New Orleans Airbnb dataset

Abhijeet Ramachandran Tatvawit

16/01/2022

# About the New Orleans Airbnb Listings and Reviews dataset:

This data set describes the listing activity of Airbnb homestays in New Orleans, Louisiana, as part of the Inside Airbnb initiative. The data set was compiled on November 7, 2021. Data credit goes to Murray Cox and Inside Airbnb. The original source for this particular New Orleans data can be found here–where you can also find information on the different listing ids and their price and availability for different calendar dates.

link for dataset : [https://www.kaggle.com/ruthgn/new-orleans-airbnb-listings-and-reviews](link)

## What we can do with this dataset :

* Can you predict how much a short-term rental in New Orleans should charge per night based on it’s location and amenities?
* Can you describe the vibe of each neighborhood in using listing descriptions?
* What are the most common amenities to have among short-term rental listings in New Orleans?
* What elements contribute to a popular or highly-rated listing?
* Is there any noticeable difference in favorability among different NOLA neighborhood/areas and what could be the reason for it?

## Imoporting Essential libraries

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.6 v dplyr 1.0.7  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.1.1 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(lubridate)

##   
## Attaching package: 'lubridate'

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

library(dplyr)  
library(readr)  
library(ggplot2)

## Importing the dataset

listing <- read.csv("new\_orleans\_airbnb\_listings.csv")

## Checking for NA values

sum(is.na(listing))

## [1] 5926

## Cleaning the dataset

### Droping the NA values

listing <- drop\_na(listing)

### Renaming the column

colnames(listing)[27] <- "price\_in\_dollars"

### Correcting the format for dataset

listing$host\_since <- as.Date(listing$host\_since, format = "%Y-%m-%d")  
listing$first\_review <- as.Date(listing$first\_review, format = "%Y-%m-%d")  
listing$last\_review <- as.Date(listing$last\_review, format = "%Y-%m-%d")  
listing$price\_in\_dollars <- gsub("[$,]", "", listing$price\_in\_dollars)  
listing$price\_in\_dollars <- as.double(listing$price\_in\_dollars)

### Separating the host\_location data

listing <- separate(listing, "host\_location", into = c("host\_state", "host\_city", "host\_country"), sep = ",")

### Removing NA values after separating the data

listing <- drop\_na(listing)

## Analyzing the data through the visualization

Now that we are ready for an exploration of our data, we can make a rule that we are going to be working from left to right. The reason some may prefer to do this is due to its set approach - some datasets have a big number of attributes, plus this way we will remember to explore each column individually to make sure we learn as much as we can about our dataset.

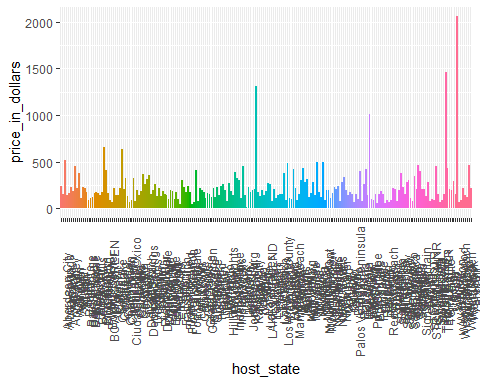
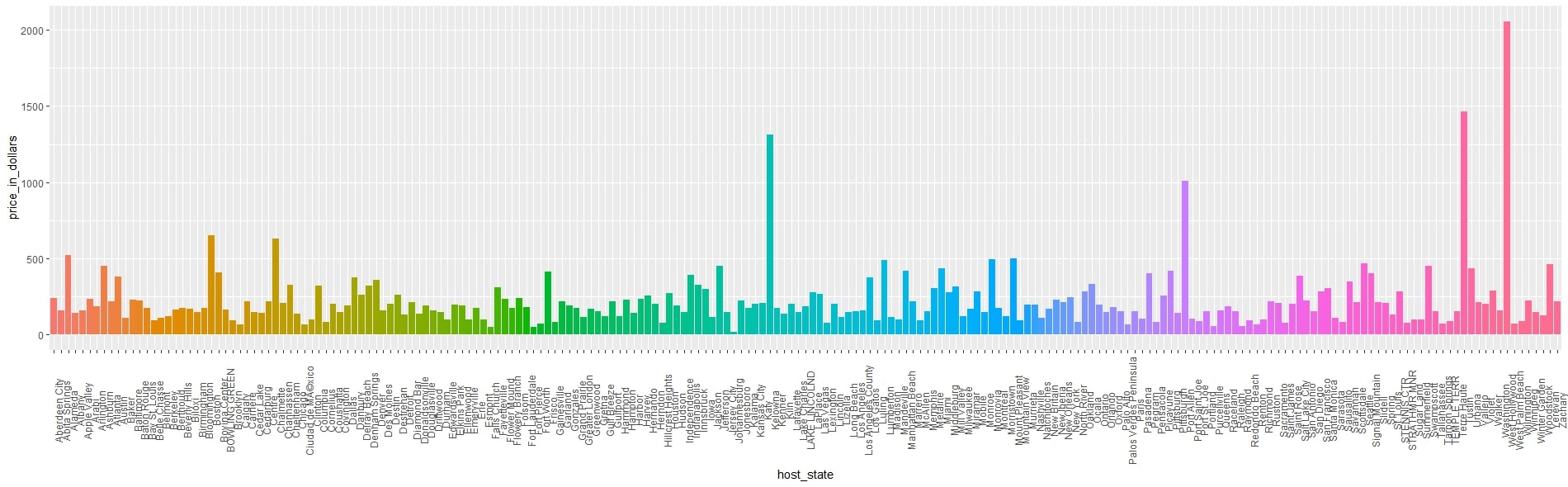
### Checking the summary of the data

summary(listing)

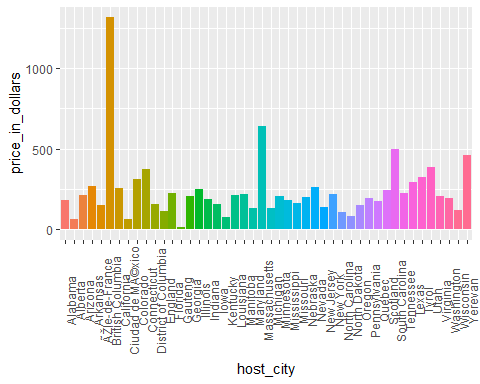
## id name description neighborhood\_overview  
## Min. : 19091 Length:4373 Length:4373 Length:4373   
## 1st Qu.:16571247 Class :character Class :character Class :character   
## Median :28201447 Mode :character Mode :character Mode :character   
## Mean :28320605   
## 3rd Qu.:41591170   
## Max. :53123547   
## host\_id host\_since host\_state host\_city   
## Min. : 5146 Min. :2008-12-13 Length:4373 Length:4373   
## 1st Qu.: 16876788 1st Qu.:2014-06-29 Class :character Class :character   
## Median : 66925019 Median :2016-04-14 Mode :character Mode :character   
## Mean :103483728 Mean :2016-03-21   
## 3rd Qu.:146815436 3rd Qu.:2017-08-22   
## Max. :425793064 Max. :2021-10-03   
## host\_country host\_response\_time host\_response\_rate host\_acceptance\_rate  
## Length:4373 Length:4373 Length:4373 Length:4373   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## host\_is\_superhost host\_listings\_count host\_verifications host\_has\_profile\_pic  
## Length:4373 Min. : 0.00 Length:4373 Length:4373   
## Class :character 1st Qu.: 1.00 Class :character Class :character   
## Mode :character Median : 2.00 Mode :character Mode :character   
## Mean : 30.52   
## 3rd Qu.: 9.00   
## Max. :3186.00   
## host\_neighbourhood host\_identity\_verified neighbourhood\_cleansed  
## Length:4373 Length:4373 Length:4373   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## latitude longitude property\_type room\_type   
## Min. :29.90 Min. :-90.14 Length:4373 Length:4373   
## 1st Qu.:29.94 1st Qu.:-90.09 Class :character Class :character   
## Median :29.96 Median :-90.07 Mode :character Mode :character   
## Mean :29.96 Mean :-90.07   
## 3rd Qu.:29.97 3rd Qu.:-90.06   
## Max. :30.16 Max. :-89.74   
## accommodates bathrooms\_text bedrooms beds   
## Min. : 1.000 Length:4373 Min. : 1.000 Min. : 1.000   
## 1st Qu.: 2.000 Class :character 1st Qu.: 1.000 1st Qu.: 1.000   
## Median : 4.000 Mode :character Median : 2.000 Median : 2.000   
## Mean : 4.852 Mean : 2.049 Mean : 2.778   
## 3rd Qu.: 6.000 3rd Qu.: 3.000 3rd Qu.: 3.000   
## Max. :16.000 Max. :18.000 Max. :27.000   
## amenities price\_in\_dollars minimum\_nights maximum\_nights   
## Length:4373 Min. : 11.0 Min. : 1.000 Min. : 1.0   
## Class :character 1st Qu.: 103.0 1st Qu.: 1.000 1st Qu.: 60.0   
## Mode :character Median : 151.0 Median : 2.000 Median :1124.0   
## Mean : 217.6 Mean : 7.959 Mean : 646.6   
## 3rd Qu.: 254.0 3rd Qu.: 3.000 3rd Qu.:1125.0   
## Max. :3450.0 Max. :365.000 Max. :5000.0   
## has\_availability availability\_30 availability\_60 availability\_90  
## Length:4373 Min. : 0.00 Min. : 0.00 Min. : 0.00   
## Class :character 1st Qu.: 2.00 1st Qu.: 9.00 1st Qu.:23.00   
## Mode :character Median :11.00 Median :30.00 Median :53.00   
## Mean :11.67 Mean :27.88 Mean :47.06   
## 3rd Qu.:19.00 3rd Qu.:44.00 3rd Qu.:72.00   
## Max. :30.00 Max. :60.00 Max. :90.00   
## availability\_365 number\_of\_reviews number\_of\_reviews\_ltm  
## Min. : 0.0 Min. : 1.0 Min. : 0.00   
## 1st Qu.: 48.0 1st Qu.: 14.0 1st Qu.: 1.00   
## Median :132.0 Median : 40.0 Median : 10.00   
## Mean :154.4 Mean : 68.6 Mean : 15.33   
## 3rd Qu.:270.0 3rd Qu.: 95.0 3rd Qu.: 23.00   
## Max. :365.0 Max. :766.0 Max. :359.00   
## number\_of\_reviews\_l30d first\_review last\_review   
## Min. : 0.000 Min. :2011-06-21 Min. :2014-04-08   
## 1st Qu.: 0.000 1st Qu.:2018-03-10 1st Qu.:2021-05-31   
## Median : 1.000 Median :2019-07-05 Median :2021-10-11   
## Mean : 1.365 Mean :2019-04-13 Mean :2021-04-27   
## 3rd Qu.: 2.000 3rd Qu.:2020-12-26 3rd Qu.:2021-10-28   
## Max. :18.000 Max. :2021-11-05 Max. :2021-11-06   
## review\_scores\_rating review\_scores\_accuracy review\_scores\_cleanliness  
## Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:4.730 1st Qu.:4.810 1st Qu.:4.740   
## Median :4.870 Median :4.920 Median :4.880   
## Mean :4.795 Mean :4.842 Mean :4.804   
## 3rd Qu.:4.960 3rd Qu.:4.980 3rd Qu.:4.970   
## Max. :5.000 Max. :5.000 Max. :5.000   
## review\_scores\_checkin review\_scores\_communication review\_scores\_location  
## Min. :1.00 Min. :1.000 Min. :1.000   
## 1st Qu.:4.89 1st Qu.:4.860 1st Qu.:4.700   
## Median :4.96 Median :4.950 Median :4.870   
## Mean :4.90 Mean :4.876 Mean :4.777   
## 3rd Qu.:5.00 3rd Qu.:5.000 3rd Qu.:4.960   
## Max. :5.00 Max. :5.000 Max. :5.000   
## review\_scores\_value license instant\_bookable reviews\_per\_month  
## Min. :1.000 Length:4373 Length:4373 Min. : 0.010   
## 1st Qu.:4.700 Class :character Class :character 1st Qu.: 0.840   
## Median :4.830 Mode :character Mode :character Median : 1.830   
## Mean :4.766 Mean : 3.267   
## 3rd Qu.:4.910 3rd Qu.: 3.410   
## Max. :5.000 Max. :292.000

### Visualizing the prices as per the host state

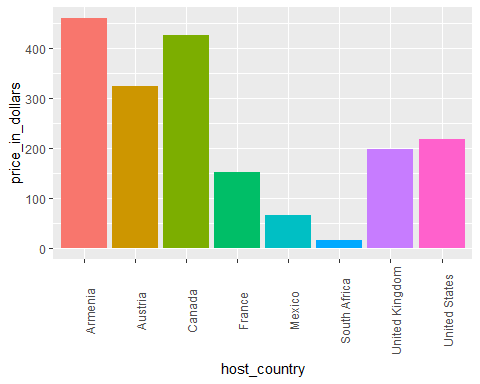
ggplot(listing, aes(x = host\_state, y = price\_in\_dollars, fill = host\_state )) + stat\_summary(fun = "mean", geom = "col") + theme(axis.text.x = element\_text(angle = 90), legend.position="none")

### Visualizing the prices as per the host city

 ### Visualization the prices per the host country

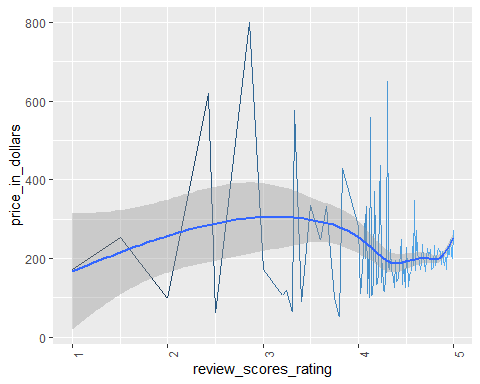
ggplot(listing, aes(x = host\_country, y = price\_in\_dollars, fill = host\_country )) + stat\_summary(fun = "mean", geom = "col") + theme(axis.text.x = element\_text(angle = 90), legend.position="none")



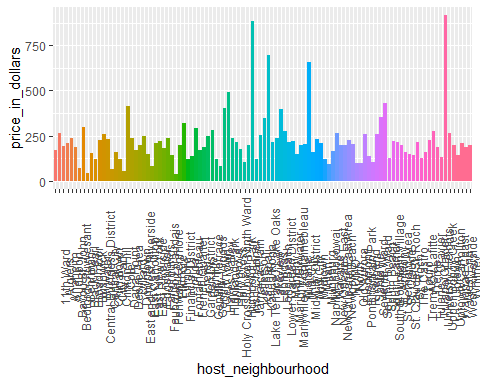
### Visualization the prices as per review\_ratings

ggplot(listing, aes(x = review\_scores\_rating, y = price\_in\_dollars, color = review\_scores\_rating )) + stat\_summary(fun = "mean", geom = "line")+ geom\_smooth() + theme(axis.text.x = element\_text(angle = 90), legend.position="none")

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

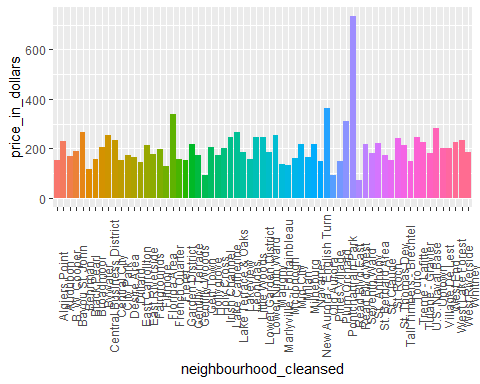
 ### Visualization the prices as per neighbourhood

ggplot(listing, aes(x = host\_neighbourhood, y = price\_in\_dollars, fill = host\_neighbourhood )) + stat\_summary(fun = "mean", geom = "col") + theme(axis.text.x = element\_text(angle = 90), legend.position="none")



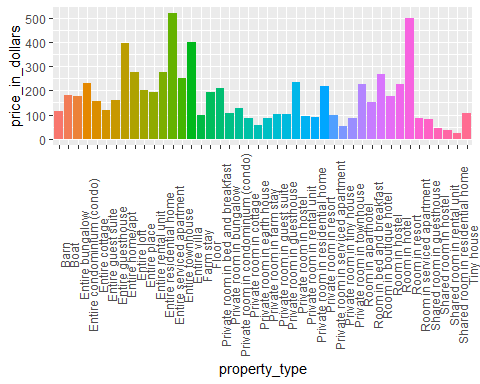
### Visualization the prices as per cleansed neighborhood

ggplot(listing, aes(x = neighbourhood\_cleansed, y = price\_in\_dollars, fill = neighbourhood\_cleansed )) + stat\_summary(fun = "mean", geom = "col") + theme(axis.text.x = element\_text(angle = 90), legend.position="none")



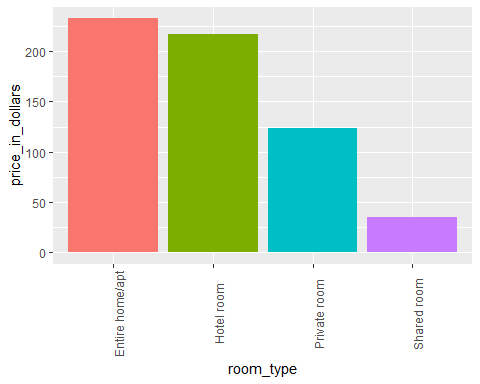
### Visualization the prices as per the property types

ggplot(listing, aes(x = property\_type, y = price\_in\_dollars, fill = property\_type )) + stat\_summary(fun = "mean", geom = "col") + theme(axis.text.x = element\_text(angle = 90), legend.position="none")



### # Visualization the prices as per the room types

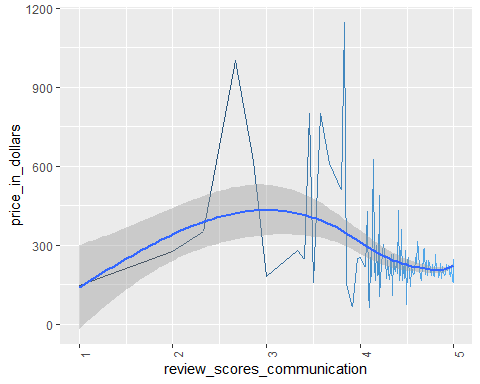
ggplot(listing, aes(x = room\_type, y = price\_in\_dollars, fill = room\_type )) + stat\_summary(fun = "mean", geom = "col") + theme(axis.text.x = element\_text(angle = 90), legend.position="none")



### Visualization the price as per review score communication

ggplot(listing, aes(x = review\_scores\_communication, y = price\_in\_dollars, color = review\_scores\_communication )) + stat\_summary(fun = "mean", geom = "line")+ geom\_smooth() + theme(axis.text.x = element\_text(angle = 90), legend.position="none")

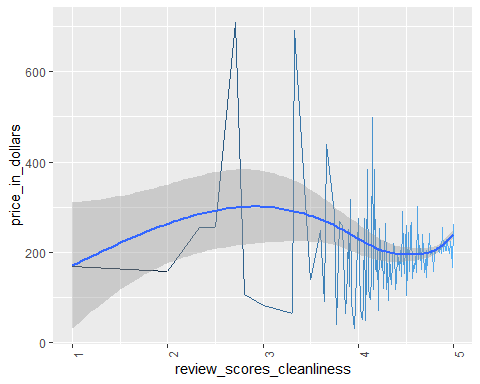
## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



### Visualization the price as per review score cleanliness

ggplot(listing, aes(x = review\_scores\_cleanliness, y = price\_in\_dollars, color = review\_scores\_cleanliness )) + stat\_summary(fun = "mean", geom = "line")+ geom\_smooth() + theme(axis.text.x = element\_text(angle = 90), legend.position="none")

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



### Conclusion

This Airbnb (‘New Orleans Airbnb Listings and Reviews’) dataset for the 2019 year appeared to be a very rich dataset with a variety of columns that allowed us to do deep data exploration on each significant column presented.