Predicting Churn at QWE

Alex Osterbuhr

2023-09-01

Preparation:

Data Import & Cleanse:

```
qwe_data <- read_excel("Predicting Customer Church at QWE Inc.xlsx", sheet=2)

table(qwe_data$`Churn (1 = Yes, 0 = No)`)
```

```
##
##    0    1
## 6024  323
```

```
head(qwe_data)
```

```
## # A tibble: 6 × 13
##      ID Custom…¹ Churn…² CHI S…³ CHI S…⁴ Suppo…⁵ Suppo…⁶ SP Mo…⁷ SP 0-…⁸ Login…⁹
##   <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1     1       67       0       0       0       0       0       0       0       0
## 2     2       67       0      62       4       0       0       0       0       0
## 3     3       55       0       0       0       0       0       0       0       0
## 4     4       63       0     231       1       1      -1       3       0     167
## 5     5       57       0      43      -1       0       0       0       0       0
## 6     6       58       0     138     -10       0       0       0       0      43
## # … with 3 more variables: `Blog Articles 0-1` <dbl>, `Views 0-1` <dbl>,
## #   `Days Since Last Login 0-1` <dbl>, and abbreviated variable names
## #   ¹`Customer Age (in months)`, ²`Churn (1 = Yes, 0 = No)`,
## #   ³`CHI Score Month 0`, ⁴`CHI Score 0-1`, ⁵`Support Cases Month 0`,
## #   ⁶`Support Cases 0-1`, ⁷`SP Month 0`, ⁸`SP 0-1`, ⁹`Logins 0-1`
```

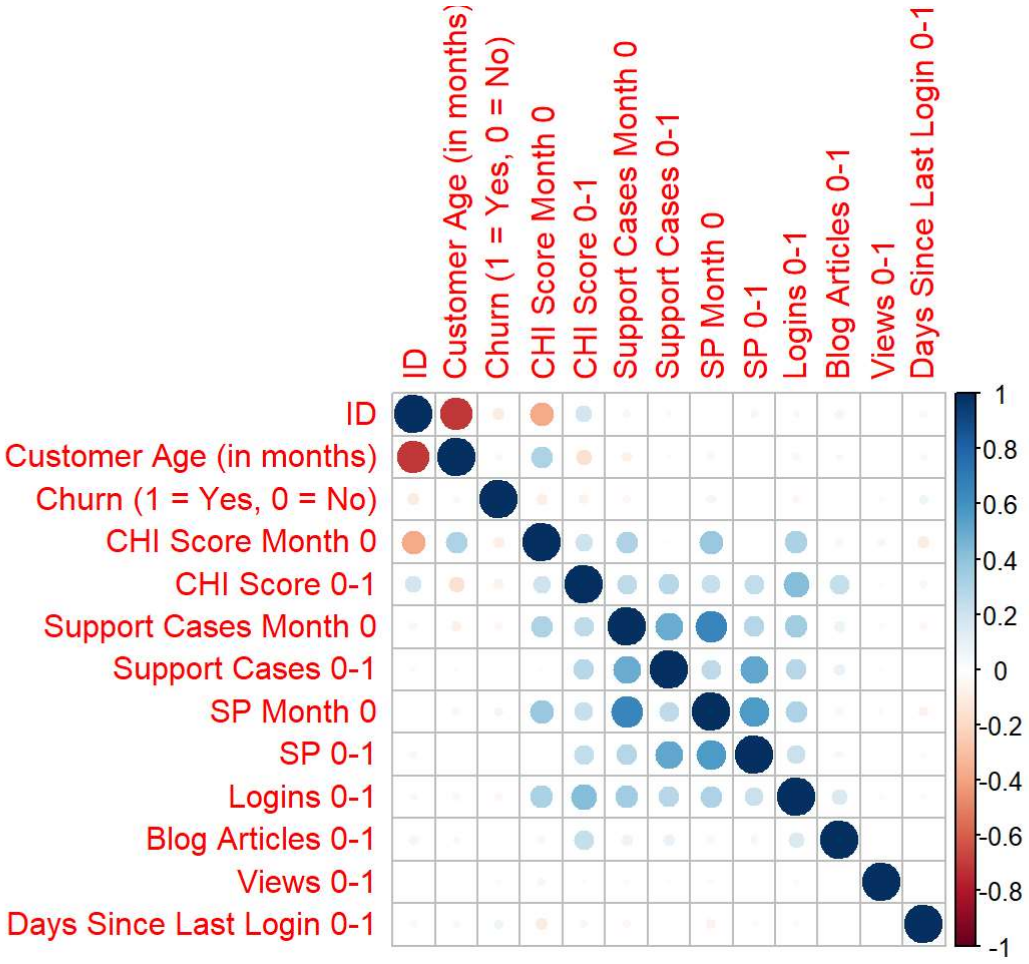
```
summary(qwe_data)
```

```
##        ID       Customer Age (in months) Churn (1 = Yes, 0 = No)
##  Min.   :   1   Min.   : 0.0             Min.   :0.00000
##  1st Qu.:1588   1st Qu.: 5.0             1st Qu.:0.00000
##  Median :3174   Median :11.0             Median :0.00000
##  Mean   :3174   Mean   :13.9             Mean   :0.05089
##  3rd Qu.:4760   3rd Qu.:20.0             3rd Qu.:0.00000
##  Max.   :6347   Max.   :67.0             Max.   :1.00000
##  CHI Score Month 0 CHI Score 0-1      Support Cases Month 0
##  Min.   :  0.00    Min.   :-125.000   Min.   : 0.0000
##  1st Qu.: 24.50    1st Qu.:  -8.000   1st Qu.: 0.0000
##  Median : 87.00    Median :   0.000   Median : 0.0000
##  Mean   : 87.32    Mean   :   5.059   Mean   : 0.7063
##  3rd Qu.:139.00    3rd Qu.:  15.000   3rd Qu.: 1.0000
##  Max.   :298.00    Max.   : 208.000   Max.   :32.0000
##  Support Cases 0-1      SP Month 0         SP 0-1           Logins 0-1
##  Min.   :-29.000000   Min.   :0.0000   Min.   :-4.00000   Min.   :-293.00
##  1st Qu.:  0.000000   1st Qu.:0.0000   1st Qu.: 0.00000   1st Qu.:  -1.00
##  Median :  0.000000   Median :0.0000   Median : 0.00000   Median :   2.00
##  Mean   : -0.006932   Mean   :0.8128   Mean   : 0.03017   Mean   :  15.73
##  3rd Qu.:  0.000000   3rd Qu.:2.6667   3rd Qu.: 0.00000   3rd Qu.:  23.00
##  Max.   : 31.000000   Max.   :4.0000   Max.   : 4.00000   Max.   : 865.00
##  Blog Articles 0-1    Views 0-1         Days Since Last Login 0-1
##  Min.   :-75.0000   Min.   :-28322.00   Min.   :-648.000
##  1st Qu.:  0.0000   1st Qu.:   -11.00   1st Qu.:   0.000
##  Median :  0.0000   Median :     0.00   Median :   0.000
##  Mean   :  0.1572   Mean   :    96.31   Mean   :   1.765
##  3rd Qu.:  0.0000   3rd Qu.:    27.00   3rd Qu.:   3.000
##  Max.   :217.0000   Max.   :230414.00   Max.   :  61.000
```

```
str(qwe_data)
```

```
## tibble [6,347 × 13] (S3: tbl_df/tbl/data.frame)
##  $ ID                       : num [1:6347] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Customer Age (in months) : num [1:6347] 67 67 55 63 57 58 57 46 56 56 ...
##  $ Churn (1 = Yes, 0 = No)  : num [1:6347] 0 0 0 0 0 0 0 0 0 0 ...
##  $ CHI Score Month 0        : num [1:6347] 0 62 0 231 43 138 180 116 78 78 ...
##  $ CHI Score 0-1            : num [1:6347] 0 4 0 1 -1 -10 -5 -11 -7 -37 ...
##  $ Support Cases Month 0    : num [1:6347] 0 0 0 1 0 0 1 0 1 0 ...
##  $ Support Cases 0-1        : num [1:6347] 0 0 0 -1 0 0 1 0 -2 0 ...
##  $ SP Month 0               : num [1:6347] 0 0 0 3 0 0 3 0 3 0 ...
##  $ SP 0-1                   : num [1:6347] 0 0 0 0 0 0 3 0 0 0 ...
##  $ Logins 0-1               : num [1:6347] 0 0 0 167 0 43 13 0 -9 -7 ...
##  $ Blog Articles 0-1        : num [1:6347] 0 0 0 -8 0 0 -1 0 1 0 ...
##  $ Views 0-1                : num [1:6347] 0 -16 0 21996 9 ...
##  $ Days Since Last Login 0-1: num [1:6347] 31 31 31 0 31 0 0 6 7 14 ...
```

```
corrplot(cor(qwe_data))
```

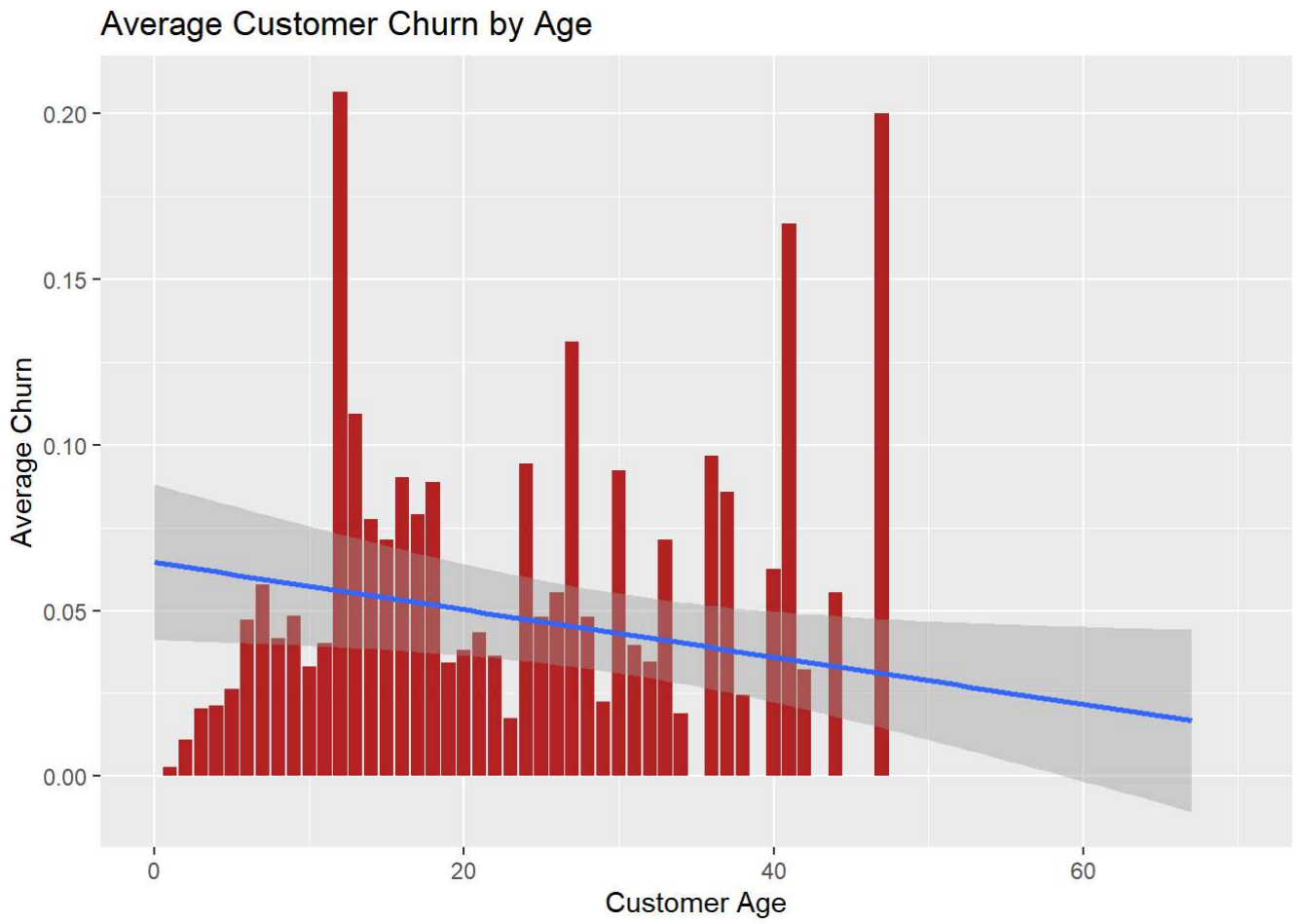



```
churn_avg <- qwe_data %>%
  group_by(`Customer Age (in months)`) %>%
  summarize(churn_avg_age = mean(`Churn (1 = Yes, 0 = No)`))

qwe_data <- qwe_data %>%
 mutate(no_churn = (`Churn (1 = Yes, 0 = No)`==0)) %>%
 mutate(yes_churn = (`Churn (1 = Yes, 0 = No)`==1))

qwe_data$ID <- as.factor(qwe_data$ID)
```

Create a graph summarizing average churn by customer age.

Bar graph shows trend line.

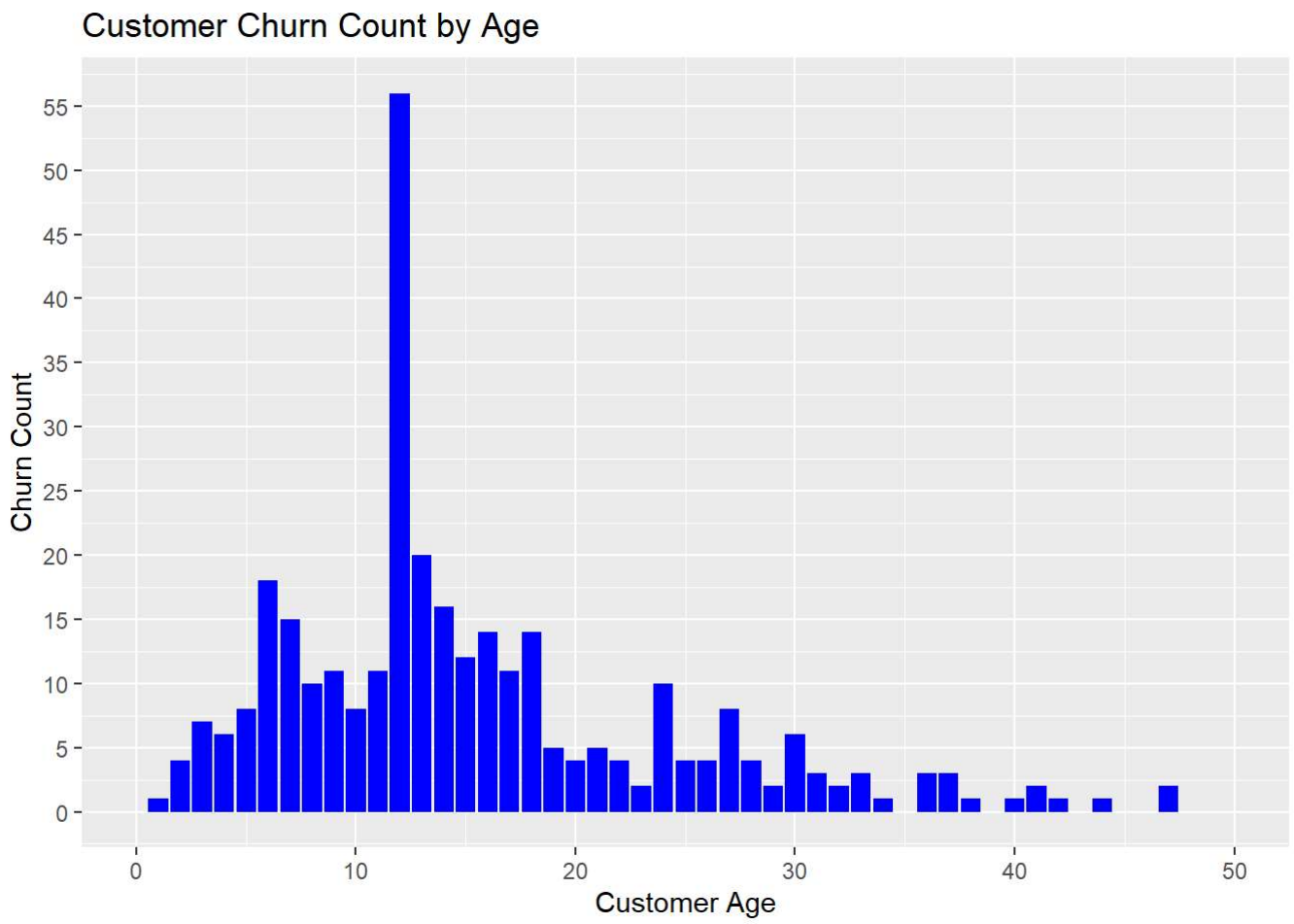
```
churn_avg %>%
  ggplot(aes(x= `Customer Age (in months)`, y= churn_avg_age)) +
  geom_bar(stat = 'unique', fill = 'firebrick') +
  geom_smooth(method='lm')+
  labs(x = 'Customer Age', y = 'Average Churn', title = 'Average Customer Churn by Age') +
  scale_x_continuous(limits= c(0, 70))
```



Create a graph summarizing the number of customers who churn by customer age.

Bar graph shows up to age 50 as ages past 47 show no churn.

```
qwe_data %>%
  ggplot(aes(x= `Customer Age (in months)`, y= `Churn (1 = Yes, 0 = No)`)) +
  geom_bar(stat = 'identity', fill = 'blue') +
  labs(x = 'Customer Age', y = 'Churn Count', title = 'Customer Churn Count by Age') +
  scale_x_continuous(limits= c(0, 50))+
  scale_y_continuous(breaks = scales::pretty_breaks(n=13))
```



What is the customer age in months with the highest average churn?

The customer age in months with the highest average churn is 12 at .22 average churn.

```
desc_avg_churn <- churn_avg %>%
  arrange(desc(churn_avg_age))

head(desc_avg_churn, 5)
```

```
## # A tibble: 5 × 2
##   `Customer Age (in months)` churn_avg_age
##                        <dbl>         <dbl>
## 1                         12         0.207
## 2                         47         0.2
## 3                         41         0.167
## 4                         27         0.131
## 5                         13         0.109
```

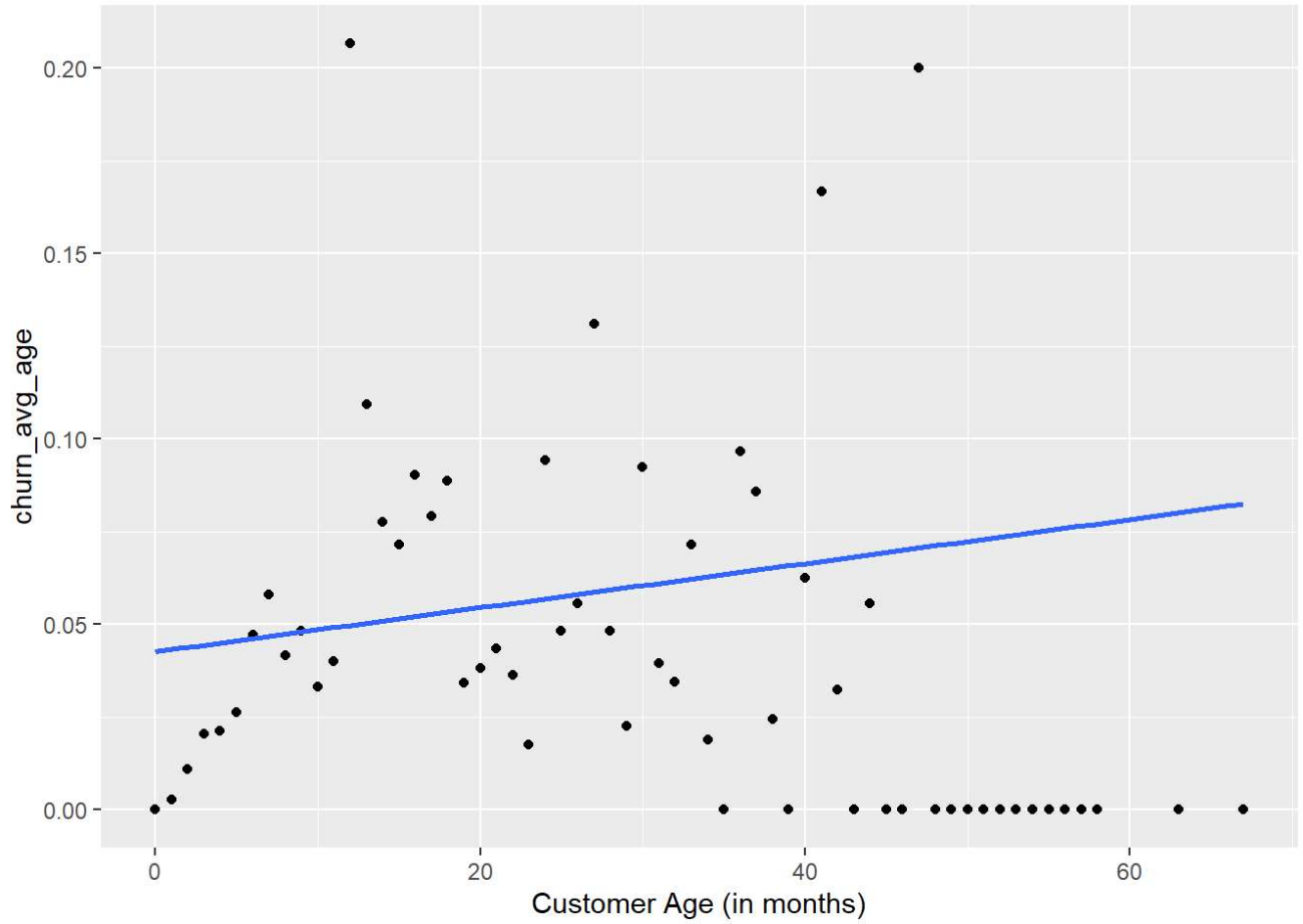
Is Wall's intuition that the churn rates depend on customer age confirmed by your graphs?

The scatterplot below shows the coefficients of a generalized linear regression model based on Age to Churn Average. A look at this graph, the low r squared of the model, and the graphs above indicate that there is a slight significant dependence of age on churn.

```
age_churn_lm <- glm(data=qwe_data, formula = `Churn (1 = Yes, 0 = No)`~`Customer Age (in months)
`)
summary(age_churn_lm)
```

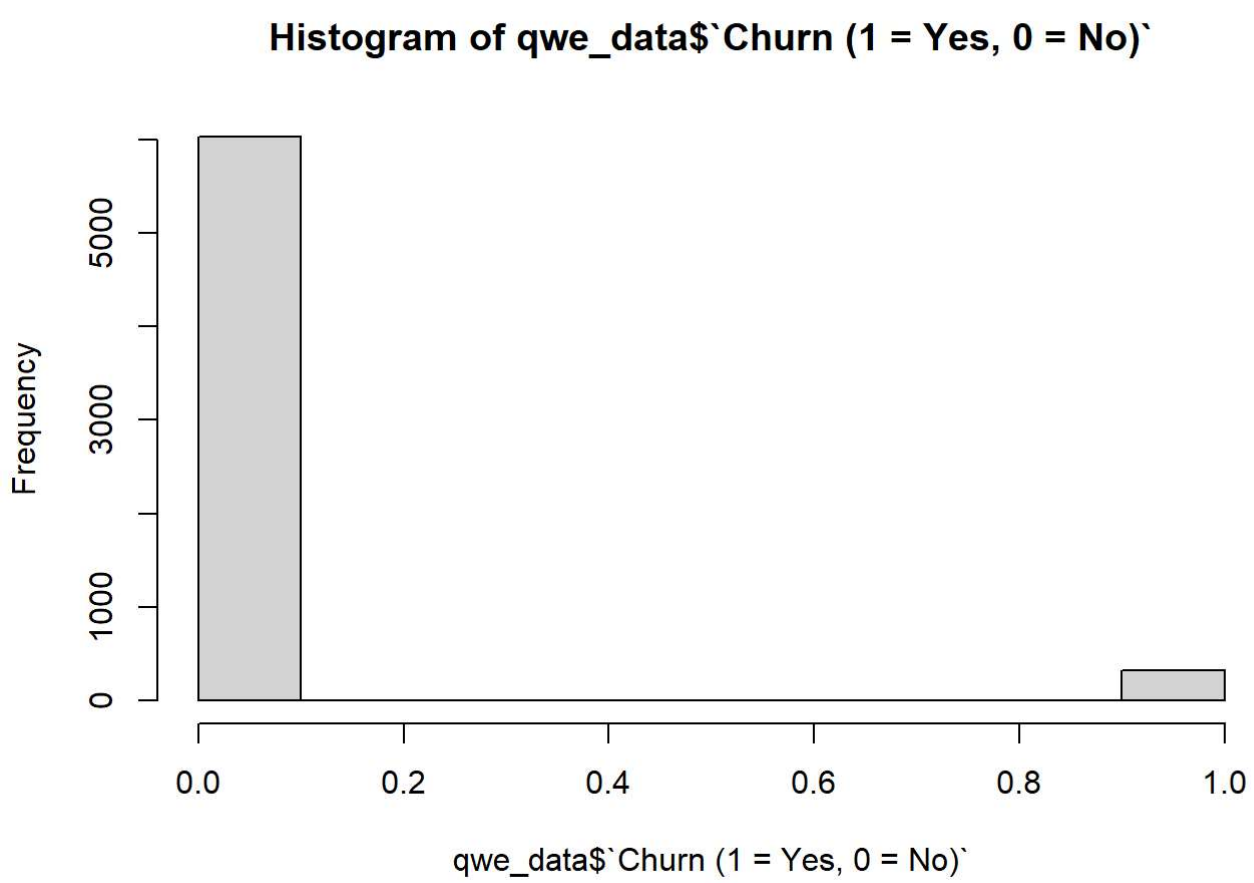
```
##
## Call:
## glm(formula = `Churn (1 = Yes, 0 = No)` ~ `Customer Age (in months)`,
##     data = qwe_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.08249  -0.05393  -0.04857  -0.04500   0.95678
##
## Coefficients:
##                              Estimate Std. Error t value            Pr(>|t|)
## (Intercept)                 0.0426206  0.0044046   9.676 <0.0000000000000002 ***
## `Customer Age (in months)` 0.0005951  0.0002471   2.408              0.0161 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.04827149)
##
##     Null deviance: 306.56  on 6346  degrees of freedom
## Residual deviance: 306.28  on 6345  degrees of freedom
## AIC: -1221.2
##
## Number of Fisher Scoring iterations: 2
```

```
qwe_data <- left_join(qwe_data, churn_avg)

ggplot(qwe_data, aes(x = `Customer Age (in months)`, y = churn_avg_age))+
  geom_point()+
  geom_smooth(method = 'lm', se = FALSE)
```



Univariate Testing:

```
hist(qwe_data$`Churn (1 = Yes, 0 = No)`)
```




```
churn_model <- glm(`Churn (1 = Yes, 0 = No)`~ .-ID- no_churn - yes_churn-churn_avg_age, data=qwe
_data, family='binomial')
summary(churn_model)
```

```
##
## Call:
## glm(formula = `Churn (1 = Yes, 0 = No)` ~ . - ID - no_churn -
##     yes_churn - churn_avg_age, family = "binomial", data = qwe_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0047  -0.3542  -0.2957  -0.2328   3.0660
##
## Coefficients:
##                               Estimate  Std. Error z value
## (Intercept)                 -2.76266978  0.10690964 -25.841
## `Customer Age (in months)`   0.01270574  0.00537044   2.366
## `CHI Score Month 0`         -0.00465732  0.00122317  -3.808
## `CHI Score 0-1`             -0.01027455  0.00247427  -4.153
## `Support Cases Month 0`     -0.15237026  0.10491954  -1.452
## `Support Cases 0-1`          0.17026293  0.09049865   1.881
## `SP Month 0`                 0.01592897  0.10217346   0.156
## `SP 0-1`                    -0.05193696  0.07851520  -0.661
## `Logins 0-1`                 0.00028933  0.00209233   0.138
## `Blog Articles 0-1`          0.00029049  0.01959896   0.015
## `Views 0-1`                 -0.00010978  0.00004071  -2.697
## `Days Since Last Login 0-1`  0.01724209  0.00428876   4.020
##                                        Pr(>|z|)
## (Intercept)                 < 0.0000000000000002 ***
## `Customer Age (in months)`               0.01799 *
## `CHI Score Month 0`                      0.00014 ***
## `CHI Score 0-1`                        0.0000329 ***
## `Support Cases Month 0`                  0.14643
## `Support Cases 0-1`                      0.05992 .
## `SP Month 0`                             0.87611
## `SP 0-1`                                 0.50830
## `Logins 0-1`                             0.89002
## `Blog Articles 0-1`                      0.98817
## `Views 0-1`                              0.00700 **
## `Days Since Last Login 0-1`            0.0000581 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2440.3  on 6335  degrees of freedom
## AIC: 2464.3
##
## Number of Fisher Scoring iterations: 7
```

```
binom.test(x=6024, n=6347, p=.3, alternative="less")
```

```
##
##  Exact binomial test
##
## data:  6024 and 6347
## number of successes = 6024, number of trials = 6347, p-value = 1
## alternative hypothesis: true probability of success is less than 0.3
## 95 percent confidence interval:
##  0.0000000 0.9535725
## sample estimates:
## probability of success
##              0.9491098
```

Based on univariate testing, which attributes are significant predictors of churn? List them out by name.

According to the glm above, significant predictors of churn may be: -CHI Score Month 0 -CHI Score 0-1 -Days Since Last Login 0-1 -Customer Age (in months) -Views 0-1

Logistic Regression:

Run a logistic regression predicting churn with all variables.

Model was ran in the previous section without taking into account features created through this analysis as well as “ID” as that is a unique identifier for each row which does not repeat.

```
summary(churn_model)
```

```
##
## Call:
## glm(formula = `Churn (1 = Yes, 0 = No)` ~ . - ID - no_churn -
##     yes_churn - churn_avg_age, family = "binomial", data = qwe_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0047  -0.3542  -0.2957  -0.2328   3.0660
##
## Coefficients:
##                               Estimate  Std. Error z value
## (Intercept)                 -2.76266978  0.10690964 -25.841
## `Customer Age (in months)`   0.01270574  0.00537044   2.366
## `CHI Score Month 0`         -0.00465732  0.00122317  -3.808
## `CHI Score 0-1`             -0.01027455  0.00247427  -4.153
## `Support Cases Month 0`     -0.15237026  0.10491954  -1.452
## `Support Cases 0-1`          0.17026293  0.09049865   1.881
## `SP Month 0`                 0.01592897  0.10217346   0.156
## `SP 0-1`                    -0.05193696  0.07851520  -0.661
## `Logins 0-1`                 0.00028933  0.00209233   0.138
## `Blog Articles 0-1`          0.00029049  0.01959896   0.015
## `Views 0-1`                 -0.00010978  0.00004071  -2.697
## `Days Since Last Login 0-1`  0.01724209  0.00428876   4.020
##                                        Pr(>|z|)
## (Intercept)                 < 0.0000000000000002 ***
## `Customer Age (in months)`               0.01799 *
## `CHI Score Month 0`                      0.00014 ***
## `CHI Score 0-1`                        0.0000329 ***
## `Support Cases Month 0`                  0.14643
## `Support Cases 0-1`                      0.05992 .
## `SP Month 0`                             0.87611
## `SP 0-1`                                 0.50830
## `Logins 0-1`                             0.89002
## `Blog Articles 0-1`                      0.98817
## `Views 0-1`                              0.00700 **
## `Days Since Last Login 0-1`            0.0000581 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2440.3  on 6335  degrees of freedom
## AIC: 2464.3
##
## Number of Fisher Scoring iterations: 7
```

What is the AIC of the full model?

```
print(paste("The AIC of the full model is:", round(churn_model$aic, 2)))
```

```
## [1] "The AIC of the full model is: 2464.33"
```

Reduce the model to only include variables you find consequential.

```
reduce_model <- glm(data = qwe_data, formula = `Churn (1 = Yes, 0 = No)`~
                      `CHI Score Month 0`+
                      `CHI Score 0-1`+
                      `Days Since Last Login 0-1`+
                      `Customer Age (in months)`+
                      `Views 0-1`, family=binomial(link='logit'))
```

List the variables that are included.

Chosen variables are similar to those from the univariate testing section. -CHI Score Month 0 -CHI Score 0-1 -Days Since Last Login 0-1 -Customer Age (in months) -Views 0-1

What is the AIC of the reduced model you came up with?

```
print(paste("The AIC of the reduced model is:", round(reduce_model$aic, 2)))
```

```
## [1] "The AIC of the reduced model is: 2459.42"
```

Does the AIC value of the reduced model align with your expectations? Why or why not?

I expected the AIC to drop significantly. With only a drop of about 5, reducing the model does not significantly improve the model based on AIC even though we only included consequential variables. ## Prep for predictions Using liberal prediction threshold due to nature of what we are predicting for.

```
preds <- predict.glm(reduce_model, newdata=qwe_data, type='response')
qwe_data$predicted_churn <- if_else(preds>=.3, 'Yes (1)', 'No (0)')
qwe_data$predicted_values <- preds
```

What is the predicted probability that customer 1023 will leave? Is that high or low? Did the customer leave?

At .02 the predicted probability is low. The customer did not leave as predicted.

```
qwe_data %>%
  select(ID, `Customer Age (in months)`, `Churn (1 = Yes, 0 = No)`, predicted_churn, predicted_v
alues) %>%
  filter(ID==1023)
```

```
## # A tibble: 1 × 5
##   ID    `Customer Age (in months)` `Churn (1 = Yes, 0 = No)` predicted…¹ predi…²
##   <fct>                      <dbl>                     <dbl> <chr>         <dbl>
## 1 1023                           4                         0 No (0)       0.0237
## # … with abbreviated variable names ¹predicted_churn, ²predicted_values
```

What is the predicted probability that customer 3769 will leave? Is that high or low? Did the customer leave?

At .08 the predicted probability is low. The customer did leave. This was not predicted.

```
qwe_data %>%
  select(ID, `Customer Age (in months)`, `Churn (1 = Yes, 0 = No)`, predicted_churn, predicted_v
alues) %>%
  filter(ID==3769)
```

```
## # A tibble: 1 × 5
##   ID    `Customer Age (in months)` `Churn (1 = Yes, 0 = No)` predicted…¹ predi…²
##   <fct>                      <dbl>                     <dbl> <chr>         <dbl>
## 1 3769                           9                         1 No (0)       0.0828
## # … with abbreviated variable names ¹predicted_churn, ²predicted_values
```

What is the predicted probability that customer 4168 will leave? Is that high or low? Did the customer leave?

At .06 the predicted probability is low. The customer did not leave as predicted.

```
qwe_data %>%
  select(ID, `Customer Age (in months)`, `Churn (1 = Yes, 0 = No)`, predicted_churn, predicted_v
alues) %>%
  filter(ID==4168)
```

```
## # A tibble: 1 × 5
##   ID    `Customer Age (in months)` `Churn (1 = Yes, 0 = No)` predicted…¹ predi…²
##   <fct>                      <dbl>                     <dbl> <chr>         <dbl>
## 1 4168                          18                         0 No (0)       0.0650
## # … with abbreviated variable names ¹predicted_churn, ²predicted_values
```

What is the predicted probability that customer 357 will leave? Is that high or low? Did the customer leave?

At .3 the predicted probability is high (based on the set threshold). The customer did leave as predicted.

```
qwe_data %>%
  select(ID, `Customer Age (in months)`, `Churn (1 = Yes, 0 = No)`, predicted_churn, predicted_v
alues) %>%
  filter(ID==357)
```

```
## # A tibble: 1 × 5
##   ID    `Customer Age (in months)` `Churn (1 = Yes, 0 = No)` predicted…¹ predi…²
##   <fct>                      <dbl>                     <dbl> <chr>         <dbl>
## 1 357                           12                         1 Yes (1)       0.329
## # … with abbreviated variable names ¹predicted_churn, ²predicted_values
```

Subset the data per Wall's intuition and re-run the analysis in section 3 above. Your output should include the AIC of the full and reduced models, the predicted probability for customers 1023, 3769, 4168, and 357, and summary comments.

Wall's intuition from the reading states he believed the following factors could be used to predict churn: age (0-6mo learners, 6mo-14mo is riskiest, >14mo probably least likely), CHI (particularly low CHI and those who have dropped score recently), service (many service requests and high priority requests most likely) and usage (less logs, less blogs, and less views). All final prediction values of old models ran on new subset plus new models created below will follow the final predictions table. Although Wall predicted Age range of 6-14 months, subset includes a range of 4-18 so that all four customers above will be included in the subset.

```
subset_segment <- qwe_data$`Customer Age (in months)` >= 4 & qwe_data$`Customer Age (in months)`
<= 18

churn_subset <- qwe_data[subset_segment, ]
```

```
new_reduce_model <- glm(data = churn_subset, formula = `Churn (1 = Yes, 0 = No)`~
                      `CHI Score Month 0`+
                      `CHI Score 0-1`+
                      `Days Since Last Login 0-1`+
                      `Customer Age (in months)`+
                      `Views 0-1`, family=binomial(link='logit'))

new_churn_model <- glm(`Churn (1 = Yes, 0 = No)`~ .-ID- no_churn - yes_churn-churn_avg_age, data
=churn_subset, family='binomial')

new_reduce_model$aic
```

```
## [1] 1565.25
```

```
new_churn_model$aic
```

```
## [1] 1549.218
```

Based on the model you think performs the best:

```
age_model <- glm(data = churn_subset, formula = `Churn (1 = Yes, 0 = No)`~`Customer Age (in mont
hs)`, family=binomial(link='logit'))

CHI_model <- glm(data = churn_subset, formula = `Churn (1 = Yes, 0 = No)`~ `CHI Score Month 0`,
family=binomial(link='logit'))

CHIchange_model <- glm(data = churn_subset, formula = `Churn (1 = Yes, 0 = No)`~ `CHI Score 0-1
`, family=binomial(link='logit'))

service_model <- glm(data = churn_subset, formula = `Churn (1 = Yes, 0 = No)`~`Support Cases Mon
th 0`, family=binomial(link='logit'))

priority_model <- glm(data = churn_subset, formula = `Churn (1 = Yes, 0 = No)`~`SP Month 0`, fam
ily=binomial(link='logit'))

login_model <- glm(data = churn_subset, formula = `Churn (1 = Yes, 0 = No)`~`Logins 0-1`, family
=binomial(link='logit'))

blog_model <- glm(data = churn_subset, formula = `Churn (1 = Yes, 0 = No)`~`Blog Articles 0-1`,
family=binomial(link='logit'))

views_model <- glm(data = churn_subset, formula = `Churn (1 = Yes, 0 = No)`~`Views 0-1`, family=
binomial(link='logit'))

Wall_model <- glm(data = churn_subset, formula = `Churn (1 = Yes, 0 = No)`~`CHI Score Month 0`+`
CHI Score 0-1`+`Logins 0-1`+`Customer Age (in months)`+`Views 0-1`+`Support Cases Month 0`+`SP M
onth 0`+`Blog Articles 0-1`, family=binomial(link='logit'))

AIC_table <- data.frame(Model=c('age_model', 'CHI_model', 'CHIchange_model', 'service_model', 'p
riority_model', 'login_model', 'blog_model', 'views_model', 'Wall_model', 'churn_model', 'reduce
_model'), AIC=c(age_model$aic,CHI_model$aic,CHIchange_model$aic, service_model$aic, priority_mod
el$aic, login_model$aic, blog_model$aic, views_model$aic, Wall_model$aic, churn_model$aic, reduc
e_model$aic))

AIC_table[order(AIC_table$AIC,decreasing=FALSE),]
```

```
##              Model      AIC
## 9       Wall_model 1564.342
## 2        CHI_model 1618.698
## 5   priority_model 1646.267
## 4    service_model 1656.858
## 1        age_model 1665.669
## 3  CHIchange_model 1676.755
## 6      login_model 1686.417
## 8      views_model 1693.953
## 7       blog_model 1698.875
## 11    reduce_model 2459.419
## 10     churn_model 2464.332
```

```
added_predictions <- churn_subset %>%
  mutate(reduce_predict=predicted_values) %>%
  mutate(new_reduce_predict=predict.glm(new_reduce_model, newdata=churn_subset, type='respons
e')) %>%
  mutate(churn_predict=predict.glm(churn_model, newdata=churn_subset, type='response')) %>%
  mutate(new_churn_predict=predict.glm(new_churn_model, newdata=churn_subset, type='response'))
%>%
  mutate(Wall_predict=predict.glm(Wall_model, newdata=churn_subset, type='response')) %>%
  mutate(CHI_predict=predict.glm(CHI_model, newdata=churn_subset, type='response'))

final_predictions <-added_predictions %>%
  select(ID,`Churn (1 = Yes, 0 = No)`, reduce_predict,new_reduce_predict, churn_predict, new_chu
rn_predict, Wall_predict, CHI_predict)

final_predictions %>% filter(ID %in% c(1023, 3769, 4168, 357))
```

```
## # A tibble: 4 × 8
##   ID    Churn (1 = Yes, 0 = No…¹ reduc…² new_r…³ churn…⁴ new_c…⁵ Wall_…⁶ CHI_p…⁷
##   <fct>                     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 357                           1  0.329   0.489   0.350   1.00    0.423   0.0193
## 2 1023                          0  0.0237  0.0208  0.0257  0.0338  0.0290  0.0523
## 3 3769                          1  0.0828  0.0929  0.0870  0.109   0.0853  0.0881
## 4 4168                          0  0.0650  0.226   0.0640  0.188   0.248   0.131
## # … with abbreviated variable names ¹`Churn (1 = Yes, 0 = No)`,
## #   ²reduce_predict, ³new_reduce_predict, ⁴churn_predict, ⁵new_churn_predict,
## #   ⁶Wall_predict, ⁷CHI_predict
```

Although we took a subset, the new_churn_model seems to be the only model that performs significantly well in terms of predictability accuracy.

Which 10 customers are the most likely to churn?

```
final_predictions %>%
  select(ID, new_churn_predict) %>%
  arrange(desc(new_churn_predict)) %>%
  slice_head(n=10)
```

```
## # A tibble: 10 × 2
##    ID    new_churn_predict
##    <fct>             <dbl>
##  1 357               1.00
##  2 1672              0.464
##  3 1616              0.434
##  4 1574              0.427
##  5 299               0.422
##  6 2546              0.407
##  7 1693              0.391
##  8 1021              0.350
##  9 335               0.335
## 10 1563              0.334
```

What is their predicted probability of churn?

Probability of churn are as follows (rounded): 357- 100% 1672- 46% 1616- 43% 1574- 43% 299- 42% 2546- 41% 1693- 39% 1021- 35% 335- 33% 1563- 33%

Did they churn?

```
final_predictions %>%
  select(ID, new_churn_predict, `Churn (1 = Yes, 0 = No)`) %>%
  arrange(desc(new_churn_predict)) %>%
  slice_head(n=10)
```

```
## # A tibble: 10 × 3
##    ID    new_churn_predict `Churn (1 = Yes, 0 = No)`
##    <fct>             <dbl>                     <dbl>
##  1 357               1.00                          1
##  2 1672              0.464                         1
##  3 1616              0.434                         0
##  4 1574              0.427                         0
##  5 299               0.422                         1
##  6 2546              0.407                         0
##  7 1693              0.391                         0
##  8 1021              0.350                         1
##  9 335               0.335                         1
## 10 1563              0.334                         1
```

357- Yes 1672- Yes 1616- No 1574- No 299- Yes 2546- No 1693- No 1021- Yes 335- Yes 1563- Yes

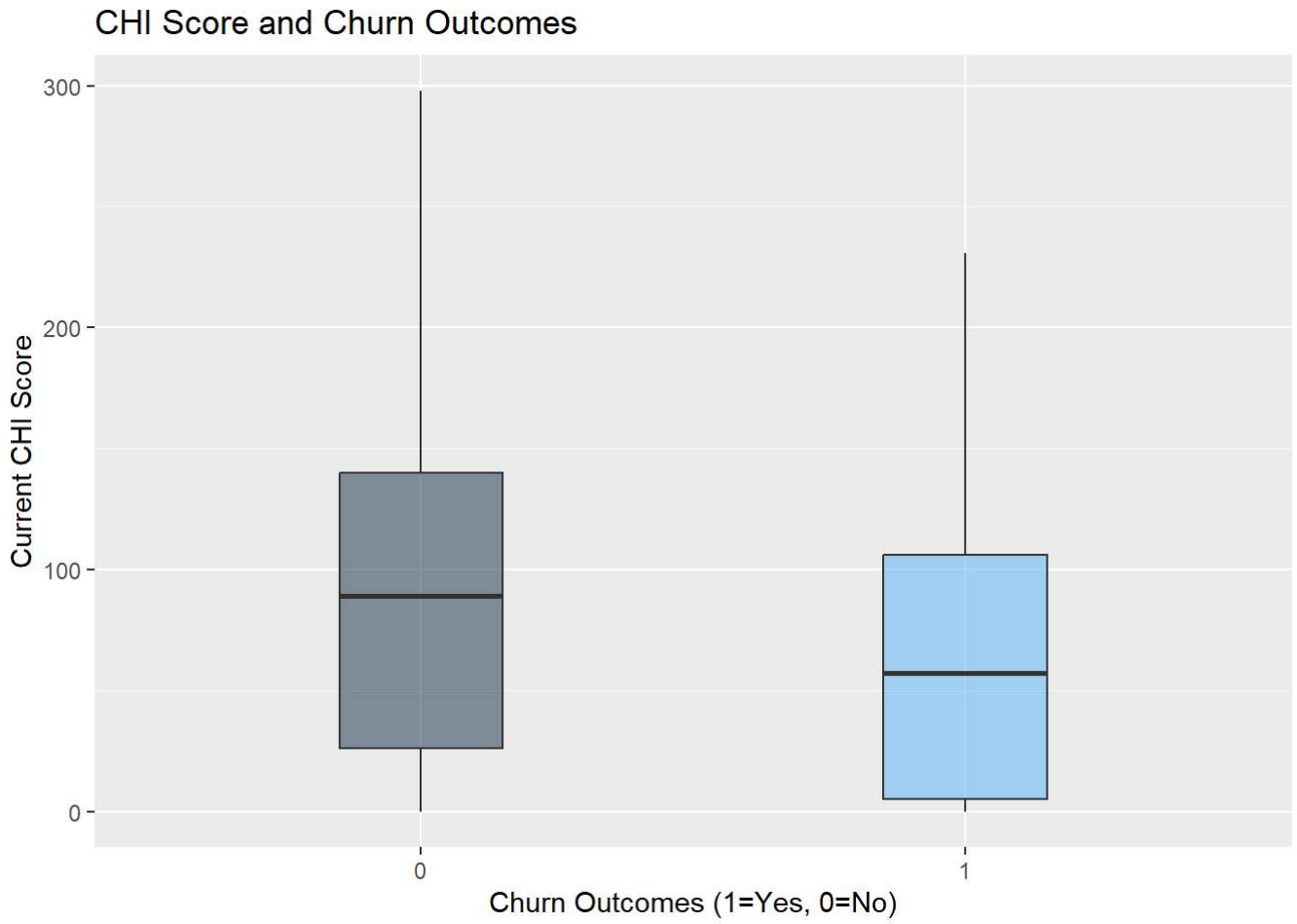
Why did you select your solution?

I chose the original model ran against the subset data as it seemed to perform the best against the four sample customers. It also had the best AIC and coefficients upon review. This model may have performed best as it included all features (except ID and any I manipulatively created through analysis). Since no features seemed to have any strong correlation with Churn, the use of all seemed to strengthen the model's performance.

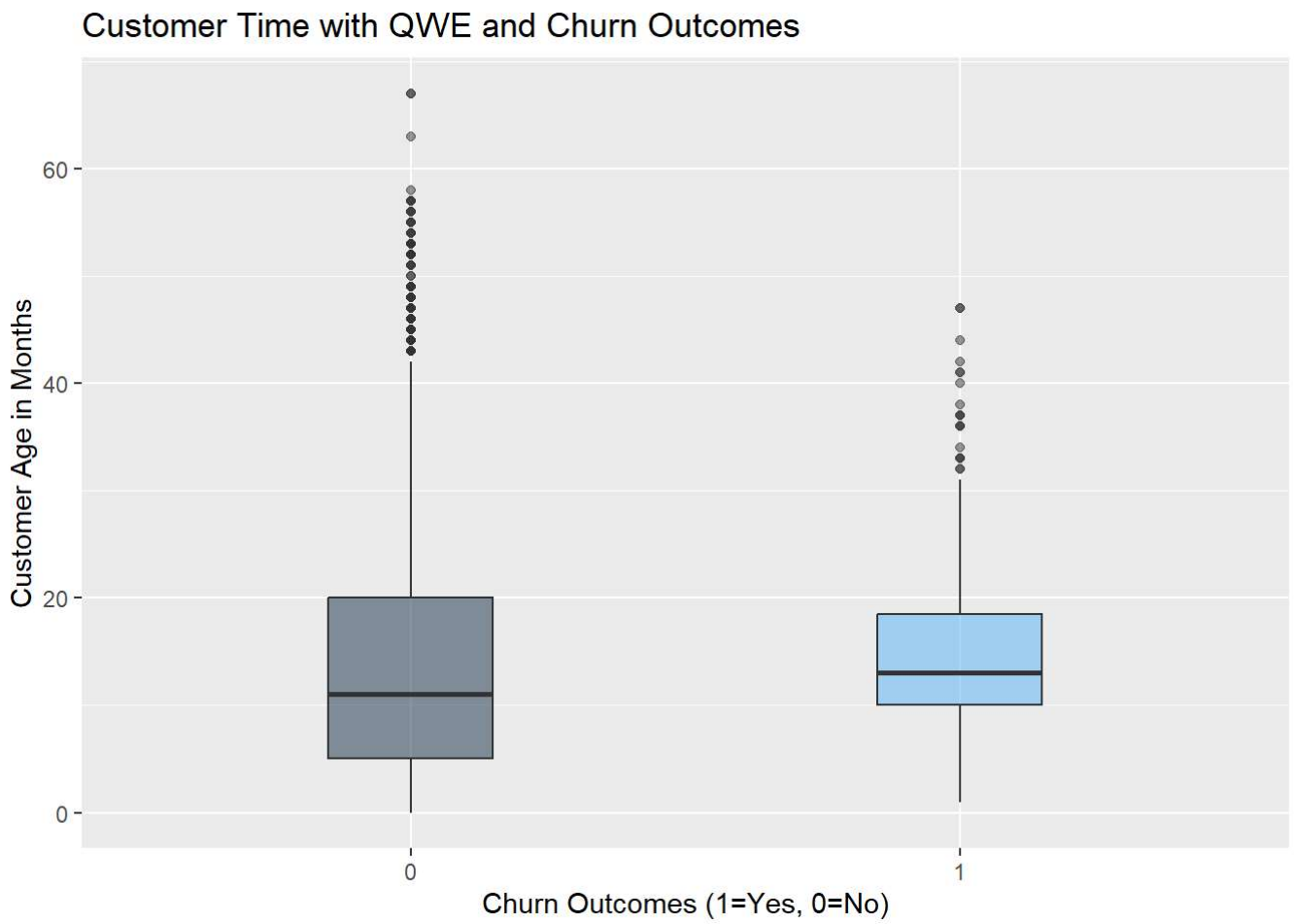
Extras

Visualization

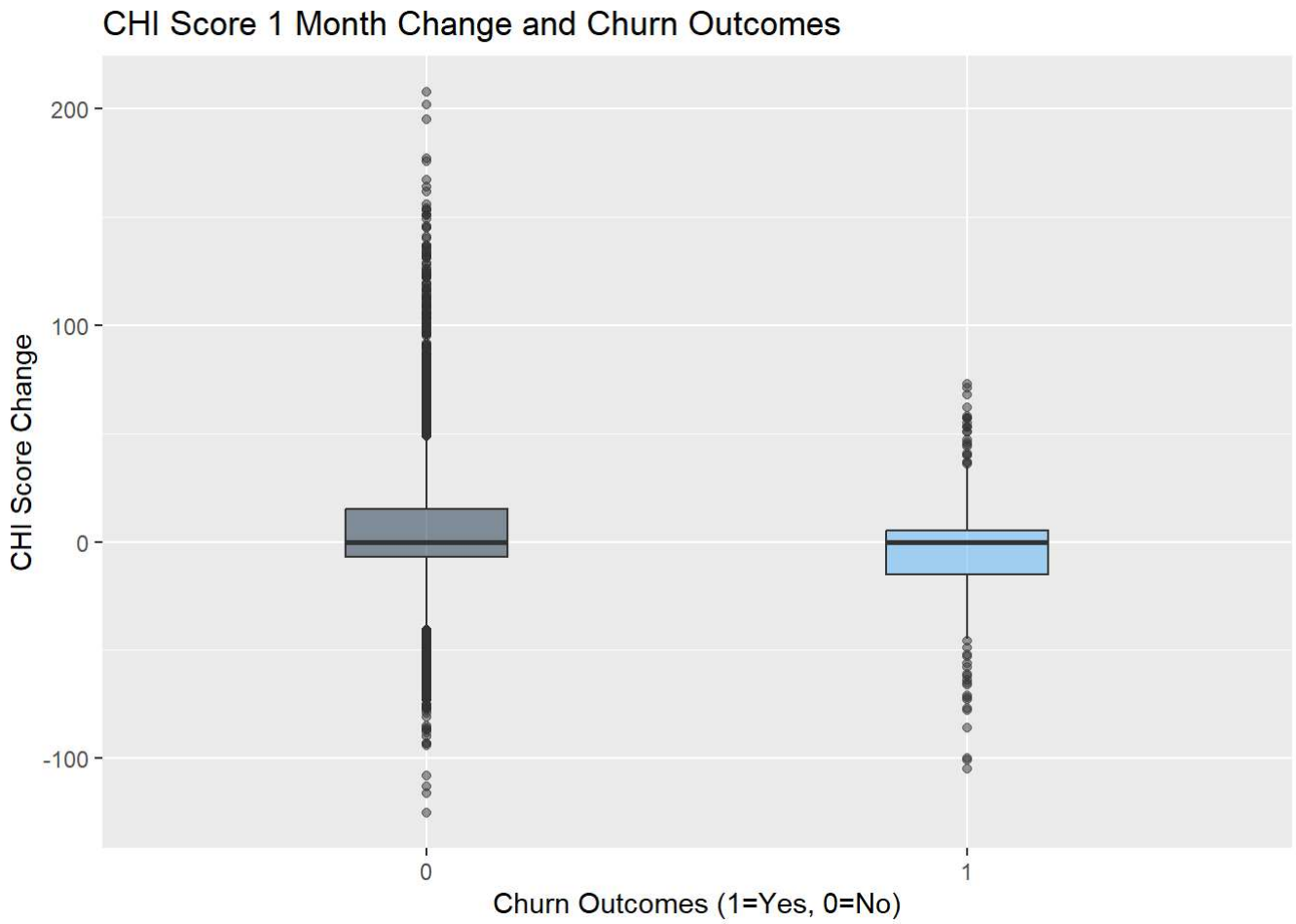
```
qwe_data %>%
  ggplot(aes(x=as.factor(`Churn (1 = Yes, 0 = No)`),y=`CHI Score Month 0`,fill=`Churn (1 = Yes, 0 = No)`))+
  geom_boxplot(alpha=.5, width=.3, position="identity", show.legend = FALSE)+
  labs(title = "CHI Score and Churn Outcomes",x="Churn Outcomes (1=Yes, 0=No)",y = "Current CHI Score")
```



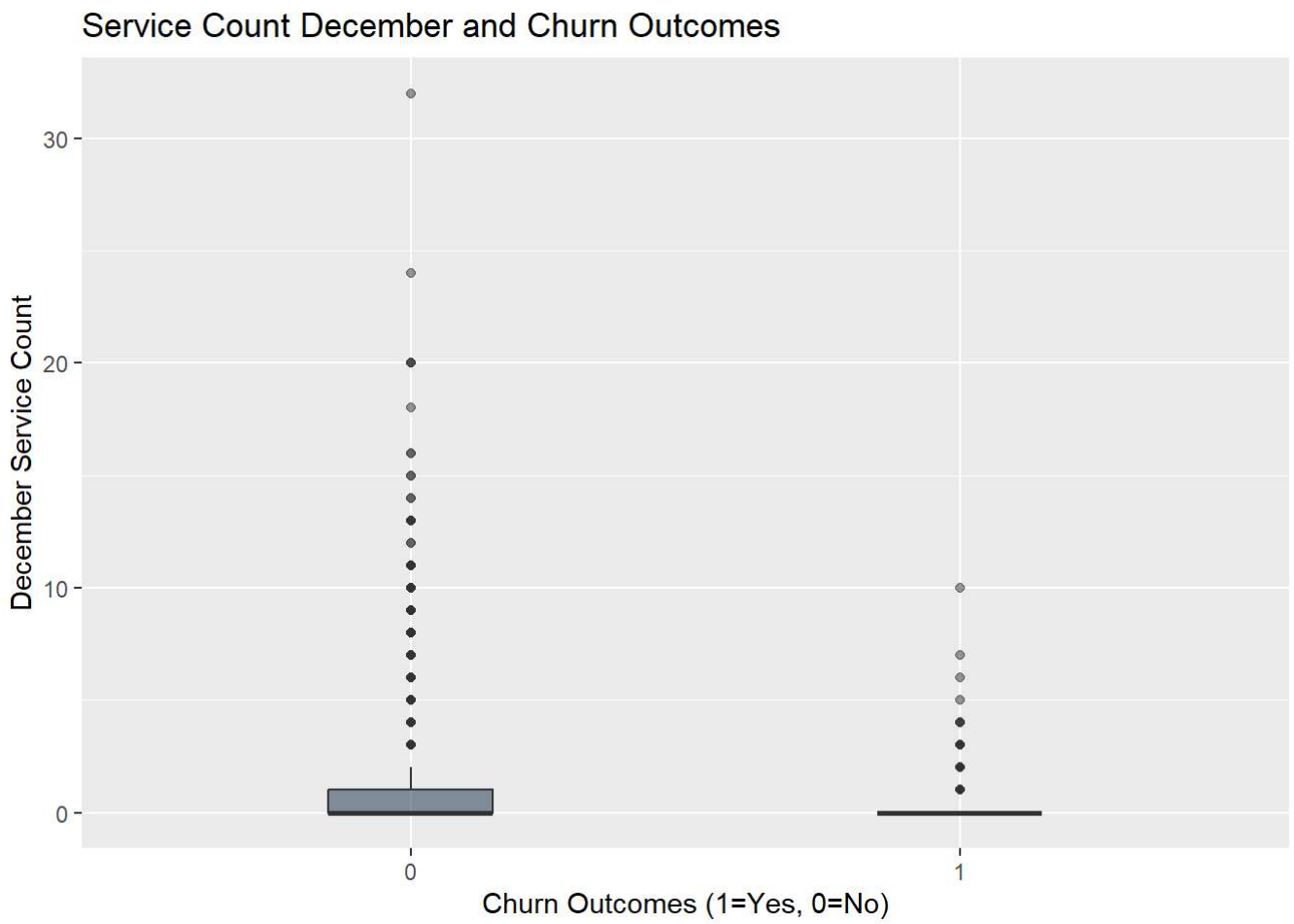

```
qwe_data %>%
  ggplot(aes(x=as.factor(`Churn (1 = Yes, 0 = No)`),y=`Customer Age (in months)`,fill=`Churn (1 = Yes, 0 = No)`))+
  geom_boxplot(alpha=.5, width=.3, position="identity", show.legend = FALSE)+
  labs(title = "Customer Time with QWE and Churn Outcomes",x="Churn Outcomes (1=Yes, 0=No)",y = "Customer Age in Months")
```



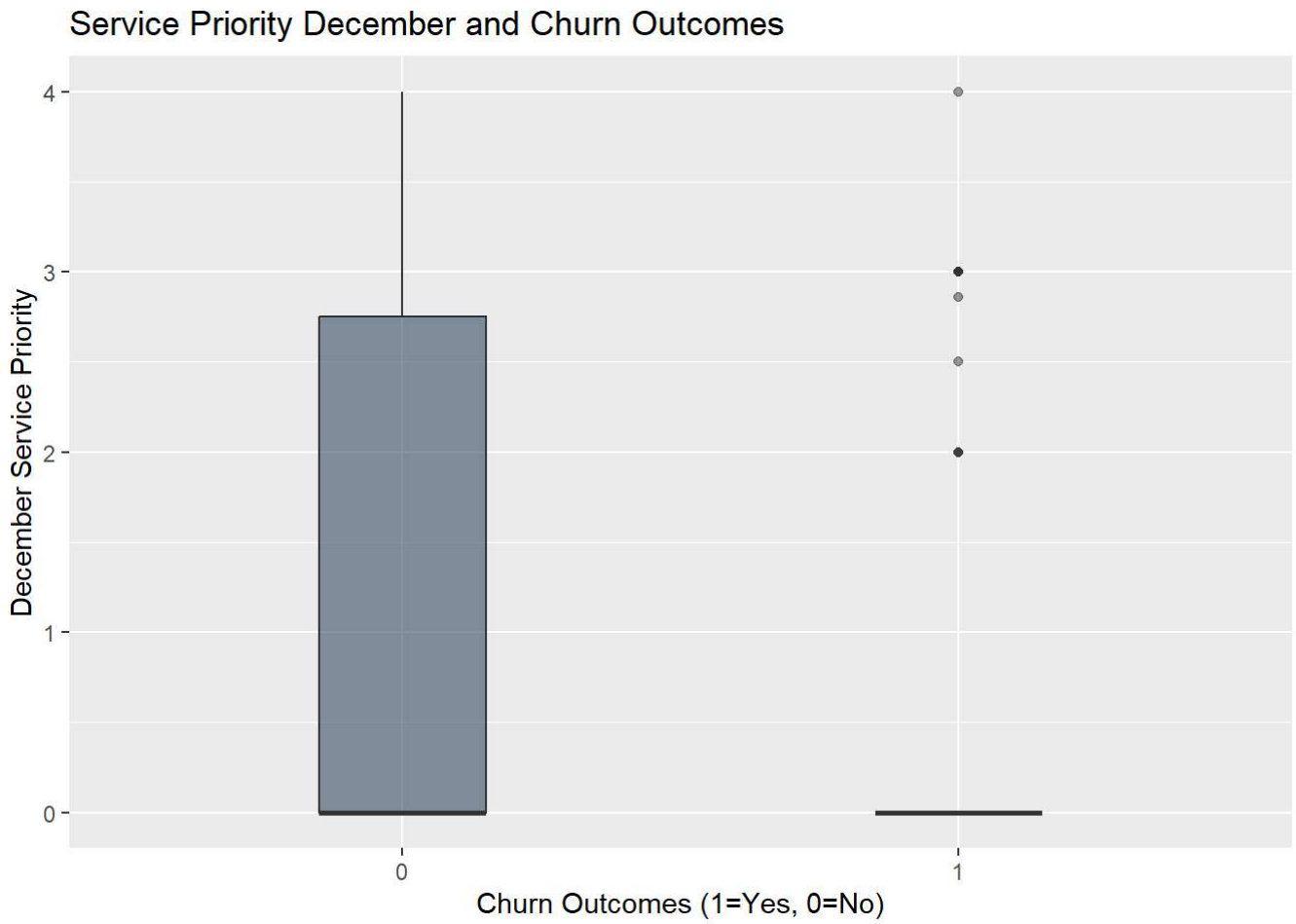

```
qwe_data %>%
  ggplot(aes(x=as.factor(`Churn (1 = Yes, 0 = No)`),y=`CHI Score 0-1`,fill=`Churn (1 = Yes, 0 = No)`))+
  geom_boxplot(alpha=.5, width=.3, position="identity", show.legend = FALSE)+
  labs(title = "CHI Score 1 Month Change and Churn Outcomes",x="Churn Outcomes (1=Yes, 0=No)",y = "CHI Score Change")
```



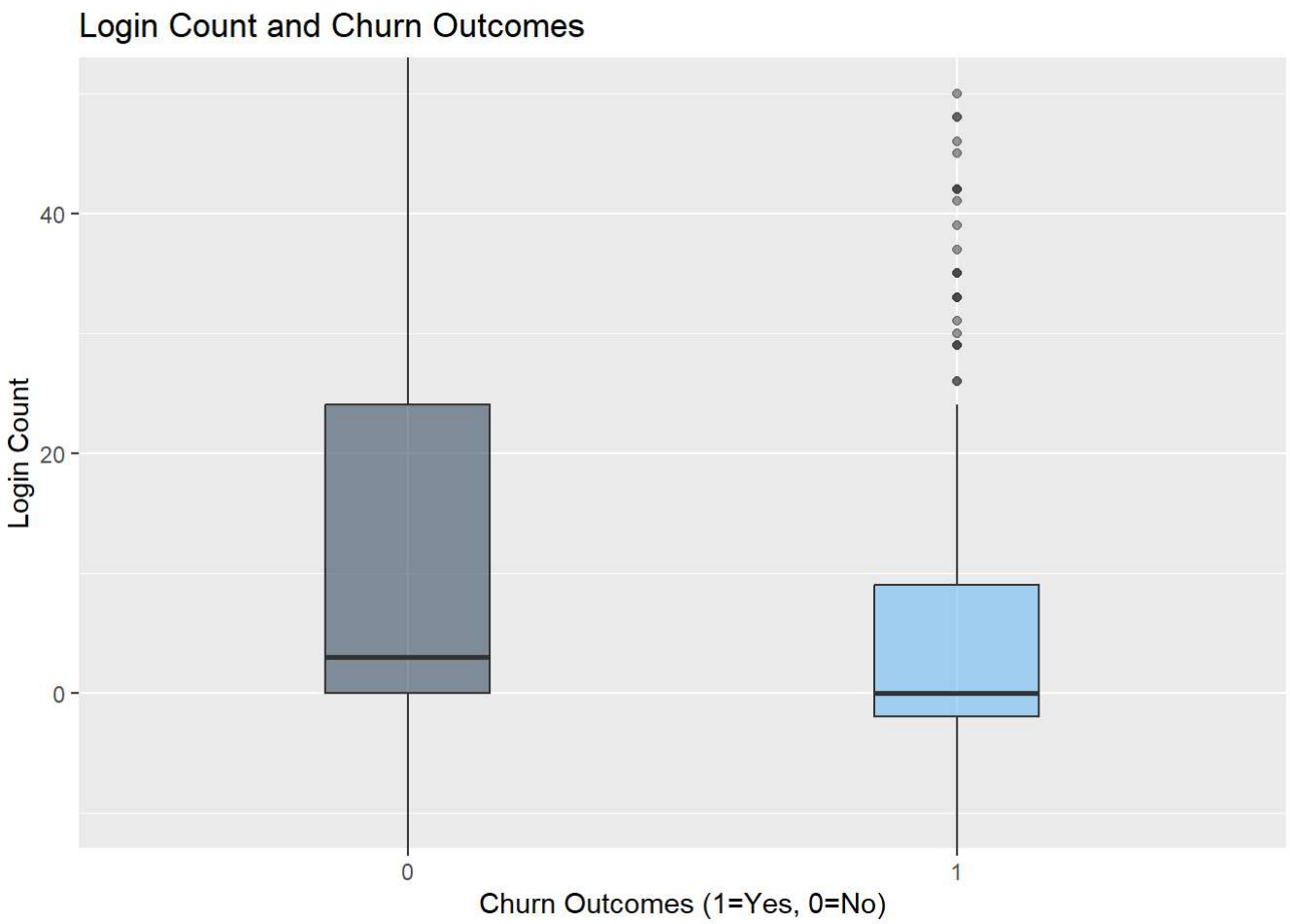

```
qwe_data %>%
  ggplot(aes(x=as.factor(`Churn (1 = Yes, 0 = No)`),y=`Support Cases Month 0`,fill=`Churn (1 = Y
es, 0 = No)`))+
  geom_boxplot(alpha=.5, width=.3, position="identity", show.legend = FALSE)+
  labs(title = "Service Count December and Churn Outcomes",x="Churn Outcomes (1=Yes, 0=No)",y =
"December Service Count")
```



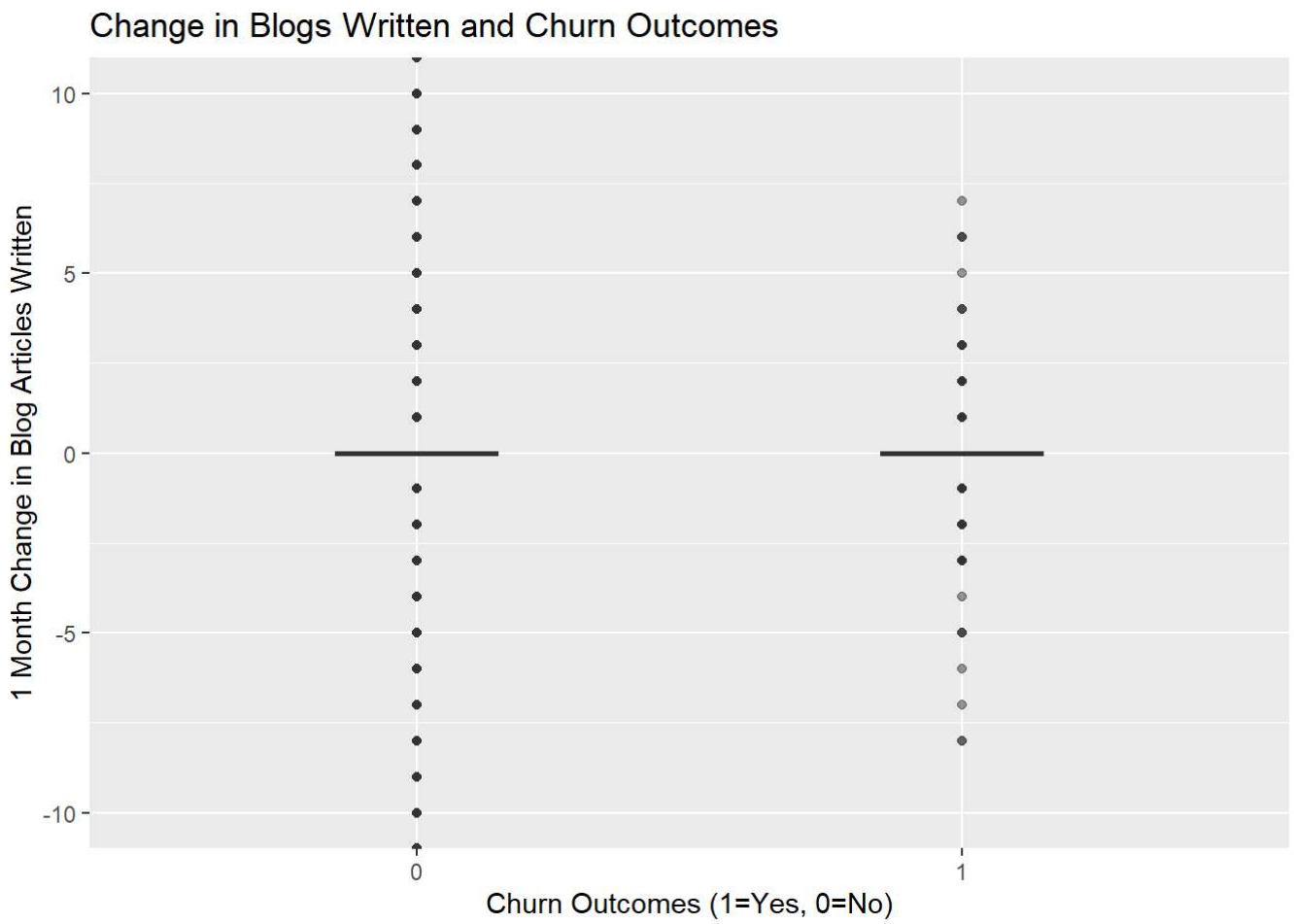

```
qwe_data %>%
  ggplot(aes(x=as.factor(`Churn (1 = Yes, 0 = No)`),y=`SP Month 0`,fill=`Churn (1 = Yes, 0 = No)
`))+
  geom_boxplot(alpha=.5, width=.3, position="identity", show.legend = FALSE)+
  labs(title = "Service Priority December and Churn Outcomes",x="Churn Outcomes (1=Yes, 0=No)",y
= "December Service Priority")
```



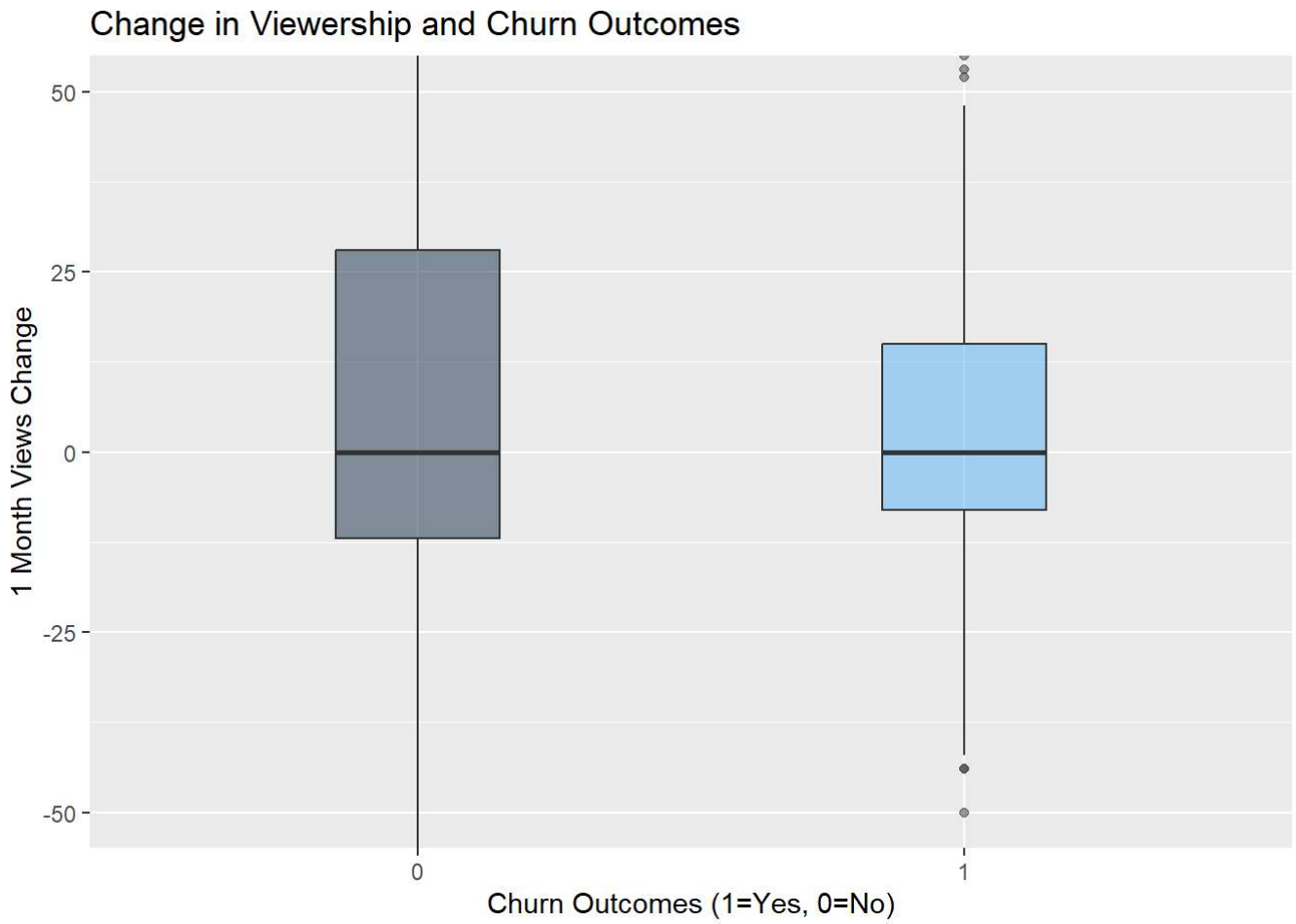

```
qwe_data %>%
  ggplot(aes(x=as.factor(`Churn (1 = Yes, 0 = No)`),y=`Logins 0-1`,fill=`Churn (1 = Yes, 0 = No)
`))+
  geom_boxplot(alpha=.5, width=.3, position="identity", show.legend = FALSE)+
  coord_cartesian(ylim=c(-10,50))+
  labs(title = "Login Count and Churn Outcomes",x="Churn Outcomes (1=Yes, 0=No)",y = "Login Coun
t")
```




```
qwe_data %>%
  ggplot(aes(x=as.factor(`Churn (1 = Yes, 0 = No)`),y=`Blog Articles 0-1`,fill=`Churn (1 = Yes,
0 = No)`))+
  geom_boxplot(alpha=.5, width=.3, position="identity", show.legend = FALSE)+
  coord_cartesian(ylim=c(-10,10))+
  labs(title = "Change in Blogs Written and Churn Outcomes",x="Churn Outcomes (1=Yes, 0=No)",y =
"1 Month Change in Blog Articles Written")
```




```
qwe_data %>%
  ggplot(aes(x=as.factor(`Churn (1 = Yes, 0 = No)`),y=`Views 0-1`,fill=`Churn (1 = Yes, 0 = No)
`))+
  geom_boxplot(alpha=.5, width=.3, position="identity", show.legend = FALSE)+
  coord_cartesian(ylim=c(-50,50))+
  labs(title = "Change in Viewership and Churn Outcomes",x="Churn Outcomes (1=Yes, 0=No)",y = "1
Month Views Change")
```


Change in Viewership and Churn Outcomes
50
25
0
-25
-50
1 Month Views Change
0
1
Churn Outcomes (1=Yes, 0=No)