MKTG 212 - Case 4: Logistic Regression & Clustering Methods

Juan Manubens

University of Pennsylvania

1 Brand Awareness

There is a lot of interest in how TV viewing (and other demographics) relates to awareness of new brands. A new detergent brand CSG collected some data on households. The data is available on Canvas (Assignments / Assignment4 / Tv_viewing.txt). The variables in the dataset for a random sample of 500 households are as follows:

Variable Name	**Description**
HH Number:	Household ID
Hours TV:	# Hours of TV Viewing per week
DeterPur:	# of detergent purchases in the last calendar year
Gender:	Gender of Head of Household (1=Male, 0=Female)
Income:	Household Income in thousands of dollars
Aware CSG:	Awareness or not of new Brand CSG (1 = aware of CSG, 0 otherwise)

- **[1a] Which of the variables is the most significant *single predictor* (i.e. one variable at a time) of AwareCSG, awareness or not of CSG? State how you arrived at this conclusion? Is it a significant predictor at the 1% significance level?**

Run an appropriate regression with dependent variable = aware CSG and independent variables = Hours TV, DeterPur, Gender and Income. Based on this regression, answer the following questions

Table 1: Results - Single Predictor Models

Predictor Variable	β_j	$Pr(>\mid z \mid)$	Significant at $\alpha = 0.02$	Significant at $\alpha = 0.01$
x_1 : Hours TV	≈ 0.0821	$\approx 2.16 \cdot 10^{-8}$	Yes	Yes
x_2 : DeterPur	≈ 0.0272	≈ 0.58	No	No
x_3 : Gender	$\approx 4.01 \cdot 10^{-15}$	1	No	No
x_4 : Income	$\approx -1.96 \cdot 10^{-5}$	≈ 0.72	No	No

The most statistically significant single predictor is **'Hours TV'**, with $P(>\mid z \mid) \approx 2.16 \cdot 10^{-8}$. Thus, it is statistically significant at $\alpha = 0.01$ (see Table 1).

- **[1b] Which of the independent variables are significant predictors of "aware CSG" at the 2% significance level? State how you arrived at this conclusion based on your regression output.**

The only statistically significant single predictor at $\alpha = 0.02$ is **'Hours TV'** (see Table 1). See Appendix 4.1 for the R code output for these models.

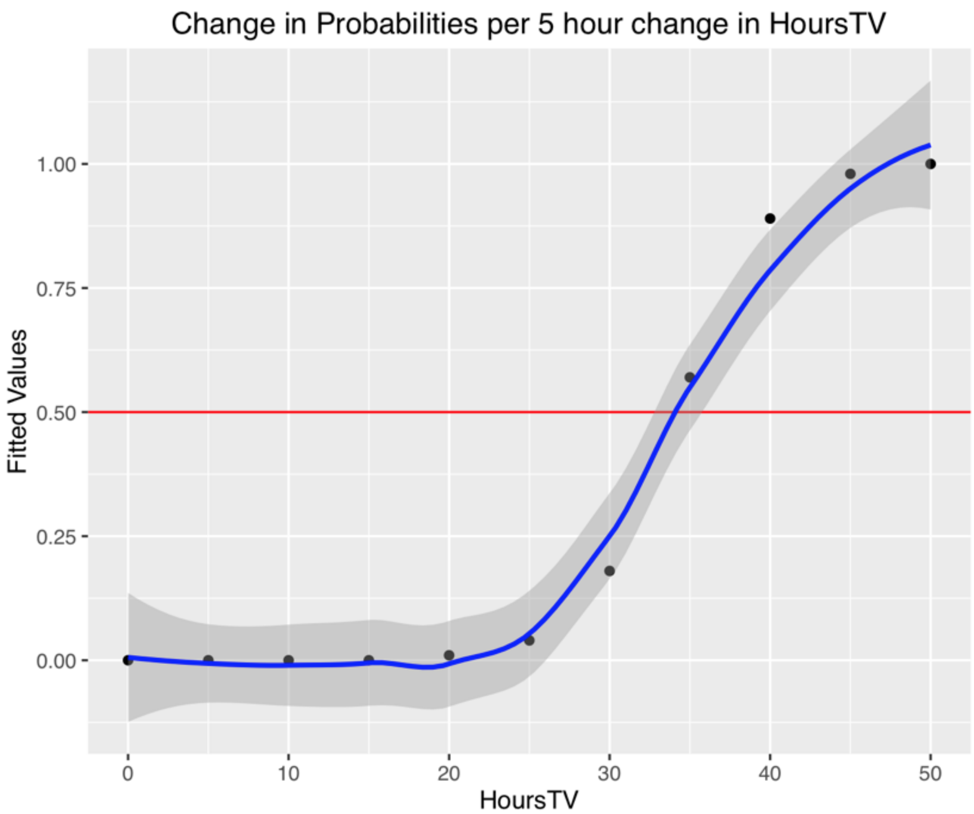
- **[1c] What is the probability of awareness for someone who watches 20 hours of TV, made 8 detergent purchases last year, is female, and makes $60,000?**

$$\hat{Y} = \hat{p}_{AwareCSG} = \hat{p}(x_1, x_2, x_3, x_4) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4}}$$

$$\Rightarrow \hat{p}_{AwareCSG} = \hat{p}(20, 8, 0, \$60k) = 0.006447225 \approx 0.64\%$$

See Appendix 4.2 for R code.

- **[1d] How does the probability of awareness change with more hrs of TV watching? Plot the probability of awareness as the amount of TV watching varies from 0-50 hrs per week. You can plot the probability of awareness using increments of 5 hrs (for the analysis, assume 8 detergent purchases last year, female, and income of $60,000).**



From the model coefficients, we know that a unit increase in **'Hours TV'** results in an increase of ≈ 0.35 in log-odds. As we can see in the plot above, there is a steep increase in $\hat{p}(20, 8, \Delta x_3, \$60k)$ for $x_3 > 25$, with $\hat{p} \geq 0.5$ when $x_3 \geq 34$ (See Appendix 4.3 for R code)

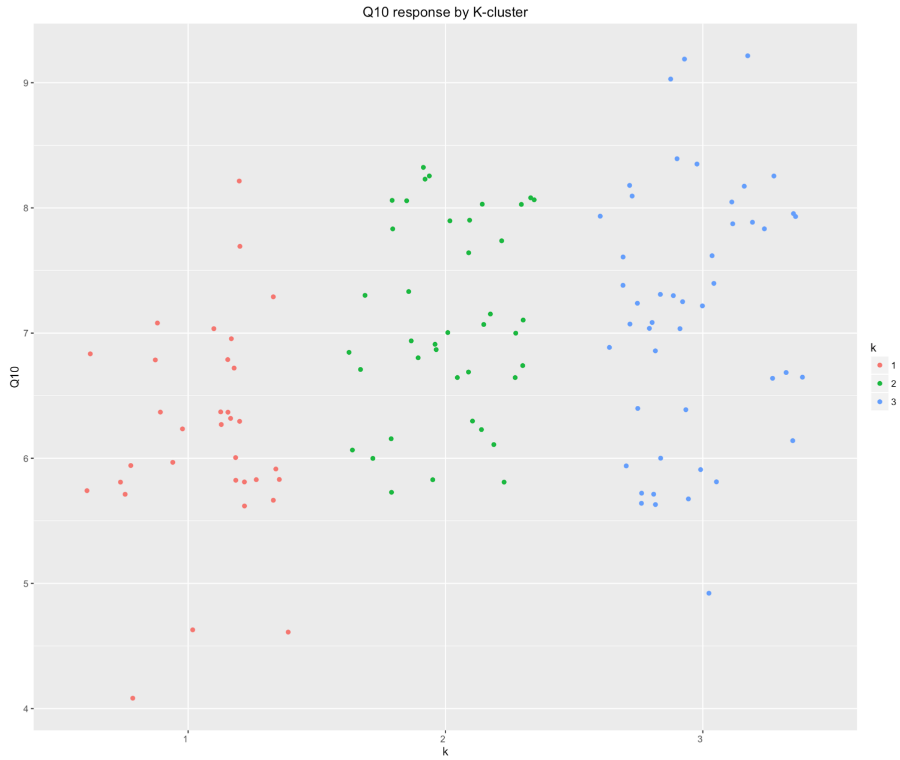
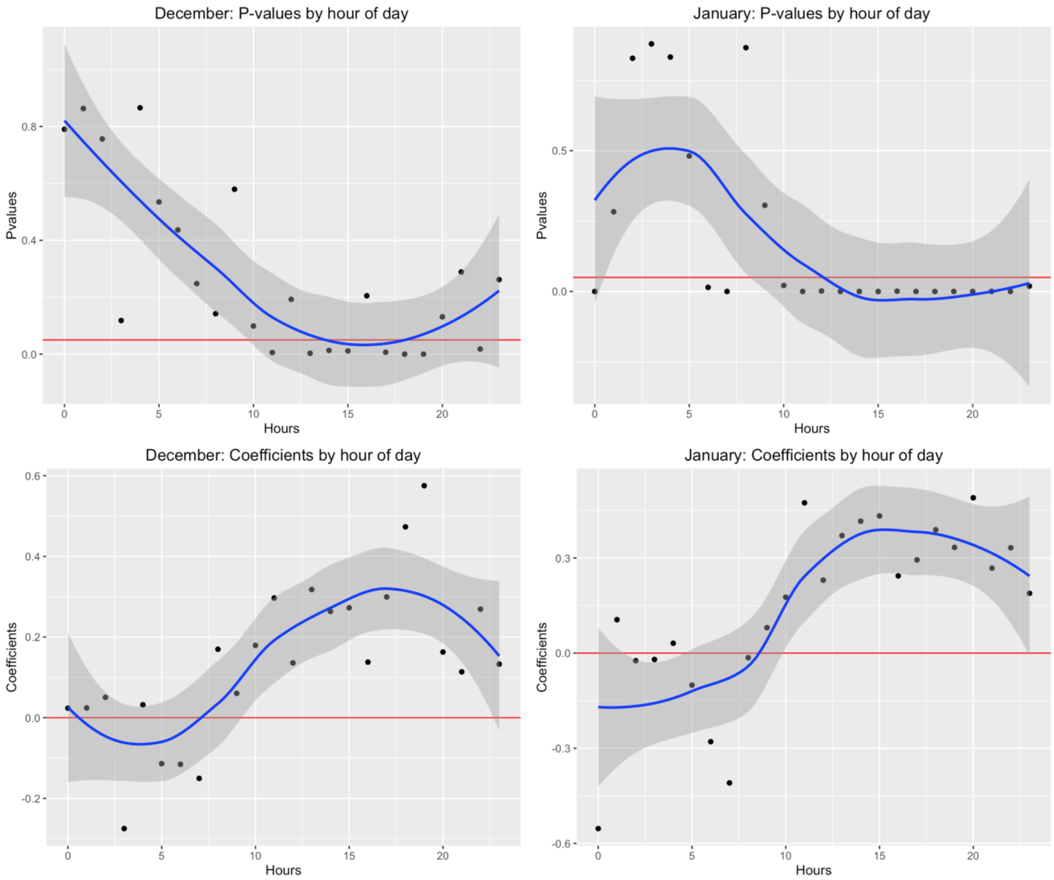
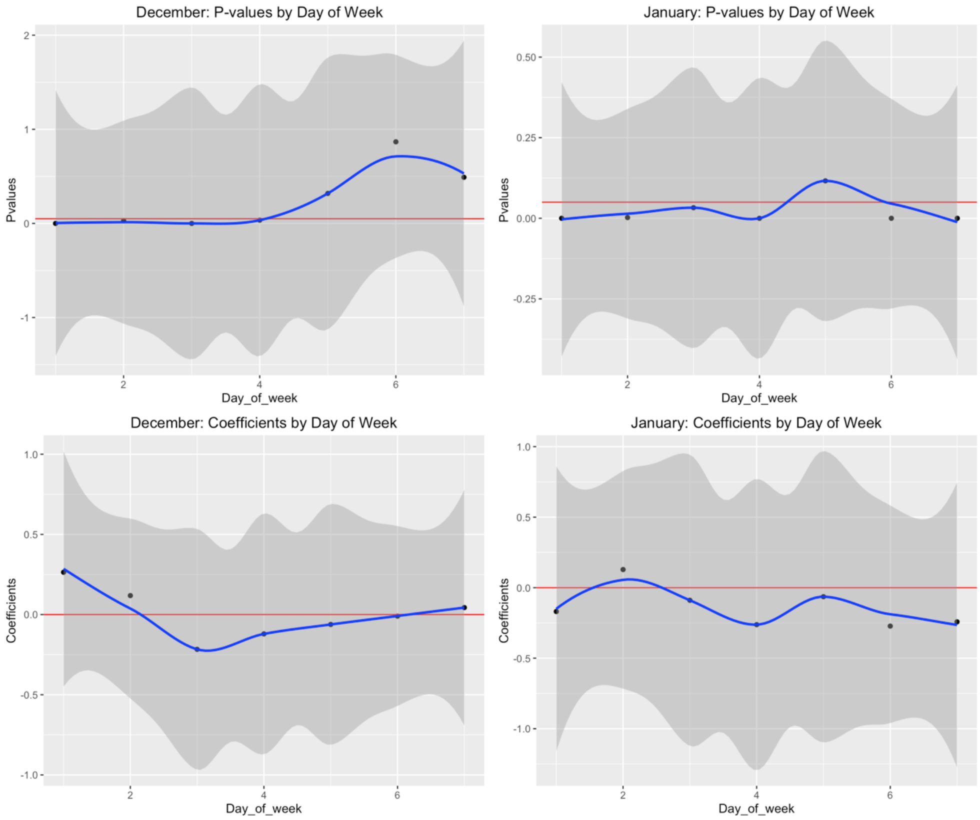
- **[1e] Do the independent variables as a whole, in your model, provide statistically significant explanatory power at 5% level for predicting the probability that someone is aware of CSG? State how you arrived at this conclusion based on your regression output.**

$$H_0 : \text{Logistic Regression Model with all variables adequately fits the data}$$

$$H_a : \text{Logistic Regression Model with all variables does not provide an adequate fit}$$

To measure goodness of fit, a Hosmer–Lemeshow (H.L.) test is a good method for this particular. An H.L. test is a statistical test for goodness of fit for logistic regression models, developed by Hosmer and Lemeshow (1980), frequently used in risk prediction models. It assesses whether or not the observed event rates match expected event rates in subgroups of the model population. H_0 can be redefined in this situation as *"actual and predicted event rates are similar across 10 deciles"*. If $p_{H.L.} > \alpha; \alpha = 0.05$, we reject H_0.

````
library("ResourceSelection")
hoslem.test(views$AwareCSG , views$predicted_q1ALL)
```

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  views$AwareCSG, views$predicted_q1ALL
X-squared = 7.5677783, df = 8, p-value = 0.4767865
````

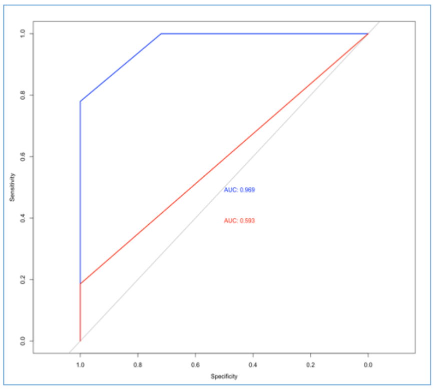
The test finds that our full model, as indicated by $p_{H.L.} \approx 0.477 >> 0.05$, ***does not properly fit the data***. In consequence, we should rethink our model, and include interactions to see if the fit improves.
````

- **[1f] State two ways that CSG can use the results of this regression equation in a managerial way. Be as specific as you can be.**

Table 2: Results - Full Model

Predictor Variable	β_j	$Pr(>\mid z \mid)$	Significant at $\alpha = 0.02$	Significant at $\alpha = 0.01$
x_1 : Hours TV	≈ 0.3531	$\approx 2.04 \cdot 10^{-21}$	Yes	Yes
x_2 : DeterPur	≈ -0.9967	$\approx 1.00 \cdot 10^{-16}$	Yes	Yes
x_3 : Gender	≈ -0.4792	$\approx 2.33 \cdot 10^{-2}$	No	No
x_4 : Income	$\approx -4.12 \cdot 10^{-4}$	$\approx 6.65 \cdot 10^{-9}$	Yes	Yes

The results of our full model are described in Table 2 above. The sorted coefficients in terms of their absolute values are $\beta_2 > \beta_3 > \beta_1 > \beta_4$. While X_2 is not significant at $\alpha = 0.02$, it is fairly close. See Appendix 4.4 for R code. This information overall can be used in several ways, such as:

1. As seen in 1(d) and as reflected by β_1, on average, people who watch more TV (controlling for other factors) are much more likely to be aware of the new detergent brand. ***Thus, based on this information, marketing campaigns should target demographics with high exposure to media through television (reflected by a high X_1).***
2. As reflected by β_3, controlling all factors but gender, men are less likely to be aware of new detergent brands. ***Thus, based on this information, marketing campaigns should target women either exclusively or to a much larger extent than men.***

- [1g] **Managers also care about how two variables may interact with other each. Include an interaction between the two variables "Gender" and "TV watching" in the above model and estimate the regression coefficients. Describe your results. Using the coefficients, calculate the following:**

Table 3: Results - Full Model with interactions between 'HoursTV' and 'Gender'*

Predictor Variable	β_j	$Pr(>\mid z \mid)$	Significant at $\alpha = 0.02$	Significant at $\alpha = 0.01$
x_1 : Hours TV	≈ 0.3114	$\approx 3.25 \cdot 10^{-15}$	Yes	Yes
x_2 : DeterPur	≈ -0.9908	$\approx 2.8 \cdot 10^{-16}$	Yes	Yes
x_3 : Gender	≈ -2.530	$\approx 1.08 \cdot 10^{-3}$	Yes	Yes
x_4 : Income	$\approx 4.16 \cdot 10^{-4}$	$\approx 1.01 \cdot 10^{-8}$	Yes	Yes
$x_5* : x_1 \leftrightarrow x_3$	≈ 0.093	≈ 0.0056	Yes	Yes

The results of our full model with interactions are described in Table 3 above. The sorted coefficients in terms of their absolute values are $\beta_3 > \beta_2 > \beta_1 > \beta_5 > \beta_4$. After including the interaction between X_1 and X_3, this model adds more weight to the effect of gender on the log-odds of $\hat{Y} = \hat{p}_{AwareCSG}$. Running another H.L. test gives us $p_{H.L.} \approx 0.1494$, showing that fit has improved compared to the previous model, but is still not sufficient (a model with all interactions, $x_1 \leftrightarrow x_2, x_3, x_4$ outputs $p_{H.L.} < 0.05$). Nevertheless, for the purpose of this exercise, this model will suffice. See Appendix 4.4 for the corresponding R code.

1. **What is the probability of awareness for someone who watches 20 hours of TV, made 8 detergent purchases last year, is female, and makes $60,000?**

$$\Rightarrow \hat{p}_{AwareCSG} = \hat{p}(20, 8, 0, \$60k) = 0.006994688203 \approx 0.70\%$$

2. **What is the probability of awareness for someone who watches 20 hours of TV, made 8 detergent purchases last year, is male and makes $60,000?**

$$\Rightarrow \hat{p}_{AwareCSG} = \hat{p}(20, 8, 1, \$60k) = 0.003592422942 \approx 0.36\%$$

As we can see in this particular example, our model predicts that the probability of awareness for a woman who watches 20 hours of TV, made 8 detergent purchases last year, is female, and makes $60,000 is almost twice that of a man with the same characteristics.

2 Survey Analysis

We will revisit a dataset that we used in the previous assignment – the survey of a sample of supermarket owners for their opinions on the raisin industry, and conduct a deeper analysis.

- **[2a] Run a 3-means clustering on $Q_1, ..., Q_9$. What appear to be the significant drivers of variation between the groups? How would you statistically test what the significant drivers are between groups?**




```
K-means clustering with 3 clusters of sizes 33, 40, 47

Cluster means:
           Q1          Q2          Q3          Q4          Q5          Q6          Q7          Q8          Q9
1 3.787878788 4.030303030 4.393939394 4.606060606 4.515151515 4.848484848 5.030303030 5.787878788 5.939393939
2 3.250000000 3.125000000 3.650000000 4.000000000 5.750000000 5.925000000 6.200000000 6.625000000 6.575000000
3 4.851063830 5.127659574 5.425531915 5.617021277 5.659574468 5.936170213 6.446808511 6.446808511 6.978723404
```

We ran a 3-means clustering on $X : Q_1, ..., Q_9$ (See Appendix 4.6). The results and a plot of $Y : Q_{10}$ per each k-cluster show differences in X, Y. At a glance, some question's mean response per cluster do not seem to follow the same ordinal ranking as the response. Q_4, *"Giving away in-package free gifts is a strong driver of brand sales."*, for instance, is predictive when regressing the joint sample, but this could change once we run a regression on each cluster - comparing the t-values would allow us to statistically test what the significant drivers are between clusters.

- **[2b] For each of the 3 segments, which variables (out of $Q_1, ..., Q_9$) are significant predictors, at the 10% significance level, of Q10, the overall profitability in the raisin category? Clearly, describe how the 3 segments are different in terms of what matters for profitability. Also, compare the above results with those from the entire dataset. What are the big differences?**

	t-values									Correlations with Q10								
Cluster	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**
1	0.10	0.83	0.88	0.00	0.57	0.38	0.22	0.00	0.02	-0.23	-0.06	0.08	0.48	-0.06	0.28	0.23	0.41	0.13
2	0.11	0.13	0.03	0.00	0.60	0.71	0.28	0.00	0.00	0.09	0.21	0.23	0.38	0.05	-0.07	0.00	0.51	0.53
3	0.77	0.19	0.45	0.05	0.78	0.80	0.01	0.03	0.00	0.00	-0.08	-0.10	0.16	0.31	0.15	0.24	0.28	0.63
All	0.00	0.00	0.06	0.00	0.35	0.35	0.00	0.00	0.00	0.05	0.04	0.07	0.26	0.35	0.31	0.37	0.46	0.57

	Significant at 10%									Correlation Direction (1: + ; 0: -)								
Cluster	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**
1	1	0	0	1	0	0	0	1	1	0	0	1	1	0	1	1	1	1
2	0	0	1	1	0	0	0	1	1	1	1	1	1	1	0	1	1	1
3	0	0	0	1	0	0	1	1	1	1	0	0	1	1	1	1	1	1
All	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1

We ran separate OLS regressions for $X : Q_1, ..., Q_9$, for each of the $k = 1, 2, 3$ clusters and for the joint data. The results are summarized on the tables above. The first table shows the t-values and correlations with Q_10 for each cluster and for all clusters. The second table indicates whether or not Q_j is statistically significant for a given k or all k's, with $X_{k,Q_j} = [1 : \alpha > Pr(>| t_{k,Q_j} |)]; 0 : \alpha < Pr(>| t_{k,Q_j} |)$ for $\alpha = 0.1$, and whether the correlation is positive or negative. We make the following observations:

1. Q_4,Q_8 and Q_9 are significant at $\alpha = 0.1$ for any k as well as for the joint data, and Q_5 and Q_6 are not significant at $\alpha = 0.1$ for the same groups.

2. While Q_2 is statistically significant at $\alpha = 0.1$ for the joint data, it is not significant in any of the k-cluster regressions.

3. **Thus, the likely drivers of variation between the k-clusters are Q_1 for $k = 1$, Q_3 for $k = 2$, and Q_7 for $k = 3$, as indicated in the second table with the cells outlined in bold.**

Now that we have identified our likely drivers, we need to interpret the implications for each cluster based on the available information. We can make some inferences regarding the differences between clusters in Supermarket Owner's strategies, based on the t-values, coefficients and correlations from our different regression models:

1. For owners in $k = 1$, the more value the raisins' color, on average, the lower their Q_{10} response. This is evidenced by the t-value, the negative correlation, and $\beta_{k=1,Q_1} \approx -0.3 < 0$. We could interpret this as a sign of these owner's over-emphasizing details of their merchandise that have a small effect on their sales, explaining the disconnect with profitability.

2. On average, owners in $k = 2$ that signal caring about offering a variety of package sizes, or allow customers to purchase custom quantities, charging per pound (or kilo) sold, for example, will tend to have a higher Q_{10} response. This is evidenced by the t-value, the positive correlation, and the coefficient, $\beta_{k=2,Q_3} \approx 0.284$.

3. On average, owners in $k = 3$ that see raisins as sales catalysts (complementary to the sales of other fruits), will tend to have a higher Q_{10} response. This is evidenced by the t-value, the positive correlation, and the coefficient, $\beta_{k=3,Q_7} \approx 0.24$.

There are several caveats to this analysis - small sample size and non-standardized values, for instance, do not allow us to compare coefficients directly. But there is clear value to be gained on using these statistical methods on larger samples.

- **[2c] Based on the above comparison, please explain (as simply as possible) why regression analysis and segmentation together can provide a lot more insight into underlying drivers as opposed to a single regression model for the entire market place.**

As we see in the results and the analysis on 2(a) and 2(b), there are non-negligible differences between the joint analysis of Storeowner's responses and the cluster-level analyses. Performing regression analysis and segmentation together introduces a higher level of heterogeneity to the exercise, allowing for deeper insights and consequently, more precise marketing. It also helps avoid reaching the wrong conclusions (and acting on them).

3 Customer Engagement Data

We are very interested in understanding what the variables are for predicting whether a customer will engage with an ad or email ("Engagement"). Please answer the following questions.

- **[3a] How does the timing of the impression (when it is sent) impact the probability that a customer will engage? How does the probability of engagement vary by month, time of day, day of the week? In other words, please document the temporal variation of engagement. This analysis can help managers decide when to send an impression.**

We split the data set by month to compare and contrast, and ran logistic regressions predicting Y : Engagement with $X_{(i)}$: Hour of Day and $X_{(ii)}$: Day of the Week. We then ran a logistic regression with $X_{(iii)}$: By Month using the joint data.

(i) Engagement by Hour of Day:



(ii) Engagement by Day of the Week:



(iii) Engagement by Day of the month:

```r
```{r}
# By Month
glm.3a.M.lm  <- glm(Engagement ~ impression_month,family = binomial(link = "logit"), data = annalect.t)
glm.3a.M.summary<- summary(glm.3a.M.lm);
#annalect.t$impression_month %>% unique %>% sort
df.M <- data.frame(Month=c(12), Pvalues= coef(glm.3a.M.summary)[,4], Coefficients=coef(glm.3a.M.summary)[,1])
df.M
```
```

	Month <dbl>	Pvalues <dbl>	Coefficients <dbl>
(Intercept)	12	1.669330e-170	-0.2901814
impression_month12	12	1.263676e-176	0.5141336

Through the scatter plots showing how engagement varied at different times of day, for both January and December separately, we can identify the key times of day to increase probability of engagement. It seems as though engagement is generally more variable and higher at peak hours in January than in December. For both months, late afternoon (6:00 pm - 8:00 pm) was the time of highest engagement, and engagement increased markedly after around 10:00 am on both months. The lowest engagement was found at midnight in January, and around 4:00 am in December. In both December and January the trend is a general increase in engagement as the 24-hour day progresses, but there's a dip in engagement in the early hours of the morning (before about 7:00 am).

It's also useful to look at how engagement varied by day of the month. For both December and January, engagement was highest at the beginning of the month and declined all the way to the end (with a slight bump up in the last five or so days of the month). However, the first five days of January have a steep decline that is only seen from the first day to the second day of December.

Finally, we examine variation by day of the week. In December, engagement rapidly decreases from Sunday to Monday and Monday to Tuesday, reaching its low on Tuesday. For there there is a steady increase in engagement until the end of the week. Engagement by weekday is more variable in January - engagement is at its highest on Monday and decreases to a Wednesday low before spiking back up to near Tuesday levels on Thursday, reaching a low again on Friday, and increasing slightly through Sunday. From this data we can draw the conclusion that the best time to send an impression would be around 7:00 pm on a Sunday at the beginning of December or on a Monday at the beginning of January.

- **[3b] We also want to understand the impact of creatives on engagement. Since there are many creatives (which you can check if you assess the distribution), it will be difficult to carry out an analysis for every creative. Thus, I would like you to focus on the creatives that were sent to more than 5% of people (each line in the dataset is a person). Consider all creatives that are sent to less than 5% of people as "the baseline level". Of the creatives that were sent to more than 5% of people, I would like you to quantify how the creative impacts the probability of their engagement (whether they engage or not).**

1. **Are some creatives better than others for engaging customers? If so, which ones?**

2. **Are email-based creatives better than other types of creatives? Please summarize your findings about the importance of different types of creatives as best as you can.**

Please note that this last question is a bit open ended and not as structured (capturing the spirit of what clients typically ask from companies doing analytics). I am interested in seeing how you approach the problem.

3b(1) Impact of creatives on Engagement

````
```{r}
which( 100*sort((table(annalect.c$creative_name)/nrow(annalect.c)), decreasing = TRUE) > 5)
cID2  <- c("Secure DMP", "Direct" , "Psearch_Other" , "Secure_DMP_Pixel" , "Email_Past" , "Guest Sale" ,  "Osearch_Google" )
cID2.f <-  function(i){ which(annalect.c$creative_name == cID2[i]) };cID2.index <- lapply(1:length(cID2), cID2.f) %>% unlist %>% as.vector
annalect.c$creativenames.comps <- annalect.c$creative_name; annalect.c[-c(cID2.index),8] <- rep("Baseline",(nrow(annalect.c)-length(cID2.index)))
100*sort((table(annalect.c$creativenames.comps)/nrow(annalect.c)),decreasing = TRUE);
annalect.c$creativenames.comps <- annalect.c$creativenames.comps  %>% as.factor
```

      Secure DMP        Baseline          Direct   Psearch_Other Secure_DMP_Pixel   Osearch_Google
       30.827390       25.232810       23.168356        7.373049        7.301553        6.096842
````

````
```{r}
# Compare creatives against baseline
glm.3bi2  <- glm(Engagement ~ creativenames.comps   , family = binomial(link = "logit"), data = annalect.c)
glm.3bi2.summary <- summary(glm.3bi2);  glm.3bi2.summary
#                                    Estimate Std. Error z value Pr(>|z|)
#(Intercept)                         -0.36308    0.01711 -21.219   <2e-16 ***
#creativenames.compsDirect           20.92915  155.73341   0.134    0.893
#creativenames.compsOsearch_Google   20.92915  303.58264   0.069    0.945
#creativenames.compsPsearch_Other    20.92915  276.06149   0.076    0.940
#creativenames.compsSecure DMP      -20.20299  135.00847  -0.150    0.881
#creativenames.compsSecure_DMP_Pixel -20.20299  277.40978  -0.073    0.942
```
````

We created a custom variable, $'creativenames.comps'$, where all $'creative\_names'$ aggregating less than 5% of engagements are encoded as "*Baseline*". None of the non-baseline creatives have significant p-values, but $\beta_{0,3b(1)}$ is quite significant. This was quite odd at first, and we will explain why we think this is the case after our results for **3b(2)**.

3b(2) Impact of email-based creatives on Engagement

````
```{r}
annalect.c$email <- (annalect.c$creative_description == "Email" )  %>% as.numeric
annalect.c$email <- annalect.c$email  %>% as.factor
glm.3bii  <- glm(Engagement ~ email   , family = binomial(link = "logit"), data = annalect.c)
glm.3bii.summary <- summary(glm.3bii);  glm.3bii.summary
#            Estimate Std. Error z value Pr(>|z|)
#(Intercept) -0.326107   0.008969 -36.361   <2e-16 ***
#email1      17.892176  56.609362   0.316    0.752
```
````

We created a custom variable, $'email'$, where all email-based creatives are encoded as **"1"** and the remaining 44 types are encoded as **"0"**. Just as in **3b(1)**, none of the email-based creatives have significant p-values, but $\beta_{0,3b(2)}$ is significant. We believe this is due to a problem with the data, as we will explain in the next page.

3b Discussion: possibility of an unrepresentative sample

By Creative Category, Email Category					
Creative Category	Email	Y	Y=1 per Category	Total	Y=1 %
Baseline	0	0	0	8326	0%
Baseline	0	1	907	907	100%
Baseline	1	1	4884	4884	100%
Direct	0	1	12962	12962	100%
Osearch_Google	0	1	3411	3411	100%
Psearch_Other	0	1	4125	4125	100%
Secure DMP	0	0	0	17247	0%
Secure_DMP_Pixel	0	0	0	4085	0%
% Total Y=1	47%				



By Creative Category				
Creative Category	Y	Count	Total	Y=1 %
Baseline	0	8326	14117	41%
	1	5791		
Direct	0	0	12962	100%
	1	12962		
Osearch_Google	0	0	3411	100%
	1	3411		
Psearch_Other	0	0	4125	100%
	1	4125		
Secure DMP	0	17247	17247	0%
	1	0		
Secure_DMP_Pixel	0	4085	4085	0%
	1	0		

By Email Category				
Email Category	Y	Count	Total	Y=1 %
0	0	29658	51063	42%
	1	21405		
1	0	0	4884	100%
	1	4884		

```
annalect.c$base0 <- (annalect.c$creativenames.comps != (annalect.c$creativenames.comps %>% unique)[1]) %>% as.numeric
glm.test <- glm(Engagement ~ email + base0  , family = binomial(link = "logit"), data = annalect.c)
glm.test.summary <- summary(glm.test);  glm.test.summary
#              Estimate Std. Error z value Pr(>|z|)
#(Intercept)  -2.21700    0.03497 -63.404   <2e-16 ***
#email1       19.78306   56.60937   0.349    0.727
#base0         2.17712    0.03631  59.962   <2e-16 ***
```
```

By aggregating the results via R's *'dplyr'* package, we are able to get a sense of what the problem could be. The first red flag is the proportion of engagements, $\frac{\sum_{i=1}^{55947} Y_i = 1}{55947} \approx 47\%$. It seems unrealistic to expect almost half of the exposures in a representative sample to lead to engagements, so either the dataset does not contain random users, or contains too few observations to be representative. It is extremely unlikely that non-baseline creative have either 0% or 100% success rates in a representative sample. The same applies to email-based creatives having a 100% success rate in a representative sample. We note that only baseline creatives with $< 5\%$ of activity sent email advertisements. Running a final logistic regression on comparing baseline versus non-baseline as well as email and non-email as binary predictor variables, tells us is that email has a higher success rate for baseline creatives, as we can see by the regression results. The fitted values for baseline creatives, $\hat{p}_B$ are $\hat{p}_{B,E} \approx 99.9\%$ and $\hat{p}_{B,NE} \approx 9.8\%$ for email and non-email based types, respectively, and $\hat{p}_{NB} \approx 49\%$ for non-baseline creatives (none of which sent email-ads). So while it seems that email works better as a channel for $\approx 25.2\%$ of the sample (baseline creatives), our doubts regarding the quality of the data do not allow us to conclude whether or not non-baseline creatives necessarily outperform baseline ones.
```

4 Appendix

4.1 R Outputs for Q1(a,b)

```r
q1i.col2 <- glm(AwareCSG ~  HoursTV, family = binomial(link = "logit"), data = views)
q1i.col2.summary <- summary(q1i.col2); q1i.col2.summary$coefficients
#                Estimate Std. Error   z value     Pr(>|z|)
#(Intercept) -1.78006428 0.32995509 -5.394868 6.857378e-08
#HoursTV      0.08212197 0.01466808  5.598686 2.159822e-08
q1i.col3 <- glm(AwareCSG ~  DetergentPur, family = binomial(link = "logit"), data = views)
q1i.col3.summary <- summary(q1i.col3); q1i.col3.summary$coefficients
#                 Estimate Std. Error    z value  Pr(>|z|)
#(Intercept)  -0.08413269 0.17509974 -0.4804844 0.6308830
#DetergentPur  0.02719424 0.04865845  0.5588802 0.5762435
q1i.col4 <- glm(AwareCSG ~  as.factor(Gender), family = binomial(link = "logit"), data = views)
q1i.col4.summary <- summary(q1i.col4); q1i.col4.summary$coefficients
#                         Estimate Std. Error       z value Pr(>|z|)
#(Intercept)         -2.025748e-15  0.1264911 -1.601494e-14        1
#as.factor(Gender)1  4.011775e-15  0.1788854  2.242651e-14        1
q1i.col5 <- glm(AwareCSG ~  Income, family = binomial(link = "logit"), data = views)
q1i.col5.summary <- summary(q1i.col5); q1i.col5.summary$coefficients
#                  Estimate   Std. Error    z value  Pr(>|z|)
#(Intercept)  1.216290e-02 0.0955763115  0.1272585 0.8987358
#Income      -1.957972e-05 0.0000543336 -0.3603611 0.7185771


q1all <- glm(AwareCSG ~  HoursTV + DetergentPur + Gender + Income, family = binomial(link = "logit"), data = views)
q1all.summary <- summary(q1all); q1all.summary$coefficients
#                   Estimate   Std. Error   z value      Pr(>|z|)
#(Intercept)  -4.1019947073 4.836342e-01 -8.481605 2.221078e-17
#HoursTV       0.3531541866 3.716322e-02  9.502787 2.043471e-21
#DetergentPur -0.9967498367 1.200227e-01 -8.304675 1.000932e-16
#Gender       -0.4791504419 2.112639e-01 -2.268018 2.332809e-02
#Income       -0.0004121295 7.106113e-05 -5.799648 6.645422e-09
```

4.2 R Outputs for Q1(c)

```r
q1all <- glm(AwareCSG ~  HoursTV + DetergentPur + Gender + Income,
  family = binomial(link = "logit"), data = views)
q1all.summary <- summary(q1all); q1all.summary$coefficients
#                   Estimate   Std. Error   z value      Pr(>|z|)
#(Intercept)  -4.1019947073 4.836342e-01 -8.481605 2.221078e-17
#HoursTV       0.3531541866 3.716322e-02  9.502787 2.043471e-21
#DetergentPur -0.9967498367 1.200227e-01 -8.304675 1.000932e-16
#Gender       -0.4791504419 2.112639e-01 -2.268018 2.332809e-02
#Income       -0.0004121295 7.106113e-05 -5.799648 6.645422e-09

q1c <- data.frame( HHNumber = c("q1d"), HoursTV=c(20),
  DetergentPur=c(8), Gender=c(0), Income=c(60)  )
CIpred.q1c <- predict(q1all, newdata = q1c,  type = "response");
CIpred.q1c # 0.006447225
```

4.3 R Outputs for Q1(d)

```r
q1d <- data.frame( HHNumber = rep(c("q1d"),11), HoursTV= seq(0,50,5), DetergentPur= rep(c(8),11), Gender= rep(c(0),11), Income = rep(Income=c(60),11)  )
q1d$pred_1d <- round(predict(q1all, newdata = q1d,  type = "response") , digits = 2)

q1d %>% ggplot(aes(x=HoursTV,y=pred_1d)) + geom_point() + geom_abline(slope = 0, intercept = 0.5, col = "red") + stat_smooth(method = "loess", col = "blue") + ylab("Fitted Values") + ggtitle("Change in Probabilities per 5 hour change in HoursTV")

# Get the fitted default probability
views$predicted_q1ALL <- predict(q1all, type = "response")
ROCALL1all <- roc(AwareCSG ~ predicted_q1ALL, data = views) # Calculate the ROC curve
plot(ROCALL1all)
#Data: predicted_q1ALL in 250 controls (AwareCSG 0) < 250 cases (AwareCSG 1).
#Area under the curve: 0.7766
```

4.4 R Outputs for Q1(f,g)

```r
```{r}
q1full.int <- glm(AwareCSG ~  HoursTV + DetergentPur + Gender + Income
  + HoursTV*Gender , family = binomial(link = "logit"), data = views)
q1full.int.summary <- summary(q1full.int); q1full.int.summary$coefficients
#                      Estimate       Std. Error       z value                      Pr(>|z|)
#(Intercept)    -3.231890532898 0.55873061021974 -5.784344859 0.00000000727954689033055810
#HoursTV         0.311369072746 0.03950920798393  7.880924185 0.00000000000000324968447400
#DetergentPur   -0.990767510674 0.12109629359348 -8.181650167 0.00000000000000027998289990
#Gender         -2.530082223654 0.77405994389859 -3.268586940 0.00108085965444370711227570
#Income         -0.000415596951 0.00007253816788 -5.729355499 0.00000001008129238407265380
#HoursTV:Gender  0.093016885269 0.03354975924882  2.772505298 0.00556266138830633413009390

# Get the fitted default probability
views$predicted_q1full.int   <- predict(q1full.int, type = "response")
ROCfull.int <- roc(AwareCSG ~ predicted_q1full.int, data = views) # Calculate the ROC curve
plot(ROCfull.int) #0.777648
q1g <- data.frame( HHNumber = c("q1d","q1d"), HoursTV=c(20,20), DetergentPur=c(8,8),
  Gender=c(0,1), Income=c(60)  )
CIpred.q1g <- predict(q1full.int, newdata = q1g,  type = "response"); CIpred.q1g
#              1              2
#0.006994688203 0.003592422942
hoslem.test(views$AwareCSG , views$predicted_q1full.int)

```

```
	Hosmer and Lemeshow goodness of fit (GOF) test

data:  views$AwareCSG, views$predicted_q1full.int
X-squared = 12.040956, df = 8, p-value = 0.1493859
```
```

4.5 Supermarket Owner Questionnaire

Please answer each of the following questions on a 1-10 scale where 1 indicates that you disagree completely with the given statement and 10 indicates perfect agreement.

1. Raisin color is a key determinant of raisin category sales.
2. Raisin aroma is a key determinant of raisin category sales.
3. Raisin consumers have differing preferences for varying raisin package sizes.
4. Giving away in-package free gifts is a strong driver of brand sales.
5. Raisin consumers are price sensitive.
6. Raisin chewiness is an important determinant of raisin preference.
7. Raisin sales are highly linked to the sales of other fruits.
8. Marketing for the raisin category can significantly affect people's purchasing habits.
9. I enjoy raisins as part of my daily diet.
10. The raisin category is extremely profitable in my store.

Additional Questions:

What is the monthly dollar volume of your store? How much marketing expenditure, per month, do you spend on the raisin category? How many different SKUs of raisins does your store carry?

4.6 Q2(a,b) R Code

```r
survey.knn <- kmeans(survey[,1:9],3, nstart = 20)
survey$k <- as.factor(survey.knn$cluster)

survey.k1 <- survey[which(survey$k == unique(survey$k)[3]), ]# k = 1
survey.k2 <- survey[which(survey$k == unique(survey$k)[2]), ]# k = 2
survey.k3 <- survey[which(survey$k == unique(survey$k)[1]), ] # k = 3

lm.k1 <- lm(Q10 ~ Q1 + Q2 + Q3 + Q4 + Q5 + Q6 + Q7 + Q8 + Q9, data = survey.k1)
lm.k2 <- lm(Q10 ~ Q1 + Q2 + Q3 + Q4 + Q5 + Q6 + Q7 + Q8 + Q9, data = survey.k2)
lm.k3 <- lm(Q10 ~ Q1 + Q2 + Q3 + Q4 + Q5 + Q6 + Q7 + Q8 + Q9, data = survey.k3)

lm.k1.summary <- summary(lm.k1)
lm.k2.summary <- summary(lm.k2)
lm.k3.summary <- summary(lm.k3)

survey %>%
  ggplot(aes(x=k, y = Q10, fill=k)) +
  geom_boxplot() +
  theme(axis.text.x=element_text(angle=0,hjust=1))+
  labs(title = "Q10 response by K-cluster")
```

```r
# Create Dataframe for results
paste(paste0(names(survey)[1:(len(names(survey))-2)],"=c(0,0,0,0)"), collapse = " , ")
q2.lms <- data.frame(Q1=c(0,0,0,0) , Q2=c(0,0,0,0) , Q3=c(0,0,0,0) , Q4=c(0,0,0,0) , Q5=c(0,0,0,0) ,
  Q6=c(0,0,0,0) , Q7=c(0,0,0,0) , Q8=c(0,0,0,0) , Q9=c(0,0,0,0) )
rownames(q2.lms) <- c("K1","K2","K3","All Clusters")
# Input t-values
q2.lms[1,] <- lm.k1.summary$coefficients[-c(1),4]; q2.lms[2,] <- lm.k2.summary$coefficients[-c(1),4]
q2.lms[3,] <- lm.k3.summary$coefficients[-c(1),4]; q2.lms[4,] <- lm.k123.summary$coefficients[-c(1),4]
# Extract data
WriteXLS(q2.lms, ExcelFileName = "Q2tvals.xls")
```

```r
# Calculate correlations with response, Q10
survey.k1.cor <- as.data.frame(cor(survey.k1[,-c(11)])); survey.k2.cor <- as.data.frame(cor(survey.k2[,-c(11)]))
survey.k3.cor <- as.data.frame(cor(survey.k3[,-c(11)])); survey.k123.cor <- as.data.frame(cor(survey[,-c(11)]))
# Create Dataframe for results
q2.cors <- data.frame(Q1=c(0,0,0,0) , Q2=c(0,0,0,0) , Q3=c(0,0,0,0) , Q4=c(0,0,0,0) , Q5=c(0,0,0,0) ,
  Q6=c(0,0,0,0) , Q7=c(0,0,0,0) , Q8=c(0,0,0,0) , Q9=c(0,0,0,0) )
rownames(q2.cors) <- c("K1","K2","K3","All Clusters")
# Input correlations
q2.cors[1,] <- survey.k1.cor[10,1:9]; q2.cors[2,] <- survey.k2.cor[10,1:9];
q2.cors[3,] <- survey.k3.cor[10,1:9]; q2.cors[4,] <- survey.k123.cor[10,1:9]
# Extract data
WriteXLS(q2.cors, ExcelFileName = "Q2cors.xls")
```