COLUMBIA UNIVERSITY
School of Professional Studies

Deep Analysis on the Dataset of Airbnb Property from Texas

by
Wei Yu

Final Deliverable Project for the Course of Applied Analytics Frameworks & Methods

Instructor: Vishal Lala

April 26th, 2018

Contents

1 Background and Data Cleaning Review **1**

2 Density Map and Data Analysis **3**

2.1 Density Map 3

2.2 Data Analysis 4

3 Text Mining on Property Descriptions **5**

4 Conclusion **7**

Appendices **8**

Bibliography **17**

Chapter 1

Background and Data Cleaning Review

Airbnb, founded in 2008, is a private US company that offers a online platform for landlord to rent their houses and/or rooms to visitors who are traveling to those cities. Our group has found some Airbnb listing data in Texas, US. The data is downloaded from Kaggle, which is a platform that contains lots of source data for data mining. The website address which can be used to download the source data is stated in reference [1]

In our first project, our group found some problems like unreadable codes, missing data and unformatted data. We used different tools and codes to fix them in order to get a more comprehensive dataset. Then we described our solutions to clean up our dataset and processed our new corrected dataset for this project. However, it is worth noticing that we did not focus on text columns in the original data in the previous project, whereas we will pay much more attention to in this project. To prepare for the unformatted characters in the text columns, we amended the methods to read the csv data and transformed to UTF-8 to avoid any unreadable characters.

The Airbnb company is experiencing its 10th-year celebration in 2018. It is important to review the past developments and set up new goals for the future. In this project, the company focuses more on the past 10 years development in Texas, especially on the increasing number of property listings, i.e. the property owners' attitudes to this method of earning money. Recall that we will be trying to analyze on a few research questions. It could be valuable to analyze the data collected from Texas about different aspects including property distributions, number of listings in terms of the year and how do the property owners attract potential customers to choose their properties.

The specific amended research questions are as follows:

1 How is the distribution of listed properties in Texas?

2 What is the trend of the number of listings change each year in Texas?

3 What would be the potential main factors for property owners to attract customers?

The next few chapters will be discussing based on the above research questions and use R language to generate relevant results. All the output will be attached in the appendices and cited in the appropriate position in the contents.

Chapter 2

Density Map and Data Analysis

2.1 Density Map

This original dataset contains more than 10 thousands Airbnb properties in Texas. It may be worth exploring the distributions of these property locations. In order to draw a density map of the properties, we use Google API to fetch the hybrid map and use relevant R code to show the distributions.

In order to draw this map, the first step is to decide the size of the region in Texas. We used the mean of all the longitudes and latitudes (excludes the NA's) as the center point of the map. We then adjust the scale of the map and decide the most appropriate size of the map. Next, by using the "ggmap" library, we are able to draw a heat map based on the valid locations of all the properties in the dataset. The output map is attached in the appendix. 1

In the density map, the red area represents the high density, whereas the blue area represents the low density. From the map, we can also observe that though having more than 10 thousands of properties in Texas, most of them locates next to each other.

There are mainly four gathering points containing the majority of the properties, which are Dallas, Austin, San Antonio and Houston. We may conclude that they are also the main residential quarters in Texas.

2.2 Data Analysis

As the data includes the properties listed from 2008, we are interested in the growing trend of Airbnb in the past ten years as well. In order to demonstrate this, we can use R code to draw a line chart to represent the new listings in each year.

In this step, we use "ggplot2" and "lubridate" libraries to help generate the chart. The code helps counts the number of new listings for each different months. In the month and year order, the code then plots the number in the right order of date. Consequently, because the data is gathered from 2008 to 2017, there would be more than 100 months if we present all in the x axis. Thus, instead of presenting every month, we use the beginning of each year to show the trend of the number of listings over the past ten years.

In the output picture 2, all the black points represent the number of new listings in that month and they are connected by red dashed lines . We can observe that the number of properties started from nearly being zero at the end of 2008. The number does not change much until 2011, where this industry started to grow rapidly. In the next six years, the number of new listings is tripled and even reached a peak of 800 properties in Jan 2017. This strongly proves that the leasing industry developed well in Texas.

On the other hand, as this line chart represents the number of new listings appeared each year, the total number of listings are more than presented. If we consider the cumulative situation, the number of current listings would be more than a thousand. If considering from the managing level perspective, it would be happy for them to see Airbnb developing well in Texas.

Chapter 3

Text Mining on Property Descriptions

In this dataset, there are a few columns containing lots of texts. The column that contains the most texts is called "description", which gives detailed descriptions of the listed properties to attract potential customers. This is also one of the main objectives set from previous project to work on the text. It is valuable as the results of the text mining will give key points about how property owners try to attract potential customers and what people particularly focuses on while selecting Airbnb properties.

Before generating any result from the data mining, it is necessary to get rid of unnecessary characters in the original dataset. This includes numbers, punctuations, extra white spaces and any other unformatted characters.

After the original data has been cleaned, the first step is to generate 10 most used words that appear in "description" excluding stop words. Stop words are usually meaningless articles such as "a" and "the". Excluding these words could give a more accurate representation on the most frequent words in descriptions.

The R code generates 10 most used words appeared in a horizontal bar chart 3. The 10 words are: home, place, room, private, downtown, minutes, close, located, bedroom and amp. From this result, we may guess that "feeling at HOME" is what most property owners use in descriptions and want to attract customers. Describing the places of the properties should also be important to customers as they need to know the exact location. As a result, words related to the location are also highly used such as "downtown", "minutes", "close" and "located".

If we want to show more words in terms of usage frequencies, we can also draw a word cloud picture showing more than a hundred words, where their scales depends on the usage frequency of the words.

In this case, we use "qdap" library along with some Java functions and apply with the "wordcloud" function. The code helps transform the frequency of texts into a matrix, and then into a data frame. Here we have also used a seed to lock the word cloud to be generated with the same picture every time. A proper seed selection could also avoid not including all the frequent words in the output picture.

In the output picture attached in the appendix 4, it is clear to see that "home" and "place" are the two biggest words. Recall that from previous chart about 10 most frequent used words, "home" and "place" are also sticked on top. They are then surrounded by other words which are all smaller than them. This means, the more times the word has been used in the "description" column as a text, the bigger the word will be in the output.

Chapter 4

Conclusion

In this project, we use the data cleaned by last project, and explore further analytics about Airbnb properties in Texas. This project could be a report to the managing levels of the Airbnb company, illustrating our business status in Texas over the past ten years.

The first key point is that the leasing industry is developing very well in Texas. This could be a strong signal to the managing level of Airbnb that the business model is well developed and can be copied to other areas in the future.

Since the model of Texas developed very well, we would recommend the company to apply this model in other undeveloped areas as well. It is essential to do enough research before actually applying it. The research also includes finding out the most appropriate areas for developing the business, which are usually residential areas.

Furthermore, text mining helps us analyze frequent words that customer most concern about. Top frequent words like "home" and "place" could be the key factors that help Airbnb further improve services and attract new customers.

Overall, visualizing images from dataset is one of our strength in our methodologies, which helps readers better understand our business status in Texas. However, we did not include much related to the prices of the leasing properties. This is because the price are related to different types of the properties. Also, as the dataset is collected for over ten years, there are many other aspects affecting the prices, such as global economics status or federal policies. Analyzing on this topic would add much complexity and unnecessary inconveniences to this project.

Appendices

Pictures Generate by R Code

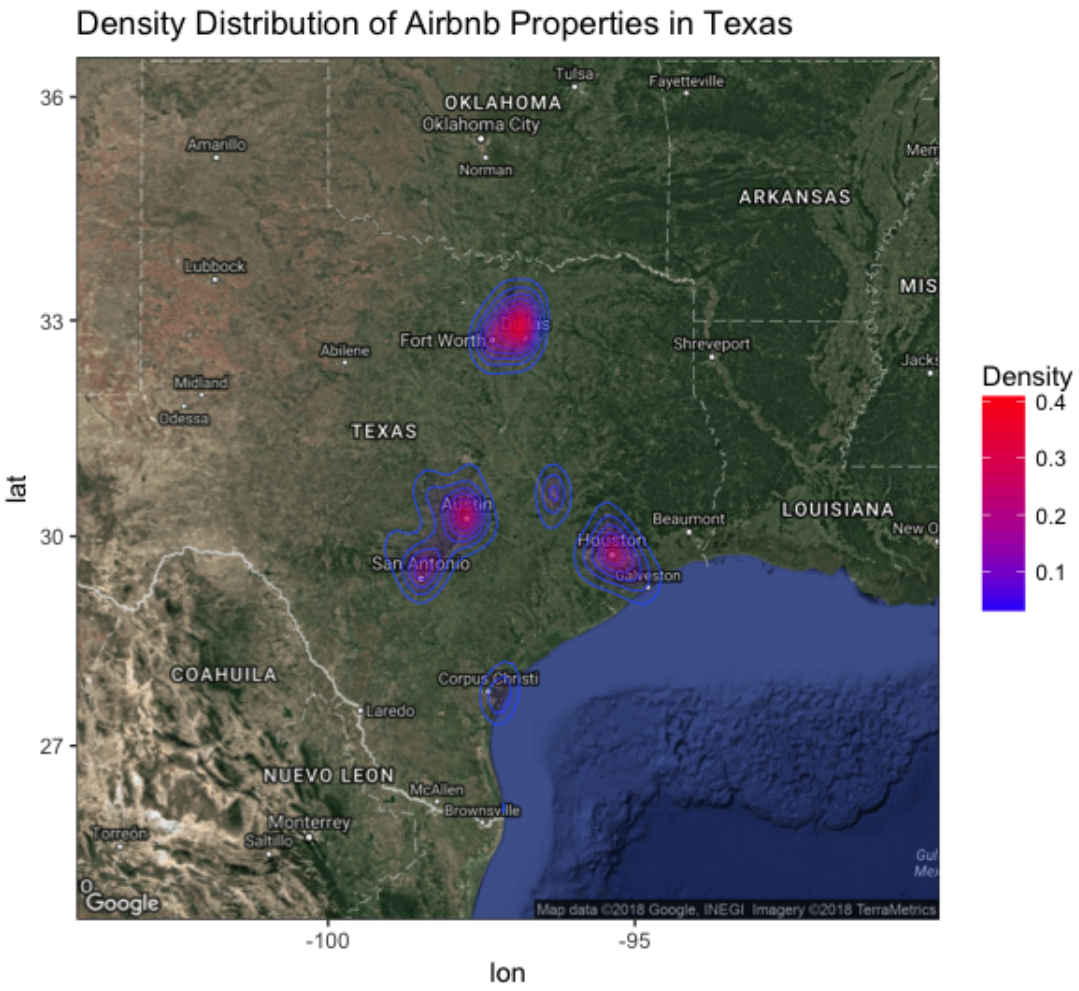

Figure 1: Density

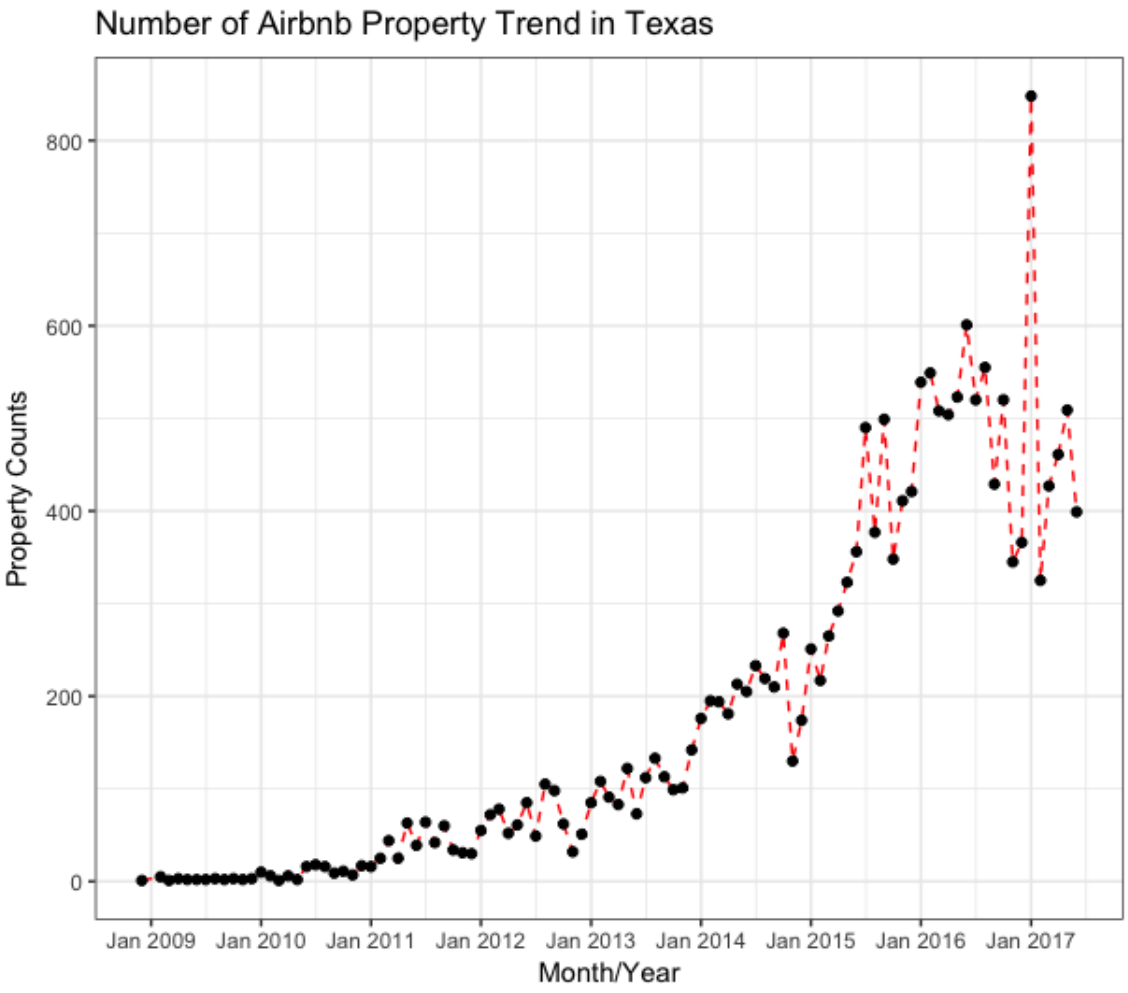

Figure 2: Number of Airbnb Properties in Texas 2009 - 2017

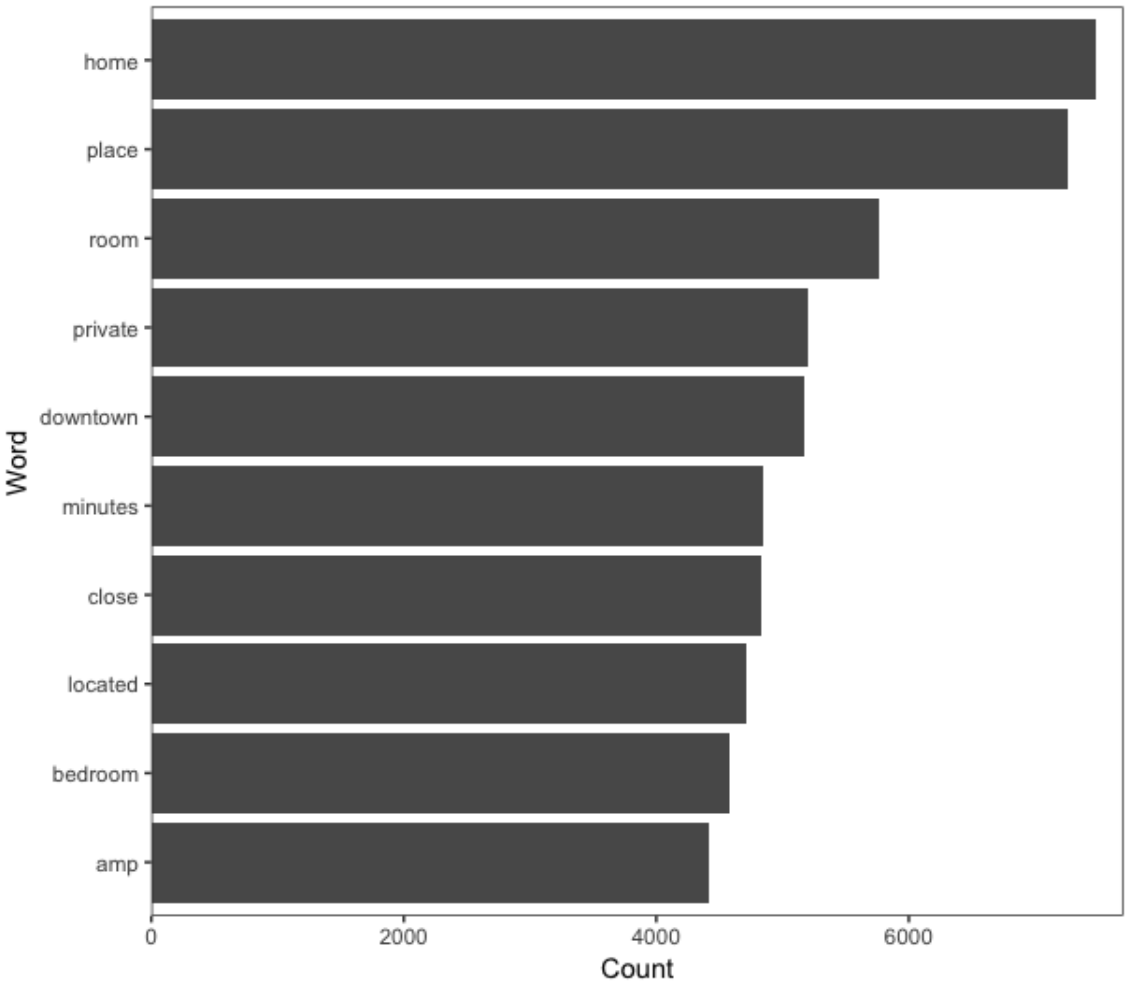

Figure 3: Histogram of Words

Figure 4: Word Cloud

R Code Screen Shots

```r
# This assignment is completed by Wei Yu (UNI: wy2314) and Shihong Song (ss5540)

csv_str = readLines("/Users/helenyu/Desktop/Columbia University/Courses/5200 Framework & Methods
                    /Group Assignment/Deliverable 2/Airbnb_Texas_Rentals.csv", warn = FALSE)
csv_str = iconv(csv_str, from = "WINDOWS-1252", to = "UTF-8", sub = " ")
Encoding(csv_str) <- "UTF-8"
csv = read.csv(text = csv_str, encoding = "UTF-8")

airbnb_data = csv
# identify key data issues
# 1. Unreadable data
# 2. Unformatted data: when measuring number of bedrooms,
# studio is a misleading term -> change to 0
# 3. missing data (rate, no. of bedroom, latitute and longitude);
head(airbnb_data, 10)
summary(airbnb_data)

# Examine the dataset
unique(airbnb_data$city)
# By using the above function, we have found that some output are not proper English
which(airbnb_data$city == "诺斯莱克")
which(airbnb_data$city == "阿纳瓦克")
# Identify the entries of "studio"
which(airbnb_data$bedrooms_count == "Studio")
# Identify the incompatible symbol "-"
which(airbnb_data$city == "Bryan-College Station")

# Code for identifying all the missing values within the dataset
airbnb_data[!complete.cases(airbnb_data), ]
# Identify missing values in each main column seperately
which(airbnb_data$average_rate_per_night == "")
which(airbnb_data$bedrooms_count == "")
which(airbnb_data$city == "")
which(airbnb_data$date_of_listing == "")
which(airbnb_data$description == "")
which(airbnb_data$latitude == "NA")
which(airbnb_data$longitude == "NA")

# Now start processing the data
# Before rewrite, we need to convert them into character type
airbnb_data$city <- as.character(airbnb_data$city)
# After locating the unreadable entries, we need to rewrite them in proper English.
airbnb_data$city[which(airbnb_data$city == "诺斯莱克")] = "Northlake"
airbnb_data$city[which(airbnb_data$city == "阿纳瓦克")] = "Anawak"

# In order to find out the studios, first need to change all the counts to "character" type,
# because "Studio" is in "character"
airbnb_data$bedrooms_count <- as.character(airbnb_data$bedrooms_count)
# Studios are counted as 0.7 bedroom
airbnb_data$bedrooms_count[which(airbnb_data$bedrooms_count == "Studio")] = "0.7"
# Now change the data type back to numeric
airbnb_data$bedrooms_count <- as.numeric(airbnb_data$bedrooms_count)
```

```
53
54  # Convert all the entries from dollars to numeric format
55  airbnb_data$average_rate_per_night <- as.numeric(sub('$','',as.character(
56    airbnb_data$average_rate_per_night),fixed=TRUE))
57  is.numeric(airbnb_data$average_rate_per_night)
58
59  # Get rid of the symbol "-" which is incompatible
60  airbnb_data$city[which(airbnb_data$city == "Bryan-College Station")] = "Bryan College Station"
61
62  # Now deal with missing values in the column of average_rate_per_night and bedrooms_count
63  # Use the data found from the website
64  airbnb_data$average_rate_per_night[104] = 85
65  airbnb_data$average_rate_per_night[105] = 30
66  airbnb_data$average_rate_per_night[168] = 210
67  airbnb_data$average_rate_per_night[173] = 49
68  airbnb_data$average_rate_per_night[181] = 32
69  airbnb_data$average_rate_per_night[182] = 49
70  airbnb_data$average_rate_per_night[343] = 79
71  airbnb_data$average_rate_per_night[344] = 109
72  airbnb_data$average_rate_per_night[345] = 99
73  airbnb_data$average_rate_per_night[948] = 250
74  airbnb_data$average_rate_per_night[1123] = 160
75  airbnb_data$average_rate_per_night[1215] = 120
76  airbnb_data$average_rate_per_night[1217] = 115
77  airbnb_data$average_rate_per_night[1219] = 129
78
79  airbnb_data$bedrooms_count[14238] = 0.7 # as it is a studio
80  airbnb_data$bedrooms_count[16812] = 1 # as it is a 1-bedroom property
81
82  # Calculate the mean and determine the temporary value
83  # for missing values in column average_rate_per_night
84  sumRate <- sum(airbnb_data$average_rate_per_night, na.rm = TRUE)
85  sumRoom <- sum(airbnb_data$bedrooms_count, na.rm = TRUE)
86  meanvalue <- sumRate / sumRoom
87  print(meanvalue)
88  airbnb_data$average_rate_per_night[which(is.na(airbnb_data$average_rate_per_night) )] =
89    meanvalue * airbnb_data$bedrooms_count[which(is.na(airbnb_data$average_rate_per_night) )]
90  # Round up the numbers to integers, i.e. no decimal places
91  airbnb_data$average_rate_per_night <- round(airbnb_data$average_rate_per_night, digits = 0)
92
93  # Fill the missing entry in "bedroom_count" column
94  airbnb_data$bedrooms_count[6876] = 1
95
96  # Finallly check if there is any missing value left in the columns of average_rate_per_night and bedrooms_count
97  which(airbnb_data$average_rate_per_night == "")
98  which(airbnb_data$bedrooms_count == "")
99
100 library(ggmap)
101 #Create Texas map
102 latavg = mean(na.omit(airbnb_data$latitude))
103 longavg = mean(na.omit(airbnb_data$longitude))
104 tx_map = get_map(location = c(lon=longavg, lat=latavg), zoom = 6, scale = 2, maptype = "hybrid")
105
```

```r
# Create the heat map
airbnb_data_no_na = na.omit(airbnb_data)
ggmap(tx_map, extent = "panel") +
  geom_density2d(data = airbnb_data_no_na, aes(x = longitude, y =latitude),
                 size = 0.3) +
  stat_density2d(data = airbnb_data_no_na, aes(x = longitude, y = latitude,
                                       fill = ..level.., alpha = ..level.. ), size = 0.001,
                 bins = 16, geom = "polygon") +
  scale_fill_gradient(low = "blue", high = "red",name = "Density") +
  scale_alpha(range = c(0,0.3), guide = FALSE) +
  ggtitle("Density Distribution of Airbnb Properties in Texas")

library(ggplot2)
library(lubridate)
library(plyr)
# number of airbnb growing with month/year
count_airbnb <- count(airbnb_data$date_of_listing)
#fre_airbnb <- frequency(count_airbnb)
#theme_set(theme_bw())

count_airbnb$x <- as.Date(count_airbnb$x)
plot1 <- ggplot(data=count_airbnb, aes(x=count_airbnb$x, y=count_airbnb$freq, group=1)) +
  geom_line(linetype = "dashed", color="red") + scale_x_date(date_breaks = "1 year", date_labels = "%b %Y") +
  geom_point()
plot1  + labs(title = "Number of Airbnb Property Trend in Texas", x = "Month/Year", y = "Property Counts")

library(qdap)
# 10 most frequent words used
plot(freq_terms(airbnb_data$description,top = 10,at.least = 3,stopwords = tm::stopwords('english')))

# text mining
library(RColorBrewer)
library(wordcloud)
library(tm)
library(SnowballC)

corpus<- Corpus(VectorSource(airbnb_data$description))

```

```
143
144 #inspect(corpus)
145 toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
146 corpus <- tm_map(corpus, toSpace, "/")
147 corpus <- tm_map(corpus, toSpace, "@")
148 corpus <- tm_map(corpus, toSpace, "\\|")
149 # Text cleaning
150 corpus <- tm_map(corpus, tolower)
151 corpus <- tm_map(corpus, removeNumbers)
152 corpus <- tm_map(corpus, removeWords, stopwords("english"))
153 # Remove punctuations
154 corpus <- tm_map(corpus, removePunctuation)
155 # Eliminate extra white spaces
156 corpus <- tm_map(corpus, stripWhitespace)
157 # Text stemming
158 # docs <- tm_map(docs, stemDocument)
159 dtm <- TermDocumentMatrix(corpus)
160 m <- as.matrix(dtm)
161 v <- sort(rowSums(m),decreasing=TRUE)
162 d <- data.frame(word = names(v),freq=v)
163 head(d, 10)
164 set.seed(600)
165 wordcloud(words = d$word, freq = d$freq, min.freq = 1,
166           max.words=120, random.order=FALSE, random.color = FALSE, rot.per=0.5,
167           colors=brewer.pal(12, "Paired"))
168
```

Bibliography

[1] Inc. Airbnb. About us - airbnb newsroom. https://press.atairbnb.com/about-us/, 2018.

[2] Wikipedia contributors. Airbnb — wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Airbnb&oldid=828363115, 2018. [Online; accessed 2-March-2018].

[3] etc. Donyoe, Faraz92. Airbnb property data from texas - kaggle. https://www.kaggle.com/PromptCloudHQ/airbnb-property-data-from-texas, 2018.

[4] David Kahle and Hadley Wickham. ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1):144–161, 2013.

[5] STHDA.com. Text mining and word cloud fundamentals in R: 5 simple steps you should know. http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know, 2018. Online; accessed Apr. 26 2018.