Airbnb business insights

true

October 4, 2019

Contents

Problem Statement

What key factors affect the demand of Airbnb in US cities?

We are interested in the business domain of home-sharing services. We decide to work on analyzing some key factors resulting in the different demand among US cities. By knowing analyzing these questions, we believe it will help improve Airbnb's performance in the cities with low demand.

Data

```
knitr::opts_chunk$set(echo = TRUE,
                      results = 'hide',
                      message = FALSE,
                      warnings = FALSE)

library(RPostgreSQL)
```

```
## Loading required package: DBI
```

```
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```
library(readr)
library(tidyr)
library(stringi)
```

Airbnb listings in US

The data comes from OpenDataSoft (https://public.opendatasoft.com/explore/dataset/airbnb-listings/table/?disjunctive.host_verifications&disjunctive.amenities&disjunctive.features) and contains the airbnb listing details alongwith host information, property description and reviews given by the customers across various countries in the world. In this dataset, we will assume the number of reviews to be the number of bookings for any particular airbnb property.

Column	Meaning
host_id	Unique identifier for host
property_type	Type of the airbnb unit
minimum_nights	Minimum night booking required
number_of_reviews	Number of reviews attained by the property
review_score_rating	Rating score out of 100
cancellation_policy	Cancellation policy for the unit
city	city where the unit is present
zipcode	zipcode of the area
price	price of the property

```
if (!dir.exists("data/input/")) {
  if (!dir.exists("data")) {
    dir.create("data")
  }
  dir.create("data/input")
}

if (!file.exists("data/input/airbnb-listings.csv")) {
  download.file("https://public.opendatasoft.com/explore/dataset/airbnb-listings/download/?format=csv&t
```

```
  "data/input/airbnb-listings.csv")
}
```

Zipcode and location data for United States

The data comes from SimpleMaps (https://simplemaps.com/) and contains the list of zipcodes, cities and states in the United States. We have taken this dataset to basically map the correct cities to airbnb listings data via zipcode.

Column	Meaning
zipcode	zipcode of all the areas in US
city	Cities in the US
state	states in the US

```
if (!dir.exists("data/input/")) {
  if (!dir.exists("data")) {
    dir.create("data")
  }
  dir.create("data/input")
}

if (!file.exists("data/input/simplemaps_uszips_basicv1.6.zip")) {
  download.file("https://simplemaps.com/static/data/us-zips/1.6/basic/simplemaps_uszips_basicv1.6.zip",
  "data/input/simplemaps_uszips_basicv1.6.zip")
  unzip("data/input/simplemaps_uszips_basicv1.6.zip", exdir = "data/input")
}
```

United States population

The data comes from US government data website (https://www.data.gov/) and contains the population for last 3 years for each city in the United States. Zipcode is the unique identifier in this dataset.

Column	Meaning
zipcode	zipcode of all the areas in US
city	Cities in the US
population	Population
Year	Year

```
if (!dir.exists("data/input/")) {
  if (!dir.exists("data")) {
    dir.create("data")
  }
  dir.create("data/input")
}

if (!file.exists("data/input/pop-by-zip-code.csv")) {
  download.file("https://query.data.world/s/alfepb4cfqwqaiegamf25yfoprleq2",
  "data/input/pop-by-zip-code.csv")
}
```

Establishing connection with the database

Before running the below code, the relevant database should be created in the Postgres server. The below code establishes a connection with that specific database. Additionally, the db_connect.txt file that contains the credentials to connect to the database should also be present in the current working directory.

```
# Build will fail if the file isn't there.
assertthat::assert_that(file.exists('db_connect.txt'),
  msg = "Your connection file is missing.")

con_file <- readr::read_lines('db_connect.txt')

con <- RPostgreSQL::dbConnect(
             PostgreSQL(),
             host = con_file[1],
             port = con_file[2],
             user = con_file[3],
         password = con_file[4],
           dbname = con_file[5])
```

Initial Data Cleaning

We want to remove NA values, and pare down the input data into a managable chunk.

Airbnb Listings Data

We are going to remove some of the data. In particular, all the irrelevant columns will be filtered and other columns will be renamed. As we will be doing analysis only for United States, we will be filtering the data based on country code. We will also filter the property types which are irrelavant or seem to be an outlier for our analysis.

```
if (!file.exists('data/output/airbnb-listings.rds')) {
  file_raw <- read.csv("data/input/airbnb-listings.csv",header = TRUE, sep = ';')
  file_reduced <- file_raw %>%
    #filter(Country.Code == 'US')
    select(host_id = 'Host.ID',
           host_since = "Host.Since",
           host_location = "Host.Location",
           host_response_time = "Host.Response.Time",
           host_listings_count = "Host.Listings.Count",
           host_total_listings_count = "Host.Total.Listings.Count",
           property_type = "Property.Type",
           room_type = "Room.Type",
           minimum_nights = "Minimum.Nights",
           number_of_reviews = "Number.of.Reviews",
           review_score_rating =  "Review.Scores.Rating",
           review_score_accuracy = "Review.Scores.Accuracy",
           review_score_cleanliness = "Review.Scores.Cleanliness",
           review_score_checkin = "Review.Scores.Checkin",
           review_score_location = "Review.Scores.Location",
           cancellation_policy = "Cancellation.Policy",
           city = "City",
```

```
            state = "State",
            zipcode = "Zipcode",
            latitude = "Latitude",
            longitude = "Longitude",
            accommodates = "Accommodates",
            price = "Price",
            country_code = "Country.Code"
            ) %>%
                 filter(country_code == 'US') %>%
                   filter(!(property_type == 'Casa particular' |
                          property_type == 'Train' |
                          property_type == 'Plane' |
                          property_type == 'Parking Space' |
                          property_type == 'Van' |
                         property_type == '2017-04-02' |
                         property_type == 'Car'|
                          property_type == 'Boat')) %>%
                 filter(stri_length(zipcode) == 5) %>%
                     mutate_each(state, funs = toupper)

  if (!file.exists('data/output')) {
    dir.create('data/output')
  }
  saveRDS(file_reduced, 'data/output/airbnb-listings.rds')
} else {
  file_reduced <- readRDS('data/output/airbnb-listings.rds')
}


#Create a new data frame that exclude Price == NA
file_cna2 <- file_reduced[!is.na(file_reduced$price),]
file_cna2$host_since <- as.Date(file_cna2$host_since, format = '%Y-%m-%d')
```

Adding column property_id in the airbnb dataset

The below code adds a column named 'property_id' which will uniquely identify each property.

```
file_cna2$property_id <- seq.int(nrow(file_cna2))
```

Population Data

Here, we include population data only for the year 2016 and rename the columns.

```
if (!file.exists('data/output/population.rds')) {
  population_raw <- readr::read_csv("data/input/pop-by-zip-code.csv")

  population <- population_raw %>%
    select(zip_code_pop = 'zip_code',
           pop = 'y-2016')

  if (!file.exists('data/output')) {
```

```
    dir.create('data/output')
  }
  saveRDS(population, 'data/output/population.rds')
} else {
  population <- readRDS('data/output/population.rds')
}
```

US Zipcodes Data

Here, we keep the entire dataset, just rename teh columns.

```
if (!file.exists('data/output/match_zip_city.rds')) {
  match_zip_city <- readr::read_csv("data/input/uszips.csv")

  us_zip <- match_zip_city %>%
    select(zip = 'zip',
           city = 'city',
           state_id = 'state_id',
           state_name = 'state_name',
           county_name = 'county_name'
           )

  if (!file.exists('data/output')) {
    dir.create('data/output')
  }
  saveRDS(us_zip, 'data/output/match_zip_city.rds')
} else {
  us_zip <- readRDS('data/output/match_zip_city.rds')
}
```

Table Creation

When we push data into the database, we do the same thing each time. We can wrap this in a function. The function `post_data()` deletes the table and any keys or indexes we have made, before creating the table again.

Creating function that will be called to write all the tables in the database

```
post_data <- function(con, x, tablename = "") {
  if (dbExistsTable(con, tablename)) {
    dbExecute(con,
      paste0("DROP TABLE ", tablename,
             " CASCADE"))
  }

  dbWriteTable(con,
               tablename,
               x,
               row.names = FALSE,
```

```
                overwrite = TRUE)
}
```

Write tables to the database

The three datasets downloaded at the beginning have been broken down into normalized tables. These tables will be inserted into the database using the connection established earlier.

Table: Host

After inserting this table into the database, we create an index on the host_id for fast retrieval of data.

```
host_data <- unique(file_cna2[c('host_id', 'host_since')])

post_data(con,
            host_data,
            "host")
dbExecute(con,
  "CREATE UNIQUE INDEX host_idx ON host (host_id)")
```

Table: Host Details

As the airbnb listings dataset can have the same host repeated multiple times, we retieve the unque host details and insert it into host_desc table.

```
host_desc <- unique(file_cna2[c('host_id', 'host_location', 'host_response_time', 'host_total_listings_

post_data(con,
            host_desc,
            "host_details")

dbExecute(con, "ALTER TABLE host_details ADD CONSTRAINT hidx FOREIGN KEY (host_id) REFERENCES host (hos
```

Table: Location

After inserting this table into the database, we create an index on the zipcode for fast retrieval of data.

```
post_data(con,
            us_zip,
            "location")

dbExecute(con,
  "ALTER TABLE location ADD PRIMARY KEY (zip)")
```

Table: Property

After inserting this table into the database, teh column property_id is set as the primary key for the table. We also create an index on this column for fast retrieval of data.

```
property <- file_cna2[c('property_id', 'zipcode', 'host_id', 'minimum_nights', 'property_type', 'room_ty

post_data(con,
            property,
            "property")

dbExecute(con,
  "ALTER TABLE property ADD PRIMARY KEY (property_id)")

dbExecute(con, "ALTER TABLE property ADD CONSTRAINT hid FOREIGN KEY (host_id) REFERENCES host (host_id)

dbExecute(con,
  "CREATE UNIQUE INDEX property_idx ON property (property_id)")
```

Table: Reviews

While inserting this table into the database, we create a column called review_id. This is later set as the primary key for the table. We also create an index on the review_id for fast retrieval of data.

```
reviews <- file_cna2[c('property_id', 'review_score_rating', 'review_score_cleanliness', 'review_score_a

post_data(con,
            data.frame(review_id = 1:nrow(reviews),reviews),
            "reviews")

dbExecute(con,
  "ALTER TABLE reviews ADD PRIMARY KEY (review_id)")

dbExecute(con, "ALTER TABLE reviews ADD CONSTRAINT pid FOREIGN KEY (property_id) REFERENCES property (pr

dbExecute(con,
  "CREATE UNIQUE INDEX review_idx ON reviews (review_id)")
```

Table: Population

After inserting this table into the databse, we set an index on the zipcode column for fast retrieval of the data.

```
post_data(con,
            data.frame(population),
            "population")
dbExecute(con,
  "CREATE UNIQUE INDEX zip_code_popx ON population (zip_code_pop)")
```

List all tables present in the database

```
dbListTables(con)
```

Get data from database that will be used to do the analysis and plot graphs

This is the first dataset that we retrieve from our database. It summarizes the city wise number of bookings, population and price of the airbnb units.

```
dataset1 = dbGetQuery(con, "
  SELECT L.city,
    SUM(PR.number_of_reviews) AS no_of_bookings,
    ROUND(CAST(AVG(POP.pop) AS NUMERIC), 2) AS pop_density,
    ROUND(CAST(AVG(PR.price) AS NUMERIC), 2) AS price
  FROM property PR
    INNER JOIN location L ON L.zip = PR.zipcode
    INNER JOIN population POP ON L.zip = POP.zip_code_pop
  GROUP BY L.city
")
```

Our second dataset retrieves data about the host, property price, property location and its characteristics like room type, minimum nights, number of bookings made for each property and the review scores given by customers for each property.

```
dataset2 = dbGetQuery(con, "
  SELECT H.host_id,
    HD.host_total_listings_count,
    PR.property_id,
      PR.price,
      PR.room_type,
      PR.minimum_nights,
      PR.number_of_reviews AS no_of_bookings,
      PR.cancellation_policy,
      R.review_score_rating,
      R.review_score_cleanliness,
      R.review_score_accuracy,
      R.review_score_checkin,
      R.review_score_location
  FROM property PR
    INNER JOIN reviews R ON R.property_id = PR.property_id
    INNER JOIN host H ON H.host_id = PR.host_id
    INNER JOIN host_details HD ON H.host_id = HD.host_id
")
```

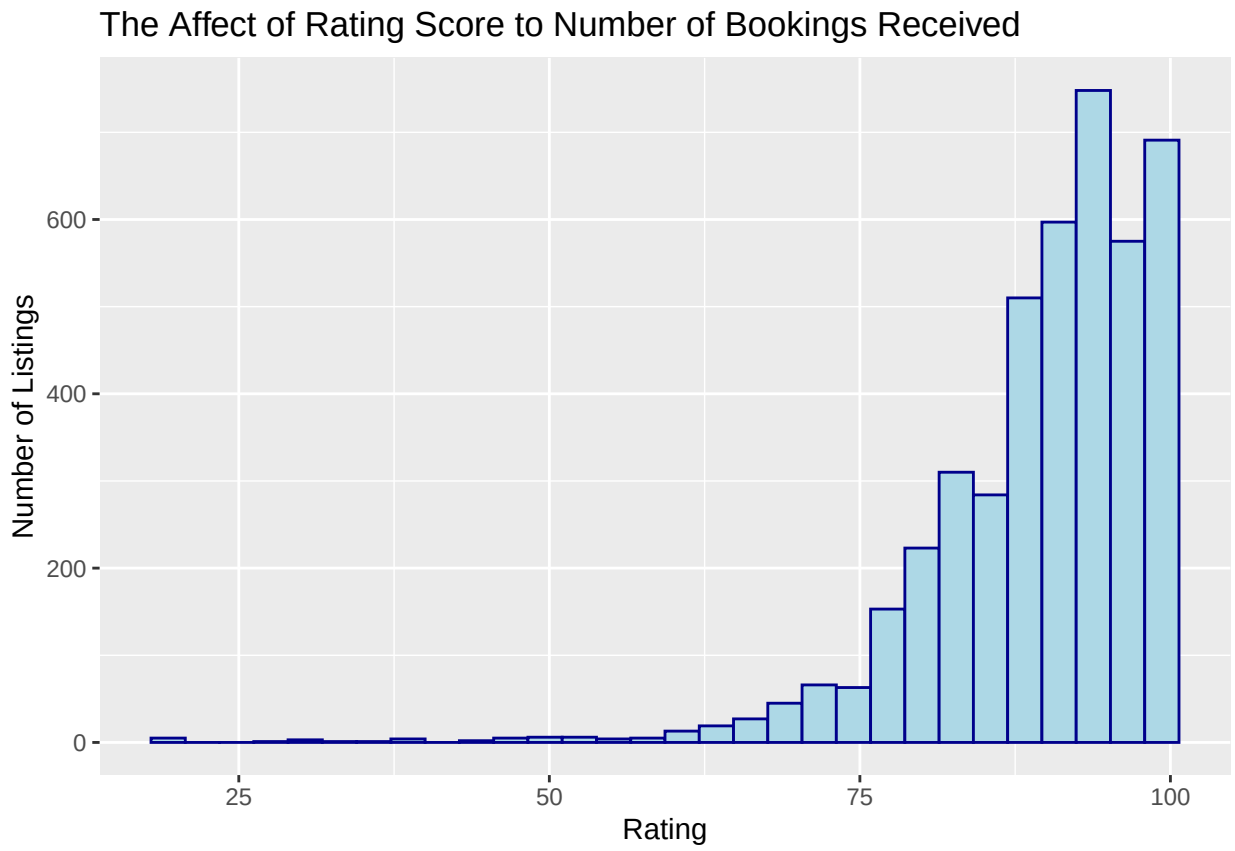
Analysis

Question One

Does the review score rating affect the demand of Airbnb(total listings)? How does it affect the demand? Positively or negatively? Slightly or heavily?

As part of this question. we want to explore the relationship between the the number of bookings and the rating score given by the customers. According to the graph, it is obvious that the number of bookings and the rating score has a positive effect to each other. From this analysis, we can conclude that the good rating score affects the Airbnb demand. From Airbnb management's perspective, the company can encourage the hosts to improve the quality of their services, which will lead to good review scores as well as an increase in the demand for that airbnb unit.

```
rating_by_count <- dataset2 %>%
                  filter(!is.na(review_score_rating) & !is.na(no_of_bookings)) %>%
                  count(no_of_bookings, review_score_rating)
ggplot(rating_by_count, aes(x = review_score_rating)) + geom_histogram(color = 'darkblue', fill = 'ligh
```



Question Two

Does price have a significant impact on the demand of Airbnb (total bookings) in NYC? In medium-sized cities? And small-sized ones?

This graph illustrates the relationship between the demand and price range of airbnb units. We can observe that the price range for highest demand is between $40-$60, and then the demand decreases as the price increases. However, for the price under $40, the demand is lower than the range $40-$120 because the customer may have minimum standard for the safety, quality, etc for the listing. We can conclude that from the Airbnb management perspective, Airbnb can focus on providing more listings in the $40-$100 price range. Furthermore, for the price around $500, it is popular compared to $400 and $600's listings because the customers are not only looking for economical accommodation, but also the quality and environment of the listings. Therefore, Airbnb can focus on specific customer group, earn profit from customers have high consuming ability.

```
price_vs_numberReviews <- dataset2 %>%
                    filter(!is.na(price) & !is.na(no_of_bookings))

ggplot(price_vs_numberReviews, aes(x = price)) + geom_histogram(stat = 'density', binwidth = 30, col =
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

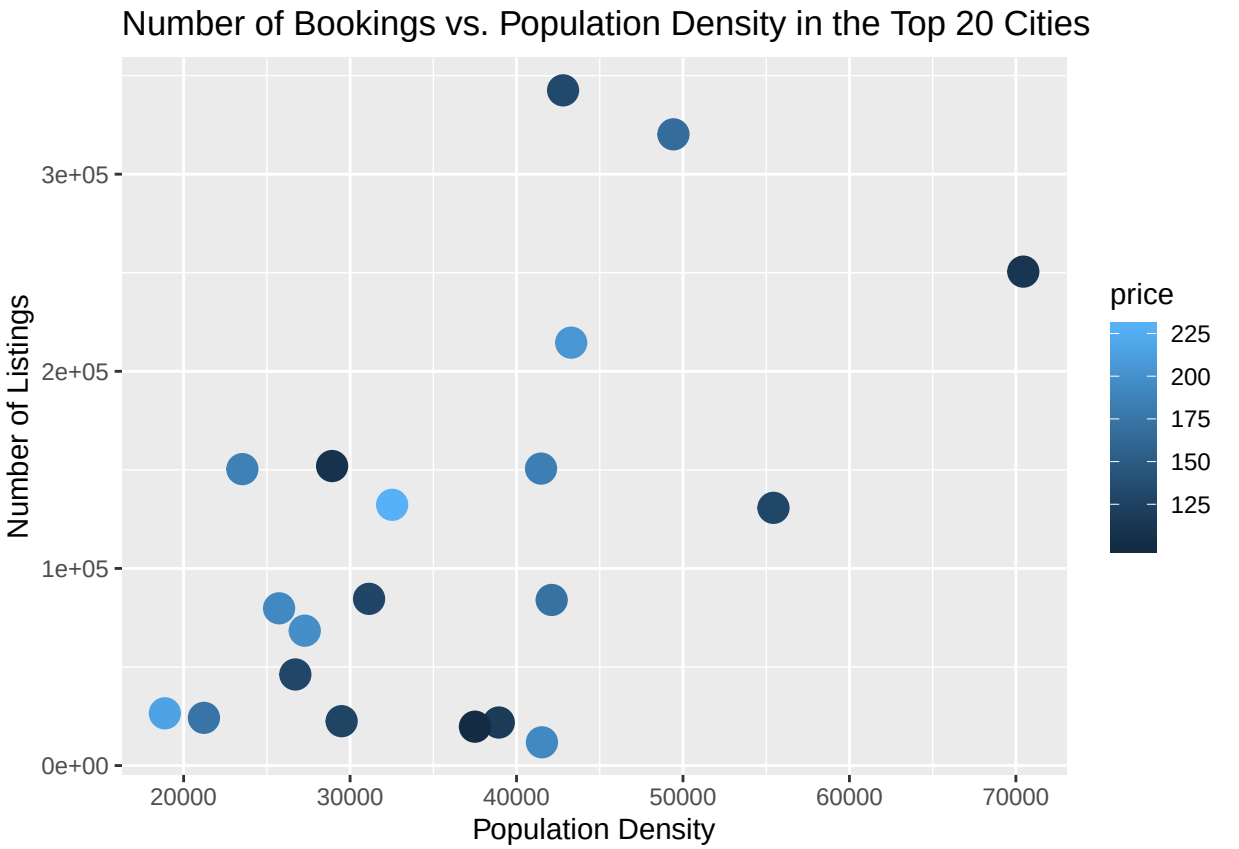
Question Three

Do cities with larger population lead to more total bookings as compared to cities with lesser population? What is the trend?

This graph aims to analyze the relationship between the population in the cities and the number of bookings. The darker the color of the bubbles, more is the price of airbnb there. Our assumption for analysis was to identify if the population of a specific city plays any role in affecting the demand or the pricing of airbnbs in that particular city. However, from this graph, we see that the points are quite scattered and there is no consistent relation between population, price and demand. It can be due to outliers in the data as well. We can safely assume a slightly positive correlation between demand (number of bookings) and population. However, we can not make any business perspective suggestions from this analysis.

```
dataset1_sorted <- dataset1 %>%
  arrange(desc(no_of_bookings)) %>%
  head(20)

ggplot(dataset1_sorted, aes(x = pop_density, y = no_of_bookings, color = price)) +geom_point(size = 5) +
```



Question Four

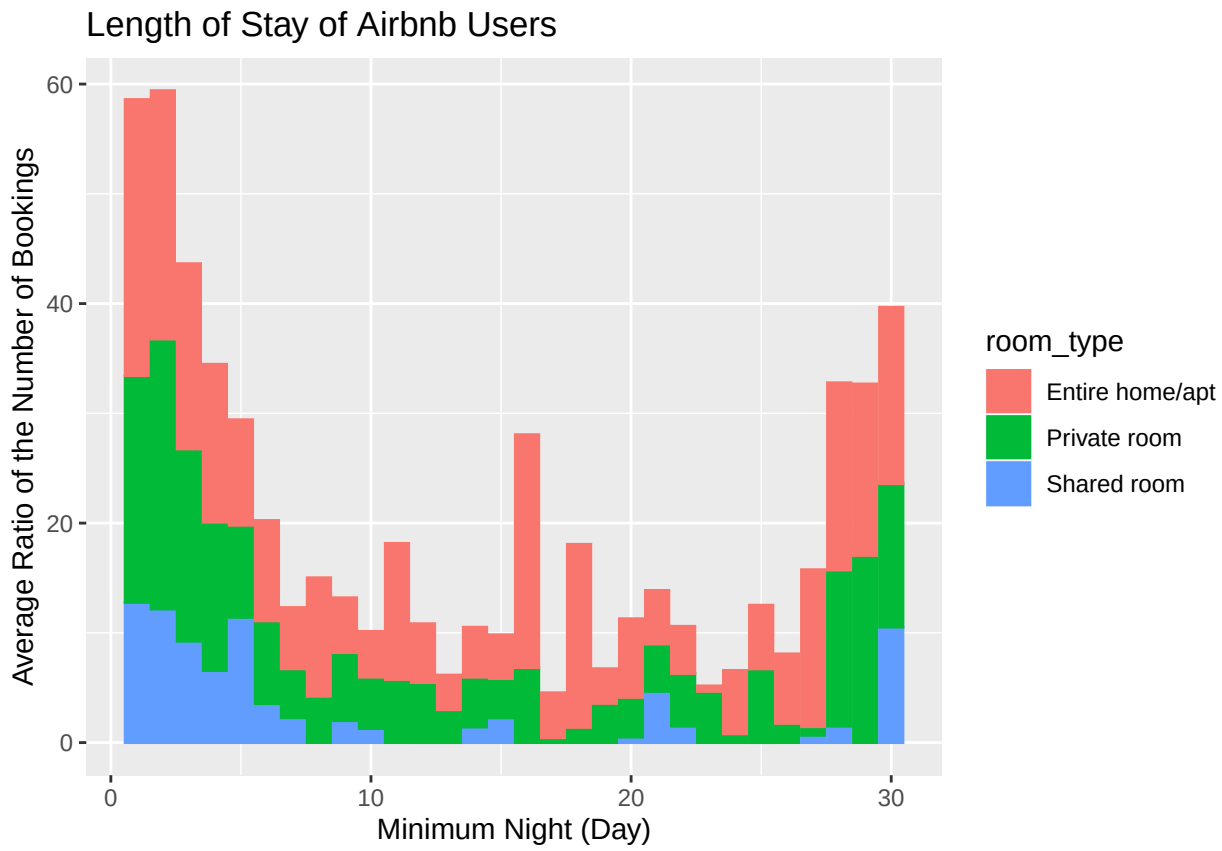
Comparing the minimum nights for 1/2/3/4/5/6, does specific minimum nights requirement have more demand as compared to others? Do customers intend to stay for longer prefer using Airbnb or customers with the need of one night stay?

This graph showcases the relationship between average number of bookings and minimum night required to stay. As per the analysis, the demand for 1 night stay is the highest. However, the graph shows a decrease in the demand for the units that require minimum 5-25 nights. It again increases when the minium nights required approaches 30 days. From this analysis, we can conclude that the demand for 1-4 nights of staying is the highest, the Airbnb can recruit more hosts with the short-term available properties. On the other hand, Airbnb can also recruit more hosts with long-term (at least one-month) properties available. Another perspective of analysis from this graph is the type pf property that is in demand. Overall, it can be observed that the private rooms and entire homes are more in demand as compared to the shared rooms, irrespective of the minimum nights criteria. So, either the listings for private rooms and entire homes should be increased so as to be more profitable. On the other hand, the reason for less demand of shared rooms should be identified and the hosts can improve on it. In this way, there is a possibility to get more demand for shared rooms as well.

```
minimum_nights_booking <- dataset2 %>%
                    filter(!is.na(minimum_nights) & !is.na(no_of_bookings & no_of_bookings != 0)) %>%
                    filter(minimum_nights <= 30) %>%
                    group_by(minimum_nights, room_type) %>%
                    summarise(avg_book = mean(no_of_bookings))
                    #summarise(mean_night = mean(no_of_bookings))
```

```
                    #count(no_of_bookings, minimum_nights)


ggplot(minimum_nights_booking, aes(minimum_nights, avg_book, color = room_type, fill = room_type)) + ge
```



Conclusions

In this project, we have worked on analyzing the business domain in a big picture, finalizing the industry and a business problem, gathering and preprocessing the data that relates to the problem statement, connecting and creating the database, retrieving data from database and plotting graphs to help us analyze the data visually. We have broken down our general question into four specific questions and tried to answer them in a data-driven way so that we can offer valuable business insights. We have included sufficient details in our rmd file for each question.

As a brief conclusion, based on the research and our analysis in this project, we can provide the following business insights to Airbnb. 1. Encourage more customers to leave positive reviews by offering promos via emails because we could figure out that there is a direct relationship between positive ratings and the number of bookings. 2. We also know that renting an entire house or apartment or private room is the most popular option which occupies the most market demand. We suggest Airbnb to add more listings on entire property renting instead of shared-room type. On the other hand, the reason for less demand for shared rooms can be identified and the hosts can work and improve on it. 3. Based on the relationship between pricing and demand, airbnb can focus on specific consumer group, earn profit from customers having high comsuming ability. This would also depend on the data which helps us analyse the relation between between consumer demographics based on their affluence and the demand for airbnb. However, as this data is unavailable, we cannot comment anything on this perspective of the analysis.

Thanks for reviewing our project!

References

https://public.opendatasoft.com/explore/dataset/airbnb-listings/information/?disjunctive.host_verifications&disjunctive.amenities&disjunctive.features

https://data.world/lukewhyte/us-population-by-zip-code-2010-2016

https://ipropertymanagement.com/airbnb-statistics

https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml