Online retail Promotions - GLM

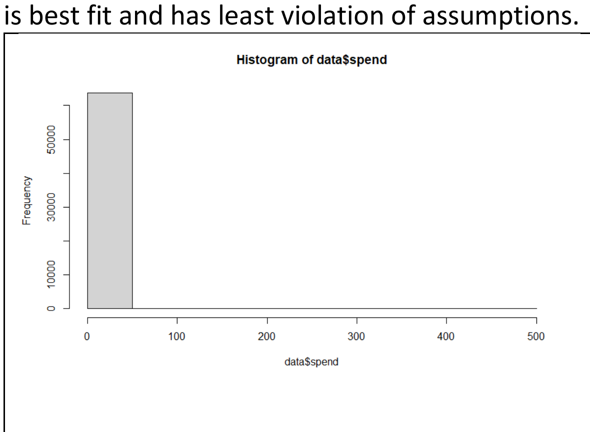
Predictor table:

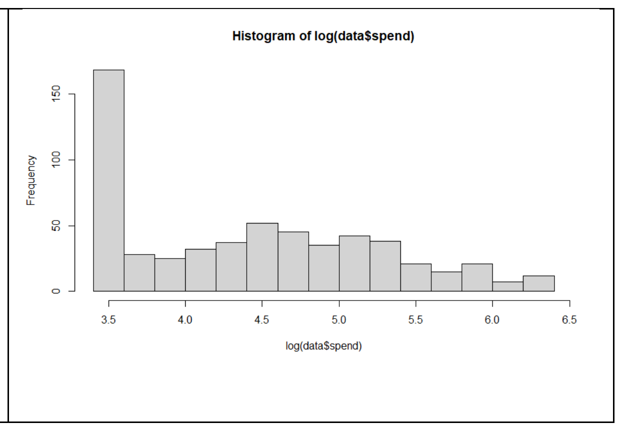
Predictor	Effect	Rationale
DV: spend		
recency	+/-	Recency can be a predictor in behavior pattern of a consumer and if they haven't purchased anything from a longer duration then that can be a contributor in their spending pattern
historysegment	+	Past Spending pattern may be helpful in determining the spend
zipcode	+	Zipcode provides the demography of the customer thus helpful, whether the customer is rural, Urban or Suburban
newcustomer	+/-	Whether a customer is a repeater or a new customer, it helps in understanding the behavior pattern
channel	+/-	Channel is the mode of promotion and would help in understanding the promotion mode
campaign	+/-	Which campaign is working more on the select demographic
Gender	+/-	This is a category created by me using Column men and women, and has factors, Men, Women and Both.
conversion	+	If a potential customer is converted into a client then it is a contributing factor in spend
visit	+	Have they visited the store can be a determining factor in spending pattern
Excluded: Men,Women as a new category column of gender was created containing variable both,men and women. History: as history segment is range of history		

Exploratory Data Analysis

Histograms of DV:

The distributions of **Spend** are right-skewed, and hence, OLS regression will not be suitable. The distributions of **log-Spend** are close to normal, and therefore, more suited for MLS regression.The variable is Poison distribution hence we'll have to use MLS models. Then we can determine which model is best fit and has least violation of assumptions.

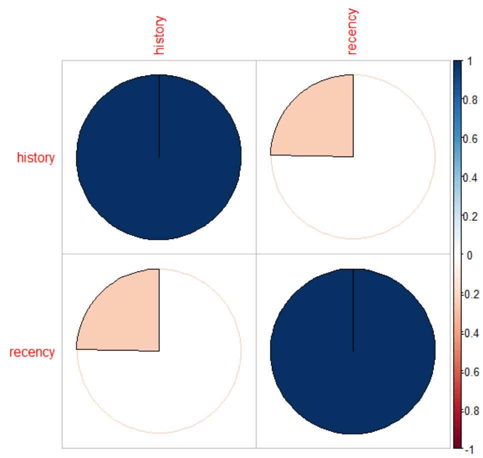




Correlations:
Almost every predictor is a factor variable; the only two continuous variables are history and Recency. However, the correlation between them is 0.24. So, we can use recency and history as a predictor of Spend.

```
correlation_matrix <- cor(data[c("history", "recency")])
corrplot(correlation_matrix, method = "pie")
print(correlation_matrix)
```

```
          history    recency
history 1.0000000 -0.2465913
recency -0.2465913  1.0000000
```



Regression Analysis

```
glm_model1 <- glm(spend ~ recency + historysegment + Gender + conversion + visit + zipcode
+ newcustomer + channel + campaign, data = data,  family = gaussian(link = "identity"))
glm_model2 <- glm(spend ~ recency + historysegment + Gender + conversion + visit + zipcode
+ newcustomer + channel + campaign, data = data, family = gaussian(link = "log"))
glm_model3 <- glm(spend ~ recency + historysegment + Gender + conversion + visit + zipcode
+ newcustomer + channel + campaign, data = data,  family = gaussian(link = "inverse"))
```

	Dependent variable:		
	spend		
	normal	glm: gaussian link = log	glm: gaussian link = inverse
	(1)	(2)	(3)
recency	0.007 (0.012)	-0.0001 (0.001)	0.0001*** (0.00001)
1,000 +	0.921*** (0.310)	0.271*** (0.016)	-0.002*** (0.0001)
200	-0.073 (0.109)	-0.099*** (0.011)	0.001*** (0.0001)
350	-0.174 (0.123)	-0.199*** (0.011)	0.002*** (0.0001)
500	-0.135 (0.157)	-0.103*** (0.013)	0.001*** (0.0001)
750	-0.076 (0.183)	-0.123*** (0.016)	0.001*** (0.0001)
1,000	-0.655** (0.266)	-0.554*** (0.028)	0.006*** (0.0003)
GenderFemale	-0.143 (0.155)	-0.094*** (0.012)	0.0004*** (0.0001)
GenderMen	0.089 (0.155)	0.193*** (0.011)	-0.002*** (0.0001)
conversionYes	115.355*** (0.440)	4.793*** (0.109)	-0.991*** (0.106)
visitYes	0.024 (0.119)	0.029 (0.117)	-0.0003 (0.114)
zipcodeSurburban	0.134 (0.121)	0.147*** (0.011)	-0.001*** (0.0001)
zipcodeUrban	0.118 (0.123)	0.125*** (0.011)	-0.001*** (0.0001)
newcustomer	0.0001 (0.089)	0.017** (0.009)	-0.0001 (0.0001)
channelPhone	0.032 (0.147)	-0.010 (0.012)	0.001*** (0.0001)
channelWeb	0.034 (0.147)	-0.012 (0.011)	0.001*** (0.0001)
campaignNo E-Mail	0.025 (0.100)	0.006 (0.010)	0.00003 (0.0001)
campaignWomens E-Mail	0.086 (0.099)	0.114*** (0.008)	-0.001*** (0.0001)
Constant	0.879*** (0.262)	-0.170*** (0.049)	1.000*** (0.043)
Observations	64,000	64,000	64,000
Log Likelihood	-239,696.400	-238,277.300	-238,009.400
Akaike Inf. Crit.	479,430.800	476,592.500	476,056.900
Note:			*p<0.1; **p<0.05; ***p<0.01

Model 2 i.e, the log model is the best fit out of the three models

Interpretation

1. *How did the promotion campaigns work relative to the control group? Did the men's promotions work better than the women's promotion (or vice versa) and by how much?*
   Campaign men email being the base, campaign no-email is 0.6% draws more spendings. Campaign women email has 11.4% more spending than men email campaign. Thus, Women's promotion is working better than men's email promotion. If men spend 100$ then females through email campaigns will spend 111.4$

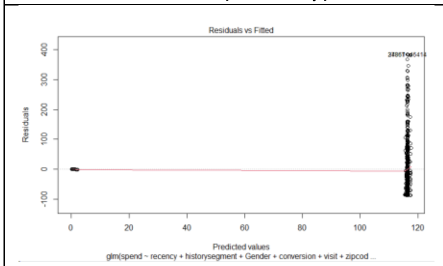
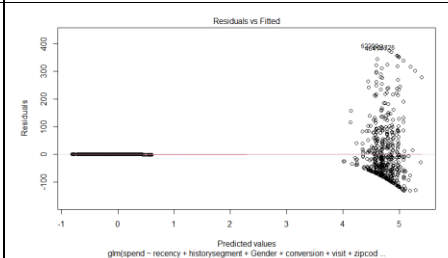
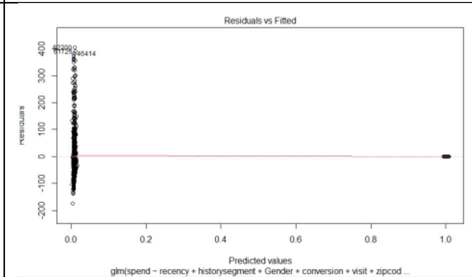
2. *Should we target these promotions to new customers (who joined over the last 12 months) rather than to established customers, or vice versa?*
   Compared to returning customers, the new customers spend 1.7% more on every dollar spent. We can target these promotions on new as well as old customers as the difference in spending pattern is low.

3. *Should we target these promotions to customers who have a higher (or lower) history of spending over the last year?*
   Yes, we should target these promotions to the revisiting customers with a spending pattern over 1000$ as the revisiting customers spend 27%. With other spending range categories, we observe a negative growth i.e. 200-350$- -ve9.9%, $350 - $500: -19.9%, $500 - $750: 10.3%, $750 - $1,000: 12.3%.
4. *Did the promotions work better for phone or web channel?*
   The Promotion worked best for both the channels when the multichannel approach is used. When we take multichannel as a base and compare it with other channels in accordance with spending, with each dollar spent: there is a -1.0% change for channel phone and -1.2% change for web. That is if 100$ are spent using multichannel then 99$ would be spend if only channel phone is used and 98.8$ for web.

5. *Will the promotions work better if the men's promotion is targeted at customers who bought men's merchandise over the last year (compared to those who purchased women's merchandise), and if the women's promotion would work better if targeted at customers who bought women's merchandise over the last year?*
   19.3% more spending will happen if the men's are target compared to promotional targeting of both the genders. If both the gender's are targeted then the female sales drop by 9.3%.Based on the data provided men have traditionally spent more and the new customers when being men spent more, hence if promotion is targeted at men's then more spending will happen.

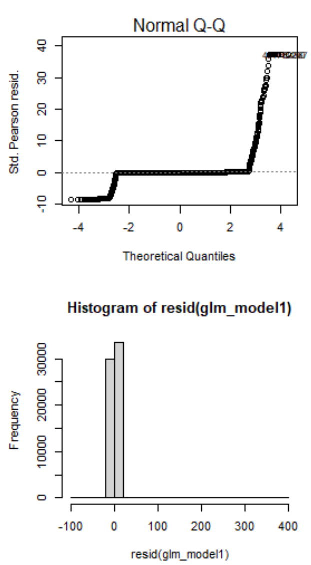
Assumptions

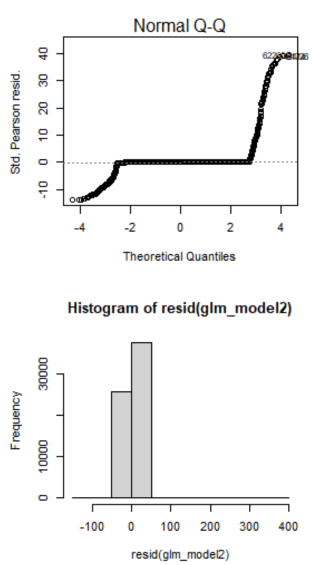
Linearity(Fail): relationship NOT required between Y and Xs

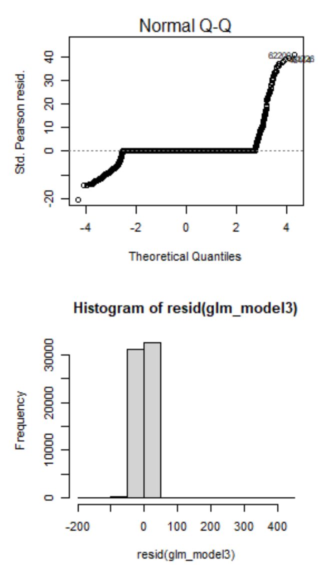
Model 1(Identity)	Model 2(Log)	Model 3 (Inverse)
		

Normality (Fail): There is deviations from normality for residuals from all the three models

Model 1(Identity)	Model 2(Log)	Model 3 (Inverse)

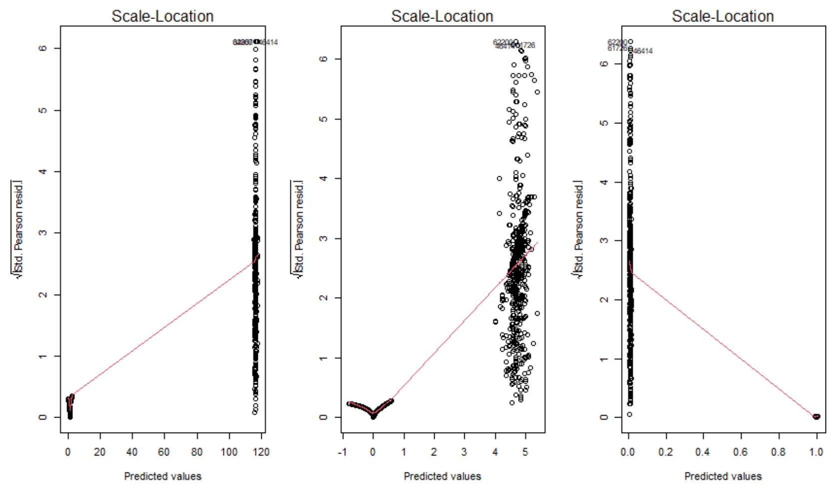






Homoscedasticity (Fail):

Model 1(Identity)	Model 2(Log)	Model 3 (Inverse)



Multicollinearity (Pass): VIF tests shows that all independent variables in both count and amount models have GVIF^(1/(2*Df)) values less than 5, indicating no significant multicollinearity.

Model 1(Identity)

	GVIF	Df	GVIF^(1/(2*Df))
recency	1.068898	1	1.033875
historysegment	1.873337	6	1.053702
Gender	1.223569	2	1.051737
conversion	1.056394	1	1.027811
visit	1.091214	1	1.044612
zipcode	1.002892	2	1.000722
newcustomer	1.211644	1	1.100747
channel	1.283674	2	1.064422
campaign	1.008555	2	1.002132

Model 2(Log)

	GVIF	Df	GVIF^(1/(2*Df))
recency	1.121472	1	1.058996
historysegment	2.733496	6	1.087410
Gender	1.401243	2	1.087999
conversion	7.159681	1	2.675758
visit	7.161389	1	2.676077
zipcode	1.060651	2	1.014830
newcustomer	1.525993	1	1.235311
channel	1.326059	2	1.073101
campaign	1.088333	2	1.021387

Model 3 (Inverse)

	GVIF	Df	GVIF^(1/(2*Df))
recency	1.118639	1	1.057657
historysegment	3.845563	6	1.118785
Gender	1.269601	2	1.061492
conversion	7.193965	1	2.682157
visit	7.193965	1	2.682157
zipcode	1.309800	2	1.069797
newcustomer	1.885030	1	1.372964
channel	1.690366	2	1.140237
campaign	1.178288	2	1.041868

Independence (Pass): Durbin-Watson test shows residuals in both count and amount models have DW statistic in the [1.5-2.5] range, indicating no severe violation of the independence assumption.

Model 1(Identity)	Model 2(Log)	Model 3 (Inverse)
data: glm_model1 DW = 1.9994, p-value = 0.4689	data: glm_model2 DW = 1.9994, p-value = 0.4689	data: glm_model3 DW = 1.9994, p-value = 0.4689