

Projet_ADD

Lucie Guillaumin, Lyes Demni, Jeremy Hazan

10 janvier 2020

Contents

1. Introduction	2
Importation des packages	3
Importation des données	3
Quelques corrections préliminaires des données	4
Gestion des valeurs manquantes et de la perte d'information	4
Evolution de la popularité de Airbnb au cours du temps	7
2. Analyse descriptive (et analyse des variables) du jeu de données	9
Description des données	9
Etude des variables	10
3. Etude Manhattan/Brooklyn	14
Création des data set	14
Etude du prix par nuit d'un Airbnb	14
Prix moyen par nuit des locations	18
Comparaison Manhattan/Brooklyn	22
A propos des hôtes	24
Map des airbnbs à Brooklyn et Manhattan	26
4. Explorations des données textuelles (Traitement automatique du langage naturel)	27
Echantillonage des données, nettoyage des données et résultats	27
5. Conclusion	29

1. Introduction

Nous sommes une équipe de trois personnes - Lucie Guillaumin, Lyes Demni, et Jérémy Hazan. Nous vous présentons ci-dessous notre analyse de données exploratoires, avec comme objectif d'utiliser les outils étudiés en cours de la library tidyverse.

Avant de commencer notre projet concernant les renseignements sommaires et mesures pour les Airbnb à New York, nous allons expliquer comment et pourquoi nous avons choisis ce sujet.

Tout d'abord, pourquoi avons nous choisis ce groupe ?

Fort d'une expertise en NLP à la suite d'un stage en tant que NLP Data Analyst, Jérémy a pu apporter toute sa connaissance et son expertise dans ce domaine singulier du machine learning et surtout l'appliquer grâce aux commentaires présent dans le jeu de données. La connaissance du cleaning des données textuelles et les notions de bag of words nous ont permis de faire un data mining pertinent. Lyes, quant à lui avait déjà fait l'usage du package leaflet dans le cadre d'un ancien projet. Il nous a dès lors proposer la pertinence d'utiliser les coordonnées géographiques présentes dans le jeu de données. Enfin, Lucie avait déjà été confrontée à devoir gérer des jeux de données désordonnés. Elle nous a grandement aider à trier les variables et trouver les plus pertinentes d'entre elles.

En outre, qu'est ce qu'Airbnb ?

Airbnb est une plateforme communautaire qui met en relations deux types d'agents; les personnes qui cherchent un hébergement (Airbnb 'guests') et ceux qui cherchent à louer leurs biens sur le court ou moyen-terme (Airbnb 'hosts'). Les biens locatifs peuvent être aussi bien des appartements, que des maisons entières, des péniches... Depuis sa création en 2008, Airbnb s'est considérablement développé, aussi bien financièrement que d'un point de vue de la gamme des services proposés. En 2019, il y a plus de 150 millions d'usagers des services Aribnb dans 191 pays, le faisant désormais un acteur majeur du maelström qu'a connu le tourisme au cours de cette dernière décennie.

Airbnb créé son chiffre d'affaire en prenant une commission sur les deux agents pour les séjours. Les 'hosts' versent une commission à hauteur de 3% de la valeur de la location, alors que les 'guests' se voient octroyer entre 6 et 12% de leurs biens locatifs à la plateforme. Comme écosystème de locations de biens, Airbnb génère un très grand nombre de données intéressantes à analyser.

Lien data

Sur le lien inside Airbnb donné ci-dessus, nous avons accès aux base de données de plus d'une centaines de villes dans le monde. Nous nous sommes concentrés sur les données de la ville de New York pour plusieurs raisons. Premièrement, New York est une ville de rang mondiale avec des quartiers et des cultures à la fois unique et extrêmement diversifiées. New York City dispose d'un des marchés Airbnb les plus denses sur la planète; plus de 48,000 logements listés en Août 2019 (ce qui correspond à une densité locative de 102 locations par miles carré).

Par ailleurs, le premier appartement en location sur Airbnb à New York date de 2008, année de création de la start-up californienne. En disposant de cette base de données, nous avons alors une possibilité d'étudier plus en profondeur évolution de la popularité de cette jeune entreprise sur plus de 10 ans, contrairement aux villes européennes arrivées beaucoup plus tardivement sur le marché. Qui plus est, 2008 est une date clé, et plus particulièrement à New York, avec l'avènement d'une des plus grandes crises économiques mondiales.

Enfin, la décennie a été marquée par d'importants changements au niveau des arrondissements. Brooklyn, et en particulier des quartiers tels Williamsburg, Red Hook, Dumbo et Brooklyn Heights ont pris en popularité... A tel point que Manhattan est désormais comparable à Brooklyn au niveau du tourisme. Plus loin encore, de nombreux touristes risquent de choisir Brooklyn au détriment de Manhattan pour son calme, sa vue imprenable, et sa diversité de cultures et de restauration. C'est en ce sens que nous comptons nous adosser à une comparaison Manhattan/Brooklyn dans la suite du projet, avec pour objectif de reléguer les autres arrondissements à des bruits statistiques.

Quelques questions auxquelles nous cherchons à répondre à travers notre analyse:

- Comment et à quelle intensité les prix des logements varient par location?
- Comment peut-on quantifier l'évolution de la popularité de Airbnb à New York ?
- Comment la demande pour les locations d'Airbnb ont elles fluctué au cours de l'étude ?
- Quels sont les types de logements sur NYC? Varient-ils de façon conséquente par quartier ?
- Quels quartiers à NYC font le plus souvent l'objet de commentaires par les hôtes ?
- Quels sont les mots qui représentent le mieux les différents quartiers de NYC ? Quel est le sentiment général des hôtes en fonction des quartiers ?

Importation des packages

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library("Hmisc")
library(wordcloud)
library(RColorBrewer)
library(wordcloud2)
library("tm")
library(SnowballC)
library(tidytext)
library(textdata)
library(stringr)
library(corrplot)
library(choroplethrMaps)
library('gridExtra')
library(cowplot)
library(leaflet)
```

Importation des données

Nous allons importer nos données grâce aux fonctions d'importation de tidyverse qui transforme directement en tibble, le séparateur , est initialisé automatiquement et reconnaît les formats **data, double et chaîne de caractères** automatiquement.

```
listings <- read_csv(file = 'C:/Users/lulu/Documents/M1/S1/1_PROJET ADD/ABNB/listings.csv', col_names = TRUE)
reviews <- read_csv(file = 'C:/Users/lulu/Documents/M1/S1/1_PROJET ADD/ABNB/reviews.csv')

attach(listings)
attach(reviews)

#Voici ci-dessous les différentes données afin que vous puissiez les voir si vous le désirez
#view(listings)
#view(reviews)

dim(listings)
```

```
[1] 50599    106
```

On remarque que notre jeu de données possède 106 variables.

Nous avons décidé de directement supprimer des variables (55) que ne nous semblent peu voire pas pertinentes en vu de notre analyse.

```

dataAB1 <- select(listings, c(1,5,20,22,23,26,27,29,35,40,41,49,50,53:69,78:93,97,99:102))

dim(dataAB1)

[1] 50599      51

```

Quelques corrections préliminaires des données

1. `comment` (`reviews`): Nous avons effectué l'analyse des commentaires des hôtes dans le cadre de notre analyse. Le jeu de données contenait de nombreux commentaires dans d'autres langues que l'anglais (Chinois, Japonais, Espagnol). Nous avons éliminé ces données pour ne récupérer que les données écrites en anglais. Nous avons par la suite réalisé un travail de nettoyage des commentaires, c'est à dire du filtrage de texte à l'aide de la fonction stop words.
2. `price` (`listings`): La colonne `price` contenait des données en format string avec le symbole \$, ainsi qu'une virgule. Cette colonne, à l'aide du chapitre portant sur les 'regular expressions' a été corrigé afin qu'elle contienne des valeurs entières pour l'analyse des données.

```

dataAB2 <- mutate(dataAB1, prix = parse_number(price), prix_semaine = parse_number(weekly_price), prix_mois = parse_number(monthly_price))

dataABNB <- select(dataAB2, -c(22:26,28)) %>% filter(prix > 0)

#Si vous souhaitez voir le data set :
#view(dataABNB)

```

Gestion des valeurs manquantes et de la perte d'information

Par ailleurs le jeu de données contenait des données non indiquées (NA). Pour travailler de façon pertinente nous avons enlevé toutes lignes et les colonnes qui contenaient des valeurs NA. Par ailleurs, certaines données faussaient les résultats d'analyse tels que les prix ou les tailles d'appartements fixées à 0.

```
dataABNB <- dataABNB %>% filter(prix > 0, minimum_nights > 0)
```

Nous prenons toutes les variables quantitative du jeu de données :

```

dataABNB %>% select_if(is.numeric) -> dataAB_num
head(dataAB_num)

```

```

# A tibble: 6 x 32
  id host_id latitude longitude accommodates bathrooms bedrooms beds
  <dbl>   <dbl>    <dbl>     <dbl>        <dbl>     <dbl>    <dbl> <dbl>
1 2595     2845     40.8    -74.0         1         1       0     1
2 3831     4869     40.7    -74.0         3         1       1     4
3 5099     7322     40.7    -74.0         2         1       1     1
4 5121     7356     40.7    -74.0         2        NA       1     1
5 5178     8967     40.8    -74.0         2         1       1     1
6 5203     7490     40.8    -74.0         1         1       1     1
# ... with 24 more variables: square_feet <dbl>, guests_included <dbl>,
#   minimum_nights <dbl>, maximum_nights <dbl>, availability_30 <dbl>,
#   availability_60 <dbl>, availability_90 <dbl>, availability_365 <dbl>,
#   number_of_reviews <dbl>, number_of_reviews_ltm <dbl>,
#   review_scores_rating <dbl>, review_scores_accuracy <dbl>,
#   review_scores_cleanliness <dbl>, review_scores_checkin <dbl>,
#   review_scores_communication <dbl>, review_scores_location <dbl>,
#   review_scores_value <dbl>, calculated_host_listings_count <dbl>,
#   prix <dbl>, prix_semaine <dbl>, prix_mois <dbl>, prix_caution <dbl>,
#   ...

```

```
#   frais_nettoyage <dbl>, personne_supp <dbl>
```

Nous voulons vérifier s'il y a des valeurs manquantes (NA) dans le data set, ainsi qu'une idée du nombre de valeurs omises :

```
dataAB_num[] %>% is.na() %>% any()
```

```
[1] TRUE
```

```
dataAB_num[] %>% is.na() %>% sum()
```

```
[1] 247379
```

```
dataAB_num[] %>% is.na() %>% colSums()
```

id		host_id
0		0
latitude		longitude
0		0
accommodates		bathrooms
0		51
bedrooms		beds
63		132
square_feet		guests_included
50197		0
minimum_nights		maximum_nights
0		0
availability_30		availability_60
0		0
availability_90		availability_365
0		0
number_of_reviews		number_of_reviews_ltm
0		0
review_scores_rating		review_scores_accuracy
11161		11197
review_scores_cleanliness		review_scores_checkin
11183		11213
review_scores_communication		review_scores_location
11194		11217
review_scores_value	calculated_host_listings_count	
11216		0
prix		prix_semaine
0		44929
prix_mois		prix_caution
45667		17469
frais_nettoyage		personne_supp
10490		0

```
dataAB_num <- dataAB_num %>% na.omit()
```

```
dim(dataAB_num)
```

```
[1] 121 32
```

Essayons désormais de construire une matrice de corrélation entre les valeurs numériques du jeu de données afin de chercher à trouver une corrélation potentielles entre variables quantitatives auxquels nous n'aurions pas pensé directement.

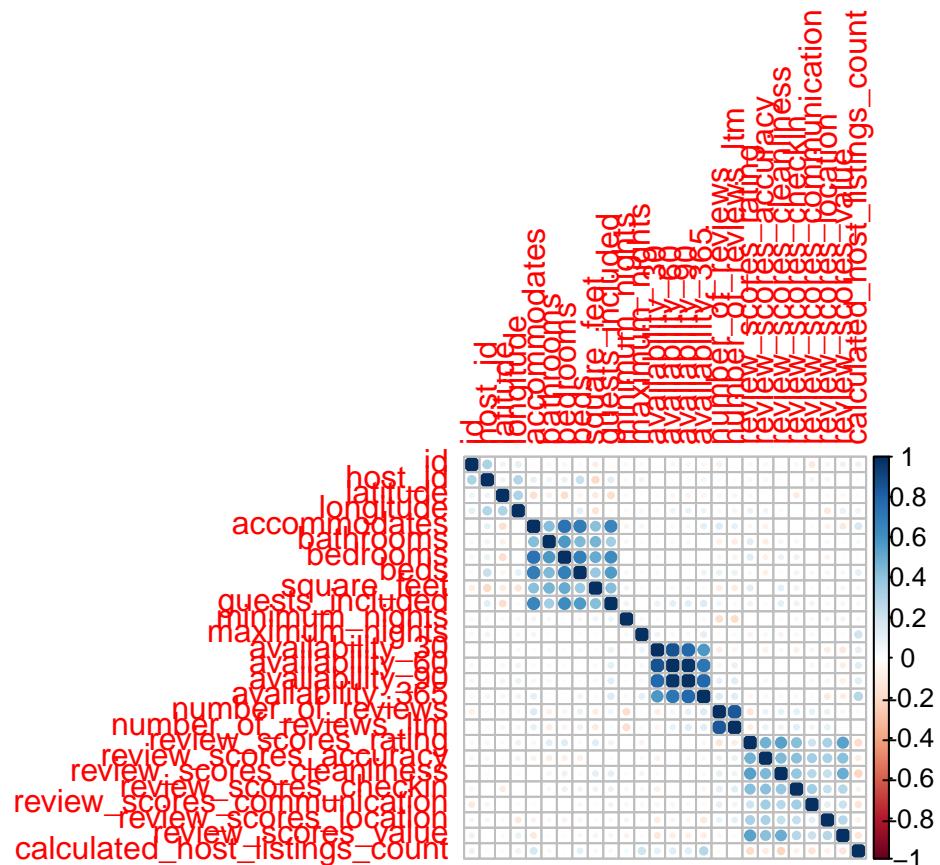
```

dataAB1 %>% select_if(is.numeric) -> data_num
data_num %>% head()
data_num[] %>% is.na() %>% any()
data_num[] %>% is.na() %>% sum()
data_num[] %>% is.na() %>% colSums()
data_num <- data_num %>% na.omit()
data_num.cor = data_num %>% cor()

data_num.rcorr = data_num %>% as.matrix() %>% rcorr()
data_num.rcorr %>% head()

data_num.cor %>% corrplot()

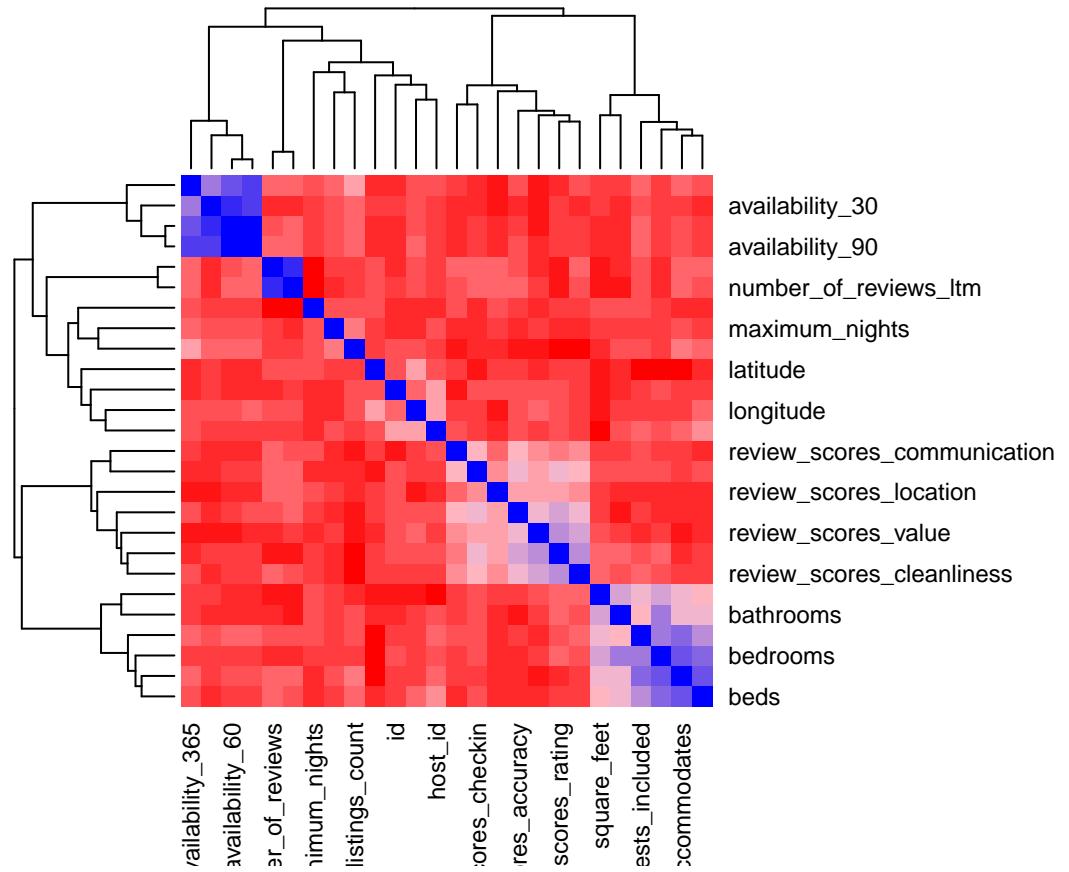
```



```

palette = colorRampPalette(c("red", "pink", "blue"))(20)
heatmap(x = data_num.cor, col = palette, symm = TRUE)

```



Evolution de la popularité de Airbnb au cours du temps

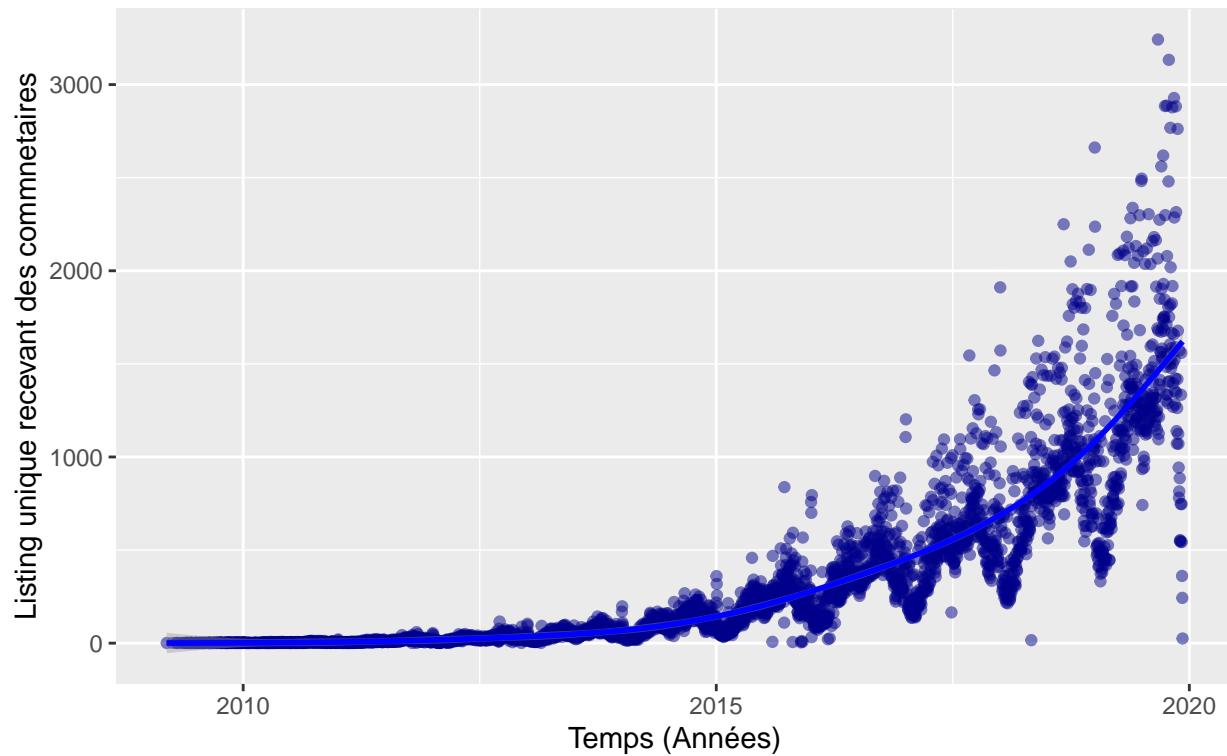
Nous n'étions pas en mesure d'obtenir les données sur le nombre de réservations faites sur Airbnb sur la période étudiée. Pour étudier cela, nous avons du faire preuve d'un peu plus de créativité. Nous avons utilisé la variable "number of reviews" comme une variable non négligeable de la demande. En ce qui concerne l'entreprise, environ la moitié des hôtes écrivent des commentaires. C'est ainsi qu'en étudiant l'évolution du nombre de commentaires que nous pouvons avoir et formuler une estimation pertinente de la demande et de l'évolution de la popularité de ce site.

```
reviewsNum <- reviews %>% group_by(date = reviews$date) %>% summarise(number = n())

ggplot(reviewsNum, aes(date, number)) +
  geom_point(na.rm=TRUE, color = "darkblue", alpha=0.5) +geom_smooth(color = "blue")+
  ggtitle("Evolution de la popularité de Airbnb (2008-2019)", subtitle = "Nombre de commentaires des hôtes") +
  labs(x = "Temps (Années)", y = "Listing unique recevant des commentaires") +
  theme(plot.title = element_text(face = "bold")) +
  theme(plot.subtitle = element_text(face = "bold", color = "grey")) +
  theme(plot.caption = element_text(color = "grey"))
```

Evolution de la popularité de Airbnb (2008–2019)

Nombre de commentaires des hôtes



2. Analyse descriptive (et analyse des variables) du jeu de données

Description des données

```
dataABNB %>% str
```

```
Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 50583 obs. of 51 variables:
 $ id                               : num 2595 3831 5099 5121 5178 ...
 $ name                             : chr "Skylit Midtown Castle" "Cozy Entire Floor of Brownstone" "La...
 $ host_id                           : num 2845 4869 7322 7356 8967 ...
 $ host_name                         : chr "Jennifer" "LisaRoxanne" "Chris" "Garon" ...
 $ host_since                        : Date, format: "2008-09-09" "2008-12-07" ...
 $ host_response_time                : chr "within a day" "within an hour" "N/A" "within a few hours" ...
 $ host_response_rate               : chr "85%" "100%" "N/A" "100%" ...
 $ host_is_superhost                : logi FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ host_verifications              : chr "[email]", 'phone', 'reviews', 'offline_government_id', 'kba' ...
 $ neighbourhood_cleansed           : chr "Midtown" "Clinton Hill" "Murray Hill" "Bedford-Stuyvesant" ...
 $ neighbourhood_group_cleansed    : chr "Manhattan" "Brooklyn" "Manhattan" "Brooklyn" ...
 $ latitude                          : num 40.8 40.7 40.7 40.7 40.8 ...
 $ longitude                         : num -74 -74 -74 -74 -74 ...
 $ room_type                         : chr "Entire home/apt" "Entire home/apt" "Entire home/apt" "Private...
 $ accommodates                      : num 1 3 2 2 2 1 3 2 2 1 ...
 $ bathrooms                         : num 1 1 1 NA 1 1 1 1 1 1.5 1 ...
 $ bedrooms                          : num 0 1 1 1 1 1 1 1 1 1 ...
 $ beds                             : num 1 4 1 1 1 1 2 1 0 1 ...
 $ bed_type                           : chr "Real Bed" "Real Bed" "Real Bed" "Futon" ...
 $ amenities                          : chr "{TV,Wifi,\\"Air conditioning\\",Kitchen,\\"Paid parking off pre...
 $ square_feet                       : num NA 500 NA NA NA NA NA NA NA NA ...
 $ guests_included                   : num 1 1 2 1 1 1 2 1 1 1 ...
 $ minimum_nights                     : num 10 1 3 29 2 2 1 3 4 180 ...
 $ maximum_nights                    : num 1125 730 21 730 14 ...
 $ availability_30                  : num 1 1 19 30 3 0 1 10 0 29 ...
 $ availability_60                  : num 1 1 19 60 12 0 1 10 0 59 ...
 $ availability_90                  : num 1 1 19 90 40 0 1 10 0 89 ...
 $ availability_365                 : num 1 1 19 365 242 0 1 10 0 271 ...
 $ calendar_last_scraped            : Date, format: "2019-12-07" "2019-12-07" ...
 $ number_of_reviews                 : num 48 295 78 49 454 118 161 204 175 27 ...
 $ number_of_reviews_ltm             : num 7 75 8 0 47 0 9 36 13 0 ...
 $ first_review                      : Date, format: "2009-11-21" "2014-09-30" ...
 $ last_review                        : Date, format: "2019-11-04" "2019-11-22" ...
 $ review_scores_rating              : num 94 90 90 90 84 98 94 97 94 97 ...
 $ review_scores_accuracy            : num 9 9 10 8 9 10 10 10 10 10 ...
 $ review_scores_cleanliness        : num 9 9 9 8 7 10 9 10 10 9 ...
 $ review_scores_checkin             : num 10 10 10 10 9 10 10 10 10 9 ...
 $ review_scores_communication      : num 10 9 10 10 9 10 10 10 10 10 ...
 $ review_scores_location            : num 10 10 10 9 10 10 9 10 10 10 ...
 $ review_scores_value               : num 9 9 9 9 8 10 9 10 10 9 ...
 $ instant_bookable                 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ cancellation_policy               : chr "strict_14_with_grace_period" "moderate" "moderate" "strict_1...
 $ require_guest_profile_picture     : logi TRUE FALSE TRUE FALSE FALSE TRUE ...
 $ require_guest_phone_verification: logi TRUE FALSE TRUE FALSE FALSE TRUE ...
 $ calculated_host_listings_count   : num 1 1 1 1 1 4 1 3 1 ...
 $ prix                             : num 225 89 200 60 79 79 150 85 89 120 ...
 $ prix_semaine                      : num 1995 575 NA NA 470 ...
```

```

$ prix_mois : num NA 2100 NA NA NA NA NA NA 3600 ...
$ prix_caution : num 350 500 300 450 NA NA 0 200 200 500 ...
$ frais_nettoyage : num 95 NA 125 0 15 NA 40 0 67 150 ...
$ personne_supp : num 0 0 100 30 12 0 10 0 22 10 ...

```

Le jeu de données comprend deux tableaux de données principaux :

- **listings** - Il s'agit d'un listing détaillé riche de 106 variables. Nous avons décidé - voir ci-dessus-d'en enlever plusieurs qui ne paraissaient pas pertinentes en vue du projet d'analyse de données. La majeure partie des attributs qui nous avons utilisé pour l'analyse sont **price** (variable continue), **longitude** (continue), **latitude** (continue), **listing_type** (qualitative), **is_superhost** (qualitative), **neighbourhood** (qualitative), **ratings** (continue).
- **reviews** - Des commentaires détaillés données par les hôtes contenant 6 variables. Les principales variables utilisées sont **date** (datetime), **listing_id** (variable discrète), **reviewer_id** (variable discrète) et **comment** (variable textuelle). Cette dernière variable est relativement pertinente pour faire du Natural Language processing (traitement automatique du langage naturel) à travers des méthodes de WordClouds...

Une prise de vue rapide au sein des données nous montre que:

- 50599 listing uniques à NYC au total. La première location Airbnb à New York City remonte en Avril 2008, dans la quartier de Harlem à Manhattan.
- Plus d'1.2 Millions de commentaires ont été rédigé par les hôtes depuis.
- Le prix des location varie de \$10 par nuit à \$10,000 par nuit. Les locations de \$10.000 se situent à Greenpoint à Brooklyn; Astoria dans le Queens, et le très huppé Upper West Side à Manhattan.

Etude des variables

Nous proposons, au lieu d'étudier nos variables une par une, d'étudier les variables pertinentes en fonction d'une autre.

On pourra, par exemple, regarder le prix et leur fluctuation en fonction de chaque quartier.

Moyenne des prix par quartier puis par arrondissements

```

dataABNB %>% group_by(id, host_id, neighbourhood_group_cleaned, neighbourhood_cleaned) %>% summarise(
  # A tibble: 50,583 x 5
  # Groups:   id, host_id, neighbourhood_group_cleaned [50,583]
    id host_id neighbourhood_group_cleaned neighbourhood_cleaned moy_par_quartier
    <dbl> <dbl> <chr> <chr> <dbl>
  1 2595 2845 Manhattan Midtown 225
  2 3831 4869 Brooklyn Clinton Hill 89
  3 5099 7322 Manhattan Murray Hill 200
  4 5121 7356 Brooklyn Bedford-Stuyvesant 60
  5 5178 8967 Manhattan Hell's Kitchen 79
  6 5203 7490 Manhattan Upper West Side 79
  7 5238 7549 Manhattan Chinatown 150
  8 5441 7989 Manhattan Hell's Kitchen 85
  9 5803 9744 Brooklyn South Slope 89
 10 6090 11975 Manhattan West Village 120
# ... with 50,573 more rows
grouped_data <- group_by(dataABNB, neighbourhood_group_cleaned)

summarise(grouped_data, moy_par_arrond = mean(prix, na.rm = TRUE))

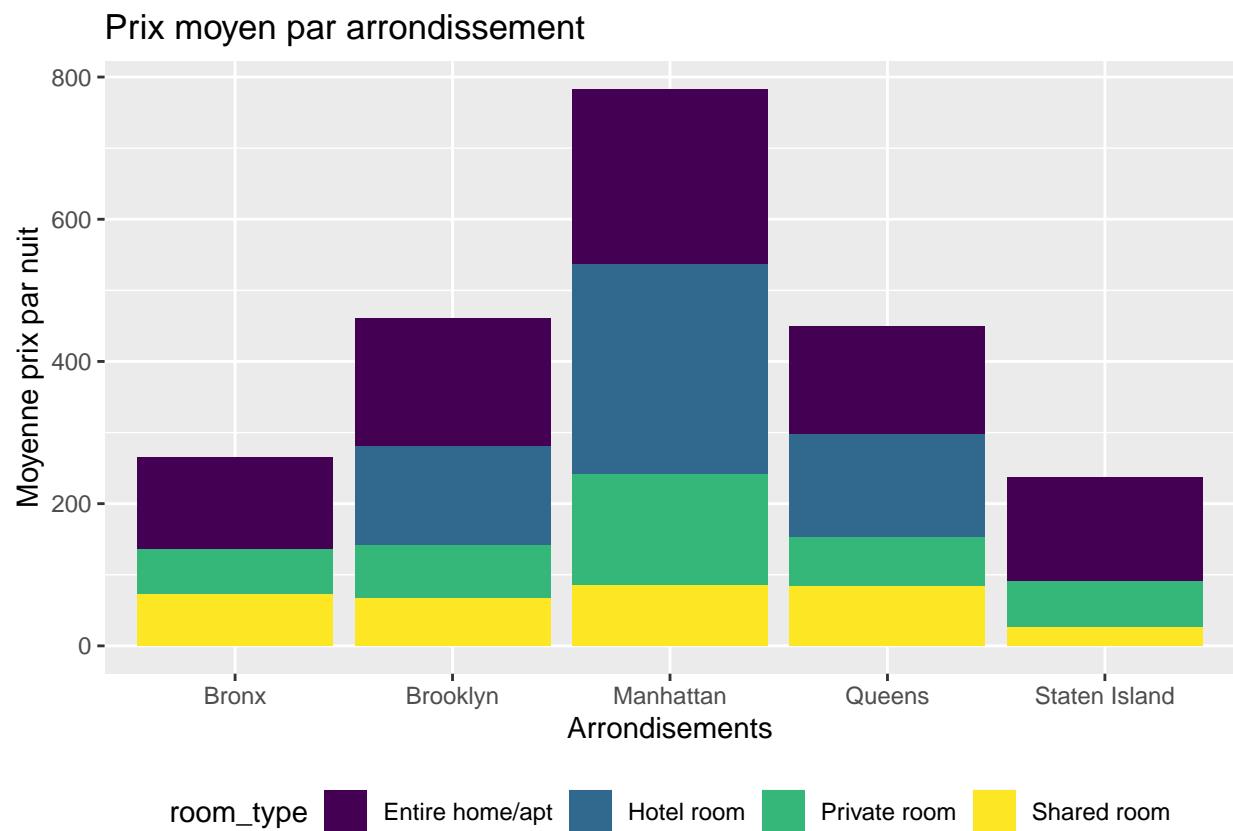
```

```

# A tibble: 5 x 2
  neighbourhood_group_cleansed moy_par_arrond
  <chr>                      <dbl>
1 Bronx                         86.9
2 Brooklyn                       124.
3 Manhattan                      211.
4 Queens                         100.
5 Staten Island                  105.

dataABNB %>%
  group_by(neighbourhood_group_cleansed, room_type) %>%
  summarise(moy_prix = mean(prix)) %>%
  ggplot() +
  geom_bar(mapping = aes(x = neighbourhood_group_cleansed, y = moy_prix, fill = room_type), stat = "identity")
  scale_fill_viridis_d(option = "viridis") +
  theme(legend.position = 'bottom') +
  ggtitle("Prix moyen par arrondissement") +
  labs(x = "Arrondissements", y = "Moyenne prix par nuit")

```



```
dataABNB %>% count(prix < 1250)
```

```

# A tibble: 2 x 2
`prix < 1250`      n
<lgl>              <int>
1 FALSE             230
2 TRUE              50353

```

On remarque qu'en moyenne le prix par nuit d'un Airbnb est le moins cher dans le Bronx, et le plus cher à

Manhattan, il est quasiment le même pour les arrondissements de Brooklyn, Staten Island et du Queens. De plus, on voit que le prix d'une nuit des Airbnb sur New York sont quasiment tous en dessous de 1 250\$, toutefois il faudrait considérer les valeurs extrême tel un AirBNB à 10,000\$ qui n'est pas à la porté de tout les consommateurs.

Comparaison entre prix par nuit en fonction des pieds carrés

```
grouped_data <- group_by(dataABNB, neighbourhood_group_cleaned)

grouped_data %>% summarise(moy_prix = mean(prix, na.rm = TRUE),
                             moy_square_feet = mean(square_feet, na.rm = TRUE),
                             moy_metre_carre = moy_square_feet*0.09290304,
                             prix_metre_carre = moy_prix / moy_metre_carre)
```

```
# A tibble: 5 x 5
  neighbourhood_group... moy_prix moy_square_feet moy_metre_carre prix_metre_carre
  <chr>              <dbl>        <dbl>            <dbl>           <dbl>
1 Bronx             86.9         197.            18.3            4.74
2 Brooklyn          124.         727.            67.6            1.84
3 Manhattan          211.         720.            66.8            3.16
4 Queens             100.         562.            52.2            1.92
5 Staten Island     105.         811.            75.3            1.40
```

Nous observons un tableau concernant les arrondissements de New York, où l'on a pour chaque arrondissement la moyenne du prix par nuit des Airbnb, la moyenne en pieds carré puis convertit en mètres carré et enfin le prix du mètre carré (qui ici est calculé à partir du prix par nuit).

Les appartements proposés par les hôtes sont plutôt spacieux au niveaux de chaque arrondissements mais bénéficie d'un très bon rapport qualité/prix. Un détail se porte sur les mesures de chambres partagés qui sont disproportionnées. On peut déduire que certain propriétaire mentionne la superficie de l'appartement et non de la chambre, ce qui nous laisse penser que nos hôtes procède à des techniques commerciale pour mieux attirer les éventuels touristes (établir peut être un meilleur contrôle et des sanctions restrictives en cas de fausse publicité).

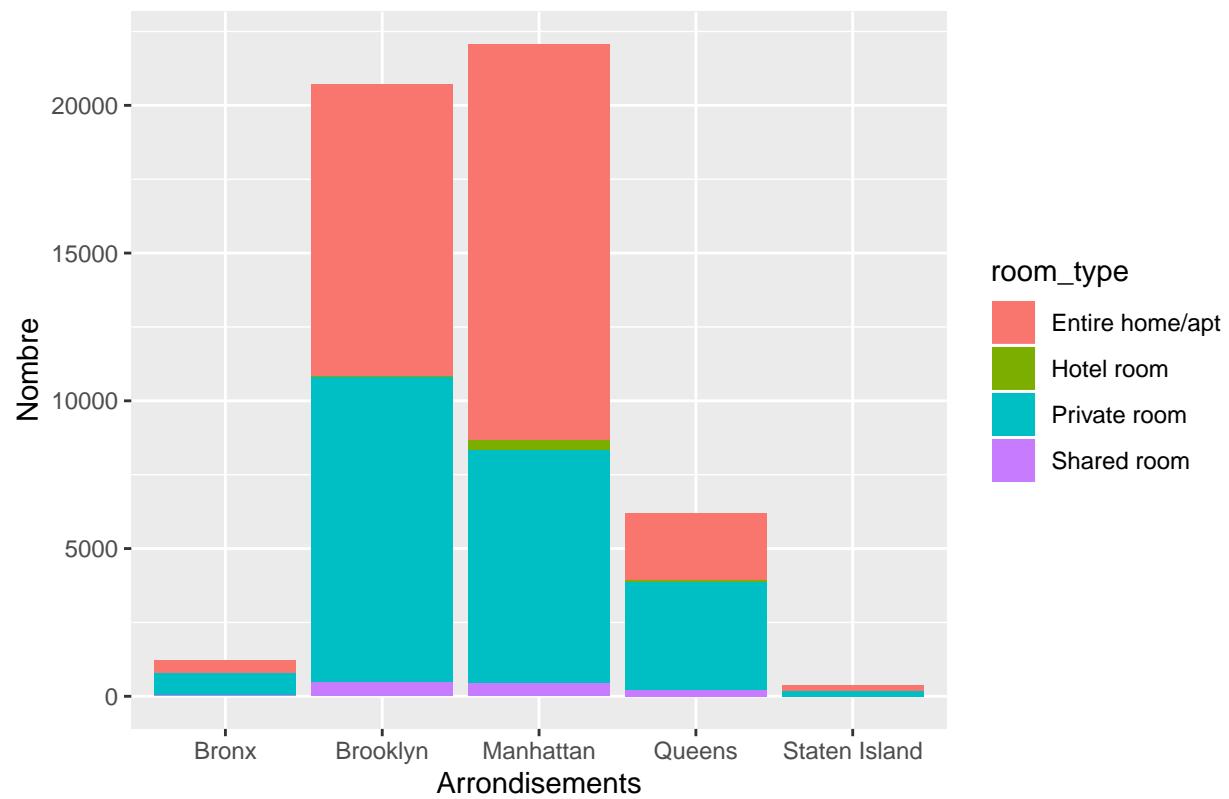
Etude de la variable neighbourhood_group_cleaned :

```
dataABNB %>% count(neighbourhood_group_cleaned)

# A tibble: 5 x 2
  neighbourhood_group_cleaned     n
  <chr>                      <int>
1 Bronx                       1215
2 Brooklyn                     20729
3 Manhattan                    22065
4 Queens                        6196
5 Staten Island                 378

ggplot(dataABNB) +
  geom_bar(mapping = aes(x = neighbourhood_group_cleaned, fill = room_type)) +
  ggtitle("Graphique du nombre de type de référencement par arrondissements") +
  labs(x = "Arrondissements", y = "Nombre")
```

Graphique du nombre de type de référencement par arrondissements



On voit bien que New York possède 5 arrondissements : Bronx, Brooklyn, Manhattan, Queens et Staten Island. De plus, on remarque que les locations se concentre beaucoup plus à Manhattan et Brooklyn, et un peu dans le Queens.

Nous proposons donc de se concentrer plutôt sur l'étude des arrondissements de Manhattan et Brooklyn.

3. Etude Manhattan/Brooklyn

Création des data set

```
datamanh <- dataABNB %>% filter(neighbourhood_group_cleansed == 'Manhattan')
databrook <- dataABNB %>% filter(neighbourhood_group_cleansed == 'Brooklyn')
data_man_bro <- dataABNB %>% filter(neighbourhood_group_cleansed == c('Manhattan', 'Brooklyn'))

#Si vous souhaitez voir le début des data set
#head(datamanh)
#head(databrook)
#head(data_man_bro)
```

Etude du prix par nuit d'un Airbnb

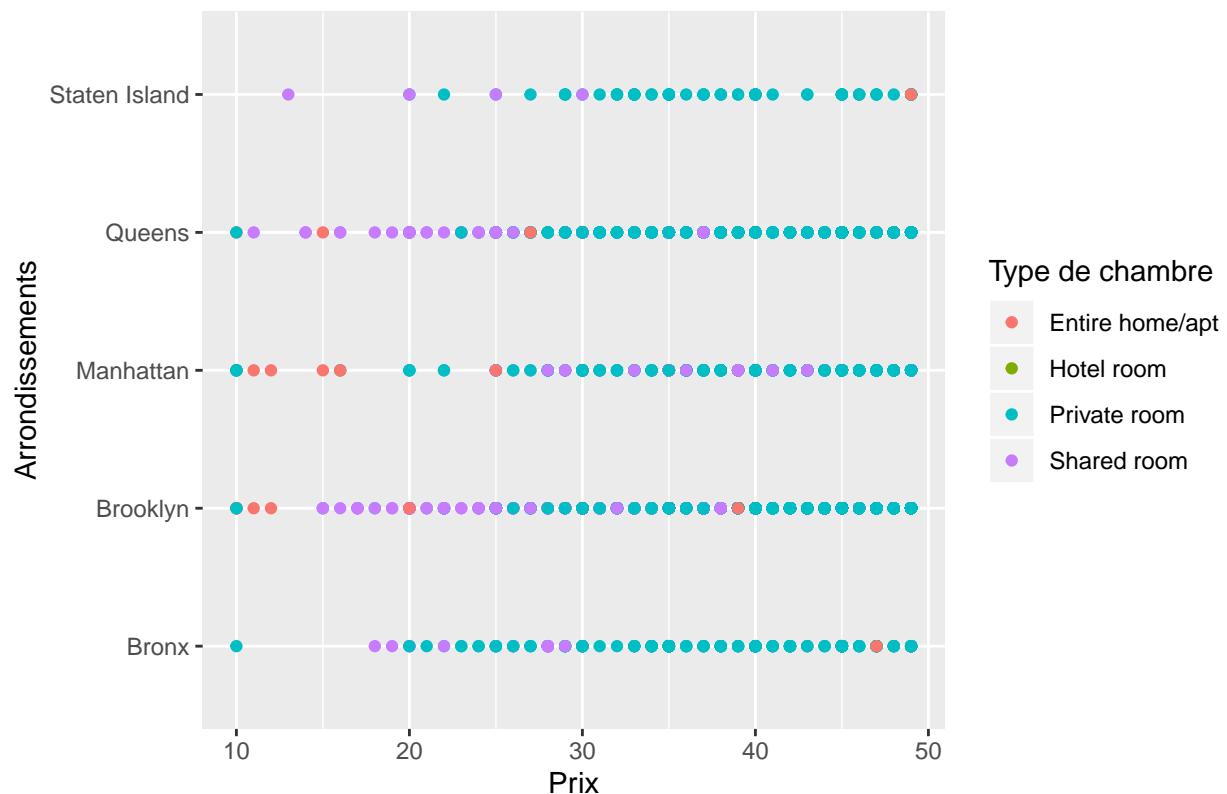
Avant de commencer l'étude sur Manhattan et celle sur Brooklyn, nous allons étudier en profondeur les prix par nuit des Airbnb.

On décide de plus, de considérer que le prix d'un Airbnb est très peu cher quand la nuit coûte moins de 50\$, est pas très cher quand la nuit coûte moins de 200\$, il est de prix moyen en dessous de 500\$ la nuit et plutôt cher pour la nuit à plus de 500\$.

```
verycheapAB <- dataABNB %>% filter(prix < 50)
cheapAB <- dataABNB %>% filter(between(prix, 50, 200))
mediumAB <- dataABNB %>% filter(between(prix, 200, 500))
expensiveAB <- filter(dataABNB, prix >= 500)

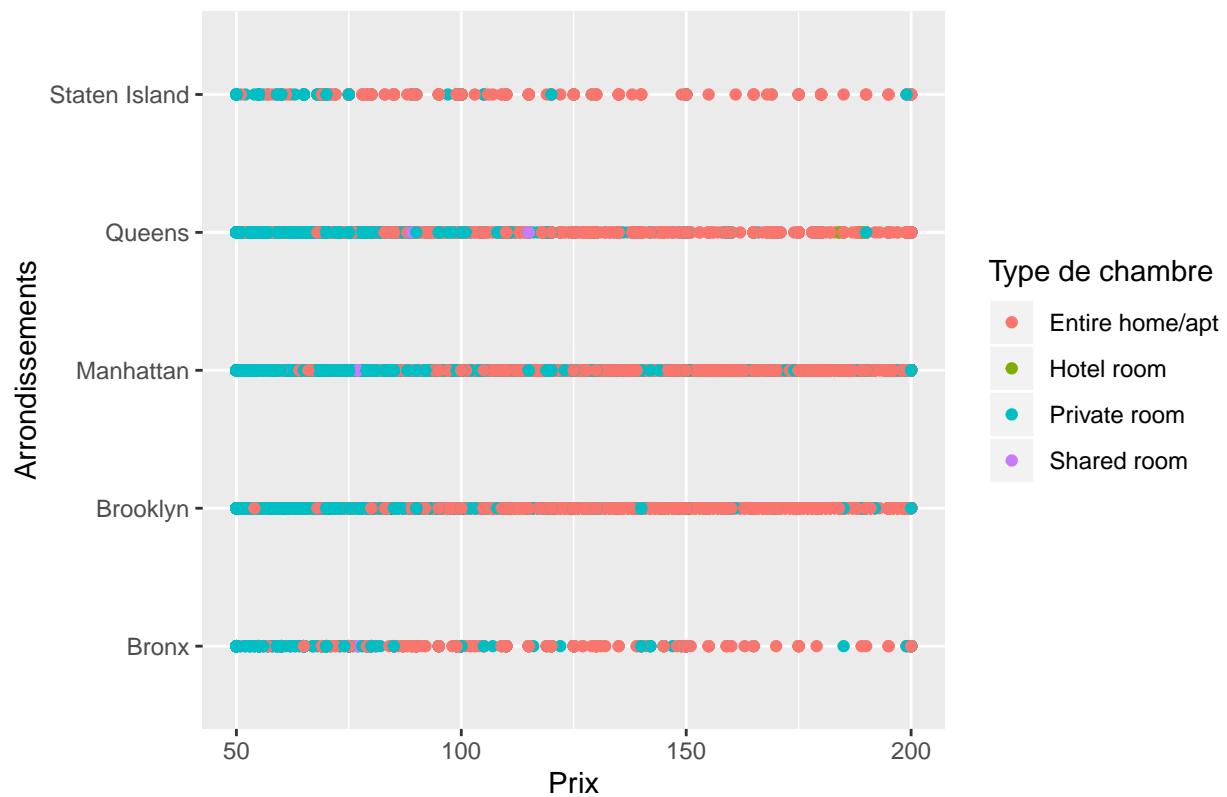
ggplot(verycheapAB) +
  geom_point(mapping = aes(x = prix, y = neighbourhood_group_cleansed, color = room_type)) +
  labs(x = 'Prix', y = 'Arrondissements', colour = 'Type de chambre') +
  ggtitle('Appartements les moins cher (prix < 50$)')
```

Appartements les moins cher (prix < 50\$)



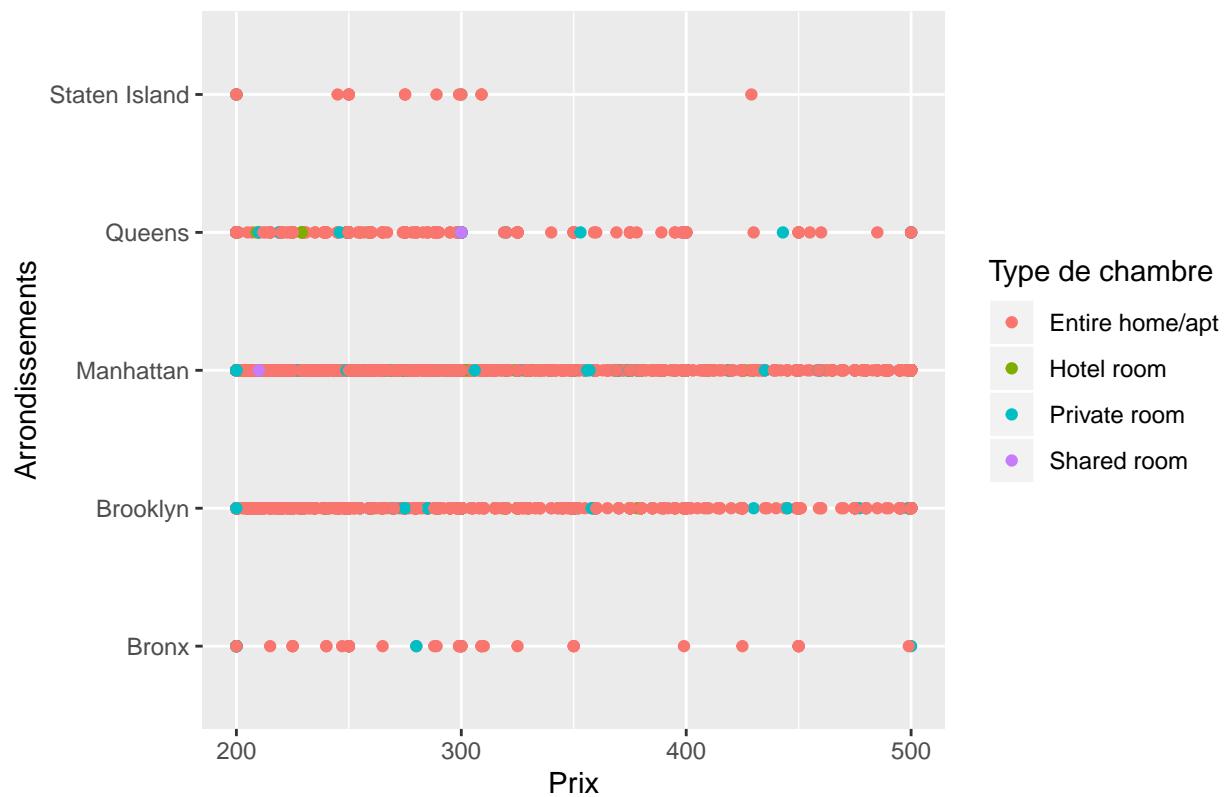
```
ggplot(cheapAB) +  
  geom_point(mapping = aes(x = prix, y = neighbourhood_group_cleansed, color = room_type)) +  
  labs(x = 'Prix', y = 'Arrondissements', colour = 'Type de chambre') +  
  ggtitle('Appartements à prix abordable (entre 50 et 200$)')
```

Appartements à prix abordable (entre 50 et 200\$)



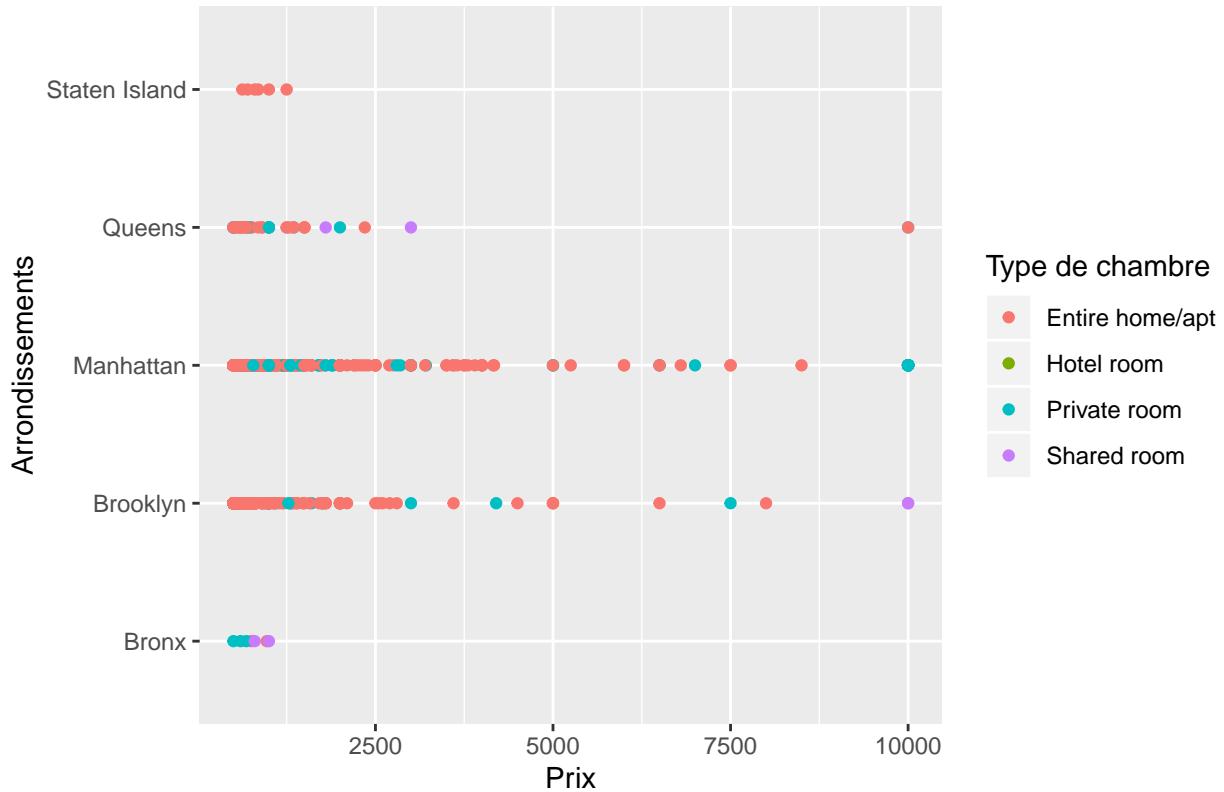
```
ggplot(mediumAB) +  
  geom_point(mapping = aes(x = prix, y = neighbourhood_group_cleansed, color = room_type)) +  
  labs(x = 'Prix', y = 'Arrondissements', colour = 'Type de chambre') +  
  ggtitle('Appartements moyennement cher (entre 200 et 500$)')
```

Appartements moyennement cher (entre 200 et 500\$)



```
ggplot(expensiveAB) +  
  geom_point(mapping = aes(x = prix, y = neighbourhood_group_cleansed, color = room_type)) +  
  labs(x = 'Prix', y = 'Arrondissements', colour = 'Type de chambre') +  
  ggtitle('Appartements les plus chers (prix > 500$)')
```

Appartements les plus chers (prix > 500\$)



Au vu des graphes, on remarque une distinction très claire à 100\$; peu importe l'arrondissement dans lequel on se trouve il apparaît qu'un client ne pourra louer qu'une chambre privée. A l'inverse, on voit qu'un client avec un budget supérieur ou égal à 100\$ par nuit peut se permettre un logement complet.

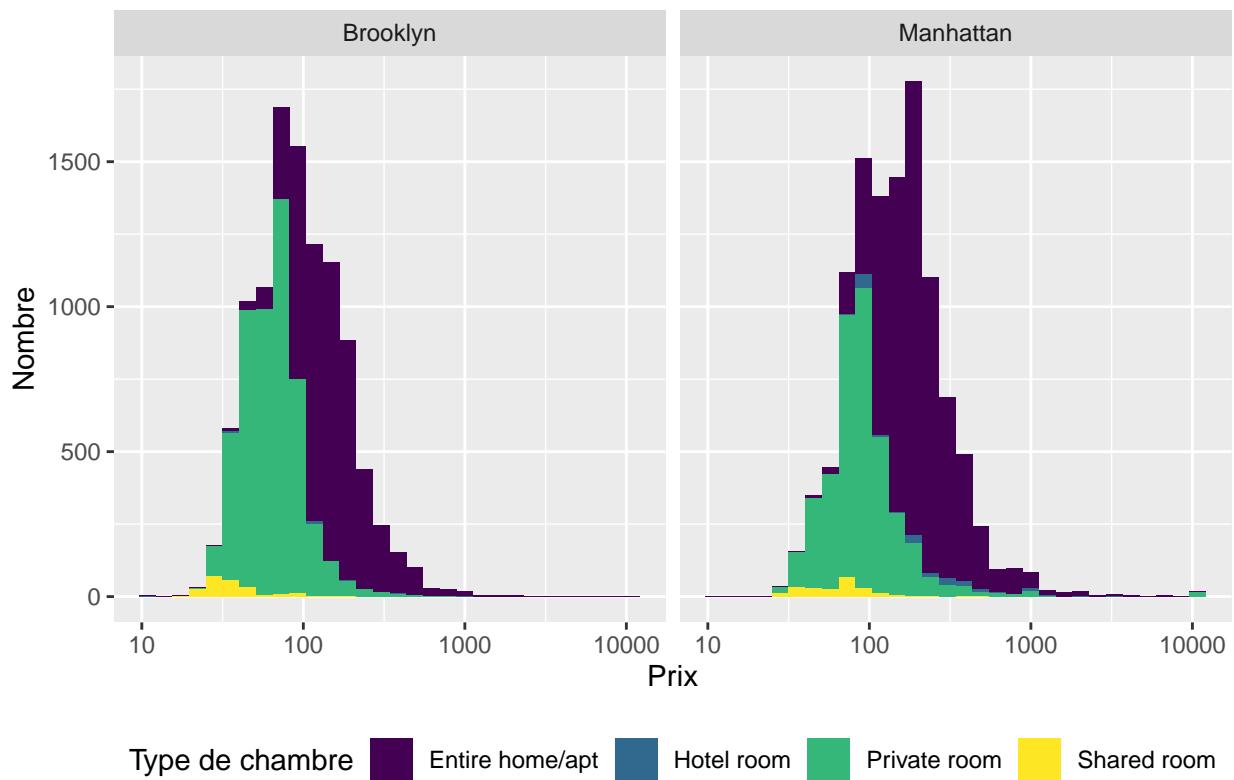
Toutefois le plus intéressants pour un consommateur moyen seraient de chercher sur un intervalle de prix plus large comme au niveau des prix abordables afin d'espérer un logement pouvant satisfaire des critères de satisfactions.

Prix moyen par nuit des locations

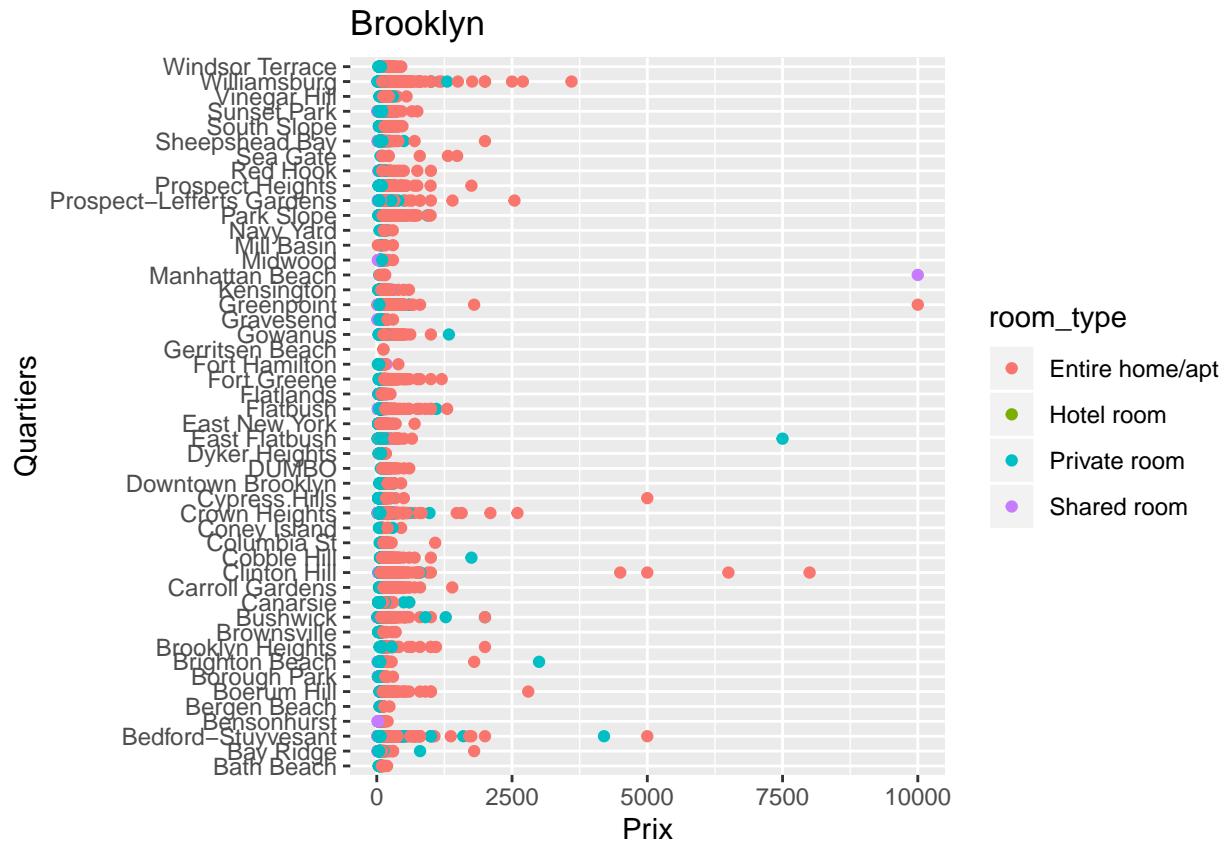
```
ggplot(data_man_bro, mapping = aes( x= prix, fill = room_type))+
  geom_histogram()+
  scale_x_log10()+
  scale_fill_viridis_d(option = "viridis")+
  ggtitle("Histogramme des prix en fonction des arrondissements")+
  labs(x = "Prix", y = "Nombre", fill = 'Type de chambre') +
  theme(legend.position = 'bottom') +
  facet_grid(~neighbourhood_group_cleansed)

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

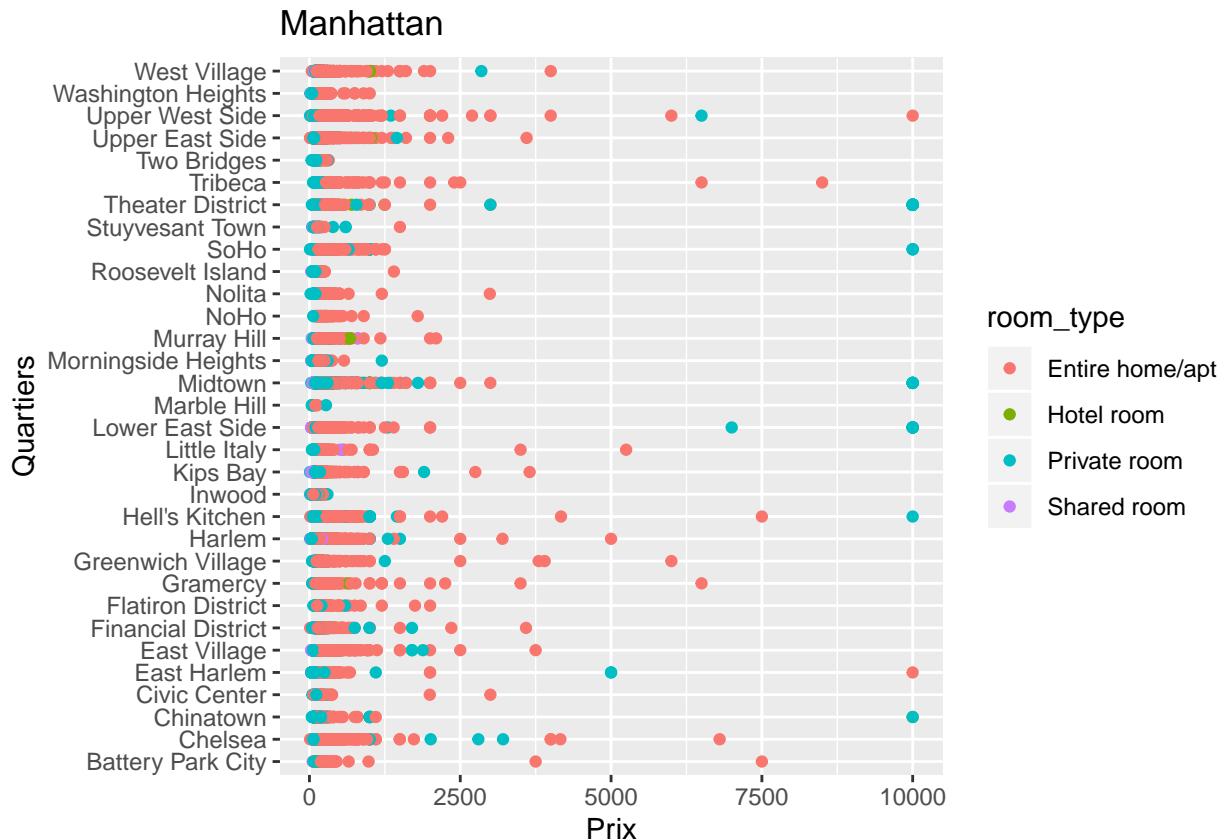
Histogramme des prix en fonction des arrondissements



```
ggplot(data = databrook, mapping = aes(y = prix, x = neighbourhood_cleanse)) +  
  geom_point(mapping = aes(color = room_type)) +  
  labs(x = 'Quartiers', y = 'Prix') +  
  ggtitle('Brooklyn') +  
  coord_flip()
```



```
ggplot(data = datamanh) +
  geom_point(mapping = aes(x = prix, y = neighbourhood_cleanse, color = room_type)) +
  labs( x = 'PrixC', y = 'Quartiers') +
  ggtitle('Manhattan')
```



Manhattan

```
datamanh %>%
  group_by(neighbourhood_cleansed) %>%
  summarise(moy_prix = mean(prix, na.rm = TRUE),
            nbr_location = n())
```

```
# A tibble: 32 x 3
  neighbourhood_cleansed moy_prix nbr_location
  <chr>                  <dbl>      <int>
1 Battery Park City       361.        72
2 Chelsea                  242.      1161
3 Chinatown                 226.       376
4 Civic Center                276.       44
5 East Harlem                 142.      1155
6 East Village                 183.      1883
7 Financial District          213.       732
8 Flatiron District           317.       75
9 Gramercy                   244.       354
10 Greenwich Village          261.       371
# ... with 22 more rows
```

Nous observons un tableau comprenant les quartiers de Manhattan, avec pour chaque quartier sa moyenne de prix pour une nuit et son nombre de location.

On remarque que le nombre de location est plus élevé dans les quartiers où le prix moyen d'une nuit est entre 100 et 250\$. Les locations qui coûtent le plus cher se situent à Theater District et à Tribeca, concernant

celles qui coûtent le moins cher elles se trouvent à Marble Hill et Washington Heights.

Brooklyn

```
databrook %>%
  group_by(neighbourhood_cleaned) %>%
  summarise(
    moy_prix = mean(prix, na.rm = TRUE),
    nbr_location = n())

# A tibble: 48 x 3
  neighbourhood_cleaned moy_prix nbr_location
  <chr>                 <dbl>        <int>
1 Bath Beach             89.3          29
2 Bay Ridge               101.          153
3 Bedford-Stuyvesant     107.         3969
4 Bensonhurst              71.5          81
5 Bergen Beach            102.          17
6 Boerum Hill              209.          180
7 Borough Park             61.8          148
8 Brighton Beach            153.          76
9 Brooklyn Heights          206.          150
10 Brownsville              79.7          77
# ... with 38 more rows
```

Nous observons un tableau comprenant les quartiers de Brooklyn, avec pour chaque quartier sa moyenne de prix pour une nuit et son nombre de location.

Contrairement au prix moyen d'une nuit à Manhattan, on voit tout de suite que les prix sont moins élevés à Brooklyn. Cependant, dans le quartier de Manhattan Beach il y a quand même une location qui coûte 10 000\$. On remarque de plus, que tous les prix moyens se situent entre 100 et 200\$.

Comparaison Manhattan/Brooklyn

Dans cette partie nous choisissons de faire différentes études concernant les variables `prix_caution`, `frais_nettoyage`, `personne_supp`, `review_scores_cleanliness`, `review_scores_checkin`, `review_scores_communication`, `review_score_location`; afin de pouvoir comparer nos deux grands arrondissements.

```
group_data <- group_by(data_man_bro, neighbourhood_group_cleaned)

group_data %>% summarise( moy_caution = mean(prix_caution, na.rm = TRUE),
                           moy_frais = mean(frais_nettoyage, na.rm = TRUE),
                           moy_plusUne_persn = mean(personne_supp, na.rm = TRUE))

# A tibble: 2 x 4
  neighbourhood_group_cleaned moy_caution moy_frais moy_plusUne_persn
  <chr>                      <dbl>       <dbl>        <dbl>
1 Brooklyn                     231.        60.8        13.9
2 Manhattan                    333.        80.6        14.4
```

Nous observons un tableau comprenant les quartiers de Brooklyn et Manhattan, avec pour chaque quartier la moyenne de la caution, du prix des frais de nettoyage et celle lorsque une personne est invitée. On remarque la moyenne des frais de nettoyage ainsi le prix de la caution sont plus élevés à Manhattan qu'à Brooklyn (respectivement de 44% et 33%). Il semble en revanche que l'ajout d'une personne ne diffère pas entre les deux arrondissements.

```

group_data %>% summarise( moy_cleanliness = mean(review_scores_cleanliness, na.rm = TRUE),
                            moy_checkin = mean(review_scores_checkin, na.rm = TRUE),
                            moy_communication = mean(review_scores_communication, na.rm = TRUE),
                            moy_location = mean(review_scores_location, na.rm = TRUE))

# A tibble: 2 x 5
#>   neighbourhood_group... moy_cleanliness moy_checkin moy_communicati... moy_location
#>   <chr>                  <dbl>          <dbl>          <dbl>          <dbl>
#> 1 Brooklyn                9.29           9.78          9.77           9.51
#> 2 Manhattan               9.24           9.70          9.72           9.70

```

Nous observons un tableau comprenant les quartiers de Brooklyn et Manhattan, avec pour chaque quartier la note moyenne de l'hôte concernant la propreté, le checkin, la communication et la localisation.

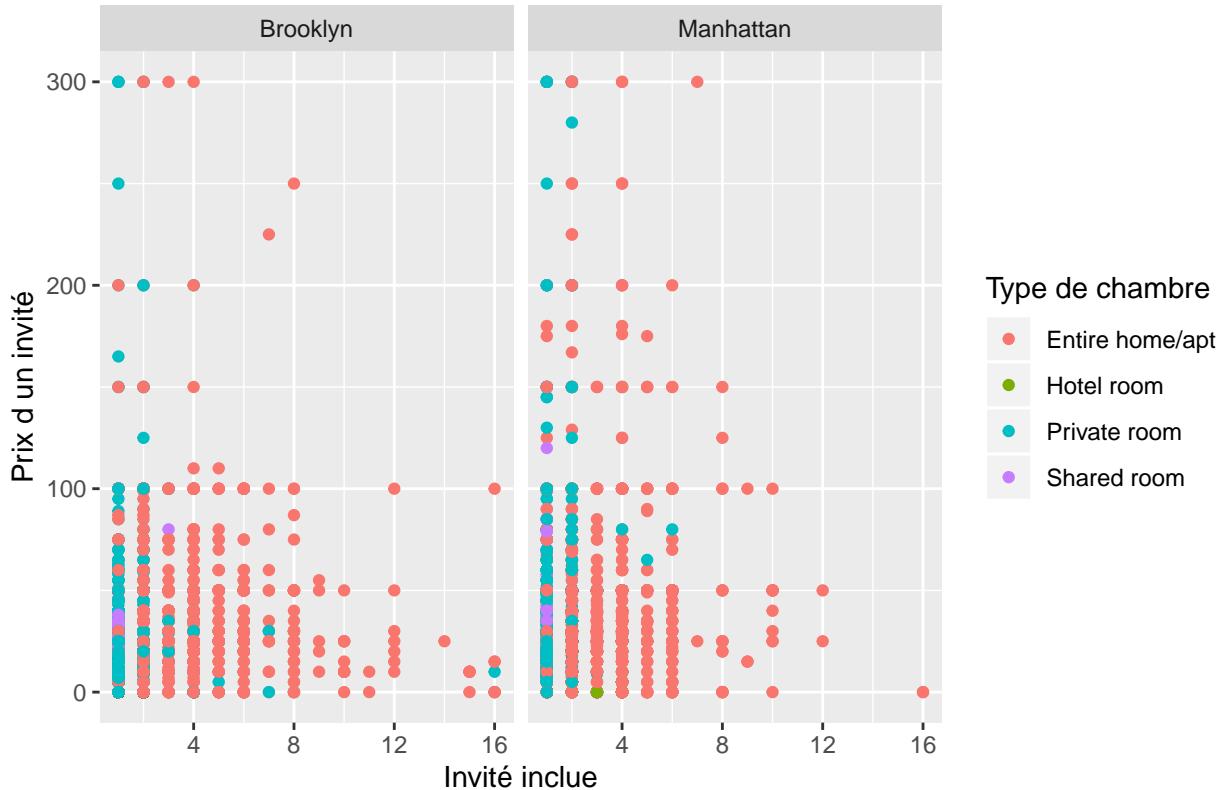
On constate que les notes moyennes des logements concernant la propreté, le checkin process et l'échange hôte/client sont similaires. En outre, il semble que la localisation est légèrement mieux noté à Manhattan.

```

ggplot(data_man_bro) +
  geom_point(mapping = aes(x = guests_included, y = personne_supp, colour = room_type)) +
  labs(x = 'Invité inclue', y = 'Prix d un invité', colour = 'Type de chambre') +
  ggtitle('Nombre d invité possible en fonction du prix d un seul invité') +
  facet_wrap(~neighbourhood_group_cleansed)

```

Nombre d invité possible en fonction du prix d un seul invité



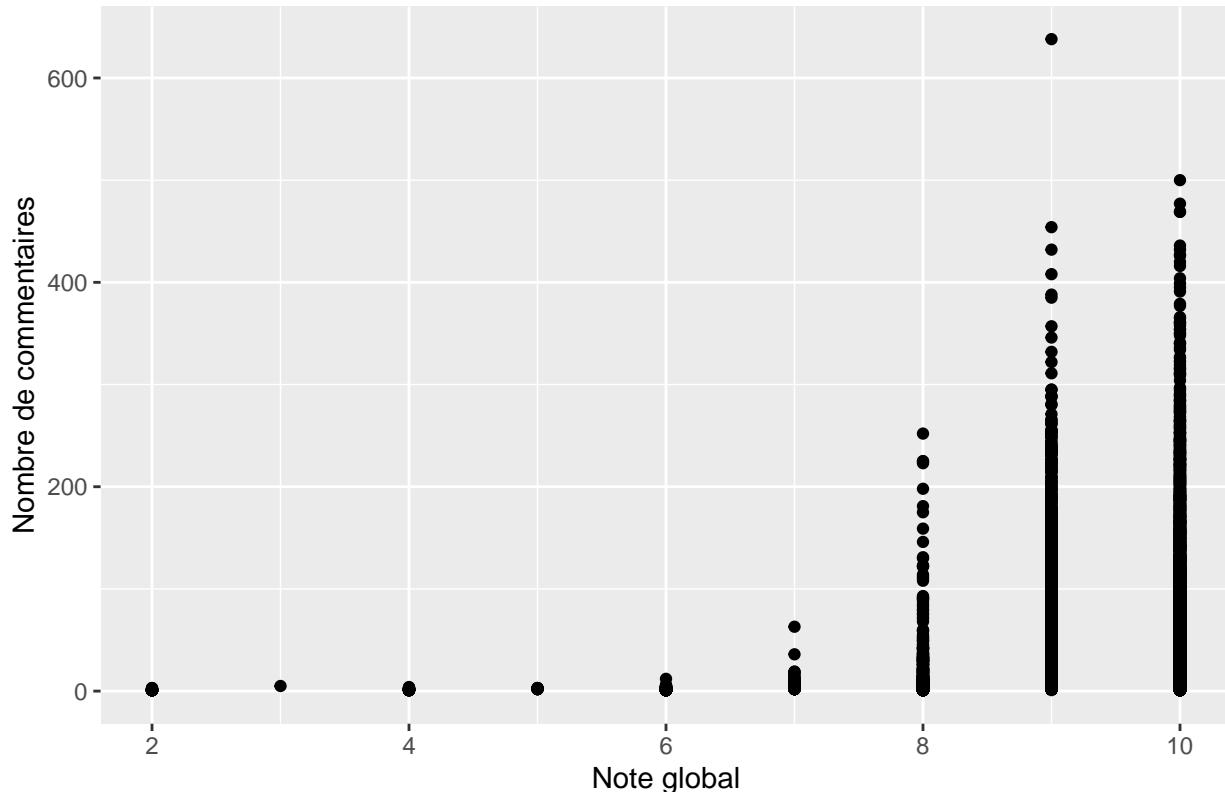
Au vu du graphe ci-dessus, on peut dire qu'à Brooklyn, un grand nombre d'invités par appartement entraîne statistiquement un coût moins élevé par tête. Pour nuancer, on observe que Manhattan dispose d'un plus grand nombre de logements complets à prix très coûteux.

A propos des hôtes

```
dataABNB <- dataABNB %>% filter(prix > 50)

ggplot(data_man_bro, mapping = aes(x = review_scores_accuracy, y = number_of_reviews)) +
  labs( x = 'Note global', y = 'Nombre de commentaires') +
  ggtitle('Note global des hôtes en fonction du nombre de commentaire') +
  geom_point()
```

Note global des hôtes en fonction du nombre de commentaire



Grâce au graphe ci-dessus, on pourra admettre que les hôtes possédant le plus de commentaires sont les mieux notés, toutefois ils sont jugés sur plusieurs critères aux travers de cette note, d'où l'existence de la variable `host_is_superhost` qui détermine si la personne proposant le logement est accueillante ou non.

Comparons un peu nos hôtes sur ces deux arrondissements :

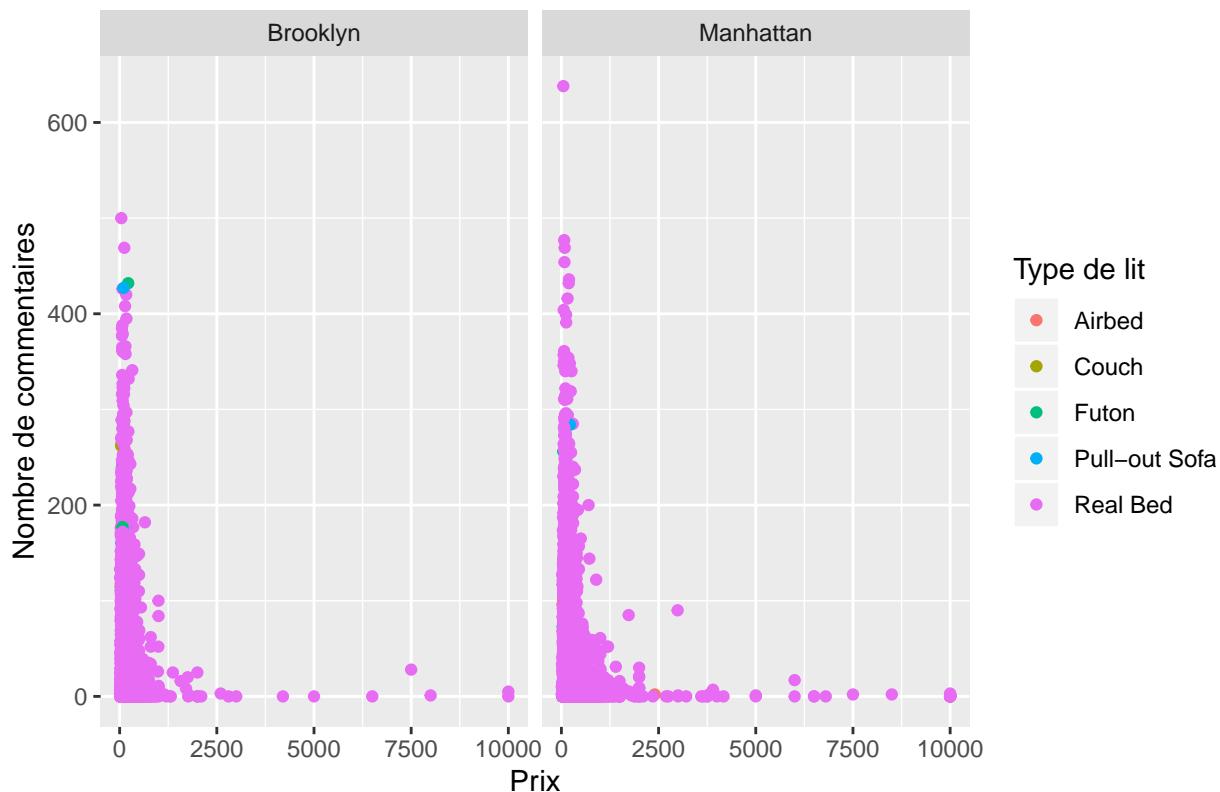
```
ggplot(data_man_bro) +
  geom_point(mapping = aes(y = number_of_reviews, x = prix, colour = host_is_superhost)) +
  labs(x = 'Prix', y = 'Nombre de commentaires', colour = 'Superhôte') +
  ggtitle('Prix par nuit d un Airbnb en fonction du nombre de commentaires et des Superhôtes') +
  facet_wrap(neighbourhood_group_cleansed ~ room_type)
```

Prix par nuit d un Airbnb en fonction du nombre de commentaires et des Si



```
ggplot(data_man_bro) +
  geom_point(mapping = aes(y = number_of_reviews, x = prix, colour = bed_type)) +
  labs(x = 'Prix', y = 'Nombre de commentaires', colour = 'Type de lit') +
  ggtitle('Prix par nuit d un Airbnb en fonction du nombre de commentaires et des types de lit') +
  facet_wrap(~neighbourhood_group_cleansed)
```

Prix par nuit d'un Airbnb en fonction du nombre de commentaires et des types de lits



A première vue, les chambres d'hôtels ont les notes les moins accueillantes sur le marché au niveau des deux arrondissements. Toutefois, on remarque que les propriétaires de chambres privées à **Manhattan** sont plus accueillants que ceux de **Brooklyn**, cependant c'est l'inverse au niveau des appartements loués entièrement. On pourrait décider d'appliquer un test de Student afin de savoir si le nombre d'hôtes les mieux notés sont ceux de Manhattan ou bien de Brooklyn.

Map des airbnbs à Brooklyn et Manhattan

Voir le lien rpubs : [lienrpubs](#)

4. Explorations des données textuelles (Traitement automatique du langage naturel)

Le jeu de données nous offre une somme conséquente de données à exploiter. Toutefois, rien de plus instructif et révélateur que les commentaires des ‘guests’ Airbnb sur leur séjour. Bien exploitées, elles peuvent nous donner de nombreuses informations sur la mentalité des guests, leurs attentes et surtout si ces dernières ont été atteintes. Pour que l’exploration de ces données aient du sens, les commentaires doivent être nettoyés en profondeur et donc faire usage de certaines fonctions et (stopwords) et mots très communs en anglais pouvant faire du bruit statistique, ou encore les guillemets, espaces et pourcentages à retirer.

Echantillonage des données, nettoyage des données et résultats

Les commentaires dans le jeu de données reviews sont supérieurs à 1.2 Millions. Pour poursuivre une analyse textuelle des données, il fallait devoir diviser ces données en ‘Bag of Words’ (Bow) de 20 millions de mots. Pour des raisons de limites computationnelles, nous avons échantillonné les données de façon aléatoire afin que le jeu de données contienne 2.5% des commentaires totaux.

```
#On va faire un échantillonage aléatoire de 30000 commentaires sur les 1 200 000 de commentaires, soit
sampleABNB <- sample(comments, size = 30000)

#Nous allons faire une séparation ou "split" à chaque espace afin de récupérer les mots et créer le bag
splitsampledreviewscolumn <- unlist(strsplit(as.character(sampleABNB), split=" "))
reviewsWordDF <- data.frame("word" = splitsampledreviewscolumn)

wordDF <- reviewsWordDF %>% count(word, sort = TRUE) %>%
ungroup()

#A laide de la librairie "tm" nous pouvons faire l'usage de certaines fonctions tel que les stopwords, la
library("tm")
docs <- Corpus(VectorSource(splitsampledreviewscolumn))
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeWords, stopwords("english"))
docs = tm_map(docs,removePunctuation)
docs <- tm_map(docs, removeWords, c("we","it", "he", "this", "i", "the", "apartment","de", "un","us","w"))

newcorpusdf <- data.frame(text=sapply(docs, identity),
stringsAsFactors=F)
newcorpusdffiltered <- newcorpusdf %>% filter(text != "")
wordDF <- newcorpusdffiltered %>% count(text, sort = TRUE) %>%
ungroup()

library(RColorBrewer)
library(wordcloud)
set.seed(789)
wordcloud(words = wordDF$text,
freq = wordDF$n,
min.freq = 1000,
max.words=500, colors = c("pink","blue", "green", "#add8e6"))
```



Une analyse du word cloud révèle des tendances intéressantes.

Tout d'abord il semble que la localisation est fondamentale, car les termes “location”, “close”, “place”, “subway” ressortent très clairement. Il ressort également que la sécurité est un facteur clé pour les clients, au vu de l'utilisation de termes fréquents tels que “sure”, ou “safe”.

De plus, il semblerait que le terme "kitchen" révèle que les clients ont envie de disposer d'une cuisine afin de se faire à manger dans l'appartement. Il y a pour contraster le terme "restaurants", sûrement pour signifier l'importance de restaurants, voire de bons restaurants dans le quartier de l'appartement.

Enfin, il semble que le confort est primordial. L'utilisation fréquente de termes comme "confortable", "space", "cozy", "friendly", "spacious" et "area" nous indique l'importance du confort. Est-ce si important pour les personnes de se sentir comme chez soi ? "Friendly" et "family" pourraient sûrement apporter une réponse à cette question.

5. Conclusion

Nous sommes arrivé avec de nombreuses questions au début du projet. Nous avons tenté d'étudier les attributs du jeu de données de façon à apporter des réponses.

On remarque que plus nous sommes au centre de NYC (Manhattan et Brooklyn) et plus le prix des logements sont élevés. Sans surprise, le type de logement impact grandement le prix des hébergements. Nous avons par exemple constaté une distinction très nette entre les appartements et les chambres privés. En effet, il paraît quasiment impossible d'avoir un appartement à Manhattan pour moins de 100 \$ par nuit.

Concernant l'évolution de la popularité d'Airbnb à NYC, nous avons dû faire une corrélation entre le nombre de commentaires et l'évolution du temps. Nous avons constaté une hausse très conséquente des commentaires pour chacun des logements proposés et nous pouvons conclure de la forte hausse de la popularité. Nous sommes passées d'environ 3 commentaires début 2008, à presque 1520 commentaires en fin 2019.

La dernière décennie on constate une forte amélioration au niveau de Brooklyn qui est dû au fait de s'ouvrir au tourisme et de proposer une meilleure alternative à Manhattan en ce qui concerne le rapport qualité prix (plus de confort, de calme, proche du centre...).

Après avoir comparé Brooklyn et Manhattan, on constate que Manhattan possède le plus de commentaires, mais aussi une fourchette de prix légèrement plus élevée que Brooklyn. Manhattan reste dominant sur les réservations de logements, quand bien même il s'agit de l'arrondissement le plus cher de NYC. Ce qui peut s'expliquer par le fait que Manhattan est beaucoup plus populaire que les autres arrondissements, et justifie les résultats des demandes principales des touristes.

Concernant le data mining des commentaires, nous avions des questions assez ambitieuses auxquelles n'ont n'avons pu répondre de façon exhaustive. Toutefois, il apparaît que le wordcloud nous a aidé à cibler des termes très fréquemment utilisés qui arrivent à transmettre l'importance des choix des logements pour les clients de la plateforme. Par exemple, la sécurité, ainsi que la proximité des loisirs et restaurants, les équipements des logements tels la cuisine et la salle de bains semblent essentiels.