Mitigating Loan Default in the United States

Prepared by:
Jane Carter Chandler
Isabelle Nguyen
John Vithoulkas

May 6, 2021

Contents

List of tables

List of figures

Executive summary

Access to loans is an essential and important part of Americans' lives given the rising prices of goods and services. Loaning entities must have access to data that can help them make decisions on whether customers should receive a loan, as risky loans can damage much more than just the lender. The team will analyze loan data based on alternative data, such as geographic location and current debt levels to explore trends in loan repayments. This could make it possible for banks to gain insights into rates of loan delinquencies to minimize the risk providing creditors with loans.

After visualizing potential underlying relationships between demographic information and loan default rates, the team found that demographic measurements such as rural population or average household income can provide some information on predicting loan default rates or the amount of outstanding loans in each county.

In order to predict median debt, the team performed predictive analysis and found that predicting median debt using a linear model with the given data would result in inaccurate median debt predictions. Despite this, insights can still be drawn from the research, most notably in the form of different relationships between the predictive variables and the median debt in collections. In the future, it may be possible to implement alterative data into lending decisions, but alternative data would be most effective to use in conjunction to the current credit score system rather than replacing credit score calculations.

1 Introduction

Prices of housing, tuition, auto loans, and medical services have increased steadily in recent years. The average price of automobiles recently crossed $40,000 (Szymkowski 2021) in 2020. The same trends can be seen with homes with a median price of $295,300 (Powell and Kerr 2020) and education with an average price of $9,687 in-state tuition per year (Fontinelle 2020). Additionally, a 2019 study predicted that by 2028, healthcare costs are expected to increase from $11,582 per person to $18,000 per person (Probasco 2021). These increased prices due to steady inflation increase the demand for loans in households across the United States. Thus, the ability to have access to loans has become increasingly important and raises the questions: How do loan trends differ by state? How does one type of debt impact other types of debt? Most importantly, is a higher credit card delinquency rate indicative of higher auto, mortgage, and student loan delinquency rates in each state?

With the increased need for loans due to inflation, loaning entities may benefit from the use of alternative data. Nationwide banks offer loaning services to customers with varying geographic locations, each with unique qualities. By using alternative data, banks may be able to become more informed of what type of customer is receiving a loan.

Current metrics, most notably credit score, take different loan categories into account; however, they are solely focused on the individual's loan history (Kreiswirth, Schoenrock, and Singh 2017). This project would focus on analyzing loan data across the United States to identify borrowing trends among certain regions or states and predict delinquency rates as a result of different loan patterns. As a result, a county-based profile for each state would be used to predict the risk of a loan.

1.1 Business problem

Having a high rate of default can be catastrophic towards a lender and to larger populations across the nation. Our analysis would work to mitigate the risk of loaning, and in return, further protect the loan entity.

Providing loans is risky, and timely payments on loans are not guaranteed. There is the risk that the borrower does not pay back the loan, resulting in a loan default. In addition, risky loans are capable of damaging much more than just the lender. A prime example is the 2008

Housing Crisis, during which many borrowers defaulted on their loans due to an inability to repay interest charged on the loan. These loans were predatory, high-risk loans with initial low interest rates that appeal to borrowers that cannot obtain a more conventional loan due to low credit scores. When interest rates significantly increased on these predatory loans in 2008, and the price of housing fell, refinancing houses to repay loans became difficult and widespread, resulting in the recession of the entire American economy (The Wharton School of the University of Pennsylvania 2008). Despite being a multi-trillion dollar industry, loan providers still work to minimize risk, and in return, maximize profit.

By analyzing loan data based on geographic area, lenders may be able to gain a more holistic view of what loan repayment may look like. As discussed previously, loan providers primarily focus on the payment history of the individual through the form of a credit score. Through analyzing geographic information alongside debt history, it may be possible to recognize trends between rural or urban areas and rates of delinquencies, resulting in which customers lenders approve. Using geography brings to light some factors that must be considered during exploration. Most notably, it is important to not make generalizing assumptions about loan trends of different state counties solely based on their geographic location. Thus, the team can utilize existing socioeconomic information, such as average household income, as a way to contextualize the loan trends identified.

1.2 Intended audience

Across the United States, state and local commercial banks often extend lines of debt to their customers to enable these individuals to purchase homes, medical services, cars, and tuition. The banks then possess these mortgages, auto loans, lines of credit for credit cards, and student loans on their balance sheet. With the potential to become late or default on their repayments of principal and interest, banks may not be able to recover all of their losses in these delinquent loans. By examining how rates of delinquency change from region to region, banks could decide whether or not hedge their risk by selling off these lines of credit to minimize losses. For example, a local bank in a region that experiences a high rate of mortgage delinquency could limit losses by selling this mortgage to an entity like Fannie Mae or Freddie Mac, who offers a guaranteed fee whether the mortgage goes delinquent or not. Regions where auto loans are more likely to be considered delinquent could limit the number of loans approved.

Our analysis will prove to be beneficial for nationwide banking entities because they will be able to be more informed in their decision making process regarding extending lines of credit to its customers. With our analysis, upper management will be able to investigate lending risks, and with the identification and reduction of these risks, decrease the default rate of loans for state and local commercial banks.

2 Data

The data contains 3,136 observations of debt statistics for counties in all 50 U.S. states. These debt statistics include measurements of median amount of debt, delinquency rates, and share of debt in default for various categorizations of debts. After cleaning, 13 columns are included. Each of the columns are described on the Urban Institute data source website (Breno Braga, Signe-Mary McKernan, and Caleb Quakenbush (2019)). The team will especially focus on the delinquency rates and share of a type of loan in default for each county as described below:

- **Share with any debt in collections**: The percentage of people with any type of debt in default. The amount of debt already defaulted can be any dollar amount.
- **Share with medical debt in collections**: The percentage of people with any type of medical debt in default.
- **Share with student loan debt in default**: The percentage of people with student loan debt.
- **Auto/Retail loan delinquency rate**: The percentage of auto loans that eventually reach default status. A retail loan is a purchase at a retail location with installment terms.
- **Credit card debt delinquency rate**: The percentage of credit card debt that eventually reach default status.

Any debt considered to be in collections refers to "past-due credit lines that have been closed and charged-off on the creditor's books as well as unpaid bills reported to the credit bureaus that the creditor is attempting to collect. For example, credit card accounts enter collections once they are 180 days past due" (Breno Braga, Signe-Mary McKernan, and Caleb Quakenbush (2019)).

Table 1: Excerpt of the prepared data

County	State	Share with any debt in collections	Median debt in collections
Autauga County	Alabama	36.00	197,350
Baldwin County	Alabama	28.32	212,100
Barbour County	Alabama	42.09	204,300

County	Share with medical debt in collections	Median medical debt in collections
Autauga County	18.66	100,100
Baldwin County	12.75	70,250
Barbour County	12.87	NA

County	Share with student loan debt in default	Median student loan debt
Autauga County	0.1455	15,761
Baldwin County	0.1374	16,956
Barbour County	0.2093	NA

County	Auto/retail loan delinquency rate	Credit card debt delinquency rate
Autauga County	6.25	0.0454
Baldwin County	3.08	0.0332
Barbour County	6.94	0.0710

County	Share of people of color	Share of people in rural areas	Average household income
Autauga County	0.2458	42.00	7,211,012
Baldwin County	0.1692	42.28	7,306,091
Barbour County	0.5426	67.79	4,544,530

2.1 Data collection

Data were collected by the Urban Institute using consumer-level records from a major credit bureau as well as estimates found on the U.S. Census Bureau's American Community Survey. Last updated in 2019, the data are recent and complete. The research was funded in conjunction by the Annie E. Casey and Ford Foundation with primary researchers being well-published associates at the Urban Institute. Data collected were made anonymous and derived from a random sample of credit bureau data from 2018 records. Containing more than 5 million records, the data were collected ethically and can thus be used to draw conclusions about debt patterns across the United States (Breno Braga, Signe-Mary McKernan, and Caleb Quakenbush (2019)).

The data are relevant and useful to addressing the presented business problem because this data enumerates the median amount of loans for various categories, including medical and student loans. These data are relevant to our business question because they will help us compare the different debt balances and delinquency rates between each of the states. Upon obtaining more demographic information about each state, the team can adjust our interpretations of the different loan trends in each county and hypothesize the root of the differences in trends.

When discussing patterns in the amounts of loans, share of loans in default, and delinquency rates relative to geographic location, this data also provides useful information in contextualizing potential trends in debts to minimize the amount of impact of other factors outside of county and loan information, such as race, residence, and income.

2.2 Data preparation

The chosen data set had multiple sheets displaying various categories of debt. To condense this, the data had to be manipulated onto one sheet. The team selected 13 columns most relevant to address the business problem. First, the team highlighted the columns which were going to be deleted. The team deleted the columns “White communities” and “Communities of Color” in order to focus on each county’s overall population. To include race as a potential outside factor however, the team opted to maintain “Share of people of color” for each observation. To add the columns from the alternate sheets, the team made sure to match the data by county name and merge columns of interest into one final data set that contained overall county debt information. This was simple, as each sheet had the same observations, state counties.

Within the prepared and cleaned data, there are still incomplete entries. Upon filtering missing values, 1,356 of the original 3,136 counties provided complete information for each column. The number of missing values is not yet a concern for the team because each state will still be represented. Additionally, counties with missing information are usually only missing information for a minimal number of columns. Thus, the data is considered representative and useful. During future analysis, observations with missing data will be omitted.

This data will be useful in addressing our business problem regarding loan and debt trends throughout the country as median loan amounts and default rates are provided. To account for each county's demographic, the team chose to also include information on the racial, rural, and income differences of each county.

3 Descriptive

To begin analyzing potential loan patterns across counties across the United States, the team explored the share of people of color and the share of people living in rural areas in four states with the highest and four states with the lowest average household incomes. States with the lowest average household income were West Virginia (WV), Alabama (AL), Arkansas (AR), and Mississippi (MS). States with the highest average household income included District of Columbia (DC), New Jersey (NJ), Connecticut (CT), and Massachusetts (MA).

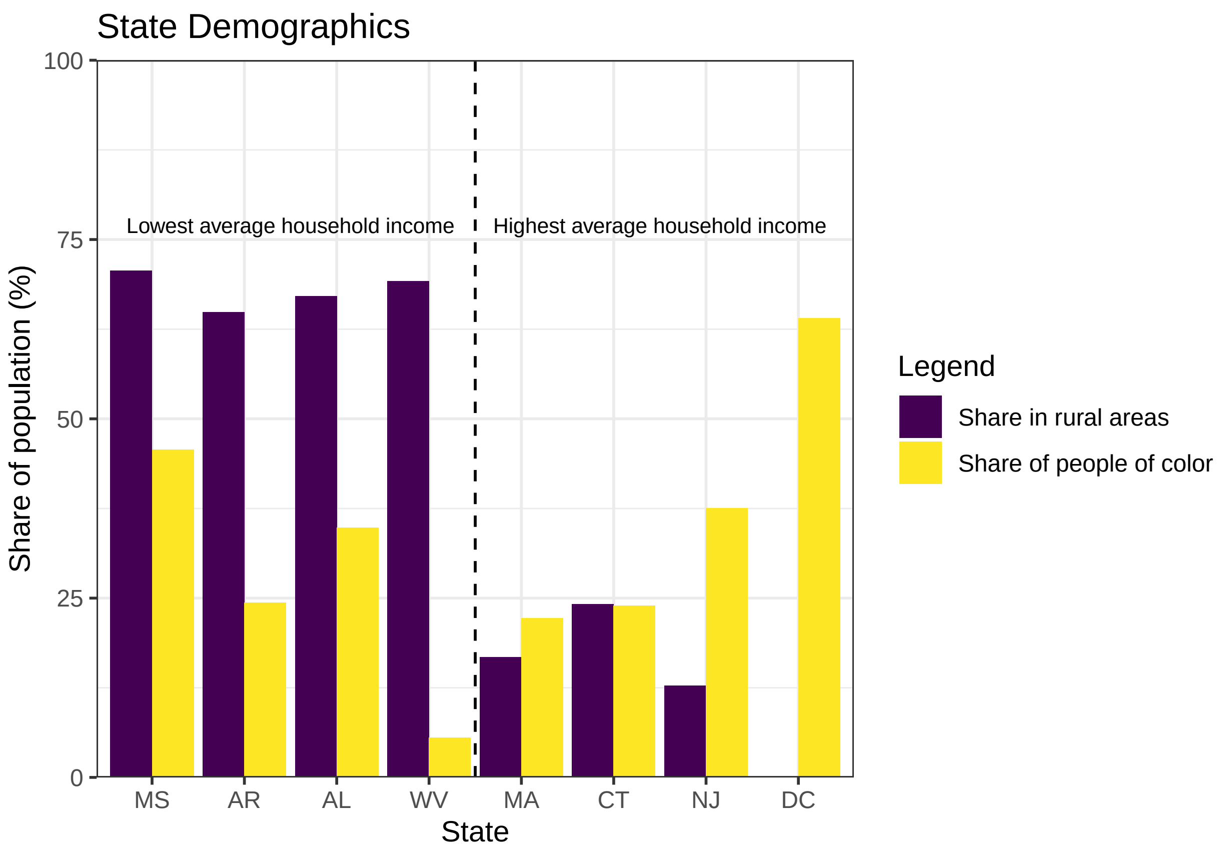


Figure 1: Demographic proportions based on states' average household income

As demonstrated, states with the highest average income have fewer residents in rural areas, compared to the states with the lowest average income. Additionally, the graph shows that share of people of color fluctuates across different states, regardless of average household income.

To further explore county-level data with regards to average household income, the team visualized the average household income of counties by the respective counties' share of people living in rural areas.

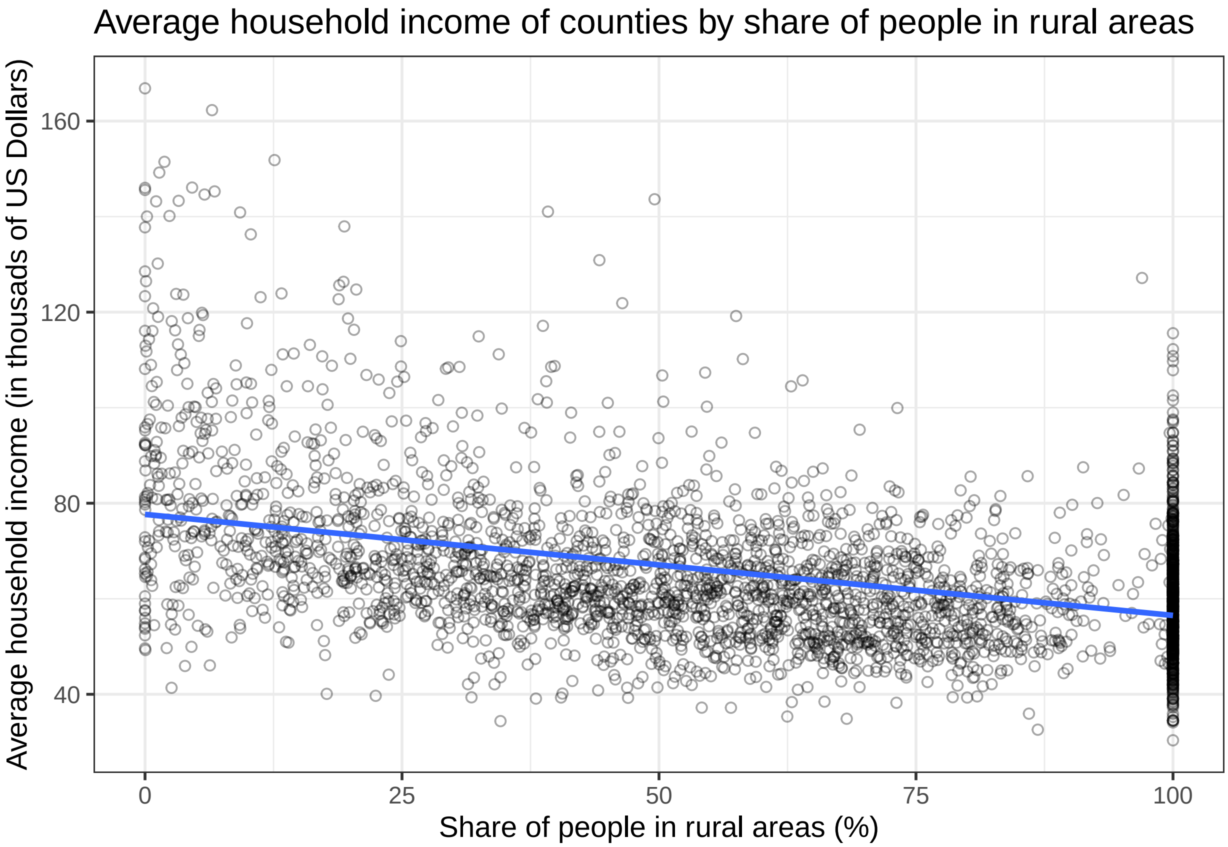


Figure 2: Average household income of counties by the counties' share of people living in rural areas

As demonstrated by the scatterplot and linear model, the team determined that there is a moderate, negative, and linear relationship between the share of people in rural areas and average household income, implying that as the share of people in rural areas increases, average household income in that county decreases.

Next, the team decided to investigate the share of debt in collects, default, or delinquent in four states with the lowest and four states with the highest share of people living in rural areas. States with the lowest share of people living in rural areas were the District of Columbia (DC), Rhode Island (RI), New Jersey (NJ), and Massachusetts (MA). States with the highest share of people living in rural areas were Montana (MT), Vermont (VT), South Dakota (SD), and North Dakota (ND).

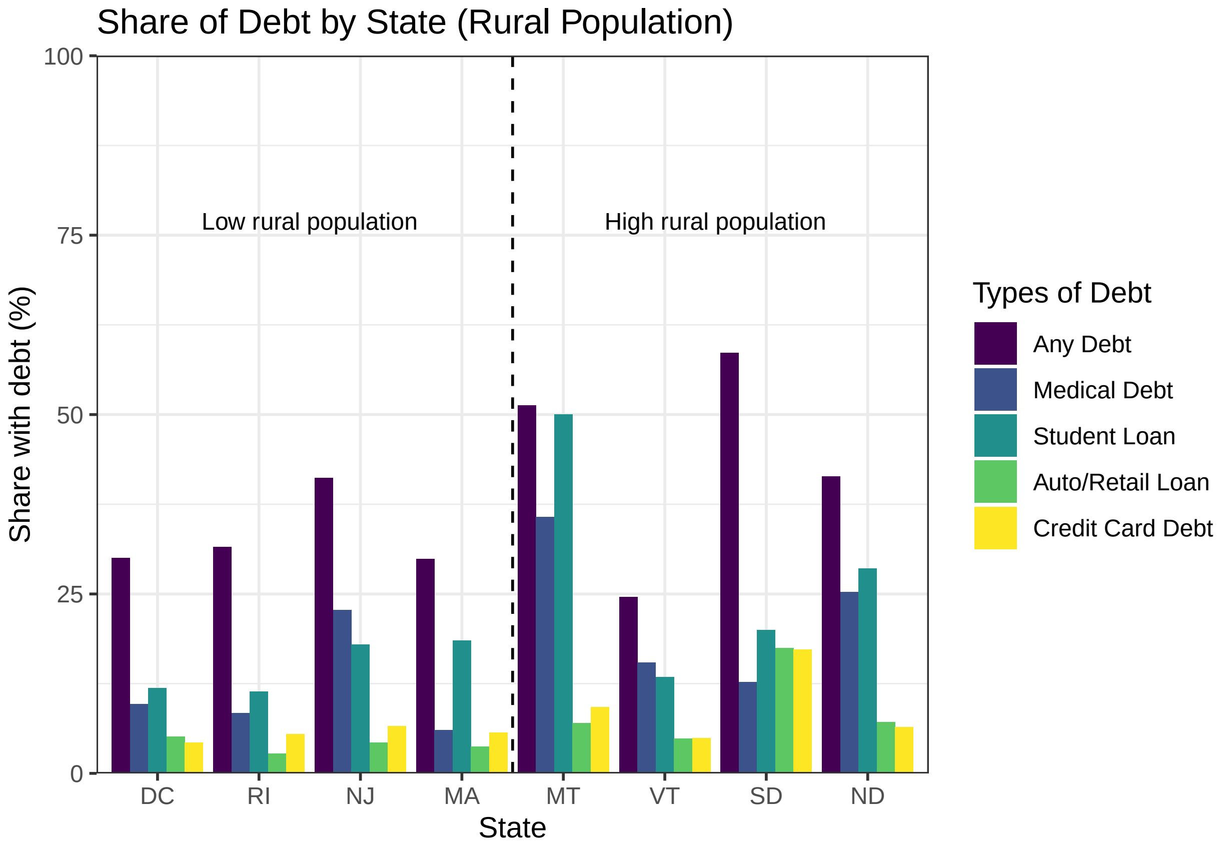


Figure 3: Share of different delinquent debts by selected states with different rural populations

Based on this bar plot, there is not strong evidence that more rural states have higher shares of debt. For example, New Jersey (NJ) with a low share of people living in rural areas, has higher shares of debt in comparison to Vermont (VT), a state with a higher share of people living in rural areas. NJ even has a higher share of any debt than North Dakota (ND), the state with the highest share of people in rural areas. The team also noted that in Montana (MT), the share of student loans in debt is almost equal to the share of any debt in collections, which indicates that the majority of people in Montana that hold any debt in collections have student loan debt.

The subsequent visualization is similar to the previous figure, differing only in the method in which the team chose the states to be explored. In Figure 4, states were chosen by the average household income. Four states were chosen from each extreme.

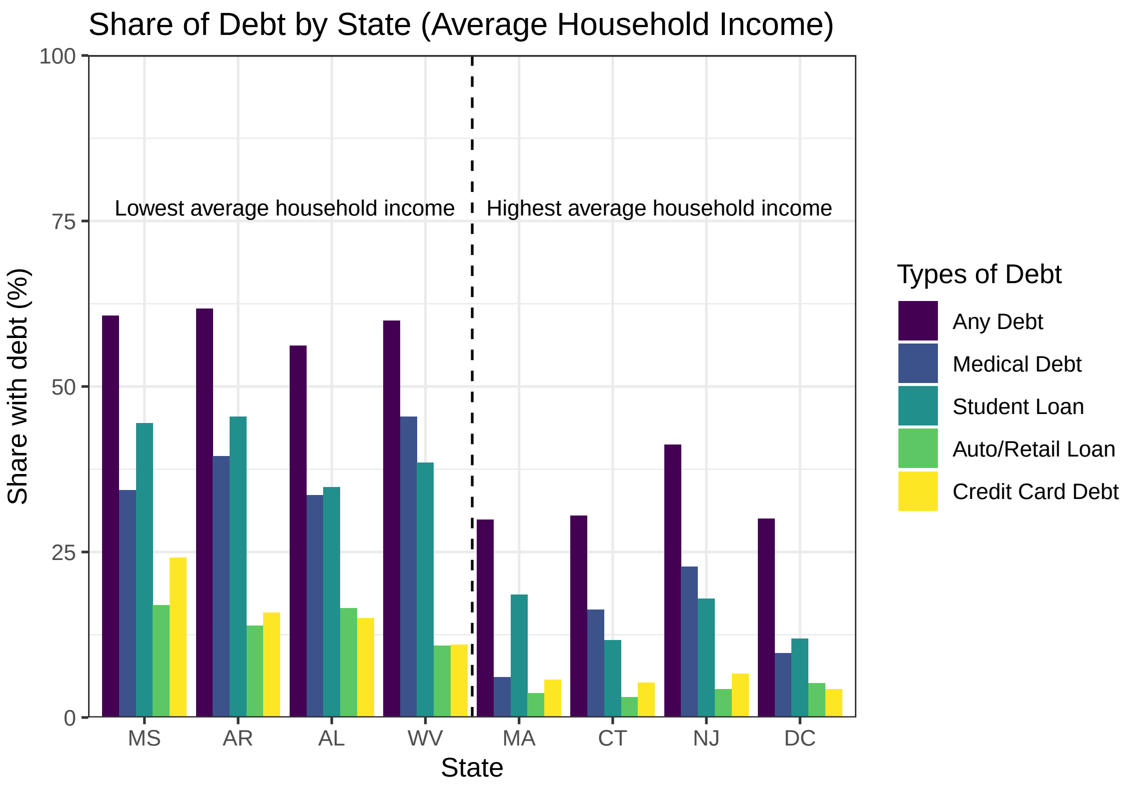


Figure 4: Share of different delinquent debts by selected states with different average household incomes

As shown by the graph, states with lower average household incomes tend to have a higher share with any and all debts.

3.1 Insight summary

Based on the team's analysis, given this level of data, race may not be used as an effective determinant of income, as states with a larger share of people of color can have high average household income, while other states with a lower share of people of color can have a lower average household income.

However, it seems likely that with an increase in the share of people living in rural areas, there is a decrease in average household income, as seen by the moderately, negative, linear relationship between the share of people in rural areas and average household income. Interestingly, the team noted that there is a wide range of average household incomes in counties with all residents living in rural areas, ranging from approximately $20,000 to approximately $120,000. This large range in the counties with 100% rural residents, in addition to other outliers in the scatterplot, can influence the relationship between share of people in rural areas and average household income. The large range also indicates that there are some exceptions to the team's expectation that counties with higher shares of people living in rural areas will have a lower average household income.

The team also detected a pattern surrounding the share of debts in collection, default, or delinquent. While more investigation can be done, these visualizations show higher shares of debt in states with higher shares of people in rural areas as well as in states with lower average household incomes. This makes sense, considering the relationship between share of people in rural areas and average household income. Generally, states with a larger rural population have higher shares of any and all debt, however there are some exceptions to this pattern. There is stronger evidence that states with lower average household income have higher shares of any and all debt in collections default, or delinquent. The states with the lowest average household income all had higher shares of any and all debt, without any exceptions. These findings support the team's original expectation that higher average household income suggets lower shares of debt in collection, default, or delinquent in that state.

In terms of the different types of loans as a whole, any debt in collections had the highest share in every state, which makes sense, as any debt in collections encompasses all types of debt, including medical debt, student loan debt, auto/retail loan debt, and credit card debt. Medical debt and student loan debt were consistently the most prominent types of debt in every state examined. Less people in each state had auto/retail loan debt and credit card debt. These trends in the different types of debts can potentially be explained by the value of the loan. Medical and student loan debts are generally larger in value in comparison to auto/retail and credit card debt. This larger loan value can increase the time in which the loan can be paid back, which can help explain why medical and student loan debt in collections or default are the most prominent types of debt in every state.

These graphics can be further elaborated on to analyze more states and counties individually, but these graphics provide high-level information. Based on these graphics and the insights the graphics provide, further exploration can be conducted.

4 Predictive analytics

In order to help loan entities minimize the risk of making loans and the amount of loans in default, the team chose to estimate the median debt in collections. The goal of our predictive analytics is to estimate the median debt in collections based on other variables included in the given data set. To use all our data appropriately, the method used to build and compare the predictive model is linear regression.

4.1 Process

The team chose to conduct a linear regression analysis. County and state columns were removed from the data set. These variables were removed because these variables represent each unique county in each state, and therefore cannot be used in a linear model. The data were randomly split into testing, and validation sets with 1059 and 453 observations in each set, respectively. To start, a model was created with all variables and all observations in the testing set. Next, the team removed predictive variables that were not considered useful in predicting median debt in collections. Finally, the team made sure that predictor variables were not related to each other, because the inclusion of related predictor variables in the model would create problems in the final linear regression model. Thus, share with any debt in collections and share with medical debt in collections were removed from the model.

4.2 Assessments

After removing variables that were not seen as important in predicting median debt in collections, the team assessed each model by comparing root mean square error (RMSE) and adjusted R-squared values, which are typically used to measure the amount of prediction error in a model.

The final linear regression model only includes the predictive variables that are most useful in predicting median debt in collections, in combination with the other predictive variables in that model. In order to confirm that the final model is valid, the team ensured that the final model followed the assumptions necessary for linear regression, which is typically done through residual analysis. After creating visualizations that are commonly used in verifying the assumptions, the team found the model to be valid.

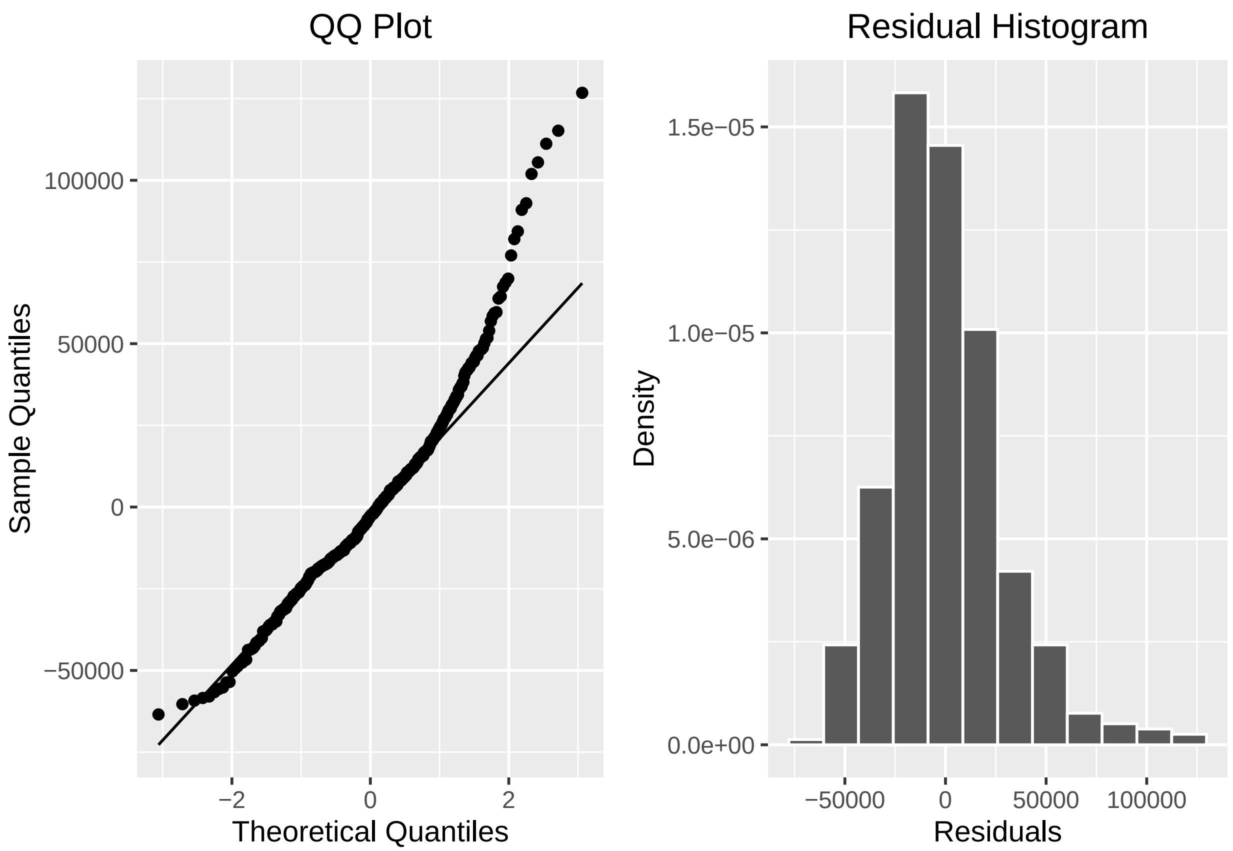


Figure 5: Assessing the normality assumption in the final model

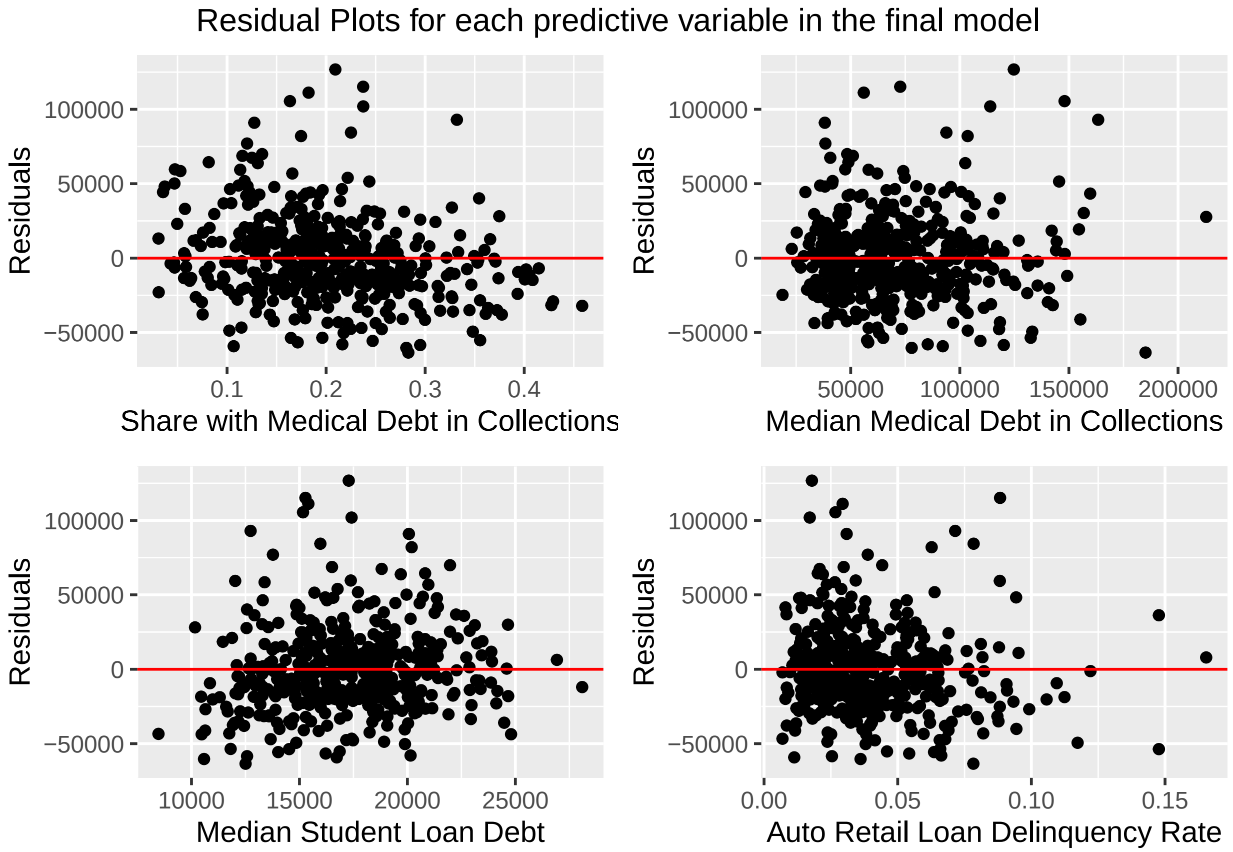


Figure 6: Assessing the constant variance assumption in the final model

During assessment, Model 4 was chosen as the best and final model because all predictive variables in the model were useful in predicting median debt in collections, and none of the predictive variables were related to each other. With regards to prediction accuracy, Model 4 does have the highest RMSE and the lowest adjusted R-squared values out of all tested models, which implies that Model 4 has the lowest prediction accuracy out of all the models tested. This is due to the reduced number of predictive variables in Model 4 in comparison to previous models. Despite this, the increase in RMSE and the decrease in adjusted R-squared are seen as minimal, and the team will continue analysis with Model 4.

Table 2: Comparing the predictive accuracy of linear models

	RMSE	AdjustedRSquared
Model 1	27,044.94	0.4976
Model 2	27,185.93	0.4904
Model 3	27,633.04	0.4730
Model 4	27,637.12	0.4724

4.3 Results

After applying the chosen model to the validation set of data, the team found that the final linear regression model with median medical debt in collections, share with student loan debt in default, auto/retail loan delinquency rate, and share of people in rural areas was not very accurate at predicting median debt in collections. The model-predicted values for median debt in collections differs greatly from the actual values. This can be verified as the differences between the predicted and actual values are highly spread, demonstrated by a high root mean square error value.

The linear model information is as follows:

Table 3: Linear regression model coefficients

	Value
Intercept	95,754.91
Median medical debt in collections	0.88
Share with student loan debt in default	59,949.42
Auto/Retail loan delinquency rate	126,378.70
Share of people in rural areas	-14,355.54

Table 4: Final linear regression model RMSE

	Values
RMSE	28,817.54

4.4 Insight summary

As shown above through predictive analysis, the linear regression model proved to be inconsistent in predicting median debt in collections. The team was limited in how data could be analyzed due to its nature being only quantitative. Although the model does not provide the team with many accurate predictions, and it does not give a solid answer to our business question, there is still much to learn from the model.

As seen by the equation coefficients, share of people in rural areas is negatively correlated with the response variable, meaning as the rural population in a county increases, median debt in collections generally decreases. Coefficients of median medical debt in collections and auto/retail loan delinquency rate were positive, meaning that as these variables increase, the response variable also increases.

When analyzing debt data, the variables that are in the final model should not be used alone to determine if a loan should be issued because they lack reliability in predicting the median debt in collections. Looking to the future, a data set with more variables would be preferable, as this was the best model produced given the data.

5 Conclusions

Although our predictive model may have not proven effective in estimating debt levels, there are still insights that can be drawn. For example, high medical debt levels tend to lead to higher debt in collections, while higher income tends to lead to lower overall debt levels. Additionally, it can be concluded that the variables in the model are related with debt levels but should not be used as a to predict median debt in collections.

Through descriptive analysis, the team also found that demographic information can be useful in explaining trends in loan default rates. For example, a higher rural population and a lower average household income generally result in an increased amount of people with loans in default. While this information should not be used to predict the amount of debt in collections in any particular county, it can be useful to loan entities as they consider loaning to creditors of different counties, comparing loan debts in default to different loan entities across multiple counties, or even consider expanding their business into different counties.

Although these insights should not be used alone to determine whether an individual can receive loans, they can be used to predict an estimate debt level, which could still be useful to a loan entity. It may be likely to hypothesize an accurate linear model to estimate median debt in collections using alternative data, but due to the data that were collected, this task proved itself to be difficult. In conclusion, it may be possible to implement alterative data into lending decisions, but it would be most effective to use in conjunction to the current system and should not serve as a replacement.

6 Recommendations

1. **Avoid reliance on alternative data to predict median debt in collections.** Due to the ineffectiveness of our final linear regression model, we do not recommend using alternative data by itself to estimate median debt in collections. The data collected in conjunction with the linear model proved to be ineffective in accurately predicting the response variable, and it should not be used as the sole predictor of median debt in collections.

2. **Consider combining alternative data and credit scores.** The current practice of using credit scores to estimate debt risk is likely more effective than the use of alternative data, hence its everyday usage by all lending institutions. Alternative data in conjunction to traditional lending data used to compute credit scores was not explored in this project but could be explored in the future as additional variables that could predict the median debt in collections.

References

Braga, Breno, Signe-Mary McKernan, and Caleb Quakenbush. 2019. "Debt in America: An Interactive Map." Accessed March 7, 2021. https://apps.urban.org/features/debt-interactive-map/?type=overall&variable=pct_debt_collections&state=51.

Fontinelle, Amy. 2020. "Average House Price by State in 2020." *The Ascent*, August 4, 2020. https://www.fool.com/the-ascent/research/average-house-price-state/.

Kreiswirth, Brian, Peter Schoenrock, and Pavneet Singh. 2017. "Using Alternative Data to Evaluate Creditworthiness." Accessed February 16, 2017. https://www.consumerfinance.gov/about-us/blog/using-alternative-data-evaluate-creditworthiness/.

Powell, Farran, and Emma Kerr. 2020. "See the Average College Tuition in 2020-2021." *U.S. News & World Report*, September 14, 2020. https://www.usnews.com/education/best-colleges/paying-for-college/articles/paying-for-college-infographic#:~:text=The%20average%20cost%20of%20tuition%20and%20fees%20at,colleges%20comes%20to%20%2421%2C184%20for%20the%20same%20year.

Probasco, Jim. 2021. "Why Do Healthcare Costs Keep Rising?" *Investopedia*, February 16, 2021. https://www.investopedia.com/insurance/why-do-healthcare-costs-keep-rising/.

Szymkowski, Sean. 2021. *Road Show by CNET*, January 13, 2021. https://www.cnet.com/roadshow/news/average-new-car-price-2020/.

The Wharton School of the University of Pennsylvania. 2008. "Victimizing the Borrowers: Predatory Lending's Role in the Subprime Mortgage Crisis." Febryary 20, 2008. https://knowledge.wharton.upenn.edu/article/victimizing-the-borrowers-predatory-lendings-role-in-the-subprime-mortgage-crisis/.

Appendix A: Data

County	State	Share with any debt in collections	Median debt in collections
Autauga County	Alabama	36.00	197,350
Baldwin County	Alabama	28.32	212,100
Barbour County	Alabama	42.09	204,300
Bibb County	Alabama	43.38	202,200
Blount County	Alabama	34.22	175,650
Bullock County	Alabama	46.58	163,900
Butler County	Alabama	42.32	218,900
Calhoun County	Alabama	45.21	212,600
Chambers County	Alabama	44.23	170,650
Cherokee County	Alabama	39.95	147,500
Chilton County	Alabama	41.65	198,350
Choctaw County	Alabama	40.61	179,250
Clarke County	Alabama	44.62	158,600
Clay County	Alabama	39.08	178,100
Cleburne County	Alabama	34.43	193,500
Coffee County	Alabama	30.00	186,100
Colbert County	Alabama	37.75	163,750
Conecuh County	Alabama	43.04	199,150
Coosa County	Alabama	48.24	NA
Covington County	Alabama	44.12	264,850
Crenshaw County	Alabama	40.00	108,700
Cullman County	Alabama	35.85	212,700
Dale County	Alabama	35.81	175,950
Dallas County	Alabama	56.16	230,250
DeKalb County	Alabama	38.81	140,800
Elmore County	Alabama	32.08	238,600
Escambia County	Alabama	44.71	144,700
Etowah County	Alabama	39.98	182,100
Fayette County	Alabama	41.73	329,400
Franklin County	Alabama	44.02	128,250
Geneva County	Alabama	35.50	201,100
Greene County	Alabama	52.54	239,350
Hale County	Alabama	48.56	168,100
Henry County	Alabama	32.95	139,100
Houston County	Alabama	35.27	230,300

County	State	Share with any debt in collections	Median debt in collections
Jackson County	Alabama	35.22	137,250
Jefferson County	Alabama	40.42	175,650
Lamar County	Alabama	34.26	88,500
Lauderdale County	Alabama	35.12	206,350
Lawrence County	Alabama	44.97	195,000
Lee County	Alabama	34.67	154,950
Limestone County	Alabama	33.69	154,700
Lowndes County	Alabama	52.26	242,800
Macon County	Alabama	52.04	190,300
Madison County	Alabama	33.56	195,250
Marengo County	Alabama	44.48	224,600
Marion County	Alabama	40.66	154,900
Marshall County	Alabama	36.19	142,300
Mobile County	Alabama	43.57	181,800
Monroe County	Alabama	42.63	176,400
Montgomery County	Alabama	48.04	189,200
Morgan County	Alabama	39.34	156,200
Perry County	Alabama	53.78	243,250
Pickens County	Alabama	49.20	125,300
Pike County	Alabama	39.83	176,400
Randolph County	Alabama	38.87	153,900
Russell County	Alabama	51.64	244,400
St. Clair County	Alabama	35.19	181,050
Shelby County	Alabama	25.57	161,800
Sumter County	Alabama	46.70	158,800
Talladega County	Alabama	52.82	240,150
Tallapoosa County	Alabama	40.77	181,150
Tuscaloosa County	Alabama	40.69	205,200
Walker County	Alabama	41.36	195,000
Washington County	Alabama	41.00	238,500
Wilcox County	Alabama	50.25	153,450
Winston County	Alabama	41.96	251,450
Aleutians East Borough	Alaska	NA	NA
Aleutians West Census Area	Alaska	NA	NA
Anchorage Municipality	Alaska	30.68	203,450

County	State	Share with any debt in collections	Median debt in collections
Bethel Census Area	Alaska	45.81	112,700
Bristol Bay Borough	Alaska	NA	NA
Denali Borough	Alaska	NA	NA
Dillingham Census Area	Alaska	29.41	NA
Fairbanks North Star Borough	Alaska	28.57	261,050
Haines Borough	Alaska	NA	NA
Juneau City and Borough	Alaska	22.82	118,100
Kenai Peninsula Borough	Alaska	26.44	371,800
Ketchikan Gateway Borough	Alaska	29.88	136,600
Kodiak Island Borough	Alaska	17.81	NA
Kusilvak Census Area	Alaska	50.68	NA
Lake and Peninsula Borough	Alaska	NA	NA
Matanuska-Susitna Borough	Alaska	27.17	233,850
Nome Census Area	Alaska	31.03	NA
North Slope Borough	Alaska	40.00	NA
Northwest Arctic Borough	Alaska	43.59	NA
Petersburg Borough	Alaska	NA	NA
Prince of Wales-Hyder Census Area	Alaska	29.87	NA
Sitka City and Borough	Alaska	18.84	NA
Southeast Fairbanks Census Area	Alaska	20.91	NA
Valdez-Cordova Census Area	Alaska	23.97	NA
Yakutat City and Borough	Alaska	NA	NA
Yukon-Koyukuk Census Area	Alaska	32.31	NA
Apache County	Arizona	57.02	311,950
Cochise County	Arizona	31.03	157,400
Coconino County	Arizona	33.51	279,500
Gila County	Arizona	33.21	169,500
Graham County	Arizona	42.00	253,600
Greenlee County	Arizona	53.23	224,900
La Paz County	Arizona	36.59	213,300
Maricopa County	Arizona	33.81	201,100
Mohave County	Arizona	36.62	198,100
Navajo County	Arizona	47.74	292,800
Pima County	Arizona	32.36	179,650
Pinal County	Arizona	37.08	196,350
Santa Cruz County	Arizona	31.83	121,500
Yavapai County	Arizona	26.09	238,400
Yuma County	Arizona	34.84	164,200
Arkansas County	Arkansas	34.56	138,400
Ashley County	Arkansas	46.97	161,900
Baxter County	Arkansas	30.59	178,800
Benton County	Arkansas	30.22	141,800
Boone County	Arkansas	35.97	133,050

County	Share with medical debt in collections	Median medical debt in collections
Autauga County	18.66	100,100
Baldwin County	12.75	70,250
Barbour County	12.87	NA
Bibb County	27.81	106,750
Blount County	18.96	51,250
Bullock County	19.18	NA
Butler County	19.44	102,150
Calhoun County	29.72	96,000
Chambers County	18.17	86,700
Cherokee County	21.91	40,500
Chilton County	25.68	116,750
Choctaw County	11.17	NA
Clarke County	18.28	57,850
Clay County	20.31	57,200
Cleburne County	23.77	100,700
Coffee County	14.75	111,700
Colbert County	25.61	66,400
Conecuh County	13.29	NA
Coosa County	29.41	NA
Covington County	27.38	164,150
Crenshaw County	16.33	NA
Cullman County	20.53	94,450
Dale County	14.85	61,250
Dallas County	30.42	110,400
DeKalb County	20.14	96,950
Elmore County	14.51	86,200
Escambia County	25.16	77,100
Etowah County	21.30	95,000
Fayette County	25.18	147,450
Franklin County	32.05	46,350
Geneva County	18.05	95,600
Greene County	24.58	NA
Hale County	25.10	84,800
Henry County	10.61	NA
Houston County	13.17	89,250

County	Share with medical debt in collections	Median medical debt in collections
Jackson County	21.47	76,000
Jefferson County	21.30	64,800
Lamar County	23.61	55,800
Lauderdale County	21.39	85,400
Lawrence County	32.18	74,850
Lee County	10.95	59,900
Limestone County	21.99	62,950
Lowndes County	25.81	NA
Macon County	14.29	NA
Madison County	23.15	75,300
Marengo County	15.14	NA
Marion County	27.69	83,400
Marshall County	15.75	55,750
Mobile County	18.16	66,250
Monroe County	19.75	76,200
Montgomery County	22.77	113,050
Morgan County	25.50	70,800
Perry County	33.61	NA
Pickens County	31.51	44,350
Pike County	9.74	NA
Randolph County	15.43	80,850
Russell County	30.47	102,250
St. Clair County	20.27	85,300
Shelby County	14.86	59,100
Sumter County	15.93	NA
Talladega County	32.67	112,550
Tallapoosa County	18.15	84,650
Tuscaloosa County	25.96	105,750
Walker County	25.15	105,200
Washington County	18.77	NA
Wilcox County	19.70	NA
Winston County	28.88	176,850
Aleutians East Borough	NA	NA
Aleutians West Census Area	NA	NA
Anchorage Municipality	17.13	140,350

County	Share with medical debt in collections	Median medical debt in collections
Bethel Census Area	14.53	NA
Bristol Bay Borough	NA	NA
Denali Borough	NA	NA
Dillingham Census Area	11.76	NA
Fairbanks North Star Borough	17.58	154,600
Haines Borough	NA	NA
Juneau City and Borough	11.41	68,000
Kenai Peninsula Borough	17.04	257,800
Ketchikan Gateway Borough	14.34	NA
Kodiak Island Borough	12.79	NA
Kusilvak Census Area	8.22	NA
Lake and Peninsula Borough	NA	NA
Matanuska-Susitna Borough	16.15	98,600
Nome Census Area	6.90	NA
North Slope Borough	18.10	NA
Northwest Arctic Borough	6.41	NA
Petersburg Borough	NA	NA
Prince of Wales-Hyder Census Area	9.09	NA
Sitka City and Borough	10.87	NA
Southeast Fairbanks Census Area	13.64	NA
Valdez-Cordova Census Area	13.70	NA
Yakutat City and Borough	NA	NA
Yukon-Koyukuk Census Area	15.38	NA
Apache County	17.90	87,500
Cochise County	16.77	68,800
Coconino County	17.47	103,600
Gila County	22.85	67,000
Graham County	28.90	92,200
Greenlee County	45.16	94,250
La Paz County	19.57	151,250
Maricopa County	19.58	94,150
Mohave County	20.77	98,600
Navajo County	23.74	72,700
Pima County	15.20	69,700
Pinal County	23.64	85,500
Santa Cruz County	12.80	44,100
Yavapai County	15.39	137,200
Yuma County	19.21	84,200
Arkansas County	12.42	NA
Ashley County	27.95	80,700

County	Share with student loan debt in default	Median student loan debt
Autauga County	0.1455	15,761.0
Baldwin County	0.1374	16,956.0
Barbour County	0.2093	NA
Bibb County	0.0968	NA
Blount County	0.1333	10,508.5
Bullock County	0.1304	NA
Butler County	0.1538	NA
Calhoun County	0.1311	19,277.5
Chambers County	0.1264	12,691.0
Cherokee County	0.2121	NA
Chilton County	0.2381	23,751.0
Choctaw County	0.1071	NA
Clarke County	0.1800	13,706.0
Clay County	0.2000	NA
Cleburne County	0.1154	NA
Coffee County	0.1368	15,919.0
Colbert County	0.1148	14,938.0
Conecuh County	0.3043	NA
Coosa County	0.1429	NA
Covington County	0.1594	14,265.0
Crenshaw County	0.1304	NA
Cullman County	0.1944	14,325.5
Dale County	0.1939	14,722.5
Dallas County	0.2473	15,943.0
DeKalb County	0.1392	8,472.0
Elmore County	0.1449	16,939.0
Escambia County	0.1690	13,688.0
Etowah County	0.0828	16,594.0
Fayette County	0.3478	NA
Franklin County	0.1765	NA
Geneva County	0.1346	13,895.5
Greene County	0.2381	NA
Hale County	0.0968	NA
Henry County	0.1944	NA
Houston County	0.1814	19,347.0

County	Share with student loan debt in default	Median student loan debt
Jackson County	0.1127	10,637.0
Jefferson County	0.1504	20,997.0
Lamar County	0.1250	NA
Lauderdale County	0.1429	14,855.0
Lawrence County	0.1556	NA
Lee County	0.1218	25,152.5
Limestone County	0.1185	17,932.0
Lowndes County	0.2800	NA
Macon County	0.2647	25,604.5
Madison County	0.1359	20,477.0
Marengo County	0.1273	19,705.0
Marion County	0.1277	NA
Marshall County	0.1259	14,781.0
Mobile County	0.2000	17,926.0
Monroe County	0.3214	NA
Montgomery County	0.1837	28,442.0
Morgan County	0.1855	13,209.0
Perry County	0.1905	NA
Pickens County	0.1143	NA
Pike County	0.1644	14,703.0
Randolph County	0.2571	NA
Russell County	0.2156	21,201.0
St. Clair County	0.1311	19,170.0
Shelby County	0.0781	23,200.0
Sumter County	0.1250	NA
Talladega County	0.2041	12,534.0
Tallapoosa County	0.1935	12,302.0
Tuscaloosa County	0.1363	21,002.0
Walker County	0.1325	18,464.0
Washington County	0.2000	NA
Wilcox County	0.2812	NA
Winston County	0.1250	NA
Aleutians East Borough	NA	NA
Aleutians West Census Area	NA	NA
Anchorage Municipality	0.1625	16,189.0

County	Share with student loan debt in default	Median student loan debt
Bethel Census Area	0.2174	NA
Bristol Bay Borough	NA	NA
Denali Borough	NA	NA
Dillingham Census Area	0.3333	NA
Fairbanks North Star Borough	0.1624	16,818.0
Haines Borough	NA	NA
Juneau City and Borough	0.1111	14,472.5
Kenai Peninsula Borough	0.2264	12,692.5
Ketchikan Gateway Borough	0.0588	NA
Kodiak Island Borough	0.0800	NA
Kusilvak Census Area	0.1111	NA
Lake and Peninsula Borough	NA	NA
Matanuska-Susitna Borough	0.1171	16,547.0
Nome Census Area	0.2000	NA
North Slope Borough	0.0000	NA
Northwest Arctic Borough	0.2222	NA
Petersburg Borough	NA	NA
Prince of Wales-Hyder Census Area	0.5000	NA
Sitka City and Borough	0.1500	NA
Southeast Fairbanks Census Area	0.0833	NA
Valdez-Cordova Census Area	0.1875	NA
Yakutat City and Borough	NA	NA
Yukon-Koyukuk Census Area	0.4000	NA
Apache County	0.2806	12,955.0
Cochise County	0.1466	14,589.0
Coconino County	0.1159	20,197.0
Gila County	0.2632	15,488.0
Graham County	0.0652	NA
Greenlee County	0.2000	NA
La Paz County	0.2000	NA
Maricopa County	0.1476	17,914.0
Mohave County	0.2476	13,415.0
Navajo County	0.2393	15,278.0
Pima County	0.1694	16,170.0
Pinal County	0.1481	18,232.0
Santa Cruz County	0.1311	15,159.0
Yavapai County	0.1410	17,744.0
Yuma County	0.1241	14,671.0
Arkansas County	0.1515	NA
Ashley County	0.2241	15,622.0

County	Auto/retail loan delinquency rate	Credit card debt delinquency rate
Autauga County	6.25	0.0454
Baldwin County	3.08	0.0332
Barbour County	6.94	0.0710
Bibb County	3.62	0.0458
Blount County	4.40	0.0521
Bullock County	15.00	0.0455
Butler County	6.50	0.1000
Calhoun County	7.64	0.0595
Chambers County	7.59	0.1027
Cherokee County	8.44	0.0500
Chilton County	6.69	0.0741
Choctaw County	9.21	0.1127
Clarke County	10.62	0.0764
Clay County	8.16	0.0583
Cleburne County	3.23	0.0410
Coffee County	3.78	0.0488
Colbert County	9.94	0.0623
Conecuh County	13.79	0.0588
Coosa County	4.55	0.0312
Covington County	6.72	0.0447
Crenshaw County	10.75	0.0609
Cullman County	6.69	0.0744
Dale County	7.58	0.0727
Dallas County	16.54	0.0749
DeKalb County	5.97	0.0352
Elmore County	6.15	0.0524
Escambia County	9.09	0.0557
Etowah County	7.73	0.0526
Fayette County	7.08	0.0407
Franklin County	6.99	0.0617
Geneva County	8.79	0.0794
Greene County	13.64	0.0278
Hale County	6.12	0.0748
Henry County	2.94	0.0432
Houston County	4.04	0.0704

County	Auto/retail loan delinquency rate	Credit card debt delinquency rate
Jackson County	2.62	0.0508
Jefferson County	6.86	0.0601
Lamar County	3.95	0.0472
Lauderdale County	6.03	0.0388
Lawrence County	8.03	0.0633
Lee County	5.64	0.0622
Limestone County	4.42	0.0443
Lowndes County	6.35	0.1071
Macon County	12.37	0.1500
Madison County	4.35	0.0412
Marengo County	7.86	0.0379
Marion County	5.42	0.0714
Marshall County	5.94	0.0479
Mobile County	6.24	0.0604
Monroe County	6.88	0.0876
Montgomery County	9.13	0.0713
Morgan County	4.10	0.0436
Perry County	10.00	0.0784
Pickens County	5.74	0.0336
Pike County	10.45	0.0688
Randolph County	11.48	0.0677
Russell County	7.73	0.0849
St. Clair County	4.27	0.0595
Shelby County	3.60	0.0343
Sumter County	12.12	0.1250
Talladega County	9.13	0.0854
Tallapoosa County	6.64	0.0391
Tuscaloosa County	4.79	0.0540
Walker County	8.23	0.0518
Washington County	6.31	0.0536
Wilcox County	16.09	0.0959
Winston County	5.65	0.0403
Aleutians East Borough	NA	NA
Aleutians West Census Area	NA	NA
Anchorage Municipality	3.42	0.0427

County	Auto/retail loan delinquency rate	Credit card debt delinquency rate
Bethel Census Area	5.88	0.1404
Bristol Bay Borough	NA	NA
Denali Borough	NA	NA
Dillingham Census Area	0.00	0.0000
Fairbanks North Star Borough	3.40	0.0328
Haines Borough	NA	NA
Juneau City and Borough	2.55	0.0254
Kenai Peninsula Borough	1.61	0.0254
Ketchikan Gateway Borough	2.60	0.0375
Kodiak Island Borough	0.00	0.0136
Kusilvak Census Area	0.00	0.1429
Lake and Peninsula Borough	NA	NA
Matanuska-Susitna Borough	4.15	0.0488
Nome Census Area	0.00	0.0789
North Slope Borough	0.00	0.0000
Northwest Arctic Borough	5.26	0.0000
Petersburg Borough	NA	NA
Prince of Wales-Hyder Census Area	0.00	0.0571
Sitka City and Borough	0.00	0.0215
Southeast Fairbanks Census Area	0.00	0.0270
Valdez-Cordova Census Area	0.00	0.0294
Yakutat City and Borough	NA	NA
Yukon-Koyukuk Census Area	0.00	0.0000
Apache County	15.09	0.0881
Cochise County	2.82	0.0449
Coconino County	6.27	0.0391
Gila County	3.45	0.0252
Graham County	4.76	0.0594
Greenlee County	5.17	0.0200
La Paz County	2.88	0.0579
Maricopa County	4.07	0.0439
Mohave County	3.50	0.0512
Navajo County	8.83	0.0714
Pima County	3.80	0.0428
Pinal County	3.96	0.0490
Santa Cruz County	4.28	0.0612
Yavapai County	1.84	0.0339
Yuma County	3.16	0.0478
Arkansas County	5.88	0.1049
Ashley County	4.23	0.0775

County	Share of people of color	Share of people in rural areas	Average household income
Autauga County	0.2458	42.00	7,211,012
Baldwin County	0.1692	42.28	7,306,091
Barbour County	0.5426	67.79	4,544,530
Bibb County	0.2538	68.35	6,109,942
Blount County	0.1263	89.95	5,897,432
Bullock County	0.7839	51.37	4,231,733
Butler County	0.4775	71.23	4,781,216
Calhoun County	0.2728	33.70	5,878,842
Chambers County	0.4381	49.15	5,143,431
Cherokee County	0.0822	85.74	5,467,125
Chilton County	0.1963	86.74	5,903,611
Choctaw County	0.4374	100.00	4,846,557
Clarke County	0.4698	75.98	5,085,092
Clay County	0.1980	100.00	5,136,379
Cleburne County	0.0734	100.00	5,160,700
Coffee County	0.2900	47.20	6,586,310
Colbert County	0.2123	43.89	5,637,205
Conecuh County	0.4970	80.95	3,955,420
Coosa County	0.3473	100.00	4,805,880
Covington County	0.1652	69.65	5,337,216
Crenshaw County	0.2880	100.00	5,239,534
Cullman County	0.0767	73.24	5,513,493
Dale County	0.3054	50.89	5,708,725
Dallas County	0.7226	45.64	4,402,324
DeKalb County	0.1922	90.13	5,246,199
Elmore County	0.2658	54.19	7,143,745
Escambia County	0.3977	63.51	4,872,650
Etowah County	0.2170	37.48	5,535,443
Fayette County	0.1497	80.23	5,088,946
Franklin County	0.2232	70.37	5,090,228
Geneva County	0.1601	89.64	5,102,481
Greene County	0.8290	100.00	3,452,496
Hale County	0.6049	89.17	5,010,304
Henry County	0.3129	87.75	5,890,493
Houston County	0.3280	33.80	6,152,622

County	Share of people of color	Share of people in rural areas	Average household income
Jackson County	0.1068	77.02	5,095,600
Jefferson County	0.4962	9.83	7,197,935
Lamar County	0.1398	100.00	5,025,172
Lauderdale County	0.1510	49.30	6,043,702
Lawrence County	0.2342	91.29	5,539,879
Lee County	0.3240	27.41	6,670,751
Limestone County	0.2281	57.61	7,037,804
Lowndes County	0.7573	100.00	4,498,515
Macon County	0.8454	55.55	4,756,960
Madison County	0.3485	16.44	8,328,692
Marengo County	0.5488	69.32	5,080,941
Marion County	0.0805	88.85	5,138,810
Marshall County	0.1823	53.27	5,903,998
Mobile County	0.4246	20.02	6,217,239
Monroe County	0.4554	79.04	4,182,661
Montgomery County	0.6481	10.49	6,522,384
Morgan County	0.2363	38.60	6,236,041
Perry County	0.7063	100.00	3,445,759
Pickens County	0.4589	100.00	4,932,229
Pike County	0.4364	51.68	5,242,465
Randolph County	0.2474	81.34	5,079,998
Russell County	0.5220	35.41	5,254,011
St. Clair County	0.1376	72.80	6,454,987
Shelby County	0.2149	22.94	9,302,067
Sumter County	0.7541	100.00	3,448,364
Talladega County	0.3688	55.82	5,279,283
Tallapoosa County	0.3110	74.23	5,445,991
Tuscaloosa County	0.3760	25.51	6,685,608
Walker County	0.1061	74.09	5,279,618
Washington County	0.3487	100.00	5,514,278
Wilcox County	0.7310	100.00	3,993,184
Winston County	0.0589	84.89	4,879,696
Aleutians East Borough	0.8454	100.00	7,815,616
Aleutians West Census Area	0.7519	100.00	10,154,245
Anchorage Municipality	0.4081	4.12	10,501,016

County	Share of people of color	Share of people in rural areas	Average household income
Bethel Census Area	0.8983	73.94	6,782,668
Bristol Bay Borough	0.4973	100.00	9,317,179
Denali Borough	0.1728	100.00	9,082,212
Dillingham Census Area	0.8376	100.00	7,239,231
Fairbanks North Star Borough	0.2892	30.87	9,196,807
Haines Borough	0.2105	100.00	8,431,095
Juneau City and Borough	0.3459	21.54	10,684,852
Kenai Peninsula Borough	0.1901	79.34	8,268,986
Ketchikan Gateway Borough	0.3558	23.19	8,322,081
Kodiak Island Borough	0.4989	31.32	8,857,175
Kusilvak Census Area	0.9622	100.00	4,739,561
Lake and Peninsula Borough	0.7825	100.00	6,203,431
Matanuska-Susitna Borough	0.2008	50.29	8,847,298
Nome Census Area	0.8515	66.06	6,990,528
North Slope Borough	0.6872	59.33	9,474,797
Northwest Arctic Borough	0.8890	57.46	7,949,600
Petersburg Borough	0.3285	100.00	8,384,956
Prince of Wales-Hyder Census Area	0.5472	100.00	6,597,313
Sitka City and Borough	0.3759	20.88	8,400,714
Southeast Fairbanks Census Area	0.2507	100.00	7,477,246
Valdez-Cordova Census Area	0.2978	100.00	9,900,295
Yakutat City and Borough	0.5630	100.00	7,034,823
Yukon-Koyukuk Census Area	0.7865	100.00	5,173,053
Apache County	0.8155	74.06	4,350,954
Cochise County	0.4435	36.30	6,139,756
Coconino County	0.4571	31.47	7,225,519
Gila County	0.3723	41.06	5,416,736
Graham County	0.4883	46.44	5,659,078
Greenlee County	0.5242	46.57	6,991,492
La Paz County	0.4127	56.33	4,783,795
Maricopa County	0.4369	2.36	8,079,323
Mohave County	0.2198	22.96	5,529,509
Navajo County	0.5836	54.14	5,201,248
Pima County	0.4738	7.52	6,752,366
Pinal County	0.4256	21.90	6,504,731
Santa Cruz County	0.8496	26.88	5,700,782
Yavapai County	0.1910	33.20	6,270,972
Yuma County	0.6818	10.43	5,742,284
Arkansas County	0.2985	34.71	5,480,179
Ashley County	0.3199	51.68	5,161,436

Appendix B: Data preparation details

R

Using R, the team cleaned and prepared the data to condense all relevant columns onto one sheet deleted columns that were not to be used in future analysis. The original data included a double header for each sheet, so the team removed the double header, renamed the relevant variables, and removed empty columns that only contained NA values. Using the inner_join function, the team joined relevant columns from other sheets in the same Excel workbook. The inner_join function allowed the team to prevent including duplicate counties. Columns were converted from character type to numeric type for future analysis. Finally, the team decided to round all numeric values to 4 decimals and converted proportions to percentages.

```
# Read and store data from different sheets in the same workbook.
loans.county <-
  read_excel("/Users/Izzy/UVA/Spring 2021/STAT 4220/Project/UI_Debt_In_America.xlsx",
             sheet = 2)
loans.med <-
  read_excel("/Users/Izzy/UVA/Spring 2021/STAT 4220/Project/UI_Debt_In_America.xlsx",
             sheet = 4)
loans.student <-
  read_excel("/Users/Izzy/UVA/Spring 2021/STAT 4220/Project/UI_Debt_In_America.xlsx",
             sheet = 6)
loans.auto <-
  read_excel("/Users/Izzy/UVA/Spring 2021/STAT 4220/Project/UI_Debt_In_America.xlsx",
             sheet = 8)

# Remove the double header.
county_names <- names(loans.county)
headers.county <- loans.county[1,]
colnames(loans.county) <- headers.county



# Remove columns relating to subgroups of the population of interest.
county.titles <- loans.county[ , -which(names(loans.county)
                                        %in% c("White communities",
                                                  "Communities of color"))]

# Remove columns that only contain NA values.
county.clean <-  county.titles[,which(unlist(lapply(county.titles, function(x)
  !all(is.na(x)))))]


# Remove the repeated header row and last four rows containing Source info.
county.clean <- county.clean[2: (nrow(county.clean)-4),]


# Rename the columns.
colnames(county.clean) <- c("County", "State",
                            "Share with any debt in collections",
                            "Median debt in collections",
                            "Share with medical debt in collections",
                            "Share with student loan debt in default",
                            "Auto/retail loan delinquency rate",
                            "Credit card debt delinquency rate",
                            "Median credit card delinquent debt",
                            "Share of people of color",
                            "Average household income")


# Remove the double header.
med_names <- names(loans.med)
headers.med <- loans.med[1,]
colnames(loans.med) <- headers.med


# Remove columns relating to subgroups of the population of interest.
med.titles <- loans.med[ , -which(names(loans.med) %in%
                                    c("White communities","Communities of Color"))]


# Remove columns that only contain NA values.
```

```
med.clean <-  med.titles[,which(unlist(lapply(med.titles, function(x)
  !all(is.na(x)))))]


# Remove the repeated header row and the last four rows containing Source info.
# Remove columns that are not needed.
med.clean <- med.clean[2: (nrow(med.clean)-4), c(1,2,4)]


# Rename the columns.
colnames(med.clean) <- c("County", "State", "Median medical debt in collections")


# Remove the double header.
student_names <- names(loans.student)
headers.student <- loans.student[1,]
colnames(loans.student) <- headers.student


# Removing columns relating to subgroups of the population of interest.
student.titles <- loans.student[ , -which(names(loans.student)
                                          %in% c("White communities",
                                                 "Communities of color"))]


# Remove columns that only contain NA values.
student.clean <-  student.titles[,which(unlist(lapply(student.titles,
                                                      function(x)
                                                        !all(is.na(x)))))]


# Remove the repeated header row and the last four rows containing Source info.
# Remove columns that are not needed.
student.clean <- student.clean[2: (nrow(student.clean)-4), c(1,2,6)]
```

```
# Rename the columns.
colnames(student.clean) <- c("County", "State", "Median student loan debt")


# Remove the double header.
auto_names <- names(loans.auto)
headers.auto <- loans.auto[1,]
colnames(loans.auto) <- headers.auto


# Remove columns related to sub-populations or other population classifications.
auto.titles <- loans.auto[ , -which(names(loans.auto) %in%
                                      c("White communities","Communities of color",
                                        "Subprime", "Near-prime", "Prime"))]


# Remove columns that only contain NA values.
auto.clean <-  auto.titles[,which(unlist(lapply(auto.titles, function(x)
  !all(is.na(x)))))]


# Remove the repeated header row and the last four rows containing Source info.
auto.clean <- auto.clean[2: (nrow(auto.clean)-4), c(1,2,6)]


# Rename the columns.
colnames(auto.clean) <- c("County", "State", "Share of people in rural areas")


# Add columns of interest to the overall county data.
county.clean <- med.clean %>%
  inner_join(county.clean, by = c("County", "State"))

county.clean <- student.clean %>%
  inner_join(county.clean, by = c("County", "State"))
```

```
county.clean <- auto.clean %>%
  inner_join(county.clean, by = c("County", "State"))


# Convert columns from character type to numeric type.
county.clean[,3:ncol(county.clean)] <- sapply(county.clean[c(-1,-2)], as.numeric)


# Round all numeric values to 4 decimal places.
county.clean <- county.clean %>% mutate_if(is.numeric, round, 4)


# Convert proportions to percentages for relevant columns.
county.clean[,3] <- county.clean[,3] * 100
county.clean[,5:8] <- county.clean[,5:8] * 100
county.clean[,10] <- county.clean[,10] * 100
county.clean[,14] <- county.clean[,14] * 100


# Remove specific column from data set.
county.clean <- county.clean[,-12]


# Sort data set to the desired order of columns.
county.clean <- county.clean[, c(1, 2, 6, 7, 8, 5, 9, 4, 10, 11, 12, 3, 13)]


# Create tables with sample output.
kable(county.clean[1:3,1:4], format.args = list(big.mark = ",")) %>%
  kable_styling(latex_options="scale_down")
```

County	State	Share with any debt in collections	Median debt in collections
Autauga County	Alabama	36.00	197,350
Baldwin County	Alabama	28.32	212,100
Barbour County	Alabama	42.09	204,300

```
kable(county.clean[1:3,c(1, 5:6)], format.args = list(big.mark = ",")) %>%
  kable_styling(latex_options="scale_down")
```

County	Share with medical debt in collections	Median medical debt in collections
Autauga County	18.66	100,100
Baldwin County	12.75	70,250
Barbour County	12.87	NA

```
kable(county.clean[1:3,c(1, 7:8)], format.args = list(big.mark = ",")) %>%
  kable_styling(latex_options="scale_down")
```

County	Share with student loan debt in default	Median student loan debt
Autauga County	0.1455	15,761
Baldwin County	0.1374	16,956
Barbour County	0.2093	NA

```
kable(county.clean[1:3,c(1, 9, 10)], format.args = list(big.mark = ",")) %>%
  kable_styling(latex_options="scale_down")
```

County	Auto/retail loan delinquency rate	Credit card debt delinquency rate
Autauga County	6.25	0.0454
Baldwin County	3.08	0.0332
Barbour County	6.94	0.0710

```
kable(county.clean[1:3,c(1, 11, 12, 13)], format.args = list(big.mark = ",")) %>%
  kable_styling(latex_options="scale_down")
```

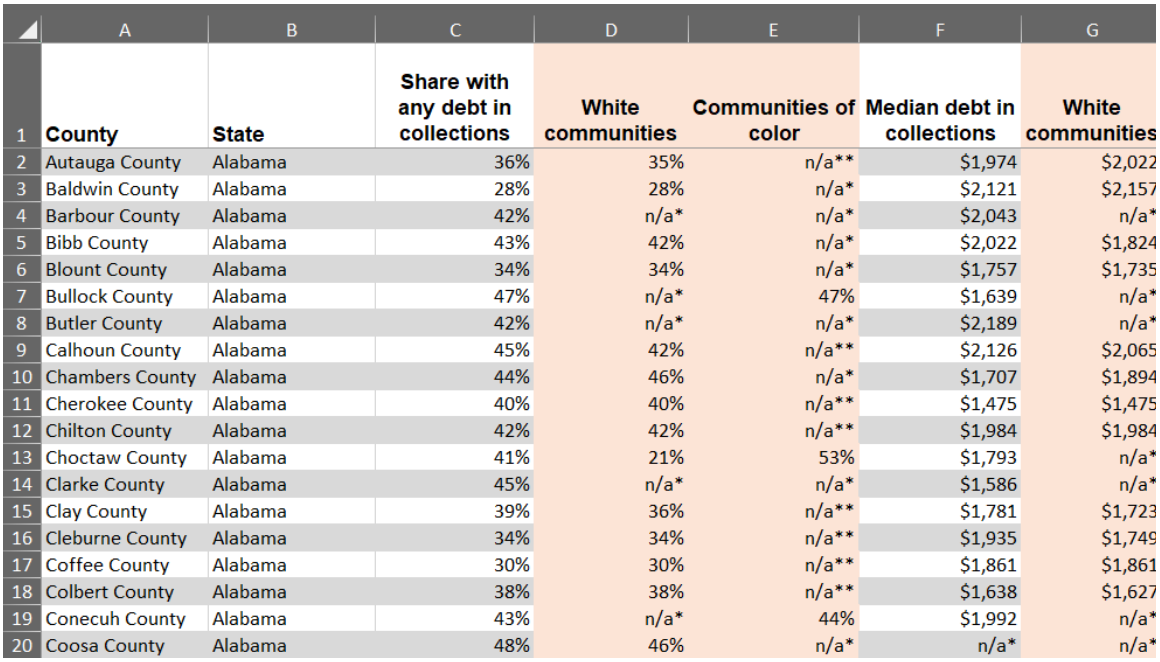
County	Share of people of color	Share of people in rural areas	Average household income
Autauga County	0.2458	42.00	7,211,012
Baldwin County	0.1692	42.28	7,306,091
Barbour County	0.5426	67.79	4,544,530

To prepare the data for linear regression, the team first removed variables not needed for linear regression, county and state. Next, all variables that related to a percentage were changed to proportions in order to maintain consistency.

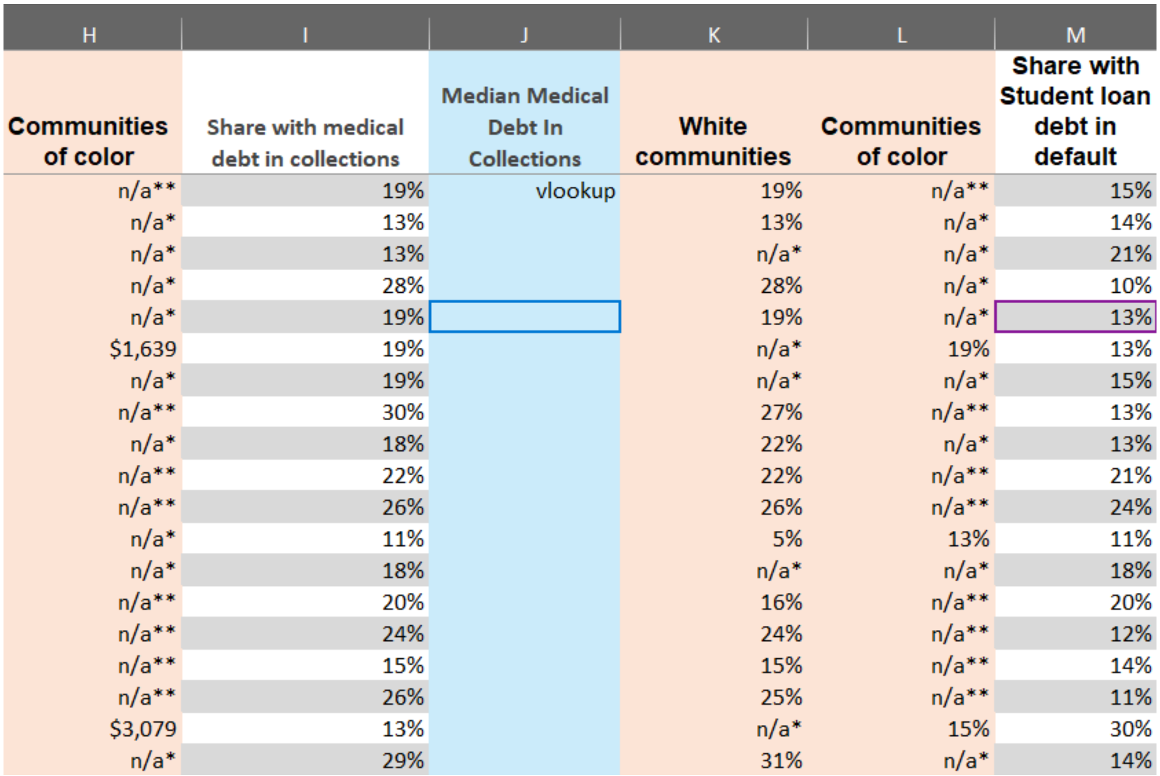
```
# Remove county and state variables.
final <- data %>%
  select(-County, -State) %>%
  drop_na()


# Convert percentages into proportions.
final$Share.with.any.debt.in.collections <-
  final$Share.with.any.debt.in.collections / 100
final$Share.with.medical.debt.in.collections <-
  final$Share.with.medical.debt.in.collections / 100
final$Auto.retail.loan.delinquency.rate <-
  final$Auto.retail.loan.delinquency.rate / 100
final$Share.of.people.in.rural.areas <-
  final$Share.of.people.in.rural.areas / 100
```

Excel

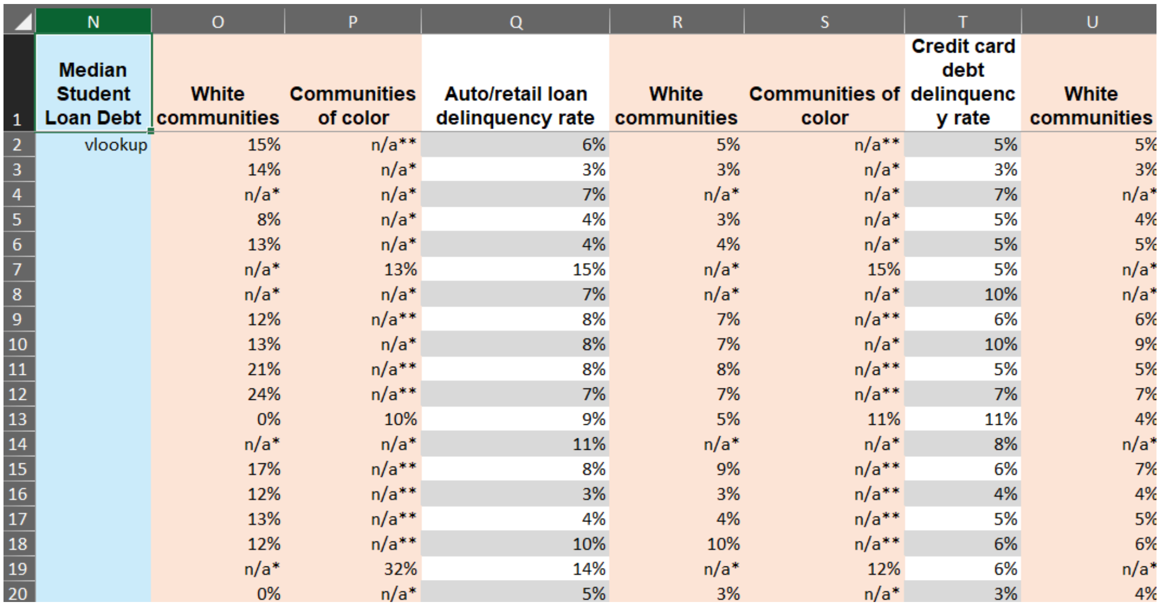
Using Excel, the team cleaned and prepared the data to condense all relevant columns onto one sheet deleted columns that were not to be used in future analysis.The Excel function =VLOOKUP was used to merge rows from other sheets into one main sheet. The columns “Median medical debt in collections,” “Median student loan debt in default,” and “Share of people in rural areas” were merged onto the main sheet. In addition, all formatting was done manually. This includes removing the double header in all sheets as well as including the entire population of the county and removing the “White communities” and “Communities of color” columns.



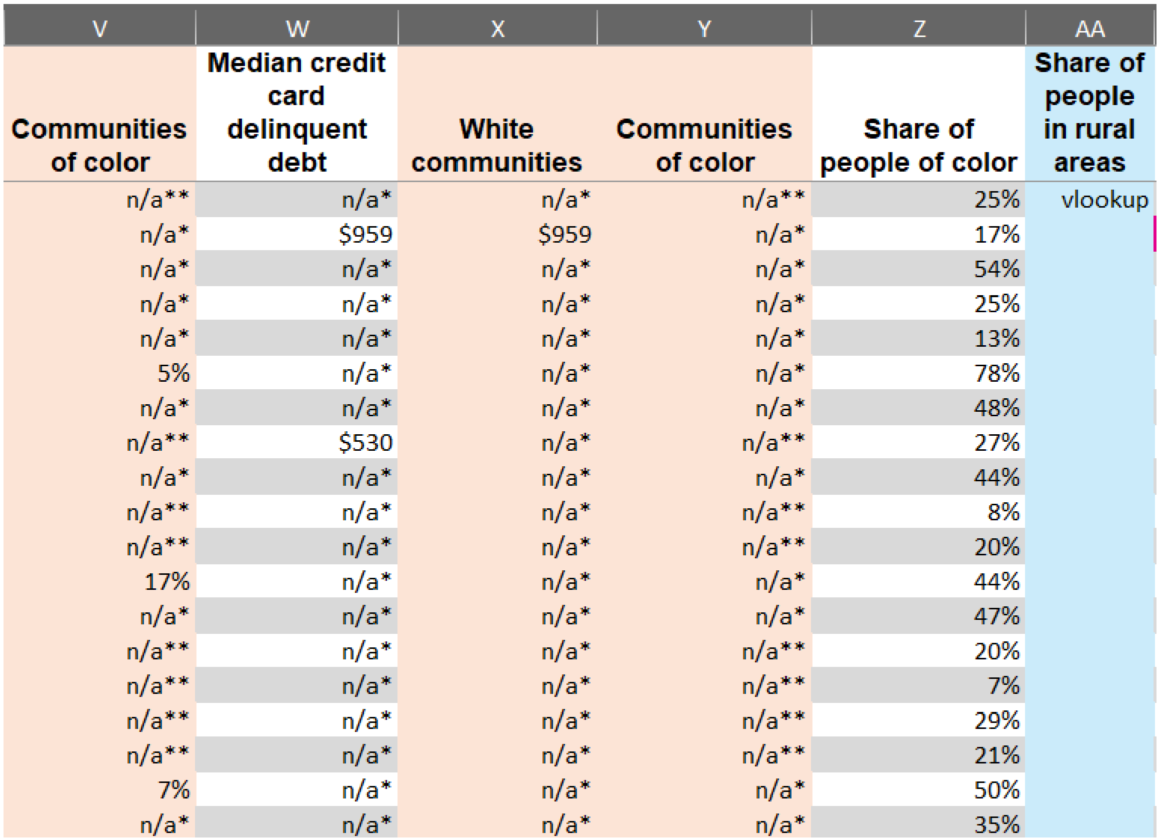
	A	B	C	D	E	F	G
1	**County**	**State**	**Share with any debt in collections**	**White communities**	**Communities of color**	**Median debt in collections**	**White communities**
2	Autauga County	Alabama	36%	35%	n/a**	$1,974	$2,022
3	Baldwin County	Alabama	28%	28%	n/a*	$2,121	$2,157
4	Barbour County	Alabama	42%	n/a*	n/a*	$2,043	n/a*
5	Bibb County	Alabama	43%	42%	n/a*	$2,022	$1,824
6	Blount County	Alabama	34%	34%	n/a*	$1,757	$1,735
7	Bullock County	Alabama	47%	n/a*	47%	$1,639	n/a*
8	Butler County	Alabama	42%	n/a*	n/a*	$2,189	n/a*
9	Calhoun County	Alabama	45%	42%	n/a**	$2,126	$2,065
10	Chambers County	Alabama	44%	46%	n/a*	$1,707	$1,894
11	Cherokee County	Alabama	40%	40%	n/a**	$1,475	$1,475
12	Chilton County	Alabama	42%	42%	n/a**	$1,984	$1,984
13	Choctaw County	Alabama	41%	21%	53%	$1,793	n/a*
14	Clarke County	Alabama	45%	n/a*	n/a*	$1,586	n/a*
15	Clay County	Alabama	39%	36%	n/a**	$1,781	$1,723
16	Cleburne County	Alabama	34%	34%	n/a**	$1,935	$1,749
17	Coffee County	Alabama	30%	30%	n/a**	$1,861	$1,861
18	Colbert County	Alabama	38%	38%	n/a**	$1,638	$1,627
19	Conecuh County	Alabama	43%	n/a*	44%	$1,992	n/a*
20	Coosa County	Alabama	48%	46%	n/a*	n/a*	n/a*



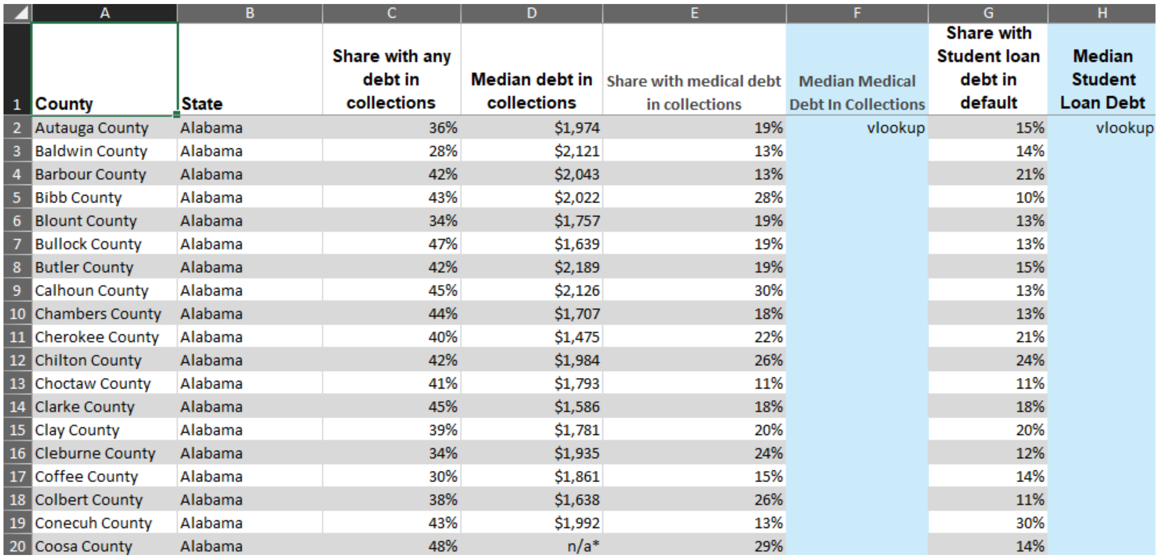
H	I	J	K	L	M
Communities of color	**Share with medical debt in collections**	**Median Medical Debt In Collections**	**White communities**	**Communities of color**	**Share with Student loan debt in default**
n/a**	19%	vlookup	19%	n/a**	15%
n/a*	13%		13%	n/a*	14%
n/a*	13%		n/a*	n/a*	21%
n/a*	28%		28%	n/a*	10%
n/a*	19%		19%	n/a*	13%
$1,639	19%		n/a*	19%	13%
n/a*	19%		n/a*	n/a*	15%
n/a**	30%		27%	n/a**	13%
n/a*	18%		22%	n/a*	13%
n/a**	22%		22%	n/a**	21%
n/a**	26%		26%	n/a**	24%
n/a*	11%		5%	13%	11%
n/a*	18%		n/a*	n/a*	18%
n/a**	20%		16%	n/a**	20%
n/a**	24%		24%	n/a**	12%
n/a**	15%		15%	n/a**	14%
n/a**	26%		25%	n/a**	11%
$3,079	13%		n/a*	15%	30%
n/a*	29%		31%	n/a*	14%



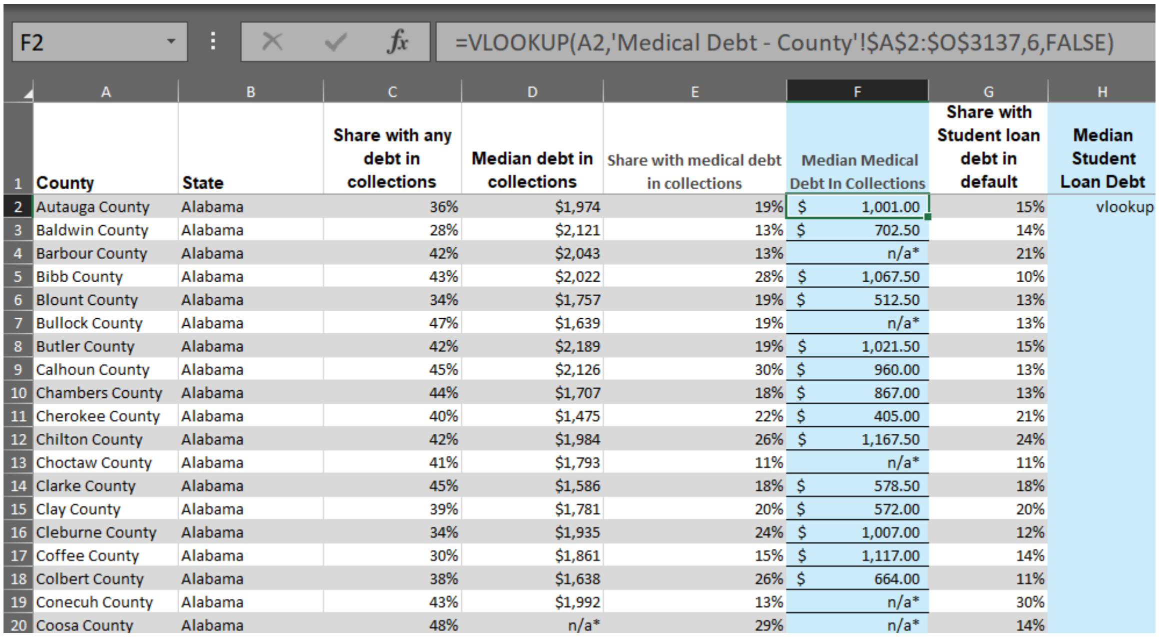
	N	O	P	Q	R	S	T	U
1	Median Student Loan Debt	White communities	Communities of color	Auto/retail loan delinquency rate	White communities	Communities of color	Credit card debt delinquency rate	White communities
2	vlookup	15%	n/a**	6%	5%	n/a**	5%	5%
3		14%	n/a*	3%	3%	n/a*	3%	3%
4		n/a*	n/a*	7%	n/a*	n/a*	7%	n/a*
5		8%	n/a*	4%	3%	n/a*	5%	4%
6		13%	n/a*	4%	4%	n/a*	5%	5%
7		n/a*	13%	15%	n/a*	15%	5%	n/a*
8		n/a*	n/a*	7%	n/a*	n/a*	10%	n/a*
9		12%	n/a**	8%	7%	n/a**	6%	6%
10		13%	n/a*	8%	7%	n/a*	10%	9%
11		21%	n/a**	8%	8%	n/a**	5%	5%
12		24%	n/a**	7%	7%	n/a**	7%	7%
13		0%	10%	9%	5%	11%	11%	4%
14		n/a*	n/a*	11%	n/a*	n/a*	8%	n/a*
15		17%	n/a**	8%	9%	n/a**	6%	7%
16		12%	n/a**	3%	3%	n/a**	4%	4%
17		13%	n/a**	4%	4%	n/a**	5%	5%
18		12%	n/a**	10%	10%	n/a**	6%	6%
19		n/a*	32%	14%	n/a*	12%	6%	n/a*
20		0%	n/a*	5%	3%	n/a*	3%	4%



V	W	X	Y	Z	AA
Communities of color	Median credit card delinquent debt	White communities	Communities of color	Share of people of color	Share of people in rural areas
n/a**	n/a*	n/a*	n/a**	25%	vlookup
n/a*	$959	$959	n/a*	17%	
n/a*	n/a*	n/a*	n/a*	54%	
n/a*	n/a*	n/a*	n/a*	25%	
n/a*	n/a*	n/a*	n/a*	13%	
5%	n/a*	n/a*	n/a*	78%	
n/a*	n/a*	n/a*	n/a*	48%	
n/a**	$530	n/a*	n/a**	27%	
n/a*	n/a*	n/a*	n/a*	44%	
n/a**	n/a*	n/a*	n/a**	8%	
n/a**	n/a*	n/a*	n/a**	20%	
17%	n/a*	n/a*	n/a*	44%	
n/a*	n/a*	n/a*	n/a*	47%	
n/a**	n/a*	n/a*	n/a**	20%	
n/a**	n/a*	n/a*	n/a**	7%	
n/a**	n/a*	n/a*	n/a**	29%	
n/a**	n/a*	n/a*	n/a**	21%	
7%	n/a*	n/a*	n/a*	50%	
n/a*	n/a*	n/a*	n/a*	35%	



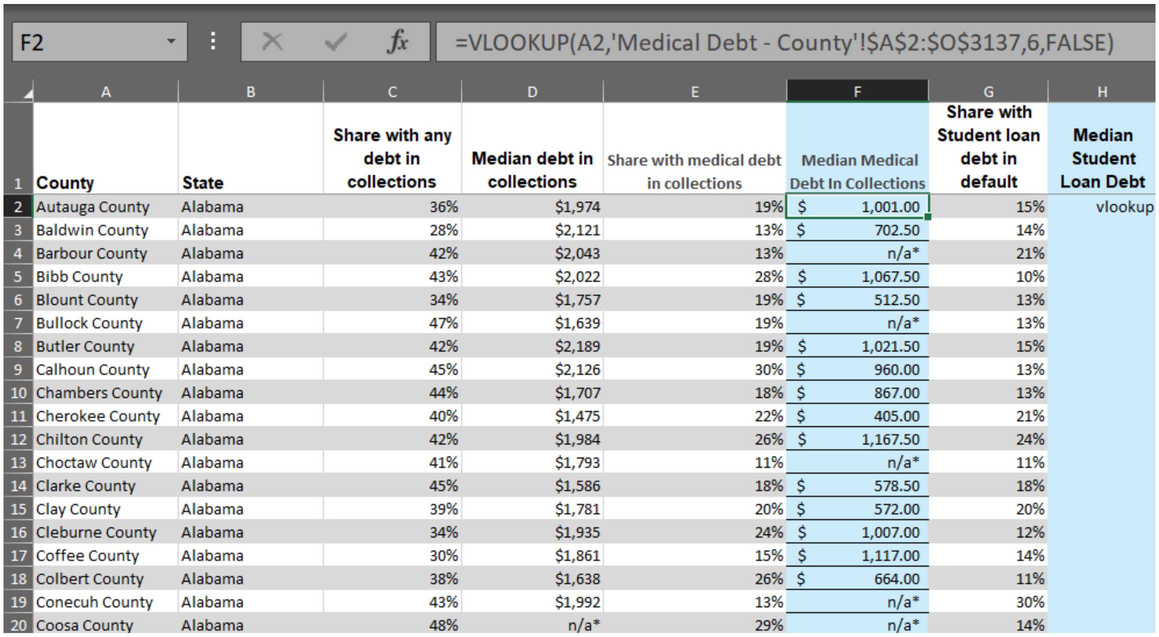
	A	B	C	D	E	F	G	H
1	County	State	Share with any debt in collections	Median debt in collections	Share with medical debt in collections	Median Medical Debt In Collections	Share with Student loan debt in default	Median Student Loan Debt
2	Autauga County	Alabama	36%	$1,974	19%	vlookup	15%	vlookup
3	Baldwin County	Alabama	28%	$2,121	13%		14%	
4	Barbour County	Alabama	42%	$2,043	13%		21%	
5	Bibb County	Alabama	43%	$2,022	28%		10%	
6	Blount County	Alabama	34%	$1,757	19%		13%	
7	Bullock County	Alabama	47%	$1,639	19%		13%	
8	Butler County	Alabama	42%	$2,189	19%		15%	
9	Calhoun County	Alabama	45%	$2,126	30%		13%	
10	Chambers County	Alabama	44%	$1,707	18%		13%	
11	Cherokee County	Alabama	40%	$1,475	22%		21%	
12	Chilton County	Alabama	42%	$1,984	26%		24%	
13	Choctaw County	Alabama	41%	$1,793	11%		11%	
14	Clarke County	Alabama	45%	$1,586	18%		18%	
15	Clay County	Alabama	39%	$1,781	20%		20%	
16	Cleburne County	Alabama	34%	$1,935	24%		12%	
17	Coffee County	Alabama	30%	$1,861	15%		14%	
18	Colbert County	Alabama	38%	$1,638	26%		11%	
19	Conecuh County	Alabama	43%	$1,992	13%		30%	
20	Coosa County	Alabama	48%	n/a*	29%		14%	



F2

=VLOOKUP(A2,'Medical Debt - County'!$A$2:$O$3137,6,FALSE)

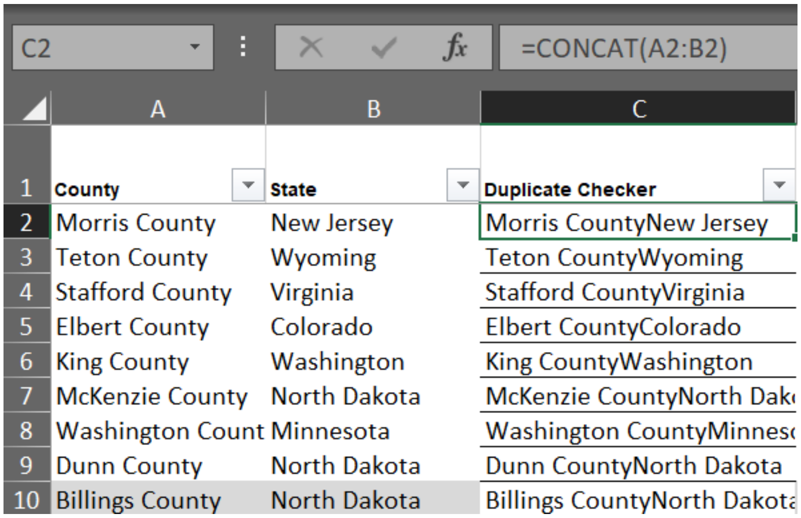
	A	B	C	D	E	F	G	H
1	County	State	Share with any debt in collections	Median debt in collections	Share with medical debt in collections	Median Medical Debt In Collections	Share with Student loan debt in default	Median Student Loan Debt
2	Autauga County	Alabama	36%	$1,974	19%	$ 1,001.00	15%	vlookup
3	Baldwin County	Alabama	28%	$2,121	13%	$ 702.50	14%	
4	Barbour County	Alabama	42%	$2,043	13%	n/a*	21%	
5	Bibb County	Alabama	43%	$2,022	28%	$ 1,067.50	10%	
6	Blount County	Alabama	34%	$1,757	19%	$ 512.50	13%	
7	Bullock County	Alabama	47%	$1,639	19%	n/a*	13%	
8	Butler County	Alabama	42%	$2,189	19%	$ 1,021.50	15%	
9	Calhoun County	Alabama	45%	$2,126	30%	$ 960.00	13%	
10	Chambers County	Alabama	44%	$1,707	18%	$ 867.00	13%	
11	Cherokee County	Alabama	40%	$1,475	22%	$ 405.00	21%	
12	Chilton County	Alabama	42%	$1,984	26%	$ 1,167.50	24%	
13	Choctaw County	Alabama	41%	$1,793	11%	n/a*	11%	
14	Clarke County	Alabama	45%	$1,586	18%	$ 578.50	18%	
15	Clay County	Alabama	39%	$1,781	20%	$ 572.00	20%	
16	Cleburne County	Alabama	34%	$1,935	24%	$ 1,007.00	12%	
17	Coffee County	Alabama	30%	$1,861	15%	$ 1,117.00	14%	
18	Colbert County	Alabama	38%	$1,638	26%	$ 664.00	11%	
19	Conecuh County	Alabama	43%	$1,992	13%	n/a*	30%	
20	Coosa County	Alabama	48%	n/a*	29%	n/a*	14%	



F2 =VLOOKUP(A2,'Medical Debt - County'!$A$2:$O$3137,6,FALSE)

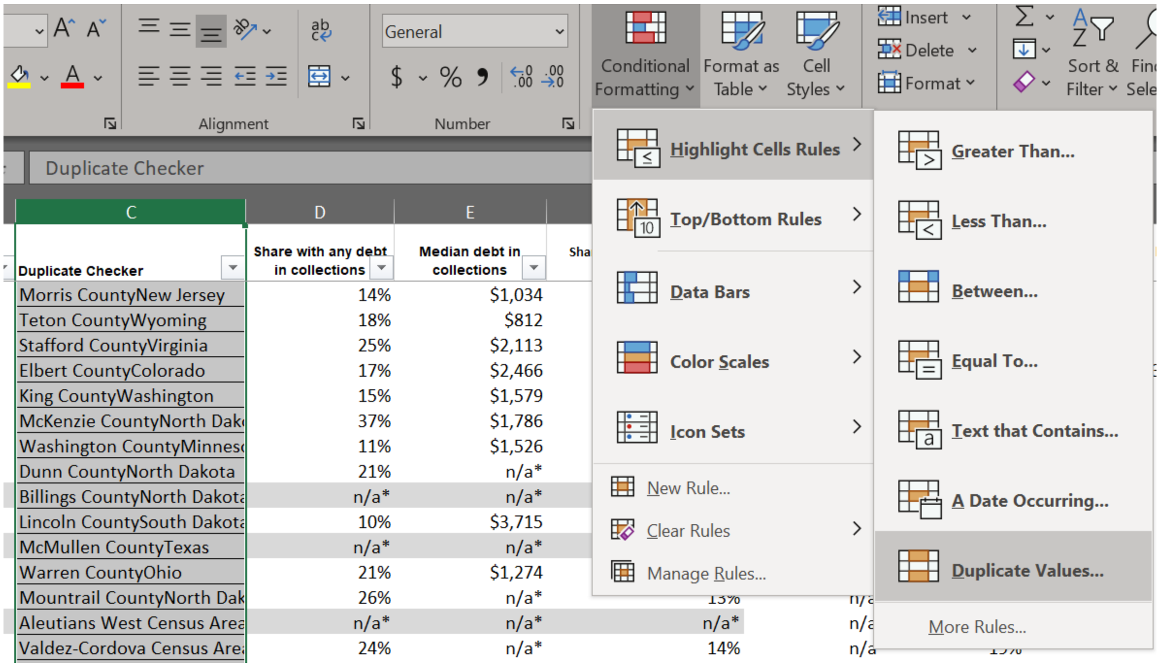
	A	B	C	D	E	F	G	H
1	**County**	**State**	**Share with any debt in collections**	**Median debt in collections**	Share with medical debt in collections	Median Medical Debt In Collections	**Share with Student loan debt in default**	**Median Student Loan Debt**
2	Autauga County	Alabama	36%	$1,974	19%	$ 1,001.00	15%	vlookup
3	Baldwin County	Alabama	28%	$2,121	13%	$ 702.50	14%	
4	Barbour County	Alabama	42%	$2,043	13%	n/a*	21%	
5	Bibb County	Alabama	43%	$2,022	28%	$ 1,067.50	10%	
6	Blount County	Alabama	34%	$1,757	19%	$ 512.50	13%	
7	Bullock County	Alabama	47%	$1,639	19%	n/a*	13%	
8	Butler County	Alabama	42%	$2,189	19%	$ 1,021.50	15%	
9	Calhoun County	Alabama	45%	$2,126	30%	$ 960.00	13%	
10	Chambers County	Alabama	44%	$1,707	18%	$ 867.00	13%	
11	Cherokee County	Alabama	40%	$1,475	22%	$ 405.00	21%	
12	Chilton County	Alabama	42%	$1,984	26%	$ 1,167.50	24%	
13	Choctaw County	Alabama	41%	$1,793	11%	n/a*	11%	
14	Clarke County	Alabama	45%	$1,586	18%	$ 578.50	18%	
15	Clay County	Alabama	39%	$1,781	20%	$ 572.00	20%	
16	Cleburne County	Alabama	34%	$1,935	24%	$ 1,007.00	12%	
17	Coffee County	Alabama	30%	$1,861	15%	$ 1,117.00	14%	
18	Colbert County	Alabama	38%	$1,638	26%	$ 664.00	11%	
19	Conecuh County	Alabama	43%	$1,992	13%	n/a*	30%	
20	Coosa County	Alabama	48%	n/a*	29%	n/a*	14%	

To check for duplicates after merging rows from other sheets, the team concatenated "County" and "State" using the =CONCAT function, applied conditional formatting to highlight duplicate values, and then sorted the columns by color. No duplicate rows were found.



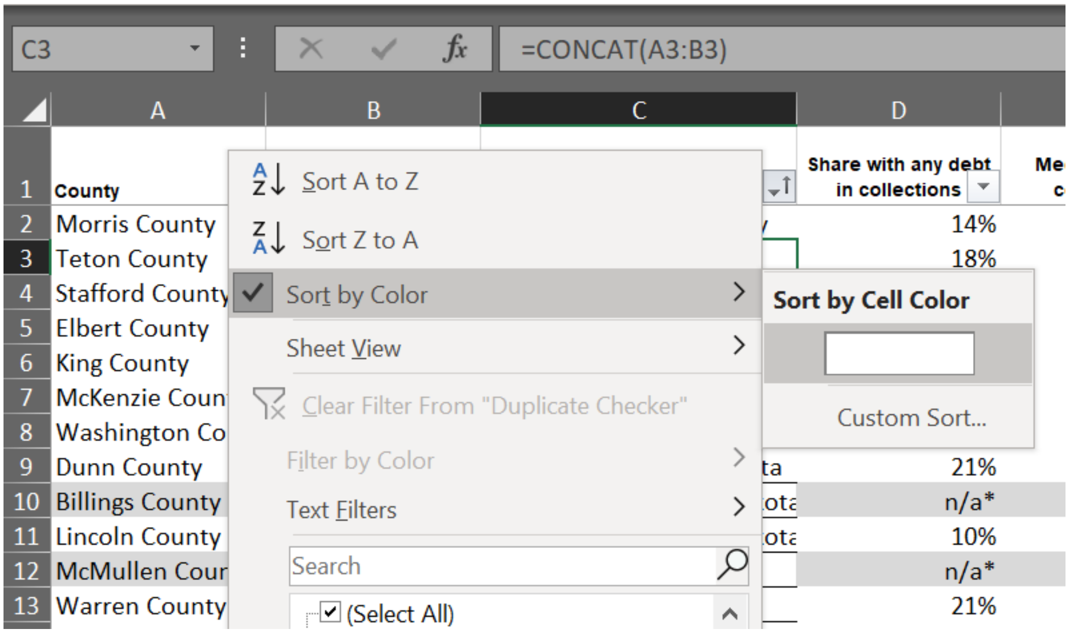
C2 =CONCAT(A2:B2)

	A	B	C
1	**County**	**State**	**Duplicate Checker**
2	Morris County	New Jersey	Morris CountyNew Jersey
3	Teton County	Wyoming	Teton CountyWyoming
4	Stafford County	Virginia	Stafford CountyVirginia
5	Elbert County	Colorado	Elbert CountyColorado
6	King County	Washington	King CountyWashington
7	McKenzie County	North Dakota	McKenzie CountyNorth Dak
8	Washington Count	Minnesota	Washington CountyMinnes
9	Dunn County	North Dakota	Dunn CountyNorth Dakota
10	Billings County	North Dakota	Billings CountyNorth Dakot



Duplicate Checker

Duplicate Checker	Share with any debt in collections	Median debt in collections
Morris CountyNew Jersey	14%	$1,034
Teton CountyWyoming	18%	$812
Stafford CountyVirginia	25%	$2,113
Elbert CountyColorado	17%	$2,466
King CountyWashington	15%	$1,579
McKenzie CountyNorth Dak	37%	$1,786
Washington CountyMinnes	11%	$1,526
Dunn CountyNorth Dakota	21%	n/a*
Billings CountyNorth Dakota	n/a*	n/a*
Lincoln CountySouth Dakota	10%	$3,715
McMullen CountyTexas	n/a*	n/a*
Warren CountyOhio	21%	$1,274
Mountrail CountyNorth Dak	26%	n/a*
Aleutians West Census Area	n/a*	n/a*
Valdez-Cordova Census Area	24%	n/a*



C3 =CONCAT(A3:B3)

	A	D
1	County	Share with any debt in collections
2	Morris County	14%
3	Teton County	18%
4	Stafford County	
5	Elbert County	
6	King County	
7	McKenzie Coun	
8	Washington Co	
9	Dunn County	21%
10	Billings County	n/a*
11	Lincoln County	10%
12	McMullen Cour	n/a*
13	Warren County	21%

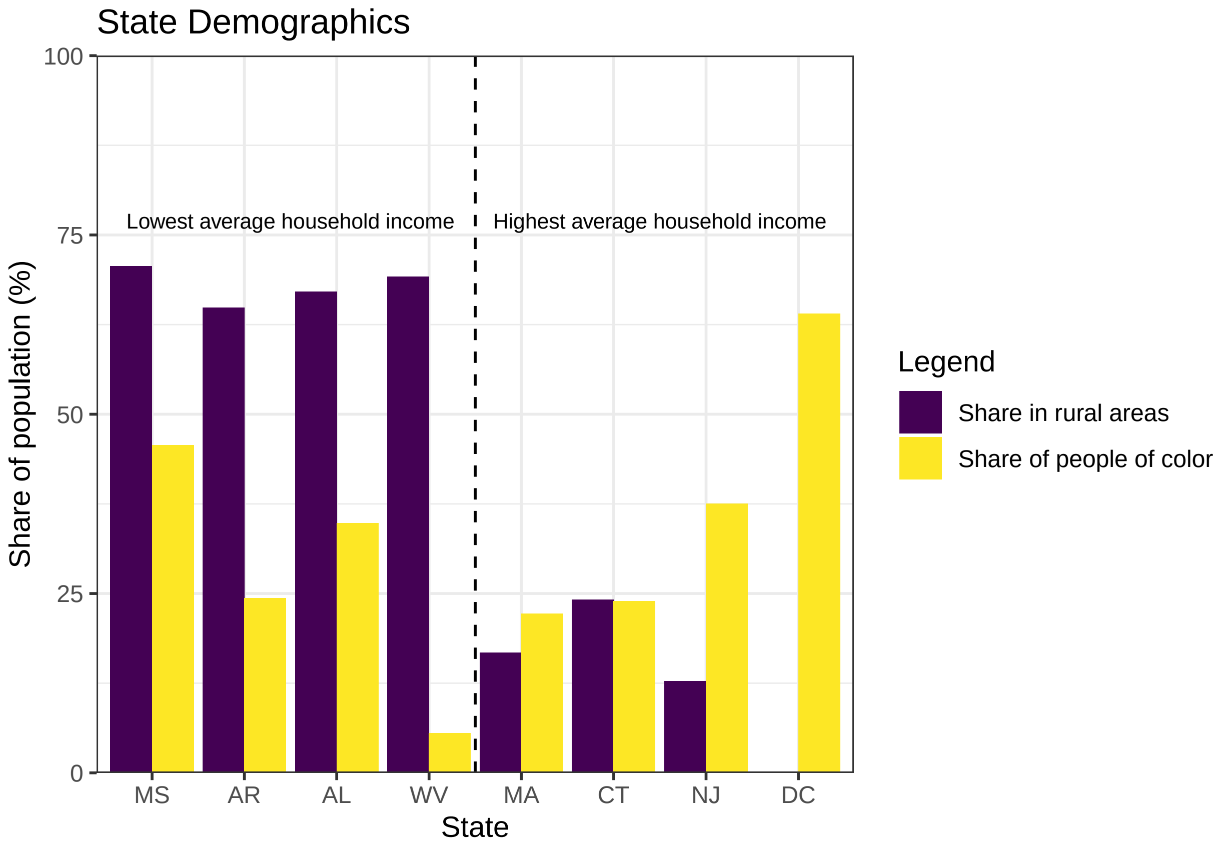
Appendix C: Analytics details

Descriptive analytics

All graphic visualizations were created in R.

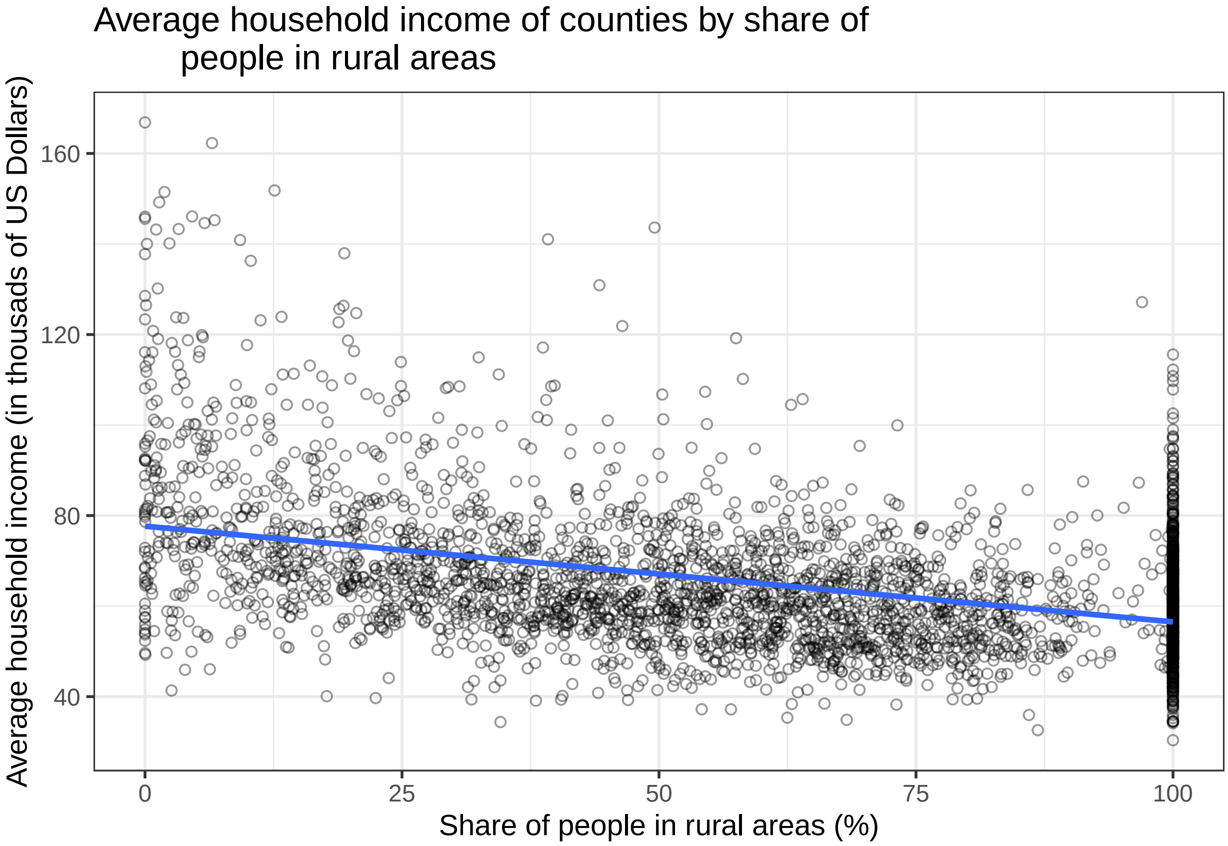
The team first calculated the mean share of people in rural areas for each state and sorted these states to find the eight desired states, creating a new data frame with only the relevant states. The team wanted to plot the different types of demographic proportions for each state side-by-side, but needed to change the structure of the data frame to be longer as opposed to wider in order to group the different types of debt. The scales of the plot were adjusted to minimize empty space on the y-axis and to change the order of the states and labels on the x-axis. The colors were also changed to the viridis color set, and labels for the different debts were adjusted in the same function. Vertical lines and labels were added in order to help readers distinguish between high and low average household incomes in each selected state.

```
# Sort states by average household income.
ahi <-group_by(county.clean, State) %>%
  summarize(mean = mean(as.numeric(‘Average household income‘, na.rm=T))) %>%
  arrange(mean)
ahi <- ahi[c(1:4,48:51),1]

# Create a new data frame with only eight states.
loans <- inner_join(county.clean, ahi, by = "State")

# Make the data frame longer using melt() function found in the reshape2 package.
loans <- melt(loans[,c('County','State', 'Share of people in rural areas',
                        'Share of people of color')],
               id.vars = c(1,2), measure.vars = -c(1,2) , na.rm=T)

ggplot(loans,aes(x = fct_rev(State), y = value)) +
  geom_bar(aes(fill = variable),stat = "summary", position = "dodge") +
  scale_y_continuous(limits = c(0,100), expand = c(0,0)) +
  scale_x_discrete(name ="State",
                   limits=c("Mississippi","Arkansas", "Alabama", "West Virginia",
                            "Massachusetts", "Connecticut", "New Jersey",
                            "District of Columbia"),
                   labels = c("MS", "AR", "AL", "WV", "MA", "CT", "NJ", "DC")) +
  theme_bw() +
  labs(title = "State Demographics",
       x = "State",
       y = "Share of population (%)") +
  scale_fill_viridis(discrete = T, name = "Legend",
                     labels = c("Share in rural areas", "Share of people of color")) +
  geom_vline(xintercept = 4.5, linetype = "dashed") +
  annotate(geom="text", x=2.5, y=77, label="Lowest average household income",
              color="black", size = 2.75) +
  annotate(geom="text", x=6.5, y=77, label="Highest average household income",
              color="black", size = 2.75)
```



Next, the team used ggplot in the ggplot2 package in order to create a scatterplot, plotting the average household income by share of people in rural areas for each county. A linear model was also added in order to better visualize any potential correlation between the two variables. The point shape and transparency were adjusted in order to better see the linear model.

```
county.clean %>%
  ggplot( aes(x = ‘Share of people in rural areas‘,
              y = ‘Average household income‘/1000) ) +
  geom_point(shape = 1, alpha = 0.4) +
  geom_smooth(method = lm, se = FALSE) +
  theme_bw() +
    labs(title = "Average household income of counties by share of
         people in rural areas",
       x = "Share of people in rural areas (%)",
       y = "Average household income (in thousads of US Dollars)")
```



In the next graphic, more data manipulation needed to be done in order to create a grouped bar plot. First, the team calculated the mean share of people in rural areas for each state and sorted these states to find the eight desired states, creating a new data frame with only the relevant states. The team wanted to plot the different types of debt for each state side-by-side, but needed to change the structure of the data frame to be longer as opposed to wider in order to group the different types of debt. The scales of the plot were adjusted to minimize empty space on the y-axis and to change the order of the states and labels on the x-axis. The colors were also changed to the viridis color set, and labels for the different debts were adjusted in the same function. Vertical lines and labels were added in order to help readers distinguish between high and low shares of people living in rural areas in each selected state.

```
# Sort states by share of people in rural areas.
rural <-group_by(county.clean, State) %>%
  summarize(mean = mean(as.numeric(‘Share of people in rural areas‘, na.rm=T))) %>%
  arrange(mean)
rural <- rural[c(1:4,48:51),1]

# Create a new data frame with only eight states.
loans <- inner_join(county.clean, rural, by = "State")

# Make the data frame longer using melt() function found in the reshape2 package.
loans1 <- melt(loans[,c(’County’,’State’,’Share with any debt in collections’,
                        ’Share with medical debt in collections’,
                        ’Share with student loan debt in default’,
                        ’Auto/retail loan delinquency rate’,
                        ’Credit card debt delinquency rate’)], id.vars = c(1,2),
               measure.vars = -c(1,2) , na.rm=T)

ggplot(loans1,aes(x = fct_rev(State), y = value)) +
  geom_bar(aes(fill = variable),stat = "identity", position = "dodge") +
  scale_y_continuous(limits = c(0,100), expand = c(0,0)) +
  scale_x_discrete(name ="State",
                   limits=c("District of Columbia","Rhode Island", "New Jersey",
                            "Massachusetts", "Montana", "Vermont", "South Dakota",
                            "North Dakota"),
                   labels = c("DC", "RI", "NJ", "MA", "MT", "VT", "SD", "ND")) +
  theme_bw() +
  labs(title = "Share of Debt by State (Rural Population)",
       x = "State",
       y = "Share with debt (%)") +
  scale_fill_viridis(discrete = T, name = "Types of Debt",
                     labels = c("Any Debt", "Medical Debt", "Student Loan",
                                "Auto/Retail Loan", "Credit Card Debt")) +
  geom_vline(xintercept = 4.5, linetype = "dashed") +
  annotate(geom="text", x=2.5, y=77, label="Low rural population",
              color="black", size = 3) +
  annotate(geom="text", x=6.5, y=77, label="High rural population",
              color="black", size = 3)
```

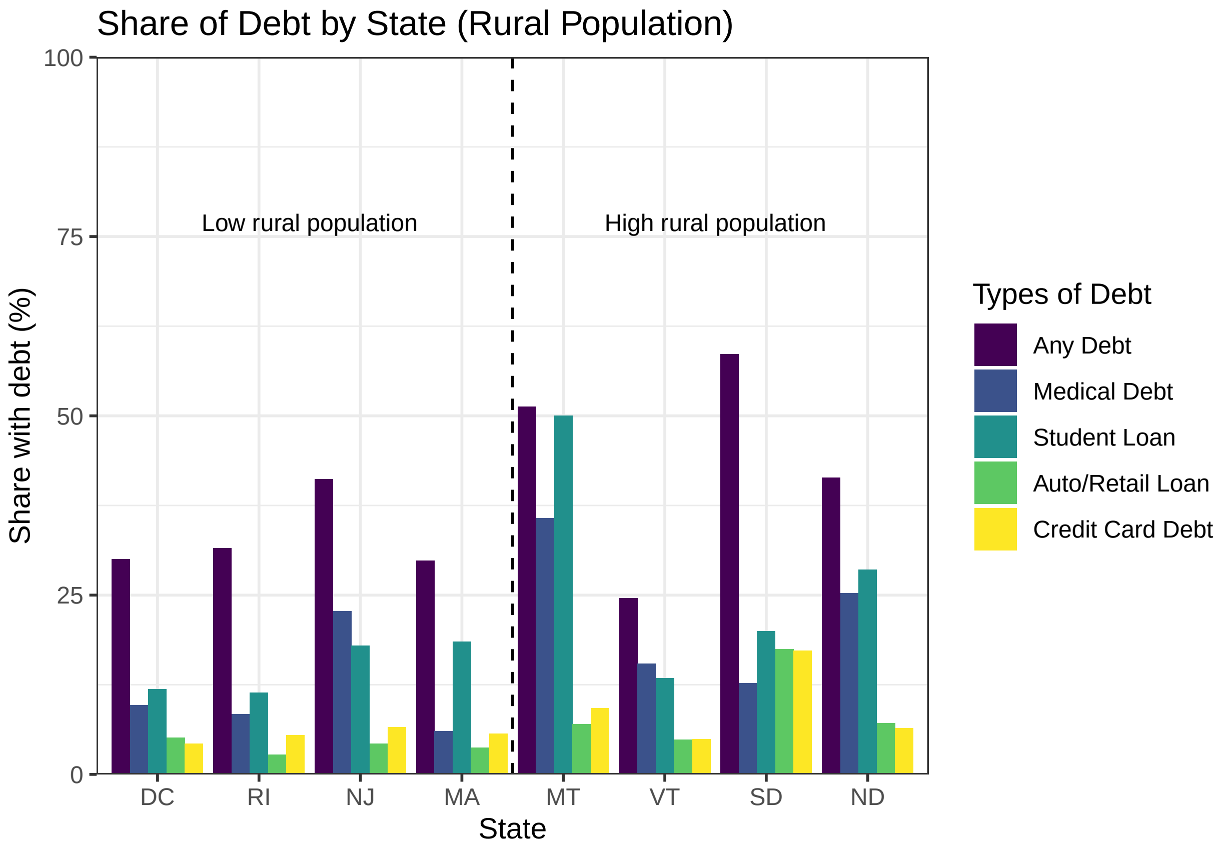
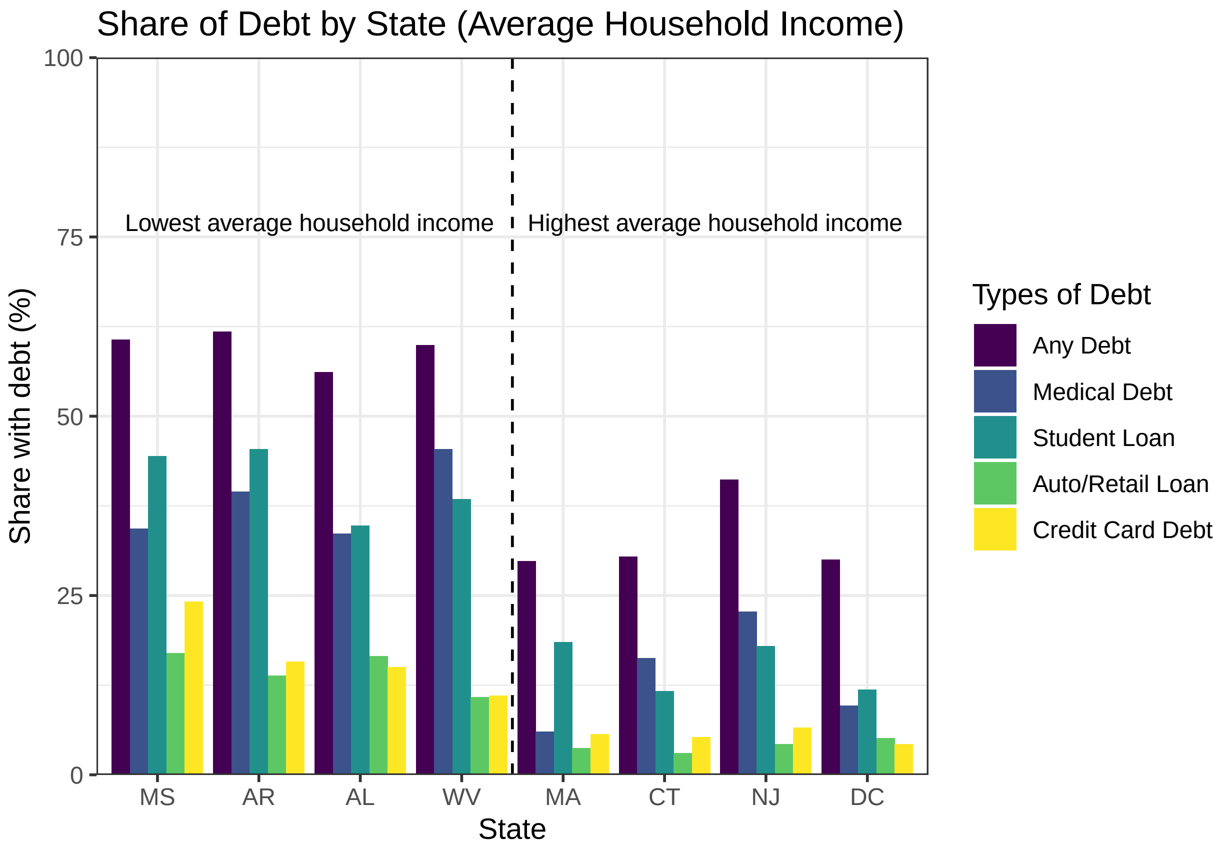


Figure 4 was created similarly to Figure 3. First, the team maintained the same eight states as Figure 1 that selected states based off of highest and lowest average household income. The team changed the structure of the data frame to be longer as opposed to wider in order to group the different types of debt. The scales of the plot were adjusted to minimize empty space on the y-axis and to change the order of the states and labels on the x-axis. The colors were also changed to the viridis color set, and labels for the different debts were adjusted in the same function. Vertical lines and labels were added in order to help readers distinguish between high and low shares of people living in rural areas in each selected state.

```
# Sort states by average househodl income.
ahi <-group_by(county.clean, State) %>%
  summarize(mean = mean(as.numeric(‘Average household income‘, na.rm=T))) %>%
  arrange(mean)
ahi <- ahi[c(1:4,48:51),1]

# Create a new data frame with only eight states.
loans <- inner_join(county.clean, ahi, by = "State")

# Make the data frame longer using melt() function found in the reshape2 package.
loans2 <- melt(loans[,c('County','State','Share with any debt in collections',
                        'Share with medical debt in collections',
                        'Share with student loan debt in default',
                        'Auto/retail loan delinquency rate',
                        'Credit card debt delinquency rate')],
               id.vars = c(1,2), measure.vars = -c(1,2) , na.rm=T)


ggplot(loans2,aes(x = fct_rev(State), y = value)) +
  geom_bar(aes(fill = variable), stat = "identity", position = "dodge") +
  scale_y_continuous(limits = c(0,100), expand = c(0,0)) +
  scale_x_discrete(name ="State",
                   limits=c("Mississippi","Arkansas", "Alabama", "West Virginia",
                            "Massachusetts", "Connecticut", "New Jersey",
                            "District of Columbia"),
                   labels = c("MS", "AR", "AL", "WV", "MA", "CT", "NJ", "DC")) +
  theme_bw() +
  labs(title = "Share of Debt by State (Average Household Income)",
       x = "State",
       y = "Share with debt (%)") +
  scale_fill_viridis(discrete = T, name = "Types of Debt",
                     labels = c("Any Debt", "Medical Debt", "Student Loan",
                                "Auto/Retail Loan", "Credit Card Debt")) +
  geom_vline(xintercept = 4.5, linetype = "dashed") +
  annotate(geom="text", x=2.5, y=77, label="Lowest average household income",
              color="black", size = 3) +
  annotate(geom="text", x=6.5, y=77, label="Highest average household income",
              color="black", size = 3)
```



Predictive analytics

The data was first divided into training and validation sets using a 70/30 split.

```
# Split data into testing and validation sets.
final.div <- final %>%
  initial_split(prop = 0.7)

final.train <- training(final.div)
final.validate <- testing(final.div)
```

Next, the team began to build various linear regression models to predict median debt in collections. The initial model included all the variables in the data set. Explanatory variables that were not significant were removed from each model. Because there are no categorical variables in the model, partial F tests were not necessary. Values with high VIF values were removed from the model to prevent multicollinearity.

```
Reg1 <- lm(Median.debt.in.collections ~ ., data = final.train)
summary(Reg1)
```

```
##
## Call:
## lm(formula = Median.debt.in.collections ~ ., data = final.train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75190 -17945  -3411  16136 129052
##
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              8.149e+04  1.133e+04   7.193 1.20e-12
## Share.with.any.debt.in.collections       1.163e+03  3.639e+04   0.032 0.974500
## Share.with.medical.debt.in.collections  -1.039e+05  3.009e+04  -3.454 0.000575
## Median.medical.debt.in.collections       1.009e+00  3.260e-02  30.949  < 2e-16
## Share.with.student.loan.debt.in.default  1.247e+05  2.362e+04   5.278 1.58e-07
## Median.student.loan.debt                 6.670e-01  2.858e-01   2.333 0.019819
## Auto.retail.loan.delinquency.rate        1.951e+05  6.710e+04   2.907 0.003724
## Credit.card.debt.delinquency.rate        1.247e+05  7.932e+04   1.573 0.116081
## Share.of.people.of.color                -1.567e+04  7.287e+03  -2.150 0.031794
## Share.of.people.in.rural.areas          -1.456e+04  4.468e+03  -3.260 0.001152
## Average.household.income                 1.682e-04  7.873e-04   0.214 0.830860
##
## (Intercept)                             ***
## Share.with.any.debt.in.collections
## Share.with.medical.debt.in.collections  ***
## Median.medical.debt.in.collections      ***
## Share.with.student.loan.debt.in.default ***
## Median.student.loan.debt                *
## Auto.retail.loan.delinquency.rate       **
## Credit.card.debt.delinquency.rate
## Share.of.people.of.color                *
## Share.of.people.in.rural.areas          **
## Average.household.income
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 27430 on 1048 degrees of freedom
## Multiple R-squared:  0.5246, Adjusted R-squared:  0.5201
## F-statistic: 115.7 on 10 and 1048 DF,  p-value: < 2.2e-16
```

```
Reg2 <- lm(Median.debt.in.collections ~ .-
             Share.with.medical.debt.in.collections-
             Median.student.loan.debt-
             Average.household.income-
             Share.of.people.of.color, data = final.train)
kable(vif(Reg2))
```

	x
Share.with.any.debt.in.collections	3.595352
Median.medical.debt.in.collections	1.094135
Share.with.student.loan.debt.in.default	1.666181
Auto.retail.loan.delinquency.rate	2.463948
Credit.card.debt.delinquency.rate	2.348304
Share.of.people.in.rural.areas	1.056137

```
Reg3 <- lm(Median.debt.in.collections ~ .-
             Share.with.medical.debt.in.collections-
             Median.student.loan.debt-
             Average.household.income-
             Share.of.people.of.color-
             Share.with.any.debt.in.collections, data = final.train)
summary(Reg3)
```

```
##
```

```
## Call:
## lm(formula = Median.debt.in.collections ~ . - Share.with.medical.debt.in.collections
##     Median.student.loan.debt - Average.household.income - Share.of.people.of.color -
##     Share.with.any.debt.in.collections, data = final.train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -68212 -18512  -3799  15300 130572
##
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              9.578e+04  3.530e+03  27.132  < 2e-16
## Median.medical.debt.in.collections       9.159e-01  3.032e-02  30.202  < 2e-16
## Share.with.student.loan.debt.in.default  7.316e+04  2.120e+04   3.451 0.000581
## Auto.retail.loan.delinquency.rate        4.367e+04  5.647e+04   0.773 0.439459
## Credit.card.debt.delinquency.rate        7.047e+03  6.875e+04   0.102 0.918382
## Share.of.people.in.rural.areas          -1.701e+04  3.569e+03  -4.766 2.14e-06
##
## (Intercept)                             ***
## Median.medical.debt.in.collections      ***
## Share.with.student.loan.debt.in.default ***
## Auto.retail.loan.delinquency.rate
## Credit.card.debt.delinquency.rate
## Share.of.people.in.rural.areas          ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28210 on 1053 degrees of freedom
## Multiple R-squared:  0.4947, Adjusted R-squared:  0.4923
## F-statistic: 206.2 on 5 and 1053 DF,  p-value: < 2.2e-16
```

```
Reg4 <- lm(Median.debt.in.collections ~ .-
             Share.with.medical.debt.in.collections-
             Median.student.loan.debt-
             Average.household.income-
             Share.of.people.of.color-
             Share.with.any.debt.in.collections-
             Credit.card.debt.delinquency.rate, data = final.train)
summary(Reg4)
```

```
##
## Call:
## lm(formula = Median.debt.in.collections ~ . - Share.with.medical.debt.in.collections
##     Median.student.loan.debt - Average.household.income - Share.of.people.of.color -
##     Share.with.any.debt.in.collections - Credit.card.debt.delinquency.rate,
##     data = final.train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -68195 -18572  -3789  15271 130570
##
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              9.591e+04  3.286e+03  29.184  < 2e-16
## Median.medical.debt.in.collections       9.157e-01  3.025e-02  30.270  < 2e-16
## Share.with.student.loan.debt.in.default  7.372e+04  2.049e+04   3.598 0.000335
## Auto.retail.loan.delinquency.rate        4.704e+04  4.593e+04   1.024 0.305985
## Share.of.people.in.rural.areas          -1.704e+04  3.560e+03  -4.786 1.94e-06
##
## (Intercept)                             ***
## Median.medical.debt.in.collections      ***
## Share.with.student.loan.debt.in.default ***
## Auto.retail.loan.delinquency.rate
## Share.of.people.in.rural.areas          ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28200 on 1054 degrees of freedom
## Multiple R-squared:  0.4947, Adjusted R-squared:  0.4928
## F-statistic:   258 on 4 and 1054 DF,  p-value: < 2.2e-16
```

```
kable(vif(Reg4))
```

	x
Median.medical.debt.in.collections	1.045211
Share.with.student.loan.debt.in.default	1.327673
Auto.retail.loan.delinquency.rate	1.296828
Share.of.people.in.rural.areas	1.041740

To compare the models, the team created a table to summarize the RMSE and adjusted R-squared values for each model. Model 4 was chosen despite having the highest RMSE and the lowest adjusted R-squared.

```
model1 <- cbind(as.numeric(summary(Reg1)[6]), as.numeric(summary(Reg1)[8]))
model2 <- cbind(as.numeric(summary(Reg2)[6]), as.numeric(summary(Reg2)[8]))
model3<- cbind(as.numeric(summary(Reg3)[6]), as.numeric(summary(Reg3)[8]))
model4<- cbind(as.numeric(summary(Reg4)[6]), as.numeric(summary(Reg4)[8]))

sum.stats <- (rbind(model1, model2, model3, model4))
colnames(sum.stats) <- c("RMSE", "AdjustedRSquared")
rownames(sum.stats) <- c("Model 1", "Model 2", "Model 3", "Model 4")
kable(data.frame(sum.stats), digits = 4, format.args = list(big.mark = ","))
```

	RMSE	AdjustedRSquared
Model 1	27,426.15	0.5246
Model 2	27,617.56	0.5161
Model 3	28,208.83	0.4947
Model 4	28,195.58	0.4947

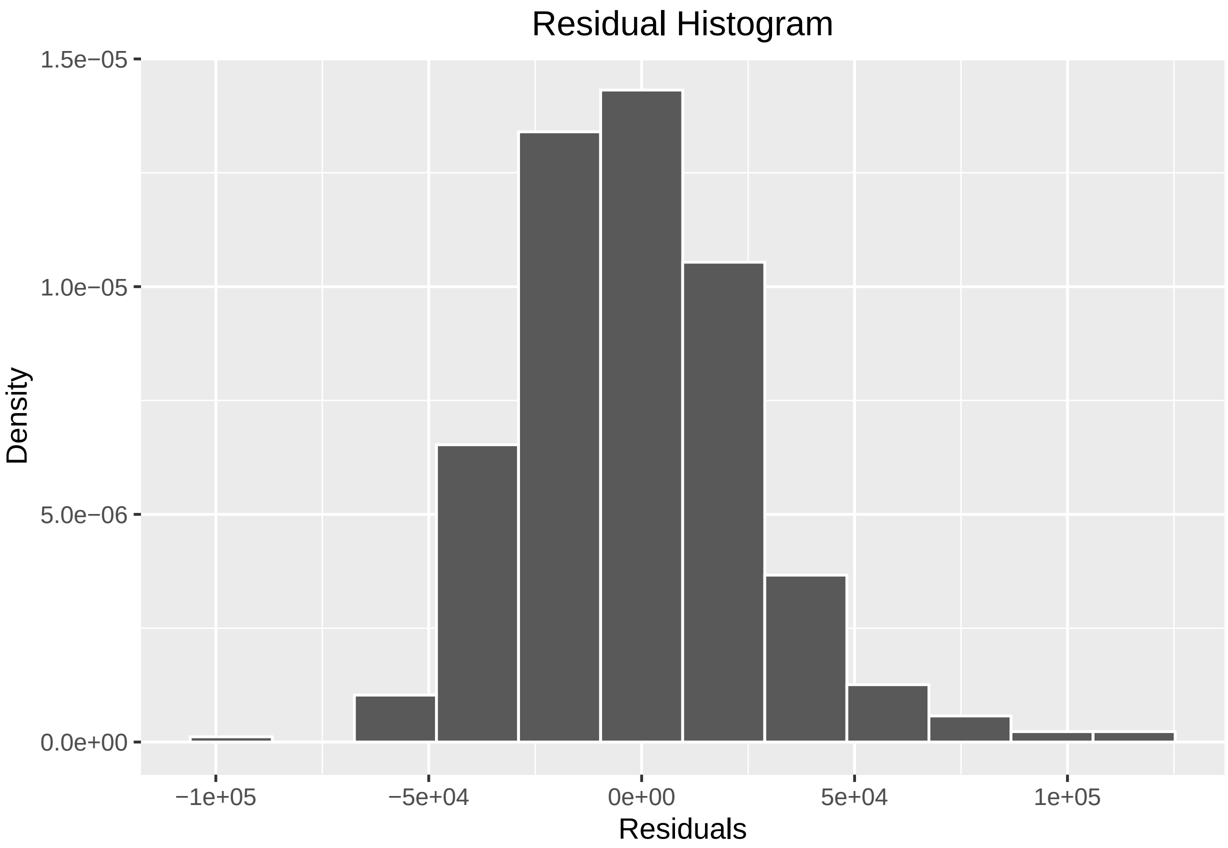
After choosing the final model, the model was assessed using the validation set. Predictions were added to the validation set, and a new RMSE value was calculated accordingly.

```
LinReg.add <- final.validate %>%
  add_predictions(Reg4) %>%
  add_residuals(Reg4) %>%
  rename(Pred_Median.debt.in.collections = pred,
         residuals = resid)

LinReg.add %>%
  rmse(truth = Median.debt.in.collections,
       estimate = Pred_Median.debt.in.collections) %>%
  kable()
```

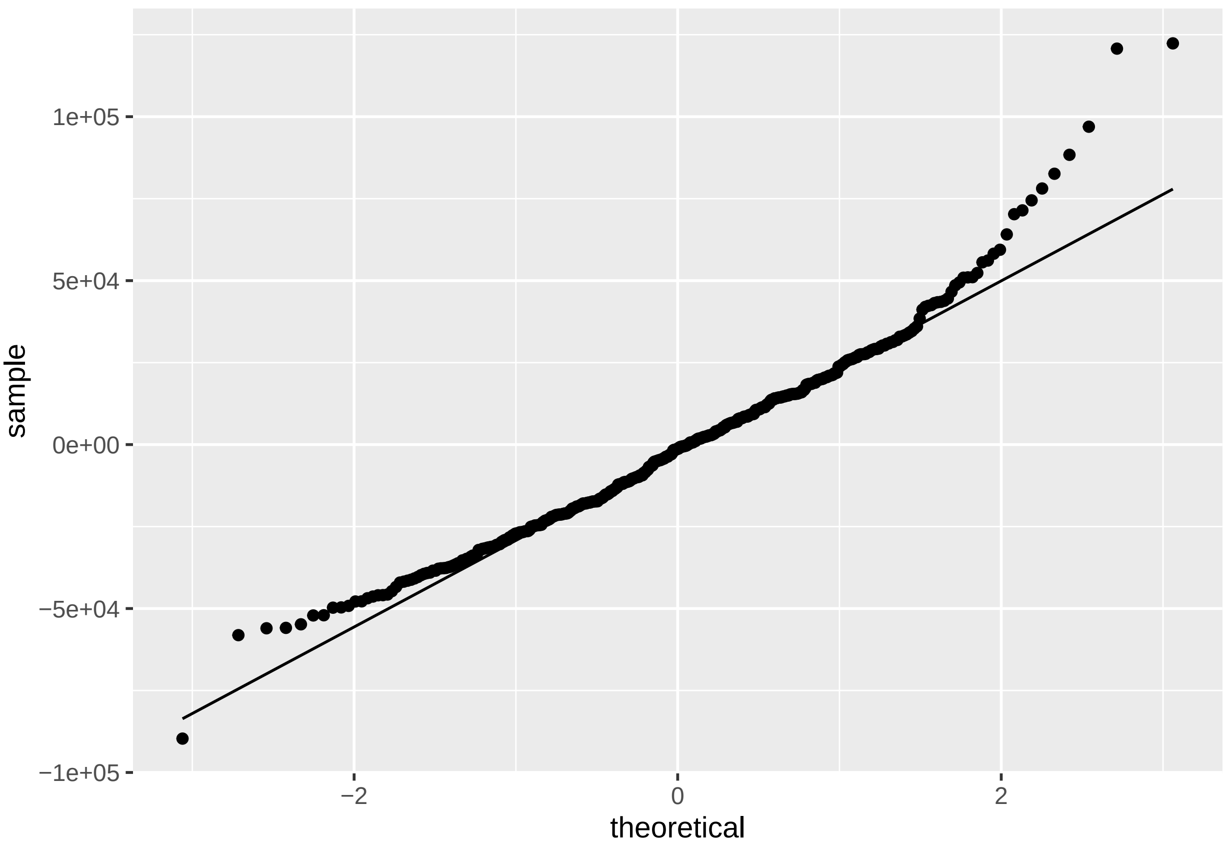
.metric	.estimator	.estimate
rmse	standard	27494.53

Finally, in order to assess the final model and verify the linear regression model assumptions, the team created a histogram of the residuals and a QQ-plot in order to verify the normality assumption. Residual plots were made for the final overall model and each explanatory variable in order to verify the constant variance assumption.

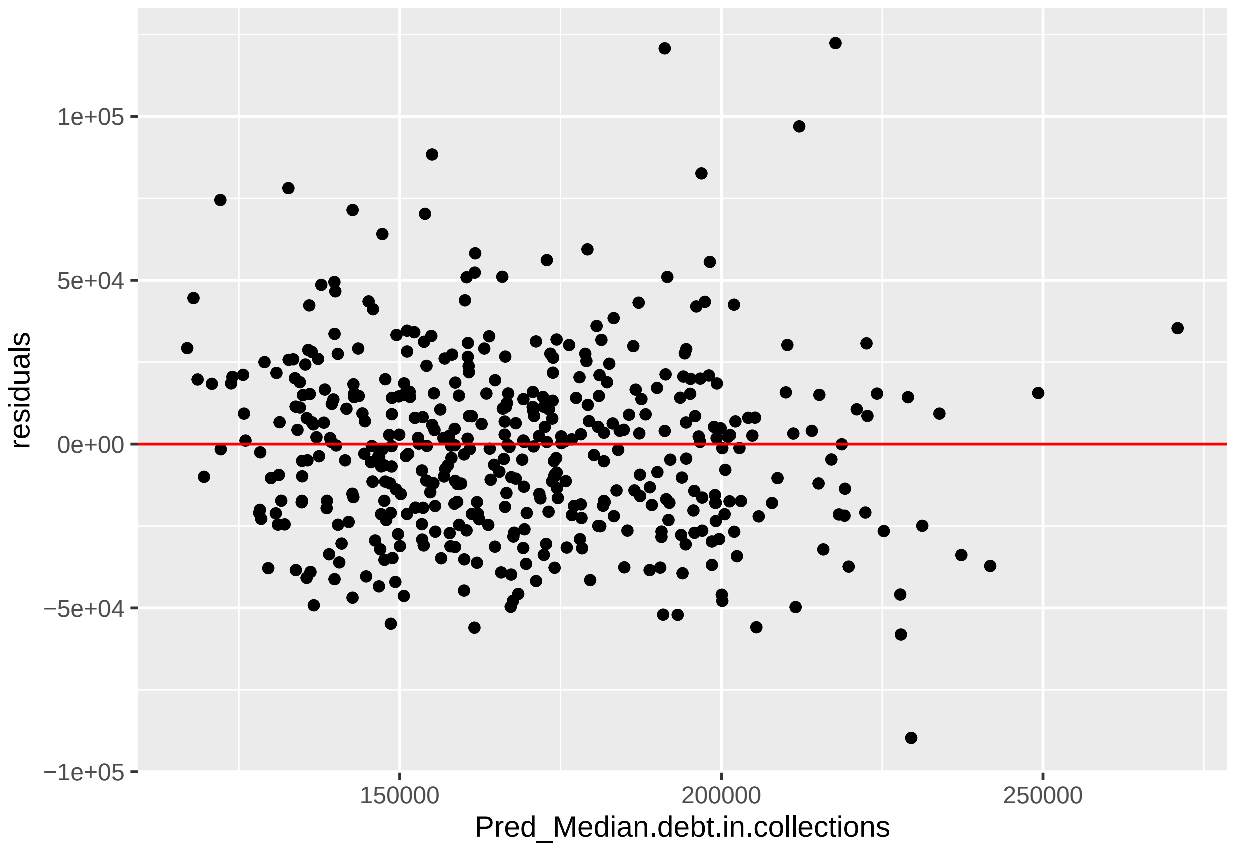
```
# Create histogram and QQ-plot to assess normality.
LinReg.add %>%
  ggplot(aes(x = residuals)) +
  geom_histogram(aes(y = ..density..), bins = 12, color = "white") +
  ggtitle("Residual Histogram") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Residuals") + ylab("Density")
```



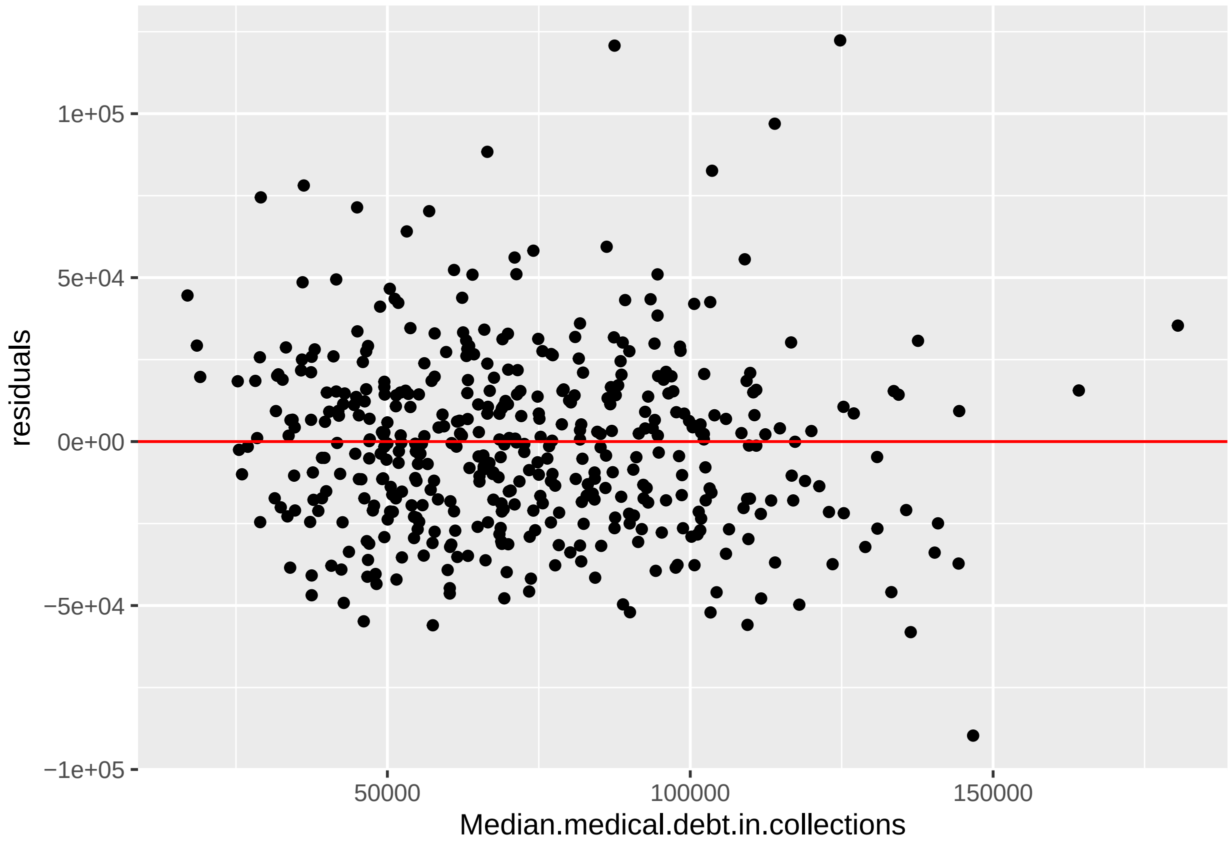
```
LinReg.add %>%
  ggplot(aes(sample=residuals)) +
  stat_qq() + stat_qq_line()
```



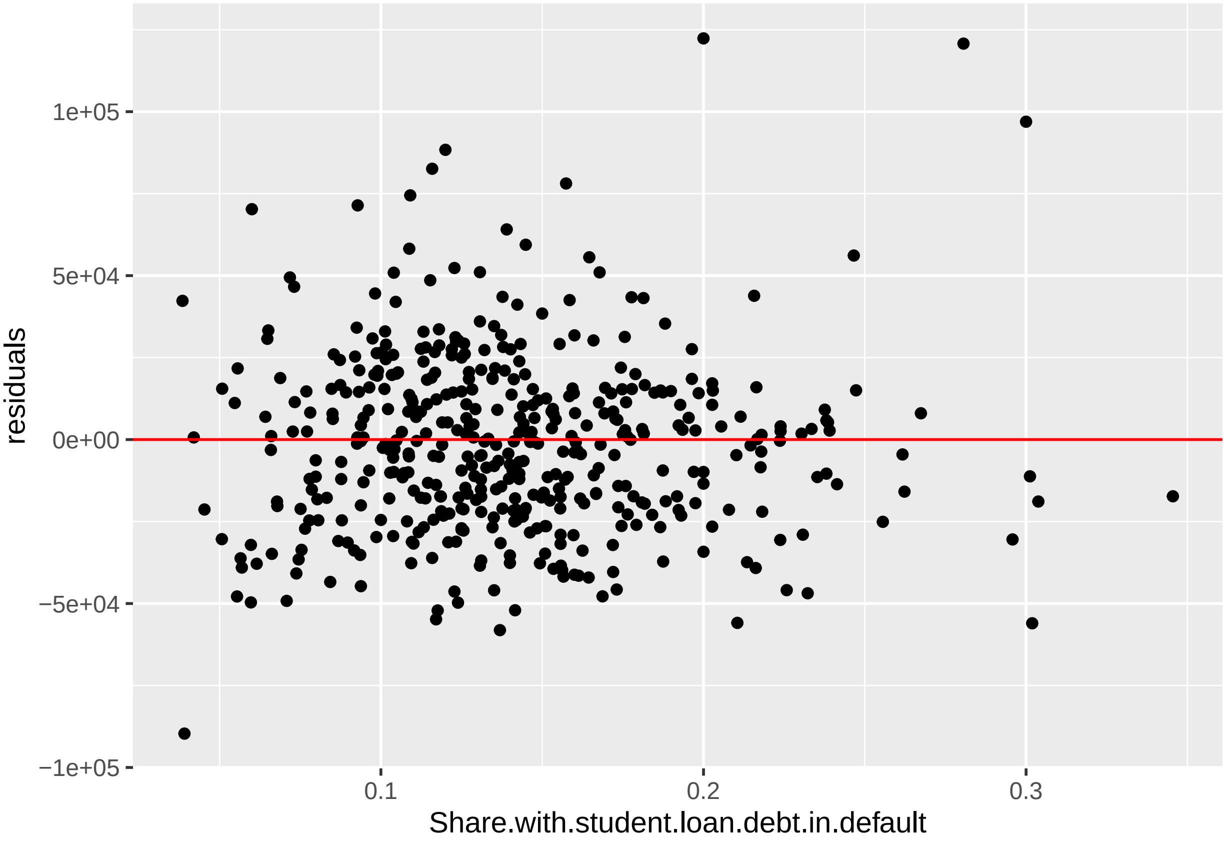
```
# Create residual plots to assess constant variance.
LinReg.add %>%
  ggplot(aes(x = Pred_Median.debt.in.collections, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red")
```



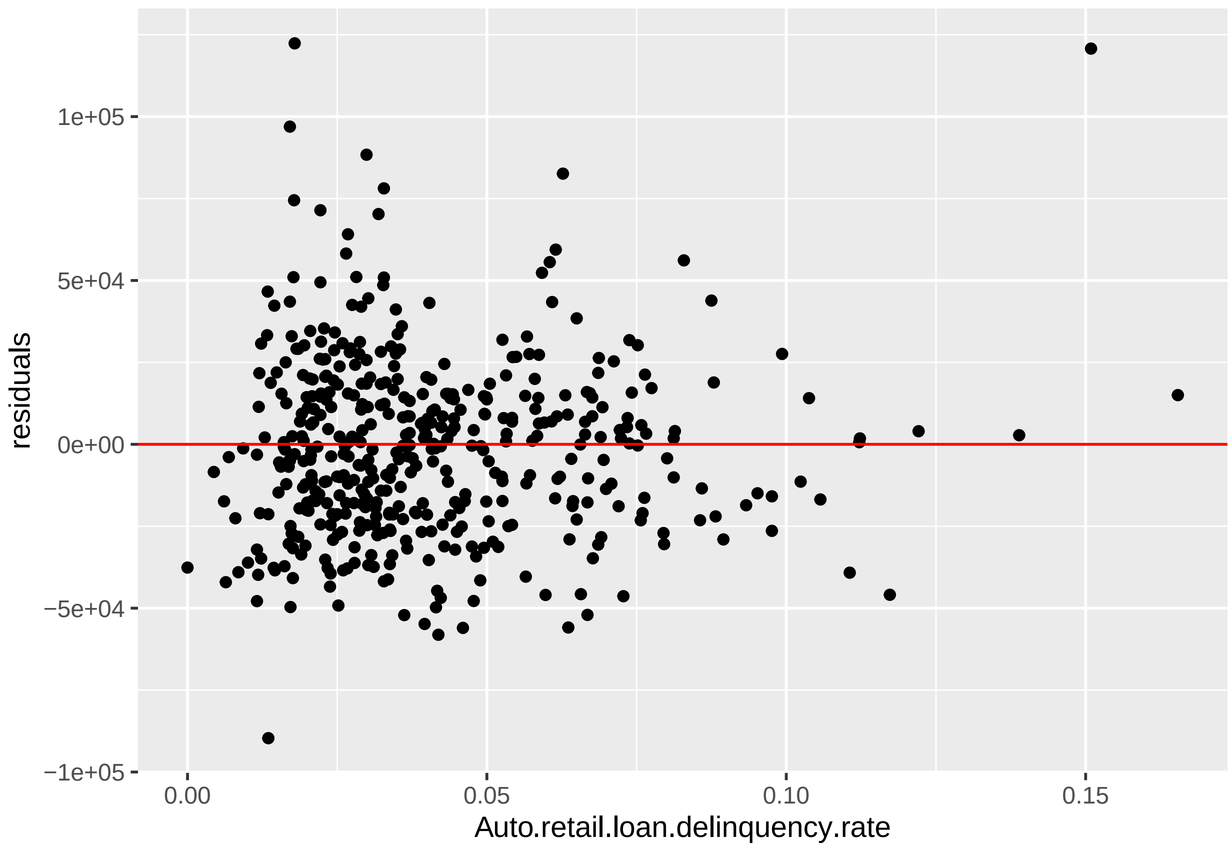
```
LinReg.add %>%
  ggplot(aes(x = Median.medical.debt.in.collections, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red")
```



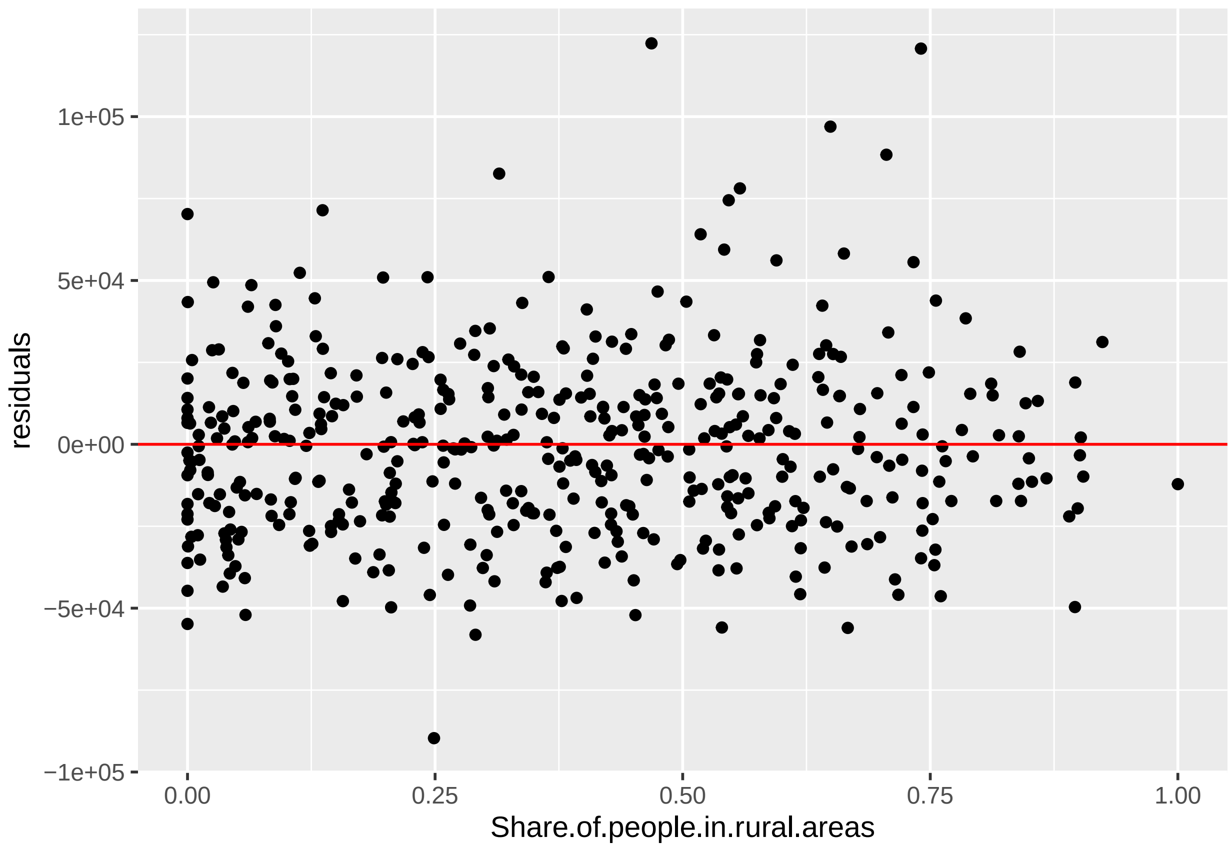
```
LinReg.add %>%
  ggplot(aes(x = Share.with.student.loan.debt.in.default, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red")
```



```
LinReg.add %>%
  ggplot(aes(x = Auto.retail.loan.delinquency.rate, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red")
```



```
LinReg.add %>%
  ggplot(aes(x = Share.of.people.in.rural.areas, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red")
```



residuals
1e+05
5e+04
0e+00
−5e+04
−1e+05
0.00
0.25
0.50
0.75
1.00
Share.of.people.in.rural.areas

Appendix D: Comment incorporation

General comments

Across all sections of the project, the team received comments regarding the formatting of the report, specifically minor errors in the readability of graphics, R code, and Excel screenshots, formatting references and citations in Chicago style format, and errors in grammar and spelling. The team addressed these concerns by increasing graphic sizes in R and Excel, adding more comments and white space to aid the reader in understanding R code, adjusting the formatting of citations to match Chicago style citations and references, and correcting grammatical and spelling errors in text explanations, captions, and graphical axis labels. Throughout the project the team also addressed comments regarding repetitive and elaborate explanations by using more concise and clear language in our explanations and summaries. Finally, the team revised the project to include non-technical language and professional language where suggested.

Deliverable 1 comments

In the Deliverable 1 Introduction section, the team received various comments to explain and elaborate on the context behind our business problem. The team addressed these comments by including more sources to aid in explaining how inflation has on loans, the increased demand for loans, and how this increased demand for loans can potentially influence loaning entities. Additionally, the team received a suggestion to isolate a single state or region rather than all states in the United States. The team chose not to incorporate this suggestion because while borrowing trends differ across the nation, the team cited examples of similar trends occurring across the entire nation, for example, the overall increases in college tuition, healthcare costs, and housing prices.

In the subsequent Business Problem section, the team added citations and a more detailed explanation on risky loans, the mortgage crisis, and predatory loans to incorporate feedback on providing more context to the team's business problem of minimizing risky loans and rates of default. Additionally, the team received a suggestion to address confounding factors when using only geographic location to explore loan trends. In order to address potential confounding factors relating to geographic locations, the team also included demographic variables to contextualize loan information and potential trends. These demographic variables include share of people in rural areas, share of people of color, and average household income for each county. Finally, a concern was raised regarding the time aspect of loan trends. The team took note of this comment but recognized that our analysis does not involve a time series analysis. All observations and measurements were recorded during the same time period. The team is more interested in analyzing geographic trends in different counties and regions across U.S. states.

The team further condensed our explanation of the intended audience to include both state and local banks to address comments regarding the wide scope of our intended audience. The team also made improvements in our justification on how our team's predictive analysis on loans in default would benefit our intended audience by providing an alternative to assessing customers' risk of default through credit scores.

Deliverable 2 comments

In the Data collection section in Deliverable 2, the team received a comment suggesting a more detailed explanation on how analysis would be carried out. However, the team chose not to incorporate this comment into our revisions because the intent of the data portion is not to analyze the data and find trends in the data, but rather to explain how the team plans to prepare and use the data in order to address our business problem.

In the Data preparation section, the team received comments to consider how missing and NA values impact the usefulness of our data. In response, the team detailed how the number of missing data were not a concern, as most states are still represented in the data and counties with missing values only had missing values for a select number of columns. The team was also asked to include more explanation as to why we chose to remove the white communities and community of color variables. We added more explanation as to why we removed these columns and referenced other columns related to race that were included in our data.

In the Data preparation details in R section, the team received a comment on code redundancy due to the length of our code to prepare our data. The team chose not to incorporate this feedback because while the code is long and some steps were repeated, this is because the team was using data from various worksheets within a single Excel file. Many of the data cleaning steps had to be repeated and applied to various sheets in order to merge the data successfully. For example, the team had to remove the double headers and rename the columns for each worksheet.

In the Data preparation details in Excel section, to improve readability of this section, the team altered the description of the data preparation process to explain the process that is specific to Excel, as opposed to relating the Excel process to the data preparation process in R. By doing so, readers will be able to understand the Excel data preparation process without having to reference the previous section. Another comment received by the group was regarding the utilization of VLOOKUP and the danger of including duplicates using only VLOOKUP. To address the duplicate observation recommendation, the team also implemented conditional formatting to ensure that there are no duplicate observations and that counties with the same name from different states will also be represented.

Deliverable 3 comments

The first comment received regarding the descriptive analytics section was suggesting the use of a correlation matrix rather than two different bar plots. The team chose to use two bar plots in our descriptive analysis because we believed that it would be more intuitive to readers without a technical background as opposed to a correlation matrix. The bar plots provided do give useful information on the general patterns between the share of debts in each county and the average household income or the share of people in rural areas.

A similar comment suggested adding the equation of the linear model in the the scatterplot. The team decided that, while adding the equation of the linear model could be useful to someone with a more technical background, we chose not to include this equation, as we are only interested in seeing the general trend between average household income and share of people in rural areas in each county. Including the equation of the trendline could confuse readers without a technical background.

The team also converted all graphics to R in order to maintain consistency in the final report after receiving a comment suggesting this. The last comment received in this section suggested adding a visual division to the bar plots to help visualize distinct groups of states. The team addressed this suggestion by adding a dividing line between the two categories of states as well as adding a text annotation to the graphics on each side that more explicitly explains what each group of states represents.

Deliverable 4 comments

In the predictive analytics section, many comments suggested explaining our justification in choosing our response variable. The team incorporated this explanation briefly into this section. Next, the team made sure to verify the assumptions for the final model as well as the RMSE for the final model after receiving a comment regarding the missing assumptions and new RMSE. The team also elaborated more on how we found the best model using linear regression after receiving similar comments in both the assessments and the results sections. We also incorporated a table with the different RMSE and adjusted R-squared models in order to demonstrate how we assessed and compared each model.