explore

Hunter York

5/24/2021

Clean up place names where possible and extract US states if mentioned in place names

Step one is to try to see what sort of location data I can get out of this. As I mentioned in my email, these data are not yet geocoded. Below is a table of frequencies of state names in the data. The bulk of the data is not associated with a state.

I can extract US state names (nothing more granular) for about 5,600/16,000 menus. Even knowing that the currency is in dollars is hard to pin down, as currency data isn't complete for all menus. However, for menus where it explicitly mentions non-US dollars, I drop those cases.

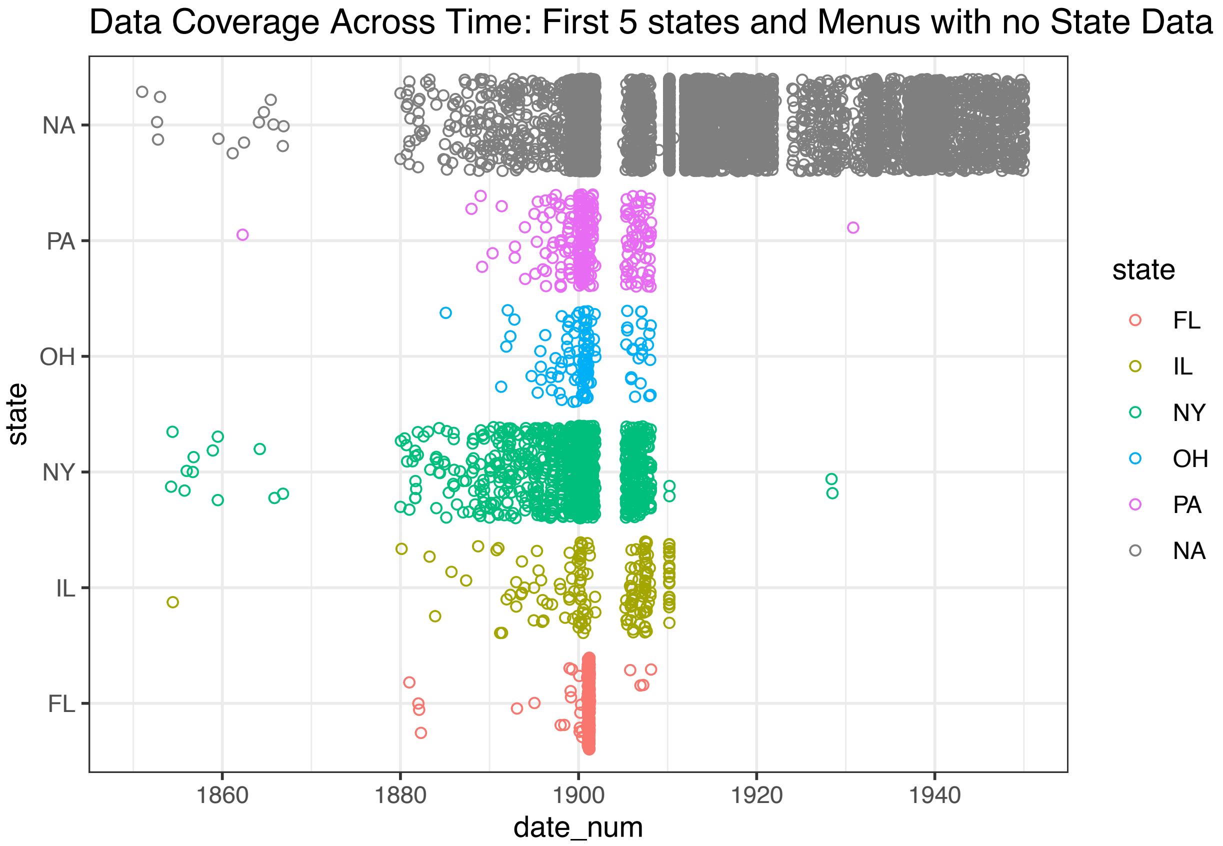
Table 1: Number of Menus by State

	state	N
1		11, 846
2	NY	1, 394
3	MA	655
4	OR	501
5	LA	361
6	UT	351
7	PA	305
8	FL	272
9	IL	164
10	RI	157
11	CA	140
12	GA	139
13	OH	124
14	NE	98
15	ND	83
16	NJ	74
17	AL	71
18	CO	64
19	DC	63
20	IN	62
21	ME	60
22	MO	58
23	MI	54
24	NC	54
25	IA	51
26	AR	48
27	VA	43
28	NH	31
29	HI	30
30	WI	27
31	WA	27
32	CT	21
33	MD	19
34	NM	13
35	TN	12
36	TX	12
37	MN	12
38	DE	10
39	AZ	8
40	VT	5
41	ID	4
42	SC	4
43	KY	4
44	MT	3
45	WV	3
46	OK	3
47	MS	2
48	AK	2
49	NV	1

Data artifacts

Seeing as this data is still not finished, I suspect the gaps in temporal coverage are due to the order in which the menus are being digitized (perhaps the catalog is discontinuous in some nonrandom way across time). Regardless, there are some glaring data quality problems, namely large missing gaps across time.

```
## Warning: Removed 1841 rows containing missing values (geom_point).
```



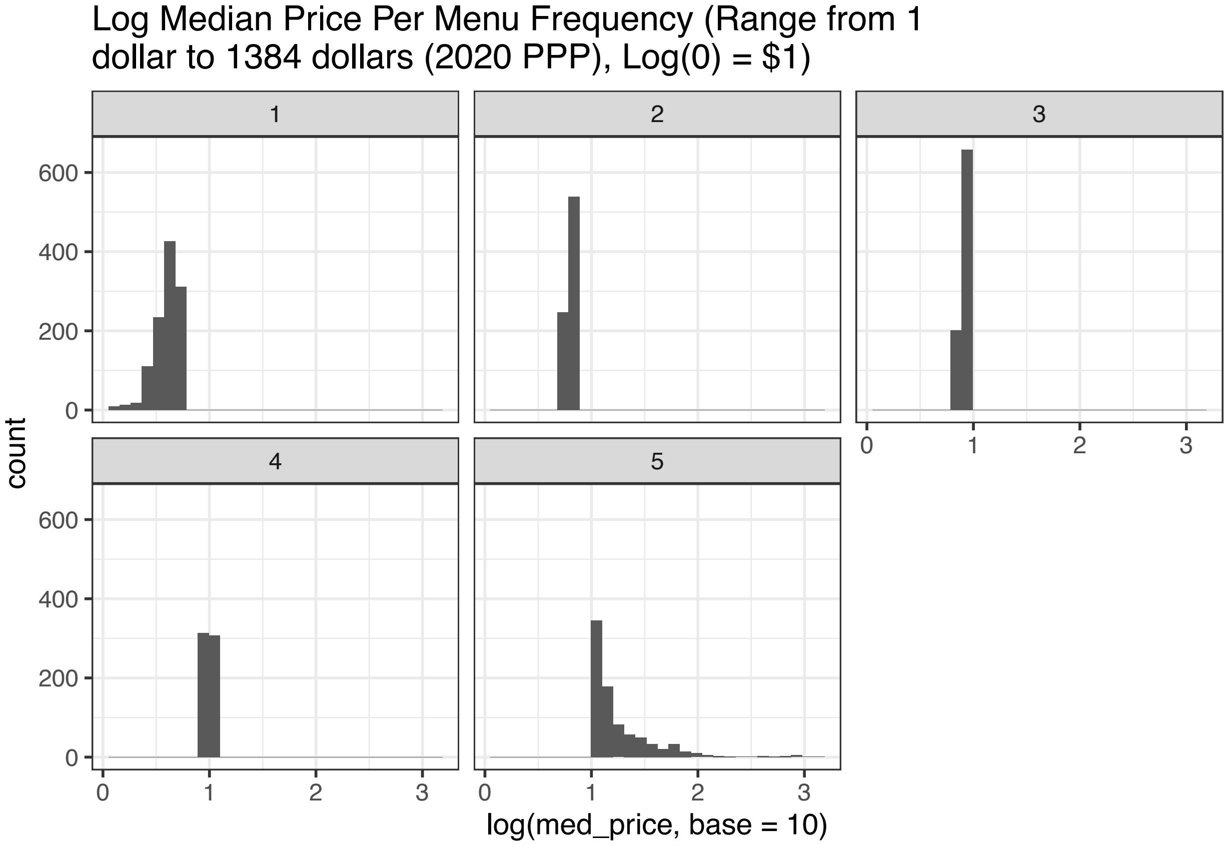
Do simple price analysis

NB only about half of the menus have prices in this DB.

As a starting point, I am doing a simple analysis by price of the menus we have. The first step is to assign each menu a median price of all the items on the menu. This serves as an ecological variable for that menu: the average price of items in a restaurant on a given day. These prices are standardized to 2020 USD PPP for comparability's sake. The prices are then split into quintiles so as to assign menus to 5 groups for easier analysis. (1 = low, 5 = high).

Below is a histogram of all prices across all menus by quintile.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
## [1] "The cutpoints for the quntiles are:"
```

```
##      20%      40%      60%      80%
##  5.00500  6.50750  8.52225 10.52720
```

Do a topic model by price quintile

This simple analysis uses LDA to create 10 topics across the entire corpus, where each menu is a “document.” That is to say that all dishes in a given menu are concatenated into one long “document” which is in turn transformed into a bag of words. Associations between words are calculated using LDA, and the 10 most salient topics are produced. These can then be analyzed across groups and across times to see if any interesting patterns emerge.

```
## A LDA_VEM topic model with 10 topics.
```

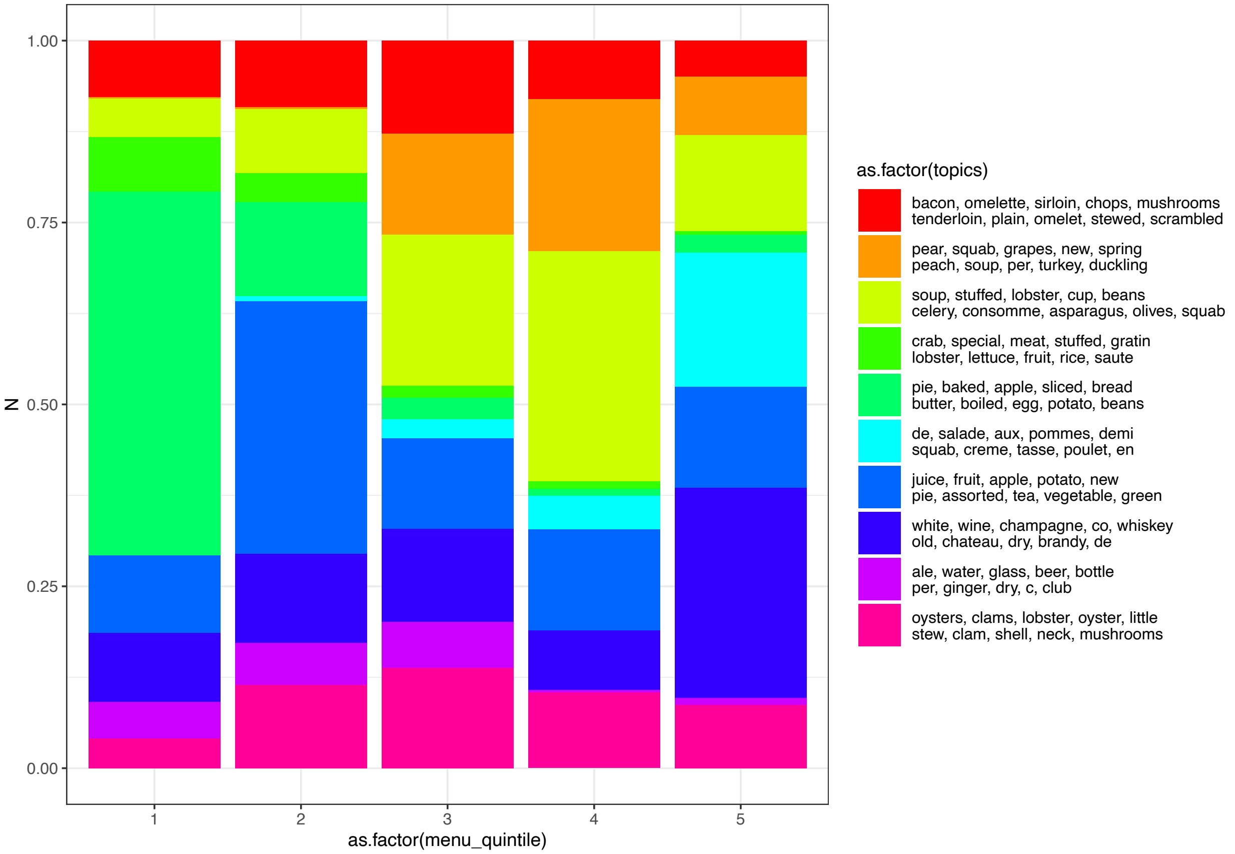
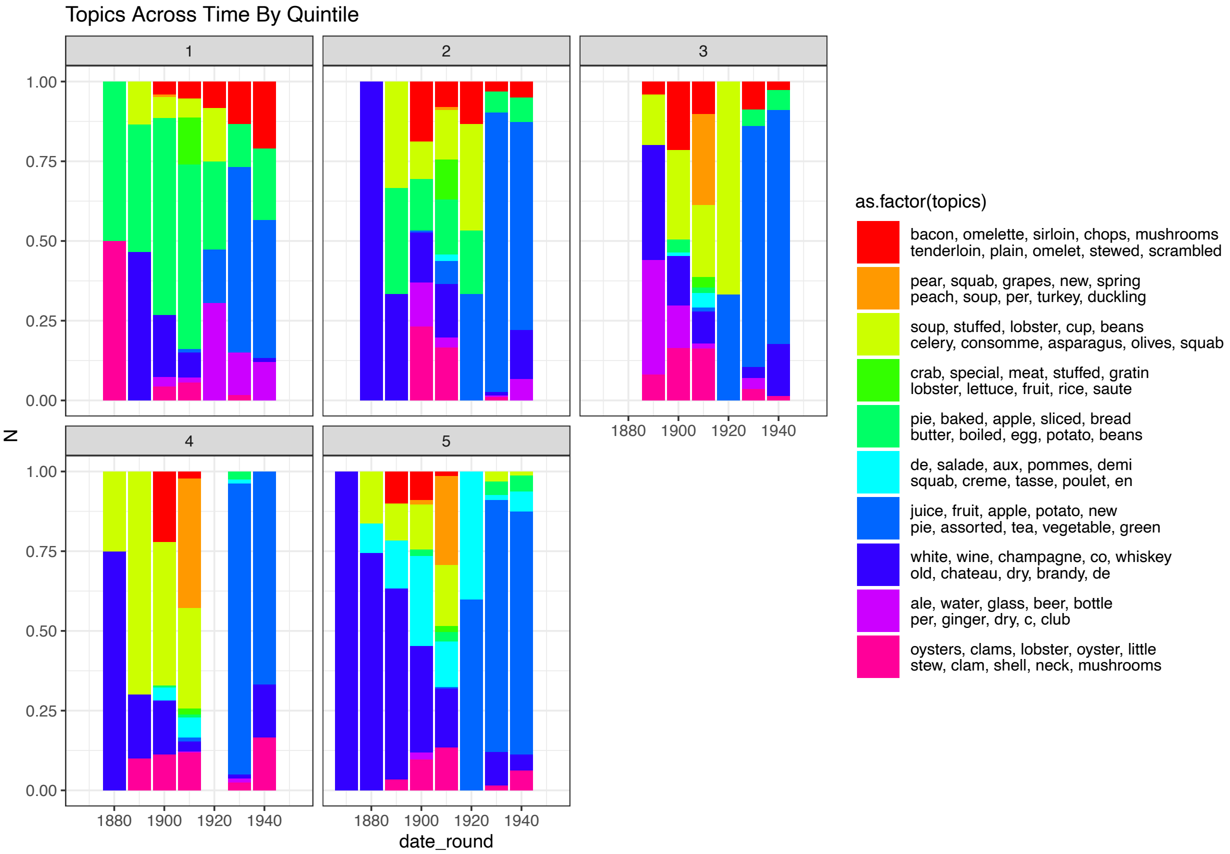


Table 2: Topics

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	bacon	pear	soup	crab	pie	de	juice	white	ale	oysters
2	omelette	squab	stuffed	special	baked	salade	fruit	wine	water	clams
3	sirloin	grapes	lobster	meat	apple	aux	apple	champagne	glass	lobster
4	chops	new	cup	stuffed	sliced	pommes	potato	co	beer	oyster
5	mushrooms	spring	beans	gratin	bread	demi	new	whiskey	bottle	little
6	tenderloin	peach	celery	lobster	butter	squab	pie	old	per	stew
7	plain	soup	consomme	lettuce	boiled	creme	assorted	chateau	ginger	clam
8	omelet	per	asparagus	fruit	egg	tasse	tea	dry	dry	shell
9	stewed	turkey	olives	rice	potato	poulet	vegetable	brandy	c	neck
10	scrambled	duckling	squab	saute	beans	en	green	de	club	mushrooms
11	tea	figs	boiled	champagne	rice	consomme	dressing	gin	soda	stewed
12	porterhouse	pie	pie	soup	tea	glace	egg	scotch	imported	sirloin
13	small	special	green	sliced	cake	volaille	soup	brut	lemonade	boiled
14	pot	crabs	sweet	wine	corned	cafe	orange	extra	white	per
15	mutton	caviar	lamb	apple	hot	terre	lettuce	st	whiskey	peas
16	extra	cherry	jelly	pie	pudding	oysters	strawberry	ale	tea	soft
17	lamb	butter	fruit	potato	two	caviar	chocolate	red	plain	green
18	two	stuffed	duck	stewed	bacon	le	butter	claret	cup	potato
19	cakes	celery	assorted	spaghetti	cakes	pois	lobster	sherry	seltzer	onions
20	hash	beans	peas	extra	corn	lamb	saute	creme	cigars	lamb

Repeat analysis across time

As I mentioned in my email, I'm afraid all the old menus are biased towards very fancy restaurants, making this a pretty useless analysis (champagne in first quintile restaurants?). Still a fun exercise. Nice to see prohibition did in fact reduce alcohol in menus, so at least the model is working at its most basic level.


Topics Across Time By Quintile
N
date_round
as.factor(topics)
bacon, omelette, sirloin, chops, mushrooms tenderloin, plain, omelet, stewed, scrambled
pear, squab, grapes, new, spring peach, soup, per, turkey, duckling
soup, stuffed, lobster, cup, beans celery, consomme, asparagus, olives, squab
crab, special, meat, stuffed, gratin lobster, lettuce, fruit, rice, saute
pie, baked, apple, sliced, bread butter, boiled, egg, potato, beans
de, salade, aux, pommes, demi squab, creme, tasse, poulet, en
juice, fruit, apple, potato, new pie, assorted, tea, vegetable, green
white, wine, champagne, co, whiskey old, chateau, dry, brandy, de
ale, water, glass, beer, bottle per, ginger, dry, c, club
oysters, clams, lobster, oyster, little stew, clam, shell, neck, mushrooms