Statistics 2 Project

Group Project

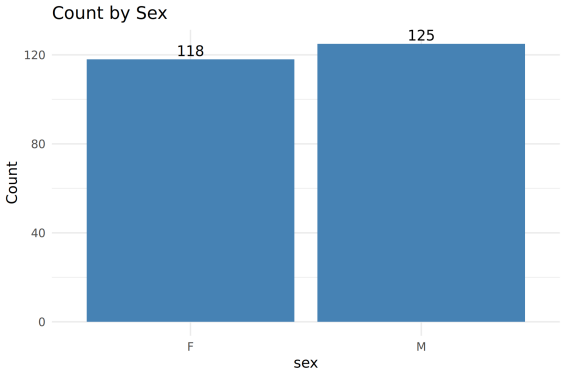
Pieter Christiaan Rotteveel **3293414**
Luca Busetti **3283258**

February 4, 2026

A.Statistics of the Sample

A.1 Variable analysis

Variable analysis: "Sex"



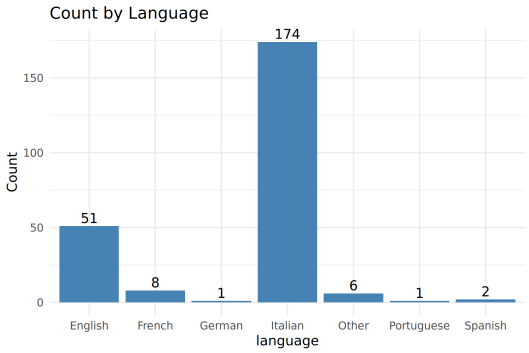
Bar Chart

category	absolute frequency	relative frequency (in percentages)
Female (F)	118	48.6
Male (M)	125	51.4
Total	243	100.0

Frequeny Table

Analysis:
"Sex" is a nominal categorical variable. The distribution can be seen to be almost perfectly heterogeneous. Almost as many male students as female students were amongst those sampled from. The relative frequencies of 48.6 and 51.4 indicate an almost perfect division. In fact, the modal category exceeds the other category by only 7 observations.

Variable analysis: "Language"



Bar Chart

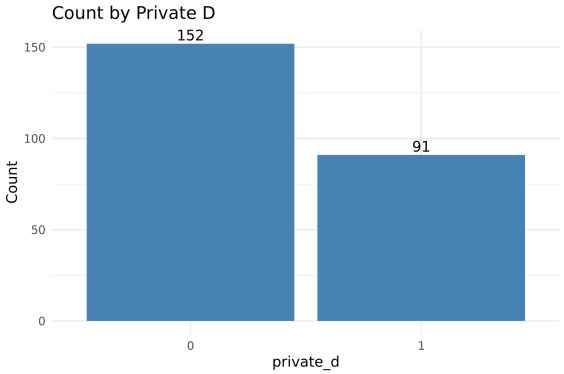
Category	Frequency	Percentage (%)
Italian	174	71.6
English	51	21.0
French	8	3.3
Other	6	2.5
Spanish	2	0.8
German	1	0.4
Portuguese	1	0.4
Total	**243**	**100.0**

Frequeny Table

Analysis:
"Language" is a nominal categorical variable as well. The distribution of this variable is unimodal. The most represented category is Italian (Mode: 174), followed by English, which is detected in 21% of the cases, the remaining languages sum up to 7.4%, which shows the somewhat limited language

diversity within the sample. The modal category is over 3 times as frequent as the runner - up, in fact, the frequency is highly concentrated in that one category. As 5 out of 7 categories have a frequency below 5%, we can conlcude that there is very limited dispersion accross categories.
Different from what was computed for the previously examinated variable, the distribution is very homogeneous in this case. Variability is minimal. There is a very low level of dispersion, with the two most frequent categories making up of more than 90% of the total observations.

Variable analysis: "Private vs Public University"



Bar Chart

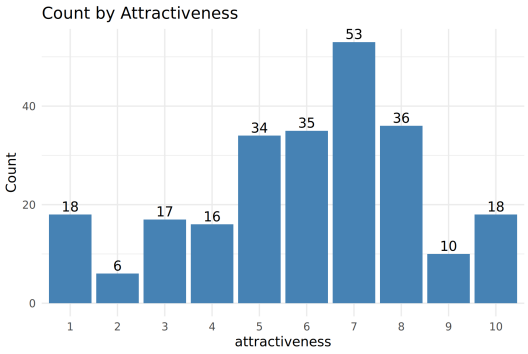
Category	absolute frequencies	relative frequencies in percetanges
No Public	152	62.6
Yes Private	91	37.4
Total	243	100.0

Frequeny Table

Analysis:
"Private-D" is the last nominal categorical variable in the dataset. The modal category is Public University, clearly identifiable due to the difference in frequencies. 62.2% of the individuals attend a public university, whereas 37.4% of the individuals attend a private one. When compared to the other binary distribution (the variable "sex"), we can observe how the distribution is not symmetric, in this case.

There is moderate level heterogeneity, as one category is notably more prevalent, with the minor group remaining sizeable.

Variable analysis: "Attractiveness"



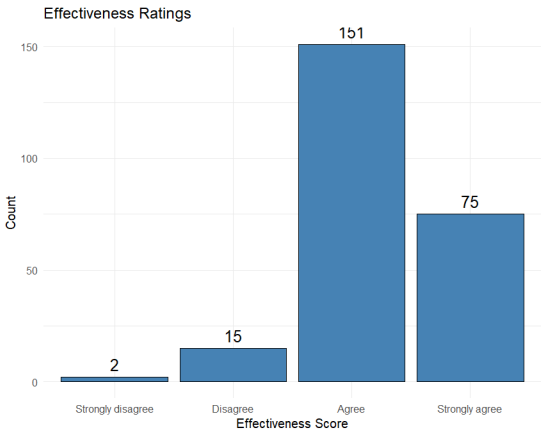
Bar Chart

Category	Absolute frequencies	Relative frequencies in percentages
1	18	7.4
2	6	2.5
3	17	7.0
4	16	6.6
5	34	14.0
6	35	14.4
7	53	21.8
8	36	14.8
9	10	4.1
10	18	7.4
Total	243	100.0

Frequeny Table

- **central tendency analysis:** "Attractiveness" is an ordinal categorical variable. This makes it not reccomendable to use the mean for an analysis of central tendency. This is the case, as the mean relies on the assumption that distances between successice values are homogeneous. For variables such as the underlying one, this assumption cannot be justified. We therefore assess central tendency through the Median and the Mode. The distribution is unimodal, with the mode being 7 (observed 53 times, 21.8%). The Median is 6, which means that half of the students have ranked themselves as intermediately to highly attractive. The bulk of the observations lies between 6 and 8, which indicates that typical attractiveness self perception lies in the upper - middle part of the scale.
- **non central tendency analysis:** The range of scores recorded is very wide, as all ratings have been recorded, indicating a strong variability in self perception about attractiveness by students. It is observable how extreme scores, such as 10 or 1 are equally frequent (observable 7.4 % of the time) and more commonly observable with respect to low intermediate or high intermediate categories such as 2,3,4 and 9. The distribution appears to be slightly left skewed.
- **analysis of dispersion:** The Interquartile Range (IQR) is: Q1 = 5 - Q3 = 8 = 3, which shows how all the observations are concentrated in a quite narrow interval. From the quartiles, we have ulterior confirmation of the fact that students typically rank themselves as medium to high when it comes to attractiveness. The relative frequencies of 7.4% of more extreme evaluations indicate that there is significant polarization in this distribution.

Variable analysis: "Effectiveness"



Bar Chart

Category	Absolute Frequencies	Rel. Frequencies (perc.)
Strongly disagree	2	0.8
Disagree	15	6.2
Agree	151	62.1
Strongly agree	75	30.9
Total	243	100.0

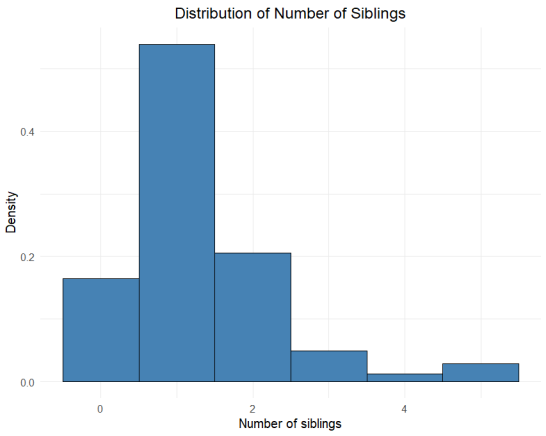
Frequency Table

- **analysis of central tendency:** The variable "Effectiveness" is an ordinal categorical variable. Its distribution is unimodal, with the modal category being "agree", which has been recorded on 151 occasions and therefore 62.1% of the time. The Mean is not meaningful in this case since it relies on the assumption that numerical distances between values are equal. Together, the categories "agree" and "strongly agree" make up of more than 90% of total observations, which indicates a strong central tendency towards agreement.
- **analysis of non central tendency:** The distribution of the variable "effectiveness" is highly concentrated towards the upper categories, as the vast majority of users gave "agree" or "strongly agree" as a response. The distribution is strongly skewed in favor of positive evaluations, with a clear clustering of responses at a high level of perceived effectiveness.
- **analysis of dispersion:** Dispersion is very limited, as the vast majority of gathered observations are concentrated in just two categories. The fact that "strongly disagree" and "disagree" have, cumulatively, been observed only 7% of the time, indicates that responses are not well spread over the scale. In fact, also the bar chart shows a narrow spread, with limited presence at the lower end of the scale. The same logic of the Mean can be applied here. The fact that we are using ordinal measurement, precludes us from using measures such as the varianes to assess dispersion, as it relies on the assumption that the numerical distances between observed valyes are homogeneous. The dispersion is, instead, measured through the distribution of frequencies.

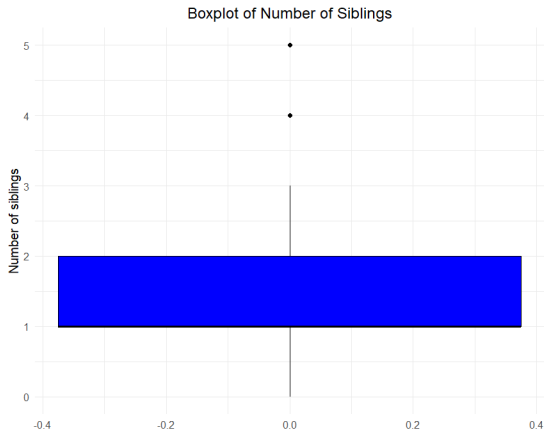
Variable analysis: "Siblings"

Descriptive Statistics: Number of Siblings

Variable	Min	Max	Q1	Median	Mean	Q3	Variance	SD
Siblings	1	5	1	1	1.29	2	1.06	1.03



Bar Chart



Boxplot

- **analysis of central tendency:** The variable "Number of Siblings" is the first discrete numeric variable put under analysis, in the dataset. It is unimodal, with the mode being 1, which can also be observed in the bar graph, as 1 is the most observed number of siblings (131 observations, 53%). The typical student has, therefore, 1 sibling. The Median is 1 and the Mean is 1.29, which is due to the fact that a small number of students has a relatively high amount of siblings (5 observed 7 times, 2%). As the Mean is notoriously sensitive to outliers, we have that those small values pull the Mean slightly up. The fact that the Mean is greater than the Median indicates a slightly right skewed distribution.

- **analysis of non central tendency:** As already mentioned, the distribution is right skewed. Most observations are clustered around lower values, with only a small number of students reporting higher values. This creates a right tail, which,in turn, explains why the Mean is larger than the Median.

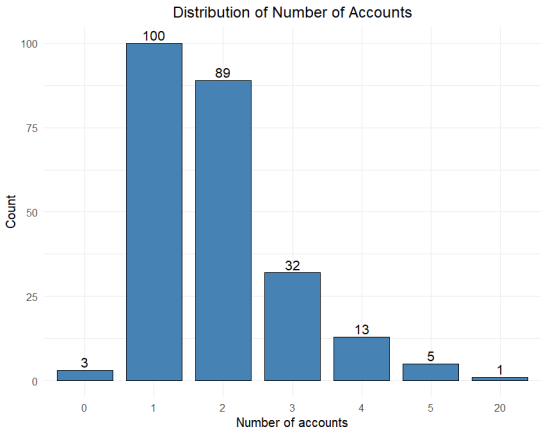
- **analysis of dispersion:** The range (Max-Min=5-0=5) shows that the number of siblings recorded varies from 0 to 5. This might give the impression of a spread out distribution. However, this value was obtained, due to the presence of a few outliers, which inflate the measure. Students with 4 or more siblings constitute the outliers, in this case. Only 4% of students have that number of siblings. The Interquartile Range (Q3-Q1=2-1=1) shows that the central 50% of students have between 1 and 2 siblings, which shows a very limited dispersion around the central part of the distribution. The Standard Deviation (1.03) tells us

that, on average, the number of siblings deviates from its mean by approximately one sibling. Given the value of the Mean (1.29), this reflects a low to medium level of variability. As the Standard Deviation is also sensitive to extreme values, we conclude that the number was inflated by the right tail.

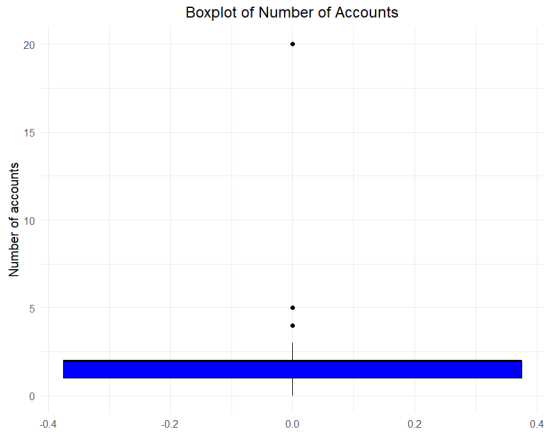
Variable analysis: "Number of accounts"

Descriptive Statistics: Number of Accounts

Variable	Min	Max	Q1	Median	Mean	Q3	Variance	SD
Number of accounts	0	20	1	2	1.94	2	2.33	1.53



Bar chart



Boxplot

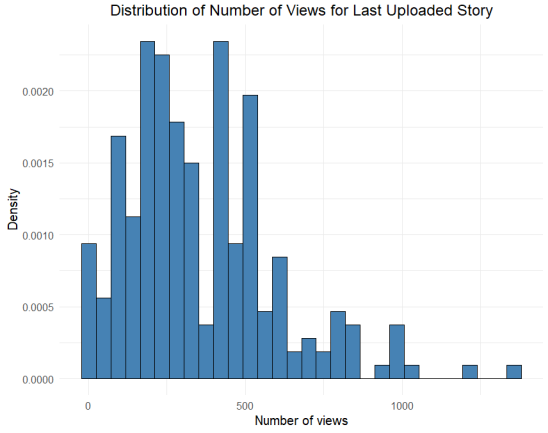
- **analysis of central tendency:** The variable "Number of Accounts" is a discrete numeric variable. It is unimodal with mode 1 (100 observations, 41%). The typical user manages just one account. However, since the Median is 2, 50% of the students under examination manage at least two accounts. The Mean of 1.94 is influenced by the strong concentration of 1 and 2 accounts. It is pulled up by the small number of outliers which manage up to 20 accounts.
- **analysis of non central tendency:** The distribution is positively skewed, as there is a concentration at 1 and 2 accounts managed, with a small number of respondents, the outliers, which pulls the tail out to the right.
- **analysis of dispersion:** The Range of 20 hints at a very high level of variability, however this is influenced by the presence of the one outlier, which manages 20 accounts, therefore the range is not a maximally reliable measure of dispersion, in this setting. The IQR is of 2, in this case. This means that the central 50% of the respondents manages between 1 and 2 accounts, which means that dispersion around the central mass of the distribution is very limited. The Standard Deviation of 1.53 hints at a moderate level of variability, given the

value of the Mean (1.94). It bears to say that, analogously to the previously examined case, the SD is influenced easily by extreme values such as the outlier of 20, in this case. It is therefore not a fully reliable measure of dispersion as it might give, similar to the range, an image of a more dispersed distribution than what would actually be the case. This findings are in line with the boxplot which showcases a narrow interquartile box, consistent with the IQR, but a long upper whisker, consistent with the presence of the outlier.

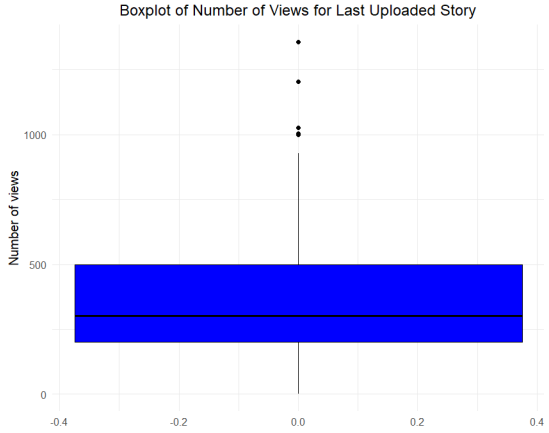
Variable analysis: "Number of views for last uploaded story"

Descriptive Statistics: Number of Views for Last Uploaded Story

Variable	Min	Max	Q1	Median	Mean	Q3	Variance	SD
Story Views	1.0	1357.0	200.0	302.5	355.6	500.0	57418.6	239.62



Enter Caption



Enter Caption

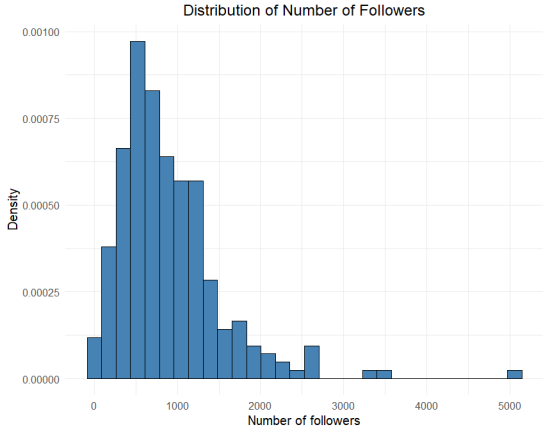
- **analysis of central tendency:** The variable "Number of visualizations" is a numeric discrete variable. The Median is of 302.5 views. In the presence of extreme outliers, such as in this case, it is a good measure as it well depicts central tendency, being very robust to outliers. In fact, the Mean (355.6) is noticeably higher than the Median, which is due to the presence of stories with many visualizations, with the maximum value obtained being 1357. The Mean overstates the typical number of views.
- **analysis of non central tendency:** The distribution is strongly right skewed. As we can see from the histogram, there is a high concentration at lower values, followed by an extended right tail. The boxplot confirms this observation. We can observe a narrow interquartile box and a long upper whisker, with a few high values beyond it, constituting the outliers.
- **analysis of dispersion:** The range of 1356 is not fully indicative, in this context. As already clarified, it is highly sensitive to extreme values, which, given the many outliers, makes it not

representative of the true dispersion of the data. 50% of students receive between 200 and 500 views, with an IQR of 300. There is a moderate level of spread for the central 50% of the observations. The Standard Devation indicates that, on average, the number of visualizations diverges from the Mean by about 240 units, hinting at a high level of variability. As was the case for the Range, SD is inflated by outliers; however, in this case, we can observe how the observations indeed fluctuate a lot, instead of clustering around a single value.

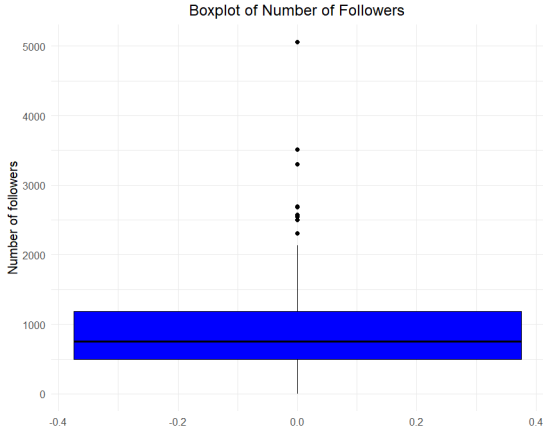
Variable analysis: "Number of followers"

Descriptive Statistics: Number of Followers

Variable	Min	Max	Q1	Median	Mean	Q3	Variance	SD
Number of followers	0	5057	500	757	901	1186	408445.4	639.10



Histogram



Boxplot

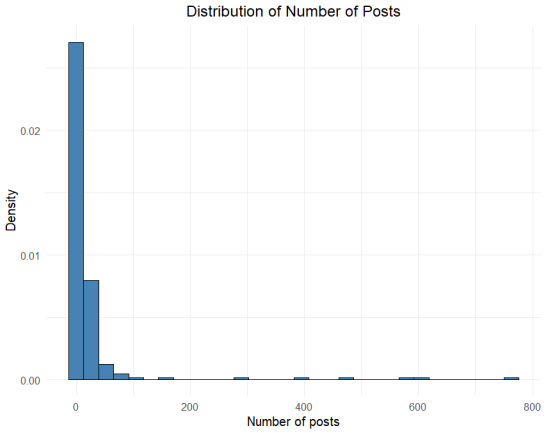
- **analysis of central tendency:** The Median is of 757 followers. In the presence of outliers it is a reliable measure of central tendency, as could be seen in the previous cases. The Mean of 901, noticeably higher, is inflated by outliers. In fact, very few profiles have a high number of followers, such as the one that has 5057, making it the most successful profile in the dataset, in terms of follower count.
- **analysis of non central tendency:** Also in this case the distribution is strongly right skewed. From the histogram, we can observe a high concentration at a relatively lower number of followers, with a few extreme values, which extend the right tail. The boxplot also sustains this point of view with a compact interquartile box and an extended upper whisked, with many outliers.
- **analysis of dispersion:** The Range and the Standard Deviation lie, respectively, at 5057 and 639.1. This shows a very high level of variability, despite the fact that they have been

inflated by a few outliers. The findings from the IQR (668) are in line with the visualizations provided by the boxplot. The center of the distribution exhibits a low to moderate level of spread, with the overall high dispersion mainly driven by a few outliers, instead of widespread extreme variability.

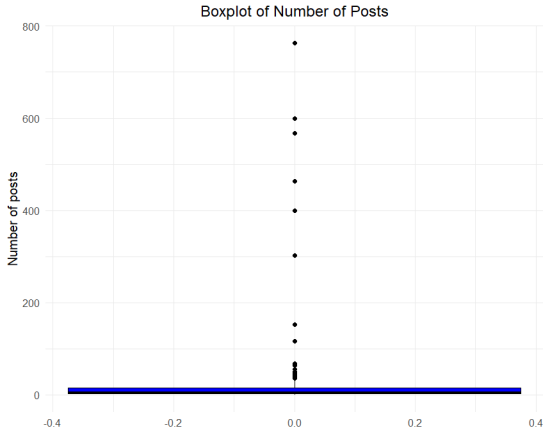
Variable analysis: "Number of Posts"

Table 1: Descriptive Statistics: Number of Posts

Variable	Min	Max	Q1	Median	Mean	Q3	Variance	SD
Number of posts	0	763	2	6	24.3	15.5	7001.30	83.67



Histogram



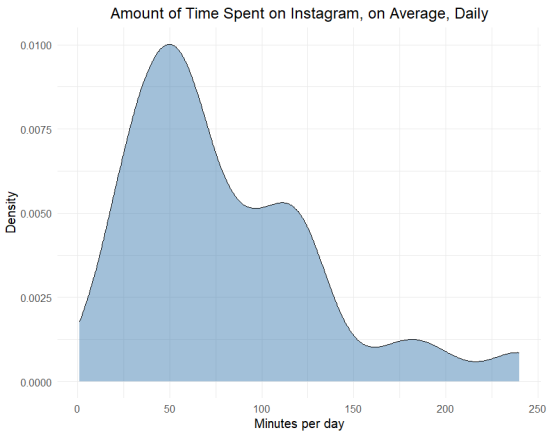
Boxplot

- **analysis of central tendency:** This is a numeric discrete variable. The Median of 6, is the most appropriate measure of central tendency, given the strong asymmetry of the distribution. This strong asymmetry is reflected in the very high value of the Mean (24.3), relative to the Median, which is inflated by the outliers. We can conclude that the Mean does not appropriately reflect the number of posts of the typical individual.
- **analysis of non central tendency:** The distribution of the number of posts is heavily right skewed. There is a high concentration at a low number of posts. The frequencies decrease steadily as the number of posts increases. The long right tail indicates the presence of a few extremely active individuals.
- **analysis of dispersion:** A reliable summary of variability is provided by the Interquartile Range of 15.5. This, as it is notoriously resistant to outliers. It shows how 50% of users are rarely present on social media, in terms of posting behavior, and a small part of the users posting very frequently. The Standard Deviation is extremely high, relative to the measures of central tendency, which is in line with what can be seen from the histogram. The data are very spread out, instead of clustering around a single value.

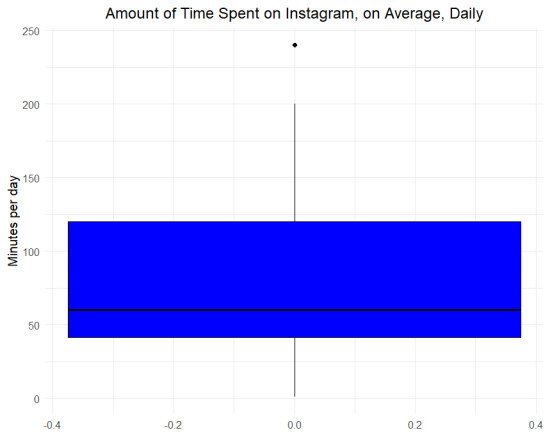
Variable analysis: "Time spent online"

Table 2: Descriptive Statistics: Daily Time

Variable	Min	Max	Q1	Median	Mean	Q3	Variance	SD
Daily Time	1.0	240.0	41.3	60.0	80.2	120.0	2769.48	52.63



Density plot



Boxplot

- **analysis of central tendency:** The variable "Time spent daily" is a numeric continous variable and the only of this type in the dataset. The Median of 60 and the Mean of 80 reflect the fact that 50% of the users spend 1 hour or less on Instagram, with some high users spending a lot, even up to 240 minutes on the app pulling the average upward.
- **analysis of non central tendency:** The findings of the previous point are consistent with the right skewed distribution. The distribution is asymmetric and righ tailed.
- **analysis of dispersion:** The Interquartile Range of about 79 minutes shows how the central 50% of the individuals spends between 41 and 120 minutes daily on the app, which signals a moderate level of disperion around the center. The Standard Deviation shows how the data are very spread out around the the mean. Precisely, they diverge, on average, from the Mean, by about 53 units. This hints at the fact that usage time is not similar accross individuals, with a few extreme users present.

A.2 Variable analysis

We first have to tae the fact into consideration that the concept of outlier only applies to numeric variables, as they are definded as values that lie very far away from the rest of the data. However, in order to label something as "far", you need: a numerical scale and a measurable distance. There is also the need for a measure to quantify how "far away" a value is. This, for example, can be done through the IQR.
Categorical variables do not have a numerically meaningful distance between each other. Consider, for example, "male" and "female". No one can be considered "further away" from something than the other.
Therefore, for this part of the task, we only take into consideration the numeric variables, as they have a measurable scale and whose distance can therefore be compared in a meaningful way.
The standard statistical way of identifying extreme values in the distribution is using the 1.5 x IQR rule. This is done by computing the values of Q1, Q3 and the IQR. Subsequently, the lower and upper limits were calculated as:

$$\text{Lower limit} = \text{Q1 - 1.5 x IQR}$$
$$\text{Upper limit} = \text{Q3 + 1.5 x IQR}$$

Then R calculates, based on the rule, the acceptable range and excludes anything that lies outside of it, labeling it as an outlier. Afterwards, R procedes cleaning out those values. We hereby removed possible distorting values from the dataset.

We took the decision to not make the mistake of removing entire rows, as for the A2 analysis we want to maximize observations.
We now display the main part of the code which helped us remove outliers.

```
no_outliers_siblings <- siblings[
  siblings > siblings_lower_limit & siblings < siblings_upper_limit]
no_outliers_account_num <- account_num[
  account_num > account_num_lower_limit & account_num < account_num_upper_
      limit]
no_outliers_story_views <- story_views[
  story_views > story_views_lower_limit & story_views < story_views_upper_
      limit]
no_outliers_num_follower <- num_follower[
  num_follower > num_follower_lower_limit & num_follower < num_follower_
      upper_limit]
no_outliers_day_time_min <- day_time_min[
  day_time_min > day_time_min_lower_limit & day_time_min < day_time_min_
      upper_limit]
no_outliers_num_post <- num_post[
  num_post > num_post_lower_limit & num_post < num_post_upper_limit]
```

Outliers are removed by applying the 1.5·IQR rule and retaining only observations that fall within the lower and upper bounds defined for each variable. This step operationally determines which values are excluded from the analysis.
After outlier removal the data count exhibits following features:

Variable	Valid Observations
Sex	243
Language	243
Private Account	243
Effectiveness	243
Attractiveness	243
Siblings	233
Number of Accounts	224
Story Views	221
Number of Followers	232
Number of Posts	222
Daily Time on Instagram (min)	230

Number of valid observations after outlier removal ($N_{total,previous} = 243$)

We can observe how outliers were removed only from numeric variables and not from categorical variables. This is the case as the concept of an outlier relies on numerical distances between observations and therefore only applies to numeric variables. Categorical variables do not have such a metric structure, making the identification of extreme values meaningless, statistically speaking. Prior to outlier removal, extreme observations had a disproportionate impact on numerical summary measures. In particular, the sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

is sensitive to large values, as outliers increase the sum of the observations. Similarly, the sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

and the corresponding standard deviation

$$s = \sqrt{s^2}$$

are amplified by extreme deviations from the mean. As a result, outliers inflate measures of dispersion and distort the representation of typical behavior, justifying their removal to obtain more valid descriptive statistics.

B. Confidence Intervals

Theorethical framework and methodolody

In this section of the report, it becomes further clear why the removal of the outliers has been necessary. In order to compute a valid confidence interval for the Mean, we had to remove extreme outliers, as they strongly distorted the distributions, by exaggerating the skewness and inflating the variance. The validity of the CI would have been compromised as, for instance its computation relies on measures such as the Standard Deviation. The SD is, in turn, sensitive to extreme observations, which explains the necessity for outliers cleansing.
A confidence interval is a range of values, calculated from the sample data, that is likely to contain the true population parameter with a specified level of confidence.
The sample size, after the removal of the outliers, is still large enough to apply the Central Limit Theorem (CLT). This theorem states, that in the presence of a large sample size, the Sample Mean is approximately normally distributed, an assumption necessary for the computation of a CI, as whitout it we would not be able to use the Standard Normal - Student T Quantiles.
The observation necessary to make in this case is that, since the population variance is unknown, we have to use the quantiles of the Student - T distribution. This is due to the fact that, an unknown variance generates additional uncertainty, other than the one already present,as it has to be approximated with the sample variance. The Student - T distribution is ideal to measure that uncertainty. It has heavier tails and thus compensates for the additional uncertainty generated.
The mathematical formula we use here to compute the confidence interval is:

$$\bar{x} \pm t_{\alpha/2,\, n-1} \frac{s}{\sqrt{n}}$$

$\bar{x}$ = sample mean
s = sample standard deviation
n = sample size
$t_{\alpha/2,\, n-1}$ = critical value from the Student's t-distribution with $n-1$ degrees of freedom

- Standard Error:

  $$\frac{s}{\sqrt{n}}$$

  The standard error measures how much the sample mean is expected to vary from sample to sample, reflecting the precision of the mean as an estimator of the population mean.

- Margin of Error:

  $$t_{\alpha/2,\, n-1} \frac{s}{\sqrt{n}}$$

  The margin of error represents the maximum expected difference between the sample mean and the true population mean at the chosen confidence level.

This paragraph is dedicated to clarify the meaning of the Confidence level $(1-\alpha)$. It does not mean that there is a $(1-\alpha)\%$ probability that the actual Mean lies within the interval computed. Instead, it means that, if we were to over and over again, draw samples from the exact same population and of the exact same size and subsequently compute a CI with them, 95% of them would contain the actual Population Mean. We can conclude that the confidence level reflects sort of the reliability

of the procedure in the Long Run, instead of some statement about the probability about each computed Confidence Interval.
Further consider that the strict significance level, hence the one, used to perform hypothesis tests, which will be of use in this task, is: $\alpha = 0.01$

B.1 Confidence Interval for the Number of Story Views

Sample Mean	$\bar{x} = 332.53$
Standard Error	$\frac{s}{\sqrt{n}} = 13.22$
Student–T quantile	$t_{\alpha/2,\, n-1} = 1.97$
Margin of Error	$t_{\alpha/2,\, n-1} \cdot \frac{s}{\sqrt{n}} = 26.05$
Confidence Interval	$[332.53 \pm 26.93] = [305.60,\ 359.46]$

As we have established the CI for the number of Story Views, we check if this average computed differs between only children and children with siblings.

Sample Mean Difference	$\bar{x}_{OC} - \bar{x}_{O} = -60.15$
Standard Error	$\frac{s}{\sqrt{n}} = 33.68$
Student–T quantile	$t_{\alpha/2,\, n-1} = 1.97$
Margin of Error	$t_{\alpha/2,\, n-1} \cdot SE = 66.02$
Confidence Interval	$[-60.15 \pm 66.02] = [-126.17,\ 5.87]$

Overall Sample: This interval provides, at the confidence level of 95 % a range for plausible values of the parameter: population mean. The Confidence Interval for the Population Mean is not very wide, considering the Sample Mean (332.53), which hints at a low level of sample variability. We can conclude that the CI provides a valid estimate interval for the population Mean. Moreover, the relatively narrow interval, considering the Mean indicates a low level of sample variability and that the estimate is precise and supported by a sufficiently large and stable sample. This might be due to a sufficiently large sample size and an overall moderate level of variability. At the 95% confidence level, the interval provides a range of plausible values for the true population mean.

Difference (Only Child vs. Siblings): The results of the analysis conducted, suggest that only children tend to have much fewer visualizations, on average, when put into comparison with students with siblings. However this cannot be said to be statistically significant, since the CI crosses zero and also contains a few positive values. We cannot say, with absolute certainty that there is a difference. Despite the sample mean difference is negative, this effect is not distinguishable from zero, statistically speaking.

B.2 Confidence Interval for the number of followers

Sample Mean	$\bar{x} = 812.66$
Standard Error	$\frac{s}{\sqrt{n}} = 39.98$
Student–T quantile	$t_{\alpha/2,\, n-1} = 2.59$
Margin of Error	$t_{\alpha/2,\, n-1} \cdot \frac{s}{\sqrt{n}} = 77.86$
Confidence Interval	$[812.66 \pm 77.86] = [734.6,\ 890.23]$

As we have established the 99% Confidence Interval for the number of followers, we now assess whether this average differs between men and women.

Sample Mean Difference	$\bar{x}_M - \bar{x}_W = -181.44$
Standard Error	$\widehat{SE} = 59.08$
Student–T quantile	$t_{\alpha/2,\, n-1} = 2.59$
Margin of Error	$t_{\alpha/2,\, n-1} \cdot SE = 154.33$
Confidence Interval	$[-181.44 \pm 154.33] = [-335.78,\ -27.11]$

Overall Sample: This is an interval which provides a possible range for the true parameter: population mean of the number of followers at the 99% confidence level. Considering the magnitude of the sample mean, it is moderately narrow. This indicates a, in turn, moderate level of precision of the estimate. It further suggests a moderate level of sampling variability consistent with a moderately large and stable sample size.

Difference (Female vs. Male): Based on the analysis conducted, we can say that there is a statistically significant difference in the number of followers between male and female users. As the Estimated Difference is of 181.45, we conclude that female students, on average, have considerably more followers than male students.
We can immediately observe how the interval does not cross zero and as a consequence does also not contain positive values. This indicates that, at a significance level of 1%, there is a statistically significant difference in the average number of followers between women and men. As the entire interval does only contain negative values, we can assess that, on average, women have more followers then men.

B.3 Confidence Interval for the Proportion of accounts in Italian

Theorethical framework and methodology

In this section we are tasked with computing a confidence interval for the proportion of accounts set in Italian. This means that we have to compute an estimate for a different parameter now. This recquires a different approach compared to before.

The sample proportion is denoted by

$$\hat{p} = \frac{x}{n},$$

In this case, as we know that both we can apply the Central Limit Theorem, as we are working with a large sample size. It states that in the presence of said large sample size, we can assume that the sample proportion $\hat{p}$ is approximately normally distributed. Since this has been verified we can work with the quantiles of the Standard Normal distribution instead of having to work with the Student - T distribution.
The standard error of the sample proportion is given by

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

A $100(1-\alpha)\%$ confidence interval for the population proportion is therefore computed as

$$\hat{p} \pm z_{\alpha/2} \cdot SE_{\hat{p}},$$

where $z_{\alpha/2}$ is the critical value from the Standard Normal distribution corresponding to the desired confidence level.
This interval provides a range of plausible values for the true population proportion p.

Computation

Sample Proportion	$\hat{p} = 0.7160$
Standard Error	$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.0519$
Z–critical value (90% CI)	$z_{0.05} = 1.645$
Margin of Error	$z_{0.05} \cdot SE_{\hat{p}} = 0.0519 \times 1.645 = 0.0854$
Confidence Interval	$[\hat{p} \pm ME] = [0.6642,\ 0.7630]$

At the confidence level of 90%, this interval provides a range of [0.66, 0.76] for the sample proportion of accounts in Italian. Considering the sample propotion estimate (0.71), we have that the interval is relatively narrow, which suggests a fairly precise estimate of the population proportion. The narrowness of the interval further suggests that there is a low level of sampling variability, due to an appropriately large level of sampling size and low level of variability in the underlying data.

C. Hypothesis Testing

Theorethical framework and methodology

The hypotheses are formalized as:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

Two Means are compared, in order to test a formulated "null" hypothesis, which states that there is no true difference between them, hence that any difference which might be observed should be due to random sampling. The alternative hypothesis states that there actually is a difference in the population means and that there is a genuine underlying effect, proved by the sample evidence. The decision rule is based on the computed p-value: if the probability of observing a certain result under the null hypothesis is smaller than a certain significance level, then we reject the null-hypothesis.

C.1 Difference in number of followers between women and men

Comparison	**Test Statistic**	**df**	**p-value**	**Conclusion** ($\alpha = 0.01$)	**Conclusion** ($\alpha = 0.1$)
Difference (Women,Men)	$t = 3.0547$	220.2	0.00253	Reject H_0	Reject H_0

As the computed p - value is below the strict significance level of $\alpha = 0.01$, we reject the null hypothesis, that there is no difference between women and men, when it comes to follower count. In fact, it appears that women, on average have more followers then men.
The computed t - statistic shows how the difference in average follower count between women and men is 3 Standard Error units away from zero. This effect is larger then what random sampling could have produced. Therefore, the effect is not only noticeable, but also statistically meaningful. We already came to this conclusion in point B.2, when we were able to show that the confidence interval for the average difference in followers between men and women, lay entirely above zero. This, together with the findings from the just computed hypothesis test, led us to the conclusion that there is actually a statistically significant difference in the average number of followers.
By amplifying the level of significance we did not change the fact that the null - hypothesis will end up being rejected. This is due to the fact that the small size of the p - value allows for the interpretation that obtaining such a result under assumption that the null - hypothesis holds is close to impossible.
This result is in line with the findings of point B.2. In fact, the 99% confidence interval for the difference in average followers lies entirely above zero. Thus we can say that zero is definitely not a plausible value for the true population difference.

C.2 Difference in views between students of private vs public universities

Comparison	**Test Statistic**	**df**	**p-value**	**Conclusion (at** $\alpha = 0.01$)
Difference (Public,Private)	$t = -2.3161$	176.44	0.0217	Not Significant

The Welch two-sample test yielded a p-value of 0.0217. This value is greater than the imposed strict significance level of $\alpha = 0.01$. Therefore we do not reject the null hypothesis that there is

a difference in views enjoyed by students, based on the fact that they attend a private instead of a public university, or viceversa. Based on the available evidence, we can say that the number of views is not influenced by the type of university attended by the individual.

C.3 Difference in time spent on Instagram daily, Italian speakers vs English speakers

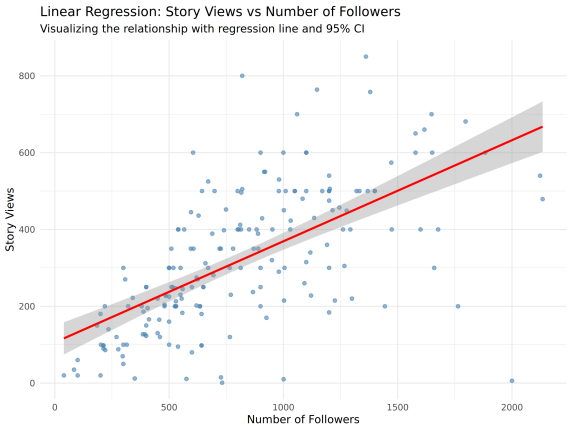
Comparison	**Test Statistic**	**df**	**p-value**	**Conclusion (at $\alpha = 0.01$)**
Difference (English,Italian)	$t = 1.121$	67.616	0.2662	Not Significant

The difference in time spent on the network between English and Italian speakers is not statistically significant, as the p-value computed, of 0.266 is above the significance level imposed of $\alpha = 0.01$. We can conclude that the evidence collected does not prove that there is a real underlying population tendency and that the effect observed (difference in sample means) is compatible with randomness.

D. Linear Regression

D.1 Simple linear regression: story_views on num_follower

Scatterplot and fitted regression line



Scatterplot of `story_views` against `num_follower` with fitted regression line

We check the relationship between the number of story views and the number of followers using the dataset `data_no_outliers` with $n = 191$ observations. The simple linear regression model is

$$\text{story\_views}_i = \beta_0 + \beta_1 \, \text{num\_follower}_i + \varepsilon_i. = 106.02 + 0.263 \, \text{num\_follower}_i + \varepsilon_i.$$

Regression table

term	estimate	std.error	statistic	p.value
(Intercept)	106.0242	22.0803	4.8018	0
num_follower	0.2632	0.0240	10.9707	0

Regression table

term	df	sumsq	meansq	statistic	p.value
num_follower	1	2433895	2433894.73	120.357	0
Residuals	189	3822013	20222.29	NA	NA

ANOVA table

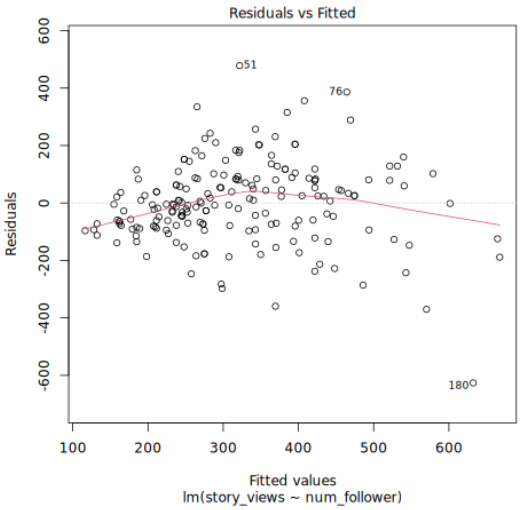
The Regression table shows that the slope for `num_follower` is positive and highly significant (p-value ≈ 0), while the intercept is also significantly different from zero. This confirms the strong upward association already visible in scatterplot and fitted regression line: accounts with more followers tend to obtain more story views.

Interpreting the magnitude, the slope estimate implies that an extra 100 followers are associated with roughly 26 additional views, and 1,000 extra followers with about 260 additional views, on average. The intercept corresponds to the predicted number of views at zero followers, which is outside the observed data range and is mainly a baseline of the linear specification.

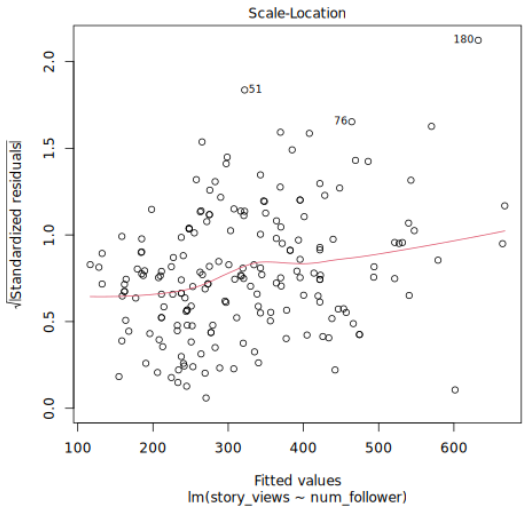
The ANOVA table shows that the variation explained by `num_follower` is very large relative to the residual variation (F-statistic around 120 with p-value ≈ 0). Together with an R^2 of about 0.39, this

indicates that the model captures a substantial, though not exhaustive, portion of the variability in story views. Followers are clearly informative, but other unobserved factors also play an important role.

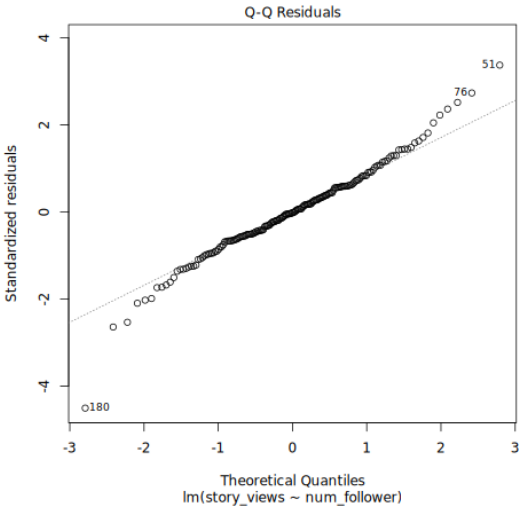
Diagnostic plots



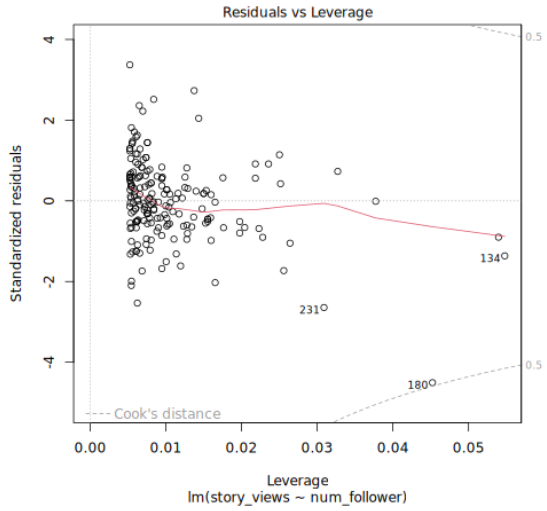
Residuals vs fitted values



Scale–location plot



Normal Q–Q plot



Residuals vs leverage

Assumption checks: linearity, normality and homoscedasticity

Linearity. The scatterplot and the residuals-vs-fitted plot in both suggest an approximately linear relationship. The loess smooth is fairly flat, with only a mild downward bend for the largest fitted values. There is no strong U-shape or S-shape pattern, so a simple linear term in `num_follower` provides a reasonable first–order approximation.

Normality. In the Q-Q plot of standardized residuals, the bulk of the points lie close to the diagonal reference line. Deviations occur mainly in the extreme lower and upper tails, where a few observations depart from the line, indicating slightly heavier tails than a perfect normal distribution. However, there is no pronounced skewness, and with $n = 191$ such moderate departures are unlikely to seriously affect inference based on t- and F-statistics.

Homoscedasticity. The scale–location plot shows the square root of the standardized residuals against the fitted values. The loess curve is gently increasing, hinting at a modest rise in variability for larger predicted numbers of views, but there is no clear "fan shape" or dramatic explosion in spread. The residuals-vs-fitted plot presents a fairly even vertical spread across the range of fitted values. Overall, these plots suggest that assumptions are reasonably satisfied.

Influential observations. The residuals-vs-leverage plot shows a few points with relatively high leverage, but their standardized residuals are not extreme and none of them lie close to the Cook's distance contours. Hence, there is no evidence that the fitted line is being unduly driven by a single or very small number of observations.

Overall assessment. Taken together, the diagnostic plots indicate that the linearity, normality and homoscedasticity assumptions are reasonably satisfied in an approximate sense. The model provides a sensible and interpretable summary of how story views vary with the number of followers, and there is no indication of severe violations that would invalidate the main conclusions.

D.2 Multiple linear regression

We now extend the model by including additional covariates that describe the account and the user's activity on the platform. The multiple linear regression is

$$\text{story\_views}_i = \beta_0 + \beta_1\,\text{num\_follower}_i + \beta_2\,\text{sexM}_i + \beta_3\,\text{account\_num}_i + \beta_4\,\text{num\_post}_i + \beta_5\,\text{day\_time\_min}_i + \varepsilon_i,$$

where `sexM` equals 1 for male accounts and 0 for female accounts (the reference group), `account_num` is the number of accounts followed, `num_post` the total number of posts, and `day_time_min` the average daily time spent on the app in minutes.

term	estimate	std.error	statistic	p.value
(Intercept)	68.0881	33.3029	2.0445	0.0423
num_follower	0.2649	0.0254	10.4114	0.0000
sexM	34.0596	22.6545	1.5034	0.1344
account_num	11.6974	6.4627	1.8100	0.0719
num_post	0.3958	1.3241	0.2989	0.7653
day_time_min	-0.0909	0.2383	-0.3812	0.7035

Estimated coefficients for the multiple regression model

This table summarizes the estimated coefficients. The coefficient for `num_follower` remains positive, large in magnitude, and highly significant even after controlling for sex, number of accounts followed, number of posts and time spent on the app. Its estimate is almost unchanged compared to the simple regression, indicating that followers are the main driver of story views among the variables considered.

The gender indicator `sexM` is positive but not statistically significant at the 5% level, so we cannot conclude that male accounts systematically receive more views than comparable female accounts once the other covariates are held fixed. The coefficient on `account_num` is marginally significant (around the 10% level), suggesting a weak tendency for accounts that follow more users to get more views, whereas `num_post` and `day_time_min` have small, clearly non–significant effects.

In terms of goodness of fit, the multiple model yields an R^2 of about 0.41 and an adjusted R^2 of roughly 0.39, only slightly higher than in the simple regression. The residual standard error decreases very little. This means that the extra variables offer only a modest improvement in explanatory power beyond the number of followers. The overall F-test for the joint significance of all regressors is highly significant, confirming that, taken together, the covariates are associated with story views; however, most of this explanatory power comes from `num_follower`.

Overall, we can say that the number of followers is the most efficient predictor, underscored by a p - value of approximately 0. It is followed by the number of accounts, which, as already clarified, has marginal level of statistical significance.

E. Prediction

E.1 Predicted expected views for the "female median account"

We use the multiple regression model from Section D2. to obtain a point prediction for a "female median account". This profile is defined as a female account (sexM = 0) whose values of `num_follower`, `account_num`, `num_post` and `day_time_min` are all set equal to their respective sample medians.

We therefore set

```
median_data_female <- data.frame(sex = "F",num_follower = median(data_no_
    outliers$num_follower, na.rm = TRUE),account_num = median(data_no_
    outliers$account_num, na.rm = TRUE),num_post = median(data_no_outliers$
    num_post, na.rm = TRUE),day_time_min = median(data_no_outliers\$day_time
    _min, na.rm = TRUE))
```

Substituting this median values into the estimated regression equation yields a fitted value of

$$\hat{y}_{\text{female median}} = 284.46 \text{ views.}$$

This should be interpreted as the expected number of story views for accounts with those median characteristics. Individual accounts with the same covariate values may deviate substantially from this value because of idiosyncratic factors captured by the error term.

E.2 95% confidence interval for the expected value

To quantify uncertainty around this point prediction, we construct a 95% confidence interval for the mean number of story views of the "female median account". Using the standard error of the fitted value from the multiple regression, the interval is

$$\hat{y} = 284.46, \qquad \text{CI}_{95\%} = [251.28,\ 317.64].$$

We are therefore 95% confident that the true population mean number of views for all accounts with these median characteristics lies between about 251 and 318 views. Note that this is a confidence interval for the *mean* response; a prediction interval for a single new account would be much wider, as it would also incorporate the residual variability (roughly 140 views).

F. Logistic Regression

F.1 Predicting private university status

We finally investigate whether social media activity can be used to predict whether an account belongs to a student attending a private university. We fit a logistic regression model with dependent variable `private_d`, equal to 1 if the student attends a private university and 0 otherwise:

$$\Pr(\text{private\_d}_i = 1 \mid X_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}, \qquad \eta_i = \beta_0 + \beta_1\,\text{num\_follower}_i + \beta_2\,\text{num\_post}_i + \beta_3\,\text{story\_views}_i.$$

term	estimate	std.error	statistic	p.value
(Intercept)	-1.5239	0.3739	-4.076	0.0000
num_follower	0.0003	0.0005	0.653	0.5138
num_post	0.0108	0.0178	0.608	0.5432
story_views	0.0024	0.0011	2.179	0.0293

Estimated coefficients for the logistic regression model

Analysis of results:
The intercept is significantly negative, implying that, for accounts with very low values of followers, posts and views, attending a private university is relatively unlikely. Among the covariates, `story_views` has a positive and statistically significant coefficient. This is supported by the computed p-value of 0.0293, which falls below the significance level. This can be interpreted by saying that observing such coefficients, under the null hypothesis that the variable has no effect, would be very unlikely. `num_follower` and `num_post`, in turn, are not statistically significant once views are controlled for. Their p-values (0.51 and 0.54, respectively) are well above significance levels, which shows that, under the null hypothesis that these variables have no effect, the observed coefficients are not sufficiently unlikely. We, therefore, do not have enough statistical evidence that they can be used as explanatory variables for private or public university attendance.

In odds-ratio terms, an increase of 100 story views corresponds to roughly a 25-30% increase in the odds of attending a private university, holding the other variables fixed, whereas changes in followers or posts have negligible additional effects.

The McFadden pseudo R^2 of about 0.05 shows that the model performs only modestly better than an intercept only specification. Thus, while students at private universities tend to have somewhat higher story views, the three social media variables considered here are not sufficient to classify accounts with high accuracy. The logistic regression provides some predictive signal, but its overall explanatory power remains fairly limited.
This can be observed also in the regression results.
As all variables enter the model, only story visualization displays a statistically significant marginal effect, once the other covariates are controlled for. Follower count and post count, instead, does not provide any additional explanatory power.

This is the R full code of the project:

```
# Download Libraries

library(ggplot2)
library(dplyr)
library(corrplot)
library(car)
library(lmtest)
library(nortest)
library(gridExtra)
library(moments)
library(ggfortify)
library(broom)
library(ggpubr)

# Clean Data by Removing NAs for Each Variable Separately

sex <- na.omit(data$sex)
language <- na.omit(data$language)
siblings <- na.omit(data$siblings)
account_num <- na.omit(data$account_num)
story_views <- na.omit(data$story_views)
day_time_min <- na.omit(data$day_time_min)
num_follower <- na.omit(data$num_follower)
num_post <- na.omit(data$num_post)
effectiveness <- na.omit(data$effectiveness)
attractiveness <- na.omit(data$attractiveness)
private_d <- na.omit(data$private_d)


# A. Descriptive Statistics

# A1. Descriptive Statistics (With Outliers)

# Summary of the Data
print(summary(data, na.rm=TRUE))

# Sex Analysis

sex_counts <- table(sex)
sex_props <- prop.table(sex_counts)
sex_mode <- names(sex_counts)[which.max(sex_counts)]


# Language Analysis

language_counts <- table(language)
language_props <- prop.table(language_counts)
language_mode <- names(language_counts)[which.max(language_counts)]
```

```r
# Siblings Analysis

siblings_counts <- table(siblings)
siblings_props <- prop.table(siblings_counts)
siblings_mode <- names(siblings_counts)[which.max(siblings_counts)]


#account_num Analysis

account_num_counts <- table(account_num)
account_num_props <- prop.table(account_num_counts)
account_num_mode <- names(account_num_counts)[which.max(account_num_counts)
    ]

account_num_mean <- mean(account_num)
account_num_median <- median(account_num)
account_num_sd <- sd(account_num)
account_num_var <- var(account_num)
account_num_IQR <- IQR(account_num)
account_num_quantiles <- quantile(account_num, probs = c(0.25, 0.5, 0.75))
account_num_Q1 <- quantile(account_num, 0.25)
account_num_Q3 <- quantile(account_num, 0.75)
account_num_IQR_value <- IQR(account_num)
account_num_lower_limit <- account_num_Q1 - (1.5 * account_num_IQR_value)
account_num_upper_limit <- account_num_Q3 + (1.5 * account_num_IQR_value)


# Story Views Analysis

story_views_mean <- mean(story_views)
story_views_median <- median(story_views)
story_views_sd <- sd(story_views)
story_views_var <- var(story_views)
story_views_IQR <- IQR(story_views)
story_views_quantiles <- quantile(story_views, probs = c(0.25, 0.5, 0.75))
story_views_Q1 <- quantile(story_views, 0.25)
story_views_Q3 <- quantile(story_views, 0.75)
story_views_lower_IQR_value <- IQR(story_views)
story_views_lower_limit <- story_views_Q1 - (1.5 * story_views_lower_IQR_
    value)
story_views_upper_limit <- story_views_Q3 + (1.5 * story_views_lower_IQR_
    value)

# Day Time Minutes Analysis

day_time_min_mean <- mean(day_time_min)
day_time_min_median <- median(day_time_min)
day_time_min_sd <- sd(day_time_min)
day_time_min_var <- var(day_time_min)
```

```r
day_time_min_IQR <- IQR(day_time_min)
day_time_min_quantiles <- quantile(day_time_min, probs = c(0.25, 0.5, 0.75)
    )
day_time_min_Q1 <- quantile(day_time_min, 0.25)
day_time_min_Q3 <- quantile(day_time_min, 0.75)
day_time_min_IQR_value <- IQR(day_time_min)
day_time_min_lower_limit <- day_time_min_Q1 - (1.5 * day_time_min_IQR_value
    )
day_time_min_upper_limit <- day_time_min_Q3 + (1.5 * day_time_min_IQR_value
    )

# Number of Followers Analysis

num_follower_mean <- mean(num_follower)
num_follower_median <- median(num_follower)
num_follower_sd <- sd(num_follower)
num_follower_var <- var(num_follower)
num_follower_IQR <- IQR(num_follower)
num_follower_quantiles <- quantile(num_follower, probs = c(0.25, 0.5, 0.75)
    )
num_follower_Q1 <- quantile(num_follower, 0.25)
num_follower_Q3 <- quantile(num_follower, 0.75)
num_follower_IQR_value <- IQR(num_follower)
num_follower_lower_limit <- num_follower_Q1 - (1.5 * num_follower_IQR_value
    )
num_follower_upper_limit <- num_follower_Q3 + (1.5 * num_follower_IQR_value
    )

# Number of Posts Analysis

num_post_mean <- mean(num_post)
num_post_median <- median(num_post)
num_post_sd <- sd(num_post)
num_post_var <- var(num_post)
num_post_IQR <- IQR(num_post)
num_post_quantiles <- quantile(num_post, probs = c(0.25, 0.5, 0.75))
num_post_min_Q1 <- quantile(num_post, 0.25)
num_post_min_Q3 <- quantile(num_post, 0.75)
num_post_IQR_value <- IQR(num_post)
num_post_lower_limit <- num_post_min_Q1 - (1.5 * num_post_IQR_value)
num_post_upper_limit <- num_post_min_Q3 + (1.5 * num_post_IQR_value)

# Effectiveness Analysis

effectiveness_counts <- table(effectiveness)
effectiveness_props <- prop.table(effectiveness_counts)
effectiveness_mode <- names(effectiveness_counts)[which.max(effectiveness_
    counts)]
```

```r
# Attractiveness Analysis

attractiveness_counts <- table(attractiveness)
attractiveness_props <- prop.table(attractiveness_counts)
attractiveness_mode <- names(attractiveness_counts)[which.max(
    attractiveness_counts)]


# Private_d Analysis
private_d_counts <- table(private_d)
private_d_props <- prop.table(private_d_counts)
private_d_counts_mode <- names(private_d_counts)[which.max(private_d_counts
    )]


# Visuals

# Sex Analysis Visuals

plot_sex <- ggplot(as.data.frame(sex_counts), aes(x = sex, y = Freq)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = Freq), vjust = -0.3) +

  labs(title = "Count by Sex", y = "Count")

ggsave("results/plot_sex.png", plot = plot_sex, width = 6, height = 4)

plot_sex_props <- ggplot(as.data.frame(sex_props), aes(x = sex, y = Freq))
    +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = paste0(" (", round(Freq / sum(Freq) * 100, 2), "%)"
      )), vjust = -0.3) +

  labs(title = "Count by Sex", y = "Count")

ggsave("results/plot_sex_props.png", plot = plot_sex_props, width = 6,
    height = 4)


# Language Analysis Visuals

plot_language <- ggplot(as.data.frame(language_counts), aes(x = language, y
     = Freq)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = Freq), vjust = -0.3) +

  labs(title = "Count by Language", y = "Count")
```

```
ggsave("results/plot_language.png", plot = plot_language, width = 6, height
    = 4)

plot_language_props <- ggplot(as.data.frame(language_props), aes(x =
   language, y = Freq)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = paste0(" (", round(Freq / sum(Freq) * 100, 2), "%)"
     )), vjust = -0.3) +

  labs(title = "Count by Language", y = "Count")

ggsave("results/plot_language_props.png", plot = plot_language_props, width
    = 6, height = 4)

# Siblings Analysis Visuals

plot_siblings <- ggplot(as.data.frame(siblings_counts), aes(x = siblings, y
    = Freq)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = Freq), vjust = -0.3) +

  labs(title = "Count by Siblings", y = "Count")

ggsave("results/plot_siblings.png", plot = plot_siblings, width = 6, height
    = 4)

plot_siblings_props <- ggplot(as.data.frame(siblings_props), aes(x =
   siblings, y = Freq)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = paste0(" (", round(Freq / sum(Freq) * 100, 2), "%)"
     )), vjust = -0.3) +

  labs(title = "Count by Siblings", y = "Count")

ggsave("results/plot_siblings_props.png", plot = plot_siblings_props, width
    = 6, height = 4)

# Account Num Analysis Visuals

plot_account_num <- ggplot(as.data.frame(account_num_counts), aes(x =
   account_num, y = Freq)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = Freq), vjust = -0.3) +

  labs(title = "Count by Account Number", y = "Count")
```

```r
ggsave("results/plot_account_num.png", plot = plot_account_num, width = 6,
    height = 4)

plot_account_num_props <- ggplot(as.data.frame(account_num_props), aes(x =
    account_num, y = Freq)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = paste0(" (", round(Freq / sum(Freq) * 100, 2), "%)"
      )), vjust = -0.3) +

  labs(title = "Count by Account Number", y = "Count")

ggsave("results/plot_account_num_props.png", plot = plot_account_num_props,
     width = 6, height = 4)

# Effectiveness Analysis Visuals

plot_effectiveness <- ggplot(as.data.frame(effectiveness_counts), aes(x =
    effectiveness, y = Freq)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = Freq), vjust = -0.3) +

  labs(title = "Count by Effectiveness", y = "Count")

ggsave("results/plot_effectiveness.png", plot = plot_effectiveness, width =
     6, height = 4)

plot_effectiveness_props <- ggplot(as.data.frame(effectiveness_props), aes(
    x = effectiveness, y = Freq)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = paste0(" (", round(Freq / sum(Freq) * 100, 2), "%)"
      )), vjust = -0.3) +

  labs(title = "Count by Effectiveness", y = "Count")

ggsave("results/plot_effectiveness_props.png", plot = plot_effectiveness_
    props, width = 6, height = 4)

# Attractiveness Analysis Visuals

plot_attractiveness <- ggplot(as.data.frame(attractiveness_counts), aes(x =
     attractiveness, y = Freq)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = Freq), vjust = -0.3) +

  labs(title = "Count by Attractiveness", y = "Count")
```

```
ggsave("results/plot_attractiveness.png", plot = plot_attractiveness, width
    = 6, height = 4)

plot_attractiveness_props <- ggplot(as.data.frame(attractiveness_props),
    aes(x = attractiveness, y = Freq)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = paste0(" (", round(Freq / sum(Freq) * 100, 2), "%)"
      )), vjust = -0.3) +

  labs(title = "Count by Attractiveness", y = "Count")

ggsave("results/plot_attractiveness_props.png", plot = plot_attractiveness_
    props, width = 6, height = 4)

# Private_d Analysis Visuals

plot_private_d <- ggplot(as.data.frame(private_d_counts), aes(x = private_d
    , y = Freq)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = Freq), vjust = -0.3) +

  labs(title = "Count by Private D", y = "Count")

ggsave("results/plot_private_d.png", plot = plot_private_d, width = 6,
    height = 4)

plot_private_d_props <- ggplot(as.data.frame(private_d_props), aes(x =
    private_d, y = Freq)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = paste0(" (", round(Freq / sum(Freq) * 100, 2), "%)"
      )), vjust = -0.3) +

  labs(title = "Count by Private D", y = "Count")

ggsave("results/plot_private_d_props.png", plot = plot_private_d_props,
    width = 6, height = 4)

# Account Num Analysis Visuals

plot_account_num <- ggplot(as.data.frame(account_num_counts), aes(x =
    account_num, y = Freq)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = Freq), vjust = -0.3) +

  labs(title = "Count by Account Number", y = "Count")
```

```
ggsave("results/plot_account_num.png", plot = plot_account_num, width = 6,
    height = 4)

plot_account_num_props <- ggplot(as.data.frame(account_num_props), aes(x =
    account_num, y = Freq)) +
  geom_col(fill = "steelblue") +
  theme_minimal() +
  geom_text(aes(label = paste0(" (", round(Freq / sum(Freq) * 100, 2), "%)"
      )), vjust = -0.3) +

  labs(title = "Count by Account Number", y = "Count")

ggsave("results/plot_account_num_props.png", plot = plot_account_num_props,
     width = 6, height = 4)

# Story Views Boxplot and Histogram

boxplot_story_views <- ggplot(as.data.frame(num_follower), aes(y=num_
    follower)) +
  geom_boxplot(fill="orange") +
  theme_minimal() + ggtitle("Boxplot of Followers (Original)")

ggsave("results/boxplot_story_views.png", plot = boxplot_story_views, width
     = 8, height = 6)

histogram_story_views <- ggplot(as.data.frame(story_views), aes(x=story_
    views)) +
  geom_histogram(fill="skyblue", color="black", bins=30) +
  theme_minimal() + ggtitle("Distribution of Story Views (Original)")

ggsave("results/histogram_story_views.png", plot = histogram_story_views,
    width = 8, height = 6)

# Number of Followers Analysis Boxplot and Histogram

boxplot_followers <- ggplot(as.data.frame(num_follower), aes(y=num_follower
    )) +
  geom_boxplot(fill="orange") +
  theme_minimal() + ggtitle("Boxplot of Followers (Original)")

ggsave("results/boxplot_followers.png", plot = boxplot_followers, width =
    8, height = 6)

histogram_followers <- ggplot(as.data.frame(story_views), aes(x=story_views
    )) +
  geom_histogram(fill="skyblue", color="black", bins=30) +
  theme_minimal() + ggtitle("Distribution of Followers (Original)")

ggsave("results/histogram_followers.png", plot = histogram_followers, width
```

```
    = 8, height = 6)

# Day Time Minutes Analysis Boxplot and Histogram

boxplot_day_time_min <- ggplot(as.data.frame(day_time_min), aes(y=day_time_
    min)) +
  geom_boxplot(fill="orange") +
  theme_minimal() + ggtitle("Boxplot of Day Time Minutes (Original)")

ggsave("results/boxplot_day_time_min.png", plot = boxplot_day_time_min,
    width = 8, height = 6)

histogram_day_time_min <- ggplot(as.data.frame(day_time_min), aes(x=day_
    time_min)) +
  geom_histogram(fill="skyblue", color="black", bins=30) +
  theme_minimal() + ggtitle("Distribution of Day Time Minutes (Original)")

ggsave("results/histogram_day_time_min.png", plot = histogram_day_time_min,
     width = 8, height = 6)

#Number of Posts Analysis Boxplot and Histogram

boxplot_num_post <- ggplot(as.data.frame(num_post), aes(y=num_post)) +
  geom_boxplot(fill="orange") +
  theme_minimal() + ggtitle("Boxplot of Number of Posts (Original)")

ggsave("results/boxplot_num_post.png", plot = boxplot_num_post, width = 8,
    height = 6)

histogram_num_post <- ggplot(as.data.frame(num_post), aes(x=num_post)) +
  geom_histogram(fill="skyblue", color="black", bins=30) +
  theme_minimal() + ggtitle("Distribution of Number of Posts (Original)")

ggsave("results/histogram_num_post.png", plot = histogram_num_post, width =
     8, height = 6)

# A2. Descriptive Statistics (Without Outliers)

no_outliers_account_num <- account_num[account_num > account_num_lower_
    limit & account_num < account_num_upper_limit]
no_outliers_story_views <- story_views[story_views > story_views_lower_
    limit & story_views < story_views_upper_limit]
no_outliers_num_follower <- num_follower[num_follower > num_follower_lower_
    limit & num_follower < num_follower_upper_limit]
no_outliers_day_time_min <- day_time_min[day_time_min > day_time_min_lower_
    limit & day_time_min < day_time_min_upper_limit]
no_outliers_num_post <- num_post[num_post > num_post_lower_limit & num_post
     < num_post_upper_limit]

# Account Number Analysis (No Outliers)
```

```
boxplot_account_num <- ggplot(as.data.frame(no_outliers_account_num), aes(y
    =no_outliers_account_num)) +
  geom_boxplot(fill="orange") +
  theme_minimal() +
  ggtitle("Boxplot of Account Number (No Outliers)") +
  ylab("Account Number")

ggsave("results/boxplot_account_num_no_outliers.png", plot = boxplot_
    account_num, width = 8, height = 6)

histogram_account_num <- ggplot(as.data.frame(no_outliers_account_num), aes
    (x=no_outliers_account_num)) +
  geom_histogram(fill="skyblue", color="black", bins=30) +
  theme_minimal() +
  ggtitle("Distribution of Account Number (No Outliers)") +
  xlab("Account Number")

ggsave("results/histogram_account_num_no_outliers.png", plot = histogram_
    account_num, width = 8, height = 6)

# Story Views Analysis (No Outliers)

boxplot_story_views <- ggplot(as.data.frame(no_outliers_story_views), aes(y
    =no_outliers_story_views)) +
  geom_boxplot(fill="orange") +
  theme_minimal() +
  ggtitle("Boxplot of Story Views (No Outliers)") +
  ylab("Story Views")

ggsave("results/boxplot_story_views_no_outliers.png", plot = boxplot_story_
    views, width = 8, height = 6)

histogram_story_views <- ggplot(as.data.frame(no_outliers_story_views), aes
    (x=no_outliers_story_views)) +
  geom_histogram(fill="skyblue", color="black", bins=30) +
  theme_minimal() +
  ggtitle("Distribution of Story Views (No Outliers)") +
  xlab("Story Views")

ggsave("results/histogram_story_views_no_outliers.png", plot = histogram_
    story_views, width = 8, height = 6)

# Number of Followers Analysis (No Outliers)

boxplot_followers <- ggplot(as.data.frame(no_outliers_num_follower), aes(y=
    no_outliers_num_follower)) +
  geom_boxplot(fill="orange") +
  theme_minimal() +
  ggtitle("Boxplot of Followers (No Outliers)") +
```

```
  ylab("Number of Followers")

ggsave("results/boxplot_followers_no_outliers.png", plot = boxplot_
    followers, width = 8, height = 6)

histogram_followers <- ggplot(as.data.frame(no_outliers_num_follower), aes(
    x=no_outliers_num_follower)) +
  geom_histogram(fill="skyblue", color="black", bins=30) +
  theme_minimal() +
  ggtitle("Distribution of Followers (No Outliers)") +
  xlab("Number of Followers")

ggsave("results/histogram_followers_no_outliers.png", plot = histogram_
    followers, width = 8, height = 6)

# Day Time Minutes Analysis (No Outliers)

boxplot_day_time_min <- ggplot(as.data.frame(no_outliers_day_time_min), aes
    (y=no_outliers_day_time_min)) +
  geom_boxplot(fill="orange") +
  theme_minimal() +
  ggtitle("Boxplot of Day Time Minutes (No Outliers)") +
  ylab("Day Time Minutes")

ggsave("results/boxplot_day_time_min_no_outliers.png", plot = boxplot_day_
    time_min, width = 8, height = 6)

histogram_day_time_min <- ggplot(as.data.frame(no_outliers_day_time_min),
    aes(x=no_outliers_day_time_min)) +
  geom_histogram(fill="skyblue", color="black", bins=30) +
  theme_minimal() +
  ggtitle("Distribution of Day Time Minutes (No Outliers)") +
  xlab("Day Time Minutes")

ggsave("results/histogram_day_time_min_no_outliers.png", plot = histogram_
    day_time_min, width = 8, height = 6)

# Number of Posts Analysis (No Outliers)

boxplot_num_post <- ggplot(as.data.frame(no_outliers_num_post), aes(y=no_
    outliers_num_post)) +
  geom_boxplot(fill="orange") +
  theme_minimal() +
  ggtitle("Boxplot of Number of Posts (No Outliers)") +
  ylab("Number of Posts")

ggsave("results/boxplot_num_post_no_outliers.png", plot = boxplot_num_post,
     width = 8, height = 6)

histogram_num_post <- ggplot(as.data.frame(no_outliers_num_post), aes(x=no_
```

```
    outliers_num_post)) +
  geom_histogram(fill="skyblue", color="black", bins=30) +
  theme_minimal() +
  ggtitle("Distribution of Number of Posts (No Outliers)") +
  xlab("Number of Posts")

ggsave("results/histogram_num_post_no_outliers.png", plot = histogram_num_
    post, width = 8, height = 6)

# Clean Data Set

data_clean <- na.omit(data)

data_no_outliers <- subset(data_clean,
    story_views > story_views_lower_limit & story_views < story_views_upper
        _limit &
    num_follower > num_follower_lower_limit & num_follower < num_follower_
        upper_limit &
    day_time_min > day_time_min_lower_limit & day_time_min < day_time_min_
        upper_limit &
    num_post > num_post_lower_limit & num_post < num_post_upper_limit
)

original_count <- nrow(data_clean)
print(original_count)
final_count <- nrow(data_no_outliers)
print(final_count)
removed_count <- original_count - final_count
print(removed_count)
removed_percentage <- (removed_count / original_count) * 100
print(removed_percentage)

write.csv(data_no_outliers, "data/cleaned_data_final_no_outliers.csv", row.
    names = FALSE)


#B. Confidence Intervals
cat("\n \n")
cat("B1. Story Views (95% Confidence Interval)")

# Story Views (95% Confidence Interval)

# Overall 95% CI

ci_story_overall <- t.test(data_no_outliers$story_views, conf.level = 0.95)
cat("\n Story Views: Overall 95% CI \n")
cat(ci_story_overall$conf.int)
cat(paste("Mean:", round(ci_story_overall$estimate, 2)))

# Compare Only Child vs Others
```

```r
views_only_child <- subset(data_no_outliers, siblings == 0)$story_views
views_others <- subset(data_no_outliers, siblings > 0)$story_views

ci_story_only <- t.test(views_only_child, conf.level = 0.95)
ci_story_others <- t.test(views_others, conf.level = 0.95)

cat("\n Story Views: Only Child 95% CI \n")
cat(ci_story_only$conf.int)
cat(paste("Mean:", round(ci_story_only$estimate, 2)))

cat("\n Story Views: Others (Has Siblings) 95% CI \n")
cat(ci_story_others$conf.int)
cat(paste("Mean:", round(ci_story_others$estimate, 2)))

# B2. Number of Followers (99% Confidence Interval)

cat("\n \n")
cat("B2. Number of Followers (99% Confidence Interval)")

# Overall 99% CI
ci_followers <- t.test(data_no_outliers$num_follower, conf.level = 0.99)
cat("\n Followers: Overall 99% CI \n")
cat(ci_followers$conf.int)
cat(paste("Mean:", round(ci_followers$estimate, 2)))

# Compare Men vs Women
followers_men <- subset(data_no_outliers, sex == 'M')$num_follower
followers_women <- subset(data_no_outliers, sex == 'F')$num_follower
ci_followers_men <- t.test(followers_men, conf.level = 0.99)
ci_followers_women <- t.test(followers_women, conf.level = 0.99)
cat("\n Followers: Men 99% CI \n")
cat(ci_followers_men$conf.int)
cat(paste("Mean:", round(ci_followers_men$estimate, 2)))
cat("\n Followers: Women 99% CI \n")
cat(ci_followers_women$conf.int)
cat(paste("Mean:", round(ci_followers_women$estimate, 2)))



# B3. Proportion of Italian Accounts (90% Confidence Interval)
cat("B3. Proportion of Italian Accounts (90% Confidence Interval)")
n_total <- nrow(data_no_outliers)
n_italian <- sum(data_no_outliers$language == "Italian")
ci_italian <- prop.test(n_italian, n_total, conf.level = 0.90)
cat("\n Language: Italian Proportion 90% CI \n")
cat(ci_italian$conf.int)
cat(paste("Proportion:", round(ci_italian$estimate, 4)))

# C1. Hypothesis Testing
```

```r
# Followers: Men vs Women
# H0: Mean followers (Men) = Mean followers (Women)
# H1: Mean followers (Men) != Mean followers (Women)

followers_men <- subset(data_no_outliers, sex == 'M')$num_follower
followers_women <- subset(data_no_outliers, sex == 'F')$num_follower
test_c1 <- t.test(followers_men, followers_women)
cat("\nC1. Followers: Men vs Women\n")
print(test_c1)

# C2. Story Views: Public vs Private University
# H0: Mean views (Public) = Mean views (Private)
# H1: Mean views (Public) != Mean views (Private)
views_public <- subset(data_no_outliers, private_d == 0)$story_views
views_private <- subset(data_no_outliers, private_d == 1)$story_views
test_c2 <- t.test(views_public, views_private)
cat("\nC2. Story Views: Public vs Private University\n")
print(test_c2)

# C3. Daily Time: English vs Italian
# H0: Mean time (English) = Mean time (Italian)
# H1: Mean time (English) != Mean time (Italian)
time_english <- subset(data_no_outliers, language == 'English')$day_time_
    min
time_italian <- subset(data_no_outliers, language == 'Italian')$day_time_
    min
test_c3 <- t.test(time_english, time_italian, conf.level = 0.99)
cat("\nC3. Daily Time: English vs Italian\n")
print(test_c3)

# D. Linear Regression Analysis

# D1. Simple Linear Regression
cat("\nD1. Simple Linear Regression Results \n")
model_simple <- lm(story_views ~ num_follower, data = data_no_outliers)
summary_simple <- summary(model_simple)
anova_simple <- anova(model_simple)
print (summary_simple)
print (anova_simple)


png("results/model_simple_plot.png")
plot(model_simple)
dev.off()

# Residuals vs Fitted
png("results/residuals_vs_fitted_simple_plot.png")
plot(model_simple, which = 1)
dev.off()
```

```r
# Normal Q-Q
png("results/normal_QQ_plot.png")
plot(model_simple, which = 2)
dev.off()

# Scale-Location
png("results/scale_location_simple_plot.png")
plot(model_simple, which = 3)
dev.off()

# Residuals vs Leverage
png("results/residuals_vs_leverage_simple_plot.png")
plot(model_simple, which = 5)
dev.off()

#Create Coefficient Table

simple_coef_df <- tidy(model_simple) %>%
  mutate(across(where(is.numeric), \(x) round(x, 4)))
table_simple_coef <- ggtexttable(simple_coef_df, rows = NULL,
                                 theme = ttheme("default"))
ggexport(table_simple_coef, filename = "results/table_simple_regression.png
    ")

#Create ANOVA Table

simple_anova_df <- tidy(anova(model_simple)) %>%
  mutate(across(where(is.numeric), \(x) round(x, 4)))

table_simple_anova <- ggtexttable(simple_anova_df, rows = NULL,
                                  theme = ttheme("default"))
ggexport(table_simple_anova, filename = "results/table_simple_anova.png")


# Visualizing the Linear Regression (Story Views vs Followers)

plot_linear_reg <- ggplot(data_no_outliers, aes(x = num_follower, y = story
    _views)) +
  geom_point(color = "steelblue", alpha = 0.6) +
  geom_smooth(method = "lm", color = "red", se = TRUE) +

  labs(title = "Linear Regression: Story Views vs Number of Followers",
       subtitle = "Visualizing the relationship with regression line and
           95% CI",
       x = "Number of Followers",
       y = "Story Views") +
  theme_minimal()

ggsave("results/plot_linear_regression_1.png", plot = plot_linear_reg,
```

```
    width = 8, height = 6)

# Visual Check: Normality (The QQ Plot)

plot_qq <- autoplot(model_simple, which = 2, ncol = 1, label.size = 3) +
  theme_minimal() +
  ggtitle("Normal Q-Q Plot")

ggsave("results/linear_model_qq_plot.png", plot = plot_qq, width = 6,
    height = 4)

# Statistical Tests for Assumptions

# Normality Test (Shapiro-Wilk)
# H0: Residuals are normally distributed.
shapiro_test <- shapiro.test(residuals(model_simple))
cat("\nNormality Assumption (Shapiro-Wilk Test)\n")
print(shapiro_test)

# Homoscedasticity Test (Breusch-Pagan)
# H0: Variance is constant.
bp_test <- bptest(model_simple)
cat("\nHomoscedasticity Assumption (Breusch-Pagan Test)\n")
print(bp_test)


# D2. Multiple Linear Regression
cat("\n D2. Multiple Linear Regression Results \n")
model_multi <- lm(story_views ~ num_follower + sex + account_num + num_post
     + day_time_min, data = data_no_outliers)
summary_multi <- summary(model_multi)
anova_multi <- anova(model_multi)
print (summary_multi)
print (anova_multi)
png("results/model_multi_plot.png")
plot(model_multi)
dev.off()

# Create Coefficient Table for Multiple Regression
multi_coef_df <- tidy(model_multi) %>%
  mutate(across(where(is.numeric), \(x) round(x, 4)))

table_multi_coef <- ggtexttable(multi_coef_df, rows = NULL,
                                theme = ttheme("default"))

ggexport(table_multi_coef, filename = "results/table_multi_regression.png")

# E. Prediction

# Median Data for a Female Account
```

```r
median_data_female <- data.frame(
  sex = "F",
  num_follower = median(data_no_outliers$num_follower, na.rm = TRUE),
  account_num = median(data_no_outliers$account_num, na.rm = TRUE),
  num_post = median(data_no_outliers$num_post, na.rm = TRUE),
  day_time_min = median(data_no_outliers$day_time_min, na.rm = TRUE)
)

# E1. Predict expected number of views

prediction_val <- predict(model_multi, newdata = median_data_female)

cat("E1. Predicted Expected Views (Female Median Account):\n")
cat(round(prediction_val, 2))

# E1. 95% Confidence Interval

prediction_ci <- predict(model_multi, newdata = median_data_female,
    interval = "confidence", level = 0.95)

cat("E1. 95% Confidence Interval:\n")
print(prediction_ci)

# F. Logistic Regression

plot_boxplot_private <- ggplot(data_no_outliers, aes(x = as.factor(private_
    d), y = story_views)) +
  geom_boxplot(fill = c("lightblue", "orange")) +
  labs(title = "Distribution of Story Views by University Type",
       subtitle = "0 = Public, 1 = Private",
       x = "University Type",
       y = "Story Views") +
  theme_minimal()

ggsave("results/plot_boxplot_private.png", plot = plot_boxplot_private,
    width = 8, height = 6)

# F1. Logistic Regression Model

model_logit <- glm(private_d ~ num_follower + num_post + story_views,
                   data = data_no_outliers,
                   family = binomial)

cat("\nF1. Logistic Regression Results\n")
print(summary(model_logit))
print(anova(model_logit, test="Chisq"))

# Odds Ratios
cat("\nOdds Ratios:\n")
```

```r
odds_ratios <- exp(coef(model_logit))
print(odds_ratios)
```