Visualising Data Assignment 2021

C17442074 Mark Higgins

The Dataset

Link: https://www.kaggle.com/kimjihoo/coronavirusdataset

The dataset used was the Data Science for COVID-19 (DS4C) dataset on Kaggle. It is an analysis of COVID-19 data for South Korea in 2020. The source of the data was the KCDC (Korea Centers for Disease Control & Prevention). The dataset is split up into 11 CSV files but only 4 were used for telling the story.

It was chosen as it suits the requirements of the assignment well. Across the files used, there are 5000+ rows of data with more than 8 attributes with qualitative, quantitative and temporal data. It also contains text data and location data but these weren't necessary for the story.

It unfortunately is not up-to-date and the updating ceased on 30th June 2020 so it really only shows how COVID-19 impacted South Korea for the first few months.

1. Case File

	case_id	province	city	group_infection	infection_case	confirmed_cases	latitude	longitude
	<int>	<chr>	<chr>	<lgl>	<chr>	<int>	<chr>	<chr>
1	1000001	Seoul	Yongsan-gu	TRUE	Itaewon Clubs	139	37.538621	126.992652
2	1000002	Seoul	Gwanak-gu	TRUE	Richway	119	37.48208	126.901384
3	1000003	Seoul	Guro-gu	TRUE	Guro-gu Call Center	95	37.508163	126.884387
4	1000004	Seoul	Yangcheon-gu	TRUE	Yangcheon Table Tennis Club	43	37.546061	126.874209
5	1000005	Seoul	Dobong-gu	TRUE	Day Care Center	43	37.679422	127.044374
6	1000006	Seoul	Guro-gu	TRUE	Manmin Central Church	41	37.481059	126.894343

The **Case.csv** file contains 8 attributes and 174 rows and contains continuous, categorical and location data. The province, city and infection_case columns are categorical, confirmed_cases is discrete and latitude and longitude are both location data. The rest of the files will be discussed too.

2. PatientInfo File

	patient_id	sex	age	country	province	city	infection_case	infected_by	contact_number	symptom_onset_date	confirmed_date	released_date
	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
1	1e+09	male	50s	Korea	Seoul	Gangseo-gu	overseas inflow	NA	75	2020-01-22	2020-01-23	2020-02-05
2	1e+09	male	30s	Korea	Seoul	Jungnang-gu	overseas inflow	NA	31	NA	2020-01-30	2020-03-02
3	1e+09	male	50s	Korea	Seoul	Jongno-gu	contact with patient	2002000001	17	NA	2020-01-30	2020-02-19
4	1e+09	male	20s	Korea	Seoul	Mapo-gu	overseas inflow	NA	9	2020-01-26	2020-01-30	2020-02-15
5	1e+09	female	20s	Korea	Seoul	Seongbuk-gu	contact with patient	1000000002	2	NA	2020-01-31	2020-02-24
6	1e+09	female	50s	Korea	Seoul	Jongno-gu	contact with patient	1000000003	43	NA	2020-01-31	2020-02-19

The PatientInfo file contains 14 attributes and 5165 rows. It discusses data related to patients affected by COVID-19. Unfortunately, there is a lot of missing data, particularly for the infected_by and symptom_onset_date columns which limits their usefulness. The file contains categorical, discrete, temporal and text data. The columns symptom_onset_date, confirmed_date and released_date are temporal.

3. Region File

	province_code	province	city	latitude	longitude	elementary_school_count	kindergarten_count	university_count	academy_ratio	elderly_population_
	<int>	<chr>	<chr>	<dbl>	<dbl>	<int>	<int>	<int>	<dbl>	<
1	10000	Seoul	Seoul	37.56695	126.9780	607	830	48	1.44	
2	10010	Seoul	Gangnam-gu	37.51842	127.0472	33	38	0	4.18	
3	10020	Seoul	Gangdong-gu	37.53049	127.1238	27	32	0	1.54	
4	10030	Seoul	Gangbuk-gu	37.63994	127.0255	14	21	0	0.67	
5	10040	Seoul	Gangseo-gu	37.55117	126.8495	36	56	1	1.17	
6	10050	Seoul	Gwanak-gu	37.47829	126.9515	22	33	1	0.89	

elderly_population_ratio	elderly_alone_ratio	nursing_home_count
<dbl>	<dbl>	<int>
15.38	5.8	22739
13.17	4.3	3088
14.55	5.4	1023
19.49	8.5	628
14.39	5.7	1080
15.12	4.9	909

The Region file contains 12 attributes and 244 rows, with categorical, continuous, discrete and location data. This is the file which contains the continuous numeric data as required by the assignment, in the form of ratios.

4. Time File

	date	time	test	negative	confirmed	released	deceased
	<date>	<int>	<int>	<int>	<int>	<int>	<int>
1	2020-01-20	16	1	0	1	0	0
2	2020-01-21	16	1	0	1	0	0
3	2020-01-22	16	4	3	1	0	0
4	2020-01-23	16	22	21	1	0	0
5	2020-01-24	16	27	25	2	0	0
6	2020-01-25	16	27	25	2	0	0

The Time file has 163 rows and 7 attributes. It shows the accumulated number of tests, negative cases, confirmed cases, people released and deceased at a specific date and time. It contains just temporal and discrete data.