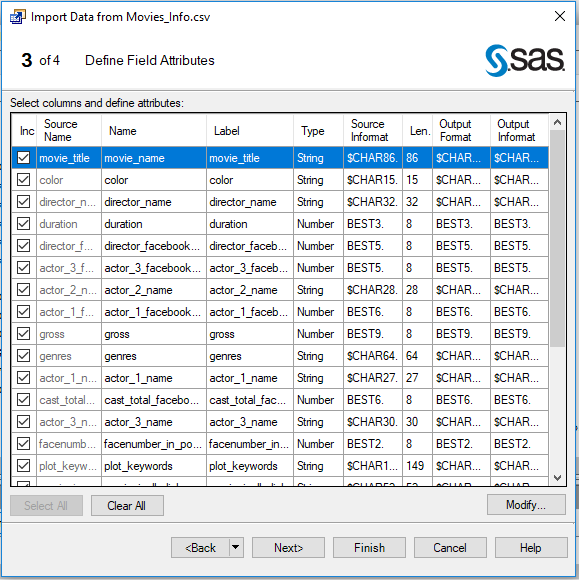
**Data Wrangling and Merging Tasks**

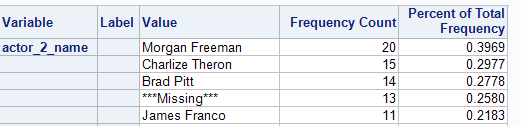
1. Import Movies\_Info.csv file into SAS Enterprise Guide. In step “3 of 4”, change the name of the column “movie\_title” to “movie\_name” to match with the IMDB\_Reviews.csv file. Also, in this step, review the column data types and make sure that the imported types are as they should be.

**Response:**

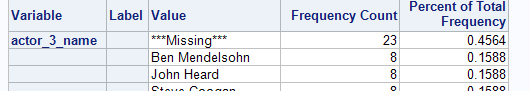


1. View the data characteristics of the various columns and list any data issues you identify. [Hint: From the Tasks, select Describe and then Characterize Data.] In particular, identify if the data has any of the following issues and support it with appropriate screenshots in a Word document:
2. Missing values

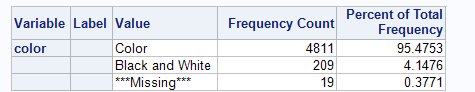
* actor\_2\_name has 13 missing values



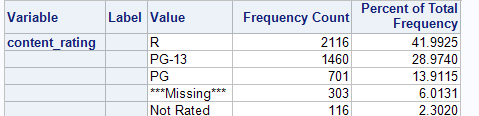
* actor\_3\_name has 23 missing values



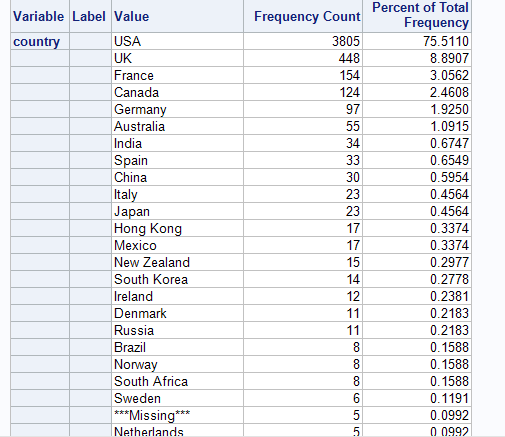
* color has 19 missing values



* content\_rating has 303 missing values



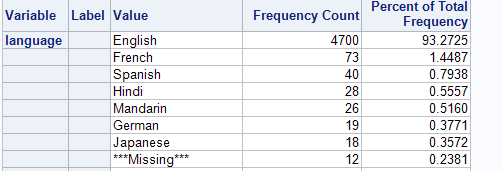
* country has 5 missing values



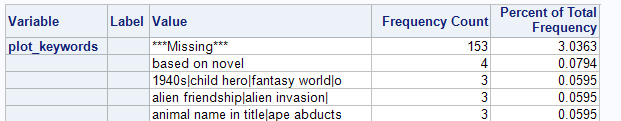
* director has 103 missing values



* language has 12 missing values



* plot\_keyword has 153 missing values



* actor\_1\_facebook\_likes has 7 missing values



* actor\_2\_facebook\_likes has 13 missing values



* actor\_3\_facebook likes has 23 missing values



* aspect\_ratio 328 missing values



* budget has 491 missing values



* director\_facebook\_likes has 103 missing values



* duration has 15 missing values



* facenumber\_in\_poster has 13 missing values



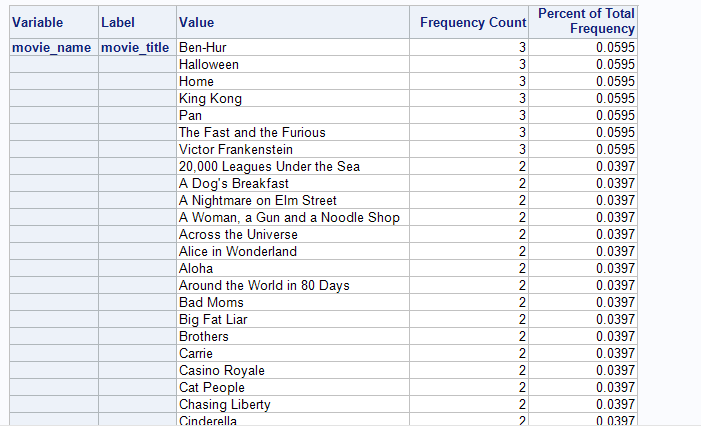
* gross has 882 missing values



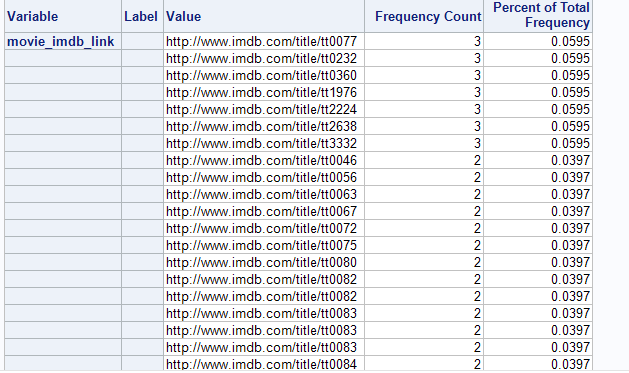
* title\_year has 107 missing values



1. Mislabeled values
   * There are no mislabeled values
2. Duplicate records
   * movie\_name has duplicate values



* + movie\_imdb\_link has duplicate values



1. Invalid range in any of the numeric or date columns

* budget has invalid range of values with min budget as 218 and max budget as 12215500000 which seems to be incorrect.

1. Suggest a plan of action for each of the above identified issue.
   1. actor\_2\_name, actor\_3\_name : Missing values are a valid case, as there are movies with just 1 or 2 actors, so this is ok.
   2. color: The number of missing color values are less than 5% of the total, so we will remove the records with missing color values (19 records).
   3. content\_rating: All the 303 missing values can be updated to ‘Not –Rated’ so as to not lose the valuable data.
   4. country: There are just 5 missing values, so we will remove the records with missing country.
   5. director\_name: The number of missing director\_name values are less than 5% of the total, so we will remove the records with missing director\_name values (103 records).
   6. language: There are just 12 missing values, so we will remove the records with missing language.
   7. plot\_keywords : The number of missing plot\_keywords values are less than 5% of the total, so we will remove the records with missing plot\_keywords values (153 records).
   8. actor\_1\_facebook\_likes, actor\_2\_facebook\_likes, actor\_3\_facebook\_likes, director\_facebook\_likes: The missing values present a valid case, as these actors and directors may not have a facebook page hence no likes.
   9. aspect\_ratio: The number of missing aspect\_ratio values are close to 6% of the total, but we cannot add any default value here as the aspect ratio varies from year to year the movie was made, so we will be dropping this field.
   10. facenumber\_in\_poster: Missing values are a valid case, as the movie poster may not have any faces of actors, so this is ok.
   11. duration: There are just 15 missing values, so we will remove the records with missing duration.
   12. gross: We will replace the missing 882 values in gross column with the mean of gross.
   13. title\_year: The number of missing title\_year values are less than 5% of the total, so we will remove the records with missing title\_year values (107 records).
   14. budget: We will remove the records with unacceptable budgets with an assumption that the acceptable range is 1400 to 300000000. So, delete records with budget < 1400 or budget >300000000, while keeping records with missing values and then replace missing budget with mean of budget (from the non-missing records).
   15. movie\_name and movie\_imdb\_link: This surely looks like there are duplicate records and resolving duplicates from movie\_name will resolve the duplicates in movie\_imdb\_link as well. We will sort the data and keep one of the duplicate record and deleting the other. If the duplicate records still exists after sorting, since we cannot determine which record has the true movie\_name without additional information we will remove both records with duplicate movie\_name. We will remove all the duplicate movie\_name records.