De-identifying Cedarville's loan data

I. Purpose

- Risk Reduction: De-identifying data reduces the risk of data breaches and unauthorized access.
- Data Sharing: De-identified data is easier to share with third parties, facilitating research and analysis.
- Legal Compliance: Organizations may not be required to report breaches involving de-identified data.
- Privacy Protection: Individuals' personal information remains confidential.

II. Columns to change

- `Student ID`
- `Student Name`
- `Loan ID`

III. Methodology

1. Edit `Master`
   - Add column after `Student ID`, `Student Name`, and `Loan ID` filled with sequential names/numbers
2. Edit `Loans`
   - Add columns respective to information I want to update
   - Update `Loans` with information from `Master` on matching columns
3. Final steps
   - Remove identifiable columns
   - Rename de-identified columns
   - Export each file as a `.csv` ## IV. De-identifying the data

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import pearsonr, chi2_contingency

loans_csv = pd.read_csv(
    r'C:\Users\jsbit\OneDrive\Documents\Coding 2023\Christie Tim Pandas 2.9.24\Modifie
    encoding_errors='replace')
master_csv = pd.read_csv(
    r'C:\Users\jsbit\OneDrive\Documents\Coding 2023\Christie Tim Pandas 2.9.24\Modifie
    encoding_errors='replace')
```

In [2]:
```
# Add new columns of de-identified information
master_csv.insert(0, 'New_ID #', range(6000, 6000 + len(master_csv)))
master_csv.insert(2, 'New_Name', 'Student')
master_csv.insert(7, 'New_Loan ID', range(90000, 90000+len(master_csv)))
master_csv['New_Name'] = master_csv['New_ID #'].apply(
    lambda x: 'Student{}'.format(x))
```

In [3]:
```
# Add columns I wish to update in 'Loans'
loans_csv.insert(0, 'New_ID #', 6)
loans_csv.insert(2, 'New_Name', 'Student')
loans_csv.insert(11, 'New_Loan ID', 'Loan ID')
```

In [4]:
```
# Update 'Loans'
loans_csv = loans_csv.set_index('Name')
master_csv = master_csv.set_index('Name')
loans_csv.update(master_csv)
loans_csv.reset_index(inplace=True)
master_csv.reset_index(inplace=True)
```

In [5]:
```
# Remove identifiable columns
clean_master = master_csv[['New_ID #', 'New_Name', 'GS Program',
                           'Unnamed: 3', 'Unnamed: 4']]
clean_loans = loans_csv[['New_ID #', 'New_Name', 'Term', 'Loan', 'Term Awd',
                         'Term Fee', 'Term Disb', 'Status', 'New_Loan ID']]
```

In [6]:
```
# Rename de-identified columns
clean_master.columns = ['ID #', 'Name', 'GS Program', 'Unnamed: 3', 'Unnamed: 4']
clean_loans.columns = ['ID', 'Name', 'Term', 'Loan', 'Term Awd',
                       'Term Fee', 'Term Disb', 'Status', 'Loan ID']
```

In [7]:
```
# Export
clean_master.to_csv('clean_master.csv', index=False)
clean_loans.to_csv('clean_loans.csv', index=False)
```