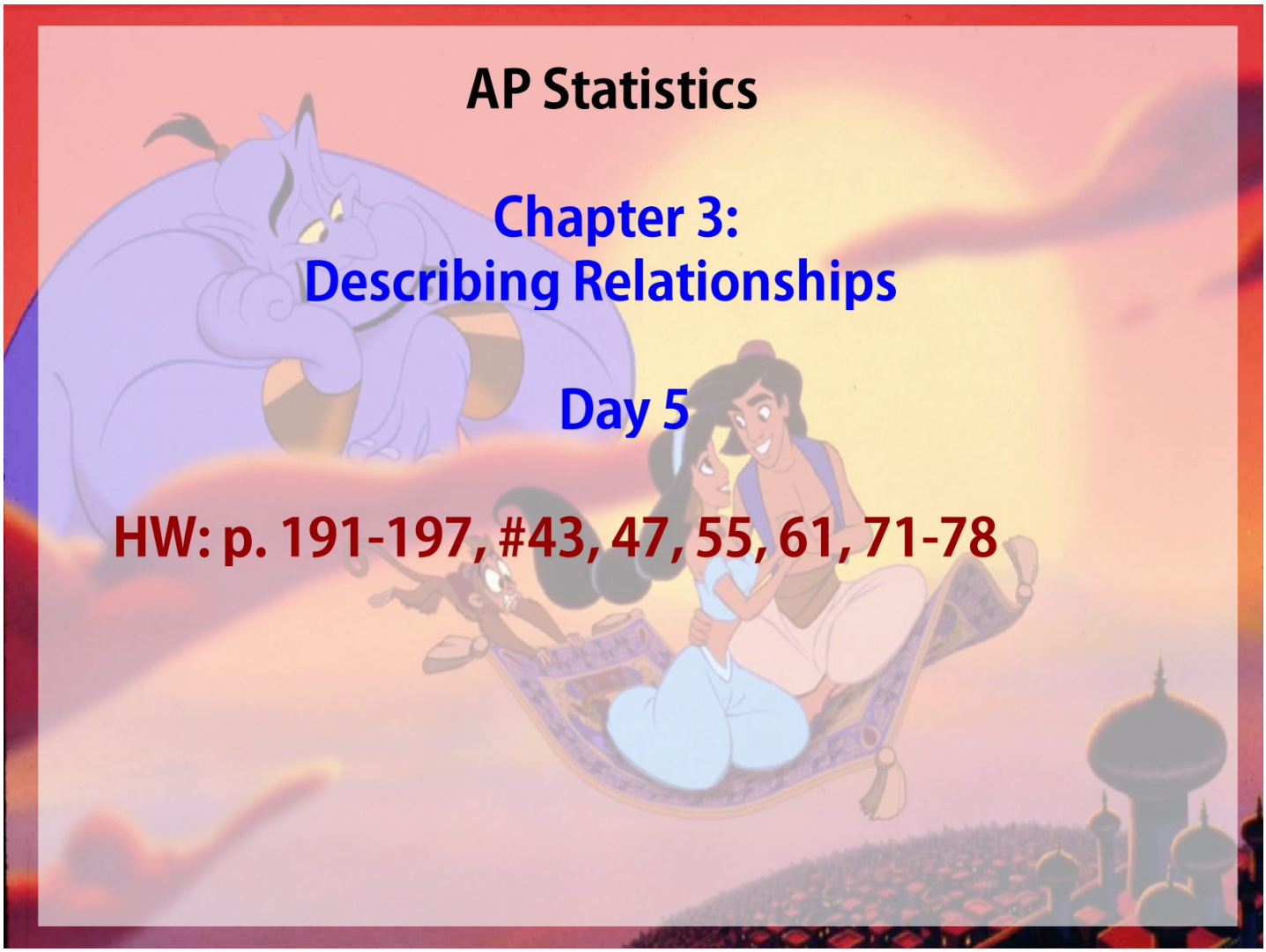


AP Statistics

Chapter 3:
Describing Relationships

Day 5

HW: p. 191-197, #43, 47, 55, 61, 71-78

p. 192-194, #49, 50, 53, 57, 63

49.

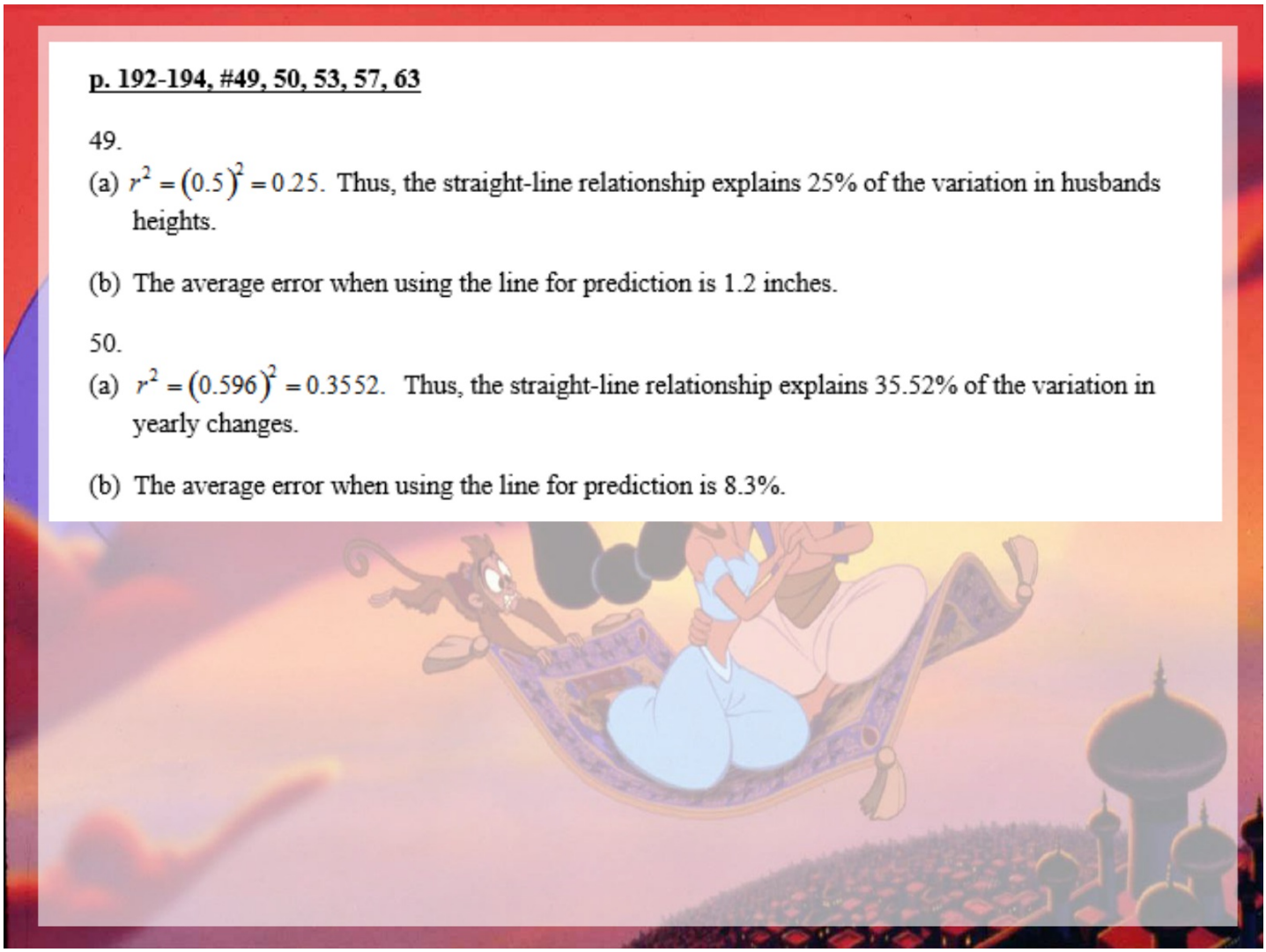
(a) $r^2 = (0.5)^2 = 0.25$. Thus, the straight-line relationship explains 25% of the variation in husbands heights.

(b) The average error when using the line for prediction is 1.2 inches.

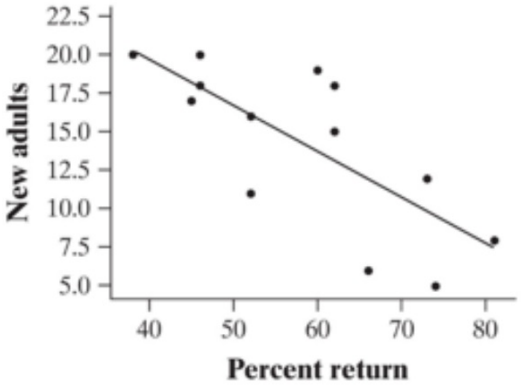
50.

(a) $r^2 = (0.596)^2 = 0.3552$. Thus, the straight-line relationship explains 35.52% of the variation in yearly changes.

(b) The average error when using the line for prediction is 8.3%.



53. (a) The scatterplot is



(b) The least squares regression line is $\hat{y} = 31.9 - 0.304x$.

(c) The slope tells us that as the percent of returning birds increases by one, we predict the number of new birds will decrease by -0.304. The y intercept provides a prediction that we will see 31.9 new adults in a new colony when the percent of returning birds is zero. This value is clearly outside the range of values studied for the 13 colonies of sparrowhawks and has no practical meaning in this situation.

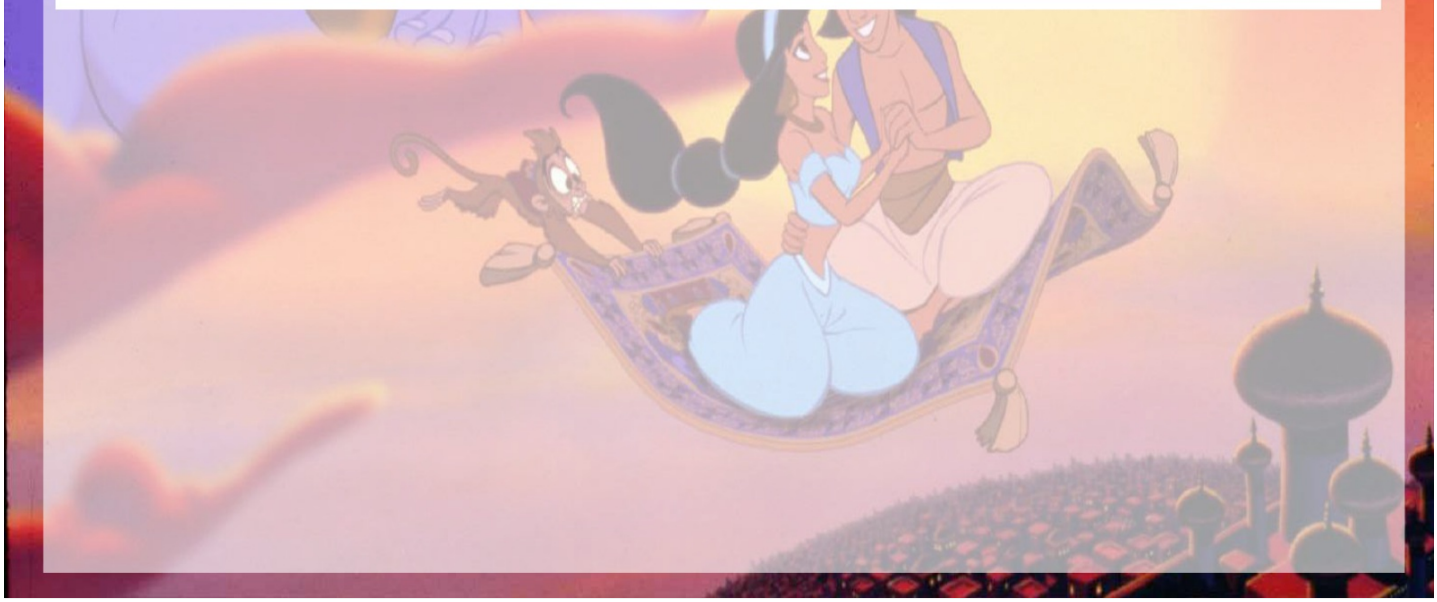
(d) The predicted value for the number of new adults is $31.9 - 0.304(60) = 13.66$ or about 14.

63.

(a) The regression line is $\hat{y} = 157.68 - 2.99x$. Following a season with 30 breeding pairs, we find $\hat{y} = 157.68 - 2.99(30) = 67.98$ so we predict that about 68% of males will return.

(b) This is given in the Minitab output as R-sq = 63.1%. The linear relationship explains 63.1% of the variation in the percent of returning males.

(c) Knowing that $r^2 = 0.631$, we find $r = -\sqrt{r^2} = -0.79$; the sign is negative because it has the same sign as the slope coefficient. (d) Since $s = 9.46$, the typical error when using the line to predict the return rate of males is about 9.46%.



Residual Plots

Residuals are the differences between the observed values of the *y* variable and the predicted value $\hat{y}$ from the regression model.

There is one residual for each point, which is calculated as:

dot – line

residual = observed value - predicted value

= y - $\hat{y}$

Residual Plots

To examine the graph of the residuals, plot the residuals on the vertical axis against the explanatory variable (x-value).

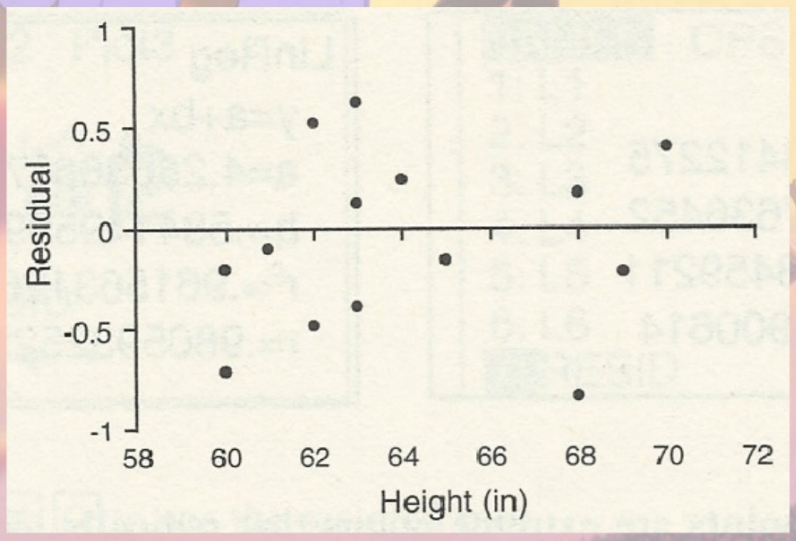
* no patterns

If the residual graph shows no curve or "U-shaped" pattern, a linear model is appropriate for the data.

However, if a linear model is not appropriate, the residual graph will have some sort of pattern or curved feature.

Residual Plots

The graph below is the residual plot for the height and shoe size data. There is no curved pattern in the plot, so a linear model is appropriate for these data.

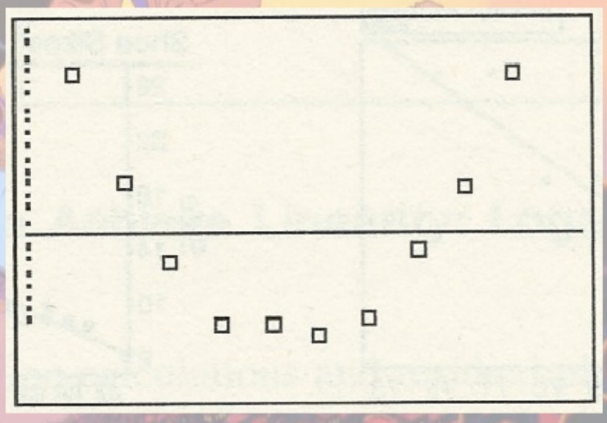


Residual Plots

The residual plot below is from a data set that is clearly not linear.

Even if the correlation coefficient is high, the residuals tell us that the pattern of the data is truly not linear.

This graph leads us to believe that the data are not well modeled by the line.



Outliers

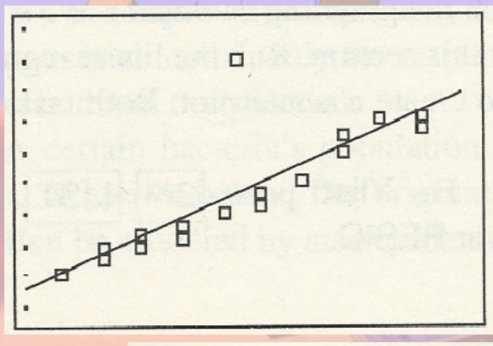
A residual plot can also reveal possible outliers.

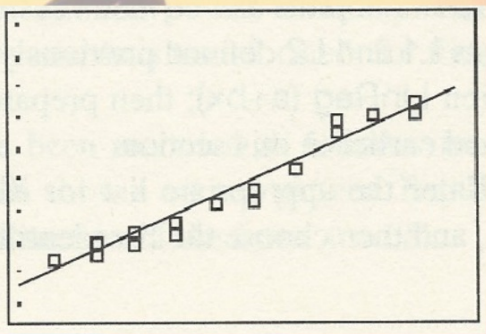
Outliers have large residuals since those points are far from the least-squares regression line.

An outlier usually does not have a great effect on the slope of a least-squares regression line, but does have and influence on the correlation.

Outliers

The scatterplots below represent a data set with an outlier and with the outlier removed. The LSRL is shown for both sets. Note that the slope did not significantly change when removing the outlier, but the correlation (r) increased substantially.





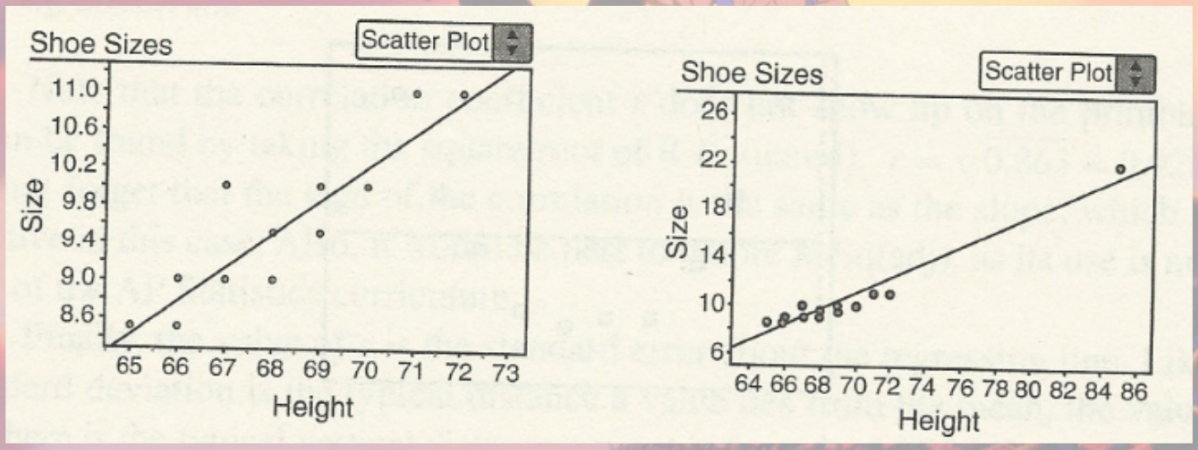
LinReg
y=a+bx
a=4.513412275
b=.5837636452
r^2=.6748459211
r=.8214900614

LinReg
y=a+bx
a=4.256366874
b=.5841495993
r^2=.9615631264
r=.9805932523

Influential Points

Influential points are extreme points that radically affect the slope of a LSRL.

The graph on the left shows data for 15 randomly selected men. The graph on the right shows the same men plus Shaquille O'Neal, an 85-inch tall basketball player with a shoe size of 22.



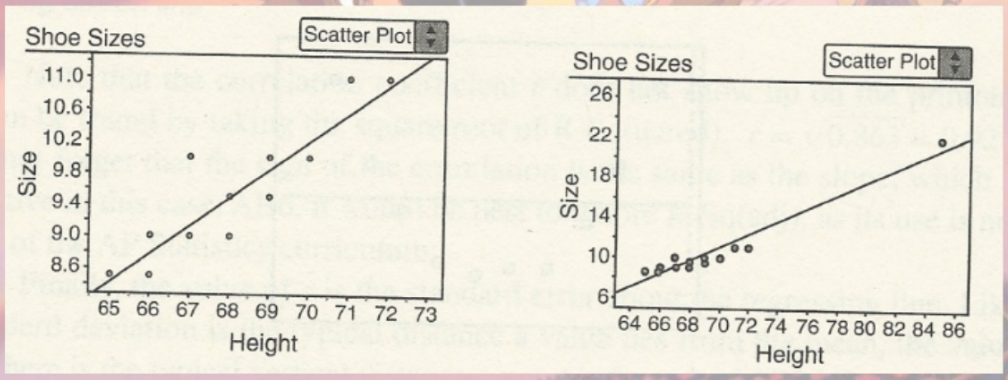
Influential Points

If O'Neal's point were removed from the data, the slope would drop from 0.676 to 0.357.

Therefore, O'Neal is an influential observation.

Note also that the removal of O'Neal would reduce r^2 from 0.94 to 0.79.

This is also a characteristic of extreme points that seem to be in the general pattern of the data - they artificially strengthen correlation and r^2.



AP Statistics
Lesson 5 U3C3 Practice

Name: _____
Period: _____ Date: _____

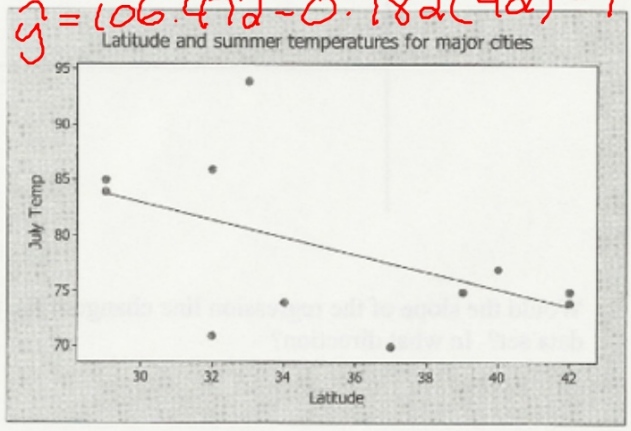
(+) residual = under estimated
(−) residual = over estimated

Linear Regression Models

1. The table and scatterplot below describes the relationship between latitude and average July temperature in the twelve largest U.S. cities.

$y - \hat{y}$ $\hat{y} = 106.472 - 0.782x$ $\hat{y} = 106.472 - 0.782(42) = 73.628$

City	Latitude (x)	July Temp (y)
New York	40	77
Los Angeles	34	74
Chicago	42	75
Houston	29	84
Philadelphia	40	77
Phoenix	33	94
San Diego	32	71
San Antonio	29	85
Dallas	32	86
San Jose	37	70
Detroit	42	74
Indianapolis	39	75

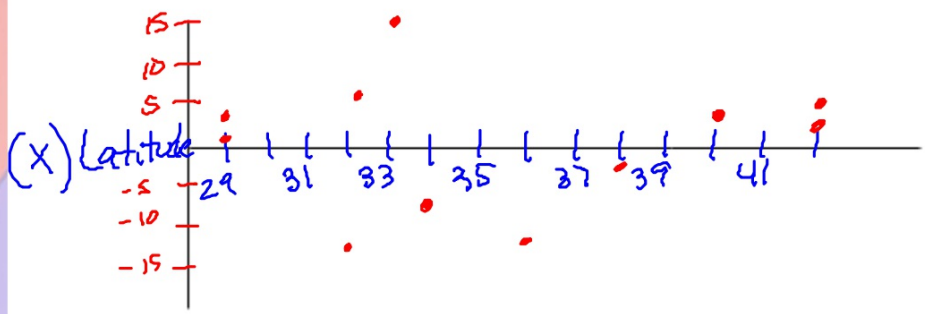


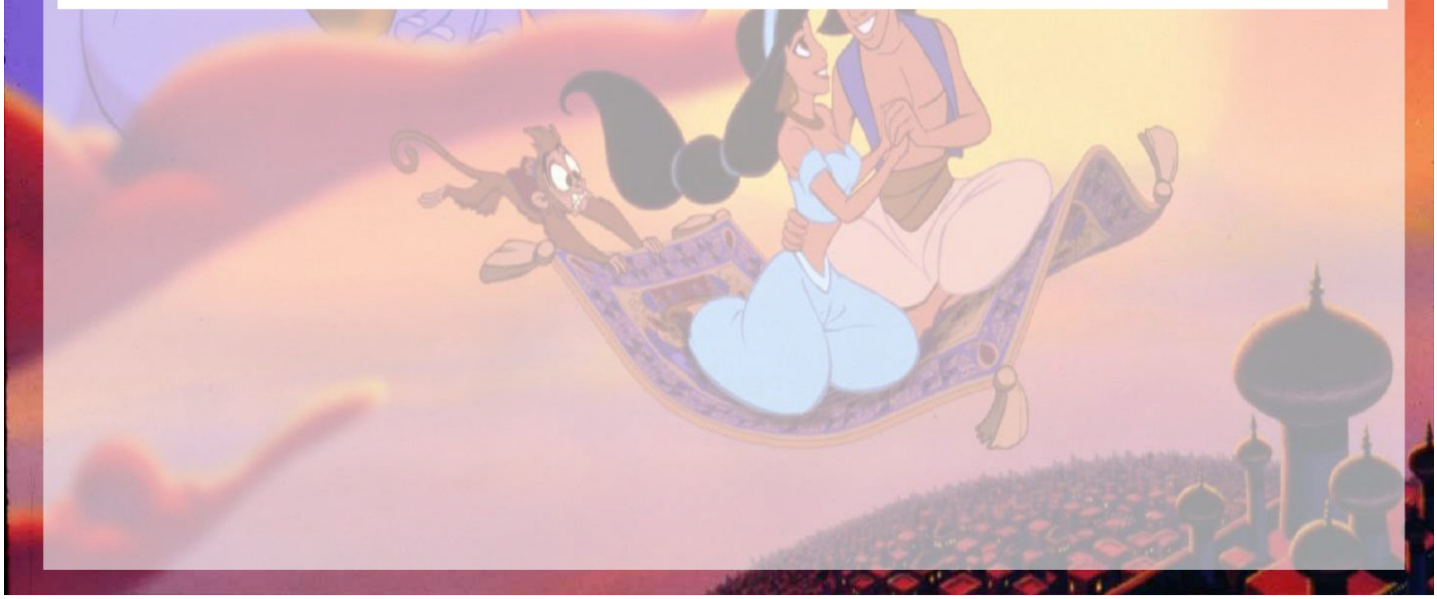
A) Find the value of the residual for Detroit. Show your work. Interpret the value of the residual in the context of the problem.

$74 - 73.628 = .372$

The average temp in July for Detroit was .372 degrees greater than the prediction line.

B) Use the scatterplot or your calculator to make a rough sketch of the residual plot for these data.





C) Phoenix has a very large positive residual. How would the slope change of the regression line if it were removed from the data set?

$\hat{y} = 106.472 - .782x$

$\hat{y} = 99.446 - .621x$

The slope would increase from $-.782$ to $-.621$.

D) Is the given least-squares regression line a good model for using latitude to predict average July temperature of U.S. cities? Support your answer with appropriate evidence from you answers above.

Yes. Our residual plot shows no pattern (no curve, no u-shape)