1A.

> colMeans(mtcars)

mpg cyl disp hp drat wt qsec vs am gear carb

20.090625 6.187500 230.721875 146.687500 3.596563 3.217250 17.848750 0.437500 0.406250 3.687500 2.812500

> apply(mtcars,2,median)

mpg cyl disp hp drat wt qsec vs am gear carb

19.200 6.000 196.300 123.000 3.695 3.325 17.710 0.000 0.000 4.000 2.000

1B.

> apply(mtcars,2,mad) (MAD)

mpg cyl disp hp drat wt qsec vs am gear

5.4114900 2.9652000 140.4763500 77.0952000 0.7042350 0.7672455 1.4158830 0.0000000 0.0000000 1.4826000

carb

1.4826000

> apply(mtcars, 2, function(x) mean(abs(x - mean(x)))) (AAD)

mpg cyl disp hp drat wt qsec vs am gear

4.7144531 1.5859375 108.7857422 56.4804688 0.4532422 0.7301875 1.3761719 0.4921875 0.4824219 0.6445312

carb

1.3007812

1C.

> meanMpg = sum(mtcars$mpg) / length(mtcars$mpg)

>

> varMpg = sum((mtcars$mpg - meanMpg)^2) / (length(mtcars$mpg) - 1)

> sdMpg = sqrt(varMpg)

>

> varMpg

[1] 36.3241

> var(mtcars$mpg)

[1] 36.3241

>

> sdMpg

[1] 6.026948

> sd(mtcars$mpg

1D.

> skew <- (sum((mtcars$mpg - mean(mtcars$mpg))^3) / (length(mtcars$mpg) \* sd(mtcars$mpg)^3))

>

> kurtosis <- (sum((mtcars$mpg - mean(mtcars$mpg))^4) / (length(mtcars$mpg) \* sd(mtcars$mpg)^4)) - 3

>

>

> skew

[1] 0.610655

> skew(mtcars$mpg)

[1] 0.610655

>

> kurtosis

[1] -0.372766

> kurtosi(mtcars$mpg)

[1] -0.372766

1E.

> apply(mtcars,2,kurtosi)

mpg cyl disp hp drat wt qsec vs am gear

-0.37276603 -1.76211977 -1.20721195 -0.13555112 -0.71470062 -0.02271075 0.33511422 -2.00193762 -1.92474143 -1.06975068

carb

1.25704307

> apply(mtcars,2,skew)

mpg cyl disp hp drat wt qsec vs am gear carb

0.6106550 -0.1746119 0.3816570 0.7260237 0.2659039 0.4231465 0.3690453 0.2402577 0.3640159 0.5288545 1.0508738

1F.

> pearsonWTMPG <- sum((mtcars$wt - mean(mtcars$wt)) \* (mtcars$mpg - mean(mtcars$mpg))) / (sqrt(sum((mtcars$wt - mean(mtcars$wt))^2)) \* sqrt(sum((mtcars$mpg - mean(mtcars$mpg))^2)))

> pearsonWTMPG = sum((mtcars$wt - mean(mtcars$wt)) \* (mtcars$mpg - mean(mtcars$mpg))) / (sqrt(sum((mtcars$wt - mean(mtcars$wt))^2)) \* sqrt(sum((mtcars$mpg - mean(mtcars$mpg))^2)))

> spearmanWTMPG = 1 - (6 \* sum((rank(mtcars$wt) - rank(mtcars$mpg))^2) / (length(mtcars$wt) \* (length(mtcars$wt)^2 - 1)))

>

> pearsonCYLMPG = sum((mtcars$cyl - mean(mtcars$cyl)) \* (mtcars$mpg - mean(mtcars$mpg))) / (sqrt(sum((mtcars$cyl - mean(mtcars$cyl))^2)) \* sqrt(sum((mtcars$mpg - mean(mtcars$mpg))^2)))

> spearmanCYLMPG = 1 - (6 \* sum((rank(mtcars$cyl) - rank(mtcars$mpg))^2) / (length(mtcars$cyl) \* (length(mtcars$cyl)^2 - 1)))

>

> pearsonGEARWT = sum((mtcars$gear - mean(mtcars$gear)) \* (mtcars$wt - mean(mtcars$wt))) / (sqrt(sum((mtcars$gear - mean(mtcars$gear))^2)) \* sqrt(sum((mtcars$wt - mean(mtcars$wt))^2)))

> spearmanGEARWT = 1 - (6 \* sum((rank(mtcars$gear) - rank(mtcars$wt))^2) / (length(mtcars$gear) \* (length(mtcars$gear)^2 - 1)))

>

> pearsonWTMPG

[1] -0.8676594

> spearmanWTMPG

[1] -0.8843475

>

> pearsonCYLMPG

[1] -0.852162

> spearmanCYLMPG

[1] -0.7794172

>

> pearsonGEARWT

[1] -0.583287

> spearmanGEARWT

[1] -0.5400477

2A.

> cor(x = anscombe$x1, y = anscombe$y1, method = "pearson")

[1] 0.8164205

> cor(x = anscombe$x1, y = anscombe$y1, method = "pearson")

[1] 0.8164205

> cor(x = anscombe$x1, y = anscombe$y1, method = "spearman")

[1] 0.8181818

> cor(x = anscombe$x2, y = anscombe$y2, method = "pearson")

[1] 0.8162365

> cor(x = anscombe$x2, y = anscombe$y2, method = "spearman")

[1] 0.6909091

> cor(x = anscombe$x3, y = anscombe$y3, method = "pearson")

[1] 0.8162867

> cor(x = anscombe$x3, y = anscombe$y3, method = "spearman")

[1] 0.9909091

> cor(x = anscombe$x4, y = anscombe$y4, method = "pearson")

[1] 0.8165214

> cor(x = anscombe$x4, y = anscombe$y4, method = "spearman")

[1] 0.5

They are all around 0.8ish

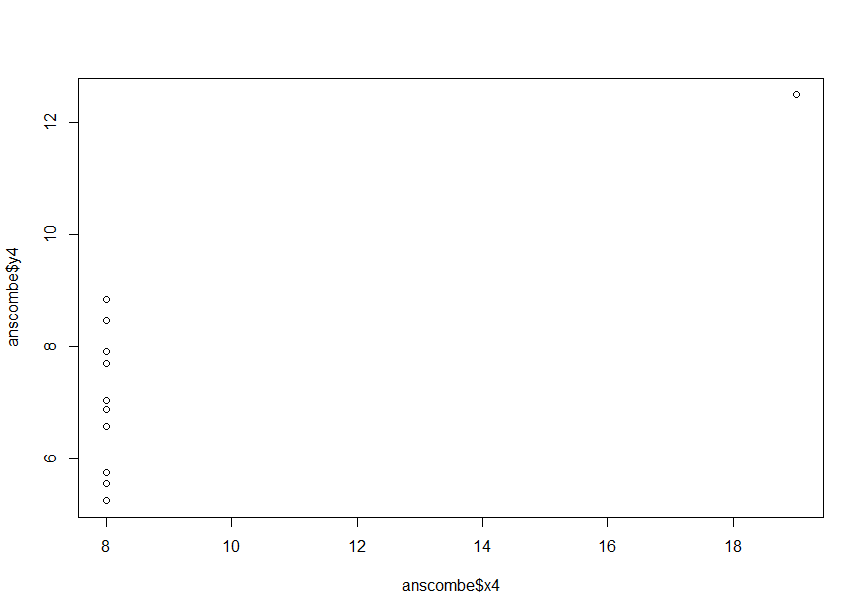
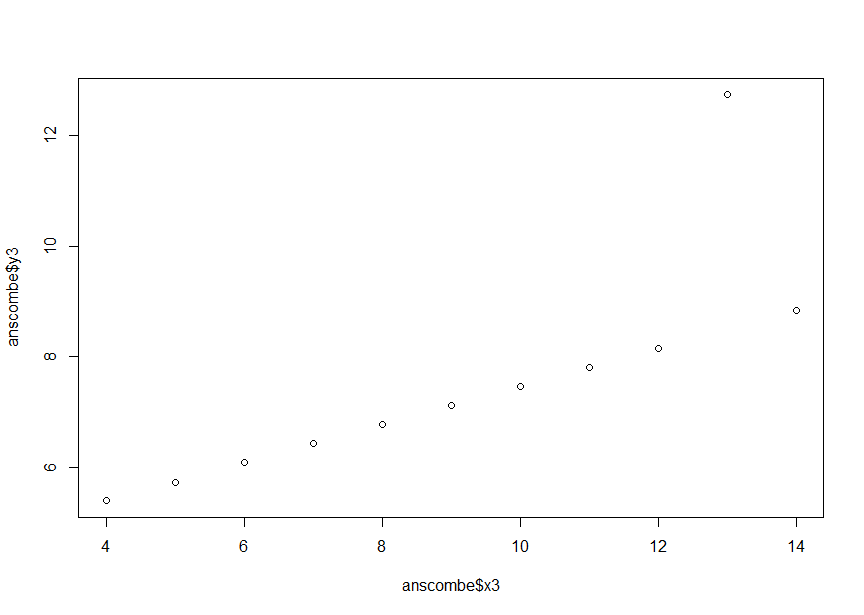
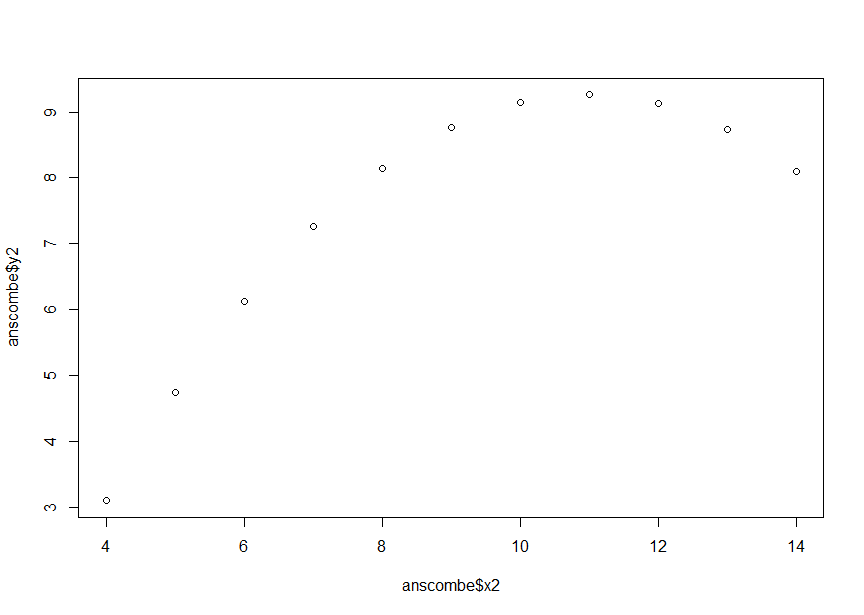
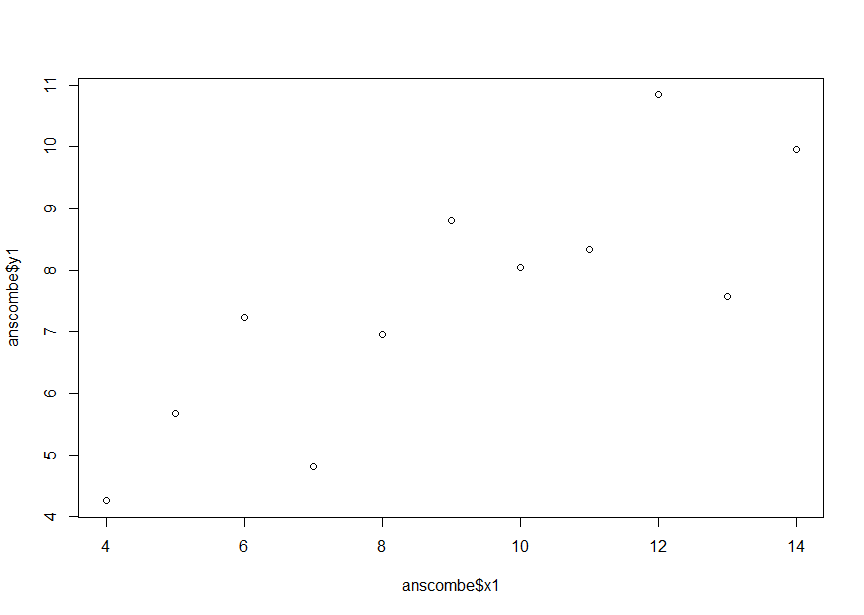
2B.

> plot(anscombe$x1,anscombe$y1)

> plot(anscombe$x2,anscombe$y2)

> plot(anscombe$x3,anscombe$y3)

> plot(anscombe$x4,anscombe$y4)



3A.

Assert that we have X and Y for every sample (S) and a constant C, Y\_S is equal to the C times X\_S.

so Y\_S = C\*X\_S for all samples.

pearson formula is r = cov(X, Y) / (sd(X) \* sd(Y))

so if we do the covariance we have cov(X, Y) = E[(X - E[X])(Y - E[Y])]

but Since Y\_S = C\*X\_S for all samples, we have:

E[Y] = E[cX] = cE[X]

so we can simplify it to C\*var(X)

the same thing applies to the standard deviation : sd(X) = sqrt(var(X))

so we can write it as r = cov(X, Y) / (sd(X) \* sd(Y)) which simplfies to 1.

so no matter the value of the constant, Y\_S = C\*X\_S for all samples, this is likely because Y is a linear function of X with a positive slope. The variables have a near perfect positive linear slope.  
  
(i’m assuming this is what i’m supposed to do.)  
  
3B.

Assert X and Y and for every sample S, Y\_S is equal to some constant C. So we have Y\_S = C for all samples (S).

Pearson formula = r = cov(X, Y) / (sd(X) \* sd(Y))

first calculate the covar of X and Y. covariance is: cov(X, Y) = E[(X - E[X])(Y - E[Y])]

where E[X] and E[Y] are the means of X and Y.

Since Y\_S = C for all S, we have:

E[Y] = E[c] = C

So the covariance can be simplified to: 0

Therefore, the Pearson correlation coefficient can be written as:

r = cov(X, Y) / (sd(X) \* sd(Y)) or 0

So the Pearson correlation coefficient between X and Y is equal to 0 when Y is a constant. This is because, the two variables do not have a linear relationship.

4.

Using only the mean or median to describe a set is often insufficient because it only provides a single measure of central tendency and does not really capture the full picture. It also doesn’t account for outliers correctly.

5.

The opposite applies here, using the mean and/or standard deviation if often sensitive to outliers or extremes. Also it’s not very good for abnormal distributions of data.