Final-project-report

Raj Deelip - Raghul Sekar - Mohammad Rafi Shak

12/13/2021

Data analysis on customer data may help the hotel management make informed decisions on staff allocations and create a personalized experience for the customers. It will also help the administration in developing an optimal pricing strategy.

We are using two datasets for our analysis. One dataset contains booking information of the hotels and includes information such as when the booking was made, length of stay, and cancellation status. The second dataset has reviews and ratings of hotels in Europe which can be used for sentiment analysis.

Data Wranglling

1. Eliminate NA and other undefined values

2. Very few booking shows the adults number is larger than 4. So, we regards those booking as abnormal value. Remove unreasonable values that have more than 4 adults in a room.

3. Subset columns that are being used for further analysis

4. Format month from string format to number

```
customerdata <- read.csv("hotel_bookings.csv",na.strings = "")
customerdata <- customerdata[!is.na(customerdata$children), ]
customerdata$children <- as.integer(customerdata$children)
customerdata <- customerdata[!is.na(customerdata$children), ]
customerdata$meal[customerdata$meal=='Undefined'] <- 'SC'
customerdata$children[is.na(customerdata$children)] <- 0
customerdata <- subset(customerdata, market_segment!='Undefined')
customerdata <- subset(customerdata, distribution_channel!='Undefined')
customerdata <- subset(customerdata, adults <= 4)
tempdataset<- customerdata[,c("adr","hotel","lead_time","is_canceled","arrival_date_month","is_repeat
tempdataset["Arrival_month"] <- match(tempdataset[,"arrival_date_month"],month.name)
tempdataset <- tempdataset %>%
  mutate(hotel_type= case_when(hotel == "Resort Hotel" ~ 1,TRUE  ~ 2)) %>%
  mutate(got_desired_roomtype= case_when(reserved_room_type == assigned_room_type ~ 1,TRUE  ~ 0))
tempdataset <- tempdataset[ , -which(names(tempdataset) %in% c("hotel","arrival_date_month","is_repea
```

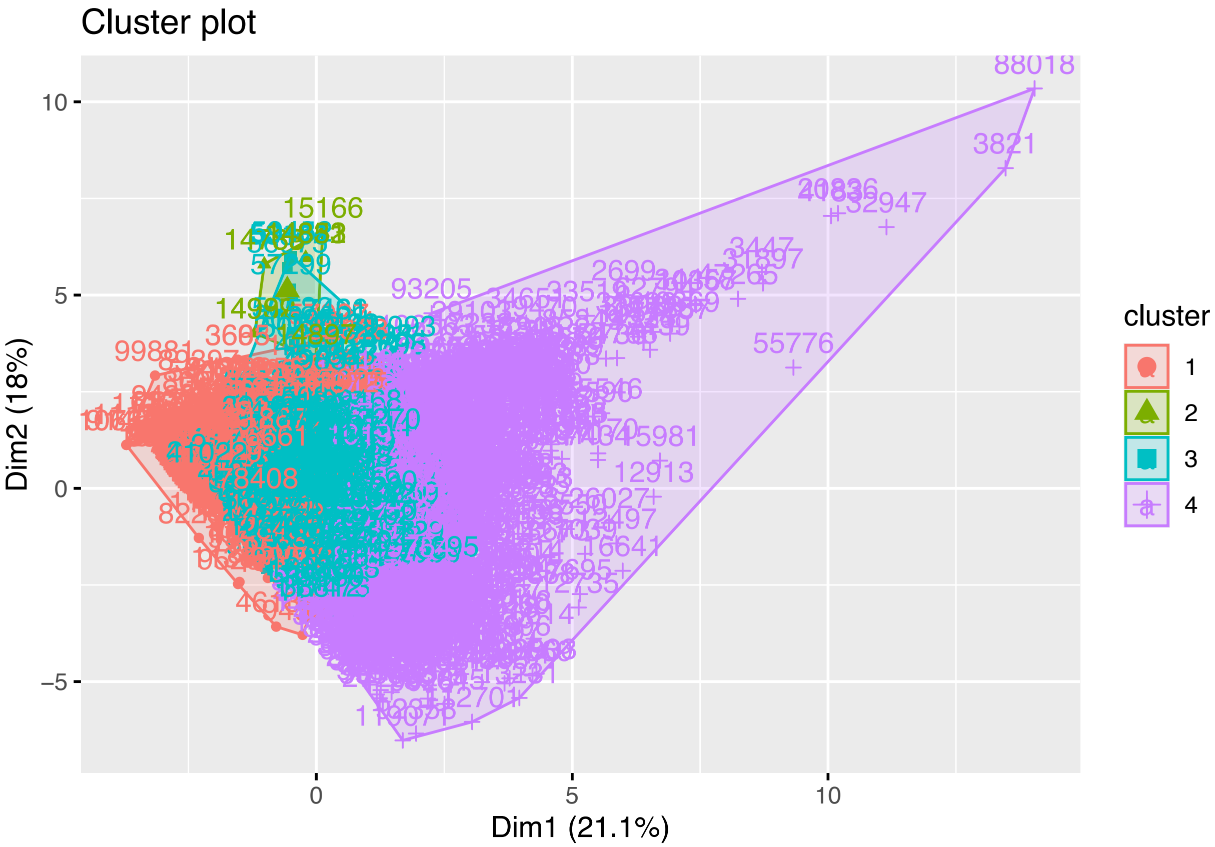
5. Classify hotel_type and got_desired_roomtype as required for clustering

Clustering

1. How to seggregate the hotel booking data? what is the optimal number of clusters?

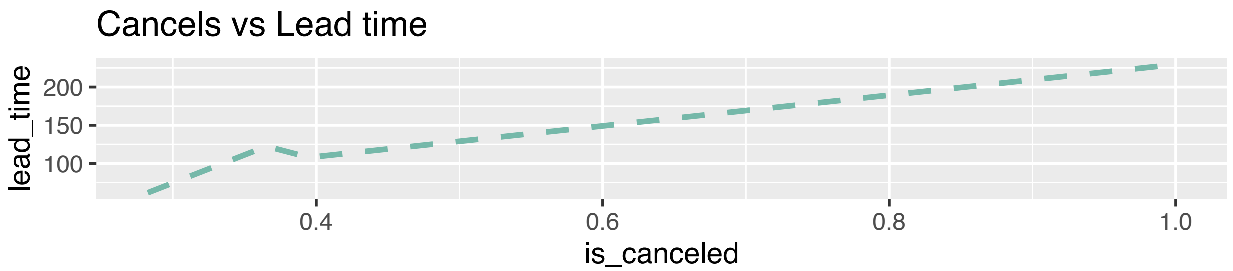
Through silhouette method, we have found that the booking data can be segregated in 4 distinct clusters

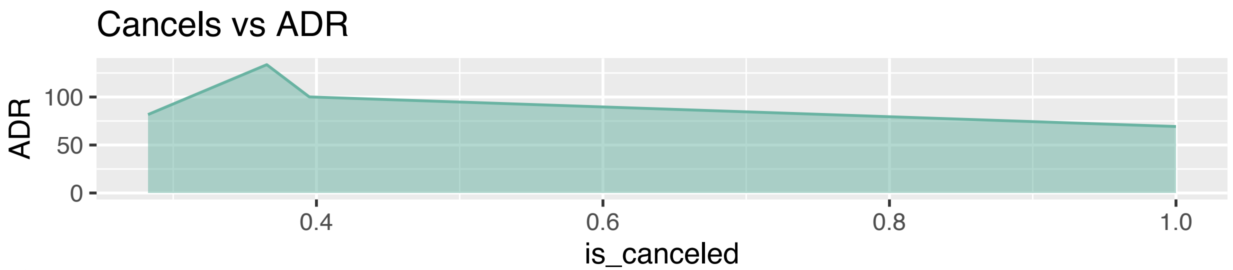
Using K-Means, visualizing clusters

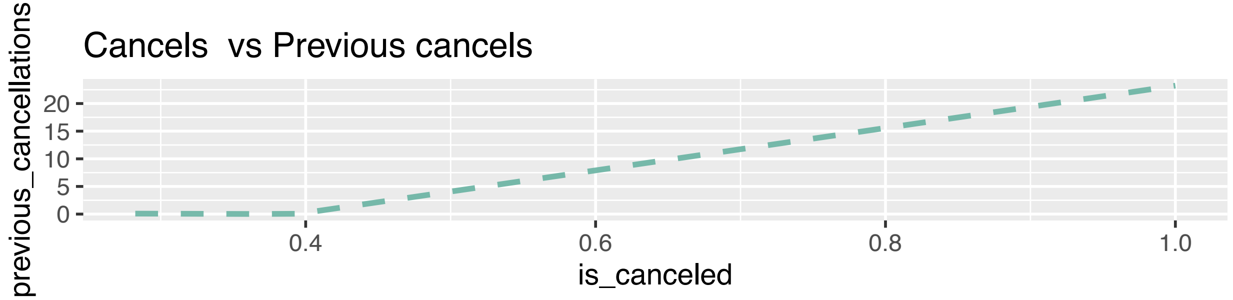


## The clustering is done through k-means method

2. From the segregated groups (clusters) find which which attributes are more associated with cancellations?


Cancels vs Lead time


Cancels vs ADR


Cancels vs Previous cancels

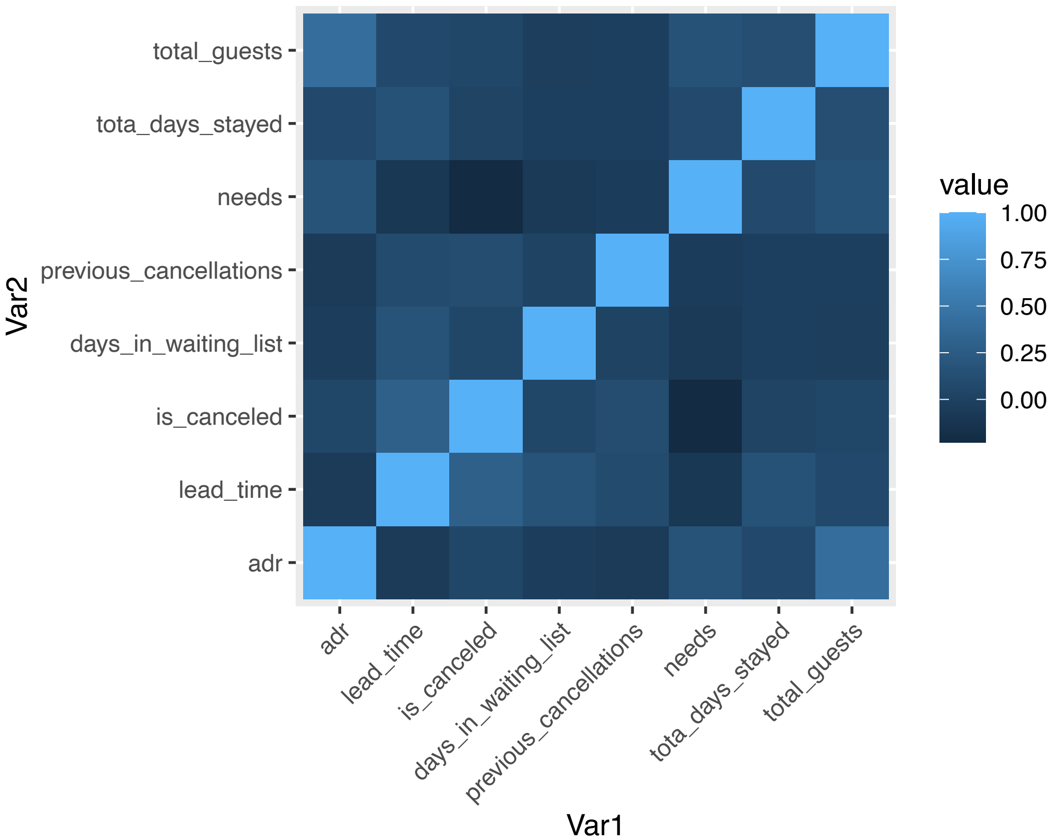
## It is evident that, Cancels are mostly likely to happen if lead-time is more, meanwhile they might have changed plans or found a better deal

ADR is inversely related to cancellation. Customers who spent/ more likely to pay high amount per night are less likely to cancel their bookings

Customers who have cancelled booking before are more likely cancel their bookings in the future

Probability

3. Which attributes are correlated to cancellation? What attributes are more liekly to be dependent on each other?



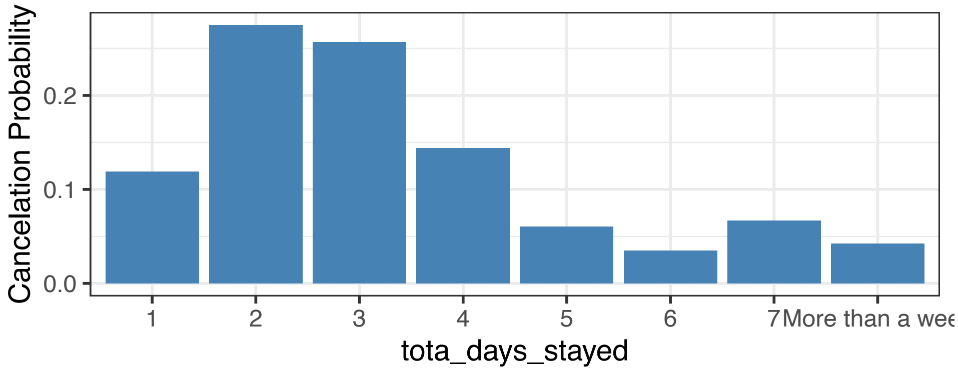
## Cancellation is not dependent on the number of requests raised during the booking. Cancellation has positive correlation with lead time.

Total number of guests and ADR have good positive corrrelation.

4. Which "duration of stay" is more likely to be cancelled?

PMF and CDF for total_days_stayed in hotel a reservation

Bar plot showing the distribution of cancellation probability vs total_days



## Customers who have booked the hotel for 2-3 days are more liekly to cancel their reservation

5. What type of bookings are likely to get cancelled? Does these customers who cancel the booking usually pay deposit?

```
## `summarise()` has grouped output by 'customer_type'. You can override using the `.groups` argument
```

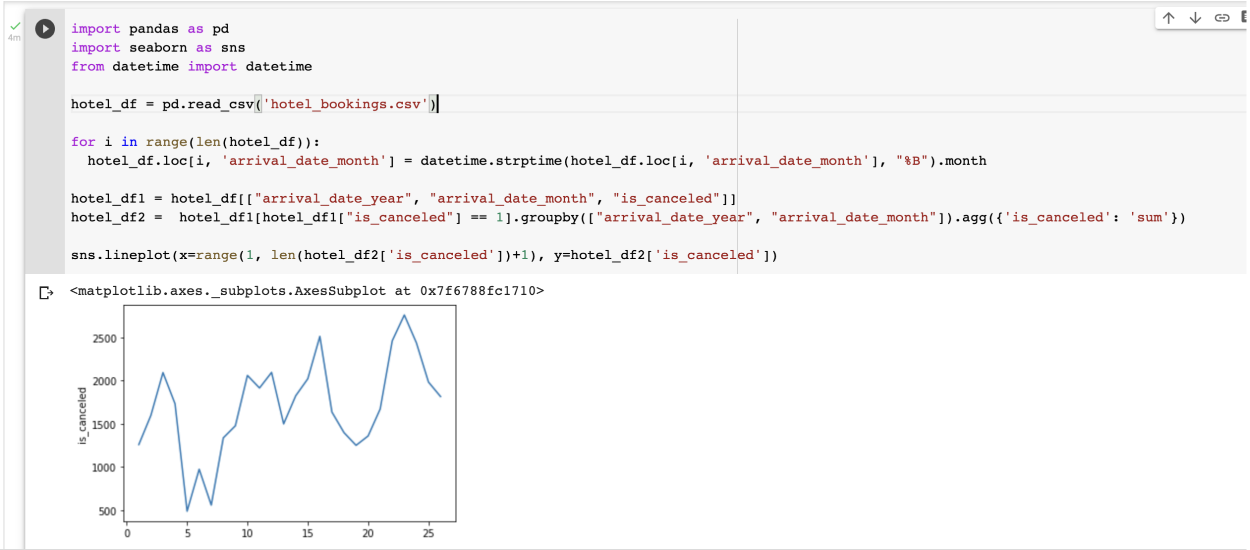
```
##     customer_type No Deposit Non Refund Refundable
## 1        Contract      0.016      0.012          0
## 2           Group      0.001         NA         NA
## 3       Transient      0.534      0.292          0
## 4 Transient-Party      0.121      0.024          0
```

Transient customers (individual booking) whodoesn't pay initial deposit are more likely to cancel their reservations. THey hold 53% of the total cancellation data

Time Series Analysis

6. Distribution of cancellations per month?

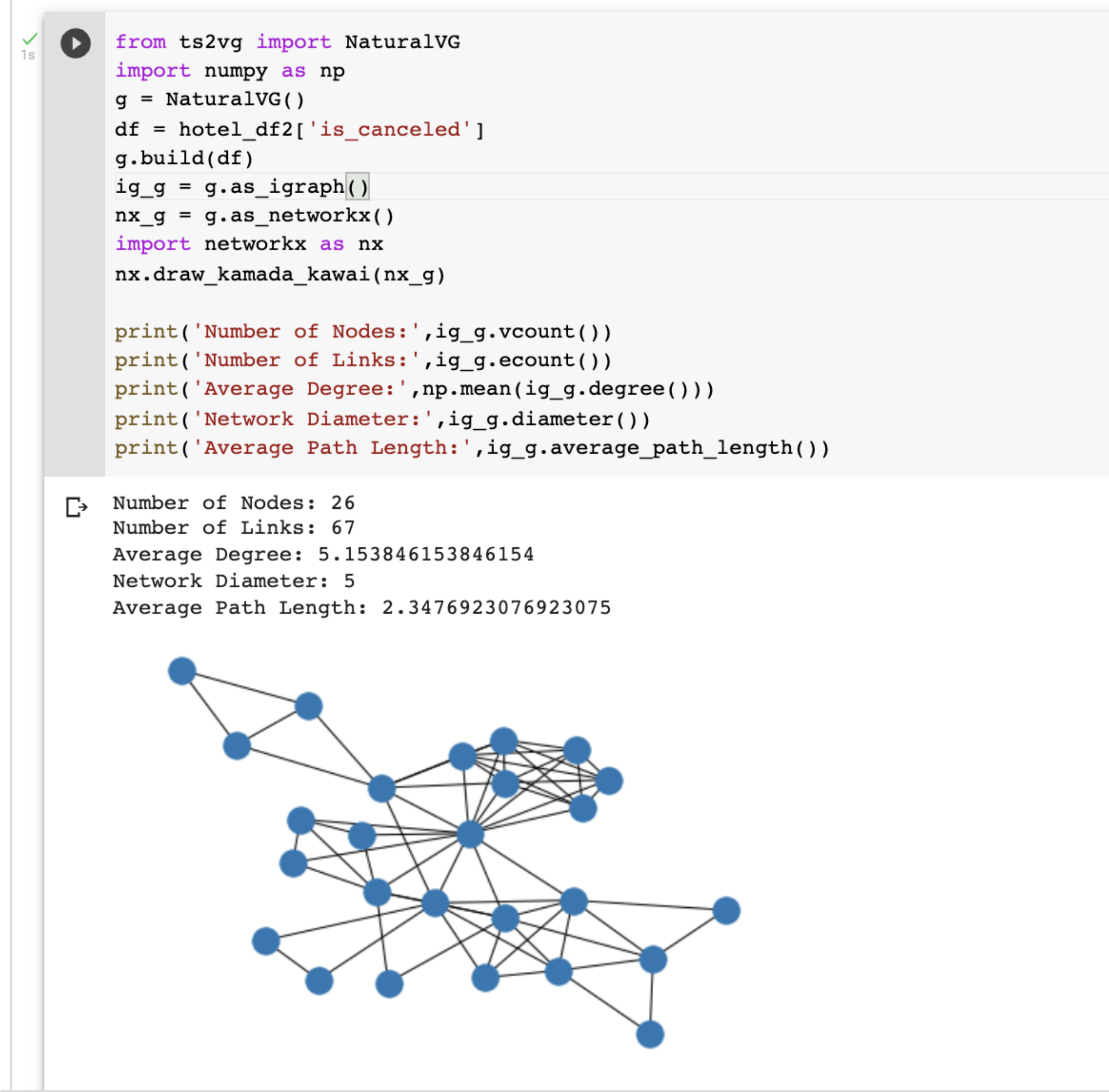
Let's check trend of Cancels over the time. Here is the SNS line plot for number of cancels per month over the time



```
import pandas as pd
import seaborn as sns
from datetime import datetime

hotel_df = pd.read_csv('hotel_bookings.csv')

for i in range(len(hotel_df)):
  hotel_df.loc[i, 'arrival_date_month'] = datetime.strptime(hotel_df.loc[i, 'arrival_date_month'], "%B").month

hotel_df1 = hotel_df[["arrival_date_year", "arrival_date_month", "is_canceled"]]
hotel_df2 =  hotel_df1[hotel_df1["is_canceled"] == 1].groupby(["arrival_date_year", "arrival_date_month"]).agg({'is_canceled': 'sum'})

sns.lineplot(x=range(1, len(hotel_df2['is_canceled'])+1), y=hotel_df2['is_canceled'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6788fc1710>
```

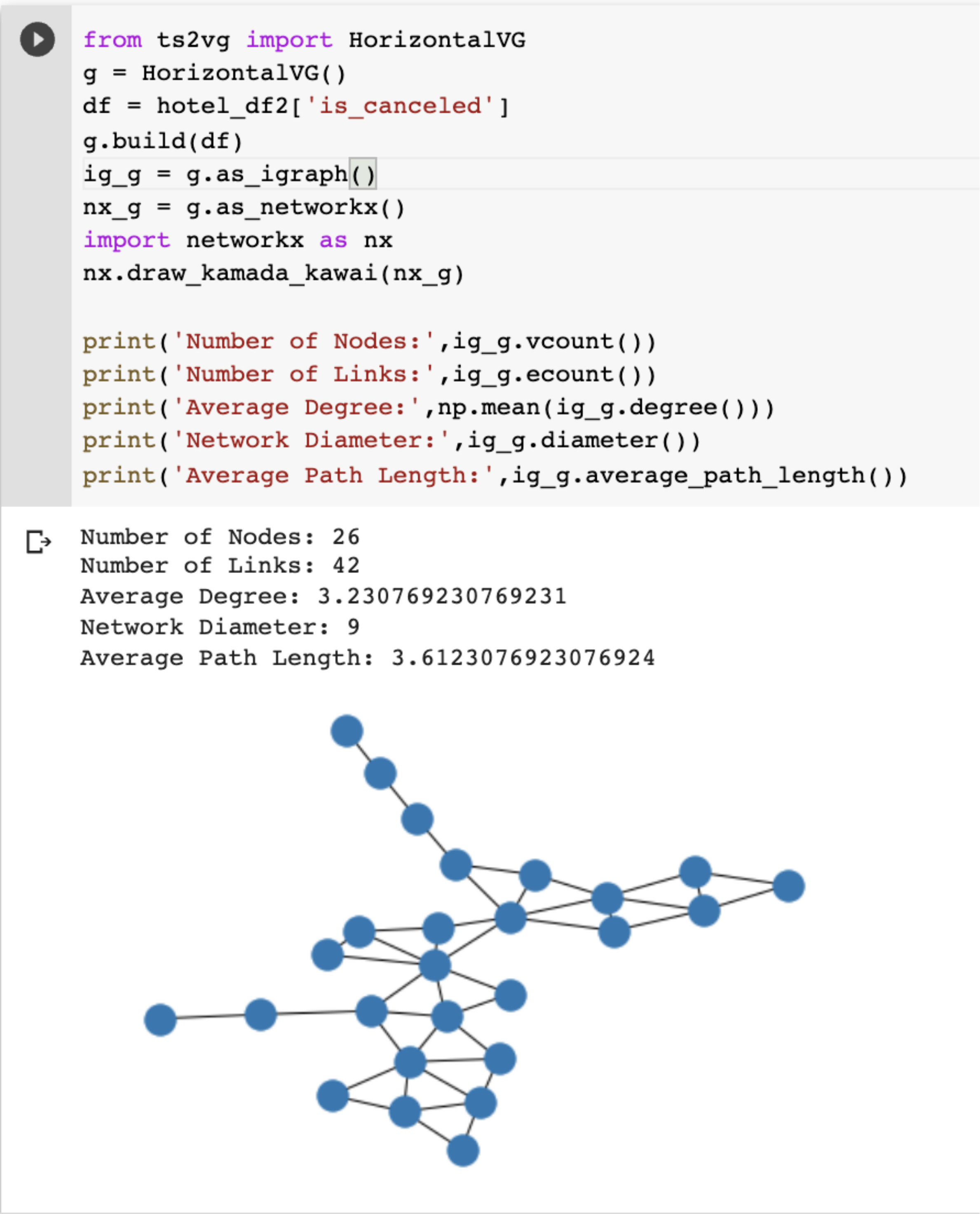
## Natural Visual Graph



```
from ts2vg import NaturalVG
import numpy as np
g = NaturalVG()
df = hotel_df2['is_canceled']
g.build(df)
ig_g = g.as_igraph()
nx_g = g.as_networkx()
import networkx as nx
nx.draw_kamada_kawai(nx_g)

print('Number of Nodes:',ig_g.vcount())
print('Number of Links:',ig_g.ecount())
print('Average Degree:',np.mean(ig_g.degree()))
print('Network Diameter:',ig_g.diameter())
print('Average Path Length:',ig_g.average_path_length())
```

```
Number of Nodes: 26
Number of Links: 67
Average Degree: 5.153846153846154
Network Diameter: 5
Average Path Length: 2.3476923076923075
```

Figure 1: Natural Visual Graph for cancels per month from Oct-2014 - Aug 2017



```
from ts2vg import HorizontalVG
g = HorizontalVG()
df = hotel_df2['is_canceled']
g.build(df)
ig_g = g.as_igraph()
nx_g = g.as_networkx()
import networkx as nx
nx.draw_kamada_kawai(nx_g)

print('Number of Nodes:',ig_g.vcount())
print('Number of Links:',ig_g.ecount())
print('Average Degree:',np.mean(ig_g.degree()))
print('Network Diameter:',ig_g.diameter())
print('Average Path Length:',ig_g.average_path_length())
```

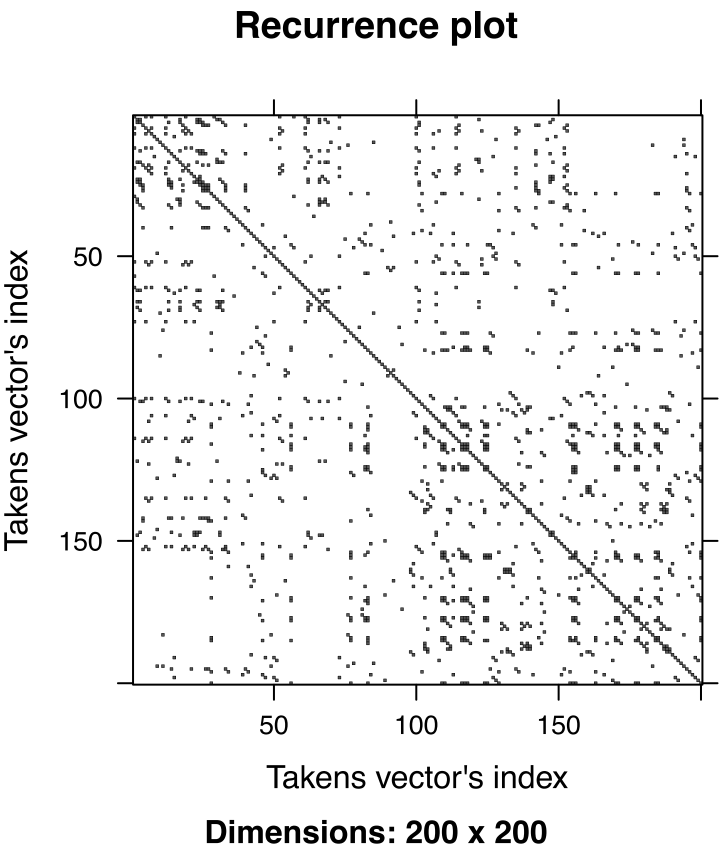
```
Number of Nodes: 26
Number of Links: 42
Average Degree: 3.230769230769231
Network Diameter: 9
Average Path Length: 3.6123076923076924
```

Figure 2: Horizontal Visual Graph for cancels per month from Oct-2014 - Aug 2017

Horizontal Visual Graph

7. RQA analysis of our data

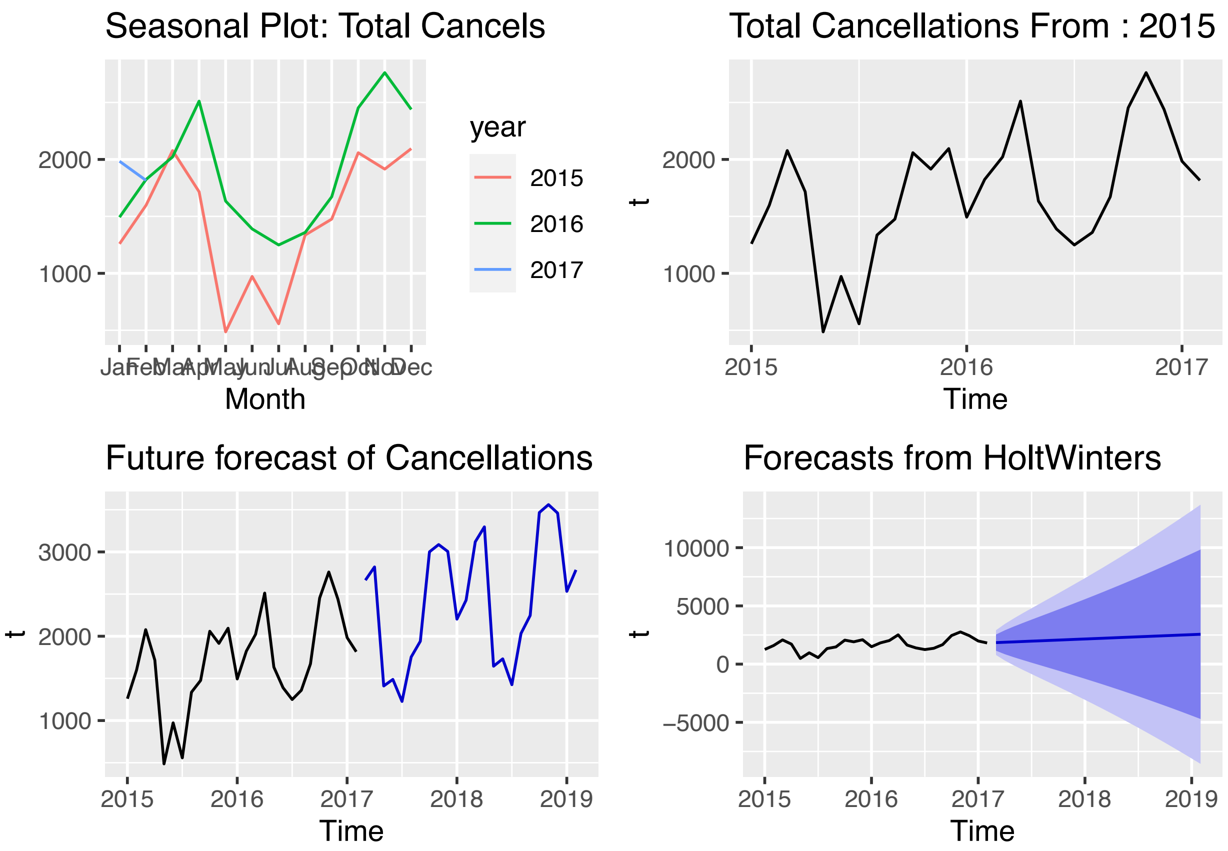
Reccurence plot of acc signals for Cancels per Day



8. Trend and Forecast of Cancellations over the time.

First two plots show the trend of Cancellations in current data

Next two plots forecast future cancellations using HoltWinters method



9. Permutation Entropy of Cancels

Entropy is ~0.93 very high indicating, randomness of cancel behaviour

Complexity is ~ 0.05

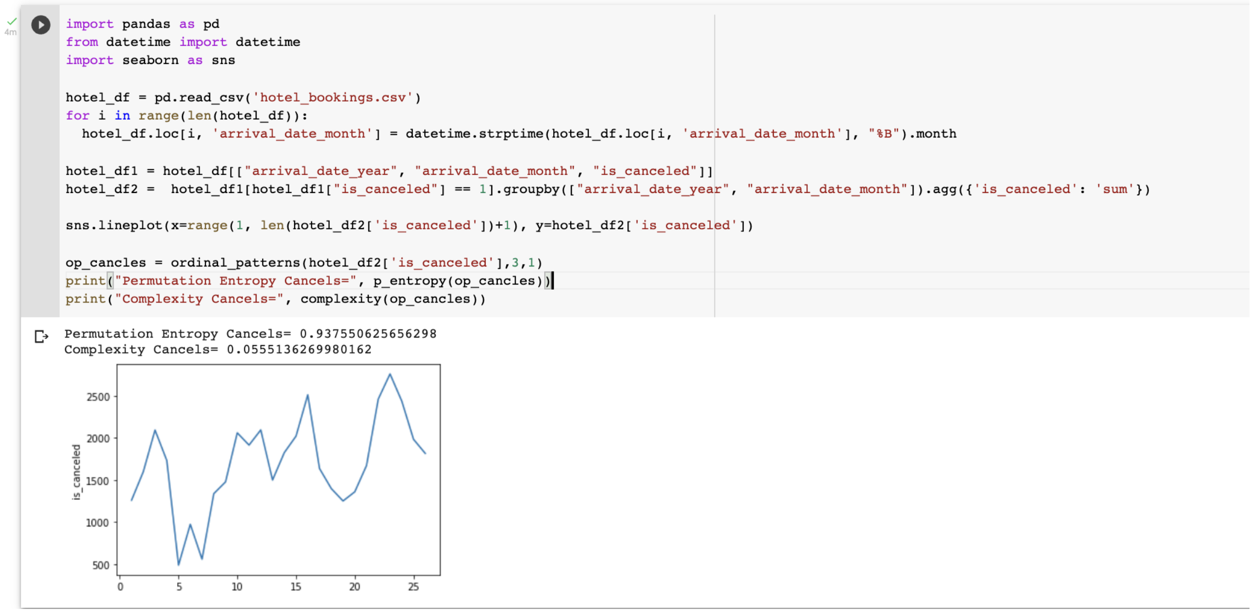
Text Analysis

Removing stop words and tokenizing the text

10. Most common topics of interest

Bar plot for most common words

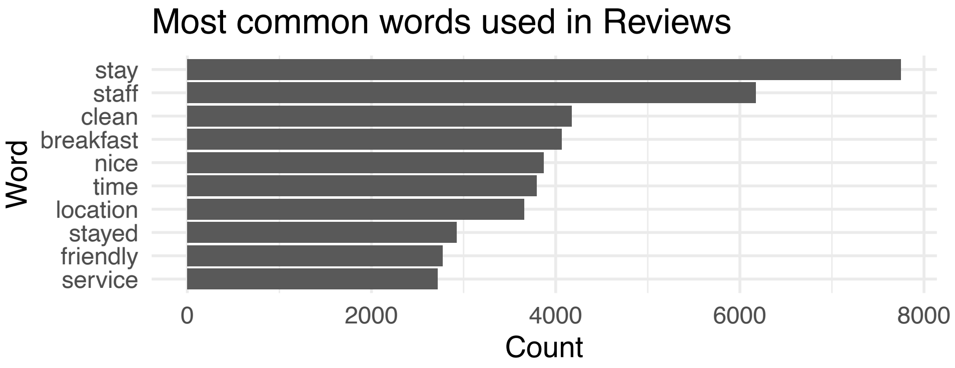
```
## Selecting by n
```



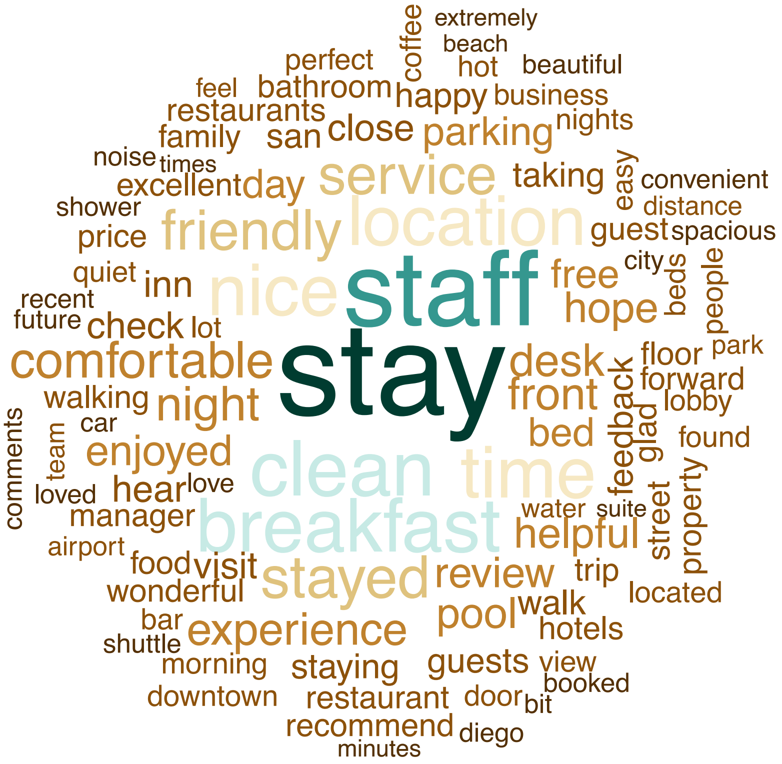
```
import pandas as pd
from datetime import datetime
import seaborn as sns

hotel_df = pd.read_csv('hotel_bookings.csv')
for i in range(len(hotel_df)):
  hotel_df.loc[i, 'arrival_date_month'] = datetime.strptime(hotel_df.loc[i, 'arrival_date_month'], "%B").month

hotel_df1 = hotel_df[["arrival_date_year", "arrival_date_month", "is_canceled"]]
hotel_df2 =  hotel_df1[hotel_df1["is_canceled"] == 1].groupby(["arrival_date_year", "arrival_date_month"]).agg({'is_canceled': 'sum'})

sns.lineplot(x=range(1, len(hotel_df2['is_canceled'])+1), y=hotel_df2['is_canceled'])

op_cancles = ordinal_patterns(hotel_df2['is_canceled'],3,1)
print("Permutation Entropy Cancels=", p_entropy(op_cancles))
print("Complexity Cancels=", complexity(op_cancles))
```

```
Permutation Entropy Cancels= 0.937550625656298
Complexity Cancels= 0.0555136269980162
```

Figure 3: Lineplot for cancels per month from Oct-2014 - Aug 2017



Wordcloud with top 100 words



As seen from the bar graph and word cloud, the most common words in reviews are stay, staff, location, time,clean, breakfast among.

So these are the topics of interest amon consumers and the hotels should look to make improvements on them.

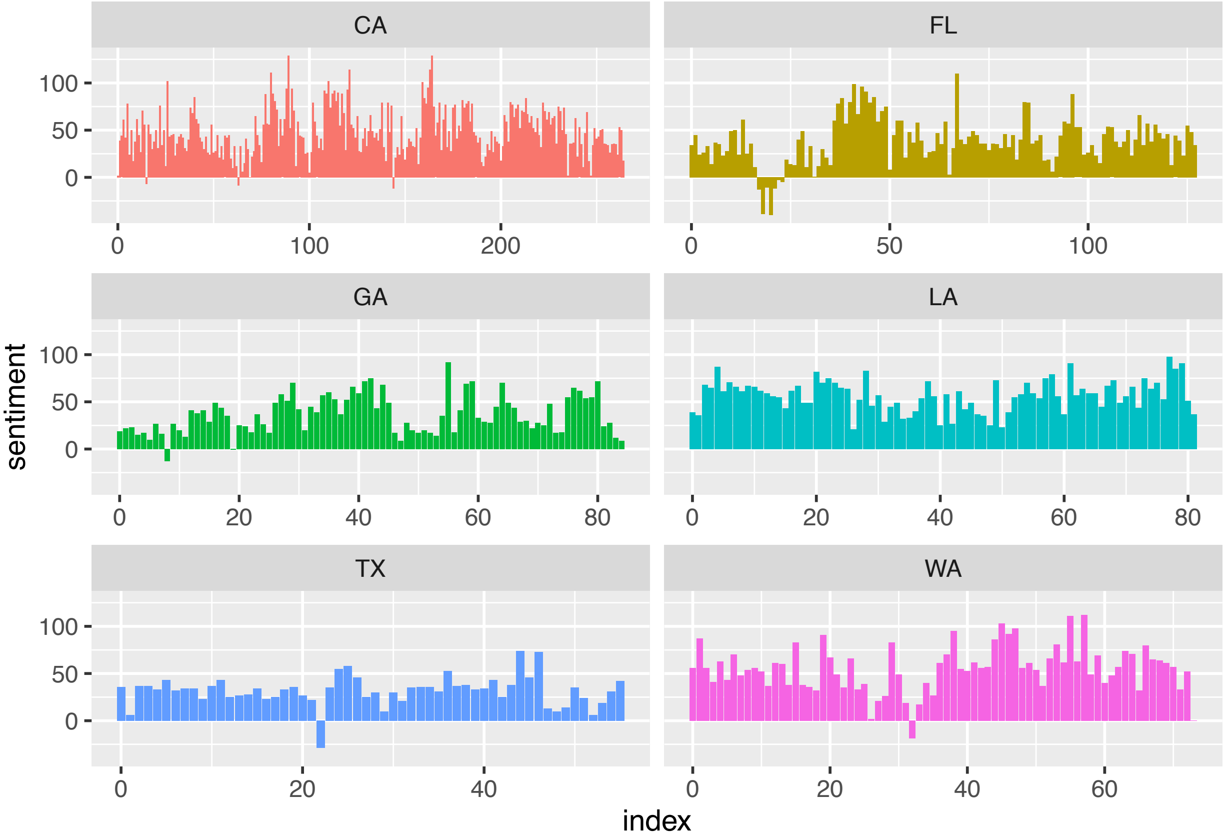
12. Sentiment analysis of customer reviews

Sentiment

```
## Selecting by n
```

```
## # A tibble: 10 x 2
##    word2     sentiment_value_from_prev_word
##    <chr>                              <dbl>
##  1 location                            3.00
##  2 stay                                2.84
##  3 service                             2.69
##  4 friendly                            2.62
##  5 clean                               2.57
##  6 staff                               2.25
##  7 time                                2.19
##   8 nice                          2.02
##   9 breakfast                     1.46
## 10 stayed                         1.38
```

The sentiments around breakfast, time, stayed, staff are a little less which the hotels should work on

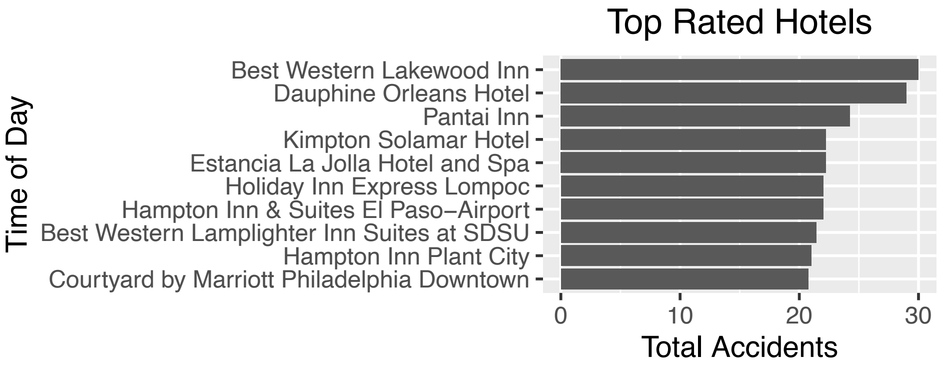


## 13. What are the hotels which have highly positive reviews?

```
## Joining, by = "word"
```

```
## Joining, by = "name"
```

```
## Selecting by sentiment
```



## These are the hotels which have reviews with high positivity

Recommendations:

Customers who have booked the hotels very early are more likely to cancel their reservations. Hotel manangement should provide timely reminders to the customers who have booked the rooms.

Double check the reservations with customers who had previously cancelled their booking.

Hotel management should impose a rule that reservations should be accompanied with minimum deposit.

Assign more staff during winter as more bookings are likely to be made for that season.

"Cleanliness", "Friendly staff" and "Quality of meal" are the major propellent for good rating for hotels.