Lab 2

Keara Dreyfuss

2/08/24

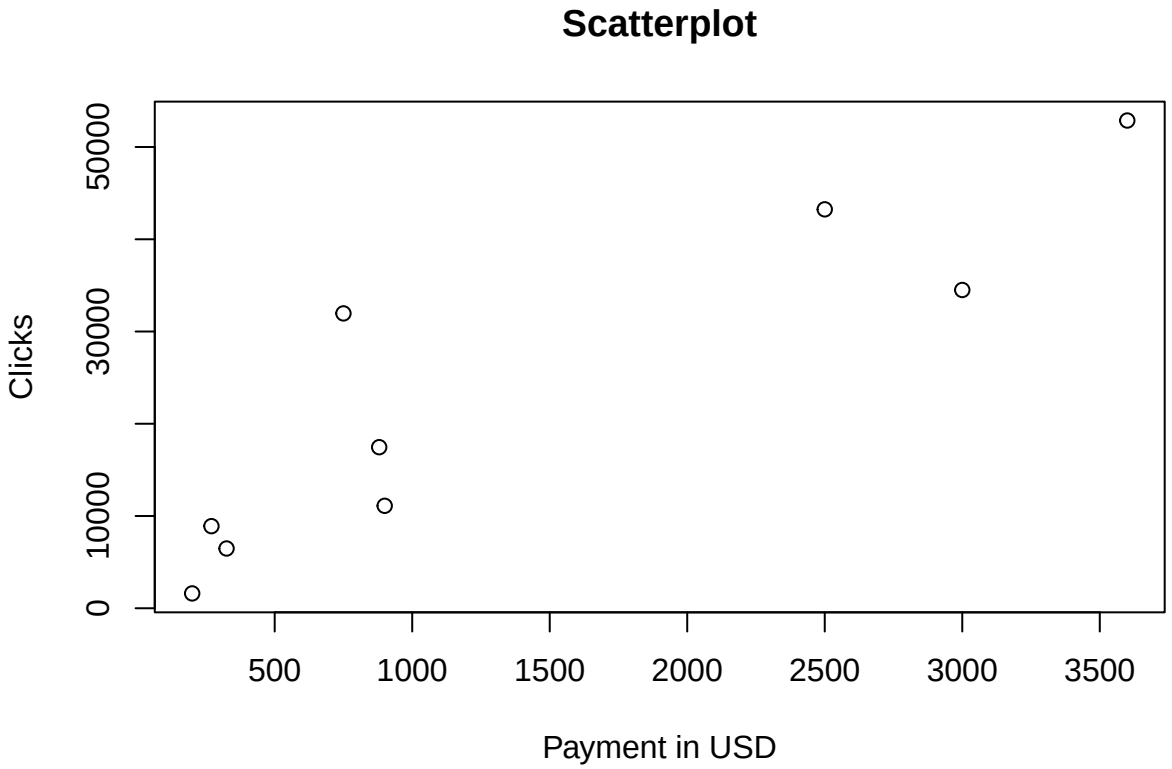
Example 1: Correlation and Scatterplot

The CEO of Zen Sports Apparel contracts social media influencers to generate clicks to their products. They send endorsements to multiple influencers and get varying number of clicks to their website. The data shown in the table below reflect the amount paid (in USD) and the resulting number of clicks.

a. Given the data, plot a scatter plot.

Variable									
x (Payment in USD)	900	325	3000	750	2500	200	880	270	3600
y (Clicks)	11106	6479	34501	31962	43242	1608	17460	8900	52870

```
x=c(900, 325, 3000,  750, 2500,  200,  880,  270, 3600)
y=c(11106, 6479,34501,31962,43242, 1608,17460, 8900,52870)

plot(x,y, xlab="Payment in USD", ylab="Clicks", main="Scatterplot")
```



c. Find the correlation coefficient (r) and the coefficient of determination (r^2).

The correlation coefficient:

```
r=cor(x,y)
r
```

```
## [1] 0.9006833
```

The coefficient of determination:

```
R2=r^2
R2
```

```
## [1] 0.8112303
```

Example 2: Linear Regression

Here we will be using the data from the previous exercise.

a. Find the linear regression line that fits the data the best and get a summary of the data.

```
reg=lm(y~x)
summary(reg)
```

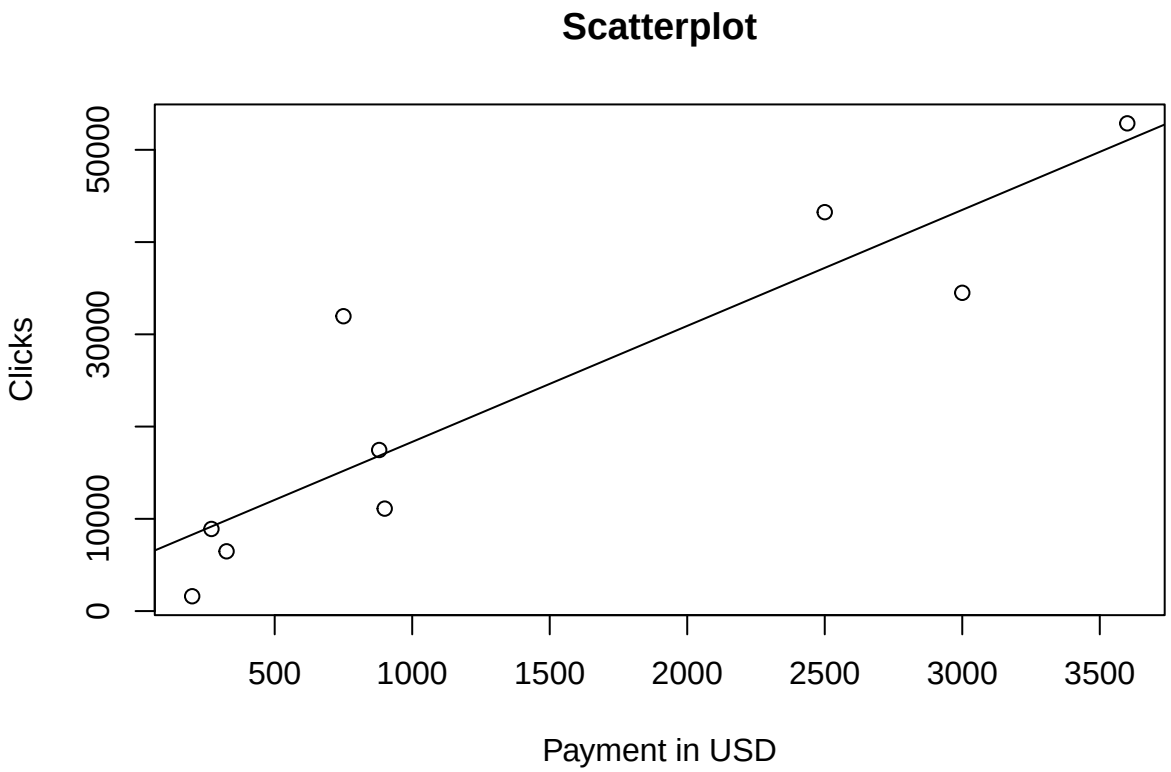
```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8987.4 -5976.8  -261.1  1837.2 16765.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5766.104   4225.582   1.365 0.214631
## x             12.574      2.293   5.485 0.000921 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8399 on 7 degrees of freedom
## Multiple R-squared:  0.8112, Adjusted R-squared:  0.7843
## F-statistic: 30.08 on 1 and 7 DF,  p-value: 0.0009213
```

Looking at the summary table of the regression line we see that the formula is:

$$\hat{y} = 5766.104 + 12.574x$$

b. Modify your labeled scatterplot in the previous problem to include the linear regression line in the graph.

```
plot(x,y, xlab="Payment in USD", ylab="Clicks", main="Scatterplot")
abline(reg)
```



c. Let's predict the number of clicks when the payment is $1000 using the predict function:

```
p=predict(reg, data.frame(x=1000))
p
```

```
##        1
## 18340.19
```

Problems

Problem 1: Age of Billionaires

A sample of 10 billionaires is selected, and the person's age and net worth are compared. The data are given here.

Variable										
X (age)	56	39	42	60	84	37	68	66	73	55
Y (net worth in billions)	18	14	12	14	11	10	10	7	7	5

a. Find the equation of the least squares line and provide a model summary. (3 points)

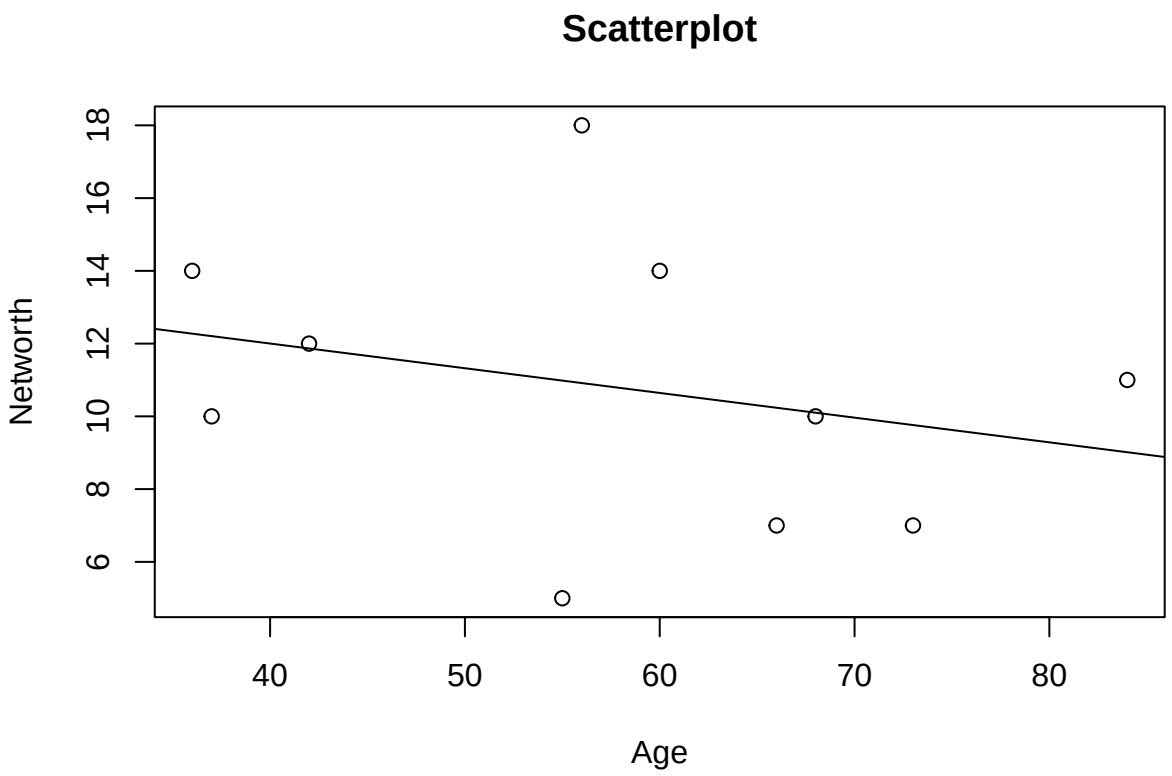
```
# Your code here
x = c(56, 36, 42, 60, 84, 37, 68, 66, 73, 55)
y = c(18, 14, 12, 14, 11, 10, 10, 7, 7, 5)

reg=lm(y~x)
summary(reg)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9833 -2.6222  0.0167  1.9210  7.0846
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.71792    4.99853   2.944   0.0186 *
## x           -0.06790    0.08383  -0.810   0.4414
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.987 on 8 degrees of freedom
## Multiple R-squared:  0.0758, Adjusted R-squared:  -0.03973
## F-statistic: 0.6561 on 1 and 8 DF,  p-value: 0.4414
```

$$\hat{y} = 14.71792 - .06790x$$

b. Create a scatter plot labeling the axes and insert the regression line in the scatter diagram of part (a). (3 points)

```
plot(x,y, xlab="Age", ylab="Networth", main="Scatterplot")
abline(reg)
```



c. Find the correlation coefficient and the coefficient of determination. (3 points)

```
r = cor(x,y)
r
```

```
## [1] -0.2753126
```

```
r^2
```

```
## [1] 0.07579703
```

Problem 2: Salary and job changes

A sociologist is interested in the relation between y = annual salary (in thousands of dollars) and x = the number of job changes for people living in the Nashville area. A random sample of 10 people employed in Nashville provided the following information:

Variable										
X (Number of job changes)	4	7	5	6	1	5	9	10	10	3
Y (Salary in $1000)	33	37	34	32	32	38	43	37	40	33

a. Find the equation of the least squares line and provide a model summary. (4 points)

```
x = c(4, 7, 5, 6, 1, 5, 9, 10, 10, 3)
y = c(33, 37, 34, 32, 32, 38, 43, 37, 40, 33)

reg=lm(y~x)
summary(reg)
```

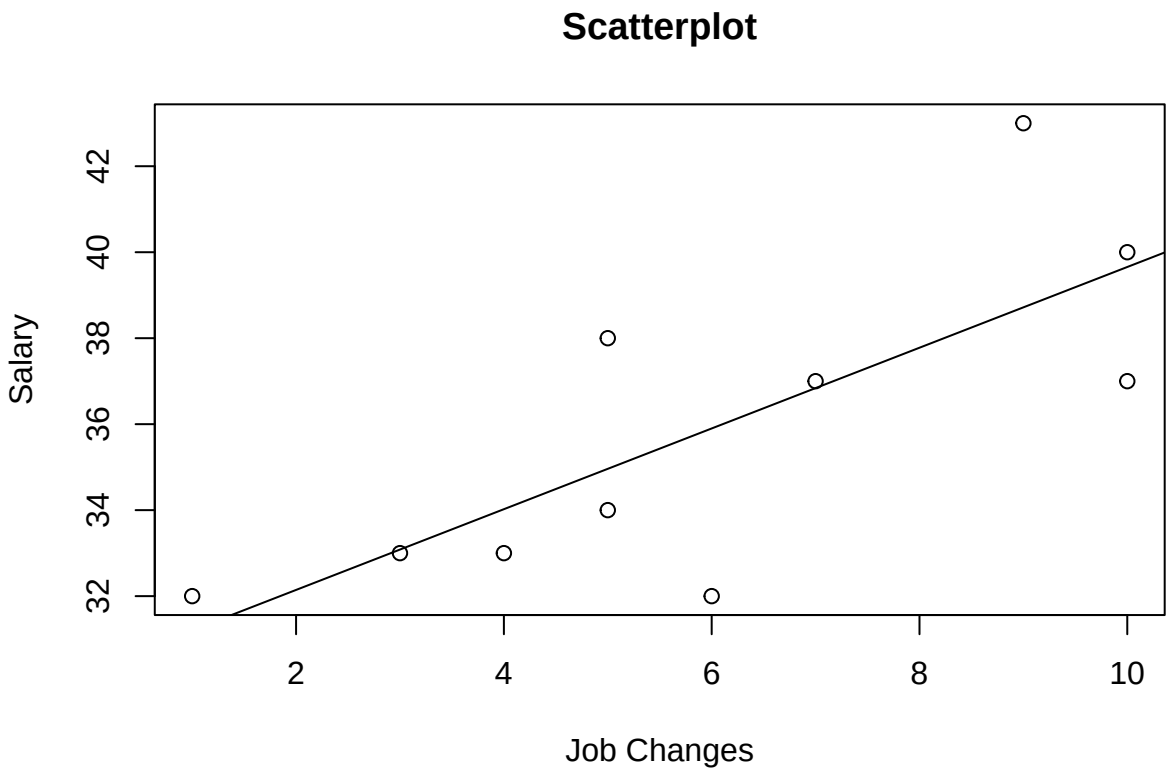
```
##
## Call:
```

```
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9000 -1.0067  0.0390  0.6823  4.2829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.2659     1.8825  16.078 2.25e-07 ***
## x             0.9390     0.2832   3.316   0.0106 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.564 on 8 degrees of freedom
## Multiple R-squared:  0.5789, Adjusted R-squared:  0.5263
## F-statistic:    11 on 1 and 8 DF,  p-value: 0.0106
```

$$\hat{y} = 30.2659 - .9390x$$

b. Create a scatter plot labeling the axes and insert the regression line in the scatter diagram of part (a). (3 points)

```
plot(x,y, xlab="Job Changes", ylab="Salary", main="Scatterplot")
abline(reg)
```



c. Find the correlation coefficient and the coefficient of determination. (3 points)

```
r = cor(x,y)
r
```

```
## [1] 0.7608562
```

```
r^2
```

```
## [1] 0.5789021
```

Problem 3: Speed and mileage

A study conducted by a department of transportation regarding driving speed and mileage for midsize automobiles resulted in the data:

Variable										
X (speed in mph)	30	50	40	55	30	25	60	25	50	55
Y (mileage in mpg)	28	25	25	23	30	32	21	35	26	25

a. Find the equation of the least squares line and provide a model summary. (4 points)

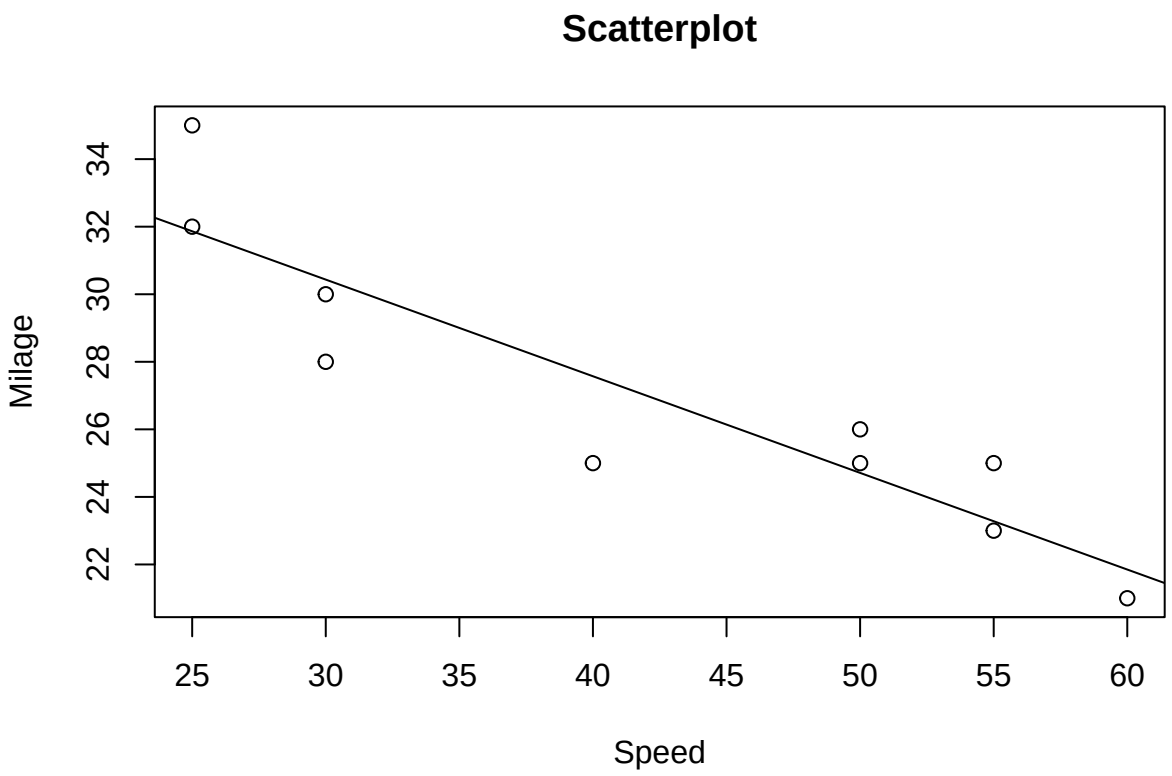
```
x = c(30, 50, 40, 55, 30, 25, 60, 25, 50, 55)
y = c(28, 25, 25, 23, 30, 32, 21, 35, 26, 25)

reg=lm(y~x)
summary(reg)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.57229 -0.74548 -0.07229  1.03916  3.13554
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.01807    2.02017  19.314 5.36e-08 ***
## x           -0.28614    0.04598  -6.223 0.000253 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.874 on 8 degrees of freedom
## Multiple R-squared:  0.8288, Adjusted R-squared:  0.8074
## F-statistic: 38.72 on 1 and 8 DF,  p-value: 0.0002531
```

$$\hat{y} = 39.01807 - .28614x$$

b. Create a scatter plot labeling the axes and insert the regression line in the scatter diagram of part (a). (3 points)

```
plot(x,y, xlab="Speed", ylab="Milage", main="Scatterplot")
abline(reg)
```



c. Find the correlation coefficient and the coefficient of determination. (3 points)

```
r = cor(x,y)
r
```

```
## [1] -0.9103694
```

```
r^2
```

```
## [1] 0.8287724
```

d. Predict the mileage for a midsize automobile traveling at 35 mph. (2 points)

```
p=predict(reg, data.frame(x=35))
p
```

```
##        1
## 29.00301
```

Problem 4: Sales and visits

Dorothy Kelly sells life insurance for the Prudence Insurance Company. She sells insurance by making visits to her clients' homes. Dorothy believes that the number of sales should depend, to some degree, on the number of visits made. For the past several years, she has kept careful records of the number of visits (X) she made each week and the number of people (Y) who bought insurance that week. For a random sample of 15 such weeks, the X and Y values follow:

Variable															
X (visits)	11	19	16	13	28	5	20	14	22	7	15	29	8	25	16
Y (sales)	3	11	8	5	8	2	5	6	8	3	5	10	6	10	7

a. Find the equation of the least squares line and provide a model summary. (4 points)

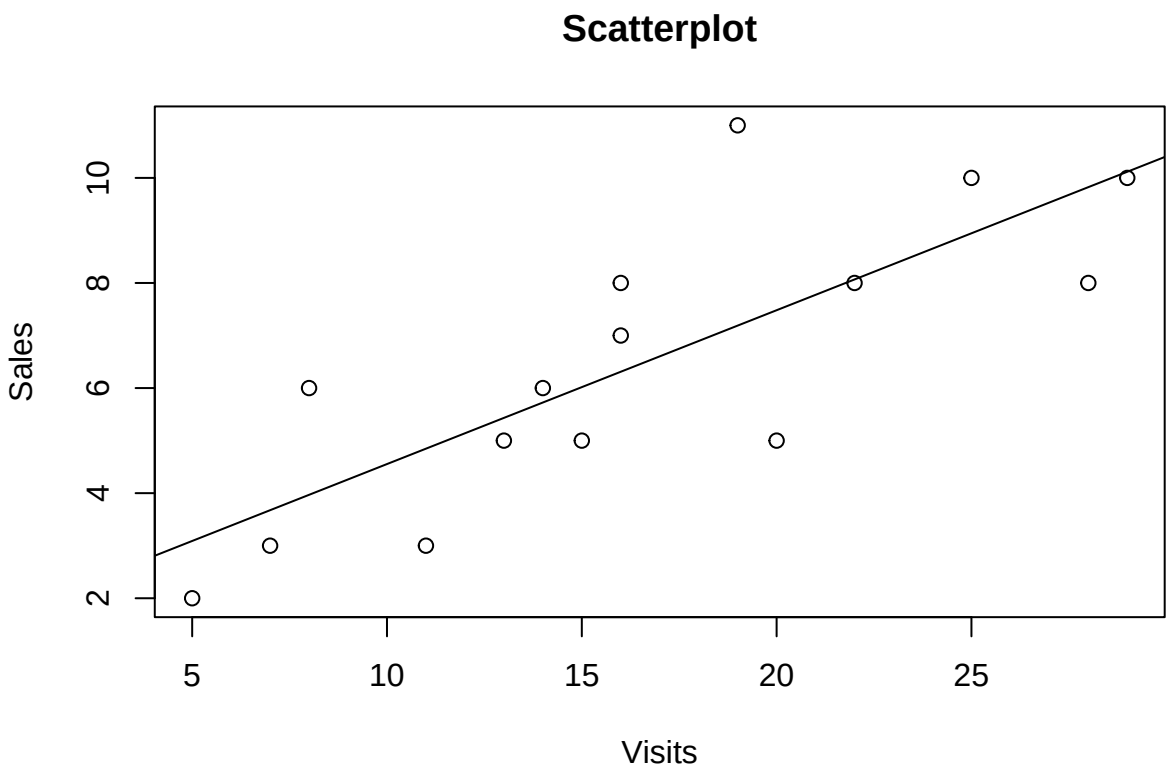
```r
x = c(11, 19, 16, 13, 28, 5, 20, 14, 22, 7, 15, 29, 8, 25, 16)
y = c(3, 11, 8, 5, 8,  2, 5, 6, 8, 3, 5, 10, 6, 10, 7)

reg=lm(y~x)
summary(reg)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4817 -1.0538 -0.1167  0.8720  3.8111
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.62597    1.13290   1.435 0.174842
## x            0.29278    0.06296   4.650 0.000455 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.731 on 13 degrees of freedom
## Multiple R-squared:  0.6245, Adjusted R-squared:  0.5956
## F-statistic: 21.62 on 1 and 13 DF,  p-value: 0.0004545
```

$$\hat{y} = 1.62597 + .29278x$$

b. Create a scatter plot labeling the axes and insert the regression line in the scatter diagram of part (a). (3 points)

```r
plot(x,y, xlab="Visits", ylab="Sales", main="Scatterplot")
abline(reg)
```



c. Find the correlation coefficient and the coefficient of determination. (3 points)

```
r = cor(x,y)
r
```

```
## [1] 0.7902646
```

```
r^2
```

```
## [1] 0.6245182
```

d. Predict the number of sales if Dorothy visits 10 times. (2 points)

```
p=predict(reg, data.frame(x=10))
p
```

```
##        1
## 4.553811
```

e. Is the prediction you made in the previous part an extrapolation or interpolation? (2 points)

The prediction is interpolation because 10 falls within the range of x.

Problem 5: Markdown

Now you will modify your document so that it is in pristine format.

Knit the document as a pdf. You will need to submit this file and the pdf you created. Submit these two files to blackboard. Partial credit will not be given for this problem. (5 points)