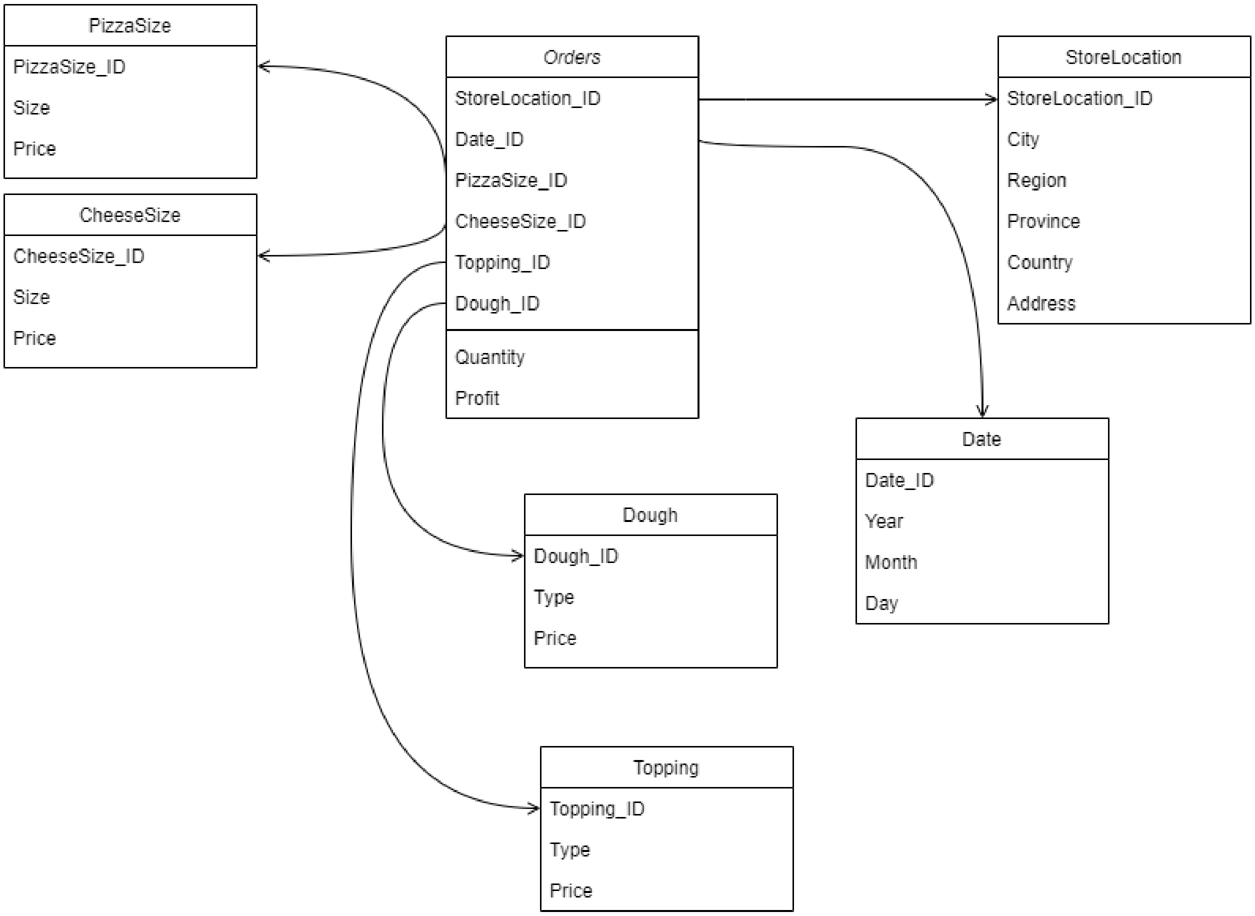
DTI5126[EG]: Fundamentals Data Science

Assignment 1

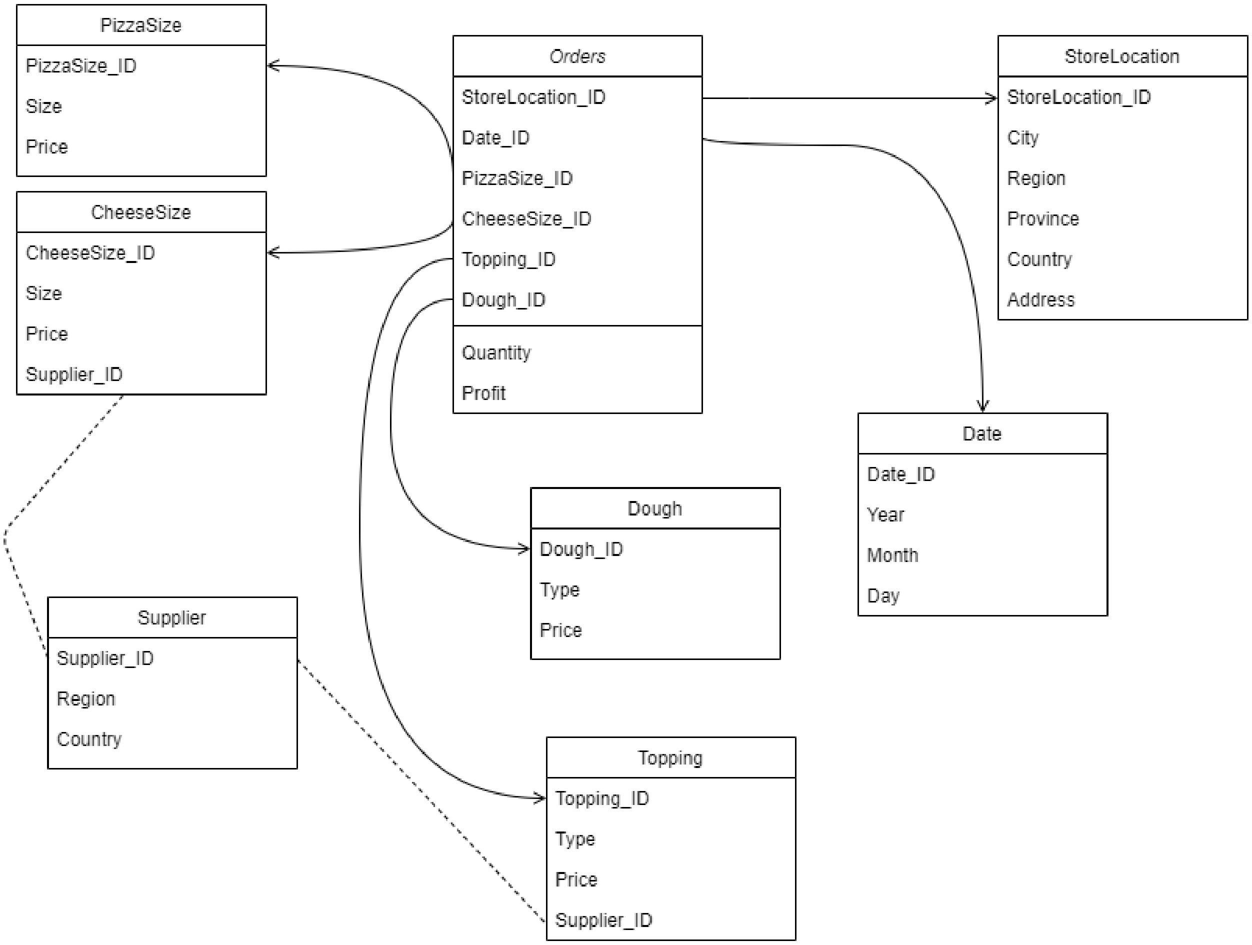
OLAP-Data Preparation

Part A:

1- A- Star schema:



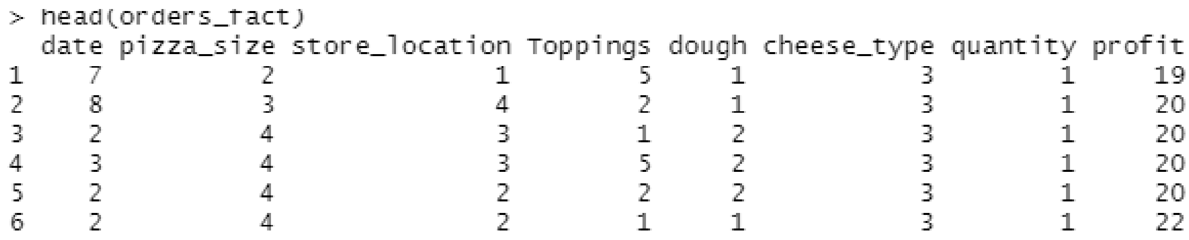
b- SnowFlake schema



Measures: quantity and profit: to get profit→

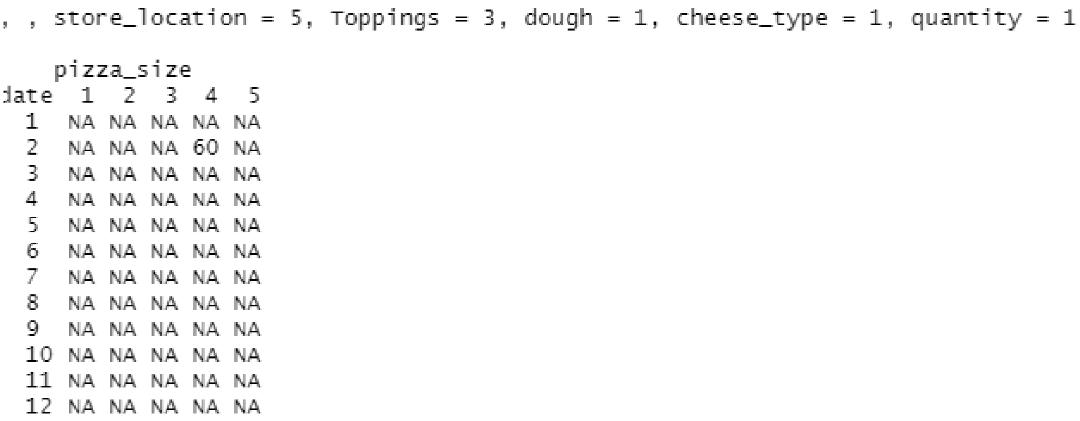
Profit = (price[cheese] + price[topping] + price[size] + price[dough]) * quantity

2- Part of cells of fact table



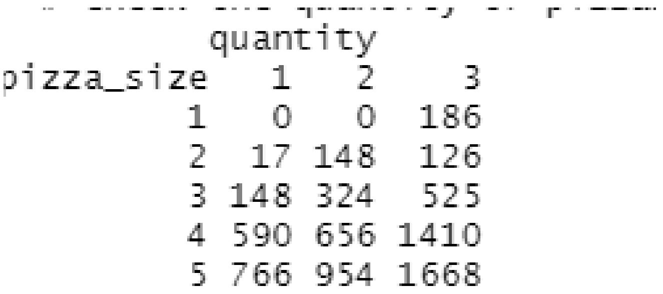
```
> head(orders_fact)
  date pizza_size store_location Toppings dough cheese_type quantity profit
1    7          2              1        5     1           3        1     19
2    8          3              4        2     1           3        1     20
3    2          4              3        1     2           3        1     20
4    3          4              3        5     2           3        1     20
5    2          4              2        2     2           3        1     20
6    2          4              2        1     1           3        1     22
```

Part of cells of cube:



```
, , store_location = 5, Toppings = 3, dough = 1, cheese_type = 1, quantity = 1

    pizza_size
date  1  2  3  4  5
  1  NA NA NA NA NA
  2  NA NA NA 60 NA
  3  NA NA NA NA NA
  4  NA NA NA NA NA
  5  NA NA NA NA NA
  6  NA NA NA NA NA
  7  NA NA NA NA NA
  8  NA NA NA NA NA
  9  NA NA NA NA NA
  10 NA NA NA NA NA
  11 NA NA NA NA NA
  12 NA NA NA NA NA
```

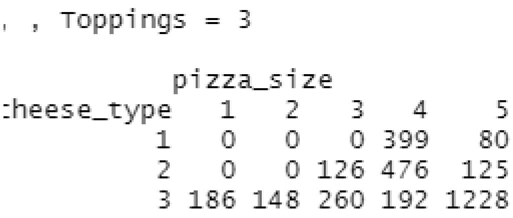
3- Rollup (between pizza_size and quantity)



```
          quantity
pizza_size   1   2    3
         1   0   0  186
         2  17 148  126
         3 148 324  525
         4 590 656 1410
         5 766 954 1668
```

As you can see for large (4) and x_large(5) pizza we've sum of count/profit as over 6000, which implies that people like the bigger pizzas more.

Drill down



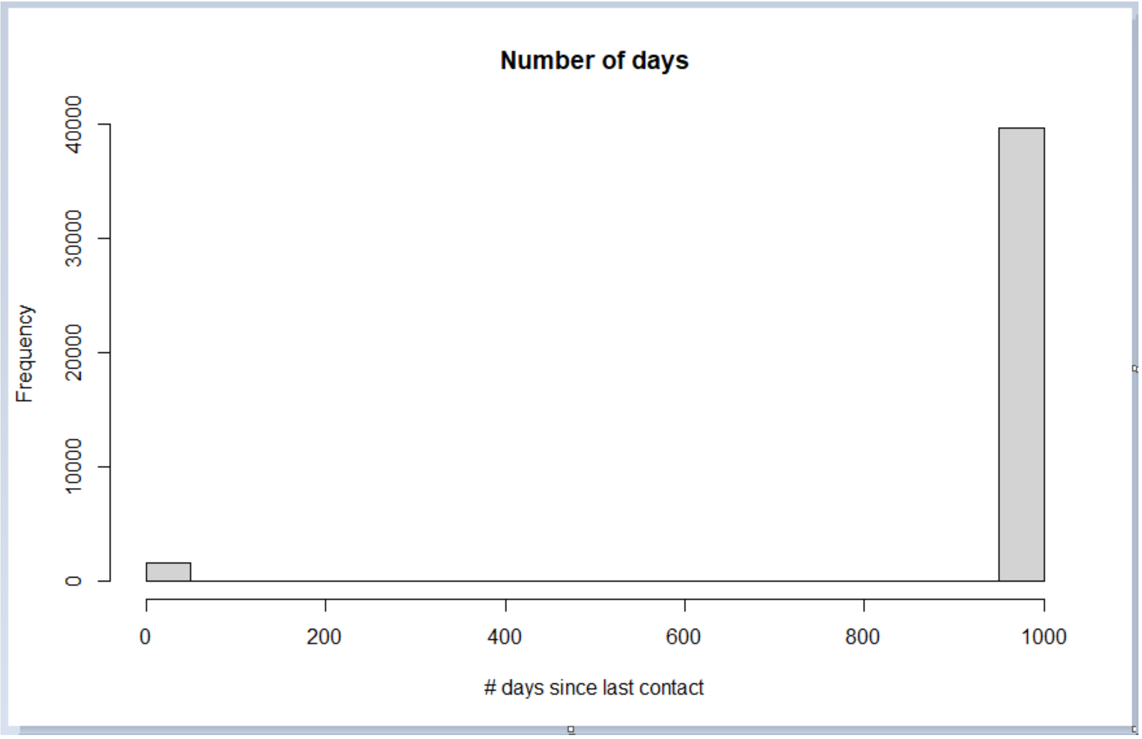
```
, , Toppings = 3

            pizza_size
:heese_type    1   2   3   4    5
          1    0   0   0 399   80
          2    0   0 126 476  125
          3  186 148 260 192 1228
```

Drill down was used to investigate which topping and cheese type do people prefer most, apparently topping number 3 “meat” along with cheese type number 3 “mozeralla” got us the highest profits, meaning that people prefer those types most, so we should buy more of them.

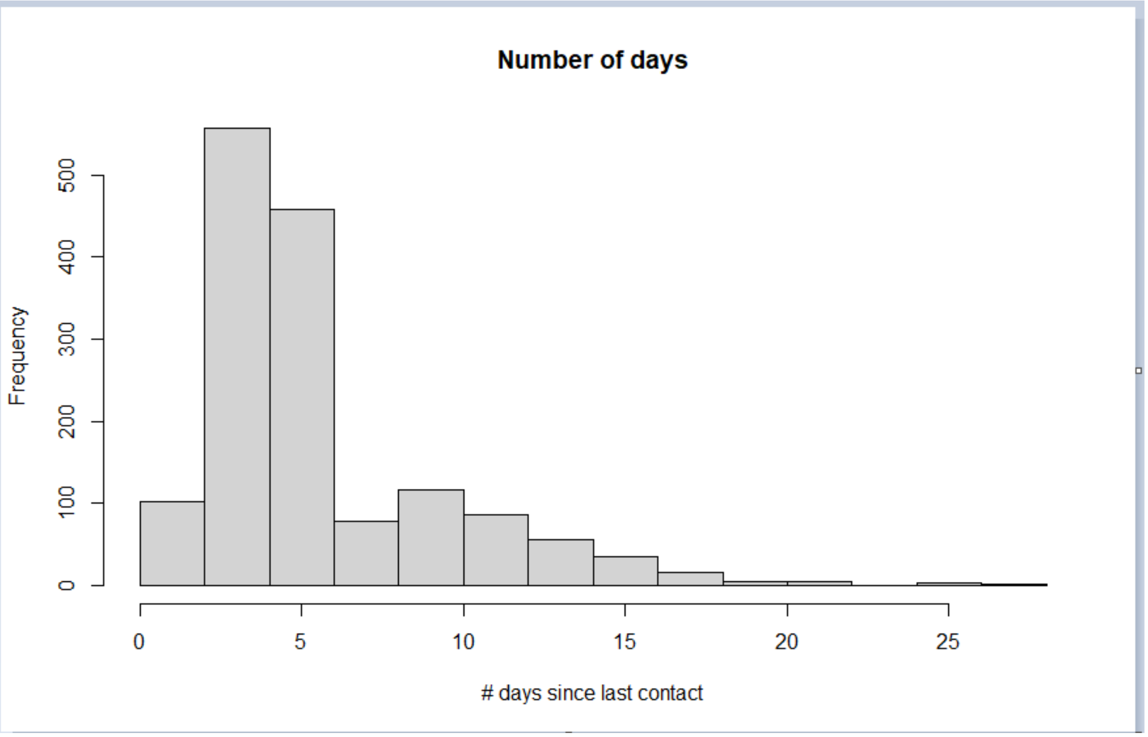
Part B:

3- If we look at the histogram before nulling the inconsistent values, we can see that most of the data is null so we have to deal with it to be able to see the distribution of the valid points that were hidden by the 999 value.

Before:



After:

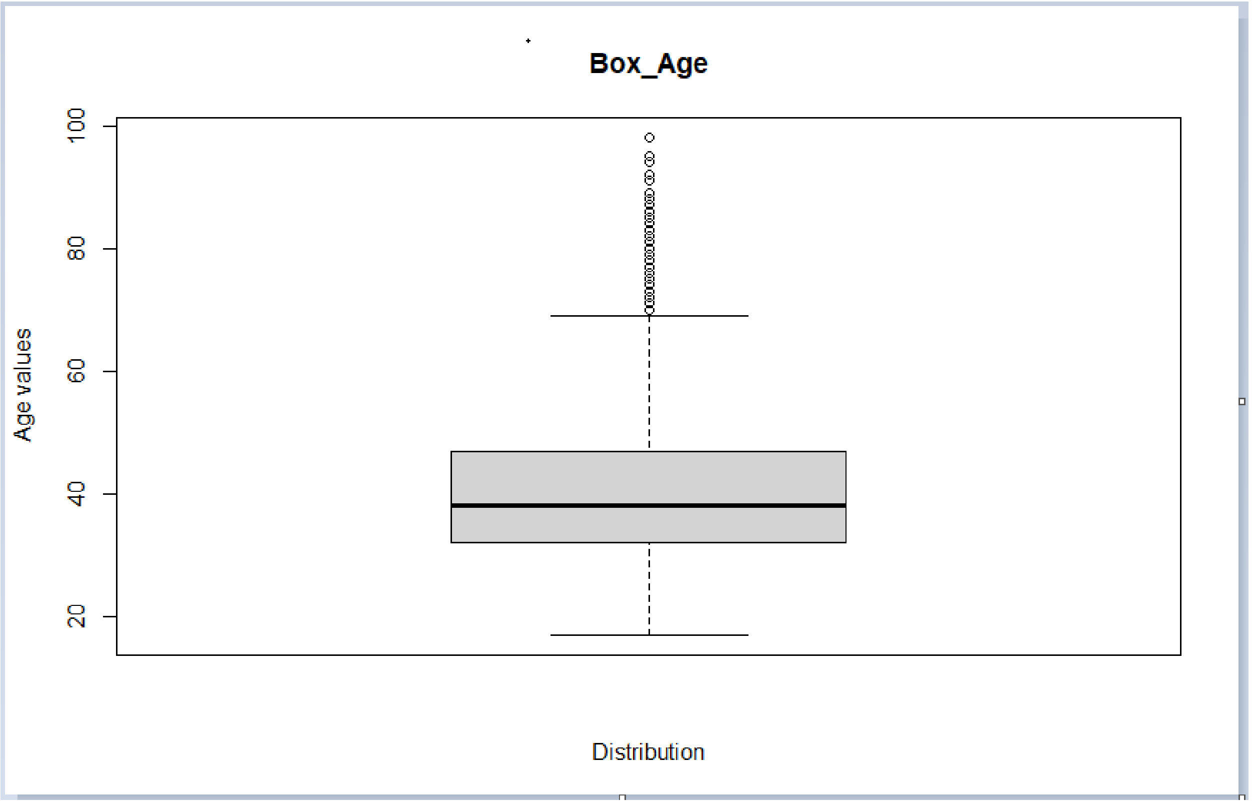


6- Mean: 40.02406

Median: 38

Mode: 31

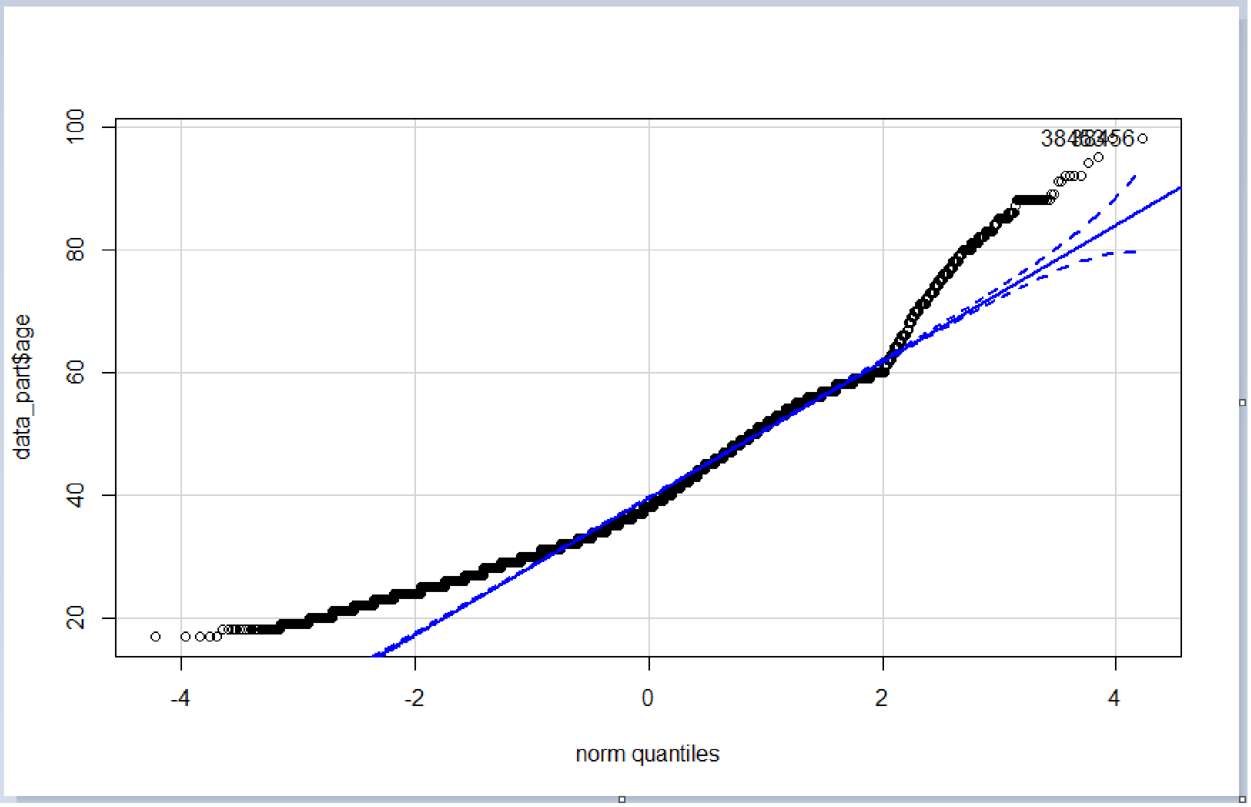
Box_plot



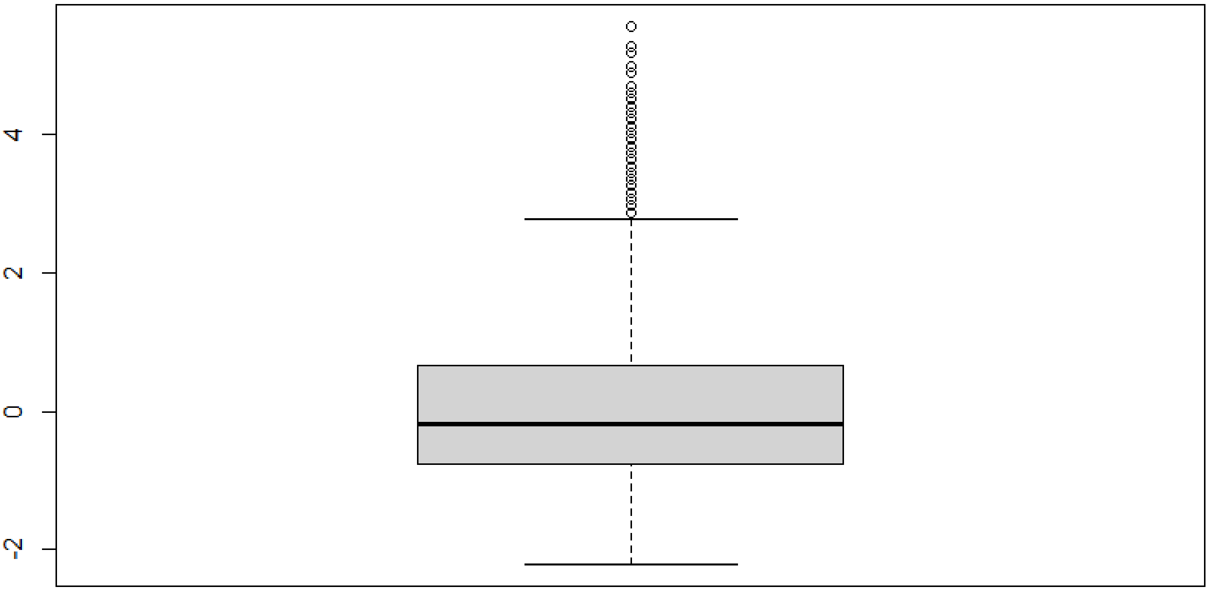
5-number summary:

Min:	17
1st quantile:	32
2nd quantile:	38
3rd quantile:	47
Max:	98

Quantile Info: QQ plot:



Outliers of standardized age:



Above 3 till 5 are considered outliers