Airbnb 2020 Listings Data Analysis

Introduction

About the Data

In this project, I examined the US Airbnb Open Data available on Kaggle. This dataset focuses on the various listings available around the country in the year 2020, some notable cities being New York, Los Angeles, and Seattle. While Kaggle functions as a crowd-sourced data science tool that contains user-uploaded data, the dataset originated from Inside Airbnb. The latter is an "independent, non-commercial[1]" website that scrapes data from the Airbnb website, and cleanses, aggregates, and visualizes it. The website's mission is to "provide data that quantifies the impact of short-term rentals on housing and residential communities; and also provides a platform to support advocacy for policies to protect our cities from the impacts of short-term rentals."
W S

Motivation

Airbnb is a well-known firm that has become synonymous with vacationing, and I wanted to explore its business through my data visualization. Airbnb connects hosts with guests in a multi-sided marketplace, an interesting business model that has been growing in popularity. The dataset contains over 200K records, allowing for the identification of patterns and trends.

Airbnb's enormous growth since its 2008 launch[2] has generated pushback and spawned legislation around the world restricting the use of the platform[3], leading to differences in availability and prices across different regions. In addition, the branding and market position of Airbnb as an alternative to corporate hotels and as a way to bring people together is mainly driven by their marketing, and not based on data, something I wanted to explore. I labeled this line of investigation the Corporate Host Phenomenon as a way to differentiate what the data was showing from what the advertisements were saying.

---

[1] *Inside Airbnb. Adding data to the debate.* (n.d.). Inside Airbnb. Retrieved March 6, 2022, from http://insideairbnb.com
[2] About Us—AirBnb. (n.d.). *Airbnb Newsroom*. Retrieved March 6, 2022, from https://news.airbnb.com/about-us/
[3] *Top Cities Where Airbnb Is Legal or Illegal*. (n.d.). Investopedia. Retrieved March 6, 2022, from https://www.investopedia.com/articles/investing/083115/top-cities-where-airbnb-legal-or-illegal.asp
[4] *Kolomatsky, M. (2021, July 15)*. What happened to airbnb during the pandemic? The New York Times. Retrieved March 10, 2022, from https://www.nytimes.com/2021/07/15/realestate/what-happened-to-airbnb-during-the-pandemic.html

Furthermore, Airbnb was impacted significantly by the COVID-19 pandemic as the “number of short-term rental listings dropped significantly between May 2020 and March 2021”[4]. Iwere interested to see what factors would heavily influence the average nightly cost of renting an Airbnb and how users could curate their Airbnb experience to fit their needs.

Key Focus

I focused my analysis on how certain features of listings correlated with the price of the listing, including the room type, the number of reviews, the number of days a listing was available throughout a year, and the number of listings per host.

While conducting this analysis, I identified a wide distribution in the number of listings per host, and Iperformed a deeper dive into this distribution, its geographic impact, and how it affects price. Part of my motivation for exploring this data - namely the widely-accepted narrative that Airbnb connects individual hosts with guests and provides greater interpersonal interactions and a means to earn supplemental income for individuals - is reflected in my analysis of the distribution of the number of listings per host.

Finally, I used my findings to provide a tool for those looking to curate their next Airbnb trip.

Design Choices, Development, and Findings

Initial Design Choices

My initial EDA process consisted of exploring different features associated with the listing price and determining the relationship between them. I plotted a subset of these features against the average price per night and added a static filter for room type. One interesting observation I found through this analysis was that shared and private rooms had a higher 365-day availability and remained on the lower end of the average price range. Additionally, the entire house/apt room type had some of the highest number of reviews at the $100/night price range. For my design choice for this graph, I used a detailed Airbnb color palette to represent the different room types. This color scheme allowed me to not only help the user differentiate between different filter types but also get a sense of the Airbnb theme. The initial EDA did not provide me with significant findings and I continued to explore the data further.

Definition of “Corporate” Airbnb Listings

By allowing almost any individual to become a host after certain requirements are met, there are many different kinds of hosts a traveler can come across on Airbnb. Many individuals have only one listing, offering up their empty vacation home for parts of the year, or maybe a single room in their house, to

make a passive income. However, there are many individuals that have 10, 50, 200, or even 500 listings on this platform. I refer to these individuals as Corporate hosts.

I encourage the user to decide how they would like to define these Corporate hosts. Whether the user believes the minimum number of listings a single host must have to be considered corporate is as little as 2 or as many as 500, the threshold on the number of listings that the user defines as corporate allows for interesting insights into the Airbnb data from 2020.

I explored this relationship between the average price per night and the number of corporate listings in my second dashboard. I included a flexible definition of the term, “Corporate Host” in my dashboard, however my focus was on letting the user define the term for themselves by including a slider that would allow the user to set the minimum number of listings that a host needed to have to be considered a Corporate Host (Corporate Host Threshold). I used a minimal coloring scheme to allow the user to see the differences between different thresholds.

How the user interprets this definition – given the relative data insights gained from my visualizations – is up to them. Many individuals want to support local hosts with only a signal listing, letting visitors into their private homes or vacation stays for possibly a more authentic cultural experience. Others might be interested in a “corporate listing '' where a single host has hundreds of listings that could possibly be less unique or authentic to the location it is in, but might offer a more standard and reliable stay with better customer service. I suspect that it is unlikely a single individual is monitoring and maintaining over 500 locations on their own and imply the word *corporate* to suggest a larger company might be running these stays owned by a “single host” in the data. This threshold on how the individual wants to define corporate listings allows the user to curate their experience in the data visualization and gain insights into how other variables in the data are affected by this threshold. Further, it allows the data to answer specific questions such as; Does the user want to support an individual local host with one or two listings in a specific area? Or do they want a stay, with a host that has 300+ listings all over the U.S? How do the different number of listings a host has changed based on location? When the Corporate host threshold increases, is there a correlation to where these listings are located? How do the prices of the stays change with the increase or decrease of the Corporate host threshold?

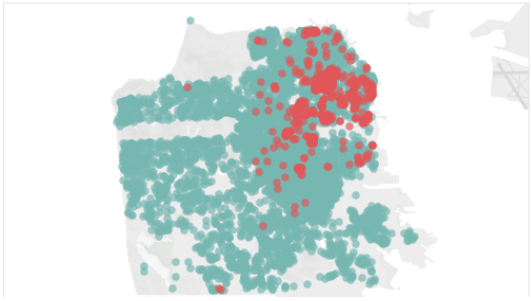
Findings

While analyzing my data and creating visualizations, I came across many insights. These would be valuable to an Airbnb user, who would like to make an informed decision regarding their next rental and its host.

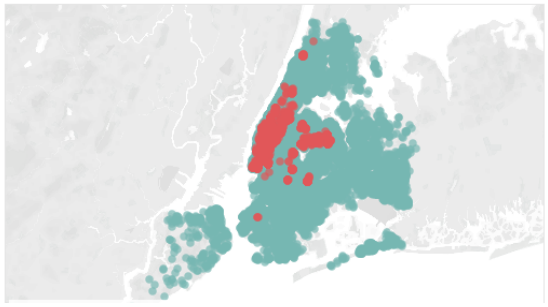
Impact of the Corporate Threshold

- As I increased the corporate threshold, I discovered that smaller cities, such as Salem, Santa Cruz County, and Oakland had a smaller corporate listing percentage. Places, i.e, Austin, Seattle, etc., that were more popular with tourists had a higher percentage of corporate listings, with Boston and Hawaii having the largest percentages.
- As the threshold increased, larger cities such as New York, Los Angeles, and Chicago remained in the middle of the "Cities with Most to Least Corporate Listings" scale.
- I found evidence of legislation restricting the use of Airbnb's platform when comparing the percentage of corporate listings in New York City and Jersey City, two localities across the river from each other but in separate jurisdictions; New York City had a much lower percentage of Corporate Listings than Jersey City, likely a result of New York's Multiple Dwelling Law[4].
- In larger cities such as San Francisco, Chicago, Los Angeles, and New York, Corporate Listings tended to cluster together geographically.

11.78% of listings in San Francisco were Corporate when the Corporate Host Threshold was set at 50



3.38% of listings in New York City were Corporate when the Corporate Host Threshold was set at 50



Impact of the Price

- While I anticipated there being a relationship between the percentage of corporate listings and the average price per night, I was not able to find one.
- As the price of a listing increased, the number of listings per host decreased. Additionally, the number of "Entire home/apartment" listings increased as well.
- Surprisingly, the highest average price of listings was displayed by Twin Cities MSA, closely followed by Broward County and Chicago.

Impact of Room Types

- "Shared room" and "Private room" listings had a higher 365-day availability and tended to remain towards the lower end of the average price range. While the "Entire home/apartment"

[4] https://www.airbnb.com/help/article/868/new-york-ny

listings averaged between $200 and $300 per night, the “Hotel room” category showcased the most price variability. Many listings in this category were also the most expensive.

- The “Entire home/apartment” had more reviews on average, and prices between $100 and $200 had the highest number of reviews, possibly because more people had stayed at rooms in this price range as opposed to more expensive options.

Usability Testing

Method

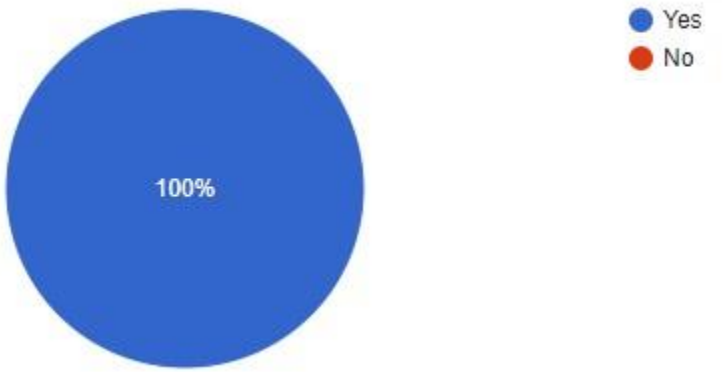
In order to conduct usability testing, I decided to survey potential users who might be interested in using Airbnb to plan their next vacation. I created a Google Form, a Google-sponsored survey tool, with ten questions. Six of these questions were multiple choice and were valuable in understanding our audience’s comprehension of the visualization. The rest of the questions were in a free-response format, allowing the users to provide more extensive specific feedback. I did not collect participant information as demographic factors did not play a role in the analysis and visualization.

Results

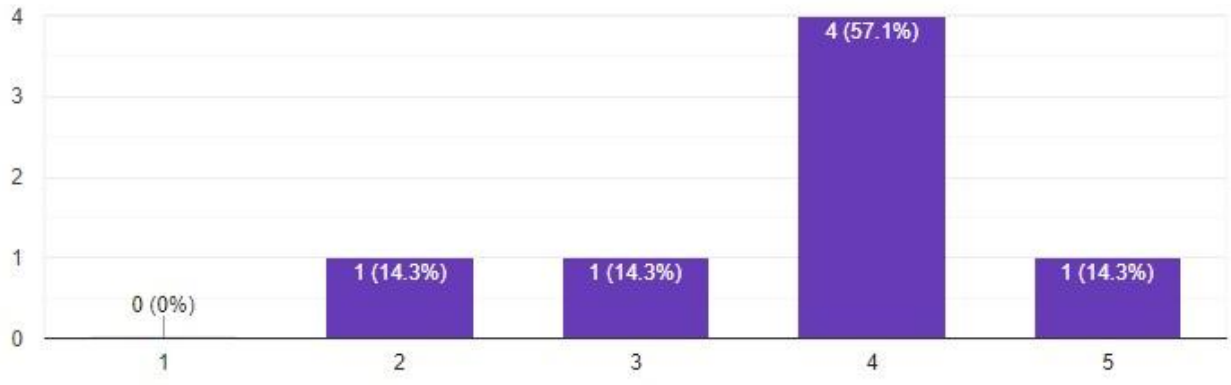
I showed my dashboard and visualizations to the users and asked them to fill out this survey once they had finished exploring the data at their own pace. The questions posed to the seven participants and their corresponding results are detailed below.

<u>Multiple Choice Questions and Results:</u>

1. **Are you familiar with Airbnb?**

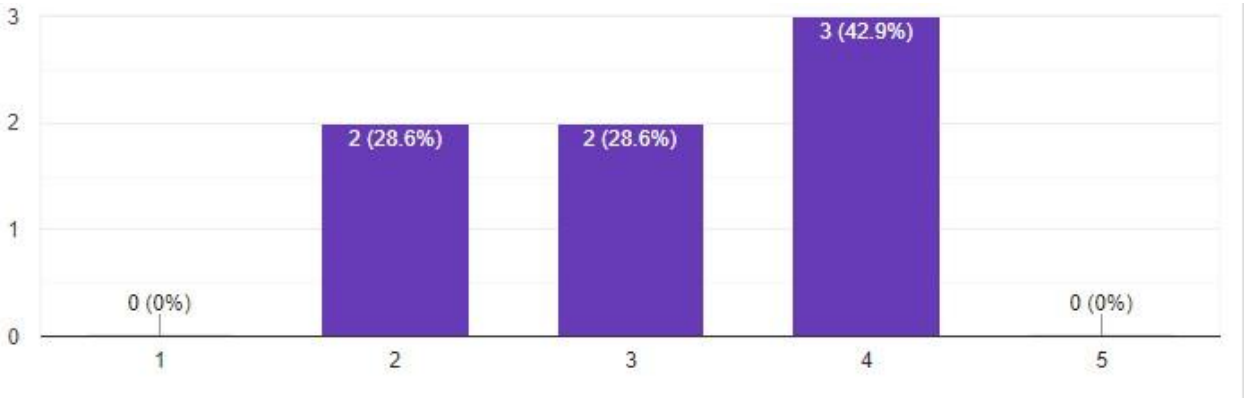


2. **How much did you enjoy the visualization overall?**



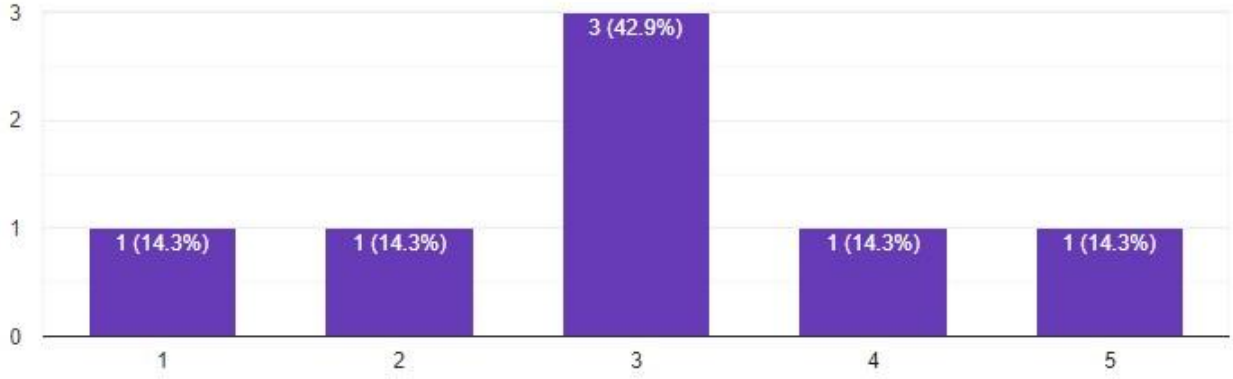
1= not very much *5= I really liked it!*

3. How much did you enjoy the narrative the visualization told?



1= Not very interesting *5= told a very compelling story*

4. How easy to use and understand were the interactive elements in the visualization?

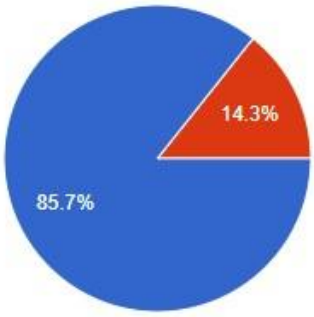


1= Very difficult to use and understand *5= super easy to use and understand*

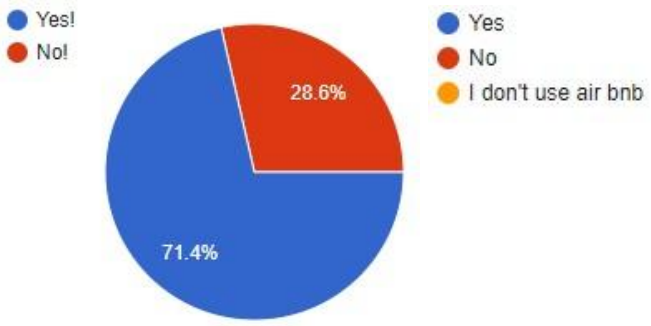
5. Did you learn something new

6. Would you use this visualization to

from the visualization?



help determine various factors in your next Airbnb stay?



Free-Response Questions and Results:

*Note: *not all responses shown, only most critical ones were selected**

What part of the visualization did you like the most?

"I thought the use of color and formatting for the graphs was very good, for the most part -- especially on the "Cities with the Most to Least Corporate Listings" section. In the "Distribution of Prices by Number Hosts" section, I thought the labeling of the specific "% of corporate per city" in addition to graphing it was a nice touch."	"The visualizations for the second tab were super aesthetic. I liked the simplicity, color schemes, and the choice of visuals to compare different cities. The third tab was very intrinsically easy to understand."	"Build your own experience I thought was really cool and well done! Also, the Geographic distribution was very clear and informative I thought."	"I liked being able to look at which cities had the best Airbnb option for me."	"I liked the adaptability of the various tables and heatmaps."

What part of the visualization did you like the least?

"I did not understand the "Corporate Threshold" filter	"For "Cities with the Most to Least Corporate Listings", I didn't understand the point of the	"Not understanding the "Number of Listings	"Probably the Relationship between prices and corporate	"The descriptions being cut off prevented me from fully

in the first listing because even after changing the value, I still got under 10, 50, etc., bins. Does the filter change the number of listings per host or something else?”	map. Moreover, the area itself sometimes looks more like a blob of color than a collection of dots, which detracts from me being able to take away any information from the map.”	Threshold" made the third view difficult to interpret.”	hosts slide, I found it a little bit confusing.”	understanding.”

What changes would you recommend to the visualization? What else would you like to learn from the data?

“Adding some more details to the spatial dimensions of the maps would allow me to better understand the significance of the concentration of the red and teal dots. Maybe each data point could have a label because I couldn't tell which city was which without having to hover over all of them. A horizontal line could help too.”	“I think the only improvement I could suggest is around the corporate host and price relationship, maybe just delving a bit more into its importance and what the data is supposed to show the observer when selecting an Airbnb.”	“Add more detail to explain the meaning of each variable. Include a line on the "Create your own experience" section to increase visual comprehension of which 'houses' belong to each city.”	“Being able to select and compare multiple cities and their prices for Airbnb”	“Allowing people to read the descriptions.”

Reflection from the results:

All of the users were familiar with Airbnb. This gave me a good reference for how to better interpret their feedback since they were acquainted with the platform and had a basic understanding of various terms and functionalities that come with the website.

Overall, through this testing, I found that the visualization had a solid foundation given the following *positive* results and feedback:

- 85% gave the visualization a 3 or above on the scale of overall enjoyment

- 85% said they learned something new
- 71% said they would use the visualization to help determine factors in their next stay.
- Enjoyed creativity in the design such as the house detailing on the final slide, color scheme, and general formatting
- Enjoyed the ability to curate their own experience as well as heatmap filtering to see geographical distribution.
- Enjoyed the combination of static visualizations – that were more intrinsic to understand– and interactive visualizations – such as the filtering on the heatmap allowing for animation based on what the user is interested in seeing.

However, I found some areas to improve the design based on the following *critical* results and feedback:

- Individuals didn't enjoy the narrative the visualization told with 57% giving it a 3 or below
    - This average-below average feedback indicated I were on the right track but could improve the storyline in the visualization
- There was also a bigger spread on the scale of the understanding of the visualization as a whole with 75% giving overall understanding a 3 or below and one individual giving it a 1
    - This very low score was important to note – even if it was only from one user – and indicated changes needed to be made for greater accessibility for the user's understanding.
- Free response feedback indicated more confusion with overall comprehension of various aspects of the visualization, explaining the lower scoring on the 5 points scale earlier in the survey
    - Some filtering features were confusing to some users
    - technical issues with text being cut off
    - Definition of Corporate Host and other key variables was noted as confusing for some individuals.
    - Some graphs were noted as too convoluted

Changes made based on Feedback

The biggest concern I noted through this user feedback all centered around the inability of the user to understand various aspects of the visualization. User feedback indicated this lack of comprehension was due to multiple elements in the visualization discussed above, so I prioritized all of the changes to hone in on making the visualization more intuitive to the average viewer. Here are a list of the changes I made to the visualization based on the user feedback:

- Created better clarity for the user through:
    - Fixing technical issues such as text being cut off and hiding the "jitter" filter from the user

    - Creating more straightforward/user-friendly graphs through deleting graphs that were confusing the user and repetitive
        - Ensured that this was not deleting any relative information the visualization wanted to portray
    - Added more detail to the titles as well as captions for the graphs
        - Specifically around the corporate host phenomenon
        - Explained filtering abilities
    - Adding consistency to colors, sizes, axis, and title names, terms and where filters are places
        - Used Airbnb Color Scheme throughout the presentation
        - Redefined “price” variable and used consistent reference to all variables throughout all visualizations
        - Started all” corporate host threshold filter” at the same number for consistency
- Improved the narrative and interactivity concerns through slowly building the number of interactive elements throughout the presentation.
    - Reordered the slides to start with the static visualizations, then next, put ones with simple filters and finally move to more complicated visualizations to attempt to fix narrative concerns.

Justification of Changes

The two main priorities I wanted to address in the adaptation of the visualization from the user feedback came from: 1. Improving clarity for the user and 2. Improving the narrative for the user.

Clarity

The most fundamental solution to this concern was fixing all of the “bugs” in the visualization. Through fixing these technical issues as listed above, the user now has access to how I originally intended the visualization to appear and therefore overall improved comprehension of the design. Fixing consistency in fonts, colors and starting threshold values also enabled a straightforward solution to creating better clarity for the users.

Through deleting the graph in the second visualization, I was able to solve a lot of other corresponding user issues that came along with this graph. Many individuals found it too complicated to understand, therefore no valuable insights were being made for the user from this visualization and therefore were providing no meaningful impact for the user experience and subsequently adding more jargon and complexity to the analysis as a whole. A worry I had about deleting this visualization was that even though the user was not fully understanding what insights were being represented, I didn't want to take out any information that users could have gained that would make for a more compelling analysis as a whole. To alleviate this concern, I ensured that the information I was trying to portray through this visualization was present in a different graph in the analysis, but this time in a cleaner and more

intuitive design. In the final visualization artifact, the second dashboard offers parts of the deleted graph's data in a simple, clean and intuitive manner that acts as an introduction to the analysis and reflects effectiveness in the changes to address the user feedback concerns mentioned above.

When adjusting the visualizations to now contain the information lost from the deleted graph, I took it upon ourselves to adjust the titles, captions, and definitions of variables throughout the analysis. By defining the variables and making them consistent throughout the analysis, it ensures less confusion over what these variables mean and when I am referring to them. Building on this, through adding more in-depth captions, explaining the corporate host phenomenon better, and walking the user through the visualization more closely, I enabled better accessibility to the designs without removing the interactivity of the user experience and knowledge.

Narrative

Addressing the second issue around narrative, which was less consequential in the feedback, began with the changes I made to allow for better clarity. By uniting aspects of the visualization (font, text, color, formatting, etc), it allowed for a more cohesive design and better storytelling analysis in general. Furthermore, through adding more details to captions and definition to terms, it walked the user through the data more which through allowing for better clarity, subsequently allows for a better narrative in the analysis since the user is now able to understand the insights the visualizations give and how this ties into the story of the analysis as a whole.

I then created a more cohesive narrative through starting with the most basic visualization with three static graphs and one filtering ability presenting the base variables and relationships in the analysis. This builds the foundation for the analysis and presents the user with information that they can now delve into more thoroughly on their own in later dashboards with more interactivity. The next dashboard shows the user the corporate threshold variable and filtering options for this definition. Having an entire dashboard and relatively simple interactive element allows the user to become more familiar with this corporate concept I created while building on the previous insights into the base variables in the analysis from the first dashboard.

After this I lead the user to a more complicated dashboard where they can gain more insights from this corporate threshold and have access to tooltips showing them different geographical locations. This complex feature allows for the user to see geographically where these corporate locations are and how they might change based on threshold value and location. This 3rd graph makes the most sense narrative wise; as the user is now familiar with the corporate threshold and interactivity elements and allows them to explore these variables mentioned in the first dashboard on their own. Lastly I combine all of the relative information the user has interacted with thus far and allow them to curate their own Aribnb experience. Ending with this dashboard concludes the narrative, and allows for the culmination of all of the previous insights gained from analysis to be presented in a fun and interactive way for the

user. Through these adaptations to how the analysis was presented, I believe the overall narrative was greatly improved from the user testing feedback .

Curating User Experience

I let the motivation and key focus guide the creation of the user experience. The aim was to enable a wide range of users to gain an understanding of the factors that influence the price of listings, build an intuitive understanding of the different classes of hosts on the platform, and use this knowledge to help the user find the experience that best suits their situation. The feedback from the usability tests was essential in curating a user experience that was pleasant, informative, and interesting.

The components of the user experience I deemed most important included providing an overview of and explanation of the data and including key terms before I showed any of the data visualizations. Next, I wanted a coherent look and feel to the dashboards that comprised the story so that users would become accustomed to where to look for information, what the different colors indicate, and gain the feeling that they were being guided through a narrative while still being able to perform an investigation on their own. Encouraging self-discovery and curiosity can pose challenges - especially for novice users - because the data can be misleading if filters are applied, or if users do not have a solid understanding of how the terms and measurements are defined. If too many filters, parameters, and other options are available, the user can get lost and find it difficult to connect back to the main thread of the narrative.

For this reason, in the first data visualization, I opted for simplicity in terms of interactivity, restricting it to highlighting room types in the distribution of various features and prices, priming users for the following visualizations in which they dive deeper into the data and eventually use their knowledge to help them select the experience that is best for them.

The second visualization introduced the concept of the “Corporate Host”. This concept was so fundamental to the accurate interpretation of the data that I decided to include the definition in the subtitle of the visualization. Next, I provided the user with a parameter that allowed them to change the definition of what a Corporate Host was, but made very clear what this parameter was doing. No other filters were available for this first view, although there were details in the tooltips.

The third visualization similarly employed the concept of a Corporate Host, and the layout and parameters were similar to the previous dashboard, although the definition of Corporate Host was not repeated. No other filters were applied, but more in-depth tooltips were used to help users understand exactly what they were looking at and to show them the geographic distribution within each city of Corporate Hosts and all other hosts, potentially providing additional insights without needing to create a separate dashboard.

The final visualization provided the user with considerable flexibility, and for the reasons explained above, I wanted to ease the user into this final dashboard by introducing terms, key concepts, and filters slowly so as not to overwhelm them. The final visualization included four filters and a parameter to help users narrow down where they should travel and which room type they should seek out.

Conclusion

User feedback provided a critical feedback loop to the data visualization process. There are countless small decisions that go into designing a data artifact that can convey insights and information, and without feedback from users, the design will not be optimal. In terms of the Airbnb data used, I found that creating my own measures to supplement what the data provides can actually become the core of the analysis, evidenced by the Corporate Host Phenomenon I investigated. However, these new terms can be overwhelming or confusing for users, and considerable effort should be made to simplify them as much as possible and provide the users with concise and intuitive definitions.