Doina Covaliu - Gender Prediction using the first name

Install the necesarry packages (remove the comment sign before running the code)

```r
#install.packages("gender")
suppressPackageStartupMessages(library("gender"))
#install.packages("genderdata", repos = "https://dev.ropensci.org", type = "source")
suppressPackageStartupMessages(library("genderdata"))
#install.packages("dplyr")
suppressPackageStartupMessages(library("dplyr"))
#install.packages("stringr")
suppressPackageStartupMessages(library("stringr"))
#install.packages("tidyverse")
suppressPackageStartupMessages(library("tidyverse"))
#install.packages("magrittr")
suppressPackageStartupMessages(library("magrittr"))
#install.packages("DescTools")
suppressPackageStartupMessages(library("DescTools"))
```

Read the csv file into a dataframe

```r
data_source<-read.csv("C:/Users/Doina/Desktop/Customers_and_Orders.csv",
                      sep =",", stringsAsFactors = F)
```

```r
head(data_source)
```

```
##             First_Name Order.Amount Gender Order_Date
## 1             Veronica     $875.00     N/A 13/04/2000
## 2                Gemma     $218.75     N/A   1/5/2000
## 3         M. Christine     $437.50     N/A   1/5/2000
## 4         Julia Claire     $875.00     N/A   1/5/2000
## 5 Estate of Jean Howso   $8,750.00     N/A 26/08/2000
## 6               Pamela   $1,312.50     N/A 23/09/2000
```

```r
tail(data_source)
```

```
##                 First_Name Order.Amount Gender Order_Date
## 33516               Harry      $87.50     N/A   2/9/2020
## 33517           Catherine      $87.50     N/A   2/9/2020
## 33518              Lana C     $437.50     N/A   2/9/2020
## 33519               Alana     $131.25     N/A   2/9/2020
## 33520                Dhol     $175.00     N/A   2/9/2020
## 33521 Jacquelin and Shawn      $87.50     N/A   2/9/2020
```

```r
str(data_source)
```

```
## ’data.frame':    33521 obs. of  4 variables:
##  $ First_Name  : chr  "Veronica" "Gemma" "M. Christine" "Julia Claire" ...
##  $ Order.Amount: chr  "$875.00 " "$218.75 " "$437.50 " "$875.00 " ...
##  $ Gender      : chr  "N/A" "N/A" "N/A" "N/A" ...
##  $ Order_Date  : chr  "13/04/2000" "1/5/2000" "1/5/2000" "1/5/2000" ...
```

Data cleaning -Removing words like : INTERNATIONAL, REFUNDS, DELETE, comments that were included in some of the names.

```
data_source<-dplyr::mutate_if(data_source, is.character,stringr::str_replace_all,
                              pattern="INTERNATIONAL ", replacement="")
data_source<-dplyr::mutate_if(data_source, is.character,stringr::str_replace_all,
                              pattern="REFUNDS ",replacement="")
data_source<-dplyr::mutate_if(data_source, is.character,stringr::str_replace_all,
                              pattern="DELETE ", replacement="")
data_source<-dplyr::mutate_if(data_source, is.character,stringr::str_replace_all,
                              pattern=" $", replacement="")
data_source<-dplyr::mutate_if(data_source, is.character,stringr::str_replace_all,
                              pattern="Estate of ", replacement="")
```

Remove "-" from the name

```
data_source$First_Name<-str_replace_all(data_source$First_Name, "-","")
```

Remove space at the begining and the end of the name

```
data_source$First_Name <- trimws(data_source$First_Name, which = c("both"))
```

Extract the data where the first name is a initial only,in a separate dataset. In this case the gender cannot be determined .

```
incomplete_names<-subset(data_source, grepl('^[a-zA-Z]{1}$|^[a-zA-Z]{1}\\.\\b|
^[a-zA-Z]{1}\\s{1}[A-Z]{1}\\b|^[a-zA-Z]{1}\\.\\s{1}[A-Z]{1}\\b
|^[a-zA-Z]{1}\\.\\s{1}[A-Z]{1}\\.\\b|^[a-zA-Z]{1}\\.[A-Z]{1}\\.\\b',
data_source$First_Name)==TRUE)
```

Removing the observations where the First Name is formed of initials.

```
data_source<-subset(data_source,
grepl('^[a-zA-Z]{1}$|^[a-zA-Z]{1}\\b|^[a-zA-Z]{1}$|^[a-zA-Z]{1}\\.\\b|
^[a-zA-Z]{1}\\s{1}[A-Z]{1}\\b|^[a-zA-Z]{1}\\.\\s{1}[A-Z]{1}\\b|
^[a-zA-Z]{1}\\.\\s{1}[A-Z]{1}\\.\\b|
^[a-zA-Z]{1}\\.[A-Z]{1}\\.\\b', data_source$First_Name)==FALSE)
```

Removing any "." from the names

```
data_source$First_Name<-str_replace_all(data_source$First_Name, "\\.","")
```

For observation with the format: Name & Name, Name and Name the value "couple" will be assigned to the Gender column

```
data_source$Gender[data_source$First_Name %like any% c("% & %","% and %")]<-"couple"
couples<-subset(data_source, grepl("couple", data_source$Gender)==TRUE)
```

Temporary remove the couple from the dataset

```
data_source<-subset(data_source, grepl("couple", data_source$Gender)==FALSE)
```

Remove the initial after the First Name or Before the first name

```
data_source$First_Name<-str_replace_all(data_source$First_Name, "^[A-Z]{1}\\.\\s|^[A-Z]{1}\\s|\\s{1}[A-
```

If the First Names has 2 names remove the second name

```
data_source$First_Name<-str_replace_all(data_source$First_Name, "\\s{1}[A-Za-z]{2,}$","")
```

Extract the unique first names from the dataset

```
names_unique<-unique(data_source[,1])
```

Predict gender and create a data frame of names & predicted genders

```
predicted_names <- data.frame(gender(names_unique, method = "ssa"))
```

Assign the gender by joining the predicted_names dataset with the original dataset

```
final_dataset<-left_join(data_source, predicted_names[,c(1,4)], by = c("First_Name" = "name"))
sapply(final_dataset, function(x) sum(is.na(x)))
```

```
##   First_Name Order.Amount       Gender   Order_Date       gender
##            0            0            0            0          864
```

```
#sapply(final_dataset2, function(x) sum(is.na(x)))
```

Remove the original gender column that has only n/a values and save the final dataset in a new csv file

```
final_dataset<-final_dataset[-3]
final_dataset<-rename(final_dataset, Gender=gender)
final<-union(final_dataset, couples)
head(final)
```

```
##   First_Name Order.Amount Order_Date Gender
## 1   Veronica      $875.00 13/04/2000 female
## 2      Gemma      $218.75   1/5/2000 female
## 3      Julia      $875.00   1/5/2000 female
## 4       Jean    $8,750.00 26/08/2000 female
## 5     Pamela    $1,312.50 23/09/2000 female
## 6      Norma    $1,750.00 23/09/2000 female
```

```
tail(final)
```

```
##                       First_Name Order.Amount Order_Date Gender
## 31755        William and Carole    $4,375.00   5/9/2020 couple
## 31756            Sandra & David      $175.00 16/09/2020 couple
## 31757            Alen and Carol       $87.50 16/09/2020 couple
## 31758              David & Liat      $875.00   2/9/2020 couple
## 31759 Kathleen Kelly and Joseph      $175.00   2/9/2020 couple
## 31760       Jacquelin and Shawn       $87.50   2/9/2020 couple
```

```
#write.csv(final_dataset,"Final_dataset.csv")
#write.csv(final,"Final.csv")
```