FRANCESCO

DETECTING OUTLIERS AFTER DATA CLEANING – IT REFERER TO df\_new

Index of all variables whose will be used

**VehOdo**

**VNZIP1**

**WarrantyCost**

**MMR\_factor\_n**

**VehBCost\_n**

I have some doubt about using variables which were normalized because in my opinion after that process we have lost the real value amount. I am processing them because I have found them after data cleaning

If someone want detect better it I am appending here the website

<https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba>

I am following these advices

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| VALUE | Q1 | Q2 | Q3 | Up-limit | out-max | Down-limit | Out-min |
| **VehOdo** | 61785.000000 | 73359.000000 | 82427.000000 | 113390 | 3 | 30822 | 260 |
| **VNZIP1** | 32124.000000 | 74135.000000 | 80022.000000 | 151869 | 0 | -39723.0  attention | 0 |
| **WarrantyCost** | 837.000000 | 837.000000 | 1623.000000 | 2802.0 | 57717 | -342attention | 0 |
| **MMR\_factor\_n** | 0.157782 | 0.211543 | 0.259000 | 0.410827 | 406 | 0.005955000000000016 | 335 |
| **VehBCost\_n** | 0.148805 | 0.183615 | 0.216506 | 0.3180575 | 141 | 0.047253499999999976 | 5 |

**DESCRIPTION OF VARIABLES ANALIZED**

**Q1,Q2,Q3 ARE THE QUANTILE 0.25,0.50,0.75.**

**THE UP LIMIT IS RAPRESENT THE TRASHOLD THAT YOU CAN FIND OVER EACH BOXPLOT.**

**THE DOWN LIMIT IS THE OPPOSITE ONE**

**OUT-MAX RAPRESENT TH NUMBER OF OUTLIERS OVER THE UP LIMIT**

**OUT-MIN RAPRESENT THE NUMBER OF OUTLIERS BELOW THE DOWN LIMIT**

**ATTENTION**

**THEY REFER TO THE FACT THAT AS YOU CAN SE FORM THE DOCUMENT I UPLOAD ON GITHUB, I HAVE FOUN TWO DOWN LIMIT NEGATIVE WHO REFER EXACTLY TO THE BOXPLOT BUT I DON’T KNOW WHY THE APPEAR NEGATIVE. MAYBE I HAVE TO VALUE THEM IN ABSOLUTE VALUE.**

**LET ME KNOW WHAT YOU THINK.**