NYCVA
A visual analysis of New York City AirBnBs

Leonardo Balzoni - 1870364
Alberto Contorno - 1873252

February 21, 2020

1 Introduction to the analysis

Housing websites like *AirBnB* [1] are amongst the sites that globally generate more traffic [2]. Even with the impressive unseen work that ensures the navigating the site is as pleasurable as possible, with the huge amount of data it can be quite hard to get a grasp on the actual informations that we, as users, are looking for. Our work will try to give some, otherwise hard to see, insights about the announcements with the help of visual analytics. Obviously doing it on a global scale would be too computationally heavy for our resources so the analysis will be limited to one of the cities that has the most announcements on the site; *New York City*.

2 Description of the dataset

The dataset that is going to be used was provided from *Kaggle* [3], and it is made of around 50000 tuples with 16 attributes each [4]. We are perfectly aware of the fact that this means that our *"AS index"* is quite higher than the suggested one, which means that we are faced with 2 options: reduce the amount of attributes and/or tuples, or build a small backend server (probably in *NodeJS*) on which to decouple the computational load to make the web browser more responsive.
Each tuple of the set represents an announcement on the AirBnB website for a house in New York City, some of the most useful attributes are: the location of the house (both its neighbourhood and its coordinates), its cost, the average availability, and the amount of reviews. These information will be used to identify patterns and gather insights about the distribution of rental housing in NYC in order to have a visual representation of the best options available.

3 Goals of the Analysis

An analyisis of this kind can be used to fulfill multiple objectives; firstly as alreay stated the multiple visualizations will allow to see otherwise hidden informations, such as for example the average price in each neighbourhood, or the estimated income of the various hosts. This means that both someone who is willing to put a announcement on the website and someone who is looking for a accomodation, will be able to find useful informations.
Moreover we could apply some dimensionality reduction algorithms such as *PCA* or *t-SNE* to gather informations about similar announcements, which

could for example be used to find an announcement similar to one that we like, but that might be unavailable.

4 Dataset preprocessing

The given dataset was already quite ready to be used, as there were no duplicates and the attribute's types were coherenent and almost always not null.
Even with this said some preprocessing was necessary, what we did was dropping some useless columns such as the name of the host or the date of the last review.
Moreover there were some announcements with missing data for important fields, so we dropped those (they were only 20 out of 50000).
Lastly there was the "Reviews per month" attribute that was null in the case of 0 reviews on the announcement. Instead of dropping these, what we did was to fill the null values with 0 in the case o no reviews for the house.

4.1 PCA preprocessing

While what we said is enough for all the analyses that we ended up doing, evaluating PCA requires for more attention. This is due to the fact that the algorithm only accepts numerical values for the various fields, and our datasets instead has some categorical ones (i.e. "neighbourhood" or "housing_type"). What we did was create some sort of one-hot encoding for those fields; for example "housing_type" could be: an entire apartment, a private room or a shared room, so we replaced the "housing_type" field with 3 new ones: "housing_type_apt", "housing_type_pvtroom" and "housing_type_shroom", and the values are all 0 except for one which is 1. Doing the same for the rest of the categorical fields we finished the preprocessing for the PCA dataset.

5 Visualizations and Interaction design

qui tutte subsections, una per ciascun graph

6 Future work

Robina su smart pricing

7 TODO

- Rimuovi TODO section lol
- Finalizza se hai usato tutto il dataset con node o se lo hai tagliato in sezione description of dataset
- prezzo per disponibiluita = income mensile

References

[1] “AirBnB’s website”
https://www.airbnb.it/.

[2] “Web traffic on AirBnB”
https://www.similarweb.com/website/airbnb.com.

[3] “Kaggle website’
https://www.kaggle.com/.

[4] “New York City AirBnB’s dataset’
https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/.