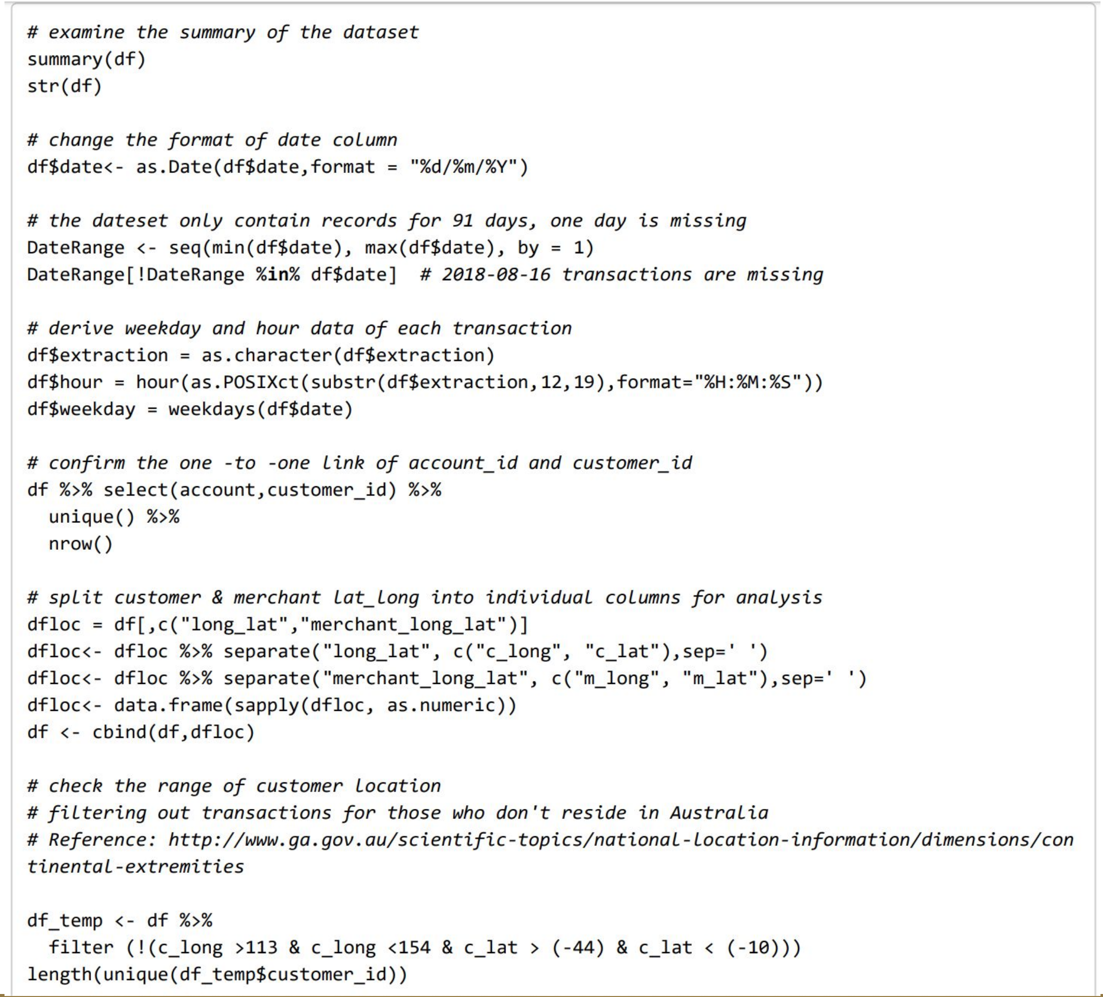
# Predictive Analysis

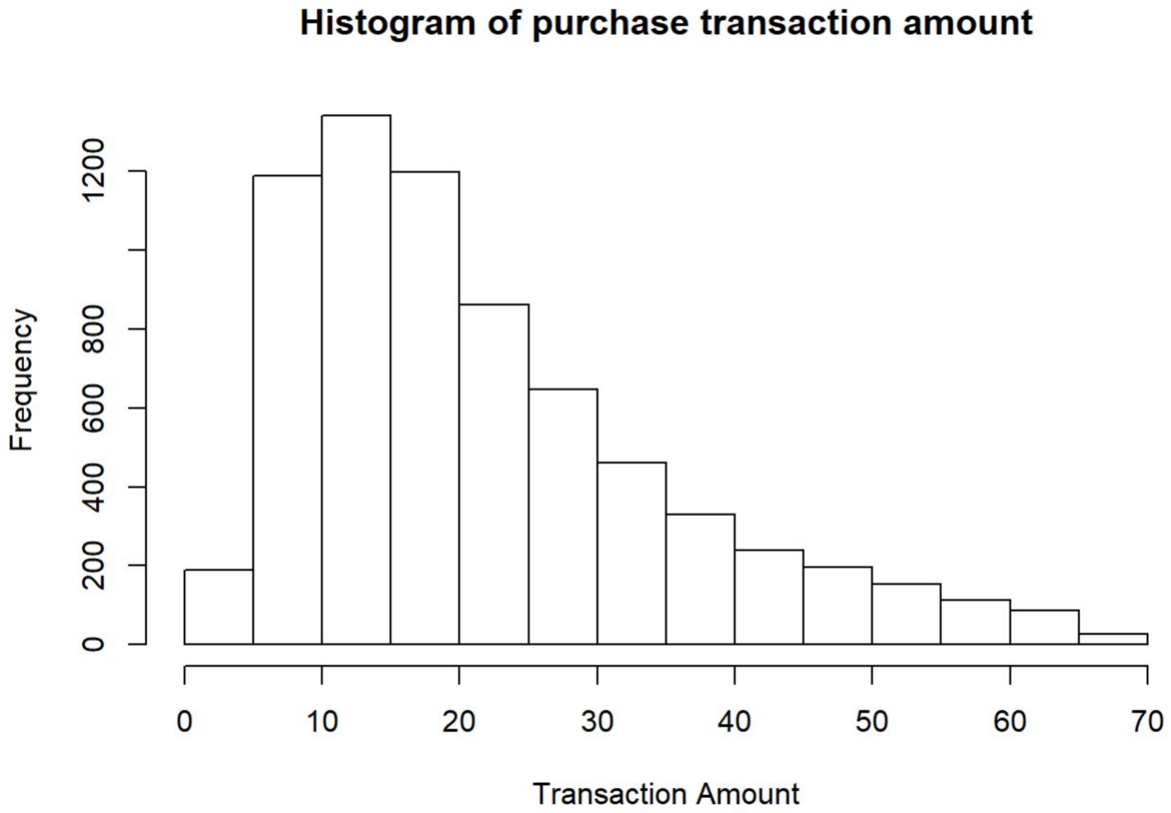
Faaz Arshad

# Data Preparation

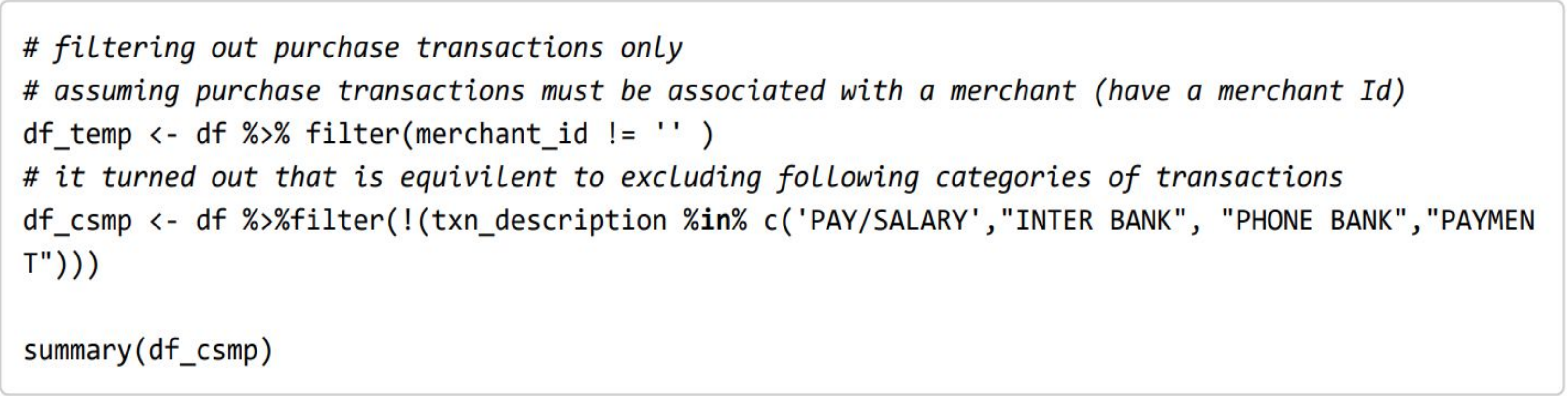
The dataset contains 12043 transactions for 100 customers who have one bank account each. Transactional period is from 01/08/2018 - 31/10/2018 (92 days duration). The data entries are unique and have consistent formats for analysis. For each record/row, information is complete for majority of columns. Some columns contain missing data (blank or NA cells), which is likely due to the nature of transaction. (i.e. merchants are not involved for InterBank transfers or Salary payments) It is also noticed that there is only 91 unique dates in the dataset, suggesting the transaction records for one day are missing (turned out to be 2018-08-16). The range of each feature should also be examined which shows that there is one customer that resides outside Australia.



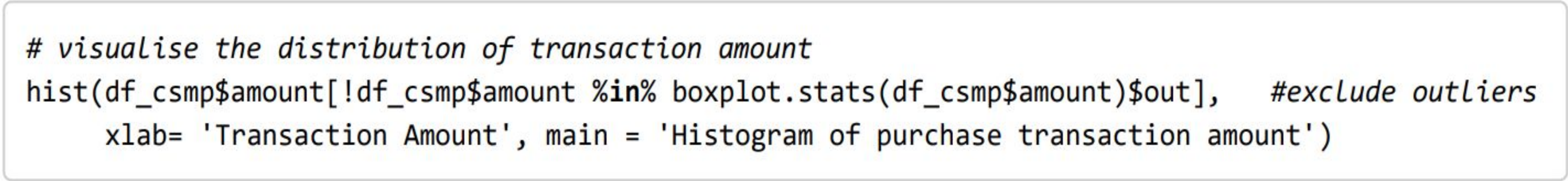
```r
# examine the summary of the dataset
summary(df)
str(df)

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

# 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]  # 2018-08-16 transactions are missing

# 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)

# confirm the one -to -one link of account_id and customer_id
df %>% select(account,customer_id) %>%
  unique() %>%
  nrow()

# 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=' ')
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
# Reference: http://www.ga.gov.au/scientific-topics/national-location-information/dimensions/con
tinental-extremities

df_temp <- df %>%
  filter (!(c_long >113 & c_long <154 & c_lat > (-44) & c_lat < (-10)))
length(unique(df_temp$customer_id))
```



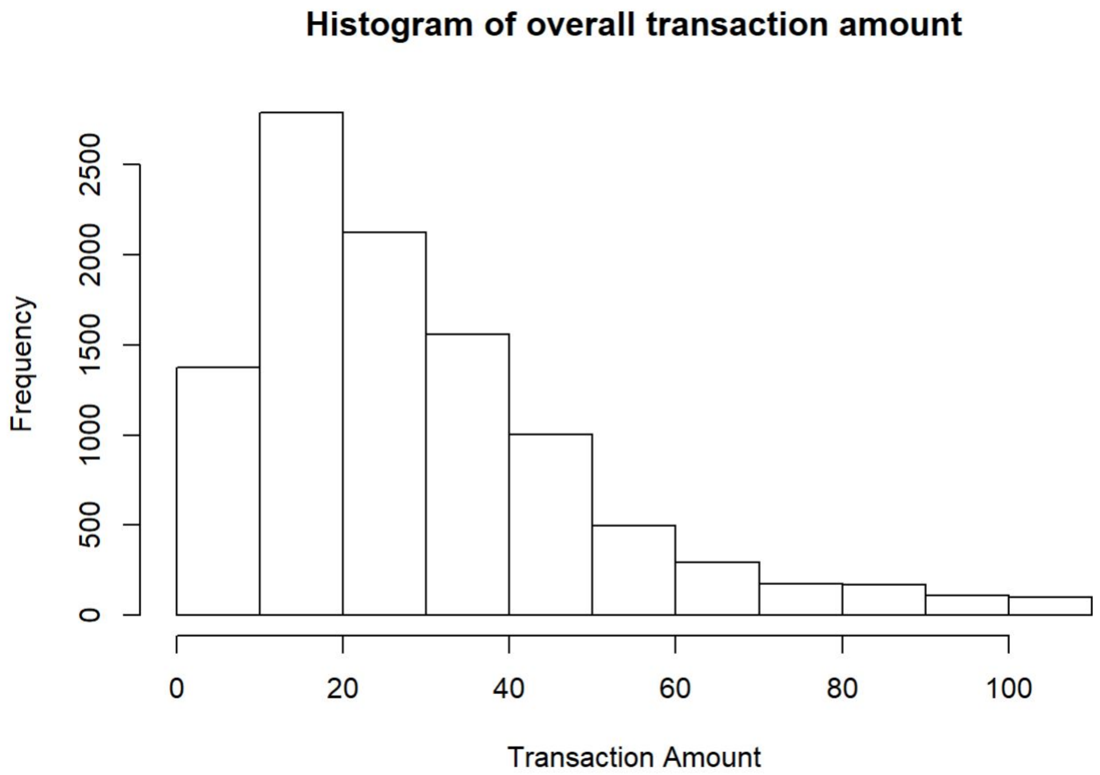
Histogram of purchase transaction amount

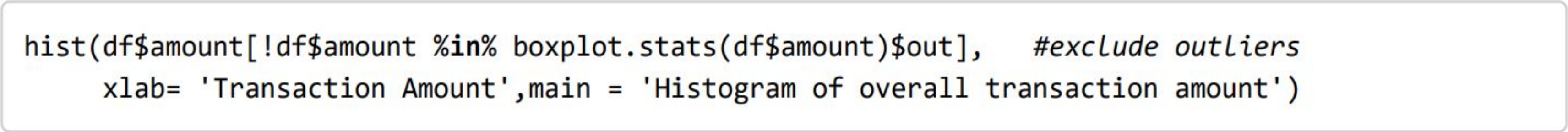


```
# filtering out purchase transactions only
# assuming purchase transactions must be associated with a merchant (have a merchant Id)
df_temp <- df %>% filter(merchant_id != '' )
# it turned out that is equivilent to excluding following categories of transactions
df_csmp <- df %>%filter(!(txn_description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

summary(df_csmp)
```

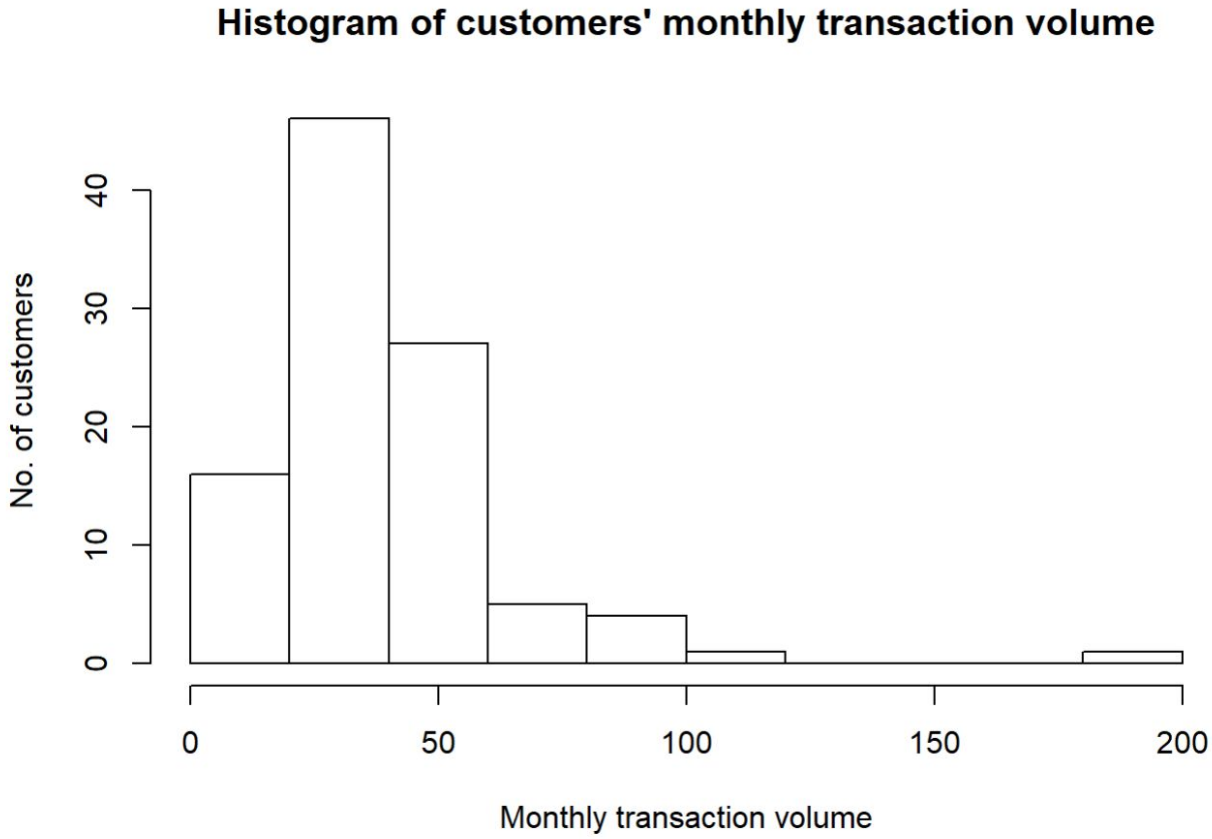


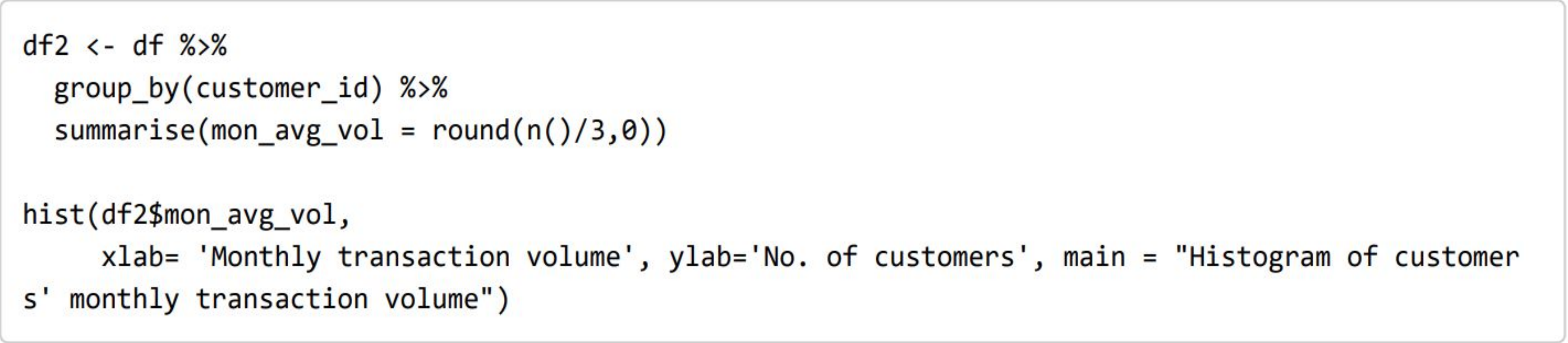
```
# visualise the distribution of transaction amount
hist(df_csmp$amount[!df_csmp$amount %in% boxplot.stats(df_csmp$amount)$out],   #exclude outliers
     xlab= 'Transaction Amount', main = 'Histogram of purchase transaction amount')
```


Histogram of overall transaction amount
Frequency
0
500
1000
1500
2000
2500
0
20
40
60
80
100
Transaction Amount

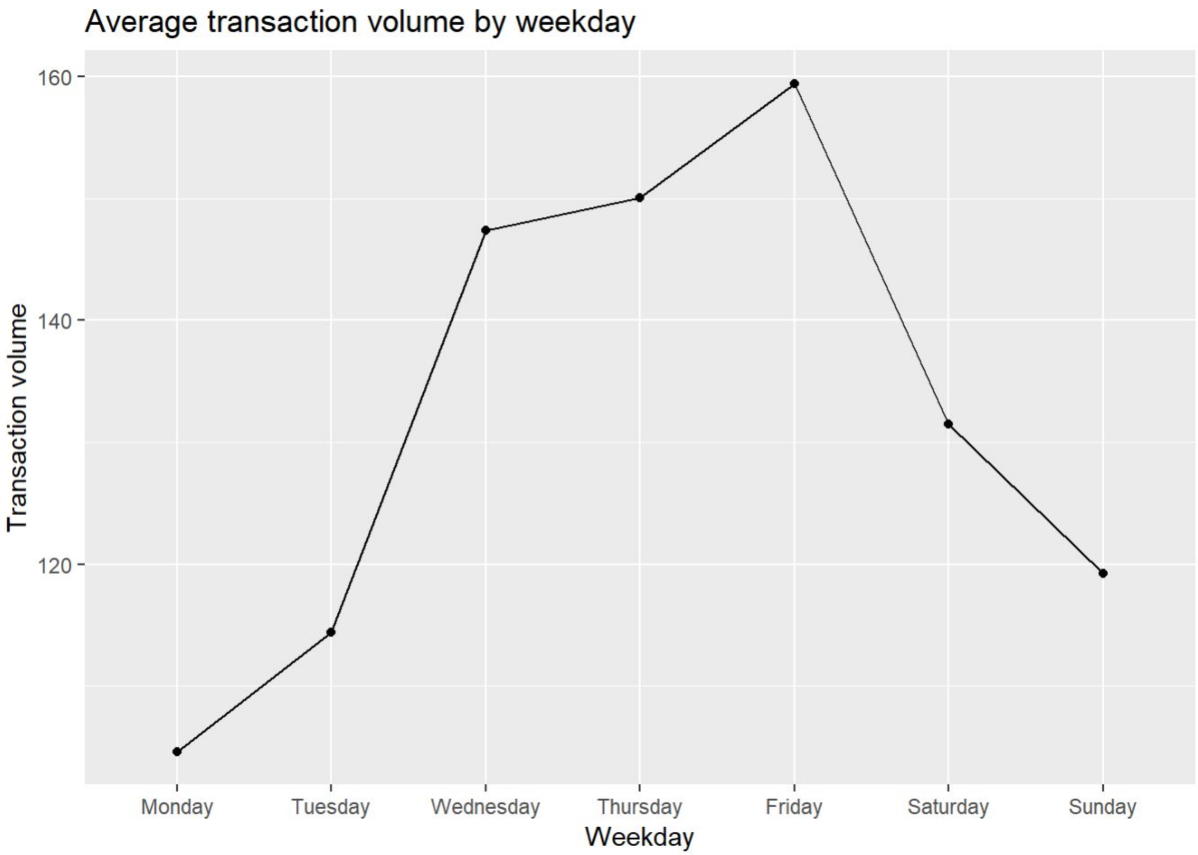


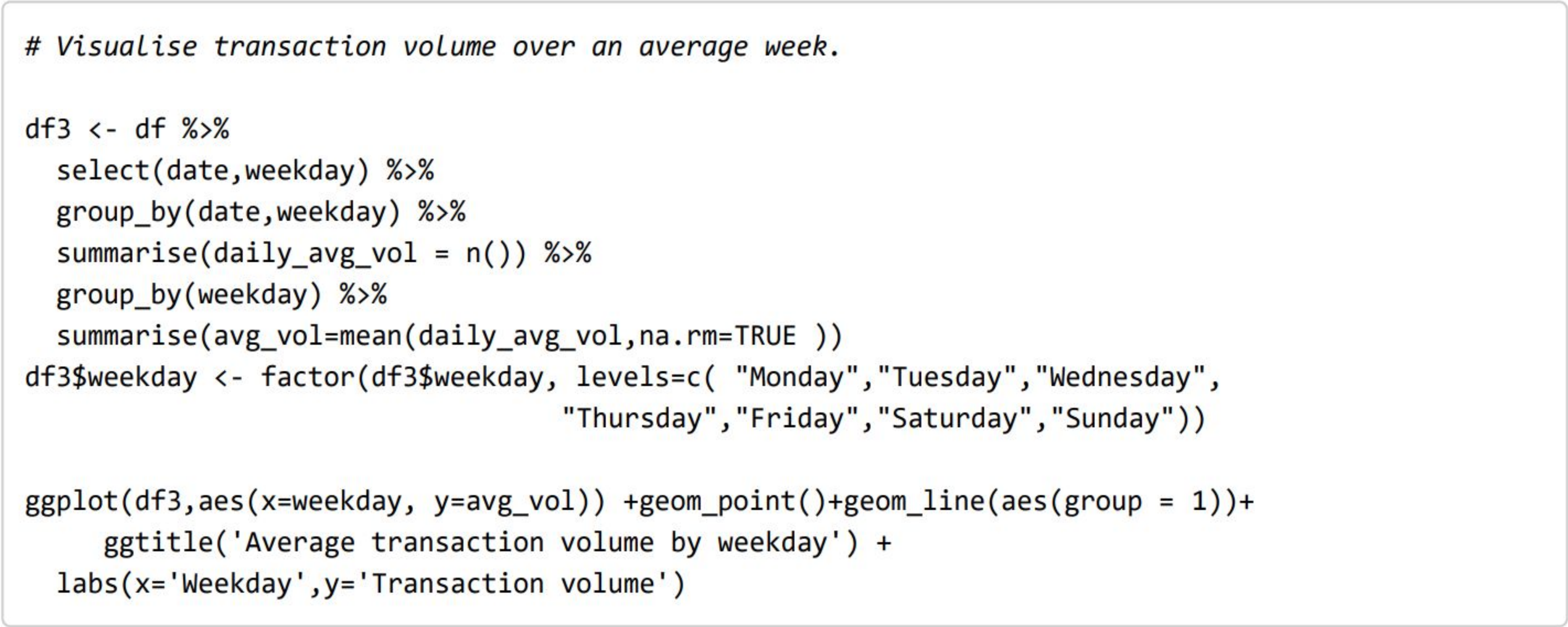
```
hist(df$amount[!df$amount %in% boxplot.stats(df$amount)$out],   #exclude outliers
     xlab= 'Transaction Amount',main = 'Histogram of overall transaction amount')
```


Histogram of customers' monthly transaction volume
No. of customers
0
10
20
30
40
0
50
100
150
200
Monthly transaction volume

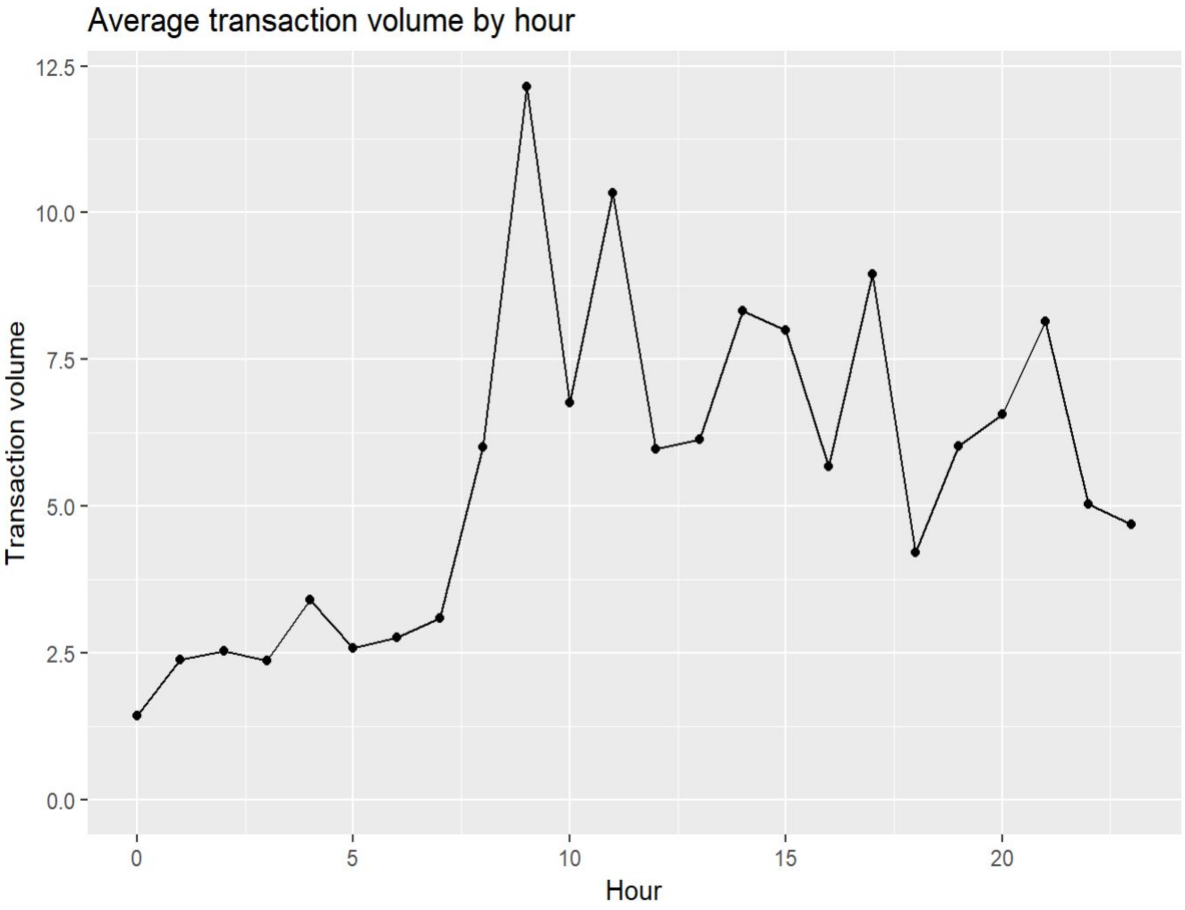


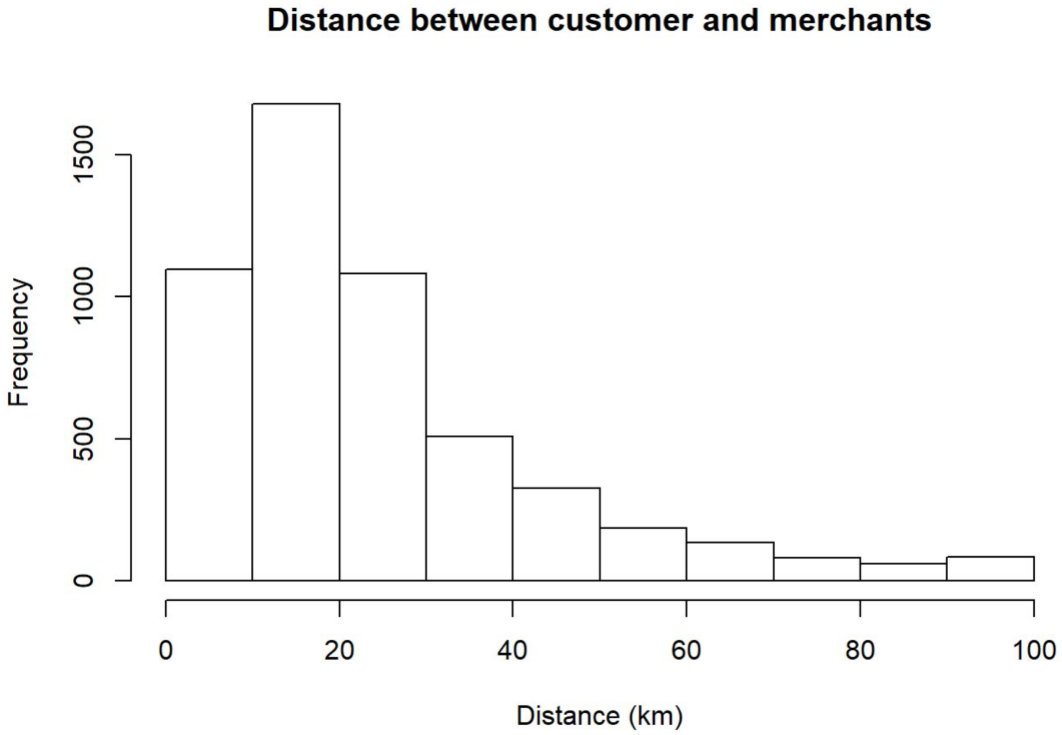
```
df2 <- df %>%
  group_by(customer_id) %>%
  summarise(mon_avg_vol = round(n()/3,0))

hist(df2$mon_avg_vol,
     xlab= 'Monthly transaction volume', ylab='No. of customers', main = "Histogram of customer
s' monthly transaction volume")
```

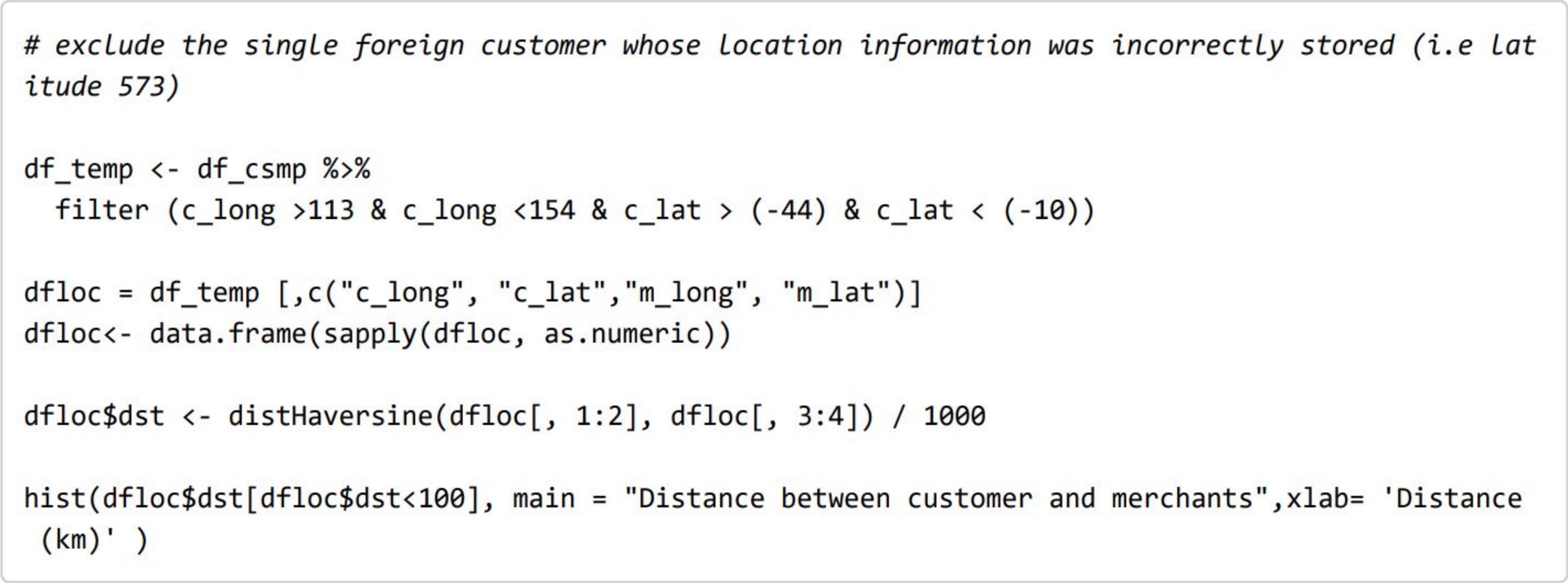

Average transaction volume by weekday
160
140
120
Transaction volume
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Weekday



```
# 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 ))
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')
```


Average transaction volume by hour
Transaction volume
12.5
10.0
7.5
5.0
2.5
0.0
0
5
10
15
20
Hour


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



```r
# exclude the single foreign customer whose location information was incorrectly stored (i.e lat
itude 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 merchants",xlab= 'Distance
 (km)' )
```

# THE END