Customer new table

December 6, 2017

1 Customer New Variables and new Table

1.0.1 load package

```
In [2]: library('dplyr',warn.conflicts = F)
        library('data.table',warn.conflicts = F)
        library('lubridate',warn.conflicts = F)
```

1.0.2 read data and view data structure

```
In [27]: setwd("/Users/sunmengnan/Documents/GitHub/thgfd/data")
         customer <- read.csv('MAIN_customer_data.csv', stringsAsFactors = F)
         country_code <- read.csv('country_code_lookup.csv', stringsAsFactors = F)
         str(customer)
         str(country_code)
```

```
'data.frame':        151888 obs. of  10 variables:
 $ Account_Key       : int  7605 5170 6412 39661 37432 36829 36503 36169 33260 26685 ...
 $ Registered_Date   : chr  "28/06/2010" "28/06/2010" "28/06/2010" "12/07/2010" ...
 $ Country           : chr  "United Kingdom" "Spain" "United Kingdom" "United Kingdom" ...
 $ PostCode          : chr  "KW1 5QQ" "8032" "DE21 7SA" "DL5 7QX" ...
 $ First_Order_Placed: chr  "29/06/2010" "28/06/2010" "28/06/2010" "12/07/2010" ...
 $ Site_Key          : chr  "121" "120" "121" "121" ...
 $ Locale            : chr  "en_GB" "es_ES" "en_GB" "en_GB" ...
 $ SCV_Key           : chr  "982885" "5187134" "2502778" "4841243" ...
 $ EDomain           : chr  "gmail.com" "gmail.com" "hotmail.co.uk" "hotmail.com" ...
 $ X                 : chr  "" "" "" "" ...
'data.frame':        248 obs. of  4 variables:
 $ Country_Code      : chr  "--" "AD" "AE" "AF" ...
 $ Country_Name      : chr  "UNKNOWN" "Andorra" "United Arab Emirates" "Afghanistan" ...
 $ Continental_Region: chr  "UNKNOWN" "Southern Europe" "Western Asia" "Southern Asia" ...
 $ Continent         : chr  "UNKNOWN" "Europe" "Asia" "Asia" ...
```

1.0.3 view number of missing values

```
In [6]: sum(is.na(customer))
        sum(is.na(country_code))
```

29
1

1.0.4 clean data, trun registered data and first order placed into date type, trun country code of GB into UK

```
In [28]: customer$Registered_Date <- as.Date(customer$Registered_Date, format = "%d/%m/%Y")
         customer$First_Order_Placed <- as.Date(customer$First_Order_Placed,format = "%d/%m/%Y
         country_code$Country_Code <- gsub("GB", "UK", country_code$Country_Code)
```

1.0.5 create a new data frame named c_code, which shows unique country code and name, view data head

```
In [29]: c_code <-data.frame(Country_Code=unique(country_code$Country_Code),
                             Country_Name=unique(country_code$Country_Name))
         head(c_code)
```

Country_Code	Country_Name
–	UNKNOWN
AD	Andorra
AE	United Arab Emirates
AF	Afghanistan
AG	Antigua and Barbuda
AI	Anguilla

1.0.6 turn the data type into character and change the column name of customer into Country_name

```
In [30]: c_code$Country_Code=as.character(c_code$Country_Code)
         c_code$Country_Name=as.character(c_code$Country_Name)
         colnames(customer)[colnames(customer)=="Country"]<- "Country_Name"
         head(customer)
```

Account_Key	Registered_Date	Country_Name	PostCode	First_Order_Placed	Site_Key	Locale
7605	2010-06-28	United Kingdom	KW1 5QQ	2010-06-29	121	en_GB
5170	2010-06-28	Spain	8032	2010-06-28	120	es_ES
6412	2010-06-28	United Kingdom	DE21 7SA	2010-06-28	121	en_GB
39661	2010-07-12	United Kingdom	DL5 7QX	2010-07-12	121	en_GB
37432	2010-07-12	United Kingdom	B21 8BE	2010-07-12	121	en_GB
36829	2010-07-11	United Kingdom	BH23 1DW	2010-07-11	121	en_GB

1.0.7 combine the customer table with country id

```
In [32]: nrow(customer)
         customer<-left_join(customer,c_code,by="Country_Name")
```

151888

1.0.8 Create a new column named joined year, which is the time interval between the registrated time and the first trading time, also we view the number of different time interval

```
In [57]: customer$Joined_years <- round((dmy("31/12/2016")-customer$Registered_Date)/365, 1)
         head(customer)
         library(stringr,warn.conflicts=F)
         expr="\\d+[\\.?]\\d?"
         customer$time_category<-str_extract(customer$Joined_years,expr)
         head(customer)
         table(customer$time_category)
```

Account_Key	Registered_Date	Country_Name	PostCode	First_Order_Placed	Site_Key	Locale
7605	2010-06-28	United Kingdom	KW1 5QQ	2010-06-29	121	en_GB
5170	2010-06-28	Spain	8032	2010-06-28	120	es_ES
6412	2010-06-28	United Kingdom	DE21 7SA	2010-06-28	121	en_GB
39661	2010-07-12	United Kingdom	DL5 7QX	2010-07-12	121	en_GB
37432	2010-07-12	United Kingdom	B21 8BE	2010-07-12	121	en_GB
36829	2010-07-11	United Kingdom	BH23 1DW	2010-07-11	121	en_GB
Account_Key	Registered_Date	Country_Name	PostCode	First_Order_Placed	Site_Key	Locale
7605	2010-06-28	United Kingdom	KW1 5QQ	2010-06-29	121	en_GB
5170	2010-06-28	Spain	8032	2010-06-28	120	es_ES
6412	2010-06-28	United Kingdom	DE21 7SA	2010-06-28	121	en_GB
39661	2010-07-12	United Kingdom	DL5 7QX	2010-07-12	121	en_GB
37432	2010-07-12	United Kingdom	B21 8BE	2010-07-12	121	en_GB
36829	2010-07-11	United Kingdom	BH23 1DW	2010-07-11	121	en_GB

```
  0.6   0.7   0.8   0.9   1.1   1.2   1.3   1.4   1.5   1.6   1.7   1.8   1.9
17773 29401 28575  3355  4729  2642  1874  1846  1615  1787  1833  1577  1402
  2.1   2.2   2.3   2.4   2.5   2.6   2.7   2.8   2.9   3.1   3.2   3.3   3.4
 2795  1524  1046  1028   895  1012  1024  1158  1030  1752  1028   919   839
  3.5   3.6   3.7   3.8   3.9   4.1   4.2   4.3   4.4   4.5   4.6   4.7   4.8
  847  1074  1080  1096  1276  1341   795   625   412   397   526   566   617
  4.9   5.1   5.2   5.3   5.4   5.5   5.6   5.7   5.8   5.9   6.1   6.2   6.3
  648  1249   649   618   537   567   429  1153   501   569   982   668   596
  6.4   6.5   6.6   6.7   6.8   6.9   7.1   7.2   7.3   7.4   7.5   7.6   7.7
  381   361   456   503   436   639  1268   474   559   522   430   670   853
  7.8
  876
```

1.0.9 upload the new dataframe to local drive

```
In [ ]: write.csv(customer, file = 'customer-new.csv', row.names = F)
```