DATA INTENSIVE COMPUTING
PROJECT PHASE – 1

PREDICTING A NEW AIRBNB USER'S FIRST DESTINATION

MANISH CHAVA – (mchava2, 50475258)
SAI SREEKAR REDDY SIDDAMREDDY – (ssiddamr, 50460534)

PROBLEM STATEMENT:

Airbnb is a multi-million-dollar company which lets user rent their home online. This company was founded in 2007 and ever since it has been a pioneer in the industry, being in the market for almost 2 decades, the company has so much data, which analyzed thoughtfully can provide very useful information to the company which can help them to sustain in the current competitive world.
The data which we have taken for this phase and the entire project is from Kaggle[1]. The data contains information about user's demographics, web sessions etc. (Columns have been explained below), using this data about the user we are trying to predict the new user's first destination. With this analysis or prediction, the company can develop their recommendation systems and can recommend homestays to a user in the predicted destination.

COLUMNS DESCRIPTION:

id: user id

date_account_created: The date user has created an account.

timestamp_first_active: The time when an activity was recorded on user's device this might be before the account was created also because the website allows users to surf without creating an account.

date_first_booking: The date on which user has made the first booking.

gender: Sex of the user

age: Age of the user

[1] https://www.kaggle.com/competitions/airbnb-recruiting-new-user-bookings/data

signup_method: Through which user has signed up, examples through facebook, gmail etc.
signup_flow: From where the user came to sign up webpage.
language: International language preference
affiliate_channel: Kind of paid marketing
affiliate_provider: Advertisement provider.
first_affiliate_tracked: Among the advertisements what's the first one which user has interacted with.
signup_app: The application used by the user for signing up.
first_device_type: The device used the user while signing up.
first_browser: The browser through which user has signed up.
country_destination: This is the target variable which we are trying to predict.

Before we start working on building a model it is essential to clean the data because a model with null values or outliers would be biased and it's predictions would not be dependable.

DATA CLEANING :

i) Dropping rows : Rows with null values were dropped to calculate the average.

ii) Updating Data Type(s) of the column : The column 'date_first_booking' which has the data type of 'object' has been updated into datetime64 datatype for EDA.

iii) Dropping Columns : Columns which have might not have significant impact on the model prediction have been dropped. Column 'signup_flow' has been dropped.

iv) Removing Outliers : Before filling the Null values of few columns, the outliers were dropped because of their impact on the mean.

v) Check/Drop for duplicate rows : As there are no duplicate rows in the dataset nothing have been dropped from the dataset.

vi) Setting the index values : After dropping certain rows, the rows indices were reset to the natural order.

vii) Label Encoding : Label Encoding the categorical variables to make the data ready to be fed to a machine learning algorithm

viii) Adding new columns : New column(s) '**time_to_first_booking**' have been added to the dataframe to perform EDA.

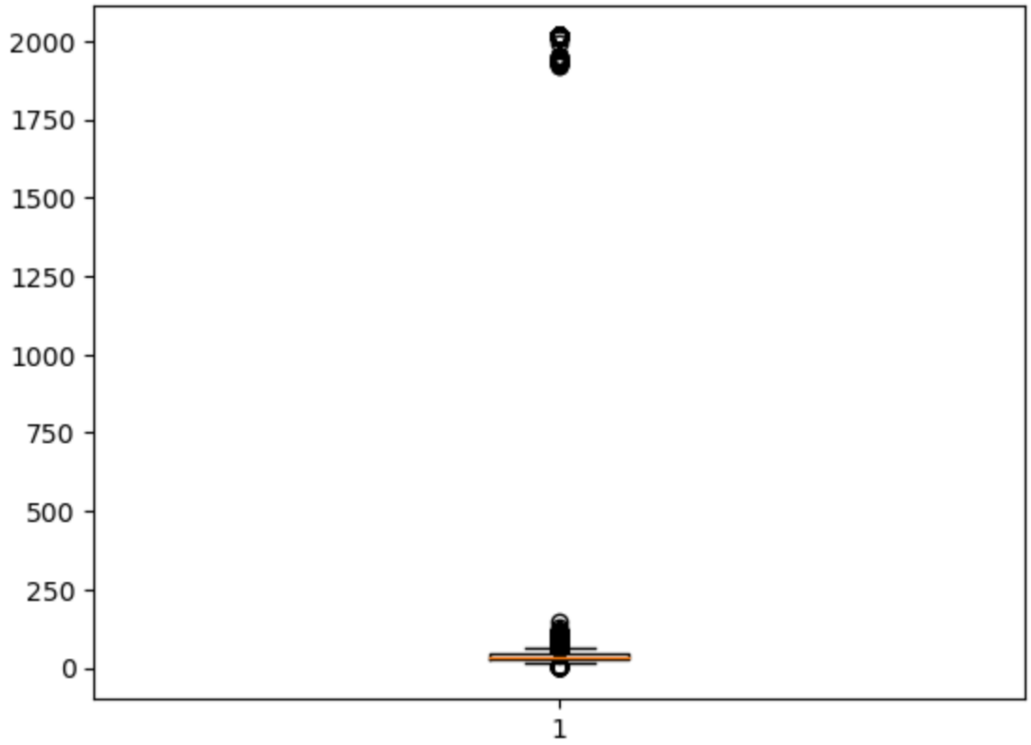
ix) Feature Encoding : Stnadardising the data to make the entire data consistent by making them have mean = 0 and variance = 1.

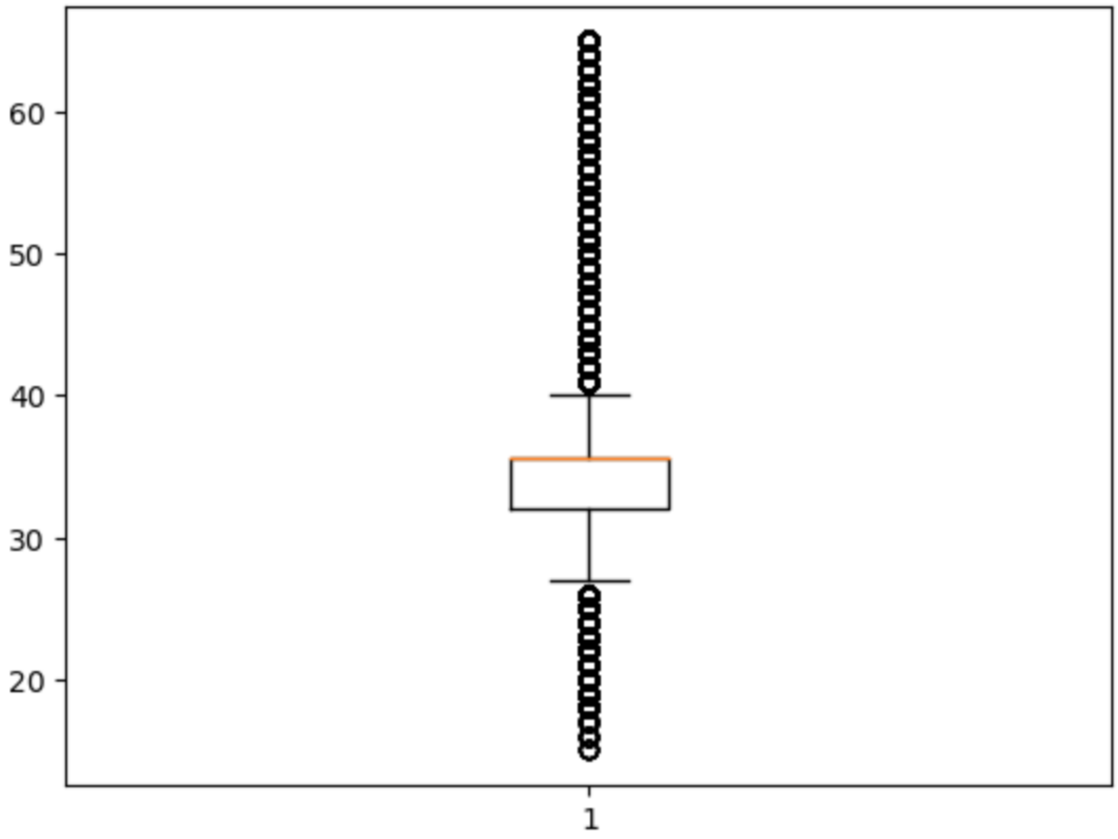
x) Null Values : In the dataset there were few null values, instead of dropping them and making the model biased we have replaced them with their column's mean value.

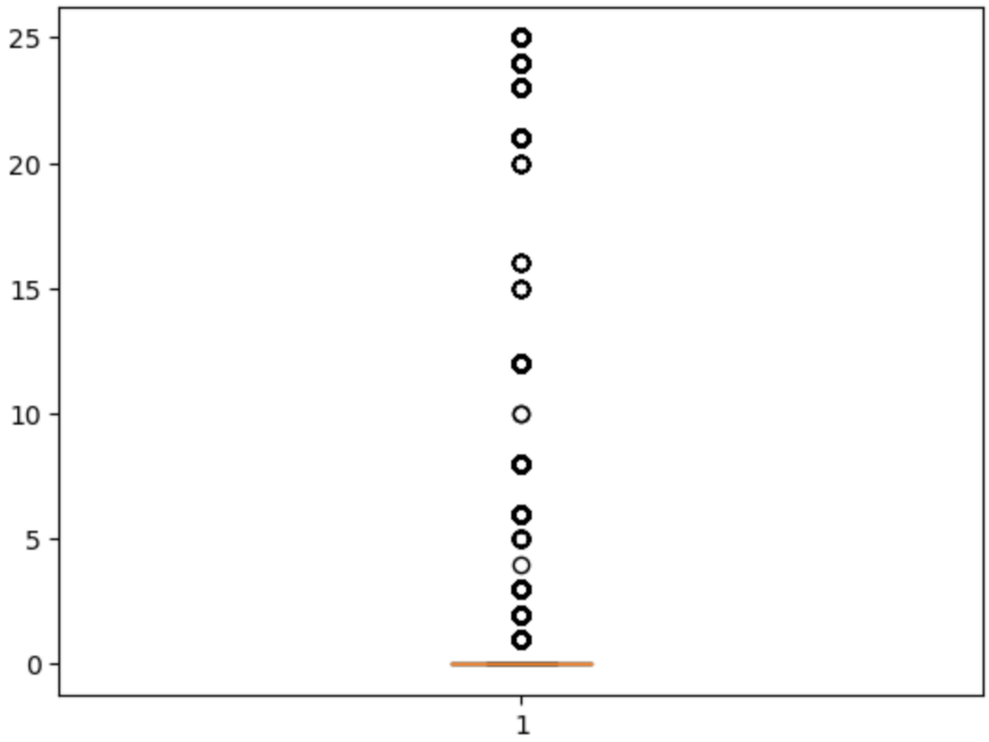
Exploratory Data Analysis (EDA) :

i) Box Plots:

Using box plots we have identified the outliers. The below plot shows the existence of outliers in the 'age' column.

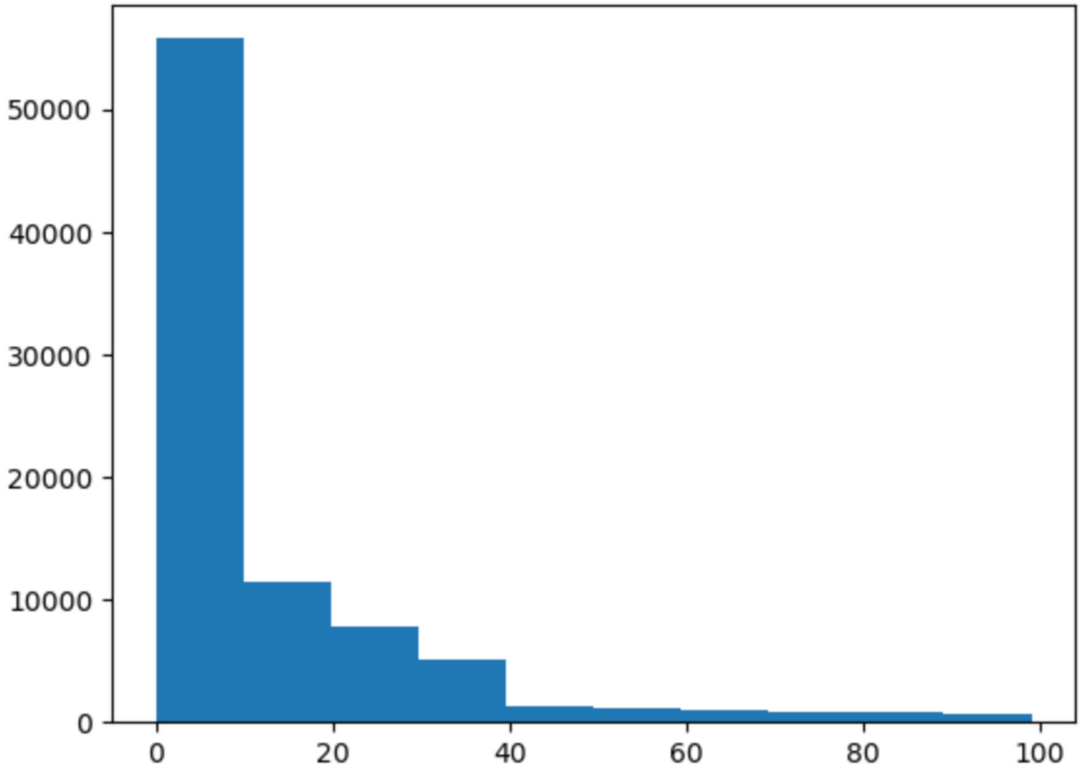


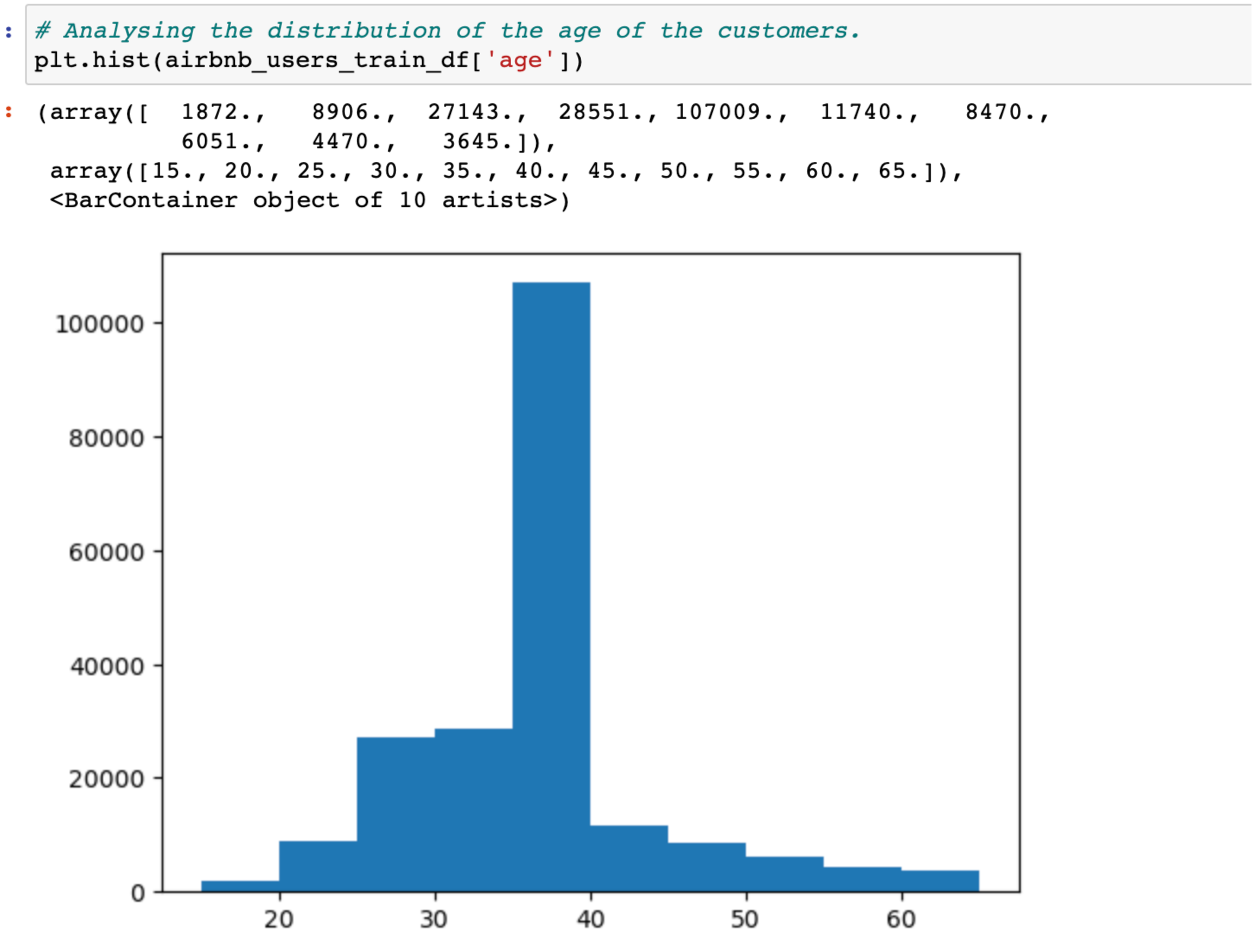

60
50
40
30
20
1


25
20
15
10
5
0
1

ii) Histograms :

Using histograms, we understood the distributions of the columns.

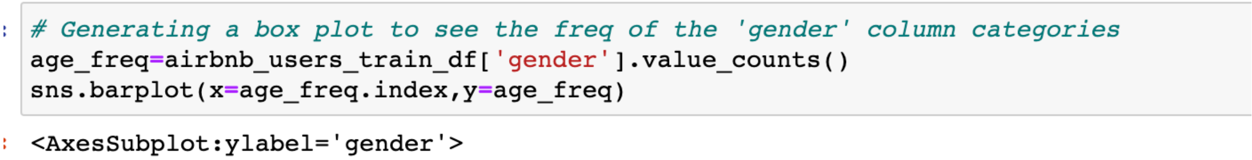




```
: # Analysing the distribution of the age of the customers.
  plt.hist(airbnb_users_train_df['age'])
```

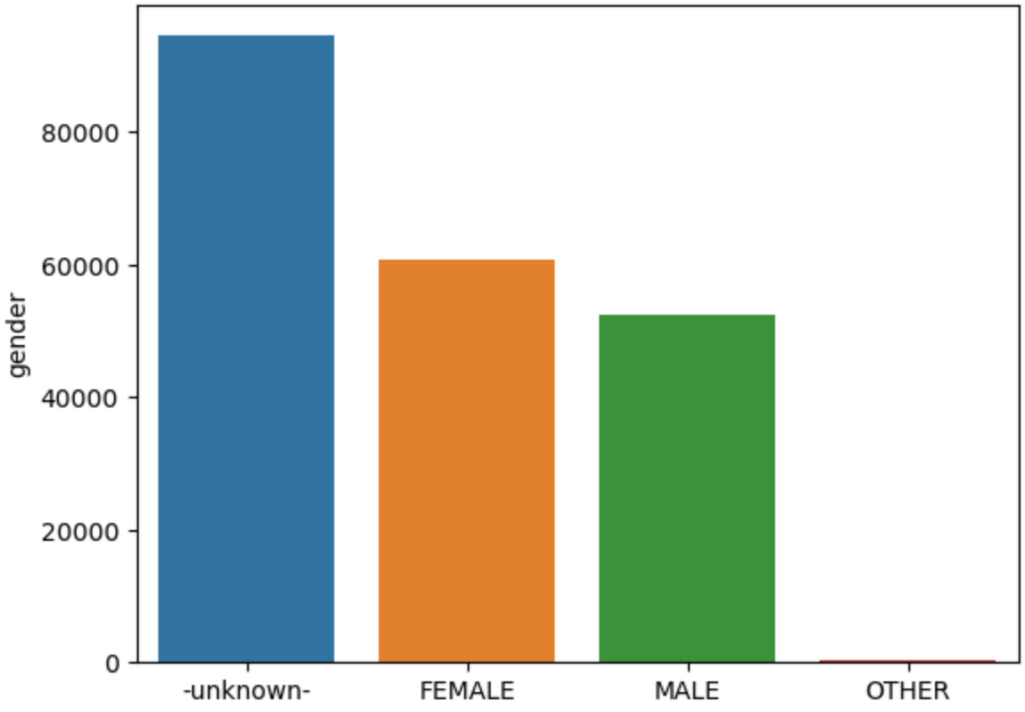
```
: (array([  1872.,   8906.,  27143.,  28551., 107009.,  11740.,   8470.,
            6051.,   4470.,   3645.]),
   array([15., 20., 25., 30., 35., 40., 45., 50., 55., 60., 65.]),
   <BarContainer object of 10 artists>)
```

iii) Barplot :
We have used barplot to find out the counts of the genders in the dataset.
Excluding ‘unknown’ there are more ‘female’ users compared to the ‘male’ users.



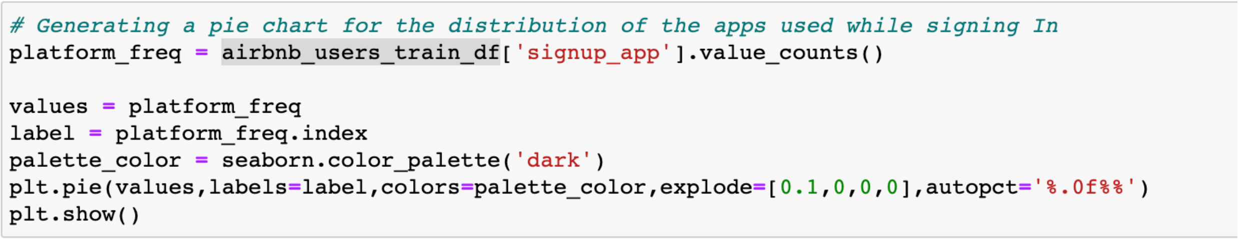
```
# Generating a box plot to see the freq of the 'gender' column categories
age_freq=airbnb_users_train_df['gender'].value_counts()
sns.barplot(x=age_freq.index,y=age_freq)
```

```
<AxesSubplot:ylabel='gender'>
```



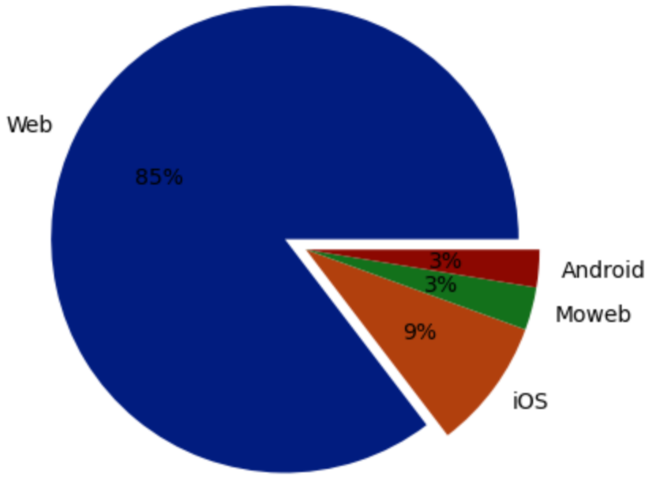
iv) Pie Chart :

We have used pie chart to understand the ratios of the ‘signup_app’ column. This shows that most of the signups have been done by ‘Web’, and the second highest is ‘iOS’.



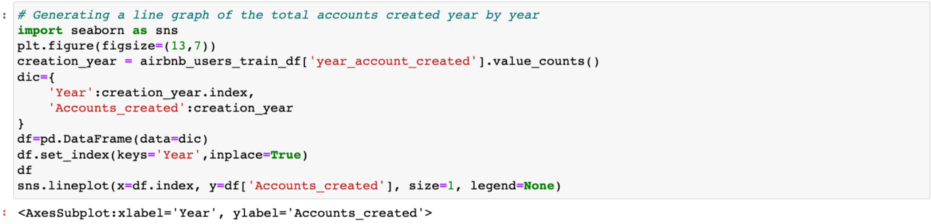
```
# Generating a pie chart for the distribution of the apps used while signing In
platform_freq = airbnb_users_train_df['signup_app'].value_counts()

values = platform_freq
label = platform_freq.index
palette_color = seaborn.color_palette('dark')
plt.pie(values,labels=label,colors=palette_color,explode=[0.1,0,0,0],autopct='%.0f%%')
plt.show()
```



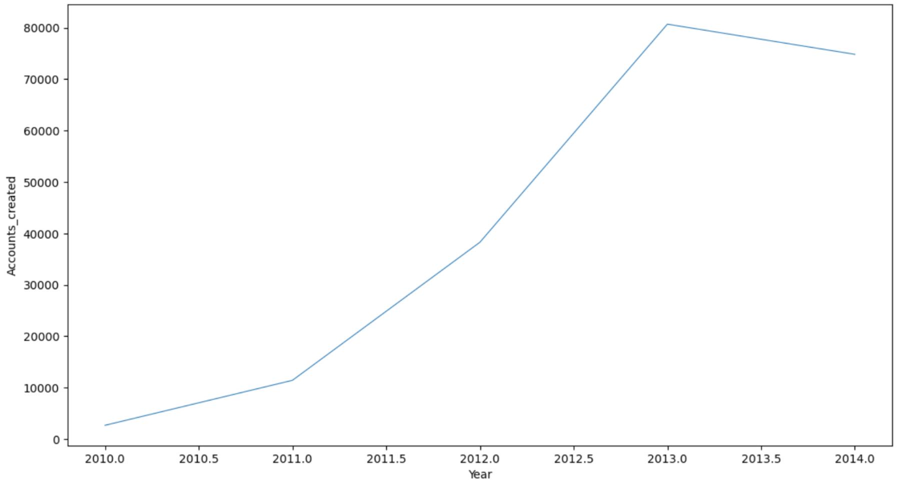
v) Line plot :

Line plot has been used to see the number of accounts created every year. This plot Shows that in year 2013 it reached the highest number of users.



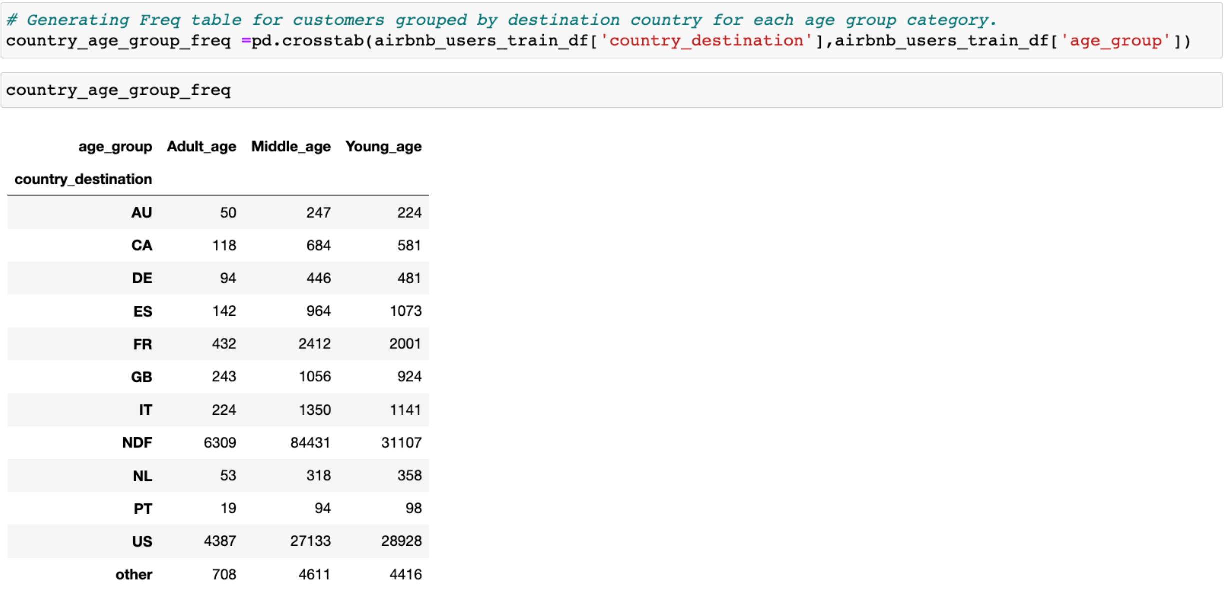
```
: # Generating a line graph of the total accounts created year by year
import seaborn as sns
plt.figure(figsize=(13,7))
creation_year = airbnb_users_train_df['year_account_created'].value_counts()
dic={
    'Year':creation_year.index,
    'Accounts_created':creation_year
}
df=pd.DataFrame(data=dic)
df.set_index(keys='Year',inplace=True)
df
sns.lineplot(x=df.index, y=df['Accounts_created'], size=1, legend=None)
```

```
: <AxesSubplot:xlabel='Year', ylabel='Accounts_created'>
```



vi) crosstab :

Crosstab has been used to get a clear understanding of the destinations grouped by the age categories.



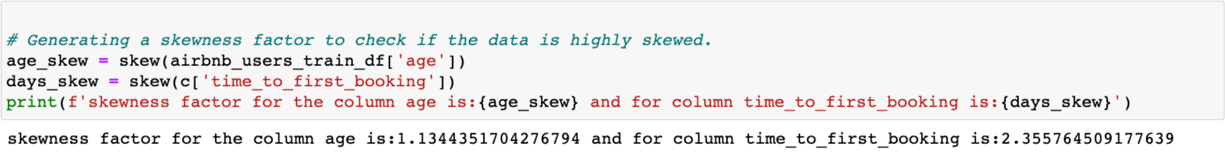
```
# Generating Freq table for customers grouped by destination country for each age group category.
country_age_group_freq =pd.crosstab(airbnb_users_train_df['country_destination'],airbnb_users_train_df['age_group'])
```

```
country_age_group_freq
```

age_group	Adult_age	Middle_age	Young_age
country_destination			
AU	50	247	224
CA	118	684	581
DE	94	446	481
ES	142	964	1073
FR	432	2412	2001
GB	243	1056	924
IT	224	1350	1141
NDF	6309	84431	31107
NL	53	318	358
PT	19	94	98
US	4387	27133	28928
other	708	4611	4416

vii) Skew :

We have used this function to understand the skewness of columns age and time_to_first_booking.

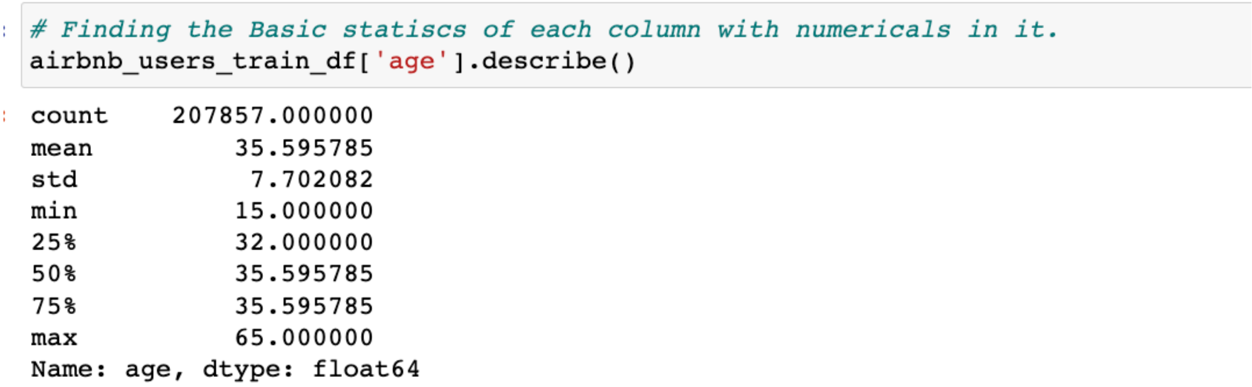


```
# Generating a skewness factor to check if the data is highly skewed.
age_skew = skew(airbnb_users_train_df['age'])
days_skew = skew(c['time_to_first_booking'])
print(f'skewness factor for the column age is:{age_skew} and for column time_to_first_booking is:{days_skew}')
```

```
skewness factor for the column age is:1.1344351704276794 and for column time_to_first_booking is:2.355764509177639
```

viii) describe :

This function has been used to understand the age column's statistics.

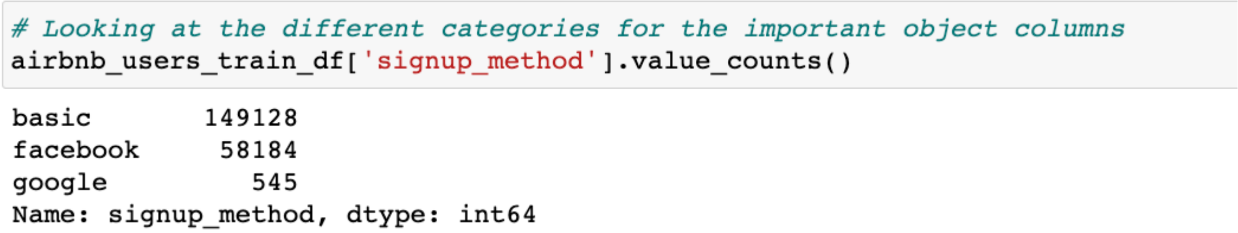


```
# Finding the Basic statiscs of each column with numericals in it.
airbnb_users_train_df['age'].describe()
```

```
count    207857.000000
mean         35.595785
std           7.702082
min          15.000000
25%          32.000000
50%          35.595785
75%          35.595785
max          65.000000
Name: age, dtype: float64
```

ix) value_counts :

This function has been used to find out the counts of the values in column 'signup_method'.

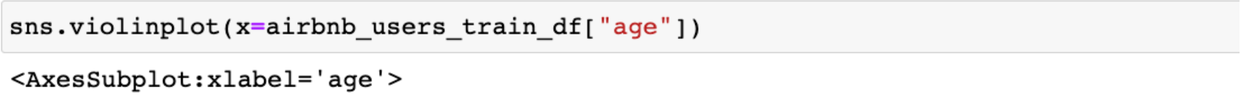


```
# Looking at the different categories for the important object columns
airbnb_users_train_df['signup_method'].value_counts()
```

```
basic       149128
facebook     58184
google         545
Name: signup_method, dtype: int64
```

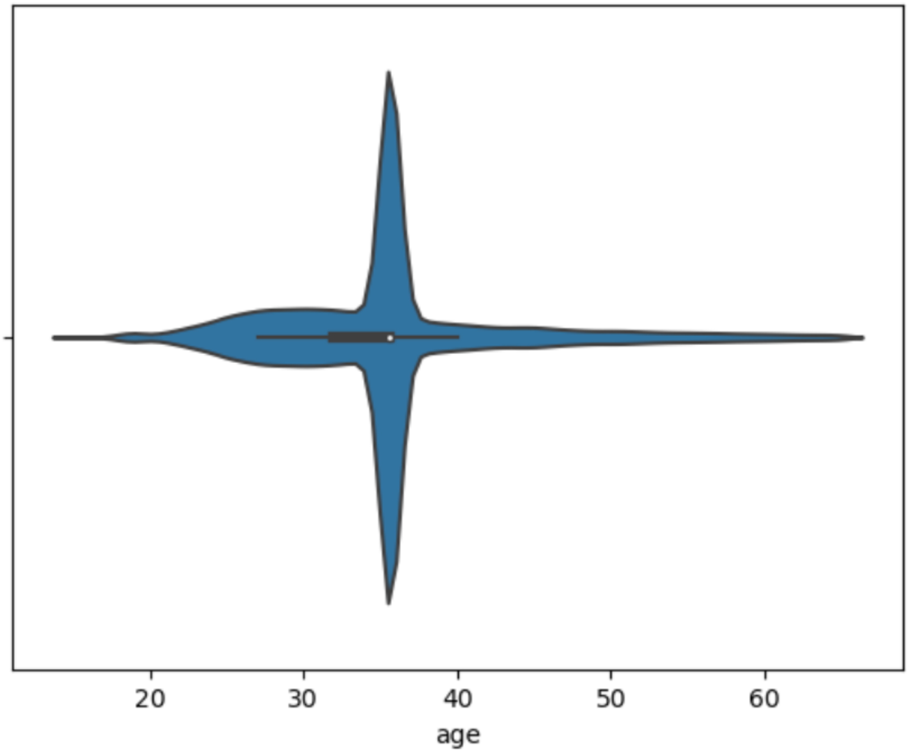
x) Violin Plot :

This plot has been used to understand the age distributions, this shows that the average age of the users is somewhere around 35.



```
sns.violinplot(x=airbnb_users_train_df["age"])
```

```
<AxesSubplot:xlabel='age'>
```


age