MAKE A COPY

Project 1: Predicting Catalog Demand

## **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 predict the Sales of product based on following parameter :

Name

Customer\_Segment

Customer\_ID

Address

City

State

ZIP

Avg\_Sale\_Amount

Store\_Number

Responded\_to\_Last\_Catalog

Avg\_Num\_Products\_Purchased

#\_Years\_as\_Customer

Assumption-

* All sales have been made with 50% profit
* Each unique customer with customer has to receive the catalog
* Each catalog price is 6.50$ each

Mainly we would like to make a decision to send the catalog to new customer or not based on minimum profit required ( 10k) by management.

1. What data is needed to inform those decisions?

* Data about the past sales is used to predicting the profit for new customers.
* Percentage of probability that a new customer will buy a catalog and purchase items?
* Information about current customers, shopping behavior, location etc.

## **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.***

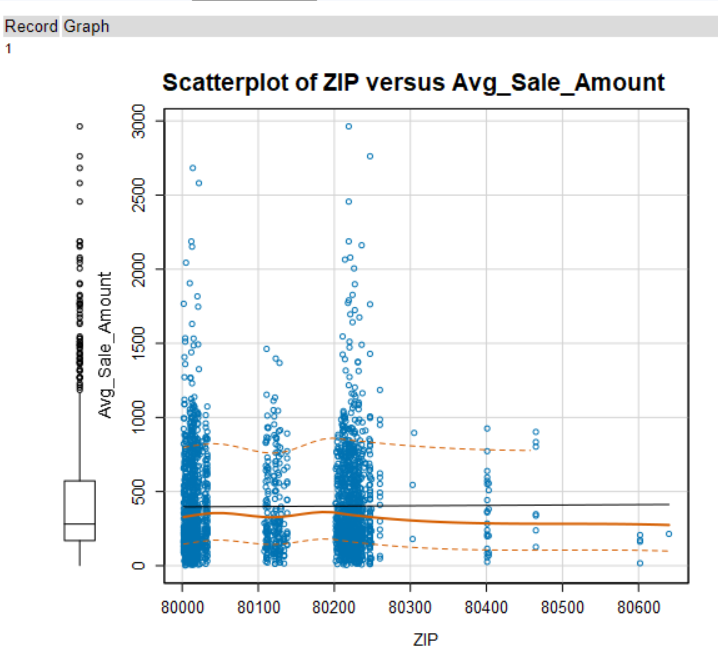
*At the minimum, answer these questions:*

1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you’ve chosen have a linear relationship with the target variable. Please refer back to the “Multiple Linear Regression with Excel” lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

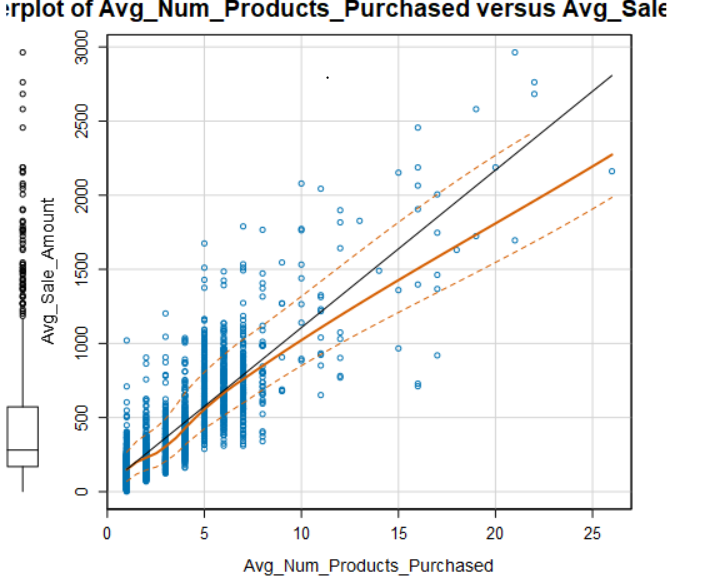
I created scatter plot between Avg sales value (y – axis) vs other variables. Except number of purchased products and customer segment others had greater than .05 p value.

As we know that, P values evaluate whether null hypothesis is true or not. As per null hypothesis, there is no effect of 1 variable on another for predicting the results values.

For ex: relation between zip code and avg sales amount

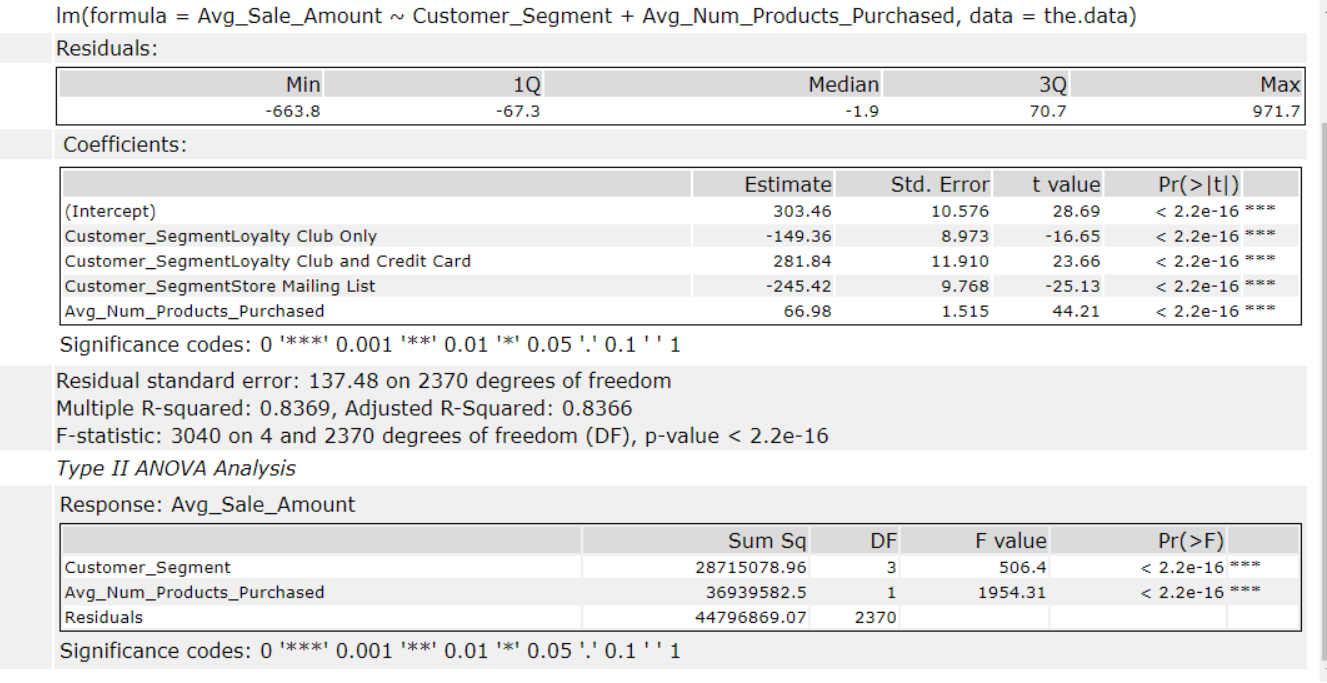


For ex: relation between number of products purchased and avg sales amount



1. 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.

Here is ANOVA diagram:

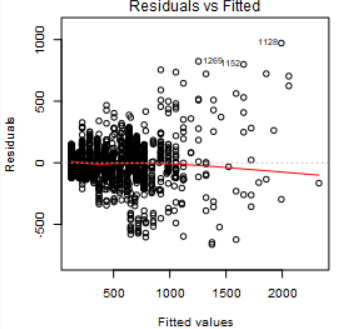


As per above ANOVA analysis, P value is significant lower for customer\_segment and Avg\_num\_products\_purchased, hence we can reject the null hypothesis for above.

R-squared is a statistical measure of how close the data are to the fitted regression line. in this case R-square is 0.8369.

The adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors. The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.  It is always lower than the R-squared. in this case R-square is 0.8366

R-squared cannot determine whether the coefficient estimates and predictions are biased, which is why we have assessed the residual plots. It looks symmetrical ( linear) on +ive and -ive so we can safely accept the above linear algorithm.



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)

**Important: The regression equation should be in the form:**

*Y = 303.46- (149.36)\* Customer\_SegmentLoyalty Club Only+ 281.84\* Customer\_SegmentLoyalty Club and Credit Card - 245.42\* Customer\_SegmentStore Mailing List + 66.98\* Avg\_Num\_Products\_Purchased+ 0\* Customer\_Segmentcustomercreditcardonly*

## **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?

Yes, we make profit of 21987.43

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

We used the linear regression to predict the Avg sales for 250 mailing list customer data by multiplying the probability of purchasing the product. then we calculated the total Sales bu adding all Sales across customer. Since assume sales makes 50% margin profit so we multiplied with .5 on sales value to calculate the cost then we subtracted the cost of printing( 6.5 $ per catalog/ per customer) the catalog to calculate the total profit.

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

we make profit of 21987.43

Before you Submit

Please check your answers against the requirements of the project dictated by the [rubric](https://review.udacity.com/#!/rubrics/186/view) here. Reviewers will use this rubric to grade your project.