CS528 Big Data Analytics Midterm

Part I: Composite questions (100%) (Select 4 questions)

Q1. The Air Quality Index (AQI) is based on the concentrations of 5 pollutants. The index is calculated from the concentrations of the following pollutants: O3, NO2, SO2, PM2.5 and PM10. The breakpoints between index values are defined for each pollutant separately and the overall index is defined as the maximum value of the index. Different averaging periods are used for different pollutants (Figure 1). Please write a program to calculate the AQI of each air pollutants.

Note 1: Here, we change SO2 to hourly mean (ug/m^3).

Note 2: You need to convert units from "ppb" to "ug/m^3" of NO2, O3 and SO2 (Figure 2)

Note 3: An example shown in Figure 3

Index	Ozone, Running 8 hourly mean (µg/m^3)	Nitrogen Dioxide, Hourly mean (µg/m^3)	Sulphur Dioxide, 15 minute mean (µg/m^3)	PM2.5 Particles, 24 hour mean (µg/m^3)	PM10 Particles, 24 hour mean (µg/m^3)
1	0-33	0-67	0-88	0-11	0-16
2	34-66	68-134	89-177	12-23	17-33
3	67-100	135-200	178-266	24-35	34-50
4	101-120	201-267	267-354	36-41	51-58
5	121-140	268-334	355-443	42-47	59-66
6	141-160	335-400	444-532	48-53	67-75
7	161-187	401-467	533-710	54-58	76-83
8	188-213	468-534	711-887	59-64	84-91
9	214-240	535-600	888-1064	65-70	92-100
10	≥ 241	≥ 601	≥ 1065	≥ 71	≥ 101

Figure 1. AQI Index

SO_2	1 ppb = 2.62 µg/m^3
NO_2	1 ppb = 1.88 µg/m^3
NO	1 ppb = 1.25 µg/m^3
O_3	1 ppb = 2.00 µg/m^3
CO	1 ppb = 1.145 µg/m^3
Benzene	1 ppb = 3.19 µg/m^3

Figure 2. Convert units from "ppb" to "ug/m^3"

input

Date	Station	Type	0	1	2	3	4	5	6	7	8	9	10	11
2011/1/1	SongShang	NO2	20	26	27	16	11	14	16	18	34	19	14	15
2011/1/1	SongShang	O3	23	17	17	22	27	23	23	20	10	21	28	28
2011/1/1	SongShang	PM10	92	83	54	54	55	57	56	54	56	66	53	63
2011/1/1	SongShang	PM2.5	32	30	34	22	26	27	22	14	10	22	30	32
2011/1/1	SongShang	SO2	6.3	3.6	3.2	3.1	2.7	2.6	3.3	4.1	4.4	3.5	3.2	3.3

12	13	14	15	16	17	18	19	20	21	22	23
14	14	15	16	28	34	35	32	24	25	21	15
29	31	30	29	19	14	13	12	18	16	18	22
61	50	48	31	49	55	30	42	52	65	64	41
27	27	23	28	27	34	28	30	30	31	29	27
2.9	2.6	2.5	2.5	3	3.2	3.4	3.8	3.9	3.7	3.1	2.9

output

Type	value	value (ug/m3)	AQI
NO2	35 ppb	65.8	1
O3	26.875 ppb	53.75	2
PM10	55.46 ug/m3	55.46	4
PM2.5	26.75 ug/m3	26.75	3
SO2	6.3 ppb	16.506	1

Figure 3 An example of this question

(a) Convert ppb to ug/m^3 of NO2, O3 and SO2 (9%)

(b) Each AQI of air pollutants

- b.1 O3 (4%)
- b.2 NO2 (3%)
- b.3 SO2 (3%)
- b.4 PM2.5 (3%)
- b.5 PM10 (3%)

Save as: SID_Q1.r (e.g s1001234_Q1.r)

Q2. The three dataset, Q2_Summary.csv, Q2_Type_info.csv, and Q2_Dist_info.csv, are the actual price registration in Taoyuan. Please integrate those tables using TYPEID and DistrictID. Please use for loop and boxplot in Plotly package to implement the data visualization. (Figure 4)

Note: Please upload the files to server.

(a) Merger Data (5%)

(b) Use for loop (5%)

(c) Use Plotly package (5%)

(d) Implement in Rshiny (10%)

Save as: SID_Q2_server.r, SID_Q2_ui.r

(e.g s1001234_Q2_server.r, s1001234_Q2_ui.r)

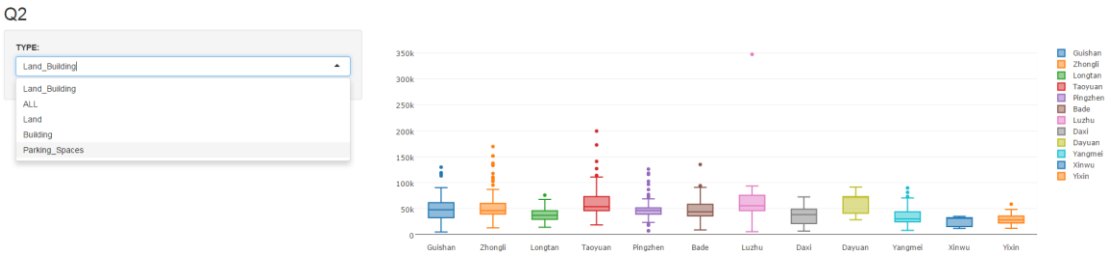

Figure 4.

Q3. The dataset, Q3_data.csv, is the MRT information in Taipei

(a) Please use Rshiny show the MRT information in the map like Figure 5. (15%)

(b) Please show the nearest 5 MRT stations when user input the lat and lon. (10%)

Hint: Use longitude and latitude information

Save as: SID_Q3_server.r, SID_Q3_ui.r

(e.g s1001234_Q3_server.r, s1001234_Q3_ui.r)

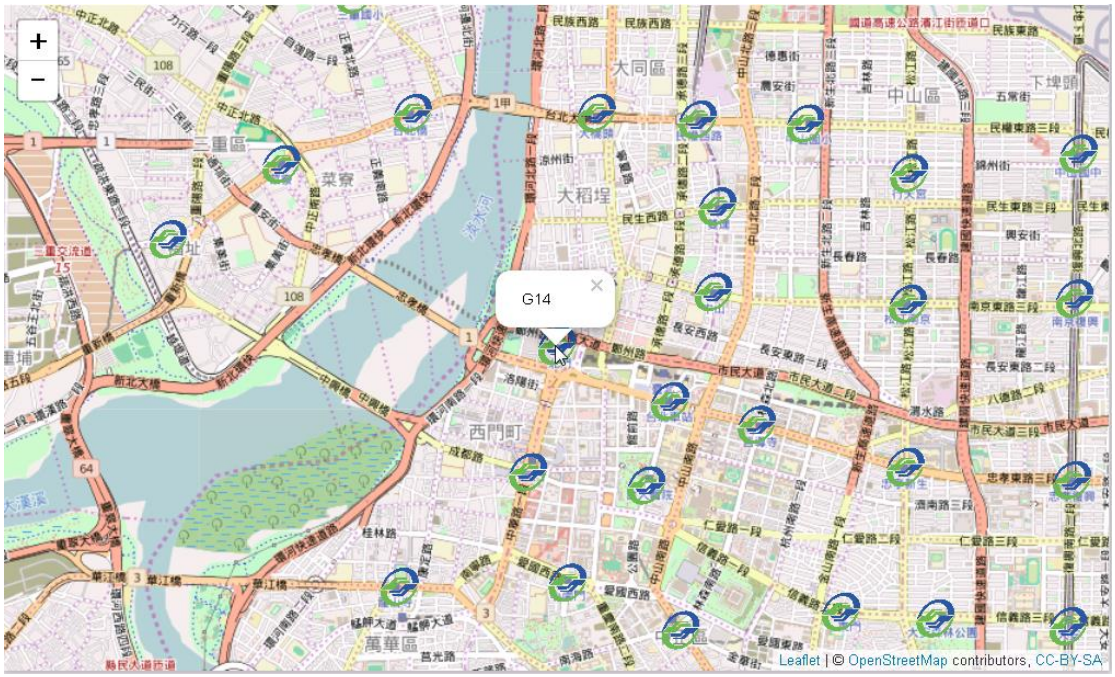

Figure 5.

Q4: In homework1, we introduce the confusion matrix. Please use the confusion matrix to calculate the accuracy, precision, sensitivity, specificity and Fscore of the clustering result.

(a) Accuracy (5%)
(b) Precision (5%)
(c) Sensitivity (5%)
(d) Specificity (5%)
(e) Fscore (5%)

Save as: SID_Q4.r (e.g s1001234_Q4.r)

		Predict			
		A	B	C	Others
Actual	A	TP_A	E_{AB}	E_{AC}	E_{AO}
	B	E_{BA}	TP_B	E_{BC}	E_{BO}
	C	E_{CA}	E_{CB}	TP_C	E_{CO}

$TP_m = TP_m$

$FP_m = \sum_{i \in K} E_{im} - TP_m$

$FN_m = \sum_{i \in K} E_{mi} - TP_m$

TN_m = #samples $-TP_m - FP_m - FN_m$

$$\text{Accuracy} = \frac{\sum_{m \in K} TP_m}{\#\ case}, K \in \{A, B, C\}$$

$$\text{Precision} = \frac{\sum_{m \in K} TP_m}{\sum_{m \in K}(TP_m + FP_m)}$$

$$\text{Sensitivity} = \frac{\sum_{m \in K} TP_m}{\sum_{m \in K}(TP_m + FN_m)}$$

$$\text{Specificity} = \frac{\sum_{m \in K} TN_m}{\sum_{m \in K}(FP_m + TN_m)}$$

$$\text{Fscore} = \frac{\sum_{m \in K} 2TP_m}{\sum_{m \in K}(2TP_m + FP_m + FN_m)}$$

Q5: The dataset, Q5_death.csv, is the cause of death in Taiwan in 2013. And the datasets Q5_cause.csv, Q5_county.csv and Q5_sex.csv are the information of cause of death, county id and sex id, respectively. Please integrate those tables by sex, cause and county.

(a) Calculate the number of cause of death in female and male and create a barplot in plotly package (Figure 5). (15%)

(b) Calculate the number of cause of death in each county and create a table like Figure 6. (10%)

Save as: SID_Q5.r (e.g s1001234_Q5.r)

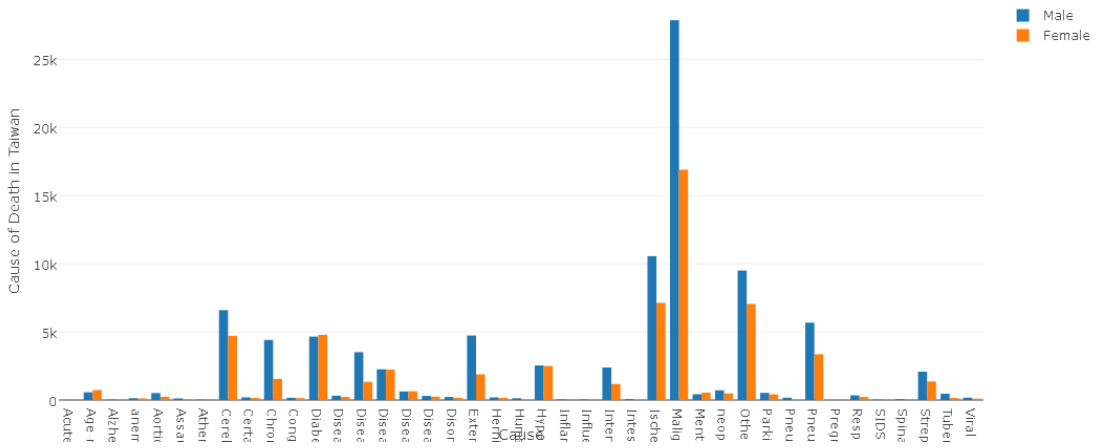

Figure 5.

```
> head(out)
```

	County1	County2	County3	County4	County5	County6	County7	County8	County9	County10	County11
Intestinal infectious diseases	1	2	0	0	2	0	1	0	0	3	2
Chronic lower respiratory diseases	39	72	32	52	56	20	58	62	49	31	57
Tuberculosis	4	8	3	6	4	1	9	9	8	5	5
Cerebrovascular diseases	95	136	59	88	109	62	147	109	103	57	123
Hypertensive diseases	46	46	20	57	31	19	50	51	29	23	60
Diseases of the musculoskeletal system and connective tissue	15	17	11	17	16	9	17	14	21	16	21
	County12	County13	County14	County15	County16	County17	County18	County19	County20	County21	
Intestinal infectious diseases	1	2	0	0	2	0	1	0	0	1	
Chronic lower respiratory diseases	45	42	20	23	22	18	8	24	18	18	
Tuberculosis	5	2	0	3	1	0	0	3	4	1	
Cerebrovascular diseases	122	70	34	37	46	35	11	40	34	23	
Hypertensive diseases	61	40	25	21	13	10	5	9	7	19	
Diseases of the musculoskeletal system and connective tissue	17	6	3	6	6	1	4	4	1	3	
	County22	County23	County24	County25	County26	County27	County28	County29	County30	County31	
Intestinal infectious diseases	0	0	0	0	0	1	0	2	0	1	
Chronic lower respiratory diseases	22	13	4	14	8	21	11	24	20	46	
Tuberculosis	0	1	1	2	2	4	1	1	0	2	
Cerebrovascular diseases	30	18	9	27	15	23	29	33	33	72	
Hypertensive diseases	16	10	7	3	5	21	15	12	16	35	
Diseases of the musculoskeletal system and connective tissue	3	3	1	1	2	7	5	5	1	7	

Figure 6.

Q6: The dataset, Q6_data.csv, is the patients' data in NHIRD (The National Health Insurance Research Database). Please create demographic table like figure 7.

Note:

id_date: Date of hospitalization

diag_amt: Diagnosis fees

room_amt: Room fees

amin_amt: Examination fees

sgry_amt: Surgical fees

age = (id_date – birthday) / 365.25

age group: <15 y/o, 15~29 y/o, 30 ~ 44 y/o, 45~64 y/o and >= 65 y/o

(a) Age_group 5%

(b) Sex 4%

(c) Average diag_amt 4%

(d) Average room_amt 4%

(e) Average amin_amt 4%

(f) Average sgry_amt 4%

	<15y/o	15~29y/o	30~44y/o	45~64y/o	>=65y/o
Male	137.00	339.000	1699.000	6110.000	5909.000
Female	112.00	214.000	601.000	2979.000	4548.000
diag_amt	12489.64	7607.325	7779.258	7669.699	8148.632
room_amt	83620.19	45173.320	45199.457	44289.199	47975.748
amin_amt	16460.43	10982.617	9100.175	8034.153	9268.775
sgry_amt	28851.13	22179.854	14674.443	10864.875	7478.414

Figure 7

Part II: Bonus (25%)

1. Please fill in the questionnaire (5%). https://goo.gl/forms/upQqfE2lfMCqIgxX2

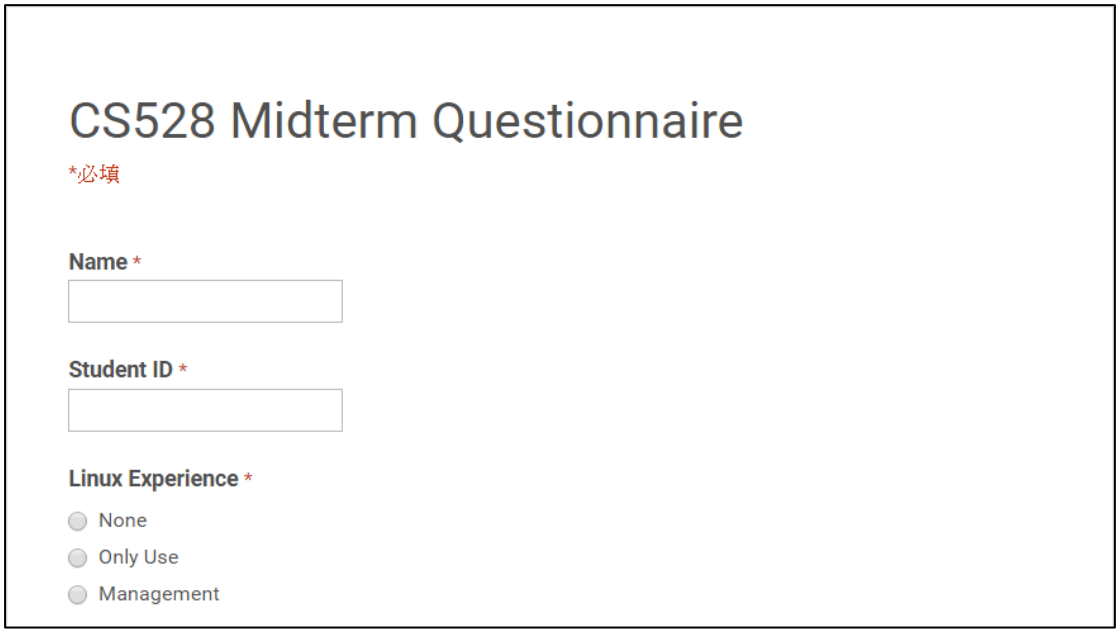

Figure 8.

2. Please describe your final project.

   (a) Title (2%)

   (b) Your idea or method (8%)

3. Please describe your homework2

   (a) Dataset (Name, URL)(5%)

   (b) Your idea (5%)

   Save as: SID_Bonus.docx (e.g s1001234_Bonus.docx)

Note:

If you select 5 questions in Part I., Score = Total – Median.

If you select 6 questions in Part I., Score = Total – Max – min

For example: Q1:20, Q2:15, Q3:25, Q4:20, Q5:10, Q6:5, Score = 95 - 25 -5 = 65