Final Examination

(Introduction to Big Data Science)

2017. 6. 9.

(Big Data Processing)

Q1

3Vs (volume, variety and velocity) are three defining properties of big data. Describe briefly what each of 3Vs refers to.

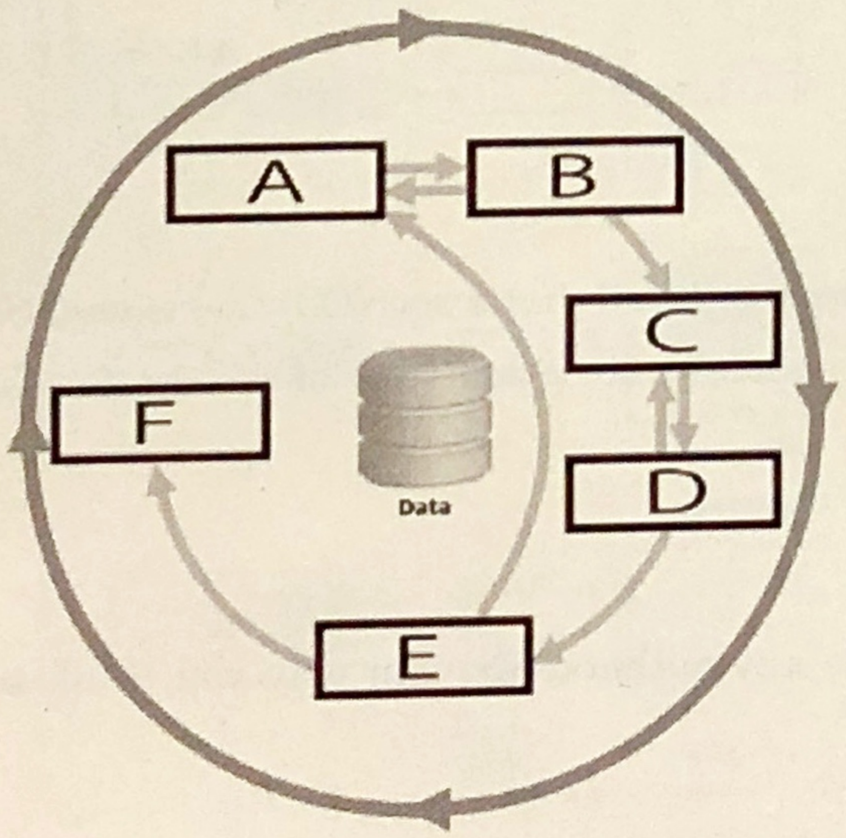
Volume:

Variety:

Velocity:

Q2

Recall that the CRISP-DM process has six phases: Deployment, Data Preparation, Data Understanding, Modeling, Business Understanding, and Evaluation. Write the name of the CRISP-DM phase corresponding to each of A—F boxes in the figure.



A _____

B _____

C _____

D _____

E _____

F _____

Q3

Recall that the CRISP-DM process has six phases: Deployment, Data Preparation, Data Understanding, Modeling, Business Understanding, and Evaluation. For each of the following meetings, explain which phase of the CRISP-DM process is represented:

(a) The data mining project manager meets with the software development manager to discuss implementation of suggested changes and improvements.

(b) The engineer needs to decide data size of outlier.

(c) The analysts meet to discuss whether a regression model or a cluster model should be applied.

Q4

Suppose that the data for analysis includes the attribute *income* as below.

ID	01	02	03	04	05	06	07	08	09	10
income ($)	2M	40K	50K	30K	0	60K	100K	200K	80K	40K

K=1,000, M=1,000,000

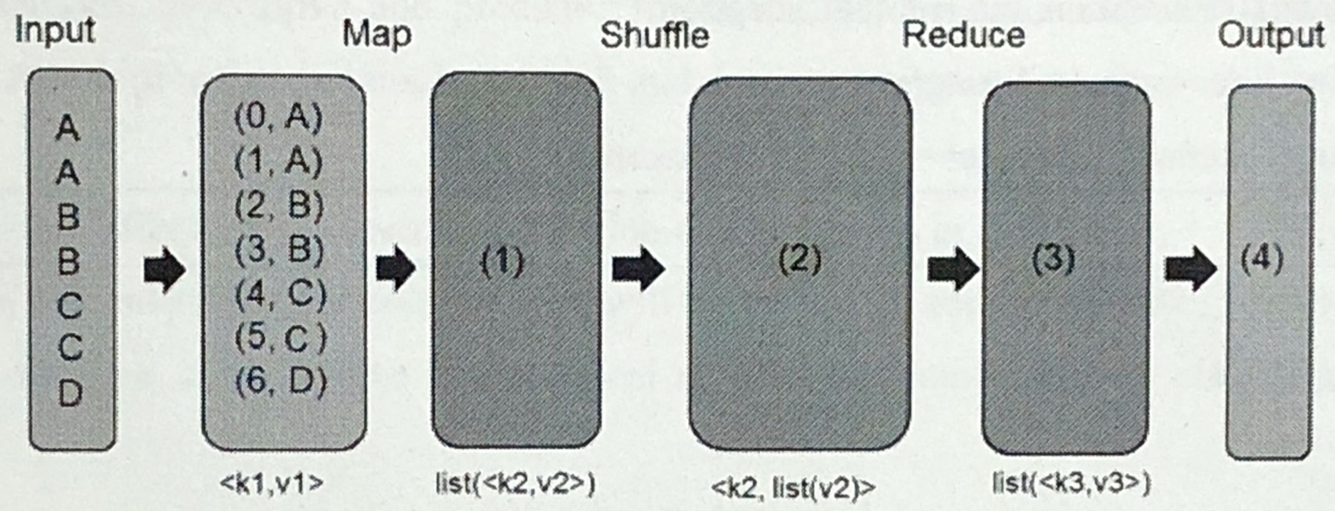
(a) What is the mean of the *income* data?

(b) Compute *z-score* of the value 60K for *income*. Recall that *z-score*(X)=(X – *mean*)/*SD*. Use the mean in (a) and assume that the standard deviation (*SD*) of income for the population is 600K.

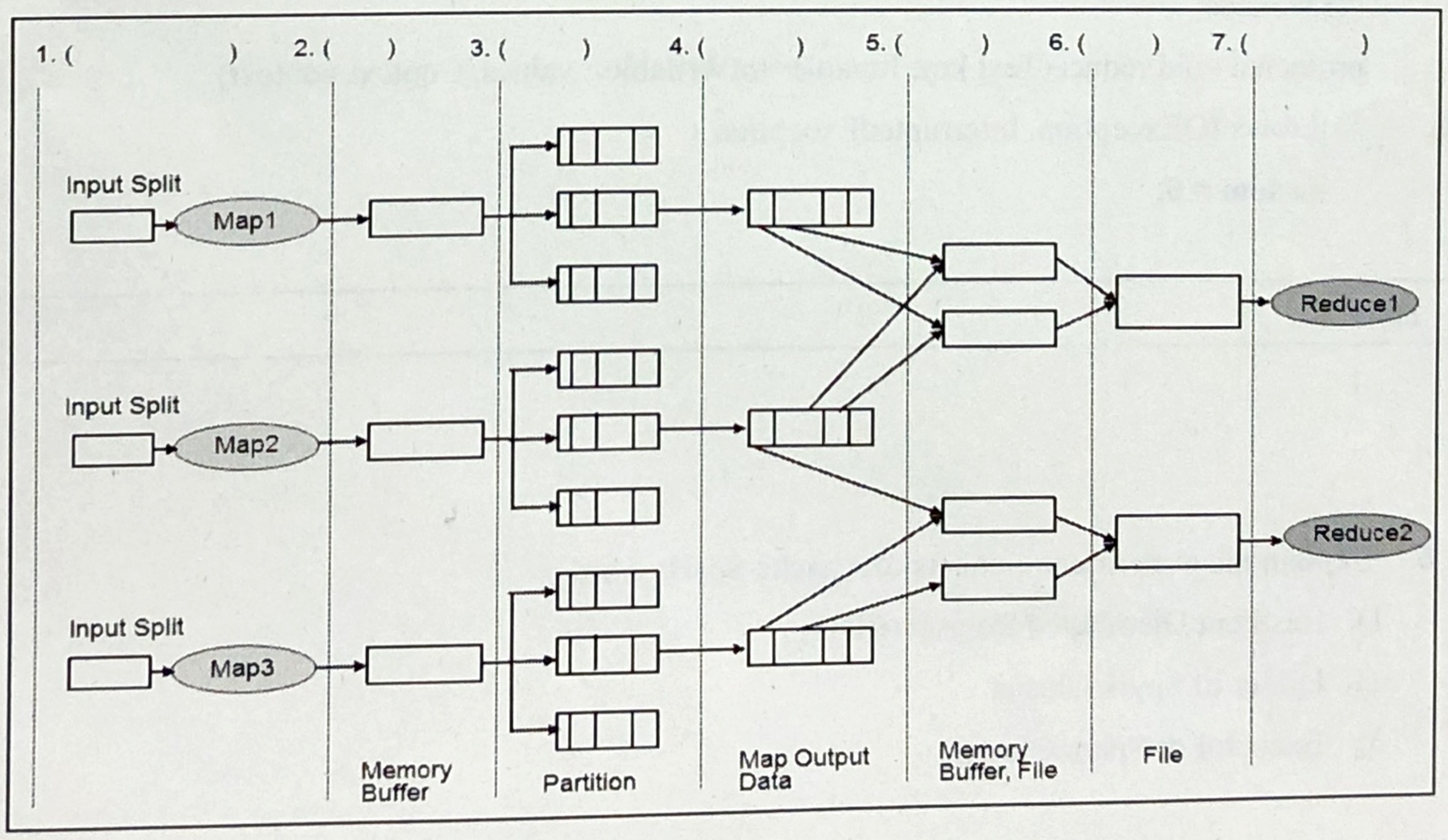
(c) Do you think the *income* data contains any outliers? Explain why you think so. Write all IDs whose income is an outlier.

(Technologies for Big Data)

1. Write characteristics of Storage-Area Network (SAN) and Network Attached Storage (NAS).
2. Write design goal of distributed file system.
3. The figure below is type and data flow of Map-Reduce operation for word count. Fill in the blank rectangular with proper key-value sets.



4. Fill in the blanks with proper process titles for the shuffle process in Map-Reduce operation.



5. The following is a part of Map-Reduce program for word count. Fill in the code blocks for Map-Reduce operation.

```
public class WordCount {
```

```
// Implementation of Reducer
public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
   private final static IntWritable one = new IntWritable(1);
   private Text word = new Text();
   @Override
      protected void map(LongWritable key, Text value, Context context)
      throws IOException, InterruptedException {
      String line = value.toString();
      StringTokenizer tokenizer = new StringTokenizer(line);
(1)
      }
   }
}

// Implementation of Reducer
public static class Reduce extends    Reducer<Text, IntWritable, Text, IntWritable> {
   private IntWritable value = new IntWritable(0);

   @Override
   protected void reduce(Text key, Iterable<IntWritable> values, Context context)
      throws IOException, InterruptedException {
      int sum = 0;

(2)
   }
}
```

6. Explain the term or components of Apache Spark shortly.
   1) Resilient Distributed Dataset (RDD)
   2) Driver of Spark Cluster
   3) Executor of Spark Cluster

7. Explain two winters in neural network of AI shortly.

8. Explain problem issues and their alternatives of neural network for deep learning shortly.

(Data Mining)

Q1. Please give a definition to the data mining, and identify the its difference between artificial intelligence, machine learning, and deep learning. An instance or conceptual figure that shows your consideration will be expected.

Q2. Two methods, descriptive and predictive methods, are the fundamental of the data mining approaches. Please give three examples of each and describe how the methods work.

Q3. What is the difference between classification and clustering in data mining? A simple and clear example that demonstrates your point of view will be helpful. In addition, please give three methods, with details on their concepts and mathematical models, that implement the classification in data mining.

Q4. Please give a definition on sharing economy using one of the successful examples.

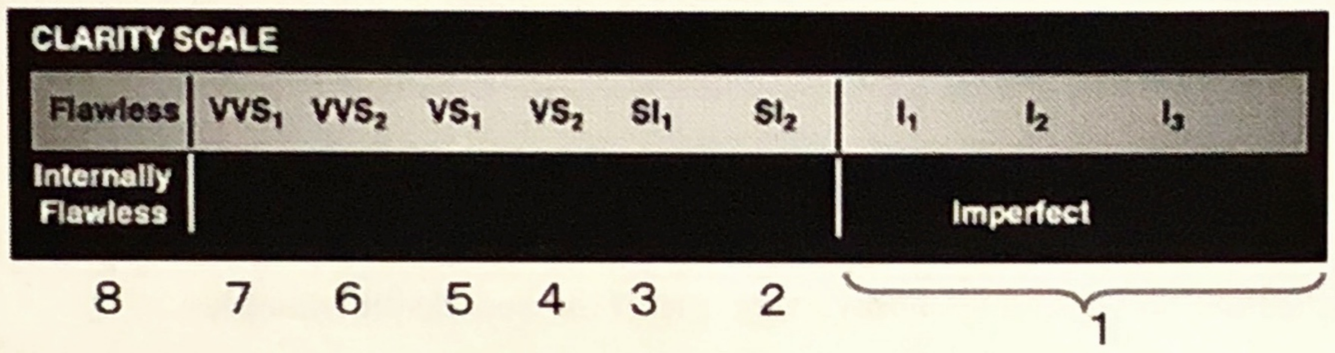
Exam problem (Ofuji part, 2017AY)

Below is the six records of the diamond price dataset **Ddat** which we took up in class,

```
> head(Ddat)
  MAN_YEN CARAT CUT CLARITY
1   22.87   0.5   3       6
2   22.64   0.5   3       5
3   22.08   0.5   3       5
4   22.08   0.5   3       5
5   22.64   0.5   3       5
6   22.64   0.5   3       5
```

where :

MAN_YEN: Price of a diamond in man (10,000) yen
CARAT: Number of carats (1 carat=0.2 grams)
CUT: Idealness rating of cut (3: Ideal, 2: Good, 1: Fair)
CLARITY: 8-level scale as below



and the dataset does not have records with **CLARITY**=1. (In other words, all diamond samples in this dataset have **CLARITY** of 2 or greater). Assume that the entire dataset contains 500 records as we saw in class.

Using this dataset, let's suppose that you obtained the following multiple regression estimates **model1**, using R's **lm** estimation command (decimal numbers are truncated for simplicity):

```
Coefficients:
            Estimate Std. Error   t value   Pr(>|t|)
(Intercept) -123     3.9          -31.3    < 2e-16   ***
CARAT        122     1.3           95.9    < 2e-16   ***
CUT          4.8     0.8           6.1      2.7e-09 ***
CLARITY      12.1    0.5           24.2     < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.65 on 496 degrees of freedom
Multiple R-squared:  0.9536, Adjusted R-squared:  0.9533
F-statistic:  3395 on 3 and 496 DF,  p-value: < 2.2e-16
```

Q1: Based on this regression result, what is the expected average impact on the diamond price in man-yen, from a one-unit increase in **CUT** ?

Q2: Suppose that you have a diamond sample with values of **CARAT** = 1, **CUT** = 2, and **CLARITY** = 4. What is the expected price of this diamond sample in man-yen ?

122
+ 58
180

9.6
+48.4
58.0

Q3: Assume that you, as a diamond retailer, are wondering whether you should apply extra cutting work on this diamond sample, so that, in exchange for reducing the **CARAT** to 0.9, the **CUT** and **CLARITY** values be improved. Suppose that, after the cutting work you are sure that the **CUT** value will result to 3, and **CLARITY** value to 5.

(a) Calculate the expected price in man-yen after performing this extra cutting work.

(b) Comparing it with the price you answered in Q2, if you want higher diamond prices, should you cut the diamond or not ?

Next suppose you built another model, **model2**, that treated **CUT** and **CLARITY** as factor variables. This model will be expressed in a mathematical form as below:

$$\begin{aligned} MAN\_YEN = c \;&+ b_1 CARAT \;+\; b_{2(2)}(CUT{=}2) \;+\; b_{2(3)}(CUT{=}3) \\ &+ b_{3(3)}(CLARITY{=}3) \;+ b_{3(4)}(CLARITY{=}4) \;+ b_{3(3)}(CLARITY{=}5) \\ &+ b_{3(6)}(CLARITY{=}6) \;+ b_{3(7)}(CLARITY{=}7) \;+ b_{3(8)}(CLARITY{=}8) \;+ e \end{aligned}$$

where

c : intercept,
$b.$: regression coefficients,
e : random error,

and the binary variable (CUT=2), for example, takes the value of 1 when **CUT**=2, and 0 otherwise.

Suppose your **model2** estimates in R looked like below (decimal numbers are again truncated for simplicity):

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         -127     3.5       -36.8   < 2e-16 ***
CARAT                126     1.1        116    < 2e-16 ***
as.factor(CUT)2      0.9     1.2        0.8    0.45
as.factor(CUT)3      8.7     1.3        6.7    8.6e-11 ***
as.factor(CLARITY)3  41      2.7        15.3   < 2e-16 ***
as.factor(CLARITY)4  59      2.7        21.8   < 2e-16 ***
as.factor(CLARITY)5  71      2.8        25.6   < 2e-16 ***
as.factor(CLARITY)6  74      3.2        23.4   < 2e-16 ***
as.factor(CLARITY)7  77      5.9        13.1   < 2e-16 ***
as.factor(CLARITY)8  93      3.5        26.7   < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.42 on 490 degrees of freedom
Multiple R-squared:  0.9689,  Adjusted R-squared:  0.9683
F-statistic:  1697 on 9 and 490 DF,  p-value: < 2.2e-16
```

Q4: Based on this result,

(a) What is the expected increase in price, in man yen, for an increase in **CLARITY** from 4 to 5 ?

74 × 6 = 444

77 × 7 = 539

539 − 444 = 95

(b) What about for a **CLARITY** increase from 6 to 7 ?

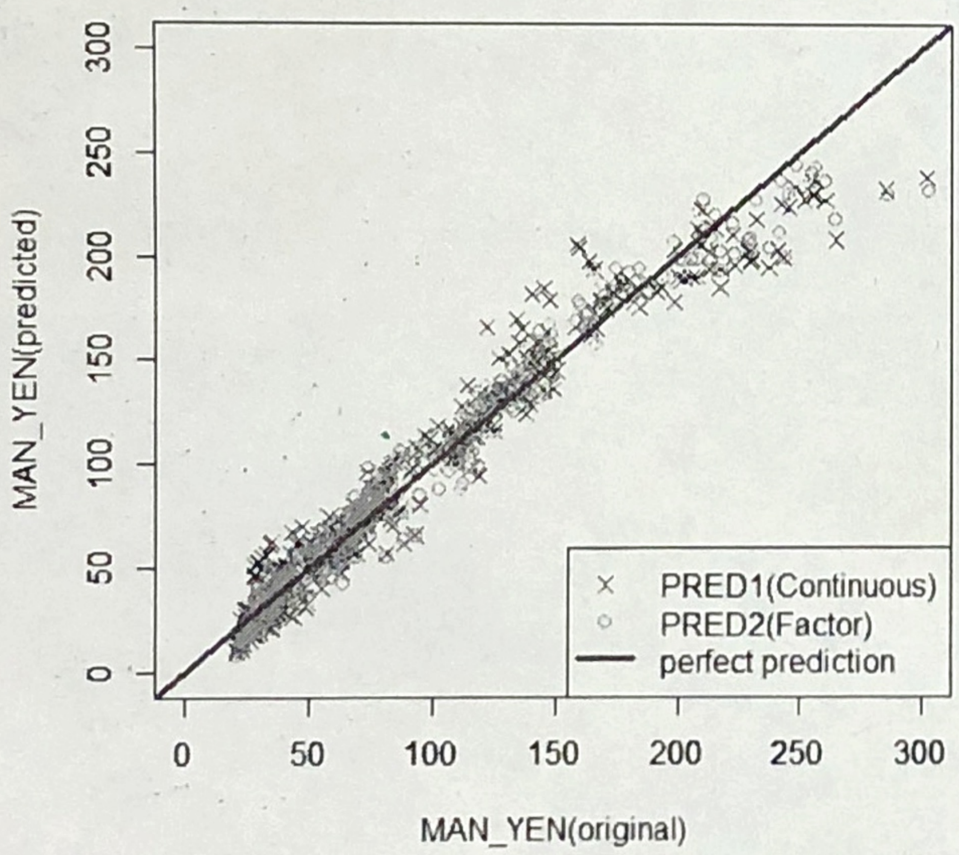
(c) What about for a **CUT** increase from 1 to 2 ? Should you view this as "statistically significant" in the conventional criteria, where you need a p-value of better than 10% ?

Q5: Re-evaluate, using this model2, the expected price of the diamond that you considered in Q3(a). How much do you expect the price to be, in man-yen, after the extra cutting work? Use the **CARAT**, **CUT** and **CLARITY** values given in Q3.

Q6: You may have obtained a different price in Q5 than in Q3(a). Suppose your AICs of the two models were :

```
> AIC(model1)
[1] 3962.881
> AIC(model2)
[1] 3774.195
```

and also suppose you generated **MAN_YEN**'s prediction values **PRED1** from **model1**, and **PRED2** from **model2**, to draw an original-prediction plot like below:



126×0.9+8.7×3+71×5 =

12.6 × 9 = 113.4

8.7 × 3 = 26.1

71 × 5 = 355

113.4 + 26.1 = 139.5

139.5 + 355.0 = 494.5

Based on the AICs and the plot, briefly discuss which expected **MAN_YEN** value you can be more confident of, the one you got in Q5 or the one in Q3(a)