Data Analysis – You're Now the Hacker

Introduction

Data Hackers Inc. specializes in identifying critical information from aggregate data sources, compiled via phishing attempts or directly from corporate data sources. The most recent heist includes a plethora of data from a popular online shopping platform. As a Principal Hacker, I have been tasked with identifying High Value Targets (HVTs) from the collected data. These are candidates that are best suited for identity and credit card theft, and their data will either be used internally by Data Hackers Inc., or sold to external bad actors.

Methodology

HVTs, as defined in the introduction, are optimal candidates from the aggregate dataset for identity and credit card theft. Based on this principle, the following evaluation protocol was developed for identifying the theft potential of a given person:

1. Likelihood of existing credit or identity fraud.
    a. This was determined through a complex analysis of social security numbers in the data set. See the Analysis section for more information.
2. Spend-to-income ratio (12 month purchase history divided by the annual income).
    a. High spend-to-income ratio is an attribute of targets that would be unlikely to notice irregular spending.
3. Credit Card Expiration date.
    a. Only targets with valid credit cards will be considered.
4. Travel frequency.
    a. Targets that frequently travel will not receive fraudulent spending alerts from their credit card provider.
5. Active account status.
    a. Targets with active accounts are less likely to identify spending on the platform as fraudulent.

The target criteria, listed above, neglects some common factors for identity theft and scams. Factors such as age and income weren't directly considered, as they are secondary to the factors above. The Federal Trade Commission identified online shopping as the highest loss category for all fraud in both the 18-59 and 60+ age groups, effectively eliminating age as a meaningful factor for the fraud being pursued ("Who Experiences Scams"). Income was not a direct consideration as the goal is to identify targets for credit and identity fraud. Individuals that spend more in proportion to their income are more likely to overlook additional charges and are more likely to default on their debt. This fact spans all age demographics, per Andrew Dorn of News Nation (Dorn).

Analysis

The merged dataset was primarily analyzed using the pandas package in python. The following steps were taken as part of the analysis. A detailed breakdown of each step will follow.

1. Identify the number of customer ID occurrences.
2. Within the set of customers with multiple customer ID entries, identify if they are the same people by validating DOB, SSN, Credit Card Number, and others are consistent across customer ID occurrences.
3. At this point, it was decided that SSN might be a better indicator of targets that may already be victims of identity fraud. The same analysis from steps 1 and 2 was completed with SSN as the key row.
4. Filter the dataset of customers with multiple SSN entries by the criteria listed in the Methodology section.
5. Extract the remaining results to identify primary HVTs.
6. Remove the targets identified in step 3 from the global dataset.
7. Repeat step 4 to identify secondary HVTs.
8. Extract the remaining results for secondary HVTs.

Step 1:

Let's take a look at the count of customer id occurrences

```
#Extract the customer ID column
cust_id = list(data["customer_id"])

#Instantiate a defaultdict to count occurrences of customer ids
occurrences = defaultdict(int)

#Loop through the set of ids and append them to the dictionary. If they already exist, add one to the id value.
for id in cust_id:
    if id in occurrences:
        occurrences[id]+=1
    else:
        occurrences[id]=1


print(max(occurrences.values()))
```

✓ 0.0s

```
3
```

Image 1: Extract Occurrences of Customer IDs

The data was previously pulled into a data frame entitled "data." The *customer_id* column was extracted from the data frame and converted to a list, and a dictionary was instantiated to capture the customer IDs as keys and the occurrence count as values. Upon looping through the *cust_id* list, the existence of the current ID key was verified. If it exists, the value was overwritten to the current value plus one, otherwise a new key was added to the dictionary with a value of 1. The max number of customer ID occurrences in the data set was found to be 3, and a histogram plot was generated (see Image 2).

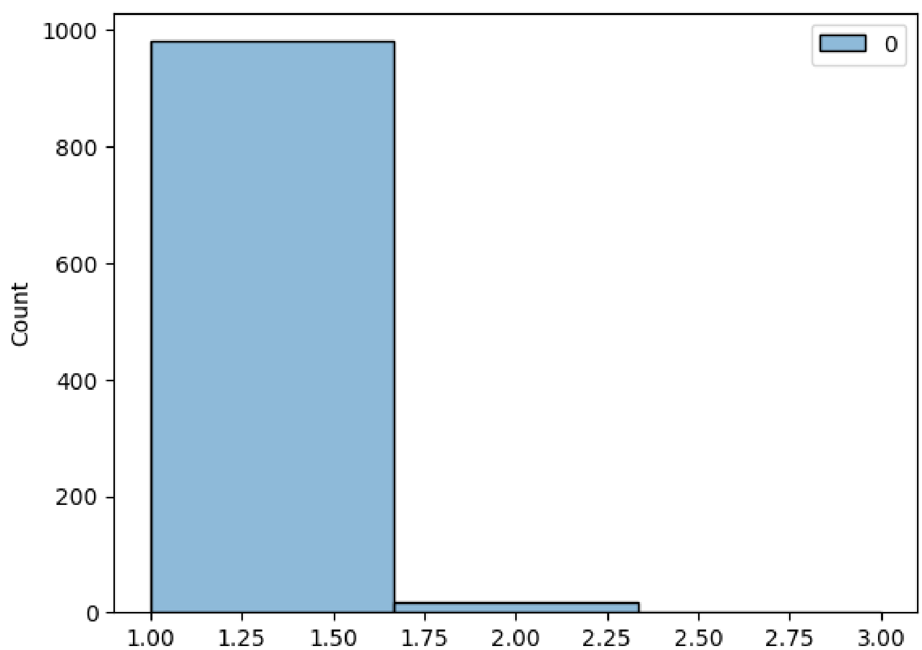
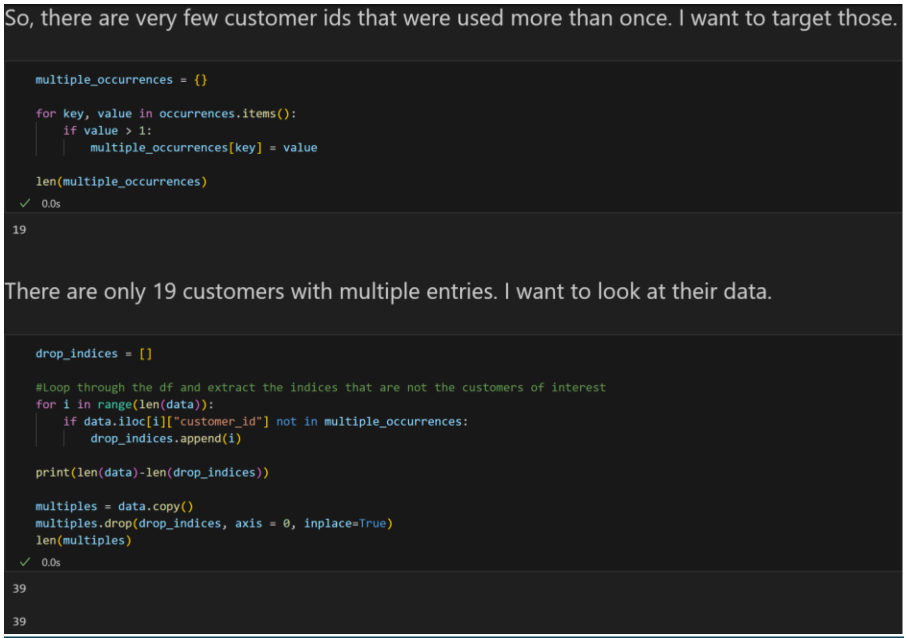


Image 2: Histogram of Customer ID Occurrences

Step 2:



So, there are very few customer ids that were used more than once. I want to target those.

```python
multiple_occurrences = {}

for key, value in occurrences.items():
    if value > 1:
        multiple_occurrences[key] = value

len(multiple_occurrences)
```

✓ 0.0s

19

There are only 19 customers with multiple entries. I want to look at their data.

```python
drop_indices = []

#Loop through the df and extract the indices that are not the customers of interest
for i in range(len(data)):
    if data.iloc[i]["customer_id"] not in multiple_occurrences:
        drop_indices.append(i)

print(len(data)-len(drop_indices))

multiples = data.copy()
multiples.drop(drop_indices, axis = 0, inplace=True)
len(multiples)
```

✓ 0.0s

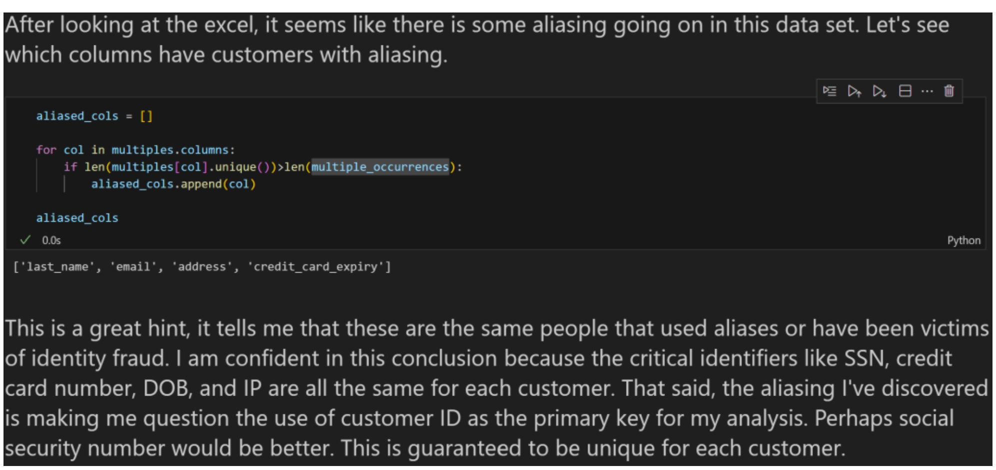
39

39

Image 3: Extract Dataset of Multiple Customer IDs

A new dictionary was instantiated to capture all cases of multiple customer ID occurrences. These customer IDs were checked against the data set, and the indices of non-multiple customer IDs were stored to a list to drop. The original data was copied to a new data frame, then the drop indices were removed to leave all rows that contain customer IDs with greater than one occurrence.

Upon inspection of the multiples dataset in excel, it was clear there were targets with multiple "aliases." An important note is that this could be a result of unclean data, however, this analysis is operating under the assumption that these targets are already victims of identity fraud. To verify the targets are, in fact, the same people, the columns with "aliases" were identified. This was done by iterating through the multiples data set columns and appending the column name to an *aliased_cols* list if there were more unique rows in the column than the total number of customer IDs with multiple occurrences.



After looking at the excel, it seems like there is some aliasing going on in this data set. Let's see which columns have customers with aliasing.

```python
aliased_cols = []

for col in multiples.columns:
    if len(multiples[col].unique())>len(multiple_occurrences):
        aliased_cols.append(col)

aliased_cols
```

✓ 0.0s Python

['last_name', 'email', 'address', 'credit_card_expiry']

This is a great hint, it tells me that these are the same people that used aliases or have been victims of identity fraud. I am confident in this conclusion because the critical identifiers like SSN, credit card number, DOB, and IP are all the same for each customer. That said, the aliasing I've discovered is making me question the use of customer ID as the primary key for my analysis. Perhaps social security number would be better. This is guaranteed to be unique for each customer.

Image 4: Find Aliased Columns

From this analysis step, it was evident that the targets with multiple customer IDs were the same people. However, this step led to a clue in the data set. The existence of multiple customer IDs hints at the existence of multiple Social Security Numbers, which is a better tool for identity theft. If multiple people share the same social security number, it is likely the SSN has been utilized in identity theft by organizations other than Data Hackers Inc.

Step 3:

```
#Extract the ssn column
ssn = list(data["ssn"])

#Instantiate a defaultdict to count occurrences of ssn
ssn_occurrences = defaultdict(int)

#Loop through the set of ids and append them to the dictionary. If they already exist, add one to the id value.
for id in ssn:
    if id in ssn_occurrences:
        ssn_occurrences[id]+=1
    else:
        ssn_occurrences[id]=1


print(max(ssn_occurrences.values()))
```

```
✓ 0.0s

4
```

Image 5: Extracting Occurrences of Social Security Numbers

See step 1 detailed breakdown. The process was identical, this time using the "ssn" column.

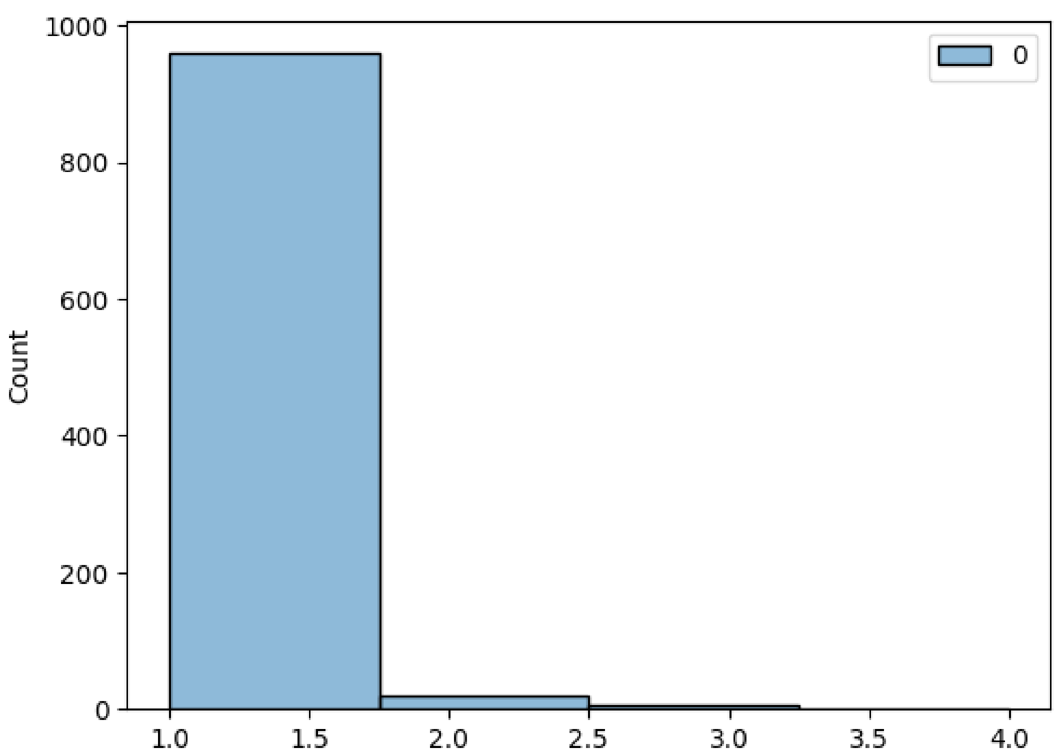


Image 6: Histogram of Social Security Number Occurrences

Step 4:

As stated in the Methodology section, the primary filtering criteria for HVTs are Spend-to-Income ratio, credit card expiration date, travel frequency, and active account status. The Spend-to-Income ratio was computed by dividing 12 month purchase history by annual income for the *multiple_ssn* dataset (seen in Image 7, below). The result was multiplied by 100 to give a percentage of annual income spent on a 12 month cycle.

Okay, I am glad I did this. SSN is a better metric for identifying repeat customers\customers with possibly fraudulent data than customer ID. Based on the excel file, there appear to be unrelated customers that share SSNs in this data set. These customers are potentially already victims of identity fraud. These will be the primary targets to screen for High Value Target status.

To identify high value targets, I want to filter by Annual Income, 12 month purchase history, travel frequency, credit card expiration, and subscription status. I want people with the highest purchase history to income ratio with active subscriptions, frequent travel, and non-expired credit cards. These people are the least likely to identify credit card fraud, and they are frequent travelers, so the credit card agency won't be tipped off if they login/purchase from another IP (me).

```python
#Compute spend to income ratio and add it as a column
multiple_ssn["spend_to_income_ratio"] = multiple_ssn["12 month purchace history"]/multiple_ssn["annual_income"]*100
multiple_ssn["spend_to_income_ratio"]
```

✓ 0.0s Python

Image 7: Compute Spend-to-Income Ratio

With Spend-to-Income ratio computed, the *multiple_ssn* dataset can now be filtered by the remaining criteria. The results of this search are the primary HVTs.

```python
#filter multiple_ssn by credit card expiration and subscription status
primary_hvts = multiple_ssn[(multiple_ssn["subscription_status"]=="Active") & (multiple_ssn["credit_card_expiry"]>datetime.today()) & (multiple_ssn["travel_freq_per_month"]>=5)]
primary_hvts.sort_values(["travel_freq_per_month", "spend_to_income_ratio"], ascending=False)
```

Image 8: Filtering for Primary HVTs

The minimum travel frequency allowed was 5, which averages to more than one trip per week.

Step 5:

The result of this query can be found in Table 1 of the Appendix.

Step 6:

With Primary HVTs identified, the remaining targets in the data set can be evaluated on the same criteria as step 4. First, the remaining targets need to be extracted from the original dataset. As a result of already computing a *multiple_ssn_occurrences* dictionary, this is far easier. As done in step 1, the data set is looped through, and this time the drop indices will only include rows that contain serial numbers with multiple entries.

```
ssn_drop_indices = []

#Loop through the df and extract the indices that are not the customers of interest
for i in range(len(data)):
    if data.iloc[i]["ssn"] in multiple_ssn_occurrences:
        ssn_drop_indices.append(i)

print(len(data)-len(ssn_drop_indices))

single_ssn = data.copy()
single_ssn.drop(ssn_drop_indices, axis = 0, inplace=True)
len(single_ssn)
```

```
✓ 0.0s
959

959
```

Image 9: Removing Duplicate SSNs

Step 7:

The process in step 4 is now repeated with the *single_ssn* dataset. First, Spend-to-Income ratio is calculated.

```
#Compute spend to income ratio and add it as a column
single_ssn["spend_to_income_ratio"] = single_ssn["12 month purchace history"]/single_ssn["annual_income"]*100
single_ssn["spend_to_income_ratio"]
```

Image 10: Compute Spend-to-Income Ratio on Non-Duplicate SSN Dataset

Next, Secondary HVTs are selected using similar filtering criteria to step 4. In an effort to more effectively identify HVTs in this larger dataset, there was an additional filter added on Spend-to-Income ratio. A minimum of 35% was set, as this is well above the 20% utilization threshold that credit bureaus identify as a "healthy" balance when evaluating an individual's credit score.

```
#filter multiple_ssn by credit card expiration and subscription status
secondary_hvts = single_ssn[(single_ssn["subscription_status"]=="Active") & (single_ssn["credit_card_expiry"]>datetime.today()) & (single_ssn["travel_freq_per_month"]>=5) & (single_ssn["spend_to_income_ratio"]>=35)]
secondary_hvts.sort_values(["travel_freq_per_month", "spend_to_income_ratio"], ascending=False)
```

Image 11: Filtering *single_ssn* by HVT Selection Criteria

Step 8:

The result of this query can be found in Table 2 of the Appendix.

Conclusion

The data analysis yielded 6 potential primary HVTs (with one duplicate) and 21 secondary HVTs, for a total of 26 High Value Targets for credit or identity theft. There is one important note regarding the duplicate primary HVT; there is likely enough information in this dataset to verify their true identity through Google or other internet resources. In the case this does not yield a result, both names and credit card information can be used in a purchase attempt to see which is correct. These 26 individuals are the best candidates for credit and identity theft, with the lowest likelihood of legal repercussions. That said, if additional income was needed to keep Data Hackers Inc. operational, the filtering criteria could be loosened to accommodate more targets. These targets are viable, however, their risk profile is greater, as they may be more likely to identify fraudulent charges, or their banks may notice unusual spending from uncommon IP addresses.

Appendix

first_name	last_name	email	phone_number	ssn	dob	credit_card_number	credit_card_expiry	credit_card_security_code	annual_income	12 month purchace history	travel_freq_per_month	spend_to_income_ratio
Paula	Williams	Paula.Williams@gmail.com	045-895-4113	243-63-5844	9/28/2001	566251192993	7/30/2025	913	167154	75707	13	45.2919
Carl	Austin	Carl.Austin@nguyen.com	001-364-375-5444x9263	878-70-5975	10/3/1973	4147411794783260	9/22/2026	822	274158	67951	5	24.7853
Mark	Davis	Mark.Davis@gmail.com	(651)062-4029	722-65-2813	12/14/1992	3581709239860870	12/25/2026	380	170393	46436	13	27.2523
Eric	Davis	Eric.Davis@williams-burton.biz	(365)634-3934x662	455-39-5032	4/27/1965	30397136951419	5/27/2027	104	70690	50817	24	71.8867
Eric	Holmes	Eric.Holmes@williams-burton.biz	(365)634-3934x662	455-39-5032	4/27/1965	30397136951419	9/2/2026	104	70690	50817	22	71.8867
Paula	Williams	Paula.Williams@gmail.com	045-895-4113	243-63-5844	9/28/2001	566251192993	6/9/2027	913	167154	75707	11	45.2919

Table 1: Primary HVTs

first_name	last_name	email	phone_number	ssn	dob	credit_card_number	credit_card_expiry	Credit card security code	Annual income	12 month purchace history	travel_freq_per_month	spend_to_income_ratio
Matthew	Rivera	Matthew.Rivera@turner.org	708-672-3512x0034	761-43-6702	6/21/1947	563556581061	7/8/2026	980	67458.6	66646	12	98.79540933
Robert	Roberts	Robert.Roberts@andrews-rocha.org	146-222-9244x564	863-81-0428	8/19/1951	4587884914303620	6/27/2025	143	145647.6	75914	25	52.12169648
Joseph	Suarez	Joseph.Suarez@garrett.com	378.508.2400	765-05-9614	6/22/1976	3538500908856380	12/31/2025	979	252584	98549	25	39.01632724
Tammy	Fuller	Tammy.Fuller@simpson-johnson.com	558.365.7424x729	835-40-6345	11/17/1955	30137699300103	8/15/2026	516	153547.6	85477	12	55.66807948
James	Lopez	James.Lopez@mueller.com	596.627.9318	700-28-7329	9/29/1966	30592118307982	8/22/2026	5135	40490.4	34618	25	85.49680912
Gregory	Quinn	Gregory.Quinn@montgomery.com	001-275-146-7885x856	684-44-0627	5/14/1952	3532807517386030	12/6/2026	666	118880.8	62917	22	52.92444196
George	Giles	George.Giles@yahoo.com	356-316-5029x620	863-44-4255	9/1/2005	30214634173869	12/20/2025	534	204590	79378	13	38.79857276

Stephanie	Meyer	Stephanie.Meyer@yahoo.com	621160838	828-89-2919	10/22/1989	4663655896653	2/12/2026	666	88989.6	56170	20	63.11973534
Amanda	Parker	Amanda.Parker@kennedy.net	653.482.0085x22150	155-84-9566	12/16/1997	4615530896708340	2/6/2027	957	164626.4	85738	25	52.08034677
Charles	Garcia	Charles.Garcia@gmail.com	+1-585-905-4322x08686	861-91-5218	12/12/1984	4529179742394	1/20/2027	892	18090.4	11263	9	62.25954097
Melissa	Atkins	Melissa.Atkins@williams.com	+1-123-608-4233x436	432-41-1377	10/2/1982	4440239852969	5/15/2027	666	199559.2	74824	20	37.49463818
Daniel	Walsh	Daniel.Walsh@hotmail.com	(442)477-7659x5582	332-16-4913	1/17/1937	3568094571801460	9/24/2025	666	32000	25331	22	79.159375
Frank	Lawrence	Frank.Lawrence@perez.com	105.790.9227x19581	292-48-3023	7/31/1994	4568814298759159808	10/6/2025	6961	38161.2	54521	9	142.8702452
Erik	Pham	Erik.Pham@yahoo.com	(984)829-3100	469-75-5046	12/28/1935	3576707737935050	6/12/2025	75	153408	62708	18	40.8766166
Darlene	Barnes	Darlene.Barnes@davis.net	+1-876-405-3493x81868	610-58-7498	8/25/1978	386439980520 32	5/10/2027	122	85460	73225	5	85.68336064

Nicholas	Lambert	Nicholas.Lambert@jimenez.com	(092)442-2697	799-97-6606	5/30/1994	571762850949	12/23/2025	666	34104.8	17604	13	51.61736764
Gregory	Gonzalez	Gregory.Gonzalez@yahoo.com	001-764-800-3385x05731	556-23-4302	10/19/1950	3595273613675770	4/14/2026	666	79025.6	33028	11	41.79405155
Stephanie	Gonzalez	Stephanie.Gonzalez@rivers.org	001-028-552-1997x274	170-36-6180	6/19/1974	4967525817572090	9/4/2026	748	70598.8	54212	11	76.7888406
Paul	Adams	Paul.Adams@mclaughlin-nelson.org	001-349-158-9165x939	029-11-9412	2/26/2006	630418047977	4/3/2027	666	192196.8	92308	8	48.02785478
Bradley	Graham	Bradley.Graham@french.biz	331492484	600-46-0911	3/22/2005	6011582404401600	10/18/2026	649	201889.6	75011	8	37.15446462
Christine	Ortiz	Christine.Ortiz@zamora.com	1-10-319-1976	477-83-4643	12/27/1968	4898753043788	12/30/2025	625	87788	76233	5	86.83760878

Table 2: Secondary HVTs

Sources

1. Dorn, Andrew. "American Credit Card Debt." *NewsNationNow*, 8 June 2023, www.newsnationnow.com/business/your-money/american-credit-card-debt/. Accessed 8 June 2025.

2. *"Who Experiences Scams? A Story for All Ages." Data Spotlight*, Federal Trade Commission, 8 Dec. 2022, www.ftc.gov/news-events/data-visualizations/data-spotlight/2022/12/who-experiences-scams-story-all-ages. Accessed 8 June 2025.