# Data Methodology

Step 1: Intro

- Inspected data for all columns data types and values that they are holding to understand the significant columns
- Prepared a raw blueprint of slides on rough notebook
- Prepared a slideshow based on blueprint prepared

Step 2: Data Wrangling

- We chose univariate analysis using Python on Google Collaboratory on the fields to see their distributions, the unique values in a field, the missing values and to check for outliers if any
- Very less number of null values are present in the dataset and will not affect our analysis. These are ignored.
- Price was highly positively skewed so median was very close the lower quartile with some outliers as seen in the boxplot below

Price for each Room Type.
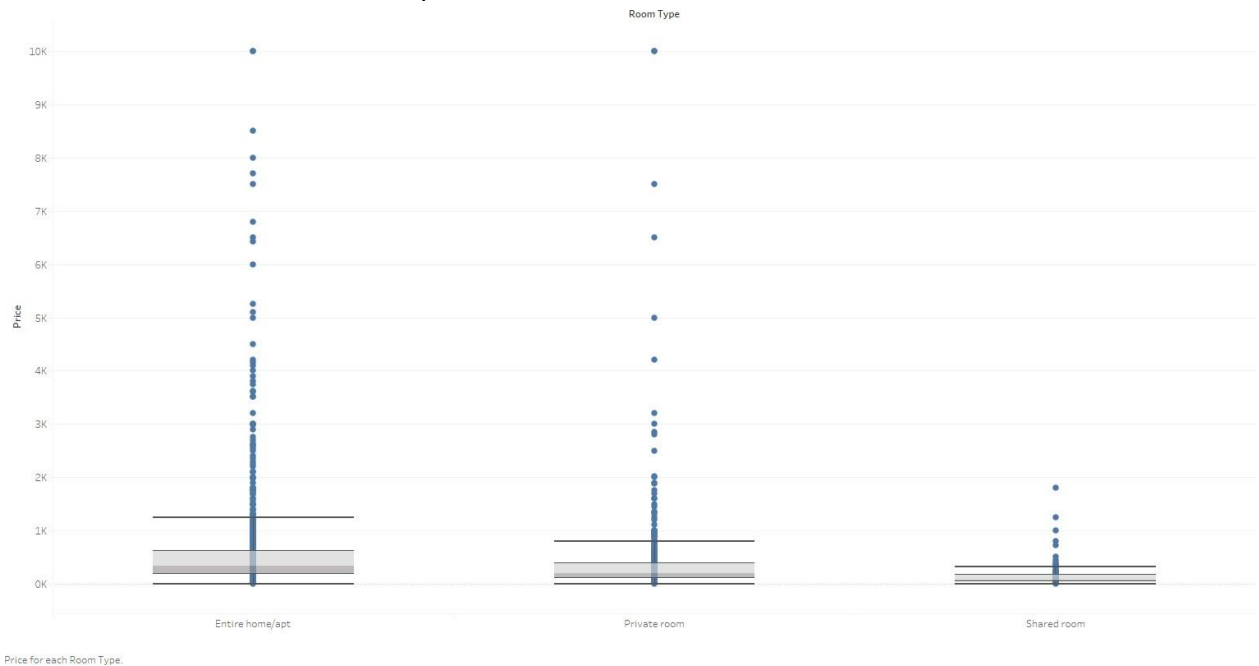
- Since price has outliers, used median instead of mean as the measure for price
- Host Listings count is maximum for entire apartment and private room and is very small for shared room.

- Created a grouped field for Minimum Number of Days assuming null values belonged to the category

```
IF [Minimum Nights]=1 THEN "1"
ELSEIF [Minimum Nights]=2 THEN "2"
ELSEIF [Minimum Nights]=3 THEN "3"
ELSEIF 4<=[Minimum Nights] AND [Minimum Nights]<=5 THEN "4-5"
ELSEIF 6<=[Minimum Nights] AND [Minimum Nights]<=7 THEN "6-7"
ELSEIF 8<=[Minimum Nights] AND [Minimum Nights]<=29 THEN "8-29"
ELSEIF 30<=[Minimum Nights] AND [Minimum Nights]<=31 THEN "30-31"
ELSE ">31" END
```

- Created a calculated field of number of reviews per listing

```
SUM([Number Of Reviews])/COUNT([Calculated Host Listings Count])
```

Step 3: Data Analysis

- Checked neighborhood grouped wise distribution of price and room type
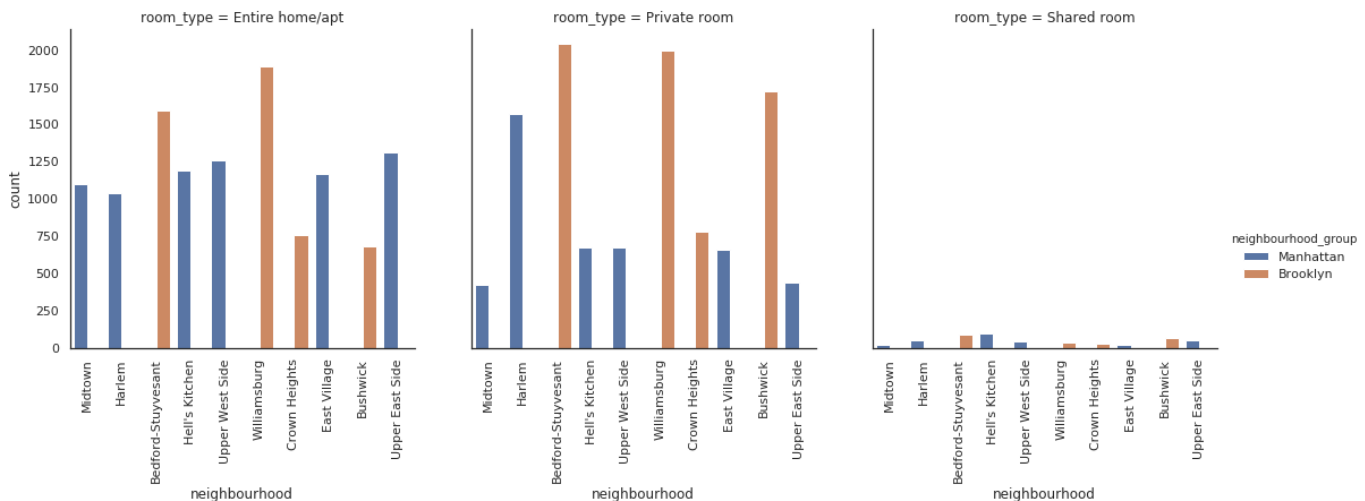
- Through these clearly:
  - Except in Manhattan, shared rooms are very unpopular
  - Entire apartments are popular and so are private rooms
  - Entire room's reviews per listing for Manhattan is 35% lower than the overall average and number of listings is the highest

Density and distribution of prices for each neighberhood_group
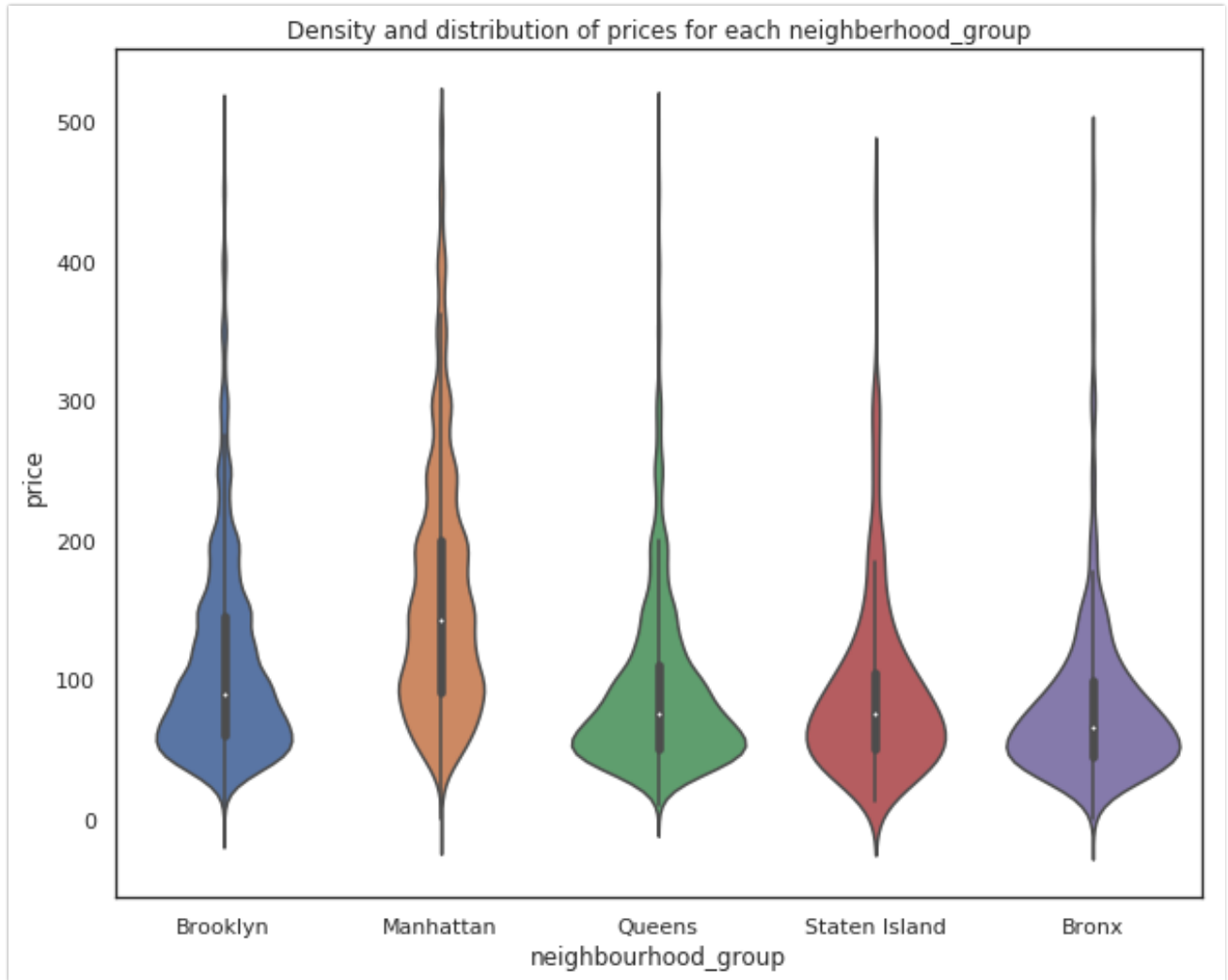
- Manhattan is costliest overall and the cheapest are:
  - Entire apt: Bronx o Private room: Staten Islands o
  Shared room: Brooklyn
- Found difference in prices between shared rooms, private rooms and entire apartments for each neighborhood as shown below

	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2			Bronx	Brooklyn	Manhattan	Queens	Staten Island		Bronx	Brooklyn	Manhattan	Queens	Staten Island
3			127.5	178.3	249.2	147.1	173.8						
4			66.8	76.5	116.8	71.8	62.3		91%	133%	113%	105%	179%
5			59.8	50.5	89	69	57.4		12%	51%	31%	4%	9%
6													
7													

I4 f_x =C3/C4-1

- Through this is clear, we can say that:

  - Entire home/apartment's price is ~100% more than private rooms except Brooklyn and Staten Island with ~150%

- Private room's price is ~10% more than shared rooms except Brooklyn and Manhattan with ~40%

Step 4: Presentation

- Made the presentation adhering to best practices and pyramid principle
- Added recommendations for the respective departments