Final Project: Green Consumption in Consumerism

Submitted by: Jaya Khan
GitHub repo: https://tinyurl.com/4ekfjj3z
Presentation link: https://tinyurl.com/yc3792cb

1. Summary

This project is inspired by an idea of encouraging sustainability in consumerism. The analysis leverages the dataset of a journal[1] published in 2020 on Elsevier Publications. The dataset used in the analysis is derived from the results of the questionnaire distributed to individuals, aged 18 years and above, in the high populated areas of Malaysia. We used this dataset to answer below questions via our analysis.

- Do people who consume more *media* content tend to exhibit high green consumption behaviour?
- Does the green consumption behaviour differ by *age* and *education*?
- Are there any *other factors* that explain the effect on green consumption behaviour?

A proportional odds model is used to answer these questions. The analysis of the proportional odds model revealed that there is an influence of *preference for newspaper, radio, and social media*, along with *ethnicity* and *number of households*, on green consumption behavior. However, we couldn't leverage *age* and *education* in the interpretation because of their low significance (low p-value) at both 90% and 95% confidence intervals.

2. Introduction

Since the outcome variable (green consumption) in the dataset is ordinal (low < medium < high), we ran both proportional odds and multinomial models to select the best fitted model in the analysis. Consequently, we ended up going with proportional odds model for it has low standard errors, explains slightly more observations, and it is parsimonious both in terms of estimation and interpretation compared to multinomial model. Via this analysis, we hope to enable policy makers to create a policy that can encourage individuals consume more green products. Moreover, this analysis also seeks to provide insights to companies about effective media channels for orchestrating green marketing.

3. Data

Questions adapted in the questionnaire are developed using Theory of Planned Behaviour as the basis of research. Questionnaire had total 9 items on demographics, 5 items on media preference, 4 items on social influence[2], 8 items on consumer novelty seeking[3] and 13 items on green consumption behavior[4]. Demographic factors included are gender, age, marital status, education, ethnicity, number of households, personal income, occupation, and work category. Data for media preference reveals the level of engagement with five media channels – newspaper, magazine, television, radio, and social media, and it is collected on the scale of 1 to 5, with 1 being seldom-used and 5 being always-used. Items for social influence, consumer novelty seeking, and green consumption behavior are also collected on a five-point scale (1 = strongly disagree, 2 = disagree, 3 = neither agree/disagree, 4 = agree, 5 = strongly agree). To enhance the interpretability of latent variables – social influence, consumer novelty seeking, and green consumption behavior, we totaled the results of their respective items into new columns. Consequently, after all data mining operations, we were left with total 375 observations and 18 variables – *preference for newspaper, preference for magazine, preference for television, preference for radio, preference for social media, gender, age, marital status, education, ethnicity, number of households, personal income, occupation, work category, social influence, consumer novelty seeking, and green consumption behavior*.

1: https://www.sciencedirect.com/science/article/pii/S2352340920311963?via%3Dihub#ecom0001
2: How much one is affected by the social factors? For instance, one believing people in its circle of friends highly value the environmental friendliness of a products.
3: How likely one seeks novelty in the products and services it consumes?
4: How often one uses a green product?

4. Exploratory Data Analysis

First, we changed the data type of all demographic variables into factor variables. Then we leveraged column of green consumption that contained the sum of the results of all 13 items on green consumption from questionnaire and divided it into three levels – Low, Medium, and High. Reponses' results from minimum value to 44 is defined as low green consumption, 45 to 55 is defined as medium green consumption, and 56 to maximum value is defined as high green consumption. We added these newly defined values into a new column and then factored them into ordered levels (low < medium < high). After this operation, we were left with 111 observations in low level, 208 observations in medium level, and 56 observations in high level.

Second, we checked for the relationship between all predictors and outcome variable. From the exploratory data analysis, we noticed a varying trend across green consumption levels for preference for newspaper, preference for magazine, preference for social media, social influence, and novelty seeking behavior.

Third, we looked at interactions among these predictors. Since we were also interested in age and education, we included these variables too in the interactions. We noticed a strong interaction between *newspaper and magazine, newspaper and social influence, newspaper and novelty seeking, newspaper and education, magazine and social media, magazine and social influence, magazine and novelty seeking, magazine and education, social influence and education,* and *social media and education*.

Last, we mean centered continuous variables in the dataset for smooth interpretation of the model coefficients.

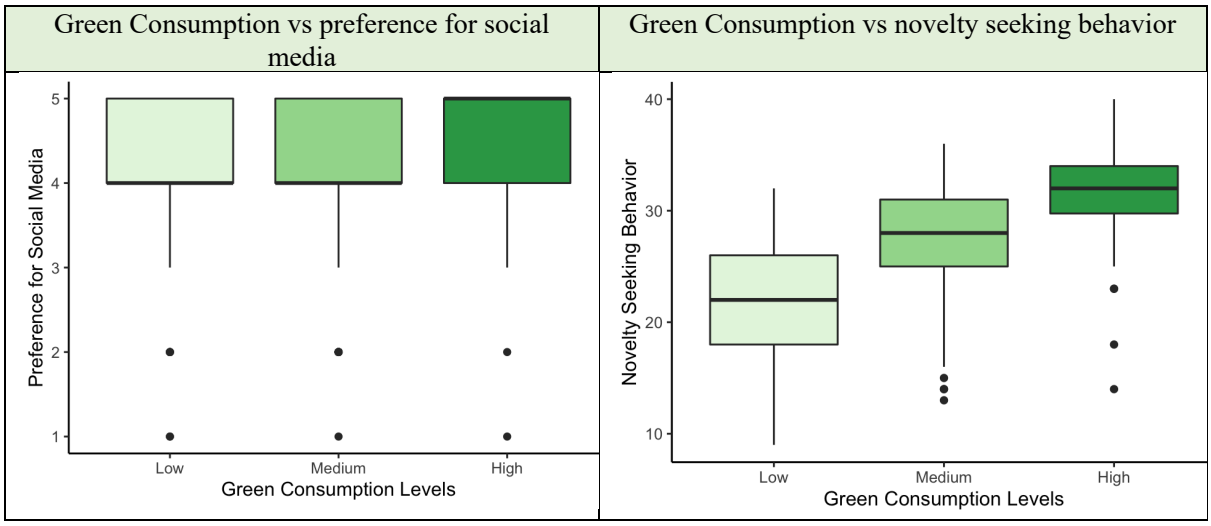


Figure 1: Different trend is observed for both graphs – Green Consumption vs preference for social media and Green Consumption vs novelty seeking behavior

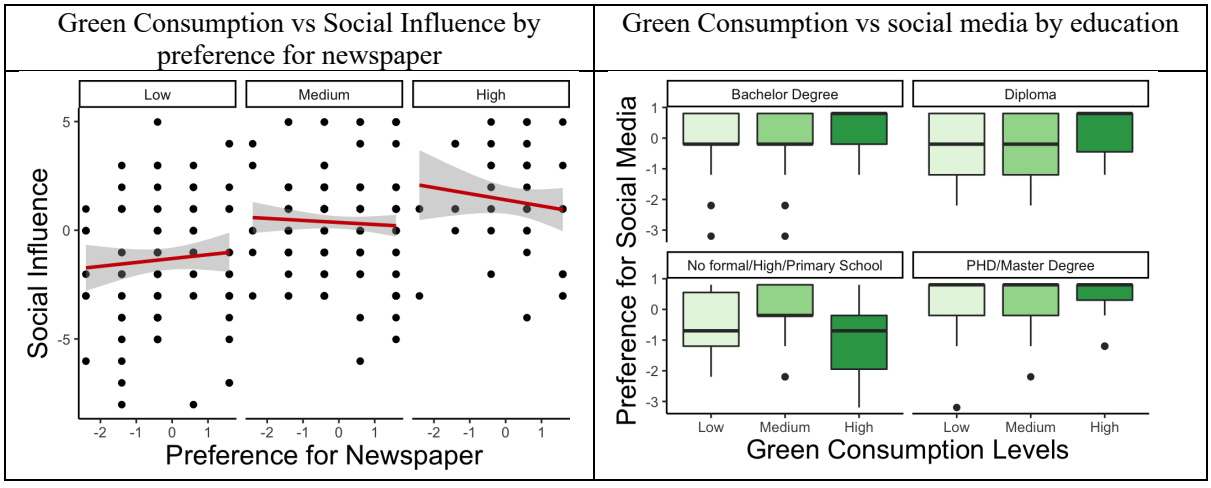


Figure 2: Interaction for both above graphs show varying trend across green consumption levels

5. Model Building and Selection

We started with fitting a proportional odds model using all predictors but no interactions. Then we ran multiple deviance tests, one at a time, to compare this model with a similar model but without a one specific predictor. From these tests, we found novelty seeking, social influence, number of households, ethnicity, newspaper, radio, and social media as statistically significant predictors. We then fitted second model using these significant predictors, along with age and education, but without interaction. Running a chi-square anova test on both models returned large p-value (0.60), indicating insignificance of additional predictors in the first model.

To test the significance of interactions, we then fitted third model using all predictors from second model and interactions among them. To select only the statistically significant interactions, we then ran both stepwise and forward aic test on third model. When compared the model returned by aic test (along with age and education variables) and second model (best model so far) via chi-square anova test, we found low p-value (0.01). This indicated the significance of interactions in third model. This allowed us to finalize the third model.

In addition to proportional odds model, we tried fitting a multinomial model using the same steps and approach used in building proportional odds model. Even though the accuracy of both proportional odds and multinomial models were same (68%), we went with proportional odds model because we wanted to incorporate low residuals and smooth interpretation in the final model. Consequently, we went with below proportional odds model as our final model:

$$\frac{\log(P_r[y_i <= j|x_i])}{\log(P_r[y_i > j|x_i])} = \beta_{0j} + \beta_{1j}\, Age_i + \beta_{2j}\, Education_i + \beta_{3j}\, NoveltySeeking_i + \beta_{4j}\, NumberofHousehold_i + \beta_{5j}\, SocialMedia_i + \beta_{6j}\, Radio_i + \beta_{7j}\, Newspaper_i + \beta_{8j}\, SocialInfluence_i + \beta_{9j}\, Ethnicity_i + \beta_{10j}\, SocialInfluence{:}Newspaper_i + \beta_{11j} SocialInfluence{:}Radio_i + \beta_{12j}\, NoveltySeeking{:}SocialMedia_i \; ; j = Low, Medium$$

6. Model Assessment

After fitting the model, we first validated the predicted probabilities for observed groups using the model. For instance, for two females with similar characteristics but different ethnicity, female with Malaysian ethnicity tend to have low probability of exhibiting high green consumption behavior than the one with non-Malaysian ethnicity.

Second, we checked the binned plot residuals against numerical variables, which are centered data for control seeking behavior. The points are all scattered randomly within the bands, except for one or two points. Since there was no trend in the residuals, linearity assumption was met.

Last, we checked for multi-collinearity which showed VIF scores for all explanatory predictors below 10 (~1). Overall accuracy of the model is 68%. Area under the curve for all three levels are – low (85%), medium (74.7%), and high (57.2%).

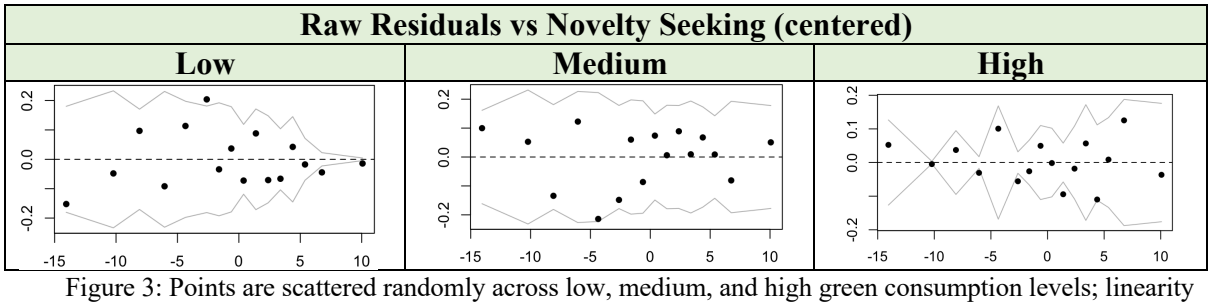


Figure 3: Points are scattered randomly across low, medium, and high green consumption levels; linearity assumption is met

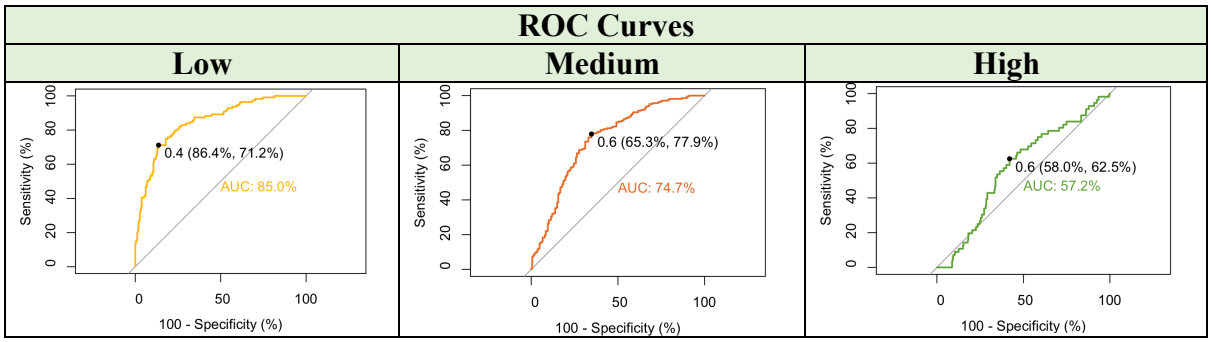


Figure 4: ROC curves for low, medium, and high green consumption levels.

7. Results

The model shows that preference for newspaper, preference for radio, preference for social media, novelty seeking behavior, number of households, ethnicity, social influence, and interaction between social influence and newspaper and between social influence and radio are all significant with p-value less than 5%.

From the model, we noticed that individuals who prefer newspaper to other media channels tend to have 2.17 times the odds of exhibiting high green consumption (vs. low or medium) compared to individuals who do not prefer newspaper (provided there is an effect of social influence on newspaper media consumption). On the contrary, preference for radio has a reversible effect on exhibiting high green consumption, i.e., individuals who prefer radio over other media channels are less likely to exhibit high green consumption. Model also shows an interaction of social media with novelty seeking behavior, but it is not significant at 5% level. Interestingly, it gets significant at 10% level, which allows us to answer the influence of social media on green consumption. At the 90% confidence interval, for someone who prefers social media and seeks novelty in its products tend to have odds of exhibiting high green consumption (vs. low or medium) in the range of 2.02 to 2.76 (on odds ratio scale: exp (0.103) + exp (-0.095), exp (0.569) + exp (-0.006)) times compared to someone who doesn't not prefer social media and doesn't seek novelty.

To answer second inferential questions, we couldn't leverage *age* and education predictor because of its low p-value, which we identified looking at inclusion of 0 in their confidence interval both at 90 and 95% levels.

Other interesting explanatory variables we discovered are *ethnicity* and *number of households*. For someone who is of non-Malaysian ethnicity tend to have 1.83 times the odds of exhibiting high green consumption (vs. low or medium) compared to someone who is of Malaysian ethnicity. Additionally, individuals who have more than 6 number of households have just 0.4 times the odds of exhibiting the high green consumption (vs. low or medium) compared to individuals who have 1-2 number of households.

	Value	Odds Ratio	Std. Error	t-value	95% Confidence Interval		90% Confidence Interval	
Intercept Low|Medium	-1.494	0.224	0.330	-4.531				
Intercept Medium|High	2.408	11.114	0.347	6.940	**2.5%**	**97.5%**	**5%**	**95%**
AgeAbove30yrs	0.090	1.094	0.246	0.367	-0.392	0.573	-0.314	0.495
EducationDiploma	0.106	1.112	0.302	0.350	-0.487	0.699	-0.391	0.603
EducationNoformal/High/PrimarySchool	0.190	1.210	0.374	0.508	-0.545	0.925	-0.427	0.806
EducationPHD/MasterDegree	0.533	1.705	0.340	1.568	-0.132	1.204	-0.025	1.096
NoveltySeeking	0.275	1.317	0.030	9.065	0.218	0.337	0.227	0.327
No_household3-5	-0.662	0.516	0.311	-2.129	-1.275	-0.054	-1.176	-0.152
No_householdAbove6	-0.910	0.403	0.360	-2.530	-1.621	-0.209	-1.506	-0.321
SocialMedia	0.334	1.396	0.141	2.363	0.060	0.614	0.103	0.569
Radio	-0.266	0.766	0.106	-2.512	-0.476	-0.060	-0.442	-0.093
SocialInfluence	0.143	1.154	0.059	2.438	0.029	0.260	0.047	0.241
Newspaper	0.235	1.265	0.111	2.127	0.020	0.453	0.054	0.418
EthnicityOthers	0.607	1.834	0.268	2.260	0.083	1.137	0.167	1.051
SocialInfluence:Newspaper	-0.102	0.903	0.046	-2.249	-0.193	-0.014	-0.178	-0.028
Radio:SocialInfluence	0.102	1.107	0.043	2.350	0.017	0.187	0.031	0.173
NoveltySeeking:SocialMedia	-0.049	0.952	0.027	-1.828	-0.103	0.003	-0.095	-0.006

Table 1: Summary Statistics for Proportional Odds Model

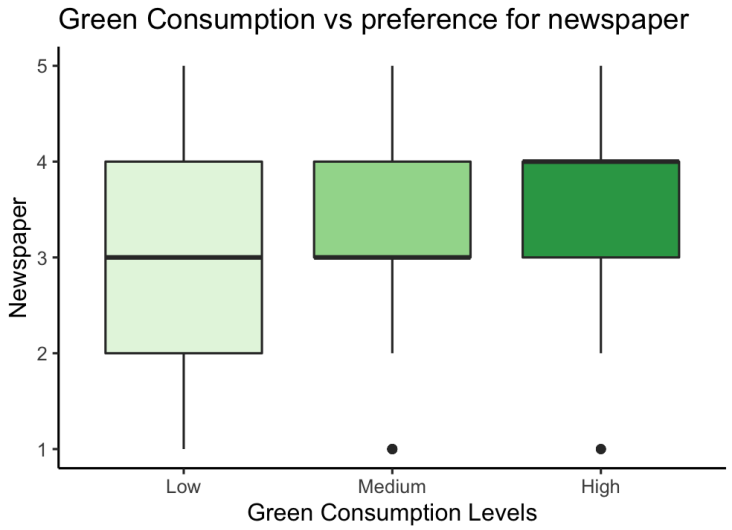
8. Conclusion

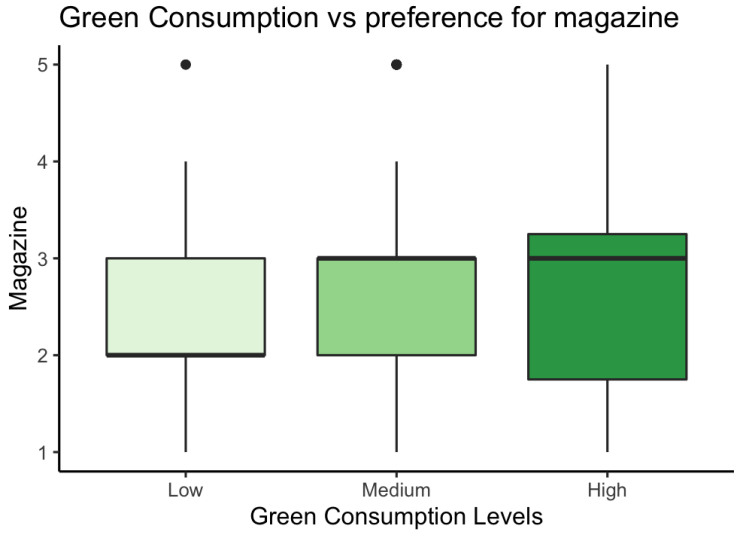
Overall, the model seems valid and helped us understand the influence of media preference and few demographic factors on green consumption in consumerism. Specifically, there seems to be a strong influence of *newspaper*, *social media*, *ethnicity*, and *number of households* on green consumption. Although we were able to identify significant explanatory predictors, there are several limitations in the study. For instance, we couldn't identify the influence of age and education on green consumption because of their low significance in the model. We also noticed a strong influence of number of households on green consumption. This may be from the fact that people who have more number of households are economically poor and cannot afford to consume green products. To assess this assumption, we tried including interaction between number of households and income as one of the explanatory parameters but couldn't leverage it in the model because of its low p-value. Thus, we believe that in the future studies we can try different combinations of age, education, and income levels to be able to include them as one of the explanatory variables.

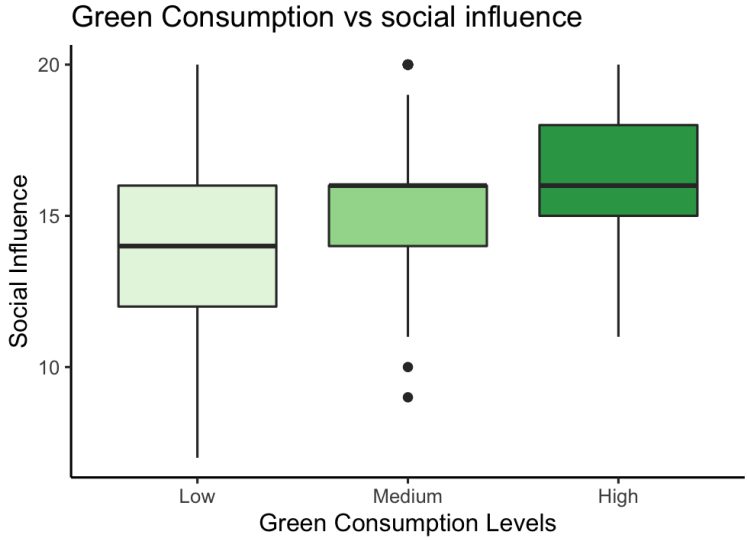
From the ROC curve, there is a scope for improvement in predictions for high green consumption. Only 16 out of 56 observations in high green consumption level can be explained by the model. Additionally, sensitivity and specificity can be improved upon in terms of a more balanced distribution among themselves across all green consumption levels. We also believe that the study should be extended globally to influence policy makers and companies across the world to be able to make informed decision at the global level. Nonetheless, the model does help us answer most of the questions in the study and can be used for future studies on the topic.

9. Appendices

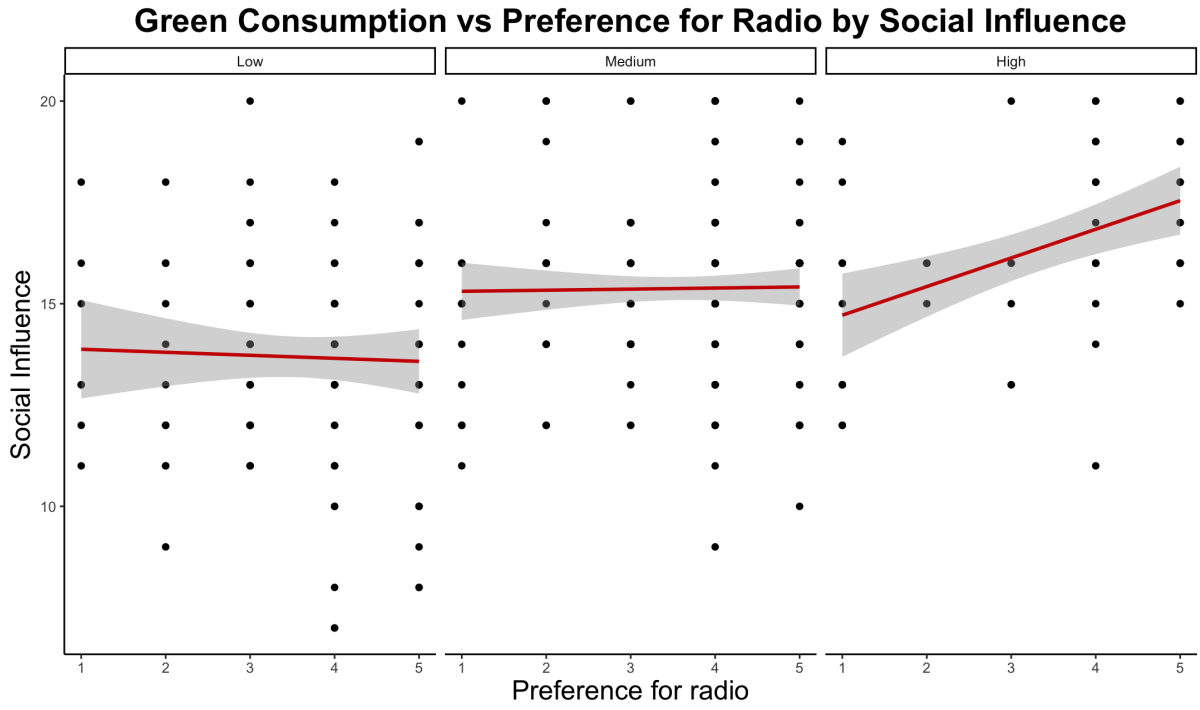
9.1 Exploratory Data Analysis

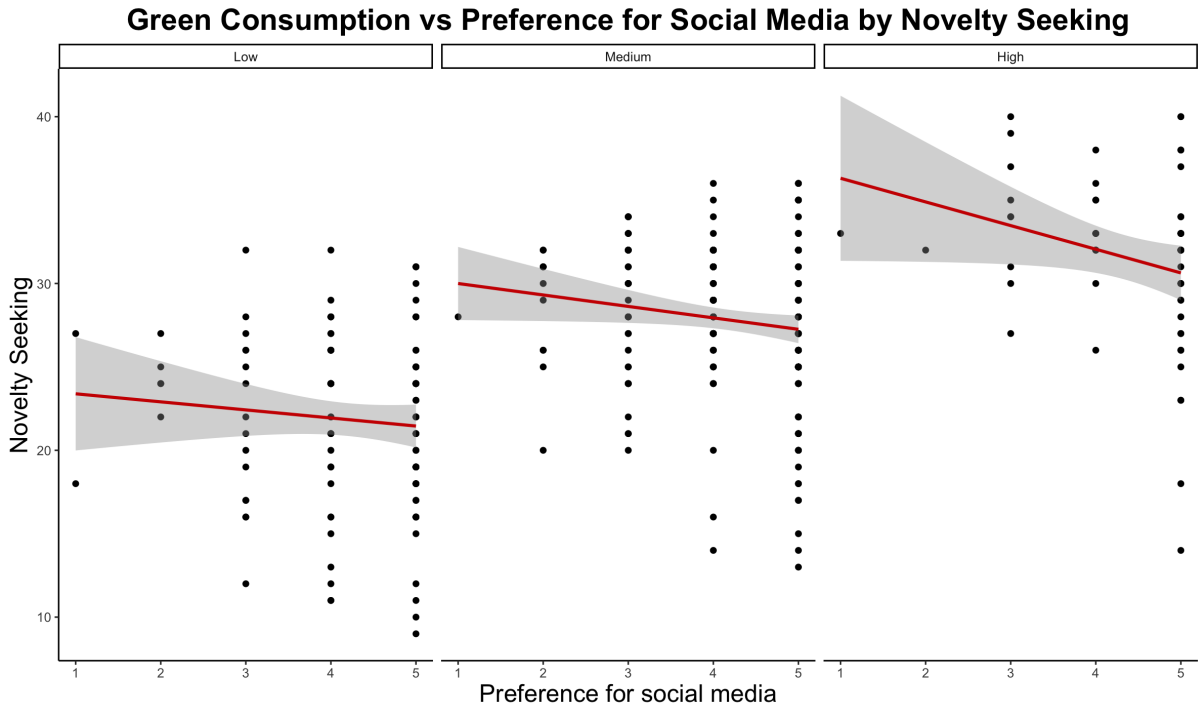






9.2 Exploratory Data Analysis (Interactions)





9.3 Dummy Dataset and Predicted Probabilities

New Data (Malaysian vs Others)		
ID	1	2
PM1	3	3
PM2	0	0
PM3	0	0
PM4	4	4
PM5	5	5
Gender	Female	Female
Age	18-30 yrs	Above 30 yrs
Marital_status	Single/Others	Married
Education	PHD/Master Degree	PHD/Master Degree
Ethnicity	Malaysian	Others
No_household	Above 6	3-5
Income	2001-4000	4001-6000
Occupation	Professional/Manager/Executive	Professional/Manager/Executive
Work_category	Arts/Media/Communication/Others	Arts/Media/Communication/Others
TotalEC	0	0
TotalSI	0	0
TotalCNS	0	0
TotalGC	48	51
GC_cat	Medium	Medium

Predictions			
Ethnicity	**Low**	**Medium**	**High**
Malaysian	0.08115758	0.7327657	0.1860768
Others	0.03318762	0.5964357	0.3703767

9.4 Confusion Matrix

Confusion Matrix			
	Low	Medium	High
Low	59	23	1
Medium	52	180	39
High	0	5	16

Confusion Matrix by Class		
	Sensitivity	Specificity
Low	0.532	0.909
Medium	0.865	0.455
High	0.286	0.984

9.5 Multi-collinearity

VIF Table			
	GVIF	Df	GVIF^(1/(2*Df))
Age	1.142	1	1.069
Education	1.229	3	1.035
TotalCNS	1.376	1	1.173
No_household	1.184	2	1.043
PM5	1.348	1	1.161
PM4	1.246	1	1.116
TotalSI	1.251	1	1.118
PM1	1.107	1	1.052
Ethnicity	1.077	1	1.038
TotalSI:PM1	1.158	1	1.076
PM4:TotalSI	1.085	1	1.042
TotalCNS:PM5	1.163	1	1.078