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 need to be made?

The decision that needed to be made was should the company send the catalog to the new customers and if so how much can it expect to earn from sending it to them.
The decision would be taken regarding the profit the company would make, if the profit prediction was less than 10,000$, the manager wouldn't send the catalogs.

2. What data is needed to inform those decisions?

The data needed to inform those decisions is taken from 2 data sets.
The first dataset "p1-customers" includes records of previous customers(2300), their addresses, their average purchases, the average number of items bought and various other details.
We will use this dataset to build our model, after choosing the appropriate predictors.
The second dataset "p1-mailing list" contains 250 new customers that we will need to predict their purchases from the catalog.
Using the model built from the first dataset we will apply it to the second dataset to get the predicted profit the company would make if it sent the catalog to the new clients.

Step 2: Analysis, Modeling, and Validation

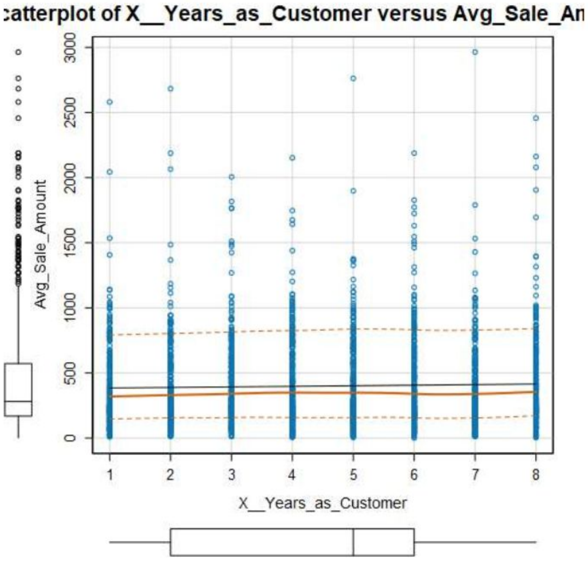
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

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

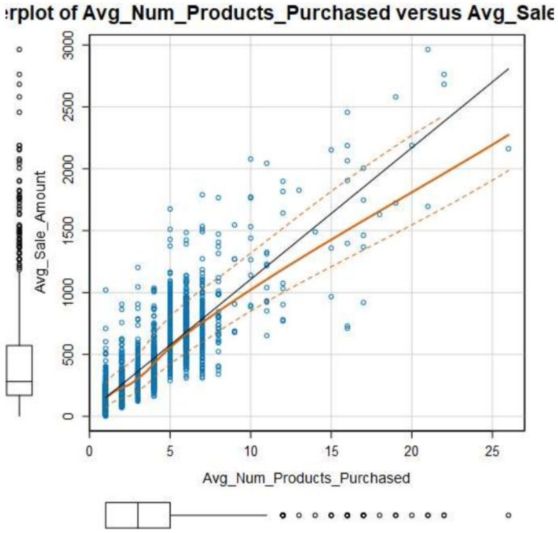
At the minimum, answer these questions:

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.

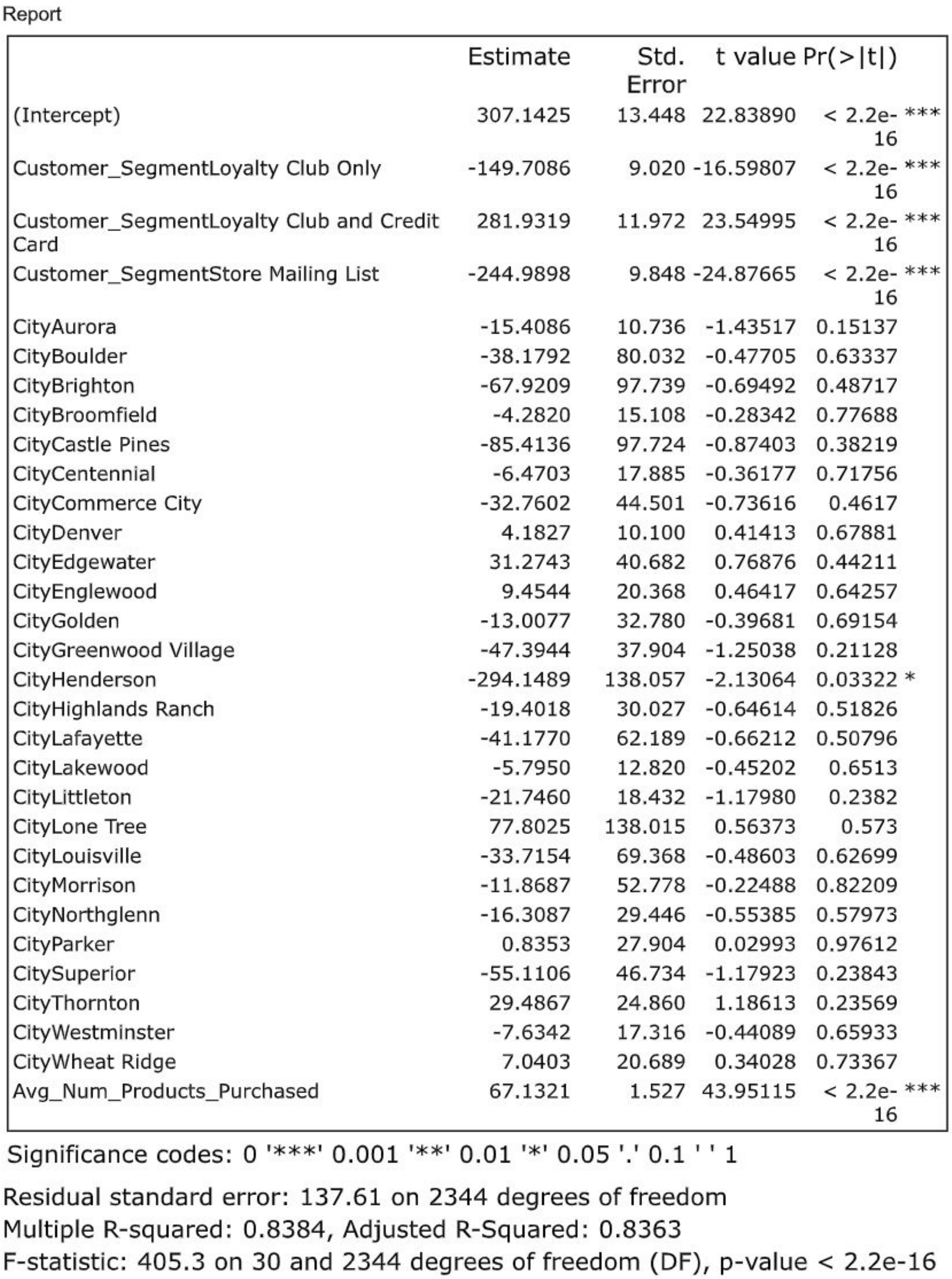
For numeric variables, I used scatterplots between the targeted variable (average_sales_amount) and an individual variable to see if the individual variable would be a good predictor variable.
When applied to the variable (years_as_a_consumer) the following scatterplots suggests that there is no correlation between it and the (average_sales_amount) and this is shown by the absence of slope.
The scatterplot is shown in the figure below:



A similar result was obtained with customer_id, zip & store_number.
I concluded that those numerical variables were not suitable for the prediction model.
On the other hand, the scatterplot between the targeted variable (average_sales_amount) and (average_num_product_purchased) shows an increasing slope which means that this variable would be a good predictor for the model, as shown in the figure below :



For categorical variables, I had to try a combination of categorical variables. I assumed at the beginning of this project that the names and address of the client won't be deterministic in the prediction model. I was then left with the city and customer_segment. The results of the R outputs of the linear regression are shown below:



Report

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	307.1425	13.448	22.83890	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-149.7086	9.020	-16.59807	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	281.9319	11.972	23.54995	< 2.2e-16	***
Customer_SegmentStore Mailing List	-244.9898	9.848	-24.87665	< 2.2e-16	***
CityAurora	-15.4086	10.736	-1.43517	0.15137	
CityBoulder	-38.1792	80.032	-0.47705	0.63337	
CityBrighton	-67.9209	97.739	-0.69492	0.48717	
CityBroomfield	-4.2820	15.108	-0.28342	0.77688	
CityCastle Pines	-85.4136	97.724	-0.87403	0.38219	
CityCentennial	-6.4703	17.885	-0.36177	0.71756	
CityCommerce City	-32.7602	44.501	-0.73616	0.4617	
CityDenver	4.1827	10.100	0.41413	0.67881	
CityEdgewater	31.2743	40.682	0.76876	0.44211	
CityEnglewood	9.4544	20.368	0.46417	0.64257	
CityGolden	-13.0077	32.780	-0.39681	0.69154	
CityGreenwood Village	-47.3944	37.904	-1.25038	0.21128	
CityHenderson	-294.1489	138.057	-2.13064	0.03322	*
CityHighlands Ranch	-19.4018	30.027	-0.64614	0.51826	
CityLafayette	-41.1770	62.189	-0.66212	0.50796	
CityLakewood	-5.7950	12.820	-0.45202	0.6513	
CityLittleton	-21.7460	18.432	-1.17980	0.2382	
CityLone Tree	77.8025	138.015	0.56373	0.573	
CityLouisville	-33.7154	69.368	-0.48603	0.62699	
CityMorrison	-11.8687	52.778	-0.22488	0.82209	
CityNorthglenn	-16.3087	29.446	-0.55385	0.57973	
CityParker	0.8353	27.904	0.02993	0.97612	
CitySuperior	-55.1106	46.734	-1.17923	0.23843	
CityThornton	29.4867	24.860	1.18613	0.23569	
CityWestminster	-7.6342	17.316	-0.44089	0.65933	
CityWheat Ridge	7.0403	20.689	0.34028	0.73367	
Avg_Num_Products_Purchased	67.1321	1.527	43.95115	< 2.2e-16	***

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

Residual standard error: 137.61 on 2344 degrees of freedom
Multiple R-squared: 0.8384, Adjusted R-Squared: 0.8363
F-statistic: 405.3 on 30 and 2344 degrees of freedom (DF), p-value < 2.2e-16

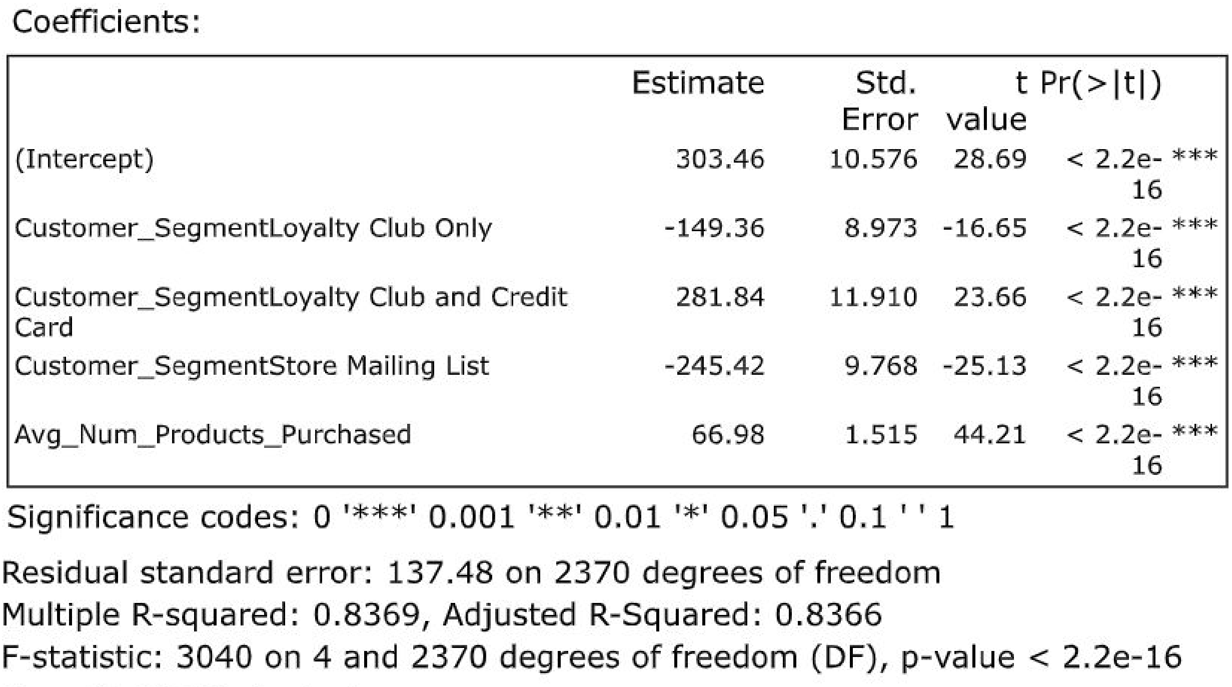
It is clear from the P-values of the city categories are relatively high (compared to 0.05) which means that this variable won't be a good predictor for the prediction model. The absence of asterisks * shows that the variable is statistically significant.

Whereas for the customer_segment @ avg_num_product_purchased, we can clearly see from the P-values and the number of asterisks that these variables are highly significant to the predictive model.

The chosen predictors are Customer_Segment & Avg_Num_Product_Purchased

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 linear model is believed to be a good model because of the P-values and R squared values.
Low P-values and high R squared values suggest that the model is highly predictive.
The values are shown in the figure below :



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

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)

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.

The regression equation is :
Y = 303.46 + 66.98 * (Average_Num_Products_Purchased) + 281.84*(if : Loyality&CreditCard) - 149.36*(if: LoyalityOnly) - 245.42*(if: MailingList)

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?
The company should send the catalog to these 250 customers because the profit gained from sending it will be higher than the 10,000$ benchmark given.

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 to this recommendation by applying the prediction model, built from the first dataset, on the second dataset which helped us make an average revenue prediction for each client. I then multiplied the average_revenue_prediction of each client by his/her yes_score which gave us a more accurate average_revenue_prediction.
I then summed all the predicted revenue of each client, multiplied the sum by 0.5 to get the gross margin and deducted the catalog printing prices which are (6.5*250).

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 would be: 21,987.44$