

MISTRuST: accommodation Short Term Rental Scanning Tool

Iván Ruiz-Rube¹, Inmaculada Arnedillo-Sánchez² and Antonio Balderas¹

¹*School of Engineering, University of Cádiz, Spain*

²*School of Computer Science and Statistics, Trinity College Dublin, Ireland*


Keywords: Accommodation Rental Platforms, Machine Learning, Web Scraping, Search Engines.


Abstract: The global irruption of ‘shared-accommodation platforms’ has ignited debate regarding the implications of the unprecedented growth of the short-term rental market. While some argue that they have generated new business, work and wealth others, highlight their wider societal, economic and legal effects. Short-term rental removes long-term housing from the market, forces rent prices up, saturates areas with tourism, generates safety and liability concerns and by and large, it is awash with likely illegal listings. This paper presents MISTRuST (accommodation Short Term Rental Scanning Tool) a computational intelligence based system aimed at uncovering whether a property is being listed in short-term rental (STR) platforms. It enables users to monitor property by scheduling automatic searches and checking listings returned by the system against the target property. The asynchronous pipeline architecture involves three stages: data ingestion, data enrichment and data matching. Preliminary test results with a set of listings are encouraging. However, further evaluation is needed to improve the accuracy of the system on a larger scale. It is hoped MISTRuST will help stakeholders such as property owners, state and agencies and others tackle the growing concern over unlawful STRs and contribute towards sustainable solutions.


1 INTRODUCTION

According to Schort (Schor, 2014) the launch of the sharing economy was marked by positive messages on how technological and economic innovation would provide financial benefits to ordinary people. In this vein, the shared-accommodation platform Airbnb, claims it helps individuals generate extra income from their underutilised accommodation to pay their rent or undertake a home renovation project. Founded in 2008, Airbnb was valued at 31 billion U.S. dollars in May 2017 and has over 3 million listings in 190 countries and 65,000 cities. These figures endorse the role of the company as an enabler of economic growth (Rae, 2018) through its direct operation and by indirectly generating tourism-related jobs (Nieuwland and van Melik, 2018). However, as demand for short-term rental grows, long-term rental is removed from market to meet that demand and rent prices of the diminishing stock are rising pushing people out of the rental market (Wachsmuth et al., 2018). Beyond the financial implications of this phe-

nomenon, the fabric of neighbourhoods is experiencing unprecedented changes. Long-term neighbours are being replaced by short-term visitors (Wachsmuth et al., 2018), bakeries, greengrocers and butchers by souvenir shops and residents are sieged by battalions of foreigners to the conquest of their neighbourhoods armed with their trolley cases in hand. While those engaging in short-term rental masquerade behind platforms and benefit from the overwhelming regulatory vacuum in the sector, local residents endure overcrowding, anti-social behaviour, safety concerns (Gurran and Phibbs, 2017) and injurious lack of enforcement of conveyance, planning permission, local and governmental regulation and common law. Most jurisdictions outlaw the use of residential accommodation for commercial short-term rental (Katz, 2015). To this end, the most comprehensive study to date on the issue, reports that 66% of revenue (\$435 million) and 45% of all New York Airbnb reservations last year were illegal because it breached New York State law (Wachsmuth et al., 2018). Nevertheless, the proliferation of STR platforms such as Airbnb.com, Booking.com, Homeaway.com, Home-stay.com, Couchsurfing.com, to mention just a few, makes enforcement an insurmountable task. For

^a  <https://orcid.org/0000-0002-9012-700X>

^b  <https://orcid.org/0000-0002-8749-9611>

^c  <https://orcid.org/0000-0003-0026-7410>

starters, unless privy of listing data, often confidential to the STR platforms and person offering the accommodation, it is difficult to locate specific listings among millions and assert they are legal or illegal. Being able to monitor listings on STR platforms should be of interest to authorities and local residents seeking enforcement of regulation but also, to landlords to ensure their property is not being used for STR without their knowledge. The STR subletting phenomenon is a real threat to landlords that can have serious consequences as, for instance, being fined by local authorities and been unable to evict tenants engaging in STR subletting. This paper presents MISTRuST, a computational intelligence based system, aimed at uncovering whether a property is being listed in short-term rental (STR) platforms. It enables users to monitor property by scheduling automatic searches and checking listings returned by the system against the target property. A methodology based on the design and creation strategy (Oates, 2005) and an agile life-cycle were adopted for its development. The rest of the paper is structured as follows. The background and related work are presented in Section 2. The requirements and features are described in Section 3. Section 4 elaborates on the asynchronous pipeline architecture involving three stages: data ingestion, data enrichment and data matching. Finally, section 5 provides discussion and future work.

2 ADVERTISEMENT LISTING DATA RESEARCH

As outlined previously, there are regulatory, financial, and societal implications of prevailing use of STR platforms. Exploring potential public policy responses may have in rental prices, Filippas (Filippas and Horton, 2017) entertains four scenarios: the decision to rent resides with the individual host, the building owner, the city or social planner. Although policy choices seem to have no detectable effect on rental prices (Filippas and Horton, 2017), ‘algorithm regulation’ would enable policymakers to draft regulations relying on data analysis to be responsive to real-time (Quattrone et al., 2016). Thus, gathering information from the online platforms and matching listing information with census and hotel data, it would be possible to determine the socioeconomic conditions of the areas that would benefit from the hospitality platforms. Personal characteristics may influence STR. Xiao (Ma et al., 2017) examine how hosts describe themselves on their Airbnb profile pages and examine their perceived trustworthiness. Along the line of trust and verification, Zhang (Zhang et al., 2016) uses

machine learning techniques to analyse the impact of having listings’ photos verified. Results illustrate that listings with verified photos are 9% more frequently booked.

In addition to the previous, attempts have been made to quantitatively characterize collaborative consumption behaviors in Airbnb (Lee et al., 2015) and analyze the economic value of trust artefacts (Teubner et al., 2016). Social features, such as responsiveness of host, number of reviews, membership seniority and ‘Super-host’ status, are significantly associated with room sales, providing economic value (Lee et al., 2015; Teubner et al., 2016).

There is a lack of peer-reviewed literature regarding property matching on STR platforms, except for a couple of commercial service providers, *subletalert.com* and *subletspy.com*, whose marketing summaries could be similar to our proposal. However, they do not disclose any technical details of their implementation and they do not seem to provide any inherent scientific contribution. However, there is work on the usage of recommender systems to help prospective tenants find a fitting property (Yuan et al., 2013).

3 MISTRuST (ACCOMMODATION SHORT TERM RENTAL SCANNING TOOL)

Since the publication of aggregated datasets of listings from Airbnb in the portal *insideairbnb.com* (Cox, 2017), the number of studies analysing the implications of STR platforms for hotels, neighbourhood cohesion and rental market have increased. The increased interest in analysing listing data and understanding the multidimensional implications of the STR phenomenon does not seem to be matched by an equal interest in the design and development of tools that help stakeholders automatically scan and monitor properties listed in such platforms. Thus, we envisaged the design and development of an accommodation STR scanning tool.

3.1 Requirements

MISTRuST functional requirements involve (i) registration of the data of the property to be monitored; (ii) issuing queries to check whether the property is listed on any STR platform; (iii) setting and receiving alerts once a potential match to the target property has been found listed in STR platforms.

With regards to the non-functional requirements, the following quality attributes were considered:

- **Security.** An authentication and authorization mechanism was included to protect the confidential data of the users.
- **Interoperability.** The system connects with online providers (alltherooms.com) of accommodation advertisements.
- **Performance.** A data cache module and a multi-thread streamed data processing scheme was included for proper performance.
- **Reliability.** An exception handling system and data cleansing and enriching tasks are applied to improve error tolerance.
- **Portability.** The system was developed with standards languages and formats, for compatibility with a regular web browser and even via an API provided for further processing.
- **Usability.** Common heuristics in user interface design have been applied to provide an adequate user experience.
- **Maintainability:** all the components of the system were developed by using well-known design patterns and principles in Software Engineering to ensure its sustainable evolution.

3.2 System Walk-through

Once the user logs on MISTRuST, the system allows access to three different modules, namely property management, search management and property finder.

From the Property Management view, the user can edit the list of properties. For each property, the user can register general information, upload pictures and annotate the property entry with the amenities or features it provides (see Figure 1).

Once the user has registered the properties, the next step is to issue searches on the Internet. There are two ways for doing that: (i) by selecting a specific time interval (check-in and check-out dates) or (ii) by scheduling the searches to be automatically launched on a regular basis. In Figure 2, a summary of each search issued for a given property is presented.

The results of the searches are incremental. Thus, regardless of the number of searches issued for each property, the listings collected are considered a single set. Therefore, in order to check whether a property is being listed, the user only has to select the property from the selector in the property finder window. Afterwards, the system will display the list of listings ranked by the computed score. The best-informed

guess of the system is ranked first so that the user will be able to check the details of the listing against the features, pictures and amenities of the target property (see Figure 3).

4 SYSTEM ARCHITECTURE

The architecture of the system which was carefully designed to fulfill its requirements follows a three-layered architecture along with an asynchronous pipeline one (see Figure 4). The whole system was developed with Java and Spring Framework.

The presentation layer consists of the different views of the system and was developed with Vaadin, a user interface framework for creating HTML5 single-page applications. The service layer includes several business components to manage the system entities and to implement the core algorithm of the system. The data layer was built on top of two data repositories, namely MySQL, for persisting the relational domain model, and Apache SOLR, for indexing advertisement data in form of flatten documents. The core of the system follows an asynchronous pipeline architecture. Three data processing stages are carried out before providing results for a given query: (a) data ingestion, (b) data enrichment and (c) data matching. However, not all three stages have to be executed at the same time.

4.1 Data Ingestion

The first data processing step involves obtaining data from Alltherooms.com. Since this web platform aggregates data from several popular STR platforms, such as Airbnb.com, Homestay.com, and Homeaway.com, the extraction process is limited to a single web platform. However currently, this web platform does not provide any structured API, so in order to find the listings the system has to directly scrape listing data from its user interface. Thus, the system issues an HTTP request to search for accommodation near the location of the user's property on the search page Alltherooms.com.

The exact location of the listing is unknown beforehand in order to guarantee the personal rights of the users. The webpage only provides us with the distance from the search point to the location of the listing for those listing further than 1 mile away (see Figure 5). Thus, to estimate the real distance (d) from a given listing (P) to the base position (O), two parallel searches are launched in equally distant locations towards West (A) and East (B). By applying the law of cosines to the formed triangles, a system of equations

General Data Pictures Amenities

Name
House1

Accommodation Type
General_Accommodation

Description
It is a magnificent area very close to the market square, the hospital, a bus stop and the beach. It is a very quite place with all the amenities.

Location
[Map showing location in Spain]

Number of bedrooms
[Slider]

Number of bathrooms
[Slider]

Save Cancel Delete

Figure 1: Editing form of properties.

Welcome to Subrenting Detector

Home Property Management Search Management Find subrenting Logout

Prepare a ad-hoc search Schedule searches

Select a property for checking their searches
House1

Check In	Check Out	Started	Finished	IsRunning	IsPropertyEnriched	Ads found	Ads enriched
2018-11-10	2018-11-15	2018-11-02T16:33:09	1970-01-01T16:38:09	false	true	40	34
2019-01-01	2019-01-05	2018-11-28T15:19:39	1970-01-01T15:22:49	false	true	0	0

Figure 2: List of searches launched for a given property.

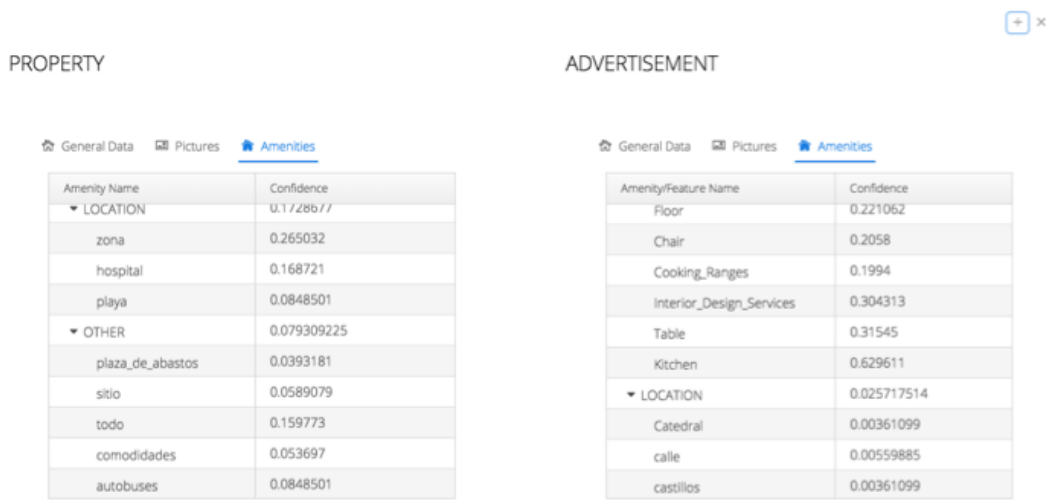


Figure 3: MISTRuST window showing the data of the user’s property (left) and a given ad (right).

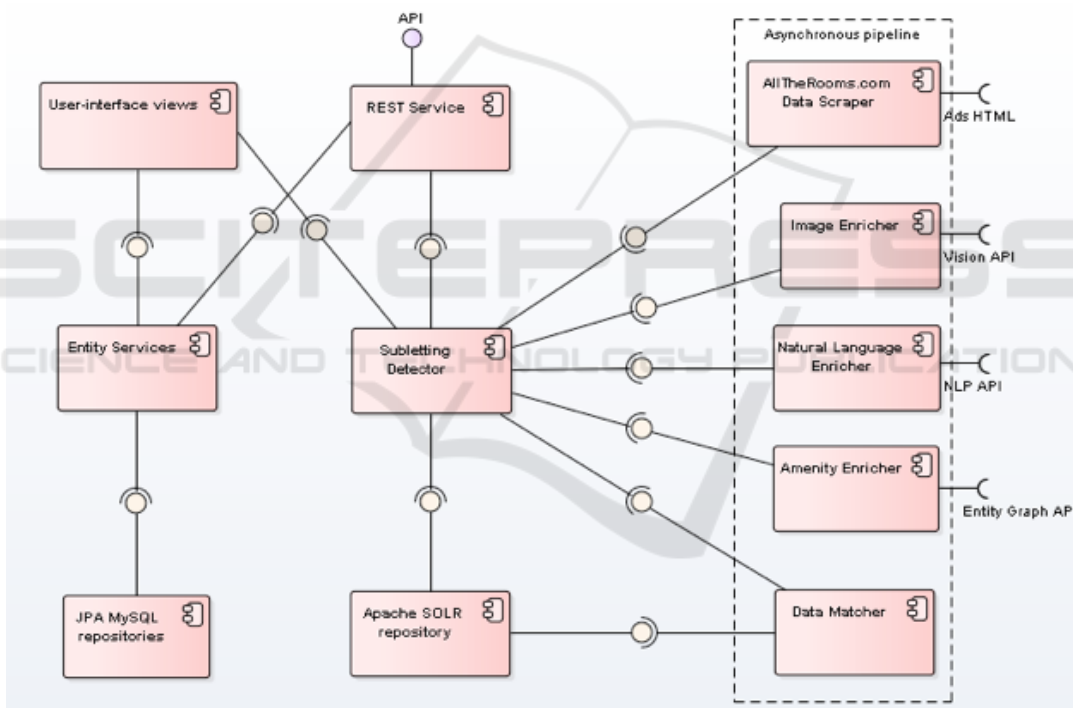


Figure 4: Three-layered partially-pipelined architecture.

(1) and (2) is generated, whose solution (3) represents the expected distance.

$$AOP : d^2 = d_A^2 + (2R)^2 - d_A(2R)\cos\alpha \quad (1)$$

$$APB : d_B^2 = d_A^2 + (4R)^2 - d_A(4R)\cos\alpha \quad (2)$$

$$d = \sqrt{\frac{d_A^2 + d_B^2}{2} - 4R^2} \quad (3)$$

In addition to the distance from the search point to the listing location, the scraper fetches (for every listing) the general description of the listing, its general features (number of rooms, bedrooms, etc.), its amenities and its pictures. Besides, due to the fact that the web provider may return a considerable number of listings in a single query, the scraper will iterate over each result page until a certain threshold distance. The listing might not be available the whole year but rather during different time intervals, for in-

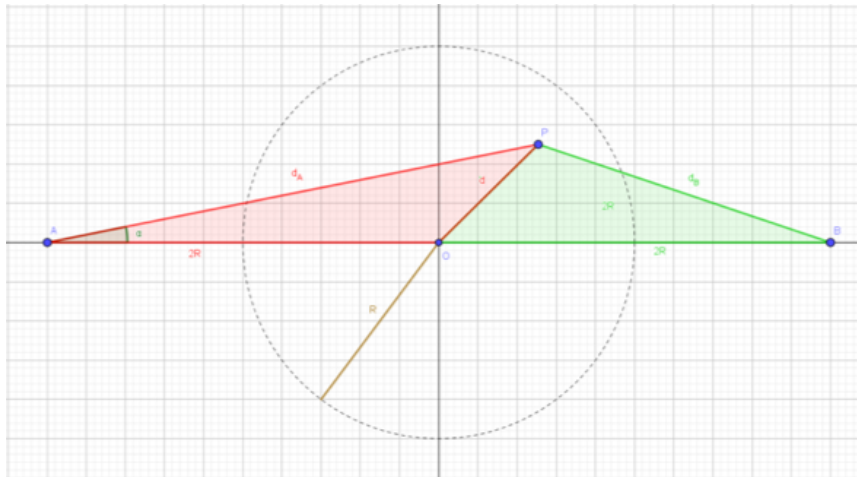


Figure 5: Estimation of the distance between the actual location of the property and that of a given advertisement.

stance, weekends or holidays. So, the scraping process can be both launched on user demand for specific dates, or scheduled in advance for specific time periods.

4.2 Data Enriching

Besides the approximated distance between the listing and the property locations, their set of features or amenities are essential to find the best listing match. Some amenities and features, such as elevators and air-conditioning systems, are explicitly stated by the user. However, there are other features which are not defined in a structured way but usually collected through textual descriptions. Moreover, the pictures uploaded by the users can also provide useful information about other features of the properties. In order to enrich the information collected from each listing, the system makes use of the cloud machine learning APIs provided by Google for processing natural language¹ and pictures². By analyzing the textual descriptions and the images of the advertisements, the system can obtain a considerable number of tags describing them. However, there is a huge variability on the kind of entities which are automatically discovered from the listings, such as house facilities, events, street names, user's rules, and so forth. So, a further classification process is required. The use of semantic web technologies (Vandenbussche et al., 2017), such as linked open vocabularies, can alleviate this task. This system uses the Google Knowledge Graph³ to convert the tags into features and amenities according

to the Schema.org⁴ vocabulary.

4.3 Data Matching

Once the listing data is downloaded and enriched by means of the machine learning services, the final stage is the data matching. The open-source search platform Apache SOLR is used for this. Unlike standard databases, this software enables us to send and index documents and to query them by using complex criteria.

Firstly, initial filter criteria, such as the number of bedrooms and bathrooms and the accommodation type (flat, house, cottage, etc.), are applied to restrict the superset of advertisements. Secondly, the results are ranked by a computed score combining (i) the computed approximated distance from the position of the listing to the location of the real property and (ii) the matching between the amenities/features of the listing and those of the property. In order to find matches between the features and amenities of listings with those of the properties, some aspects have to be considered. On the one hand, not all the entities have the same relevance. For example, a user's rule, such as 'dogs allowed' or 'towels provided', does not provide any useful insight, but amenities such as elevator or heating do. For that reason, all the amenities can include a significance factor between 0 and 1, indicating its usefulness for the matching process. On the other hand, because of the own nature of the machine learning models, it is necessary to consider a certain confidence value to their results. For example, in the case of natural language recognition, the Machine Learning (ML) model provides us with information about

¹<https://cloud.google.com/natural-language/>

²<https://cloud.google.com/vision/>

³<https://developers.google.com/knowledge-graph/>

⁴<https://schema.org/>

the relevance of each recognized entity to the full-text description of the listing. Furthermore, the ML model for image recognition returns a number representing the accuracy of the entity detection. In the case of the features and amenities explicitly defined by the users without the support of any ML model, it is considered a confidence factor of 1.0.

Both the relevance factor and the confidence value are hence taken into account to calculate the final score. However, before computing the score and ranking the results, a query expansion process is applied to deal with synonymous tags, parent/child terms and translations in different languages. This process is carried out by means of tables with recursive relationships and the WordNet lexical database⁵ (this step is still under development). Finally, the results are ranked by the score.

5 DISCUSSION AND CONCLUSION

Short-term rental removes long-term housing from the market, forces rent prices up, saturates areas with tourism, generates safety and liability concerns and by and large, it is awash with likely illegal listings. In this paper, we have presented MISTRuST (accommodation Short Term Rental Scanning Tool) a machine learning based system aimed at uncovering whether a given property is being listed in short-term rental (STR) platforms. It enables users to monitor property by scheduling automatic searches and checking listings returned by the system against the target property. During the development of MISTRuST validity considerations have been contemplated. Firstly, the manual allocation of the relevant values of the amenities and features is a time-consuming task. The required time can be reduced by assigning default values to the categories they belong. Secondly, we are subjected to the terms and conditions of the STR platforms. For example, Airbnb bans the use of automated data scrapping tools but Alltherooms.com doesn't. Nonetheless, this may change in the future. In addition, Alltheroms.com doesn't currently provide an API so we are using web scraping techniques. Any future changes in the user interface design of Alltheroms.com may impact our tool. Lastly, a large-scale evaluation, with many properties and in different locations would be necessary in order to obtain accurate results. However, no software will be able to ensure a 100% the accuracy of its guesses due to the dependence on completeness of the property

⁵<https://wordnet.princeton.edu/>

data provided by the user and the uncertainty of the ML algorithms themselves. Our next step is to apply a machine learning model to improve the ranking algorithm for the top N retrieved listings, to avoid manually adjust the relevance's value of each feature or amenity. To accomplish that, a substantial amount of training data is required. So we will introduce a user feedback system to register the success or the failure of the algorithm executions and so to train the ML model. Finally, we plan to release the software as open-source.

ACKNOWLEDGEMENTS

The work has been supported by the University of Cadiz with grant ref. EST2018-167 and funds of its Department of Computer Engineering.

REFERENCES

- Cox, M. (2017). Inside Airbnb: adding data to the debate. Accessed: 2019-09-11.
- Filippas, A. and Horton, J. J. (2017). The tragedy of your upstairs neighbors: When is the home-sharing externality internalized? *Available at SSRN 2443343*.
- Gurran, N. and Phibbs, P. (2017). When tourists move in: How should urban planners respond to Airbnb? *Journal of the American Planning Association*, 83(1):80–92.
- Katz, V. (2015). Regulating the sharing economy. *Berkeley Technology Law Journal*, 30(4):1067–1126.
- Lee, D., Hyun, W., Ryu, J., Lee, W. J., Rhee, W., and Suh, B. (2015). An analysis of social features associated with room sales of Airbnb. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, pages 219–222. ACM.
- Ma, X., Hancock, J. T., Lim Mingjie, K., and Naaman, M. (2017). Self-disclosure and perceived trustworthiness of Airbnb host profiles. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 2397–2409. ACM.
- Nieuwland, S. and van Melik, R. (2018). Regulating Airbnb: how cities deal with perceived negative externalities of short-term rentals. *Current Issues in Tourism*, pages 1–15.
- Oates, B. J. (2005). *Researching information systems and computing*. Sage.
- Quattrone, G., Proserpio, D., Quercia, D., Capra, L., and Musolesi, M. (2016). Who benefits from the "sharing" economy of Airbnb? In *Proceedings of the 25th International Conference on World Wide Web, WWW '16*, pages 1385–1394, Republic and Canton of Geneva,

- Switzerland. International World Wide Web Conferences Steering Committee.
- Rae, A. (2018). From neighbourhood to a globalhood? three propositions on the rapid rise of short-term rentals. *Area*, 0(0).
- Schor, J. (2014). Debating the sharing economy. a great transition initiative essay. *Online: Great Transition Initiative*.
- Teubner, T., Saade, N., Kawlitschek, F., and Weinhardt, C. (2016). It's only pixels, badges, and stars: On the economic value of reputation on Airbnb. In *Australian Conference on Information Systems*.
- Vandenbussche, P.-Y., Ateazing, G. A., Poveda-Villalón, M., and Vatan, B. (2017). Linked open vocabularies (LOV): a gateway to reusable semantic vocabularies on the web. *Semantic Web*, 8(3):437–452.
- Wachsmuth, D., Chaney, D., Kerrigan, D., Shillolo, A., and Basalaev-Binder, R. (2018). The high cost of short-term rentals in New York City. *School of Urban planning, McGill University: Montreal, QC, Canada*.
- Yuan, X., Lee, J.-H., Kim, S.-J., and Kim, Y.-H. (2013). Toward a user-oriented recommendation system for real estate websites. *Information Systems*, 38(2):231 – 243.
- Zhang, S., Lee, D., Singh, P. V., and Srinivasan, K. (2016). How much is an image worth? an empirical analysis of property's image aesthetic quality on demand at Airbnb.

