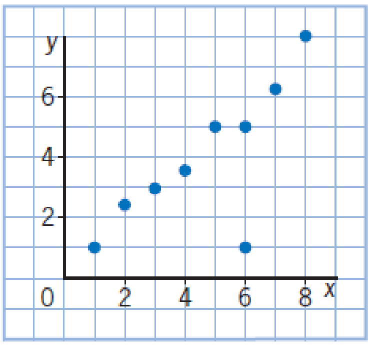
Solutions

1. Consider the correlation.



Which statement is most accurate?

A There is a strong positive correlation.

B There is a moderate positive correlation.

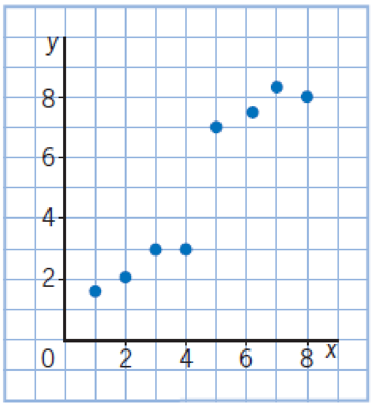
C There is a strong positive correlation with an outlier.

D There is a strong positive correlation with a possible hidden variable.

C

As x increases, y increases which indicates a positive correlation. The data would be close to a line of best fit, suggesting a strong correlation. The point (6,1) does not seem to fit the trend so it is an outlier.

2. Consider the correlation.



Which statement most accurate?

A There is a strong positive correlation.

B There is a moderate positive correlation.

C There is a strong positive correlation if the outlier is disregarded.

D There is a strong positive correlation with a hidden variable.

D

Again there is a strong positive correlation. There does appear to be two trends here though, so this would imply that there is a hidden variable.

3. What impact can a hidden variable have on a linear trend?

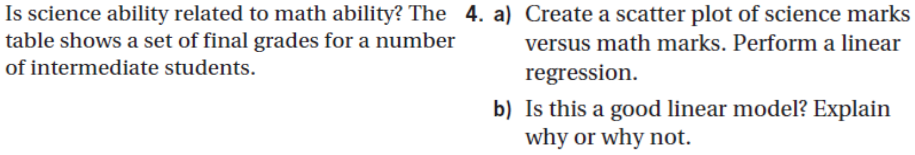
A It can hide or obscure the linearity.

B It can cause an irregularity in an otherwise linear trend.

C Both A and B are possible.

D It cannot have an impact on the trend.

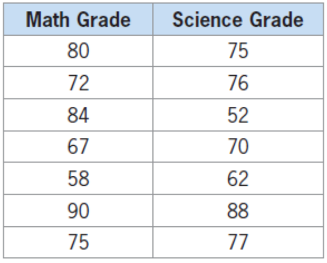
C



Is science ability related to math ability? The table shows a set of final grades for a number of intermediate students.

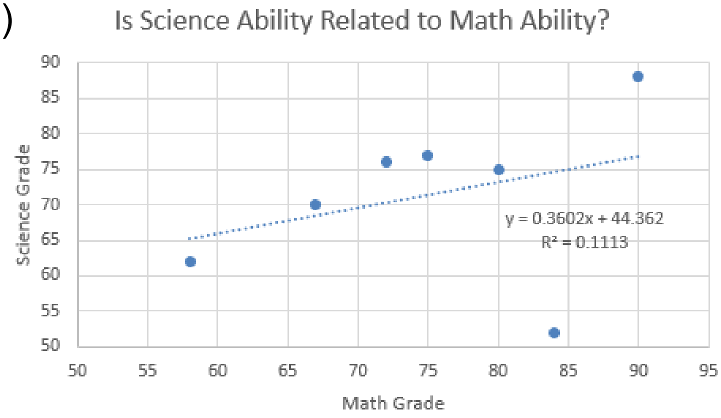
4. a) Create a scatter plot of science marks versus math marks. Perform a linear regression.

b) Is this a good linear model? Explain why or why not.

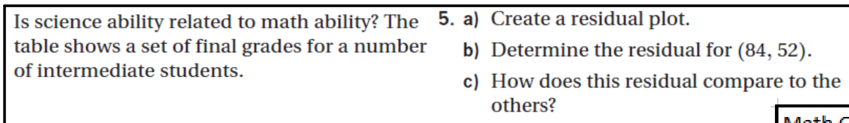


Math Grade	Science Grade
80	75
72	76
84	52
67	70
58	62
90	88
75	77

a)



b) This is not a good linear model because the r^2 value is only 0.1113 which would give an r-value of 0.334 which suggests a weak positive linear correlation. It would appear that the point (84,52) is an outlier.

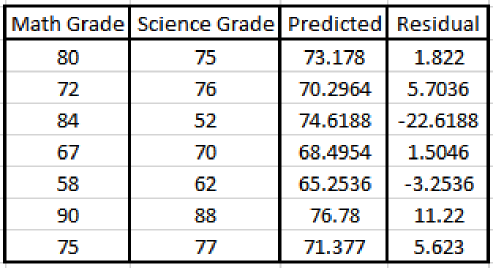


Is science ability related to math ability? The table shows a set of final grades for a number of intermediate students.

5. a) Create a residual plot.

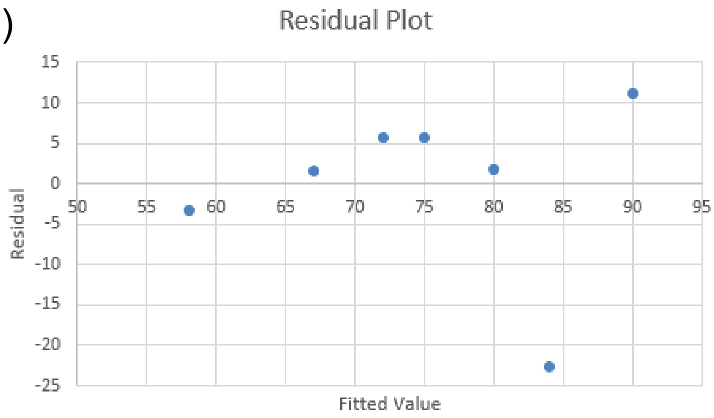
b) Determine the residual for (84, 52).

c) How does this residual compare to the others?



Math Grade	Science Grade	Predicted	Residual
80	75	73.178	1.822
72	76	70.2964	5.7036
84	52	74.6188	-22.6188
67	70	68.4954	1.5046
58	62	65.2536	-3.2536
90	88	76.78	11.22
75	77	71.377	5.623

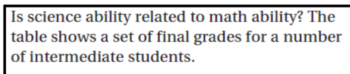
a)



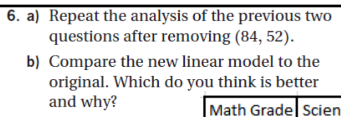
b) Use the equation from Q4 to create the PREDICTED column. Residual column = Science - Predicted.

The residual for (84,52) is -22.6188.

c) From the residual plot we can see that this residual point is much further from the residual line than the other residuals.



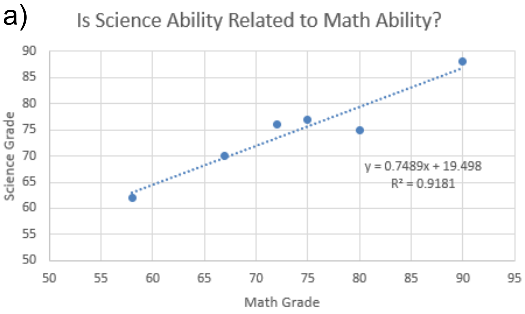
Is science ability related to math ability? The table shows a set of final grades for a number of intermediate students.

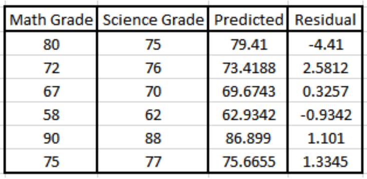


6. a) Repeat the analysis of the previous two questions after removing (84, 52).

b) Compare the new linear model to the original. Which do you think is better and why?

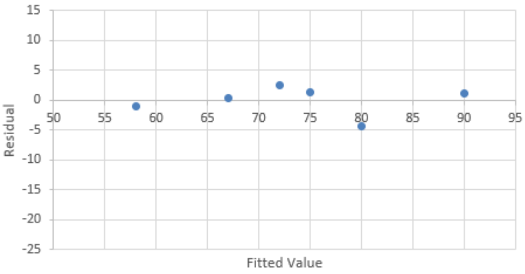
a)





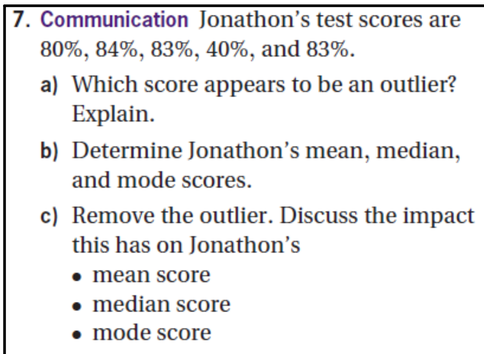
Math Grade	Science Grade	Predicted	Residual
80	75	79.41	-4.41
72	76	73.4188	2.5812
67	70	69.6743	0.3257
58	62	62.9342	-0.9342
90	88	86.899	1.101
75	77	75.6655	1.3345

This is a good linear model because the r² value is only 0.9181 which would give an r-value of 0.9582 which suggests a very strong positive linear correlation.



Using the same scales as before, we can see how the data is much closer to the residual line, which is backed up by the r-value being significantly closer to 1 than previously.

b) The original linear model did seem to be influenced by the presence of the outlying point (84,52). Their science score does seem lower than predicted. **We need to think about WHY this has happened.** Did the student do poorly in the exam, were they not able to study, did they not hand in some assignments?



7. **Communication** Jonathon's test scores are 80%, 84%, 83%, 40%, and 83%.

a) Which score appears to be an outlier? Explain.

b) Determine Jonathon's mean, median, and mode scores.

c) Remove the outlier. Discuss the impact this has on Jonathon's

- mean score
- median score
- mode score

a) The score of 40% appears to be an outlier as it is significantly lower than his other four marks.

b) Mean = (80 + 84 + 83 + 40 + 83) ÷ 5

= 370 ÷ 5

= 74%

Median = 83% Mode = 83%

c) Mean = (80 + 84 + 83 + 83) ÷ 4

= 330 ÷ 4

= 82.5%

Median = 83% Mode = 83%

Recall:

MEAN - sum of data divided by number of values

MEDIAN - middle value of the ordered data

MODE - value(s) that occur most often

The impact of removing the outlier is that the mean rose to 82.5% in line with the median and mode, which were both unchanged.

8. a) Construct a scatter plot of earnings versus time worked. Describe the correlation.

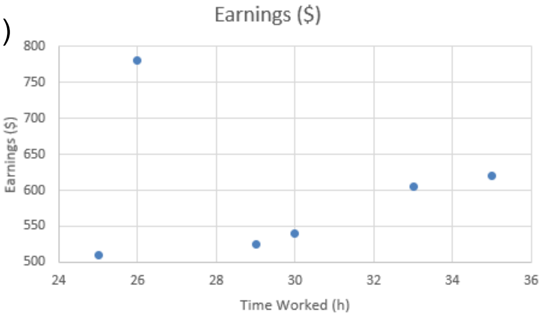
b) Perform a linear regression. Interpret the meaning of the equation of the line of best fit.

c) Is this a useful linear model? Explain.

The table shows the weekly earnings of a restaurant server, including tips.

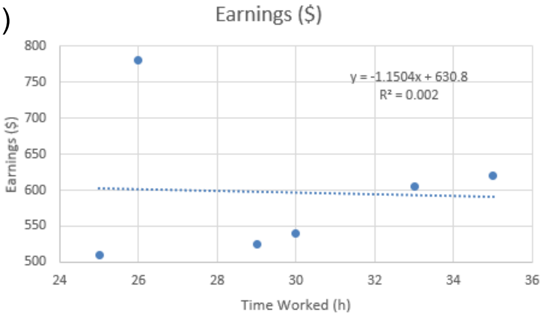
Time Worked (h)	Earnings ($)
30	540
25	510
33	605
26	780
35	620
29	525

a)



The data appears to have a strong, positive linear correlation with the exception of the point (25,510).

b)



The equation of the line of best fit is y = -1.1504x + 630.8, where y = earnings and x = hours worked.

This would imply that if you were to work 0 hours you would get paid $630.80 and that you have to give your employer back $1.15 for every hour that you work!

c) This is not a useful model as it really doesn't make any sense at all.

9. a) Construct a residual plot.

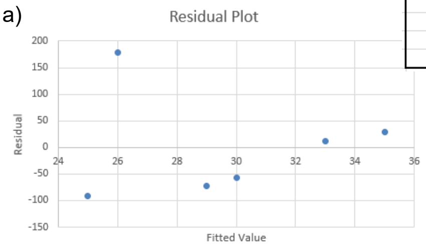
b) Identify an outlier in the data. What could account for the unusual data point?

c) Repeat the analysis of #8 with the outlier removed.

The table shows the weekly earnings of a restaurant server, including tips.

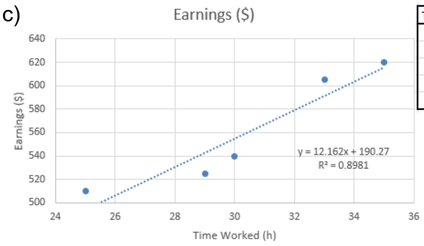
Time Worked (h)	Earnings ($)	Predicted	Residual
30	540	596.288	-56.288
25	510	602.04	-92.04
33	605	592.8368	12.1632
26	780	600.8896	179.1104
35	620	590.536	29.464
29	525	597.4384	-72.4384

a)



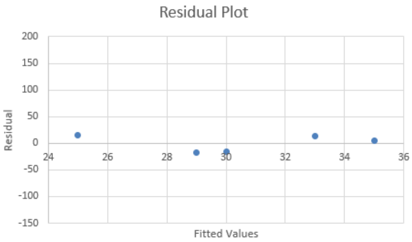
b) The residual for (26,780) is much further away from the residual line than the other data points. This could have been caused by some very generous tippers.

c)



Time Worked (h)	Earnings ($)	Predicted	Residual
30	540	555.13	-15.13
25	510	494.32	15.68
33	605	591.616	13.384
35	620	615.94	4.06
29	525	542.968	-17.968

The equation of the line of best fit is y = 12.162x + 190.27, where y = earnings and x = hours worked.



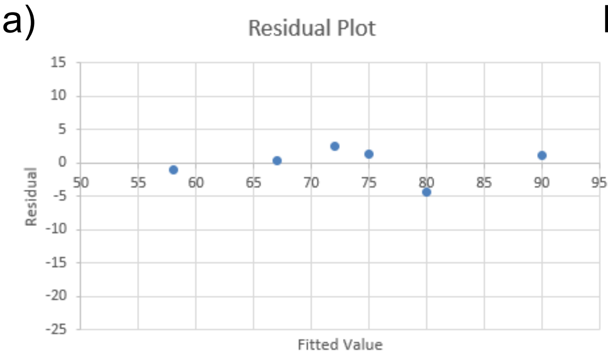
This would imply that if you were to work 0 hours you would get paid $190.27 which is more reasonable than before and that for every hour you work you will earn an extra $12.16.

The r^2 value is 0.8981 which gives an r-value of 0.948 which is a very strong, positive linear correlation.

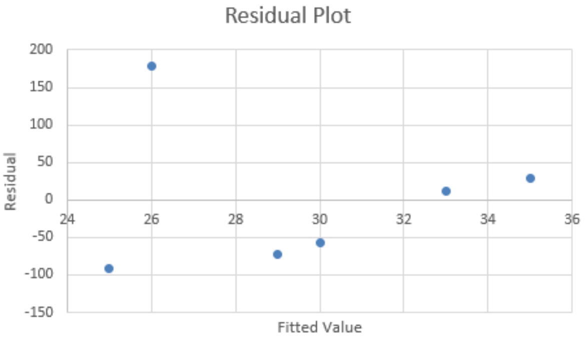
10. a) A set of two-variable data has no outliers. Draw a sketch that shows what its residual plot could look like.

b) Repeat part a) for a data set that has an outlier.

a)



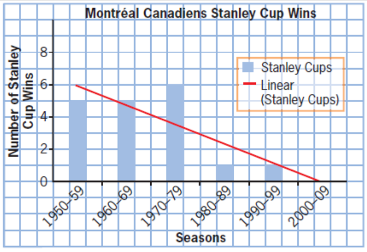
b)

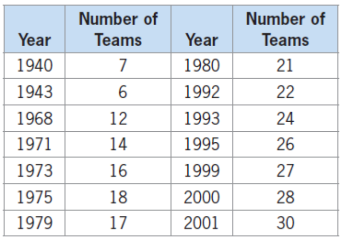


The graph illustrates the number of Stanley Cup wins by the Montréal Canadiens over time, measured in decades. For these questions, 1950 refers to the 1949–50 season, and so on.

11. a) What does this graph suggest about the performance trend of the Montréal Canadiens over the 50-year period?

b) Any team in the National Hockey League (NHL) is eligible to win the Stanley Cup. Consider the table, which shows how the number of teams in the NHL changed over time. Identify a possible hidden variable related to the correlation shown in the graph.

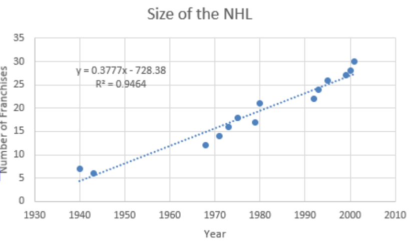




Year	Number of Teams	Year	Number of Teams
1940	7	1980	21
1943	6	1992	22
1968	12	1993	24
1971	14	1995	26
1973	16	1999	27
1975	18	2000	28
1979	17	2001	30

a) The graph suggests that Montreal have not been as successful as they previously were in 50s, 60s, and 70s.

b) The possible hidden variable could be the expansion of the NHL that started in 1967. **The more teams, the harder it is to win.** From 1942 to 1967 only 6 teams played in the NHL. In fact before this there were 7 (Brooklyn Americans stopped in 1942). In 1968, 6 teams were added. By 1975, a further 6 were added. By 1980, another 4 more (although 1 folded - Cleveland Barons if you're interested). The league has grown further still from 22 teams in 1992 to the current league size of 31 (32 for 2021-22 season).



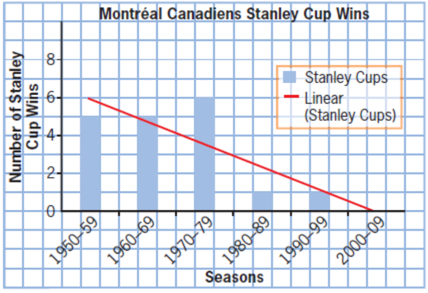
An r^2 value of 0.9464 gives an r-value of 0.973 implying a very strong, positive correlation between the number of teams and the year.

https://en.wikipedia.org/wiki/List_of_defunct_and_relocated_National_Hockey_League_te

The graph illustrates the number of Stanley Cup wins by the Montréal Canadiens over time, measured in decades. For these questions, 1950 refers to the 1949–50 season, and so on.

12. The Stanley Cup was not awarded in the 2004–05 season due to a labour disruption. Discuss how this could also represent a hidden variable in this study.

Despite the lockout in 2004-05 the NHL only lost one season in that decade (10%) and therefore would not invalidate the trends over the period of study.

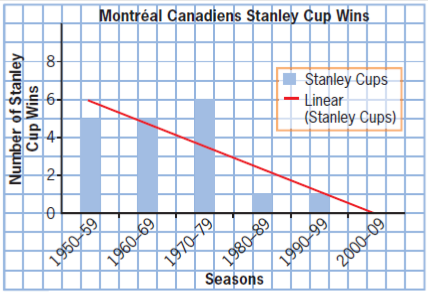


Year	Number of Teams	Year	Number of Teams
1940	7	1980	21
1943	6	1992	22
1968	12	1993	24
1971	14	1995	26
1973	16	1999	27
1975	18	2000	28
1979	17	2001	30

The graph illustrates the number of Stanley Cup wins by the Montréal Canadiens over time, measured in decades. For these questions, 1950 refers to the 1949–50 season, and so on.

13. Based on the given data, could you make an argument that the Montréal team of the 1970s was a better hockey team than those of the 1950s or 1960s? Explain.

Yes, you can make an argument that the team from the 70s were a "better" team than those of the 50s and 60s. Not only did they win more Stanley Cups, but they also were playing in a league with more teams in each of those seasons (12 to 18 instead of just 6).



Year	Number of Teams	Year	Number of Teams
1940	7	1980	21
1943	6	1992	22
1968	12	1993	24
1971	14	1995	26
1973	16	1999	27
1975	18	2000	28
1979	17	2001	30

14. The table shows the average annual attendance for a minor league baseball team.

a) Construct a scatter plot of this time series. Is baseball interest on the rise?

b) Perform a linear regression. Describe the strength of correlation.

c) What graphical evidence is there of a hidden variable?

d) In 2007, a large factory shut down due to the poor economy. How do you think this affected this correlational study?

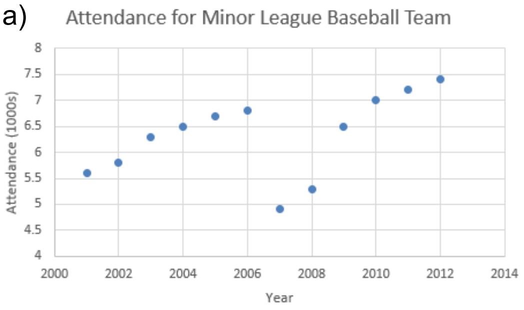
e) What do you think happened over the next few years following the plant closure?

f) Repeat the linear regression with the data points for 2007–2009 removed. Compare this linear model to the previous one.

g) Reflect on the interest in baseball now.

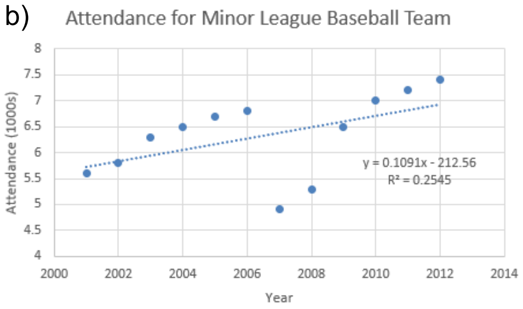
Year	Attendance (thousands)
2001	5.6
2002	5.8
2003	6.3
2004	6.5
2005	6.7
2006	6.8
2007	4.9
2008	5.3
2009	6.5
2010	7.0
2011	7.2
2012	7.4

a)



Generally, the trend is positive in that attendances are on the rise.

b)



The r^2 value is 0.2545 which gives an r-value of 0.504 leading to a moderate, positive, linear correlation.

c) There seem to be three "sections": 2001 to 2006, 2007 to 2009 and 2010 to 2012. This would lead you to believe there may be a hidden variable.

14. The table shows the average annual attendance for a minor league baseball team.

a) Construct a scatter plot of this time series. Is baseball interest on the rise?

b) Perform a linear regression. Describe the strength of correlation.

c) What graphical evidence is there of a hidden variable?

d) In 2007, a large factory shut down due to the poor economy. How do you think this affected this correlational study?

e) What do you think happened over the next few years following the plant closure?

f) Repeat the linear regression with the data points for 2007–2009 removed. Compare this linear model to the previous one.

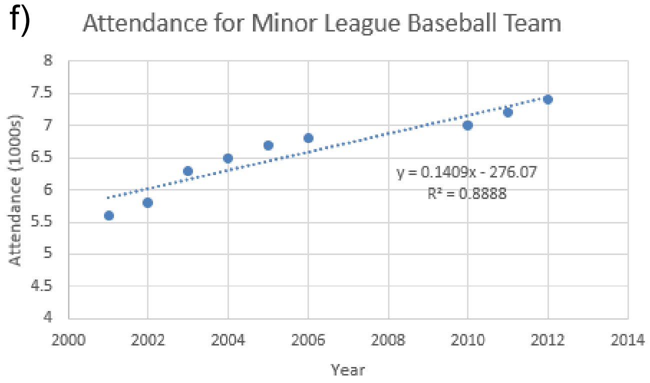
g) Reflect on the interest in baseball now.

Year	Attendance (thousands)
2001	5.6
2002	5.8
2003	6.3
2004	6.5
2005	6.7
2006	6.8
2007	4.9
2008	5.3
2009	6.5
2010	7.0
2011	7.2
2012	7.4

d) The closure of the large factory is likely to have caused the fragmentation of the trend.

e) The factory likely employed a lot of people in and around the town. With these people looking for work, they have less disposable income to spend on watching baseball, so the attendance dropped.

f)



The r^2 value is 0.8888 which gives an r-value of 0.943 leading to a very strong, positive, linear correlation.

g) With the years 2007 to 2009 removed we can see that there is a clear increase in baseball interest in this town.