**Creating a Predictive Model for Credit Delinquencies**

David Reese

Colorado State University Global Campus

MIS 581 Capstone- Business Intelligence and Data Analytics

Dr. Kimberly Ford

02-19-2021

**Creating a Predictive Model for Credit Delinquencies**

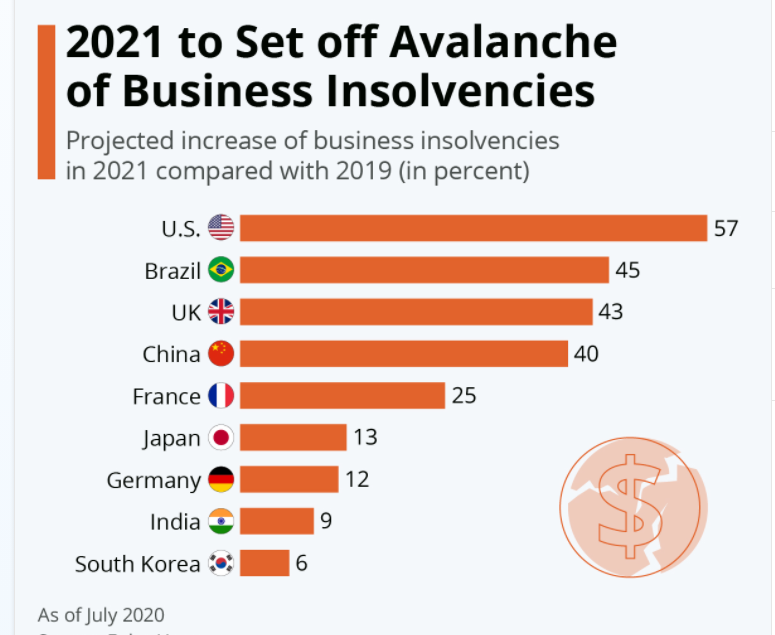
**Introduction**

Unlike the 2008-2010 economic meltdown, which impacted primarily sub-prime mortgages, COVID-19 has impacted every area of the economy. Viewing COVID-19’s impact In terms of drawdown in the Gross Domestic Product (GDP) of the United States, the housing collapse of 2008-2010 led to a five percent drawn down in GDP over two years, and the great depression saw a 30 percent drawdown in GDP over 3 years. COVID-19 saw a 32.9 drawdown in the second quarter of 2020. Lockdowns and new rules decimated service, hospitality, and restaurants as well as many other industries across the entire country.

Unlike 2008, the resulting job losses are more permanent. Small business closures will continue to grow as the impact of COVID-19 continues to wreak havoc. Figure one shows the impact on business insolvencies versus 2019.

**Figure 1**

*Business insolvencies in 2021 compared with 2019 (Buchholz, 2020)*

**

Insolvencies mean that finding a new job at or near what your last position paid will be even more difficult. This is reflected in the unemployment rate for the country during 2020. In January 2020, the unemployment rate was four percent (unadjusted). By April 2020, the rate was 14.4 percent (U.S. Department of Labor, 2021). As of January 2021, the rate had gone done to around six percent, but of that figure, nearly forty percent had been unemployed for over 27 weeks.(U.S. Department of Labor, Bureau of Labor Statistics 2021).

Figure two shows the impact of unemployment caused by the pandemic on Blacks, Latinos Whites and Women in unemployment (U.S. Department of Labor, Bureau of Labor Statistics 2021).

**Figure 2**

*Unemployment rates by demographic*

To see the impact of this, twenty-seven weeks is equivalent to thirteen bi-weekly paychecks. Assuming a regular salary equal to $2000 take home, that would equal a loss of income of $26,000. That is money which was a factor in the credit decision for car loans, home mortgages, and credit cards. It is true that there are programs to offset the loss of income, and there has been direct stimulus to the economy, those programs have limits on the amount and length of time the person may be eligible for them.

In the post COVID-19 economy, one risk area is the banking and lending sector. With trillions of dollars at risk for default due to lower earnings and protracted unemployment, a way to predict which loans are at risk of delinquency or default must be found. Using the available data, and finding which variables have predictive power in answering that question, a model can be created to protect the banks from the type of meltdown seen in 2008-2010.

**Objectives**

The goal of the project is to prove the hypothesis that there is a predictive analytic model that can predict, with a degree of accuracy, the consumers who will become delinquent. Gaining an understanding of of the consumers who were current, and are heading towards delinquency can help banks better set loss reserves, change lending and underwriting guidelines, and protect the solvency of the banks. To prove the hypothesis, data from the Federal Reserve system will be used to create a predictive model of the typical consumer who may be at risk for delinquency.

The model will be based on auto loans and credit cards. Those variables are the consumer debts that are not subject to forbearance or other legislation.

**Overview of the Project**

The first step in the process would be to identify those datasets that offer some insight as to what we are looking for. The Federal Reserve system has a treasure trove of studies and data to work with. The most valuable has been the Household Debt and Credit survey. Done quarterly, this report gives information direct form anonymous credit files and bank records regarding new delinquencies, bankruptcies, and other factors. The Federal Reserve banks track virtually every aspect of consumer behavior and sentiment. Data sets dealing with debt-to-income ratios, demographic information regarding age, income, and loan type are available and can be set up to create models. The focus of the data will be on the data with some link to consumers.

The second step would be to isolate those consumer-focused studies which have information deemed valuable, and begin to see if any variable shows any predictive power. Predictive analysis would be run on single variables and then on combinations of variables to see which variables create the best predictive model. Regression analysis and correlation done in SAS would provide the most predictive variables.

Lastly, once the model has been created, a process to check the accuracy would be developed. Given that at a certain point in time we know how much of the credit market is delinquent, it may be possible to run the model, on the older data. It would be like running the analysis backward to see how close the model came to predicting the answer.

**Hypothesis**

There are certain factors which are common to all credit delinquencies. The challenge becomes finding the factor, or combination of factors that can generate the most predictive power. For years, the gold standard for credit risk has been an individual’s credit score from one of the three reporting bureaus. COVID-19 has changed that dynamic, thanks to its impact on the economy. Using the Federal Reserve Bank of New York quarterly report on Household Debt and Credit, along with other data sources from the Federal Reserve system, a predictive model that considers other variables can be designed.

The research question is this: which factors cause a change in status from current to delinquent? Additionally, which of those factors are the ones that have the most impact on a consumer falling behind?

To test this hypothesis, it may be necessary to create a fictitious dataset. It could also be that the Federal Reserve has ways to model certain attributes and create a model based on the data from the hundreds of reports they provide. Auto loans and credit card debt will be studied, since these two debts are not subject to any forbearance.

To answer the research question, the following hypothesis have been created.

The null, and alternative hypothesis would be as follows:

HA  = Certain factors are strongly correlated with credit delinquency, and can be used to create an accurate predictive model.

Ho  = There is no single factor, or combination of factors, which create an accurate predictive model for delinquency.

These hypothesis would be tested on auto loans and credit cards to see if there is a variable, or set of variables that can predict delinquencies.

**Literature review**

The Federal Reserve System offers very little in the way of literature. It does have amazing and complex surveys regarding consumer attitudes. Understanding the numbers is one aspect of this crisis. The other side is understanding human behavior and societal issues that drive the use of credit.

As stated earlier, COVID-19 has impacted every area of the economy. Consumer debt is no exception. In reviewing the literature regarding the social impact of debt on consumers, it becomes clear that the decision to go into debt in the first place is driven more by societal expectations, than actual need. In a study by the Aspen Institute, the findings showed that people often support status-enhancing consumption through borrowing (Holt & Lucas McKay, n.d.).

Lending institutions understand this and have developed products for lower income and lower credit risk score clients. The new products led to growth in unemployed consumers having access to credit cards to borrow to cover short falls. From 1977 to 2010 the rate of consumers with uneven job tenures who had access to credit cards grew from 17% to 45% (Herkenhoff, 2019).

These two pieces paint the picture of the COVID-19 credit user. Prior to the pandemic, optimism and wages were high, borrowing rates were low, so the decision was made to purchase a car, go on a trip, refurbish a bathroom, all on borrowed money. The payments could be handled by the budget. Now, the borrowing for the trip, or bathroom is still being paid for, along with all the other bills. This causes a great deal of stress of consumers.

To maintain stability in the housing market, and ease homeowner stress, the Coronavirus Aid, Relief, and Economic Security Act (CARES act) has allowed consumers to skip mortgage payments without penalty (derogatory credit – late payment notation). A homeowner is eligible for ‘forbearance’ provided the mortgage is federally guaranteed, and they meet other requirements. For a period of 180 days, with an option for an additional 180 days, the property owner does not have to make payments. All with little or no consequence to them at the time the decision is made to take advantage of the program. The payments are either added to the end of the loan, spread across the remaining months, or left as a balloon payment due at the time the house is sold. This is in essence a form of borrowing, but one which is legislated. If the loan was reviewed, given the changes in conditions due to COVID-19 would the property owner qualify? In 2008 the property would have been repossessed after a certain number of months of non-payment. In COVID-19, Federally guaranteed mortgage payments are forgiven for a period of up to a year.

This legislation, while good for property owners, changes the actual delinquency rates in the mortgage market, and therefore the ability of any model to be accurate. Therefore, for the purposes of the predictive model; the decision was made to not use mortgage data. Figure 3 shows clearly the impact of this policy. Forbearance started in the first quarter of 2020, and delinquent mortgages fell by over 50%.

**Figure 3**

*Impact of Forbearance on mortgage delinquency. (Federal Reserve Bank of New York, 2021 pg. 15)*

**Research Design: Methodology, Methods, Limitations, and Ethical Considerations**

The project is looking to create a predictive model that can help banks find out which consumers are at risk for delinquency, but not yet delinquent. The Federal Reserve data is anonymous, so there are no real worries as to data privacy, or ethical concerns as to data collection or use. If the model works, it should be able to give a composite sketch of a consumer at risk. The addition of a bank’s client list would create an entirely different situation requiring several security protocols, and notification of clients as to what was going on.

**Methodology**

The software platform SAS will be used to generate correlation reports as well as regression analysis to try and determine the variables which add to the predictive power of the model. Being able to chose which variables have the highest correlations will make sure the model is accurate.

**Method**

The first step would be to compile all the meaningful data and begin to join the datasets together. In this way the analysis could be done easily, and rapidly in either SAS or python. The goal is to find enough information on the type of loan, or the consumer, to create an accurate picture of of the risk.

Secondly, the project will test the model for its accuracy of prediction. This will involve testing individual variables as well as combinations of variables to see which has the best response. It is critical that the right combination of variables be found so that the best model can be created.

The overall dataset is complied from the the New York Federal Reserve report on Household Debt and Credit. The period is 20 quarters, from 3rd quarter 2015 to 3rd quarter 2020. There are 48 variables dealing with the transition to delinquency, the age of the consumer, and the origination score of loans. The data headings are in Appendix A.

**Limitations and Ethical Considerations**

There is no limit to the data that the Federal Reserve has available. 12 regional banks, plus the headquarters in Washington create vast amounts of reports, and there are customizable drop-down style menus for generating reports involving different topics. One major limit is the lack of *consumer* data. In many predictive models, a database of actual people is available. With Federal Reserve data, there is a great deal of data *on* consumers, but nothing about the *actual* consumer. If the predictive model works, then it would be up to whoever uses it to make decisions regarding actual consumers is safeguarding the data. They must be sure the data is protected properly and all protections are in place to safeguard the information. Ultimately, the responsibility for the data rests with the individual using it.

**Findings**

To do the review of the dataset, the SAS platform was used. Correlation and regression analysis was run on the 48 different variables. In the end, the hypothesis was proven out by two different variables. The first was the Financial Obligation Ratio, the second was Household Debt Service Ratio (DSR). The Household Debt Ratio is defined by the Federal Reserve as “the ratio of total required household debt payments to total disposable income” (Federal Reserve Bank, 2013 paragraph 1). The Financial Obligation Ratio (FOR) is defined “is a broader measure than the Debt Service Ratio. It includes rent payments on tenant-occupied property, auto lease payments, homeowners' insurance, and property tax payments” (Federal Reserve Bank, 2013 paragraph 16). Because the FOR is broader, it was able to show correlations with new delinquencies in auto loans. In the data set the variable FINOB was used to for the Financial Obligation Ratio.

DSR was better able to predict credit card debt issues. This makes sense given the definition of the Federal Reserve. By using both, where each is a better predictor, the modeling could be done.

FINOB is the variable name for the Federal Reserve Bank’s Financial Obligation ratio. It was found to be the better predictor for new auto loan delinquencies. There is a strong positive correlation between financial obligations, and percentage changes in new auto delinquencies.

**Table 1**

*Details of correlation analysis on FINOB and New Auto Delinquencies.*

| **Simple Statistics** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Dev** | **Sum** | **Minimum** | **Maximum** | **Label** |
| **FINOB** | 21 | 15.20642 | 0.44075 | 319.33485 | 13.73775 | 15.61985 | FINOB |
| **NEW\_DEL\_AUTO\_PER** | 21 | 7.06442 | 0.39051 | 148.35281 | 5.79000 | 7.46124 | NEW\_DEL\_AUTO\_PER |

| **Pearson Correlation Coefficients, N = 21 Prob > |r| under H0: Rho=0** | | |
| --- | --- | --- |
|  | **FINOB** | **NEW\_DEL\_AUTO\_PER** |
| **FINOB FINOB** | 1.00000 | 0.85734 <.0001 |
| **NEW\_DEL\_AUTO\_PER NEW\_DEL\_AUTO\_PER** | 0.85734 <.0001 | 1.00000 |

**Figure 4**

*Graph of correlation between FINOB and New Delinquencies is Auto loans*



Regression analysis was also run on the same variables. Table 2,3, and 4 shows the results.

**Table 2,3, and 4**

*Regression analysis results*

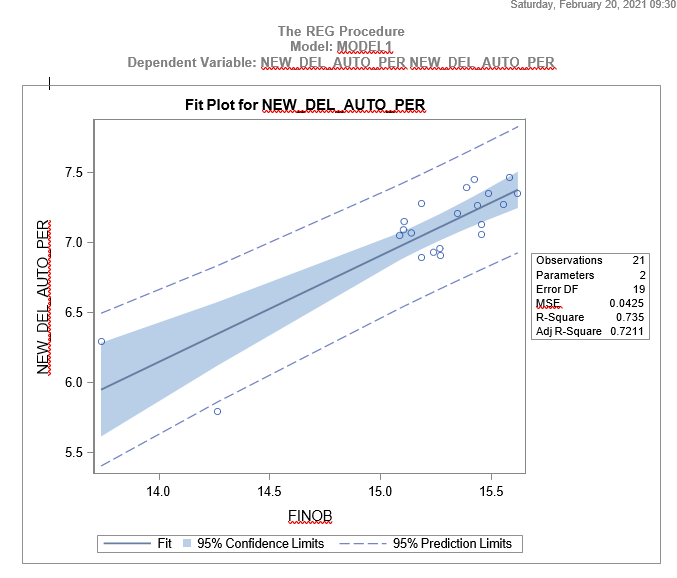
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | 2.24184 | 2.24184 | 52.70 | <.0001 |
| **Error** | 19 | 0.80818 | 0.04254 |  |  |
| **Corrected Total** | 20 | 3.05002 |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | | |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard**  **Error** | **t Value** | **Pr > |t|** |
| **Intercept** | Intercept | 1 | -4.48676 | 1.59175 | -2.82 | 0.0110 |
| **FINOB** | FINOB | 1 | 0.75963 | 0.10463 | 7.26 | <.0001 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 0.20624 | **R-Square** | 0.7350 |
| **Dependent Mean** | 7.06442 | **Adj R-Sq** | 0.7211 |
| **Coeff Var** | 2.91944 |  |  |

**Figure 5**

*Regression analysis, confidence intervals*



A linear regression established that financial obligations could statistically significantly predict changes in the percentage of new auto loan delinquencies. The prediction equation is : -4.48676 + (.75963 x quarter).

In terms of credit card debt the variable with the best predictive power was DSR. In terms of correlation, the relationship is there, as the tables five and six show. A strong correlation was shown between the consumer debt service (DSR) and changes in the rate for new delinquencies in credit card debt. The CDS statistically explained 55% of the variation in new credit card delinquencies.

**Table 5**

*Results from correlation analysis*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Simple Statistics** | | | | | | | |
| **Variable** | **N** | **Mean** | **Std Dev** | **Sum** | **Minimum** | **Maximum** | **Label** |
| **CDS\_RATIO** | 21 | 5.54683 | 0.15753 | 116.48343 | 5.08761 | 5.75578 | CDS\_RATIO |
| **NEW\_DEL\_CC\_PER** | 21 | 6.11116 | 0.62915 | 128.33440 | 5.07259 | 6.95000 | NEW\_DEL\_CC\_PER |

**Table 6**

*Results from Correlation*

|  |  |  |
| --- | --- | --- |
| **Pearson Correlation Coefficients, N = 21** **Prob > |r| under H0: Rho=0** | | |
|  | **CDS\_RATIO** | **NEW\_DEL\_CC\_PER** |
| **CDS\_RATIO CDS\_RATIO** | 1.00000 | 0.55754  0.0086 |
| **NEW\_DEL\_CC\_PER NEW\_DEL\_CC\_PER** | 0.55754  0.0086 | 1.00000 |

**Figure 6**

*Graph of Correlation of between Consumer Debt Raio and Percentage of New Deliquencies in Credit Card.*



Linear regression analysis established that CDS\_ratio could statistically significantly predict changes in percentages of new delinquencies in auto loans. The predictive equation would be -6.24042 + (2.2678 x Quarter)

The regression analysis showed the following when reviewing CDS and new\_del\_cc\_per:

**Table 7, 8 and 9**

*Regression Analysis Results*

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | 2.46087 | 2.46087 | 8.57 | 0.0086 |
| **Error** | 19 | 5.45576 | 0.28715 |  |  |
| **Corrected Total** | 20 | 7.91663 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 0.53586 | **R-Square** | 0.3108 |
| **Dependent Mean** | 6.11116 | **Adj R-Sq** | 0.2746 |
| **Coeff Var** | 8.76854 |  |  |

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | Intercept | 1 | -6.24042 | 4.22081 | -1.48 | 0.1557 |
| **CDS\_RATIO** | CDS\_RATIO | 1 | 2.22678 | 0.76065 | 2.93 | 0.0086 |

**Figure 7**

*Graph of Regression Analysis.*



**Conclusion**

Given the results, we can accept the alternative hypothesis and reject the null hypothesis. The variables FINOB and CDS\_RATIO are good predictors of potential delinquencies.

Banks can create their own versions of the ratios using the data they have and score the results. The scores can translate into risk categories. The banks can then act to minimize the risks and protect the solvency of the bank from unforeseen losses.

References

Herkenhoff, K. (2019, February). *The Impact of Consumer Credit Access on Unemployment.: Discovery Service for CSU - Global Campus*. Oclc.org; Review of Economic Studies. https://eds-a-ebscohost-com.csuglobal.idm.oclc.org/eds/pdfviewer/pdfviewer?vid=2&sid=b0ce0ee6-850a-4be0-8c95-9e06dd53d05c%40sessionmgr4007

Holt, S., & Lucas McKay, K. (n.d.). Consumer Debt: A Primer. In *aspeninstitute.org* (p. 4). Retrieved February 20, 2021, from https://www.aspeninstitute.org/wp-content/uploads/2018/03/ASPEN\_ConsumerDebt\_06B.pdf

*S.3548 - 116th Congress (2019-2020): CARES Act*. (2019). Congress.gov. https://www.congress.gov/bill/116th-congress/senate-bill/3548?q=%7B%22search%22%3A%5B%22Coronavirus+Relief+Aid%2C+Relief%2C+and+Economic+SEcurity%22%5D%7D&s=1&r=1

U.S. Department of Labor - Bureau of Labor Statistics. (2021, February 20). *Labor force statistics from the current population survey*. Bls.gov; U.S. Department of Labor. https://data.bls.gov/cgi-bin/surveymost

**Appendix A**

Data Definitions

Data was taken from Q3 of 2015 to Q3 of 2020. The following variables were used in the analysis.

**TDB\_AGE\_(RANGE)**

Total debt balance by age group

Age 1 = 18-29

Age 2 = 30-39

Age 3 = 40-49

Age 4 = 50-59

Age 5 = 60-69

Age 6 = 70 and over

**TDB\_TRIL\_(AUTO -CC)**

Total Debt Balance in trillions of dollars. Loan Category follows Tril\_

**NUM\_ACC\_(AUTO-CC)**

Number of Accounts in millions. Loan category follows ACC\_

**AVG\_BAL\_**

Average value of the credit type. Derived from dividing **TBD\_TRIL\_** / **NUM\_ACC\_.** Expressed in thousands

**AOR\_BILL\_**

Auto loan origination by credit score. In billions of dolars

1= Less than 620 credit score

2= Between 620 and 659 credit score

3= Between 660 and 719 credit score

4= Between 720 and 759 credit score

5= Above 760 credit score

**NEW\_DEL\_ (AUTO, CC) \_PER**

New delinquencies by loan type. This is a loan which previously was current. PER is the percent of new delinquencies.

**SER\_DEL\_AUTO\_PER**

Serious delinquency is an account that is 90 days or more delinquent.

**SER\_DEL\_AGE\_(range)**

Transition to serious delinquency by age range, percent. All debt

Age 1 = 18-29

Age 2 = 30-39

Age 3 = 40-49

Age 4 = 50-59

Age 5 = 60-69

Age 6 = 70 and over

**SER\_DEL\_AUTO\_AGE\_(range)**

Transition to serious delinquency by age range, percent. Auto loans

Age 1 = 18-29

Age 2 = 30-39

Age 3 = 40-49

Age 4 = 50-59

Age 5 = 60-69

Age 6 = 70 and over

**SER\_DEL\_CC \_AGE\_(range)**

Transition to serious delinquency by age range, percent. Credit cards

Age 1 = 18-29

Age 2 = 30-39

Age 3 = 40-49

Age 4 = 50-59

**SER\_DEL\_CC \_AGE\_(range)** *(cont.)*

Age 5 = 60-69

Age 6 = 70 and over

**BK\_AGE\_(range)**

Percent of new bankruptcies by age range

Age 1 = 18-29

Age 2 = 30-39

Age 3 = 40-49

Age 4 = 50-59

Age 5 = 60-69

Age 6 = 70 and over

**FINOB**

Financial obligations, seasonally adjusted, expressed as a percent. It includes rent payments on tenant-occupied property, auto lease payments, homeowners' insurance, and property tax payments.

**CDS\_RATIO**

Consumer debt service ratio expressed as a percent. The Consumer DSR is total quarterly scheduled consumer debt payments divided by total quarterly disposable personal income.

References

Buchholtz, K. (2020, September 22). *Infographic: 2021 to Set off Avalanche of Business Insolvencies.* Statistia. . https://www.statista.com/chart/22996/projected-increase-of-business-insolvencies-per-year/

Federal Reserve Bank. (2013). *Household Debt Service and Financial Obligations Ratios*. Federalreserve.gov. https://www.federalreserve.gov/releases/housedebt/about.htm

Federal Reserve Bank of New York. (2021). Quarterly report on household debt and credit. In *newyorkfed.org* (p. 15). https://www.newyorkfed.org/medialibrary/interactives/householdcredit/data/pdf/HHDC\_2020Q4.pdf

Herkenhoff, K. (2019, February 7). *The impact of consumer credit access on unemployment: Discovery service for CSU - Global Campus*. Oclc.org; Review of Economic Studies. https://eds-a-ebscohost-com.csuglobal.idm.oclc.org/eds/detail/detail?vid=4&sid=d7695bc9-beb6-41e6-8e6e-eb32cbdd8d0d%40sessionmgr4007&bdata=JnNpdGU9ZWRzLWxpdmU%3d#AN=139743878&db=bth

Holt, S., & Lucas McKay, K. (n.d.). Consumer Debt A Primer. In *aspeninstitute.org*. Retrieved February 19, 2021, from https://www.aspeninstitute.org/wp-content/uploads/2018/03/ASPEN\_ConsumerDebt\_06B.pdf

United States Department of Labor. (2021, January 5). *Labor Force Statistics from the Current Population Survey*. Bls.gov. https://beta.bls.gov/dataViewer/view