Summary Report

General information:

Requirements:

1. Implement all tasks using data. Table R library.
2. For some specific calculations (date transformations, linear regression, charts, etc.) use special R packages (lubridate, ISOweek, gam, ggplot2, etc.).
3. Write clear comments across all functions.
4. Share your intermediate and final results (R script, outputs, charts, reports) via Git repo (you can use GitHub or GitLab account).

Deliverables:

1. Prepare full R Markdown Report/R notebook that will contain a detailed description of Exploratory Data Analysis, Feature Extraction, Modeling, and Post-modeling analysis with conclusions and explanations.

Evaluation:

1. The way you're thinking -
    a. what actions you're implementing
    b. and why.
2. Code style (please follow best practices).
3. Suggestions on the additional analytics/approaches which could be applied.

Structure of the report:

Technical tasks:

Quality Check/ Exploratory Data Analysis of the raw input data

What I planned to do:

1. Check data quality interpretation with data.table library
2. Check NA's in the dataset
3. Better understand data with visualization
4. Think about normalization/removing outliers

Explanatory summary for initial dataset

```
     cat                subcat             date               value        
 Length:75000       Length:75000       Length:75000       Min.   : -1.655  
 Class :character   Class :character   Class :character   1st Qu.:  11.783  
 Mode  :character   Mode  :character   Mode  :character   Median :  25.944  
                                                          Mean   :  53.181  
                                                          3rd Qu.:  54.599  
                                                          Max.   :2247.129  
                                                          NA's   :7         
     volume             units              promo        
 Min.   : -2.3617   Min.   :  -0.0711   Min.   :0.0000  
 1st Qu.:  0.8457   1st Qu.:   3.1322   1st Qu.:0.0000  
 Median :  2.2662   Median :   7.9865   Median :0.0000  
 Mean   :  5.5351   Mean   :  19.9289   Mean   :0.4325  
 3rd Qu.:  5.6876   3rd Qu.:  18.5874   3rd Qu.:1.0000  
 Max.   :325.7443   Max.   :1276.6882   Max.   :1.0000  
 NA's   :1          NA's   :4
```

First of all, I look at the summary of the initial (rows) data set. I saw several NA`s in 'volumn', 'units', and 'value' columns.

Summary for type in the columns

```
Observations: 75,000
Variables: 7
$ cat    <chr> "beauty", "beauty", "beauty", "beauty", "beauty", "beauty", …
$ subcat <chr> "Flowers", "Flowers", "Flowers", "Flowers", "Flowers", "Flow…
$ date   <chr> "W01 17", "W02 17", "W03 17", "W04 17", "W05 17", "W06 17", …
$ value  <dbl> 89.06202, 41.43436, 42.16188, 80.66179, 48.46908, 43.06444, …
$ volume <dbl> 5.659097, 2.364016, 2.405883, 6.149799, 2.771325, 2.451478, …
$ units  <dbl> 26.37573, 11.01813, 11.21326, 28.66278, 12.91650, 11.42577, …
$ promo  <int> 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, …
```

Secondly, I look at how my data is initialized with data.table. It seems that all was done correctly. And the total number of observations is 75 000.

Explanatory summary for dataset without Na's

```
     cat                subcat             date               value        
 Length:74988       Length:74988       Length:74988       Min.   : -1.655  
 Class :character   Class :character   Class :character   1st Qu.:  11.781  
 Mode  :character   Mode  :character   Mode  :character   Median :  25.943  
                                                          Mean   :  53.179  
                                                          3rd Qu.:  54.598  
                                                          Max.   :2247.129  
     volume             units              promo        
 Min.   : -2.3617   Min.   :  -0.0711   Min.   :0.0000  
 1st Qu.:  0.8456   1st Qu.:   3.1317   1st Qu.:0.0000  
 Median :  2.2662   Median :   7.9844   Median :0.0000  
 Mean   :  5.5347   Mean   :  19.9282   Mean   :0.4325  
 3rd Qu.:  5.6875   3rd Qu.:  18.5863   3rd Qu.:1.0000  
 Max.   :325.7443   Max.   :1276.6882   Max.   :1.0000
```

After removing Na`s, it seems that it has not changed the picture at all, but it can be helpful to avoid problems with calculation new columns in the future.

Number of observations after removing Na's = 74988

NA'S Infinity - Infinity

Visualization to better understand data

'cat'

Categorical visualization (will be with grouping at the next step)

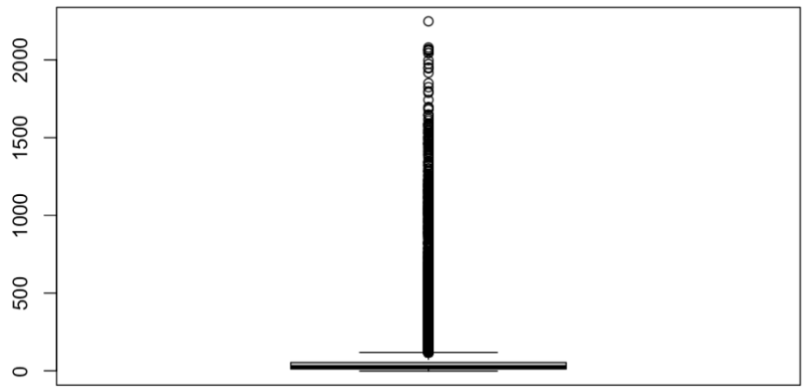
'subcat'

Categorical visualization (will be with grouping at the next step)

'date'

It needs special transformation for making visualization.

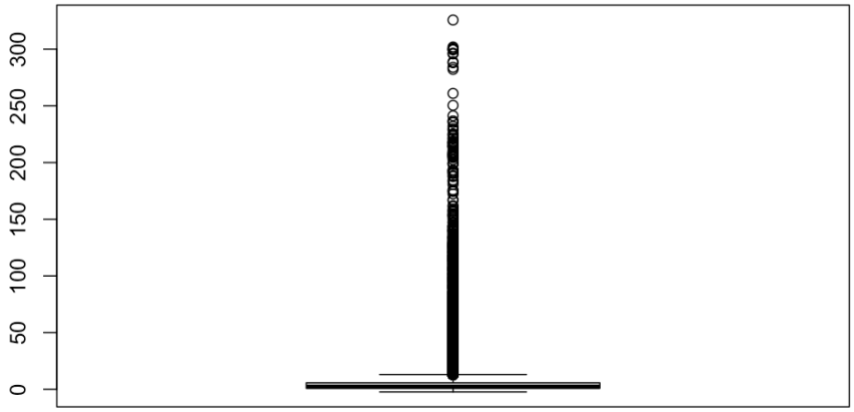
'value' column (boxplot for discovering)



For visualization, I choose boxplot. It is pretty helpful for range analysis.

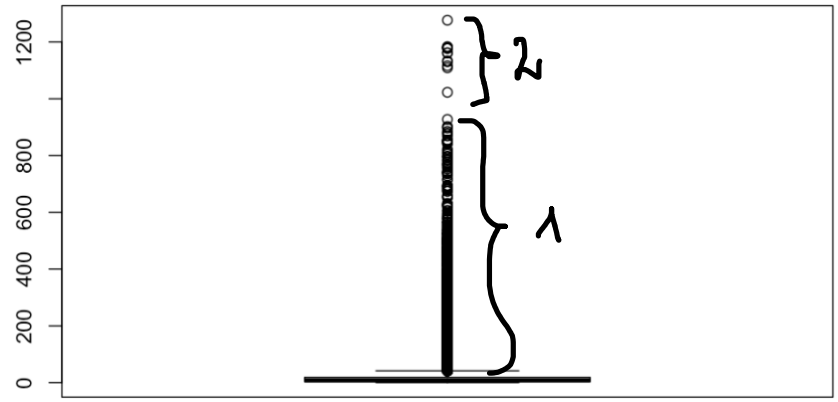
Of course, It is not easy to say about exact median or percentiles, but we see maximum, which can be possible outliers.

'volume' column



The same situation as with 'value' column.

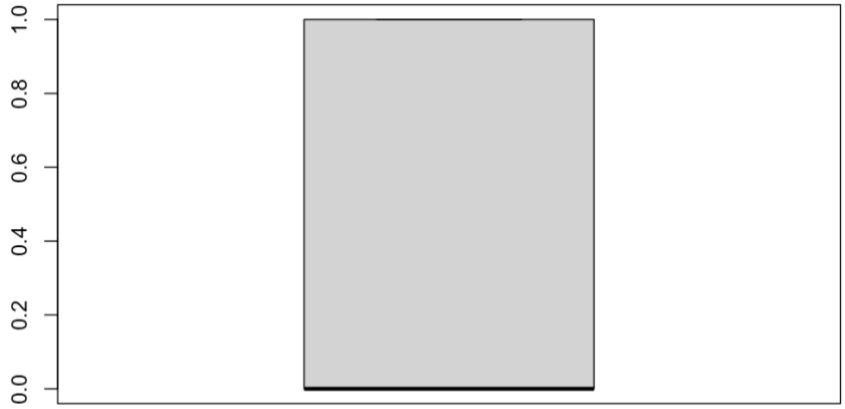
‘units’



The main difference with the previous example is that we see continuous growth to 850 units per volume and break till new growing from 1000 to ~ 1300. It did not tell as much as we wanted since a categorical column makes these clusters logical. But without categorical variables, we can see two separated groups.

Yes, I saw that value, volume, units interpret in decimal, but I decided not to round.

‘promo’ column



For the variable ‘promo’ at this stage,- boxplot is not informative. I would better use a pie chart to see distribution, but I need to make data aggregation.

Only we can conclude that we haven’t negative values.

Delating outliers and normalization/scaling/standardization

About data normalization

In the following steps, I will take the logarithm of some columns, so I decided not to normalize or make standardization of my data.

About outliers

Also, with the same situation as with data normalization, we can make EDA for the upgraded dataset and conclude outliers at that step.

Data Manipulation & Transformation

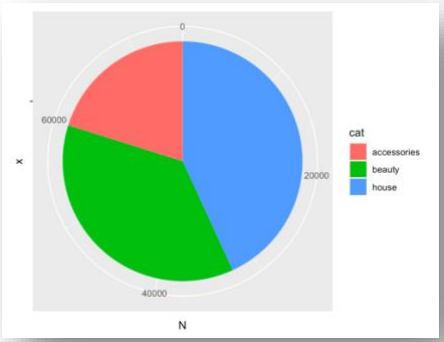
Data Manipulation

What I planned to do:

1. Aggregate data (for better understanding)
2. Visualize aggregated data (the same reason)

‘cat’ column (total three possible values in this column)

cat <chr>	N <int>
house	32356
beauty	27514
accessories	15118



‘subcat’ column (total 248 possible values in this column)

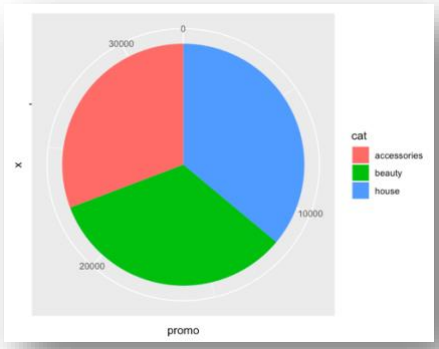
‘cat’ * ‘subcat’ (total 466 possible configuration)

Unfortunately, it is impossible to do a pie chart for subcat and cat*subcat columns distributions because there are too many groups to display.

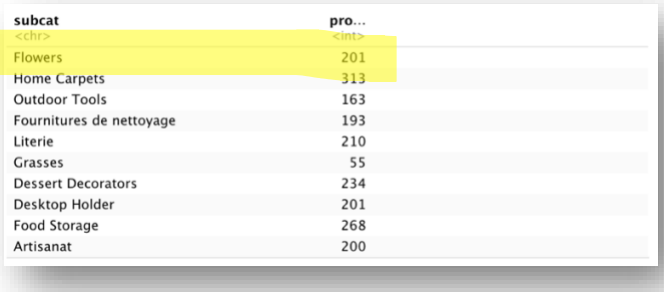
For your recommendations, I aggregate numeric variables by categorical variables and “`promo “. I obtained:

Distribution for a total sum of the promo by categories

cat <chr>	promo <int>
house	11697
beauty	10754
accessories	9984

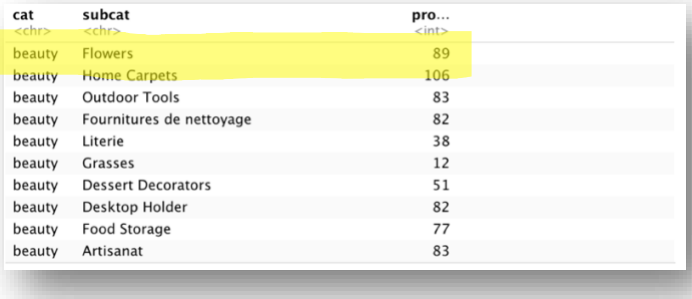


Aggregation for a total sum of the promo by subcategories



subcat	pro...
<chr>	<int>
Flowers	201
Home Carpets	313
Outdoor Tools	163
Fournitures de nettoyage	193
Literie	210
Grasses	55
Dessert Decorators	234
Desktop Holder	201
Food Storage	268
Artisanat	200

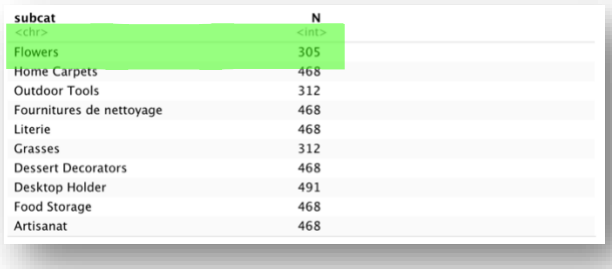
Aggregation for a total sum of the promo by subcategories and Categories



cat	subcat	pro...
<chr>	<chr>	<int>
beauty	Flowers	89
beauty	Home Carpets	106
beauty	Outdoor Tools	83
beauty	Fournitures de nettoyage	82
beauty	Literie	38
beauty	Grasses	12
beauty	Dessert Decorators	51
beauty	Desktop Holder	82
beauty	Food Storage	77
beauty	Artisanat	83

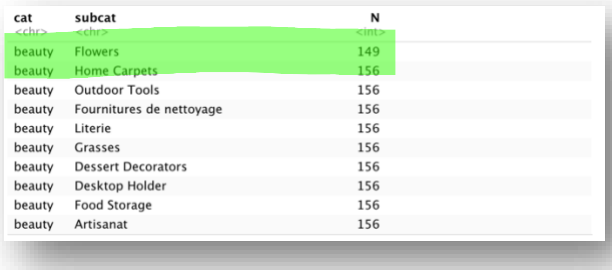
I notice that for cat==’beauty’ and subcat==’Flowers’ sum of the promo is equal 89 at that time only for subcat==’Flowers’ sum of the promo is 201. I was confused, so I decided to look at aggregation for total rows separately for subcat and cat*subcat.

Aggregation for the total number of records by subcategories



subcat	N
<chr>	<int>
Flowers	305
Home Carpets	468
Outdoor Tools	312
Fournitures de nettoyage	468
Literie	468
Grasses	312
Dessert Decorators	468
Desktop Holder	491
Food Storage	468
Artisanat	468

After data manipulation, I understand that subcategories can be in any category. And I think it will be important in feature generating or modeling.

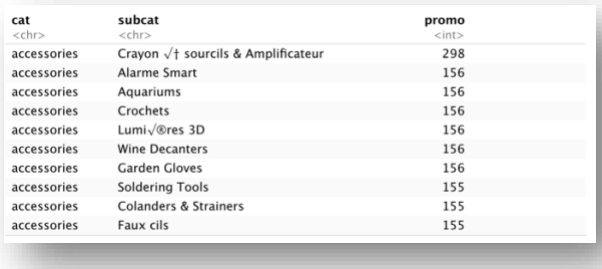


cat	subcat	N
<chr>	<chr>	<int>
beauty	Flowers	149
beauty	Home Carpets	156
beauty	Outdoor Tools	156
beauty	Fournitures de nettoyage	156
beauty	Literie	156
beauty	Grasses	156
beauty	Dessert Decorators	156
beauty	Desktop Holder	156
beauty	Food Storage	156
beauty	Artisanat	156

Ordered aggregation by promo and subcategories to see where was the most promo

subcat	pro...
<chr>	<int>
Crayon √† sourcils & Amplificateur	525
Aquariums	465
Alarme Smart	414
Garden Gloves	403
Soldering Tools	396
Wine Decanters	389
Hammocks	381
Car Seat Cover	373
Crochets	353
Jouets	347

Ordered aggregation by promo, categories, and subcategories to see where was the most promo (to understand data better)

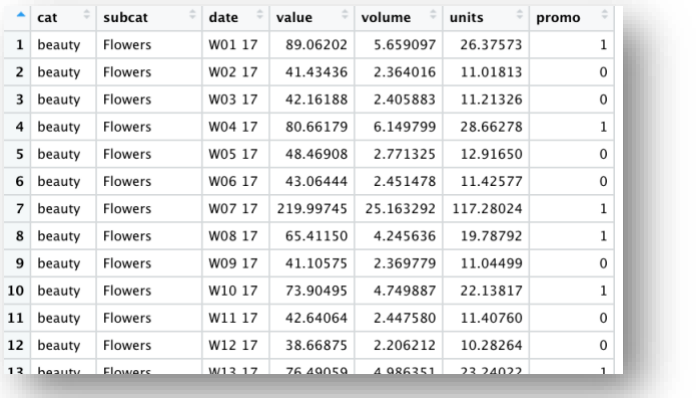


cat	subcat	promo
<chr>	<chr>	<int>
accessories	Crayon √† sourcils & Amplificateur	298
accessories	Alarme Smart	156
accessories	Aquariums	156
accessories	Crochets	156
accessories	Lumi√®res 3D	156
accessories	Wine Decanters	156
accessories	Garden Gloves	156
accessories	Soldering Tools	155
accessories	Colanders & Strainers	155
accessories	Faux cils	155

I analyzed by volume and units, the same as for promo. You can find charts and data tables in r markdown.

Transformation

Dataset preview before first part transformation



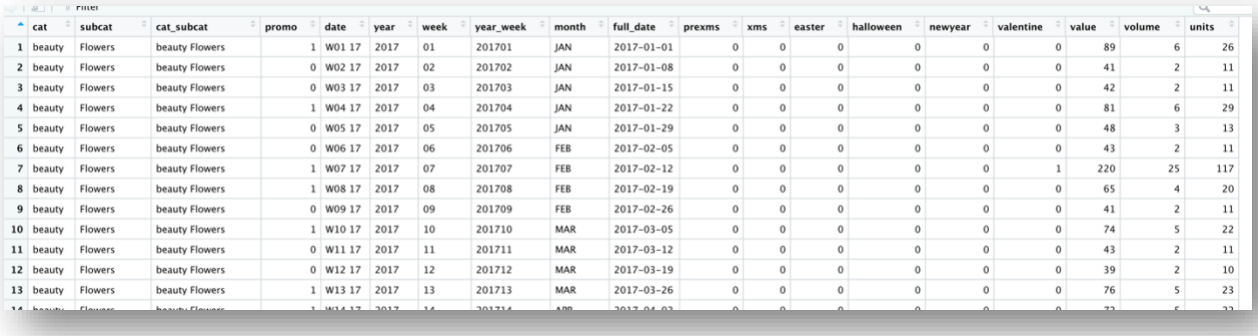
	cat	subcat	date	value	volume	units	promo
1	beauty	Flowers	W01 17	89.06202	5.659097	26.37573	1
2	beauty	Flowers	W02 17	41.43436	2.364016	11.01813	0
3	beauty	Flowers	W03 17	42.16188	2.405883	11.21326	0
4	beauty	Flowers	W04 17	80.66179	6.149799	28.66278	1
5	beauty	Flowers	W05 17	48.46908	2.771325	12.91650	0
6	beauty	Flowers	W06 17	43.06444	2.451478	11.42577	0
7	beauty	Flowers	W07 17	219.99745	25.163292	117.28024	1
8	beauty	Flowers	W08 17	65.41150	4.245636	19.78792	1
9	beauty	Flowers	W09 17	41.10575	2.369779	11.04499	0
10	beauty	Flowers	W10 17	73.90495	4.749887	22.13817	1
11	beauty	Flowers	W11 17	42.64064	2.447580	11.40760	0
12	beauty	Flowers	W12 17	38.66875	2.206212	10.28264	0
13	beauty	Flowers	W13 17	76.49059	4.986351	23.24022	1

What I added (there were some problems, I will take notice):

1. year, week, year_week, month, full_date
   Comment: all these columns are necessary for:
   1.1. Future analysis time series
   1.2. Making new dummy columns from “holidays.csv” based on date.
2. prexms, xms, easter, halloween, newyear, valentine

* During this task, I have a problem with 53 weeks, but I solve it with the ISOweek library; also, I have a problem with initialization data in incorrect time zone, but I solve it by delaminating one day from the current date in a cell.

Dataset preview before second part transformation (because of previous manipulations)



	cat	subcat	cat_subcat	promo	date	year	week	year_week	month	full_date	prexms	xms	easter	halloween	newyear	valentine	value	volume	units
1	beauty	Flowers	beauty Flowers	1	W01 17	2017	01	201701	JAN	2017-01-01	0	0	0	0	0	0	89	6	26
2	beauty	Flowers	beauty Flowers	0	W02 17	2017	02	201702	JAN	2017-01-08	0	0	0	0	0	0	41	2	11
3	beauty	Flowers	beauty Flowers	0	W03 17	2017	03	201703	JAN	2017-01-15	0	0	0	0	0	0	42	2	11
4	beauty	Flowers	beauty Flowers	1	W04 17	2017	04	201704	JAN	2017-01-22	0	0	0	0	0	0	81	6	29
5	beauty	Flowers	beauty Flowers	0	W05 17	2017	05	201705	JAN	2017-01-29	0	0	0	0	0	0	48	3	13
6	beauty	Flowers	beauty Flowers	0	W06 17	2017	06	201706	FEB	2017-02-05	0	0	0	0	0	0	43	2	11
7	beauty	Flowers	beauty Flowers	1	W07 17	2017	07	201707	FEB	2017-02-12	0	0	0	0	0	1	220	25	117
8	beauty	Flowers	beauty Flowers	1	W08 17	2017	08	201708	FEB	2017-02-19	0	0	0	0	0	0	65	4	20
9	beauty	Flowers	beauty Flowers	0	W09 17	2017	09	201709	FEB	2017-02-26	0	0	0	0	0	0	41	2	11
10	beauty	Flowers	beauty Flowers	1	W10 17	2017	10	201710	MAR	2017-03-05	0	0	0	0	0	0	74	5	22
11	beauty	Flowers	beauty Flowers	0	W11 17	2017	11	201711	MAR	2017-03-12	0	0	0	0	0	0	43	2	11
12	beauty	Flowers	beauty Flowers	0	W12 17	2017	12	201712	MAR	2017-03-19	0	0	0	0	0	0	39	2	10
13	beauty	Flowers	beauty Flowers	1	W13 17	2017	13	201713	MAR	2017-03-26	0	0	0	0	0	0	76	5	23

Summary

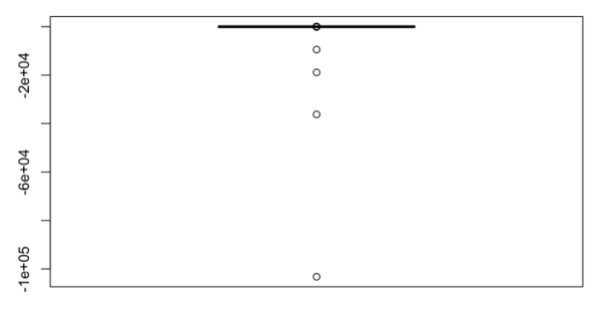
```
     cat                subcat            cat_subcat          promo              date               year          
 Length:74988       Length:74988       Length:74988       Min.   :0.0000   Length:74988       Length:74988      
 Class :character   Class :character   Class :character   1st Qu.:0.0000   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Median :0.0000   Mode  :character   Mode  :character  
                                                          Mean   :0.4325                                        
                                                          3rd Qu.:1.0000                                        
                                                          Max.   :1.0000                                        
     week            year_week            month            full_date             prexms             xms          
 Length:74988       Length:74988       Length:74988       Min.   :2016-11-20   Min.   :0.00000   Min.   :0.00000  
 Class :character   Class :character   Class :character   1st Qu.:2017-08-13   1st Qu.:0.00000   1st Qu.:0.00000  
 Mode  :character   Mode  :character   Mode  :character   Median :2018-05-06   Median :0.00000   Median :0.00000  
                                                          Mean   :2018-05-09   Mean   :0.01922   Mean   :0.01922  
                                                          3rd Qu.:2019-02-03   3rd Qu.:0.00000   3rd Qu.:0.00000  
                                                          Max.   :2019-11-03   Max.   :1.00000   Max.   :1.00000  
     easter          halloween          newyear          valentine           value               volume          
 Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   : -1.655   Min.   : -2.3617  
 1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:  11.781   1st Qu.:  0.8456  
 Median :0.00000   Median :0.00000   Median :0.0000   Median :0.00000   Median :  25.943   Median :  2.2662  
 Mean   :0.01923   Mean   :0.01926   Mean   :0.0192   Mean   :0.01923   Mean   :  53.179   Mean   :  5.5347  
 3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:  54.598   3rd Qu.:  5.6875  
 Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :2247.129   Max.   :325.7443  
      units         
 Min.   :  -0.0711  
 1st Qu.:   3.1317  
 Median :   7.9844  
 Mean   :  19.9282  
 3rd Qu.:  18.5863  
 Max.   :1276.6882
```

* Total rows - 74988

What I added to the previous data table:

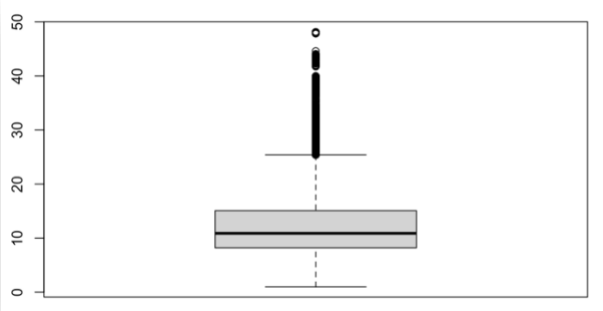
1. ‘price’ column
   It was produced 7703 Na’s.

```
     price          
 Min.   :-103239.68  
 1st Qu.:      8.20  
 Median :     10.88  
 Mean   :      9.95  
 3rd Qu.:     15.07  
 Max.   :     48.13  
 NA's   :7703
```



We suppose price can’t be negative (so I remove value which is less than 0):

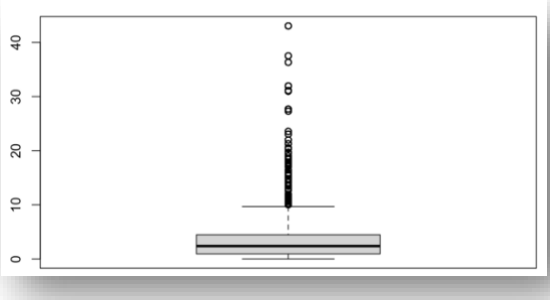
```
     price          
 Min.   : 0.9777  
 1st Qu.: 8.1971  
 Median :10.8803  
 Mean   :12.4540  
 3rd Qu.:15.0735  
 Max.   :48.1279
```



Now it looks much better.

2. ‘price_var’ column

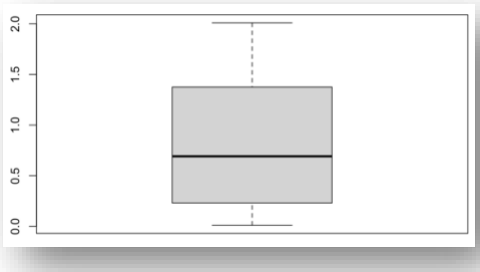
   For me, it was a bit confused to make ‘price_var’ for an unknown sample. So, I decided to sample by category*subcategory because it seems the most logical for me.



   At this point, without NA’a.

   If we delete observations where var > 2 or Na, we will have 28599 observations total. (So, at this point, I decide two save two versions with this cutting (copy) and without(without_var_cuting_copy))

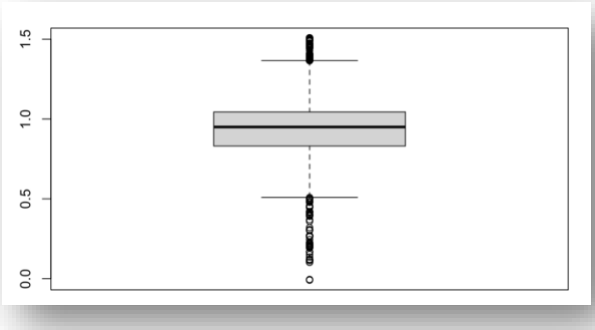
price	price_var
Min. : 0.9777	Min. :0.01021
1st Qu.: 6.7691	1st Qu.:0.22980
Median : 8.9176	Median :0.69174
Mean : 9.4824	Mean :0.85808
3rd Qu.:11.0864	3rd Qu.:1.37558
Max. :32.2901	Max. :2.00920



   * Box plot is for ‘price var’ column

3. ‘log_price’ column

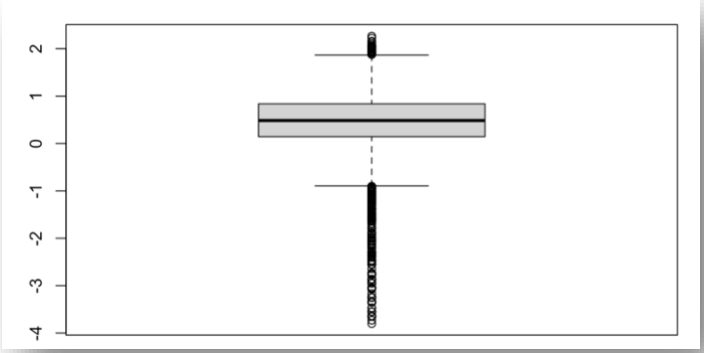
log_price
Min. :-0.009779
1st Qu.: 0.830531
Median : 0.950249
Mean : 0.944263
3rd Qu.: 1.044792
Max. : 1.509070



4. ‘avg_volume’ column

   To this column also arise questions (we will use seasonality in our modeling, but this column does not include seasonality). But anyway, I did it, and in the future, I will check this variable on statistical significance.
5. ‘log_volume’ column

```
  log_volume
Min.   :-3.8008
1st Qu.: 0.1441
Median : 0.4851
Mean   : 0.4731
3rd Qu.: 0.8375
Max.   : 2.2660
NA's   :7
```



   * Removed 7 NA’s

5.5 After made condition week>29, we have 13079 observations (from 28592)

6. ‘cpi’ column

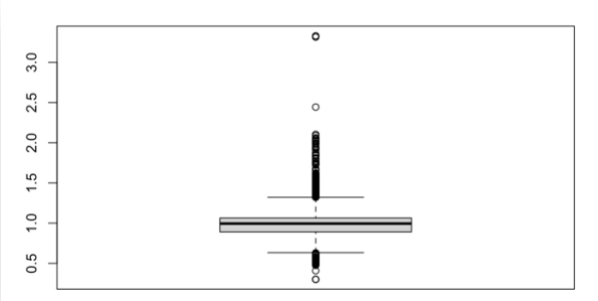
   # Statistical analysis for new columns (‘total_value’ and ‘total_volume’)

```
 total_value
Min.   :  0.0006
1st Qu.: 23.9409
Median : 58.7925
Mean   : 99.4418
3rd Qu.:127.6193
Max.   :821.7723
```

```
 total_volume
Min.   :  0.00023
1st Qu.:  2.73190
Median :  7.62576
Mean   : 12.64162
3rd Qu.: 16.00281
Max.   :200.10653
```

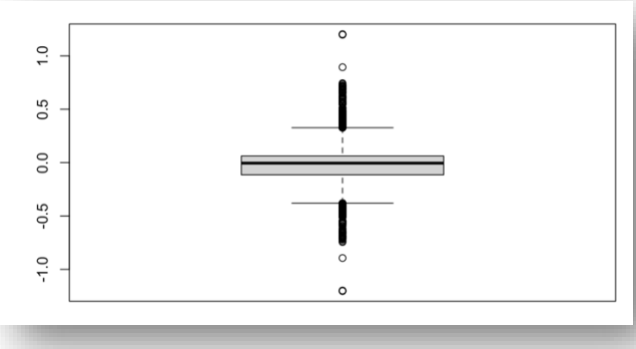
   # Statistical analysis for new column ‘cpi.’

```
      cpi
Min.   :0.300
1st Qu.:0.892
Median :0.995
Mean   :1.002
3rd Qu.:1.064
Max.   :3.329
NA's   :4903
```



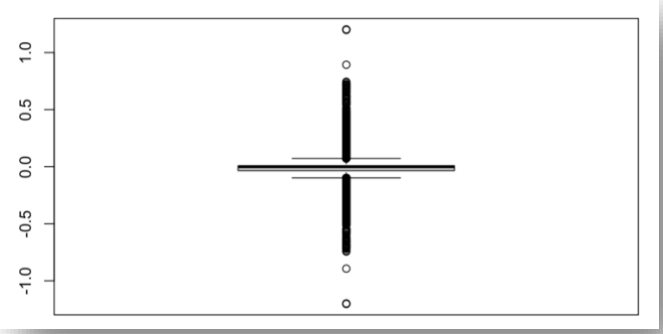
7. ‘log_cpi’column

```
   log_cpi
Min.   :-1.203
1st Qu.:-0.115
Median :-0.005
Mean   :-0.015
3rd Qu.: 0.062
Max.   : 1.203
NA's   :4903
```



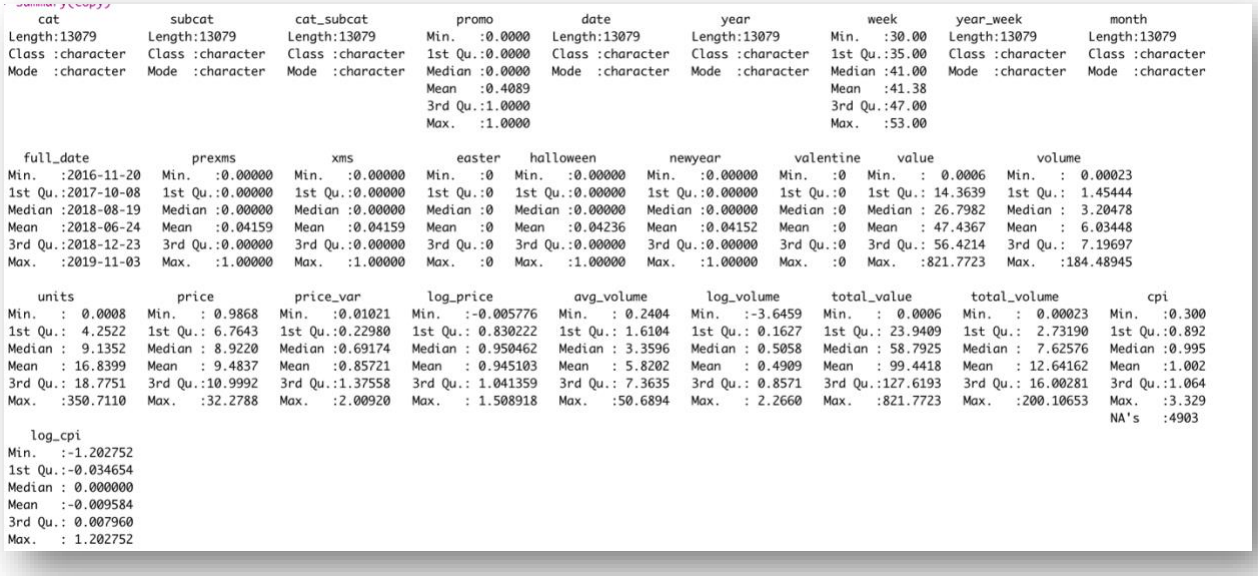
After changing NA’s on 0:

```
   log_cpi
Min.   :-1.202752
1st Qu.:-0.034654
Median : 0.000000
Mean   :-0.009584
3rd Qu.: 0.007960
Max.   : 1.202752
```



Dataset after deformation (second part)

Summary (if we make all conditions)



```
     cat                subcat            cat_subcat          promo              date               year               week          year_week            month
Length:13079       Length:13079       Length:13079       Min.   :0.0000   Length:13079       Length:13079       Min.   :30.00   Length:13079       Length:13079
Class :character   Class :character   Class :character   1st Qu.:0.0000   Class :character   Class :character   1st Qu.:35.00   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Median :0.0000   Mode  :character   Mode  :character   Median :41.00   Mode  :character   Mode  :character
                                                         Mean   :0.4089                                         Mean   :41.38
                                                         3rd Qu.:1.0000                                         3rd Qu.:47.00
                                                         Max.   :1.0000                                         Max.   :53.00

  full_date             prexms             xms              easter     halloween          newyear          valentine      value              volume
Min.   :2016-11-20   Min.   :0.00000   Min.   :0.00000   Min.   :0   Min.   :0.00000   Min.   :0.00000   Min.   :0   Min.   :  0.0006   Min.   :  0.00023
1st Qu.:2017-10-08   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0   1st Qu.: 14.3639   1st Qu.:  1.45444
Median :2018-08-19   Median :0.00000   Median :0.00000   Median :0   Median :0.00000   Median :0.00000   Median :0   Median : 26.7982   Median :  3.20478
Mean   :2018-06-24   Mean   :0.04159   Mean   :0.04159   Mean   :0   Mean   :0.04236   Mean   :0.04152   Mean   :0   Mean   : 47.4367   Mean   :  6.03448
3rd Qu.:2018-12-23   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0   3rd Qu.: 56.4214   3rd Qu.:  7.19697
Max.   :2019-11-03   Max.   :1.00000   Max.   :1.00000   Max.   :0   Max.   :1.00000   Max.   :1.00000   Max.   :0   Max.   :821.7723   Max.   :184.48945

     units              price           price_var         log_price          avg_volume        log_volume       total_value        total_volume           cpi
Min.   :  0.0008   Min.   : 0.9868   Min.   :0.01021   Min.   :-0.005776   Min.   : 0.2404   Min.   :-3.6459   Min.   :  0.0006   Min.   :  0.00023   Min.   :0.300
1st Qu.:  4.2522   1st Qu.: 6.7643   1st Qu.:0.22980   1st Qu.: 0.830222   1st Qu.: 1.6104   1st Qu.: 0.1627   1st Qu.: 23.9409   1st Qu.:  2.73190   1st Qu.:0.892
Median :  9.1352   Median : 8.9220   Median :0.69174   Median : 0.950462   Median : 3.3596   Median : 0.5058   Median : 58.7925   Median :  7.62576   Median :0.995
Mean   : 16.8399   Mean   : 9.4837   Mean   :0.85721   Mean   : 0.945103   Mean   : 5.8202   Mean   : 0.4909   Mean   : 99.4418   Mean   : 12.64162   Mean   :1.002
3rd Qu.: 18.7751   3rd Qu.:10.9992   3rd Qu.:1.37558   3rd Qu.: 1.041359   3rd Qu.: 7.3635   3rd Qu.: 0.8571   3rd Qu.:127.6193   3rd Qu.: 16.00281   3rd Qu.:1.064
Max.   :350.7110   Max.   :32.2788   Max.   :2.00920   Max.   : 1.508918   Max.   :50.6894   Max.   : 2.2660   Max.   :821.7723   Max.   :200.10653   Max.   :3.329
                                                                                                                                                        NA's   :4903

   log_cpi
Min.   :-1.202752
1st Qu.:-0.034654
Median : 0.000000
Mean   :-0.009584
3rd Qu.: 0.007960
Max.   : 1.202752
```

* Total number observations- 13079

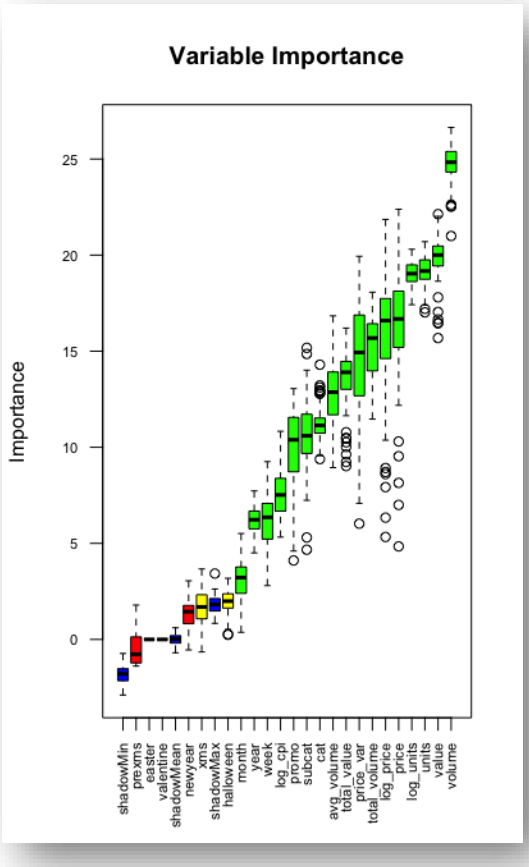
Regression analysis

Features selection

Variables with statistically significant influence on log_volume variables (Boruta library):

```
> print(boruta_signif)
 [1] "cat"          "subcat"       "value"        "volume"       "units"        "log_units"    "promo"        "year"         "week"         "month"
[11] "avg_volume"   "price"        "price_var"    "log_price"    "total_value"  "total_volume" "log_cpi"      "xms"          "halloween"
```

	meanImp	decision
volume	24.793408	Confirmed
value	19.893041	Confirmed
units	19.177472	Confirmed
log_units	19.018194	Confirmed
price	16.305416	Confirmed
log_price	15.848585	Confirmed
total_volume	15.226778	Confirmed
price_var	14.728627	Confirmed
total_value	13.527180	Confirmed
avg_volume	12.813820	Confirmed
cat	11.220164	Confirmed
subcat	10.622604	Confirmed
promo	9.869864	Confirmed
log_cpi	7.539264	Confirmed
year	6.242038	Confirmed
week	6.221052	Confirmed
month	3.058743	Confirmed
halloween	1.915930	Confirmed
xms	1.629961	Confirmed



Modeling

First model

```
lm(formula = log_volume ~ log_units + log_price + cat + subcat +
    promo + log_cpi + year + week + month + xms + halloween,
    data = copy)
```

Comment: I added all variables from feature selection which are highly important for log_price.

Second model

```
lm(formula = log_volume ~ log_units + log_price + cat + subcat +
    cat * subcat + promo + log_cpi + year + week + month + xms +
    halloween, data = copy)
```

Comment: I delete seasonality.

Third model

```
lm(formula = log_volume ~ log_units + log_price + cat + subcat +
    promo + log_cpi + year + week + month + xms + halloween +
    (prexms * xms * newyear), data = copy)
```

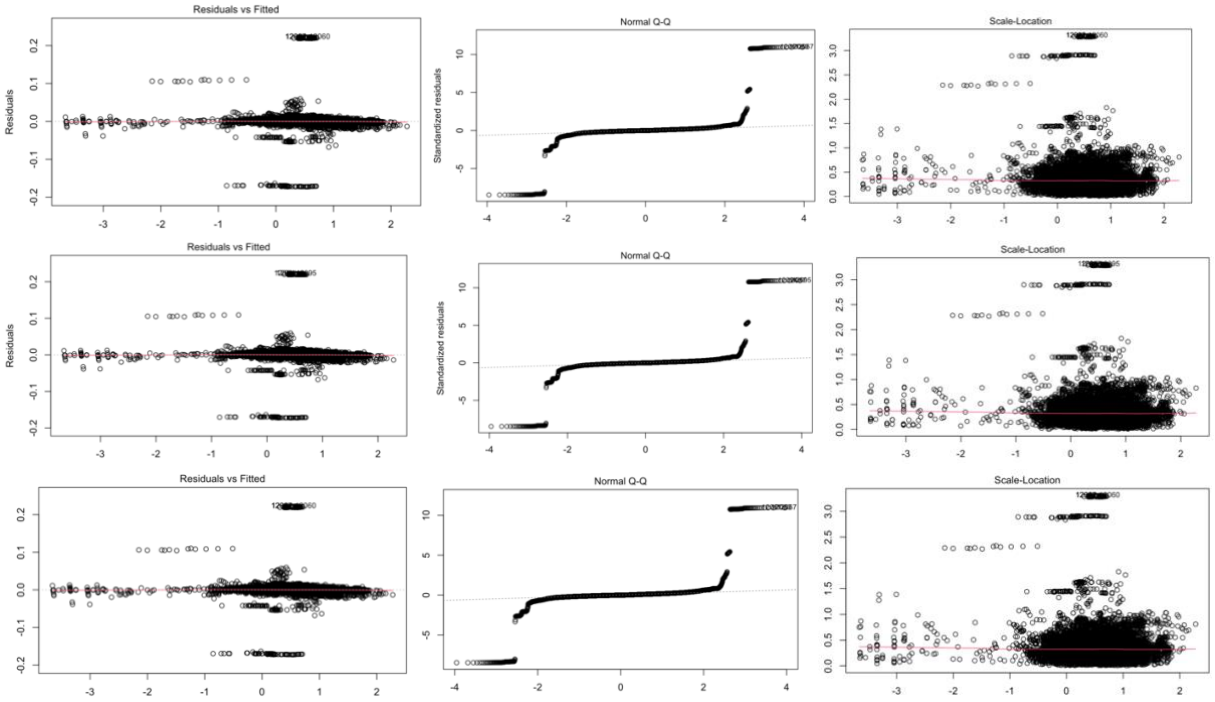
Comment: I added interaction term prexms*xms*newyear.

Results

Metrics

	First model	Second model	Third model
R^2	0.9987	0.9987	0.9987
RSS	5.400473	5.401647	5.400472
TSS	4111.76	4111.76	4111.76
MPE	-0.00123435	-0.00123435	-0.001144638
MAPE	0.05072712	0.05069987	0.05073671

Visualizations (first, second and third accordingly)



All three models are on the same level of quality. R^2 is high, MPE and RSS are small. So, we can take any of these three models.

Prepared by Yaroslav Boiko.