1. J-codes are procedure codes that start with the letter 'J'.

A. Find the number of claim lines that have J-codes.

51029

B. How much was paid for J-codes to providers for 'in network' claims?

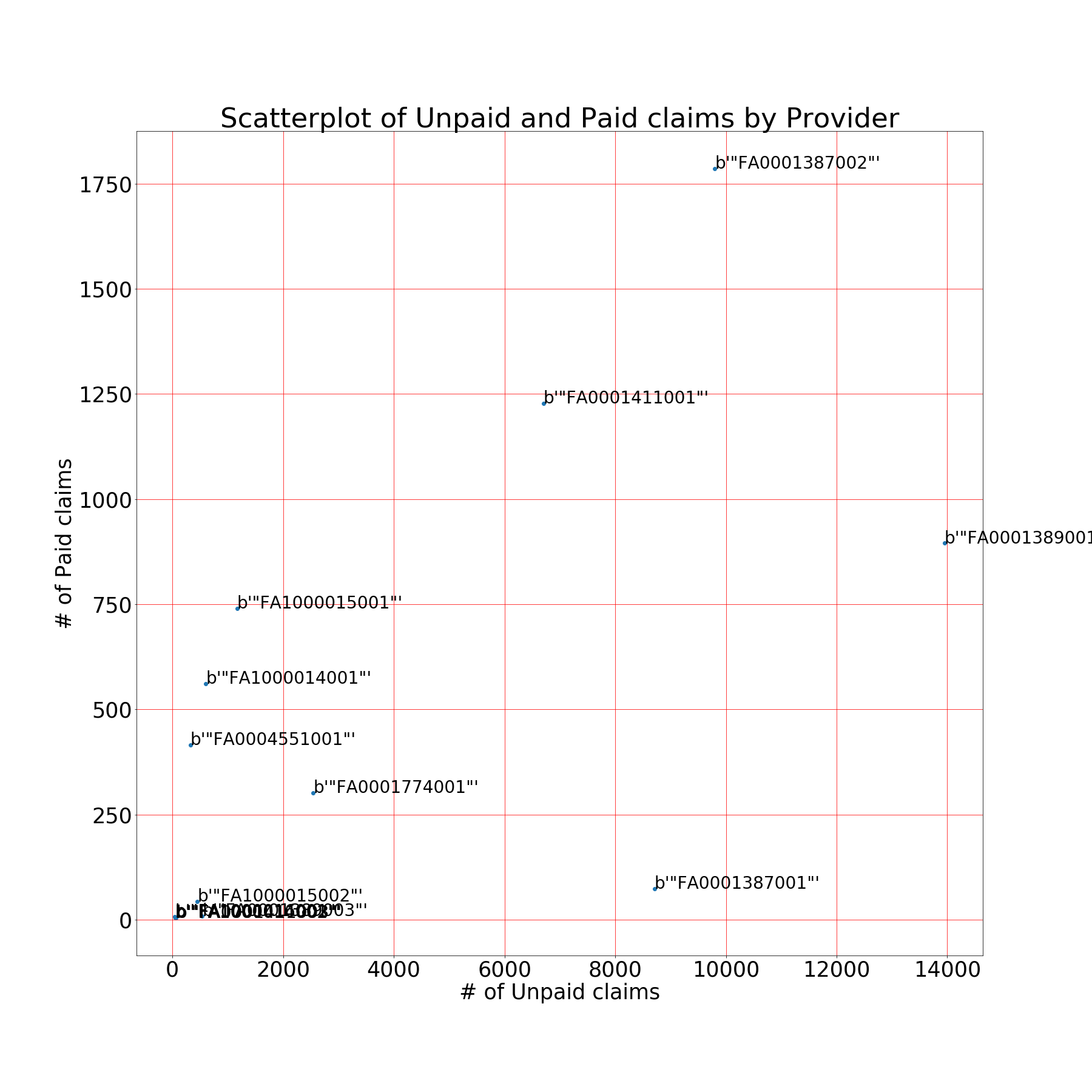
2417220.96029

C. What are the top five J-codes based on the payment to providers?

J1745, J0180, J9310, J3490, J1644

2. For the following exercises, determine the number of providers that were paid for at least one J-code. Use the J-code claims for these providers to complete the following exercises.

A. Create a scatter plot that displays the number of unpaid claims (lines where the ‘Provider.Payment.Amount’ field is equal to zero) for each provider versus the number of paid claims.



B. What insights can you suggest from the graph?

There appears to be a slight linear relationship between paid and unpaid claims for most providers. We can also see that a few providers handle a bulk of the claims represented in this dataset. FA0001387002, FA0001389001, and FA0001411001 process a much higher number of claims compared to the other providers. There also appears to 3 providers that process a very small number of claims. FA0001411003, FA1000014002, and FA100016001 all processed less than 100 claims which is proportionally small compared to the other providers.

C. Based on the graph, is the behavior of any of the providers concerning? Explain.

FA0001387001 is concerning as it paid a very small number of claims compared to its unpaid claims. This provider had 8710 unpaid claims and only 74 paid claims. It is concerning that this provider may be over rejecting claim payments. FA0001389001 is also concerning for the same reason. This provider only had 895 paid claims compared to 13947 unpaid claims.

3. Consider all claim lines with a J-code.

A. What percentage of J-code claim lines were unpaid?

88.1087224911325%

B. Create a model to predict when a J-code is unpaid. Explain why you choose the modeling approach.

I used both a random forest and logistic regression model to predict when a J-code was unpaid. I thought using 2 different models might give some insight and help me decide which was more accurate.

C. How accurate is your model at predicting unpaid claims?

The logistic regression model appears to be about 90% accurate whereas the random forest model is closer to 99% accuracy.