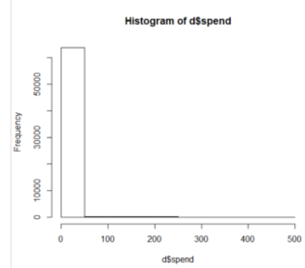
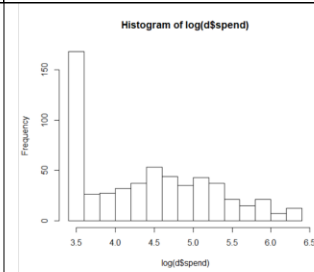
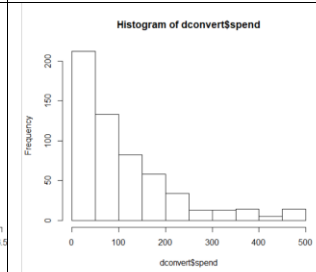
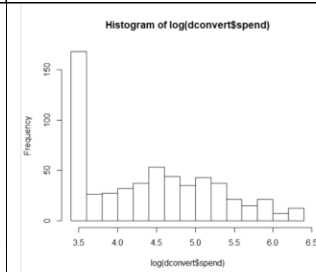
Online Retail Promotions

All spend data		Converted users only	
without log transform	with log transform	without log transform	with log transform
			
Range: 0-499, Mean 1.05		Range: 30-499 Mean 116.4	

- Spend data has a lot of zeroes in raw (all user) data.
- Converted user data has no zero, but spend is still not normal. Hence, it is perhaps best to try GLM models with non-Gaussian distributions (e.g., Poisson).

Predictor	Effect	Rationale
campaign	+	We want to examine the effect of campaigns on customer spend; customers receiving the promotional campaign are expected to spend more
recency	+	Recent customers may be predisposed to spending more
history	+	Customers with a history of high prior purchases may be expected to spend more
mens/womens	+/-	Customers who purchased mens (womens) products last year are more likely to respond to mens (womens) campaign
womens	+/-	This variable helps us understand the gender-based product bought by customer last year.
zipcode	?	Urban shoppers may have different spending patterns than rural or suburban shoppers
newcustomer	+	New customers may be more excited about online purchases
channel	+	Some shoppers may prefer web or online channels; but since we have some shoppers that used both channels, we have to split this data into separate variables for web and online channel shoppers
Excluded Factors		
historysegment	n/a	Correlated with history. Omit as continuous variable history is more granular than categorical variable historysegment
visit, conversion	n/a	Spend = 0 (constant) if visit = 0 or conversion = 0

Modeling

```
m1 = hurdle(spend ~ campaign*mens + campaign*womens + campaign*newcustomer +
            campaign*history + campaign*channelphone + campaign*channelweb +
            recency + zipcode | visit, data=d, link="logit", dist="negbin")
```

```
m2 = zeroinfl(spend ~ campaign*mens + campaign*womens + campaign*newcustomer +
              campaign*history + campaign*channelphone + campaign*channelweb +
              recency + zipcode | visit, data=d, link="logit", dist="negbin")

m0 = hurdle(spend ~ campaign + mens + womens + newcustomer + history +
            channelphone + channelweb + recency + zipcode | visit,
            data=d, link="logit", dist="negbin")
```

Model justification:

Why so many interaction terms?

- We need them to answer the questions asked in part 4 of the assignment.

Why negative binomial models?

- We ran an initial Poisson model, and the dispersion test showed overdispersion (lambda=201).

Why hurdle and zero inflated models?

- Because of excess zeroes: people who did not even visit the website (~54,000 out of 64,000 targeted customers) have spend = 0

What is/are good logit predictors for the hurdlemodel?

- Visit seems pretty reasonable because customers who did not even visit the website will have spend = 0.

	Dependent variable: spend		
	m0 (baseline) (no interactions)	m1 (hurdle) (with interactions)	m2 (zero inflated) (with interactions)
campaignMen	0.003 (0.089)	-0.096 (0.374)	-0.096 (0.374)
campaignWomen	0.104 (0.096)	0.491 (0.418)	0.491 (0.418)
mens	0.137 (0.102)	0.493** (0.238)	0.493** (0.238)
womens	-0.128 (0.101)	0.209 (0.232)	0.209 (0.232)
newcustomer	-0.005 (0.074)	-0.249 (0.184)	-0.250 (0.184)
history	0.00004 (0.0001)	-0.00005 (0.0002)	-0.00005 (0.0002)
channelphone	-0.091 (0.104)	-0.326 (0.234)	-0.326 (0.234)
channelweb	-0.073 (0.105)	-0.303 (0.230)	-0.303 (0.230)
recency	-0.008 (0.010)	-0.004 (0.010)	-0.004 (0.010)
zipcodeRural	-0.090 (0.095)	-0.118 (0.096)	-0.119 (0.096)
zipcodeSurburban	0.049 (0.076)	0.038 (0.076)	0.038 (0.076)
campaignMen:mens		-0.293 (0.279)	-0.293 (0.279)
campaignWomen:mens		-0.752** (0.311)	-0.752** (0.311)
campaignMen:womens		-0.168 (0.273)	-0.168 (0.273)
campaignWomen:womens		-0.845*** (0.304)	-0.845*** (0.304)
campaignMen:newcustomer		0.322 (0.214)	0.323 (0.214)
campaignWomen:newcustomer		0.289 (0.224)	0.290 (0.224)
campaignMen:history		0.0001 (0.0003)	0.0001 (0.0003)
campaignWomen:history		0.0001 (0.0003)	0.0001 (0.0003)
campaignMen:channelphone		0.188 (0.278)	0.188 (0.278)
campaignWomen:channelphone		0.389 (0.297)	0.389 (0.298)
campaignMen:channelweb		0.228 (0.276)	0.228 (0.276)
campaignWomen:channelweb		0.317 (0.296)	0.317 (0.296)
Constant	4.862*** (0.178)	4.797*** (0.324)	4.797*** (0.324)
Observations	64,000	64,000	64,000
Log Likelihood	-5,464.107	-5,457.133	-5,457.127

Model assumptions: GLM models are robust to linearity, multivariate normality, and homoscedasticity violations. But they are subject to multicollinearity and independence violations, in addition to overdispersion and excess zero violations of Poisson models.

Multicollinearity: Passed VIF tests shows GVIF^(1/(2*Df)) values (equivalen t to VIF values) of all variables below 5. `vif(m0)`	<pre> GVIF Df GVIF^(1/(2*Df)) campaign 5.117314 2 1.504044 history 2.656226 1 1.629793 recency 7.196000 1 2.682536 mens 5.201364 1 2.280650 womens 5.646536 1 2.376244 zipcode 2.674161 2 1.278783 newcustomer 2.174551 1 1.474636 channelphone 5.292357 1 2.300512 channelweb 6.085430 1 2.466866</pre>
Independence: Passed Durbin-Watson test shows DW statistic = 2.006 and p=0.78	`dwtest(m0)` `DW = 2.006, p-value = 0.7757`
Overdisperson: Negative binomial models and are robust to overdispersion.	
Excess zeros: Hurdle and zero inflated models are robust to excess zeroes.	

Which model is best: Models m2 is the “best” model since it passes all assumptions and will be used for interpretation below. According to this model:

```
log(spend) = 4.80 – 0.10*campaignMen + 0.49*campaignWomen + 0.49*mens +0.21*womens
             - 0.25*newcustomer - 0.33*channelphone – 0.30*channelweb - 0.00*history
             – 0.00*recency – 0.12*zipcodeRural + 0.04*zipcodeSurburban
             – 0.29*campaignMen:mens – 0.71*campaignWomen:mens
             – 0.17*campaignMen:womens – 0.85*campaignWomen:womens
             + 0.32*campaignMen:newcustomer + 0.29*campaignWomen:newcustomer
             + 0.00*campaignMen:history + 0.00*campaignWomen:history
             + 0.19*campaignMen:channelphone + 0.39*campaignWomen:channelphone
             + 0.23*campaignMen:channelweb + 0.32*campaignWomen:channelweb
```

- ***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?***
  From model m5, the marginal effects of mens’ and womens’ campaign relative to no campaign is (we ignore the interaction term of history whose beta is 0.001 and too small to be of significance):
  d(spend)/d(campaignMen) = -0.10 - 0.29*mens - 0.71*womens + 0.32*newcustomer + 0.19*channelphone + 0.23*channelweb
  d(spend)/d(campaignWomen) = 0.49 - 0.17*mens - 0.85*womens + 0.29*newcustomer + 0.39*channelphone + 0.32*channelweb
  The difference in marginal effects between men and women is:
  -0.59 – 0.13*mens + 0.68*womens + 0.03*newcustomer – 0.20*channelphone – 0.09*channelweb
  The overall effect of men’s vs women’s campaign depends on whether recipients purchased men’s or women’s products last year, whether they are a new customer, and their web/phone channel preference. If all those things are constant, then men’ campaign underperformed women’s campaign by 59%, and it even underperformed no campaign by 10% (spend on log scale).

- ***Should we target these promotions to new customers (who joined over the last 12 months) rather than to established customers, or vice versa?***
  d(spend)/d(newcustomer) = -0.25 + 0.32*campaignMen + 0.29*campaignWomen

New customers have a -25% effect compared to old customers in the no campaign group, but new customers who received the men's campaign had a 7% net increase in customer spend relative to no campaign, and those who received the women's campaign had a 4% increase in spend.

- ***Should we target these promotions to customers who have a higher (or lower) history of spending over the last year?***
  d(spend)/d(history) = -0.00 + 0.00*campaignMen + 0.00*campaignWomen
  History had zero effect on customer spend for both men's and women's campaign.

- ***Did promotions work better for phone or web channel?***
  d(spend)/d(channelphone) = -0.33 + 0.19*campaignMen + 0.39*campaignWomen
  d(spend)/d(channelweb) = -0.30 + 0.23*campaignMen + 0.32*campaignWomen
  Both phone and web channel worked poorly if customers received no campaign (-33% and -30%). Men's campaign increased phone spend to -14% and web spend to -7%, while women's campaign increased phone spend to +2% and web spend to +2%. Hence, women's campaign definitely improved customer spend over no campaign. Men's campaign reduced deficit spend compared to no campaign, but still resulted in negative spend.

- ***Will 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?***
  d(spend)/d(mens) = 0.49 – 0.29*campaignMen - 0.75*campaignWomen
  d(spend)/d(womens) = 0.21 – 0.17*campaignMen – 0.85*campaignWomen
  Men's campaign directed at customers who bought men's products last year had 29% less effect on spend relative to no campaign, while men's campaign directed at customers who bought women's product last year had a -17% effect. However, women's campaign directed at customers who bought women's products last year had a -85% effect relative to no campaign, while women's campaign directed at men's products had a -75% effect. Hence, these campaigns seem to have the best effects if directed at new customers rather than to customers who bought products over the last year. In particular, the women's campaign had significantly worse effect than men's campaign.