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?

Whether or not to send the catalog to the new customers. i.e., will it be worth sending this year's catalog to the 250 new customers?

2. What data is needed to inform those decisions?

Data on current customers and similar fields for the new customers will be needed. Data needed, in this case, to inform the decisions here include data on customer segments and the number of products they purchased.

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.

To start with, I tested for the existence of relationships between each available variable and the target variable (average sale amount). This was to ensure that I would not be leaving out any potentially significant variables. This limited the variables from all the available variables to average number of products purchased (positive relationship). To be sure of the results, I ran all the variables with type "Double" and customer segment (largely unique) through the regression model to test for and check their significance. This revealed that customer segments and average number of products purchased were the most significant, with p-values under 0.05.

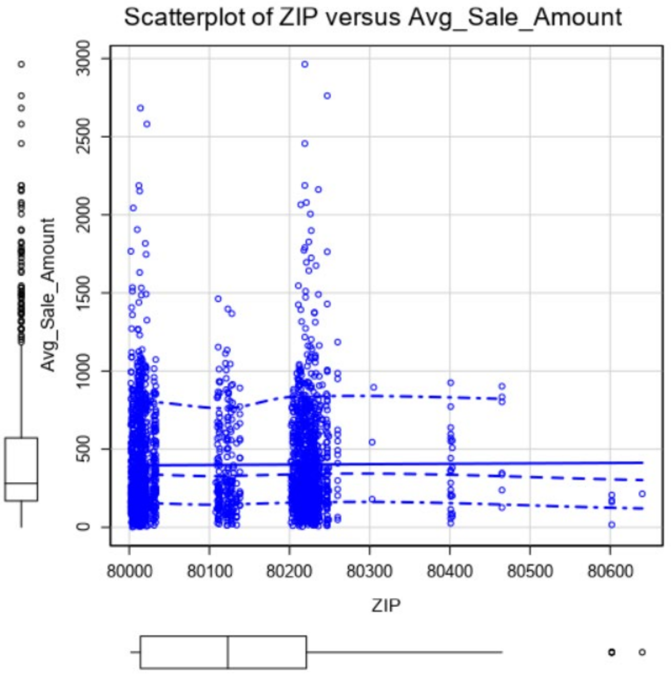


Image 1: Scatterplot of ZIP vs. Avg. Sale Amount

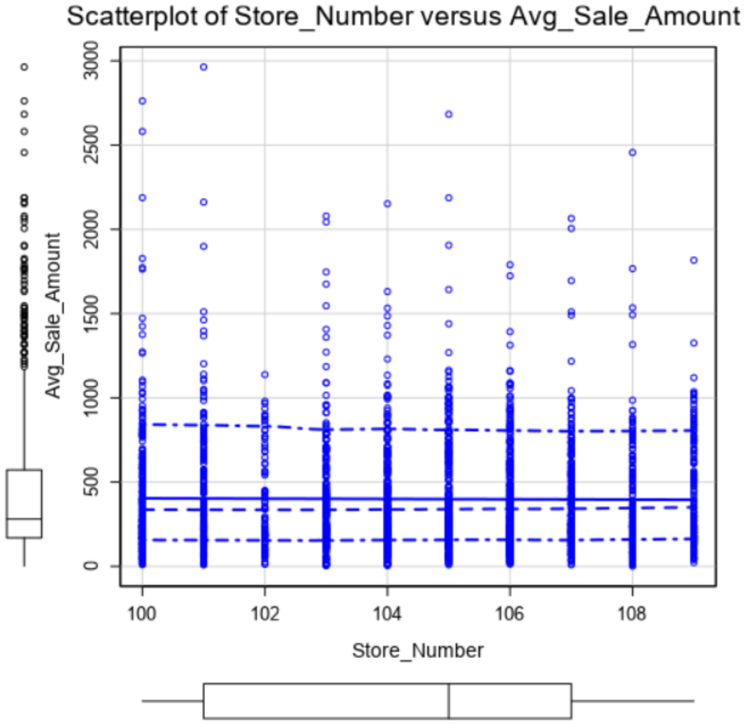


Image 2: Scatterplot of Store Number vs. Avg. Sale Amount

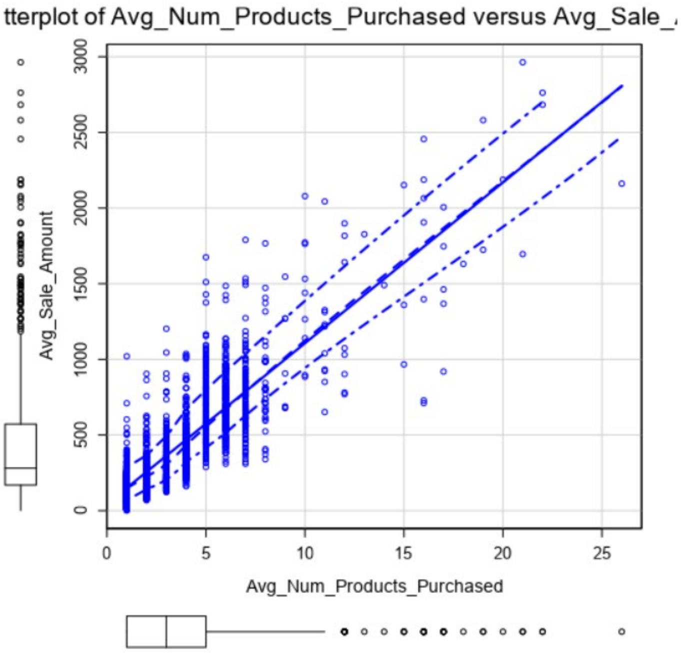


Image 3: Scatterplot of Average Num. of Products Purchased vs. Avg. Sales Amount

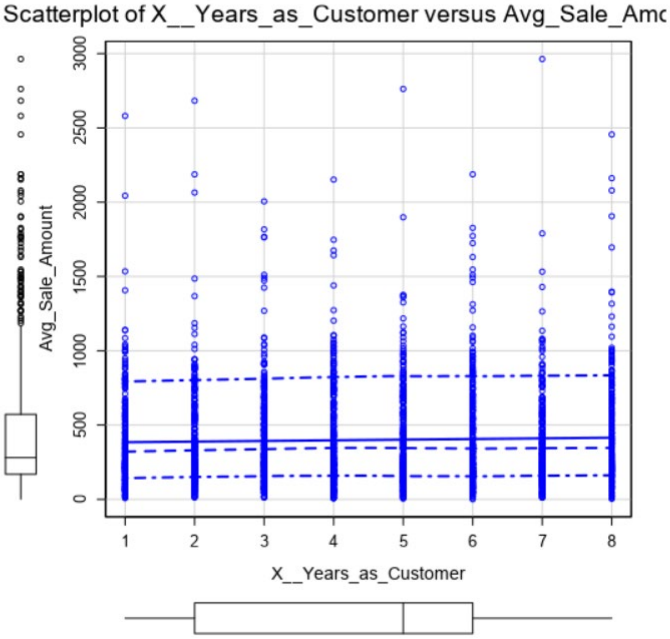
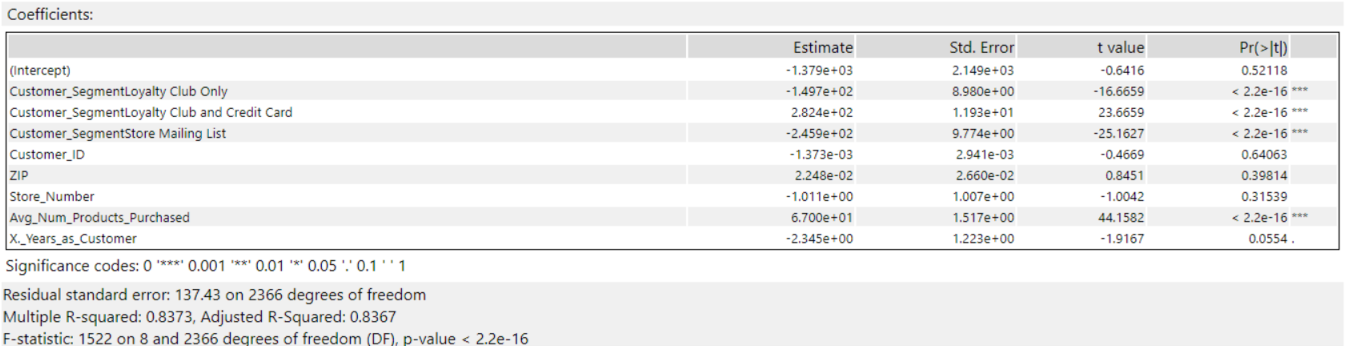


Image 4: Scatterplot of Number of Years as Customer vs. Avg. Sales Amount



Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.379e+03	2.149e+03	-0.6416	0.52118
Customer_SegmentLoyalty Club Only	-1.497e+02	8.980e+00	-16.6659	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	2.824e+02	1.193e+01	23.6659	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-2.459e+02	9.774e+00	-25.1627	< 2.2e-16 ***
Customer_ID	-1.373e-03	2.941e-03	-0.4669	0.64063
ZIP	2.248e-02	2.660e-02	0.8451	0.39814
Store_Number	-1.011e+00	1.007e+00	-1.0042	0.31539
Avg_Num_Products_Purchased	6.700e+01	1.517e+00	44.1582	< 2.2e-16 ***
X._Years_as_Customer	-2.345e+00	1.223e+00	-1.9167	0.0554 .

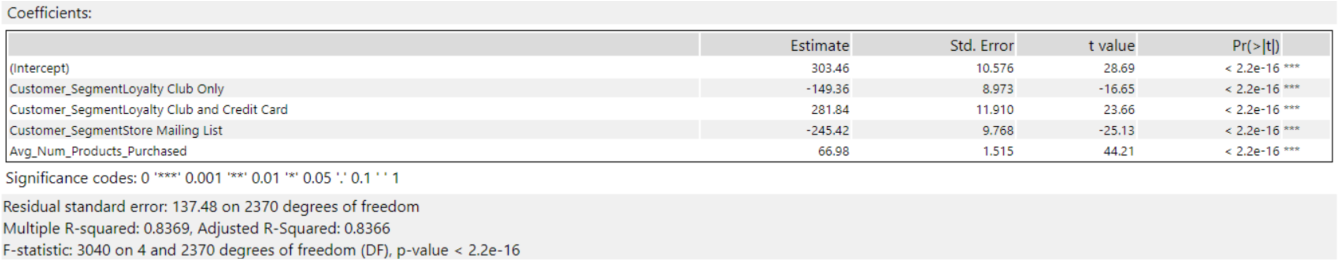
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.43 on 2366 degrees of freedom
Multiple R-squared: 0.8373, Adjusted R-Squared: 0.8367
F-statistic: 1522 on 8 and 2366 degrees of freedom (DF), p-value < 2.2e-16

Image 5: Segment of report on initial model with all variables

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 is good because it considers only significant predictor variables (per their p-values as explained above), and the adjusted r-square value of 0.8366 further highlights the strength of the model.



Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	303.46	10.576	28.69	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-149.36	8.973	-16.65	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	281.84	11.910	23.66	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-245.42	9.768	-25.13	< 2.2e-16 ***
Avg_Num_Products_Purchased	66.98	1.515	44.21	< 2.2e-16 ***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366
F-statistic: 3040 on 4 and 2370 degrees of freedom (DF), p-value < 2.2e-16

Image 6: Segment of report on improved model with significant variables

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 = 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)

Based on the available data, the best linear regression equation is:
Y = 303.46 – 149.36 * (If Customer_Segment: Loyalty Club Only) + 281.84 * (If Customer_Segment: Loyalty Club and Credit Card) - 245.42 * (If Customer_Segment: Store 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?

I recommend that the company sends the catalog to the 250 customers.

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

I came up with the recommendation based on the final predicted profit contribution of about $21,987.44 if the catalog is sent. This predicted profit contribution is twice as much as the minimum requirement of $10,000 for the decision. How did I arrive at this answer?

After getting the 'improved' model which used only significant predictor models, it was applied to the dataset of the new customers to get the expected revenue. A formula was applied to multiply the chances (probability) of a yes response from each new customer to the predicted average sale amount to give the predicted average revenue per customer i.e. Score_yes * score where Score_yes represents the chances of a yes response and score represents the predicted average sale amount. The gross margin of 0.5 was then multiplied by the predicted sale amounts, and the $6.50 cost deducted. The resulting expected individual profit contributions were then summed to give us the predicted total profit contribution from the new customers. The workflow is visualized below;

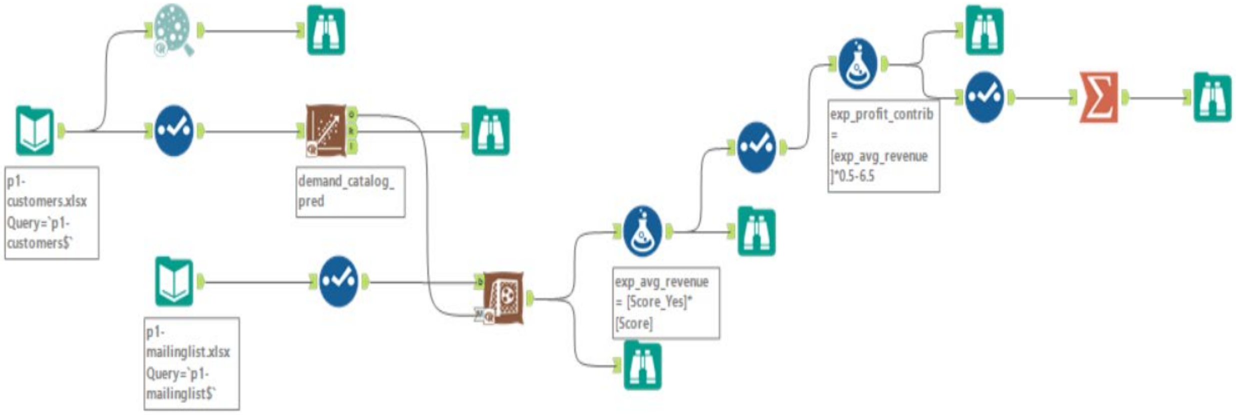
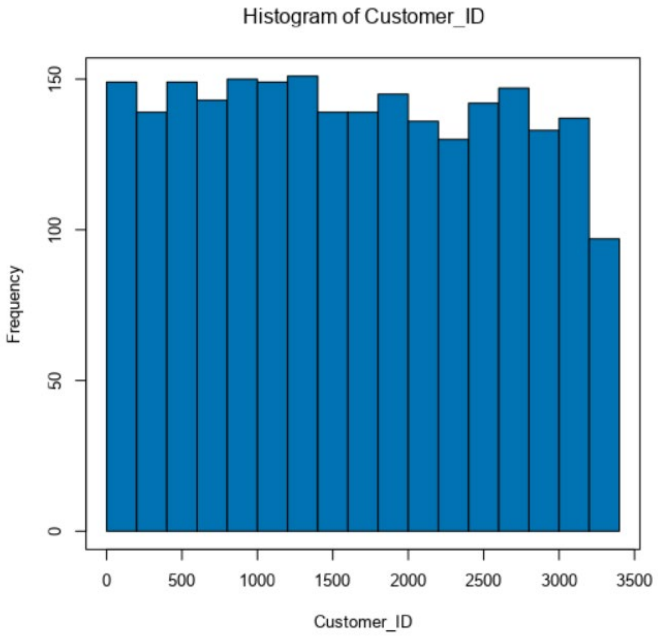

p1-customers.xlsx Query=`p1-customers$`
demand_catalog_pred
p1-mailinglist.xlsx Query=`p1-mailinglist$`
exp_avg_revenue = [Score_Yes]* [Score]
exp_profit_contrib = [exp_avg_revenue]*0.5-6.5

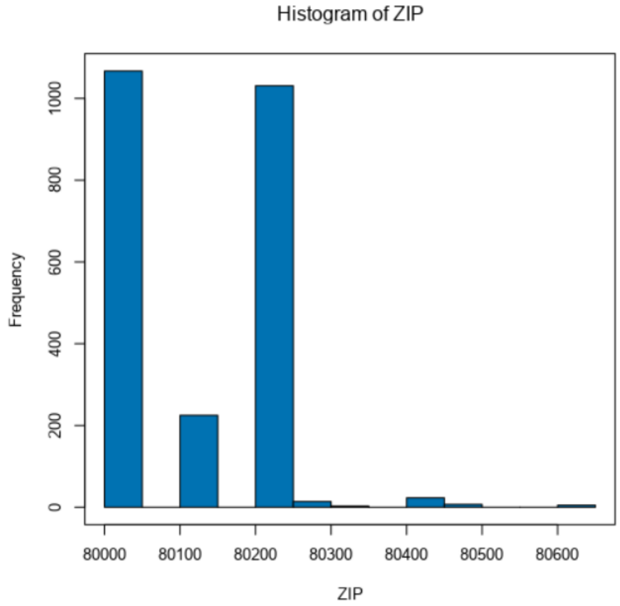
Image 7: Visualization of the Alteryx workflow to predict the profit contribution from the new customers.

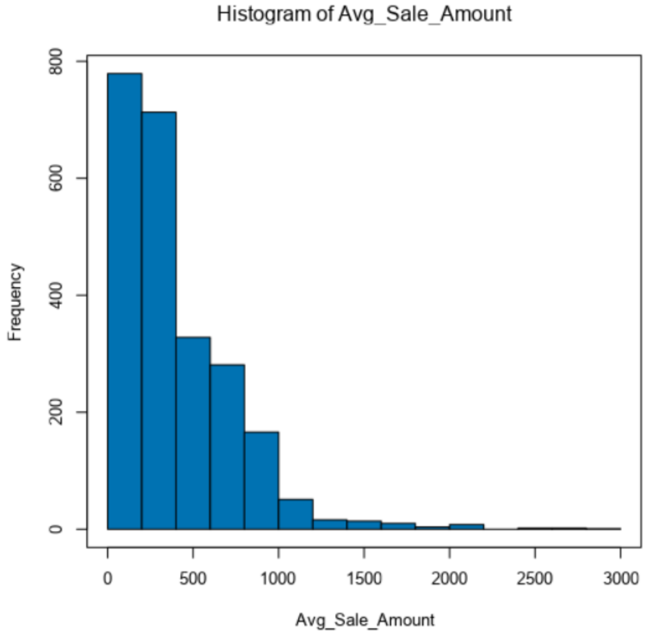
3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?

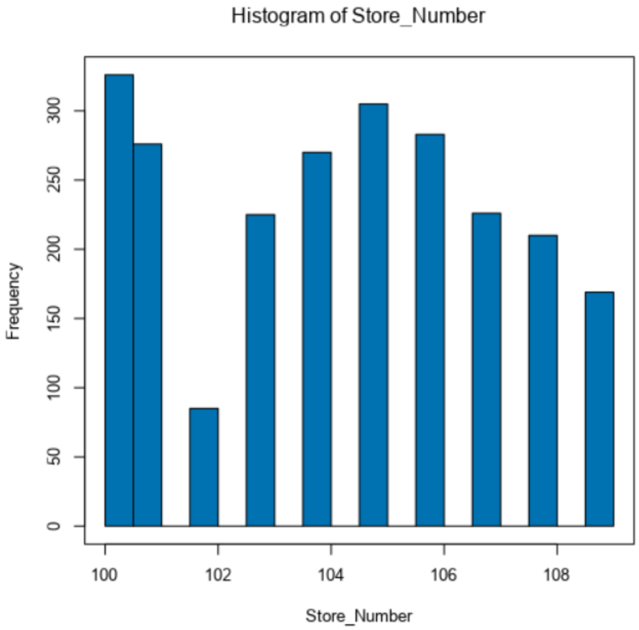
The expected profit from the new catalog if sent to the 250 customers will be $21,987.44.

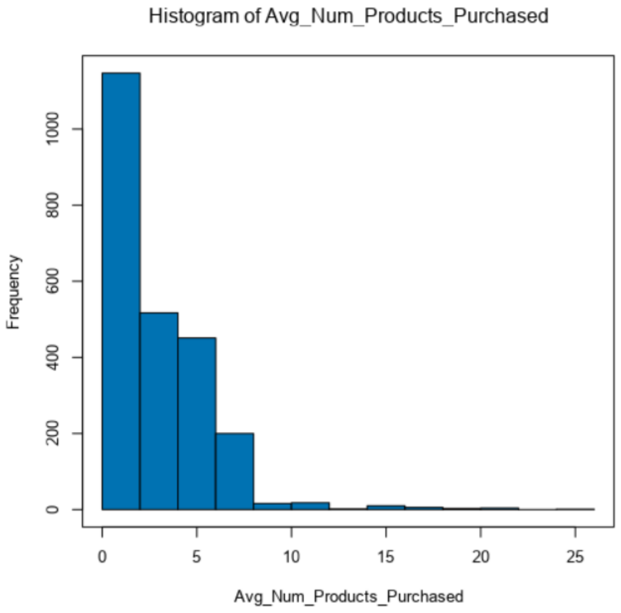
DISTRIBUTION OF VARIABLES IN CUSTOMER LIST

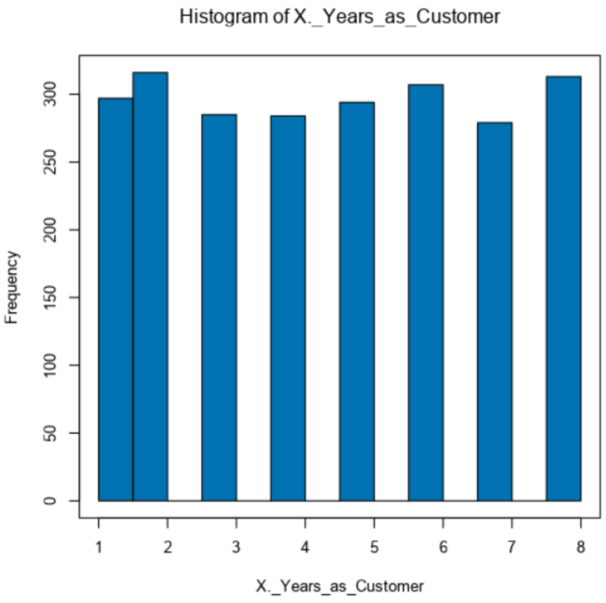

Histogram of Customer_ID
Frequency
0
50
100
150
0
500
1000
1500
2000
2500
3000
3500
Customer_ID


Histogram of ZIP
Frequency
0
200
400
600
800
1000
80000
80100
80200
80300
80400
80500
80600
ZIP


Histogram of Avg_Sale_Amount
Frequency
0
200
400
600
800
0
500
1000
1500
2000
2500
3000
Avg_Sale_Amount


Histogram of Store_Number
Frequency
0
50
100
150
200
250
300
100
102
104
106
108
Store_Number





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.