Project 1: Predicting Catalog Demand

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/c0b53068-1239-4f01-82bf-24886872f48e/project

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (500 word limit)

Key Decisions:

Answer these questions

1. What decisions needs to be made?

The manager needs to make a decision about whether to send the catalog out to the new customers.

2. What data is needed to inform those decisions?

To answer this question, we need to calculate whether the expected profit contribution exceeds $10,000.
To calculate profit, we need to know the possible revenue, then we make an equation, profit=revenue*50%-6.5*250.
To calculate the revenue, we need to predict the average sale amount of each customer and multiply the probability of whether a customer will buy or not. Then make an equation, revenue=sum of (each ave sale amount*each probability to buy).
Therefore, we need to build a model to predict “average sale amount” of each new customer.

Step 2: Analysis, Modeling, and Validation

Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

Important: Use the p1-customers.xlsx to train your linear model.

At the minimum, answer these questions:

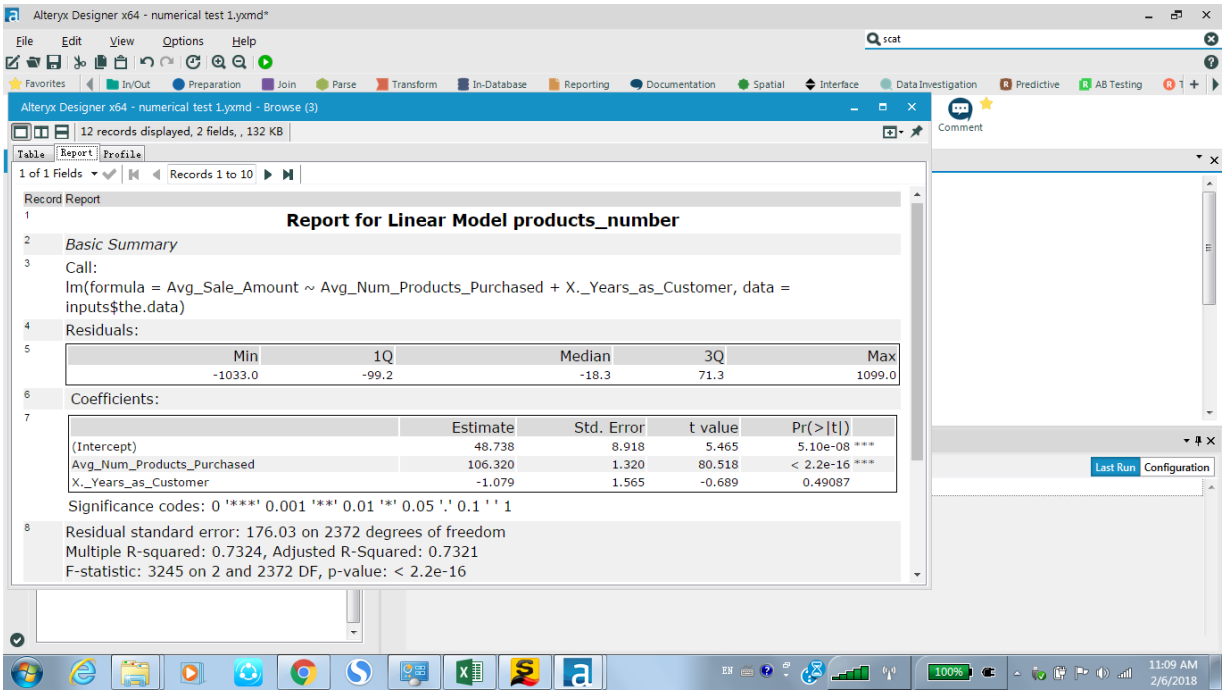
1. How and why did you select the predictor variables (see supplementary text) in your model? You must explain how your continuous predictor variables you’ve chosen have a linear relationship with the target variable. Please refer to this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

For numerical predictor, we could consider average number of products purchased, or the number of years as customer. We need to test and see the linear relationship.

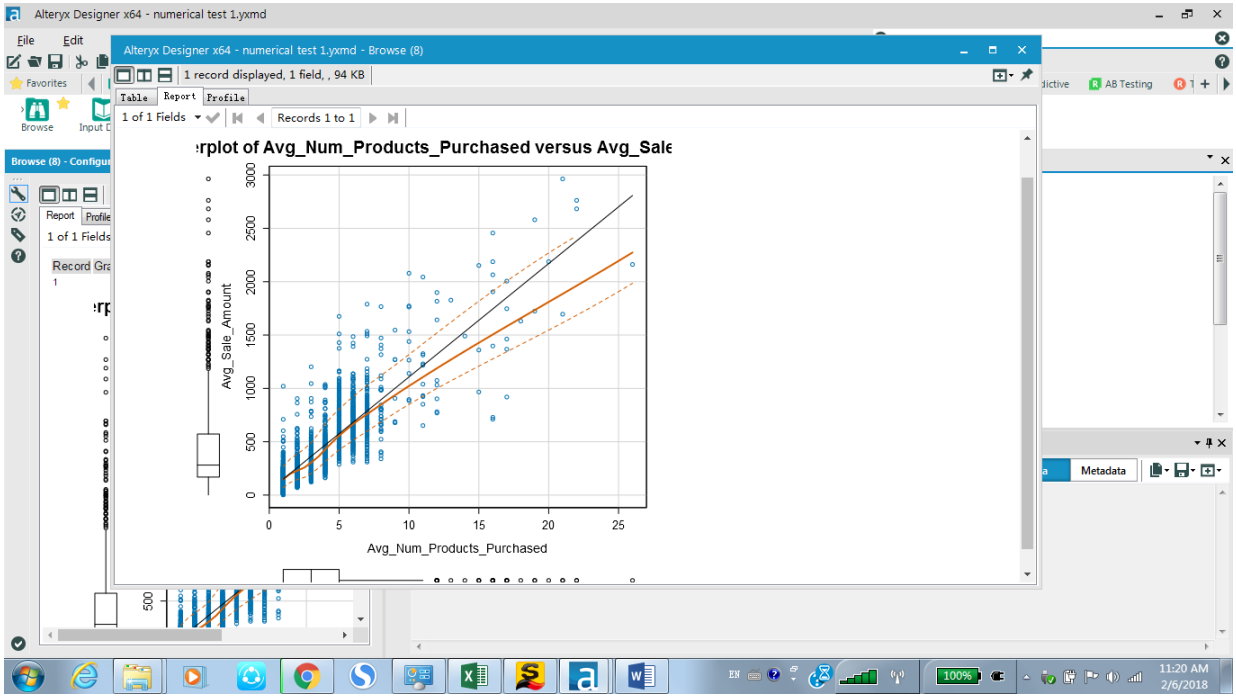
For categorical predictor, we could consider customer segment, city, zip code, or store. Though city, zip code or store seems not so possible, but they are worthy to be tested, since people from rich neighborhood may more willing to purchase high quality furniture. And neighborhood could be reflected from either zip code, or store. We also need trial and see them.

For the numerical predictor, we built a linear regression model to test both the “average number of products purchased” and the “year as customer”.

Based on the report, the average number of products’ P value is small enough to be significant. The years of customer’s P value is too big to be significant.

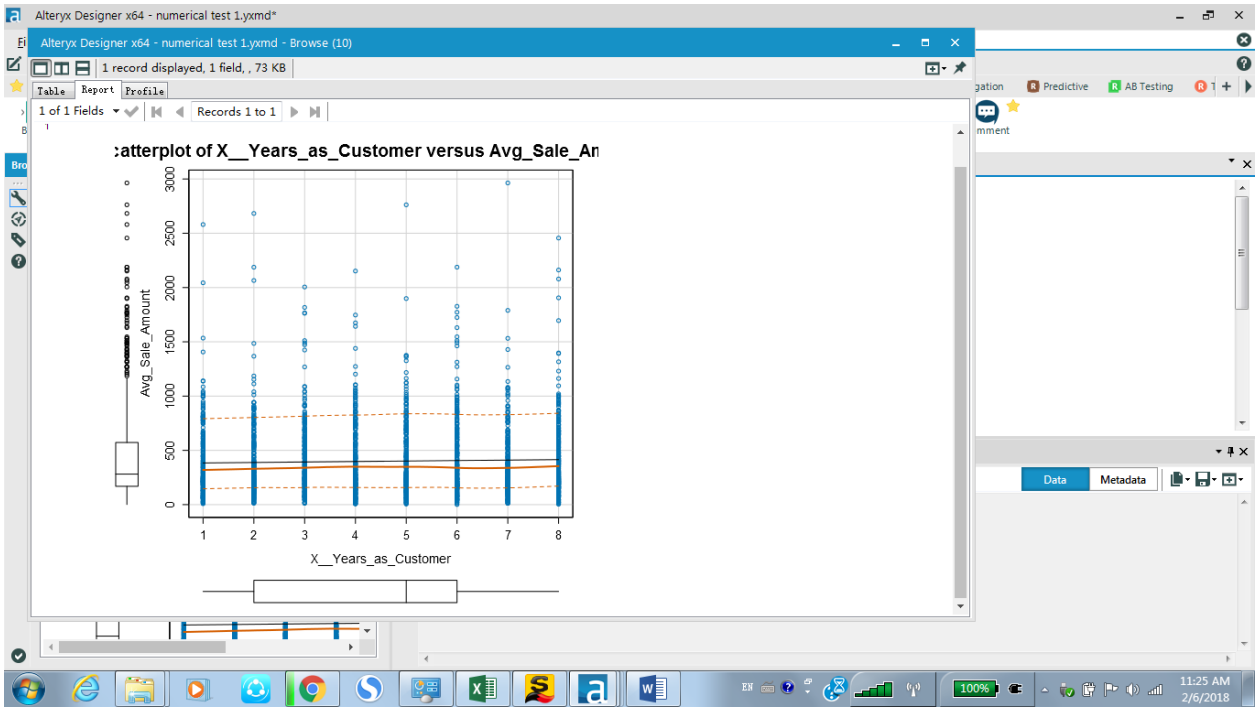


I then added scatterplot to see how the two numerical predictor do.
The average number of products looks like this:



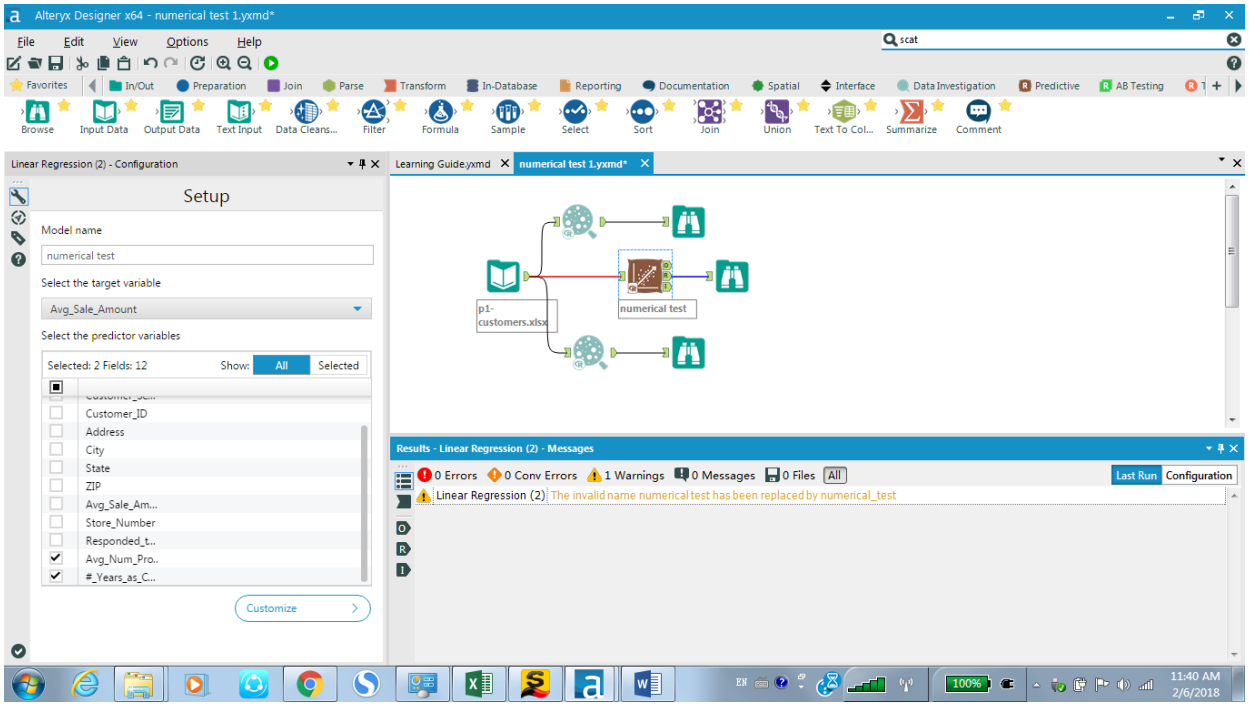
It is barely ok. At least it goes with the same trend with the sale amount. We can show there is a linear relationship between the average number of products and the sale amount.

Then I tried the year of customer.
It looks like this:

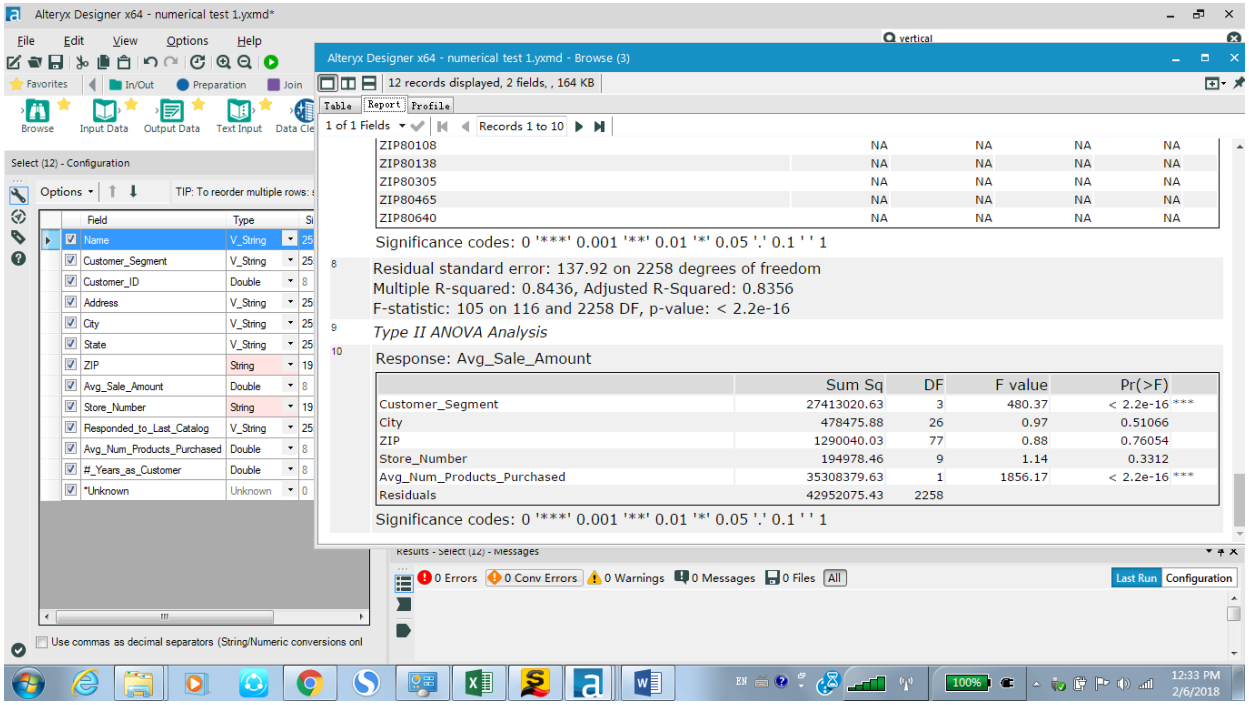


There is no linear relationship between the years of customer and the amount of sale.

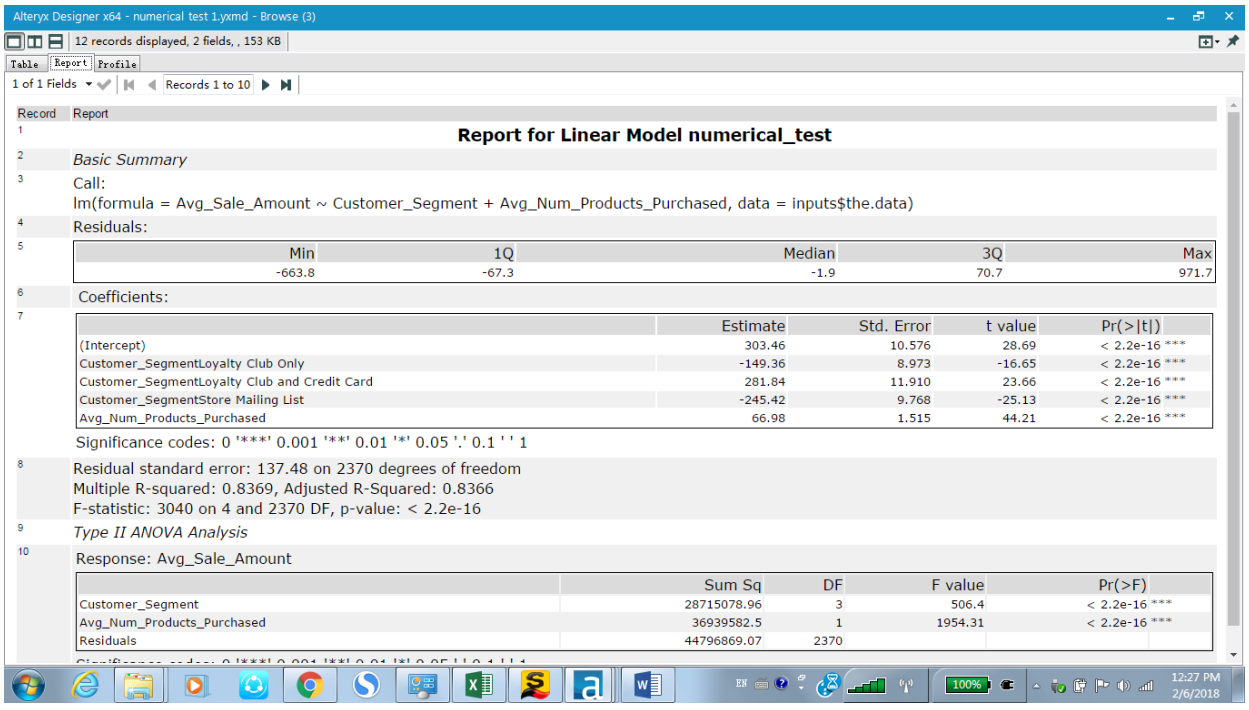
By now, I used two scatterplot functions and one linear regression function in Alteryx.
It goes like this:



For categorical predictor, I moved on to trail and see the customer segment, city, zip code, and store. I configured the linear regression function.
I added a sort function before linear regression function to change the type of “zip code” and “store number” to string, because I wanted to use the two as categorical predictors, not numerical ones. Then I added the four potential categorical predictors to the linear regression configuration and got the result like this:



We can see that the city, ZIP, and store actually have no linear relationship with the average sale amount. And I am now sure that the customer segment, as a categorical predictor, has linear relationship with the average sale amount. We now have located the categorical predictor as the customer segment, and the numerical predictor as the average number of products purchased. I left the two in the linear regression configuration and did the calculation again.



2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

We can see from above report chart, that the predictors all have significant relation with the target since their P values are much smaller than 0.05 and with three stars. For the equation, the P value is small, much smaller than 0.05. And the multiple R-squared and the adjusted R-squared are big, quite near to 1. We can say the equation is a good model.

3. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)

Based upon the above analysis, the equation shall be

Y=303.46+66.98*Ave_Num_Products_Purchased-149.36*(if Loyalty Club Only)+281.84* (if Loyalty Club and Credit Card)-245.42* (if Store Mailing List)+0 (if Credit Card Only)

Important: The regression equation should be in the form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

For example: Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

Note that we **must** include the 0 coefficient for the type Cash.

Note: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

Step 3: Presentation/Visualization

Use your model results to provide a recommendation. (500 word limit)

At the minimum, answer these questions:

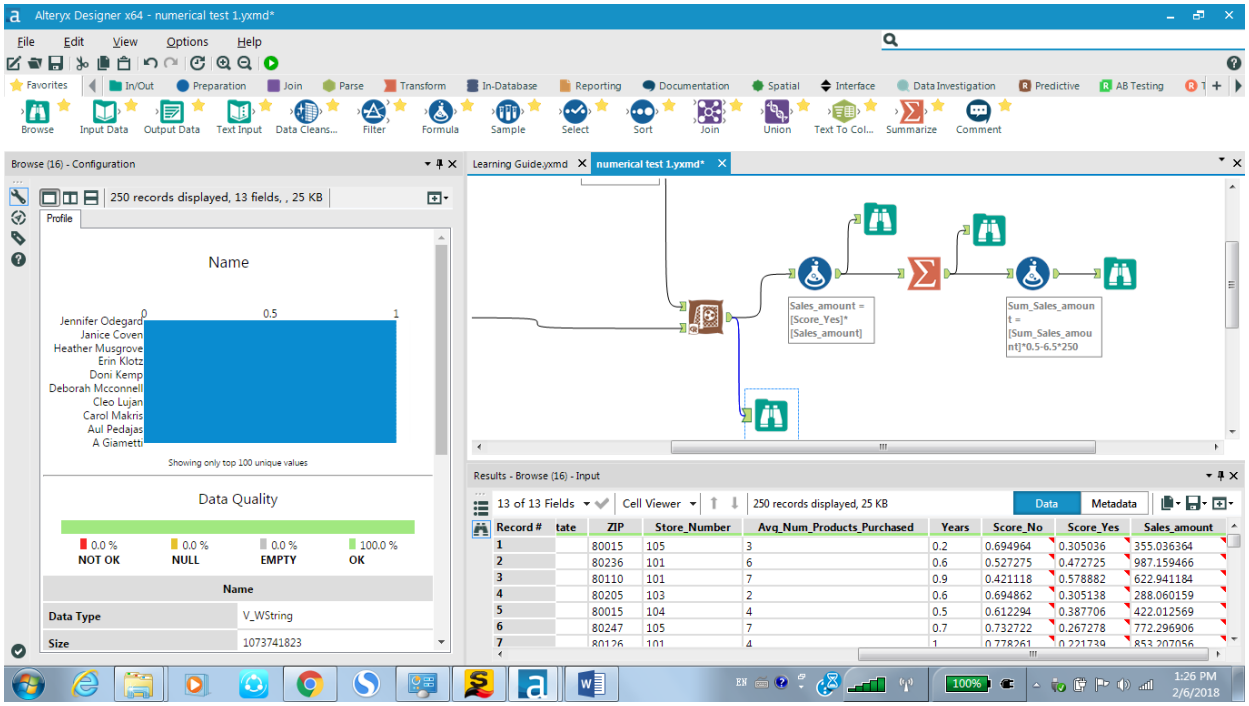
1. What is your recommendation? Should the company send the catalog to these 250 customers?

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

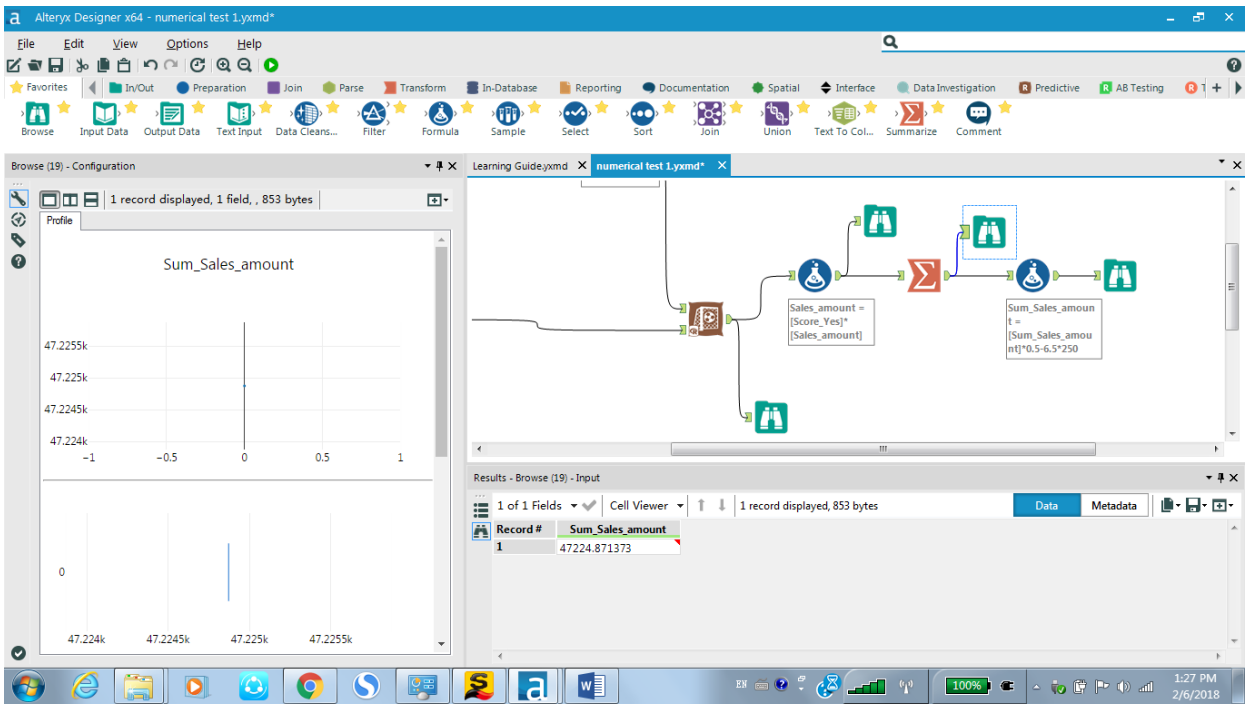
3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?

My suggestion is that we may go send the catalog out to the new customers.

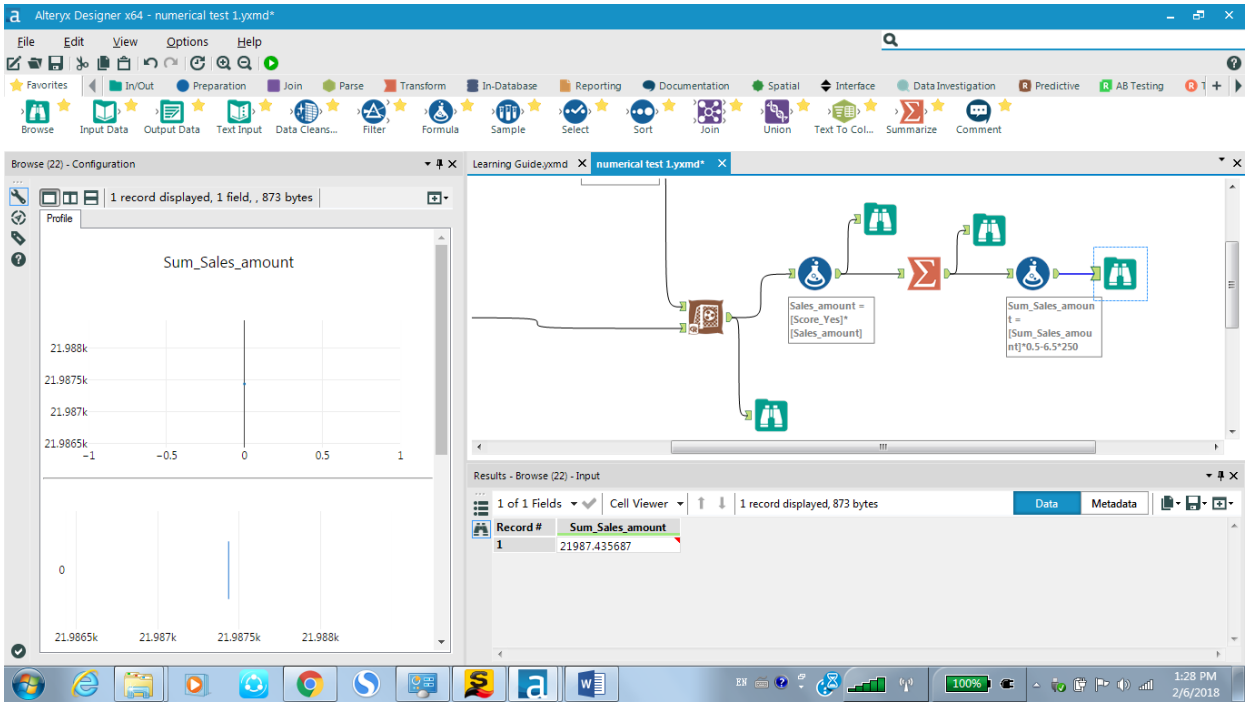
We run the above-mentioned model by inputting the mailing list of new customers and result sale amount for each new customer.



We add a formula function to multiply “the possibility to purchase” to each the sale amount and then add sum function to calculate the total possible sale amount, which is the possible revenue. It equals 47224.87

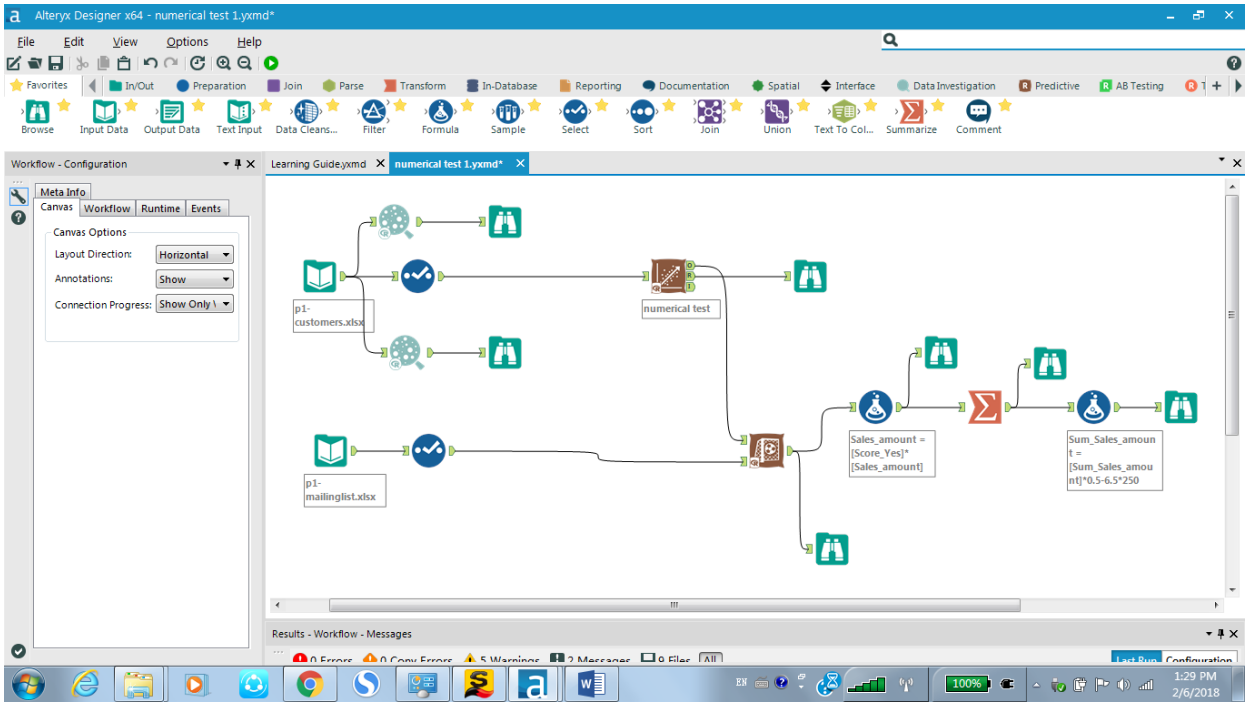


Then, we calculate the profit by adding a formula function to the end, using the equation, profit=revenue*0.5-6.5*250, and result the profit as 21987.44



Since the expected profit contribution, $21,987.44 exceeds $10,000, I would suggest send the catalog out to the new customers.

My whole picture of Alteryx looks like this (I put more-than-needed browse functions to the flow work just to double check each step):



Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.