Insights of a hotel business through data science analysis

Torben Kraft [1] & Ana Peralta[2]

[1]E20181424; E20181424@novaims.unl.pt
[2]M20180946; M20180946@novaims.unl.pt

Abstract:
This project intends to analyse data from an existing hotel using techniques from data science to get deeper insight of a hotel's business performance during the last six years and understand how this analysis can help the management decision process of the hotel. Located in the western part of Germany, the hotel attracts both business and leisure tourism. The dataset comprises information about the guests, local and foreign, occupied rooms and revenue from 01-2013 to 12-2018. The analysis stated in this report compares the number of local and foreign rooms, finds correlations and compares monthly variance (and periodicity), and does a time series analysis of occupied rooms and revenue.
Keywords: rooms; revenue; timeseries.
Statement of Contribution: As this project was developed in the way of pair-programming, both authors were involved in all parts of the project. Either the data extraction and transformation, as well as the analysis and the final report were partly done by both members as well as overall editing. The knowledge of the python language of the author Torben Kraft, shared throughout the development of the project, played a very important role on the results achieved.

I. Introduction

This project intends to analyse data from an existing hotel using techniques from data science to get deeper insight of a hotel's business performance during the last six years. The data was made available by a hotel[1] located in the western part of Germany, at the edge of a large industrial area with good transportation connections to close cities and the industrial area. Due to several cultural spots and events of interest in and around the location close to the hotel, the attractiveness for leisure tourism increases. The available data comprises information about the number of guests, local and foreign number of occupied rooms and the revenue of the hotel for a time period from 01-2013 to 12-2018. The data is read and cleaned in the beginning, further is inspected with techniques of descriptive statistics and finally the development over time is demonstrated with a time series analysis. The analysis will follow the goal of the approval of hypotheses created by inspecting the data, which are:1 - The hotel is more occupied by local than foreign guests; 2 - Seasonality has a small impact on the occupation due to the business tourism and several cultural events throughout the year close to the area were the hotel is situated; 3 - The revenue of the hotel depends on the occupation of the hotel. An increase of occupation will show an increase of the revenue. Results obtained showed us that hypotheses 1 and 2 did confirmed, while the third one was not linear. Throughout the report we explain the several steps of the analysis, illustrated by the respective graphics from the Jupiter notebook developed. The project was posted on the GitHub repository with the link: https://github.com/TeKraft/DS4HT-Final-Project.

1 – The name of the hotel will not be disclosed due to privacy reasons to not connect the analysis results with the hotel for public.

II. Data

a. Extraction

The dataset itself is saved in several .csv-files containing data for yearly data stored for each month and for each month and year. The data cleaning contains the deletion of empty rows and columns resulting from the reading process of the .csv-files and the inspection of the head and the tail of the data frame. The data frame is of type "pandas Data Frame" in which all the data from the files is combined and easily accessible throughout the analysis. Further the date is created by each year and month during the reading process as well as an index indicating the number of year and month is added to also have a look at classification in the following analysis.

b. Transformation

The data is loaded with values of type String which are type casted to floats and integers and a datetime object is needed for the time series analysis. No further data transformation is needed, because the used data only consists of numbers.

c. Description

The dataset of the hotel comprises information about the number of guests, local and foreign, number of occupied rooms and total revenue for the period 2013-2018. The descriptive summary of the data set shows that for the period of 72 months (01.2013 - 12.2018), on average local guests represent nearly 70% of rooms occupancy. The standard deviation is about 20% (1/5) of the mean for guests and rooms occupation. The revenue by room and by person has a standard deviation of around 10% to 15% which results in the assumption that the price per room changes in a year or in the total period (Figure 1).

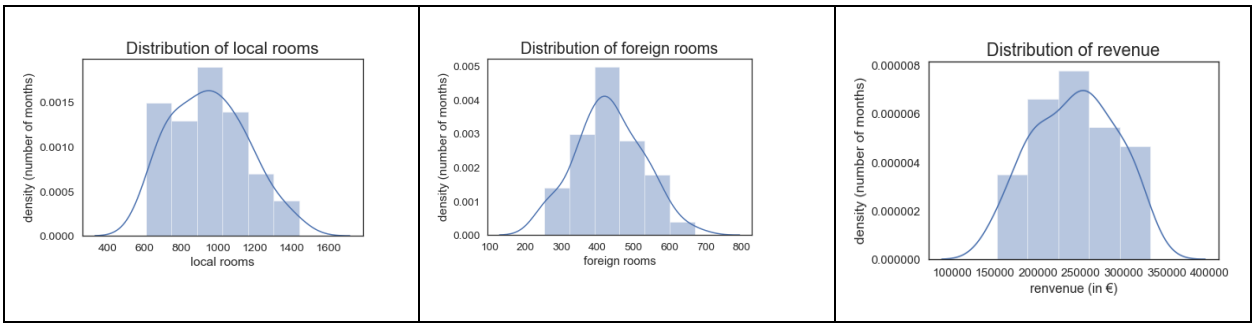
	local guests	foreign guests	local rooms	foreign rooms	revenue by room	revenue by person	revenue hotel	month idx	year idx	total rooms	total guests
count	72.000000	72.000000	72.000000	72.000000	72.000000	72.000000	72.000000	72.000000	72.00000	72.000000	72.000000
mean	1194.944444	497.402778	951.319444	432.111111	176.269167	144.144028	242286.909028	6.500000	2015.50000	1383.430556	1692.347222
std	244.763001	100.751215	204.877072	92.130275	27.615902	21.017044	47356.487178	3.476278	1.71981	234.272793	282.582216
min	706.000000	278.000000	606.000000	252.000000	118.680000	98.830000	151995.400000	1.000000	2013.00000	872.000000	1103.000000
25%	986.250000	425.500000	791.250000	365.750000	152.162500	129.387500	197730.052500	3.750000	2014.00000	1172.750000	1445.750000
50%	1226.500000	505.500000	931.500000	427.000000	178.680000	146.310000	248882.700000	6.500000	2015.50000	1404.500000	1719.000000
75%	1378.250000	572.250000	1109.750000	502.500000	191.420000	158.442500	279602.490000	9.250000	2017.00000	1582.500000	1922.250000
max	1723.000000	741.000000	1441.000000	670.000000	253.850000	194.830000	330680.470000	12.000000	2018.00000	1807.000000	2101.000000

Figure 1 Descriptive summary of the data set used for the analysis

III. Results and Discussion

The analysis results obtained indicate that hypothesis 1 (predominance of local guests) and 2 (small impact of seasonality) proved to be right, while hypothesis 3 (increase on occupation will show an increase of the revenue) showed a different behaviour from what was expected. The distribution plots visualize how the occupation of the rooms in the hotel is counted throughout the time period. The data is nearly normal distributed following a similar pattern of a normal distribution where the normal curve is bell shaped and symmetric about the mean. The range of the local room's occupation is much more with 600 to 1400 than the foreign rooms occupation with 250 to 700. This underlines the statistical description of the data as well, where the local occupation of guests is around 70% of the total guests (Figure 1 & 2). The revenue of the hotel of the last six years has its mean at ~242.300€ with a very high range and a standard deviation lying at 15% of the mean (~50.000€). Figure 3 shows the correlation between each parameter of the dataset. It is clearly visible that there is a very strong correlation between the occupation of the guests and rooms. There is even a higher correlation between the local occupation (local guests and rooms) and the total occupation (total guests and rooms) which is also related to the fact, that circa 70% of the total occupation is local. It would be expected that the revenue of the hotel is somehow dependent on the occupation, there is as well correlation, but interesting is to see less correlation between the revenue of the hotel and the occupation of the hotel with a value of around 0.7 (Figure 3).

Figure 2 Distribution of rooms (local and foreign) and revenue per months



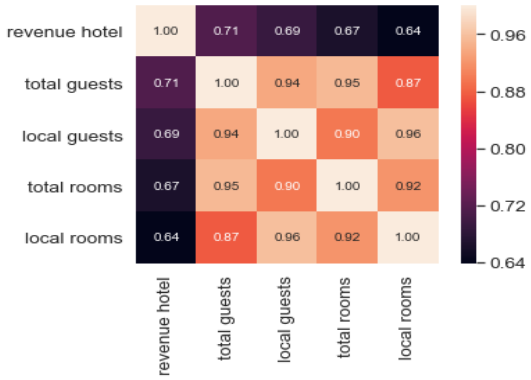


Figure 3 Correlation of each parameter used in the analysis.

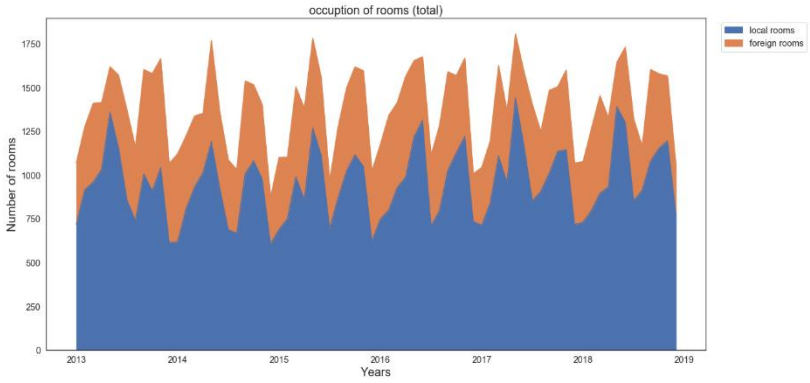


Figure 4 Time series total room's occupation (local vs foreign).

As the distribution and the time series of the guest occupation is following nearly the same pattern as the room's occupation, the rooms occupation will be used for further analysis in this report. The analysis and plots of the guest occupation are visualized in the python script only. Comparing the amount of local and foreign rooms one can see difference in volume and monthly occupation. Further, there is a large deviation of occupation throughout the years. There are a lot of low peaks in the beginning and the middle of the year and high peaks following each low peak. In general, there is no large increase or decrease visible from the separation of local and foreign guests (Figure 4). Similarities between the distribution of rooms occupation over the years show that the hotel was able to maintain a stabilized performance concerning the lodging business throughout the last six years. Highlighted through the correlation and distribution plots, the foreign occupation follows a similar pattern as the local occupation and are just a small part of the whole occupation, around 20% throughout the year. Having a closer look on the relation of the occupation and the revenue, regression plots are used to combine rooms occupation and revenue in total and rooms occupation and revenue by room. These regression plots show each month as a dot containing the value of the total room's occupation and the revenue of the hotel for this month. The dots are classified by the years and coloured in the same colour for each year to compare the yearly development. Further, a regression line created with the method of least squares shows the mean of each year and the trend of the relation of both parameters (Figure 5 and 6).

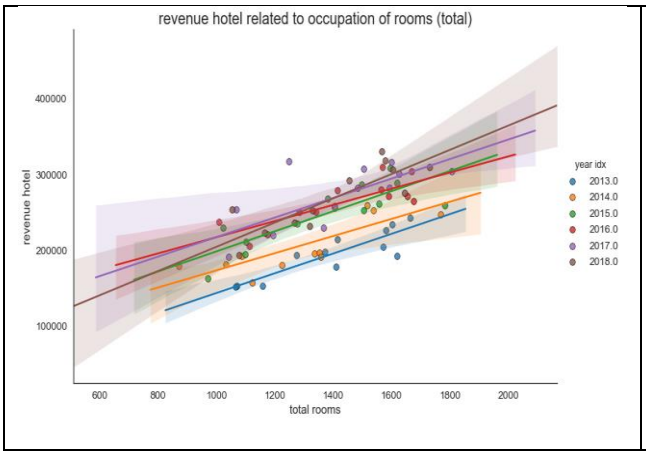


Figure 5 Dependencies of revenue from room's occupation

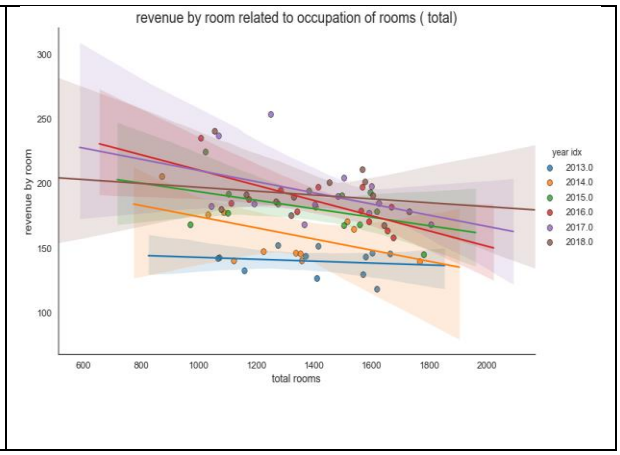


Figure 6 Revenue by room related to occupation of rooms

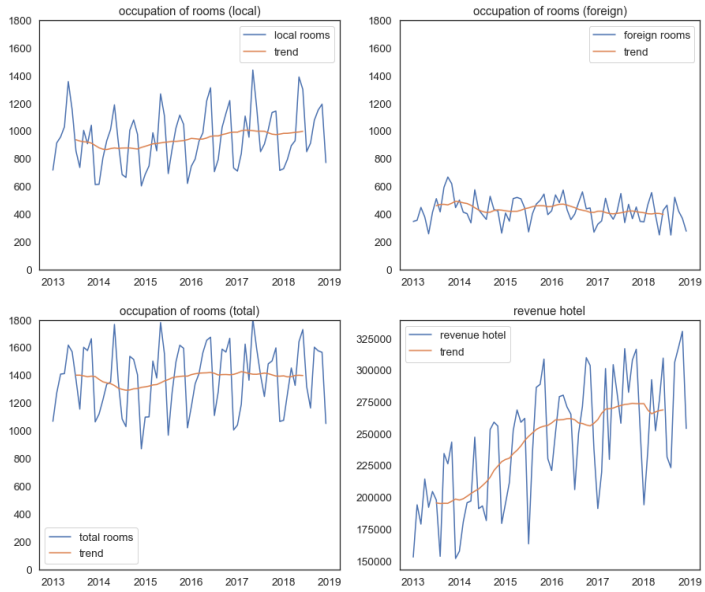


Figure 7 Time series analysis of occupied rooms (local and foreign) and revenue by years.

Concerning revenue from occupied rooms, overall there is an increase over the years, especially noted from year 2014 (orange) to 2015 (green). The pattern of relation between the two parameters is very similar in each year. This plot expresses as well that with a high occupancy of rooms the revenue is high too. Though, there are some "outliers" showing that even with the same amount of occupied rooms the revenue is the same for a month (Figure 5). However,

there is not a significant increase of revenue with the increase of occupied rooms, which shows that the dependence of one parameter from another is not strongly decisive on the overall revenue. Figure 6 shows the relation of the occupation and the revenue by room. Here at first glance a similar pattern of the yearly comparison is visible. Over the past six years the revenue by room has been increased. Further, the regression lines are different from Figure 5. It shows an optical decrease which means that the lower the revenue by room, the more rooms were occupied. This could indicate a decrease of the price of the rooms, which results in an increase of guests, because the price is lower than usual (Figure 6). Comparing both plots, it gets clear that if the revenue by room increases also the total revenue increases and vice versa. Moreover, the increase of the revenue over the year is an important fact as well as that there might be more occupation concerning less price for a room, because the revenue by room decreases with more occupation. Figure 7 shows the time series for each of the analysed parameters (occupation of rooms: local, foreign, total; revenue hotel) with the trend of each. At first glance comparing the occupation and the revenue the trend has a completely different shape. Though, there are a lot high and low peaks throughout each year and over the period, the trend of the occupation remains nearly constant with slight increase in the local occupation and slight decrease in the foreign occupation. But the revenue of the hotel increases to a doubled value over the last six years. This supports the assumption that the revenue of the hotel is not significantly dependent on the occupation, but more on the pricing of the rooms and other factors, which are not analysed in this report.

IV. Conclusions

The insights that were able to achieve from the initial dataset by using data science techniques provide a clear view of the hotel's business and performance over the last six years related to the occupation of the hotel in combination with the revenue. As discovered from the analysis there are similarities between the distribution of rooms occupancy (local and foreign) over the years, showing that the hotel has been able to maintain a stabilized performance concerning the lodging business, though, the occupation of local guests slightly increased while the occupation of foreign guests slightly decreased. This might help the hotel management in trying to find ways to also attract more foreign guests, so that this stabilized performance remains in future years. In general, the time series analysis of occupied rooms (local and foreign), and revenue shows consistency on the behaviour over the years. The revenue tends to increase, with less accentuated losses at the end of the period. Furthermore, based on the regression plots (Figure 5 and 6) the analysis concluded that the pricing of rooms changes in between years and increases over time which affects the revenue of the hotel more than the occupation which remains nearly stable. Deducing, there might be a way in further improve

the calculation of the room pricing to increase either the revenue as well as the occupation of the hotel throughout the year. Additionally, the low peaks during a year could be solved somehow.

V. Acknowledgements

We would like to thank the hotel who helped us in this project. We are very grateful for the collaboration and trust they demonstrated by giving us access to their internal data with the permission to analyse it and do assumptions upon their data for the last six years.

VI. References

Coursera. Online verfügbar unter https://www.coursera.org/, zuletzt geprüft am 18.05.2019.

Grus, Joel (2015): Data science from scratch. 1 ed. Sebastopol, CA: O'Reilly Media. Online verfügbar unter http://site.ebrary.com/lib/alltitles/docDetail.action?docID=11047336.

Kaggle. Online verfügbar unter https://www.kaggle.com/, zuletzt geprüft am 18.05.2019.

McKinney, Wes (2014): Python for data analysis. [data wrangling with Pandas, NumPy, and IPython]. 1. ed., 3. release.

Provost, Foster; Fawcett, Tom (2013): Data science for business. [what you need to know about data mining and data-analytic thinking]. 1. ed.

Python Library. Online verfügbar unter https://www.python.org/, zuletzt geprüft am 18.05.2019.

VanderPlas, Jake (2016): Python data science handbook. Essential tools for working with data. First edition.