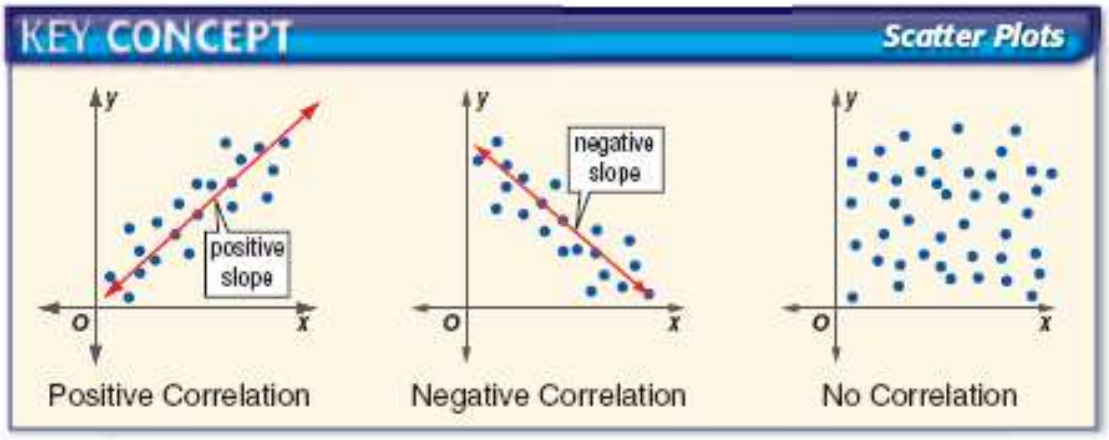
Scatter Plots and Lines of Fit

Scatter Plots and Correlation:



Line of Fit:

When you find a line that closely approximates a set of data, you are finding a line of fit for the data.

Prediction Equation:

An equation of the line of fit is called a prediction equation because it can be used to predict one of the variables given the other variable.

Formulas for Calculating Regression Lines:

$y = mx + b$

$$m = \frac{(N \cdot xy_{sum}) - (x_{sum} \cdot y_{sum})}{(N \cdot x_{sum}^2) - (x_{sum} \cdot x_{sum})}$$

$$b = \frac{(x_{sum}^2 \cdot y_{sum}) - (x_{sum} \cdot xy_{sum})}{(N \cdot x_{sum}^2) - (x_{sum} \cdot x_{sum})}$$

Example 1:

The table shows the temperature of the atmosphere at various altitudes.

Altitude (ft)	0	1000	2000	3000	4000	5000
Temp (°C)	15.0	13.0	11.0	9.1	7.1	?

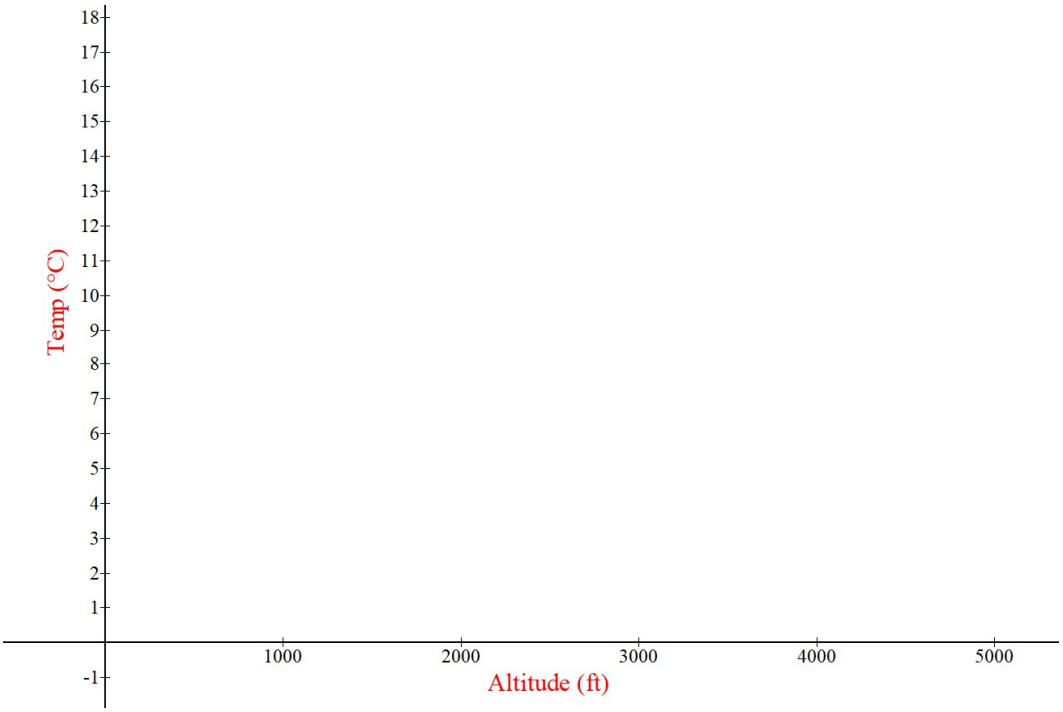
Source: NASA

a) Draw a scatter plot.
b) Sketch a line of fit.
c) Describe the correlation and what it means.
d) Write a prediction equation.
e) Use your prediction equation to predict the missing value.
f) When will the temperature be 12°C?

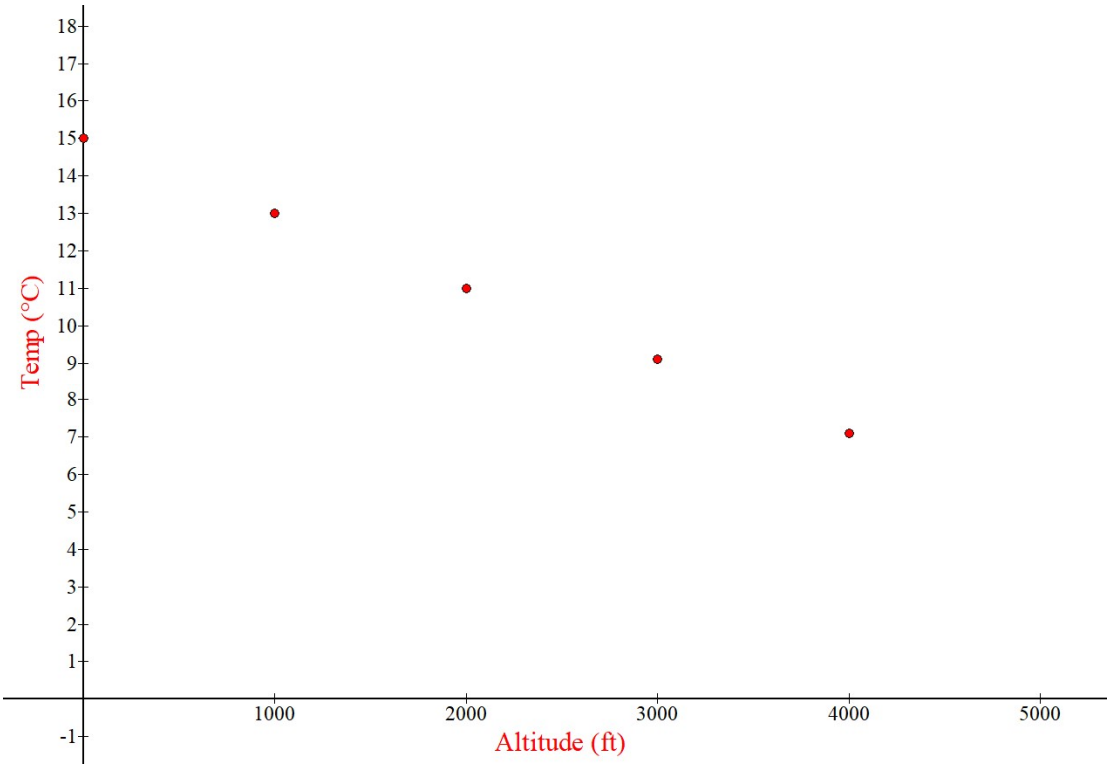
a) Draw a scatter plot.

A scatter plot is just a graph with all of these points graphed. The first thing you want to decide is which value will be on which axis. I am going to choose altitude to be on the horizontal axis and temperature to be on the vertical axis. That would be the typical way to do this since the temperature depends on the altitude and is therefore the dependent variable.

We also need to decide on a scale. Since altitude increases in increments of 1000 feet, it would make sense to have our graph increase in increments of 1000 and we need to make sure that we go to at least 4000 on the horizontal axis. The temperature decreases by about 2 units, so it would make sense to have a scale the increases by 1.

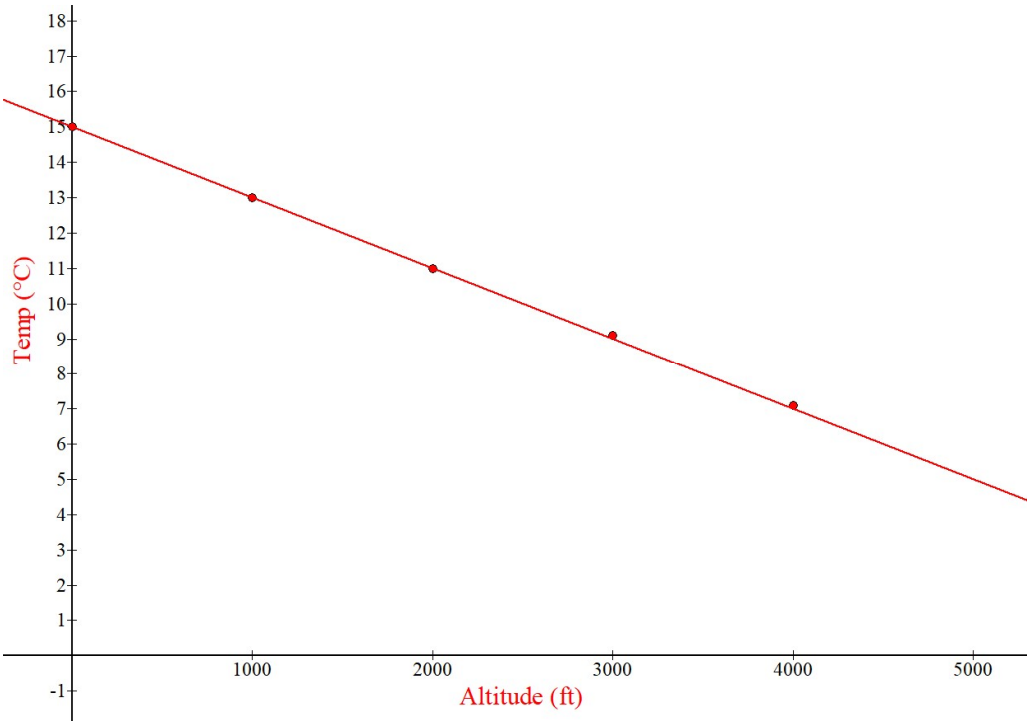


Now we just need to plot the points given in the table and we will have completed part (a).



b) Sketch a line of fit.

When we are sketching a line of fit we just want to draw a line that we think looks like it approximates the shape of the data. Once you've sketched a line that looks pretty close to the data you are done with part (b).



c) Describe the correlation and what it means.

We can see that the correlation is negative because the data decreases as we move from right to left. If we think about what that means, we would say that as the altitude increases the temperature decreases.

d) Write a prediction equation.

In order to write a prediction equation, we write a table with 4 columns and label them as follows.

x	y	xy	x^2

Then we start filling in the table. We put the altitude in the x column because we said that was the independent variable. We will fill temperature in the y column.

x	y	xy	x^2
0	15.0		
1000	13.0		
2000	11.0		
3000	9.1		
4000	7.1		

To fill in the xy column, we just multiply the x and y values in each row.

x	y	xy	x^2
0	15.0	0	
1000	13.0	13,000	
2000	11.0	22,000	
3000	9.1	27,300	
4000	7.1	28,400	

To fill in the x^2 column, we multiply the x value in each row by itself.

x	y	xy	x^2
0	15.0	0	0
1000	13.0	13,000	1,000,000
2000	11.0	22,000	4,000,000
3000	9.1	27,300	9,000,000
4000	7.1	28,400	16,000,000

The last thing we need to do is count how many rows of data we have and total each column.

x	y	xy	x^2
0	15.0	0	0
1000	13.0	13,000	1,000,000
2000	11.0	22,000	4,000,000
3000	9.1	27,300	9,000,000
4000	7.1	28,400	16,000,000
10,000	55.2	90,700	30,000,000

There are five rows of data, so $N = 5$

The totals from the last row are what we will use in our formulas:

$$m = \frac{(N \cdot xy_{sum}) - (x_{sum} \cdot y_{sum})}{(N \cdot x_{sum}^2) - (x_{sum} \cdot x_{sum})}$$

$$m = \frac{(5 \cdot 90{,}700) - (10{,}000 \cdot 55.2)}{(5 \cdot 30{,}000{,}000) - (10{,}000 \cdot 10{,}000)}$$

$$m = \frac{(453{,}500) - (552{,}000)}{(150{,}000{,}000) - (1{,}000{,}000)}$$

$$m = \frac{-98{,}500}{50{,}000{,}000}$$

$$m = -0.00197$$

$$b = \frac{(x_{sum}^2 \cdot y_{sum}) - (x_{sum} \cdot xy_{sum})}{(N \cdot x_{sum}^2) - (x_{sum} \cdot x_{sum})}$$

$$b = \frac{(30{,}000{,}000 \cdot 55.2) - (10{,}000 \cdot 90{,}700)}{(5 \cdot 30{,}000{,}000) - (10{,}000 \cdot 10{,}000)}$$

$$b = \frac{(1{,}656{,}000{,}000) - (907{,}000{,}000)}{(150{,}000{,}000) - (1{,}000{,}000)}$$

$$b = \frac{749{,}000{,}000}{50{,}000{,}000}$$

$$b = 14.98$$

$y = -0.00197x + 14.98$ **This is our prediction equation.

e) Use your prediction equation to predict the missing value.

We will use the prediction equation that we wrote in part (d).

$y = -0.00197x + 14.98$

The missing value is for an altitude of 5000 feet.

Since we chose altitude to be the x-values, we will plug 5000 feet into x in our prediction equation.

$y = -0.00197x + 14.98$

$y = -0.00197(5000) + 14.98$

$y = -9.85 + 14.98$

$y = 5.13$

Now we need to apply units to the value. Remember that y-values are temperature.

At 5000 feet the temperature would be 5.13°C.

f) When will the temperature be 12°C?

We will use the prediction equation that we wrote in part (d).

$y = -0.00197x + 14.98$

Since we chose temperature to be the y-values, we will plug 12 feet into y in our prediction equation and solve for x.

$y = -0.00197x + 14.98$

$12 = -0.00197x + 14.98$

$-14.98 \qquad -14.98$

$-2.98 = -0.00197x$

$1512.6903 \ldots = x$

The temperature would be 12°C at an altitude of about 1,512 feet.

Example 2:

The table shows the percentage of U.S. households with televisions that also had cable service.

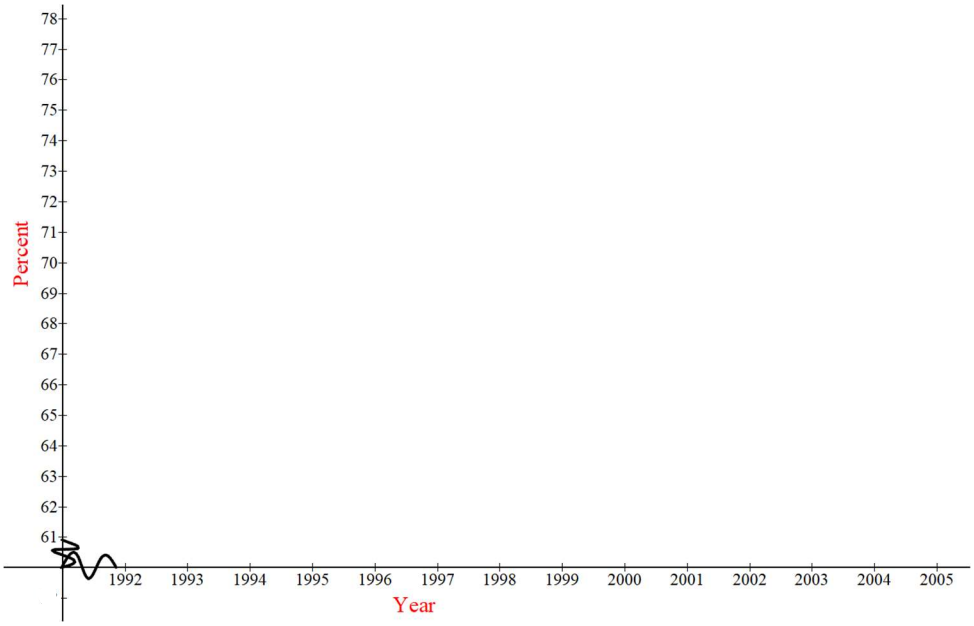
Year	1995	1997	1999	2001	2003	2015
Percent	65.7	67.3	68.0	69.2	68.0	?

a) Draw a scatter plot.
b) Sketch a line of fit.
c) Describe the correlation and what it means.
d) Write a prediction equation.
e) Use your prediction equation to predict the missing value.
f) When will 75% of homes have televisions and cable service?

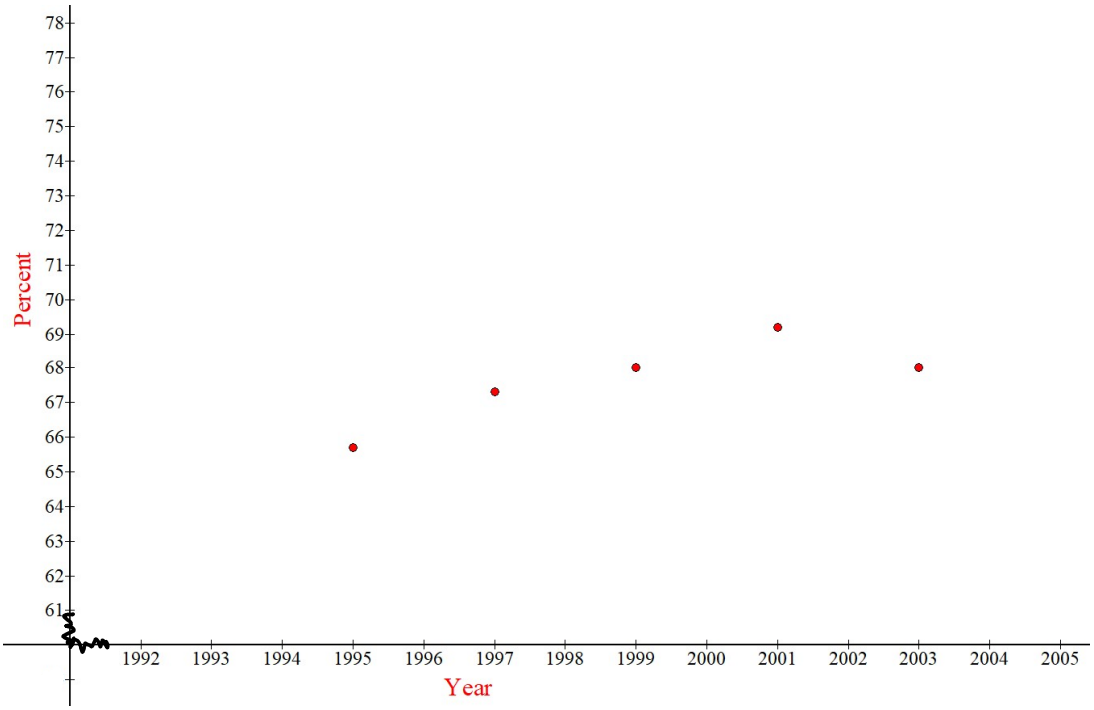
a) Draw a scatter plot.

I am going to choose year to be on the horizontal axis and percent to be on the vertical axis. That would be the typical way to do this since the percent depends on the year and is therefore the dependent variable.

We also need to decide on a scale. Since year increases in increments of 2, it would make sense to have our graph increase in increments of 1 or 2, but we will want to start our graph at 1990 instead of 0. To do that we compact a chunk of our graph, shown by the scrunched line at the beginning of the graph. The percent scale would make sense to be 1 but not starting until 60.

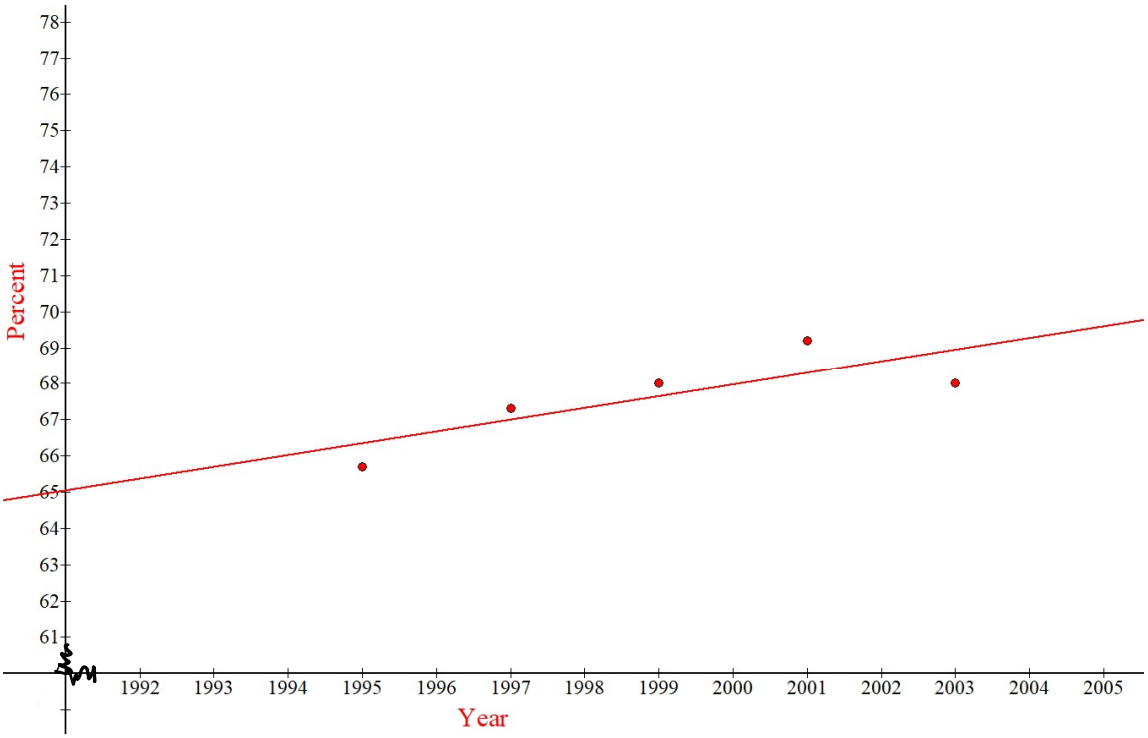


Now we just need to plot the points given in the table and we will have completed part (a).



b) Sketch a line of fit.

When we are sketching a line of fit we just want to draw a line that we think looks like it approximates the shape of the data. Once you've sketched a line that looks pretty close to the data you are done with part (b).



c) Describe the correlation and what it means.

We can see that the correlation is generally positive because the data increases as we move from right to left. If we think about what that means, we would say that as the years have increased the percent of homes with cable service has increased.

d) Write a prediction equation.

In order to write a prediction equation, we write a table with 4 columns and label them as follows.

x	y	xy	x^2

Then we start filling in the table. We put the year in the x column because we said that was the independent variable. We will fill percent in the y column.

x	y	xy	x^2
1995	65.7		
1997	67.3		
1999	68.0		
2001	69.2		
2003	68.0		

To fill in the xy column, we just multiply the x and y values in each row.

x	y	xy	x^2
1995	65.7	131,071.5	
1997	67.3	134,398.1	
1999	68.0	135,932	
2001	69.2	138,469.2	
2003	68.0	136,204	

To fill in the x^2 column, we multiply the x value in each row by itself.

x	y	xy	x^2
1995	65.7	131,071.5	3,980,025
1997	67.3	134,398.1	3,988,009
1999	68.0	135,932	3,996,001
2001	69.2	138,469.2	4,004,001
2003	68.0	136,204	4,012,009

The last thing we need to do is count how many rows of data we have and total each column.

x	y	xy	x^2
1995	65.7	131,071.5	3,980,025
1997	67.3	134,398.1	3,988,009
1999	68.0	135,932	3,996,001
2001	69.2	138,469.2	4,004,001
2003	68.0	136,204	4,012,009
9,995	338.2	676,074.8	19,980,045

There are five rows of data, so $N = 5$

The totals from the last row are what we will use in our formulas:

$$m = \frac{(N \cdot xy_{sum}) - (x_{sum} \cdot y_{sum})}{(N \cdot x_{sum}^2) - (x_{sum} \cdot x_{sum})}$$

$$m = \frac{(5 \cdot 676{,}074.8) - (9{,}995 \cdot 338.2)}{(5 \cdot 19{,}980{,}045) - (9{,}995 \cdot 9{,}995)}$$

$$m = \frac{(3{,}380{,}374) - (3{,}380{,}309)}{(99{,}900{,}225) - (99{,}900{,}025)}$$

$$m = \frac{65}{200}$$

$$m = 0.325$$

$$b = \frac{(x_{sum}^2 \cdot y_{sum}) - (x_{sum} \cdot xy_{sum})}{(N \cdot x_{sum}^2) - (x_{sum} \cdot x_{sum})}$$

$$b = \frac{(19{,}980{,}045 \cdot 338.2) - (9{,}995 \cdot 676{,}074.8)}{(5 \cdot 19{,}980{,}045) - (9{,}995 \cdot 9{,}995)}$$

$$b = \frac{(6{,}757{,}251{,}219) - (6{,}757{,}367{,}626)}{(99{,}900{,}225) - (99{,}900{,}025)}$$

$$b = \frac{-116{,}407}{200}$$

$$b = -582.035$$

$y = 0.325x - 582.035$ **This is our prediction equation.

e) Use your prediction equation to predict the missing value.

We will use the prediction equation that we wrote in part (d).

$y = 0.325x - 582.035$

The missing value is for the year 2015.

Since we chose year to be the x-values, we will plug 2015 into x in our prediction equation.

$y = 0.325x - 582.035$

$y = 0.325(2015) - 582.035$

$y = 654.875 - 582.035$

$y = 72.84$

Now we need to apply units to the value. Remember that y-values are percent

In the year 2015 72.8% of households will have cable.

g) When will 75% of homes have televisions and cable service?

We will use the prediction equation that we wrote in part (d).

$y = 0.325x - 582.035$

Since we chose percent to be the y-values, we will plug 75 into y in our prediction equation and solve for x.

$y = 0.325x - 582.035$

$75 = 0.325x - 582.035$

$+582.035 \quad + 582.035$

$657.035 = 0.325x$

$2021.6561 \ldots = x$

75% of households will have cable in the year 2021.