Data Science Through R

Topic: -Correlation and Regression Analysis

1. Six children aged 9, 11, 13, 15, 17 and 19 years old weight 20, 25, 32, 44, 48 and 52 kilograms respectively. Find the equation of the regression line of age on weight. Based on this data, what is the approximate weight of a twelve-year-old child?

Ans: -

```
> x <- c(9,11,13,15,17,19)
> y <- c(20,25,32,44,48,52)
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    -11.367        3.443

>
```

The regression equation of age on weight is y = 3.443x + (-11.367).

Based on the regression equation, the

approximate weight of a twelve-year-old child is : -

3.443*12-11.367= 29.949 kg.

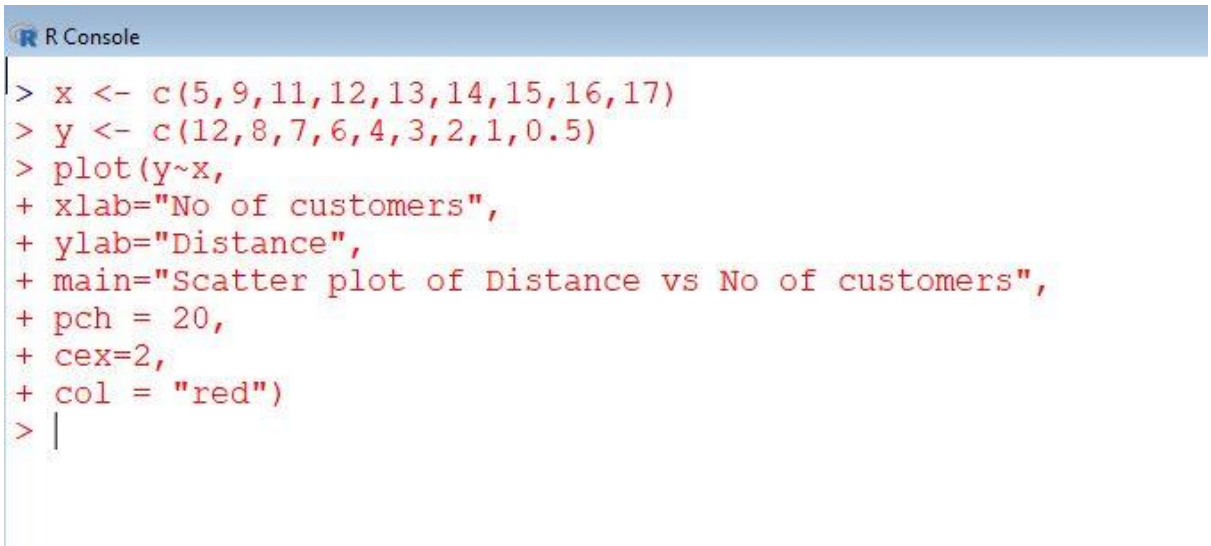
2. The success of a shopping center can be represented as a function of the distance (in km) from the center of the population and the number of clients (in hundreds of people) who will visit the centre each day. A set of collected data is given in the table below:

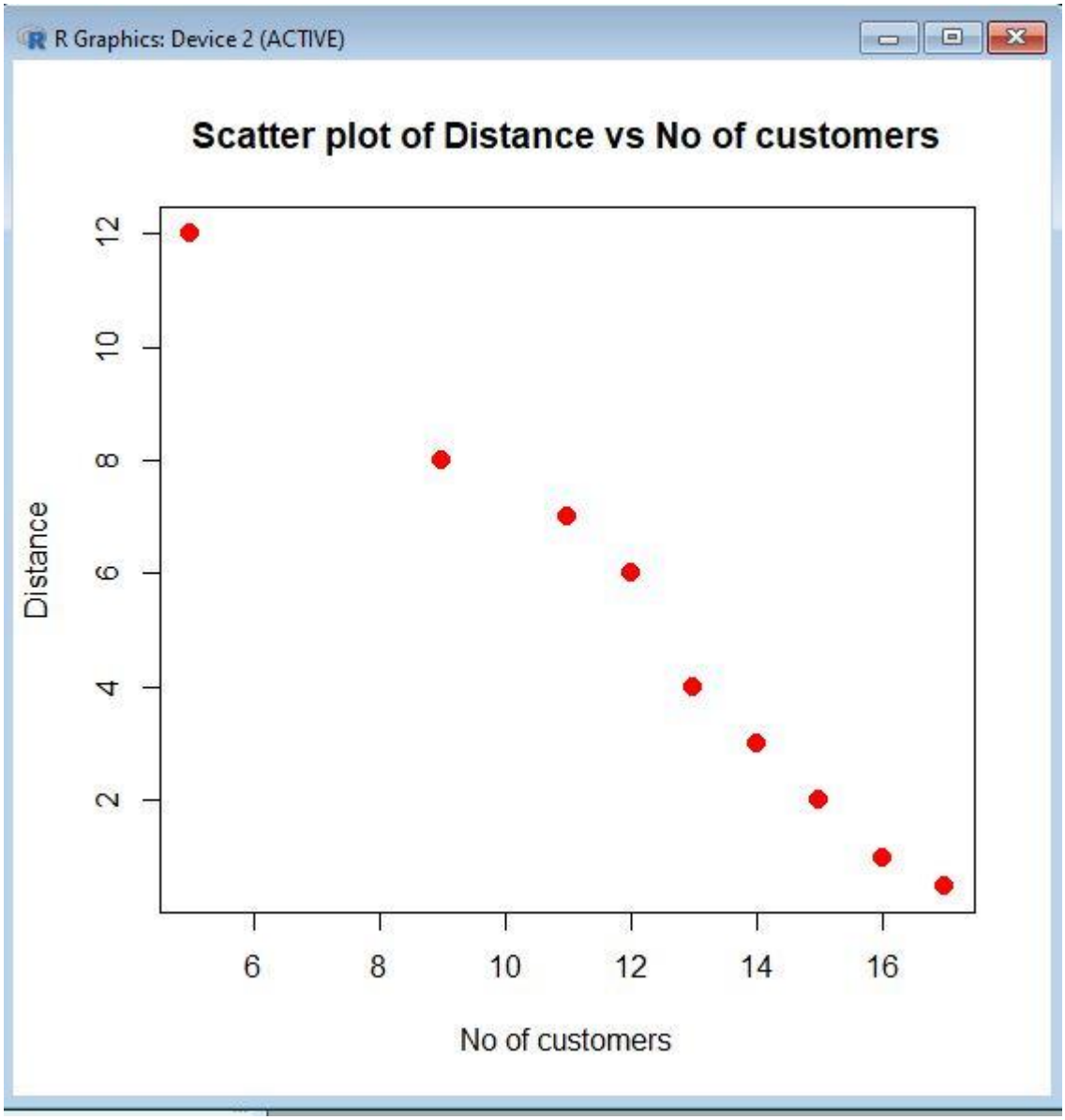
No of Customers(x)	5	9	11	12	13	14	15	16	17
Distance(y)	12	8	7	6	4	3	2	1	0.5

On the basis of the above data, perform the following:

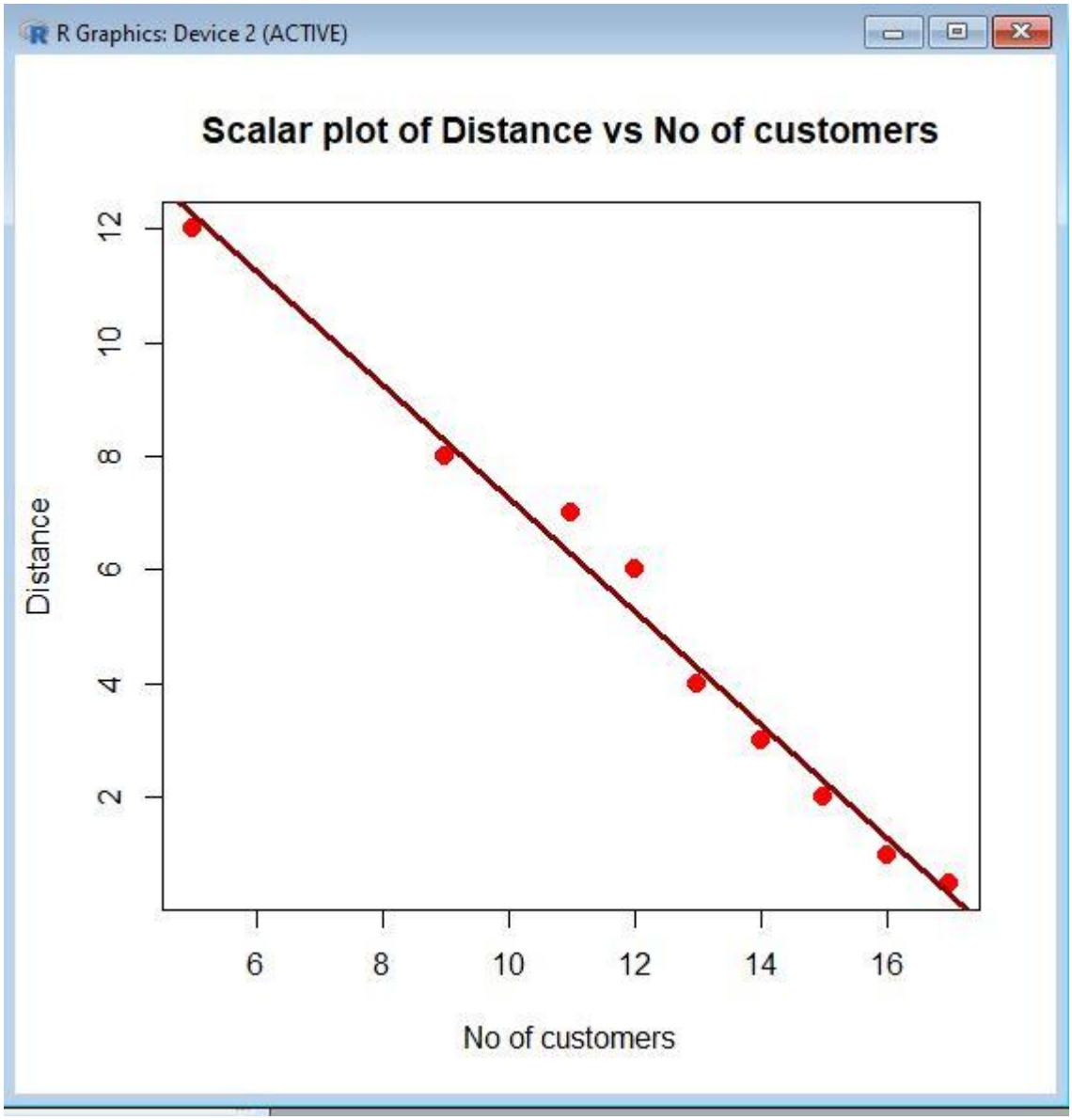
• A. Draw:

(a) Scatter plot: to visualize the linear relationship between the predictor and response





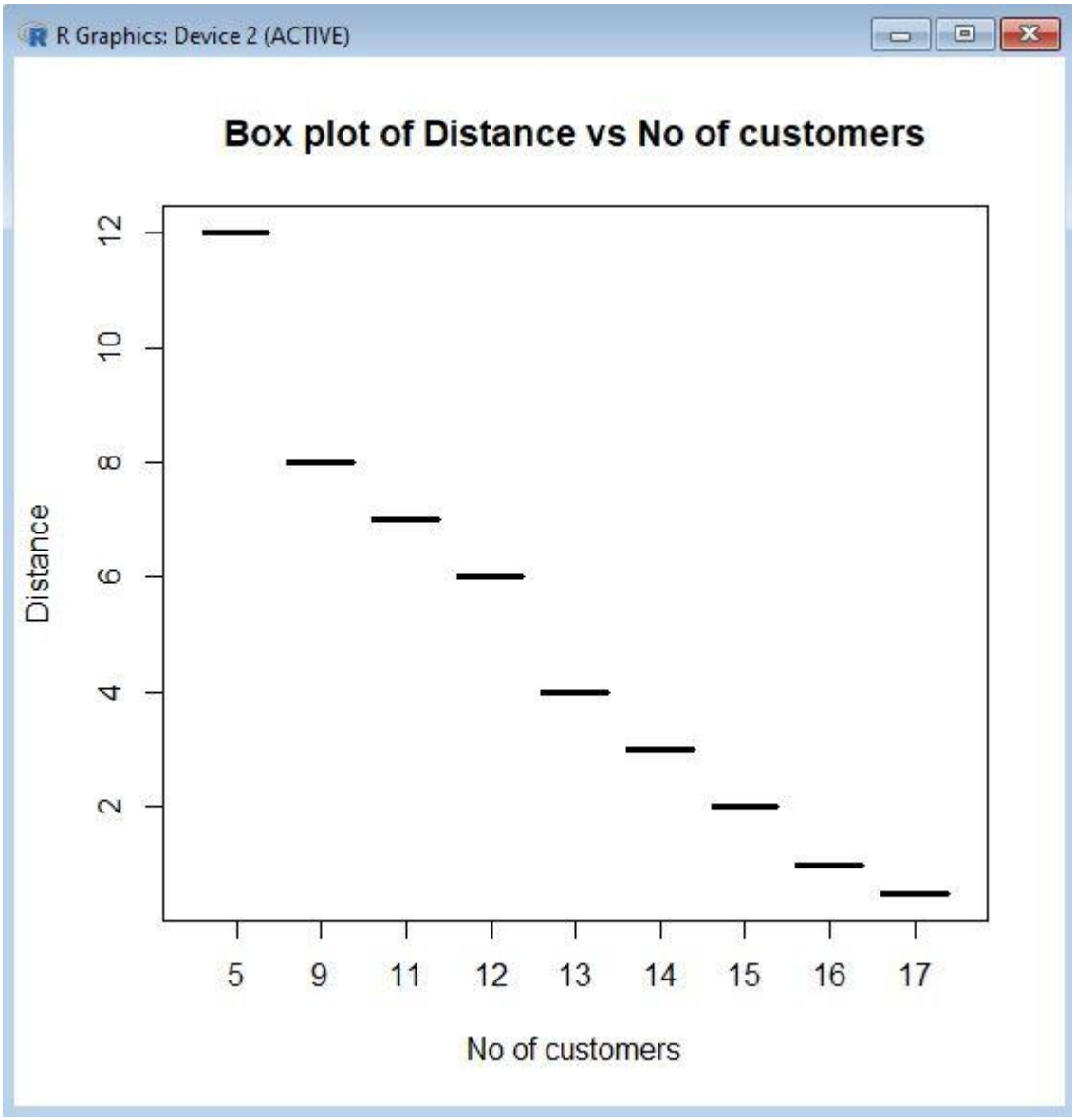

```
>  cust_dist <- lm(y~x)
> abline(cust_dist, lwd = 3, col = "darkred")
>
```



(b) Box plot: to spot any outlier observations in the variable. Having outliers in your predictor can drastically affect the predictions as they can easily affect the direction/slope of the line of best fit.

R Console

```
> boxplot(y~x,
+ xlab="No of customers",
+ ylab="Distance",
+ main="Box plot of Distance vs No of customers",
+ pch = 20,
+ cex=2,
+ col = "red")
>
```



• B.

(a) Calculate the linear correlation coefficient.

```
> cor(x,y)
[1] -0.9930992
```

(b) If the mall is located 3 miles away from the center of the population, how many customers should the shopping center expect?

```
> lm(x~y)

Call:
lm(formula = x ~ y)

Coefficients:
(Intercept)            y
    17.2279      -0.9897
```

No of customers the shopping centre should expect is : -

1.60934*3 *(-0.9897) + 17.2279 = 12.44

(c) To receive 900 customers a day, at what distance from the center of the population should the shopping centre be located?

```
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    17.2347      -0.9965

>
```

To receive 900 customers a day, the shopping centre should be located at a distance of : -

-0.9965*9+17.2347 = 8.2662 miles

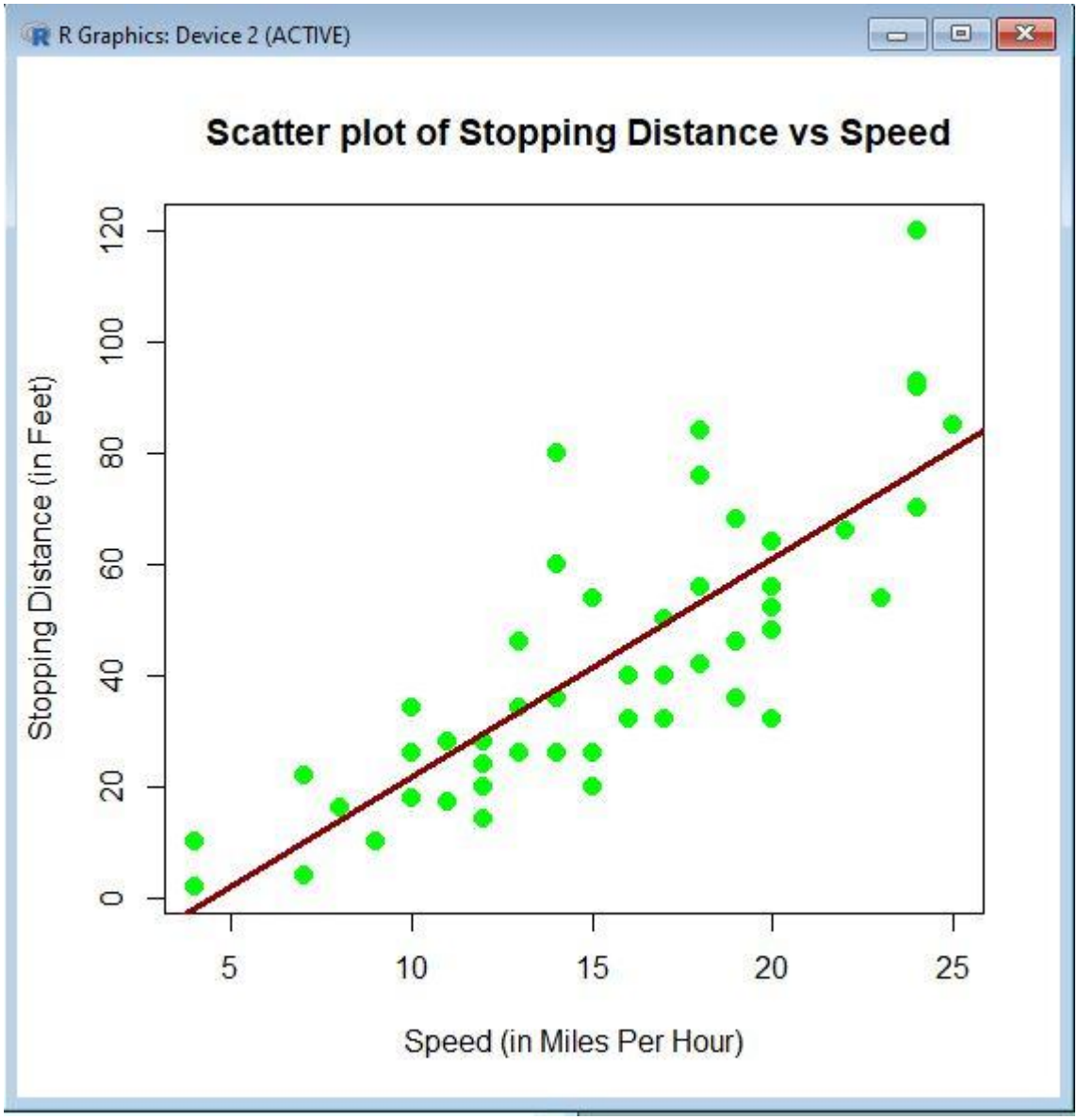
4. Let us use the cars dataset that comes with R by default as a standard built-in dataset and perform the following:

A. Draw

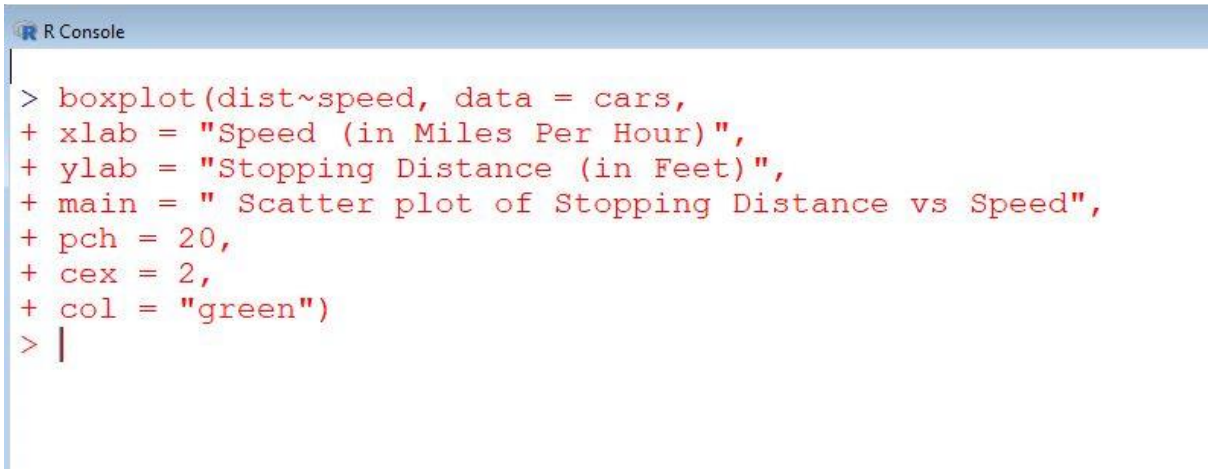
(a) Scatter plot: To Visualize the linear relationship between the predictor and response

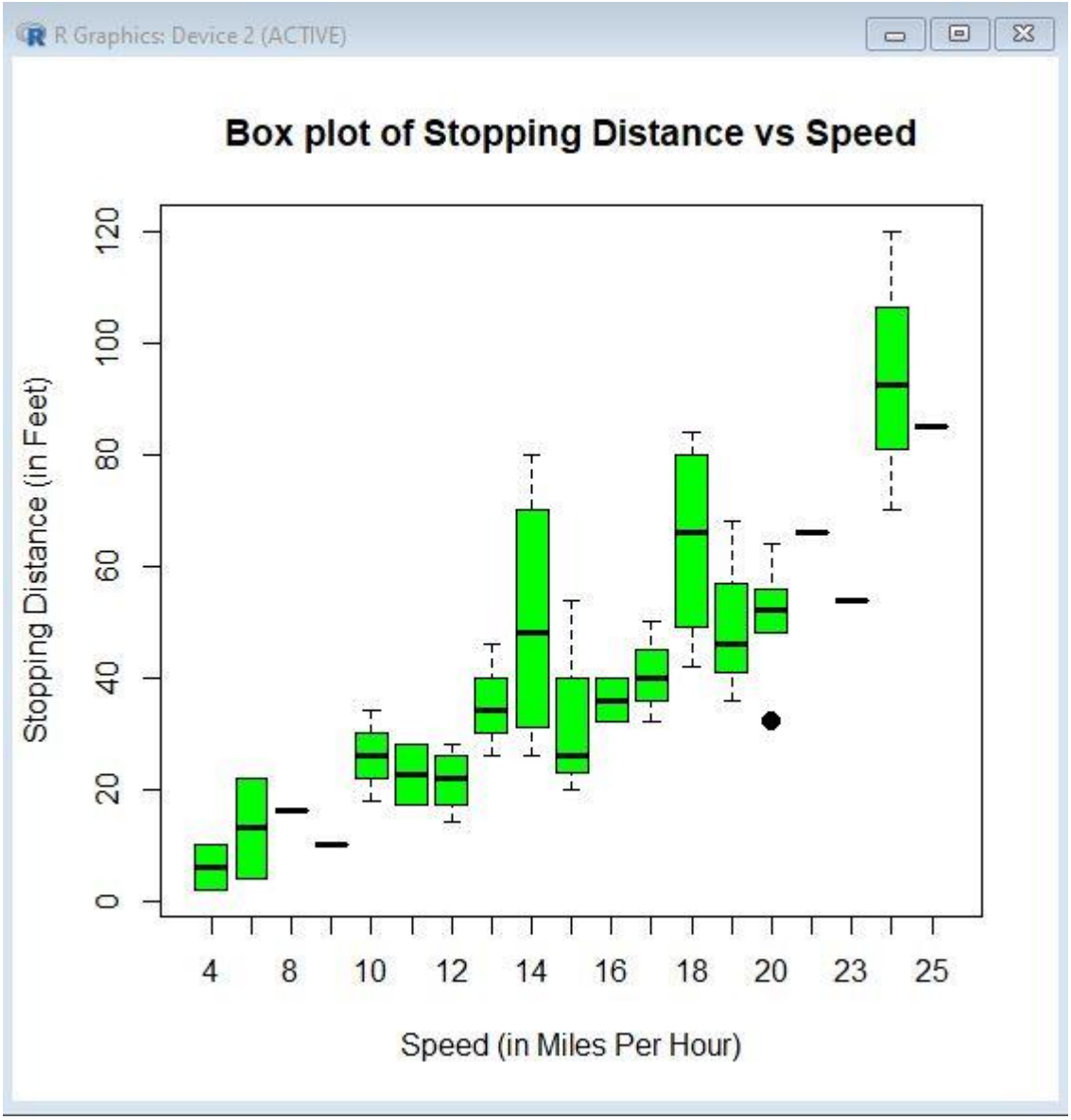
R Console

```
> Cars_model=lm(dist ~ speed,data=cars)
> plot(dist~speed, data = cars,
+ xlab = "Speed (in Miles Per Hour)",
+ ylab = "Stopping Distance (in Feet)",
+ main = " Scatter plot of Stopping Distance vs Speed",
+ pch = 20,
+ cex = 2,
+ col = "green")
> abline(Cars_model, lwd = 3, col = "darkred")
>
```



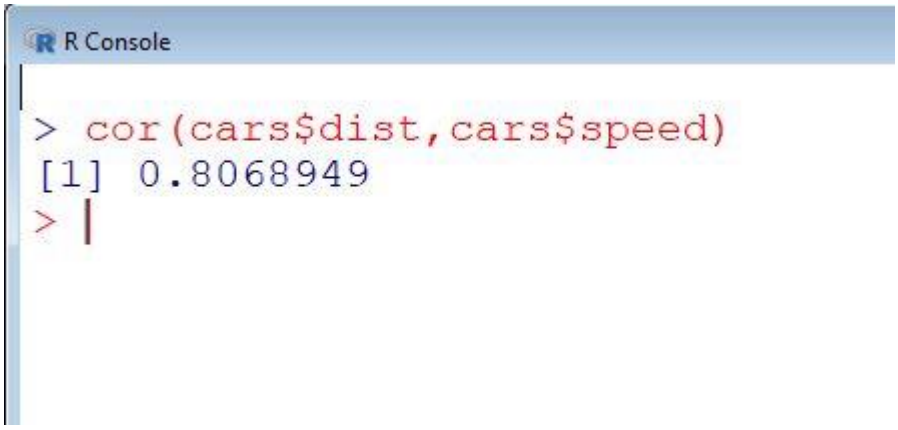
(b) Box plot: To spot any outlier observations in the variable. Having outliers in your predictor can drastically affect the predictions as they can easily affect the direction/slope of the line of best fit.



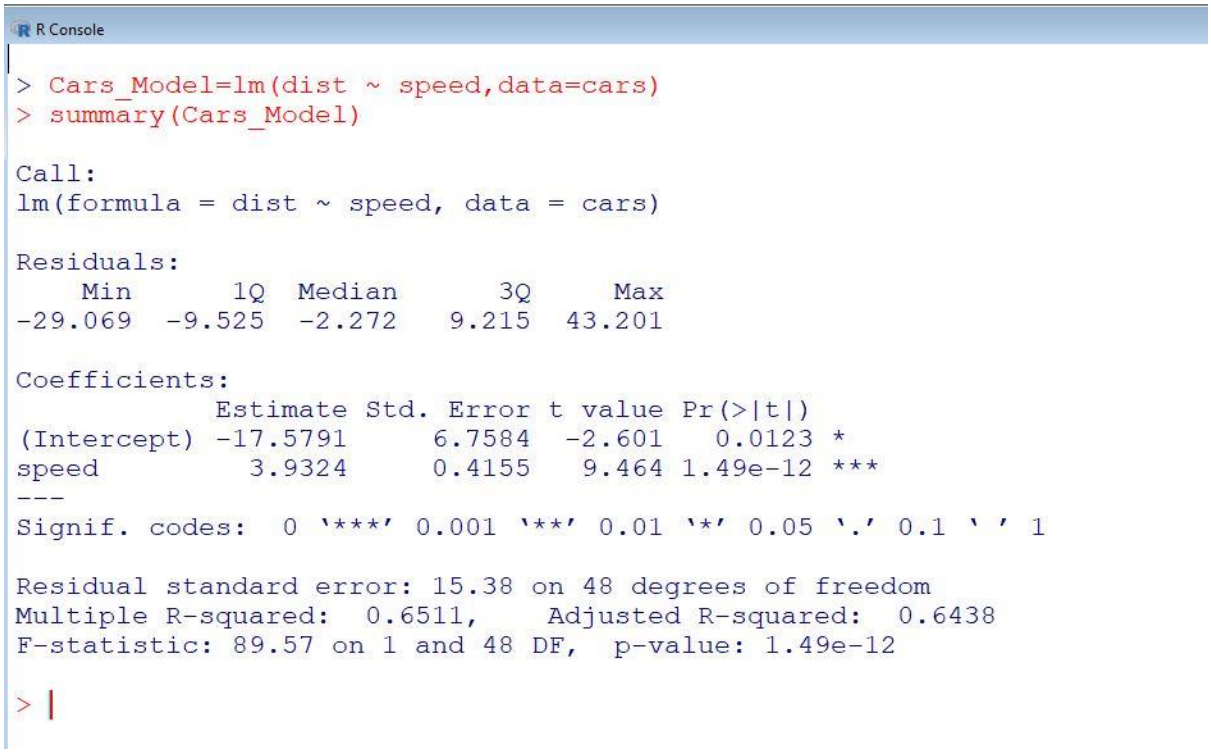

R Graphics: Device 2 (ACTIVE)
Box plot of Stopping Distance vs Speed
Stopping Distance (in Feet)
0
20
40
60
80
100
120
4
8
10
12
14
16
18
20
23
25
Speed (in Miles Per Hour)

B.

(a) Calculate correlation between speed and distance



(b) Build linear regression model on full data and Print the summary statistics for linear the Model.



6. For each problem below, state the random variables. Also, look to see if there are any outliers that need to be removed. Do the regression analysis with and without the suspected outlier points to determine if their removal affects the regression.

(a) When an anthropologist finds skeletal remains, they need to figure out the height of the person. The height of a person (in cm) and the length of their metacarpal bone 1 (in cm) were collected and are in table below ("Prediction of height," 2013).

Data of Metacarpal versus Height

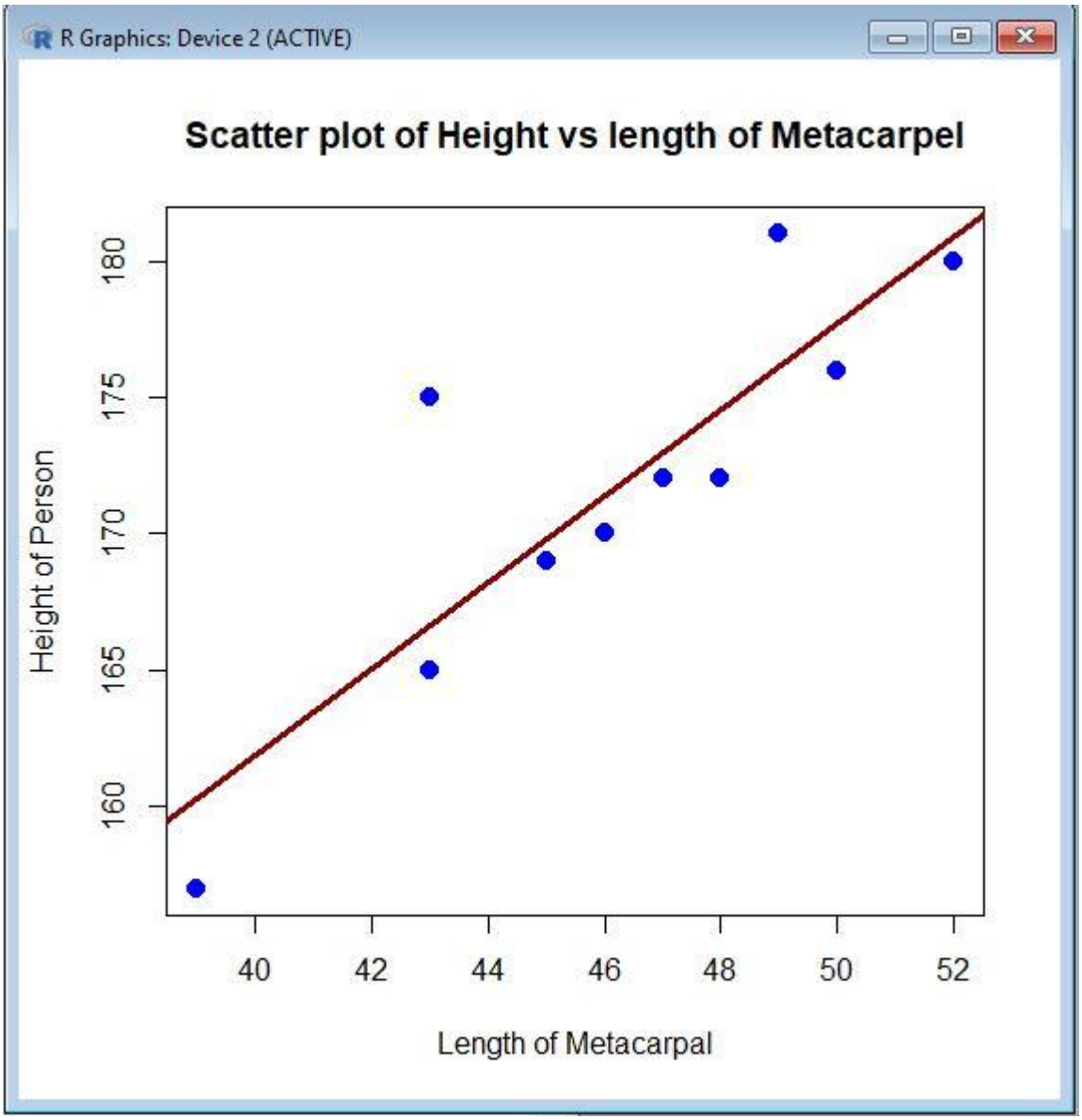
4

Length of Metacarpal (cm)	Height of Person (cm)
46	170
52	180
39	157
43	165
49	181
48	172
45	169
43	175
47	172
50	176

• Create a scatter plot and find a regression equation between the height of a person and the length of their metacarpal.

R Console

```
> a=lm(y~x)
> plot(y~x,
+ xlab="Length of Metacarpal",
+ ylab="Height of Person",
+ main="Scatter plot of Height vs length of Metacarpel",
+ pch = 20,
+ cex=2,
+ col = "blue")
> abline(a, lwd = 3, col = "darkred")
>
```



```
> lm(x~y)

Call:
lm(formula = x ~ y)

Coefficients:
(Intercept)            y
   -33.4574       0.4639

> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     98.527        1.584

> |
```

• Use the regression equation to find the height of a person for a metacarpal length of 44 cm and for a metacarpal length of 55 cm.

Ans: -

The height of a person for a metacarpal length of 44 cm is: -

1.584 * 44 + 98.527 = 168.223 cm.

The height of a person for a metacarpal length of 55 cm is: -

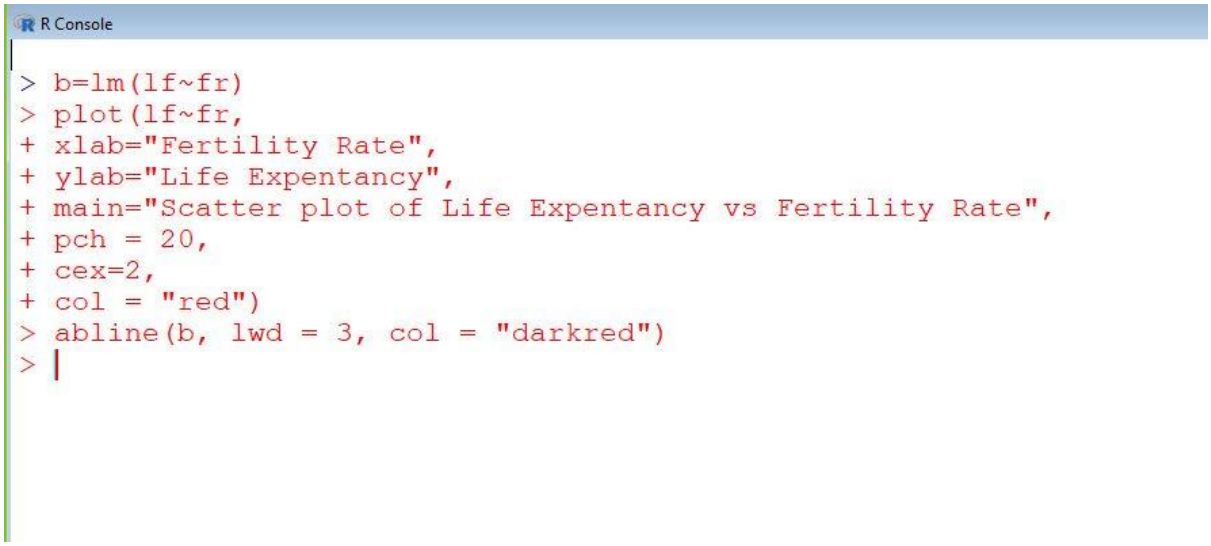
1.584*55 + 98.527 = 185.647 cm.

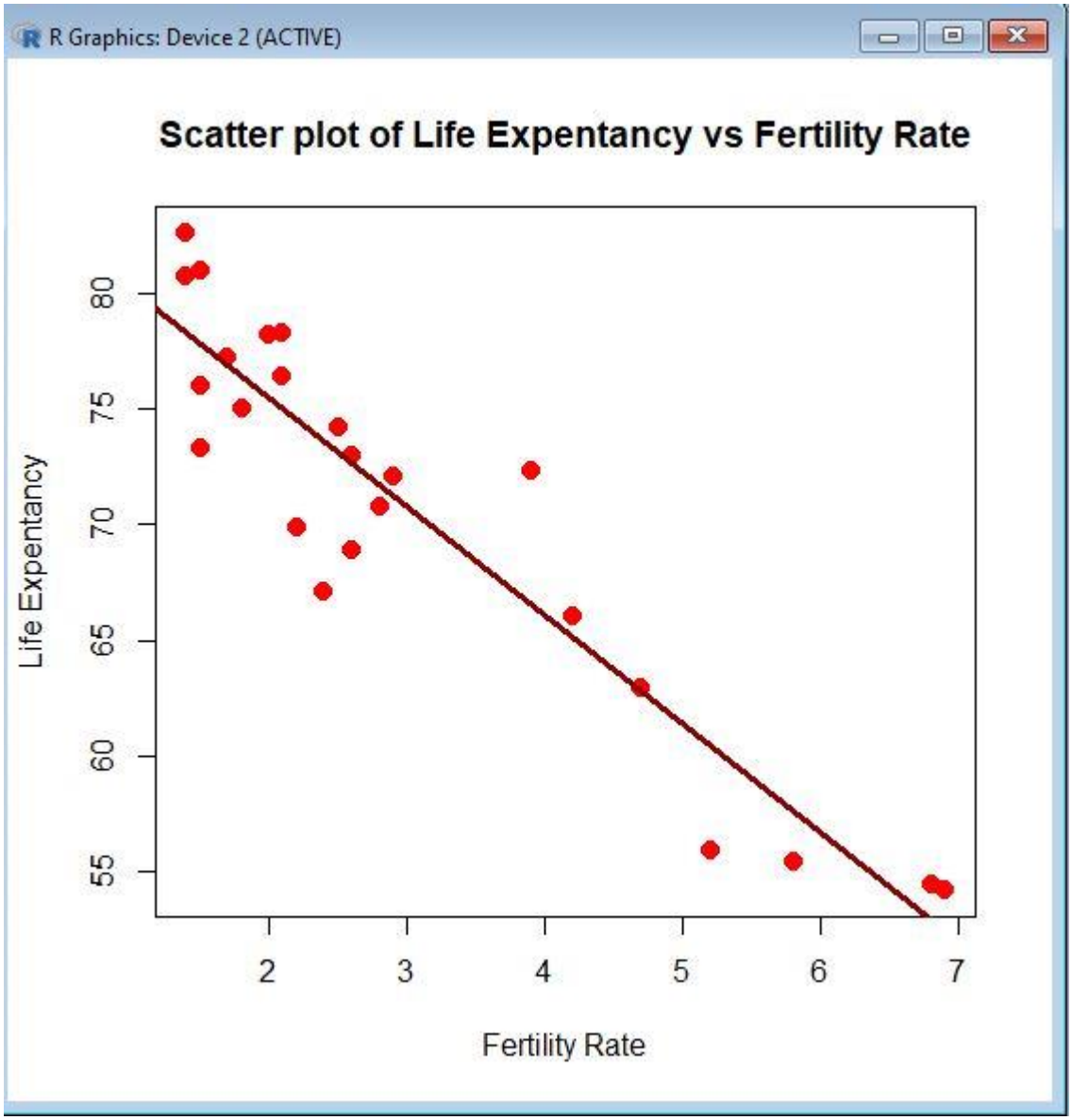
(b) The World Bank collects information on the life expectancy of a person in each country ("Life expectancy at," 2013) and the fertility rate per woman in the country ("Fertility rate," 2013). The data for 24 randomly selected countries for the year 2011 are in the table below.

Data of Fertility Rates versus Life Expectancy

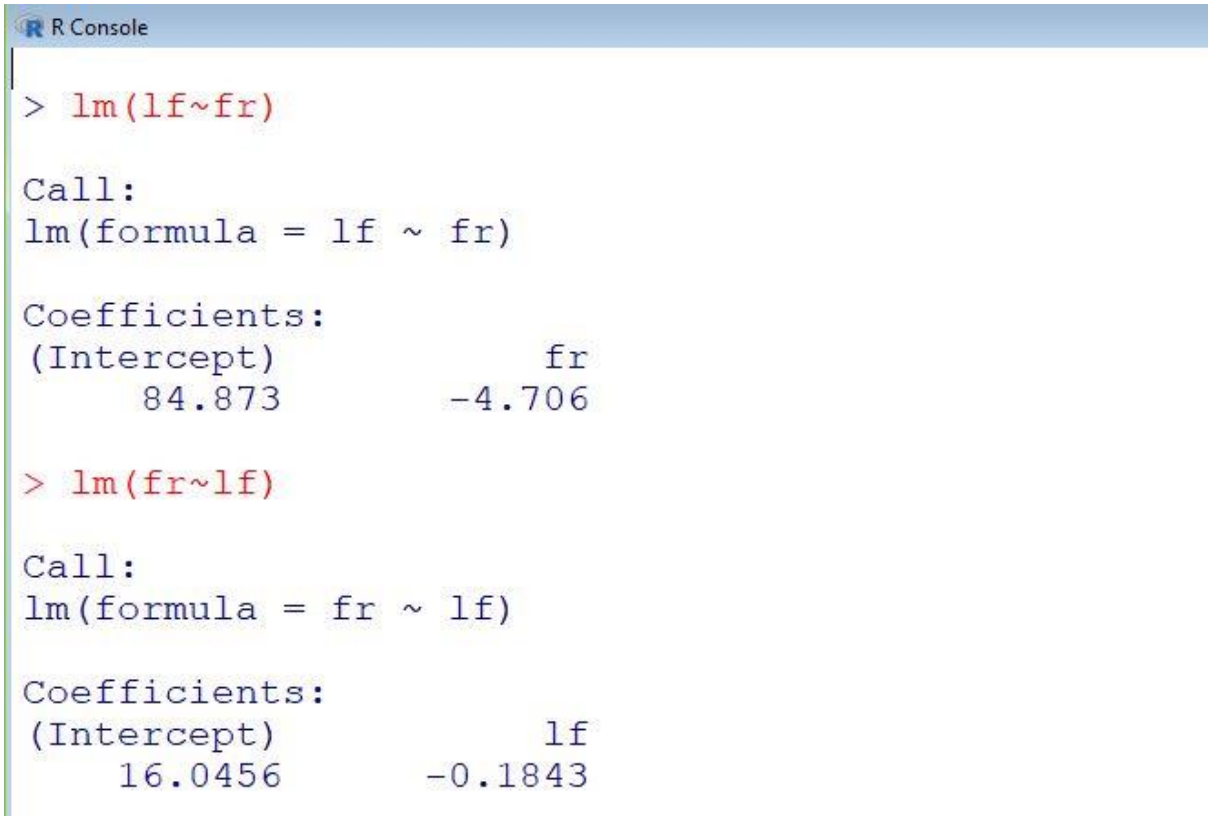
Fertility Rate	Life Expectancy
1.7	77.2
5.8	55.4
2.2	69.9
2.1	76.4
1.8	75.0
2.0	78.2
2.6	73.0
2.8	70.8
1.4	82.6
2.6	68.9
1.5	81.0
6.9	54.2
2.4	67.1
1.5	73.3
2.5	74.2
1.4	80.7
2.9	72.1
2.1	78.3
4.7	62.9
6.8	54.4
5.2	55.9
4.2	66.0
1.5	76.0
3.9	72.3

• Create a scatter plot of the data and find a linear regression equation between fertility rate and life expectancy.





• Use the regression equation to find the life expectancy for a country that has a fertility rate of 2.7 and for a country with fertility rate of 8.1.



Using the regression equation, the life expectancy for

a country that has a fertility rate of 2.7 is: -

2.7 * (-4.706) + 84.873 = 72.1668 years.

The life expectancy for a country that has a fertility rate of 8.1 is: -

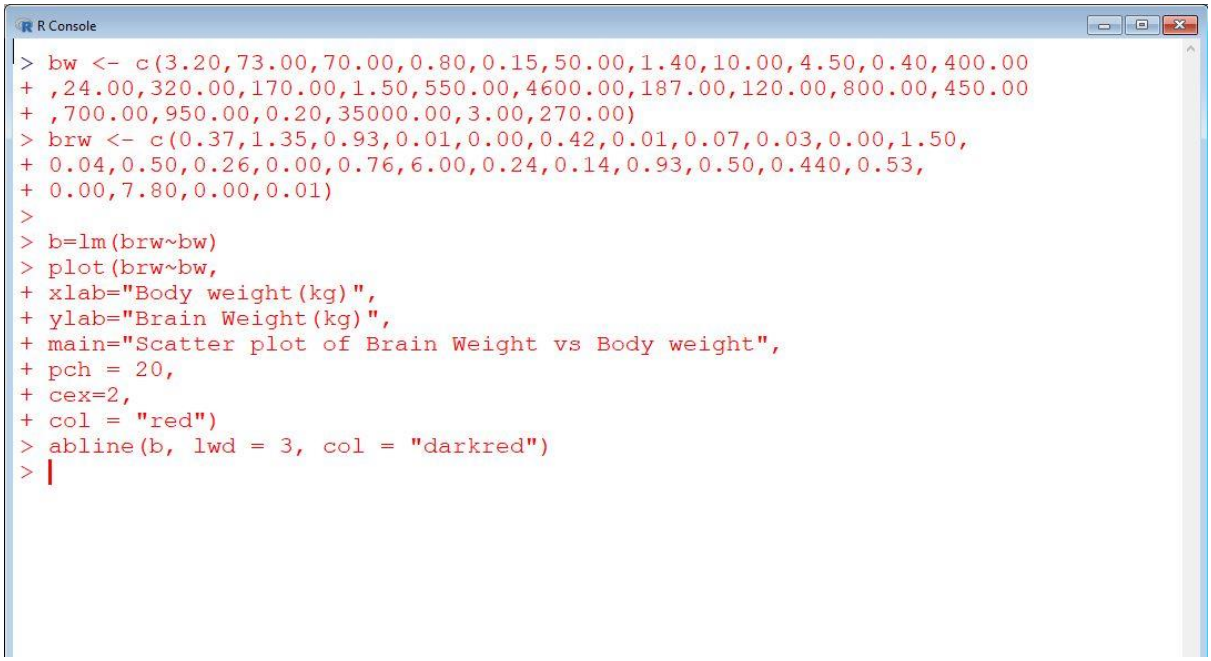
8.1 * (-4.706) + 84.873 = 46.7544 years.

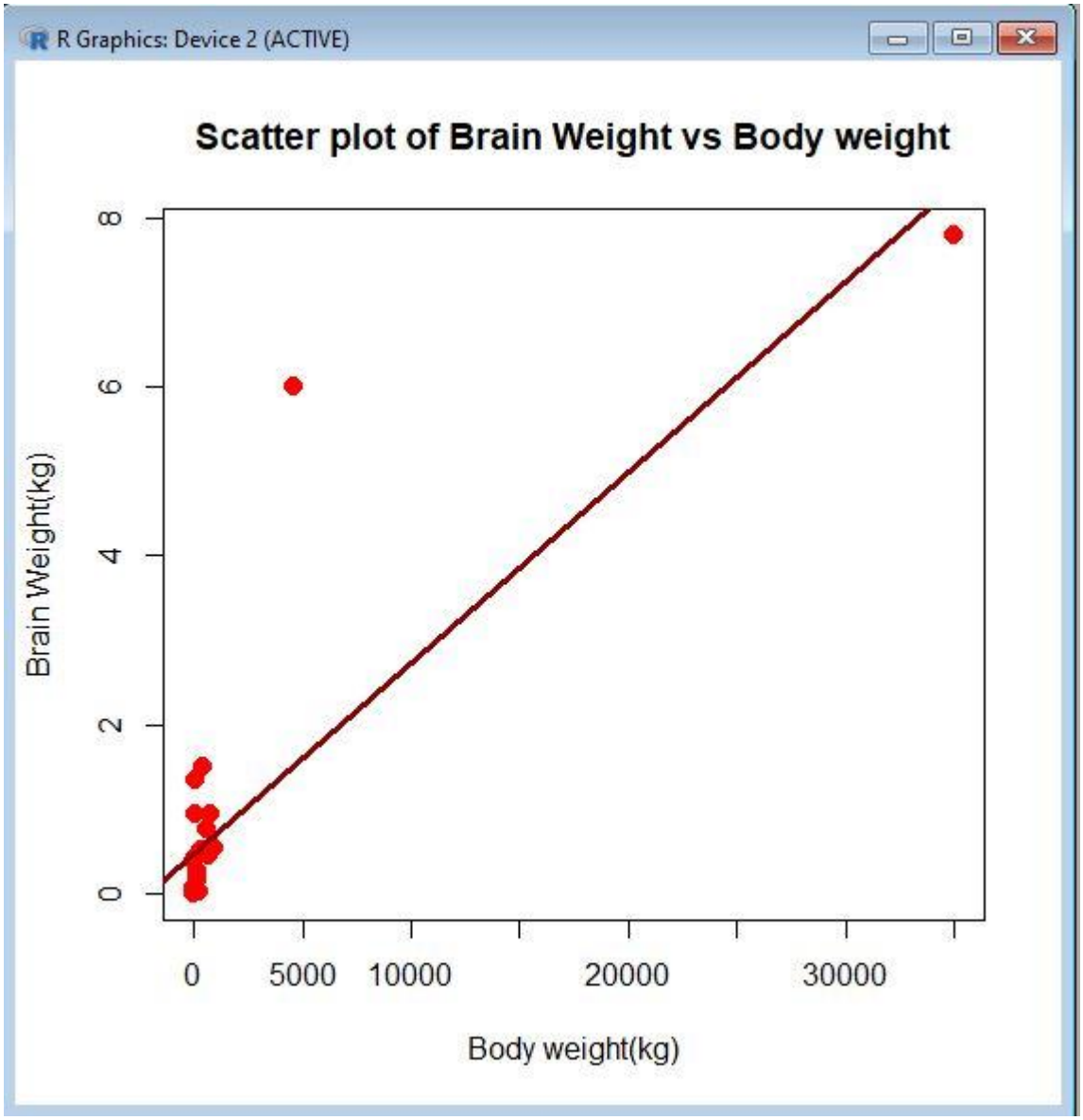
(c) Different species have different body weights and brain weights are in the table below ("Brain2bodyweight," 2013).

Body Weights and Brain Weights of Species

Species	Body Weight (kg)	Brain Weight (kg)
Newborn Human	3.20	0.37
Adult Human	73.00	1.35
Pithecanthropus Man	70.00	0.93
Squirrel	0.80	0.01
Hamster	0.15	0.00
Chimpanzee	50.00	0.42
Rabbit	1.40	0.01
Dog_(Beagle)	10.00	0.07
Cat	4.50	0.03
Rat	0.40	0.00
Bottle-Nosed Dolphin	400.00	1.50
Beaver	24.00	0.04
Gorilla	320.00	0.50
Tiger	170.00	0.26
Owl	1.50	0.00
Camel	550.00	0.76
Elephant	4600.00	6.00
Lion	187.00	0.24
Sheep	120.00	0.14
Walrus	800.00	0.93
Horse	450.00	0.50
Cow	700.00	0.440
Giraffe	950.00	0.53
Green Lizard	0.20	0.00
Sperm Whale	35000.00	7.80
Turtle	3.00	0.00
Alligator	270.00	0.01

i. Create a scatter plot and find a regression equation between body weights and brain weights.





ii. Use the regression equation to find the brain weight for a species that has a body weight of 62 kg and for a species that has a body weight of 180,000 kg.

Ans: -

```
> lm(bw~brw)

Call:
lm(formula = bw ~ brw)

Coefficients:
(Intercept)          brw
     -978.9        3116.9

> lm(brw~bw)

Call:
lm(formula = brw ~ bw)

Coefficients:
(Intercept)           bw
  0.4706281    0.0002264

>
```

Using the regression equation, the brain weight of species having body weight of 62kg is: -

62 * 0.0002264 + 0.4706281 = 0.48466 kg.

The brain weight of species having body weight of 180000 kg is: -

180000 * 0.0002264 + 0.4706281 = 41.22 kg.

7. A researcher is interested in how variables, such as GRE(Graduate Record Exam scores), GPA (Grade Point Average) and the prestige of the undergraduate institution affect admission into graduate school. The response variable, Admit/Don't Admit, is a binary variable. A set of hypothetical data is available at https://stats.idre.ucla.edu/stat/data/binary.csv Estimate a logistic regression model and use the summary() to show the details of the estimated statistics.

Sol: -

```
> library(rio)
The following rio suggested packages are not installed: 'arrow', 'feather', 'fst', 'hexView', 'pzfx', 'readODS', 'rmatio'
Use 'install_formats()' to install them
Warning message:
package 'rio' was built under R version 4.0.5
> library(caret)
Loading required package: lattice
Loading required package: ggplot2
Warning messages:
1: package 'caret' was built under R version 4.0.5
2: package 'ggplot2' was built under R version 4.0.4
> data <- import("binary.sas7bdat")
> data$ADMIT <- as.factor(data$ADMIT)
> data$RANK <- as.factor(data$RANK)
> set.seed(125)
> ind <- createDataPartition(data$ADMIT,p=0.80,list = FALSE)
> training <- data[ind,]
> testing <- data[-ind,]
> set.seed(123)
> mymodel <- glm(ADMIT~GPA + RANK,data=training,family=binomial(link = "logit"))
```

```
> summary(mymodel)

Call:
glm(formula = ADMIT ~ GPA + RANK, family = binomial(link = "logit"),
    data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4690  -0.8923  -0.6721   1.1925   2.1362

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.0613     1.1908  -2.571 0.010143 *
GPA           0.9312     0.3330   2.796 0.005167 **
RANK2        -0.6990     0.3536  -1.977 0.048046 *
RANK3        -1.2356     0.3738  -3.305 0.000950 ***
RANK4        -1.7759     0.4868  -3.648 0.000264 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 401.36  on 320  degrees of freedom
Residual deviance: 373.02  on 316  degrees of freedom
AIC: 383.02

Number of Fisher Scoring iterations: 4

>
```

8. Consider a dataset of 144 observations of household cats. The data contains the cats' gender, body weight and height. The dataset is shipped with R and is named as cats.csv. It is available at C:/Users/aarschle1/Google Drive/Projects /Blog/posts/classification_r/logistic/cats.csv Plot the data and observe the relationship between the body weight and height of a cat and its gender.

(a) Fit a logistic regression, model to the data.

```
> summary(cats)
  Sex          Bwt             Hwt
  F:47   Min.   :2.000   Min.   : 6.30
  M:97   1st Qu.:2.300   1st Qu.: 8.95
         Median :2.700   Median :10.10
         Mean   :2.724   Mean   :10.63
         3rd Qu.:3.025   3rd Qu.:12.12
         Max.   :3.900   Max.   :20.50
> cats$Sex.f <- factor(cats$Sex)
> contrasts(cats$Sex.f)
  M
F 0
M 1
```

(b) verify and test our model's performance

```
> inTrain <- createDataPartition(y = cats$Sex.f, p = .60, list = FALSE)
> training <- cats[inTrain,]
> testing <- cats[-inTrain,]
> dim(training)
[1] 88  4
> dim(testing)
[1] 56  4
```

(c) Call the predict () function make predictions of cats' gender on new data.

```
> cats.fit = glm(Sex.f ~ Bwt + Hwt, data=training, family=binomial)
> predict(cats.fit, newdata=data.frame(Bwt=c(2.8, 1.8), Hwt=c(13, 7)), type="response")
        1         2
0.85110044 0.08506393
>
```

9. By use of the logistic regression equation of vehicle transmission in the data set mtcars, estimate the probability of a vehicle being fitted with a manual transmission if it has a 120hp engine and weights 2800 lbs.

Sol: -

```
> am.glm = glm(formula=am ~ hp + wt,
+ data=mtcars,
+ family=binomial)
> newdata = data.frame(hp=120, wt=2.8)
> predict(am.glm, newdata, type="response")
        1
0.6418125
>
```

For an automobile with 120hp engine and 2800 lbs weight, the probability of it being fitted with a manual transmission is about 64%.

10. The muscle dataset is an old data from an experiment on muscle contraction in experimental animals. The data is explained below. The data consists of the variables: Strip (identifier of muscle), Conc(CaCl concentrations used to soak the section and Length (resulting length of muscle section for each concentration). Let us consider a non-linear model for this data of the form $L = \alpha + \beta e^{(-C/\theta)} + error$, where α and β may vary with the animal but θ is constant. Note that α and β are linear parameters. Muscle Contraction in Rat Hearts Description: Experiment to assess the influence of calcium chloride on the contraction of the heart muscle of 21 rats.

Variables

Name	Description	Mode
Strip	heart muscle strip (S01-S21)	factor
Conc	Concentration of calcium chloride solution	numeric
Length	Change in length	numeric

R Object: muscle Location and Source: This dataset is part of the MASS package . P.Fit a regression line to the dataset.

Sol: -

```
> A <- model.matrix(~ Strip - 1, data=muscle)
> rats.nls1 <- nls(log(Length) ~ cbind(A, rho^Conc),
+ data = muscle, start = c(rho=0.1), algorithm="plinear")
> (B <- coef(rats.nls1))
         rho .lin.StripS01 .lin.StripS02 .lin.StripS03 .lin.StripS04 .lin.StripS05 .lin.StripS06 .lin.StripS07
  0.07776401    3.08304824    3.30137838    3.44562531    2.80464434    2.60835015    3.03357725    3.52301734
.lin.StripS08 .lin.StripS09 .lin.StripS10 .lin.StripS11 .lin.StripS12 .lin.StripS13 .lin.StripS14 .lin.StripS15
  3.38711844    3.46709396    3.81438456    3.73878664    3.51332581    3.39741115    3.47088608    3.72895847
.lin.StripS16 .lin.StripS17 .lin.StripS18 .lin.StripS19 .lin.StripS20 .lin.StripS21        .lin22
  3.31863862    3.37938673    2.96452195    3.58468686    3.39628029    3.36998872   -2.96015460
```

```
> st <- list(alpha = B[2:22], beta = B[23], rho = B[1])
> (rats.nls2 <- nls(log(Length) ~ alpha[Strip] + beta*rho^Conc,
+ data = muscle, start = st))
Nonlinear regression model
  model: log(Length) ~ alpha[Strip] + beta * rho^Conc
   data: muscle
alpha..lin.StripS01 alpha..lin.StripS02 alpha..lin.StripS03 alpha..lin.StripS04 alpha..lin.StripS05 alpha..lin.StripS06
            3.08305             3.30138             3.44563             2.80464             2.60835             3.03358
alpha..lin.StripS07 alpha..lin.StripS08 alpha..lin.StripS09 alpha..lin.StripS10 alpha..lin.StripS11 alpha..lin.StripS12
            3.52302             3.38712             3.46709             3.81438             3.73879             3.51333
alpha..lin.StripS13 alpha..lin.StripS14 alpha..lin.StripS15 alpha..lin.StripS16 alpha..lin.StripS17 alpha..lin.StripS18
            3.39741             3.47089             3.72896             3.31864             3.37939             2.96452
alpha..lin.StripS19 alpha..lin.StripS20 alpha..lin.StripS21         beta..lin22             rho.rho
            3.58469             3.39628             3.36999            -2.96015             0.07776
 residual sum-of-squares: 1.045

Number of iterations to convergence: 0
Achieved convergence tolerance: 4.923e-06
```

```
> Muscle <- with(muscle, {
+ Muscle <- expand.grid(Conc = sort(unique(Conc)), Strip = levels(Strip))
+ Muscle$Yhat <- predict(rats.nls2, Muscle)
+ Muscle <- cbind(Muscle, logLength = rep(as.numeric(NA), 126))
+ ind <- match(paste(Strip, Conc),
+ paste(Muscle$Strip, Muscle$Conc))
+ Muscle$logLength[ind] <- log(Length)
+ Muscle})
> lattice::xyplot(Yhat ~ Conc | Strip, Muscle, as.table = TRUE,
+ ylim = range(c(Muscle$Yhat, Muscle$logLength), na.rm = TRUE),
+ subscripts = TRUE, xlab = "Calcium Chloride concentration (mM)",
+ ylab = "log(Length in mm)", panel =
+ function(x, y, subscripts, ...) {
+ panel.xyplot(x, Muscle$logLength[subscripts], ...)
+ llines(spline(x, y))
+ })
>
```