Boston Airbnb

Tiago Santos
2019.08.29

Introduction

Airbnb is an online marketplace for arranging homestays, acting as a broker. The seller, known as a "host" posts a listing on Airbnb's website with an offer, and customers make a selection from all the offers posted. This analysis is based on official data provided by Airbnb for the city of Boston from the years 2008 (the year in which Airbnb began operation) to 2019 (the current year).

Objectives:

There are 3 main areas we would like to study

Spatial data analysis:

- How location ratings change across the neighbourhoods?
- How is the Super Host distributed across the neighbourhoods?

Price and supply analysis:

- What is the average price of Airbnb per year, and across all the years available?
- Is there a relationship between supply and price?
- Similarly, is there a relationship considering the different room types available?
- Is it possible to stay at a Super host for under $100 per night?

Demand analysis:

- What is the total demand for Airbnb per year, and across all the years available?
- Is there a relationship between supply, price, and demand?
- Are there any trends in the demand for Airbnb?

Dataset information:

'listings.csv': This dataset has 6247 rows of data, each with 106 attributes

- Key attributes used in this analysis include: 'price' (continuous), 'Longitude' (continuous), 'Latitude' (continuous), 'host_is_superhost' (categorical), 'neighbourhood_cleansed' (categorical), 'review_scores_location' (continuous), 'host_id' (continuous), 'property_type' (categorical)

'reviews.csv': This dataset has 199,106 rows of data, each with 6 attributes

- Key attributes used in this analysis include: 'date' (continuous), and 'id' (continuous)

Tools and packages used:

- library(dplyr)
- library(ggplot2)
- require(devtools)
- library(devtools)
- library(lubridate)
- library("csv")
- library(grid)
- library(tidyr)
- library(scales)
- library("ggmap")

Prepare Data:

Price (listings): The price column contained data in string format with the currency symbol '$' and comma separator ',' attached to it. This column was manipulated to contain double values for analysis.

```r
airbnbl$price <- gsub("\\$", "", airbnbl$price)
airbnbl$price <-as.numeric(gsub(",","",airbnbl$price,fixed=TRUE))
```

```r
typeof(airbnbl$price)
is.numeric(airbnbl$price)
head(airbnbl$price)
```

```
 [1] "double"
 [1] TRUE
 [1] 125 145 169  65  99 154
```

Date (listings and reviews): The date was contained in yyyy-mm-dd format but as a string. It was transformed to R date format, so the date can be manipulated and Analysed later.

```r
airbnbl$host_since <- as.Date(airbnbl$host_since)
is.Date(airbnbl$host_since)
head(airbnbl$host_since)

airbnbr$date = as.Date(airbnbr$date)
is.Date(airbnbr$date)
head(airbnbr$date)
```

```
 [1] TRUE
 [1] "2008-12-03" "2009-02-19" "2009-02-19" "2009-05-11" "2009-07-22" "2009-07-22"
 [1] TRUE
 [1] "2015-07-10" "2015-08-09" "2015-09-01" "2015-09-30" "2015-10-30" "2016-02-22"
```

Dealing with Missing Values

Evaluate the columns that we will use in the exploratory analysis

```r
airbnbl %>% select(host_is_superhost,host_id,review_scores_rating,review_scores_location,host_name,host_since,price,neighbourhood_cleansed,property_type,latitude,longitude) %>%
sapply(function(x){sum(is.na(x))})
```

```
     host_is_superhost                host_id   review_scores_rating review_scores_location
                     0                      0                   1312                   1321
             host_name             host_since                  price neighbourhood_cleansed
                     0                      0                      0                      0
         property_type               latitude              longitude
                     0                      0                      0
```

```r
colSums(is.na(airbnbr))
```

```
    listing_id             id           date    reviewer_id  reviewer_name       comments
             0              0              0              0              0              3
```

- (listings and reviews) Most of the rows have no missing values.
- (listings) Review_scores_rating and review_scores_location have 1312 missing values what represents 21% of the data set. As a large fraction of observations and there is no way to fill the reviews because is a personal opinion of the customers, we are going to exclude these rows just for Spatial data analysis.
- (reviews) comments has only 3 rows with missing data, there is no way to fill the comments because is a personal opinion of the customers, we are going to exclude these rows.

```r
airbnbr = airbnbr %>% filter(!is.na(comments))
colSums(is.na(airbnbr))
```

```
    listing_id             id           date    reviewer_id  reviewer_name       comments
             0              0              0              0              0              0
```

additionally, we create new data frame "airbnbrNo2019" that removes all 2019 values

- we do this because 2019 is incomplete data
- from summary(airbnbr), we see that 2019-01-17 is the max value for date

```r
airbnbrNo2019 = airbnbr %>% mutate(year = year(date)) %>% filter(year < 2019) %>% arrange(desc(year))
tail(airbnbrNo2019)
```

Creating new data frames that count the number of reviews by year, and by month

```r
airbnbrNumByYear = airbnbrNo2019 %>% mutate(year = year(date)) %>% group_by(year) %>%
summarise(numberOfReviews = n()) %>% mutate(diff = numberOfReviews - lag(numberOfReviews))

airbnbrNumByYear
#airbnbrNumByYear is a data frame that shows the number of reviews in each year

airbnbrNumByMonth = airbnbr %>% mutate(year = year(date)) %>% mutate(month = month(date)) %>%
group_by(year, month) %>% summarise(numberOfReviews = n()) %>% unite("Year_Month", year, month)
airbnbrNumByMonth
#airbnbrNumByMonth is a data frame that shows the number of reviews in each month

airbnbrNumByDay = airbnbr %>% mutate(year = year(date)) %>% mutate(month = month(date)) %>%
mutate(day = day(date)) %>% group_by(year, month, day) %>% summarise(numberOfReviews = n()) %>%
unite("Year_Month_Day", year, month, day, sep = "-")
#airbnbrNumByDay is a data frame that shows the nubmer of reviews in each day

airbnbrNumByDay$Year_Month_Day = as.Date(airbnbrNumByDay$Year_Month_Day)
#converting the "Year_Month_Day" from characters to date format, so we can manipulate it later

airbnbrNumByMonth

airbnbrNumByDay
tail(airbnbrNumByDay)
```

Spatial data analysis

How location ratings change across the neighbourhoods?

This section will explore review scores rating from our dataset using spatial visualizations and will answer questions relating to changing in ratings across different locations in Boston.

Airbnb users (customers) rate their stay on the basis of location, cleanliness and other parameters. Here we work with the review scores location data that range from 1 to 10. We use the location scores as an indicator of the appeal of the neighbourhood. We predict that highly rated and concentrated neighbourhoods will tend to have better connectivity (public transportation like subway stations and bus route), it will tend to be closer to the city tourist places or business locations.

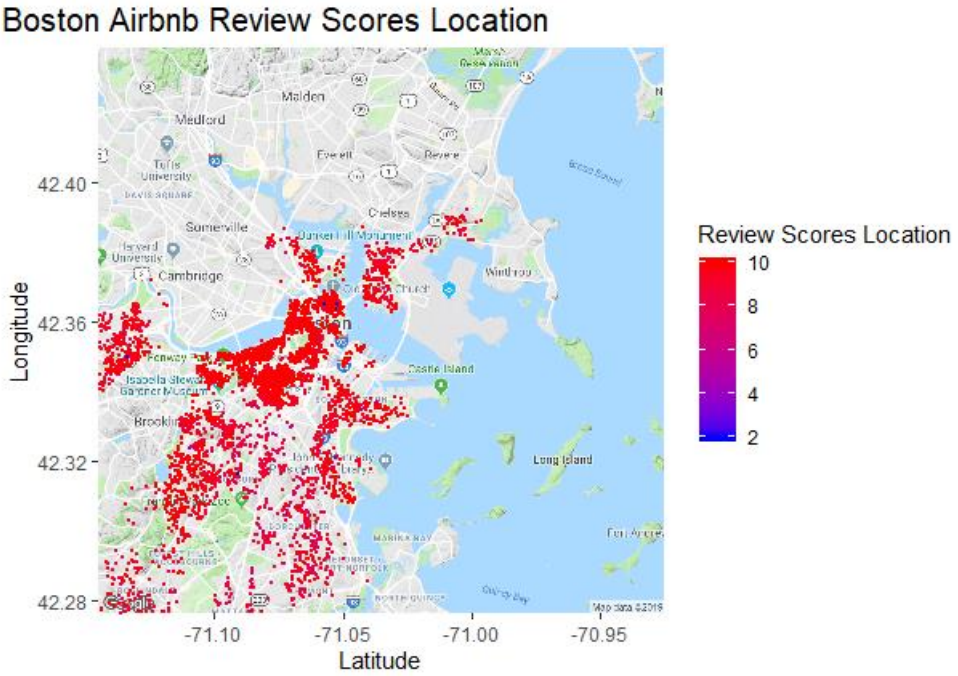


Figure 1

The graph confirms our premise that the concentration of the review locations is near tourist areas and business locations. As the Boston Common and Boston Public Garden in Beacon Hill, Boylston St in Back Bay, Old North Church and St. Stephen`s Catholic Church in North End, Faneuil Hall Marketplace and New England Aquarium in Financial District.

How is the Super Host distributed across the neighbourhoods?

Airbnb wards the title of “Super host” to a small fraction of its dependable hosts. This is designed as an incentive program that is a win-win for both the host, Airbnb, and their customers. The super host gets more business in the form of higher bookings, the customer gets improved service and Airbnb gets happy satisfied customers.

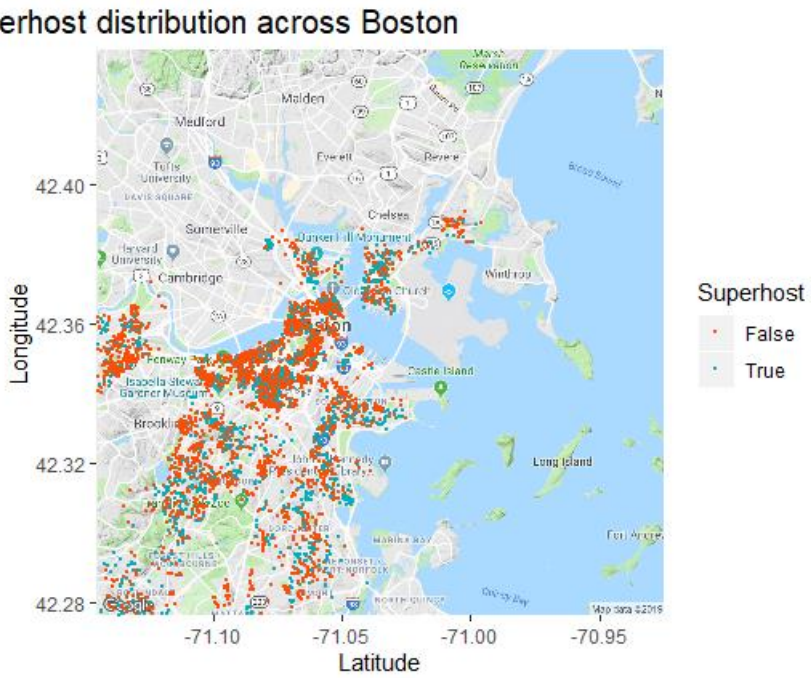


Figure 2

From the plot, we can see that super hosts are spread across all the neighbourhoods of Boston and not concentrated in one area.

Price and supply analysis

Distribution of Property type in Neighbourhoods

Through visualization, we want to study the relationship between property type and neighbourhood. The primary question we aim to answer is whether different neighbourhoods constitute of different rental property types. Though there are more than 17 types, we will be focussing on the top 5 by their total Listings in the city and understanding their distribution in each neighbourhood.

We begin by identifying which property type in Boston has the most listings.

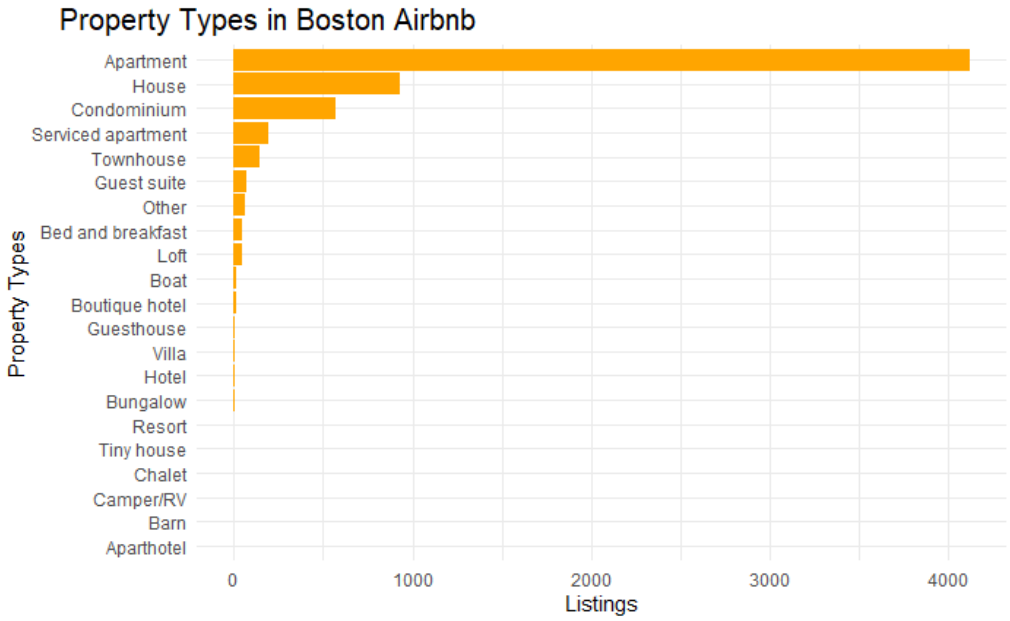


Figure 3

The preceding graph shows that the most popular listing supplied is apartments, followed remotely by Houses and Condominiums, Serviced apartments, and Townhouses. We present another graph to show the distribution of property types by neighbourhoods.

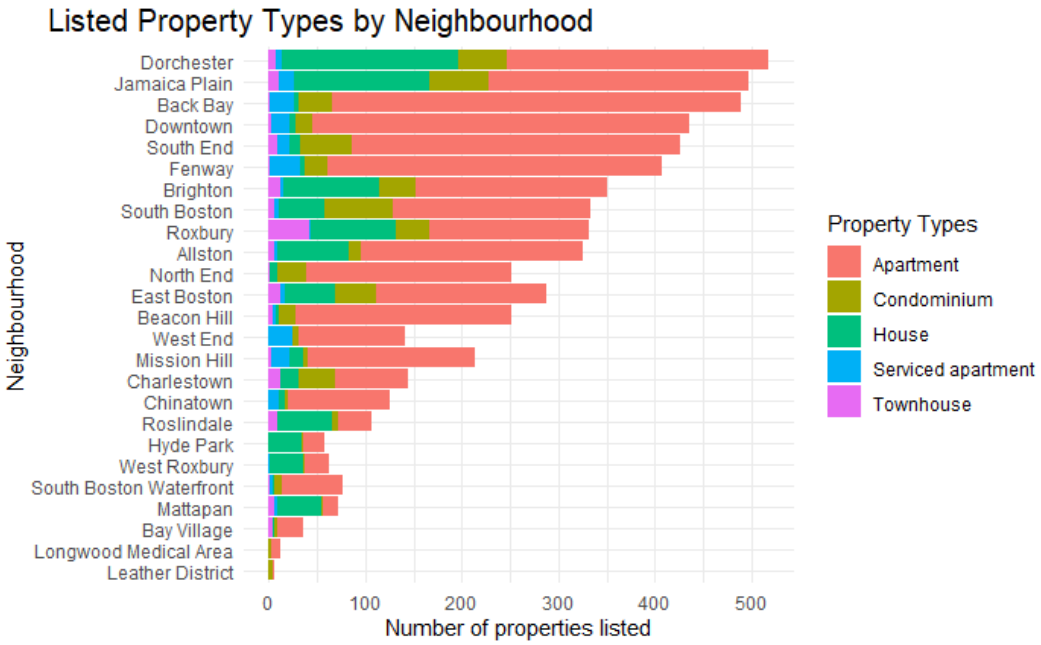
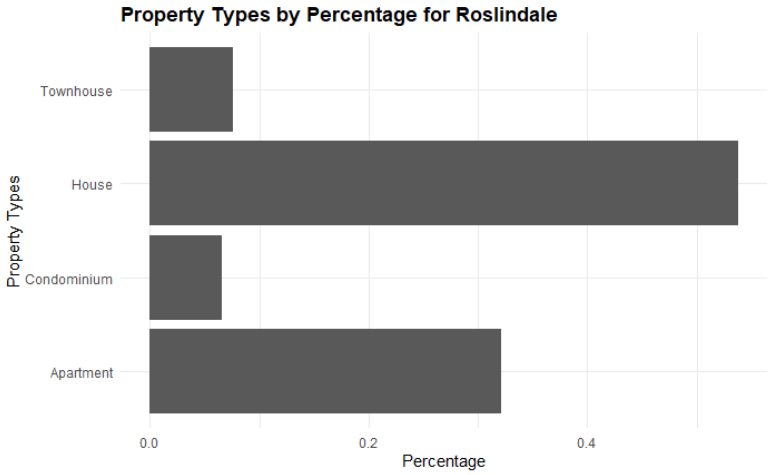


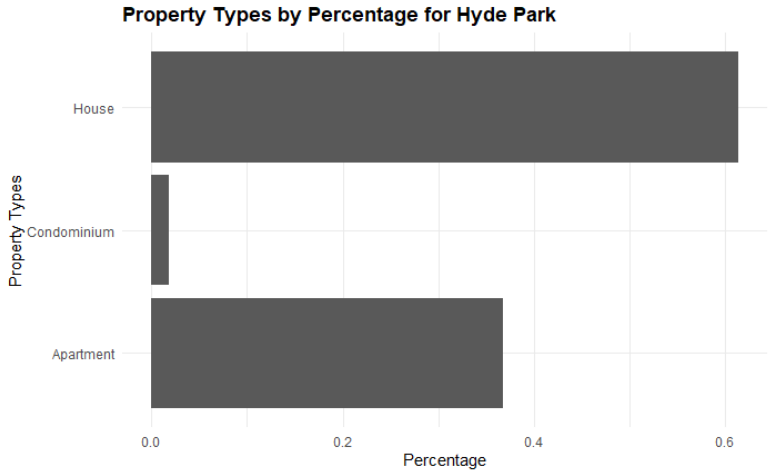
Figure 4

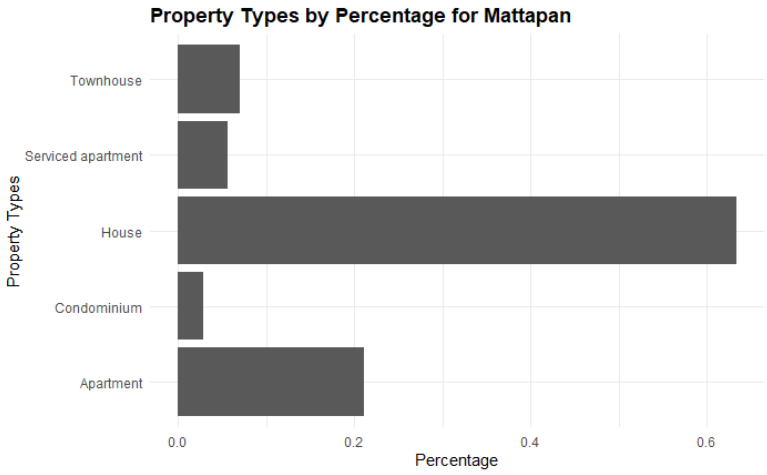
There are several key takeaways from the graph:

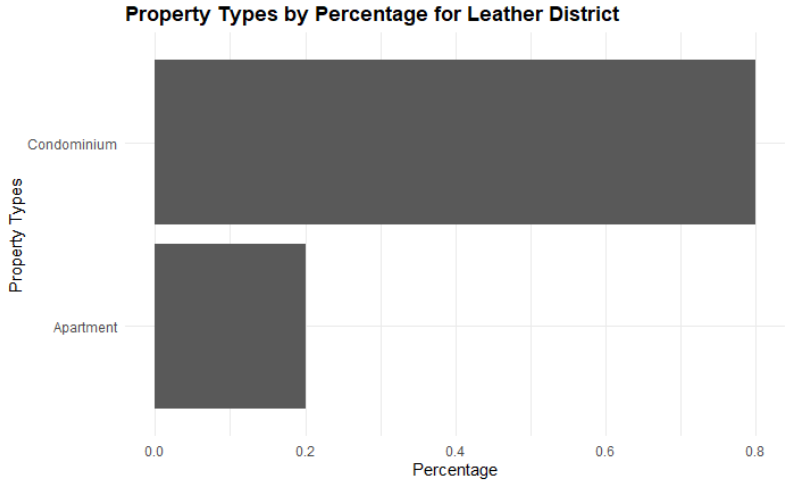
- The Apartment style listings are highest in number for most of the neighbourhoods except Roslindale, Hyde Park, West Roxbury, Mattapan and Leather District.
- The maximum apartment style listings are located in Back Bay with 424 properties in that neighbourhood in comparison with the other neighbourhoods. Next is Downtown with 390 properties followed by Fenway and South End with 348 and 340 properties respectively.
- Roslindale and Mattapan have a lot more House style listings than apartment, with 57 and 45 properties in those neighbourhood and apartment style with 34 and 15 properties respectively.
- Condominium style listings are common in South Boston with 71 properties in that neighbourhood in comparison with the other neighbourhoods. Next is Jamaica Plain with 62 properties and South End with 54 properties.
- Serviced apartment style listings are common in Fenway with 32 properties in that neighbourhood in comparison with the other neighbourhoods. Next are Back Bay and West End with 25 properties each.
- Townhouse style listings are common in Roxbury with 42 properties in that neighbourhood in comparison with the other neighbourhoods. Next are Brighton, Charlestown and East Boston with 12 properties each.

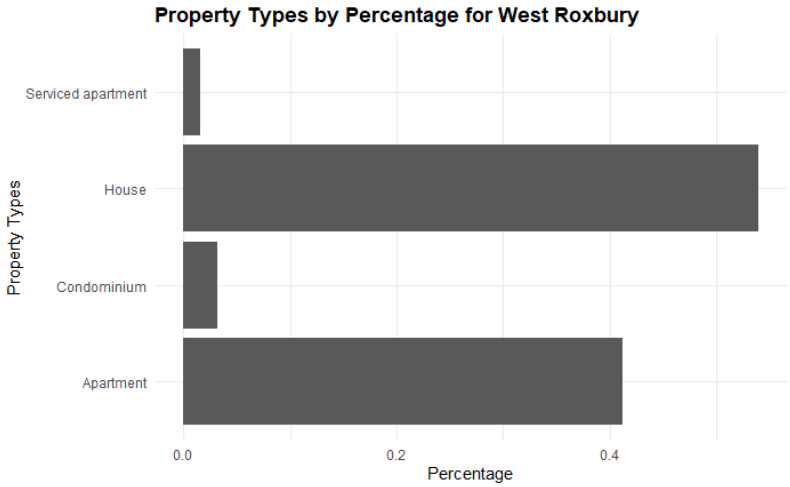
While analysing the plot, we wish to better understand the relevance of property types in the Roslindale, Hyde Park, West Roxbury, Mattapan and Leather District because it wasn't clear on the plot.


Property Types by Percentage for Roslindale
Townhouse
House
Condominium
Apartment
Property Types
0.0
0.2
0.4
Percentage


Property Types by Percentage for Hyde Park
House
Condominium
Apartment
Property Types
0.0
0.2
0.4
0.6
Percentage


Property Types by Percentage for Mattapan
Townhouse
Serviced apartment
House
Condominium
Apartment
Property Types
0.0
0.2
0.4
0.6
Percentage


Property Types by Percentage for Leather District
Condominium
Apartment
Property Types
0.0
0.2
0.4
0.6
0.8
Percentage


Property Types by Percentage for West Roxbury
Serviced apartment
House
Condominium
Apartment
Property Types
0.0
0.2
0.4
Percentage

Supply and price analysis:

We are interested in the supply and price of Airbnb listings throughout the years 2008 to 2019. In our analysis, we interpret the number of listings as an indicator of supply. This is because a listing on Airbnb corresponds to a housing type supplied to the market.

What is the average price of Airbnb per year, and across all the years available?

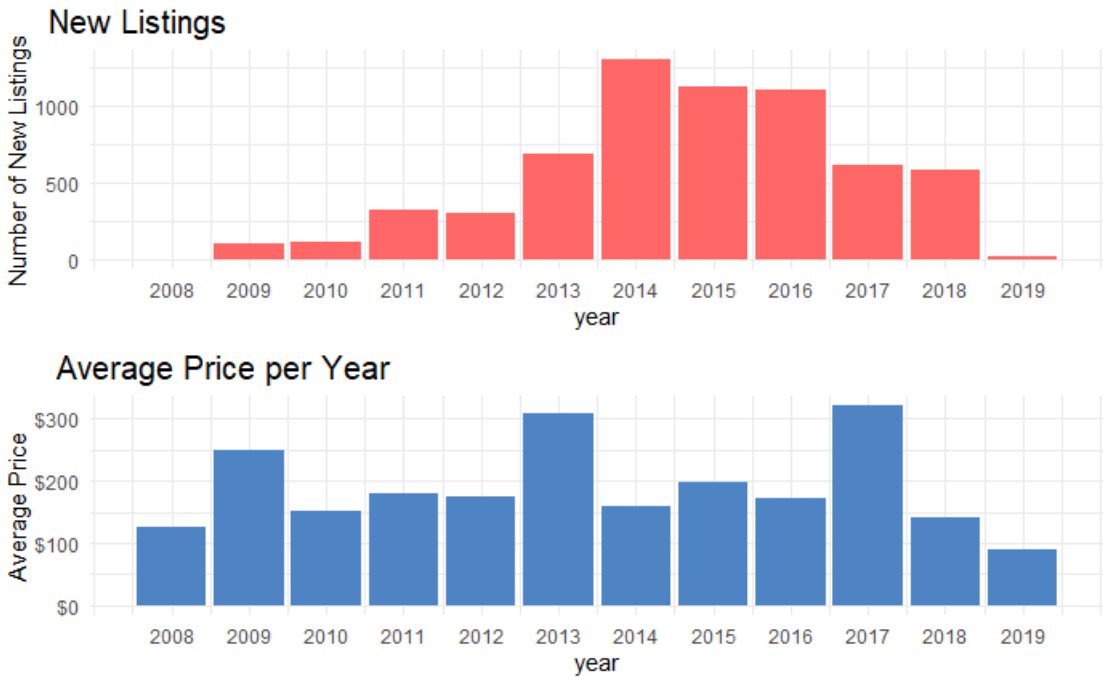


Figure 5

From the graphs, we see the following key takeaways

Analysis of supply of Airbnb listings:

- The supply of Airbnb in Boston rises steadily and consistently, until 2012, when the supply increases exponentially and peaks at 2014.
- From 2014 onwards the supply follows a consistent and linear-like decline.
- Note that the 2019 year has data only for the month of January, which explains the seemingly small supply.

Analysis of the price of Airbnb listings:

- The average price of Airbnb listings generally stays within the range of $100 to $200, with the exceptions of several peaks, where prices rise to above $200.
- The peaks in price are seen at the years 2009, 2013, and 2017. There seems to be consistent 4-year intervals between each price peak.

What is the average price of Airbnb for Entire Home apartments?

Airbnb primarily provides three types of properties, “Entire homes”, “private rooms”, and “shared rooms”. We wish to analyse the average price per year of each type of property. Since shared rooms constitute less than 2% of all room types, with only 92 points of data, the data set available is too small to make any meaningful conclusions. Thus, we disregard the shared rooms from our analysis.

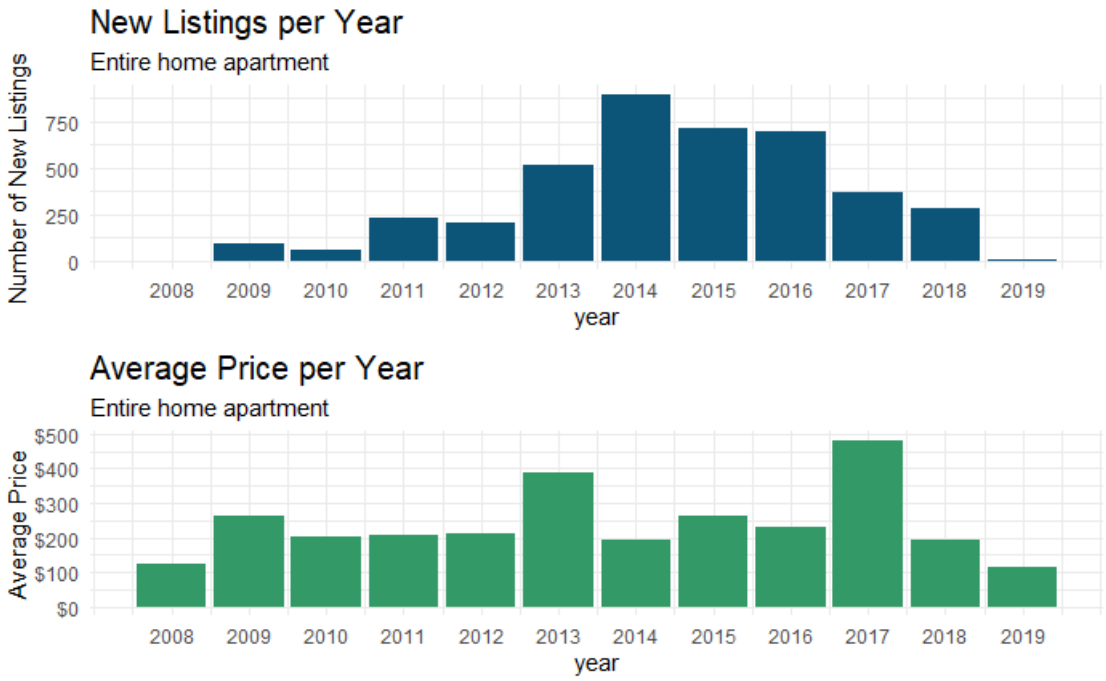


Figure 6

Analysis of supply of Airbnb Entire Home Apartment type:

- The New Listings has the same comportment as the all listings.

Analysis of the price of Airbnb Entire Home Apartment type:

- Similar to the average price of all listings, the average price of entire homes peaks every four years, with peaks occurring at 2009, 2013, and 2017.
- The range for the average price is generally between $150 and $250.
- The mean of the average prices across all years for Entire Home Apartment is greater than that of all Airbnb listings and that of private rooms. This means that Entire homes tend to cost more than private rooms, and bring up the price of all listings.

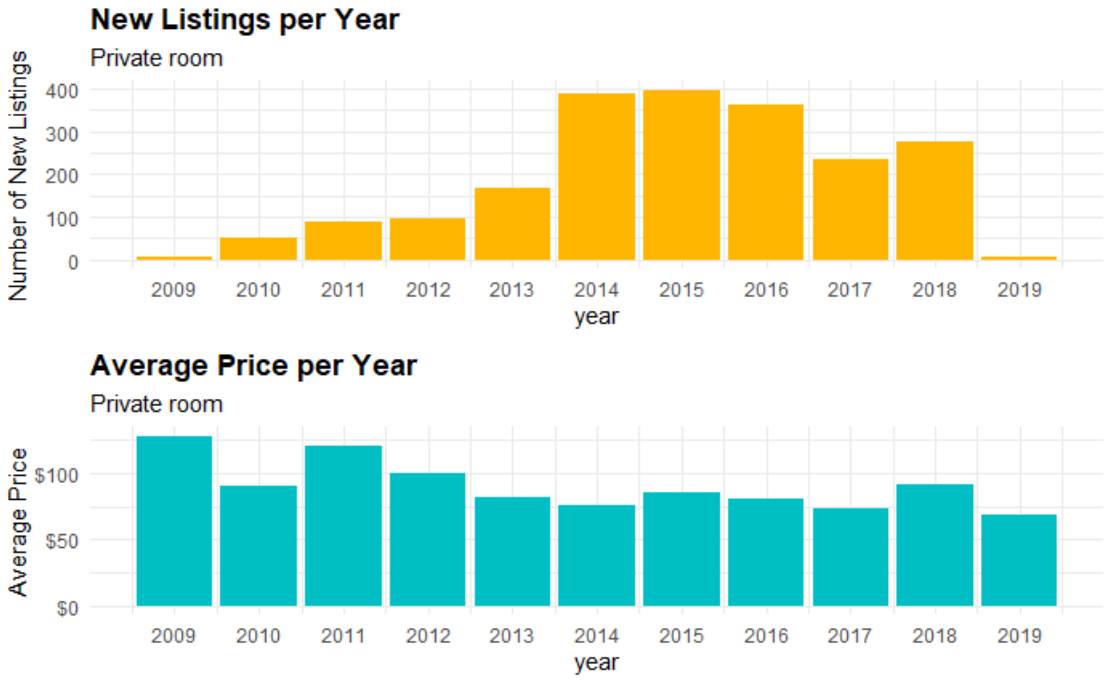


Figure 7

Analysis of supply of Airbnb Private room:

- The New Listings has a different compartment of the all listings. The peak occurs in 2015 and the price decrease until 2017 and increase in 2018.

Analysis of the price of Airbnb Private room:

- Unlike the average price per year for all listings, and for Entire Home apartments, the average price per year for private rooms does not follow a trend in price peaks every 4 years.
- The highest peak occurred during 2009, which was Airbnb's first year on the market.
- The mean of all average prices per year for private rooms is less than that of all apartments and Entire Home apartments. This means that private rooms tend to cost less than Entire homes, and bring down the average price of all listings.

Is it possible to stay at a Super host for under $100 per night?

The following analysis explores the price per night less than $100 of a stay at a super host in Boston.

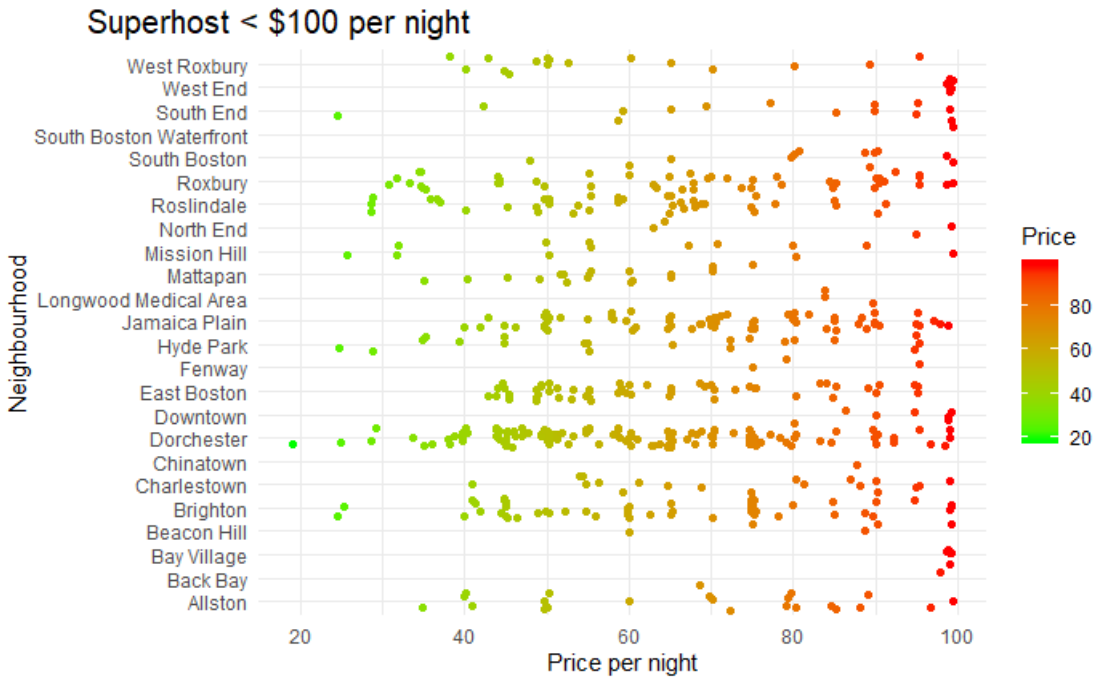


Figure 8

Analysis of the Super Host

- From the graph we can see that the range of prices starts at $19 per night.
- The prices vary by neighbourhood, with the neighbourhood of Dorchester having the greatest distribution. An explanation for his spread distribution in prices can be found in our analysis of *Figure 4*, which shows that Dorchester supplies the most varied property types out of all the neighbourhoods.
- Similarly, *Figure 4*. Indicates that the neighbourhood of Bay back supplies a very small variety of property types, with most of them being apartments. In *Figure 8.* we see that Bay back also has a very small prices distribution, with most of the price being in the upper range.

Demand analysis

Demand trend throughout 2009 to 2018

The following graph plots the date from 2009 to 2018 on the x axis, and the number of unique listings with reviews on the y axis. In our analysis we use the number of unique listings with reviews as an indication of demand, as it is implied that a reviewer has bought a listing from Airbnb.

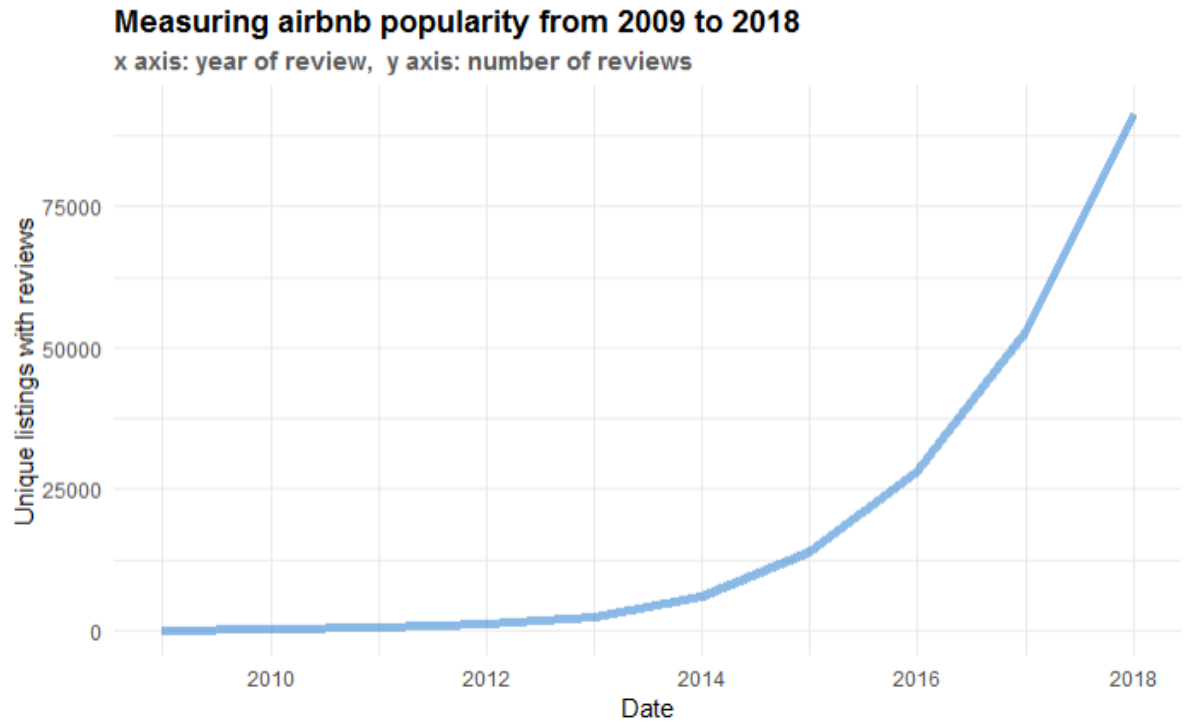


Figure 9

Analysis of the graph

- The graph shows a slow and linear increase between the years 2009 and 2012.
- From 2012 onwards, the demand for Airbnb increases dramatically.

Why does the popularity of Airbnb increase dramatically from 2012 onwards?

- The classical explanation is demand rises when price falls. However according to *Figure 5.*, *Figure 6.*, and *Figure 7.*, there is no exponential decrease in price from the year 2012.
- Thus, the explanation could be that demand increases as the popularity of Airbnb increased through publicity and word of mouth.

Demand trend throughout 2009 to 2018, by day

The following graph is similar to *Figure 9.* except the x axis is plotted with the days of the year.

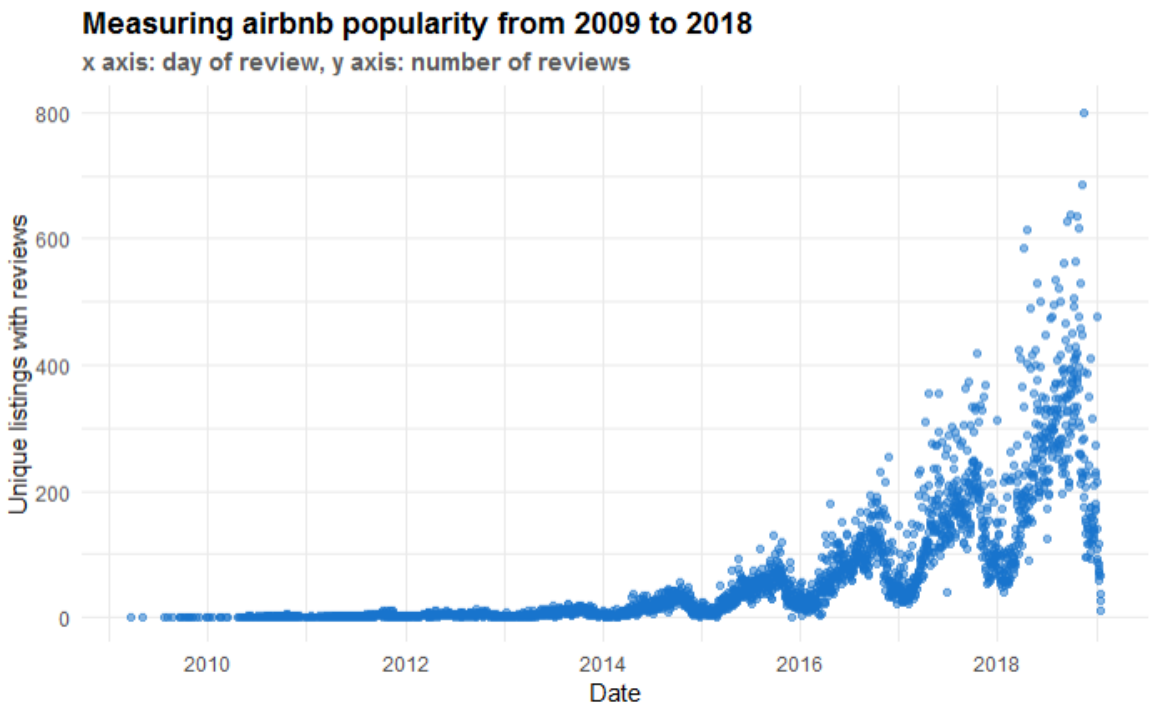


Figure 10

Analysis of the graph

- The graph shows the same increase in demand exhibited in *Figure 9*.
- Additionally, it indicates wave-like function behaviour beginning from 2014. This can be interpreted as seasonal demand, as the demand for Airbnb may rise during some months and fall during others.

A closer look at seasonal demand

The following graph plots the days of 2016 on the x axis, with listings with reviews on the y axis.

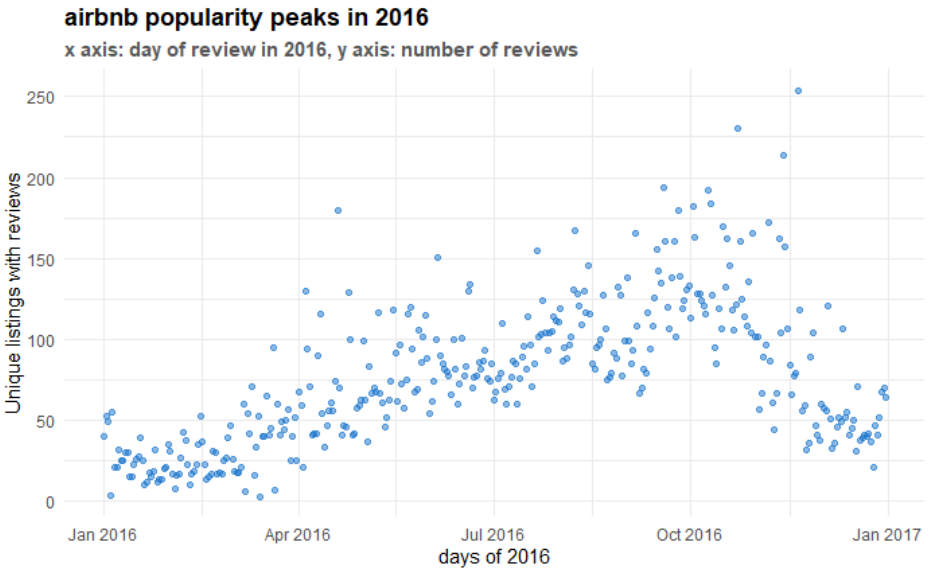


Figure 11

Analysis of the graph

- The graphs for the all years since 2014 (not shown) follow the same demand trend throughout the year.
- The demand trend follows what we expect for travel seasonality, with the summer holiday bringing in more demand during June and July, and large events in the city of Boston bringing in demand during September and October.

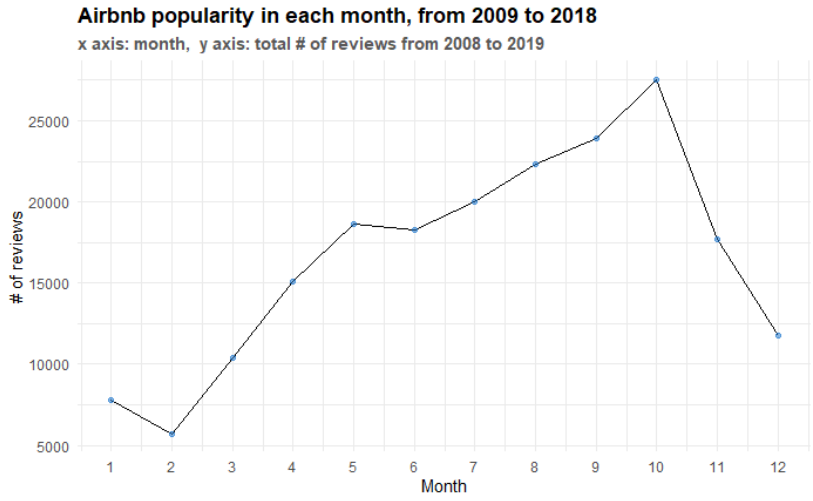


Figure 12

A closer look at the increase in demand

The following graph plots the year on the y axis increase in number of reviews on the y axis. We use this to analyse the marginal increase in demand from each year.

```
   listing_id             id                  date            reviewer_id        reviewer_name                comments
 Min.   :    3781   Min.   :     1021   Min.   :2009-03-21   Min.   :       1   Michael :  1863   Great location!:   203
 1st Qu.: 4863108   1st Qu.:113954444   1st Qu.:2016-11-14   1st Qu.: 20562998   David   :  1758   Great place!   :   179
 Median :13296735   Median :210669732   Median :2017-11-10   Median : 53617248   John    :  1507   .              :   157
 Mean   :12462051   Mean   :204844889   Mean   :2017-07-26   Mean   : 72054436   Sarah   :  1295   Great location :   146
 3rd Qu.:19209936   3rd Qu.:298584654   3rd Qu.:2018-07-29   3rd Qu.:115733849   Chris   :  1120   Great place    :   132
 Max.   :31574139   Max.   :402196828   Max.   :2019-01-17   Max.   :236100686   Jennifer:  1048   (Other)        :198286
                                                                                  (Other) :190515   NA's           :     3
   listing_id             id          date    reviewer_id reviewer_name      comments
            0              0             0              0             0             3
```

NOTE: We omit the year 2019 as the maximum value for the date column is 2019-01-17 (incomplete)

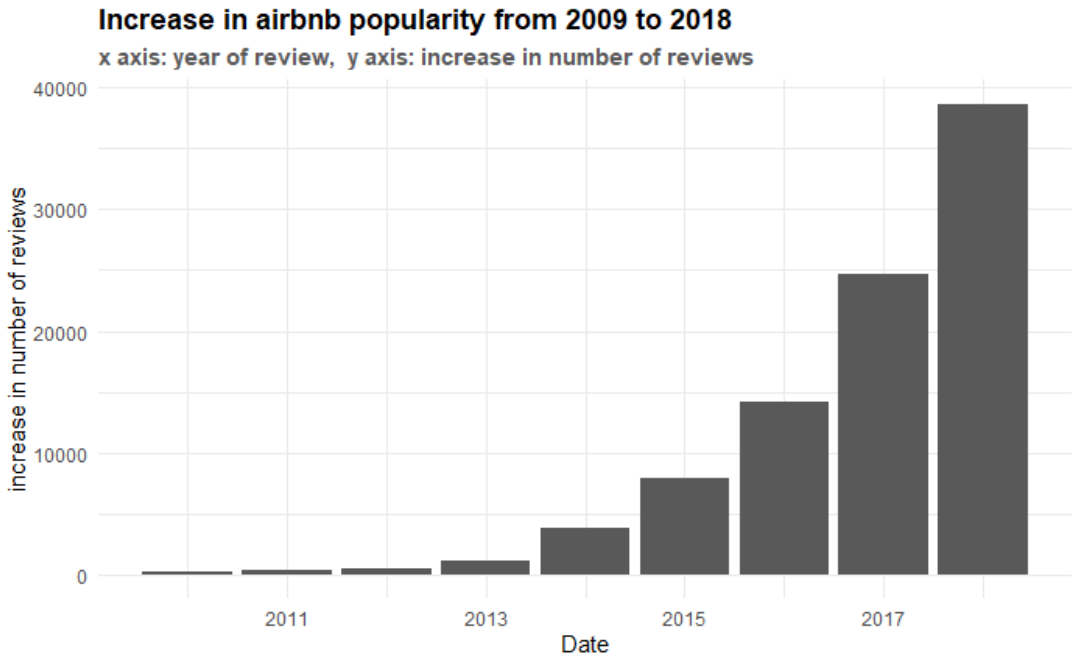


Figure 13

We get an idea of the growth of Airbnb's popularity using the following formula:

$$demand\ growth\ rate = \frac{current - past}{past}$$

2009	N/A
2010	9.3
2011	2.19
2012	0.87
2013	1.02
2014	1.6
2015	1.3
2016	1.008
2017	0.877
2018	0.73

CONCLUSIONS

Spatial data analysis:

- How location ratings change across the neighbourhoods?
  The concentration of the review locations (Figure 12) is near tourist areas and business locations. As the Boston Common and Boston Public Garden in Beacon Hill, Boylston St in Back Bay, Old North Church and St. Stephen`s Catholic Church in North End, Faneuil Hall Marketplace and New England Aquarium in Financial District.

- how is the Super Host distributed across the neighbourhoods?
  The super hosts (Figure 2) are spread across all the neighbourhoods of Boston and not concentrated in one area.

Price and supply analysis:

- What is the average price of Airbnb per year, and across all the years available?
  The average price of Airbnb listings (Figure 5) generally stays within the range of $100 to $200, with the exceptions of peaks in 2009,2013 and 2017, where prices rise to above $200.

- Is there a relationship between supply and price?
  We couldn't find any data that supports correlation between supply and price.

- Similarly, is there a relationship considering the different room types available?
  We couldn't find any data that supports correlation between supply and price considering the different room types available.
  We find that the range for the average price for Entire Home Apartment (Figure 6) is greater than that of all Airbnb listings and that of private rooms.
  And unlike the average price per year for all listings, and for Entire Home apartments, the average price per year for private rooms (Figure 7) does not follow a trend in price peaks every 4 years.

- Is it possible to stay at a Super host for under $100 per night?
  It is possible to stay in most of neighbourhood (Figure 8) in a Super Host with prices starting at $19 per night.

Demand analysis:

- What is the total demand for Airbnb per year, and across all the years available?
  There is a linear increase in demand between the years 2009 and 2012 (Figure 9) and after that the demand for Airbnb increases dramatically.

- Is there a relationship between supply, price, and demand?
  We couldn't find any data that supports correlation between supply, price and demand.

- Are there any trends in the demand for Airbnb?
  The demand trend (Figure 12) follows what we expect for travel seasonality, with the summer holiday bringing in more demand during June and July, and large events in the city of Boston bringing in demand during September and October.