Wrangling of the BackOrder dataset

**Goal**: Prepare the backorder dataset for EDA and Modeling

**Tasks performed:**

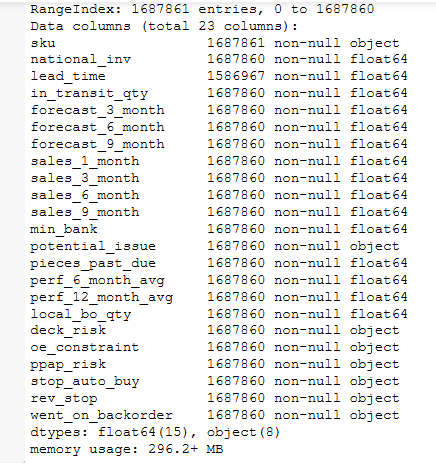
* Handling inconsistent column names and datatype
* Missing Data handling
* Removal of duplicate rows
* Handling columns with repetitive values
* Handling the outliers
* Write the clean data into a new file for further steps

**Reference**

* Details : <https://www.kaggle.com/tiredgeek/predict-bo-trial>
* Data source: <https://www.kaggle.com/tiredgeek/predict-bo-trial>

**Data load and description**

Dataset had 1687861 rows and 23 columns



**Handling inconsistent column names and datatype**

Columns were with mixed data types (was also evident from the pandas warning while loading csv). We wanted to identify and clean them up. Some of the column names had to be changed for better readability.

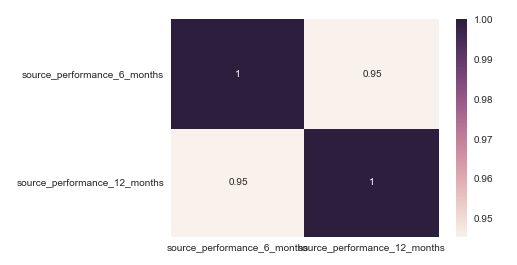
Many of the columns had just a single null entry. After investigation I realized that this all belonged to same row which was part of the footer .Removal of the footer solved the mixed datatype issue as well. I also changed the data type wherever applicable.

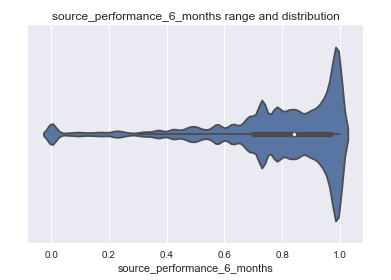
**Missing Data handling**

Missing values in columns source\_performance\_6\_months and source\_performance\_12\_months were represented as -99. I replaced -99 with NaN for the ease of processing.

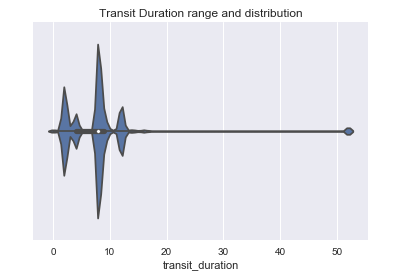
source\_performance\_6\_months had 129478 and source\_performance\_12\_months had 122050 missing values

Heatmap showed a strong correlation between source\_performance\_6\_months and source\_performance\_12\_months. So, linear regression is used to fill missing values. However another interesting point to note was that many observations had both source\_performance\_12\_months and source\_performance\_6\_months as null, so linear regression cannot fill such values so I looked for another approach. I checked for the central tendency of the data and replace the null accordingly. It’s clearly visible from violin plot that data is not distributed normally. So I picked up median to fill remaining values



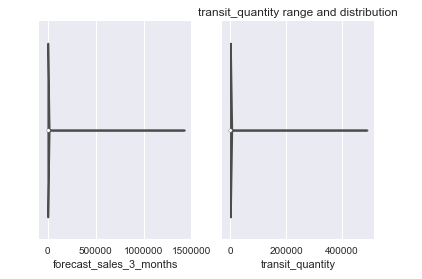


Transit Duration had 100894 null values. I didn't see correlation of this attribute with other features and also data is not distributed normally. So, again I chose median to fill the nulls.



**Handling columns with repetitive values**

Looking at the data set I realized there are many 0's in our dataset. So I decided to check on 0’s. I took approach to drop all the columns which has more than 60% 0’s.



All the below columns had more than 60% of the 0’s and I removed these features from the data set.

|  |  |
| --- | --- |
| **Features** | **0's** |
| forecast\_sales\_3\_months | 69.78% |
| forecast\_sales\_6\_months | 64.23% |
| forecast\_sales\_9\_months | 61.22% |
| source\_overdue | 98.50% |
| stock\_overdue | 98.62% |
| transit\_quantity | 79.67% |

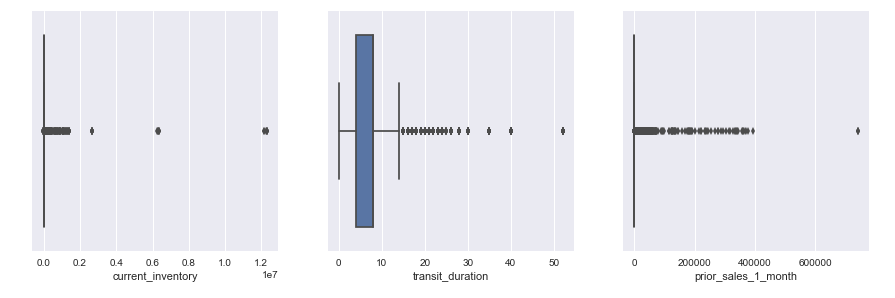
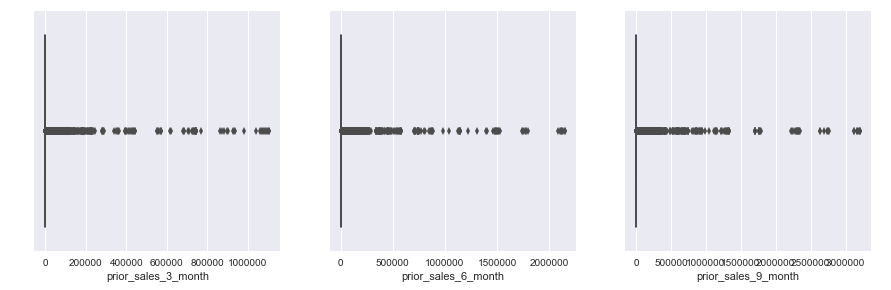
Same way we also had below categorical features which I removed-

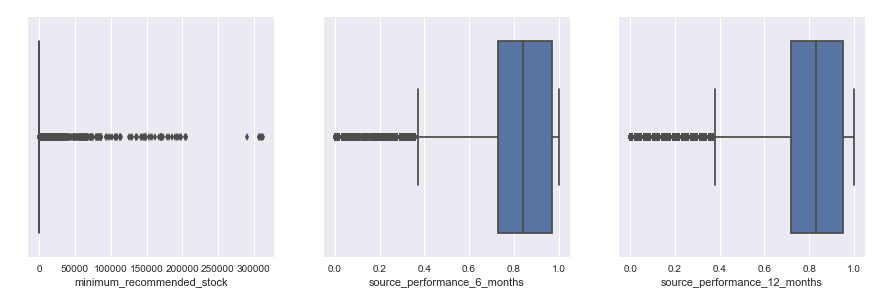
|  |  |  |
| --- | --- | --- |
| **Features** | **Yes** | No |
| oe\_constraint | 245 | 1687584 |
| rev\_stop | 731 | 1687098 |
| source\_has\_issue | 907 | 1686922 |

**Handling the outliers**

Outliers are something we need to take care at multiple phases. At this point of time I used univariate outlier detect techniques to come-up with the feature outliers.

Using the Boxplot some of the outliers can be determined and handled now itself.



Below are the outlier details - I handled

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Condition** | **Outliers count** | **Action** |
| prior\_sales\_9\_month | >100,000 unit | 3 | Removed the observation |
| transit\_duration | > 14 weeks | 48098 | Removed the observation |
| minimum\_recommended\_stock | > 210000 units | 7 | Removed the observation |
| source\_performance\_6\_months | <.3 | 95591 | Size is large. We will keep this as is for now and will decide during later stage |
| current\_inventory | > 0  <211 units | 225,000 | Further investigation is required |

**Write the clean data into a new file for further steps**

Finally I wrote the data to a new file Backorder\_clean.csv which will be used for further analysis.

Cleaned dataset has 1687829 observations and 14 features.