Laboratory Exercise Week 12

Brian Tipton - STAT380-001

11/8/23

Directions:

- Write your R code inside the code chunks after each question.
- Write your answer comments after the `#` sign.
- To generate the word document output, click the button `Knit` and wait for the word document to appear.
- RStudio will prompt you (only once) to install the `knitr` package.
- Submit your completed laboratory exercise using Blackboard's Turnitin feature. Your Turnitin upload link is found on your Blackboard Course shell under the Laboratory folder.

For this exercise, you will need to use the packages `mosaic` and `dplyr` to find numerical and graphical summaries.

```
# install packages if necessary
if (!require(mosaic)) install.packages(`mosaic`)
if (!require(dplyr)) install.packages(`dplyr`)
# load the package in R
library(mosaic) # load the package mosaic to use its functions
library(dplyr) # load the package dplyr to use data management functions
```

1. Data from the US Federal Reserve Board (2002) on the percentage of disposable personal income required to meet consumer load payments and mortgage payments for selected years are found in the data below

```
debt <- read.csv("https://www.siue.edu/~jpailde/debt.csv")
debt
```

```
##    Consumer_Debt Household_Debt
## 1           7.88           6.22
## 2           7.91           6.14
## 3           7.65           5.95
## 4           7.61           5.83
## 5           7.48           5.83
## 6           7.49           5.85
## 7           7.37           5.81
## 8           6.57           5.79
## 9           6.24           5.73
## 10          6.09           5.95
## 11          6.32           6.09
## 12          6.97           6.28
## 13          7.38           6.08
## 14          7.52           5.79
## 15          7.84           5.81
```

i) Compute the mean and standard deviation of the two variables in the data set. Describe each variable using these sample characteristics.

ii) What is the value of the correlation coefficient for this data set?

iii) Construct a scatterplot for this data set.

iv) Based on parts (ii) and (iii), is it reasonable to conclude in this case that there is no strong relationship between the variables (linear or otherwise)?

Code chunk

```r
# start your code

# i) Compute the mean and standard deviation of the two variables in the data set.
# Describe each variable using these sample characteristics.
favstats( ~ Consumer_Debt, data = debt)
```

```
##   min   Q1 median   Q3  max     mean        sd  n missing
##  6.09 6.77   7.48 7.63 7.91 7.221333 0.6235826 15       0
```

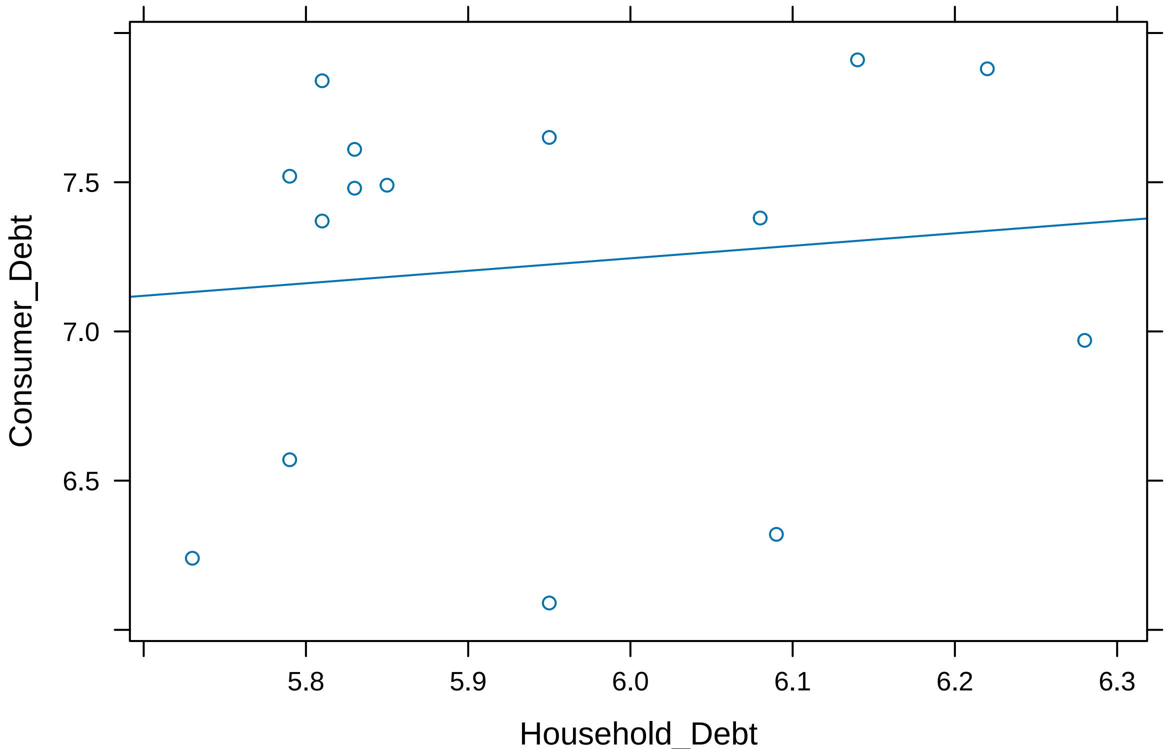
```r
favstats( ~ Household_Debt, data = debt)
```

```
##   min   Q1 median    Q3  max     mean        sd  n missing
##  5.73 5.81   5.85 6.085 6.28 5.943333 0.1755264 15       0
```

```r
# ii) What is the value of the correlation coefficient for this data set?
cor(debt$Consumer_Debt, debt$Household_Debt)
```

```
## [1] 0.1178129
```

```r
# iii) Construct a scatterplot for this data set.
xyplot(Consumer_Debt ~ Household_Debt,
       data = debt,
       type = c("p", "r"))
```




```
# iv) Based on parts (ii) and (iii),
# is it reasonable to conclude in this case that there is
# no strong relationship between the variables (linear or otherwise)?

# Yes there is no strong relationship between the variables.
# The correlation coefficient is very low.
# As well as on the scatter plot the points are very far from the correlation line.

# last R code line
```

2. The data below contains sale price, size, and land-to-building ratio for 10 large industrial properties

```
saleprice <- read.csv("https://www.siue.edu/~jpailde/saleprice.csv")
saleprice
```

```
##    Property price...M. Size..k.sq.ft. Ratio
## 1         1       10.6           2166   2.0
## 2         2        2.6            751   3.5
## 3         3       30.5           2422   3.6
## 4         4        1.8            224   4.7
## 5         5       20.0           3917   1.7
## 6         6        8.0           2866   2.3
## 7         7       10.0           1698   3.1
## 8         8        6.7           1046   4.8
## 9         9        5.8           1108   7.6
## 10       10        4.5            405  17.2
```

i) Construct scatterpots for `sale price versus size` and `sale price versus land-to-building` ratio. Be sure to fit regression lines on the scatterplots.

ii) Calculate and interpret the value of the correlation coefficients on `sale price versus size` and `sale price versus land-to-building ratio`.

iii) Use the `lm` function to estimated the equations of each fitted line for for `sale price versus size` and `sale price versus land-to-building ratio`. Give an interpretation of the slopes for each line.

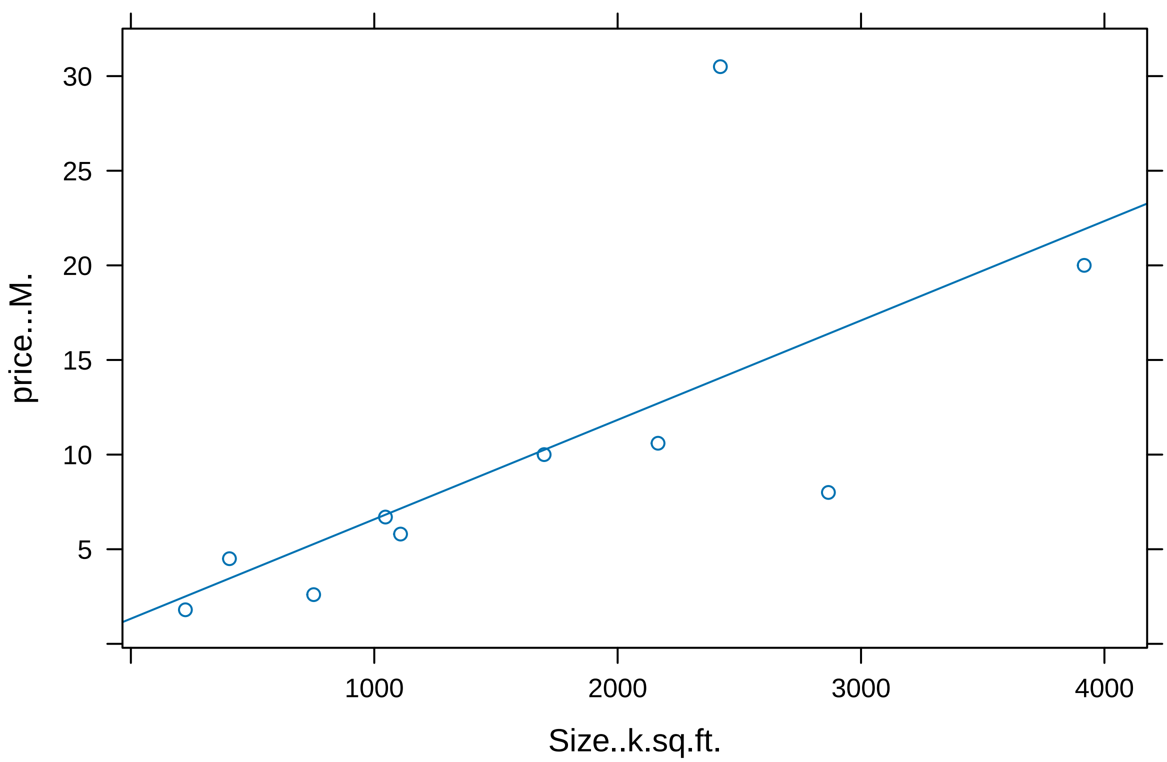
iv) Which model would you use to explain the behavior of sale price? Why?

Code chunk

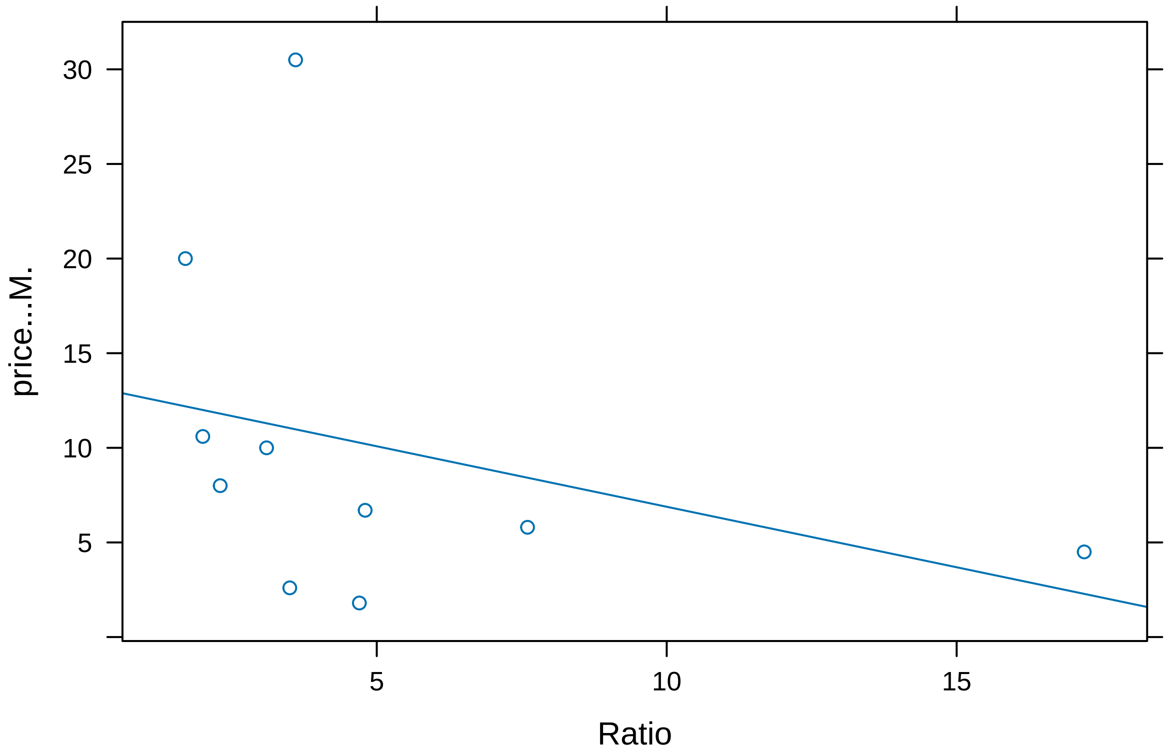
```
# start your code

# i) Construct scatterpots for `sale price versus size` and `sale price versus land-to-building ratio`.
# Be sure to fit regression lines on the scatterplots.

# a. Sale Price Versus Size
xyplot(price...M. ~ Size..k.sq.ft.,
       data = saleprice,
       type = c("p", "r"))
```




```
# b. Sale Price Versus land-to-building Ratio
xyplot(price...M. ~ Ratio,
       data = saleprice,
       type = c("p", "r"))
```




```
# ii) Calculate and interpret the value of the correlation coefficients
# on `sale price versus size` and `sale price versus land-to-building ratio`.
cor(saleprice)
```

```
##                  Property price...M. Size..k.sq.ft.      Ratio
## Property        1.0000000 -0.2878978     -0.2621620  0.6668957
## price...M.     -0.2878978  1.0000000      0.7001976 -0.3324335
## Size..k.sq.ft. -0.2621620  0.7001976      1.0000000 -0.5766567
## Ratio           0.6668957 -0.3324335     -0.5766567  1.0000000
```

```
# iii) Use the `lm` function to estimated the equations of each fitted line for for `sale price versus s
# Give an interpretation of the slopes for each line.
model1 <- lm(price...M.~Size..k.sq.ft., data = saleprice)
print(model1)
```

```
##
## Call:
## lm(formula = price...M. ~ Size..k.sq.ft., data = saleprice)
##
## Coefficients:
##    (Intercept)  Size..k.sq.ft.
##       1.328142        0.005253
```

```
# y = 0.005253x + 1.328142
# Price is positively correlated to the size.

model2 <- lm(price...M. ~ Ratio, data = saleprice)
print(model2)
```

```
##
## Call:
## lm(formula = price...M. ~ Ratio, data = saleprice)
##
## Coefficients:
## (Intercept)        Ratio
##     13.2787      -0.6393
```

```r
# y = -0.6393x + 13.2787
# Price is negatively correlated to the ratio.

# iv) Which model would you use to explain the behavior of sale price? Why?
summary(model1)$r.squared
```

```
## [1] 0.4902767
```

```r
summary(model2)$r.squared
```

```
## [1] 0.110512
```

```r
# Model 1 has a higher r squared value.
# So we would pick Model 1.
# The Model 1 has a better fitted line then Model 2.

# last R code line
```