

Homework 1 Solutions

1) Read Chapter 1 (all) and Chapter 2 (only sections 2.1, 2.2 and 2.3).

2) Many answers are possible.

3)

(a) Time in terms of AM or PM.

Binary, qualitative, nominal (most people consider binary attributes to be nominal)

(b) Brightness as measured by a light meter.

Continuous, quantitative, ratio

(c) Brightness as measured by people's judgments.

Discrete, qualitative, ordinal (assuming we make them choose from a discrete set of ratings)

(d) Angles as measured in degrees between 0° and 360° .

Continuous, quantitative, ratio

(e) Bronze, Silver, and Gold medals as awarded at the Olympics.

Discrete, qualitative, ordinal

f) Height above sea level.

Continuous, quantitative, interval/ratio (depends on whether sea level is regarded as an arbitrary origin)

(g) Number of patients in a hospital.

Discrete, quantitative, ratio

(h) ISBN numbers for books. (Look up the format on the Web.)

Discrete, qualitative, nominal (but ISBN numbers do have some order information so it could be ordinal if you use that information)

(i) Ability to pass light in terms of the following values: opaque, translucent, transparent.

Discrete, qualitative, ordinal

(j) Military rank.

Discrete, qualitative, ordinal

(k) Distance from the center of campus.

Continuous, quantitative, interval/ratio (depends)

(l) Density of a substance in grams per cubic centimeter.

Continuous, quantitative, ratio

(m) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)

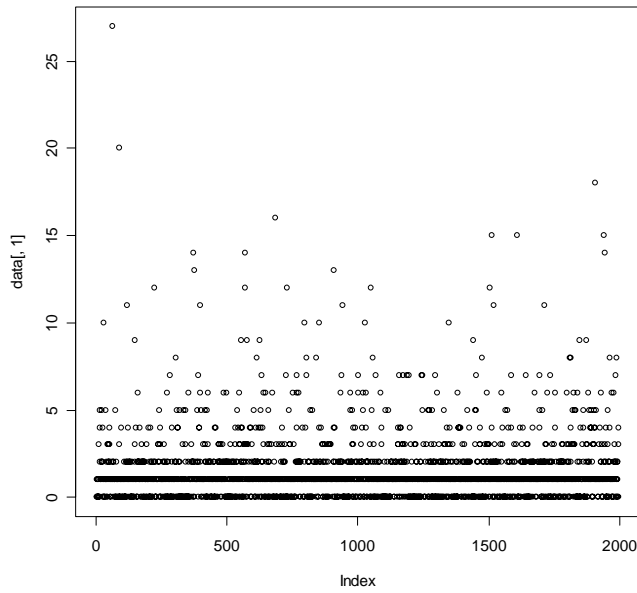
Discrete, qualitative, nominal (or ordinal if you are using the order information)

4)

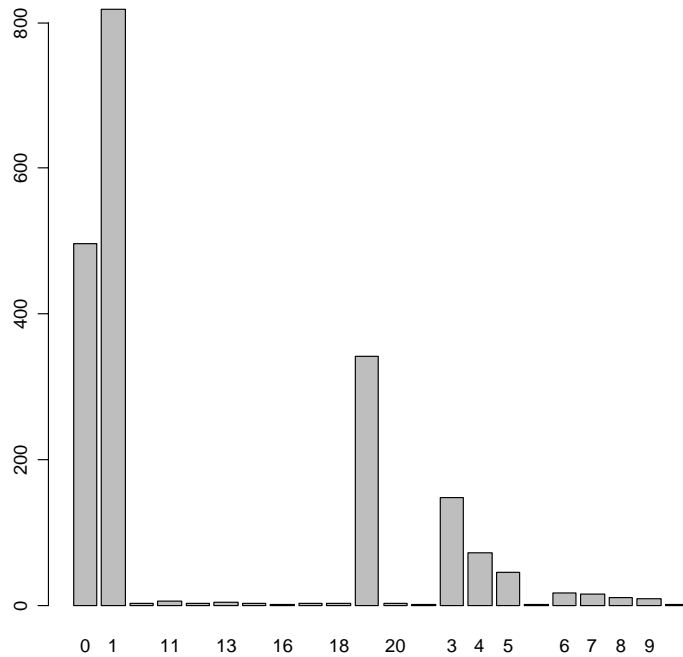
a) The first column is quantitative (numeric) while the second column is qualitative (categorical). There are many ways to tell in R but one is to use the `is.factor()` function. This will indicate that the first column is not a factor but that the second column is a factor.

b) The observation in row 1463 says “two” instead of being a number.

c) The plot for column 1 shows the row numbers on the x axis and the column 1 values on the y axis. A point is drawn for each row.



The plot for column 2 shows all the possible values in column 2 on the x axis and the bar height on the y axis gives the frequency of occurrence.



d) Excel says “#VALUE!”.

5)

a)

```
> data<-read.csv("twomillion.csv",header=F)
> sam<-sample(seq(1,2000000),10000,replace=TRUE)
> sample_data<-data[sam,]
```

b) Answers will vary (depending on what 10,000 observations are in your sample).

```
> mean(sample_data)
[1] 9.460049

> max(sample_data)
[1] 17.43476

> var(sample_data)
[1] 4.096144

> quantile(sample_data,.25)
      25%
8.121058
```

c) You should have the answers below exactly. These should not differ much from your answers in part b, with the exception of the maximum.

```
> mean(data[,1])
[1] 9.451468

> max(data[,1])
[1] 18.96657

> var(data[,1])
[1] 4.001822

> quantile(data[,1],.25)
      25%
8.10388
```

d) Note, again your answers will vary depending on what 10,000 observations are in your sample.

The Excel function for mean is “AVERAGE”. The result is 9.460048737.

The Excel function for maximum is “MAX”. The result is 17.43476417.

The Excel function for variance is “VAR”. The results is 4.096143627.

The Excel function for the first quartile is “QUARTILE” with the argument “1”. The result is 8.121057662.

e) Excel complains “File not loaded completely” and then loads only the first 65,536 rows (or 1,048,576 in Excel 2007).

6) Read Chapter 3 (only sections 3.1, 3.2 and 3.3).

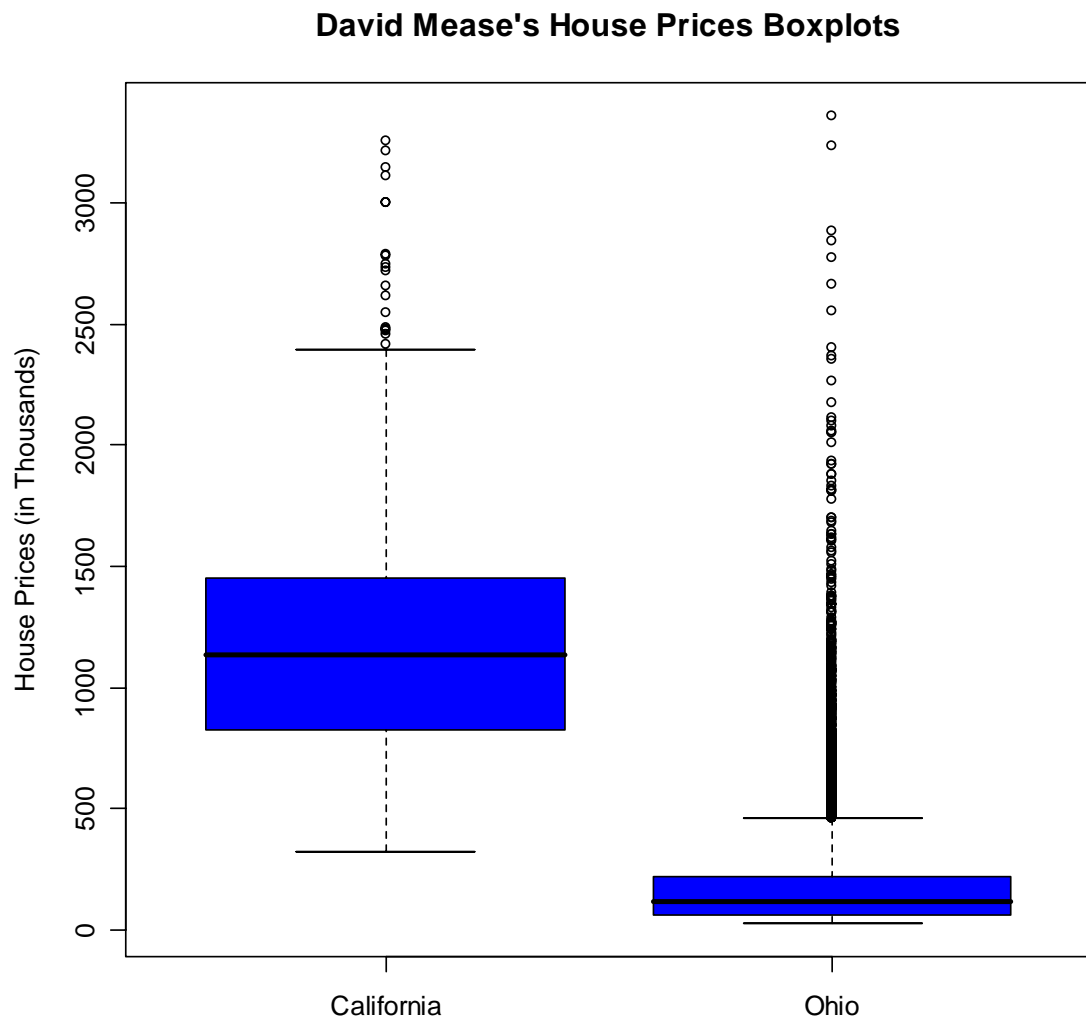
7)

a)

```
ca<-read.csv("CA_house_prices.csv",header=F)
```

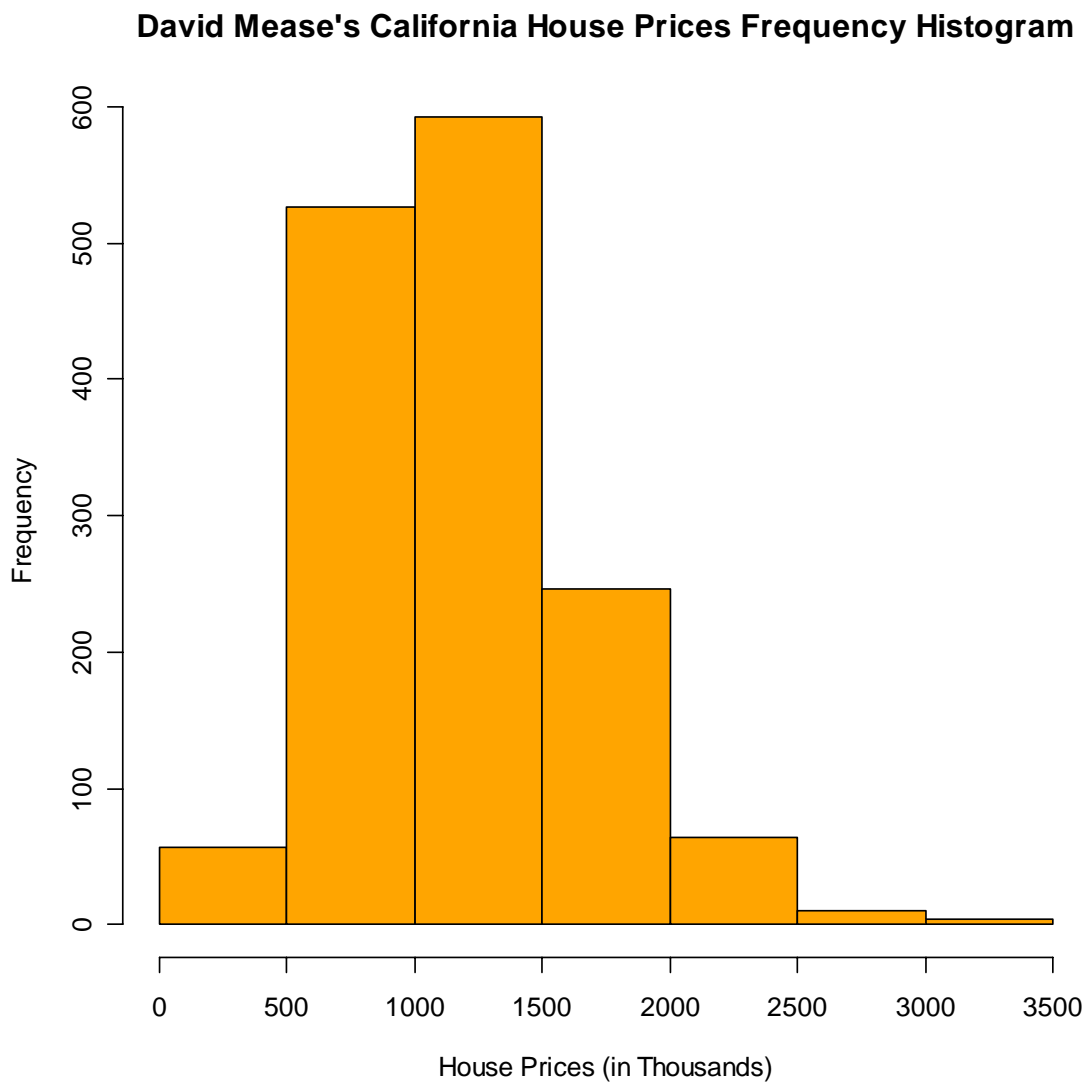
```
oh<-read.csv("OH_house_prices.csv",header=F)
```

```
boxplot(ca[,1],oh[,1],col="blue",  
main="David Mease's House Prices Boxplots",  
names=c("California","Ohio"),ylab="House Prices (in Thousands)")
```



b)

```
hist(ca[,1],breaks=seq(0,3500,by=500),  
     col="orange",  
     xlab="House Prices (in Thousands)", ylab="Frequency",  
     main="David Mease's California House Prices Frequency Histogram")
```

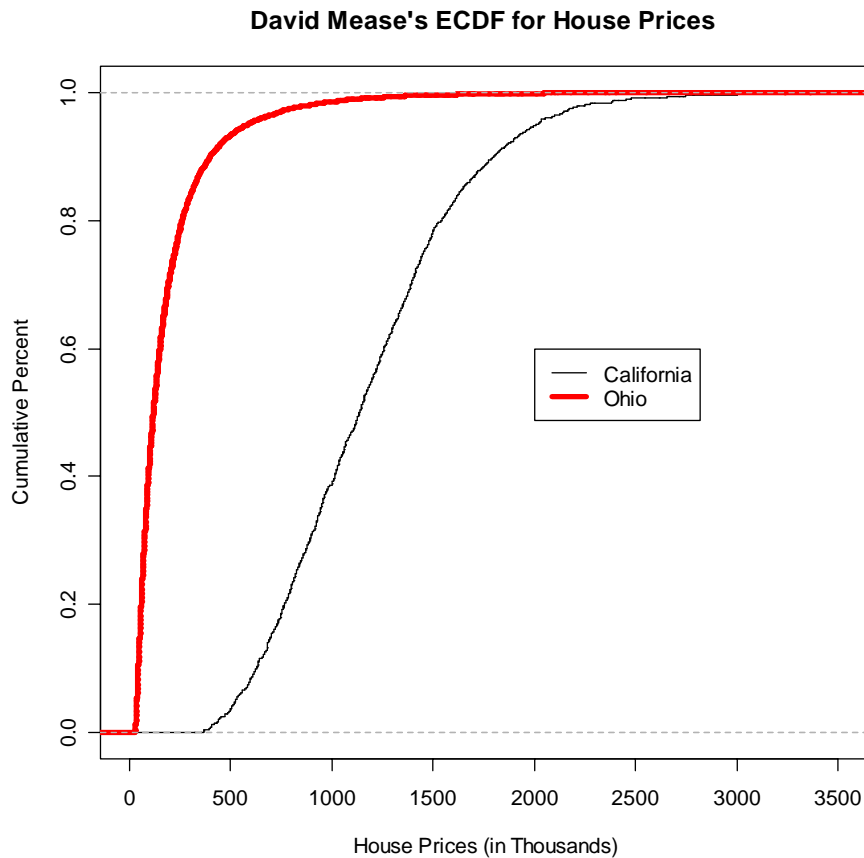


c)

```
plot(ecdf(ca[,1]),  
     verticals= TRUE,do.p = FALSE,  
     main ="David Mease's ECDF for House Prices",  
     xlab=" House Prices (in Thousands)",  
     ylab="Cumulative Percent",  
     xlim=c(0,3500))
```

```
lines(ecdf(oh[,1]),  
      verticals= TRUE,do.p = FALSE,  
      col.h="red",col.v="red",lwd=4)
```

```
legend(2000,.6,c("California","Ohio"),  
      col=c("black","red"),lwd=c(1,4))
```

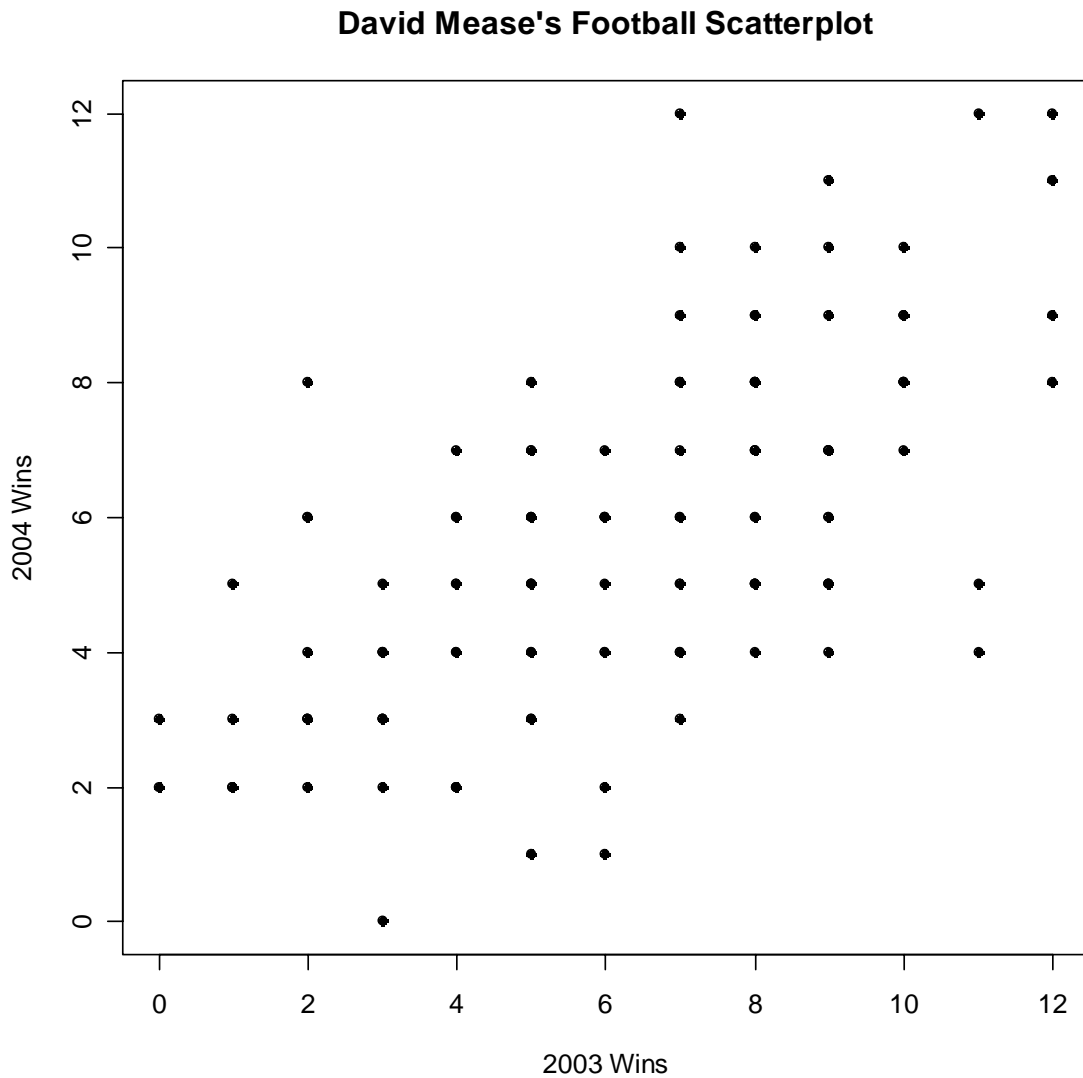


8)

a)

```
data<-read.csv("football.csv")
```

```
plot(data[,2],data[,3],  
      xlim=c(0,12),ylim=c(0,12),  
      pch=19,  
      main="David Mease's Football Scatterplot",  
      xlab="2003 Wins",ylab="2004 Wins")
```



b) There are fewer than 117 points because there are many ties due to the fact that the number of wins is discrete. The solution we discussed in class is to add a small amount of noise to jitter the points.

c) The correlation is 0.6537691 found by:

```
data<-read.csv("football.csv")  
cor(data[,2],data[,3])
```

d) It does not change.

```
cor(data[,2],data[,3]+10)
```

e) It does not change.

```
cor(data[,2],data[,3]*2)
```

f) It changes its sign to become negative.

```
cor(data[,2],data[,3]*-2)
```

9)

a) The median is \$118,000 found by:

```
data<-read.csv("OH_house_prices.csv",header=F)
median(data[,1])
```

This is much smaller than the mean of \$190,317.60 found by

```
mean(data[,1])
```

b) This suggests the data is right skewed.

c) The median increases by 10 (thousand dollars) to \$128,000.

```
median(data[,1]+10)
```

d) The median doubles.

```
median(2*data[,1])
```

10) a) 8.315218 found by

```
ages<-c(19,23,30,30,45,25,24,20).  
sd(ages)
```

b)

a. $\bar{X} = 216/8=27$

$$s = \sqrt{\frac{(19-27)^2 + (23-27)^2 + (30-27)^2 + (30-27)^2 + (45-27)^2 + (25-27)^2 + (24-27)^2 + (20-27)^2}{7}}$$

$$s = \sqrt{\frac{484}{7}} = 8.315$$

c) It does not change.

```
sd(ages+10)
```

d) It is multiplied by 100.

```
sd(100*ages)
```