# Data Analysis Using R Programming- Mini Project Report

**Jeeva Mary Loui**

**Metro College Of Technology**

## Data Analysis Of Airbnb Data using R Programming

## Abstract

Inside Airbnb is an independent, non-commercial set of tools and data that is not associated with or endorsed by Airbnb or any of Airbnb's competitors. The dataset here used is from the Toronto Airbnb data. The project will be dealing with comparison of various factors such as price, neighbourhood, room type ,bed type , amenities provided ,most expensive listing , time of year , time of week , reviews etc.It uses techniques like ggplot ,textmining ,word cloud etc.

**Dataset Source:**

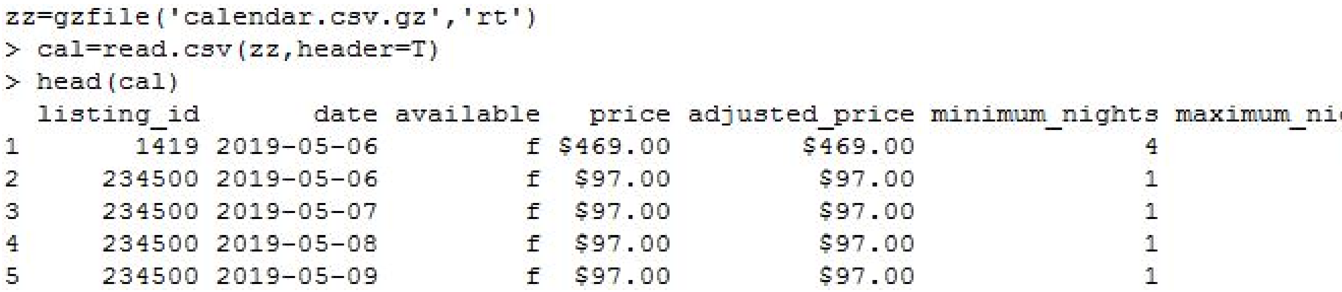
The dataset is from Inside Airbnb.Inside Airbnb is an independent, non-commercial set of tools and data that is not associated with or endorsed by Airbnb or any of Airbnb's competitors.

The Dataset used in this analysis is the Airbnb Dataset of Toronto which has files in csv format with data on listings reviews and calendar.

## THE ANALYSIS

## CALENDAR

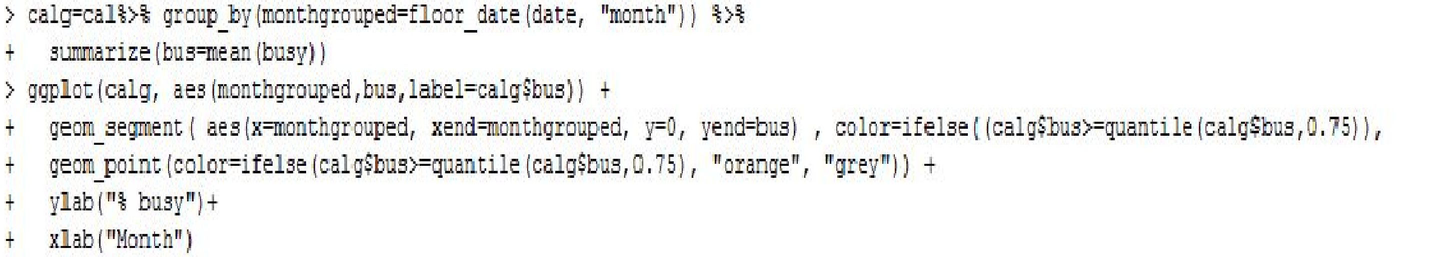
The calendar file which has records of the availability data is unzipped and read to a variable.There are 7410929 rows and 7 columns.In which there are 20303 unique host listings availability data spread across 367days, a year from May 2019 -the last month to May 2020.



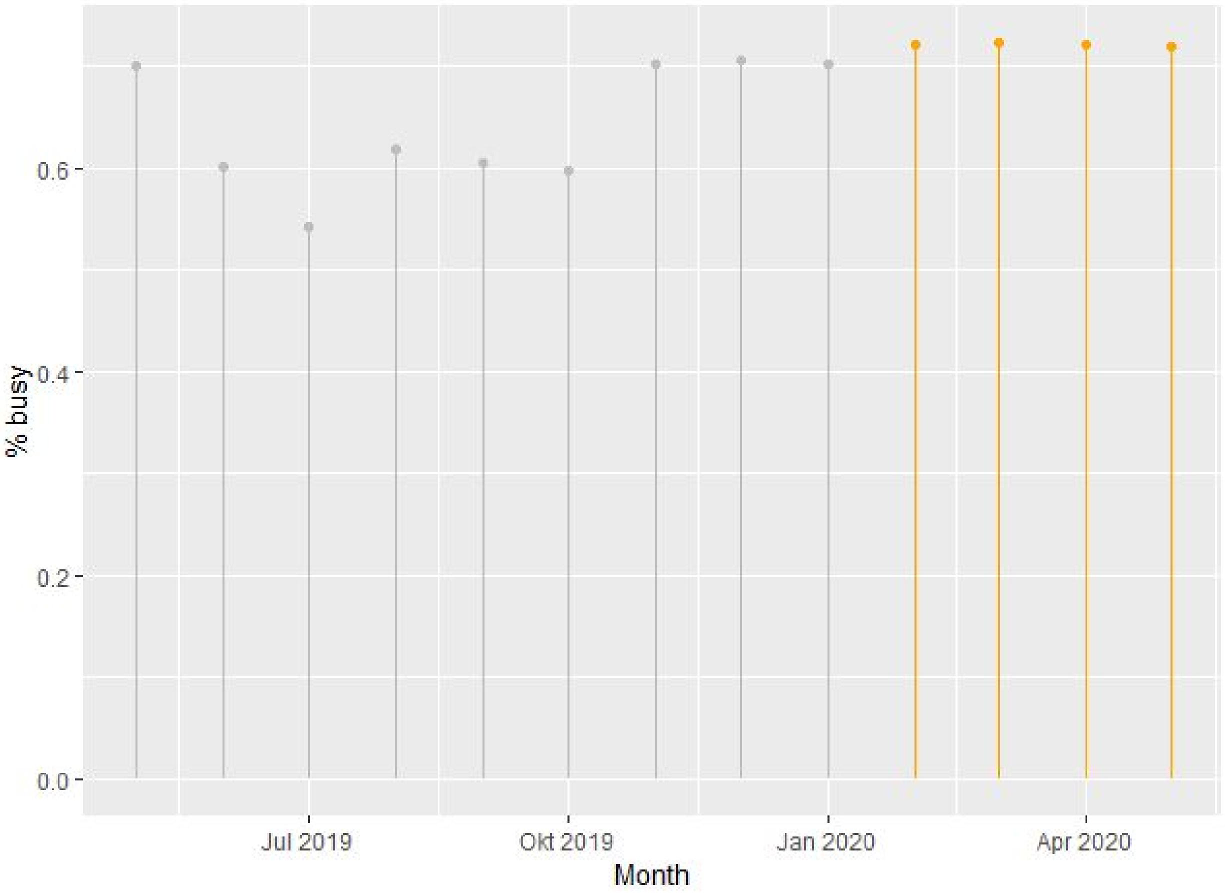
```
zz=gzfile('calendar.csv.gz','rt')
> cal=read.csv(zz,header=T)
> head(cal)
  listing_id       date available   price adjusted_price minimum_nights maximum_ni
1       1419 2019-05-06         f $469.00        $469.00              4
2     234500 2019-05-06         f  $97.00         $97.00              1
3     234500 2019-05-07         f  $97.00         $97.00              1
4     234500 2019-05-08         f  $97.00         $97.00              1
5     234500 2019-05-09         f  $97.00         $97.00              1
```

### #1 How busy is Airbnb host in Toronto?

The busy months of the hosts can be found using lollipop plot.



```
> calg=cal%>% group_by(monthgrouped=floor_date(date, "month")) %>%
+   summarize(bus=mean(busy))
> ggplot(calg, aes(monthgrouped,bus,label=calg$bus)) +
+   geom_segment( aes(x=monthgrouped, xend=monthgrouped, y=0, yend=bus) , color=ifelse((calg$bus>=quantile(calg$bus,0.75)),
+   geom_point(color=ifelse(calg$bus>=quantile(calg$bus,0.75), "orange", "grey")) +
+   ylab("% busy")+
+   xlab("Month")
```

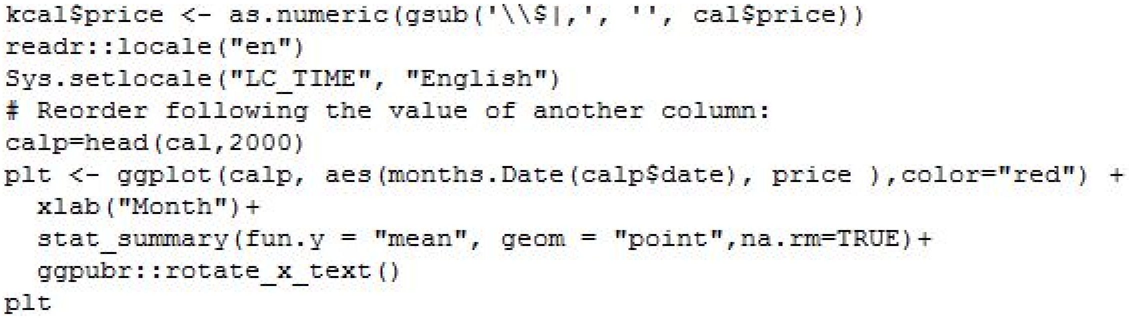


The months from February to June are busy and has low chance  to get Airbnb Rooms.

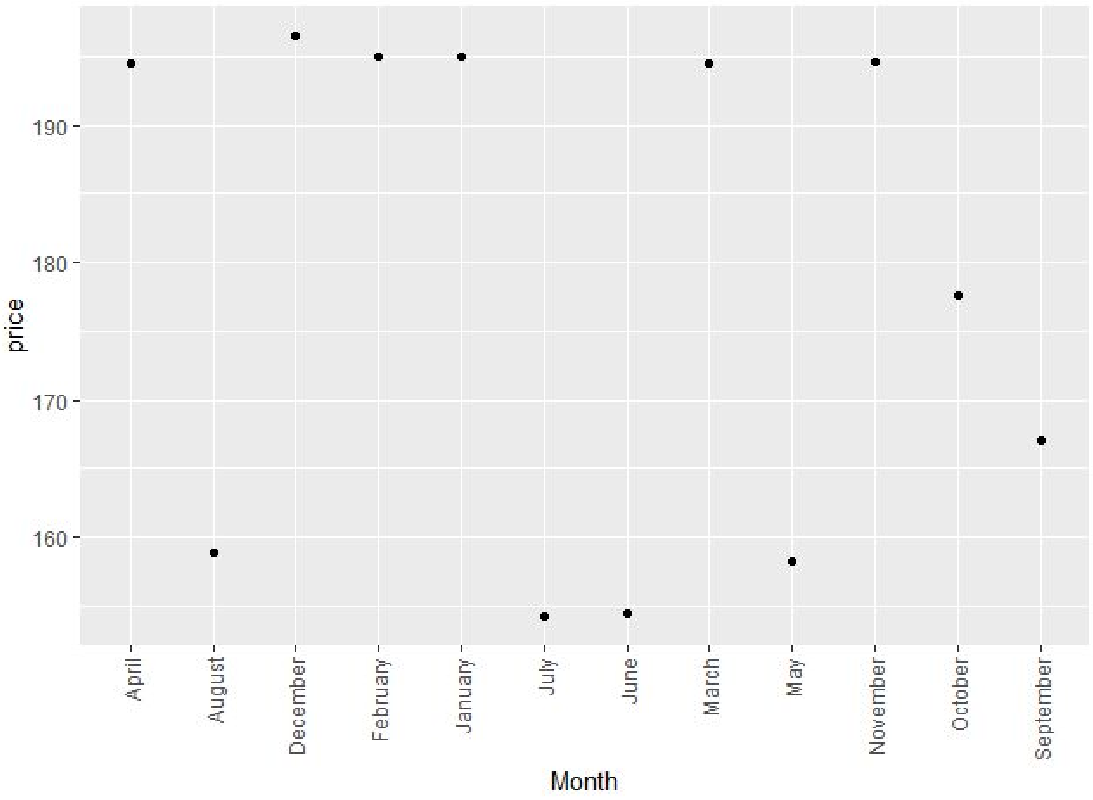
## Price on the Calendar

### #2 How price changes over the year by month?

We remove “$” symbol in price column and convert it to numeric, and convert date to datetime data type. The months of the year and price is evaluated using  dplyr library.



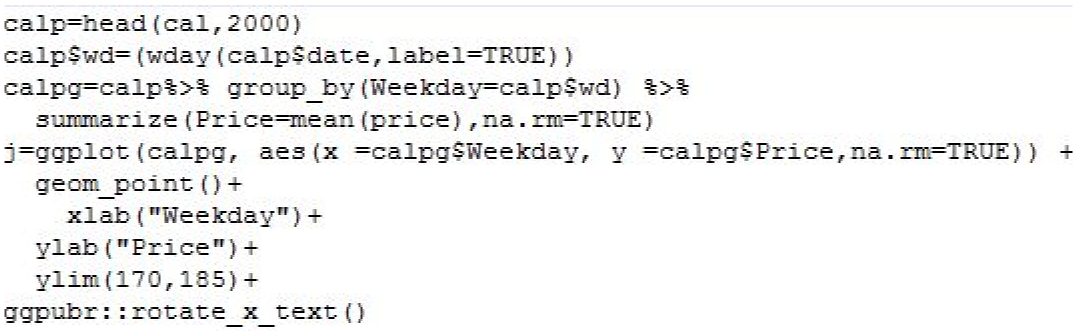
```
kcal$price <- as.numeric(gsub('\\$|,', '', cal$price))
readr::locale("en")
Sys.setlocale("LC_TIME", "English")
# Reorder following the value of another column:
calp=head(cal,2000)
plt <- ggplot(calp, aes(months.Date(calp$date), price ),color="red") +
  xlab("Month")+
  stat_summary(fun.y = "mean", geom = "point",na.rm=TRUE)+
  ggpubr::rotate_x_text()
plt
```



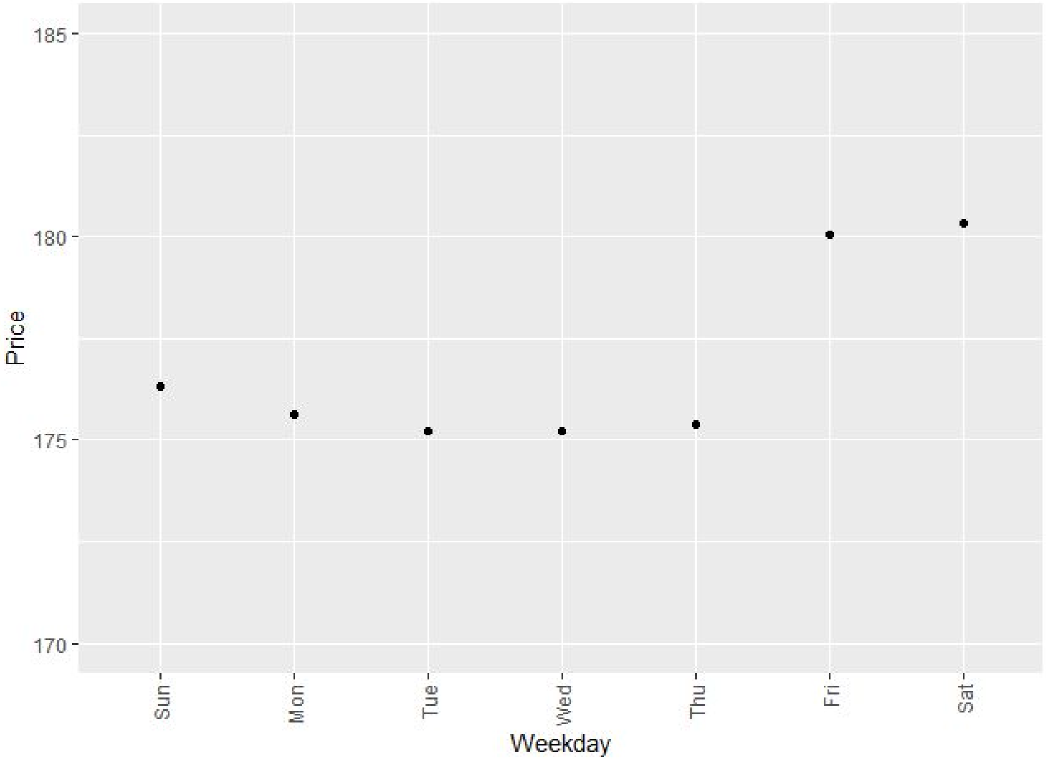
The price over the year has a variation of about 40$ maximum as interpreted from the plot.The price is higher December compared to all other months.

**#3 How price changes during day of week?**

For the analysis first 2000 rows are extracted considering fast computing,using head function.



```
calp=head(cal,2000)
calp$wd=(wday(calp$date,label=TRUE))
calpg=calp%>% group_by(Weekday=calp$wd) %>%
  summarize(Price=mean(price),na.rm=TRUE)
j=ggplot(calpg, aes(x =calpg$Weekday, y =calpg$Price,na.rm=TRUE)) +
  geom_point()+
    xlab("Weekday")+
  ylab("Price")+
  ylim(170,185)+
ggpubr::rotate_x_text()
```

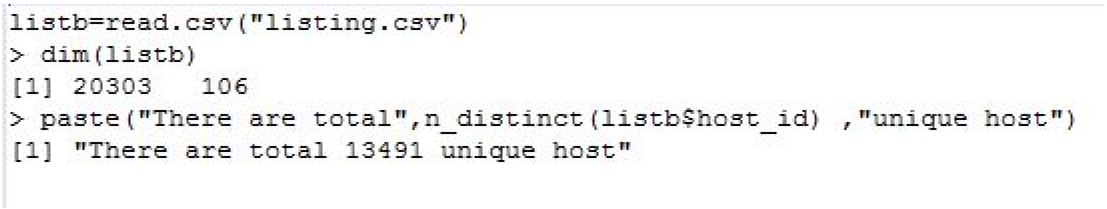


Fridays and Saturdays are over $10 more expensive than the rest of the week.

**LISTING**

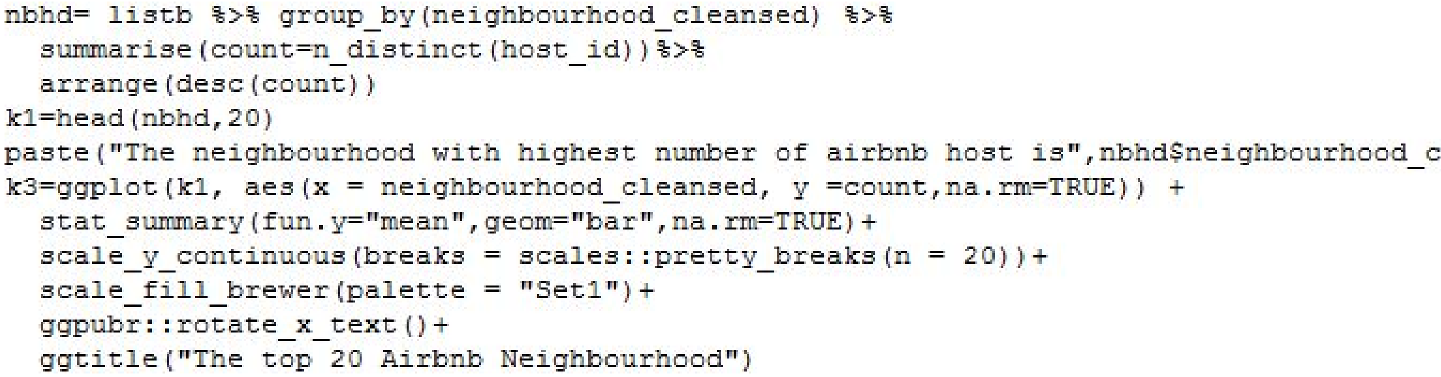
Listing is a large dataset with 20303 rows and 106 columns.

**#4 Number of listings in each neighbourhood**



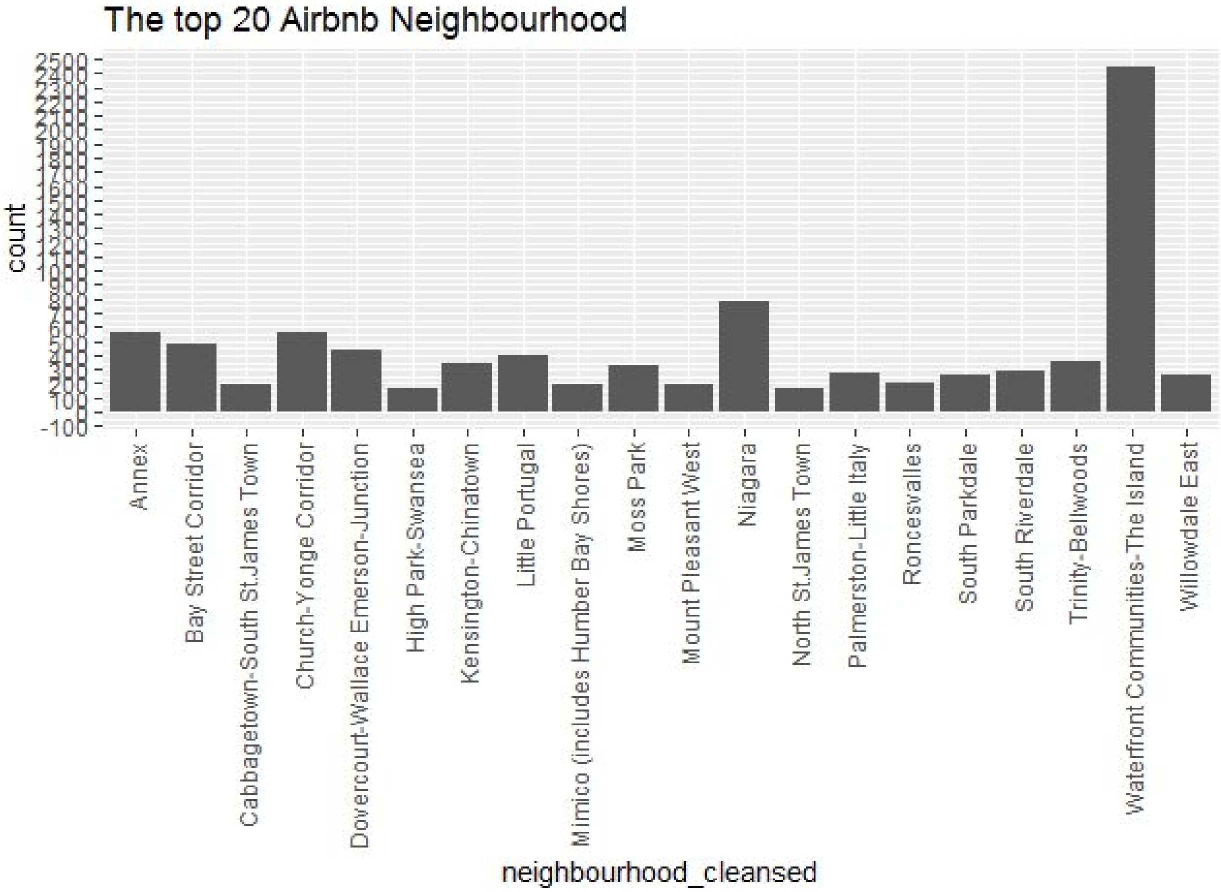
```
listb=read.csv("listing.csv")
> dim(listb)
[1] 20303   106
> paste("There are total",n_distinct(listb$host_id) ,"unique host")
[1] "There are total 13491 unique host"
```

The top 20 neighbourhoods with highest number of Airbnb host are found using the following codes.



```
nbhd= listb %>% group_by(neighbourhood_cleansed) %>%
  summarise(count=n_distinct(host_id))%>%
  arrange(desc(count))
k1=head(nbhd,20)
paste("The neighbourhood with highest number of airbnb host is",nbhd$neighbourhood_c
k3=ggplot(k1, aes(x = neighbourhood_cleansed, y =count,na.rm=TRUE)) +
  stat_summary(fun.y="mean",geom="bar",na.rm=TRUE)+
  scale_y_continuous(breaks = scales::pretty_breaks(n = 20))+
  scale_fill_brewer(palette = "Set1")+
  ggpubr::rotate_x_text()+
  ggtitle("The top 20 Airbnb Neighbourhood")
```

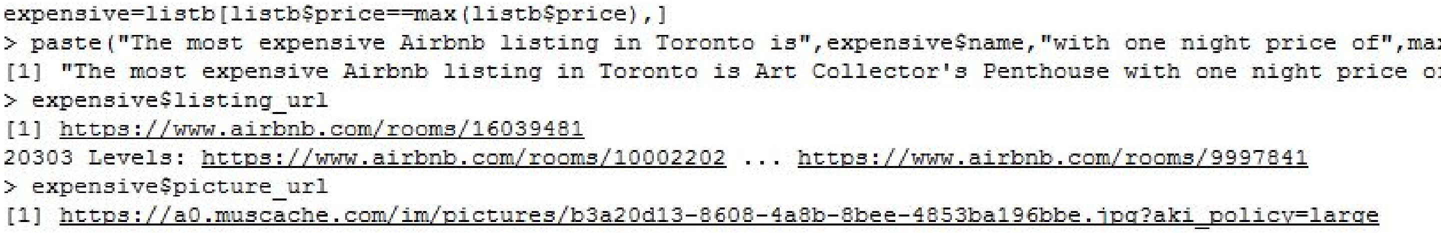
The neighbourhood that has the highest number of listings is Waterfront Communities-The Island, and almost four times more than the second most neighbourhood (Niagara).



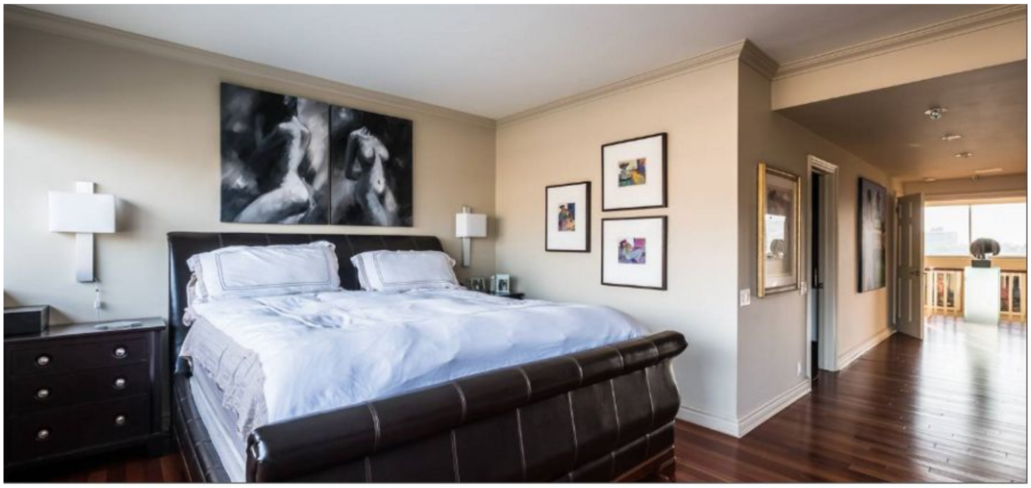
### #5 The most expensive Airbnb Host in Toronto

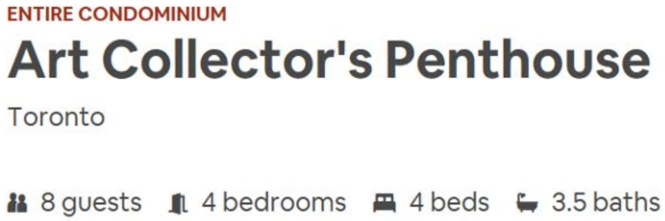
"The most expensive Airbnb listing in Toronto is Art Collector's Penthouse with price of

$**13426/night .**



```
expensive=listb[listb$price==max(listb$price),]
> paste("The most expensive Airbnb listing in Toronto is",expensive$name,"with one night price of",ma
[1] "The most expensive Airbnb listing in Toronto is Art Collector's Penthouse with one night price o
> expensive$listing_url
[1] https://www.airbnb.com/rooms/16039481
20303 Levels: https://www.airbnb.com/rooms/10002202 ... https://www.airbnb.com/rooms/9997841
> expensive$picture_url
[1] https://a0.muscache.com/im/pictures/b3a20d13-8608-4a8b-8bee-4853ba196bbe.jpg?aki_policy=large
```





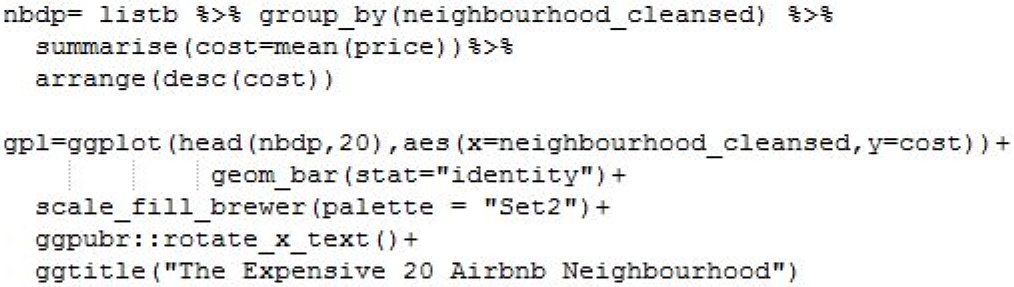
ENTIRE CONDOMINIUM

Art Collector's Penthouse

Toronto

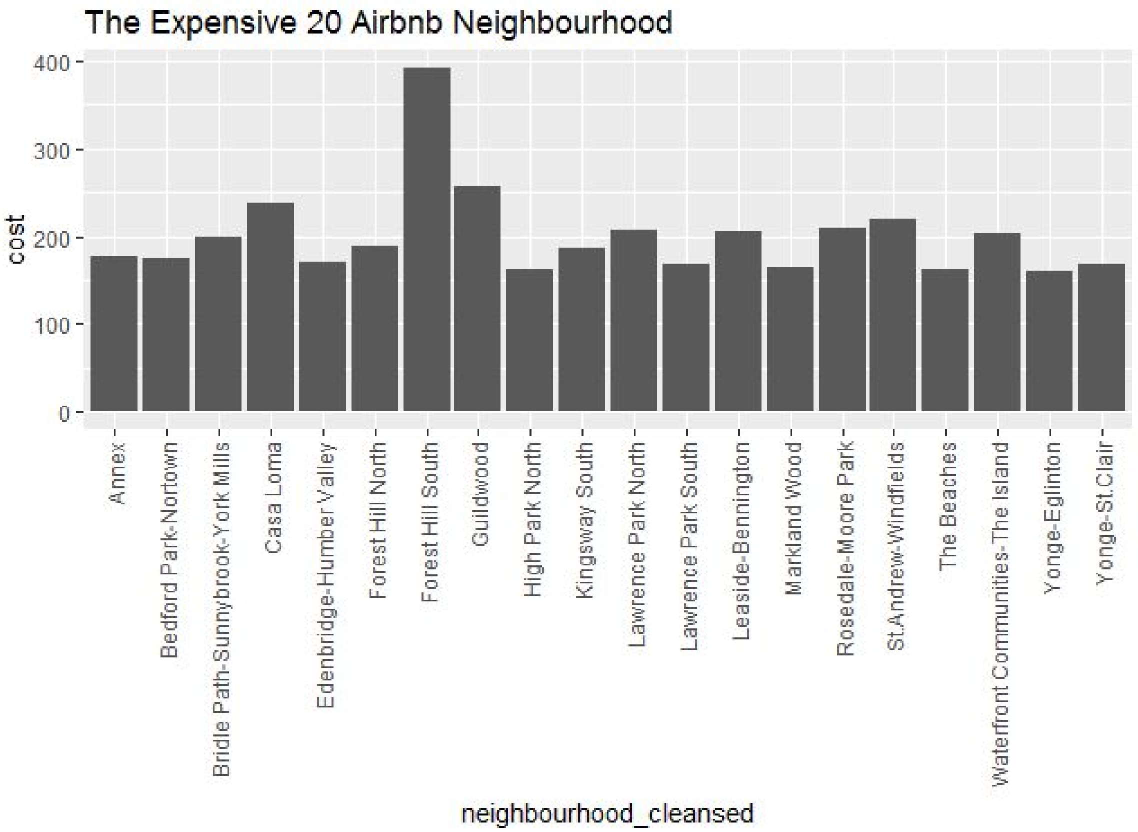
8 guests 4 bedrooms 4 beds 3.5 baths

### #6 Expensive 20 Neighbourhood



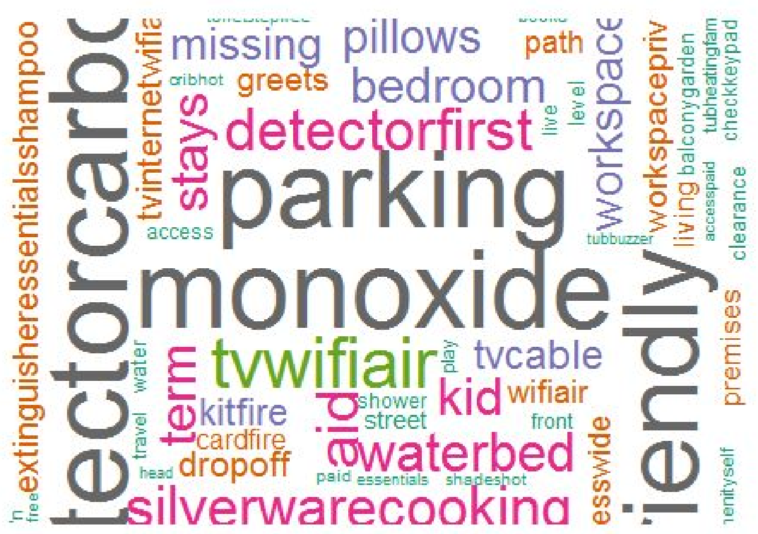
```
nbdp= listb %>% group_by(neighbourhood_cleansed) %>%
  summarise(cost=mean(price))%>%
  arrange(desc(cost))

gpl=ggplot(head(nbdp,20),aes(x=neighbourhood_cleansed,y=cost))+
             geom_bar(stat="identity")+
  scale_fill_brewer(palette = "Set2")+
  ggpubr::rotate_x_text()+
  ggtitle("The Expensive 20 Airbnb Neighbourhood")
```

Without removing the outliers Forest Hill South is the most expensive followed by Guildwood as interpreted from the following barplot.


The Expensive 20 Airbnb Neighbourhood
cost
400
300
200
100
0
Annex
Bedford Park-Nortown
Bridle Path-Sunnybrook-York Mills
Casa Loma
Edenbridge-Humber Valley
Forest Hill North
Forest Hill South
Guildwood
High Park North
Kingsway South
Lawrence Park North
Lawrence Park South
Leaside-Bennington
Markland Wood
Rosedale-Moore Park
St.Andrew-Windfields
The Beaches
Waterfront Communities-The Island
Yonge-Eglinton
Yonge-St.Clair
neighbourhood_cleansed

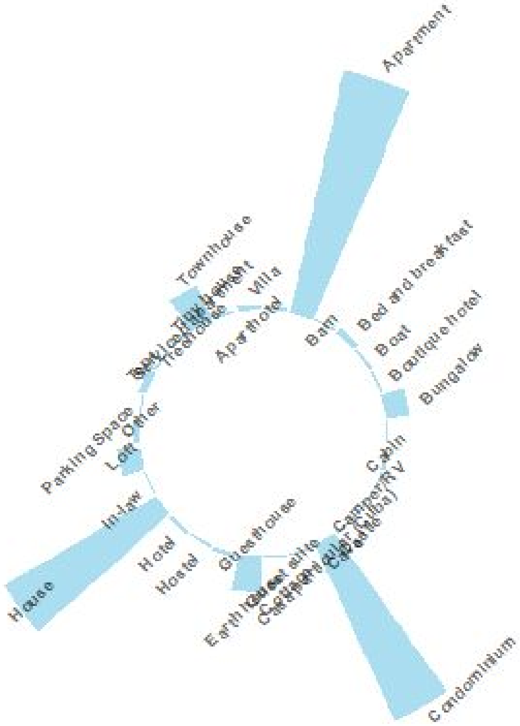
**#7 Most listed Amenities**

The most listed amenities can be found using text mining the amenities column of lising and forming the word cloud.



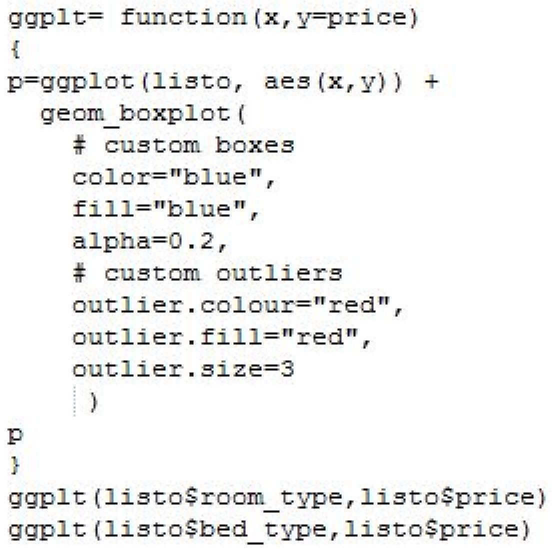
**#8 Most listed Room Type**

The most listed room type is found from circular bar plot.



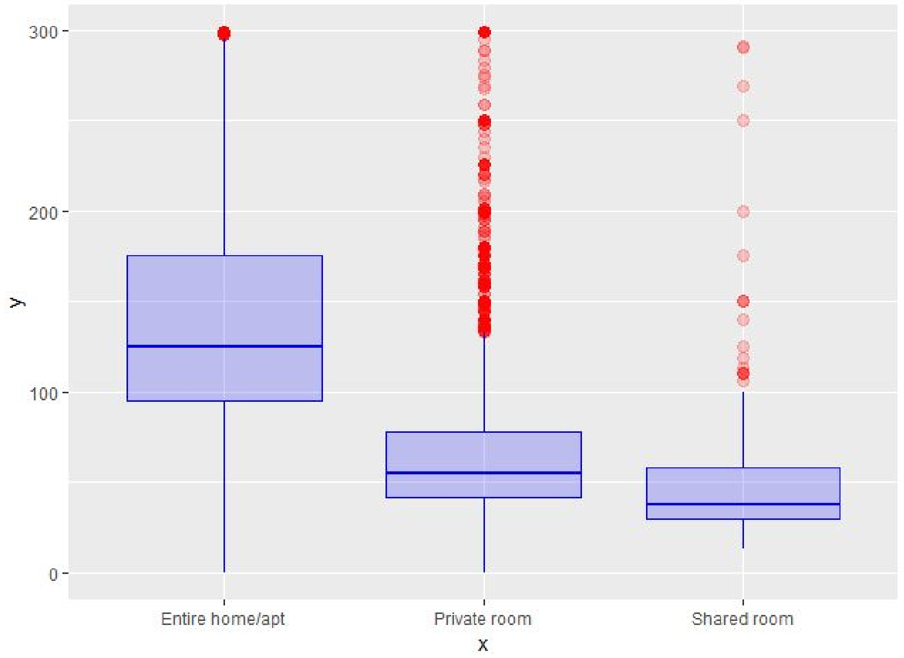
Condominium is the most listed room type followed by Apartment and Houses

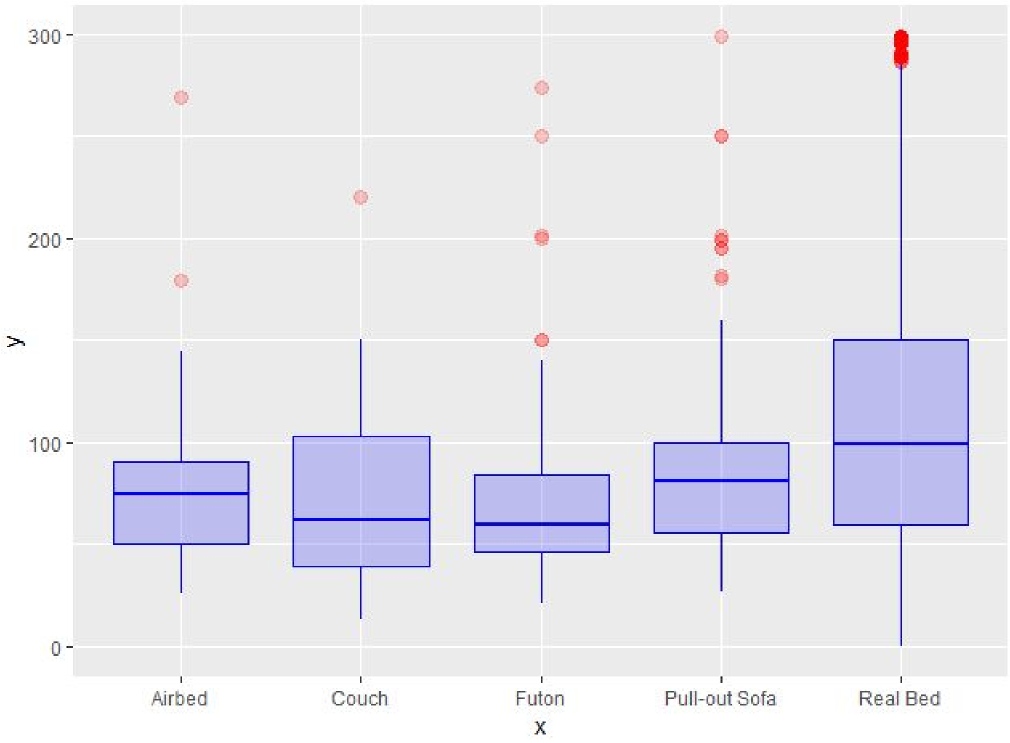
**#9 Room Type Vs Price & Bed Type Vs Price**



```
ggplt= function(x,y=price)
{
p=ggplot(listo, aes(x,y)) +
  geom_boxplot(
    # custom boxes
    color="blue",
    fill="blue",
    alpha=0.2,
    # custom outliers
    outlier.colour="red",
    outlier.fill="red",
    outlier.size=3
     )
p
}
ggplt(listo$room_type,listo$price)
ggplt(listo$bed_type,listo$price)
```

Using ggplot- box plot can be found to get the most prominent room type and bed type to the pricing.



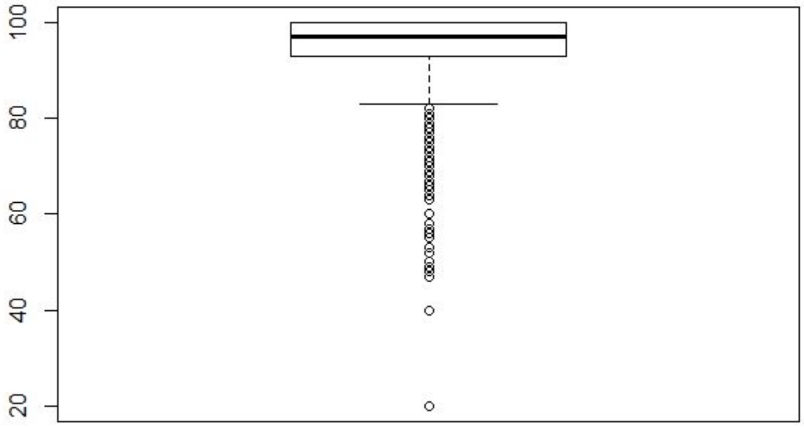


**REVIEWS**

Reviews is a large dataset which has data regarding reviews of customers, scores, date and host id etc

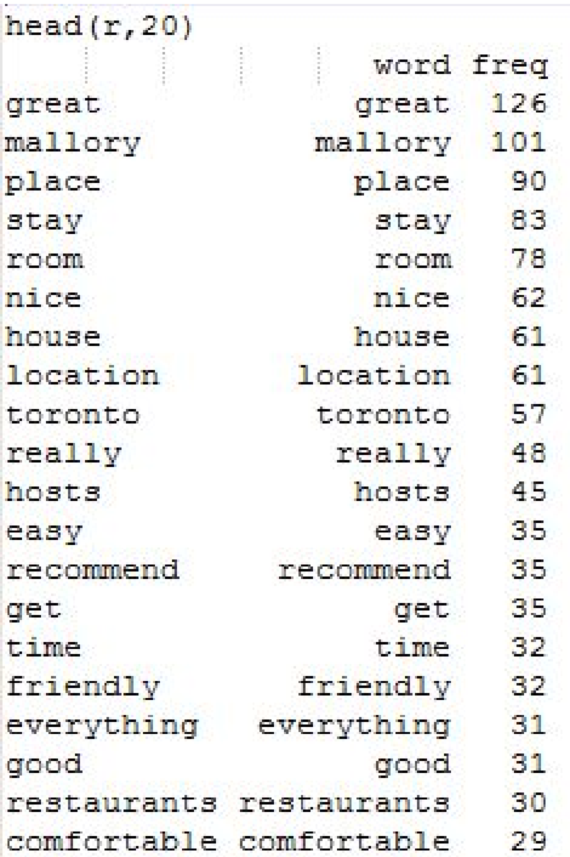
```
> dim(rvw)
[1] 508754      6
```

**#10 Box plot of review scores:**



As expected, most of reviewers leave high scores.

**#11 The most popular words used in best reviews with scores equal to 100 in reviews are:**



```
head(r,20)
                   word freq
great             great  126
mallory         mallory  101
place             place   90
stay               stay   83
room               room   78
nice               nice   62
house             house   61
location       location   61
toronto         toronto   57
really           really   48
hosts             hosts   45
easy               easy   35
recommend     recommend   35
get                 get   35
time               time   32
friendly       friendly   32
everything   everything   31
good               good   31
restaurants restaurants   30
comfortable comfortable   29
```