Data visualisation project

Process Book

Airbnb in Switzerland, visualized
Website: https://arnauddhaene.github.io/airbnb-visualized/
Team CAM

Subject and Goal of our project

Due to the current pandemic, the **hospitality sector** has been hit hard. Regardless of recent relaxations of safety measures and the increase in vaccinations, we expect that this year again, holidays in Switzerland will be favored and encouraged by the government for the Swiss population instead of going abroad. To encourage this, we wanted to give an insight into the best Airbnb locations in three major regions of Switzerland: Geneva, Vaud and Zurich.

Through our visualisations, local tourists have the ability to learn more about Geneva, Vaud and Zurich Airbnb offers and find the best regions corresponding to their criteria.

We also wanted to give local Airbnb hosts some benefits when visualizing our project. Therefore, we planned on providing them with more in-depth knowledge about the clients' demands at a regional level, allowing them to tailor their listing to the gathered data. In other words, local hosts will be able to gain insights about the environment where they own a listing and will be able to adapt their offer to the regional demands.

Starting out

The dataset:
The chosen dataset contains listings and reviews of Airbnb listings in Switzerland pulled from **Inside Airbnb** (last scraped in April 2021). The available regions in Switzerland were the canton of Vaud, Geneva and the city of Zurich (unfortunately there was no cantonal information available for Zurich).

Structure of the dataset:
The data was scraped separately from **Inside Airbnb** for each region. Each of the pulled data contains the following datasets:

- Listings: detailed information about the amenities and price of listings
- Calendar: predictions in time of some information about listings (e.g. price, minimum number of nights, etc)
- Reviews: information about reviews
- Neighbourhoods: list of neighbourhoods and GeoJSON file for each

Initial ideas:
As previously mentioned, we wanted our visualization to be appealing to both Airbnb hosts and clients. While perusing through our website, users should first get some general statistics about swiss Airbnb listings. For this we planned a map where they can hover over regions and gather general informations about them. This way tourists could for example know which regions are the cheapest at which time of the year.

For the more avid statisticians, we wanted users of our website to also find a a visualization of the relationship between different features e.g., correlation between price and number of beds. .

Finally, we aimed to give information to hosts about which amenities are most common in the price range of their listing and thus, what they should add to compete with similar places.

Implementation Process

Maps:

To implement all maps on our website, we used version 6 of d3 and SVGs. Regions and locations were plotted by iterating through the datasets with the Promise.all method.

To simplify the datasets and visualisations, we prepared csv files containing all necessary information for listings per region and time period. This way, our website handled small files with only the crucial information and we could easily store them on GitHub. Outliers were also removed e.g., listings with a price per night above $500.

For the first map, users can choose between four features (price per night, number of reviews and min/max number of nights). Users can also choose a specific month from June to December 2021 and a region (e.g., Vaud, Geneva or Zurich). For the second visualisation with maps, users can choose more features to see their correlation.

The maps are colored according to the feature chosen by the user. We chose to color them in a range from light green to dark blue depending on the average value of a feature of a neighborhood.

Sankey flow diagram:

To implement the Sankey flow diagram on our website, we used version 6 of d3 and SVGs. The sankey diagram was plotted by iterating through the formatted datasets with the Promise.all method. A d3-sankey external script was used to aid in the creation of the sankey plots.

To simplify the visualisation, the csv files containing link, node and source information were prepared through extensive pre-processing with the goal of simplifying the information for our website visitors -- indeed, not all amenities are pertinent to hosts. By doing so, we can easily store these files on GitHub. To be coherent with the map visualization, listings with a price per night above $500 were dropped. Individual listings were grouped together in five price ranges. Calculated probabilities with less than 0.10 occurrences for amenities were dropped in order to render a more "breathable" sankey flow chart.

The user can navigate between the 3 Sankey flow diagrams via a selecting menu. When the mouse hovers over a link, its probability of occurrence is displayed and the link between the price category and amenity is highlighted.

All nodes from "Places" are of the same color while all nodes of "Rooms" are different color to distinguish the different rental possibilities.

01
Implementation

02
Data
Pre-processing

03
Interactivity

04
Design

CHALLENGES FACED

40%

DATA

The dataset was considerably large and we had to filter through it to find things we could actually use. We also faced missing data as unfortunately the predictions of features in time (c.f. first map)) was only available for four features unlike the dataset of listings in April that contained a lot more features (c.f. correlation map). This forced us to reduce the choice for users for the map in time.

30%

TECHNICAL

We came with little prior knowledge in JS and had to adapt our ideas slightly to implement them. One thing we had to change due to technical failures was that instead of giving users a continuous time range to choose from, we switched to average information per month. We realised that otherwise it would be too difficult to handle a dataset for each time point and did not find an alternative solution.

30%

IMPLEMENTATION

Overall maps and the Sankey flow diagram were straightforward to implement. What we found more difficult was implementing more than one button for users to select options and having them synchronise. In regard to implementing the Sankey diagrams, it was tricky to narrow down which amenities we wanted to display to keep a pleasing graph, sufficiently simple for visualisation.

DESIGN DECISIONS

Map

The initial state of the map is the canton of Vaud for June 2021 as this is the closest month to the end of the semester.
For the color palette, we chose a neutral sequential color scale so as not to distract the user too much. This color scale also reflected the colors used in our website layout.
We chose a dark blue color for the largest values of features as the latter are generally outliers and so they would stand out.

Sankey diagram

Given the large quantity of data, and its richness-- especially for the amenities (over 80 collected)-- we needed to make a decision on the tradeoff between providing users with an overview of *all* amenities and only a subset of it. In the end, we decided to showcase only pertinent amenities (like basic kitchenware, wifi, ...). This way, hosts can appreciate the breadth of choices available to them. Also, having places and rooms to rent on the outermost part of the graph, with the amenities centered allows to assess the differences for each amenity, between each rental type.

Website

For our website we chose to put all visualisations on the same page so as to facilitate the use. We used a basic website template and applied the same color scheme as in the visualisations. At the beginning of the website, we decided to provide the users with some basic insights into the goal of our project, the pandemic and Airbnb in Switzerland. A menu is provided on the left for seamless navigation on the single page. We provide some of our friends' testimonials from visiting our website.
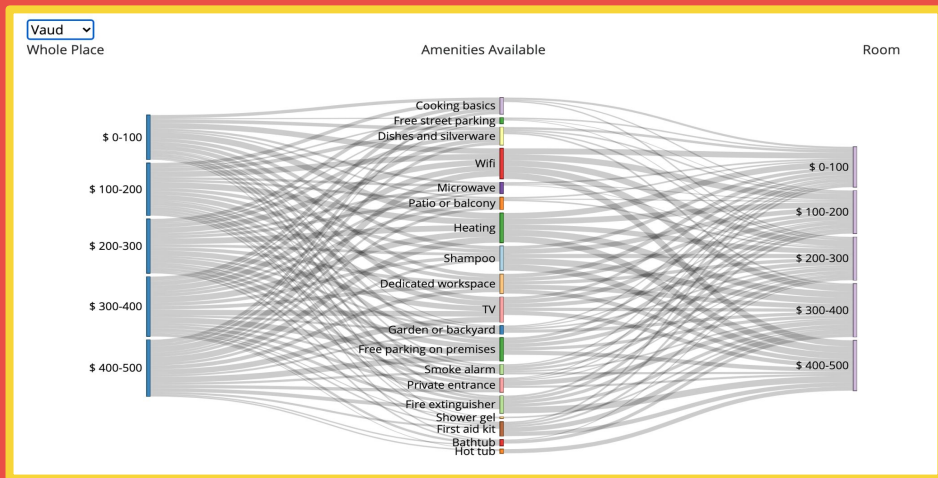
Final result
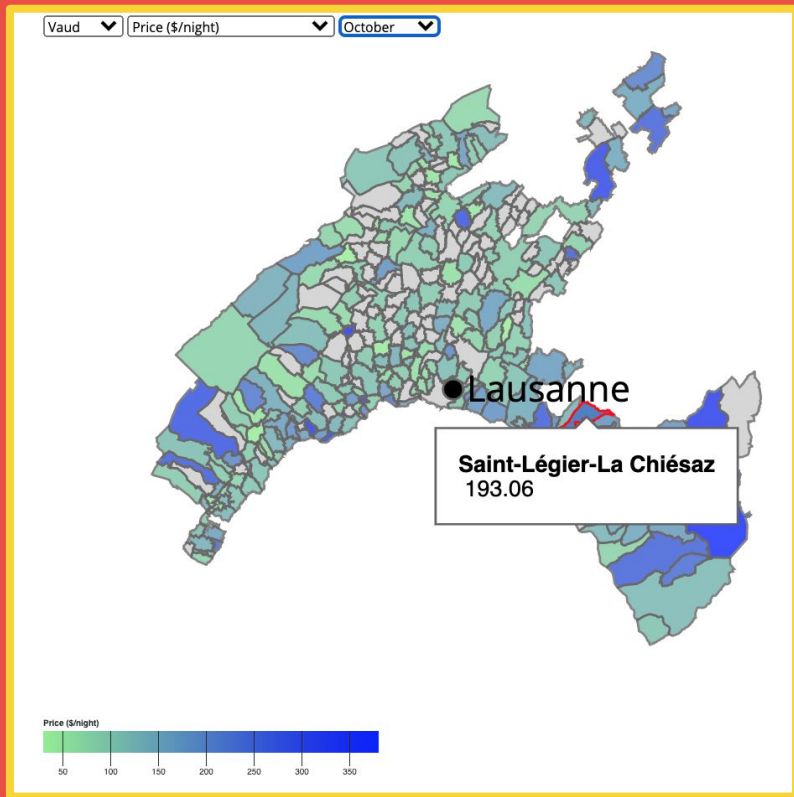
Map 1: a visual of features of Airbnb listings per region and time
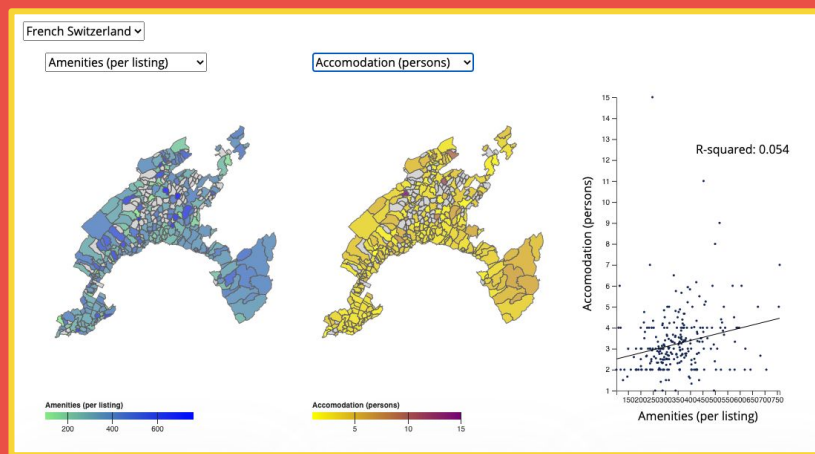
Map 2: a visual of the correlation between features of Airbnb listings

Sankey : a visual of the different amenities' probable availability, by price range, in either room rentals or full place rentals

PEER ASSESSMENT

ARNAUD

Data pre-processing, EDA, implementation of the side-by-side maps and correlation plot, managed overall website implementation and GitHub repositories

CHARLYNE

Data pre-processing, EDA implementation of Sankey flow visualisation, recorded the screencast, completed the process handbook

MARIJN

Data pre-processing, EDA, implementation of first map, wrote the process handbook