Statistical Analysis of Instagram Story Performance: Insights from University Student Profiles

TABLE OF CONTENTS:

Davide Fraschini *(3243178)*

Carlotta Greco *(3241535)*

Viviana Locatelli *(3242659)*

Introduction

This report analyzes Instagram usage among university students to identify factors influencing the performance of the Stories feature, particularly in terms of views. Using a database of 243 observations, updated as of October 31, 2024, and key statistical methods as descriptive statistics, confidence intervals, hypothesis testing and regression analysis, the study examines variables such as, for example, the number of followers, daily usage time or sex. Findings reveal differences in story views across user segments, with moderate correlations between views and variables like follower count and daily time spent on the app. Predictive models helped understanding the key drivers of engagement, while acknowledging variability in data interpretation.

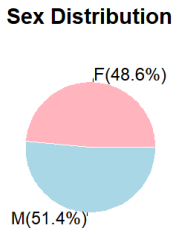
Moreover, under every part is written and briefly explained the Rcode used to interpret data and find the results. In the appendix is available the full Rcode along with the database.

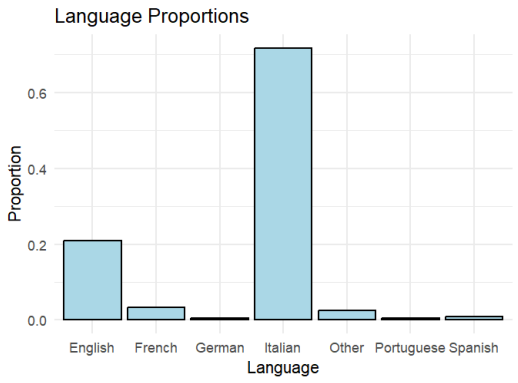
Descriptive Statistics of the sample

A.1.

For categorical variables, frequency distribution tables and bar charts are used to summarize and visualize the data. Discrete and continuous numeric variables are mainly analyzed using measures of central tendency such as the mean and median, as well as measures of dispersion like the standard deviation and interquartile range, allowing detection of outliers or skewness patterns. Graphical representations, including histograms and boxplots, are particularly useful in examining the spread and symmetry of continuous variables. By considering both central and non-central measures, the analysis offers a comprehensive view of the data.

The dataset's **sex** variable was identified as categorical. It shows a fairly balanced distribution, with 118 females ('F': 48.6%) and 125 males ('M': 51.4%). The mode of the variable is male, indicating the slight predominance of male respondents. The sample is representative for analyzing gender-related trends in Instagram usage, as visually highlighted by the near-equal distribution of sexes in the pie chart.

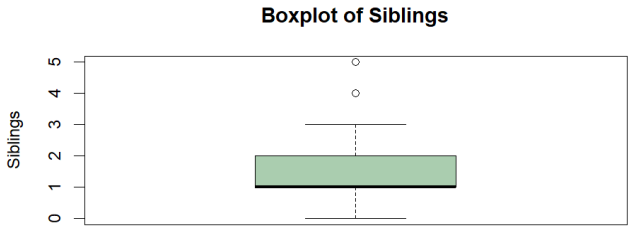


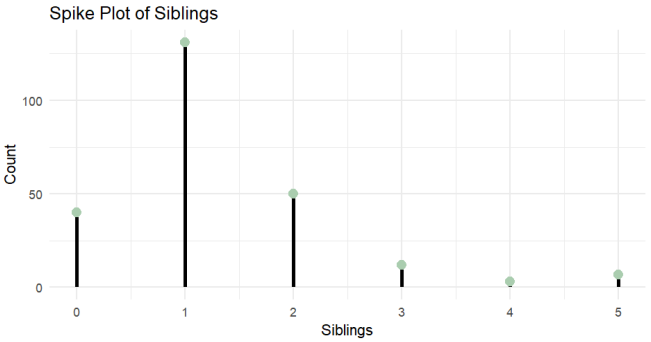


Language, another categorical variable, demonstrates a strong predominance of Italian speakers, representing 71.6% of respondents. Other languages, such as English (21.0%), French (3.3%), and Spanish (0.8%), contribute to a smaller proportion of the dataset, while German, Portuguese, and other languages are minimally represented. The distribution of languages in the sample is heavily uneven, as suggested by the bar chart.

The **siblings** variable, just like **account_num**, was treated as a discrete variable. It shows a range from 0 (only child) to 5 siblings. The mean is

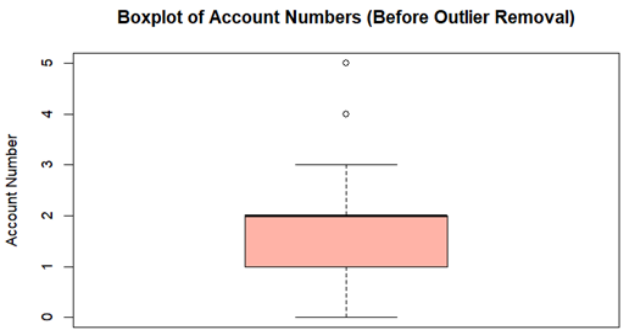
approximately 1.292, the median is 1, and the mode is also 1, indicating that most respondents have one sibling. The data distribution is right-skewed, with over half of the respondents having one sibling (53.9%) and a notable proportion being only children (16.5%). Variability, as indicated by the standard deviation, is moderate. A box plot and frequency distribution chart visualize the spread of the data, confirming the presence of 4 and 5 as mathematical outliers. However, these two have been retained as they do not significantly distort the results or impact the analysis, since they are not part of the

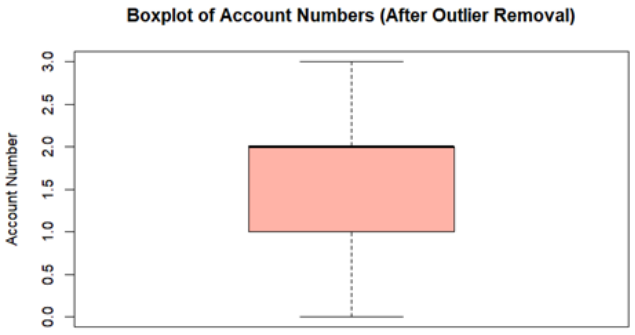




variables used for linear regression.

The **account number** variable originally ranged from 0 to 20. However, 20 was interpreted as an error in the survey and replaced with 2 in the original dataset. Therefore the variable ranged from 0 to 5 in the original dataset, with most respondents accessing one account (41.2%) or two accounts (37.0%), making one the mode. Mathematical outliers identified and visualizable from the boxplot, notably those above three accounts, were removed to ensure consistency with the treatment used for the other variables. After this adjustment, the mean number of accounts is 1.671, with a median and third quartile of 2. The filtered data offers a clearer picture of account usage trends, with the extremes values transformed in NAs.

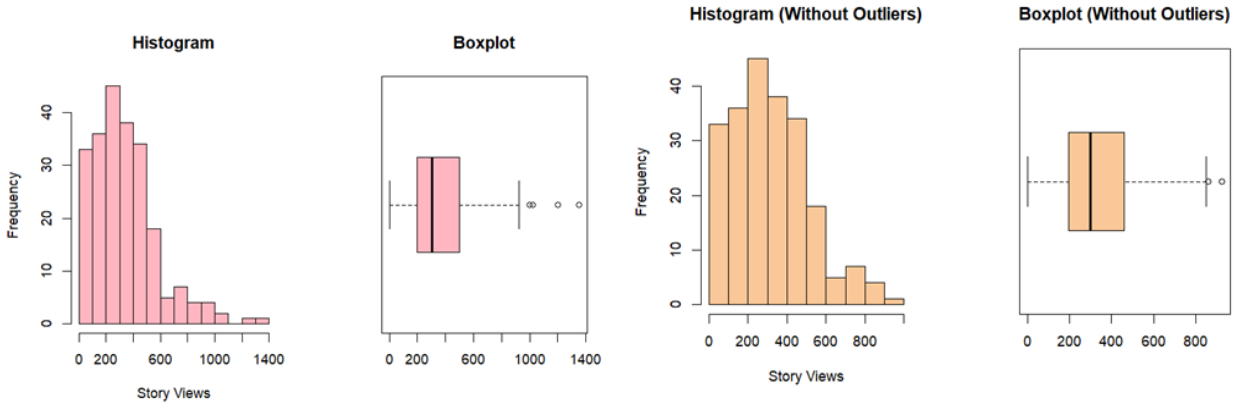




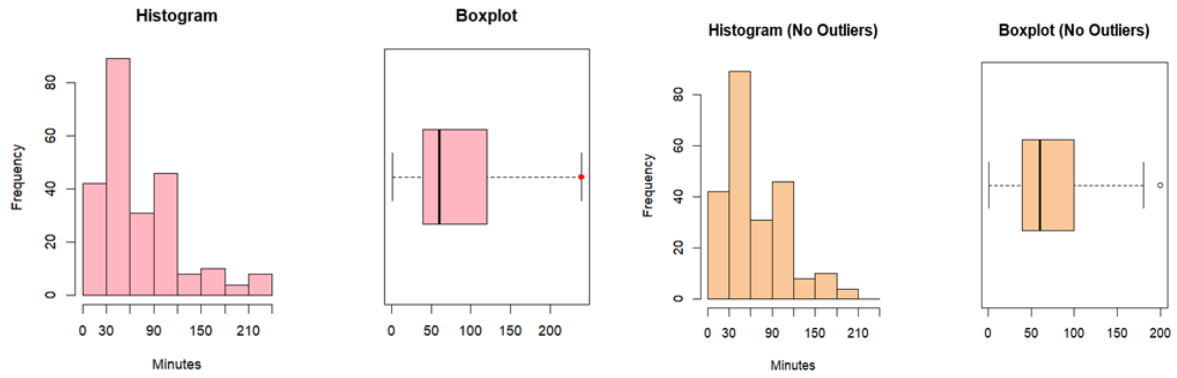
The following variables were analyzed as continuous numeric: story_views, day_time_min, num_follower, num_post.

The **story_views** variable originally showed a wide range, with values spanning from 1 to over 1,357, indicating substantial variability. The mean story views was 355.618, with a median of 302.5 and an interquartile range (IQR) of 300. Outliers, defined as values exceeding 927 based on the IQR method, were removed to improve robustness. After this adjustment, the mean lowered to 332.5, and the range narrowed significantly. For improved readability data was

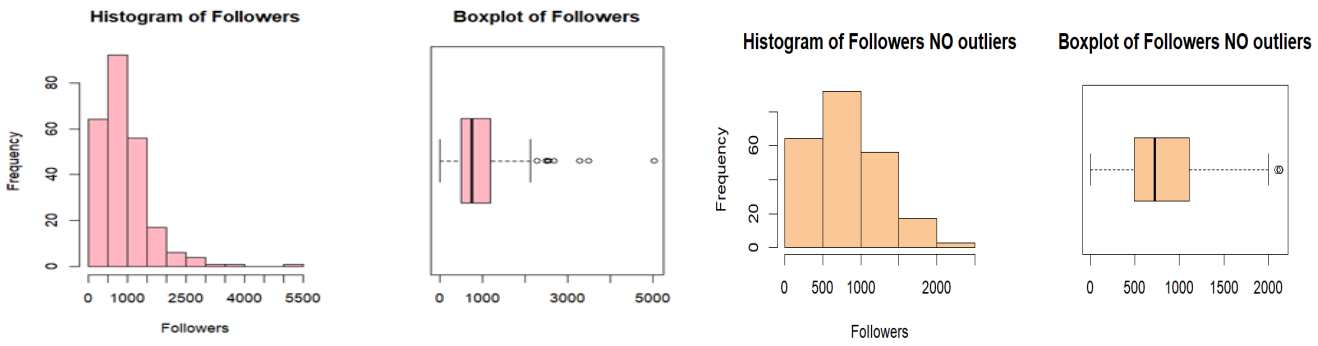
grouped in bins of 100. Histograms and boxplots, both before and after outlier removal, illustrate a skewed distribution, emphasizing the concentration of story views in the lower ranges.



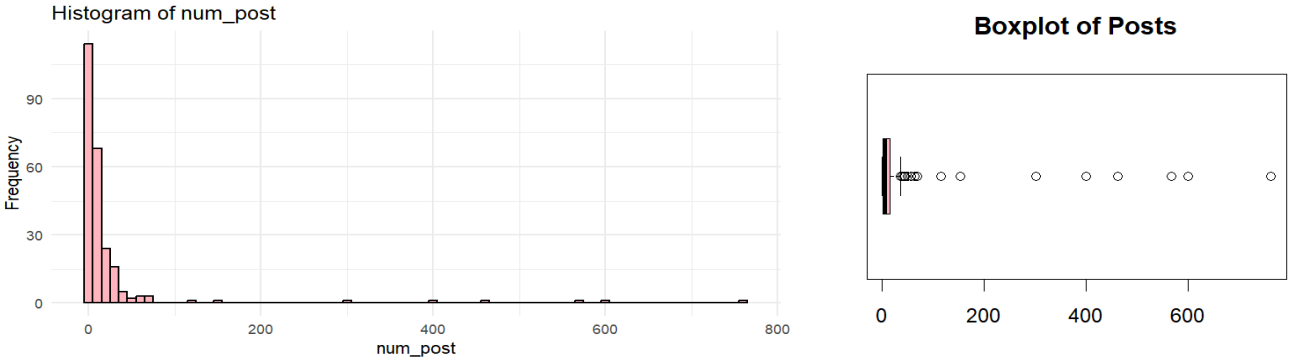
The **day_time_min** variable reflects the average daily minutes spent on Instagram, with a mean of approximately 74.6 minutes and a median of 60 minutes after outliers were removed. The data exhibits considerable variability, as indicated by a standard deviation of 44.076 minutes and a coefficient of variation of about 59%. Outliers, primarily values at 240 minutes, were excluded. The updated histogram and boxplot reveal a concentration of usage between 40 and 100 minutes, aligning with the interquartile range of 60 minutes. Data was grouped into bins of 30-minutes each.

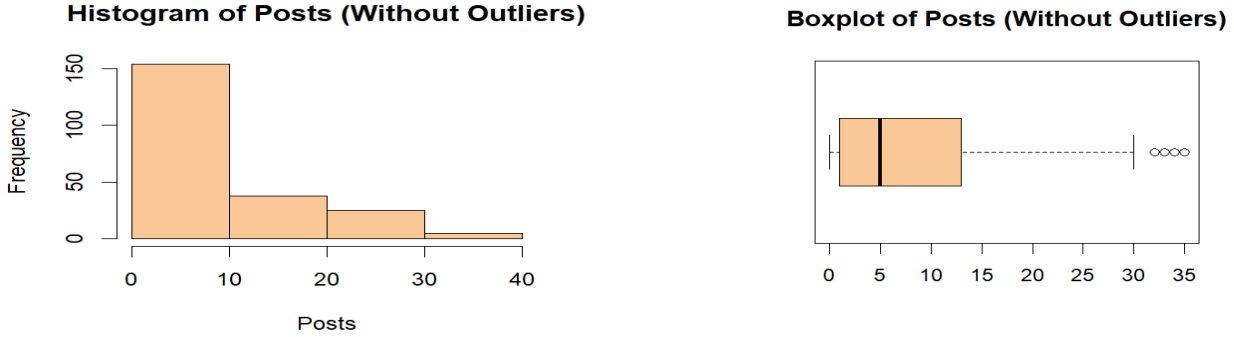


The **number of followers** variable is highly right-skewed, with a mean of approximately 812.5 and a median of 726.5 after outliers removal, indicating that most users have follower counts below the mean. The interquartile range (IQR) is 608, and the data show considerable variability with a standard deviation of 457.6 and a coefficient of variation of 56.3%. Outliers, ranging from 2,577 to 5,057 followers, were identified and removed. However the distribution remains heavily concentrated in the lower ranges. Moreover, it is to mention that data has first been filtered, removing the non-existing accounts with no followers and no posts and the NA values.

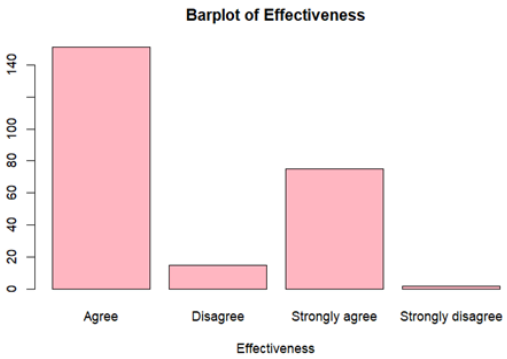


The **number of posts** variable has also been filtered for the non-existing accounts with no followers and no posts and for the NAs values. For easier readability data has been grouped in classes of 10 posts. Distribution is extremely right-skewed, with a mean of approximately 24.33 posts and a median of 6 posts, indicating that the majority of users post less frequently than the mean suggests. The interquartile range (IQR) is 13.50, reflecting a moderate spread, while the coefficient of variation (342.9%) highlights extreme variability in the dataset. To assess the problem, outliers, ranging up to 763 posts, were identified and removed. Post-outlier removal, the distribution remains skewed, but with a significant decrease in the differences between mean (8.396) and median (5) and a drastic reduction in the coefficient of variation(106.6%), which remains, however, extremely high.

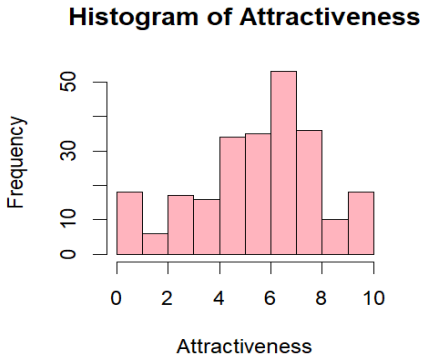




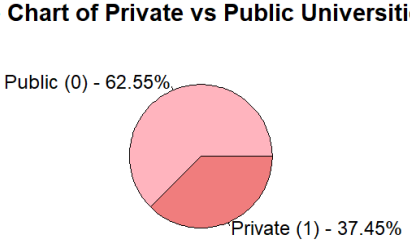
Effectiveness is an ordinal variable, measuring the level of agreement with the statement "*Instagram is an effective way to get in touch with other people*". Responses are ranked on a Likert scale from "*Strongly Disagree*" to "*Strongly Agree*". The survey is dominated by the "*Agree*" category, accounting for 62.1% of responses. This is followed by "*Strongly agree*" at 30.9%, while "*Disagree*" (6.2%) and "*Strongly disagree*" (0.8%) are much less frequent. The mode of the variable is "*Agree*," indicating general consensus among users regarding Instagram's effectiveness. A barplot highlights this distribution, showcasing the strong positive sentiment in the dataset.



The **attractiveness** variable captures users' self-perceptions of their profile's appeal rated on a scale from 1 (very bad) to 10 (very good). It has a mean score of approximately 6.7 and a median of 7, indicating a generally positive self-assessment among users. The mode is also 7 with 21.8% of responses. The distribution is relatively symmetrical but slightly skewed toward higher ratings. The interquartile range (IQR) is moderate, reflecting a balanced spread of responses, while the coefficient of variation suggests moderate variability in attractiveness scores across the sample.



Private_d was considered as a dummy variable. It indicates whether a respondent's university is public (0) or private (1). The majority of respondents, 62.6%, attend public universities, while 37.4% are from private institutions. The mode is 0, reflecting the higher representation of public university students in the dataset. A pie chart visually emphasizes this proportional difference, highlighting the skew toward public universities within the sample. This might be a potential structural limitation of the dataset, further affecting the analysis results.



A.2. Outliers Removal

Outliers were identified and removed based on the analysis conducted in Section A.1. To illustrate the impact of this process, two tables are provided: the first summarizes key measures of central tendency and dispersion for numerical variables in the original dataset, while the second presents the same analysis after outlier removal. Variables highlighted in orange indicate a minor reduction in variability and centrality results post-filtering, whereas the one highlighted in green demonstrates a significant improvement in robustness, reflecting the benefits of outlier exclusion. These tables provide a clear overview of the changes made, emphasizing how filtering has enhanced the dataset's structure and reliability in the results of the analysis.

Original	Range	Quartiles (Q1, Q3)	IQR	Median	Mean	Var	Sd	CV	NAs
siblings	(0, 5)	(1, 2)	1.00	1.0	1.292	1.059	1.029	0.796	0
account_num	(0, 5)	(1, 2)	1.00	2.0	1.864	0.977	0.989	0.530	0
story_views	(1, 1357)	(200, 500)	300.00	302.5	355.618	57418.600	239.622	0.674	15
day_time_min	(1, 240)	(41.25, 120)	78.75	60.0	80.164	2769.479	52.626	0.656	5
num_follower	(0, 5057)	(500, 1186)	686.00	757.0	900.690	408445.400	639.097	0.710	1
num_post	(0, 763)	(2, 15.5)	13.50	6.0	24.333	7001.298	83.674	3.439	0
attractiveness	(1, 10)	(5, 8)	3.00	6.0	5.984	5.669	2.381	0.398	0

Filtered	Range	Quartiles (Q1, Q3)	IQR	Median	Mean	Var	Sd	CV	NAs
siblings	(0, 5)	(1, 2)	1.00	1.0	1.292	1.059	1.029	0.769	0
account_num	(0, 3)	(1, 2)	1.00	2.0	1.671	0.534	0.731	0.437	18
story_views	(1, 927)	(195, 460)	265.00	300.0	332.500	41264.740	203.137	0.611	22
day_time_min	(1, 200)	(40, 100)	60.00	60.0	74.600	1942.690	44.076	0.591	13
num_follower	(0, 2134)	(496.5, 1104.5)	608.00	726.5	812.500	209429.700	457.635	0.563	11
num_post	(0, 35)	(1, 12.75)	11.75	5.0	8.396	80.114	8.951	1.066	21
attractiveness	(1, 10)	(5, 8)	3.00	6.0	5.984	5.669	2.381	0.398	0

Code explanation

Key features of the analysis included detection and handling of outliers, to improve reliability, the use of visualizations (i.e. pie charts, bar graphs, histograms and boxplots) and adjustment through data cleaning and binning. The main function used have been:

- table() to compute absolute frequencies.
- prop.table() to calculate relative frequencies.
- pie() to visualize the split of data in dummy and categorical variables.
- ggplot() for generating barplots and histograms to help the visualization of the data. Visualizations also used boxplot().
- na.omit() to exclude missing values.
- Summary statistics like mean(), median(), mode()...
- fivenum () to identify outliers.

Find here attached some lines of code from the **day_time_min** variable, to use as a reference.

```
#DAY_TIME_MIN VARIABLE
day_time_min <- na.omit(database$day_time_min)

mean_value <- mean(day_time_min)
median_value <- median(day_time_min)
mode_value <- as.numeric(names(table(day_time_min)[table(day_time_min) == max(table(day_time_min))]))
range_value <- range(day_time_min)
iqr_value <- IQR(day_time_min)
variance_value <- var(day_time_min)
sd_value <- sd(day_time_min)
cv_value <- (sd_value / mean_value) * 100

breaks <- seq(0, 240, by = 30)  # Bin breaks at 30-minute intervals
bin_labels <- as.character(seq(0,240, by= 30))
bins <- cut(day_time_min, breaks = breaks, include.lowest = TRUE, right = TRUE)

freq_table <- table(bins)
rel_freq_table <- prop.table(freq_table)
```

```
      #Outlier Detection using fivenum
fivenum_stats <- fivenum(day_time_min)
Q1 <- fivenum_stats[2]
Q3 <- fivenum_stats[4]
IQR_value <- diff(fivenum_stats[c(2, 4)])
lower_bound <- Q1 - 1.5 * iqr_value
upper_bound <- Q3 + 1.5 * iqr_value
```

In some variables a few lines of code were added to bin data in order to improve readability of graphs (e.g. **day_time_min** variable) or to replace mistaken entries in the original dataset (**account_num** variable).

Confidence Interval

B.1. 95% CI for the Number of Views

To analyze potential differences in the number of story views between only children and individuals with siblings, we computed 95% confidence intervals for the respective groups:

- Only Child: The 95% confidence interval for the number of story views is (224.34, 370.77).
- Others (Siblings): The 95% confidence interval for the number of story views is (310.28, 368.40).

The confidence intervals for the two groups partially overlap (310–370), indicating no clear significant difference in average story views between only children and those with siblings. However, the only child group shows a lower bound (224.34 vs. 310.28), suggesting greater variability or a potentially smaller average in that group. This finding could reflect underlying differences in engagement levels or behavioral patterns between the groups. The overlap suggests the importance of the interpretation of the results of the t-test to determine if the observed differences are statistically significant.

Furthermore, external factors such as content type or posting frequency could contribute to the observed patterns, deepening the relevance of the analysis.

B2. 99% CI for the Number of Followers

To investigate differences in the number of followers between male and female users, we computed 99% confidence intervals for each group:

- Men: The confidence interval for the average number of followers is (625.33, 825.97).
- Women: The confidence interval for the average number of followers is (788.15, 1026.05).

The intervals do overlap, with the lower bound for women (788.15) being lower than the upper bound for men (825.97). This overlap suggests that the difference in average follower counts between males and females may not be statistically significant at the 99% confidence level. While female users appear to have a higher average follower count, an interpretation of the p-value produced by the t-test is needed to confirm if the difference is statistically significant or not. The current analysis does not provide robust evidence for the difference, as the potential for a Type I error (false positive) still exists.

B.3. 90% CI for the Proportion of Accounts in Italian

To determine the prevalence of Italian-language accounts in the dataset, we computed the proportion of accounts labeled as Italian and derived a 90% confidence interval.

- The proportion of Italian accounts is estimated to fall between 66.8% and 76.4%, with a midpoint of 71.6%.

The relatively narrow confidence interval reflects a high level of precision in the estimate, likely due to a sufficiently large sample size. Given the significant majority of accounts in Italian, it indicates that Italian-language profiles dominate the dataset. Depending on the analysis goal, this could be relevant for interpreting the results.

Explanation of Code

To compute the confidence intervals the following steps were taken: the data was first filtered according to the directions as seen in B.1. Indeed, only_child_views filters story views for respondents with no siblings (siblings == 0) and sibling_views filters for respondents with one or more siblings (siblings > 0).

```
only_child_views <- database_filtered$story_views[database_filtered$siblings == 0]
sibling_views <- database_filtered$story_views[database_filtered$siblings > 0]
```

In particular, in B.3. proportion_italian accounts is calculated by dividing the number of Italian-language accounts by the total number of observations. To assess the reliability of this value, the function qnorm(0.95) was implemented to calculate the margin of error for a 90% confidence level and retrieve the critical z-value corresponding to this confidence level.

```
error_margin <- qnorm(0.95) * sqrt((proportion_italian * (1 - proportion_italian)) / n_total)
```

The margin of error reflects the range of variability around the observed proportion, ensuring that the true proportion of Italian-language accounts in the population is likely to fall within this range.

When dealing with NA values the na.omit() function was used to ensure accurate statistical calculations.

Finally, the t.test() function was employed to compute the confidence intervals for both groups. The confidence level chosen reflects the precision of the function: the higher the percentage, the greater the precision. It reflects varying needs for reliability across analyses.

Hypothesis Testing

C.1. Comparing the Number of Followers Between Men and Women

To assess whether the number of followers significantly differs between men and women, we conducted both an F-test for equality of variances and a pooled-variance t-test, and compared the results with 99% confidence intervals computed in Point B.2.

F-Test for Equality of Variances:

F-statistic: 0.7777, p-value: 0.1778

Since the p-value is greater than both **α = 0.05** and **α = 0.01**, we fail to reject the null hypothesis of equal variances. This justifies the use of a pooled-variance t-test.

Pooled-Variance T-Test:

T-statistic: -3.0713, p-value: 0.0024

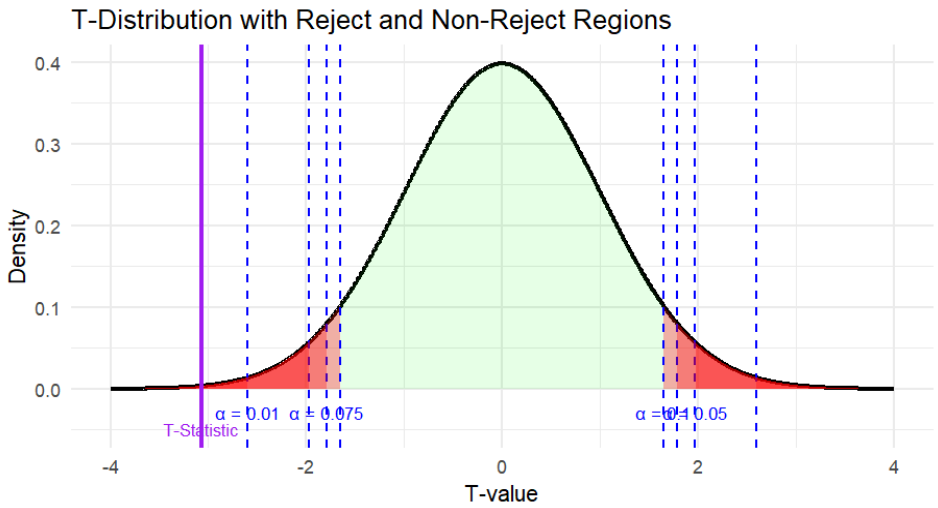
At **α = 0.05**, we reject the null hypothesis, indicating a significant difference in average follower counts between men and women. Similarly, at **α = 0.01**, the p-value remains below the threshold, providing sufficient evidence to reject the null hypothesis even at this stricter significance level.

- **95% Confidence Interval for the Difference in Means**: [-297.85, -65.04]

Since the interval does not include 0, it indicates a statistically significant difference at **α = 0.05**. Moreover, the p-value of 0.0024 confirms this significance at the stricter **α = 0.01** level, providing sufficient evidence that a meaningful difference exists.

For what it concerns the comparison with point B.2. and the 99% confidence interval, recall that:
Men: [625.33, 825.97]
Women: [788.15, 1026.05]

The pooled-variance t-test indicates a significant difference in follower counts at both **α = 0.05** and at **α = 0.01**. However, the 99% confidence intervals in Point B.2 overlap slightly which weakens the evidence for a difference compared to stricter non-overlapping intervals. Nevertheless, the p-value of 0.0024 and the t-test support the robustness of the finding, suggesting that women have significantly higher follower counts than men. This result may reflect gender-based differences in engagement levels or content creation trends.



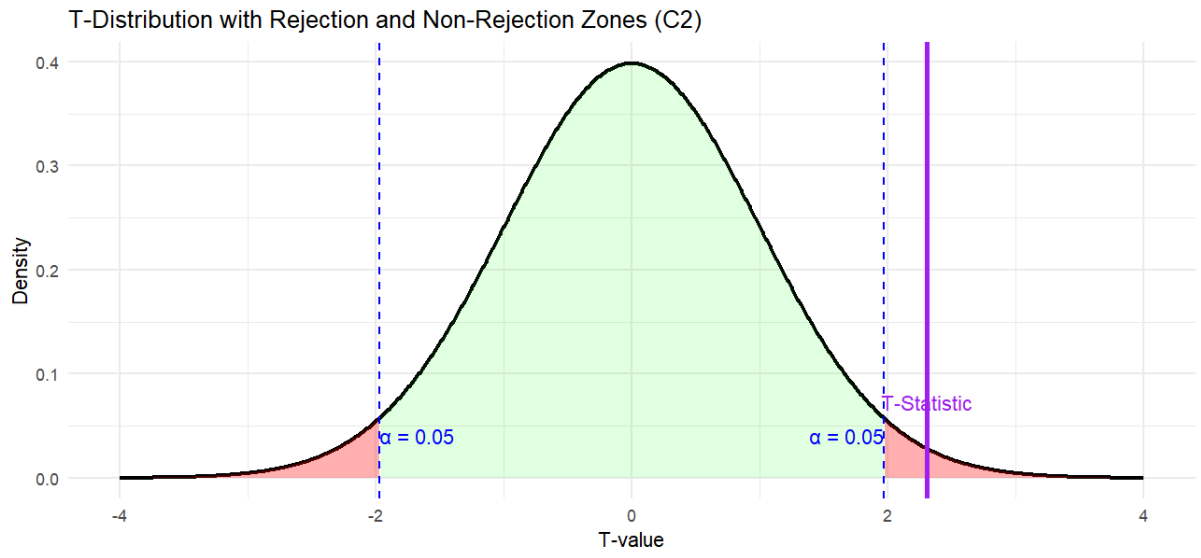
C.2. Analyzing Differences in Story Views Between Private and Public University Students Using P-Values

To investigate whether the number of story views differs significantly between private and public university students, we first conducted an F-test to compare the variances of the two groups. The F-test yielded a **p-value of 0.8055**, which is well above the 0.05 significance level. This suggests there is no significant difference in the variances of story views between the two groups, justifying the use of a pooled-variance t-test.

Next, a two-sample t-test was conducted assuming equal variances. The t-test produced a **t-statistic of 2.301** and a **p-value of 0.02233**, which is less than 0.05. This indicates that the null hypothesis (which posits no difference in means) is rejected, meaning there is a statistically significant difference in the average number of story views between private and public university students.

The 95% confidence interval for the difference in means ranged from 9.23 to 119.38, and since this interval does not include 0, it further supports the conclusion of a significant difference in story views between the two groups. The mean number of views for private university students was 372.69, while the mean for public university students was 308.38, showing a higher average for private university students.

In conclusion, the results suggest that university type (private vs. public) significantly influences the number of story views, with private university students having a higher average number of views compared to their public university counterparts.



C.3. Evaluating Differences in Daily Instagram Usage Between English and Italian Speakers at $\alpha = 0.01$

The analysis comparing the average daily time spent on Instagram between English and Italian speakers reveals the following results:

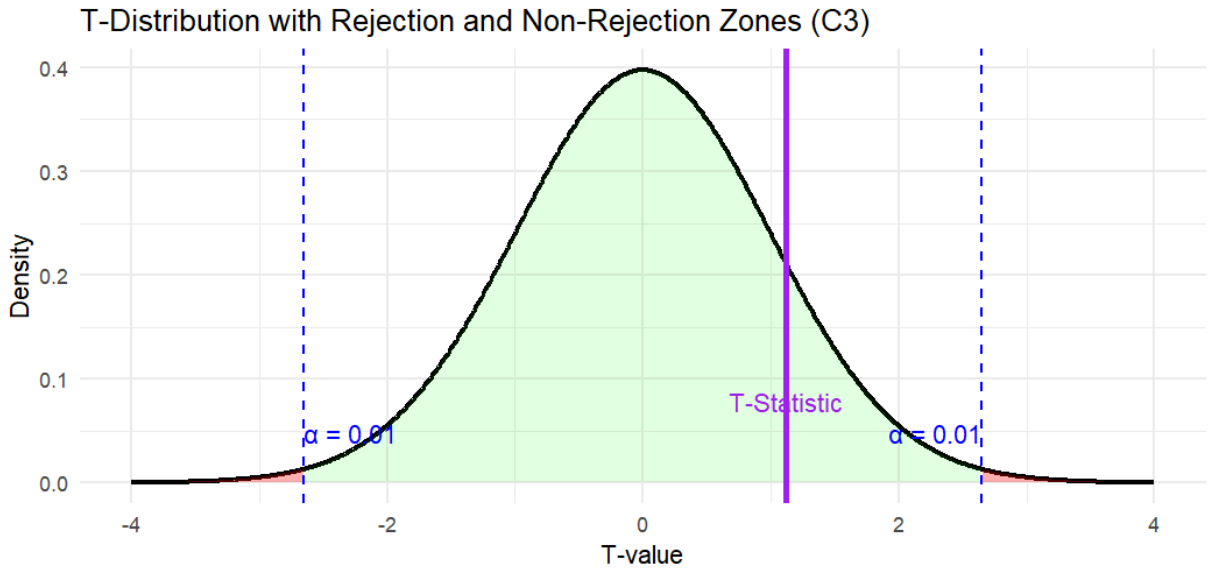
The F-test for equality of variances returned a **p-value of 0.01477**, suggesting that the variances between the two groups are significantly different. Indeed, p-value is compared to $\alpha = 0.05$ and being the p-value less than the level of significance taken into account, Welch's t-test, which accounts for unequal variances, is applied.

The results from Welch's t-test showed a t-statistic of 1.121 and a p-value of 0.2662, which is greater than the significance level of $\alpha = 0.01$. Therefore, we fail to reject the null hypothesis, indicating that there is no statistically significant difference in the average time spent on Instagram between English-speaking (mean = 82.26 minutes) and Italian-speaking (mean = 73.04 minutes) users.

The 95% confidence interval for the difference in means is [-7.19, 25.63], which includes 0, further supporting the conclusion that there is no significant difference in Instagram usage between the two language groups.

These results suggest that language preference (English vs. Italian) does not significantly affect the average daily time spent on Instagram.

The graph of the t-distribution illustrates the rejection and non-rejection zones, with the t-statistic falling well within the non-rejection region, reinforcing the conclusion.



Explanation of Code

The purpose of the code in the points C1, C2 and C3 was, respectively, to:

- Determine whether the number of followers differs significantly between male and female users
- Test if the number of story views differs significantly between students at private and public universities
- Analyze whether daily Instagram usage differs significantly between English and Italian speakers

To do so, in point C1 follower counts were extracted for males (sex == 'M') and females (sex == 'F') using na.omit() to exclude missing values. In point C2, story views for private (private_d == 1) and public (private_d == 0) universities were separated, excluding missing values. In point C3, daily usage times for English (language == 'English') and Italian (language == 'Italian') speakers were extracted, ensuring sufficient observations.

The command var.test() was then used to determine whether variances between the two groups are equal. If they were (var.equal = TRUE), then a pooled-variance t-test was used. Otherwise (var.equal = FALSE), a Welch's t-test was run.

Tests at different α levels, which depended on what was asked in the assignment, have been conducted on the p-value that was given by the t-test.

For visualization purposes a t-distribution plot was then used to highlight rejection and non-rejection zones with critical values and the observed t-statistic.

Linear Regression

D.1. Linear Regression

The simple linear regression analysis investigates the relationship between the number of followers and story views, as outlined in question D.1 of the assignment. The regression model estimates that the number of story views is equal to:

$$132.29 + 0.244 \times Num\ Followers.$$

This equation suggests that for every additional follower, the number of story views increases by approximately 0.244 on average. The intercept value of 132.29 represents the predicted story views for a profile with no followers. Both the intercept and slope coefficients are highly significant ($p < 0.001$), as demonstrated by the t-tests on the model coefficients.

The model achieves an R^2 of 0.3171, meaning it explains approximately 31.7% of the variability in story views, which is a moderate fit for the data. However, the residual standard error of 160.6 indicates that there is considerable variability in story views not accounted for by the number of followers alone.

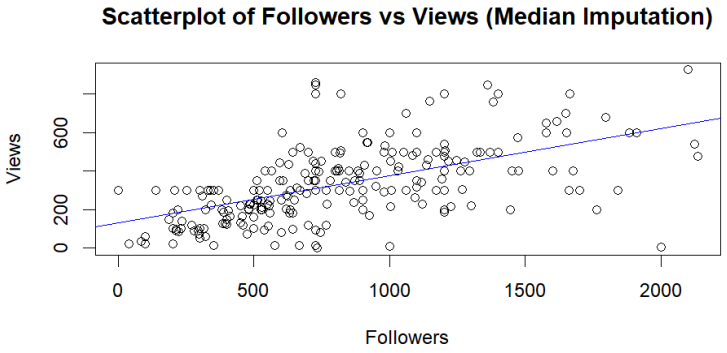
Residual diagnostics reveal some violations of key assumptions. The scatterplot of followers versus story views demonstrates a generally linear relationship, as indicated by the correlation coefficient *r,* supporting the use of linear regression, though minor deviations suggest potential non-linearity.

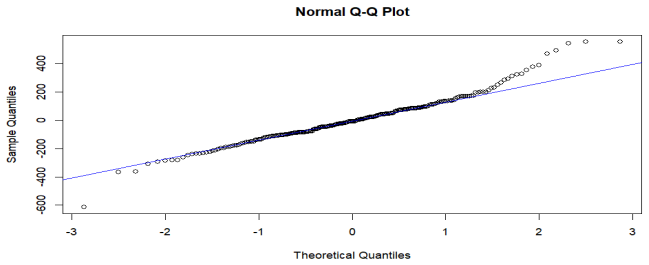
The normality of residuals is assessed through a Q-Q plot and the Shapiro-Wilk test (**$W = 0.9668, p < 0.001$**), which strongly rejects the null hypothesis of normality, indicating that the residuals deviate significantly from a normal distribution.

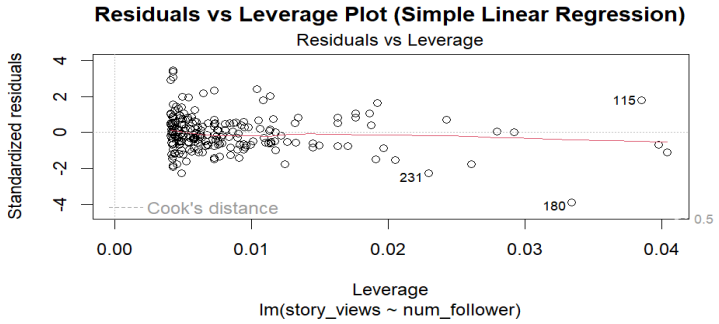
The residuals vs. fitted values plot reveals a non-random pattern, suggesting mild heteroscedasticity, as the spread of residuals appears uneven across fitted values. A scale-location plot further confirms some variation in residual spread across fitted values.

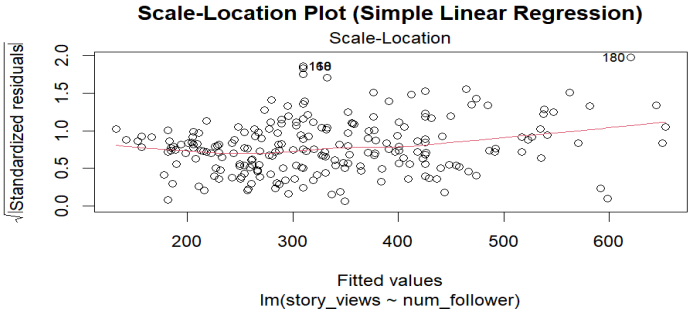
Finally, the residuals vs. leverage plot highlights no extreme outliers or points with undue influence, but the zoomed-in version of this plot underscores that most data points lie within a low-leverage range.

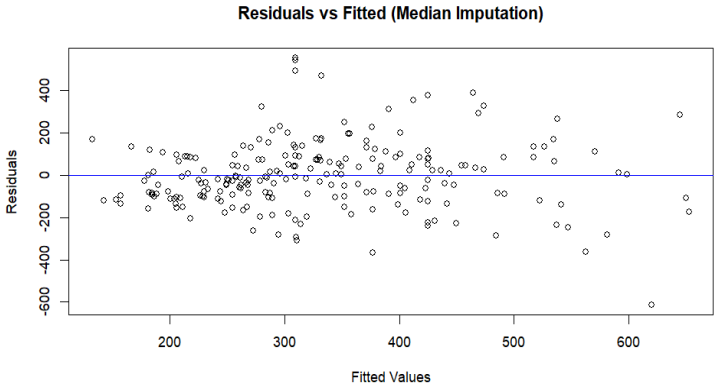
In conclusion, the simple linear regression model provides a reasonable explanation of the relationship between followers and story views, capturing a meaningful portion of the variability. However, the normality and homoscedasticity assumptions are not fully met, and potential refinements, such as transformations of the dependent variable, could improve model performance. Otherwise, more advanced regression techniques, such as weighted least squares or generalized linear models, could be implemented to better meet the assumptions and improve prediction accuracy.

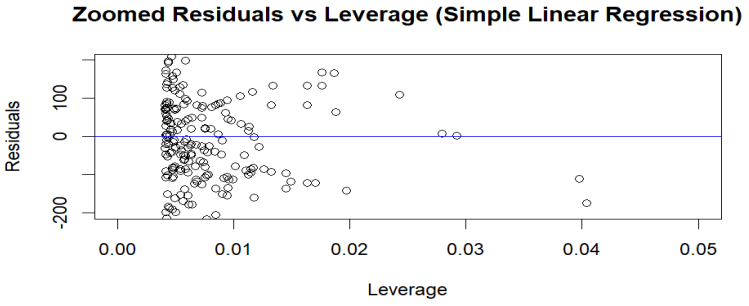












D.2. Multiple Linear Regression

The multiple linear regression analysis builds upon the simple model by incorporating additional predictors, as required by question D.2 of the assignment. The estimated regression equation is:

$$Story\ Views = 80.90 + 0.237 \times Num\ Followers + 25.01 \times Sex(M) + 24.60 \times Account\ Num + 0.53 \times Num\ Post - 0.029 \times Day\ Time\ Min.$$

The strongest predictor remains the number of followers **($p < 0.001$)**, with each additional follower associated with a 0.237 increase in story views. Interestingly, none of the other predictors are statistically significant contributors to the model, as indicated by their respective p-values ($p > 0.05$).

The R^2 value of 0.3274 indicates that 32.7% of the variability in story views is explained by the model, representing only a marginal improvement over the simple regression model ($R^2 = 0.3171$). The ANOVA test comparing the simple and multiple regression variances yields an F-statistic of 0.9135 **($p = 0.4567$)**, indicating that the additional predictors do not significantly enhance the model's explanatory power.

The residual standard error (160.7) remains nearly unchanged from the simple model, suggesting that the added predictors do not substantially reduce prediction errors.

Residual diagnostics for the multiple regression model reveal similar issues to those observed in the simple model.

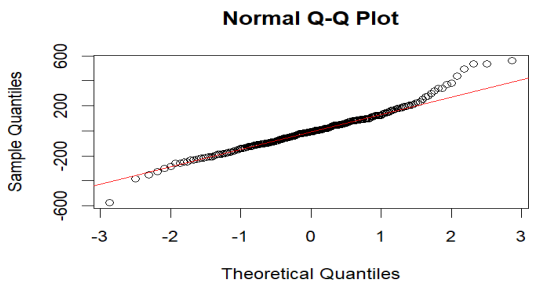
A scatterplot of residuals vs. fitted values indicates that the relationship between predictors and story_views is approximately linear, though slight curvature hints at possible nonlinearities in the data.

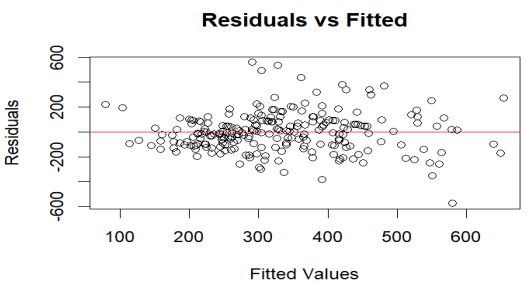
The Q-Q plot suggests mild deviations from normality in the residuals, particularly in the tails.

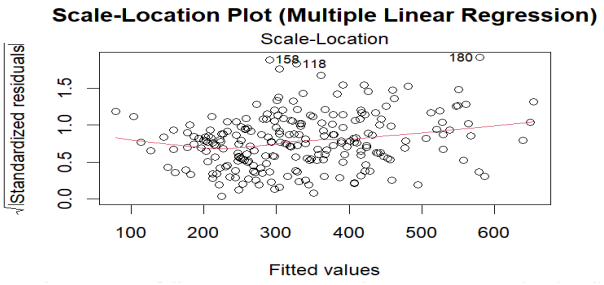
The residuals vs. fitted values plot displays slight patterns, hinting at heteroscedasticity. The scale-location plot supports this finding, showing variation in residual spread across fitted values.

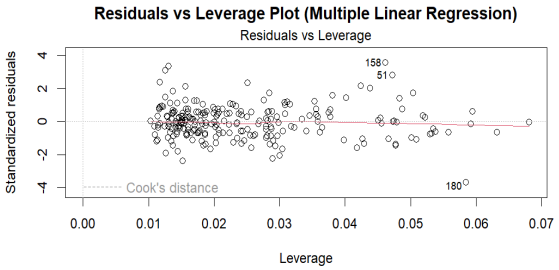
The residuals vs. leverage plot, along with its zoomed-in version, confirms that while there are no points with undue influence, most data points are concentrated in the low-leverage region.

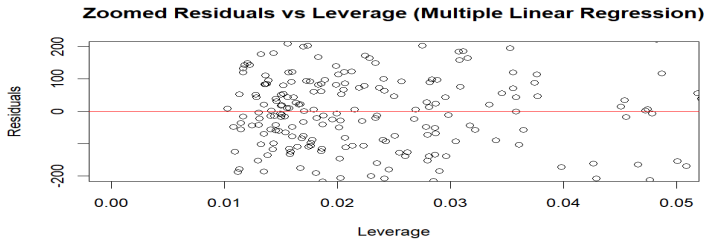
Overall, the multiple regression model provides only a slight improvement in explaining the variability in story views compared to the simple model. The dominant role of the number of followers as a predictor is reaffirmed, while the lack of significance of other predictors suggests limited additional explanatory power. Despite improvements in explanatory power, the residual analyses indicate persistent violations of normality and homoscedasticity assumptions.











Explanation of Code

In order to avoid problems given by the different length of vectors of the variables taken into account when running both the simple and multiple linear regression, data imputation was implemented.

In the simple linear regression model, missing values in story_views and num_follower are replaced with given values to address skewness and ensure a complete dataset. Given that the median is based solely on the position of the observations in a sorted dataset, extreme values, whether very high or very low, do not affect its calculation. Unlike the mean, which can be heavily susceptible to outliers, the median remains stable and accurately reflects the central tendency of a dataset. Therefore, our values of choice for the imputation was the respective median (comprehensive of the outliers to reduce information loss) of each variable. This action left the dataset length to its original one: 243.

```
for (col in columns_to_impute) {
  database_imputed[[col]][is.na(database_imputed[[col]])] <- median(database_imputed[[col]], na.rm = TRUE)
}

lm_simple_imputed <- lm(story_views ~ num_follower, data = database_imputed)
summary(lm_simple_imputed)
```

Then a series of checks were performed, such as the linearity check, normality of residuals (with a Q-Q plot along with the shapiro test as illustrated below), homoscedasticity check and a residuals vs leverage plot.

```
      #Check normality of residuals
qqnorm(residuals(lm_simple_imputed))
qqline(residuals(lm_simple_imputed), col = "blue")
shapiro.test(resid(lm_simple_imputed))

      # Residuals vs Leverage plot
plot(lm_simple_imputed, which = 5, main = "Residuals vs Leverage Plot (Simple Linear Regression)")
```

The same was applied to the multiple linear regression model incorporating additional predictors (sex, account_num, num_post, day_time_min) alongside num_follower to explain story views. Each NA was found and substituted with the respective variable's median..

```
for (col in columns_to_impute) {
  database_imputed[[col]][is.na(database_imputed[[col]])] <- median(database_imputed[[col]], na.rm = TRUE)
}

lm_multiple_imputed <- lm(story_views ~ num_follower + sex + account_num + num_post + day_time_min, data = database_imputed)
summary(lm_multiple_imputed)
```

Finally, the two models can be compared using ANOVA.

```
anova(lm_simple_imputed, lm_multiple_imputed)
```

To complete the analysis the previously mentioned checks are performed again: Q-Q plot and Shapiro-Wilk test for normality, residuals vs. fitted and scale-location plots for homoscedasticity, residuals vs. Leverage plot to identify high-influence points.

What would change without imputation?

Median imputation was applied to ensure data completeness and mitigate the effects of skewness. However, alternative approaches to handling missing data were explored during the modeling process. In the initial model, no imputation was performed. Without imputation, the total number of observations was reduced to 178, compared to the current 243. This substantial reduction in sample size limited the reliability and informativeness of the model.

In a subsequent iteration, we sought to improve the model by applying imputation selectively. The simple linear regression was computed as in the first model and it utilized 214 observations. While, to enable a direct comparison through ANOVA, the multiple linear regression dataset underwent median imputation to ensure equal lengths between the two models.

This updated approach led to improvements in the R-squared values, but a significant portion of the variability remained unexplained. For the simple linear regression, the R-squared increased from 31.71 to 45.64, with a reduction in the standard error from 160.6 to 144.1. For the multiple linear regression, the R-squared improved from 32 to 45. While these adjustments provided marginal gains in significance, the overall improvement was modest. Hence, we chose to keep in the code only the linear regression with median imputation.

Prediction

E.1. Predicted Story Views for the Female Median Account

To estimate the expected number of story views for a female account with median characteristics, we used the multiple linear regression model developed earlier. The predictors included in the model were the number of followers, gender, number of accounts, number of posts, and daily time spent on Instagram. For this prediction, all variables were set to their respective medians for female accounts within the dataset. The specific median values used were:

- **Median number of followers**: median_followers

- **Median number of accounts**: median_account_num
- **Median number of posts**: median_num_post
- **Median daily time spent on the app (in minutes)**: median_day_time_min

The prediction was generated using the predict() function in R, with the lm_multiple_imputed regression model applied to a data frame containing these median values for a female user. The resulting point estimate for the predicted story views was approximately 303.49. This aligns with the assumption that story views are influenced by a combination of these factors, as modeled in the regression.

The relatively moderate number of story views implies that engagement is proportional to the median account's characteristics, particularly the number of followers. This suggests that follower count and content activity significantly contribute to story visibility. Notably, this prediction highlights that accounts with typical median characteristics are expected to perform moderately well, though individual deviations may occur based on unmodeled variables or outlier effects.

The accuracy of this prediction depends on the validity of the regression model: it is to be checked if the dataset is well balanced and representative of the population accounts. However, some limitations are intrinsic to the analysis, as the model does not include potential additional variables such as content quality, hashtags, specific times of posting, etc.

E.2. Confidence Interval for the Predicted Story Views

To assess the reliability of the predicted value, a 95% confidence interval for the prediction was calculated. The confidence interval provides a range within which the true average story views for a female account with median characteristics is expected to fall. The interval obtained was:

- **Lower bound**: 267.99
- **Upper bound**: 338.98

This relatively narrow range indicates that the regression model provides a reasonably precise estimate for the number of story views. The narrowness of the interval suggests that the model captures the underlying relationships in the data effectively, despite minor violations of normality and heteroscedasticity noted during residual diagnostics.

The precision of this confidence interval depends on the explanatory power of predictors (*num_follower, sex, account_num)* in the regression model. Significant predictors generally lead to narrower confidence intervals: the dominant one is the number of followers, while the other variables contribute to fine_tuning the prediction. While this interval is specific to the female median account, it may overlap with intervals for other groups or account types and not generalize that are highly active or inactive.

The model assumes that the relationships between predictors and the dependent variable (story views) are linear and that the predictors adequately capture the variability in story views. While these assumptions were tested earlier, mild violations were noted, particularly regarding residual normality and homoscedasticity. Thus, while the confidence interval is narrow, the

results should be interpreted with some caution, particularly for accounts that deviate significantly from the median characteristics used in the prediction.

Prediction median and mean in comparison with other variables'

The predicted story views for a median female account are approximately 303.49, closely aligning with the dataset's median of 302.5 but lower than the mean of 332.5 (after outlier removal). This highlights that predictions based on median values reflect typical account behavior and are less influenced by outliers compared to mean-based estimates.

Similarly, the prediction uses the median follower count of 726.5, which is lower than the dataset's mean of 812.5 due to a skewed distribution. For posts, the predicted median of 6 aligns with typical activity, contrasting the mean of 24.33, which is inflated by extreme values. Median daily Instagram usage of 60 minutes also reflects typical patterns, as the dataset mean of 74.6 is impacted by heavy users.

Overall, predictions based on medians provide a robust representation of central trends, avoiding distortions caused by variability and outliers, while confidence intervals indicate model reliability for typical accounts.

Explanation of Code

The tasks to fulfill for points E1 and E2 were respectively:

- To estimate the expected number of story views for a female user with median characteristics
- To calculate a 95% confidence interval for the predicted story views of the female median account

In the first point, the median values of key predictors (num_follower, account_num, num_post, day_time_min) from the filtered dataset were computed and a data frame representing a female user with these median values was created. The multiple linear regression model (lm_multiple_imputed) was then used to predict story views (predict(lm_multiple_imputed, newdata=female_median_data))

In the second point, the predict() function was used with the interval = "confidence" parameter to compute the lower and upper bounds of the interval.

Logistic Regression Analysis

F.1. Predicting if an Account Belongs to a Student Attending a Private University

A logistic regression model was employed to assess whether the number of followers, number of posts, and number of views could predict whether an account belongs to a student attending a private university.

Initial Logistic Regression Model Results

The initial logistic regression model, based on raw variables (number of followers, number of posts, and story views), yielded the following results. The **intercept estimate** of -1.3618 (**p-value < 0.001**) was statistically significant, indicating a strong baseline for predicting private university affiliation when all predictors are at their mean values. However, the estimates for the **number of followers** (0.0003937, **p-value = 0.386**) and **number of posts** (0.0076846, **p-value = 0.659**) were not statistically significant, suggesting that these variables had minimal influence on the prediction. The estimate for the **number of story views** (0.0017396, **p-value = 0.093**) was marginally significant, indicating a weak but potentially meaningful association with the outcome.

So, it can be said that the initial model revealed that its **predictive power** was limited. **McFadden's pseudo R-squared** value of 0.036 suggested that the model explained only a small fraction of the variance in the data. The **odds ratios** for all predictors were close to 1, reinforcing that the effect sizes were negligible. The model's **accuracy** was 61.31%, with **specificity** at 26.25%, which points to difficulties in correctly identifying non-private accounts. However, the model achieved a **sensitivity** of 84.87%, demonstrating strong performance in identifying private accounts.

Second Model with Feature Engineering

To improve upon the initial model, **feature engineering** was employed by introducing **log-transformed variables** and **interaction terms** to better capture non-linear relationships and reduce data skewness. **Log(number of followers)** (estimate = 1.6553, **p-value = 0.002**) and the **follower engagement ratio** (estimate = 3.2834, **p-value = 0.035**) emerged as statistically significant predictors. Both showed positive relationships, indicating that a higher **follower count** and greater **engagement** were associated with a greater likelihood of an account belonging to a private university student. In contrast, **log(story views)** (estimate = -0.7841, **p-value = 0.061**) showed marginal significance in the negative direction, suggesting that fewer views relative to followers might indicate a private university affiliation. Neither **log(number of posts)** nor **story views per post** were significant predictors.

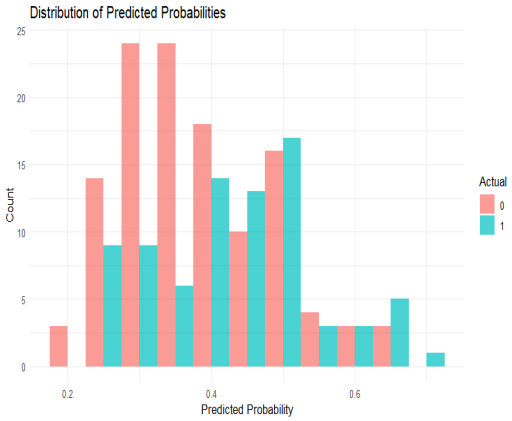
Interpretation of the model coefficients revealed that a **one-unit increase** in **log(number of followers)** increased the odds of the account coming from a private university student by a factor of **5.23**. The **follower engagement ratio** also significantly influenced the likelihood of an account belonging to a private university, increasing the odds by a factor of **26.67**. However, a decrease in probability can be drawn by the marginally negative coefficient for **log(story views)**. The higher **story views** relative to followers, the lower the probability of the owner of the account being a private university student.

Despite the feature engineering efforts, the model's overall performance showed only slight improvement. The **McFadden's pseudo R-squared** increased to 0.0558, indicating a marginal

improvement in explanatory power. The model's **accuracy** rose to **62.31%**, with a confidence interval between **55.18%** and **69.07%**. The **sensitivity** increased to **86.55%**, improving the model's ability to identify private university students, but **specificity** remained low at **26.25%**, reflecting challenges in correctly identifying non-private accounts.

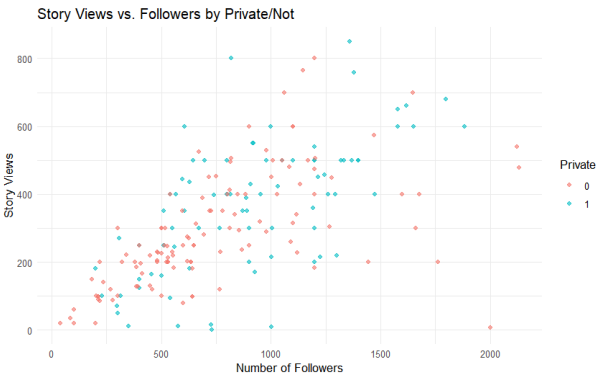
While the logistic regression model with feature engineering provided some improvements, its predictive power remains limited. The follower-engagement ratio and log(number of followers) emerged as meaningful predictors, yet the low specificity and modest McFadden's pseudo R-squared suggest that the model still struggles with accurate prediction. Future research may focus on incorporating additional features or exploring more complex models to further enhance predictive performance and address the issues with classification accuracy.

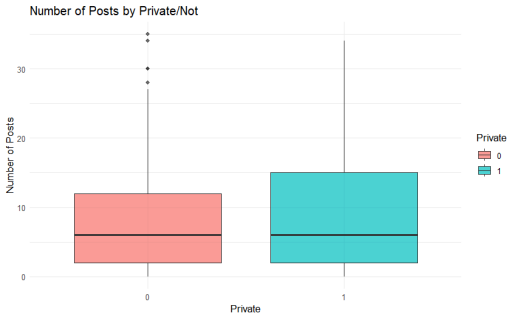
Graphical Analysis



The histogram of predicted probabilities showed the distribution of predicted probabilities for both private and non-private accounts, indicating that the model tends to predict more private accounts than non-private ones.

The **scatter plot of story views vs. followers**, colored by the private variable, revealed that private accounts generally have a higher number of followers and more story views compared to non-private accounts, with some overlap. This visual corroborates the model's findings, suggesting that both the number of followers and story views have a potential impact on the likelihood of an account being private.





Finally, the **boxplot of number of posts by private status** showed that private accounts tend to have a higher median number of posts, with some variance within each category. This insight is consistent with the model's findings that num_post

was not a significant predictor but may still be associated with private status at a descriptive level.

In conclusion, the logistic regression analysis reveals some limited predictive power for determining if an account belongs to a private university student based on the number of followers, posts, and views. While the model with feature engineering shows some improvement, it still only explains a small fraction of the variance. Further exploration with additional features or more sophisticated models might yield better results. Additionally, the graphical analyses provide valuable insights into the relationships between the predictors and the outcome, supporting the interpretation of the logistic regression results.

Explanation of Code

In the initial logistic regression, the variables num_follower, num_post, and story_views were used as predictors in a logistic regression model (glm()).

```
logistic_data <- database_filtered %>%
  filter(!is.na(num_follower) & !is.na(num_post) & !is.na(story_views))

logistic_model <- glm(private_d ~ num_follower + num_post + story_views,
                      data = logistic_data, family = binomial)
summary(logistic_model)
```

The Pseudo R-squared (pR2()) and a confusion matrix evaluated model performance. Predictions were reintroduced into the dataset.

```
pseudo_r2 <- pR2(logistic_model)
cat("\nPseudo R-squared:\n")
print(pseudo_r2)
```

With feature engineering, Log transformations (log1p()) were applied to predictors (num_follower, num_post, story_views) to reduce skewness.

```
logistic_data <- logistic_data %>%
  mutate(
    follower_engagement_ratio = ifelse(num_follower > 0, story_views / num_follower, 0),
    story_views_per_post = ifelse(num_post > 0, story_views / num_post, 0),
    log_num_follower = log1p(num_follower),
    log_num_post = log1p(num_post),
    log_story_views = log1p(story_views)
  )
```

New features were introduced such as:

- follower_engagement_ratio (story views per follower).
- story_views_per_post (views divided by posts).

In the enhanced logistic regression, the updated model included transformed and new interaction variables. Moreover, odds ratios (exp(coef())) were computed to interpret the impact of predictors on private account likelihood.

```
exp_coef <- exp(coef(logistic_model))
cat("\nOdds Ratios for Logistic Regression with New Features:\n")
print(exp_coef)
```

For visualization purposes, an histogram of predicted probabilities, a scatter plot (story views vs. followers by private_d), and a boxplot (posts by private_d) were done to analyze variable distributions.

Conclusion

This report comprehensively examines factors influencing Instagram Story performance among university students, leveraging robust statistical tools such as confidence intervals, hypothesis testing, and regression analysis. It reveals significant differences in follower counts between genders, with women generally commanding higher engagement levels, and a notable disparity in story views based on university type, favoring private institutions. However, daily Instagram usage and language preference showed limited impact on engagement metrics, suggesting other unexamined factors at play.

Regression analyses highlighted the dominant influence of follower count on story views, reaffirming its critical role in predicting user engagement. However, the limited explanatory power of the models and persistent violations of key assumptions, such as normality and homoscedasticity, underscore the need for more advanced modeling techniques. Logistic regression further demonstrated the potential of engineered features like follower engagement ratios, though challenges in classification accuracy, particularly specificity, remain.

Future research should explore integrating richer datasets, including external metrics like Instagram's algorithmic preferences or visual content analysis, to refine engagement predictions. Additionally, adopting machine learning models could enhance predictive accuracy, offering deeper insights into the nuanced dynamics of Instagram Story performance and overcoming limitations such as data skewness, reliance on self-reported metrics, and exclusion of qualitative factors like content quality or posting time.

Appendices

Full R code

Database