WHERE SHOULD YOU BOOK YOUR NEXT AIRBNB?

Data Visualization Report by Team 5: Sameen Haroon | Abhishek Saha | Sharon Xu | Zeyu Zhu

Introduction

The online travel and hospitality space has a large availability of data. This data is flowing in real-time to consumers along with an immediate feedback loop. There is also increased competition (e.g. from online travel agencies) amongst companies which has resulted in more choices for consumers. In this context, consumers are faced with the need to rapidly make informed decisions. As everyday consumers, we've encountered this multiple times when we try to find where to stay, or how to travel for a particular trip.

Within this space, Airbnb is a disruptive and growing force. Statistics show that the number of people using Airbnb in the United States is currently 38 million and is forecasted to reach 43.3 million by 2020.[1] At the same time, there is an opportunity to better enable Airbnb consumer to make decisions about where to stay. For these reasons, and the practical availability of abundant data in the public domain, we chose to use Airbnb listing information as our primary data set.

To help fuel our creative process, we explored a variety of blog posts, relevant visualizations, and more. For instance, we came across TripHappy's use of clustering to help identify neighborhoods to stay in during a trip.[2] In conducting research on the accommodation features that consumers care about when choosing an Airbnb listing, we found helpful articles such as from the Boston Hospitality Review.[3] We also drew inspiration from some visualizations on Tableau public which showcased Airbnb data.[4] Of course, our design choices were also strongly influenced by data visualization principles we covered across class modules.

Data selection and visualization goal

In the early stage of data exploration, we evaluated potential research questions and the audiences we could serve using Airbnb data that was available at the zip-level across the US or with specific location coordinates by city[5]. We thought about combining external data with Airbnb information to find interesting patterns or correlations. For instance, by looking into time series economic data such as loan rates in different areas across the United States, we could research into the disruptive effects of Airbnb in the real estate market space. We also considered mapping listing data over time to relevant social and demographic indicators to identify trends in listing growth. In addition, we looked at ways to enrich listing attribute data by adding in external information such as Yelp data to understand location in terms of restaurant density or crime data to gauge safety.

While we thought through the perspectives of Airbnb hosts, business managers and consumers, we realized that the features available in public data were most relevant and abundant for Airbnb

[1] https://www.statista.com/statistics/346589/number-of-us-airbnb-users/

[2] https://triphappy.com/blog/finding-the-best-places-to-stay-anywhere/15

[3] http://www.bu.edu/bhr/2017/06/07/airbnb-guest-pricing-value/

[4] https://community.tableau.com/thread/207661

[5] Specifically, opendatasoft provides listing data that includes zip-code, and has combined this data across the U.S. Inside Airbnb provides the same feature set with latitude longitude added, but only for a specific city.

consumers. With this in mind, we chose to focus in on building a dashboard for Airbnb consumers, an audience we can relate to, and researched what this audience cared about to inform our visualization purpose. Through our research, we realized that zooming in on a user's booking experience would be the most relevant, and granular data that provided each listing's location within a city would be the best dataset for our purposes.

As such, the specific source of our Airbnb data is the Inside Airbnb website, which provides independent, non-commercial and up-to-date data scraped from Airbnb website for listings in cities around the world.[6] This data includes rich information (100+ columns) for each listing such as users reviews and ratings. For our city, we decided to focus on Los Angeles, a destination that benefits from high tourism, has a large number of Airbnb listings (43K+) but is not as widely studied as other cities such as New York and San Francisco.

With our data, audience, and city selected, we began researching what Airbnb consumers really care about in order to come up with our visualization plan. In addition to using our experience and that of our friends, we sought to minimize bias by randomly sampling 100,000 reviews from all listings. To do this, we used the *ntlk* library in Python for natural language processing and found the relevant descriptive words (over 2 million words after tokenization and filtering out stop words). To "visualize" people's thinking, we used WordCloud and matplotlib in Python to make a word map. As shown in the image below, the frequency of top words or phrases are represented by their sizes so we can get a sense of what consumers value the most. We can see that most reviews are very positive and the word "location" is the biggest one.

Wordcloud showing highest frequency words in 100,000 airbnb reviews for LA listings

This finding aligns with our research. As travelers, we know that location is a major factor in finding a place to stay when visiting a new city. While Airbnb provides neighborhood information of listings to users, it can be challenging to tell apart better vs. worse neighborhoods, and desirable locations can

[6] http://insideairbnb.com/index.html

span across predefined boundaries. Looking at the map view on Airbnb's interface, there's no functionally to differentiate areas and recommend which ones a consumer should stay in.

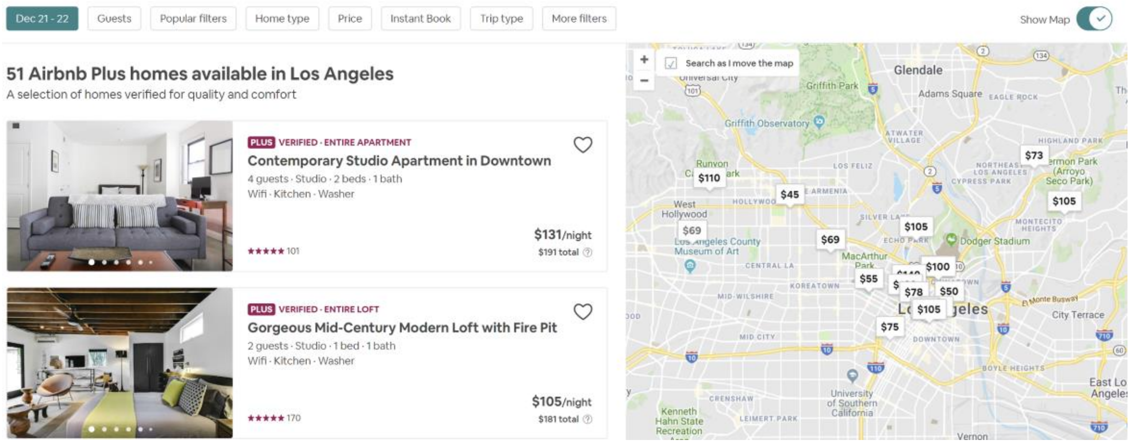


Image of Airbnb listing view for Los Angeles[7]

Moreover, there is limited ability to instantly filter listings. To understand how well some of these filters map to consumer preference on listing features, in addition to external research and friendly conversations, we generated a word-map of 50,000 reviews from listings with high locations (to exclude this as a variable) and filtered out the word location. In the resulting view, we saw that words such as "host", "room" and "space" stood out, but no immediate information on hosts is available in the home page. In addition, there is limited ability to rapidly compare multiple listings across variables.

Wordcloud showing highest frequency words in 50,000 reviews with high locations

With the limitations in the Airbnb consumer experience, we opted to produce a visualization that helps users find the best area to stay in, and identify the right listing within that area to book. Since this dashboard is built using static data (versus real-time), the goal of this visualization is to show a prototype of the content an enriched Airbnb or related-app user interface could provide.

---

[7] Image taken on December 14, 2018 using www.airbnb.com

Data preparation and transformation

To enable this goal, a number of data preparation and transformation activities were conducted outside of Tableau. Our base data preparation was primarily done using Python. We cleaned the listings data and parsed the amenities column to extract the various amenities that were being provided by a listing. After that, we selected the amenities which we wanted to showcase in our visualization (based on research from before). We also selected the other attributes of a listing that we wanted to include in our visualization and excluded the rest. In addition, we introduced listing ranks based on both individual variables such as price and a weighted combination of average ratings and number of reviews (to rank by rating). This was done in Python instead of Tableau given limitations in Tableau's rank function.

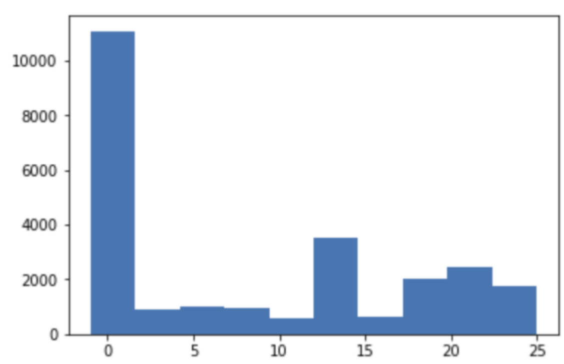
With the cleaned and processed data, we then used the Hierarchical Density-based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm to come up with clusters that represented areas with great locations. To choose the basis for identifying listings with a high review rating on location, we explored the data and observed that many of the ratings were high, with an average rating of 9.6. To only look at the best locations, we filtered the data for listings rated 10/10 on location. Similar patterns were observed on rating data about value for money, and the same filter was applied. These subsetted data were separately passed to the HDBSCAN algorithm to generate clusters of high-density areas.

```
clustering_location = hdbscan.HDBSCAN(min_cluster_size=80, gen_min_span_tree=True).fit(data_location_latlong)
```

```
labels_location = clustering_location.labels_
labels_location
```

```
array([-1, -1, 21, ..., 23, -1, -1], dtype=int64)
```

```
import matplotlib.pyplot as plt
plt.hist(labels_location)
plt.show()
```




```
clustering_location
```

```
HDBSCAN(algorithm='best', allow_single_cluster=False, alpha=1.0,
    approx_min_span_tree=True, cluster_selection_method='eom',
    core_dist_n_jobs=4, gen_min_span_tree=True, leaf_size=40,
    match_reference_implementation=False, memory=Memory(cachedir=None),
    metric='euclidean', min_cluster_size=80, min_samples=None, p=None,
    prediction_data=False)
```

Python code excerpt for cluster creation

To enrich location and listing comparison, we explored a number of additional datasets during our data preparation process with varying levels of success. We tried different packages to scrape data on hotels (lat/long, price etc.) from multiple websites including Hotels.com, TripAdvisor etc. to be able to compare airbnb listing price to hotel price by location. Unfortunately, we were unable to get data at the appropriate granularity (location and price), and instead, only generated insights on overall hotel versus Airbnb prices for Los Angeles.

Another attempt, which was more successful, was to get data on major attractions by city. This data was compiled manually using sources such as Planetware, Tripadvisor to get the top tourist attractions for LA. After this, we calculated distance matrices which captured the radial distance of each major attraction to the cluster centroid. We then set a cutoff of 5 miles to determine the 'nearby' attractions. Finally, we ranked these 'nearby' attractions in order to choose the top 3 nearest attractions for each cluster.

cluster_loc	cent_longit	cent_latitu	Universal S	Griffith Par	Disneyland	Hollywood	The Getty (	Santa Mon	Natural His	California S	The Broad	The Origina	Los Angele	Venice Bea	Long Beach	Page Muse	The Nether	Museum of	1st_closest	2nd_closes	3rd_closest
21	-118.355	34.10379	2.38	3.26	32.15	1.70	6.96	9.74	7.13	7.26	6.91	2.24	2.77	10.54	25.05	2.78	15.38	6.94	Hollywood	The Origina	Universal St
14	-118.466	34.00127	11.45	12.39	34.02	10.09	6.07	1.92	10.20	10.35	12.89	7.75	7.49	1.08	22.79	7.66	21.16	12.84	Venice Bea	Santa Monica	
5	-118.399	33.87938	18.09	17.35	27.97	15.31	15.12	11.05	11.40	11.43	14.80	13.49	12.97	8.56	14.56	12.99	29.83	14.72			
19	-118.35	34.08839	3.44	3.45	31.26	1.26	7.17	9.39	6.08	6.22	6.18	1.30	1.77	9.95	23.96	1.74	16.48	6.19	Hollywood	The Origina	Page Museu
9	-118.222	34.10405	7.90	4.58	26.64	6.17	14.56	16.51	7.17	7.14	3.81	8.27	8.35	16.53	23.11	8.15	19.73	3.90	The Broad	Museum of	Griffith Parl
24	-118.333	34.17034	2.51	4.17	34.34	5.37	9.90	13.81	10.94	11.03	9.33	7.00	7.51	15.01	28.93	7.49	12.06	9.39	Universal S	Griffith Park and Griffit	
8	-118.12	34.08013	13.98	10.65	21.83	12.01	20.38	21.71	10.65	10.54	7.70	13.83	13.77	21.26	21.63	13.56	25.20	7.75			
13	-118.407	34.03272	7.90	8.40	31.88	6.10	5.52	4.94	6.84	6.99	9.07	3.77	3.47	4.96	22.21	3.63	19.27	9.04	Los Angele	Page Muse	The Origina
4	-118.661	34.26842	19.80	23.18	52.99	22.59	16.38	19.79	27.54	27.70	27.81	21.94	22.35	22.27	44.03	22.53	11.60	27.83			
3	-118.611	34.1621	14.84	18.06	46.50	16.86	9.29	12.02	21.02	21.18	21.96	15.63	15.94	14.48	36.57	16.15	13.09	21.97			
1	-118.225	33.68261	32.37	30.33	19.74	29.00	31.56	27.89	23.40	23.32	25.76	28.04	27.49	25.43	6.65	27.41	45.33	25.67			
25	-118.255	34.08302	6.82	3.50	26.88	4.30	12.64	14.25	5.00	5.00	2.00	6.13	6.14	14.18	21.95	5.93	19.59	2.09	The Broad	Museum of	Griffith Parl
15	-118.338	34.12279	1.36	2.22	32.27	2.15	8.18	11.30	7.88	7.99	6.92	3.75	4.24	12.17	25.91	4.21	14.64	6.97	Universal S	Hollywood	Griffith Parl
6	-118.128	34.16611	13.06	10.46	27.27	12.57	20.60	23.16	13.86	13.81	10.45	14.85	15.02	23.35	27.46	14.84	21.56	10.53			
12	-118.461	34.07502	7.54	9.62	36.02	7.67	1.25	4.22	10.65	10.81	12.14	5.73	5.87	6.14	26.41	6.09	16.06	12.13	The Getty (	Santa Monica	
11	-118.458	34.04755	8.66	10.21	34.96	8.04	3.01	2.72	9.92	10.08	11.90	5.81	5.77	4.29	24.83	5.98	17.96	11.87	Santa Mon	The Getty (	Venice Beac
16	-118.38	34.1499	1.71	5.10	35.28	4.91	6.92	11.05	10.59	10.72	9.93	5.51	6.06	12.49	28.54	6.13	11.90	9.97	Universal S	Hollywood	
18	-118.383	34.06985	5.03	5.75	32.07	3.52	5.41	7.09	6.56	6.71	7.70	1.31	1.45	7.70	23.67	1.67	17.05	7.69	The Origina	Los Angele	Page Museu
10	-118.448	34.16189	5.65	9.01	38.81	8.33	5.31	10.16	13.56	13.71	13.54	7.98	8.47	12.20	31.03	8.62	10.09	13.57			
17	-118.293	34.06808	5.95	3.38	27.84	2.66	10.53	11.84	3.57	3.65	2.63	3.88	3.79	11.74	21.49	3.58	19.21	2.64	The Broad	Museum of	Hollywood
2	-117.828	34.03422	30.97	27.68	16.24	29.00	37.30	38.06	26.47	26.33	24.27	30.67	30.54	37.15	27.24	30.32	41.03	24.29			
23	-118.266	34.04562	8.13	5.29	25.65	4.86	12.36	13.04	2.40	2.38	1.08	5.73	5.49	12.55	19.57	5.28	21.35	1.01	Museum of	The Broad	California S
0	-118.194	34.6108	33.95	34.74	57.45	36.66	39.54	44.30	41.47	41.51	38.64	38.49	39.00	46.03	58.06	38.97	26.06	38.73			
22	-118.194	34.04839	11.04	7.70	22.73	8.30	16.35	17.15	5.86	5.75	3.25	9.69	9.52	16.55	19.16	9.31	23.64	3.27	The Broad	Museum of Contempor	
7	-118.054	34.11436	17.21	14.10	22.31	15.79	24.18	25.88	15.05	14.95	11.98	17.81	17.81	25.57	24.80	17.61	26.97	12.03			
20	-118.31	34.085	4.42	2.25	29.34	1.18	9.45	11.31	4.89	4.99	4.04	3.03	3.16	11.54	22.91	2.98	17.70	4.07	Hollywood	Griffith Par	Page Museu

Distance matrix and top 3 attractions by cluster

Tableau visualization buildout: design choices and functionality

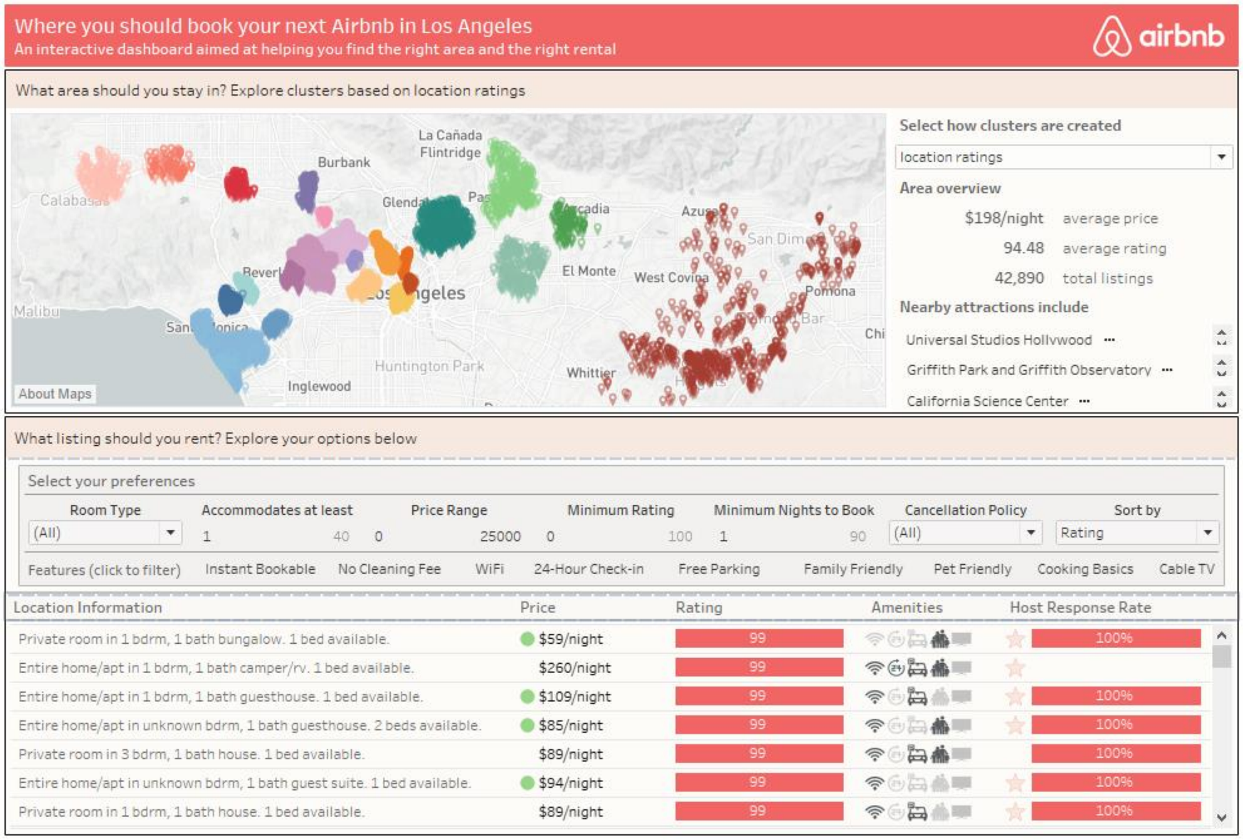


Image showing full dashboard with default settings

To build our visualization, we decided to leverage a warm palette of pinks, yellows, and grays that complemented the base pink color associated with the Airbnb brand. We used the Tableau font family as our base font since it is well suited for visualizations, with font size, shade and/or boldface used to establish content hierarchy. In addition, tiled containers were used to nest relevant content and further emphasis hierarchy where appropriate through background shading and/or borders[8]. For instance, since our visualization is designed to address two specific customer needs, namely finding the right area to stay in and then finding the right listing, we treated these as two separate parts of our dashboard using section headings and relevant text. Design and interactivity choices for the key visual elements associated with each customer need are explained in detail below.

Finding the right area:

As mentioned earlier, to help consumers find the best areas to stay at, we leveraged a clustering algorithm (HDBScan) to identify sets of relevant areas based on listing location ratings and ratings on value for money respectively. Since geographic data is best represented on a map, we plotted each point within a cluster on a map zoomed in on Los Angeles, our city of interest. The added value of showcasing each point, versus a radius circle or polygon, is that it lets you see each specific location as

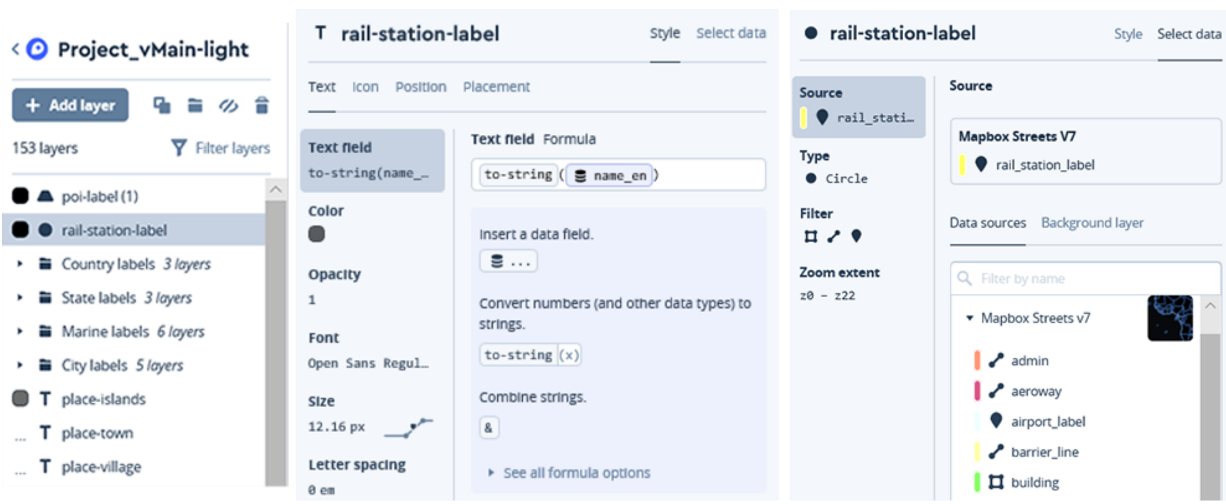
---

[8] Nested containers are also better suited for automatic resizing of the dashboard. The dashboard was designed for the laptop it would be displayed on, but this enables useability across devices and screens.

well as the true size and shape of the cluster. Since each listing in the cluster represents the best location rating i.e. has a rating of 10, the density of the points in a cluster helps show the consistency/strength of the cluster's location as well as the quantity of available listings that have great locations. Using the specific location also allowed us to display listing neighborhood in the tooltip, to help a user orient where a listing is based on known neighborhood classifications as well.

To derive additional information from the map, we opted to use a Mapbox custom template that was ported in using the Tableau mapbox API. The default light template was selected as a starting point, since this minimizes the use of color to drive attention to any data visualized on the map. To add in relevant information for our specific use, we created custom layers that updated the map to show points of interest as well as metro stops with intuitive icons. We further adjusted city, town, village etc. label sizes and fonts at different zoom settings to ensure these were sufficiently light/minimal for our visualization.

Image showing map overall (with tooltip) view versus zoom-in view



Customizing map designs with 153 layers of geo-information

Unsurprisingly, in a large sprawling city such as LA, our clustering analysis produced a large number of areas e.g. 25 based on location ratings. While each of these represents a separate area cluster (that usually spanned multiple neighborhoods), the need for 25 categorical variables encoded with unique colors made for some challenging visualization decisions. Since all clusters were based on listings with the same location rating, it wasn't possible to use color to further differentiate between clusters based on location quality metrics. As such, we opted to use similar colors with different shades for clusters that fell into various 'buckets' created based on our understanding of LA and general proximity. For instance, clusters near Hollywood or Universal Studios were colored in various shades of purples and pinks (for glamor), while clusters near Santa Monica or Venice Beach used blue (for water). A limitation of this approach was that since we assigned cluster colors manually based on the default numbers associated with the location rating classification, these colors were preserved for each number in the view that shows clusters based on value for money as well. Visually inspecting the resulting colors, however, we saw that the buckets translated well across both views, so we preserved the coloring as is, although distinct cluster numbers for each clustering basis could be used and mapped to separate clusters as a workaround. An additional idea for possible enhancement would be to use a monochromatic scheme that updates cluster color based on a dynamic variable (i.e. parameter) selected by the user, such as average listing price or average overall review.

Speaking of dynamic variables, an important part of our cluster map visualization involved interactivity. To start, users are able to select either location ratings or value for money ratings as a way to find the best areas, as mentioned earlier. At the same time, recognizing that these views introduce substantial filters in the data, a third view that lets users see all listings grouped by neighborhood was included as a parameter value as well. For the neighborhood argument of the parameter, a monochromatic color scheme was used to avoid adding in substantial contrasting colors while enabling some visual differentiation of neighborhoods in a small space on the map, although color-based differentiation was not prioritized for this view. This parameter is placed immediately to the right of the visualization to help users become aware of their ability to update the map clusters right after looking at the map.

Additional interactivity is introduced through selecting a point on the map. The cluster the point is part of is highlighted visually, and is used to filter all other content on the dashboard. Moreover, the section to the right of the map shows relevant information about the selected area (the default is data for all of LA otherwise). The average price and average rating of a listing in the selected area are shown as relevant summary information, as well as the total number of listings in the map. Recognizing that de-averaged ratings can be useful, the tooltip for average rating includes a breakdown of the rating by all available subcategories e.g. accuracy, cleanliness. In addition to summary statistics, we also added in descriptive information that shows the nearby tourist attractions each cluster is near to, to help users distinguish between areas further.

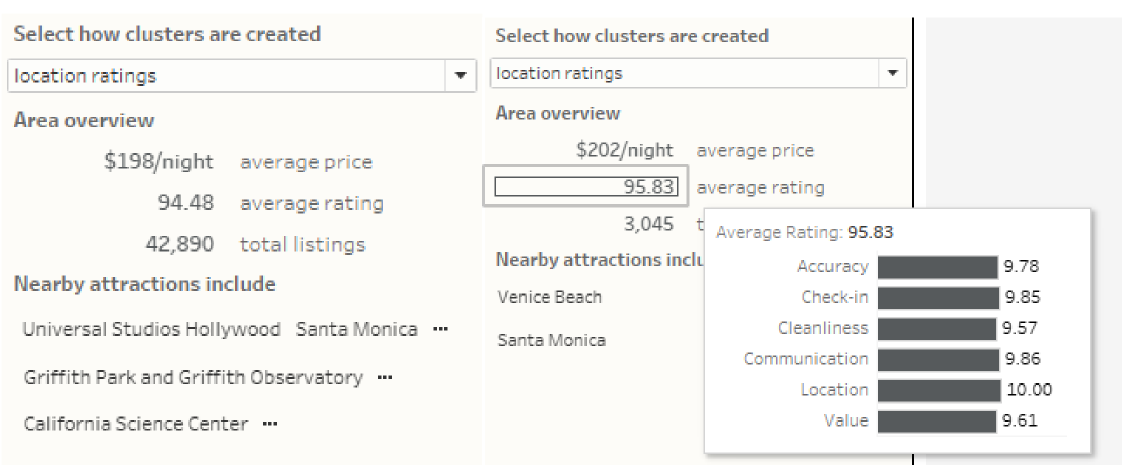


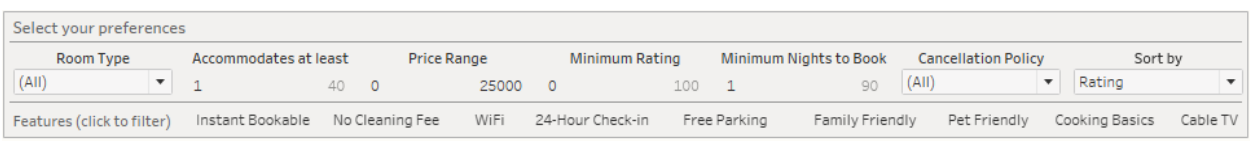
Image showing listings and tourism information of the selected area

Finding the right listing

The lower half of the dashboard (ideally to be used after an area is selected on the map) is focused on enabling users to compare and choose the right listing. The first part of this section is dedicated to an interactive panel that lets users select listings based on their needs. To help view the panel as an embedded tool, container spacing, shading and borders are used. Moreover, the panel is divided into two subsections using lines (that are actually text boxes with fill shading and a 1px height). This panel looks to add value through additional features that aren't available for filtering on airbnb's website such as 24-hour check-in, and also provides prominence to a broader list of relevant features than shown in airbnb's default view.

In the first subsection, users can curate listings based on relevant criterion such as the room type needed (e.g. entire home vs. shared room), the price range, and the minimum rating. The specific variables were included and their order was determined based on the interpretation of external research and from the team's own experience with Airbnb listing selection. While the default listing view sorts on rating (and the number of reviews), users are also given the choice to reorder listings based on the lowest price.

The second subsection of the panel shows listing features that rely on a single user click to filter i.e. represent yes/no values. Since each of these features represent a distinct column in the data and multi-level parameter selection is not supported in Tableau, custom text fields were created for each relevant variable to indicate its presence, displayed in a separate sheet and then assigned an individual filter action. This approach was preferred to using separate filters for each column since the latter would take up substantial dashboard real estate.



Panel containing filtering and sorting functions

Based on the filtering and sorting done using the panel, the view that provides relevant listing options and information is updated. To be able to readily compare variable values across listings, a tabular format is used, inspired by the dashboards discussed in Stephen Few's Information Dashboard Design.

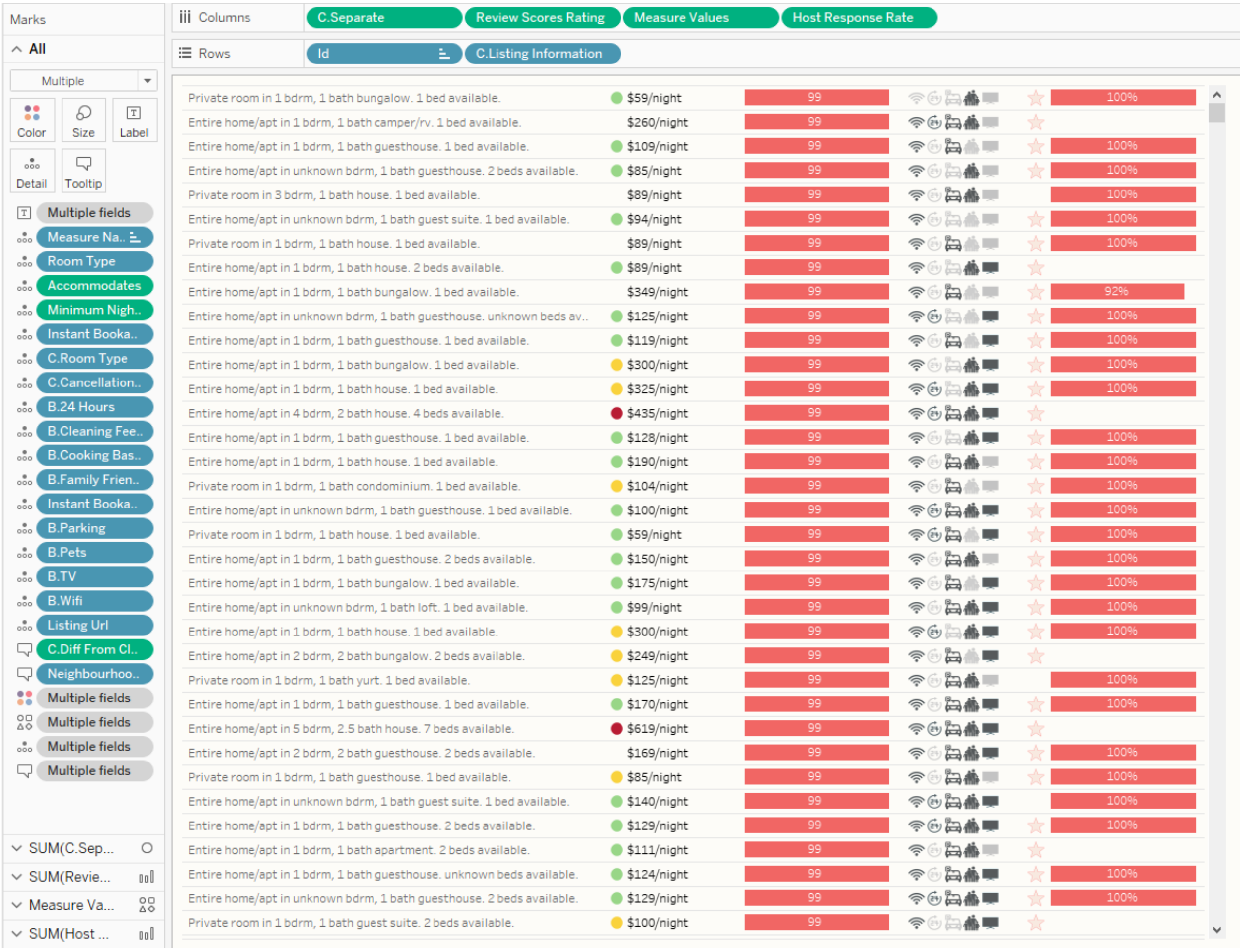


Table sheet entire view

For each listing, basic relevant information is summarized through a calculated field that combined the room type, number of bedrooms, number of beds and number of bathrooms into a single string. Additional relevant information on price, listing rating, available amenities and the rental host was included, chosen once again based on a combination of research and experience.

The listing price is displayed in text format since the price value is important to users, but we also included a colored indicator for relative performance. More specifically, this shows if a listing price is a good deal (green, price below 20%), a somewhat high price (yellow, price above 20%) or an extremely high price (red, price above 60%) when compared to the average price for listings of that room type in the cluster area. The specific relative value of the price is also included in the tooltip, along with the cleaning fee, if any.

The rating is shown as a bar chart, that enables rapid visual comparison when values are different across listings, and can be used to indicate that listings have comparable customer feedback when the bars are of similar length. As with average area rating, a tooltip is included that shows the de-averaged rating by subcategory.

Although a direct filter on select amenities is available in the earlier panel, being able to quickly gauge the presence of key useful features such as wifi, 24-hour parking and family-friendliness was deemed valuable. To show this content efficiently and consistently across listings, an icon was used to encode each feature, with a tooltip that lists the full set of nine features a listing has available.

The last section of this table shows information on host performance, which was seen as useful based on our natural language processing analysis of 50,000 reviews for listings in LA. More specifically, the host response rating is included as a bar chart, as well as a star when the host is known to be a superhost. Hovering on the hostname also provides this information, and also shows the number of Airbnb listings a host has.

Displaying all of this information in a single sheet proved challenging, since Tableau doesn't support using so many different visual encodings in a table. To get around this, we tried a few different methods, and ultimately settled on an approach that leveraged a combination of dummy calculated fields, measure names, values, and more. For instance, for showing the indicator before the text, a fixed value field was used as a column with the indicator displayed using the mark. For the amenities column, we created custom fields that assigned a specific value to each amenity and then displayed those values in a single plot using appropriate shapes. The star shape indicating someone is a superhost was also added to this plot (with space inbetween), because this approach required even column widths across the table.

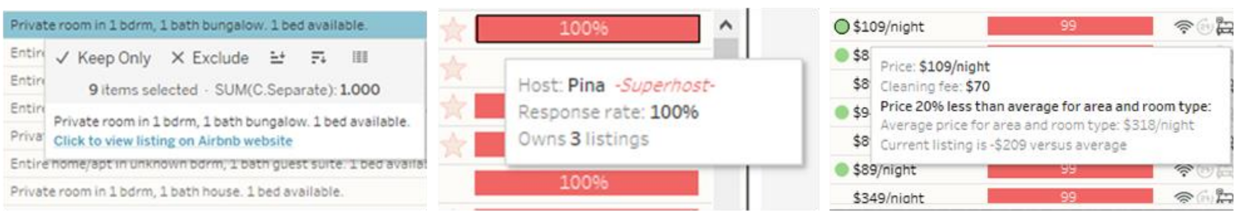


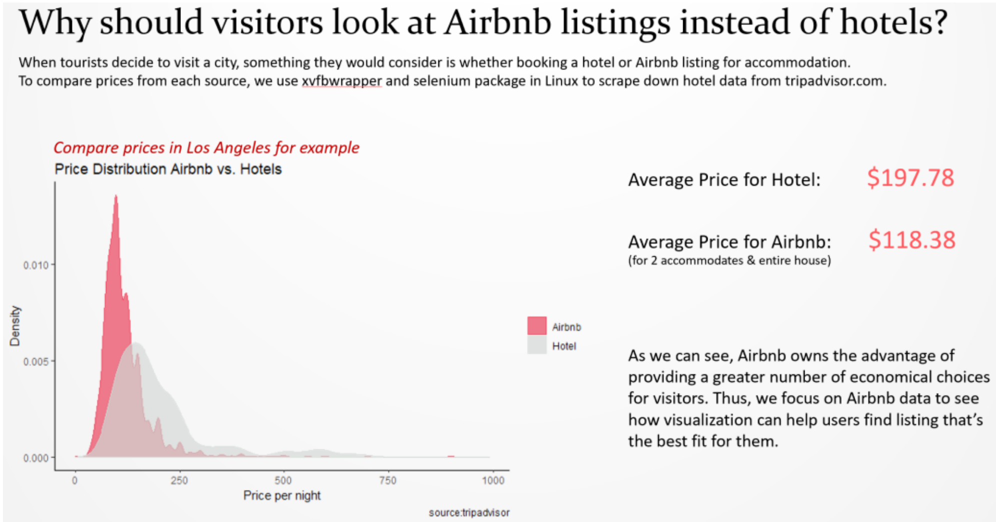
Image showing tooltips in listing table

It is to hoped that this content sufficiently enables a user to narrow their choice of both area and listing. Of course, there are limitations to the full content on a single listing that can be displayed in one dashboard, and even when right potential listing is found through the listing table, users need a way to look up further information and lock in their choice. To enable this, the tooltip for each listing also includes a URL action that opens up the specific listing on Airbnb in a user's web browser. Through this, users can review a listing in additional detail, explore available images, and make the final booking.

Conclusion

It is to be hoped that this visualization provides a compelling demonstration of the way Airbnb users be further enabled in their decision-making process. In the process for building this dashboard, we learned a number of valuable lessons worth sharing.

Firstly, while there are any number of interesting questions for a given topic, it can be difficult to obtain all the data you need to effectively answer them. In our unsuccessful efforts to scrape hotel data, we learned that acquiring relevant data can be challenging, possibly unfruitful, but always informative. For instance, we validated our intuition that Airbnb prices are often lower than those for hotels (when specific location isn't factored in).



What we learned through unsuccessful web-scraping attempts

Secondly, identifying the target audience early is a critical part of successful dashboard design, and factoring in data availability in making this decision is helpful. While we considered building visualizations that focused on Airbnb business managers, hosts, or consumers, the fact that we had abundant rich data for the latter drove our planning, and enabled our final design.

At the same time, in our efforts to execute our design, we learned there are important tradeoffs between rich dashboard functionality and the variety of quantitative messages and dimensions that can be included. For instance, we had a version of our dashboard that also had data for Washington DC and Chicago included, with a dynamic button used to update the view for a specific city. Ultimately, the added load times on dashboard interactivity made this level of detail infeasible, and since our overall goal was effectively captured in content for a single city, we made the decision to exclude DC and Chicago.

At the same time, in an improved version of this dashboard, additional data could still be added. For instance, our current visualization is based on static data that is updated monthly, and has limited utility for real-time trip planning. A live version of this dashboard could look to directly scrape Airbnb's website to enable users to make location and listing choices, and add in relevant information on location availability. It would also be helpful to add in additional hotel data for more comprehensive comparisons if possible. Moreover, while we included data on top tourist attractions and categorized which ones a cluster was close to, there is room for much richer content here. For instance, adding in information such as restaurant proximity (using Yelp data) or transit proximity (using public city data) could add value. Ideally, if this information and additional content on neighborhoods and listings could be used to provide the key features associated with each cluster, this could be represented as a word cloud in our dashboard.