

Title : Financial Literacy and Credit Card Debt Before and After COVID-19

Contributors: Wenqi Shan, Jun Kim

Institution: University of Illinois Urbana–Champaign

Version: Final Release 1.0

Repository URL: <https://github.com/oliviawudi/IS477-WJ/releases/tag/final-project>

Description: Analysis of SHED 2020 & 2024 microdata integrated with CFPB Consumer Credit Trends to evaluate post-COVID financial literacy and household credit card outcomes.

Link to box: <https://uofi.box.com/s/ietmow5znmh9ewrgu13qb431de4x3euce>

Summary

The COVID-19 pandemic has had a profound and long-lasting impact on the global economy; therefore, household credit card debt is also influenced. This project aims to investigate how financial literacy and household credit card debt have changed after COVID-19. Our team will be using two large-scale datasets from the years 2020 and 2024 Federal Reserve's Survey of Household Economics and Decisionmaking (SHED).

By comparing the 2020 and 2024 SHED datasets, this project will explore how levels of financial literacy, credit access, and repayment behavior have evolved in response to the pandemic's lasting economic effects. The analysis will examine demographic variables such as income, age, and education level to identify which groups experienced the most significant changes in credit card usage and debt management. We will also consider economic influences to interpret shifts in household financial behavior.

Our approach combines data integration, cleaning, and statistical analysis to ensure a transparent and reproducible workflow. The findings will provide insights into how households adapted their financial strategies in the years following COVID-19 and last year of 2024 after COVID-19.

The overall goal is to produce a reproducible, data-integrated analysis that would demonstrate how the pandemic and economic disruptions affect individuals' financial decision-making and credit outcomes. Through the process of merging these two large federal datasets, we will illustrate multiple aspects of the data lifecycle and other techniques that were learned in class.

To address feedback encouraging integration of our data from multiple sources, we expanded our project by adding an external dataset from the Consumer Financial Protection Bureau (CFPB). Unlike the Federal Reserve's SHED data, which mostly captures household-level experiences and perceptions, the CFPB dataset would provide macro-level indicators, including things like monthly trends in credit card originations, inquiries, and borrowing activity.

In addition to the technical aims, this project will contribute to a broader understanding of financial resilience: how knowledge, education, and access to certain financial resources influence households to make financial decisions and manage debt in a time of crisis. And by

highlighting the difference between 2020 and 2024, we hope to address the importance of promoting household stability in an unstable economic setting.

In order to answer our research questions, we analyzed SHED indicators such as emergency savings, response categories, and credit card categories. We paired these features from SHED with indicators from CFPB representing the total credit card volume of new credit card limits issued to different credit score groups. This would allow us to examine whether financial literate households would behave differently than financially stressed households, and whether there is a shift in behavior between 2020 and 2024.

Our findings suggest that financial literacy and financial stress indicators remained remarkably stable across the two SHED datasets. (2020 & 2024) Especially the emergency saving rate is close to 50% in both years. Good emergency responses and stressful responses changed only by a few percentage points, and the credit card usage patterns showed minimal variation. Therefore, pandemic didn't fundamentally change household resilient behaviors, even as the economic condition has a drastic downward trend.

Data Profile

1. Dataset Descriptions

- **Survey of Household Economics and Decisionmaking (SHED) 2020 & 2024 :**
We used individual level from the years 2020 and 2024 before and after COVID-19. SHED is an annual, nationally representative survey of U.S. adults that has been fielded since 2013. It is designed to measure the economic well-being of households and identify potential risks to their finances. The survey covers topics such as credit access and behavior, savings, retirement preparedness, economic fragility, housing, education, and student loans. SHED microdata are publicly available through the Federal Reserve Board and are anonymized on names, addresses, and direct identifiers are not included. The data are provided under terms that allow research and teaching use, with the expectation that results are reported in aggregate.

The 2024 survey was fielded in October 2024 and includes over 12,000 respondents and the 2020 has a similar order of magnitude. In our analysis, we harmonized the two datasets by aligning variable names and response encodings. Because the 2024 sample is larger than the 2020 sample, we drew a random subsample of 2024 respondents to match the 2020 sample size, so that year to year comparisons are not dominated by unequal actual size.

Feature:

credit card

- C2A — Credit card ownership

- C4A — Frequency of carrying unpaid balance

Financial Literacy

- EF1 — Has emergency savings
- EF2 — Ability to cover 3 months if income is lost

EF3 - Suppose that you have an emergency expense that costs \$400. Based on your current financial situation, How would you pay for this expense.

- EF3_a — Put it on my credit card and pay it off in full at the next statement
- EF3_b — Put it on my credit card and pay it off over time
- EF3_c — With the money currently in my checking/savings account or with cash
- EF3_d — Using money from a bank loan or line of credit
- EF3_e — By borrowing from a friend or family member
- EF3_f — Using a payday loan, deposit advance, or overdraft
- EF3_g — By selling something
- EF3_h — I wouldn't be able to pay for the expense right now

- **Consumer Credit Trends: Credit Cards (CFPB):**

To complement the reported SHED data, we used aggregate credit market data from consumer credit trends of credit cards series published by Consumer Financial Protection Bureau (CFPB). These data are derived from a nationally representative sample of credit records maintained by one of the three nationwide consumer reporting agencies (NCRAs). Before being provided to the CFPB, the records are stripped of identifying information such as names, addresses, and Social Security numbers.

The CFPB notes that these data are intended to monitor conditions in consumer credit markets, track lending activity over time, and identify emerging risks. Because the underlying sample is tied to a single NCRA, the CFPB cautions that shifts in market share across credit bureaus cannot be fully controlled for. The credit records are anonymized, and the published aggregates are made available as open data for research and policy analysis.

The panel spans from January 2007 through at least April 2025. We restrict the data to credit cards only and derive annual totals by credit score group by summing the monthly volumes for 2020 and 2024. We then take the total volume to obtain a more interpretable scale using **log_val** feature. These year by score

group aggregates are merged with SHED's credit score group categories, enabling us to relate household level financial behaviors.

Feature:

- month - Months observation count from 2007 Jan.
- date - Observation date
- vol - Clean up version of vol_unadj that remove seasonal effects.
- vol_unadj - The actual total dollar volume of new credit cards opened that month, which the actual money the bank gave out.
- credit_score_group - Divided each group by the credit score as:
 - Deep subprime : <580
 - Subprime : 580 ~ 619
 - Near-prime : 620 ~ 659
 - Prime : 660 ~ 719
 - Super-prime : 720+

-

2. Ethical & Legal Constraints

Copyright, Licensing, Terms of Use of CFPB makes Consumer Credit Trends available under an open public license, but use must follow the terms of use and attribution requirements.

- SHED microdata are federal public-use files, but FRB requires proper citation, prohibits implying endorsement, and disallows altering variable meaning in a misleading way.
- Therefore, we will cite both federal agencies explicitly in the report and codebook and preserve all original SHED variable labels and carefully documented transformations.

FAIR Assessment

- Findable: We provide GitHub repository with clear directory structure and metadata; Box links for outputs.
- Accessible: Public GitHub repository; Box folder with permission verified for UofI accounts.
- Interoperable: Open formats (CSV, Markdown, Python scripts), standard Python libraries, and a Snakemake workflow.
- Reusable: MIT-licensed code, explicit instructions for rerunning the entire pipeline.

3. Acquire Data

SHED 2020 and 2024 :

https://www.federalreserve.gov/consumerscommunities/shed_data.htm

- Source : Board of Governors of the Federal Reserve System
- Format : SAS file of **publicxxxx.sas7dbat**
- Dataset :
[https://www.federalreserve.gov/consumerscommunities/files/SHED_public_use_data_2020_\(SAS\).zip](https://www.federalreserve.gov/consumerscommunities/files/SHED_public_use_data_2020_(SAS).zip) (2020),
[https://www.federalreserve.gov/consumerscommunities/files/SHED_public_use_data_2024_\(SAS\).zip](https://www.federalreserve.gov/consumerscommunities/files/SHED_public_use_data_2024_(SAS).zip) (2024)
- Action : Download SAS file and documentation, converting the SAS file to CSV with analyzing its documentation.
- Documentation :
 - https://www.federalreserve.gov/consumerscommunities/files/SHED_2024codebook.pdf (2024)
 - https://www.federalreserve.gov/consumerscommunities/files/SHED_2020codebook.pdf (2020)

CFPB :

https://www.consumerfinance.gov/data-research/consumer-credit-trends/credit-cards/borrower-risk-profiles/#anchor_lending-levels

- Source : CFPB Consumer Credit Trends of borrower risk profiles (Lending Levels)
- Format : CSV
- Dataset :
https://files.consumerfinance.gov/data/consumer-credit-trends/credit-cards/volume_data_Score_Level_CRC.csv
- Action : Download and save as raw CSV for later aggregation and integration.

Data Quality

SHED 2024

1.1 Initial structure and feature selection

- Shape before subsetting: **12,295 rows × 753 columns**.

- Each row represents a survey respondent; each column is a survey question or derived variable.
- For this project, we extracted the variables most closely related to financial literacy and credit card behavior:
 - Credit card:
 - C2A – Credit card ownership (Yes/No)
 - C4A – Frequency of carrying an unpaid balance
 - Financial resilience:
 - EF1 – Has emergency savings (Yes/No)
 - EF2 – Ability to cover 3 months if income is lost (Yes/No)
 - Emergency expense handling:
 - EF3_a – Put it on credit card, pay in full next statement
 - EF3_b – Put it on credit card, pay over time
 - EF3_c – Pay with money in checking/savings/cash
 - EF3_d – Use bank loan or line of credit
 - EF3_e – Borrow from friend/family
 - EF3_f – Payday/advance/overdraft
 - EF3_g – Sell something
 - EF3_h – Would not be able to pay

1.2 Derived risk score and proxy credit group

We constructed a **risk_score** as an intermediate measure of credit risk / financial vulnerability:

- Start with risk_score = 0.
- Add 1 point for each “risky” or vulnerable condition:

- No emergency savings: EF1 == "No".
- Cannot cover 3 months of income: EF2 == "No".
- Risky emergency responses:
 - EF3_b == "Yes" (credit card, pay over time).
 - EF3_e == "Yes" (borrow from friends/family).
 - EF3_d == "Yes" (bank loan/line of credit).
- Limited credit access: C2A == "No".
- Frequent unpaid balances (from C4A):
 - Once → +1
 - Some of the time → +2
 - Most or all of the time → +3

We then mapped risk_score into a categorical **proxy_credit_group**:

- 0–1 → *Super-prime*
- 2–3 → *Prime*
- 4–5 → *Near-prime*
- 6–7 → *Subprime*
- ~8 → *Deep subprime*

The **proxy_credit_group** is what we use downstream. The numeric risk_score is treated as an intermediate feature used to derive the category and is not central to the final interpretation.

1.3 Missing value assessment and syntactic checks

- **Explicit missing values** (NaN) from df_shed24.isna().sum() were negligible for the selected EF and C variables.
- We then checked for **implicit missing values** (e.g., empty strings, invalid codes):

- **EF2:**
 - We discovered a large number of empty strings (""), not coded as NaN.
 - Over **50%** of responses in EF2 were empty.
 - Because EF2 is critical but heavily incomplete and would require dropping over half the sample to use it reliably, we **dropped EF2** from both SHED 2024 and SHED 2020 to avoid biasing the sample.
- **C4A:**
 - We found a set of records where C4A == "" (blank).
 - By cross-checking with C2A, we observed that these blanks correspond to respondents who **do not own a credit card**.
 - Instead of treating them as missing, we recoded:
 - C4A == "" → "No credit card ownership".
 - Final allowed C4A categories:
 - "Never carried an unpaid balance (always pay in full)"
 - "Once"
 - "Some of the time"
 - "Most or all of the time"
 - "No credit card ownership"

1.4 Semantic consistency checks

We performed **semantic checks** to ensure that responses across related variables were logically consistent:

- EF2 vs EF3_ (ability to cover vs how they would pay)*
 - Logic: If a respondent says they can cover 3 months of expenses (EF2 == "Yes"), it is unlikely that **all** EF3 options are "No" (i.e., they say they have no way to pay a \$400 emergency expense).
 - We identified rows where:

- EF2 == "Yes" and
 - all EF3_a–EF3_h are "No".
- This yielded **three inconsistent rows** (e.g., likely data entry or response errors).
- We **dropped these 3 rows** from SHED 2024, as they appear to be semantically impossible.
- **C2A vs C4A (ownership vs behavior)**
 - Logic: A respondent without a credit card (C2A == "No") should not report any of the unpaid balance behaviors (e.g., "Once", "Some of the time").
 - We checked for rows with:
 - C2A == "No" and C4A in unpaid-balance categories.
 - No such rows were found after the recording step, so **no further action** was needed.

2. SHED 2020 (public_use_data_2020)

2.1 Initial structure and feature selection

- Raw file: public2020.sas7bdat (converted to shed2020.csv).
- Shape before subsetting: **11,648 rows × 372 columns**.
- Selected variables (2020 uses slightly different coding and capitalization):
 - C2A – Credit card ownership (0/1)
 - C4A – Unpaid balance frequency (0–3, -1, NaN)
 - EF1 – Has emergency savings (0/1)
 - EF2 – Ability to cover 3 months if income is lost (0/1)
 - EF3_A–EF3_H – Emergency expense handling (0/1)

We use a parallel logic to SHED 2024, but 2020 is coded numerically.

2.2 Derived risk score and proxy credit group

We computed a numeric **risk_score** using the 0/1 indicators:

- Start with risk_score = 0.
- Add 1 point for:
 - EF1 == 0 (no emergency savings)
 - EF2 == 0 (cannot cover 3 months)
 - EF3_B == 1, EF3_E == 1, EF3_D == 1 (riskier emergency responses)
 - C2A == 0 (no credit card access)
 - C4A == 1 to +1
C4A == 2 to +2
C4A == 3 to +3

We then mapped risk_score to the same **proxy_credit_group** categories as in 2024 (Super-prime, Prime, Near-prime, Subprime, Deep subprime). As with 2024, risk_score is primarily used as a working variable to derive proxy_credit_group.

2.3 Missing values and refusal codes

- C4A initially contained:
 - Codes: 0, 1, 2, 3, -1, NaN.
 - NaN values accounted for ~**13.8%** of responses.
 - -1 indicated “refused to answer”.

To handle this:

1. We filled true missing values in C4A with a sentinel value:
 - NaN to -2 (distinct from valid responses and refusal).
2. We removed all rows where any of the EF/C variables had **-1** (refusal) using:
 - `df_shed20 = df_shed20[~df_shed20.eq(-1).any(axis=1)]`
3. After this, C4A was cast to an integer type (Int64), and we defined valid values as:

- { -2, 0, 1, 2, 3 }
- No invalid codes remained.

2.4 Type checks and syntactic consistency

Initially, many binary features and C4A were stored as floats (e.g., 1.0, 0.0). This caused syntactic checks using string or integer comparisons to fail.

- **Binary variables:** C2A, EF1, EF3_A–EF3_H:
 - We first detected “syntactic errors” because values like 1. and 0. did not match the string set {"1.0", "0.0"}.
 - To enforce consistent numeric coding, we converted these columns to integers:
 - `df_shed20[binary_cols] = df_shed20[binary_cols].astype(float).astype("Int64")`
 - We then defined valid binary values as {0, 1} and verified that no invalid values remained.
- **C4A:**
 - We cast C4A to Int64 and accepted { -2, 0, 1, 2, 3 } as the valid set.
 - This aligned the syntactic representation with the semantic meaning (never, once, some, most/all, special codes).

As with SHED 2024, **EF2** was dropped from the final cleaned dataset due to consistency considerations and to keep both years directly comparable on the remaining variables.

2.5 Row-level quality impact

- After removing rows with any -1 (refusals) and enforcing data-type corrections, we dropped **137 rows** in SHED 2020.
- The cleaned dataset was saved as **shed20_cleaned2.csv**.

3. CFPB Consumer Credit Trends (credit cards – borrower risk profiles)

3.1 Structure and variable meanings

- Raw file: volume_data_Score_Level_CRC.csv (later saved as cfpd_cleaned2.csv).
- Shape: **1,100 rows × 5 columns**.
- Each row represents a **credit score group in a given month**.
- Variables:
 - month – Index of month since baseline.
 - date – Year–month string (e.g., "2025-01").
 - vol – Seasonally adjusted total dollar volume of new credit limits.
 - vol_unadj – Unadjusted (actual) total volume of new credit card limits.
 - credit_score_group – One of:
 - Deep subprime (<580)
 - Subprime (580–619)
 - Near-prime (620–659)
 - Prime (660–719)
 - Super-prime (720+)

This dataset provides the macro-level credit environment broken down by risk category.

3.2 Missing values, type checks, and semantic validation

- df_crc.isna().sum() showed **no missing values** in any column.
- dtypes confirmed that:
 - month is integer-like,
 - vol and vol_unadj are numeric (float),
 - date and credit_score_group are strings.
- We verified semantic and syntactic validity:
 - **Dates:**

- `df_crc["date"].value_counts()` showed properly formatted YYYY-MM strings, with no unexpected codes.
- **Credit score groups:**
 - `df_crc["credit_score_group"].value_counts()` contained only the five expected categories.
- **Volumes:**
 - Checked for negative values:
 - `(df_crc[["vol", "vol_unadj"]] < 0).any() → all False.`
 - Spot-checked rows for specific dates (e.g., "2025-01") and verified a full set of credit score groups with plausible volumes.

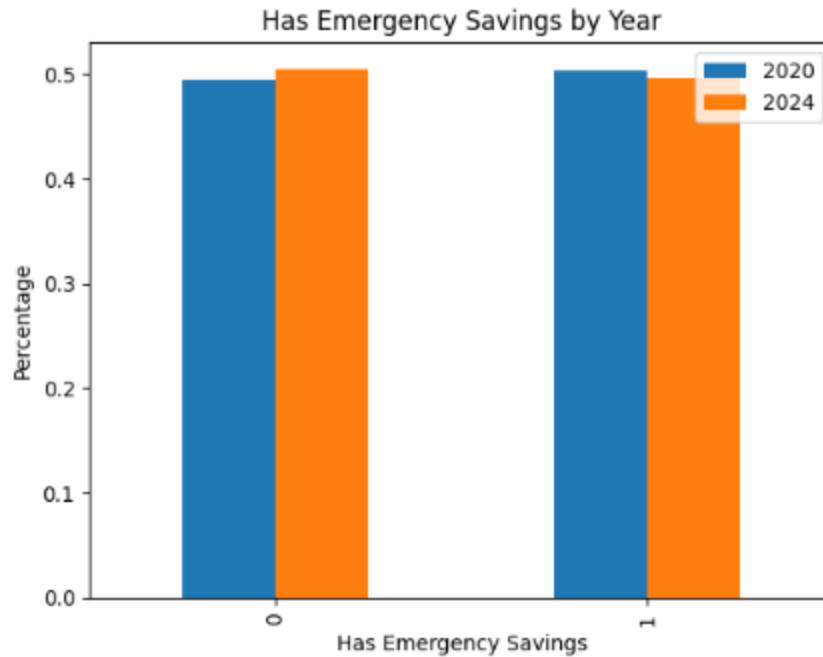
Because the CFPB dataset is already well-structured and uses clean categorical labels, no major recoding was needed at this stage. We saved this cleaned version as **cfpd_cleaned2.csv** for downstream aggregation.

Data Findings

Visualization & Analysis

Financial Literacy

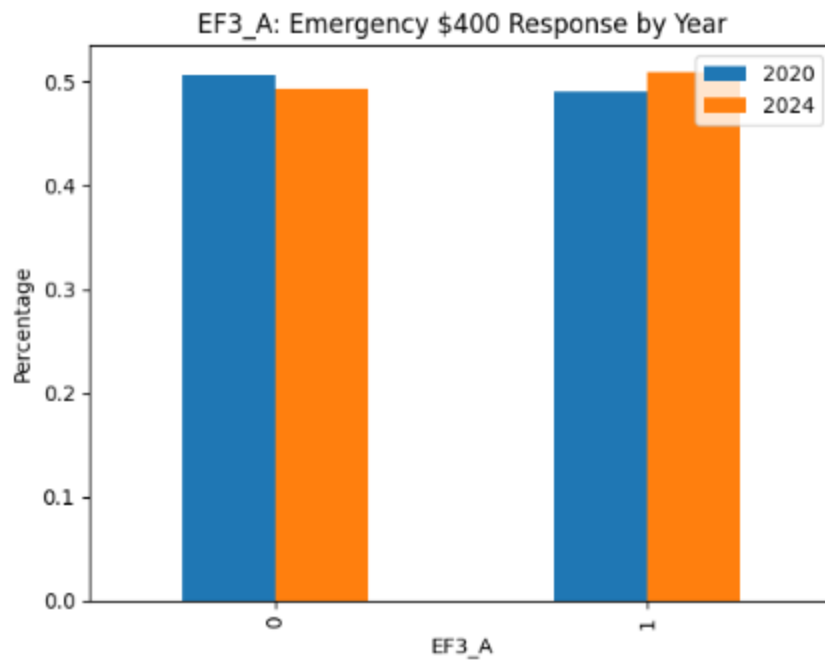
1. Emergency Savings (EF1)



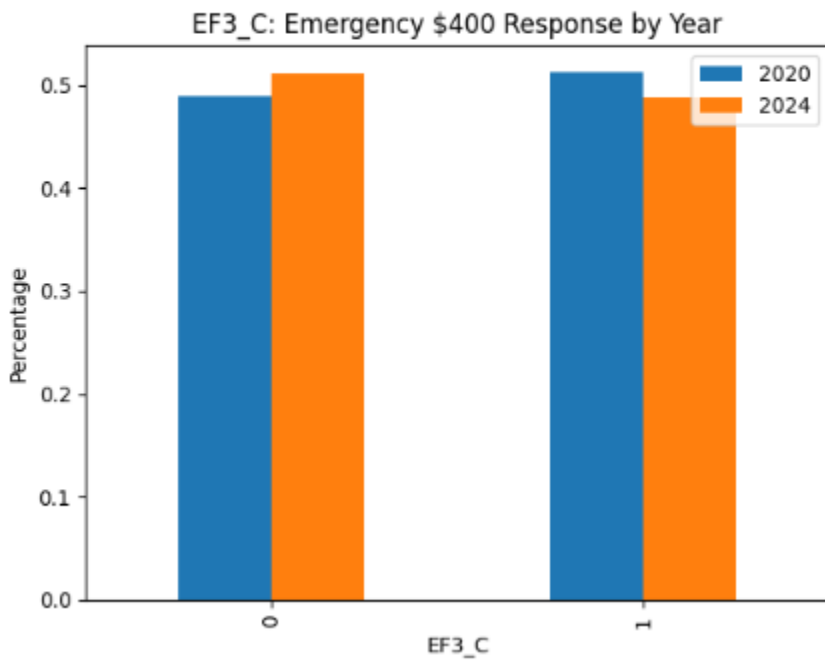
There is no big difference in emergency savings behavior between 2020 and 2024. Although SHED represents millions of United States adults, the observed change in EF1 between 2020 and 2024 is less than one percentage point and likely reflects sampling noise rather than a real change in national emergency savings rates.

2. Emergency Savings with financially stable (EF3_A, EF3_C)

year	2020	2024
EF3_A		
0	0.506092	0.493908
1	0.490628	0.509372



year	2020	2024
EF3_C		
0	0.489696	0.510304
1	0.512709	0.487291



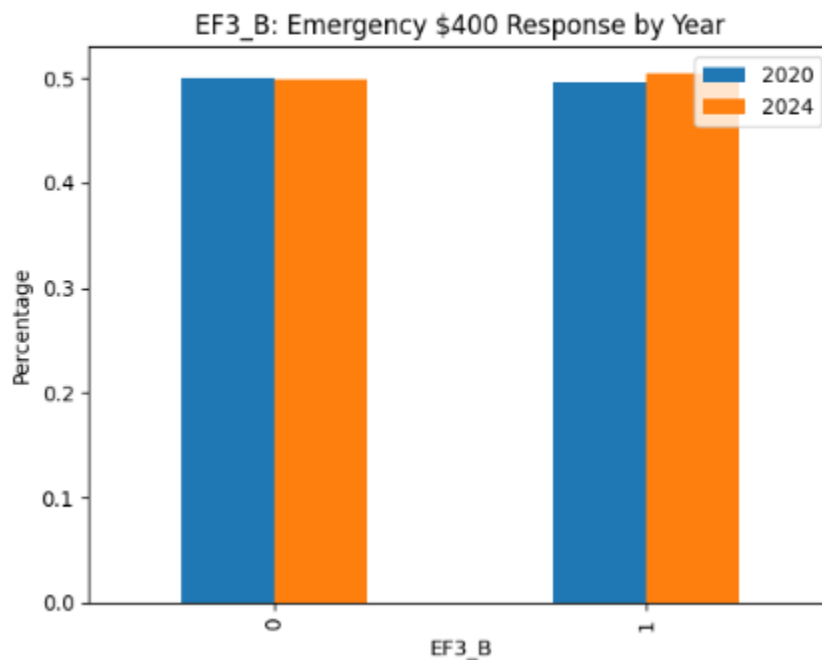
By looking at this chart, people who put it on their credit card and pay it off in full at the next statement have increased about 3% over 4 years, and people who are not decreased about 2%.

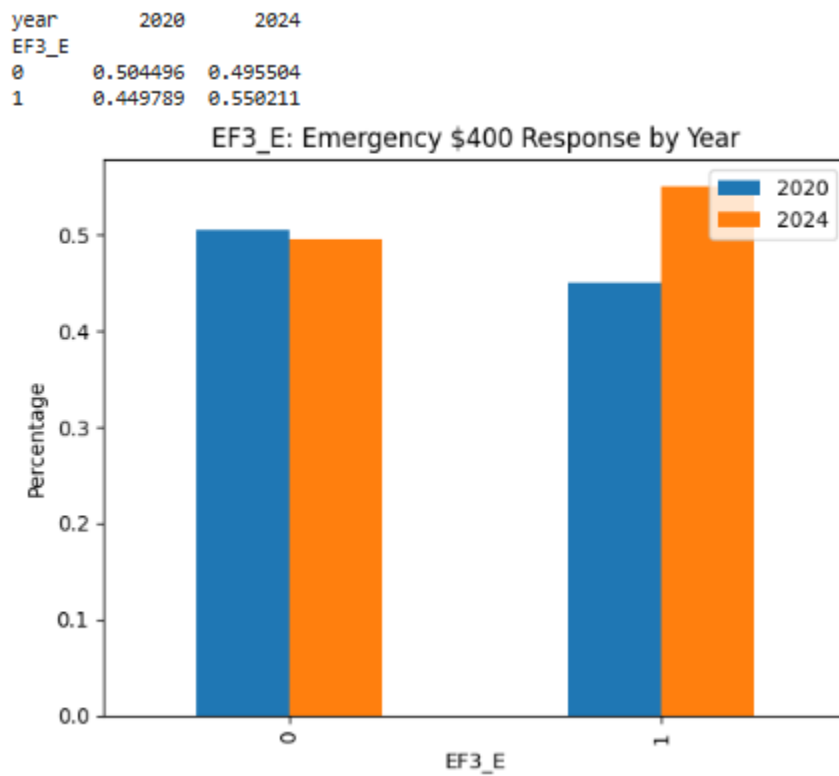
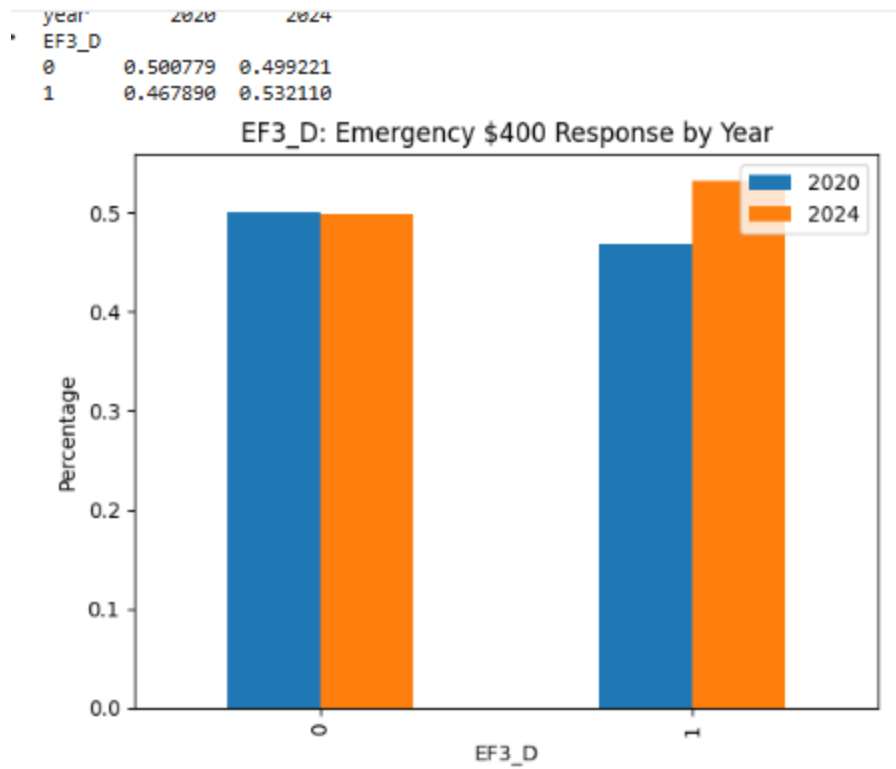
The people with the money currently in my checking/savings account or with cash have decreased about 1.5% and people who are not paying with the EF3_C method increased about 1.3% over 4 years.

People who are financially strong with no stress have similar status before and after COVID-19.

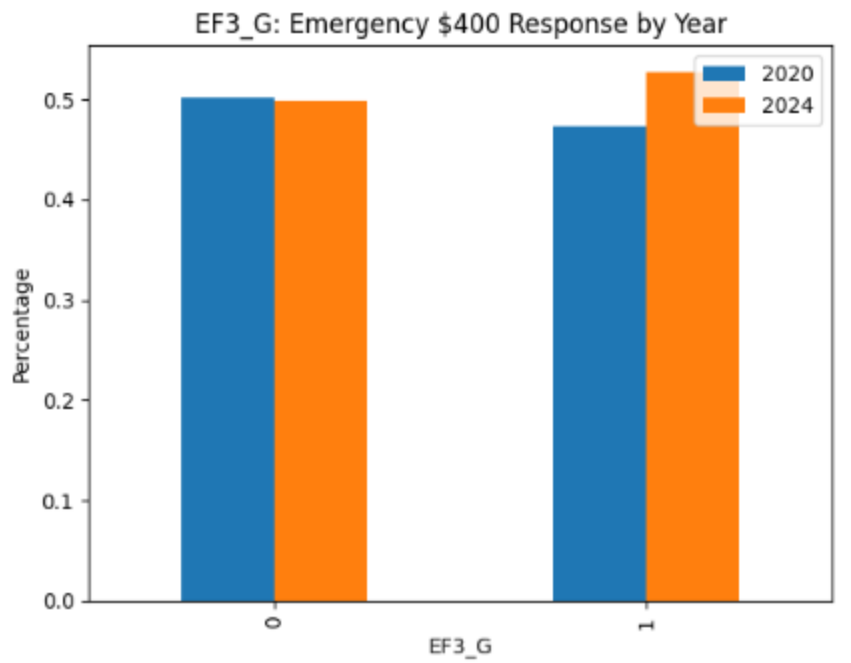
3. Emergency Savings with financially unstable (EF3_B, EF3_D, EF3_E, EF3_G / EF3_F, EF3_H)

year	2020	2024
EF3_B		
0	0.500789	0.499211
1	0.495402	0.504598



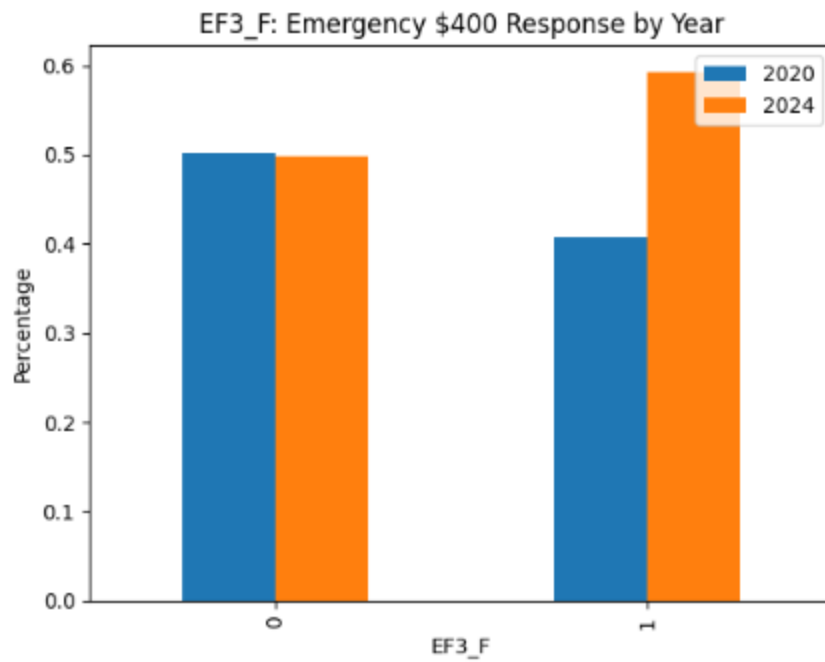


year	2020	2024
EF3_G		
0	0.501586	0.498414
1	0.472684	0.527316

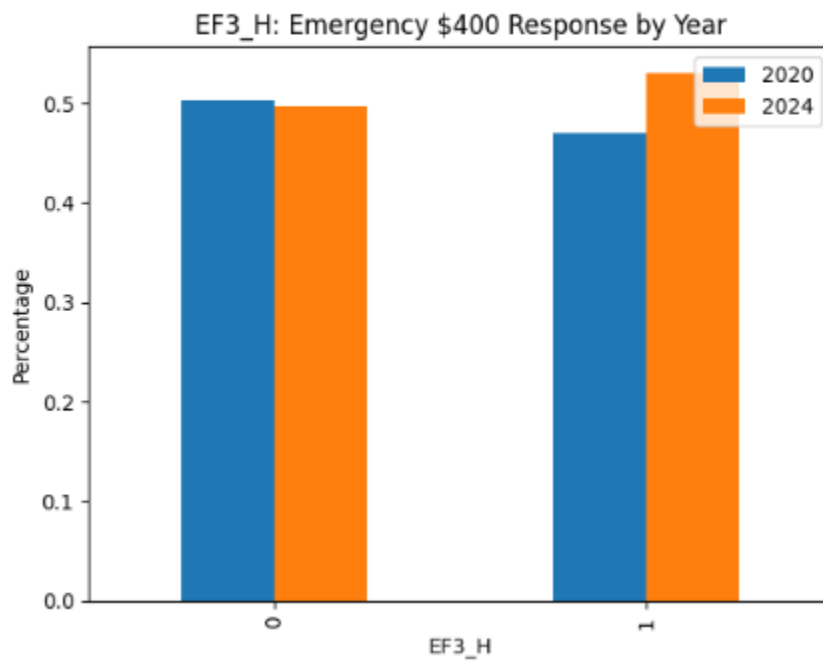


EF3 of B, D, E, and G are the people who are financially stressed. Overall, number of people have been similar or increased who are financially stressed, especially people who are borrowing from their friends or family have been increased before and after COVID-19 finished.

year	2020	2024
EF3_F		
0	0.501344	0.498656
1	0.407295	0.592705



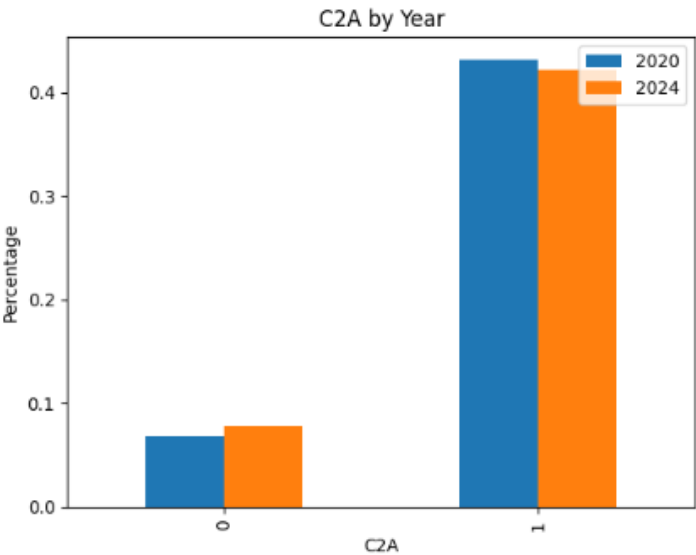
year	2020	2024
EF3_H		
0	0.503677	0.496323
1	0.469642	0.530358



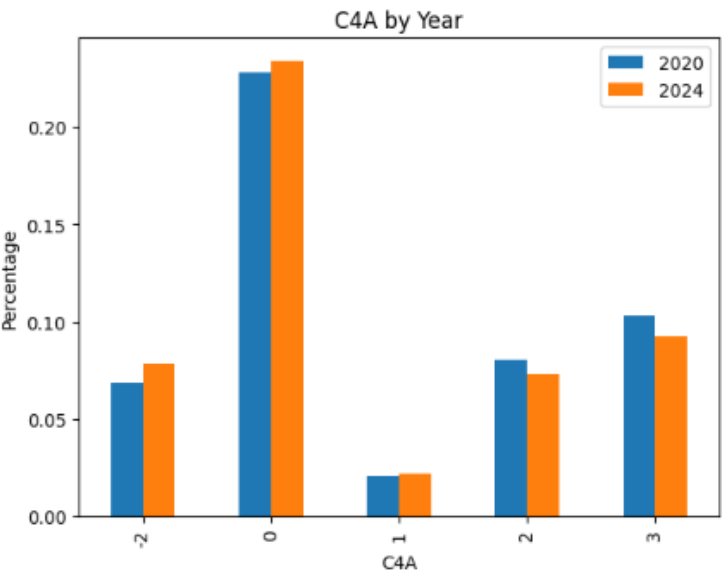
EF3 of F and G are the people who are facing the high risk of dealing with financial literacy. The number of people has increased before and after COVID-19, especially the people who are in payday loan or overdraft.

4.Credit Card

year	2020	2024
C2A		
0	0.068369	0.078664
1	0.431631	0.421336



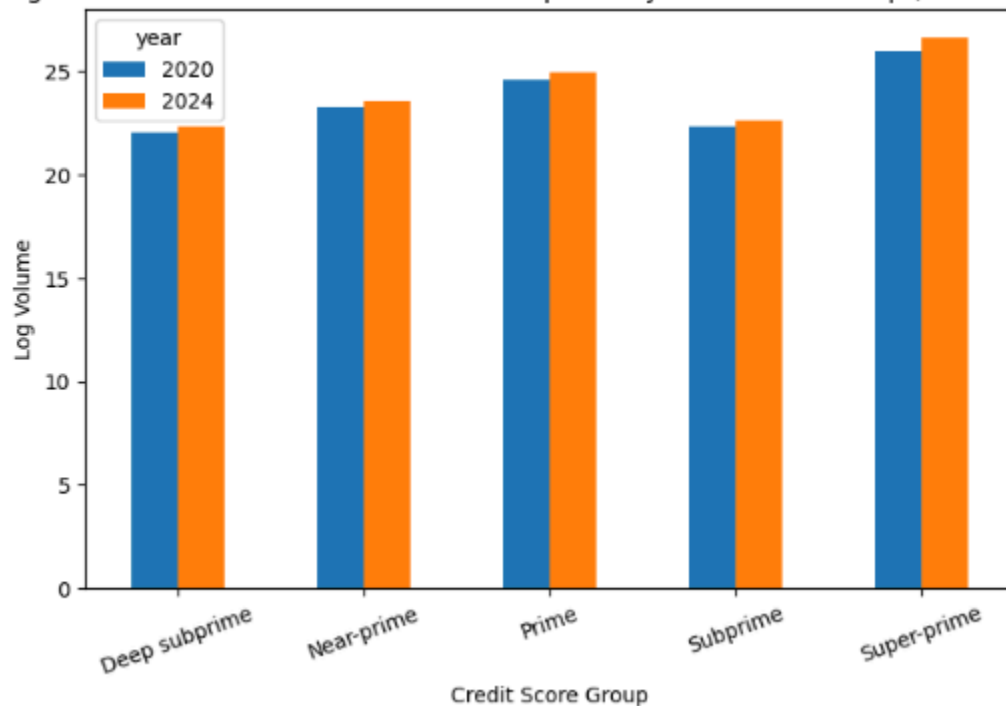
year	2020	2024
C4A		
-2	0.068369	0.078664
0	0.227695	0.233907
1	0.020806	0.021979
2	0.080228	0.073017
3	0.102902	0.092433



Credit card ownership decreased slightly between 2020 and 2024 (about one percentage point), indicating a small rise in households without access to revolving credit, though the change is small.

Credit card ownership (C2A) declined modestly from 2020 to 2024, with the share of non-credit card holders increasing by about one percentage point. The distribution of unpaid balance behavior **C4A** remained largely stable across years. Small improvements are evident slightly more respondents report paying their balance in full, and slightly fewer report carrying a balance frequently—but these shifts are within a narrow range. Overall, both credit access and repayment behavior exhibit minimal change between the two SHED waves.

Log of Total Dollar Volume of Credit Cards Opened by Credit Score Group (2020 vs 2024)

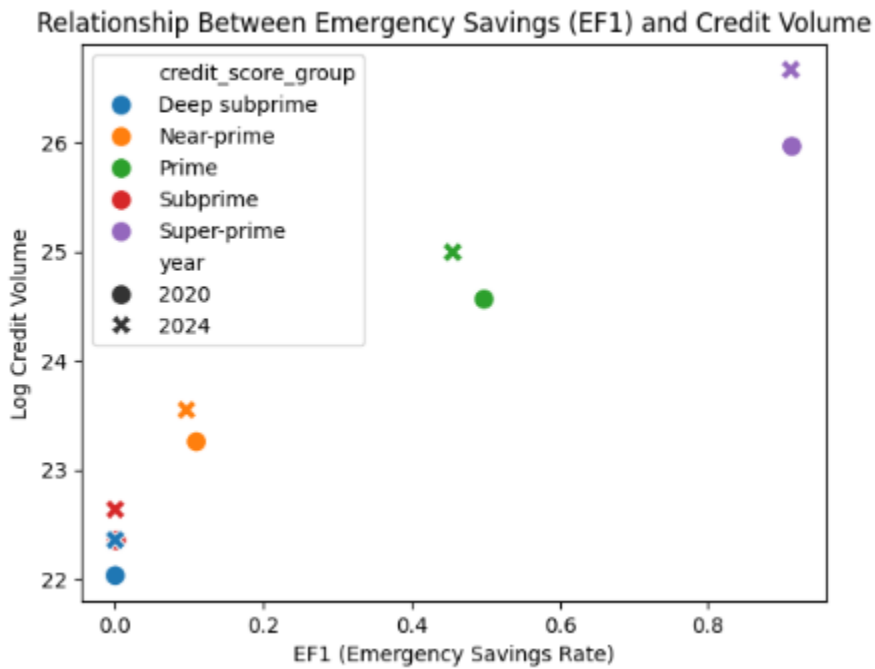


We can identify that lenders issued more credit in 2024 than in 2020, regardless of borrower credit quality. We can see the patterns of how the credit balances are distributed as it explained:

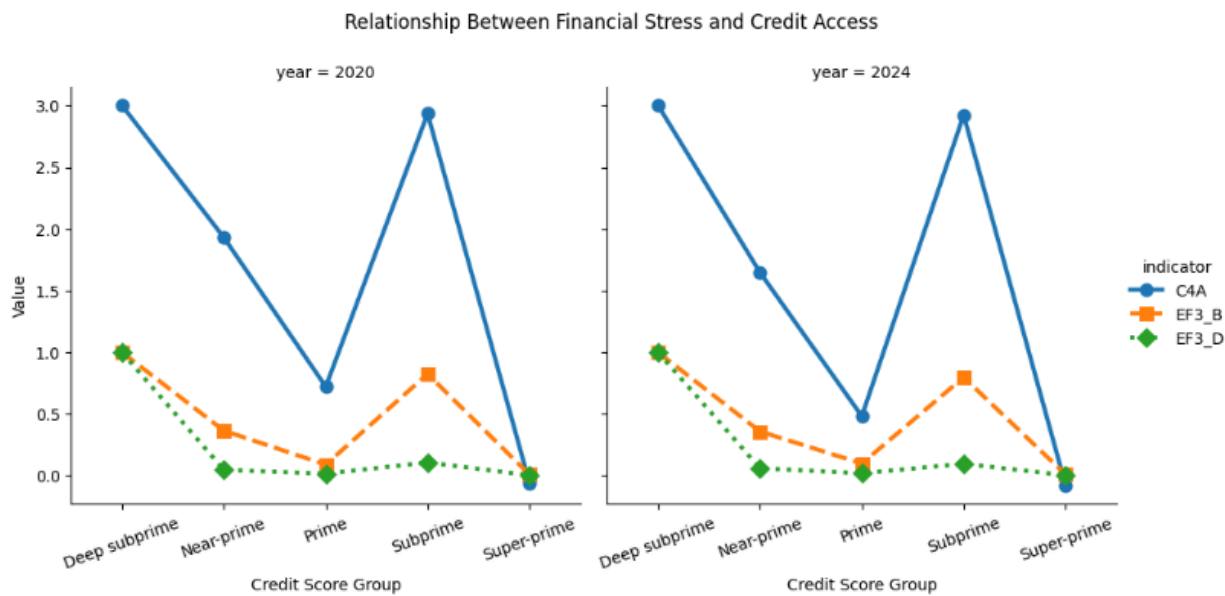
- Deep subprime : <580
- Subprime : 580 ~ 619
- Near-prime : 620 ~ 659
- Prime : 660 ~ 719
- Super-prime : 720+

The group that is closer to Super-prime has more volume of credit card usage.

5. Merged SHED & CFPB



The credit access is strongly aligned with financial stability that credit score groups with higher emergency savings rates of EF1 consistently receive much higher credit volumes in both 2020 and 2024. However, Deep Subprime and Subprime groups where bottom-left shows the weakest financial stability with receiving the least new credit. Overall, all groups in 2024 have high credit volume then 2020 showing that financial stability has been better after COVID-19.



Carrying unpaid balance C4A was the highest for Deep Subprime and Subprime in both years. The Super-prime had near zero indicating very strong financial habits. As the EF3_B of using credit card to pay emergencies over time again showing highest in Subprime and Deep Subprime. This indicates that lower credit score groups depend more on revolving credit during emergencies than higher credit score groups. EF3_D of using loans / borrowing for emergencies shows the highest for Deep Subprime, lowest on Super-prime. This indicates financial stress behaviors cluster heavily in lower credit tiers.

Deep Subprime and Subprime households rely more on borrowing and revolving credit and carry unpaid balances at much higher rates. Prime and Super-prime households, in contrast, demonstrate stronger financial resilience and better credit management. This pattern did not change from 2020 to 2024, despite the economic recovery.

Research Question

1. Did the strength of the relationship between financial literacy and household credit card change after COVID-19?

From SHED (2020 vs 2024),

- EF1 (has emergency savings) is almost identical in 2020 and 2024 ($\approx 50/50$).
- “Good” emergency responses (EF3_A, EF3_C) and “bad/stressful” responses (EF3_B, EF3_D/E/F/G/H) changed only by $\sim 1-3$ percentage points.
- Credit card ownership (C2A) and unpaid balance behavior (C4A) are also very stable, with only small shifts ($\sim 1\%$).

So financial literacy / resilience indicators and credit-card behaviors are basically stable across years in the SHED data.

From **merged SHED and CFPB**,

- In both 2020 and 2024, groups with higher EF1 and more “good” EF3 behavior (Prime & Super-prime) have much higher credit volumes.
- Deep Subprime & Subprime have:
 - lowest EF1 and “good” EF3_A/C
 - highest stress indicators (C4A, EF3_B, EF3_D)
 - lowest credit volume.

Overall, there is no evidence that the strength of the relationship between financial literacy and household credit card debt changed after COVID-19. In both 2020 and 2024, credit-score groups with higher financial resilience (higher emergency savings rates and more cash-based emergency responses) consistently received much larger volumes of new credit card limits, while financially stressed groups (Deep Subprime and Subprime) had both weaker financial literacy indicators and lower credit volumes. Although total credit volume increased for all groups between 2020 and 2024, the association between financial literacy and credit access remained strong and qualitatively similar across both years.

2. Did households with higher financial literacy show smaller increases in debt post-COVID-19?

From **CPPB vol** as proxy for new credit card debt,

- The bar chart of log_vol by credit_score_group (2020 vs 2024) shows:
 - All credit-score groups saw higher credit volume in 2024 than in 2020.
 - The largest increases are in Prime and Super-prime groups.
 - Deep Subprime and Subprime also increased, but by a smaller amount.

From **EF1 vs log_vol**,

- Points for 2024 are above the 2020 points for every credit_score_group → more credit in 2024.
- The positive relationship between EF1 and log_vol is clear: higher EF1 → higher credit volume, and this holds in both years.

Overall, not that households with higher financial literacy did not show smaller increases in credit-card debt; instead, they experienced larger increases in credit access. Between 2020 and 2024, total credit volumes increased across all credit-score groups, but the largest growth occurred among Prime and Super-prime borrowers—those with the strongest emergency savings and lowest financial stress indicators. Deep Subprime and Subprime groups, which show weaker financial literacy and higher financial stress, also saw some growth in new credit card limits, but their increases were smaller in magnitude. This suggests that post-COVID credit expansion primarily benefited financially stronger households rather than reducing debt growth among them.

Future Work

While this project provides an integrated view of household financial literacy, credit card behavior, and national credit-market conditions before and after COVID-19, the findings also highlight several important directions for future work. Our current analysis shows that financial literacy indicators (such as emergency savings and emergency-response behavior) remained relatively stable between 2020 and 2024, and the relationship between financial resilience and credit access appears consistent across years. However, these results should be interpreted cautiously due to several data and methodological limitations that future research could address.

The first limitation would be that SHED doesn't really offer local or regional data regarding credit outcomes, which creates some barriers for us to examine geographical disparities. A major direction that we may take next step would be having some county-level or zip-code-level indicators. And merging datasets that would provide county level insights such as unemployment rate, local price indices, and regional price constraints would be salient to get a better understanding of how place-based economic conditions are shaping their roles in long term debt accumulation as well.

Second, our current CFPB only captures credit volume and credit score groups of credit cards. Therefore, consider adding additional CFPB series such as utilization ratios, repayment patterns, and delinquency rates for future work so that we can have a more structured understanding of debt growth. Because these new indicators will enable us to have a full cycle-analysis of household debt. Credit access -> Credit Use -> Repayment -> Distress. Integrating multiple CFPB series would also enable classification of borrowers into dynamic risk categories and allow deeper investigation of how credit behavior evolves across economic cycles.

Third, the SHED survey structure limits how financial literacy can be operationalized beyond high-level proxies such as EF1 and EF3. More nuanced measures, such as numeracy questions, risk assessments, or multi-item literacy scales would enable richer modeling of how

financial knowledge influences credit outcomes. Future surveys or experimental designs could help isolate causal pathways between literacy, behavior, and borrowing.

In addition, our analysis right now used descriptive comparisons and simple aggregated visualizations. Future work could apply statistical modeling, such as logistic regression, hierarchical models, or matching techniques, to quantify differences in credit outcomes across demographic groups while controlling for income, education, or employment shifts.

If there is a chance, making this project a longitudinal study would be important since it determines whether financial resilience has other behaviors over time (does it fluctuate or not?) rather than discretely between 2020 and 2024. With some additional waves of data from SHED, we could possibly model trends across the full post-pandemic recovery period plus the situation before the pandemic.

Incorporating qualitative and mixed methods approaches would add more rigor to our project. Analyzing some survey comment fields, open-ended responses or interviews. These would conceptualize our findings with actual lived experience and offer insights into household decision-making.

Therefore, we would love to expand our data source to add more features and years, analytical methods, and granularity of financial literacy measures would be able to allow us for a deeper understanding of how households navigate credit markets in periods of economic disruption.

Reproducing

To ensure full reproducibility and transparency, we have created a complete reproducible research package that allows anyone to re-run our entire workflow

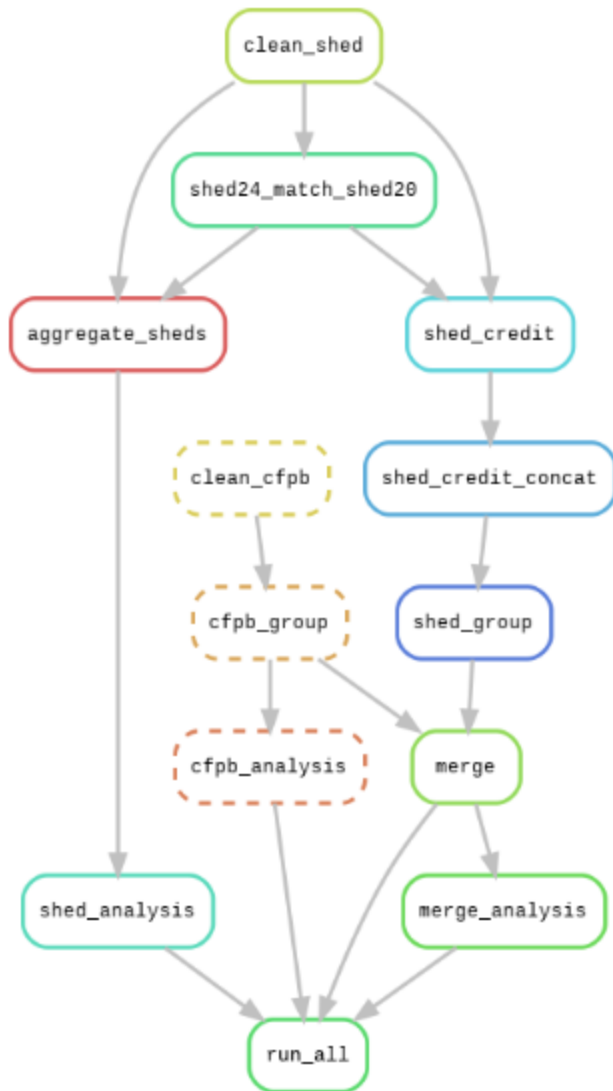
1. Documentation describing how to reproduce the analysis We have posted a dedicated **reproduce.md** that outlines each step required to reproduce the workflow. This includes instructions for:
 - Setting up the Python environment
 - Downloading SHED data from the Federal Reserve website
 - Running the Snakemake workflow
 - Locating and interpreting the output files
2. Box Storage Link: <https://uofi.box.com/s/ietmow5zmf9ewrgu13qb431de4x3euce>
3. Following the course requirement, there are :
 - All output files and original csv from CFPB
 - Verified permissions 12+ hours before the deadline.
 - Addition the Box data directory to **.gitignore**.

This ensures reproducibility while respecting licensing rules.

3. All code, workflow scripts, and automation tools
4. Actual results including output data, visualizations, and tables
 - These are included in the jupyter notebook we provided.
 - As well as the box folder
5. Specification of software dependencies
 - A requirements.txt listing all libraries needed to run the workflow.
 - A pip freeze snapshot for exact reproducibility of the environment.
 - The environment documentation includes:
 - a) Python version
 - b) Operating system details
 - c) Any optional tools (Jupyter, Snakemake version)

Workflow & Provenance

This project was implemented using a fully reproducible Snakemake workflow that documents every step in the data lifecycle. From acquisition and cleaning to integration, analysis, and final output generation. The workflow ensures transparency, traceability, and automation by encoding all data transformations in rule-based procedures. The figure above illustrates the Directed Acyclic Graph (DAG) of the workflow, showing how raw inputs flow through intermediate processing steps to produce the merged analytical dataset from SHED and CFPB and final visualizations.



clean_shed

Takes the raw SHED SAS files (shed2024.sas7bdat, public2020.sas7bdat) and runs scripts/clean_shed.py to produce cleaned CSVs shed24_cleaned2.csv and shed20_cleaned2.csv. This step implements the missing-value handling, type normalization, and recoding decisions described in the data cleaning section.

clean_cfpb

Takes the raw CFPB CSV (volume_data_Score_Level_CRC (4).csv) and runs scripts/clean_cfpb.py to produce cfpd_cleaned2.csv.

shed24_match_shed20

Runs scripts/shed24_match_shed20.py on shed24_cleaned2.csv and creates df_shed24ver1.csv,

which subsamples 2024 respondents to match the 2020 sample size.

aggregate_sheds

Takes df_shed24ver1.csv and shed20_cleaned2.csv and concatenates them into a combined SHED dataset df_shed.csv using scripts/aggregate_sheds.py.

shed_credit and shed_credit_concat

Filter both years to respondents who own credit cards and then concatenate them into shed2.csv. These steps implement the logic of removing non-card-holders when the analysis focuses on credit card debt behavior.

shed_group

Groups shed2.csv by proxy credit risk group and year using scripts/shed_group.py, producing shed_year.csv for integration.

cfpb_group

Aggregates cfpd_cleaned2.csv by CFPB credit score group and year via scripts/cfpb_group.py, producing crc_year.csv.

cfpb_analysis

Runs scripts/cfpb_analysis.py on crc_year.csv to create analysis_cfpb.PNG, summarizing changes in lending volume over time by credit score group.

shed_analysis

Runs scripts/shed_analysis.py on df_shed.csv and generates multiple SHED-only diagnostic plots of analysis PNG used for data profiling and exploratory checks.

merge

Takes the grouped datasets shed_year.csv and crc_year.csv and merges them into the final integrated dataset merged.csv using scripts/merge.py. This is where the individual-level SHED behavior is aligned with the macro-level CFPB credit environment.

merge_analysis

Runs scripts/merge_analysis.py on merged.csv and produces the final merged analysis figures analysis_merged 1.PNG and analysis_merged 2.PNG, which directly address the research question about changes in the relationship between financial literacy and credit card debt before vs. after COVID-19.

References

Federal Reserve Board. (2020). *Survey of Household Economics and Decisionmaking (SHED), 2020 Public-Use Microdata*. Retrieved from:
https://www.federalreserve.gov/consumerscommunities/shed_data.htm

Federal Reserve Board. (2024). *Survey of Household Economics and Decisionmaking (SHED), 2024 Public-Use Microdata*. Retrieved from:
https://www.federalreserve.gov/consumerscommunities/shed_data.htm

Consumer Financial Protection Bureau. (2025) *Consumer Credit Trends: Credit Card Originations and Credit Limits*. Retrieved from:
https://www.consumerfinance.gov/data-research/consumer-credit-trends/credit-cards/borrower-risk-profiles/#anchor_lending-levels

Snakemake Authors. (2024). *Snakemake Workflow Management System*.
<https://snakemake.readthedocs.io/>