SIX WEEKS SUMMER TRAINING REPORT

On

(Data Visualization on Airbnb)

Submitted by

Anil Kushwaha

11503173

Dual Degree BCA-MCA

Under the Guidance of

Coursera

School of Computer Application Lovely Professional University, Phagwara

(02/JUN/2019 to 06/JUL/2019)

DECLARATION

I hereby declare that I have completed my six weeks summer training at Coursera (Online) from 02/JUN/2019 to 06/JUL/2019 under the guidance of Mr. Joseph Santarcangelo. I hereby undertake that the project undertaken by me is the genuine work of mine.

(Signature of student)

Anil Kushwaha
11503173

Date: _____

Acknowledgement

In preparation of my project, I had to take the help and guidance of some respected persons, who deserve my deepest gratitude. As the completion of this assignment gave me much pleasure, I would like to show my gratitude Mr. Joseph Santarcangelo, Course Instructor, on Cousera for giving me a good guidelines for assignment throughout numerous consultations. I would also like to expand my gratitude to all those who have directly and indirectly guided me in writing this assignment.

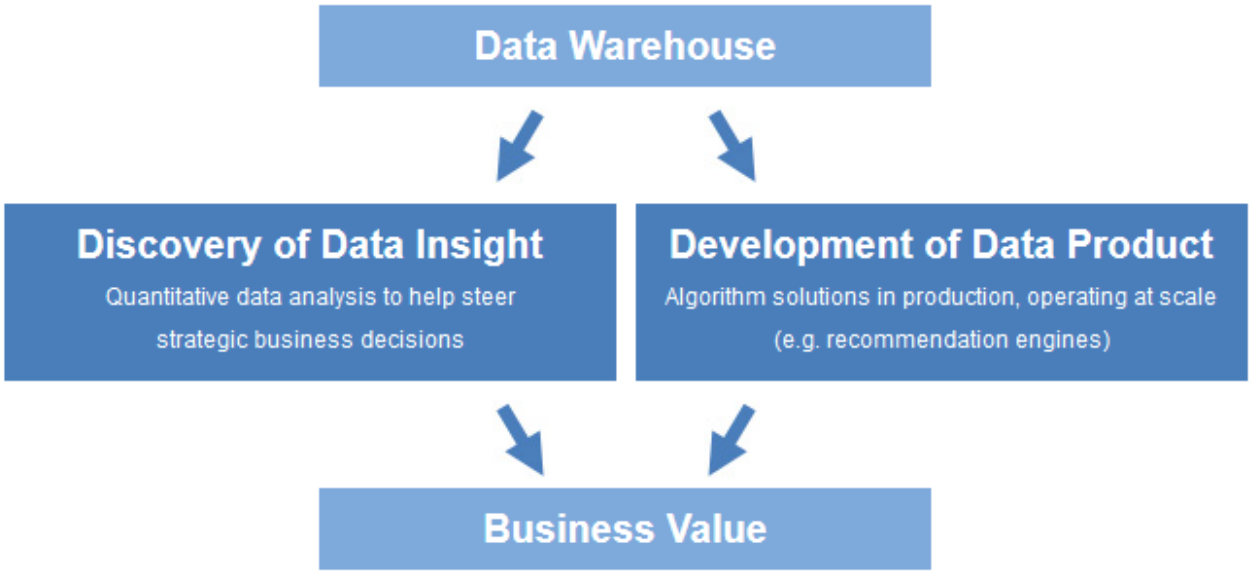
In addition, a thank you to Professor Mr. Sarabjit , who introduced me to the Methodology of work, and whose passion for the “underlying structures” had lasting effect? I also thank you to Professor Mr. Pranjal Jain who guided me about the project and helped wherever I faced difficulties.

Many people, especially my classmates have made valuable suggestions on my paper which gave me an inspiration to improve the quality of the assignment.

Table of contents

Introduction

Data science is a multidisciplinary blend of data inference, algorithm development, and technology in order to solve analytically complex problems. At the core is data. Troves of raw information, streaming in and stored in enterprise data warehouses. Much to learn by mining it. Advanced capabilities we can build with it. Data science is ultimately about using this data in creative ways to generate business value:



This aspect of data science is all about uncovering findings from data. Diving in at a granular level to mine and understand complex behaviours, trends, and inferences. It's about surfacing hidden insight that can help enable companies to make smarter business decisions. For example:

- Netflix data mines movie viewing patterns to understand what drives user interest, and uses that to make decisions on which Netflix original series to produce.

- Target identifies what are major customer segments within it's base and the unique shopping behaviours within those segments, which helps to guide messaging to different market audiences.
- Proctor & Gamble utilizes time series models to more clearly understand future demand, which help plan for production levels more optimally.

Profile of the Problem

Airbnb is an online marketplace and hospitality service, enabling people to lease or rent short-term lodging including vacation rentals , apartment rentals, homestays , hostels beds, or hotel rooms. New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. By accurately predicting where a new user will book their first travel experience, Airbnb can share more personalized content with their community, decrease the average time to first booking, and better forecast demand. We need to predict the first travel destination of a new user based on his personalized content.

Design

- Data pre-processing

  Data preprocessing is a **data** mining technique that involves transforming raw **data** into an understandable format. Real-world **data** is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. **Data preprocessing** is a proven method of resolving such issues.

  (a) Loading the csv files into pandas data frame

  (b) Dividing data into Train and Test data

  (c) Data cleaning

- Exploratory Data Analysis

  Exploratory data analysis (**EDA**) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

Coding

Importing Libraries ¶

[1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib as mpl
%matplotlib inline
from sklearn.model_selection import train_test_split
```

Loading CSV File and showing the table

[2]:
```
data_train_org = pd.read_csv("data/train_users_2.csv")
#print(data_train_org.columns)
#data_train_org=data_train_org.sort_values(by='timestamp_first_active')
data_train_org
```

[2]:

	id	date_account_created	timestamp_first_active	date_first_booking	gender	age	signup_method	signup_flow	language	affiliate_channel	affiliate
0	gxn3p5htnn	2010-06-28	20090319043255	NaN	-unknown-	NaN	facebook	0	en	direct	
1	820tgsjxq7	2011-05-25	20090523174809	NaN	MALE	38.0	facebook	0	en	seo	
2	4ft3gnwmtx	2010-09-28	20090609231247	2010-08-02	FEMALE	56.0	basic	3	en	direct	
3	bjjt8pjhuk	2011-12-05	20091031060129	2012-09-08	FEMALE	42.0	facebook	0	en	direct	

Showing no. of Rows and Columns

In [3]:
```
print(data_train_org.shape)
```

```
(213451, 16)
```

Spliting the Data in ratio

In [4]:
```
data_train, data_test = train_test_split(data_train_org, test_size=0.2)
data_train_copy = data_train
print("%d items in training data, %d in test data" % (len(data_train), len(data_test)))
```

```
170760 items in training data, 42691 in test data
```

Removing the data_first_booking column from data_train , data_test

In [5]:
```
print(data_train.columns)
data_train.drop('date_first_booking',1)
data_test.drop('date_first_booking',1)
data_train=data_train.sort_values(by='timestamp_first_active')
data_test=data_train.sort_values(by='timestamp_first_active')
```

```
Index(['id', 'date_account_created', 'timestamp_first_active',
       'date_first_booking', 'gender', 'age', 'signup_method', 'signup_flow',
       'language', 'affiliate_channel', 'affiliate_provider',
       'first_affiliate_tracked', 'signup_app', 'first_device_type',
       'first_browser', 'country_destination'],
      dtype='object')
```

Replacing gender and age values which are not present to Nan

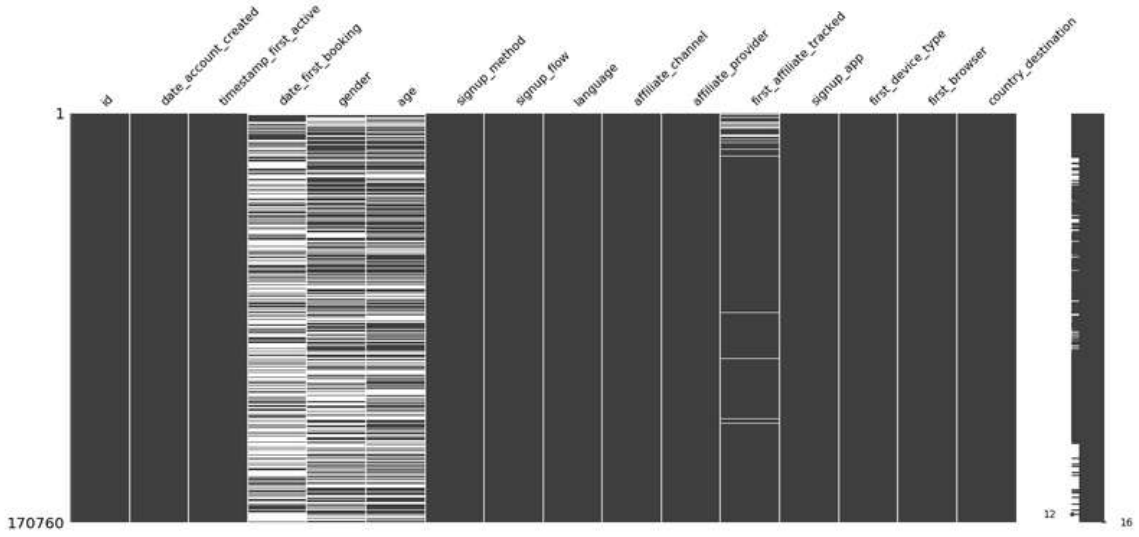
In [6]:
```
data_train.gender.replace('-unknown-',np.nan, inplace=True)
data_test.gender.replace('-unknown-',np.nan, inplace=True)
data_train.age.replace('NaN', np.nan, inplace=True)
data_test.age.replace('NaN',np.nan, inplace=True)
print(data_train.head())
```

```
           id date_account_created  timestamp_first_active date_first_booking  \
0  gxn3p5htnn           2010-06-28          20090319043255                NaN
1  820tgsjxq7           2011-05-25          20090523174809                NaN
3  bjjt8pjhuk           2011-12-05          20091031060129         2012-09-08
4  87mebub9p4           2010-09-14          20091208061105         2010-02-18
7  0d01nltbrs           2010-01-03          20100103191905         2010-01-13

   gender   age signup_method  signup_flow language affiliate_channel  \
0     NaN   NaN      facebook            0       en             direct
1    MALE  38.0      facebook            0       en                seo
3  FEMALE  42.0      facebook            0       en             direct
4     NaN  41.0         basic            0       en             direct
7  FEMALE  47.0         basic            0       en             direct

  affiliate_provider first_affiliate_tracked signup_app first_device_type  \
0             direct               untracked        Web       Mac Desktop
1             google               untracked        Web       Mac Desktop
3             direct               untracked        Web       Mac Desktop
4             direct               untracked        Web       Mac Desktop
7             direct                     omg        Web       Mac Desktop

  first_browser country_destination
0        Chrome                 NDF
1        Chrome                 NDF
3       Firefox               other
4        Chrome                  US
7        Safari                  US
```

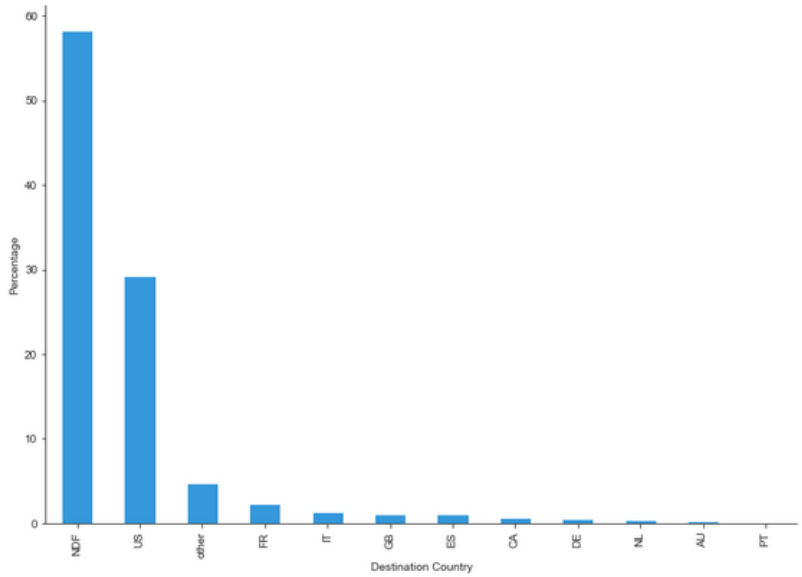
Showing the plot lot of missing values in gender , age

In [7]:
```
import missingno as msno
msno.matrix(data_train)
```

Out[7]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x26a4442fac8>
```



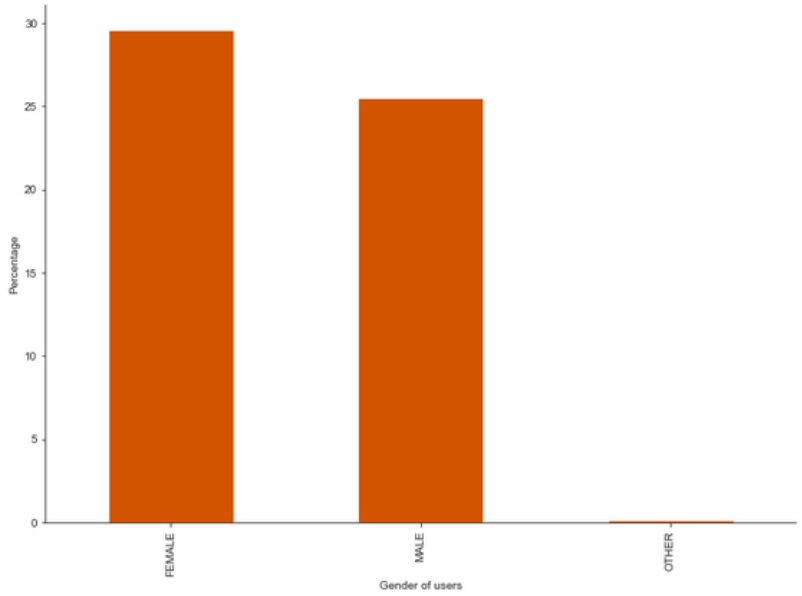
In [8]:
```
sns.set_style('ticks')
fig, ax = plt.subplots()
fig.set_size_inches(11.7, 8.27)
destination_percentage = data_train.country_destination.value_counts() / data_train.shape[0] * 100
destination_percentage.plot(kind='bar',color='#3498DB')
plt.xlabel('Destination Country')
plt.ylabel('Percentage')
sns.despine()
```



1 . 57% of users in Train data set did not travel anywhere .

2 . 28 % of users travelled in their home country i.e ..,U.S .

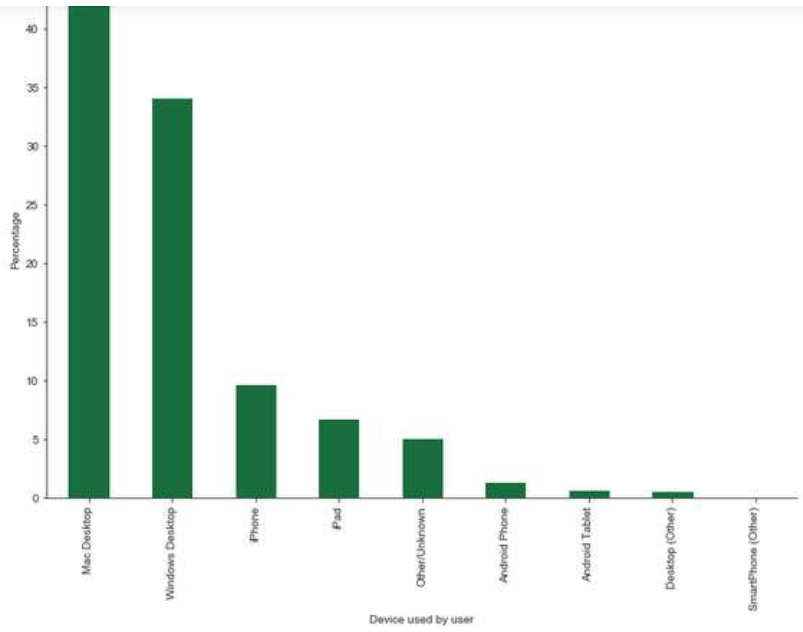
In [9]:
```
sns.set_style('ticks')
fig, ax = plt.subplots()
fig.set_size_inches(11.7, 8.27)
gender_percentage = data_train.gender.value_counts() / data_train.shape[0] * 100
gender_percentage.plot(kind='bar',color='#D35400')
plt.xlabel('Gender of users')
plt.ylabel('Percentage')
sns.despine()
```



1 . 45 % of user's gender information is not present .

2 . There is less difference between Female and Male users.

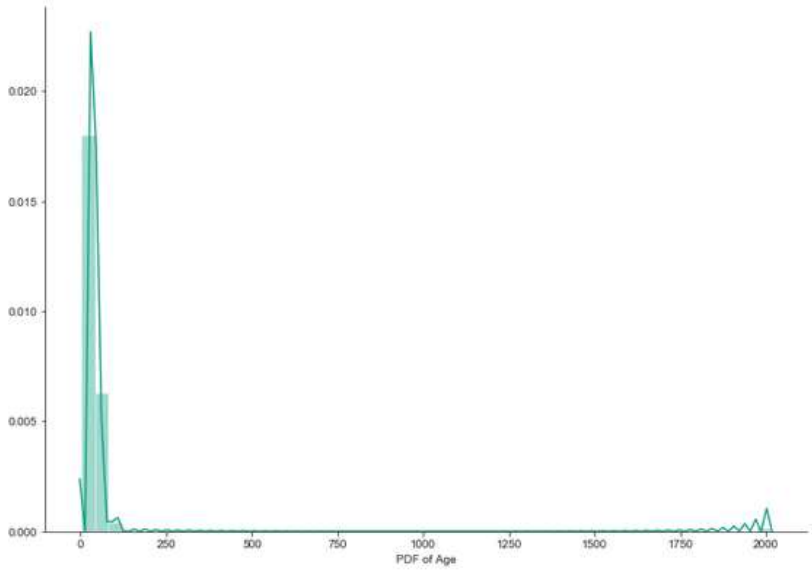
In [10]:
```
sns.set_style('ticks')
fig, ax = plt.subplots()
fig.set_size_inches(11.7, 8.27)
device_percentage = data_train.first_device_type.value_counts() / data_train.shape[0] * 100
device_percentage.plot(kind='bar',color='#196F3D')
plt.xlabel('Device used by user')
plt.ylabel('Percentage')
sns.despine()
```



1 . 58% users are using Apple products .

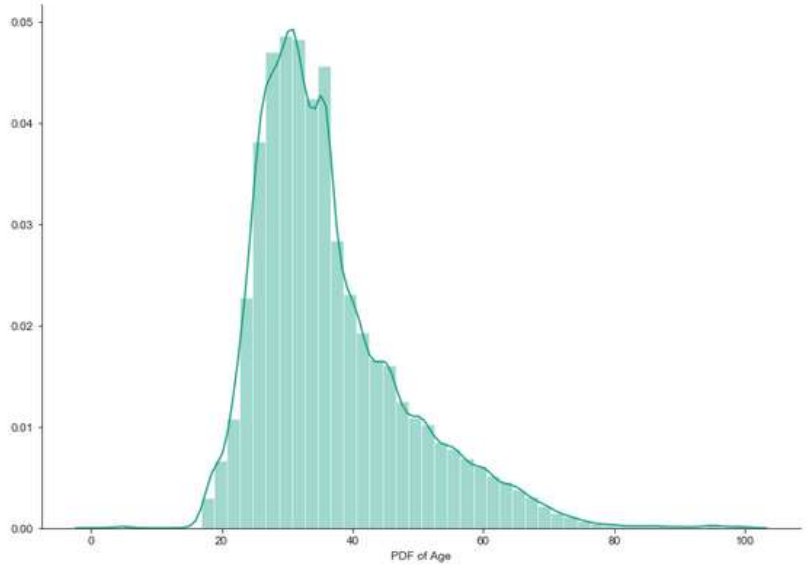
2 . Out of 71,719 users who travelled atleast once,31660 users are apple users [44.15%] which implies Mac users are booking more frequently .

In [11]:
```
sns.set_style('ticks')
fig, ax = plt.subplots()
fig.set_size_inches(11.7, 8.27)
sns.distplot(data_train.age.dropna(), color='#16A085')
plt.xlabel('PDF of Age')
sns.despine()
```



Some age values are incorrect, like close to 2000 , so cleaning such data [0.0035%]

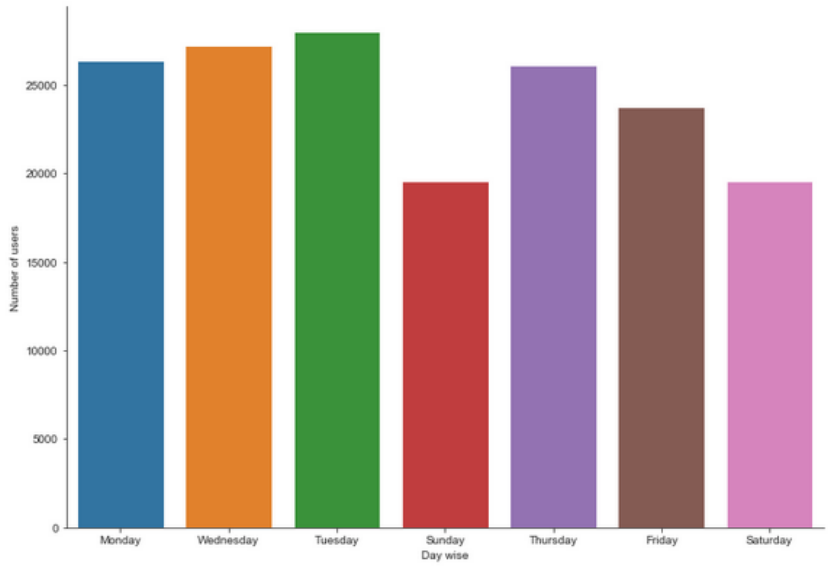
In [12]:
```
sns.set_style('ticks')
fig, ax = plt.subplots()
fig.set_size_inches(11.7, 8.27)
data_train['age']=data_train['age'].apply(lambda x : 36 if x>100 else x)
sns.distplot(data_train.age.dropna(), color='#16A085')
plt.xlabel('PDF of Age')
sns.despine()
```



1 . Majority of the users are between age 25 and 40 years . [72%]

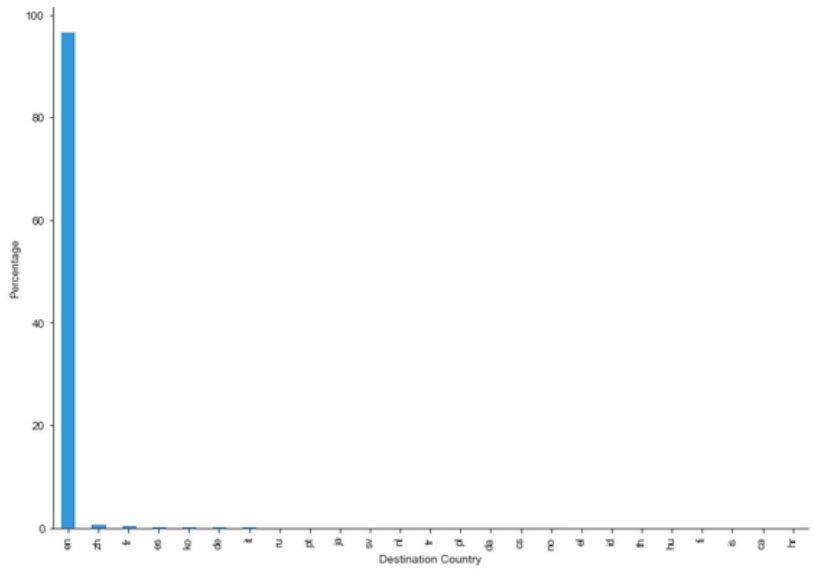
2 . There are some age values which are less than 18 years [0.006%](not allowed)

In [13]:
```
data_train['date_account_created_new'] = pd.to_datetime(data_train['date_account_created'])
data_train['date_first_active_new'] = pd.to_datetime((data_train.timestamp_first_active // 1000000), format='%Y%m%
data_train['date_account_created_day'] = data_train.date_account_created_new.dt.weekday_name
data_train['date_account_created_month'] = data_train.date_account_created_new.dt.month
data_train['date_account_created_year'] = data_train.date_account_created_new.dt.year
sns.set_style('ticks')
fig, ax = plt.subplots()
fig.set_size_inches(11.7, 8.27)
data_without_NDF = data_train[data_train['country_destination']!='US']
data_without_NDF1= data_without_NDF[data_without_NDF['country_destination']!='NDF']
sns.countplot(x='date_account_created_day',data=data_train)
plt.xlabel('Day wise')
plt.ylabel('Number of users')
sns.despine()
```



User activity is low on saturday and sunday . So chance of booking on saturdays , sundays is pretty low

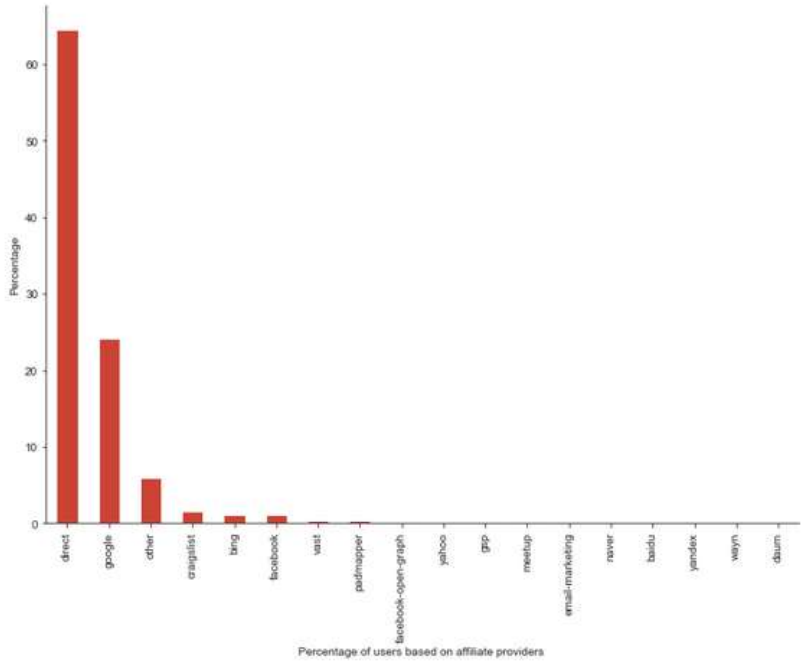
In [14]:
```
sns.set_style('ticks')
fig, ax = plt.subplots()
fig.set_size_inches(11.7, 8.27)
destination_percentage = data_train.language.value_counts() / data_train.shape[0] * 100
destination_percentage.plot(kind='bar',color='#3498DB')
plt.xlabel('Destination Country')
plt.ylabel('Percentage')
sns.despine()
```



1 .Majority of the user's language preference is English (96.67%) . But it is still qu-estionable because most of users are from US

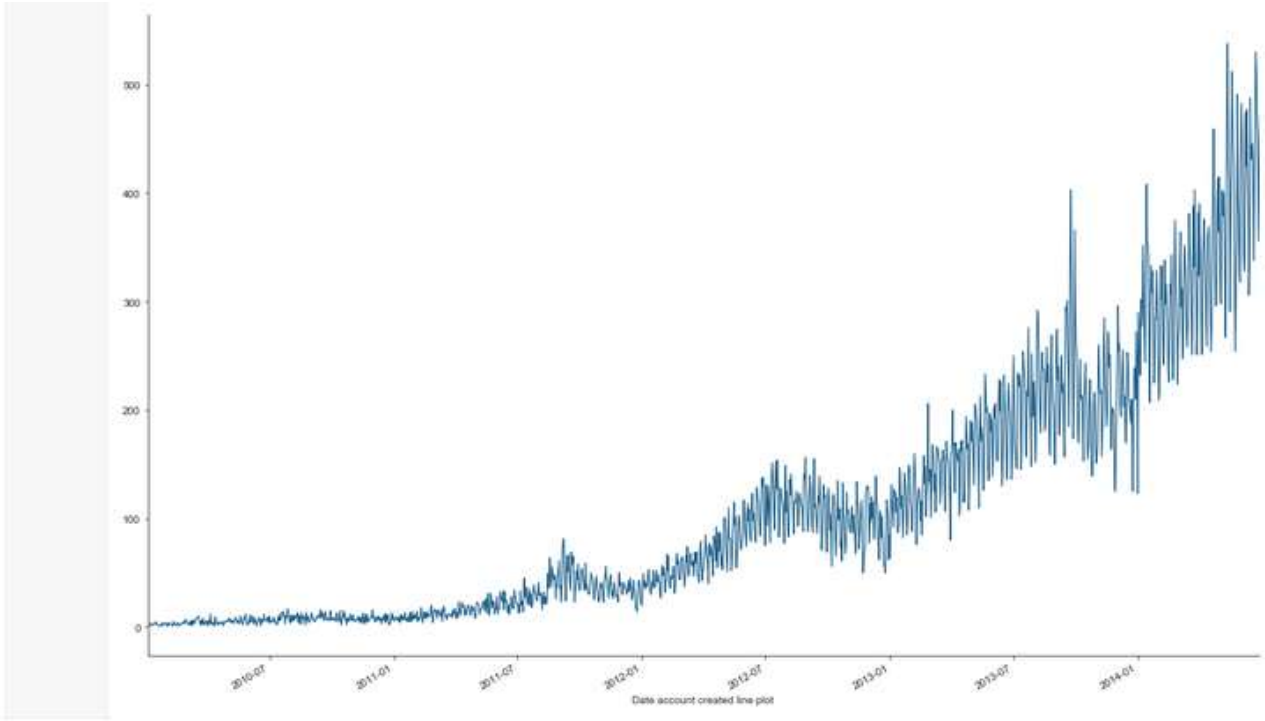
2 .Predicting geo location of users based on language preference may be useful

In [15]:
```
sns.set_style('ticks')
fig, ax = plt.subplots()
fig.set_size_inches(11.7, 8.27)
affiliate_provider_percentage = data_train.affiliate_provider.value_counts() / data_train.shape[0] * 100
affiliate_provider_percentage.plot(kind='bar',color='#CB4335')
plt.xlabel('Percentage of users based on affiliate providers ')
plt.ylabel('Percentage')
sns.despine()
```



In this plot we observe that most of users are coming from which source

In [16]:
```
sns.set_style('ticks')
fig, ax = plt.subplots()
fig.set_size_inches(18.7, 12.27)
data_train.date_account_created_new.value_counts().plot(kind='line', linewidth=1.2, color='#1F618D')
plt.xlabel('Date account created line plot ')
sns.despine()
```



1 . Every year beween September and October there is increase in Activity of users on Airbnb .

2 . Basic study on this lead to interesting phenomena that users are trying to book for Superbowl , Labor day.

Bibliography

- https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/
- https://www.kaggle.com/rajsankhe03/airbnb-analysis
- https://ai.google/tools/datasets/