Recommender\_System

SHIVAM VATS

SUSHIL DEORE

ANSH SACHDEVA

2022-12-07

#### Question of Interest

What features reflect some pattern that would be usefulto target relevant audience, so that the customer engagement can be increased.

#### Data:

##### Source: Data is taken from Kaggle: <https://www.kaggle.com/competitions/airbnb-recruiting-new-user-bookings/data>

Dataset is a collection of users along with their demographics, web session records, and some summary statistics. All the users in this dataset are from the USA. Here is a short description of the features of data:

1. id: user id
2. date\_account\_created: the date of account creation
3. timestamp\_first\_active: timestamp of the first activity, note that it can be earlier than date\_account\_created or date\_first\_booking because a user can search before signing up
4. date\_first\_booking: date of first booking
5. gender
6. age
7. signup\_method - Platform used for signup
8. signup\_flow: the page a user came to signup up from
9. language: international language preference
10. affiliate\_channel: what kind of paid marketing
11. affiliate\_provider: where the marketing is e.g. google, craigslist, other
12. first\_affiliate\_tracked: whats the first marketing the user interacted with before the signing up
13. signup\_app
14. first\_device\_type
15. first\_browser
16. country\_destination

### Analysis

Importing data and checking its result

data <-read.csv("Airbnb New User Bookings.csv")  
print(head(data, n = 15))

## id date\_account\_created timestamp\_first\_active date\_first\_booking  
## 1 gxn3p5htnn 2010-06-28 2.009032e+13   
## 2 820tgsjxq7 2011-05-25 2.009052e+13   
## 3 4ft3gnwmtx 2010-09-28 2.009061e+13 2010-08-02  
## 4 bjjt8pjhuk 2011-12-05 2.009103e+13 2012-09-08  
## 5 87mebub9p4 2010-09-14 2.009121e+13 2010-02-18  
## 6 osr2jwljor 2010-01-01 2.010010e+13 2010-01-02  
## 7 lsw9q7uk0j 2010-01-02 2.010010e+13 2010-01-05  
## 8 0d01nltbrs 2010-01-03 2.010010e+13 2010-01-13  
## 9 a1vcnhxeij 2010-01-04 2.010010e+13 2010-07-29  
## 10 6uh8zyj2gn 2010-01-04 2.010010e+13 2010-01-04  
## 11 yuuqmid2rp 2010-01-04 2.010010e+13 2010-01-06  
## 12 om1ss59ys8 2010-01-05 2.010011e+13   
## 13 k6np330cm1 2010-01-05 2.010011e+13 2010-01-18  
## 14 dy3rgx56cu 2010-01-05 2.010011e+13   
## 15 ju3h98ch3w 2010-01-07 2.010011e+13   
## gender age signup\_method signup\_flow language affiliate\_channel  
## 1 -unknown- NA facebook 0 en direct  
## 2 MALE 38 facebook 0 en seo  
## 3 FEMALE 56 basic 3 en direct  
## 4 FEMALE 42 facebook 0 en direct  
## 5 -unknown- 41 basic 0 en direct  
## 6 -unknown- NA basic 0 en other  
## 7 FEMALE 46 basic 0 en other  
## 8 FEMALE 47 basic 0 en direct  
## 9 FEMALE 50 basic 0 en other  
## 10 -unknown- 46 basic 0 en other  
## 11 FEMALE 36 basic 0 en other  
## 12 FEMALE 47 basic 0 en other  
## 13 -unknown- NA basic 0 en direct  
## 14 FEMALE 37 basic 0 en other  
## 15 FEMALE 36 basic 0 en other  
## affiliate\_provider first\_affiliate\_tracked signup\_app first\_device\_type  
## 1 direct untracked Web Mac Desktop  
## 2 google untracked Web Mac Desktop  
## 3 direct untracked Web Windows Desktop  
## 4 direct untracked Web Mac Desktop  
## 5 direct untracked Web Mac Desktop  
## 6 other omg Web Mac Desktop  
## 7 craigslist untracked Web Mac Desktop  
## 8 direct omg Web Mac Desktop  
## 9 craigslist untracked Web Mac Desktop  
## 10 craigslist omg Web Mac Desktop  
## 11 craigslist untracked Web Mac Desktop  
## 12 craigslist untracked Web iPhone  
## 13 direct Web Other/Unknown  
## 14 craigslist linked Web Mac Desktop  
## 15 craigslist untracked Web iPhone  
## first\_browser country\_destination  
## 1 Chrome NDF  
## 2 Chrome NDF  
## 3 IE US  
## 4 Firefox other  
## 5 Chrome US  
## 6 Chrome US  
## 7 Safari US  
## 8 Safari US  
## 9 Safari US  
## 10 Firefox US  
## 11 Firefox US  
## 12 -unknown- NDF  
## 13 -unknown- FR  
## 14 Firefox NDF  
## 15 Mobile Safari NDF

# Checking dimensions of data  
dim(data)

## [1] 213451 16

Dropping columns of related to timestamps since it is not relevant and useful in this scenario.

# dropping id and timestamp related features  
data <- data[,-(1:4), drop= FALSE]

Checking for null values

# Checking Column wise null values  
colSums(is.na(data))

## gender age signup\_method   
## 0 87990 0   
## signup\_flow language affiliate\_channel   
## 0 0 0   
## affiliate\_provider first\_affiliate\_tracked signup\_app   
## 0 0 0   
## first\_device\_type first\_browser country\_destination   
## 0 0 0

# Checking structure of data  
str(data)

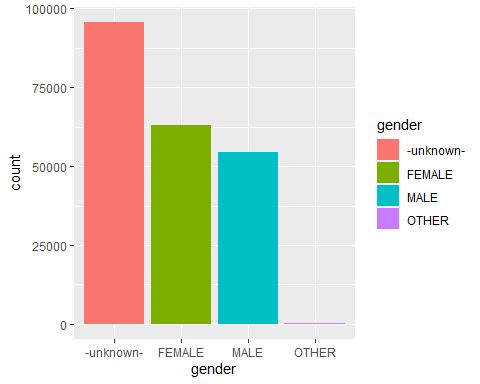
## 'data.frame': 213451 obs. of 12 variables:  
## $ gender : chr "-unknown-" "MALE" "FEMALE" "FEMALE" ...  
## $ age : num NA 38 56 42 41 NA 46 47 50 46 ...  
## $ signup\_method : chr "facebook" "facebook" "basic" "facebook" ...  
## $ signup\_flow : int 0 0 3 0 0 0 0 0 0 0 ...  
## $ language : chr "en" "en" "en" "en" ...  
## $ affiliate\_channel : chr "direct" "seo" "direct" "direct" ...  
## $ affiliate\_provider : chr "direct" "google" "direct" "direct" ...  
## $ first\_affiliate\_tracked: chr "untracked" "untracked" "untracked" "untracked" ...  
## $ signup\_app : chr "Web" "Web" "Web" "Web" ...  
## $ first\_device\_type : chr "Mac Desktop" "Mac Desktop" "Windows Desktop" "Mac Desktop" ...  
## $ first\_browser : chr "Chrome" "Chrome" "IE" "Firefox" ...  
## $ country\_destination : chr "NDF" "NDF" "US" "other" ...

# Checking summary of data  
summary(data)

## gender age signup\_method signup\_flow   
## Length:213451 Min. : 1.00 Length:213451 Min. : 0.000   
## Class :character 1st Qu.: 28.00 Class :character 1st Qu.: 0.000   
## Mode :character Median : 34.00 Mode :character Median : 0.000   
## Mean : 49.67 Mean : 3.267   
## 3rd Qu.: 43.00 3rd Qu.: 0.000   
## Max. :2014.00 Max. :25.000   
## NA's :87990   
## language affiliate\_channel affiliate\_provider  
## Length:213451 Length:213451 Length:213451   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## first\_affiliate\_tracked signup\_app first\_device\_type   
## Length:213451 Length:213451 Length:213451   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## first\_browser country\_destination  
## Length:213451 Length:213451   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##   
##

Distribution of Gender

# Distribution of Gender  
ggplot(data, aes(x=gender)) +  
 geom\_bar(aes(fill=gender))

 % missing values in age

# Checking % missing values in column age  
sum(is.na(data$age))\*100/dim(data)[1]

## [1] 41.22258

Summary Statistics of Age

# Checking summary statistics of age  
summary(data$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.00 28.00 34.00 49.67 43.00 2014.00 87990

From the above statistics, we can see that the maximum value of age is 2014. Hence, we shall remove outliers/errors while analyzing the same

# Removing outliers in age  
age\_filtered = data %>% filter(age <= IQR(age,na.rm = TRUE)\*1.5 + 50)

Distribution of age with respect to gender

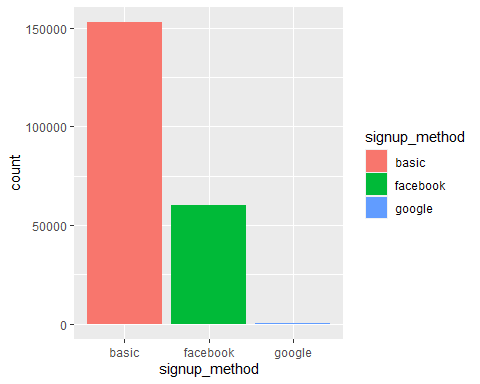
plt1 = ggplot(age\_filtered,aes(x=age,na.rm=TRUE))+  
 stat\_density(aes(fill=gender))+  
 ggtitle("Distribution of Age with respect to Gender")+  
 theme(plot.title = element\_text(hjust=0.5))

# Possible values in Signup method  
unique(data$signup\_method)

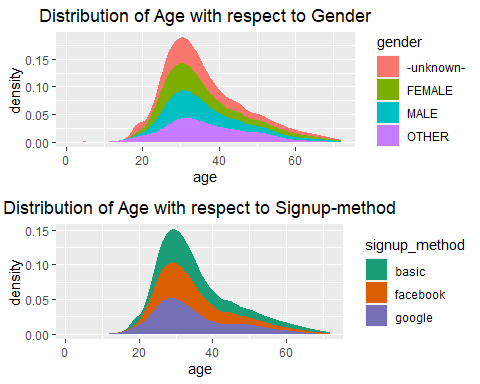
## [1] "facebook" "basic" "google"

Distribution of Signup method

ggplot(data, aes(x=signup\_method)) +  
 geom\_bar(aes(fill=signup\_method))

 Distribution of age with respect to signup\_method

plt2 = ggplot(age\_filtered,aes(x=age,na.rm=TRUE))+  
 stat\_density(aes(fill=signup\_method))+  
 scale\_fill\_brewer(palette="Dark2")+  
 ggtitle("Distribution of Age with respect to Signup-method")+  
 theme(plot.title = element\_text(hjust=0.5))  
grid.arrange(plt1,plt2,nrow=2,ncol=1)



# Column  
unique(data$language)

## [1] "en" "fr" "de" "es" "it" "pt" "zh" "ko" "ja" "ru" "pl" "el" "sv" "nl" "hu"  
## [16] "da" "id" "fi" "no" "tr" "th" "cs" "hr" "ca" "is"

# Distribution of Languages  
ggplot(data, aes(x=language)) +  
 geom\_bar()



#### Bias and its cause

From the above distribution, we can observe that the dataset’s language feature is biased towards a single language. The possible source of this bias shall be because of the users’ database. Since all the users in the database are from the United States(as mentioned in source of the data), language is highly skewed towards English.

#From the above bar plot, we can infer that the most of the data is for a single  
#language, that is English  
  
# Checking column affiliate\_channel  
unique(data$affiliate\_channel)

## [1] "direct" "seo" "other" "sem-non-brand"  
## [5] "content" "sem-brand" "remarketing" "api"

# Checking column affiliate\_provider  
unique(data$affiliate\_provider)

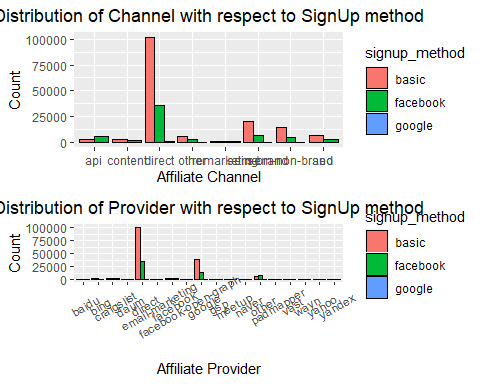
## [1] "direct" "google" "other"   
## [4] "craigslist" "facebook" "vast"   
## [7] "bing" "meetup" "facebook-open-graph"  
## [10] "email-marketing" "yahoo" "padmapper"   
## [13] "gsp" "wayn" "naver"   
## [16] "baidu" "yandex" "daum"

Distribution affiliate\_channel

#Distribution affiliate\_channel  
plt1= ggplot(data, aes(x=affiliate\_channel)) +  
 geom\_bar(aes(fill=signup\_method),color='black',position="dodge")+  
 ggtitle("Distribution of Channel with respect to SignUp method")+  
 ylab("Count")+  
 xlab("Affiliate Channel")+  
 theme(plot.title = element\_text(hjust=0.5))

Distribution affiliate\_provider

#Distribution affiliate\_provider  
plt2 = ggplot(data, aes(x=affiliate\_provider)) +  
 geom\_bar(aes(fill=signup\_method),color='black',position="dodge")+  
 ggtitle("Distribution of Provider with respect to SignUp method")+  
 ylab("Count")+  
 xlab("Affiliate Provider")+  
 theme(plot.title = element\_text(hjust=0.5),axis.text.x = element\_text(angle=30,vjust=0.8))  
  
grid.arrange(plt1,plt2,nrow=2,ncol=1)



# Column - first\_affiliate\_tracked  
# Checking column first\_affiliate\_tracked  
  
unique(data$first\_affiliate\_tracked)

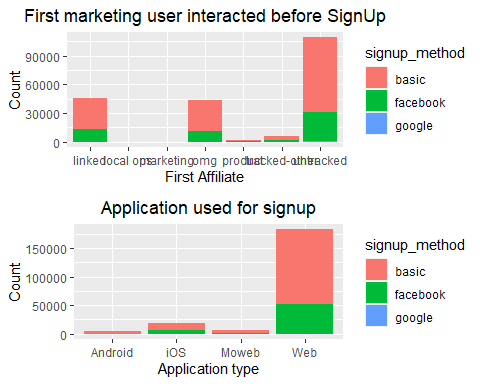
## [1] "untracked" "omg" "" "linked"   
## [5] "tracked-other" "product" "marketing" "local ops"

# Plotting barplot for different levels of column first\_affiliate\_tracked in data  
plt1 = data %>%   
 filter(first\_affiliate\_tracked != "") %>%   
 ggplot(aes(x=first\_affiliate\_tracked)) +  
 geom\_bar(aes(fill=signup\_method))+  
 ggtitle("First marketing user interacted before SignUp")+  
 ylab("Count")+  
 xlab("First Affiliate")+  
 theme(plot.title = element\_text(hjust=0.5))

# Checking column signup\_app  
unique(data$signup\_app)

## [1] "Web" "Moweb" "iOS" "Android"

# Plotting barplot for different levels of sign\_up in data  
plt2 = ggplot(data=data,aes(x=signup\_app)) +  
 geom\_bar(aes(fill=signup\_method))+  
 ggtitle("Application used for signup")+  
 ylab("Count")+  
 xlab("Application type")+  
 theme(plot.title = element\_text(hjust=0.5))  
grid.arrange(plt1,plt2,nrow=2,ncol=1)

 ### CHI Squared Analysis

Here we have performed an analysis to determine the nature dependency of categorical variables on each other. Since most of the features are categorical Chi squared analysis is a good fit. Chi Squared analysis is used to determine if two or more categorical features are related to each other. In this analysis, we perform a hypothesis test. If the p-value of the hypothesis test is less than 0.05, then we reject our Null hypothesis, otherwise, we cannot reject it.

#### Null Hypothesis:

There is no relation between method used to signup and the type of app #### Alternate Hypothesis: There is some relation between method used to signup and the type of app #### Significance Level: 0.05

data$signup\_method = factor(data$signup\_method)  
data$signup\_app = factor(data$signup\_app)  
  
contigency\_table = table(data$signup\_method,data$signup\_app)  
chisq.test(contigency\_table)

##   
## Pearson's Chi-squared test  
##   
## data: contigency\_table  
## X-squared = 20952, df = 6, p-value < 2.2e-16

Here, p-value is 2.2e-16, which is lower than 0.05 (significance level). Hence we can reject the null hypothesis and infer that there is some relation between the features signup\_app and signup\_method.

Distribution of feature first\_browser

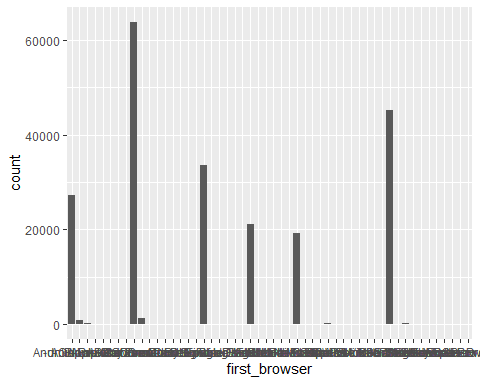
# Checking column first\_browser  
str(data$first\_browser)

## chr [1:213451] "Chrome" "Chrome" "IE" "Firefox" "Chrome" "Chrome" "Safari" ...

# Checking summary of first\_browser column  
summary(data$first\_browser)

## Length Class Mode   
## 213451 character character

# Plotting barplot for different levels of column first\_browser in data  
ggplot(data, aes(x=first\_browser)) +  
 geom\_bar()

 Distribution of destination country

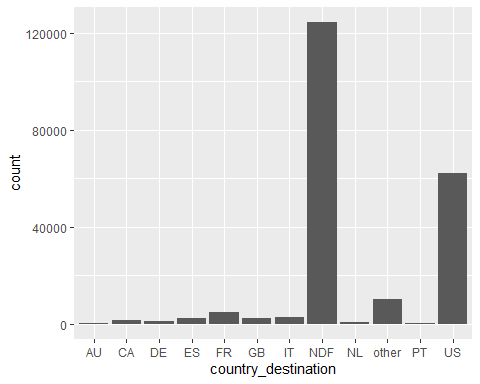
# Checking column country\_destination  
str(data$country\_destination)

## chr [1:213451] "NDF" "NDF" "US" "other" "US" "US" "US" "US" "US" "US" "US" ...

# Checking summary of country\_destination column  
summary(data$country\_destination)

## Length Class Mode   
## 213451 character character

# Plotting barplot for different levels of column country\_destination in data  
ggplot(data, aes(x=country\_destination)) +  
 geom\_bar()

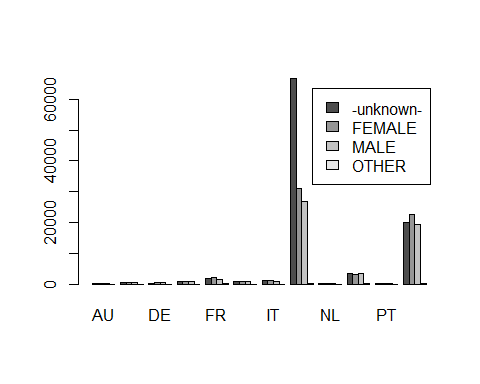


Bi-variate analysis of features

# Defining function for two categorical column visualization  
bivariate <- function(arg\_1, arg\_2) {  
 count <- table(arg\_1, arg\_2)  
 barplot(count,beside = TRUE,legend.text = TRUE)  
}

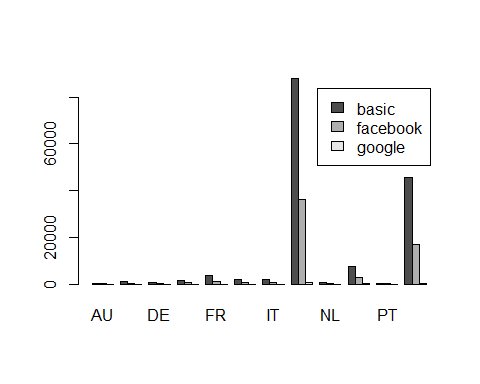
Gender vs. Country Destination

# Plotting two categorical columns  
bivariate(data$gender, data$country\_destination)



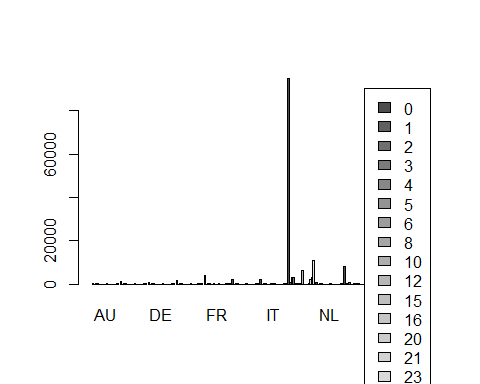
Signup method vs. Country Destination

# Plotting two categorical columns  
bivariate(data$signup\_method, data$country\_destination)



Signup\_flow vs Country Destination

# Plotting two categorical columns  
bivariate(data$signup\_flow, data$country\_destination)



### CONCLUSION

Because we now know who to target and what kind of hotel is best for the users, our analysis helped us learn more about the demographic makeup of the users and their appropriate choices when they make a reservation on Airbnb. Our system will be able to distribute more personalized material within the community to be more predictive, reduce the average time towards first booking, and better forecast demand by accurately predicting where a new user will book their first hotel experience.

##### Github Link - <https://github.com/shivam360d/DTSC-5301-Project-Report>

##### By - Shivam Vats, Sushil Deore and Ansh Sachdeva