Find Your Next Airbnb Vacation in Seattle

Link to the Final Tableau Dashboard

Ameya Bhamare, Poornima Muthukumar, Logan O'Brien, Marques Chacon,

Carolina Mack, Saumya Nauni

Summary

The purpose of this data visualization project is to use Airbnb listings and reviews data to develop a cutting-edge dashboard that will help travel enthusiasts or frequent Airbnb users make informed decisions when booking listings in Seattle. Using our visualizations, any user can compare listings and neighborhoods, and choose the best listing for them based on their individualized travel wants and needs. We also want to give users more control over search criteria and provide unique, data-driven insights that you will not find on Airbnb.

Background

Research

We conducted background research to determine if similar projects already exist. In doing so, we discovered that most Airbnb analytics projects/software available are catered toward Airbnb hosts. Examples of such projects are 'Airbnb Calculator', 'AirDNA', and 'AllTheRooms'. There didn't seem to be much work centered around providing higher-level analytics to potential Airbnb renters to help them better understand the market and make smarter bookings. Our project will attempt to provide that service.

Data Source

To meet our objective, we pulled data from Inside Airbnb, which contained information for various cities around the world. Each city contained seven files about different aspects of Airbnb listings. Listings.csv.gz, calendar.csv.gz, and reviews.csv.gz were zip files providing detailed information describing listing characteristics, availability over the next year, and reviews respectively. We primarily focused on these files for our analysis as they had the most relevant use. Furthermore, we limited our scope to Seattle listings only, but our analysis applies to other cities as well. The next section will profile each dataset in more detail.

Data Profile

The listings zip folder contains an 18MB dataset, with rows for each listing in the Seattle area. The columns describe 96 attributes for each listing, including metadata (when the data was scraped and the source of the scrape). Some of the attributes used in the analysis are *price* (continuous), *longitude* (continuous), *latitude* (continuous), *listing_type* (categorical), *is_superhost* (categorical), *neighborhood* (categorical), and *ratings* (continuous), among others. After sifting through the data, we determined that there were around 20 columns most relevant for our purposes, including the ones listed above.

The calendar zip folder contained a large 96MB dataset, with 7 attributes. Each row depicted the *availability* (boolean) and *price* (continuous) for each *listing_id* (discrete) and *date* (datetime). The other columns indicated the *minimum* and *maximum* nights (integer).

The reviews zip folder contained a 110MB dataset with 6 attributes, where each row was a detailed review given by a guest. Key attributes include *date* (datetime), *listing_id* (discrete), *reviewer_id* (discrete), and *comment* (textual), which is an individual comment for a listing.

Data Cleaning

Most of the data cleaning was performed on the listings dataset, as it had the bulk of the messy data. Columns such as *bathrooms_text* and *price* needed to be parsed and converted to numeric data. For example, the string “2 baths" was converted to the integer 2. Similarly “$45.00” was converted to 45.00. Additionally, taking a closer look at the *bedrooms* and *beds* column revealed that they occasionally had different values. This was particularly confusing because we could not distinguish between the two columns. However, the *name* column, which described the listing’s title on Airbnb, often referred to a “3-bedroom apartment” or “2-bedroom home.” After filtering on listings that had mismatched *bedrooms* and *beds* values, we were able to conclude that the information from the name column described the *bedroom* column, so we decided to remove the *beds* column.

The most difficult data-cleaning task involved parsing the *amenities* column. Within each entry of this column was a list of amenities that the listing was reported to have. However, these lists were not clean, as each entry was encased with quotes, and some entries included escape characters or unicode characters. To remedy this issue, we used a combination of regular expressions that eliminated the special characters and made sure that each entry described a particular amenity. After parsing the lists in the *amenities* column, we put this in a separate table for easier readability and merged it as necessary when creating the dashboard.

Design Process

User Tests

For our design evaluation experiment, we created prototypes on paper. We printed a map of both the US and New York City and encoded the listings using dots. For the experiment, we picked three classmates and three people outside of class. Specifically, outside of class, we chose a travel / Airbnb Enthusiast, a Data Analyst, and a Software Engineer. We first gave the participants a high-level overview of our idea, and then gathered feedback for the prototypes based on usability, functionality, and their ability to complete three tasks. We simulated user interaction by showing the prototypes — one after another — based on user selection. First, we had them select New York City. Then, we had them filter listings for different ratings. Finally, we asked them to view the statistics for a particular listing.

Our prototype uses clusters of dots to depict all listings in the US. The users can then click on a particular city to find the listings within that city. When the user selects New York, we give the users a zoomed-in view of all listings in New York City (once again distributed as dots). The user can further filter listings based on - Cost, Rating, Availability, Popularity, and different amenities such as - A/C, Washer/Dryer, Wifi, etc. The first two visualizations help the user explore listings geographically based on the City/Neighbourhood that they would like to visit. It gives the user an overall sense of how the listings are distributed across neighborhoods in a city. Finally, the user can click on any listing to see a pop-up of listing statistics such as - price trend over time, reviews word cloud, histogram of the cost compared to similar listings, etc. We asked users to complete three tasks:

1. Navigate to the listings in New York City
2. Find ratings information for the listings
3. Find the price history for a 4-star listing of your choosing

User Feedback Summary

Throughout most of our evaluations, it was clear that users could intuitively navigate through each of our pages. People also liked our word cloud idea and found it novel and useful. However, it was also clear through the feedback we got that we could improve certain aspects. For one, the heatmap depicting different listings would be best utilized with a monochromatic scale. Furthermore, almost everyone suggested adding or altering our current filters. Shweta suggested adding a neighborhood filter, while Ayesha suggested adding filters for “Unique Stays” and “Rare Finds”. Others suggested changing the filtering scheme so that we can filter for specific values as opposed to color coding for different values on an applied filter. This was an excellent suggestion, especially since it reduces the number of listings users have to sift through. In a similar vein, people suggested better organization and layout for the statistics page to make it easier to understand what each graph was depicting. This mostly included calls for better axis labels, titles, and layout.

Challenges Faced + Solutions

- Originally, we set out to create a dashboard that summarized Airbnb listings trends and helped users find listings in multiple cities. However, our data source provided each city's data as an individual set of datasets. Combining and cleaning upwards of 15 datasets became too large of a task for our timeline. Thus, we decided to pivot and focus only on Seattle. However, in the future, our project could be expanded to provide a similar service to users looking for Airbnbs in multiple cities across the country and world.
- We wanted to create a dashboard where all sheets change on a common filter. However, we ran into a redundancy issue where some values were linked to the variable ID and some were linked to the variable Listing ID. ID and Listing ID referred to the same data but represented two separate columns and thus could not operate on a singular filter. To work around this, we created a relationship between ID and Listing ID across datasets. Once the values were linked, they could operate as a singular filter.
- The amenities filters were initially all over the place in terms of position on the dashboard, and consistency (sliders, drop-down, etc). We decided to incorporate the amenities into one drop-down list with the option to select multiple values. This streamlined the cluttered design and provided more organization to the filters.
- In the dashboard we presented to the class, color schemes across visualizations were heterogeneous which seemed to confuse the audience. We incorporated changes following the final presentation to operate all visualizations on a single color scheme for consistency.
- In creating the word cloud, we ran into issues with drilling it down to an ID level. The dataset structure wasn't conducive for a world cloud at first because reviews contained multiple words, while a word cloud in Tableau functions by counting individual words. We solved this issue by counting the number of times each word appeared in the reviews of a listing and plotted the top 50 words for each listing in the word cloud.
- Another comment we received was that the heatmaps displayed side by side were cluttered and hard to look at. We condensed it down to the 3 most interesting heatmaps and added a neighborhood filter option that simultaneously applies to each heatmap. This gives the user less to look at and guides them in how to interact with the data.

Insights

- We had hypothesized that the number of new hosts joining the AirBnB family would have dipped substantially during COVID-19. A time series chart verified the hypothesis.
- Average price is higher during weekends, holidays, and summer months which is when people prefer going on vacation most. During the middle of the week and colder months, prices tend to be low since the demand is low. This confirms our hunch that demand and price are directly proportional.
- Availability is lower over weekends and holiday seasons. Thus, it is inversely proportional to price.
- We wanted to see if the host being verified affects ratings. On plotting a bar graph relating verified status to average rating, we saw that there was no significant effect.
- 80% of room types are “entire rooms”. It outnumbers all other types of rooms by a large margin.
- The highest frequency of listings accommodates two people, as we concluded from a unimodal distribution that we plotted.
- Broadway, Belltown, Fremont, U-district, and Wallingford are the most popular neighborhoods for Airbnb bookings in Seattle.
- Entire rental unit, entire home, entire guest suite, entire condo, and private room in house are the most popular property types in Seattle.
- In every neighborhood, there is a large spread of prices per night. There is no significant trend in price by neighborhood.
- Most listings fall between an average review of 4 and 5 stars. This makes it difficult to compare listings by average review score, as there is very little variance.

Critical Evaluation

In this section, we evaluate our overall process from initial ideation to visual exploration of the data to the different prototypes we created to the final visualization product we built. To build our visualization we used a combination of principles, methods, and techniques we learned in class as well as from the different reading assignments specifically employing the ones highlighted below -

Expressiveness and Effectiveness

One of our primary goals for the project was to stay true to Mackinlay’s principle of expressiveness and effectiveness (Mackinlay, 1986). This helped us iterate over multiple ideas until we could build visualizations that effectively helped the user navigate the Airbnb data set.

Graphical Integrity

Tufte's principles of graphical integrity helped us evaluate our ideas. While initially we wanted to build a single dashboard view that helped users navigate through all the complexities of finding the right Airbnb listing, we couldn't find the right way to build a single dashboard that encompassed all the information without sacrificing clarity and hence sacrificing graphical integrity (Tufte 1973). Instead, we chose to prioritize clear, detailed, and thorough labeling to increase the clarity of our visualizations and decided to create multiple tabs to allow space for detailed legends and captions.

Schneiderman's Taxonomy

Schneiderman's taxonomy of overview first, zoom and filter, then details on demand guided the organization of our project (Shneiderman 1996). Our dashboard begins with a view of all listings in Seattle. Then, users can zoom and filter based on what features are relevant to them. Finally, they can pan over a listing to obtain additional information, and for even more information, they can click on a listing and be taken directly to the detailed Airbnb page. We also provide users the ability to obtain information on overarching general trends through heatmaps and citywide graphs of features and prices. Thus, we built a dashboard that not only helps the user understand general Airbnb listings in Seattle, but also filters to specific listings that meet their needs and obtain more information about those listings on demand.

5 design sheets

The 5 design sheet method (Roberts et al. 2016) helped us iterate over prototypes. From the early sheets, we were clear we wanted to build a map view to help customers navigate large amounts of listings data that were predominantly geographical in nature. This method helped us identify our strong points and focus on those for the final product.

Conclusion

Our team followed the visualization principles and encoding best practices taught in class to organize our data and build effective dashboards to help users understand Airbnb market trends for Seattle and find the best listing based on their individual preferences. We categorized our dashboard into four different tabs for effective grouping of visualizations based on purpose. From the beginning stages of design ideation, our design underwent many adjustments based on user and class feedback that guided our dashboard toward higher levels of effectiveness, expressiveness, and graphical integrity. The final product is an insightful story of Airbnb listings in Seattle that informs users on neighborhood and date trends citywide and provides more detailed information on each listing on demand.

References

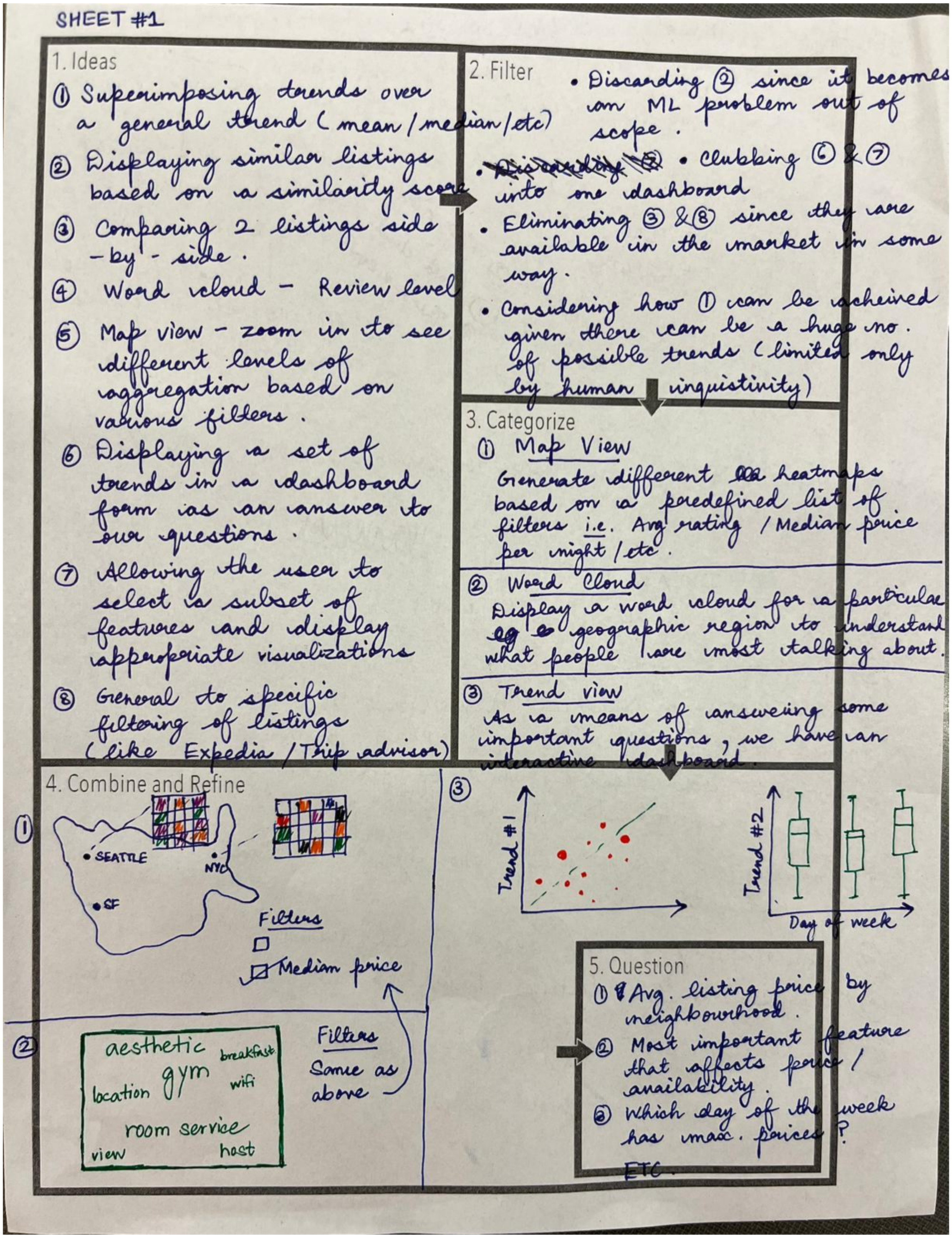
1. “Graphical Integrity”, Tufte E. R., 1973

2. J. Mackinlay, “Automating the design of graphical presentations of relational information,” *ACM Transactions on Graphics*, vol. 5, no. 2, pp. 110–141, 1986.

3. B. Shneiderman, “The eyes have it: A task by data type taxonomy for information visualizations,” *Proceedings 1996 IEEE Symposium on Visual Languages*.

4. J. C. Roberts, C. Headleand and P. D. Ritsos, "Sketching Designs Using the Five Design-Sheet Methodology," in IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 1, pp. 419-428, 31 Jan. 2016, doi: 10.1109/TVCG.2015.2467271.

Appendix A: Design Ideation



SHEET #1

1. Ideas

① Superimposing trends over a general trend (mean / median / etc)

② Displaying similar listings based on a similarity score

③ Comparing 2 listings side-by-side.

④ Word cloud – Review level

⑤ Map view – zoom in to see different levels of aggregation based on various filters.

⑥ Displaying a set of trends in a dashboard form as an answer to our questions.

⑦ Allowing the user to select a subset of features and display appropriate visualizations

⑧ General to specific filtering of listings (like Expedia / Trip advisor)

2. Filter

- Discarding ② since it becomes an ML problem out of scope.
- ~~Discarding ②~~
- Clubbing ⑥ & ⑦ into one dashboard
- Eliminating ③ & ⑧ since they are available in the market in some way.
- Considering how ① can be acheived given there can be a huge no. of possible trends (limited only by human inquistivity)

3. Categorize

① Map View
Generate different heatmaps based on a predefined list of filters i.e. Avg rating / Median price per night / etc.

② Word Cloud
Display a word cloud for a particular eg geographic region to understand what people are most talking about.

③ Trend view
As a means of answeing some important questions, we have an interactive dashboard.

4. Combine and Refine

5. Question

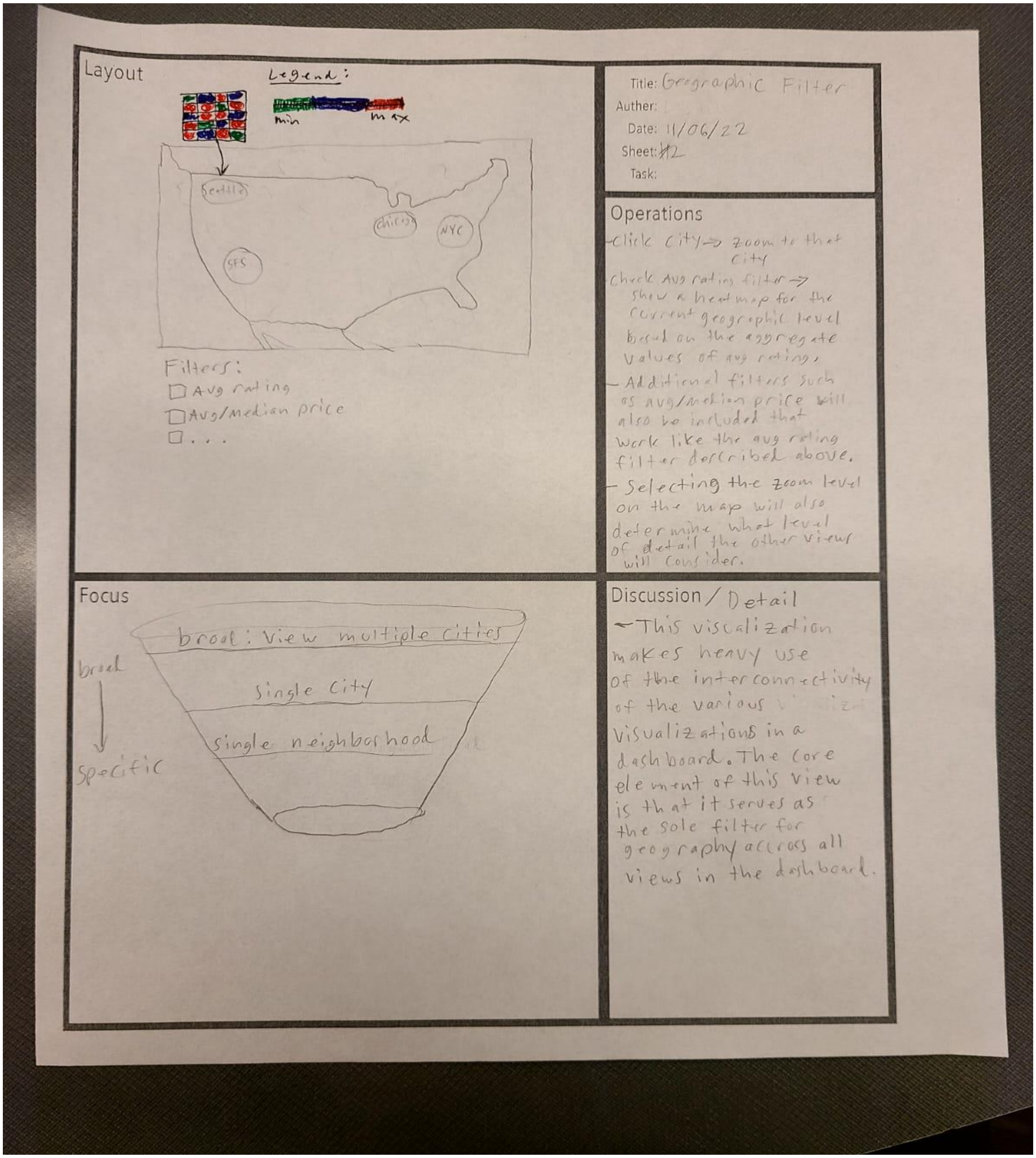
① Avg. listing price by neighbourhood.

② Most important feature that affects price / availability.

③ Which day of the week has max. prices?

ETC.

Figure 1: Sheet 1



Filters:

□ Avg rating

□ Avg/median price

□ . . .

Title: Geographic Filter

Auther:

Date: 11/06/22

Sheet: #2

Task:

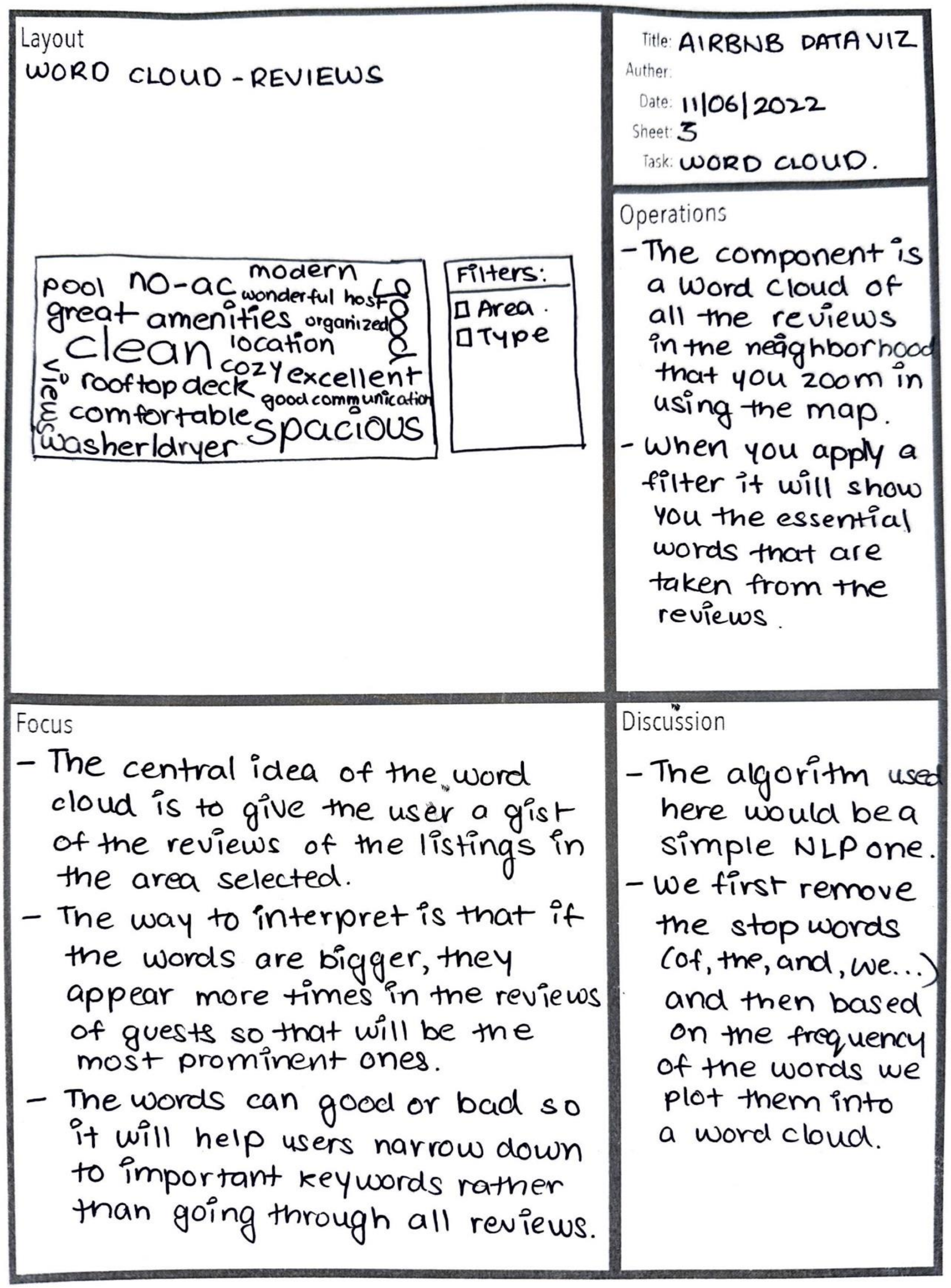
Operations

- click city → zoom to that city
- check Avg rating filter → show a heat map for the current geographic level based on the aggregate values of avg rating.
- Additional filters such as avg/median price will also be included that work like the avg rating filter described above.
- Selecting the zoom level on the map will also determine what level of detail the other views will consider.

Discussion / Detail

- This viscalization makes heavy use of the interconnectivity of the various visualizations in a dashboard. The core element of this view is that it serves as the sole filter for geography accross all views in the dashboard.

Figure 2: Sheet 2



Title: AIRBNB DATA VIZ

Auther:

Date: 11|06|2022

Sheet: 3

Task: WORD CLOUD.

Layout

WORD CLOUD - REVIEWS

pool
NO-AC
modern
wonderful host
great
amenities
organized
clean
location
cozy
excellent
rooftop deck
good communication
comfortable
washer/dryer
spacious
views
Good

Filters:

□ Area.

□ Type

Operations

- The component is a Word Cloud of all the reviews in the neighborhood that you zoom in using the map.
- When you apply a filter it will show you the essential words that are taken from the reviews.

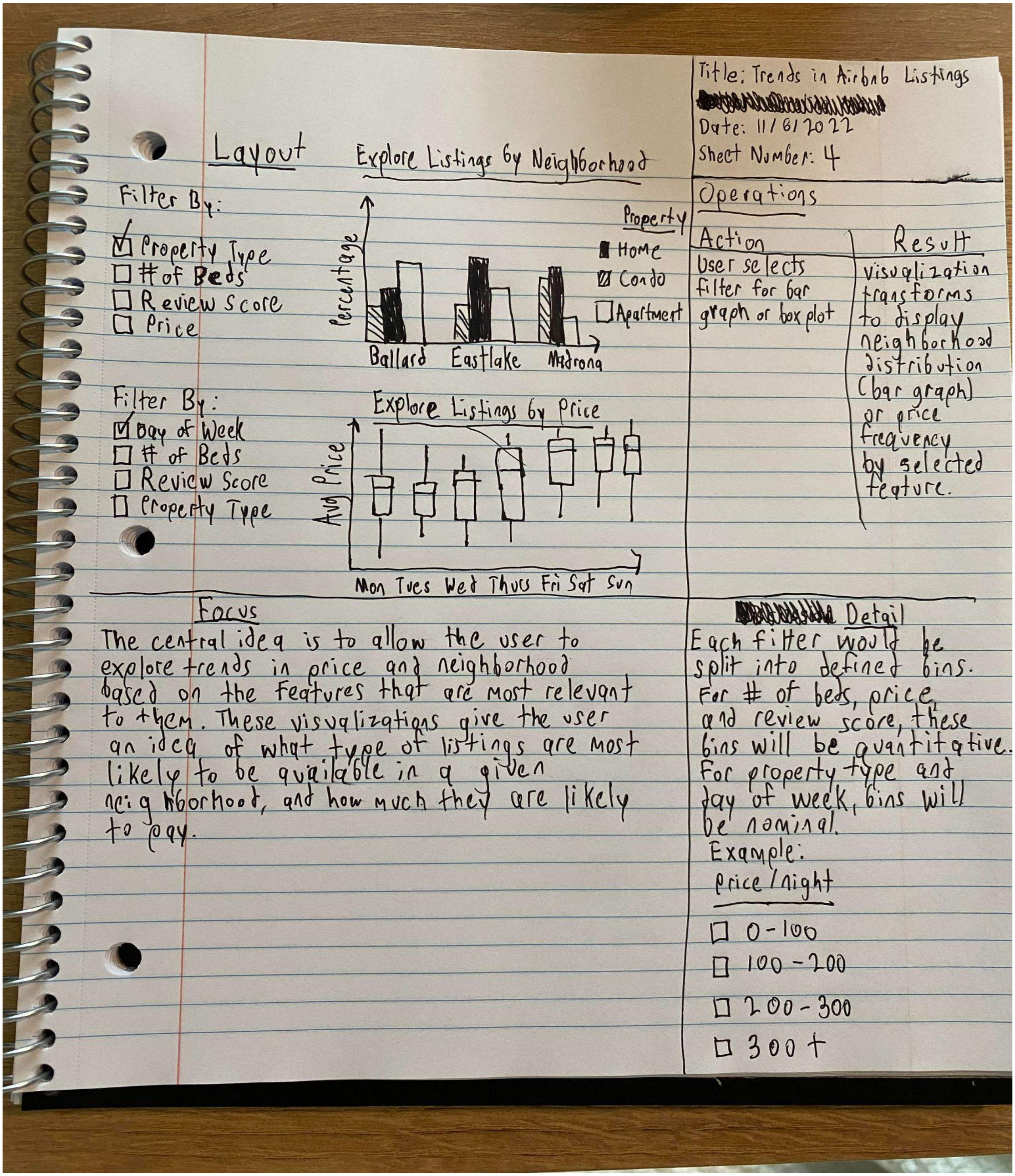
Focus

- The central idea of the word cloud is to give the user a gist of the reviews of the listings in the area selected.
- The way to interpret is that if the words are bigger, they appear more times in the reviews of guests so that will be the most prominent ones.
- The words can good or bad so it will help users narrow down to important keywords rather than going through all reviews.

Discussion

- The algoritm used here would be a simple NLP one.
- We first remove the stop words (of, the, and, we...) and then based on the frequency of the words we plot them into a word cloud.

Figure 3: Sheet 3



Title: Trends in Airbnb Listings

[redacted]

Date: 11/6/2022

Sheet Number: 4

Layout

Filter By:

☑ Property Type
□ # of Beds
□ Review Score
□ Price

Explore Listings by Neighborhood
Percentage
Property
■ Home
▨ Condo
□ Apartment
Ballard
Eastlake
Madrona

Filter By:

☑ Day of Week
□ # of Beds
□ Review Score
□ Property Type

Explore Listings by Price
Avg Price
Mon
Tues
Wed
Thurs
Fri
Sat
Sun

Operations

Action	Result
User selects filter for bar graph or box plot	visualization transforms to display neighborhood distribution (bar graph) or price frequency by selected feature.

Focus

The central idea is to allow the user to explore trends in price and neighborhood based on the features that are most relevant to them. These visualizations give the user an idea of what type of listings are most likely to be available in a given neighborhood, and how much they are likely to pay.

[redacted] Detail

Each filter would be split into defined bins. For # of beds, price, and review score, these bins will be quantitative. For property type and day of week, bins will be nominal.

Example:

Price/night

□ 0-100
□ 100-200
□ 200-300
□ 300+

Figure 4: Sheet 4

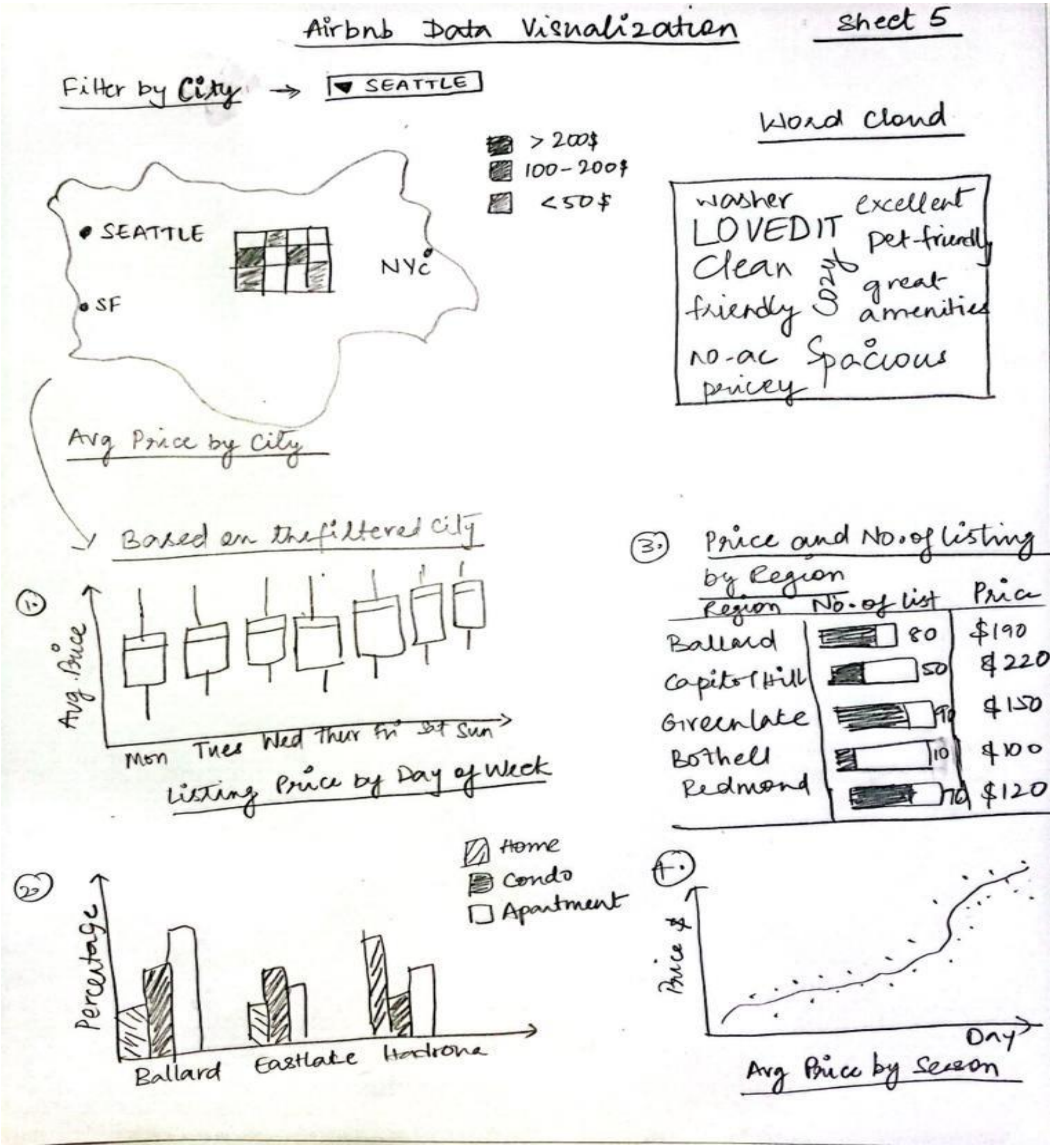


Figure 5: Sheet 5

Appendix B: User Tests

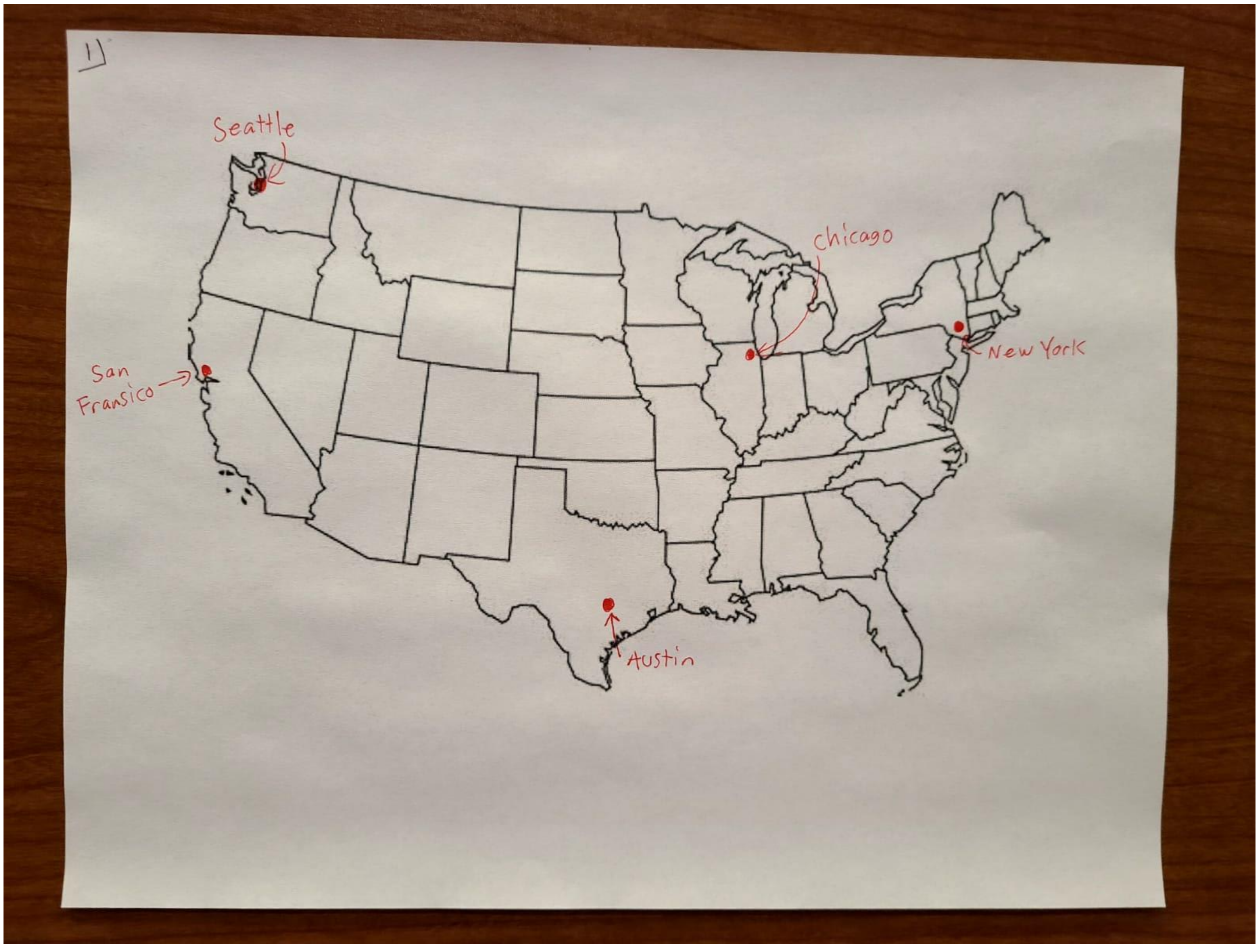


Figure 6: User Task 1 (Find listings in NYC)

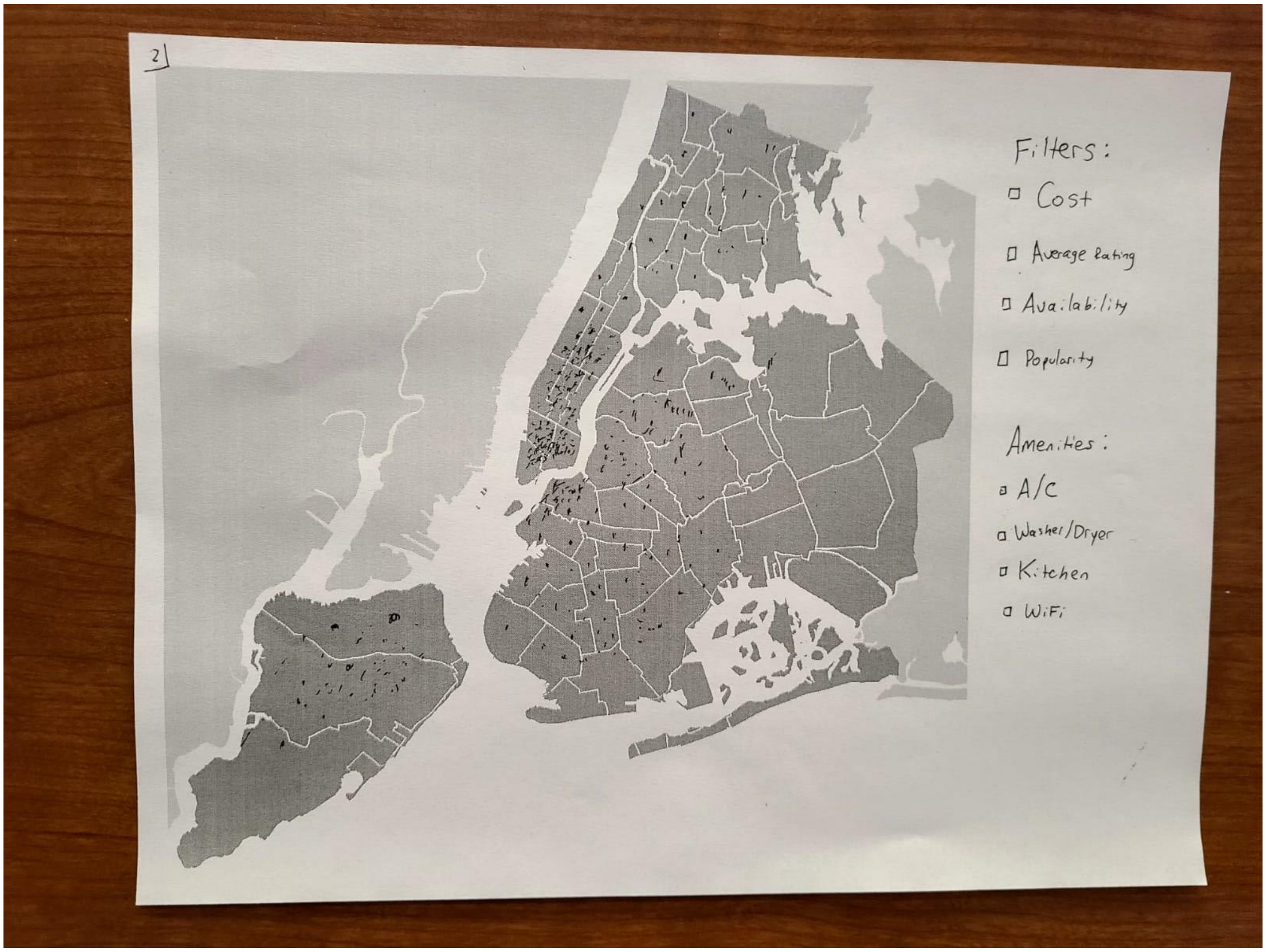


Figure 7: User Task 2 (Find ratings info)

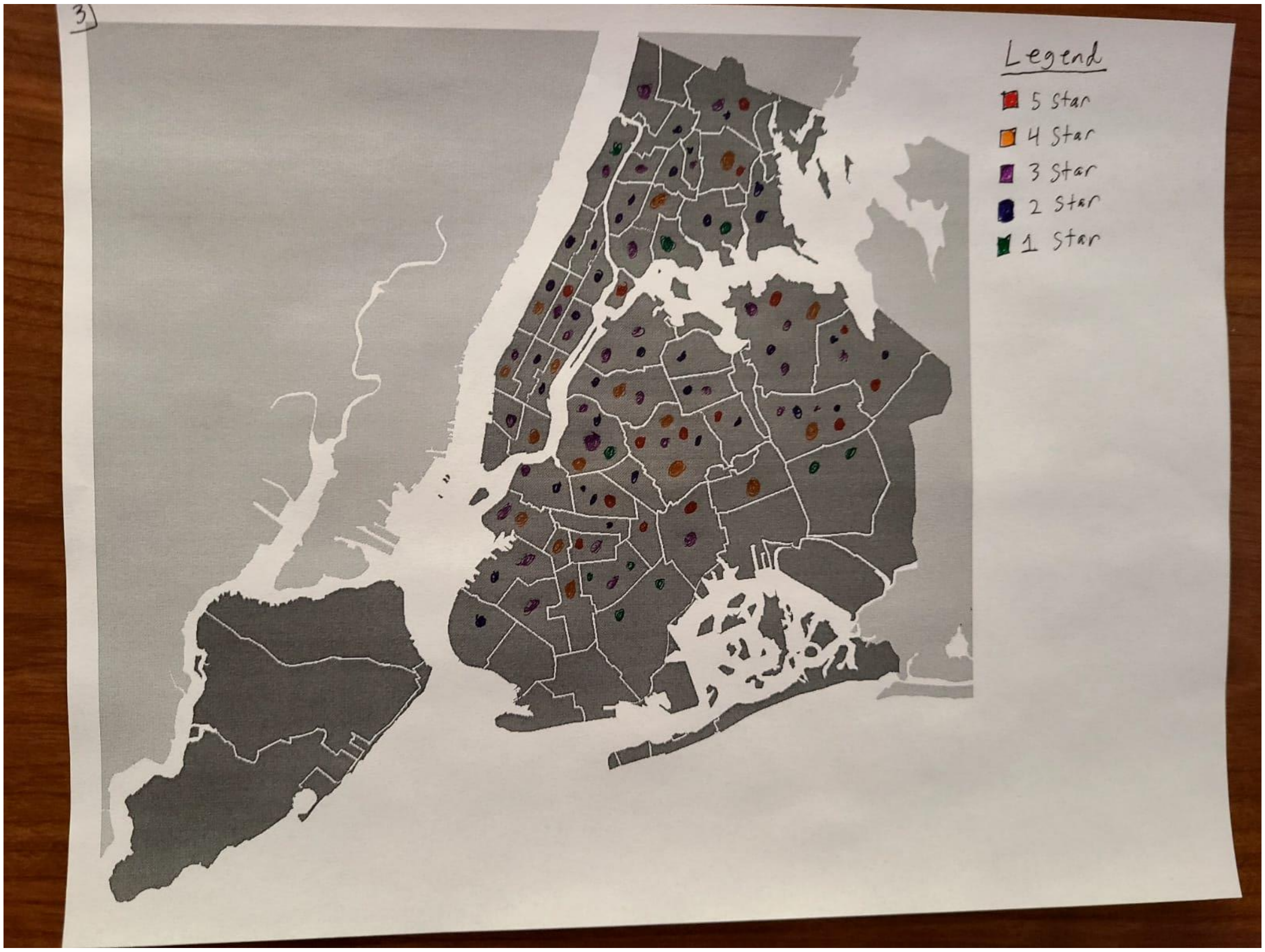


Figure 8: User Task 3 (Find price history for a 4-star listing)

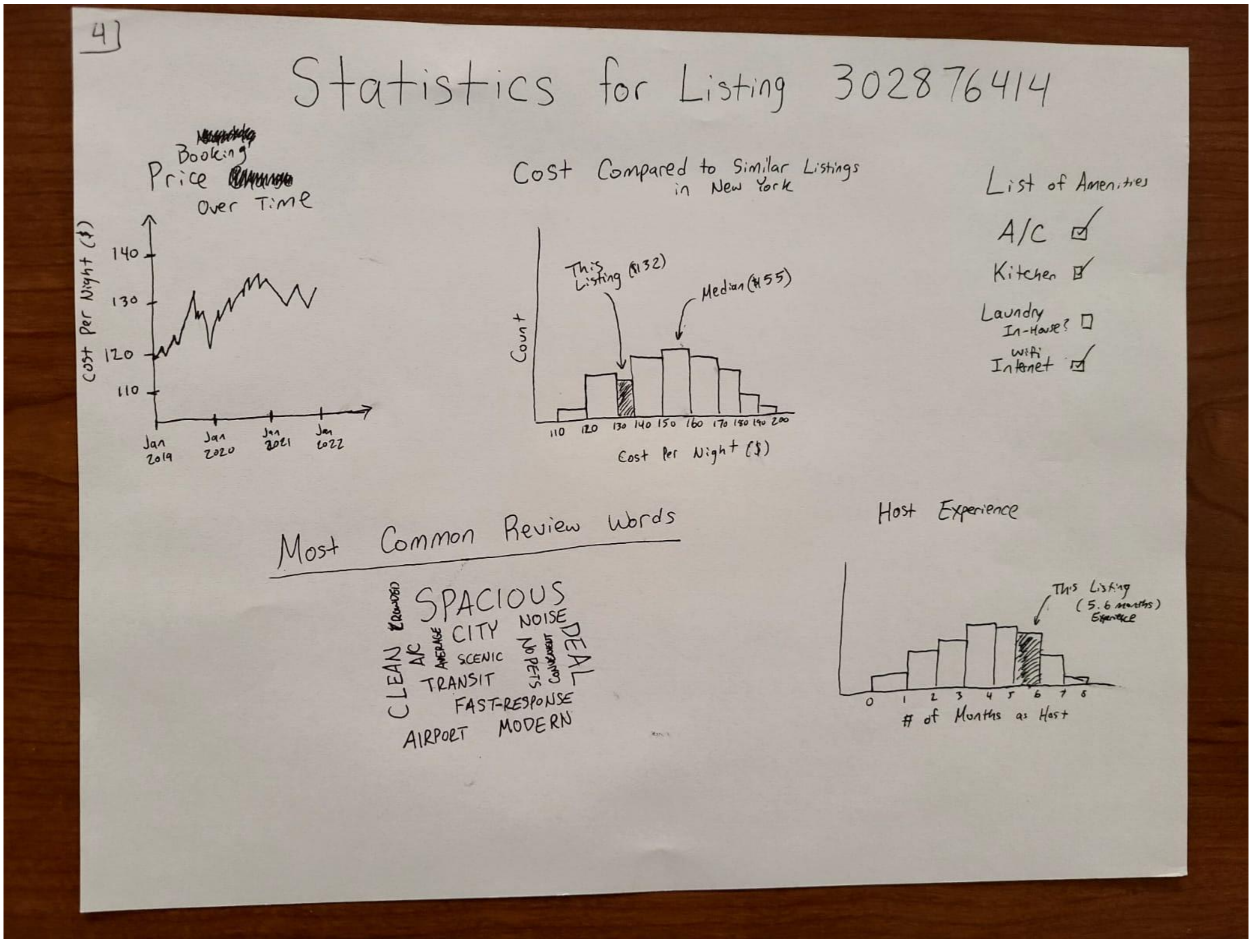


Figure 9: Prototype for Individual Listings Details Dashboard