l Project. We've made it this far, and this will put us across the finish line.

Your task is to learn something while also demonstrating what you've learned. Hopefully, you've learned enough of Spark's inner workings to appreciate it's elegance and it's tediousness.

Your final project will be in multiple parts. This project is open-ended in an attempt to promote some creativity. You will be judged based on the question that you ask and how well you answer it using data.

Part 1: Pick a Data Set

I'm going to recommend several data sets, but you may select your own. The requirement here is that it must have at least 10,000 records. I want something that is not so small as to be ineffective yet not so big as to frustrate you. I do provide suggestions as to how you might use this data, but feel free to ignore these suggestions.

Good for Binary Classification

- Census Income
- Skin Segmentation This is perfect for Logistic Regression.
- Nursery

Good for Regression

- Power Consumption
- Electricity Readings
- Bike Sharing

Good for recommendation tasks.

- Online Retail
- Plants
- Drugs

Part 2: Formulate a question based on data.

There's no programming to this part. I want you to look at several data sets (or more, should you be so inclined to investigate further) and ask yourself questions that will be answered by the data.

Pick one question. Make it narrow. And then answer it. I give you some template questions below.

So here are some questions that you might ask,

1. Is variable Z dependent on independent variables W, X, and Y using Linear Regression?

Reports of Linear Regression should include the Root-Mean-Squared-Error (RMSE) (lower is better) and the Pearson R^2 result (higher is better). In these problems, you are trying to

estimate a value based on the belief that the value is dependent on other values. For instance, can you determine the sales of a store based on the number of customers who walk in per hour, the size of the floor mats, or the amount of hair on the manager's head? I assure you that some of these variables may impact the results more than you think. Is the R^2 reliable in getting the desired estimation?

2. Can variable Z (a binary result) be determined by examining variables W, X, and Y using Logistic Regression?

Now, be careful with this one. Can you determine who will win a football game based on the final point scores of the two teams? Yes, of course. That's how wins are determined. Pick a question that is less obvious, yet still somewhat obvious. Can you determine who will win a football game based on which team has the highest average yards per play? More than likely, yes, this is true. But how accurate is this claim? Can you determine who will win a basketball game based only on rebounds?

A famous example of this was when a statistician for the Oakland A's discovered that the on-base percentage was the most critical stat in baseball, so the team hired only players who performed well with that stat. The result was that the A's played extremely boring baseball yet improved their rate of winning games.

Any use of the Logistic Regression should come with an explanation of the accuracy of the training and the test data.

3. Can you build a recommendation engine based on a data set?

In other words, can you discover a commonality between two items that would cause someone to agree that the two items are similar? For example, if I said that I wanted to watch more movies like "Star Wars," which would be a better recommendation for me: "Young Frankenstein" or "Star Trek" or "Lord of the Rings." This question is more subjective than the other two questions and requires a degree of interpretation.

Part 4. Code

Write the Spark code which will help you to answer your question. I would expect you to have a training set and a test set in your code in order to properly test your idea.

Part 5. Write a report.

Write a 1 to 2-page report based on your findings. Since you are only asking and answering one question, you should need about than 0.75 to 2 pages to address this. First, state your question. Then formulate a hypothesis or a prediction to your question. This prediction should demonstrate your intuition. I want you to attempt to show a positive result. Next, describe your procedure, but don't show any code. (In computer science papers, academics don't show any code unless the topic is about code. This paper shouldn't be about code. It should be about your question and the data.) Finally, present your results and conclude on how close your result came to your prediction.

Part 6. Tack on your code.

But this is still my class, and I want to see your code. :)