RFM Exercise 11 (10 points)
To be done individually

Use the RFM_Exercsie_data file for this assignment. The variable names and the description of the variables are as follows:

Variable Name	Description
CUST_ID	Customer identification number
Last_Response	1 if "Customer responded favorably to the next year mailing campaign"
Gender	"1"=male "0"=female
Tot_Dollars	Total money (in dollars) spent by the customer in last 3 years not counting the next year campaign. (Money Variable)
Last_pur	Number of months since customer's last purchase by the customer in last 3 years not counting the next year campaign. (Recency Variable)
Num_pur	Total number of purchases by the customer in last 3 years not counting the next year campaign. (Frequency Variable)
Flag_express	1 if "Customer prefers express shipping"
Flag_multibuyer	1 if "Customer is an identified multi-buyer"
Flag_bigCity	1 if "Customer lives in a big city"

1. Use appropriate statistical methods to analyze relationships between Last_response and the recency, frequency and monetary variables in the data set. Based on this, do you think Recency should come first, then Frequency and then Monetary values when you create RFM bins? Why or why not? **(2 Points)**

Solution: To understand the relationship between Last_response, recency, frequency and monetary variables, a logistic regression model was run with Last_response as the dependent variable and recency, frequency & money as independent variables. Based on the values of regression output in *Table (1)*, Recency with the highest value of absolute standardized coefficient (0.3690) is the most important variable followed by Frequency (0.1848) and Money (0.0491). Based on the order of variable importance, recency should come first followed by frequency and money for creating RFM bins.

Table 1

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1945.7541	3	<.0001
Score	1808.9868	3	<.0001
Wald	1698.7846	3	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-1.9889	0.0458	1887.6777	<.0001	
Recency	1	-0.0821	0.00257	1023.3304	<.0001	-0.3690
Frequency	1	0.0964	0.00492	384.5684	<.0001	0.1848
Money	1	0.000878	0.000184	22.8535	<.0001	0.0491

2. Use the Recency, Frequency and Monetary variables and attempt to create bins (use similar settings that were used in the demonstrations) for R, F and M. Report the bin category ranges for each of the R, F and M variable as well as number of observations in each bin for each of those 3 variables . **(2 Points)**

Solution:

Bins and category ranges for Recency

Customers were divided into five buckets (quantiles) based on their values for recency. The recency values were multiplied with -1 before forming the buckets to ensure that customers with smallest recency end up in the highest numbered bucket. The proportion of customers in each of the recency buckets is as shown in *Table (2.a)*

Table 2.a

Rank for Variable Trf_recency				
rank_R	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	10183	20.37	10183	20.37
2	12413	24.83	22596	45.19
3	12412	24.82	35008	70.02
4	7484	14.97	42492	84.98
5	7508	15.02	50000	100.00

Table 2.b

Quantiles for Recency					
Min	Bin 1	Bin 2	Bin 3	Bin 4	Max (Bin 5)
-36	-18	-14	-10	-6	-2

The cutoff values of recency used for distributing customers into buckets are as shown in *Table (2.b)*. The maximum and minimum values of recency are 2 and 36.
Customers with recency between 36 to18 are put in Bin 1, 18 to 14 in Bin 2, 14 to 10 in Bin 3, 10 to 6 in Bin 4 and 6 to 2 in Bin 5.

Bins and category ranges for Frequency

Customers were divided into four buckets based on their values for frequency. The proportion of customers in each of the frequency buckets is as shown in *Table (2.c)*

Rank for Variable Frequency				
rank_F	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	15120	30.24	15120	30.24
2	14935	29.87	30055	60.11
4	10042	20.08	40097	80.19
5	9903	19.81	50000	100.00

Table (2.c)

Quantiles for Frequency				
Min	Bin 1	Bin 2	Bin 4	Max (Bin 5)
1	1	2	7	12

Table (2.d)

The cutoff values of frequency used for distributing customers into buckets are as shown in Table (2.d). The maximum and minimum frequency values are 1 and 12.

Customers with a frequency of 1 are put in Bin 1, frequency 2 in Bin 2, 2 to 7 in Bin 4, 7 to 12 in Bin 5.

Bins and category ranges for Money

Customers were divided into five buckets based on their values of money spent. The proportion of customers in each of the money buckets are as shown in *Table (2.e)*

Rank for Variable Money				
rank_M	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	10062	20.12	10062	20.12
2	9966	19.93	20028	40.06
3	10091	20.18	30119	60.24
4	9919	19.84	40038	80.08
5	9962	19.92	50000	100.00

Table (2.e)

Quantiles for Money					
Min	Bin 1	Bin 2	Bin 3	Bin 4	Max
15	110	178	239	298	479

Table (2.f)

The cutoff values of spend used for distributing customers into buckets are as shown in *Table (2.f)*. The maximum and minimum for money spent are 15 and 479.

Customers with spend between 15 to 110 are put in Bin 1, 110 to 178 in Bin 2, 178 to 239 in Bin 3, 239 to 298 in Bin 4 and 298 to 479 in Bin 5.

3. Create RFM cells by concatenating R, F and M bins. Which is the best (highest number) RFM cell? Which is the worst (lowest number) RFM cell? Describe the characteristics of the best and the worst RFM cell (i.e., profile them) using Gender, Flag_express, Flag-Multibuyer and Flag_bigcity. **(3 points)**

Solution: The highest RFM cell is 555 with a customer count of 780 and the lowest RFM cell is 111 with a customer count of 1019.

On profiling RFM cells 555 & 111 across gender we can observe that males and females are equally distributed in the both the cells. For Express shipping flag, we see that cell 111 doesn't have any customer who uses express shipping while 31.79% of customers from cell 555 use express shipping.

Similarly, cell 111 doesn't have any customers from big cities while 14.62% of all cell 555 customers come from big cities. None of the customers from cell 111 are multi-buyers whereas 10.26% of all cell 555 customers have shopped multiple times.

RFM Cell	Gender	
	Female	Male
111	708 69.48%	311 30.52%
555	564 72.31%	216 27.69%

RFM Cell	Express Shipping Flag	
	No	Yes
111	1019 100%	0 0%
555	532 68.21%	248 31.79%

RFM Cell	Bigcity Flag	
	No	Yes
111	1019 100%	0 0%
555	666 85.38%	114 14.62%

RFM Cell	Multibuyer Flag	
	No	Yes
111	1019 100%	0 0%
555	700 89.74%	80 10.26%

Table (3)

4. Assume the average profit from a sale is $15 and cost of promotion is $1.25. How many (and which) of the RFM cells exceed the break-even cut-off? Report a table that shows following: RFM bin number, Count (number of observations in that RFM cell) and the % yes response in the RFM Cell. (**3 points**)

Solution: Based on the given value of average profit and cost of promotion, break-even cutoff value is 0.083 ($1.25 / $15). Out of all the RFM cells, 51 cells have the percentage of responding customers more that the cutoff value (8.33%) indicating that these cells are profitable. The RFM cell numbers, number of customers and percentage of responding customers in each cell are shown in the table below –

#	RFM	Number of Customers	% Yes Response
1	5 5 4	287	24.04%
2	4 5 4	300	24.00%
3	5 5 5	780	23.72%
4	5 5 2	136	23.53%
5	5 5 3	302	22.85%
6	5 4 2	343	22.45%
7	4 5 5	757	21.93%
8	5 2 5	255	20.00%
9	4 5 3	290	19.66%
10	5 4 4	300	19.33%
11	5 4 1	180	18.33%
12	4 5 2	144	18.06%
13	3 5 4	503	17.89%
14	5 2 4	429	17.25%
15	4 4 4	281	17.08%
16	5 1 5	120	16.67%
17	5 4 5	367	16.35%
18	3 5 5	1215	16.05%
19	4 4 3	310	15.16%
20	5 4 3	299	15.05%
21	3 5 3	502	14.94%
22	4 4 5	380	13.95%
23	5 2 3	467	13.92%
24	5 2 1	583	13.89%
25	2 5 5	1248	13.38%
26	4 4 2	361	13.30%
27	4 1 5	113	13.27%
28	3 4 5	620	13.06%
29	5 1 4	445	12.81%
30	5 1 2	514	12.65%
31	2 5 3	495	12.53%

32	5	2	2	535	12.15%
33	3	4	4	509	11.79%
34	3	4	3	497	11.67%
35	4	2	3	457	11.60%
36	5	1	1	715	11.33%
37	3	5	2	225	11.11%
38	4	2	2	499	10.62%
39	4	1	3	452	10.62%
40	4	4	1	179	10.61%
41	3	4	2	504	10.52%
42	4	1	4	461	10.20%
43	5	1	3	451	9.76%
44	3	2	4	782	9.72%
45	2	5	4	471	9.55%
46	2	4	2	568	9.51%
47	2	5	2	242	9.50%
48	4	2	1	605	9.09%
49	4	2	5	244	9.02%
50	2	4	5	640	8.91%
51	4	2	4	435	8.74%

Deliverables:
As you complete the exercise, create a report in Microsoft Word and in this report answer the questions in the exercise description. Copy and paste supporting tables/diagrams as needed to justify any of your answer. Make sure you *print your name, student ID#, student email on the cover page* of the report and turn-in the report as communicated by your instructor. Please also put a running *header/footer with your name, on each page of your exercise* solution report. Failure to follow these instructions will result in deduction of points