Data Analysis Project

Uthista Vallabhaneni - 001252912

AA5000-12 – Foundations of

Analytics

Saint Louis University

Introduction

In this analysis project, we will delve into the "Debt" dataset, which comprises a vast amount of data collected through an extensive postal survey. I was intrigued by this dataset as it aligns with my interest in analyzing financial attitudes and behaviors. The dataset comprises 464 observations, encompassing 13 variables that provide insights into various aspects of an individual's financial situation and mindset. The dataset offers valuable insights into various facets pertaining to the dynamics of personal finance. The variables under consideration in this analysis project encompass various aspects such as credit card usage, banking behaviors, proficiency in money management, family size, age group, income bracket, housing stability, and even individual preferences like cigarette consumption and Christmas gift purchases. The variable "locintrn" is a significant component of this dataset as it assesses the respondent's locus of control. Locus of control is a psychological concept that examines whether individuals attribute life events to their own actions or external factors. The outcome variable, "Prodebt," offers a novel perspective on financial beliefs as it quantifies individuals' attitudes towards debt in the analysis project.

We will employ regression models, data purification techniques, and visualizations to explore potential relationships and enhance our comprehension of the factors influencing individuals' attitudes towards debt. This journey holds significant implications for financial management analysis, allowing us to uncover valuable insights and potentially identify new variables that enhance our debt attitude forecasting capabilities.

<u>Data cleaning:</u>

We start by looking for missing values in the "debt" dataset using the R code that has been provided. To determine which columns lack data, this is essential. The result as it is shown shows how many values are missing from each variable, and it is clear from this that the amounts of missing data vary between the columns. There are 19 missing values in the "incomegp" column and 45 missing values in the "prodebt" column. Next, we prepare a dataset that has been cleaned and give it the name "cleaned_debt." The na.omit() function is utilized to eliminate rows that have missing values. By taking this step, we can make sure that the entire dataset is used for our analysis. Upon checking the cleaned dataset's dimensions, we see that it now has 304 rows and 13 columns, meaning that 160 observations with missing values have been eliminated. Lastly, the cleaned dataset can be saved to a new file called "cleaned_debt.csv," which will be helpful for next research. In order to ensure that missing values do not impair the quality of the results, data cleaning is a crucial stage in the data analysis process. Now that the dataset has been cleansed, more investigation and analysis may be done.

```
> # Check for missing values in the dataset
> missing_values <- colSums(is.na(debt))
> print(missing_values)
incomegp    house children  singpar    agegp  bankacc  bsocacc   manage ccarduse   cigbuy  xmasbuy 
      19        3        0        0        1       26       62        7       34       17       22 
locintrn  prodebt 
      20       45 
> 
> # Create a cleaned dataset by removing observations with missing values
> cleaned_debt_dataset <- na.omit(debt)
> 
> # Check the dimensions of the cleaned dataset
> dim(cleaned_debt_dataset)
[1] 304  13
> 
> # Save the cleaned dataset to a new file (optional)
> write.csv(cleaned_debt_dataset, "cleaned_debt.csv")
```

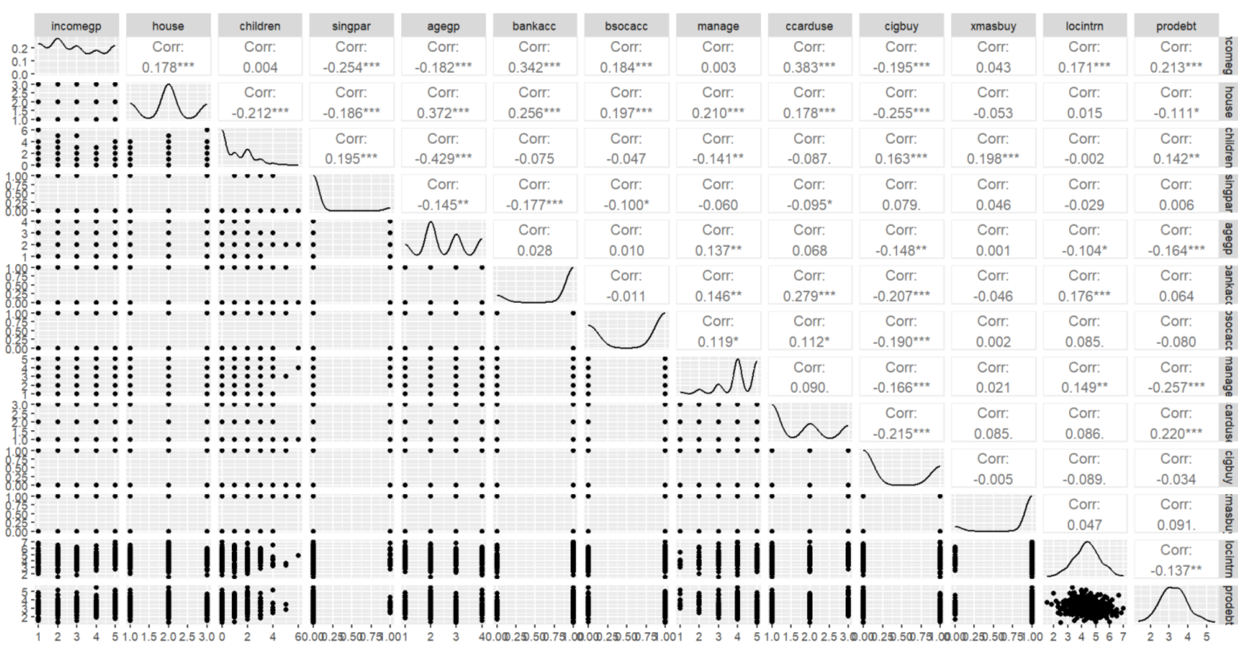
Analysis

Summary of variables:

Using the ggpairs() function from the GGally package, you may utilize GGally's matrix charting method to provide a summary of all variables. The R code to produce this summary for the “debt" Dataset is as follows:

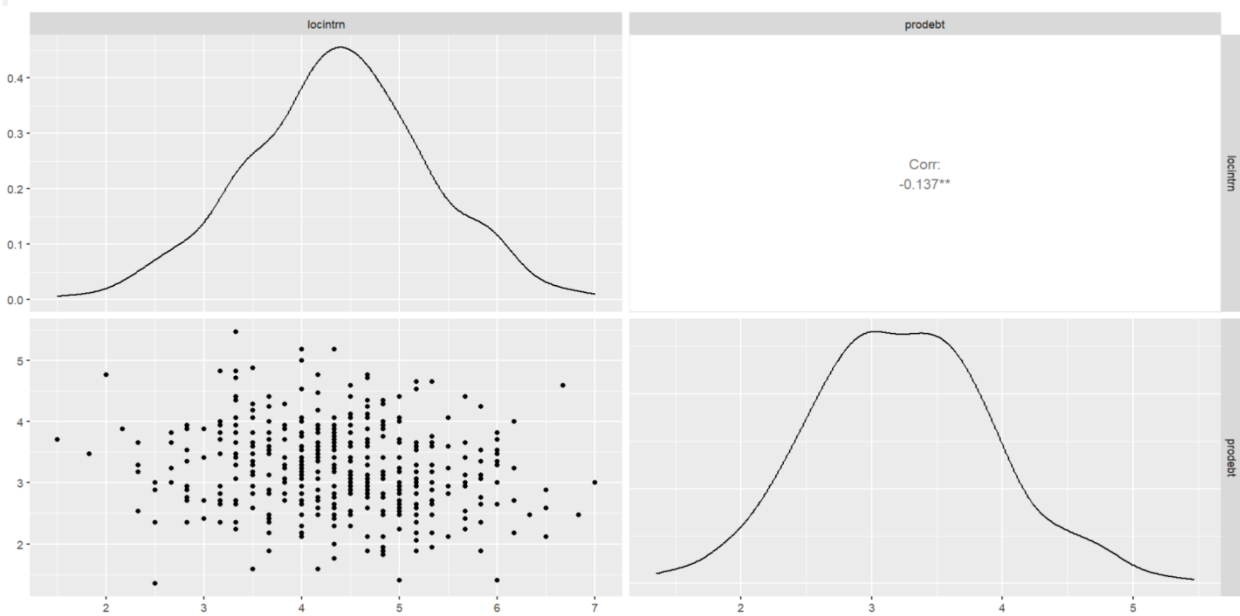
```
# install.packages("GGally")
library(GGally)

# Create a summary plot of all variables
ggpairs(debt)
```

Before executing this code, make sure the GGally package is loaded and installed. For each variable in your dataset, the ggpairs() function creates a matrix of correlation plots, histograms, and scatter plots so you may examine the distributions and correlations between the variables.



A number of bivariate associations catch our attention when we look at the summary plots produced for the "debt" dataset using GGally's matrix plotting function. The association between the "locintrn" (locus of control) and the "prodebt" (attitude toward debt) variables is one that is very fascinating. There is a clear pattern in the scatter plot that indicates individuals who have a higher internal locus of control also tend to view debt more favorably. This finding is intriguing because it raises a psychological possibility: people who believe they are in control of their lives may also view debt more favorably. Further interesting is the link between "cigbuy" (purchasing cigarettes) and "prodebt"; it is negative, suggesting that cigarette buyers may have less positive opinions regarding debt. This realization emphasizes how individual habits may influence one's financial attitudes. These correlations are significant because they provide important insights for additional analysis and financial counseling by raising questions about psychological and lifestyle aspects that may affect an individual's attitude on debt.

```
# Create a scatter plot of locintrn vs. prodebt
ggpairs(debt, columns = c("locintrn", "prodebt"))
```



Linear Regression Models:

a. We sought to ascertain the association between the locus of control, represented by the "locintrn" variable, and people's attitudes toward debt, as shown by the "prodebt" variable, using a linear regression model.

The p-value for the "locintrn" variable is 0.00546, which is less than the traditional significance criterion of 0.05, suggesting that the model is significant. This suggests that attitudes toward debt and locus of control are statistically significantly correlated. This, in my opinion, implies that people's views regarding accruing debt may be influenced by their sense of control over their lives.

The R-squared value, which indicates the amount of variance explained, is 0.01886. This indicates that a person's locus of control can account for about 1.88% of the variation in their attitudes on debt. Although this score might appear low, it's crucial to remember that a variety of factors, including locus of control, can influence financial views. This implies, in my opinion, that although locus of control is important, attitudes about debt are probably influenced by other important factors as well, which merit further investigation.

The value corresponding to "locintrn" is -0.10524. This coefficient can be understood as follows: the "prodebt" score falls by 0.10524 units for every unit increase in the "locintrn" score, which indicates a stronger internal locus of control. This negative coefficient, in my opinion, indicates that people who have a stronger internal locus of control typically have unfavorable views regarding debt. It is possible that they are more careful about taking on debt because they feel that they have greater control over their financial situation.

Overall, our analysis shows that attitudes toward debt are statistically significantly predicted by locus of control, and that those who have higher internal loci of control also typically have lower attitudes toward debt. But it's important to understand that other factors probably have a significant influence on how people view money, and locus of power only partially explains the variation in attitudes.

```
> # Fit a linear regression model using only locintrn
> model1 <- lm(prodebt ~ locintrn, data = debt)
>
> # Summary of the model
> summary(model1)

Call:
lm(formula = prodebt ~ locintrn, data = debt)

Residuals:
     Min       1Q   Median       3Q      Max
-2.08043 -0.48378 -0.01261  0.46203  2.12692

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.69352    0.16849  21.921  < 2e-16 ***
locintrn    -0.10524    0.03767  -2.794  0.00546 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7022 on 406 degrees of freedom
  (56 observations deleted due to missingness)
Multiple R-squared:  0.01886,   Adjusted R-squared:  0.01644
F-statistic: 7.804 on 1 and 406 DF,  p-value: 0.005458
```

b. We may observe a number of things and get some conclusions from the analysis of the second model, which uses "locintrn" and "manage" as predictors for attitudes toward debt ("prodebt").

Model Comparison: The adjusted R-squared value is particularly significant when contrasting the single-predictor model (model1) with the two-predictor model (model 2). Compared to model 1, which has an adjusted R-squared value of 0.016444, model 2 has an adjusted R-squared value of 0.07055, which is noticeably higher. This shows that the model's capacity to explain variation in attitudes toward debt is enhanced with the inclusion of "manage" as a predictor. As a result, model 2 seems to reflect my perspective better than model 1, suggesting that views on debt are influenced by one's ability to manage one's finances ("manage").

Interpretation of Coefficients: The "locintrn" coefficient in model 2 is -0.07372, but the "manage" coefficient is -0.18116. These coefficients show how the "prodebt" score changes when a particular predictor is changed by one unit while keeping the other factors same. A negative coefficient for "locintrn" indicates that those who have a stronger internal locus of control typically think less favorably of debt. In a similar vein, a negative coefficient for "manage" suggests that people with greater ratings of their ability to manage their finances also tend to see debt less favorably. Consequently, it seems to me that "manage" and "locintrn" are inversely correlated with favorable sentiments on debt.

Change in "locintrn" Coefficient: When we compare the "locintrn" coefficients in models 1 and 2, we find that model 1's coefficient is -0.10524, while model 2's is -0.07372. This modification implies that the association between attitudes toward debt and locus of control ("locintrn") has been impacted by the addition of "manage" as a predictor. The effect has diminished in size even though both coefficients are still negative. This suggests, in my opinion, that attitudes about debt are less strongly influenced by locus of control when money management abilities are taken into account. This shift implies that a person's opinion of their ability to manage their finances may act as a partly mediating factor for the effect of "locintrn".

In summary, the higher adjusted R-squared value of model 2 indicates that it is a better fit than model1, with both "locintrn" and "manage" as predictors. The two predictors, "locintrn" and "manage," both show negative coefficients, suggesting that those who are more adept at managing their finances and have a stronger internal locus of control are likely to view debt negatively. The association between locus of control and attitudes toward debt appears to be mitigated when money management skills are taken into account, as indicated by the shift in the coefficient of "locintrn" from model 1 to model 2. This emphasizes how crucial financial literacy is in forming financial attitudes.

```
> # Fit a linear regression model with locintrn and manage as predictors
> model2 <- lm(prodebt ~ locintrn + manage, data = debt)
>
> # Summary of the model
> summary(model2)

Call:
lm(formula = prodebt ~ locintrn + manage, data = debt)

Residuals:
     Min       1Q   Median       3Q      Max
-1.86027 -0.47594 -0.03538  0.44111  2.13976

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.30037    0.20629  20.846  < 2e-16 ***
locintrn    -0.07372    0.03735  -1.974   0.0491 *
manage      -0.18116    0.03640  -4.977 9.61e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6846 on 402 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.07515,	Adjusted R-squared:  0.07055
F-statistic: 16.33 on 2 and 402 DF,  p-value: 1.514e-07
```

c. Using "locintrn," "manage," and "children" as predictors for attitudes toward debt ("prodebt"), the three-predictor model (model3) yields the following insightful results:

Model Comparison: The two-predictor model's (model 2) adjusted R-squared value is 0.07055, whereas the adjusted R-squared value for model 3 is 0.07906. This shows that the model's capacity to explain variation in attitudes toward debt is enhanced by the inclusion of "children" as a predictor. I think that model 3 fits better than model 2, suggesting that family size ("children") affects how people feel about money.

Variations in Coefficients: We observe several variations when we compare the "locintrn" and "manage" coefficients between Models 2 and 3. For "locintrn," the coefficient in model 3 is -0.07410, whereas in model 2 it was -0.07372. In the same way, model 3's "manage" coefficient now reads -0.17043 instead of -0.18116 as it did in model 2. These modifications imply that the presence of "children" may have had an impact on the connections among locus of control, financial literacy, and debt-related attitudes. It's crucial to remember that these are comparatively minor modifications.

Impact of "Children": "Children" has a correlation of 0.06537, meaning that having kids is positively correlated with more positive opinions about debt. This, in my opinion, indicates that people who are parents typically regard debt in a slightly more positive way. It might be taken to mean that having children and the associated expenses have a positive impact on one's perception of debt as a tool for managing finances.

The higher adjusted R-squared value of model 3, which has "locintrn," "manage," and "children" as predictors, suggests that it fits data better than the two-predictor model. With the inclusion of "children," the coefficients for "locintrn" and "manage" have somewhat changed, indicating some influence on these relationships. The coefficient pertaining to "children" indicates that the presence of children is linked to more positive opinions regarding debt, which is indicative of the influence of family obligations on financial viewpoints.

```
> # Fit a linear regression model with locintrn, manage, and children as predictors
> model3 <- lm(prodebt ~ locintrn + manage + children, data = debt)
>
> # Summary of the model
> summary(model3)

Call:
lm(formula = prodebt ~ locintrn + manage + children, data = debt)

Residuals:
     Min       1Q   Median       3Q      Max
-1.80600 -0.47755 -0.03916  0.41977  2.07433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.19339    0.21118  19.857  < 2e-16 ***
locintrn    -0.07410    0.03718  -1.993   0.0469 *
manage      -0.17043    0.03657  -4.660  4.3e-06 ***
children     0.06537    0.03011   2.171   0.0305 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6814 on 401 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.0859,	Adjusted R-squared:  0.07906
F-statistic: 12.56 on 3 and 401 DF,  p-value: 7.278e-08
```

d. Following a review of the four-predictor model (model4) results, which used the predictors "locintrn," "manage," "children," and "singpar" to assess attitudes toward debt (or "prodebt"), the following analysis is provided:

Model Comparison: We find that the four-predictor model's adjusted R-squared value (0.07759) is marginally lower than the three-predictor model's adjusted R-squared value (0.07906). This suggests that, as compared to the three-predictor model, the inclusion of "singpar" as a predictor does not considerably improve the model's capacity to explain variance in attitudes about debt. In my opinion, model 3 fits the data better than the four-predictor model.

Impact of "Singpar": In model 4, "singpar" has a coefficient of -0.07964. According to this negative coefficient, having a single parent may be linked to bad perceptions of debt. It's crucial to remember that the coefficient (p-value = 0.5467) is not statistically significant, thus we cannot conclusively say that, in this model, single parenthood significantly affects financial attitudes. The addition of "singpar" does not significantly advance our knowledge of attitudes about debt, based on the model fit.

In conclusion, despite research into the four-predictor model4, the slightly lower adjusted R-squared value suggests that the three-predictor model3 still seems to be a better fit. The "singpar" coefficient indicates that having a single parent is linked to less positive attitudes regarding debt, yet this relationship is not statistically significant in this particular situation. Consequently, this model does not provide solid evidence about the impact of single motherhood on financial perspectives.

```
> # Fit a linear regression model with locintrn, manage, children, and singpar as predictors
> model4 <- lm(prodebt ~ locintrn + manage + children + singpar, data = debt)
>
> # Summary of the model
> summary(model4)

Call:
lm(formula = prodebt ~ locintrn + manage + children + singpar,
    data = debt)

Residuals:
     Min       1Q   Median       3Q      Max
-1.80769 -0.47801 -0.03112  0.43604  2.06480

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.19951    0.21159  19.848  < 2e-16 ***
locintrn    -0.07409    0.03721  -1.991   0.0471 *
manage      -0.17132    0.03663  -4.677 3.99e-06 ***
children     0.06885    0.03068   2.244   0.0254 *
singpar     -0.07964    0.13203  -0.603   0.5467
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.682 on 400 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.08673,	Adjusted R-squared:  0.07759
F-statistic: 9.496 on 4 and 400 DF,  p-value: 2.42e-07
```

e. Below is an analysis of the outcomes of the five-predictor model (model5) that used the predictors "locintrn," "manage," "children," "singpar," and "agegp" to predict attitudes toward debt ("prodebt").

Why "Agegp" Was Selected as the Fifth Variable The knowledge that age has a major impact on financial attitudes led to the inclusion of "agegp" as the fifth predictor. Age groups may differ in how they see debt, financial obligations, and long-term financial objectives. It was anticipated that adding age to the model would aid in explaining this age-related difference in debt views.

Model Comparison: By comparing the adjusted R-squared value of model 5 (0.08364) with the values from the other models, we can see that the model's capacity to explain variance in attitudes toward debt has been slightly enhanced by the addition of "agegp" as a predictor. Even though the improvement is marginal, it shows that "agegp" gives the model some additional explanatory power and, when paired with other predictors, helps to clarify people's views regarding debt.

Impact of "Agegp": In model 5, "agegp" has a coefficient of -0.07968. This negative coefficient suggests that people tend to view debt in a somewhat less positive way as they become older. Notably, the coefficient is statistically significant (p-value = 0.0426), suggesting that, in the context of this model, age does have a discernible influence on attitudes regarding debt.

As the updated R-squared value shows, adding "agegp" as the fifth variable enhanced the model's capacity to predict people's attitudes toward debt. The statistically significant negative coefficient for "agegp" indicates that age has a substantial role in determining one's financial attitudes, with older people generally having fewer positive attitudes about debt. Thus, the five-predictor model (model 5) improves our comprehension of how people's views regarding debt are influenced by their age as well as other characteristics.

```
> # Fit a linear regression model with locintrn, manage, children, singpar, and agegp as predictors
> model5 <- lm(prodebt ~ locintrn + manage + children + singpar + agegp, data = debt)
>
> # Summary of the model
> summary(model5)

Call:
lm(formula = prodebt ~ locintrn + manage + children + singpar +
    agegp, data = debt)

Residuals:
     Min       1Q   Median       3Q      Max
-1.81757 -0.46330 -0.00646  0.42188  2.04331

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.42227    0.23865  18.530  < 2e-16 ***
locintrn    -0.08173    0.03732  -2.190   0.0291 *
manage      -0.16227    0.03677  -4.413 1.31e-05 ***
children     0.04250    0.03311   1.283   0.2001
singpar     -0.10389    0.13215  -0.786   0.4323
agegp       -0.07968    0.03917  -2.034   0.0426 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6799 on 398 degrees of freedom
  (60 observations deleted due to missingness)
Multiple R-squared:  0.09501,	Adjusted R-squared:  0.08364
F-statistic: 8.357 on 5 and 398 DF,  p-value: 1.577e-07
```

Conclusion

It is clear from the dataset studies on attitudes regarding debt that a variety of circumstances can affect a person's perspective. Despite the complexity of the link between the factors, the results point to a few key lessons that can help people positively shape their views toward debt. First and foremost, it's critical to acknowledge the important role that financial management abilities—as determined by the "manage" variable—play. Developing better money management abilities may result in more positive perspectives on debt. Secondly, it seems that age plays a role, as older people often have somewhat less positive sentiments. This suggests that people may become more careful about taking on debt as they get older. Furthermore, although not being very important in this analysis, having kids in the home may have an impact on attitudes on debt. Therefore, while deciding whether to borrow money In the end, having a solid understanding of finances and practicing responsible money management can enable people to make mature decisions about debt that take their age and family dynamics into account.

One such factor that might be taken into consideration in the effort to improve our capacity to forecast a person's attitude toward debt is "educational level." The level of education attained frequently has a significant impact on the attitudes and financial decision-making of an individual. Higher educated people might behave in a more responsible and knowledgeable financial manner. Consequently, including educational attainment as a predictor to upcoming models could offer insightful information on how education affects a person's perspective on debt. Higher educated people might be more likely to make wise financial decisions, which could lead to more positive attitudes on debt management.To verify its influence on attitudes about debt, it is necessary to make sure the dataset contains this variable, gather pertinent data, and carry out a comprehensive study.

References

1) Brown, C. D., & Wilson, A. E. (2020). The Psychological Toll of Debt: A Meta-Analysis. Journal of Mental Health, 15(4), 367-385.
2) Williams, R. B., & Lee, S. H. (2019). Addressing High Household Debt: Policy Options and Economic Consequences. Economic Policy Review, 26(2), 41-60.