Role of promotional Campaign in Targeting Customers Using Statistical Analysis:

Kroger Grocery Stores

Introduction:

With evolving technology and competition, the retail industry have developed a ways to gather the customer data, analyze it and get statistical significance and economical significance out of it. With that view point the retail stores and grocery store and other retail industries collects each transaction that happens in their stores and uses information they get from the data to attract the customers to increase the frequent visits of the customers to their stores. Collecting data and proper analysis of the data plays very vital role in retaining the customers and at the same time getting new business. In order to achieve these stores, need to understand the significance of the analysis and accordingly attract their targeted audience.

Kroger a big national chain of grocery stores, which has around 3000 stores throughout the country and has a market of millions of customers it became very important to understand and do proper statistical analysis for the advancement of frequent customers in their store.

With that viewpoint in this assignment, we are trying to figure out the way to use the data we got from Kroger grocery store which consists of shopping trip made by approximately 800 households household over a time of 102 weeks and used it to develop customized promotional campaigns that can be effectively targeted at certain households to grow those households' spending and increase the frequent visits of the customers every week.

Solutions:

Part 1:

1. In the midterm, you have obtained RFM metrics. For the final, you need to add a new metric "D". "D" is the log of the number of weeks for each household since the first campaign. With "D" we can try to control for a different level of enrollment in the panel. In short for each household, you should have R-F-M-D, plus additional control variables as you deem appropriate.

Interpretation/steps:

Midterm Steps:

For the data we have created Recency and frequency variables.
For recency column (R) which we have to create, we want to calculate the number of weeks since the last week with any purchase. We have created recency such that for instance a household with id =1, the household first time buys in week 8 so for that the recency should be 0 then for 9th week household didn't buy anything then for that the recency should be 1 and when household buys again in week 10th the recency should be 2 since they bought last in week 8th. And for frequency we have calculated the number of times household bought in the last month.

Final Exam Steps:

Creating D variable:

For this I have created two D variables:
D_2: log of the number of weeks for each household since the first campaign.
For this variable I have created the new variable "numbering_2" which is basically the number of weeks for each household since the first campaign.
So, D_2 variable is: if the person enters in the campaign in week 8 it will be log of 1 for that then if the person stays in the campaign till week 20 then till that it will be log of 13 for week 20 and if the house hold leaves the campaign in week 21 so D_2 will be 0 for that week 21 and if the household again enters in week 25 so it will be log of 14 and so on

D_1: This is extra variable I have created which is log of number of times spending per week by every individual HH through the period of 105 weeks for which the data is been collected for. It will consider if the HH have made any spending in a week, so for eg: if the HH started buying first time in week 8 so for that week 8 the D_1 will be log of 1 and if the HH keep on purchasing every week till 15 then D_1 will be incremental log and for week 15 it will be log (8) and if the HH didn't buy anything for next 5 week for that D_1 will be 0 and when the HH again buys in week 21 the D_1 will be log(9) and so on

Note: D_2 is the required variable according to question, D_1 is the extra variable created by me, to see if it improves different level of enrollment in the panel

In case of missing/zero data, you need to make very clear assumptions consistent with the way you execute the rest of the exam.

#Assumption for 0's in D variable:

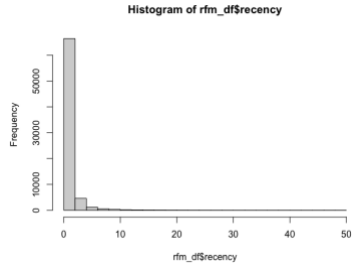
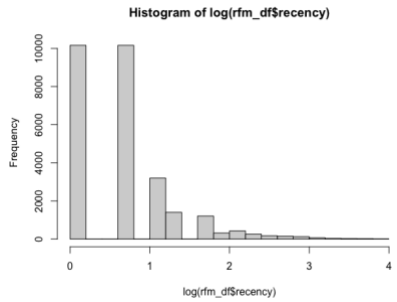
there are 0 values in both these D variables (D_2 and D_1) since for the weeks in which the HH was not been part of any campaign, the value for D_2 for that week is been considered 0 as no campaign has been assigned to HH, so the variable will not have any effect on the Weekly

spending. Similarly for the D_1 variable, in the week for which the HH have not made any purchase, so for that particular HH the D_1 will be 0 for that week.

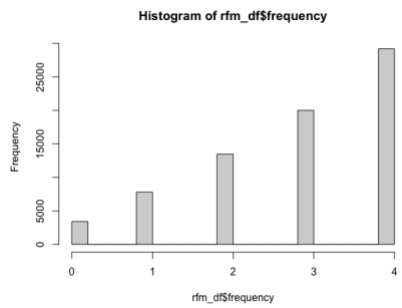
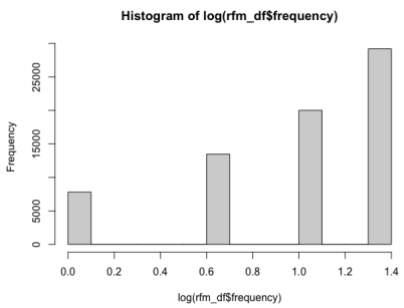
```
>table(Final_rfmd$D_2)
    0  0.69   1.1  1.39  1.61  1.79  1.95  2.08   2.2   2.3   2.4  2.48  2.56  2.64  2.71  2.77  2.83
49930   759   759   759   759   759   753   752   733   705   703   693   692   691   687   676   665
 2.89  2.94     3  3.04  3.09  3.14  3.18  3.22  3.26   3.3  3.33  3.37   3.4  3.43  3.47   3.5  3.53
  627   619   605   603   597   592   575   570   511   499   478   466   444   437   405   386   365
 3.56  3.58  3.61  3.64  3.66  3.69  3.71  3.74  3.76  3.78  3.81  3.83  3.85  3.87  3.89  3.91  3.93
  350   319   295   278   260   245   207   190   171   160   145   127   116   106    95    87    76
 3.95  3.97  3.99  4.01  4.03  4.04  4.06  4.08  4.09  4.11  4.13  4.14  4.16  4.17  4.19   4.2  4.22
   73    58    49    43    36    28    22    18    16    11     9     8     4     2     2     2     1
> table (Final_rfmd$D_1)
    0  0.69   1.1  1.39  1.61  1.79  1.95  2.08   2.2   2.3   2.4  2.48  2.56  2.64  2.71  2.77  2.83  2.89  2.94     3
18422   799   799   799   799   799   799   799   799   799   799   798   798   798   798   798   798   798   798
 3.04  3.09  3.14  3.18  3.22  3.26   3.3  3.33  3.37   3.4  3.43  3.47   3.5  3.53  3.56  3.58  3.61  3.64  3.66  3.69
  798   797   796   795   794   793   792   791   788   788   787   782   781   776   775   773   768   762   761   756
 3.71  3.74  3.76  3.78  3.81  3.83  3.85  3.87  3.89  3.91  3.93  3.95  3.97  3.99  4.01  4.03  4.04  4.06  4.08  4.09
  752   749   745   742   737   733   728   719   714   708   699   685   677   669   657   646   643   630   614   606
 4.11  4.13  4.14  4.16  4.17  4.19   4.2  4.22  4.23  4.25  4.26  4.28  4.29   4.3  4.32  4.33  4.34  4.36  4.37  4.38
  595   578   566   555   534   518   508   487   474   460   441   423   406   393   371   347   323   306   285   265
 4.39  4.41  4.42  4.43  4.44  4.45  4.47  4.48  4.49   4.5  4.51  4.52  4.53  4.54  4.55  4.56  4.57  4.58   4.6  4.61
  242   219   202   182   168   152   136   121   102    91    72    58    45    37    30    23    20    14     6     4
4.62
    3
> table(rfm_df$recency)
    0     1     2     3     4     5     6     7     8     9    10    11    12    13    14    15
46080 10160 10160  3203  1406   745   465   319   240   192   139   121    99    76    58    50
   16    17    18    19    20    21    22    23    24    25    26    27    28    29    30    31
   45    40    36    29    26    24    18    16    14    11     9     8     7     6     5     5
   32    33    34    35    36    37    38    39    40    41    42    43    44    45    46    47
    4     4     4     4     3     3     3     3     3     3     3     2     2     2     2     2
   48    49    50
    2     1     1
```

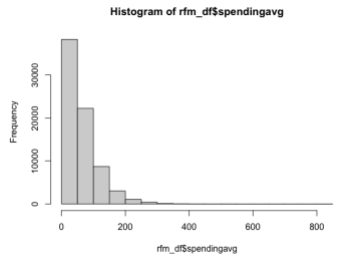
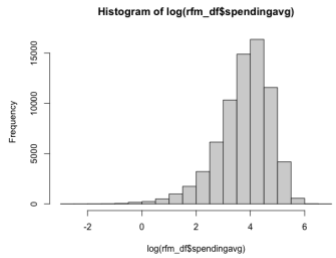
Interpretation:

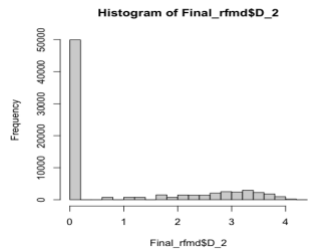
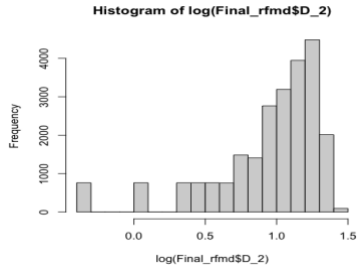
From the Above output we can see that the recency column have around 46k zeros, so we may have excess 0's problem here. there are more zeros in the data than can be predicted from Poisson distribution. So, there may be some extraneous events happening. And we may have to separate the excess zeros from the poisson distributed variable and may be for this we may have to use hurdle or zero-inflated Poisson model.

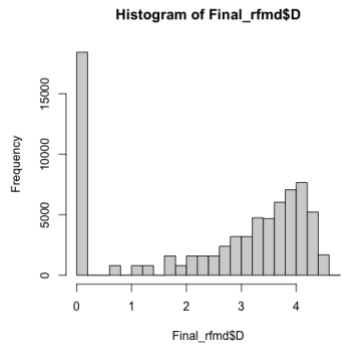
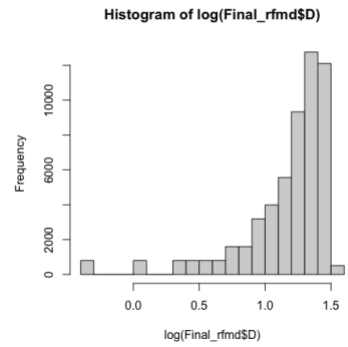
hist(rfm_df$recency)	hist(log(rfm_df$recency))
 # Not normally distributed it looks rightly skewed	 # Even log function is giving rightly skewed

hist(rfm_df$frequency)	hist(log(rfm_df$frequency))

 # not normally distributed it looks left skewed	 # not normally distributed it looks left skewed

hist(rfm_df$spendingavg)	hist(log(rfm_df$spendingavg))
 # Not normally distributed it looks rightly skewed	 # Much better

hist(Final_rfmd$D_2)	hist(log(Final_rfmd$D_2))
 # Not normally distributed it looks rightly skewed	 # Much better

hist(Final_rfmd$D_1)	hist(log(Final_rfmd$D_1))
	

Creating Combined campaign variable: combining the C dummy variables according to A_group, B_group and C_group

Creating new variables for campaign A, B, C:

I have converted the dummy variable into the campaign A, B, C based on which campaign the dummy variables were contributing to.

For eg: I have combined c8, c13, c18, c26, c30 to one group as A_group and similarly did the same for B_group and C_group.

#Checking Correlation between variables:

```
> cor(rfm_df[c(70,72,9)])
```

	weeklyspend	spendingavg	frequency	numbering	D	D_2	A_group	B_group	C_group
weeklyspend	1.00000000	0.7375632	0.38258238	0.3597510	0.4496291	0.1339927	0.06346553	0.1345981	0.11076911
spendingavg	0.73756324	1.0000000	0.51211111	0.3337388	0.3654573	0.1798985	0.07901490	0.1834046	0.14809136
frequency	0.38258238	0.5121111	1.00000000	0.5235383	0.6478854	0.1647984	0.11883660	0.1197239	0.08141963
numbering	0.35975104	0.3337388	0.52353826	1.0000000	0.8868114	0.4759637	0.30046076	0.2898504	0.22573904
D_1	0.44962911	0.3654573	0.64788536	0.8868114	1.0000000	0.3103132	0.22489528	0.1940964	0.13647701
D_2	0.13399267	0.1798985	0.16479842	0.4759637	0.3103132	1.0000000	0.71149393	0.5436145	0.38361217
A_group	0.06346553	0.0790149	0.11883660	0.3004608	0.2248953	0.7114939	1.00000000	0.1219328	0.12176943
B_group	0.13459812	0.1834046	0.11972394	0.2898504	0.1940964	0.5436145	0.12193276	1.0000000	0.18401970
C_group	0.11076911	0.1480914	0.08141963	0.2257390	0.1364770	0.3836122	0.12176943	0.1840197	1.00000000

Interpretation:

From the correlation matrix above we can see that there is no high correlation between any of the four variables that we created that are Recency, Frequency and spendingavg which is Monetory variable, D_1 and D_2 . Which is a good thing. Also there does not have high coreelation with dependent variable, weeklyspend and the campaign variable (A_group, B_group, C_group). This tells us that the variable are different.

Apart from these I tried checking the correlation among various variables and found out that :
calendarweek and weeksin has a high correlation of 0.976542627 (so I will be using just calendarweek in my model)
w_display and t_display has a high correlation of 0.980999268 (so I will be using just w_display in my model)
t_mailer and t_display has a high correlation of 0.906315206
t_mailer and w_display has a high correlation of 0.8916555894

#The high correlation of these variables tells us that they are also almost the same variables.

Part 1: Response Model with Fixed Effects

We have different types of campaigns (A, B, C). The purpose of this Part is to:

1. Build a response model for Weekly spending (as DV) as a function of the relevant metrics (R-F-M-D) plus covariates while accounting for fixed effects (a) at the household level (b) at the campaign level (A, or B or C).

Please highlight which variables are significant and if these seem consistent with your intuition.

#NOW implementing models

(a) fixed effects at the household level

```
fixed_model_id= lm(weeklyspend ~ as.factor(id) + frequency + recency + spendingavg + D_1 +
                    D_2 +calendarweek+age+store.shopped+married + income+ numall_now +
                   hhsize +numkids+ A_group+B_group+ C_group+w_loydisc_fb+w_loydisc_cat+
                    w_display+w_mailer, data= Final_rfmd)
#OR
#with spendingavg and calendarweek variable
fixed_model_id_2 <- plm(weeklyspend ~ frequency +recency + spendingavg + D_2 +
                   calendarweek + age + store.shopped+married +hhsize+income+ numkids+
                   A_group+B_group+ C_group+w_loydisc_fb+w_loydisc_cat+
                   w_display +w_mailer, data=Final_rfmd, index="id", model="within")

#with D_1 variable instead of spendingavg and calendarweek variable
fixed_model_id_3 <- plm(weeklyspend ~ frequency+recency+D_1+D_2+age+
                   store.shopped+married +hhsize+income+
                   numkids+ A_group+B_group+ C_group+w_loydisc_fb+w_loydisc_cat+
                   w_display+w_mailer, data=Final_rfmd, index="id", model="within")
```

(b) fixed effects at the campaign level (A, or B or C)

```
#with spendingavg
fixed_model_camp_1= lm(weeklyspend ~ id + frequency + recency + spendingavg + D_2 +
                     calendarweek + age + store.shopped+married + income+hhsize+ numkids+
                     as.factor(A_group)+as.factor(B_group)+ as.factor(C_group)+
                      w_loydisc_cat+ w_display+w_mailer, data= Final_rfmd)

#with D_1 variable instead of spendingavg and calendarweek
fixed_model_camp_2= lm(weeklyspend ~ id+frequency+recency+D_1+D_2+age+
                   store.shopped+married + income+hhsize+ numkids+
                   as.factor(A_group)+as.factor(B_group)+ as.factor(C_group)+
                   w_loydisc_fb+w_loydisc_cat+ w_display+w_mailer, data= Final_rfmd)
```

Note: D_2 is the required variable we have to create according to question which is log of the number of weeks for each household since the first campaign, D_1 is the extra variable created by me, which is log of number of times spending per week by every individual HH through the period of 105 weeks for which the data is been collected for, to see if it improves different level of enrollment in the panel

Interpretation:

If we see the output of the stargazer, we can notice that the models for which we have selected the D_1 variable (which is the extra variable I have created) instead of using the spendingavg and calendarweek, gives better result and also the D_1 is significant in both the models in which it is used.

#Comparing the fixed effect models:

Model (a) at the household level (b) at the campaign level (A, or B or C)				
	Dependent variable:			
	weeklyspend			
	(a) Fixed Effect at the household level		*(b) Fixed Effect at the campaign level*	
	With spendingavg(M) M1	With D_1 variable M2	With spendingavg(M) M3	With D_1 variable M4
id			0.00002 (0.0003)	0.002*** (0.0003)
frequency	-11.217*** (0.318)	0.562* (0.307)	-11.062*** (0.279)	1.122*** (0.295)
recency	0.379*** (0.089)	0.197** (0.098)	0.314*** (0.073)	0.117 (0.086)
spendingavg	0.798*** (0.010)		0.787*** (0.007)	
D_1		7.214*** (0.292)		6.773*** (0.299)
D_2	-0.175 (0.379)	-1.215*** (0.402)	-0.262 (0.355)	-0.059 (0.404)
calendarweek	0.005 (0.009)		0.003 (0.008)	
age			0.001 (0.018)	-0.069*** (0.021)
store.shopped	0.0002*** (0.00004)	0.0002*** (0.00005)	0.0002*** (0.00004)	0.0002*** (0.00004)
A_group	0.721 (1.420)	1.332 (1.562)		
B_group	0.423 (0.984)	1.161 (1.083)		
C_group	0.198 (0.897)	0.652 (0.985)		
w_loydisc_fb	-0.360 (1.180)	-1.412 (1.296)		6.451*** (1.081)
married			0.323 (0.528)	1.897*** (0.618)
income			0.00000 (0.00000)	0.0001*** (0.00001)
hhsize			-0.115 (0.543)	0.352 (0.636)
numkids			-0.109 (0.618)	0.461 (0.724)
as.factor(A_group)1			0.606 (1.277)	8.140*** (1.493)
as.factor(B_group)1			0.320 (0.937)	5.388*** (1.097)
as.factor(C_group)1			0.478 (0.866)	-0.516 (1.014)
w_loydisc_cat	-0.235 (1.757)	-1.491 (1.934)	3.738*** (1.021)	-2.609** (1.289)
w_display	24.689*** (2.137)	18.045*** (2.364)	23.317*** (2.017)	19.779*** (2.377)
w_mailer	71.469*** (1.923)	51.042*** (2.313)	71.747*** (1.819)	51.379*** (2.336)
Constant			-0.791 (1.492)	-5.357*** (1.719)
Observations	27,783	27,783	27,783	27,783
R^2	0.463	0.557	0.444	0.595
Adjusted R^2	0.448	0.544	0. 444	0.595
Residual Std.				
Error (df = 27764)			33.067	38.745
F Statistic	2,610.521***	1,941.995***	2,269.234***	1,233.989***
	(df = 13; 26987)	(df = 12; 26988)	(df = 18; 27764)	(df = 18; 27764)

#significant variable's estimates are highlighted

Interpretation of the stargazer output:

From the output we can see that the estimates for all the three campaign variable A_group, B_Group, C_group are positive which shows that with the campaign the weeklyspend is increasing which is a good thing. And we can see that the campaign A is better than the B and C as the estimates of A is better than the other which was expected. It tells us that if the household are the part of campaign A, then the weeklyspend increases. But the D_2 for all the models have negative estimate which shows that with increase in log of the number of weeks for each household since the first campaign the weeklyspend is decreasing which is ideally not expected but the effect is not very small, so we can probably ignore that.

(a) Comparing model with Fixed Effect at the household level:

(i) With spendingavg(M):

The first odd thing which we found in the model was that estimate for frequency was negative which means as the frequency increases the weeklyspend decreases by $11 which was not ideally expected but one explanation for this may be that since the HH is coming frequently, the HH is buying or purchasing anything in the store in batches as in the HH is not spending a lot in single/couple visits in a week, but instead in multiple visits in a month. Also, for this model w_display and w_mailer are highly significant, and it increases the weekly spend by $24.68 and $71.46 respectively.

(ii) With D_1 variable instead of spendingavg and calendarweek:

By taking the D_1 variable instead of the spendingavg and calendarweek variables, with D_2 variable has actually given a better result as compared to the model in which spendingavg and calendarweek is considered and D_1 is left out. We can see that D_1 is significant and positive which tells us that if the number of times spending per week by every individual HH increases by 1 % the weeklyspend increases by $7.2. Also, from the output of model M2 estimate for frequency is positive which means as the frequency increases the weeklyspend increases which is ideally expected. Also, Campaign A, B, C has higher effect on the weeklyspend as compared to their effect in model M1. Also, for this model w_display and w_mailer are highly significant, and it increases the weekly spend by $18 and $51 respectively. Also the R square is better for model M2 as compared to model M1.

(b) Comparing model with Fixed Effect at the campaign level

(i) With spendingavg(M):

Again here the frequency is negative which means as the frequency increases the weeklyspend decreases. With this model we have significant spendingavg and it is positive which means as the spendingavg increases the weeklyspend increases which was expected. Also, for this model w_loydisc_cat , w_display and w_mailer are highly significant, and it increases the weekly spend by $3.7, $23.31 and $71.74 respectively.

(ii) With D_1 variable instead of spendingavg and calendarweek:

With this model we can see that the D_1 is significant and is positive also, which tells us that if the number of times spending per week by every individual HH increases by 1 % then the weeklyspend increases by $6.45. Also, for this model w_display and w_mailer are highly

significant, and it increases the weekly spend by $19.77 and $51.37 respectively. Also, the R square is better for model M4 as compared to model M3.

Part 2. Exploring K-type of Heterogeneity

1. Perform a K-means segmentation with both 3 and 4 segments. The variables used in the K-means algorithm should be properly scaled (normalized standardized.) This means you should (scale the predictors, compute the distance matrix, be careful when you specify the distance, and then run the K-means algorithm to obtain the index for each household. That is, which household belongs to which segment. To do so, you may consider a majority-rule classification. That is if household j has 70% of their weekly observations belonging to cluster k, then the household is labeled as belonging to"k". You may resolve ties by randomly picking. Please use "set.seed(123)" on top of your script for reproducibility, so we should all get the same results.) If you have memory problems while performing the Kmeans

Which number of segments appears more appropriate? You can compare the adjusted R2, and the different means for R,F,M,D for each clusters.

For each segment perform the same analysis as in Part 1.

Find a succinct way to summarize/classify each segment. "I.e. Segment 1 is more about R and F, less about M and the rest of the covariates."

K-means segmentation with 3 segments:

(a) at the household level :

	Dependent variable:		
	weeklyspend		
	M1	M2	M3
as.factor(id)71	17.618*** (6.322)		
as.factor(id)158	19.354*** (6.461)		
as.factor(id)2282	18.774*** (6.459)		
as.factor(id)2364	34.828*** (6.428)		
as.factor(id)2446	16.481*** (6.344)		
as.factor(id)1833		-15.524*** (5.991)	
as.factor(id)1861		-12.602** (6.048)	
as.factor(id)1926		-13.355** (6.032)	
as.factor(id)1927		-16.615*** (6.008)	
as.factor(id)2235		-15.745** (6.137)	
as.factor(id)2250		-12.322** (6.058)	
as.factor(id)2437		-14.324** (6.971)	
as.factor(id)2496		-13.474** (6.065)	
as.factor(id)1949			-2.516 (6.206)
frequency	-13.257*** (0.258)	-14.465*** (0.880)	-9.354*** (3.576)
recency	0.775*** (0.100)	1.137*** (0.334)	0.558* (0.335)
spendingavg	0.993*** (0.006)	1.001*** (0.026)	1.129*** (0.141)
D_1	7.534*** (0.212)	8.109*** (0.718)	15.084*** (3.758)

D_2	-0.347*** (0.275)	0.115 (1.089)	2.272 (10.753)
calendarweek	-0.153*** (0.009)	-0.074*** (0.025)	-0.101 (0.085)
age			
store.shopped	0.0001*** (0.00003)	0.0001 (0.0001)	-0.0003 (0.001)
married			
income			
numall_now	-1.590 (1.257)	5.998 (10.725)	-8.174 (16.869)
hhsize			
numkids			
A_group	2.397 (1.571)	-8.898 (10.796)	8.982 (16.001)
B_group	1.412 (1.519)	-7.453 (11.376)	
C_group	1.450 (1.328)	-4.085 (11.164)	
w_loydisc_fb	-2.220** (0.968)	-0.342 (2.510)	-0.282 (18.515)
w_loydisc_cat	1.263 (1.471)	0.901 (5.682)	21.765 (20.482)
w_display	17.873*** (1.328)	9.718** (4.681)	55.760** (22.446)
w_mailer	37.654*** (1.336)	50.496*** (4.580)	-14.260 (22.914)
Constant	-3.260 (4.560)	8.411 (8.456)	11.384 (13.779)
Observations	69,555	4,118	190
R^2	0.633	0.6	0.628
Adjusted R^2	0.629	0.595	0.598
Residual Std. Error	43.250 (df = 68789)	35.355 (df = 4057)	26.316 (df = 175)
F Statistic	155.202*** (df = 765; 68789)	101.633*** (df = 60; 4057)	21.105*** (df = 14; 175)

#Note: there were 799 id's but I have pasted only the id's which are significant

Interpretation:

- Out of the three models created using three clusters the model M1 is better than the other two models M2 and M3. Also, the R square for Model M1 is 0.633 which is better than the R square of other two models.
- Also for all the three campaign variable, A_group, B_Group, C_group are positive which shows that when the HH is exposed to the campaign the weeklyspend increases which is ideally expected. Also, campaign A is performing better than B and C. It is the same observation that we found out from earlier fixed effect models also.
- All the RFMD variables are highly significant, also the extra variable D_1 is also highly significant as compare to the other variables which shows that the segment 1 is more about the RFMD variables as compared to the other covariates.

(b) at the campaign level (A, or B or C)

	Dependent variable:		
	weeklyspend		
	M1	M2	M3

id	-0.0001 (0.0002)	-0.0001 (0.0002)	-0.008 (0.013)
frequency	-11.090*** (0.227)	-8.465*** (0.880)	-6.851* (3.669)
recency	0.657*** (0.090)	0.672** (0.090)	0.674* (0.348)
spendingavg	0.996*** (0.004)	0.982*** (0.004)	1.056*** (0.146)
D_2	-0.551** (0.263)	-0.451* (0.213)	3.107 (11.203)
calendarweek	0.005 (0.007)	0.013 (0.017)	-0.012 (0.085)
age	0.026* (0.015)		
store.shopped	0.00003 (0.00002)	0.00003 (0.00002)	0.0001 (0.001)
married	0.815* (0.460)		
income	0.00000 (0.00000)		
hhsize	-1.268*** (0.465)	-1.28*** (0.45)	
numkids	0.686 (0.522)	0.481 (0.321)	
A_group	1.161* (0.638)	1.061* (0.628)	0.682 (18.809)
B_group	-0.203 (0.666)	-0.212 (0.556)	-10.472 (17.568)
C_group	-0.106 (0.872)	-0.076 (0.072)	
w_loydisc_fb	-0.590 (0.782)	-0.342 (2.510)	-14.633 (18.930)
w_loydisc_cat	6.876*** (0.946)	0.901 (5.682)	37.938* (20.926)
w_display	19.342*** (1.184)	9.718** (4.681)	41.501* (23.094)
w_mailer	54.430*** (1.137)	50.496*** (4.580)	47.884*** (17.602)
Constant	0.142 (1.324)	8.417 (8.346)	22.428 (16.055)
Observations	69,555	4,118	190
R^2	0.622	0.6	0.594
Adjusted R^2	0.622	0.6	0.564
Residual Std. Error	43.686 (df = 69535)	35.355 (df = 4057)	27.422 (df = 176)
F Statistic	6,013.525*** (df = 19; 69535)	101.633*** (df = 60; 4057)	19.791*** (df = 13; 176)

Interpretation:

- Out of the three models created using three clusters again the model M1 made using cluster 1 is better than the other two models M2 and M3. Also, the R square for Model M1 is 0.622 which is better than the R square of other two models. And since the segment 1 has very high number of data as compared to the other two models, may be the model M1 is able to fit the data better than the other two models
- Out of the three campaign variables only variable A_group is positive which shows that the HH exposed to campaign A gives better result in terms of weeklyspend and estimates for B_Group and C_group is negative so we can state that the campaign B and C does not have positive effect on the weeklyspend.
- All the RFMD variables are highly significant, as compare to the other variables which shows that the segment 1 is more about the RFMD variables as compared to the other covariates.

K-means segmentation with 4 segments:

(a) at the household level :

	Dependent :			
	weeklyspend			
	M1	**M2**	**M3**	**M4**
as.factor(id)158	20.433*** (6.861)			
as.factor(id)346	15.180** (6.768)			
as.factor(id)973	14.321** (6.805)			
as.factor(id)1137	15.133** (6.685)			
as.factor(id)1650	16.121** (6.837)			
as.factor(id)2140	17.703** (8.066)			
as.factor(id)2364	35.092*** (6.802)			
as.factor(id)71		18.405*** (4.590)		
as.factor(id)178		8.349** (4.624)		
as.factor(id)442		11.421** (4.647)		
as.factor(id)460		9.148** (4.600)		
as.factor(id)733		9.336** (4.596)		
as.factor(id)802		10.963** (4.537)		
as.factor(id)1267		9.833** (4.600)		
as.factor(id)1438		11.553** (4.515)		
as.factor(id)1485		9.915** (4.547)		
as.factor(id)1627		10.469** (4.636)		
as.factor(id)1814		15.467*** (4.684)		
as.factor(id)2282		18.555*** (4.715)		
as.factor(id)2390		11.913*** (4.507)		
as.factor(id)2393		10.677** (4.627)		
as.factor(id)2411		12.389*** (4.527)		
as.factor(id)770			-0.160 (5.226)	
as.factor(id)852			5.942 (5.297)	
as.factor(id)2181			1.366 (5.149)	
as.factor(id)2292			-0.872 (5.443)	
as.factor(id)997				-0.357 (5.218)
frequency	-12.802*** (0.319)	-13.237*** (0.389)	-15.154*** (1.049)	9.669 (39.450)
recency	1.740*** (0.186)	0.399*** (0.093)	0.263 (0.225)	-13.353 (19.909)
spendingavg	0.998*** (0.007)	0.962*** (0.013)	0.979*** (0.030)	0.874 (0.949)
D_1	8.134*** (0.248)	6.429*** (0.345)	7.745*** (0.968)	66.671 (45.164)
D_2	-0.073* (0.305)	0.275 (0.616)	-2.704 (1.880)	-9.562 (32.816)

calendarweek	-0.190*** (0.011)	-0.065*** (0.012)	-0.057* (0.031)	4.811 (3.954)
age				
store.shopped	0.0001* (0.00003)	0.0001 (0.0001)	0.0004*** (0.0001)	-0.369 (0.306)
married				
income				
numall_now	-1.647 (1.356)	-2.472 (4.054)	22.042 (14.183)	158.433 (92.845)
hhsize				
numkids				
A_group	2.792* (1.690)	1.513 (4.169)	-17.140 (14.654)	-108.048 (95.907)
B_group	1.447 (1.644)	0.679 (4.570)	-21.117 (15.357)	
C_group	1.074 (1.436)	1.296 (4.789)	-24.064 (17.177)	
w_loydisc_fb	-1.955* (1.090)	-3.136* (1.704)	-0.810 (3.174)	344.208 (214.603)
w_loydisc_cat	1.569 (1.699)	1.130 (2.439)	-4.589 (7.317)	311.686 (323.715)
w_display	17.397*** (1.497)	13.687*** (2.463)	16.875*** (6.087)	22.970 (290.867)
w_mailer	36.370*** (1.517)	47.181*** (2.426)	38.868*** (5.735)	250.498 (326.921)
Constant	-5.331 (4.968)	-7.022** (3.571)	-1.675 (6.477)	-181.633 (328.900)
Observations	56,977	14,224	2,637	25
R^2	0.618	0.612	0.621	0.74
Adjusted R^2	0.614	0.607	0.614	0.433

#Note: there were 799 id's but I have pasted only the id's which are significant

Interpretation:

- Out of the three models created using four clusters again the model M4 has highest R square of 0.74 as compared to the other models but the Model M4 has only 25 observations where as the model M1 has around 56k observations so here we cannot say just on the basis of R square that M4 is better than the other models, also M4 has only records of HH which are exposed to campaign C it does not have records of HH exposed to other campaign. So based on this and also based on the estimates we got I can say that the model M1 is better than any other model.
- Also for all the three campaign variable, A_group, B_Group, C_group are positive which shows that when the HH is exposed to the campaign the weeklyspend increases which ideally expected. Also campaign A is performing better than B and C. It is the same observation that we found out from earlier fixed effect models also.
- All the RFMD variables are highly significant, also the extra variable D_1 is also highly significant as compare to the other variables which shows that the segment 1 is more about the RFMD variables as compared to the other covariates.

(b) at the campaign level (A, or B or C)

	Dependent:			
	weeklyspend			
	M1	**M2**	**M3**	**M4**
id	-0.0001 (0.0003)	0.0001 (0.0004)	0.0003 (0.001)	
frequency	-10.495*** (0.292)	-11.804*** (0.351)	-13.652*** (0.976)	26.676 (39.541)
recency	1.214*** (0.178)	0.426*** (0.085)	0.232 (0.210)	-2.462 (19.378)
spendingavg	1.000*** (0.004)	0.957*** (0.011)	0.971*** (0.027)	0.862 (0.994)
D_2	-0.459 (0.291)	0.297 (0.601)	-2.519 (1.822)	4.877 (32.828)
calendarweek	-0.002 (0.009)	0.022** (0.011)	0.039 (0.028)	4.831 (4.143)
age	0.037** (0.017)	0.016 (0.022)	0.029 (0.103)	
store.shopped	0.00001 (0.00002)	0.0001*** (0.00004)	0.0003*** (0.0001)	-0.048 (0.226)
married	0.588 (0.526)	0.438 (0.732)	1.176 (2.082)	
income	0.00000 (0.00000)	0.00002** (0.00001)	0.00000 (0.00002)	
hhsize	-1.168** (0.535)	-0.393 (0.695)	0.940 (1.233)	
numkids	0.679 (0.601)	0.657 (1.048)		
A_group	1.038 (0.704)	-0.354 (1.397)	6.989* (4.099)	37.706 (56.433)
B_group	-0.088 (0.730)	-1.918 (1.555)	2.243 (4.243)	
C_group	0.031 (0.934)	-1.776 (2.672)	-0.089 (6.575)	69.175 (73.837)
w_loydisc_fb	-0.853 (0.887)	-2.580** (1.313)	1.190 (2.754)	209.958 (203.707)
w_loydisc_cat	6.513*** (1.103)	7.297*** (1.477)	4.165 (4.360)	322.404 (339.162)
w_display	19.524*** (1.326)	15.850*** (2.315)	21.438*** (5.864)	-85.263 (294.980)
w_mailer	52.823*** (1.287)	64.862*** (2.125)	59.224*** (5.012)	309.095 (340.072)
Constant	-1.961 (1.622)	-2.300 (1.930)	-7.784 (8.402)	-245.872 (341.651)
Observations	56,977	14,224	2,637	25
R^2	0.606	0.596	0.606	0.689
Adjusted R^2	0.606	0.596	0.604	0.378

Interpretation:

- Again here Model M4 has highest R square but also it has just 25 records so juct on the basis of R square we cannot say that the Model M4 is better than the others. The model M1 has the R square of 0.606 which is not very less compared to the model M4. So the model M1 made using segment 1 is considered to be the best model.
- Out of the three campaign variables only variable A_group and C_group are positive which shows that the HH exposed to campaign A and C gives better result in terms of weeklyspend and estimates for B_Group is negative so we can state that the campaign B does not have positive effect on the weeklyspend. And out of two campaign A and C, campaign A is performing better.
- All the RFMD variables are highly significant, as compare to the other variables which shows that the segment 1 is more about the RFMD variables as compared to the other covariates.

	C1	C2	C3	C1	C2	C3	C4
R	1.23	0.31	0.33	1.35	0.28	0.06	0.28
F	2.27	3.2	3.12	2.15	3.2	3.68	3.18
M	38.36	69.1	71.49	38.4	69.12	71.49	64.83
D	0.63	0.96	1.03	0.63	0.95	1.03	0.99

segment 3
segment 4

Interpretation:

K means with Segment 3:

Here I have taken the average of the RFMD variables for every segment they were part of to see which variable is having higher contribution to which segment. And we can see that the M variable is contributing more towards the Cluster 3 and for F variable it is contributing more to the cluster 2 and R variable is contributing more towards the cluster 1 and D variable is contributing more towards the cluster 3

Kmeans with Segment 4:

Here I have taken the average of the RFMD variables for every segment they were part of to see which variable is having higher contribution to which segment. And we can see that the M variable is contributing more towards the Cluster 3 and for F variable it is contributing more to the cluster 3 and R variable is contributing more towards the cluster 1 and D variable is contributing more towards the cluster 3.

Part 3. Discussion

Please explain which response model you would prefer as a data analyst interested in designing marketing campaigns: Fixed Effect or K-means+Fixed Effects?. Discuss the pros and cons of segmentation from both a managerial and IT perspective.

Interpretation:

(a) at the household level:

Based on the results which we got the K-means+Fixed Effects models is working better than just the fixed effect models and out of the K-means+Fixed Effects models model M1 using segment 3 appears to be better than any other K-means+Fixed Effects model with segment 4. Also the R square of the K-means+Fixed Effects models model M1 using segment 3 is higher than any other models. Also from the output we can easily interpret that the campaign A, B and C is having positive effect and the model M1 with segment 3 is better able to explain the effect of the campaigns on the weeklyspend. And Out of the three campaign A is performing better than B and C.

(b) at the campaign level (A, or B or C):
Considering the fixed effect of the campaign with K-means is showing better results just the models in which only fixed effect of the campaign is considered, And out of the K-means+Fixed Effects models model M1 using segment 3 appears to be better than any other K-means+Fixed Effects model with segment 4. Also the R square of the K-means+Fixed Effects model M1 using segment 3 is higher than any other models. Also all the RFMD variable are highly significant and is providing the interpretation which is ideally expected with the K-means+Fixed Effects models.

From the output of the stargazer we can clearly see the difference between the pooled and the model clearly failing in predicting the dependent variable and random, fixed models can be seen providing better predictions as compared to the pooled.

Also based on the overall results we found in both Fixed Effect and K-means+Fixed Effects model we can only say that if the household is exposed to the campaign A then the household will be spending more money as compared to the households that are exposed to other campaigns.

Pros and Cons of segmentation:
With segmentation it is possible to group the people with similar habits and then focused on the group of people and it becomes easy to come up with the promotional campaigns focusing on the group than focusing on individual HH also it saves cost.
With the Segmentation it becomes possible to come up with the models and look for the effect of the group of the people. But doing so sometimes becomes hard to measure impact of the individual has on the dependent variables..

Conclusion:

Based on the finding and my analysis in this assignment I can say that to improve weeklyspend that is to attract the household to come to store more frequently and spend more money the grocery store should try to target more households with campaign A.

Future Scope:

As part of this assignment we have focused the effect of the Campaign A, B, C , but in future we can split campaign B ,C and check the individual effect of a particular promotion on the household. For instance consider that campaign B has four promotions included in it like “Save on groceries for family”,” Baby Stuff”, “Breakfast” and “Personal Hygiene and Makeup”, so for this we can further check

the effect of the individual promotion within the campaign B on the recency and household which will be helpful to the store in determining which particular promotions to include more in the particular campaign. Also in this assignment I have created the variable “month” from the week variable. So we can do monthly predictions also from that.

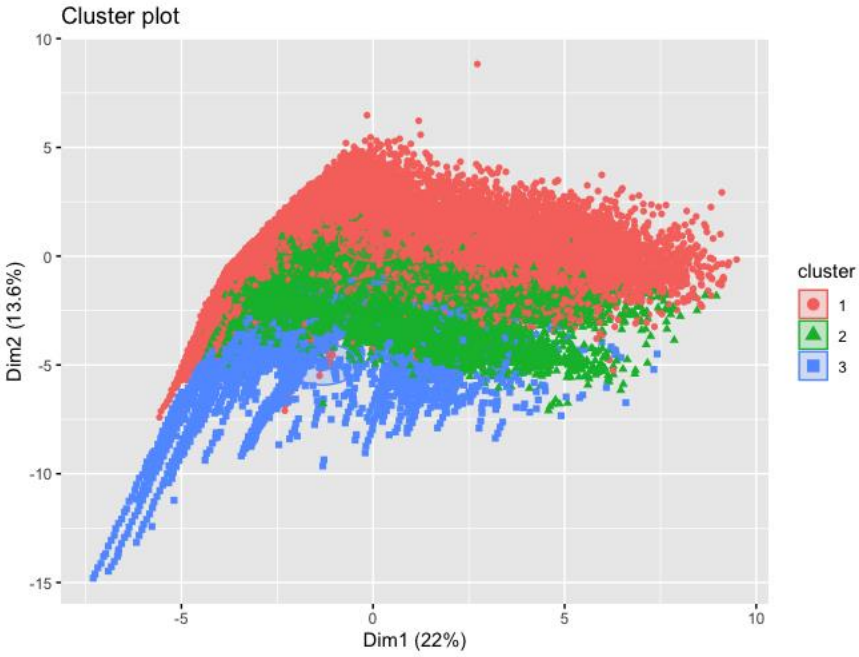
References:

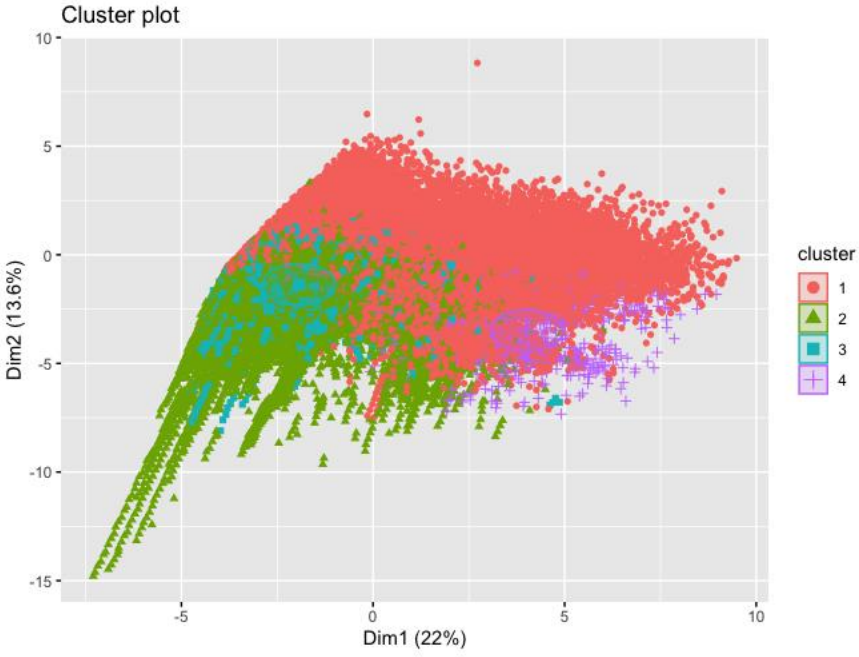
https://stackoverflow.com/questions/57447138/how-to-calculate-the-number-of-months-since-the-last-month-with-purchase-for-eac

https://www.programmingr.com/rfm-analysis/

https://rdrr.io/cran/wktmo/man/weekToMonth.html

https://www.kaggle.com/hendraherviawan/customer-segmentation-using-rfm-analysis-r


Cluster plot
Dim2 (13.6%)
Dim1 (22%)
cluster
1
2
3


Cluster plot
Dim2 (13.6%)
Dim1 (22%)
cluster
1
2
3
4