Transaction Classification Update

William Lee

2023-08-17 14:11:18.34485

Contents

What have I been up to?

- Rewriting the code that identifies employers
  - Now can separate between higher ed, local ed, and local.
  - Broken down federal categories as much as possible by description string (e.g. now sort DFAS into Navy/Army/Air Force whenever possible)
  - Instead of doing a broad search and eliminating case by case, I manually write out each case (hopefully fewer false positives)
  - Runs in a more efficient manner on the AWS cluster
- I've come to the conclusion that Yodlee's transaction classification and primary_merchant was very bad (some salary payments show up as deposits, insurance, etc. and sometimes payments that obviously fit the same pattern as others are not given the same primary_merchant tag.)

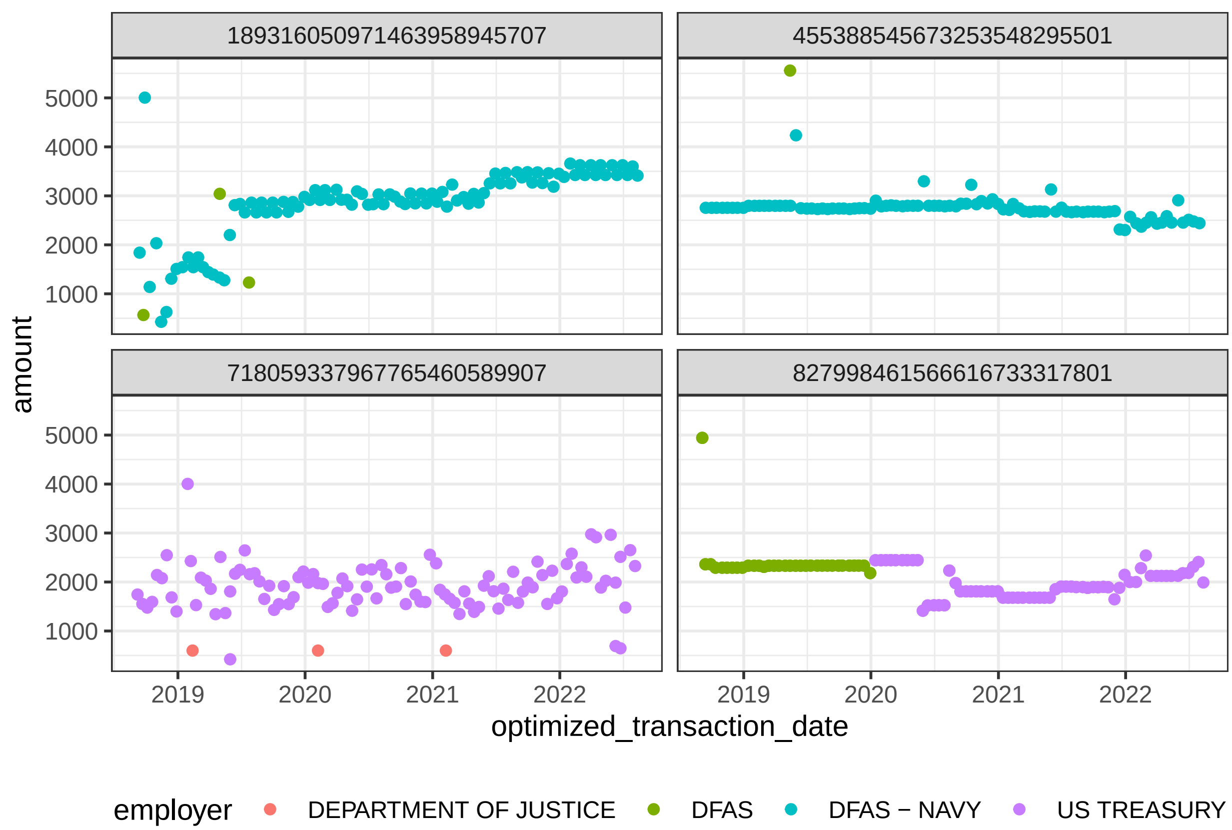
- Went through each state to make sure we were not getting unemployment or child support benefits as income because Yodlee will mark those as coming from 'State of XX'. I hope I've ruled out all retirement benefits, but I can't be sure. Most of the state and federal payroll strings have 'Salary/Payroll/Misc' in them. I made sure to exclude 'ERS/RETIR' etc from consideration.
- Implemented new income cutoffs as per your suggestions (discussed more later)
- Downloaded a bigger sample
- Re-ran regressions

One thing that's a bit odd in the Yodlee data is that sometimes the strings patterns will abruptly switch for no discerable reason (or at least a reason I have not been able to discern). For example, an individual might receive most of their pay from with one pattern that identifies them as DFAS but will then have a few strings that identify them as DFAS- NAVY. After discussing this with Scott, we have decided to mark this person as DFAS-NAVY for all of their transactions.

For some of the US Treasury employees, it looks like they receive their base salary from *'FED SAL'* and then receive *'MISC PAY'* from an TREAS 310 account that shows their department. Again, after discussing this with Scott, we have decided to mark all of their transactions as coming from the agency in the MISC PAY.

I've included a few examples in the plot below where each panel is a different unique_mem_id. I'm particularly interested in bottom right panel where it looks like the person goes from DFAS to US TREASURY but maintains almost the exact same salary. Do we think this is a job switch or just a switch in the formatting. I'm just a switch in the formatting since there are others with the same pattern. Currently, I'm assigning this person to US TREASURY for all the payments.

There is this enormous guide that seems like it should have the answers if I read enough. https://www.dfas.mil/contractorsvendors/miscpaymentguide/



String Search Update

After our last meeting, I realized that I needed to put some effort into improving the method of finding eligible employers. I was especially concerned with the cases where payroll transactions were mis-classified or where non-wage income might appear to be Salary/Regular Income (i.e. child support or OPM pension payments looking like regular income). I've learned a decent amount by just looking at the strings:

- We can determine which agency a person works for in the federal government whenever they receive MISC/TVL pay.
- Went state by state to find their payroll strings and common benefit strings (unemployment, retirement, child support) to make sure we are only getting payroll. We now have coverage for almost 40 states instead of the 10 for the last update.
- Found the payment strings for the largest public universities (and smaller universities that were clearly marked in the Yodlee data) to make sure state payroll is identified separately from state higher ed
- Separated local strings by local and local_education for similar reasons - education spending is more volatile, different employment dynamics, etc. We are missing a few of the top largest cities (possibly because they use a benefit manager).

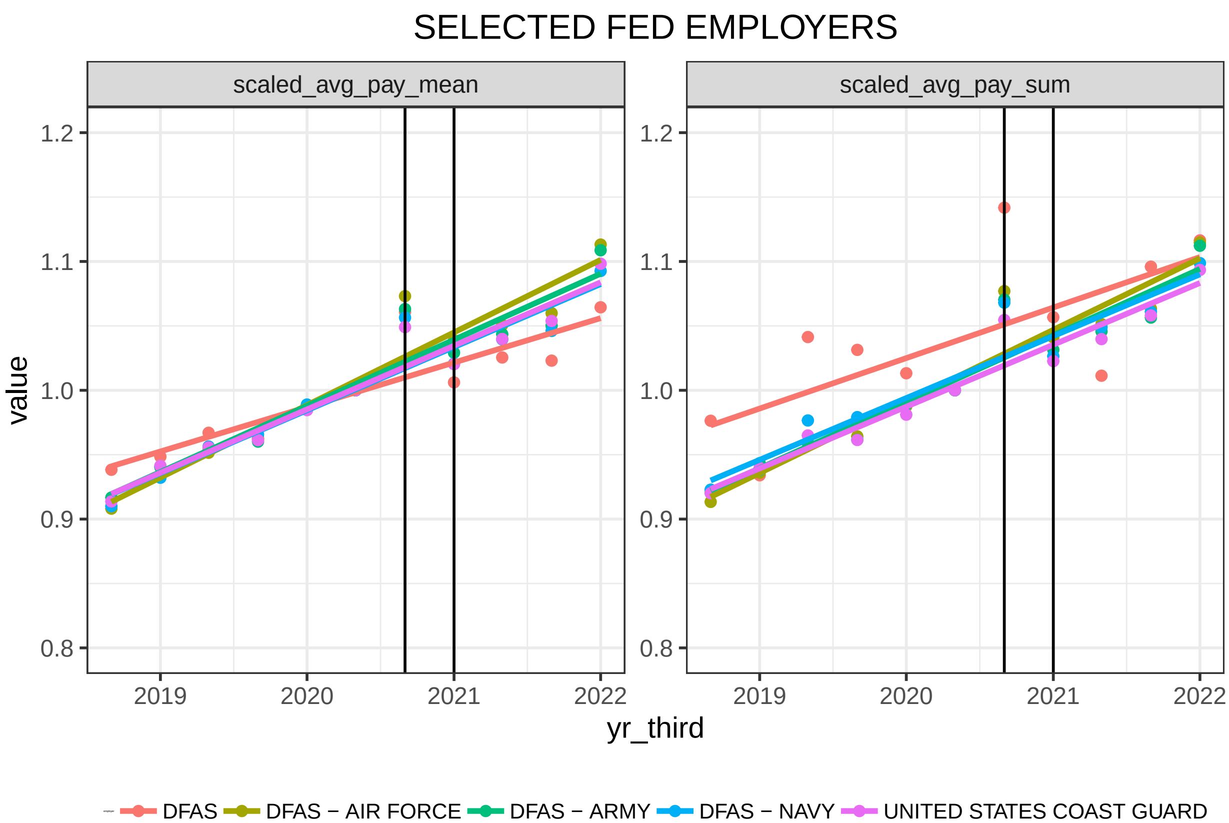
Overall, the underlying strategy of the algorithm shifted. Instead of starting with all primary_merchants that match 'STATE OF' and then manually removing non-state vendors like 'ALLSTATE OF IL', I now only consider strings/merchants that match the **when primary_merchant_name ilike '%STATE OF ILLINOIS%' and description similar to '(%Payroll%|%Deposit%)' and description not similar to '(%Commercial%|%Tax%)' then 'IL'**. I think this is probably what I should have done in the first place, but we can always go back to the original method if you think that's better.

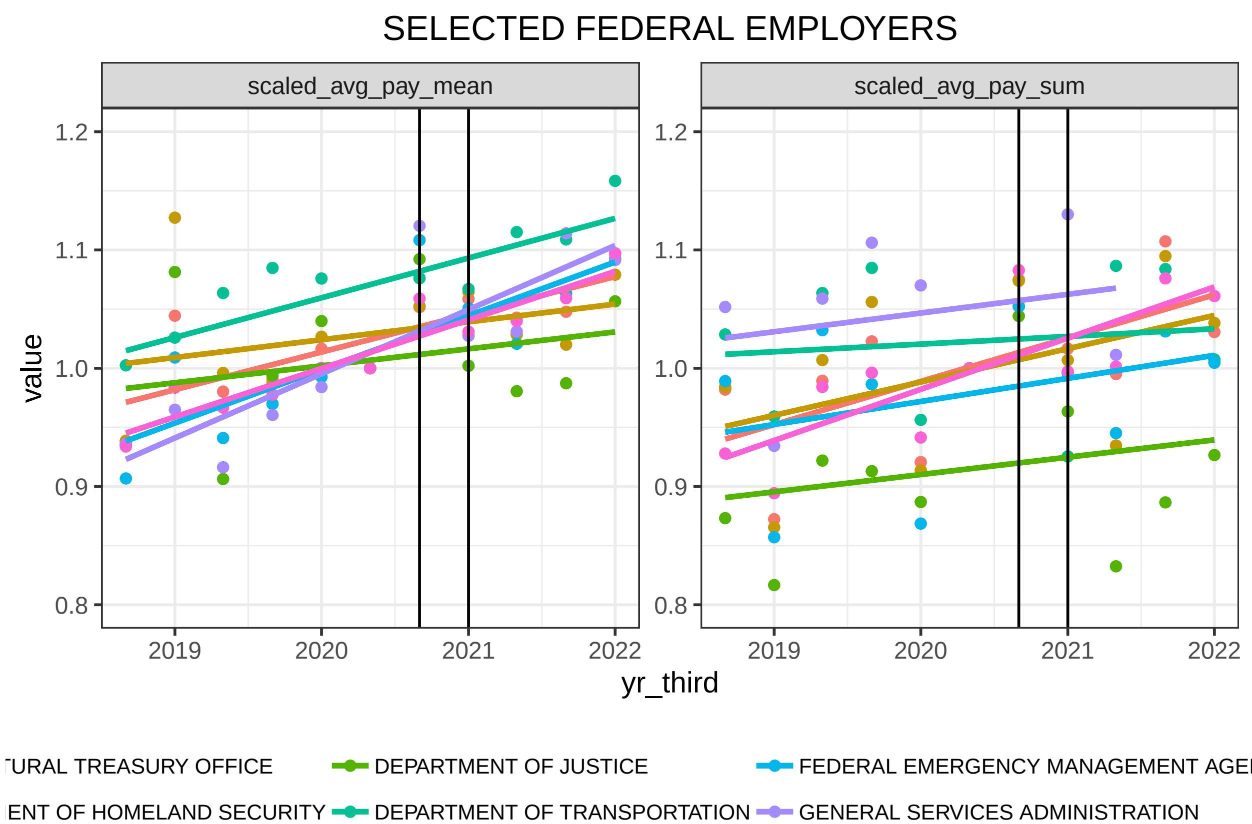
Federal

- Dropped Amtrak since it looks like they got paid by hours worked which was obviously off during the pandemic
- Still need to figure out why the pattern switches between DFAS and DFAS-NAVY or at least figure out if it corresponds to any significant economic event (reassignment, active duty pay, off-base housing etc.)

employer	num_indiv
DFAS - ARMY	4788
DFAS - AIR FORCE	4266
US TREASURY	1898
DFAS	1528
DFAS - NAVY	1186
UNITED STATES COAST GUARD	712
AGRICULTURAL TREASURY OFFICE	86
DEPARTMENT OF HOMELAND SECURITY	59
GENERAL SERVICES ADMINISTRATION	35
DEPARTMENT OF TRANSPORTATION	19
DEPARTMENT OF JUSTICE	17
FEDERAL EMERGENCY MANAGEMENT AGENCY	17
UNITED STATES SECRET SERVICE	17
DEPARTMENT OF THE INTERIOR	12
U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES	5
BONNEVILLE POWER ADMINISTRATION	1
DEPARTMENT OF HEALTH AND HUMAN SERVICES	1
DEPARTMENT OF STATE	1

- You get different results if you use per period total pay or if you use average paycheck size. Would like your thoughts?





- Why is the GSA so off in sums but not in means?
- Removed reservists. The logic being that it's not a regular income, not a large fraction of income anyways.
- Unsure why service branch is mentioned in some but not others. Would be concerning if the change had something to do with retirement.

State (Higher Ed and Traditional)

- Hopefully fewer false positives (manually inspected each string to see what was included).
- Allows for further investigation of within-employer variation along the lines of some other previous work
- Still probably below the actual fraction of state employees because university/corrections/hospital all are likely paid directly from their subsidiary. Unless you really want to be thorough and search for these sub-departments, we will probably have to be fine with missing them. At least we can be fairly sure that the basic payroll strings are most likely to capture our regular office workers.
- Initially, I was worried because we had the states showing up way out of proportion to their population. Now it seems much more reasonable. PA, OH, FL, NY, CA are all near the top.
- Some of these universities capture entire systems (e.g. University of Illinois is probably capturing University of Illinois in Chicago), but I would need to double-check that.
- I'd love to know why the coverage in some states is so bad. Is it due to gaps in regional coverage in the Yodlee data? Is it variation in how states handle payroll (departments vs. centralized system)? Somebody may ask why Maryland is higher than California in our number of observations? Of course, the selection in Yodlee only matters if it's correlated with member observables or outcome variables. My only concern would be that state employees might be incentivized to work with a local bank (eg.

Illinois State Employees Credit Union) that doesn't have a contract with Yodlee. I believe Yodlee has some details on type of institutions, so we could check that at a later date.

employer	num_indiv
MD	88
PA	74
OH	40
CA	29
NY	20
WA	20
FL	18
MA	17
CT	15
IN	15
WV	15
MT	10
NV	8
AZ	7
AL	6
AR	6
NC	5
UT	5
ID	4
MN	4
VT	4
LA	3
MO	3
KS	2
NH	2
WY	1

employer	num_indiv
Penn State University	50
University Of Michigan	42
University Of Illinois	22
University Of Colorado	19
University Of Kentucky	15
University Of California	13
University Of North Carolina	12
University Of Delaware	11
University Of Florida	10
University Of Texas	7
University Of Texas Southwestern	7
Michigan State University	5
Bowling Green State University	4
Texas Tech University	4
University Of Arkansas	4
University Of Utah	4
Kent State University	3
University Of Alabama	3
University Of Arizona	3
Arizona State University	2
Nc State University	2
Northern Arizona University	2
Ohio State University	2
University Of Nebraska	2
University Of Wisconsin	2
Florida A&m University	1
Florida State University	1
Indiana University	1
Montana State University	1
Texas A&m University	1
University Of Massachusetts	1
University Of Missouri	1
University Of South Carolina	1
University System Of New Hampshire	1
Utah State University	1
Wayne State University	1
Weber State University	1

Local (City/County/Education)

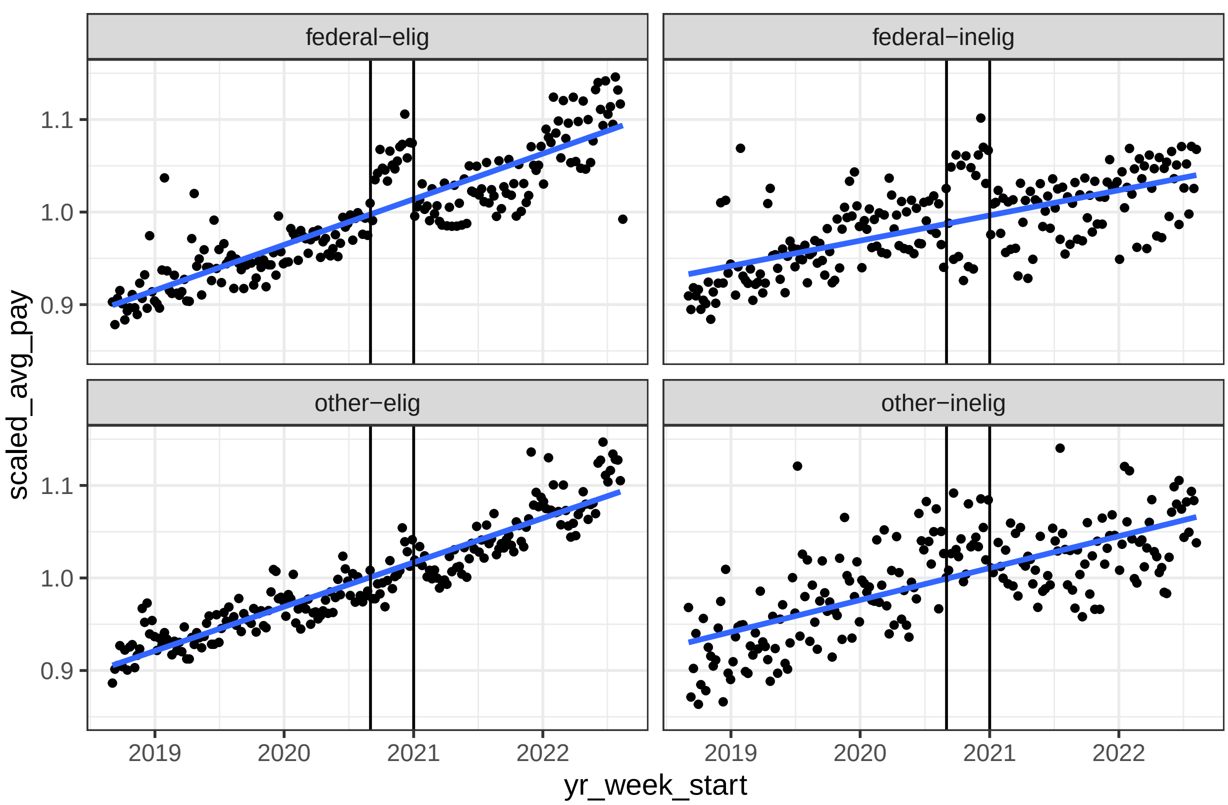
We have 263 number of local non-education employers and 15 number of local non-education employers which is not a substantial change from before. I went throughout the list to see how many of the top 100 cities we had, and I would estimate we have 60% of the top 100 cities in the US here. Not all employees survive the filters as we discussed earlier – likely due to variable pay/hours worked/etc.

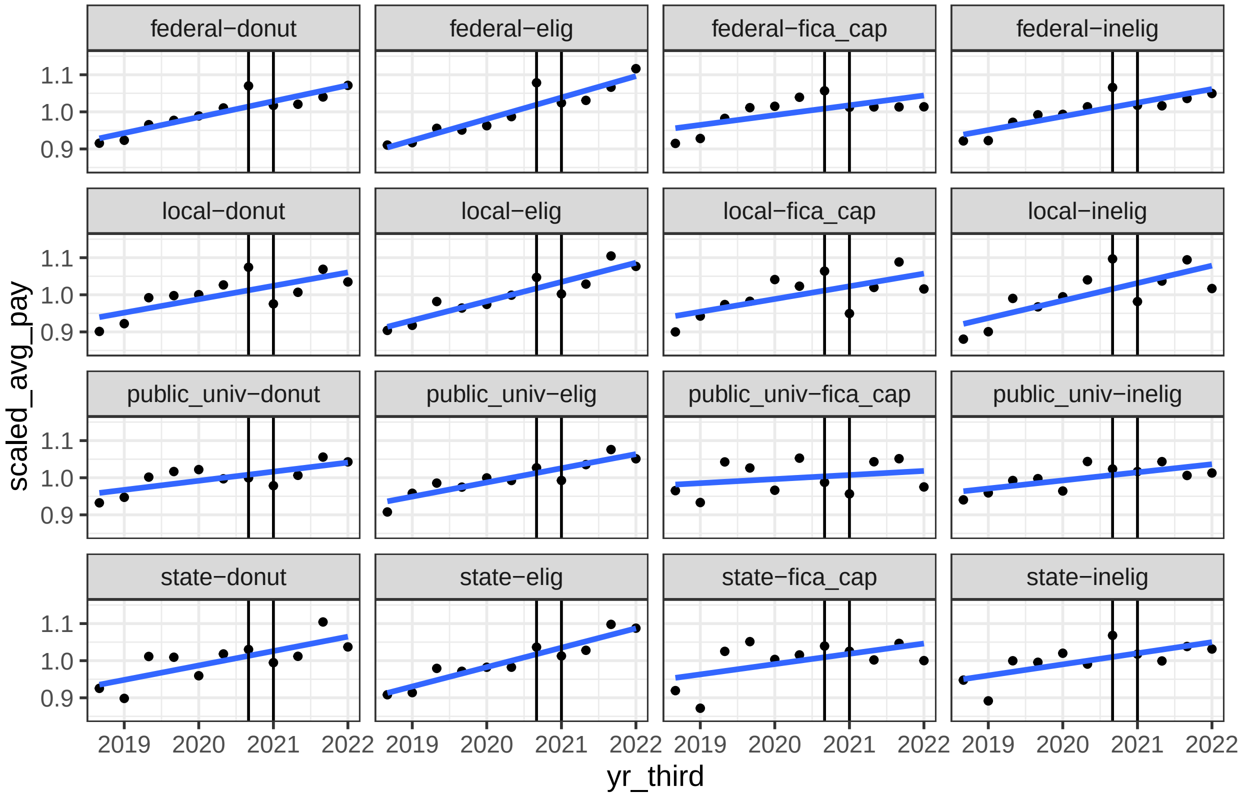
employer_type	employer	num_indiv
local	City Of New York	31
local	Montgomery County Government	29
local	Allegheny County	18
local	Anne Arundel County Board	17
local	City Of Phila	17
local	City Of Cincinnati	13
local	Cook County	13
local	Howard County Government	13
local	City Of Chicago	12
local	City Of Los Angeles	11
local	Prince William County	11
local	Baltimore County	9
local	City Of Columbus	9
local	Dekalb County	9
local	Harris County	9
local	City Of Cleveland	8
local	Arlington County	7
local	City Of Alexandria	7
local	City of Houston	7
local	County Of Orange	7
local	Cuyahoga County	7
local	City Of Dallas	6
local	City Of Garland	6
local	City of Phoenix	6
local	City Of Seattle	6
local	County Of Riverside	6
local	Franklin County	6
local	Palm Beach County Sheriff's Office	6
local	Bexar County	5
local	City of Charlotte	5
local	City Of Springfield	5
local	County Of San Diego	5
local	San Bernardino County	5
local	Charles County Government	4
local	City Of Milwaukee	4
local	City of San Diego	4
local	City of Tacoma	4
local	County Of Bucks	4
local	County Of Hamilton	4
local	County Of Tulare	4
local	King County	4
local	Palm Beach County	4
local	Allen County Government	3
local	Autauga County	3
local	Broward County	3
local	City Of Benton	3
local	City Of Boston	3
local	City Of Colorado Springs	3
local	City Of Dayton	3
local	City Of El Paso	3
local	City Of Griffin	3
local	City Of Joliet	3
local	City Of Lakewood	3
local	City Of Mesa	3
local	City Of Newton	3
local	City Of Norfolk, Nebraska	3
local	City Of Orlando	3

employer_type	employer	num_indiv
local_educ	Loudoun County Public Schools	17
local_educ	School District Of Palm Beach County	17
local_educ	Howard County Public School	15
local_educ	Prince George's County Public Schools	11
local_educ	Fulton County Schools	7
local_educ	Gwinnett County Board Of Education	7
local_educ	Hillsborough County Public Schools	7
local_educ	Broward County Public Schools	6
local_educ	Cobb County School	6
local_educ	Pinellas County Schools	6
local_educ	Charles County Public Schools, Elementary Schools	5
local_educ	St Mary's County Public Schools	4
local_educ	The San Joaquin County Office Of Education	3
local_educ	Collier County Public School	1
local_educ	Rankin County School District	1

Private Employers

You floated the idea of including large public employers as an additional control group, so I wanted to give you a bit of info on that option. Large private employers seem to have very spotty coverage as measured by *primary_merchant_name*. Large employers like Cisco, Microsoft, Apple, and others are included in the Yodlee data, but we would have to do a rigorous search of large private employers and potentially do some re-balancing to make them reflect the broader population.





- Federal ineligible also experience a decent pay raise

Updated Waterfall

Recall the definitions. Please forgive me for the numbering errors last time. I really hate that there are now multiple documents with different numbering and definitions, so please don't reference across documents.

Definition of Federal Worker/Eligibilty

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 2019 - Dec 2022. (This is Yodlee's suggested value for a 'stable' user.)
2. Has qualifying payment observations (based on description, primary_merchant, and amount fields).[1]
3. Single Classification (i.e. only ever receive a paycheck from federal employers, state employers, or local employers but never a combination of categories).

[1]transaction must be greater than $200, from an identified vendor, marked as transaction_base_type = 'credit' and transaction_category_name in ('Salary/Regular Income', 'Investment/Retirement Income', 'Deposits', 'Education', 'Services/Supplies', 'Other Income', 'Utilities') by Yodlee and not marked as a duplicate transaction

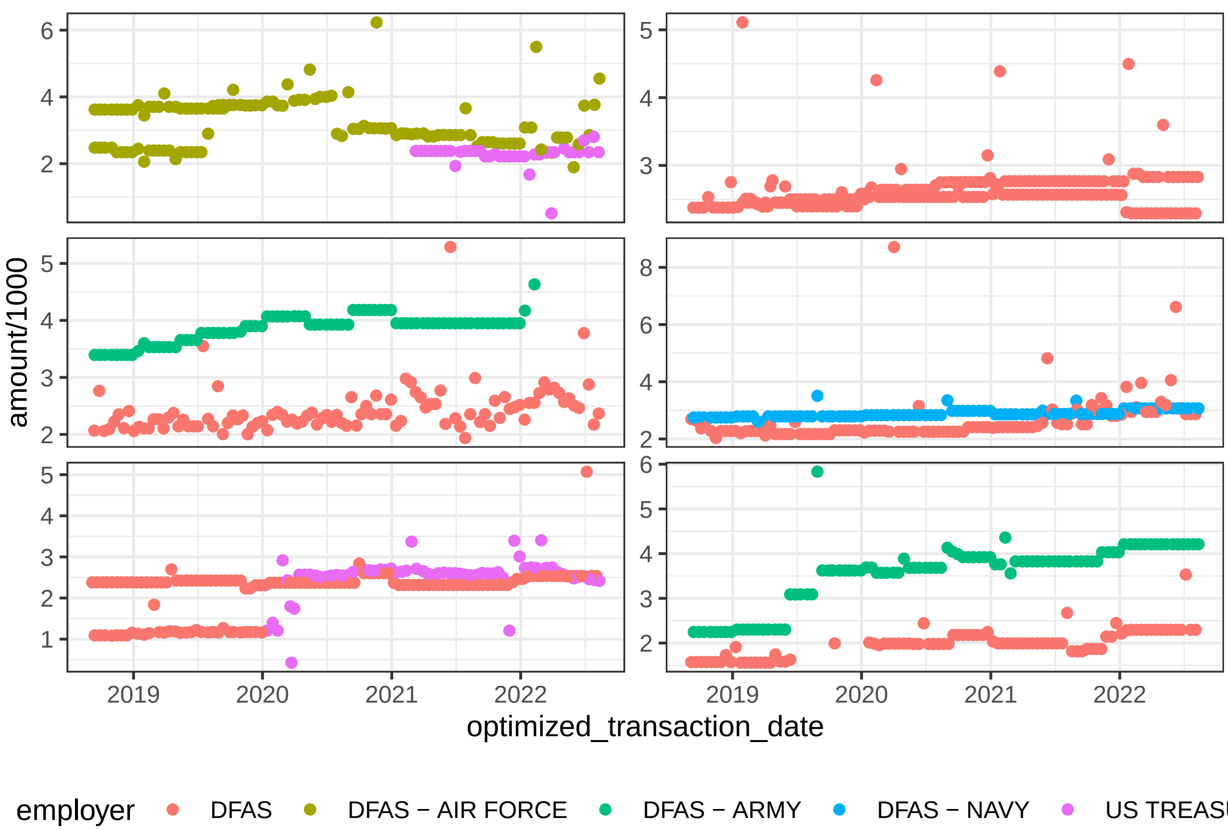
4. No more than 1 employer within a category, excluding the cases where DFAS/US TREASURY switched patterns.(Note: This was done to allow for idiosyncrasies in the description string patterns and Yodlee's determination of the primary_merchant.)
5. At least 30 qualifying paychecks over the sample period.
6. Earn at least 10k in each 4 month period throughout the sample period.
7. A median paycheck frequency between 7 and 35 days (i.e. most paychecks come at a regular interval)
8. Observe no more than 40% of total inflows from other sources of income (outside income = Gov Income + Interest Income + Other Income + Deposits + Insurance + Investment/Retirement Income + Taxes + Salary/Regular Income + Sales/Services Income)
9. Observe no more than 35% volatility between paychecks to rule out employees with varying hours worked,travel reimbursements,etc.

type	elig	f2	class	emp	chqs	pds10k	freq	gov	vol
exempt_federal	<30k	47887	44750	44750	4503	0	0	0	0
exempt_federal	elig	6162	4640	4640	4556	1168	1129	699	638
exempt_federal	donut	281	110	110	110	63	48	27	20
exempt_federal	inelig	192	64	64	63	39	27	15	12
exempt_federal	fica	136	23	23	23	11	4	3	1
exempt_federal	**total**	**54658**	**49587**	**49587**	**9255**	**1281**	**1208**	**744**	**671**
federal	<30k	445450	422062	419807	98894	0	0	0	0
federal	elig	213473	198044	193225	188667	84989	83403	56487	41647
federal	donut	45215	42013	40689	40430	26070	25180	18229	13375
federal	inelig	35968	33099	31833	31581	21391	20046	14903	10230
federal	fica	40179	35892	33624	33161	23461	19186	15265	7995
federal	**total**	**780285**	**731110**	**719178**	**392733**	**155911**	**147815**	**104884**	**73247**
local	<30k	165778	152802	152661	20234	0	0	0	0
local	elig	31410	23465	23388	21830	6309	6173	3424	2951
local	donut	3256	1764	1755	1706	822	754	474	364
local	inelig	2350	1029	1023	992	486	417	287	205
local	fica	2746	940	922	823	371	277	193	133
local	**total**	**205540**	**180000**	**179749**	**45585**	**7988**	**7621**	**4378**	**3653**
local_educ	<30k	21395	19700	19700	3436	0	0	0	0
local_educ	elig	4643	3617	3615	3339	1001	980	549	502
local_educ	donut	328	104	104	102	51	35	27	24
local_educ	inelig	275	50	50	50	26	9	7	7
local_educ	fica	415	41	41	40	25	2	2	2
local_educ	**total**	**27056**	**23512**	**23510**	**6967**	**1103**	**1026**	**585**	**535**
public_univ	<30k	130551	119261	119082	11694	0	0	0	0
public_univ	elig	15353	10305	10215	7561	1860	1802	1114	932
public_univ	donut	2086	986	972	783	350	318	214	154
public_univ	inelig	1725	710	697	539	237	203	132	83
public_univ	fica	3311	1465	1442	1170	537	444	378	182
public_univ	**total**	**153026**	**132727**	**132408**	**21747**	**2984**	**2767**	**1838**	**1351**
state	<30k	79028	68969	68969	10468	0	0	0	0
state	elig	16331	10653	10652	9971	3288	3208	1896	1732
state	donut	1640	715	715	694	391	359	222	157
state	inelig	1248	438	438	413	242	208	141	81
state	fica	1446	408	408	380	211	158	120	61
state	**total**	**99693**	**81183**	**81182**	**21926**	**4132**	**3933**	**2379**	**2031**

You asked to see whether these people with pay-cycles of one week are actually two earners. Judging by these plots, it's definitely possible, especially when there are clearly two paycheck amounts. It could also be

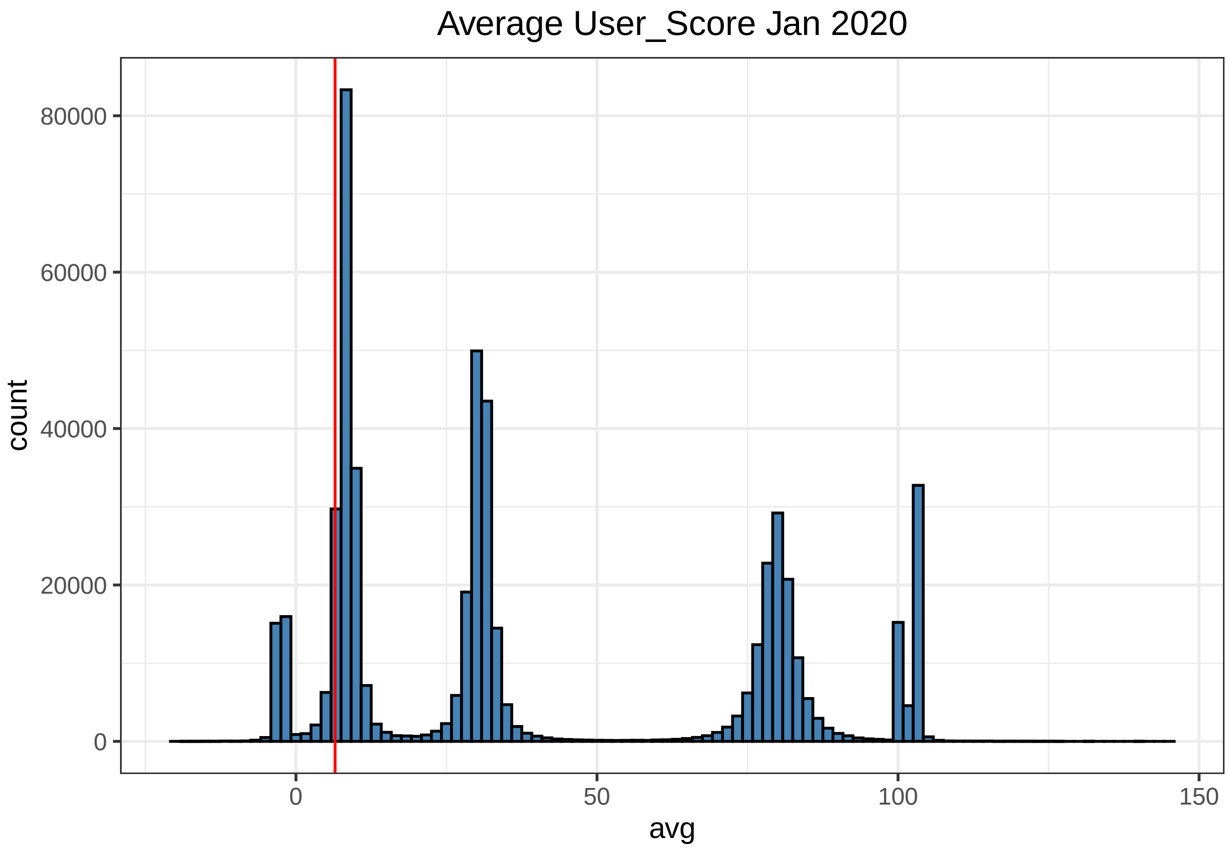
elig	f2	class	emp	chqs	pds10k	freq	gov	vol
<30k	890089	827544	824969	149229	0	0	0	0
elig	287372	250724	245735	235924	98615	96695	64169	48402
donut	52806	45692	44345	43825	27747	26694	19193	14094
inelig	41758	35390	34105	33638	22421	20910	15485	10618
fica	48233	38769	36460	35597	24616	20071	15961	8374
total	**1320258**	**1198119**	**1185614**	**498213**	**173399**	**164370**	**114808**	**81488**

a housing allowance or other types of compensation. Would love your opinion. Regardless, it doesn't seem to change the regression results (not a honest reason for keeping them but still. . .)

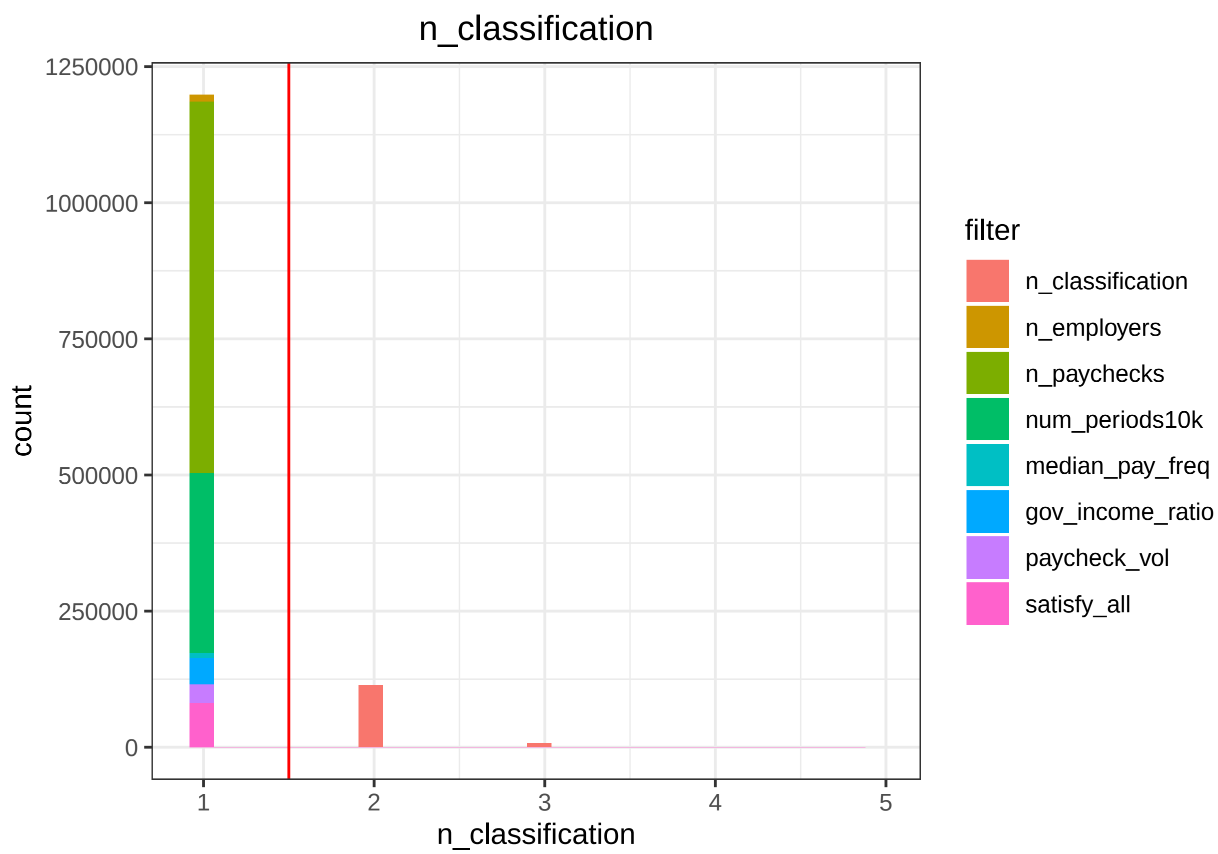


Some thoughts and key highlights:

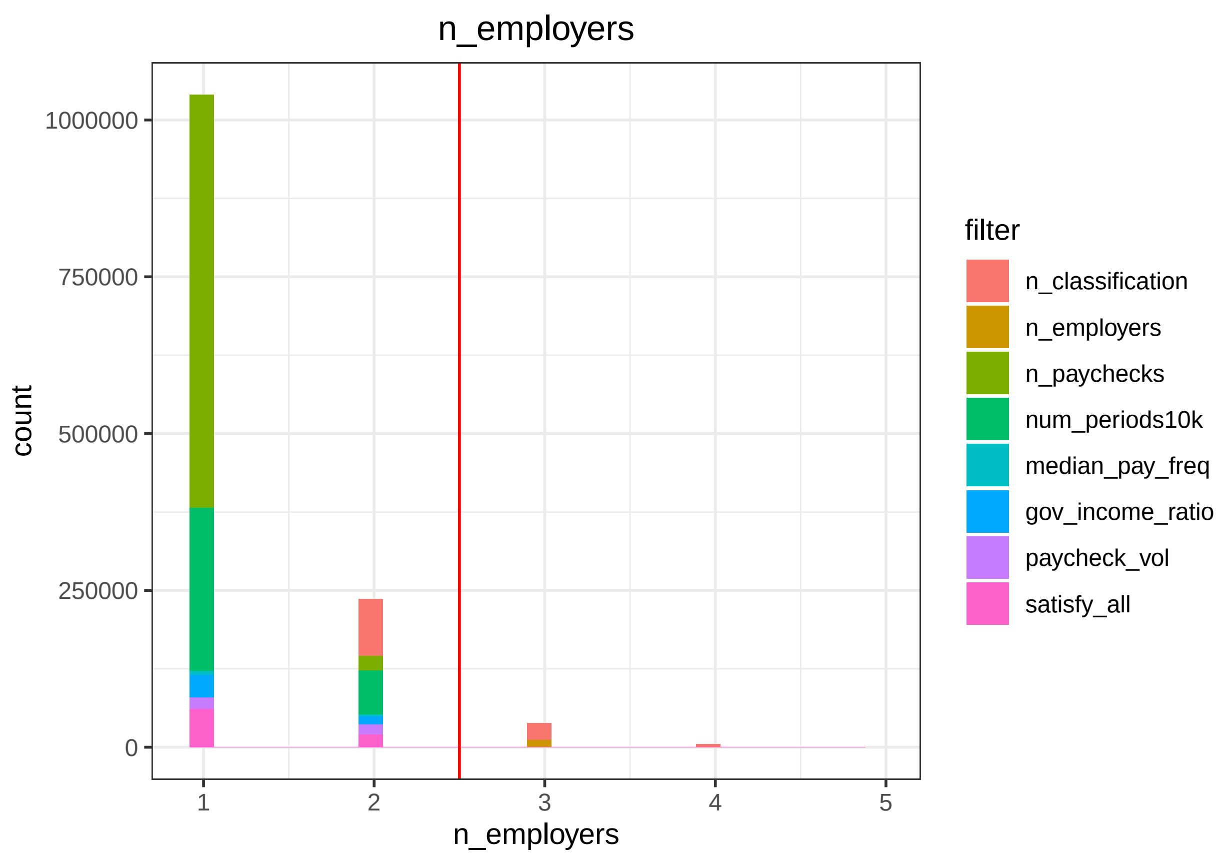
- Changed definition of outside income (see next section). Yodlee misclassifies a small amount of transactions, but it's enough to matter. The difference is quite apparent in the regression analysis to follow.
- Dropped filter for credit card linkage (we were losing too many users), and it's not necessary for the first stage anyways. When we get to regressions involving credit card usage, we can just restrict the sample then.
- The definition of regular intervals changed a bit.


Average User_Score Jan 2020
count
80000
60000
40000
20000
0
0
50
100
150
avg

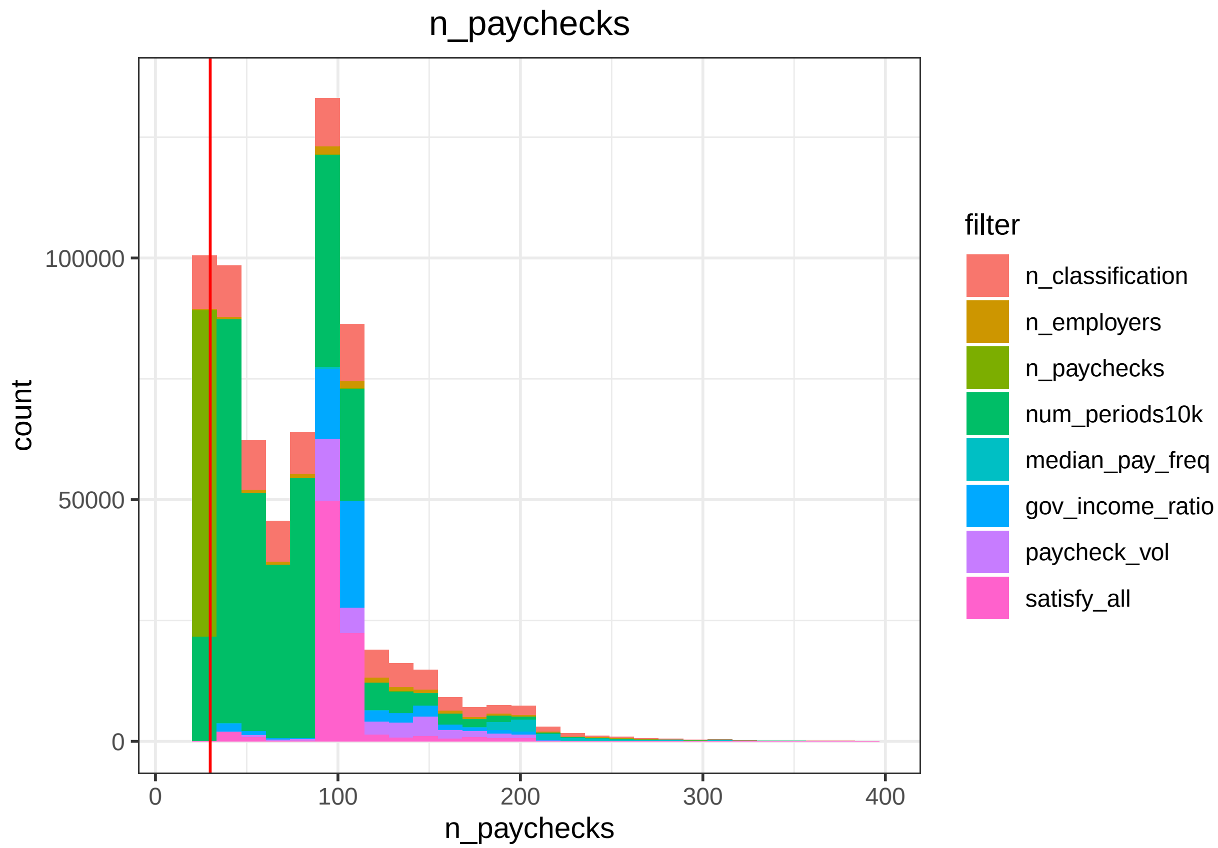
```
## [[1]]
```


n_classification
1250000
1000000
750000
500000
250000
0
count
1
2
3
4
5
n_classification
filter
n_classification
n_employers
n_paychecks
num_periods10k
median_pay_freq
gov_income_ratio
paycheck_vol
satisfy_all

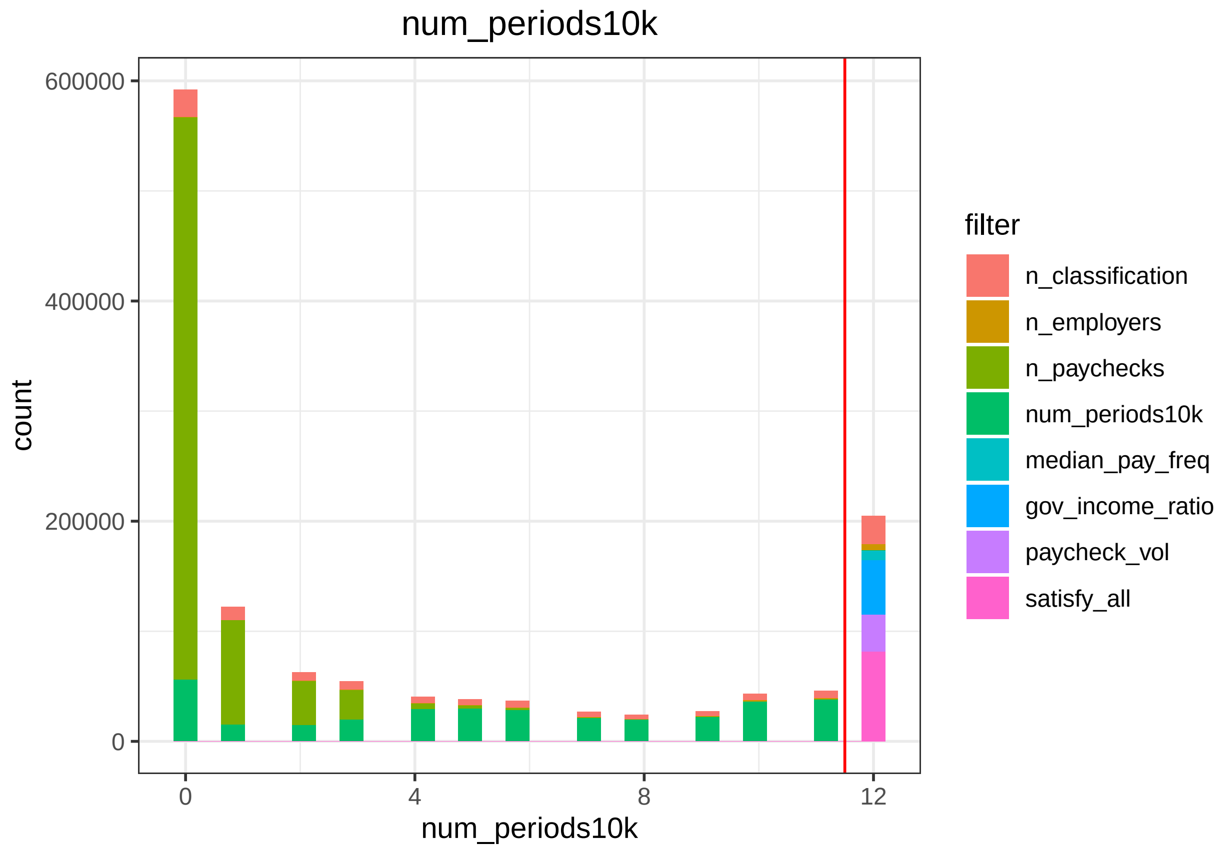
```
##
## [[2]]
```


n_employers
count
1000000
750000
500000
250000
0
1
2
3
4
5
n_employers
filter
n_classification
n_employers
n_paychecks
num_periods10k
median_pay_freq
gov_income_ratio
paycheck_vol
satisfy_all

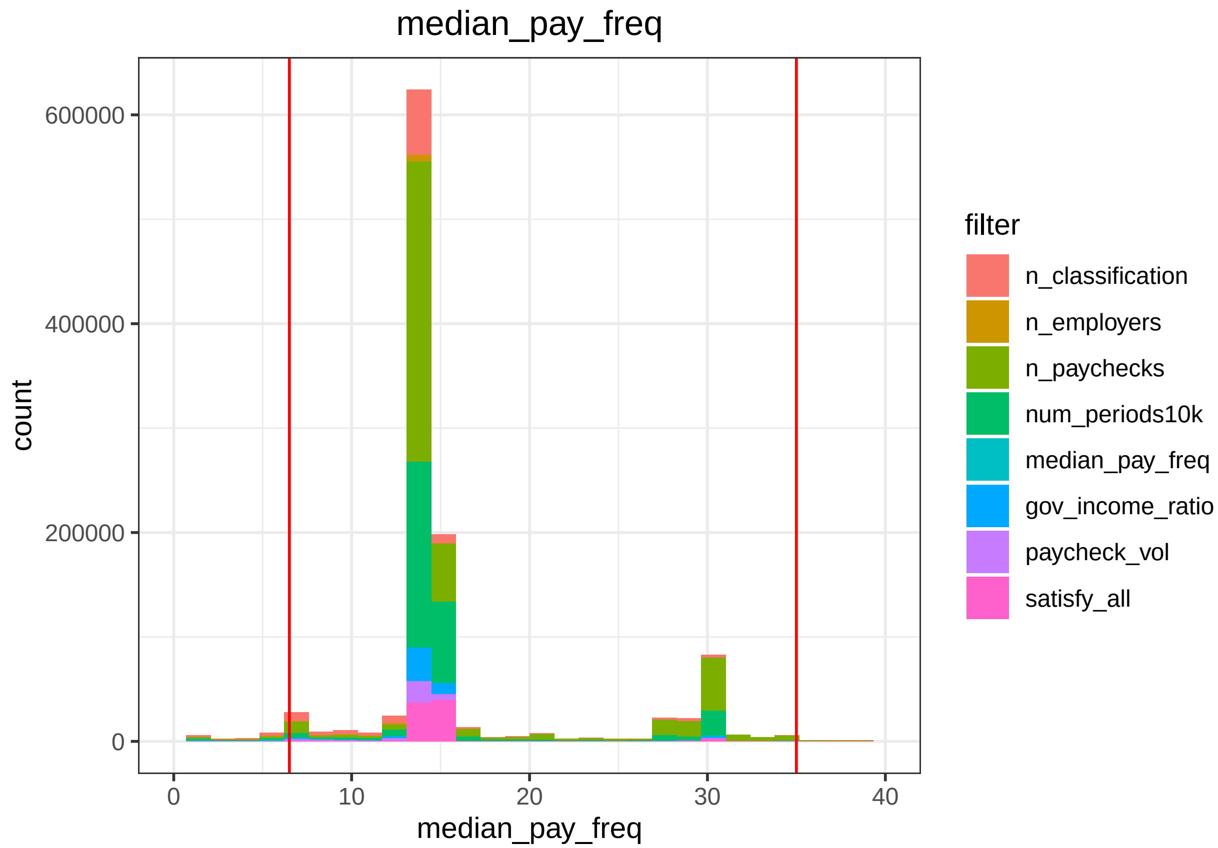
```
##
## [[3]]
```


n_paychecks
count
100000
50000
0
0
100
200
300
400
n_paychecks
filter
n_classification
n_employers
n_paychecks
num_periods10k
median_pay_freq
gov_income_ratio
paycheck_vol
satisfy_all

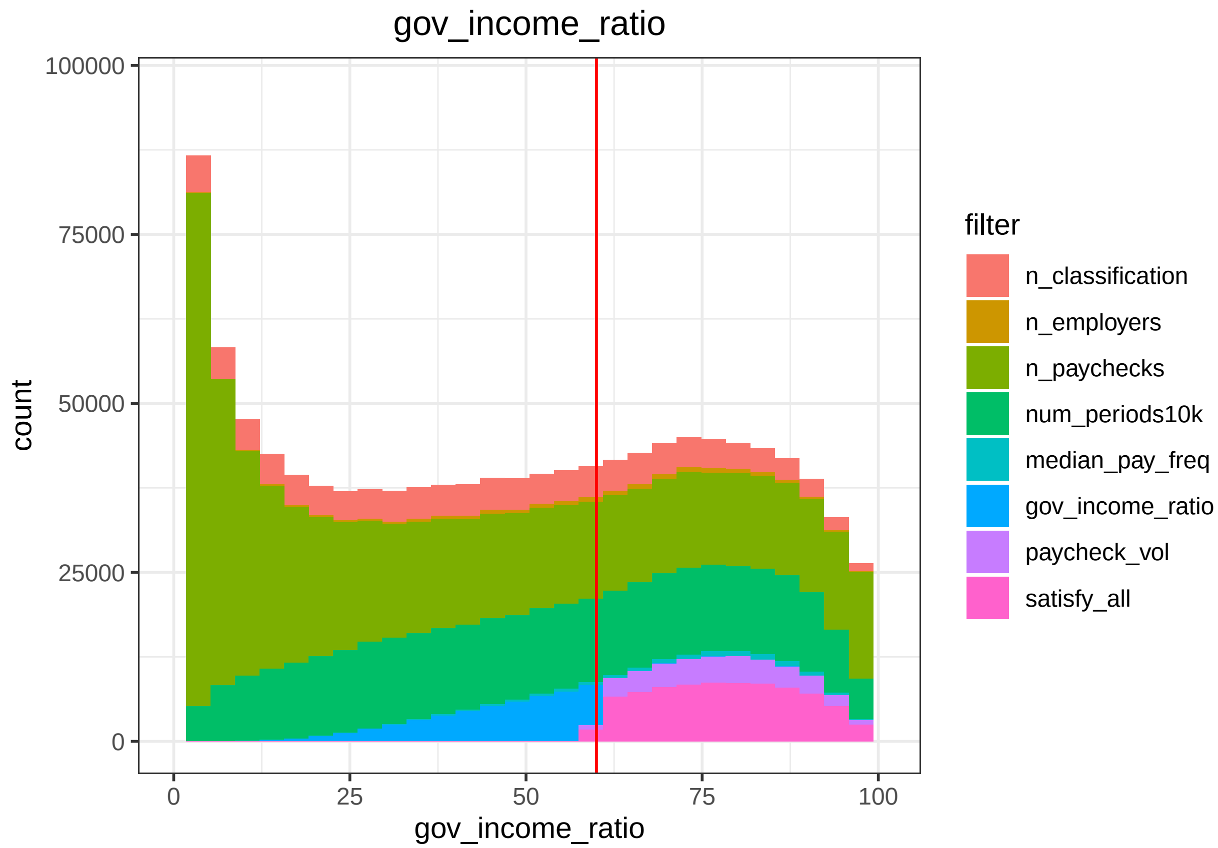
```
##
## [[4]]
```


num_periods10k
600000
400000
200000
0
count
0
4
8
12
num_periods10k
filter
n_classification
n_employers
n_paychecks
num_periods10k
median_pay_freq
gov_income_ratio
paycheck_vol
satisfy_all

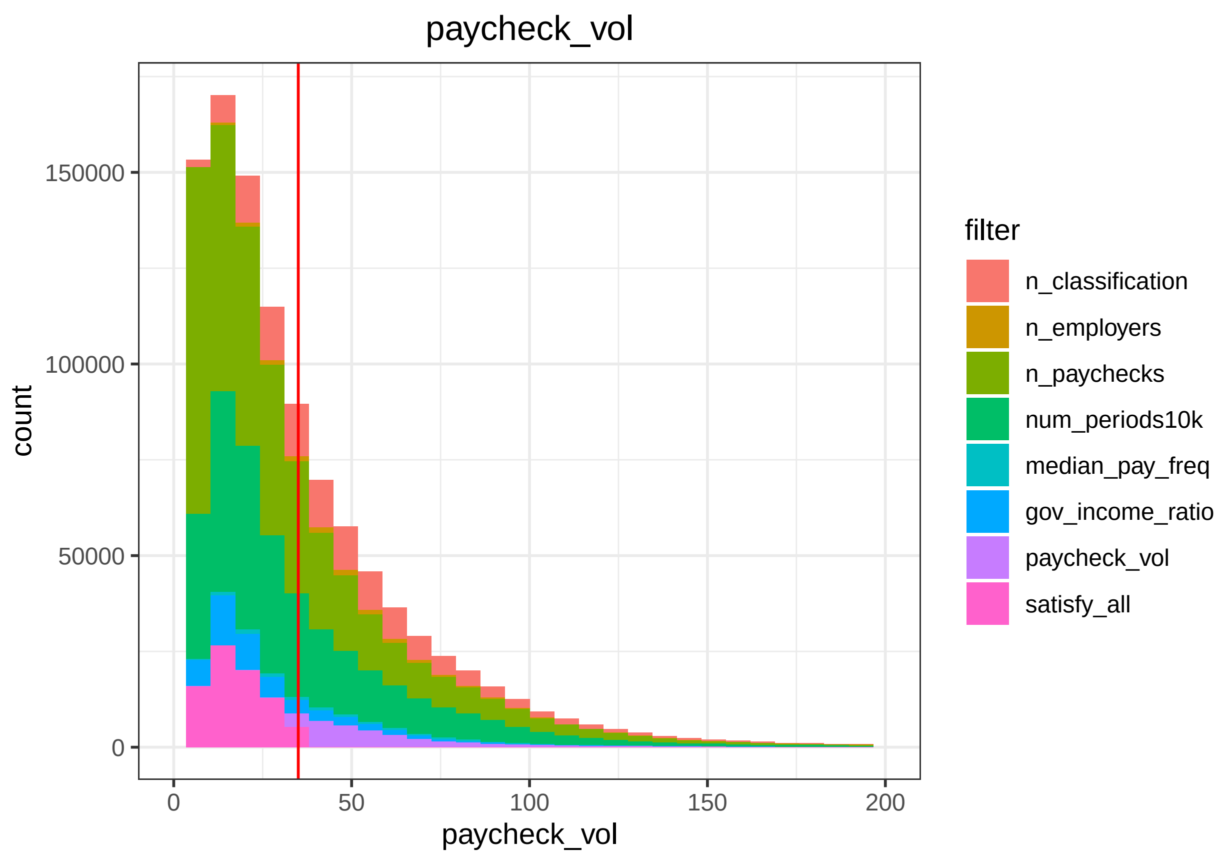
```
##
## [[5]]
```

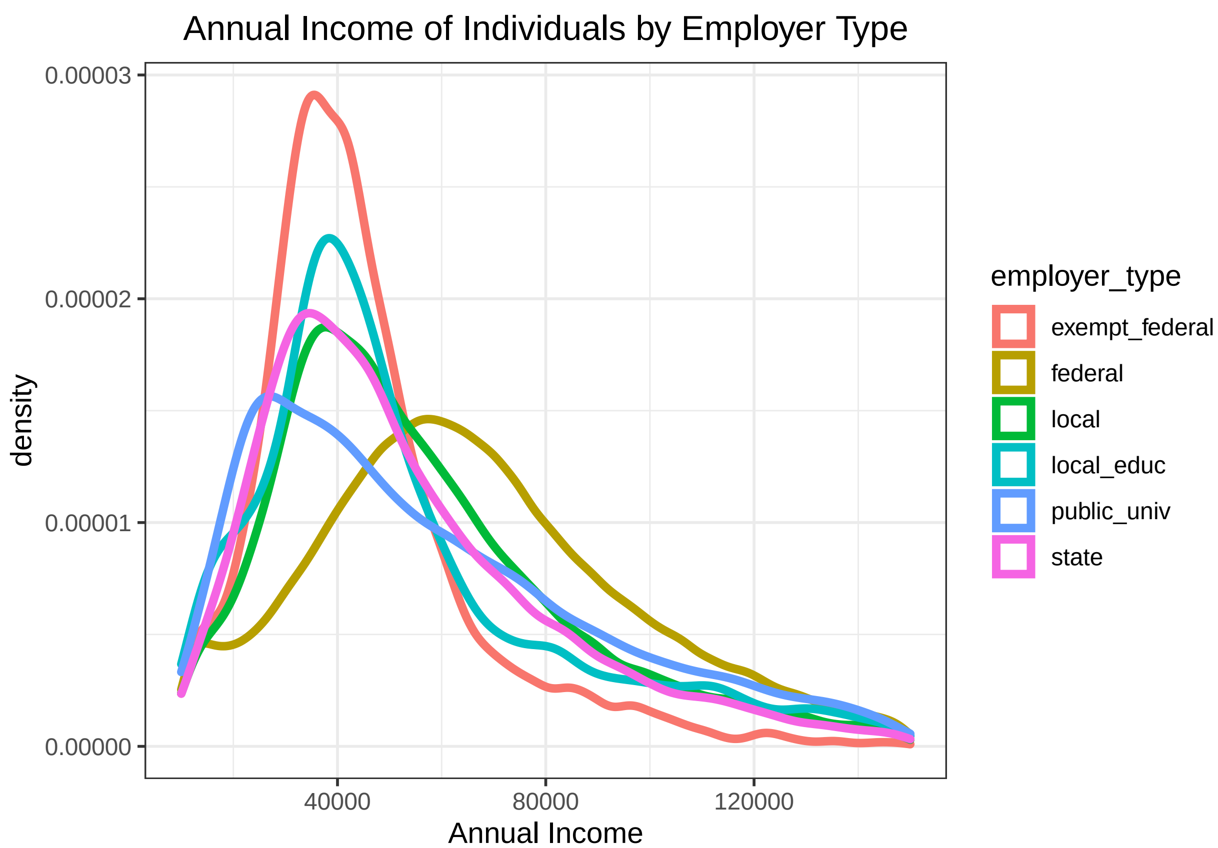

median_pay_freq
600000
400000
200000
0
count
0
10
20
30
40
median_pay_freq
filter
n_classification
n_employers
n_paychecks
num_periods10k
median_pay_freq
gov_income_ratio
paycheck_vol
satisfy_all

```
##
## [[6]]
```


gov_income_ratio
100000
75000
50000
25000
0
count
0
25
50
75
100
gov_income_ratio
filter
n_classification
n_employers
n_paychecks
num_periods10k
median_pay_freq
gov_income_ratio
paycheck_vol
satisfy_all

```
##
## [[7]]
```


paycheck_vol
count
150000
100000
50000
0
0
50
100
150
200
paycheck_vol
filter
n_classification
n_employers
n_paychecks
num_periods10k
median_pay_freq
gov_income_ratio
paycheck_vol
satisfy_all



Change to Outside Income Definition

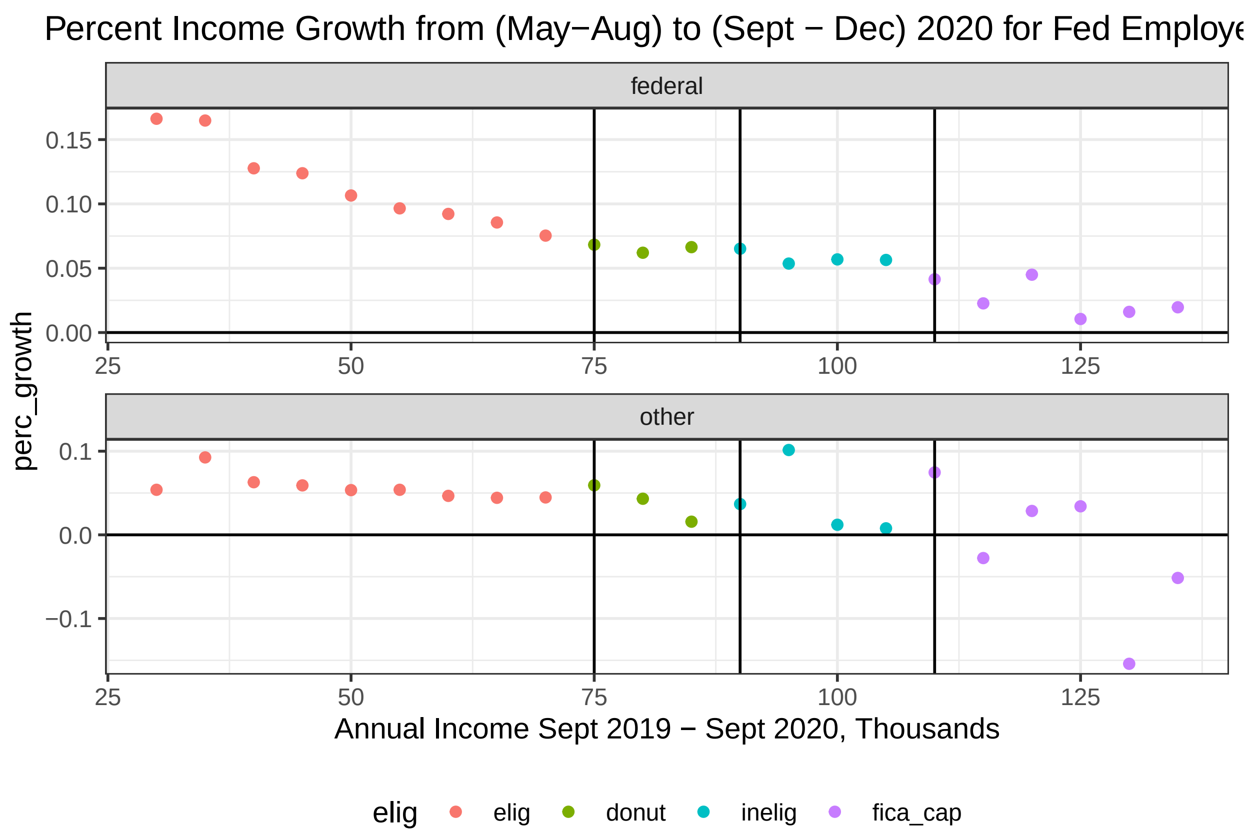
I have come to the realization that the Salary/Regular Income is not a very good definition of salary. Plenty of transactions that clearly fit into the approved strings are classified into Deposits, Insurance, or Other Income. I have expanded the search for qualifying strings, AND I have broadened the scope of categories for outside income while lowering the threshold to no more than 40% of income coming from outside sources. Outside Income = Interest Income + Other Income + Deposits + Insurance + Investment/Retirement Income + Taxes + Salary/Regular Income + Sales/Services Income

$$\frac{GovIncome}{OutsideIncome} > 60$$

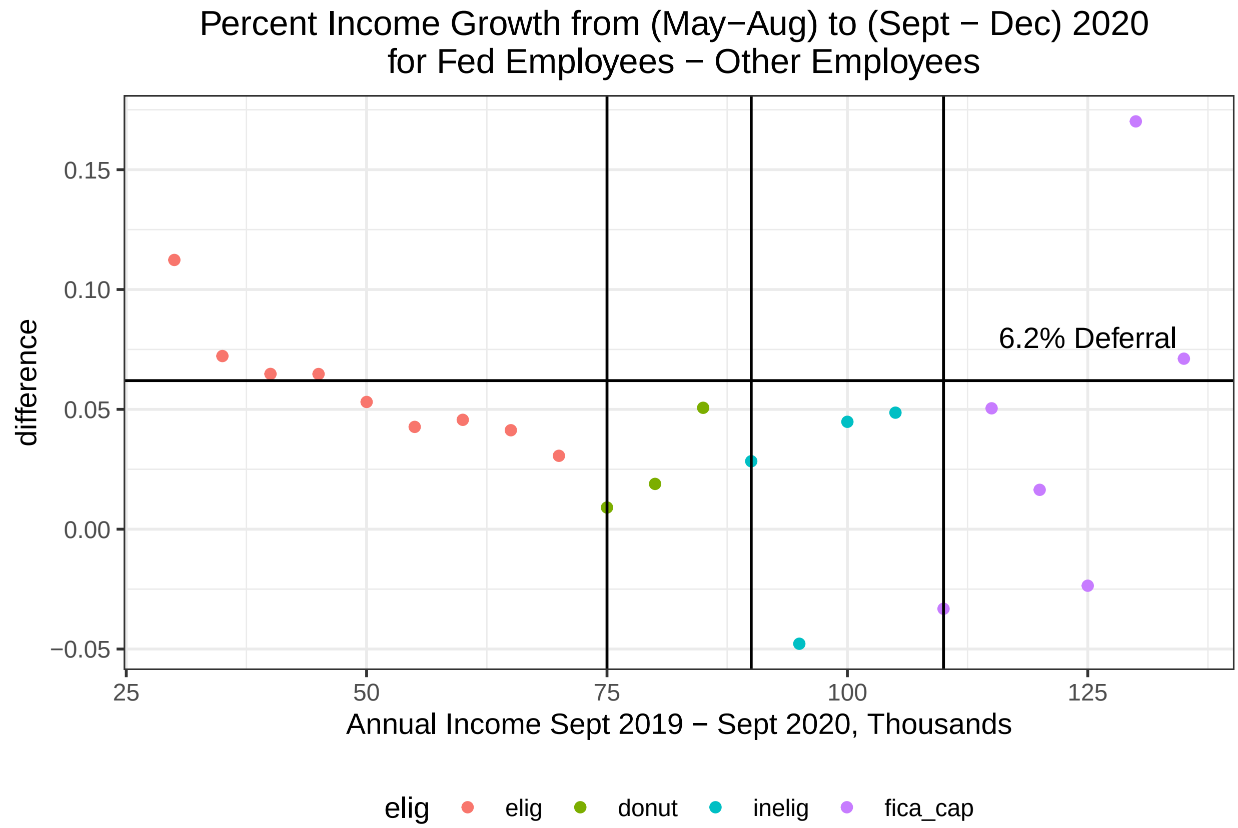
Examining the Waterfall Cutoffs

Examining the Income Thresholds

What happened toward the lower end of the federal spectrum?



The next graph shows the difference between federal and other (exempt fed/state/higher ed/local/local ed) employees

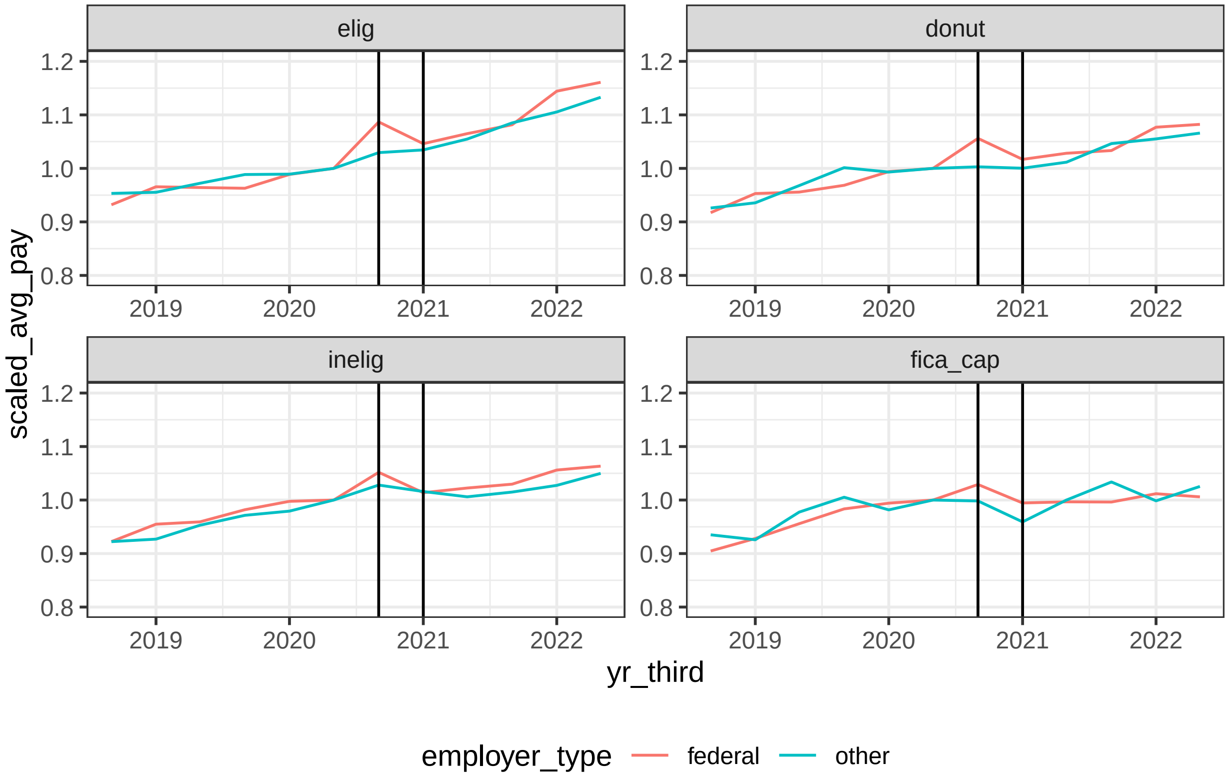


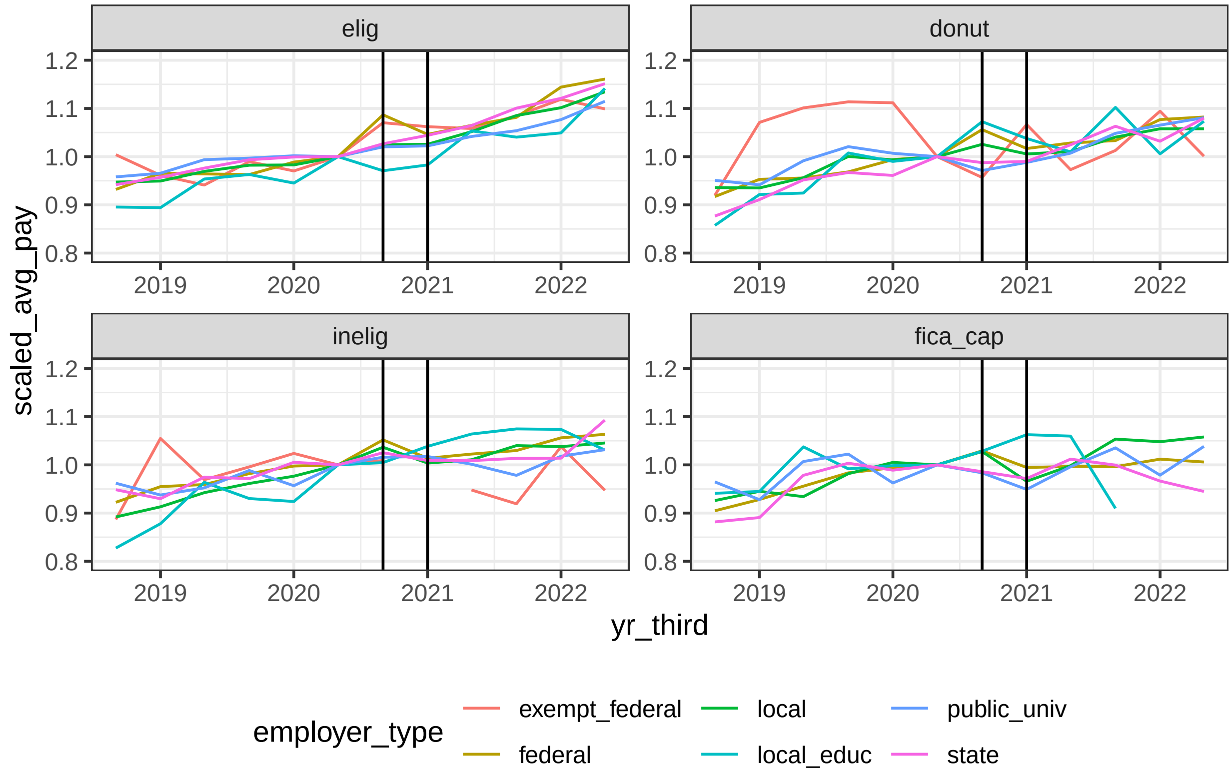
I don't think USPS is a good control group. There is **a lot** of variability, probably due to seasonal variation, variation in hours worked, and holiday pay.

New First Stage & Parallel Trends

We are primarily concerned with making sure there are parallel trends for the *elig* and *inelig* groups. It would be nice to see parallel trends for the donut and fica group as well. The time interval for this graph is 4 month chunks. The values plotted are weekly average pay during the 4 month chunks (excluding 0s). All groups are indexed to 1 at '2020-05-01' so that that the next value can be interpreted as a percentage increase in average weekly pay. The *federal-elig* group is exactly where we expected it to be with an increase of 7% during the deferral period and an approximately 3% decrease during the next 4 month period.

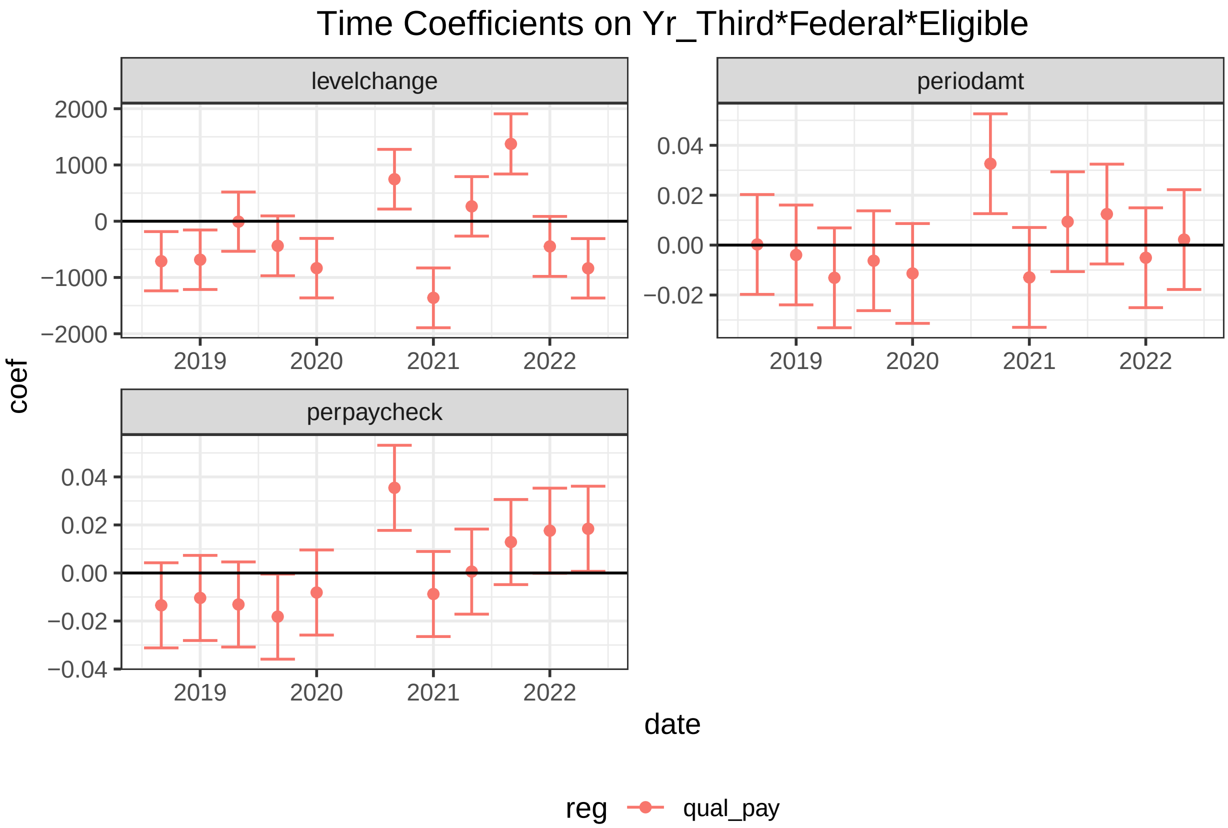
- I'm a bit surprised to see that the downturn in the *federal-elig* is so small. I guess enough people got raises at the same time the repayment was extended so that the average weekly payment only reverted to trend instead of going below it.
- I think it looks pretty good overall. In the second graph, it's clear that the education groups are driving some of the fluctuations in the trends.

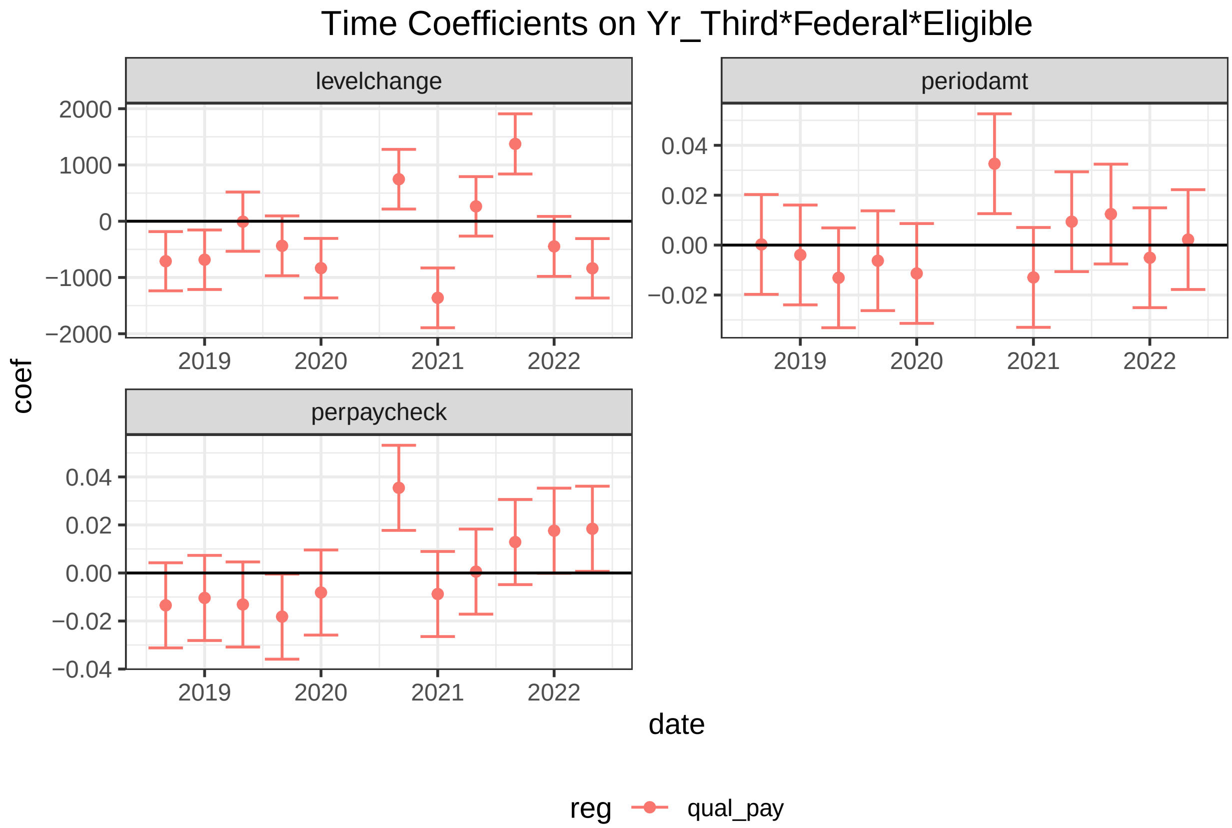

elig
donut
inelig
fica_cap
1.2
1.1
1.0
0.9
0.8
2019
2020
2021
2022
scaled_avg_pay
yr_third
employer_type
federal
other

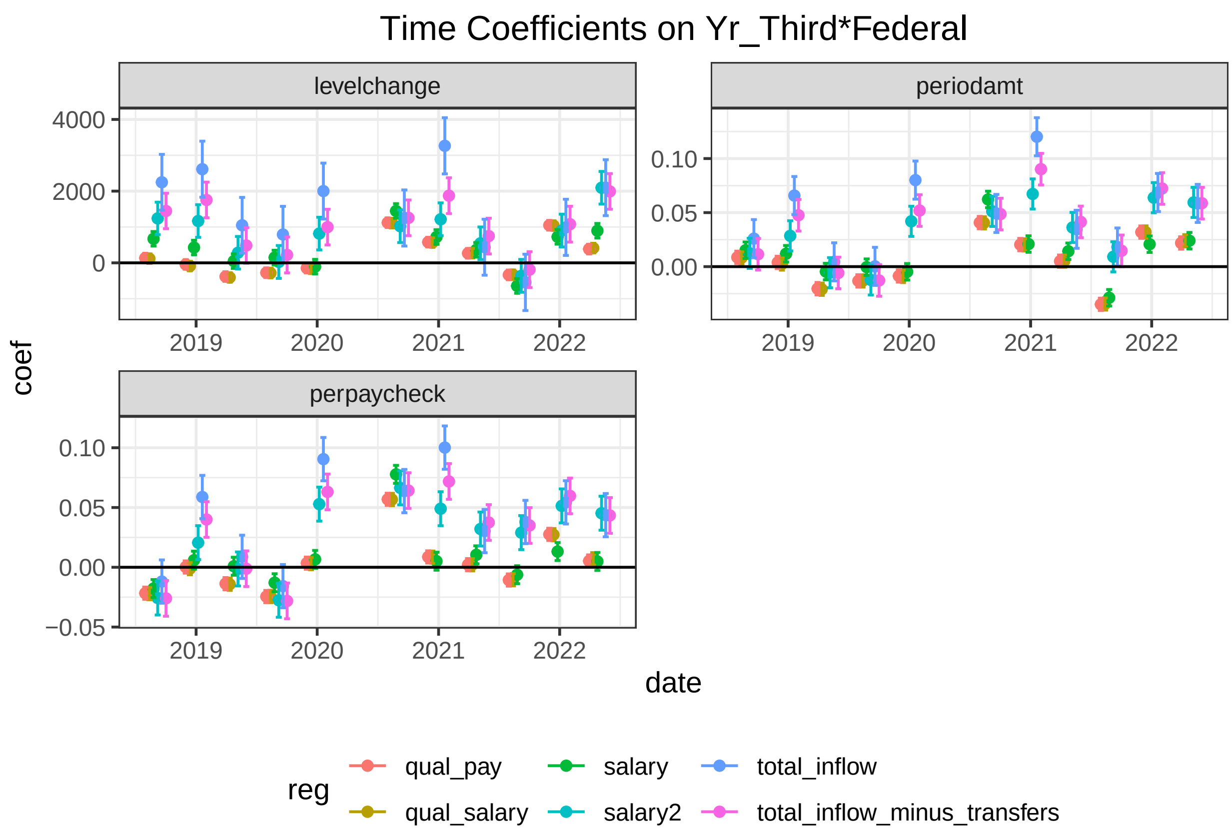


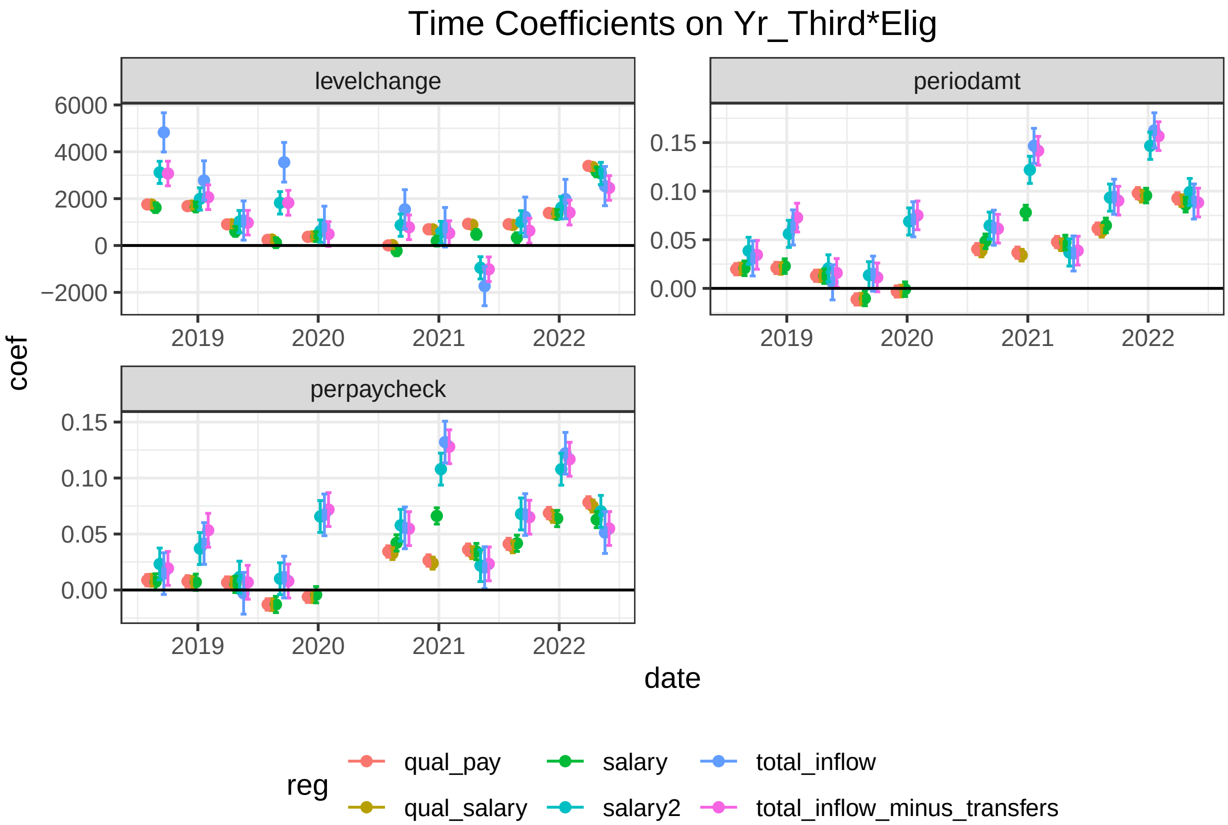
Regressions

All of these are without the USPS (federal_exempt) employees.


Time Coefficients on Yr_Third*Federal*Eligible
levelchange
periodamt
perpaycheck
coef
date
2019
2020
2021
2022
2000
1000
0
−1000
−2000
0.04
0.02
0.00
−0.02
−0.04
reg
qual_pay


Time Coefficients on Yr_Third*Federal*Eligible
levelchange
periodamt
perpaycheck
2000
1000
0
−1000
−2000
0.04
0.02
0.00
−0.02
−0.04
2019
2020
2021
2022
coef
date
reg
qual_pay

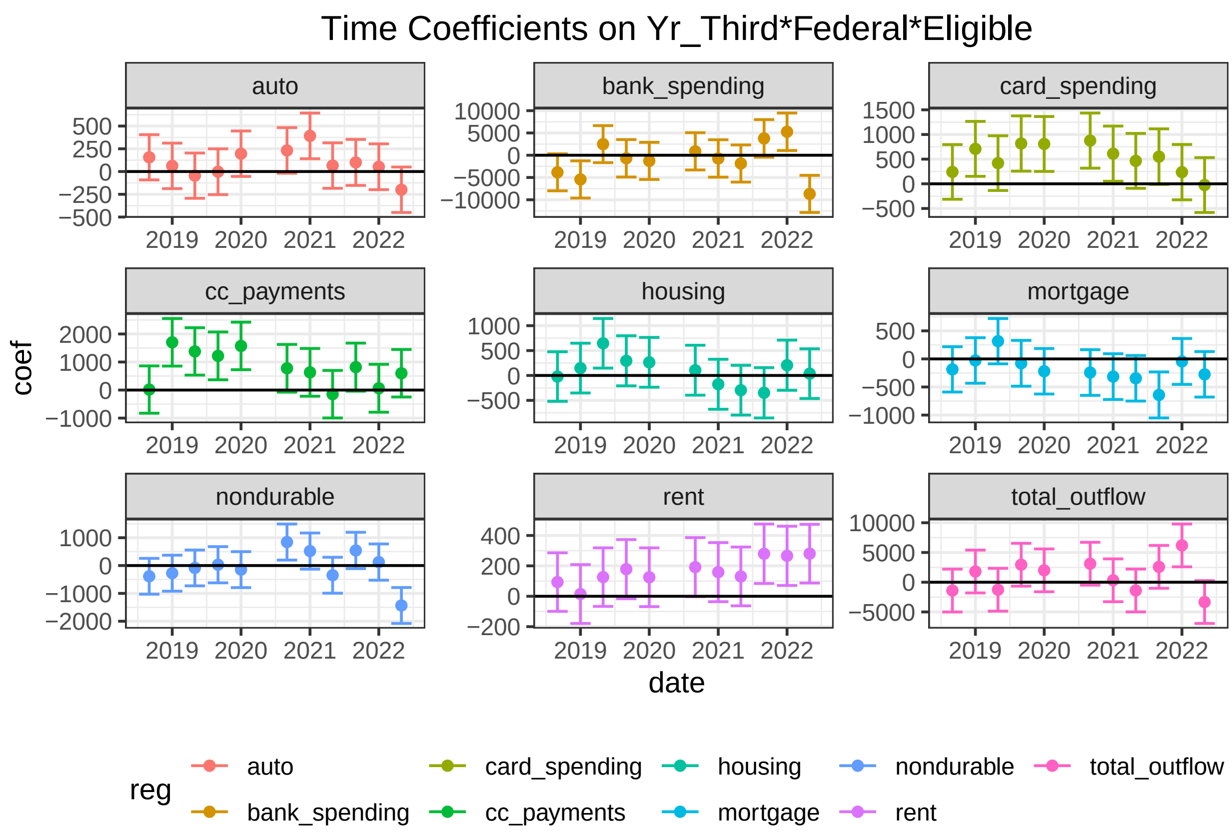

Time Coefficients on Yr_Third*Federal
levelchange
periodamt
perpaycheck
coef
date
2019
2020
2021
2022
4000
2000
0
0.10
0.05
0.00
−0.05
reg
qual_pay
qual_salary
salary
salary2
total_inflow
total_inflow_minus_transfers



Played around with all sorts of different versions of expenditures (percent changes, levels, etc.) ... they all seem to have very wide confidence intervals. Tell me more about how you think winsorizing would work in this situation.

transaction_category_name	qual_amount	amount
Salary/Regular Income	84.493550902917	37.3907799
Investment/Retirement Income	9.74669404492665	6.3001995
Deposits	3.72351779796079	18.6513508
Services/Supplies	0.727266350664436	1.6201425
Education	0.671632232035301	0.2765123
Other Income	0.36196299164184	4.5121244
Utilities	0.275375679853985	0.1231712
Service Charges/Fees		0.0135192
Taxes		1.4818162
Charitable Giving		0.0439299
Electronics/General Merchandise		0.1796566
Securities Trades		0.3310600
Subscriptions/Renewals		0.0019033
Savings		0.1356426
Restaurants		0.0599539
Loans		0.0001079
Refunds/Adjustments		1.3136227
Entertainment/Recreation		0.0409277
Mortgage		0.0011862
Retirement Contributions		0.0020371
Personal/Family		0.0981740
		0.0000263
Transfers		26.4526918
Sales/Services Income		0.1146222
Credit Card Payments		0.0093850
Other Expenses		0.0539100
Home Improvement		0.0624709
Healthcare/Medical		0.0683548
Pets/Pet Care		0.0023116
Automotive/Fuel		0.0571280
Interest Income		0.0441276
Postage/Shipping		0.0142860
ATM/Cash Withdrawals		0.0002562
Travel		0.0752875
Office Expenses		0.0146950
Rent		0.0021977
Insurance		0.0501269
Groceries		0.0506843
Gifts		0.0016310
Rewards		0.0566103
Expense Reimbursement		0.2759296
Cable/Satellite/Telecom		0.0154496

transaction_category_name	**amount**
Transfers	31.7863415
Credit Card Payments	17.1908917
Check Payment	8.6930743
Services/Supplies	5.0894952
Mortgage	4.3486523
Loans	4.0098404
Other Expenses	3.8471709
Electronics/General Merchandise	3.5575245
Securities Trades	2.3752663
Groceries	2.2744358
Restaurants	2.2189115
Personal/Family	1.7814128
Automotive/Fuel	1.4144218
Cable/Satellite/Telecom	1.3795776
ATM/Cash Withdrawals	1.3163581
Taxes	1.2787174
Utilities	1.1933567
Insurance	1.1523130
Entertainment/Recreation	0.8047283
Home Improvement	0.7941705
Rent	0.7745800
Travel	0.7262816
Healthcare/Medical	0.5056640
Education	0.3053068
Service Charges/Fees	0.2720946
Savings	0.2325871
Charitable Giving	0.1978518
Pets/Pet Care	0.1978067
Postage/Shipping	0.1226028
Subscriptions/Renewals	0.0564820
Office Expenses	0.0528943
Gifts	0.0361581
Retirement Contributions	0.0127198
Salary/Regular Income	0.0003057
	0.0000034
Deposits	0.0000006
Rewards	0.0000001
Investment/Retirement Income	0.0000000



Covariate Analysis

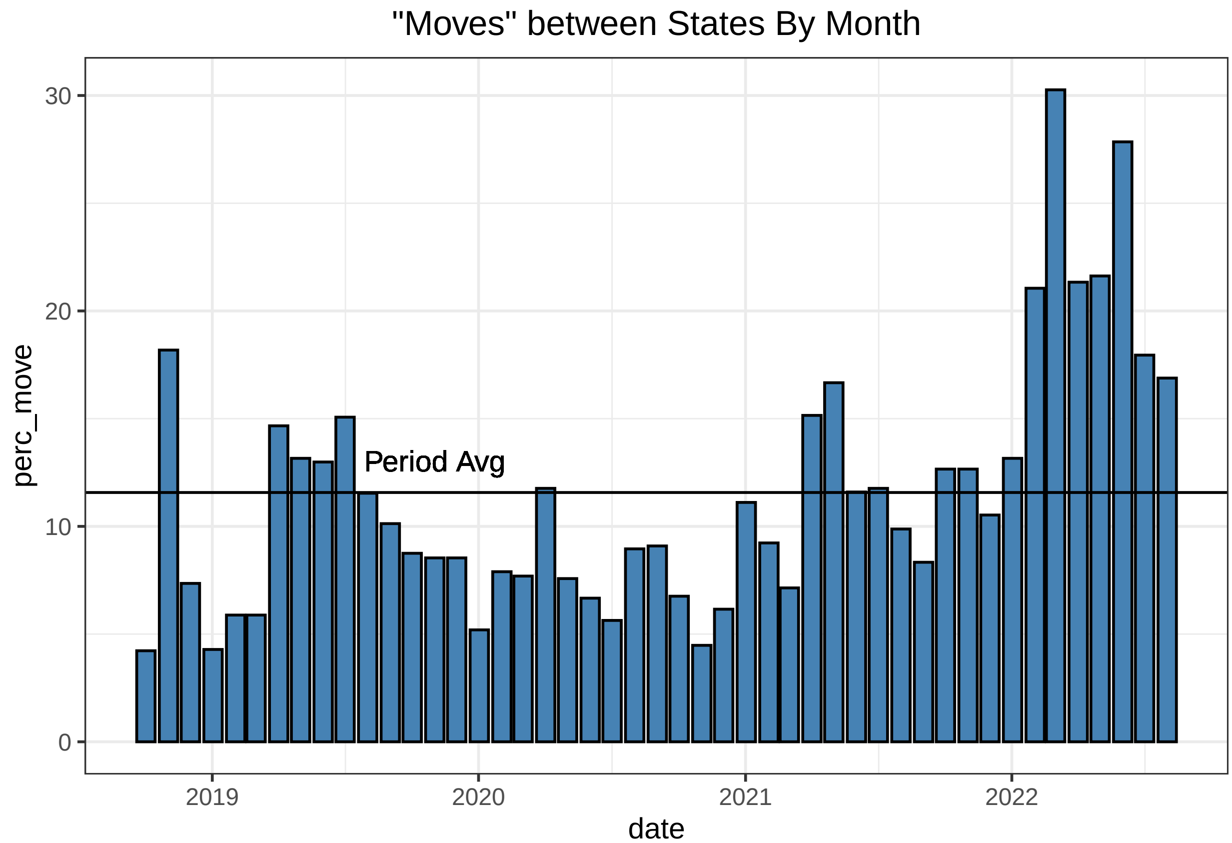
Mobility

After receiving feedback from Scott on the measure of mobility, I've updated the definitions to:

- preperiod_geo: most common geography from 2020-01 to 2020-06 (inclusive).
- postperiod_geo: most common geography from 2021-06 to 2022-01 (inclusive).
- moved: (indicator) for preperiod_geo $\neq$ postperiod_geo

Hopefully, this cuts down on the noise in the Yodlee's geography measures.

ever_fed	elig	move_state	move_city	dc_area_pre
federal	donut	0.4175824	0.1586429	0.0496250
federal	elig	0.3770025	0.1671057	0.0268520
federal	fica	0.4205526	0.1550943	0.0846367
federal	inelig	0.4179514	0.1559895	0.0636063
other	donut	0.0809061	0.2443366	0.0145631
other	elig	0.0753390	0.2427591	0.0169094
other	fica	0.0983051	0.1762712	0.0033898
other	inelig	0.0845921	0.2326284	0.0060423



I created the previous graph to show just how volatile Yodlee's geographic assignments are. They have an overall moving rate of 11.571468 per month which is obviously way too high. It should be around 1-2% per month, and I'm not sure the dramatic increase in changes in geographies (i.e. "moves") matches actual moving data. At least the summer seasonal spike is present which matches reality. Also worth noting is that roughly 10% of the geo observations are not spaced one month apart. So beware of this fact when doing a month-over-month measure in the geography data.

- UI Payments, overdraft fees, balance snapshot, length of time in the dataset
- Worth linking to home price indices for those with mortgage payments (home prices were incredibly volatile during this time period) We could use percentile ranking of property tax payments in the same city as a proxy for home value since levy rate is usually stable within a city.
- How should we think about 401k contributions/retirement plans since we don't see most of the activity? We do not know how much they are contributing pre-tax, and we don't necessarily see what type of account money is transfered from if it is indeed transfered.
- How should we think about contemporaneous wealth effects due to retirement accounts/401k/housing? Not a problem econometrically unless there is endogeneity in location, home ownership, etc.
- We can probably come up with a way to identify refinances, maybe even distinguish between cash out and rate refi (i.e. no change of city in geo, no uhaul or mover transactions, significant change in mortgage amount (but still positive to avoid picking up defaulters),)
- Looked at CPS, SCF, and SIPP to see what we can benchmark our data too (will have more on that shortly)