Assignment Three

1. **How many rows/observations exist in the dataset?**

   ```
   print(nrow(A3Dataset))
   ```

   Ans: 39644

2. **For a classification problem, two of the variables in the dataset will not be relevant. Create a *new* dataframe that has the url and timedelta variables excluded.**

   ```
   df<- subset(A3Dataset , select=-c(url,timedelta))
   ```

3. **Even though the dataset in question does not contain a binary response variable, we will create one for our classification problem. Use the median() function on the shares variable to calculate the median of thee number of times the article has been shared.**

   ```
   median(df$shares)
   ```

   Ans: 1400

4. **Now, create a binary response variable that will take a *numerical* value of 0 if shares <= median value, or 1 if it is greater. Append this newly created variable to the modified dataframe you created in Question 2**

   ```
   df$binaryshare <- ifelse(df$shares <= 1400, 0, 1)
   ```

5. **Run a logistic regression using the glm() function as was done in class. Be sure to exclude from your formula the shares variable, as it will be redundant to have shares as an independent variable while the corresponding encoded variable is the response variable. Also be sure to split your dataset into a training and test set, and run the logistic regression on this training set.**

   ```
   set.seed(1212)

   inds <- sample(1:nrow(df), 0.80*nrow(df))
   ```

```
trdf <- df[inds,]

tedf <- df[-inds,]

logreg1<- glm(binaryshare~.-shares, data = trdf, family = binomial)

summary(logreg1)

Call:
glm(formula = binaryshare ~ . - shares, family = binomial, data = trdf)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.5948  -1.0361  -0.6401   1.0751   2.1586

Coefficients: (2 not defined because of singularities)
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                6.990e+02  1.252e+03   0.558 0.576627
n_tokens_title             7.195e-03  5.936e-03   1.212 0.225485
n_tokens_content           1.570e-04  4.809e-05   3.264 0.001099 **
n_unique_tokens           -1.681e-02  3.999e-01  -0.042 0.966468
n_non_stop_words          -2.812e-01  1.205e+00  -0.233 0.815536
n_non_stop_unique_tokens  -5.915e-01  3.393e-01  -1.743 0.081336 .
num_hrefs                  7.764e-03  1.467e-03   5.293 1.21e-07 ***
num_self_hrefs            -2.155e-02  3.803e-03  -5.666 1.46e-08 ***
num_imgs                   1.161e-03  1.880e-03   0.618 0.536725
num_videos                 5.045e-05  3.275e-03   0.015 0.987711
average_token_length      -1.265e-01  5.030e-02  -2.514 0.011931 *
num_keywords               4.186e-02  7.747e-03   5.404 6.53e-08 ***
data_channel_is_lifestyle -1.823e-01  8.225e-02  -2.216 0.026698 *
data_channel_is_entertainment -3.374e-01  5.228e-02  -6.454 1.09e-10 ***
data_channel_is_bus       -3.279e-01  7.976e-02  -4.111 3.93e-05 ***
data_channel_is_socmed     7.229e-01  7.932e-02   9.114  < 2e-16 ***
data_channel_is_tech       3.926e-01  7.693e-02   5.103 3.33e-07 ***
data_channel_is_world     -6.407e-02  7.772e-02  -0.824 0.409681
kw_min_min                 1.684e-03  3.450e-04   4.881 1.06e-06 ***
kw_max_min                 1.797e-05  1.333e-05   1.348 0.177678
kw_avg_min                -1.567e-04  8.274e-05  -1.894 0.058278 .
kw_min_max                -5.715e-07  2.406e-07  -2.375 0.017536 *
kw_max_max                -3.368e-07  1.218e-07  -2.765 0.005687 **
kw_avg_max                -5.929e-07  1.729e-07  -3.429 0.000605 ***
kw_min_avg                -7.951e-05  1.613e-05  -4.931 8.19e-07 ***
kw_max_avg                -8.395e-05  5.519e-06 -15.212  < 2e-16 ***
kw_avg_avg                 6.826e-04  3.200e-05  21.330  < 2e-16 ***
```

```
self_reference_min_shares      2.747e-06  1.932e-06   1.422 0.155033
self_reference_max_shares      1.156e-06  1.001e-06   1.155 0.248240
self_reference_avg_sharess     7.474e-07  2.511e-06   0.298 0.765945
weekday_is_monday             -6.390e-01  5.529e-02 -11.557  < 2e-16 ***
weekday_is_tuesday            -7.907e-01  5.454e-02 -14.496  < 2e-16 ***
weekday_is_wednesday          -7.768e-01  5.450e-02 -14.254  < 2e-16 ***
weekday_is_thursday           -7.243e-01  5.459e-02 -13.269  < 2e-16 ***
weekday_is_friday             -5.533e-01  5.642e-02  -9.807  < 2e-16 ***
weekday_is_saturday            2.312e-01  6.961e-02   3.322 0.000895 ***
weekday_is_sunday                 NA         NA      NA       NA
is_weekend                        NA         NA      NA       NA
LDA_00                        -6.993e+02  1.252e+03  -0.559 0.576456
LDA_01                        -7.005e+02  1.252e+03  -0.559 0.575837
LDA_02                        -7.006e+02  1.252e+03  -0.560 0.575751
LDA_03                        -7.004e+02  1.252e+03  -0.559 0.575855
LDA_04                        -7.000e+02  1.252e+03  -0.559 0.576119
global_subjectivity            9.028e-01  1.769e-01   5.105 3.32e-07 ***
global_sentiment_polarity      2.243e-01  3.439e-01   0.652 0.514228
global_rate_positive_words    -3.301e+00  1.485e+00  -2.223 0.026189 *
global_rate_negative_words     3.404e+00  2.851e+00   1.194 0.232428
rate_positive_words            1.080e+00  1.177e+00   0.918 0.358686
rate_negative_words            7.171e-01  1.187e+00   0.604 0.545747
avg_positive_polarity         -5.237e-01  2.838e-01  -1.845 0.065000 .
min_positive_polarity         -2.867e-01  2.387e-01  -1.201 0.229703
max_positive_polarity          3.570e-03  8.949e-02   0.040 0.968175
avg_negative_polarity         -3.906e-01  2.627e-01  -1.487 0.137083
min_negative_polarity          8.517e-02  9.593e-02   0.888 0.374623
max_negative_polarity          2.913e-01  2.180e-01   1.336 0.181406
title_subjectivity             2.124e-01  5.706e-02   3.723 0.000197 ***
title_sentiment_polarity       2.550e-01  5.194e-02   4.910 9.11e-07 ***
abs_title_subjectivity         3.885e-01  7.569e-02   5.132 2.86e-07 ***
abs_title_sentiment_polarity  -9.379e-02  8.231e-02  -1.140 0.254488
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43959  on 31714  degrees of freedom
Residual deviance: 39907  on 31658  degrees of freedom, kw, data_channel
AIC: 40021

Number of Fisher Scoring iterations: 8
```

6. **Which predictor variables are most relevant with respect to their influence on the response variable?**

   Ans :title_sentiment_polarity, global_subjectivity , weekdays , abs_title_subjectivity title_subjectivity , num_hrefs and num_self_hrefs : These predictor variables are most releveant based on the result of the logistic regression. All have p-value< 0.01

7. **Calculate the classification accuracy of the model by predicting against the test set.**

```
preds <- predict(logreg1, newdata = tedf,  type = "response")
preds_encoded <- ifelse(preds < 0.5, 0, 1)
sum(preds_encoded == tedf$binaryshare)/nrow(tedf)
```

   Ans: Prediction Accuracy: 0.6604868= 66.04%

8. **Create a confusion matrix as we did in class. What can you say about the**

   predictive power of the logistic regression model?

```
table(preds_encoded, tedf$binaryshare)
```

```
preds_encoded   0       1
        0          2756   1458
        1          1234   2481
```

   Ans: based on the confusion matrix 2756 true negative and 2481 true positive. Moreover 1438 false negative and 1234 false positive. Based on the results of the confusion matrix our model has low predictive power.