Customer Transaction

Ayodeji Yekeen

12/31/2021

ANZ Customer Transaction Data Analysis

This analysis is based on a synthesised transaction dataset containing 3 months worth of transactions for 100 hypothetical customers. It contains purchases, recurring transactions, and salary transactions.

The dataset is designed to simulate realistic transaction behaviours that are observed in ANZ's real transaction data.

load required libraries

```
library(stringr)
library(lubridate)

## 
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
## 
##     date, intersect, setdiff, union

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.
3.1 ──

## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.6     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.4     ✓ forcats 0.5.1
## ✓ readr   2.1.1

## ── Conflicts ────────────────────────────────────────── tidyverse_conflict
s() ──
## x lubridate::as.difftime() masks base::as.difftime()
## x lubridate::date()        masks base::date()
## x dplyr::filter()          masks stats::filter()
## x lubridate::intersect()   masks base::intersect()
## x dplyr::lag()             masks stats::lag()
## x lubridate::setdiff()     masks base::setdiff()
## x lubridate::union()       masks base::union()

library(modelr)
library(sp)
library(leaflet)
library(geosphere)
library(knitr)
library(rpart)
```

Exploratory data analysis

Read the transaction dataset

```
df <- read.csv('ANZ_synthesised_transaction_dataset.csv')

#list of column names
colnames(df)

##  [1] "status"            "card_present_flag" "bpay_biller_code"
##  [4] "account"           "currency"          "long_lat"
##  [7] "txn_description"   "merchant_id"       "merchant_code"
## [10] "first_name"        "balance"           "date"
## [13] "gender"            "age"               "merchant_suburb"
## [16] "merchant_state"    "extraction"        "amount"
## [19] "transaction_id"    "country"           "customer_id"
## [22] "merchant_long_lat" "movement"

# how many rows are in the dataframe?
nrow(df)

## [1] 12043

# what is the size (column, row) of the dataframe?
dim(df)

## [1] 12043    23

# see list columns and data types
str(df)

## 'data.frame':    12043 obs. of  23 variables:
##  $ status           : chr  "authorized" "authorized" "authorized" "authori
zed" ...
##  $ card_present_flag: int  1 0 1 1 1 NA 1 1 1 NA ...
##  $ bpay_biller_code : chr  "" "" "" "" ...
##  $ account          : chr  "ACC-1598451071" "ACC-1598451071" "ACC-12223005
24" "ACC-1037050564" ...
##  $ currency         : chr  "AUD" "AUD" "AUD" "AUD" ...
##  $ long_lat         : chr  "153.41 -27.95" "153.41 -27.95" "151.23 -33.94"
"153.10 -27.66" ...
##  $ txn_description  : chr  "POS" "SALES-POS" "POS" "SALES-POS" ...
##  $ merchant_id      : chr  "81c48296-73be-44a7-befa-d053f48ce7cd" "830a451
c-316e-4a6a-bf25-e37caedca49e" "835c231d-8cdf-4e96-859d-e9d571760cf0" "485146
82-c78a-4a88-b0da-2d6302e64673" ...
##  $ merchant_code    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ first_name       : chr  "Diana" "Diana" "Michael" "Rhonda" ...
##  $ balance          : num  35.39 21.2 5.71 2117.22 17.95 ...
##  $ date              : chr  "01/08/2018" "01/08/2018" "01/08/2018" "01/08/2
018" ...
##  $ gender            : chr  "F" "F" "M" "F" ...
##  $ age               : int  26 26 38 40 26 20 43 43 27 40 ...
##  $ merchant_suburb   : chr  "Ashmore" "Sydney" "Sydney" "Buderim" ...
##  $ merchant_state    : chr  "QLD" "NSW" "NSW" "QLD" ...
##  $ extraction        : chr  "2018-08-01T01:01:15.000+0000" "2018-08-01T01:1
3:45.000+0000" "2018-08-01T01:26:15.000+0000" "2018-08-01T01:38:45.000+0000"
...
##  $ amount            : num  16.25 14.19 6.42 40.9 3.25 ...
##  $ transaction_id    : chr  "a623070bfead4541a6b0fff8a09e706c" "13270a2a902
145da9db4c951e04b51b9" "feb79e7ecd7048a5a36ec889d1a94270" "2698170da3704fd981
b15e64a006079e" ...
##  $ country           : chr  "Australia" "Australia" "Australia" "Australia"
...
##  $ customer_id       : chr  "CUS-2487424745" "CUS-2487424745" "CUS-21426011
69" "CUS-1614226872" ...
##  $ merchant_long_lat: chr  "153.38 -27.99" "151.21 -33.87" "151.21 -33.87"
"153.05 -26.68" ...
##  $ movement          : chr  "debit" "debit" "debit" "debit" ...
```

```
# view first 6 rows of the dataframe
head(df)
```

```
##       status card_present_flag bpay_biller_code        account currency
## 1 authorized                 1                  ACC-1598451071      AUD
## 2 authorized                 0                  ACC-1598451071      AUD
## 3 authorized                 1                  ACC-1222300524      AUD
## 4 authorized                 1                  ACC-1037050564      AUD
## 5 authorized                 1                  ACC-1598451071      AUD
## 6     posted                NA                  ACC-1608363396      AUD
##        long_lat txn_description                          merchant_id
## 1 153.41 -27.95             POS 81c48296-73be-44a7-befa-d053f48ce7cd
## 2 153.41 -27.95       SALES-POS 830a451c-316e-4a6a-bf25-e37caedca49e
## 3 151.23 -33.94             POS 835c231d-8cdf-4e96-859d-e9d571760cf0
## 4 153.10 -27.66       SALES-POS 48514682-c78a-4a88-b0da-2d6302e64673
## 5 153.41 -27.95       SALES-POS b4e02c10-0852-4273-b8fd-7b3395e32eb0
## 6 151.22 -33.87         PAYMENT
##   merchant_code first_name balance       date gender age merchant_suburb
## 1            NA      Diana   35.39 01/08/2018      F  26         Ashmore
## 2            NA      Diana   21.20 01/08/2018      F  26          Sydney
## 3            NA    Michael    5.71 01/08/2018      M  38          Sydney
## 4            NA     Rhonda 2117.22 01/08/2018      F  40         Buderim
## 5            NA      Diana   17.95 01/08/2018      F  26   Mermaid Beach
## 6            NA     Robert 1705.43 01/08/2018      M  20
##   merchant_state                   extraction amount
## 1            QLD 2018-08-01T01:01:15.000+0000  16.25
## 2            NSW 2018-08-01T01:13:45.000+0000  14.19
## 3            NSW 2018-08-01T01:26:15.000+0000   6.42
## 4            QLD 2018-08-01T01:38:45.000+0000  40.90
## 5            QLD 2018-08-01T01:51:15.000+0000   3.25
## 6                2018-08-01T02:00:00.000+0000 163.00
##                     transaction_id   country    customer_id merchant_long_
lat
## 1 a623070bfead4541a6b0fff8a09e706c Australia CUS-2487424745     153.38 -27
.99
## 2 13270a2a902145da9db4c951e04b51b9 Australia CUS-2487424745     151.21 -33
.87
## 3 feb79e7ecd7048a5a36ec889d1a94270 Australia CUS-2142601169     151.21 -33
.87
## 4 2698170da3704fd981b15e64a006079e Australia CUS-1614226872     153.05 -26
.68
## 5 329adf79878c4cf0aeb4188b4691c266 Australia CUS-2487424745     153.44 -28
.06
## 6 1005b48a6eda4ffd85e9b649dc9467d3 Australia CUS-2688605418
##   movement
## 1    debit
## 2    debit
## 3    debit
## 4    debit
## 5    debit
## 6    debit
```

Data Summary

```
# statistical summary of data
summary(df)
```

```
##     status          card_present_flag bpay_biller_code     account
##  Length:12043       Min.   :0.000     Length:12043       Length:12043
##  Class :character   1st Qu.:1.000     Class :character   Class :character
##  Mode  :character   Median :1.000     Mode  :character   Mode  :character
##                     Mean   :0.803
##                     3rd Qu.:1.000
##                     Max.   :1.000
##                     NA's   :4326
##    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   Length:12043
##  1st Qu.:0       Class :character   1st Qu.:  3158.58   Class :character
##  Median :0       Mode  :character   Median :  6432.01   Mode  :character
##  Mean   :0                          Mean   : 14704.20
##  3rd Qu.:0                          3rd Qu.: 12465.94
##  Max.   :0                          Max.   :267128.52
##  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
##
##
##
##
```

Date Cleaning & Manipulation

```
# change the format of date column
df$date<- as.Date(df$date,format = "%d/%m/%Y")

# derive weekday and hour data of each transaction
df$extraction = as.character(df$extraction)
df$hour = hour(as.POSIXct(substr(df$extraction,12,19),format="%H:%M:%S"))
df$weekday = weekdays(df$date)

# the dateset only contain records for 91 days, one day is missing
DateRange <- seq(min(df$date), max(df$date), by = 1)
DateRange[!DateRange %in% df$date]

## [1] "2018-08-16"
```

confirm the one -to -one link of account_id and customer_id

```
df %>% select(account,customer_id) %>%
  unique() %>%
  nrow()

## [1] 100
```

split customer & merchant lat_long into individual columns for analysis

```
dfloc = df[,c("long_lat","merchant_long_lat")]
dfloc<- dfloc %>% separate("long_lat", c("c_long", "c_lat"),sep=' ')
dfloc<- dfloc %>% separate("merchant_long_lat", c("m_long", "m_lat"),sep=' ')

## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 4326 rows [
6, 10,
## 11, 12, 14, 15, 17, 18, 19, 26, 27, 41, 42, 43, 44, 45, 46, 51, 52, 62, ..
.].

dfloc<- data.frame(sapply(dfloc, as.numeric))
df <- cbind(df,dfloc)
```

check the range of customer location

```
# filtering out transactions for those who don't reside in Australia
df_temp <- df %>%
  filter (!(c_long >113 & c_long <154 & c_lat > (-44) & c_lat < (-10)))
length(unique(df_temp$customer_id))

## [1] 1
```

check the distribution of missing values

```
apply(df, 2, function(x) sum(is.na(x)| x == ''))

##            status card_present_flag  bpay_biller_code           account
##                 0              4326             11158                 0
##          currency          long_lat   txn_description       merchant_id
##                 0                 0                 0              4326
##     merchant_code        first_name           balance              date
##             11160                 0                 0                 0
##            gender               age   merchant_suburb    merchant_state
##                 0                 0              4326              4326
##        extraction            amount    transaction_id           country
##                 0                 0                 0                 0
##       customer_id merchant_long_lat          movement              hour
##                 0              4326                 0                 0
##           weekday            c_long             c_lat            m_long
##                 0                 0                 0              4326
##             m_lat
##              4326

# check the number of unique values for each column
apply(df, 2, function(x) length(unique(x)))

##            status card_present_flag  bpay_biller_code           account
##                 2                 3                 4               100
##          currency          long_lat   txn_description       merchant_id
##                 1               100                 6              5726
##     merchant_code        first_name           balance              date
##                 2                80             12006                91
##            gender               age   merchant_suburb    merchant_state
##                 2                33              1610                 9
##        extraction            amount    transaction_id           country
##              9442              4457             12043                 1
##       customer_id merchant_long_lat          movement              hour
##               100              2704                 2                24
##           weekday            c_long             c_lat            m_long
##                 7                87                85               719
##             m_lat
##               670
```

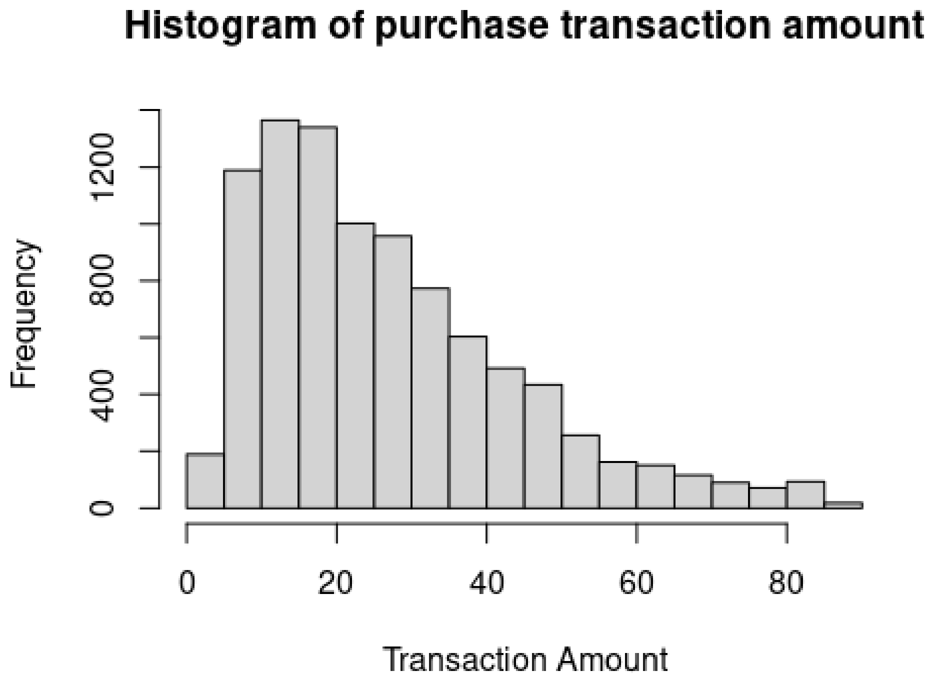
filtering out purchase transactions only

```
# assuming purchase transactions must be associated with a merchant (have a m
erchant Id)
df_temp <- df %>% filter(merchant_id != '' )
# it turned out that is equivilent to excluding following categories of trans
actions
df_csmp <- df %>%filter(!(txn_description %in% c('PAY/SALARY',"INTER BANK", "
PHONE BANK","PAYMEN
T")))
summary(df_csmp)
```

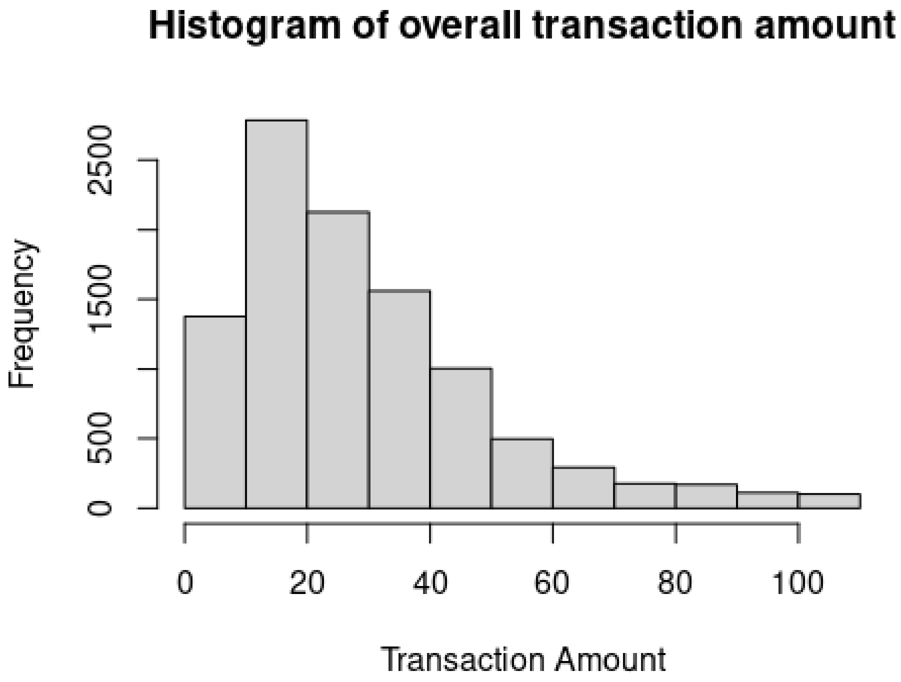
```
##     status          card_present_flag bpay_biller_code     account
##  Length:10317       Min.   :0.0000    Length:10317       Length:10317
##  Class :character   1st Qu.:1.0000    Class :character   Class :character
##  Mode  :character   Median :1.0000    Mode  :character   Mode  :character
##                     Mean   :0.8026
##                     3rd Qu.:1.0000
##                     Max.   :1.0000
##                     NA's   :2600
##    currency           long_lat          txn_description    merchant_id
##  Length:10317       Length:10317       Length:10317       Length:10317
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  merchant_code    first_name          balance               date
##  Min.   : NA     Length:10317       Min.   :     0.24   Min.   :2018-08-01
##  1st Qu.: NA     Class :character   1st Qu.:  3035.41   1st Qu.:2018-08-25
##  Median : NA     Mode  :character   Median :  6026.23   Median :2018-09-16
##  Mean   :NaN                        Mean   : 13691.17   Mean   :2018-09-15
##  3rd Qu.: NA                        3rd Qu.: 11757.93   3rd Qu.:2018-10-09
##  Max.   : NA                        Max.   :267093.66   Max.   :2018-10-31
##  NA's   :10317
##     gender               age          merchant_suburb    merchant_state
##  Length:10317       Min.   :18.00   Length:10317       Length:10317
##  Class :character   1st Qu.:23.00   Class :character   Class :character
##  Mode  :character   Median :28.00   Mode  :character   Mode  :character
##                     Mean   :30.36
##                      3rd Qu.:38.00  
##                      Max.   :78.00  
##                                     
##   extraction            amount        transaction_id       country         
##  Length:10317       Min.   :   0.10   Length:10317       Length:10317      
##  Class  :character  1st Qu.:  14.46   Class  :character  Class  :character  
##  Mode   :character  Median :  25.55   Mode   :character  Mode   :character  
##                     Mean   :  49.59                                         
##                     3rd Qu.:  43.16                                         
##                     Max.   :7081.09                                         
##                                                                             
##  customer_id        merchant_long_lat    movement              hour      
##  Length:10317       Length:10317       Length:10317       Min.   : 0.00  
##  Class  :character  Class  :character  Class  :character  1st Qu.: 9.00  
##  Mode   :character  Mode   :character  Mode   :character  Median :14.00  
##                                                           Mean   :13.34  
##                                                           3rd Qu.:19.00  
##                                                           Max.   :23.00  
##                                                                          
##    weekday              c_long          c_lat              m_long     
##  Length:10317       Min.   :114.6   Min.   :-573.00   Min.   :113.8  
##  Class  :character  1st Qu.:138.7   1st Qu.: -37.66   1st Qu.:144.7  
##  Mode   :character  Median :145.4   Median : -33.87   Median :145.8  
##                     Mean   :143.7   Mean   : -38.54   Mean   :143.4  
##                     3rd Qu.:151.2   3rd Qu.: -28.80   3rd Qu.:151.2  
##                     Max.   :255.0   Max.   : -12.37   Max.   :153.6  
##                                                       NA's   :2600   
##      m_lat       
##  Min.   :-43.31  
##  1st Qu.:-37.71  
##  Median :-33.84  
##  Mean   :-32.75  
##  3rd Qu.:-29.44  
##  Max.   :-12.33  
##  NA's   :2600
```

visualise the distribution of transaction amount

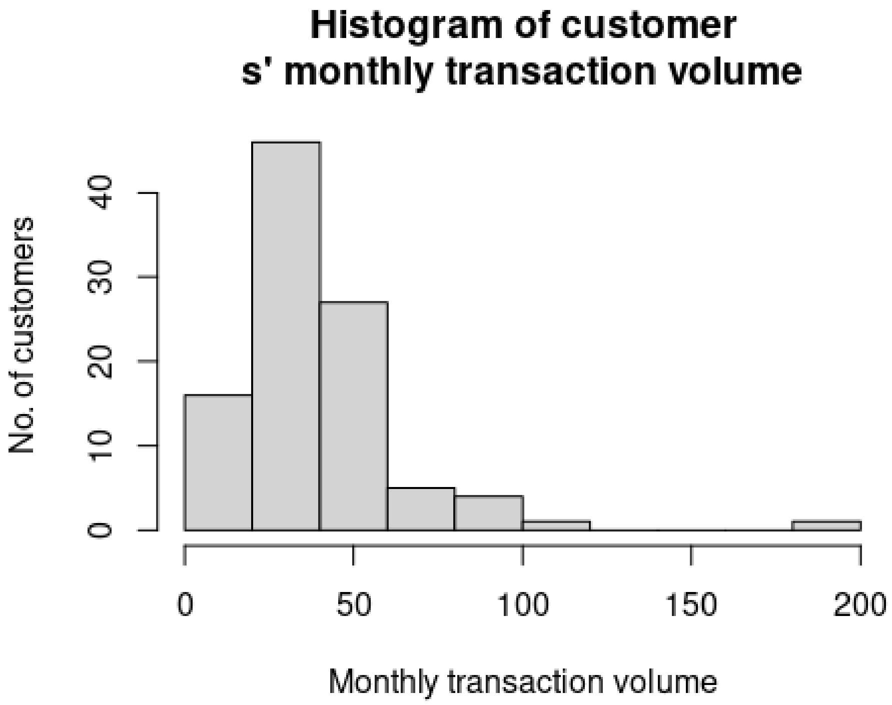
```
hist(df_csmp$amount[!df_csmp$amount %in% boxplot.stats(df_csmp$amount)$out], 
#exclude outliers
     xlab= 'Transaction Amount', main = 'Histogram of purchase transaction am
ount')
```



```
hist(df$amount[!df$amount %in% boxplot.stats(df$amount)$out], #exclude outlie
rs
     xlab= 'Transaction Amount',main = 'Histogram of overall transaction amou
nt')
```



```
df2 <- df %>%
  group_by(customer_id) %>%
  summarise(mon_avg_vol = round(n()/3,0))
hist(df2$mon_avg_vol,
     xlab= 'Monthly transaction volume', ylab='No. of customers', main = "His
togram of customer
s' monthly transaction volume")
```

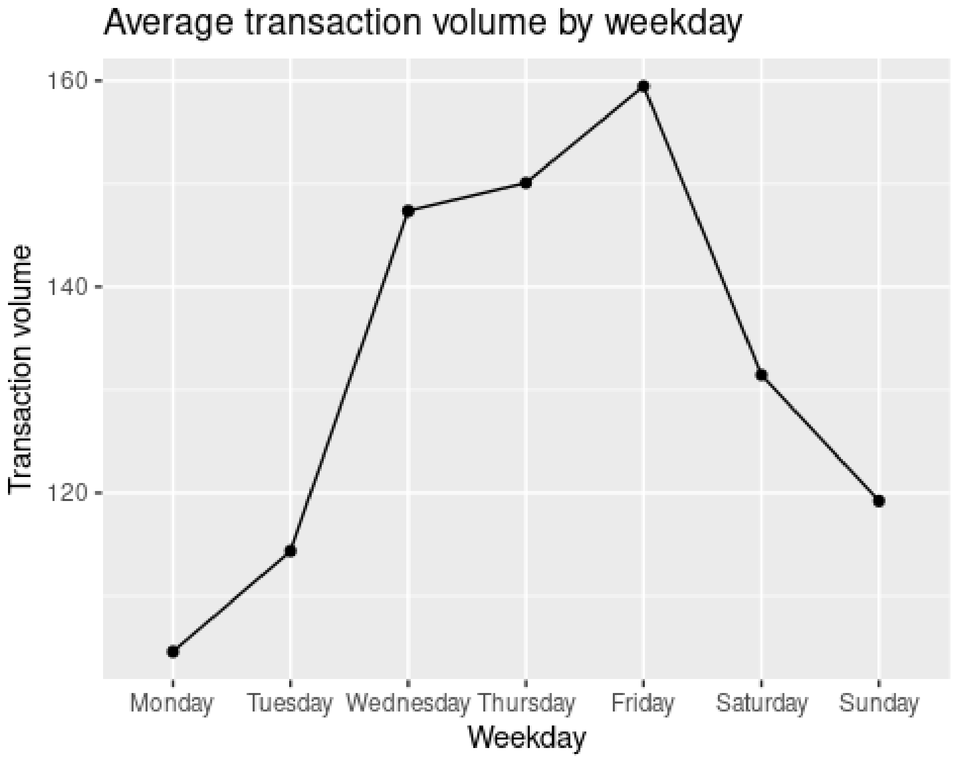


### Visualise transaction volume over an average week.

```
df3 <- df %>%
  select(date,weekday) %>%
  group_by(date,weekday) %>%
  summarise(daily_avg_vol = n()) %>%
  group_by(weekday) %>%
  summarise(avg_vol=mean(daily_avg_vol,na.rm=TRUE ))
```

```
## `summarise()` has grouped output by 'date'. You can override using the `.g
roups` argument.
```

```
df3$weekday <- factor(df3$weekday, levels=c( "Monday","Tuesday","Wednesday",
                                             "Thursday","Friday","Saturday","
Sunday"))
ggplot(df3,aes(x=weekday, y=avg_vol)) +geom_point()+geom_line(aes(group = 1))
+
  ggtitle('Average transaction volume by weekday') +
  labs(x='Weekday',y='Transaction volume')
```

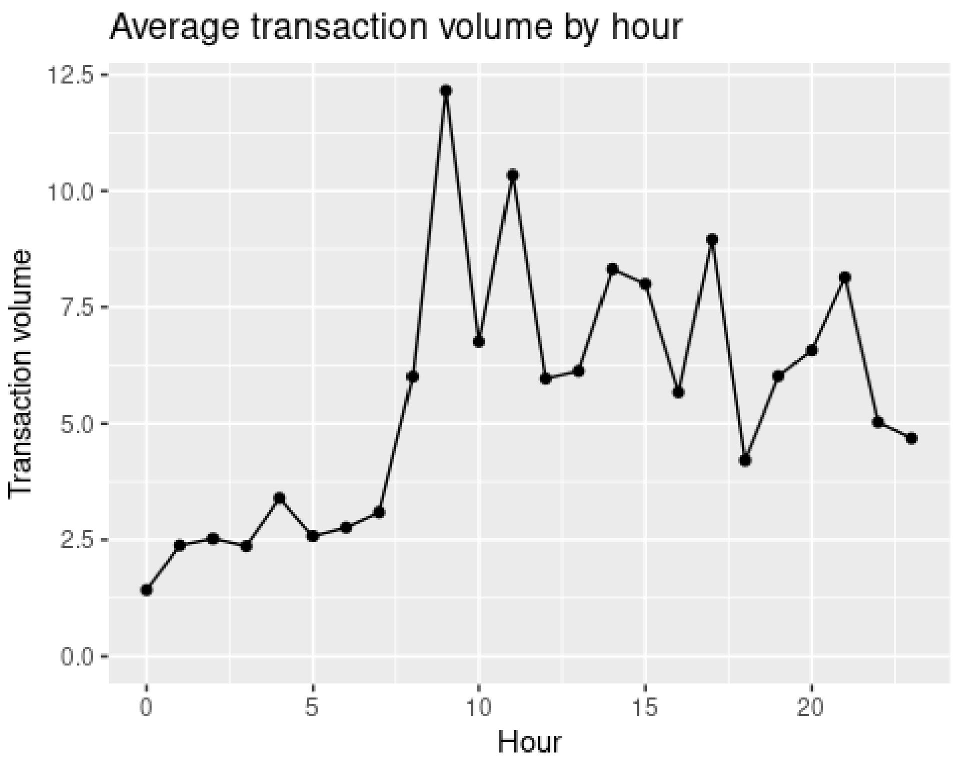


### visualize transaction volume over an average week.

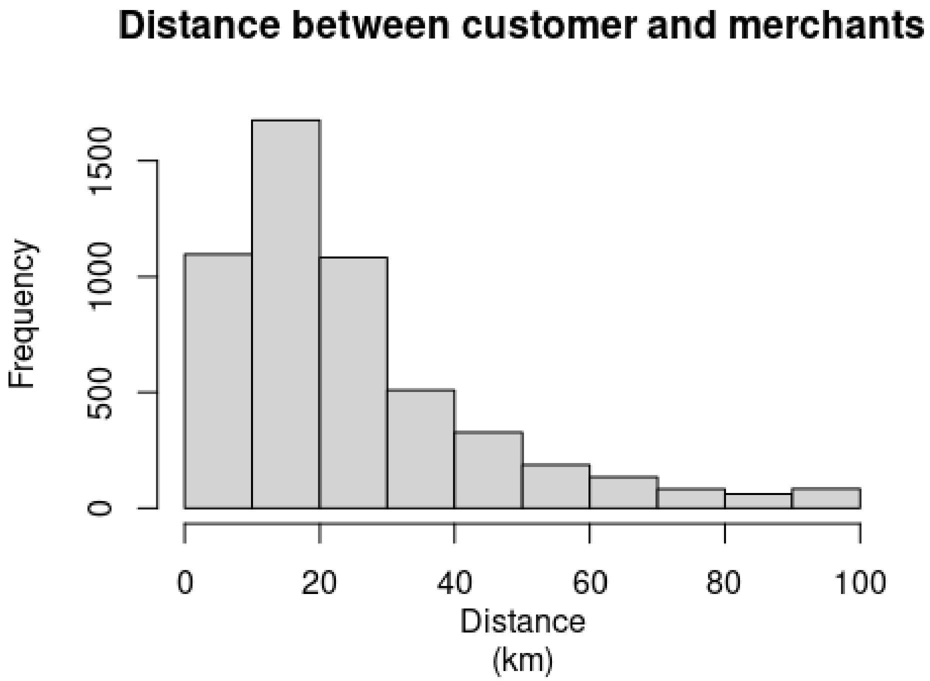
```
df4 <- df %>%
  select(date,hour) %>%
  group_by(date,hour) %>%
  summarize(trans_vol=n()) %>%
  group_by(hour) %>%
  summarize(trans_vol_per_hr = mean(trans_vol,na.rm=TRUE))
```

```
## `summarise()` has grouped output by 'date'. You can override using the `.g
roups` argument.
```

```
ggplot(df4,aes(x=hour,y=trans_vol_per_hr))+geom_point()+geom_line(aes(group =
1))+
  ggtitle('Average transaction volume by hour') +
  labs(x='Hour',y='Transaction volume') + expand_limits( y = 0)
```



```
# exclude the single foreign customer whose location information was incorrec
tly stored (i.e latitude 573)
df_temp <- df_csmp %>%
  filter (c_long >113 & c_long <154 & c_lat > (-44) & c_lat < (-10))
dfloc = df_temp [,c("c_long", "c_lat","m_long", "m_lat")]
dfloc<- data.frame(sapply(dfloc, as.numeric))
dfloc$dst <- distHaversine(dfloc[, 1:2], dfloc[, 3:4]) / 1000
hist(dfloc$dst[dfloc$dst<100], main = "Distance between customer and merchant
s",xlab= 'Distance
(km)' )
```


Distance between customer and merchants
Frequency
0
500
1000
1500
0
20
40
60
80
100
Distance
(km)