DSCI 320: Milestone II Written Report

By Cassandra Zhang, Selena Shew, Jamie Jiang

Group Name: CSJ

December 2, 2023

Project Title: The Visualization of AirBnB Data For Travellers

Project Focus: We aim to identify the relationships between different data attributes and the price per day for AirBnB locations in Vancouver. More particularly, we want to learn which data attribute classes or values are associated with a more expensive daily rate. As well, we hope to learn which values and classes of different attributes lead to a higher or lower review rating for each AirBnB location.

Intended Audience: Young adults (from 18-30 years of age) with disposable income who are travelling for leisure or work.

Overall Cohesion

Tasks

Task 1: Exploring how the daily rate of the AirBnB varies with the number of beds, bathrooms, and people that can be accommodated.

Task 2: Exploring how the property type and room type of the AirBnB relates to the review ratings.

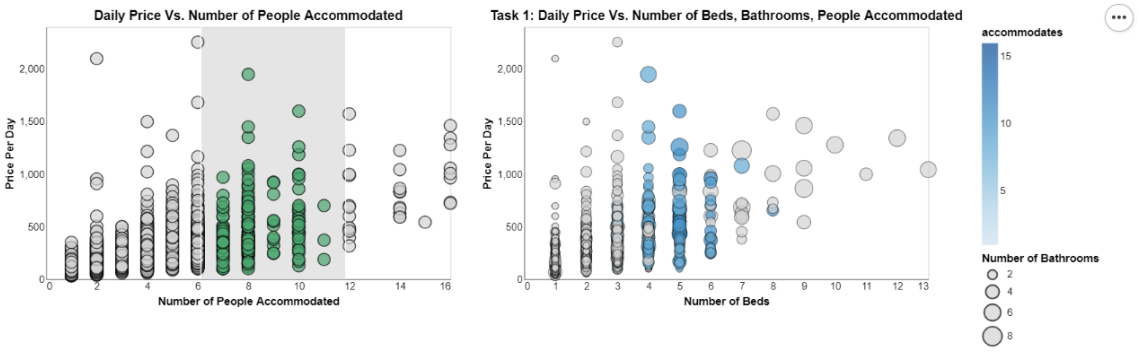
Task 3: Exploring how the AirBnB host's response time as well as whether they are a designated superhost or not affects their review ratings & number of reviews.

Task 4: Exploring how the daily rate of the AirBnB varies with the neighbourhood location and room type.

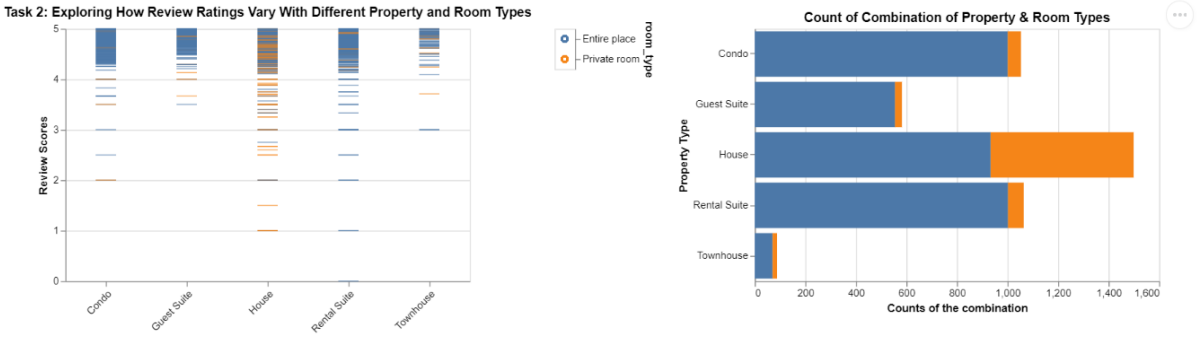
Dashboard Overview

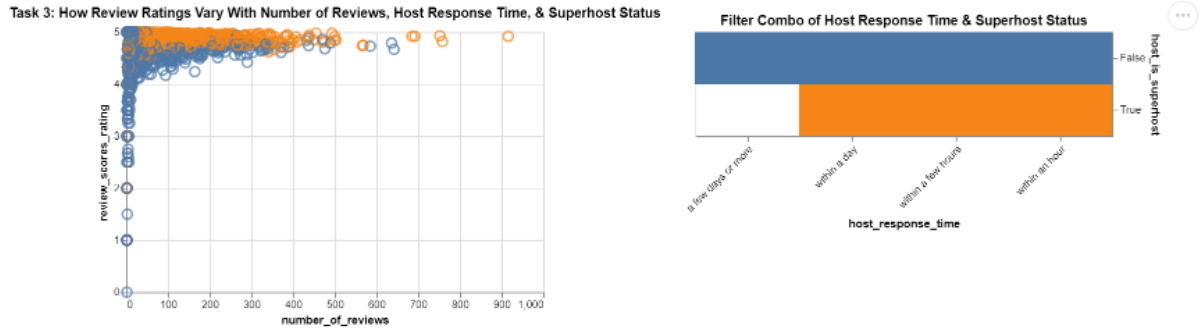
For this project, we aimed to visualize AirBnB data in order to get a better understanding of how various attributes relate to the daily price as well as review rating for each AirBnB listing. All five of our visualizations are directly tied to this overall theme as three of them (two in Altair, one novel) examine the relationships between price and attributes such as the the number of beds, bathrooms, neighbourhood, and room type, whereas the other two examine the relationships between review ratings and attributes such as the property type, room type, superhost status, and host response time to enquiries. From all of our visualizations, the audience can clearly tell which values or categories of attributes are more likely to lead to a higher daily price or higher review rating for the AirBnBs, which will aid them in their decision-making when they need to decide where to stay while on their travels. The final dashboard can be seen below.

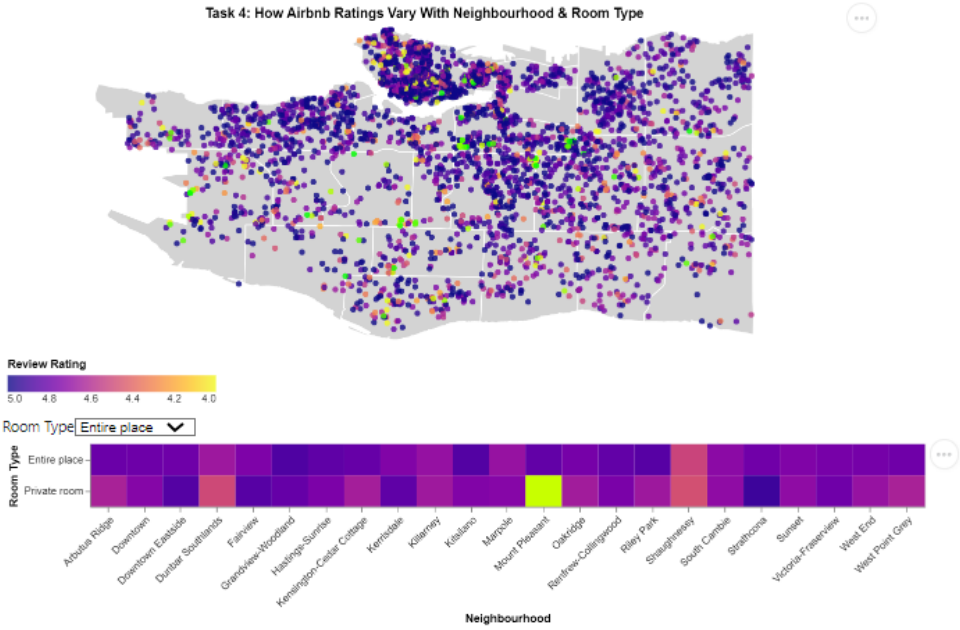
The visualizations that contain bidirectional linking are the two made for Task 1. The chart on the left displays the price per day against the number of people that can be accommodated per listing. Meanwhile, the chart on the right displays the price per day against the number of beds on the x-axis and the number of bathrooms in the size encoding. Both visualizations contain a selection interval. When the selection interval is activated for the right side chart, the left side chart displays the corresponding values of the number of people that can be accommodated for the selected listings. When the selection interval is activated for the left, the right side chart displays the corresponding values of the number of beds and bathrooms that allow for those numbers of people to be accommodated.


Daily Price Vs. Number of People Accommodated
Price Per Day
Number of People Accommodated
Task 1: Daily Price Vs. Number of Beds, Bathrooms, People Accommodated
Price Per Day
Number of Beds
accommodates
Number of Bathrooms

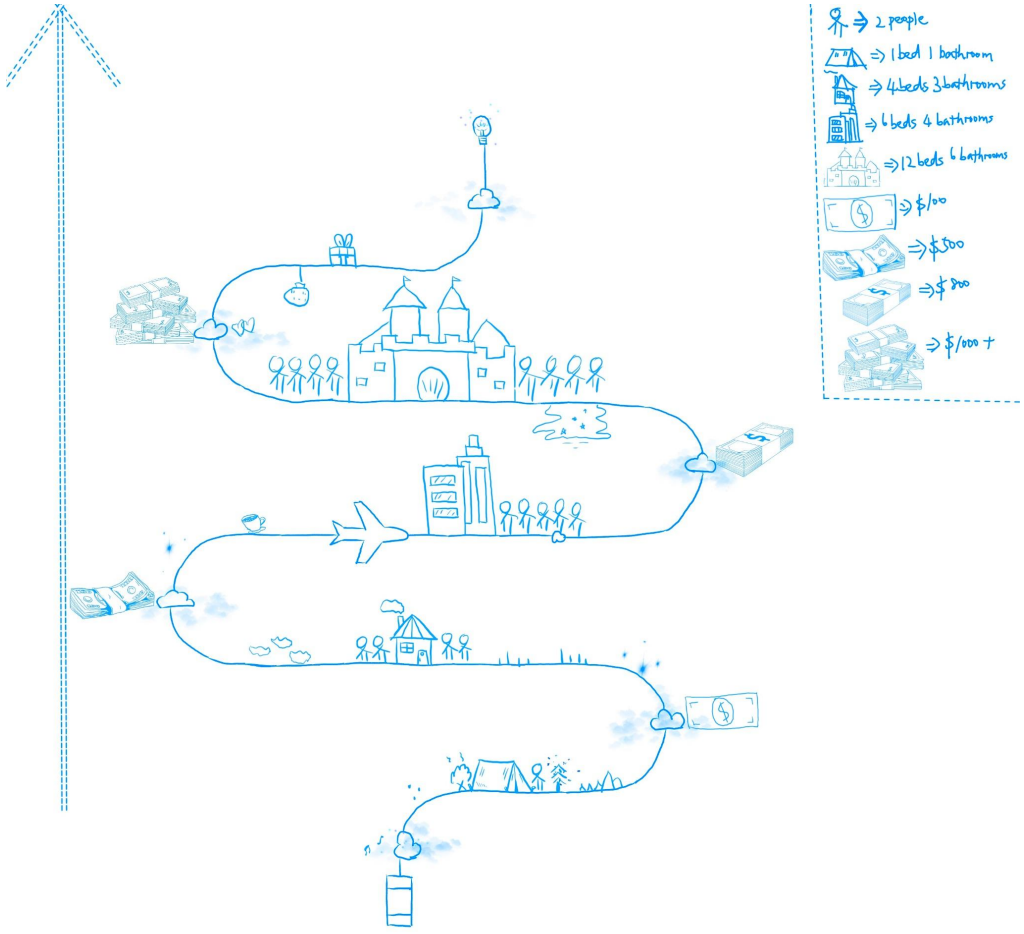
Opacity: 0.7


Task 2: Exploring How Review Ratings Vary With Different Property and Room Types
Review Scores
Entire place
Private room
room_type
Condo
Guest Suite
House
Rental Suite
Townhouse
Count of Combination of Property & Room Types
Property Type
Counts of the combination


Task 3: How Review Ratings Vary With Number of Reviews, Host Response Time, & Superhost Status
review_scores_rating
number_of_reviews
Filter Combo of Host Response Time & Superhost Status
host_is_superhost
False
True
a few days or more
within a day
within a few hours
within an hour
host_response_time


Task 4: How Airbnb Ratings Vary With Neighbourhood & Room Type
Review Rating
5.0
4.8
4.6
4.4
4.2
4.0
Room Type
Entire place
Private room
Abbotsford
Downtown
Downtown Eastside
Dunbar Southlands
Fairview
Grandview-Woodland
Hastings-Sunrise
Kensington-Cedar Cottage
Kerrisdale
Killarney
Kitsilano
Marpole
Mount Pleasant
Oakridge
Renfrew-Collingwood
Riley Park
Shaughnessy
South Cambie
Strathcona
Sunset
Victoria-Fraserview
West End
West Point Grey
Neighbourhood

Novel & Non-Altair Visualization

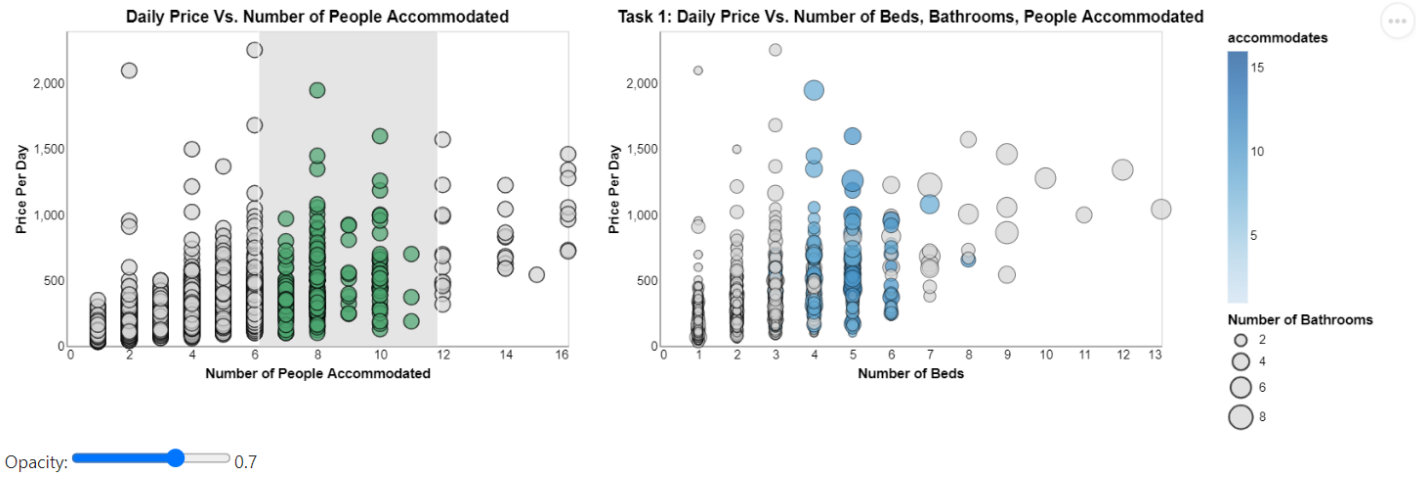


In this hand-drawn illustration, we can see a creative representation of the general patterns that we identified for Task 1: as the number of beds, bathrooms, and people that can be accommodated increases, so too does the daily price of the AirBnB. We included the visual of an airplane to drive home the importance of our tasks in aiding travellers on their adventures. Moreover, along the airplane's flight path, we drew increasing numbers of people as well as increasing sizes of the physical accommodations to represent the aforementioned relationship. The legend at the right-top shows the meaning of each level's small graphs, representing that increasing the number of people, beds, and bathrooms will result in an increase in price.

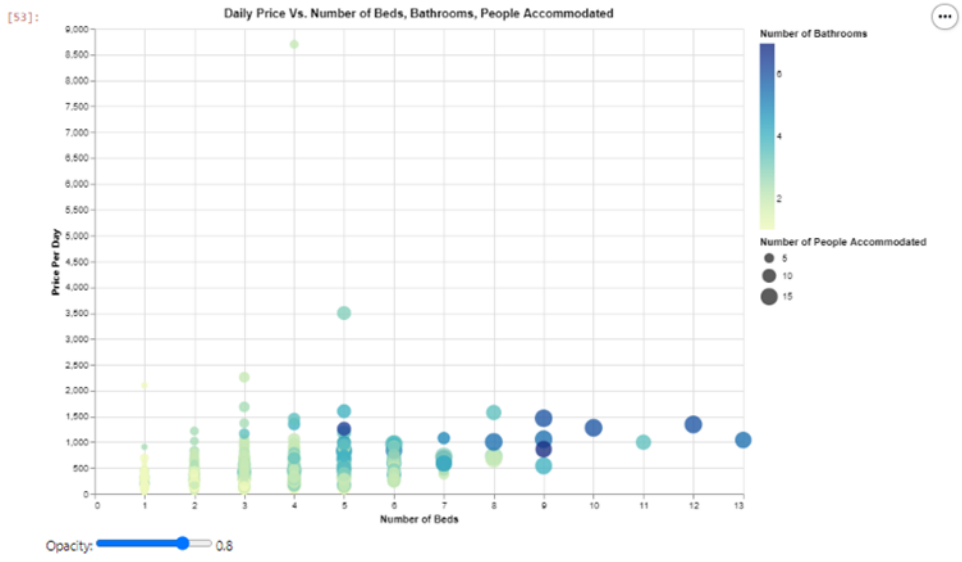
Justifications For Each Visualization

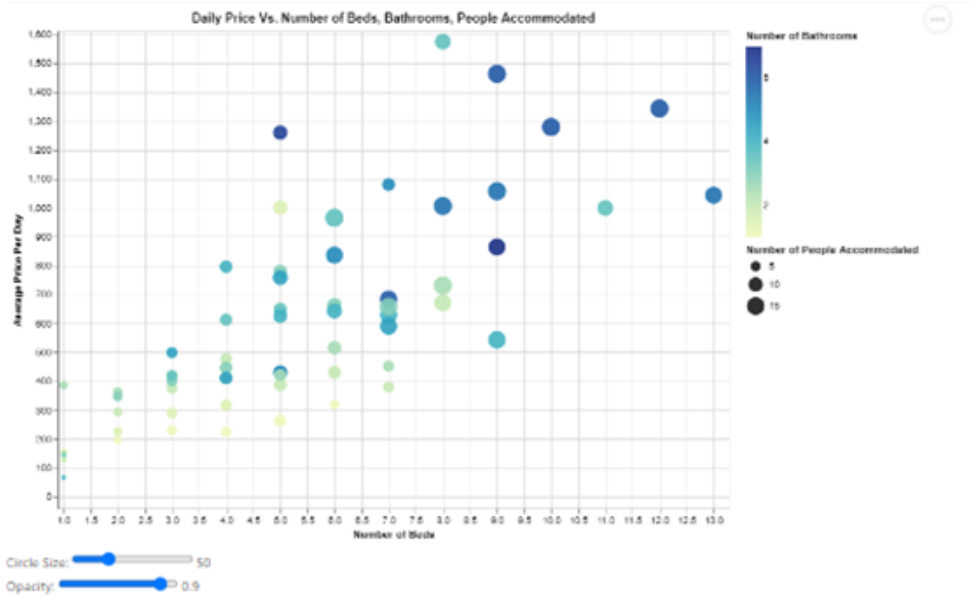
Task 1: Exploring how the daily rate of the AirBnB varies with the number of beds, bathrooms, and people that can be accommodated.

Final Screenshot:



Previous Iterations:





Explanation of Visualization

In the first task, the audience can clearly see that there appears to be a positive linear relationship between the price per day on the y-axis, the number of beds on the x-axis, and the number of bathrooms in the size encoding for the plot on the right. As the number of beds and the number of bathrooms increases, so too does the price. On the left hand side, we can also see that as the number of people that can be accommodated increases, the price generally appears to increase as well.

The marks that were used are circles, and the data semantics that they represent are the individual AirBnB listings in Vancouver. The channels that are used are the colour as well as the size/area. The colour channel on the right side chart showcases the number of people that can be accommodated, with the darker hues responding to more people and the lighter hues responding to less. On the other hand, the size encoding represents the number of bathrooms, with the larger area corresponding to a higher number of bathrooms for that listing.

Critique

There are several areas that the visualization is lacking and can be improved in. The first area is the visualization's expressiveness. The number of bathrooms is a quantitative data type, but was encoded in the size channel. Humans inherently are not well apt at determining differences in size or area, making it difficult for the user to accurately tell at first glance the number of bathrooms that each listing has. As well, the number of bathrooms had to be grouped up into classes of 2 (2, 4, 6, 8), which is not a

precise depiction of the number of bathrooms that the individual listing has. Unfortunately, this is due to the fact that there were computational limitations to the number of ways that Altair would allow for multiple semantics to be displayed on the same chart. Due to the need to try and make the charts smaller to fit on the same dashboard, the number of bathrooms had to be grouped together into classes in order to save space. Secondly, another area of concern is the interference between the size and colour channels. Some of the points are extremely small in area due to having a lower number of bathrooms, which then makes it difficult to discern the number of people that can be accommodated for that listing via the colour. Additionally, the colour that was intended to be used for the number of the people that can be accommodated is green in order to differentiate the information shown on the graph on the left from the graph on the right. However, this may also cause confusion as the colour scale used to display the number of people accommodated on the graph on the right is blue instead, meaning there are now two colours for the same variable. This was unintentionally done as Altair automatically put in the blue colour scale on the right after bidirectionally linking the two graphs together, and nothing I tried could remove it or overwrite the default colour scheme. Finally, with more time, the visualization for Task 1 could definitely be more innovative and eye-catching.

Explanation & Justification of Interactions

- **Interactions Present:**
    - Event type: hover, click, drag
    - Reaction type: selection, highlight
    - Views: Juxtapose
    - Interaction coupling: bidirectional
    - Views data: share data
    - Interactivity: immediate, on-demand, spread to both graphs

As mentioned previously, the advanced interaction encoded here is the bidirectional linking between the two charts for Task 1. This interaction is activated by a selection interval on either chart. When the selection interval is activated for the right side chart, the left side chart displays the corresponding values of the number of people that can be accommodated for the selected listings. When the selection interval is activated for the left, the right side chart then displays the corresponding values of the number of beds and bathrooms.

Furthermore, another interaction tool that was used was the tooltip. While this is not visible in the screenshot, when a user hovers their mouse over the live image on right, an informational window appears that displays the daily price, the number of beds, and the number of bathrooms per listing. Similarly, when the user hovers their mouse

over the live image on the left, an informational window pops up that displays the daily price and the number of people that can be accommodated for the listing. The use of the tooltip greatly aids the user in being able to see the exact values for each listing, as it is difficult to tell the exact values from just examining the x and y-axes. Lastly, the third interaction tool that was used was the range slider for the opacity of each mark. Since there are thousands of listings within the dataset even after the data cleaning step, it is extremely difficult to see each one as they overlap within the visualizations. Therefore, the opacity slider allows for the user to more easily discern data points that are obscured behind other points.

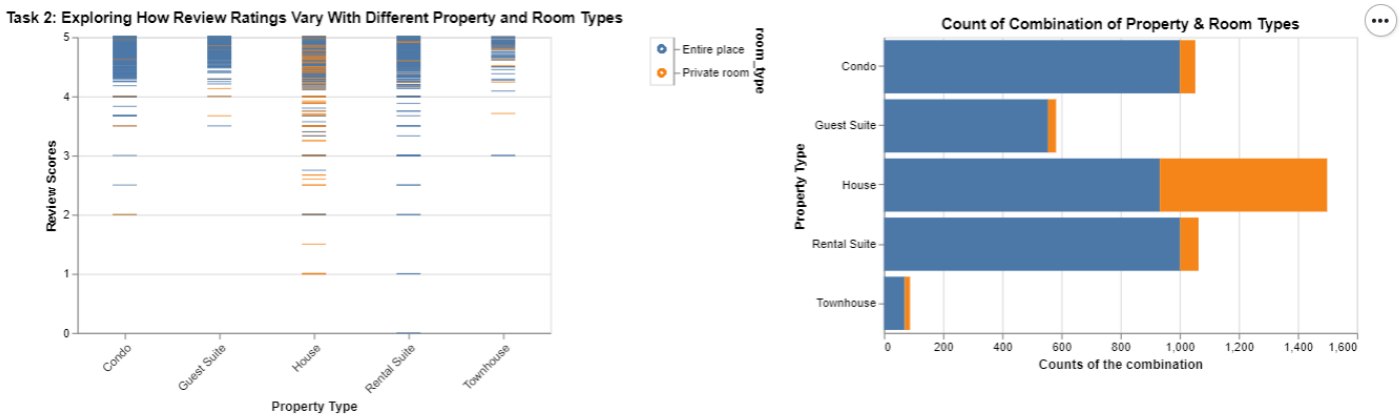
So why make this visualization interactive rather than static? To begin, we wanted to examine the relationship between many different variables all at the same time: price, number of beds, number of bathrooms, and the number of people that can be accommodated. As seen in one of the earlier iterations, there is an informational overload that occurs trying to examine the relationship between all four variables at the same time, which makes it extremely difficult for the viewer to see how price varies with each variable's values. By separating out this task into two charts and then linking them together, it then became much easier for users to examine subsets of the data at a time and makes the answers to these questions possible that would not be for a static image:

- When the number of beds and bathrooms are low, what happens to the price and the number of people that can be accommodated?
- If we now select listings with a higher number of beds and bathrooms, how does the number of people that can be accommodated and the price change?

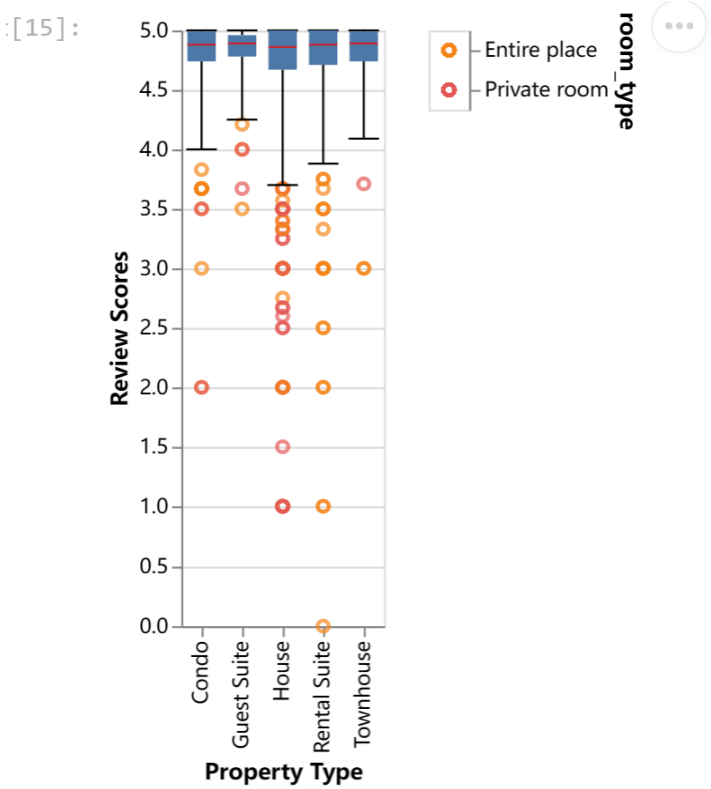
The answers to these questions become much clearer and more apparent by providing the interactive element between the two charts for Task 1.

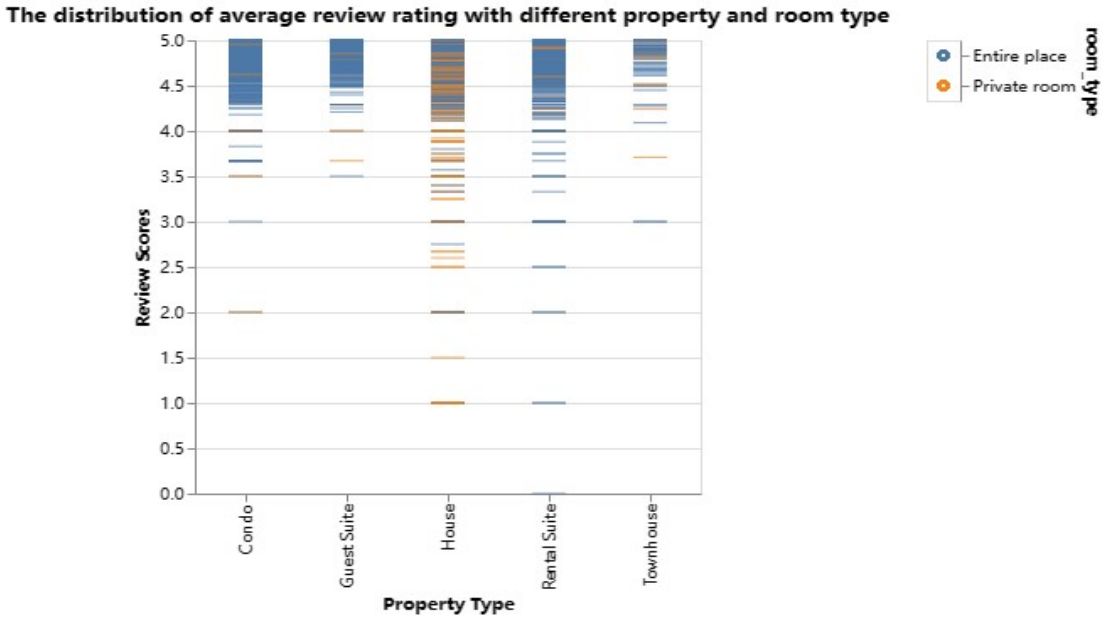
Task 2: Exploring how the property type and room type of the AirBnB relates to the review ratings.

Final Screenshot:



Previous Iterations:





Explanation of Visualization

Our second task is to explore the relationship between average review ratings and various combinations of property types and room types. Given the involvement of two nominal features and one quantitative feature, our preference for visualization is a tick chart over a scatter plot. Ticks not only reveal the density in a certain interval/area better than points/circles but also can convey the relationship between a quantitative feature and a categorical feature more effectively. This preference arises from the fact that scatter plots are more adept at presenting trends between two quantitative features. In addition, we employ a stacked bar chart to illustrate the counts of diverse combinations of two nominal features. As underscored in this course, when the task is interested in the frequency of multiple categorical variables, a stacked bar chart emerges as the most suitable option. Notably, bar charts prove advantageous for univariate visualizations, while stacked bar charts are recommended for bivariate visualizations.

Here, the left-side plot takes the form of a tick chart, designed to provide a comprehensive overview of the relationship between average review ratings, property types, and room types. Complementing this, on the right, a stacked bar graph intends to demonstrate the distribution of diverse combinations of property types and room types within a specified review ratings interval.

The tick chart employs the x-axis to denote five property types and the y-axis for average review ratings, with ticks as marks to represent each Airbnb's property type alongside its corresponding average review rating. Meanwhile, in the stacked bar graph, the y-axis indicates the categories of Airbnb's property types, and the x-axis illustrates the frequency of each property type within a defined review rating range. Both visualizations leverage the color channel to signify the room type of each Airbnb, either "entire place" or "private room." Given that the third requisite feature for this task is a categorical variable, the color channel proves optimal in terms of expressiveness and effectiveness for channel selection. Specifically, the utilization of color channels can effectively communicate the room type associated with each tick or stacked bar, as colors can instantly capture the reader's attention. Moreover, with two distinctive room types, the color channel is capable of conveying all the information in the dataset attributes, and the chosen color palette ensures accessibility for individuals with color blindness.

From the finished visualization, we can see that the property types of house and rental suite have a large variation in their review scores, ranging from 1 to 5 stars for both no matter the room type (whether it's the entire place or a private room). On the other hand, the property types of guest suites and townhouses have the lowest amount of variation in their review scores no matter the room type, only ranging from at worst 3 stars to 5 stars. When the variable of room type is specifically examined, we can see that in general, private rooms seem to have lower ratings compared to when the entire place is booked out across all property types. On a different note, across all of the AirBnB listings examined in the dataset, a majority of them appear to be the room type of 'Entire Place,' with only the property type of houses allowing for more private rooms to be booked.

Critique

Nevertheless, one potential limitation of the plot is that it doesn't facilitate a direct comparison among room types within a selected rating interval. For example, if readers want to study the distribution of room types instead of the property types or the combinations of the two in a fixed rating scale, they must mentally aggregate the segments of stacked bars in the same color to ascertain the counts of each room type. As a result, calculating the count of each room mentally, alongside memorizing and comparing the counts of two room types within the chosen review rating interval simultaneously, becomes a challenging task.

Explanation & Justification of Interactions

- **Interactions Present:**
    - Action: select, filter, aggregate

- Event type: hover, click, drag
- Reaction type: highlight
- Views: Juxtapose
- Interaction coupling: unidirectional
- Views data: share data
- Interactivity – Action Elements: Focus for legend, Presence for tooltip, rectangular brush
- Interactivity - Reaction Elements: Spread for rectangular brush, Activation for selection

As mentioned above, we use a tick chart and a stacked bar chart to delve into the correlation between average review ratings and Airbnb's property and room types.There are a few interactions incorporated in the graphs. In the tick chart, a small point chart acts as an interactive legend to highlight the ticks on the tick chart. Clicking on the legend's points initiates a selection for the corresponding lines on the tick plot, specifying the room type with its distinctive color. This interactive legend enables users to discern the distribution of average review ratings across various property types with a chosen room type more easily than the static version of the tick chart. Furthermore, it can deliver an immediate response to users' actions, which consequently, enhances engagement throughout the analysis process. Among drop-down menus, radio buttons, and sliders, we opt for the legend because it stands out more effectively, utilizing color and allowing for a larger presentation size.

In addition, a selection interval is added to the tick chart, empowering users to define a selection region for interaction with the stacked bar chart on the right. This allows users to delve into how the counts evolve in the stacked bar chart based on the size and position of the selection within the tick chart. In this scenario, the interaction unveils the fluctuations in counts for combinations of property and room types within the stacked bar chart, aligned with the chosen review rating interval in the tick chart. The preference for the selection interval (and brush) in this visualization stems from its support for exploratory data analysis, enabling users to visually compare various combined categories within a specific review ratings interval.

What's more, we incorporate tooltips to annotate the room type, property type, and counts of each stacked bar. This addition aids in offering a detailed breakdown of each segment within the stacked bars, thereby improving the clarity of the visualizations.

In conclusion, through these interactions, audiences can gain valuable insights not only into the variations in average review ratings across different property types for

each of the two room types but also into the changes in the distribution of property and room types within various selected rating intervals.

- Questions the interaction makes possible that a static version wouldn't have:
    - For the tick chart:
        - Which property type and room type dominate in the review scores interval of 4.5 - 5.0?
        - In which review score intervals do houses combined with two room types predominate in the charts?"
        - Which room type is more popular: private room or entire space?

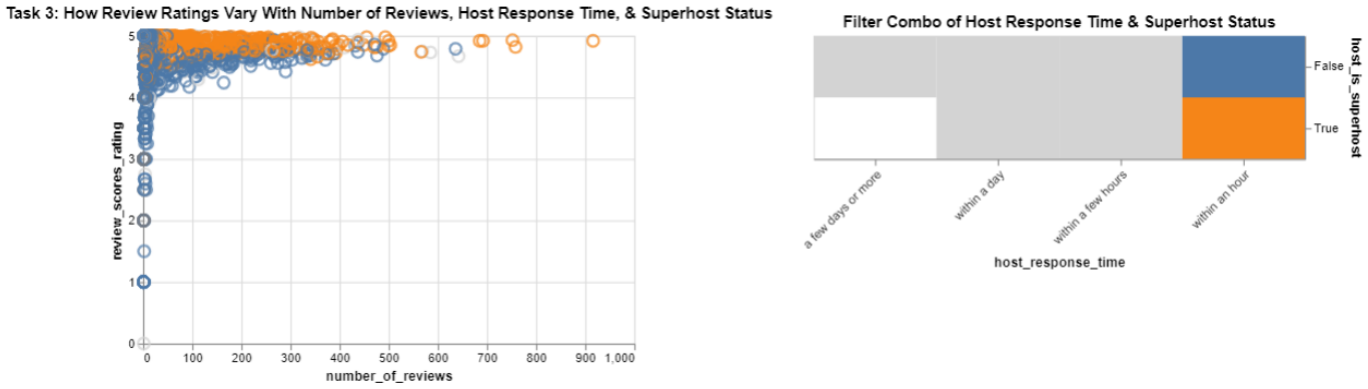
The interaction in the tick chart allows us to filter out the data points we need for different room types. In other words, we can discern the distribution of average review ratings across various property types with a chosen room type more easily than with the static version of the tick chart.

    - For the stacked bar chart:
        - What's the distribution of the property type in the review scores interval of below 3?
        - Which property type and room type contribute the lowest review scores in this data set?
        - How does the "condo" and the "private room" evolve across different rating intervals?

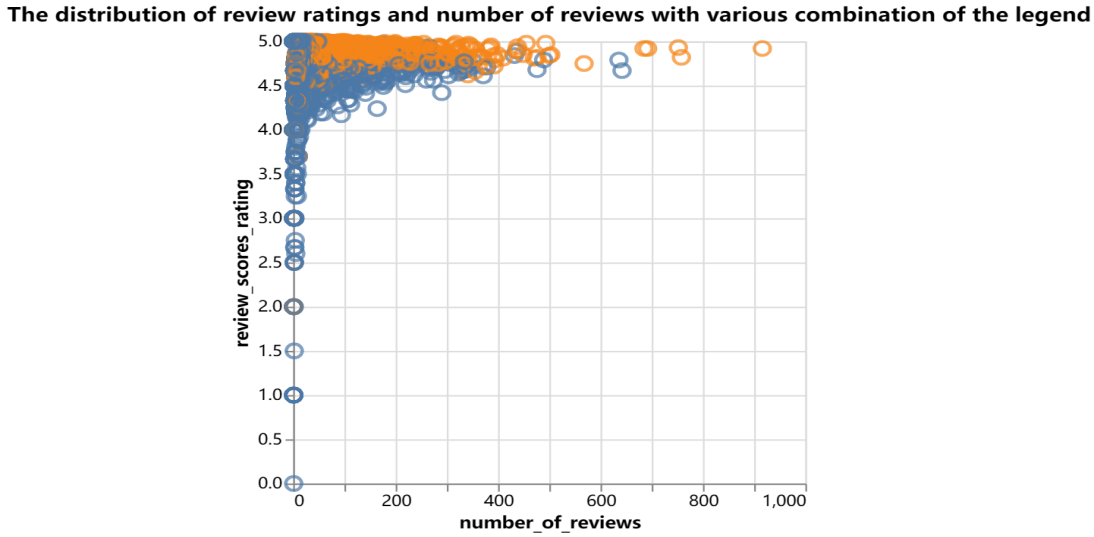
The interactive version enables the readers to observe the variation in the distribution of property and room types within different selected rating intervals. On the other hand, the static version cannot provide the changes in distribution over various review scores.

Task 3: Exploring how the AirBnB host's response time as well as whether they are a designated superhost or not affects their review ratings & number of reviews.

Final Screenshot:



Previous Iterations:



Explanation of Visualization

For the second task, we intend to study how an AirBnb host's response time and designated superhost status influence their average review ratings and number of reviews. To achieve this, we employ a scatterplot to depict the relationship between two

quantitative features: average review ratings and review counts. Because scatterplots can effectively capture trends in bivariate numerical features and fulfill the criteria for chart and mark selection: expressiveness and effectiveness. Here, the x-axis represents the number of reviews for each Airbnb, while the y-axis indicates the average review ratings. In addition, points on the plot illustrate each Airbnb's review counts and corresponding average review rating.

Beyond the scatterplot, we utilize a heatmap-formatted legend to manage combinations of two nominal features: the host's response time and their super host status. This heatmap legend effectively organizes these combinations in a matrix, with the x-axis introducing four categories of host response time and the y-axis denoting whether the host is a designated super.

Furthermore, the color channel is applied in both charts to distinguish categories of the binomial feature: the status of designated super hosts. This allows both charts to clearly highlight the points' categories. More importantly, the use of only two colors simplifies information processing for the human eye, facilitating the conversion of data into easily comprehensible messages during visualization.

Analyzing the visualizations reveals the significance of being a super host in attaining a high review score, evident in the majority of orange points dominating the upper section of the scatterplot. Additionally, it's noteworthy that a substantial number of super hosts respond to customers' messages within an hour, with very few extending their response time beyond one day. This pattern may serve as a potential criterion for achieving superhost status. (Learned from the graphs)

Critique

However, a potential drawback arises from the challenge when making direct comparisons of review ratings and numbers of reviews across four types of host response times while holding the super host status constant. To maintain the clarity of the chart, we opt for the color channel to encode the binary feature (super host status of each Airbnb host) and fix the category of host response time on the y-axis. In other words, if readers are interested in the distribution of points on the scatterplot with respect to various host response times (and ignore whether the hosts are super hosts or not), they need to memorize the general pattern of previous categories and compare them mentally. Consequently, this limitation may add complexity for the readers when assessing the importance of response time for Airbnb hosts.

Explanation & Justification of Interactions

- **Interactions Present:**

    - Action: select, filter
    - Event type: hover, click
    - Reaction type: highlight
    - Views: Juxtapose
    - Interaction coupling: unidirectional
    - Views data: share data
    - Interactivity – Action Elements: Focus for legend, Presence for tooltip
    - Interactivity - Reaction Elements: Activation for selection(legend)

In this analysis, we employ a scatterplot and a heatmap to explore how an Airbnb host's response time, as well as whether they are designated super hosts or not, affects their average review ratings and number of reviews. These two plots integrate several interactive features. First, as explained in the previous section, the heatmap functions as an interactive legend, allowing users to activate a selection for all points based on various combinations of host response time and super host status. The color channels assist audiences in observing the patterns of review scores and review counts while altering combinations of these two nominal features. It's essential to note that the selection is based on four types of host response time. In other words, users can only select columns and cannot choose rows or individual squares in the heatmap/legend.

Additionally, tooltips have been added to both plots to enhance convenience for readers, allowing them to focus on the chart without regularly referring to the axis and legend. What's more, the tooltip in the heatmap/legend includes the frequency of the combination of nominal features—host response time and whether or not the host is a designated super host. Therefore, tooltips facilitate readers in obtaining information about each point/square more easily and quickly.

In conclusion, these interactions empower users to explore the impact of host response time and the status of designated super hosts on the average review rating and the number of reviews for Airbnbs.

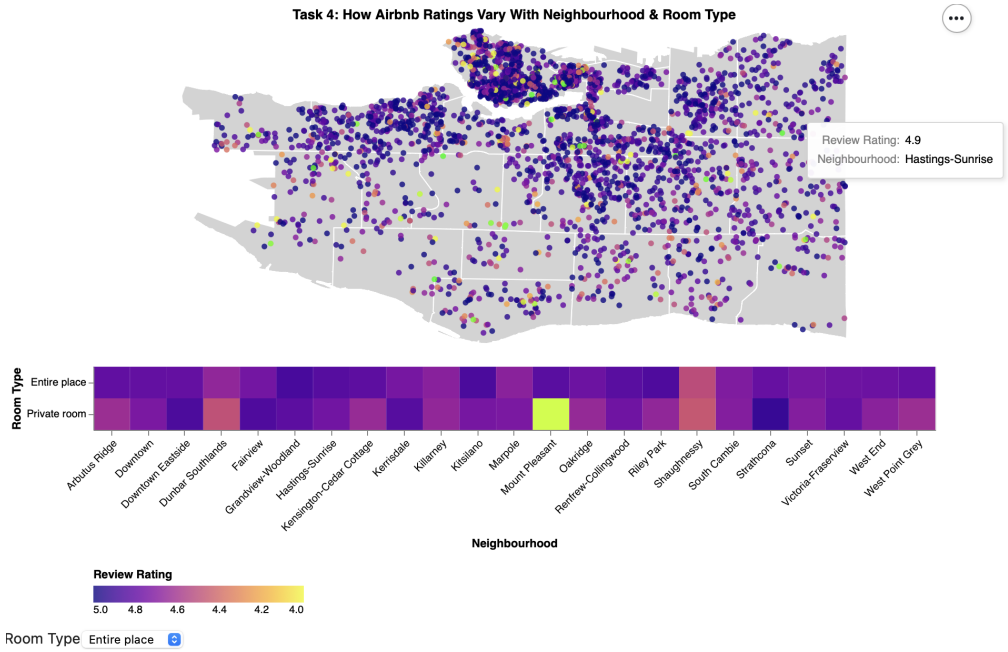
- Questions the interaction makes possible that a static version wouldn't have:
    - For the interaction between the legend and the scatter plot:
        - Which combinations of host response time and the status of designated super hosts occupy the largest portion with a score higher than 4.7?
        - If the host is not a superhost and responds to messages within a day, what average review score will he/she have?
        - What is the impact of the host's response time on Airbnb's average review rating and review counts?

- Which combinations of host response time and the status of designated super hosts have the highest counts in this dataset? (can achieve an answer from the tooltip in heatmap/legend)

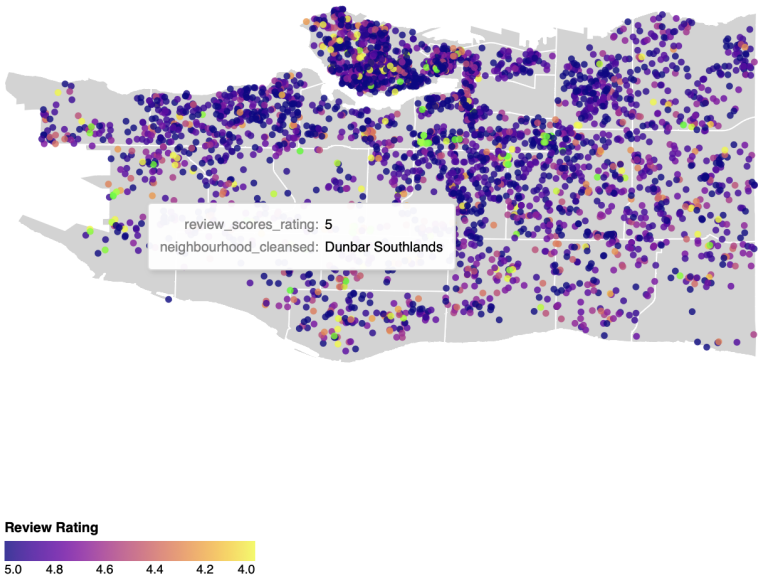
In summary, the interactive version enables the readers to delve into how host response time and the designation as a super host influence the average review rating and the number of reviews associated with each Airbnb.

Task 4: Exploring how the daily rate of the AirBnB varies with the neighbourhood location and room type.

Final Screenshot:



Previous Iterations:



Explanation of Visualization:

From the task, we can see that there is no significant variation in review rating according to Airbnb's neighbourhood or the room type, as most of the Airbnbs have a rating between 4 to 5 (so the green points are Airbnbs having a rating less than 4). Looking at the heat map, the combination of the Mount Pleasant neighbourhood with private rooms has the lowest average rating of 3.73. We can see that most of the AirBnB's are centered around Downtown Vancouver and the West Side of Vancouver. Private room AirBnBs in Strathcona have the highest average rating of 4.91.

The map visualization uses the longitude and latitude as the x-and y-channels, which then help to create a map of the Vancouver area. The color channel represents the review rating for this Airbnb on a scale from 4 to 5 since most of the ratings fall within this range. Darker hues represent higher ratings. The point/circle mark is used to represent each of the Airbnb locations. The data semantics that they represent are the longitude and latitude of Airbnb in Vancouver.

The heat-map visualization uses the neighbourhood of Airbnb as the x-channel and the room type of Airbnb as the y-channel. The colour channel represents the average review rating for each Airbnb room type in the neighbourhood. The rectangular mark is used to represent the room type in the neighbourhood. The data semantics represent the Airbnb room type in Vancouver neighbourhoods.

The colour scheme used in both visualizations is the same.

The color hue used in the visualizations may lead to a problematic understanding of the differences between each rating since the color scheme provides different colors according to the rating and we have many different ratings.

Critique

The map visualization uses points to represent each airbnb in Vancouver but it is hard to see a pattern of changes in review rating between the room type and neighbourhood. Using only the map visualization is hard to answer the task question directly since it shows a general overview of the airbnb reviewing rating changes in Vancouver and the location-spreads as well. Also, since we are using the color for review rating changes, it is hard to clearly see the differences between each point especially when the points have close ratings, i.e., the distinction of different review ratings in color is hard to interpret. It would be better if we select around 4 distinct colors instead of a sequential color scheme, and use one color for a specific review rating range i.e red for 4.8-5.0, etc.

Explanation & Justification of Interactions

- **Interactions Present:**
    - Event type: hover, click
    - Reaction type: selection, filter
    - Views: Juxtapose
    - Interaction coupling: N/A (map & heat map not linked)
    - Views data: share data
    - Interactivity: immediate, on-demand

Both the map and heat-map visualizations have mouse-based interaction of tooltip. The map visualization has the review rating and the neighbourhood of the airbnb, and the heat-map visualization has the average review rating.

The map visualization uses a dropdown filter to filter the room type between the entire place and private room. The marks on the map will then change accordingly to show the selected room type of airbnb on the map.

The interaction makes it easier for the audience to see the specific rating for each of the Airbnb in the map and the average rating for a specific room type in a specific neighbourhood. Hiding the actual rating to the tooltip is a way to emphasize the rating changes through color instead of looking through a great amount of numbers in the chart. The map interaction of allowing the audience to select different room types and changing the corresponding mark points on the graph makes the visualization playable and clearer to see the positions of different Airbnb. A heat map at the bottom

part provides a summary of review rating for each Airbnb room type in different neighbourhoods. The tooltip also helps the audience to see the actual number.

- What questions the interaction makes possible that a static version wouldn't have:
    - For the map visualization:
        - What are some of the neighbourhoods that have more of one room type than the other? Which room types are they?
        - What is the specific review rating for each of the airbnb points?
        - What are the specific names of the neighbourhoods?
    - For the heat-map visualization:
        - What is the specific average review rating of the selected room type in the selected neighbourhood?
        - Can we compare two room types in the same neighbourhood and find out if there are any differences in their ratings on average?

Group Reflection

3 Strengths

- Strong communication between all three group members
    - We immediately established a Discord server to communicate
    - We all respond quickly to each other's messages (within 3 hours maximum)
- Strong project organization
    - We established a Google Drive folder to contain all of our relevant project files & reports
    - We split up tasks evenly and everyone finished their sections within the self-imposed deadlines
- We chose interesting & relevant tasks to explore within the dataset that are highly relevant to the overall theme

2 Weaknesses

- We encountered many technical issues with attempting to code our visualizations (especially with the advanced interactions)
- We ran out of time to really polish up our visualizations

2 Things to Do Differently

- In the future we definitely will start much earlier on our projects to give ourselves more time
- We will ask for help from the teaching team sooner when we run into bugs & technical difficulties so that we don't waste as much time spending hours trying to debug

Work Distribution

High Level Overview				
Project Milestone	**Selena Shew**	**Cassandra Zhang**	**Jamie Jiang**	Total
Milestone 1	~33%	~33%	~33%	100%
Milestone 2	~33%	~33%	~33%	100%
Breakdown				
Milestone Description	**Selena Shew**	**Cassandra Zhang**	**Jamie Jiang**	**Percentage of Total Work for the Deliverable**
Milestone 1 - EDA	5%	90%	5%	100%
Milestone 1 - Project Scope	90%	5%	5%	100%
Milestone 1 - Visualization Ideas	10%	0%	90%	100%
Milestone 2 - Data Cleaning	90%	5%	5%	100%
Milestone 2 - Novel Visualization	5%	90%	5%	100%
Milestone 2 - Coding Visualizations Per Task (4)	25% (Task 1)	25% (Task 4)	50% (Tasks 2 & 3)	100%
Milestone 2- Written Report	~33%	~33%	~33%	100%