PayrollTax

William Lee

2023-06-26

Introduction

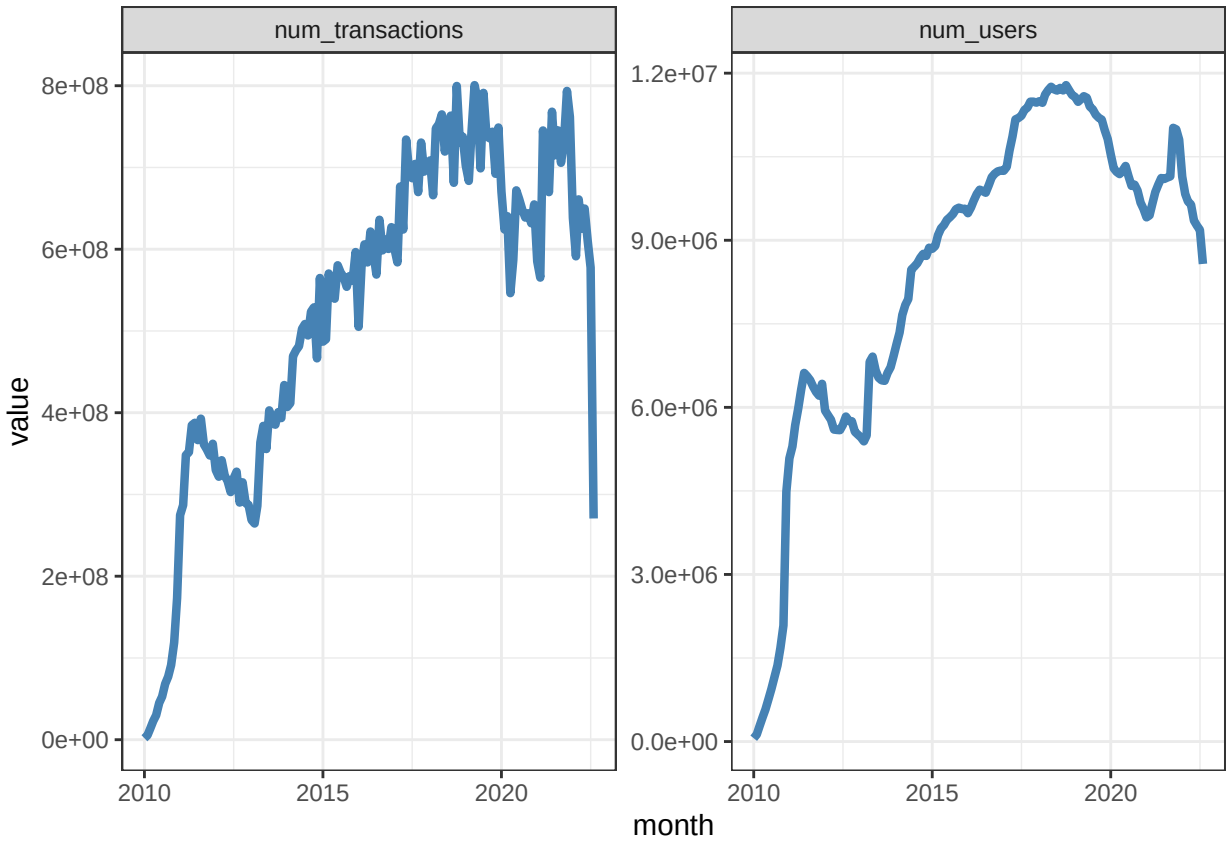
Given the importance of carefully constructing the treatment and control groups for this research project, I figured it would be a good idea to write up the procedure and discuss the results all in one document, with particular emphasis on definitions for eligibility and sample sizes.

Simple Data Set Statistics

The Yodlee data set is indexed by unique_mem_ids which are generated by Yodlee after merging the data from all of their subcontractors. We don't exactly know how the number of unique Yodlee members translates to unique human beings or households. Naturally, Yodlee doesn't reveal how they translate their raw data into unique member ids, however we will assume that each unique member id is truly one individual (or perhaps a joint account). Since Yodlee matches across data sets from different banks and credit card companies, they are likely matching on SSNs.

As of the August 2022 Yodlee batch, there are **56,474,837** distinct unique_mem_ids. Below is the plot showing how the count of unique ids by month changes over time. The sample size grows considerably over time with a few minor dips presumably due to vendor contracts expiring. Throughout our analysis, we will have to keep in mind that we do not have a perfectly balanced sample, and that an individual's presence in the data is not necessarily random.[1]

[1] As an exercise for a later date, we can look to see how the composition of the users changes from regional/internet/brokerage/credit union/national bank using the account_type variable



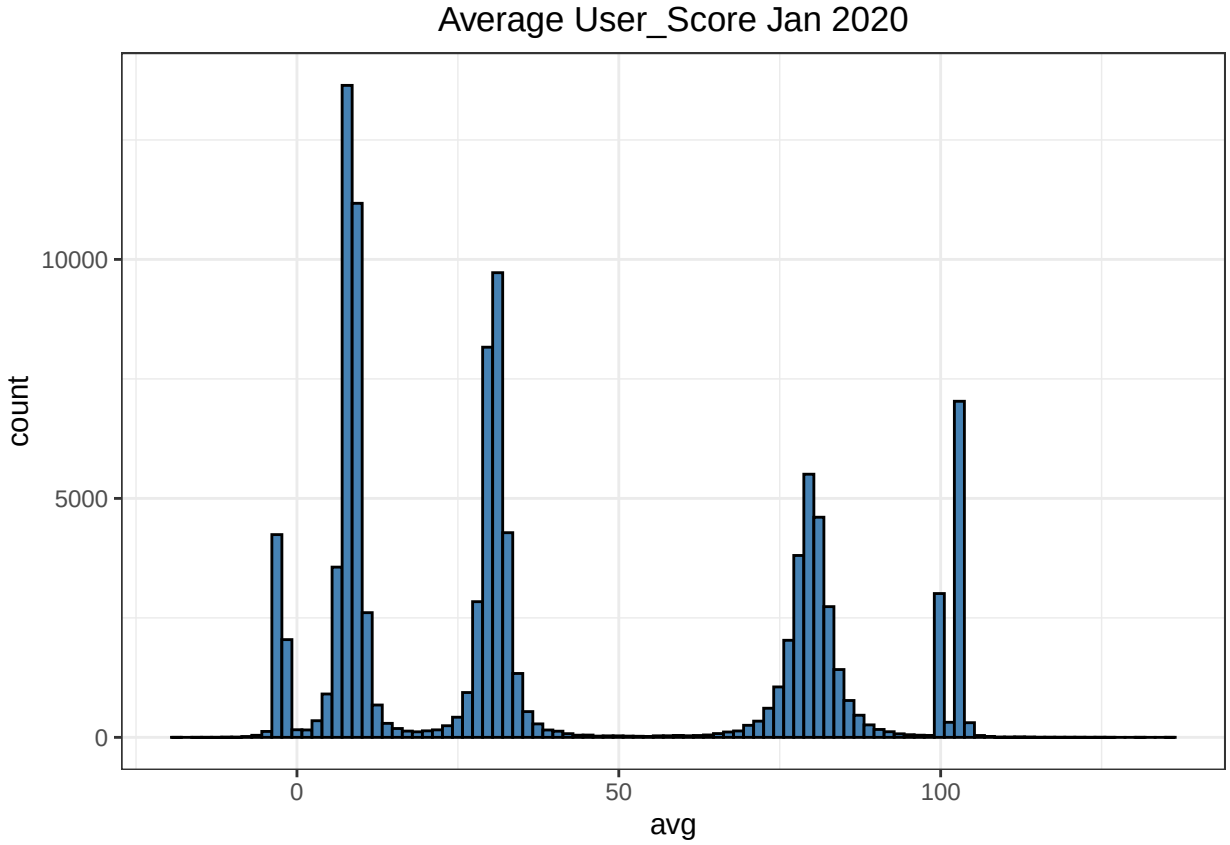
Since we are not able to directly observe whether or not a given user was eligible/enrolled in the Payroll Tax Deferment, we have to infer eligibility from the transaction data and refine our sample so that we have high-quality treatment and control groups. So far, we are working under the assumption that an eligible federal worker is one who:

1. Has an average Yodlee Score of at least 6.5 from August 2018 - present. (This is Yodlee's suggested value for a 'stable' user.)
2. Has qualifying payment observations (based on description, primary_merchant, and amount fields).
3. Has qualifying payment observations from a single employer at regular intervals (weekly, biweekly, or monthly) for 2 years before the deferral went into effect.
4. Makes between \$2,500 - \$8666.67 per month (\$1,153.85 - \$4,000 for biweekly). We rule out individuals making too little as they are unlikely to be engaged in stable, full-time work. (Consider lowering the upper threshold by a certain amount to account for typical withholding). QUESTION: Does eligibility change during the deferral? For example, if you receive a scheduled pay raise in October 2020 that puts you above the threshold, do you still participate in the program?)
5. Observe no more than 35% volatility between paychecks to rule out employees with varying hours worked,travel reimbursements,etc.
6. Observe no more than 20% of total inflows from other sources of income (i.e. SSI, Venmo transactions, transfers from other accounts) (Get feedback on this!)
7. Be able to link all credit card payments to credit cards in the Yodlee sample. Otherwise, we will not be observe debt-funded consumption.

A state/local employee, will match the exact same definition but with a qualifying state/local string.

Sample Size Estimates

In this section, I try to estimate what percent of users are disqualified at each successive component of the definition. Starting with a 1% random sample of the 20220816 Yodlee panel of transactions on or after '2018-08-01', I have **307,380** unique individuals, **619,830** bank accounts **334,199,899** transactions.



The first layer of filtering is for the user_score. Below, I plot the histogram of the average monthly user_score for January 2020 (a random month before Covid). I'm not sure how Yodlee decides these user_scores, but there are some clear clusters of user_scores. The user_score varies a lot even within a single month (average range of user_score was 45.54). I would love to know why user_scores vary so much in a short amount of time since things like "User History, presence of a linked account, and number of merchants" shouldn't change dramatically in a short amount of time.

After removing all individuals with an average user_score of less than 6.5 over the sample period, we are down to 290,825 users, representing a 5.39 percent reduction in the sample. Perhaps it's worth exploring a higher threshold, especially since the main analysis may involve daily time coefficients.

The second layer of filtering is to sort each individual into federal/state/local based on the description and primary merchant. The full .sql used is included in the appendix, but uses all of the patterns we have found up to this point. After filtering, we are left with 17,400 individuals which represents a 94.02 percent reduction in eligible individuals. 12753 are federal, 1919 are state, and 3203 are local.

I tried to match these figures with outside estimates of the federal/state/local workforce. Despite the lack of a single definition for federal worker, this Brookings piece has similar estimates to our data (~15% in some sort of federal/state/local government, ~6% in federal).

After the third round of filtering, we are left with 3,560 individuals which represents a 79.54 percent reduction in eligible individuals. Most of the reduction is coming from the restriction on having four years of consistent paychecks from a single employer. Should we relax the definition a bit on the pre-period?

Below is a list of the top 10 employers by each category:

- The empty primary_merchant_name for the federal category is the generic "Fed Sal" category. Not sure what to do with it
- I believe we can tell which branch of the military the DFAS is coming from based on the city
- Do we want to keep obscure federal employers like the Bonneville Power Authority?
- I forgot what we decided on school boards and public education for local employers. I decided to keep for now
- For the state employers, I'm pretty certain we are missing something because California, Texas, and Florida should be the largest employers, not Maryland. Perhaps some states centralize payroll and others decentralize. Probably something worth figuring out at some point.

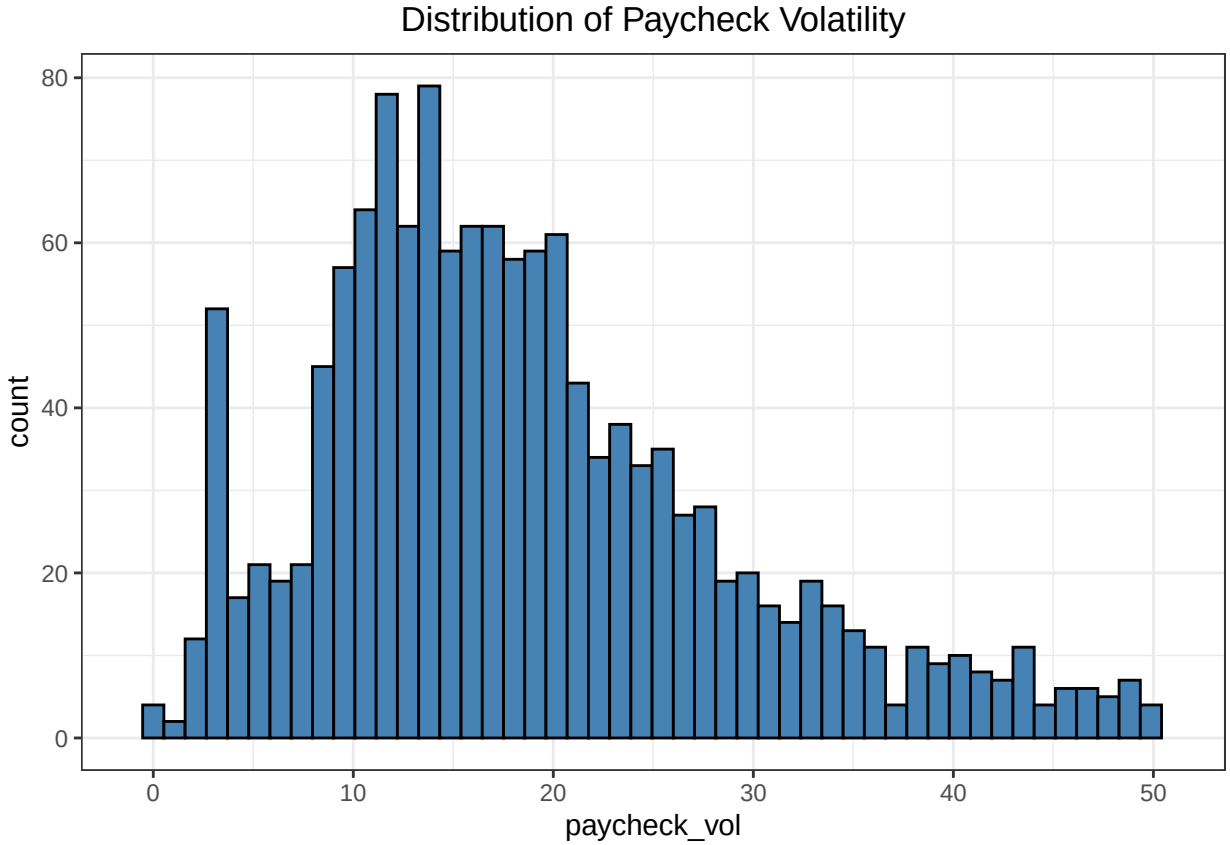
fed	primary_merchant_name	n_unique_id	rank
federal	DFAS	2432	1
federal	U.S. Department Of The Treasury	340	2
federal	The U.S. Office Of Personnel Management	212	3
federal	United States Coast Guard	143	4
federal	Us Treasury	23	5
federal		14	6
federal	Agricultural Treasury Office	9	7
federal	General Services Administration	7	8
federal	U.s. Department Of Health And Human Services	3	9
federal	Bonneville Power Administration	1	10
local	Anne Arundel County Board	15	1
local	City Of New York	14	2
local	Montgomery County Government	10	3
local	City Of Columbus	7	4
local	City Of Chicago	6	5
local	Allegheny County	5	6
local	Cook County	5	7
local	Franklin County	5	8
local	Fulton County Schools	5	9
local	Gwinnett County Board Of Education	5	10
state	State Of Maryland	31	1
state	Commonwealth of Pennsylvania	19	2
state	State Of California	13	3
state	West Virginia State Treasurer	8	4
state	State Of Indiana	6	5
state	Washington State Treasurer	5	6
state	State Of Alaska	4	7
state	State Treasurer	4	8
state	State Of Alabama	3	9
state	State Of Arizona	3	10

I'd be interested to hear your thoughts on whether restricting our analysis to people who do not switch jobs or take side-jobs limits the external validity of our sample. On one hand, a refined sample will allow us to cleanly estimate the effects of the deferral on individual, but on the other hand, we will not be to make strong statements about the overall impact of the program.

After the fourth layer of filtering which classes individuals by income thresholds, we are left with the following number of unique_id_mem's. The proportions seem about right, given that most government employees that have regular paychecks should make between 30,000 - 104,000 USD per year.

elig	federal	local	state
elig	1045	164	82
inelig	129	9	4

Below is a histogram of percentage volatility (std(paycheck_amt)*100/mean(paycheck_amt)). The threshold is entirely arbitrary, so I wanted to show the general distribution. I need to examine these individuals more to see why they have such volatility, but I think it's better to exclude them for now until we can better understand the trends. I'm guessing the common causes for paycheck volatility are pay raises, bonuses (are these common for government employees), changes in withholding, reimbursements, etc.



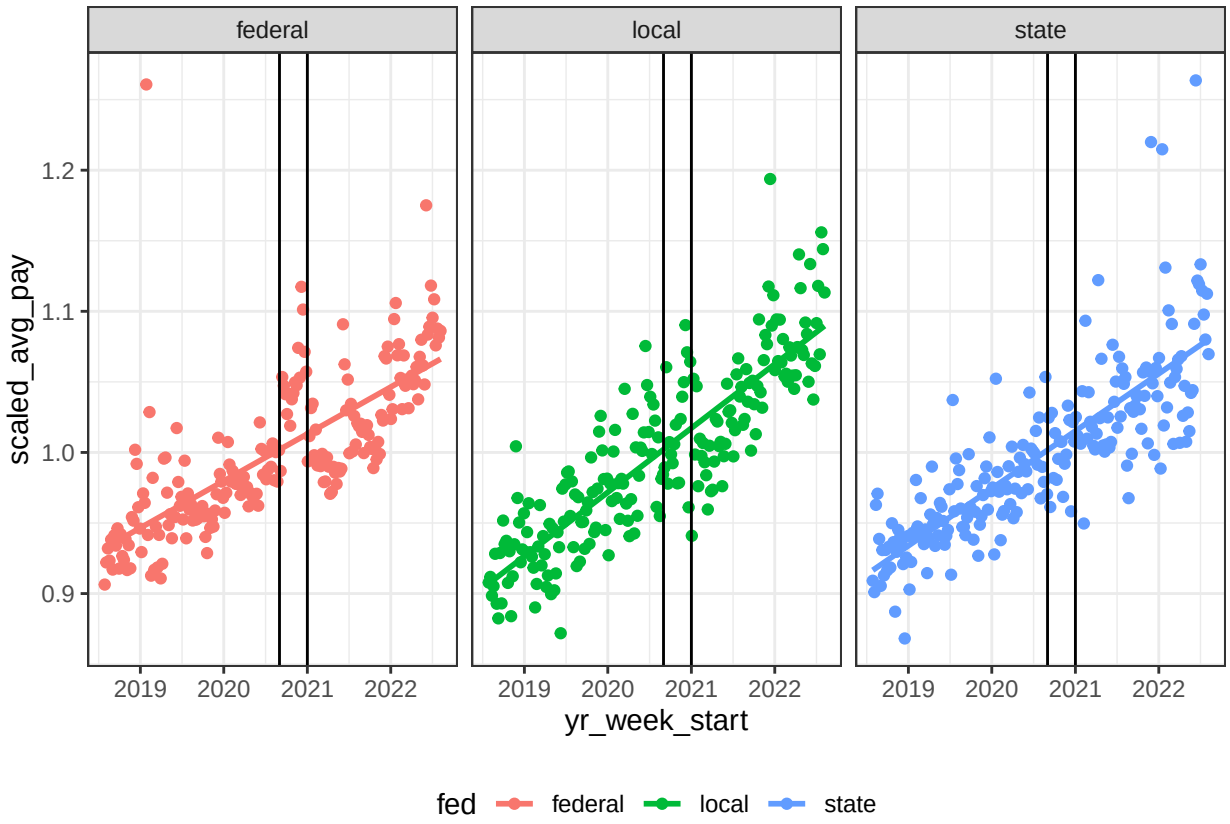
After the fifth round of filtering which removes individuals with high volatility in paychecks, we are left with the following number of individuals in each category.

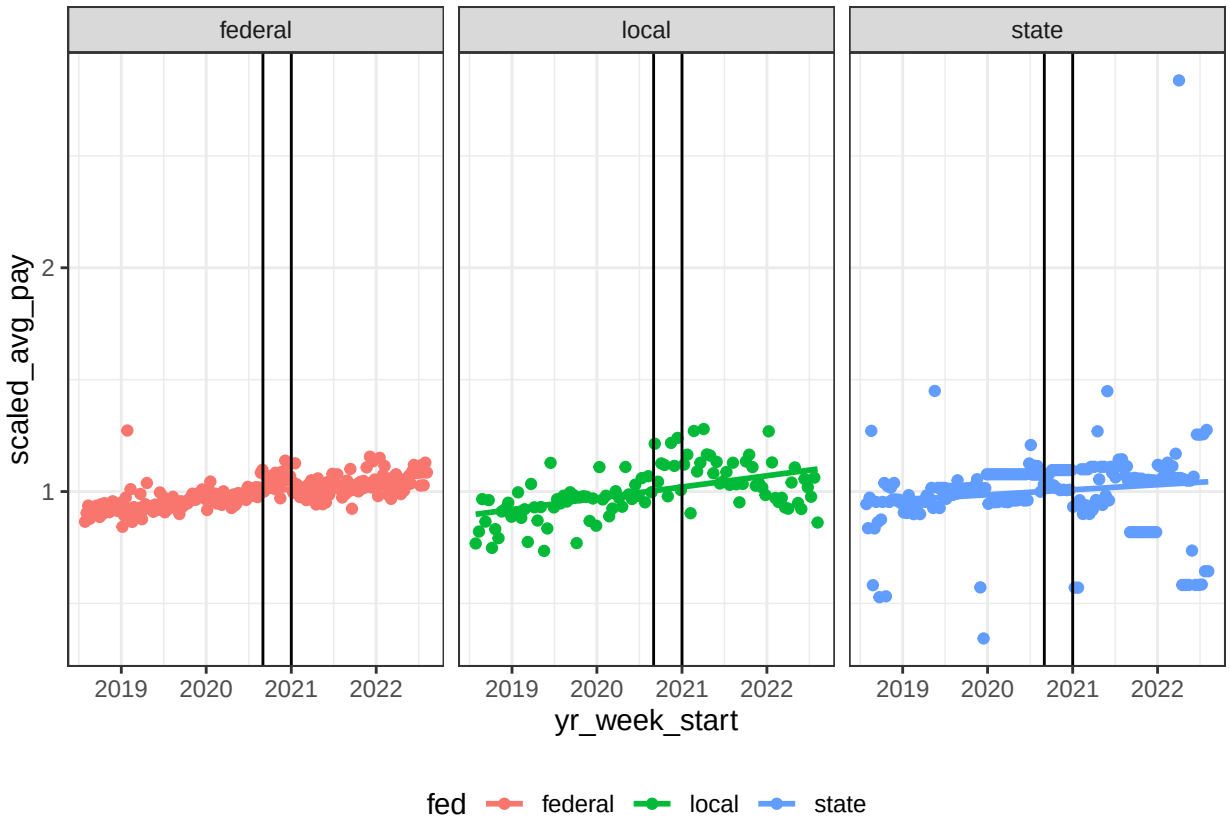
elig	federal	local	state
elig	916	149	75
inelig	97	4	4

There are two more layers of filter: one for ruling out other sources of income and another for ensuring link-ability of Yodlee data. I have some ideas on how to do this efficiently, but I'd like your feedback first.

Paycheck Graph

Here is the most important piece of the whole exercise. After all of this cleaning, we need to see if the deferral is visible in a simple weekly aggregate plot, and I think it looks pretty good. The trend is clearly visible for eligible federal employees but absent for state and local employees. However, I'm concerned to see a small increase in the ineligible (due to the salary cutoff) federal employees during the payroll period.


federal
local
state
scaled_avg_pay
1.2
1.1
1.0
0.9
2019
2020
2021
2022
yr_week_start
fed
federal
local
state



Appendix

You can find my code and outputs on my personal github page (https://github.com/WilliamHLee104/PayrollTax)[https://github.com/WilliamHLee104/PayrollTax].

```
-- Simple Count for Total Population
select count(distinct unique_mem_id) as num_users
from yi_xpanelov6_20220816.bank_panel

-- Count By Month (saved to count_by_month.csv)
select count(distinct unique_mem_id) as num_users, count(*) as num_transactions, month
from (select substring(optimized_transaction_date, 1, 7) as month, unique_mem_id
      from yi_xpanelov6_20220816.bank_panel) as month_create
GROUP BY month
ORDER BY month

-- Checking if IDs are truly random. They are
select count(*), enddigits
from (select mod(unique_mem_id, 100) as enddigits from yi_xpanelov6_20220816.bank_panel)
GROUP BY enddigits
ORDER BY enddigits

-- Create 1% sample and store in my temp directory
CREATE TABLE temp_132.onepercsample AS (SELECT *
                                        FROM yi_xpanelov6_20220816.bank_panel
```

```
                                        WHERE mod(unique_mem_id, 100) = 1
                                          AND optimized_transaction_date >= '2018-08-01')

-- Count Users/Transactions in the 1% sample
select count(distinct unique_mem_id) as num_users,
       count(distinct unique_bank_account_id),
       count(*)                      as num_transactions
from temp_132.onepercsample

-- User Score Dist (saved to user_score_dist.csv)
select unique_mem_id, avg(user_score), min(user_score), max(user_score), count(*)
from temp_132.onepercsample
where optimized_transaction_date >= '2020-01-01'
  AND optimized_transaction_date < '2020-02-01'
group by unique_mem_id


-- Filter By User Score
-- NB: Next round of filtering uses filter1 as a base table
create table temp_132.filter1 as (SELECT b.*
                                  from (select unique_mem_id,
                                               avg(user_score)                                    as avg,
                                               min(user_score)                                    as min,
                                               substring(min(optimized_transaction_date), 1, 7) as min_
                                               substring(max(optimized_transaction_date), 1, 7) as max_
                                        from temp_132.onepercsample
                                        group by unique_mem_id) a
                                           inner join temp_132.onepercsample b
                                                      on a.unique_mem_id = b.unique_mem_id
                                  where avg > 6.5)


select count(distinct unique_mem_id)
from temp_132.filter1


-- Find all Qualifying Federal/State/Local Payroll Transactions
select *
from (select *,
             CASE
                 WHEN ((upper(primary_merchant_name) like '%DFAS%' OR
                        upper(primary_merchant_name) like '%U.S. DEPARTMENT OF THE TREASURY%' OR
                        upper(primary_merchant_name) like '%US TREASURY%' OR
                        upper(primary_merchant_name) like '%GOVERNMENT%' OR
                        upper(primary_merchant_name) like '%GSA%' OR
                        upper(primary_merchant_name) like '%THE GENERAL SERVICES ADMINISTRATION%' OR
                        upper(primary_merchant_name) like '%THE U.S. OFFICE OF PERSONNEL MANAGEMENT%' O
                        upper(primary_merchant_name) like '%UNITED STATES COAST GUARD%' OR
                        upper(primary_merchant_name) like '%U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
                        upper(primary_merchant_name) like '%AGRICULTURAL TREASURY OFFICE%' OR
                        upper(primary_merchant_name) like '%CENSUS%' OR
                        upper(primary_merchant_name) like '%SOCIAL SECURITY ADMINISTRATION%' OR
                        upper(primary_merchant_name) like '%FARM SERVICE AGENCY%' OR
```

```
   description ilike '%FED SAL%'
or description ilike '%FAA TREAS 310%'
or description ilike '%EPA TREAS 310%'
or description ilike '%GSA TREAS 310%'
or description ilike '%DOI1 TREAS 310%'
or description ilike '%DOT4 TREAS 310%'
or description ilike '%NIH  TREAS 310%' or description ilike '%NIH. TREAS 310%'
or description ilike '%OPM1 TREAS 310%'
or description ilike '%DHS  TREAS 310%'
or description ilike '%LOC1 TREAS 310%'
or description ilike '%USSS TREAS 310%'
or description ilike '%CBP  TREAS 310%'
or description ilike '%DOJ  TREAS 310%'
or description ilike '%USSS TREAS 310%'
or description ilike '%US HOUSE OF REPR%'
or description ilike '%US SENATE FED SAL%'
or description ilike '%TENN VALLEY AUTH TRPDFEDSL%'
or description ilike '%TENN VALLEY AUTH ACH: TRPDFEDSL%'
or description ilike '%US SENATE FED SAL%'
or description ilike '%USPS%'
or description ilike '%IN AF PAY%'
or description ilike '%IN ARMY ACT%'
or description ilike '%IN AF PAY%'
or description ilike '%IN AF RES%'
or description ilike '%IN ARMY RC%'
or description ilike '%NAVY ACT%'
or description ilike '%NAVY ALT%'
or description ilike '%NAVY RES%')
AND description not ilike '%SSA  TREAS 310%'
AND description not ilike '%SOC SEC%'
AND description not ilike '%VA BEN%'
AND description not ilike '%TREASURY PMN%'
AND description not ilike '%SERV F%'
AND description not ilike '%SUPP SEC%'
AND description not ilike '%US TREASURY CF%'
AND description not ilike '%TAX%'
AND description not ilike '%RET%'
AND description not ilike '%FED PAYMENT%'
AND description not ilike '%ALLT%'
AND description not ilike '%PPTAS%'
AND description not ilike '%BENEFIT PAYMENT%'
AND description not ilike '%TRAVEL PAY%'
AND description not ilike '%UI BEN%'
AND description not ilike '%USCIS%'
AND description not ilike '%VACP%'
AND description not ilike '%DCPS%'
AND description not ilike '%CASH%'
AND description not ilike '%IATS PAY%'
AND description not ilike '%MISC PAY%'
AND description not ilike '%NJ SDU%'
AND description not ilike '%TREAS 449%'
AND description not ilike '%SDP%'
AND description not ilike '%CHILD%'
```

```
                AND description not ilike '%FAIRFAX%'
                AND description not ilike '%GOVERNMENT SOLUTIONS%'
                AND description not ilike '%GOVERNMENT SERVICES%'
                AND description not ilike '%GOVERNMENT VI%'
                AND description not ilike '%COUNTY%'
                AND description not ilike '%EITX%'
                AND description not ilike '%CITY%'
                AND description not ilike '%ASI GOV%'
                AND description not ilike '%STUDENT LN%'
                AND description not ilike '%STATE%'
                AND description not ilike '%POLICE%'
                AND description not ilike '%NY %'
                AND description not ilike '%OHIO%'
                AND description not ilike '%AR.GOV%'
                AND description not ilike '%NJMONT%'
                AND description not ilike '%EDUCATION%'
                AND description not ilike '%KANSAS%'
                AND description not ilike '%NEWYORK%'
                AND description not ilike '%SSA TREAS 310%'
                AND description not ilike '%SBAD TREAS 310%'
                AND description not ilike '%RRB  TREAS 310%'
                AND description not ilike '%RRB TREAS 310%'
                AND description not ilike '%DOEP TREAS%'
                AND description not ilike '%DFEC TREAS 310%'
                AND primary_merchant_name not ilike '%COUNTY%'
                AND primary_merchant_name not ilike '%USPS%'
                AND primary_merchant_name not ilike '%LOUISIANA%'
                AND primary_merchant_name not ilike '%ACCO BRANDS%'
                AND primary_merchant_name not ilike '%KEYPOINT GOVERNMENT SOLUTIONS%'
                AND primary_merchant_name not ilike '%ASCENSUS TRUST%'
                AND primary_merchant_name not ilike '%US NAVY NSA PC MORALE WELFARE & RECREATION%'
                AND primary_merchant_name not ilike '%SOCIAL SEC%'
                AND primary_merchant_name not ilike '%NATIONAL GOVERNMENT SERVICES%'
                AND primary_merchant_name not ilike '%FEDERAL RESERVE%') THEN 'federal'
            WHEN ((upper(primary_merchant_name) like '%STATE TREASUR%' or
                   upper(primary_merchant_name) like '%STATE COMPTROL%' or
                   upper(primary_merchant_name) like '%STATE CONTROLLER%' or
                   upper(primary_merchant_name) like '%ST OF%' or
                   upper(primary_merchant_name) like '%STATEOF%' or
                   upper(primary_merchant_name) like '%STATE OF%' or
                   upper(primary_merchant_name) like '%COMMONWEALTH OF%' or
                   upper(primary_merchant_name) like '%DEPARTMENT OF%' or
                   upper(primary_merchant_name) like '%STATE DEPARTMENT%' or
                   (upper(primary_merchant_name) like '%STATE OF COLORADO' and
                    description not ilike '%COLORADO STATE U%') or
                   (upper(primary_merchant_name) like '%STATE OF ILLINOIS' and description ilike '
                   upper(primary_merchant_name) like '%LOUISIANA GOVERNMENT%')
                and description not ilike '%TAX%'
                and description not ilike '%UI%'
                and description not ilike '%UNEMP%'
                and description not ilike '%DSS%'
                and description not ilike '%REFUND%'
                and description not ilike '%BENEFIT%'
```

```
                    and description not ilike '%CHILD%'
                    and description not ilike '%EITX%'
                    and description not ilike '%SUPP%'
                    and upper(primary_merchant_name) not like '%U.S. DEPARTMENT OF THE TREASURY%'
                    and upper(primary_merchant_name) not like '%US TREASURY%'
                    and upper(primary_merchant_name) not like '%US DEPARTMENT OF EDUCATION%'
                    and upper(primary_merchant_name) not like '%U.S. DEPARTMENT OF HEALTH AND HUMAN SE
                    and upper(primary_merchant_name) not like '%DEPARTMENT OF VETERAN%'
                    and upper(primary_merchant_name) not like '%UNITED STATES%'
                    and upper(primary_merchant_name) not like '%STAR OF%'
                    and upper(primary_merchant_name) not like '%TAX%'
                    and upper(primary_merchant_name) not like '%BLUE CROSS%'
                    and upper(primary_merchant_name) not like '%POWER%'
                    and upper(primary_merchant_name) not like '%HEALTH%'
                    and upper(primary_merchant_name) not like '%MEDIC%'
                    and upper(primary_merchant_name) not like '%UNIV%'
                    and upper(primary_merchant_name) not like '%ELECTRIC%'
                    and upper(primary_merchant_name) not like '%CORP%'
                    and upper(primary_merchant_name) not like '%AIRLINE%'
                    and upper(primary_merchant_name) not like '%PACIFIC%'
                    and upper(primary_merchant_name) not like '%TOOL%'
                    and upper(primary_merchant_name) not like '%CLEANER%'
                    and upper(primary_merchant_name) not like '%DINER%'
                    and upper(primary_merchant_name) not like '%EDISON%'
                    and upper(primary_merchant_name) not like '%EXPRESS%'
                    and upper(primary_merchant_name) not like '%SOUTHERN%'
                    and upper(primary_merchant_name) not like '%LOTTERY%'
                    and upper(primary_merchant_name) not like '%PRIME%'
                    and upper(primary_merchant_name) not like '%VISION%'
                    and upper(primary_merchant_name) not like '%EMC%'
                    and upper(primary_merchant_name) not like '%CENTRAL%'
                    and upper(primary_merchant_name) not like '%LIFE%'
                    and upper(primary_merchant_name) not like '%INSUR%'
                    and upper(primary_merchant_name) not like '%BERGEN%'
                    and upper(primary_merchant_name) not like '%ROADHOUSE%'
                    and upper(primary_merchant_name) not like '%CHILD%'
                    and upper(primary_merchant_name) not like '%TECH%'
                    and upper(primary_merchant_name) not like '%SPCA%'
                    and upper(primary_merchant_name) not like '%TOYOTA%'
                    and upper(primary_merchant_name) not like '%TIMES%'
                    and upper(primary_merchant_name) not like '%ZOO%'
                    and upper(primary_merchant_name) not like '%CENTER%'
                    and upper(primary_merchant_name) not like '%MADE%'
                    and upper(primary_merchant_name) not like '%DENT%'
                    and upper(primary_merchant_name) not like '%CNG%'
                    and upper(primary_merchant_name) not like '%SOURCE%'
                    and upper(primary_merchant_name) not like '%HOTEL%'
                    and upper(primary_merchant_name) not like '%RAILR%'
                    and upper(primary_merchant_name) not like '%FRESH%'
                    and upper(primary_merchant_name) not like '%YORKER%'
                    and upper(primary_merchant_name) not like '%THEATRE%'
                    and upper(primary_merchant_name) not like '%GRILL%'
                    and upper(primary_merchant_name) not like '%GENESEE%'
```

```
                    and upper(primary_merchant_name) not like '%FURNITURE%'
                    and upper(primary_merchant_name) not like '%EAST%'
                    and upper(primary_merchant_name) not like '%COLLEGE%'
                    and upper(primary_merchant_name) not like '%GAS%'
                    and upper(primary_merchant_name) not like '%UTILIT%'
                    and upper(primary_merchant_name) not like '%COFFEE%'
                    and upper(primary_merchant_name) not like '%HOSPITAL%'
                    and upper(primary_merchant_name) not like '%RETIR%'
                    and upper(primary_merchant_name) not like '%REVENUE%') THEN 'state'
                WHEN ((upper(primary_merchant_name) like '%COUNTY OF%' or
                       upper(primary_merchant_name) like '%COUNTY%' or
                       upper(primary_merchant_name) like '%CITY OF%' or
                       upper(primary_merchant_name) like '%DEPARTMENT OF%' or
                       upper(primary_merchant_name) like '%CITY DEPARTMENT%' or
                       upper(primary_merchant_name) like 'PUBLIC SCHOOLS')
                    and description not ilike '%TAX%'
                    and description not ilike '%UI%'
                    and description not ilike '%DSS%'
                    and description not ilike '%UNEMP%'
                    and description not ilike '%REFUND%'
                    and description not ilike '%BENEFIT%'
                    and description not ilike '%CHILD%'
                    and upper(primary_merchant_name) not like '%U.S. DEPARTMENT OF THE TREASURY%'
                    and upper(primary_merchant_name) not like '%US DEPARTMENT OF EDUCATION%'
                    and upper(primary_merchant_name) not like '%U.S. DEPARTMENT OF HEALTH AND HUMAN SE
                    and upper(primary_merchant_name) not like '%US TREASURY%'
                    and upper(primary_merchant_name) not like '%DEPARTMENT OF VETERAN AFFAIRS%'
                    and upper(primary_merchant_name) not like '%DEPARTMENT OF VETERANS AFFAIRS%'
                    and upper(primary_merchant_name) not like '%ELECTRIC%'
                    and upper(primary_merchant_name) not like '%GAS%'
                    and upper(primary_merchant_name) not like '%UTILIT%'
                    and upper(primary_merchant_name) not like '%Prince William County Service Authorit
                    and upper(primary_merchant_name) not like '%UNIVERSITY%'
                    and upper(primary_merchant_name) not like '%HOSPITAL%'
                    and upper(primary_merchant_name) not like '%RETIR%'
                    and upper(primary_merchant_name) not like '%REVENUE%') THEN 'local'
                ELSE 'other' END as fed
      from temp_132.filter1
      where transaction_base_type = 'credit'
        and transaction_category_name = 'Salary/Regular Income'
        and amount > 500
        and is_duplicate = 0)
where fed not like 'other'
```