STAT 184 project

STAT 184 Final Course Project

Hanzhe Jiang

2025-08-13

Analysis of Airbnb New User Booking

Airbnb is one of the top American companies that holds a marketplace for logging, vacation rentals, and tourism activities. From 2012 to 2014, the user population has increased from 6 million to 50 million. The most recent report shows the number of nights booked has increased to 356.9 million. (2 million users in worldwide) The main idea for this project is how to predict new user's intentions for their first booking through my motivation and method:

1. Support Destination-Specific-Advertising through User Accommodation

2. Explore Distributions of the Values of Interest through EDA

3. Predict or Infer User's First Book Destination from their Information

This process was done by solving Airbnb's Kaggle problems where they wanted Kaggleusers to predict where their users were most likely going to travel based on data from their website, and there's also data from Airbnb where they offer a detailed overview of Airbnb listings worldwide:

- Data source: Kaggle Competition, Airbnb reported source.
- Link: https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data?select=age_gender_bkts.csv.zip
- Link: https://www.kaggle.com/datasets/lovishbansal123/
- 213451 US users, 16 columns, Label Column: Country Destination (12 possible outcomes), and detailed overview of Airbnb listings worldwide

(@misc{airbnb-recruiting-new-user-bookings,

author = {alokgupta and Anna Montoya and LizSellier and Meghan O'Connell and Wendy Kan},

title = {Airbnb New User Bookings}, year = {2015}, howpublished = {https://kaggle.com/competitions/airbnb-recruiting-new-user-bookings}, note = {Kaggle} })

There are 5 datasets in cvs form from this two data source I believe could be supplied for this project:

1.Train_users_2: Train users contains data on 213.451 users from Airbnb, all data used for identification is replaced by an ID field.

2.Test users: Test users contains 62.096 users on which the prediction is supposed to be made, it uses the same format as Train_users with the exception of the date_first_booking and country_destination fields.

3.Sessions: Contains records of user's actions on the website

4.Age_gender_bkts: This dataset groups users in the training set into age groups of 5 years difference and shows information about each group's gender and decided country destination.

5.Countries: A summary of the different country destinations and various data on

Dataset Preview

```
##I'm going to use the BOAST Style Guide
library(ggplot2)
library(dcData)
library(tidyr)
library(tibble)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.1 --

v readr   2.1.2     v stringr 1.4.0
v purrr   0.3.4     v forcats 0.5.1

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
library(readr)

train_users <- read_csv("train_users_2.csv")
```

```
Rows: 213451 Columns: 16

-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (11): id, gender, signup_method, language, affiliate_channel, affiliate...
dbl   (3): timestamp_first_active, age, signup_flow
date  (2): date_account_created, date_first_booking

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(train_users)
```

```
# A tibble: 6 x 16
  id         date_account_created timestamp_first~ date_first_book~ gender   age
  <chr>      <date>                         <dbl> <date>           <chr>  <dbl>
1 gxn3p5htnn 2010-06-28                  2.01e13 NA               -unkn~    NA
2 820tgsjxq7 2011-05-25                  2.01e13 NA               MALE      38
3 4ft3gnwmtx 2010-09-28                  2.01e13 2010-08-02       FEMALE    56
4 bjjt8pjhuk 2011-12-05                  2.01e13 2012-09-08       FEMALE    42
5 87mebub9p4 2010-09-14                  2.01e13 2010-02-18       -unkn~    41
6 osr2jwljor 2010-01-01                  2.01e13 2010-01-02       -unkn~    NA
# ... with 10 more variables: signup_method <chr>, signup_flow <dbl>,
#   language <chr>, affiliate_channel <chr>, affiliate_provider <chr>,
#   first_affiliate_tracked <chr>, signup_app <chr>, first_device_type <chr>,
#   first_browser <chr>, country_destination <chr>
```

```r
age_gender_bkts <- read_csv("age_gender_bkts.csv")
```

```
Rows: 420 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): age_bucket, country_destination, gender
dbl (2): population_in_thousands, year

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
head(age_gender_bkts)
```

```
# A tibble: 6 x 5
  age_bucket country_destination gender population_in_thousands  year
  <chr>      <chr>               <chr>                    <dbl> <dbl>
1 100+       AU                  male                         1  2015
2 95-99      AU                  male                         9  2015
3 90-94      AU                  male                        47  2015
4 85-89      AU                  male                       118  2015
5 80-84      AU                  male                       199  2015
6 75-79      AU                  male                       298  2015
```

```r
countries <- read_csv("countries.csv")
```

```
Rows: 10 Columns: 7
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (2): country_destination, destination_language
dbl (5): lat_destination, lng_destination, distance_km, destination_km2, lan...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
head(countries)
```

```
# A tibble: 6 x 7
  country_destinati~ lat_destination lng_destination distance_km destination_km2
  <chr>                        <dbl>           <dbl>       <dbl>           <dbl>
```

```
1 AU                            -26.9          133.        15298.         7741220
2 CA                             62.4          -96.8        2828.         9984670
3 DE                             51.2           10.5        7880.          357022
4 ES                             39.9           -2.49       7731.          505370
5 FR                             46.2            2.21       7683.          643801
6 GB                             54.6           -3.43       6884.          243610
# ... with 2 more variables: destination_language <chr>,
#   language_levenshtein_distance <dbl>
```

```
TestUsers <- read_csv("test_users.csv")
```

```
Rows: 62096 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (10): id, gender, signup_method, language, affiliate_channel, affiliate...
dbl   (3): timestamp_first_active, age, signup_flow
lgl   (1): date_first_booking
date  (1): date_account_created

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(TestUsers)
```

```
# A tibble: 6 x 15
  id         date_account_created timestamp_first~ date_first_book~ gender   age
  <chr>      <date>                         <dbl> <lgl>            <chr>  <dbl>
1 5uwns89zht 2014-07-01                  2.01e13 NA               FEMALE    35
2 jtl0dijy2j 2014-07-01                  2.01e13 NA               -unkn~    NA
3 xx0ulgorjt 2014-07-01                  2.01e13 NA               -unkn~    NA
4 6c6puo6ix0 2014-07-01                  2.01e13 NA               -unkn~    NA
5 czqhjk3yfe 2014-07-01                  2.01e13 NA               -unkn~    NA
6 szx28ujmhf 2014-07-01                  2.01e13 NA               FEMALE    28
# ... with 9 more variables: signup_method <chr>, signup_flow <dbl>,
#   language <chr>, affiliate_channel <chr>, affiliate_provider <chr>,
#   first_affiliate_tracked <chr>, signup_app <chr>, first_device_type <chr>,
#   first_browser <chr>
```

EDA visualization

Device usage

Close to 30% of the population use Facebook as their sign up method, and over 70% of the population use basic. Google is not preferred by users as their sign up method.

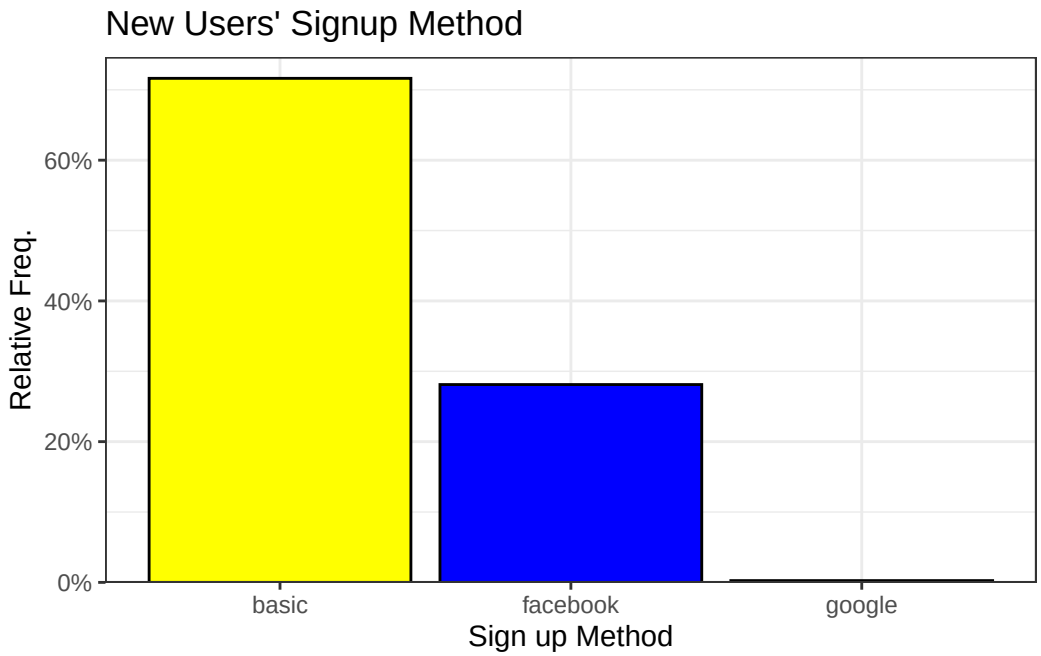
```
## New user's sign up method
train_users %>%
  count(signup_method) %>%
  mutate(prop = n / sum(n)) %>%
  arrange(desc(n))
```

```
# A tibble: 3 x 3
  signup_method      n    prop
  <chr>          <int>   <dbl>
1 basic         152897 0.716
2 facebook       60008 0.281
3 google           546 0.00256
```

```
  select(signup_method) %>%
  group_by(signup_method)%>%

  ggplot(
    data = train_users,
    mapping = aes(
      x = signup_method,
      y = after_stat(prop),
      group = 1)
  ) +
  geom_bar(
    color = "black",
    fill = c("basic" = "yellow",
             "facebook"   = "blue",
             "google"  = "red")
  ) +
  labs(
    title = "New Users' Signup Method",
    x = "Sign up Method",
    y = "Relative Freq."
  ) +
  theme_bw() +
```

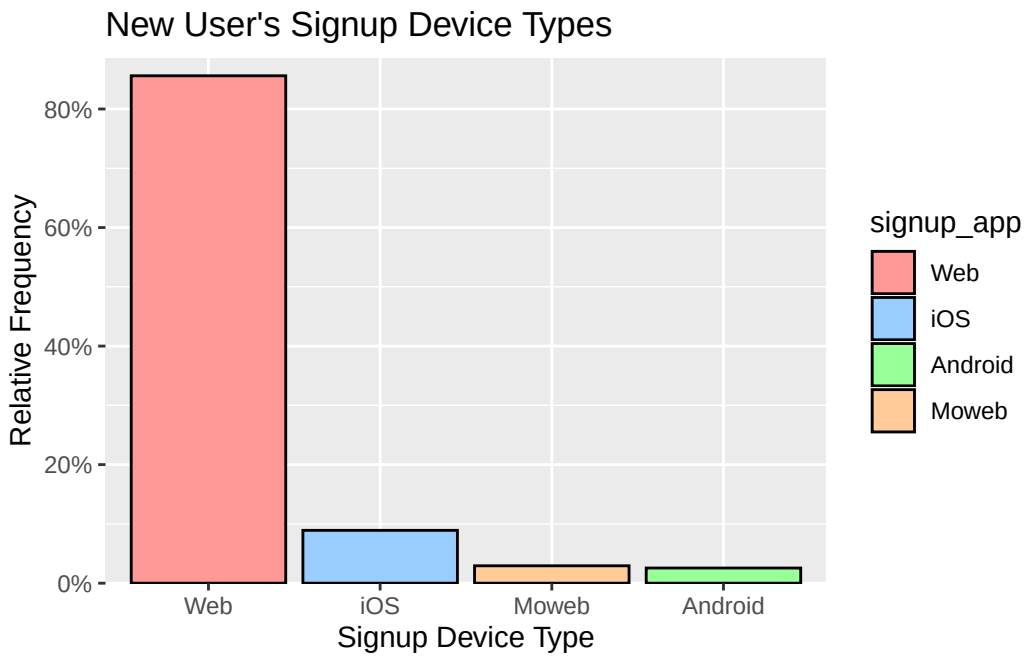
```
  scale_y_continuous(
    labels = scales::percent,
    expand = expansion(add = c(0, 0.03))
  )
```



In the training set, Most people use Desktop web for their first device sign up

```
train_users %>%
  count(signup_app) %>%
  mutate(prop = n / sum(n)) %>%
  ggplot(
    aes(
      x = reorder(signup_app,desc(prop)),
      y = prop,
      fill = signup_app)) +
  geom_bar(
    stat = "identity",
    color = "black")+
  scale_fill_manual(values = c(
    "Web"    = "#FF9998",
    "iOS"    = "#99CCFF",
    "Android" = "#99FF99",
    "Moweb"   = "#FFCC99"
```

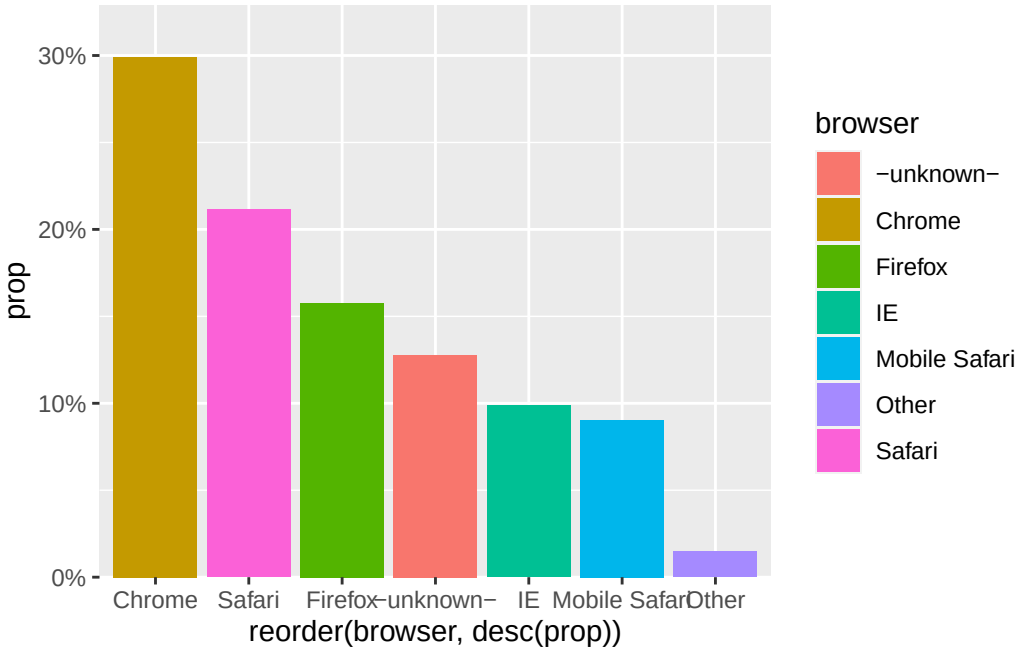
```
  )) +
  labs(
    title = "New User's Signup Device Types",
    x = "Signup Device Type",
    y = "Relative Frequency"
  ) +
  scale_y_continuous(
    labels = scales::percent,
    expand = expansion(add = c(0, 0.03))
  )
```



More people using Chrome than using Safari as their first browser for signing up.

```
train_users %>%
  count(first_browser) %>%
  mutate(prop = n / sum(n),
         browser = if_else(
           prop < 0.02,
           "Other",
           first_browser)) %>%
  group_by(browser) %>%
  summarise(n = sum(n)) %>%
  mutate(prop = n / sum(n)) %>%
```

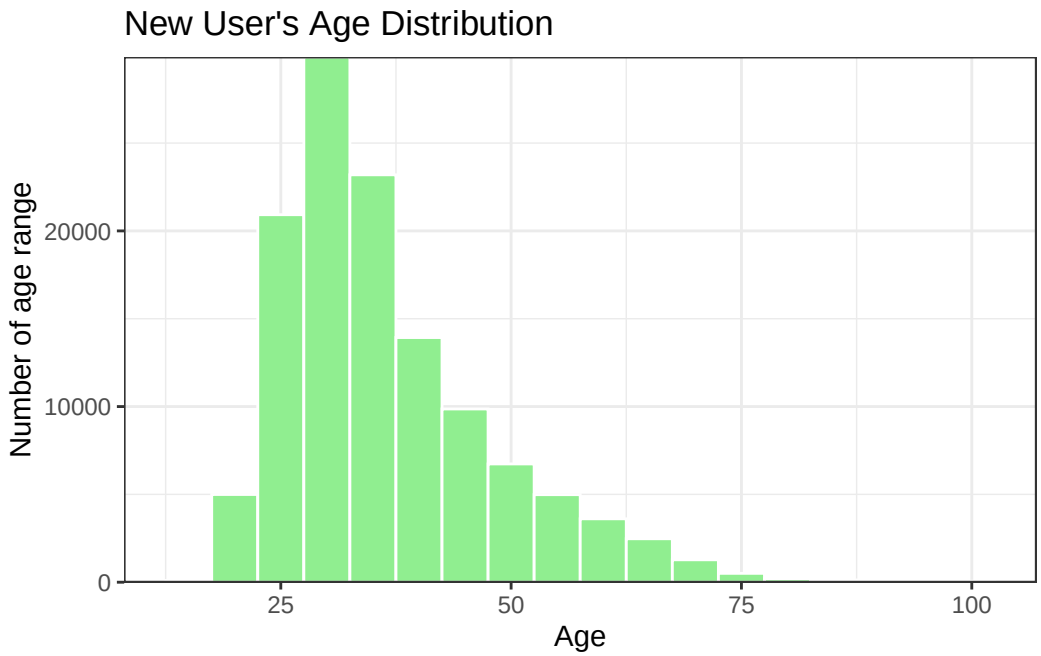
```
  ggplot(
    aes(
      x = reorder(browser, desc(prop)),
      y = prop,
      fill = browser)) +
  geom_col() +
  scale_y_continuous(
    labels = scales::percent,
    expand = expansion(add = c(0, 0.03))
  )
```



User Population

```
# 3.Age distribution
train_users %>%
  select(age)%>%
  filter(age >= 15 & age <= 100) %>%
  ggplot(
    aes(x = age))+
  geom_histogram(
    color = "white",
```

```
    fill = "light green",
    binwidth = 5
    ) +
  labs(
    title = "New User's Age Distribution",
    x = "Age",
    y = "Number of age range")+
  theme_bw() +
  scale_y_continuous(
    expand = expansion(add = c(0, 0.03))
  )
```

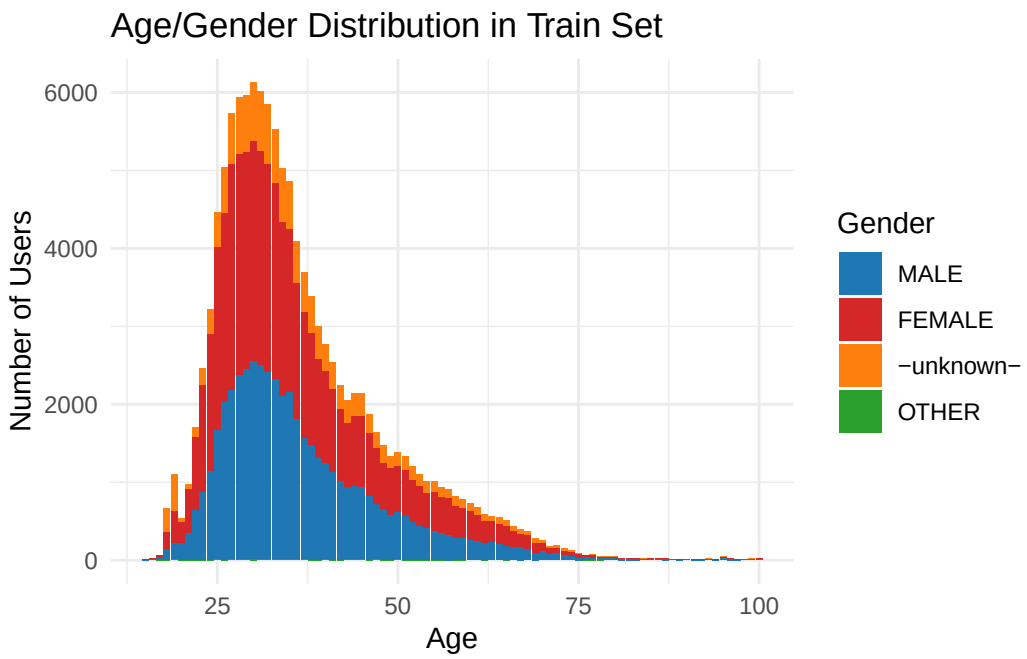


Combine the User's Age and gender together.

```
age_gender <- train_users %>%
  filter(age >= 15, age <= 100) %>%
  group_by(age, gender) %>%
  summarise(n = n())
```

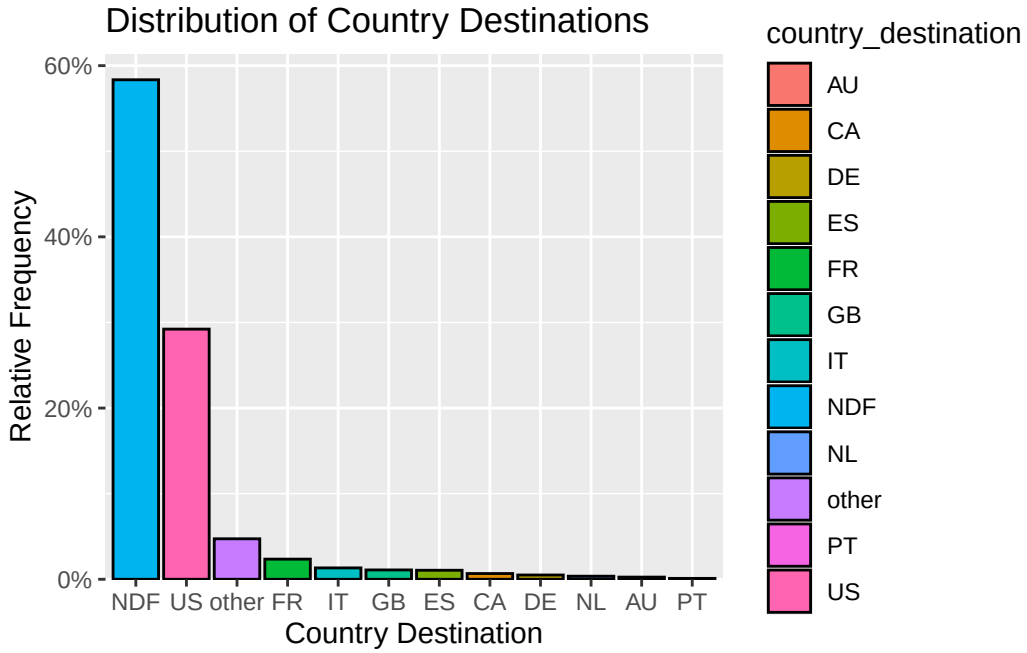
```
`summarise()` has grouped output by 'age'. You can override using the `.groups`
argument.
```

```r
ggplot(
  data = age_gender,
  aes(
    x = age,
    y = n,
    fill = gender)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(
    values = c(
    "MALE" = "#1f77b4",
    "FEMALE" = "#d62728",
    "-unknown-" = "#ff7f0e",
    "OTHER" = "#2ca02c"
  )) +
  labs(
    title = "Age/Gender Distribution in Train Set",
    x = "Age",
    y = "Number of Users",
    fill = "Gender"
  ) +
  theme_minimal()
```



Countries

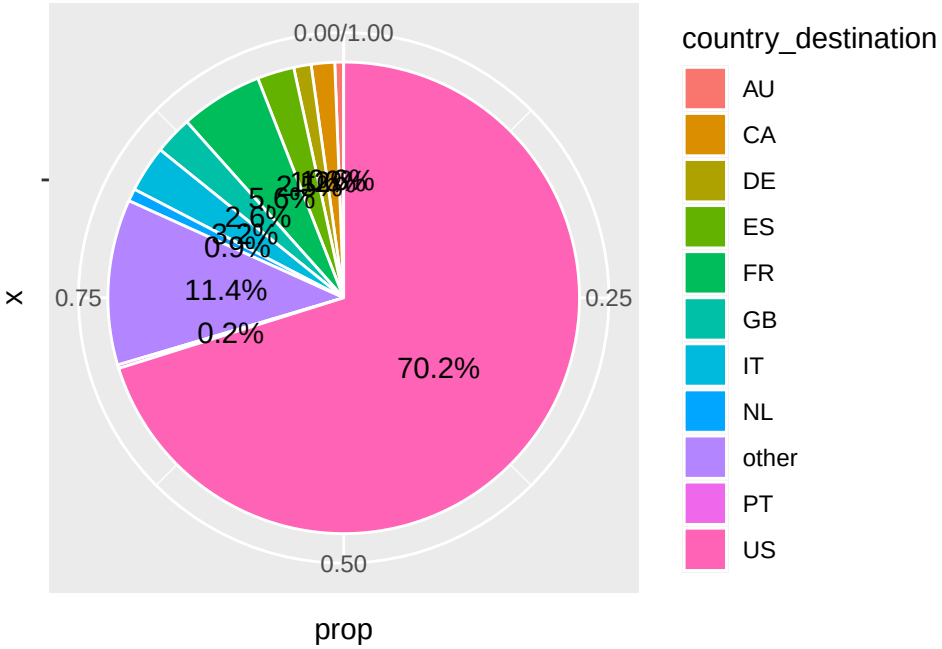
```
train_users %>%
  count(country_destination) %>%
  mutate(prop = n / sum(n)) %>%
  arrange(desc(prop)) %>%
  ggplot(
    aes(
      x = reorder(country_destination, desc(prop)),
      y = prop,
      fill = country_destination)) +
  geom_col(color = "black") +
  labs(
    title = "Distribution of Country Destinations",
    x = "Country Destination",
    y = "Relative Frequency"
  ) +
  scale_y_continuous(
    labels = scales::percent,
    expand = expansion(add = c(0, 0.03))
  )
```



Noticed that there's almost 60% of the training data with Not defined as user's first country

destination. So I removed the NDF value and made a pie chart. 70.2% of new user would prefer US as their first destination.

```
train_users %>%
  filter(country_destination != "NDF") %>%
  count(country_destination) %>%
  mutate(prop = n / sum(n)) %>%
  ggplot(
    aes("",
        prop,
        fill = country_destination)) +
  geom_col(
    color = "white") +
  coord_polar(theta = "y") +
  geom_text(aes(label = scales::percent(prop, 0.1)),
            position = position_stack(vjust = 0.5))
```



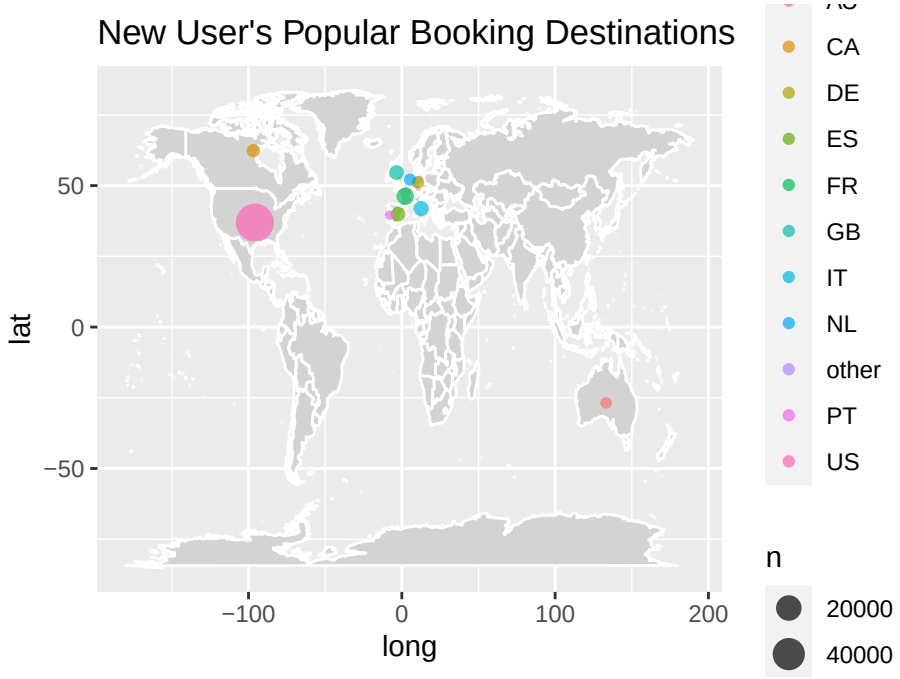
A world map would also be clear way to visualize user's prefernce for the first destination countries

```
DestCount <- train_users %>%
  filter(country_destination != "NDF") %>%
  count(country_destination)
DestMap <- DestCount %>%
```

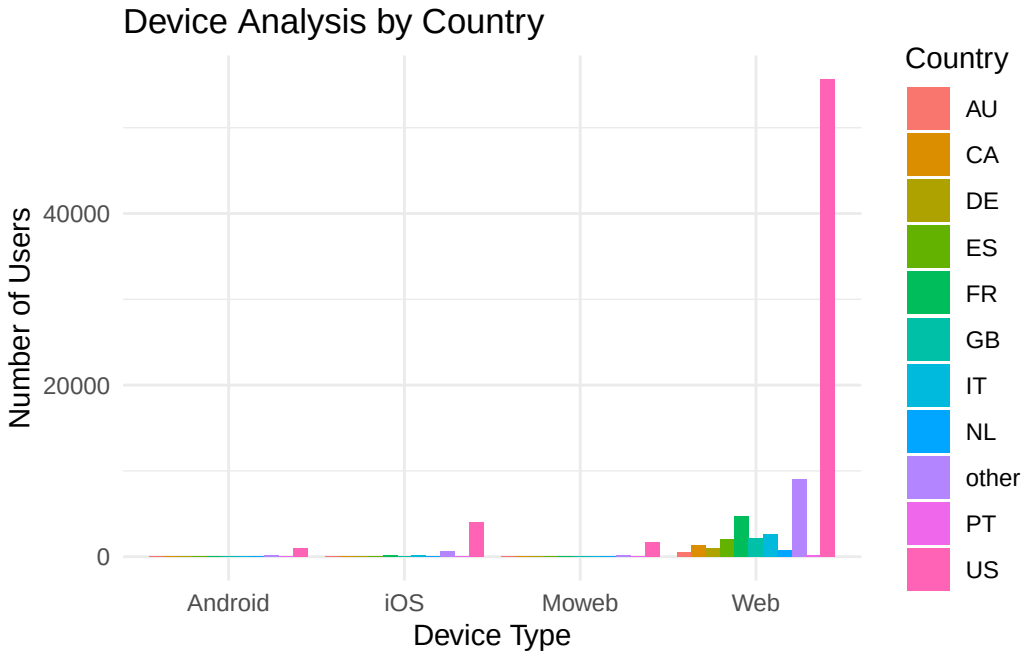
```
  left_join(countries, by = "country_destination")
world_map <- map_data("world")
ggplot() +
  geom_map(
    data = world_map,
    map = world_map,
    aes(
      long,
      lat,
      map_id = region),
    color = "white",
    fill = "lightgray"
  ) +
  geom_point(
    data = DestMap,
    aes(x = lng_destination,
        y = lat_destination,
        size = n,
        color = country_destination),
    alpha = 0.7
  ) +
  labs(title = "New User's Popular Booking Destinations")
```

```
Warning: Ignoring unknown aesthetics: x, y
```

```
Warning: Removed 1 rows containing missing values (geom_point).
```




```
##device analyse by countries
train_users %>%
  filter(country_destination != "NDF") %>%
  count(signup_app, country_destination) %>%
  ggplot(
    aes(x = signup_app,
        y = n,
        fill = country_destination)) +
  geom_col(position = "dodge") +
  labs(
    title = "Device Analysis by Country",
    x = "Device Type",
    y = "Number of Users",
    fill = "Country"
  ) +
  theme_minimal()
```



In Summary

- From Analysis, We see majority of the users are people from Age 20 to 40
- Majority of Users use Web Desktop, corresponding to Google Chrome and Safari
- Instead of NDF destination, the most first booking destination is US.

Therefore, we would like to recommend Airbnb to advertise Local Tourism Locations, primarily to Young Age user population through Ads on mainly on websites, and optimize their phone applications.

By accurately predicting where a new user will book their first travel experience, Airbnb can improve more personalized content in their community, lower the average waiting time of new customer's first booking, and better forecast new user's demand.