Project 1 – Group 12

Financial Analysis Project Overview

Caleb Meinke, Luisa Murillo,

Gavin Bozan, and Charles Arnold

Database Analytics Bootcamp – 12/11/24

Introduction:

For project 1, our group chose to research and analyze a financial dataset of 2,000 individual consumers to determine what relationships existed between consumer attributes, consumer habits, and overall financial health. Data for this effort is sourced from the “Financial Transactions Dataset: Analytics” dataset located at https://www.kaggle.com/datasets/computingvictor/transactions-fraud-datasets/code.

Inspiration for this research stemmed from multiple resources, including the Center for Economic and Policy Research study, “Before and After the Pandemic: Income Volatility, Health Care Affordability, and Debt” and the Financial Health Network’s study, “Financial Health Pulse 2023 U.S. Trends.”

Using the Kaggle dataset, the project team isolated three key categories of consumer information from the dataset, which served as the foundation for subsequent research questions. These categories include the following:

- Consumer demographic information:
    - Age
    - Gender
    - Income
- Consumer Financial Attributes:
    - Total debt
    - Credit score
    - Total number of payment cards
- Geographic information
    - State of residence
    - Region of residence

Objectives:

To determine the impact of these variables on individual financial health, the team leveraged the total debt-to-income ratio (DTI) for each consumer; calculated by dividing the individual consumer’s total debt by their yearly income.

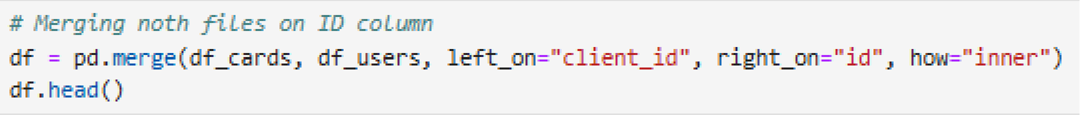
Various forms of the debt-to-income ratio are used in the financial industry to measure general financial health, creditworthiness, and financial stress. This is a known consideration in lending, credit scoring, and consumer risk assessment; providing a glance at individual debt payment obligations compared against income to determine overall financial health.

For the purposes of this study, analysis of individual variables in each consumer category, their interrelationships, and their correlation to DTI provided the basis for assessing their influence or impact on individual financial health. Following the above outline, this research hypothesized the following:

1. Customer demographics including age, gender, and income influence individual financial health (as measured by the DTI ratio)
2. Individual financial attributes including total debt, total card count, and credit score influence individual financial health (as measured by the DTI ratio)
3. Customer geographic information including state, region, and per-capita-income influence individual financial health (as measured by the DTI ratio)

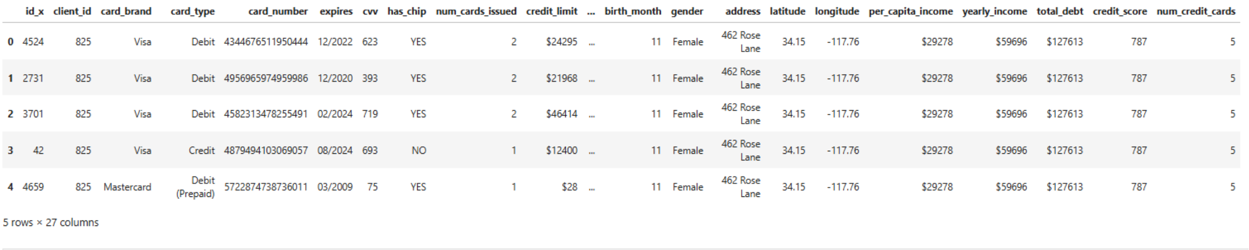
Data organization and cleaning:

The dataset required significant reorganization and cleaning. Initially, the dataset contained two CSV files, one file containing individual customer attributes and another containing payment transaction attributes. Merging these files into a single dataframe was necessary to achieving the stated research objectives. As a result, the team merged the CSVs on the consumer client ID columns contained in both databases.



```
# Merging noth files on ID column
df = pd.merge(df_cards, df_users, left_on="client_id", right_on="id", how="inner")
df.head()
```

Python code for merging CSVs



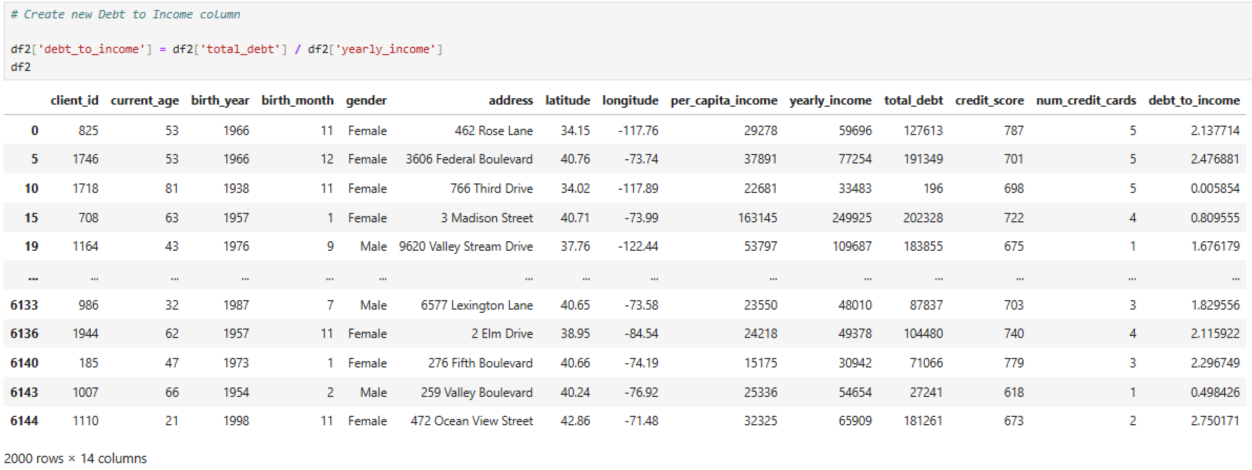
	id_x	client_id	card_brand	card_type	card_number	expires	cvv	has_chip	num_cards_issued	credit_limit	...	birth_month	gender	address	latitude	longitude	per_capita_income	yearly_income	total_debt	credit_score	num_credit_cards
0	4524	825	Visa	Debit	4344676511950444	12/2022	623	YES	2	$24295	...	11	Female	462 Rose Lane	34.15	-117.76	$29278	$59696	$127613	787	5
1	2731	825	Visa	Debit	4956965974959986	12/2020	393	YES	2	$21968	...	11	Female	462 Rose Lane	34.15	-117.76	$29278	$59696	$127613	787	5
2	3701	825	Visa	Debit	4582313478255491	02/2024	719	YES	2	$46414	...	11	Female	462 Rose Lane	34.15	-117.76	$29278	$59696	$127613	787	5
3	42	825	Visa	Credit	4879494103069057	08/2024	693	NO	1	$12400	...	11	Female	462 Rose Lane	34.15	-117.76	$29278	$59696	$127613	787	5
4	4659	825	Mastercard	Debit (Prepaid)	5722874738736011	03/2009	75	YES	1	$28	...	11	Female	462 Rose Lane	34.15	-117.76	$29278	$59696	$127613	787	5

5 rows × 27 columns

Merged dataframe

From there, the data also required data cleaning on several columns. This included removing dollar signs from monetary values in the “credit_limit,” “per_capita_income,” “yearly_income,” and “total_debt” columns to convert these figures from strings to integers. Additionally, data cleaning efforts resulted in dropping several columns unessential to the research objectives and removing duplicate entries of client ID financial transactions. The refined data frame provided a consolidated view of the variables essential to the research objectives, which included calculation of the DTI for each client

ID in the dataframe. Code generated to achieve this objective saved the output to a new "debt_to_income" column, which staged the dataframe for data visualization.



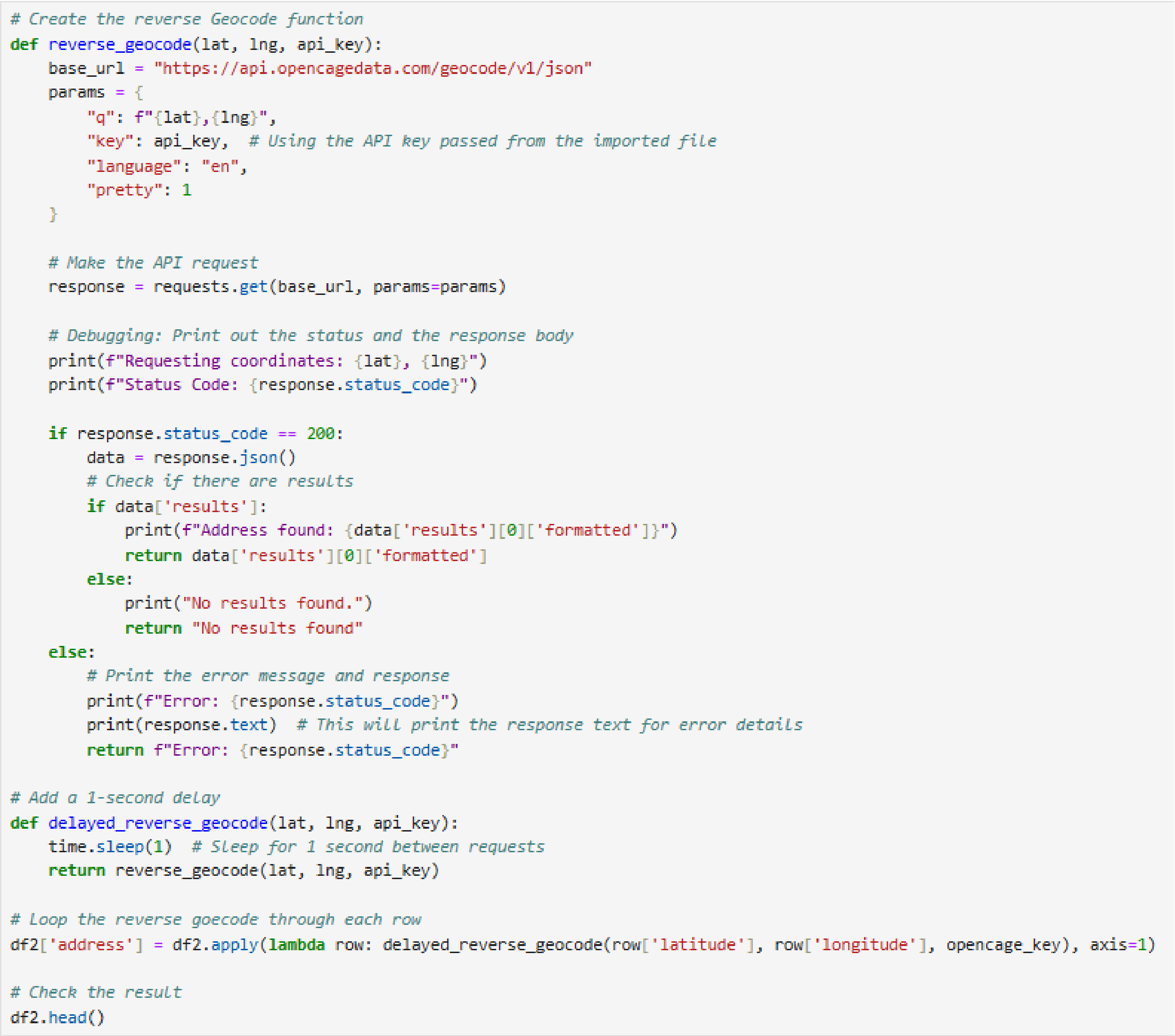
```
# Create new Debt to Income column

df2['debt_to_income'] = df2['total_debt'] / df2['yearly_income']
df2
```

	client_id	current_age	birth_year	birth_month	gender	address	latitude	longitude	per_capita_income	yearly_income	total_debt	credit_score	num_credit_cards	debt_to_income
0	825	53	1966	11	Female	462 Rose Lane	34.15	-117.76	29278	59696	127613	787	5	2.137714
5	1746	53	1966	12	Female	3606 Federal Boulevard	40.76	-73.74	37891	77254	191349	701	5	2.476881
10	1718	81	1938	11	Female	766 Third Drive	34.02	-117.89	22681	33483	196	698	5	0.005854
15	708	63	1957	1	Female	3 Madison Street	40.71	-73.99	163145	249925	202328	722	4	0.809555
19	1164	43	1976	9	Male	9620 Valley Stream Drive	37.76	-122.44	53797	109687	183855	675	1	1.676179
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6133	986	32	1987	7	Male	6577 Lexington Lane	40.65	-73.58	23550	48010	87837	703	3	1.829556
6136	1944	62	1957	11	Female	2 Elm Drive	38.95	-84.54	24218	49378	104480	740	4	2.115922
6140	185	47	1973	1	Female	276 Fifth Boulevard	40.66	-74.19	15175	30942	71066	779	3	2.296749
6143	1007	66	1954	2	Male	259 Valley Boulevard	40.24	-76.92	25336	54654	27241	618	1	0.498426
6144	1110	21	1998	11	Female	472 Ocean View Street	42.86	-71.48	32325	65909	181261	673	2	2.750171

2000 rows × 14 columns

Cleaned Dataframe

Following the addition of the DTI ratio information to the data frame, the final step in the data cleaning effort required reverse geocoding of the latitude and longitude information in the original CSV files. This effort produced the state of residence for client ID, a critical component to the assessment of geographical relationship to DTI. Using the OpenCage Geocoding API, the project team generated code to loop through the data frame and extract the exact address of each latitude and longitude intersection via the OpenCage database, saving this address off to a new column (ChatGPT, 2024). This finalized the dataframe preparation efforts, permitting research and data visualization for the first of the project team's three research questions.

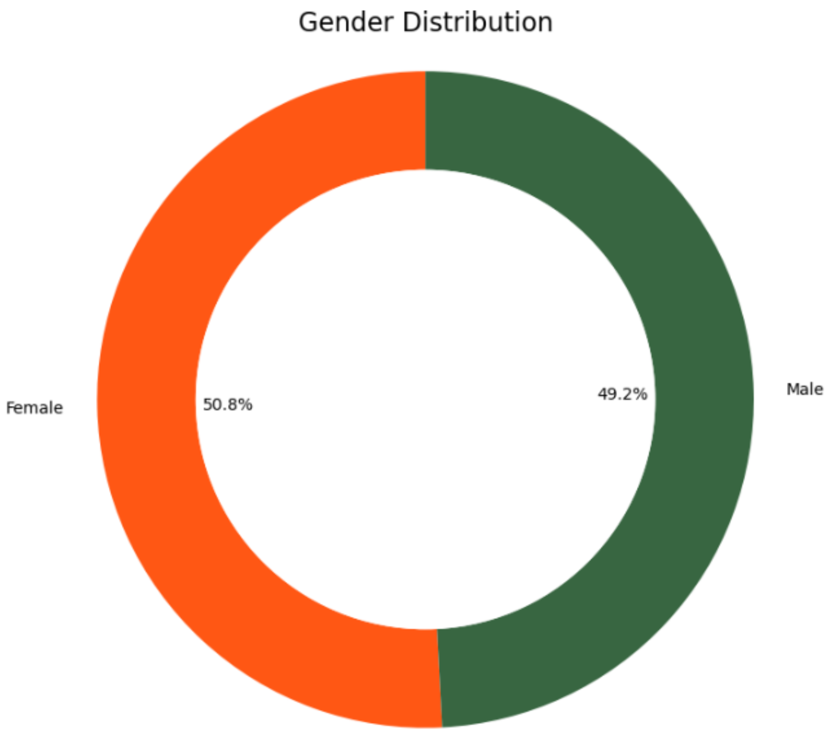


```python
# Create the reverse Geocode function
def reverse_geocode(lat, lng, api_key):
    base_url = "https://api.opencagedata.com/geocode/v1/json"
    params = {
        "q": f"{lat},{lng}",
        "key": api_key,  # Using the API key passed from the imported file
        "language": "en",
        "pretty": 1
    }

    # Make the API request
    response = requests.get(base_url, params=params)

    # Debugging: Print out the status and the response body
    print(f"Requesting coordinates: {lat}, {lng}")
    print(f"Status Code: {response.status_code}")

    if response.status_code == 200:
        data = response.json()
        # Check if there are results
        if data['results']:
            print(f"Address found: {data['results'][0]['formatted']}")
            return data['results'][0]['formatted']
        else:
            print("No results found.")
            return "No results found"
    else:
        # Print the error message and response
        print(f"Error: {response.status_code}")
        print(response.text)  # This will print the response text for error details
        return f"Error: {response.status_code}"

# Add a 1-second delay
def delayed_reverse_geocode(lat, lng, api_key):
    time.sleep(1)  # Sleep for 1 second between requests
    return reverse_geocode(lat, lng, api_key)

# Loop the reverse goecode through each row
df2['address'] = df2.apply(lambda row: delayed_reverse_geocode(row['latitude'], row['longitude'], opencage_key), axis=1)

# Check the result
df2.head()
```

Reverse geocoding function

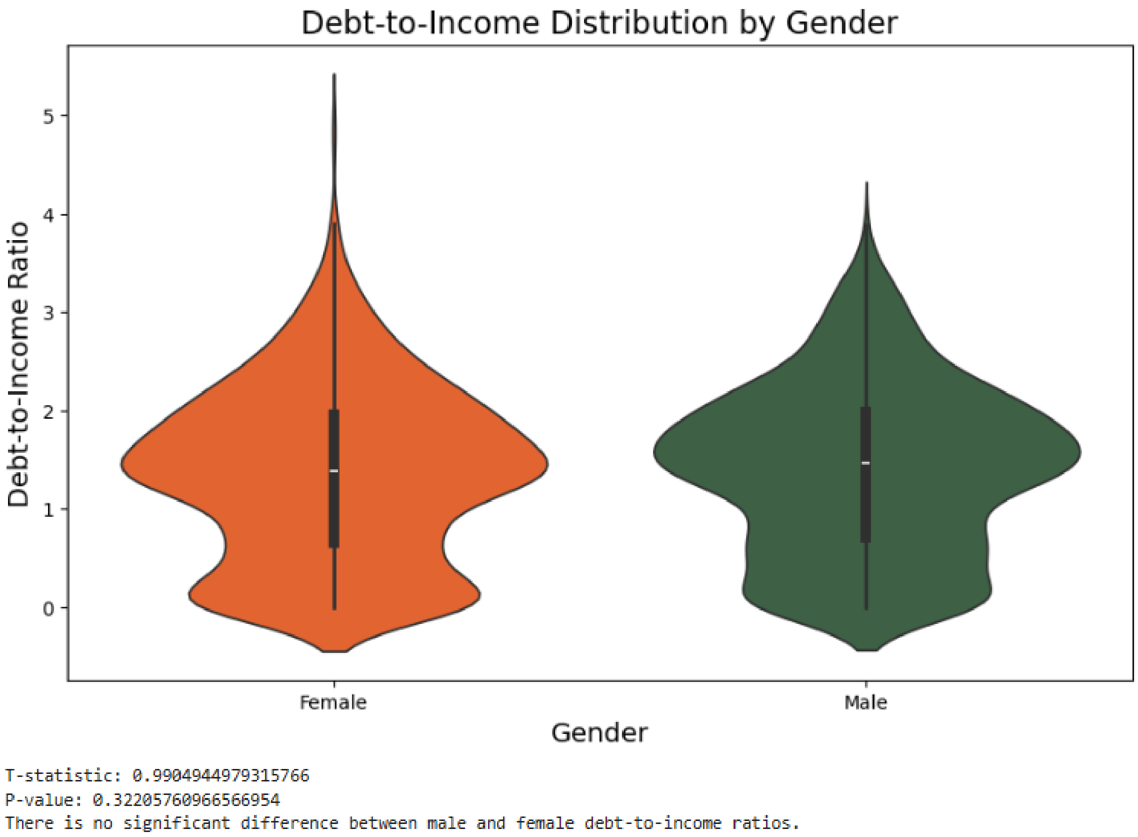
Question 1: How do customer demographics correlate with customer financial health?

The goal of this analysis was to explore the distribution of debt-to-income across genders and identify any significant differences between women and men.



We took the combined dataset to breakout the men and women from the total population and displayed the results in a doughnut plot. The dataset had 50.8% female and 49.2% males. While we were hoping for a 50/50 distribution of men and women in the dataset, the ratio produced is in line with U.S. population distribution of men and women.

To further analyze the impact of gender on financial health, the team decided to do a violin plot to visualize the distribution of debt-to-income ratios for both men and women.

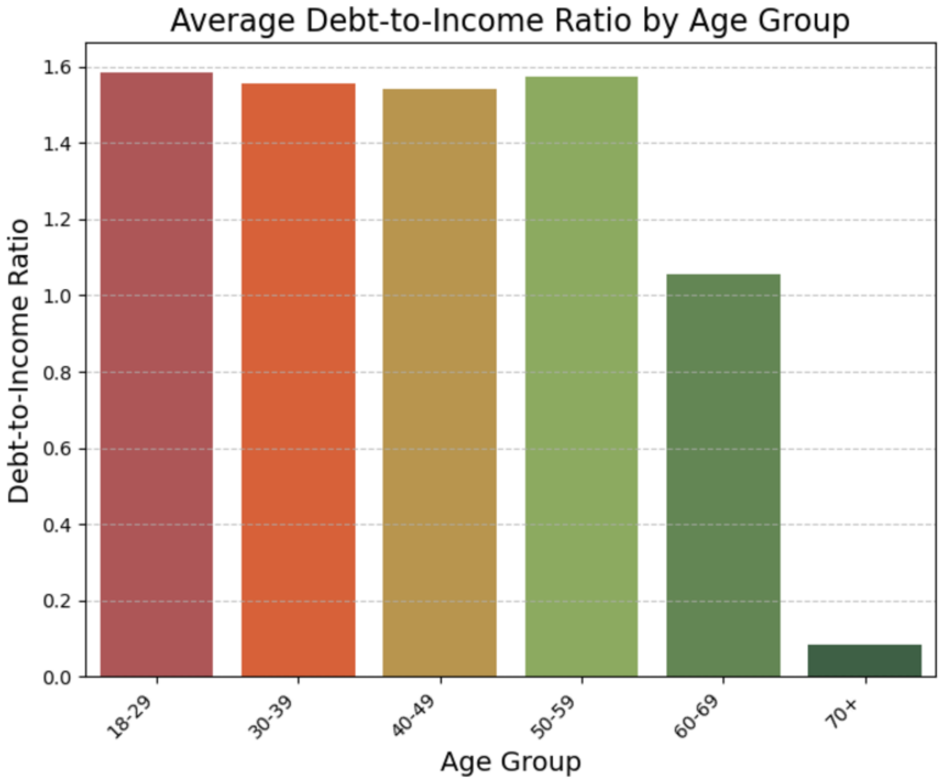


```
T-statistic: 0.9904944979315766
P-value: 0.32205760966566954
There is no significant difference between male and female debt-to-income ratios.
```

The shape of the violin plot provides insight into the density and spread of debt-to-income distribution by gender, with the wider parts of the plot indicating higher proportions of each gender population falling in those debt-to-income ranges. Assessment of the two populations showed a high degree of similarity in the distribution of debt-to-income by gender. The only exceptions to this observation are a slight degree of variation in the 0.25 to 1.0 range between genders and in the high end of the debt-to-income range, where a small population of females showed debt-to-income ratios exceeding the upper range of the debt-to-income ratios observed in the male segment of the population. Given these differences, the team conducted a T-test to determine the degree to which these populations showed statistically significant differences.

The T-Test gauging the similarities between the male and female populations produced a 99% confidence variable, indicating the debt-to-income ratio across genders revealed no significant difference men and women. This suggests that gender alone may not be a key determinant in the distribution of debt-to-income ratios in this dataset.

Final assessment of demographic information included analysis of how the debt-to-income ratio varies across different age group. This view helps understand if age plays a role is shaping financial behavior, particularly in relation to debt levels.



We divided the dataset into six age bins: 18-29, 30-39, 40-49, 50-59, 60-69, and 70+, and computed the average debt-to-income ratio for each age group.

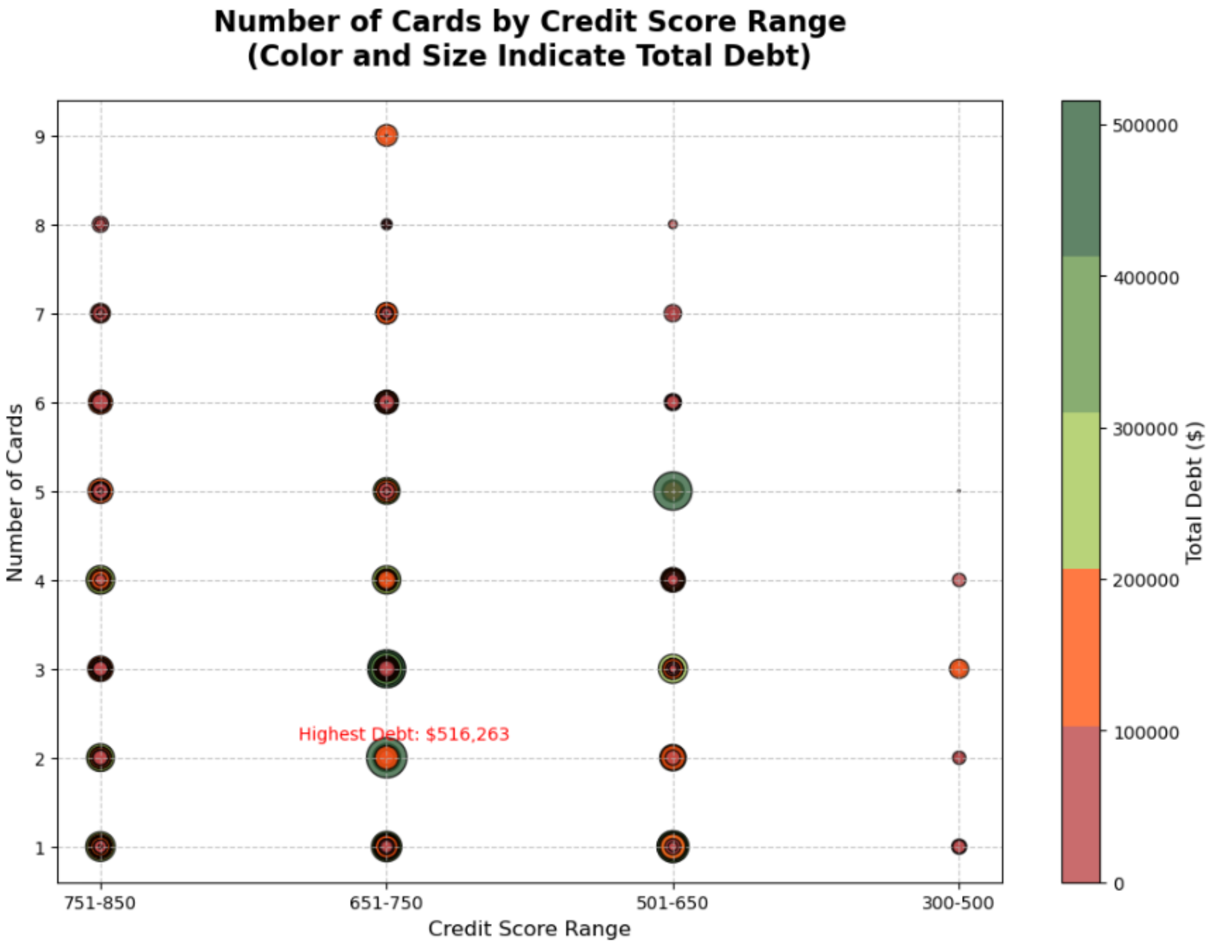
From the chart, we observe that the average debt-to-income ratio is higher for younger age groups (18-29) and tends to decrease with age, starting in the 60's with a major drop specially in the 70+ category. This reflects lower debt levels as people approach retirement.

We noticed a downward trend in the data leading up to the 50-59 age range, at which point the DTI experienced a slight increase. This may be a result of individuals in this age group incurring more debt as their children are starting college, resulting in a slight rise in the debt-to-income ratio. However, this observation is temporary, as debt-to-income levels drop dramatically in subsequent age ranges, suggesting a possible correlation between debt-to-income and age.

In summary, debt-to-income ratios tend to decrease with age, reflecting financial stability as individuals progress through their life, careers and approach retirement.

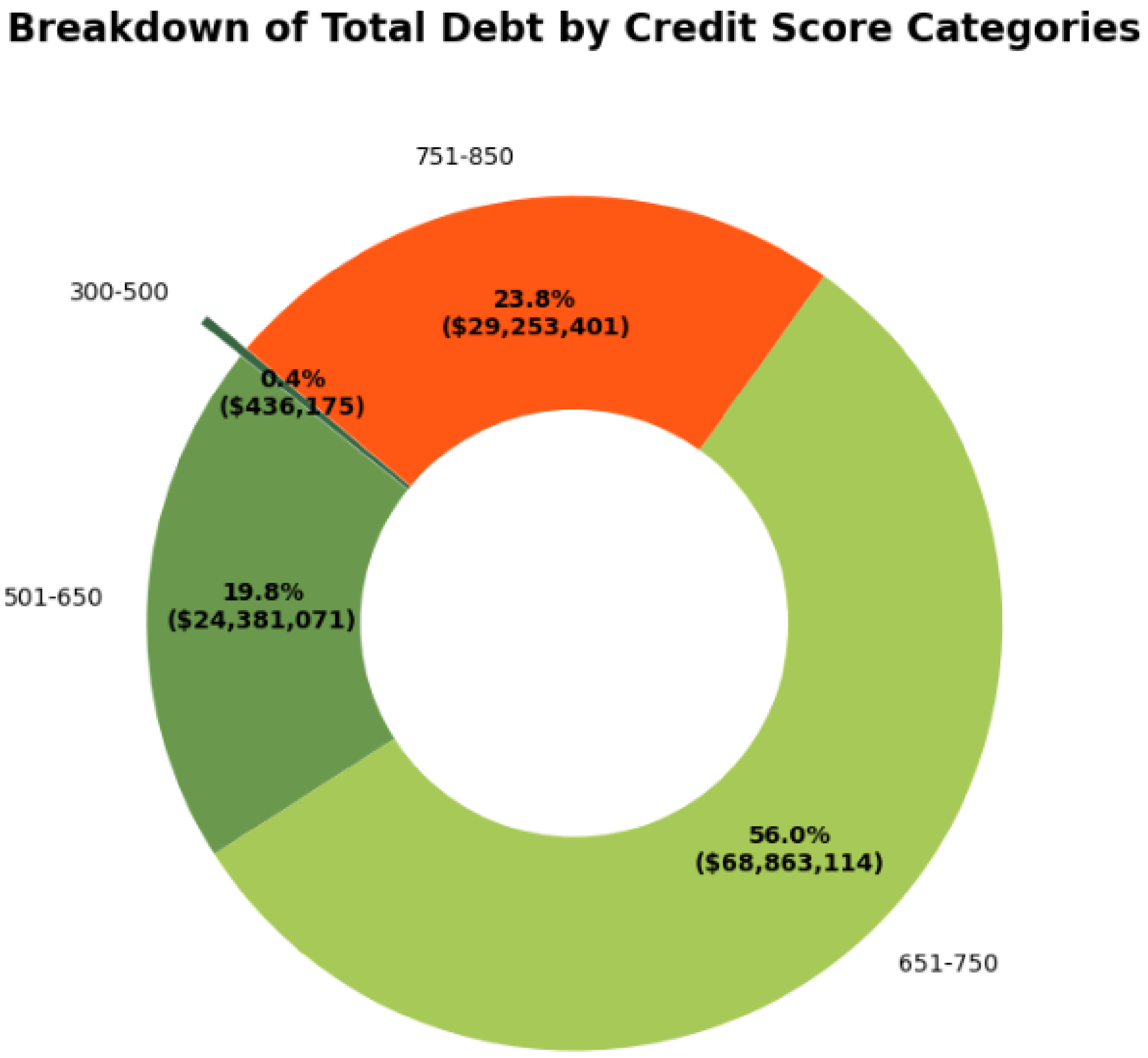
Question 2: How do financial attributes correlate to customer financial health?

In the second phase of this research, the project team assessed how financial factors like credit scores, total debt, and the number of credit cards relate to customer financial health. Particularly, the team focused on how well people manage debt and credit, which is generally measured by the aggregation of financial risk elements that appear in each consumer's individual credit score. Focus on the credit score provides insight into the individual spending habits, borrowing habits, and the risks or benefits associated with different credit score groups. The project team hypothesized that there would be a negative relationship between these variables; namely, as credit scores rise, the DTI declines.

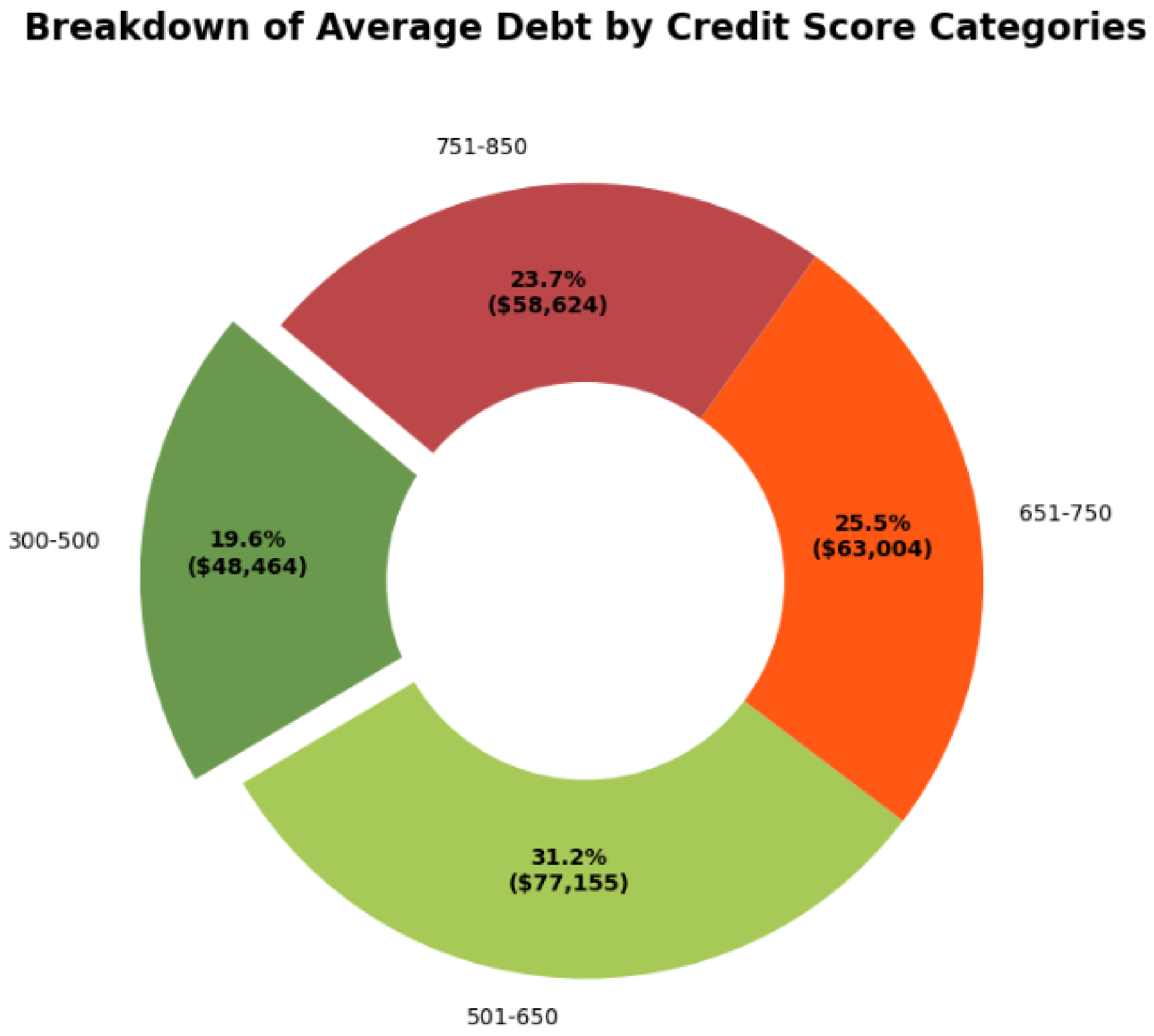


To test this hypothesis, the team analyzed the data frame from a variety of perspectives. The first view from this research lens included an assessment of the relationship between credit score and the number of payment cards possessed by individual consumers. The above chart plots this data, displaying the number of credit cards individuals have, broken down by credit score ranges, with color and size showing total debt. This visualization resulted in several key observations. First, people in the 651–750 range have the highest total debt, even though the individuals in this credit score range don’t always have the most payment cards. One interesting outlier is a case with over $500,000 in debt. This suggests that higher debt doesn’t necessarily depend on having more cards but could be influenced by income, credit limits, or spending habits. That said, this chart provides compelling evidence that the number of payment cards is not a clear predictor of financial health or debt levels. Other factors, such as income and financial behavior, play a more significant role.

To continue this analysis, the team proceeded to assess the relationship of total debt levels and credit score ranges to test the assumption that higher debt levels equated to lower overall credit scores.



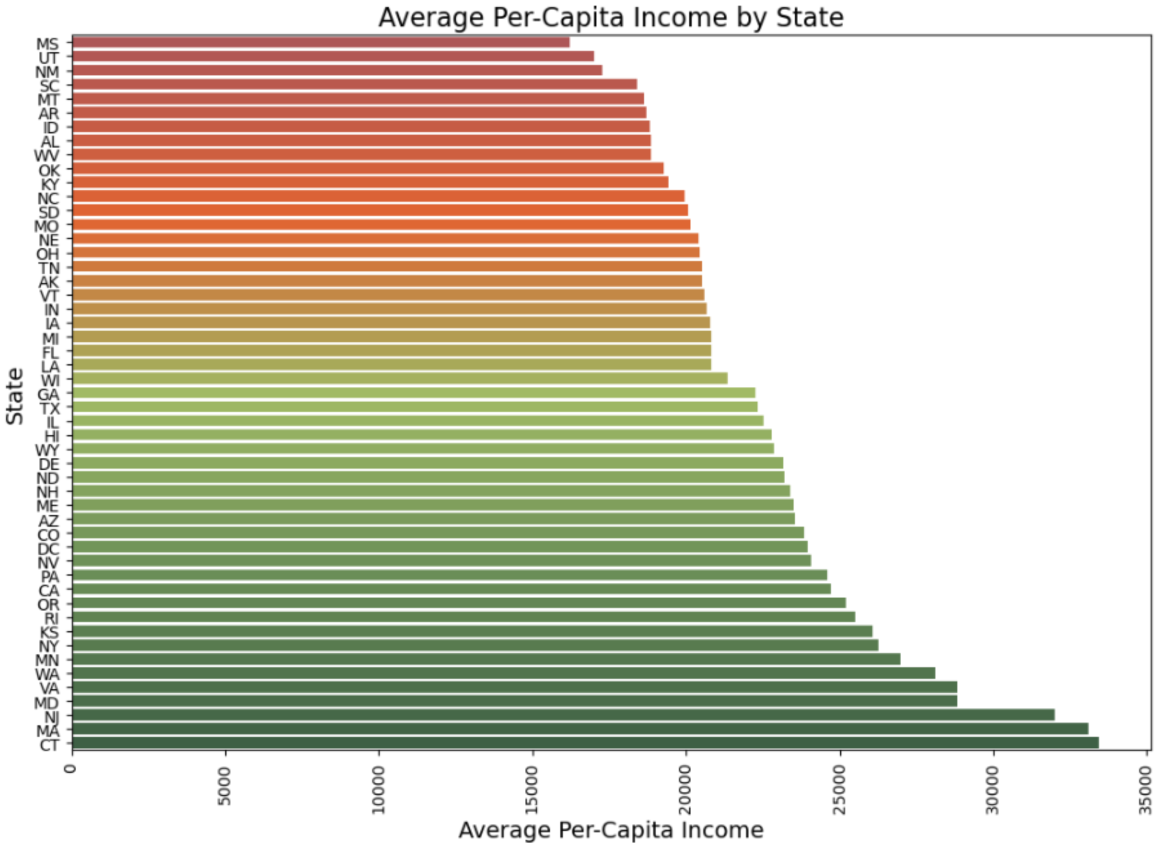
This donut chart displays the distribution of cumulative total debt across the specified credit score ranges, which revealed some key findings. First, over 56% of total debt is held by individuals in the 651–750 credit score range. This group is often considered reliable by lenders, explaining their access to higher credit limits. Second, the 300–500 credit score range holds less than 1% of the total debt. This is likely due to limited credit access rather than better financial health.



To further refine the review of debt and credit score, the team proceeded to assess the average debt for each credit score range. This showed some slight variation from the assessment of cumulative total debt by credit score range; however, the observed patterns remained mostly the same. Of note, the 501–650 score group has the highest average debt at over $77,000, suggesting this group borrows heavily, possibly facing higher interest rates and credit costs. Comparable to the previous chart, the 300–500 score range has the lowest proportion of debt in the population, with average debt at $48,464. This is likely due to restricted access to credit, making it difficult to accumulate substantial debt. Finally, the heaviest total average debt loads occurred among the mid-range credit scores. This analysis resulted in a few key takeaways. First, mid-range credit scores (651–750) often come with high total and average debt. This isn't necessarily indicative of poor financial health but could reflect higher income and spending habits. Second, the number of credit cards doesn't automatically mean higher debt. It can indicate greater reliance on credit but not necessarily financial instability. Further, lower credit scores often mean limited access to credit, resulting in a lower total concentration of debt among lower credit class individuals, which runs counter to the team's original hypothesis about the relationship between financial health indicators and credit scores.

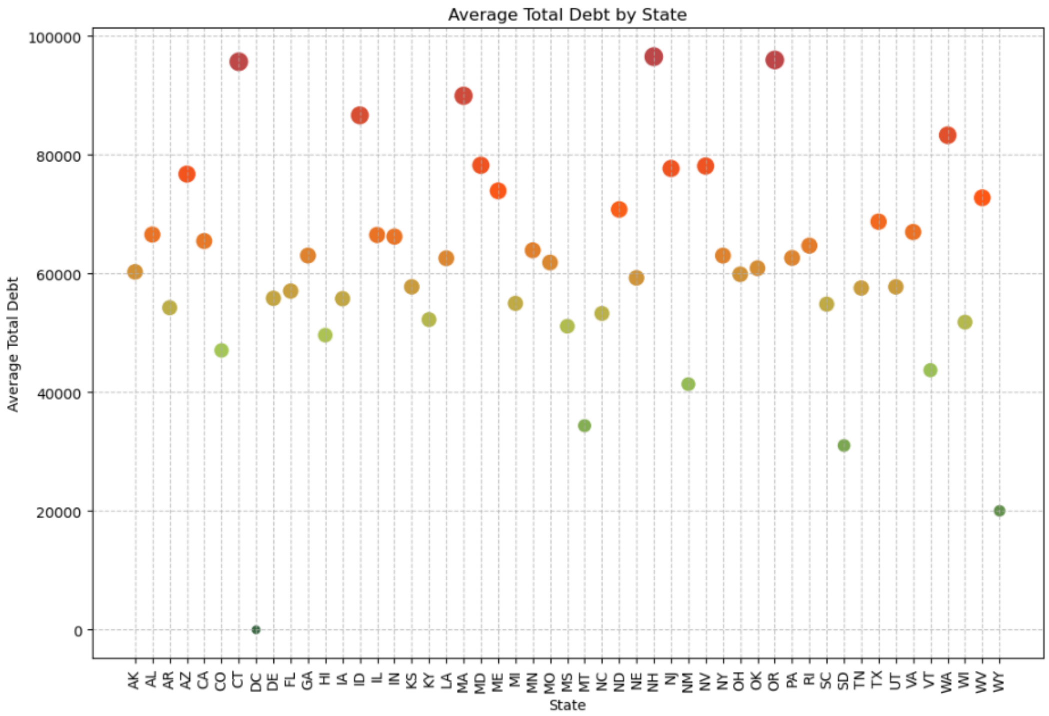
Question 3: How does customer geographic information correlate with customer financial health?

The initial assessment of the relationship between geographic location and financial health sought to establish a baseline understanding of the regional distribution of income in the United States. To achieve this, the project team created visualization of per capita income by state using a horizontal bar chart with an embedded color gradient signifying the lowest per capita income states (red) and the highest per capita income states (green).

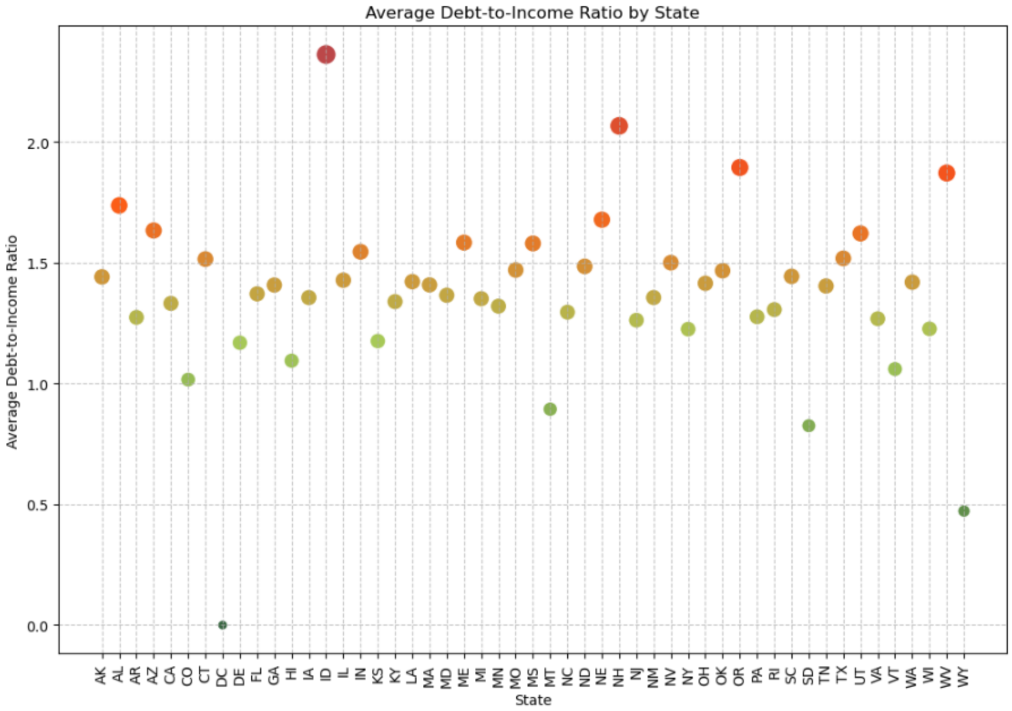


The initial assumption was that states and regions with lower per-capita income may rely on debt spending more than regions with notably higher per-capita-income, which would result in a higher DTI. As a result, the next iteration of the analysis looked at total debt by state, using the data frame population with debt and income outliers removed to smooth the data. This scatterplot used the same gradient scale to show the highest total debt (red) to lowest total debt (green) across the 51 geographic locations in the assessment.

Initially, observations showed clear distinctions between the total debt from one state to the next; however, there was no clear regional trends that could observed in this data. To better understand how these debt figures compared against total income, the next iteration of the analysis looked at average DTI by state. This approach demonstrated much less variance in the data by state, which suggested that there may not be a strong geographical influence on financial health in the United States.

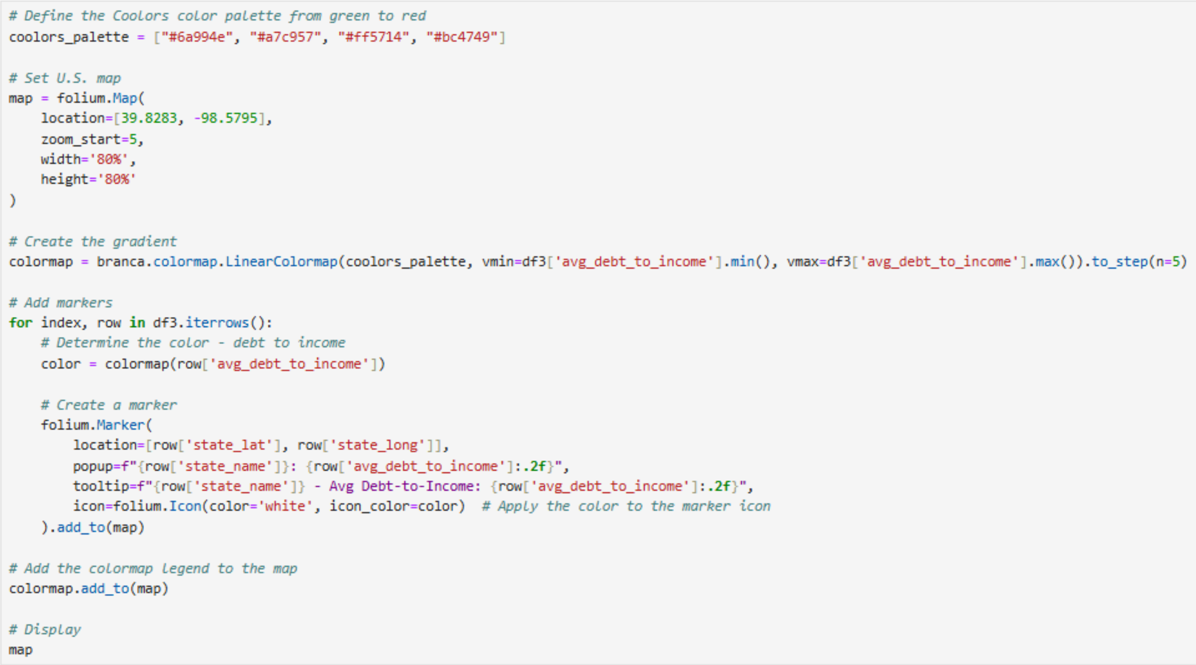


Debt by state



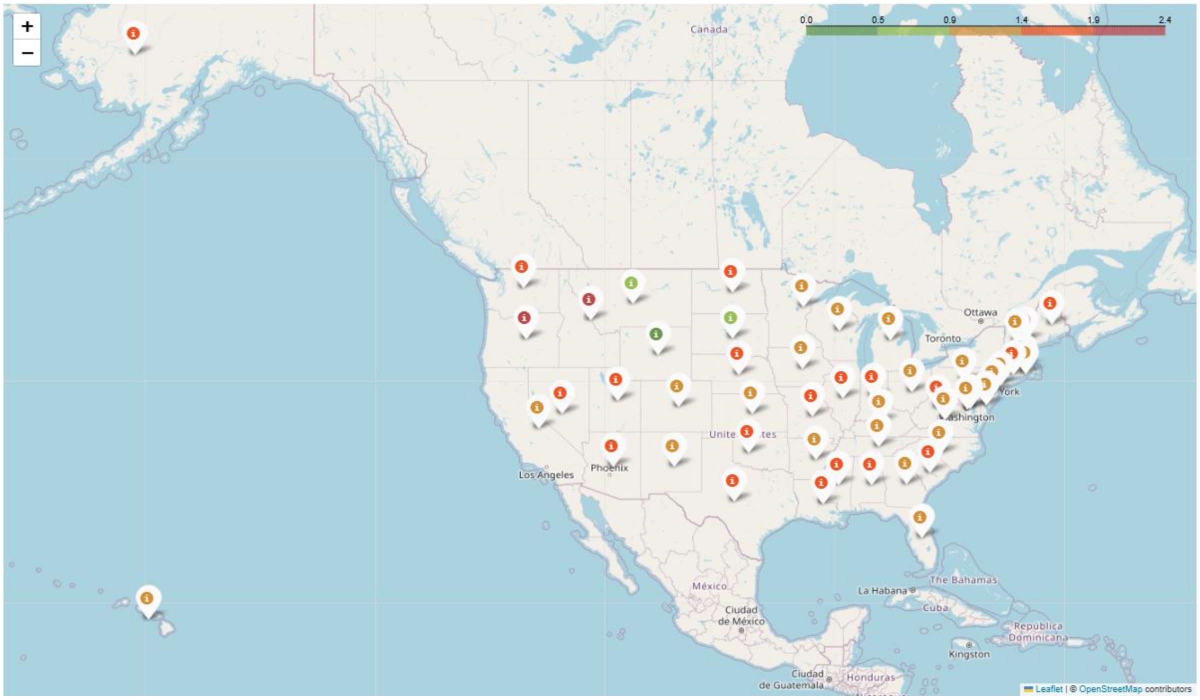
Debt to Income by state

To further test this refined hypothesis, the team developed Folium code to create a geographic plot of average DTI by state (Kaggle, 2024), which required the creation of a new data frame to merge centralized state latitude and longitude with the existing data frame and plot the average DTI and state information successfully.



```
# Define the Coolors color palette from green to red
coolors_palette = ["#6a994e", "#a7c957", "#ff5714", "#bc4749"]

# Set U.S. map
map = folium.Map(
    location=[39.8283, -98.5795],
    zoom_start=5,
    width='80%',
    height='80%'
)

# Create the gradient
colormap = branca.colormap.LinearColormap(coolors_palette, vmin=df3['avg_debt_to_income'].min(), vmax=df3['avg_debt_to_income'].max()).to_step(n=5)

# Add markers
for index, row in df3.iterrows():
    # Determine the color - debt to income
    color = colormap(row['avg_debt_to_income'])

    # Create a marker
    folium.Marker(
        location=[row['state_lat'], row['state_long']],
        popup=f"{row['state_name']}: {row['avg_debt_to_income']:.2f}",
        tooltip=f"{row['state_name']} - Avg Debt-to-Income: {row['avg_debt_to_income']:.2f}",
        icon=folium.Icon(color='white', icon_color=color)  # Apply the color to the marker icon
    ).add_to(map)

# Add the colormap legend to the map
colormap.add_to(map)

# Display
map
```

Folium geo mapping code



Gradient Average DTI map

Geographic plotting of the average DTI by state did not reveal any apparent regional influence on the DTI ratio and financial health. Overall, there is relatively even distribution of high and low DTI across the United States, with few exceptions observed. Included in those exceptions are a pocket of low DTI (strong financial health) in northern plains region containing Wyoming, Montana, and South Dakota, and the North-central region of the eastern seaboard containing Virgina, Maryland, Pennsylvania, New Jersey, and New York.

The northern Great Plains region contains a pocket of the lowest DTI states in the nation, which may suggest a regional influence on financial health initially; however, low sample sizes from these states may be responsible for this outcome. The North-central region of the eastern seaboard showed more potential for regional influence on the DTI ratio, with a larger sample size showing a greater concentration of average DTI by state when compared to other regions in the United States. Testing this observation requires further research using broader, real-world data, however.

Ultimately, the assessment of geographical influence on the average DTI ratio did not provide any clear indication of a relationship between location and financial health. Analysis of the data indicates that there is general trend toward populations leveraging debt as a proportion of their total income rather than leveraging debt more frequently in regions where per-capita income is lower. Variations in cost of living may help explain part of this relative consistency in average financial health distribution across the United States.

Bias and Limitations:

The above research, and subsequent regression analysis must acknowledge inherent dataset influences on the observations and conclusions. These influences include the following:

Limitations -

- Fictitious Data: The dataset used is entirely fictitious and was created solely for exercise purposes. As a result, the findings are not reflective of real-world statistics.
- Sample Size: Some states had very small sample sizes, which may not accurately represent broader trends. For instance, rural or low population states might show skewed averages due to a lack of data points.
- Simplified Metrics: The debt-to-income ratio used is a simplified measure. It does not account for important variables like cost of living, savings, or credit utilization, which could impact financial health significantly.

Bias -

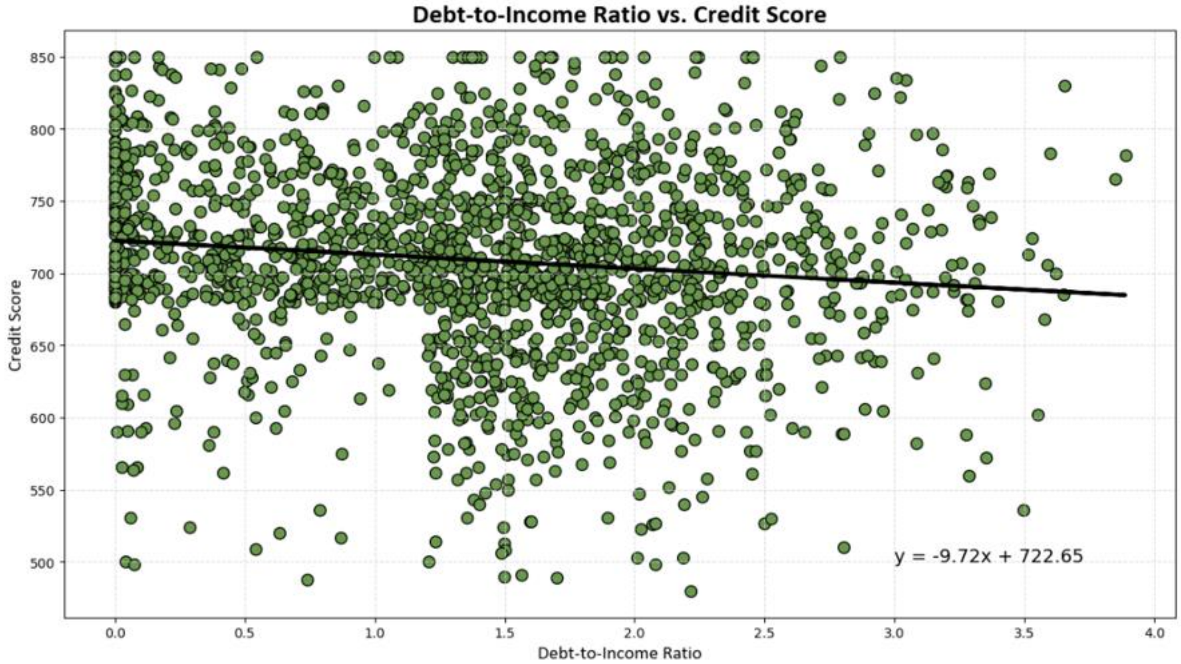
- Gender Distribution: The dataset's gender distribution is 49.2% male and 50.8% female. While close to an even split, it may not fully account for the nuanced financial behaviors influenced by gender.
- Outlier Adjustments: Outliers in total debt and yearly income were adjusted or removed to smooth the DTI ratio. While this helps reduce noise in the data, it may have unintentionally introduced bias by excluding valid but extreme data points.
- Geographic Representation: The inclusion of geographic data may overlook regional disparities such as varying costs of living, income distribution, and state level financial policies.

Additional Considerations -

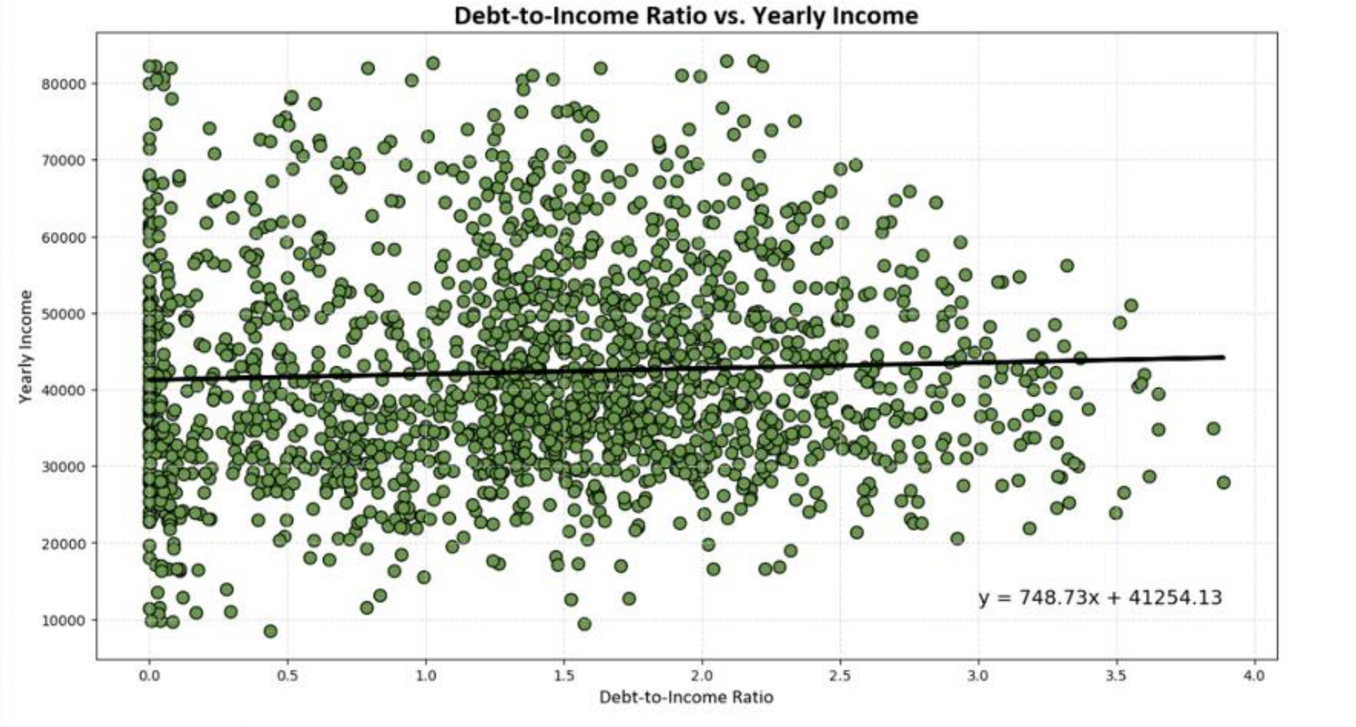
- The reliance on a single dataset limits the scope of the conclusions. Real world validation using diverse and larger datasets is essential.
- Differences in credit practices, such as regional banking policies or lender behaviors, were not accounted for and could influence debt and credit score relationships.

Regression Analysis:

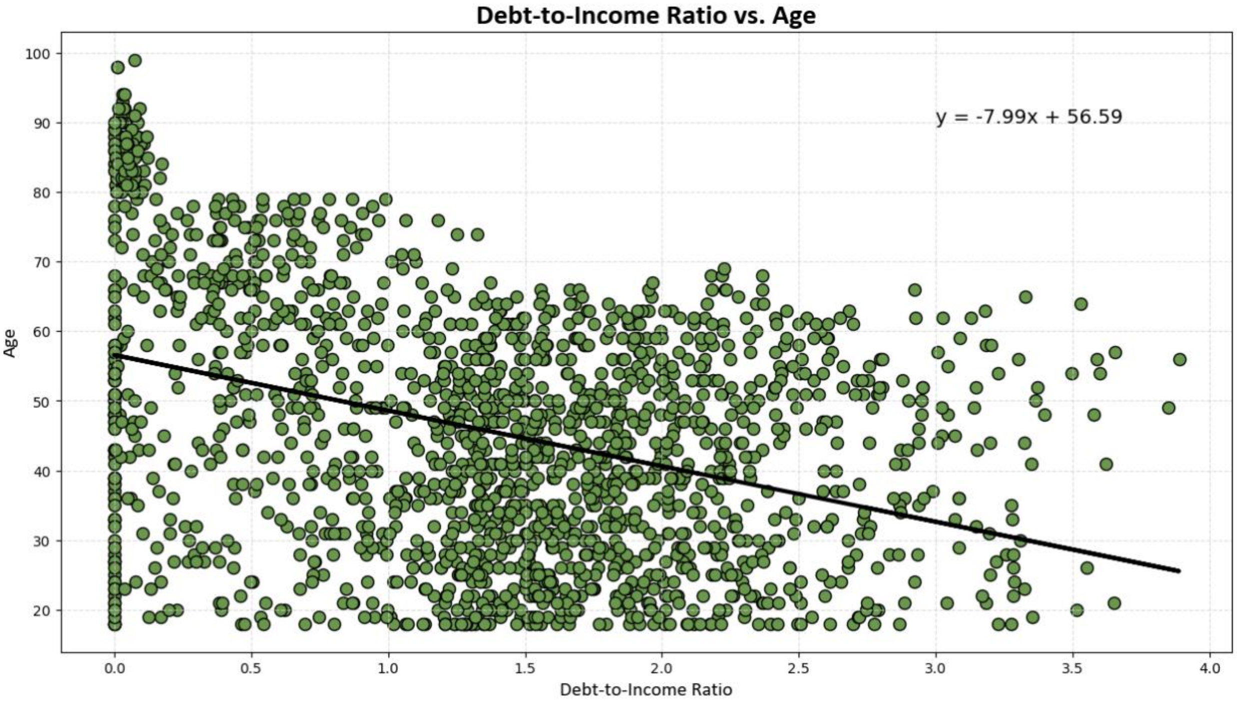
We conducted a regression analysis to examine the relationships between the debt-to-income ratio (DTI), demographic factors (such as age), and financial attributes (including total debt and yearly income). By plotting these relationships, we aimed to identify any significant trends or patterns that could provide insights into how these variables are connected. Below, we analyze each of these three graphs in detail.



The first graph shows no significant relationship between credit score and DTI. The regression line displays a very slight negative slope (y = -0.0x + 719.88), indicating that there is almost no measurable relationship between debt-to-income levels and credit scores in this population. This outcome was somewhat unexpected given the known role debt and income play in the calculation of credit scores. However, our analysis shows these variables may take a smaller role in the calculation of credit scores, with payment history, length of credit history, frequency of credit applications, and credit mix playing a larger role in the determination of this key financial indicator.



The second regression analysis between yearly Income and DTI. The regression line shows a slope that is nearly flat, with a slight incline across the x axis (y = 748.73x + 41254.13). Such a slope indicates no correlation between these variables. While the team initially hypothesized that a rise in income would reduce the debt-to-income ratio among the observed population, this did not occur. This indicates that debt loads remain relatively proportional to income levels, even as incomes rise. As individuals earn more money, their spending habits change, in kind.



In the last regression analysis, the team tested the previously observed potential relationship between debt-to-income and age. This analysis reflected the hypothesis generated by the previous plot for these variables, with the linear model showing a moderate negative relationship between age and DTI. The scatterplot shows a negative slope in the regression line (y = -7.99x + 56.59). suggesting that as individuals age, their DTI tends to decrease.

From this, the team determined that younger individuals may have higher DTIs due to lower incomes early in their careers, combined with higher debt, such as student loans and mortgages. Conversely, older individuals tend to have greater financial stability, higher incomes, and lower debts as they gradually pay off loans over time.

Conclusions:

For most variables analyzed, review of the DTI ratio across demographics, financial attributes, and geographic location did not reveal any significant relationships between these characteristics and overall financial health. The only exception to these results was age vs. DTI, which revealed a moderate negative relationship between the DTI ratio and age. This suggests that financial health, as measured by DTI, is influenced by the age of an individual, with financial health improving noticeably as people approach retirement age. This finding aligns with general observations that can be made in everyday life. In most cases, older people rely on debt less often as loans, mortgages, and other financial obligations are paid off with time, thus improving financial health.

The limitations and potential bias of this dataset must be factored into the conclusions, however. The stated hypotheses about the influence of demographic, financial, and geographic influence on financial health would benefit from further research using real-world data containing a larger and more diverse sample Americans.

Works Cited:

Financial Transactions Dataset: Analytics

https://www.kaggle.com/datasets/computingvictor/transactions-fraud-datasets/code

Kaggle: Introduction to Folium

https://www.kaggle.com/code/imdevskp/folium

ChatGPT – Python AI Assistance

https://chatgpt.com/