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?

   We need to decide if we send out the catalogues to the new customers or not.

2. What data is needed to inform those decisions?

   What is the expected profit from the 250 new customers? Will it exceed $10,000? We will need the cost to send out the catalogues, as well as the expected profit for each new customer based on demographic/past purchase data on those customers. We will need the same demographic/past purchase history on older customers, in order to make a predictive model.

Step 2: Analysis, Modeling, and Validation

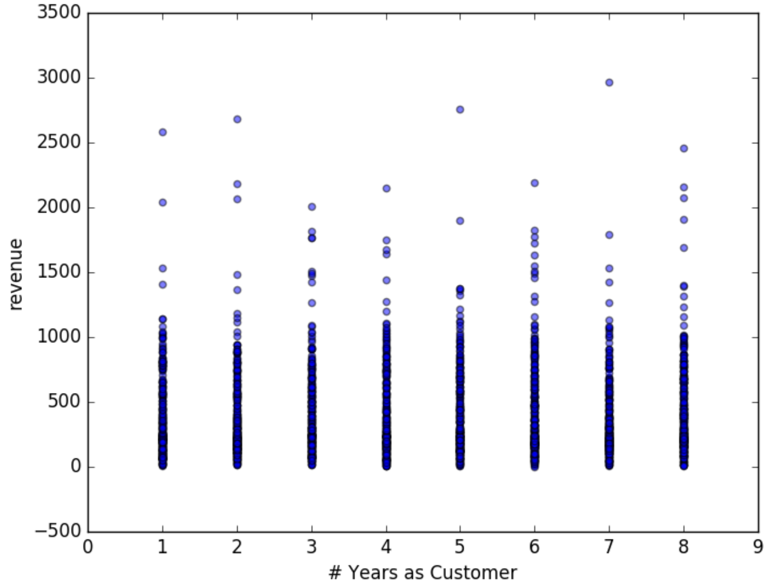
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

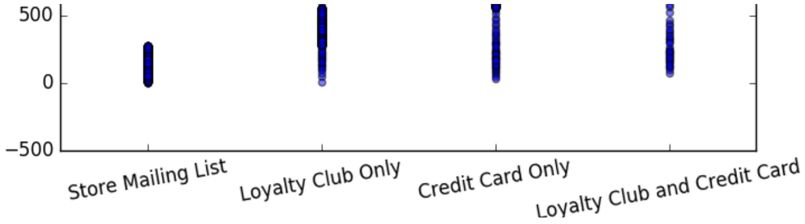
Important: Use the p1-customers.xlsx to train your linear model.

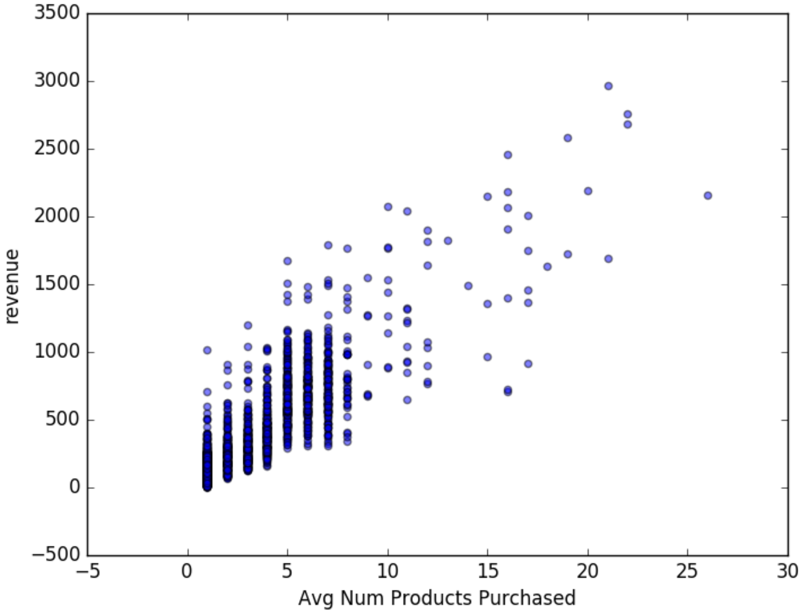
At the minimum, answer these questions:

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.

I selected customer segment and avg num products purchased. These appear to have a linear relationship with profit from the scatterplots. I also looked at avg num years as a customer, and this did not appear related to profit. The p-value from a linear regression showed years as customer had a p-value of 0.104 (> 0.05), which means we accept the null hypothesis, which is that the predictor variable has no relationship to the target.



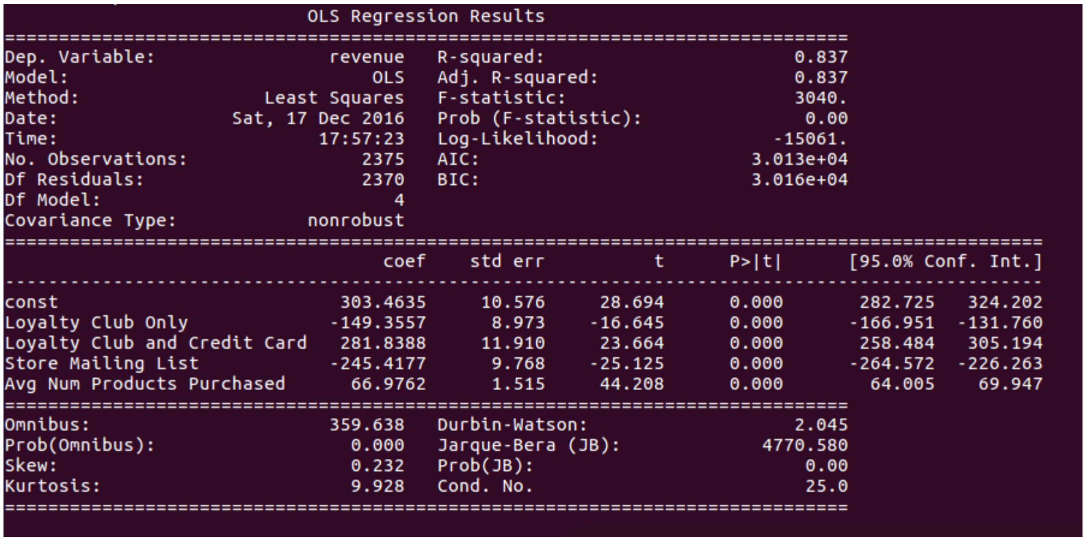




2. Explain why you believe your linear model is a good model. You must justify your

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.

The model has a decent adjusted r-squared value (0.837) and the p-values of the coefficients are 0.0, meaning they have a meaningful relationship to the target variable. The regression summary from statsmodels (Python) follows:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                revenue   R-squared:                       0.837
Model:                            OLS   Adj. R-squared:                  0.837
Method:                 Least Squares   F-statistic:                     3040.
Date:                Sat, 17 Dec 2016   Prob (F-statistic):               0.00
Time:                        17:57:23   Log-Likelihood:                -15061.
No. Observations:                2375   AIC:                         3.013e+04
Df Residuals:                    2370   BIC:                         3.016e+04
Df Model:                           4
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------------------------
const                          303.4635     10.576     28.694      0.000       282.725   324.202
Loyalty Club Only             -149.3557      8.973    -16.645      0.000      -166.951  -131.760
Loyalty Club and Credit Card   281.8388     11.910     23.664      0.000       258.484   305.194
Store Mailing List            -245.4177      9.768    -25.125      0.000      -264.572  -226.263
Avg Num Products Purchased      66.9762      1.515     44.208      0.000        64.005    69.947
==============================================================================
Omnibus:                      359.638   Durbin-Watson:                   2.045
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4770.580
Skew:                           0.232   Prob(JB):                         0.00
Kurtosis:                       9.928   Cond. No.                         25.0
==============================================================================
```

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)

   Revenue =
   CC_only * 0
   + Loyaly_club_only * -149.36
   + loyalty_club_and_cc * 281.84
   + store_mailing_list * -245.42
   + Avg_num_purch_prods * 66.98
   + 303.46

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?

I recommend sending the catalogue, because the model predicts about $22K -- this is greater than the $10K threshold that was set to determine go/no go.

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

   I made the model using dummied variables for the Customer Segment data, the Avg num prods purchased, and an intercept term. I fit the model to the past customer data, and used it to predict the total revenue from the new customers. I then multiplied each prediction by the Score_Yes for each customer, multiplied by 50% for the profit margin, subtracted 6.50 for each customer (for cost of catalogue production/mailing) and summed up the results to get $22K.

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

   $21987.44