Evaluating the influence of Airbnb listings' descriptions on demand

Bram Janssens[a], Matthias Bogaert[a], Dirk Van den Poel[a]

[a]Ghent University, Department of Marketing, Innovation and Organisation, Tweekerkenstraat 2, 9000 Ghent, Belgium
Bram.Janssens@UGent.Be, Matthias.Bogaert@UGent.Be (Corresponding author), Dirk.VandenPoel@UGent.Be

Abstract

Hosts list their accommodations on Airbnb aspiring to attract guests. Extant research on the drivers of guests' booking behaviour has solely considered structured information on the Airbnb platform, thereby omitting the rich information provided in the unstructured textual listing description. This work adds to the stream of research on Airbnb demand determinants by identifying the latent topics used in these unstructured descriptions as drivers of listing demand. Both our empirical model and follow-up experimental study indicate that Airbnb guests value unique accommodation aspects of which hosts can convince their potential guests by using the textual description. Guests especially value enthusiastic home experiences and a unique local city guide accompanying the listing. However, when hotel-like properties are conveyed in the description, prospective guests are dissuaded.

Keywords: Airbnb, Sharing Economy, Topic Modelling, Latent Dirichlet Allocation, Demand Determinants

1. Introduction

In 2008, the hospitality industry witnessed the entrance of a new disruptive player: Airbnb. Airbnb is an online platform that gives people around the world (hosts) the opportunity to rent out property as a hospitality service for which they receive a fee. Hosts are free to set their own price, of which Airbnb receives a certain percentage. Interested visitors (guests) have the option to check the available offering per geographical region, with the option to use various filters to narrow the offering (e.g., date, price ranges, and accommodation type). Hosts can give details about themselves (host description) as well as about the listing (listing description).

Our research addresses how these listing descriptions influence listing demand, as textual information is one of the elements under control of the host that shows great potential in explaining differences in demand. To this end, we will estimate the listing demand based on the topics included

in the listing descriptions and several control variables. To determine these topics, we propose to use Latent Dirichlet Allocation (LDA; Blei, Ng & Jordan, 2003). LDA tries to find the underlying latent topics in the listing descriptions. After identifying these hidden topics, we determine the topic distribution for each listing description. Next, we regress these topic distributions against listing demand, while controlling for other covariates that were proven to be significant in prior research. To study the impact of each topic on listing demand, we fit a Stochastic Frontier Analysis (SFA; Battese & Coelli, 1995) model. The model is interpreted and used to evaluate whether and how listing topics influence demand.

While academic interest in Airbnb has been growing ever since the platform was founded (Guttentag, 2019), this question has not been previously addressed. As such, we advance current research in several respects. Several studies exist on how guests choose their Airbnb accommodation. These prior studies provide insights into the different aspects that guests consider when selecting their accommodation. Research has identified that both non-influenceable and influenceable aspects drive demand for Airbnb listings. The former encompasses characteristics that cannot be easily adapted, such as location (e.g., Varma et al., 2016) and size (e.g., Gunter & Önder, 2018). The latter are those that can be more easily changed, such as price (e.g., Visser, Erasmus & Miller, 2017), and verified identification (e.g., Abrate & Viglia, 2017). From a strategic perspective influenceable aspects are more interesting to the host. For instance, adding professional photos requires only a small investment, but can drive up demand significantly. Hence, knowing more about the influenceable factors of listing demand is key to the success of an Airbnb listing. The main contribution of this paper is to be situated in the derivation of the topics that are currently used in listing descriptions by the use of LDA. Besides simply deriving the topics, we determine the influence they have on listing demand, thereby further developing the insight in these important influenceable factors. Making changes to these descriptions is not resource nor labour intensive and, yet, they can be seen as potential key differentiators between Airbnb listings' demand.

Secondly, we contribute by validating whether customer segment targeting is applied by Airbnb hosts. The Airbnb customer market consists of multiple diverse segments that have different desires (Guttentag et al., 2018; Lutz & Newlands, 2018). Yet, no prior research has checked whether these segments are addressed accordingly. By deriving the latent topics in the richest signal on the Airbnb platform (i.e., the listing description) we control whether these topics align to the desires of certain previously identified segments and what impact this has on demand.

The potential that the open, unstructured format of the listing description offers to the hosts may also have a potential downside. When too many topics are used in a description, this could confuse prospective guests as they interpret this as an incomplete alignment to their desires. The final

contribution is thus focussed on determining whether such a dispersed use of topics influences demand.

This study yields several interesting findings. First, great variation is detected in the topics used in the listing descriptions. These topics do significantly impact demand, with both the potential to enhance and decrease demand. Guests seem to desire listings sufficiently different to hotel accommodations. Demand-decreasing topics have an effect which is similar to what is observed with the omittance of key information. Second, several topics are linkable to segment targeting being applied by Airbnb hosts. Third, we observe that the usage of multiple simultaneously used topics does not decrease demand, thereby allowing hosts to combine multiple successful strategies. The results also suggest that topics used in listing descriptions are much more diverse in nature and have a more significant impact on demand as their host description counterparts. Finally, we provide evidence for the existence of a 'too-good-to-be-true-effect' (Maslowska, Malthouse, & Bernritter, 2017), where interested guests distrust overly optimistic review ratings.

The rest of the study is organized as follows. Section 2 presents a review of current academic literature and its shortcomings. Section 3 includes the used mathematical models. Empirical results are presented and discussed in Section 4. To validate the representativeness of our results, robustness checks are performed in Section 5. Section 6 elaborates on a follow-up experimental study. We conclude the study, with its implications and limitations during Section 7.

2. Literature review

While Airbnb was only founded in 2008 (Guttentag, 2019), it heavily impacted the hospitality industry due to its enormous growth. Within only a decade, it changed the entire industry and became one of the largest players in hospitality (e.g., Zervas et al., 2017; McGowan & Mahon, 2018). Its ever-growing market share and importance also attracted a large number of researchers leading to an exponential growth in the number of annually published papers (Guttentag, 2019).

The majority of research on Airbnb focussed on the determinants of guests' decision making and their impact on demand. That is, when guests book a listing through Airbnb, this can be seen as a transaction with asymmetric information (Mauri et al., 2018). The potential guests have nearly no prior knowledge about the listing, while the host knows everything about his or her own property. In such a case, Signalling Theory (Spence, 2002) suggests that the party with (nearly) complete information tries to convince the opposite party by giving a limited number of signals about the true value of their product/service. In the case of Airbnb, we observe that the host tries to convince potential guests by the information available on the listing's webpage. This page is essentially a collection of signals. This collection of signals is critical for the competitiveness of a listing in the market. A host needs to know which signals to use in order to be competitive.

Many important signals constitute non-influenceable aspects such as facilities (e.g., Mauri et al., 2018), location (e.g., Varma et al., 2016) and capacity (e.g., Xie & Mao, 2017). However, of even greater importance to hosts are the influenceable aspects as their corresponding strategic range of choices is not limited to deciding upon which information to share, but also on what and how to share this information. For instance, setting a fair price (e.g., Visser et al., 2016) or deciding how many pictures to display (e.g., Gunter & Önder, 2018) can easily influence demand. Knowing which of these low-effort changes significantly enhance listing demand enables hosts to get the most out of their accommodation's potential. The use of advanced statistical models and web data (e.g., Abrate & Viglia, 2019) facilitated the research around several possible signals used on the listing page. Previous studies (Xie & Mao, 2017; Gunter & Önder, 2018; Mauri et al., 2018; Abrate & Viglia, 2019) leveraged this and analysed all feasible predetermined information fields such as Superhost status, number of bathrooms, and listing capacity. Unfortunately, the list of predetermined fields is finite, limiting the range of researchable aspects. This was solved by recent studies, researching various easy-to-encode aspects such as photo quantity (e.g., Gunter & Önder, 2018).

Nevertheless, one potential highly-influenceable signal has been consistently overlooked: the textual description of the listing. The listing description shows huge potential for reducing information asymmetry as it allows a host to provide very detailed information to the guest about the property across a variety of topics. This allows the hosts to differentiate their property from nearby competitor listings and convey their unique selling points in a much more nuanced way than predetermined information fields.

However, the diversity and unstructuredness of the listing description makes it more complex to analyse. This translates into a high required level of pre-processing for the information to be usable in research. For instance, Mauri et al. (2018) added textual information from the *host* description (as opposed to *listing* description) and argued that storytelling in the host description led to higher popularity. Their sample was limited to 502 listings and their approach required human intervention (i.e., coding of this information into mathematical features), making it not feasible to be applied on large sample sizes and thereby limiting its generalisability to other data sets. These difficulties probably resulted in no prior study addressing the influence of listing descriptions on listing demand, while previous research proved that large differences exist in listing description topics (Lutz & Newlands, 2018). Therefore, the main objective of this study is to determine the various topics used in Airbnb listing descriptions and whether they have the potential to influence demand.

Determining the drivers of Airbnb listing demand is further complicated by the fact that Airbnb guests are not one homogeneous group, but rather a heterogeneous population with different customer segments (Guttentag et al., 2018; Lutz & Newlands, 2018). Hence, it is important to identify the differential drivers of listing demand for the different customer segments. For instance, Guttentag

et al. (2018) segmented Airbnb customers based on their primary motivation to use Airbnb. They came up with five customer segments with varying demographics. *Money savers* are mainly motivated by low cost, whereas *home seekers* want to have a more 'homely' experience than going to a hotel. *Collaborative consumers* are drawn to the sustainable aspect of the sharing economy. *Pragmatic novelty seekers* and *interactive novelty seekers* are attracted towards the novelty and innovativeness of the Airbnb concept. The difference between both lays in the fact that *pragmatic novelty seekers* are looking for a 'homely' experience (similar to *home seekers*), whereas *interactive novelty seekers* are interested in the social interactions with others. These findings have direct impact on our study as the different motivations of the different customer segments are opportunities for hosts to use in their descriptions. Closely related to this, Lutz and Newlands (2018) investigated whether hosts were aware of market segmentation within Airbnb and were positioning their listings accordingly. They found that shared room listings have a specific focus towards price sensitive customers and are thus more focused towards the *money savers* segment. However, the study did not check if the segment-focused descriptions led to an increase in demand nor did it analyse the topics of the listing in a quantitative way. It is plausible that targeting a certain specific market reaches more guests from that segment but does not increase the overall reach. Therefore, a second objective of this study is to check whether customer segmentation is applied by hosts and whether these segment-focused descriptions influence listing demand.

A downside of targeting segments is that these segments are inherently different and possibly value opposing things. Hence, it could be that a host targets too many segments, as a multi-segment strategy may send confusing signals to interested guests, thereby making the listing less attractive than listings from hosts that are targeting one specific segment. This would be operationalized as having a listing description that belongs to various topics at once. Therefore, a third and final objective is to investigate whether including too many topics simultaneously leads to a decrease in demand.

3. Methodology

3.1. Data handling

For this study, we used publicly available data on Airbnb listings as is provided by the online provider Inside Airbnb. The data was scraped from the Airbnb website during 4 and 5 December 2019. This means that all used data is information as it displayed on Airbnb platform at that moment in time. For instance, the number of reviews represent the number of received reviews between the initial date the listing was created on Airbnb and December 4 (or 5, depending on moment of scraping). For our empirical analysis, we focus on the Airbnb listings in San Francisco, a popular tourist destination and the location where Airbnb was founded. A first analysis of the 8,533 listings in San Francisco revealed that several hosts have multiple listings. These hosts are usually professional entities, renting out over

250 listings. Interestingly, we noticed that some of these 'industrial hosts' had the exact same description for either all or a part of their listings. This could be a general description about the company as well as a generic description of the building if multiple listings resided in the same building. To prevent topics from being biased towards these descriptions, we dropped all exact duplicates from the data, leaving only the first occurrence of the description. After visual and statistical exploratory analysis other special cases that could induce noise in the models were removed as well: listings from one professional host that had very similar descriptions and listings consisting of simple string repetition. This resulted in a further reduction towards 7,330 listings under observation. This reduction and subsequent processing steps are visualized in Figure 1. Note that our final SFA model is restricted to 5,580 observations since some listings did not receive any review or changed their minimal night policy. However, we deliberately chose to model the topics on the bigger sample as research has shown that the performance of topic models asymptotically increases with the sample size (Canini, Shi & Griffiths, 2009). Our motivations to use different samples for our topic model and SFA model is further explained in Section 3.2.

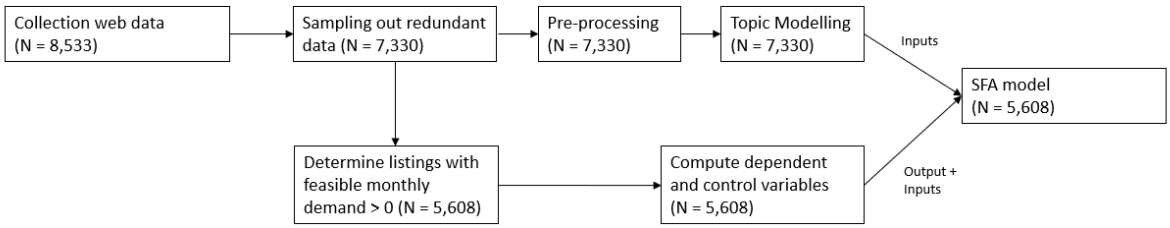

Figure 1: Overview main research methodology

For our topics modelling, we only used textual information from listing descriptions (as opposed to host descriptions). Many hosts had multiple listings with unique listing descriptions but retained the same personal host description. As a result, the number of unique observations for the host description topic model was limited. As topic modelling requires large amounts of unique data (Tang et al., 2014), the host description was not well-suited in our case. However, to inspect the impact on demand it was added in a robustness check (see Section 5).

To perform topic modelling on the listing descriptions, we used Latent Dirichlet Allocation (LDA; Blei, Ng & Jordan, 2003). LDA is part of a subgroup of topic models, known as mixed-membership models. As opposed to assuming that a document contains one topic, they assume a document to simultaneously belong to several topics. The distribution of various words across documents is assumed to result in the distribution of various topics across those documents. Simply said, documents that have similar words are assumed to have the same topic(s).

Before fitting the LDA model, extensive pre-processing is required. The LDA model needs a vocabulary of words per document (i.e., listing description) as an input. First, we counted all the unigrams, which are one-word phrases. However, if we would simply use single words (i.e., unigrams), the model would not pick up constructions such as 'san francisco' or 'square feet'. To overcome this issue, we also added bigrams (i.e., a sequence of two words). To account for infrequent items, bigrams should at least occur 5 times in the overall corpus and meet the following criterium: $\frac{(bigram-5)}{w_1 * w_2} * l >$ 10, with bigram the occurrence of the bigram, w_1 the occurrence of the first word in the bigram, w_2 the occurrence of the second word, and l the number of unique words in the vocabulary. These criteria ensure the inclusion of popular bigrams while filtering out very popular unigrams coincidentally occurring after each other. No threshold was used for unigrams as it is argued that each word (unigram) on its own is a meaningful textual construct. To make sure that our model discovers meaningful hidden topics and does not learn noise in the data, several other pre-processing steps are required: punctuation removal, lower case conversion, stop word removal, and matching 'sf' with 'san francisco'.

Mathematically, the corner stone of the LDA model is the fact that it assumes that documents consist of latent topics, thereby assuming that the documents are in fact probability distributions of those latent topics (Blei, Ng & Jordan, 2003). This probability distribution is modelled as $\mathrm{Dir}(\alpha)$, with $\mathrm{Dir}(\alpha)$ being a Dirichlet distribution with symmetric parameter α and K the number of topics. To derive these topics from the descriptions, it assumes each description to be a distribution of words (the previously mentioned unigrams and bigrams). LDA assumes that these words can be linked to various topics. Topics on their turn are characterized by a probability distribution of words, following a $\mathrm{Dir}(\beta)$ distribution, with $\mathrm{Dir}(\beta)$ being a Dirichlet distribution with symmetric parameter β and number of dimensions equal to the number of unique words (unigrams and bigrams) in the vocabulary. Both distributions are then fitted to the data for parameters K, α, and β using an online variational Bayes algorithm (Hoffman, Bach & Blei, 2010). This allows for fitting the model for various candidate settings within reasonable computation time. Table 1 gives an overview of the candidate settings of the parameters. The candidate values are optimized for the C_V value as proposed by Röder, Both and Hinneburg (2015). The C_V value is a combination of prior developed coherence measures and tends to have near-human performance on topic coherence (Röder et al., 2015). In essence, the C_V value evaluates if various words that contribute to a topic coherently lead to the same topic being identified. This evaluation was done for both 75% and 100% of the data, to ensure the best average fit and thus making it less dependent on data-specific structures.

Table 1: Candidate Settings LDA model

Parameter	Candidate Settings
K	{2, 3, 4, …, 20}
α	{0.001, 0.201, 0.401, 0.601, 0.801, 1, 1/K}

β	{0.001, 0.201, 0.401, 0.601, 0.801, 1}

Optimal performance was obtained with K = 16, α = 0.201, and β = 0.801. After refitting the model with these parameters, a C_V = 0.588 was obtained, indicating a very good fit. Based on the 20 words that contribute the most to each topic (depicted in Appendix A1), we derived the meaning of each topic. Following the assumptions of LDA, each listing description can then be assigned with a probability of belonging to each topic.

3.2. Variables and analysis

We created 16 variables (one for each topic) that represent the amount that each description belongs to each of these topics according to the probability distribution $\mathrm{Dir}(\alpha)$. Note that this method automatically causes topic variable values to lay in the range (0,1). Topic 8 was removed from the analysis as this was the most common topic and its inclusion would cause multicollinearity. To check if having a dispersed listing description influences demand, we created the variable *Topic dispersion.* *Topic dispersion* is operationalized as $\frac{1}{\sigma_{topics}}$, with σ_{topics} the standard deviation of the topic probability values (including topic 8).

Airbnb has a standardized format for entering listing description, with a general section and three important subsections: the space, guest access, and other things to note. Interestingly, out of our sample of 5,608 used listings, many omitted at least one of these fields. More specifically, information on the Space was omitted 701 times (12.5%), while Guest Access was missing 1,416 times (25.2%), which was still lower than the high missing rate of Other Things To Note, with 2,029 omitted values (36.2%). The impact of omitting such information could be important, as this strongly enhances information asymmetry and should theoretically dissuade guests from booking the listing. Therefore, we also added indicators whether such a field was empty (referred to as omittance indicators).

These 19 topic-related variables (15 topics + 3 omittance indicators + topic dispersion) are then regressed against guest demand. Demand was approximated by calculating the minimum demand that occurred by multiplying the minimum nights to stay by the average number of monthly reviews:

$$Demand = Minimum\ Nights\ x\ Average\ Reviews\ per\ Month \quad (1)$$

For example, consider a listing that has been online for 2 months (from October to December 2019), received 8 reviews and has a minimal night policy of 3 nights, the estimated monthly demand in our case would then become 12 nights = $3\ nights\ minimal\ x\ \frac{8\ reviews}{2\ months}$. We acknowledge that this methodology approximates (minimum) demand. However, several popular insights websites such as AirDNA or InsideAirbnb define their occupancy rates in the same way. Other studies defined their own proxies for demand. For instance, Abrate and Viglia (2019) randomly selected 1,014 listings in 5 cities and then closely monitored the availability during the subsequent month, with non-availability interpreted as a booking. We chose not to follow this method since it is a simple approximation of

demand, very time-consuming and limits both the sample size and observed time period. Therefore, we decided to use a method that also approximates demand but can easily be implemented on a large scale. Other studies aggregate constructs such as popularity instead of demand as their dependent variable (Mauri et al., 2018). We believe that demand is more relevant as a dependent variable since demand is directly linked to revenue. Revenue is observed as the most important motivation for hosts to be active on the Airbnb platform (Karlsson & Dolnicar, 2016; Visser et al., 2017). Directly modelling revenue was infeasible as price was used as an explanatory variable. As our metric is closely aligned to AirDNA and InsideAirbnb's definitions of demand, it also has a direct link to key performance indicators used by actual hosts in the industry.

One downside of this approach is the fact that some hosts changed their minimum night policy, resulting in a large number of reviews per month and a high number of minimum nights. This caused some observations to have an unreasonably high monthly demand. These observations were excluded from the sample. Since we are interested in explaining guest conduct, we also only considered premises that were already rented (i.e., already received a review). This led to a further sample reduction to 5,608 observations. As can be seen in Figure 1, all listings were used for fitting the LDA model as this model requires large amounts of textual data for representative results, while listings that had unreasonably high or no demand, were excluded from the regression analysis.

Several control variables were added to the model to ensure valid conclusions. *Location* and *price* were identified by most studies as critical determinants (e.g., Visser et al., 2017). *Location* was operationalized as the distance in km to Fisherman's wharf, a major tourist location in the city centre. *Price* was computed as the average price per person if the full capacity of the listing was used, with cleaning fees included. Service fees were excluded as they are booking dependent. Host responsiveness was also included by using the *host response rate*. A dummy variable indicating whether the host had *Superhost* status (Mauri et al., 2018) was included as well, alongside average *review rating* (Xie & Mao, 2017). Listing *capacity* (Xie & Mao, 2017) was operationalized as the maximal number of guests. Following Abrate and Viglia (2019), following variables were added as well: indicators for *entire home*, *washer/dryer*, and *host identification*. In case of missing data, the variable value was imputed by the average value of this variable for observable cases (Howell, 2007). This resulted in the following variables as described in Table 2.

Table 2: Used Variables

Variable	Description
Ln(Demand)	Monthly demand of listing
Topic1	Location focus, city centre
Topic2	Hotel
Topic3	Luxury
Topic4	Rules & pricing mechanism

Topic5	City centre hotels
Topic6	Housing projects or hotels with amenities
Topic7	(Premium) amenities listing
Topic9	Chinatown location
Topic10	Focus on key elements
Topic11	Homely feeling in quiet neighbourhood
Topic12	Enthusiastic home experience
Topic13	Using listing as residence + according legislation
Topic14	Business focus
Topic15	Focus on building structure & event offering
Topic16	Neighbourhood touring
Topic dispersion	Topic dispersion
Space	Indicator whether space description field was omitted
Guest Access	Indicator whether guest access description field was omitted
Notes	Indicator whether 'Other things to note' was omitted
Review rating	Average review rating
*Capacity**	Maximum number of guests listing can accommodate
*Location**	Distance to Fisherman's Wharf
Superhost	Indicator if host has Superhost status
Response rate	Host response rate
*Price**	Listing price per person at full capacity
Hots identification	Indicator whether host has verified his identity
*Washer/Dryer**	Indicator whether washer/dryer was provided
*Entire home**	Indicator whether the listing is an entire home (vs. private/shared room)

** denotes variables assumed to contribute to the efficiency frontier*

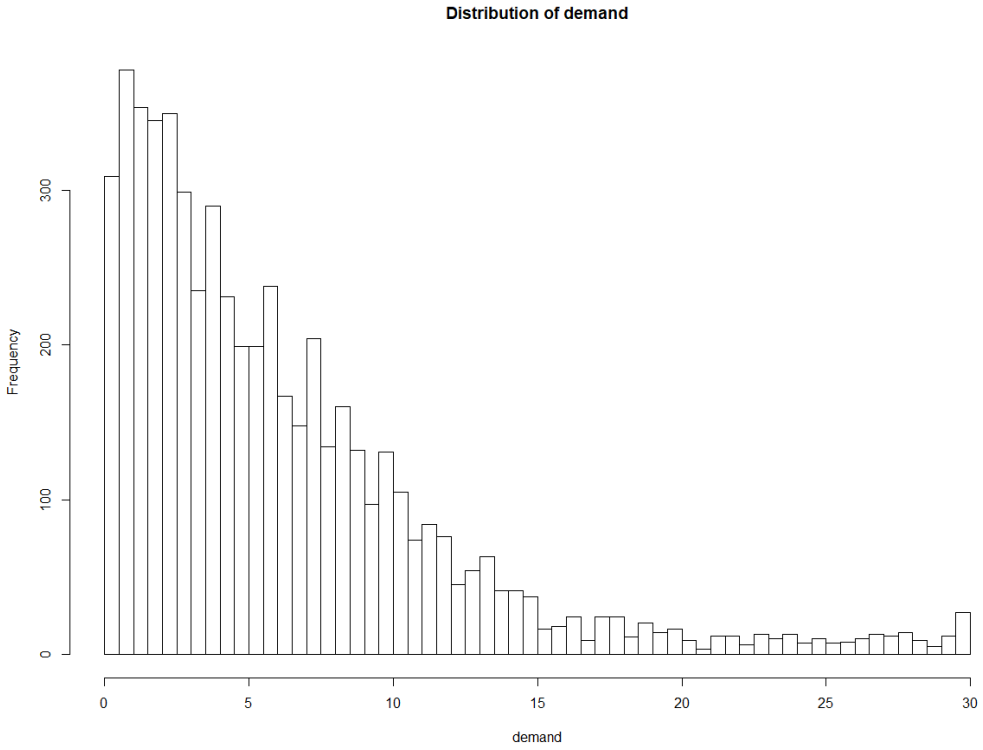

Figure 2: Distribution of demand

To estimate demand, we use a stochastic frontier analysis (SFA; Battese & Coelli, 1995) model. The choice of this approach is motivated by previous research that claims its superior performance in this setting (Abrate & Viglia, 2019)[1]. SFA models the efficiency frontier of a listing, which is how much demand it could be generating based on its specifications, such as price and listing size. This efficiency frontier is then compared to host-specific variables that are assumed to hinder the listing from generating this demand (inefficiencies). For instance, a host using the wrong topic in his/her description could scare of interested guests, which is regarded by SFA as an inefficiency since it hinders the host from generating his/her potential demand. The aforementioned information asymmetry is a large driver of inefficiency as potential guests are not capable of making fully informed decisions. Following Abrate and Viglia (2019), we include capacity, price, location, entire home, and washer/dryer in the frontier model.

Figure 2 depicts the distribution of listing demand. Since the data is highly skewed towards low values, we take the natural logarithm of demand $\ln(demand)$ as dependent variable rather than $demand$. For the SFA model, we model the relationship between demand and covariates according to Eq. (2) and Eq. (3), where the vector of control values for listing *i* is split up in control variables influencing the efficiency frontier ($\boldsymbol{ControlSpecs_i}$) and variables influencing inefficiency ($\boldsymbol{ControlOther_i}$). $\boldsymbol{Topic_i}$ stands for the probability scores the listing description received for each

[1] A standard linear model (OLS) was also tested, but performed worse with an AIC of 16264.50 compared to the SFA's AIC of 15731.26.

of the determined latent topics, as well as the indicator whether the description was a mix of various topics.

$$\ln(demand_i) = f(\boldsymbol{ControlSpecs_i}) - Inefficiency_i + v_i \qquad (2)$$

$$Inefficiency_i = \boldsymbol{\beta_{other}}\, \boldsymbol{ControlOther_i} - \boldsymbol{\beta_{topic}}\, \boldsymbol{Topic_i} + \sigma_i \qquad (3)$$

To ensure no multicollinearity issues, variance inflation factors were computed with satisfactory maximal values around 1.5. As there was no evidence for positively skewed residuals in the SFA model (γ_{σ_v} = -0.766), the SFA model could be interpreted.

4. Empirical results

4.1. Demand determinants

Table 3: Results SFA model

VARIABLE	COEFFICIENT	STANDARD ERROR	P-VALUE
FRONTIER MODEL			
(Intercept)	2.6775	0.0508	p<0.0001***
Capacity	-0.0992	0.0071	p<0.0001***
Price	-0.0005	0.0001	p<0.0001***
Location	0.0125	0.0053	0.0187**
Entire home	0.2450	0.0274	p<0.0001***
Washer/Dryer	-0.0522	0.0268	0.0516*
MODEL OF INEFFICIENCY			
Topic1	-0.0879	0.2188	0.6879
Topic2	4.8734	0.5777	p<0.0001***
Topic3	4.7061	1.0782	p<0.0001***
Topic4	2.1697	1.4216	0.1269
Topic5	1.5275	0.3844	0.0001***
Topic6	1.9147	0.2812	p<0.0001***
Topic7	0.3207	0.6366	0.6144
Topic9	3.9023	0.5696	p<0.0001***
Topic10	0.3266	0.7652	0.6695
Topic11	-0.0654	0.1313	0.6185
Topic12	-4.0020	1.8825	0.0335**
Topic13	2.1628	0.5995	0.0003***
Topic14	-1.2584	2.4224	0.6034
Topic15	0.8415	0.3979	0.0344**
Topic16	-2.0535	0.6408	0.0018***
Topic dispersion	-0.0340	0.0405	0.4007
Superhost	-1.1986	0.0965	p<0.0001***
Review rating	0.0089	0.0037	0.0166**
Response rate	-0.0097	0.0034	0.0045***
Space	0.3641	0.1263	0.0040***
Guest access	-0.1927	0.1032	0.0619*
Notes	0.2771	0.0893	0.0019***
Host identification	0.6577	0.0908	p<0.0001***
σ_u	1.5001	0.0478	p<0.0001***
σ_v	0.5412	0.0170	p<0.0001***

λ	2.7717	0.1089	p<0.0001***
Mean Efficiency	0.4381		

****p-value<0.01, **p-value<0.05, *p-value<0.10*

Table 3 summarizes the results of the SFA model. When looking at the coefficients of the control variables we notice some opposing effects. For instance, the results suggest that review rating has a positive effect on inefficiency, thereby hindering demand. While we might expect that higher review ratings will convince guests, a possible explanation could be the presence of the 'too-good-to-be-true-effect' (Maslowska et al., 2017). This effect states that customers value some fraction of nonperfect reviews, as otherwise it may seem that all reviews are created by acquaintances of the host. To test for this non-linear effect and other non-linear effects, we implemented a Generalized Additive Model (GAM; Hastie & Tibshirani, 1990) with thin plate regression splines for all continuous predictors. This type of model allows modelling of highly non-linear relationships. In essence, a GAM is a generalized linear model that assumes a flexible relationship between an additive set of predictors and the dependent variable. Each predictor can be transformed using a smoothing spline (in our case thin plate regression splines). By doing so, a GAM combines both flexibility with interpretability (Wood, 2004). The resulting model performed worse than the SFA model (AIC = 15867.80). Interestingly, a visual inspection of the regressed spline of review rating in Figure 3 does indicate a downwards sloping bend in the curve at very high values, indicating the presence of a 'too-good-to-be-true-effect'. This while, for lower values, there seems to exist a positive effect between review rating and listing demand as these review ratings correspond to believable customer evaluations. This effect may be even more pronounced for Airbnb when compared to other settings. Airbnb listings have high average review ratings (Zervas et al., 2017; Guttentag, 2017), with an average rating of 95.80% in this case. When all listings have high review ratings, the positive effect of a high rating may be limited, while the suspicion towards overly positive reviews remains.

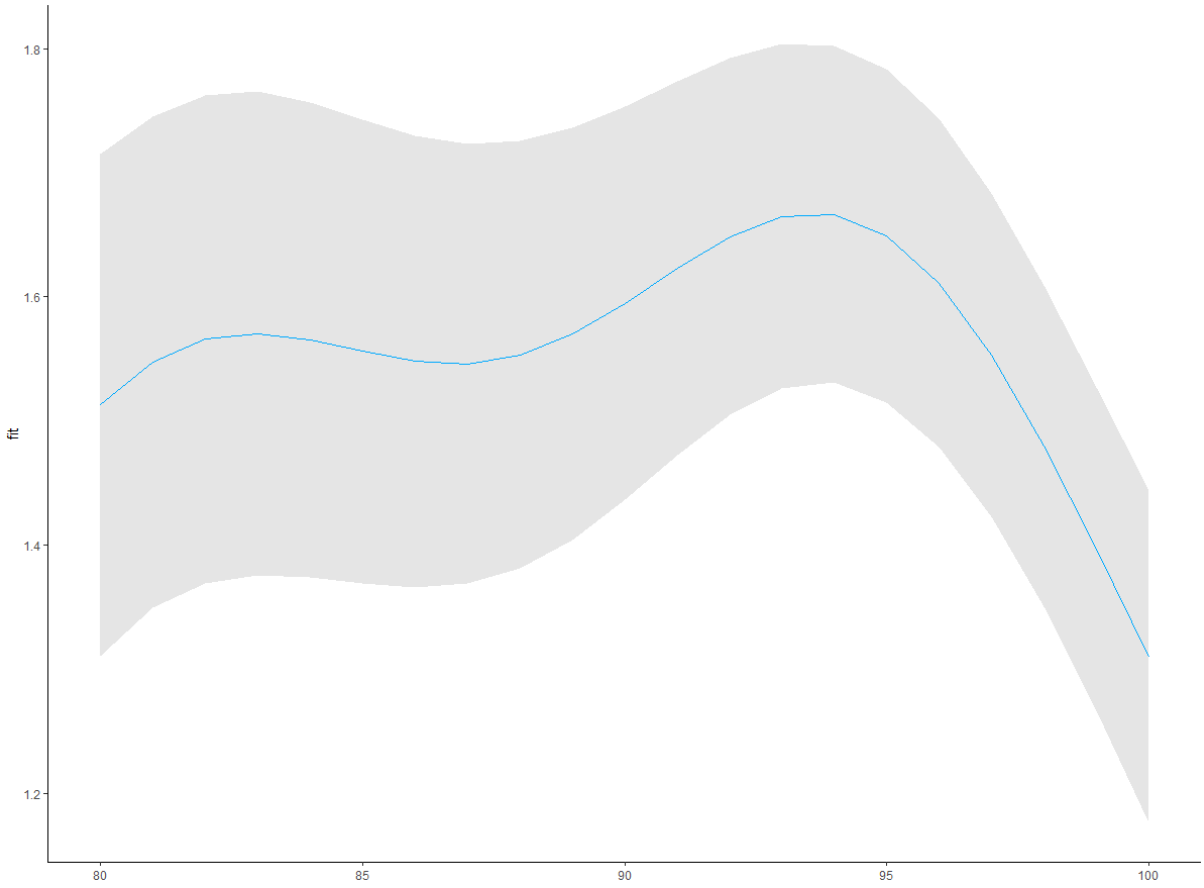

Figure 3: Fitted regression spline of 'Review rating' variable in GAM model[2]

The effects of capacity and washer/dryer are also interesting to point out, as they are opposed to what was observed in Abrate and Viglia (2019). This difference can be explained by the nature of the dependent variable. Abrate and Viglia (2019) modelled revenue, whereas we model the monthly demand for a listing. This is an interesting distinction, as capacity shows to be positively impacting revenue, but negatively impacting demand. Simply said, you will have less bookings, but will make more money out of it.

This reasoning holds less for the washer/dryer amenity. A possible explanation could lie in the fact that this type of amenity could be linked to long-term residential stays. This type of listing seems to be of lesser demand in our sample, as suggested by the effect of topic 13. We should, however, be cautious as the variable is only significant on the 10% significance level.

With regard to location, we find evidence in our SFA model and by inspecting the important words in our LDA model, that Fisherman's Wharf is not the primary interest towards tourists. Location does significantly impact the efficiency frontier, but the effect is positive. This means that locations not near

[2] The figure only depicts the range (80,100) to give a more clear visual inspection of the relationship within a commonly observed range of ratings.

Fisherman's Wharf are in higher demand than locations near it. This also translates to the important dimensions of various topics, which include various city locations (e.g., Union Square, Presidio), but no mention of Fisherman's Wharf. This means guests do not value proximity to Fisherman's wharf and that hosts are aware of it.

The most puzzling effect was the one observed for host identification, as host identification reduces information asymmetry. It thereby should instigate further trust in the host and decrease inefficiency. However, the opposite effect is observed. A possible explanation could be that hosts who fail to generate enough demand, try to enhance their demand and use host identification to do so.

The main research objective of this paper is to assess the effect of listing descriptions on demand. Interestingly, 7 out of 15 observed topic variables are significant on the 1% significance level, and 2 more on the 5% significance level. This demonstrates that listing descriptions do influence guests in their decision making.

Also worth noticing is the fact that most signs for the significant topics are positive. Only topics 12 and 16 have a negative sign. This means that most atypical descriptions (i.e. not topic 8) increase inefficiency and lower demand. Topics 12 and 16, however, decrease inefficiency and positively impact listing demand. Topic 12 is the 'enthusiastic home experience' topic (for a detailed description we refer the reader to Appendix A2). Guests are potentially drawn to the enthusiastic way a homely accommodation is described, which is a clear distinction from the more formal interactions at traditional accommodations.

Topic 16 also decreases inefficiency and thereby increases demand. In descriptions scoring high on this topic (see Appendix A2), we see hosts acting as tour guides for the neighbourhood. It is thus useful to inform interested guests about the neighbourhood. It is likely that guests have no or limited experience in the city and by explaining them what your neighbourhood and its adjacent areas have to offer, you can further convince guests to stay at your residence. The specific focus on public transportation options is especially noticeable. This positive effect can again be explained based on Signalling Theory. By including information about the neighbourhood, users get the necessary signals that the host truly lives in the area and is a city-local. Hence, this increases the authenticity and trust of the user in the renter.

Both enthusiastic home experience (topic 12) and neighbourhood touring (topic 16) are topics that are very distinct to experiences offered in the traditional hospitality industry. They can be interpreted as signals that a guest will be experiencing a 'true' Airbnb experience, rather than a 'substitute hotel'. This indicates that guests value experiences that are truly unique and different to what is understood under classical hospitality management and search this (among other desires) through the signals indicated in the textual descriptions of listings. This also shows that Airbnb remains inherently different

when compared to other accommodation options. This confirms earlier work that highlights how Airbnb guests place high importance on unique experiences (Paulauskaite et al., 2017).

However, most topics increase inefficiency: topics 2, 3, 5, 6, 9, 13, and 15 hinder the demand for a listing and hosts should thus avoid these topics. Especially topics 2 (hotel) and 3 (luxury) have very high coefficients. This demonstrates that San Francisco Airbnb guests are not interested in reading about the special amenities a location has to offer and are more interested in what the neighbourhood has to offer. Topics 5 (city centre hotels) and 6 (hotels and housing projects), also increases inefficiency, expressing the mismatch of hotels with the Airbnb market segment. The unpopularity of these topics can also be linked to the previous statement that Airbnb remains inherently different when compared to other accommodation options. Luxury is traditionally linked to the high-end hotel market segment. Previous research (Guttentag & Smith, 2017) indicates that this high-end segment is receiving the least direct competition by Airbnb. This also translates into our observation that Airbnb guests do not value amenities linked to the hotel industry, especially the ones linked to the more upscale segments.

Topic 9 also has a very high positive coefficient. Guests seem to value other tourist neighbourhoods more than Chinatown. This shows that it is crucial for hosts to know what guests value in the city. Hence, researchers should continue determining the interests of guests across the globe, as these interests are location-specific and should be addressed accordingly.

The positive coefficient of topic 13 hints that the residence market remains significantly different from the Airbnb market, while topic 15 hints that Airbnb guests have limited interest in the building properties of a listing or in the event offering.

The addition of the omittance indicators for certain description proves useful as the omittance of 'Space' and 'Other things to note' increases inefficiency. From a theoretical point-of-view this behaviour could be expected as omittance of important information enlarges the information asymmetry between host and guests. Surprisingly, the omittance of 'Guest access' has an inefficiency-reducing effect. This field often contains information on limitations to guest access, possibly discouraging interested guests. However, the effect is only significant on the 10% significance level and has a much smaller coefficient than the two other indicators, so the results should be interpreted carefully. Overall, open communication seems to be advised as omitting certain description fields generally leads to an increase in inefficiency and decrease in demand.

4.2. Market segmentation & topic dispersion

With regard to the second research objective, we observe some clear market segmentations in the topics. For example, topics 2, 3, and 7 focus on a more premium segment. By targeting a segment that is underrepresented in the total guest population (i.e., the premium segment), hosts will receive less bookings than by targeting the entire population (i.e., use a generic description). Of course, it could

well be that some of these segments lead to higher margins and need less demand in order to generate sufficient profit.

The LDA model clearly picks up the listing descriptions targeted towards some of the segments identified by Guttentag et al. (2018). Topic 4 (rules & pricing mechanism) seems to be targeting the *money savers*¸ while topics 11 and 12 (more homely feeling in quiet neighbourhood & enthusiastic home experience) can be seen as a call towards the *home seekers* and *pragmatic novelty seekers*. The *interactive novelty seekers* receive less attention but could be drawn to the event offering aspect of topic 15. The only segment that is not linkable to any of the topics from the LDA model, are the *collaborative consumers.* Based on the data, it seems that this eco-focused segment is not being targeted adequately in San Francisco, which is especially noticeable given the progressive image of the city.

With respect to the effect on demand, we see that the effect is rather neutral, with topics 4 (*money savers)* and 11 (*home seekers* and *pragmatic novelty seekers*) having no significant effect and topics 12 and 15 having opposing effects. These results should be interpreted with regard to the other topics as discussed above. Most topics decrease demand by giving potential guests information and signals they do not desire, thereby increasing inefficiency. By targeting the correct segments you will receive the amount of customers your listing can generate based on its specificities (i.e., be efficient), but if you target underrepresented segments (e.g., premium segment) or fail to make a clear link between the segment and your description (i.e., *interactive novelty seekers* and topic 15), you will generate less demand than feasible. Topic 12 seems to differentiate itself here by the enthusiastic language usage.

Interestingly, topic 14 (business) does not increase inefficiency, hinting that the *business* segment is also present in the Airbnb population, while not being observed in earlier research (Guttentag et al., 2018). This could signify an emerging trend of more business people using Airbnb accommodations. This could signify an emerging trend of more business people using Airbnb accommodations. Caution should be in place when interpreting these results as the situation might be specific to San Francisco. The San Francisco Travel Association (SFTA) partners up with Airbnb and markets the city as a leisure, convention and business travel destination (San Francisco Travel Association, 2015)[3]. As a consequence, Airbnb could be growing as a business traveling platform due to the efforts of SFTA.

Some of these segments can be targeted without causing significant differences in demand. This means that hosts can differentiate their concept without getting rewarded or punished for doing so, leaving them relatively free in their business conduct. They can choose freely which segments to target based on their personal desires and assets. This remark can only be made with regard to the demand these segments generate, as it is likely that not all segments will be as profitable to each host as others.

Finally, there is no evidence to support the claim that using multiple topics simultaneously influences demand as the *topic dispersion* variable is observed to be highly insignificant. This offers opportunities to hosts as they can combine various successful strategies in their description. A host could for instance give an enthusiastic home experience and guide about the area.

5. Robustness check

To make sure that the results are not due to the SFA model overfitting on the data, we fit a GAM with the same variable list as used in the SFA model (Table 2). Remember that in the GAM implementation, there is no assumed parametric relationship between dependent and independent variables, which allows highly complex nonlinear relationships, while leaving open the possibility of statistical inference testing due to the model's additive nature (Hastie & Tibshirani, 1990). In that regard is a GAM perfectly suited as a robustness check since it can identify for each variable whether it should be ideally modelled as a linear relationship or a non-linear relationship. Accordingly, a GAM allows us to test whether the shapes fitted by the SFA model are correct. We add the results of the GAM for the significant topic variables as depicted in Figure 4. The figure depicts the fitted relationship between topic probability (independent variable, depicted on x-axis) and demand (dependent variable, depicted on y-axis).

Conclusions drawn from the SFA model are somewhat tempered by the regression splines of the GAM model. Highly non-linear relationships seem to be absent. Topics 2 (hotel) and 16 (neighbourhood touring) exhibit the exact relationship to demand as modelled by the SFA model, validating our conclusions. The relationship between topic 12 (enthusiastic home experience) and demand is less pronounced due to the wider confidence bands. However, topic 12 does follow the demand-enhancing pattern observed in the SFA model. Only topic 3 (luxury) and 5's (city centre hotel) curves are opposed to what was observed in the SFA model. This could indicate some interest in luxury and hotels by Airbnb guests, but caution is in place given the large confidence bands. Other topics are identified as of lesser influence towards demand by the GAM model. In general the GAM puts more nuance on the observed statistically significant relationships in the SFA model. It seems to reaffirm the positive impact of neighbourhood touring and (to a lesser extent) enthusiastic home experience descriptions, but does (partially) question the negative effect of hotel-like properties. We added some typical examples per topic in Appendix A2 as we consider these our main outcomes.

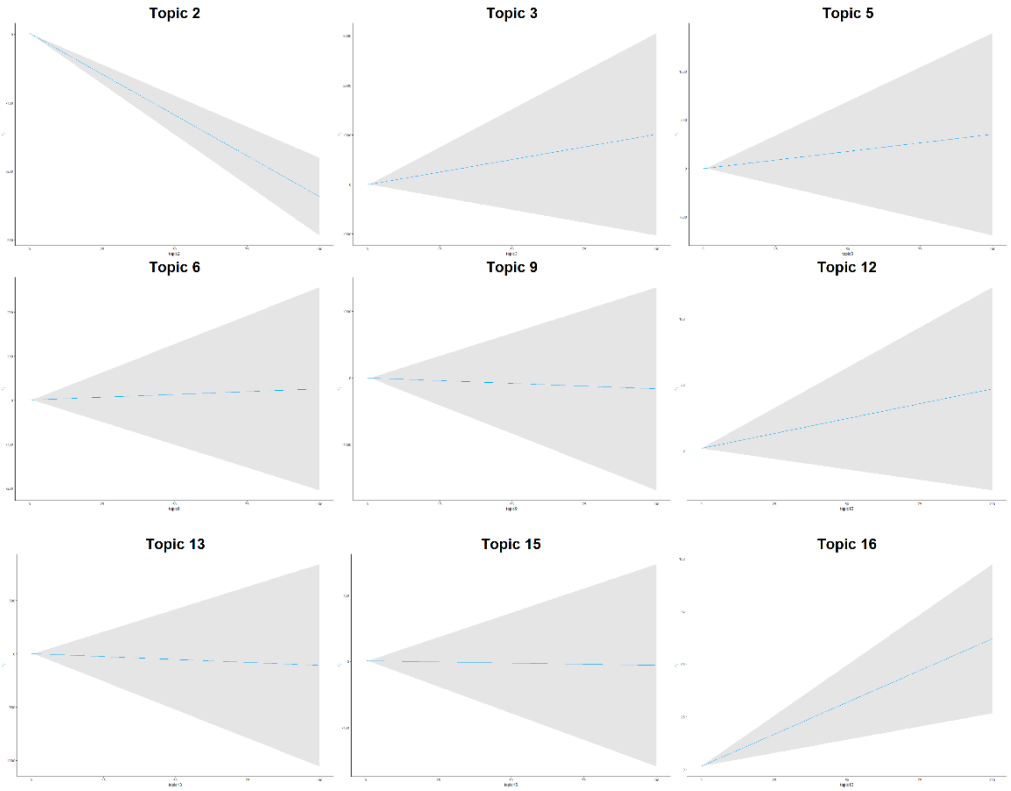

Figure 4: Fitted regression spline of topics determined by SFA model as significant.

Other studies (e.g., Mauri et al., 2018; Abrate & Viglia, 2019) state the importance of personal reputation alongside product reputation. Product reputation was captured in this study by the textual information from the listing description. However, we did not include personal reputation as determined by the personal host description given the limited variance in these descriptions. To test whether host description does not alter our results, we repeat the analysis with host description topics (optimized with the same candidate settings from Table 1). The derived optimal number of topics was 2 (α = 0.5, β = 0.801, C_V = 0.770), showing very limited variation in what hosts put in their personal descriptions. This can be explained by the large number of industrial hosts having at least two listings on the market, validating our choice to leave the variables out of the main analysis. While distinctions are less fine-grained, the topics can be split up between personal descriptions by authentic hosts and more professional descriptions, with the personal descriptions being more prevalent ($\overline{x}$= 0.844) than the business descriptions ($\overline{x}$= 0.155).

As the two topic variables should sum to one, they are perfectly correlated, which is why only the amount of *business description* (i.e., probability of host description topic 2) was added to the model as control variable. Correlations to other predictors remained limited, with topic 15 being the only predictor having a correlation higher than 0.2 (ρ = 0.242), indicating the distinctiveness of the measured construct.

Adding the *business description* variable to the model had little effect. When comparing the coefficients of the variables used in the main model (i.e., those stated in Table 2), no changes where

noted both with regard to the magnitude of the observed effects as well as to the significance levels. The added business description variable, on the other hand, was highly insignificant (p > 0.5), indicating personal descriptions to be of little impact in our sample.

As a third, and final check, we have created the various topics per predetermined subfield (i.e., 'Guest access', 'Other things to note', and 'Space') using the same parameter settings as used for the overall topics (see. Table 1). All received optimal coherence score for the parameters α = 1/K, β = 0.801. This resulted in 6 topics for 'Guest access' (C_V = 0.762), 5 topics for 'Other things to note' (C_V = 0.758), and 6 topics for 'Space' (C_V = 0.719). The related topic probability variables (minus a reference topic for each field) were added to the SFA model as extra control variables, of which the results are summarized in Table 4. Multicollinearity was more problematic with several VIFs exceeding the VIF ≤ 2.5 threshold, indicating considerable collinearity (Johnston, Jones & Manley, 2018). Therefore, results were interpreted as a robustness check rather than as main effect.

Table 4: Results extended SFA model

VARIABLE	COEFFICIENT	STANDARD ERROR	P-VALUE
FRONTIER MODEL			
(Intercept)	2.7037	0.0502	p<0.0001***
Capacity	-0.0964	0.0072	p<0.0001***
Price	-0.0005	0.0001	p<0.0001***
Location	0.0087	0.0053	0.1034
Entire home	0.2426	0.0290	p<0.0001***
Washer/Dryer	-0.0486	0.0269	0.0707*
MODEL OF INEFFICIENCY			
AccessTopic1	-0.2224	0.1710	0.1933
AccessTopic3	0.1826	0.1528	0.2320
AccessTopic4	0.4500	0.2515	0.0736*
AccessTopic5	-0.5592	0.2701	0.0384**
AccessTopic6	0.0853	0.5772	0.8825
NotesTopic2	-0.1165	0.2048	0.5692
NotesTopic3	0.2095	0.2499	0.4019
NotesTopic4	-0.6083	0.2221	0.0062***
NotesTopic5	0.5891	0.2440	0.0158**
SpaceTopic2	-1.5147	0.4515	0.0008***
SpaceTopic3	-0.5239	0.1634	0.0013***
SpaceTopic4	-0.0673	0.1314	0.6086
SpaceTopic5	0.0747	0.4130	0.8565
SpaceTopic6	-2.4051	0.5868	p<0.0001***
Topic1	-0.0457	0.2196	0.8350
Topic2	4.9297	0.5957	p<0.0001***
Topic3	4.2784	1.0578	0.0001***
Topic4	2.0995	1.3905	0.1311
Topic5	1.7659	0.3855	p<0.0001***
Topic6	2.7874	0.3243	p<0.0001***
Topic7	0.9499	0.7169	0.1852
Topic9	4.5309	0.6064	p<0.0001***

Topic10	0.1081	0.7144	0.8797
Topic11	-0.0989	0.1460	0.4984
Topic12	-3.8779	1.6118	0.0161**
Topic13	4.4046	0.7976	p<0.0001***
Topic14	-1.5145	2.4822	0.5418
Topic15	0.6177	0.6791	0.3630
Topic16	-1.8874	0.6499	0.0037***
Topic dispersion	-0.0061	0.0142	0.6684
Superhost	-1.1511	0.0856	p<0.0001***
Review rating	0.0111	0.0039	0.0046***
Response rate	-0.0091	0.0034	0.0072***
Space	0.4540	0.1337	0.0007***
Guest access	-0.1403	0.1162	0.2271
Notes	0.1810	0.1323	0.1712
Host identification	0.5855	0.0852	p<0.0001***
σ_u	1.4500	0.0433	p<0.0001***
σ_v	0.5317	0.0158	p<0.0001***
λ	2.7271	0.1100	p<0.0001***
Mean Efficiency	0.4331		

****p-value<0.01, **p-value<0.05, *p-value<0.10*

The results in Table 4 are interesting, as they confirm the relationships proposed by the main SFA model. Slight deviations occur, with location no longer estimated as a significant covariate. All topic probability variables remain to have the same effect direction, as well as similar significance levels. Interesting to note is how topics 5 and 6 experience a significant reduction in estimated coefficient value, which confirms the nuance put on these variables in the GAM model. Topics 2, 12, and 16 are once again confirmed as having a large and significant impact on demand.

The effect of the pre-determined fields is nuanced by this robustness check. While the omittance indicators had (marginally) significant effects for all indicators in Table 3, we now only observe the omittance of 'Space' omittance to be of significant influence. The elevated importance of this subfield is also reflected in the coefficients of the respective topic probability variables. While the topic variables of 'Guest access' and 'Other things to note' have limited effect size compared to the general description topic variables, this does not hold for 'Space' topics 2 and 6, which is why they will receive special attention. Space topics 2 and 6 are labelled as 'Luxury amenities' and 'Legislation in order', respectively. Examples are added in Appendix A3. Note how these are closely related to general description topics 3 and 13, which further indicates the possible issues with multicollinearity. It is interesting, however, how luxury has a positive effect in the predetermined 'Space' field. The relationship between luxury and demand seems to be more complex, which was already indicated by the inconclusive regression spline.

6. Follow-up experimental study

Overall, the results indicate a significant influence of topics 2, 12, and 16 on demand. To validate these findings, we checked these outcomes in a controlled experimental setting. Respondents were asked to indicate how likely they were on a 7-point Liker scale (from extremely likely to extremely unlikely) to book four accommodations: one without a description (control), one that scores high on topic 2 (hotel), one that scores high on topic 12 (enthusiastic home experience), and one that scores high on topic 16 (neighbourhood touring). A value of 1 corresponds to 'extremely likely' while 7 corresponds to 'extremely unlikely'. The used descriptions were actual descriptions used in our San Francisco sample. Note that our control group is information asymmetry-enhancing and thus not truly neutral. A full overview of the questionnaire is provided in Appendix A4. The order of the questions was randomized to overcome possible order effects. Our final analysed sample consisted of 116 participants of which 100 (86.2%) already used Airbnb as accommodation means. The average age of the respondents was equals to 28.3 years and 61 (52.6%) respondents were female.

On average, enthusiastic home experience ($M_{topic12}$ = 3.02) and neighbourhood touring ($M_{topic16}$ = 2.60) were preferred to hotel descriptions (M_{topic2} = 3.91) and the control condition ($M_{control}$ = 3.88). Since responses were not normally distributed, we tested for statistical differences using the non-parametric Friedman test ($p < 0.0001$), followed by pairwise comparisons using Wilcoxon tests with Bonferroni corrections for repeated testing. The results are reported in Table 7. All groups were significantly different on the 5% significance level, besides the couples enthusiastic home experience - neighbourhood touring ($p = 0.0550$) and hotel – control ($p > 0.5$). In general, we can say that using either the enthusiastic home experience or the neighbourhood touring approach enhances listing demand, while using no description or a hotel-related description will reduce demand. This validates our earlier findings.

Table 7: Results Wilcoxon tests

Group 1	Group 2	W	Adjusted p-value
Control	Topic 12	3126	0.0006***
Control	Topic 16	3690	p < 0.0001***
Control	Topic 2	2100	p > 0.5000
Topic 12	Topic 16	2412	0.0550*
Topic 12	Topic 2	908	p < 0.0001***
Topic 16	Topic 2	872	p < 0.0001***

****p-value<0.01, **p-value<0.05, *p-value<0.10*

7. Conclusion

7.1. Theoretical contribution

This work contributes to the growing field of literature on Airbnb interactions by investigating textual listing descriptions and their influence on listing demand. The main research objective was to check the influence of listing description on listing demand. We show that hosts use various topics in their descriptions and that these topics have various outcomes on demand. This study is therefore the first to show the impact of unstructured signals on Airbnb demand.

Three topics are identified as main influencers on listing demand in our setting: hotel descriptions, enthusiastic home experience, and neighbourhood touring. The first topic reduces demand, while the latter two enhance demand. Note that these two are inherently very different to a classical hotel setting, while the hotel description of course is very closely linked to hotel accommodations. This would signify that a lot of the guests in our San Francisco sample chose Airbnb because of the unique nature of the accommodation rather than as a substitute of hotels. Airbnb guest were previously identified as valuing unique accommodation aspects (Paulauskaite et al., 2017). This observation seems to hold while the Airbnb business concept matures. Hosts should be aware of this distinction and design and signal their accommodation as being unique. Guests could be drifted away by the impression of having a hotel-like experience.

The omittance of important fields in the descriptions, or even the entire description, dissuades guests as well because of the enhanced information gap and thereby induced risk. In general, listing descriptions are a richer source of information than other preformatted structured fields provided on the Airbnb platform and should receive careful attention of hosts.

A second objective was to investigate possible segment targeting used in descriptions. While several topics could be linked to certain previously identified segments (Guttentag et al., 2018) as well as new ones, the impact on demand was limited compared to omittance of information or to usage of topics 2, 12, or 16.

It was also identified that using multiple topics in a description does not lead to decreased demand. This implies that a host can target a segment that is best suited to his or her business concept, while

using demand-enhancing techniques such as enthusiastically signalling a homely experience and/or acting as virtual guide.

In accordance with previous literature on online retailers (Maslowska et al, 2017), we found evidence of the existence of the 'too-good-to-be-true-effect' in Airbnb reviews, where guests seem to distrust overly positive ratings which makes the relationship between the average review rating and listing demand negative.

Interestingly, when controlling for host description topics next to the mainly investigated listing description topics, we found host descriptions to be of little to no influence. This could be explained by the limited variance in San Francisco host descriptions. If this observation would be confirmed in other settings (e.g., other geographical locations), this would signify that listing description are more important in Airbnb guest decision making than personal host descriptions. This counterargues previous findings that personal reputation is equally important to product reputation (e.g., Mauri et al., 2018; Abrate & Viglia, 2019). This could imply two things: (1) previous studies did not include the correct information on product reputation as textual information from listing descriptions was not included, or (2) textual descriptions also act as a form of personal description since they reflect the host.

7.2. Managerial implications

What is written in a listing's textual description clearly impacts the demand for that listing. Hosts should be aware of that and should know which strategies are beneficial and which ones are harmful. Signalling true Airbnb experiences are valued by guests, while hotel descriptions or information omittance decrease demand. Hosts can even combine various strategies, given the fact that using multiple topics at once does not lead to significant changes in demand.

Since we confirm the presence of a 'too-good-to-be-true-effect', we encourage hosts not to spend efforts or funds in synthetically enhancing their average rating, since the overall effect is rather detrimental than beneficial. This openness is also advised with regard to listing descriptions, as the omittance of information has the potential to reduce demand to a similar extent as using undesired topics. Hence to sum up, open communication about your listing towards the guests pays off.

Section 4.1 clearly indicated that several of our outcomes deviate from previous literature. For example, host identification seems to have lost its demand-increasing effect in our analysed sample, probably due to the fact that this has become standard policy for Airbnb hosts. Given these deviations, we believe that managers should keep track of changing preferences in customer demand. Such preference changes are common to any market and plausibly even further aggravated by global shocks such as the COVID-19 pandemic. The pandemic is likely to have changed the used descriptions, with perhaps a larger emphasis on related topics such as contactless transfers and hygiene. A similar shift seems possible to have occurred with the customer preferences.

The study shows the methodological feasibility of our approach, which uses latent Dirichlet allocation to derive topics from textual descriptions of products. The detected topic distributions are then included in a stochastic frontier model, which uses a proxy of demand as dependent variable. The detected patterns are also validated in a follow-up experimental study. While the approach can be used to measure the influence of above described shocks such as the COVID-19 pandemic, the approach can also be translated to other settings, thereby identifying the important aspects of a product's textual description. This could potentially identify unknown drivers of customer choice. The range of applications is wide, with any full-text type of hospitality marketing element being a valid candidate. One could think of very similar applications to Airbnb such as ride-hailing and food delivery platforms (e.g., Uber).

7.3. Limitations

A first limitation is that the results are based on San Francisco only. It may well be that different effects are observed in other cities. Outcomes may not only vary per setting, but also over time as customer preferences might change, as discussed in 7.2. Managerial implications.

This is currently of even greater importance, facing the global COVID-19 pandemic. The influence of this event is not considered in our study, as the used sample is from a period where the pandemic had no impact on the considered geographical area.

Next, we often refer to the difference between demand and revenue as an explanation for some observed outcomes. In principle, revenue should be the main goal, but does not consider cost structure. Also, we can only use historical revenue data, which does not allow for price experimentation. Nevertheless, it is plausible that in some situations a higher price could have been asked, while retaining maximal bookings.

Despite its limitations, this study is the first to examine the effect of topic usage in Airbnb listings' textual descriptions on demand, while empirically evaluating the targeting of certain customer segments. Accordingly, we believe our study makes a significant contribution to extant literature.

7.4. Future research

As the study was solely focused on one geographical area, future research should perform a cross-city examination of these topics and assess their impact on demand. Since we are the first to incorporate listing descriptions, we believe that our analysis is significant and that our work can stimulate practitioners and researchers to perform a similar analysis for other cities. Furthermore, this could also be beneficial to validate the observed limited importance of host descriptions.

A similar analysis could be repeated on data which includes the period of the COVID-19 pandemic. Such a study could identify which aspects are becoming more prevalent during different phases of the pandemic and different regulations (e.g., travel restrictions, and group size limitations). These possible

patterns can be identified in both the textual descriptions, as well as in the customer responses to these descriptions.

One of the above discussed limitations considers the discrepancy between our used dependent variable (i.e., demand) and revenue. In order to facilitate Airbnb hosts in conducting more profitable Airbnb transactions, it seems interesting to conduct a study on profit maximization for Airbnb listings, as was also deployed in the hotel industry (e.g., Choi & Cho, 2000).

References

Abrate, G., & Viglia, G. (2019). Personal or product reputation? Optimizing revenues in the sharing economy. *Journal of Travel Research*, 58(1), 136-148.

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017*). When should you adjust standard errors for clustering?* (No. w24003). National Bureau of Economic Research.

Battese, G. E., & Coelli, T. J. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical economics*, 20(2), 325-332.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Canini, K., Shi, L., & Griffiths, T. (2009, April). Online inference of topics with latent Dirichlet allocation. *Artificial Intelligence and Statistics* (pp. 65-72). PMLR.

Choi, T. Y., & Cho, V. (2000). Towards a knowledge discovery framework for yield management in the Hong Kong hotel industry. *International Journal of Hospitality Management*, *19*(1), 17-31.

Gunter, U., & Önder, I. (2018). Determinants of Airbnb demand in Vienna and their implications for the traditional accommodation industry. *Tourism Economics*, 24(3), 270-293.

Guttentag, D. (2017). Regulating innovation in the collaborative economy: an examination of Airbnb's early legal issues. In *Collaborative Economy and Tourism* (pp. 97-128). Springer, Cham.

Guttentag, D. (2019). Progress on Airbnb: a literature review. *Journal of Hospitality and Tourism Technology.*

Guttentag, D., & Smith, S. (2017). Assessing Airbnb as a disruptive innovation relative to hotels: Substitution and comparative performance expectations. *International Journal of Hospitality Management*, *64*, 1-10.

Guttentag, D., Smith, S., Potwarka, L., & Havitz, M. (2018). Why tourists choose Airbnb: A motivation-based segmentation study. *Journal of Travel Research*, 57(3), 342-359.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *Advances in neural information processing systems* (pp. 856-864).

Howell, D. C. (2007). The treatment of missing data. *The Sage handbook of social science methodology*, 208-224.

Johnston, R., Jones, K., & Manley, D. (2018). Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. *Quality & quantity*, *52*(4), 1957-1976.

Karlsson, L., & Dolnicar, S. (2016). Someone's been sleeping in my bed. *Annals of Tourism Research*, 58, 159-162.

Lutz, C., & Newlands, G. (2018). Consumer segmentation within the sharing economy: The case of Airbnb. *Journal of Business Research*, 88, 187-196.

Maslowska, E., Malthouse, E. C., & Bernritter, S. F. (2017). Too good to be true: the role of online reviews' features in probability to buy. *International Journal of Advertising*, *36*(1), 142-163.

Mauri, A. G., Minazzi, R., Nieto-García, M., & Viglia, G. (2018). Humanize your business. The role of personal reputation in the sharing economy. *International Journal of Hospitality Management*, 73, 36-43.

McGowan, R., & Mahon, J. (2018). David versus Goliath: Airbnb and the New York hotel industry. Archives of Business Research, *6*(4).

Paulauskaite, D., Powell, R., Coca-Stefaniak, J. A., & Morrison, A. M. (2017). Living like a local: Authentic tourism experiences and the sharing economy. *International Journal of Tourism Research*, *19*(6), 619-628.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the eighth ACM international conference on Web search and data mining* 399-408. ACM.

San Francisco Travel Association. (2015, July 22). *SAN FRANCISCO TRAVEL ASSOCIATION PARTNERS WITH AIRBNB TO MEET NEEDS OF VISITORS*. https://www.sftravel.com/article/san-francisco-travel-association-partners-airbnb-meet-needs-visitors

Spence, M. (2002). Signaling in retrospect and the informational structure of markets. *American Economic Review*, 92(3), 434-459.

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. In *International Conference on Machine Learning*, 190-198.

Varma, A., Jukic, N., Pestek, A., Shultz, C. J., & Nestorov, S. (2016). Airbnb: Exciting innovation or passing fad?. *Tourism Management Perspectives*, 20, 228-237.

Visser, G., Erasmus, I., & Miller, M. (2017). Airbnb: the emergence of a new accommodation type in Cape Town, South Africa. *Tourism Review International*, 21(2), 151-168.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), 673-686.

Xie, K., & Mao, Z. (2017). The impacts of quality and quantity attributes of Airbnb hosts on listing performance. *International Journal of Contemporary Hospitality Management*.

Zervas, G., Proserpio, D., & Byers, J. W. (2017). The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry. *Journal of marketing research*, 54(5), 687-705.

Appendix A1

Topic	Words	Topic Description
1	{san francisco, apartment, city, located, heart, restaurants, one, location, union square, building, home, downtown, stay, enjoy, great, neighbourhood, kitchen, north beach, place, bedroom}	Location focus, city centre
2	{rooms, hotel, san francisco, one, check, square feet, donatello, performance arts, king bed, foot ceiling, many, luxuries, largest, availability, sofa sleeper, queen sleeper, space, attractions, sink compact, amenities}	Hotel
3	{controls, bath products, individual climate, hairdryers makeup, designer desks, drawers inch, bathrooms feature, comforters pillows, refrigerators, chest, rooms feature, provide beddings, month deal, completing, ask questions, plush serta, mattresses thread, count linens, high definition, rooms}	Luxury
4	{rates based, october, confirm availability, additional fees, please send, fees see, offer inclusive, submitting reservation, submit special, rental terms, night minimum, april, sections, inquiry prior, rates may, inquiry, stay discounted, request, contact host, room}	Rules & pricing mechanism
5	{room, san francisco, private, hotel, guests, close, city, heart, access, free, union square, bathroom, lobby, space, cable car, bed, mini fridge, also, full, site}	City centre hotels
6	{check, front desk, room, minute drive, time, kitchen, shared, rooms, hotel, home, space, minute, walk, san francisco, pm, parking, downtown, guests, private, stay, designed}	Housing projects or hotels with amenities
7	{apartment, bedroom, one, fully equipped, san francisco, kitchen, home, speaker, blueground apartment, apartment include, smart tv, superior quality, premium wireless, start living, living room, unique, enjoy, building amenities, also offers, laundry amenities}	(Premium) amenities listing
8	{room, bedroom, kitchen, private, san francisco, home, bathroom, large, bed, living room, one, space, full, apartment, access, house, guests, two, unit, located}	Generic description
9	{san francisco, building dating, union square, cable car, hotel style, stops, chinatown nob, exciting urban, units, location worldmark, occupies historic, stockton streets, bush, manor, short walk, lower level, corner, reservations, hill, orange dog}	Chinatown location
10	{great, studio, tenderloin, required, apartment, site laundry, close, union square, san francisco, downtown, signed lease, neighborhood, wifi smart, nitty gritty, business, whether, fully equipped, building, accordance local, tv}	Focus on key elements
11	{san francisco, park, house, golden gate, neighborhood, restaurants, home, private, room, street, kitchen, downtown, city, great, located, bedroom, access, bathroom, quiet, one}	Homely feeling in quiet neighbourhood
12	{room, san francisco, shared, minutes, house, kitchen, home, living room, one, private, restaurants, neighborhood,	Enthusiastic home experience

	bedroom, rooms, location, living, available, bathroom, free, https://www.airbnb.com}	
13	{san francisco, short term, residential rental, ordinance, rental registration, registration certificate, agreed, comply, certificate certifies, terms, holder, registration number, inspection, potential building, housing fire, require, administrative code, code violations, city, section}	Using listing as residence + according legislation
14	{private mailbox, shared lobby, lively stretch, regular, regular business, secured entry, property management, team, farrell, maintenance, credit operated, shared debit, san francisco, laundry facilities, stay, community, hours, fidi etc, downtown, innovation, heart}	Business focus
15	{room, house, shared, outpost club, space, listing, san francisco, located, red Victorian, move, self check, nd floor, bunk, offers, bed, invite, music nights, attend events, access, art shows}	Focus on building structure & event offering
16	{apartment, close, san francisco, marina, golden gate, street, presidio, union, park, palace, union square, neighborhood, located, bridge, fine arts, financial district, place, city, restaurants, away}	Neighbourhood touring

Appendix A2

Topic 2: Hotel

SPECIAL RATE FOR 3 NIGHTS OR MORE

Private HOTEL room w/private bathroom, timeshare hotel property at THE DONATELLO in UNION SQUARE (sleeps 4) in the heart of San Francisco

You are not obligated to attend a timeshare presentation.

Requests for early check-in and/or late check-out are subject to availability; please feel free to contact the front desk directly and make arrangements.

The space
On-site valet garage parking available & paid by the guest. $38+TAX/day for cars
$48+TAX/day for larger vehicles (SUV's) no self parking

The Donatello is a hotel/timeshare and is one of the top rated in the city.

The space
The Donatello is a hotel/timeshare and is one of the top rated in the city.

Availability at this popular location changes quickly. I try to keep the calendar updated as best as I can. If you see availability, please submit a reservation request quickly.

Just show up and check in like any other hotel, no arranging to get keys necessary.

DESCRIPTION
Discover this modern hotel, The Donatello, inspired by the renowned Renaissance painter who shares its name. Located in the heart of San Francisco's fashionable shopping and theater district near numerous attractions, the Union Square SF hotel boasts a charming Italian atmosphere, superior amenities and attentive staff.

With 400 square feet of space and 10 foot ceilings, these rooms are some of the largest in the San Francisco area. Rooms are stunningly elegant and give you many of the luxuries of home including wet-bar and sink, compact refrigerator and microwave, sofa sleeper, cd player and in-room safe. All rooms have one KING BED and one QUEEN SLEEPER SOFA and sleep maximum of 4 people. No rollaways are available.

Topic 12: Enthusiastic home experience

hacknsleep,com

Location! Location! Location! It doesn't get much better than this. WalkScore gives our place top ratings: You are in the heart of where things are happening. No matter whether you work out of a co-working space like Runway, Galvanize or RocketSpace, intern at one of the tech companies attend a coding camp, or simply want to grab a bite, it's all easily accessible by foot and bike. WholeFoods, Safeway, Starbucks, Jamba Juice – also just around the corner.

The space
Guerrero and Market
15 bedrooms
6 bathrooms
2 Kitchens
2 Living rooms

When you enter your home for the first time, you'll notice the modern tones throughout, from the quartz countertops to the perfectly paired white cabinets. Our furniture and decor are expertly selected for this individual layout, and the mounted Samsung Smart TV's are ready for your enjoyment. Your home's filled with convenience amenities from the luxurious bedding to the fully stocked kitchen. Every element of your new home invites you to step right into a hassle-free living experience.

Topic 16: Neighbourhood touring

In prestigious Nob Hill and close to what you love about The City. Steps away from both the California Street cable car and the Powell Street line for quick hops to Market Street, the Embarcadero and Fisherman's Wharf. Just up the hill is the Top of the Mark cocktail lounge in the Mark Hopkins with 360 degree views and the world-famous Fairmont Hotel. This fully equipped apartment is conveniently close to Union Square, the Financial District, Chinatown and North Beach. And it's pet friendly!

The space

Reach out and touch the Transamerica building and SF skyline icons from this elegant snow white tower. The awe-inspiring views stretch from the Bay Bridge to the Golden Gate. Located directly across from the Ritz, this address is as swanky as it gets. Hardwood floors, chic kitchen design, combo washer/dryer and proximity to the best of the City make this a dream home.

The neighborhood is quiet, friendly, sunny and light, located one block from the Palace of Fine Arts, the St Francis Yacht Club, Crissy Field Beach (famous for windsurfing) and is also walking distance to the Golden Gate Bridge, Presidio Park, Walt Disney's museum, hiking trails, Fort Point and Fisherman's wharf. This is one of the nicest neighborhood in the city.
Few blocks away is a buzzing Chestnut Street fun friendly shopping area filled with small family operated cafes, restaurants, bars, delis, market, banks, pharmacies, boutiques, theaters, bus stops for transportation all over the city.

There are taxis, Uber and Lift all over the city for transportation. The bus system is great and affordable. It costs $2 for a 3 hour window which can take you all over the city. The cable cars are close by and are a good way to visit famous Ghirardelli and Union squares.

Appendix A3

Space Topic 2: Luxury ammenities

The space
Thoughtfully designed with bespoke finishes, modern furnishings, and a fully-equipped kitchen, you'll enjoy that "I'm home" feeling with this Blueground apartment. Whether you're lounging in your sophisticated living room streaming the latest and greatest entertainment on the smart TV or premium wireless speaker, or getting some well-earned rest on the superior quality mattress with luxury linens, you'll fall in love with everything this Pacific Heights apartment has to offer. This apartment also offers in-floor laundry.

Amenities

Building amenities unique to this studio apartment include an on-site:

- In-Floor Laundry
- Pet Friendly
- Swimming Pool
- Gym
- Indoor Parking
- Garden
- Courtyard

- Elevator

The space
Got pool and jacuzzi with gym

Space Topic 6: Legislation in order

The space
STR0000118 ---San Francisco Registration Number

The space
"SAN FRANCISCO SHORT-TERM RESIDENTIAL RENTAL REGISTRATION NUMBER: STR-0000674
Possession of a San Francisco Short-Term Rental Registration Certificate certifies that the registration certificate holder has agreed to comply with the terms of the San Francisco Short-Term Residential Rental Ordinance (San Francisco Administrative Code Section 41A). This ordinance does not require an inspection of the unity by the City for potential Building, Housing, Fire, or other Code Violations."

The preceding text is provided by the City and County of San Francisco and is required to be displayed here in the listing.

Appendix A4

Hi,

Thank you for participating to this survey. The purpose is to check how you respond to certain unexpected signals. Before we go ahead, we are curious if you have ever used Airbnb as holiday accommodation?

- Yes
- No

To quickly check whether you are not a bot: could you type 'A'?

How old are you?

What sex are you?

- Male
- Female
- Do not wish to answer

Assume that you and three friends were planning a trip to San Francisco for the summer. The plan is to stay from Wednesday to Sunday (four nights). Assume you found an apartment on Airbnb, a social marketplace that offers individuals (buyers) the opportunity to rent accommodations worldwide, that fits within all the parameters you selected (e.g., price range, capacity, etc.). Without reading the description, how likely will you be to book the listing?

1. Extremely likely
2. Moderately likely
3. Slightly likely
4. Neither likely nor unlikely
5. Slightly unlikely
6. Moderately unlikely
7. Extremely unlikely

You will now read a couple of descriptions. All accommodations are entire homes, meaning that all rented areas are private for you and your friends.

Now, let us assume you read following description: how likely are you to book this listing?

SPECIAL RATE FOR 3 NIGHTS OR MORE

Private HOTEL room w/private bathroom, timeshare hotel property at THE DONATELLO in UNION SQUARE (sleeps 4) in the heart of San Francisco

You are not obligated to attend a timeshare presentation.

Requests for early check-in and/or late check-out are subject to availability; please feel free to contact the front desk directly and make arrangements.

The space
On-site valet garage parking available & paid by the guest. $38+TAX/day for cars $48+TAX/day for larger vehicles (SUV's) no self parking

1. Extremely likely
2. Moderately likely
3. Slightly likely
4. Neither likely nor unlikely
5. Slightly unlikely
6. Moderately unlikely
7. Extremely unlikely

Now, let us assume you read following description: how likely are you to book this listing?

When you enter your home for the first time, you'll notice the modern tones throughout, from the quartz countertops to the perfectly paired white cabinets. Our furniture and decor are expertly selected for this individual layout, and the mounted Samsung Smart TV's are ready for your enjoyment. Your home's filled with convenience amenities from the luxurious bedding to the fully stocked kitchen. Every element of your new home invites you to step right into a hassle-free living experience.

1. Extremely likely
2. Moderately likely
3. Slightly likely
4. Neither likely nor unlikely
5. Slightly unlikely
6. Moderately unlikely
7. Extremely unlikely

Now, let us assume you read following description: how likely are you to book this listing?

The neighborhood is quiet, friendly, sunny and light, located one block from the Palace of Fine Arts, the St Francis Yacht Club, Crissy Field Beach (famous for windsurfing) and is also walking distance to the Golden Gate Bridge, Presidio Park, Walt Disney's museum, hiking trails, Fort Point and Fisherman's wharf. This is one of the nicest neighborhood in the city.
Few blocks away is a buzzing Chestnut Street fun friendly shopping area filled with small family operated cafes, restaurants, bars, delis, market, banks, pharmacies, boutiques, theaters, bus stops for transportation all over the city.

There are taxis, Uber and Lift all over the city for transportation. The bus system is great and affordable. It costs $2 for a 3 hour window which can take you all over the city. The cable cars are close by and are a good way to visit famous Ghirardelli and Union squares.

1. Extremely likely
2. Moderately likely
3. Slightly likely
4. Neither likely nor unlikely
5. Slightly unlikely
6. Moderately unlikely
7. Extremely unlikely