Tecnologia de Processamento de Dados – 2019/20

Mestrado em Ciência de Dados

Project	Student	Student ID	Contribution (in hours)
Data Warehouse Planning (Phase I)	Beatriz Lima David Almeida João Castanheira Pedro Cotovio	49377 54120 55052 55053	10h 10h 10h 10h

1. Describe your original data set. Identify the most important information. Identify missing or incomplete data, and identify possible strategies to use (or discard) them

The main dataset was obtained from http://insideairbnb.com/get-the-data.html, on *airbnb* listings for Lisbon, Portugal. Data is a result of web scraping, and was compiled on the 28th of January, 2020.

- *listings.csv* contains a detailed list of homestays available for booking on *airbnb.com*. A majority of records contain a registration code for the National Register of Local Housing (Registo Nacional de Alojamento Local, RNAL). The document has 106 fields, of which a few will be removed for lack of relevance (e.g.: fields with technical information about data retrieval, fields with identical information for every record), and others will be restructured into creating the facts and dimensions tables.
- *reviews.csv* is a compilation of all reviews for the listings.
- *calendar.csv* has the confirmed bookings for the listings over the next year, describing their availability day by day.

In our exploratory data analysis we encountered the following issues:

- Certain records have some missing location data which could be obtained from our secondary dataset from RNAL, or, if this proves of little use, seeing as only a minority of records has this problem, we may choose to simply discard these from our warehouse.
- In terms of measures, we saw that the weekly and monthly prices for listings were missing in ~95% of entries. In these cases, we will assume that the daily price is scalable for the week and month as well.
- Data pertaining to the description of the listing's property, geographical location, publication date and host is, for the most part, complete, with only a negligible number of records missing it.

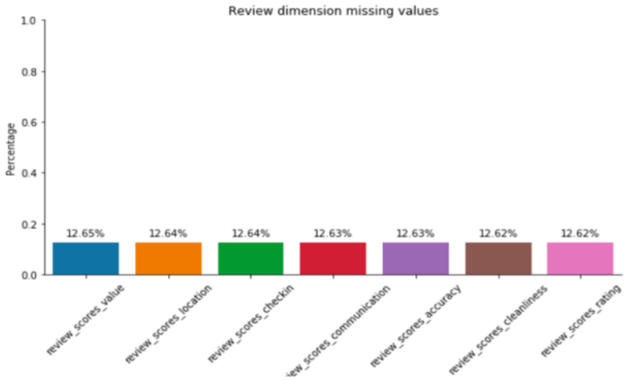


Figure 1 - Review scores' missing values

- As for data about reviews of the listings, the review scores' percentage of missing values form an interesting pattern, suggesting that some listings might not have any type of review measure available. We will further consider if removing ~12% of the data is a safe solution.

2. Describe what other data sets (if any) are going to be used for the construction of the data warehouse and how they will complement the existing information

To enrich the main dataset, we can cross it with the dataset from RNAL (*https://dadosabertos.turismodeportugal.pt/datasets/alojamento-local*) to obtain further information regarding each listing's property, as well as refine already available data, particularly in the case of location data.
We will also use data on Lisbon's neighbourhoods' average price/m_2 (available in *https://geohab.ine.pt/*) in order to have some information on real estate valuation in different parts of the city.

3. Identify and characterize the facts table. Identify and characterize the grain of each element. If more than one fact table is present identify the grain for each of them.

As facts tables, we will have the set of listings available in *airbnb.com*, as well as a set of dates in which the listing is booked.

The Listing facts correspond to each property (located in Lisbon) registered in the *airbnb* platform. The grain is fine, since we have one fact for each property, which is the unit of the sales, and not for a group of properties, for example. We also hold varied information around the listing, such as its host, property information, among other things.

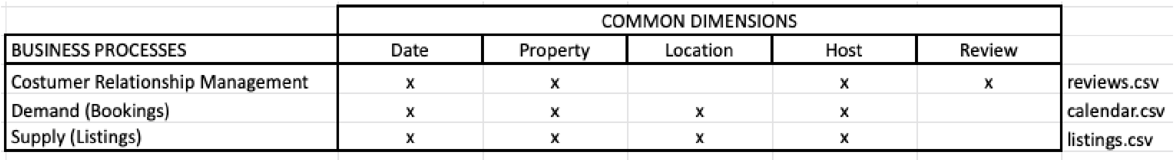
The Availability facts represent whether a listing was booked for each day of the year. Here we have a fine grain, since we will keep a daily calendar of bookings, rather than an aggregation of days in which the listing was booked. We chose this type of grain because it enables us to analyse the bookings at the day level (day of the week, holiday) but also to roll up and analyse at the week, month and year level.
Also, we do not keep information on individual bookings made by each user, only if in a certain day the listing was booked or not. The reason for this is that *airbnb* does not have this data available for open source (since it is monetizable).

4. Identify all the possible dimensions with the data set and the accessory data found. Here we just need to identify the Dimensions, not define their structure.

The dimensions defined are Date, Property, Location, Host and Review.

5. Identify how many data marts are going to be used and what types of business processes might use them. If more than one data mart is present, define the Bus matrix and identify their position in a feasibility/value matrix

We will have two data marts: one centred on the listings fact table, and the other centred on the availability fact table.



	COMMON DIMENSIONS					
BUSINESS PROCESSES	Date	Property	Location	Host	Review	
Costumer Relationship Management	x	x		x	x	reviews.csv
Demand (Bookings)	x	x	x	x		calendar.csv
Supply (Listings)	x	x	x	x		listings.csv

Figure 2 - Bus matrix

- **Customer Relationship Management (CRM):** This business process focuses on customer feedback. It will make use of a data mart centered on the "listing" fact table, in order to evaluate each listing available based on customer reviews.
- **Demand (Bookings):** This business process focuses on the bookings made on the platform. It will use a data mart centered on the "availability" facts table. The goal is to understand what is the market demand in terms of properties, calendar dates and location. A defining question for this process would be: "What are our clients looking for?" (what kind of housing, in what areas and in what times of year).
- **Supply (Listings):** This business process focuses on the description of the listings. The goal is to describe the properties already available for rental. In short, "What kind of listing can we offer our clients?".

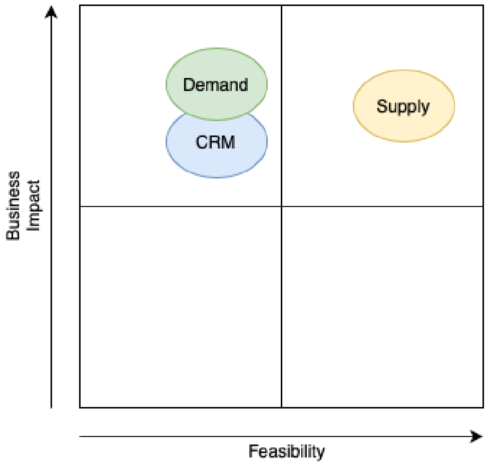


Figure 3 - Feasibility / value matrix

Supply analysis should be easily accomplished since we already have plenty of data available from our hosts and listings as a product of registration on the website.

Demand analysis is not so feasible since we need to parse the calendar of bookings (the dataset is not organized in the most practical way).

CRM is also not so easily done since it could possibly require NLP analysis. There is also the added difficulty of defining what are the different quality ratings in reviews, not only due to the inherently subjective nature of the scoring process, but also because the platform ensures poorly rated listings are removed from the website. Therefore, it is necessary to discriminate amongst relatively highly rated homestays**.**

6. Define the star schema of the proposed data warehouse

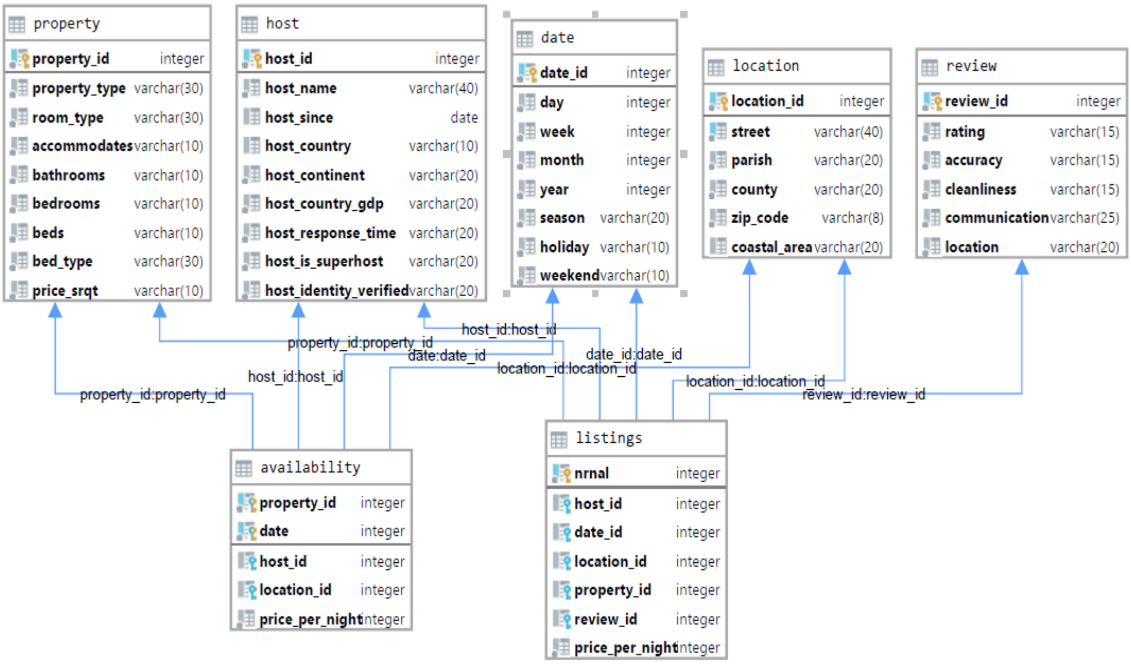


Figure 4 - Star schema

a. For each dimension identify and describe its columns

The **Date** dimension represents the days of the year, and so one entry exists for each given day. The columns day, week, month, and year contain integers directly extrapolated from given date. *Season* is a string with possible values {'Spring', 'Winter', 'Summer', 'Autumn'}, representing the season of the year (in Portugal) for a given date. The columns *Holiday*, and *Weekend* contain strings with a boolean meaning, that indicate if a given date is a holiday or a weekend, in this order. The primary key is the column *date_id* that is obtained by the concatenation of the date into a single integer.

The **Location** dimension represents a given location with a maximum precision at the street level. The column Street is a string containing a given street, with *Parish*, *County* and *Zip Code* being columns that can be extrapolated. *Coastal Area* is a string column that indicates if a given location is near the coast. The primary key is the column *location_id* that is an auto-incremented column.

The **Host** dimension represents the owner of a certain listing. The *Host Name* column indicates the legal name of the host, *Host Country*, *GDP per capita* and *Continent* give geographical context about the host's living conditions. The *Host Since* column gives temporal information about the host, representing the account creation date. *Host Response Time*, *Is Superhost* and *Is Verified* give practical information about the host. The first is categorical and indicates how much time a given host typically takes to answer a rental request. The other two indicate if a host has been awarded Superhost status, and if that host has a verified account. *host_id* is the table's primary key (an incremental integer). In this case we don't have access to confidential information. In a real situation, this could possibly be a citizen identification card number, or social security code.

The **Property** dimension represents the physical location listed in the advertising. All features represent distinct aspects of a given property, represented as strings with values in:

- *Property Type* - {'Apartment', 'GuestHouse', 'House', 'Hostel', 'Room'}
- *Room Type* - {'Entire Property', 'Private Room', 'Hotel Room', 'Shared Room'}
- *Accommodates* - {'0-2','2-4','4-6','>6'}
- *Bathrooms* - {'0','1','2','3','>=4'}
- *Bedrooms* - {'T0','T1','T2','T3','T>=4'}
- *Beds* - {'0','1','2','3','>=4'}
- *Bed Type* - {'Real Bed', 'Pull-out Sofa', 'Futon', 'Couch', 'Airbed'}
- *Price By Square Meter* - {'Expensive', 'Medium', 'Cheap'}

The primary key is a serial integer, *property_id*.

The **Review** dimension represents the average results of reviews conducted by clients. The *Ratings* column represents the average overall rating of a given advertisement, and *Accuracy*, *Cleanliness*, *Location*, and *Communication* represent ratings regarding specific factors of customer satisfaction. All columns are categorical with classifications distributed in bins. The primary key is an incremental integer, *review_id*.

b. Identify and characterize the measures of the facts table

We have two fact tables, Listings and Availability. The first is a representation of an advertisement in the airbnb website, while the second indicates if a given property was not available on a specific date. Both our fact tables have a single shared measured *Price per night* which indicates the price of a given property per night. The Listing's primary key is the advertisement license number, and Availability has a complex primary key.

c. Estimate its size and growth over time (in number of lines in tables)

Locations, Hosts, Properties and Reviews will grow linearly with Listings. From Listings, a steady but slow increase is expected, since it's not everyday that someone chooses to put a house for rent, and most hosts will only have one listing. Date is a representation of days of the year so they increase daily by a factor of 1. Availability will have one entry per each day of the year for each listing, and so it may increase daily by a factor of 1 and exponentially with the growth of properties.

7. Identify the main usage of the data system and propose different analysis queries for the system. It's just necessary to write them down in plain language.

One possible use of this data system in a broad sense would be strategy planning for expansion and monitoring supply and demand.

This data system could be used to do different analysis of our facts, based on each dimension that we identified. Each dimension corresponds to a different perspective in which we could analyse our facts. Based on the dimensions that we have defined, we could analyse both fact tables from a temporal perspective, with the *Date* dimension, and from a location perspective with the *Location* dimension. Furthermore, we have the

Host and *Property* dimensions that could give us insights about the listing host's profile and the property's characteristics, respectively. The Review dimension will only apply to the *Listings* fact table.

With the schema identified, we could do varied analytics. For the *Listings* fact table we could, for example, look into:

- Number of listings by parish or county, and their evolution over time;
- Use the Review dimension to check in which parishes the listings with high review scores are located;
- Check price evolution across differents locations;
- Locations where hosts have listings in comparison to their listed residence ("Are hosts residents in foreign countries?", "Are they investors from richer countries?");
- Analyse the evolution over time and across locations, of types of properties using the *Property* dimension, providing an overview on what the platform can supply.

For the Availability fact time we could analyse, for instance:

- The calendar period in which properties are available or not;
- Locations where the properties are most available;
- Correlation between host profile and demand for their listings;
- The types of properties that are most sought after.