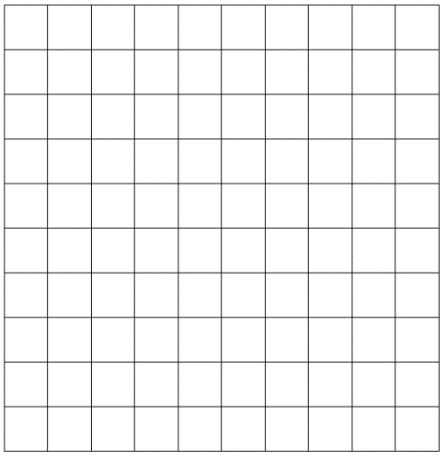
CHAPTER 6 – EXPLORING DATA RELATIONSHIPS

A ***response variable*** measures an outcome or result of a study.
An ***explanatory variable*** is a variable that we think explains or causes changes in the response variable.

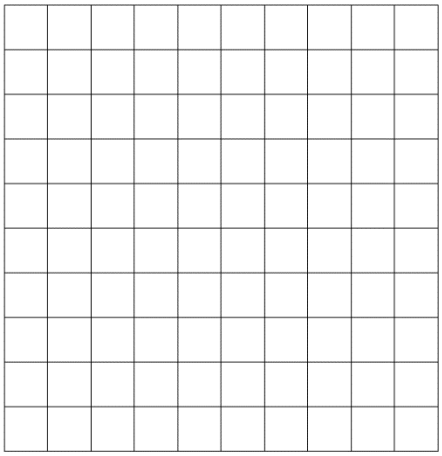
For the studies shown below, decide which is the explanatory variable and which is the response variable. Then display the data in a scatterplot.
Note: Data used in these notes are fictitious.

(a) The average speed of the car and the average miles per gallon for the trip were recorded as follows:



MPG	36	33	27	24
Speed of Car	55	60	70	75

(b) A study was done on the number of hours students spent studying and their grade on the exam. The results were:



Hours of study	2	4	6	7	9	6
Grade on Exam	65	70	85	86	92	94

Two variables are ***positively associated*** if an *increase* in one variable tends to accompany an *increase* in the other variable.

Two variables are ***negatively associated*** if an *increase* in one variable tends to accompany a *decrease* in the other variable.

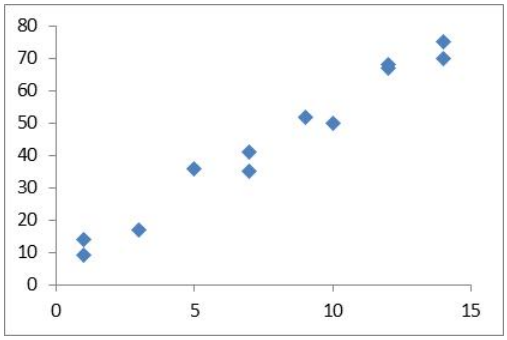
How is the speed of the car related to gas mileage?

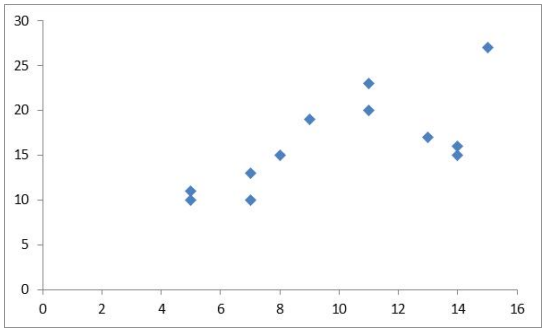
How is the amount time spent studying related to the grade on the exam?

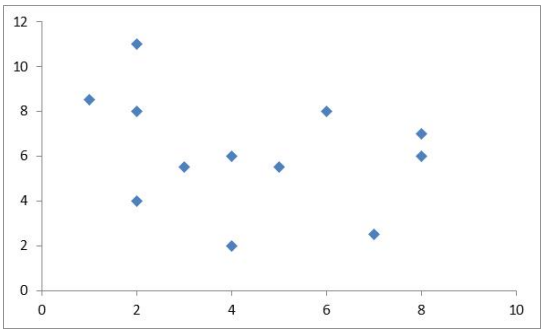
A ***regression line*** is a straight line that describes how the response variable changes as the explanatory variable changes.

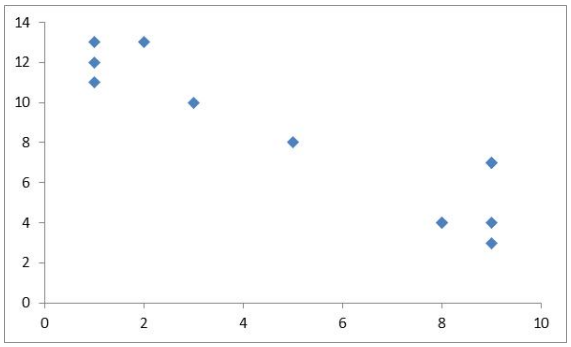
The regression line is a line that is as close as possible to all the points.

Sketch the regression line for the scatterplots below:









The ***correlation*** measures the direction and strength of the *straight line* relationship between two numerical variables.

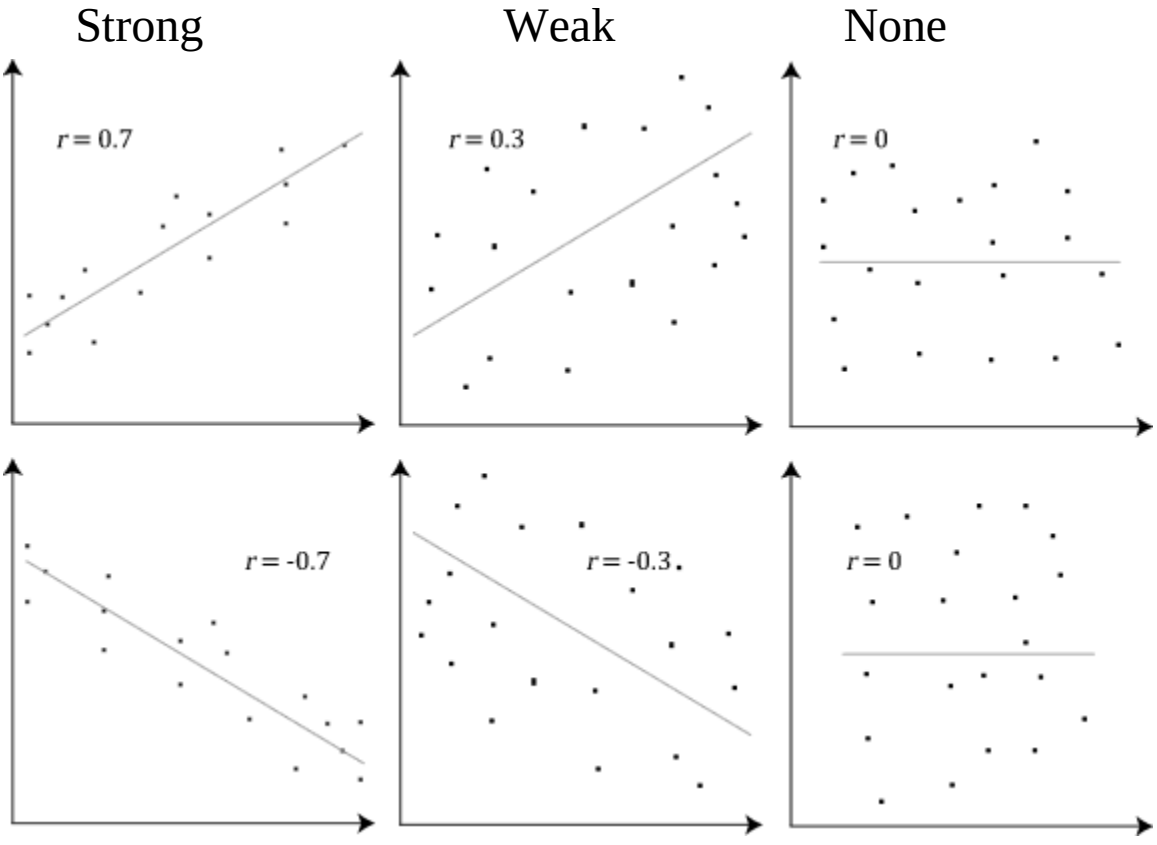
The value of the correlation is a number r (called the correlation coefficient) that is between -1 and 1, inclusive. That is, $-1 \le r \le 1$.

For positive association, $r > 0$.

For negative association, $r < 0$.

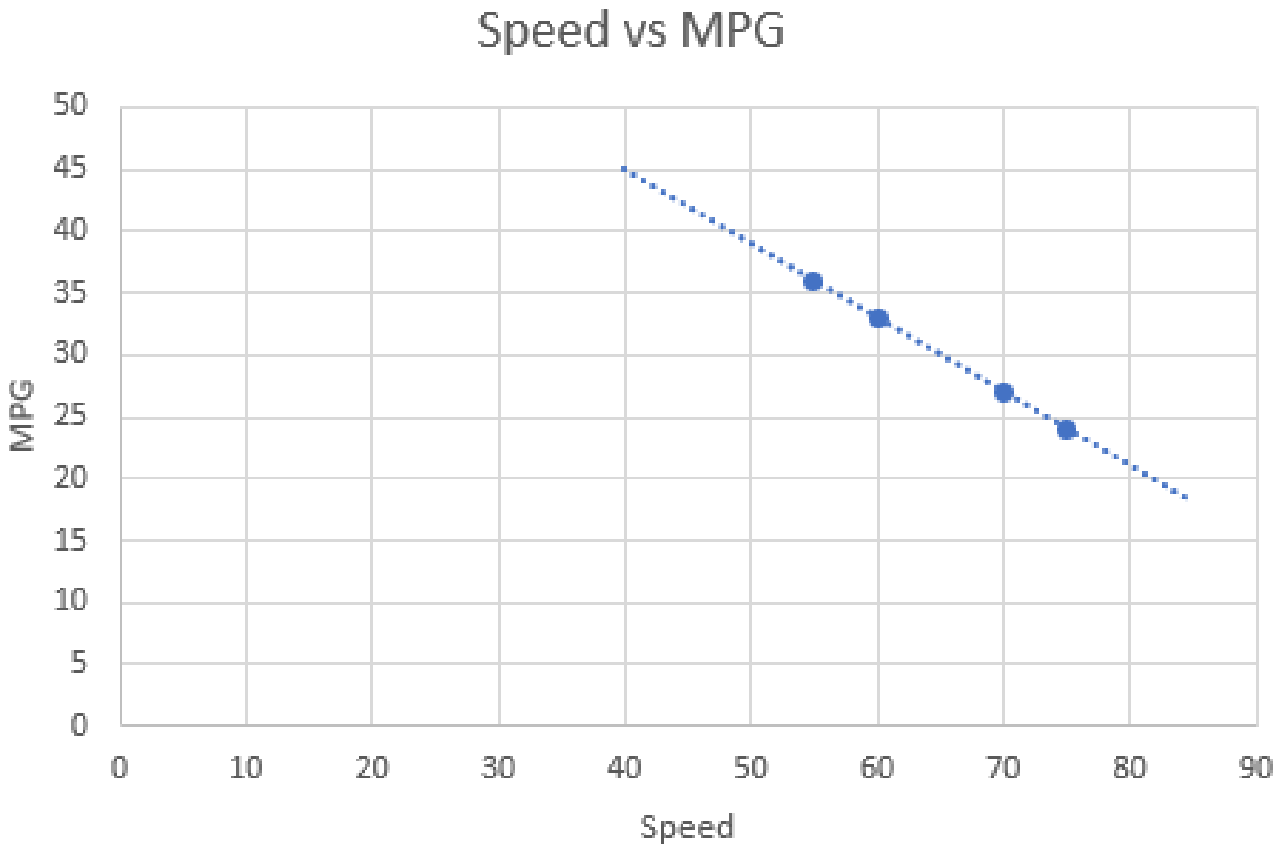
For no linear association, $r = 0$.

The closer $|r|$ is to 1, the stronger the association.



Interpolation is using the regression line to find values *between* the minimum and maximum data values.
Extrapolation is using the regression line to find values that are *outside* the minimum and maximum values.

Use the given regression line to make predictions about the variables and tell whether you are using interpolation or extrapolation.

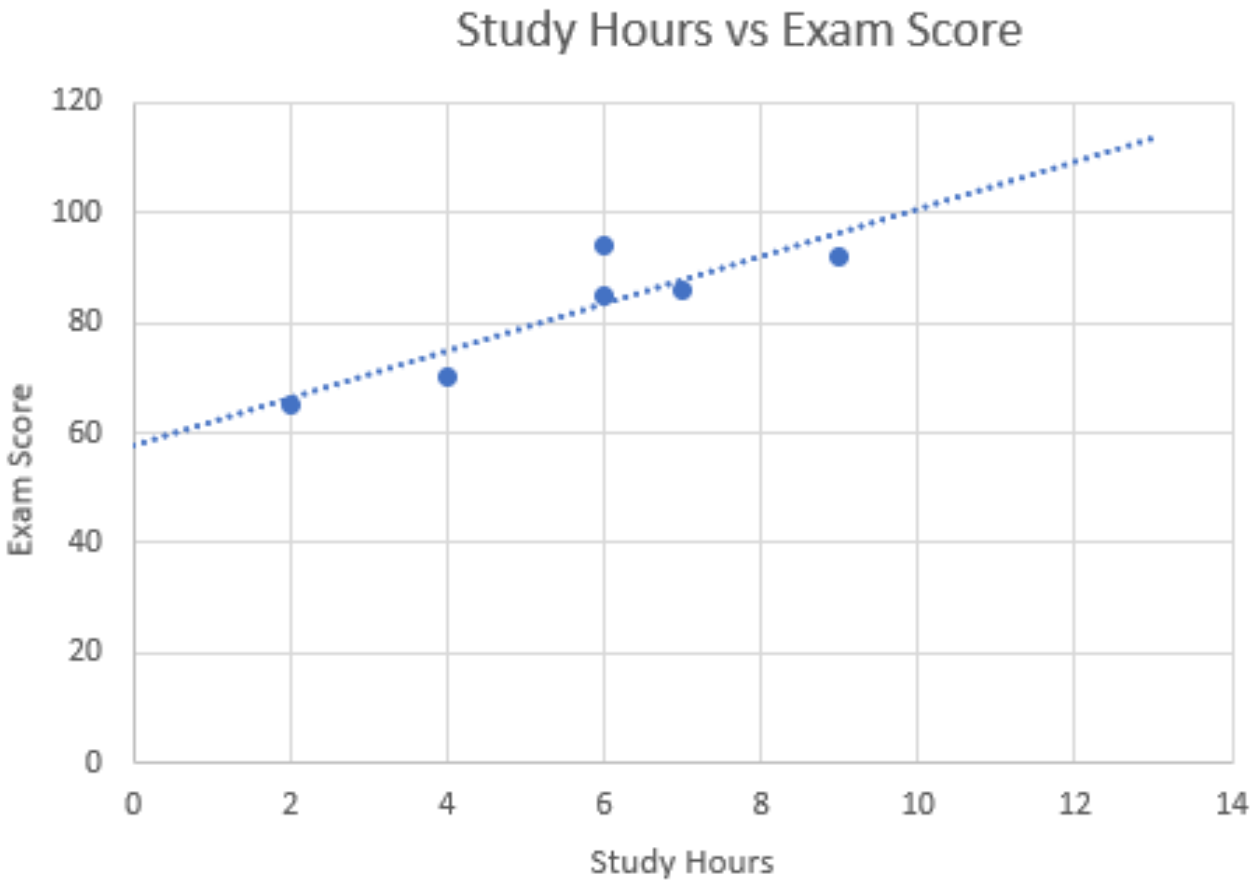


(a) How many MPG if you travel 80 miles per hour?

(b) How many MPG if you travel 30 miles per hour?

(c) How many miles per hour should you travel to get 30 MPG?

Use the given regression line to make predictions about the variables and tell whether you are using interpolation or extrapolation.

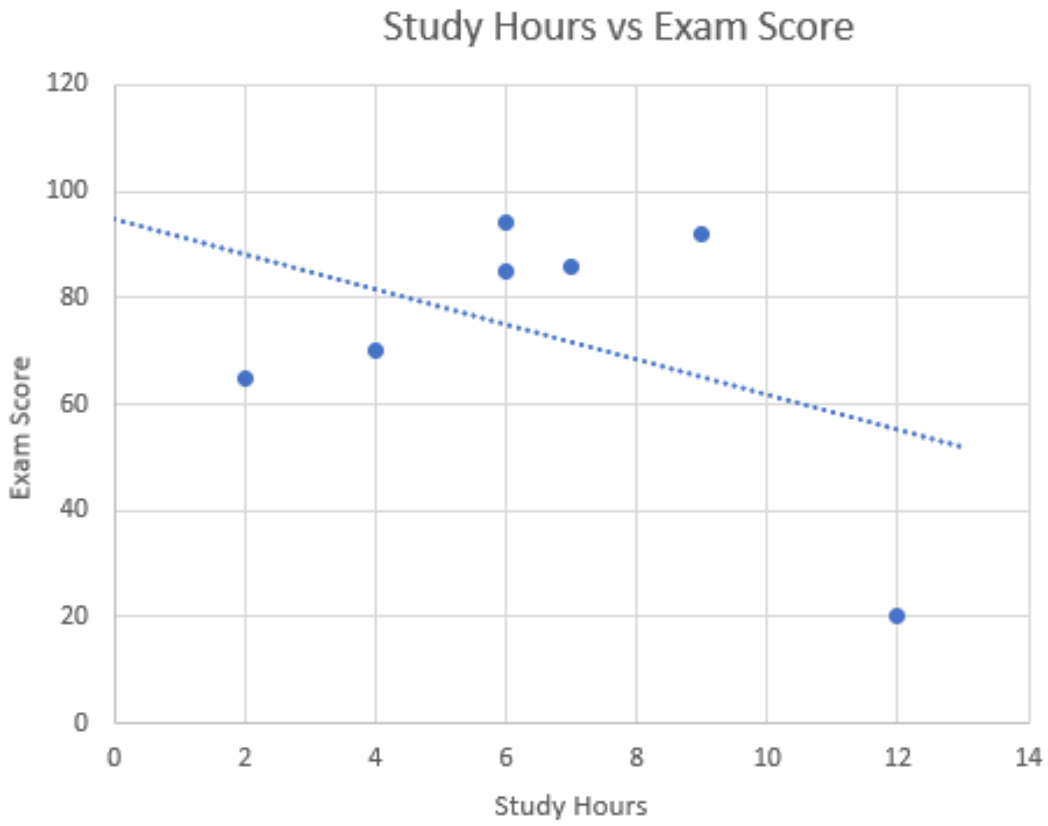


(a) If a person studied for 8 hours, what score would the regression line predict?

(b) How many hours would a person need to study for the regression line to predict a score of 100?

(c) If a person studied for 4 hours, what score would the regression line predict?

How would the graph change if we found out that a person studied for 12 hours and scored a 20 on the exam?

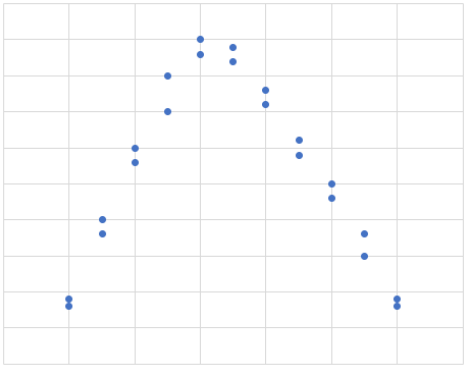


Things to keep in mind when using linear regression:

- Outliers can change things a lot
- Linear regression is only appropriate for linear relationships
- Interpolation is more likely to be accurate than extrapolation

CORRELATION DOES NOT IMPLY CAUSATION

SAMPLE EXAM QUESTIONS FROM CHAPTER 6



1. To the left is the scatterplot for data collected on two variables. Would a least squares linear regression equation be useful in describing the relationship between the variables? Why or why not?

(A) Yes because the data is approximately linear.
(B) Yes because the data has a strong relationship.
(C) No because the data does not have a linear relationship.
(D) No because the data does not have a strong relationship.

2. Suppose the children of an elementary school are weighed. What type of association would you expect between their ages and their weights?

(A) No association between the ages and weights
(B) A positive association between the ages and weights
(C) A negative association between the ages and weights

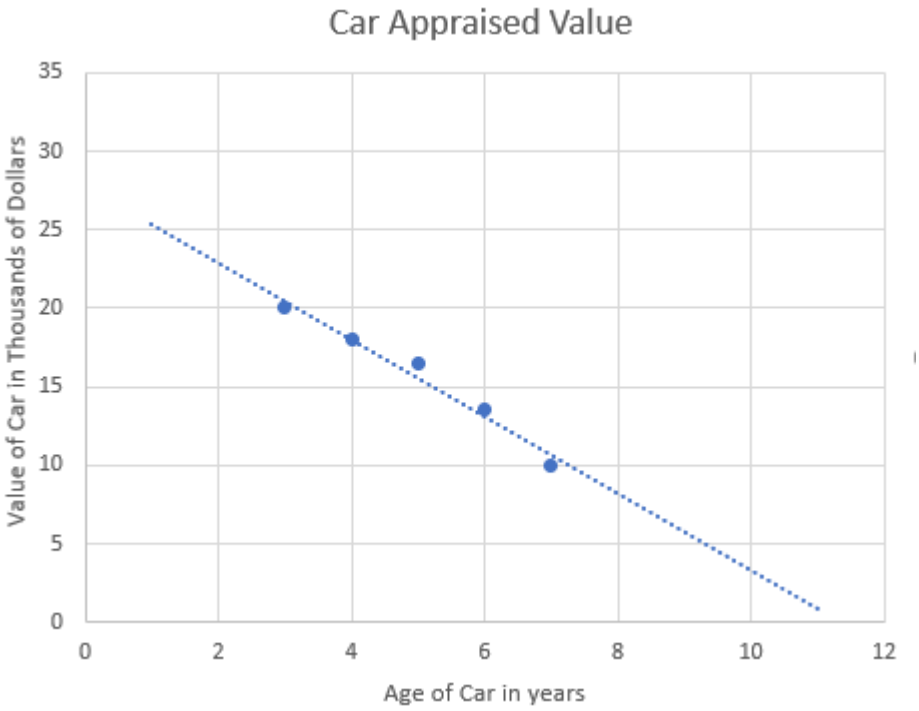
3. To choose advertising media, a marketing analyst studies the relationship between a consumer's age and the amount spent on restaurant dining. Which variable, consumer's age or dining expenditures, would be the explanatory variable for a scatterplot and least squares regression equation?

(A) Consumer's age
(B) Dining expenditures
(C) Neither of these
(D) Need more information

4. When the regression line is used to estimate values outside the data points this is called

(A) extrapolation (B) interpolation (C) extraordinary
(D) interdisciplinary (E) interception

5. A car owner has his car appraised annually to find the value of the car. The results of the appraisal are shown in the scatterplot below along with the regression line. Use the regression line to answer the following questions:



(a) What is the value of the car when it is 12 years old?

(b) What was the initial value of the car?

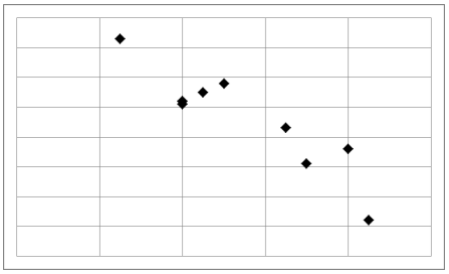
6. Match the scatter plots to the correlations below:

(a) Which scatterplot has a correlation of $r = 0.5$?

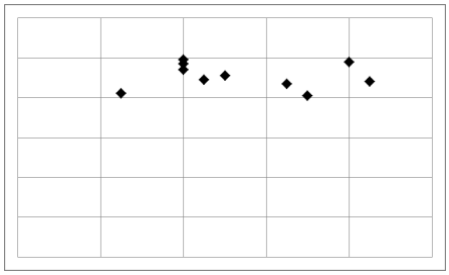
(b) Which scatterplot has a correlation of $r = -0.8$?

(c) Which scatterplot has a correlation of $r = 0$?

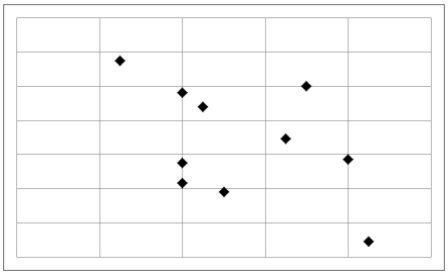
(i)



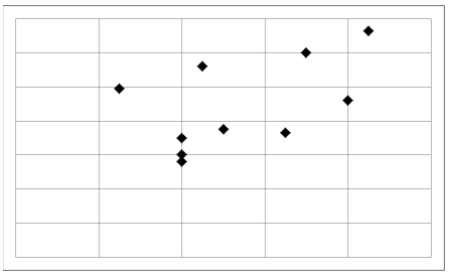
(ii)



(iii)



(iv)



(v)