DATA606 Homework for 9-2-2018

John K. Hancock

September 1, 2018

1.8 Smoking Habits of UK Residents

Answers: (a) Each row in the data matrix represents an observation collected through a survey.
(b) There were 1,691 participants in the survey.
(c) The variables in the survey are as follows:
Sex, a Categorical, Nominal variable
Age, a Numerical, Discrete variable
Marital Status, a Categorical, Nominal variable
Gross Income, a Categorical, Ordinal variable
Smoke, a Categorical, Nominal variable
amtWeekends,a Categorical, Ordinal variable
amtWeekdays,a Categorical, Ordinal variable

1.10 Cheaters Scope of Influence

Answers:
(a) The population of interest in the study were children between the ages of 5 and 15. To make inferences about this population, the researchers selected a sample of 160 children between the ages of 5 and 15.
(b) Given the limited amount of information as to how the sample was selected, where the sample was collected from, unknown biases in the sample, and the amount of observations in the sample, the sample is not adequate to make any generalizations about the population. The findings should not be used to infer any casual relationships between honesty, age, and self-control.

1.28 Reading the Paper

Answers:
(a) Assumming that the research methods are sound, we cannot conclude from the results that smoking Causes demntia later in life. The survey shows a strong correlation between different levels of smoking and increased levels of dimentia later in life. This study shows a correlation between observed smoking levels and dimentia. To prove causation, one would have to design an experiment where a non-smoker would smoke and the effects of smoking could be directly linked to dimentia.
(b) The conclusion drawn from the study shows a correlation between sleep disorders and bullying. It does not show that sleep disorders are the reason or cause for bullying. There could be a confounding variable that's correlated with both sleep disorders and bullying which can better explain the bullying. For example, children experiencing extreme domestic stress and anxiety at home could answer both sleep disorders and bullying.

1.36 Exercise and Mental Health

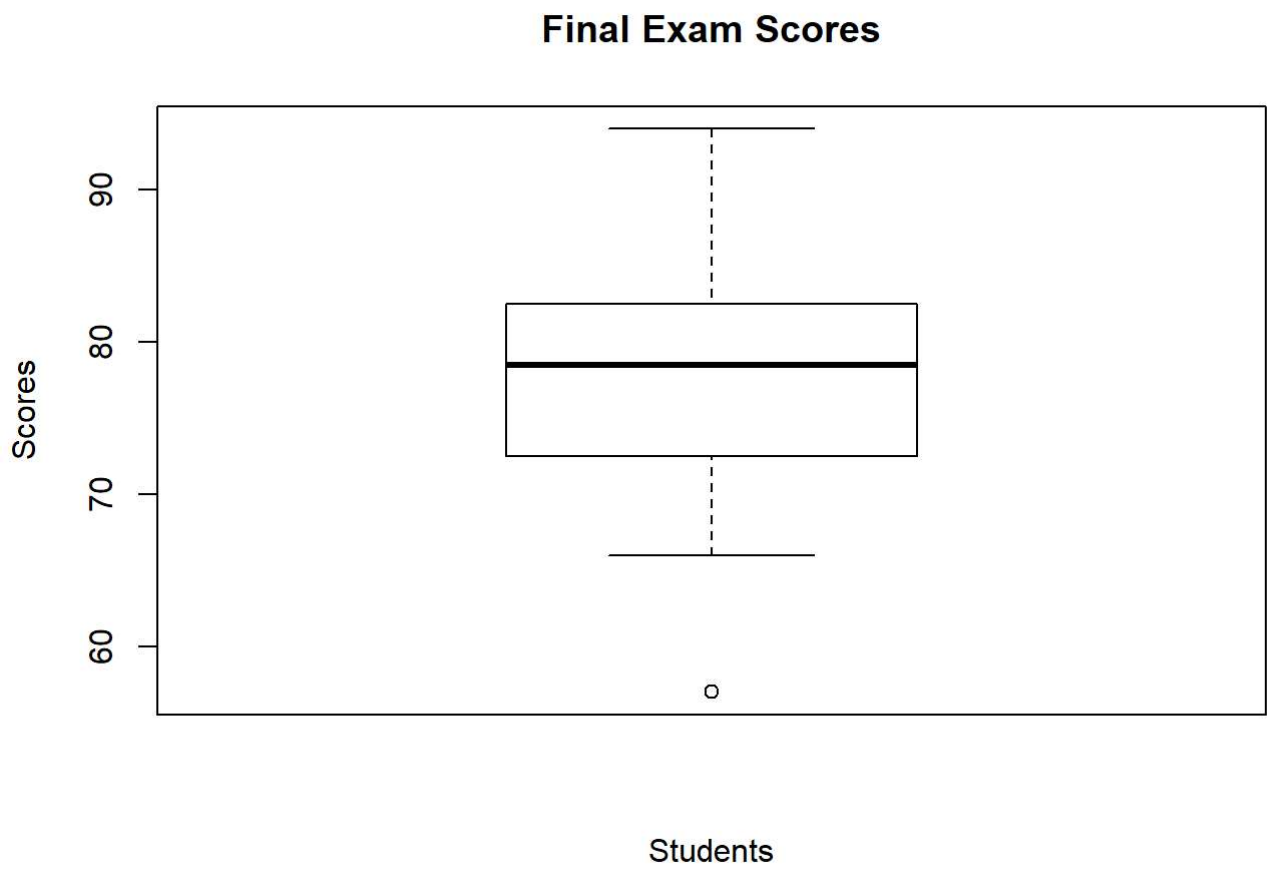
Answers:
(a) This is an experiment.
(b) Treatment group is the one that exercises twice a week. The Control group is the one instructed to do nothing.
(c) Yes. Age is used for blocking.
(d) Yes. Single blinding. The researchers know which group received treatment.
(e) Yes, assuming the overall sample size is representative of the popultation, the process of using a randomm, stratified sample based on age can be used to make generalizations about the overall population because age

bias has been accounted for in the sample.
(f) Yes. If the overall size of the sample is representative of the population and all groups are given an equal chance of participating in the sample.

1.48 Stats scores.

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

boxplot(x=scores, data=scores, main="Final Exam Scores",
    xlab="Students", ylab="Scores")
```



####1.50 Mix and Match

a. Is a symmetrical, unimodal distribution and matches boxplot at 2.
b. Is a uniform, multimodal distribution and matches boxplot at 3.
c. Is a right skewed, unimodal distribution and matches boxplot at 1.

1.56 Distributions and appropriate statistics, Part II.

a. Based on the plots of hypothetical data below, the data skews to the right due to the higher prices above 6,000,000. Given the skewness, the median would be best at representing the price of a typical home. Finally, the IQR would be better at explaining variability than standard deviation also due to the higher prices of a few homes.

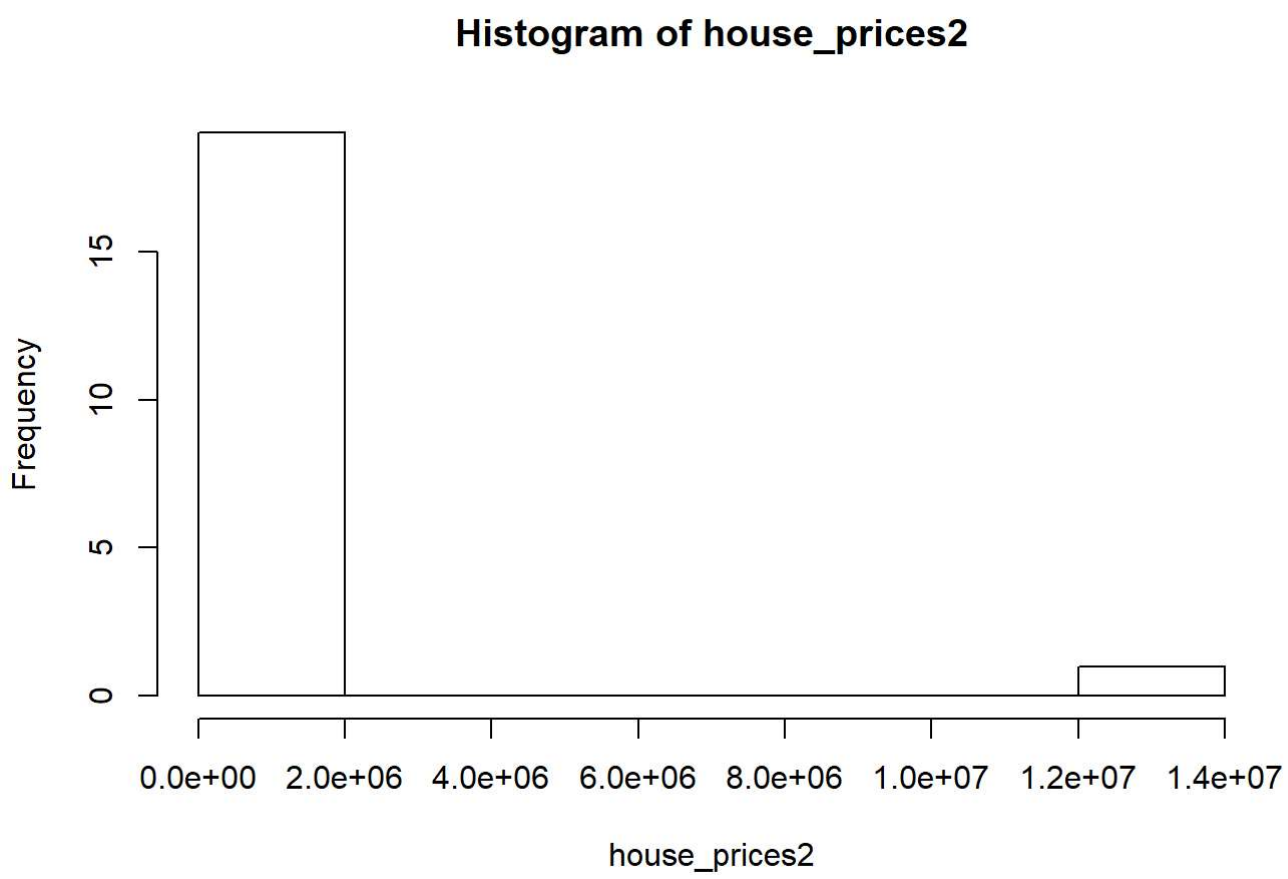
```
house_prices <-c(350000,325000,300000,280000,450000,410000,420000,400000,1000000,980000,900000,8
00000,6000000,6250000,6500000,6750000,10000000)
boxplot(x=house_prices , data=house_prices , main="Housing Prices",
    xlab="Houses", ylab="Housing Prices", ylim=c(200000, 10000000))
```


```
hist(house_prices)
```

(b)Based on the plots of hypothetical data below, the data skews to the right due to the higher prices above 1,200,000. Given the skewness, the median would be best at representing the price of a typical home. Finally, the IQR would be better at explaining variability than standard deviation also due to the higher prices of a few homes.

```
house_prices2 <-c(300000,300000,290000,280000,600000,500000,420000,400000,900000,880000,890000,8
00000,880000,890000,800000,720000,700000,750000,1250000,13500000)
boxplot(x=house_prices2 , data=house_prices2 , main="Housing Prices2",
    xlab="Houses2", ylab="Housing Prices2", ylim=c(10000, 2000000))
```


```
hist(house_prices2)
```



c. Given the age constraint, the data will be skewed to the right. Because of the skew, the median would be a better representation, and the IQR would be better at explaining variability.

d. Given that a few executives will make most of the money, annual salaries of the employees will be skewed to the right, the median and IQR are better at representing the data and explaining variability.

1.70 Heart Transplants

```
library(openintro)
```

```
## Please visit openintro.org for free statistics materials
```

```
##
## Attaching package: 'openintro'
```

```
## The following objects are masked from 'package:datasets':
##
##     cars, trees
```

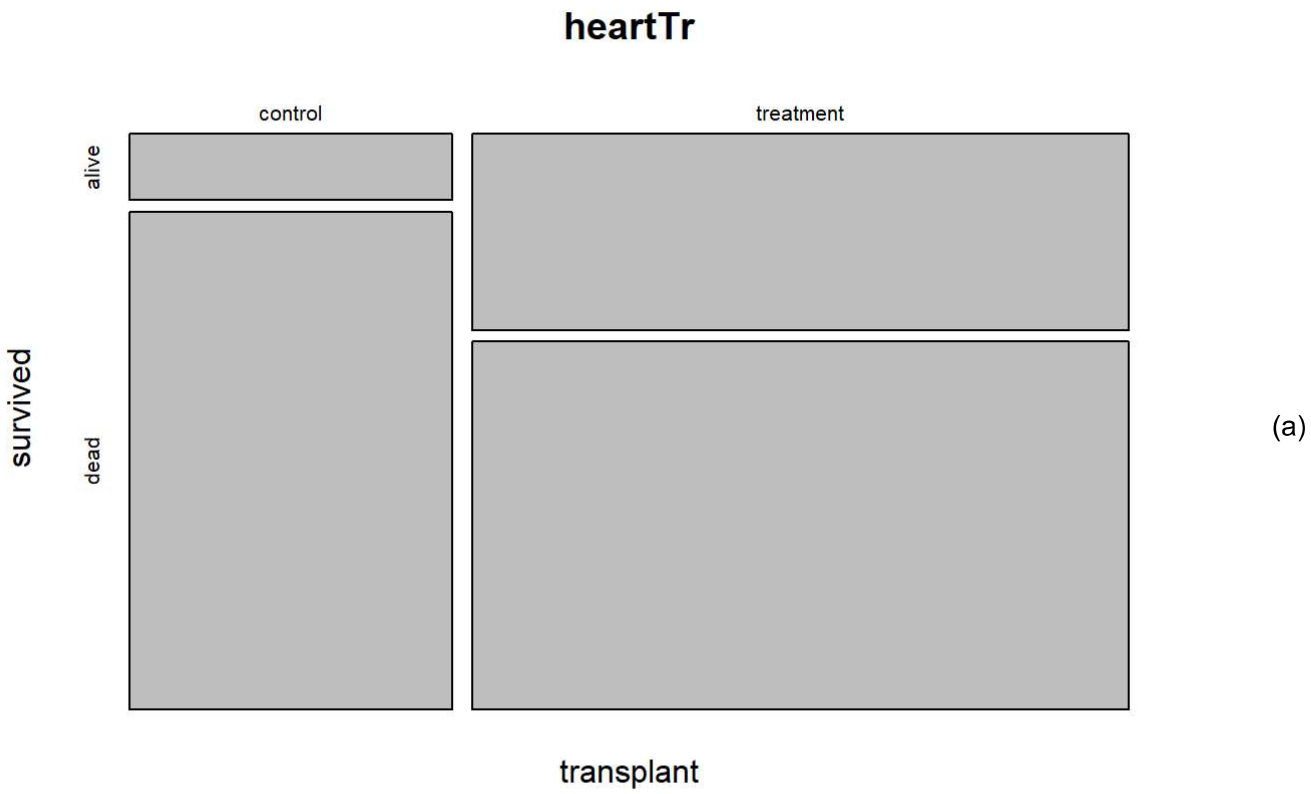
```
data(heartTr)
str(heartTr)
```

```
## 'data.frame':    103 obs. of  8 variables:
##  $ id        : int  15 43 61 75 6 42 54 38 85 2 ...
##  $ acceptyear: int  68 70 71 72 68 70 71 70 73 68 ...
##  $ age       : int  53 43 52 52 54 36 47 41 47 51 ...
##  $ survived  : Factor w/ 2 levels "alive","dead": 2 2 2 2 2 2 2 2 2 2 ...
##  $ survtime  : int  1 2 2 2 3 3 3 5 5 6 ...
##  $ prior     : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ transplant: Factor w/ 2 levels "control","treatment": 1 1 1 1 1 1 1 2 1 1 ...
##  $ wait      : int  NA NA NA NA NA NA NA 5 NA NA ...
```

```
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.5.1
```

```
mosaicplot(~ transplant + survived, data = heartTr)
```



(a)

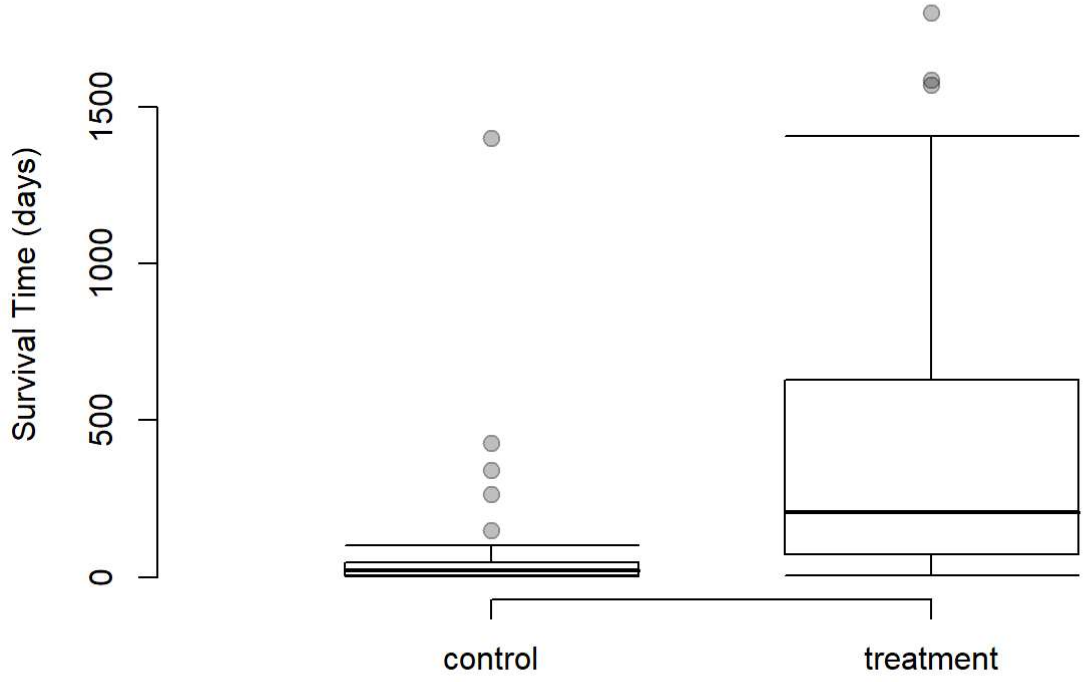
Based on the mosaic plot, the height and width of the treament/alive area is greater than the height/width of the control/alive area. One could conclude, that survival is greater if you were in the treatment group. In the frequency table below, 35% in the treatment group are alive while only 12% are alive in the control group.

```
count(heartTr, vars=c("transplant","survived"))
```

```
##   transplant survived freq
## 1    control    alive    4
## 2    control     dead   30
## 3  treatment    alive   24
## 4  treatment     dead   45
```

b. The boxplot shows that the median survival time in days is higer for the treatment group. The data is skewed so the median is a better measure. Also, the control group had a number of outliers which also skew the results.

```
boxPlot(heartTr$survtime, heartTr$transplant,
    ylab = 'Survival Time (days)')
```



(c)

Approximately 65% of those in the treatment group died while 88% in the control group died.

d. i. The claim or null hypothesis being tested is that the heart transplant will not extend the survival time for patients that received them versus those that do not receive them. Any extension of survival time is due to error or chance.

ii. 28 cards 75 cards 69 representing treatment 34 representing control 0 distribution high/low

iii. The simulations show that the fraction simulations where the simulated di↵erences in proportions are indeed low. Therefore, we can reject the null hypothesis that the treatment does extend longevity and that the results are not just by chance.