Exploratory Data Analytics

Daniel Ma

30/05/2021

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'tibble' was built under R version 4.0.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
## Warning: package 'readr' was built under R version 4.0.2
```

```
## Warning: package 'purrr' was built under R version 4.0.2
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
## Warning: package 'stringr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.0.5
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.0.3
```

```
my_data <- read_excel("ANZ synthesised transaction dataset.xlsx")
```

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting numeric in C3052 / R3052C3: got 'THE DISCOUNT CHEMIST GROUP'
```

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting numeric in C4360 / R4360C3: got 'LAND WATER & PLANNING East Melbourne'
```

```
my_data$date <- as.Date(my_data$date, "yyyy-mm-dd") # will convert the date column into a time value
```

```
## Warning in as.POSIXlt.POSIXct(x, tz = tz): unknown timezone 'yyyy-mm-dd'
```

```
summary(my_data)
```

```
##     status           card_present_flag bpay_biller_code   account
##  Length:12043       Min.   :0.000     Min.   :0        Length:12043
##  Class :character   1st Qu.:1.000     1st Qu.:0        Class :character
##  Mode  :character   Median :1.000     Median :0        Mode  :character
##                     Mean   :0.803     Mean   :0
##                     3rd Qu.:1.000     3rd Qu.:0
##                     Max.   :1.000     Max.   :0
##                     NA's   :4326      NA's   :11160
##    currency           long_lat         txn_description    merchant_id
##  Length:12043       Length:12043       Length:12043       Length:12043
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  merchant_code    first_name           balance               date
##  Min.   :0       Length:12043       Min.   :     0.24   Min.   :2018-08-01
##  1st Qu.:0       Class :character   1st Qu.:  3158.58   1st Qu.:2018-08-24
##  Median :0       Mode  :character   Median :  6432.01   Median :2018-09-16
##  Mean   :0                          Mean   : 14704.20   Mean   :2018-09-15
##  3rd Qu.:0                          3rd Qu.: 12465.94   3rd Qu.:2018-10-09
##  Max.   :0                          Max.   :267128.52   Max.   :2018-10-31
##  NA's   :11160
##     gender               age        merchant_suburb    merchant_state
##  Length:12043       Min.   :18.00   Length:12043       Length:12043
##  Class :character   1st Qu.:22.00   Class :character   Class :character
##  Mode  :character   Median :28.00   Mode  :character   Mode  :character
##                     Mean   :30.58
##                     3rd Qu.:38.00
##                     Max.   :78.00
##
##   extraction            amount        transaction_id       country
##  Length:12043       Min.   :   0.10   Length:12043       Length:12043
##  Class :character   1st Qu.:  16.00   Class :character   Class :character
##  Mode  :character   Median :  29.00   Mode  :character   Mode  :character
##                     Mean   : 187.93
##                     3rd Qu.:  53.66
##                     Max.   :8835.98
##
##  customer_id        merchant_long_lat    movement
##  Length:12043       Length:12043       Length:12043
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
```

```
##
##
```

The warning that exists is because of the unexpected values in the biller_code column.

```
library(factoextra)
fact_cols <- names(dplyr::select_if(subset(my_data,select = -c(account,long_lat,merchant_id,first_name,
my_data[,fact_cols] <- lapply(my_data[,fact_cols], factor)
```

Findings

```
summary(my_data$amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.10   16.00   29.00  187.93   53.66 8835.98
```

From this summary, we can see that the average transaction amount is $187.93.

```
my_data %>% filter(is.na(card_present_flag)) %>% filter(status=="posted") %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  4326
```

```
my_data %>% filter(status=="posted") %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  4326
```

The two above lines of code indicate that missing values for the sveral columns are when the status is "posted". Data does not need to be cleansed as merchant details will be NA for posted transactions

```
unique(my_data$first_name)
```

```
##  [1] "Diana"       "Michael"     "Rhonda"      "Robert"      "Kristin"
##  [6] "Tonya"       "Fernando"    "Isaiah"      "Ricky"       "Jeffrey"
## [11] "Patrick"     "Karen"       "Ruth"        "Kimberly"    "Joseph"
## [16] "Tiffany"     "Emily"       "Christine"   "Ryan"        "Michelle"
## [21] "Richard"     "Jessica"     "Ronald"      "Kaitlyn"     "Lori"
## [26] "Virginia"    "Andrew"      "Susan"       "Luis"        "Gregory"
## [31] "Barry"       "Daniel"      "Renee"       "Amy"         "Christopher"
## [36] "Marissa"     "Eric"        "Natasha"     "Edward"      "Craig"
## [41] "Sandra"      "Debra"       "Michele"     "Antonio"     "Tyler"
## [46] "Lucas"       "Jonathan"    "Matthew"     "James"       "Mackenzie"
## [51] "Linda"       "Dustin"      "Heather"     "Derek"       "Scott"
```

```
## [56] "Charles"     "Tim"         "Melissa"     "Darren"      "Jacqueline"
## [61] "Cindy"       "Stephanie"   "Rachael"     "Mary"        "Maria"
## [66] "Timothy"     "Nathaniel"   "Elizabeth"   "Paul"        "Sarah"
## [71] "Alexander"   "Donald"      "Kenneth"     "Ashley"      "Catherine"
## [76] "Billy"       "Abigail"     "Brian"       "David"       "Robin"
```

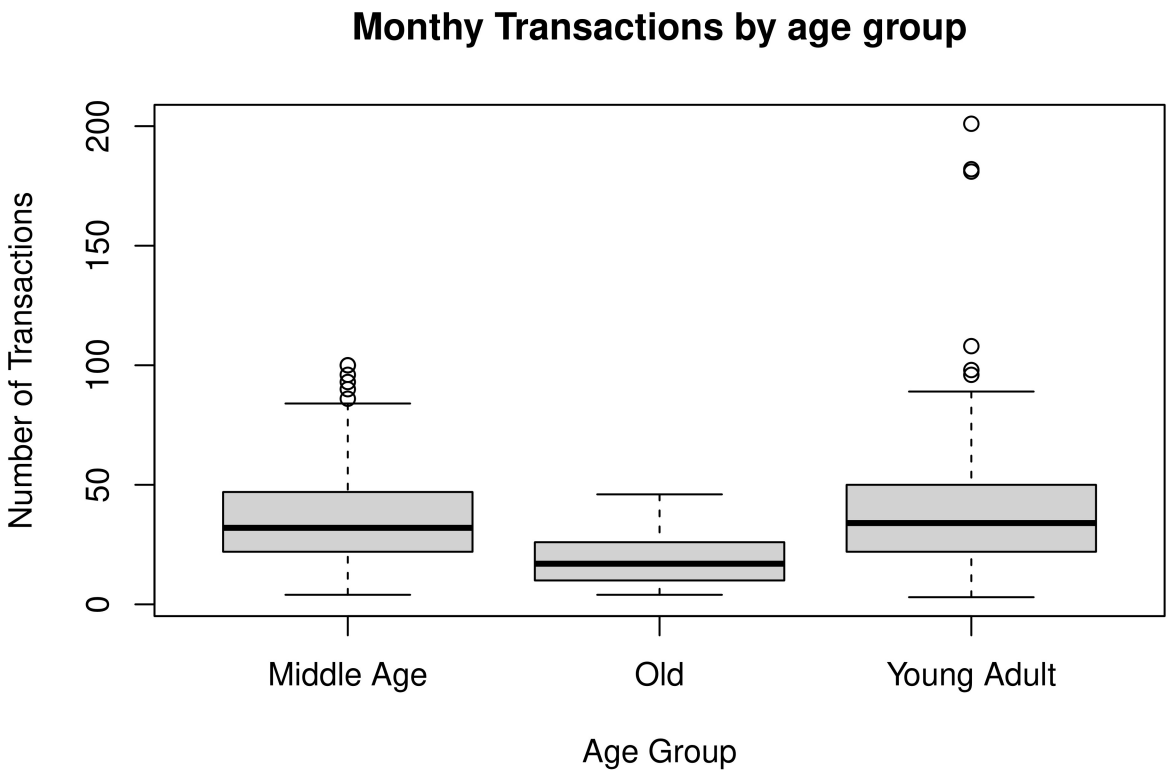
```
my_data %>% group_by(first_name)%>% summarise(unique(account)) %>% count() %>% filter(n>1)
```

```
## # A tibble: 14 x 2
## # Groups:   first_name [14]
##    first_name      n
##    <chr>       <int>
##  1 Christopher     2
##  2 Eric            2
##  3 Jeffrey         2
##  4 Jessica         2
##  5 Joseph          2
##  6 Kenneth         2
##  7 Kimberly        3
##  8 Linda           2
##  9 Michael         6
## 10 Richard         3
## 11 Robert          2
## 12 Ryan            2
## 13 Susan           2
## 14 Tyler           2
```

There are many unique first names but we notice that Michael is the most common first name, of which has 6 different account users.

```
# Separating into different age groups
my_data1 <- mutate(my_data, Age_group = ifelse(age <= 30 ,'Young Adult',ifelse(age >30 & age <50, 'Middl
# Looking at total number of transactions per month excluding credit
my_data_transaction <- my_data1 %>% filter(movement!='credit')%>% group_by(account,first_name,format(da

# Outlier Analysis
boxplot(my_data_transaction$n ~ my_data_transaction$Age_group,xlab = "Age Group",ylab = "Number of Trans
```



We see that young adults and middle aged people made more transactions. However, we need to take into account that these age groups also has large outliers, namely, one young adult who made over 200 transaction during the period of which this data was taken.

```
boxplot(my_data_transaction$n ~ my_data_transaction$gender,xlab = "Gender",ylab = "Number of Transactio
```

Once again, we notice that females made more transactions during this time period compared to males, but there is once again an outlier of a female who made over 200 transactions. From both boxplots (this one and the previous), we can deduce that a young female adult made over 200 transaction over the course of this period.

Regardless of age or gender, the existence of outliers have increase the mean of the number of transactions for each age group or gender.

```
#The plot was created with this code (but cannot be used in a pdf format)
#library(plotly)
#my_data_location <- my_data %>%
# separate(merchant_long_lat, c("merchant_long_lat_1", "merchant_long_lat_2"), sep = " ")
#my_data_location %>%
# plot_geo() %>%
# add_trace(x = ~merchant_long_lat_1,
#           y = ~merchant_long_lat_2)
```

While the plot below does not focus in on Australia (only because the image was saved as a pdf, not sure how to actually do that), zooming into Australia, we find that many of the merchants operate in Queensland, New South Wales and Victoria (the eastern side of Australia). Most merchants also operate on or close to the coast, with very few operating in inland Australia.

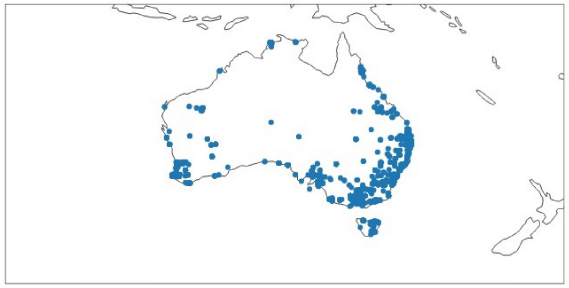


Figure 1: Some cool caption