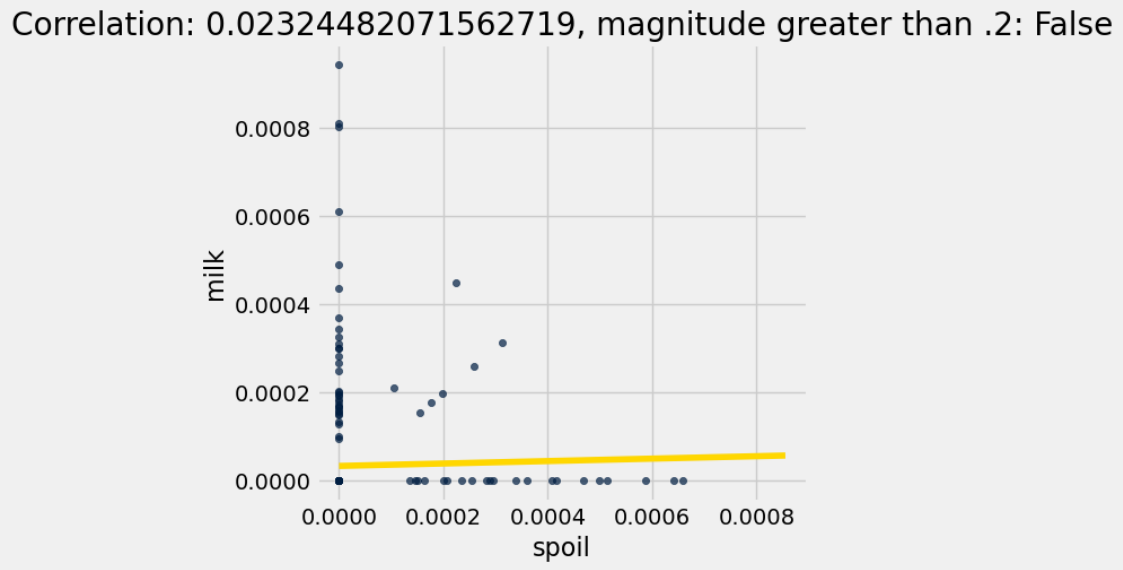

Question 1.2.2

Choose two *different* words in the dataset with a magnitude (absolute value) of correlation higher than 0.2 and plot a scatter plot with a line of best fit for them. Please do not pick "outer" and "space" or "san" and "francisco". The code to plot the scatter plot and line of best fit is given for you, you just need to calculate the correct values to `r`, `slope` and `intercept`.

Hint 1: It's easier to think of words with a positive correlation, i.e. words that are often mentioned together. Try to think of common phrases or idioms.

Hint 2: Refer to Section 15.2 of the textbook for the formulas. For additional past examples of regression, see Homework 9.

In [156]:
```
word_x = "spoil"
word_y = "milk"

# These arrays should make your code cleaner!
arr_x = movies.column(word_x)
arr_y = movies.column(word_y)

x_su = (arr_x - np.mean(arr_x)) / np.std(arr_x)
y_su = (arr_y - np.mean(arr_y)) / np.std(arr_y)

r = np.mean(x_su * y_su)

slope = r * np.std(arr_y) / np.std(arr_x)
intercept = np.mean(arr_y) - slope * np.mean(arr_x)

# DON'T CHANGE THESE LINES OF CODE
movies.scatter(word_x, word_y)
max_x = max(movies.column(word_x))
plots.title(f"Correlation: {r}, magnitude greater than .2: {abs(r) >= 0.2}")
plots.plot([0, max_x * 1.3], [intercept, intercept + slope * (max_x*1.3)], color='gold');
```

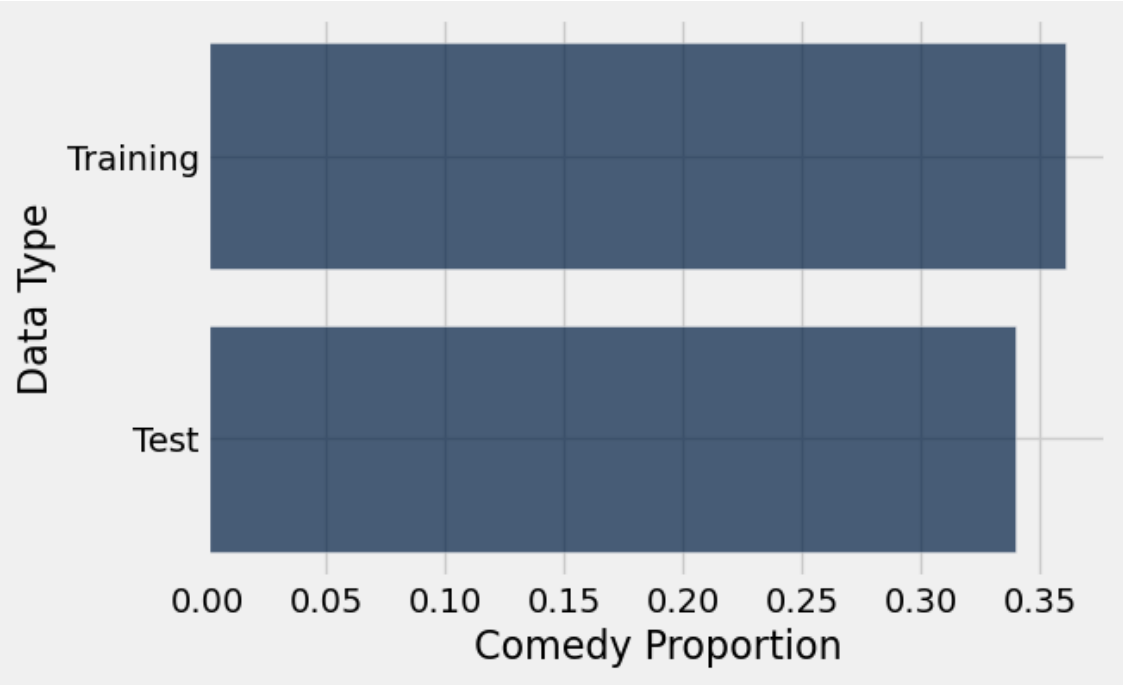

Correlation: 0.02324482071562719, magnitude greater than .2: False
0.0008
0.0006
0.0004
0.0002
0.0000
milk
0.0000
0.0002
0.0004
0.0006
0.0008
spoil

Question 1.3.1

Draw a horizontal bar chart with two bars that show the proportion of Comedy movies in each dataset (`train_movies` and `test_movies`). The two bars should be labeled "Training" and "Test". Complete the function `comedy_proportion` first; it should help you create the bar chart.

Hint: Refer to Section 7.1 of the textbook if you need a refresher on bar charts.

```
In [160]: def comedy_proportion(table):
              # Return the proportion of movies in a table that have the comedy genre.
              return table.where("Genre", "comedy").num_rows / table.num_rows

          # The staff solution took multiple lines.  Start by creating a table.
          # If you get stuck, think about what sort of table you need for barh to work
          Table().with_columns(
              "Data Type", make_array("Training", "Test"),
              "Comedy Proportion", make_array(comedy_proportion(train_movies), comedy_proportion(test_m
          ).barh("Data Type")
```



Question 3.1.7

In two sentences or less, describe how you selected your features.

I tried to look for words that were either common enough to be included in most, if not all, movies or seemed to be more common in one genre than in the other. For example, I chose the word "it" to ensure I wouldn't have any movies with zeroes across the board for all the words I selected, and I selected words like "kill" and "homicid(e/al/etc.)" because I'm sure they'd be more common in thriller movies than in comedy movies.

Question 3.3.3

Do you see a pattern in the types of movies your classifier misclassifies? In two sentences or less, describe any patterns you see in the results or any other interesting findings from the table above. If you need some help, try looking up the movies that your classifier got wrong on Wikipedia.

Many of the misclassified movies have more than one genre or incorporate themes and elements found in both comedy and horror movies; these movies are more nuanced than movies that fall into clear "comedy" and "horror" categories, such as "Beetlejuice", a well-known movie that is both comedy and horror, or "Storytelling" and "Smoke", comedy films whose plots involve themes like murder and abuse. Additionally, many of these movies' second or alternate genres are romance, sci-fi, or fantasy, such as "Intolerable Cruelty" and "The Atomic Submarine".

Question 4.2

Do you see a pattern in the mistakes your new classifier makes? How good an accuracy were you able to get with your limited classifier? Did you notice an improvement from your first classifier to the second one? Describe in two sentences or less.

Hint: You may not be able to see a pattern.

Once again, many of the misclassified movies fall under multiple genres and are quite nuanced, with most movies' second genre being action/adventure or mystery along with a couple of misclassifed romance movies. My second classifier performed a little worse than my first classifier (my first classifier's proportion correct is 0.76 while my second's is around 0.73), which I partly think is due to the constraint on the number of features being looked at by the second classifier.

Question 4.3

Given the constraint of five words, how did you select those five? Describe in two sentences or less.

I kept the fact that many of the movies my first classifier misclassified were romcoms or romance/thriller movies in mind, so a few of the five words I selected are meant to account for those films, like "love", "roman(ce, tic, etc.)", and "happi/y(ness, etc.)". I kept "kill" because I think it helps pinpoint thrillers, and "the", like "it" in my ten-feature list, is included so no movies have zeroes for their script-word proportions.