Instagram Reach Analysis

Predictive and Inferential Insights

Ogechikanma Chelsea Anosike

Master of Science in Data Analytics

Binghamton University

December 14, 2024

OBJECTIVE

The objective of this project is to analyze Instagram reach data to determine factors that significantly influence the number of impressions a post receives. The analysis aims to build a predictive model for impressions and provide insights for content optimization strategies.

#SIGNIFICANCE Optimizing content strategies can directly impact brand visibility, audience retention, and conversion rates. Insights from this analysis enable informed decision-making for targeted marketing and improved ROI on social media campaigns.

DATA DESCRIPTION

The dataset consists of Instagram post performance metrics collected over a defined period, including:

- Response Variable: Impressions (total views).
- Predictors: Engagement metrics (likes, shares, saves), Comments and profile visits

PREPROCESSING STEPS

Standardized column names for clarity (e.g., "From.Home" → "Home"). Verified and handled missing data (none identified). Created an Engagement Score composite metric to consolidate multiple engagement variables.

METHODOLOGY

PREDICTIVE MODELING APPROACH

A multiple linear regression model was selected due to its interpretability and ability to quantify the contribution of each predictor.

Model Design

- Response Variable: Impressions.
- Predictors: Engagement Score and Profile Visits.
- Rationale: Engagement metrics are strong indicators of user interest, while profile visits reflect broader reach dynamics.

EVALUATION METRICS

- Adjusted R^2: Measures explained variance while accounting for model complexity.
- Coefficient significance: Assessed via p-values (<0.05 deemed significant).
- Residual diagnostics: Ensures model assumptions are met.

STEPS AND ANALYSIS

STEP 1: LOAD THE DATASET

The dataset was loaded into R using the read.csv() function. Initial checks on structure and data types were conducted to understand the dataset's composition

```
data <- read.csv("~/Downloads/instagram_data.csv")
head(data)
```

```
##   Impressions From.Home From.Hashtags From.Explore From.Other Saves Comments
## 1        3920      2586          1028          619         56    98        9
## 2        5394      2727          1838         1174         78   194        7
## 3        4021      2085          1188            0        533    41       11
## 4        4528      2700           621          932         73   172       10
## 5        2518      1704           255          279         37    96        5
## 6        3884      2046          1214          329         43    74        7
##   Shares Likes Profile.Visits Follows
## 1      5   162             35       2
## 2     14   224             48      10
## 3      1   131             62      12
## 4      7   213             23       8
## 5      4   123              8       0
## 6     10   144              9       2
```

```
##
## 1
## 2                    Here are some of the best data science project ideas on heal
## 3
## 4 Here\x92s how you can write a Python program to detect whether a sentence is a
## 5                              Plotting annotations while visualizing your data is
## 6
##
## 1
## 2
## 3
## 4
## 5 #datavisualization\xa0#datascience\xa0#data\xa0#dataanalytics\xa0#machinelearn
## 6
```

DATA CLEANING

```r
colnames(data) <- gsub("From\\.", "", colnames(data))

# Verify the column names
colnames(data)
```

```
##  [1] "Impressions"    "Home"           "Hashtags"       "Explore"
##  [5] "Other"          "Saves"          "Comments"       "Shares"
##  [9] "Likes"          "Profile.Visits" "Follows"        "Caption"
## [13] "Hashtags"
```

```r
head(data)
```

```
##   Impressions Home Hashtags Explore Other Saves Comments Shares Likes
## 1        3920 2586     1028     619    56    98        9      5   162
## 2        5394 2727     1838    1174    78   194        7     14   224
## 3        4021 2085     1188       0   533    41       11      1   131
## 4        4528 2700      621     932    73   172       10      7   213
## 5        2518 1704      255     279    37    96        5      4   123
## 6        3884 2046     1214     329    43    74        7     10   144
##   Profile.Visits Follows
## 1             35       2
```

```
## 2              48      10
## 3              62      12
## 4              23       8
## 5               8       0
## 6               9       2
##
## 1
## 2                  Here are some of the best data science project ideas on heal
## 3
## 4 Here\x92s how you can write a Python program to detect whether a sentence is a
## 5                            Plotting annotations while visualizing your data is
## 6
##
## 1
## 2
## 3
## 4
## 5 #datavisualization\xa0#datascience\xa0#data\xa0#dataanalytics\xa0#machinelearn
## 6
```

```
colSums(is.na(data))
```

```
##     Impressions            Home        Hashtags         Explore           Other
##               0               0               0               0               0
##           Saves        Comments          Shares           Likes Profile.Visits
##               0               0               0               0               0
##         Follows         Caption        Hashtags
##               0               0               0
```

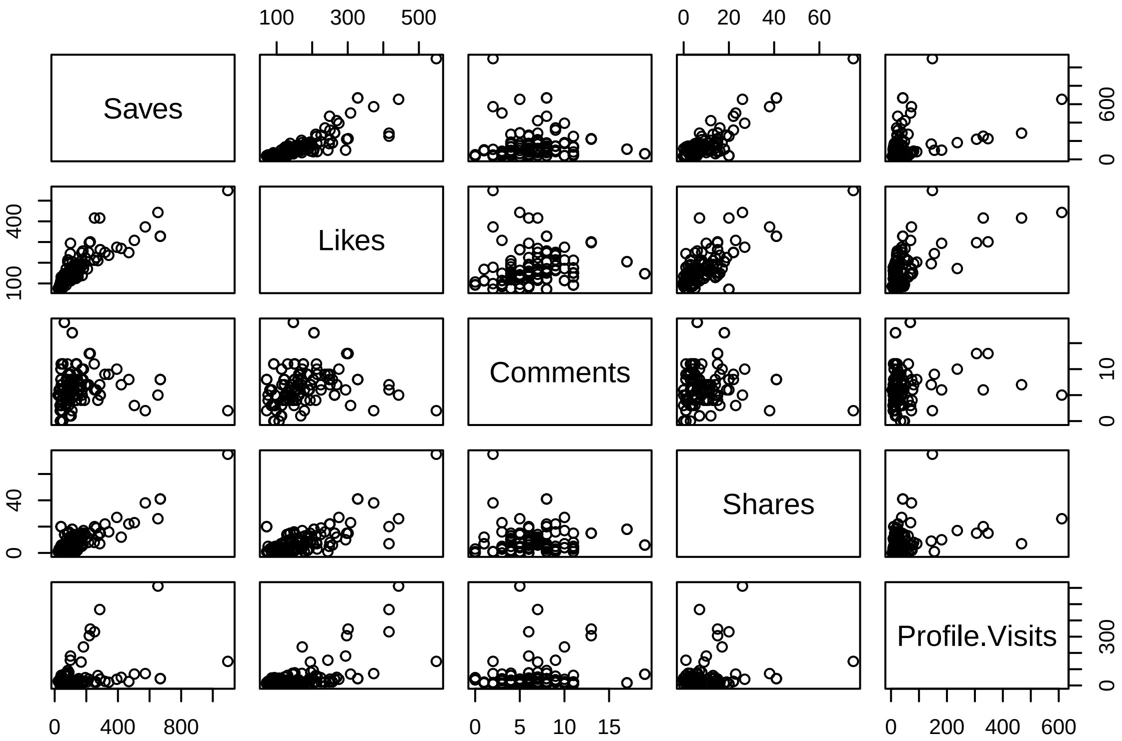
STEP 2: EXPLORATORY DATA ANALYSIS

```
response <- "Impressions"
predictors <- c("Saves","Likes","Comments","Shares","Profile.Visits")
```

DATA VISUALIZATION/CLEANING

Scatterplots and correlation matrices were used to explore relationships between the response variable (Impressions) and predictors.

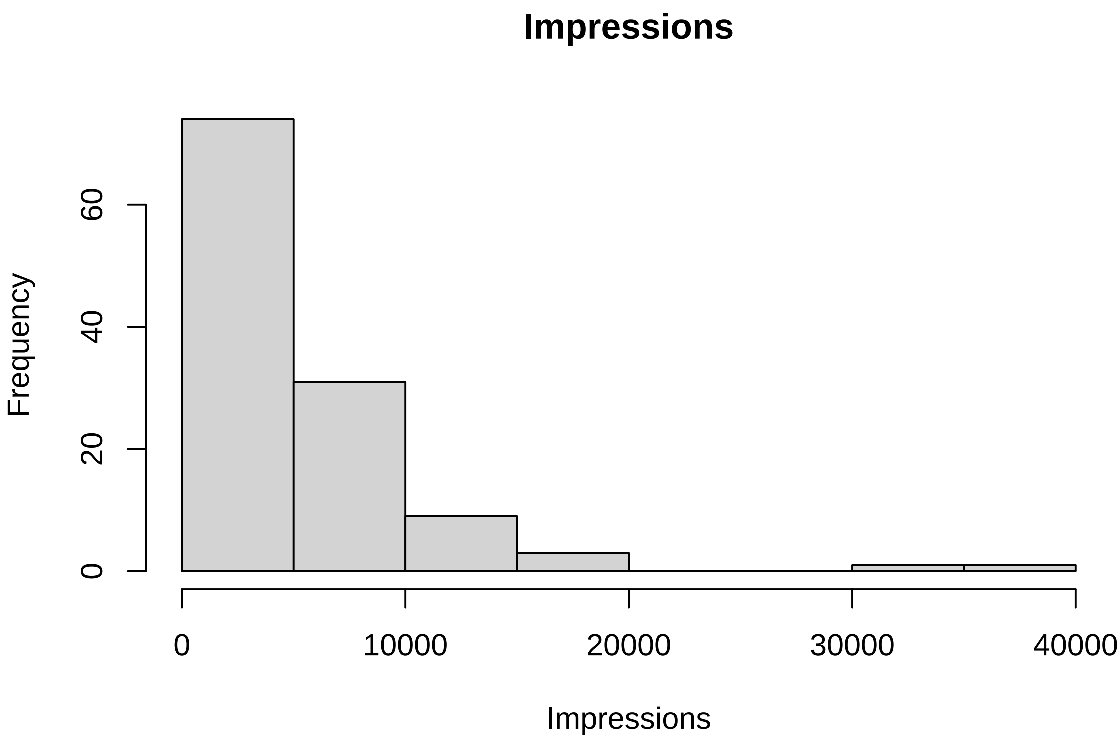
```r
#Scatterplot and Correlation Matrix
pairs(data[, predictors])
```




```r
cor(data[, predictors])
```

```
##                      Saves     Likes    Comments     Shares Profile.Visits
## Saves           1.00000000 0.8456433 -0.02691226 0.86032419     0.36062821
## Likes           0.84564329 1.0000000  0.12358610 0.70779400     0.62610703
## Comments       -0.02691226 0.1235861  1.00000000 0.01693253     0.09671424
## Shares          0.86032419 0.7077940  0.01693253 1.00000000     0.24536082
## Profile.Visits  0.36062821 0.6261070  0.09671424 0.24536082     1.00000000
```

```r
#Response variable histogram
hist(data$Impressions, main = "Impressions", xlab = "Impressions")
```



Key Findings:

- Strong positive correlations were observed between Likes, Shares, and Impressions.
- Comments and Profile Visits also showed moderate correlations with Impressions.
- The distribution of Impressions is slightly right-skewed

STEP 3: REGRESSION MODELING

CREATING COMPOSITE METRIC

```
# Normalize the variables
data$Saves_scaled <- scale(data$Saves)
data$Likes_scaled <- scale(data$Likes)
data$Shares_scaled <- scale(data$Shares)

# Create the Engagement Score
data$Engagement_Score <- data$Saves_scaled + data$Likes_scaled + data$Shares_scale
```

FIT MULTIPLE REGRESSION MODEL

```
model <- lm(Impressions ~ Engagement_Score + Comments + Profile.Visits, data = dat
summary(model)
```

```
##
## Call:
## lm(formula = Impressions ~ Engagement_Score + Comments + Profile.Visits,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4935.3  -830.1  -139.2   761.2  6769.0
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5190.497    358.113  14.494  < 2e-16 ***
## Engagement_Score 1017.645     65.305  15.583  < 2e-16 ***
## Comments         -139.260     46.496  -2.995  0.00336 **
## Profile.Visits     28.476      2.106  13.521  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1782 on 115 degrees of freedom
## Multiple R-squared:  0.8681, Adjusted R-squared:  0.8647
## F-statistic: 252.3 on 3 and 115 DF,  p-value: < 2.2e-16
```

Coefficients

- Intercept:$Estimate = 5190.497$. This represents the expected Impressions when all predictors (Engagement_Score, Comments, Profile.Visits) are 0.
- Engagement_Score:$Estimate = 1017.645$. For every 1-unit increase in Engagement_Score, Impressions increase by approximately 1018, holding other predictors constant. Highly significant(p<2e−16).
- Comments:$Estimate = -139.260$. For every 1 additional comment, Impressions decrease by approximately 139, holding other predictors constant. This negative relationship is counterintuitive and I will have to investigate further

- Profile.Visits:*Estimate* = 28.476. For every 1 additional profile visit, Impressions increase by approximately 28.5, holding other predictors constant. Highly significant (p<2e−16).

Model Diagnostics

- Residuals:Residuals are the differences between observed and predicted Impressions. The summary shows a range from -4935.3 to 6769.0, indicating variability in prediction errors.
- Significance: All predictors are statistically significant (p<0.01), meaning they contribute meaningfully to explaining "Impressions".
- Goodness of Fit: *R-squared = 0.8681: Approximately 86.81% of the variability in Impressions is explained by the predictors.* Adjusted R-squared = 0.8647: Adjusts for the number of predictors, still very high, indicating a good fit. *F-statistic = 252.3 (p < 2.2e-16):

MODEL WITHOUT COMMENTS VARIABLE

```r
model_no_comments <- lm(Impressions ~ Engagement_Score + Profile.Visits, data=data
summary(model_no_comments)
```

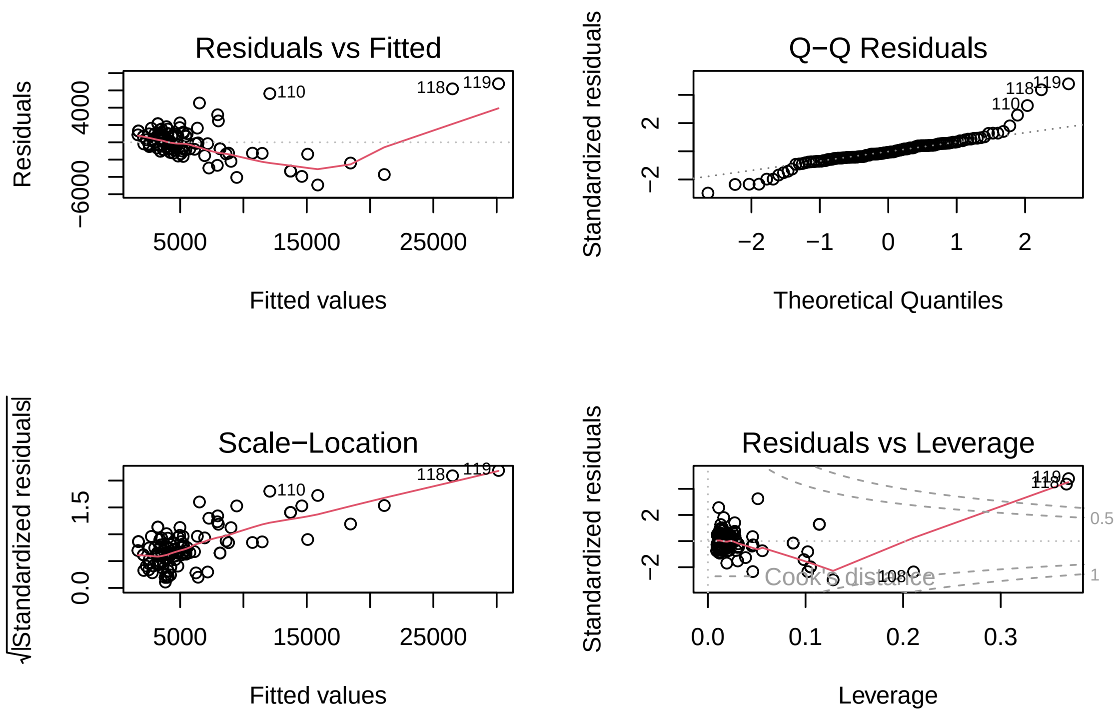
```
##
## Call:
## lm(formula = Impressions ~ Engagement_Score + Profile.Visits,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5654.5  -649.9    27.0   684.7  7307.8
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4290.545    201.417   21.30   <2e-16 ***
## Engagement_Score 1018.079     67.511   15.08   <2e-16 ***
## Profile.Visits     27.922      2.169   12.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1842 on 116 degrees of freedom
## Multiple R-squared:  0.8578, Adjusted R-squared:  0.8554
## F-statistic: 349.9 on 2 and 116 DF,  p-value: < 2.2e-16
```

- The R-squared has reduced by 1% when comments has been removed as a predictor variable; meaning it has some sort of influence in predicting 'Impressions'. This means i will keep Comment as a predictor

RESIDUAL DIAGNOSTICS

```r
par(mfrow=c(2,2))
plot(model)
```



STEP 4:DATA CLEANING (OUTLIERS)

There are 3 outliers consistently appearing in the residuals vs. fitted plot, Q-Q plot, scale-location plot, and residuals vs. leverage plot.

```
outlier_test <- which(abs(rstandard(model)) > 2)
print(outlier_test)
```

Identifying Outliers

```
##  55  68 100 106 108 110 118 119
##  55  68 100 106 108 110 118 119
```

```
data[outlier_test,]
```

```
##     Impressions  Home Hashtags Explore Other Saves Comments Shares Likes
## 55        10667  3152     6564     617   187   219       13     15   297
## 68        10933  3152     6610     623   334   225       13     15   301
## 100        5409  2643     2006    1068   230   393       10     27   275
## 106       11068  2099     2986    5634   122   214        7      8   250
## 108       17396  1817    10008    5192   251   285        7      7   416
## 110       17713  2449     2141   12389   561   504        3     23   308
## 118       32695 11815     3147   17414   170  1095        2     75   549
## 119       36919 13473     4176   16444  2547   653        5     26   443
##     Profile.Visits Follows
## 55             306      74
## 68             347      94
## 100             38      14
## 106             39      34
## 108            467     260
## 110             70      96
## 118            148     214
## 119            611     228
##
## 55
## 68
## 100
## 106
## 108 Here is a list of 100+ Machine Learning Algorithms and Models explained usi
## 110
## 118
## 119
##
```

```
## 55
## 68
## 100                              #data\xa0#datascience\xa0#dataanalysis\xa0#da
## 106                                          #data\xa0#datascience\xa0#dataana
## 108                                            #machinelearning\xa0#machinelea
## 110 #sql\xa0#mysql\xa0#datascience\xa0#datasciencejobs\xa0#datasciencetraining\
## 118                  #datascience\xa0#datasciencejobs\xa0#datasciencetraining\
## 119
##     Saves_scaled Likes_scaled Shares_scaled Engagement_Score
## 55     0.4202279     1.495752     0.5588801         2.474860
## 68     0.4586113     1.544308     0.5588801         2.561800
## 100    1.5333454     1.228694     1.7482701         4.510309
## 106    0.3882418     0.925218    -0.1349308         1.178529
## 108    0.8424449     2.940296    -0.2340466         3.548694
## 110    2.2434376     1.629281     1.3518068         5.224526
## 118    6.0241987     4.554786     6.5058305        17.084815
## 119    3.1966244     3.268050     1.6491543         8.113828
```

```r
cooks_d <- cooks.distance(model)
influential_points <- which(cooks_d > (4 / nrow(data))) #influential points
print(influential_points)
```

Using Cook's distance to identify Oulier Impact

```
##  27  41  50  55  68  91  96 100 108 110 118 119
##  27  41  50  55  68  91  96 100 108 110 118 119
```

All outliers except 106 are influential points. I am going to transform the variables

VARIABLE TRANSFORMATION

```r
data$log_Impressions <- log(data$Impressions)
log_model <- lm(log_Impressions ~ Engagement_Score + Comments + Profile.Visits, da
summary(log_model)
```

```
##
## Call:
## lm(formula = log_Impressions ~ Engagement_Score + Comments +
##     Profile.Visits, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51510 -0.17294 -0.01933  0.16899  0.73243
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.3191079  0.0502762 165.468  < 2e-16 ***
## Engagement_Score 0.1142967  0.0091682  12.467  < 2e-16 ***
## Comments         0.0036957  0.0065277   0.566    0.572
## Profile.Visits   0.0025566  0.0002957   8.646 3.64e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2502 on 115 degrees of freedom
## Multiple R-squared:  0.7795, Adjusted R-squared:  0.7738
## F-statistic: 135.5 on 3 and 115 DF,  p-value: < 2.2e-16
```

- I may have to take out Comments as a predictor in this transformed model because the p value = 0.572 meaning it is not significant

```r
clean_model_no_comments <- lm(log_Impressions ~ Engagement_Score + Profile.Visits,

summary(clean_model_no_comments)
```

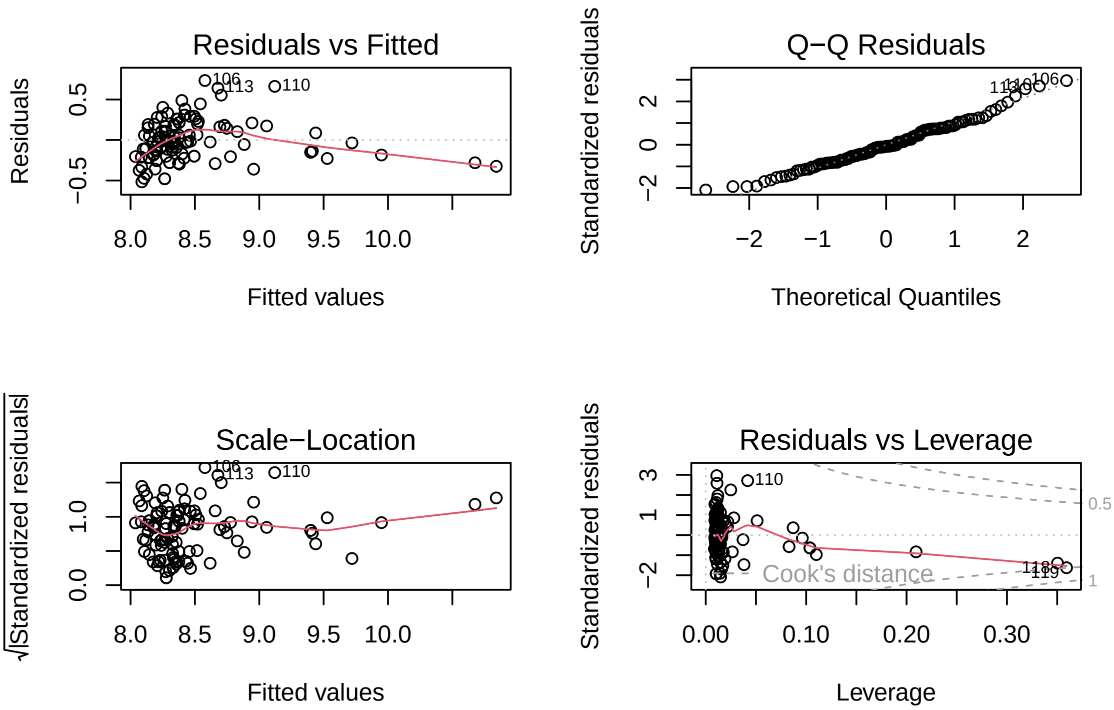
EXCLUDING Comments variable

```
##
## Call:
## lm(formula = log_Impressions ~ Engagement_Score + Profile.Visits,
##     data = data)
##
## Residuals:
```

```
##       Min       1Q   Median       3Q      Max
## -0.51695 -0.17998 -0.01991  0.17885  0.73385
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.3429906  0.0272729 305.907  < 2e-16 ***
## Engagement_Score 0.1142852  0.0091413  12.502  < 2e-16 ***
## Profile.Visits   0.0025713  0.0002937   8.756 1.92e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2494 on 116 degrees of freedom
## Multiple R-squared:  0.7789, Adjusted R-squared:  0.7751
## F-statistic: 204.3 on 2 and 116 DF,  p-value: < 2.2e-16
```

Diagnostic Plots To Test If Better Without 'Comment'

```r
par(mfrow = c(2,2))
plot(clean_model_no_comments)
```



- The plots look better than the diagnostic plots that included comments

STEP 5: INTERACTION AND HYPOTHESIS TESTING

Null Hypothesis (H_0): There is no interaction effect between Engagement_Score and Profile.Visits. $\beta_3 = 0$ Alternative Hypothesis (H_A): There is an interaction effect between Engagement_Score and Profile.Visits. $\beta_3 \neq 0$

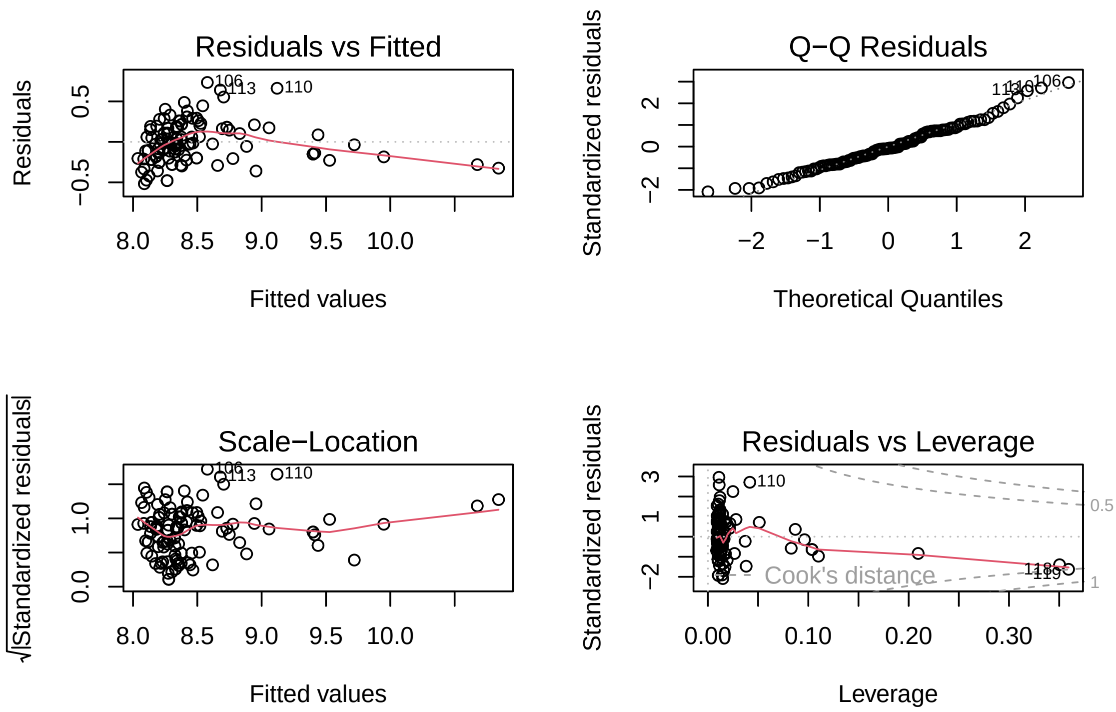
```
data$Interaction <- data$Engagement_Score * data$Profile.Visits

interaction_model <- lm(log_Impressions ~ Engagement_Score * Profile.Visits, data =

summary(interaction_model)
```

```
##
## Call:
## lm(formula = log_Impressions ~ Engagement_Score * Profile.Visits,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44190 -0.15629 -0.02094  0.15800  0.70180
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     8.309e+00  2.703e-02 307.429  < 2e-16 ***
## Engagement_Score                1.383e-01  1.046e-02  13.217  < 2e-16 ***
## Profile.Visits                  3.929e-03  4.362e-04   9.009 5.26e-15 ***
## Engagement_Score:Profile.Visits -3.238e-04  8.048e-05  -4.023 0.000103 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2345 on 115 degrees of freedom
## Multiple R-squared:  0.8062, Adjusted R-squared:  0.8011
## F-statistic: 159.4 on 3 and 115 DF,  p-value: < 2.2e-16
```

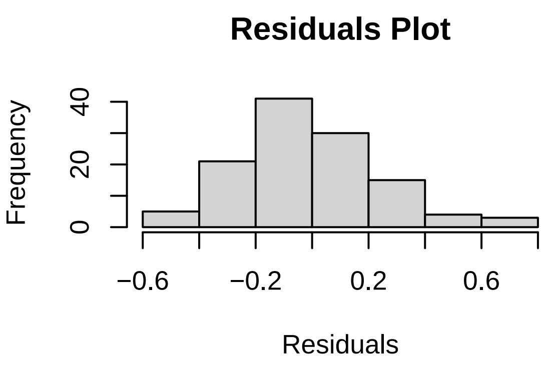
- The interaction term 'Engagement_Score:Profile.Visits' has a p_value of 0.000103 which is less than $\alpha = 0.05$. We conclude that there is a statistically significant interaction between Engagement_Score and Profile.Visits, this means that the relationship between Engagement_Score and log_Impressions changes with different levels of Profile.Visits.

STEP 6: DIAGNOSTICS AND ASSUMPTIONS

```r
par(mfrow = c(2,2))
plot(clean_model_no_comments)
```




```r
residuals <- resid(clean_model_no_comments)
hist(residuals, main = "Residuals Plot", xlab = "Residuals")
```



MODEL PERFORMANCE

```r
summary(clean_model_no_comments)$adj.r.squared
```

```
## [1] 0.7750769
```

An Adjusted R_square value of 0.7751 indicates that approximately 77.5% of the variance in Impression variable is explained by the predictors.

```r
data$Saves_scaled <- as.numeric(scale(data$Saves))
data$Likes_scaled <- as.numeric(scale(data$Likes))
data$Shares_scaled <- as.numeric(scale(data$Shares))

data$Engagement_Score <- data$Saves_scaled + data$Likes_scaled + data$Shares_scale
```

REFITTING MODEL

```r
clean_model_no_comments <- lm(Impressions ~ Engagement_Score + Profile.Visits, dat
```

STEP 7: CONFIDENCE INTERVAL AND PREDICTION

```r
new_data <- data.frame(Saves = 30, Likes = 200, Shares = 15, Profile.Visits = 25)

# Normalize the new data
new_data$Saves_scaled <- (new_data$Saves - mean(data$Saves)) / sd(data$Saves)
new_data$Likes_scaled <-(new_data$Likes - mean(data$Likes)) / sd(data$Likes)
new_data$Shares_scaled <- (new_data$Shares - mean(data$Shares)) / sd(data$Shares)


new_data$Engagement_Score <- as.numeric(new_data$Saves_scaled + new_data$Likes_sca

confint(clean_model_no_comments)
```

```
##                       2.5 %     97.5 %
## (Intercept)       3891.61374 4689.47707
## Engagement_Score  884.36566 1151.79260
## Profile.Visits     23.62603   32.21729
```

```
predict(clean_model_no_comments, newdata = new_data, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 5078.482 4724.564 5432.401
```

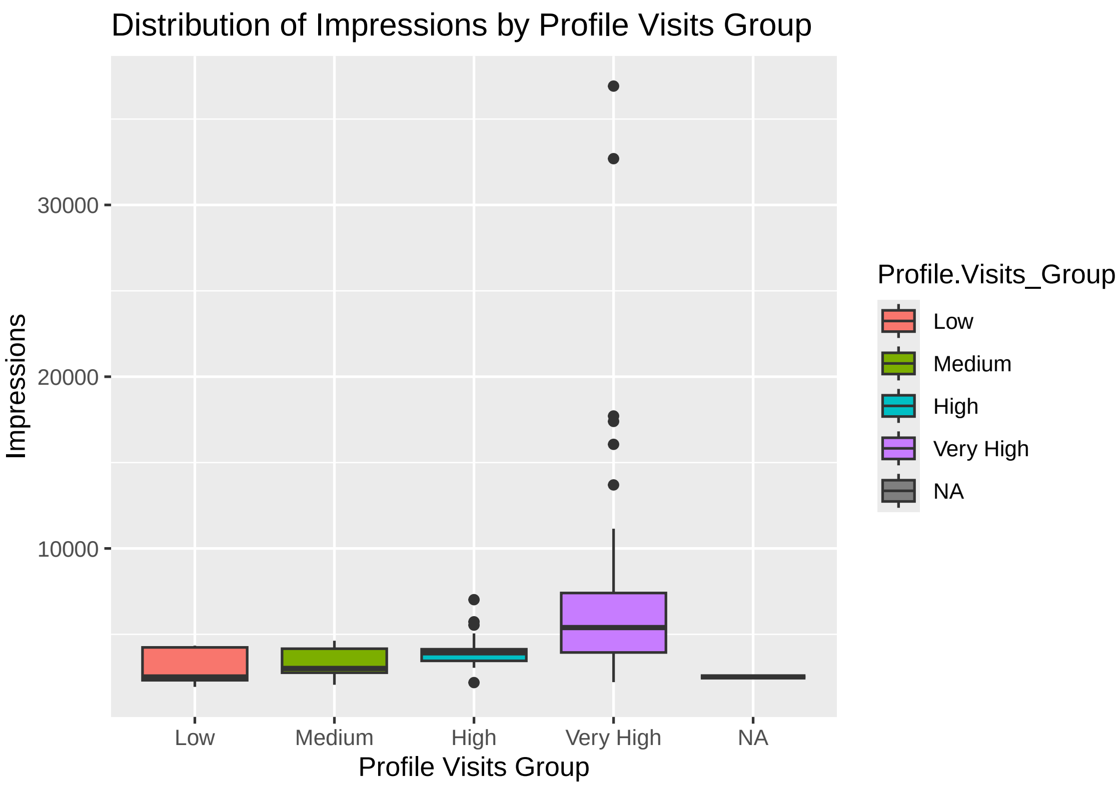
```
names(data) <- make.unique(names(data))
names(data)
```

```
##  [1] "Impressions"      "Home"             "Hashtags"         "Explore"
##  [5] "Other"            "Saves"            "Comments"         "Shares"
##  [9] "Likes"            "Profile.Visits"   "Follows"          "Caption"
## [13] "Hashtags.1"       "Saves_scaled"     "Likes_scaled"     "Shares_scaled"
## [17] "Engagement_Score" "log_Impressions"  "Interaction"
```

STEP 8: VISUALIZATION OF MODEL RESULTS

```
data$Profile.Visits <- as.numeric(as.character(data$Profile.Visits))
data$Profile.Visits_Group <- cut(data$Profile.Visits, breaks = c(4, 8, 12, 20, Inf
                                 labels = c("Low", "Medium", "High", "Very High"))
```

```
library(ggplot2)
ggplot(data, aes(x = Profile.Visits_Group, y = Impressions, fill = Profile.Visits_
  geom_boxplot() +
  labs(
    title = "Distribution of Impressions by Profile Visits Group",
    x = "Profile Visits Group",
    y = "Impressions"
  )
```



MODEL PERFORMANCE

Adjusted R^2:0.7751 (77.51% variance explained). F-statistic: Significant ($p < 2.2e-16$), confirming the model's predictive power.

LIMITATIONS AND CONSIDERATIONS

LIMITATIONS

1. Data Scope: The analysis is limited to historical data from a single source, potentially reducing generalizability.
2. Comments Anomaly: Requires further investigation to understand its unexpected effect.
3. Feature Limitations: Excludes visual and contextual factors like post aesthetics or timing.

FUTURE WORK

1. Include temporal features (e.g., time of posting).
2. Experiment with nonlinear models or machine learning algorithms (e.g., Random Forests) for potential performance gains.
3. Analyze post captions and hashtags using text analytics for deeper insights.

CONCLUSION

This analysis investigates key factors influencing Instagram post impressions using historical reach data. By leveraging predictive modeling, significant drivers of impressions were identified, including user engagement metrics (likes, shares, saves) and profile visits. The model explains 77.51% of the variation in impressions, offering actionable insights for optimizing content strategies. Recommendations include focusing on boosting engagement metrics and refining caption and hashtag strategies to amplify reach.