Project 1 – Group 12

Financial Analysis Project Overview

Caleb Meinke, Luisa Murillo,

Gavin Bozan, and Charles Arnold

Database Analytics Bootcamp – 12/06/24

**Introduction:**

For project 1, our project 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 project team leveraged the total debt-to-income ratio (DTI) for each consumer; calculated by dividing the individual total debt of each consumer by said consumer’s 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:**

Prior to conducting the analysis, the dataset required significant reorganization and cleaning. Initially, the dataset containing two CSV files that separated the records into 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 project team merged the CSVs on the consumer client ID columns contained in both databases.

A black and red text

Description automatically generated

Python code for merging CSVs

A screenshot of a computer

Description automatically generated

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 dataframe 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.

A screenshot of a computer

Description automatically generated

Cleaned Dataframe

Following the addition of the DTI ratio information to the dataframe, 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 dataframe 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.

A screenshot of a computer program

Description automatically generated

Reverse geocoding function

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

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

**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).

A graph of a number of income

Description automatically generated

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 dataframe 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.

A graph with numbers and dots

Description automatically generated

Debt by state

A graph of a number of dots

Description automatically generated with medium confidence

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 dataframe to merge centralized state latitude and longitude with the existing dataframe and plot the average DTI and state information successfully.

A computer screen shot of text

Description automatically generated

Folium geo mapping code

A map of the north america with a map of the north

Description automatically generated

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:**

**Regression Analysis:**

**Conclusions:**

For the majority of 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/>