Visualizing Airbnb Rentals in New York City

Kabir Chaturvedi, Monisha Patro, Nigel Mills

Abstract

Over the past decade, Airbnb has profoundly reshaped the global hospitality landscape, offering flexible, affordable, and often unique lodging alternatives to traditional hotels. In New York City—one of the world's most visited and culturally diverse destinations—this platform has flourished, providing visitors with countless accommodation options and hosts with lucrative business opportunities. Yet, the sheer scale and complexity of NYC's Airbnb market make it challenging for stakeholders to fully understand the underlying patterns, trends, and opportunities hidden within the data.

Our project seeks to transform raw Airbnb data from New York City into clear, insightful, and interactive visualizations. By analyzing a wide range of attributes—such as location, pricing, listing characteristics, and guest reviews—we aim to help hosts, guests, and investors identify demand hotspots, refine pricing strategies, and anticipate shifting market conditions. Through user-friendly dashboards and transparent visuals, we hope to empower stakeholders with immediate insights, ensuring that navigating the NYC short-term rental landscape becomes both more intuitive and more strategic.

Introduction

Airbnb's presence in New York City has grown at a remarkable pace, redefining how travelers find accommodation and altering the dynamics of local neighborhoods. With thousands of listings spanning the five boroughs—Manhattan, Brooklyn, Queens, the Bronx, and Staten Island—Airbnb offers a unique window into the city's hospitality market. Property owners and investors can harness data-driven insights to boost profitability, while city planners, tourism boards, and local communities can better understand the evolving relationship between short-term rentals and neighborhood ecosystems.

Market Segmentation:
By identifying high-demand neighborhoods and property types, hosts and investors can focus their resources where they will generate the greatest returns. Visualizing the geographic distribution of listings, amenities, and ratings helps pinpoint niche markets and emerging trends in different parts of the city.

Pricing Strategy Insights:
Dynamic pricing strategies are crucial for maximizing revenues, especially in a competitive environment like NYC. By visualizing seasonal pricing trends and correlating rates with attributes—such as proximity to attractions or customer reviews—stakeholders can set optimal nightly prices with confidence.

Competitor Analysis:
Our visualizations will help stakeholders compare Airbnb's rental landscape against other hospitality options. Understanding how Airbnb listings stack up against traditional hotels or rival platforms reveals new competitive advantages and opportunities.

Demand Forecasting:
Identifying booking patterns over time can guide hosts to better prepare for peak seasons or mitigate low-demand periods. By forecasting occupancy rates and seasonal demand shifts, stakeholders can make more informed operational decisions.

Investment Opportunities:
Our analysis empowers real estate investors to spot untapped market segments. Whether it's recognizing a borough on the rise or seeing where luxury listings command premium rates, strategic investors can rely on these insights to identify and capitalize on emerging opportunities.

Revenue Optimization:
Pinpointing the factors correlated with high earnings—from guest ratings to host responsiveness—helps stakeholders tailor their offerings. This data-driven approach ensures resources are allocated to the most impactful aspects of the property listing.

Brand Strategy for Hosts:
In a crowded marketplace, standing out is critical. By observing patterns in successful listings' branding efforts, hosts can refine their marketing strategies, differentiate their offerings, and engage more effectively with potential guests.

The primary dataset we used, “New York City Airbnb Open Data”, is a collection of n = 48,895 Airbnb listings hosted on Kaggle. Each entry includes information about a listing’s host name and ID; description; physical location within NYC both for borough/neighborhood and latitude/longitude; nightly price & minimum nights required; availability; and review counts.

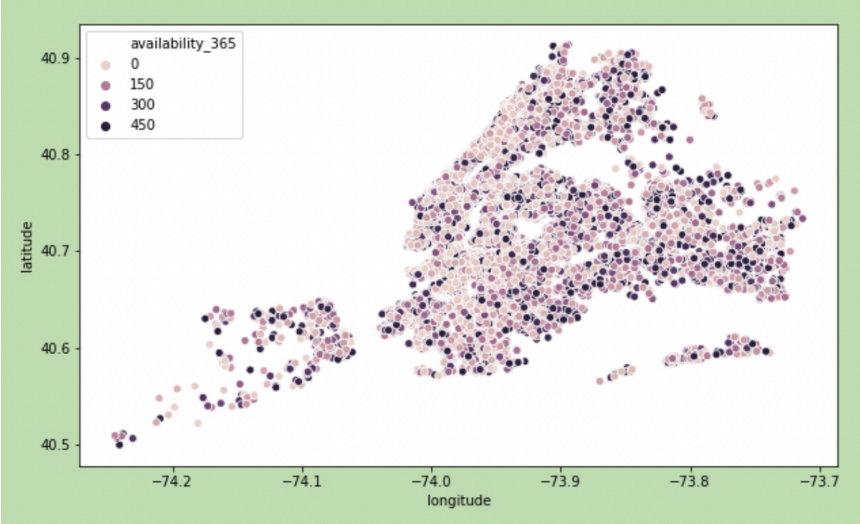
In order to compare the listed prices with actual costs to rent in different parts of NYC, an auxiliary dataset was also employed: “DOF Condominium Comparable Rental Income in NYC”, an official dataset from the City of New York local government containing n = 31,575 condominium buildings treated as rental apartments. Attributes within this dataset include address, borough, units per building, and estimates for total gross income, expenses, and net income. Taking gross income as a measure of yearly rental income, an estimate of each borough’s average monthly rent was calculated.

Data Dictionary

Column Name	**Description**
id	*int*: A unique identifier for each Airbnb listing. This ID is assigned by Airbnb and can be used to reference specific listings.
name	*string*: The title or name of the Airbnb listing as provided by the host. This field may contain descriptive information about the property.
host_id	*int*: A unique identifier for each host offering listings on Airbnb. Hosts with multiple listings will have the same host_id across those listings.
host_name	*string*: The name of the host. This is the public name chosen by the host and may not be unique across different hosts.
neighbourhood_group	*string*: The borough or administrative area of New York City where the listing is located. The possible values are: - Bronx - Brooklyn - Manhattan - Queens - Staten Island
neighbourhood	*string*: The specific neighborhood within the borough where the listing is located. Examples include Williamsburg, Harlem, Midtown, etc. There are over 200 unique neighborhoods represented in the dataset.
latitude	*float*: The latitude coordinate of the listing's location. This can be used for mapping and spatial analysis. Coordinates are in decimal degrees and correspond to the WGS84 coordinate system.

longitude	*float*: The longitude coordinate of the listing's location. Like latitude, this is used for mapping and spatial analysis in decimal degrees (WGS84).
room_type	*string*: The type of room being offered. Possible values are: - Entire home/apt: The guest has the whole place to themselves. - Private room: Guests have a private room but share some spaces. - Shared room: Guests share sleeping spaces with others.
price	*int*: The rental price per night in US dollars. This is the cost to book the listing for one night and is set by the host.
minimum_nights	*int*: The minimum number of nights a guest must book when reserving the listing. Hosts set this requirement to control the length of stays.
number_of_reviews	*int*: The total number of reviews that the listing has received from guests. This is an indicator of the listing's popularity and level of guest engagement.
last_review	*string (date)*: The date of the most recent review for the listing, in YYYY-MM-DD format. If the listing has never been reviewed, this field may be blank or null.
reviews_per_month	*float*: The average number of reviews the listing receives per month. Calculated by dividing the total number of reviews by the number of months the listing has been active. If there are no reviews, this value may be zero or null.
calculated_host_listings_count	*int*: The total number of active listings that the host has on Airbnb. Hosts with multiple properties will have higher counts. This can indicate professional hosts versus casual ones.
availability_365	*int*: The number of days within the next 365 days that the listing is available for booking. A value of 0 means the listing is not available at all in the next year, while 365 means it's available every day.

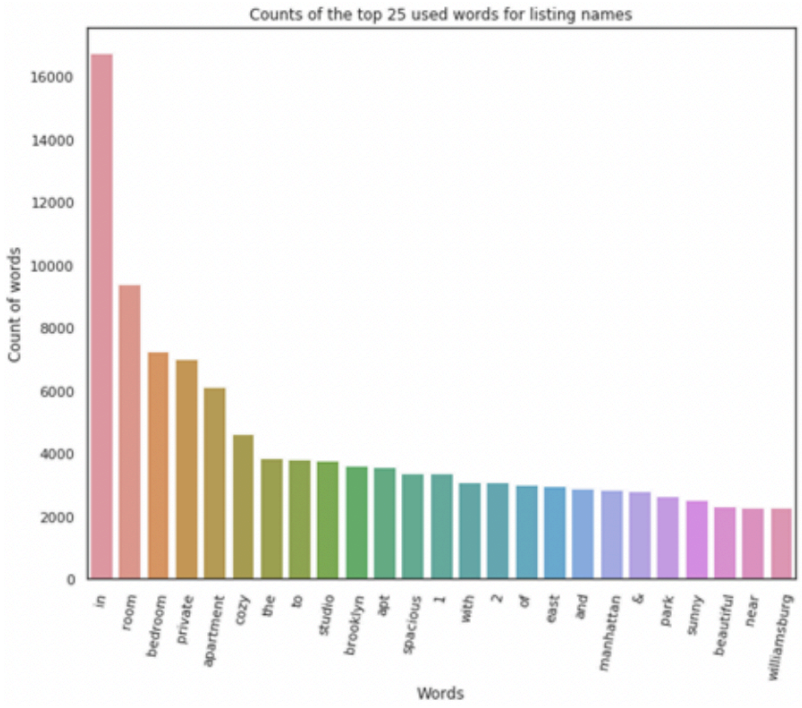
Existing Visualizations



Author: Chirag Samal

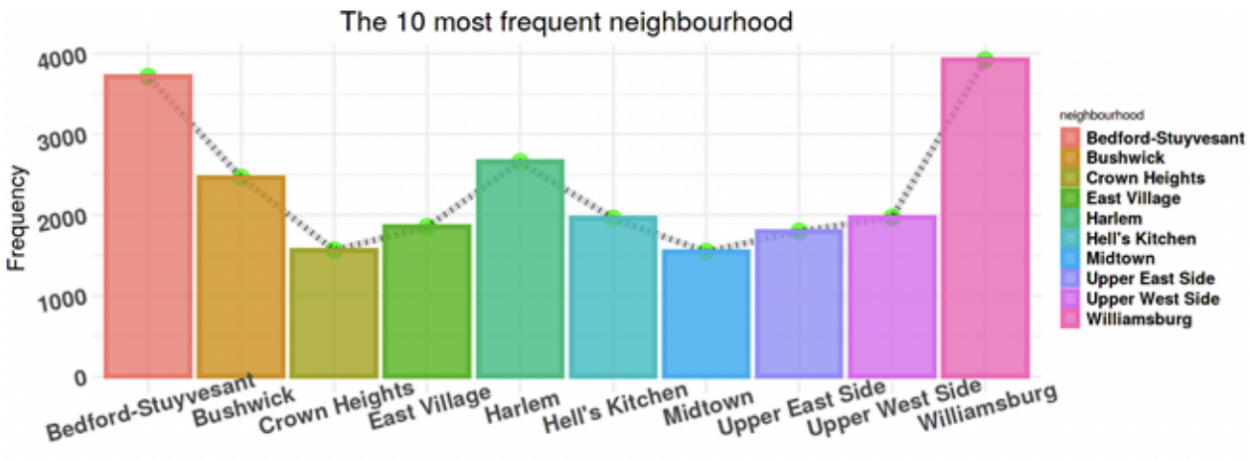
The above visualization by Chirag Samal shows how each rental is distributed geographically, with yearly availability used to color the points. While this does make it appear that some less dense boroughs, such as Queens and the Bronx, have more long-term rentals (indicated by darker points) on average, the large point size and white borders make it so there is an extreme amount of overlap between points which makes this hard to verify. Even if these issues were addressed, this information would be better portrayed using a choropleth map, which can much more effectively summarize the statistics for each borough or other areal unit. Lastly, while the color scale used does effectively combine hue and value to make different values distinguishable, the way the discrete colormap is presented makes it appear that *only* those four options for availability are present within the dataset, which is not true. In fact, it is impossible for a rental to be available for 450 days of the year.

-



Author: Dgomonov

A second visualization by Dgomonov portrays the most commonly-seen words in listing descriptions using a bar graph. This graph is relatively clear and visually attractive, but many of the entries are seemingly unimportant to track. Words such as “in”, “the”, and “&” carry little to no meaning without their surrounding context and add clutter to the graph. Additionally, the sheer quantity of the words “in” and “room” compared to the others makes it difficult to visually tell the difference between other words’ representations. Adding number labels to the bars and applying logscale to the y-axis are two potential ways to address this. Lastly, the color scheme, too, carries no actual meaning and could potentially be utilized to distinguish something such as word category.



Author : Murilão

A third and final visualization by Murilão shows the ten most common neighborhoods for Airbnb listings to occur in. Much of the visual information here is extraneous; the legend and the dotted line connecting all counts are both redundant. Additionally, the ordering of bars is alphabetical rather than quantitative, making it more difficult to compare similar values to one another. Spatial context of where these neighborhoods can be found is also critical and critically missing from the graph. Rather than employing the rainbow color for each bar, it may be worth coloring each of them based on which borough they can be found in.

Objectives and Expected Outcomes

Current Airbnb data visualizations often suffer from a lack of clarity and interactivity, limiting their usefulness for rapid decision-making. Our goal is to create accessible, transparent visualizations that allow users to quickly absorb complex information and draw meaningful insights. We will design our dashboards so that stakeholders—be they hosts, potential investors, tourists, or policymakers—can easily navigate from a high-level overview of the entire NYC market down to the granular details of individual neighborhoods.

By focusing on the city's five main boroughs, we aim to provide a holistic understanding of New York's Airbnb market landscape. Through descriptive analytics, we will unearth patterns in listing density, pricing tiers, host dominance, and guest reviews. We also plan to investigate whether properties in premium locations, like Manhattan, truly yield higher profits, and to uncover which attributes correlate most strongly with listing success. The overarching objective is to combine robust data analysis with elegant, intuitive visualization methods that empower users to gain actionable insights within seconds. This comprehensive perspective can drive better operational decisions, highlight investment opportunities, and ultimately lead to a more nuanced

appreciation of the Airbnb ecosystem in New York City

Process

Our data analysis and visualization process began with exploratory data cleaning and preprocessing steps to ensure the dataset's integrity. Initially, we addressed missing values in critical fields such as `reviews_per_month` by substituting zeroes and filtered out rows lacking essential categorical attributes like `name`, `host_name`, `room_type`, or `neighbourhood_group`. Converting certain columns to categorical types helped streamline downstream analyses.

Data Exploration and Candidate Visualizations:

We started by examining high-level distributions using boxplots and kernel density estimates to get a sense of price variability across neighborhoods and room types. This initial approach helped us understand the skewed nature of the data (e.g., Manhattan listings having notably higher price ranges) but was limited in providing multidimensional insights.

To uncover more complex patterns—such as how location correlates with price or how number of reviews ties into pricing strategies—we explored jointplots, hexbin maps, and 2D KDE plots. The spatial distribution and 2D KDE of price vs. number_of_reviews allowed us to visualize interactions between multiple variables. Yet, these simpler static plots lacked interactivity and did not easily scale to more complex relationships.

Dimensionality Reduction and Advanced Visualization Methods:

As we ventured beyond simple univariate and bivariate plots, we experimented with dimensionality reduction techniques like PCA and UMAP. PCA gave us linear combinations of features, but we found that it did not cluster listings in a way that clearly differentiated categories. UMAP proved more adept at revealing a more nuanced grouping of listings by room type, although the results were still somewhat amorphous, suggesting that price, reviews, and availability are not strongly separable along simple embeddings.

Failed Experiments and Lessons Learned:

- **Boundary-based Choropleth Maps:** We initially tried choropleth maps to show borough-level aggregates. Without precise neighborhood polygons or shapefiles directly integrated into the notebook environment, the choropleths lacked clarity and geographic precision.
- **High-Dimensional Parallel Coordinates:** While parallel coordinates helped us visualize multiple features at once, the default approach resulted in cluttered lines and made it challenging to discern patterns at a glance. Color-coding by borough partially mitigated this, but still required careful interpretation.

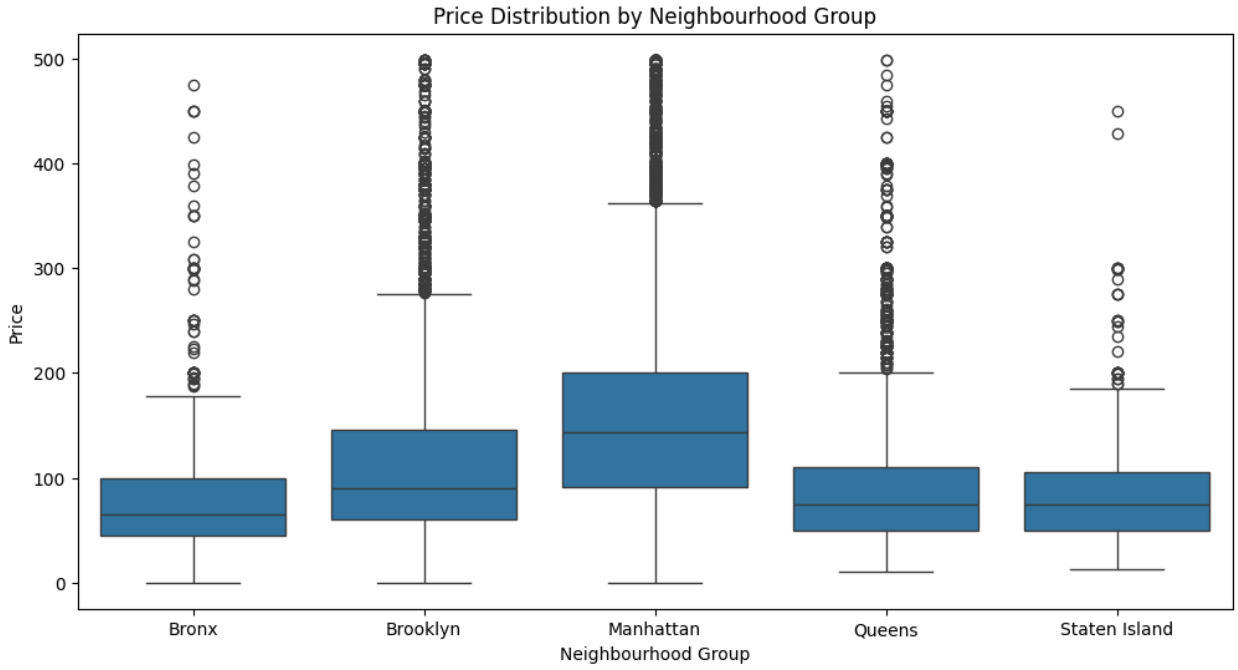
- **Complex Interactivity:** Some attempts at advanced interactive dashboards (like using Altair with complex conditional filtering) either became slow or too dense to navigate. We realized simplicity and thoughtful tooltips are key to effective interactivity.

Why Certain Methods Worked Better:

- **Hexbin Spatial Plots:** These helped distill thousands of points into smoothed density representations, making it easier to identify hotspots where listings cluster.
- **2D KDE:** Provided a smoothed contour that revealed dense clusters of listings by price and review volume, showing that most high-review listings cluster at lower price points.
- **Overlaying Listings on a NYC Map:** A simple scatter overlay on a map background was more intuitive than trying to force a choropleth. The point distribution clearly showed that the densest areas align with well-known borough centers.
- **Interactive Plots:** Adding interactivity (via Altair and Plotly) allowed users to hover over points or lines to get tooltips and additional data. For example, the host activity bubble plot (Image 9) and the yearly trend lines by borough and room type (Image 10) gave users the ability to explore temporal trends and host metrics in greater detail.

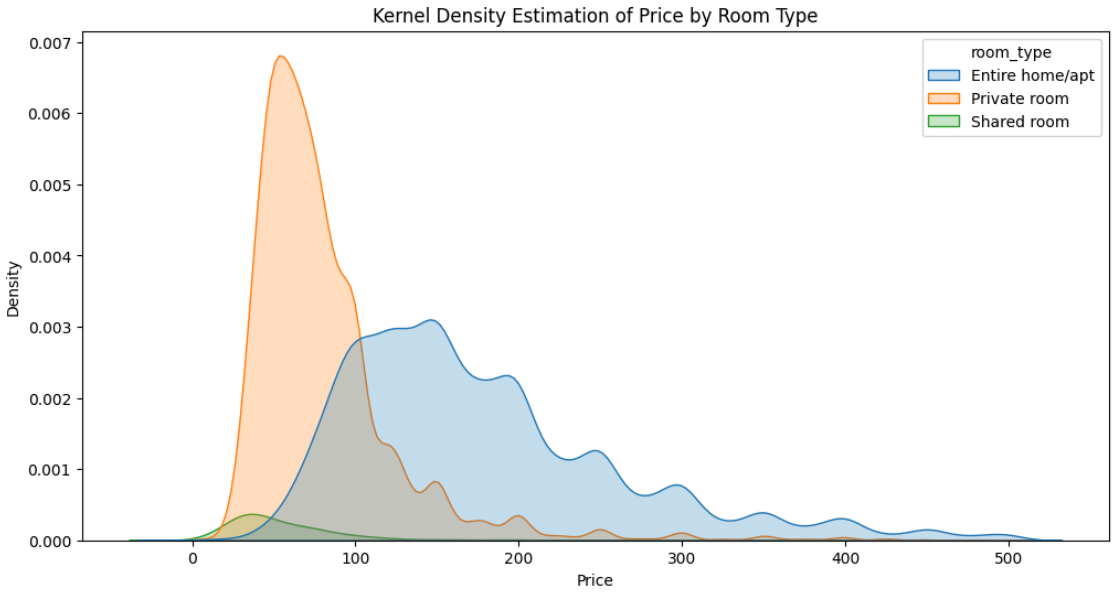
Results and Insights

Boxplots by Neighborhood Group



This visualization is a boxplot where the x-axis categorizes listings by their neighbourhood groups (Bronx, Brooklyn, Manhattan, Queens, Staten Island), and the y-axis displays the listing price. Each box represents the price distribution within that particular borough, including the median (the line inside the box), the interquartile range (the box), and any outliers (individual points). The height of the box and the position of the whiskers give a quick view of how spread out prices are, while the outliers highlight unusually high-priced listings.

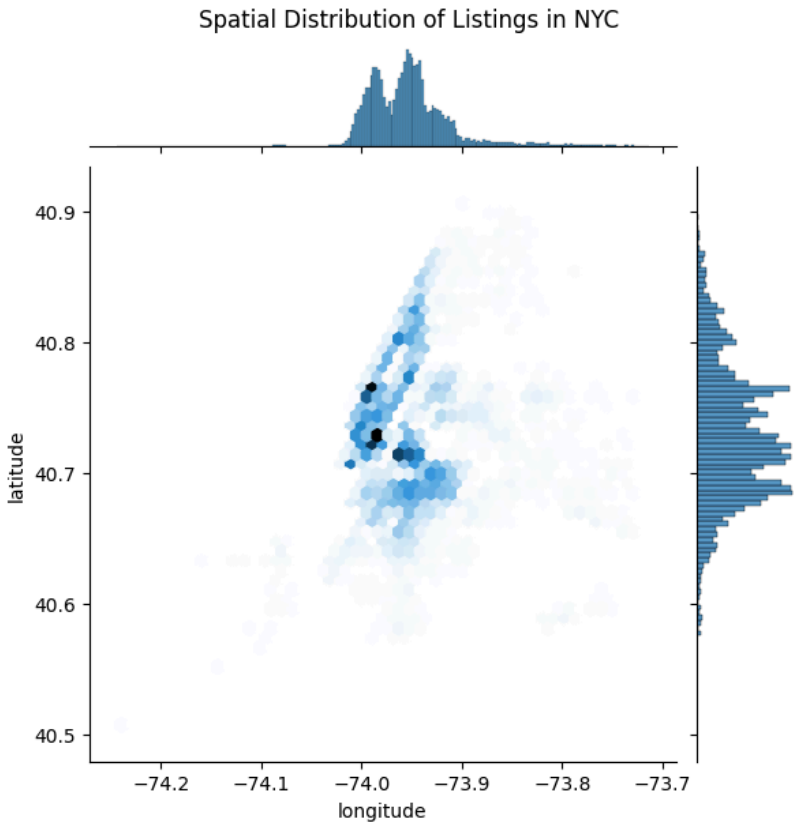
From this plot, we can understand that Manhattan listings generally command higher median prices and have a greater number of expensive outliers compared to the other boroughs. Areas like the Bronx and Staten Island appear more affordable with tighter clustering around lower median prices. Overall, the plot suggests that location heavily influences rental costs, and Manhattan stands out as the most premium market segment.



This plot shows price distributions using kernel density estimation curves for different room types on the x-axis (price) with the density on the y-axis. Each line (with shaded area) corresponds to a specific room type (Entire home/apt, Private room, Shared room), and the height of the curve at any price point indicates the relative frequency of listings at that price.

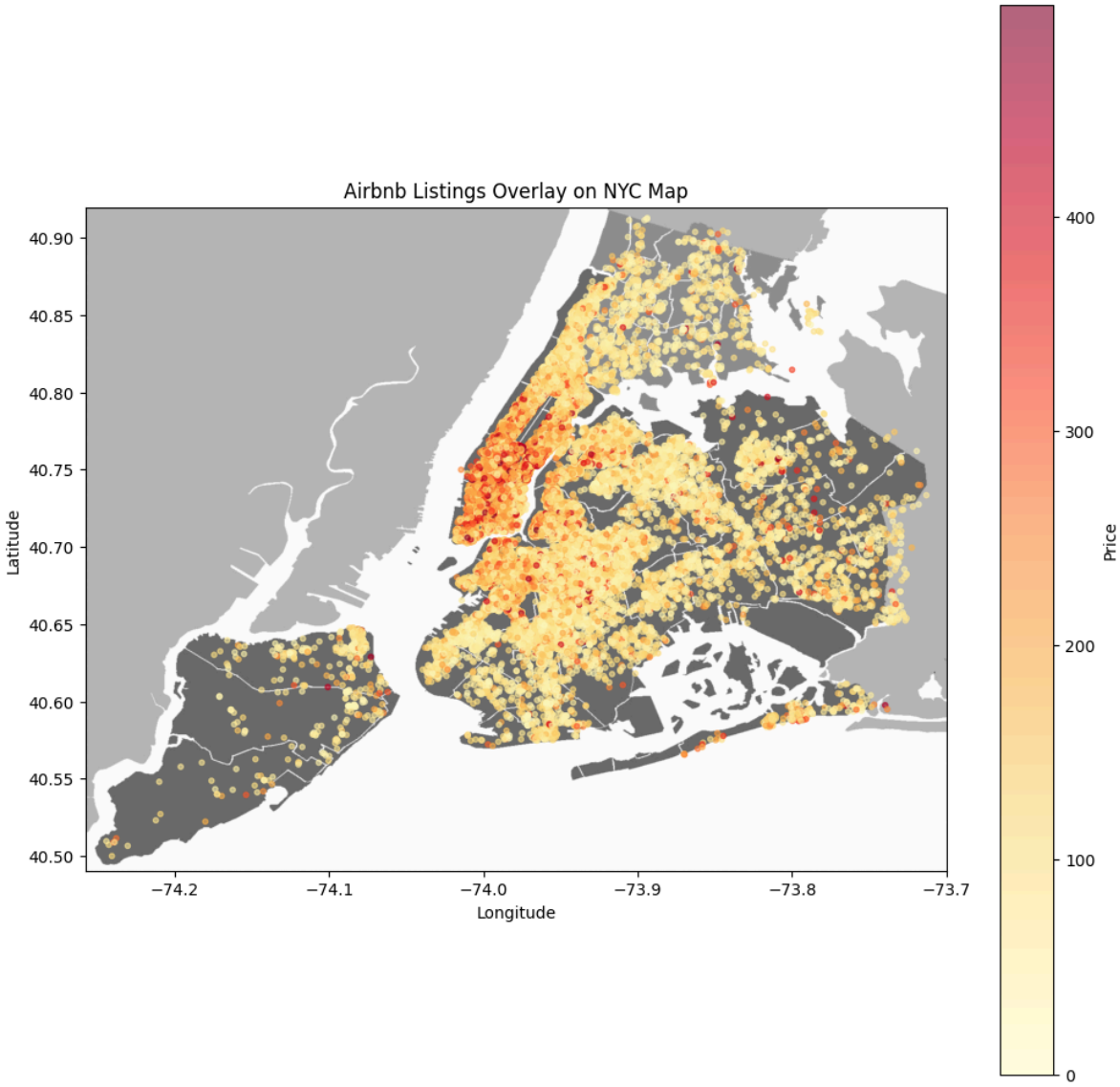
We see that private rooms cluster heavily at lower price points, suggesting budget-friendly options, while entire homes/apartments have a broader and more spread-out distribution, reaching higher price tiers. Shared rooms remain consistently low in price. This indicates a clear distinction in pricing structure between room types: private and shared rooms cater to cost-sensitive guests, while entire homes/apartments capture both moderate and high-end segments.

Spatial Insights



This jointplot uses hexagonal bins to map listing density over a geographical coordinate system, where the x-axis represents longitude and the y-axis represents latitude. Each hexagon's shade indicates how many listings are clustered in that particular region—darker colors mean denser concentrations of listings.

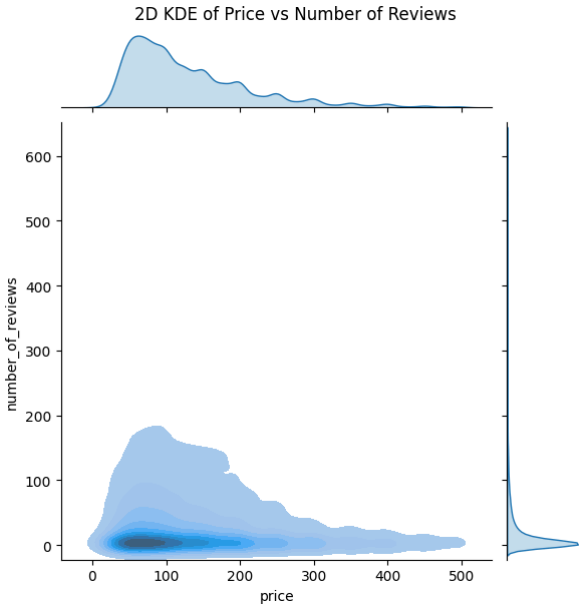
The plot reveals that listings are not evenly distributed throughout NYC. Instead, they cluster in specific hotspots—likely central Manhattan and parts of Brooklyn—while some areas remain relatively sparse. This helps identify neighborhoods with higher competition and possibly higher demand. Stakeholders could focus on these hotspots for investment or analyze why certain areas remain less populated.



The map plots actual listing locations over a grayscale image of New York City. Each point's position is determined by its longitude and latitude. The points are colored according to listing price, forming a gradient from lower-priced (lighter colors) to higher-priced (darker reds).

Seeing the distribution directly on a map underscores the relationship between location and price. Heavily touristed, central areas (like central Manhattan) show concentrated pockets of high-priced listings, whereas outlying neighborhoods display more moderately priced points. This spatial perspective confirms that geographic desirability and proximity to attractions or transit hubs drive pricing power.

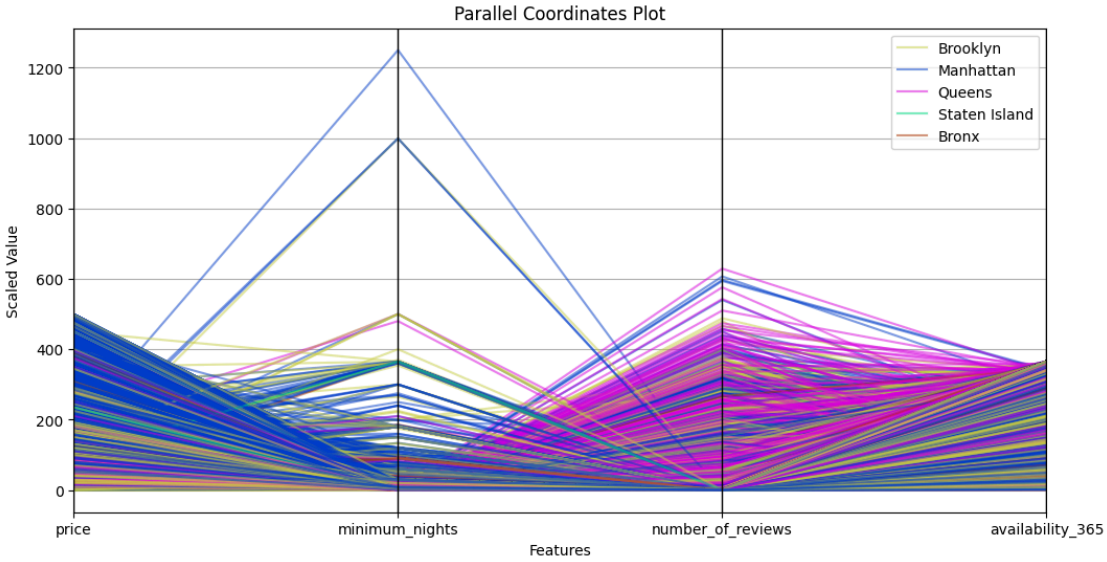
Price vs. Popularity



This visualization plots price on the x-axis and the number of reviews on the y-axis, using 2D kernel density estimation to show contours of dense regions. The contours represent areas where listings with certain combinations of price and review counts are more common, allowing us to see how these two variables interact.

From the contours, we can see that most listings with a high volume of reviews cluster around lower to moderate prices, suggesting that affordability might drive booking frequency and guest turnover. High-priced listings are present but generally have fewer reviews. This insight helps hosts understand that while luxury accommodations exist, the most reviewed (and likely busiest) properties tend to be more budget-friendly.

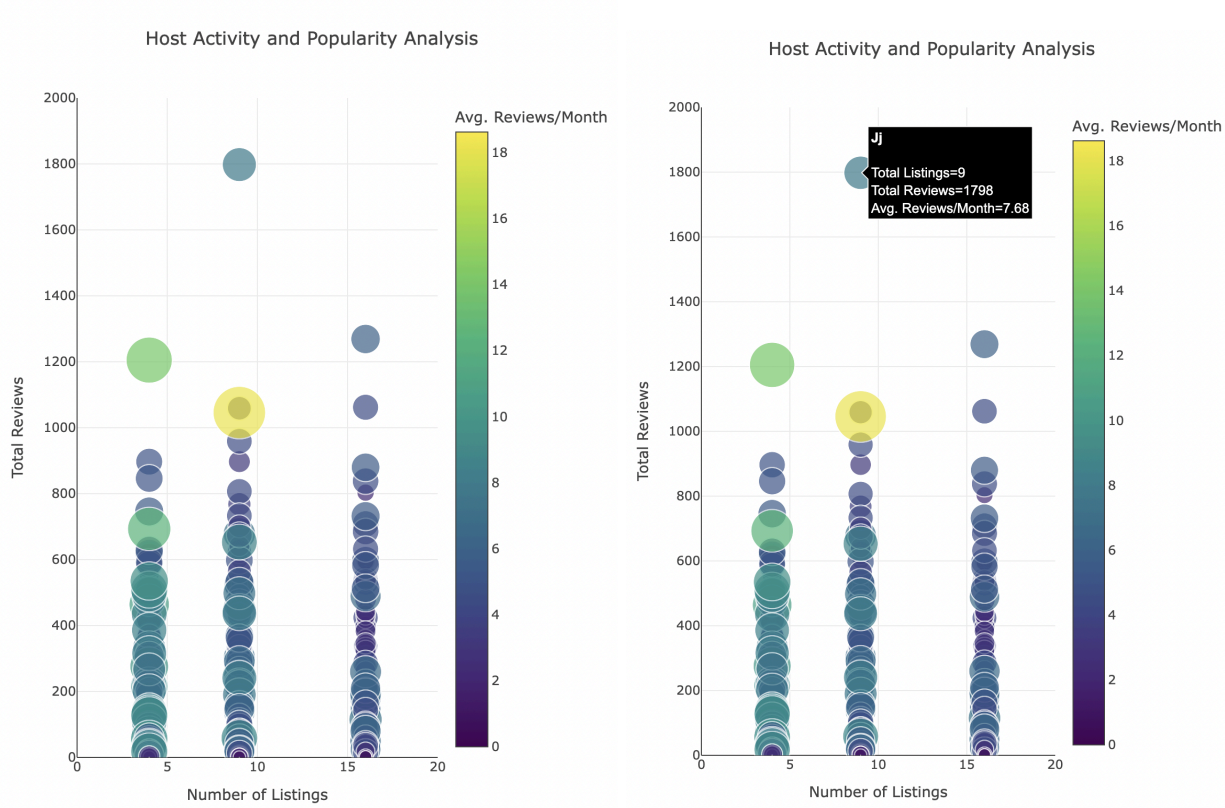
Multidimensional Patterns



Each line in the parallel coordinates plot represents one listing, passing through multiple vertical axes—each axis is a different numerical feature (like price, minimum_nights, number_of_reviews, availability_365). Lines are colored by neighbourhood group, allowing us to see how listings differ across multiple dimensions at once.

This visualization shows that boroughs differ not just in price, but also in minimum stay requirements, review counts, and availability. While the plot can look busy, patterns such as consistently higher prices or differing availability ranges in Manhattan versus Brooklyn become visible. It suggests that boroughs offer distinct market "profiles," influencing both guest experience and hosting strategy.

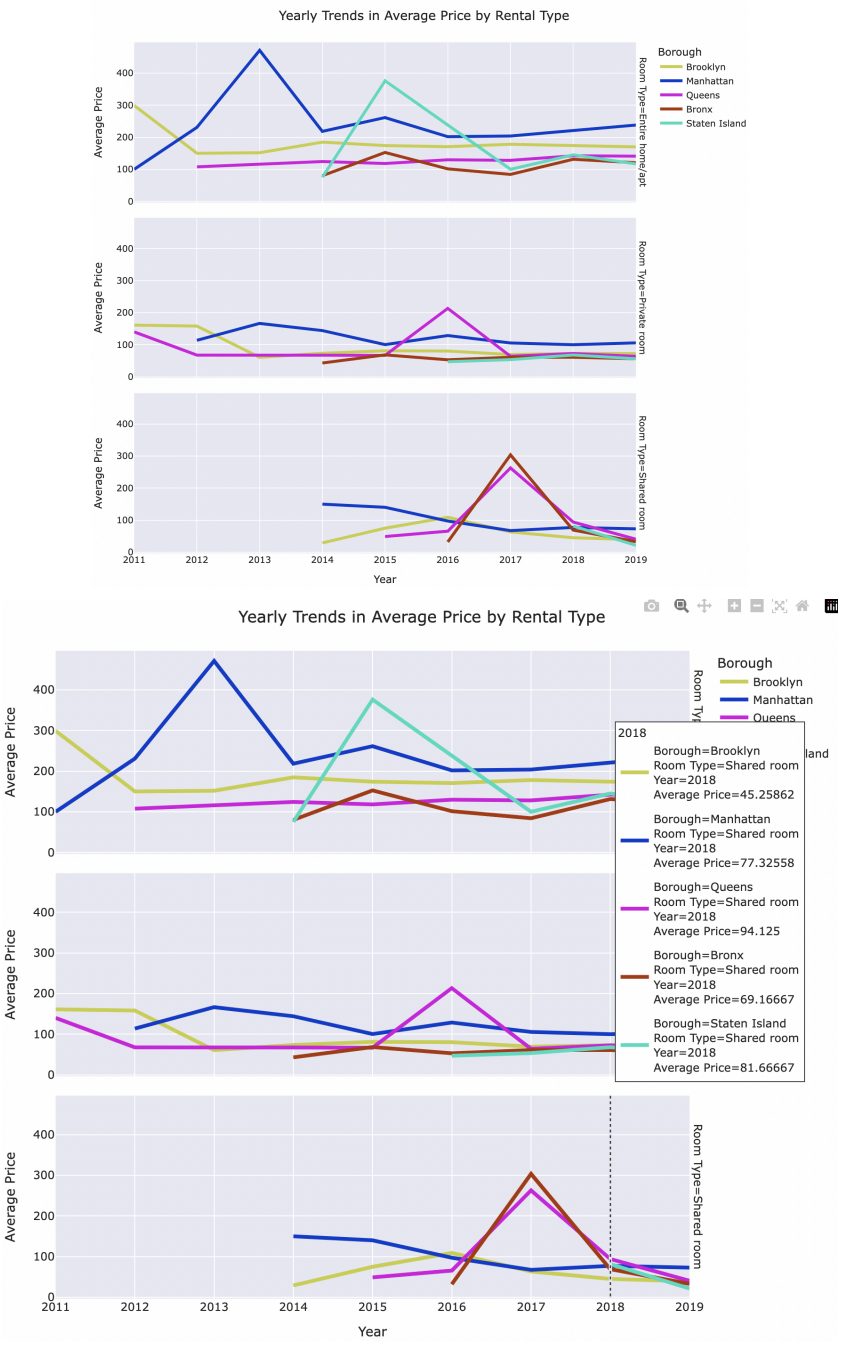
Host Activity and Popularity



This bubble chart compares hosts by plotting the total number of their listings on the x-axis and their total reviews on the y-axis. Each bubble's size and color reflect the average reviews per month. Hovering over a bubble reveals the host's name and exact metrics, giving immediate context.

For the host activity visualization, the interactive tooltips transform a static bubble plot into a dynamic exploration tool. By hovering over each bubble, users see the host's total listings, total reviews, and average reviews per month, making it easy to compare hosts at a glance. This interactivity highlights that hosts with around 5–10 listings often hit a "sweet spot" of strong engagement—too few listings might limit visibility, while too many might dilute attention per listing. The direct feedback provided by the tooltips and adjustable filtering options allow users to identify standout hosts and benchmark their performance in real time.

Temporal and Market Evolution

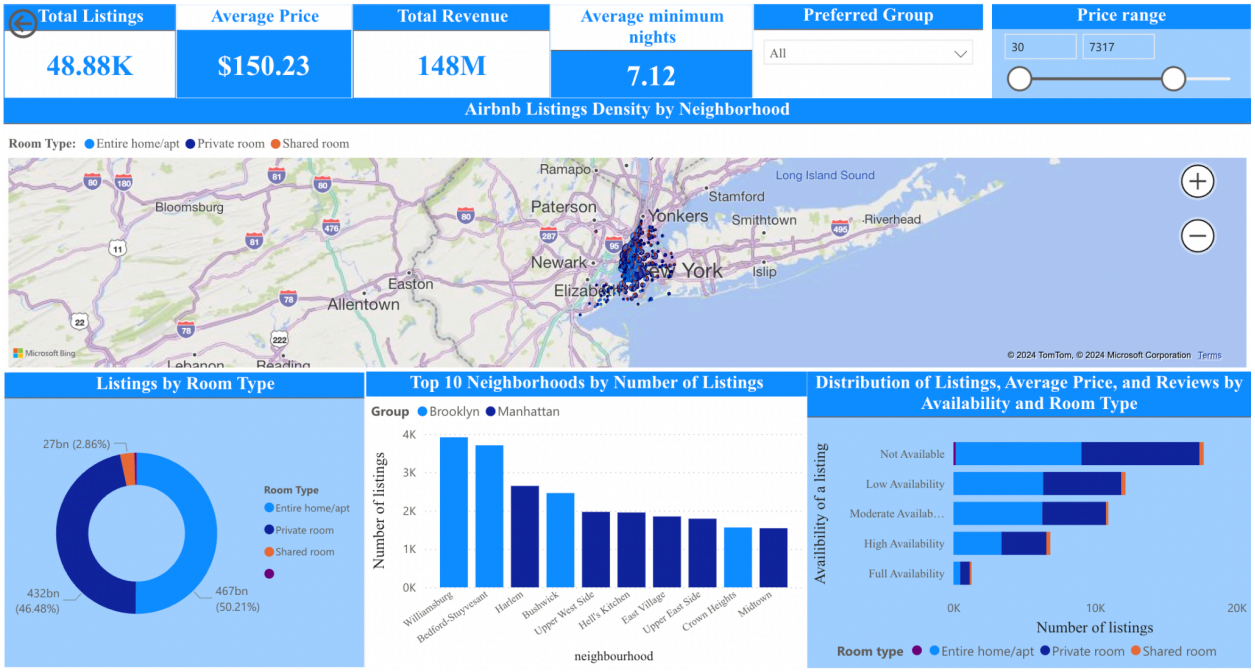


This set of [line charts]shows average price (y-axis) over time (x-axis, by year) for each borough and room type. Each line represents a borough, and the charts are faceted by room type, providing separate views for entire homes, private rooms, and shared rooms. Hovering over the lines at any point displays a tooltip with the exact price and year, making temporal comparisons easier.

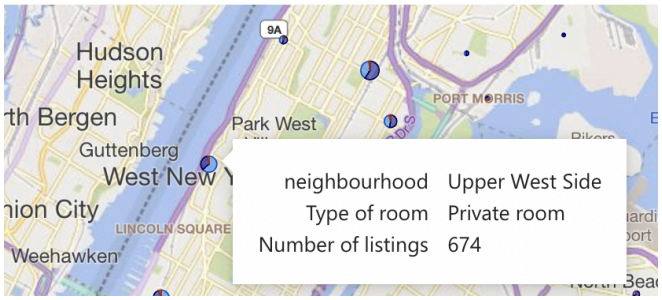
For the temporal pricing trends, interactivity and tooltips similarly enhance interpretation. By

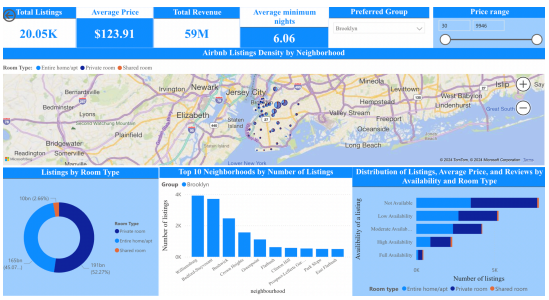
hovering over any data point, users can instantly see the exact average price for a specific borough and room type in a given year. This immediate numeric context helps confirm hypotheses about market volatility. For instance, while Manhattan and Brooklyn occasionally exhibit price spikes for entire homes/apartments—possibly due to seasonal demand or regulatory shifts—private and shared rooms remain more stable. With interactive filtering and zooming, users can zero in on a particular borough-year combination, making pattern recognition and strategic comparisons both intuitive and data-driven.

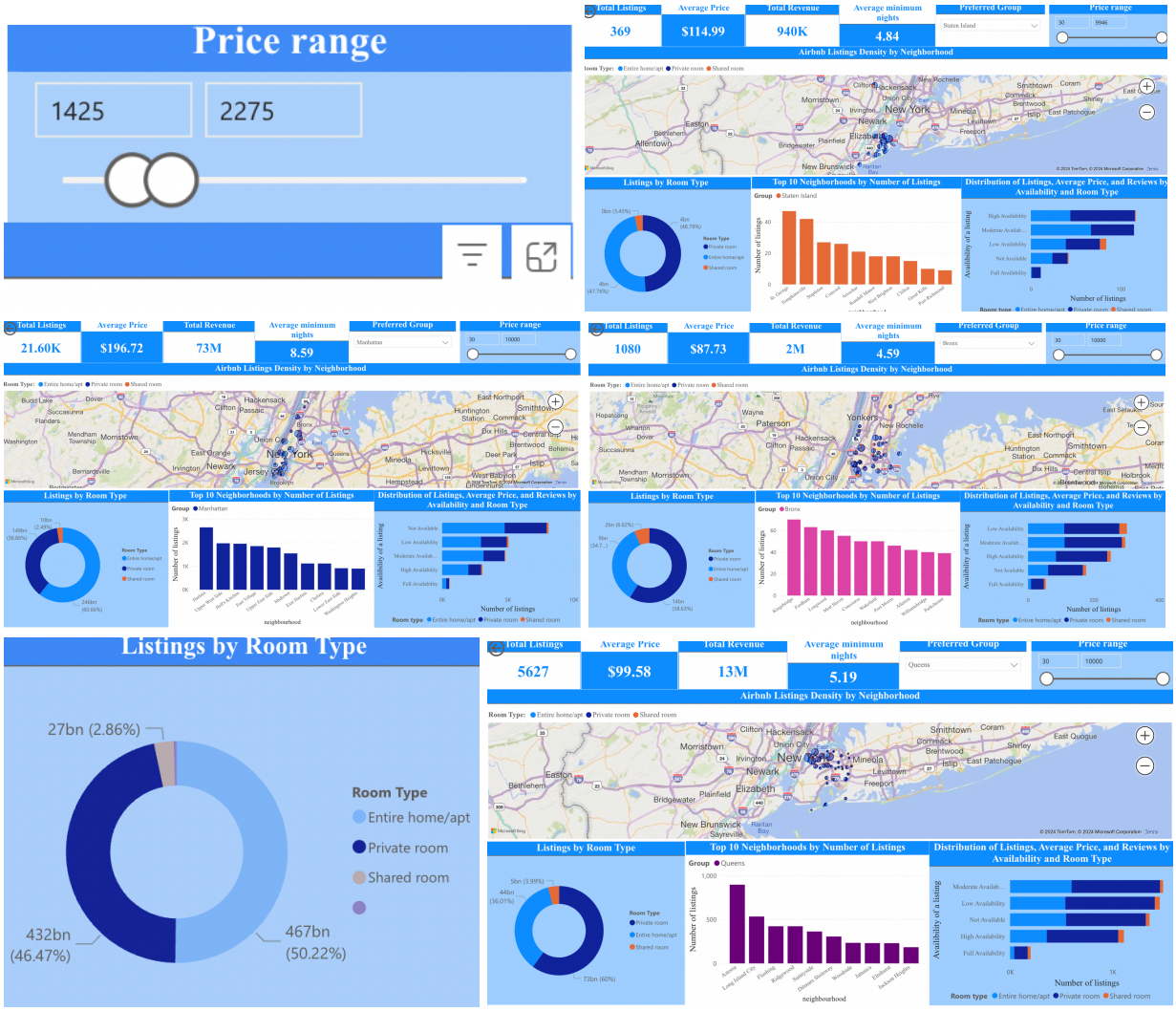
Our Power BI dashboard [custom tour]



The custom Power BI dashboard provides a highly interactive and user-friendly environment for exploring the Airbnb dataset in real time. By seamlessly integrating data filtering, geographic visualizations, and key performance indicators (KPIs), it empowers stakeholders to derive actionable insights more efficiently than static reports or spreadsheets could ever allow. Below are the key elements and benefits of the dashboard, along with the value of the two newly created metrics and why they matter.







The Power BI dashboard offers an intuitive and interactive way to explore New York City's Airbnb market. At the top, key metrics such as total listings, average price, total revenue, and average minimum nights provide an overall summary of the market. These metrics allow users to quickly understand the scale of Airbnb operations in NYC and the revenue potential.

The map at the center of the dashboard dynamically displays the distribution of listings across the city. You can filter by borough—for example, selecting Manhattan—and instantly see only listings in that area. The map adjusts to show the density of listings in Manhattan, while other visuals, such as charts for room type distribution and top neighborhoods, automatically update to reflect your selection. This makes it easy to explore specific boroughs and focus on areas of interest.

For deeper insights, you can apply additional filters. For example, by selecting "Private room" as the room type, the dashboard adjusts to show the number of private rooms in Manhattan, their price ranges, and their availability levels. If you're interested in a specific neighborhood, such as Harlem, you can filter further to focus on listings in that area. The charts for average prices,

number of reviews, and availability immediately adapt, helping you understand how private rooms in Harlem compare to other neighborhoods.

The dashboard also includes interactive charts for room types and availability. A donut chart shows the proportion of listings for different room types, while bar charts display the distribution of listings by availability and reviews. By hovering over these charts, you can see additional details, such as the exact number of listings that are highly available or the average price for listings with a high number of reviews. These features make it simple to identify patterns, such as how availability affects pricing and guest engagement.

Another key feature is the ability to compare top neighborhoods. For instance, the bar chart highlighting the top 10 neighborhoods by number of listings allows you to see which areas dominate the market. If you select Brooklyn instead of Manhattan, the dashboard instantly updates to display Brooklyn's top neighborhoods, such as Williamsburg or Bedford-Stuyvesant, and their respective metrics.

The interactive design of the dashboard enables users to focus on specific aspects of the market and uncover insights quickly. Whether you're analyzing price trends, exploring listing density, or comparing neighborhoods, the dashboard provides all the necessary tools in a clear and accessible format. It simplifies complex data, making it easy to understand how various factors, such as room type, availability, and location, influence the Airbnb market in NYC.

Descriptive Word Clouds

To start, we created a series of word clouds by extracting and combining the text within each listing's description, to see if there are any trends worth pointing out between them. The following code does this while also eliminating any stopwords and punctuation which were seen as having little value and only cluttering the visualizations:

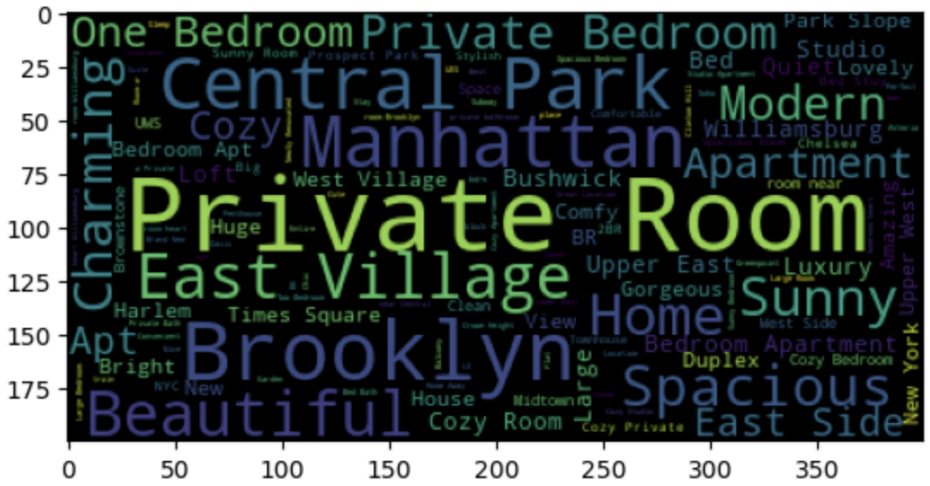
```
names = ""
for desc in airbnb["name"]:
    names = names + str(desc) + " " #add spaces between each entry
stopwords = ["a", "an", "of", "and", "NY", "NYC", "ny", "nyc",
             "the", "in", "by", "with", "to", "New York", "city",
             "for", "s", "very"] #establish words to remove from analysis
names_split = names.split()
names = ' '.join([word for word in names_split if word.lower() not in stopwords])
```

```
# Quick and dirty way to remove punctuation by replacing with spaces:
translator = str.maketrans(string.punctuation, ' '*len(string.punctuation)) #map punctuation to space
names = names.translate(translator)
```

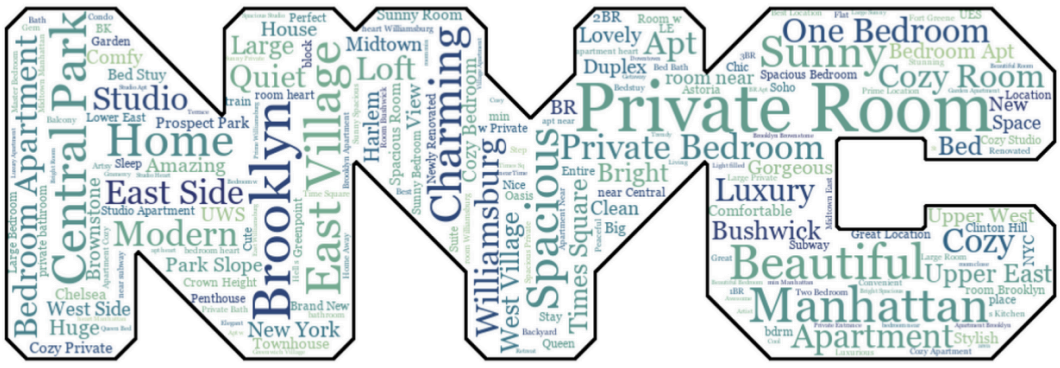
A basic word cloud was created as an initial test but not used due to there being high potential for interesting visualization techniques:

```
wc = WordCloud().generate(names)
plt.imshow(wc)
```

```
<matplotlib.image.AxesImage at 0x2cb8474ef50>
```



Word clouds can be given masks, which are essentially a shape that the words are forced to stay within. A masked word cloud can also use contours to outline the shape of the mask more clearly. The input for a mask can be any black shape on a white background, converted into a NumPy array. We used this tactic to create a word cloud for all of NYC's Airbnb descriptions using the official NYC logo as a mask. Other masks for the Airbnb logo and the Statue of Liberty were also tried, but ultimately rejected. The Airbnb logo provided less space to work with, while the Statue of Liberty outline was repurposed for Manhattan later.



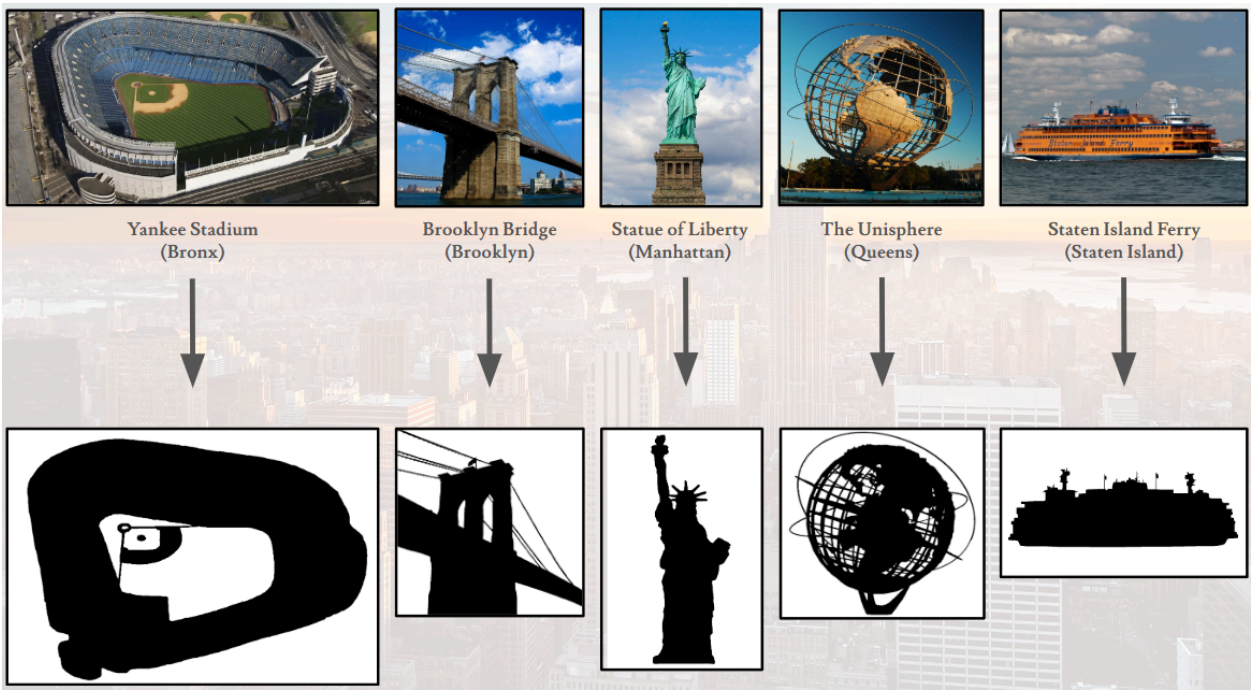
This visualization is the first of several that provides a more qualitative analysis of the dataset, and it makes the most common words and phrases stand out with larger fonts. Overall, while word clouds like these cannot be used for any direct numerical analysis, they still have value due

to their visual attractiveness and ease in understanding how to use them and what their message is. In this word cloud, we can see many mentions of different boroughs and neighborhoods (such as “Brooklyn”, “Midtown”, and “Bushwick”); rental types (such as “Apartment” and “Private Room”), and aesthetic qualities (such as “Cozy”, “Sunny”, and “Spacious”). It provides a good basis on which to judge other borough-specific word clouds which were made next.

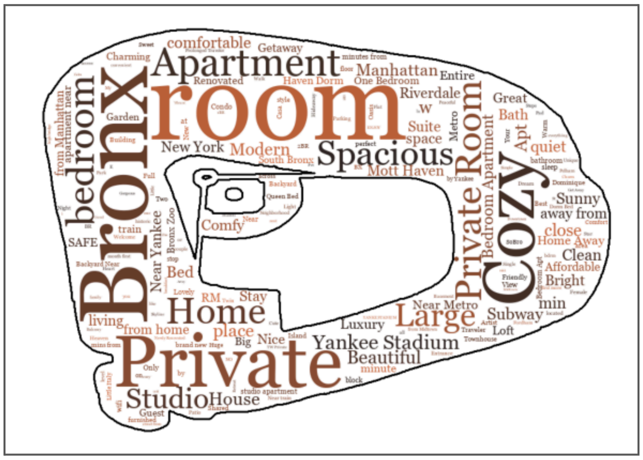
The color scheme here, Seaborn’s “Crest”, is not borough-specific and was chosen for its perceptual uniformity as well as its interesting color variation. There is a slight issue with some words being both smaller and light green, which can make them hard to see against the white background. However, this was a consistent problem with all of the color schemes which were tried, and Crest was chosen in part because it was one of the least affected by this.

The same process was repeated for each borough. The dataframe was separated by borough, and each separate dataframe had the same text manipulation applied to it. Each was also given its own mask based on a photo of a famous landmark in each borough by using Krita (a drawing & image editing program similar to GIMP or Photoshop) to turn it into a usable outline:Bronx: Yankee Stadium

- Brooklyn: the Brooklyn Bridge
- Manhattan: the Statue of Liberty
- Queens: the Unisphere
- Staten Island: a Staten Island Ferry

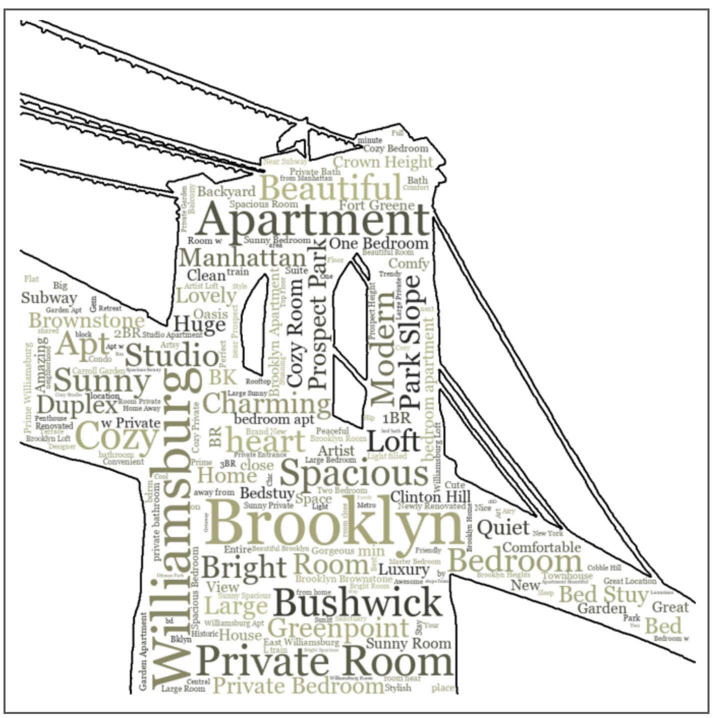


Bronx



The Bronx has a heavy emphasis on large, spacious, yet cozy and homey rental spaces. There is a larger focus on private rooms than full apartments or houses, and a proximity to New York's metro system, Yankee Stadium, and Manhattan are often cited. Some jagged edges and artifacting is present in the mask outline, but these could potentially be smoothed out with more time investment. As with other visualizations in this project, the Bronx is represented primarily by dark reds and browns.

Brooklyn



Brooklyn has lots of mentions of different neighborhoods, likely done due to the borough's high volume of listings and population density. There is a greater frequency here of some building and

living space types, such as “duplex”, “studio”, and “brownstone”. As for adjective usage, listings are often described as bright, sunny and charming. Brooklyn is represented by yellow hues which are made paler here to maximize their visibility against the white background.

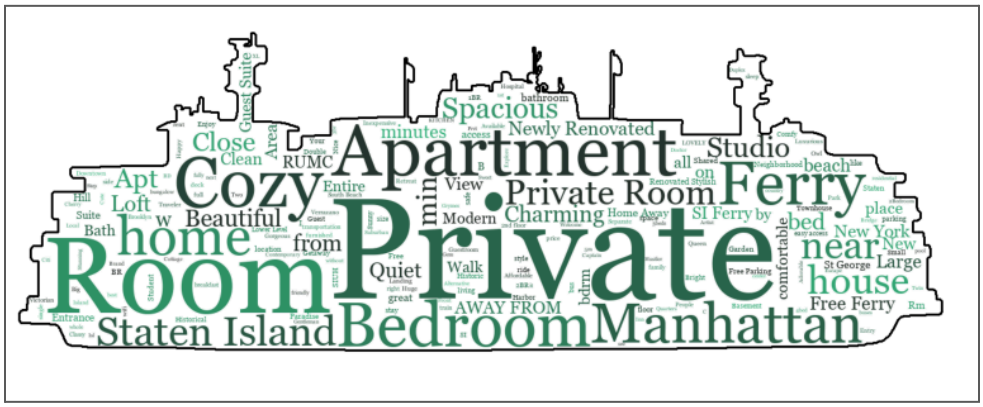
Manhattan

Manhattan has many of the same trends as Brooklyn, primarily in that lots of different neighborhoods and landmarks are mentioned. Listings here appear to be more focused on full apartments than private rooms. Common descriptors include “modern”, “cozy”, “beautiful”, and “luxury”, fitting the higher price tags found here than elsewhere in the city. Manhattan is represented using dark blues as it is in other visualizations here, which proved easy to work with and required no substantial color modifications here.

Queens

Queens is an interesting case due to how many of its listings mention proximity to different places. In fact, Manhattan is mentioned at about the same rate as Queens is in its own listings; the same also holds true for Astoria, one of the most prominent neighborhoods in the borough. Additionally, there are frequent mentions of “LGA” and “JFK”, referring to LaGuardia Airport and John F. Kennedy International Airport, respectively. Queens is also more apartment-focused than some other areas. This borough is represented using purples, pinks, and magentas, which were greyed out somewhat for this word cloud to prevent any extremely bright colors in this hue range from being physically painful to look at.

Staten Island



Staten Island is perhaps the most unique of the five NYC boroughs, and that is reflected in its word cloud. There is an extreme representation not only of private rooms but also of privacy and quietness in general, and of being "away from" certain nuisances and loud areas. Many unique amenities are mentioned as well, most notably beaches and access to the Staten Island Ferry system which is used to travel in and out of the borough. Some listings even advertise free ferry access to sweeten the deal for potential renters. Staten Island is represented using sea greens and teals, which are made brighter and bluer than other green shades to help distinguish them from the Bronx's reds and Brooklyn's yellows for colorblind viewers. Like Manhattan, this color scheme proved quite salient for the word cloud and required no modification.

Adjective Proportionality Analysis

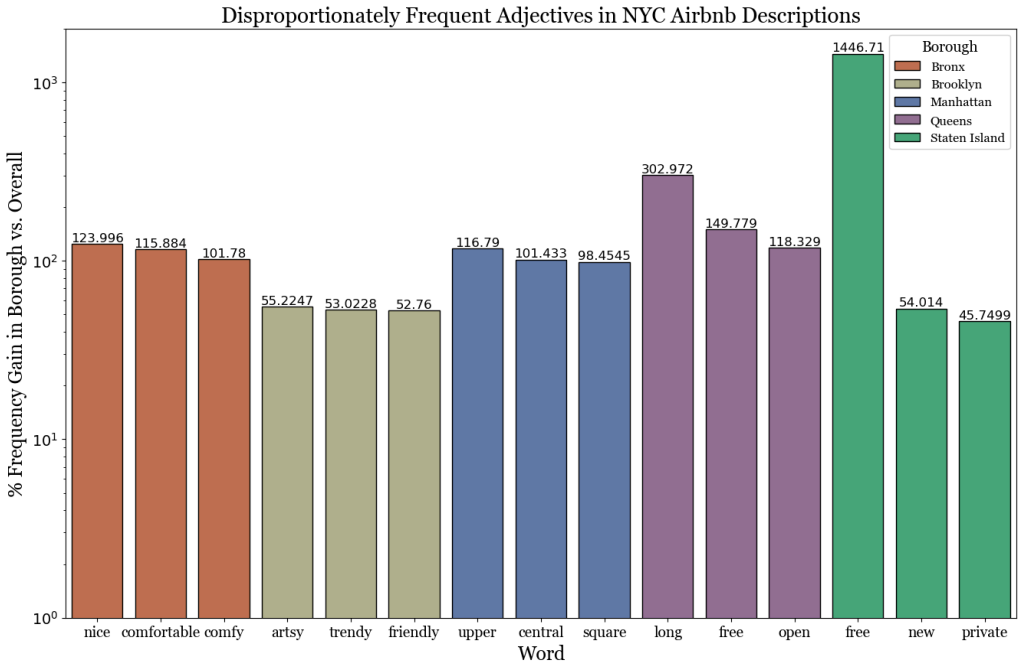
While word clouds were great for a qualitative analysis on listings, we became interested as well in portraying them quantitatively. We decided to use bar charts for this since they could provide the most direct comparison of which words showed up the most often.

First, a series of dataframes was constructed for NYC and each borough showing word presence. Using the Counter package in Python, the number of times each word showed up in the sectioned and overall text strings was counted. Each entry in the dataframe corresponded to a different word, and included information about its overall count & proportion, borough-specific count & proportion, the percent change in prevalence between borough and overall, and word type. For example, below is part of the constructed dataframe for Brooklyn's word prevalence:

```
brooklyn_counter.head(10)
```

	Word	Count	Proportion	Overall Count	Overall Proportion	Word Type	Borough % Diff
64	bedstuy	314	0.003052	314	0.001208	propn	152.641607
33	greenpoint	559	0.005434	559	0.002151	propn	152.641607
28	slope	606	0.005891	606	0.002332	propn	152.641607
131	bedford	126	0.001225	126	0.000485	noun	152.641607
141	carroll	119	0.001157	119	0.000458	verb	152.641607
72	greene	281	0.002732	281	0.001081	noun	152.641607
29	prospect	593	0.005764	594	0.002285	noun	152.216284
5	williamsburg	2728	0.026518	2734	0.010519	noun	152.087163
50	stuy	416	0.004044	417	0.001604	adj	152.035751
58	crown	341	0.003315	342	0.001316	verb	151.902888

Adjectives were much more interesting to look at and easier to combine into one overall graph. This also solved a coloring conundrum where the colors chosen for word types could be confused for the colors representing each borough. For example, the sea green used for adjectives looks strikingly similar to that used for Staten Island in other graphs, even in cases like the Bronx chart above where it has nothing to do with Staten Island.



Each borough is ordered within itself, but clustered in its own section for better readability. This approach was chosen over ordering them overall to ensure that each borough would get the same number of adjectives shown. If these fifteen were ordered by percent gain alone, it could also cause confusion if there are any words that fit into the overall top 15 but are missing from the visualization (such as if the fourth most frequent Queens adjective has a higher rate than the most frequent Brooklyn adjective). At the same time, a logscale was applied to the y-axis to preserve the height variation between bars despite some extreme values, most notably “long” in Queens and “free” in Staten Island.

Economic Analysis

The final part of our visualization incorporated official economic data from the City of New York government in order to predict how profitable it would be to run an Airbnb depending on one’s location within NYC. As such, both the Airbnb and Comparable Rental Income datasets were combined for this analysis.

Important variables kept include borough, total units, gross square footage, and estimated gross income, though some others were also added just in case we ended up needing them.

	Borough	Address	Neighborhood	Total Units	Gross SqFt	Estimated Gross Income	Gross Income per SqFt
0	Manhattan	60 WEST 13 STREET	GREENWICH VILLAGE-CENTRAL	70	82017	4452703	54.29
1	Manhattan	1360 6 AVENUE	MIDTOWN WEST	183	141738	7113830	50.19
2	Manhattan	77 PARK AVENUE	MURRAY HILL	109	158571	7329152	46.22
3	Manhattan	712 GREENWICH STREET	GREENWICH VILLAGE-WEST	20	53943	2132906	39.54
4	Manhattan	35 EAST 38 STREET	MURRAY HILL	113	88230	4288860	48.61

These were then aggregated across borough to come up with average stats for gross income per square foot, average unit square footage, and estimated mean unit rent for each borough.

	Borough	Total Units	Gross SqFt	Estimated Gross Income	Gross Income per SqFt	Average Unit SqFt	Estimated Mean Unit Rent
0	Bronx	197019	204706742	3259070091	15.920678	1039.020308	1378.492299
1	Brooklyn	363493	426698503	11349532139	26.598481	1173.883687	2601.960271
2	Manhattan	1239777	1489771174	65554743259	44.003230	1201.644468	4406.353136
3	Queens	322985	316728279	7397633712	23.356404	980.628447	1908.662867
4	Staten Island	31003	30526117	496227828	16.255845	984.618166	1333.816695

A similar process was done with the Airbnb rentals dataset, condensing it by borough and using the aggregate statistics to calculate the mean monthly price of listings:

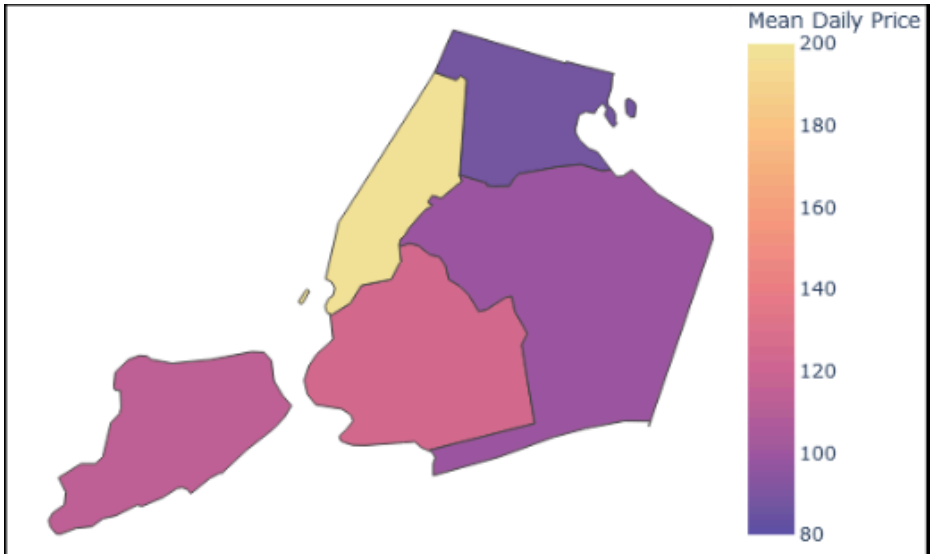
	Borough	Total Rentals	Mean Daily Price	Mean Monthly Price
0	Bronx	1091	87.496792	2663.183604
1	Brooklyn	20104	124.383207	3785.913873
2	Manhattan	21661	196.875814	5992.407579
3	Queens	5666	99.517649	3029.068446
4	Staten Island	373	114.812332	3494.600369

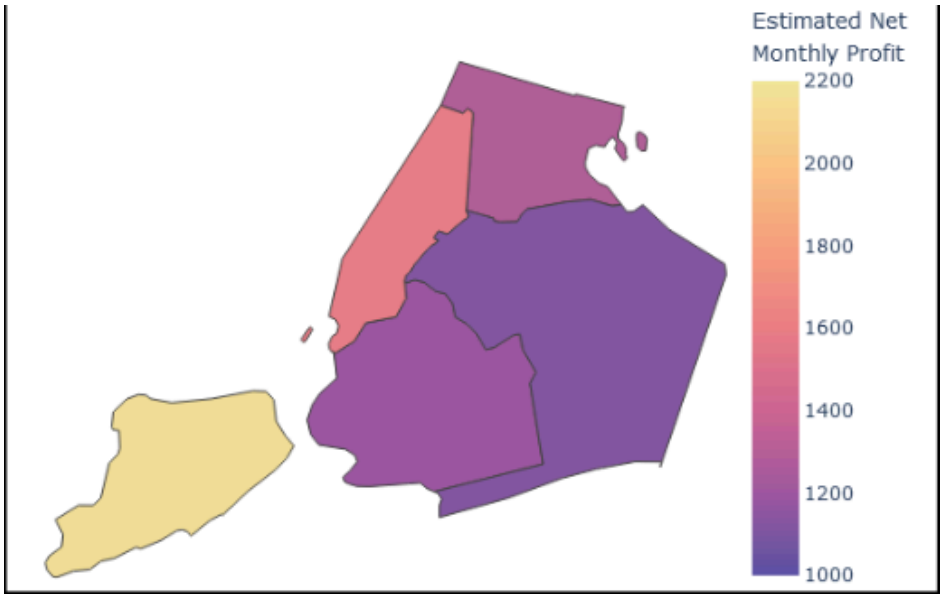
These two datasets were combined in order to calculate a final “Estimated Net Monthly Profit” attribute by subtracting Estimated Mean Unit Rent from Mean Monthly Price. As a finishing touch, two smaller versions of this dataframe were created for easier visualization later:

	Borough	Mean Daily Price
2	Manhattan	196.875814
1	Brooklyn	124.383207
4	Staten Island	114.812332
3	Queens	99.517649
0	Bronx	87.496792

	Borough	Estimated Net Monthly Profit
4	Staten Island	2160.783673
2	Manhattan	1586.054443
0	Bronx	1284.691306
1	Brooklyn	1183.953602
3	Queens	1120.405579

The above two dataframes were fed into Plotly to create choropleth maps of the most expensive boroughs for Airbnb renters and the most profitable boroughs for hosts:





Just like with many prior visualizations, Manhattan stands out as the most expensive option for Airbnb renters, followed by Brooklyn. However, looking at the estimated profit from running an Airbnb, it is actually the Staten Island lessors who seem to earn the most after monthly rent is taken into account. In fact, aside from Manhattan having high metrics in both, these two maps seem to rank each borough quite differently. The Bronx is another example of a borough that seems unassuming from its nightly prices, but because of how cheap its monthly rents are by comparison, Airbnb hosts there can likely make quite a bit of money – oftentimes even more than their counterparts in a more pricey area such as Brooklyn.

Note that these estimated profit values assume that all listings have full rental uptime, i.e. every one of them is rented at all times of the year. This is of course not always realistic, and as such, actual profits are likely lower in general than they can be portrayed using a map such as this. However, this is unlikely to significantly affect the comparative analysis performed here, and the

way the map itself is shown would likely not need any major changes. Similarly, the color scheme – Plotly's "Sunset_r" – does not appear to pose any issues and would not need to be changed in such a case.

Discussion

This project provided critical insights into the Airbnb market in New York City, highlighting patterns and trends in listing distribution, pricing, and availability across boroughs. Through dynamic visualizations and interactive dashboards, we explored how room types, geographic location, and host activity influence market performance. The ability to filter and drill down into specific boroughs, neighborhoods, or room types significantly enhanced our understanding of the data, making it easier to derive actionable insights.

However, some limitations became apparent during the analysis. For example, the absence of precise geographic boundaries (e.g., shapefiles of neighborhoods or zoning data) meant that spatial visualizations lacked the ability to clearly delineate specific neighborhoods. This restricted the granularity of the insights. Moreover, the analysis could benefit from incorporating external data, such as tourist footfall, seasonal trends, and major events, to better contextualize the observed pricing and availability patterns.

Conclusion

This project demonstrated the power of interactive visualizations and dashboards in making complex datasets accessible and actionable. By analyzing borough-specific trends, room type distributions, and availability patterns, we gained a deeper understanding of how various factors contribute to the success of Airbnb listings in New York City. Key insights included the dominance of Manhattan and Brooklyn in pricing and density, the affordability of shared and private rooms, and the importance of availability in driving guest engagement. These findings can guide hosts in optimizing their listings, investors in identifying high-potential neighborhoods, and policymakers in better understanding short-term rental dynamics.

Future Work

1. **Detailed Geographic Analysis:**
   Incorporating shapefiles of neighborhood boundaries or zoning areas could enhance spatial analyses, allowing for more precise visualizations and better identification of micro-market trends within boroughs.

2. **Integration of External Data Sources:**
   Adding data on tourism patterns, local events, or hotel occupancy rates could provide additional context for pricing and availability trends. This would allow for a richer understanding of market dynamics and seasonal fluctuations.

3. **Sentiment Analysis on Reviews:**
   Analyzing guest reviews for sentiment and common themes could offer qualitative insights into what guests value most. Hosts could use this information to refine their offerings and improve guest satisfaction.

4. **Predictive Modeling:**
   Developing machine learning models to predict future trends, such as demand, pricing, or optimal listing strategies, could provide forward-looking insights for stakeholders.

5. **Comparative Market Analysis:**
   Expanding the project to compare NYC's Airbnb market with other major cities could highlight what makes the NYC market unique and identify shared patterns or outliers across global markets.

6. **Enhanced Host Analysis:**
   Further analysis of host behavior, such as pricing strategies, response times, and listing descriptions, could reveal what differentiates high-performing hosts from the rest.

References

1. Inside Airbnb Dataset: *AB_NYC_2019.csv dataset*
2. City of New York dataset: *Comparable Rental Income.csv dataset*
3. Images used for word clouds:
   a. *Unisphere (Queens)*
   b. *Brooklyn Bridge (Brooklyn)*
   c. *Staten Island Ferry (Staten Island)*
   d. *Yankee Stadium (Bronx)*
   e. *Statue of Liberty (Manhattan)*
4. Borough boundary data: *US FIPS GeoJSON via Plotly*
5. McKinney, Wes. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and*

IPython. O'Reilly Media, 2017.
6. Pedregosa, Fabian, et al. *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 2011.
7. UMAP Library Documentation: *UMAP Documentation*
8. Plotly Express: *Plotly Interactive Graphing*
9. Airbnb Official Website: *www.airbnb.com*
10. Tableau & Power BI Best Practices: *Microsoft Power BI Documentation*