|  |
| --- |
|  |

DATA MINING GROUP PROJECT

DATA SCIENCE AND BUSINESS INFORMATICS

ASSESSING DATA QUALITY – OUTLIERS PART

During these papers we are going to detect outliers which are discovered in our dataset during the analysis. Before starting the procedure we have to say that our detection has been conducted over a modified dataset called df\_new, due to the fact that we made data cleaning before. Consequently the variables under the analysis are:

1. **VehOdo**

2. **VNZIP1**

3. **WarrantyCost**

4. **MMR\_factor\_n**

5. **VehBCost\_n**

The outliers analysis will be done using the method of IQR score. Let’s print a table where will be presented many of the usufeful informations.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| VARIABLE | Q1 0,25 | Q2 0,5 | Q3 0,75 | UP-LIMIT | OUTLIERS -UP | DOWN-LIMIT | OUTLIERS –  DOWN |
| **VehOdo** | 61785.000000 | 73359.000000 | 82427.000000 | 113390.0 | 3 | 30822.0 | 260 |
| **VNZIP1** | 32124.000000 | 74135.000000 | 80022.000000 | 151869.0 | 0 | -39723.0 | 0 |
| **WarrantyCost** | 837.000000 | 1155.000000 | 1623.000000 | 4057.263327 | 669 | -2434.105545 | 0 |
| **MMR\_factor\_n** | 0.157782 | 0.211543 | 0.259000 | 0.410827 | 406 | 0.005955000000000016 | 355 |
| **VehBCost\_n** | 0.148805 | 0.183615 | 0.216506 | 0.3180575 | 141 | 0.047253499999999976 | 5 |

Where

Q1 is the 0,25 quantile

Q2 is the 0,5 quantile

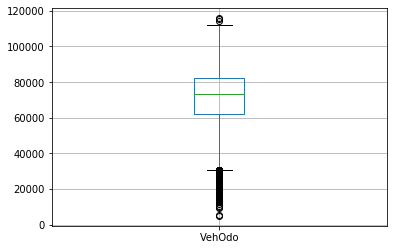
Q3 is the 0,75 quantile

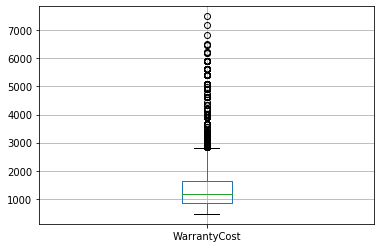
Up-Limit due to IQR score method is the value above which a point Is considerated outliers.

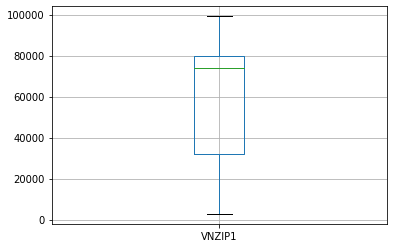
Down-Limit due to IQR score method is the value blow which a point is considerated outliers.

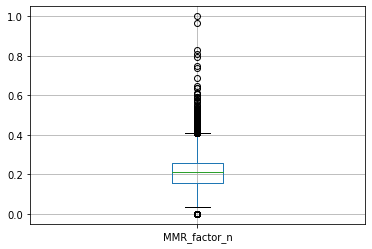
We would put attention on the fact that two variables assume the negative value in contrast to the fact that this value has not been show on the boxplot we are going to present now.

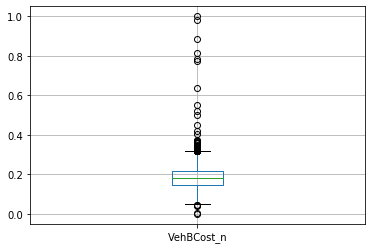
Anyway given the total number of data collected the amount of the outliers detected has not a relevant impact on data.











REMOVING OUTLIERS

Due to the fact that the numbero of ouliers don’t represent a bid deal for the prosecution of the analysis we don’t really need to adjust it filling up or removing them but just keep them and trying to remove them in the next stage to be precise clustering.