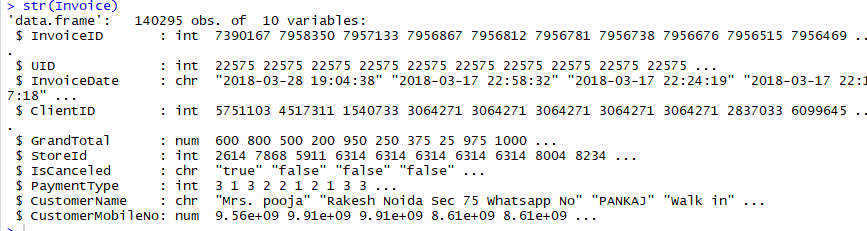
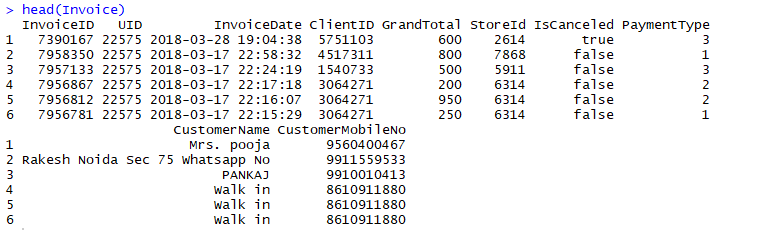
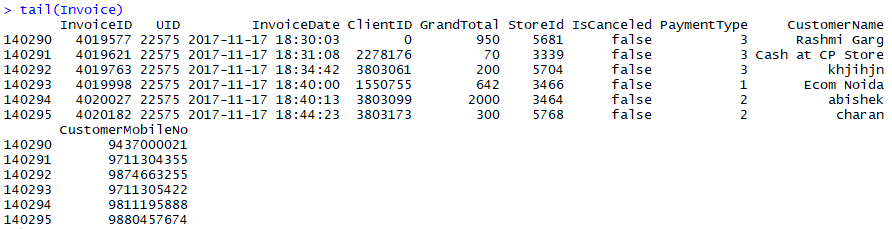
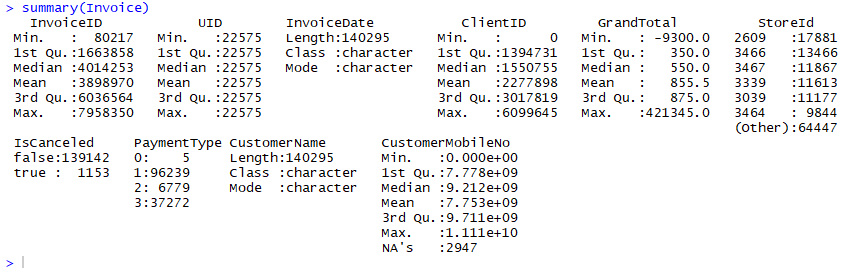
1. Total Number of records:

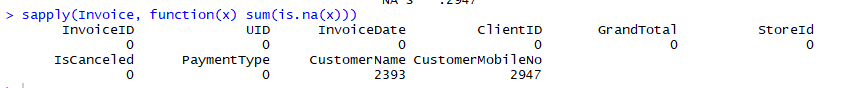






1. After making IsCanceled, PaymentType and StoreId as factor





1. Validity of phone number : Done using below regular expression:

"(?:\\s+|)((0|(?:(\\+|)91))(?:\\s|-)\*(?:(?:\\d(?:\\s|-)\*\\d{9})|(?:\\d{2}(?:\\s|-)\*\\d{8})|(?:\\d{3}(?:\\s|-)\*\\d{7}))|\\d{10})(?:\\s+|)"

##(?:\s+|) // leading spaces

##((0|(?:(\+|)91)) // prefixed 0, 91 or +91

## (?:\s|-)\* // connecting space or dash (-)

## (?:(?:\d(?:\s|-)\*\d{9})| // 1 digit STD code & number with connecting space or dash

## (?:\d{2}(?:\s|-)\*\d{8})| // 2 digit STD code & number with connecting space or dash

## (?:\d{3}(?:\s|-)\*\d{7})| // 3 digit STD code & number with connecting space or dash

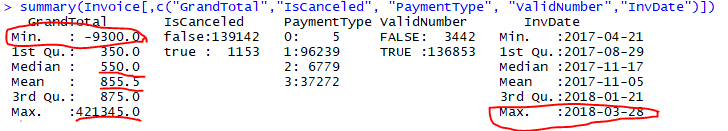
## \d{10}) // plain 10 digit number

## (?:\s+|) // trailing spaces

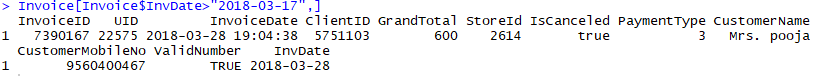
Out of total 32501 unique entries for phone numbers, 32435 seems to correct



1. Create Date field and summarize the data



Out of total invoices, 6779 are already cancelled, so we will remove it from our analysis. Also for 1 invoice, invoice date is coming as 28-Mar-18, which should not be possible so we will not use this.

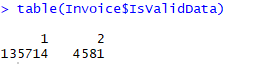


Negative amounts are in two cases:

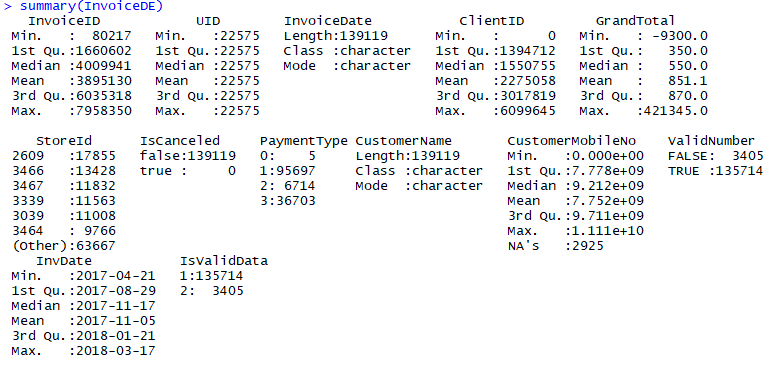
* Where invoice is cancelled: We are not taking the cancelled invoices into consideration.
* In this case reversal of original invoice is done. We will take these invoices for aggregation, but will not consider these for calculating frequency/recency.

Out of total invoices, few are having GrandTotal as “0”, we will remove these are these will not be used in study

Out of total transactions, 4581 are not valid for Study.

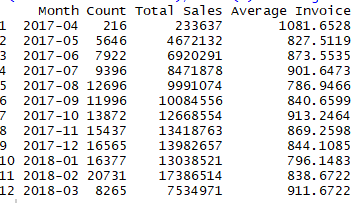


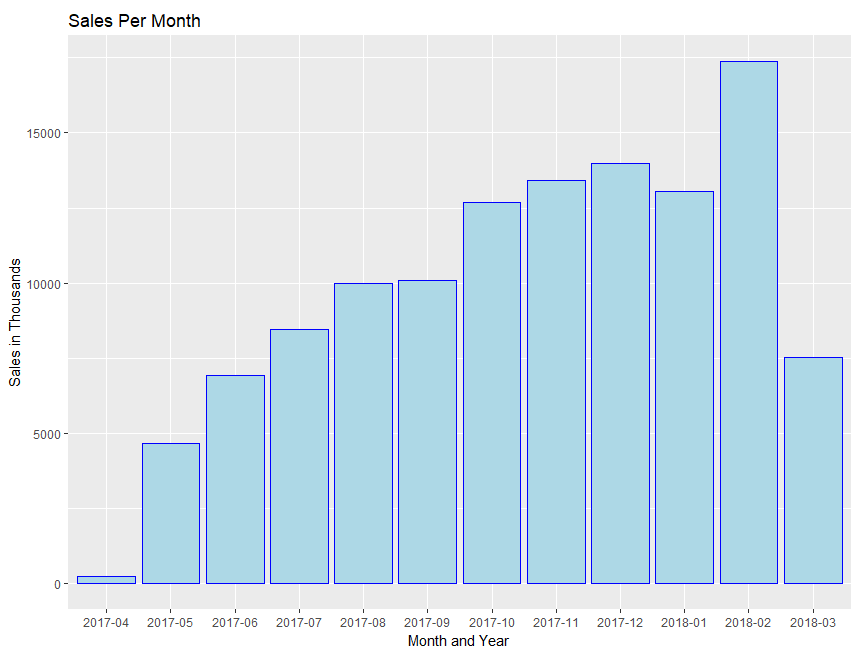
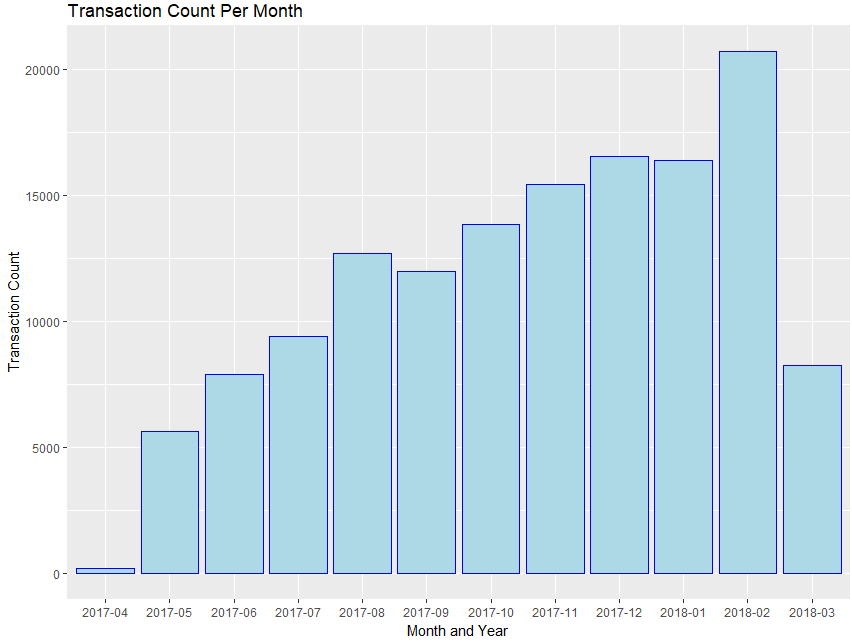
Removed cancelled ,future dated and 0 invoices

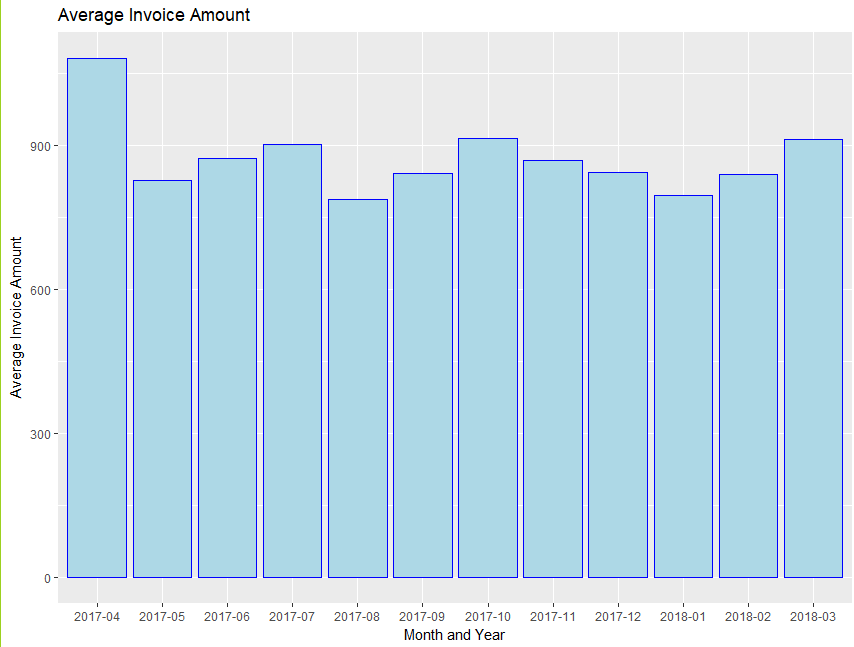


Now InvalidData is only because of invalid numbers.

1. Sales

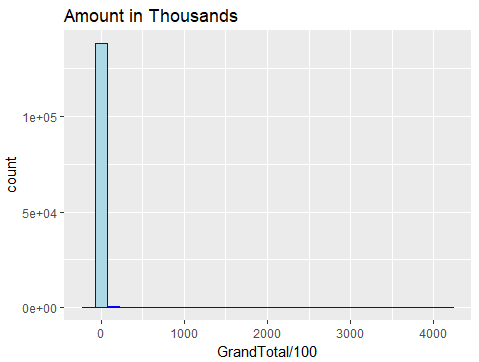




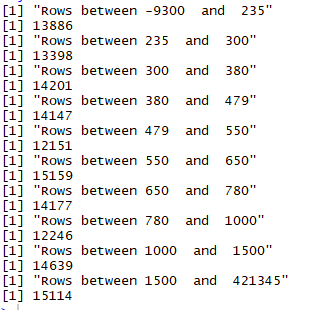


1. Data Exploration on Invoice Amount

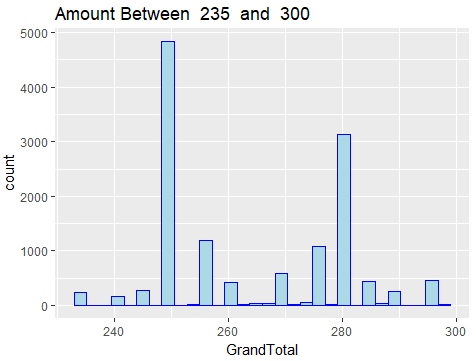
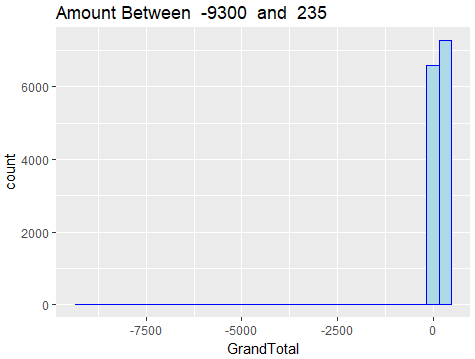


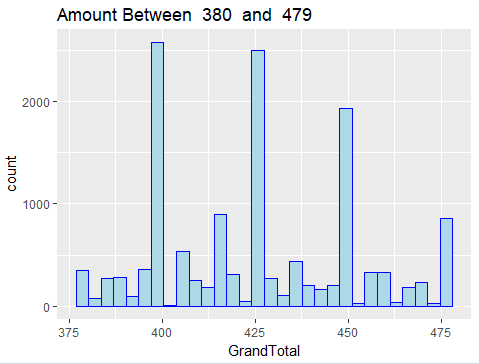
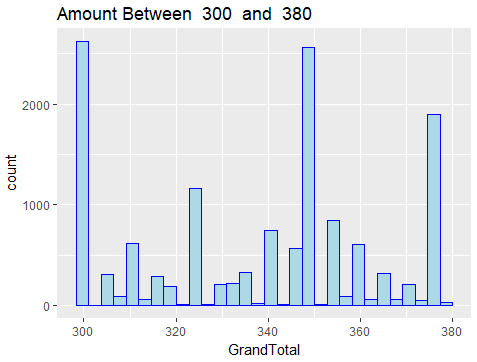


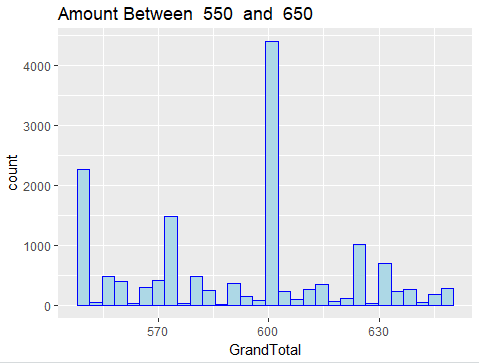
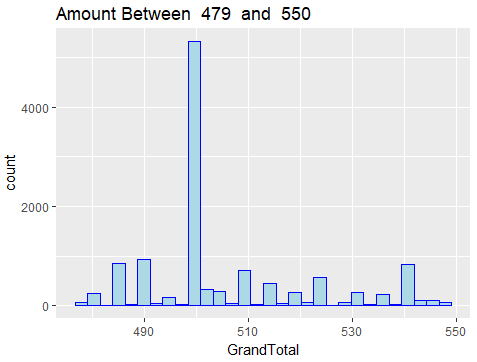
Variance is very high, though most of invoices are of low amounts

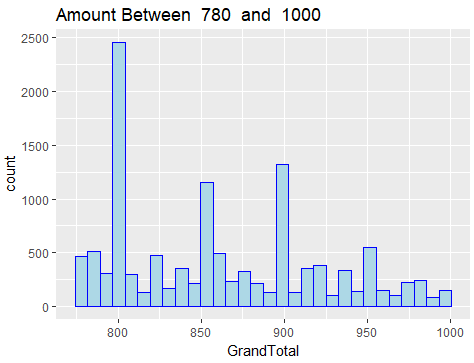
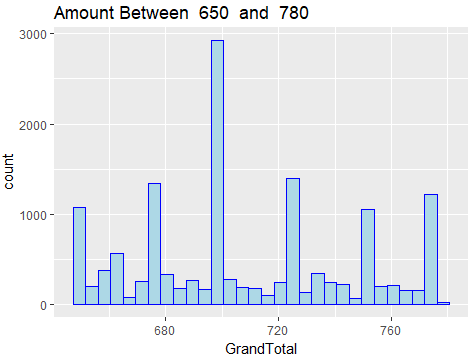


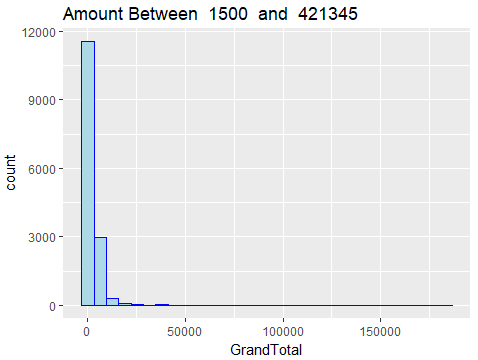
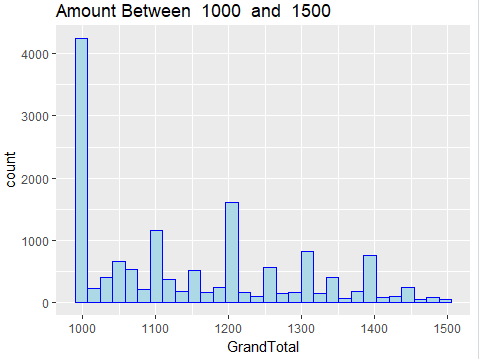
90% records are having invoice amount as less than 1500 and 95% within 2200



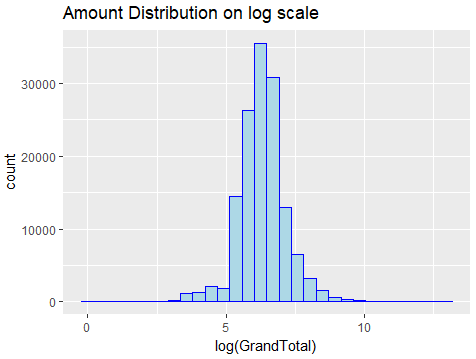






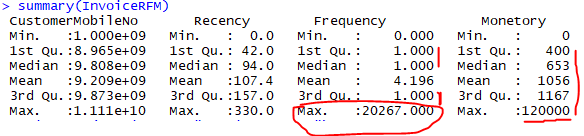


Logarthimic scale:

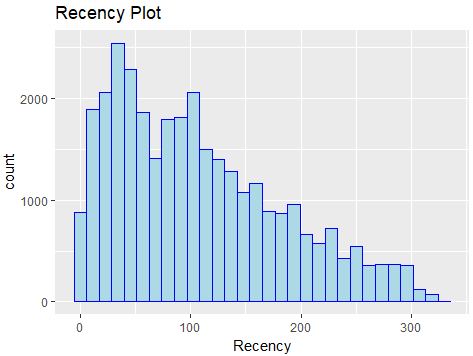


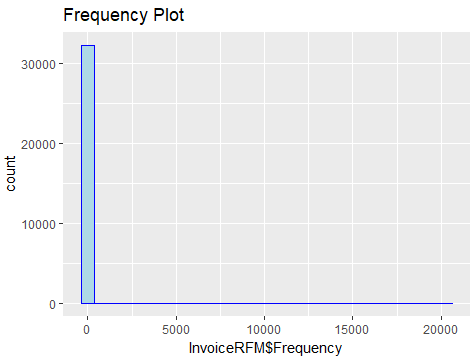
1. Preparing Data for Recency, Frequency and Monitory

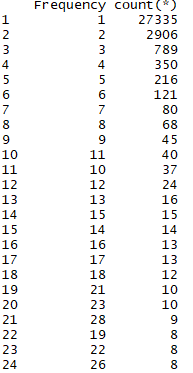
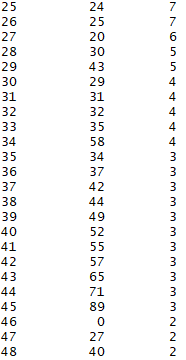
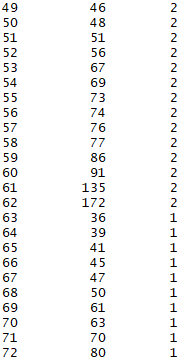


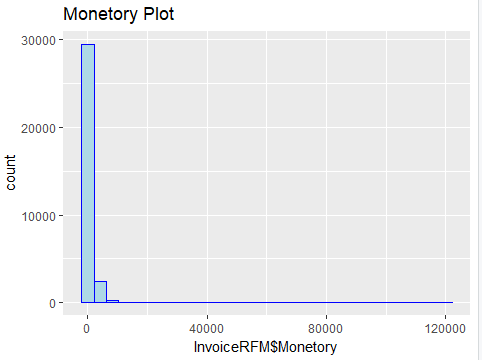


Recency is having good distribution. 75% of population has frequency of 1 only. There is spread in average invoice amount, but few stands out with amount as big as 1.2 Lac

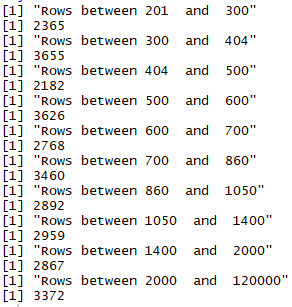
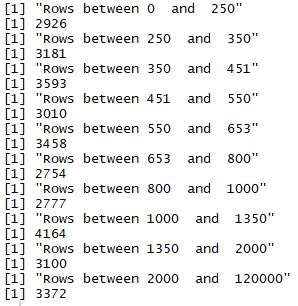


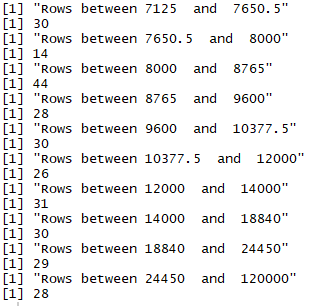
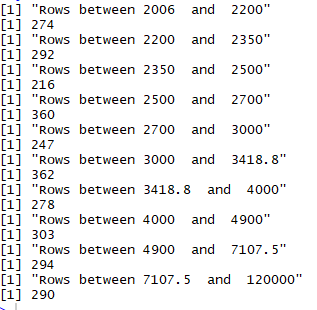










Log Plot

