Group 4

Benjamin Loo
Jerald Seow
Jin Dong Yang
Leng Kenn Siang
Mark Jeremiah Robert

Contents

.

Our Motivation

In today's fast paced and tech savvy world, there has been an exponential increase in online shopping and e-commerce. It is projected that global business-to-consumer e-commerce sales will hit USD$1.92 trillion by 2016 (Statista, 2015). For its launch into India in 2013, Amazon has committed $2 billion to tap on India's growing middle class (Ali Jaafar, 2015). With rising e-commerce sales and a larger market to tap on, there has been greater emphasis on the efficiency and quality of online stores. As such, in order to improve the predictive ability of online websites, our project aims to identify significant correlations between product categories and certain demographic variables. By predicting and modelling the patterns between different categories of goods from the comScore database, we can provide e-Commerce platforms the ability to, with certain statistical accuracy, suggest categories of products to consumers registered in their database immediately. We can then repeat the implementation of this process onto the remaining product categories. This system removes the need for substantial search history regarding a particular consumer before suggesting additional products. For this project, our group has decided to focus on three distinct product categories (Apparels, Health and Beauty and Music) from the comScore dataset to study the relationship between the purchases for products from these categories.

Descriptive analysis of the dataset would provide us with the rough idea of the nature of each demographic variable that we are analysing and would likely reveal the correlations between those variables. This may give us further insights to eliminate certain variables from future regression/ mining. Predictive analytics through multiple linear regression and data mining would provide us with the main products for this project. We expect to draw correlations between significant demographic variables, and eventually draw conclusions on when and who to suggest Apparels, Health and Beauty and Music products to online shoppers.

Data Source

Our dataset is obtained from the 2004 comScore Disaggregate Dataset from the Wharton Research Data Services, which tracked the web-wide visitation and transaction behaviour of more than 1.5 million Internet users over the course of 12 months. The panelist-level database used in this report captures the detailed browsing and buying behaviour by over fifty thousand Internet users across the United States. This panel is based on a random sample from a cross-section of more than 1.5 million global Internet users who have given comScore explicit permission to confidentially capture their Web-wide activity. We expect this data to have been collected without any bias involved.

In our investigation into a method to suggest products to users before they even build up a browsing history, this dataset provides us with great insight on the historic data with details of the demographics of users who purchased products online. By studying the trends and patterns within this dataset for certain product categories, we can gain greater insight as to the types of demographics that purchase such products and their expenditure on these products. With such information, models can be generated to allow e-commerce websites to effectively recommend products to their users based off their demographics. In this study, for each of the product categories analysed, a random sample size of 1000 was taken from the main dataset.

Descriptive Analysis

Instead of focussing our efforts on running data analytics on every single product category (60 in total), the group decided to narrow our scope down to the Top 3 products in terms of transaction count. We discovered that the Top 3 product categories were - #1 Apparel, #16 Health & Beauty and #22 Music. The following PivotChart displays our findings.

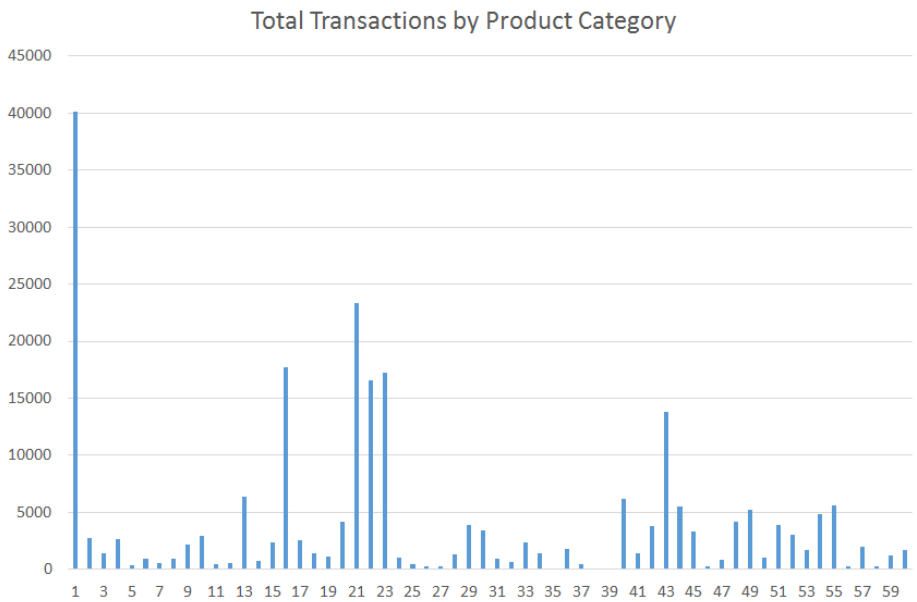


Figure 1: Number of transactions by product category

After identifying the three product categories, we then generated graphs to determine the transactions by month to see which month generates the highest amount of profit for the three product categories. From Figure 3,4 and 5, we concluded that the highest transactions within a month for all three product categories is in the month of December, This led us to focus on data collected within the month of December to perform our predictive analytics, as it provides the highest amount of profit.

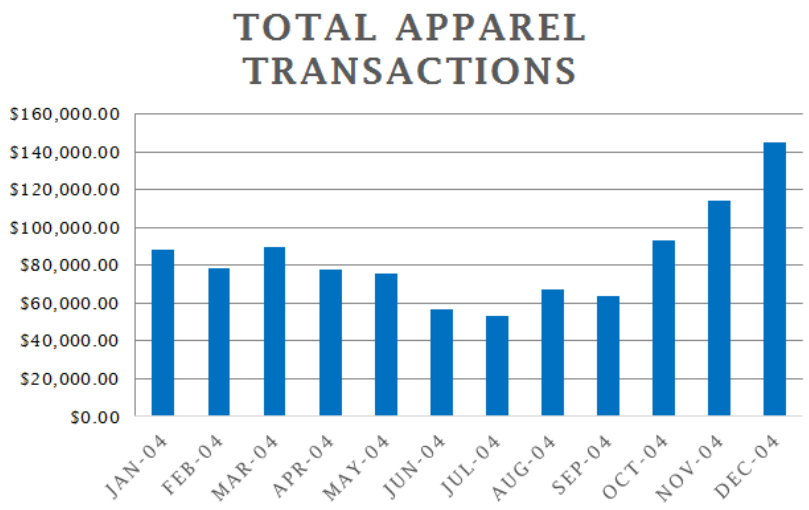


Figure 2: Total apparel transactions by month

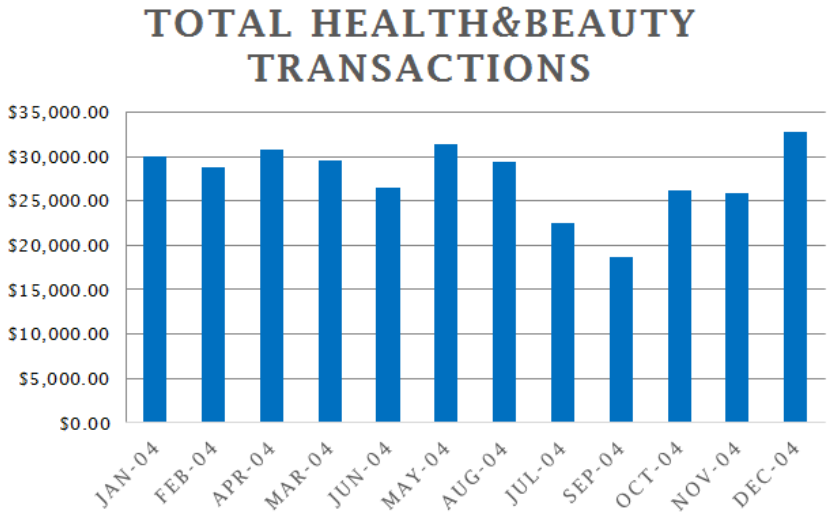


Figure 3: Total health & beauty transactions by month

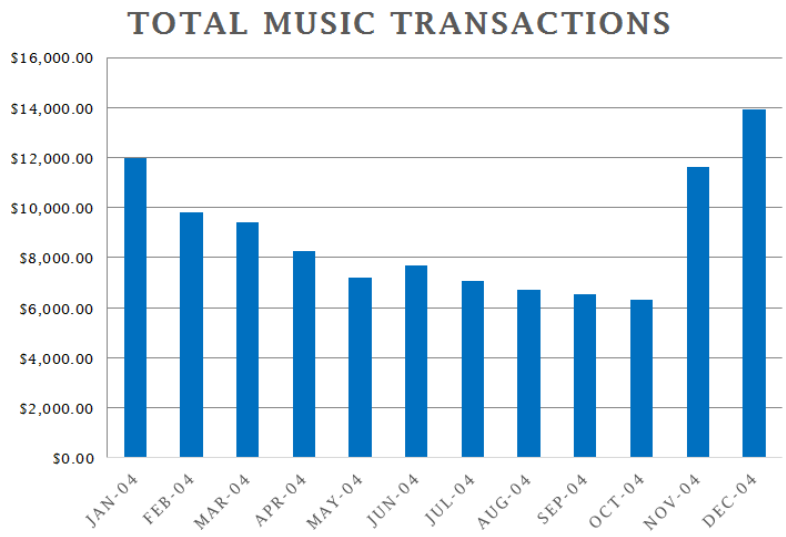


Figure 4: Total music transactions by month

Predictive Analysis (Regression)

Apparel Regression Model

	prod_qty	od_totpri	basket_tot	nost_educ	h_oldest	cnsus_regi	usehold_s	sehold_inc	children	al_backgr	nection_sp	ntry_of_or	nsusDumm	nsusDumm	nsusDumm	nsusDumm	Race1	Race2	Race3	Race5	Edu0	Edu1	Edu2	Edu3	Edu4	Edu5	Edu99
prod_qty	1																										
prod_totp	0.032094	1																									
basket_to	-0.03954	0.372058	1																								
hoh_most	-0.03859	-0.01913	0.039363	1																							
hoh_olde	0.039084	0.08798	0.014948	0.045959	1																						
census_re	0.04422	0.038813	-0.09866	-0.10439	-0.03643	1																					
househol	0.045511	-0.02173	0.099664	0.103259	-0.06989	0.037545	1																				
househol	-0.05971	0.109449	0.088865	0.037162	0.078405	0.006781	0.157534	1																			
children	-0.01592	-0.01316	0.089698	0.120045	-0.10783	0.058987	0.633915	0.220228	1																		
racial_bac	-0.03301	-0.03665	0.038973	0.041289	-0.05797	-0.01829	0.048251	-0.12264	-0.06716	1																	
connectio	-0.08097	0.045186	0.136486	0.055719	-0.10597	-0.0713	0.043701	0.190546	0.052531	0.00841	1																
country_o	-0.01797	-0.04531	-0.07889	0.053161	-0.02077	0.136103	0.070118	-0.01233	-0.00054	0.354511	-0.05842	1															
CensusDu	-0.0259	-0.05186	0.081091	0.122442	0.045276	-0.80522	-0.05289	0.023274	-0.05694	0.050305	0.115965	-0.08561	1														
CensusDu	-0.00578	0.014678	0.006996	-0.05561	-0.02934	-0.21705	0.049087	-0.02486	-0.00261	-0.00805	-0.0592	-0.06741	-0.31755	1													
CensusDu	-0.01571	0.03212	-0.01897	-0.01118	0.006425	0.401347	-0.02235	-0.03575	0.031932	-0.08456	-0.05984	0.054562	-0.434	-0.39225	1												
CensusDu	0.056855	0.003552	-0.07874	-0.06749	-0.02864	0.674688	0.03589	0.045954	0.028996	0.05775	0.006149	0.105934	-0.26636	-0.24073	-0.32901	1											
Race1	0.01532	0.021338	-0.10071	-0.02183	0.041237	0.073397	-0.03623	0.128227	0.067884	-0.81925	-0.00673	-0.19229	-0.07722	-0.02825	0.100988	-0.00671	1										
Race2	-0.06748	0.019916	0.146295	-0.01825	0.023447	-0.11238	-0.01871	-0.07976	-0.05394	0.21649	-0.00871	-0.07928	0.078236	0.04861	-0.05356	-0.07763	-0.70578	1									
Race3	0.002243	-0.04204	-0.03187	0.046129	-0.05458	0.026348	0.054352	0.06656	0.021158	0.193651	0.060626	-0.00196	0.009165	-0.00555	-0.07359	0.089249	-0.2647	-0.02693	1								
Race5	-0.00141	-0.03485	0.004312	0.036954	-0.057	0.007135	0.045171	-0.12363	-0.0579	0.935152	-0.00374	0.393768	0.026876	-0.01915	-0.05645	0.061762	-0.59139	-0.06017	-0.02257	1							
Edu0	0.134705	-0.04881	-0.00327	-0.11289	0.045827	0.00968	0.033131	-0.08747	0.013128	-3.4E-05	-0.0195	0.002248	-0.04596	0.090094	-0.05179	0.019635	-0.06192	0.114137	-0.01361	-0.0304	1						
Edu1	0.0144	0.053861	0.026014	-0.2827	0.042369	-0.02426	-0.01838	-0.06689	-0.10454	0.025149	-0.06981	0.047642	0.021448	0.01483	-0.03171	-0.00124	0.005903	-0.02881	-0.03497	0.042442	-0.04711	1					
Edu2	-0.01165	-0.02703	-0.07638	-0.41521	-0.04888	0.104709	-0.06963	-0.06169	-0.01367	-0.02188	-0.06073	-0.03075	-0.12831	0.039884	0.054122	0.037127	0.060164	-0.08234	0.020515	-0.00418	-0.07106	-0.18264	1				
Edu3	-0.0209	-0.01112	-0.03204	-0.23458	-0.01119	-0.04487	-0.04033	-0.05319	-0.02012	-0.03463	-0.03902	-0.08165	0.056406	-0.0513	0.032611	-0.05064	-0.04635	0.13345	-0.03063	-0.06844	-0.04126	-0.10605	-0.15998	1			
Edu4	0.007644	0.025398	0.045282	-0.30564	-0.0785	0.074138	-0.00477	0.082029	-0.03953	-0.0088	0.065363	-0.02002	-0.0663	0.007624	0.00655	0.060941	0.025979	-0.02949	-0.01235	0.001817	-0.0553	-0.14214	-0.21441	-0.1245	1		
Edu5	-0.03497	0.022485	-0.00493	-0.17023	0.047908	0.003864	-0.03996	0.137337	-0.01513	-0.02246	0.054999	0.00714	0.004929	0.015191	-0.04864	0.039321	-0.00608	0.045324	-0.02354	-0.03152	-0.03171	-0.08149	-0.12293	-0.07138	-0.09567	1	
Edu99	-0.03825	-0.01952	0.038956	0.999751	0.046977	-0.1047	0.103389	0.033443	0.119479	0.042007	0.053371	0.053878	0.122214	-0.05488	-0.01114	-0.06808	-0.02175	-0.01864	0.046246	0.037684	-0.10532	-0.27068	-0.40832	-0.2371	-0.31776	-0.18219	1

Figure 5: Correlation Table of Apparel Variables

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	14.16256	40.05308	0.353595	0.723658	-64.3592	92.6843312	4237835
hoh_oldest_age	0.466832	0.235202	1.984812	0.04722	0.005732	0.92793247	7548.481
household_size	-2.11265	0.587338	-3.597	0.000325	-3.2641	-0.9612104	153.1786
household_income	2.288019	0.366526	6.242452	4.67E-10	1.569467	3.00657217	73716.92
children	4.97837	1.522073	3.270783	0.00108	1.994433	7.96230761	13902.3
connection_speed	4.323822	1.193199	3.623724	0.000293	1.984624	6.66301945	15503.49
country_of_origin	-5.14709	1.933826	-2.66161	0.007802	-8.93825	-1.355939	7289.914
CensusDummy1	68.88134	48.95296	1.407092	0.159463	-27.0881	164.850824	4882.023
CensusDummy2	73.61448	48.96522	1.503403	0.132799	-22.379	169.607996	4440.15
CensusDummy3	70.65987	48.97088	1.442895	0.149113	-25.3447	166.664484	3951.483
CensusDummy4	73.58595	48.98186	1.50231	0.133081	-22.4402	169.612085	4.903936
Race1	-76.7942	28.52951	-2.69175	0.007132	-132.725	-20.863751	8829.752
Race2	-71.3058	28.5615	-2.49657	0.012572	-127.299	-15.312604	321.9377
Race3	-83.7023	28.94001	-2.89227	0.003841	-140.438	-26.967098	10024.58
Race5	-71.1268	28.64676	-2.48289	0.013065	-127.287	-14.966482	7226.22
Edu0	0	0	N/A	N/A	0	0	0
Edu1	19.1818	4.613373	4.157869	3.27E-05	10.13755	28.2260575	58707.26
Edu2	6.397474	4.448574	1.438096	0.15047	-2.3237	15.1186481	3039.764
Edu3	8.503232	4.702351	1.808294	0.070621	-0.71546	17.7219223	10.60571
Edu4	8.826411	4.525894	1.950203	0.051208	-0.04635	17.6991672	215.1733
Edu5	7.652041	4.736063	1.615697	0.106224	-1.63274	16.9368212	121.5542
Edu99	8.722676	4.378249	1.992275	0.046396	0.139369	17.3059823	6264.724

Figure 6: Regression model from Apparel data with all demographic variables

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	24.71871	2.551188	9.689097	5.25E-22	19.71725	29.7201643	4237835
household_size	-2.01846	0.5852	-3.44919	0.000567	-3.16571	-0.8712125	329.3454
household_income	2.384868	0.356087	6.697434	2.36E-11	1.686781	3.08295568	77728.22
children	4.764595	1.512649	3.149836	0.001643	1.799135	7.73005463	12663.94
connection_speed	3.833773	1.173373	3.267309	0.001093	1.533443	6.13410285	13466.71
country_of_origin	-4.96588	1.807852	-2.74684	0.006039	-8.51006	-1.4216894	7634.85
Race1	-5.36549	1.787858	-3.00107	0.002704	-8.87048	-1.8604947	7697.182
Race3	-11.2604	5.042413	-2.23314	0.025584	-21.1458	-1.3750497	8619.314
Edu1	11.42211	1.842983	6.197622	6.2E-10	7.809053	15.035174	60766.81
Adjusted R^2	0.021899						

Figure 7: Regression model derived from Apparel data

*prod_totprice = -2.02*household_size + 2.38*household_income + 4.76*children + 3.83*connection_speed - 4.97*country_of_origin - 5.47*Race1 - 11.26*Race3 + 11.42*Edu1*

From the results of our Multiple Linear Regression Analysis, it was observed that there were several significant variables that could be included into our regression model which were not autocorrelated nor collinear (shown through the correlation table). However, in the spirit of parsimony, we narrowed our regression model to 8 independent variables as highlighted in Figure 7 above.

Household income and having children were statistically significant variables unsurprisingly, bearing positive coefficients which suggested that having a higher household income, having more children or having faster internet led to an in increase in expenditure on clothing. What was interesting was the negative coefficient of our household_size variable, which suggested that having a bigger household though statistically significant, led to decreased expenditure on apparel. A possible explanation for this negative coefficient could be one of two scenarios: 1.) that a larger household size suggests that the buyer is staying with his or her parents / children, and thus, clothes are bought for the user by someone else and hence their expenditure on apparel could be lower, or 2.) that a smaller household size suggests an independent user living on their own, thus being more likely to purchase their apparel online than offline as compared to those staying in a larger household. It is also surprising that the model suggests that being hispanic results in lesser expenditure on clothing, though literature suggests that clothing is the most popular retail item purchased by hispanics (89%) as compared to non-hispanics (86%) (Tagliani, H. 2013). A possible rationale to the negative coefficient could be that the expenditure on apparel by non-hispanics is higher than that of the hispanics, as

hispanics tend to purchase clothing only when they are on sale (60.5% of males and 69.7% of females) (*ibid*).

An interesting variable that arose as statistically significant was having an education level that of edu1 (high school diploma or equivalent). Those with this specific highest education level are expected to spend $11 more on apparel than those who have a higher or lower highest education level. The users who fall within this education level are likely to be teenagers, of which 53% of teens aged 14-17 have been recorded to have purchased things online (Lenhart, A. *et al.* 2010). The last two significant variables (Race1 and Race3) which insinuate that being White or Asian results in lowered expenditure on clothing are somewhat unexplainable, though Asian Americans have been recorded to spend an average of 0.3% less of their household income on clothing as compared to other American households. (U.S. Bureau of Labor Statistics, 2005). It is surprising that the model suggests that being white connotes decreased expenditure on apparel, as suggested by the -5.36 coefficient on the Race1 variable.

Health and Beauty Regression Model

	prod_qty	rod_totpricbasket_totmost_educh_oldest_a	nsus_regicusehold_sisehold_incc	children	al_backgronection_spntry_of_ori	edu0	edu1	edu2	edu3	edu4	edu5	edu99	census1	census2	census3	census4	race1	race2	race3	race5
prod_qty	1																			
prod_totp	0.129268	1																		
basket_tot	-0.00594	0.23321	1																	
hoh_most	-0.03689	-0.0286	-0.02369	1																
hoh_oldes	0.125885	0.062117	-0.10524	-0.07562	1															
census_re	0.00189	-0.07687	0.103524	-0.13919	0.019016	1														
household	-0.03915	-0.09227	0.003898	0.017924	-0.12338	0.105386	1													
household	-0.00142	-0.01262	-0.04245	0.107273	0.049376	-0.1837	0.071213	1												
children	0.020775	-0.09211	-0.10295	0.028377	-0.10822	0.042415	0.65038	0.059098	1											
racial_bac	-0.03593	0.026278	0.024405	0.047901	-0.10314	0.030705	0.162486	-0.09061	-0.04011	1										
connectio	-0.01501	-0.00131	0.001808	0.033008	-0.04299	-0.14515	0.008636	0.120038	-0.02711	0.002137	1									
country_o	-0.06226	-0.01847	0.11048	-0.00677	-0.09555	0.186124	0.24475	-0.09589	0.051692	0.403508	-0.02936	1								
edu0	-0.03226	0.022603	0.259718	-0.0922	-0.08947	0.073903	-0.01941	-0.10607	0.031866	-0.03641	-0.10162	-0.05164	1							
edu1	0.016375	-0.0363	-0.05016	-0.28957	0.028159	-0.11428	0.063318	-0.04787	0.089926	-0.0635	-0.05419	-0.10408	-0.04383	1						
edu2	0.061947	0.015765	0.032499	-0.44628	0.08575	0.143648	-0.01426	-0.08486	-0.03233	-0.0515	-0.01275	0.060758	-0.06935	-0.19957	1					
edu3	-0.02325	0.013499	-0.04423	-0.20534	-0.08771	0.087617	0.142143	-0.04763	0.036424	0.209902	-0.09311	0.114986	-0.0328	-0.09441	-0.14938	1				
edu4	-0.01185	-0.00345	-0.0056	-0.28129	0.015771	0.070402	-0.1236	-0.00073	-0.05968	-0.05765	0.025637	-0.04555	-0.04621	-0.133	-0.21045	-0.09955	1			
edu5	-0.00294	0.053267	-0.02884	-0.19552	0.069199	-0.04461	-0.06364	0.107605	-0.0846	-0.04002	0.135912	0.011423	-0.03307	-0.09517	-0.15059	-0.07124	-0.10036	1		
edu99	-0.0357	-0.028	-0.0198	0.999749	-0.0729	-0.14034	0.020008	0.101923	0.027417	0.04802	0.032196	-0.0068	-0.096	-0.27627	-0.43714	-0.20679	-0.29133	-0.20846	1	
census1	-0.02803	0.041278	-0.0813	0.174545	-0.0272	-0.78785	-0.12088	0.185477	-0.1122	0.02874	0.15951	-0.06276	-0.04812	0.031057	-0.13022	-0.04058	-0.068	-0.00453	0.17582	1
census2	0.010025	0.056434	-0.01243	-0.00053	0.002655	-0.26709	-0.01394	0.01698	0.058431	-0.09576	-0.09905	-0.1704	-0.00421	0.123952	-0.05915	-0.05032	-0.04693	0.055217	0.00162	-0.28598
census3	0.053698	-0.03293	0.002435	-0.14942	0.024892	0.28775	0.107573	-0.10663	0.102643	0.01927	0.077609	0.038499	-0.03317	-0.03729	0.12368	-0.00443	0.103291	0.018413	-0.15364	-0.38117
census4	-0.04158	-0.05975	0.090957	-0.00832	-0.00308	0.729942	0.014854	-0.08367	-0.059	0.043139	-0.1499	0.186326	0.089378	-0.11054	0.050375	0.094761	-0.00121	-0.0699	-0.00655	-0.2987
race1	0.04632	-0.04345	-0.00588	-0.0701	0.037776	0.042468	-0.1256	0.056056	0.083227	-0.86134	-0.03794	-0.31472	0.042268	0.05958	0.047356	-0.14544	0.088545	-0.01259	-0.0701	-0.0909
race2	-0.04316	0.03653	-0.01669	0.054573	0.062414	-0.11433	0.027751	0.040368	-0.09377	0.169401	0.058254	0.033465	-0.02647	-0.01435	-0.01781	-0.01775	-0.08034	0.07902	0.054449	0.113787
race3	0.016177	0.036908	-0.02817	0.046058	0.058194	-0.05636	-0.07221	-0.05631	-0.0414	0.168314	0.0489	-0.03758	-0.01108	-0.03189	0.002103	-0.02387	-0.03363	0.020832	0.045954	0.083949
race5	-0.0303	0.011337	0.034529	0.027632	-0.13197	0.06896	0.175175	-0.09127	-0.01158	0.956589	-0.02086	0.414951	-0.02919	-0.05571	-0.04933	0.224761	-0.03423	-0.0634	0.027806	-0.01316

Figure 8: Correlation Table of Health and Beauty Variables

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	0	0	N/A	N/A	0	0	0
hoh_oldest_ag	0.69493	0.290342463	2.393484814	0.016806	0.125429905	1.26443065	642029.4
household_siz	-1.24185	0.693298606	-1.79121734	0.073451	-2.60173849	0.11804152	4109.11
household_inc	0.07592	0.469947053	0.161549816	0.871681	-0.84587097	0.99771069	15597
children	-0.83596	1.89327316	-0.44153988	0.658883	-4.54956938	2.87765816	2867.833
connection_sp	0.345267	1.404712112	0.245791834	0.805876	-2.41004522	3.10057876	918.0834
country_of_ori	-1.20991	2.194272342	-0.55139662	0.58144	-5.5139314	3.09410272	84.98402
edu0	0	0	N/A	N/A	0	0	0
edu1	-8.35586	6.452531783	-1.29497401	0.19552	-21.0123605	4.30063859	113.438
edu2	-6.76788	6.308869386	-1.07275592	0.283546	-19.1425858	5.60683194	28.25146
edu3	-2.10842	6.650659205	-0.31702355	0.751268	-15.1535377	10.9367065	4314.998
edu4	-4.70141	6.446889829	-0.72925255	0.465956	-17.3468438	7.9440221	1203.336
edu5	-2.14438	6.684432843	-0.32080252	0.748403	-15.2557513	10.9669854	3150.508
edu99	-7.63896	6.25890569	-1.2204945	0.222461	-19.9156662	4.6377462	10494.64
census1	27.99466	9.99654851	2.800432532	0.005166	8.386649068	47.6026702	651.8459
census2	29.4408	10.02653087	2.936289313	0.00337	9.77397511	49.1076158	4545.702
census3	25.64542	10.03133438	2.556531071	0.010666	5.96917571	45.3216603	243.2823
census4	25.35205	9.876804097	2.566827426	0.010355	5.978917088	44.7251862	8555.93
race1	-1.99212	7.449469506	-0.26741792	0.789183	-16.6040927	12.6198493	4149.653
race2	3.616651	7.998377139	0.452173162	0.651207	-12.0719898	19.3052928	3.366779
race3	0	0	N/A	N/A	0	0	0
race5	4.707324	8.138063525	0.578432947	0.563055	-11.255309	20.6699571	232.1207

Figure 9: Regression model from Health and Beauty data with all demographic variables

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	0	0	N/A	N/A	0	0	0
hoh_oldest_age	0.624764	0.28424708	2.197960546	0.028096	0.067223	1.182305	642029.4
household_size	-1.4201	0.502173005	-2.82790509	0.004744	-2.40509	-0.4351	4109.11
census1	21.44961	2.89571541	7.407362656	2.09E-13	15.76977	27.12946	4213.775
census2	22.59413	3.006817117	7.514301	9.53E-14	16.69636	28.4919	8187.256
census3	19.25445	2.970154432	6.482641308	1.2E-10	13.42859	25.0803	3555.521
census4	19.01796	2.995306074	6.349255359	2.82E-10	13.14277	24.89315	28072.22
Adjusted R^2	0.009751						
P-Value	0.001						

Figure 10: Regression model derived from Health and Beauty data

*prod_totprice = 0.62*hoh_oldest_age - 1.42*household_size + 22.45*census1 + .22.59*census2 + 19.25*census3 + 19.02*census4*

From the results of our Multiple Linear Regression Analysis, it was observed that there were several significant variables that could be included into our regression model which were not

autocorrelated nor collinear (shown through the correlation table). We narrowed our regression model to 6 independent variables as highlighted in Figure 10 above.

Significant factors that affect the total amount spent on Health and Beauty products are the oldest age in a household, household size as well as the region from which a consumer is from. When there are older people present in a household, there is more likely to be greater expenditure on health and beauty products as this group of consumers is generally more health conscious or require specialised health products. This is shown through a positive coefficient value for hoh_oldest_age which is 0.62. When there is a larger household size, this would most likely mean that the household has children. With more children to feed(i.e. a larger household size), parents are unlikely to spend on health and beauty products due to lesser disposable income to spend on such products. This is reflected by the negative coefficient of household_size, which is -1.42. From the regression model, consumers from census1(North East region) and census2 (North Central Region region) are more likely to purchase Health and Beauty products as compared to their Southern and Western counterparts. This is due to the fact that states in the North East and North Central parts of USA (New York, Washington D.C etc.) generally hold a wealthier population (Ali Zifan, 2015). With a wealthier population, they are able to spend more on Health and Beauty products. Therefore substantiating that the coefficients of census1 (22.45) and census2 (22.59) are higher than census3 (19.25) and census4 (19.02).

Music Regression Model

	prod_qty	prod_totprice	basket_tot	most_educa	oldest_	nsus_reg	usehold	ehold_inc	children	l_backg	nection	ntry_of_c	edu0	edu1	edu2	edu3	edu4	edu5	edu99	census1	census2	census3	census4	race1	race2	race3	race5
prod_qty	1																										
prod_totprice	0.104396109	1																									
basket_tot	0.15536974	0.306144734	1																								
hoh_most_education	-0.021399591	-0.0369881	-0.072581215	1																							
hoh_oldest_age	-0.006195921	0.090852983	0.005089152	-0.055213	1																						
census_region	0.063060768	-0.046137627	-0.092253103	-0.0106111	-0.0696	1																					
household_size	0.074240496	0.023624819	0.035403459	0.1423022	0.0874	-0.0159	1																				
household_income	0.01138245	0.119184861	0.048535421	0.0937898	0.1715	0.0438	0.0176	1																			
children	0.029604336	0.008346252	0.03143747	0.0420318	0.0127	-0.0319	0.66	-0.0141	1																		
racial_background	-0.021290079	-0.042051982	-0.025516124	0.0312626	-0.0392	0.0294	0.0974	-0.0896	0.0716	1																	
connection_speed	-0.00151677	0.004764487	-0.073547724	-0.0180506	-0.0495	0.0609	0.0603	0.0711	0.0651	-0.0705	1																
country_of_origin	-0.007160174	-0.051381075	-0.037909316	0.1725687	-0.0591	0.0917	0.0987	-0.0251	0.067	0.3237	0.0267	1															
edu0	-0.025566441	-0.028467366	0.018223228	-0.2880692	0.0404	0.0405	0.0317	-0.1201	0.0139	-0.0148	-0.1494	-0.1098	1														
edu1	-0.025566441	-0.028467366	0.018223228	-0.2880692	0.0404	0.0405	0.0317	-0.1201	0.0139	-0.0148	-0.1494	-0.1098	1	1													
edu2	-0.00081347	0.006056595	0.0500884	-0.4119912	0.028	-0.0296	-0.1916	-0.0726	-0.1331	-0.0482	0.0218	0.0109	-0.1998	-0.1998	1												
edu3	-0.022273428	0.023098738	0.028799532	-0.2370855	-5E-05	0.017	0.0416	-0.0331	0.0671	0.1076	0.0584	-0.0072	-0.1183	-0.1183	-0.1741	1											
edu4	-0.002227849	0.019229593	-0.014157813	-0.2666868	-0.0135	-0.0297	-0.024	0.0356	0.0157	-0.056	0.0803	-0.0683	-0.1371	-0.1371	-0.2017	-0.1195	1										
edu5	-0.017571898	0.021942791	-0.024043825	-0.176091	0.0688	0.0304	-0.0798	0.1438	0.0079	0.0094	0.0211	-0.0817	-0.0934	-0.0934	-0.1373	-0.0813	-0.0943	1									
edu99	-0.019750966	-0.037457176	-0.071267601	0.9997392	-0.0559	-0.0104	0.1437	0.0895	0.0413	0.0307	-0.0208	0.1728	-0.2757	-0.2757	-0.4055	-0.2402	-0.2783	-0.1895	1								
census1	-0.016989668	0.061129048	0.13697551	0.0226132	0.0999	-0.7729	0.0238	0.0083	-0.0045	0.0264	-0.0165	-0.0765	-0.0388	-0.0388	0.0841	-0.0561	-0.0116	-0.0689	0.0239	1							
census2	-0.055608162	-0.006935721	-0.058399122	-0.0097712	-0.0092	-0.296	-0.0145	-0.0164	-0.0291	-0.0974	-0.0202	-0.0181	-0.0256	-0.0256	-0.0242	0.0444	0.0601	0.0068	-0.0116	-0.2764	1						
census3	-4.92835E-05	-0.04243726	-0.044628337	-0.0175539	-0.0862	0.2736	-0.0004	-0.0913	0.1349	0.0368	-0.0589	0.0237	0.0545	0.0545	-0.1089	0.0291	-0.009	0.0983	-0.0183	-0.366	-0.3781	1					
census4	0.071114082	-0.004850677	-0.025011729	0.0073575	0.0088	0.7318	-0.0082	0.11	-0.1175	0.0292	0.1014	0.0651	0.0019	0.0019	0.0641	-0.0219	-0.0379	-0.0498	0.0087	-0.2852	-0.2947	-0.3902	1				
race1	0.016848283	0.094573141	0.072705418	-0.0103968	-0.029	0.0234	-0.1046	0.0129	-0.0438	-0.8038	0.0667	-0.1886	-0.0304	-0.0304	0.044	-0.0802	0.1013	-0.0663	-0.0103	-0.0537	0.1034	-0.0956	0.0568	1			
race2	-0.002021765	-0.113337122	-0.088166374	0.0065673	0.1057	-0.1093	0.0724	0.0827	0.0051	0.205	-0.0081	-0.0063	0.0157	0.0157	-0.0387	0.0295	-0.0906	0.1136	0.0061	0.0732	-0.042	0.1306	-0.175	-0.7182	1		
race3	-0.011136096	0.000228481	-0.01907224	-0.0695434	-0.0498	0.1225	-0.003	-0.0295	-0.0241	0.2443	-0.0891	-0.0668	0.1585	0.1585	0.0509	-0.0295	-0.0597	-0.0407	-0.0665	-0.0358	-0.0858	-0.0319	0.1545	-0.3113	-0.0595	1	
race5	-0.018187123	-0.006724086	0.008484868	0.0529108	-0.0605	0.0282	0.0795	-0.1123	0.0814	0.9066	-0.0431	0.3642	-0.0714	-0.0714	-0.0541	0.1128	-0.0098	-0.0148	0.0516	0.015	-0.0613	0.0056	0.0395	-0.5085	-0.0971	-0.0421	1

Figure 11: Correlation Table of Music Variables

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	1.367488	2.11753546	0.64579203	0.51856	-2.78792736	5.52290243	43570.4
hoh_oldest_age	0.260772	0.11597632	2.24849667	0.02477	0.033182449	0.48836229	586.946
household_size	0.302818	0.26873209	1.12683877	0.26009	-0.22453746	0.83017295	18.2548
household_income	0.619184	0.17456124	3.54708715	0.00041	0.276627981	0.96173989	757.059
children	-0.41267	0.75740048	-0.5448491	0.58598	-1.89897855	1.07364062	0.66595
connection_speed	-0.05222	0.57258837	-0.0912077	0.92735	-1.17586198	1.07141299	0.00966
country_of_origin	-1.02751	0.81900937	-1.2545821	0.20993	-2.63472427	0.57969532	164.016
edu0	-0.1417	0.90866277	-0.1559416	0.87611	-1.92484243	1.6414457	41.665
edu1	0	0	N/A	N/A	0	0	0
edu2	0.474601	0.72948492	0.65059786	0.51546	-0.95692724	1.90612989	11.2022
edu3	1.148721	0.98169929	1.17013481	0.24223	-0.77774916	3.07519017	49.3624
edu4	0.27679	0.88533111	0.31263956	0.75462	-1.4605689	2.01414795	22.1933
edu5	0.94975	1.20386591	0.78891666	0.43035	-1.41269572	3.31219547	4.70833
census1	1.646885	0.82136781	2.00505222	0.04523	0.035047394	3.25872329	181.386
census2	0.425701	0.80623764	0.52800935	0.59761	-1.15644575	2.00784777	2.16421
census3	0.606846	0.76231174	0.79605951	0.42619	-0.88910184	2.10279285	2.01667
census4	0	0	N/A	N/A	0	0	0
race1	-0.20791	1.74974142	-0.1188214	0.90544	-3.641569	3.22575549	673.246
race2	-3.99348	1.91783685	-2.0822851	0.03757	-7.75701281	-0.22995322	737.562
race3	0	0	N/A	N/A	0	0	0
race5	0.074052	2.08362391	0.03554021	0.97166	-4.01481502	4.16291988	0.08737

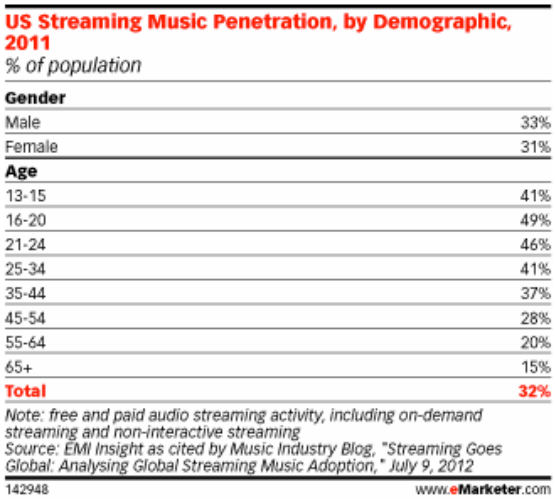
Figure 12: Regression model from Music data with all demographic variables

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	2.269108	0.980162	2.315035	0.020814	0.345687	4.19253	43570.43
hoh_oldest_age	0.282814	0.11333	2.495488	0.012739	0.06042	0.505208	586.9463
household_income	0.607059	0.167098	3.632961	0.000295	0.279155	0.934963	757.6224
census1	1.302442	0.647915	2.010204	0.044679	0.031005	2.573879	207.3721
race2	-3.55492	0.811527	-4.38053	1.31E-05	-5.14742	-1.96242	1317.436
Adjusted R^2	0.036453						

Figure 13: Regression model derived from Music data

*prod_totprice = 0.28*hoh_oldest_age + 0.61*household_income + 1.30*census1 - 3.55*race2*

From the results of our Multiple Linear Regression Analysis, it was observed that there were several significant variables that could be included into our regression model which were not autocorrelated nor collinear (shown through the correlation table). We narrowed our regression model to 4 independent variables as highlighted in Figure 13 above.



US Streaming Music Penetration, by Demographic, 2011
% of population

Gender	
Male	33%
Female	31%
Age	
13-15	41%
16-20	49%
21-24	46%
25-34	41%
35-44	37%
45-54	28%
55-64	20%
65+	15%
Total	**32%**

Note: free and paid audio streaming activity, including on-demand streaming and non-interactive streaming
Source: EMI Insight as cited by Music Industry Blog, "Streaming Goes Global: Analysing Global Streaming Music Adoption," July 9, 2012

142948 www.eMarketer.com

Figure 14: US Music Penetration by Age

Significant factors that affect the total amount spent on Music products are the oldest age in a household, household income, whether they were from North East America and whether they were of Black ethnicity. Age is a significant factor as there is differing levels of music consumption across the various age groups. From our regression model, we can see that the higher the oldest age in the household, the more the consumer would spend on music transactions. One possible reason is the fact that youths stream their music instead of purchasing them (refer to Figure 14) as they lack purchasing power. Conversely, adults are more inclined to purchase their music.

Household income is another significant factor. We can see that the higher the household income, the more the consumer would spend on music transactions. This is given by the +0.61 coefficient for household income. This is to be expect as with a higher level of household income, there is more disposable income that can be used for the purchase of entertainment goods such as Music. From out model, we found that living in the North-East Region of USA is a another significant factor in determining the amount spent on music consumption. This is given by the +1.30 coefficient for census1. As mentioned earlier, states in the North East regions of USA (New York, New Jersey, Maine etc.) generally hold a wealthier population. Thus there is likely to be greater expenditure on entertainment. Lastly, we found out that if a consumer is African-American, he/she are less likely to spend on music transactions. This is given by the -3.55 coefficient for race2. However research in African American behaviour has shown that African Americans actually spend more than the general population of Americans (AdWeek, 2015). On further inspection of our dataset, we realised that of the 122 transactions that were from African Americans, 77 of them (63.1%) of them were transactions that had a price value of $0.00. The ratios for the other races were as follows: White (43.0%), Asians (42.3%) and Others (34.8%). We feel that since more than half of the African American dataset were transactions of $0.00, this could have affected the linear regression model and thus produced a negative coefficient.

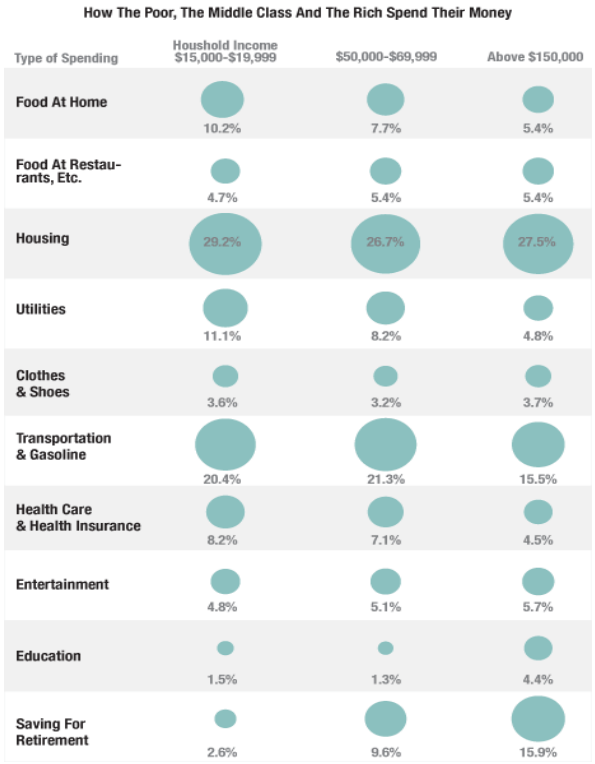
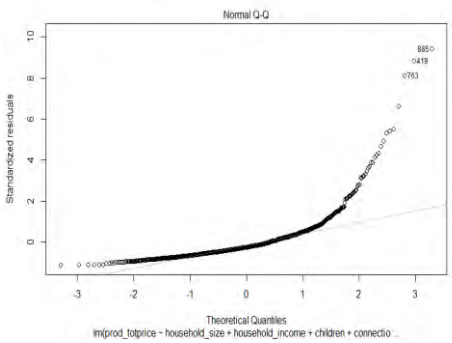


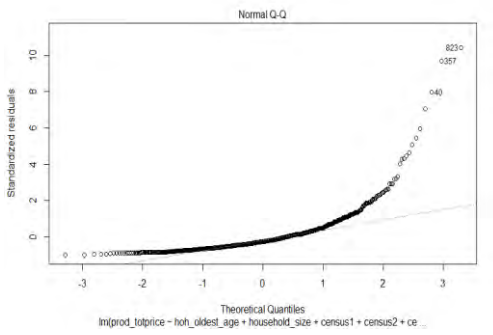
Figure 15: How the Poor, the Middle Class and the Rich in US Spend their Money

Linear Regression Assumptions

First Assumption - Variables are Normally Distributed



Apparel



Health and Beauty

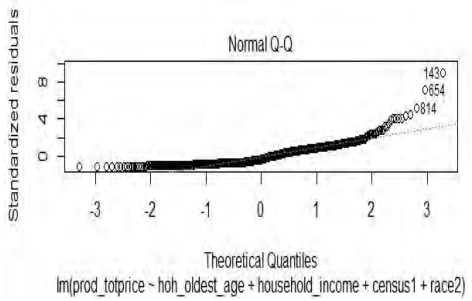
 Music	If the points are close to the plotted line of the Normal Q-Q graph, we can conclude that the variables are normally distributed. We can see this holds for all 3 product categories

Figure 16: Normal Quantile - Quantile plots of regression models

Second Assumption – Linear Relationship between the Independent and Dependent Variable(s)

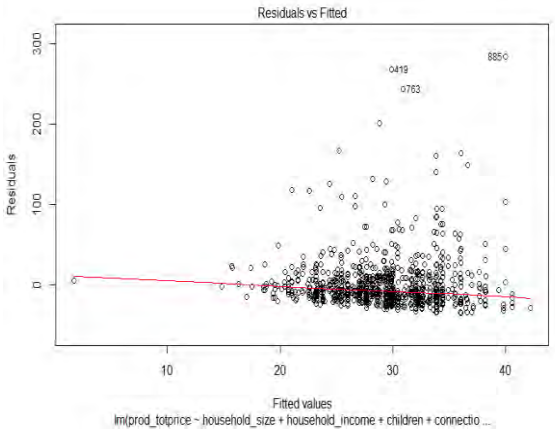
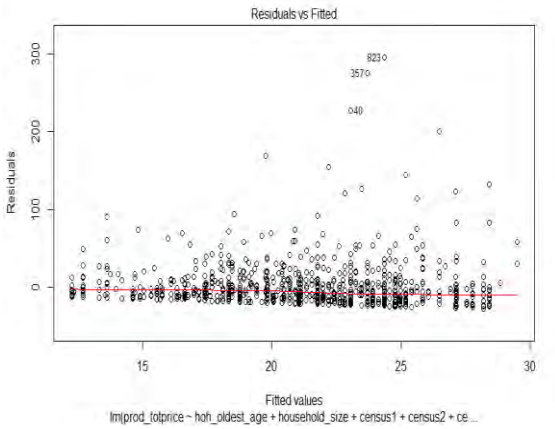
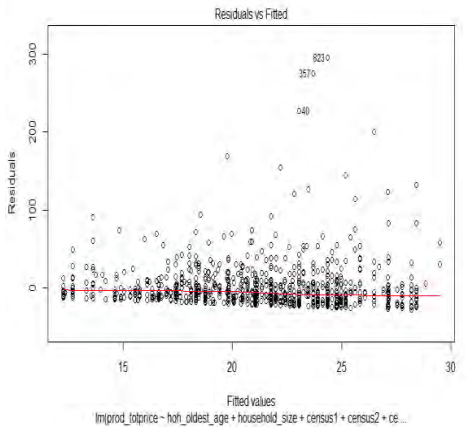
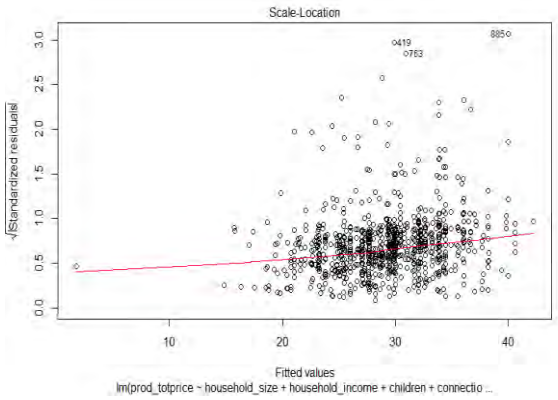
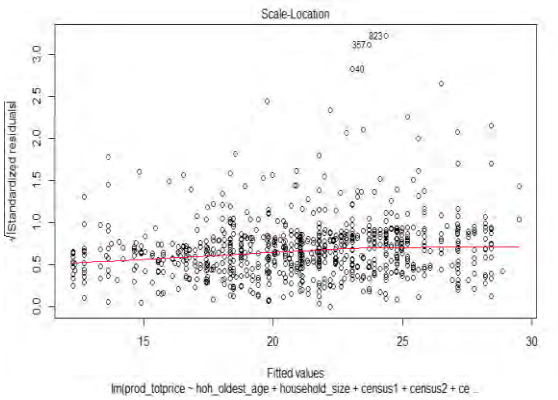
 Apparel	 Health and Beauty
 Music	We can conclude from these graphs that there is a linear relationship between the independent and dependent variable as the line plotted resembles the expected straight line of linear association.

Figure 17: Residuals vs Fitted plots of regression models

Third Assumption – Assumption of Homoscedasticity



Apparel



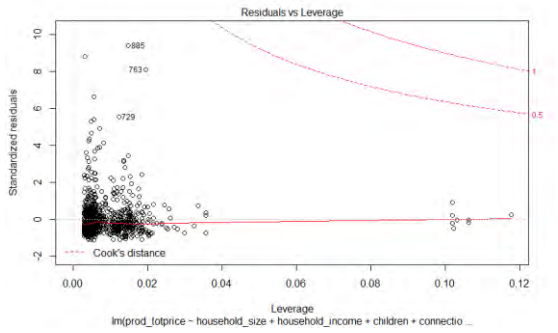
Beauty and Health

Music

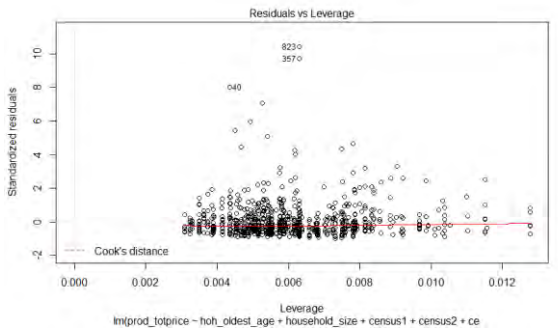
If the spread of the standardized residuals against fitted values is constant, we can conclude that homoscedasticity is valid for the data. We can make a visual observation that this is true for the data above.

Figure 18: Scale-Location plots of regression models

Presence of Outliers:



Apparel



Beauty and Health

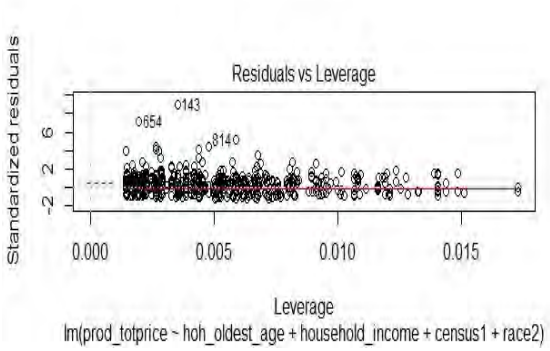
 Music	A generally low cook's distance for majority of the points suggest that there are only few datapoints that bear a greater influence on the other points (2-3 per model). These outliers were not excluded due to the nature of their low leverages, and are expected from large sample sizes.

Figure 19: Residuals vs Leverage plots of regression models

Multicollinearity - VIF values:

household_size household_income children connection_speed country_of_origin Race1 Race3 1.709199 1.115928 1.755559 1.047873 1.053980 1.158185 1.094323 Edu1 1.023735 Apparel	A general rule of thumb is that if VIF values are more than 7, then multicollinearity is high. The VIF values for across all 3 product categories are below 7 showing that there is little or no multicollinearity
hoh_oldest_age household_size census1 census2 census3 1.017930 1.036988 1.556891 1.521017 1.637884 Beauty and Health	
hoh_oldest_age household_income census1 race2 1.048964 1.035424 1.014394 1.020243 Music	

Figure 20: VIF test statistics of regression models

Auto Correlation - Durbin Watson Test:

lag Autocorrelation D-W Statistic p-value 1 -0.01352126 2.026402 0.664 Alternative hypothesis: rho != 0 Apparel	An ideal statistic for the Durbin Watson Test is 2. Where there is minimal auto correlation. The Durbin Watson test statistic across all 3 product categories are very close to 2, indicating a minimum level of auto correlation.
lag Autocorrelation D-W Statistic p-value 1 -0.04514795 2.089567 0.118 Alternative hypothesis: rho != 0 Health and Beauty	
lag Autocorrelation D-W Statistic p-value 1 0.02760635 1.943945 0.362 Alternative hypothesis: rho != 0 Music	

Figure 21: Durbin Watson Test statistics of regression models

Predictive Analysis (Data Mining)

Clustering

Apparel:

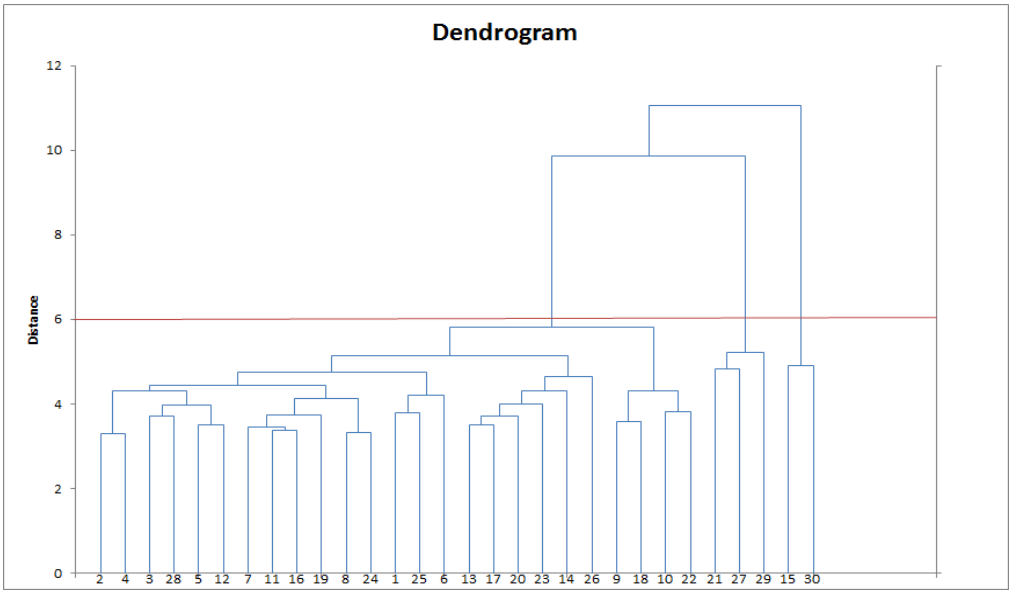


Figure 22: Dendrogram of Apparel clusters

	Price	Household Size	Household Income	With Children	Connection Speed	Country of Origin	Caucasian	Asian	High School Diploma or Equivalent	n
Cluster 1										
Mean	28.47051672	2.903748734	4.680851064	0.396149949	0.401215805	0.106382979	0.883485309	0	0.10739615	987
SD	26.92400508	1.276983058	1.668971117	0.489344233	0.490393042	0.308483478	0.321003767	0	0.309773208	
Cluster 2										
Mean	16.351	3.6	5.8	0.5	0.7	0.1	0	1	0	10
SD	12.54882504	1.173787791	1.751190072	0.527046277	0.483045892	0.316227766	0	0	0	
Cluster 3										
Mean	299	2.666666667	4.666666667	0.333333333	0.333333333	0	0.666666667	0	0.666666667	3
SD	24.51530134	1.154700538	2.516611478	0.577350269	0.577350269	0	0.577350269	0	0.577350269	

Figure 23: Descriptive summary of Apparel clusters

Taking a closer look at the dendrogram illustrated above, our group has decided to cluster into 3 main groups. After analysing the mean and standard deviation for each variable, it is evident that characteristics such as being Caucasian or not, Asian or not and having a High School Diploma or Equivalent or not are significantly different among the three clusters. Cluster 1, which has a mean price of $28.47, is composed of mainly Caucasians and people with no high school diploma or equivalent (This does not necessitate that they have a lower education degree) on the average. Cluster 2, which has a mean price of $16.35, is composed of all Asians and people with no high school diploma or equivalent on the average. Cluster 3, which has a mean price of $299, is composed of mainly Caucasians and people with high school diploma or equivalent on the average.

These clustering interpretations show that people who fall under those clusters due to similar characteristics would have a higher propensity to purchase apparels and have expenditure on those apparels close to the average price of each cluster. However, looking at the data, Cluster 1 hold close to the entire proportion of the consumers sampled who bought apparel (98.7%) as compared to Cluster 2 (1%) and Cluster 3 (0.3%). This shows that since majority of the sample population belong to Cluster 1, people who fall within Cluster 1 should have a higher probability of getting a recommendation of apparel as compared to people who fall within other clusters.

Health & Beauty:

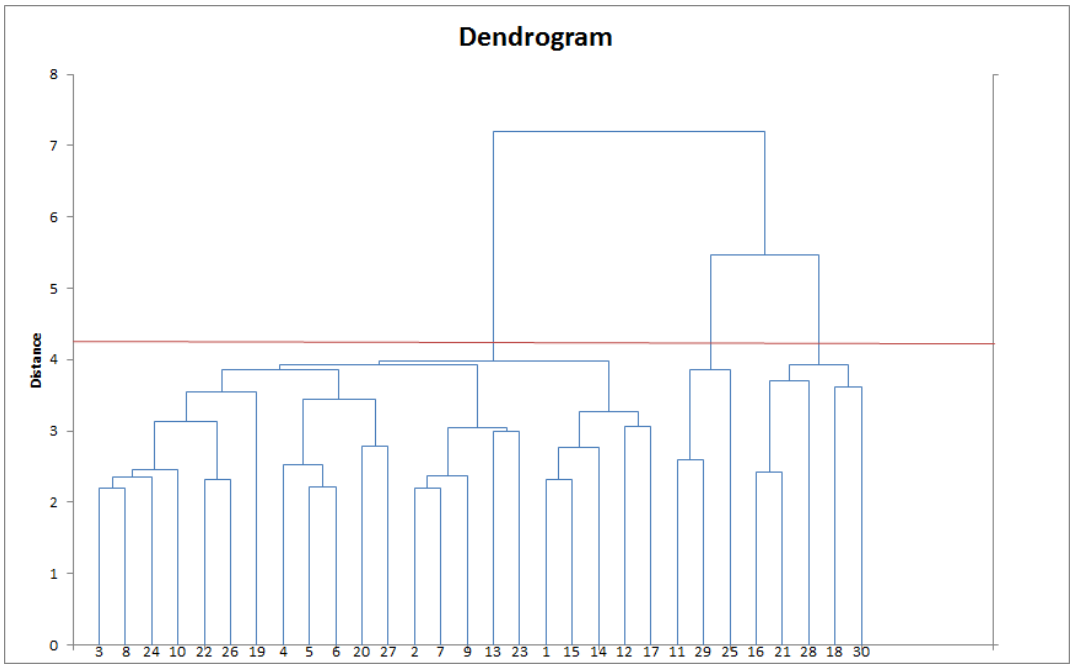


Figure 24: Dendrogram of Health and Beauty clusters

	Price	Oldest person in Household	Household Size	North East	North Central	South	West	n
Cluster 1								
Mean	19.03968	6.95951417	2.951417004	0.229757085	0.212550607	0.326923	0.230769	988
SD	20.0375	2.356817821	1.329909998	0.420889613	0.409319464	0.469326	0.421538	
Cluster 2								
Mean	289.6467	9.666666667	2.666666667	0	0.333333333	0.666667	0	3
SD	35.9446	1.154700538	1.154700538	0	0.577350269	0.57735	0	
Cluster 3								
Mean	167.3722	7.666666667	2	0.444444444	0.333333333	0.111111	0.111111	9
SD	27.56525	2.061552813	0.707106781	0.527046277	0.5	0.333333	0.333333	

Figure 25: Descriptive summary of Health and Beauty clusters

From the dendrogram above, our group clustered the data into 3 groups. Looking at the table analysing each variable above, only one characteristic, which is the age of the oldest person in the household, is significantly different among the three clusters. Cluster 1, which has a mean price of $19.04, is composed of mainly people who are aged 45-49. Cluster 2, which has a mean price of $289.65, is composed of all Asians and people with no high school diploma or equivalent on the average. Cluster 3, which has a mean price of $167.37, is composed of mainly Caucasians and people with high school diploma or equivalent on the average.

These clustering interpretations show that people who fall under those clusters due to similar characteristics would have a higher propensity to purchase apparels and have expenditure on those apparels close to the average price of each cluster. However, looking at the data, Cluster 1 hold close to the entire proportion of the consumers sampled who bought apparel (98.7%) as compared to Cluster 2 (1%) and Cluster 3 (0.3%). This shows that since majority of the sample population belong to Cluster 1, people who fall within Cluster 1 should have a higher probability of getting a recommendation of apparel as compared to people who fall within other clusters.

Music:

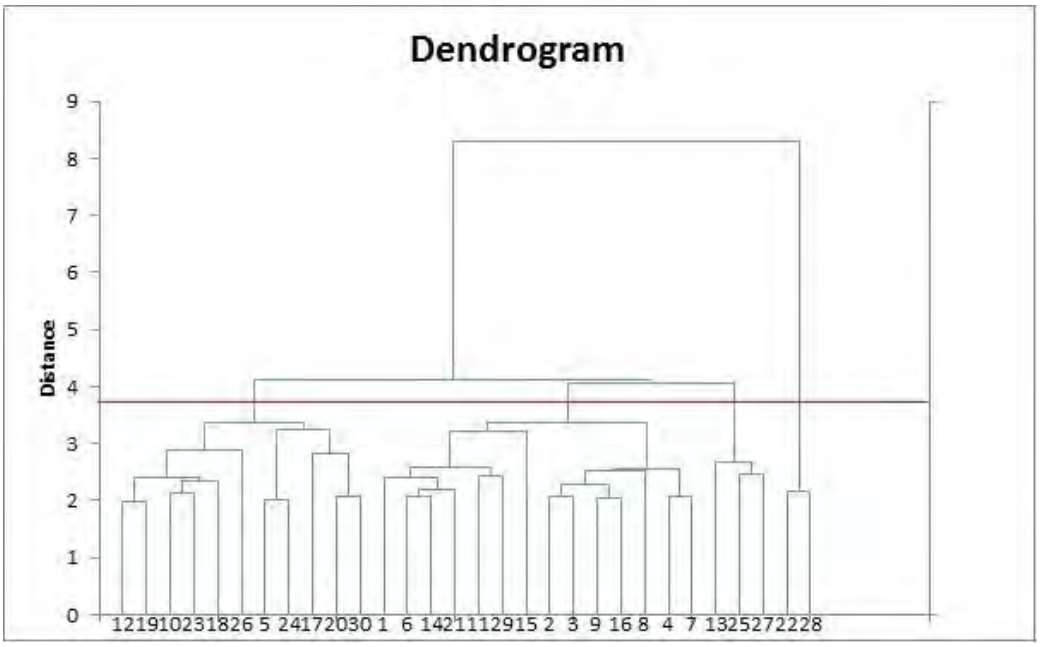


Figure 26: Dendrogram of Music clusters

	Price	hoh_oldest_age	household_income	census1	race2	n
Cluster 1						
Mean	6.227334	6.541327125	4.236321304	0.203725262	0	859
SD	6.937254	2.392997478	1.57598039	0.403001673	0	
Cluster 2						
Mean	3.993719	7.330578512	4.661157025	0.289256198	1	242
SD	6.308169	2.203740872	1.65606823	0.4553024	0	
Cluster 3						
Mean	34.49389	7.111111111	4.666666667	0	0	54
SD	7.896512	1.409584373	1.909727421	0	0	
Cluster 4						
Mean	73.69	8.5	4	0	0	2
SD	10.30962	0.707106781	1.414213562	0	0	

Figure 27: Descriptive summary of Music clusters

From the dendrogram above, we had clustered the data into 4 groups. Looking at the table, we analysed each variable above, race2, household income, household oldest age and price were differentiating factors among the 4 clusters. Firstly, we compared race2 among the four clusters and concluded that Cluster 2 were customers of Black American ethnicity. We then compared Cluster 1 and Cluster 3 and concluded that the only differentiating factor between them was that Cluster 3 had a higher household income bracket than Cluster 1. Cluster 3 had an average household income of between $50,000 - $75,000 per annum while Cluster 1 had an average household income of between $35,000 - $50,000 per annum. Cluster 4 remained as the cluster with the highest spending and oldest household age (55-59 yrs old). We observed that although Cluster 2 and 3 shared the same characteristics other than race, there was a huge difference in amount spent. That is, Black Americans spend less on music transactions as compared to their peers from other races of the same socio-economic background. This is supported by our linear regression model. We also observed that although Cluster 1 and 4 shared the same characteristics other than household oldest age, there was again a huge difference in amount spent. That is, the older the household age, the more money spent on music transactions. This is again supported by our linear regression model.

These clustering interpretations show that people who fall under those clusters due to similar characteristics would have a higher propensity to purchase music and have expenditure on those music transactions close to the average price of each cluster. However, looking at the data, Cluster 1 holds close to the entire proportion of the consumers sampled who bought apparel (74.2%) as compared to Cluster 2 (20.9%) , Cluster 3 (4.67%) and Cluster 4 (0.23%). This shows that since majority of the sample population belong to Cluster 1, people who fall within Cluster 1 should have a higher probability of getting a recommendation of music as compared to people who fall within other clusters.

Association:

	Apparel	Shoes	Accessories	Food & Beverages	Other home & living items	Books & magazines	Movies & videos	Other bmv
Apparel		937	551	190	411	88	211	1
Health & Beauty	390	55	70	537	561	56	80	1
Music	65	5	9	12	18	596	1176	197

Figure 28: Top 3 secondary purchases by customers who purchased each of the 3 product categories

In its raw form, our data was grouped by transactions; this means that the transactions from each session were broken up into several entries. As such, the first step of our association analysis was to compile the transaction data into session level.

Therefore, we used a PivotTable to modify and group the data by session ID. We chose to group the data by session ID over machine ID as the same machine could have been used over multiple sessions by different users. As such, this would prevent the situation where there is an inflated number of items in each market basket. This would also help us differentiate each basket according to the time of purchase when session ID instead of machine ID is used.

Instead of performing association rule analysis on the whole data, we chose to carry out the analysis on individual sites as association rules across multiple websites would not be as valid. The number of product categories sold by each website was different and some sites only sold products from specific product categories. For instance, a website owned by an airliner will not be selling other items other than air tickets (product category 43). Hence, we chose to conduct association rule analysis on Amazon and eBay; which have sold products from most of the product categories of interest.

We converted the data into binary form which allowed us to perform association rule analysis on it. Using a minimum cut off at 10% support and confidence, we obtained two valid rules for Amazon and no valid rules for eBay. Valid rules are defined as rules that had a lift value of more than 1. However, these two valid rules did not fall under any of the three product categories that we are using for our project.

The lack of rules might be due to the data being too small after it is grouped according to sites. Hence, the rules do not have a high enough support to be considered by us. Although we would be able to obtain more valid rules if we lowered support below 10%, we did not lower the support further as the dataset is already very small. The Amazon dataset only contains 1727 baskets which

means at 10% support there must be a minimum of 173 baskets that fulfils the rules we obtained. The situation is also similar for eBay which only contains 196 baskets.

As such, we decided to identify the three product categories with the highest number of purchases premised by the user purchasing a product from one of the three product categories which we analysed in this project.

From Figure 28, most of the results we obtained are quite logical and expected, with a few exceptions; Health & Beauty to Food & Beverages and Music to Books & Magazines. We think that customers who purchased health and beauty products are more likely to be health conscious so they are pickier with regards to what they eat and drink. This may result them to purchase what they consume online as it provides more control over what they consume as compared to eating out. As for customers who purchased music we suspect that users who prefer to purchase music online will also prefer to buy books and magazines they need online.

Application

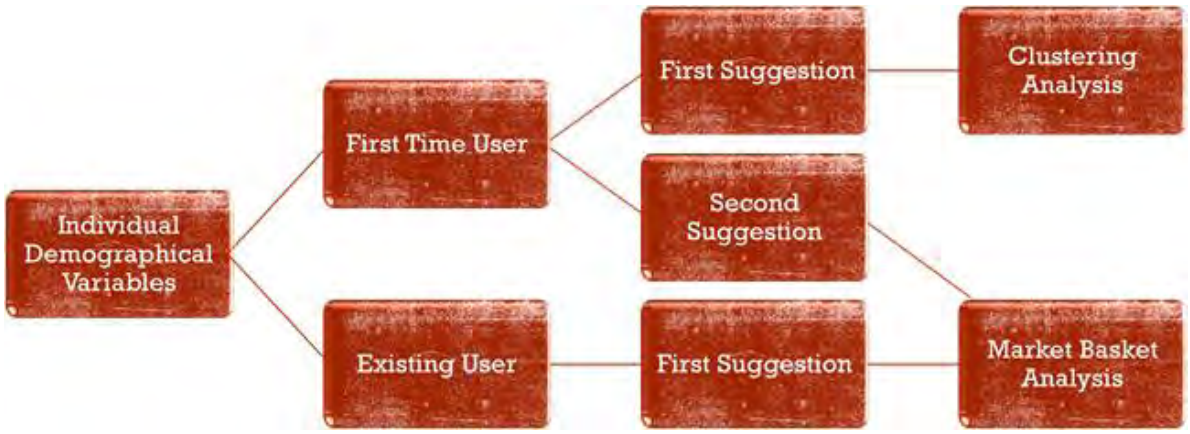


Figure 29: Flow chart of the application of our findings

This chart illustrates a bird's eye view on how we are going to apply our findings from our research on the comScore data. As we recognise the fact that the usage of comScore data would be limited to the demographics of United States, our group decided to take one step further and outline the process of analysing and implementing the data findings into actual application for ecommerce platforms, so that our findings from this research would not be futile even when thrown into a different context.

Figure 29 depicts the flowchart on what happens behind-the-scenes when an online consumer key in his or her demographic variables when signing up for a new account on an e-commerce platform. Starting at the first box, when the individual's demographic variables are keyed in, the system would recognize if the user is a first-time user or an existing user, depending on the existence of any purchase or browsing history. Thus, this means that even if the account isn't new, but there is no trace of purchase history (as the user had not bought anything or browsing history as the user

cleared the cookies or set his or her privacy to high), the user will still be recognized as a new user. When this happens, the system would recommend its first batch of suggestions via the usage of Clustering Analysis.

Firstly, we shall go more in depth into Clustering Analysis- mainly on how our team wants to implement it ideally. We decided that a fair system to group the individuals under certain predetermined clusters would have to consider two factors: the relevance and popularity of the cluster assigned. The relevance of the cluster refers to the how well the individual demographic variables are matched to the predetermined clusters' distinct demographic variables' means. The popularity of the clusters refers to the significance of the cluster as compared to the entire sample size of consumers who bought the product we are targeting at. Thus, we would rank the clusters according to these two factors in a certain method of summation and the total points given would be the status of how applicable the cluster would be for that individual, and eventually, which recommendations would take priority and earn their spot in the top 5 most likely recommendations which our system would display to the individual. These recommendations would be the best possible solution when there is no previous history of the customer to fall back on.

Subsequently, when the first time user has sufficiently accumulated enough browsing history or an existing user has his or her purchase or browsing history available, our system would be able to use Market Basket Analysis to generate the next batch of recommendations for the user. This would require the usage of Association rules which we generated previously, and match the best rules to the most frequent or most recent product browsed or purchased by the user. The rules will thus also be ranked by the lift value and again, only the top 5 recommendations are displayed to the user. On those rare occasions when the Market Basket Analysis is not able to generate rules which are applicable either due to the user browsing very rare products or products with no association with other products, Clustering Analysis would then be used again to generate the next batch of suggestions instead. However, such situations would be rare as we are planning to expand our analysis to all of the product types listed in our data provided eventually, so there will be more that sufficient rules to pick from to dish out recommendations.

Limitations

Firstly, one of the data limitations we faced was the limited number of demographic variables in the data. With more variables such as gender and more specific variables such as state in which the consumer lives in, we would be able to construct a more accurate model to predict the spending of the consumer on a good. Furthermore, the limitations to our application are also affected by the willingness of online consumers sharing their personal information. When an online consumer sign up for an account on an ecommerce platform, they do not wish to fill up information on how many siblings they have, or what is the oldest age out of all the members in their household as that would undermine the credibility of the ecommerce platforms since these information are of no direct

relevance to the user and their account usage. Thus, the scope of individual demographic variables becomes more myopic and that would undermine the accuracy of our suggestions by a long mile.

An attempt to mitigate the above limitations would be to allow the consumers to be aware that additional demographic variables is required to improve on the system in place to predict their choices of items they are planning to look for. This could be made optional so that consumers would not be burdened by the fact of exposing too much sensitive information online. Other variables can also be available when people link their social media accounts to their shopping accounts, so certain information can be passed within legal means.

Secondly, the scale on how we can determine the relevance or closeness between the individual demographic variables and the predetermined clusters can only be set with accuracy given more market research and it depends on the local context as well. This is because the predetermined clusters are defined the means of their distinct demographic variables, there should be a range to determine how far apart from the means of these clusters should it be considered close or not. This range cannot be merely set on a whim, but given more market research either before or after implementation, it can be fine-tuned so that profits for the ecommerce platforms would be maximised.

Thirdly, the summation of the two factors as mentioned previously on how our group decided to rank the clusters (relevance and popularity of the clusters) has to be determined through further market research as well. As of right now, we are unsure that which factor should hold more weight in determining the ranking of the clusters for each individual and how we should compute the factors to determine the points as there might be some formula to follow. By researching more in-depth, the summation of these two factors would generate more accurate predictions which would in turn maximise the profits of the ecommerce platforms which are using our system.

Lastly, the small R^2 values of our regression models was a sign that our model might be accurate only to a small degree. This suggests that the linear regression model may not be the most suitable model to predict the expenditure based on demographic variables. Other existing models could be further used to fine tune the results of our findings.

Overall Lessons Learnt

After seeing through this project, our team has learnt many valuable lessons on descriptive analytics, data mining and its applications. The main lesson we had learnt was that in reality, the data we collected from the large majority might not be as intuitive or logical as we thought it might be, as there isn't always a clear cut association or trend seen. Thus, one cannot blindly perform analytics unto data without a clear cut objective to study and research on, so the other intruding factors can be eliminated to isolate one's objective. For example, the association rules we initially generated were inconclusive as the support value was extremely low due to the wide range of products displayed. Thus, we decided on a different approach to evaluate the association rules which were able to allow us to rank the rules more efficiently, disregarding the support ratio.

Secondly, as the data are not manipulated, results might not be what one has in mind initially. This means that one has to constantly update his or her objectives as the project continues so as to not lose focus on what kind of data should one depict in the end. For example, our team has initial thoughts of showing positive trend between household size and product expenditure, but that was not the case for some of the products we researched on like Health and Beauty. However, this does not mean that the data is inaccurate, but our conclusions now are merely different and further research can be done on our part to find out the explanation behind such a trend.

Lastly, our team has learnt that coming up with an overall structure especially on the implementation of our data would be more sustainable than mere interpretation of the data we are provided. This would mean that our research would not be limited to the data used, and it could be referenced by subsequent research on the implementation of this system on other regional contexts like Asia which will undoubtedly drive the global economy in years to come. What was more important to our team was the viability of our research as our main objective is to extrapolate the process in which find our results to every single product type, and subsequently the whole process implemented for every region in the world.

Conclusion

In conclusion, the application of our refined system will provide e-commerce sites with the opportunity to further meet the needs of their users, allowing them to stay competitive in their advertisement of relevant products for the consumer even before the user enters his/her first query. With the sales through ecommerce sites expected to continue rising through the 21st century, this constantly updating system will maximise the customer base captured by a ecommerce website, maximising its revenue while improving customer service.

References

Statista, (2015). *B2C e-commerce sales worldwide from 2012 to 2018 (in billion U.S. dollars).* Retrieved November 10, 2015 from http://www.statista.com/statistics/261245/b2c-e-commerce-sales-worldwide/

Ali Jaafar, (2015, 20 July). *Amazon to launch prime service in India, readies itself to invest up to $5 Billion*. Retrieved November 10, 2015, from http://deadline.com/2015/07/amazon-india-prime-jeff-bezos-sony-e-commerce-1201482192/

Tagliani H, (2013), *Hispanics & Clothing,* Retrieved November 10 2015 from http://www.blog.thegroupadvertising.com/hispanics-clothing/

Ali Zifan (2015), Map of states by median household income in 2014, Retrieved November 10, 2015 from https://commons.wikimedia.org/wiki/File:Map_of_states_by_median_household_income_in_2014.svg

Lenhart A., Purcell K., Smith A., Zickhur K (2010, 3 February), *Social Media and Young Adults*, Retrieved November 10 2015 from http://www.pewinternet.org/2010/02/03/social-media-and-young-adults/

US Bureau of Labor Statistics, (2005 September), Issues in Labour Statistics, Retrieved November 10 2015 from http://www.bls.gov/opub/btn/archive/spending-by-asian-families-pdf.pdf

AdWeek,. (2015). AOL Black Focus Goes Online. Retrieved 13 November 2015, from http://www.adweek.com/news/advertising/aol-black-focus-goes-online-64535