sample.bib

XPL Big Data Group Project

Ben Trovato[*]
G.K.M. Tobin[*]
trovato@corporation.com
webmaster@marysville-ohio.com
Institute for Clarity in Documentation
Dublin, Ohio, USA

Jeff Pan
New York University
Tandon School of Engineering
New York, United States
jp4839@nyu.edu

Hansheng Li
New York University
Tandon School of Engineering
New York, United States
hl4346@nyu.edu

Figure 1: Big Data Analysis on Covid Data.

ACM Reference Format:
Ben Trovato, G.K.M. Tobin, Jeff Pan, and Hansheng Li. 2021. XPL Big Data Group Project. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROJECT DESCRIPTION

The wide spreading of COVID over world has resulted in serious impacts among all areas, including economy, social life style, health and so on. This project focus on analysis of COVID impact on some Economic trends, such as stock price, gas price, fund, real estate, etc. How are those connected with COVID stats/spread? How was unemployment rate, individual saving and investment impacted or changed during this period? With those questions as background, this project will be implementing some ideas/techniques on analysis toward those areas.

2 LIST OF DATASET

In general, we will use below datasets to analyze the trend of American economics under the effect of Covid-19. The datasets are all cleaned and ready to be analyzed in the next phrase of the project.

[*] Both authors contributed equally to this research.

However, it may be the case that the datasets are too much or too less for our actual analysis. So, the final datasets that will be used may be slightly vary from the current ones.

2.1 COVID case and death count from NY Times

This dataset contains a series of data files with cumulative counts of coronavirus cases in the United States at the state level.

2.2 Unemploy rate for each states in United States

This dataset contains a series of data files with cumulative counts of Unemploy rate in the United States at the state level.

2.3 Average housing price and mortgage loan trend

One dataset contains historical avrage housing price in the U.S. And another dataset contains the corresponding mortgage loan for housing.

2.4 American stock market indexes

There are three datasets for Nasdaq, SP500, and Dow Jone's indexes. Each of them has date and index rate columns.

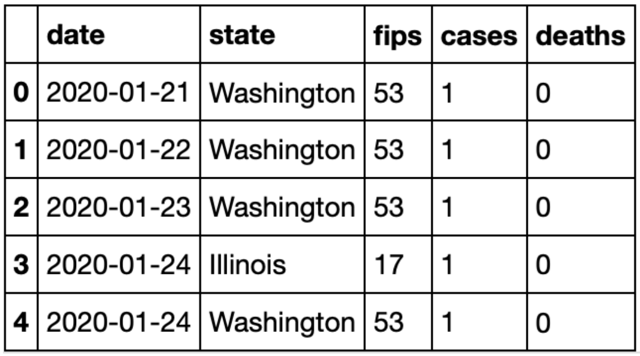
2.5 American gasoline trend

The dataset, "gas_price.csv", is the historical gas price recorded historically in the U.S.

3 DATA CLEANING

3.1 COVID case and death count dataset

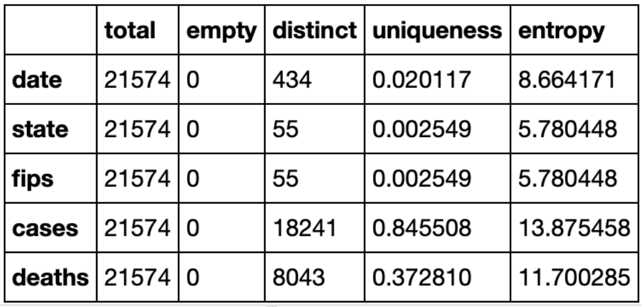
3.1.1 Data Loading. As shown in Figure 2, this dataset contains five columns: date, state, fips, cases and deaths. More precisely, 'date' is in date format; 'state' is in string format; and 'fips', 'cases' and 'deaths' are in integer format.



	date	state	fips	cases	deaths
0	2020-01-21	Washington	53	1	0
1	2020-01-22	Washington	53	1	0
2	2020-01-23	Washington	53	1	0
3	2020-01-24	Illinois	17	1	0
4	2020-01-24	Washington	53	1	0

Figure 2: Initial Data Overview

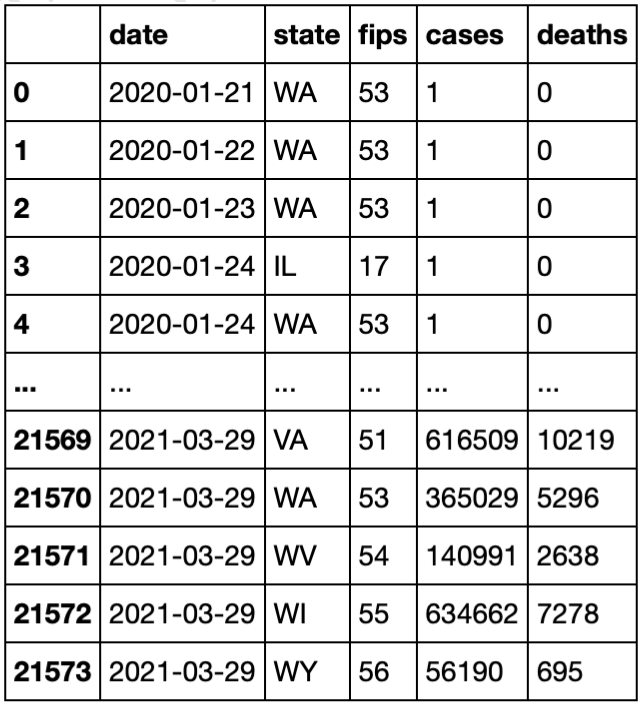
3.1.2 Data Profiling. Figure 3 shows the data stats of this dataset. According to the stats, no missing value detection was needed in the cleaning process. Also, the distinct state values match with the distinct fips, where fips code is unique to its assigned state.



	total	empty	distinct	uniqueness	entropy
date	21574	0	434	0.020117	8.664171
state	21574	0	55	0.002549	5.780448
fips	21574	0	55	0.002549	5.780448
cases	21574	0	18241	0.845508	13.875458
deaths	21574	0	8043	0.372810	11.700285

Figure 3: Data Stats

3.1.3 Data Convention. The 'state' column value, name of states, was converted into Two-Letter Postal Abbreviation. since they are capitalized letters and unique, it helps join with other dataset for future operations. The final dataset view is shown in Figure 4.

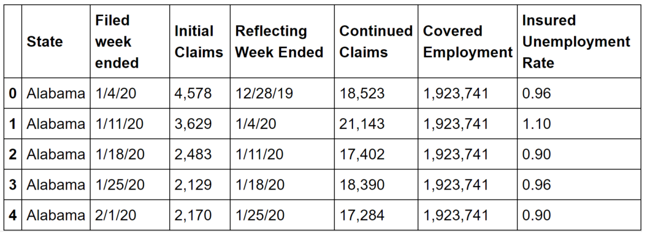


	date	state	fips	cases	deaths
0	2020-01-21	WA	53	1	0
1	2020-01-22	WA	53	1	0
2	2020-01-23	WA	53	1	0
3	2020-01-24	IL	17	1	0
4	2020-01-24	WA	53	1	0
...	...	...	...	...	...
21569	2021-03-29	VA	51	616509	10219
21570	2021-03-29	WA	53	365029	5296
21571	2021-03-29	WV	54	140991	2638
21572	2021-03-29	WI	55	634662	7278
21573	2021-03-29	WY	56	56190	695

Figure 4: Data Stats

3.2 Unemploy rate for each states in United States dataset

3.2.1 This is a dataset for unemploy in each state of United States, we find there are 53 states, normally should be 50. By compare with



	State	Filed week ended	Initial Claims	Reflecting Week Ended	Continued Claims	Covered Employment	Insured Unemployment Rate
0	Alabama	1/4/20	4,578	12/28/19	18,523	1,923,741	0.96
1	Alabama	1/11/20	3,629	1/4/20	21,143	1,923,741	1.10
2	Alabama	1/18/20	2,483	1/11/20	17,402	1,923,741	0.90
3	Alabama	1/25/20	2,129	1/18/20	18,390	1,923,741	0.96
4	Alabama	2/1/20	2,170	1/25/20	17,284	1,923,741	0.90

Figure 5: Unemploy

the normal 50 states, it shows that the dataset have include: "District of Columbia, Puerto Rico, Virgin Islands". Therefor this need to be remember for later join and comparison operation. We need to decided to keep these three place or not.

4 GITHUB REPOSITORY LINK

https://github.com/Hansheng-Li/BigDataGroupProject-XPL

Below are the references for the datasets [1] [3] [2] [5] [4]

REFERENCES

[1] Datahub. *gas price.* URL: https://datahub.io/core/natural-gas.

[2] FHFA. *house price.* URL: https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index-Datasets.aspx.

[3] Fred. *stock market indexes.* URL: https://fred.stlouisfed.org/series.

[4] NYTimes. *covid-19 data.* URL: https://github.com/nytimes/covid-19-data/blob/master/us-states.csv.

[5] USDL. *unemployment rate.* URL: https://oui.doleta.gov/unemploy/claims.asp.