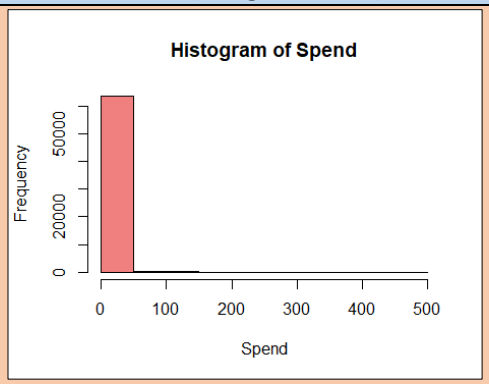
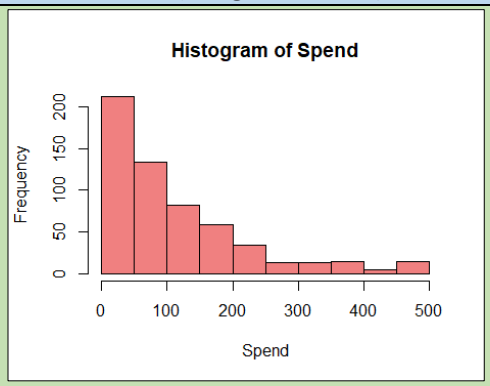
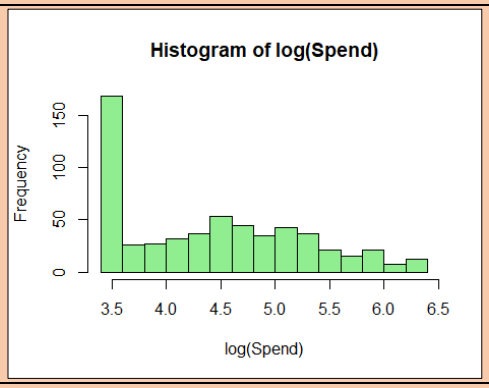
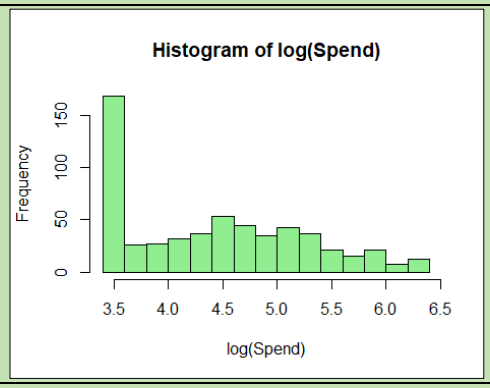
Retail Customer Spend Prediction:

I. Examine the "spend" variable that we want to predict and explain step-by-step what you would do to create a model to explain customer spend (bullet points are fine). What model(s) is(are) appropriate for this analysis and why. Run appropriate visualizations if necessary and document your work in your answer. Be sure to read Question 4 below to get a sense of the analysis the client is looking for.

Spend: It is the actual amount spent (In dollars) in the following two weeks of the e-mail campaign to track the customers purchase behavior. In this case, the retailer is targeting three groups where in we have two treatment groups (Men and Women who received e-mails) and one control group (Not received an e-mail).

Before (Range: 0 – 499)	After (Range: 30 – 499)
	
Without Log Transform	Without Log Transform
	
With Log Transform	With Log Transform

- Dependent variable (spend) seems to follow a non-normal distribution. Transformations does not seem to have any effect.
- Based on the above histograms its very clear that OLS will not work. We will have to use GLM models.
- On top of that, our DV has zeros up until at least the third quartile. This means we will have to correct for excess zeros and round off to get a perfect count DV.
- From Question 4, to find if the e-mail campaign was successful, we need to have some interaction terms to understand what effect modification a variable has on the other with respect to the DV. This is an example of "Does X1 effect the relationship between X2 and y".
- Our X1 = campaign categorical variable. X2 = All other categorical variables and y is the dependent variable (spend).

II. Create a table of predictors for our dependent variable, listing all relevant predictors, the sign of their hypothesized effects, and a short 1-sentence rationale for each effect.

Factor Effects:
Positive Effect: If there is a direct proportionality (+X then +y / -X then -y) between the predictor variable (X) and the response variable (y), we can say there is a positive effect.

Negative Effect: If there is an inverse proportionality (+X then -y / -X then +y) between the predictor variable (X) and the response variable (y), we can say there is a negative effect.

Relevant Factors for Spending		
Predictor	**Effect**	**Rationale**
recency	+	Months since last purchase is included as customers tend to spend more if there is no spending for an extended period.
history	+/-	Higher spending in the past year could mean the customer is a frequent shopper and could be expected to spend more with better promotional efforts.
mens	+/-	Customer purchased Men merchandize based on the e-mail campaign.
womens	-/+	Customer purchased Women merchandize based on the e-mail campaign.
zipcode	+	Customers in Urban areas tend to spend more due to higher accessibility of online services than those in the rural areas and vice versa.
newcustomer	+	This could help us understand what we can do to improve our campaign efforts.
channel	+	This variable is refactored into two new variables to see the individual relationship between channels and spending based on the promotional effect.
campaign	+	Key predictor as we are trying to find out if the promotional efforts had any impact on the spending.

Irrelevant Factors		
Predictor	**Effect**	**Rationale**
historysegment	No Effect	This variable is similar to history but shows the spending in terms of range. Can be omitted.
visit	No Effect	Visits will not have any effect on the spending.
conversion	No Effect	Conversion is similar to our dependent variable so can be omitted.

III. Run alternative models to test for the effects of the hypothesized predictors. Be sure to test the assumptions of these models and modify them as necessary. Present the best 3 models and their output in a nice, compact manner. Also justify your choice of these models and include your assumptions testing results.

Summary:
I started with a log transformed OLS regression model. Then moved to Poisson and QuasiPoisson models which had high β and log likelihood values. Since our model has high dispersion (~94.8), I tried Negative Binomial models which seemed like a good estimate. Finally, I tried correcting our model for excess zeros by using hurdle and zero inflated models. Based on the analysis, the best 3 models are as follows.

Model m5: Hurdle model with key variables and interaction terms (Hurdle method builds two models where the first is with zero count and the other is with positive count).
Model m7: Zero inflated model with key variables and interaction terms.
Model m8: Zero inflated model with all variables without interaction terms.

```r
#Regression models
m5 <- hurdle(spend ~ recency + campaign*history campaign*zipcode + campaign*newcustomer
+ campaign*phone + campaign*web + campaign*mens + campaign*womens | visit + conversion,
data=orc_rd, link="logit", dist="negbin")

m7 <- zeroinfl(spend ~ recency + campaign*history campaign*zipcode + campaign*newcustomer
+ campaign*phone + campaign*web + campaign*mens + campaign*womens | visit + conversion,
data=orc_rd, link="logit", dist="negbin")

m8 <- zeroinfl(spend ~ recency + history + zipcode + newcustomer + phone + web + campaign + mens
+ womens | visit + conversion, data=orc_rd, link="logit", dist="negbin")
```

```r
#Stargazer
stargazer(m5, m7, m8, type='text', single.row = TRUE)
```

```
##
## ==============================================================================
##                                                 Dependent variable:
##                                         --------------------------------------
##                                                        spend
##                                              hurdle         zero-inflated
##                                                              count data
##                                                (1)           (2)        (3)
## ------------------------------------------------------------------------------
## recency                                  0.003 (0.010)     0.003      0.008
## campaignMens E-Mail                      0.183 (0.438)     0.183      0.003
## campaignWomens E-Mail                    0.698 (0.471)     0.697      0.104
## history                                 -0.0001 (0.0002)  -0.0001    0.00004
## zipcodeSurburban                         0.429** (0.199)   0.428      0.139
## zipcodeUrban                             0.202 (0.203)     0.201      0.090
## newcustomer                             -0.279 (0.184)    -0.279     -0.005
## phone                                   -0.309 (0.235)    -0.308     -0.090
## web                                     -0.285 (0.231)    -0.284     -0.072
## mens                                     0.531** (0.246)   0.530      0.136
## womens                                   0.254 (0.239)     0.254     -0.128
## campaignMens E-Mail:history              0.0001 (0.0003)   0.0001
## campaignWomens E-Mail:history            0.0003 (0.0003)   0.0003
## campaignMens E-Mail:zipcodeSurburban    -0.387 (0.246)    -0.386
## campaignWomens E-Mail:zipcodeSurburban  -0.317 (0.259)    -0.317
## campaignMens E-Mail:zipcodeUrban        -0.269 (0.250)    -0.269
## campaignWomens E-Mail:zipcodeUrban       0.097 (0.260)     0.097
## campaignMens E-Mail:newcustomer          0.346 (0.214)     0.345
## campaignWomens E-Mail:newcustomer        0.307 (0.224)     0.306
## campaignMens E-Mail:phone                0.183 (0.279)     0.183
## campaignWomens E-Mail:phone              0.336 (0.299)     0.335
## campaignMens E-Mail:web                  0.215 (0.276)     0.214
## campaignWomens E-Mail:web                0.273 (0.297)     0.272
## campaignMens E-Mail:mens                -0.301 (0.285)    -0.300
## campaignWomens E-Mail:mens              -0.867*** (0.317)  -0.865
## campaignMens E-Mail:womens              -0.199 (0.278)    -0.199
## campaignWomens E-Mail:womens            -0.930*** (0.309)  -0.928
## Constant                                4.436*** (0.383)   4.438      4.666
## ------------------------------------------------------------------------------
## Observations                                 64,000        64,000     64,000
## Log Likelihood                             -3,282.181    -3,282.781 -3,293.330
## ==============================================================================
## Note:                                              *p<0.1; **p<0.05; ***p<0.01
```

The modified assumptions for our models are multicollinearity, independence, overdispersion and excess zeros.

Assumption	DV Model: m5 (Spend)
Multicollinearity: Passed • **Variance Inflation Factor (VIF – GVIF^(1/(2*Df)))** 1. **VIF = 1/T (Where T = 1 – R^2 , T < 0.1 is indicative of multicollinearity).** 2. **VIF > 5 indicates multicollinearity. VIF > 10 is strong evidence of multicollinearity.**	``` ## GVIF Df GVIF^(1/(2*Df)) ## recency 3.558708 1 1.886454 ## campaign 2828.549789 2 7.292744 ## history 14.096475 1 3.754527 ## zipcode 130.093484 2 3.377255 ## newcustomer 13.844696 1 3.720846 ## phone 27.676793 1 5.260874 ## web 30.603797 1 5.532070 ## mens 31.469302 1 5.609751 ## womens 32.387879 1 5.691035 ## campaign:history 43.420619 2 2.566989 ## campaign:zipcode 876.995869 4 2.332785 ## campaign:newcustomer 33.293455 2 2.402092 ## campaign:phone 147.570908 2 3.485380 ## campaign:web 200.282004 2 3.761928 ## campaign:mens 163.997799 2 3.578570 ## campaign:womens 211.939442 2 3.815513 ``` **This model passed the VIF test as most of these are either control variables or are product of two variables all values are below 10.**

Assumption	DV Model: m5 (Spend)
Independence: Passed • **Durbin-Watson's Test (DW)** 1. **Ho: Residuals are not linearly auto-correlated.** 2. **DW ~ [0, 4]; values around 2 (i.e., 1.5 to 2.5) suggests no autocorrelation.**	**DW = 2.0062, p-value = 0.7824** **This model passed the Durbin-Watson's test.**

Assumption	DV Model: m5 (Spend)
Overdispersion: Passed	Observed variance (dispersion) for Poisson family is 1. The dispersion test returns 94.8 for the Poisson model so we have used Negative Binomial models for our analysis.

Assumption	DV Model: m5 (Spend)
Excess Zeros: Passed	To correct for excess zeros, we used hurdle model on Negative Binomial (Dispersion adapted).

IV. Based on your analysis, answer the following questions (using marginal effects, not statistical significance).
[Selected Model = m5]

- **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 our analysis, Men's campaign performed better by 18.3% from the control group. Women's campaign performed better by 69.8% from the control group. Men's campaign performed worse by 51.5% from Women's campaign.

- **Should we target these promotions to new customers (who joined over the last 12 months) rather than to established customers, or vice versa?**

  Men's campaign effect is 6.7% higher than the control group with respect to spending.
  Women's campaign effect is 2.8% higher than the control group with respect to spending.

  Ln(spend) = -0.279*(newcustomer) + 0.346*(campaignMens E-Mail)*(newcustomer)
  Ln(spend) = -0.279*(newcustomer) + 0.307*(campaignWomens E-Mail)*(newcustomer)

- **Should we target these promotions to customers who have a higher (or lower) history of spending over the last year?**

  Considering history there does not seems to be any significance effect. When we consider the interaction terms there is still no effect. So, targeting promotion efforts towards customers with high/low spending history won't have any marginal effect on spending.

- **Did the promotions work better for phone or web channel?**

	On Phone with respect to spending	On Web with respect to spending
Men's campaign effect	-12.6% (less than No campaign)	+2.7% (more than No campaign)
Women's campaign effect	-7% (less than No campaign)	-1.2% (less than No campaign)

  Ln(spend) = -0.309*(phone) + 0.183*(campaignMens E-Mail)*(phone) + 0.336*(campaignWomens E-Mail)*(phone)
  Ln(spend) = -0.285*(web) + 0.215*(campaignMens E-Mail)*(web) + 0.273*(campaignWomens E-Mail)*(web)

- **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?**

	On customers that purchased Mens merchandize	On customers that purchased Womens merchandize
Men's campaign effect	+23% (more than No campaign)	+5.5% (more than No campaign)
Women's campaign effect	-33.6% (less than No campaign)	-67.6% (less than No campaign)

  Ln(spend) = 0.531*(men) + 0.254*(women) - 0.301*(campaignMens E-Mail)*(men) - 0.199*(campaignMens E-Mail)*(women)
  Ln(spend) = 0. 531*(men) + 0.254*(women) - 0.867*(campaignWomens E-Mail)*(men) - 0.930*(campaignWomens E-Mail)*(women)

V. Reflect on the quality of your analysis, and comment on things you can do to further improve this analysis.

The final models from our analysis seems to give the most accurate estimations of the promotional efforts of the retailer. Based on the data we have decided to go with GLM rather than a log transformed OLS which was an appropriate choice. To improve this model, other combinations of interaction terms could possibly lower the log likelihood and result in better β values.