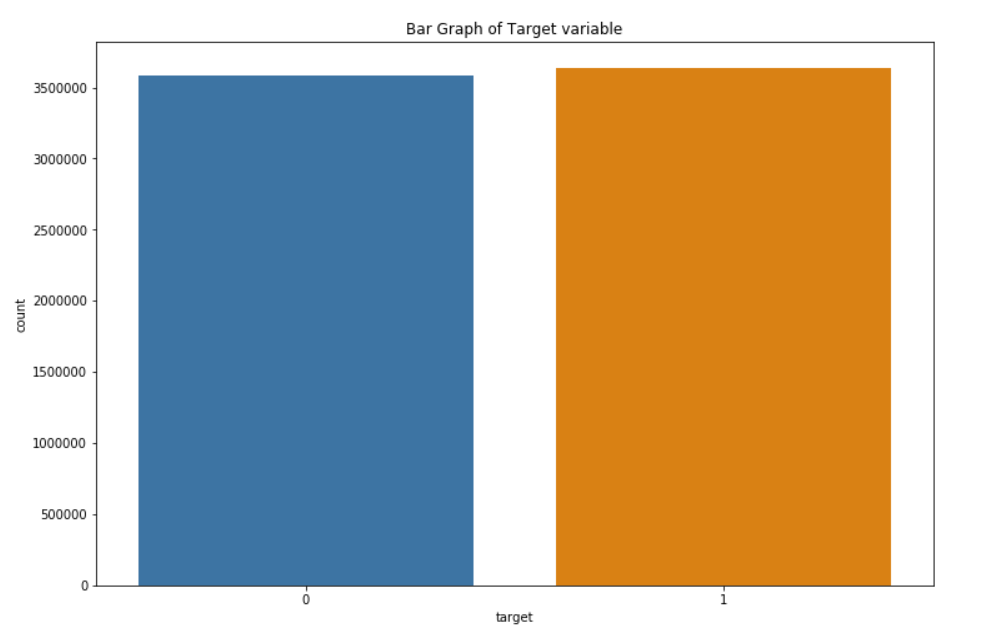
**Data Preparation:**

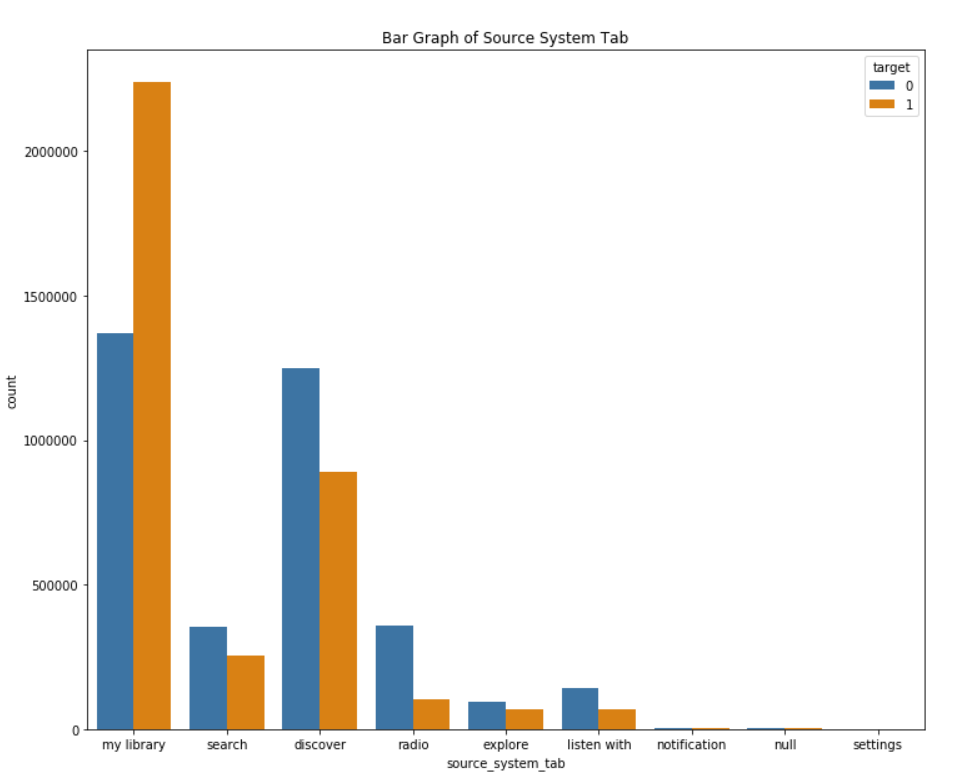
* Handled Missing Values
* Removed Outliers-

Age column – removed the rows that had values as 0 and above 60

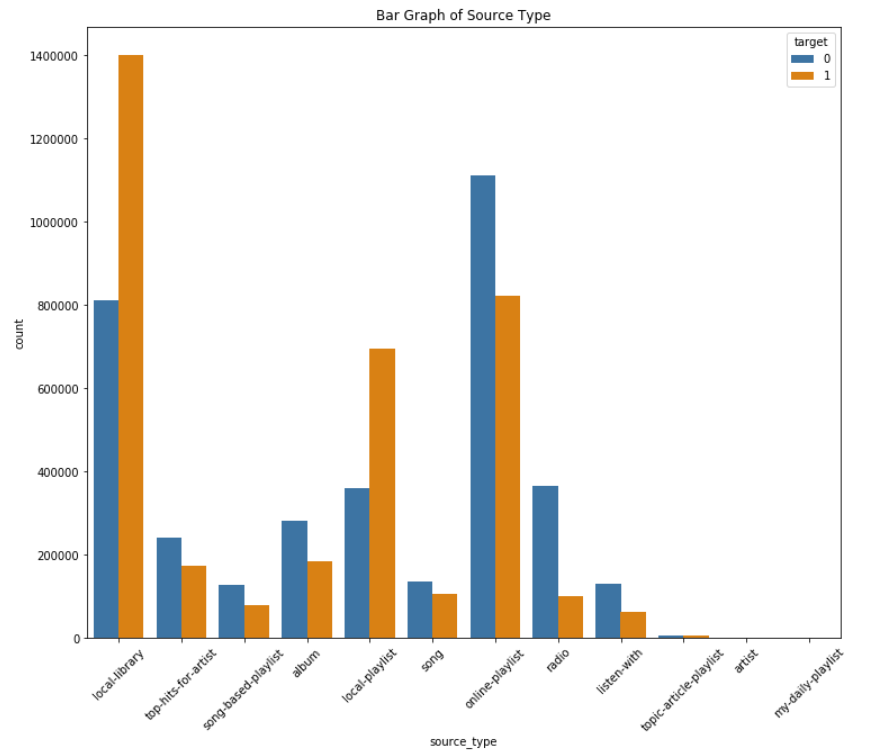
* Merging the datasets
* One Hot Encoding – for categorical attributes like 'source\_system\_tab', 'source\_screen\_name', 'source\_type', 'gender', 'language', 'genre\_ids', 'city'
* Feature Engineering -
* String Indexer-

**Visualization:**

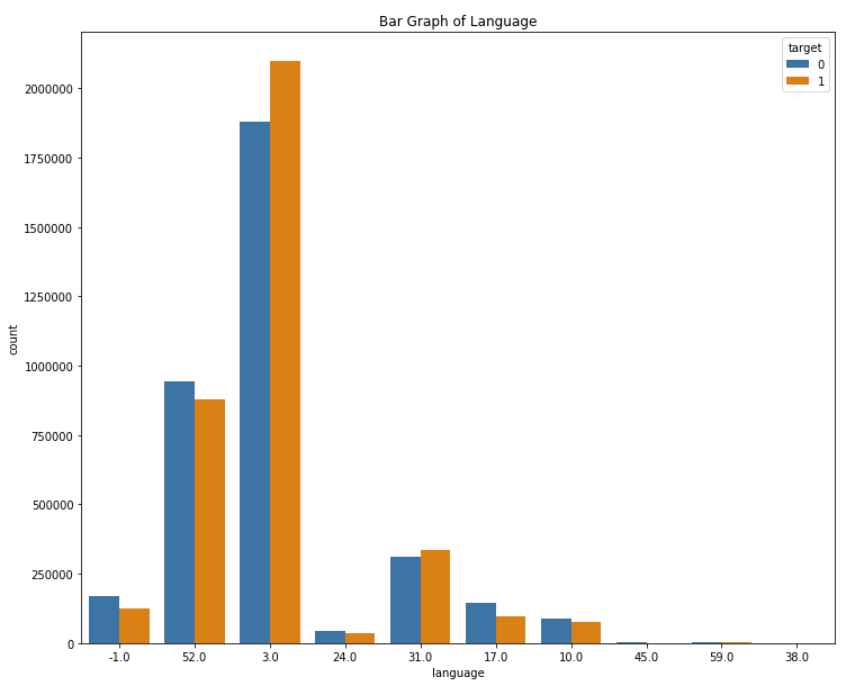
****As we can see that the ‘target’ variable is equally split between the value 0 and 1. Target=1 means there are recurring listening event(s) triggered within a month after the user’s very first observable listening event, target=0 otherwise. Hence, it is a balanced data.



We can observe that the ‘my-library’ and the ‘discover’ features of the app have the highest count of users from where they play their music .



The above graph tells us that most of the users prefer playing from their local playlist or local library when they open their app.



The above graph shows us that the 3 languages with codes 3.0 (Taiwanese), 52.0 (English) and 31.0 (Korean) make up for most of the data.