**Supply Chain Management – Model preprocessing and Feature engineering stage**

* The features Warehouse\_Id and WH\_Manager\_Id has been dropped because all entries are unique.
* The datatype for flood\_impacted, flood\_proof, electric\_supply, temp\_reg\_mach has been modified to object type from integer datatype as they have values that belong to only 2 categories: 0 and 1.
* Missing values are treated using median value.
* Outliers were present in the data and treated using capping method.
* From the bar plot it is found that there is no much variations among values belonging to:

**a**) Mid and Large category in WH\_capacity\_size. So, it can be clubbed together. After clubbing WH\_capacity\_size will have only 2 categories: Mid and Large.

**b)** B and B+ categories in approved\_wh\_govt\_certificate. These 2 features can be binned together. Now approved\_wh\_govt\_certificate has only 4 categories: A+, A, B, C.

* One hot encoding is performed on categorical variables to convert it into numeric.
* Features are then split into X and y for the machine learning algorithms to use.
* Multicollinearity is checked using variance inflation factor (VIF). The features that were categorical earlier shows a high multicollinearity. Numerical variables do not show any correlation between other variables.