**Project Part 2: Basic Data Analysis (Due 12/7/21)**

**Description:** This phase of the project extends part 1 with three principal tasks:  
1) Preprocess the dataset into useful form\* and load it into the AUList structure **(40 pts)**  
2) Perform simple correlation analysis\* of the dataset. **(40 pts)**  
3) Discuss a related application to your dataset for which you believe an ADT from the 2nd half of the semester could be appropriate.\* **(20 pts)**

\*Specific details and requirements for the above follow (if you have already done the required preprocessing in Part 1, simply state as much in your report):  
1) Preprocessing entails taking the raw dataset you found online and converting it to one that satisfies the following properties:  
- **2-3 numeric variables** and **one discrete target variable of no more than 5 or so values** (sample rules for converting non-numeric to numeric data and continuous to discrete may be found in the Data Analysis session – If there are more than 5 values for the target variable, you should consider keeping only commonly occurring ones)   
- 50 to 1000 data samples. (If there are more, you should trim them; if there are less, make sure you clear things with me in advance).

2) A C++ program should be used to read the dataset into a compatible structure. You should modify the design you used in part 1 (or create one, if you did not finish!) For each value the target variable may possess, estimate the mean and standard deviation of at least two of the numeric variables. Write about *½ page* explaining your preprocessing, your sample structure, and whether or not you believe the 2-3 non-target variables in your dataset are correlated with the target variable.

For the 2nd portion of project, I ran into an issue with my dataset not meeting the criteria. The available variables I picked during the first part of the project did not have much coloration with each other, so I had to go back to the original dataset from Kaggle and pick data sets that correlate with each other. I decided to go with Platform, Name, NA\_Sales, EU\_Sales, and a column I added column called “Platform\_Count” to assigned a value from 1-21 based on the platform (i.e Xbox360). All variables I picked relate to my target variable, Platform\_count. NA\_Sales, EU\_Sales relate to my target variable in that it provided me on how many platforms were sold in North America. All selected data aided me into calculating standard deviation as well as mean. I chose to use both locations to better help me understand the data I had chosen. Name related to my target variable in that it provided me crucial information what kind of game were played on each platform.

3) (Note: I will be very flexible about your choice for this one.) Write about one paragraph detailing an application broadly related to your dataset that would be best handled with an ADT we looked at in the 2nd half of the semester (BST, priority queue, hash table). An example for the credit-card fraud dataset is below:

Modified video games sales could include “genre” variable. An appropriate ADT would be BST in which we would add name and genre of the game to the right child of the tree if it is great than the year we selected as our node (year between 1980 to 2016). Name and genre of the game would go to the left child of the tree if the year was less than the selected node. In this way I can better be organized to see a shift of customer love of genre over the years and which half of the years one genre thrived.

*Modified credit-card records could include a “risk” variable based on how likely the transaction is deemed to be fraudulent based on information like price, location, time, etc. The appropriate ADT for this data would be then be the priority queue, with higher risk transactions being placed higher in the queue. Investigators (manual or automated) would then “dequeue” records in order of risk when investigating potential fraud.*

***Unless you already have a plan set out, it is highly recommended that you use the files included with this document (including the credit-card fraud code) in conjunction with the structure you made for Part 1 as a starting point. Adapting the existing code to your data will be MUCH easier than writing it from scratch.***

***Your submission should be a .zip file that includes a .csv file for the final, preprocessed dataset, a series of .h/.cpp source files, and a .doc(x) or .pdf format for the report document. Upload the file as “LN\_FN\_ProjP2.zip” where LN is your last name and FN is your first name.***