Finder Project
By: Tyler Gimple

Introduction: For this project I was instructed to look for correlation between variables in the 2019 Household Debt and Credit Report from the Center for Microeconomic Data of the Federal Reserve Bank of New York. To calculate correlations between all variables in the data set, I created a R-script program that calculated every variables' correlation with every other variable in the data set. The complete dataset, R-script, and results tables are available at https://github.com/tylergimple/Correlation_Project. To accomplish this, I first had to clean the data into a new spreadsheet entitled 'Base Data' to ensure that all of the data aligned in terms of time as some data began in 2003 and some in 2000 or 1999. The data then had to be loaded into the R program so it could calculate the correlations. The program resulted in the calculation of 49,749 Pearson R correlation values. See the full spreadsheet "Hmisc Corr Table.xlsx" which is previewed in the screen capture below (Figure 1).

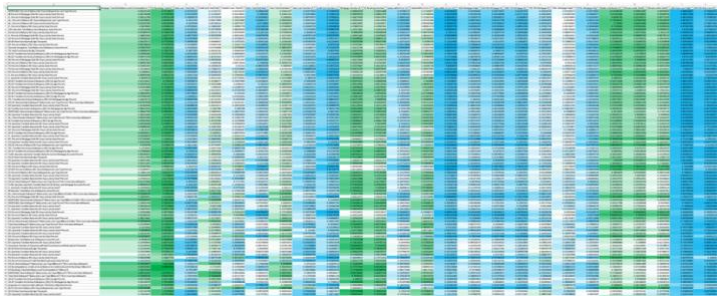

Figure 1: Screen capture showing correlation values between variables in microeconomic data. Green indicates positive Pearson correlations and blue indicates negative Pearson correlations.

Analysis: In order to identify correlations of potential interest, the correlations were visualized in descending order by correlation value. Figure 2 shows interesting positive (green) and negative (blue) correlations related to student debt as discussed further in section 2.c. In the analysis phase, I focused on 3 areas of interest: Foreclosures/Bankruptcies, Student Loans, and a 'Pre-Data' hypothesis regarding Debt and Foreclosures. The correlation values calculated are Pearson Correlation Coefficients where 1 is perfect positive linear correlation, 0 is no linear correlation, and -1 is perfect inverse linear correlation.

A	B	C	D
Index	row	column	Pearson R Correlation Value
246	Student Loan.Total Debt Balance and Its Composition in Trillions	70+.Total Debt Balance by Age Trillions of Dollars	0.979410022
291	Student Loan.Total Debt Balance and Its Composition in Trillions	STUDENT LOAN.New Seriously Delinquent* Balances by Loan Type Billions of Dollars	0.971675634
803	Student Loan.Total Debt Balance and Its Composition in Trillions	Total.Total Debt Balance and Its Composition in Trillions of $	0.775304432
3165	Student Loan.Total Debt Balance and Its Composition in Trillions	620-659.Mortgage Origination Volume by Riskscore in Billions $	-0.778364646
3170	Student Loan.Total Debt Balance and Its Composition in Trillions	TX.Quarterly Transition Rates into 30+ Days Late by State*	-0.797007511
3171	Student Loan.Total Debt Balance and Its Composition in Trillions	PA.Percent of Consumers* with New Bankruptcies by State Percent	-0.805224914
3173	Student Loan.Total Debt Balance and Its Composition in Trillions	<620.Mortgage Origination Volume by Riskscore in Billions $	-0.814513721
3224	Student Loan.Total Debt Balance and Its Composition in Trillions	720-759.Mortgage Origination Volume by Riskscore in Billions $	-0.830189561
3252	Student Loan.Total Debt Balance and Its Composition in Trillions	inquiry within 6 mo.Total Number of New and Closed Accounts and Consumer Credit	-0.83977147

Figure 2: Screen capture of correlation table showing Pearson values between selected variables related to student debt.

Attached Items:

Base Data: Data from the 2019 Household Debt and Credit Report from the Center for Microeconomic Data of the Federal Reserve Bank of New York that was formatted for input into the custom R-script.

Hmisc Corr Table: The first output of the R program is a table showing the Pearson correlation values for each pairwise comparison. Values are color-coded based on the strength of positive or negative correlation.

Flat Hmisc: The second output of the R program is a table showing the Pearson correlation values for each pairwise comparison in a 'flat' format.

Correlation Program.R: R-script for calculation of all Pearson correlation values from the "Base Data" dataset. All data, scripts and results are available at:
https://github.com/tylergimple/Correlation_Project

Interpretation of Results

1. Examination of Foreclosures and Bankruptcies.
    a. Foreclosures
        i. **Age Groups:** The correlation coefficients between overall foreclosures and new foreclosures among 30-39 year-olds, 40-49 year-olds, and 50-59 year-olds are all very high (0.98,0.99,0.97) indicating that in quarters where the three mentioned age ranges are high we also see equally high levels of overall foreclosures. 70+ year olds are also positively correlated but to a slightly lesser extent (Figure 3).

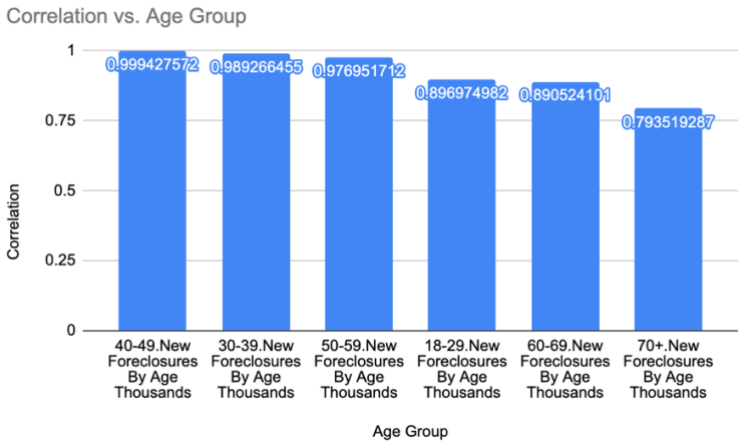

Figure 3: Pearson correlation coefficients between overall foreclosures and new foreclosures in various age ranges are displayed.

b. Bankruptcies
    i. **Debt:** Interestingly, there is an inverse correlation between the overall total debt balance and the overall number of national bankruptcies with a correlation coefficient of -0.47 (Figure 4). This indicates that in cases of higher debt balance, we actually see lower levels of bankruptcies. However, -0.47 indicates a relatively weak correlation.

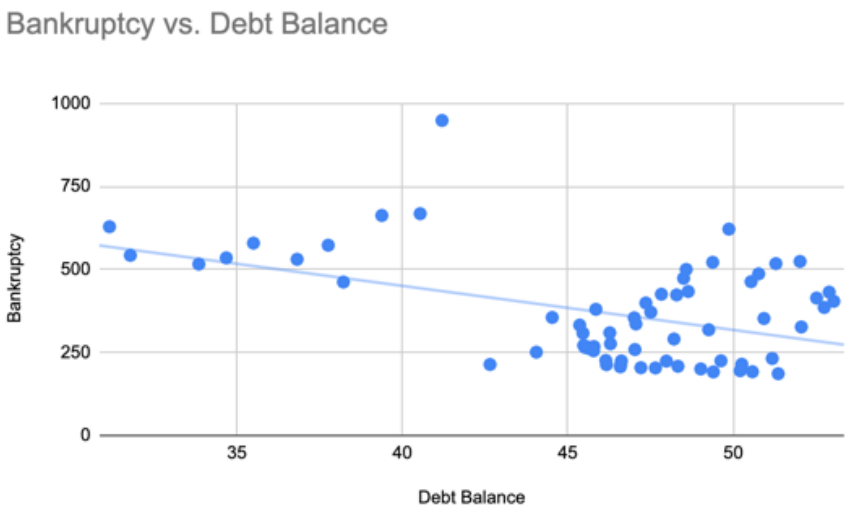

Figure 4: Correlation between overall total debt balance and the number of overall bankruptcies over time.

2. **Student Loans**: The issue of student loans is a contentious topic as there are many younger people facing high levels of loan debt in the United States. Let's examine their correlations.
    a. **Age Groups:** When studying overall student loans, we counterintuitively see that student loan debt is most positively correlated with older age groups, namely 70+ and 60-69 year-olds at 0.97 and 0.96 correlation coeffcients respectively. It can also be observed that in the younger generations of 18-29 year-olds we see a low and insignificant correlation value of 0.33 (Figure 5).

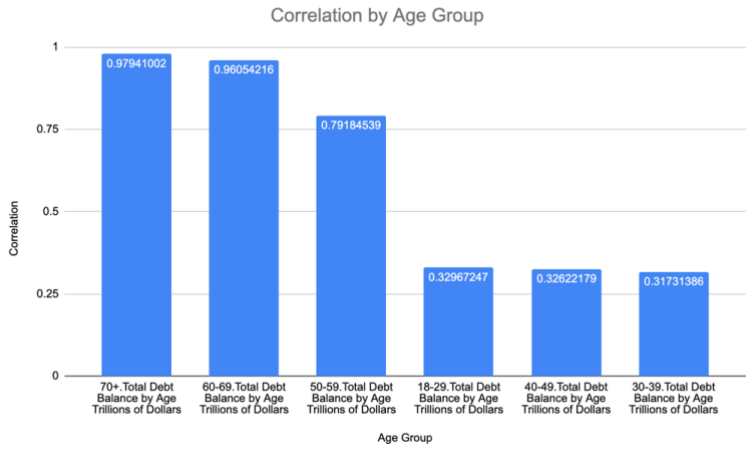

Figure 5: Correlations between overall student loans and total debt balance by age group.

b. **States:** When looking at states it can be observed that Texas has the highest correlation to Total debt balance with a 0.91 correlation. The next highest states are Pennsylvania and New York with levels of 0.80 and 0.71. Michigan and Nevada actually have negative correlation coefficients indicating that as Nevada's student loan debt increases we observe lower levels of overall student debt (Figure 6).

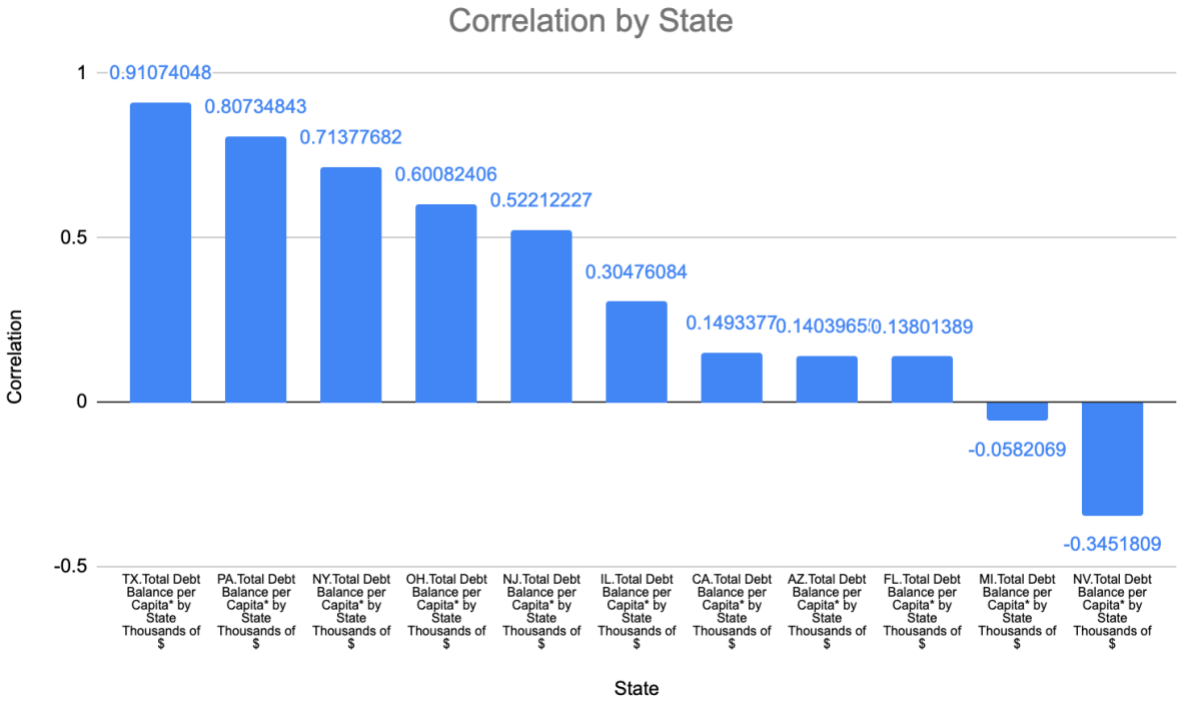

Figure 6: Correlation values between overall student loans and total debt balance by state.

c. **Total Debt/Mortgage**: When looking at how student loan debt correlates to other metrics, there is a positive correlation of 0.78 between student loan debt and total debt balance but a negative correlation coefficient of -0.83 between student loan debt and mortgage origination volume within the risk score of 720-759 (Figure 7).

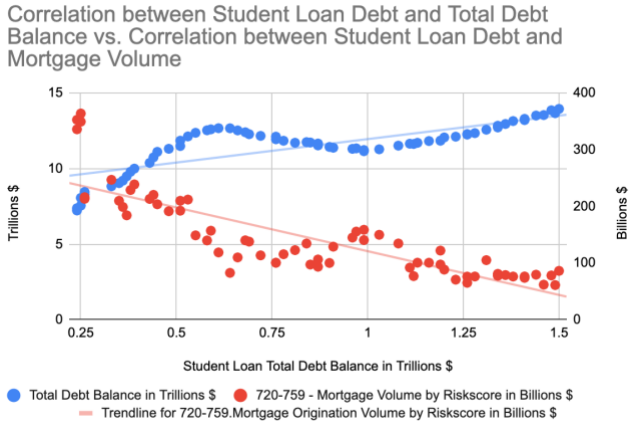

Figure 7: (Blue) Correlation between student loan debt and total debt balance. (Red) Correlation between student loan debt and mortgage origination volume.

3. **Pre-Data Hypothesis:** For this analysis, I tested the hypothesis that there would be a strong correlation between total debt owned by 30-39 year-olds and new bankruptcy levels in the same age group, that is to say that as debt increases we will see a higher level of new bankruptcy and vice-versa. The data shows that there is actually a weak *negative* correlation of -0.55 between the two data sets indicating that as debt levels increase we actually observe lower levels of bankruptcies. However, we calculate an R^2 value of just 0.31, which indicates that less than a third of the variation of new bankruptcy can be explained by total debt (Figure 8). Thus this proves the hypothesis to be false as the two entities are, in fact, weakly negatively correlated.

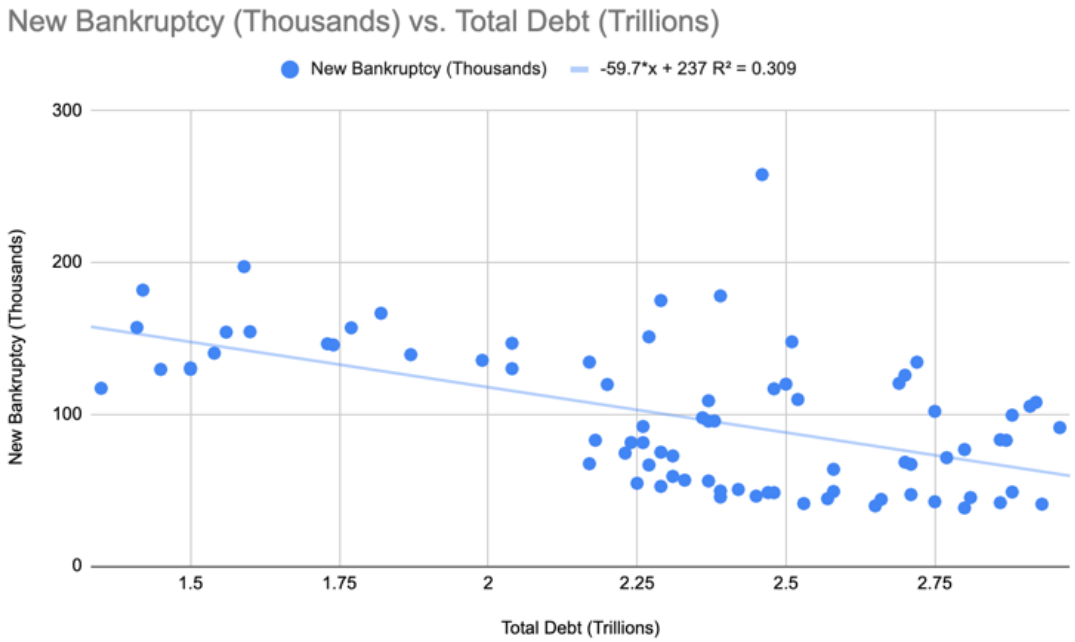

Figure 8: Correlation between total debt in the 30-39 year-old age group and new bankruptcies in the same age range.