Predicting AIRBNB Sales with Google Searches in a Customer Journey Context

Krarup, Mads Zacho; Buus Lassen, Niels; Madsen, Rene

Document Version
Final published version

Published in:
Symposium i anvendt statistik

Publication date:
2018

License
Unspecified

Citation for published version (APA):
Krarup, M. Z., Buus Lassen, N., & Madsen, R. (2018). Predicting AIRBNB Sales with Google Searches in a Customer Journey Context. In P. Linde (Ed.), *Symposium i anvendt statistik: 22.-24. januar 2018* (pp. 32-49). Institut for Fødevare- og Ressourceøkonomi, Københavns Universitet og Det Nationale Forskningscenter for Arbejdsmiljø.
http://nfa.dk/api/PdfRelay/Get?id=http://pure.ami.dk/ws/files/5005424/Linde_P_Symposium_i_anvendt_statistik_2018.pdf

Predicting AIRBNB Sales with Google Searches in a Customer Journey Context

Mads Zacho Krarup, Niels Buus Lassen, and Rene Madsen

Journal article (Published version)

Please cite this article as:
Krarup, M. Z., Buus Lassen, N., & Madsen, R. (2018). Predicting AIRBNB Sales with Google Searches in a Customer Journey Context. I P. Linde (red.), *Symposium i anvendt statistik: 22.-24. januar 2018* (s. 32-49). København: Københavns Universitet.

Predicting AIRBNB sales with Google searches in a customer journey context

Mads Zacho Krarup, Dept. of Digitalization, CBS

Niels Buus Lassen, Dept. of Digitalization, CBS

René Madsen, Dept. of Digitalization, CBS

Abstract

This paper presents a predictive model of Airbnb sales in Copenhagen, build on a dataset with over a 33-months of Airbnb bookings in Copenhagen and related searches on Google. Moreover, geospatial patterns of the AIRBNB listings are detected by constructing a 100 m x 100 m grid cells in UTM 32 coordinates of the city of Copenhagen. The predictive models built are a stepwise regression model, which is compared with a neural network model, both performing well on training as well as on k fold cross validation data with an R^2 value on 86%-96%.

1. Introduction

AIRBNB is a Peer to Peer online marketplace and a homestay platform where travelers around the world, can book short term accommodations in residential properties. AIRBNB is currently represented in 34,000 cities with above 2 million listings worldwide [1]. The equity funding had in June 2015 a total value of 2.3 billion USD, however, AIRBNB has so far kept the information about the total revenue and the number of reservations secret [2]. In Denmark, the government has had a liberal approach to AIRBNB, by allowing Danish residential to sublet their property without taxation to an amount of 24.000 DKK per year [3].

Research Questions:

1: To what extend can search engine data describe the customers journey process in terms of AIRBNB bookings in Copenhagen?

2: What kind of geospatial patterns can be seen computing AIRBNB listings in a 100 m. x 100 m. grid cells of Copenhagen?

2. Briefly on the existing literature

Forecasting in tourism enable business participant or destinations to allocate resources that meets the demand at a period of time [4]. Tourism products are

described as perishable and the demand have traditionally been predicted using time series or econometric analysis. However, research have concluded that no single method is superior to other models. It depends on the evaluation criteria and the data set employed that allow certain models to perform better than others [5]. Some researchers have successfully used Holt-Winters method to forecast daily hotel room demands [6], while others have used dynamic linear models [7] or ARMA models with high predictive accuracy [8]. With the increase of internet searching, researchers have successfully incorporated Google Trend data, to find a high correlation of queries from influenza patients and certain search words on Google Trend in order to generate an accurate forecasting model [9].

In terms of peer to peer research papers, concerning AIRBNB, relatively few has occurred despite the rise of the sharing economy. Analysis from Texas City concludes that low-end hotels, not marked against business customers, are vulnerable due to the shift in the patterns of accommodation booking [10].

3. The data and methodology

3.1 AIRBNB data

The Airbnb dataset provided by AIRDNA [11] contains 20,5 million lines of listings over a period of time from September 2014 to May 2017. Each line in the dataset represent a listing that has been marked with a status of being; (R) reversed, (B) blocked or (A) available on a given day.

The primary focus in this paper is to look at the reserved listings, which in total aggregate to 2.6 million reservations, with in other words is the amount of reservation days in the dataset. 11.391 unique property ID's are detected among the reserved listings. The dataset contains a verity of descriptive parameters, more specific 51 in total, are assigned to each reservation.

The data set host endless of opportunities for further research, however to specify this research the main focus will be on following parameters; Sum of sales which is 288 million USD for the whole period and geospatial coordinates in terms of longitude/latitude.

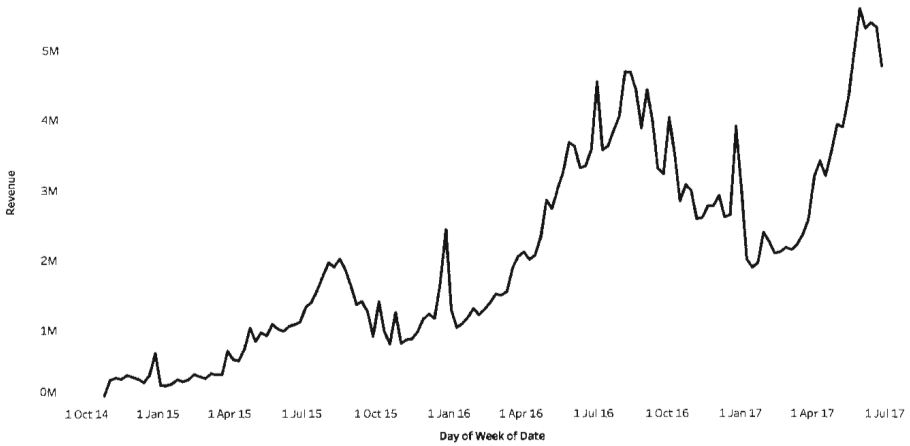

Figure 1: Tableau output, AIRBNB revenue Copenhagen 288 mio USD
From 26.10.2014 – 25.06.2017 [11].

3.2 Google search engine data

Google trend is a public tool that provides an index of search terms relative to the total number of Google searches over time. The index can include multiple queries to obtain a relative comparison among them [7].

The use of google search quires are applied to describe the customer journey corresponding to the reservation of AIRBNB bookings. The idea is to use travel related queries to build a predictive model relevant for the information searching process. More specific in terms of transportation options, accommodation alternatives and for general city information.

Travel related category	Search Queries
• Accommodation	"Hostel Copenhagen", " Hotel Copenhagen", "AIRBNB Copenhagen"
• Transportation	"Flight Copenhagen", "Train Copenhagen", "Metro Copenhagen"
• City information	"Travel Copenhagen", "Sightseeing Copenhagen"

Figure 2: Search Queries from Google Trend [12].

The queries have moreover been translated into 13 languages representing the countries which had the highest number of overnights in Denmark in 2016 [13]. A preliminary analysis indicated that the highest explorative power in terms of building a predictive model is obtained by using English search quires worldwide.

As Google trend only allows 5 queries in one Google Trend Search, the query with the highest index value ("Hotel Copenhagen") are used as a baseline, meaning that the query is always included when applying new search quires on google trend [12].

Travel related category	Search Queries
• Language	English
• Region	Worldwide
• Search time	01.09.2014 to 30.06.2017
• Category	All
• Dates when data gathered	Weekly average

Figure 3: Search settings from Google Trend [12].

The Google Trend index time range from 01.09.2014 to 30.06.2017 due to the possibilities of time lag. This is one of the advantages of using search engine data, as they represent real-time consumer behavior [14].

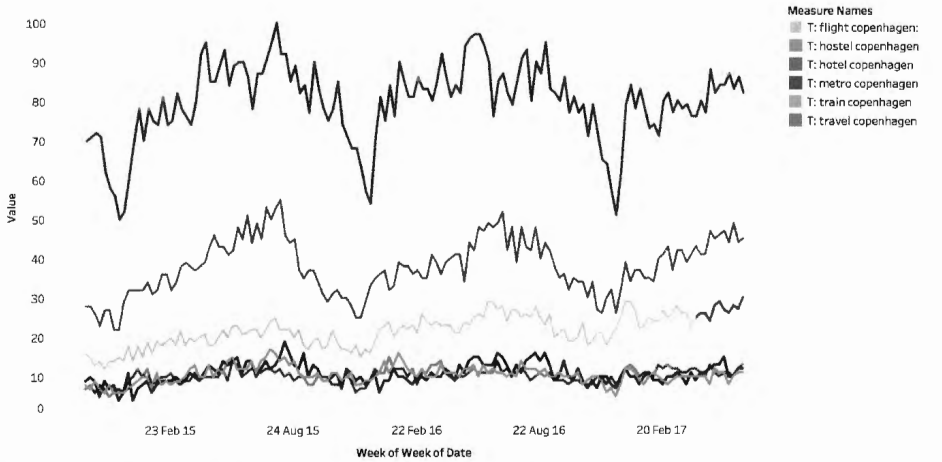

Figure 4: Tableau output, Preliminary Subplot of Google Trend search quires from 20.10.2014 to 19.06.2017 [12].

Report from Emarketer states that 48.4% of the consumers in Canada and USA use search engine when ones begin to research about upcoming trips and 20.4 % uses property websites [15]. According to net market share, Google holds 75.94 % of the current search engine market share [16]. It would be preferred to include all search engines; however, Google is the only company that allow the public to export data from their database. Furthermore, it would be ideal to have a higher percent of usage of search engines, when they are planning research on a trip. However, is it believed that Google Trends with its limitations can provide a significant contribution to the purpose of this paper. Despite the fact that Google Trends may

contain approximation methods in the sample data, which leads to inaccuracies [5].

3.2 Data preprocessing and transformation

In our data preprocessing we have used Alteryx Designer [17], IPython Jupyter notebook [18] and libraries from Scitools [19], Osgeo [20], Sharpely [21] and Matplotlib [22] to compute and visualize UTM 32 coordinates on a map of Copenhagen. Tableau [23] has been used to visualize and explore the dataset and for statistical computing SAS/JMP [24] has been used.

As seen in the preliminary revenue plot [figure 1], trend and seasonality are seen in the graph. In order to deseasonalise and make the trend stationary, a logarithm transformation has been performed of the AIRBNB revenue. Moreover, the data has been aggregated to weekly level, in order to compare the values with the google search quires. The google search quires has as well been time lagged up to 7 weeks back in time. This creates the possibility to describe at what time travelers respond to certain explorative variables.

4. Models for AirBnB sales

4.1 Stepwise regression model

Stepwise regression is a method for developing a regression model, where the selection of input variables is done by an automatic procedure. The method has received a lot of criticism, for overfitting, not adjusting for degrees of freedom and in general generating too many input variables.

We have dealt with overfitting by choosing 20% Hold-out data in Cross 5 fold validation. This automatic function for Cross fold validation in SAS JMP, was actually the main argument for us to choose stepwise regression, as this method in SAS JMP allowed us to use exactly the same 20% Hold-out data in Cross 5 fold validation for both stepwise regression and neural networks. The stepwise regression in SAS JMP is resulting in 21 input variables, which can be considered too many. We defend our 21 input variables as being only 7 Google searches in several timelagged versions. This brings in some intercorrelation in our model, but also allows us to start developing a preliminary Customer Journey model in this article with Google searches over time, linking Google searches on a timeline to different phases in the Customer Journey model. We see the Stepwise regression model as a starting point for comparing multiple regression with neural networks. But as future work we would develop a classic multiple regression model with much fewer input variables, and compare that model to neural networks on the same 20% Hold-out data in Cross 5 fold validation. But SAS JMP did not allow us to do Cross fold validation on a classic multiple regression model automatic.

4.2 Neural Network

Applying artificial neural network on sales modelling could be seen as shooting birds with a cannon, but artificial neural networks have performed very good on many datasets within sales modelling in many cases. So we feel it is a natural and logic process to compare regression models with neural networks within sales modelling. We also chose the neural network modelling approach, because we have an Airbnb dataset with 20,5 million lines of data, and see a good potential in general for the neural network model on such a large and complex dataset in our future work on this dataset.

The neural network used is a standard feed forward network with 1 hidden layer and using back propagation

4.3 Model evaluation, validation and comparison

The above listed models are evaluated by RMSE where the smallest values between the two models is preferred as well as the highest R^2. A K fold cross validation is as well performed.

We made one model for the weekly data with stepwise multiple regression, and K-5fold cross validation on 20% of the dataset. We also applied Neural Network models for comparison, with exactly the same K-5fold cross validation on 20% of the dataset.

5. Models for AIRBNB sales

5.1 Model comparison

The Multiple Regression model obtained a Rsquare on 0,87 on the validation data identifying 21 of 80 Google searches after validation Rsquare optimum method. The 80 Google searches were 10 searches positively linked to Airbnb sales, and then timelagged from 0-7 weeks back in time.

The Neural Network obtained a Rsquare on 0,96 on the validation data – with exactly the same set of 21 Google searches as input variables. So the Neural Network performs significantly better than Multiple Regression on both Rsquare and RMSE, refer to below outputs from SAS JMP.

	Neural Network		**Stepwise regression**	
	Full model	Hold out	Full model	Holdout
Rsquared	0,9465	0,9568	0,9190	0,8721
N	112	28	112	28
RMSE	15,5221%	15,1117%	28,8389%	N/A

Table 5: Summarized output form SAS/JMP from Stepwise regression and Neural Network.

Other combinations of input variables for the Neural Network were tested, but none of them performed better on validation data, than the Neural Network with 21 Google searches as input variables.

5.2 Model results: Stepwise regression

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC	RSquare K-Fold
9,314887	112	0,2883897	0,9190	0,8995	2,2394695	28	91,71697	161,2064	0,8721

Figure 6: SAS JMP output, Multiple Regression on Log of Sales, with 21 Google searches as input variables.

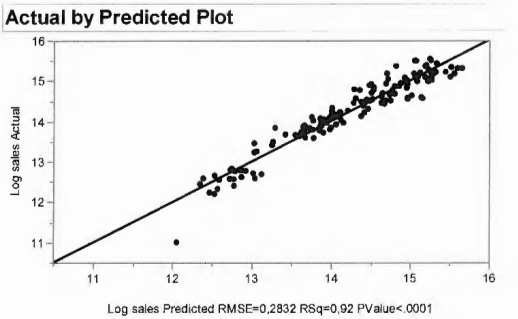

Figure 7: SAS JMP output, Multiple Regression on Log of Sales, with 21 Google searches as input variables.

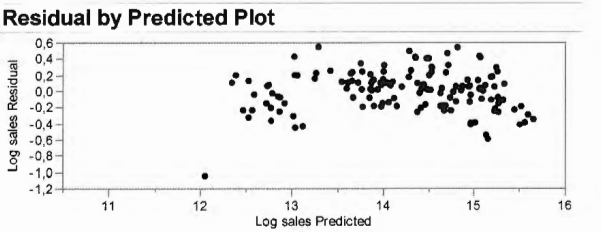

Figure: 8: SAS JMP output, Multiple Regression on Log of Sales, with 21 Google searches as input variables.

5.3 Model results: Neural Network

Training		Validation	
Log sales		**Log sales**	
Measures	**Value**	**Measures**	**Value**
RSquare	0,9464885	RSquare	0,9567919
RMSE	0,2115284	RMSE	0,1814072
Mean Abs Dev	0,1552189	Mean Abs Dev	0,1511751
-LogLikelihood	-15,05924	-LogLikelihood	-8,066025
SSE	5,0113573	SSE	0,9214404
Sum Freq	112	Sum Freq	28

Figure 9: SAS JMP output, Neural Network on Log of Sales, with 21 Google searches as input variables.

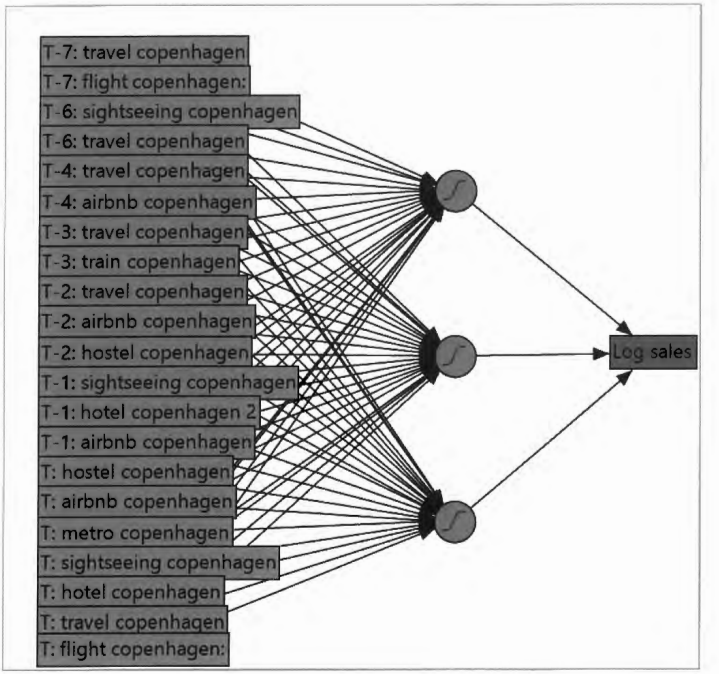

Figure 10: SAS JMP output Diagram for the Neural Network of Log Sales with 21 Google searches as input variables.

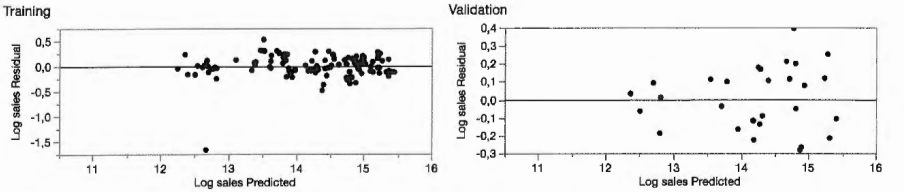

Figure 11: SAS JMP output, Residual by Predicted Plot of Neural Network on Log of Sales, with 21 Google searches as input variables.

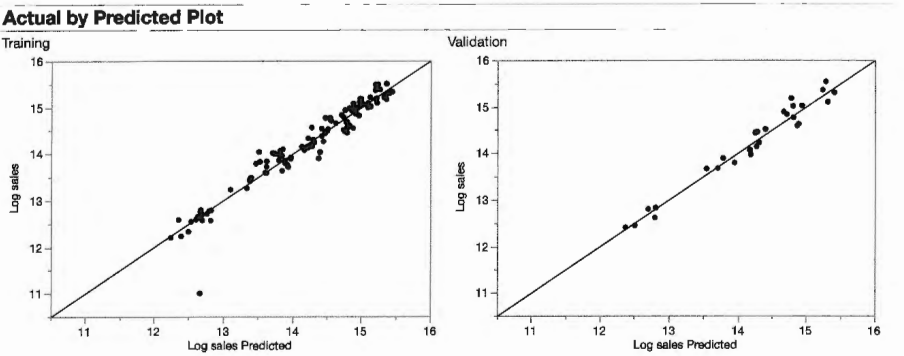

Figure 12: SAS JMP output, Actual by Predicted Plot of Neural Network on Log of Sales, with 21 Google searches as input variables.

5.4 Model interpretation: Customer Journey

The 21 Google searches identified by the stepwise regression model are visualized in below preliminary Customer Journey model.

T-minus 7weeks.	Travel Copenhagen	A cluster of customers starts their customer journey 7 weeks before arriving in Copenhagen,
	Flight Copenhagen	looking at both Flights & other ways to travel to Copenhagen.
T-minus 6weeks.	Travel Copenhagen	A cluster of customers starts their customer journey 6 weeks before arriving in Copenhagen,
	Sightseeing Copenhagen	Here we are closer to purchase, as the customer already starts looking at what can be explored in Copenhagen.
T-minus 4weeks.	Travel Copenhagen	A cluster of customers start their customer journey 4 weeks before arriving in Copenhagen,

	Airbnb Copenhagen		After looking at ways to travel to Copenhagen, many have bought train/flight ticket already here, and then customers start exploring Airbnb possibilities.
T-minus 3weeks.	Travel Copenhagen		This exploring of travel possibilities is both before and after customers have bought Airbnb, Hostel or Hotel.
	Train Copenhagen		This exploring of travel possibilities is both before and after customers have bought Airbnb, Hostel or Hotel.
T-minus 2weeks.	Travel Copenhagen		A cluster of customers starts their customer journey 2 weeks before arriving in Copenhagen,
	Airbnb Copenhagen		After looking at ways to travel to Copenhagen, many have bought train/flight ticket already here, and then customers start exploring Airbnb possibilities.
	Hostel Copenhagen		After looking at ways to travel to Copenhagen, many have bought train/flight ticket already here, and then customers start exploring HOSTEL possibilities.
T-minus 1weeks.	Sightseeing Copenhagen		Here we are closer to purchase, as the customer already starts looking at what can be explored in Copenhagen.
	Airbnb Copenhagen		After looking at ways to travel to Copenhagen, many have bought train/flight ticket already here, and then customers start exploring Airbnb possibilities.
	Hotel Copenhagen		Potential hotel customers find hotel much later in customer journey, compared to Hostel & Airbnb customers who find it in better time.
T, travel week.	Travel Copenhagen		Customers who explore & maybe buy travel with short timespan
	Airbnb Copenhagen		Customers who explore and maybe buy Airbnb with short timespan
	Hostel Copenhagen		Customers who explore & maybe buy Hostel with short timespan
	Hotel Copenhagen		Potential hotel customers find hotel much later in customer journey, compared to average Hostel & Airbnb customers who find it in better time.
	Sightseeing Copenhagen		Customer starts looking at what can be explored in Copenhagen, just before or after they arrived in Copenhagen.
	Metro Copenhagen		Customer starts looking at how to use Metro in Copenhagen, just before or after they arrived in Copenhagen.
	Flight Copenhagen		Customers who explore & maybe buy flight with short timespan

Figure 13: Customer journey process. Summarized findings from significant Google searches in stepwise regression from JMP/SAS.

All Google searches will belong to one of the phases in a Customer Journey model, and this is the explanation why Airbnb relevant Google searches have predictive power up to Airbnb sales. The pattern in the timeline is:

1. explore and maybe purchase train or flight
2. explore and maybe purchase hostel, Airbnb or hotel
3. explore possibilities with Metro and sightseeing

6. Geospatial analysis

To analyse the dataset for geospatial patterns, the WGS84 latitude, longitude coordinates for the bookings are converted into a local UTM32 equivalent coordinate system and then aggregated to 100x100 meter grid cells in UTM32 for a normalized equal area presentation.
The color maps are produced as 5 equal quantiles of 20% to get the best visual representation of the geospatial distribution within the map.
The maps are produced in Jupyer based from 2.6 mill booking data rows in a SQL database.

Press service: It can be possible upon request, to get access to the geospatial maps in color and high resolution. Please contact Niels Buus Lassen nbl.digi@cbs.dk

Figure 14: Output from Jupyter notebook. Avg. Revenue for 100 m x 100 m grid cells.

In figure 14, it can be seen that the highest revenues are in the center of Copenhagen.

Figure 15: Output from Jupyter Notebook. Reservations days by 100 m x 100 m grid cells.

In figure 15, it can be seen that the highest level of reservation days are in center, Vesterbro, Nørrebro, Østerport and the parts of Amager close to the center.

Figure 16: Output from Jupyter Notebook. Price per guest on 100 m x 100 m grid.

In figure 16, it can be seen that the highest level of price per guest is in a little broader center area (compared to figure 14), and also in Islands Brygge.

Figure 17: Output from Jupyter Notebook. Number of Properties per 100 m x 100 m grid cell.

In figure 17, it can be seen that the highest level of properties per grid cell, are in the center, Vesterbro, Nørrebro, Østerport and the parts of Amager close to the center. The pattern it quite similar to figure 15, showing level of reservation days. A figure showing number of guests per grid, is also very similar to this figure 17.

Figure 18: Output from Jupyter Notebook. Sum revenue per 100 m x 100 m grid.

In figure 18, it can be seen that the highest level of revenue per grid, are in the center, Vesterbro, Nørrebro, Østerport and the parts of Amager close to the center. The pattern it quite similar to figure 15, showing level of reservation days, and figure 17 showing properties per grid.

7. Conclusion

We have proven the significant predictive power in relevant Google searches up to Airbnb sales in a Customer Journey context. The predictive power of the Google searches were shown in both a stepwise regression and neural network model, and the two models were compared in relation to modelling and predicting Airbnb sales in Copenhagen with relevant Google searches. Stepwise regression had Rsquare on 0,87 and Neural Network had 0,96 on 20% hold-out data in cross 5fold validation.

By computing 2,6 million of AIRBNB reservation days in a 100 m. x 100 m. grids of Copenhagen, we also showed new insights of the Airbnb data patterns in Copenhagen. Like in many other cities a central location of the property is preferred and the price and avg. revenue per property is highest in the central Copenhagen. However, the areas just outside center, Vesterbro, Nørrebro, Østerport and Islands Brygge (bro areas) has more days of booking and higher density of bookable properties per grid cell than the center resulting in higher total capacity than the center. The result is that the revenue per grid cell is as high in the bro areas as in central Copenhagen.

References

[1] AIRBNB, about us, Accessed 18-12-2016 https://www.airbnb.dk/about/about-us

[2] Statista, AIRBNB, 2016, Accessed 18-12-2016 https://www.statista.com/topics/2273/airbnb/

[3]Skat.dk, Accessed 18-12-2016https://www.skat.dk/SKAT.aspx?oId=1615284

[4] Frechtling, D 1996, Practical tourism forecasting, Elsevier. —— 2001, Forecasting tourism demand: methods and strategies, ButterworthHeinemann.

[5] Song, H, Witt, SF & Li, G2009, Travelers' Use of internet, 2009 Edition, Travel Industry Association of Amercia, Washington D.C

[6] Mihir Rajopadhye et al. Forecasting uncertain hotel room demand, information sciences volume 132, issues 1-4, pages 1-11 (2001) Accessed 18-12-2016 http://www.sciencedirect.com/science/article/pii/S0020025500000827

[7] Roberto Rivera. A dynamic line model to Forecast Hotel Registations in Puerto Rico Using Google Trend data. (2016): Accessed 18-12-2016 https://arxiv.org/pdf/1512.08097.pdf 2016

[8] Pan et al, Bing Pan, Chenguang Doris, Haiyan Song. Forecasting Hotel Room

Demand Using Search Engine Data. 2012 Accessed 18-12-2016 https://pdfs.semanticscholar.org/3dce/ecd911bb8abb2464a70b8db2f0f609c283e4.pdf

[9] (Ginsberg et al) Ginsberg, J, Mohebbi, MH, Patel, RS, Brammer, L, Smolinski, MS & Brilliant, L 2009,'Detecting influenza epidemics using search engine query data', Nature, vol. 457, no. 7232, pp. 1012-4.

[10] G. Zervas, D. Properpio, J. W. Byers, The Rise of the Sharing Economy: Estimating the Impact og Airbnb on the hotel Industry (2016), http://people.bu.edu/zg/publications/airbnb.pdf

[11] AIRDNA: Airbnb dataset from Copenhagen https://www.airdna.com/

[12] GOOGLE TREND SEARCH: https://trends.google.dk/trends/explore?date=2014-09-01%202017-06-30&q=airbnb%20copenhagen,Hotel%20copenhagen,hostel%20copenhagen

[13] Visit Denmark, Status på turisternes overnatninger I Danmark 2016, 2016, Accessed 18-12-2016 http://www.visitdenmark.dk/da/analyse/turisternes-overnatninger-i-danmark-0

[14] Torsten Schmidt & Simeon Vosen, Forecasting Private Consumption: Survey — based indicators vs. Google Trends, Ruhr, Economic Papers #155, 2009, Accessed 18-12-2016 https://core.ac.uk/download/pdf/6326969.pdf

[15] E Marketer, Most Travelers Use Search Engines When Planning a Trip, 2016, Accessed 18-12-2016 https://www.emarketer.com/Article/Most-Travelers-Use-Search-Engines-Planning-Trip/1013745

[16] Net market share, Desktop Search Engine Market Share, 2016, Accessed 18-12-2016 https://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0

[17] Software; Alteryx Designeri , https://www.alteryx.com/

[18] Software; Jupiter, http://jupyter.org/

[19] Python Library; Scitools, https://scitools.com/

[20] Python Library; Osgeo http://www.osgeo.org/

[21] Python Library; Sharpely, https://pypi.python.org/pypi/Shapely

[22] Python Library; Matplotlib, https://matplotlib.org/

[23] Software; Tableau, https://www.tableau.com/

[24] Software; SAS/JMP, https://www.jmp.com/en_dk/home.html