DSBA 6100: Data Wrangling with SAS Enterprise Guide (30 points)

Instructions: THIS IS AN INDIVIDUAL ASSIGNMENT. Do not work on this assignment in pairs. For this data wrangling assignment with SAS Enterprise Guide, you will use data on 5000+ movies scrapped from IMDB website by user chuansun76 and posted on Kaggle (https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset). The columns are self-explanatory for the most part and are derived from the movie info and the reviews on IMDB. You will clean one data file and merge it with the second data file to create a single dataset.

Download the two data files – Movies_Info.csv and IMDB_Reviews.csv - from Canvas and copy the files to Google Drive. Open a new Word document and provide your answers or screenshots on the Word document. After performing the data wrangling tasks described below, you will be asked to submit the final dataset (in both SAS and csv format) and the SAS Enterprise Guide project (submission instructions given on last page), in addition to the Word document.

Launch SAS Enterprise Guide and open a new project. Then, use the "Assign Project Library" under Tools menu to set up a new library and point it to the folder where you had copied the data files.

Data Wrangling Tasks to Perform

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.
2. 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:
    a. Missing values
    b. Mislabeled values
    c. Duplicate records
    d. Invalid range in any of the numeric or date columns

    For each issue identified, describe it in the Word document and provide appropriate screenshots to show the data issues.

3. Discuss a plan of action for each of the above identified issue. Write your answers in the Word document.
4. Use SAS Enterprise Guide to perform the following data wrangling actions. You do not have to implement all the actions described in step 3 above.
    a. Keep only one row if there are duplicates of the entire row. [Hint: **Tasks → Data → Sort Data**]
    b. Keep only rows with English language movies.
    c. Delete the rows with unacceptable values for budget. DO NOT delete rows with missing values.
    d. Replace the missing values in budget column with the mean of budget.

5. Import IMDB_Reviews.csv file into SAS Enterprise Guide. In step "3 of 4", review the column data types and make sure that the imported types are as they should be.
6. Merge the imported IMDB_Reviews data and the cleaned Movies Data into one new dataset containing only the following columns and in the given sequence. Name the output IMDB_Final_xxxx (where xxxx is your ninernet username).
    a. movie_name
    b. title_year
    c. genre
    d. language
    e. duration
    f. director
    g. budget (modified to replace missing values according to step 4d)
    h. movie_imdb_link
    i. content_rating
    j. imdb_score
    k. num_voted_users
    l. num_user_for_reviews
7. Export the merged dataset as a comma separated file and name it IMDB_Final_xxxx (where xxxx is your ninernet username).
8. In the Word document, write your name (or names) on top of the first page. Save the Word document as Answers_xxxx (where xxxx is your ninernet username).
9. Save the SAS Enterprise Guide project as Project_xxxx (where xxxx is your ninernet username).
10. Upload the following through the assignment upload link on Canvas.
    a. The final output dataset in SAS data format
    b. The final output dataset in csv format
    c. The Word document with your name and answers
    d. The SAS Enterprise Guide project (Project_xxxx.egp)

This assignment is due **October 7th 11.45pm.**