Ramkishore Rao, HW #s 4 and 5

This memorandum presents the exploratory analysis, preparation of data in support of the development of predictive models, and the results of the modeling of data that provides information on customer's website site behavior for an online retailer.

In order to follow the format required for this homework assignment, snippets of R code, explaining the process, observations, results, and findings are included. Notations are provided on the tables and the graphics to explain the observations and findings. Only minimal text is provided when necessary to clarify the presented information.

1.0 Exploratory Data Analysis (Part a)

Preliminary data analysis included assessment of number, domain types of predictor variables, select scatter plots, evaluation of missingness of data, estimation of potential outliers, and summary of stats of select numeric variables.

35 variables in strain dataset with 70,071 records. Select character variables were converted to factors to facilitate subsequent analysis.

```
glimpse(train)
## Rows: 70,071
## Columns: 35
## $ sessionId                    <dbl> 200000120, 400000140, 600000160, 700...
## $ custId                       <int> 1795, 1797, 1799, 1800, 1801, 1803, ...
## $ date                         <chr> "2017-04-25", "2016-09-04", "2016-12...
## $ channelGrouping              <chr> "Social", "Social", "Organic Search"...
## $ visitStartTime               <int> 1493117200, 1473037945, 1483011213, ...
## $ visitNumber                  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, ...
## $ timeSinceLastVisit           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 16825, 0,...
## $ browser                      <chr> "Chrome", "Safari", "Chrome", "Safar...
## $ operatingSystem              <chr> "Windows", "Macintosh", "Windows", "...
## $ isMobile                     <int> 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, ...
## $ deviceCategory               <chr> "desktop", "desktop", "desktop", "de...
## $ continent                    <chr> "Asia", "Americas", "Asia", "Africa"...
## $ subContinent                 <chr> "Southern Asia", "Northern America",...
## $ country                      <chr> "India", "United States", "India", "...
## $ region                       <chr> "Tamil Nadu", "", "", "", "Californi...
## $ metro                        <chr> "", "", "", "", "San Francisco-Oakla...
## $ city                         <chr> "Chennai", "", "", "", "San Francisc...
## $ networkDomain                <chr> "airtel.in", "comcast.net", "", "ipb...
## $ topLevelDomain               <chr> "in", "net", "", "na", "", "fr", "",...
## $ campaign                     <chr> "", "", "", "", "", "", "", "", "", ...
## $ source                       <chr> "quora.com", "youtube.com", "google"...
## $ medium                       <chr> "referral", "referral", "organic", "...
## $ keyword                      <chr> "", "", "", "", "", "", "", "", "", ...
## $ isTrueDirect                 <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, ...
## $ referralPath                 <chr> "/How-can-one-get-a-Google-T-shirt-i...
## $ adContent                    <chr> "", "", "", "", "", "", "", "", "", ...
## $ adwordsClickInfo.page        <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ adwordsClickInfo.slot        <chr> "", "", "", "", "", "", "", "", "", ...
## $ adwordsClickInfo.gclId       <chr> "", "", "", "", "", "", "", "", "", ...
## $ adwordsClickInfo.adNetworkType <chr> "", "", "", "", "", "", "", "", "", ...
## $ adwordsClickInfo.isVideoAd   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ pageviews                    <int> 1, 1, 1, 1, 6, 6, 1, 1, 1, 1, 5, 1, ...
## $ bounces                      <int> 1, 1, 1, 1, NA, NA, 1, 1, 1, 1, NA, ...
## $ newVisits                    <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, NA, 1, 1,...
## $ revenue                      <dbl> 0.00000, 0.00000, 0.00000, 0.00000, ...
```

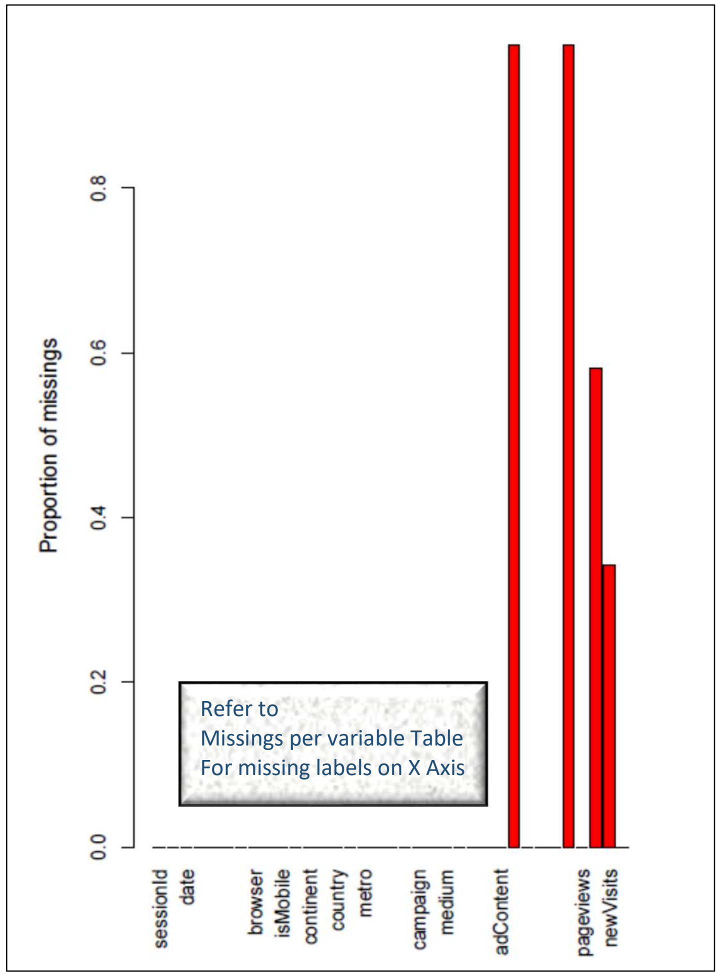
Missings per variable:

```
a<-aggr(dfTrain)
summary(a)
```

Missing values in pageviews, bounces, and newVisits are integer and will be imputed later. A few character variables (e.g., country, etc.) have missing data although not categorized as missing based on " ".

Ramkishore Rao, HW #s 4 and 5

```
##  Missings per variable:
##                       Variable Count
##                      sessionId     0
##                         custId     0
##                           date     0
##                channelGrouping     0
##                 visitStartTime     0
##                    visitNumber     0
##             timeSinceLastVisit     0
##                        browser     0
##                operatingSystem     0
##                       isMobile     0
##                 deviceCategory     0
##                      continent     0
##                   subContinent     0
##                        country     0
##                         region     0
##                          metro     0
##                           city     0
##                  networkDomain     0
##                 topLevelDomain     0
##                       campaign     0
##                         source     0
##                         medium     0
##                        keyword     0
##                   isTrueDirect     0
##                   referralPath     0
##                      adContent     0
##            adwordsClickInfo.page 68260
##            adwordsClickInfo.slot     0
##           adwordsClickInfo.gclId     0
##   adwordsClickInfo.adNetworkType     0
##       adwordsClickInfo.isVideoAd 68260
##                      pageviews     8
##                        bounces 40719
##                      newVisits 23944
##                        revenue     0
```



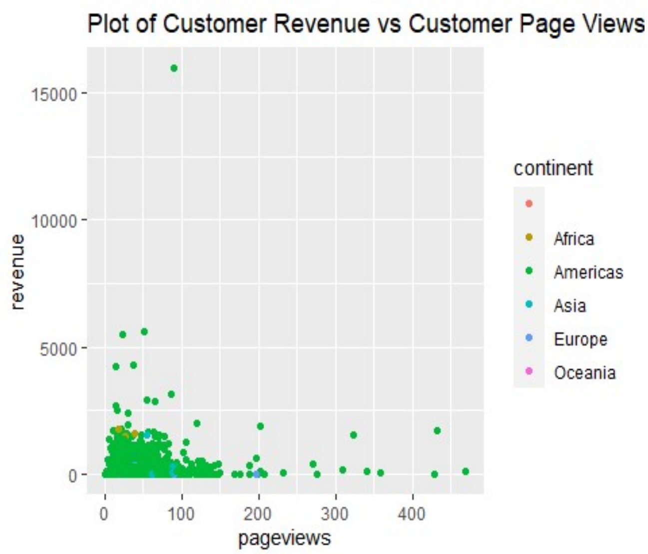

```
trainfull<-na.omit(dfTrain)
class(trainfull)
nrow(trainfull)
```

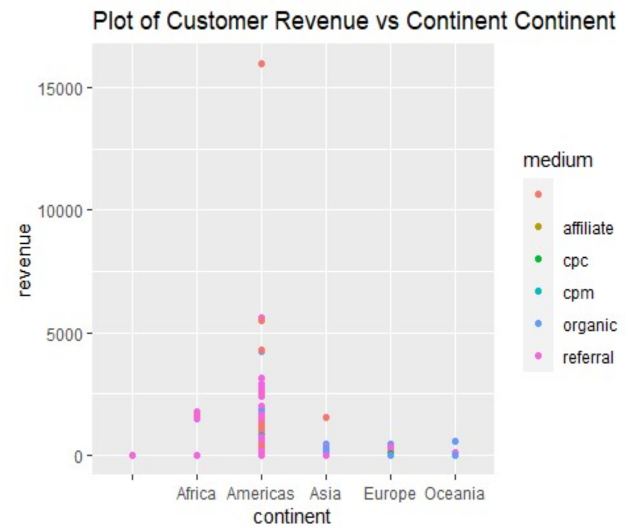
Command to check total number of records that have all data available.

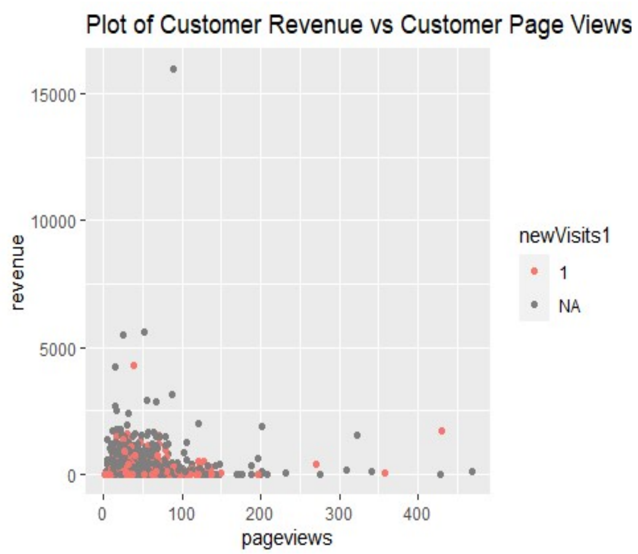
```
## [1] 264 (Number of records with complete data)
```

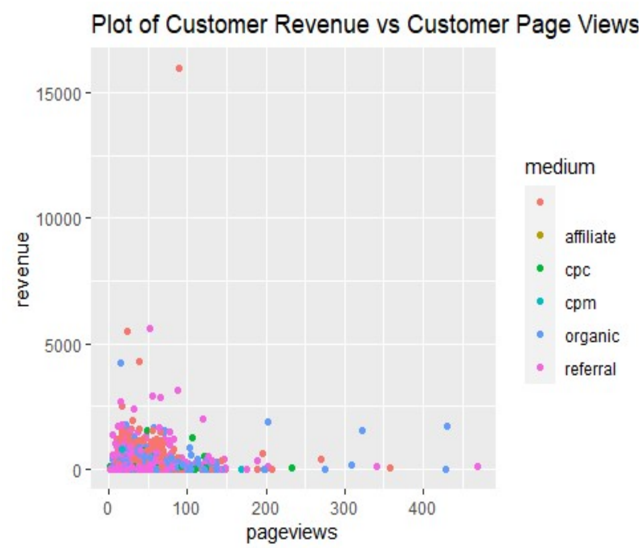
```
ggplot(data = dfTrain) +
  geom_point(mapping = aes (x = pageviews, y= revenue, color = continent))+
  ggtitle("Plot of Customer Revenue vs Customer Page Views")

ggplot(data = dfTrain) +
  geom_point(mapping = aes (x = continent, y= revenue, color = medium))+
  ggtitle("Plot of Customer Revenue vs Continent Continent")

dfTrain$newVisits1 <- as.factor(dfTrain$newVisits)
ggplot(data = dfTrain) +
  geom_point(mapping = aes (x = pageviews, y= revenue, color = newVisits1 ))+
  ggtitle("Plot of Customer Revenue vs Customer Page Views")

ggplot(data = dfTrain) +
  geom_point(mapping = aes (x = pageviews, y= revenue, color = medium))+
  ggtitle("Plot of Customer Revenue vs Customer Page Views")
```

Revenue vs pageviews and medium plotted. Some charts of revenues vs. pageviews are plotted with points colors of points clustered by continent, newVisits and medium. Majority of the other variables do not have significant x values to facilitate plotting.





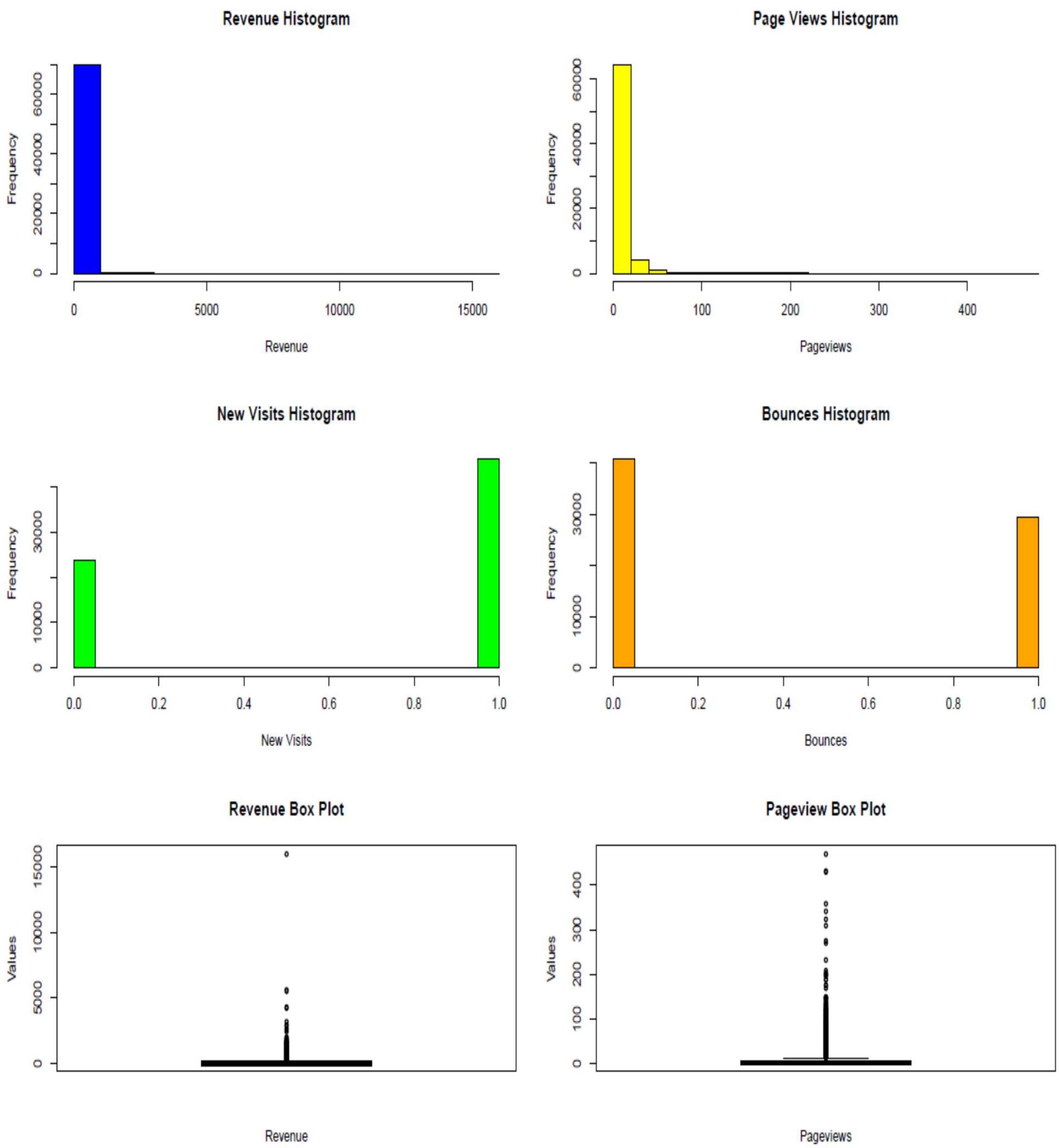





```
outlier(dfTrain$revenue)
## [1] 15980.79
grubbs.test(dfTrain$revenue)
Grubbs test for one outlier
data:  dfTrain$revenue
## G = 160.45368, U = 0.63257, p-value < 2.2e-16
## alternative hypothesis: highest value 15980.79 is an outlier
outlier(dfTrain$pageviews)
## [1] 469
grubbs.test(dfTrain$pageviews)
##  Grubbs test for one outlier
## data:  dfTrain$pageviews
## G = 39.57003, U = 0.97765, p-value < 2.2e-16
## alternative hypothesis: highest value 469 is an outlier
```

Outlier assessment on revenue and pageViews. Only one each are noted in this analysis. Action was postponed until after the initial modeling analysis where outliers on residuals were handled.

Ramkishore Rao, HW #s 4 and 5




```
meanrevenue <- mean(dfTrain$revenue, na.rm = TRUE)
sdrevenue<-3*sd(dfTrain$revenue, na.rm =TRUE)
count(filter(dfTrain, revenue >(meanrevenue+sdrevenue)))
##     n
## 1 502
pageviews <-na.omit(dfTrain$pageviews)
meanpageviews <- mean(pageviews)
sdpageviews<-3*sd(pageviews)
count(filter(dfTrain, pageviews >(meanpageviews+sdpageviews)))
##      n
## 1 1370
```

Summary of stats for revenue and pageViews.

Number of records above Mean +3 Standard Deviation noted.

Too many records identified and no action proposed.

2.0 Data Preparation (Part b)

Data preparation included missing value imputation, aggregation of data to customer level, further analysis of the aggregated data at the customer level to aid in feature extraction/selection, assessment of features relevant for subsequent regression modeling, and transformation of select variables.

```
dfTrain$newVisits[is.na(dfTrain$newVisits)] <-0
dfTrain$bounces[is.na(dfTrain$bounces)] <-0
#Imputing values for continent, Only Imputing if found for that customer
dfTrain$continent[dfTrain$custId == 61056] <- 'Asia'
dfTrain$continent[dfTrain$custId == 86024] <- 'Asia'

#Imputing values for subContinent, , Only Imputing if found for that customer
dfTrain$subContinent[dfTrain$custId == 61056] <- 'Eastern Asia'
dfTrain$subContinent[dfTrain$custId == 86024] <- 'Southeast Asia'

#Imputing values for country, , Only Imputing if found for that customer
dfTrain$country[dfTrain$custId == 61056] <- 'Japan'
dfTrain$country[dfTrain$custId == 86024] <- 'Indonesia'
```

Imputed NAs in newVisits and bounces from null to 0 as these variables carry 2 values: 1 or 0.

Imputed values for factor variables, continent, subcontinent, and country only if such data was available for a given customer.

```
ModelTrain %>% group_by(country) %>% summarize(n=n()) %>% arrange(desc(n))
ModelTrain <- mutate(ModelTrain, country1 = fct_lump(fct_explicit_na(country), n=4))
SummaryTable <-ModelTrain %>% mutate(country1 = fct_lump(fct_explicit_na(country), n=4)) %>%
 group_by(country1) %>%
 summarize(n = n(),
        meanRev = round(mean(revenue),2),
        stdevRev = round(sd(revenue),2),
        meanpageviews = round(mean(pageviews),2),
        sdpageviews = round(sd(pageviews),2),
        totbounces = sum(bounces),
        totnewvisits = sum(newVisits))%>%
 arrange(desc(n))
```

Summarize by 4 main countries and place the rest of the data in other category for countries to facilitate subsequent feature extraction/regression.

Table 1: SUMMARY BY COUNTRY

country1	n	meanRev	stdevRev	meanpageviews	sdpageviews	totbounces	totnewvisits
United States	36941	18.02	132.66	9.21	13.40	9808	18334
Other	25838	1.04	28.53	2.88	8.66	15693	21943
India	3044	0.14	4.88	2.70	4.21	1781	2686
United Kingdom	2330	0.24	5.59	2.66	4.70	1414	1954
Canada	1918	9.75	96.79	6.63	10.64	656	1210

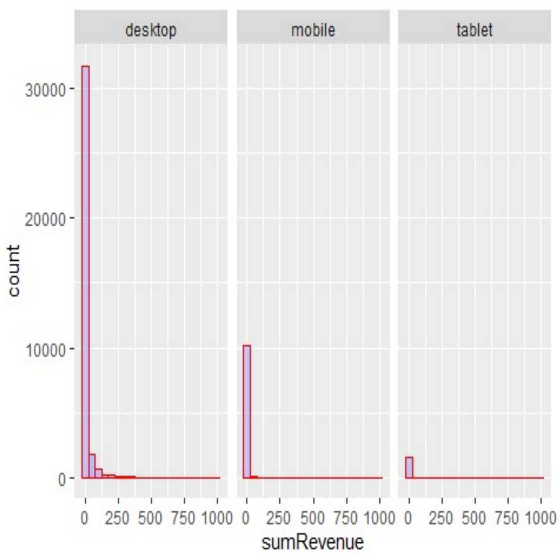
```
abc1<-ModelTrain %>% group_by(custId) %>%
 summarise(sumRevenue = sum(revenue), sumviews = sum(pageviews), medium = last(medium),
        device = last(deviceCategory), isTrueDirect = last(isTrueDirect),
        isMobile = last(isMobile), op = last(operatingSystem), bounces = last(bounces),
        newvisit= last(newVisits), country = last(country1))
Rows: 47,249
Columns: 11
$ custId       <int> 1795, 1797, 1799, 1800, 1801, 1803, 1804, 1807, 1810, 1812, 1813, 1815,
$ sumRevenue   <dbl> 0.00000, 0.00000, 0.00000, 0.00000, 0.00000, 0.00000, 0.00000, 0.00000,
$ sumviews     <dbl> 1, 1, 1, 1, 6, 6, 1, 1, 2, 5, 1, 1, 8, 24, 14, 2, 3, 1, 1, 1, 1, 1, 14,
$ medium       <fct> referral, referral, organic, referral, , organic, organic, organic,
```

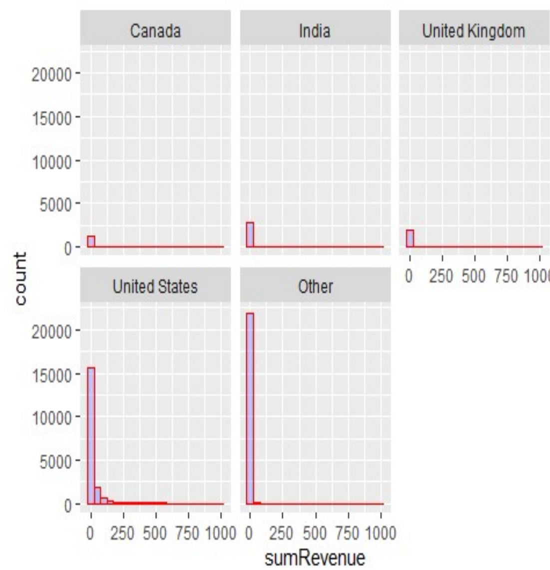
Aggregated Train Data at Customer Level, Key Numeric and Categorical Vaiables

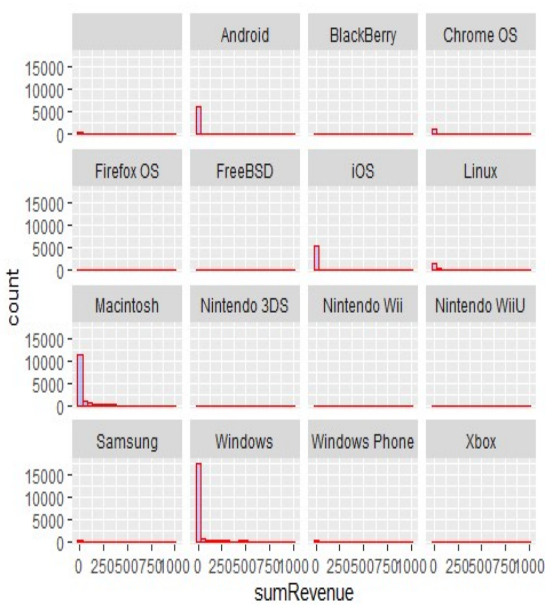
```
$ device       <fct> desktop, desktop, desktop, desktop, mobile, desktop, mobile, desktop,
$ isTrueDirect <int> 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0,
$ isMobile     <int> 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0,
$ op           <fct> Windows, Macintosh, Windows, Macintosh, Android, Macintosh, iOS, Windows,
$ bounces      <dbl> 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0,
$ newvisit     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1,
$ country      <fct> India, United States, India, Other, United States, Other, India, United
```

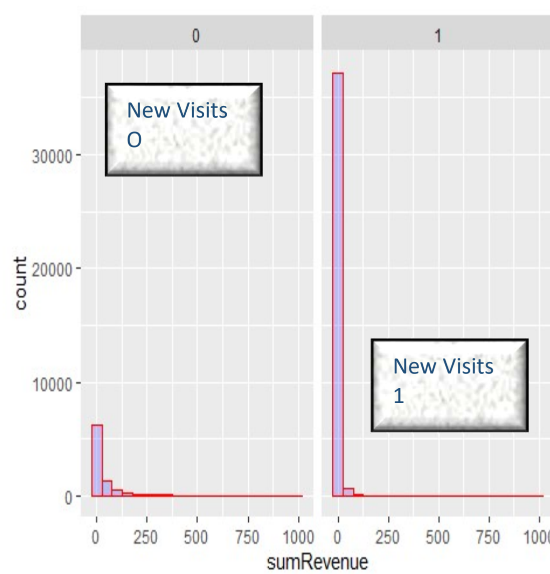
To facilitate review, histograms plotted for a subset of Aggregated Dataset with revenue values less than 1,000.

```
TrainforHist <-filter(abc1, sumRevenue <1000)
ggplot(data = TrainforHist, aes(sumRevenue)) +
  geom_histogram(col= "red", fill="blue", alpha =0.2, binwidth=50 )+facet_wrap(~device)
```



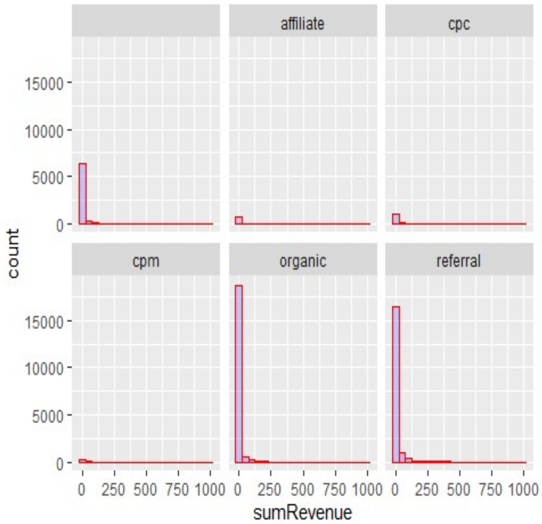


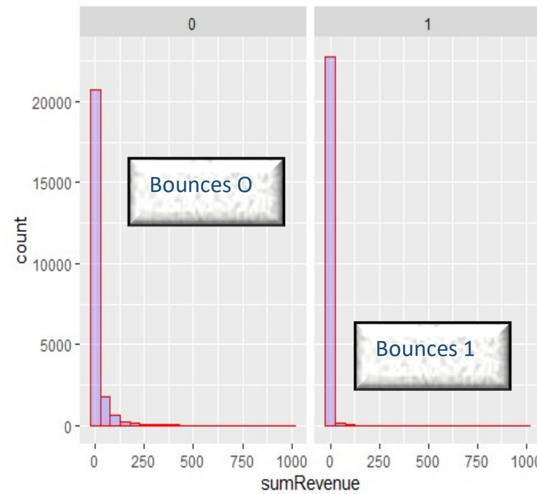




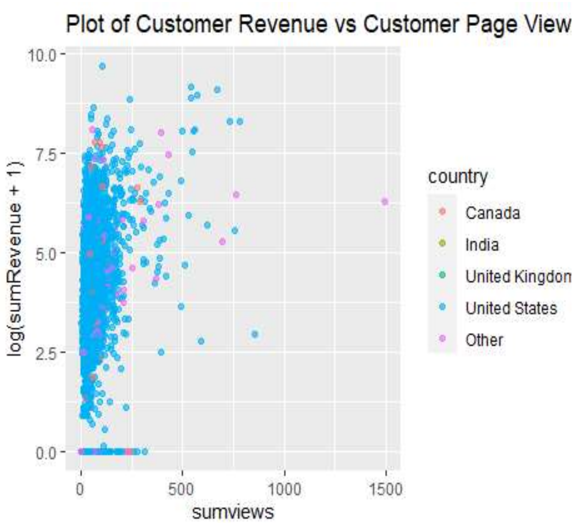
Based on plotted histograms, number of transactions in which sumRevenues are plotted are higher when customers use a desktop, are in United states, make a new visit, and use Windows or Mcintosh. Histograms for a dataset that only includes revenue greater than zero may result in providing data that could be used to assess whether revenues that are generated are also higher.

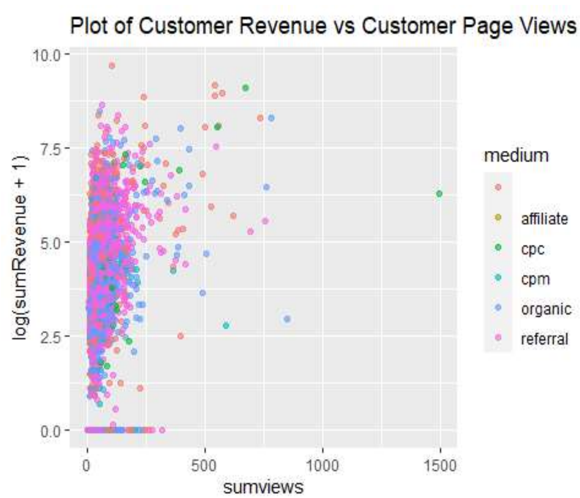
Ramkishore Rao, HW #s 4 and 5

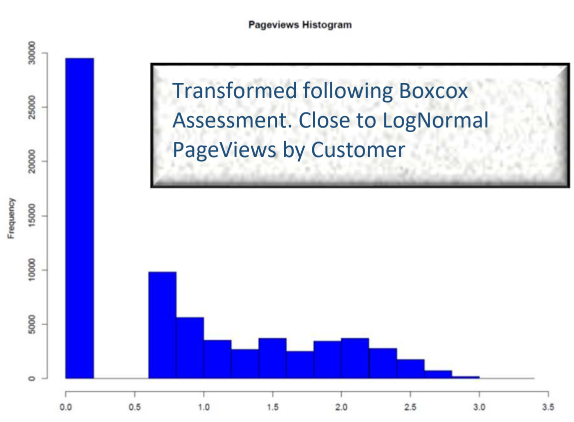


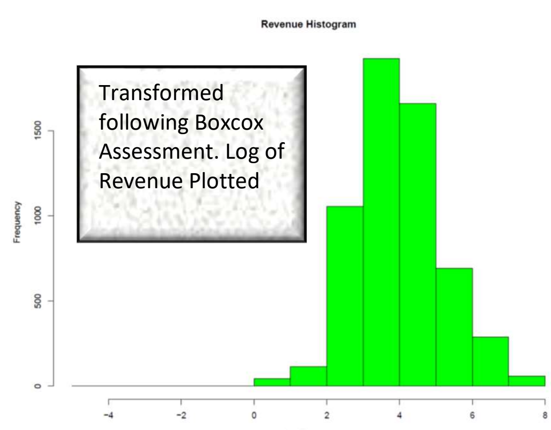



```
ggplot(data = abc1) +
  geom_point(mapping = aes (x = sumviews, y= log(sumRevenue+1), color = country),
alpha = 0.6)+
  ggtitle("Plot of Customer Revenue vs Customer Page Views")
```



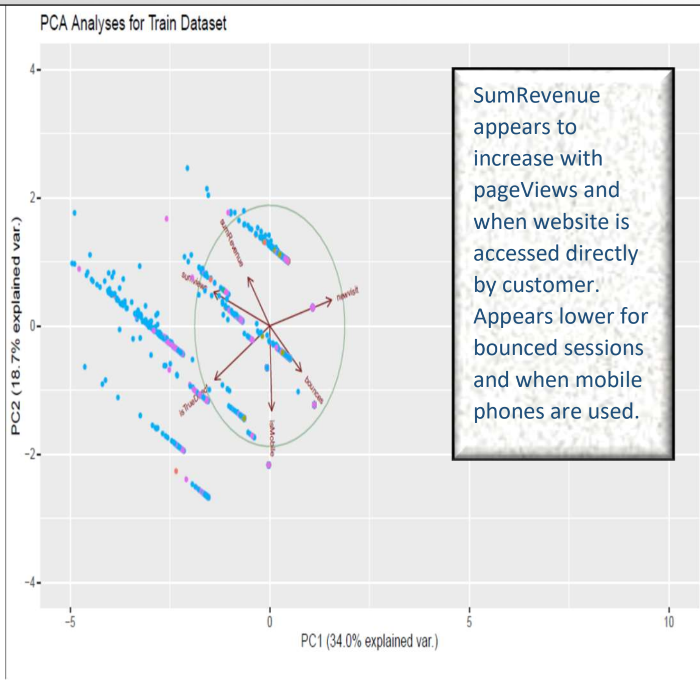


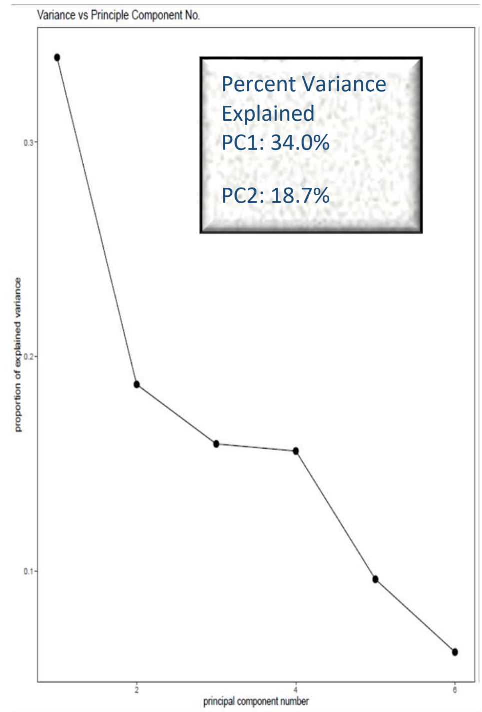


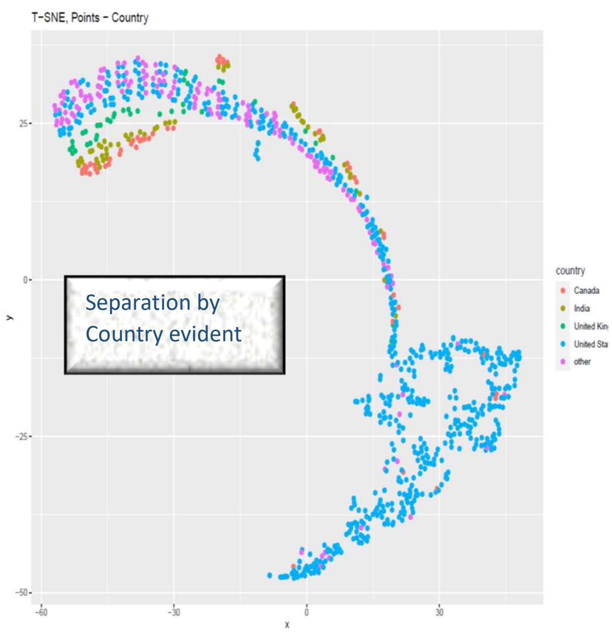


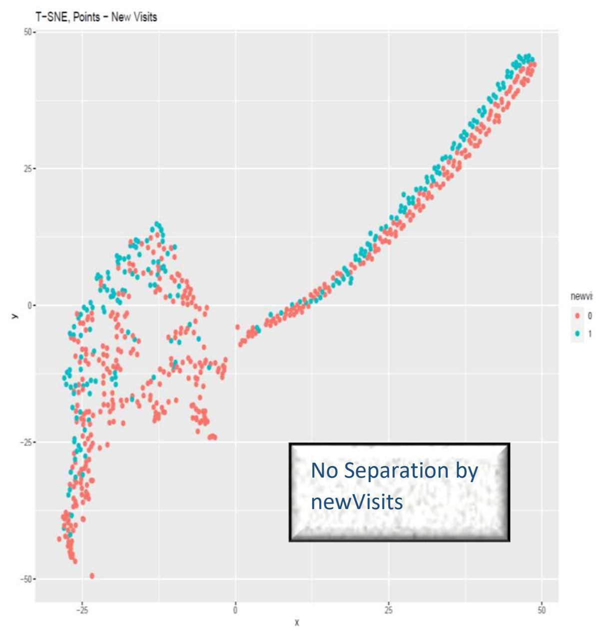
Ramkishore Rao, HW #s 4 and 5

To reduce computing resources, a random sample of 2000 and 4000 were collected from aggregated train dataset for PCA and T-SNE, respectively









Feature extraction by PCA indicates that sumRevenue appears to increase with total pageViews by customer and with isTrueDirect, perhaps, implying that sumRevenue is higher when website was access directly by the customer. SumRevenue appears to decrease for bounced sessions and also when mobile phones are used to access the website. T-SNE indicates some level of class separation for predicted outcomes based on country of origin, which validates the higher means for United States and Canada as indicated in Table 1 of this report.

3.0 Modeling

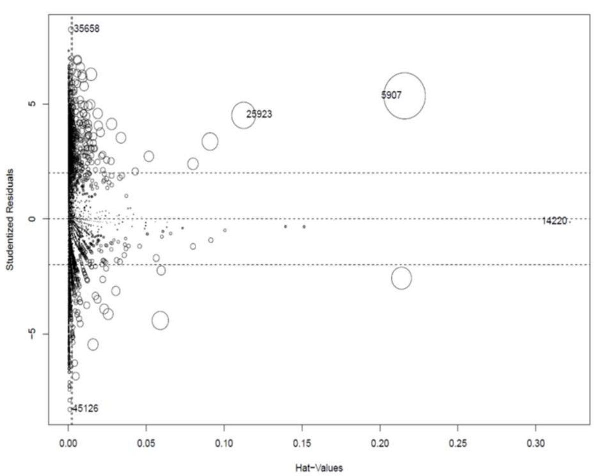
Following initial data exploration and preparation, the aggregated Train Dataset (abc1 dataframe) was used for modeling. An initial linear regression model was developed on this data set using several predictor variables. Sumviews was transformed log normally and so was sumRevenues. Country and device were coded as dummy variables. Although medium was not included in the linear regression modeling, it was considered in the final MARS model and its variant that was used for predicting customer revenues.

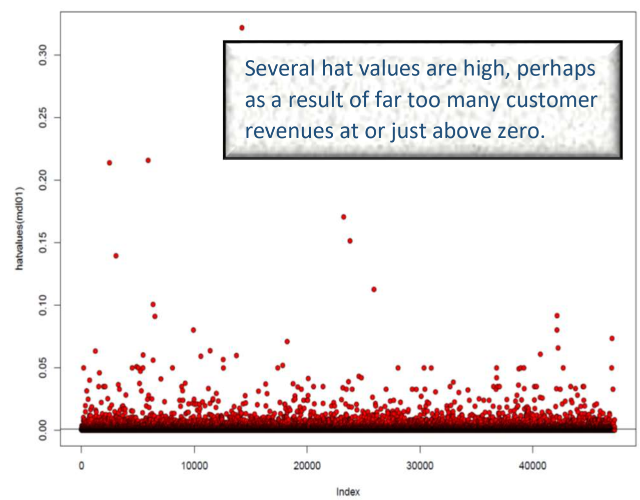
```
mdl01<-lm(data=abc1, log(sumRevenue+1)
~log(sumviews+1)+bounces*device*log(sumviews+1)+newvisit*log(sumviews+1)*device+country*log(sumviews+1
)+device*country*log(sumviews+1))
```

3.1 Continued Iterative Data Preparation (Part c)

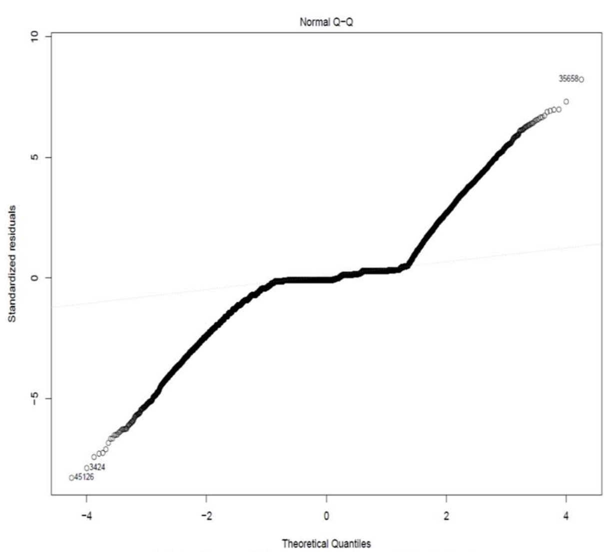
Extreme values were removed based on outlier test on model md101 (see abc11, used for subsequent modeling).

Diagnostic assessment on the results of the linear regression model presented above were conducted.





Several hat values are high, perhaps as a result of far too many customer revenues at or just above zero.




```
outlierTest(mdl01)

      rstudent unadjusted p-value Bonferroni p
45126 -8.290567   1.1565e-16   5.4642e-12
35658  8.237653   1.8008e-16   8.5087e-12
3424  -7.886015   3.1865e-15   1.5056e-10
5942  -7.427465   1.1256e-13   5.3184e-09
11599  7.316275   2.5899e-13   1.2237e-08
12841 -7.286448   3.2320e-13   1.5271e-08
29768 -7.254890   4.0815e-13   1.9285e-08
17312 -7.100641   1.2593e-12   5.9500e-08
30541  6.991649   2.7527e-12   1.3006e-07
26801  6.980835   2.9729e-12   1.4047e-07

abc11<-abc1[-c(45126, 35658,
3424,5942,12841,29768,17312,11599,30541,26801),]

summary(abc11$sumRevenue)
Min. 1st Qu. Median Mean  3rd Qu.  Max.
0.00  0.00   0.00   14.91  0.00 15980.79

summary(abc11$sumviews)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

1.00    1.00    2.00    9.32   6.00 1496.00
```

Predictor Variables Based on Feature Extraction/Analysis Performed in 2.0. Early rounds included additional variables in the linear regression; reduced to noted selection based on observed trial error.

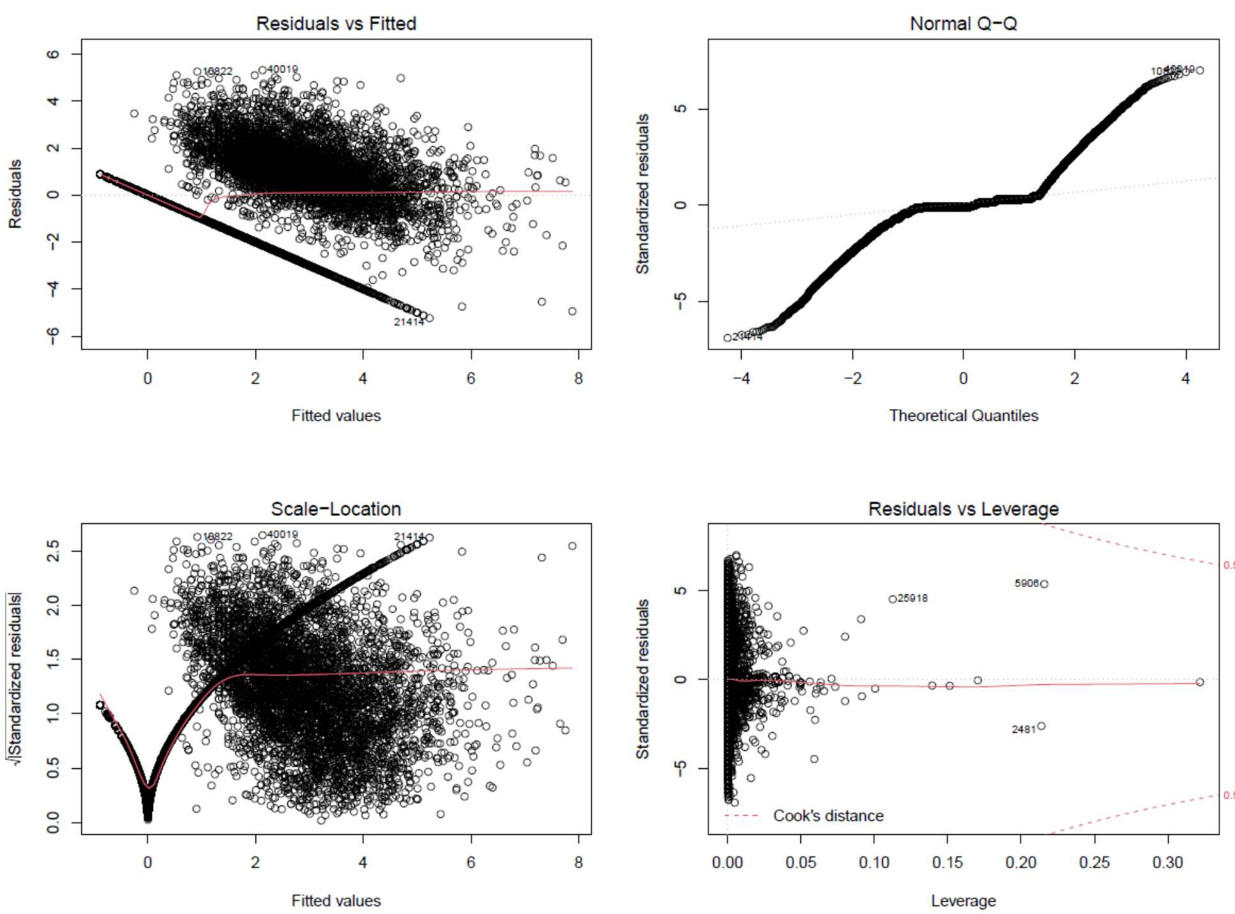
3.2 Modeling Part d, (i)

Following the final data cleaning performed on the train dataset as outlined in Section 3.1, regression modeling was performed using several types of models, which included linear regression modeling, and penalized regression modeling approaches such as lasso, ridge, and elastic net. Finally, multivariate adaptive regression spine modeling was also performed on the dataset. Observations gained from these models and the results are summarized below.

<u>Linear Regression Modeling</u>

```
mdl02<-lm(data=abc11, log(sumRevenue+1)
~log(sumviews+1)+bounces*device*log(sumviews+1)+newvisit*log(sumviews+1)*device+country*log(s
umviews+1)+device*country*log(sumviews+1))
summary(mdl02)
```

```
Residual standard error: 0.7613 on 47197 degrees of freedom
Multiple R-squared:  0.6616, Adjusted R-squared:  0.6613
F-statistic:  2250 on 41 and 47197 DF,  p-value: < 2.2e-16; Coefficients and Intercept in the
attached R script
```



Cooks distance (See Leverage vs Std Residuals plot) not significantly high for any observations. In general, apart from a few values, residuals do not appear to follow a pattern and appear uncorrelated.

Penalized Regression Modeling

Penalized regression modeling using the Caret package was conducted. Modeling was conducted using ridge, lasso, and elastic net approaches. 5-fold cross validation technique was used for this modeling. The objective function was to minimize the RME for each of the models. Hperparameter chosen for ridge regression was lambda, for lasso regression was fraction, for elastic net regression was fraction. For the GLMNET method for elasticnet modeling, alpha and lamda were chosen as the hyperparameters for model tuning.

Ridge Regression

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}^{p}\hat{\beta}_j^2$$

OBJECTIVE FUNCTION:

Hyperparameter Lambda, Alpha = 0; GLMNET METHOD:

```
fitControl <- trainControl(method="cv",number=5)
ridgefit1 <- train( log(sumRevenue+1) ~log(sumviews+1)+bounces*device*log(sumviews+1)+
newvisit*log(sumviews+1)*device+country*log(sumviews+1)+device*country*log(sumviews+1), data = abc11,
method = "glmnet", trControl=fitControl, tuneGrid = expand.grid(alpha = 0, lambda = seq(0,.8,length=20))
lambda      RMSE      Rsquared  MAE
0.08421053  0.7801767  0.6477863  0.3929564
```

Lasso Regression

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}^{p}|\hat{\beta}_j|$$

OBJECTIVE FUNCTION:

Hyperparameter Fraction

```
lassoGrid <- expand.grid(fraction=seq(0.7,1.0,length=16))
lassofit <- train(log(sumRevenue+1) ~log(sumviews+1)+bounces*device*log(sumviews+1)+
newvisit*log(sumviews+1)*device+country*log(sumviews+1)+device*country*log(sumviews+1),
data=abc11, method="lasso", trControl=fitControl, tuneGrid=lassoGrid)
fraction  RMSE      Rsquared  MAE
0.98      0.7620675  0.6605471  0.3992677
```

Elastic Net Regression

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda_1\sum_{j=1}^{p}|\hat{\beta}_j|+\lambda_2\sum_{j=1}^{p}\hat{\beta}_j^2$$

OBJECTIVE FUNCTION:

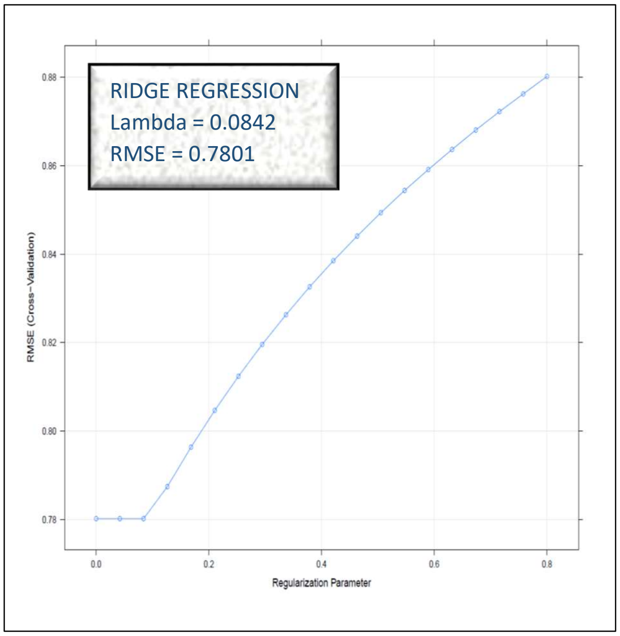
GLMNET MODEL: lamba1 = (alpha) x lambda and lambda2 = (1-alpha) x lambda

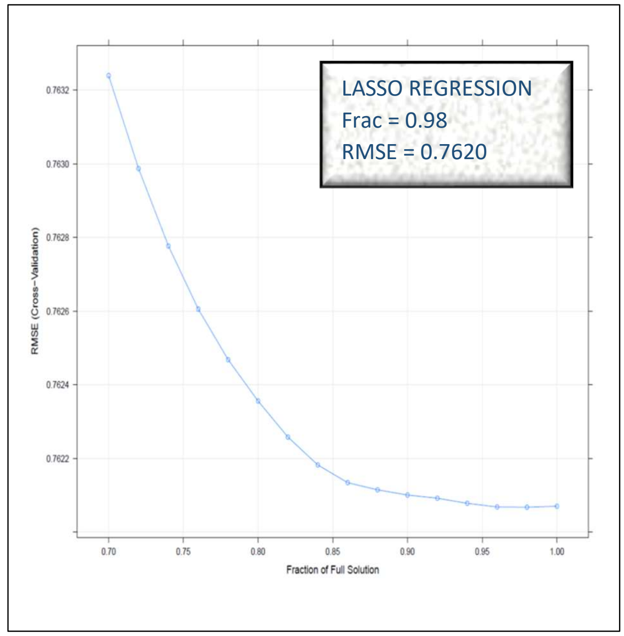
Caret Package

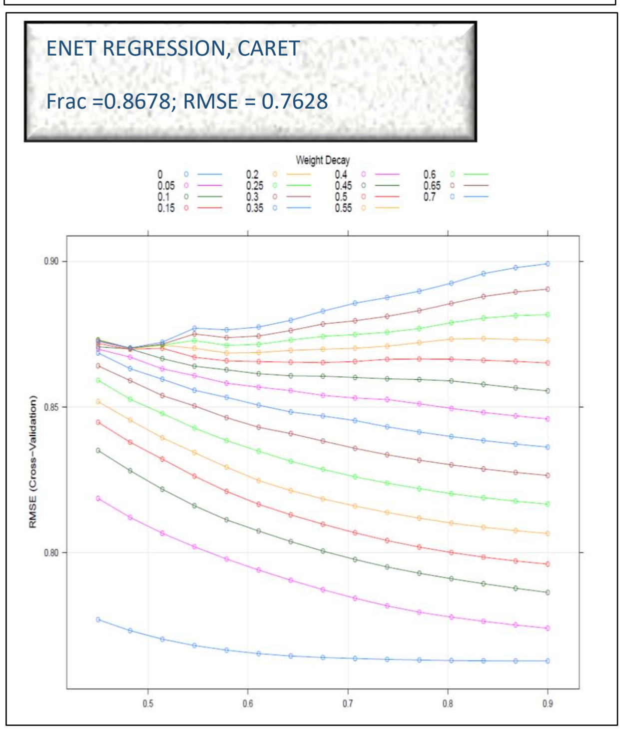
```
enetGrid <- expand.grid(lambda=seq(0,.7,length=15), fraction=seq(0.45,.9,length=15))
fitenet <- train(log(sumRevenue+1) ~log(sumviews+1)+bounces*device*log(sumviews+1)+
newvisit*log(sumviews+1)*device+country*log(sumviews+1)+device*country*log(sumviews+1),
data=abc11, method="enet", trControl=fitControl, tuneGrid=enetGrid)
fraction  RMSE      Rsquared  MAE
0.8678    0.7628827  0.6600626  0.3990483
```

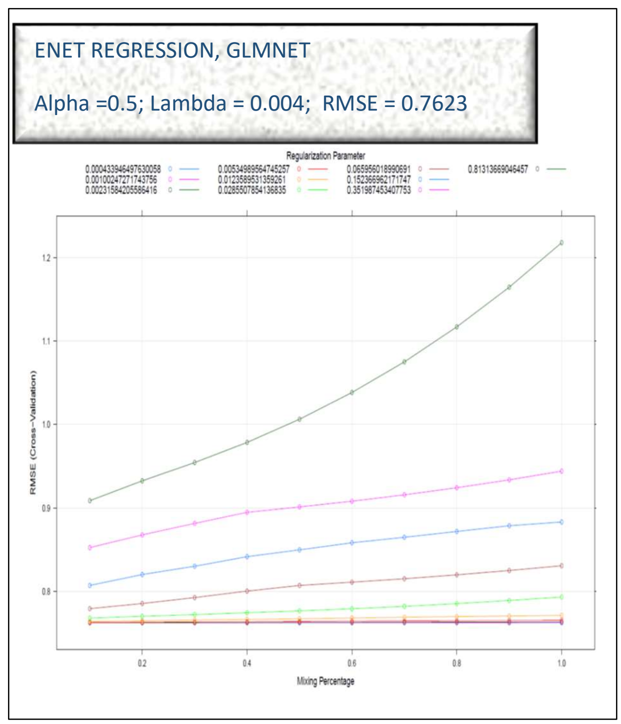
Caret Package – GLMNET Method

```
fitenet1 <- train(log(sumRevenue+1) ~log(sumviews+1)+bounces*device*log(sumviews+1)+
newvisit*log(sumviews+1)*device+country*log(sumviews+1)+device*country*log(sumviews+1),
data=abc11, method="glmnet", trControl=fitControl, tuneLength=10)

alpha  lambda        RMSE      Rsquared  MAE
0.5    0.0004339465  0.7623360  0.6601843  0.3983044
```









Mars Regression Modeling

Mars regression modeling was performed using the variables noted below. A degree of freedom of 3 was imposed on the model. Upon completion of the modeling, the earth function of the MARS model identified relative importance of predictor variables and the interactions that have the most impotance and used those to develop the regression model (please see Equation 1 below).

```
marsFit1<- earth(log(sumRevenue+1) ~log(sumviews+1)+medium +device+ isTrueDirect+ isMobile+
op+bounces+newvisit+country, data=abc11, degree=3,nk=50,pmethod="cv",nfold=5,ncross=5)
```

summary(marsFit1)

```
                                                       coefficients
(Intercept)                                              -0.0019053
h(log(sumviews + 1)-2.19722)                              0.8781058
h(log(sumviews + 1)-3.63759)                              1.8390540
h(log(sumviews + 1)-2.19722) * mediumorganic             -0.1777946
h(log(sumviews + 1)-2.19722) * isMobile                  -0.6805525
h(log(sumviews + 1)-2.19722) * opWindows                 -0.2274912
h(log(sumviews + 1)-2.19722) * countryUnited States       1.6644160
h(log(sumviews + 1)-3.63759) * countryUnited States      -3.2991466

Selected 8 of 9 terms, and 5 of 31 predictors (pmethod="cv")
Termination condition: RSq changed by less than 0.001 at 9 terms
Importance: log(sumviews + 1), countryUnited States, isMobile, opWindows, mediumorganic,
mediumaffiliate-unused, mediumcpc-unused, ...
Number of terms at each degree of interaction: 1 2 5
GRSq 0.7085151  RSq 0.708731  mean.oof.RSq 0.7087974 (sd 0.0107)

pmethod="backward" would have selected the same model:
    8 terms 5 preds,  GRSq 0.7085151  RSq 0.708731  mean.of.RSq 0.7087974
```

EQUATION 1

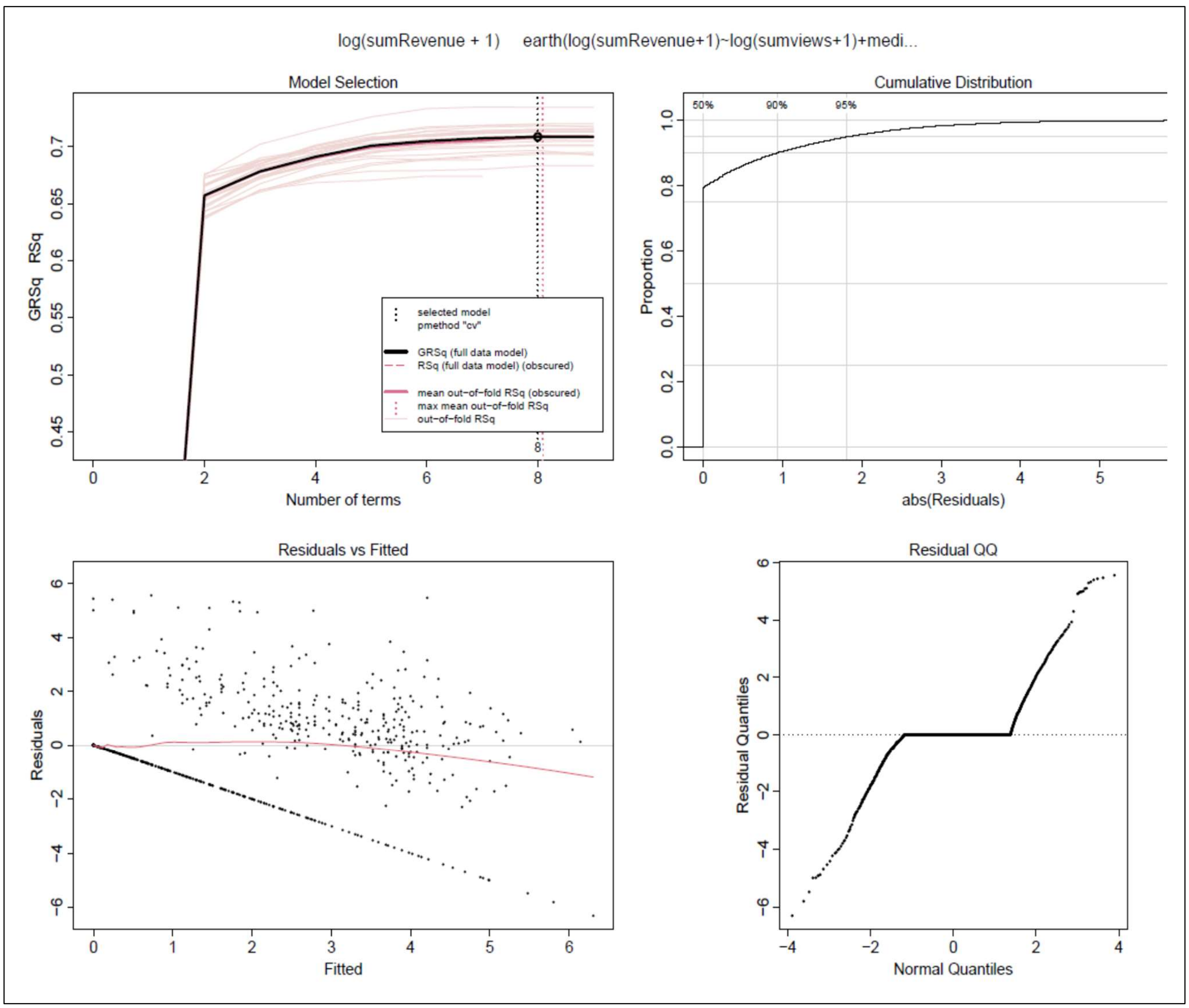


TABLE 2: SUMMARY OF MODEL RESULTS

Model	Method	Package	Hyperparameter	Selection	CV Peformance	
					RMSE	R-Squared
OLS	lm	Stats	NA	NA	0.7613	0.6616
Lasso	lasso	Caret	Fraction	0.98	0.7620	0.6605
Ridge	ridge	Caret - Glmnet	Lambda	0.0842	0.7801	0.6477
Elastic Net 1	enet	Caret	Fraction/Lambda	0.90	0.7628	0.6600
Elastic Net-2	enet	Caret – Glmnet	Aplpha and Lamda	0.5/0.004	0.7623	0.6601
MARS	earth	Earth	Degree	3	**0.7059**	**0.7087**

3.3 Modeling Approach/Selected Model, Part d, ii

For the modeling, we implemented the following steps:

1) Initial Data Exploration – select variables were converted from character to factor variables to ensure inclusion in the regression modeling.
2) Data Preparation – missing values for pageViews, bounces, newVisits were imputed as outlined in Section 2.0. Also, a feature extraction/engineering was performed to identify the predictor variables that could potentially have most impact on the customer revenue.
3) Modeling was performed iteratively by first including several target variables in each of the selected models and then pared down to the key variables. The selected variables for the final step for each of these models have been presented for each type of regression. It is worth noting that, the MARS model includes several predictor variables on the command line, but the model identifies key parameters for inclusion. The results of the modeling in Section 3.2 identify sumviews, medium, isMobile, Windows operating system, and country United States as key predictor variables, which were preliminarily determined as important variables during the feature extraction phase.
4) Further, to simplify the modeling the log transformation of sumviews was removed in the final selected model (See below). The key predictor variables where were sumviews, operating system Mcintosh, medium referral, is Mobile, and country United States, which were once again preliminarily determined as important variables during the feature extraction phase
5) Final Mars Model was selected because it is less complex than the previous model version and had lower test error when uploaded to the Kaggle web site.

3.4 Final Selected Model, Part d, iii

Final MARS model that was selected was based on the terms noted in marsFit2 below. The log transformation of sumviews was removed, and MARS was allowed to provide the optimum parameters, including the best interaction between variables that leads to the optimum GRSq (see Equation 2). Further, RMSEs for the initial MARS model (marsFit1) and the final selected model(marsFit2) were also computed. Although marsFit 1 resulted in lower train error than marsFit 2, when uploaded to the Kaggle site, marsFit 2 resulted in a marginally lower test error, and hence marsFit2 was chosen for prediction of customer revenues.

```
res1<- marsFit1$residuals^2
avgres1<-mean(res1)
```

RMSEMarsFit1<-sqrt(avgres1)

[1] 0.705926 (RMSE for MARS Model 1 presented in Section 3.2)

marsFit2<- earth(log(sumRevenue+1) ~sumviews+medium +device+ isTrueDirect+isMobile+ op+bounces+newvisit+country, data=abc11,degree=2,nk=50,pmethod="cv",nfold=5,ncross=5)

```
                                       coefficients
(Intercept)                             -0.52161349
mediumreferral                           0.47343444
countryUnited States                    -0.73077420
h(sumviews-6)                            0.03682355
h(sumviews-28)                           0.06630938
h(47-sumviews)                           0.01162619
h(sumviews-47)                          -0.08871064
h(sumviews-267)                         -0.01055617
opMacintosh * countryUnited States       0.13156912
h(27-sumviews) * mediumreferral         -0.01910653
h(sumviews-27) * mediumreferral         -0.00934538
h(sumviews-43) * mediumreferral          0.00801494
h(sumviews-6) * isMobile                -0.02706115
h(sumviews-47) * isMobile                0.02558087
h(sumviews-8) * countryUnited States     0.12311513
h(25-sumviews) * countryUnited States   -0.04940710
h(sumviews-25) * countryUnited States   -0.12608015
h(47-sumviews) * countryUnited States    0.04075790
```

EQUATION 2

```
Selected 18 of 19 terms, and 5 of 31 predictors (pmethod="cv")
Termination condition: RSq changed by less than 0.001 at 19 terms
Importance: sumviews, countryUnited States, isMobile, mediumreferral, opMacintosh,
mediumaffiliate-unused, mediumcpc-unused, ...
Number of terms at each degree of interaction: 1 7 10
GRSq 0.7081476  RSq 0.7086726  mean.oof.RSq 0.705561 (sd 0.0109)

pmethod="backward" would have selected the same model:
    18 terms 5 preds,  GRSq 0.7081476  RSq 0.7086726  mean.oof.RSq 0.705561
```

res2<- marsFit2$residuals^2

avgres2<-mean(res2)

RMSEMarsFit2<-sqrt(avgres2)

[1] 0.7060094 (RMSE for Final MARS Model presented in this section)

Prediction, Part d(iii), posted to Kaggle web site