Part 3
Correlation Matrix & Heatmap

Zheng HE (Frederick)

University of Chicago

MACS 30122 Final Project

Three tasks

- Create the final dataframe
- Create the correlation matrix
- Create the heatmap

Special Packages

- pandas: A package for handling dataframes and generating the correlation matrix.
- dataframe_image: A package to embed pandas DataFrames as images when generating output files.
- seaborn: A package for generating the heatmap

Creating the final dataframe

- data (four quarterly keyword lists from Naiyu's output files):

> 0_word_frequency.txt

> 1_word_frequency.txt

> 2_word_frequency.txt

> 3_word_frequency.txt

Top 20 threshold

emergency	34	emergency	71	relief	41	relief	8
relief	27	protection	61	emergency	29	safe	8
response	17	relief	48	protection	22	recovery	7
care	13	care	38	assistance	21	care	7
business	13	response	29	business	18	business	6
public	11	business	27	care	18	protection	6
assistance	9	public	23	public	16	access	5
testing	9	access	21	rural	16	protecting	4
security	8	amend	21	response	15	emergency	4
federal	8	assistance	20	national	13	vaccine	4
financial	8	national	18	protecting	12	support	4
secretary	7	recovery	16	access	12	data	4
credit	7	support	15	support	11	supporting	3
access	6	protecting	14	fund	10	federal	3
housing	6	federal	13	transparency	10	response	3
cost	6	secretary	13	recovery	10	security	3
defense	6	housing	12	worker	9	increase	3
loan	6	credit	12	extension	9	healthy	3
charitable	6	community	12	child	9	economic	3
aid	5	direct	12	community	9	expansion	3
coverage	5	flexibility	12	safe	8	direct	3
community	5	loan	11	federal	8	learning	3
disaster	5	human	10	united	7	strengthening	3
protection	5	accountability	10	tax	7	fund	3
amend	5	economic	9	essential	7		
family	5			training	7		
crisis	5			transportation	7		
supporting	5			coverage	7		

Creating the final dataframe

- In this new dataframe, only the keywords that appear in all the four top-20 keyword dataframes as well as their quarterly frequencies are kept.
- There are six keywords after the filtering: emergency, relief, response, care, business, access
- Six medical and socio-economic parameters are also added; sources are NYTimes Github dataset, OECD and US Bureau of Labor Statistics.

Creating the final dataframe

	keyword	2020_qt1	2020_qt2	2020_qt3	2020_qt4
0	emergency	34	71	29	4
1	relief	27	48	41	8
2	response	17	29	15	3
3	care	13	38	18	7
4	business	13	27	18	6
5	access	6	21	12	5
6	us covid positive	188461	2464860	4609413	12764156
7	us covid positive case growth rate %	100	1207.89	87.0051	176.915
8	us covid death	4304	123158	79390	139198
9	us covid death growth rate %	100	2761.48	-35.5381	75.3344
10	us unemployment rate %	3.8	13.06667	8.8	6.766667
11	us productivity %	-0.8	11.1	4.2	-4.2

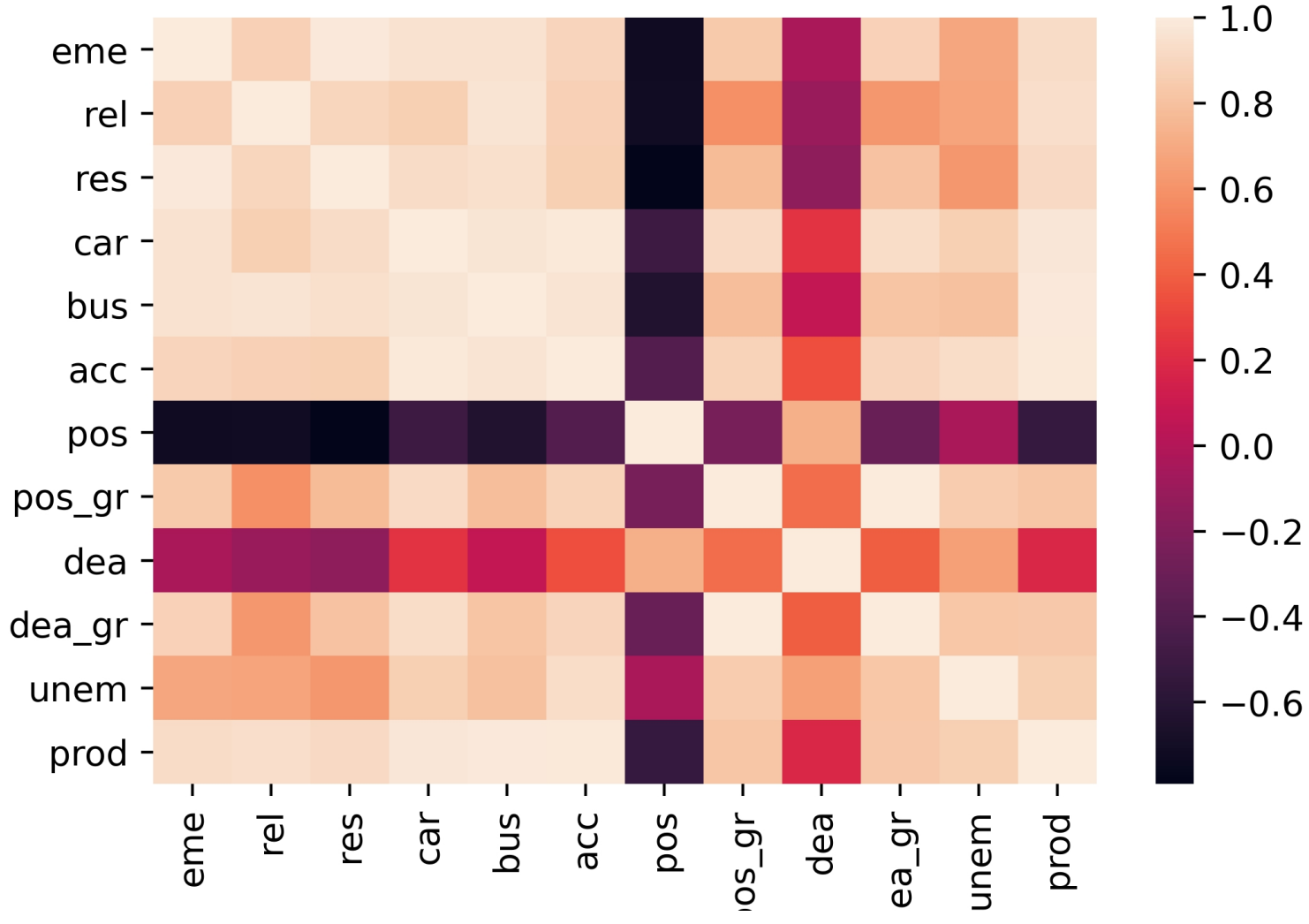
Creating the correlation matrix

- **A correlation matrix** is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables.
- **A correlation matrix** is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

Creating the correlation matrix

	eme	rel	res	car	bus	acc	pos	pos_gr	dea	dea_gr	unem	prod
eme	1.000000	0.867134	0.992116	0.958009	0.951531	0.894205	-0.714897	0.844691	-0.037771	0.875323	0.682441	0.925875
rel	0.867134	1.000000	0.897920	0.861564	0.960428	0.868932	-0.711612	0.587309	-0.103665	0.618126	0.676284	0.939105
res	0.992116	0.897920	1.000000	0.927137	0.950140	0.860705	-0.790622	0.771124	-0.147134	0.807840	0.618783	0.913176
car	0.958009	0.861564	0.927137	1.000000	0.968750	0.982055	-0.507421	0.916874	0.231918	0.930844	0.863344	0.976904
bus	0.951531	0.960428	0.950140	0.968750	1.000000	0.960453	-0.633048	0.789278	0.067930	0.812559	0.797934	0.992882
acc	0.894205	0.868932	0.860705	0.982055	0.960453	1.000000	-0.399725	0.883040	0.341418	0.888952	0.931021	0.985203
pos	-0.714897	-0.711612	-0.790622	-0.507421	-0.633048	-0.399725	1.000000	-0.247362	0.720413	-0.309152	-0.039526	-0.536459
pos_gr	0.844691	0.587309	0.771124	0.916874	0.789278	0.883040	-0.247362	1.000000	0.453653	0.997881	0.846202	0.819373
dea	-0.037771	-0.103665	-0.147134	0.231918	0.067930	0.341418	0.720413	0.453653	1.000000	0.395094	0.653282	0.183869
dea_gr	0.875323	0.618126	0.807840	0.930844	0.812559	0.888952	-0.309152	0.997881	0.395094	1.000000	0.828169	0.835266
unem	0.682441	0.676284	0.618783	0.863344	0.797934	0.931021	-0.039526	0.846202	0.653282	0.828169	1.000000	0.863572
prod	0.925875	0.939105	0.913176	0.976904	0.992882	0.985203	-0.536459	0.819373	0.183869	0.835266	0.863572	1.000000

Creating the heatmap



THANKS FOR WATCHING!