Report Data Visualization: Airbnb Discoverer

Discover data from Airbnb. Visually. Dynamically. Simple.

Jan Lennartz & Hornella Fakem Fosso

Project Structure

Algorithms

The project is structured into the following files:

- **ui.R**: User interface logic
- **server.R**: Server logic (calculations)
- **global.R**: Globally available variables
- **preprocessing.R**: Used for manual preprocessing of calendar data, only needed if new data is imported
- **read_data.R**: Methods to read the input files

Data

All data is contained in the Data folder. It contains four subfolders:

- **Calendar**: Reservations for one year per listing for the given city
- **Listings**: The listings data for one city, e.g. price, host, location, etc.
- **Neighbourhoods**: Neighbourhoods data as geoJSON files. Boundaries of all neighbourhoods in the city
- **POIs**: Data for the points of interest

The Calendar folder and POIs folder each contain a subfolder **raw**. Here are the raw calendar data sets (unpacked) and raw POI data sets, respectively. As it takes a while to process them, each of them has to be preprocessed in order to avoid long waiting times. This can be done within the **preprocessing.R** file with the methods *preprocess_calendar* and *preprocess_pois*.

If you want to add another city you can simply download the files: **calendar.csv.gz**, **listings.csv**, **neighbourhoods.geojson** for a city from the source below and rename them to the cityname. Then you copy them to the appropriate folders and run the preprocessing.

To add the files for the POI to this app, the appropriate file has to be downloaded (based on the country or region). The dowloaded file needs to be extracted and the corresponding csv file (in a subfolder of the extracted folder) needs to be compressed into a .gz format. With the naming convention: nameofthecity-pois.osm.csv.gz. Note: During the preprocessing the input data gets reduced to only the points present in a circle around the city center. This cuts the data size down by a lot depending on how big the initial region of the file was. This file needs to go into the folder Data/POIs/raw and then the preprocessing for this city needs to be done. Analogue to the preprocessing of the calendar data.

The last step is to add the city's name to the *supportedCities* list in **global.R**.

App Structure

Data

To allow for this analysis we made use of two data sources, each delivering multiple input files. In the following the reader gets and overview of these data sources.

Airbnb

This app allows you to interactively discover data from Airbnb. The available data is from the year 2019/2020. The data source is: *http://insideairbnb.com/get-the-data.html*. There were three files used for each city: listings.csv, calendar.csv.gz and neighbourhoods.geojson. The corresponding calendar data is available for one year in advance (i.e. 2021).

POIs

Additionally, there is a second source of data for the points of interests. It is scraped from Open Street Maps (*https://www.openstreetmap.org*) and converted to csv files that are available here: *http://download.slipo.eu/results/osm-to-csv/europe/*. The description of the data can be found here: *http://www.slipo.eu/?p=1551*.

Limitations

It can be easily extended to more cites. The only limitation is your RAM. The only version of this app (*https://madnex.shinyapps.io/Airbnb/*) is limited to 1GB RAM. This is why some big cities would by themselves demand most of the RAM here and only some "minor" cities are shown.

Pages

On top of each page there is a header which allows to switch dynamically between the available cities.

Info Page

An overview over the app and where to find what.

Data Page

Here can be found descriptive statistics visually prepared for the reader's eyes. It is subdivided into six tabs:

- Summary: A simple summary output to get an overview of each variable in the listings data.
- Details: Depending on the selected data set (Calendar, Listings or POIs) this data set is ready to be explored in a dynamic data table.
- Descriptive Stats: Three plots can be found on this tab. The first plot is a horizontal bar plot of the variables "Room Type" or Neighbourhood. The second plot is a histogram of the variable "Price", "Minimum Nights", "Number of Listings" or "Number of Reviews". Additionally, one can adjust the number of bins and a cutoff point. The cutoff allows to exclude outliers and look only at the data points below the q-th quantile. The last plot is a scatter plot which allows for the same selection of variables as the histogram and as well offers a cutoff point. This time the cutoff can be separately adjusted for the x variable and y variable.

- Calendar: The calendar allows to select a host, which is pre-filtered to have between 3-10 listings, so that the plot will be visually appropriate. For this host then, the availability of all his listings will be shown in a gantt diagram.
- Neighbourhoods: On this tab two plots can be found. The first is a pi chart which shows the proportion of number of listings by neighourhood for the top 15 neighbourhoods (based on the number of listings). The second plot is a bar plot of the average price per neighbourhood. Both plots use the same color scheme.
- Hosts: Here, a bar plot of the number of listings per host can be seen. Additionally, a boxplot is shown where the host can be selected and it will show the distribution of the prices of the listings owned by this host.

Explore Page

On this page you can explore the data visually on a map and filter it by some extend.

Linear Model Page

This page allows you to interactively conduct a linear model for the target variable price.