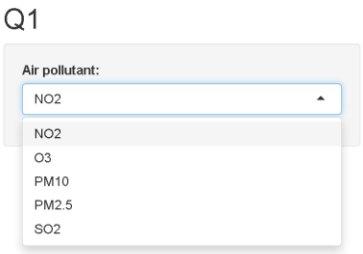
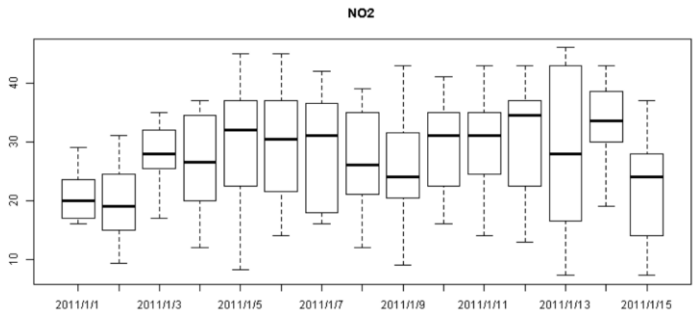
CS528A Big Data Analytics Midterm

Part I: Composite questions (130%)

Q1: The dataset, Q1.csv, is the air pollution data in Chungli in 2011. Please use shiny to implement the data visualization.

(a) User can select the type of air pollutants. (5%)
(b) The Boxplot displays daily values based on user selection.
    b.1 x-axis: date (4%)
    b.2 y-axis: air pollution concentration (3%)
    b.3 main title: air pollutant (3%)
(c) The value is displayed correctly. (5%)
(d) User can access the app by URL. (5%)





Q2. Term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a term-document matrix, columns correspond to documents in the collection and rows correspond to terms. Please write a program to generate the term-doc matrix. (20%)

Note: All articles are in Q2.

Save as: SID_Q2.r (e.g s1001234_Q2.r)

Q3. The datasets in Q3 are an example of the school curriculum database. Please write the R program to solve the following problems.

Note:

The student.csv is student information (student ID, gender, email)

The class.csv is class information (course ID, credit)

The score.csv is the record of student grades (student ID, course ID, score)

(a) How many male and female students are in the class (4%)

(b) How many students are attending the course AA101? (4%)

(c) How many students are failed in course AA101 (4%)

(d) How many credits did the student s1000001 pass in this semester (6%)?

(e) How many male and female students are in each course (7%)

Save as: SID_Q3.r (e.g s1001234_Q3.r)

Q4. Q4.csv is derived from 311 New York Open Data (ref: R programming.v2 p127).

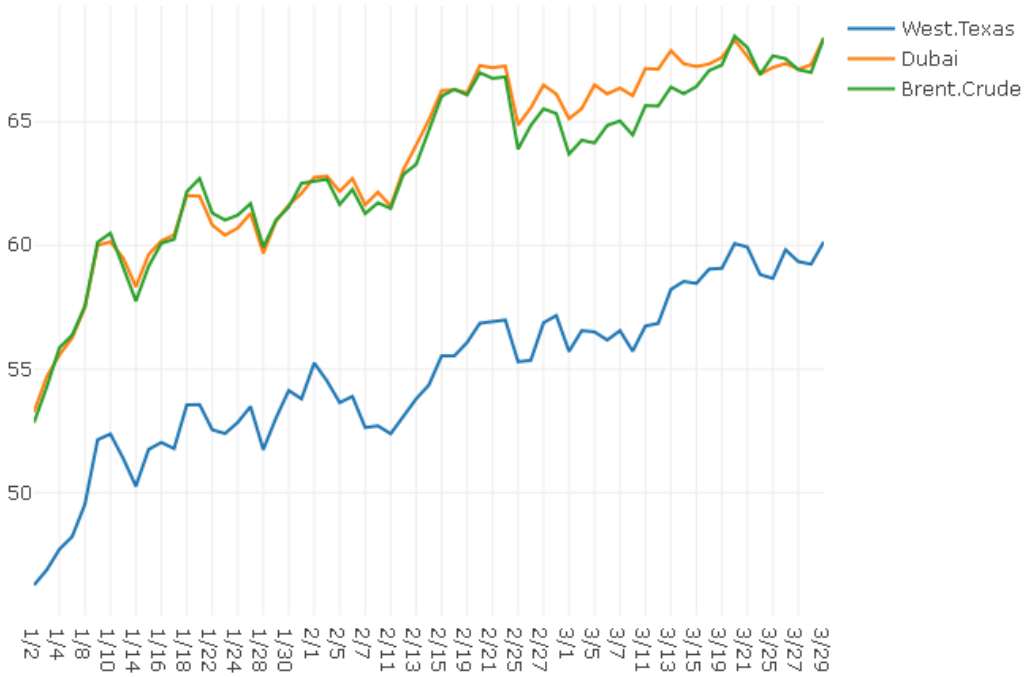
(a) Please use map (leaflet) show the complaint information in the map.

a.1 If the complaint type is “Blocked Driveway”, please use red dot in the map; If the complaint type is “Street Condition”, please use green dot in the map; If the complaint type is “Illegal Parking”, please use yellow dot in the map; And the others are blue dot. (15%)

a.2 Please use popup to show the “Resolution Description”. (10%)

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

Q5. Q5.csv is the international crude oil price, which include West Texas, Dubai and Brent Crude. Please use Plotly package to show the line chart like Figure. (15%) Save as: SID_Q5.r (e.g s1001234_Q5.r)



Q6. The dataset, Q6.csv, is an example of equipment sensor data from the manufacturing industry. Due to the limiting of sensor technology, the equipment data can't be collected in every second. To fulfill missing value, we can use linear interpolation (Equation 1) to quality of data.

$$X = X_l + (X_u - X_l)\frac{t - t_l}{t_u - t_l} \qquad \text{Equation (1)}$$

, where X_l, X_u are the equipment sensor value detected at time t_l, t_u, respectively, and X is equipment sensor value detected at time t between t_l and t_u

(a) Please write the R function to implement linear interpolation to quality the data. The output format like the following figure. (20%)

Save as: SID_Q6.r (e.g s1001234_Q6.r)

Hint:

a=as.POSIXlt(x$Time[2],format="%H:%M:%S")

b=as.POSIXlt(x$Time[1],format="%H:%M:%S")

dt=as.integer(a-b)

Time	A1	A2	A3
00:00:00	100.01	120.51	110.21
00:00:02	100.03	120.55	110.17
00:00:05	100.06	120.58	110.23
00:00:07	100.04	120.54	110.29
00:00:11	100.08	120.62	110.21

Input

```
    Time      A1      A2      A3
00:00:00 100.01 120.51 110.21
00:00:01 100.02 120.53 110.19
00:00:02 100.03 120.55 110.17
00:00:03 100.04 120.56 110.19
00:00:04 100.05 120.57 110.21
00:00:05 100.06 120.58 110.23
00:00:06 100.05 120.56 110.26
00:00:07 100.04 120.54 110.29
00:00:08 100.05 120.56 110.27
00:00:09 100.06 120.58 110.25
00:00:10 100.07 120.60 110.23
00:00:11 100.08 120.62 110.21
```

Output

Part II: Bonus (20%)

1. Please write down your suggestion or advice for this class. (10%)

   Save as: SID_Bonus.txt (e.g s1001234_Bonus1.txt)

2. Please describe your final project.

   (a) Title (2%)

   (b) Your idea or method (8%)

   Save as: SID_Bonus.txt (e.g s1001234_Bonus2.txt)