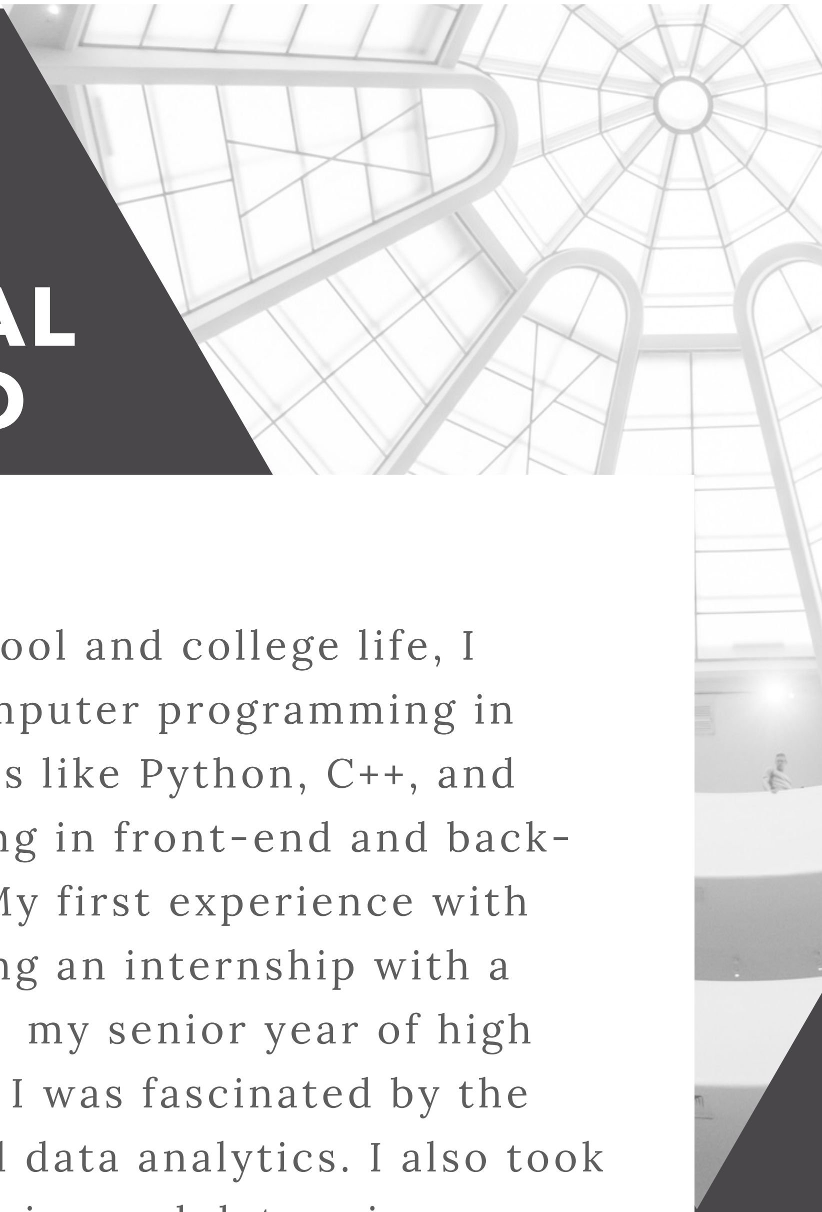
DATA ANALYSIS PORTFOLIO



MY PROFESSIONAL BACKGROUND

For most of my high school and college life, I focused on studying computer programming in many different languages like Python, C++, and Java, while experimenting in front-end and back-end web development. My first experience with Data Analytics was during an internship with a startup company during my senior year of high school. Ever since then, I was fascinated by the field of data science and data analytics. I also took a class on Machine Learning and data science as an elective which only increased my drive to be in this field. I've recently graduated with Cum Laude at Queens College, receiving a Bachelor's degree of Arts in Computer Science. While my work experience is still rather minimal and limited, my goal is to not only change that soon but to find my place in the field of Data.

TABLE OF CONTENTS

UDEMY PROJECT

CONTEXT

Udemy is a popular learning platform that hosts courses taught by experts which are taken to either hone a current skill set or to learn a new skill entirely. Regardless, of the student's purpose, there is a wide variety of courses present on Udemy whether it be music, coding, art, etiquette, or something completely new. No matter the genre, a surplus of courses are available on the platform with different levels of difficulty to consider.

DESCRIPTION

This project focuses on the relationship between the popularity of courses in the Web Development category and their level of difficulty. The goal of this project is to explain what can be done to better net a profit from these courses based on the relationship between enrollments, difficulty and pricing.

After some data cleanup and visualization, by looking into the total number of subscribers for each difficulty, the median pricing for each difficulty, and the average rating, the main takeaway was that the most popular Web Development courses are typically labeled for beginners or of all skills.

THE PROBLEM

What is the problem?

- How can we further increase Udemy's revenue using the data of the company's courses?

How long should this project be?

- Approximately two to three weeks.

What data should be collected to understand this problem?

- We should look into the data that shows the revenue of each courses. I would want to see the different of revenue in more popular courses compared to the less popular.

How should the data be presented?

- A graph that can easily compare and visualize the revenue of each course over a period of time should be utilized. A bar graph can easily achieve this, for example.

What questions should be asked to better understand the business problem?

- Who (age, occupation, etc.) buys these courses and when do they buy them?
- Are certain courses being bought more during the timeframe of a school semester or are they being bought when more people are free?
- Why are people buying these courses (discounts, or other incentives)?

DATA DESIGN

Originally, each Udemy category and their course data was in their own spreadsheet. After some data consolidation, I was able to import the data of all Udemy courses into a single spreadsheet.

This however would mean that the single spreadsheet is more likely to contain rows with null or blank values representing a column. As such with the process of data cleanup, I removed any duplicates and rows that contained null/blank values inside.

Using the refined spreadsheet, I decided to use Tableau to visualize most of my tables as I feel the visualization options presented in Tableau is better suited for comparing attributes of course difficulties compared to Google Sheets.

FINDINGS AND ANALYSIS

Figure 1: Subscribers by Udemy Category

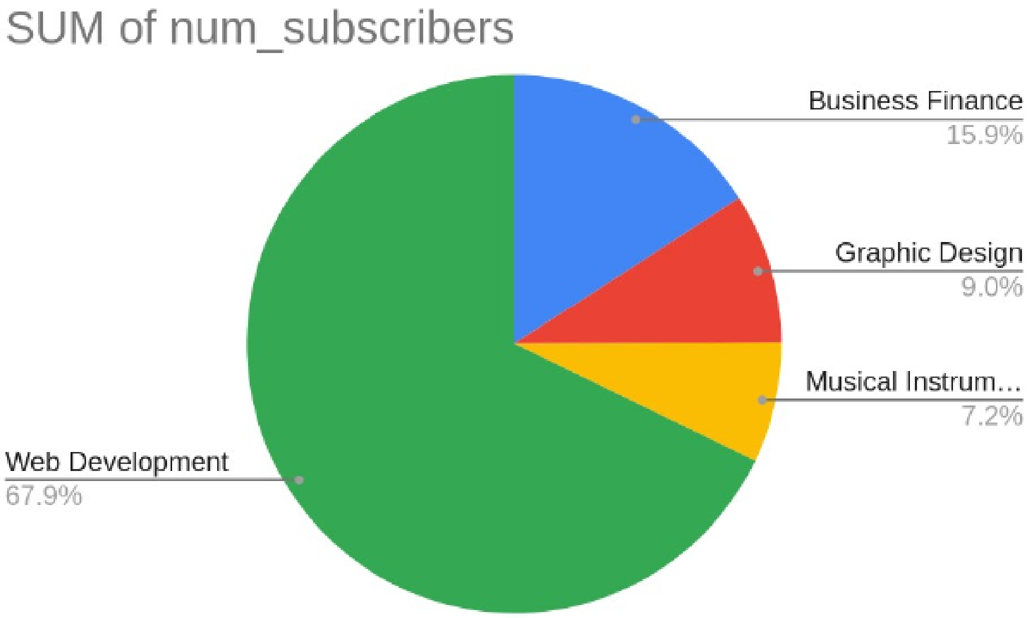


Figure 1 visualizes the total amount of subscribers depending on the course category. Web Development covers the majoirty of with 67.9% of the courses.

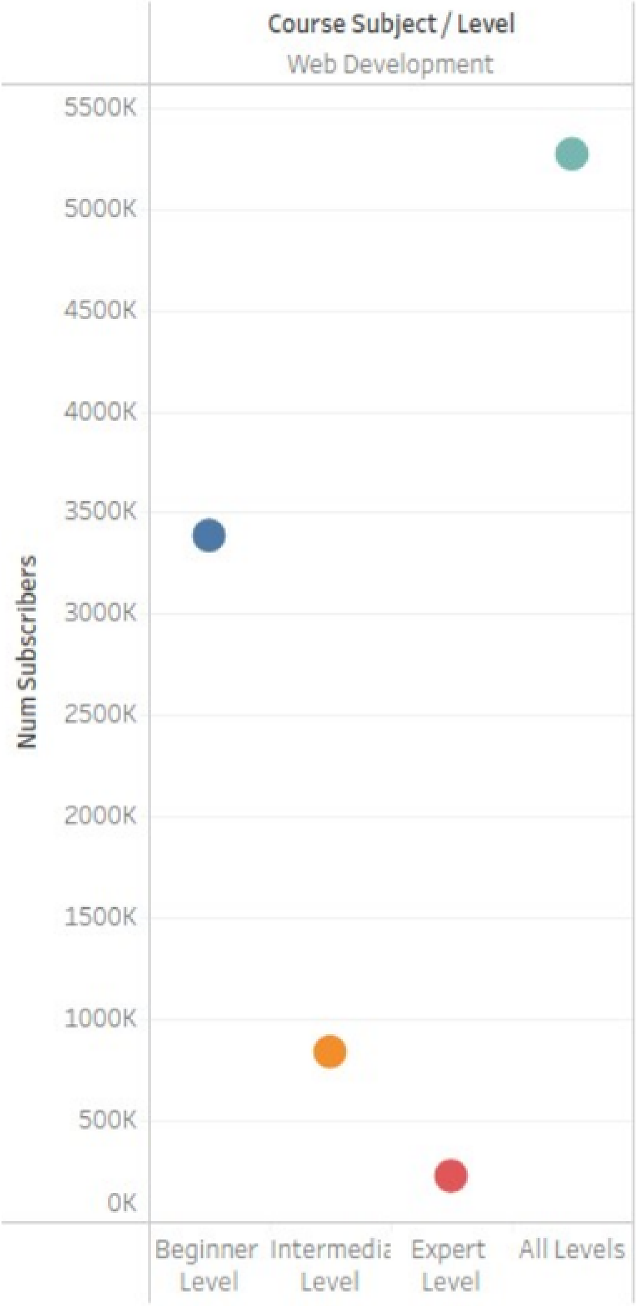
Q: Why is the majority for Web Development?
With the rising dependence on technology, it would make sense that more people are becoming more interested in coding, specifically the Internet. The popularity of the field means more competition which leads to more desire to better themselves in related topics like Web Development versus the other subjects present in the above table.

As mentioned previously, due to the rising popularity of the topic, this project will focus on the courses associated with Web Development.

This should go without saying but the wide field of topics and lessons relating to Web Development informs us that not everyone can handle the same types of topics, hence the existence of different difficulties. So why don't we compare the popularity between the ones available?

FINDINGS AND ANALYSIS

Figure 2: Subscriber Count by Web Development Course Difficulty



- The category titled "All Level" has a hefty lead over the specified levels being Beginner,
- Intermediate, and Expert as it the sum of subscribers for this level sits well over 5000K (or 5 million subscribers). Beginner has the second highest count sitting a little below 3500K (3.5 million).
- Intermediate and Expert levels however are way behind Beginner with both levels sitting under a million subscribers (Expert hasn't even hit half a million).

Q: "Why the gap between beginner and intermediate/expert"

- The difference between Beginner and Intermediate/Expert does make sense from a learner's perspective as it's a lot easier to pick up a skill with an introductory course than to hone an existing one with a harder leveled course.

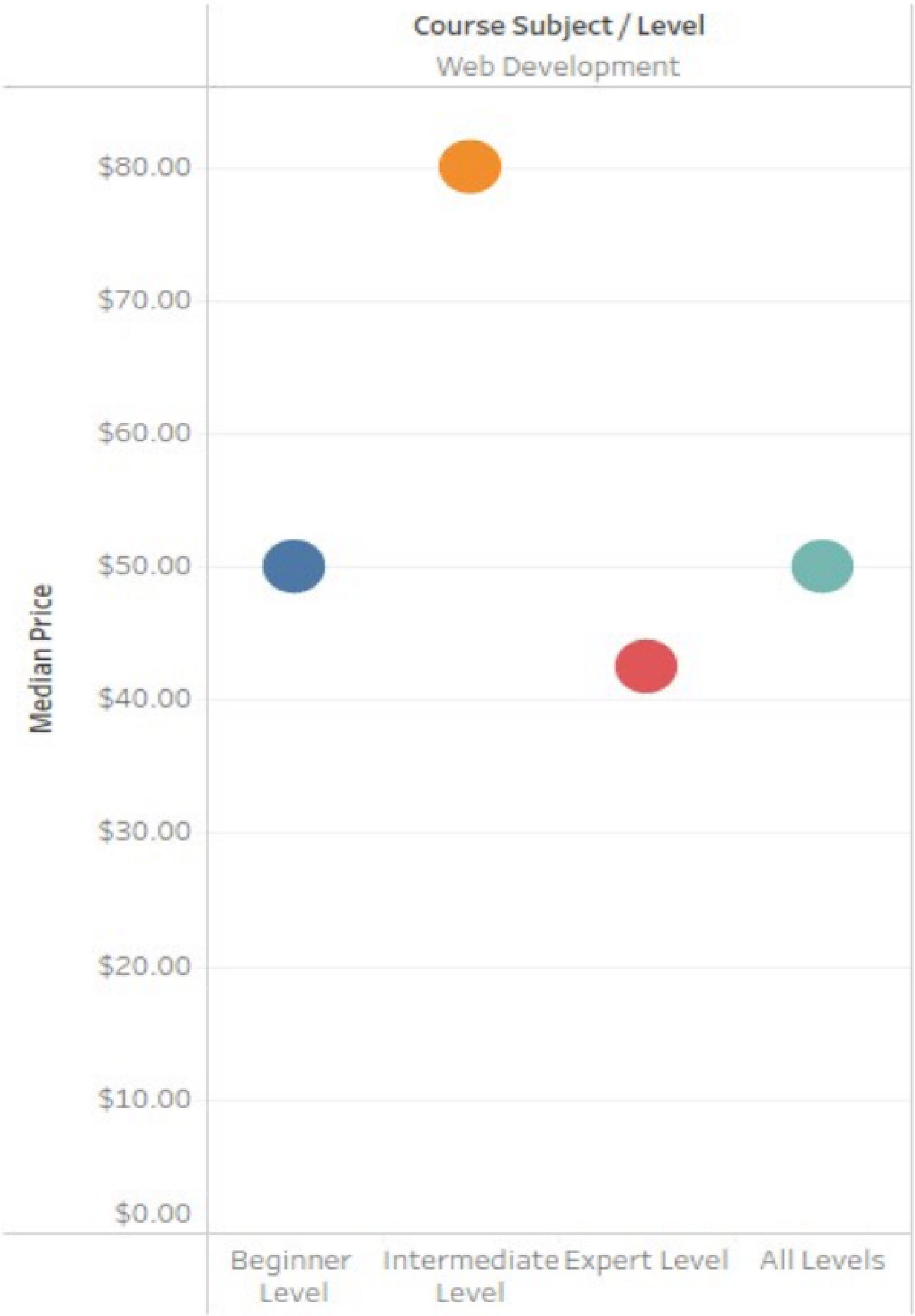
Q: Then why is there another gap between 'All Levels' and Beginner.

- In this case, courses that have something for all levels of skill would bring the attention of a lot more people than a course catering to one level..

As such, the next logical step to consider is that the level of a course is most likely an important factor to consider when in comes to pricing as a wide audience is more likely to pay a set price for courses that cater to either beginners or all levels.

FINDINGS AND ANALYSIS

Figure 3: Median Pricing of Courses Relative to Course Difficulty

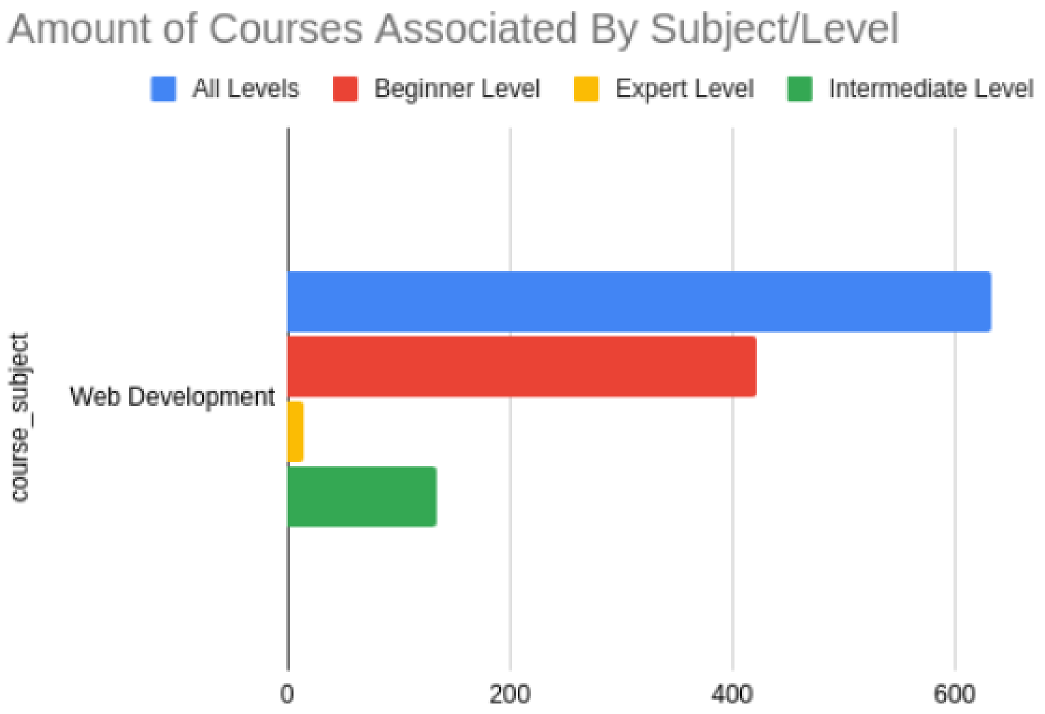


FINDINGS AND ANALYSIS

Notice in Figure 3 that the Intermediate courses (represented by the orange circle) is much higher than the other levels (a $30 jump from both Beginner and 'All Levels'!). Meanwhile Expert courses
aren't too far behind the median of Beginner/All. But why is this the case?

First, let's look at the amount of courses by level.

Figure 4: Web Development Course Amount By Level



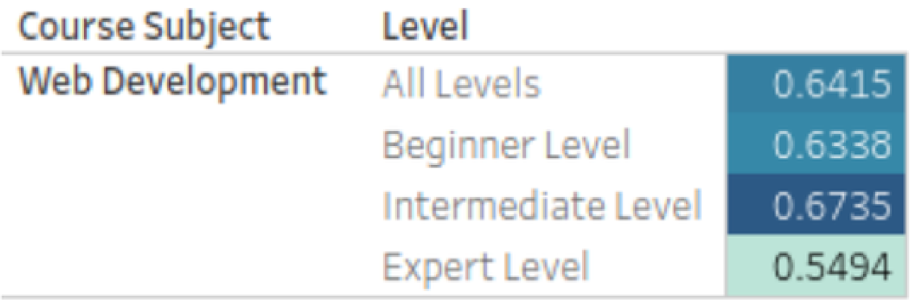
According to Figure 4, there are more than a hundred intermediate-leveled courses which tells us that it likely isn't the work of one outlier with a highly expensive course.

On the other hand, the amount of expert courses and the amount of subscribers for those courses are very low
which makes sense that it's typically cheaper than every other level.

Perhaps, the price relates to the quality of the course? Let's check the average ratings by level.

FINDINGS AND ANALYSIS

Figure 5: Web Development Ratings by Level



Course Subject	Level	
Web Development	All Levels	0.6415
	Beginner Level	0.6338
	Intermediate Level	0.6735
	Expert Level	0.5494

Here, the Intermediate leveled courses are the also hold the highest average rating out of any course level while Expert leveled courses are rated lower. However, in regards to intermediate courses, the gap between ratings is still much smaller than the gap between pricing.

Q: So why is the price of Intermediate classes much higher than the rest?
Possible reasons include:

- Higher production quality/budget which in turn makes a more expensive class
- Instructors are heavily experienced in the field so they demand higher admission fees
- The topics in the course while significant are seldomly talked about elsewhere.

Therefore, when going back to the topic of the $30 median gap from Figure 3, while it’s important to note that the ratings for intermediate level courses are the highest out of four possible categories.
However, not a lot of people are subscribed to those courses, especially when compared to easier courses. As such, the higher price gap may be considered as a deal-breaker for many.

CONCLUSION

Knowing the success and popularity of courses that are either for beginners or everyone, the price of admission should be slightly increased for those courses.

To compensate, another possible suggestion is to slighly lower the pricing of Intermediate level courses to make those types of courses more approachable to buy. While the second suggestion isn't a priority, the first suggestion should be considered for the possibility of netting a higher profit.

CAPSTONE PROJECT

CONTEXT

Airbnb is an online platform for selling and renting spaces for a set period of time. The popularity of Airbnb revolves around the convenience of renting or hosting a space in an area that might be otherwise difficult to find a spot in.

DESCRIPTION

The goal of this project is to figure out the NYC boroughs that are the most popular in Airbnb and how those boroughs are affected based on the various conditions of the area. Borough and room type are primarily compared as a basis for this project. Upon looking at Airbnb representation (host representation and customer bookings) and pricing within each borough, the main takeaway was that most Airbnb users in New York City prioritize privacy and space over pricing.

DATA DESIGN

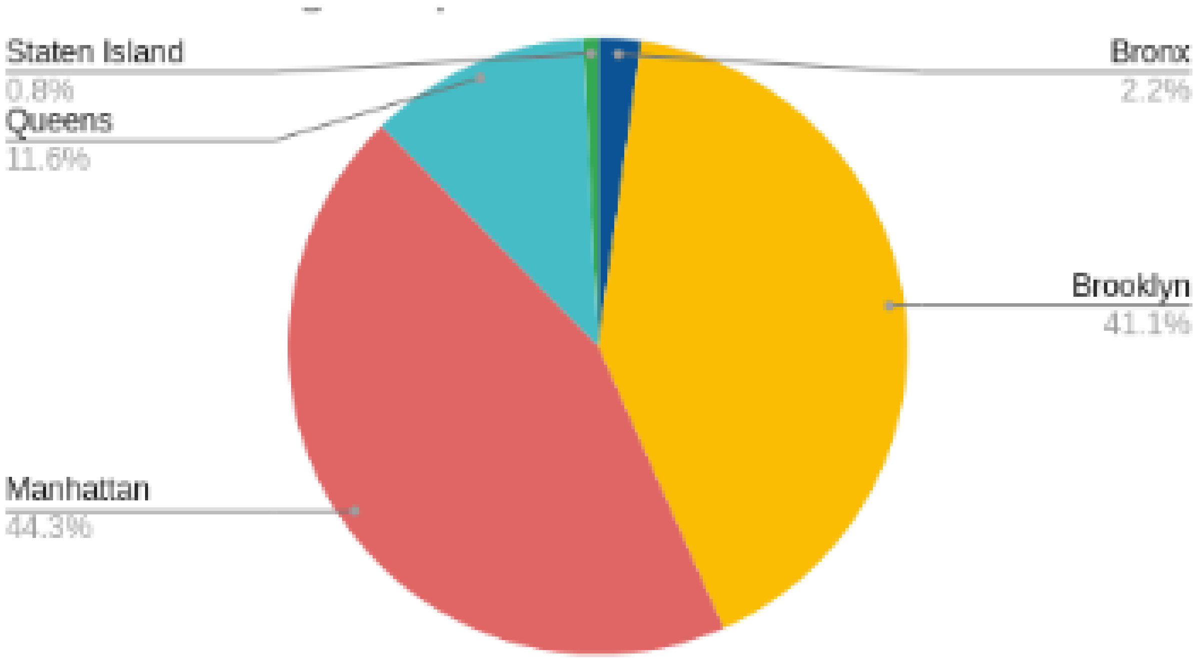
Only one dataset was used for this project so data consolidation wasn't implemented for this project.

[Source of dataset]

Both Tableau and Google Sheets were utilized for visualization purposes. This also includes the creation of charts and graphs referenced in this project.

FINDINGS AND ANALYSIS

Figure A: NYC Borough Representation by Rental Location



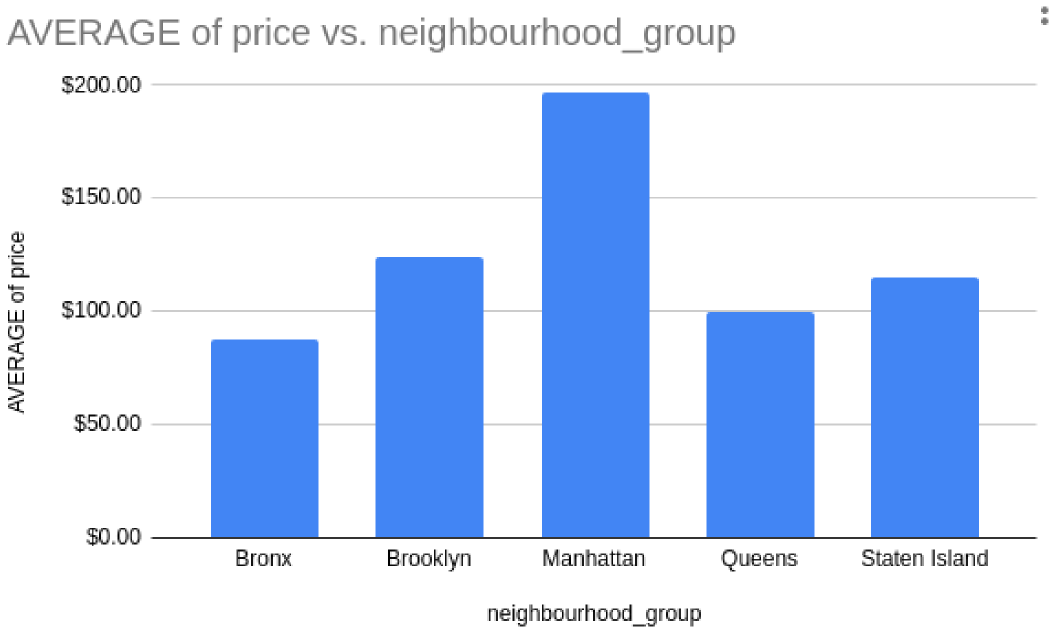
According to Figure A, Manhattan and Brooklyn take up the majority (44.3% and 41.1% respectively) with a total of 85.4% of the Airbnb rentals. Staten Island on the other hand, doesn't even possess at least 1% of the Airbnb rentals.

Manhattan being the most represented makes sense knowing that it is the most densely populated borough in NYC. Many popular places, stores, and landmarks (Times Square, Central Park, etc.) reside in Manhattan making it a popular place to host rentals, especially when considering tourists.

Q: So why is Brooklyn close to Manhattan in representation?
Despite not being as dense as Manhattan, Brooklyn still holds the highest population count out of any borough according to 2021 Census (2,641,052 vs Manhattan's 1,576,876). The amount of rentals in Brooklyn just might relate to how many people reside in that borough. Landmarks like Coney Island, and the Brooklyn Bridge, for example are likely places that draw people into Brooklyn.

It would seem that price would have a correlation with representation. However, Staten Island holds the #3 highest average pricing of rentals beating Queens and the Bronx despite not only having less available rentals than both but having the smallest population count out of the five boroughs based on the 2021 Census.

FINDINGS AND ANALYSIS



Manhattan and Brooklyn still hold the #1 and #2 spots respectively. However, Staten Island isn't too far behind Brooklyn.

Figure B: Average Airbnb Rental Pricings by Borough

Q: Why might the average rental in Staten Island be as high as it is, despite being the least populated and represented borough of NYC?

With Staten Island being a smaller population, the average price of a rental would likely be increased by any outlier would push hosts to increase the value of rentals in the borough by a considerable amount. It’s also likely that population/representation in Airbnb doesn’t correlate too much with pricing overall.

So what if instead of by boroughs, we compare by room type?

FINDINGS AND ANALYSIS

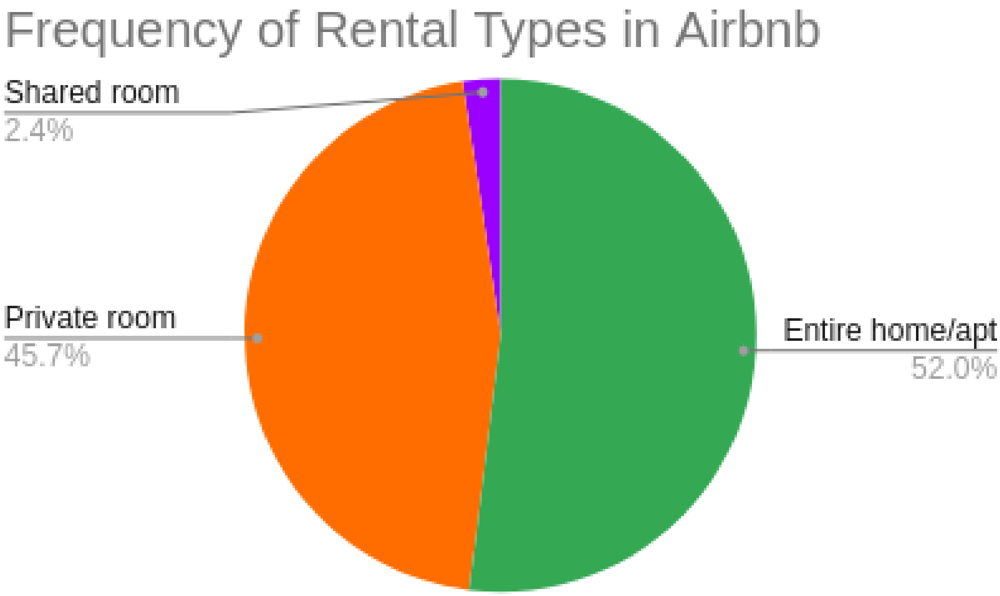


Figure C: Room/Rental Representation in NYC

Between private rooms and entire home rentals, the majority of rental representation lies from the latter with 52% which is closely followed by private room rentals at 45.7%. Shared room rentals are significantly smaller, only representing 2.4% of available rentals.

Q: Why might entire home rentals be offered the most frequently?
Offering someone an entire home for a short period of time likely leads to the most profit. Thus, it would make sense that most people would try to offer their entire house/apartment to someone when possible.

However, not everyone is capable or confident of doing so which is why offering a private room is a nearly as popular of an option.

Let’s see if the popularity of each room type and borough stays consistent for the customers.

FINDINGS AND ANALYSIS

Given the current dataset, we'll assume that each customer review represents a single customer booking. As such, we'll be looking at the total number of reviews per borough. That way, we'll be to have an idea on how often people stay at a certain type of spot.

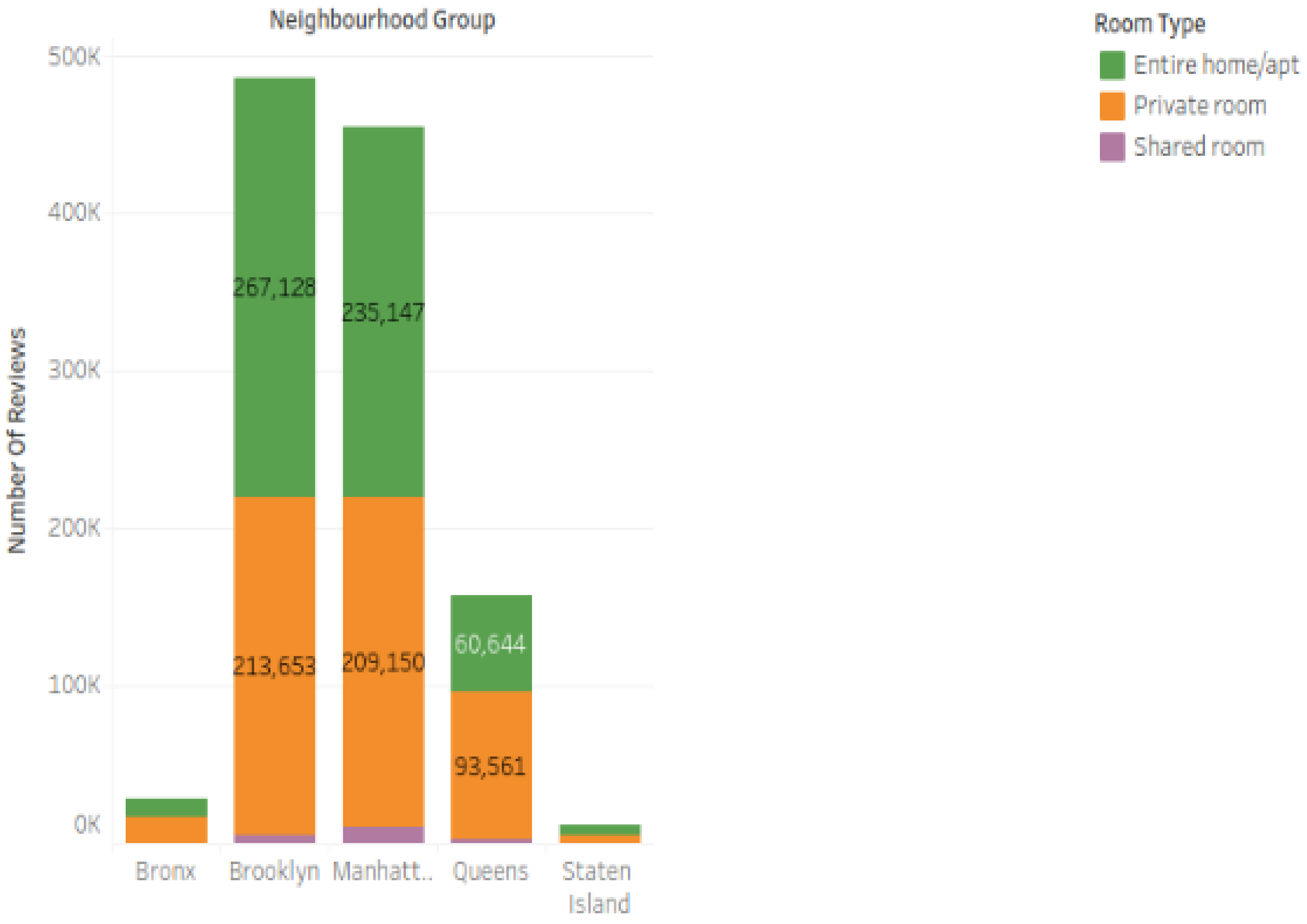
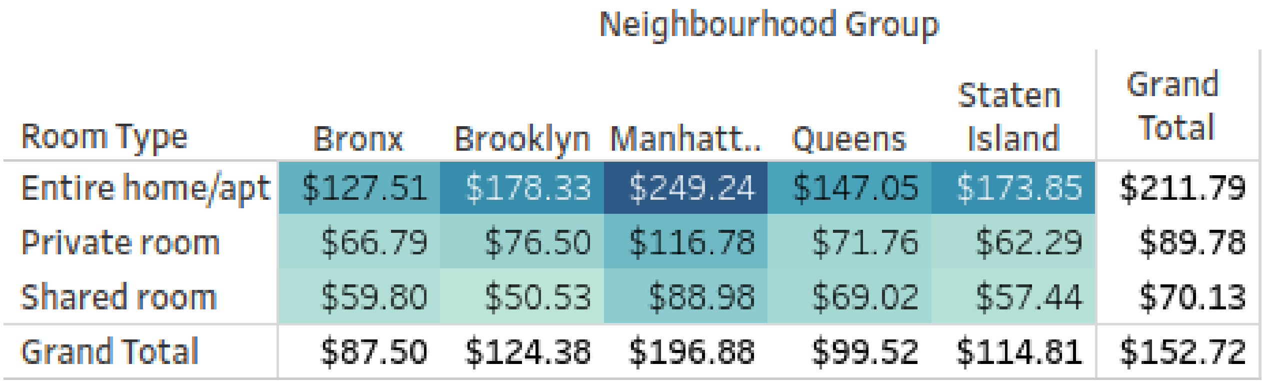


Figure D: Total Number of Customer Reviews by Borough

Ultimately, Brooklyn possesses the highest total average amount of reviews out of the five boroughs followed by Manhattan (despite having the most available rooms on Airbnb) in second, Queens in third, the Bronx in fourth, and Staten Island in last.

FINDINGS AND ANALYSIS

Even despite being a lot higher in pricing than a private room, Entire home rentals still represent more than half of the rentals in NYC.



	Neighbourhood Group					
Room Type	Bronx	Brooklyn	Manhatt..	Queens	Staten Island	Grand Total
Entire home/apt	$127.51	$178.33	$249.24	$147.05	$173.85	$211.79
Private room	$66.79	$76.50	$116.78	$71.76	$62.29	$89.78
Shared room	$59.80	$50.53	$88.98	$69.02	$57.44	$70.13
Grand Total	$87.50	$124.38	$196.88	$99.52	$114.81	$152.72

Figure E: Table Representation of Average Pricing by Room Type and Borough
Note: Grand Total represents the room type's average within all five boroughs (Not the sum)

Based on Figure E, Manhattan has the highest average pricing, at $196.88 while Brooklyn has the second highest, Staten Island has third, followed by Queens at fourth, and lastly the Bronx.

Q: Why shouldn't we assume that popularity is associated with prices?
Price alone doesn't affect a borough's popularity heavily as seen by the Bronx. Despite having the lowest average pricing out of any borough, it also has the second lowest amount of customer reviews.

Q: If Manhattan has the most available spaces, why does Brooklyn have a higher number of customers/reviews?
It's worth mentioning that Manhattan and Brooklyn are close to each other in terms of available rentals on Airbnb. It would make sense that a borough like Brooklyn would take the slight lead in terms of customer bookings. While we shouldn't assume a borough's popularity on price alone, the relation between review amount and a borough's number of rentals still has its similarities to the relation between rentals and average pricing.

CONCLUSION

Key Points:

- Despite its typically high pricing, Manhattan remains at least in the top 2 boroughs when it comes to Airbnb host representation and popularity with customers.
- A borough having a higher amount of hosts correlates with the total amount of customer reviews/bookings.
- (Average) pricing has some correlation between both hostings and customer bookings/reviews
- Most people are willing to rent an entire home despite the higher pricing and the thought of sharing is undesired for many
- Based on Brooklyn exceeding Manhattan in customer reviews, it’s a possibility that other factors like outdoor activities, crime rates, traffic, and other surroundings are to be considered when it comes to a borough’s popularity.
- **Privacy and convienience are significant and pricing to a lesser degree for most users**