Homework #2

1. Pizza was original invented in Naples, Italy in the early 19th century. It is a kind of flat bread baked by oven and is usually topped with cheese, tomato sauce, meat and vegetables. Pizza has become a common delicacy around the world.

Suppose the dataset **pizza2.txt** contains data on pizzas sold at a US pizzeria, which could provide insights into factors that influence pizza ratings. The table below shows some key information about the data.

Data	pizza2.txt	
Description	Data about pizza	
Variables descriptions	rating	Rating for the pizza
	cost	Cost per slice
	heat	Heat source used (Gas/Coal/Wood)
	brick	The use of brick oven (TRUE/FALSE)
	area	The location of pizzeria
	heat_re 這是一個錯誤的編碼	Same as the **heat** variable but **it is just numerically coded instead of using strings.** 0 – Coal 1 – Wood 2 – Gas

Tasks:

I. "Using coal to bake pizzas yields different ratings with those baked by using gas or wood".

We wish to verify this statement by providing some statistical evidences:

a. Compute each of the average ratings of the pizzas baked by coal, wood and gas, along with the standard deviations of the ratings. Comment the results. *[hint: you could use codes like **pizza[pizza[,"heat"]=="Coal", ratings]** **OR** **sapply()** and a self-defined function to do so]*

b. Perform an ANOVA test to find out if the ratings of the pizzas baked by different heat sources are equal in average. Comment the results.

c. Fit a simple linear regression by using **rating** as the response variable and **heat** as the predictor variable. Interpret the estimated regression coefficients and the corresponding p-values.
d. Compare and contrast the results in (a), (b) and (c). In other words, what information are shown from both analyses, *OR* from one analysis, but not from the others?

(a 小題是單變量分析, b 小題是 ANOVA, c 小題是迴歸. 所以 d 小題要你比較三種方法得出來的結果)

II. Fit two multiple linear regression by using **rating** as the response variable, and

e. **heat**, **area** and **cost** as the predictor variables.
f. **heat_re**, **area** and **cost** as the predictor variables.

Assume that coal-baked pizzas produce the highest ratings, followed by using wood, and then gas, compare the two models. It is more reasonable to use dummy (indicator) variables in model fitting (as in 1b.), why? Justify your answer by comparing the interpretations of the regression coefficients of **heat** and **heat_re**.

Then, predict the rating for a coal baked pizza that costs $2.50 per slice in LittleItaly and find the corresponding prediction interval using both of the models built in 3a. and 3b. *[hint: use **predict()**]*

("heat" 為 categorical variable, "heat_re" 則是 numerical. 在這題用兩個方法跑迴歸, 再比較兩者結果.)

III. Construct the 95% t-based confidence intervals for the mean rating for each pizzeria location (**area**). Plot **all** of the intervals in a single plot and briefly comment the results. (*Hint: you could make use of **plot(), lines()** and **points()** **OR** search online[1] for some ways to plot confidence intervals.*)

練習畫信賴區間 (後來的 project 可以使用類似方法呈現資料的比較)

[1] *http://stackoverflow.com/questions/14069629/plotting-confidence-intervals*

2. Suppose we are interested in studying the effect of different types of music on people's moods. We collect data on 60 participants and record their mood score (out of 10) after listening to one of three types of music: classical, jazz, or pop. In the file "mood.csv" The data look like

Participant	MusicType	Gender	MoodScore
1	Pop	Female	6.19639294
2	Pop	Female	6.415425525
3	Pop	Female	7.84785533
4	Jazz	Female	6.818517618
5	Pop	Male	7.925675055
6	Jazz	Male	7.888697976
7	Jazz	Female	8.234856128
8	Jazz	Male	7.144747139
9	Pop	Male	6.06780208
10	Classical	Male	7.964462651

Let Y be "MoodScore" and "MusicType" and "Gender" be explanatory variables.

a. Draw side-by-side boxplots to compare the distribution of mood scores based on the music type. Also side-by-side boxplots to compare the distribution of mood scores based on gender.

b. First, test whether the type of music has a significant effect on mood scores while ignoring gender in the model. (Hint: Use a partial F test. Write down the two regression model expressions used in the analysis.)

c. Then, test whether the type of music has a significant effect on mood scores while including gender in the model. (Hint: Use a partial F test. Write down the two regression model expressions used in the analysis.)

d. Finally, test whether there exists an interaction effect between gender and the type of music on mood scores. (Hint: Use a t test. Write down the regression model expression used in the analysis.)