COVID-19 Data Warehouse for Chicago and Los Angeles

By: Panthi Patel, Adam Miller, Austin Kartheiser

Datasets

- COVID testing by demographic from Chicago
- COVID testing data from LA County
- COVID case data from LA county
- Reasonable amount of data (750~ days)
- Interesting format differences between sets

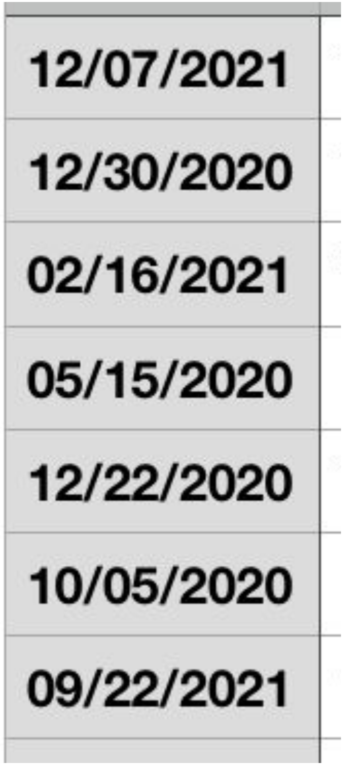
Identified Problems

- Vastly different schemas among datasets
  - Chicago set has 59 attributes
- Entirely empty/redundant data columns in LA cases set
- Presence of negative deaths/cases in LA cases set
- Inconsistent date format
- Disagreed numbers between LA sets
- Chicago data isn’t even sorted [dates are not in order]

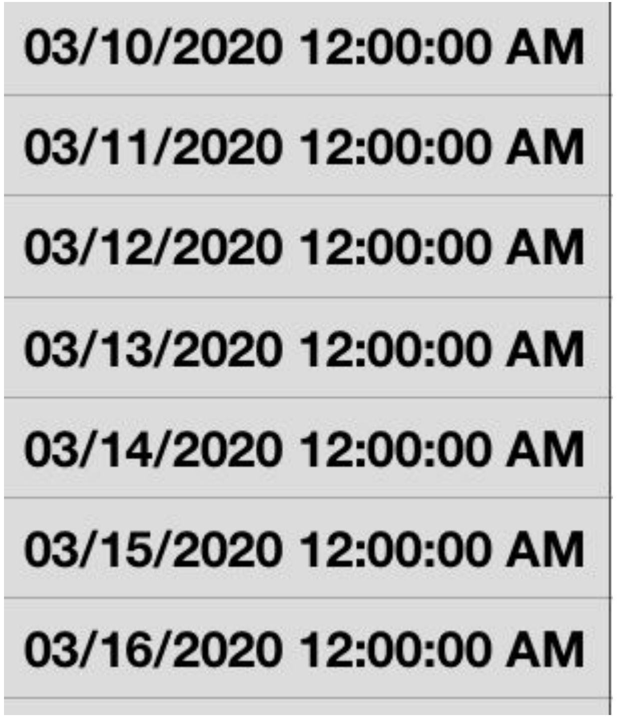
Solutions to Identified Problems

- Sorted Chicago data by date
- Visualized California data to see which set is more consistent
    - Use that set when figures disagree
- Added Chicago & Illinois columns to distinguish
- Unified date format
- Eliminated unnecessary columns from all sets
- Handle any negative and nonsensical values

Date Formats



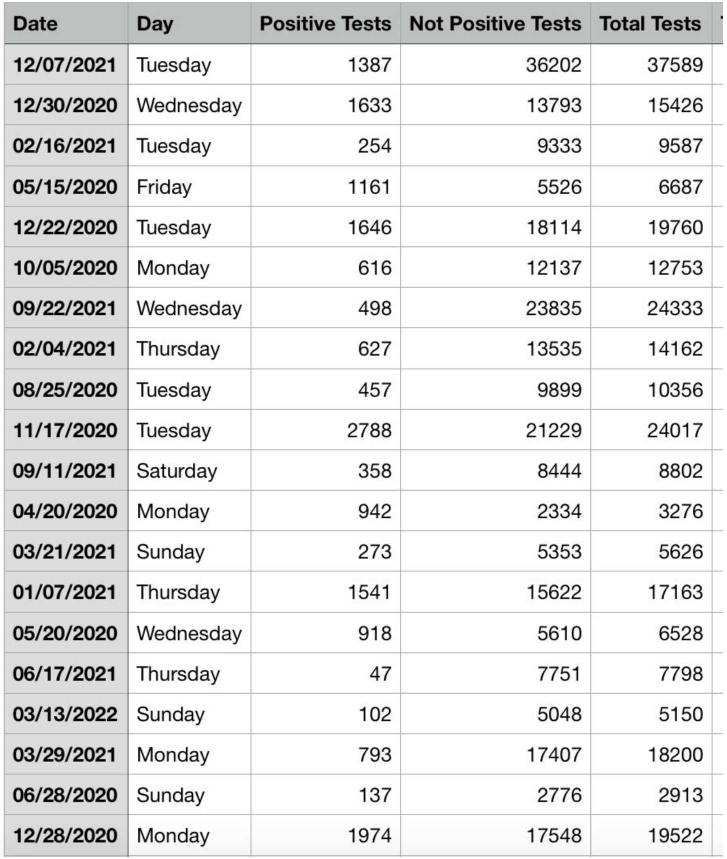
12/07/2021
12/30/2020
02/16/2021
05/15/2020
12/22/2020
10/05/2020
09/22/2021



03/10/2020 12:00:00 AM
03/11/2020 12:00:00 AM
03/12/2020 12:00:00 AM
03/13/2020 12:00:00 AM
03/14/2020 12:00:00 AM
03/15/2020 12:00:00 AM
03/16/2020 12:00:00 AM

BEFORE : COVID testing by demographic from Chicago

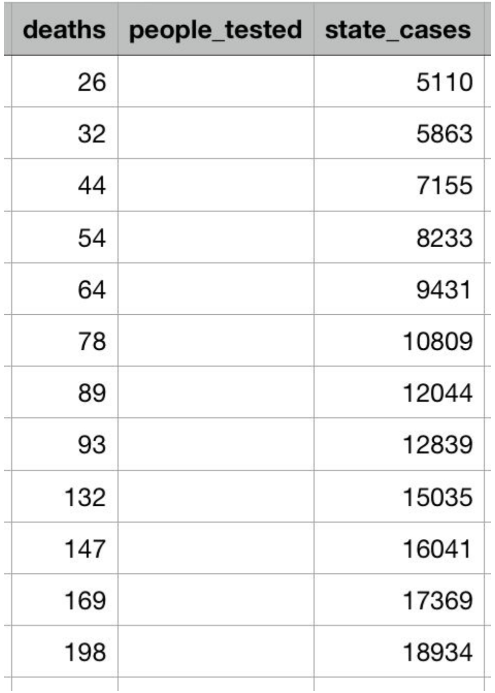
Unsorted



Date	Day	Positive Tests	Not Positive Tests	Total Tests
12/07/2021	Tuesday	1387	36202	37589
12/30/2020	Wednesday	1633	13793	15426
02/16/2021	Tuesday	254	9333	9587
05/15/2020	Friday	1161	5526	6687
12/22/2020	Tuesday	1646	18114	19760
10/05/2020	Monday	616	12137	12753
09/22/2021	Wednesday	498	23835	24333
02/04/2021	Thursday	627	13535	14162
08/25/2020	Tuesday	457	9899	10356
11/17/2020	Tuesday	2788	21229	24017
09/11/2021	Saturday	358	8444	8802
04/20/2020	Monday	942	2334	3276
03/21/2021	Sunday	273	5353	5626
01/07/2021	Thursday	1541	15622	17163
05/20/2020	Wednesday	918	5610	6528
06/17/2021	Thursday	47	7751	7798
03/13/2022	Sunday	102	5048	5150
03/29/2021	Monday	793	17407	18200
06/28/2020	Sunday	137	2776	2913
12/28/2020	Monday	1974	17548	19522

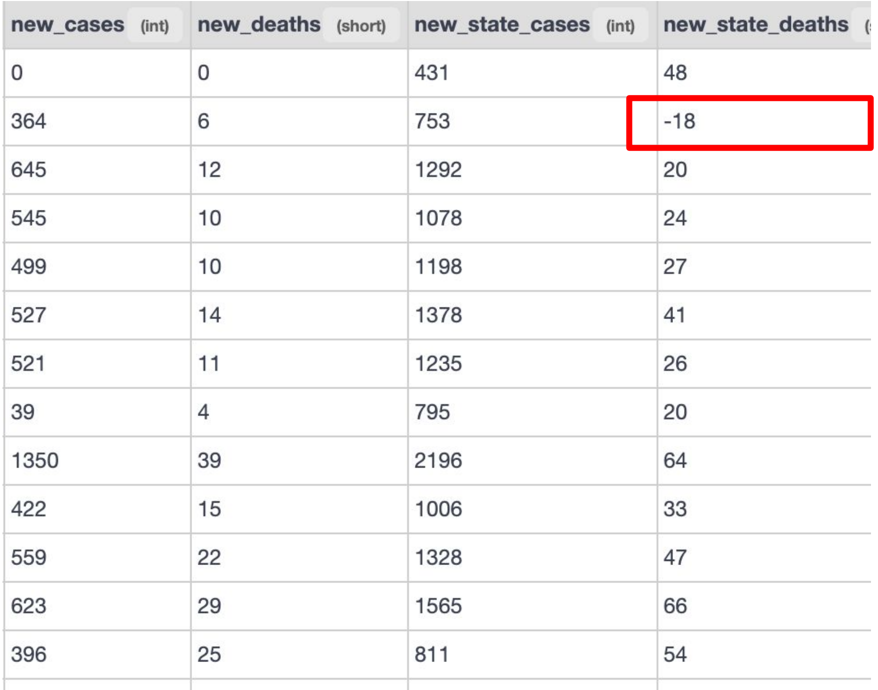
BEFORE : COVID case data from LA county

Empty Columns



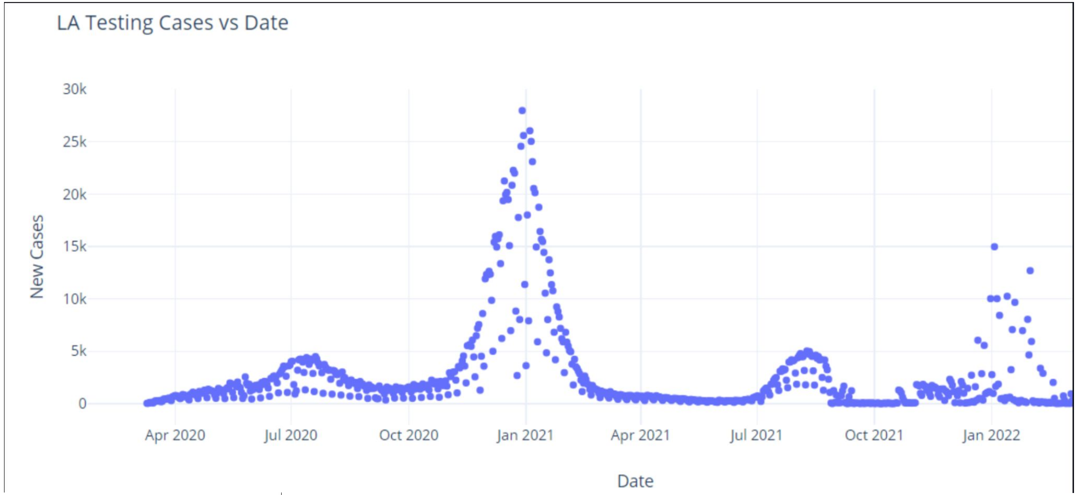
deaths	people_tested	state_cases
26		5110
32		5863
44		7155
54		8233
64		9431
78		10809
89		12044
93		12839
132		15035
147		16041
169		17369
198		18934

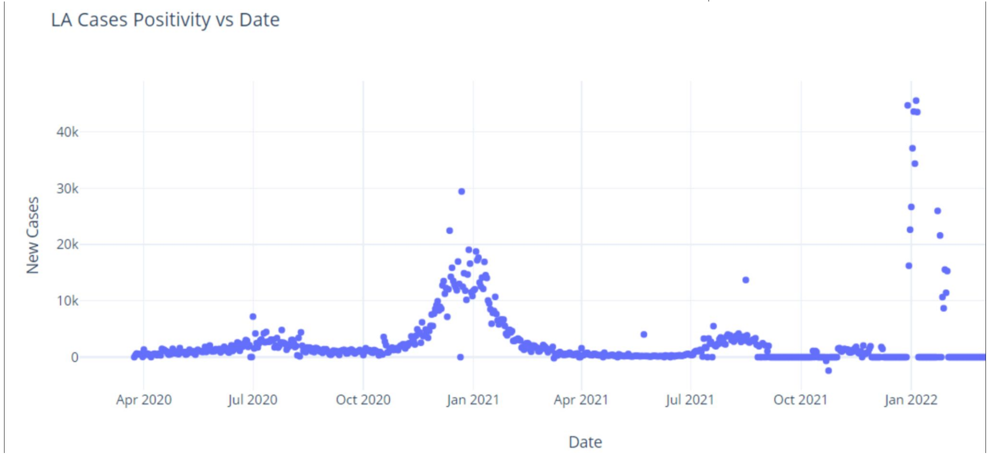
Negative Deaths



new_cases (int)	new_deaths (short)	new_state_cases (int)	new_state_deaths (
0	0	431	48
364	6	753	-18
645	12	1292	20
545	10	1078	24
499	10	1198	27
527	14	1378	41
521	11	1235	26
39	4	795	20
1350	39	2196	64
422	15	1006	33
559	22	1328	47
623	29	1565	66
396	25	811	54

Inconsistent Information

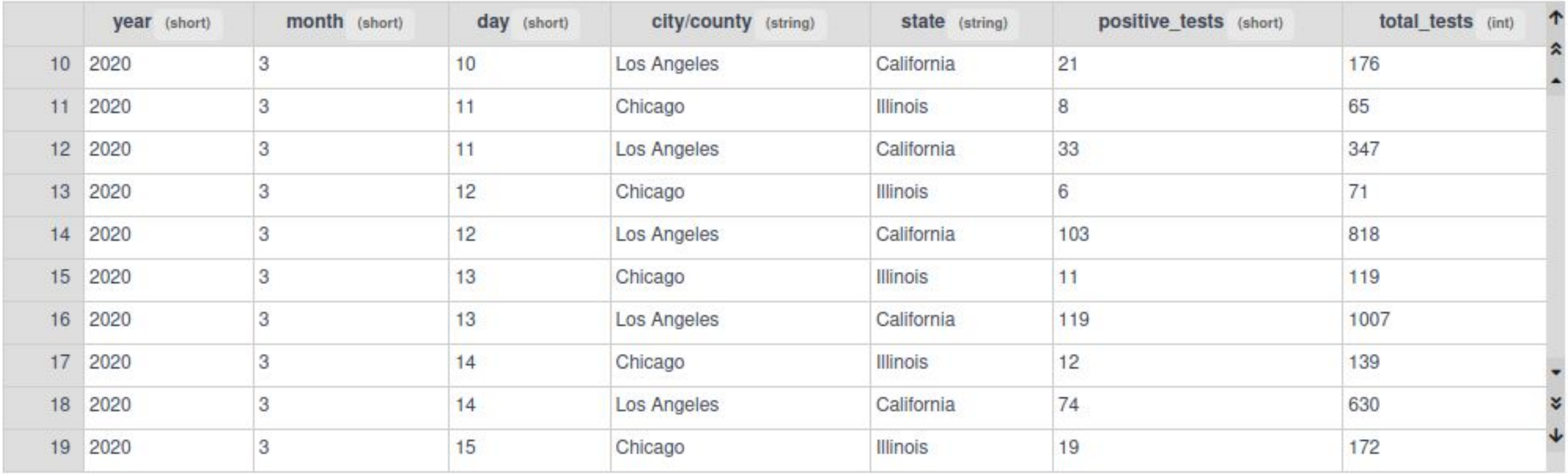




After: Explanation

- COVID data from LA Testing set was chosen
- No info was used from LA Cases set
    - Did help with the schema for county & state
- Date was normalized into 3 columns
- Sorted on year, month, day, city

After: Resulting Schema



	year (short)	month (short)	day (short)	city/county (string)	state (string)	positive_tests (short)	total_tests (int)
10	2020	3	10	Los Angeles	California	21	176
11	2020	3	11	Chicago	Illinois	8	65
12	2020	3	11	Los Angeles	California	33	347
13	2020	3	12	Chicago	Illinois	6	71
14	2020	3	12	Los Angeles	California	103	818
15	2020	3	13	Chicago	Illinois	11	119
16	2020	3	13	Los Angeles	California	119	1007
17	2020	3	14	Chicago	Illinois	12	139
18	2020	3	14	Los Angeles	California	74	630
19	2020	3	15	Chicago	Illinois	19	172