# B2W Labs Pricing Challenge

Miller Horvath

## Data Analysis and Preprocessing

Exploring the data was my first thought since I do not have experience dealing with sales related info. I started organizing the data in the sales.csv file, joining transactions made on the same day. As a result, I got a file with daily quantity of sales, mean price and median price. To perform this preprocessing step, I wrote scripts in python.

The next step was performing a data statistical analysis. I decided to separate the data by product id to see the behavior difference between the products. Using R, I got the summary of the data and some graph plots that helped me to make decisions about how to manipulate the data.

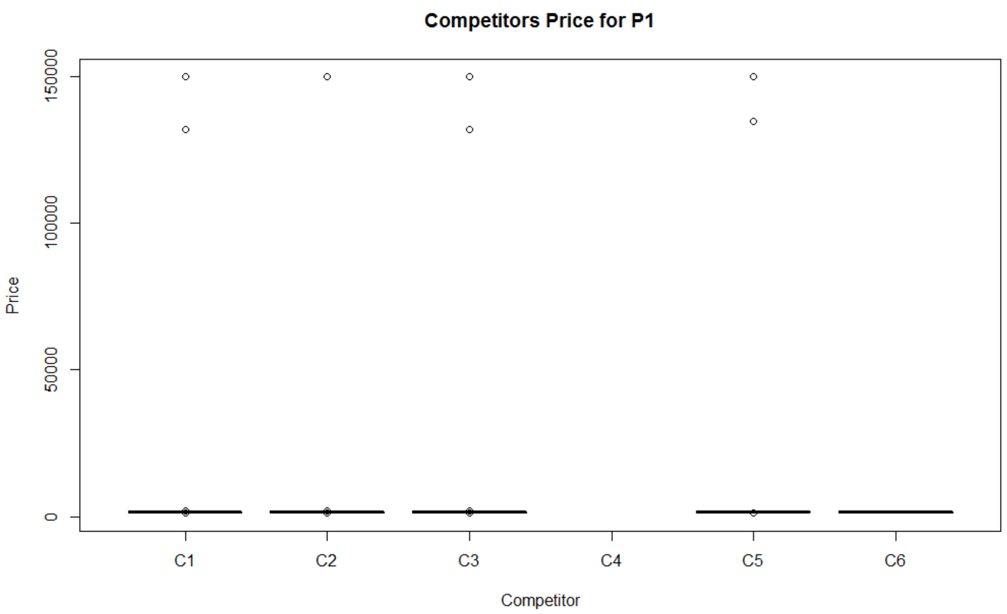
Let's look at the competitors' prices first.

```
> summary(cprices.p1)
    prod_id      comp            date             time           price          pay_type
 P1      :2713   C1:765   2015-04-13:  20   20:10:19: 122   Min.   :  1090   PT1:1358
 P2      :   0   C2:703   2015-04-14:  20   08:11:25: 120   1st Qu.:  1424   PT2:1355
 P3      :   0   C3:396   2015-04-15:  20   08:11:35: 116   Median :  1499
 P4      :   0   C4:  0   2015-04-16:  20   08:11:36: 110   Mean   :  1908
 P5      :   0   C5:713   2015-04-17:  18   20:10:17: 100   3rd Qu.:  1499
 P6      :   0   C6:136   2015-04-22:  18   08:11:27:  96   Max.   :149900
 (Other):   0             (Other)   :2597   (Other) :2049
> summary(cprices.p1)
    prod_id      comp            date             time           price          pay_type
 P1      :2713   C1:765   2015-04-13:  20   20:10:19: 122   Min.   :  1090   PT1:1358
 P2      :   0   C2:703   2015-04-14:  20   08:11:25: 120   1st Qu.:  1424   PT2:1355
 P3      :   0   C3:396   2015-04-15:  20   08:11:35: 116   Median :  1499
 P4      :   0   C4:  0   2015-04-16:  20   08:11:36: 110   Mean   :  1908
 P5      :   0   C5:713   2015-04-17:  18   20:10:17: 100   3rd Qu.:  1499
 P6      :   0   C6:136   2015-04-22:  18   08:11:27:  96   Max.   :149900
 (Other):   0             (Other)   :2597   (Other) :2049
> summary(cprices.p2)
    prod_id      comp              date             time           price             pay_type
 P2      :8755   C1:1122   2015-02-17:  72   08:10:24: 346   Min.   :  506.9   PT1:4379
 P1      :   0   C2:1128   2015-02-16:  46   20:10:07: 324   1st Qu.:  677.9   PT2:4376
 P3      :   0   C3:2089   2015-03-05:  46   20:10:08: 309   Median :  729.5
 P4      :   0   C4:1959   2015-04-24:  46   08:10:23: 300   Mean   :  819.9
 P5      :   0   C5: 494   2015-01-05:  40   08:10:22: 290   3rd Qu.:  799.0
 P6      :   0   C6:1963   2015-01-08:  40   20:10:05: 258   Max.   :79900.0
 (Other):   0              (Other)   :8465   (Other) :6928
> summary(cprices.p3)
    prod_id      comp              date             time           price              pay_type
 P3      :5853   C1:1646   2015-03-16:  40   08:10:24: 218   Min.   :   879.1   PT1:2930
 P1      :   0   C2:1652   2015-03-17:  40   20:10:07: 218   1st Qu.:  1099.0   PT2:2923
 P2      :   0   C3: 835   2015-04-24:  37   20:10:08: 194   Median :  1214.1
 P4      :   0   C4: 639   2015-03-18:  36   08:11:25: 188   Mean   :  1421.6
 P5      :   0   C5: 286   2015-03-19:  36   20:10:19: 188   3rd Qu.:  1312.3
 P6      :   0   C6: 795   2015-03-20:  36   08:10:23: 182   Max.   :119900.0
 (Other):   0              (Other)   :5628   (Other) :4665
> summary(cprices.p4)
    prod_id      comp              date             time           price             pay_type
 P4      :1689   C1:   0   2015-05-19:  16   20:10:08:  77   Min.   :  431.1   PT1:845
 P1      :   0   C2:   0   2015-05-20:  16   20:10:07:  76   1st Qu.:  497.0   PT2:844
 P2      :   0   C3:   0   2015-05-18:  14   08:10:24:  74   Median :  499.9
 P3      :   0   C4:1085   2015-05-21:  14   08:10:23:  64   Mean   :  643.0
 P5      :   0   C5:  16   2015-05-27:  14   08:11:25:  60   3rd Qu.:  569.0
 P6      :   0   C6: 588   2015-06-03:  14   08:11:35:  58   Max.   :49700.0
 (Other):   0              (Other)   :1601   (Other) :1280
> summary(cprices.p5)
    prod_id      comp            date             time           price             pay_type
 P5      :1896   C1:628   2015-07-09:  18   08:11:25:  96   Min.   :  674.1   PT1:948
 P1      :   0   C2:636   2015-05-05:  12   08:11:35:  90   1st Qu.:  809.1   PT2:948
 P2      :   0   C3:632   2015-05-07:  12   08:11:36:  78   Median :  886.5
 P3      :   0   C4:  0   2015-05-08:  12   08:11:27:  76   Mean   : 1142.3
 P4      :   0   C5:  0   2015-05-09:  12   20:10:18:  72   3rd Qu.:  933.2
 P6      :   0   C6:  0   2015-05-11:  12   20:10:19:  72   Max.   :84890.0
 (Other):   0             (Other)   :1818   (Other) :1412
```

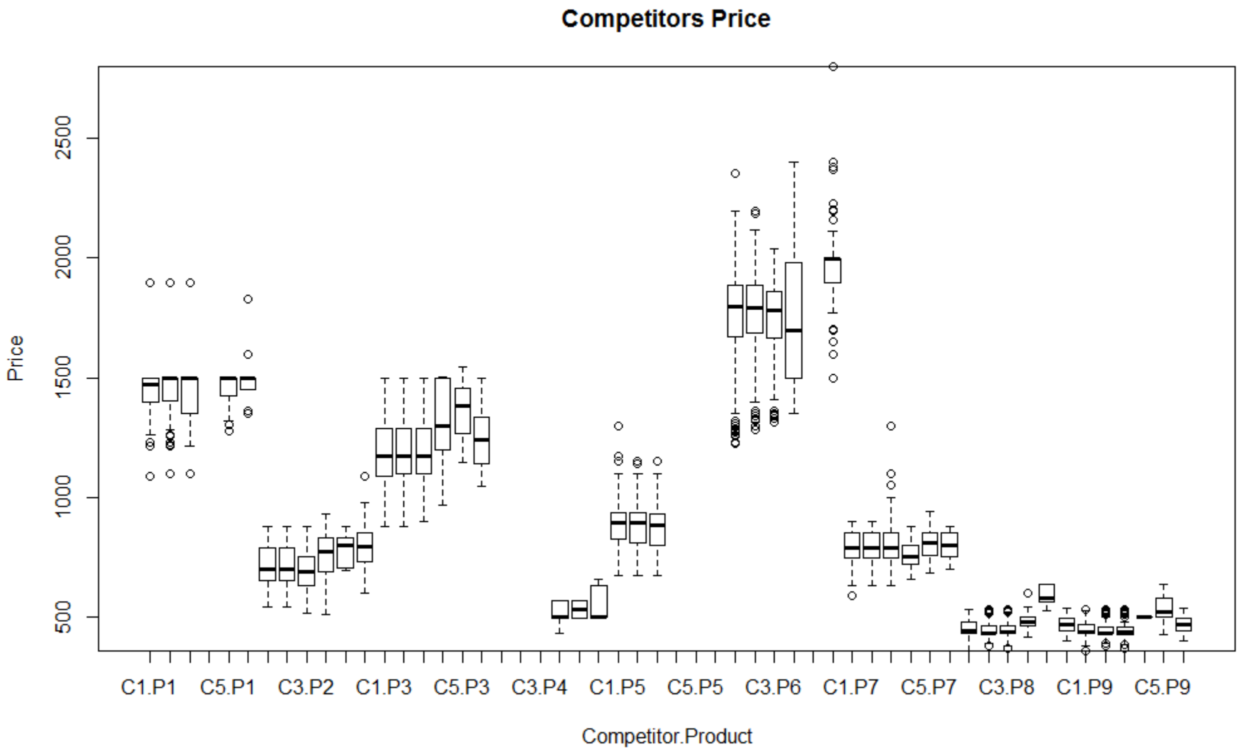
```
> summary(cprices.p6)
    prod_id       comp               date               time             price             pay_type
 P6     :9542   C1:2210   2015-02-17:  80   20:10:05: 504   Min.   :  1226   PT1:4771
 P1     :   0   C2:2142   2015-01-08:  72   08:10:24: 460   1st Qu.:  1674   PT2:4771
 P2     :   0   C3:2166   2015-02-16:  50   08:10:22: 438   Median :  1799
 P3     :   0   C4:1190   2015-09-02:  50   20:10:07: 414   Mean   :  1952
 P4     :   0   C5:   0   2015-09-21:  50   08:10:23: 406   3rd Qu.:  1928
 P5     :   0   C6:1834   2015-01-05:  48   20:10:08: 406   Max.   :149900
 (Other):   0             (Other)   :9192   (Other) :6914
> summary(cprices.p7)
    prod_id       comp               date               time             price               pay_type
 P7     :7748   C1: 991   2015-01-08:  48   08:11:25: 226   Min.   :   588.7   PT1:3877
 P1     :   0   C2: 979   2015-02-17:  40   20:10:19: 212   1st Qu.:   745.0   PT2:3871
 P2     :   0   C3:1077   2015-01-05:  36   20:10:14: 206   Median :   788.0
 P3     :   0   C4:2249   2015-03-05:  36   08:10:24: 200   Mean   :   893.5
 P4     :   0   C5:1549   2015-02-16:  34   08:11:27: 196   3rd Qu.:   849.0
 P5     :   0   C6: 903   2015-03-08:  34   08:11:35: 190   Max.   :104900.0
 (Other):   0             (Other)   :7520   (Other) :6518
> summary(cprices.p8)
    prod_id       comp               date               time             price              pay_type
 P8     :5795   C1:1253   2015-09-02:  45   20:10:07: 408   Min.   :  359.1   PT1:2900
 P1     :   0   C2:1263   2015-05-15:  42   20:10:08: 379   1st Qu.:  431.1   PT2:2895
 P2     :   0   C3:1253   2015-05-05:  40   08:10:24: 374   Median :  448.0
 P3     :   0   C4: 863   2015-05-14:  40   08:10:23: 318   Mean   :  509.5
 P4     :   0   C5:  14   2015-05-16:  40   20:10:06: 154   3rd Qu.:  479.0
 P5     :   0   C6:1149   2015-05-17:  40   08:10:22: 148   Max.   :39999.0
 (Other):   0             (Other)   :5548   (Other) :4014
> summary(cprices.p9)
    prod_id       comp               date               time             price              pay_type
 P9     :6123   C1:1253   2015-09-21:  50   20:10:07: 432   Min.   :  359.1   PT1:3064
 P1     :   0   C2:1267   2015-09-02:  45   20:10:08: 401   1st Qu.:  431.1   PT2:3059
 P2     :   0   C3:1247   2015-05-18:  40   08:10:24: 400   Median :  448.2
 P3     :   0   C4:   4   2015-05-19:  40   08:10:23: 336   Mean   :  533.7
 P4     :   0   C5:1215   2015-05-20:  40   20:10:06: 162   3rd Qu.:  496.0
 P5     :   0   C6:1137   2015-05-21:  40   08:10:22: 160   Max.   :56900.0
 (Other):   0             (Other)   :5868   (Other) :4232
```

The data summary points to some important details. Looking to the competitors, there are some that seems less important than others for each product. For instance, analyzing the product P1, there is no pricing record of competitor C4 for this product. Furthermore, competitors C3 and C6 have a lot less prices recorded than the rest of competitors but C4. All the products were analyzed similarly.

Now looking to number of monitored records, some products have more records than others, which might be relevant to the prediction quality. Finally, the max price of all the products is huge. I used boxplots to visualize this extreme values. The following figure represents the boxplot of product P1. The boxplots of all the products look similar as all of them has extreme values.

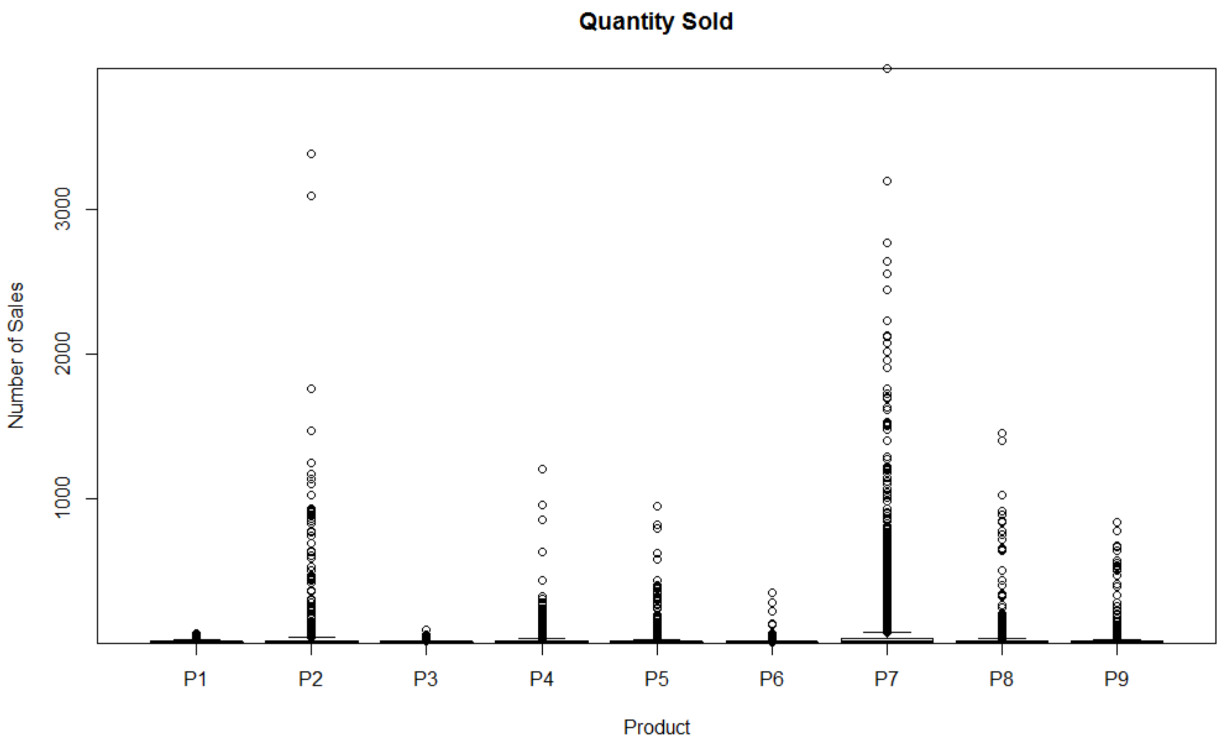



This boxplot shows that there are some price records a lot higher than the regular distribution of prices. Looking to the data I could observe that the record of all competitors were wrong in some specific days. It may be caused by a human error or a problem in the system accountable for monitoring the competitors' prices. Since they are wrong records, I removed them from the data. The following figure represents the boxplot analyzing the competitors' prices after removing the wrong values.




This plot shows us the difference in the price distribution between the products. For example, the product P6 has higher prices and also more variation of prices than the others; the products P8 and P9 have similar distribution of prices and, in both of them, competitor C5 practices higher prices than the others. We can see that we still have outliers in our data, represented by the small circles above and below the boxes. I decided to keep these outliers as they are because I believe they may represent aggressive market strategies adopted by the competitors.

The next figure shows the boxplot of the quantity sold per day for all the products.



Dealing with these outliers was one of the biggest challenges that I came across. I did not realize a reasonable way to reduce their impact on the prediction. When I try to make predictions with the outliers, some products presented really high error. The only approach I could use was removing those outliers. I am aware that this is not the best strategy to deal with outliers, but as I said, I did not find out a better way to do so. After removing them, I believe I achieved decent results that will be explained later on.

Finally, I use a python script to create a .csv file combining all the data to create a dataset for prediction. Each line of this file represents a day of sales. The structure of the data and its column description is the following:

| PROD_ID | DATE | SALES_COUNT | MEDIAN_PRICE | MEAN_PRICE | DAY_OF_WEEK | … |
|---|---|---|---|---|---|---|
| P7 | 2015-03-09 | 375 | 849.00 | 847.53 | Monday | … |
| P7 | 2015-03-10 | 422 | 849.00 | 847.27 | Tuesday | … |

| … | C1_PRICE_1 | C1_PRICE_2 | C2_PRICE_1 | C2_PRICE_2 | C3_PRICE_1 | C3_PRICE_2 | … |
|---|---|---|---|---|---|---|---|
| … | 849.00 | 849.00 | 849.00 | 764.10 | 849.00 | 849.00 | … |
| … | 849.00 | 849.00 | 849.00 | 764.10 | 849.00 | 849.00 | … |

| … | C4_PRICE_1 | C4_PRICE_2 | C5_PRICE_1 | C5_PRICE_2 | C6_PRICE_1 | C6_PRICE_2 |
|---|---|---|---|---|---|---|
| … | 849.00 | 785.33 | 849.00 | 806.55 | 849.00 | 849.00 |
| … | 849.00 | 721.65 | 849.00 | 806.55 | 849.00 | 849.00 |

PROD_ID: Product ID. We provide data for 9 different products, P1 to P9;
DATE: Sales Date, under YYYY-MM-DD format;
SALES_COUNT: The quantity sold on a respective DATE;
MEDIAN_PRICE: The median price value of all sales registered on a respective DATE;
MEAN_PRICE: The mean price value of all sales registered on a respective DATE;
DAY_OF_WEEK: Day of week equivalent to a DATE;
C1_PRICE_1: Mean price of all the monitored prices of the competitor C1 for a respective DATE with PAY_TYPE=1;
C2_PRICE_1: Mean price of all the monitored prices of the competitor C2 for a respective DATE with PAY_TYPE=1;
C3_PRICE_1: Mean price of all the monitored prices of the competitor C3 for a respective DATE with PAY_TYPE=1;
C4_PRICE_1: Mean price of all the monitored prices of the competitor C4 for a respective DATE with PAY_TYPE=1;
C5_PRICE_1: Mean price of all the monitored prices of the competitor C5 for a respective DATE with PAY_TYPE=1;
C6_PRICE_1: Mean price of all the monitored prices of the competitor C6 for a respective DATE with PAY_TYPE=1;
C1_PRICE_2: Mean price of all the monitored prices of the competitor C1 for a respective DATE with PAY_TYPE=2;
C2_PRICE_2: Mean price of all the monitored prices of the competitor C2 for a respective DATE with PAY_TYPE=2;
C3_PRICE_2: Mean price of all the monitored prices of the competitor C3 for a respective DATE with PAY_TYPE=2;
C4_PRICE_2: Mean price of all the monitored prices of the competitor C4 for a respective DATE with PAY_TYPE=2;
C5_PRICE_2: Mean price of all the monitored prices of the competitor C5 for a respective DATE with PAY_TYPE=2;

C6_PRICE_2: Mean price of all the monitored prices of the competitor C6 for a respective DATE with PAY_TYPE=2;

When I look to this .csv file I realized that a lot of competitors' price cells were empty, which means that my data has a lot of missing values. One approach that I tried was filling in those missing values signing the last known price of the competitors to the empty cell. It seemed to be a good idea, but the results I got with this data were worse than the result I got with the missing values dataset. Unfortunately, I did not see another plausible way to deal with this missing values.

I also considered combining data using, instead of the actual competitors' price, the difference between my sale price and the competitors' price. I created the dataset but I could not test it in time for this deliverable.

## Prediction Model

To build the prediction model, I decided to use methods that I learnt in a data mining class. In that class, we used decision trees in R to make predictions for categorical variables. For this challenge, as decision trees do not work with numerical prediction, I used regression trees as an alternative.

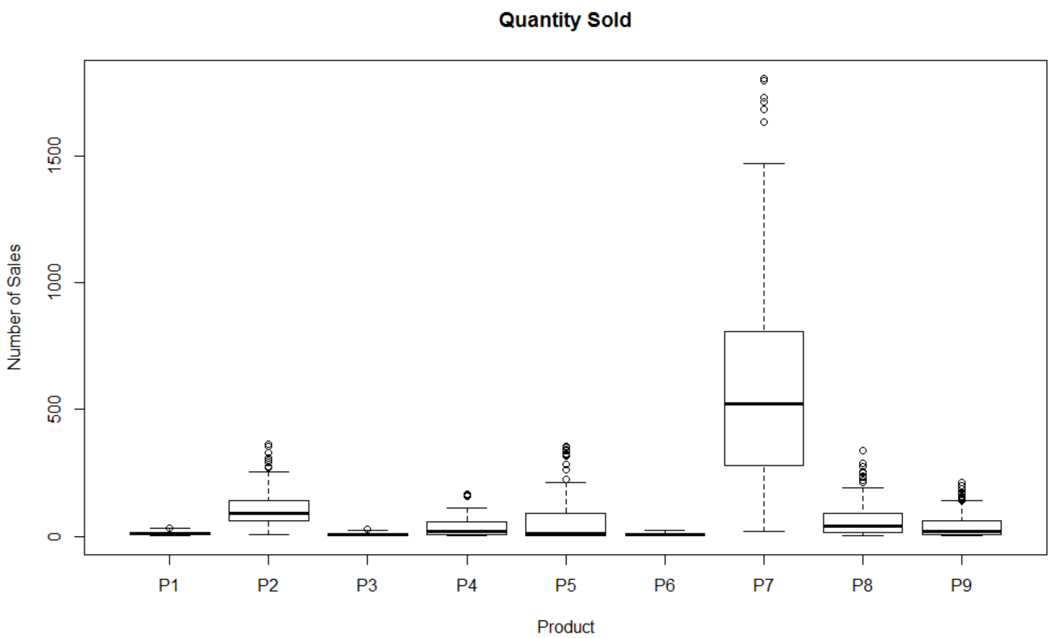
For sampling, I used 80% of the data for training the models and 20% for validation. The metrics I used to evaluate the models were: mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE).

I adopted an empirical process to decide which variables should be used on the model for each product. For instance, I have some similar variables in the dataset, as DATE and DAY_OF_WEEK or MEAN_PRICE and MEDIAN_PRICE. So I just created a few regression trees, removing and adding variables to the dataset, and I picked the one that had the lowest error according to the metrics previously mentioned.

The following table presents the evaluation of the models.

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 |
|---|---|---|---|---|---|---|---|---|---|
| MAE | 5.500733 | 39.37441 | 3.638551 | 9.820261 | 25.4165 | 4.054855 | 239.8189 | 29.53881 | 16.15692 |
| MSE | 48.53849 | 2901.531 | 20.0386 | 183.4174 | 1824.488 | 27.20912 | 122783.4 | 1733.852 | 662.5908 |
| RMSE | 6.966957 | 53.86586 | 4.476449 | 13.54317 | 42.71403 | 5.216236 | 350.4046 | 41.63955 | 25.74084 |

The great difference of the error between the models is highly related to the mean quantity of sales as we can see on the following boxplot:



The prediction for model P7 has the highest error. However, their quantity of sales distribution has higher values as well. Consequently, the difference between the observed and predicted values are more likely to be high, causing the higher error.

I also considered using neural networks to make the predictions, but I also could not implement it in time for this deliverable. One common technique on recommender systems is combining different models to make recommendations. If I had done the neural network model, the next step would be trying to combine the models as a linear weighted model.