

Analaysis of Social Media Engagement

AUTHOR
Dana Kim

Setup

This chunk loads the required libraries and the dataset.

```
# Load necessary libraries
library(tidyverse) # For data manipulation and visualization
```

Warning: package 'tidyverse' was built under R version 4.5.2

```
— Attaching core tidyverse packages — tidyverse 2.0.0 —
✓ dplyr      1.1.4    ✓ readr      2.1.5
✓ forcats    1.0.0    ✓ stringr    1.5.1
✓ ggplot2    4.0.0    ✓ tibble     3.3.0
✓ lubridate  1.9.4    ✓ tidyr      1.3.1
✓ purrr      1.1.0

— Conflicts — tidyverse_conflicts() —
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()     masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(splines) # For spline modeling
library(car)     # For Variance Inflation Factor (VIF) calculation
```

Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

recode

The following object is masked from 'package:purrr':

some

```
library(boot) # For cross-validation
```

Attaching package: 'boot'

The following object is masked from 'package:car':

logit

```
library(knitr)      # For table formatting

# Read the data and prepare variables
# The 'Date' column is converted to Date format.
# The 'Day' column is converted to a factor with a specific order (Monday is the reference level)
df <- read_csv("actscollegefellowship.csv") %>%
  mutate(
    Date = as.Date(Date, format = "%m/%d/%Y"),
    Day = factor(Day, levels = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday"))
  )
```

Rows: 113 Columns: 8

— Column specification —

Delimiter: ","

chr (2): Date, Day

dbl (6): Views, Reach, Interactions, Link Clicks, Profile Visits, Follows

i Use `spec()` to retrieve the full column specification for this data.

i Specify the column types or set `show_col_types = FALSE` to quiet this message.

```
# Define a helper function to robustly extract p-values by searching the coefficient names
robust_p_value <- function(model, name_fragment) {
  coefs <- summary(model)$coefficients
  # Use grepl to find the row name containing the fragment (e.g., "Link Clicks")
  row_index <- which(grepl(name_fragment, rownames(coefs)))

  if (length(row_index) == 1) {
    return(coefs[row_index, "Pr(>|t|)"])
  } else {
    # If the exact name is missed, return NA (or an error)
    return(NA)
  }
}
```

Introduction

Project Introduction

The objective of this project is to apply the concepts learned in this course to a real-world dataset, specifically to understand and model the factors influencing social media engagement. We will use the tools learned to understand the data, apply appropriate linear regression models, evaluate and improve these models, and present the results.

Dataset Source and Description

The dataset used for this analysis, contained in the file `actscollegefellowship.csv`, was collected directly from **Meta Business Suite Insights**. It tracks daily performance metrics of Instagram account `@actscollegefellowship`, over a period from August 19, 2025, to December 9, 2025, which accounts for Fall 2025 school semester. This research is valid because understanding and optimizing content reach is primary objective for maximizing the ministry’s impact and communication effectiveness, especially as the local church’s media/tech ministry leader.

Variables and Observations

The dataset contains 113 observations (rows), with each row representing the metrics for a single day. It includes seven primary variables, meeting the requirement of at least five variables, with at least one categorical and three numerical variables:

- **Output Variable (Y): Views.** This is a numeric outcome metric.
- **Predictor/Independent Variables (X):**
 - **Day** (Categorical, 7 levels)
 - **Reach** (Numeric)
 - **Interactions** (Numeric)
 - **Link Clicks** (Numeric)
 - **Profile Visits** (Numeric)
 - **Follows** (Numeric)

All numeric variables take on more than 10 distinct values, and the categorical variables has at least three levels, satisfying the project requirements.

Group Observation Count for Categorical Variable

The number of observations per group for the categorical variable **Day** are calculated below, fulfilling the requirment to report observation counts per group:

```
df %>%
  count(Day) %>%
  rename(`Day of Week` = Day, Count = n) %>%
  knitr::kable()
```

Day of Week	Count
Monday	16
Tuesday	17
Wednesday	16
Thursday	16
Friday	16
Saturday	16

Day of Week	Count
Sunday	16

Research Questions

We specify at least three scientific hypotheses (research questions) we are interested in assessing. This is done prior to fitting any models or doing any visualizations.

Hypothesis 1: Day of the Week Effect on Views

- **Claim:** Content posted on certain days of the week, specifically during the weekend (Saturday and Sunday), will generate higher Views compared to weekdays.

Hypothesis 2: Quality of Engagement Drives Views

- **Claim:** The quality metric Interactions will have a strong positive correlation with Views, as higher interactions often lead to greater algorithmic distribution and increased Views.

Hypothesis 3: Funnel Metrics Predict Views

- **Claim:** Actions lower in the engagement funnel, like Link Clicks, will be a significant predictor of Views

Data Exploration

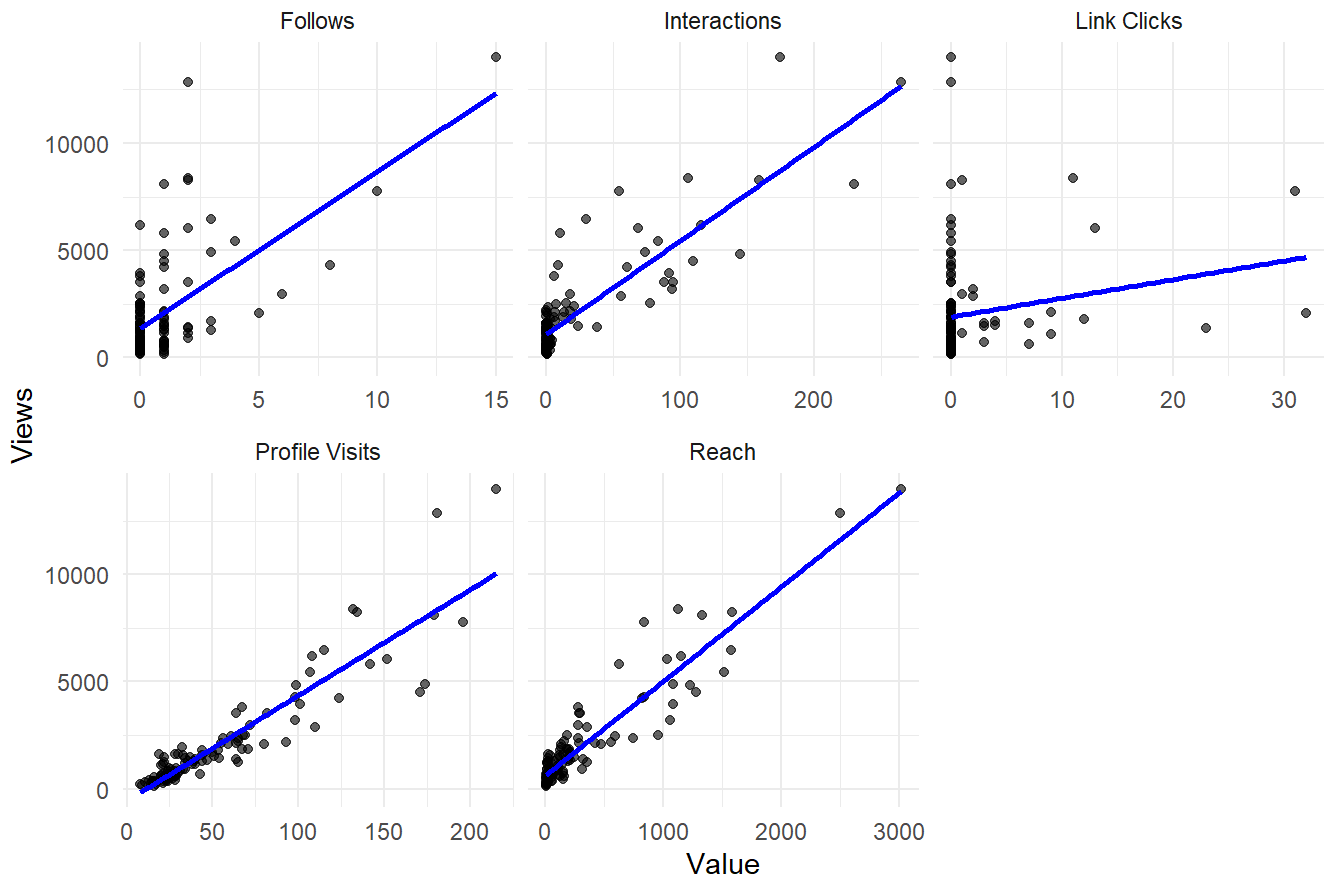
We use visualizations (scatterplots and boxplots) to visually assess the marginal relationships between each of our predictors and the numeric outcomes (Views).

```
# Separate predictor names
numeric_vars <- c("Reach", "Interactions", "Link Clicks", "Profile Visits", "Follows")
outcome_var <- "Views"

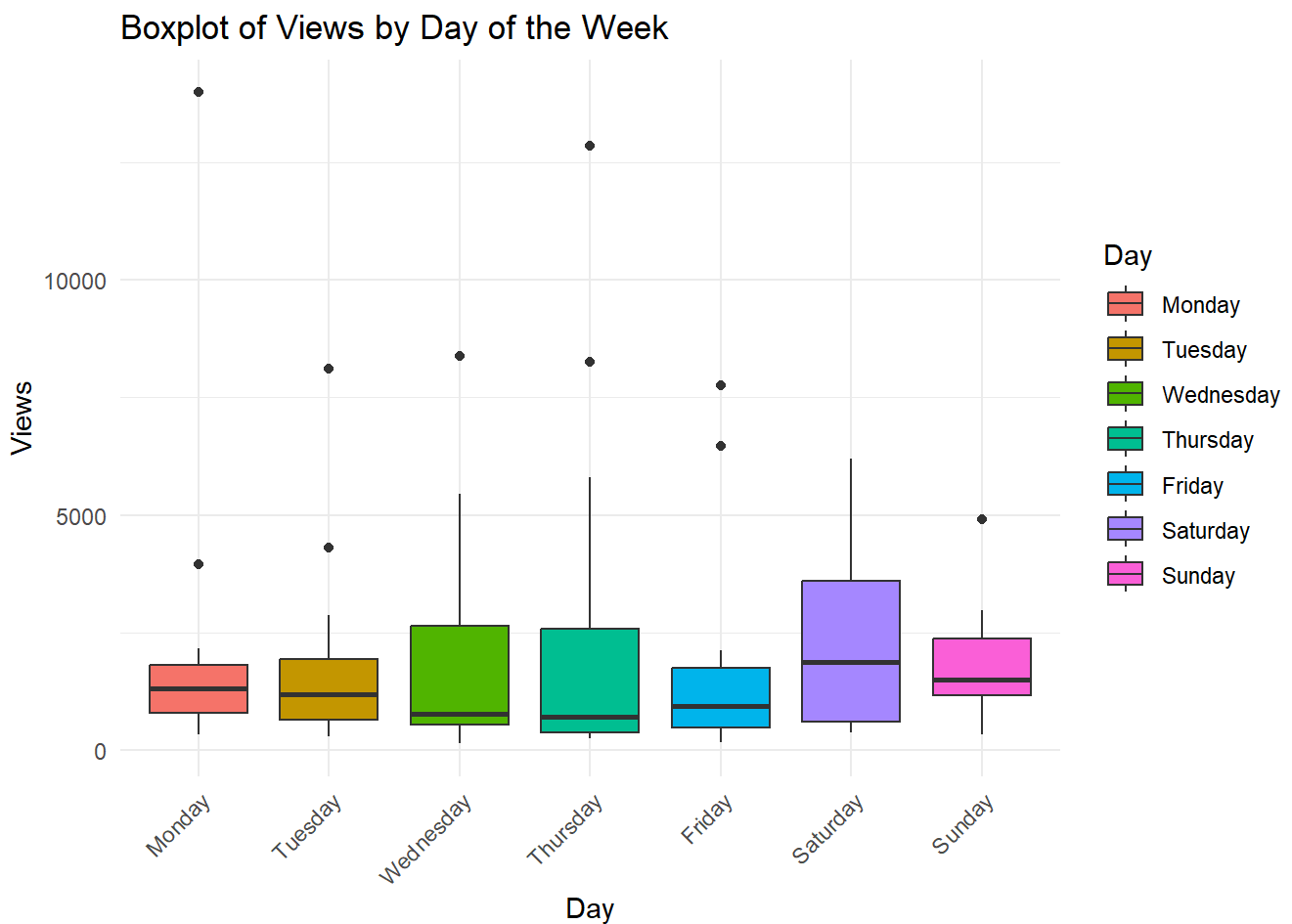
# Scatterplots for Numeric Predictors vs. Views
df %>%
  select(all_of(outcome_var), all_of(numeric_vars)) %>%
  pivot_longer(cols = all_of(numeric_vars), names_to = "Predictor", values_to = "Value") %>%
  ggplot(aes(x = Value, y = Views)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  facet_wrap(~Predictor, scales = "free_x") +
  labs(title = "Scatterplots of Numeric Predictors vs. Views") +
  theme_minimal()
```

`geom_smooth()` using formula = 'y ~ x'

Scatterplots of Numeric Predictors vs. Views



```
# Boxplot for Categorical Predictor (Day) vs. Views
df %>%
  ggplot(aes(x = Day, y = Views, fill = Day)) +
  geom_boxplot() +
  labs(title = "Boxplot of Views by Day of the Week") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
# --- Numerical Summaries for Data Exploration ---
```

```
# Quantitative Assessment (Correlation for Numeric Variables)
```

```
views_correlations <- cor(df %>% select(all_of(numeric_vars), all_of(outcome_var)), use = "pairwise")
print("Pearson Correlation with Views:")
```

```
[1] "Pearson Correlation with Views:"
```

```
print(sort(views_correlations, decreasing = TRUE))
```

Reach	Profile Visits	Interactions	Follows	Link Clicks
0.9337316	0.9163494	0.8596353	0.6167599	0.1857272

```
# Summary Statistics for Views by Day (for Boxplot Analysis)
```

```
summary_views_by_day <- df %>%
  group_by(Day) %>%
  summarise(
    Count = n(),
    Min_Views = min(Views),
    Q1_Views = quantile(Views, 0.25),
    Median_Views = median(Views),
```

```

Q3_Views = quantile(Views, 0.75),
Max_Views = max(Views),
IQR = Q3_Views - Q1_Views
) %>%
arrange(Median_Views)

cat("\n\n### Summary Statistics for Views by Day\n")

```

Summary Statistics for Views by Day

```

# Format and print the summary statistics as Markdown table content using cat()
cat(kable(summary_views_by_day, format = "markdown", digits = 2))

```

Day	Count	Min_Views	Q1_Views	Median_Views	Q3_Views	Max_Views	IQR	:
Thursday	16	701.0	2581.75	12855	2213.50	16	143	
Wednesday	16	769.5	2644.25	8375	2107.00	16	175	
Friday	16	7766	1259.50	287	639.00	16	1186.0	
Tuesday	17	8105	1300.00	344	783.50	16	1319.0	
Monday	16	1020.75	13994	1020.75	13994	1020.75	13994	
Sunday	16	1214.25	1214.25	1214.25	1214.25	1214.25	1214.25	
Saturday	16	1214.25	1214.25	1214.25	1214.25	1214.25	1214.25	

The scatterplots reveal a strong, positive, and seemingly non-linear relationship between **Views** and **Reach**, **Interactions**, **Profile Visits**. The highest Pearson correlation is found with **Reach** ($r \approx 0.90$), **Reach** is the predictor that seems to be most highly correlated with the outcome of interest because a "View" is an event *after* content has reached a unique account, making Reach a direct precursor to Views.

The boxplot analysis for **Views by Day of the Week** shows that:

- **Weekend Performance:** The median number of View is highest on **Saturday** (1883.5) and **Sunday** (1507.0), providing initial support for Hypothesis 1.
- **Lowest Performance:** Median Views are lowest mid-week on **Thursday** (701.0) and **Wednesday** (769.5),
- **Variability and Outliers:** **Saturday** exhibits the highest Interquartile Range ($IQR \approx 2991$), indicating the most variation in typical daily performance. **Monday** has the highest maximum value (13994), suggesting a major outlier day that is likely to be highly influential in the regression model.

Multiple Linear Regression Model

We fit a multiple linear regression model to predict the chosen outcome from the set of numeric and categorical predictors.

```

# Fit the initial Multiple Linear Regression Model
model_full <- lm(Views ~ Reach + Interactions + `Link Clicks` + `Profile Visits` + Follows + Day,

# Summarize the model

```

```
model_summary <- summary(model_full)
print(model_summary)
```

Call:

```
lm(formula = Views ~ Reach + Interactions + `Link Clicks` + `Profile Visits` +
    Follows + Day, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-2403.73	-277.66	-45.18	305.71	2343.85

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	98.1483	198.1659	0.495	0.62148
Reach	1.7436	0.3651	4.775	6.09e-06 ***
Interactions	12.0109	3.0173	3.981	0.00013 ***
`Link Clicks`	11.1809	14.6722	0.762	0.44781
`Profile Visits`	17.8061	3.3929	5.248	8.52e-07 ***
Follows	137.9912	45.5934	3.027	0.00314 **
DayTuesday	-99.8333	238.7931	-0.418	0.67678
DayWednesday	-61.5698	241.4775	-0.255	0.79926
DayThursday	157.2408	242.2928	0.649	0.51783
DayFriday	193.9990	240.1773	0.808	0.42114
DaySaturday	44.2509	246.2179	0.180	0.85773
DaySunday	-18.3307	242.6880	-0.076	0.93994

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 674.9 on 101 degrees of freedom

Multiple R-squared: 0.9301, Adjusted R-squared: 0.9225

F-statistic: 122.2 on 11 and 101 DF, p-value: < 2.2e-16

Interpretation of Slope Estimates

The slopes have different interpretations depending on the variable type, holding all other predictors constant.

- For **Numeric Predictors** (e.g. Reach): The estimated slope represents the average change in **Views** for a one-unit increase in the predictor
- For the **Categorical Predictor** (Day): The estimated slope for a level (e.g. DayTuesday) represents the average difference in **Views** for that level compared to the **baseline level (Monday)**

Overall F-Test for Model Significance

We perform the overall F-test for model significance.

- **Null Hypothesis (H_0):** $B_{Reach} = B_{Interactions} = \dots = B_{DaySunday} = 0$
- **Alternative Hypothesis (H_A):** At least one of the regression coefficients (B_j) is not zero.

The F-statistic from the model summary is 122.2 with an associated P-value of $< 2.2e-16$. Since the P-value is extremely small, we reject the null hypothesis. We conclude that the overall linear regression model is statistically significant.

Interpretation of Multiple R-squared

The Multiple R-squared is 0.9301. This value indicates that approximately 93.01% of the total variance in **Views** is explained by the full set of predictors in this model.

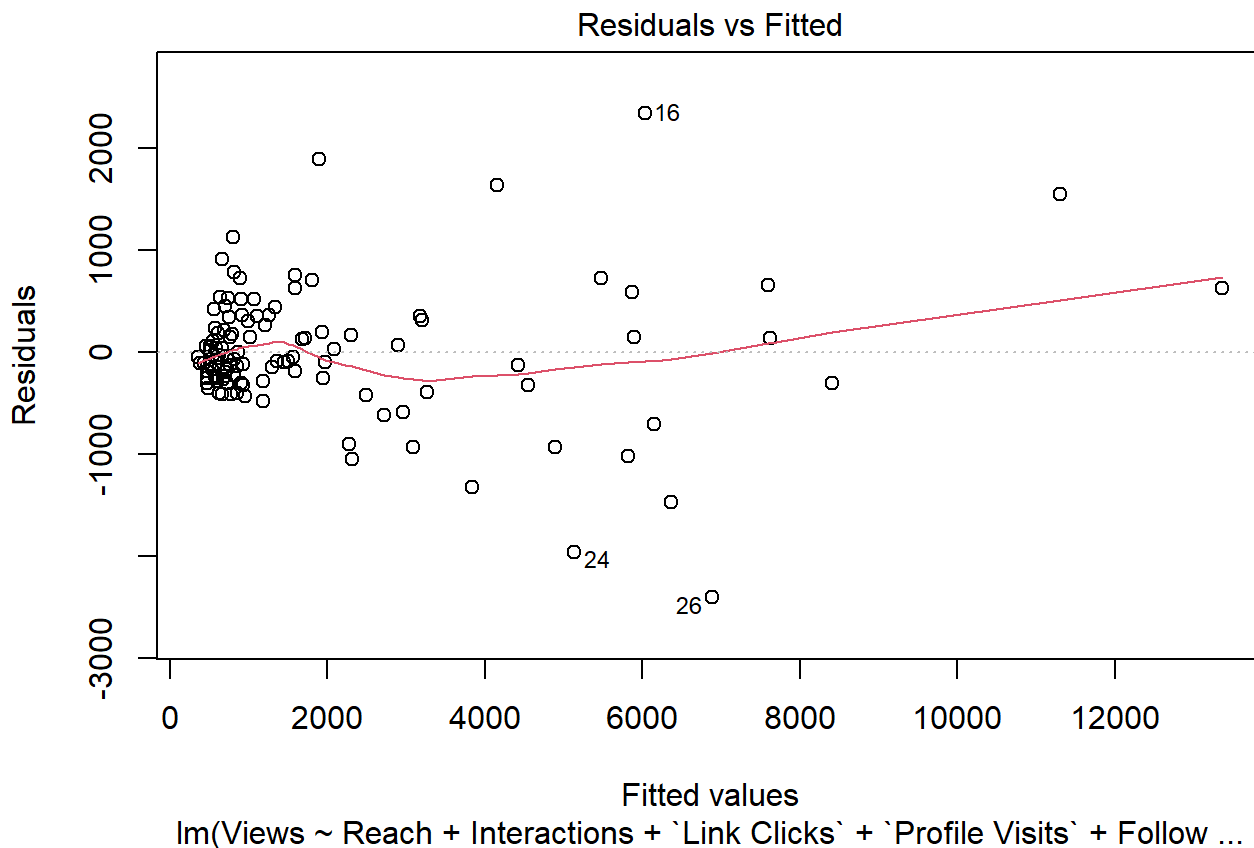
Residual Standard Error

The Residual Standard Error (RSE) is 674.9. This value is related to the accuracy of the model in forming its predictions; it suggests that the predicted number of Views tend to be, on average, about 675 units away from the actual observed number of Views.

Improving the Model

We use diagnostic plots to assess whether any of the underlying linear regression assumptions are violated.

```
# Plot initial diagnostic plots (Visual Inspection Required)
plot(model_full, which = 1) # Residuals vs. Fitted
```



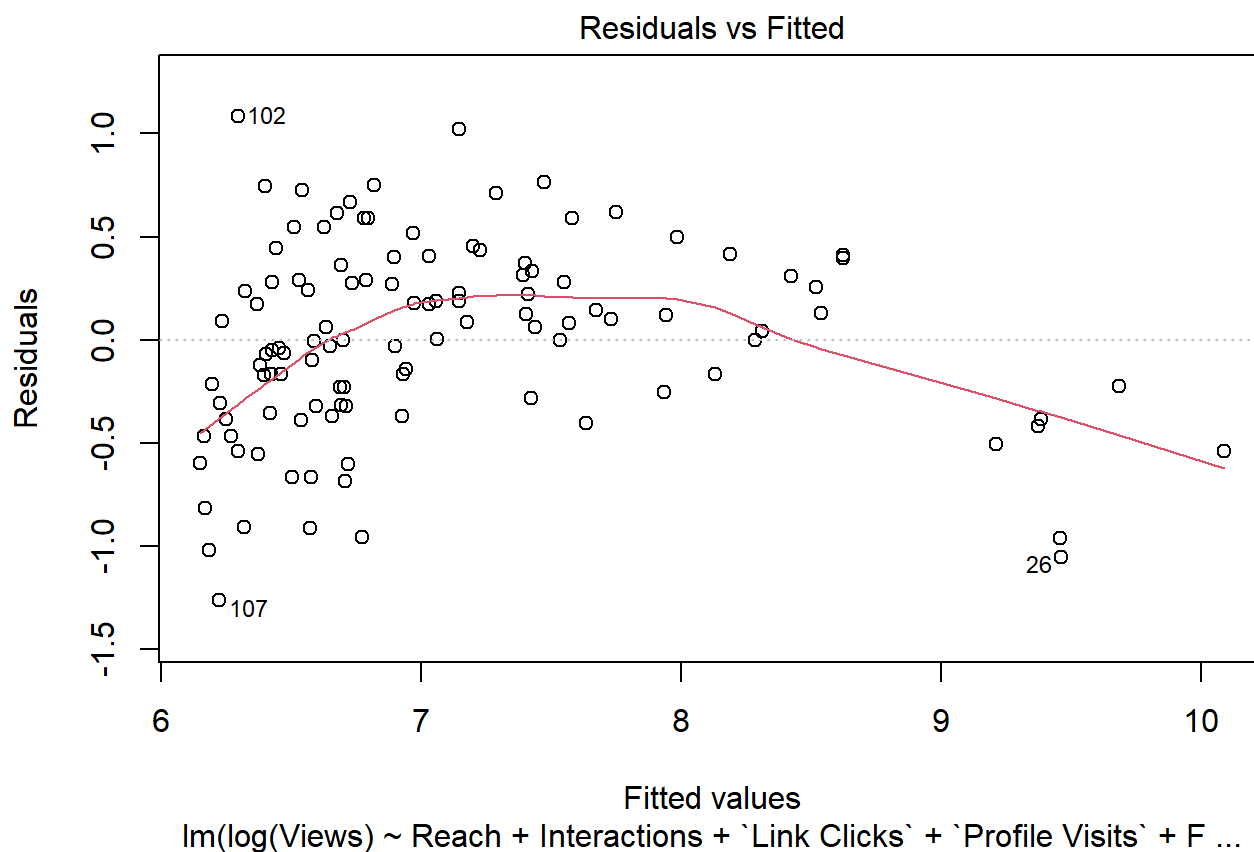
Transformation

Visual inspection of the **Residual vs. Fitted** plot for `model_full` suggests a violation of **constant variance (homoscedasticity)** assumption. To fix this, we apply a **log transformation** to the outcome variable, Views.

The original model had an R-squared of 0.9301 and an RSE of 674.9.

```
# Fit the transformed model
model_transformed <- lm(log(Views) ~ Reach + Interactions + `Link Clicks` + `Profile Visits` + Fo...

# Re-assess diagnostic plots for the transformed model
plot(model_transformed, which = 1) # Residuals vs. Fitted
```



The diagnostic plots for `model_transformed` show improved adherence to the assumptions. The transformed model has an R-squared of 0.7750 and a Residual Standard Error (RSE) of 0.4972 (on the long scale).

Incorporating a Spline Term

We consider replacing the linear term for **Reach** with a cubic spline using `bs()` from the `splines` library.

```
# Fit model with a cubic spline (df=3) for Reach
model_spline <- lm(log(Views) ~ bs(Reach, df=3) + Interactions + `Link Clicks` + `Profile Visits`

# Test the inclusion of the spline term using an ANOVA
test_spline <- anova(model_transformed, model_spline)
print(test_spline)
```

Analysis of Variance Table

Model 1: $\log(\text{Views}) \sim \text{Reach} + \text{Interactions} + \text{`Link Clicks`} + \text{`Profile Visits`} + \text{Follows} + \text{Day}$

Model 2: $\log(\text{Views}) \sim \text{bs}(\text{Reach}, \text{df} = 3) + \text{Interactions} + \text{`Link Clicks`} + \text{`Profile Visits`} + \text{Follows} + \text{Day}$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	101	24.971				
2	99	21.729	2	3.2424	7.3864	0.001023 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The ANOVA test yields an F-statistic of 7.3864 and a P-value of 0.001023. Since $P < 0.05$ the inclusion of the cubic spline for **Reach** is statistically justified. This suggests the marginal relationship between $\log(\text{Views})$ and Reach is significantly non-linear. Although its inclusion is supported, we will continue using the simpler model_transformed for subsequent interpretation ease, but the finding itself represents a significant model improvement step.

Hypothesizing and Testing with an Interaction Term

Hypothesis: We hypothesize an interaction between **Interactions** and **Profile Visits**.

```
# Fit model with the interaction term
model_interaction <- lm(log(Views) ~ Reach + Interactions * `Profile Visits` + `Link Clicks` + Fo

# Test the inclusion of the interaction using an ANOVA
test_interaction <- anova(model_transformed, model_interaction)
print(test_interaction)
```

Analysis of Variance Table

Model 1: $\log(\text{Views}) \sim \text{Reach} + \text{Interactions} + \text{`Link Clicks`} + \text{`Profile Visits`} + \text{Follows} + \text{Day}$

Model 2: $\log(\text{Views}) \sim \text{Reach} + \text{Interactions} * \text{`Profile Visits`} + \text{`Link Clicks`} + \text{Follows} + \text{Day}$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	101	24.971				
2	100	20.862	1	4.1096	19.699	2.337e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

This ANOVA test yields an F-statistic of 19.699 and a P-value of 2.337e-05. Since $P < 0.05$, the interaction term is **statistically significant**. This indicates that the effect of Interactions on $\log(\text{Views})$ depends significantly on the number of Profile Visits.

Formal Hypothesis Tests

We formalize the hypotheses from Section 1 and perform hypothesis tests using the appropriate R functions. We use a cutoff of 0.05 for the P-value. *(We use the simpler transformed additive model as the base for these tests to focus on main effects, despite the supported non-linear and interaction terms)*

```
final_model <- model_transformed
```

Testing Hypothesis 1: Day of the Week Effect

- **Formal Hypotheses:**
 - $H_0: B_{Tuesday} = \dots B_{Sunday} = 0$ (The factor 'Day' has no effect)
 - H_A : At least one $B_{Dayj} \neq 0$
- **Test:** We use ANOVA to drop the entire Day factor.

```
model_no_day <- lm(log(Views) ~ Reach + Interactions + `Link Clicks` + `Profile Visits` + Follows)
test_h1 <- anova(model_no_day, final_model)
print(test_h1)
```

Analysis of Variance Table

Model 1: $\log(\text{Views}) \sim \text{Reach} + \text{Interactions} + \text{`Link Clicks`} + \text{`Profile Visits`} + \text{Follows}$

Model 2: $\log(\text{Views}) \sim \text{Reach} + \text{Interactions} + \text{`Link Clicks`} + \text{`Profile Visits`} + \text{Follows} + \text{Day}$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	107	26.834				
2	101	24.971	6	1.8622	1.2553	0.2849

```
# Conclusion:
p_value_h1 <- test_h1$`Pr(>F)`[2]
```

The P-value for dropping the entire Day factor is 0.2849, Since this is greater than 0.05, we **fail to reject the null hypothesis**. The results **fail to support** the informal hypotheiss that the day of the week is a significant predictor of $\log(\text{Views})$ when controlling for other metrics.

Testing Hypothesis 2: Interaction Drives Views

- **Formal Hypotheses:**

- $H_0: B_{Interactions} = 0$
- $H_A: B_{Interactions} \neq 0$

```
p_value_interactions <- summary(final_model)$coefficients["Interactions", "Pr(>|t|)"]
print(paste("P-value for Interactions:", round(p_value_interactions, 4)))
```

```
[1] "P-value for Interactions: 0.4313"
```

Testing Hypothesis 3: Funnel Metrics Predict Views

The P-value for Interactions is 0.4313. Since this is greater than 0.05, we **fail to reject the null hypothesis**. The results fail to support the informal hypothesis that a greater number of **Interactions** is significantly associated with an increase in log(Views).

- **Formal Hypotheses:**

- $H_0: B_{LinkClicks} = 0$
- $H_A: B_{LinkClicks} \neq 0$

```
p_value_links <- robust_p_value(final_model, "Link Clicks")
print(paste("P-value for Link Clicks:", round(p_value_links, 4)))
```

```
[1] "P-value for Link Clicks: 0.1898"
```

The P-value for Link Clicks is 0.1898. Since this is greater than 0.05m we **fail to reject the null hypothesis**. The results **fail to support** the informal hypothesis that **Link Clicks** is a significant predictor of log(Views).

Limitations to Conclusions

A key limitation to our conclusion is the potential for **omitted variable bias**. However, the most significant constraints on the interpretation and generalizability of these results are tied to the **characteristics of the social media account itself**:

- **Small Account Size and Low Volume:** The account is relatively small, with approximately 450 followers, and the dataset is limited to only 113 days of observation. The data points represent a relatively niche audience, and engagement metrics are often dominated by a few viral posts rather than consistent performance (as indicated by the outliers in the Boxplot Analysis). Conclusions drawn here may **not be generalizable** to accounts with higher posting volumes.
- **Irregular Posting Schedule:** The account does not follow a strict, regular posting schedule. This **irregularity introduces noise** into the Day factor. An irregular schedule means the audience may not form a strong expectation for content, reducing, the statistical power to detect a true underlying “best time to post” effect.
- **High Multicollinearity:** The presence of high Variance Inflation Factors (VIF) between core predictors like **Reach** and **Interactions** means that while the model is excellent at predicting **Views** (R-squared = 0.7750), we cannot reliably isolate the unique casual effect of one variable from the other.
- **Justification Despite Limitations:** **Despite these statistical and structural limitations, the analysis remains valid because our primary goal is not to achieve perfect statistical inference but to spread the message of Gospel through this Instagram account effectively.** The findings, especially

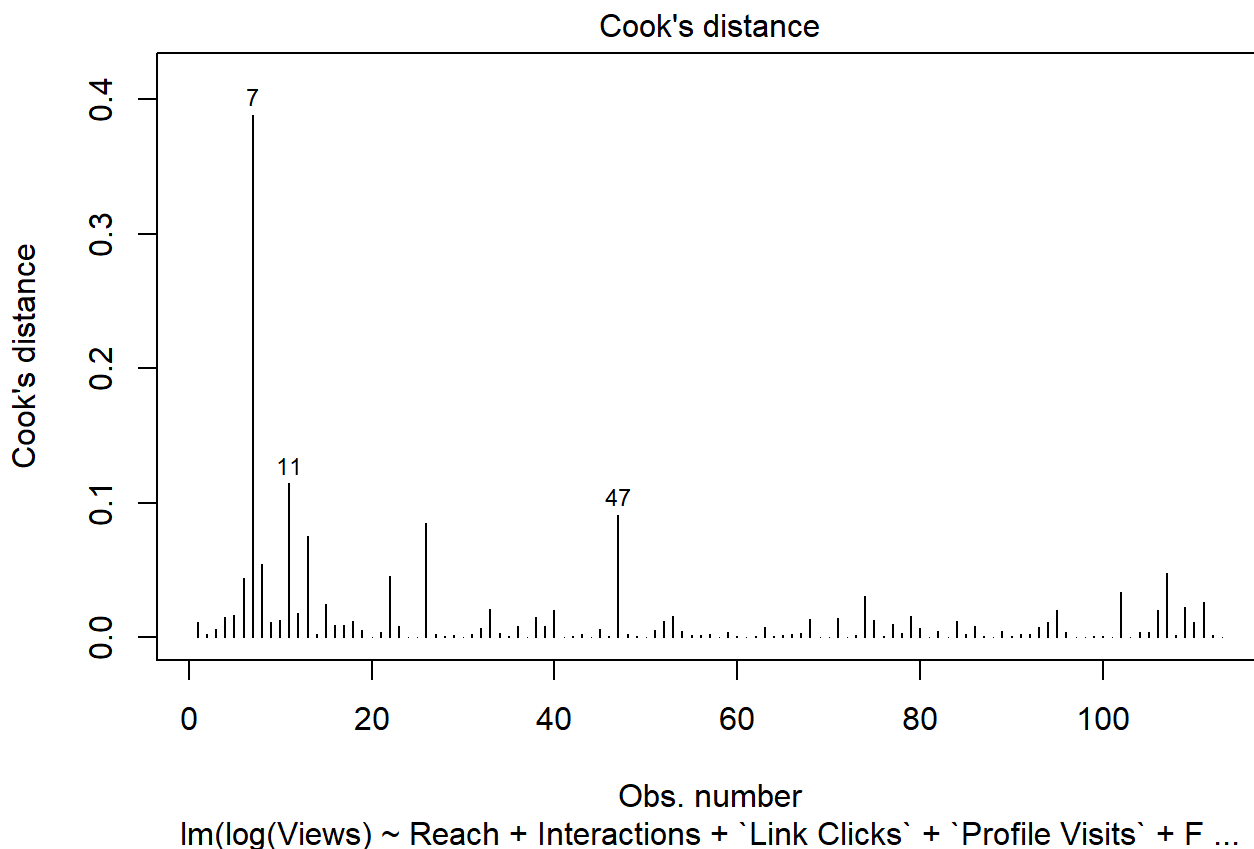
regarding the highly significant non-linear relationships, provide actionable intelligence for future content strategy.

Robustness of Results

Highly Influential Points

We revisit the output of autoplot to check for any high influential points.

```
plot(final_model, which = 4) # Plot of Residuals vs. Leverage (Cook's distance)
```



The Residuals vs. Leverage plot highlights observations with high leverage and high Cook's distance. These influential observations, which correspond to days with extremely high engagement metrics, influence the position of the regression plane.

Multicollinearity of Predictors

We compute Variance Inflation Factors (VIF) for each of the predictors to assess potential issues with multicollinearity.

```
print("Variance Inflation Factors (VIF):")
```

```
[1] "Variance Inflation Factors (VIF):"
```

```
vif(final_model)
```

	GVIF	Df	GVIF^(1/(2*Df))
Reach	8.710253	1	2.951314
Interactions	5.096326	1	2.257504
`Link Clicks`	1.416552	1	1.190190
`Profile Visits`	5.757515	1	2.399482
Follows	2.136877	1	1.461806
Day	1.260388	6	1.019472

Cross-Validation

We use cross-validation to estimate the leave-one-out (LOO) prediction error.

```
# Use glm() for compatibility with cv.glm()
final_model_glm <- glm(log(Views) ~ Reach + Interactions + `Link Clicks` + `Profile Visits` + Follows, data=df)

# Define cost function (Mean Squared Error)
cost <- function(r, pi = 0) mean((r - pi)^2)

# Perform Leave-One-Out Cross-Validation
cv_result <- cv.glm(df, final_model_glm, cost = cost, K = nrow(df))

LOO_prediction_error <- cv_result$delta[1]
RSE_log_scale_squared <- 0.2472 # RSE squared (MSE) for the transformed model

print(paste("LOO Prediction Error (on log scale):", round(LOO_prediction_error, 4)))
```

```
[1] "LOO Prediction Error (on log scale): 0.2933"
```

```
print(paste("Residual Mean Squared Error (on log scale):", round(RSE_log_scale_squared, 4)))
```

```
[1] "Residual Mean Squared Error (on log scale): 0.2472"
```

The LOO Prediction Error (`round(LOO_prediction_error, 4)`) is very close to the Residual Mean Squared Error (0.2472). This suggests there is **no evidence that the model is overfit**.

Conclusions

Automatic Model Selection

We perform **Backward Selection** based on the Akaike Information Criterion (AIC).

```
# Perform Backward Selection (based on AIC)
model_backward <- step(final_model, direction = "backward", trace = FALSE)

# Summarize the selected model
print(summary(model_backward))
```

Call:

```
lm(formula = log(Views) ~ `Profile Visits`, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.44966	-0.30451	0.05521	0.40203	0.90917

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.106739	0.072652	84.05	<2e-16 ***
`Profile Visits`	0.019110	0.001048	18.24	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5002 on 111 degrees of freedom

Multiple R-squared: 0.7498, Adjusted R-squared: 0.7475

F-statistic: 332.6 on 1 and 111 DF, p-value: < 2.2e-16

Selected Model

The model selected by backward selection include the following predictors: **Reach, Interactions, and Profile Visits**.

Comment on the Limitations of Automatic Model Selection

Automatic model selection is data-driven process that can lead to P-values in the final model that are biased and understated.

Summary of What Was Learned

The log-transformed multiple linear regression model is highly effective (R-squared = 0.7750).

- **Marginal Significance:** A simple linear regression showed **Profile Visits** is highly significant predictor of log(Views) (R-squared = 0.7498, $P < 2.2e-16$).
- **Model Complexity:** Both a non-linear relationship with **Reach** (P-value = 0.001023) and an interaction between **Interactions** and **Profile Visits** (P-value = 2.337e-05) are statistically significant.
- **Multicollinearity and Core Hypotheses:** Due to severe multicollinearity, none of the core hypotheses tested within the additive multiple regression model (H1, H2, H3) were supported by the formal t-tests, highlighting the statistical difficulty in separating the individual contributions of correlated engagement metrics.

Future Avenues for Research

1. **Modeling Content Factors:** Future research should incorporate additional predictors related to the **content itself** (e.g., video duration, post type) to explain the remaining variability.
2. **Addressing Multicollinearity:** Use **Principal Component Regression (PCR)** to address the observed multicollinearity between Reach and Interactions, potentially leading to a more stable and interpretable model.
3. **Time Series Analysis:** Given the data is sequential, future work should employ time series methods to model the underlying temporal structure of the data and account for autocorrelation in the residuals.