Trnsact

2023-10-13

The summary table below shows all the key statistics that are calculated for the first 100,000 rows of trnsact data, among them we could find that the column **sprice** and **amt** seems to contain the same value, but with deeper observations, we find the association that the value of **sprice** times the value of **quantity** will become the value of **amt**.

```r
transact <- read.csv("trnsact.csv", nrows = 100000, sep=",", header = FALSE,
                     strip.white = TRUE, quote = "", na.strings='NA',
                     stringsAsFactors = TRUE, fill = TRUE,
                     col.names = c('SKU', 'STORE', 'REGISTER', 'TRANNUM', 'SEQ',
                                   'SALEDATE', 'STYPE', 'QUANTITY', 'ORGPRICE',
                                   'SPRICE', 'AMT', 'INTERID', 'MIC', 'Unknown'))
summary(transact)
```

```
##       SKU           STORE         REGISTER        TRANNUM
##  Min.   :   3   Min.   : 102   Min.   :  1.0   Min.   :  100
##  1st Qu.:4310   1st Qu.:2104   1st Qu.:190.0   1st Qu.: 1000
##  Median :7367   Median :4104   Median :373.0   Median : 2200
##  Mean   :6311   Mean   :4460   Mean   :404.2   Mean   : 3241
##  3rd Qu.:7915   3rd Qu.:7104   3rd Qu.:580.0   3rd Qu.: 4100
##  Max.   :9633   Max.   :9909   Max.   :993.0   Max.   :99500
##
##       SEQ                  SALEDATE     STYPE        QUANTITY    ORGPRICE
##  Min.   :        0   2005-02-26: 1060   P:91732   Min.   :1   Min.   :  0.00
##  1st Qu.:        0   2005-02-23:  918   R: 8268   1st Qu.:1   1st Qu.: 19.50
##  Median :        0   2005-02-25:  899             Median :1   Median : 30.00
##  Mean   :193188898   2005-07-30:  830             Mean   :1   Mean   : 41.77
##  3rd Qu.:351600786   2005-02-24:  825             3rd Qu.:1   3rd Qu.: 50.00
##  Max.   :999906136   2005-02-19:  651             Max.   :1   Max.   :788.00
##                      (Other)   :94817
##      SPRICE            AMT            INTERID                MIC
##  Min.   :  0.00   Min.   :  0.00   Min.   :       17   Min.   :  1.0
##  1st Qu.: 13.99   1st Qu.: 13.99   1st Qu.:244300023   1st Qu.:205.0
##  Median : 19.50   Median : 19.50   Median :496900030   Median :358.0
##  Mean   : 26.86   Mean   : 26.86   Mean   :496055089   Mean   :437.5
##  3rd Qu.: 32.00   3rd Qu.: 32.00   3rd Qu.:747600091   3rd Qu.:680.0
##  Max.   :695.00   Max.   :695.00   Max.   :999900097   Max.   :999.0
##
##     Unknown
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.0203
##  3rd Qu.:0.0000
##  Max.   :1.0000
```

```
##
```

TRANNUM is typically another point that we probably focus on when coming up with our business question. Basically, we draw the two diagrams above to show the distribution of the total amount of transaction charge to the customer, with one in plain and the other grouping by the trannum, as shown below. Since we did not include all the observation in this dataset, thus we could not arbitraily conclude that the distribution plot is similar to Chi-Square/Exponential distribution respectively, but these two plots give us a hint and subtle directions to move forward!

```
library(ggplot2)
ggplot(transact, aes(x = AMT, fill = STYPE)) +
  geom_histogram(binwidth = 20, position = "identity", alpha = 0.5) +
  labs(title = "Distribution of Total amount of the Transaction Charge to the Customer",
       x = "Total Transaction Charge",
       y = "Frequency",
       fill = "STYPE") +
  scale_fill_manual(values = c("P" = "blue", "R" = "red")) +
  guides(fill = guide_legend(title = "Transaction Type"))



library(dplyr)
```

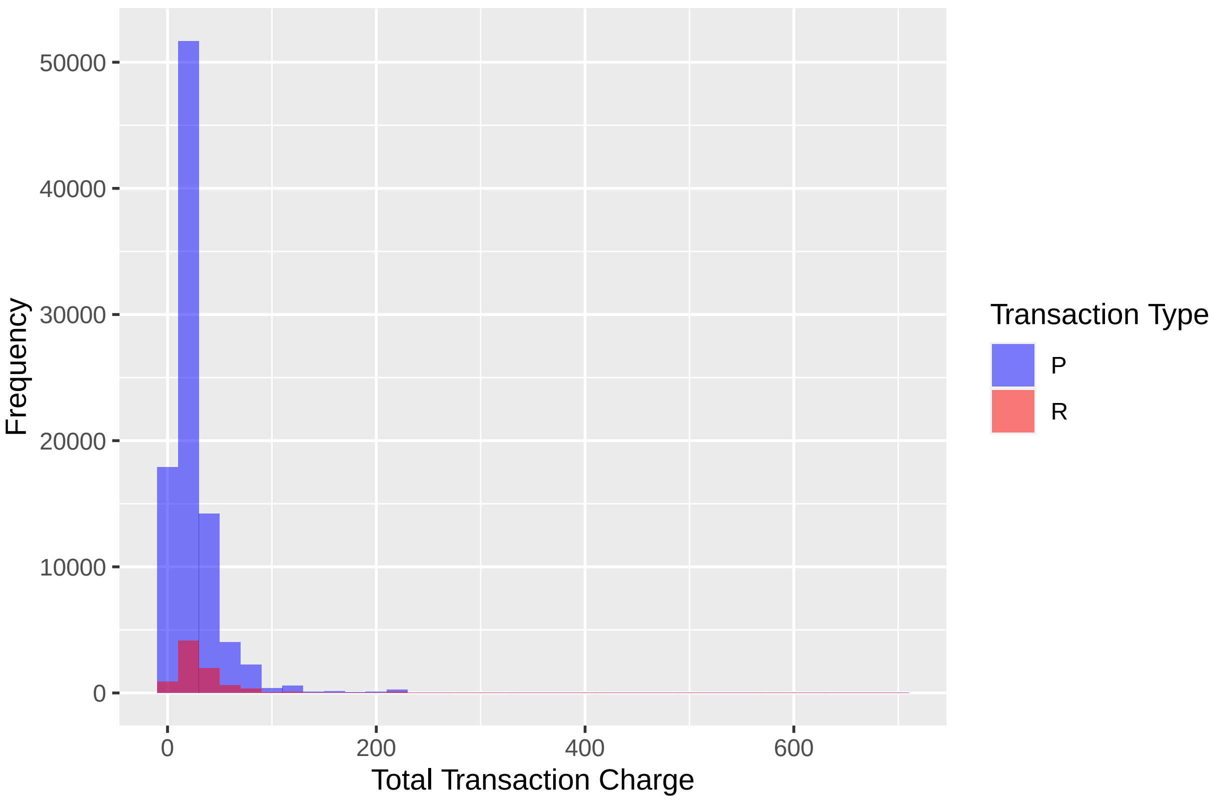
```
##
## Attaching package: 'dplyr'

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

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

```
transact_groupby_trannum <- transact %>%
  group_by(TRANNUM, STYPE) %>%
  summarize(Sum_AMT = sum(AMT))
```

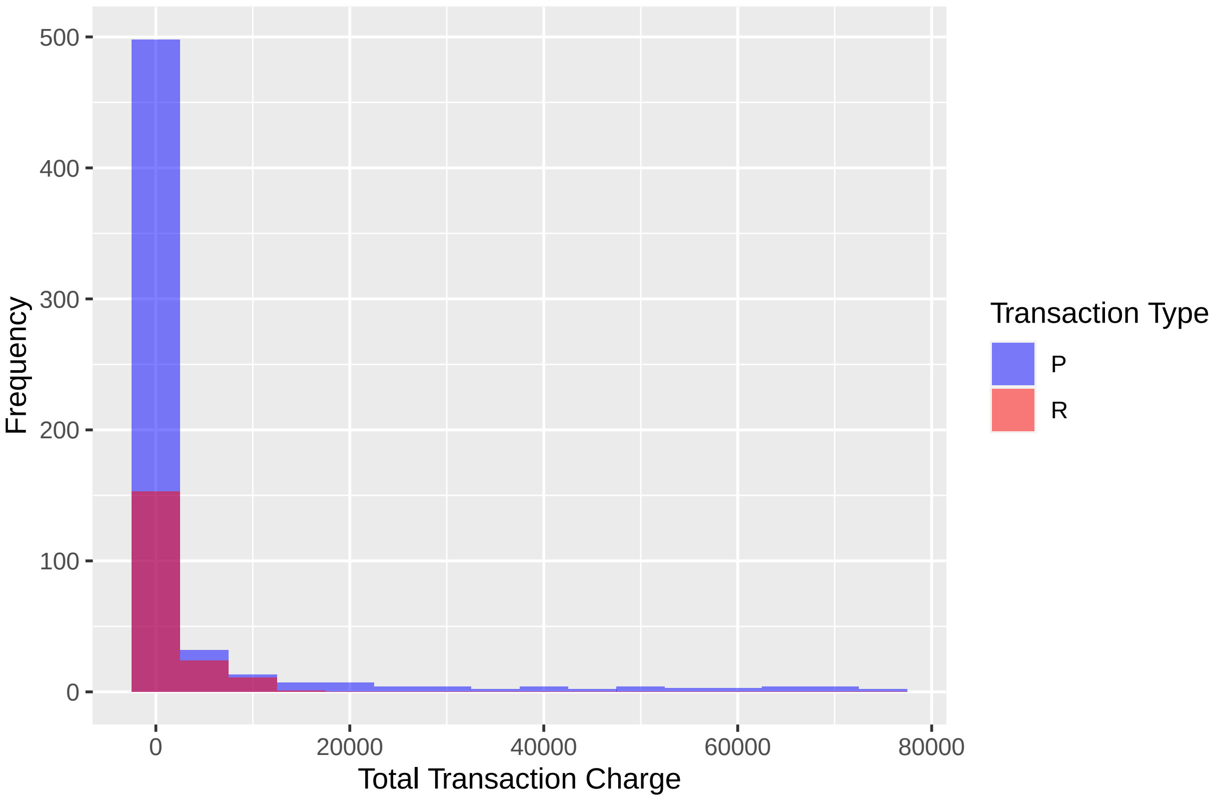
```
## `summarise()` has grouped output by 'TRANNUM'. You can override using the
## `.groups` argument.
```

```
ggplot(transact_groupby_trannum, aes(x = Sum_AMT, fill = STYPE)) +
  geom_histogram(binwidth = 5000, position = "identity", alpha = 0.5) +
  labs(title = "Distribution of Total amount for Each Transaction",
       x = "Total Transaction Charge",
       y = "Frequency",
       fill = "STYPE") +
  scale_fill_manual(values = c("P" = "blue", "R" = "red")) +
  guides(fill = guide_legend(title = "Transaction Type"))
```

Distribution of Total amount of the Transaction Charge to the Customer



Distribution of Total amount for Each Transaction


Frequency
Total Transaction Charge
Transaction Type
P
R