Capstone: Airbnb New Users

Youzhu Shi

10/13/2018

Backgroud and Objective

Airbnb is an online market place for home rentals. It was founded in 2008, since then, it has had exponential growth. Airbnb has become a serious competitor and a threat to the hotel industry. You can find Airbnb listings in most cities and countries in different styles and price points. When it first started, people use it for personal travel to save money or when no hotel rooms are not available. Nowaday, people still use it for that purpose, but more importantly, a lot of people see it as the first choice, because it has what hotels can't offer: a sense of home.

The goal of this analysis is to provide insignts in to Airbnb's new user base. The data was originally posted by Airbnb for Kaggle Competition in 2015. Even if it is three years old, it still provides us some insights into Airbnb user profiles and destination preferences. Let take a quick look at the data. What important fields are available to us to analyze?

```
names(train_user)
```

```
##  [1] "id"                      "date_account_created"
##  [3] "timestamp_first_active"  "date_first_booking"
##  [5] "gender"                  "age"
##  [7] "signup_method"           "signup_flow"
##  [9] "language"                "affiliate_channel"
## [11] "affiliate_provider"      "first_affiliate_tracked"
## [13] "signup_app"              "first_device_type"
## [15] "first_browser"           "country_destination"
## [17] "age_cat"                 "gender_age"
## [19] "age_gender"              "language_full"
## [21] "language_combo"          "time_diff"
```

Age and gender can be important contributing factors to the country someone decides to travel to. One of the other two categories is the time related: timestamp_first_active and date_first_booking. The two attributes may not be very useful if we analyze each one by itself. It can be if we find the difference between the two and create a new attribute time_diff, which captures the number of day it takes for someone to make their first reservation after they first signed up as an user. Thet next group of attribute is related to users choice of techonology: signup_app, signup_method, affiliate_provider, first_browser, etc. These can be useful for marketing team for evaluating the effectiveness of marketing dollors. The last attribute I want to mention is language. Most users preferred language is English; however, for those whose primary language is not English, we can decide whether it has any correlation with the primary language of the country destination.

We must know that there are limitation to our dataset. * We do not know how many guests are traveling with the person making the booking. We do not know the gender and age of those guests. * There are a lot of missing data for important attributes such as age and gender, which lead to varying sample sizes and predictibility. * We can only determine the correlation not the causation the country destinaton. If the number of booking is low in a certain area, it could be due to the lack of supply in the country or the low quality of listings in the country.

Cleaning and Wrangling

In order to prepare the dataset for analysis, I have cleaned and wrangled the dataset which are described in detail below.

In the age attribute, there are lots of numbers that fall out of a reasonable range. Many people have entered their year of birth instead of age. For those intances, I have calculated their ages in the year of 2015. Additionally, to make it eaiser to analyze, I placed different ages in groups.

```
train_user$age_cat <- as.numeric(train_user$age)
train_user$age_cat [is.na(age_cat)] <- -1
train_user$age_cat  <- ifelse(age_cat >1000, 2015 - age_cat, age_cat)
train_user$age_cat  <- ifelse(age_cat >65 & age_cat < 1000,"over 65", age_cat)
train_user$age_cat  <- ifelse(age_cat >55 & age_cat <= 65 ,"from 56 to 65", age_cat)
train_user$age_cat  <- ifelse(age_cat >45 & age_cat <= 55 ,"from 46 to 55", age_cat)
train_user$age_cat  <- ifelse(age_cat >35 & age_cat <= 45 ,"from 36 to 45", age_cat)
train_user$age_cat  <- ifelse(age_cat >25 & age_cat <= 35 ,"from 26 to 35", age_cat)
train_user$age_cat  <- ifelse(age_cat >18 & age_cat <= 25 ,"from 18 to 25", age_cat)
train_user$age_cat  <- ifelse(age_cat > 0 & age_cat <= 18,"below 18", age_cat)

train_user$age_cat <- as.character(train_user$age_cat)
train_user$age_cat <- ifelse(train_user$age_cat == 7, "below 18", train_user$age_cat)
train_user$age_cat <- ifelse(train_user$age_cat == -1, "-unknown-", train_user$age_cat)
```

Now that we have categorized ages, it becomes easier to combine with gender and create a new varaiable to analyze: age_gender.

```
train_user$gender_age = as.character(paste(train_user$gender, train_user$age_cat))
```

Next, we can take a look at users' choice of technology. We can see that there are many first_device_types for Airbnb users. We can simplify that by put them in three categories "Desktop", "Portable", and "others".

```
device_table <-  data_frame(
  first_device_type = c("Mac Desktop","Windows Desktop", "iPhone", "iPad" , "Android Pho
ne", "Android Tablet ","Desktop (Other)","SmartPhone (Other)"),
  device_type = c("Desktop","Desktop","Portable","Portable","Portable", "Portable","Desk
top","Portable"))

train_user <- left_join(train_user, device_table, by = "first_device_type", copy = "devi
ce_table")
```

```
## Warning: Column `first_device_type` joining factor and character vector,
## coercing into character vector
```

```
train_user$device_type[is.na(train_user$device_type)] <- "others"
```

Similarly, brownser types can be simplified to make it easier to interpret.

```
browser_table <-  data_frame(
  first_browser = c("Chrome","Chrome Mobile", "Safari", "Safari Mobile", "Firefox", "IE"
),
  browser_type = c("Chrome","Chrome","Safari","Safari","Firefox", "IE"))

train_user <- left_join(train_user, browser_table, by = "first_browser", copy = "browser
_table")
```

```
## Warning: Column `first_browser` joining factor and character vector,
## coercing into character vector
```

```
train_user$browser_type[is.na(train_user$browser_type)] <- "others"
```

In the language column, language codes were stored instead of full language names. I added a column to show the full name of these languages.

```
language_code <- read.csv(file="language_code.csv",header=TRUE,sep=",")
train_user <- left_join(train_user, language_code, by = "language", copy = "language_cod
e")
```

```
## Warning: Column `language` joining character vector and factor, coercing
## into character vector
```

We can potentially examine the relationship between the primary language of a user and the primary language of a country destination; thereby, a new attribute that represents the primary language the country destination is added to the train_user table.

```
destination_language <-  data_frame(
  country_destination = c("US", "FR", "IT", "GB", "ES", "CA", "DE", "NL", "AU","PT"),
  country_primary_language = c("English", "French", "Italian", "English", "Spanish", "En
glish", "German", "Dutch", "English","Portuguese"))

train_user <- left_join(train_user, destination_language, by = "country_destination", co
py = "destination_language")
```

```
## Warning: Column `country_destination` joining factor and character vector,
## coercing into character vector
```

```
train_user$country_primary_language [is.na(train_user$country_primary_language)] <- "oth
ers"
```

We will create a new variable that combine the primary language of a user and the primary language of a country destination to see if there’s a link between the two.

```
train_user$language_combo = as.factor(paste(train_user$language_full,"-", train_user$cou
ntry_primary_language))
```

Last but not least, the difference between date_account_created and date_first_booking is created as a new attribute to examine how long it will take someone to make a book after they signup.

Statistical Analysis

Let's start with age as it is one of the most important attributes of a new user.

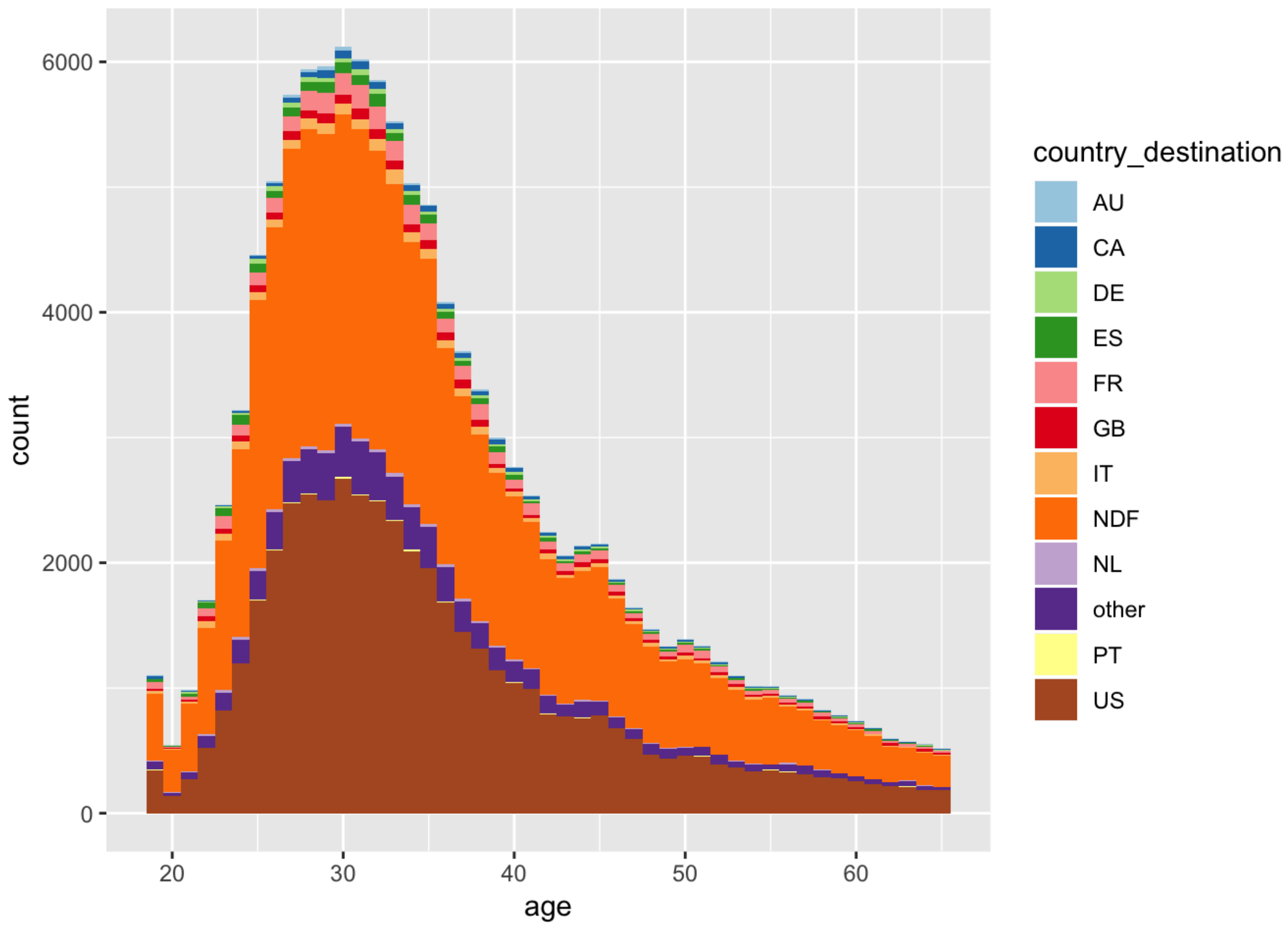
Age Summary

```
summary(train1$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   19.00   28.00   33.00   35.71   41.00   65.00 
```

Age Plot

```
train1 <- train %>% filter(age > 18 & age <= 65)
ggplot(train1, aes(x = age, fill = country_destination)) + geom_histogram(binwidth = 1,
 position = "stack") + scale_fill_brewer(palette="Paired")
```

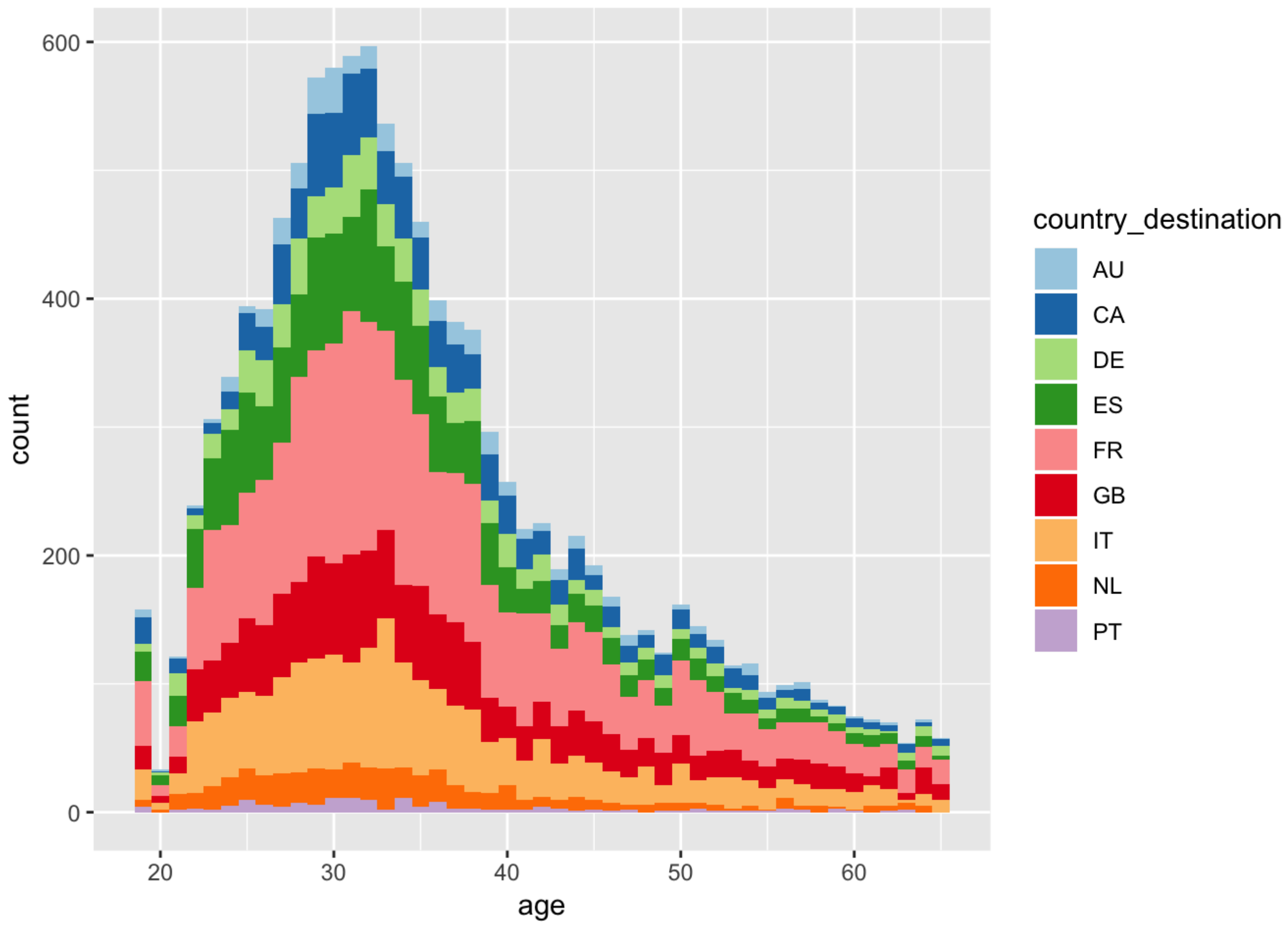


The Age of guests and the number of booking made on Airbnb appear to be a right skewed distribution. After filtering out guests who are less than 18 years old or over 99 years old, the mean age is 36.6, and the median age is 34. It has a first quartile of 28 years old and a 3rd quartile of 42 years old. It is safe to say that majority of new users' ages fall between late twenties and early forties. As people get older, they are less likely to make reservations on Airbnb. An outlier is around age 18 and 19, there's a small spike on bookings, this is probably because after graduating high school, many students decide to travel before starting college. Regardless of age, traveling within in the US is the top choice.

Outside of US

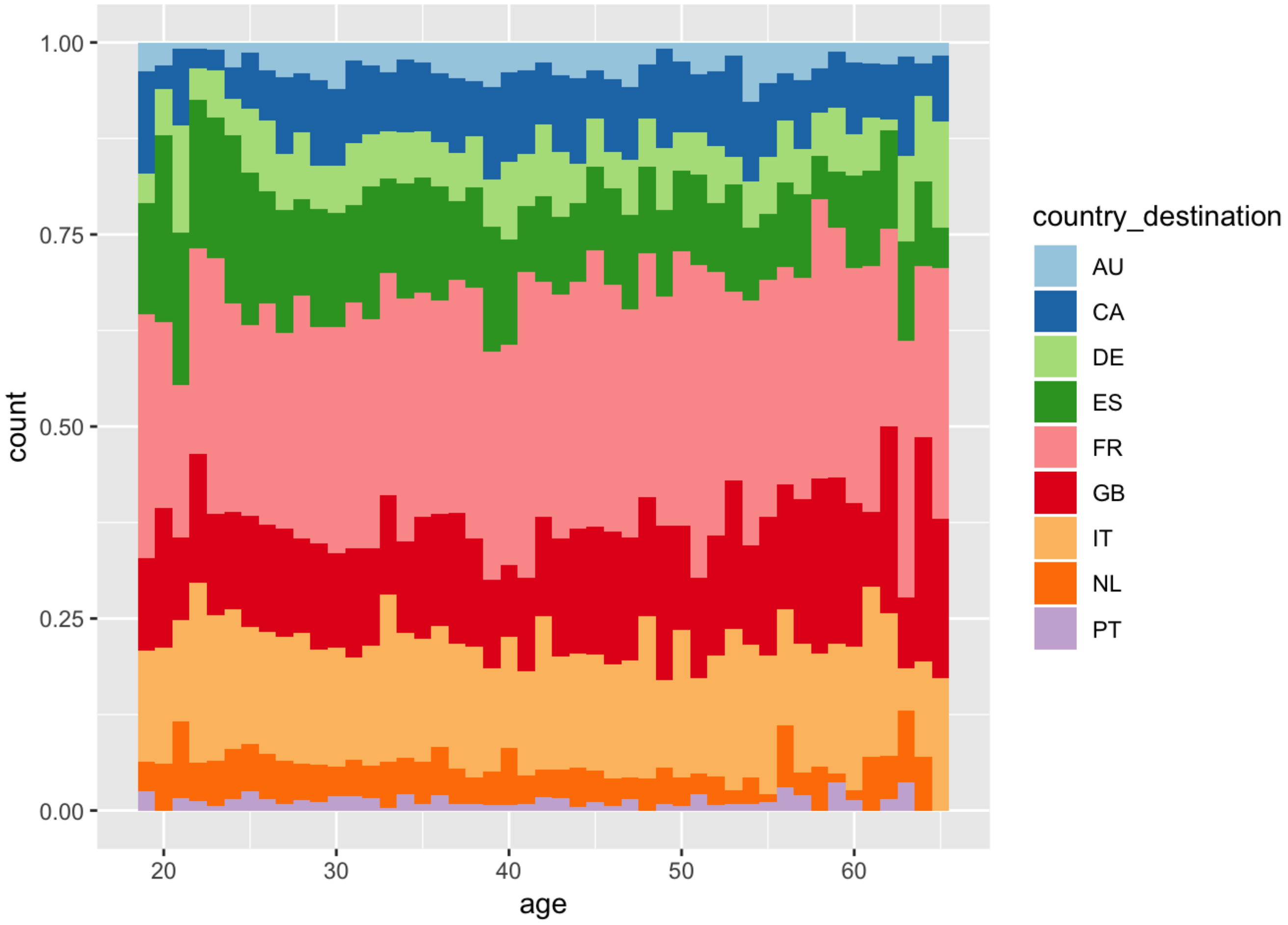
Now, let's take a look at those who chose countries outside of the US. What are their characteristics?

```
train2 <- train1 %>% filter(country_destination != "US" & country_destination!= "NDF" &
 country_destination != "other")
ggplot(train2,aes(x = age, fill = country_destination)) + geom_histogram(binwidth = 1, p
osition = "stack") + scale_fill_brewer(palette="Paired")
```



Outside of the US, France is the most popular destination followed by Italy and Great Britain.

```
ggplot(train2,aes(x = age, fill = country_destination)) + geom_histogram(binwidth = 1, p
osition = "fill") + scale_fill_brewer(palette="Paired")
```

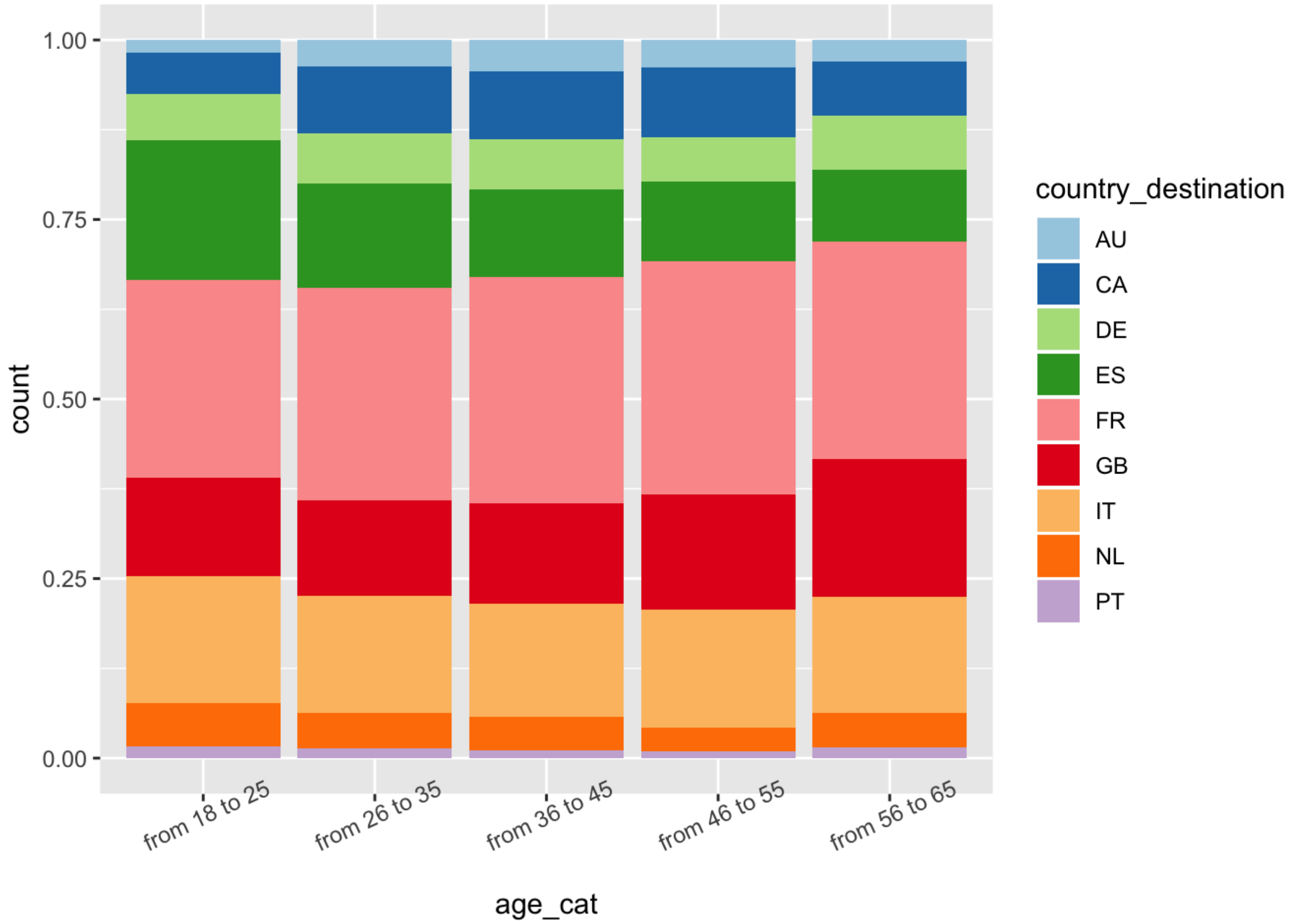


Country preferences appear fairly consistent amount different ages.

Outside of US - Age Category

What if we put age into five different brackets, will that show us a clearer picture?

```
ggplot(train2, aes(x = age_cat , fill = country_destination)) + geom_bar(position = "fil
l") + theme(axis.text.x = element_text(angle = 25)) + scale_fill_brewer(palette="Paired"
)
```

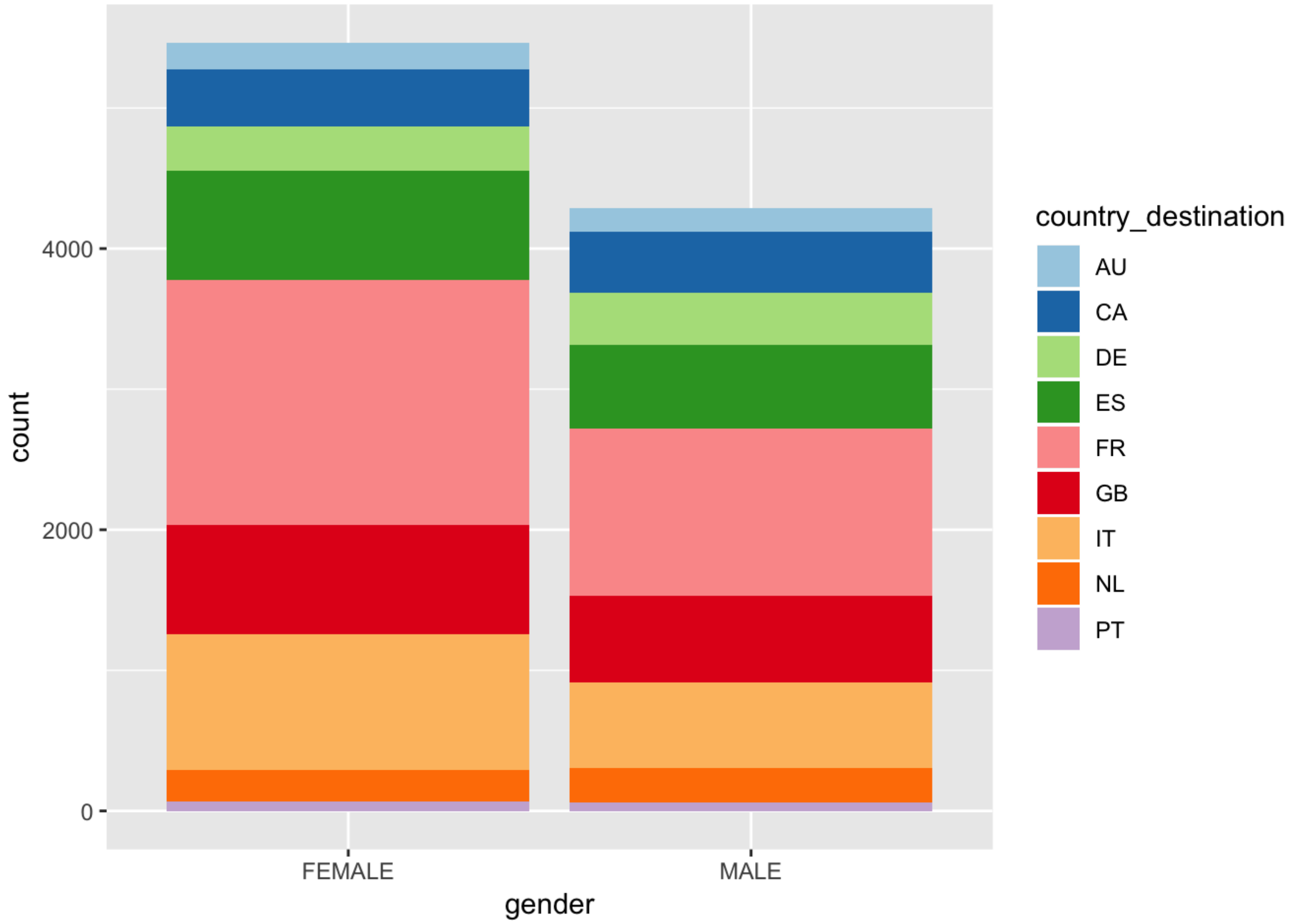


Spain is more popular among younger users; the opposite is true for Great Britain. Canada is more popular among middle aged users, this is probably because people who are in their working age do more business travels between the two countries as US and Canada have very close trade relations. They are more likely to work in the other country instead of where they are originally from.

Gender

Let's take a look at gender. Which gender is represent a higher percentage of users who made their first bookings on Airbnb?

```
train2_1 <- train2 %>% filter((gender == "FEMALE" | gender == "MALE") & (gender_age !=
"FEMALE below 18" & gender_age != "MALE below 18" & gender_age != "FEMALE over 65" & gen
der_age != "MALE over 65" ))
ggplot(train2_1, aes(x = gender, fill = country_destination)) + geom_bar() +  scale_fill
_brewer(palette="Paired")
```

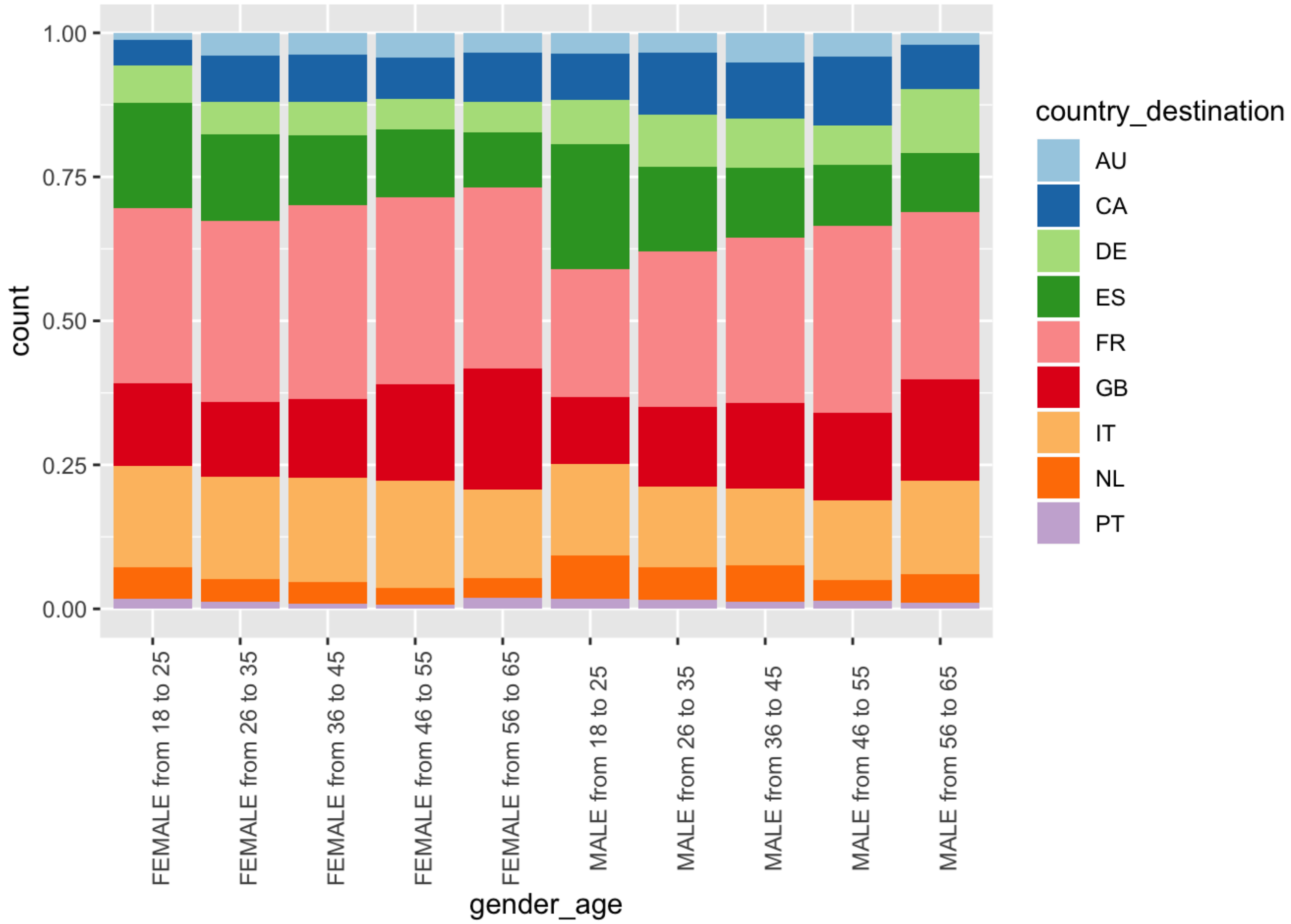


There are more female users than male users. It appears that female users have a much stronger preference for France.

Age & Gender

What if we combine age category and gender to create a new variable gender_age, will that provide us some unique insights?

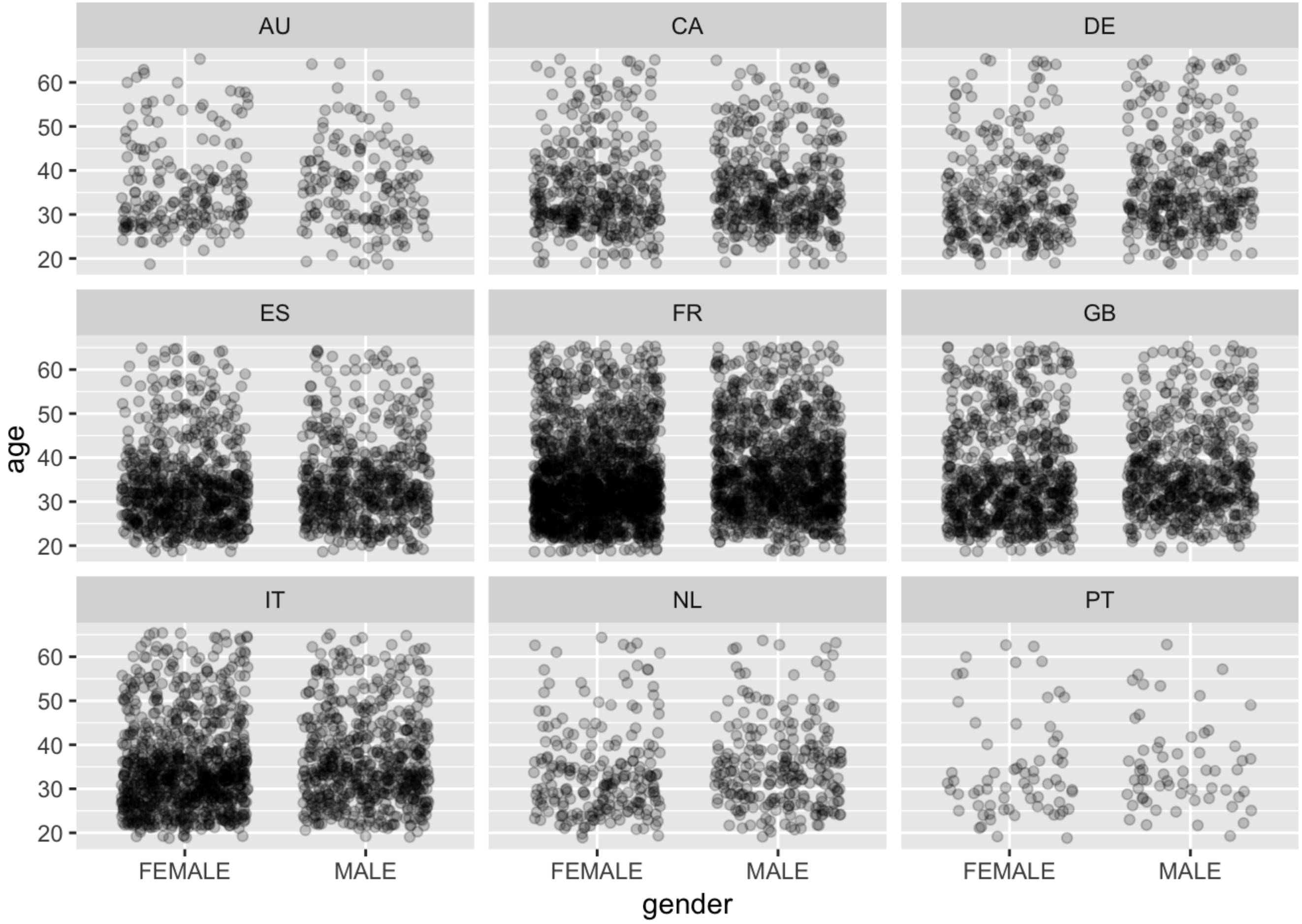
```
ggplot(train2_1, aes(x = gender_age, fill = country_destination)) + geom_bar(position =
"fill") + theme(axis.text.x = element_text(angle = 90)) + scale_fill_brewer(palette="Pai
red")
```



Both females and males are less likely to travel to Spain as they age, males show a stronger correlation. Both females and males are more likely to travel to Great Britain as they get older, females show a stronger correlation.

The Big Picture - Age & Gender

```
ggplot(train2_1 , aes(x = gender, y = age)) + geom_jitter(alpha = 0.2, width = 0.35 ) +
 facet_wrap(~country_destination)
```

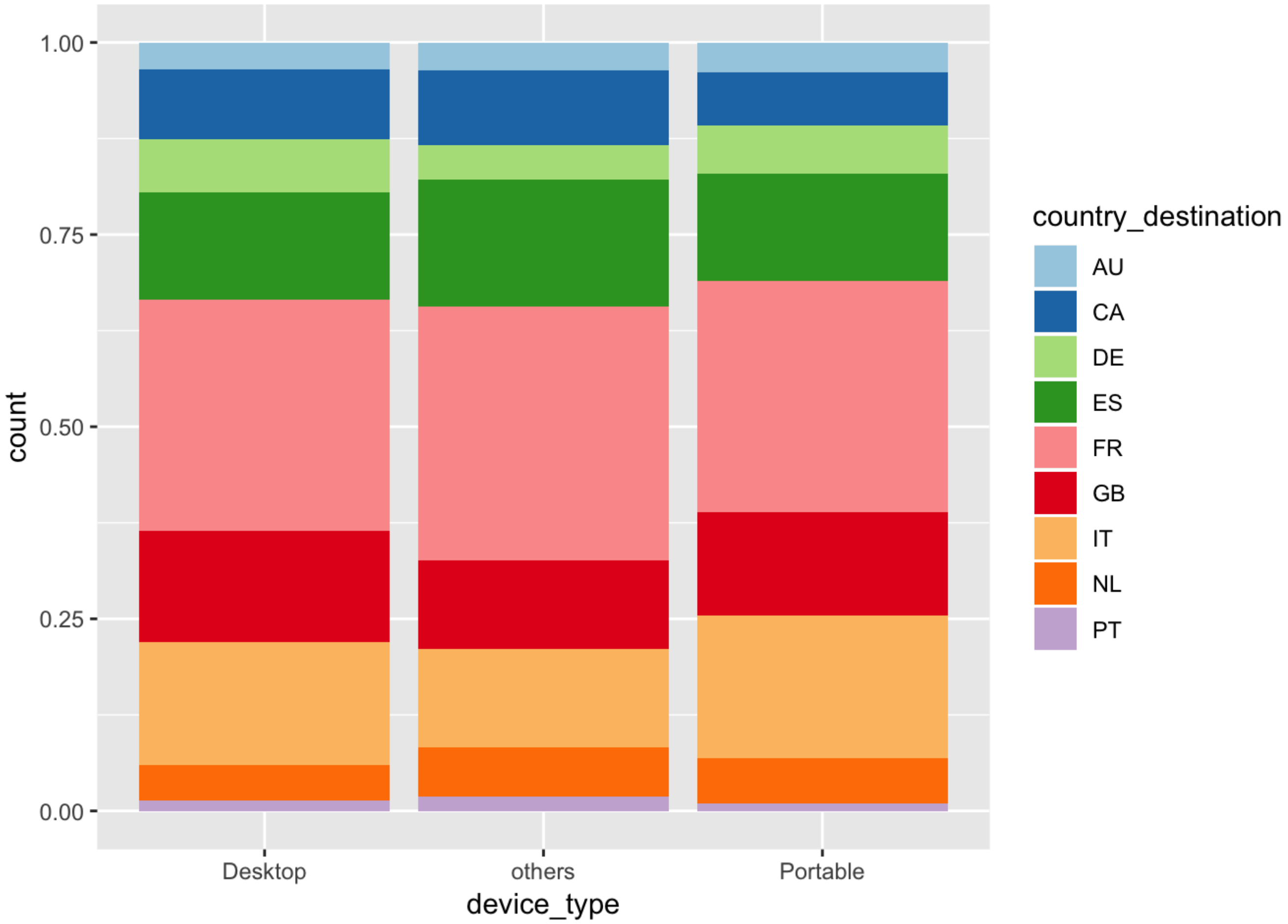


This facet grid reaffirmed us of our previous analysis. It does provide us additional information. The travel age for males is slight high than the travel age for female. This tendency appears strong in Great Britain, Germany, and France.

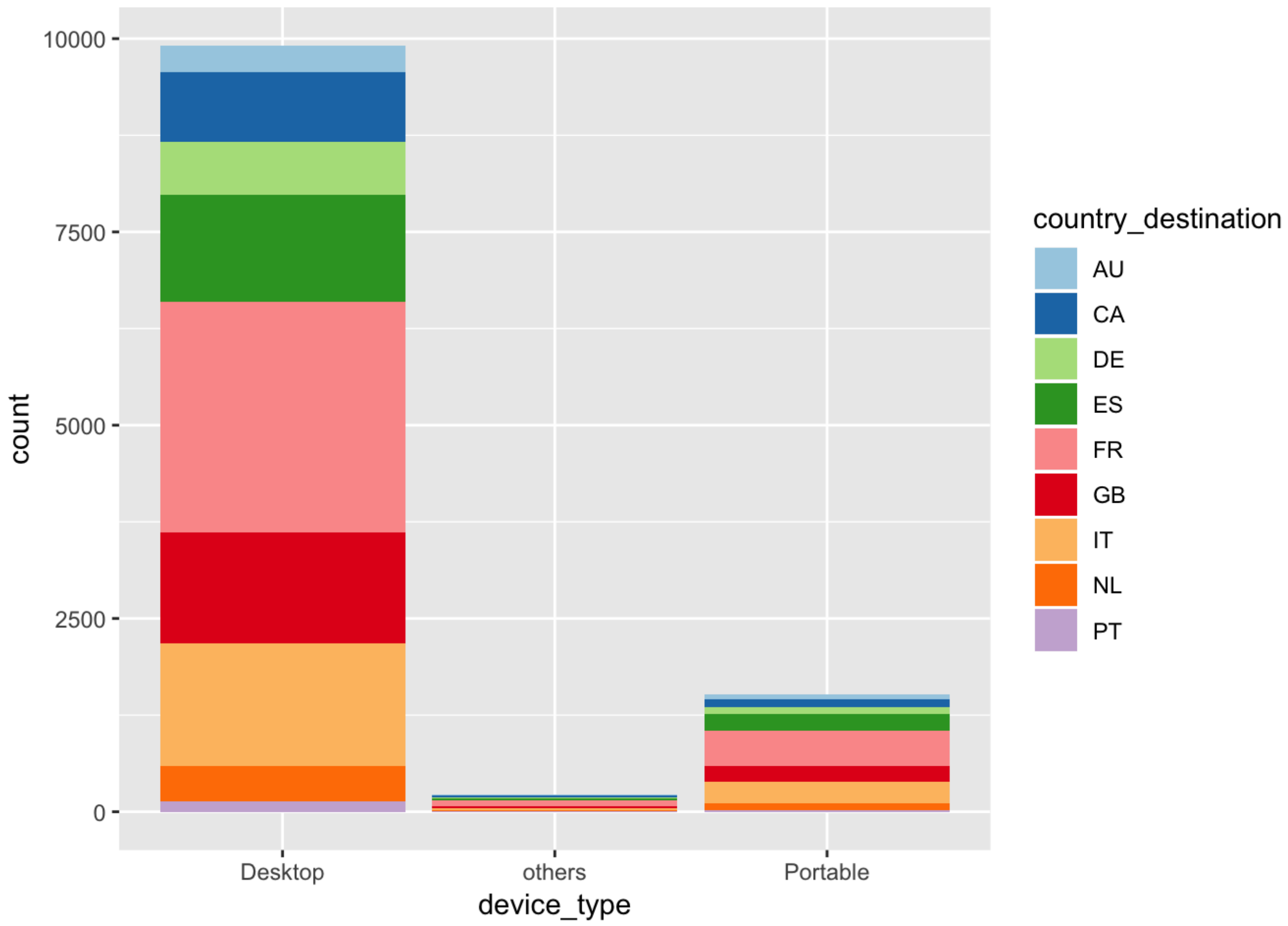
Choice of Technology (Device, Browser, OS)

After exploring age and gender, we can take a look at devise type, browser type, and signup methods. Do people use different browser type have different preference?

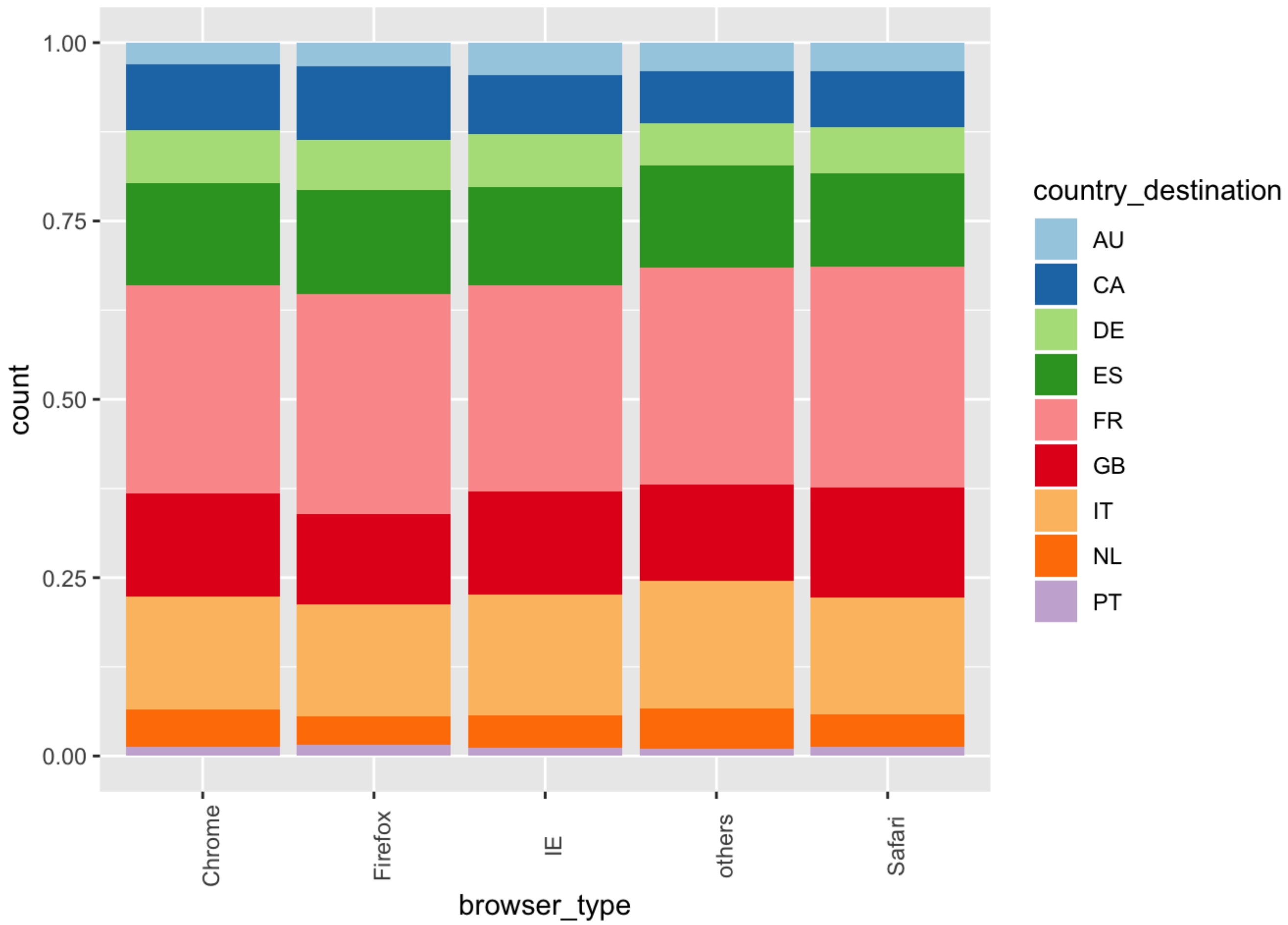
```
ggplot(train2, aes(x = device_type, fill = country_destination)) + geom_bar(position =
"fill") + scale_fill_brewer(palette="Paired")
```



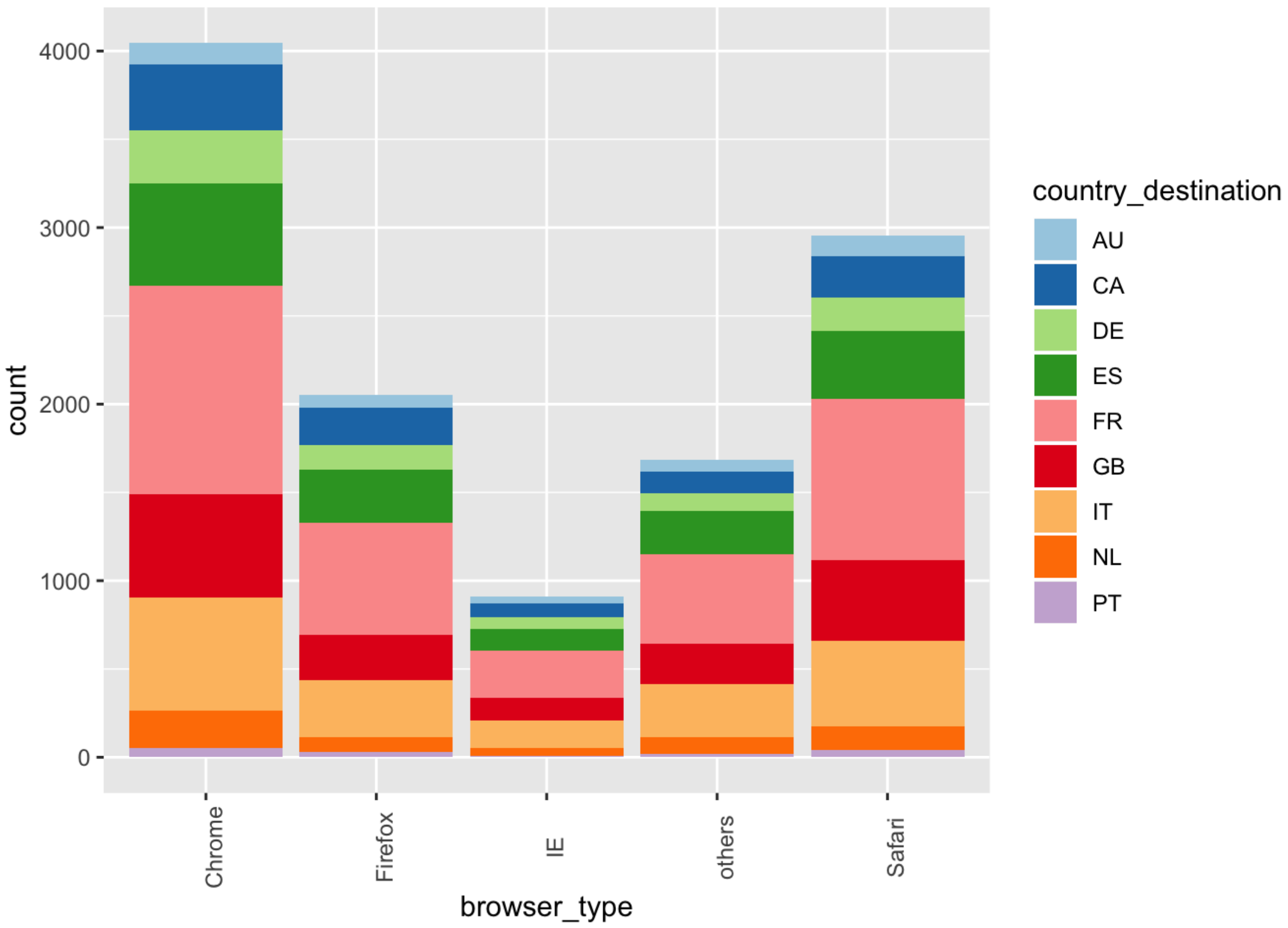
```
ggplot(train2, aes(x = device_type, fill = country_destination)) + geom_bar(position =
"stack") + scale_fill_brewer(palette="Paired")
```



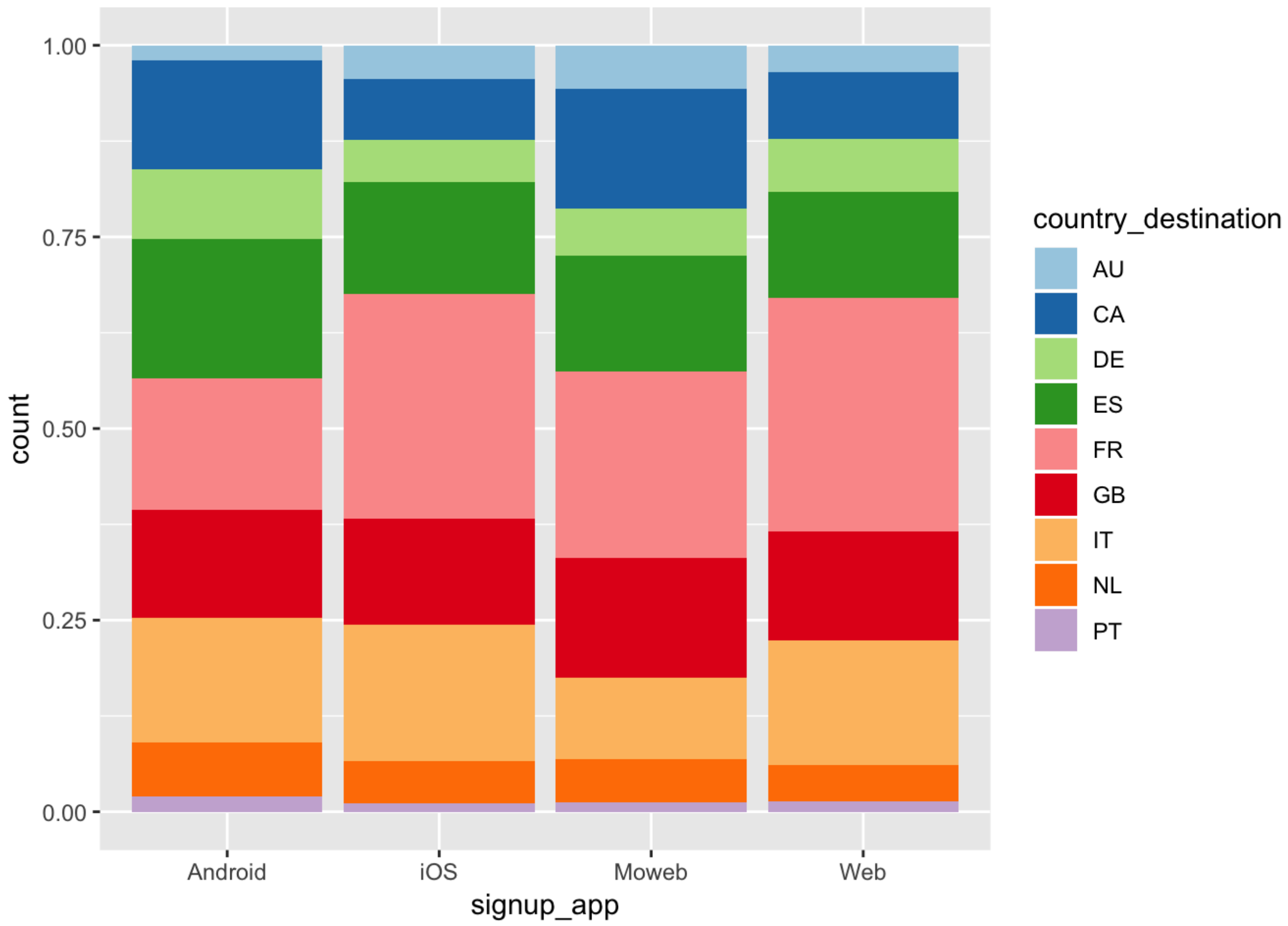
```
ggplot(train2, aes(x = browser_type, fill = country_destination)) + geom_bar(position =
"fill") + theme(axis.text.x = element_text(angle = 90)) + scale_fill_brewer(palette="Pai
red")
```



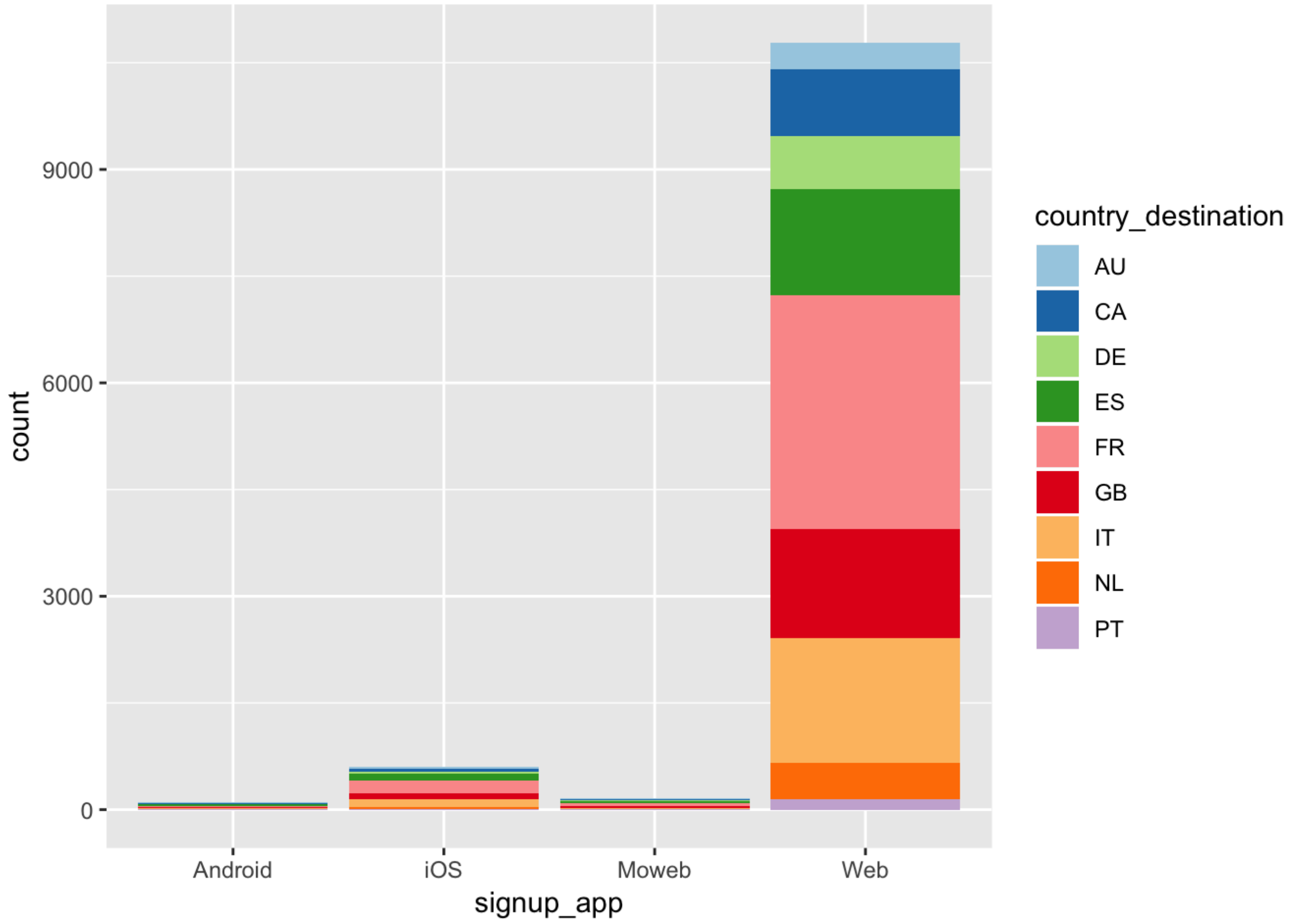
```
ggplot(train2, aes(x = browser_type, fill = country_destination)) +
geom_bar(position = "stack") + theme(axis.text.x = element_text(angle = 90)) + scale_fil
l_brewer(palette="Paired")
```



```
ggplot(train2, aes(x = signup_app, fill = country_destination)) + geom_bar(position = "fill") + scale_fill_brewer(palette="Paired")
```



```
ggplot(train2, aes(x = signup_app, fill = country_destination)) + geom_bar(position = "stack") + scale_fill_brewer(palette="Paired")
```

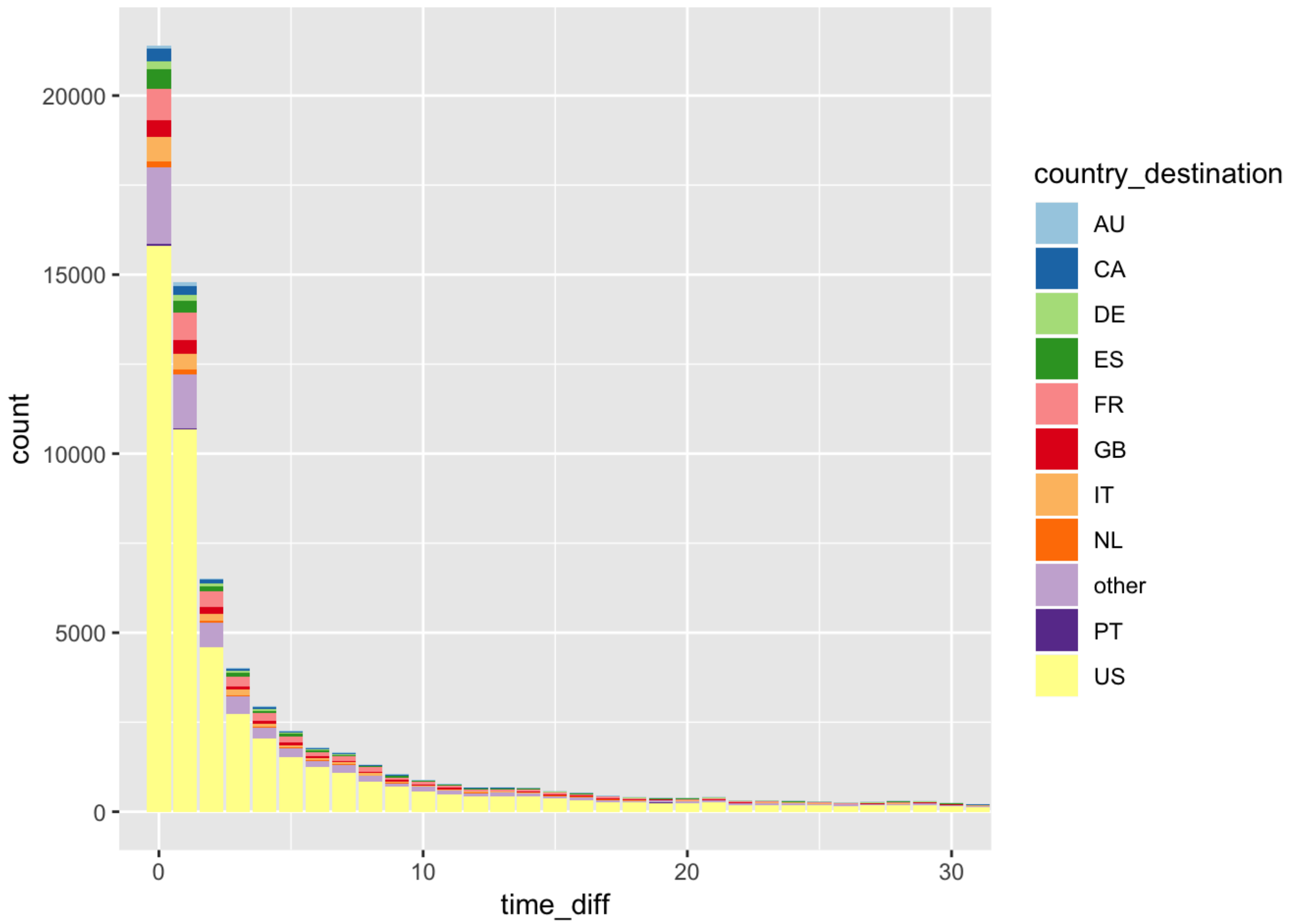


It does not seem like there's a strong correlation between device type / signup_app / browser type and country destination. Android users does show different proportion of country of destination; however, since the total population is a so small compared to other signup app, the difference should be disregarded.

Time

We have wwo important time-related attributes - timestamp_first_active and date_first_booking. We are not considering date_account_created because this is the same date as timestamp_first_active. I created a column time_diff to show the number of days for a user to make the first reservation after the date they are first active on Airbnb. You'll find that more than 50% make their first booking in the first three days.

```
train3 <- train %>% filter(time_diff >= 0)
ggplot(train3, aes(x = time_diff, fill = country_destination)) + geom_bar(position = "st
ack") + coord_cartesian(xlim=c(0,30)) + scale_fill_brewer(palette="Paired")
```



This graph appears to be a Poisson distribution. As time goes by, it is much less likely for someone to make a reservation on Airbnb.

Language

Last attribute I want to explore is the language preference of new user. English is chosen by more than 97% users to display on Airbnb. This doesn’t mean it is the primary language spoken by Airbnb users.

For those people who did choose other languages to display may give us some additional knowledge of the ethnicity or the country of origin of Airbnb users. Chinese is the second most used language on Airbnb. At the same time, top destination for these Chinese users is others, which means the destination countries they chose are not US or popular European counties.

The third and fourth language chosen are French and Spanish. And both groups of users preferred countries that have English as primary language. If we dig deeper in their second choice of country destinaton. These second choices tend to have a primary language that is the same as the user’s language. This is true for French, German and Italian users. We don’t know if this is the case for Chinses or Spanish users; however, they second most preferred destination is other, which could very well be a country with the same spoken language as the user’s.

```
train4 <- train1 %>% filter(country_destination != "NDF")
language_combo1 <- train4 %>%
  group_by(language_combo) %>%
  tally()
display_language_combo1 <- arrange(language_combo1,desc(n))
print(display_language_combo1,n=30)
```

```
## # A tibble: 105 x 2
##    language_combo                   n
##    <chr>                        <int>
##  1 English - English            48299
##  2 English - others              6960
##  3 English - French              3414
##  4 English - Italian             1849
##  5 English - Spanish             1570
##  6 English - German               763
##  7 English - Dutch                550
##  8 Chinese - English              328
##  9 French - English               212
## 10 Spanish; Castilian - English   151
## 11 English - Portuguese           147
## 12 German - English               142
## 13 Korean - English               110
## 14 Russian - English               58
## 15 Chinese - others                56
## 16 Italian - English               54
## 17 French - French                 48
## 18 Japanese - English              47
## 19 Portuguese - English            37
## 20 Spanish; Castilian - others     29
## 21 Korean - others                 26
## 22 Swedish - English               25
## 23 German - German                 18
## 24 German - others                 17
## 25 Spanish; Castilian - Spanish    16
## 26 Dutch - English                 15
## 27 French - others                 15
## 28 Italian - Italian               15
## 29 Portuguese - others             15
## 30 Danish - English                12
## # ... with 75 more rows
```

```
language1 <- train4 %>%
  group_by(language_full) %>%
  tally()
arrange(language1,desc(n))
```

```
## # A tibble: 23 x 2
##    language_full          n
##    <chr>              <int>
##  1 English            63552
##  2 Chinese              400
##  3 French               305
##  4 Spanish; Castilian   220
##  5 German               199
##  6 Korean               150
##  7 Italian               86
##  8 Russian               85
##  9 Japanese              61
## 10 Portuguese            54
## # ... with 13 more rows
```