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)*

Business understanding is observed as there are some retail store and they sell some items. They also record their customer activity. Dataset is of customer activity and their activity aggregated at customer level. The data set contains 12 features with avg\_sales as target feature and no missing values.

### **Key Decisions:**

1. What decisions needs to be made?

Ans. Decisions is to find potential sales if the new catalog of items were introduced to the listed customers and by using likelihood score per customer to estimate sales value.

1. What data is needed to inform those decisions?

Ans. Historical data of the customers with their purchase and interest in our store. This data allows us the estimate the sales given the catalog.

## **Step 2: Analysis, Modeling, and Validation**

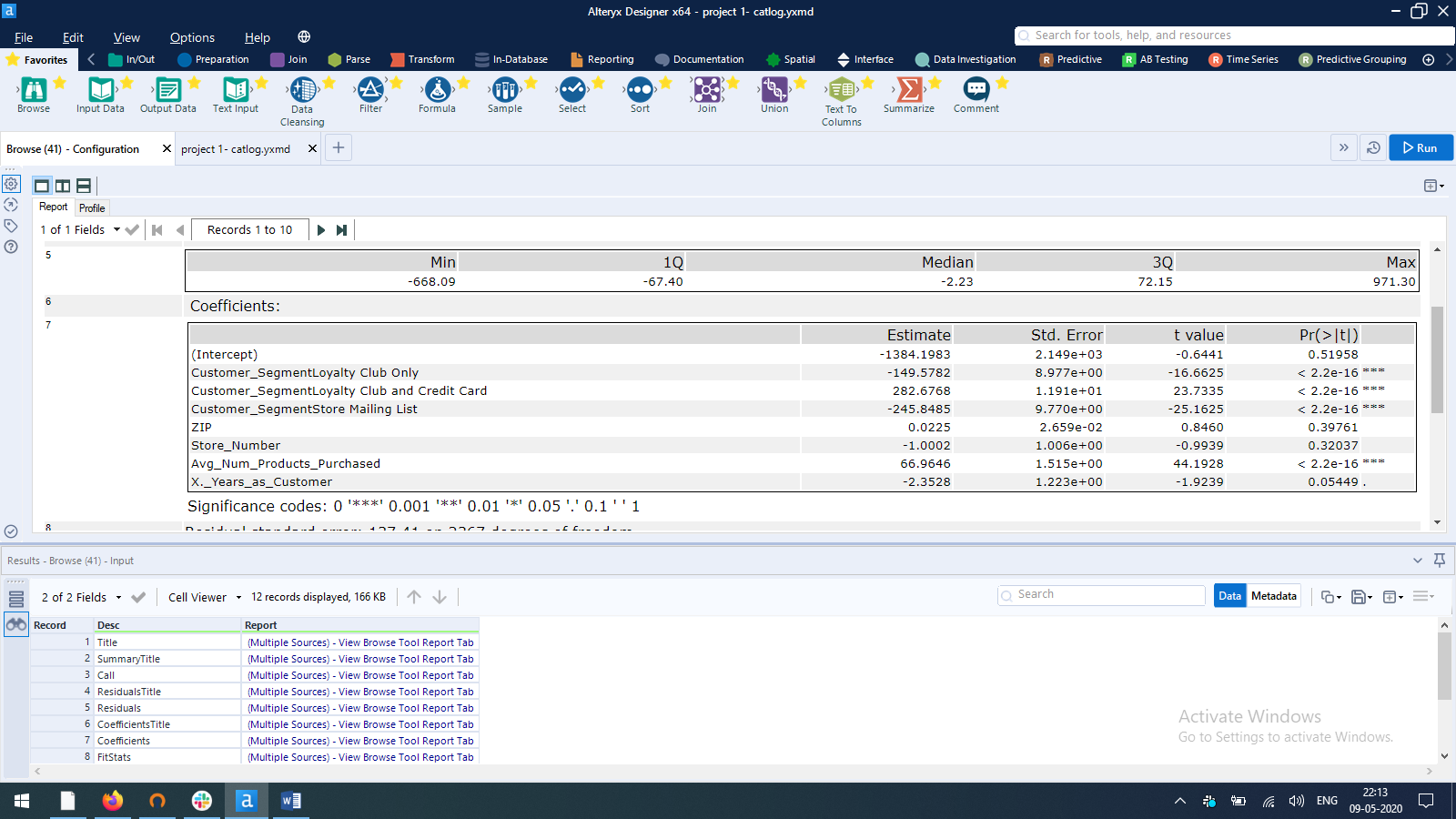
*Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)*

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

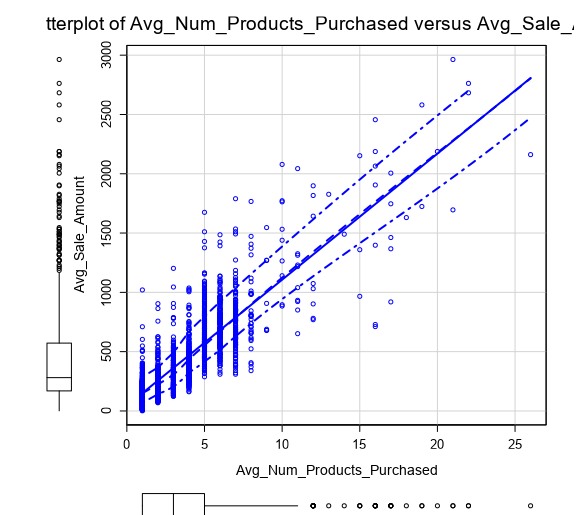
*At the minimum, answer these questions:*

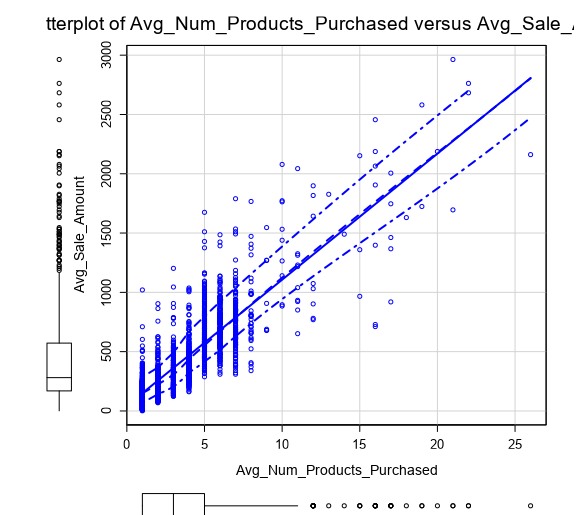
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you’ve chosen have a linear relationship with the target variable. Please refer back to the “Multiple Linear Regression with Excel” lesson to help you explore your data and use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

Ans.



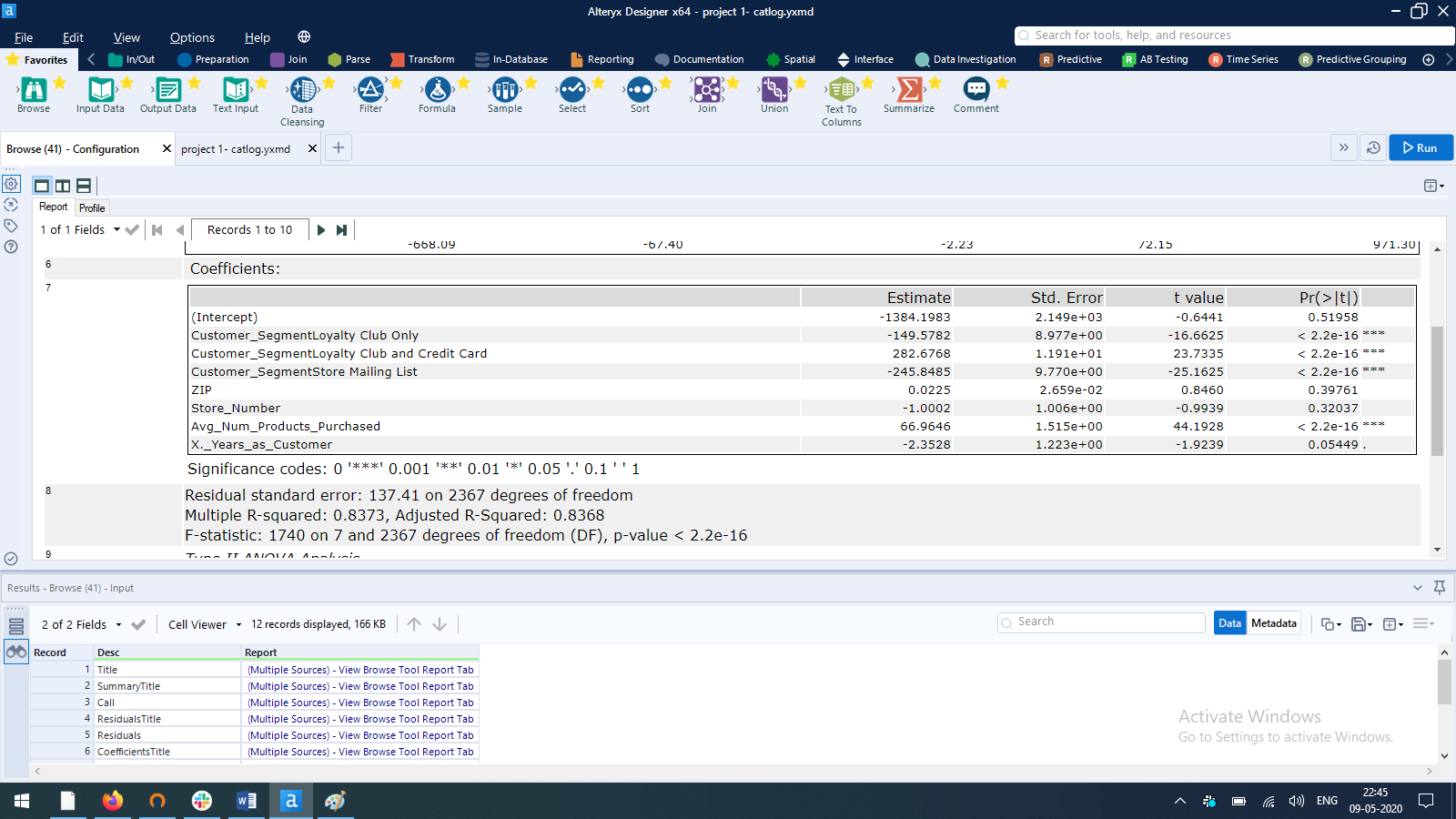
From the above report we can find out that only 2 variables are <= 0.05 p-value in other word statistically significant.

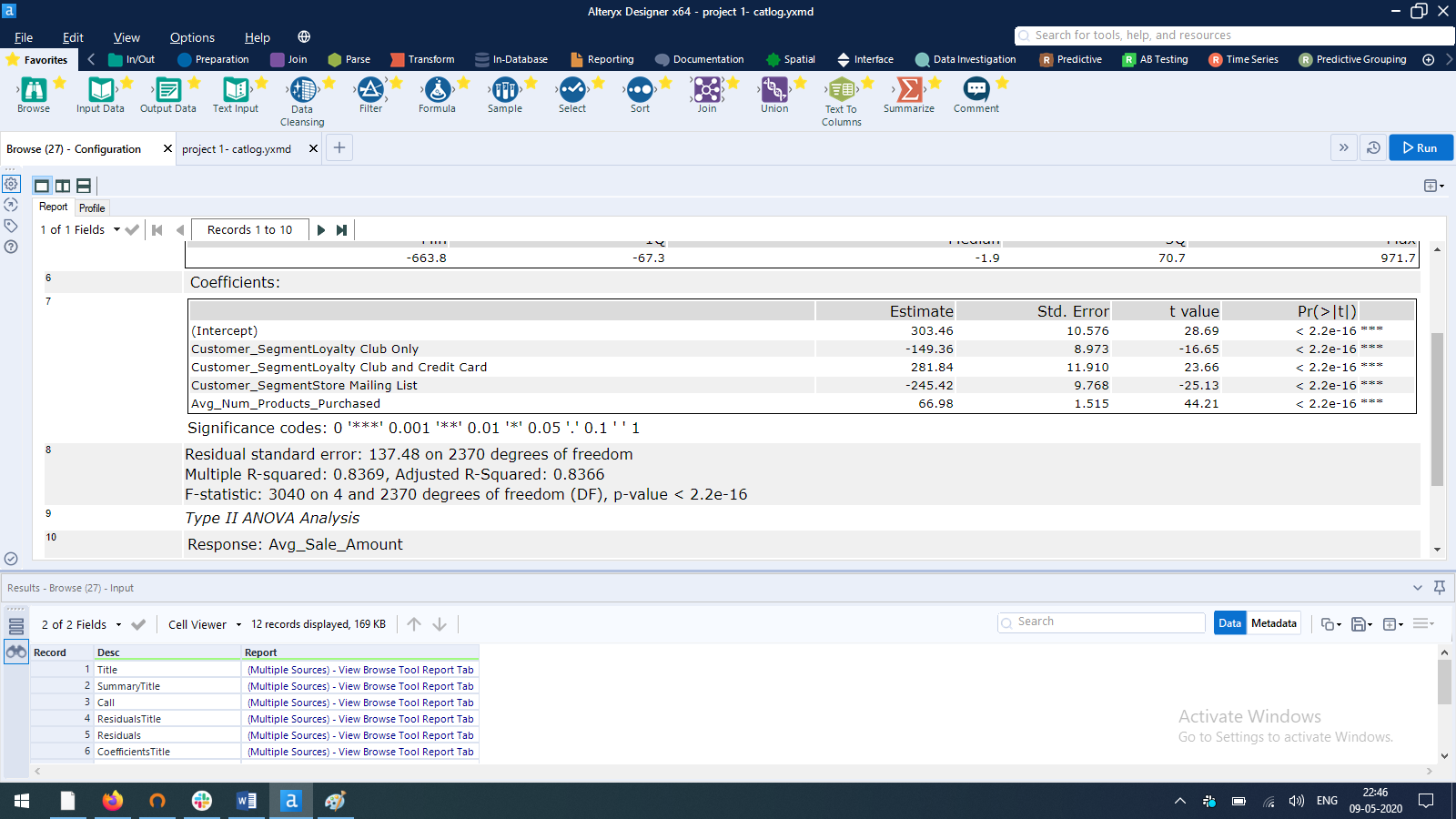




1. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that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

Ans.





“In **statistics**, the **p**-**value** is the probability of obtaining results as extreme as the observed results of a **statistical** hypothesis test, assuming that the null hypothesis is correct. ... A smaller **p**-**value** means that there is stronger evidence in favor of the alternative hypothesis”

\_\_ Google

As, p -value have that significance in regression then I tried to use that p-value as parameter to remove a feature and see if I can improve previous accuracy (R2\_score value) but it is seen that by removing the feature we able to improve that significate but a insight emerges that the feature which are removed are contributing anything much.

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)

*Ans. Est Sales Value =* ***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***

## 

## **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?

Ans. Yes I recommend to introduce our new catalog to the customers

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

Ans.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| bin | 0-20 | 20-40 | 40-60 | 60-80 | 80-100 |
| Avg | 37.89186 | 69.89235 | 128.2626 | 182.4287 | 242.54 |

| Bin Score Yes | Frequency | Percent | Cumulative Frequency | | Cumulative Percent |
| --- | --- | --- | --- | --- | --- |
| 20-40 | 150 | 60.00 | 150 | 60.00 | |
| 0-20 | 40 | 16.00 | 190 | 76.00 | |
| 40-60 | 31 | 12.40 | 221 | 88.40 | |
| 60-80 | 17 | 6.80 | 238 | 95.20 | |
| 80-100 | 12 | 4.80 | 250 | 100.00 | |

| Bin Score Yes | Frequency | Acquired % | Avg net profit per cust. | | | total |
| --- | --- | --- | --- | --- | --- | --- |
| 20-40 | 150 | 70.00 | | 88.00 | 9240 | |
| 0-20 | 40 | 10.00 | | 88.00 | 325 | |
| 40-60 | 31 | 80.00 | | 88.00 | 2200 | |
| 60-80 | 17 | 90.00 | | 88.00 | 1408 | |
| 80-100 | 12 | 100.00 | | 88.00 | 1056 | |
| total | 250 | 66.00 | | 88.00 | 14229 | |

It is fair to assume that there is very less probability that customer from bin 0-20 will buy from over catalog but will recommend to introduce to them also. From above insight I have taken global avg to calculate expected profit which inherently have 72.40 per cust. as std. deviations due to high ticket size for high probability bins.

Note that the Acquired % is a assumption due to lack of historical data and insights.

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

Ans. The profit estimation are 14,229 to 21,987 if we introduce our catalog to 250 customers which almost 30 to 50% of over investment

|  |  |
| --- | --- |
| Metric | Value |
| Sum\_Predicted\_Sales | 138292.1301 |
| Sum\_Sales\_yes\_value | 47224.87134 |
| Sum\_average gross margin | 23612.43567 |
| Sum\_net profit | 21987.43567 |
| Count | 250 |
| Median\_net profit | 64.98418427 |
| Avg\_net profit | 87.94974269 |
| StdDev\_net profit | 72.39440262 |

.