Course: DATA 606
Homework 1
Student Name: Lung Tze Fung
ID: 23637639

Chapter 1 - Introduction to Data

- Practice: 1.7 (available in R using the `data(iris)` command), 1.9, 1.23, 1.33, 1.55, 1.69
- Graded: 1.8, 1.10, 1.28, 1.36, 1.48, 1.50, 1.56, 1.70

1.8 a

Answer:

```
> smoking
# A tibble: 1,691 × 12
   gender   age maritalStatus highestQualification nationality ethni
city      grossIncome
    <chr> <int>         <chr>                <chr>       <chr>     <
chr>            <chr>
1     Male    38      Divorced      No Qualification     British     W
hite   2,600 to 5,200
2   Female    42        Single      No Qualification     British     W
hite      Under 2,600
3     Male    40       Married                Degree     English     W
hite 28,600 to 36,400
4   Female    40       Married                Degree     English     W
hite 10,400 to 15,600
5   Female    39       Married          GCSE/O Level     British     W
hite   2,600 to 5,200
```

……

```
1.8b
Answer:
> summary(smoking)
    gender               age          maritalStatus      highestQualif
ication nationality
 Length:1691        Min.   :16.00   Length:1691        Length:1691
        Length:1691
 Class :character   1st Qu.:34.00   Class :character   Class :charac
ter     Class :character
 Mode  :character   Median :48.00   Mode  :character   Mode  :charac
ter     Mode  :character
```

```
                    Mean   :49.84

                    3rd Qu.:65.50

                    Max.   :97.00



  ethnicity          grossIncome            region             smoke
          amtWeekends
 Length:1691        Length:1691        Length:1691        Length:169
1         Min.   : 0.00
 Class :character   Class :character   Class :character   Class :cha
racter   1st Qu.:10.00
 Mode  :character   Mode  :character   Mode  :character   Mode  :cha
racter   Median :15.00

         Mean   :16.41

         3rd Qu.:20.00

         Max.   :60.00

         NA's   :1270
  amtWeekdays         type
 Min.   : 0.00   Length:1691
 1st Qu.: 7.00   Class :character
 Median :12.00   Mode  :character
 Mean   :13.75
 3rd Qu.:20.00
 Max.   :55.00
 NA's   :1270
```

1.10 Cheaters, scope of inference.

(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

Answer:

a)

The age from 5-15 who may have an intention to get the reward by reporting white, it depend on the probabilities to get white, intention to cheat and if they get the instruction not to cheat.

b)

The result can not be generalized to the population, because it have intention to cheat to get the reward in nature.

The study will be tend to cheat in order to get the reward, the causal relationship like as follow:

Students	Ages	Outcome	Reward	Cheat	Instruction
1	5	White	Yes	No	Yes
2	10	Black	Yes	Yes	Yes
3	15	White	Yes	No	No
4s	8	Black	No	No	Yes

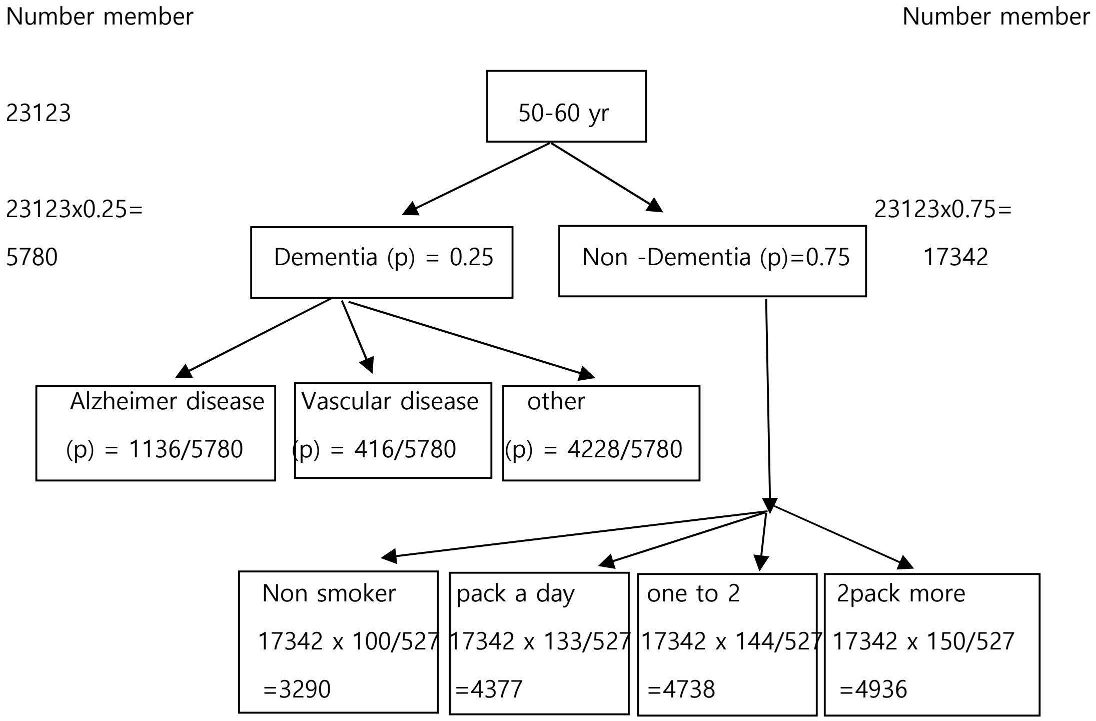
Course: DATA 606
Homework 1
Student Name: Lung Tze Fung
ID: 23637639

1.28 Reading the paper.

Below are excerpts from two articles published in the NY Times:

(a) Answer:

Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning.



If x = non smoker

Y = pack a day smoker

Z = one to 2 pack

W = 2 pack more =

X : Y : Z : W = 100 : 133 : 144 : 150

Ratio:

X = 100 / 527

Y = 133/527

Course: DATA 606
Homework 1
Student Name: Lung Tze Fung
ID: 23637639

Z = 144/527

W = 150/527

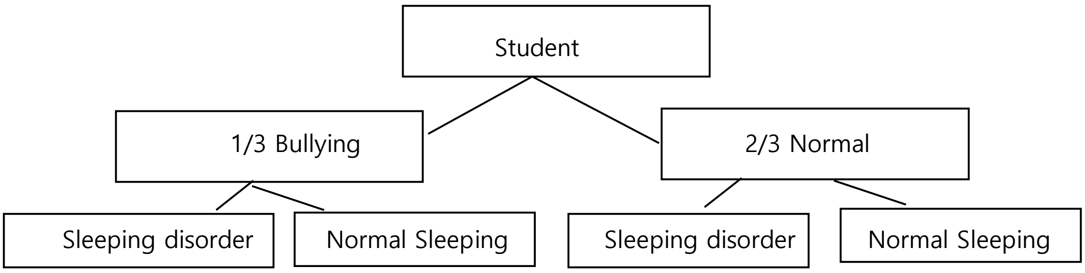
According to the probability from 23123, if all non-smoker was without dementia around 17342 person, but smoker make more (4377+4738+4936) 14051 person, In conclusion, smoking cause dementia in life.

(b) Another article titled The School Bully Is Sleepy states the following:[62]

The statement ”sleep disorders lead to bullying in school children is not justified.

The best describe the conclusion should be that ”Bullying in school children is relative to sleep disorders.
Because it does not mention the proportion of Sleeping disorder with normal student, bullying are twice as sleeping disorder that does not represent any in this case.

From the below analysis:



1.36 Exercise and mental health.

(a) What type of study is this?

Mental health exam for different country.

(b) What are the treatment and control groups in this study?

(c) Does this study make use of blocking? If so, what is the blocking variable?

Yes, this study blocking some information about the subject of exercise and mental health. From the data shown, it cannot reflect any information if only checking the data.

(d) Does this study make use of blinding?

Yes, this study make use of blinding for hiding the critical information of mental health.

(e) Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether or not the conclusions can
be generalized to the population at large.

No, the result of the study can not be used to be reflected between exercise and mental health by using only data.

The Conclusion can not be generalized to the population at large, because the topic is not easy to be reflected for exercise and mental health.

(f) Suppose you are given the task of determining if this proposed study should get funding.

Would you have any reservations about the study proposal?

1.48 Stats scores.

Below are the final exam scores of twenty introductory statistics students.

57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

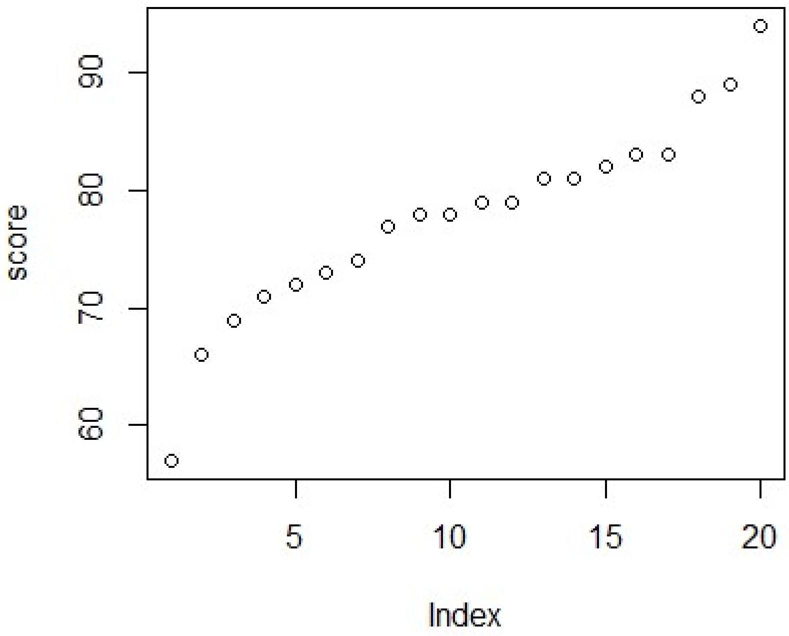
Min Q1 Q2 (Median) Q3 Max

57 72.5 78.5 82.5 94

Answer:

```
> score <- c(57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94)
> plot (score)
```

Course: DATA 606
Homework 1
Student Name: Lung Tze Fung
ID: 23637639



```
> quantile (score)
   0%    25%    50%    75%   100%
57.00 72.75 78.50 82.25 94.00
```

1.50 Mix-and-match.

Describe the distribution in the histograms below and match them to the box plots.

a) The group between 50 and 70, appropriate center is 60, the max number is 60.
b) The group between 0 and 100, the maximum is 40, and the minimum is 30.
c) The group between 0 and 6, appropriate center is 1, the max number is 1.
1) The group between 0 and 3.8, appropriate center is 1.5.
2) The group between 50 and 67, appropriate center is 60.
3) The group between 0 and 100, appropriate center is 50.

1.56 Distributions and appropriate statistics, Part II .

Course: DATA 606
Homework 1
Student Name: Lung Tze Fung
ID: 23637639

(a) Housing prices in a country where 25% of the houses cost below $350,000, 50% of the houses cost below $450,000, 75% of the houses cost below $1,000,000 and there are a meaningful number of houses that cost more than $6,000,000.

(b) Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000 and very few houses that cost more than $1,200,000.

(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

(d) Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than the all other employees.

a)

Q1=$350,000

Q2=$450,000

Q3=$1,000,000

IQR = 1000000-350000=650000

More than $6,000,000 does not show the meaningful data.

It would be best represented using the standard deviation or IQR.

b)

Q1=$300,000

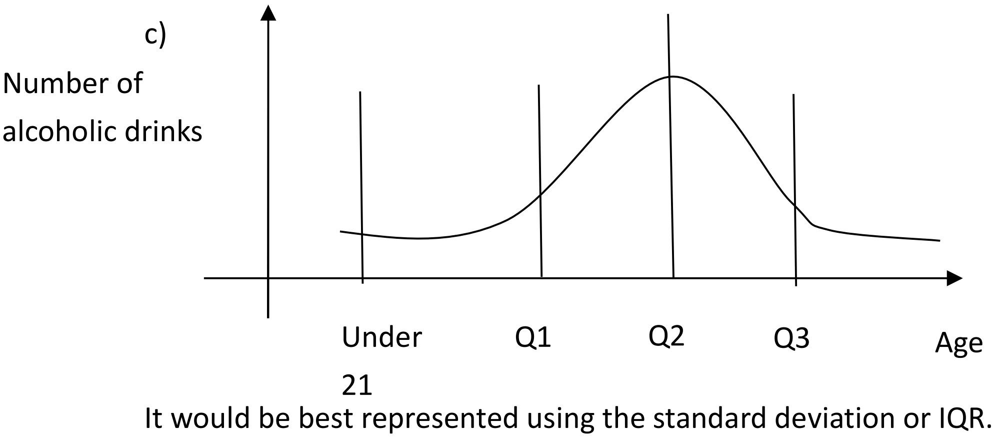
Q2=$600,000

Q3=$900,000

IQR=900000-300000=600000

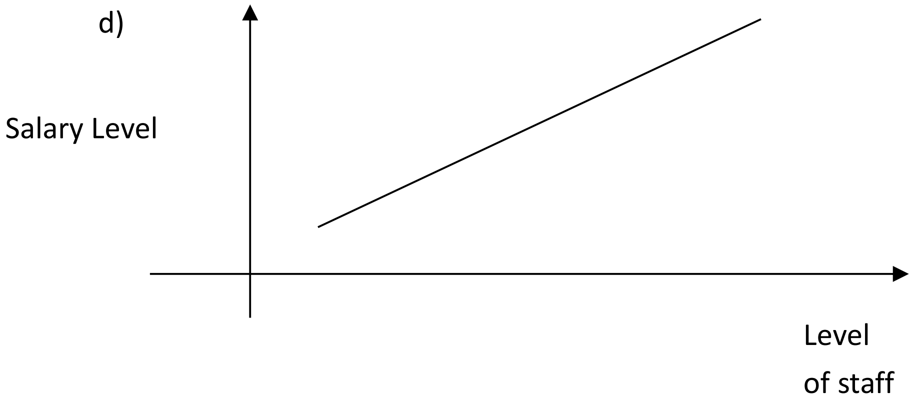
It would be best represented using the standard deviation or IQR.

c)



It would be best represented using the standard deviation or IQR.

d)



It would not be best represented using the standard deviation or IQR.

1.70 Heart transplants.

(a) Based on the masaic plot, it is definitely the survival independent on patient who got a transplant, For example, the control Q2 (mean of control) is less than 100 days on survival time. For the treatment, the mean is more than 100 days and Q3 (75%) is larger than 500 days on survival time.

(b) The effectiveness of heart transplant treatment are suggested according to the survival time increasing.

(c) More than 80% proportion of patients in the control group died.

Around 30% proportion of patients in treatment group died.

(di) The survival time should be tested to investigate whether treatment is effective.

(dii)

iii) The mean of simulation is 0, it does not show the effectiveness of the transplant program.