trivago Case Study: Which hotel is better for nightlife?

Jac Davis

9 July 2018

This report was written in rmarkdown. The full report can be reproduced using the rmarkdown file. Code chunks can be run separately within the rmarkdown file if needed. If you want to view the report with the code included, set echo=TRUE in the above code chunk.

Which hotel is better for nightlife?

Problem. Make it easier for travellers interested in "nightlife" to find a suitable hotel.
Solution. Create a score to rank hotels according to their suitability for travellers interested in "nightlife".

Intended audience. The intended audience for this analysis and presentation is trivago product managers, with a strong consideration of how useful the score will be for travellers using trivago.

Considerations. The problem requires estimation of a "nightlife" score for hotels. This "nightlife" score is not directly observable, and so it must be estimated from the data that we do have. These data may be skewed, missing, or unreliable. To discover some of the potential issues with the data, an exploratory analysis is presented first.

Summary of the approach. The following sections are presented in order. An **exploratory analysis** presents a description of the dataset, summary statistics,

Part 1. Exploratory analysis of the input data

This section presents an exploration and visualisation of the data.

Two .csv files contribute data to this analysis. The **hotels** csv lists hotels in four cities, their geo-coordinates, and information about the hotel. The **pois** csv lists points of interest (POIs) in the same four cities, their geo-coordinates, and information about the POIs.

The data exploration is presented in three parts: 1. Summary statistics and distributions of the hotel data 2. Missing data analysis 3. Geographic exploration

1.1. Summary Statistics

1.1. Summary statistics. Summary statistics are useful for understanding the data, and distributions are important because some statistical techniques require data to be distributed in a particular way (e.g., normal or Gaussian distribution is required for OLS regression). We need to know what techniques might be suitable for these data.

Hotels. The **hotels** dataset has 32 variables and 400 rows. The variables are: hotel_id, city_id, hotel_type, basename, distance_to_center, longitude, latitude, overall_rating, impression_level, interaction_level, car_park, club_club_hotel, designer_hotel, attraction_hotel, luxury_hotel, beach_front_hotel, convention_hotel, spa_hotel, country_hotel, airport_hotel, senior_hotel, eco_friendly_hotel, party_people, business_people, honeymooners, singles, large_groups, family_hotel, gay_friendly, wifi_lobby, wifi_room, city_name.

There are four cities included in the dataset: Amsterdam, Los Angeles, Hong Kong, and Thessaloniki. These cities may show different patterns of travel, so we should look at them separately.

Let's take a first look at the hotel data.

Here is the information we know about hotels

Types of hotels in different cities:

	Amsterdam	Hong Kong	Los Angeles	Thessaloniki
Bed & Breakfast	4	0	2	1
Home / Apartment	5	2	8	17
Hostel	1	7	2	3
Hotel	86	88	78	74
Inn	2	1	2	0
Motel	0	0	6	0
Resort	0	2	0	0
Serviced Apartment	2	0	2	5

Hotels with different characteristics, such as “luxury hotels”:

	No	Yes	Missing
club_club_hotel	9	0	391
designer_hotel	10	44	346
attraction_hotel	10	10	380
luxury_hotel	9	44	347
beach_front_hotel	21	6	373
convention_hotel	11	84	305
spa_hotel	14	23	363
country_hotel	9	1	390
airport_hotel	10	35	355
senior_hotel	1	2	397
eco_friendly_hotel	9	30	361

Hotels targeting specific people, such as “party people”:

	No	Yes	Missing
party_people	14	59	327
business_people	1	227	172
honeymooners	1	85	314
singles	1	104	295
large_groups	3	136	261
family_hotel	15	142	243
gay_friendly	5	77	318

Hotel wifi:

	No	Yes	Missing
wifi_lobby	4	338	58
wifi_room	4	338	58

We also know how some information about the hotels' performance:

	Mean	Standard Deviation	Min	Max	Missing
overall_rating	3.20612813370474	1.30827460909766	1	5	41
impression_level	1.93	1.76433619401569	1	10	0
interaction_level	1.6925	1.36109337959228	1	10	0

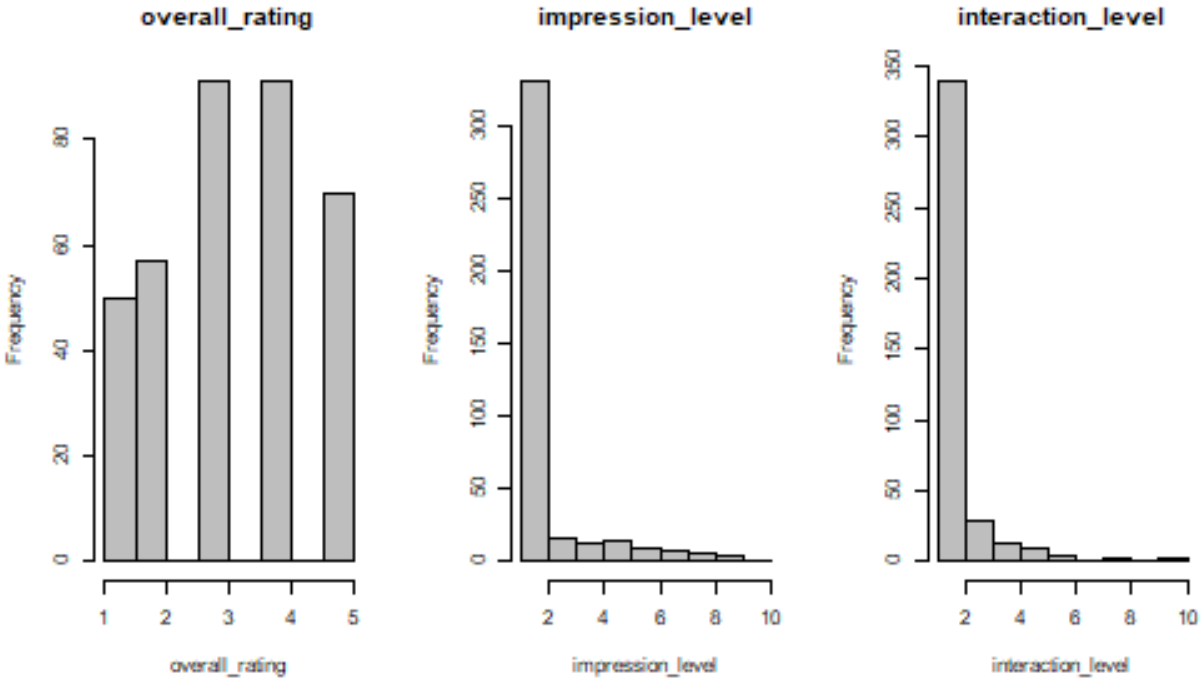

Figure 1: Hotel rating, impressions, and interactions

Finally, we have some spatial information: latitude, longitude, and distance to center for each hotel. It will be best to explore these in a map (see section 1.3).

1.2. Missing Data

1.2. Missing data analysis. It is important to diagnose and correct for missing data. Missing data is common in real-world analytic scenarios, but the amount and type of missing data can affect the analytic results (in this case, the nightlife score) in different ways. Very small amounts of missing data, or data that are missing at random, can be easily assumed to not affect the result, and can be dealt with by deletion or imputation. However, large amounts of missing data, or data missing "not at random"" - where the missing values might be systematically different to the values that we observe in the dataset - cannot be dealt with through simple deletion or imputation. The best way to deal with data that are missing not at random is to collect more data.

Are the data in our dataset missing at random?

party_people	Percent missing	81.75
business_people	Percent missing	43
honeymooners	Percent missing	78.5
singles	Percent missing	73.75
large_groups	Percent missing	65.25
family_hotel	Percent missing	60.75
gay_friendly	Percent missing	79.5

As seen in the table, there is a very high percentage of missing data in these hotel variables. For example, "party_people" has 82% missing data. This high proportion of missing data is a problem - even if we impute the missing values, we are likely to impute low-quality estimates, because the hotels that have this information may be systematically different to hotels that do not have it. For example, hotels that provide information about "party_people" may also be systematically better for nightlife than hotels that do not provide this information. If we train our dataset only on this subset, we will get biased results. Similarly, if we assume that any missing value is actually zero, we will get biased results, because it could actually be 1.

But these variables, such as "party people" and "business people", seem intuitively very useful for a nightlife score. We could combine several of them into a single score, to give us more data. We could assume that hotels with a 1 for "party people" or "singles" might be good for nightlife, but hotels with a 1 for "business people" or "family hotel" might be bad for nightlife.

Let's take a look at our combined variable.

Business or Family	Party or Single
258	20

Not missing	Missing
278	122

Now we have many more observations, but the total amount of missing data is still 43.88% missing data.

Additionally, the combined score is strongly skewed to "business or family" hotels over "party or single" hotels. There just aren't many hotels in the dataset that are marked as "party people" or for "singles" - only 20 hotels in total. When we split this number into different cities, we will have even fewer observations.

So the hotel's description as a "party people" or "singles" hotel isn't a good choice to base a nightlife score on. For example, if we used a machine learning algorithm to predict hotels' scores on this variable, it would most likely just predict that every hotel was a "business or family" hotel (n = 258) and no hotels were "party or singles" hotels (n = 20). Guessing that every hotel was not a party or singles hotel would give the model

very high predictive power, but would be close to useless for a score that we want to use to distinguish hotels.

As a final check, let's see whether hotels with missing values on this combined variable, are actually different to hotels that don't have missing values. We can do this by comparing the interaction level and impression level between hotels that have data for the combined score, and hotels that have missing data. Statistically speaking: Is there a significant relationship between interaction or impression level, and the probability that the hotel data are missing, for the combined variable?

Table 9: Table continues below

	Coefficient	SE
Impression level	-0.0631702682683326	0.0127081898666855
Interaction level	-0.068298886262095	0.0166280504248665

p value
9.92540468159338e-07
4.85812346217986e-05

The missing data analysis shows that *there is a significant statistical relationship between impression and interaction levels, and the probability that hotels have missing data* on the combined variable. Specifically, *hotels with lower impression levels, or lower interaction levels, are more likely to have missing data.* One possible explanation for this relationship would be that hotels that have a higher Internet presence in general, have more impressions and interactions, and are also likely to have given themselves a rating for "party people", "business people", etc. So if we based our analysis on these scores, we might actually just be analysing a hotel's online presence, and not whether the hotel is actually better for nightlife.

Therefore, we need to look at other data. Fortunately, we also have information about the hotels' locations, and points of interest in these cities.

1.3. Geographic Information

1.3. Geographic exploration. Since we have the geographic coordinates of the hotels and the POIs, we can visualise this information geographically. A geographic visualisation may give clues to the best approach for analysis. (a) Hotels on 4 maps: Amsterdam, Thessaloniki, Hong Kong, Los Angeles (b) POIs overlaid on hotel maps (c) Colour hotels with "nightlife" identifiers and colour POIs with "nightlife" characteristics

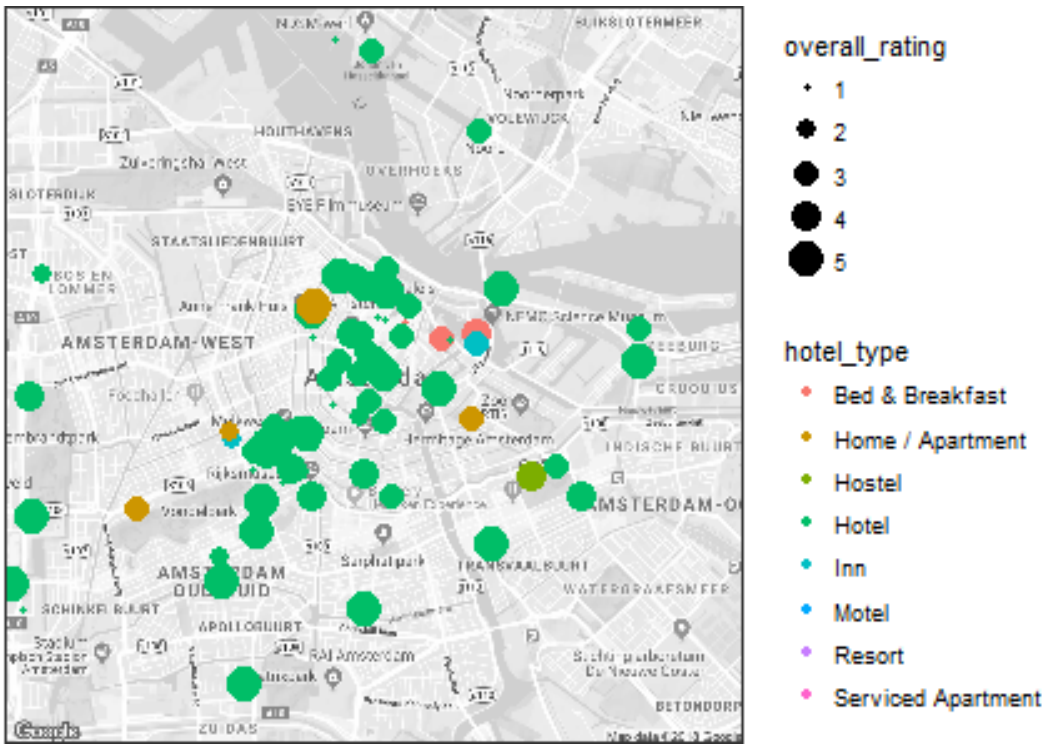

Figure 2: Hotel locations in Amsterdam

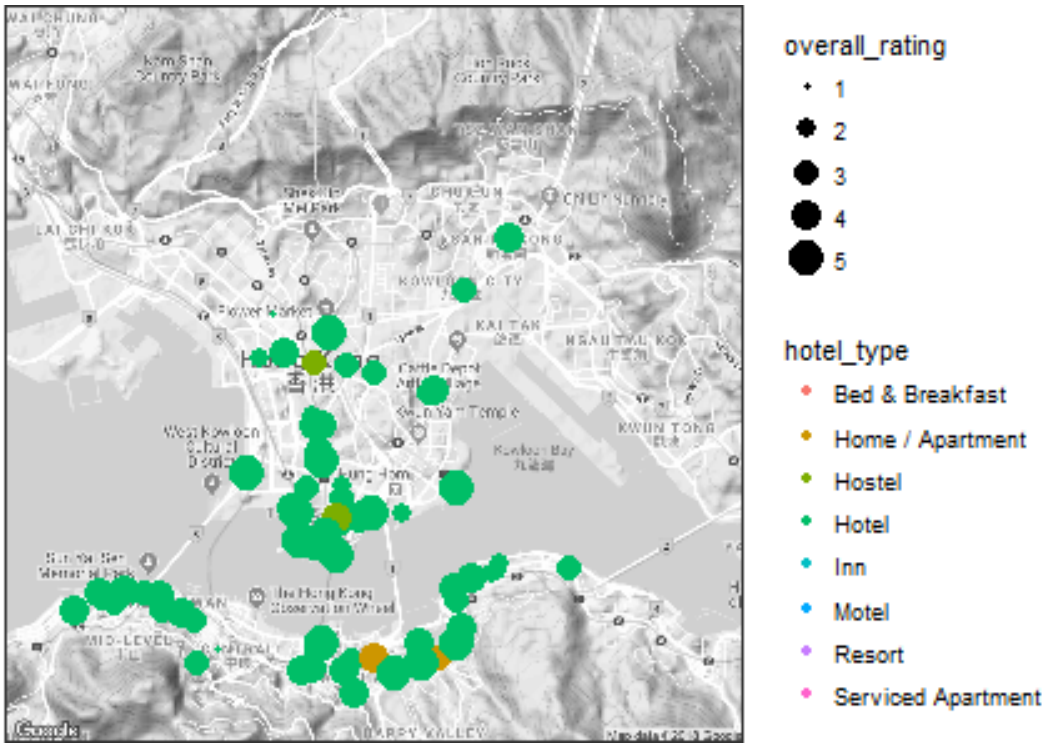

Figure 3: Hotel locations in Hong Kong

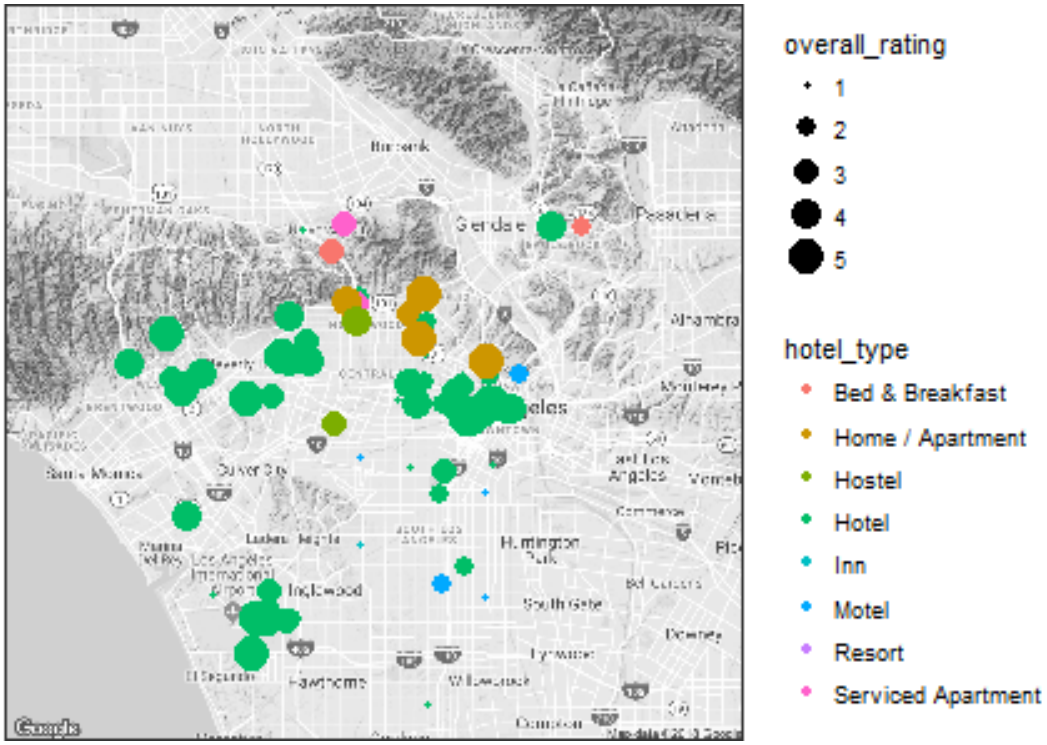

Figure 4: Hotel locations in Los Angeles

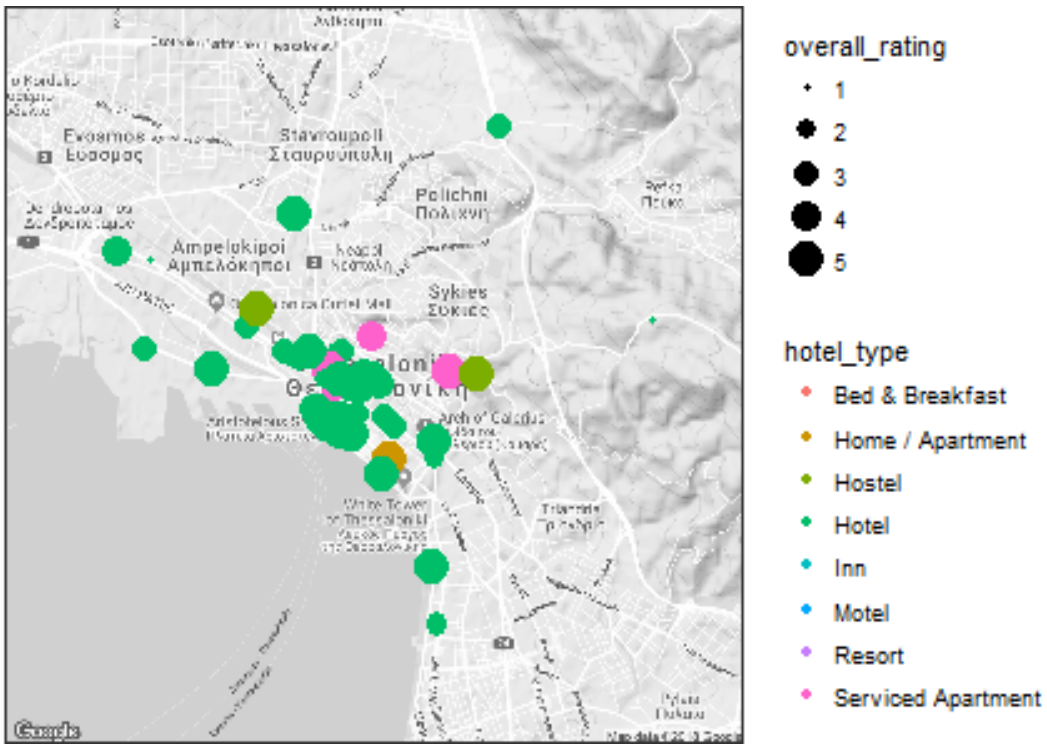

Figure 5: Hotel locations in Thessaloniki

1.4. POI data

1.4. POI data. We also have data on points of interest (POIs) in each city. These are geographically specified as well, so we can map them with the hotel locations. However, mapping all of the POIs with the hotel locations creates quite dense maps, because there are many more POIs than hotels.

Initial explorations of the POI data show that multiple types of POI are contained in the *poi_types* variable, so we should split these to get a clearer look at the data.

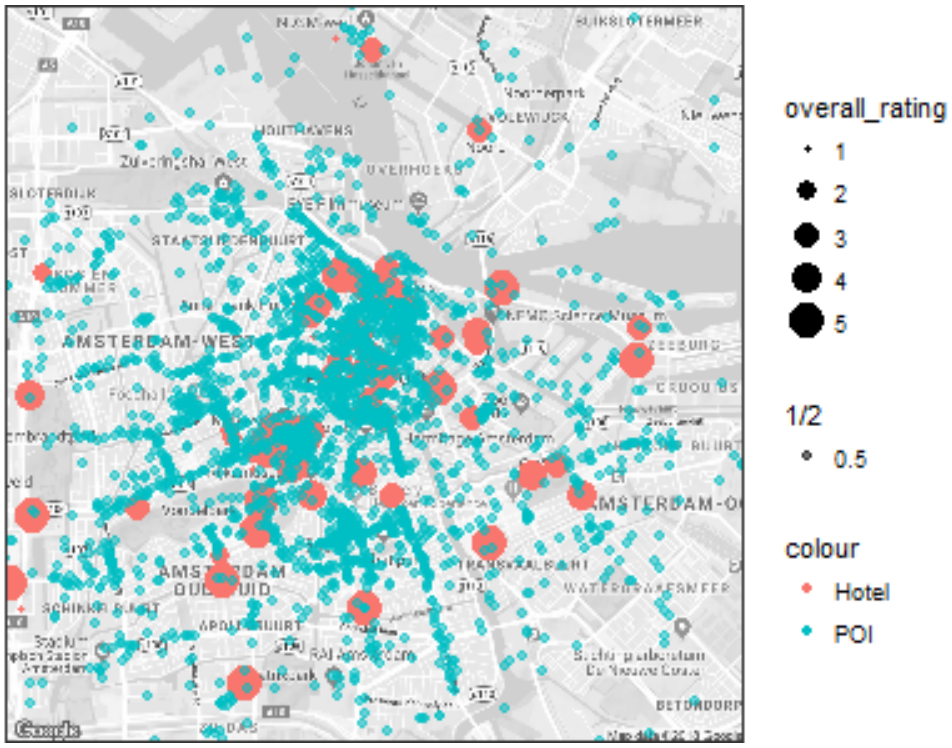

Figure 6: Hotels and POIs in Amsterdam

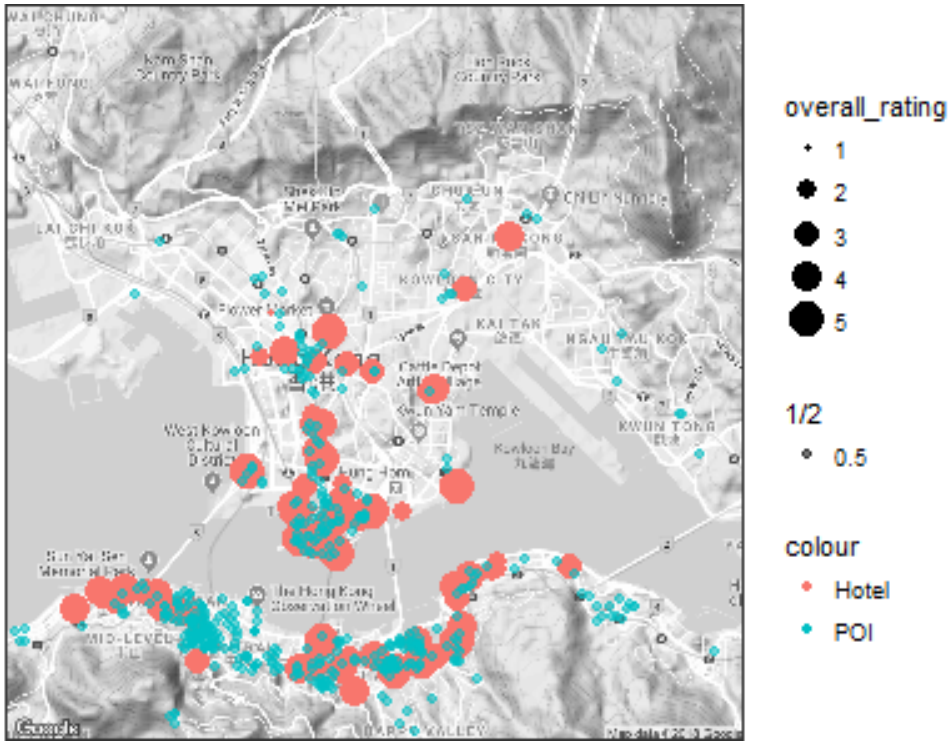

Figure 7: Hotels and POIs in Hong Kong

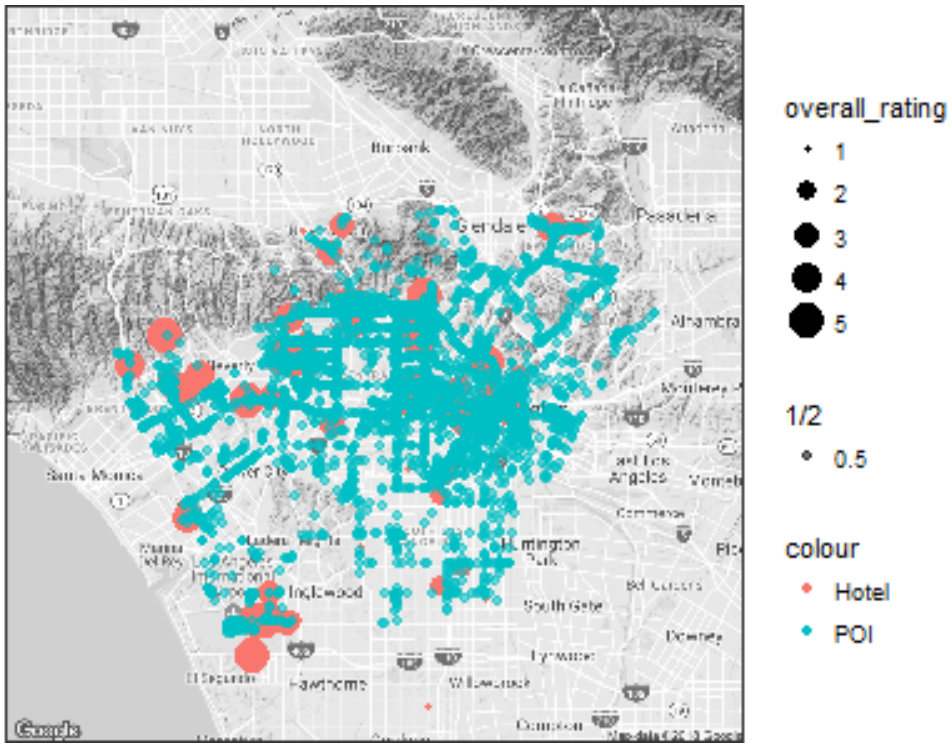

Figure 8: Hotels and POIs in Los Angeles

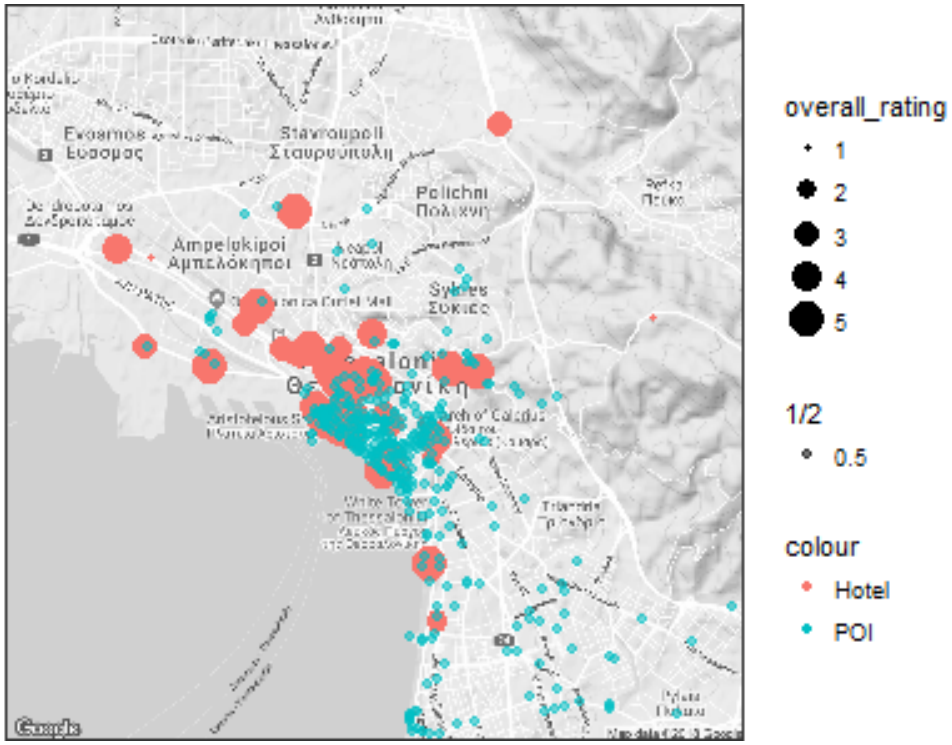

Figure 9: Hotels and POIs in Thessaloniki

We might want to know what types of POIs are available.

Table 11: Table continues below

Airports	Amusement Park	Architectural Buildings	Art Galleries
37	13	198	281

Table 12: Table continues below

Bar / Pub	Beach	Boardwalk / Promenade / Street	Botanical Gardens
1749	10	176	6

Table 13: Table continues below

Bowling	Bus Stations	CafÃ©	Car Rental	Casino	Cinema
10	20	2802	9	3	53

Table 14: Table continues below

Classes / Workshops	Disco / Nightclub	Event / Entertainment
78	834	1405

Table 15: Table continues below

Festival Area	Flea / Street Markets	Food & Drink	Game Centers
108	48	7960	53

Table 16: Table continues below

Gift Shop	Golf Area	Gym / Fitness Center	Harbors	Historic Sites
41	10	88	5	119

Table 17: Table continues below

Islands	Lake	Lookout	Malls	Metro Stations	Mountain	Museums
3	10	24	73	128	5	177

Table 18: Table continues below

National Parks	Nature	Outfit Shops	Palaces / Castles	Parks
6	151	440	3	126

Table 19: Table continues below

Performing Arts Venue	Religious Sites	Restaurants	River	Services
331	116	7242	3	1492

Table 20: Table continues below

Shopping	Spas	Sports	Stadium	Taxi Stand	Tour Provider
1965	49	192	30	1	14

Table 21: Table continues below

Tourist / Visitor Centers	Tourist Attractions	Trade Fair	Trails / Tracks
18	669	19	24

Table 22: Table continues below

Train Stations	Tram Stations	Transportation	Water Parks
18	15	197	1

Windsurfing Spot	Zoos / Aquariums
1	7

From this list, we can **heuristically** select the types of POIs that might be relevant to "nightlife": **"Bar / Pub"**, **"Festival Area"**, **"Disco / Nightclub"**, **"Casino"** .

note: As a later step, we could check sensitivity to including certain types of POIs and not others, e.g., "Bar/Pub" but not "Restaurant".

Then we can map nightlife “hotspots” in each city. These are geographic regions where travellers might easily find “nightlife” POIs.

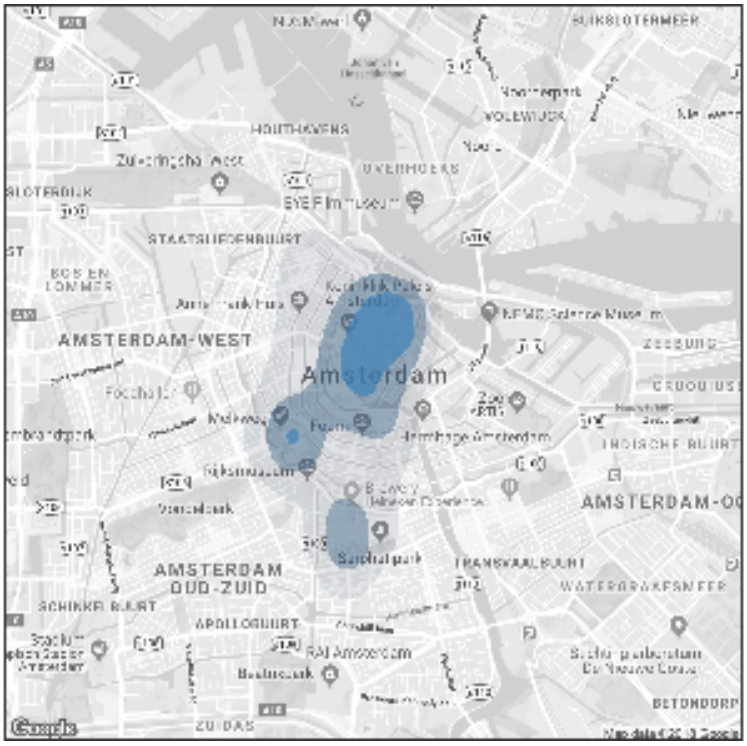

Figure 10: Nightlife density in Amsterdam

Figure 11: Nightlife density in Hong Kong

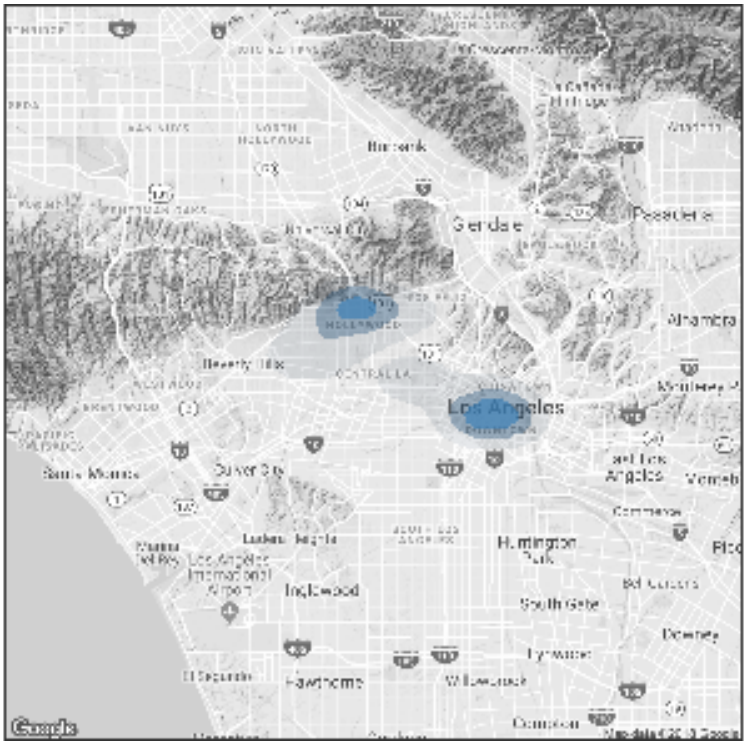

Figure 12: Nightlife density in Los Angeles

Figure 13: Nightlife density in Thessaloniki

2. Creating a nightlife score

We could use a machine learning approach to rank the hotels, but ML would require a **ground truth** of a hotel's *"nightlife"* score, with which to train and test the model. Here, we do not have a ground truth. What we do have are several variables that might predict a hotel's *nightlife* score. We can use these variables to create a score.

Based on the previous data exploration, we know that hotel features, such as rating of "party people" hotels, have a lot of missing data, and so may not be the best choice to base an entire score on. However, we may use these features later, to check whether our score is good.

Solution: Nightlife Hot-Spot Score

Based on the previous geographic analysis, we can see several nightlife "hot-spots", identified by a high density of POIs related to nightlife: bars/pubs, discos, casinos, etc. We can use this geographic density to create a score for each hotel's geographic nightlife density, from its geographic location.

Step 1. Estimate nightlife density

Nightlife density can be calculated from the POI data. Using two-dimensional kernel density estimation, we can estimate the density of nightlife hotspots across latitude and longitude. We use non-parametric local regression to estimate the relationship between latitude+longitude, and nightlife density.

We want to know *relative* nightlife density for each city: travellers looking for places to stay in Thessaloniki are unlikely to change their plans just because there are more bars on average in Amsterdam.

We will use a statistical method that accounts for differences in absolute nightlife density between cities, but that can be generalised to more cities than appear in the current dataset.

Step 2. Calculate nightlife density for each hotel

We use the predictive function created in step 1 to calculate a nightlife density score for each hotel, based on its latitude and longitude.

Step 3. Rescale nightlife density score to min = 0 and max = 10

The final nightlife score needs to range from 0 to 10, with 10 representing the highest nightlife density in the dataset.

Results

Distribution of nightlife score

The final nightlife score has a minimum of 0, a maximum of 10, a mean of 7.25, and a standard deviation of 2.9.

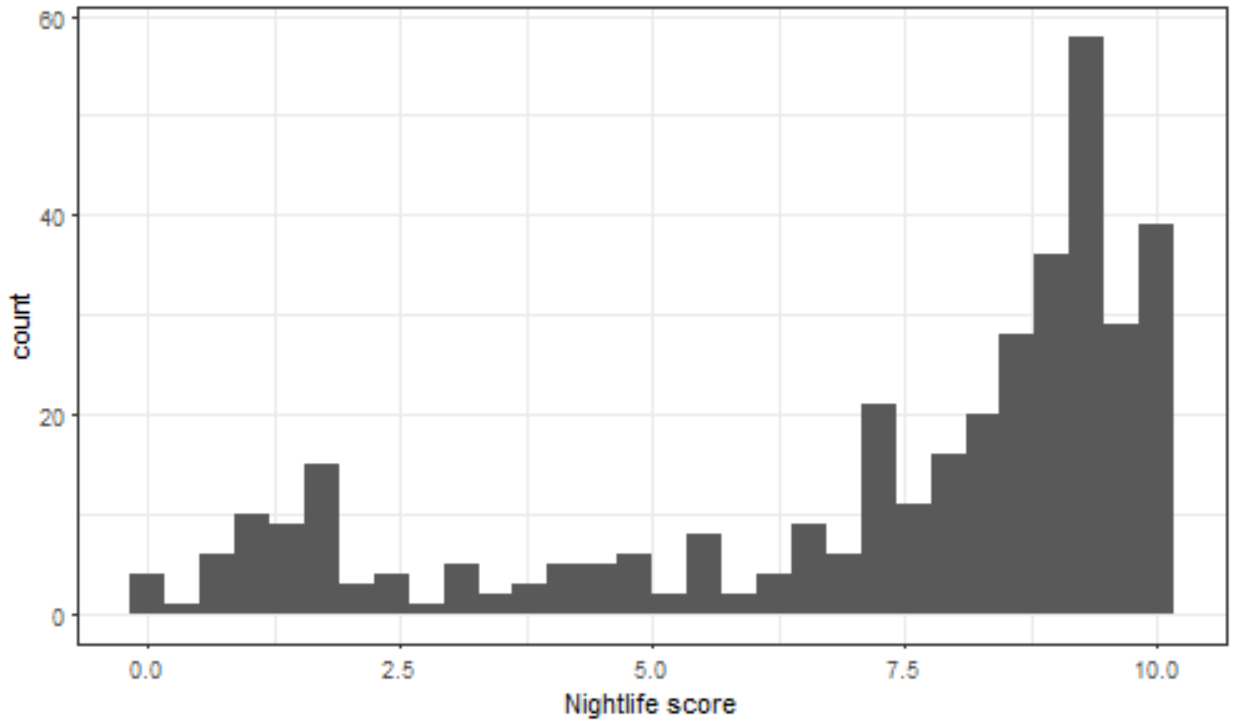

Figure 14: Nightlife score distribution, across all hotels

From the distribution of nightlife scores across the whole dataset, we can see that nightlife is not equally distributed among all hotels. Some hotels have a very low score, and some hotels have a very high score. This property makes the nightlife score very easy to use for recommendation - it discriminates well between hotels that are good for nightlife, and hotels that are bad for nightlife.

Distribution of nightlife score in each city

Now we can see which cities, and which hotels, are better for nightlife.

Some cities have, on average, higher nightlife scores than others. This is to be expected, as some cities have higher relative density of clubs, bars, etc. close to hotels, than other cities. However, we do see some overlap, and a good range of scores within each city.

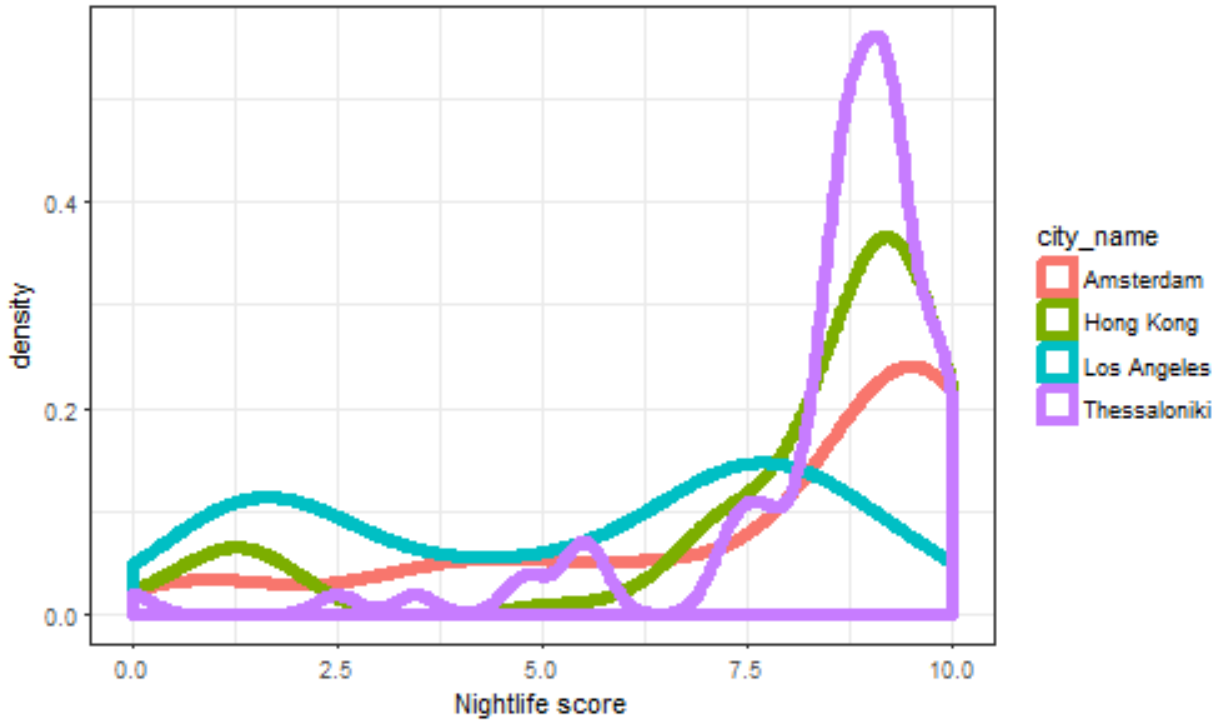

Figure 15: Nightlife score distribution, in each city

Top 5 nightlife hotels in each city

Hotel	City	Nightlife score (0-10)
NH Collection Doelen	Amsterdam	10.0
Radisson Blu Hotel Amsterdam	Amsterdam	10.0
NH Amsterdam Caransa	Amsterdam	10.0
Rembrandt Square	Amsterdam	10.0
NH Carlton Amsterdam	Amsterdam	10.0
Novotel Hong Kong Century	Hong Kong	10.0
Lanson Place	Hong Kong	10.0
Ease Access Wan Chai	Hong Kong	10.0
Crowne Plaza Hong Kong Causeway Bay	Hong Kong	10.0
Mini Causeway Bay	Hong Kong	10.0
Boutique Hollywood	Los Angeles	9.3
Hollywood Hotel	Los Angeles	9.5
Avenue Hotel	Los Angeles	9.8
Beautifull Studio Close To Hollywood	Los Angeles	9.9
Melrose Inn	Los Angeles	10.0
Olympic Bibis	Thessaloniki	10.0
Alcyone Apartments	Thessaloniki	9.9
Studio Floral	Thessaloniki	9.9
RentRooms Thessaloniki	Thessaloniki	9.9
Penthouse With City View!	Thessaloniki	9.9

How good is the nightlife score?

Validity - how well does the nightlife score discriminate between "party people" hotels (should have a high nightlife score) and "business" or "family" hotels (should have a low nightlife score)?

For the nightlife score to be useful, it should distinguish between hotels that have been rated elsewhere as good for party people, and good for business travellers or families. Specifically, "party people" hotels, or hotels for "singles", should have a high nightlife score, while "business" or "family" hotels should have a low nightlife score.

Note that we did not use "party people", "singles", "business", or "family" hotel characteristics when creating the nightlife score. This is because these characteristics have a very high level of missing data, and therefore, a nightlife score created on this very small subset of the actual data would probably not generalise well to all the hotels on trivago.

However, we can use this subset of hotels to test whether our score is useful. And in future, if more data became available on these variables, we could add them to the predictive model, and incorporate them into the nightlife score.

First, we create a new variable that combines our information about "party people", "business", and "family" hotels (it will be dichotomous, with 1 = "party people", 0 = "business" or "family", and NA = missing information):

Looking at the descriptive statistics for this new variable, including missing data: Even when we combine all the information about "party people", "singles", "business", and "family" hotels, we still have 43.88% missing data.

Business or Family or Party or Single	Missing
278	122

But we have enough data on the different types of hotels to use it for testing our nightlife score.

Business or Family	Party or Single
258	20

We can use a simple regression to see if our nightlife score correctly predicts whether a hotel will be identified as "party people" or "business/family" (for the subset of hotels that have this information):

Results of the validity test.

Comparing geo score 1 (latitude and longitude only) with geo score 2 (calculated separately for each city)

Score	Coefficient	SE	p value
Score 1	0.163004877402301	0.101226225206373	0.107331760695892
Score 2	0.229271303565388	0.0797225136789808	0.00402920128986543
Score 3	0.228532266448966	0.124306929482634	0.0659958958159508

We can see that geo score 2 (calculated separately for each city) **significantly predicts** whether a hotel will be labelled as "party people / singles" or "business people / families". Score 1, however, does not significantly predict the hotel label. Score 3 falls just under the threshold of statistical significance (p = .066), but gives a better resolution of hotel rankings within cities than Score 2, and has a similar effect size (coefficient). On balance, the most useful score to travellers would be Score 3. So Score 3 is selected for this presentation.

In general, we can see that hotels labelled as “party people” or “singles”, tend to have higher nightlife scores than hotels for “business people” or “families”. But some “business people” and “family” hotels also have high nightlife scores. There is some overlap between geographical density of nightlife, and hotels that might be suitable for business or family travellers.

Reliability - how well do the hotels map onto nightlife hotspots?

Since our nightlife score is based on geographic denisty of nightlife POIs, we should see a close match between a hotel's nightlife score, and where the hotel is on a map, relative to nightlife hotspots.

We should see that hotels with darker colours (higher nightlife scores) are clustered in the nightlife "hot spots" we identified in the beginning.

From a visual inspection, the hotels with high nightlife scores seem to be clustered in the nightlife hot spots for each city, with some variation. Hotels further outside the city and further from nightlife hotspots have lower nightlife scores, as expected.

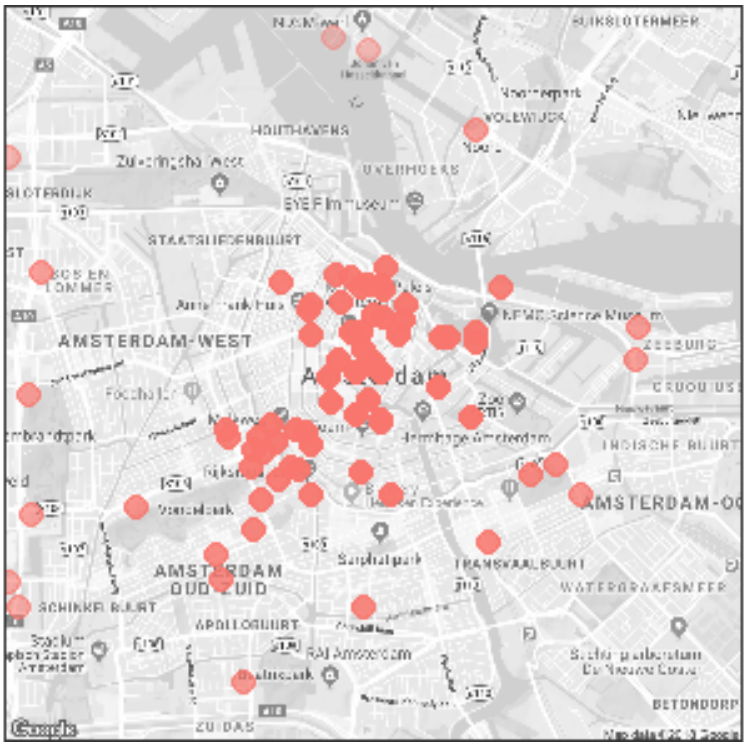

Figure 16: Hotels in Amsterdam, with nightlife scores

Creating a csv with the nightlife score

Finally, create a .csv file with the following columns: hotel_id, city_id, score.

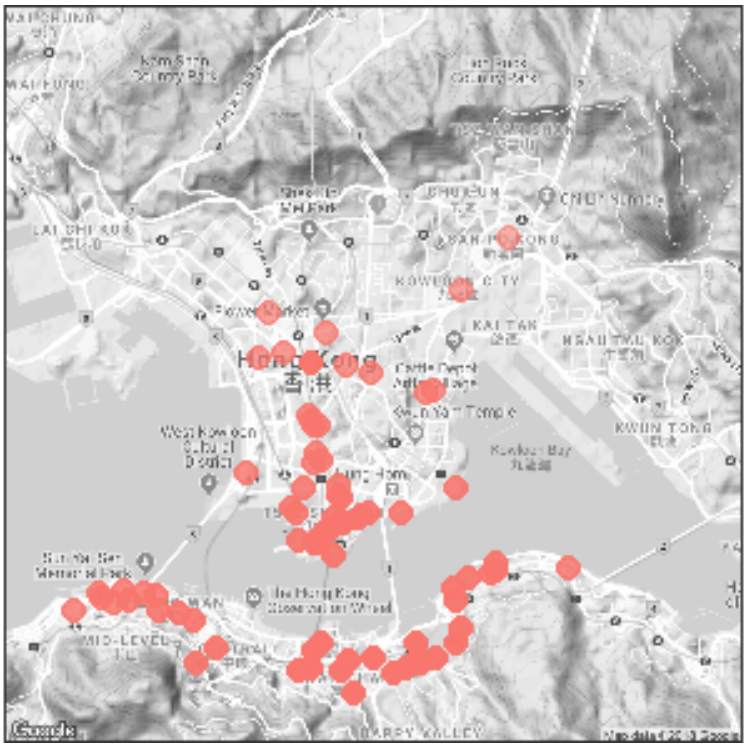

Figure 17: Hotels in Hong Kong, with nightlife scores

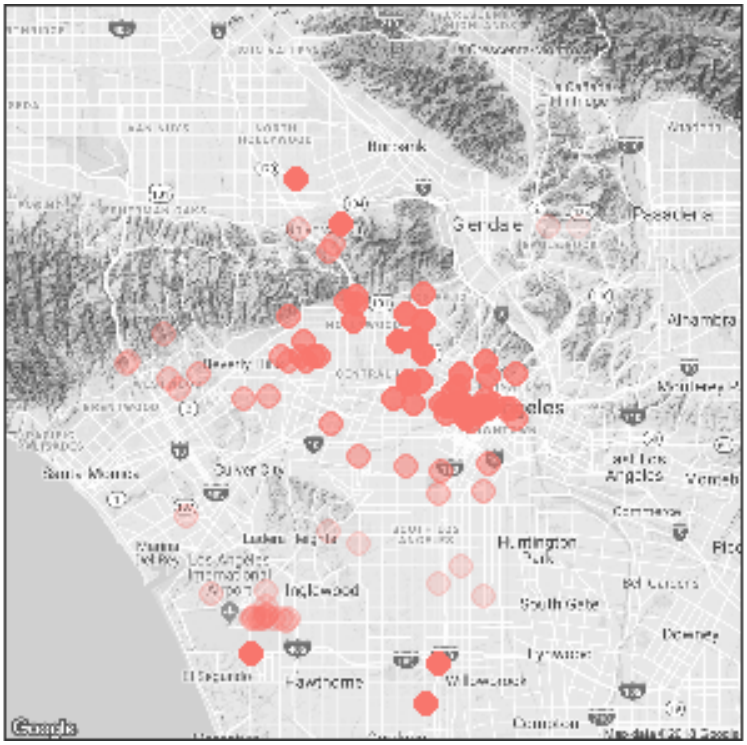

Figure 18: Hotels in Los Angeles, with nightlife scores

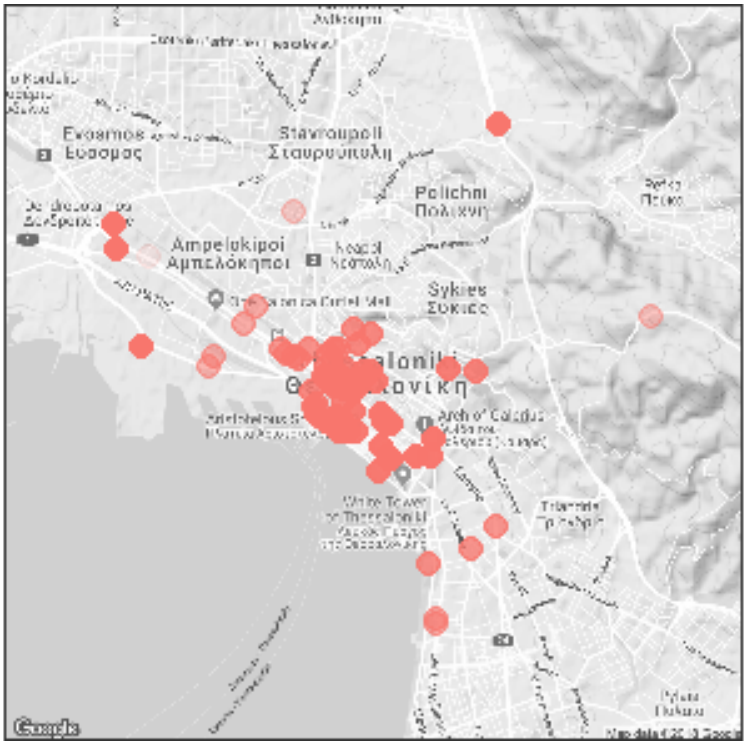

Figure 19: Hotels in Thessaloniki, with nightlife scores