Regression Analysis

Howard Nguyen

——— R ———

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
# Load data
df <- read.csv(file="WA_Fn-UseC_-Marketing-Customer-Value-Analysis.csv", header=TRUE, sep=",")
head(df)
```

```
##   Customer      State Customer.Lifetime.Value Response Coverage Education
## 1  BU79786 Washington                2763.519       No    Basic  Bachelor
## 2  QZ44356    Arizona                6979.536       No Extended  Bachelor
## 3  AI49188     Nevada               12887.432       No  Premium  Bachelor
## 4  WW63253 California                7645.862       No    Basic  Bachelor
## 5  HB64268 Washington                2813.693       No    Basic  Bachelor
## 6  OC83172     Oregon                8256.298      Yes    Basic  Bachelor
##   Effective.To.Date EmploymentStatus Gender Income Location.Code Marital.Status
## 1           2/24/11         Employed      F  56274      Suburban        Married
## 2           1/31/11       Unemployed      F      0      Suburban         Single
## 3           2/19/11         Employed      F  48767      Suburban        Married
## 4           1/20/11       Unemployed      M      0      Suburban        Married
## 5            2/3/11         Employed      M  43836         Rural         Single
## 6           1/25/11         Employed      F  62902         Rural        Married
##   Monthly.Premium.Auto Months.Since.Last.Claim Months.Since.Policy.Inception
## 1                   69                      32                             5
## 2                   94                      13                            42
## 3                  108                      18                            38
## 4                  106                    18                            65
## 5                   73                    12                            44
## 6                   69                    14                            94
##   Number.of.Open.Complaints Number.of.Policies    Policy.Type       Policy
## 1                         0                  1 Corporate Auto Corporate L3
## 2                         0                  8  Personal Auto  Personal L3
## 3                         0                  2  Personal Auto  Personal L3
## 4                         0                  7 Corporate Auto Corporate L2
## 5                         0                  1  Personal Auto  Personal L1
## 6                         0                  2  Personal Auto  Personal L3
##   Renew.Offer.Type Sales.Channel Total.Claim.Amount Vehicle.Class Vehicle.Size
## 1           Offer1         Agent           384.8111  Two-Door Car      Medsize
## 2           Offer3         Agent          1131.4649 Four-Door Car      Medsize
## 3           Offer1         Agent           566.4722  Two-Door Car      Medsize
## 4           Offer1   Call Center           529.8813           SUV      Medsize
## 5           Offer1         Agent           138.1309 Four-Door Car      Medsize
## 6           Offer2           Web           159.3830  Two-Door Car      Medsize
```

```
dim(df)
```

```
## [1] 9134   24
```

```
# Encode Response as 0s and 1s
df$Response <- ifelse(df$Response=="Yes",1,0)
df$Engaged <- as.integer(df$Response)
```

```
engagementRate <- df %>%
  group_by(Engaged) %>%
  summarise(Count=n()) %>%
  mutate(Percentage=Count/nrow(df)*100.0)

engagementRate
```

1. Engagement Rate

```
## # A tibble: 2 x 3
##   Engaged Count Percentage
##     <int> <int>      <dbl>
## 1       0  7826       85.7
## 2       1  1308       14.3
```

```
# Transpose
transposed <- t(engagementRate)

colnames(transposed) <- engagementRate$Engaged
transposed <- transposed[-1,]
transposed
```

```
##                      0          1
## Count      7826.00000 1308.00000
## Percentage   85.67988   14.32012
```

```
renewalOfferType <- df %>%
  group_by(Engaged, Type=Renew.Offer.Type) %>%
  summarise(Count=n())
```

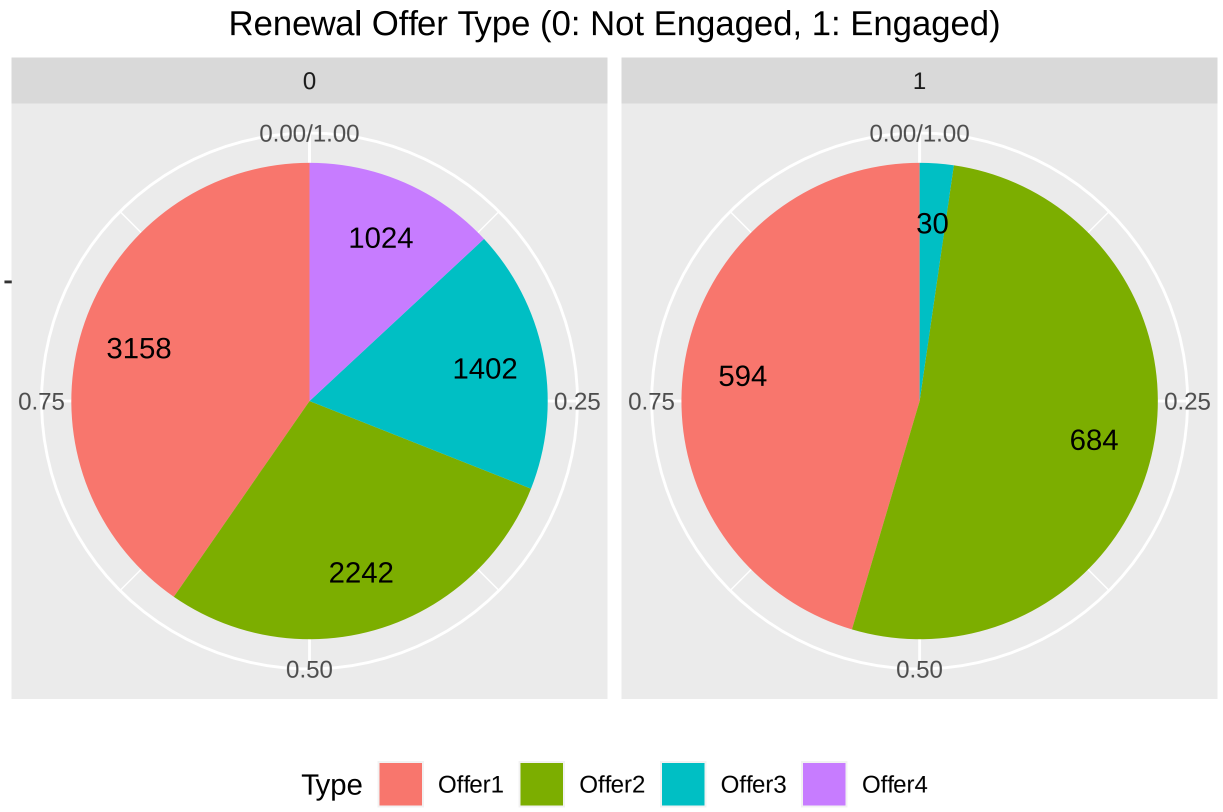
2. Renewal Offer Type

```
## `summarise()` has grouped output by 'Engaged'. You can override using the
## `.groups` argument.
```

```
renewalOfferType
```

```
## # A tibble: 7 x 3
## # Groups:   Engaged [2]
##   Engaged Type   Count
##     <int> <chr>  <int>
## 1       0 Offer1  3158
## 2       0 Offer2  2242
## 3       0 Offer3  1402
## 4       0 Offer4  1024
## 5       1 Offer1   594
## 6       1 Offer2   684
## 7       1 Offer3    30
```

```
# pie chart
ggplot(renewalOfferType, aes(x="", y=Count, fill=Type)) +
  geom_bar(width=1, stat = "identity", position=position_fill()) +
  geom_text(aes(x=1.25, label=Count), position=position_fill(vjust = 0.5)) +
  coord_polar("y") +
  facet_wrap(~Engaged) +
  ggtitle('Renewal Offer Type (0: Not Engaged, 1: Engaged)') +
  theme(
    axis.title.x=element_blank(),
    axis.title.y=element_blank(),
    plot.title=element_text(hjust=0.5),
    legend.position='bottom'
  )
```



```
salesChannel <- df %>%
  group_by(Engaged, Channel=Sales.Channel) %>%
  summarise(Count=n())
```

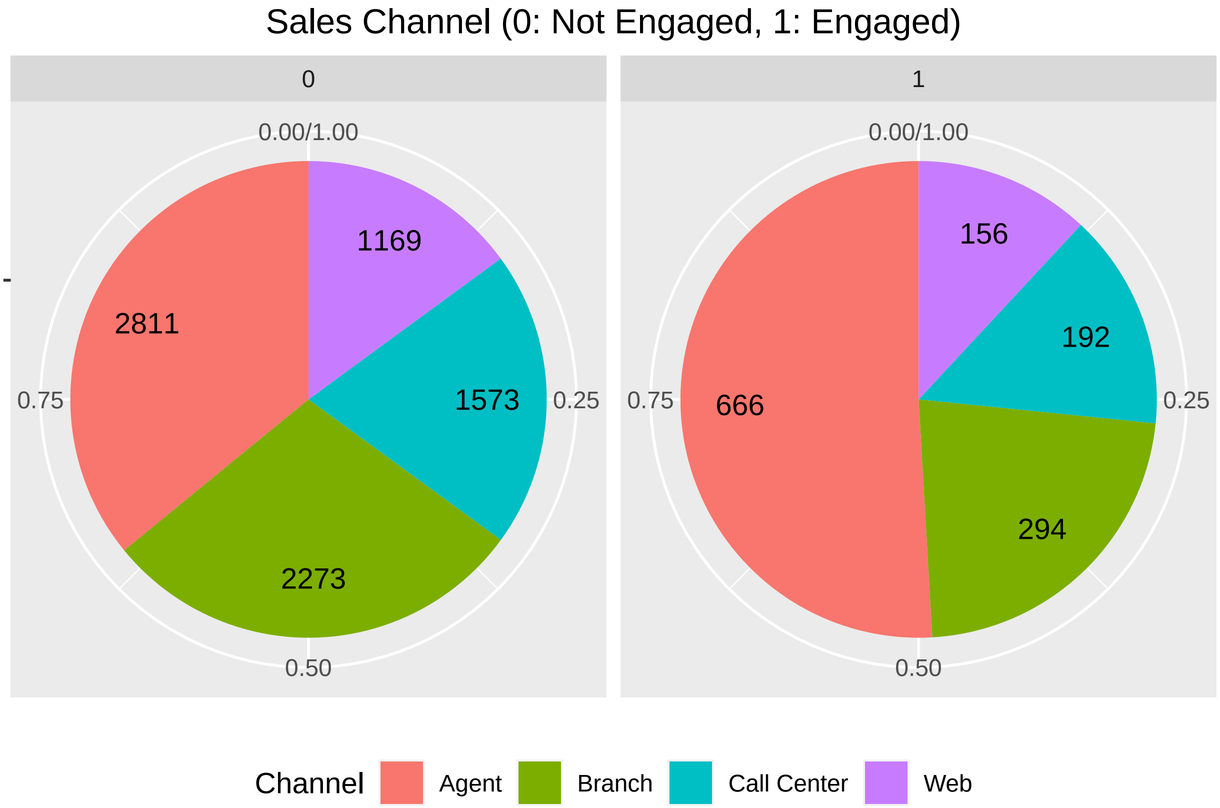
3. Sales Channel

```
## `summarise()` has grouped output by 'Engaged'. You can override using the
## `.groups` argument.
```

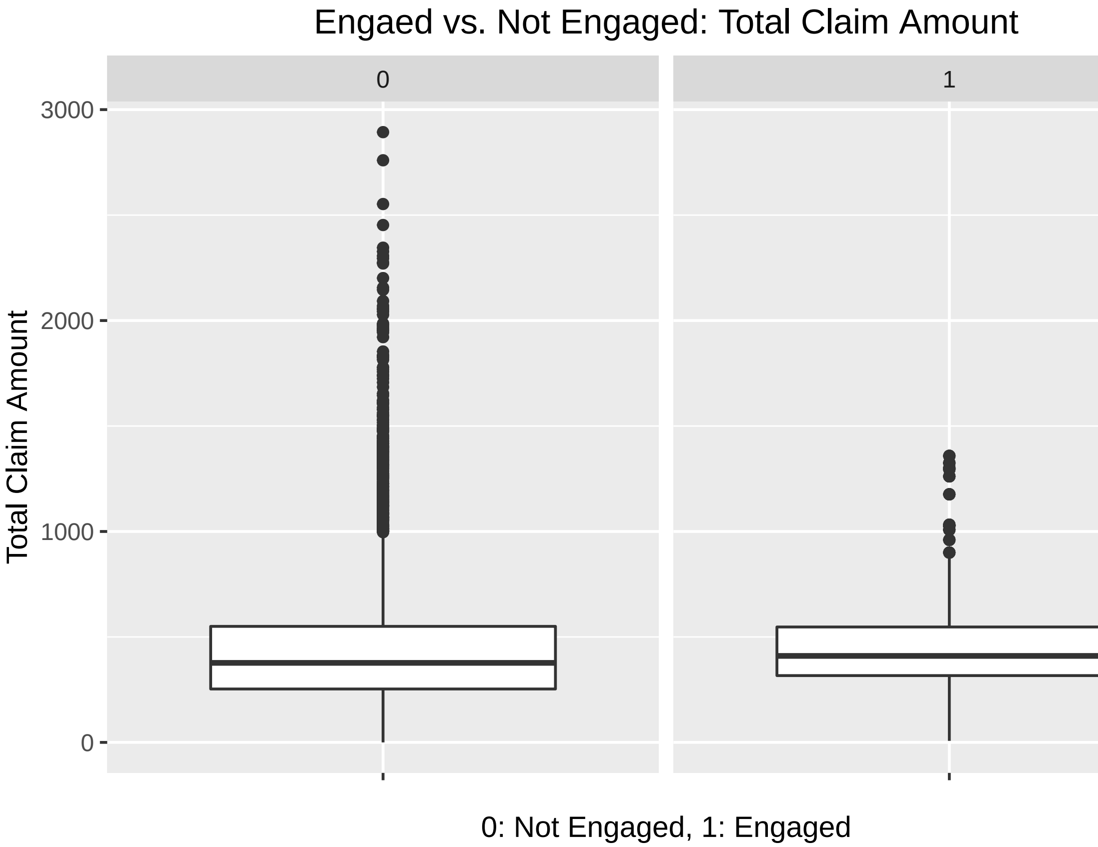
```
salesChannel
```

```
## # A tibble: 8 x 3
## # Groups:   Engaged [2]
##   Engaged Channel     Count
##     <int> <chr>       <int>
## 1       0 Agent        2811
## 2       0 Branch       2273
## 3       0 Call Center  1573
## 4       0 Web          1169
## 5       1 Agent         666
## 6       1 Branch        294
## 7       1 Call Center   192
## 8       1 Web           156
```

```
# pie chart
ggplot(salesChannel, aes(x="", y=Count, fill=Channel)) +
  geom_bar(width=1, stat = "identity", position=position_fill()) +
  geom_text(aes(x=1.25, label=Count), position=position_fill(vjust = 0.5)) +
  coord_polar("y") +
  facet_wrap(~Engaged) +
  ggtitle('Sales Channel (0: Not Engaged, 1: Engaged)') +
  theme(
    axis.title.x=element_blank(),
    axis.title.y=element_blank(),
    plot.title=element_text(hjust=0.5),
    legend.position='bottom'
  )
```



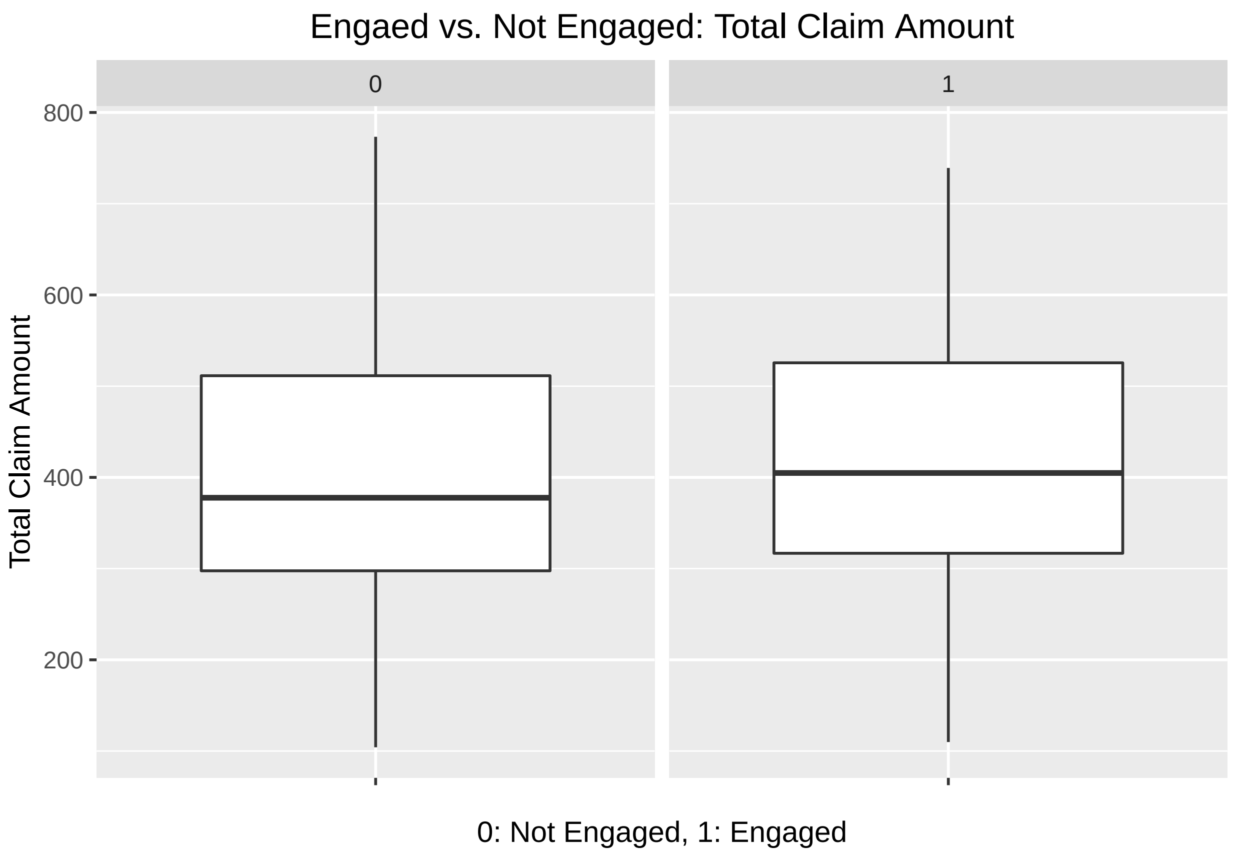
```
ggplot(df, aes(x="", y=Total.Claim.Amount)) +
  geom_boxplot() +
  facet_wrap(~Engaged) +
  ylab("Total Claim Amount") +
  xlab("0: Not Engaged, 1: Engaged") +
  ggtitle("Engaed vs. Not Engaged: Total Claim Amount") +
  theme(plot.title=element_text(hjust=0.5))
```



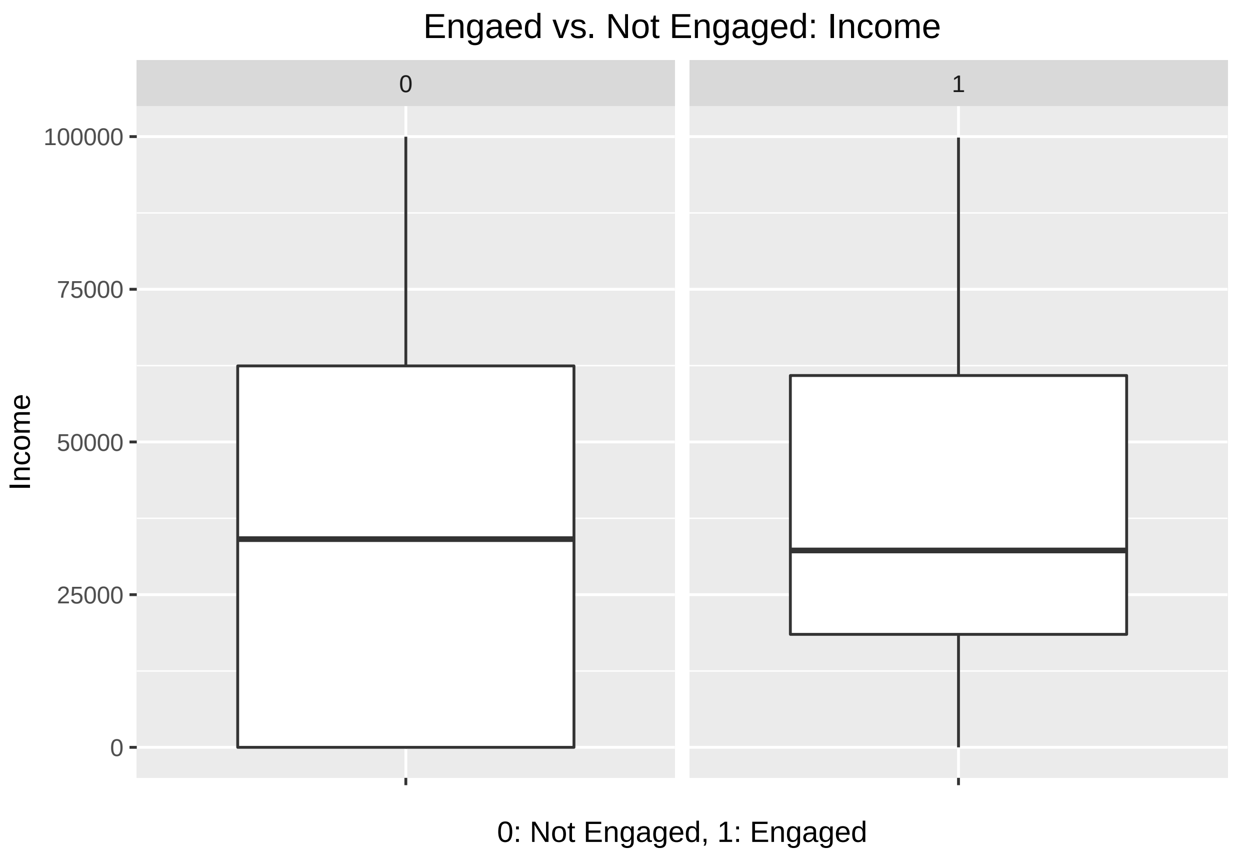
4. Total Claim Amount

```
# without outliers
ggplot(df, aes(x="", y=Total.Claim.Amount)) +
  geom_boxplot(outlier.shape = NA) +
  scale_y_continuous(limits = quantile(df$Total.Claim.Amount, c(0.1, 0.9))) +
  facet_wrap(~Engaged) +
  ylab("Total Claim Amount") +
  xlab("0: Not Engaged, 1: Engaged") +
  ggtitle("Engaed vs. Not Engaged: Total Claim Amount") +
  theme(plot.title=element_text(hjust=0.5))
```

```
## Warning: Removed 1828 rows containing non-finite values (stat_boxplot).
```



```
# boxplot
ggplot(df, aes(x="", y=Income)) +
  geom_boxplot() +
  facet_wrap(~Engaged) +
  ylab("Income") +
  xlab("0: Not Engaged, 1: Engaged") +
  ggtitle("Engaed vs. Not Engaged: Income") +
  theme(plot.title=element_text(hjust=0.5))
```



5. Income

```
# summary statistics
incomeDescription <- df %>%
  group_by(Engaged) %>%
  summarise(
    Min=min(Income), Q1=quantile(Income, 0.25),
    Median=median(Income), Q3=quantile(Income, 0.75),
    Max=max(Income)
  )

incomeDescription
```

```
## # A tibble: 2 x 6
##   Engaged   Min    Q1 Median     Q3   Max
##     <int> <int> <dbl>  <dbl>  <dbl> <int>
## 1       0     0     0  34091 62454. 99981
## 2       1     0 18495  32234 60880  99845
```

```
# summary statistics per column
summary(df)
```

6. Regression Analysis

```
##    Customer             State           Customer.Lifetime.Value    Response
##  Length:9134        Length:9134        Min.   : 1898           Min.   :0.0000
##  Class :character   Class :character   1st Qu.: 3994           1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median : 5780           Median :0.0000
##                                        Mean   : 8005           Mean   :0.1432
##                                        3rd Qu.: 8962           3rd Qu.:0.0000
##                                        Max.   :83325           Max.   :1.0000
##    Coverage          Education         Effective.To.Date  EmploymentStatus
##  Length:9134        Length:9134        Length:9134        Length:9134
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     Gender              Income      Location.Code      Marital.Status
##  Length:9134        Min.   :    0   Length:9134        Length:9134
##  Class :character   1st Qu.:    0   Class :character   Class :character
##  Mode  :character   Median :33890   Mode  :character   Mode  :character
##                     Mean   :37657
##                     3rd Qu.:62320
##                     Max.   :99981
##  Monthly.Premium.Auto Months.Since.Last.Claim Months.Since.Policy.Inception
##  Min.   : 61.00       Min.   : 0.0            Min.   : 0.00
##  1st Qu.: 68.00       1st Qu.: 6.0            1st Qu.:24.00
##  Median : 83.00       Median :14.0            Median :48.00
##  Mean   : 93.22       Mean   :15.1            Mean   :48.06
##  3rd Qu.:109.00       3rd Qu.:23.0            3rd Qu.:71.00
##  Max.   :298.00       Max.   :35.0            Max.   :99.00
##  Number.of.Open.Complaints Number.of.Policies Policy.Type
##  Min.   :0.0000            Min.   :1.000      Length:9134
##  1st Qu.:0.0000            1st Qu.:1.000      Class :character
##  Median :0.0000            Median :2.000      Mode  :character
##  Mean   :0.3844            Mean   :2.966
##  3rd Qu.:0.0000            3rd Qu.:4.000
##  Max.   :5.0000            Max.   :9.000
##     Policy          Renew.Offer.Type   Sales.Channel      Total.Claim.Amount
##  Length:9134        Length:9134        Length:9134        Min.   :   0.099
##  Class :character   Class :character   Class :character   1st Qu.: 272.258
##  Mode  :character   Mode  :character   Mode  :character   Median : 383.945
##                                                           Mean   : 434.089
##                                                           3rd Qu.: 547.515
##                                                           Max.   :2893.240
##  Vehicle.Class      Vehicle.Size          Engaged
##  Length:9134        Length:9134        Min.   :0.0000
##  Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :0.0000
##                                        Mean   :0.1432
##                                        3rd Qu.:0.0000
##                                        Max.   :1.0000
```

```r
# get data types of each column
sapply(df, class)
```

```
##                      Customer                         State
##                   "character"                   "character"
##       Customer.Lifetime.Value                      Response
##                     "numeric"                     "numeric"
##                      Coverage                     Education
##                   "character"                   "character"
##             Effective.To.Date              EmploymentStatus
##                   "character"                   "character"
##                        Gender                        Income
##                   "character"                     "integer"
##                 Location.Code                Marital.Status
##                   "character"                   "character"
##          Monthly.Premium.Auto       Months.Since.Last.Claim
##                     "integer"                     "integer"
## Months.Since.Policy.Inception     Number.of.Open.Complaints
##                     "integer"                     "integer"
##            Number.of.Policies                   Policy.Type
##                     "integer"                   "character"
##                        Policy              Renew.Offer.Type
##                   "character"                   "character"
##                 Sales.Channel            Total.Claim.Amount
##                   "character"                     "numeric"
##                 Vehicle.Class                  Vehicle.Size
##                   "character"                   "character"
##                       Engaged
##                     "integer"
```

6.1. Continuous Variables

```
# get numeric columns
continuousDF <- select_if(df, is.numeric)
colnames(continuousDF)
```

```
##  [1] "Customer.Lifetime.Value"       "Response"
##  [3] "Income"                        "Monthly.Premium.Auto"
##  [5] "Months.Since.Last.Claim"       "Months.Since.Policy.Inception"
##  [7] "Number.of.Open.Complaints"     "Number.of.Policies"
##  [9] "Total.Claim.Amount"            "Engaged"
```

```
# Fit regression model with continuous variables
logit.fit <- glm(Engaged ~ ., data = continuousDF, family = binomial)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
summary(logit.fit)
```

```
##
## Call:
## glm(formula = Engaged ~ ., family = binomial, data = continuousDF)
##
## Deviance Residuals:
```

```
##        Min          1Q      Median          3Q         Max
## -2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06   2.409e-06
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -2.657e+01  1.558e+04  -0.002    0.999
## Customer.Lifetime.Value      -2.204e-17  5.919e-01   0.000    1.000
## Response                      5.313e+01  1.065e+04   0.005    0.996
## Income                        1.129e-17  1.371e-01   0.000    1.000
## Monthly.Premium.Auto         -9.399e-15  1.539e+02   0.000    1.000
## Months.Since.Last.Claim       5.080e-14  3.705e+02   0.000    1.000
## Months.Since.Policy.Inception -2.210e-14  1.337e+02   0.000    1.000
## Number.of.Open.Complaints    -3.029e-13  4.096e+03   0.000    1.000
## Number.of.Policies           -7.612e-14  1.560e+03   0.000    1.000
## Total.Claim.Amount            7.330e-16  1.849e+01   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7.5033e+03  on 9133  degrees of freedom
## Residual deviance: 5.2992e-08  on 9124  degrees of freedom
## AIC: 20
##
## Number of Fisher Scoring iterations: 25
```

6.2. Categorical Variables

```
# a. Education
# Fit regression model with Education factor variables
logit.fit <- glm(Engaged ~ factor(Education), data = df, family = binomial)
summary(logit.fit)
```

```
##
## Call:
## glm(formula = Engaged ~ factor(Education), family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6211  -0.5746  -0.5440  -0.5287   2.0184
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -1.83575    0.05538 -33.146   <2e-16 ***
## factor(Education)College              0.11816    0.07719   1.531   0.1258
## factor(Education)Doctor               0.28819    0.15258   1.889   0.0589 .
## factor(Education)High School or Below -0.06137    0.08019  -0.765   0.4441
## factor(Education)Master               0.19191    0.11407   1.682   0.0925 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 7503.3  on 9133  degrees of freedom
## Residual deviance: 7492.4  on 9129  degrees of freedom
## AIC: 7502.4
##
## Number of Fisher Scoring iterations: 4
```

```
# b. Education + Gender
# Fit regression model with Education & Gender variables
logit.fit <- glm(Engaged ~ factor(Education) + factor(Gender), data = df, family = binomial)
summary(logit.fit)
```

```
##
## Call:
## glm(formula = Engaged ~ factor(Education) + factor(Gender), family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6247  -0.5713  -0.5409  -0.5256   2.0238
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -1.84803    0.06257 -29.537   <2e-16 ***
## factor(Education)College             0.11782    0.07720   1.526   0.1269
## factor(Education)Doctor              0.28759    0.15259   1.885   0.0595 .
## factor(Education)High School or Below -0.06173    0.08019  -0.770   0.4415
## factor(Education)Master              0.19223    0.11407   1.685   0.0919 .
## factor(Gender)M                      0.02534    0.05979   0.424   0.6717
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7503.3  on 9133  degrees of freedom
## Residual deviance: 7492.3  on 9128  degrees of freedom
## AIC: 7504.3
##
## Number of Fisher Scoring iterations: 4
```

6.3. Continuous & Categorical Variables

```
continuousDF$Gender <- factor(df$Gender)
continuousDF$Education <- factor(df$Education)
```

```
# Fit regression model with Education & Gender variables
logit.fit <- glm(Engaged ~ ., data = continuousDF, family = binomial)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
summary(logit.fit)
```

```
## 
## Call:
## glm(formula = Engaged ~ ., family = binomial, data = continuousDF)
## 
## Deviance Residuals: 
##        Min          1Q      Median          3Q         Max  
## -2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06   2.409e-06  
## 
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.657e+01  1.683e+04  -0.002    0.999
## Customer.Lifetime.Value       -5.968e-18  5.921e-01   0.000    1.000
## Response                       5.313e+01  1.065e+04   0.005    0.996
## Income                         2.473e-18  1.372e-01   0.000    1.000
## Monthly.Premium.Auto          -2.375e-15  1.548e+02   0.000    1.000
## Months.Since.Last.Claim        1.087e-14  3.708e+02   0.000    1.000
## Months.Since.Policy.Inception -3.975e-15  1.338e+02   0.000    1.000
## Number.of.Open.Complaints     -4.359e-14  4.098e+03   0.000    1.000
## Number.of.Policies            -2.890e-14  1.561e+03   0.000    1.000
## Total.Claim.Amount             2.593e-16  1.881e+01   0.000    1.000
## GenderM                       -1.266e-13  7.495e+03   0.000    1.000
## EducationCollege              -2.397e-13  9.674e+03   0.000    1.000
## EducationDoctor               -1.944e-13  2.048e+04   0.000    1.000
## EducationHigh School or Below -2.262e-13  9.763e+03   0.000    1.000
## EducationMaster               -2.866e-13  1.482e+04   0.000    1.000
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 7.5033e+03  on 9133  degrees of freedom
## Residual deviance: 5.2992e-08  on 9119  degrees of freedom
## AIC: 30
## 
## Number of Fisher Scoring iterations: 25
```

REPORT

We fitted a logistic model (estimated using ML) to predict Engaged with Customer.Lifetime.Value, Response, Income, Monthly.Premium.Auto, Months.Since.Last.Claim, Months.Since.Policy.Inception, Number.of.Open.Complaints, Number.of.Policies, Total.Claim.Amount, Gender and Education (formula: Engaged ~ Customer.Lifetime.Value + Response + Income + Monthly.Premium.Auto + Months.Since.Last.Claim + Months.Since.Policy.Inception + Number.of.Open.Complaints + Number.of.Policies + Total.Claim.Amount + Gender + Education). The model's explanatory power is substantial (Tjur's R2 = 1.00). The model's intercept, corresponding to Customer.Lifetime.Value = 0, Response = 0, Income = 0, Monthly.Premium.Auto = 0, Months.Since.Last.Claim = 0, Months.Since.Policy.Inception = 0, Number.of.Open.Complaints = 0, Number.of.Policies = 0, Total.Claim.Amount = 0, Gender = F and Education = Bachelor, is at -26.57 (95% CI [-33004.41, 32951.28], p = 0.999). Within this model:

- The effect of Customer Lifetime Value is statistically non-significant and negative (beta = -5.97e-18, 95% CI [-1.16, 1.16], p > .999; Std. beta = -8.93e-14, 95% CI [-7973.61, 7973.61])
- The effect of Response is statistically non-significant and positive (beta = 53.13, 95% CI [-20828.75, 20935.02], p = 0.996; Std. beta = 18.61, 95% CI [-7296.25, 7333.48])

- The effect of Income is statistically non-significant and positive (beta = 2.47e-18, 95% CI [-0.27, 0.27], p > .999; Std. beta = 3.27e-13, 95% CI [-8170.49, 8170.49])
- The effect of Monthly Premium Auto is statistically non-significant and negative (beta = -2.38e-15, 95% CI [-303.44, 303.44], p > .999; Std. beta = -3.57e-13, 95% CI [-10440.80, 10440.80])
- The effect of Months Since Last Claim is statistically non-significant and positive (beta = 1.09e-14, 95% CI [-726.66, 726.66], p > .999; Std. beta = 3.62e-13, 95% CI [-7319.87, 7319.87])
- The effect of Months Since Policy Inception is statistically non-significant and negative (beta = -3.97e-15, 95% CI [-262.15, 262.15], p > .999; Std. beta = -6.34e-13, 95% CI [-7315.69, 7315.69])
- The effect of Number of Open Complaints is statistically non-significant and negative (beta = -4.36e-14, 95% CI [-8032.86, 8032.86], p > .999; Std. beta = -2.87e-13, 95% CI [-7312.98, 7312.98])
- The effect of Number of Policies is statistically non-significant and negative (beta = -2.89e-14, 95% CI [-3059.46, 3059.46], p > .999; Std. beta = -3.46e-13, 95% CI [-7312.66, 7312.66])
- The effect of Total Claim Amount is statistically non-significant and positive (beta = 2.59e-16, 95% CI [-36.86, 36.86], p > .999; Std. beta = 2.95e-13, 95% CI [-10708.22, 10708.22])
- The effect of Gender [M] is statistically non-significant and negative (beta = -1.27e-13, 95% CI [-14690.68, 14690.68], p > .999; Std. beta = -8.63e-13, 95% CI [-14690.68, 14690.68])
- The effect of Education [College] is statistically non-significant and negative (beta = -2.40e-13, 95% CI [-18961.08, 18961.08], p > .999; Std. beta = -1.64e-12, 95% CI [-18961.08, 18961.08])
- The effect of Education [Doctor] is statistically non-significant and negative (beta = -1.94e-13, 95% CI [-40130.84, 40130.84], p > .999; Std. beta = -1.57e-12, 95% CI [-40130.84, 40130.84])
- The effect of Education [High School or Below] is statistically non-significant and negative (beta = -2.26e-13, 95% CI [-19135.04, 19135.04], p > .999; Std. beta = -1.38e-12, 95% CI [-19135.04, 19135.04])
- The effect of Education [Master] is statistically non-significant and negative (beta = -2.87e-13, 95% CI [-29052.10, 29052.10], p > .999; Std. beta = -2.69e-12, 95% CI [-29052.10, 29052.10])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using.