SUMMARY WITH MAIN RESULTS

Il progetto ha avuto come obiettivo l'esplorazione di un dataset e l'utilizzo di modelli di machine learning. L'analisi esplorativa del dataset considerato è stata svolta andando ad evidenziare le principali statistiche descrittive e realizzando come supporto dei grafici esplicativi.

Nella fase del progetto relativa al machine learning è stata utilizzata la classificazione secondo due approcci diversi: alberi di classificazione e knn. Le due metodologie, seppur diverse, hanno portato a risultati simili.

INTRODUCTION

Gli account fake e spam rappresentano uno dei problemi principali per le varie piattaforme social media. Tra queste Instagram.

L'obiettivo del progetto è quello di analizzare il dataset selezionato e riuscire ad indentificare queste tipologie di account tramite tecniche di machine learning.

Il dataset è stato scaricato da: https://www.kaggle.com/free4ever1/instagram-fake-spammer-genuine-accounts#test.csv.

Per identificare i possibili profili fake si cercherà di analizzare le principali caratteristiche di un profilo Instagram, quali il numero di followers, di follows, di posts, evidenziandone possibili relazioni tra esse.

DATA ACQUISITION

Il dataset inizialmente si è presentato diviso in due file csv: train e test. Mediante opportune funzioni in R è stato creato un unico dataset contente il totale delle osservazioni.

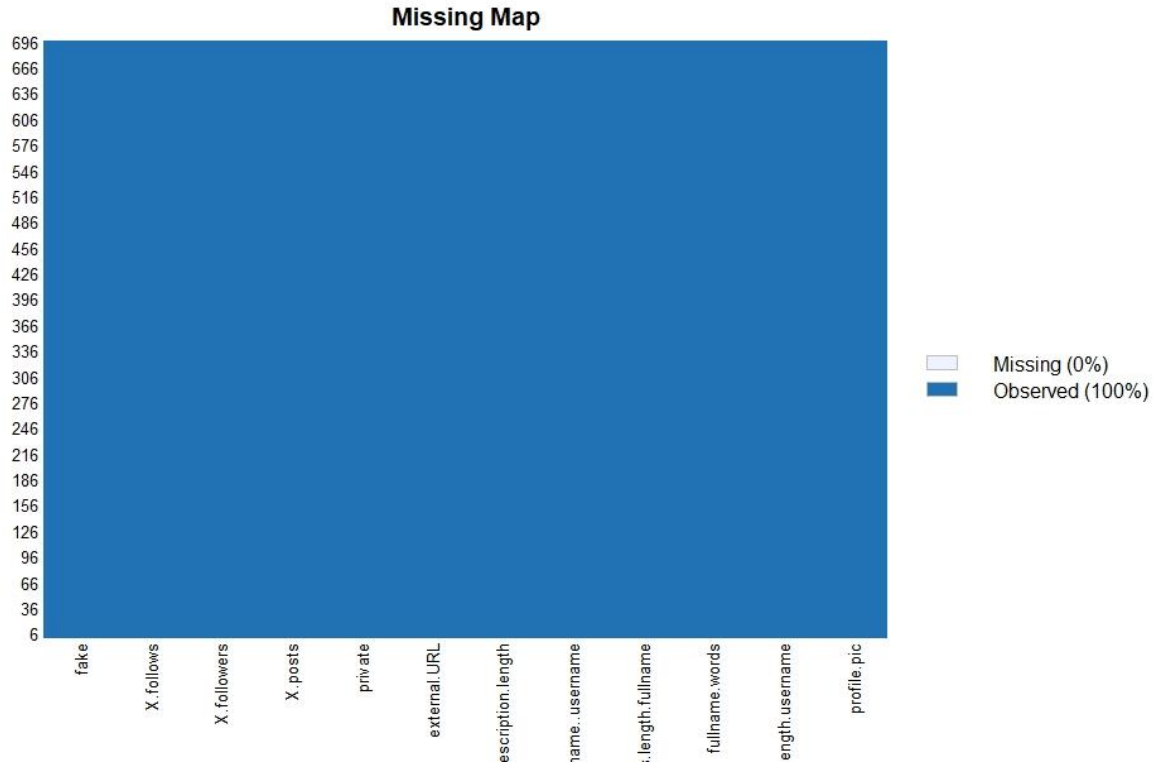
Il dataset è costituito da 696 osservazioni ed è caratterizzato da 12 variabili.

Quest'ultime risultano essere:

profile.pic	*È un valore booleano che indica se lo user ha la foto del profilo (1) oppure no (0)*
nums.length.username	*Rappresenta il rapporto dei caratteri numerici su tutta la lunghezza dello username*
fullname.words	*Indica il numero di tokens che costituiscono il nome completo dello user. Il nome completo può non essere presente.*
nums.length.fullname	*Rappresenta il rapporto dei caratteri numerici su tutta la lunghezza del nome completo dello user*
name..username	*È un valore booleano che indica se lo username coincide con il nome completo dello user (1) oppure no (0)*
description.length	*Indica la lunghezza in caratteri della bio dello user*
external.URL	*È un valore booleano che indica la presenza di un link esterno sul profilo dello user (1) oppure no (0)*
private	*È un valore booleano che indica se il profilo dello user è privato (1) oppure no (0)*
X.posts	*Indica il numero di post presenti sul profilo dello user*
X.followers	*Indica il numero di seguaci del profilo dello user*
X.follows	*Indica il numero di persone seguite dal profilo dello user*
fake	*È un valore booleano che indica se il profilo dello user è fake (1) oppure no (0)*

DATA PRE-PROCESSING

Tramite opportune funzioni in R si è verificata la presenza di missing values ma si è ottenuto un esito negativo: nel dataset non ci sono valori mancanti o valori errati.



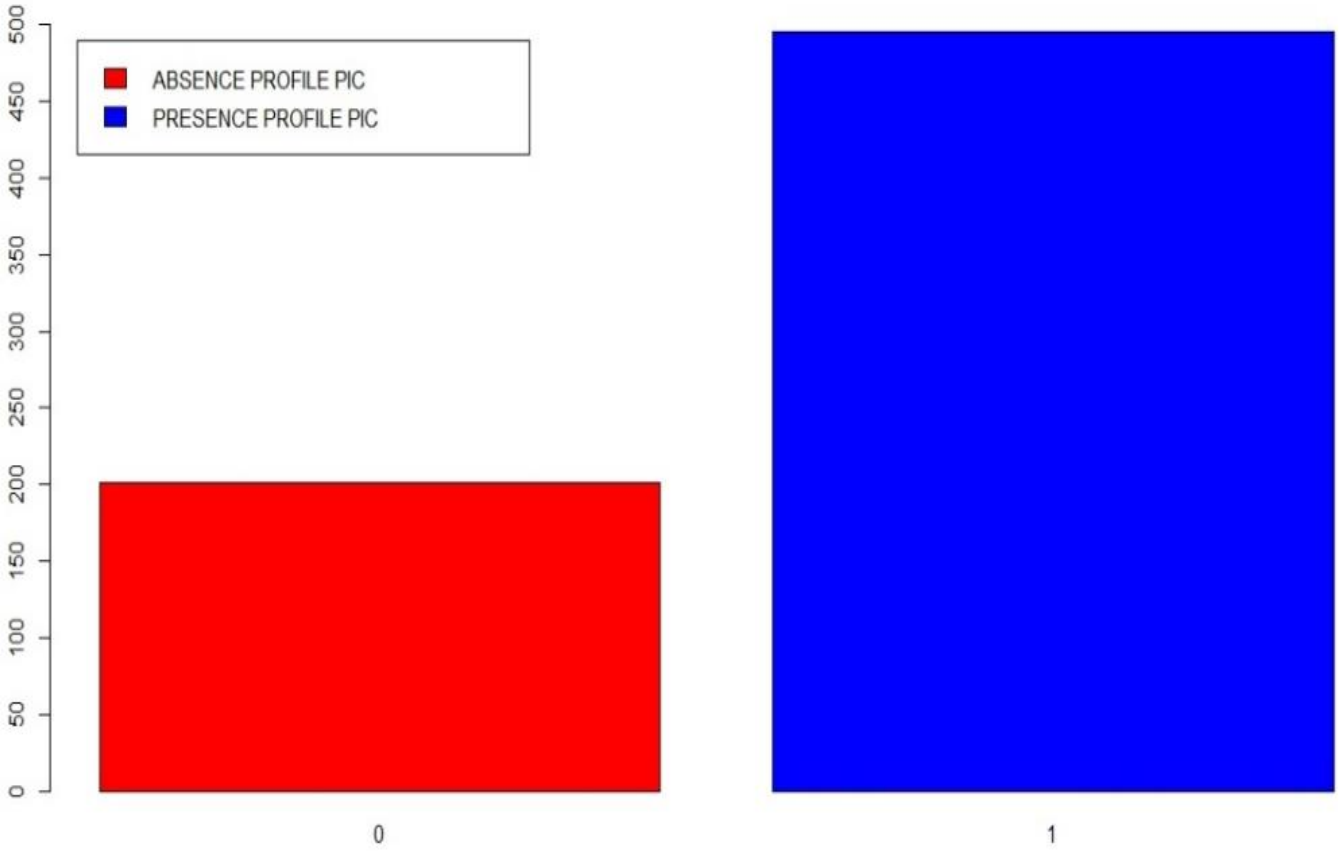
Gli attributi *profile.pic*, *fullname.words, name..username*, *external.URL*, *private* e *fake* essendo di tipo int sono stati convertiti in factor per avere una migliore gestione del dataset.

EXPLORATORY ANALYSIS

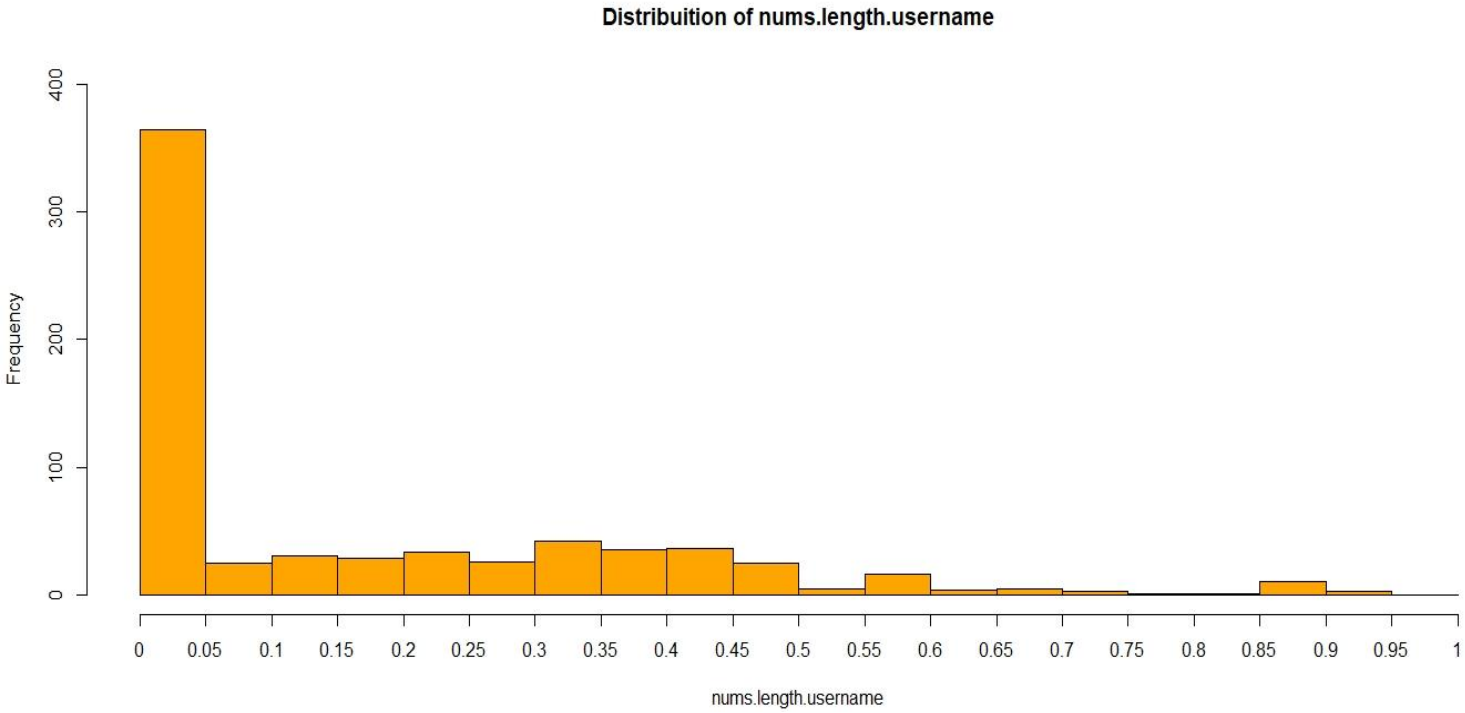
Successivamente si è passati ad un'analisi esplorativa più mirata del dataset facendo uso delle principali statistiche descrittive. Per rendere più chiari i concetti sono stati realizzati anche dei grafici.

Considerando le singole variabili:

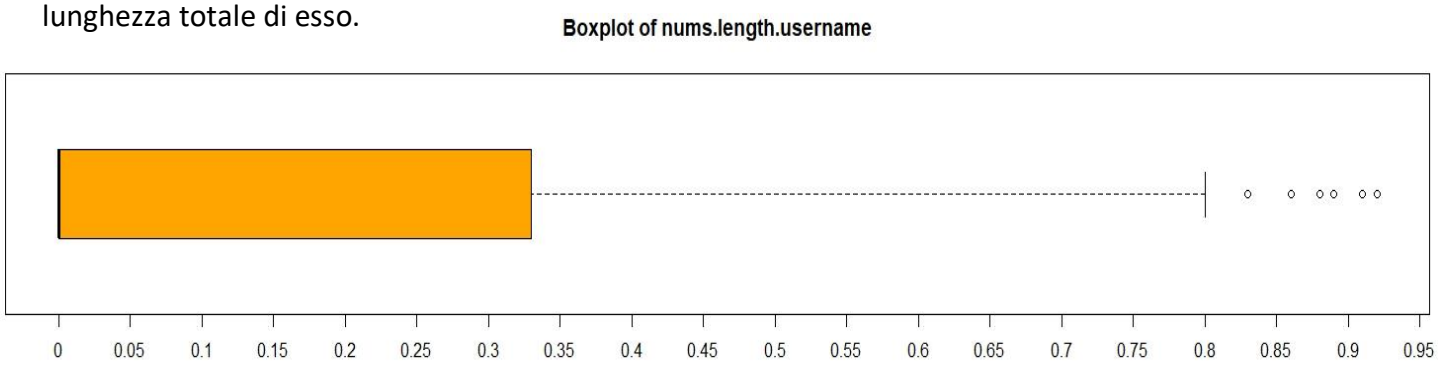
-**profile.pic**: 495 profili su un totale di 696 hanno una foto profilo. Dunque, più della metà dei profili osservati possiede una foto profilo

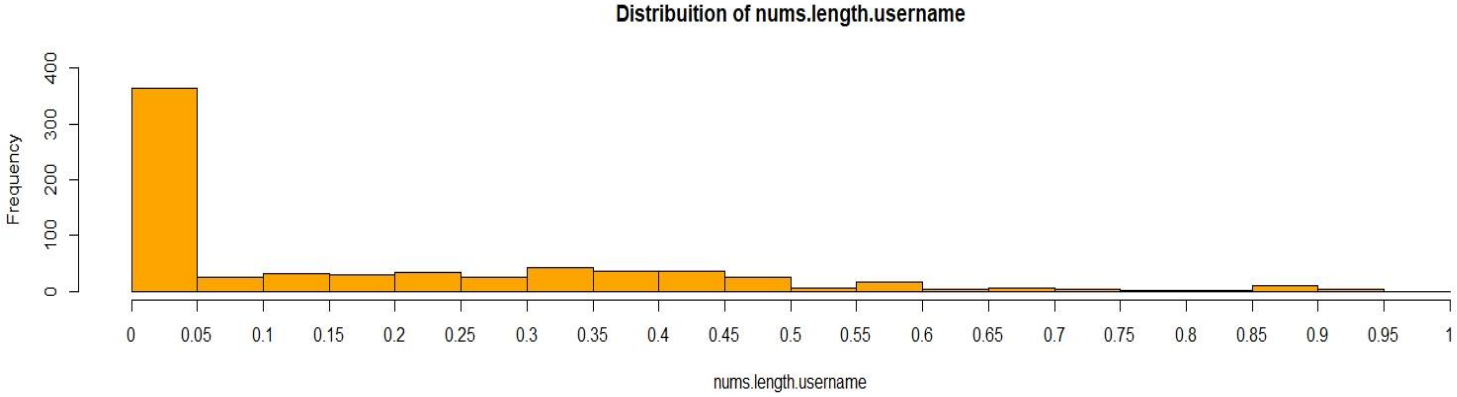


-**nums.length.username**: il valore minore risulta essere 0, il valore massimo 0.92. Viene assunto il valore 0 in 363 osservazioni su un totale di 696. La moda e la mediana confermano ciò. La media è pari a 0.1666092 e la varianza è pari a 0.04794503



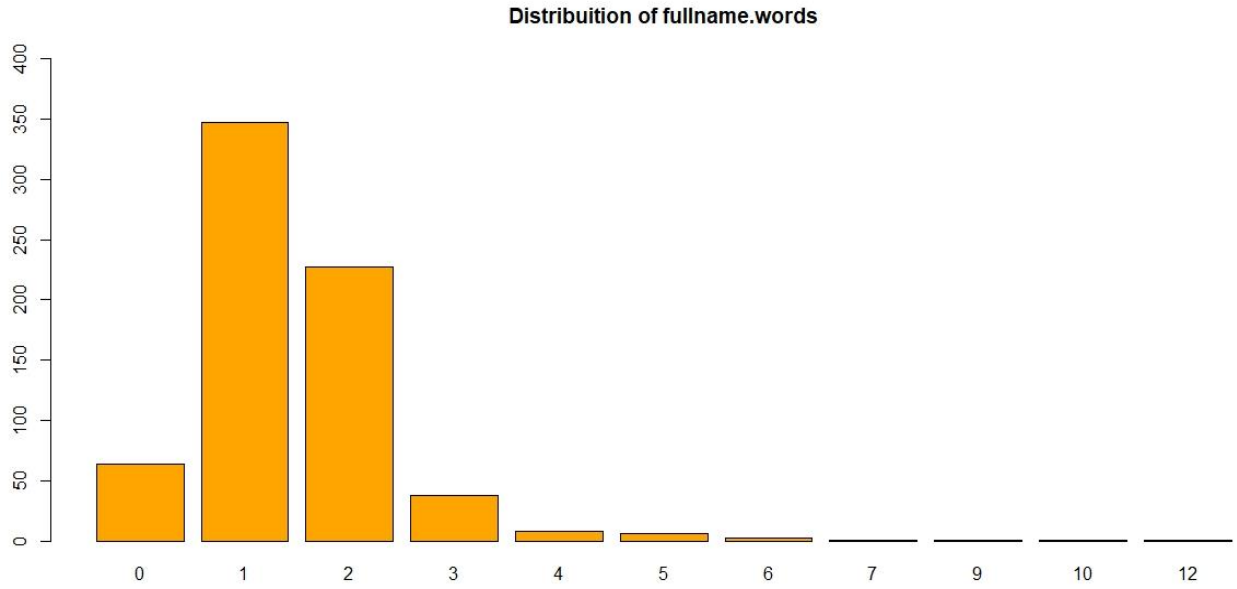
È stato generato anche un box plot per comprendere meglio la distribuzione della variabile considerata. Come si può osservare il 50% delle osservazioni si colloca tra 0 e un valore inferiore a 0.35, rispettivamente il primo e il terzo quartile. Questo risultato significa che circa metà degli utenti osservati hanno uno username la cui quantità di caratteri numerici non supera il 33% della lunghezza totale di esso.



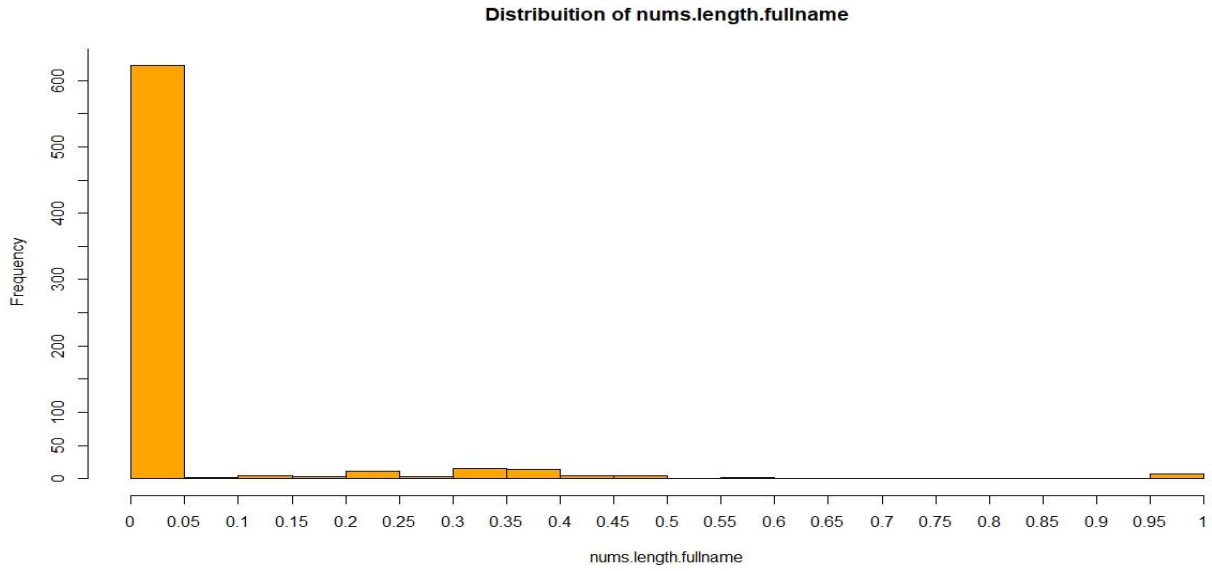


-**fullname.words**: il valore minore risulta essere 0, il valore massimo 12 e quello medio 1.475575 (da approssimare ad un valore discreto). La mediana e la moda sono entrambe pari ad 1 mentre la varianza assume il valore di 1.159115. La maggior parte degli utenti, perciò, utilizza una sola parola come fullname (ad esempio si può pensare che si preferisce inserire solo il nome e non il cognome)

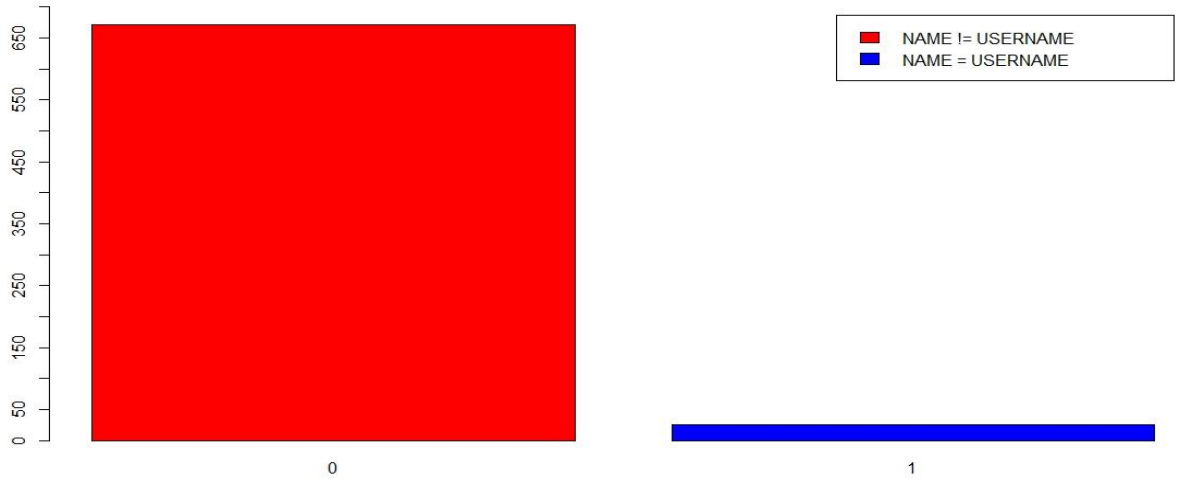
n° tokens	0	1	2	3	4	5	6	7	9	10	12
n° obs	64	347	227	38	8	6	2	1	1	1	1



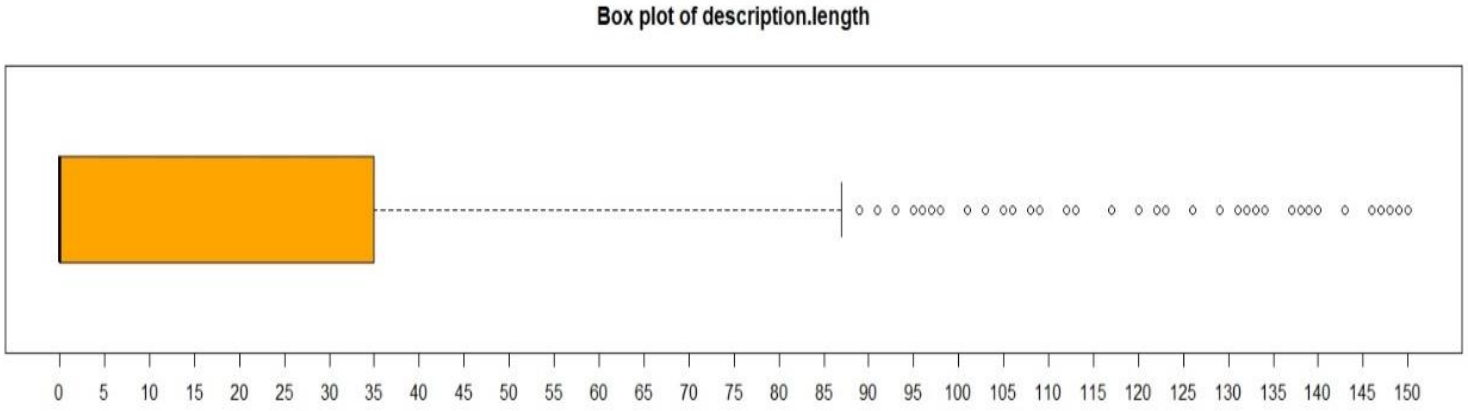
-**nums.length.fullname**: il valore minore risulta essere 0, il valore massimo 1. Su un totale di 696 osservazioni in 622 casi la variabile assume il valore 0. La varianza è pari a 0.02063946. Questi risultati suggeriscono che nella quasi totalità degli utenti considerati è raro trovare dei caratteri numerici all'interno del fullname. È più probabile trovarli all'interno dello username come visto in precedenza

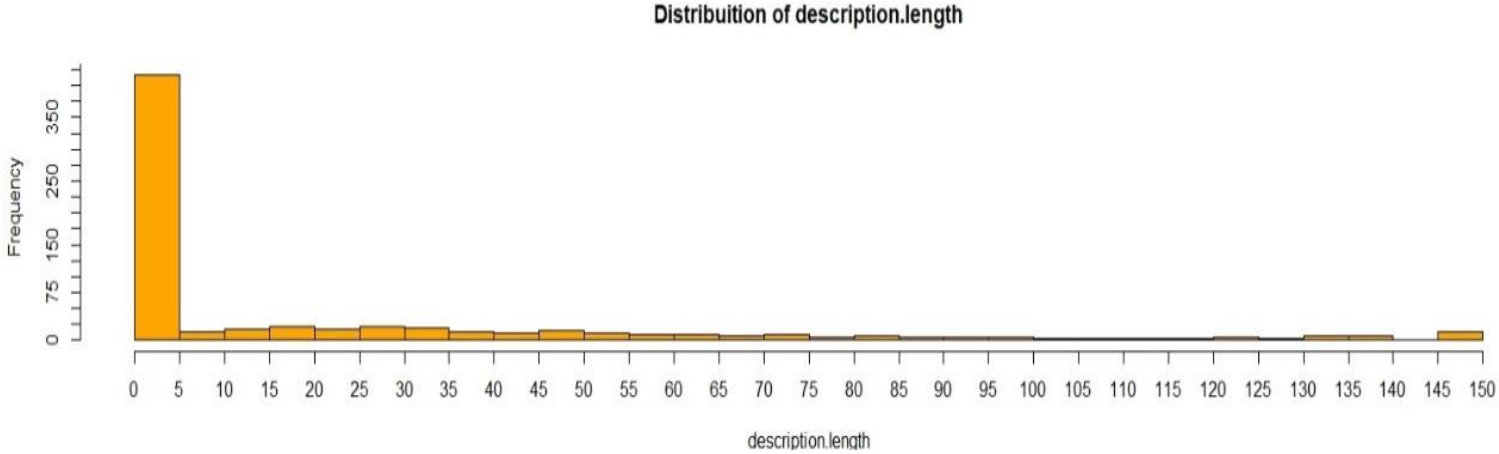


-**name..username**: solo in rari casi (25 su 696) lo username corrisponde con il fullname del profilo dello user. Questa osservazione vale considerando sia profili fake sia profili non fake. La quasi totalità degli utenti tende a ud utilizzare un username che si discosta dal fullname

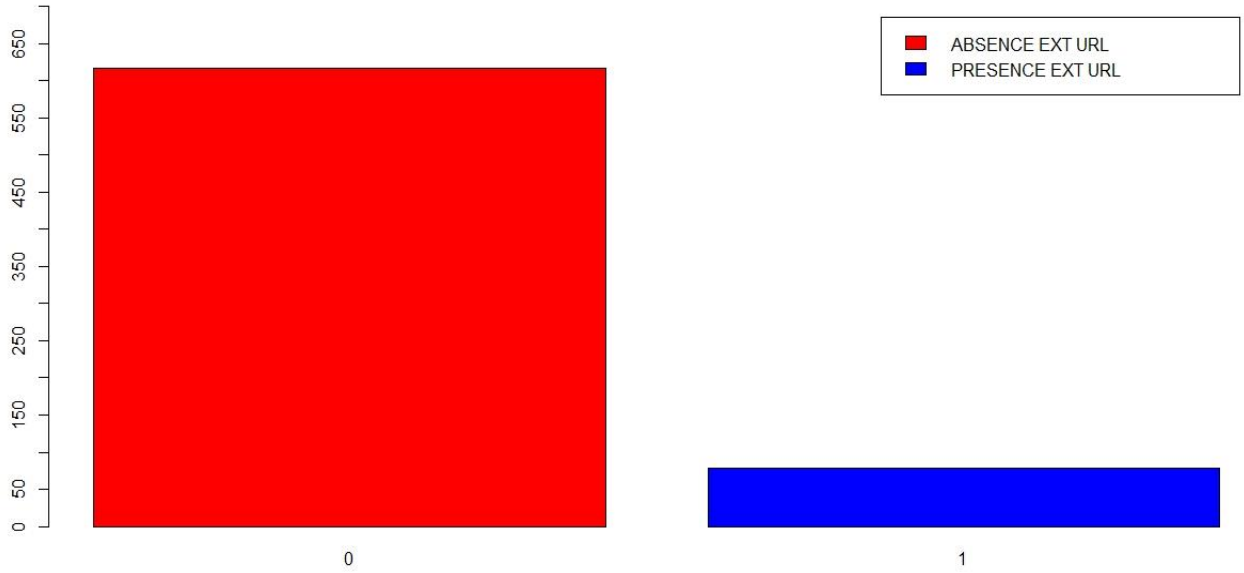


-**description.length**: il valore minore risulta essere 0, il valore massimo 150. La moda risulta essere pari a 0. Infatti, più della metà dei profili analizzati non possiede una description. Come si può notare dal grafico del box plot per metà dei profili osservati la description ha una lunghezza minore di 35 (che risulta essere il terzo quartile)

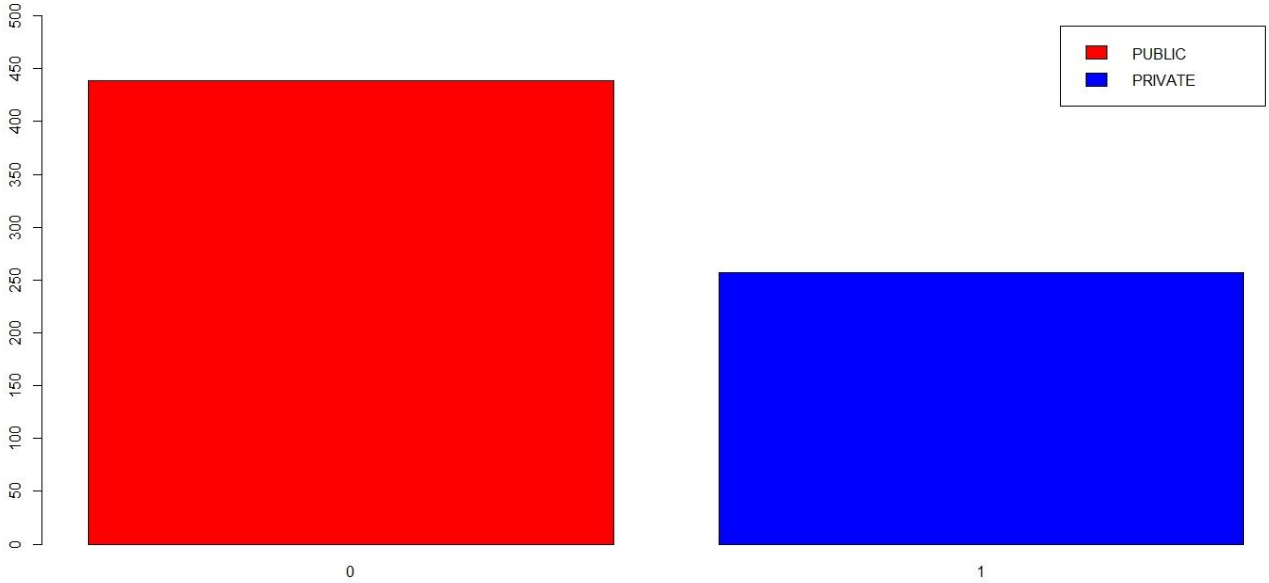




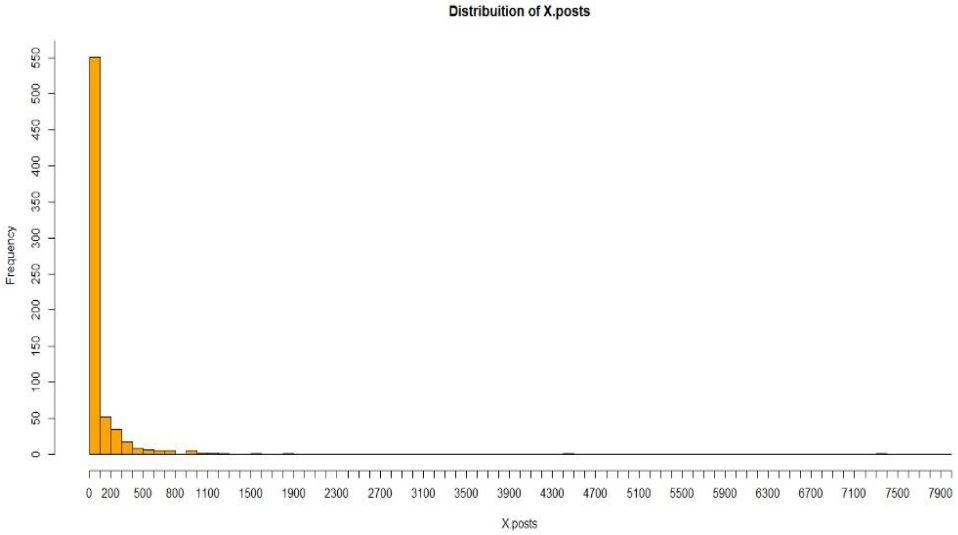
-**external.URL**: la maggior parte dei profili (617) non possiede un link esterno



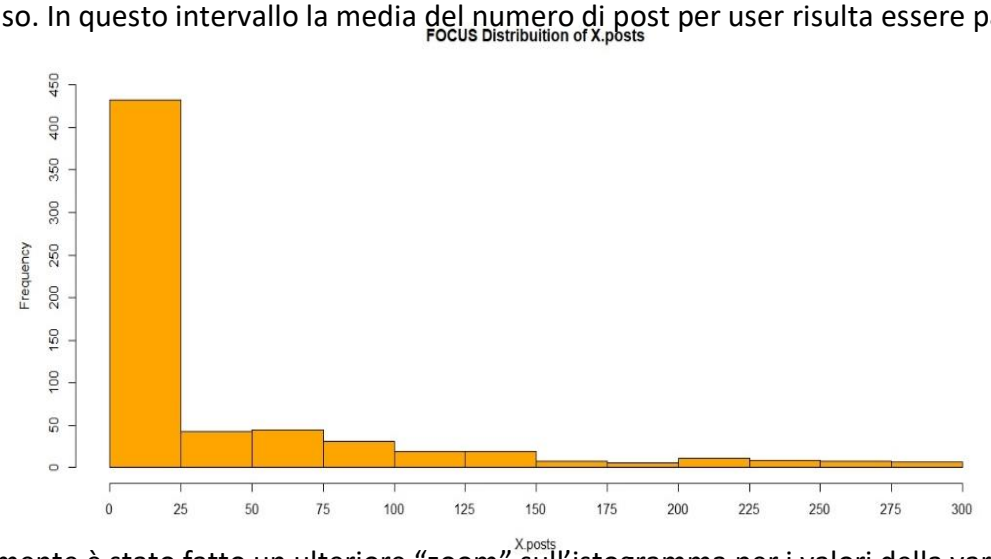
-**private**: sono presenti in numero maggiore i profili pubblici (439) rispetto a quelli privati (257)



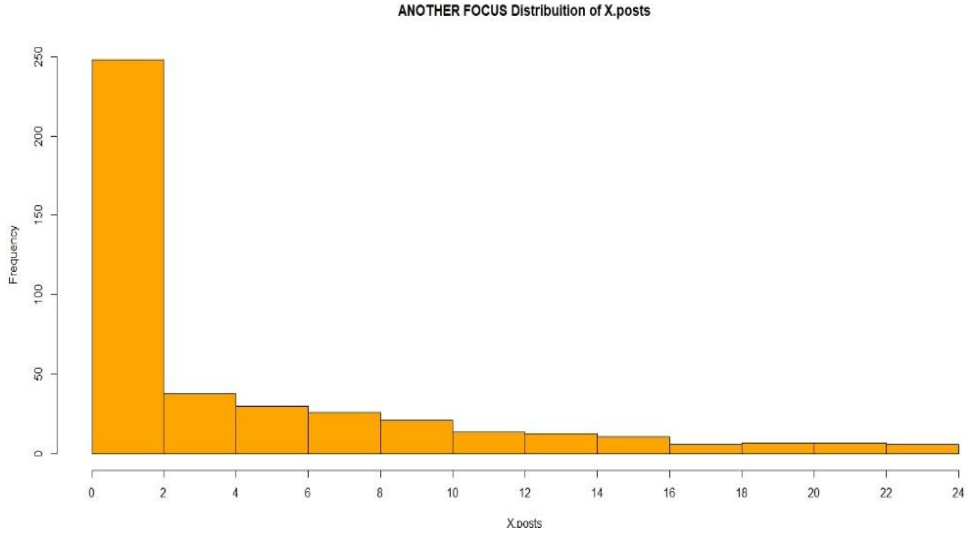
-**X.posts**: il valore minore risulta essere 0, il valore massimo 7389. La variabile assume in 185 osservazioni il valore pari a 0, che risulta essere la moda



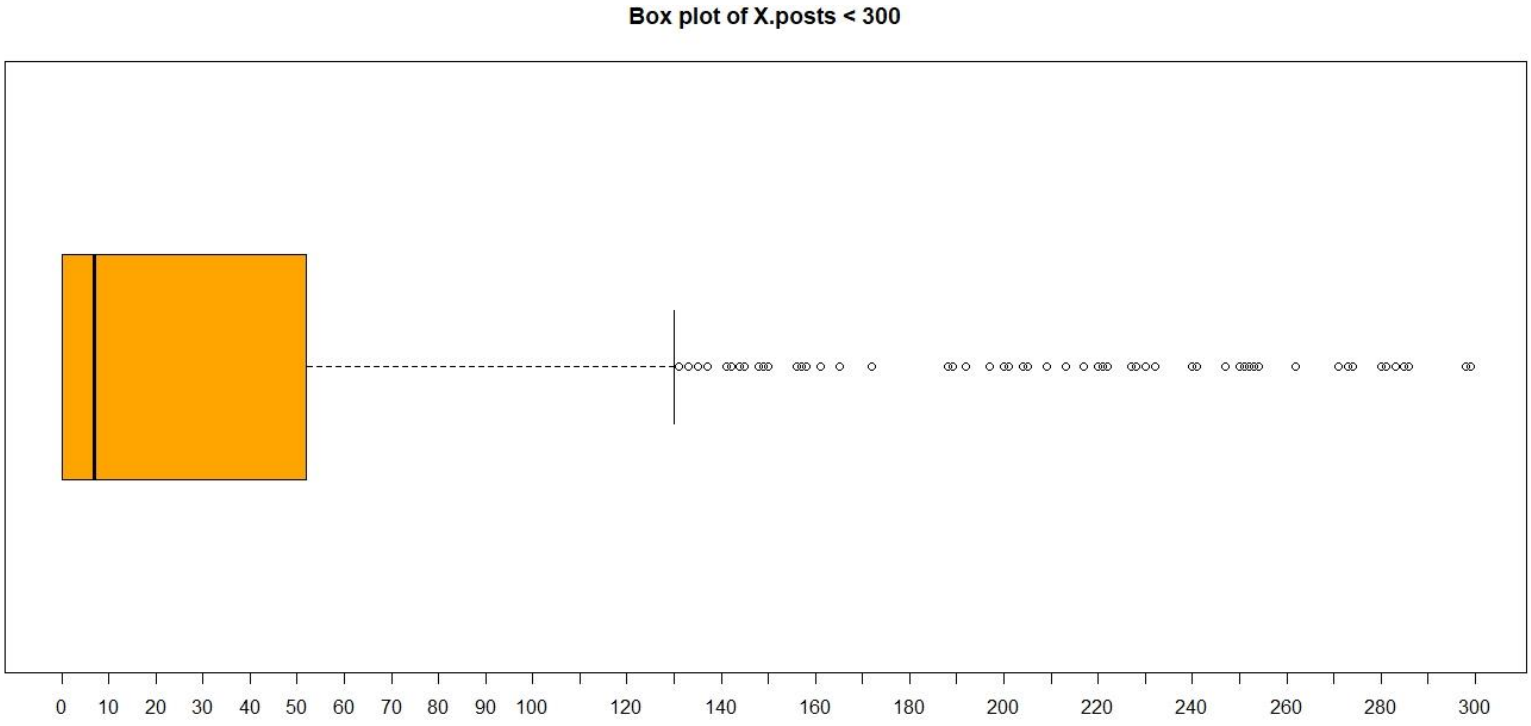
Dal momento in cui la maggior parte delle osservazioni (638 su 696) si possono individuare per valori inferiori a 300 è stata analizzata la parte di istogramma ove il valore della variabile varia da 0 a 300 escluso. In questo intervallo la media del numero di post per user risulta essere pari a 39.



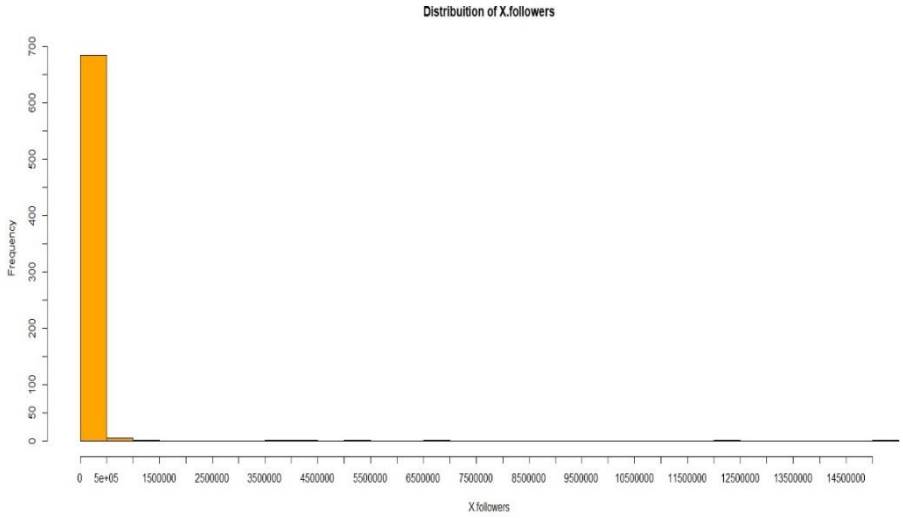
Successivamente è stato fatto un ulteriore “zoom” sull’istogramma per i valori della variabile che vanno da 0 a 25 escluso, rappresentando gran parte delle osservazioni sull’intero dataset (427 su 696)



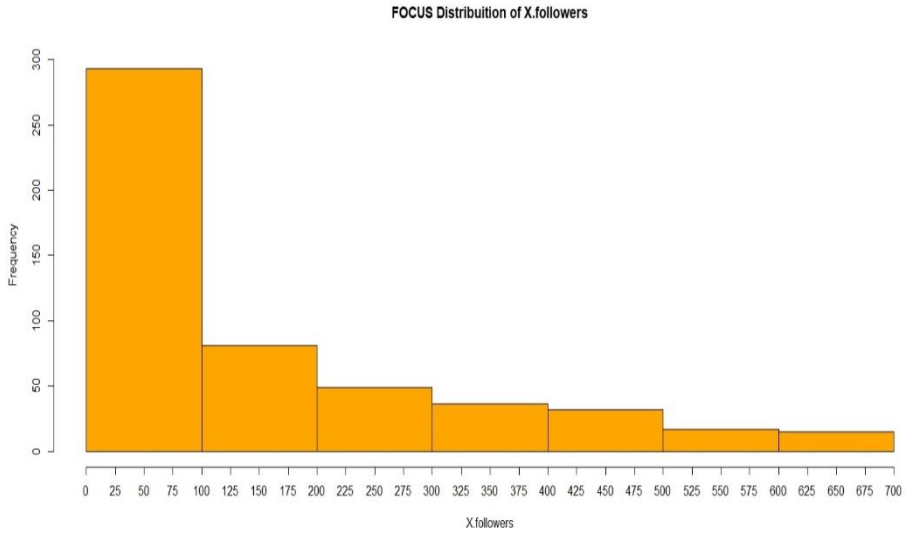
Per avere maggior chiarezza e una conferma delle osservazioni fatte in precedenza è stato generato un box plot. Dal grafico possiamo dunque constatare che il 50% delle osservazioni (considerando solo le variabili che assumono un valore che varia da 0 a 300 escluso) si collocano tra 0 e 52. Dunque, è possibile osservare che almeno la metà dei profili hanno un numero di posts nel proprio profilo il cui numero va da un minimo di 0 ad un massimo di 52.



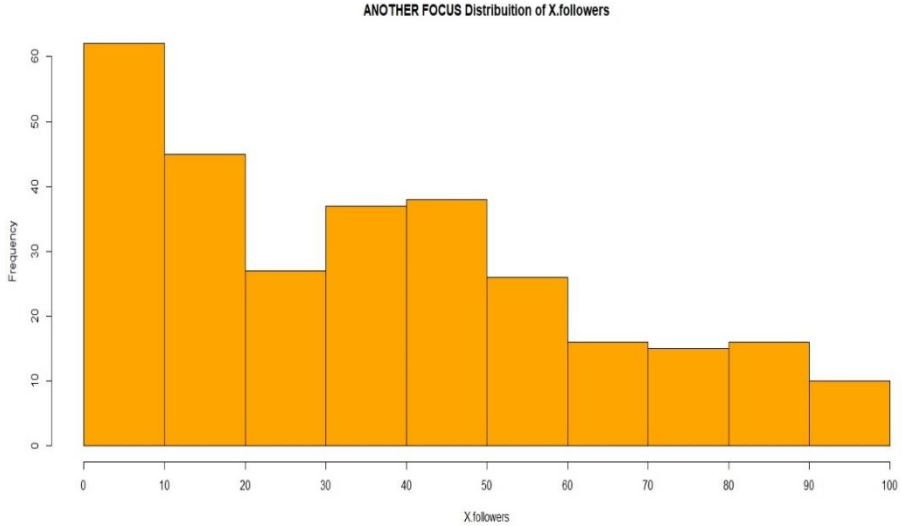
-**X.followers**: il valore minore risulta essere 0, il valore massimo 15338538. La variabile assume in 20 osservazioni il valore pari a 0, che risulta essere la moda



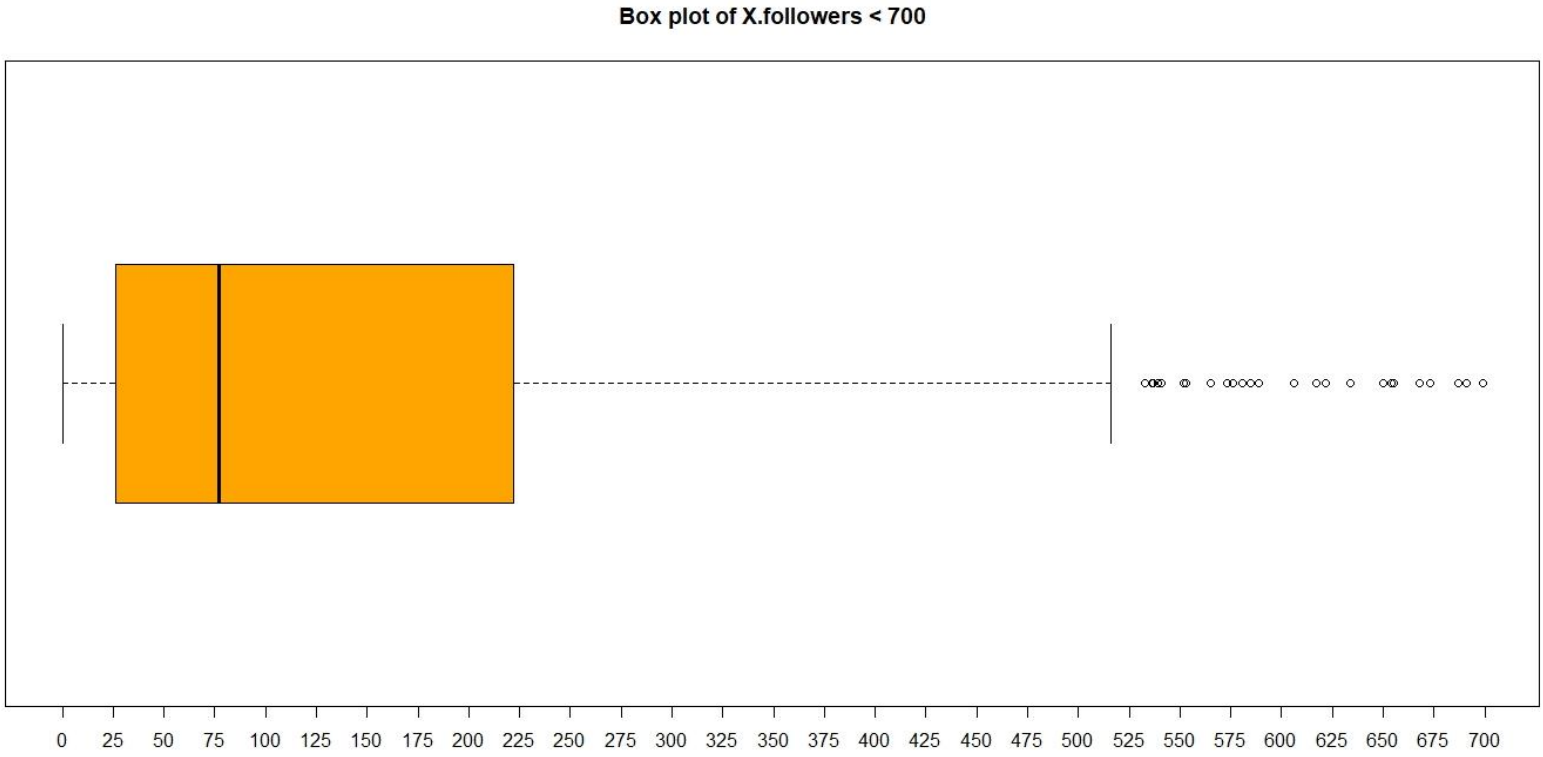
Dal momento in cui la maggior parte delle osservazioni (523 su 696) si possono individuare per valori inferiori a 700 è stata analizzata la parte di istogramma ove il valore della variabile varia da 0 a 700 escluso. In questo intervallo la media risulta essere pari a 154, ovvero i profili tendono ad avere in media 154 followers.



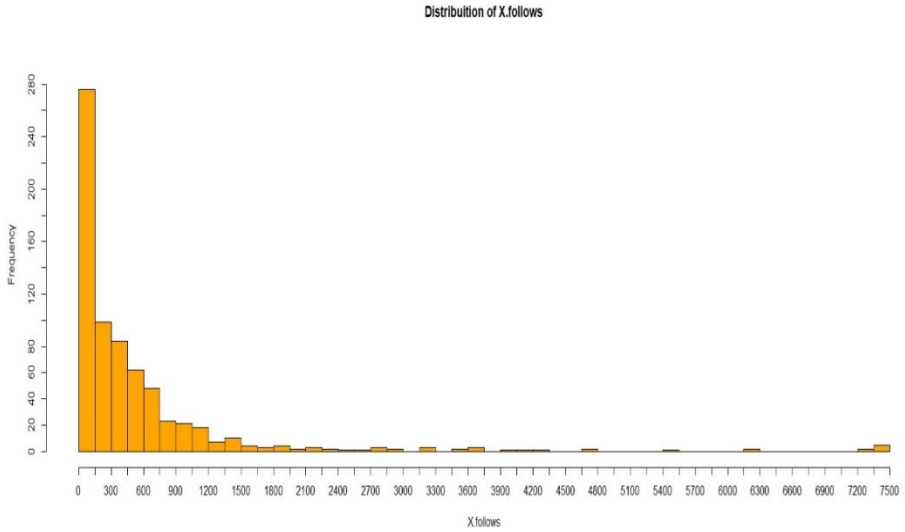
Successivamente è stato fatto un ulteriore “zoom” sull’istogramma per i valori della variabile che vanno da 0 a 100 escluso, rappresentando 292 osservazioni su un totale di 696



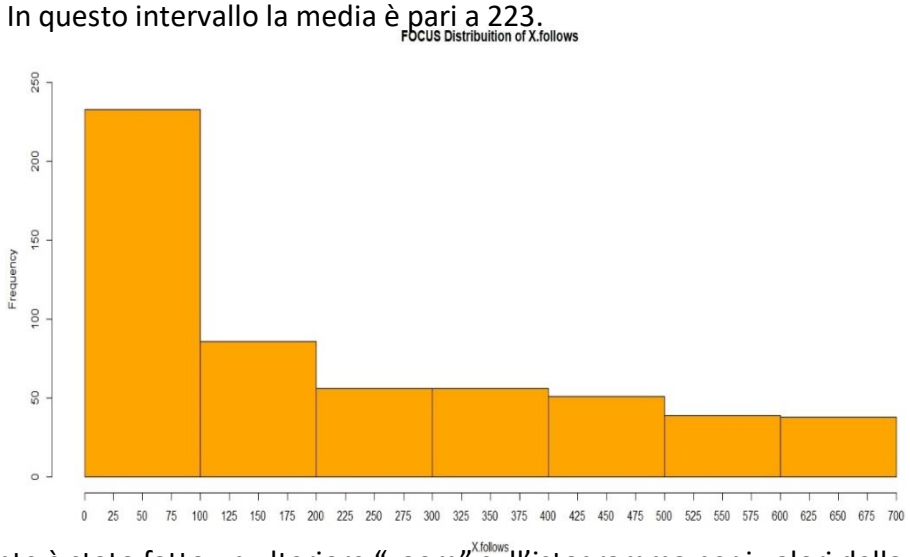
Come nel caso della variabile X.posts per una maggior chiarezza è stato generato un box plot. Dal grafico possiamo dunque constatare che il 50% delle osservazioni (considerando solo le variabili che assumono un valore che varia da 0 a 700 escluso) si colloca tra circa 26 e 222. Il valore minimo risulta essere 0 mentre quello massimo supera 500. Si possono osservare anche diversi outliers, profili il cui numero di followers è maggiore di 500 circa.



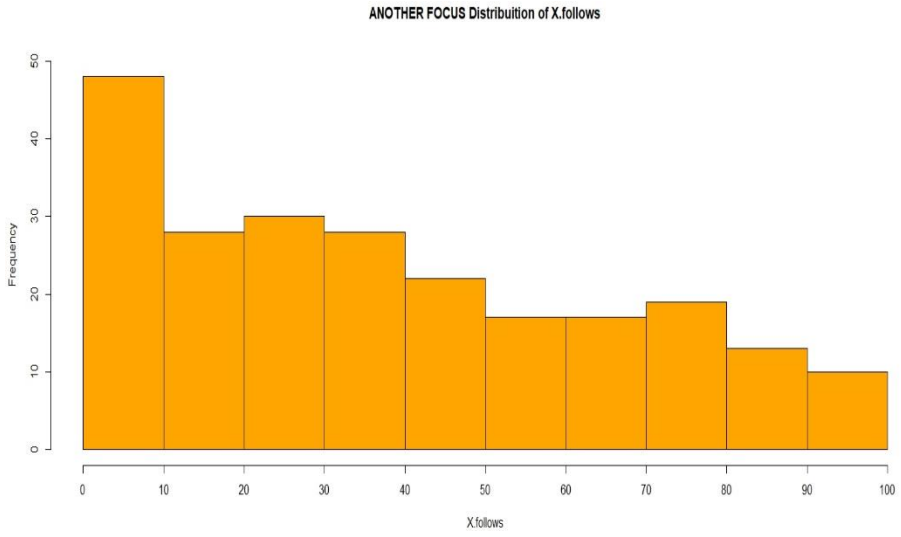
-**X.follows**: il valore minore risulta essere 0, il valore massimo 7500. La variabile assume in 11 osservazioni il valore pari a 0, che risulta essere la moda



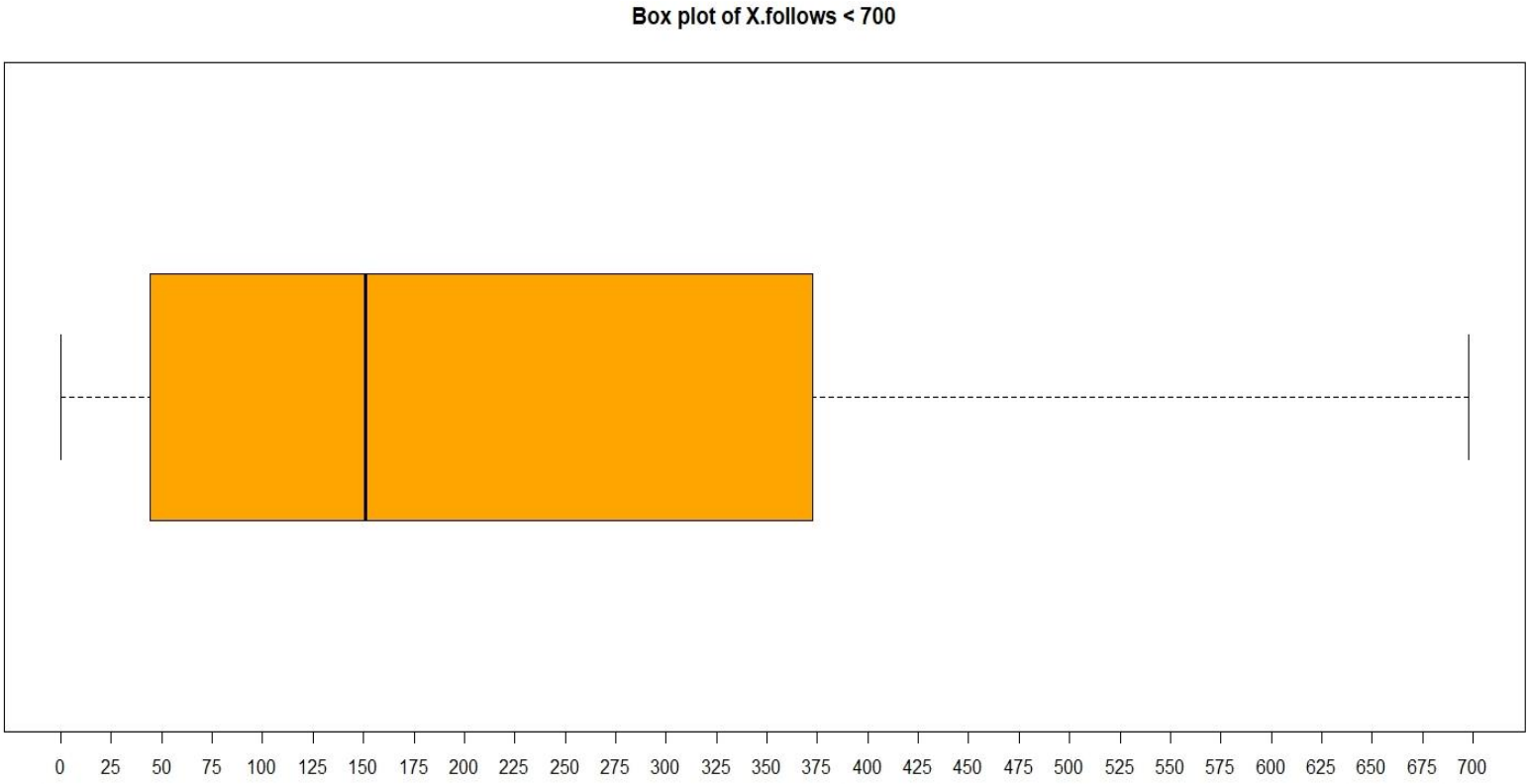
Dal momento in cui la maggior parte delle osservazioni (559 su 696) si possono individuare per valori inferiori a 700 è stata analizzata la parte di istogramma ove il valore della variabile varia da 0 a 700 escluso. In questo intervallo la media è pari a 223.



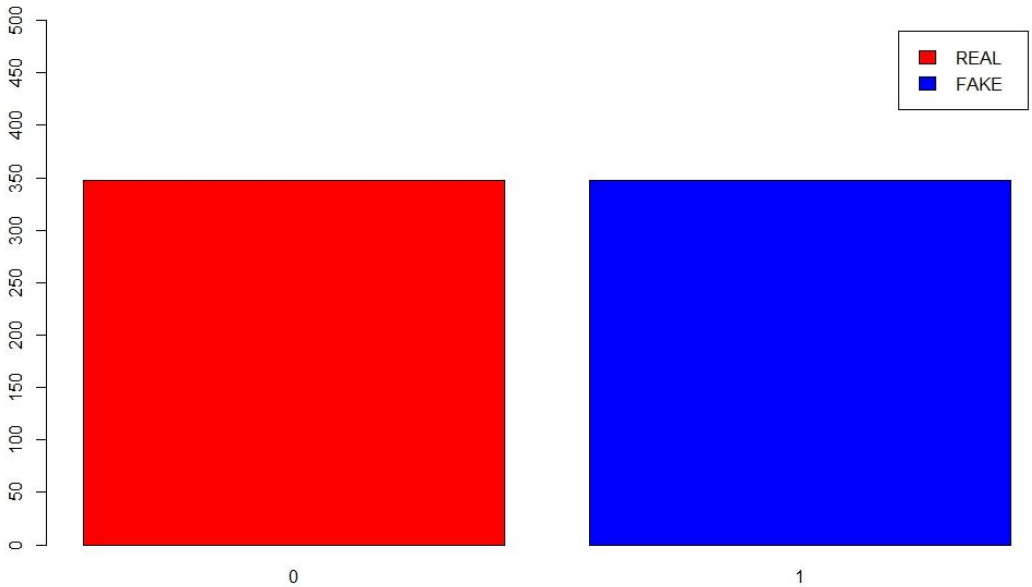
Successivamente è stato fatto un ulteriore "zoom" sull'istogramma per i valori della variabile che vanno da 0 a 100 escluso, rappresentando 232 osservazioni su un totale di 696



È stato successivamente generato un box plot considerando l'intervallo che va da 0 a 700 della variabile X.follows. In quest'ultimo intervallo il 50% degli user seguono un numero di profili che varia da un valore di 44 fino ad valore pari a 372.5.



-**fake**: sono presenti in quantità uguali all'interno del dataset profili reali (348) e profili fake (348)

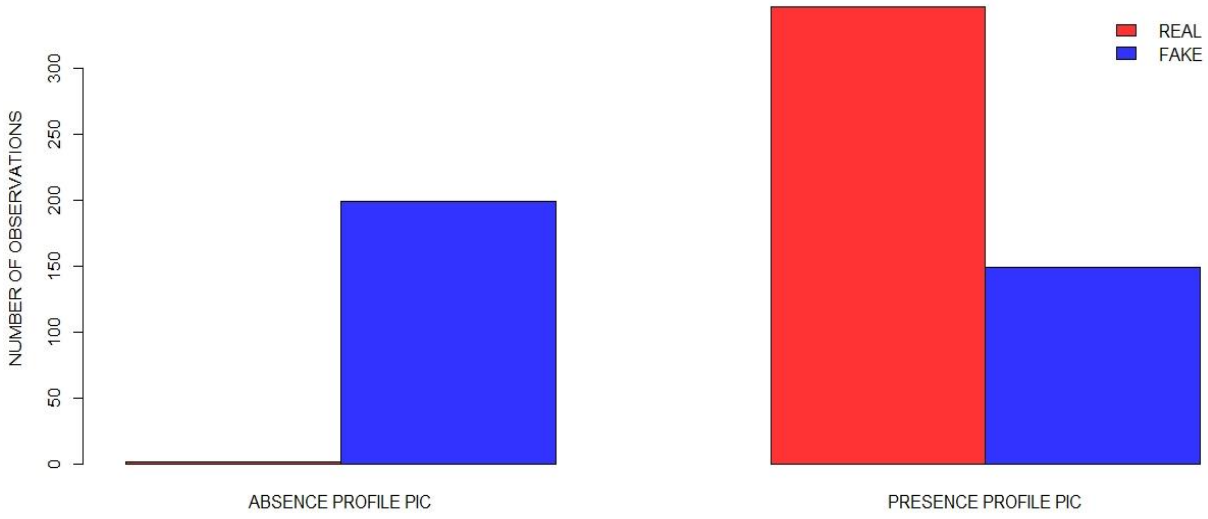


RELATIONSHIP BETWEEN VARIABLES

A seguire sono state analizzate le relazioni tra la variabile fake e le restanti variabili categoriche:

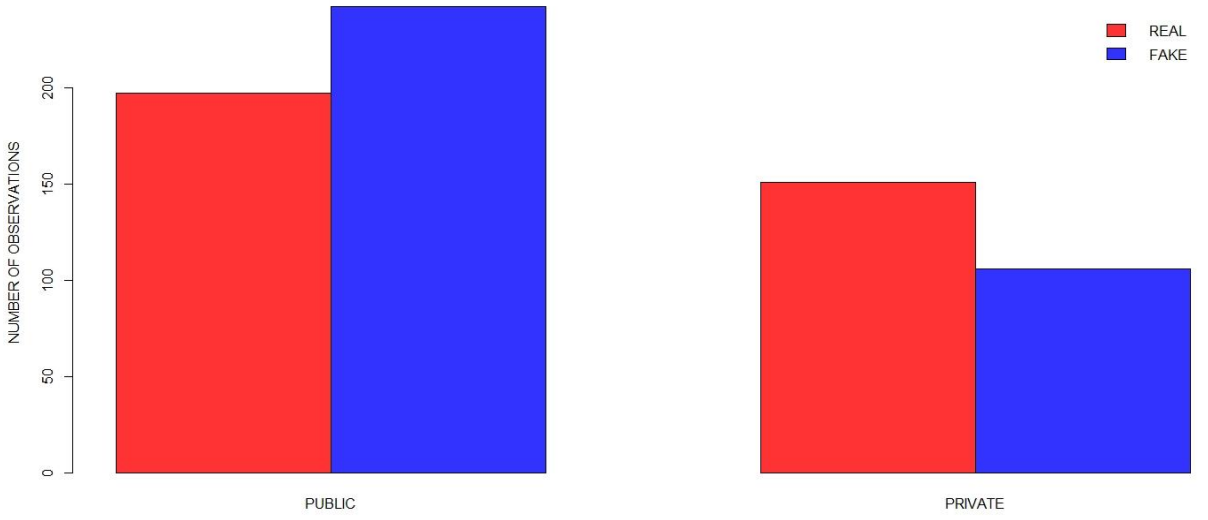
-**fake e profile.pic:** tra i profili autentici solo 2 (su 346) non hanno una foto profilo. Dunque, la maggior parte, se non quasi la totalità, dei profili veri hanno una foto profilo. Tra i profili fake, invece, sono più frequenti i profili che non possiedono una foto (199 su 346)

	NO PROFILE PIC	PROFILE PIC
NO FAKE	2	346
FAKE	199	149



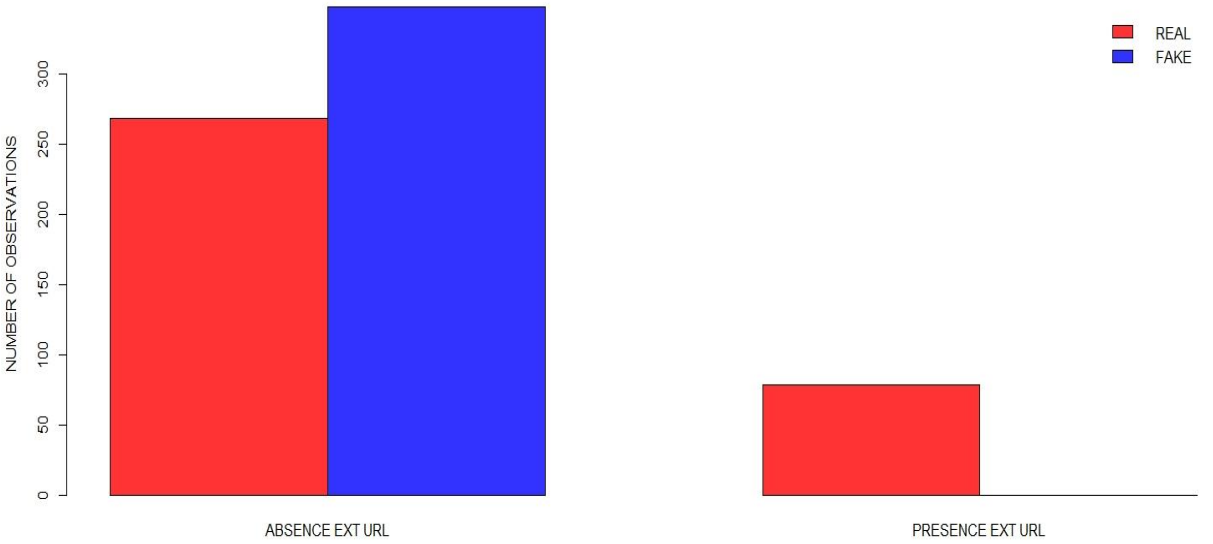
-**fake e private:** su 346 profili autentici 197 sono pubblici mentre i restanti 151 sono privati. Nel caso dei profili fake la maggior parte sono profili pubblici

	PUBBLICO	PRIVATO
NO FAKE	197	151
FAKE	242	106



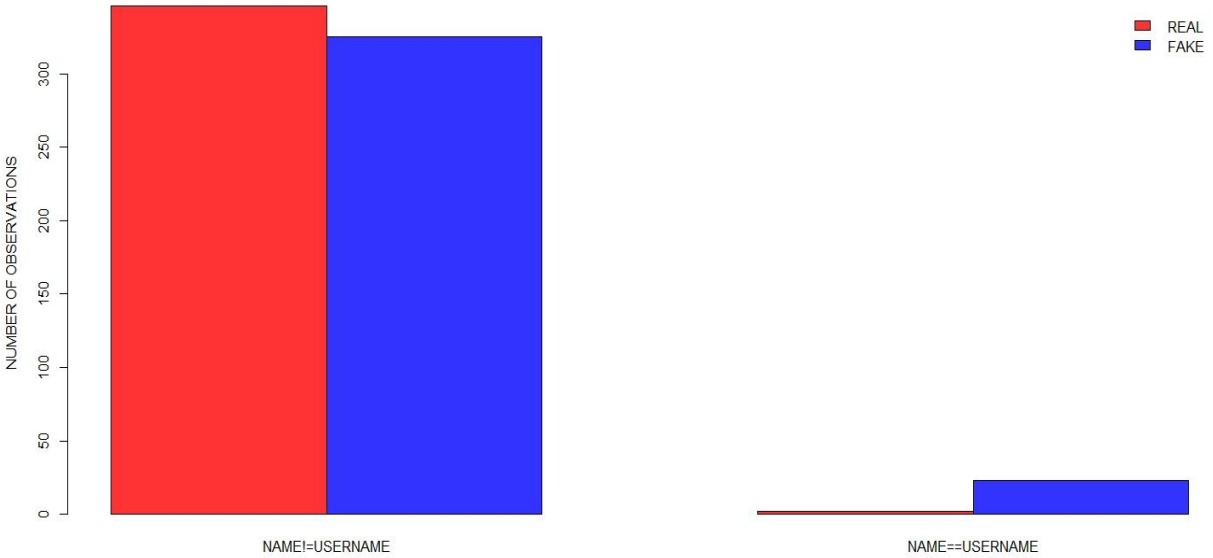
-**fake e external url:** nessun profilo fake possiede un link esterno, a differenza dei profili autentici

	NO EXT URL	EXT URL
NO FAKE	269	79
FAKE	348	0

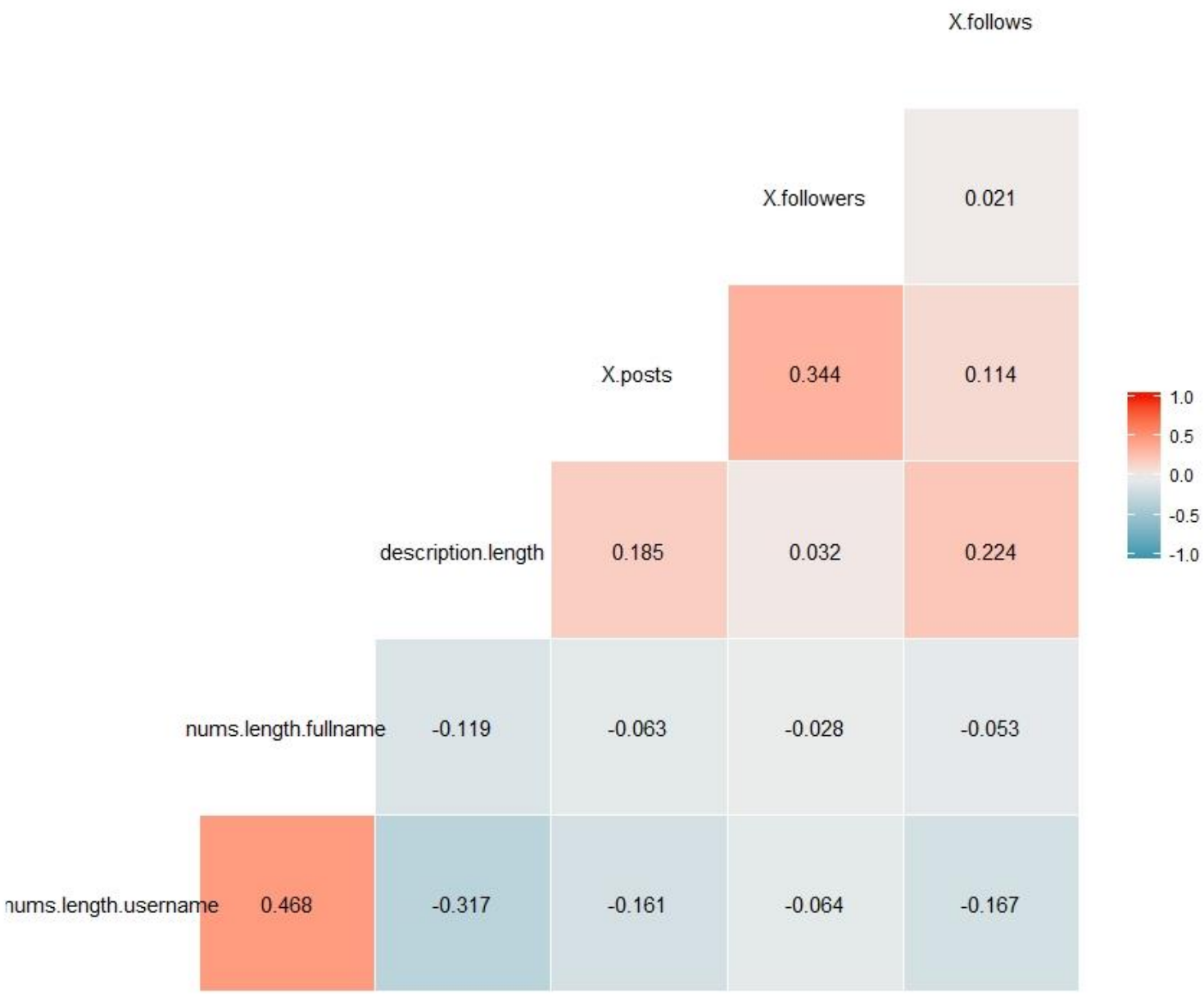


-**fake e name..username:** sia per profili fake sia per profili autentici è raro che lo username coincida con il fullname

	NAME!=USERNAME	NAME==USERNAME
NO FAKE	346	2
FAKE	325	23

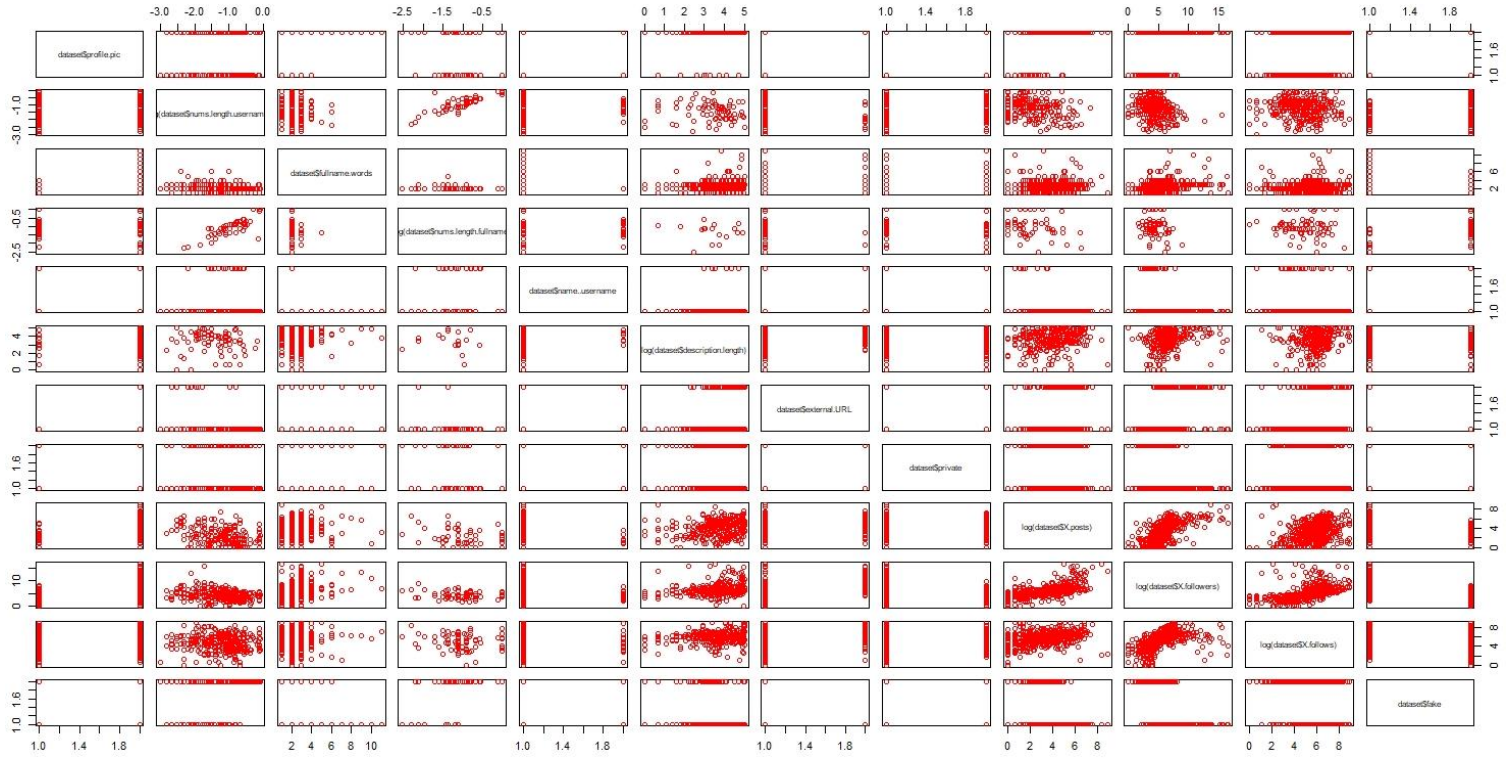


A seguire è stata analizzata la correlazione tra le variabili non categoriche.

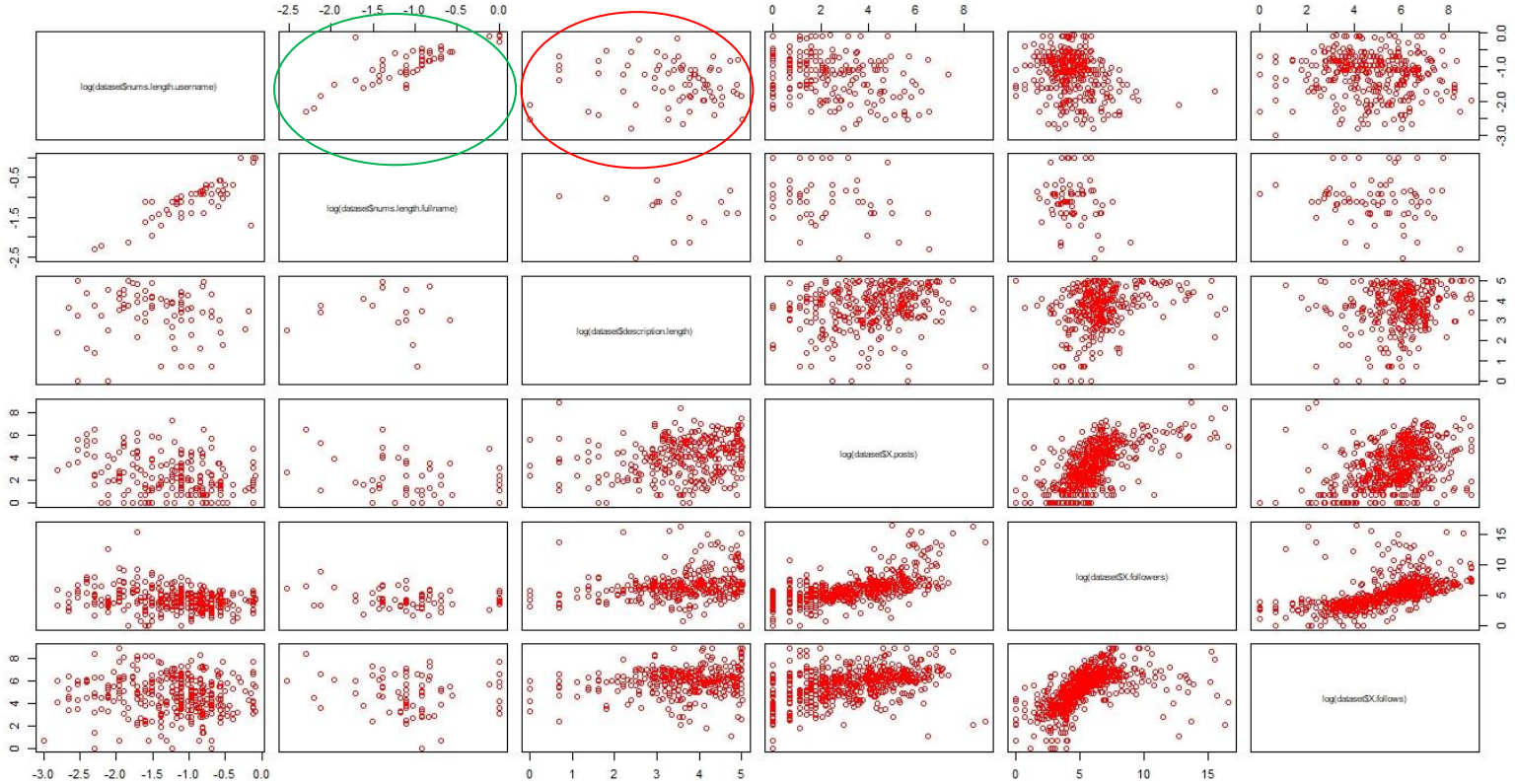


Come si può osservare le variabili maggiormente correlate sono nums.length.username e nums.length.fullname. Questo potrebbe essere spiegato dal fatto che un utente che inserisce diversi caratteri numerici nello username sarà più propenso a inserirli anche nel fullname, rispetto ad un utente che non usa caratteri numerici nello username o viceversa. Altre variabili correlate sono X.posts e X.followers: un utente molto attivo che pubblica dunque tanti posts genera più facilmente un seguito e perciò maggiori followers rispetto ad utente poco attivo. Le due variabili meno correlate risultano essere, invece, nums.length.username e description.length. Ciò potrebbe essere dovuto al fatto che si tratta di due porzioni totalmente distinte con scopi differenti del profilo di un utente. Come si può osservare non c'è nessuna coppia di variabili perfettamente correlate o perfettamente non correlate.

Successivamente è stato generato un grafico per studiare meglio l'andamento tra le variabili e confermare i risultati precedentemente ottenuti.



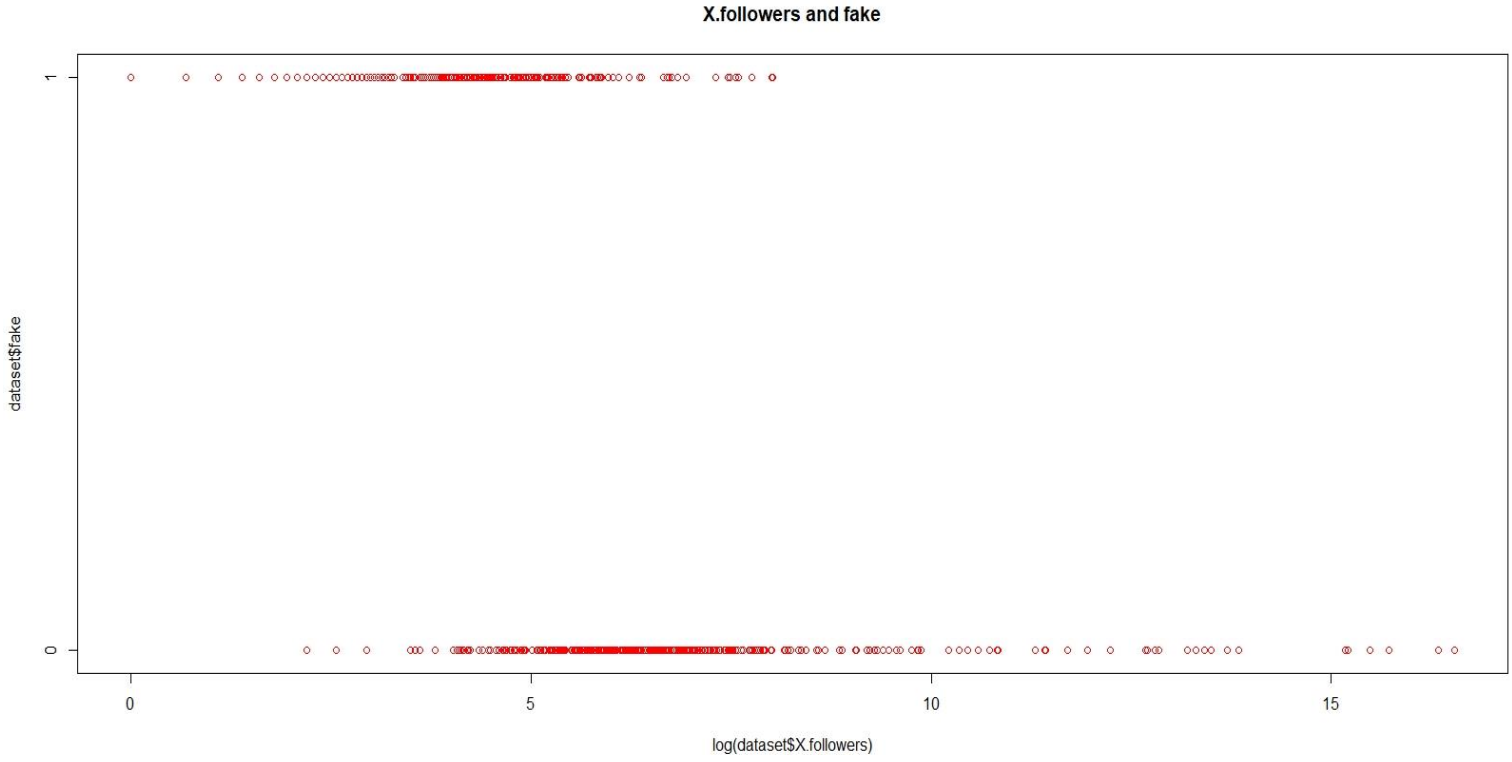
In questo secondo grafico sono state considerate le sole variabili non categoriche:

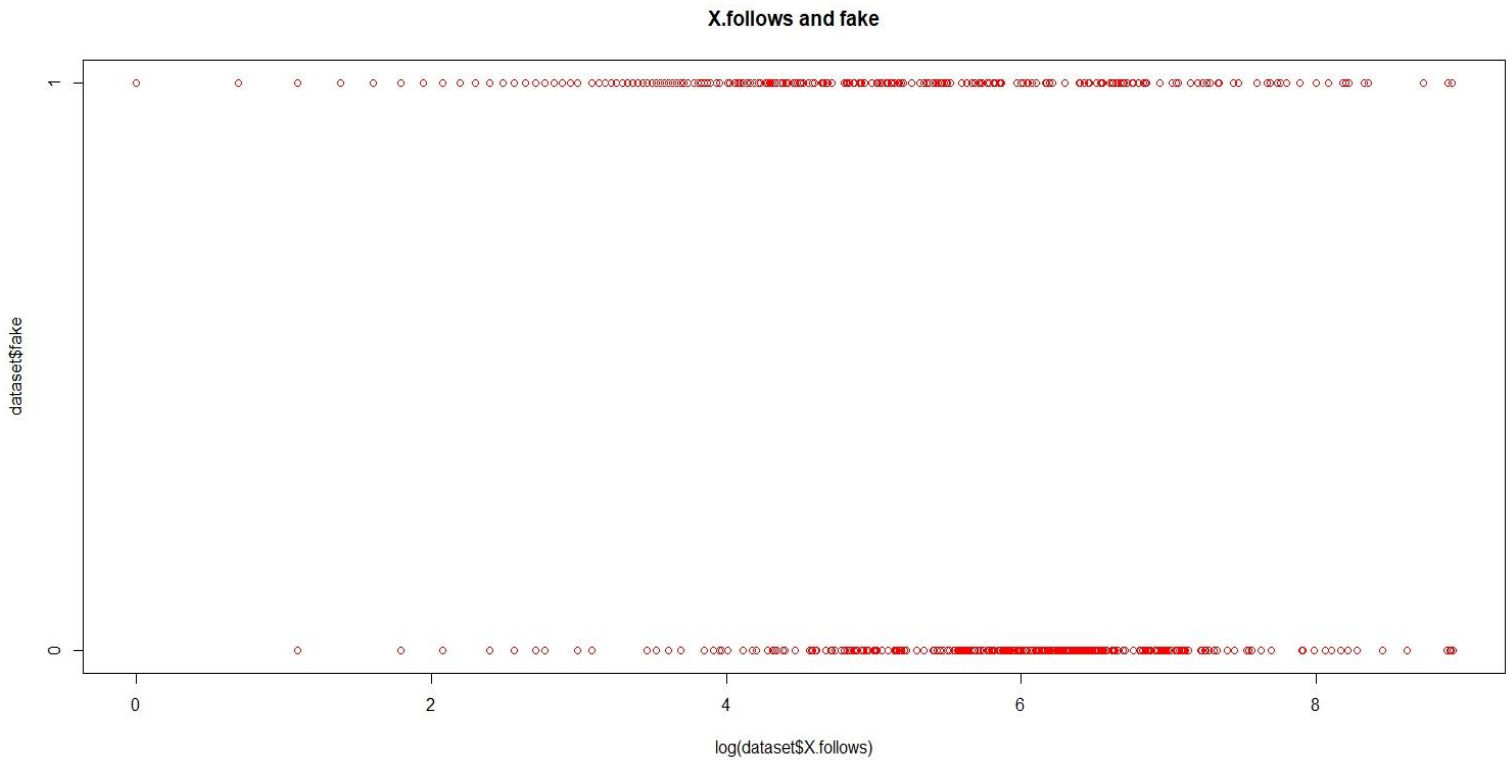


È stato applicato il logaritmo sulle variabili considerate per meglio evidenziare i possibili rapporti tra di esse. Anche da questo grafico è possibile notare la maggior correlazione tra

nums.length.username e nums.length.fullname e la minor correlazione tra nums.length.username e description.length

Successivamente si è analizzata la relazione tra X.followers e fake e X.follows e fake. È emerso che è maggiormente probabile trovare un profilo fake con un numero elevato di follows (profili seguiti dall'user fake) rispetto che trovare un profilo fake con un numero elevato di followers. La spiegazione relativa a ciò potrebbe essere data dal fatto che esistono diversi espedienti per i profili fake per seguire a loro volta in modo automatico profili (bot ecc.). Inoltre, risulterebbe piuttosto anomalo il fatto che un utente reale segua volontariamente un profilo fake.





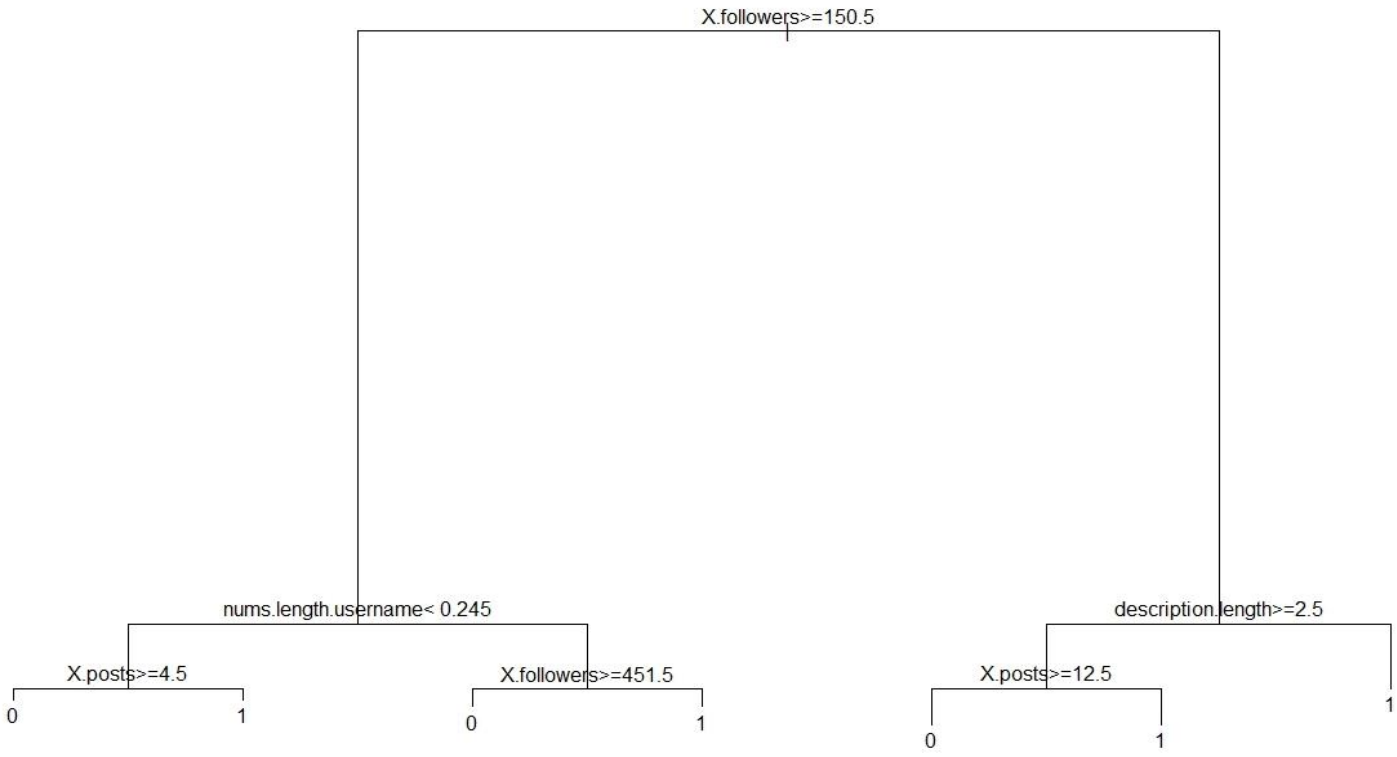
ANALYSIS

Per la parte finale relativa al machine learning è stata utilizzata come tecnica la classificazione: un profilo è fake o no?

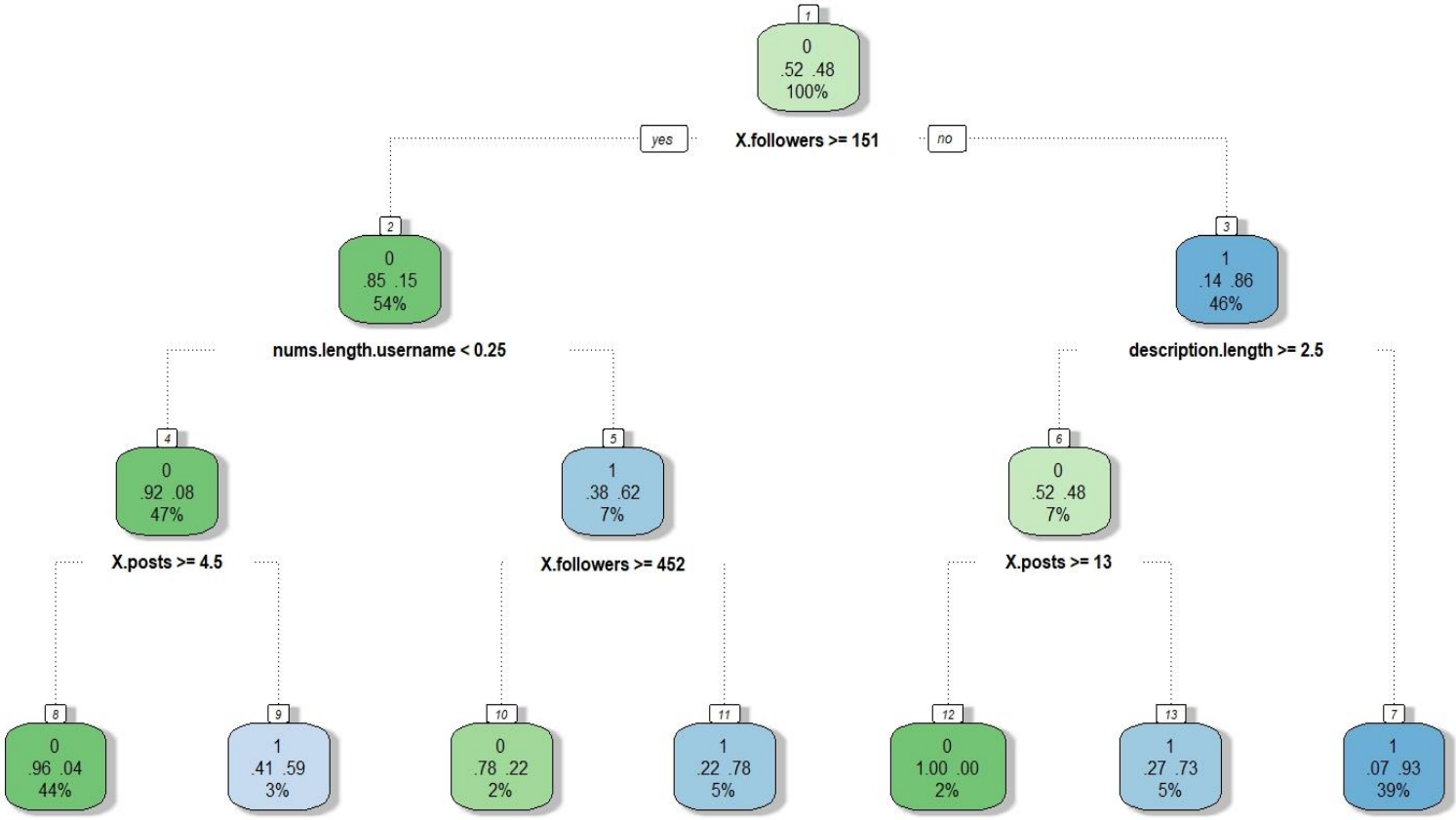
In particolare, come tecnica di classificazione è stata utilizzata inizialmente quella che fa uso degli alberi di decisione in quanto rappresentano un metodo supervisionato per la costruzione di un modello che ha come obiettivo la previsione di una variabile target in funzione di un insieme di variabili indipendenti. Nel contesto analizzato la variabile target risulta essere la variabile fake mentre le variabili indipendenti risultano essere le rimanenti, ovvero profile.pic, nums.length.username ecc. . In particolare, è stato utilizzato un albero di classificazione essendo la variabile target (fake) quantitativa.

Tramite opportune funzioni in R il dataset è stato diviso in training set (70% delle osservazioni) e testing set (30% delle osservazioni).

Sulla base del training set è stato creato un albero di decisione (di classificazione).



È stato poi realizzato un grafico equivalente più chiaro rispetto al precedente:

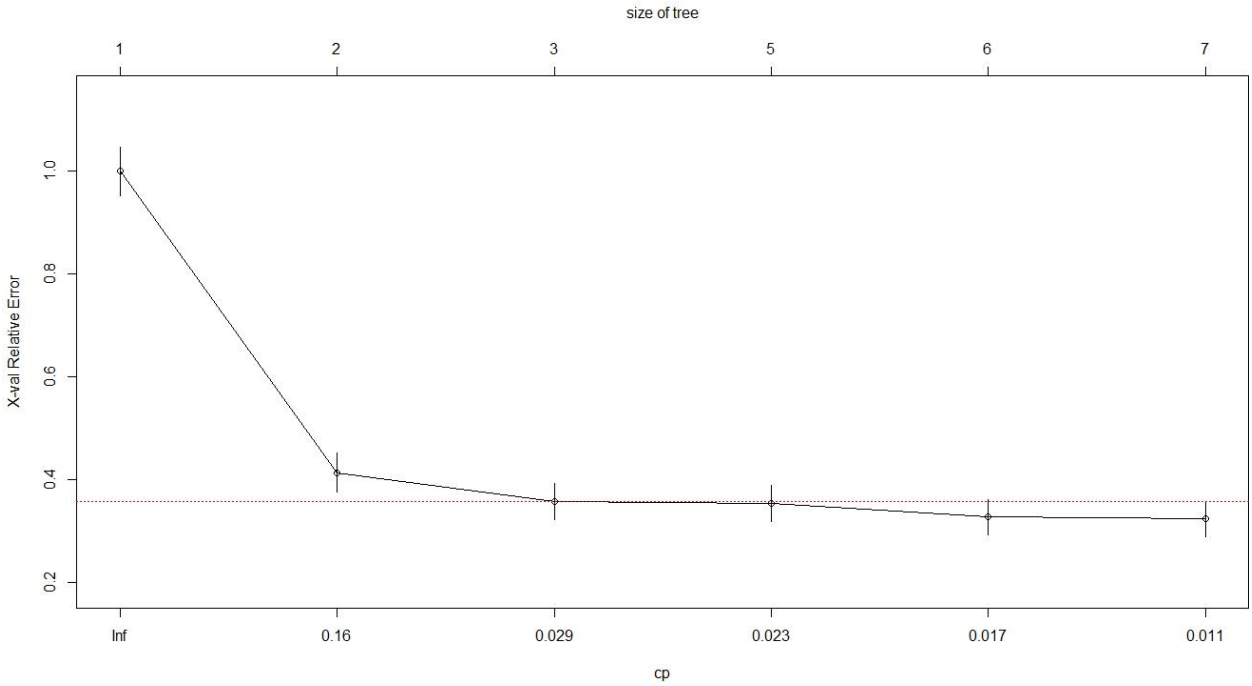


A seguire è stato validato il modello appena creato sul testing set e si è ottenuto un valore di accuratezza pari a: 0.8846. L'accuratezza rappresenta una delle metriche di valutazione del modello creato. Essa equivale al rapporto tra il numero di stime corrette e il numero totale di campioni di input e più si avvicina ad 1 migliore sarà il modello. Sono state calcolate anche altre metriche quali: precision (0.9157), recall (0.8172), f-measure (0.8636).

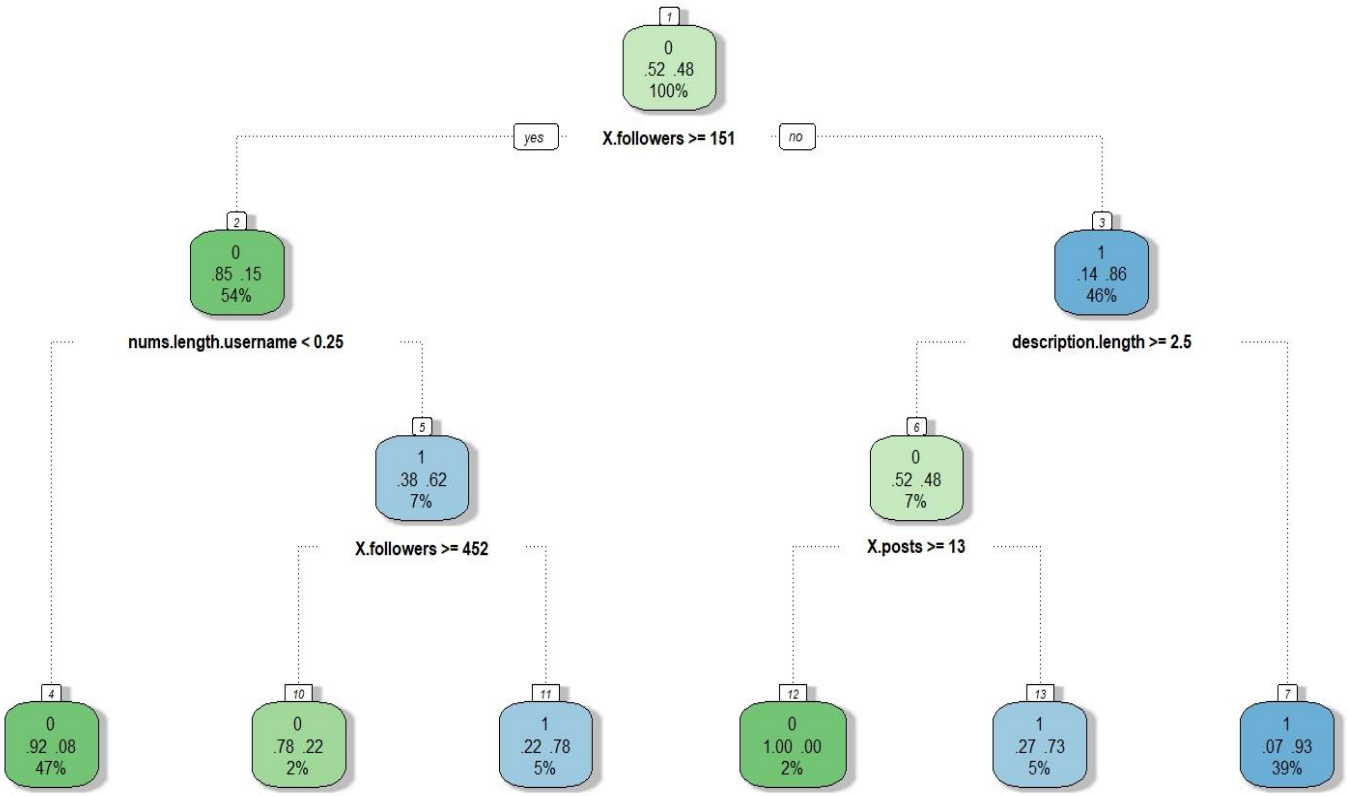
In particolare, su 93 profili reali ne sono stati predetti correttamente 76 reali e 17 sono stati classificati erroneamente fake. Tra 115 profili fake ne sono stati individuati correttamente 108 fake e 7 sono risultati in maniera errata profili reali. Di seguito viene presentata la matrice di confusione che riassume il tutto.

	Predicted real	**Predicted fake**
Actually real	76	17
Actually fake	7	108

Inoltre, per evitare problemi di overfitting si è passati alla potatura dell'albero cercando ottenere il valore di CP (complexity parameter) per cui fosse minimo l'errore di cross validation.



Facendo riferimento al grafico sopra è stato selezionato il valore di CP tale per cui è risultato minimo l'errore di cross validation, ovvero 0.017. Tale CP è stato poi considerato per la realizzazione del nuovo albero potato. L'obiettivo della potatura dell'albero è stato quello di evitare l'overfitting dei dati.



Oltre all'albero di classificazione è stata utilizzata un'altra tecnica di classificazione ovvero il K-nearest Neighbors Algorithm.

I dati che compongono il dataset sono stati in un primo momento normalizzati. A seguire sono stati creati il training set (70%) e il testing set (30%). Utilizzando la funzione "knn" con parametro k=22 (radice quadrata del numero delle osservazioni del training set) sono stati predetti i valori relativi alla variabile target fake.

Si è ottenuto come risultato finale un'accuratezza pari a: 0.8413. Come in precedenza sono stati calcolati anche i valori di altre metriche: precision (0.783), recall (0.8925), f-measure (0.8342).

In particolare, su 93 profili reali ne sono stati predetti correttamente 83 reali mentre su 115 profili effettivamente fake ne sono stati predetti correttamente 92 fake mentre 23 sono stati classificati in maniera errata come profili reali. Di seguito la matrice di confusione:

	Predicted real	**Predicted fake**
Actually real	83	10
Actually fake	23	92

CONCLUSIONS

L'analisi esplorativa del dataset ha avuto come obiettivo l'esplorazione delle distribuzioni delle variabili. Sono state messe in evidenza alcune delle loro principali statistiche descrittive come la media, la moda, la mediana e la varianza oltre che massimo e minimo. Per rendere più chiari i concetti sono stati realizzati dei grafici, tra i quali istogrammi, box plot e scatterplot. È stata analizzata anche la correlazione tra le variabili facenti parte del dataset andando ad evidenziare così delle relazioni tra di esse. A seguire nella parte di machine learning con l'obiettivo di predire se un profilo fosse fake o meno sono state utilizzate due diverse tecniche legate alla classificazione ovvero alberi di decisione e knn. I due approcci hanno portato a risultati di accuratezza accettabili, non troppo differenti. Il migliore tra i due metodi è stato l'utilizzo di alberi di classificazione, il quale ha presentato un valore di accuratezza maggiore. Infine, per visualizzare in maniera sintetica i risultati ottenuti durante l'esecuzione del progetto è stata realizzata un'applicazione web Shiny.