Cryptocurrency Transactions EDA

Thomas Jung

Exploratory Data Analysis

# getwd()  
file\_name <- "eda.xlsx"  
excel\_sheets(file\_name)

## [1] "test" "pivot"   
## [3] "FSA\_Province" "Price Limit Anomaly"   
## [5] "Price Limit anomaly - BTC" "Subtotal higher for buyers"  
## [7] "Subtotal lower for sellers"

data\_raw <- read\_excel(file\_name, sheet = "test")  
  
# str(data\_raw)  
# summary(data\_raw)  
  
data\_clean <- data\_raw %>%  
 rename(Order\_Datetime = "Order Time (UTC)", Assets\_Ordered = "Assets Ordered", Assets\_Requested = "Assets Requested", Price\_Limit = "Price Limit") %>%  
 mutate\_at(vars(UserID), as.factor) %>%  
 mutate\_at(vars(Order\_Datetime), ymd\_hms) %>%  
 mutate\_if(is.character, as.factor)  
  
# Check for missing TXNID  
TXNID\_min <- min(as.numeric(data\_clean$TXNID))  
TXNID\_max <- max(as.numeric(data\_clean$TXNID))  
identical(data\_clean$TXNID, as.numeric(seq(TXNID\_min, TXNID\_max)))

## [1] TRUE

# Check for missing UserID  
UserID\_min <- min(as.numeric(data\_clean$UserID))  
UserID\_max <- max(as.numeric(data\_clean$UserID))  
identical(as.numeric(levels(data\_clean$UserID)), as.numeric(seq(UserID\_min, UserID\_max)))

## [1] TRUE

# Missing Data  
sum(is.na(data\_clean))

## [1] 0

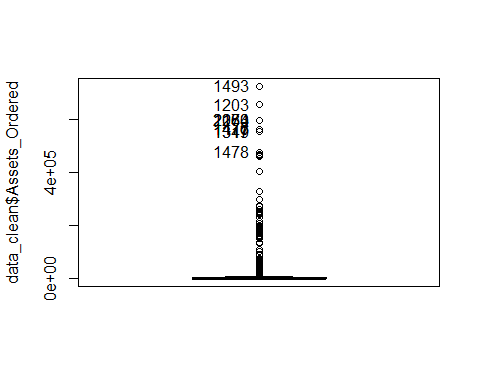
str(data\_clean)

## tibble [2,333 x 14] (S3: tbl\_df/tbl/data.frame)  
## $ TXNID : num [1:2333] 1 2 3 4 5 6 7 8 9 10 ...  
## $ UserID : Factor w/ 251 levels "1","2","3","4",..: 245 156 156 156 156 102 188 102 102 21 ...  
## $ FSA : Factor w/ 189 levels "E1E","E3G","H1Z",..: 156 37 37 37 37 21 98 21 21 90 ...  
## $ Province : Factor w/ 8 levels "AB","BC","MB",..: 2 5 5 5 5 5 5 5 5 5 ...  
## $ Order\_Datetime : POSIXct[1:2333], format: "2021-02-24 04:02:21" "2021-02-24 02:16:51" ...  
## $ Market\_Purchase : Factor w/ 38 levels "ADA","ALGO","BAT",..: 6 1 18 13 6 15 18 13 31 15 ...  
## $ Market\_Tradeoff : Factor w/ 4 levels "BTC","CAD","ETH",..: 2 2 2 2 2 2 2 2 3 2 ...  
## $ Type : Factor w/ 3 levels "ceiling","limit",..: 2 2 2 2 2 2 3 2 2 3 ...  
## $ Side : Factor w/ 2 levels "buy","sell": 2 1 2 2 2 1 1 2 2 1 ...  
## $ Assets\_Ordered : num [1:2333] 0.2 9100 40 0.68 0.139 ...  
## $ Assets\_Requested: num [1:2333] 0.2 9100 40 0.68 0.139 ...  
## $ Price\_Limit : num [1:2333] 64400 1.24 34.2 2030 63430 ...  
## $ Subtotal : num [1:2333] 12880 11418 1352 1364 8693 ...  
## $ Fees : num [1:2333] 0 16.93 2.05 2.07 13.2 ...

summary(data\_clean)

## TXNID UserID FSA Province   
## Min. : 1 245 : 161 v3b : 161 ON :1075   
## 1st Qu.: 584 133 : 79 M3C : 79 BC : 660   
## Median :1167 134 : 68 m1w : 68 AB : 397   
## Mean :1167 1 : 60 V3K : 65 QC : 174   
## 3rd Qu.:1750 215 : 52 L0K : 64 MB : 16   
## Max. :2333 76 : 50 M3H : 52 SK : 6   
## (Other):1863 (Other):1844 (Other): 5   
## Order\_Datetime Market\_Purchase Market\_Tradeoff Type   
## Min. :2021-01-01 00:27:21 HBAR :722 BTC : 232 ceiling: 18   
## 1st Qu.:2021-01-21 14:03:30 BTC :539 CAD :2013 limit :1255   
## Median :2021-02-05 12:00:14 ETH :224 ETH : 51 market :1060   
## Mean :2021-02-02 03:39:24 ADA :215 USDT: 37   
## 3rd Qu.:2021-02-13 20:43:39 DOGE : 88   
## Max. :2021-02-24 04:02:21 LINK : 63   
## (Other):482   
## Side Assets\_Ordered Assets\_Requested Price\_Limit   
## buy :1868 Min. : 0.0 Min. : 0.0 Min. : 0.0   
## sell: 465 1st Qu.: 0.5 1st Qu.: 0.5 1st Qu.: 0.1   
## Median : 135.6 Median : 135.6 Median : 0.4   
## Mean : 13662.7 Mean : 13662.7 Mean : 11943.3   
## 3rd Qu.: 3500.0 3rd Qu.: 3500.0 3rd Qu.: 2134.3   
## Max. :727242.9 Max. :727242.9 Max. :520026.0   
##   
## Subtotal Fees   
## Min. : 0.0 Min. : 0.0000   
## 1st Qu.: 43.7 1st Qu.: 0.0015   
## Median : 350.8 Median : 0.2181   
## Mean : 2493.7 Mean : 1.5256   
## 3rd Qu.: 1942.7 3rd Qu.: 1.4726   
## Max. :95903.1 Max. :142.8596   
##

# Outliers  
outlier\_rows <- Boxplot(data\_clean$Assets\_Ordered)



data\_clean[outlier\_rows,]

## # A tibble: 10 x 14  
## TXNID UserID FSA Province Order\_Datetime Market\_Purchase  
## <dbl> <fct> <fct> <fct> <dttm> <fct>   
## 1 1493 133 M3C ON 2021-01-29 00:34:07 DOGE   
## 2 1203 133 M3C ON 2021-02-04 15:32:48 DOGE   
## 3 2183 133 M3C ON 2021-01-07 00:17:05 SC   
## 4 2270 133 M3C ON 2021-01-03 23:54:33 SC   
## 5 1064 133 M3C ON 2021-02-07 01:34:50 DOGE   
## 6 1476 133 M3C ON 2021-01-29 03:11:08 DOGE   
## 7 1477 133 M3C ON 2021-01-29 03:09:56 DOGE   
## 8 1319 1 L0K ON 2021-02-01 12:29:55 STMX   
## 9 1347 1 L0K ON 2021-01-31 16:59:47 STMX   
## 10 1478 133 M3C ON 2021-01-29 03:00:25 DOGE   
## # ... with 8 more variables: Market\_Tradeoff <fct>, Type <fct>, Side <fct>,  
## # Assets\_Ordered <dbl>, Assets\_Requested <dbl>, Price\_Limit <dbl>,  
## # Subtotal <dbl>, Fees <dbl>

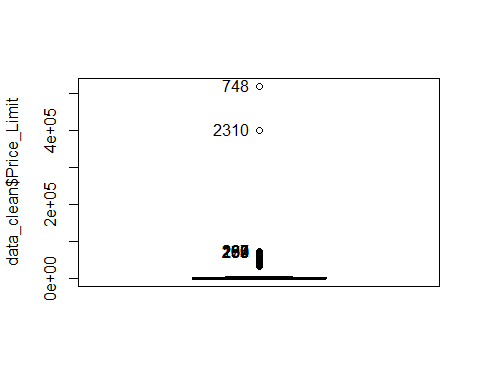
data\_clean %>%  
 filter(UserID == 133) %>%  
 arrange(desc(Assets\_Ordered)) %>%  
 select(-c(TXNID, FSA, Province, Order\_Datetime, Market\_Tradeoff, Type, Side, Assets\_Requested))

## # A tibble: 79 x 6  
## UserID Market\_Purchase Assets\_Ordered Price\_Limit Subtotal Fees  
## <fct> <fct> <dbl> <dbl> <dbl> <dbl>  
## 1 133 DOGE 727243. 0.000001 0.730 0.00109   
## 2 133 DOGE 657149. 0.00000131 0.861 0.00129   
## 3 133 SC 599101. 0.000004 2.40 0.00359   
## 4 133 SC 599101. 0.005 3037. 4.56   
## 5 133 DOGE 598937. 0.0000016 0.958 0.00144   
## 6 133 DOGE 564488. 0.00000141 0.796 0.00119   
## 7 133 DOGE 564488. 0.0000014 0.790 0.00119   
## 8 133 DOGE 475747. 0.0000013 0.618 0.000928  
## 9 133 DOGE 470890. 0.00000178 0.838 0.00126   
## 10 133 DOGE 468666. 0.00000119 0.553 0.000830  
## # ... with 69 more rows

data\_clean %>%  
 filter(UserID == 1) %>%  
 arrange(desc(Assets\_Ordered)) %>%  
 select(-c(TXNID, FSA, Province, Order\_Datetime, Market\_Tradeoff, Type, Side, Assets\_Requested))

## # A tibble: 60 x 6  
## UserID Market\_Purchase Assets\_Ordered Price\_Limit Subtotal Fees  
## <fct> <fct> <dbl> <dbl> <dbl> <dbl>  
## 1 1 STMX 554723. 0.00000011 0.0610 0.0000914  
## 2 1 STMX 554723. 0.00000009 0.0499 0.0000749  
## 3 1 SC 262764. 0.00000021 0.0552 0.0000828  
## 4 1 SC 262764. 0.00000019 0.0499 0.0000748  
## 5 1 SC 249626. 0.00000025 0.0624 0.0000936  
## 6 1 SC 249626. 0.0000002 0.0499 0.0000749  
## 7 1 SC 217066. 0.00000026 0.0564 0.0000846  
## 8 1 SC 217066. 0.00000023 0.0499 0.0000749  
## 9 1 STMX 199999 0.00000026 0.0525 0.0000787  
## 10 1 XVG 110945. 0.0000005 0.0555 0.0000832  
## # ... with 50 more rows

# Assets Ordered: no outliers since users 1 & 133 frequently order large amounts of assets at very low price, resulting in small subtotals.  
  
  
outlier\_rows <- Boxplot(data\_clean$Price\_Limit)



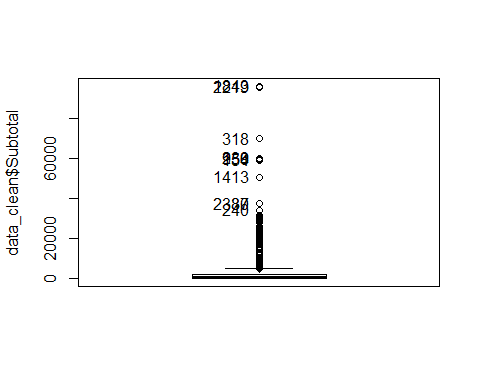
data\_clean[outlier\_rows,]

## # A tibble: 10 x 14  
## TXNID UserID FSA Province Order\_Datetime Market\_Purchase  
## <dbl> <fct> <fct> <fct> <dttm> <fct>   
## 1 748 71 V4W BC 2021-02-11 23:43:19 BTC   
## 2 2310 133 M3C ON 2021-01-02 17:00:31 BTC   
## 3 197 74 V5K BC 2021-02-21 23:31:16 BTC   
## 4 263 134 m1w ON 2021-02-20 17:21:08 BTC   
## 5 262 111 V3M BC 2021-02-20 17:38:08 BTC   
## 6 259 74 V5K BC 2021-02-20 18:11:30 BTC   
## 7 270 29 V2Y BC 2021-02-20 14:36:12 BTC   
## 8 254 74 V5K BC 2021-02-20 19:35:02 BTC   
## 9 255 74 V5K BC 2021-02-20 19:30:55 BTC   
## 10 189 109 V3C BC 2021-02-22 01:35:00 BTC   
## # ... with 8 more variables: Market\_Tradeoff <fct>, Type <fct>, Side <fct>,  
## # Assets\_Ordered <dbl>, Assets\_Requested <dbl>, Price\_Limit <dbl>,  
## # Subtotal <dbl>, Fees <dbl>

data\_clean %>%  
 filter(Market\_Purchase == "BTC") %>%  
 arrange(desc(Price\_Limit)) %>%  
 select(-c(TXNID, FSA, Province, Order\_Datetime, Market\_Tradeoff, Type, Side, Assets\_Requested))

## # A tibble: 539 x 6  
## UserID Market\_Purchase Assets\_Ordered Price\_Limit Subtotal Fees  
## <fct> <fct> <dbl> <dbl> <dbl> <dbl>  
## 1 71 BTC 0.00192 520026 118. 0.177   
## 2 133 BTC 0.008 400000 336. 0.504   
## 3 74 BTC 0.04 72615. 2939. 4.36   
## 4 134 BTC 0.028 72165. 2045. 3.03   
## 5 111 BTC 0.0274 72104. 1996. 2.96   
## 6 74 BTC 0.0275 72000 2004. 2.97   
## 7 29 BTC 0.0274 71931. 1993. 2.95   
## 8 74 BTC 0.0126 71930 917. 1.36   
## 9 74 BTC 0.0275 71900 2001. 2.97   
## 10 109 BTC 0.000337 71787. 24.5 0.0363  
## # ... with 529 more rows

# Price Limit: no outliers since BTC is frequently purchased at high price, and price limit can be set higher than market price by users  
  
  
outlier\_rows <- Boxplot(data\_clean$Subtotal)



data\_clean[outlier\_rows,]

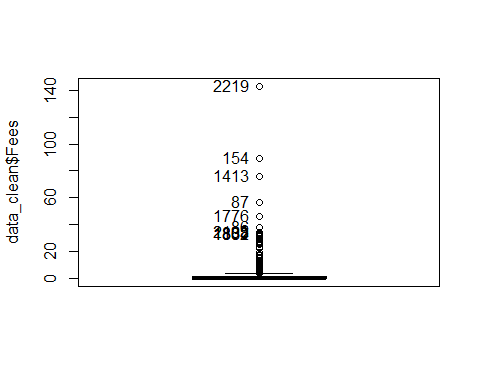
## # A tibble: 10 x 14  
## TXNID UserID FSA Province Order\_Datetime Market\_Purchase  
## <dbl> <fct> <fct> <fct> <dttm> <fct>   
## 1 1843 99 V4N BC 2021-01-20 03:21:49 BTC   
## 2 2219 99 V4N BC 2021-01-06 02:50:20 BTC   
## 3 318 245 v3b BC 2021-02-19 16:42:59 BTC   
## 4 239 99 V4N BC 2021-02-21 02:35:54 BTC   
## 5 956 99 V4N BC 2021-02-09 02:58:22 BTC   
## 6 154 175 T8W AB 2021-02-22 14:41:39 BTC   
## 7 1413 90 M3H ON 2021-01-30 04:44:39 HBAR   
## 8 87 90 M3H ON 2021-02-23 05:10:55 HBAR   
## 9 2330 245 v3b BC 2021-01-01 09:17:53 BTC   
## 10 240 245 v3b BC 2021-02-21 02:14:09 BTC   
## # ... with 8 more variables: Market\_Tradeoff <fct>, Type <fct>, Side <fct>,  
## # Assets\_Ordered <dbl>, Assets\_Requested <dbl>, Price\_Limit <dbl>,  
## # Subtotal <dbl>, Fees <dbl>

subtotal\_outliers <- data\_clean[outlier\_rows,]  
outlier\_users <- subtotal\_outliers$UserID %>%  
 droplevels %>%  
 levels %>%  
 as.numeric  
  
data\_clean %>%  
 filter(UserID %in% outlier\_users) %>%  
 group\_by(UserID) %>%  
 summarize(median\_Subtotal = median(Subtotal))

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 4 x 2  
## UserID median\_Subtotal  
## <fct> <dbl>  
## 1 90 4805.  
## 2 99 59481.  
## 3 175 58888.  
## 4 245 18901.

# Subtotal: no outliers since outlier\_users frequently purchased or sold large sums of assets, and because most of the transactions involved BTC  
  
  
outlier\_rows <- Boxplot(data\_clean$Fees)



fees\_outliers <- data\_clean[outlier\_rows,]  
fees\_outliers %>%  
 select(Subtotal) %>%  
 range

## [1] 21754.03 95239.72

# Fees: no outliers since large fees were accompanied by large subtotals  
  
  
# Which user had the most diverse portfolio?  
data\_clean %>%  
 group\_by(UserID) %>%  
 summarize(unique\_Assets = n\_distinct(Market\_Purchase)) %>%  
 arrange(desc(unique\_Assets))

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 251 x 2  
## UserID unique\_Assets  
## <fct> <int>  
## 1 1 14  
## 2 118 12  
## 3 134 12  
## 4 84 10  
## 5 133 10  
## 6 76 9  
## 7 24 8  
## 8 140 8  
## 9 217 8  
## 10 62 7  
## # ... with 241 more rows

# User 1 with 14 unique assets  
  
  
# write\_excel\_csv(data\_clean, "test\_clean.csv")