Internship_1

```
library(stringr)
library(readxl)
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
#library(tidyverse)
#install.packages("modelr")
#library(modelr)
library(sp)
```

```
## Warning: package 'sp' was built under R version 3.6.3
```

```
#install.packages("leaflet")

library(leaflet)
```

```
## Warning: package 'leaflet' was built under R version 3.6.3
```

```
#install.packages("geosphere")
library(geosphere)
```

```
## Warning: package 'geosphere' was built under R version 3.6.3
```

```
#library(knitr)
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 3.6.3
```

```
Transactional_data <- read_xlsx("ANZ Internship/ANZ synthesised transaction dataset.xlsx")
#After loading the transactional dataset we get the below issue Expecting #numeric in C3052 / R3052C3:
#In the row no :C3052 and C4360 we found there are non-numeric elements which are replaced by zero as n


#The dataset contains 12043 transactions for 100 customers who have one #bank account each. Trasactiona
#analysis. For each record/row, information is complete for majority of columns. Some columns contain m
#data (blank or NA cells), which is likely due to the nature of transaction. (i.e. merchants are not in
#InterBank transfers or Salary payments) It is also noticed that there is #only 91 unique dates in the
#The range of each feature should also be examined which shows that there is one customer that resides
```

```
#learn about the data
summary(Transactional_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   :11158
##    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
##  Min.   :0       Length:12043       Min.   :     0.24
##  1st Qu.:0       Class :character   1st Qu.:  3158.59
##  Median :0       Mode  :character   Median :  6432.01
```

```
##  Mean   :0                          Mean   : 14704.20
##  3rd Qu.:0                          3rd Qu.: 12465.94
##  Max.   :0                          Max.   :267128.52
##  NA's   :11160
##       date                          gender              age
##  Min.   :2018-08-01 00:00:00   Length:12043       Min.   :18.00
##  1st Qu.:2018-08-24 00:00:00   Class :character   1st Qu.:22.00
##  Median :2018-09-16 00:00:00   Mode  :character   Median :28.00
##  Mean   :2018-09-15 21:27:39                      Mean   :30.58
##  3rd Qu.:2018-10-09 00:00:00                      3rd Qu.:38.00
##  Max.   :2018-10-31 00:00:00                      Max.   :78.00
##
##  merchant_suburb    merchant_state      extraction            amount
##  Length:12043       Length:12043       Length:12043       Min.   :   0.10
##  Class :character   Class :character   Class :character   1st Qu.:  16.00
##  Mode  :character   Mode  :character   Mode  :character   Median :  29.00
##                                                           Mean   : 187.93
##                                                           3rd Qu.:  53.66
##                                                           Max.   :8835.98
##
##  transaction_id       country          customer_id        merchant_long_lat
##  Length:12043       Length:12043       Length:12043       Length:12043
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    movement
##  Length:12043
##  Class :character
##  Mode  :character
##
##
##
##
```

```
str(Transactional_data)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    12043 obs. of  23 variables:
##  $ status           : chr  "authorized" "authorized" "authorized" "authorized" ...
##  $ card_present_flag: num  1 0 1 1 1 NA 1 1 1 NA ...
##  $ bpay_biller_code : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ account          : chr  "ACC-1598451071" "ACC-1598451071" "ACC-1222300524" "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" "830a451c-316e-4a6a-bf25-e37caedca4
##  $ merchant_code    : num  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             : POSIXct, format: "2018-08-01" "2018-08-01" ...
##  $ gender           : chr  "F" "F" "M" "F" ...
##  $ age              : num  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" "2018-08-01T01:13:45.000+0000" "2018-08-01
##  $ amount           : num  16.25 14.19 6.42 40.9 3.25 ...
##  $ transaction_id   : chr  "a623070bfead4541a6b0fff8a09e706c" "13270a2a902145da9db4c951e04b51b9" "fe
##  $ country          : chr  "Australia" "Australia" "Australia" "Australia" ...
##  $ customer_id      : chr  "CUS-2487424745" "CUS-2487424745" "CUS-2142601169" "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" ...
```

```
# Format Date
Transactional_data$date<- as.Date(Transactional_data$date,format = "%d/%m/%Y")
#To find which date was missing
DateRange <- seq(min(Transactional_data$date), max(Transactional_data$date), by = 1)
DateRange[!DateRange %in% Transactional_data$date]
```

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

```
# 2018-08-16 date  transactions are missing
#derive weekday and hour data of each transaction
Transactional_data$extraction = as.character(Transactional_data$extraction)
Transactional_data$hour = hour(as.POSIXct(substr(Transactional_data$extraction,12,19),format="%H:%M:%S"
Transactional_data$weekday = weekdays(Transactional_data$date)

#Split customer's logitude and latitude information  and merchant's long an #d lat using 'seperate'
dfloc = Transactional_data[,c("long_lat","merchant_long_lat")]
dfloc<- dfloc %>% separate("long_lat", c("cust_long", "cust_lat"),sep=' ')
dfloc<- dfloc %>% separate("merchant_long_lat", c("mer_long", "mer_lat"),sep=' ')
dfloc<- data.frame(sapply(dfloc, as.numeric))
df <- cbind(Transactional_data,dfloc)
# 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         cust_long          cust_lat          mer_long
##                 0                 0                 0              4326
##           mer_lat
##              4326
```

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

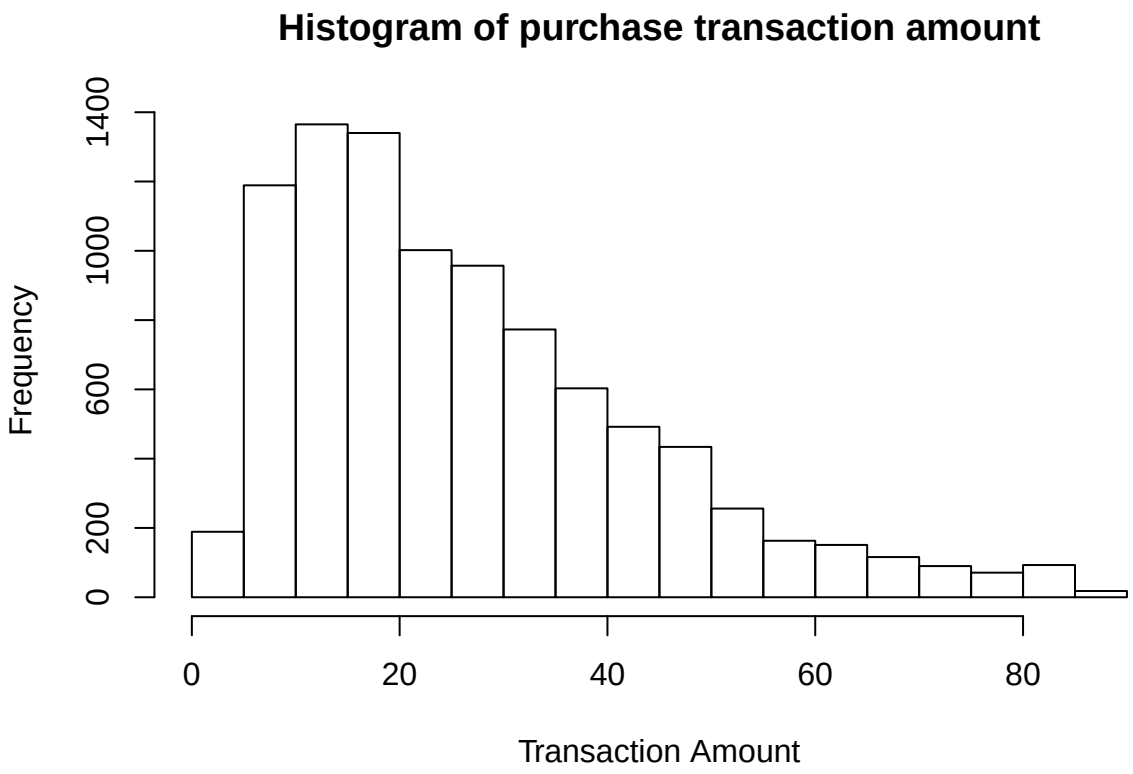
```
##            status card_present_flag  bpay_biller_code           account
##                 2                 3                 2               100
##          currency          long_lat   txn_description       merchant_id
##                 1               100                 6              5726
##     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         cust_long          cust_lat          mer_long
##                 7                87                85               719
##           mer_lat
##               670
```

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

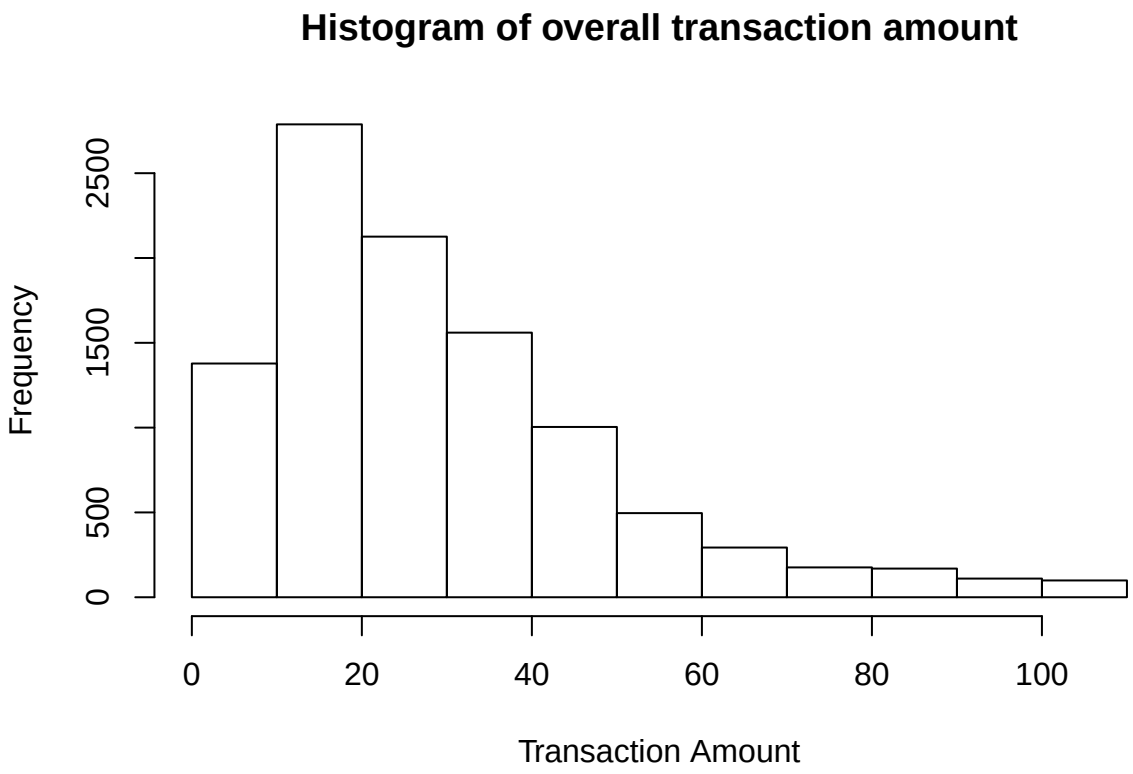
```
##     status           card_present_flag bpay_biller_code   account
##  Length:10317       Min.   :0.0000    Min.   :0        Length:10317
##  Class :character   1st Qu.:1.0000    1st Qu.:0        Class :character
##  Mode  :character   Median :1.0000    Median :0        Mode  :character
##                     Mean   :0.8026    Mean   :0
##                     3rd Qu.:1.0000    3rd Qu.:0
##                     Max.   :1.0000    Max.   :0
##                     NA's   :2600      NA's   :10315
##    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            cust_long        cust_lat          mer_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
##     mer_lat
##  Min.   :-43.31
##  1st Qu.:-37.71
##  Median :-33.84
##  Mean   :-32.75
##  3rd Qu.:-29.44
##  Max.   :-12.33
##  NA's   :2600
```

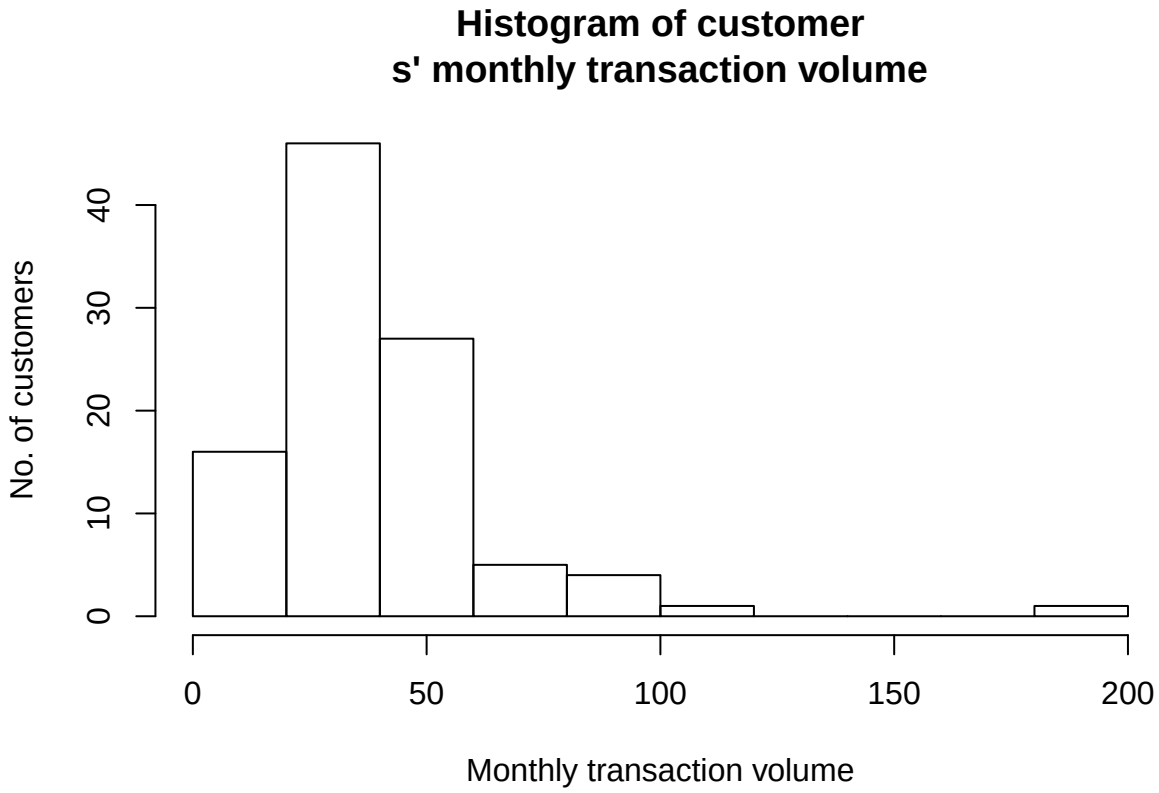
```r
# visualise the distribution of transaction amount
hist(df_csmp$amount[!df_csmp$amount %in% boxplot.stats(df_csmp$amount)$out], #include outliers
xlab= 'Transaction Amount', main = 'Histogram of purchase transaction amount')
```



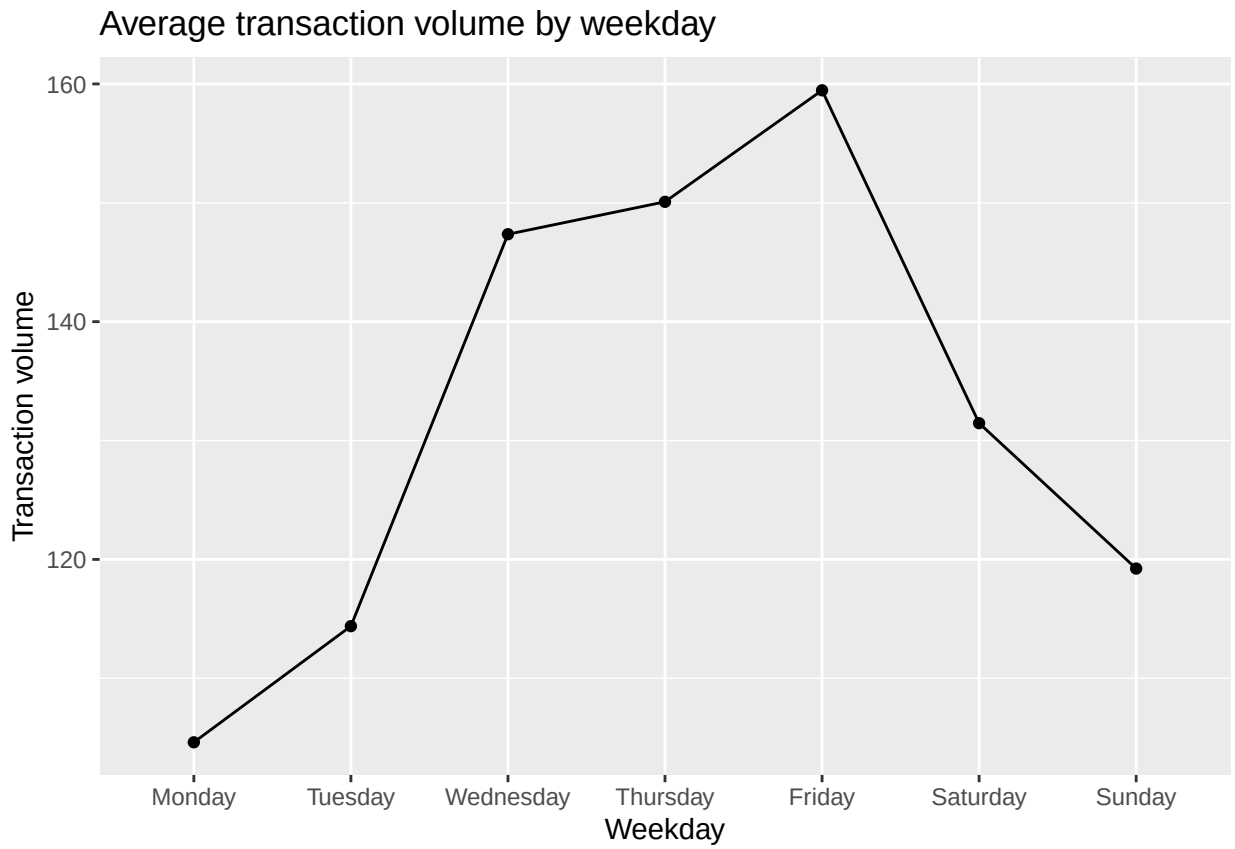

```
hist(df$amount[!df$amount %in% boxplot.stats(df$amount)$out], #exclude outliers
xlab= 'Transaction Amount',main = 'Histogram of overall transaction amount')
```



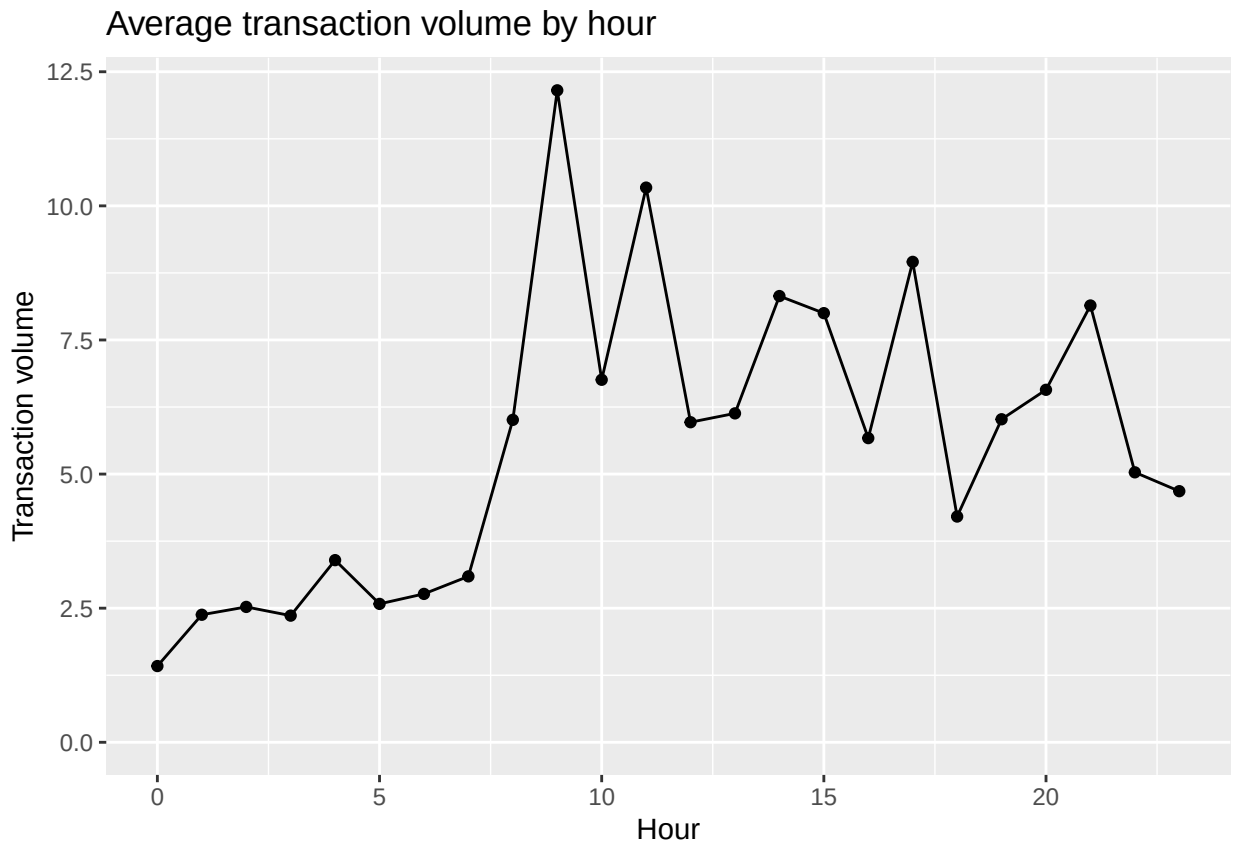

```
df2 <- df %>%
group_by(customer_id) %>%
summarise(mon_avg_vol = round(n()/3,0))
df2 <- df %>%
group_by(customer_id) %>%
summarise(mon_avg_vol = round(n()/3,0))
hist(df2$mon_avg_vol,
xlab= 'Monthly transaction volume', ylab='No. of customers', main = "Histogram of customer
s' monthly transaction volume")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



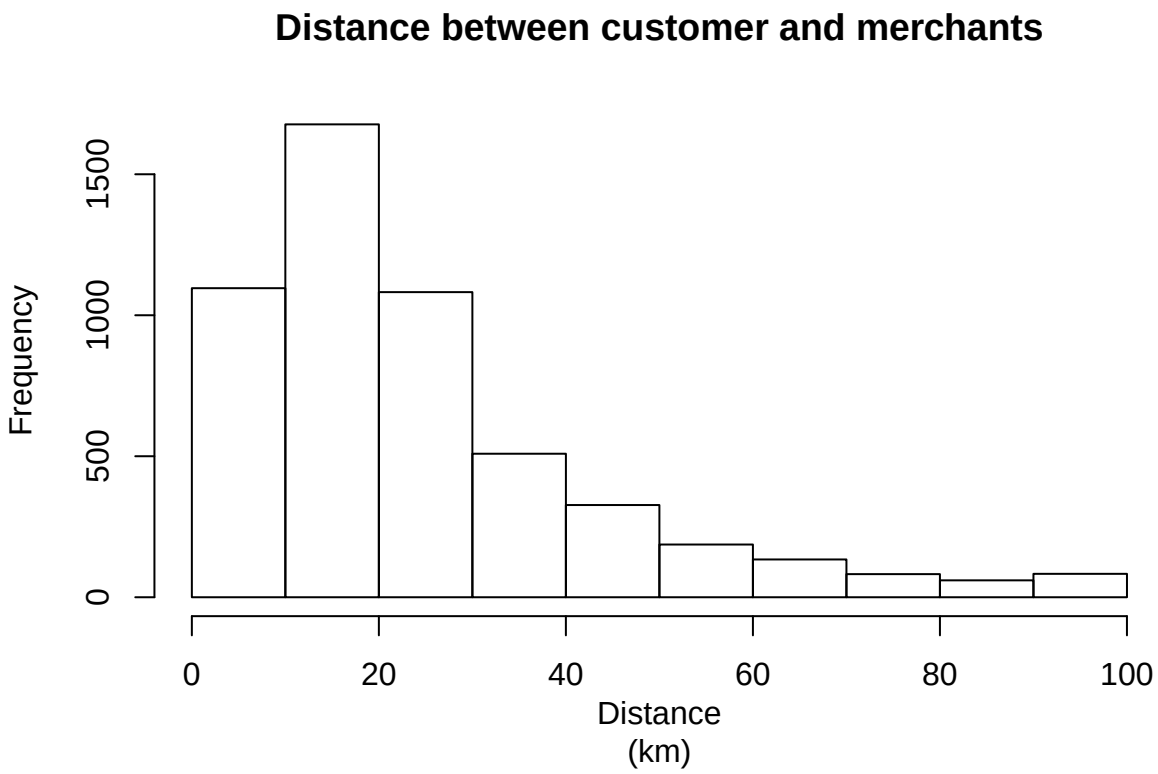

```
#Handling date and time
df3 <- df %>%
select(date,weekday) %>%
group_by(date,weekday) %>%
summarise(daily_avg_vol = n()) %>%
group_by(weekday) %>%
summarise(avg_vol=mean(daily_avg_vol,na.rm=TRUE ))
df3$weekday <- factor(df3$weekday, levels=c( "Monday","Tuesday","Wednesday",
"Thursday","Friday","Saturday","Sunday"))
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')
```



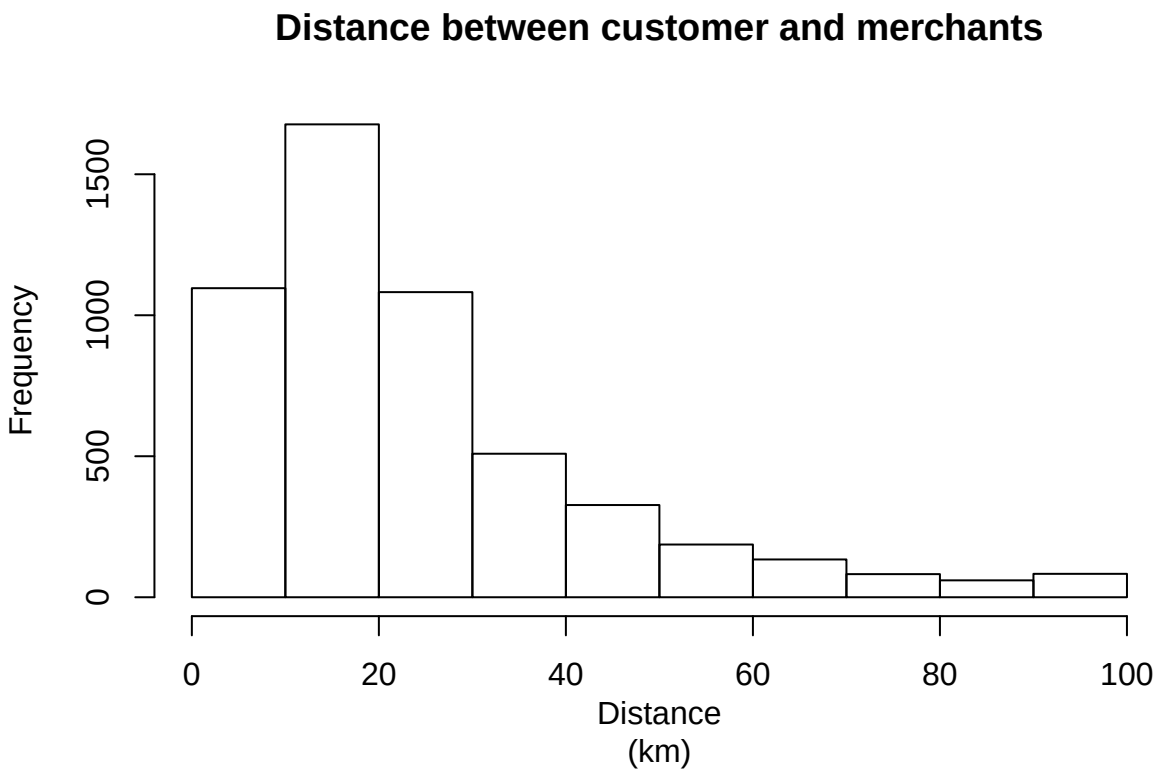

```
# visualize transaction volume over an average week.
df4 <- df %>%
select(date,hour) %>%
group_by(date,hour) %>%
summarize(trans_vol=n()) %>%
group_by(hour) %>%
summarize(trans_vol_per_hr = mean(trans_vol,na.rm=TRUE))
ggplot(df4,aes(x=hour,y=trans_vol_per_hr))+geom_point()+geom_line(aes(group = 1))+
ggtitle('Average transaction volume by hour') +
labs(x='Hour',y='Transaction volume') + expand_limits( y = 0)
```




```
#Location details
df_temp <- df_csmp %>%
filter (cust_long >113 & cust_long <154 & cust_lat > (-44) & cust_lat < (-10))
dfloc = df_temp [,c("cust_long", "cust_lat","mer_long", "mer_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 merchants",xlab= 'Distance
(km)' )
```




```
df_temp <- df_csmp %>%
filter (cust_long >113 & cust_long <154 & cust_lat > (-44) & cust_lat < (-10))
dfloc = df_temp [,c("cust_long", "cust_lat","mer_long", "mer_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 merchants",xlab= 'Distance
(km)' )
```


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