Room, Revenue and Rate:

(Sujita Kapali 07-27-2019)

Table of Contents:

Introduction:

The title of the project is “Room, Revenue and Rate”. The purpose of this project is to forecast the number of rooms sold and find the rate for a hotel room in order to maximize room revenue.

Having worked in a hotel industry I have found that there is a big challenge in forecasting rooms to be sold and determining the daily rate of a room. The rates are usually set by the brand. They are then accessed by a revenue manager and finally predicted to go in forecast plans by a hotel manager. By the time it drills down to the forecast plan, most of the time, we find the forecasts are off than what was originally budgeted. Brands use algorithms to determine the rate; revenue managers use industry expertise and hotel managers know their hotels!

This is why I thought it would be a great idea if I could do a project on this topic so I can understand what it takes to determine that magic number.

The rate of a hotel room is defined as the total room revenue of the hotel divided by the total rooms sold.

The factors that contribute to the average rate are:

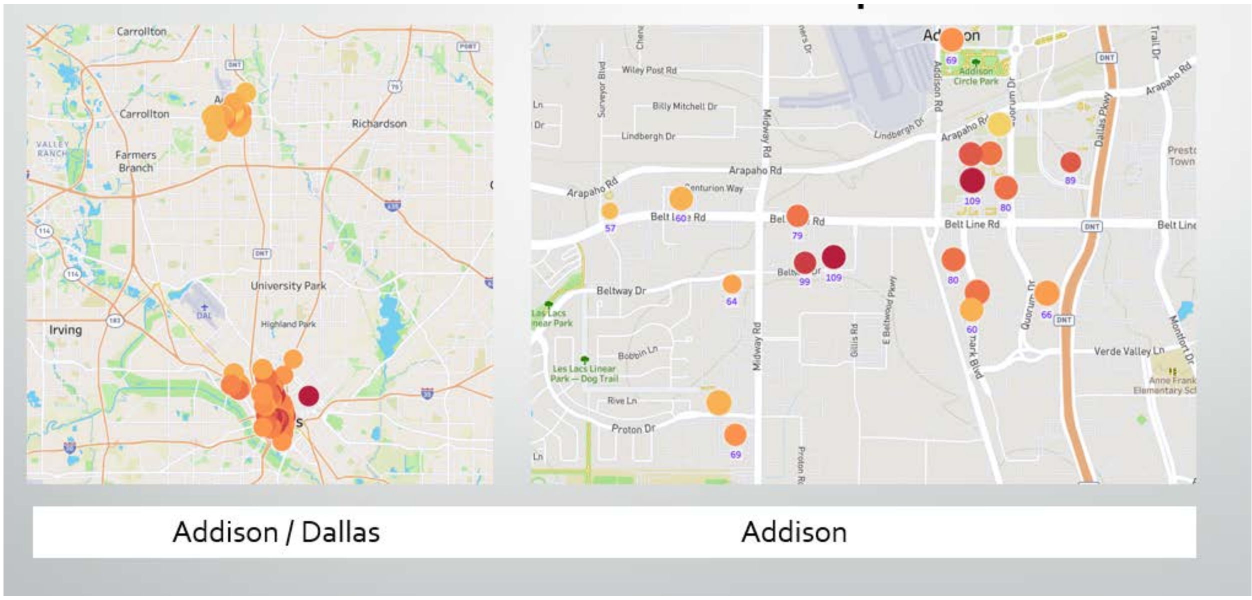
Rooms Available
**Demand and Supply
* Location
*Type of room
Competition
Season
Day of the week.

*For the purpose of this project, I am using hotels that are located within a certain acceptable range of Addison and Galleria. I am only using rate for 1 room per 1 guest per night. I picked this category because this is the first rate that comes up when you open your browser to search for hotels.

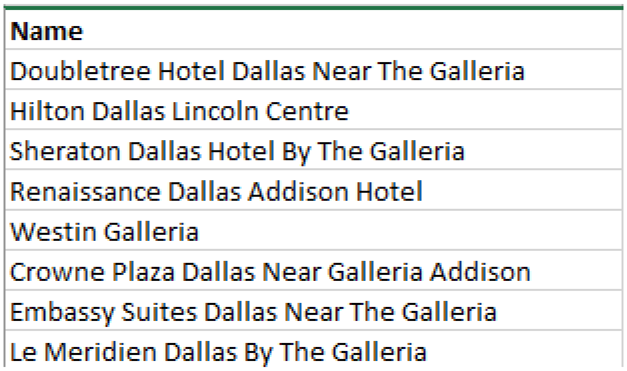
**The limitation that I have in this analysis is getting an accurate measure of demand. In order to compensate for this, I am using historical data and making assumptions that demand for this certain segment will follow historical trend.

Description of the Subject:

For the purpose of this study, a hotel was picked in the Addison Dallas Galleria area. Reason for picking this particular hotel is because of familiarity with the location and the hotel itself.



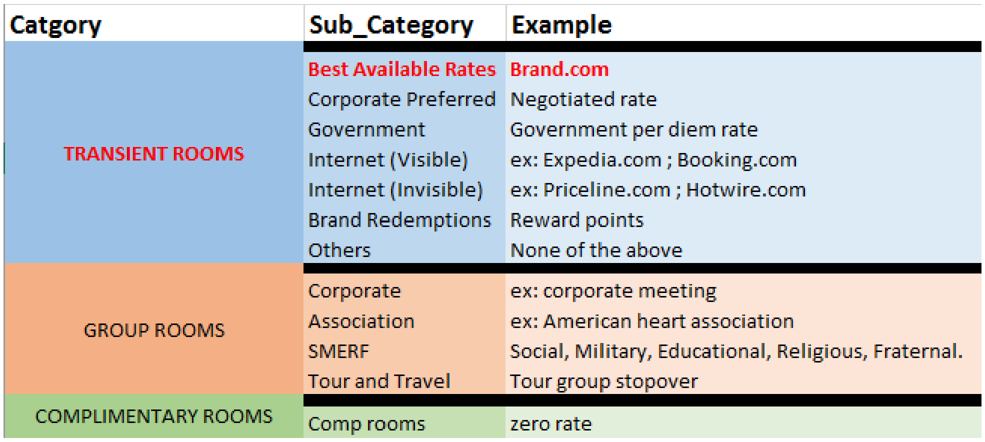
The following hotels were used as a base for competitor pricing as provided by the official STAR report.



Name
Doubletree Hotel Dallas Near The Galleria
Hilton Dallas Lincoln Centre
Sheraton Dallas Hotel By The Galleria
Renaissance Dallas Addison Hotel
Westin Galleria
Crowne Plaza Dallas Near Galleria Addison
Embassy Suites Dallas Near The Galleria
Le Meridien Dallas By The Galleria

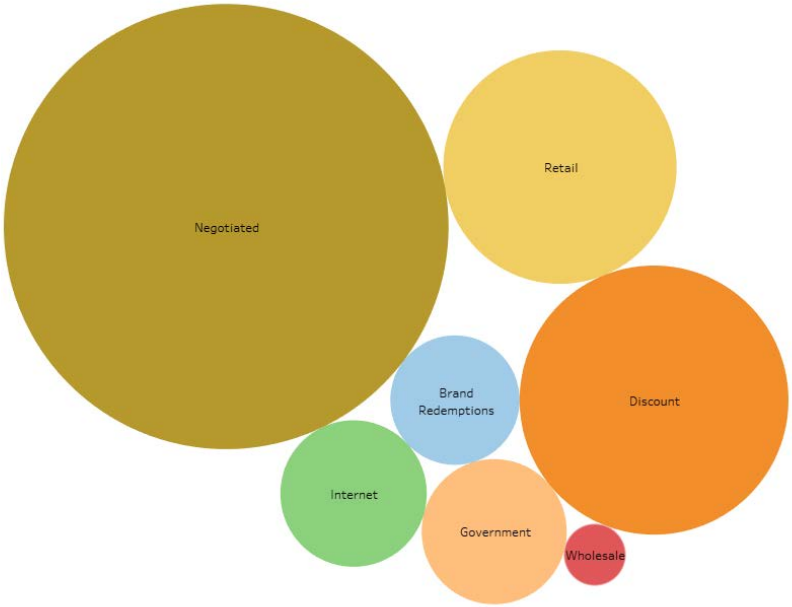
Only one sub -category of the market segment was used for the study.

The following table lists the different market segments for rooms.



Catgory	Sub_Category	Example
TRANSIENT ROOMS	Best Available Rates	Brand.com
	Corporate Preferred	Negotiated rate
	Government	Government per diem rate
	Internet (Visible)	ex: Expedia.com ; Booking.com
	Internet (Invisible)	ex: Priceline.com ; Hotwire.com
	Brand Redemptions	Reward points
	Others	None of the above
GROUP ROOMS	Corporate	ex: corporate meeting
	Association	ex: American heart association
	SMERF	Social, Military, Educational, Religious, Fraternal.
	Tour and Travel	Tour group stopover
COMPLIMENTARY ROOMS	Comp rooms	zero rate

The plot below shows the distribution of rooms sold for each sub-category in Transient rooms. Although retail rooms occupy a smaller portion of the segment, we will look at the Best Available Rates because of the availability of competitor information online.



Description of Data:

Data for this project was obtained from the following sources:

1. Historical Rooms, Revenue and Rate data are based upon actual data. They are transformed for this project in order to protect proprietary information.
2. Competitor rates for the month of August was obtained by web scraping TripAdvisor data.
3. Rates for neighborhood hotels were also obtained by web scraping TripAdvisor data.
4. Due to the limitations of data that could be scraped from TripAdvisor, a different website hotelplanner.com had to be used to obtain location information.

Below are the links to the websites:

https://www.tripadvisor.com/
https://www.hotelplanner.com/

Challenges in obtaining information from the sources:

The hotel industry is not very open to sharing their data. I tried various websites including brand websites. They do not allow users to make an API call to pull data. TripAdvisor will provide data by request but they are not to be used for academic research purposes. They would not let you web scrape their data. I had to create a roundabout to get that information.
headers={'User-Agent': 'Mozilla/5.0 (iPad; U; CPU OS 3_2_1 like Mac OS X; en-us) AppleWebKit/531.21.10 (KHTML, like Gecko) Mobile/7B405'}

Also the link would not refresh after selections are made. This made it impossible to use the same website to obtain selected information. It also did not provide hotel addresses in the same page. For this reason, I had to go to a different website to pull the needed information.

Analysis of Data:

Data was analyzed in three stages:

1. **Exploration:**
   A brief study of the data was done by plotting a correlation matrix. The purpose was to find the feature that is most correlated to room revenue. Details on this is explained in the next section.

2. **Prediction:**
   Once it was determined that the room revenue is correlated to rooms and rate, the next step was to predict the rooms sold. Details are discussed in a different section.

3. **Prescription:**
   Now that we have predicted the rooms sold, the next step was to plug the rate that will maximize revenue.
   Details are discussed in a different section.

Methodology Used to Analyze Data:

1. **Exploration:**
   A correlation matrix was used to determine the relationship between room revenue and other features.

2. **Prediction:**
   The data to be analyzed is a time series data. After doing much research, I ended up using Python fbprophet package.

   The other tools I tried were ARIMA modelling for R, fbprophet for R, LSTM (Long Short-Term Modelling) time series in Python.

   The challenges I had with all the models were removing outliers, filling in missing values and inverse transforming the log of values. I manually removed the outliers and imputed missing values based upon historical values. As for inversing the values, I found out after much research that the models transform the log values (add additive or multiplicative component) while making predictions and hence is impossible to invert them back to its original state.

   I also used several regression models to test the hypothesis that room revenues are determined by rooms sold and rates.

3. **Prescription:**

   Excel Solver was used to plug in the constraints and output the maximum revenue.

   The major challenge in using this tool is that unlike other functions in Excel, this particular solver function does not let you copy and paste formulas. In order to calculate revenue for each day you have to create a different sheet. For the purpose of this project, I calculated for one day only.

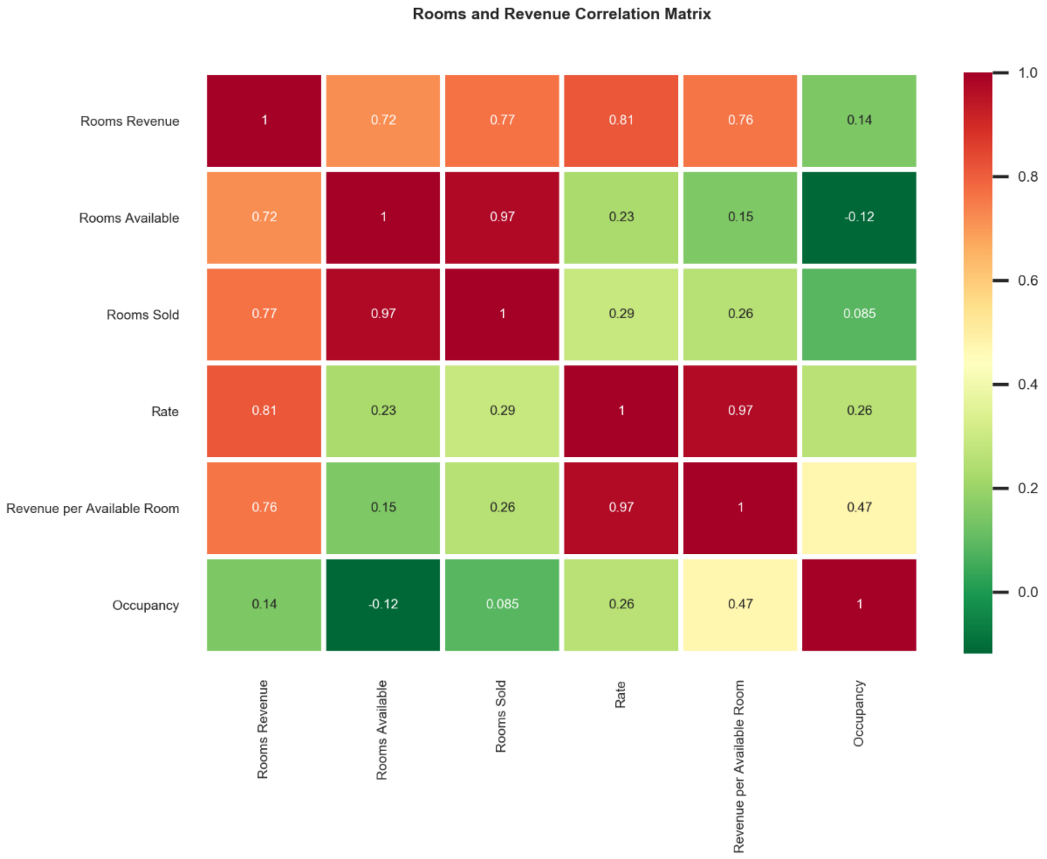
<u>Correlation between Room Revenue and other features:</u>

This is a generalization report and is not for the particular hotel under study.

Data was obtained for a sample of 10 hotels. The selection of hotels included luxury brands, economy brands, mid-scale, select service and full-service brands.

The variables selected were rooms sold, rooms available, occupancy, average daily rate, revenue per available room and room revenue. The purpose of this exercise was to find out correlation between room revenue and other features.

A correlation matrix (shown below) was plotted in Python using seaborn.



The matrix shows that there is high correlation between room revenue and rooms and rate and is least correlated with occupancy. This does make sense because revenue goes up as you increase the sale and rate of a room. A hotel has fixed number of available rooms. The revenue can be maximized by increasing the rate. The hotel can have full occupancy but if they are selling at a lower cost, they are not making any revenue. Note that there is difference between occupancy and rooms sold. Rooms sold are rooms sold that generate revenue. Occupancy means how many rooms were occupied. It could include complimentary rooms and rooms for in-house use.

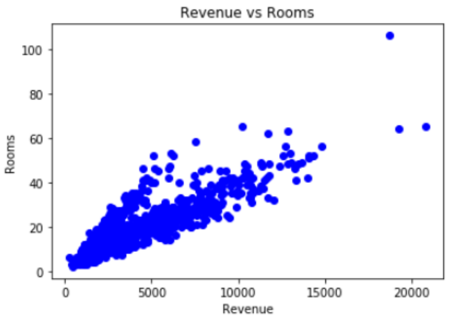
Model Testing:

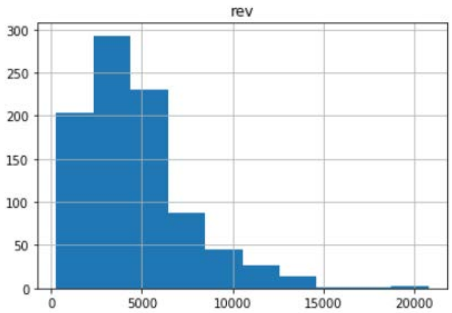
Several regression models were fit in order to test the hypothesis if we can predict rooms revenue with rooms sold and rate.

For the purpose of this exercise, initially a three-year monthly data was used. None of the models showed a positive relationship between revenue and rooms or rate.

The data was then replaced with a two-year daily data and outliers were removed manually.

The initial plot showed a positive correlation between revenue and rooms.

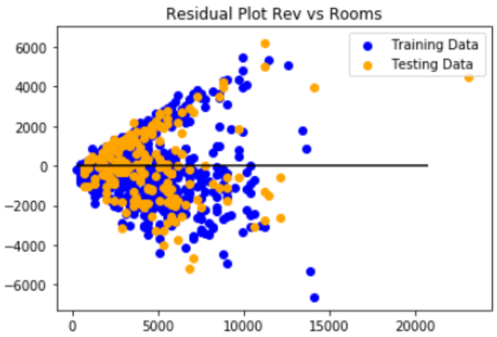




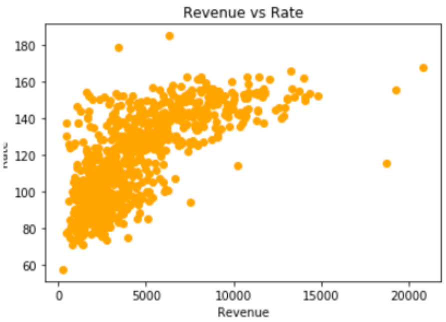
The dataset was split in to train and test data.

I started with a simple linear regression model. The train and test scores came to be 0.68 and 0.69. This could be better.

```
Training Score Rev vs Rooms: 0.686455142640532
Testing Score Rev vs Rooms: 0.6980897278384115
```

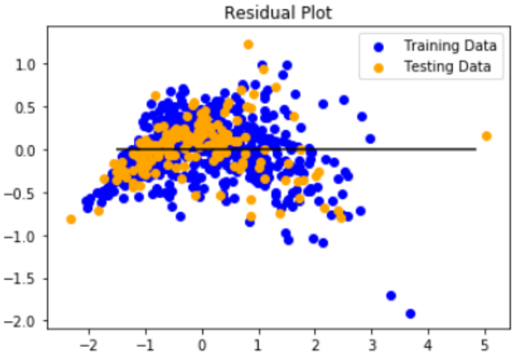


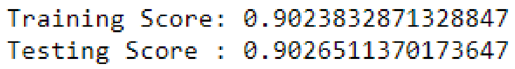
The same test was run for rate. It gave a negative score. Although the initial plot showed a positive relati on between revenue and rate, the statistics were not satisfactory.



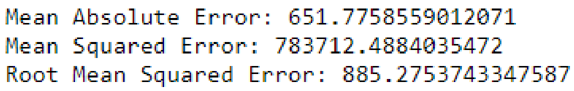
Next, a series of tests were run using a combination of rooms and rate.

The data was split into train and test and scaled.
Started with linear regression. Although the initial plot and residual score looked promising the statistics showed a big percentage of error.



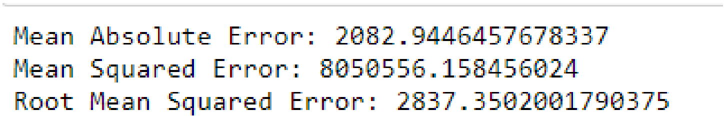


```
Training Score: 0.9023832871328847
Testing Score : 0.9026511370173647
```



```
Mean Absolute Error: 651.7758559012071
Mean Squared Error: 783712.4884035472
Root Mean Squared Error: 885.2753743347587
```

The error from average of the data provided the result below:



```
Mean Absolute Error: 2082.9446457678337
Mean Squared Error: 8050556.158456024
Root Mean Squared Error: 2837.3502001790375
```

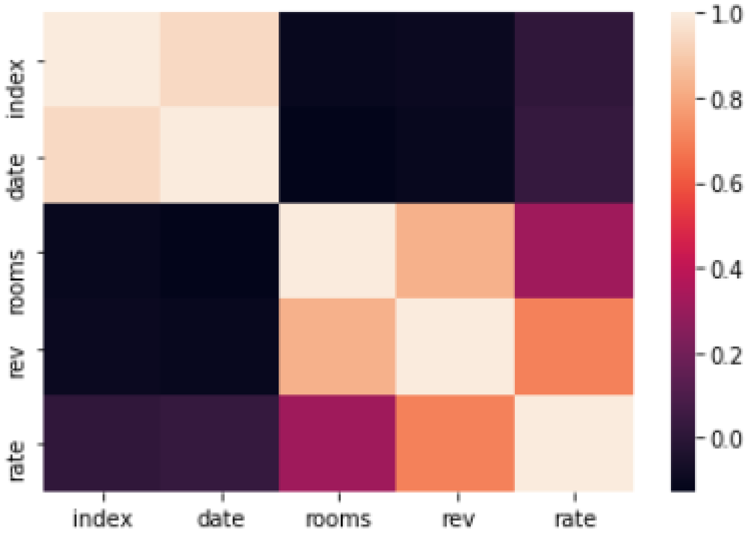
This model was disregarded because the average scored better than our actual model.

A series of other tests were performed and the following results obtained:

Model	Score	Result
RandomForestRegressor	`0.8867228481757251`	`[(0.6204741594030869, 'rooms'), (0.37952584059691324, 'rate')]`
GradientBoostingRegressor	`0.9018650858921968`	`[(0.6260904097082505, 'rooms'), (0.37390959029174947, 'rate')]`

Both the results above show that they are a better model than the linear regression. The scores are above 85 percent and both show that rooms sold define revenue more than the rate.

This is also confirmed by the correlation plot below.

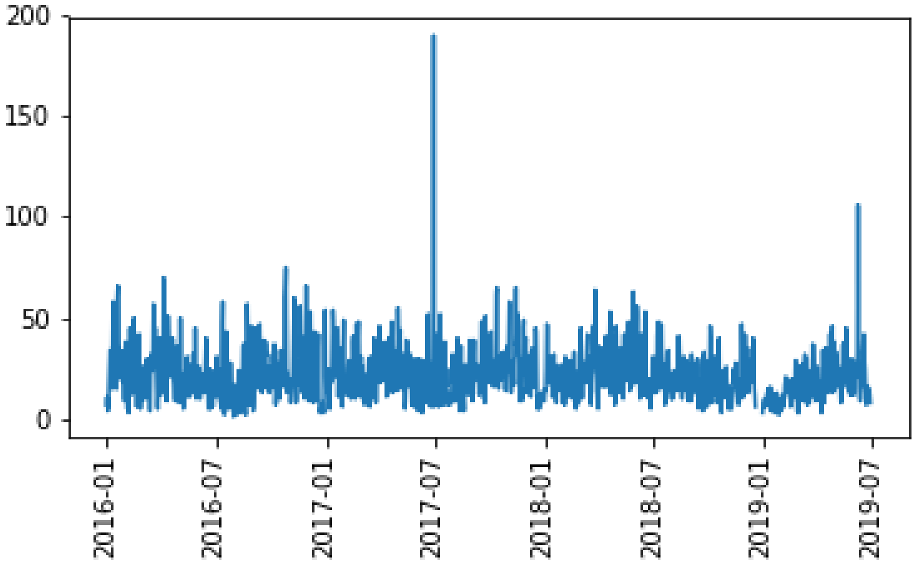


Forecasting Rooms using fbprophet

The rooms were predicted from July through December 2019 by using fbprophet.

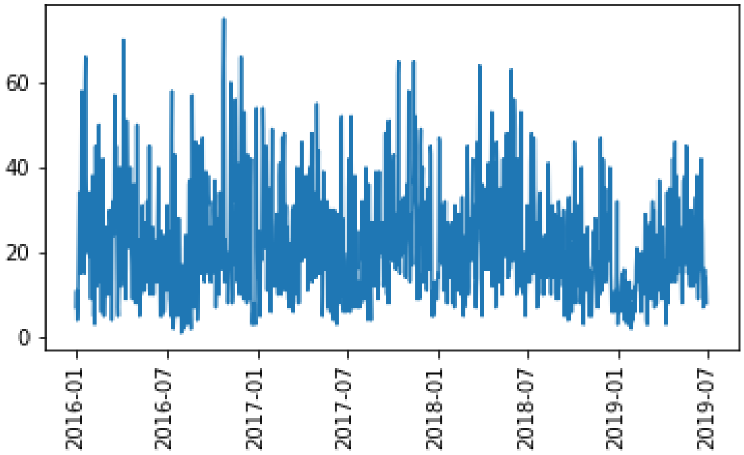
Below are the steps that were taken:

A plot was created using the original data series. The original plot shows two outliers. The data also shows seasonality and some missing values at the end of 2018.



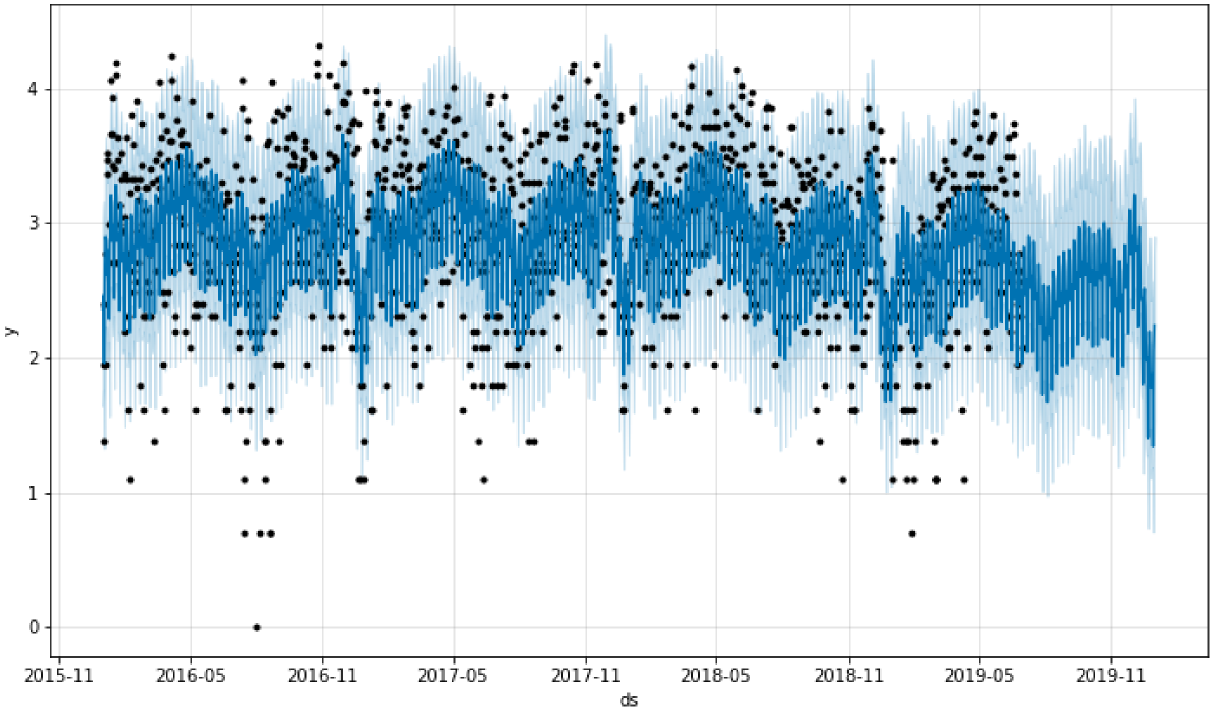
The outliers might have been a coding error. Because Prophet can handle outliers in historical data only by fitting trend changes, these two outliers were corrected manually by taking historical average.

There were missing data at the end of 2018. These were imputed manually by taking historical averages for the same period. This must have been a date entry error. This hotel underwent construction during the latter part of the year but did not actually completely shut down. Below is the plot after outliers were corrected and missing values were imputed.

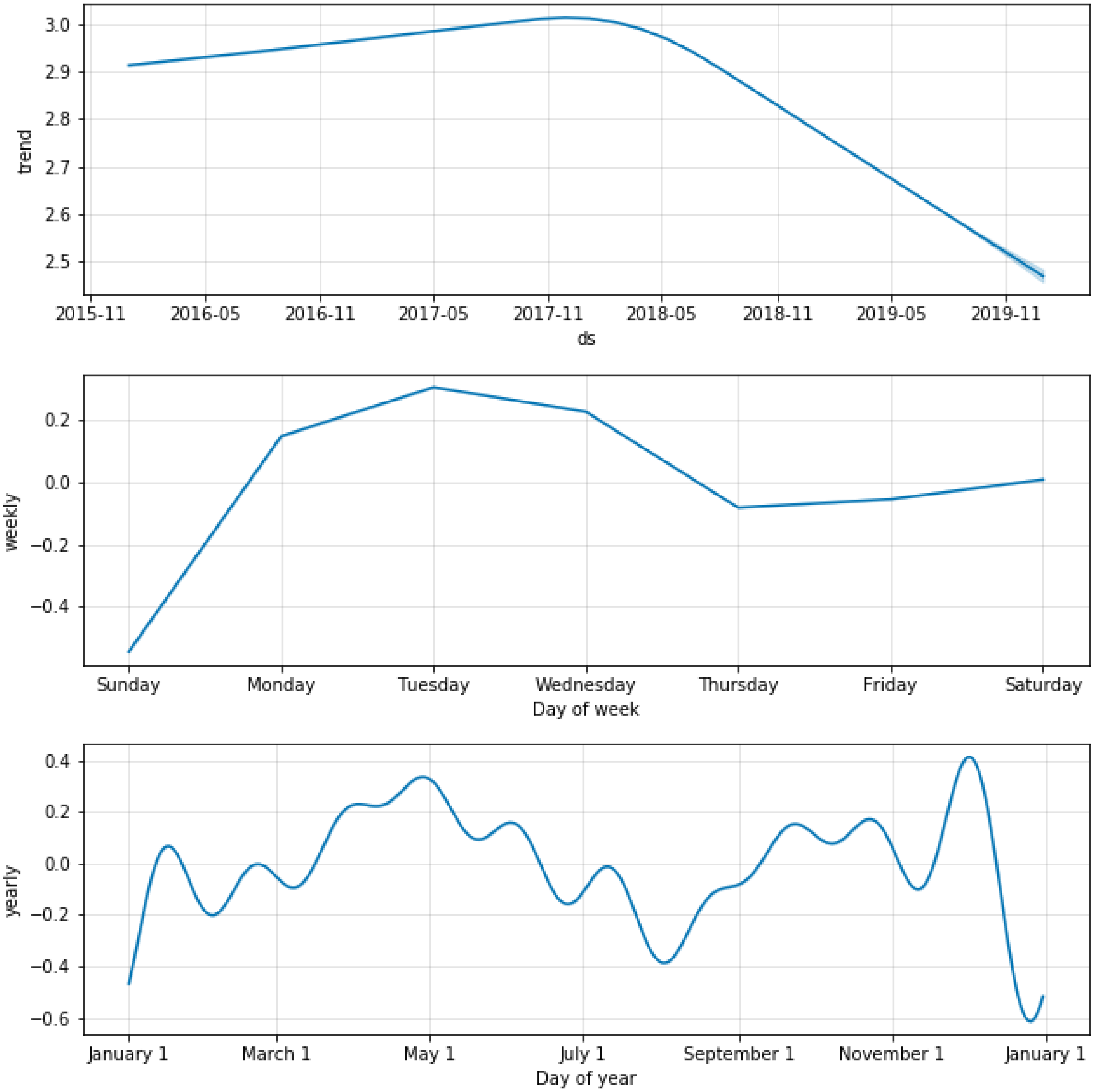


The data was converted to fbprophet friendly columns, date as ds and values as y. The values were converted to log to make them stationary. Original values were saved for later use.

Model was fit and values forecasted until the end of 2019. In the figure below, the black dots represent original values and the blue lines at the end represent forecast.



Next the components were plotted.

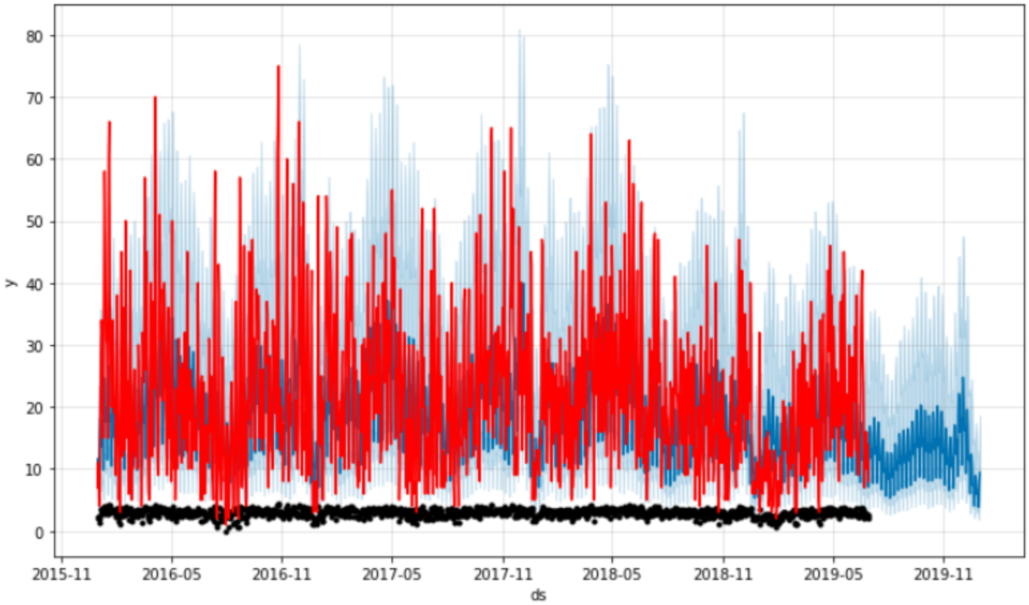


Trend - It shows a downward trend. This is because the hotel underwent renovation. This is where it is important to understand the context of the data.

Weekly – It shows a weekly seasonality. The rooms are more occupied during business days Mon – Wed.

Yearly – It shows some yearly trend mid-year, around October and December. This is probably due to holidays.

Next the forecast data was plotted. But it was plotted with original data in log format. The black line is the original data. The full data set was exported to a csv file



The Mean Absolute Error is 8 which means our predictions are off by 8 rooms.

The limitation of this exercise was converting the data back to the source form. It was not possible to invert the data back to its original number. The original data gets converted twice. First, it gets converted to log. Second time, it gets converted when predict function is used.

Below is an excerpt from stack overflow regarding this issue:

“You take log on the initial values of your dataframe df, but then apply the np.exp on the forecast of your time series.
The function fitted by your Prophet instance to the training set is not the identical function.
It's an additive model whose weights are learned as part of an optimization problem.

Therefore, the predicted values (values on your forecast dataframe) are not exactly equal to those of np.log(df[y]). That's why when you're inverting them you don't end up with the same values as the original ones.
NB: Your forecast function is certainly "better" than the identity function as it avoids overfitting (and actually learns something

In Short:

Let's say y is your original values and y_hat is your forecasted values. Then:
np.exp(np.log(y)) different from np.exp(np.log(y_hat))

Source: https://github.com/facebook/prophet/issues/220

The forecasted data comes with a confidence interval that specifies the probability (by default 80%) that your true data will lie within it.

Optimization using Excel Solver:

We have seen from the correlation matrix that room revenue is highly correlated to rate. Now that we have predicted the rooms that will be sold during a certain month, (for this exercise we are only taking the month of August), what rate should the rooms be set at in order to maximize revenue.

Assuming that the other segments are pre-determined, we set up the variables and constraints as follows:

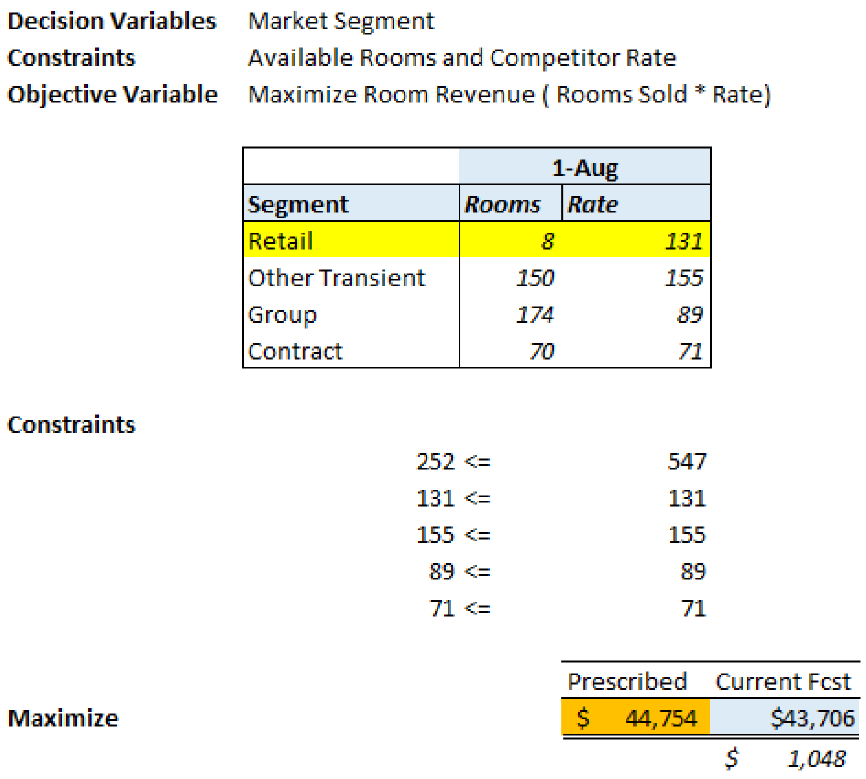
Decision variables: Market Segment (Transient, Group, Contract)

Constraints: Available rooms and Competitor Rate
Competitor rates were obtained by web scraping TripAdvisor data.

Objective Variable: Maximize room revenue (rooms sold * rate)

Inequality constraints: All rooms sold and rates must be positive

The limitation of this exercise is that Excel doesn't allow to copy and calculate the formula. The cell has to be calculated for each day!



Decision Variables	Market Segment
Constraints	Available Rooms and Competitor Rate
Objective Variable	Maximize Room Revenue (Rooms Sold * Rate)

	1-Aug	
Segment	***Rooms***	***Rate***
Retail	*8*	*131*
Other Transient	*150*	*155*
Group	*174*	*89*
Contract	*70*	*71*

Constraints

252	<=	547
131	<=	131
155	<=	155
89	<=	89
71	<=	71

	Prescribed	Current Fcst
Maximize	$ 44,754	$43,706
		$ 1,048

Findings and Conclusions:

For some reason, this particular hotel hasn't forecasted any rooms for August 1st. If they forecast 8 rooms as predicted by our ts model and sell at a rate of $131 they could be making $1,048 per day.

For Later:

This project is very basic and is based upon the "Classic" method of revenue management of a hotel. The rate structure of a hotel room is much more complex. The rates of a hotel depend upon numerous other factors like the type of room, number of beds, number of guests, check in and check out dates etc. The next step would be to build this model up.

The other thing I would like to do is to check for an alternative optimization tool in place of Excel Solver.