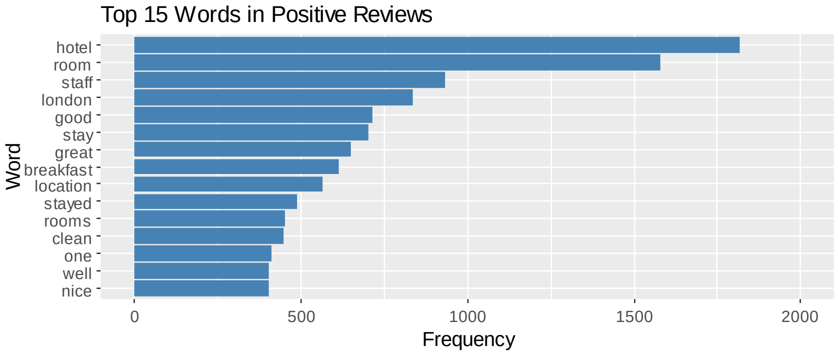
The reviews are categorized according to their score as following:

Review Score	Category
Review Score >= 4	Positive
Review Score = 3	Neutral
Review Score <= 2	Negative

For analysis, neutral reviews are not included as it would introduce further noise in the interpretation.

4.1 Positive Reviews

A text corpus of clean positive reviews and its document term matrix is created.



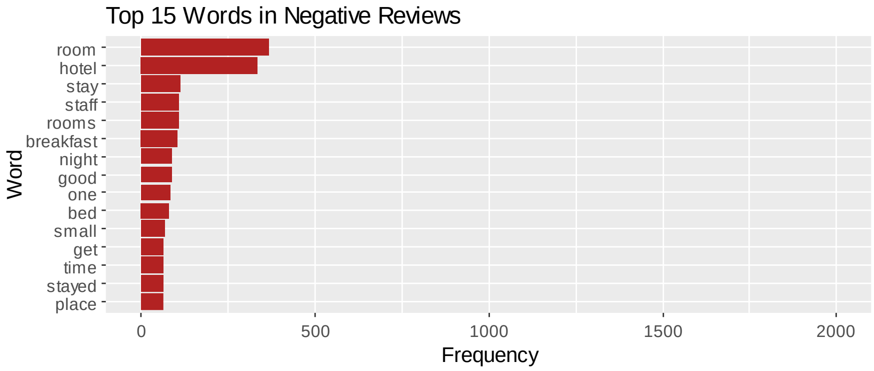
Below are the top 30 more frequent words in positive reviews:

Word	Count	Word	Count	Word	Count
hotel	1815	room	1580	staff	931
london	836	good	713	stay	704
great	651	breakfast	613	location	564
stayed	489	rooms	451	clean	448
one	410	nice	405	friendly	366
service	354	well	333	just	369
comfortable	354	helpful	333	station	315
night	312	really	322	walk	292
excellent	323	also	328	bed	285
tube	270	small	270	will	292

‘Hotel’ and ‘room’ are significantly mentioned in positive reviews than the others. This indicates that people are generally satisfied with the overall hotel and room experience. ‘Staff’, ‘friendly’, and ‘helpful’ suggests that the staff interacts well with the customers. The words ‘london’, ‘location’, and ‘station’ suggests that the location of the hotel is convenient. The ‘breakfast’ in the hotel is good and is often praised by the customers. All the above seem to be primary factors in customer satisfaction.

4.2 Negative Reviews

A text corpus of clean negative reviews and its document term matrix is created.



Below are the top 30 more frequent words in negative reviews:

Word	Count	Word	Count	Word	Count
room	367	hotel	336	staff	108
breakfast	106	stay	113	one	83
night	90	good	87	rooms	107
bed	80	london	62	get	66
Great	57	small	68	just	54
Floor	53	like	51	bathroom	60
Stayed	64	even	58	place	63
Time	65	clean	55	also	48
Service	47	booked	51	nice	57
Shower	51	back	47	back	47

The words ‘room’ and ‘hotel’ occur more in negative reviews. However the frequency is much lesser than positive reviews, which suggests that either customer are not vocable in their feedback with negative experience or they are fewer in number. The word ‘small’ suggests the room size is an issue for customers. The word ‘night’ suggested certain problems that affect customers in night or noise that may affect their sleep.

‘Room’, ‘hotel’, ‘staff’, and ‘breakfast’ occur in both the lists that suggest different perspectives of customers and the need for further in-depth analysis to identify underlying factors.

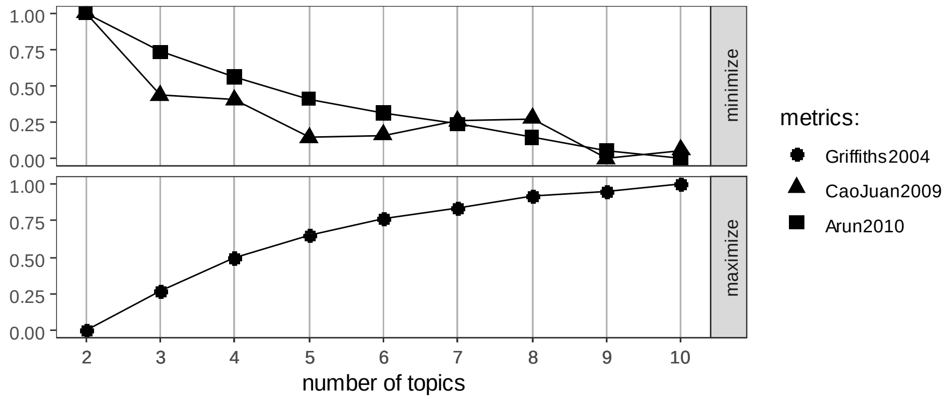
5. TOPIC MODELLING BY LATENT DIRICHLET ALLOCATION (LDA)

.

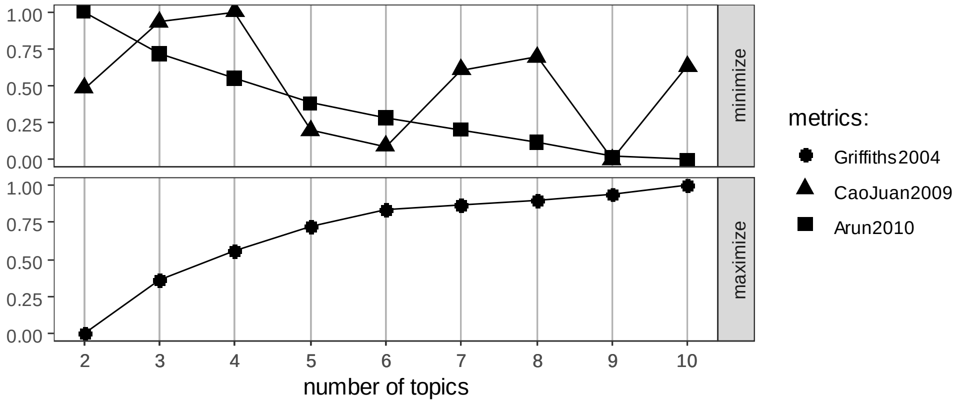
LDA is applied to both positive and negative reviews.

To find optimal number of topics for positive and negative reviews, multiple evaluation metrics are applied.

Positive Reviews:



Negative Reviews:



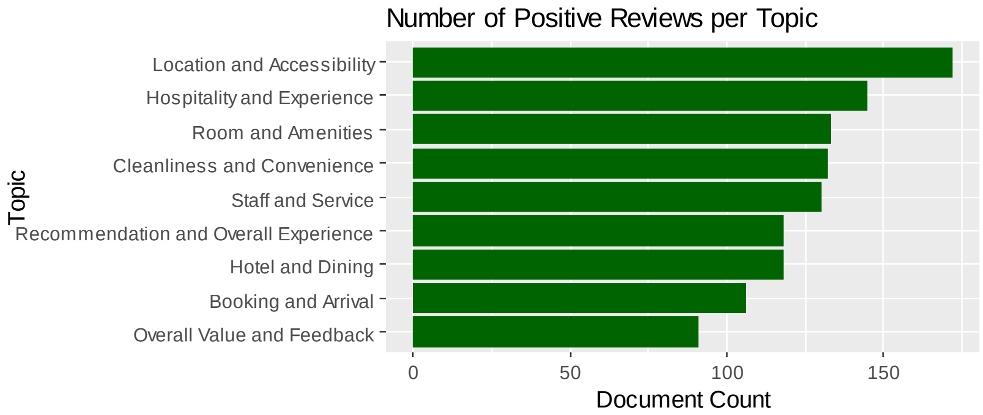
The goal is to minimize the criteria Arun2010 and CaoJuan2009 and maximize the Griffiths2004. Based on this, k=9 is chosen for both the models.

Top 10 positive terms for the topics:

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
room	Well	good	hotel	stay	staff	room	station	Time
bed	Nice	breakfast	london	lovely	great	night	walk	Get
bathroom	Hotel	rooms	stay	staff	stayed	one	tube	much
floor	Also	clean	recommend	everything	location	just	london	Just
shower	service	reception	definitely	always	friendly	booked	close	Like
breakfast	Food	free	will	made	helpful	went	park	place
enough	Bar	small	comfortable	tea	excellent	got	easy	Can
view	Really	central	business	will	clean	didnt	minutes	One
small	restaurant	quiet	trip	wonderful	rooms	arrived	street	Bit
two	Area	nights	found	amazing	breakfast	back	away	Back

The following labels are assigned to the topics manually based on the words under it:

Topic	Label	Reason
Topic 1	Room and Amenities	Terms around room units and features
Topic 2	Hotel and Dining	Terms about broader hotel service and dining experience
Topic 3	Cleanliness and Convenience	Terms about fundamental hotel features and amenities
Topic 4	Recommendation and Overall Experience	Terms about overall evaluation and recommendation
Topic 5	Hospitality and Experience	Abstract terms around service and experience
Topic 6	Staff and Service	Terms about the service quality by staff and location
Topic 7	Booking and Arrival	Terms about the general logistics of booking and arrival in the hotel
Topic 8	Location and Accessibility	Terms that refer to the location and evaluate the accessibility
Topic 9	Overall Value and Feedback	Terms about the overall value assessment and expectations feedback



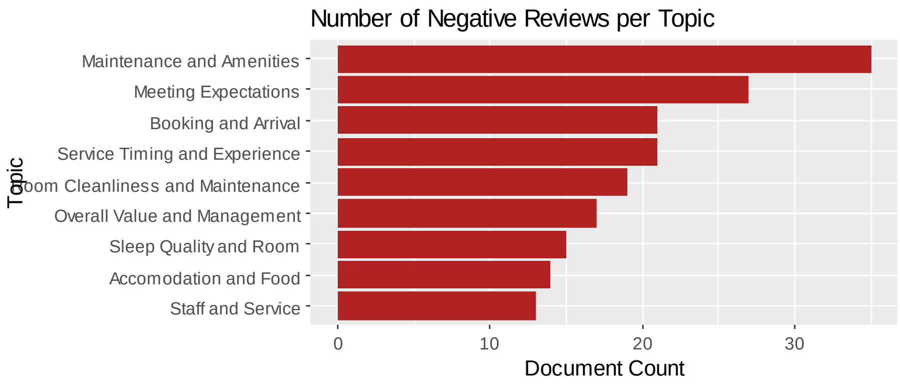
Top 10 negative terms for the topics:

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
bathroom	Room	good	great	night	breakfast	staff	rooms	hotel
like	Time	floor	location	booked	small	hotel	one	stay
bed	However	water	nice	back	place	lobby	clean	asked
area	Much	hot	service	got	bed	close	get	manager
people	experience	better	also	first	nights	put	sleep	london
can	Told	shower	just	didnt	will	find	given	room
really	Minutes	old	well	morning	restaurant	bar	helpful	even
dirty	Wasn’t	day	stayed	dont	wall	quite	extremely	stayed
went	Found	even	friendly	next	made	paid	london	long
door	Ready	stay	work	told	time	another	look	money

The following labels are assigned to the topics manually based on the words under it:

Topic	Label	Reason
Topic 1	Room Cleanliness and Maintenance	Terms around room conditions and maintenance
Topic 2	Service Timing and Experience	Terms about service delays and general dissatisfaction
Topic 3	Maintenance and Amenities	Terms about infrastructure maintenance and its evaluation
Topic 4	Meeting Expectations	Positive and negative terms suggesting flaws in positive aspects of stay
Topic 5	Booking and Arrival	Terms about the general logistics of timing, booking and arrival in the hotel

Topic 6	Accommodation and Food	Terms about room, size, and food quality
Topic 7	Staff and Service	Terms about the customer service quality by staff
Topic 8	Sleep Quality and Room	Terms about sleep disturbance and room allocation
Topic 9	Overall Value and Management	Terms about the overall value assessment and hotel management



6. KEY FACTORS AND INTERPRETATION

After analyzing both positive and negative reviews, the key themes are:

- Room
- Dining
- Amenities
- Staff
- Service
- Maintenance
- Cleanliness
- Booking
- Location
- Sleep Quality
- Overall Value

These themes affect the overall customer satisfaction and dissatisfaction. Below are the key elements in both the positive and negative reviews for a detailed interpretation:

6.1 Key Elements Affecting Customer Satisfaction

- **Room Structure –** ‘bed’, ‘bathroom’, ‘shower’, ‘floor’ suggests that customers are satisfied with the overall structure and condition of different room units.
- **Dining Experience –** ‘food’, ‘restaurant’, ‘bar’, ‘breakfast’ suggests a positive dining experience and good quality of food.
- **Staff and Service –** ‘lovely’, ‘wonderful’, ‘helpful’, ‘friendly’ suggest the staff are capable and experienced in fulfilling the needs of the customers.
- **Booking –** Several adjectives suggest seamless experience in booking and arrival by the customers.
- **Location –** ‘station’, ‘tube’, ‘close’, ‘easy’, ‘minutes’ suggest that the hotel is in a convenient location with proximity to shops and availability of good transportation channels.

6.2 Key Elements Affecting Customer Dissatisfaction

- **Cleanliness and Maintenance –** ‘bathroom’ and ‘dirty’ suggests lack of cleanliness in rooms and ‘small’ suggests complaints about the general size of the room. The words ‘hot’, ‘shower’ suggests lack of basic amenities in room.
- **Service –** ‘experience’, ‘ready’, ‘time’, ‘minutes’ suggest service delays.
- **Sleep Quality –** ‘night’, ‘sleep’ suggests low quality of sleep by customers at night.
- **Value Assessment –** ‘money’ suggests that the customer doesn’t think the overall experience is worth the money paid. Either the fees are huge, or the experience needs to be improved more.

6.3 Strategies for Management

- Maintenance of the rooms should be considered, and efforts need to be made to keep it clean and hygienic for the customers.
- Staff need to be trained more in meeting the needs of customers and without any delays.
- The sleep disturbances experienced by customers at the hotel need to be addressed.
- Breakfast seems to attract neutral reviews, and the offerings need to be looked into.
- Basic amenities like hot showers need to be checked.

7. CONCLUSION

This report presents analysis 2000 English hotel reviews using word frequency analysis and Latent Dirichlet Allocation (LDA) topic modelling. The key themes in reviews include room and amenities, dining, staff and service, maintenance and cleanliness, booking, location, sleep quality, and overall value. Room structure, positive dining experience, good service by staff, seamless booking, and convenient location leads to customer satisfaction. Unhygienic environment, service delays, poor sleep quality, and lower value assessment leads to customer dissatisfaction.

References

1. Robert I. Kabacoff 2015. R in Action : Data analysis and graphics with R. Manning.

2. Foreman, J.W. 2014. Data smart : using data science to transform information into insight. Hoboken, New Jersey: John Wiley & Sons.

3. Hyndman, R.J. and Athanasopoulos, G. 2014. Forecasting : principles and practice. Heathmont, Vic: OTexts.

4. Fortino, A., 2021. Text Analytics for Business Decisions. Business Expert Press.

5. Atkinson-Abutridy, J., 2022. Text Analytics. Springer.

6. U.S. Bureau of Labor Statistics, n.d. Comparing Consumer Expenditure Surveys (CE) and Personal Consumption Expenditures (PCE). [online] Available at: https://www.bls.gov/cex/cecomparison/pce_profile.htm [Accessed 3 April 2025].

7. Santra, R., 2019. Tests for stationarity in time series — Dickey Fuller test & Augmented Dickey Fuller (ADF) test. [online] Medium. Available at: https://medium.com/@ritusantra/tests-for-stationarity-in-time-series-dickey-fuller-test-augmented-dickey-fuller-adf-test-d2e92e214360 [Accessed 3 April 2025].

8. Analytics Vidhya, 2021. Text Preprocessing in NLP with Python Codes. [online] Available at: https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/ [Accessed 16 April 2025].