Homework 1

Statistics 109

Due February 2, 2021 at 5:50 pm EST

Homework policies. *Please provide concise, clear answers for each question. Note that only writing the result of a calculation (e.g., "SD = 3.3") without explanation is not sufficient. For problems involving* R*, include the code in your solution, along with any plots.*

Please submit your homework assignment via Canvas as a PDF file.

We encourage you to discuss problems with other students (and, of course, with the course head and the TFs), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution.

Max points: 100 (Each problem is 20 points.)

Student Name: Erin Lopez

SOLUTION 1

(a) str(possum)

```
'data.frame':   104 obs. of  14 variables:
 $ case    : num  1 2 3 4 5 6 7 8 9 10 ...
 $ site    : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Pop     : Factor w/ 2 levels "Vic","other": 1 1 1 1 1 1 1 1 1 1 ...
 $ sex     : Factor w/ 2 levels "f","m": 2 1 1 1 1 1 2 1 1 1 ...
 $ age     : num  8 6 6 6 2 1 2 6 9 6 ...
 $ hdlngth : num  94.1 92.5 94 93.2 91.5 93.1 95.3 94.8 93.4 91.8 ...
 $ skullw  : num  60.4 57.6 60 57.1 56.3 54.8 58.2 57.6 56.3 58 ...
 $ totlngth: num  89 91.5 95.5 92 85.5 90.5 89.5 91 91.5 89.5 ...
 $ taill   : num  36 36.5 39 38 36 35.5 36 37 37 37.5 ...
 $ footlgth: num  74.5 72.5 75.4 76.1 71 73.2 71.5 72.7 72.4 70.9 ...
 $ earconch: num  54.5 51.2 51.9 52.2 53.2 53.6 52 53.9 52.9 53.4 ...
 $ eye     : num  15.2 16 15.5 15.2 15.1 14.2 14.2 14.5 15.5 14.4 ...
 $ chest   : num  28 28.5 30 28 28.5 30 30 29 28 27.5 ...
 $ belly   : num  36 33 34 34 33 32 34.5 34 33 32 …
```

(b) > possum[!complete.cases(possum),]

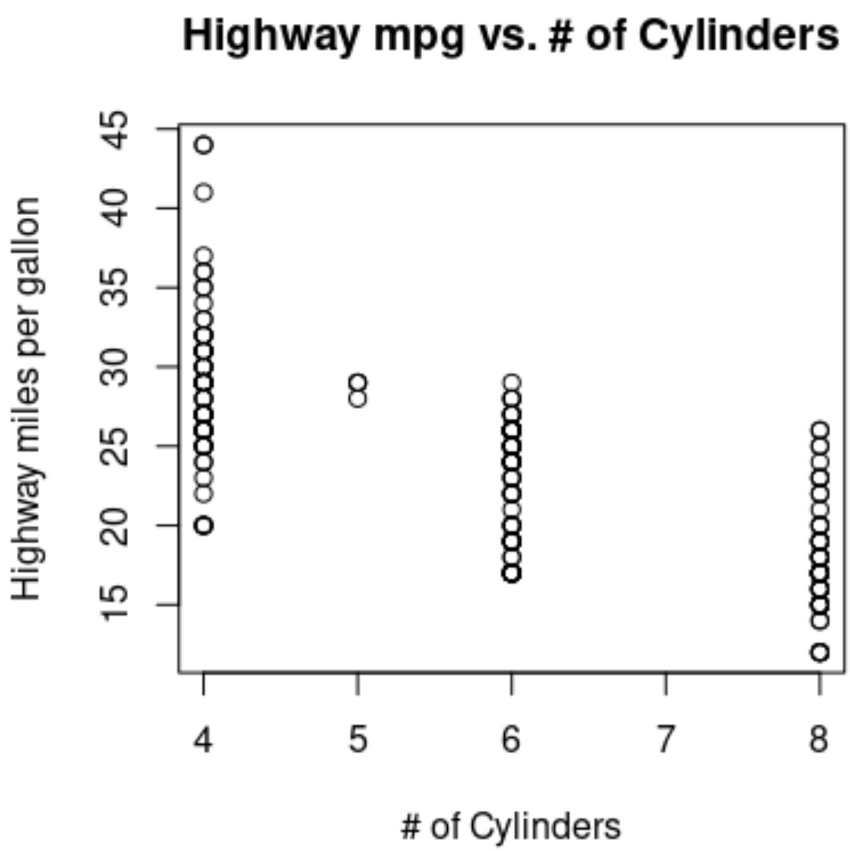
```
     case site Pop sex age hdlngth skullw totlngth
BB36   41    2 Vic   f   5    88.4   57.0       83
BB41   44    2 Vic   m  NA    85.1   51.5       76
BB45   46    2 Vic   m  NA    91.4   54.4       84
     taill footlgth earconch  eye chest belly
```

BB36	36.5	NA	40.3	15.9	27.0	30.5
BB41	35.5	70.3	52.6	14.4	23.0	27.0
BB45	35.0	72.8	51.2	14.4	24.5	35.0

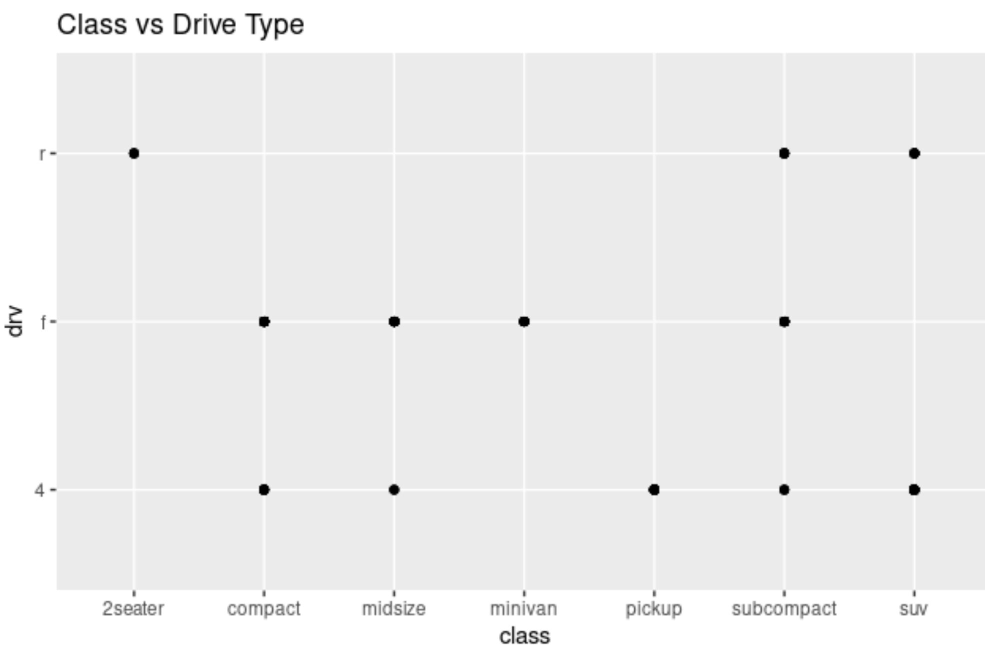
Row 36: missing footlngth
Rows 41, 45: missing age

SOLUTION 2

(a) 234 rows x 11 columns
(b) drv: the type of drive train, where f = front-wheel drive, r = rear wheel drive, 4 = 4wd
(c) plot(hwy ~ cyl, data=mpg, xlab='# of Cylinders', ylab='Highway miles per gallon', main='Highway mpg vs. # of Cylinders')
    (i) This has created a scatter plot showing the range of highway miles per gallon for each number of cylinders. Cars with 8 cylinders have the lowest range of approximately 5-25 highway mpg, and cars with 4 cylinders have highest range of approximately 20-45 highway mpg.

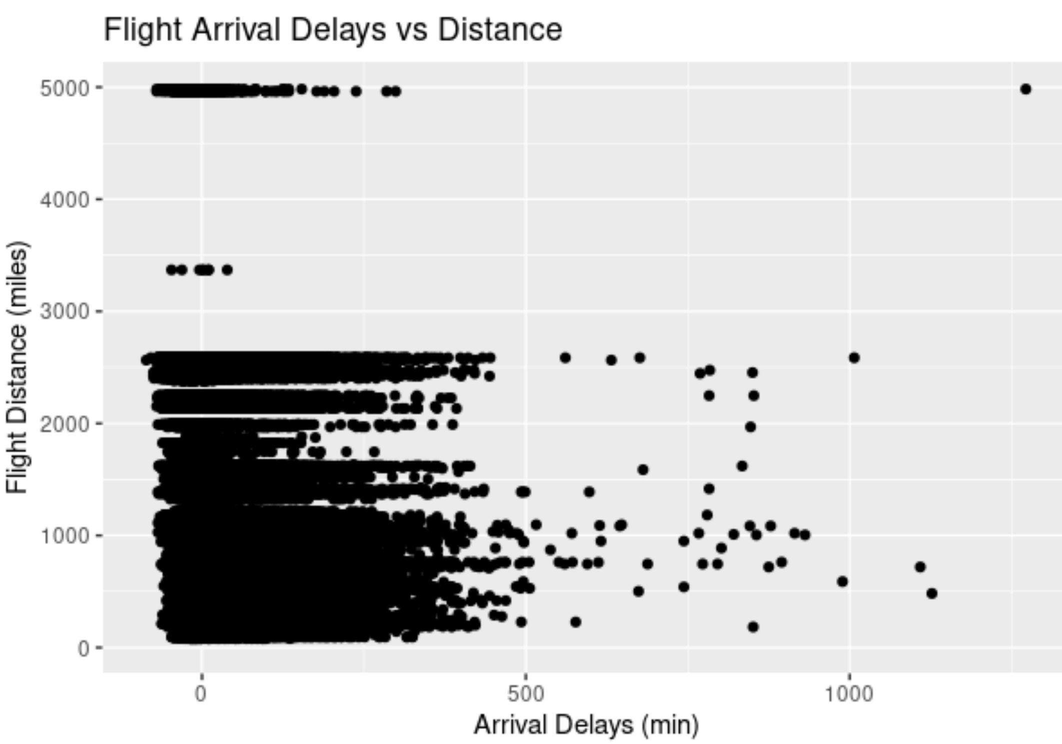


(d) ggplot(mpg, aes(class, drv)) + geom_point()
    (i) Class has 7 options and drv has only 3 options, so there are very limited possible points on the plot.



SOLUTION 3

(a) ggplot(flights, aes(arr_delay, distance)) + geom_point() + xlab('Arrival Delays (min)') + ylab('Flight Distance (miles)') + ggtitle('Flight Arrival Delays vs Distance')

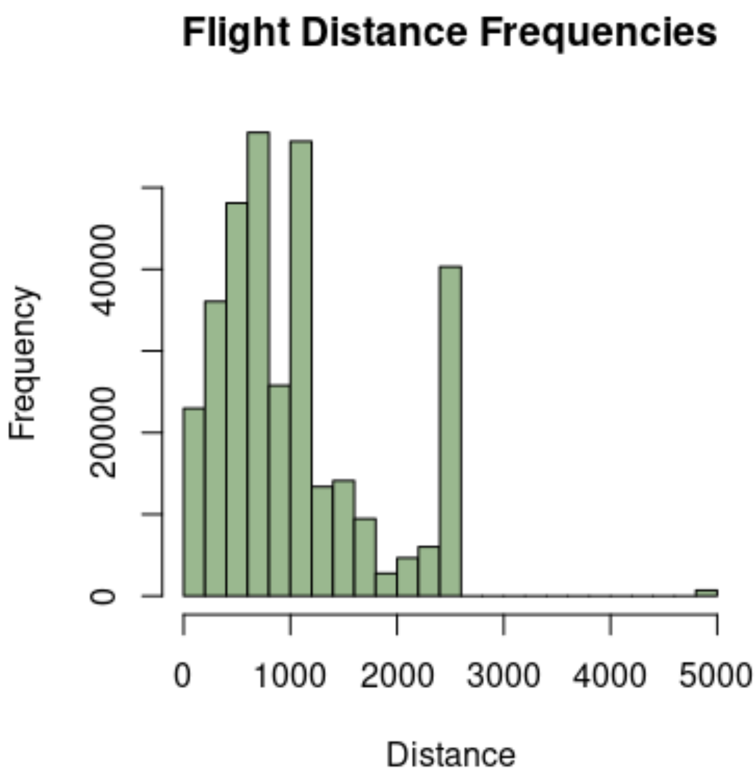


(b) hist(flights$distance,

col = 'darkseagreen',

xlab = 'Distance',

main = 'Flight Distance Frequencies')



(c)

logdist <- log(flights$distance)

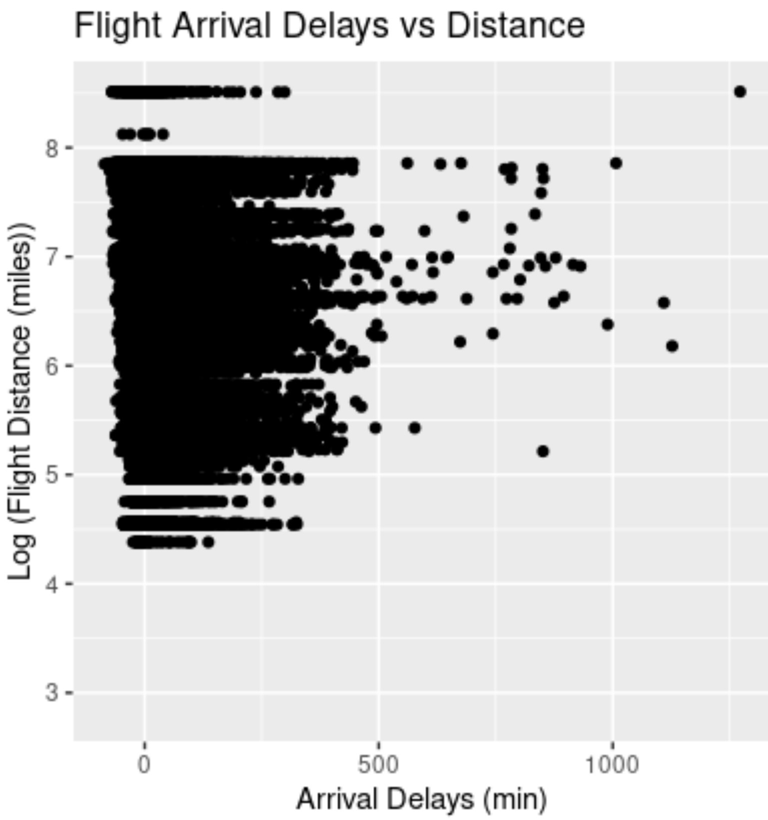
ggplot(flights, aes(arr_delay, logdist)) + geom_point() + xlab('Arrival Delays (min)') + ylab('Log (Flight Distance (miles))') + ggtitle('Flight Arrival Delays vs Distance')

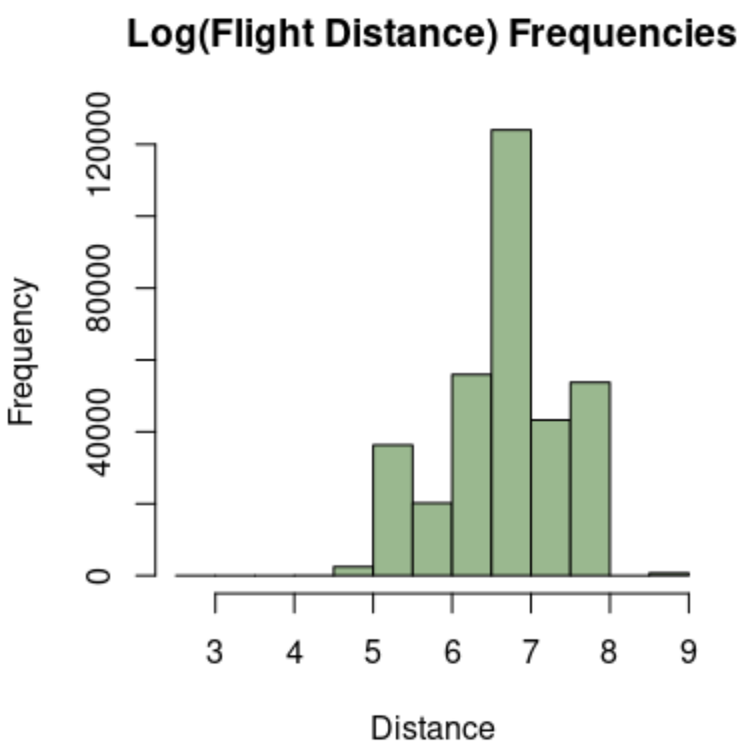
hist(logdist,

col = 'darkseagreen',

xlab = 'Distance',

main = 'Log(Flight Distance) Frequencies')





(d) The log(distance) histogram has fewer bars than the distance histogram. However, the distance histogram has lower frequencies than the log(distance) histogram.

SOLUTION 4

(a) ?Boston 506 rows and 14 columns

(i) Columns:

`crim`

per capita crime rate by town.

`zn`

proportion of residential land zoned for lots over 25,000 sq.ft.

`indus`

proportion of non-retail business acres per town.

`chas`

Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

`nox`

nitrogen oxides concentration (parts per 10 million).

`rm`

average number of rooms per dwelling.

`age`

proportion of owner-occupied units built prior to 1940.

`dis`

weighted mean of distances to five Boston employment centres.

`rad`

index of accessibility to radial highways.

`tax`

full-value property-tax rate per \$10,000.

`ptratio`

pupil-teacher ratio by town.

`black`

1000(Bk - 0.63)^2 where *Bk* is the proportion of blacks by town.

`lstat`

lower status of the population (percent).

`medv`

median value of owner-occupied homes in \$1000s.

The Boston dataset contains data on housing values in Boston suburbs. Each row represents a different Boston suburb town and each column represents a different data point of the suburb, such as medv, which represents the median value of homes in thousands that are owner-occupied.

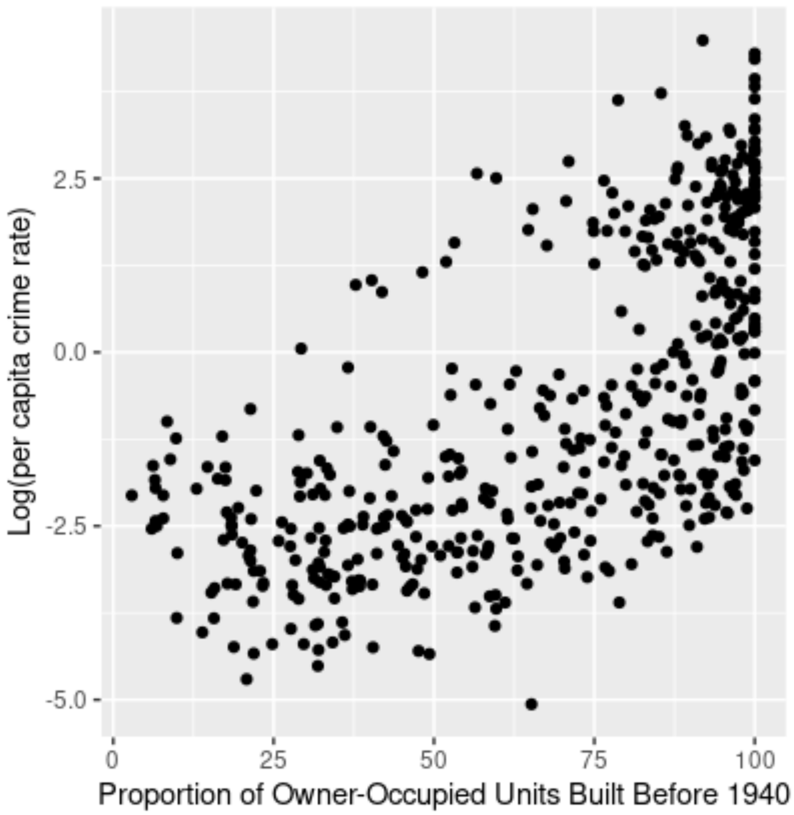
(b) > sum(Boston[,'chas'] == 1)
   (i) 35

SOLUTION 5

(a) ggplot(Boston, aes(medv, crim)) + geom_point() +xlab('Median Value of Owner-Occupied Homes in $1000s') + ylab('per capita crime rate') +ggtitle('Home Value vs Crime Rate')

Plotting median owner-occupied home value versus per capita crime rate shows that towns with lower home values show high crime rates, and towns with the highest home values have low per capita crime rates.

ggplot(Boston, aes(age, logcrim)) + geom_point() +xlab('Proportion of Owner-Occupied Units Built Before 1940') + ylab('Log(per capita crime rate)')



Plotting the age of homes versus the log of the crime rate shows that towns with older homes are associated with higher crime rates.

(b) Two of the predictors that are associated with per capita crime rate are the age of the homes and the value of homes. Older and less valuable homes are associated with higher crime rates. Newer homes and very expensive homes are associated with lower crime rates per capita.