Statistical Analysis - Airbnb New Users

Youzhu Shi

10/14/2018

The Data

Age

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

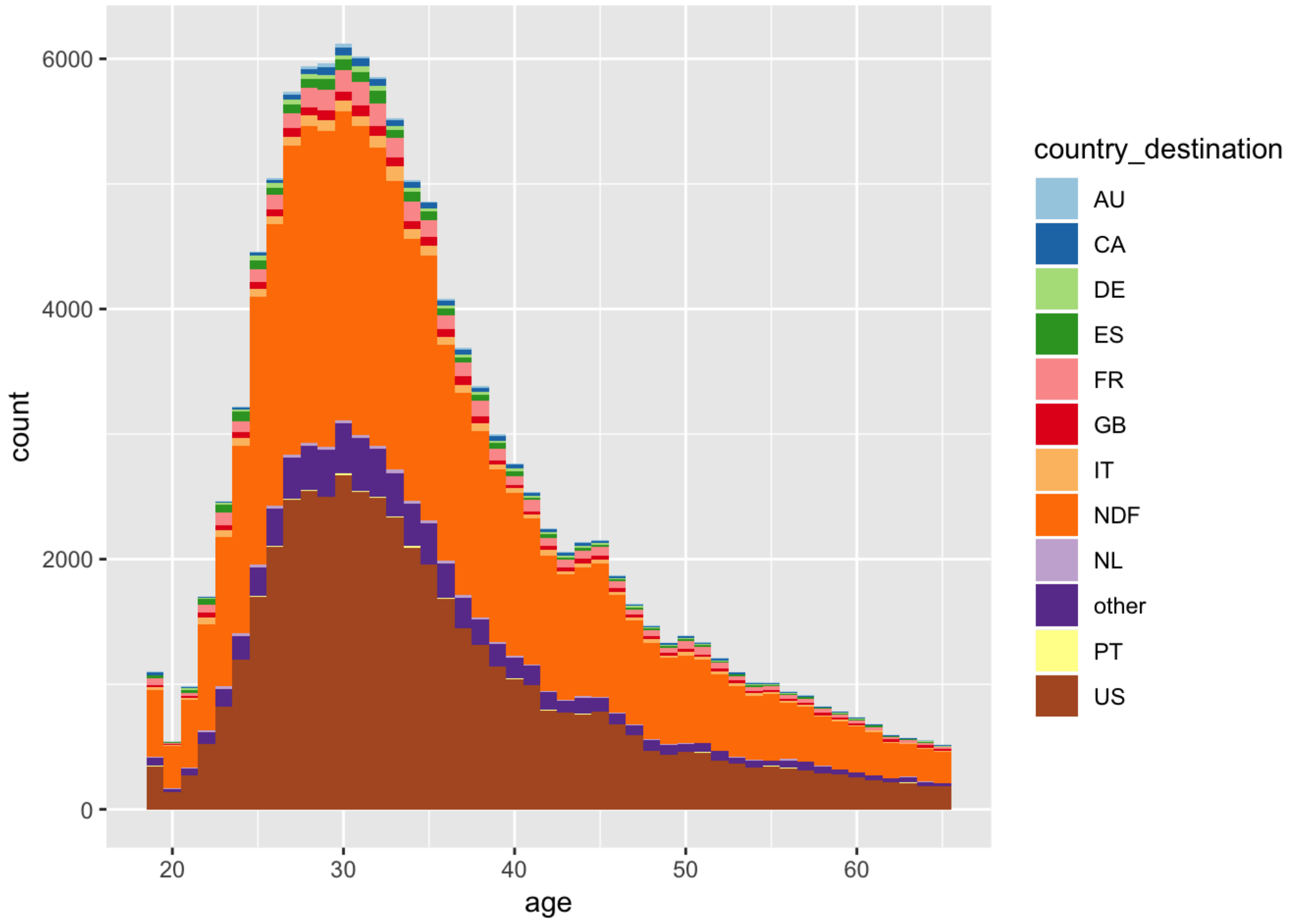
Age Summary

```
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

```
ggplot(train1, aes(x = age, fill = country_destination)) + geom_histogram(binwidth = 1,
 position = "stack") + scale_fill_brewer(palette="Paired")
```

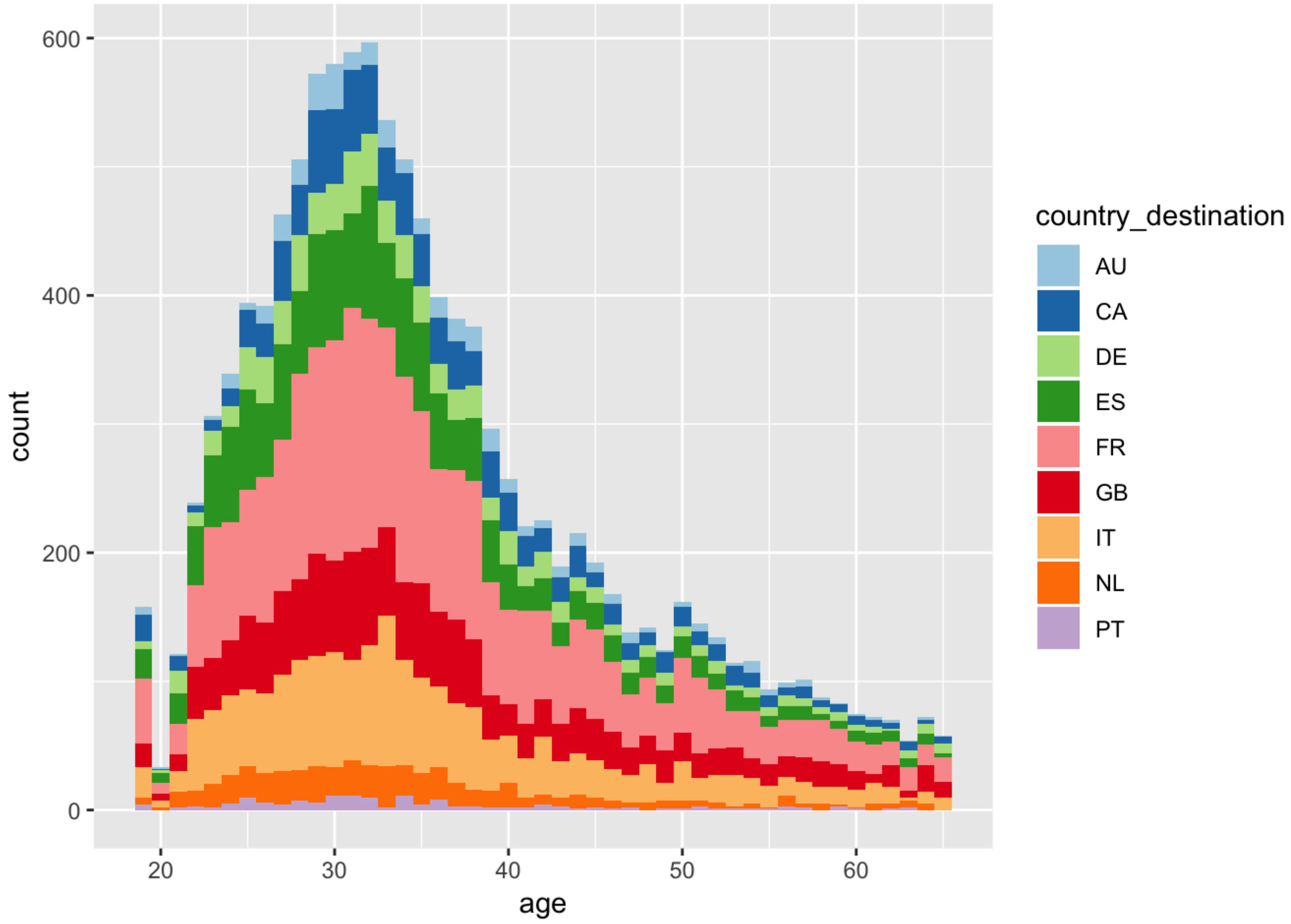


The Age of guests and the number of booking made on Airbnb appear to be a right skewed distribution. After filtering out guests who are less than 18 years old or over 99 years old, the mean age is 36.6, and the median age is 34. It has a first quartile of 28 years old and a 3rd quartile of 42 years old. 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

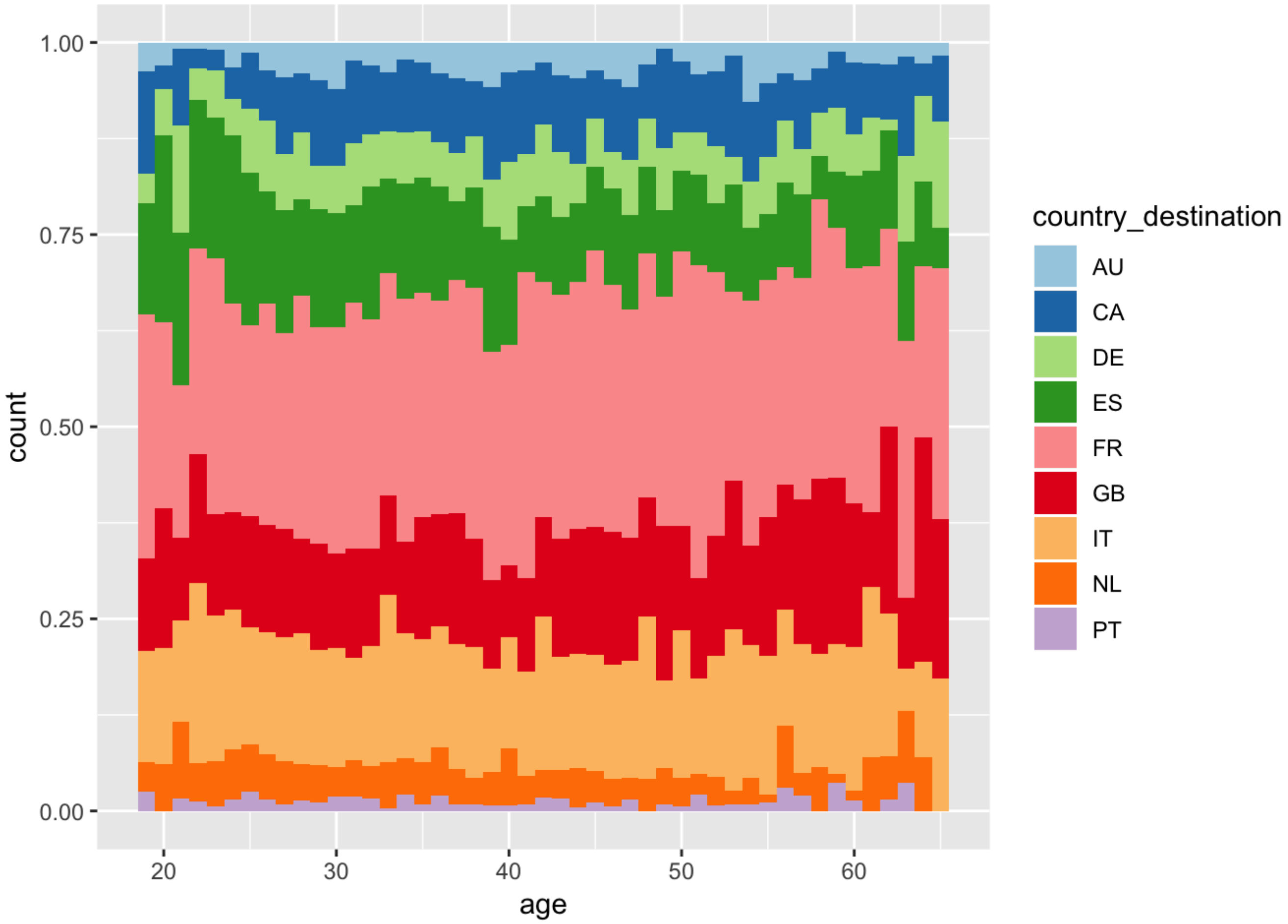
Now, let's take a look at those who chose countries outside of the US. What are their characteristics?

```
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")
```

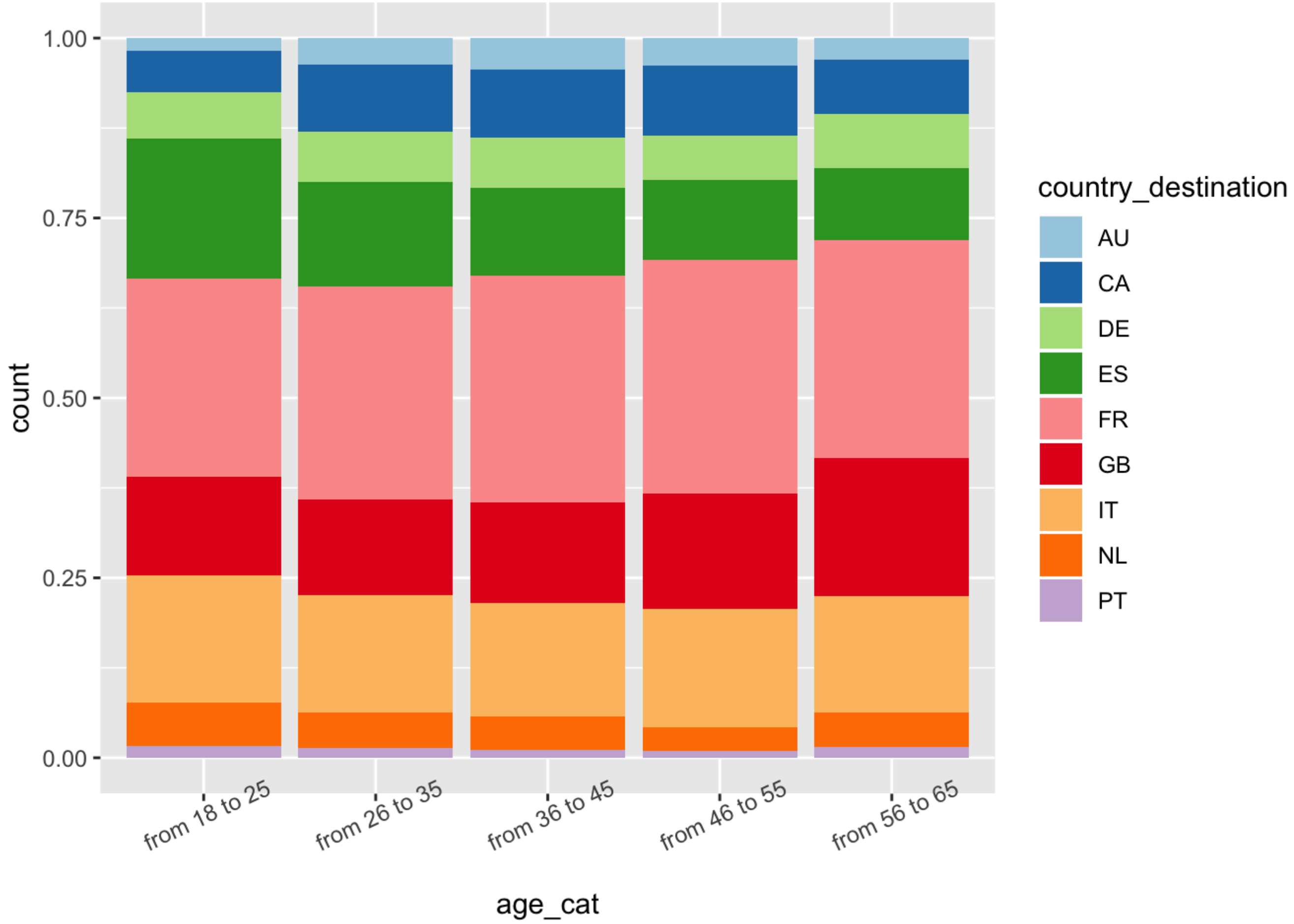


Country preferences appear fairly consistent amount different ages.

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"
)
```

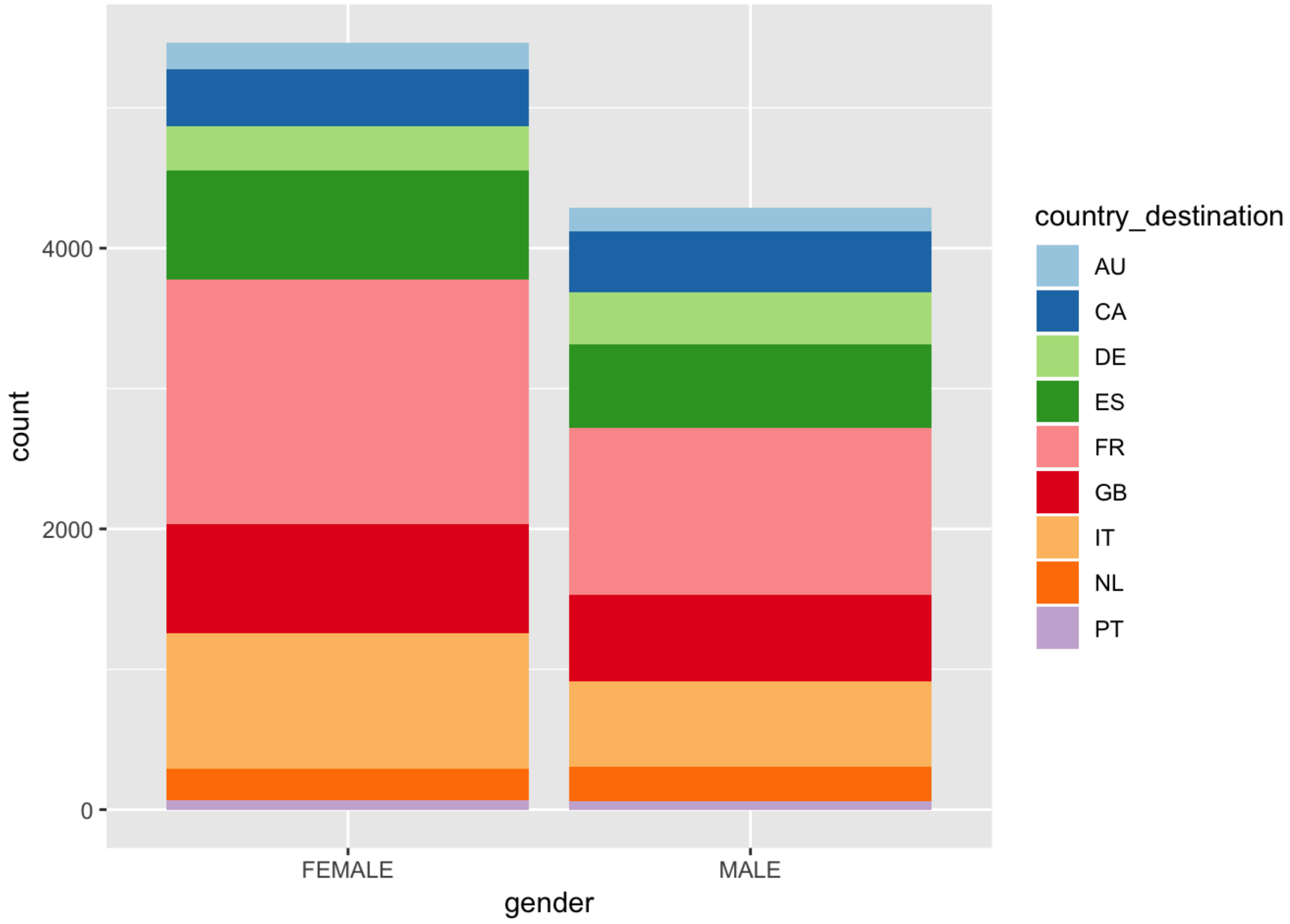


Spain is more popular among younger users; the opposite is true for Great Britain. Canada is more popular among middle aged users, this is probably because people who are in their working age do more business travels between the two countries as US and Canada have very close trade relations. They are 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?

```
ggplot(train2_1, aes(x = gender, fill = country_destination)) + geom_bar() +  scale_fill
_brewer(palette="Paired")
```

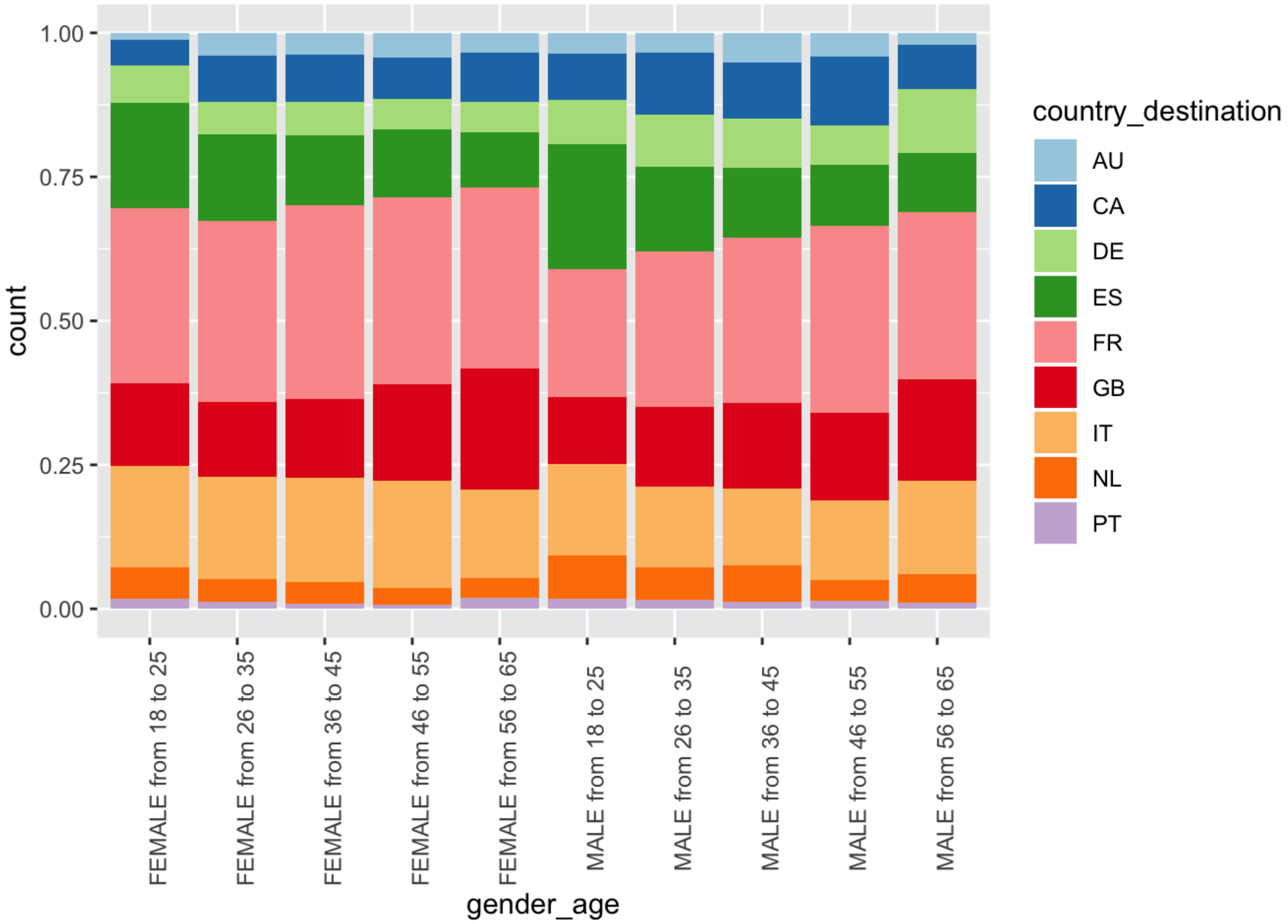


There are more female users than male users. It appears that female users have a much stronger preference for France.

Age & Gender

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

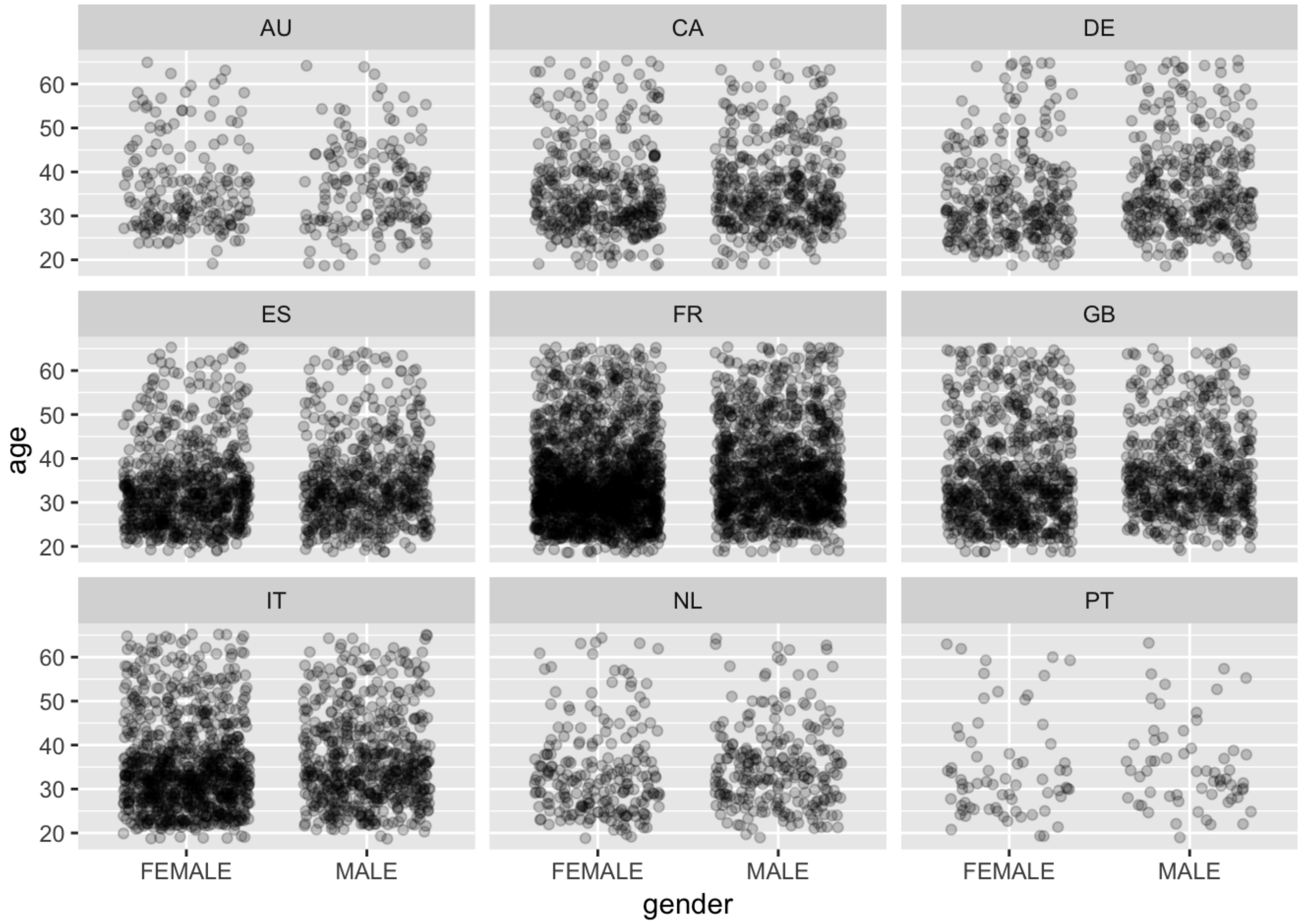
```
ggplot(train2_1, aes(x = gender_age, fill = country_destination)) + geom_bar(position =
"fill") + theme(axis.text.x = element_text(angle = 90)) + scale_fill_brewer(palette="Pai
red")
```



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)
```

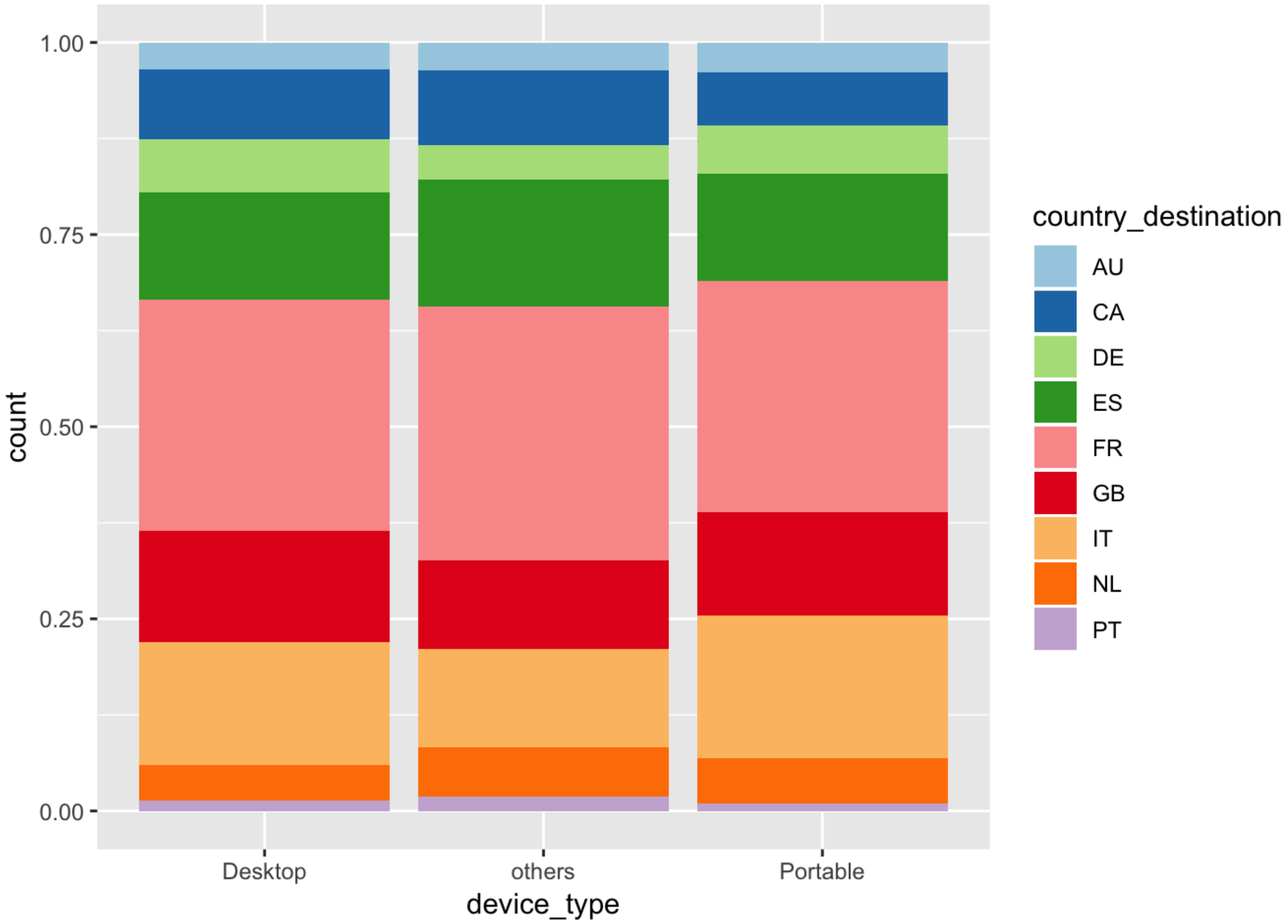


This facet grid reaffirmed us of our previous analysis. It does provide us additional information. The travel age for males is slight high than the travel age for female. This tendency appears strong in Great Britain, Germany, and France.

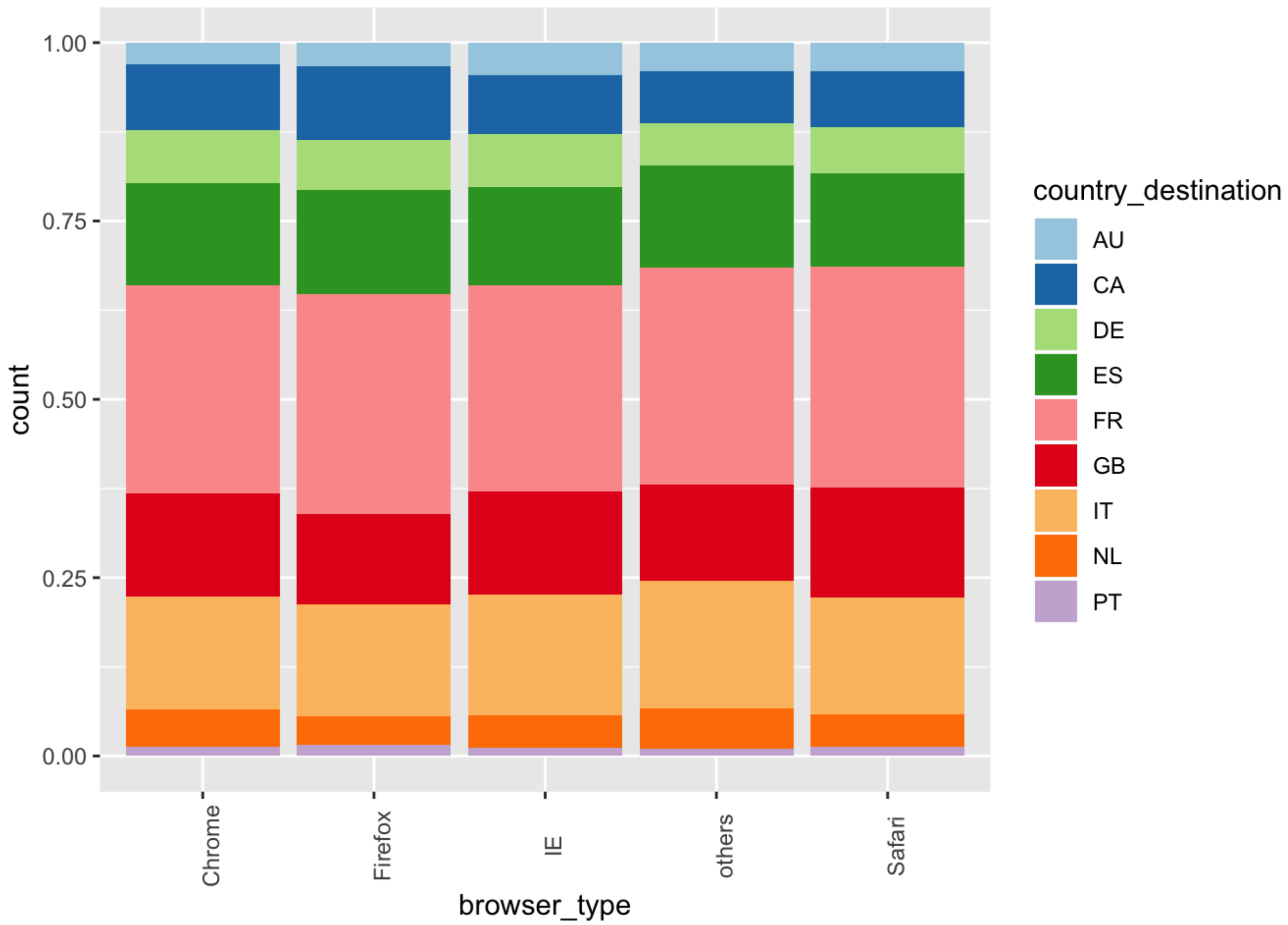
Choice of Technology (Device, Browser, OS)

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?

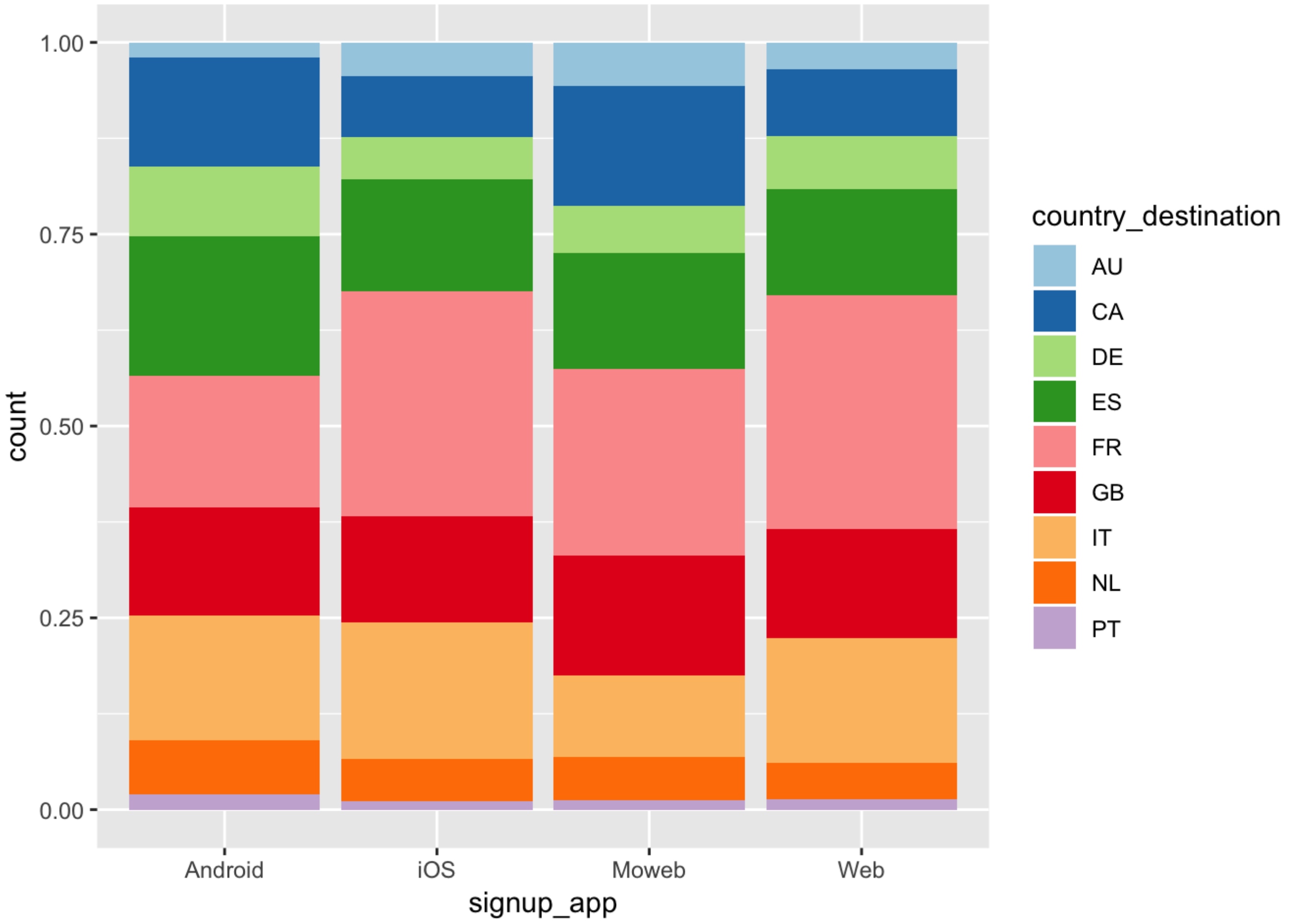
```
ggplot(train2, aes(x = device_type, fill = country_destination)) + geom_bar(position =
"fill") + scale_fill_brewer(palette="Paired")
```



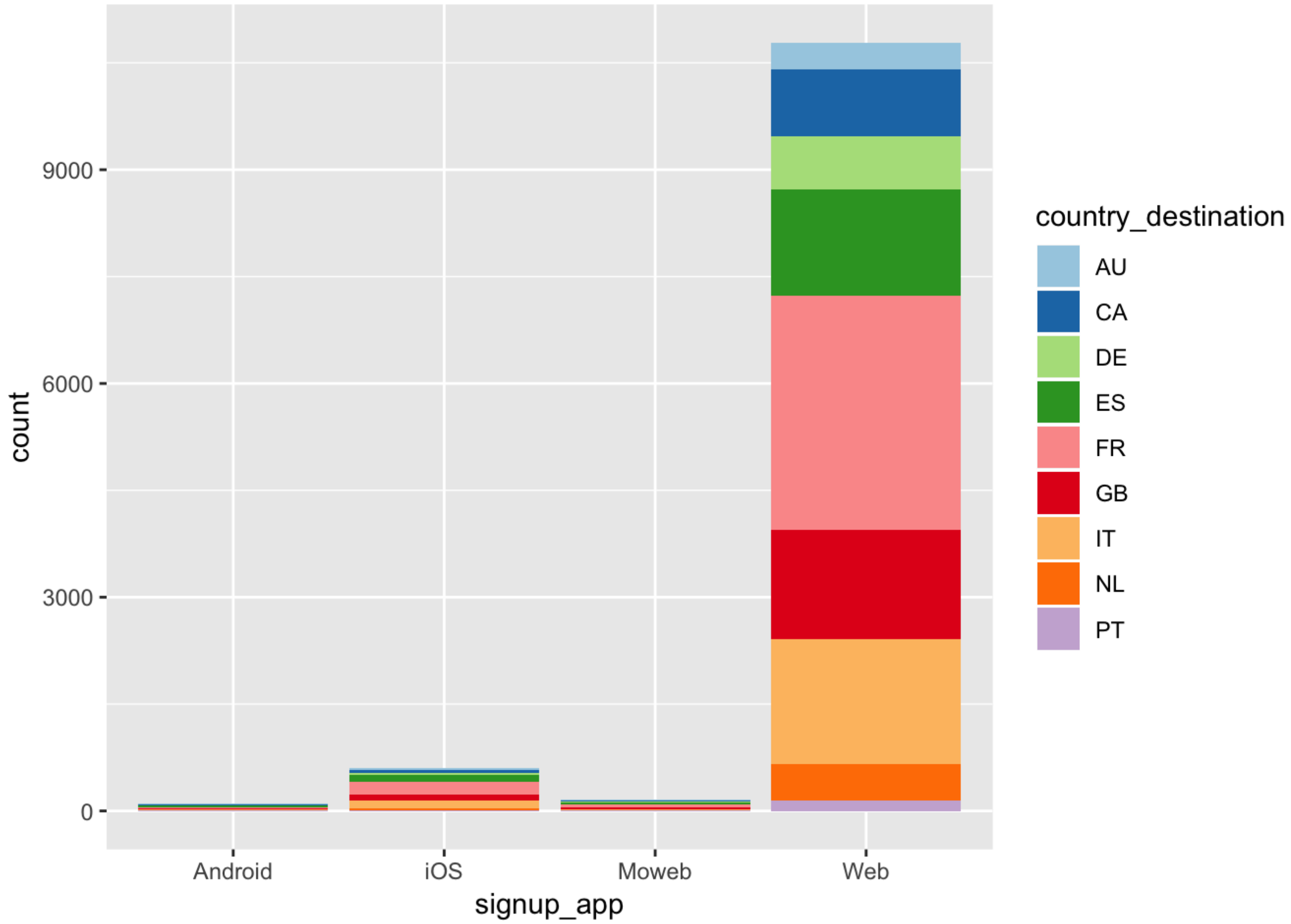
```
ggplot(train2, aes(x = browser_type, fill = country_destination)) + geom_bar(position =
"fill") + theme(axis.text.x = element_text(angle = 90)) + scale_fill_brewer(palette="Pai
red")
```



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



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

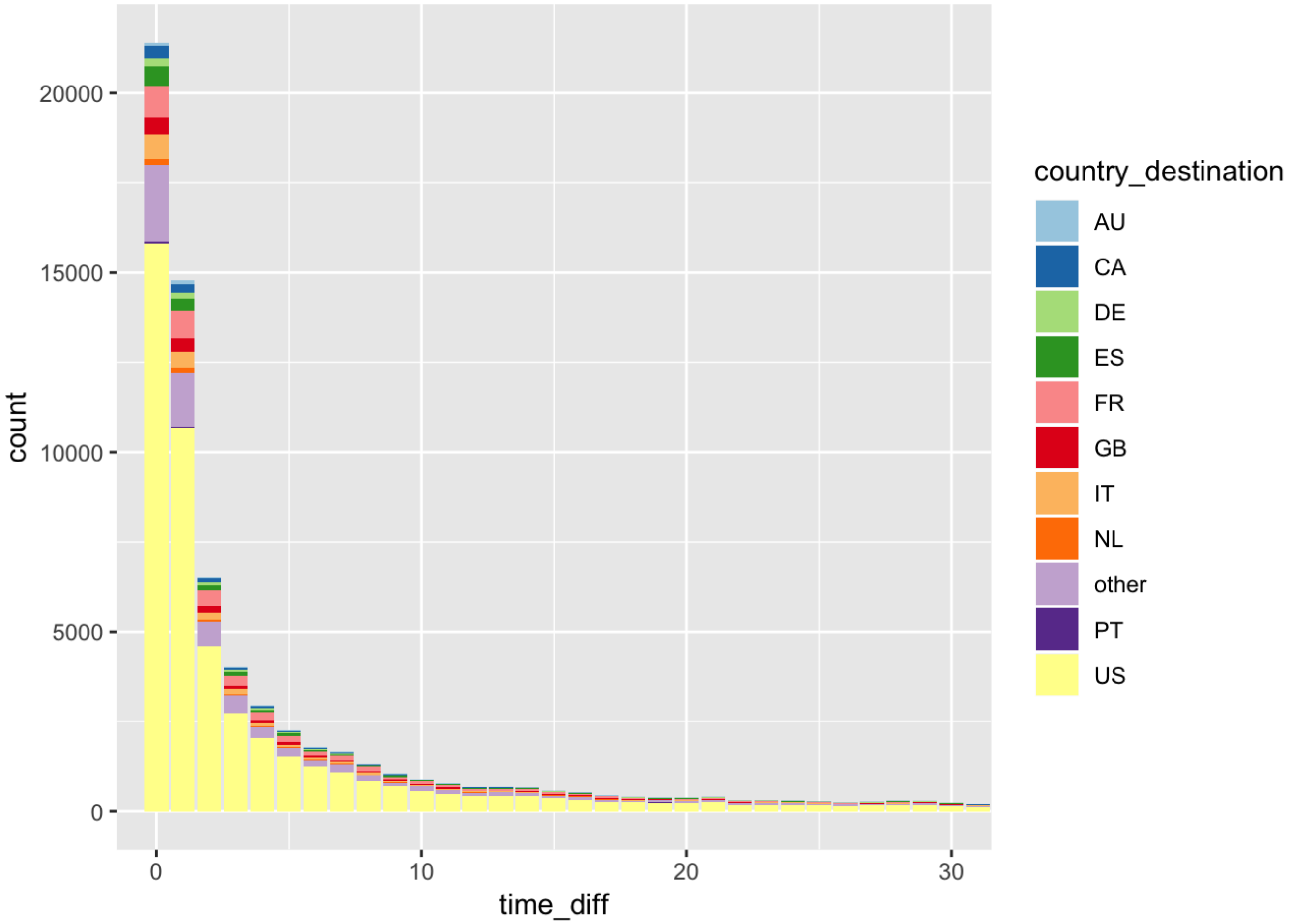


It does not seem like there's a strong correlation between device type / signup_app / browser type and country destination. Android users does show different proportion of country of destination; however, since the total population is a so small compared to other signup app, the difference should be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

```
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.

```
language_combo1 <- train4 %>%
  group_by(language_combo) %>%
  tally()
arrange(language_combo1,desc(n))
```

```
## # 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
## # ... with 95 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
```