Project Summary

In this project, Microsoft Excel or Google Sheets is needed to open the original Excel files.

The original Excel files can be found at: <https://catalog.data.gov/dataset/national-student-loan-data-system-722b0> .

In this project, the Jupyter Notebooks software was used to open and run notebook files with the .ipynb extension.

The following libraries were imported:

csv: For reading xls files as csv.

pandas: For loading data into dataframes.

numpy: For handling arrays.

matplotlib and seaborns: For plotting data observations.

statistics: For exploratory analysis and checking basic information of the table.

xlrd: This library must be installed (via "!pip install xlrd") AND imported to read XLS files.

re: For using regular expressions (regex) for string manipulation.

The xlrd package must be installed via the command “!pip install xlrd” in order to utilize the xlrd library.

The following questions were asked of the data:

1. Do community colleges or technical colleges originate more loans overall?

2. Which states have the highest total loan disbursements?

3. Do private or public schools disburse more in federal student loans?

4. Do institutions with “college” in the name receive more loans than those with “university”?

5. Are unsubsidized loans more common at private colleges than public ones?

The following insights were discovered:

1. Community Colleges originate more loans than technical colleges. Various factors may account for this, such as there being a higher number of community colleges than technical colleges. It is also possible that more students attend community colleges rather than technical colleges.

2. Pennsylvania, Texas, and New York have the highest loan disbursements. This could be due to a variety of factors such as state population, student population, amount of colleges per state, and differences in school funding between the states.

3. When observing the original Excel files, it should be noted that community colleges have missing data concerning the Federal Grad Plus loans. When cleaning the data for missing values, caution is advised. By removing rows containing 2 or more missing values, all of the community colleges are dropped from the data.

4. Public schools disburse more in federal student loans. This may be due to a higher amount of students attending public schools rather than private schools.

5. Colleges with "University" in the name disburse more loans than colleges with "College" in the name. However, this could be due to a variety of possible factors, such as more colleges having the word "university" in their name overall. Universities have higher tuition, and thus more students apply for student loans in order to pay for the tuition.

6. Public colleges disburse more unsubsidized loans than private colleges. However, this may be due to more students attending public schools.

If presenting this information to a client, it is recommended to automatically remove all foreign schools from the data in order to focus on American schools. State codes FC and PR should be removed, as well as the school types “FOREIGN PUBLIC” and “FOREIGN PRIVATE”. Proprietary schools could also be lumped under the “Private” school type for higher accuracy.

For future work, graduate loans (Federal Grad Plus) could be analyzed to determine patterns in which schools or states give the most amount of graduate loans.