Project #1
Predicting Catalogue Demand

Claudia Dai

22th September, 2019

1 Business and data understanding

The manager needs to know if the company can expect a profit contribution of more than $10,000 if the catalogues are sent out to 250 new customers. Based on the expected profit, the manager has to decide if they should send out the catalogues. Corporate management has communicated that only if the expected profit is more than $10,000 the catalogs should be sent – if it is less, management does not want the catalogs to be sent to the new customers.

To inform the manager's decision, we need to predict the expected profit from the new 250 customers. For this, we are given data about 2300 customers with information on past sales, shopping behaviour, their location, etc. We are also given the cost of catalog print and distribution and the average gross margin of 50%. We have a numeric, data rich problem at hand, so we can use a linear regression to predict the outcome.

2 Analysis, Modelling and Validation

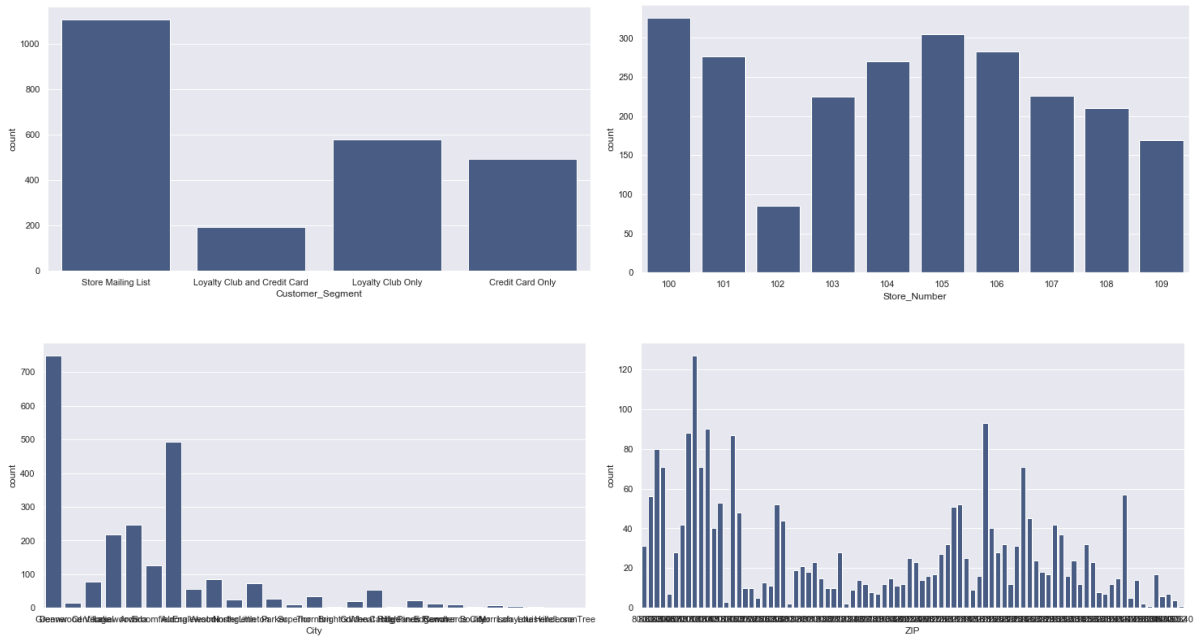
We already know from common sense that a few features are not informative for our model:

- `Name`: is an identifier
- `Customer_ID`: is an identifier
- `Address`: is an identifier
- `State`: only takes "CO" as value
- `Responded_to_Last_Catalog`: is not in the other dataset, hence we cannot use it. I suspect this column may have been used by the data science team to predict the probability scores of whether or not a new customer will respond to the catalog, which we will use later on to calculate the expected profit.

This leaves us with the following features to predict `Avg_Sale_Amount`.

- `Customer_Segment` (category)
- `City` (category)
- `ZIP` (category)
- `Store_Number` (category)
- `Avg_Num_Products_Purchased` (numeric)
- `#_Years_as_Customer` (numeric)

Let's first take a look at the distribution of the categorical features.

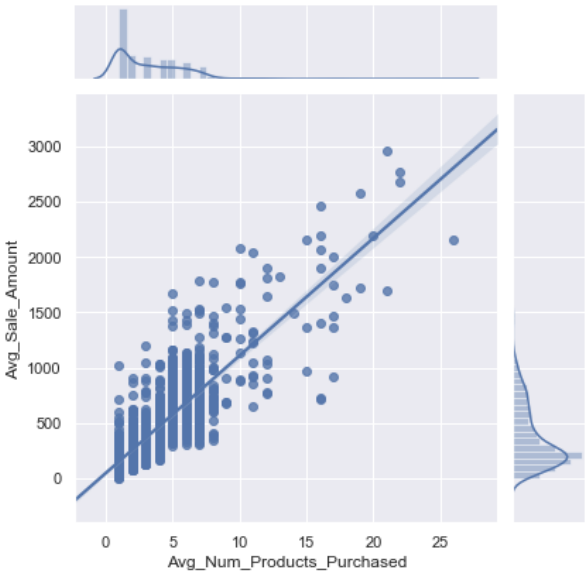


There seems to be unequal distributions in all four categorical variables. Most customers seem to fall in the store mailing list segment, and very few are part of the loyalty club and credit card. This makes common sense. Looking at the store, city and ZIP, we can see that these location variables are also imbalanced. Common sense may infer that some stores may have certain other features that make a visit more popular to the customers.

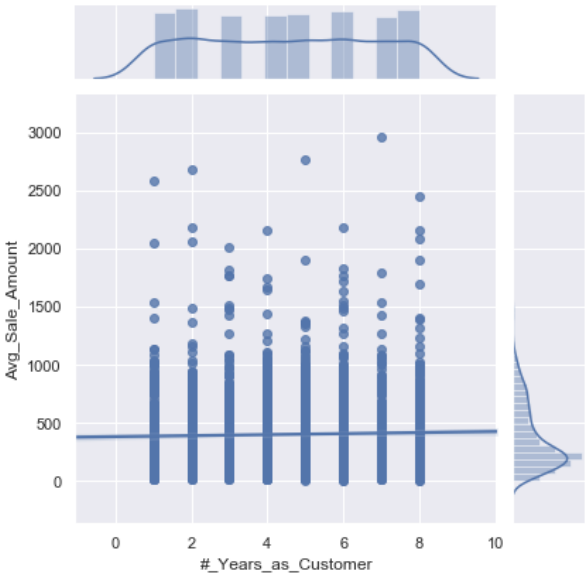
There are a couple of ideas that could enrich our currently given data and resulting model. For example, we could retrieve information about the city populations, seasonal data, weather data, the types of items purchased, or other information that may have an influence on customer shopping behaviour and spend.

Let's check the scatterplots between the target variables and numeric features to see if they have some sort of linear relationship with another.

`Avg_Sale_Amount vs. Avg_Num_Products_Purchased:` There seems to be a linear relationship between these two variables.



`Avg_Sale_Amount vs. #_Years_as_Customer:` There seems to be no linear relationship between the two variables.



After running a multiple linear regression for average sales amount as a function of average number of products purchased, number of years as customer, customer segment, city, zip, and store number, the results of the ordinary least squares (OLS) regression (see Appendix A) was analyzed on the p-values of the regression coefficients for statistical significance. The following variables were identified to be informative for the model:

- `Customer_Segment:` $p = 0.000$ for all three segment categories, and
- `Avg_Num_Products_Purchased:` $p = 0.000$

The other variables have a p-value > 0.05. The multiple linear regressions was adjusted with the three variables above, and yielded an adjusted R-squared of 0.837 (see Appendix B), indicating

that 83.70% of the average sale amount can be explained by the three chosen predictor variables, which is quite a strong model.

The linear regression equation based on the model with the given data is:

```
Avg_Sale_Amount = 303.46
                  - 149.36 (Loyalty Club Only)
                  + 281.84 (Loyalty Club and Credit Card)
                  - 245.42 (Store Mailing List)
                  + 66.98 (Avg_Num_Products_Purchased)
```

3 Presentation/Visualization

In conclusion, the company should send the catalog to the 250 new customers. The expected profit from each customer is calculated by multiplying the expected sales amount with the probability that they will reply to the catalog (indicated in the column `Score_Yes`). We need to substract the cost of printing and distributing the catalogs to 250 customers from the expected profit.

```
Expected Profit = (Sum Expected Sales * Gross Margin) - (Catalog Cost * 250)
                = (47224.87 * 0.5) — (6.50 * 250)
                = 21987.44
```

The expected profit is $**21,987.44**. Because this is more than $10,000, the manager should decide to send out the catalogs.

Appendix A: OLS Regression Results

OLS Regression Results

Dep. Variable:	Avg_Sale_Amount	**R-squared:**	0.844
Model:	OLS	**Adj. R-squared:**	0.836
Method:	Least Squares	**F-statistic:**	104.3
Date:	Sat, 22 Sep 2018	**Prob (F-statistic):**	0.00
Time:	15:43:44	**Log-Likelihood:**	-15009.
No. Observations:	2375	**AIC:**	3.025e+04
Df Residuals:	2257	**BIC:**	3.093e+04
Df Model:	117		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	349.1013	31.347	11.137	0.000	287.628	410.574
Customer_Segment[T.Loyalty Club Only]	-147.6879	9.201	-16.052	0.000	-165.731	-129.645
Customer_Segment[T.Loyalty Club and Credit Card]	288.3455	12.286	23.470	0.000	264.253	312.438
Customer_Segment[T.Store Mailing List]	-243.2747	10.022	-24.275	0.000	-262.927	-223.622
City[T.Aurora]	-21.9297	212.766	-0.103	0.918	-439.168	395.309
City[T.Boulder]	-65.0334	59.603	-1.091	0.275	-181.916	51.849
City[T.Brighton]	1.8711	217.025	0.009	0.993	-423.718	427.460
City[T.Broomfield]	-38.0776	195.380	-0.195	0.845	-421.221	345.066
City[T.Castle Pines]	-68.3883	52.024	-1.315	0.189	-170.408	33.632
City[T.Centennial]	-3.6113	214.465	-0.017	0.987	-424.181	416.959
City[T.Commerce City]	-28.4660	25.611	-1.111	0.266	-78.689	21.757
City[T.Denver]	2.6701	161.421	0.017	0.987	-313.878	319.218
City[T.Edgewater]	38.8112	164.865	0.235	0.814	-284.492	362.115
City[T.Englewood]	1.9278	221.982	0.009	0.993	-433.383	437.238
City[T.Golden]	26.2311	119.498	0.220	0.826	-208.107	260.569
City[T.Greenwood Village]	-104.9781	224.310	-0.468	0.640	-544.853	334.897
City[T.Henderson]	-145.7189	70.357	-2.071	0.038	-283.690	-7.747

City[T.Highlands Ranch]	24.0451	297.505	0.081	0.936	-559.367	607.458
City[T.Lafayette]	-39.6286	33.558	-1.181	0.238	-105.436	26.179
City[T.Lakewood]	28.1064	158.299	0.178	0.859	-282.319	338.532
City[T.Littleton]	52.6882	259.271	0.203	0.839	-455.746	561.123
City[T.Lone Tree]	158.1580	325.292	0.486	0.627	-479.746	796.062
City[T.Louisville]	-15.8274	49.357	-0.321	0.748	-112.616	80.962
City[T.Morrison]	-18.3509	29.468	-0.623	0.534	-76.138	39.436
City[T.Northglenn]	-41.4181	168.389	-0.246	0.806	-371.631	288.795
City[T.Parker]	-33.4534	28.998	-1.154	0.249	-90.320	23.413
City[T.Superior]	-38.2602	39.298	-0.974	0.330	-115.324	38.804
City[T.Thornton]	65.6241	167.497	0.392	0.695	-262.840	394.088
City[T.Westminster]	-26.9809	192.295	-0.140	0.888	-404.074	350.112
City[T.Wheat Ridge]	-88.0394	166.051	-0.530	0.596	-413.669	237.590
ZIP[T.80003]	-33.3113	31.204	-1.068	0.286	-94.503	27.880
ZIP[T.80004]	-25.3720	29.276	-0.867	0.386	-82.783	32.038
ZIP[T.80005]	-19.8075	29.953	-0.661	0.509	-78.547	38.932
ZIP[T.80007]	-19.5822	57.723	-0.339	0.734	-132.778	93.614
ZIP[T.80010]	5.4398	216.470	0.025	0.980	-419.062	429.941
ZIP[T.80011]	-2.6647	215.998	-0.012	0.990	-426.240	420.911
ZIP[T.80012]	-27.6771	214.446	-0.129	0.897	-448.209	392.855
ZIP[T.80013]	-26.3521	215.281	-0.122	0.903	-448.522	395.818
ZIP[T.80014]	6.7528	215.565	0.031	0.975	-415.974	429.479
ZIP[T.80015]	-7.7615	215.889	-0.036	0.971	-431.123	415.600
ZIP[T.80016]	-47.7890	216.432	-0.221	0.825	-472.215	376.637
ZIP[T.80017]	-27.6022	215.751	-0.128	0.898	-450.692	395.488
ZIP[T.80018]	-174.6915	229.424	-0.761	0.446	-624.596	275.213
ZIP[T.80020]	8.4416	196.454	0.043	0.966	-376.807	393.691
ZIP[T.80021]	-20.3470	197.939	-0.103	0.918	-408.508	367.814
ZIP[T.80022]	-28.4660	25.611	-1.111	0.266	-78.689	21.757
ZIP[T.80023]	6.3890	200.779	0.032	0.975	-387.341	400.119

	ZIP[T.80026]	-39.6286	33.558	-1.181	0.238	-105.436	26.179
	ZIP[T.80027]	-54.0876	33.040	-1.637	0.102	-118.880	10.705
	ZIP[T.80030]	32.3900	198.253	0.163	0.870	-356.388	421.168
	ZIP[T.80031]	-26.8036	194.791	-0.138	0.891	-408.792	355.185
	ZIP[T.80033]	102.8804	169.179	0.608	0.543	-228.882	434.643
	ZIP[T.80108]	-68.3883	52.024	-1.315	0.189	-170.408	33.632
	ZIP[T.80110]	-16.8022	225.950	-0.074	0.941	-459.895	426.290
	ZIP[T.80111]	38.9614	224.792	0.173	0.862	-401.860	479.783
	ZIP[T.80112]	-64.1124	219.658	-0.292	0.770	-494.866	366.641
	ZIP[T.80113]	-25.1019	225.557	-0.111	0.911	-467.422	417.218
	ZIP[T.80120]	-120.2460	263.060	-0.457	0.648	-636.112	395.620
	ZIP[T.80121]	-9.9891	221.141	-0.045	0.964	-443.650	423.672
	ZIP[T.80122]	-86.4103	221.129	-0.391	0.696	-520.048	347.227
	ZIP[T.80123]	-83.6004	261.905	-0.319	0.750	-597.200	429.999
	ZIP[T.80124]	-95.8104	296.448	-0.323	0.747	-677.150	485.529
	ZIP[T.80126]	-6.0866	294.785	-0.021	0.984	-584.164	571.991
	ZIP[T.80127]	-106.8188	263.188	-0.406	0.685	-622.935	409.297
	ZIP[T.80128]	-137.3501	263.628	-0.521	0.602	-654.328	379.628
	ZIP[T.80129]	-96.5670	302.699	-0.319	0.750	-690.164	497.030
	ZIP[T.80130]	-128.5427	303.318	-0.424	0.672	-723.353	466.268
	ZIP[T.80134]	-52.8200	31.084	-1.699	0.089	-113.776	8.136
	ZIP[T.80138]	19.3666	29.744	0.651	0.515	-38.962	77.696
	ZIP[T.80202]	-24.7261	168.722	-0.147	0.884	-355.593	306.141
	ZIP[T.80203]	-16.2288	168.412	-0.096	0.923	-346.488	314.030
	ZIP[T.80204]	-43.6306	165.764	-0.263	0.792	-368.697	281.436
	ZIP[T.80205]	5.6856	165.981	0.034	0.973	-319.805	331.176
	ZIP[T.80206]	-33.7001	167.744	-0.201	0.841	-362.649	295.249
	ZIP[T.80207]	-37.2339	167.273	-0.223	0.824	-365.259	290.791
	ZIP[T.80209]	-57.3508	167.166	-0.343	0.732	-385.165	270.463
	ZIP[T.80210]	-56.2185	165.811	-0.339	0.735	-381.377	268.940

	ZIP[T.80211]	11.7432	165.209	0.071	0.943	-312.235	335.721
	ZIP[T.80212]	-9.3262	164.394	-0.057	0.955	-331.706	313.053
	ZIP[T.80214]	-21.3565	161.962	-0.132	0.895	-338.967	296.254
	ZIP[T.80215]	-91.4767	162.762	-0.562	0.574	-410.655	227.702
	ZIP[T.80216]	23.9784	169.688	0.141	0.888	-308.782	356.739
	ZIP[T.80218]	13.1012	167.317	0.078	0.938	-315.010	341.213
	ZIP[T.80219]	-12.9848	164.303	-0.079	0.937	-335.185	309.215
	ZIP[T.80220]	-0.4987	165.163	-0.003	0.998	-324.386	323.389
	ZIP[T.80221]	-11.5064	165.444	-0.070	0.945	-335.945	312.932
	ZIP[T.80222]	-31.5209	165.700	-0.190	0.849	-356.460	293.419
	ZIP[T.80223]	-39.0041	168.451	-0.232	0.817	-369.339	291.331
	ZIP[T.80224]	5.3267	165.783	0.032	0.974	-319.776	330.429
	ZIP[T.80226]	-52.7760	161.340	-0.327	0.744	-369.167	263.614
	ZIP[T.80227]	-21.4207	162.743	-0.132	0.895	-340.563	297.721
	ZIP[T.80228]	-76.0817	163.004	-0.467	0.641	-395.735	243.572
	ZIP[T.80229]	-57.0905	168.061	-0.340	0.734	-386.660	272.479
	ZIP[T.80230]	-12.7153	167.279	-0.076	0.939	-340.753	315.322
	ZIP[T.80231]	-28.2466	165.274	-0.171	0.864	-352.352	295.859
	ZIP[T.80232]	-59.6730	162.062	-0.368	0.713	-377.478	258.132
	ZIP[T.80233]	-4.4364	172.102	-0.026	0.979	-341.931	333.059
	ZIP[T.80234]	4.3975	167.748	0.026	0.979	-324.560	333.355
	ZIP[T.80235]	-33.0885	167.175	-0.198	0.843	-360.921	294.744
	ZIP[T.80236]	-78.4329	165.628	-0.474	0.636	-403.232	246.366
	ZIP[T.80237]	-41.4873	166.508	-0.249	0.803	-368.012	285.037
	ZIP[T.80238]	-15.4498	170.825	-0.090	0.928	-350.440	319.540
	ZIP[T.80239]	-99.8228	172.170	-0.580	0.562	-437.450	237.804
	ZIP[T.80241]	-54.7830	174.141	-0.315	0.753	-396.277	286.711
	ZIP[T.80246]	-54.3111	167.773	-0.324	0.746	-383.318	274.695
	ZIP[T.80247]	-14.9840	164.964	-0.091	0.928	-338.481	308.513
	ZIP[T.80249]	-22.4718	175.293	-0.128	0.898	-366.224	321.281

		ZIP[T.80260]	-122.7280	169.774	-0.723	0.470	-455.657	210.201
		ZIP[T.80303]	-0.4804	80.235	-0.006	0.995	-157.822	156.861
		ZIP[T.80305]	-64.5531	98.098	-0.658	0.511	-256.925	127.819
		ZIP[T.80401]	-66.3482	127.250	-0.521	0.602	-315.886	183.190
		ZIP[T.80403]	-13.8402	100.749	-0.137	0.891	-211.411	183.730
		ZIP[T.80465]	-18.3509	29.468	-0.623	0.534	-76.138	39.436
		ZIP[T.80602]	-109.7946	195.639	-0.561	0.575	-493.446	273.857
		ZIP[T.80640]	-145.7189	70.357	-2.071	0.038	-283.690	-7.747
		Store_Number[T.101]	0.7083	14.549	0.049	0.961	-27.823	29.240
		Store_Number[T.102]	20.0352	23.317	0.859	0.390	-25.690	65.761
		Store_Number[T.103]	0.8548	18.536	0.046	0.963	-35.495	37.205
		Store_Number[T.104]	-11.7545	14.538	-0.809	0.419	-40.265	16.756
		Store_Number[T.105]	-13.2166	12.506	-1.057	0.291	-37.741	11.307
		Store_Number[T.106]	-19.2877	14.728	-1.310	0.190	-48.169	9.594
		Store_Number[T.107]	-24.6093	16.047	-1.534	0.125	-56.077	6.858
		Store_Number[T.108]	-16.1201	17.853	-0.903	0.367	-51.131	18.890
		Store_Number[T.109]	9.3323	21.408	0.436	0.663	-32.648	51.313
		Avg_Num_Products_Purchased	67.1945	1.557	43.144	0.000	64.140	70.249
		Num_Years_as_Customer	-2.5310	1.264	-2.002	0.045	-5.010	-0.052

Omnibus:	345.264	**Durbin-Watson:**	2.028
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4523.256
Skew:	0.177	**Prob(JB):**	0.00
Kurtosis:	9.752	**Cond. No.**	9.96e+15

Appendix B: OLS Regression Results with selected variables

OLS Regression Results

Dep. Variable:	Avg_Sale_Amount	**R-squared:**	0.837
Model:	OLS	**Adj. R-squared:**	0.837
Method:	Least Squares	**F-statistic:**	3040.
Date:	Sun, 23 Sep 2018	**Prob (F-statistic):**	0.00
Time:	18:30:54	**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|	[0.025	0.975]
Intercept	303.4635	10.576	28.694	0.000	282.725	324.202
Customer_Segment[T.Loyalty Club Only]	-149.3557	8.973	-16.645	0.000	-166.951	-131.760
Customer_Segment[T.Loyalty Club and Credit Card]	281.8388	11.910	23.664	0.000	258.484	305.194
Customer_Segment[T.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