

# **UNIVERSITÀ DEGLI STUDI DI SALERNO**



**Progetto di Statistica e analisi dei dati**

**Francesco Maria Torino  
Matricola: 0522501879**

# Indice

<b>1. Introduzione.....</b>	<b>7</b>
1.1. Research Question 1 .....	7
1.2. Research Question 2 .....	7
1.3. Research Question 3 .....	7
1.4. Scelta del dataset.....	8
1.4.1. Variabili quantitative del dataset.....	9
1.4.2. Variabili qualitative del dataset .....	9
<b>2. Analisi dei dati.....</b>	<b>10</b>
<b>2.1. Analisi univariata .....</b>	<b>10</b>
2.1.1. Call Failures .....	10
2.1.2. Complains.....	17
2.1.3. Subscription Length.....	18
2.1.4. Charge Amount.....	23
2.1.5. Seconds of Use.....	29
2.1.6. Seconds of use intervals (Feature aggiunta) .....	34
2.1.7. Frequency of use .....	36
2.1.8. Frequency of SMS .....	42
2.1.9. Distinct Called Numbers .....	48
2.1.10. Age Group .....	55
2.1.11. Tariff Plan .....	61
2.1.12. Status .....	62
2.1.13. Age (Feature Rimossa).....	64
2.1.14. Churn.....	70
2.1.15. Customer Value .....	72
<b>2.2. Analisi Bivariata .....</b>	<b>78</b>
2.2.1. Customer Value VS. Frequency of SMS.....	78
2.2.2. Customer Value VS. Frequency of use.....	80
2.2.3. Customer Value VS. Frequency of SMS & Frequency of use .....	82
2.2.4. Frequency of use VS. Seconds of use .....	83
<b>3. Clustering .....</b>	<b>85</b>
3.1.1. Complains.....	91
3.1.2. Status .....	91
3.1.3. Seconds of use .....	92
<b>3.2. Clustering con rimozione degli outliers .....</b>	<b>93</b>
3.2.1. Rimozione degli outliers dalla variabile Seconds of use.....	93
3.2.2. Complains.....	94
3.2.3. Status .....	95
<b>3.3. Clustering con rimozione degli outliers ed utilizzo k-means++.....</b>	<b>96</b>
<b>3.4. Informazioni ottenute dal clustering .....</b>	<b>97</b>
<b>4. Studio proiezione popolazione di Churn su scala globale.....</b>	<b>98</b>

<b>5.</b>	<b>Studio popolazioni divise per le variabili utilizzate nel clustering .....</b>	<b>102</b>
5.1.1.	Confronto delle popolazioni sulla base della feature Complains .....	103
<b>6.</b>	<b>Generazione dei Dati Sintetici con LLM con scripting Few-shot con anomalie.....</b>	<b>111</b>
<b>6.1.</b>	<b>Analisi univariata Dataset Sintetico .....</b>	<b>113</b>
6.1.1.	Call Failures .....	113
6.1.2.	Complaints .....	117
6.1.3.	Subscription Length .....	119
6.1.4.	Charge Amount.....	123
6.1.5.	Second of use.....	127
6.1.6.	Seconds of use intervals .....	131
6.1.7.	Frequency of use .....	132
6.1.8.	Frequency of sms .....	135
6.1.9.	Distinct Call Numbers.....	138
6.1.10.	Age Group .....	141
6.1.11.	Tariff plan .....	145
6.1.12.	Status .....	147
6.1.13.	Churn.....	149
6.1.14.	Customer value .....	151
<b>7.</b>	<b>Analisi dei Risultati .....</b>	<b>154</b>
<b>7.1.</b>	<b>Confronto tra Dati Reali e Sintetici .....</b>	<b>154</b>
<b>8.</b>	<b>Conclusioni.....</b>	<b>155</b>
<b>8.1.</b>	<b>Research Question 1 .....</b>	<b>155</b>
<b>8.2.</b>	<b>Research Question 2 .....</b>	<b>155</b>
<b>8.3.</b>	<b>Research Question 3 .....</b>	<b>155</b>
<b>9.</b>	<b>Appendice.....</b>	<b>157</b>

---

# Indice delle figure

Figura 1: Boxplot Call Failures.....	11
Figura 2 Istogramma Call Failures.....	12
Figura 3 Funzione di distribuzione empirica (discreta) Call Failures.....	13
Figura 4 Diagramma a torta Call Failures.....	14
Figura 5 Diagramma di Pareto Call Failures .....	15
Figura 6 Distribuzione di frequenza Call Failures .....	16
Figura 7 Diagramma a torta e FDE Complains.....	17
Figura 8 Boxplot Subscription Length.....	18
Figura 9 Istogramma Subscription Length.....	19
Figura 10 FDE Subscription Length .....	20
Figura 11Diagramma a torta Subscription Length.....	20
Figura 12 Diagramma di Pareto Subscription Length .....	21
Figura 13 Distribuzione di frequenza Subscription Length.....	22
Figura 14 Boxplot Charge Amount.....	24
Figura 15 Istogramma Charge Amount .....	25
Figura 16 FDE Charge Amount.....	25
Figura 17 Diagramma a torta Charge Amount.....	26
Figura 18 Diagramma di Pareto Charge Amount .....	27
Figura 19 Distibuzione di frequenza Charge Amount .....	28
Figura 20 Boxplot Seconds of use .....	30
Figura 21 Istogramma Seconds of use .....	31
Figura 22 FDE Subscription Length .....	32
Figura 23 Distribuzione di frequenza Seconds of use.....	33
Figura 24 Distribuzione di frequenza Seconds of use Interval .....	34
Figura 25 Diagramma di Pareto Second of Use Interval .....	35
Figura 26 Boxplot Frequency of use.....	37
Figura 27 Istogramma Frequency of use.....	38
Figura 28 Funzione di distribuzione empirica (discreta) Frequency of use.....	39
Figura 29 Diagramma di Pareto Frequency of Use.....	40
Figura 30 Distribuzione di frequenza Frequency of use .....	41
Figura 31 Boxplot frequency of SMS .....	43
Figura 32 Istogramma Frequency of use.....	44
Figura 33 Funzione di distribuzione empirica (discreta) Frequency of SMS .....	45
Figura 34 Diagramma di Pareto Frequency of SMS .....	46
Figura 35 Distribuzione di frequenza Frequency of SMS .....	47
Figura 36 Boxplot Distinct Called Numbers.....	49
Figura 37 Istogramma Distinct Called Numbers .....	50
Figura 38 Funzione di distribuzione empirica (discreta) Distinct Called Numbers .....	51
Figura 39 Diagramma a torta Distinct Call Numbers .....	52
Figura 40 Diagramma di Pareto Distinct Called Numbers .....	53
Figura 41 Distribuzione di frequenza Distinct Called Numbers.....	54
Figura 42 Boxplot Age Group .....	55
Figura 43 Istogramma Age Group .....	56
Figura 44 FDE Age Group.....	57
Figura 45 Diagramma a torta Charge Amount.....	58
Figura 46 Diagramma di Pareto Age Group .....	59
Figura 47 Distibuzione di frequenza Age Group .....	60
Figura 48 Diagramma a torta e FDE Tariff plan.....	61

---

Figura 49 Diagramma a torta e FDE Status .....	63
Figura 50 Boxplot Age.....	64
Figura 51 Istogramma Age .....	65
Figura 52 Funzione di distribuzione empirica (discreta) Age.....	66
Figura 53 Diagramma a torta Age.....	67
Figura 54 Diagramma di Pareto Age .....	68
Figura 55 Distribuzione di frequenza Age.....	69
Figura 56 Diagramma a torta e FDE Churn.....	71
Figura 57 Boxplot Customer Value .....	73
Figura 58 Istogramma Customer Value.....	74
Figura 59 Funzione di distribuzione empirica (discreta) Customer Value .....	75
Figura 60 Diagramma di Pareto Customer Value.....	76
Figura 61 Distribuzione di frequenza Customer value .....	77
Figura 62 Correlazione Customer Value & Frequency of SMS .....	78
Figura 63 Correlazione Customer Value & Frequency of use.....	80
Figura 64 Customer Value in funzione di Frequency of use e Frequency of SMS .....	82
Figura 65 Correlazione Frquency of use & Seconds of use.....	83
Figura 66 Impatto variabili indipendenti su variabile Churn.....	88
Figura 67 Risultato Elbow method .....	89
Figura 68 Silhouette per calcolo k .....	90
Figura 69 Risultato del clustering su seconds of use senza outliers .....	<b>Errore. Il segnalibro non è definito.</b>
Figura 70 Rappresentazione normale della variabile di Churn.....	99
Figura 71 Rappresentazione normale standard .....	100
Figura 72 Popolazione di churn su dataset con lamentele .....	106
Figura 73 Popolazione di churn su dataset senza lamentele .....	108
Figura 74 Rappresentazione normale ottenuta con confronto popolazioni.....	110
Figura 75 Boxplot Call Failures Sintetico.....	113
Figura 76 Istogramma Call Failures Sintetico .....	114
Figura 77 Pie chart Call Failures Sintetico .....	115
Figura 78 FDE Call of failures Sintetico .....	116
Figura 79 Diagramma a torta e FDE Complains Sintetico .....	117
Figura 80 Boxplot Subscription Length Sintetico.....	119
Figura 81 Istogramma Subscription Length Sintetico .....	120
Figura 82 Pie chart Subscription Length Sintetico .....	121
Figura 83 FDE Subscription length Sintetico .....	122
Figura 84 Boxplot Charge Amount Sintetico .....	123
Figura 85 Istogramma Charge amount Sintetico .....	124
Figura 86 Pie chart Charge amount Sintetico .....	125
Figura 87 FDE Charge amount Sintetico .....	126
Figura 88 Boxplot Seconds of use Sintetico .....	127
Figura 89 Istogramma Seconds of use Sintetico .....	128
Figura 90 FDE Seconds of use Sintetico .....	130
Figura 91 Distribuzione di frequenza Seconds of use Interval .....	131
Figura 92Boxplot Frequency of use Sintetico.....	132
Figura 93 Istogramma Frequency of use Sintetico .....	133
Figura 94 FDE Frequency of use Sintetico .....	134
Figura 95 Boxplot Frequency of sms Sintetico.....	135
Figura 96 Istogramma Frequency of sms Sintetico .....	136
Figura 97 FDE Frequency of sms Sintetico .....	137
Figura 98 Boxplot Distinct Call Numbers Sintetico .....	138

Figura 99 Istogramma Distinct call numbers Sintetico.....	139
Figura 100 Distribuzione di frequenza Distinct call numbers Sintetico .....	140
Figura 101 Boxplot Age Group Sintetico .....	141
Figura 102 Istogramma Age Group Sintetico .....	142
Figura 103 Pie chart Age Group Sintetico .....	143
Figura 104 FDE Age Group Sintetico.....	144
Figura 105 Diagramma a torta e FDE Tariff plan Sintetico .....	145
Figura 106 Diagramma a torta e FDE Status Sintetico .....	147
Figura 107 Diagramma a torta e FDE Churn Sintetico.....	149
Figura 108 Boxplot Customer Value Sintetico .....	151
Figura 109 Istogramma Customer Value Sintetico.....	152
Figura 110 FDE Customer Value Sintetico .....	153
Figura 111 Pairs dataset sintetico.....	154

# 1. Introduzione

Oggiorno si parla sempre di più di scarsità di dati raccolti per effettuare degli studi su larga scala. Questi dati sono utili per effettuare analisi statistiche e per osservare dei fenomeni.

Per via del GDPR in Italia la tutela della privacy è uno dei pilastri dei diritti inalienabili e ciò porta ad avere una bassa disponibilità di dataset consistenti, di fatti, la tutela della privacy impone che i dataset seppur anonimizzati o pseudo-anonimizzati possono violare la privacy dei cittadini. Ultimamente, infatti, si parla sempre più spesso di dataset sintetici. La crescente disponibilità di modelli di intelligenza artificiale, in particolare i Large Language Model (LLM), ha aperto nuove strade per la generazione di dati sintetici. Questo progetto si pone l'obiettivo di confrontare le proprietà statistiche di dati reali e dati sintetici generati da LLM, valutando se i dati sintetici possono essere utilizzati come alternativa o integrazione ai dati reali in contesti di analisi applicate.

La generazione di dati sintetici è particolarmente utile in situazioni in cui i dati reali sono scarsi, sensibili, o soggetti a restrizioni legali. Tuttavia, è essenziale garantire che i dati sintetici mantengano le stesse proprietà statistiche dei dati reali e che siano affidabili per scopi pratici come la modellazione o l'addestramento di sistemi di machine learning. Questo studio esplora due Research Question (RQ) fondamentali riguardanti i dati sintetici:

## 1.1. Research Question 1

“I dati sintetici generati dai Large Language Model mantengono le stesse proprietà statistiche dei dati reali?”

Per rispondere a questa research question andremo a creare un dataset sintetico con l'ausilio di ChatGPT 4-0 il quale verrà poi confrontato con il dataset di partenza per verificare se i dataset sintetici mantengono o meno le proprietà del dataset di partenza.

[\[Conclusione Research question 1\]](#)

## 1.2. Research Question 2

Perché gli utenti abbandonano il servizio del servizio descritto dal dataset analizzato?

Per rispondere a questa research question andremo a studiare nel dettaglio quali sono le features che più impattano l'abbandono degli utenti abbonati al servizio comprendendo come sarebbe possibile abbassare quello che è l'abbandono generale del servizio.

[\[Conclusione Research question 2\]](#)

## 1.3. Research Question 3

Con i dati forniti dal dataset Iranian Churn e **progettando il servizio su scala mondiale**, quale sarebbe il tasso di abbandono degli utenti entro 9 mesi?

In altre parole, sarebbe conveniente investire in tecnologia per espandere il provider a livello globale e ottenere un profitto?

[\[Conclusione Research question 3\]](#)

## 1.4. Scelta del dataset

Il dataset scelto per questo studio è il dataset “Iranian Churn”.

Il **Churn rate**, o **abbandono dei clienti**, è una metrica importante che le aziende devono monitorare quando cercano di espandere il proprio business.

Questa metrica rappresenta il numero di clienti che hanno smesso di utilizzare il prodotto o il servizio durante un determinato periodo di tempo.

In ultima analisi, il Churn rate di un’azienda identificherà il tasso complessivo di retention dei clienti.

Il Churn rate è inversamente correlato al tasso di retention dei clienti. Un’azienda con un alto Churn rate avrà un basso tasso di retention, il che significa che non riesce a mantenere i clienti. Al contrario, un basso Churn rate implica un alto tasso di retention, suggerendo che l’azienda è efficace nel mantenere i suoi clienti nel tempo.

Monitorare il Churn rate aiuta le aziende a identificare aree di miglioramento e a sviluppare strategie per aumentare la soddisfazione e la fidelizzazione dei clienti.

Questo dataset è stato raccolto in modo casuale dal database di una compagnia telefonica iraniana nel corso di 12 mesi.

Contiene un totale di 3150 righe di dati, ciascuna rappresentante un cliente, e presenta informazioni su 14 colonne. Le variabili presenti in questo dataset includono:

- **Call Failures (Fallimenti di Chiamata)**: numero di fallimenti di chiamata del fruitore;
- **Complains (Lamentela)**: Lamentele riportate dal fruitore;
- **Subscription Length (Durata della sottoscrizione)**: totale mesi di fruizione del servizio;
- **Charge Amount (Importo addebitato)**: fascia di costo del servizio mensile;
- **Seconds of Use (Secondi di Utilizzo)**: totale secondi di chiamate effettuate dagli utenti;
- **Frequency of use (Frequenza di Utilizzo)**: numero totale di chiamate da parte del fruitore del servizio (**questa è un’assunzione** dato che dalla documentazione non è chiaro. È stata data una spiegazione con [l’analisi bivariata](#));
- **Frequency of SMS (Frequenza di SMS)**: numero totale di messaggi di testo da parte del fruitore del servizio;
- **Distinct Called Numbers (Numeri Chiamati Distinti)**: numero totale di chiamate distinte da parte del fruitore del servizio;
- **Age Group (Gruppo di Età)**: gruppo d’età a cui appartiene il fruitore del servizio;
- **Tariff Plan (Piano Tariffario)**: piano tariffario del servizio specifico dell’utente;
- **Status (Stato)**: stato dell’attivazione del servizio;
- **Age (Età)**: Età del fruitore del servizio;
- **Churn (Abbandono)**: Abbandono del servizio da parte dell’utente;
- **Customer Value (Valore del Cliente)**: il valore calcolato del cliente considerando le feature associategli (**questa è un’assunzione** dato che non è chiaro nella documentazione esposta dai creatori del dataset come questa feature sia stata calcolata tramite [l’analisi bivariata](#) è stata data una spiegazione);

Tutte le variabili, eccetto l’attributo **churn** (abbandono), sono dati aggregati dei primi 9 mesi. Le etichette di **churn** indicano lo stato dei clienti alla fine dei 12 mesi. I tre mesi rappresentano un intervallo di pianificazione designato.

Dalla documentazione prodotta dai creatori del dataset, questo dataset **non presenta valori mancanti**.

Possiamo poi dividere le colonne del dataset analizzato in variabili quantitative e qualitative.

### 1.4.1. Variabili quantitative del dataset

Le variabili quantitative di questo dataset sono:

- **Call Failures**: variabile numerica intera;
- **Subscription Length**: variabile numerica intera;
- **Charge Amount**: variabile ordinale (quantitativa discreta) (0: importo più basso, 10: importo più alto);
- **Seconds of Use**: variabile numerica intera;
- **Frequency of use**: variabile numerica intera;
- **Frequency of SMS**: variabile numerica intera;
- **Distinct Called Numbers**: variabile numerica intera;
- **Age**: variabile numerica intera;
- **Age Group**: quantitativa discreta, rappresenta gruppi di età (1: età più giovane, 5: età più anziana);
- **Customer Value**: variabile numerica intera;

### 1.4.2. Variabili qualitative del dataset

Le variabili qualitative di questo dataset vengono espresse tramite dati binari e sono le seguenti:

- **Complains**: variabile binaria (0: Nessuna lamentela, 1: lamentela);
- **Tariff Plan**: variabile binaria (1: Pay to go, 2: Pagamento contrattuale);
- **Status**: variabile binaria (1: Attivo, 2: Non attivo);
- **Churn**: variabile binaria (0: Non abbandonato il servizio, 1: Abbandonato il servizio);

## 2. Analisi dei dati

In questo capitolo andremo ad effettuare una analisi generale sui dati che compongono il dataset.

Nello specifico verranno analizzate tutte le features, una per una per averne una migliore comprensione dopodiché verrà effettuata un'analisi bivariata per comprendere come le features sono correlate tra loro.

### 2.1. Analisi univariata

Il dataset reale è stato sottoposto a un processo di pre-elaborazione, che include l'analisi e la creazione di visualizzazioni grafiche per un'analisi esplorativa preliminare.

Partendo dal presupposto che i creatori del dataset abbiano dichiarato l'assenza di valori mancanti, non saranno necessarie operazioni di gestione in tal senso.

Procederemo inizialmente con un'analisi dettagliata di ciascun campo.

#### 2.1.1. Call Failures

La feature “[Call Failures](#)” è una variabile quantitativa discreta espressa in numeri interi, rappresentante il numero di fallimenti di chiamata per ogni fruitore del servizio. Per una completa caratterizzazione statistica della variabile, si procederà con un'analisi delle sue misure di centralità e dispersione, seguita da un'analisi grafica.

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Call Failures” risulta pari a 7,63.
- **Mediana campionaria [2]**: La mediana è pari a 6.
- **Moda campionaria [3]**: La moda invece risulta essere 0.

La predominanza della moda pari a 0 indica una distribuzione unimodale, con un picco concentrato sul valore zero. Dalle misure di media e mediana possiamo desumere che la distribuzione sia asimmetrica positiva (sbilanciata a destra).

In altre parole:

- **Asimmetria verso destra**: La distribuzione presenta una "coda" estesa a destra a causa della presenza di valori più elevati di Call Failures.
- **Concentrazione attorno allo zero**: La maggior parte degli utenti riporta pochi fallimenti di chiamata, con valori di media spostati a destra rispetto alla mediana e alla moda.

Un boxplot della variabile *Call Failures* permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.

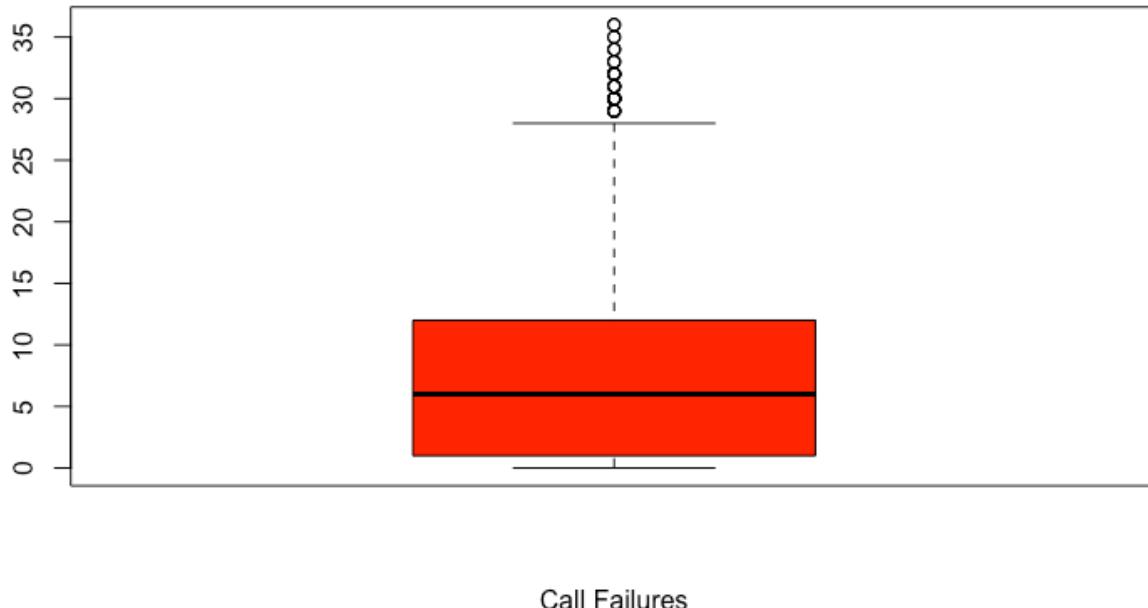


Figura 1: Boxplot Call Failures

Possiamo notare dall'immagine che abbiamo molteplici outliers.

Utilizzando lo scarto interquartile, abbiamo rilevato i seguenti outliers: **29, 30, 31, 32, 33, 34, 35, 36.**

Tramite poi una funzione apposita confermiamo che il **primo quartile** è **1.000** mentre il **terzo quartile** è **12.00**.

Inoltre, abbiamo il **minimo** uguale a **0.00** ed un **massimo** uguale a **36.00**.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] delle chiamate fallite dei fruitori.

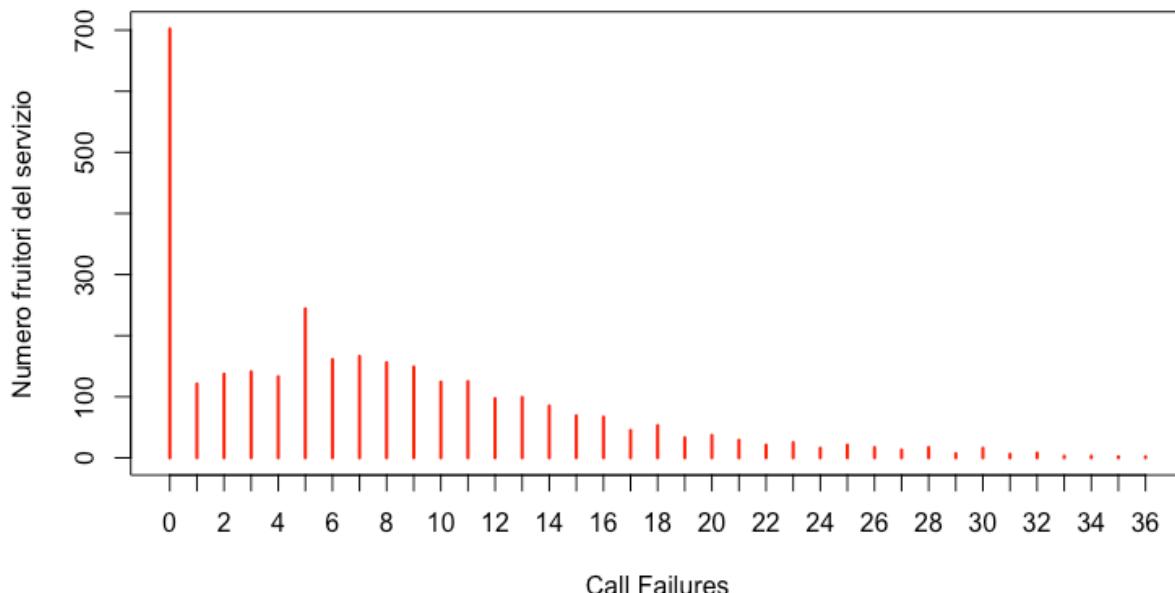


Figura 2 Istogramma Call Failures

Un istogramma della variabile *Call Failures* mostra la frequenza assoluta [9] dei fallimenti di chiamata per ciascun valore osservato. Le ascisse rappresentano il numero di fallimenti di chiamata, mentre le ordinate indicano la quantità di utenti corrispondenti.

Il grafico conferma una distribuzione asimmetrica, con una concentrazione di osservazioni attorno a valori bassi e una coda verso destra.

Un'analisi delle frequenze relative [10] tramite **Funzione di Distribuzione Empirica (discreta)** evidenzia ulteriormente come una larga porzione degli utenti presenti valori prossimi allo zero.

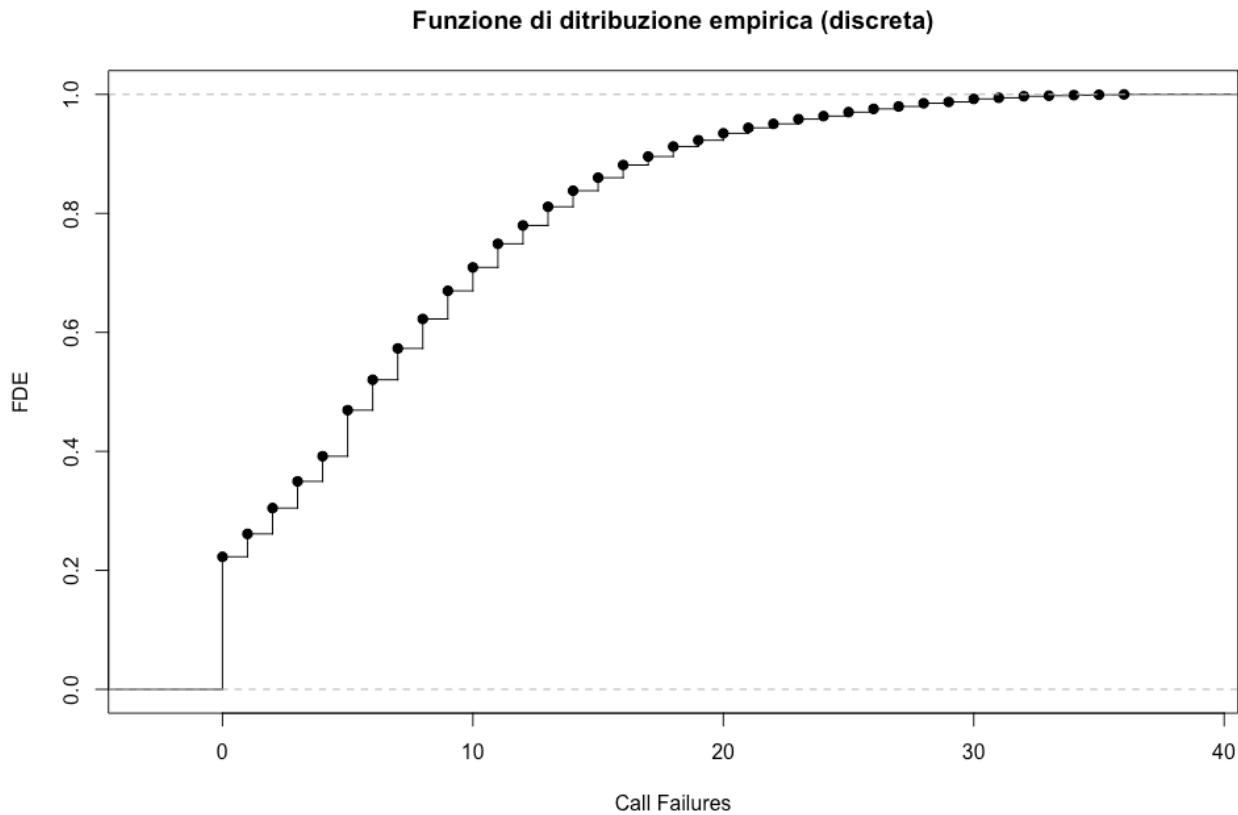


Figura 3 Funzione di distribuzione empirica (discreta) Call Failures

L'analisi della **Funzione di Distribuzione Empirica (FDE)** conferma ulteriormente l'asimmetria menzionata precedentemente: mostra infatti una rapida crescita iniziale (data dalla frequenza elevata di valori prossimi allo zero), seguita da un incremento più graduale in corrispondenza dei valori più elevati.

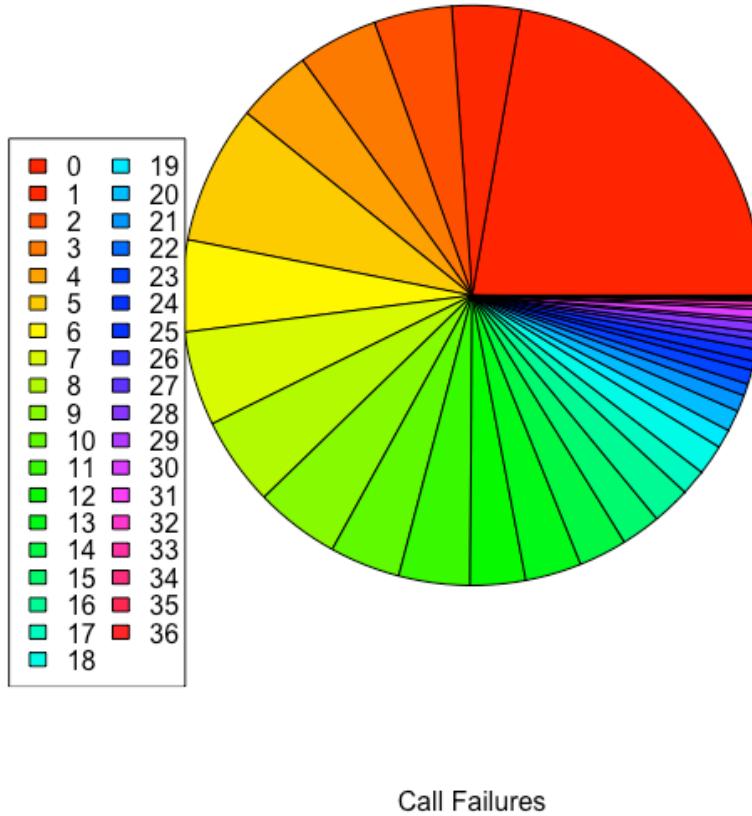


Figura 4 Diagramma a torta Call Failures

Il **diagramma a torta** illustra che la maggior parte degli utenti non ha registrato fallimenti di chiamata.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza [4]: 52.76**
- **Deviazione standard [5]: 7.26**
- **Coefficiente di variazione [6]: 95.22%**

L'elevato coefficiente di variazione indica una **forte dispersione dei valori** rispetto alla media, segnalando una significativa variabilità nel numero di fallimenti di chiamata tra gli utenti.

L'analisi tramite diagramma di Pareto permette di visualizzare come le frequenze assolute [9] siano associate alla frequenza relativa cumulativa, sottolineando la predominanza di utenti con pochi fallimenti di chiamata e il peso cumulativo degli utenti con più fallimenti.

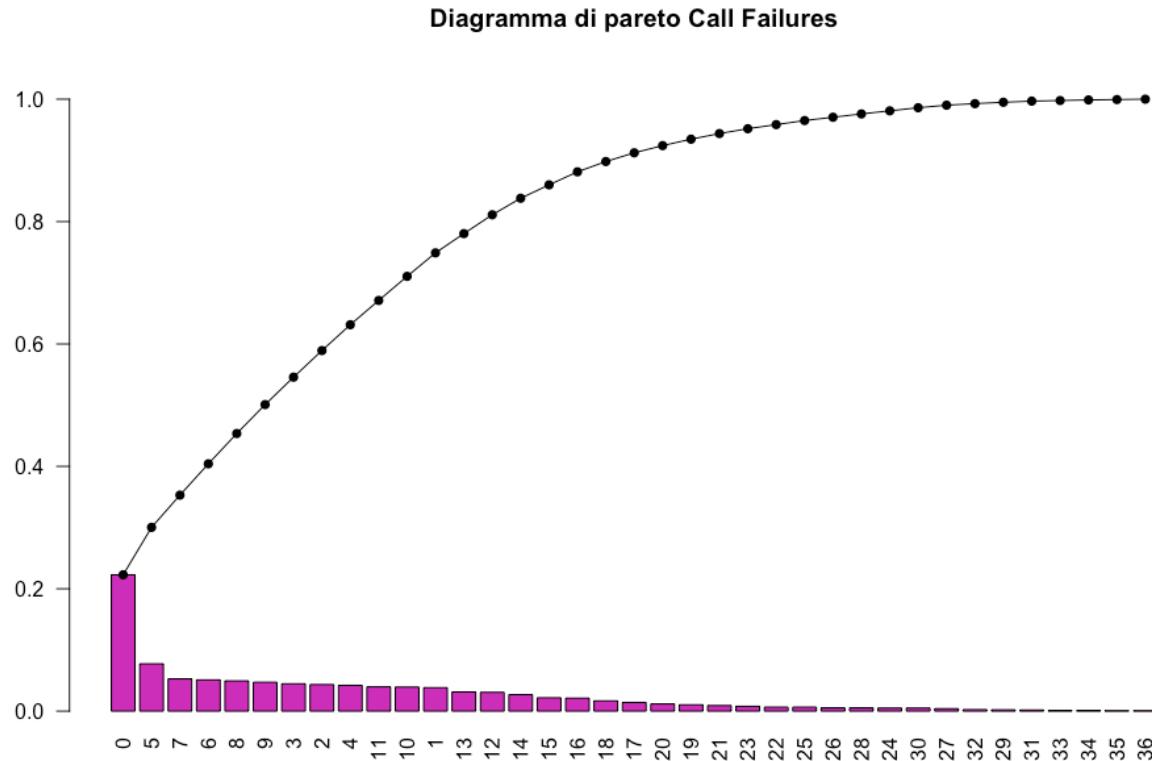


Figura 5 Diagramma di Pareto Call Failures

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 1.09**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 3.90**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza dei fallimenti di chiamata, confermando le caratteristiche sopra descritte.

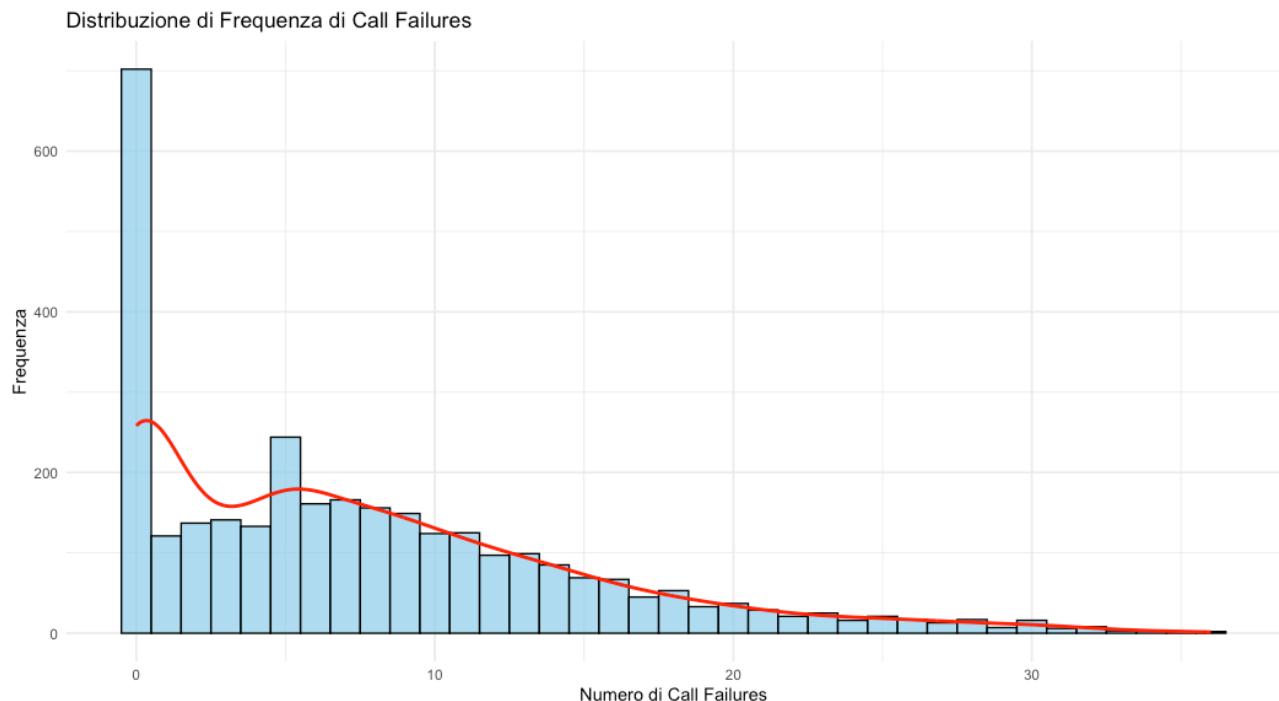


Figura 6 Distribuzione di frequenza Call Failures

## 2.1.2. Complains

La feature “[Complains](#)” rappresenta una variabile binaria che indica la presenza o assenza di una lamentela registrata dal fruitore del servizio (**0: Nessuna lamentela, 1: Lamentela**).

Data la natura qualitativa della variabile, l’analisi procederà tramite lo studio delle frequenze e delle distribuzioni.

Analizziamo quindi le **frequenze assolute** [9] dei valori assunti dalla variabile Complains:

Valore	Frequenza
0: Nessuna lamentela	<b>2909</b>
1: Lamentela	<b>241</b>

Andiamo inoltre a vedere le **frequenze relative** [10]:

Valore	Frequenza
0: Nessuna lamentela	0.92
1: Lamentela	0.08

Possiamo quindi notare che il **92.34%** dei fruitori non ha espresso alcuna lamentela riguardante il servizio.

Mentre il restante **7,66%** ha espresso una lamentela.

Per avere un’idea più chiara possiamo osservare il diagramma a torta e il diagramma rappresentante la funzione di distribuzione empirica (discreta) sottostanti:

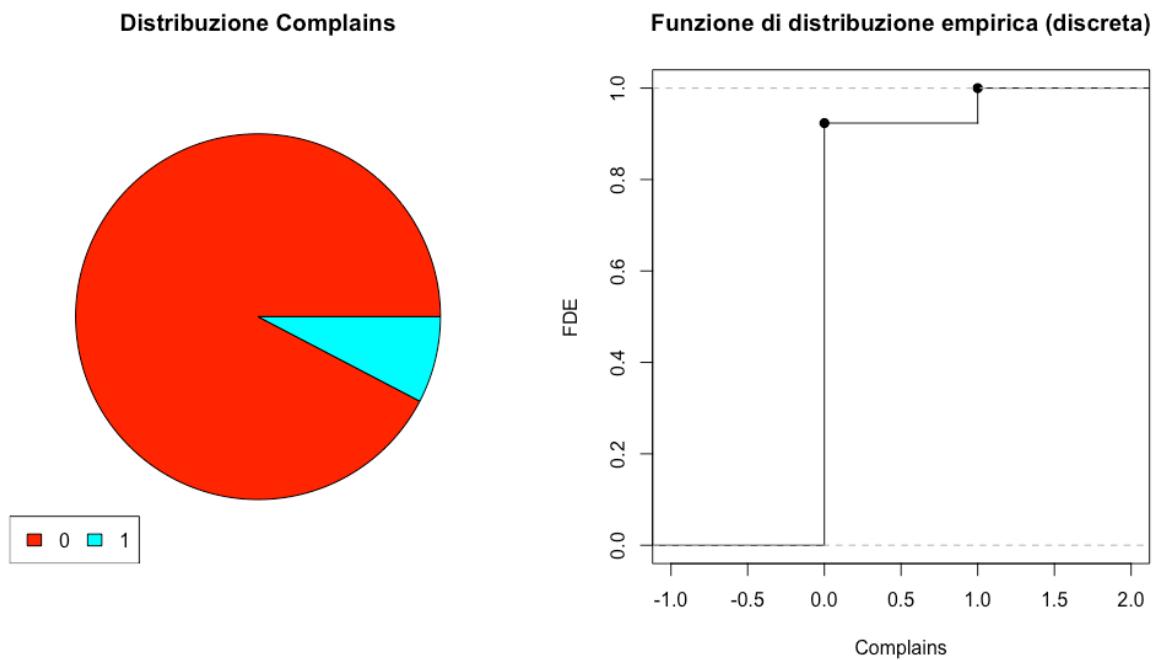


Figura 7 Diagramma a torta e FDE Complains

### 2.1.3. Subscription Length

La feature “[Subscription Length](#)” è una variabile quantitativa discreta espressa in numeri interi, rappresentante il numero di mesi di fruizione del servizio. Per una completa caratterizzazione statistica della variabile, si procederà con un’analisi delle sue misure di centralità e dispersione, seguita da un’analisi grafica.

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Subscription Length” risulta pari a **32.54**.
- **Mediana campionaria [2]**: La mediana è pari a **35**.
- **Moda campionaria [3]**: La moda invece risulta essere **36**.

La prevalenza della moda a 36 suggerisce una distribuzione unimodale, con un picco concentrato intorno a questo valore. La relazione tra media, mediana e moda indica una distribuzione asimmetrica negativa (sbilanciata a sinistra). Nello specifico:

- **Asimmetria verso sinistra**: La distribuzione è caratterizzata da una coda a sinistra, che rappresenta la presenza di valori bassi di Subscription Length.
- **Moda**: Essendo il valore più alto tra le tre misure di centralità, conferma la concentrazione dei dati su valori elevati di Subscription Length.

Un boxplot della variabile Subscription length permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.

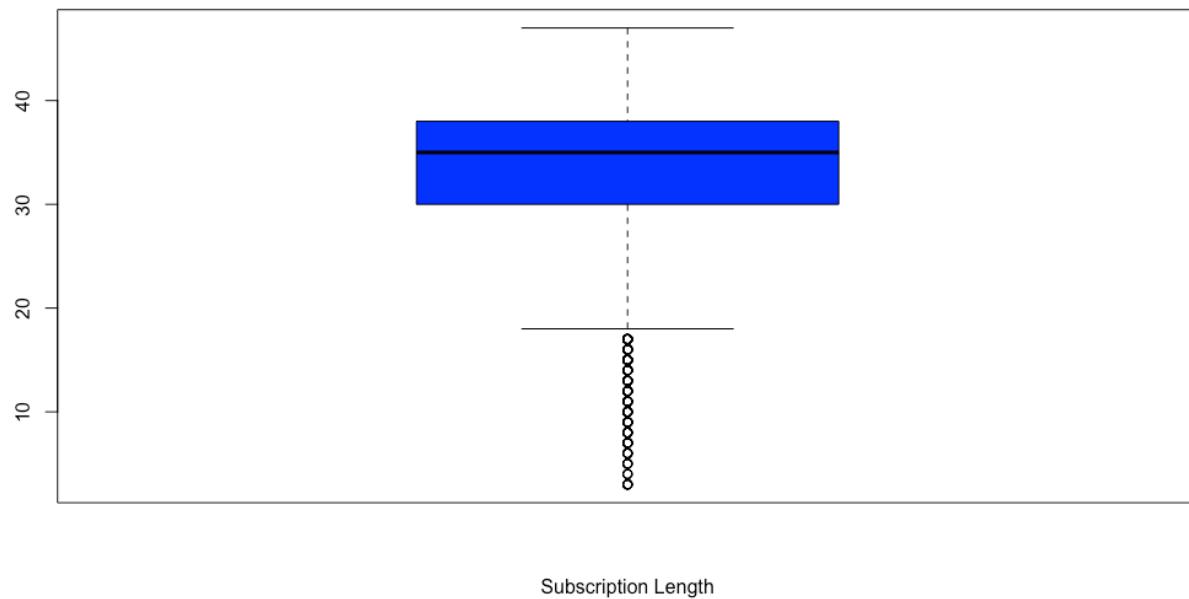


Figura 8 Boxplot Subscription Length

Possiamo notare dall’immagine che abbiamo molteplici outliers inferiori alla mediana. Utilizzando lo scarto interquartile, abbiamo rilevato i seguenti outliers:

**3,4,5,6,7,8,9,10,11,12,13,14,15,16,17**.

Tramite poi una funzione apposita confermiamo che il **primo quartile** è **30.00** mentre il **terzo quartile** è **38.00**.

Inoltre, abbiamo il **minimo** uguale a **3.00** ed un **massimo** uguale a **47.00**.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] del dato in questione.

L'istogramma della variabile Subscription Length fornisce una visualizzazione delle frequenze assolute [9] dei mesi di fruizione tra gli utenti. Le ascisse rappresentano il numero di mesi, mentre le ordinate mostrano il numero di utenti corrispondenti.

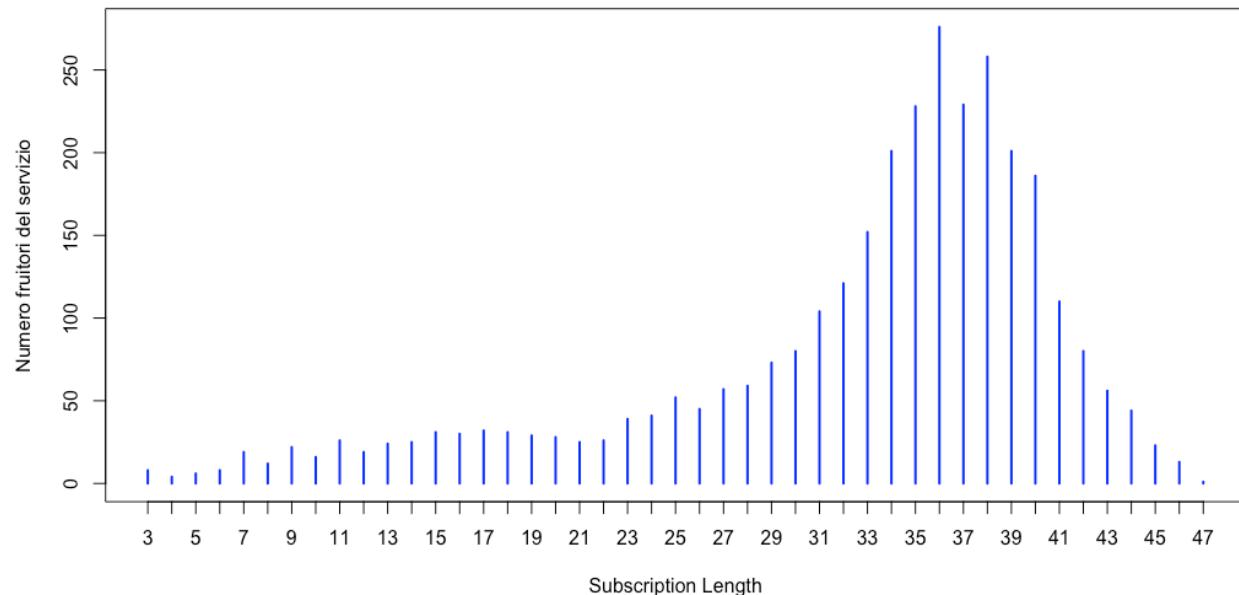
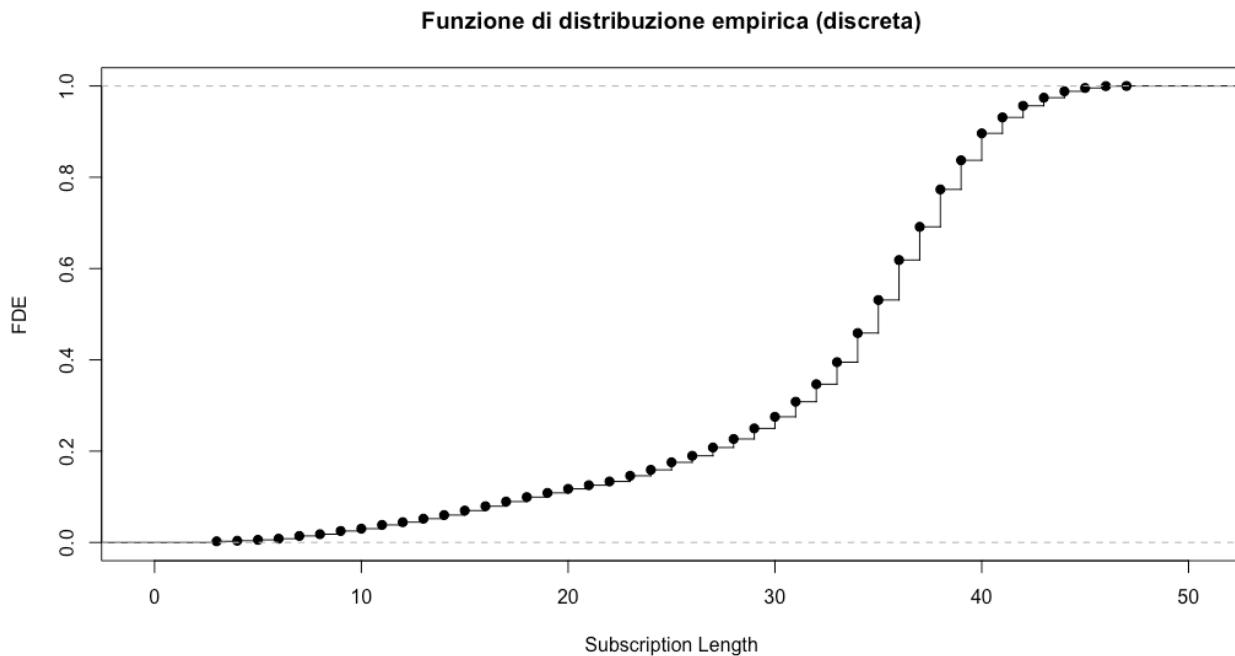


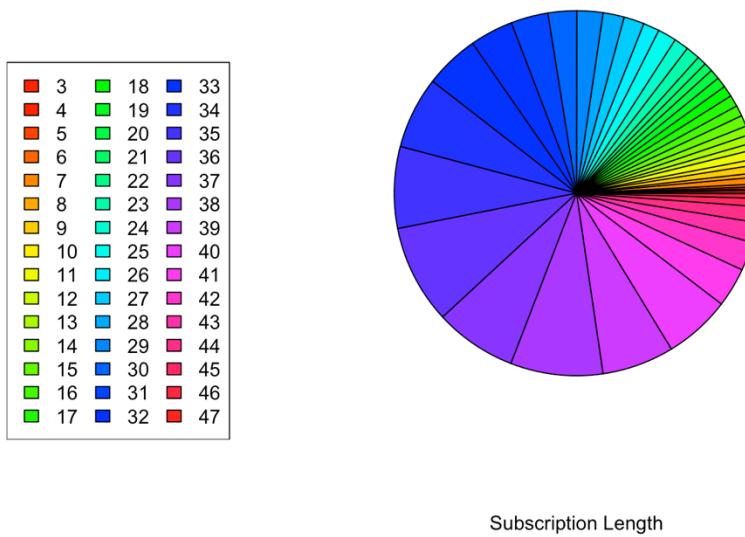
Figura 9 Istogramma Subscription Length

Questo grafico evidenzia l'asimmetria della distribuzione, con un'alta concentrazione di osservazioni per valori elevati e una coda verso sinistra.

Un'analisi delle frequenze relative [10] attraverso la **Funzione di Distribuzione Empirica (discreta)** conferma che i valori di Subscription Length si distribuiscono tra 3 e 47, rappresentando adeguatamente l'ampiezza della fruizione del servizio.



*Figura 10 FDE Subscription Length*



*Figura 11 Diagramma a torta Subscription Length*

Il **diagramma a torta** illustra anche qui la distribuzione equa dei valori.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza [4]: 73.501**
- **Deviazione standard [5]: 8.57**
- **Coefficiente di variazione [6]: 26.35%**

Un coefficiente di variazione tra il 15% e il 30% indica una moderata dispersione rispetto alla media.

L'analisi tramite diagramma di Pareto permette di visualizzare come le [9] siano associate alla frequenza relativa cumulativa, sottolineando la dispersione dei valori equa.

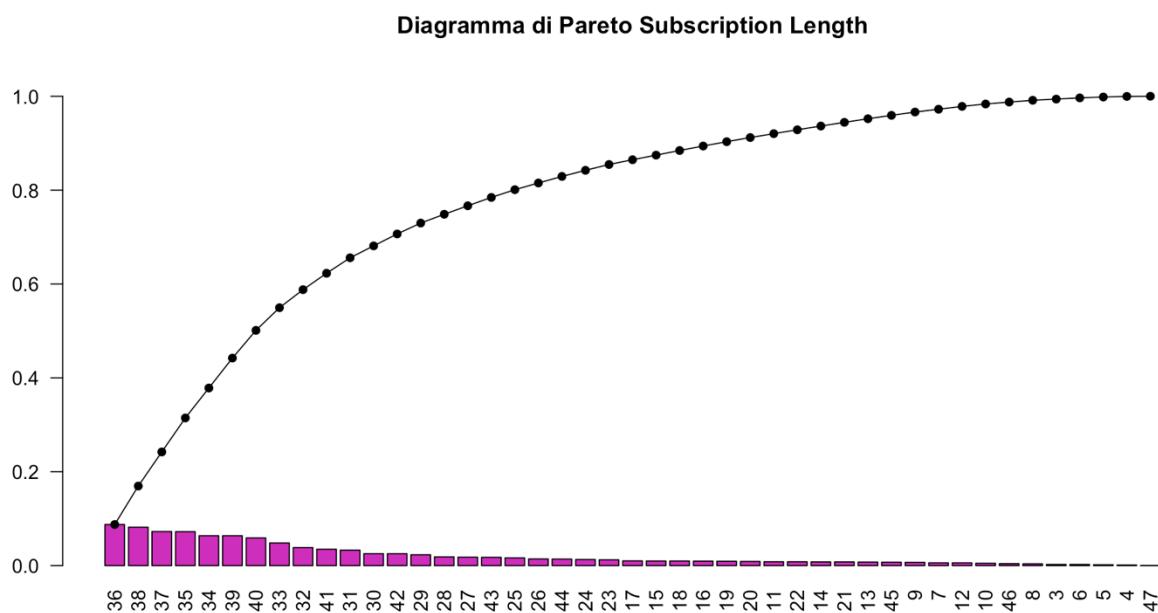
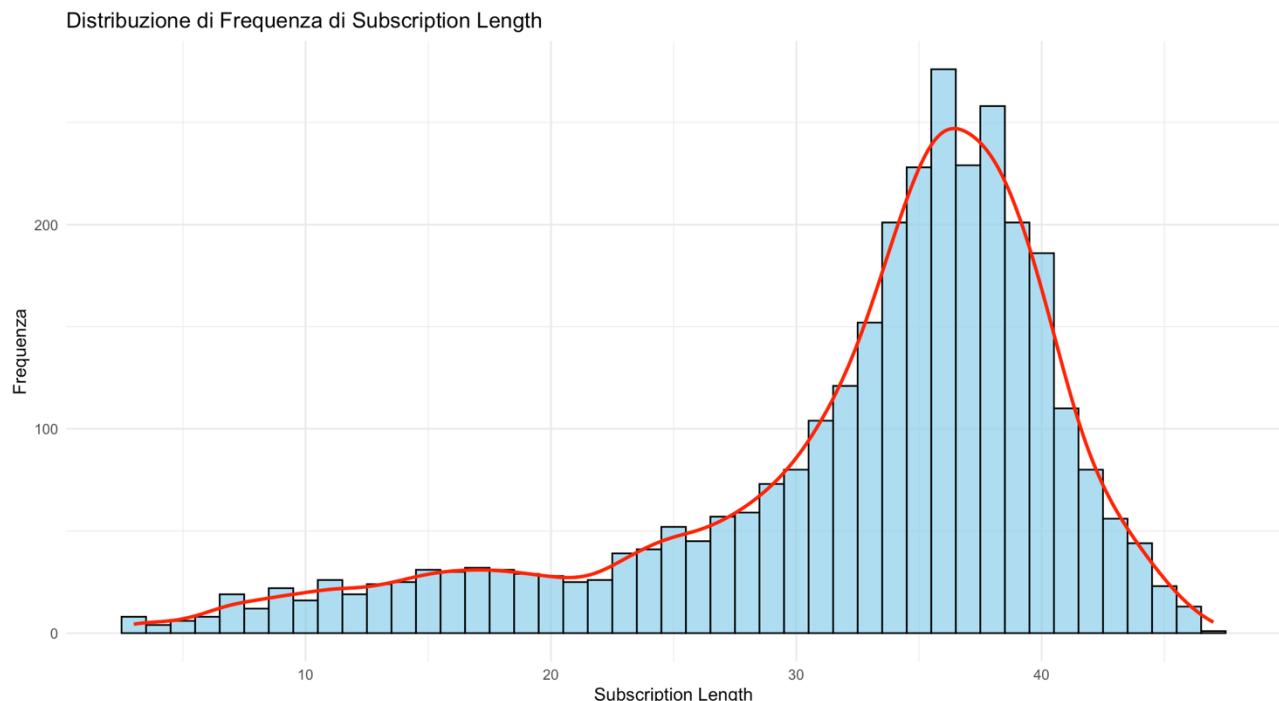


Figura 12 Diagramma di Pareto Subscription Length

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: -1.30**, che conferma l'asimmetria verso sinistra.
- **Curtosi [8]: 4.21**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza dei mesi di sottoscrizione, confermando le caratteristiche sopra descritte.



## 2.1.4. Charge Amount

La feature “Charge Amount” è una variabile ordinale (quantitativa discreta) (0: importo più basso, 10: importo più alto) che rappresenta l’importo addebitato al fruitore del servizio. Per una completa caratterizzazione statistica della variabile, si procederà con un’analisi delle sue misure di centralità e dispersione, seguita da un’analisi grafica.

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

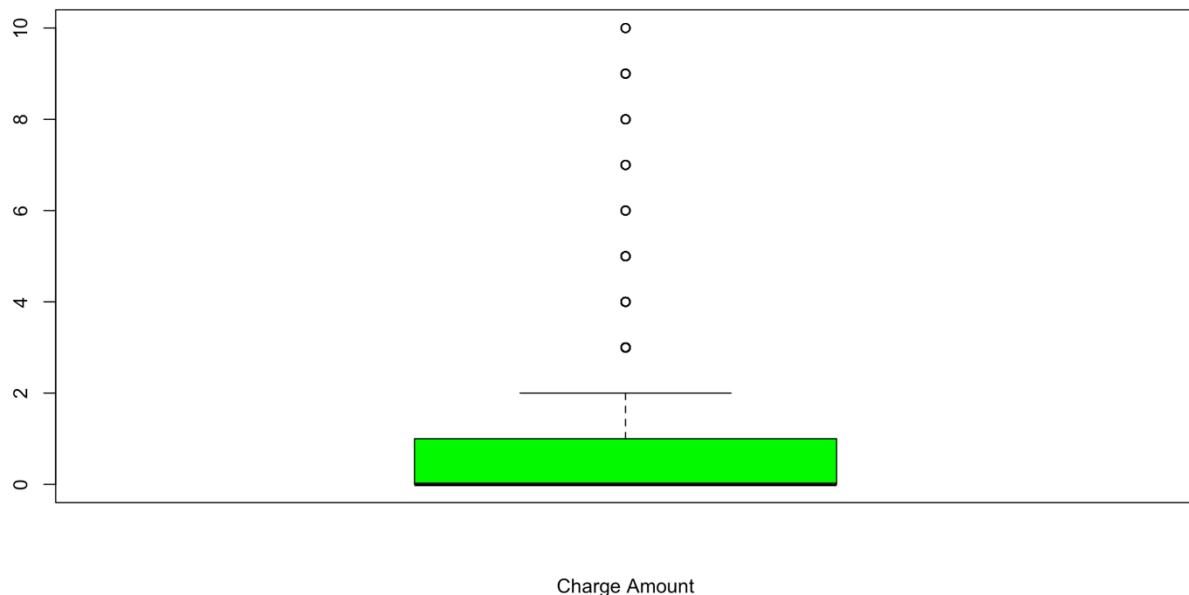
- **Media campionaria [1]**: La media del campo “Charge Amount” risulta pari a **0.94**.
- **Mediana campionaria [2]**: La mediana è pari a **0**.
- **Moda campionaria [3]**: La moda invece risulta essere **0**.

La predominanza della moda pari a 0 indica una distribuzione unimodale, con un picco concentrato sul valore zero. Dalle misure di media e mediana possiamo desumere che la distribuzione sia asimmetrica positiva (sbilanciata a destra).

In altre parole:

- **Asimmetria verso destra**: La distribuzione presenta una "coda" estesa a destra a causa della presenza di valori più elevati di Charge Amount.
- **Concentrazione attorno allo zero**: La maggior parte degli utenti riporta un valore di importo addebitato basso, con valori di media spostati a destra rispetto alla mediana e alla moda.

Un boxplot della variabile *Charge Amount* permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.



*Figura 14 Boxplot Charge Amount*

Possiamo notare dall'immagine che abbiamo molteplici outliers superiori alla mediana. Utilizzando lo scarto interquartile, abbiamo rilevato i seguenti outliers: **3, 4, 5, 6, 7, 8, 9, 10.**

Tramite poi una funzione apposita confermiamo che il **primo quartile è 0.00** mentre il **terzo quartile è 1.00**.

Inoltre, abbiamo il **minimo uguale a 0.00** ed un **massimo uguale a 10.0000**.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] del dato in questione.

L'istogramma della variabile Charge Amount fornisce una visualizzazione delle frequenze assolute [9] degli importi addebitati ai fruitori del servizio. Le ascisse gli importi addebitati, mentre le ordinate mostrano il numero di utenti corrispondenti.

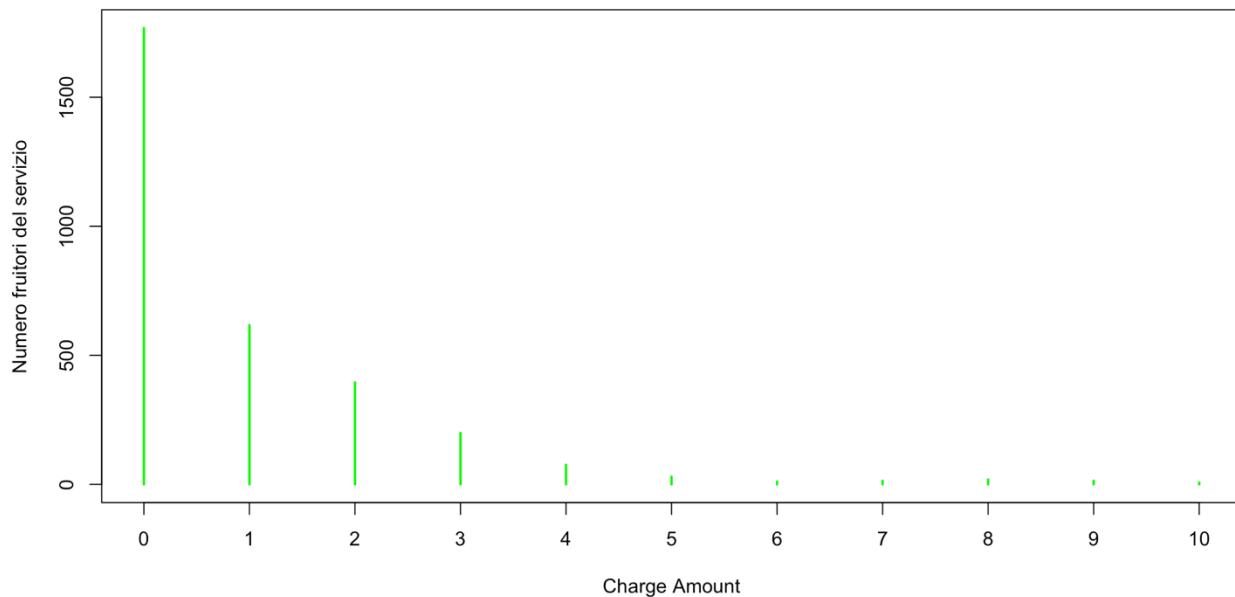


Figura 15 Istogramma Charge Amount

Questo grafico evidenzia l'asimmetria della distribuzione, con un'alta concentrazione di osservazioni per valori pari a zero e una coda verso destra.

Un'analisi delle frequenze relative [10] tramite **Funzione di Distribuzione Empirica (discreta)** evidenzia ulteriormente come una larga porzione degli utenti presenti valori prossimi allo zero.

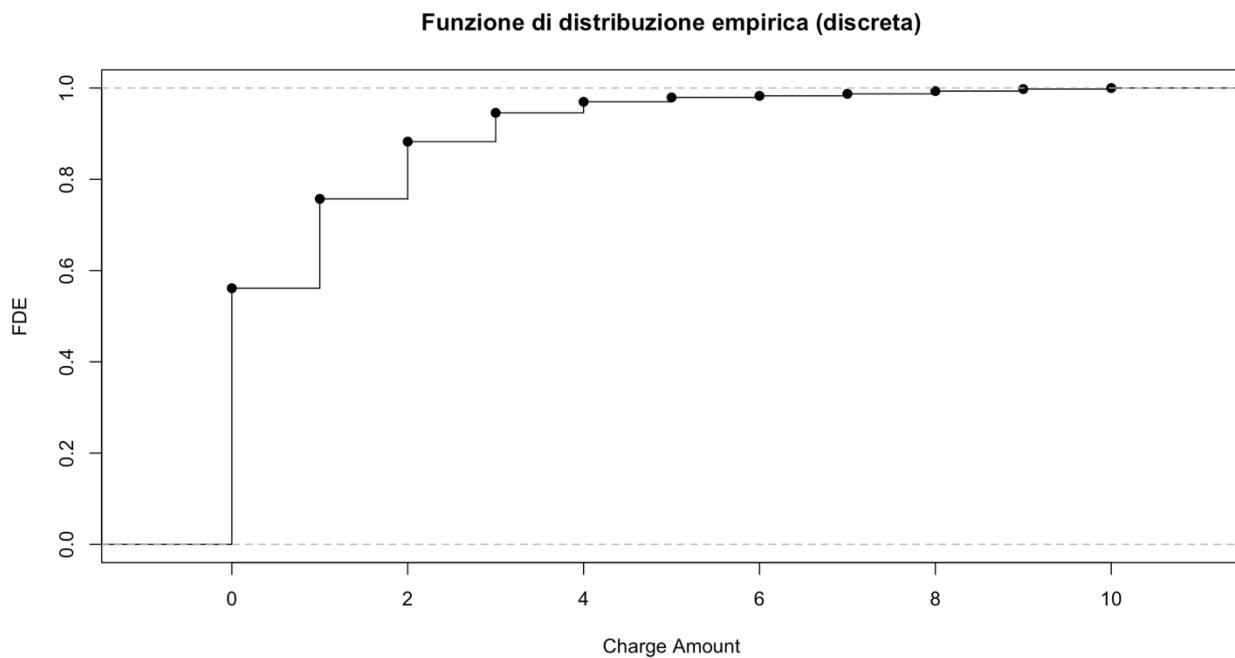


Figura 16 FDE Charge Amount

L'analisi della **Funzione di Distribuzione Empirica (FDE)** conferma ulteriormente l'asimmetria menzionata precedentemente: mostra infatti una rapida crescita iniziale (data

dalla frequenza elevata di valori prossimi allo zero), seguita da un incremento più graduale in corrispondenza dei valori più elevati.

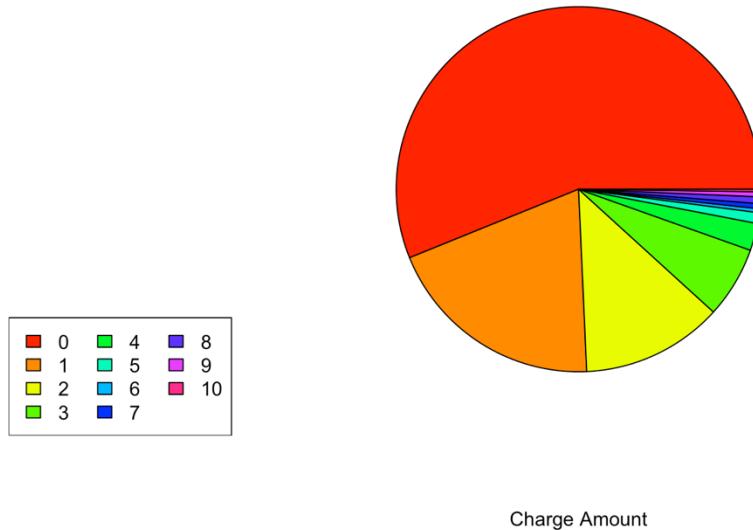


Figura 17 Diagramma a torta Charge Amount

Una rappresentazione tramite **diagramma a torta** illustra che la maggior parte degli utenti ha un basso importo addebitato.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza [4]: 2.31**
- **Deviazione standard [5]: 1.52**
- **Coefficiente di variazione [6]: 161.33%**

L'elevato coefficiente di variazione indica una **forte dispersione dei valori** rispetto alla media, segnalando una significativa variabilità nel numero di Charge Amount tra gli utenti.

**Distribuzione di Frequenza tramite Diagramma di Pareto:** L'analisi tramite diagramma di Pareto permette di visualizzare come le frequenze assolute [9] siano associate alla frequenza relativa cumulativa, sottolineando la predominanza di utenti importi più bassi e il peso cumulativo degli utenti con più fallimenti.

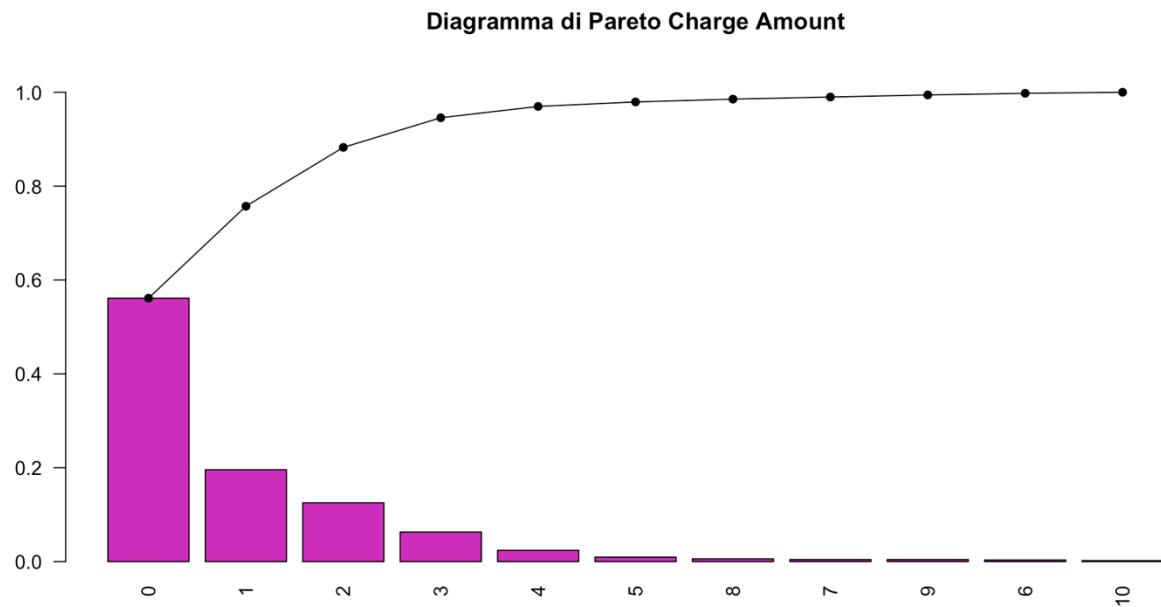


Figura 18 Diagramma di Pareto Charge Amount

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 2.58**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 11.84**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza degli importi addebitati, confermando le caratteristiche sopra descritte.

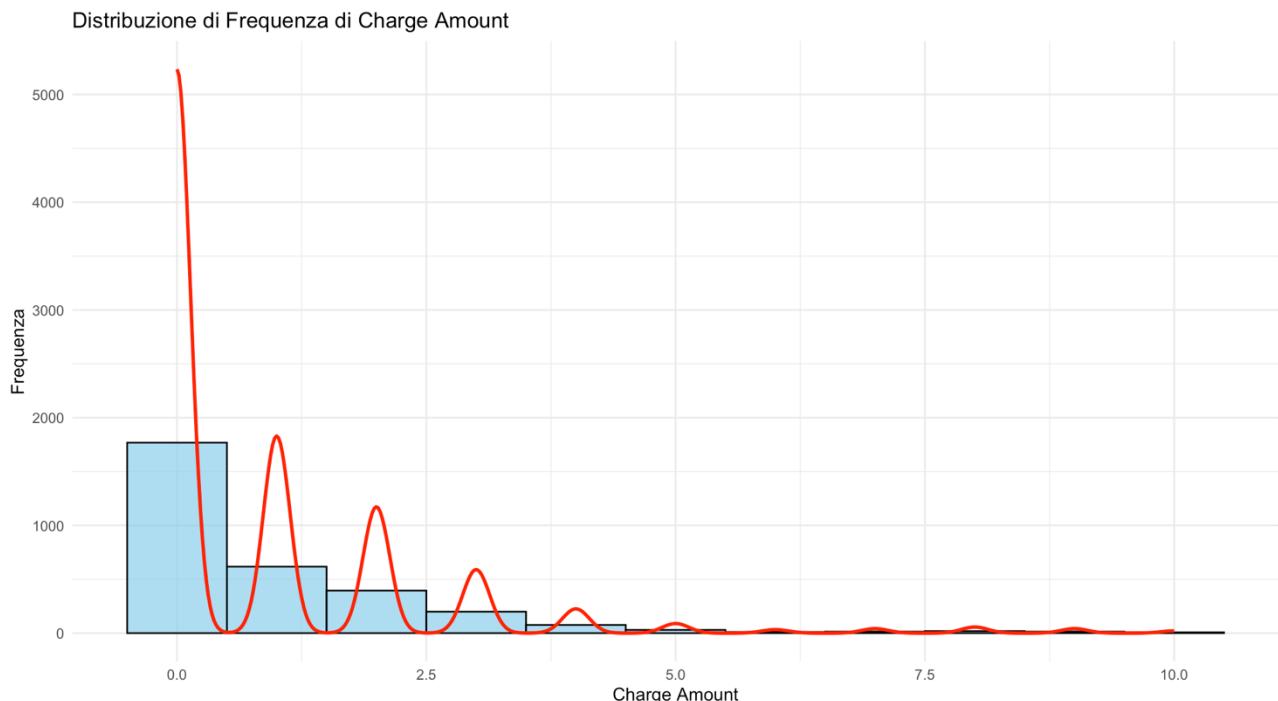


Figura 19 Distibuzione di frequenza Charge Amount

## 2.1.5. Seconds of Use

La feature “Second of use” è una variabile quantitativa discreta espressa in numeri interi, rappresentante il numero di secondi di fruizione del servizio.

Per una completa caratterizzazione statistica della variabile, si procederà con un’analisi delle sue misure di centralità e dispersione, seguita da un’analisi grafica.

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

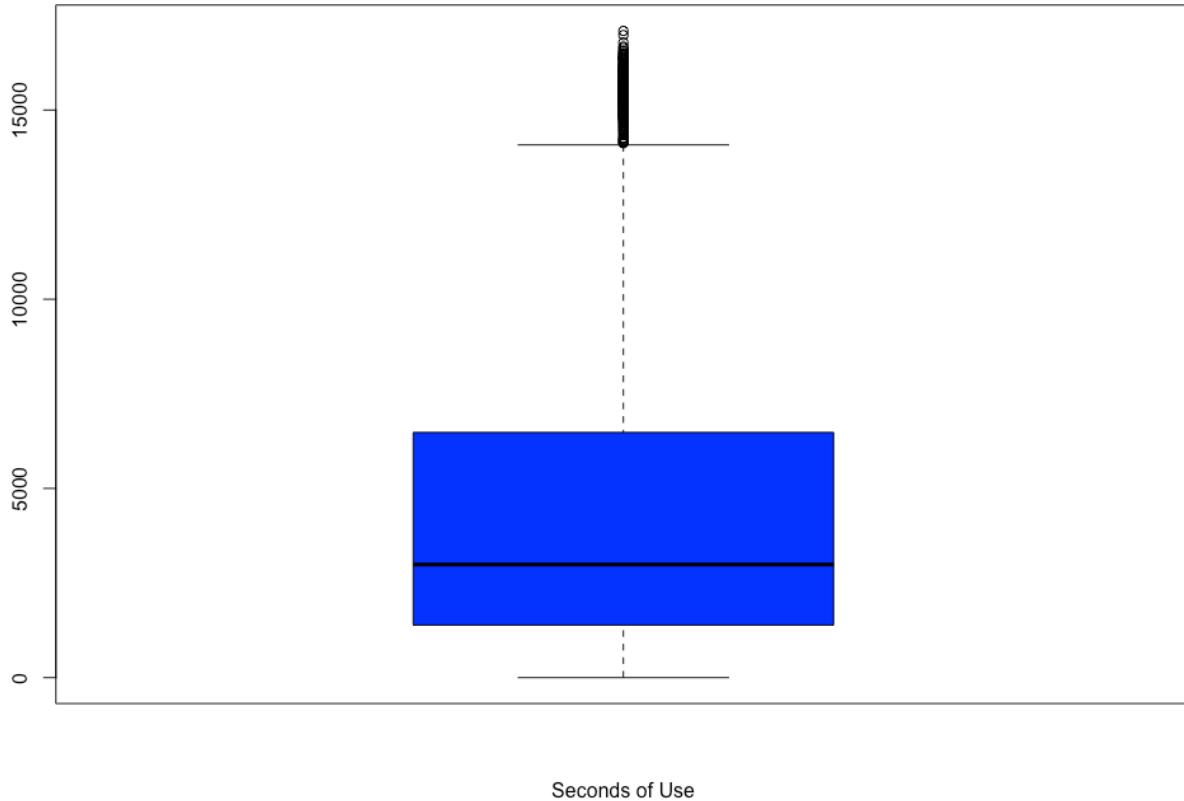
- **Media campionaria [1]**: La media del campo “Seconds of use” risulta pari a **4472.46**
- **Mediana campionaria [2]**: La mediana è pari a **2990**.
- **Moda campionaria [3]**: La moda invece risulta essere **0**.

La prevalenza della moda a 0 suggerisce una distribuzione unimodale, con un picco iniziale intorno a questo valore. La relazione tra media, mediana e moda indica una distribuzione asimmetrica positiva (sbilanciata a sinistra).

Nello specifico:

- **Asimmetria verso destra**: La distribuzione è caratterizzata da una coda a destra, che rappresenta la presenza di valori alti rispetto alla moda che è **0**.
- **Moda**: Essendo la moda a 0, rappresenta il valore più frequente, confermando che la maggior parte dei dati si concentra su valori relativamente bassi di Seconds of use e quindi molti dei fruitori non hanno effettivamente utilizzato il servizio.

Un boxplot della variabile *Seconds of use* permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.



*Figura 20 Boxplot Seconds of use*

Possiamo notare dall'immagine che abbiamo molteplici outliers superiori alla mediana. Utilizzando lo scarto interquartile, abbiamo rilevato i seguenti outliers: 15140, 15485, 16075, 15200, 15545, 16135, 15080, 15425, 16015, 15220, 15565, 16155, 15060, 15405, 15995, 15400, 15745, 16335, 14880, 15225, 15815, 15320, 15665, 16255, 14960, 15305, 15895, 14373, 15740, 16085, 16675, 14540, 14885, 15475, 15255, 15600, 16190, 15025, 15370, 15960, 15330, 15675, 16265, 14950, 15295, 15885, 14483, 15850, 16195, 16785, 14430, 14775, 15365, 14835, 15180, 15770, 14895, 15240, 15830, 15120, 15710, 14915, 15260, 15690, 15095, 15440, 16030, 14575, 14920, 15510, 15015, 15360, 15950, 14655, 15000, 15590, 15435, 15780, 16370, 14235, 14580, 15170, 14720, 15065, 15655, 14990, 15580, 14178, 16480, 14125, 14470, 15445, 15790, 16380, 14138, 15505, 16440, 15385, 15730, 16320, 14158, 15525, 15870, 16460, 14338, 15705, 16050, 16640, 15185, 15530, 16120, 14258, 15625, 15970, 16560, 15265, 15610, 16200, 14678, 16045, 16390, 16980, 14845, 15190, 15905, 16495, 16570, 14788, 16500, 17090, 14735, 15670. Tramite poi una funzione apposita confermiamo che il **primo quartile** è 1391 mentre il **terzo quartile** è 6478.

Inoltre, abbiamo il **minimo** uguale a 0 ed un **massimo** uguale a 17090.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] del dato in questione.

L'istogramma della variabile Seconds Of Use fornisce una visualizzazione delle frequenze assolute [9] dei secondi di fruizione tra gli utenti. Le ascisse rappresentano il numero di secondi, mentre le ordinate mostrano il numero di utenti corrispondenti.

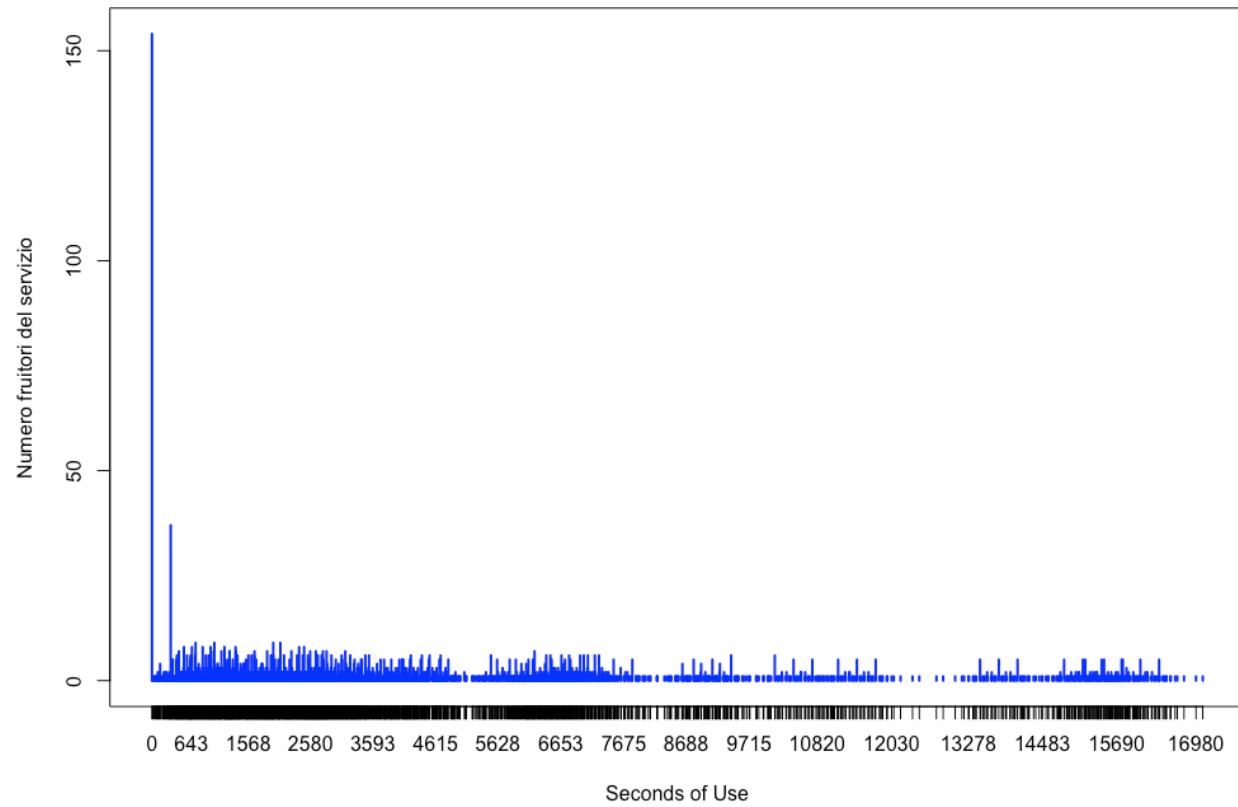
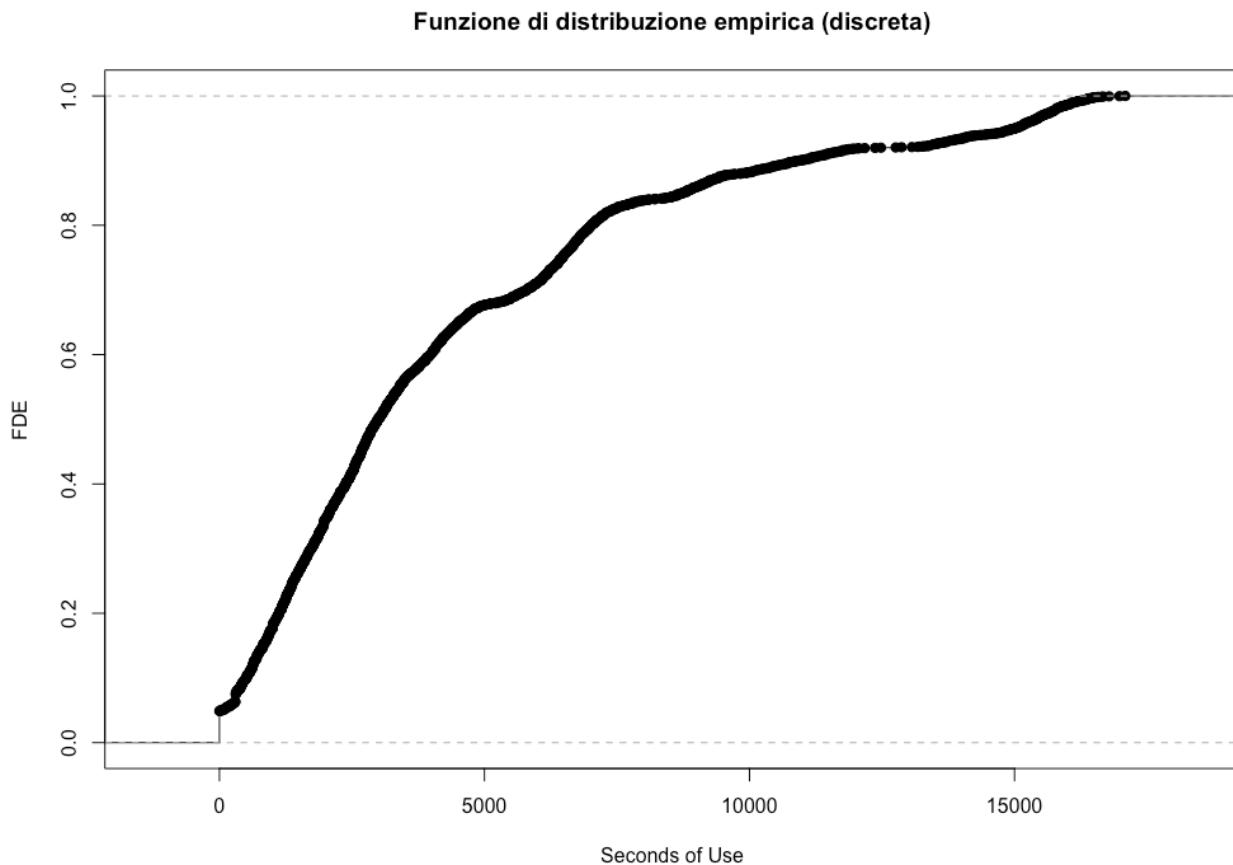


Figura 21 Istogramma Seconds of use

Questo grafico evidenzia l'asimmetria della distribuzione, con un'alta concentrazione di osservazioni per valori elevati e una coda verso destra.

Inoltre, si può notare anche la varietà di dati assunti dalla variabile.

Un'analisi delle frequenze relative [10] attraverso la **Funzione di Distribuzione Empirica (discreta)** conferma che i valori di Seconds of use si distribuiscono tra 0 e 17090, rappresentando adeguatamente l'ampiezza della fruizione del servizio.



*Figura 22 FDE Subscription Length*

Un diagramma torta nel caso di questa variabile sarebbe eccessivamente confusionario data l'eccessivo numero di stati che assume la variabile.

Per la stessa motivazione il diagramma di pareto sarebbe superfluo dato che risulterebbe circa come la FDE.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza [4]: 17622437**
- **Deviazione standard [5]: 4197.91**
- **Coefficiente di variazione [6]: 93.86%**

Un coefficiente di variazione così alto era scontato data la sensibilità agli outliers del coefficiente di variazione.

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 1.32**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 3.99**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza dei mesi di sottoscrizione, confermando le caratteristiche sopra descritte.

Ovviamente il picco della funzione non sarà evidente data la varianza dei valori.

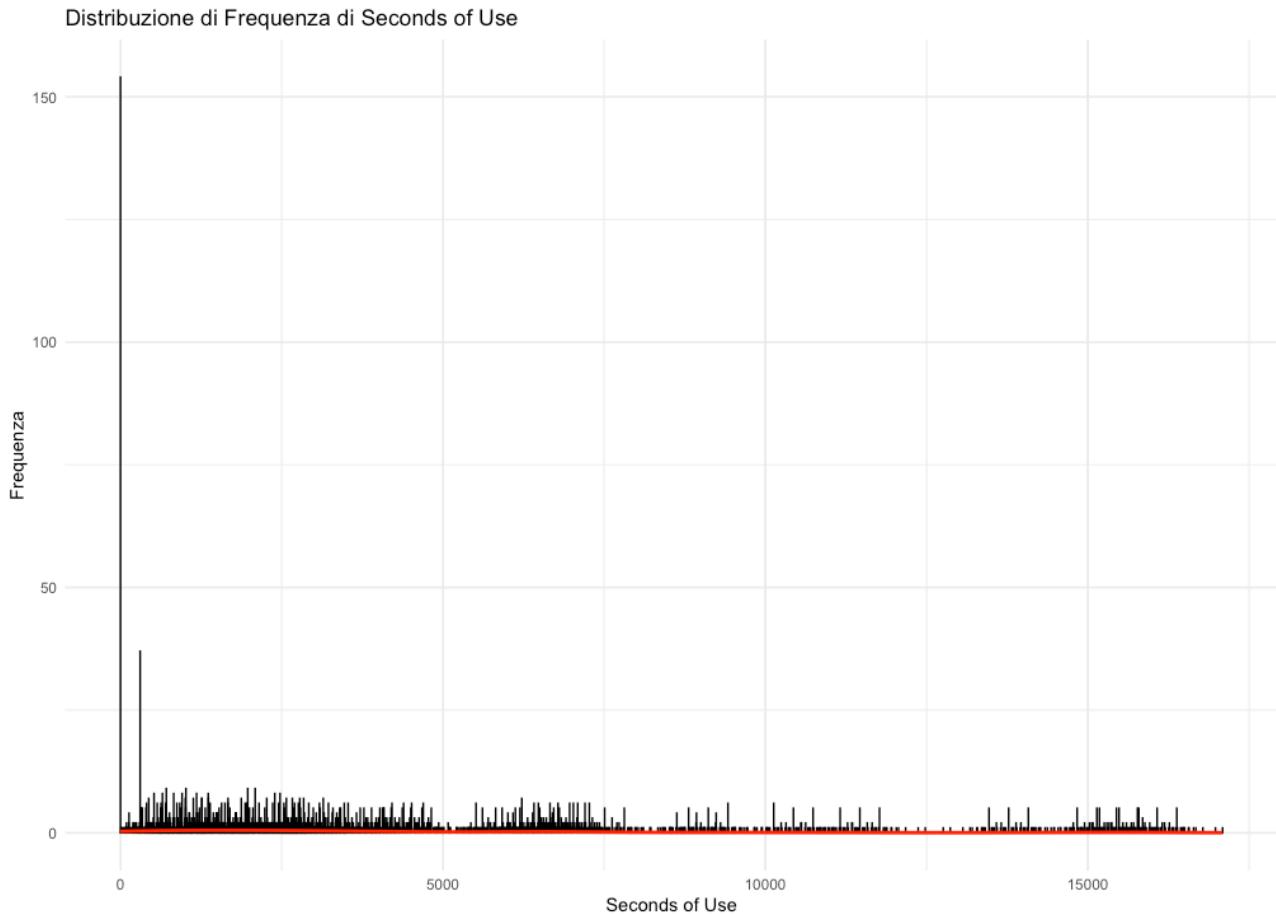


Figura 23 Distribuzione di frequenza Seconds of use

Data la scarsa rappresentanza dei dati e la conseguente difficoltà nell'analizzarli, tentiamo di aggiungere una nuova feature al dataset chiamata Second of use intervals la quale andrà a suddividere i valori assunti dalla variabile in più intervalli. Per dividere tutto in intervalli è stato deciso di utilizzare un numero di intervalli tali che permetta anche una rappresentazione soddisfacente pari a 50.

## 2.1.6. Seconds of use intervals (Feature aggiunta)

Seconds of use intervals è una Colonna risultante dalla suddivisione in 50 intervalli della feature Seconds of use per avere una maggiore precisione di analisi grafica.

Come possiamo notare quindi dalla seguente distribuzione di frequenza, riusciamo subito ora a notare la distribuzione dei valori.

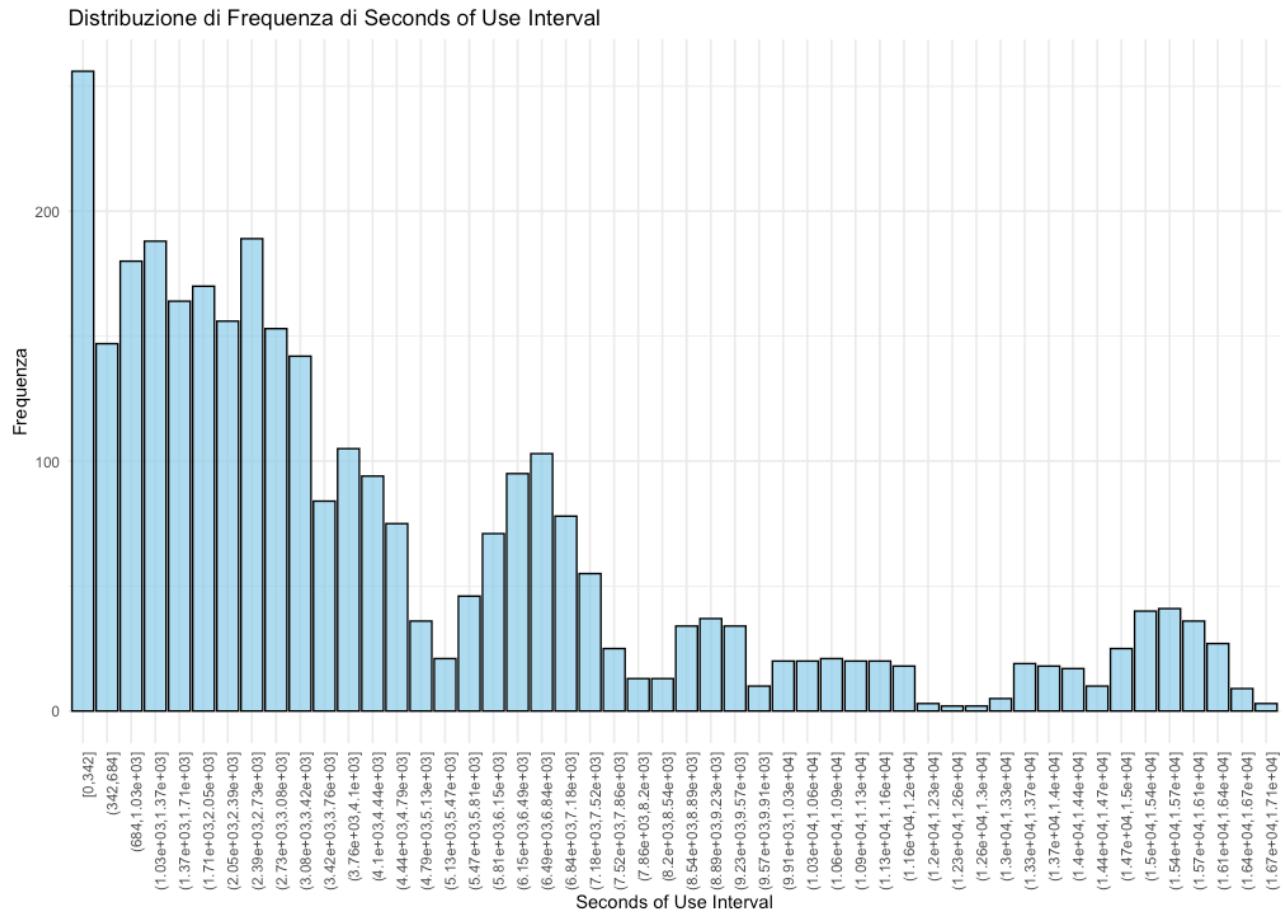
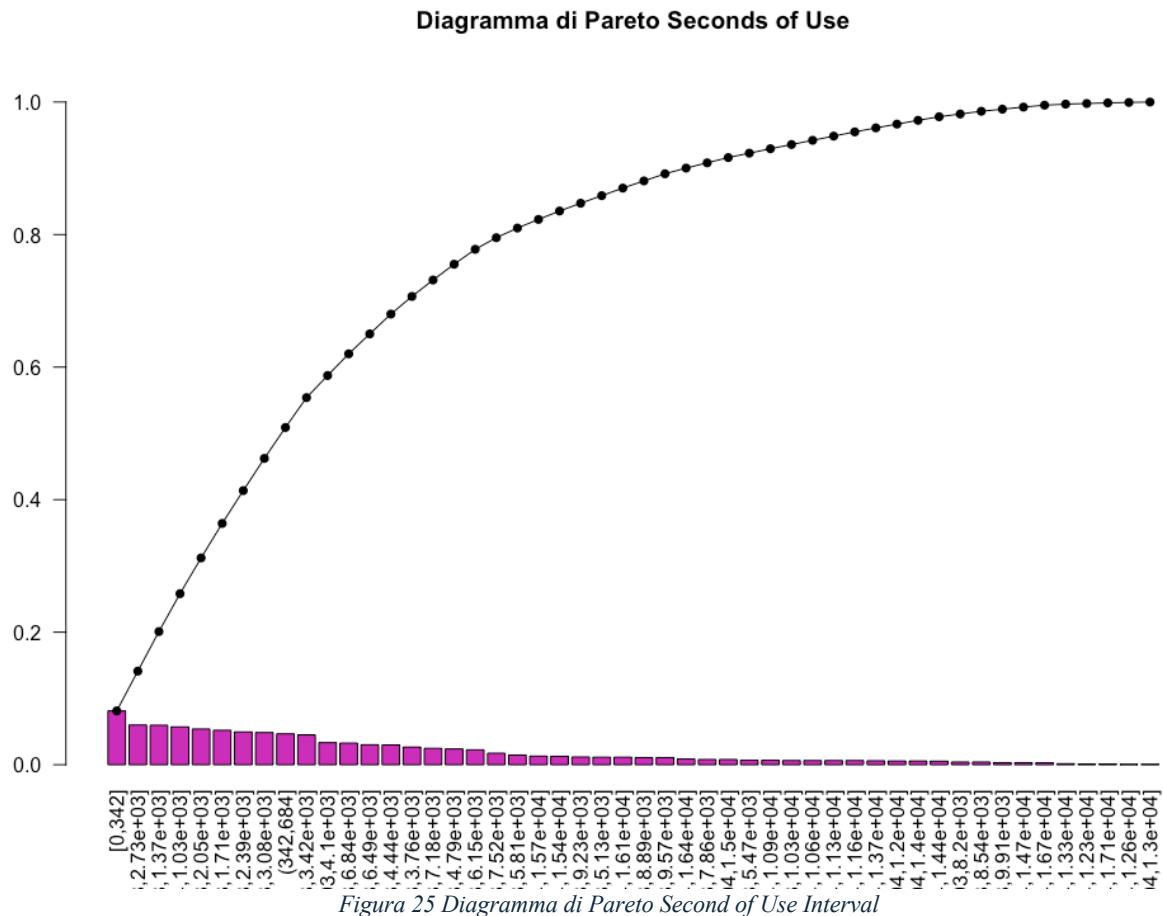


Figura 24 Distribuzione di frequenza Seconds of use Interval

Possiamo inoltre andare a studiare come si comportano le frequenze assolute [9] in correlazione alle frequenze assolute cumulative tramite il diagramma di Pareto.



## 2.1.7. Frequency of use

La feature “Frequency of use” è una variabile quantitativa discreta espressa in numeri interi, la quale non è ben specificata dai creatori del dataset. Da questo momento in poi assumiamo che questa variabile viene utilizzata per valorizzare il numero di chiamate distinte per ogni fruitore del servizio. Per una completa caratterizzazione statistica della variabile, si procederà con un’analisi delle sue misure di centralità e dispersione, seguita da un’analisi grafica.

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Frequency of user” risulta pari a **69.46**.
- **Mediana campionaria [2]**: La mediana è pari a **54**.
- **Moda campionaria [3]**: La moda invece risulta essere **0**.

La predominanza della moda pari a 0 indica una distribuzione unimodale, con un picco concentrato sul valore zero. Dalle misure di media e mediana possiamo desumere che la distribuzione sia asimmetrica positiva (sbilanciata a destra).

In altre parole:

- **Asimmetria verso destra**: La distribuzione presenta una "coda" estesa a destra a causa della presenza di valori più elevati di Frequency of use.
- **Concentrazione attorno allo zero**: La maggior parte degli utenti riporta 0 chiamate distinte, con valori di media spostati a destra rispetto alla mediana e alla moda.

Un boxplot della variabile *Frequency of use* permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.

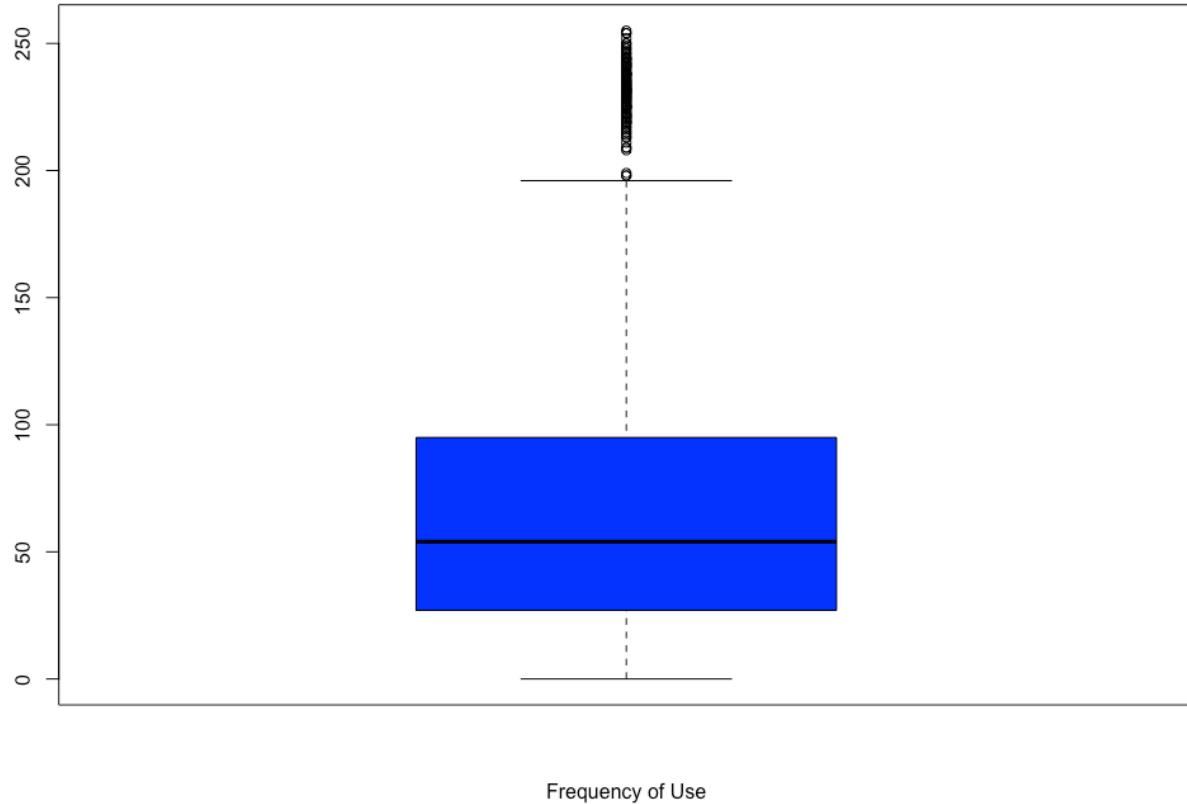


Figura 26 Boxplot Frequency of use

Possiamo notare dall'immagine che abbiamo molteplici outliers.

Utilizzando lo scarto interquartile, abbiamo rilevato i seguenti outliers: 198, 199, 208, 209, 211, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 252, 254, 255.

Tramite poi una funzione apposita confermiamo che il **primo quartile** è 27.00 mentre il **terzo quartile** è 95.00.

Inoltre, abbiamo il **minimo** uguale a 0.00 ed un **massimo** uguale a 255.00.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] delle chiamate totali distinte dei fruitori.

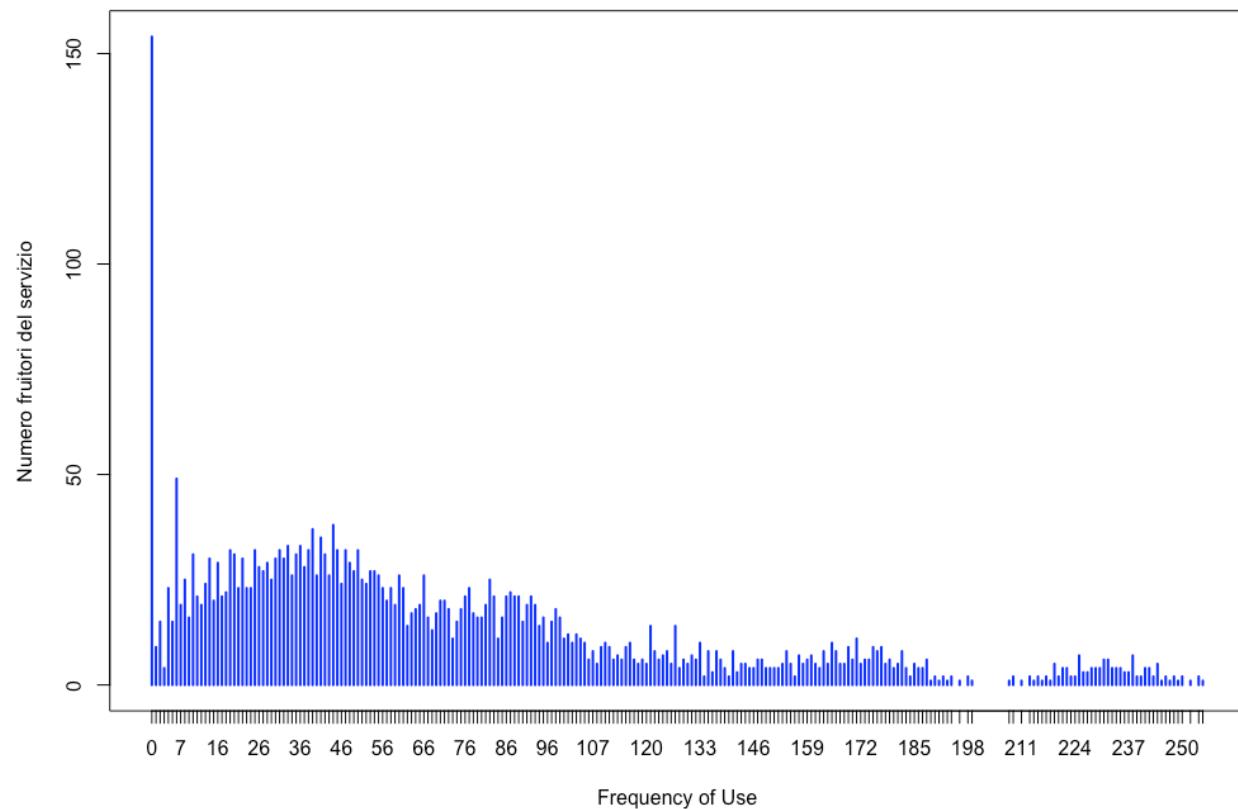


Figura 27 Istogramma Frequency of use

Un istogramma della variabile *Frequency of use* mostra la frequenza assoluta [9] delle chiamate per ciascun valore osservato. Le ascisse rappresentano il numero di chiamate distinte, mentre le ordinate indicano la quantità di utenti corrispondenti.

Il grafico conferma una distribuzione asimmetrica, con una concentrazione di osservazioni attorno a valori bassi e una coda verso destra.

Un'analisi delle frequenze relative [10] tramite **Funzione di Distribuzione Empirica (discreta)** evidenzia ulteriormente come una larga porzione degli utenti presenti valori prossimi allo zero.

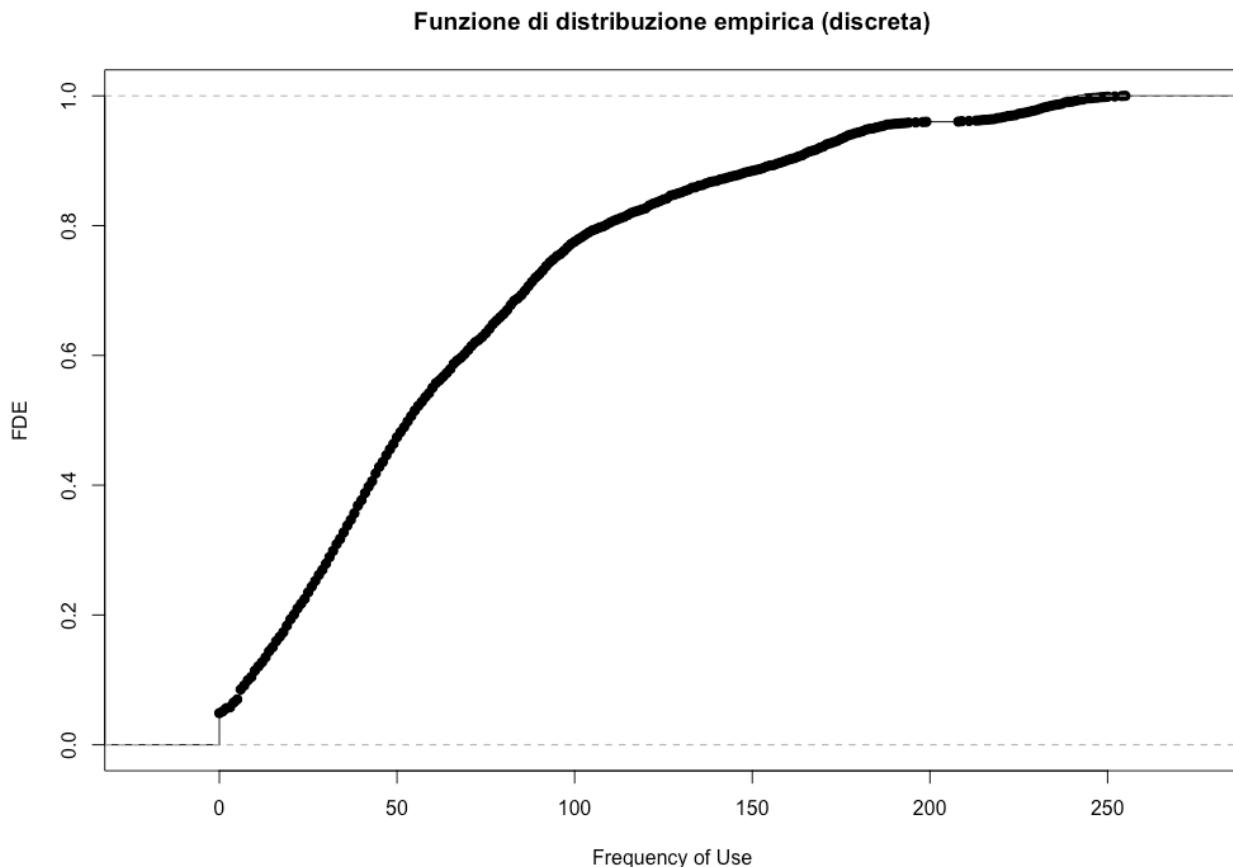


Figura 28 Funzione di distribuzione empirica (discreta) Frequency of use

L'analisi della **Funzione di Distribuzione Empirica (FDE)** conferma ulteriormente l'asimmetria menzionata precedentemente: mostra infatti una rapida crescita iniziale (data dalla frequenza elevata di valori prossimi allo zero), seguita da un incremento più graduale in corrispondenza dei valori più elevati.

Data l'elevata dispersione dei valori sarebbe superfluo avere un diagramma a torta per la variabile in questione.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza [4]: 3296.288**
- **Deviazione standard [5]: 57.41331**
- **Coefficiente di variazione [6]: 82.65589%**

L'elevato coefficiente di variazione indica una **forte dispersione dei valori** rispetto alla media, segnalando una significativa variabilità nel numero di chiamate tra gli utenti.

**Distribuzione di Frequenza tramite Diagramma di Pareto:** L'analisi tramite diagramma di Pareto permette di visualizzare come le frequenze assolute [9] siano associate alla frequenza relativa cumulativa, sottolineando la predominanza di utenti con poche chiamate e il peso cumulativo degli utenti con più chiamate.

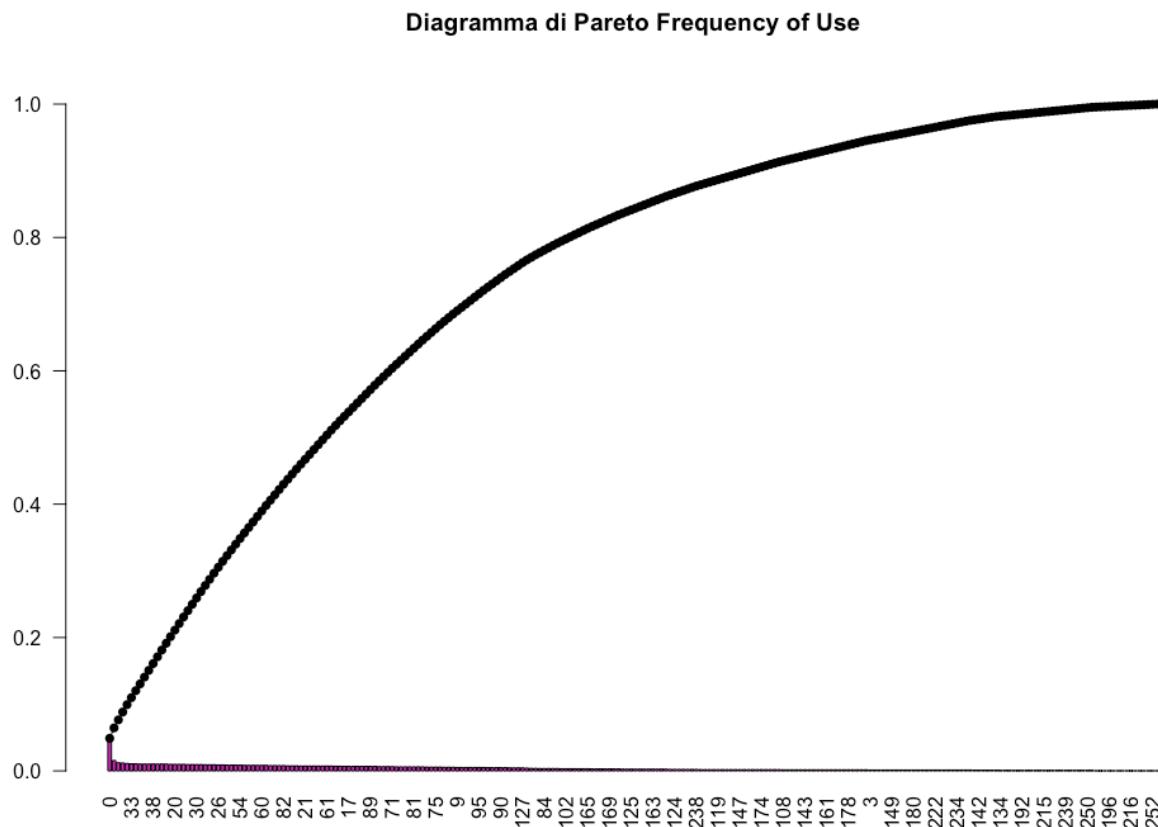


Figura 29 Diagramma di Pareto Frequency of Use

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 1.143622**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 3.816919**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza del numero di chiamate distinte, confermando le caratteristiche sopra descritte.

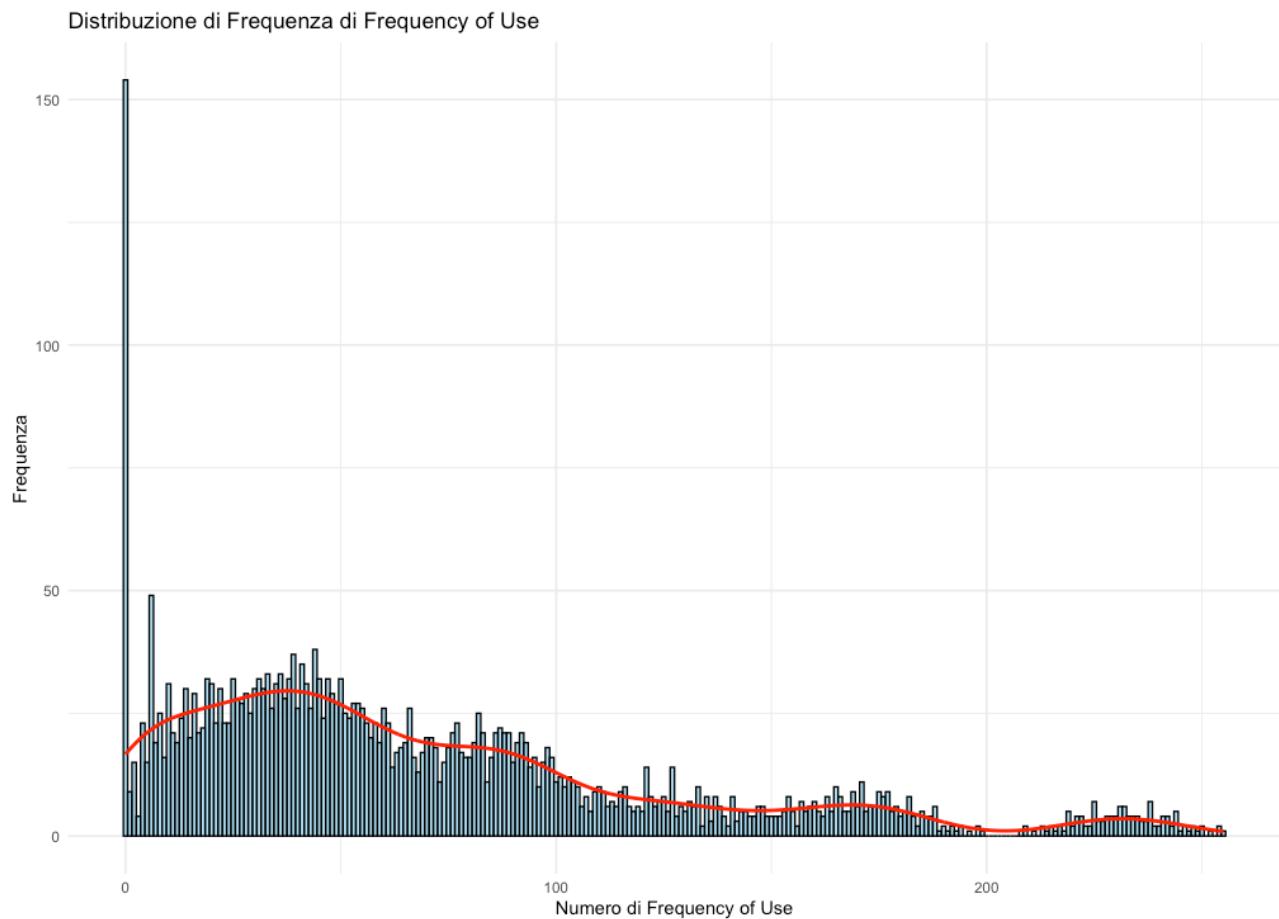


Figura 30 Distribuzione di frequenza Frequency of use

## 2.1.8. Frequency of SMS

La feature “Frequency of sms” è una variabile quantitativa discreta espressa in numeri interi, il quale compito non è ben specificato dai creatori del dataset. Da questo momento in poi assumiamo che questa variabile viene utilizzata per valorizzare il numero di sms distinti per ogni fruitore del servizio. Per una completa caratterizzazione statistica della variabile, si procederà con un’analisi delle sue misure di centralità e dispersione, seguita da un’analisi grafica.

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Frequency of SMS” risulta pari a 73.18.
- **Mediana campionaria [2]**: La mediana è pari a 21.
- **Moda campionaria [3]**: La moda invece risulta essere 0.

La predominanza della moda pari a 0 indica una distribuzione unimodale, con un picco concentrato sul valore zero. Dalle misure di media e mediana possiamo desumere che la distribuzione sia asimmetrica positiva (sbilanciata a destra).

In altre parole:

- **Asimmetria verso destra**: La distribuzione presenta una "coda" estesa a destra a causa della presenza di valori più elevati di Frequency of SMS.
- **Concentrazione attorno allo zero**: La maggior parte degli utenti riporta e chiama distinte, con valori di media spostati a destra rispetto alla mediana e alla moda.

Un boxplot della variabile *Frequency of SMS* permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.

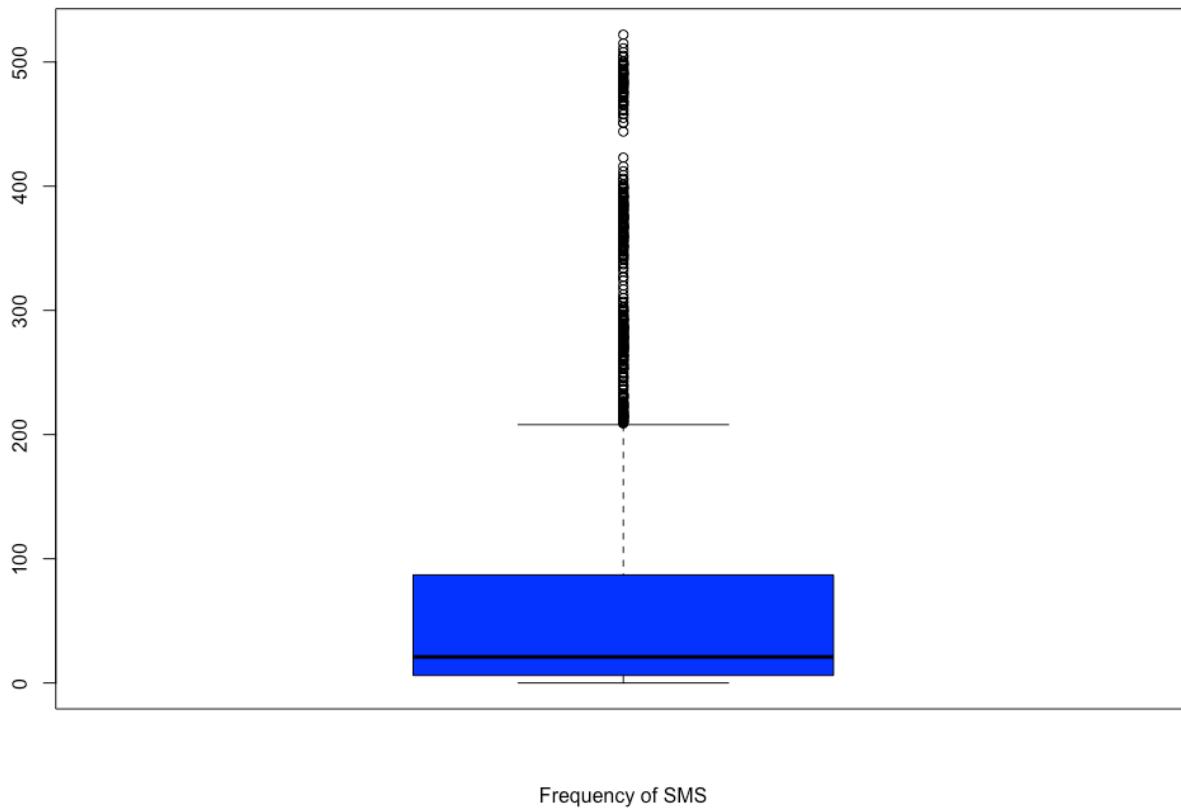


Figura 31 Boxplot frequency of SMS

Possiamo notare dall'immagine che abbiamo molteplici outliers.

Utilizzando lo scarto interquartile, abbiamo rilevato molti outliers(troppi per essere trascritti sottoforma di lista) che vano da **222 a 522**.

Tramite poi una funzione apposita confermiamo che il **primo quartile è 6.00** mentre il **terzo quartile è 87.00**.

Inoltre, abbiamo il **minimo** uguale a **0.00** ed un **massimo** uguale a **522.00**.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] degli SMS totali distinte dei fruitori.

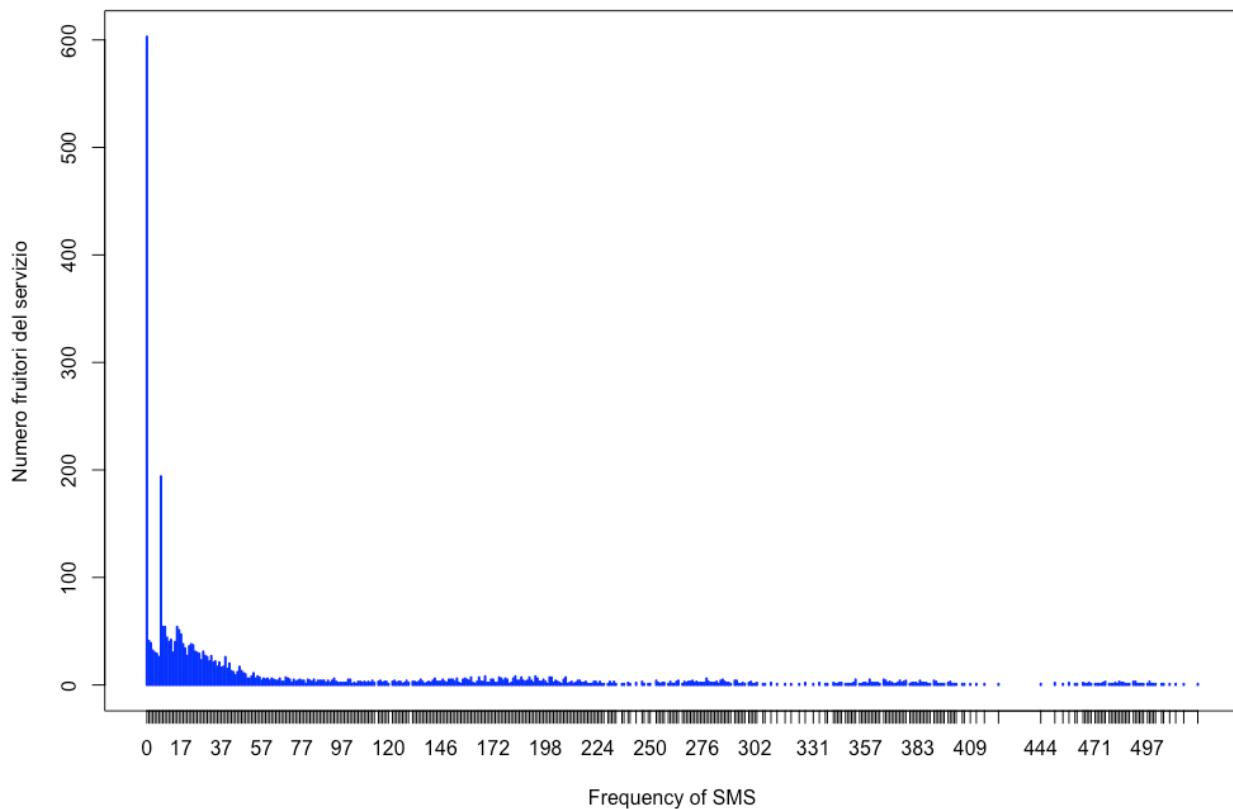


Figura 32 Istogramma Frequency of use

Un istogramma della variabile *Frequency of SMS* mostra la frequenza assoluta [9] degli SMS per ciascun valore osservato. Le ascisse rappresentano il numero di SMS, mentre le ordinate indicano la quantità di utenti corrispondenti.

Il grafico conferma una distribuzione asimmetrica, con una concentrazione di osservazioni attorno a valori bassi e una coda verso destra.

Un'analisi delle frequenze relative [10] tramite **Funzione di Distribuzione Empirica (discreta)** evidenzia ulteriormente come una larga porzione degli utenti presenti valori prossimi allo zero.

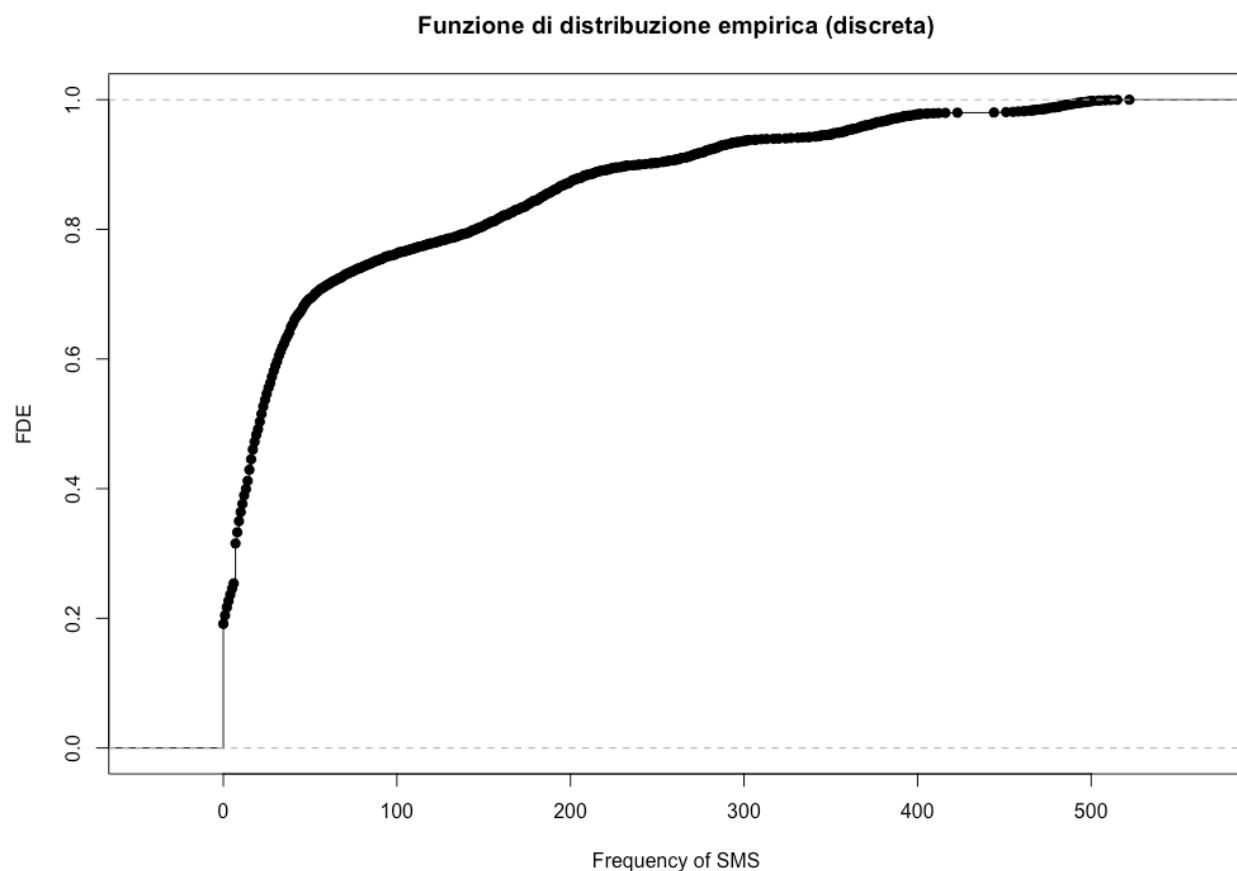


Figura 33 Funzione di distribuzione empirica (discreta) Frequency of SMS

L'analisi della **Funzione di Distribuzione Empirica (FDE)** conferma ulteriormente l'asimmetria menzionata precedentemente: mostra infatti una rapida crescita iniziale (data dalla frequenza elevata di valori prossimi allo zero), seguita da un incremento più graduale in corrispondenza dei valori più elevati.

Data l'elevata dispersione dei valori sarebbe superfluo avere un diagramma a torta per la variabile in questione.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza [4]: 12597.27**
- **Deviazione standard [5]: 112.24**
- **Coefficiente di variazione [6]: 153.38%**

L'elevato coefficiente di variazione indica una **forte dispersione dei valori** rispetto alla media, segnalando una significativa variabilità nel numero di SMS tra gli utenti.

**Distribuzione di Frequenza tramite Diagramma di Pareto:** L'analisi tramite diagramma di Pareto permette di visualizzare come le frequenze assolute [9] siano associate alla frequenza relativa cumulativa, sottolineando la predominanza di utenti con pochi SMS inviati e il peso cumulativo degli utenti che hanno inviato più SMS.

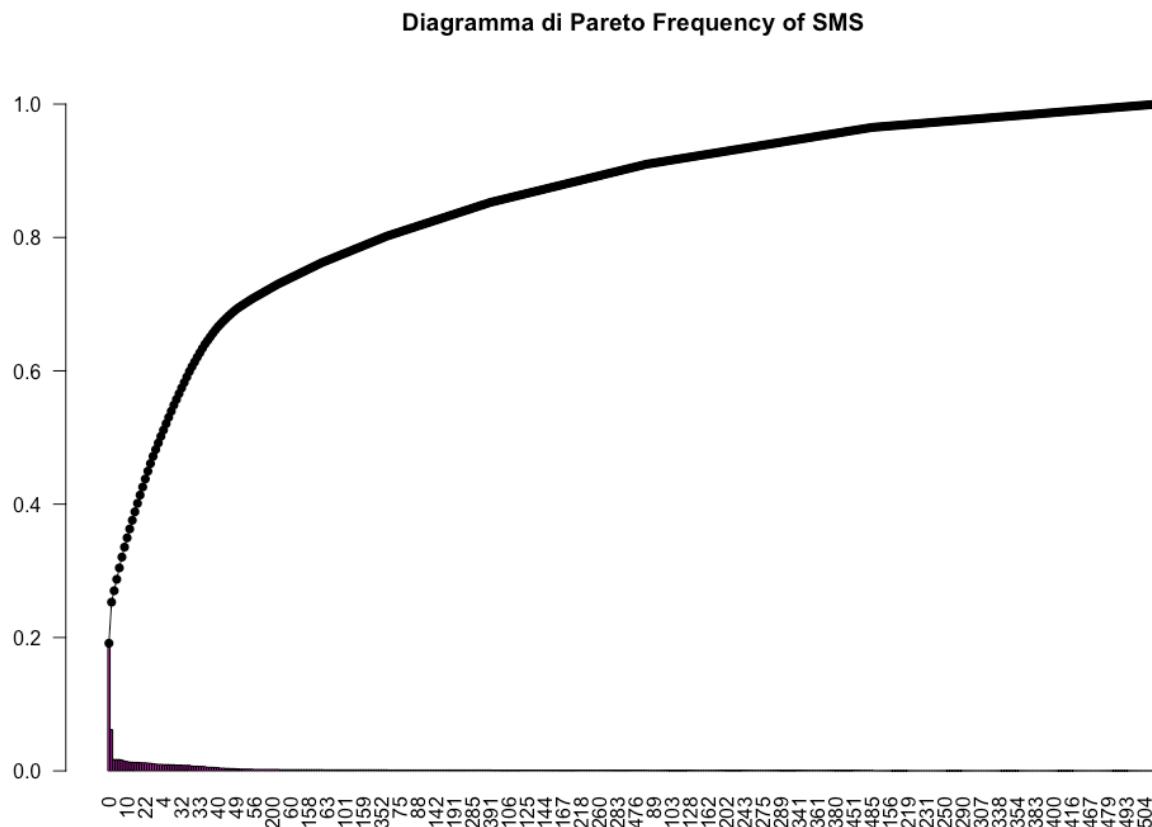


Figura 34 Diagramma di Pareto Frequency of SMS

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 1.97**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 6.25**, indicando una distribuzione leptocurtica, caratterizzata da un picco molto elevato.

Il seguente grafico riassume la distribuzione di frequenza del numero di SMS, confermando le caratteristiche sopra descritte.

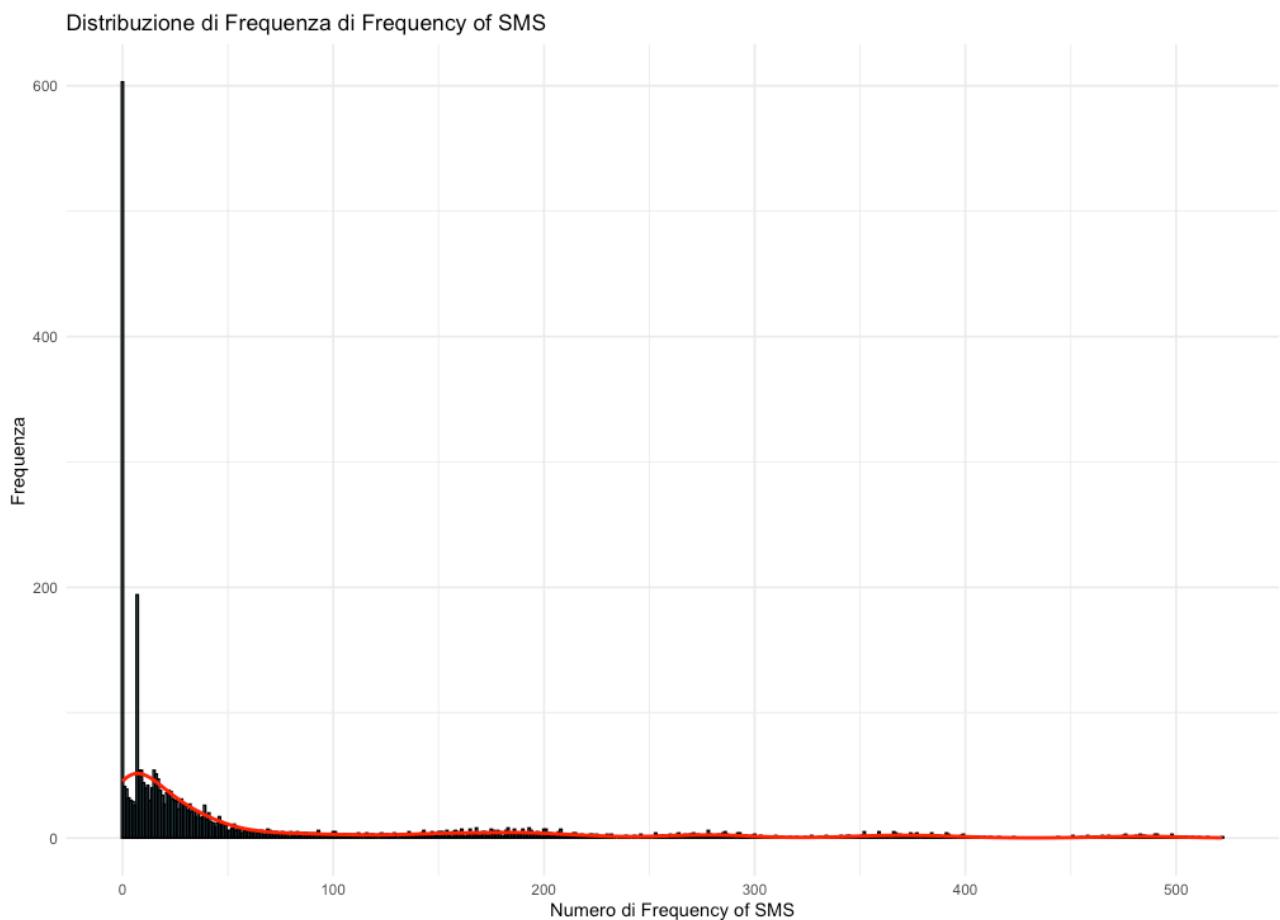


Figura 35 Distribuzione di frequenza Frequency of SMS

## 2.1.9. Distinct Called Numbers

La feature “[Distinct Call Numbers](#)” è una variabile quantitativa discreta espressa in numeri interi, rappresentante la quantità di numeri telefonici distinti chiamati per ogni fruitore del servizio. Per una completa caratterizzazione statistica della variabile, si procederà con un’analisi delle sue misure di centralità e dispersione, seguita da un’analisi grafica.

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

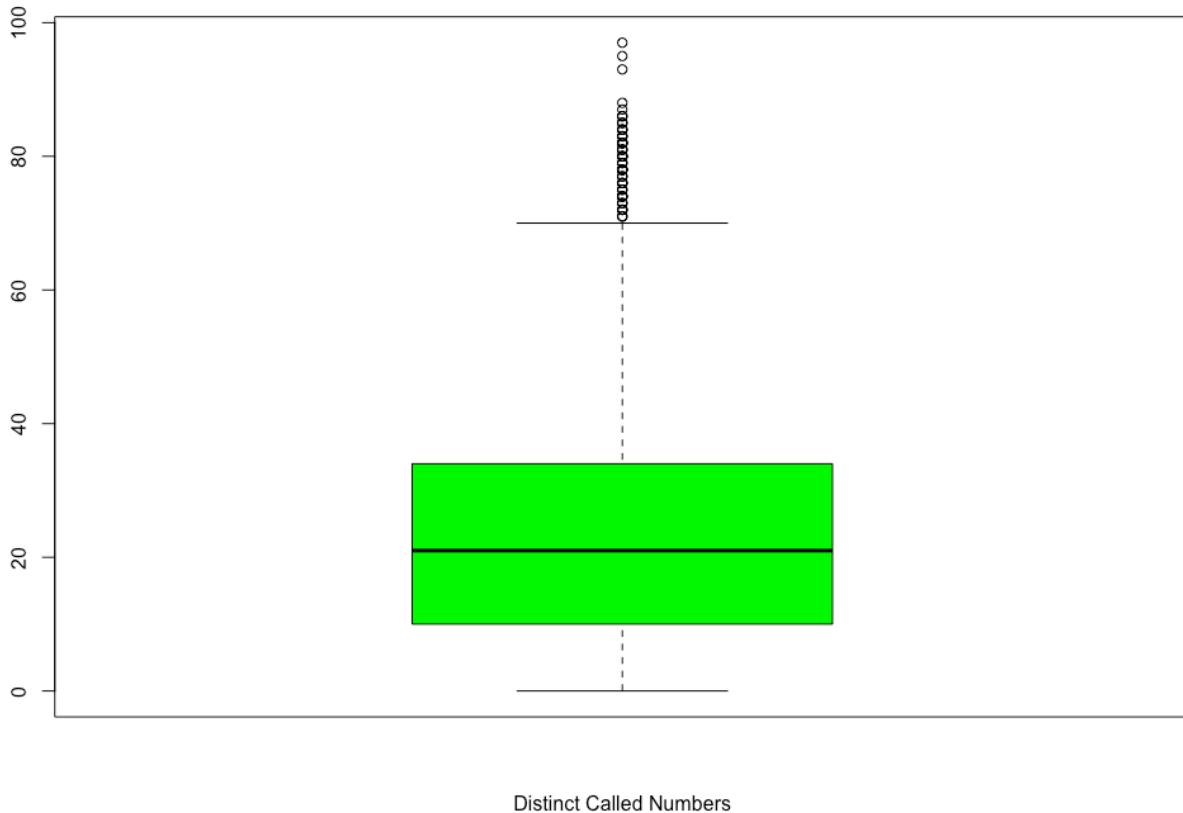
- **Media campionaria [1]**: La media del campo “Distinct Called Numbers” risulta pari a 23.51.
- **Mediana campionaria [2]**: La mediana è pari a 21.
- **Moda campionaria [3]**: La moda invece risulta essere 0.

La predominanza della moda pari a 0 indica una distribuzione unimodale, con un picco concentrato sul valore zero. Dalle misure di media e mediana possiamo desumere che la distribuzione sia asimmetrica positiva (sbilanciata a destra).

In altre parole:

- **Asimmetria verso destra**: La distribuzione presenta una "coda" estesa a destra a causa della presenza di valori più elevati di Distinct Called Numbers.
- **Concentrazione attorno allo zero**: La maggior parte degli utenti riporta pochi numeri chiamati, con valori di media spostati a destra rispetto alla mediana e alla moda.

Un boxplot della variabile *Distinct Called Numbers* permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.



*Figura 36 Boxplot Distinct Called Numbers*

Possiamo notare dall'immagine che abbiamo pochi outliers.

Utilizzando lo scarto interquartile, abbiamo rilevato i seguenti outliers: 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 93, 95, 97.

Tramite poi una funzione apposita confermiamo che il **primo quartile** è 1.00 mentre il **terzo quartile** è 12.00.

Inoltre, abbiamo il **minimo** uguale a 0.00 ed un **massimo** uguale a 97.00.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [q] dei numeri chiamati distinti dei fruitori.

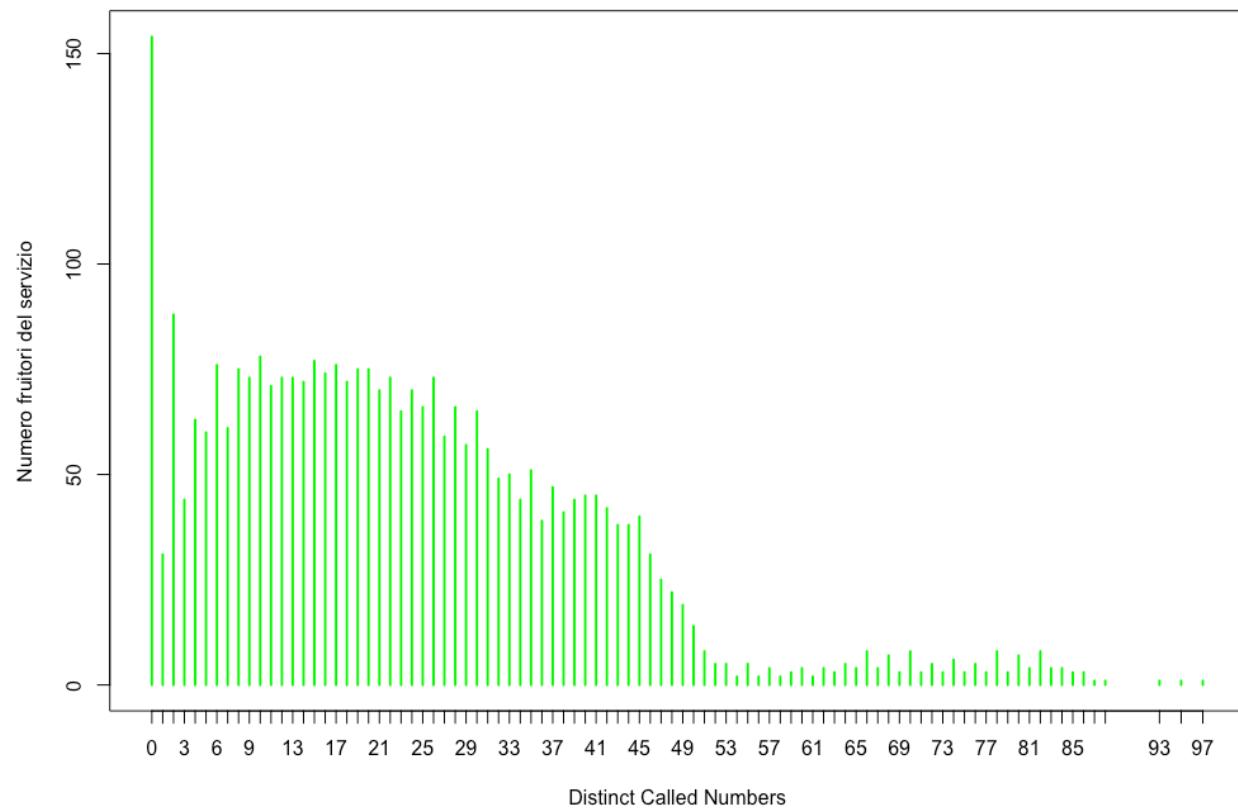


Figura 37 Istogramma Distinct Called Numbers

Un istogramma della variabile *Distinct Called Numbers* mostra la frequenza assoluta [9] dei numeri telefonati per ciascun valore osservato. Le ascisse rappresentano la quantità dei numeri telefonici contattati, mentre le ordinate indicano la quantità di utenti corrispondenti.

Il grafico conferma una distribuzione asimmetrica, con una concentrazione di osservazioni attorno a valori bassi e una coda verso destra.

Un'analisi delle frequenze relative [10] tramite **Funzione di Distribuzione Empirica (discreta)** evidenzia ulteriormente come una larga porzione degli utenti presenti valori prossimi allo zero.

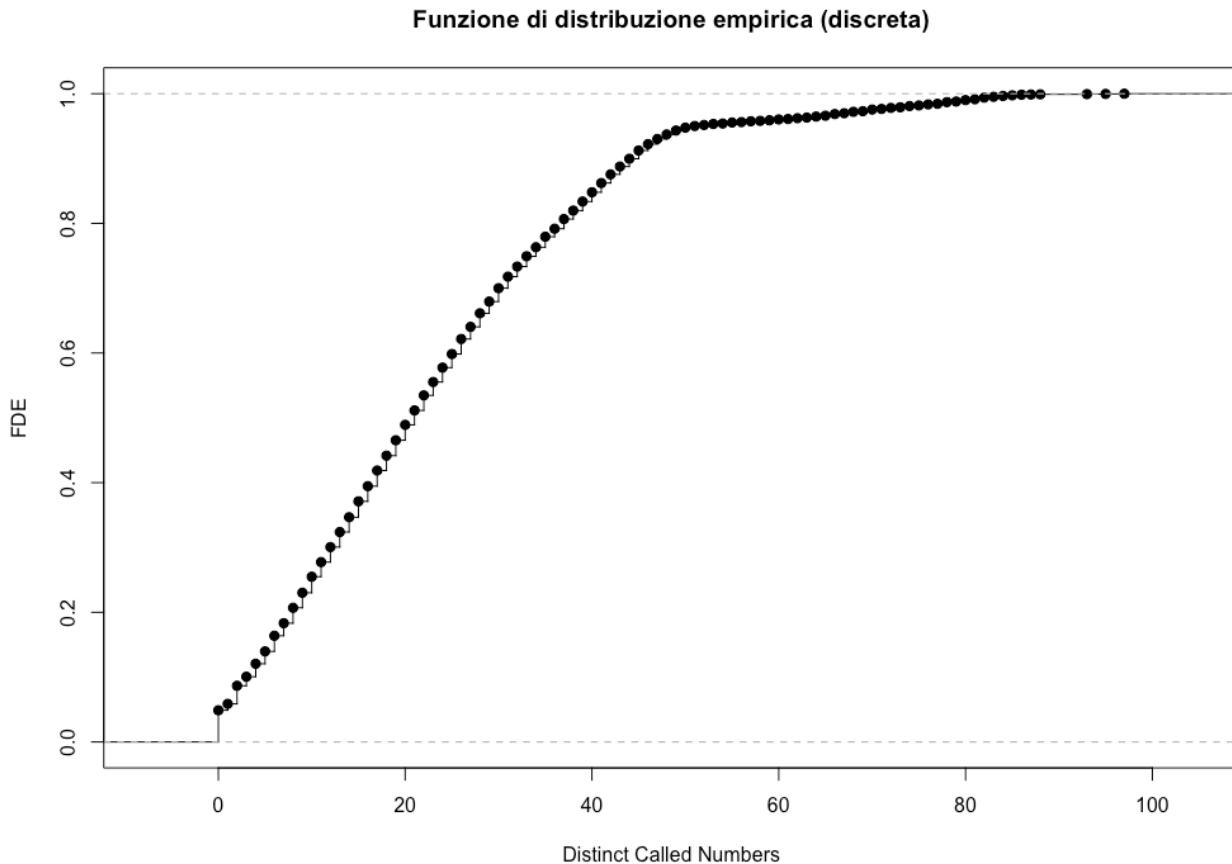


Figura 38 Funzione di distribuzione empirica (discreta) Distinct Called Numbers

L'analisi della **Funzione di Distribuzione Empirica (FDE)** conferma ulteriormente l'asimmetria menzionata precedentemente: mostra infatti una rapida crescita iniziale (data dalla frequenza elevata di valori prossimi allo zero), seguita da un incremento più graduale in corrispondenza dei valori più elevati.

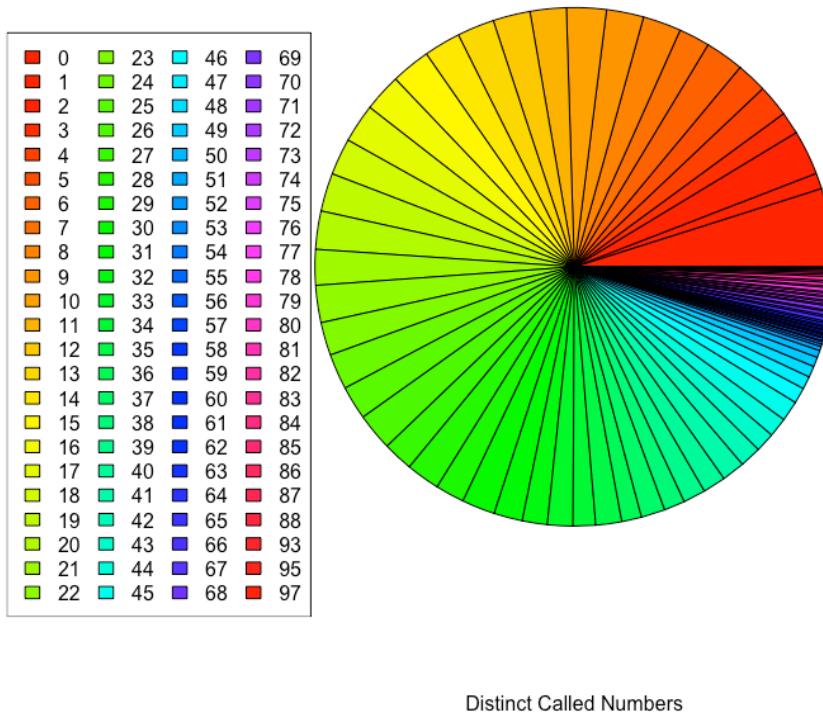


Figura 39 Diagramma a torta Distinct Call Numbers

Una rappresentazione tramite **diagramma a torta** illustra che la maggior parte degli utenti non ha contattato nessun numero.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza [4]: 296.44**
- **Deviazione standard [5]: 17.22**
- **Coefficiente di variazione [6]: 73.25 %**

L'elevato coefficiente di variazione indica una **forte dispersione dei valori** rispetto alla media, segnalando una significativa variabilità nella quantità dei numeri telefonati dagli utenti.

**Distribuzione di Frequenza tramite Diagramma di Pareto:** L'analisi tramite diagramma di Pareto permette di visualizzare come le frequenze assolute [9] siano associate alla frequenza relativa cumulativa, sottolineando la predominanza di utenti con pochi numeri distinti telefonati e il peso cumulativo degli utenti con più numeri contattati.

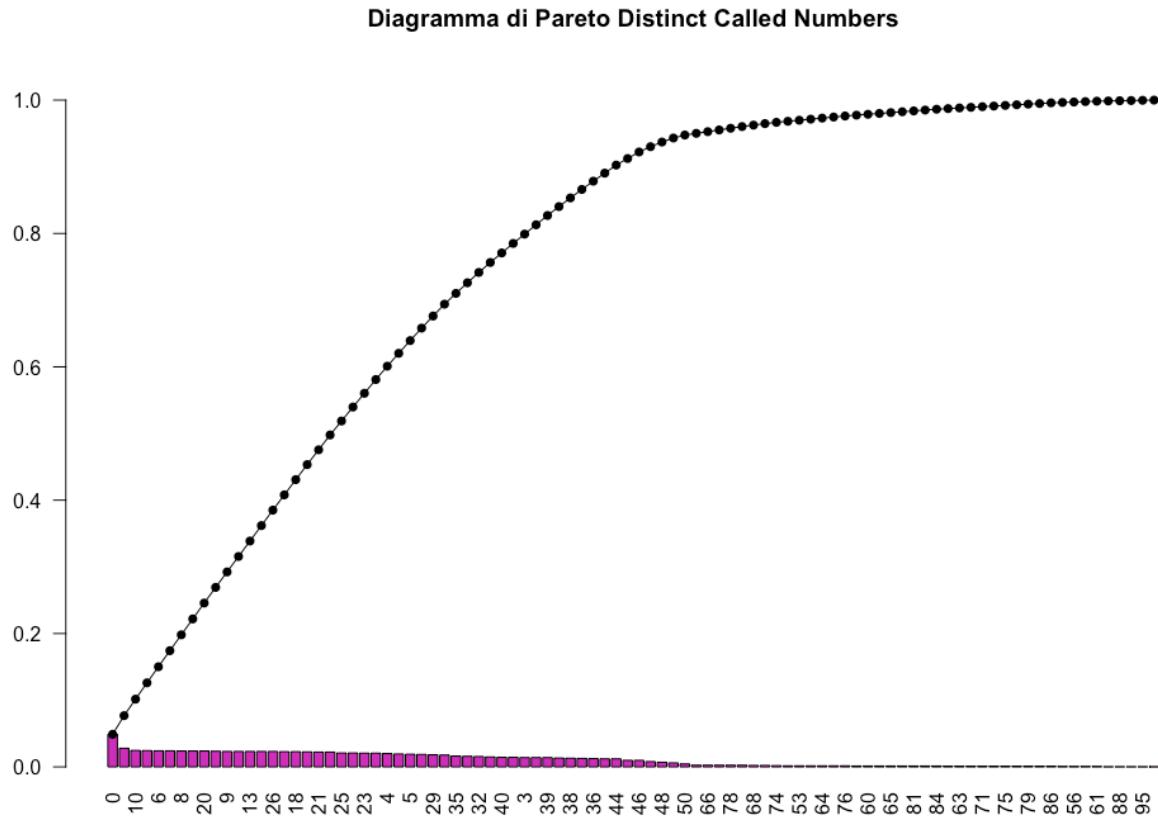
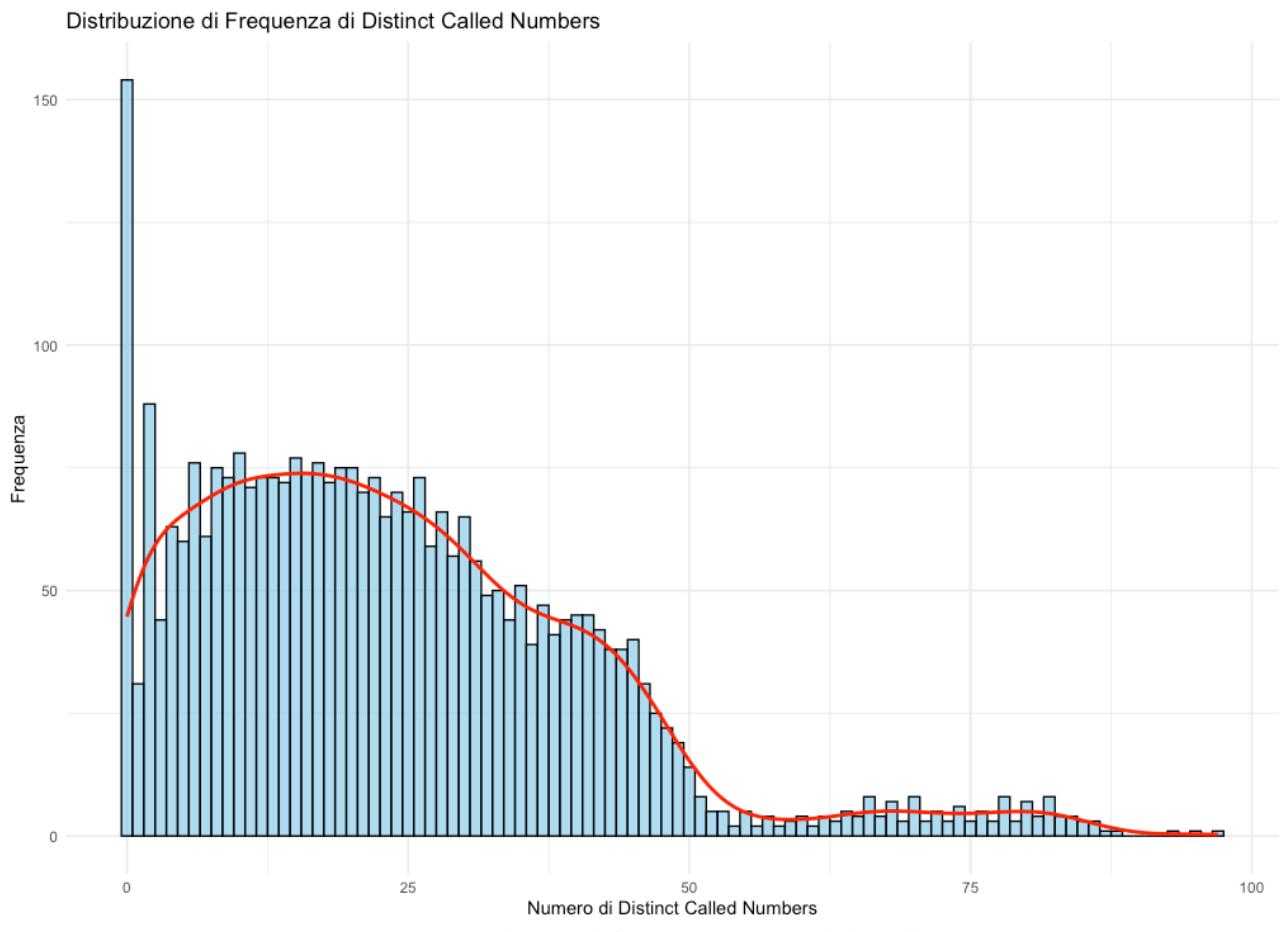


Figura 40 Diagramma di Pareto Distinct Called Numbers

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 1.03**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 4.36**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza dei numeri chiamati dai fruitori, confermando le caratteristiche sopra descritte.



## 2.1.10. Age Group

La feature “Age Group” variabile ordinale (quantitativa discreta) (1: fascia d’età più bassa, 5: fascia d’età più alta) che rappresenta l’età dei fruitori divisa in fasce. Per una completa caratterizzazione statistica della variabile, si procederà con un’analisi delle sue misure di centralità e dispersione, seguita da un’analisi grafica.

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Age Group” risulta pari a **2.83**.
- **Mediana campionaria [2]**: La mediana è pari a **3**.
- **Moda campionaria [3]**: La moda invece risulta essere **3**.

La predominanza della moda pari a 3 indica una distribuzione unimodale, con un picco concentrato sul valore 3. Dalle misure di media e mediana possiamo desumere che la distribuzione sia quasi simmetrica ma che presenta un’asimmetria a destra.

Un boxplot della variabile *Age Group* permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.

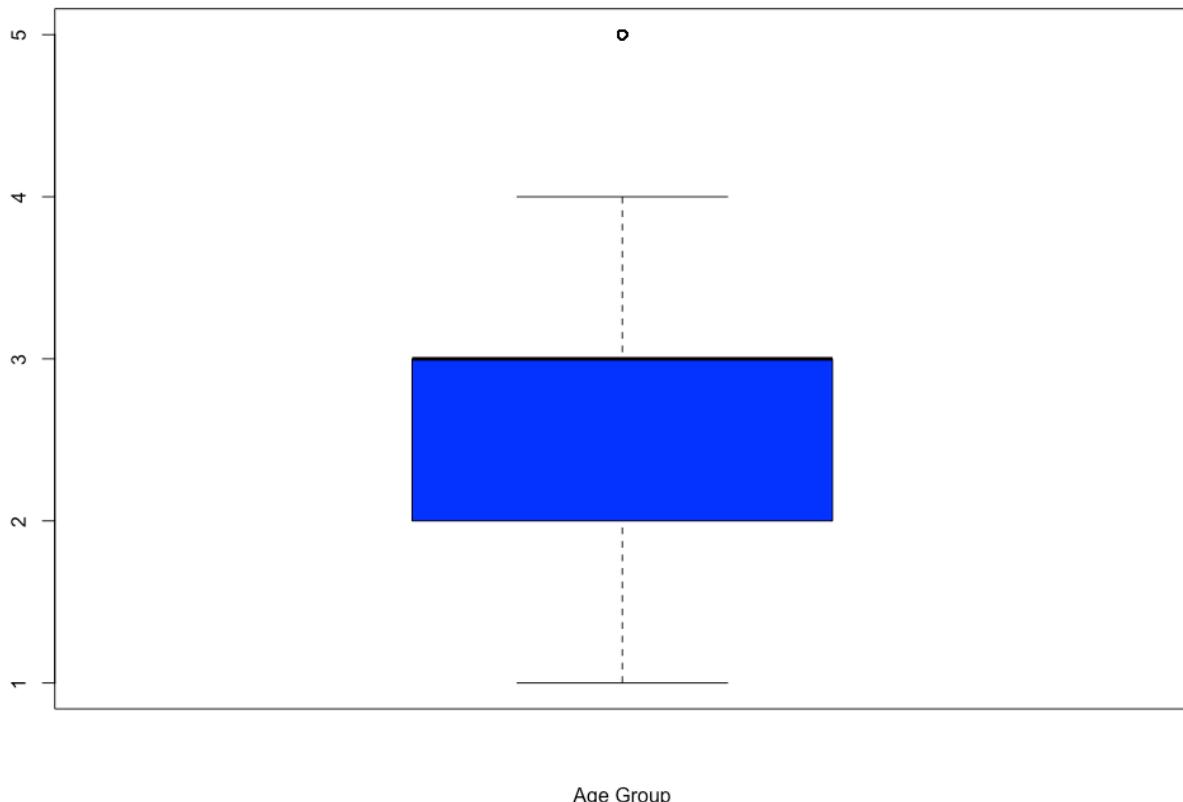


Figura 42 Boxplot Age Group

Possiamo notare dall'immagine che abbiamo un unico outlier e che la mediana coincide con il terzo quartile.

Utilizzando lo scarto interquartile, abbiamo rilevato l'outlier che è **5**.

Tramite poi una funzione apposita confermiamo che il **primo quartile** è **2.00** mentre il **terzo quartile** è **3.00**.

L'istogramma della variabile Age Group fornisce una visualizzazione delle frequenze assolute [9] delle fasce d'età dei fruitori. Le ascisse le fasce d'età, mentre le ordinate mostrano il numero di utenti corrispondenti.

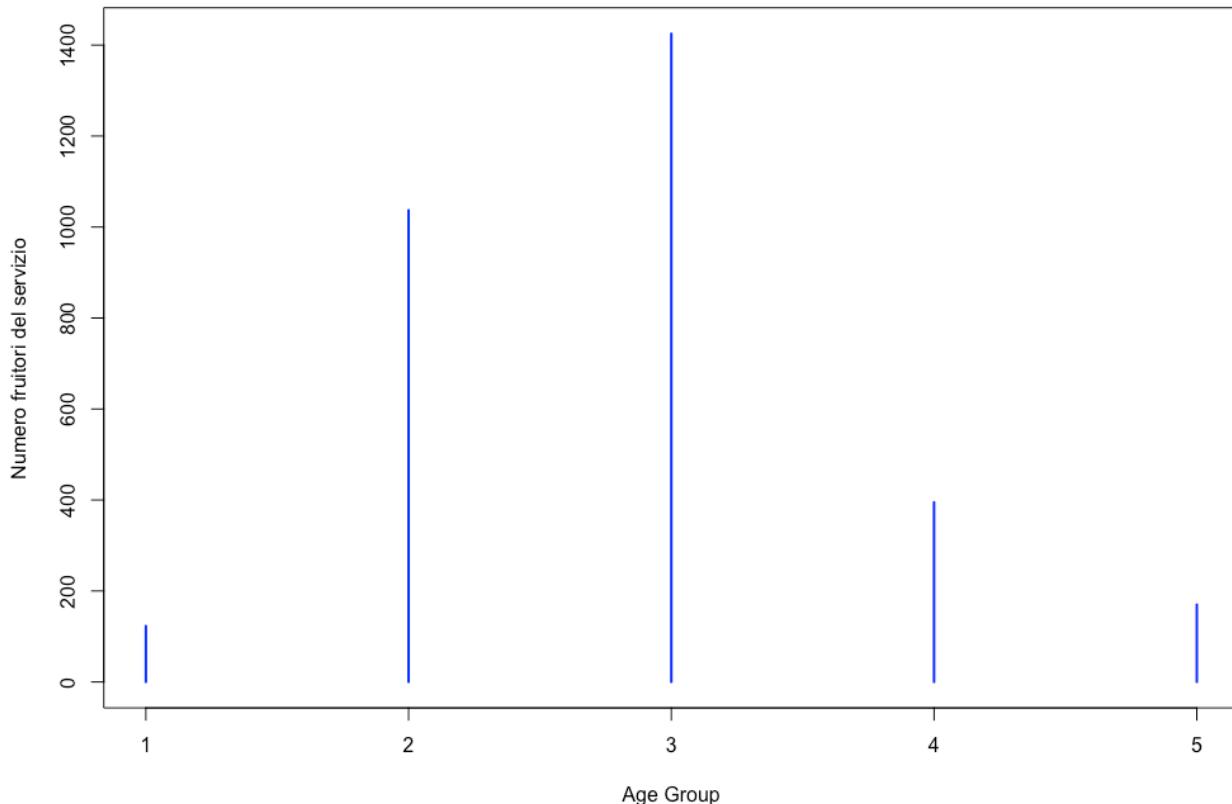


Figura 43 Istogramma Age Group

Questo grafico evidenzia l'asimmetria della distribuzione, con un'alta concentrazione di osservazioni al centro in corrispondenza del valore 3.

Un'analisi delle frequenze relative [10] tramite **Funzione di Distribuzione Empirica (discreta)** evidenzia ulteriormente come una larga porzione degli utenti presenti valori prossimi allo zero.

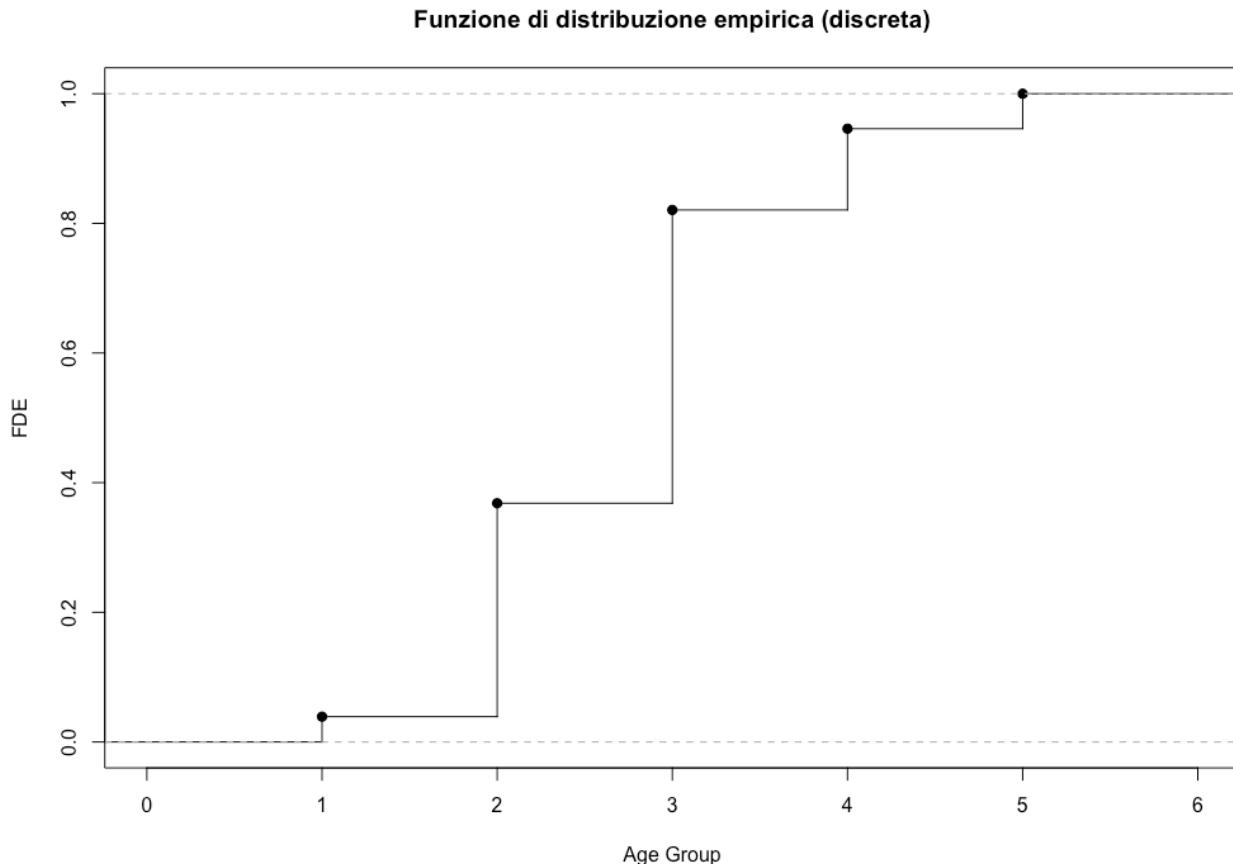


Figura 44 FDE Age Group

L'analisi della **Funzione di Distribuzione Empirica (FDE)** conferma ulteriormente le assunzioni fatte precedentemente.



*Figura 45 Diagramma a torta Charge Amount*

Una rappresentazione tramite **diagramma a torta** illustra che la maggior parte degli utenti ha una fascia d'età uguale a 3.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza [4]: 0.80**
- **Deviazione standard [5]: 0.89**
- **Coefficiente di variazione [6]: 31.58%**

Il coefficiente di variazione che si aggira intorno al 30% esprime che i valori non sono molto differenti tra loro.

**Distribuzione di Frequenza tramite Diagramma di Pareto:** L'analisi tramite diagramma di Pareto permette di visualizzare come le frequenze assolute [9] siano associate alla frequenza relativa cumulativa, sottolineando l'appartenenza ad una determinata fascia d'età e il peso cumulativo della fascia.

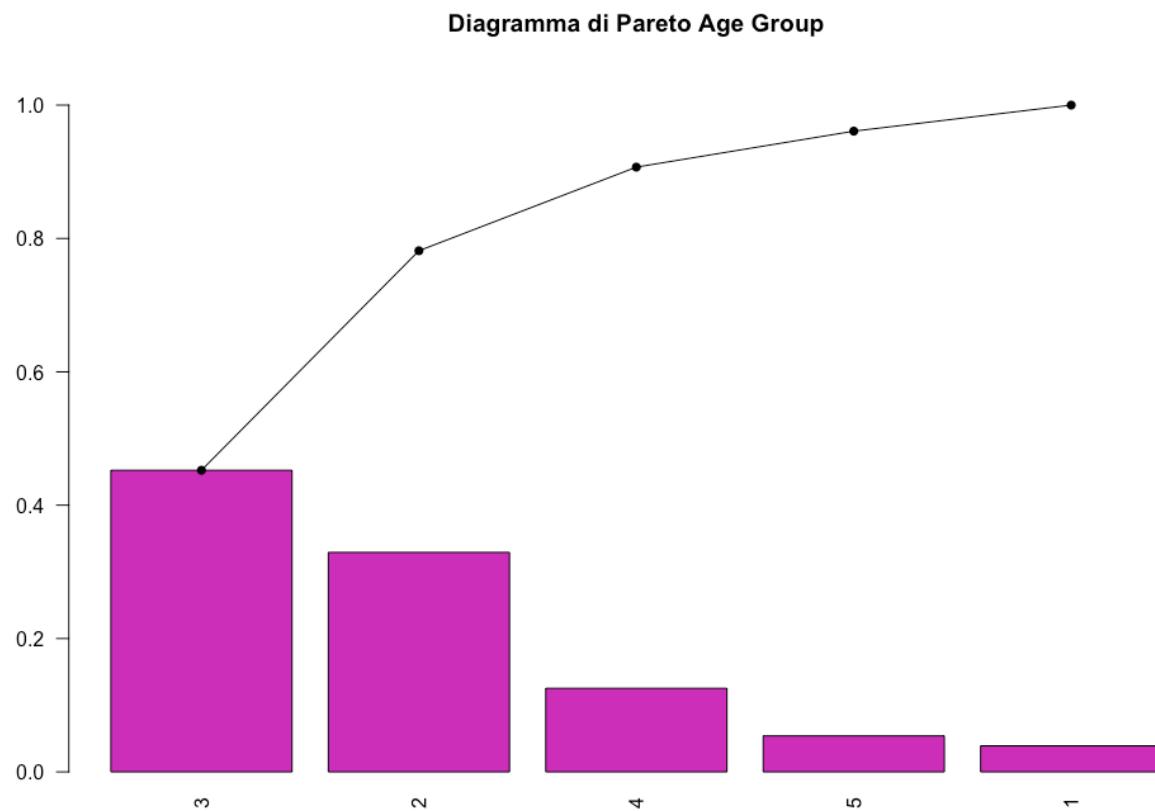


Figura 46 Diagramma di Pareto Age Group

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 0.47**, che conferma la leggera asimmetria verso destra.
- **Curtosi [8]: 3.20**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza delle fasce d'età, confermando le caratteristiche sopra descritte.

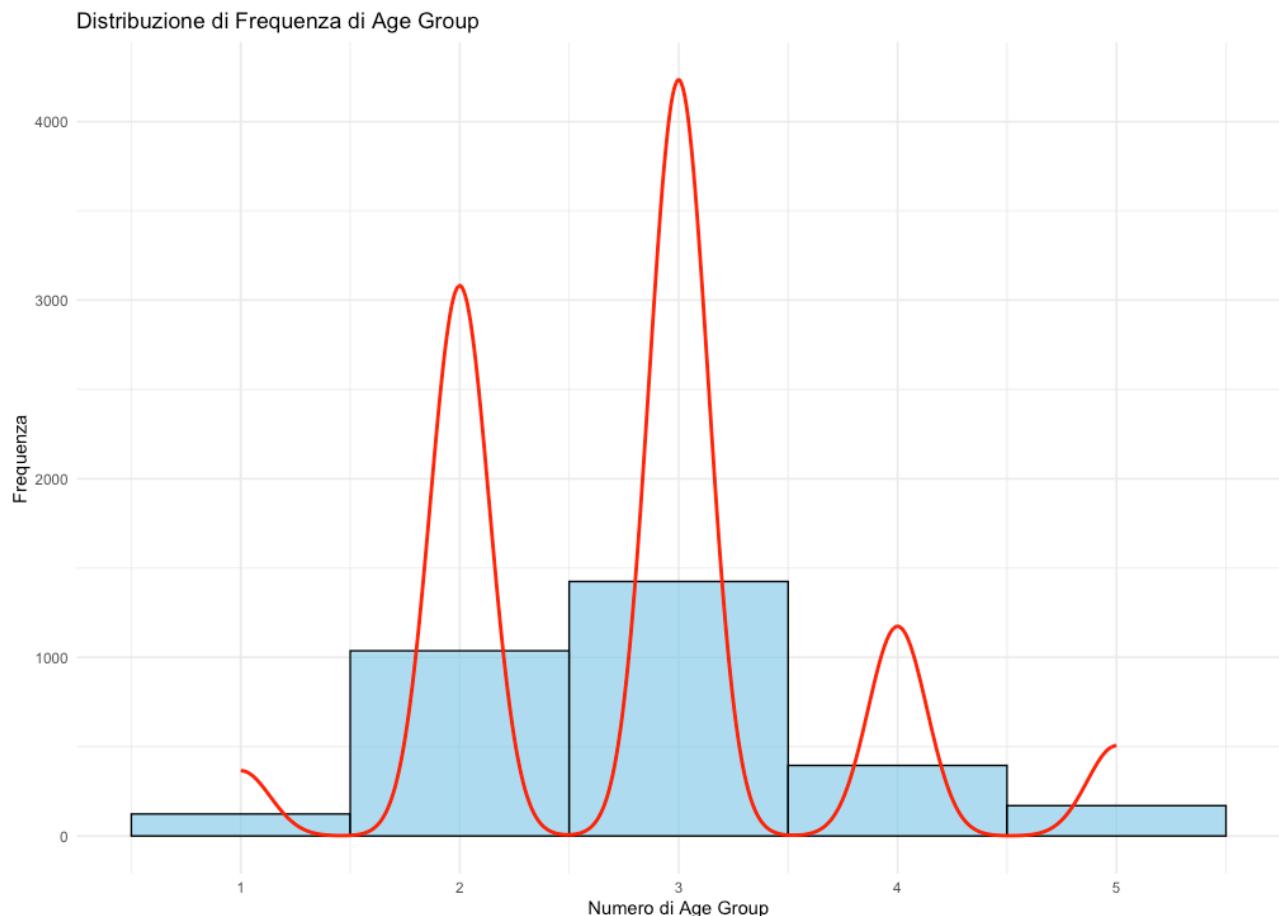


Figura 47 Distibuzione di frequenza Age Group

## 2.1.11. Tariff Plan

La feature “Tariff plan” rappresenta una variabile binaria che indica il tipo di piano tariffario scelto dal fruitore del servizio (**1: Pay to go, 2: Pagamento contrattuale**). Data la natura qualitativa della variabile, l’analisi procederà tramite lo studio delle frequenze e delle distribuzioni.

Analizziamo quindi le **frequenze assolute** [9] dei valori assunti dalla variabile Tariff plan:

Valore	Frequenza
1: Pay to go	<u>2905</u>
2: Pagamento contrattuale	<u>245</u>

Andiamo inoltre a vedere le **frequenze relative** [10]:

Valore	Frequenza
1: Pay to go	<u>0.92</u>
2: Pagamento contrattuale	<u>0.08</u>

Possiamo quindi notare che il **92.22%** dei fruitori usufruisce di un servizio Pay To go. Mentre il restante **7.77%** usa un pagamento contrattuale.

Per avere un’idea più chiara possiamo osservare il diagramma a torta e il diagramma rappresentante la funzione di distribuzione empirica (discreta) sottostanti:

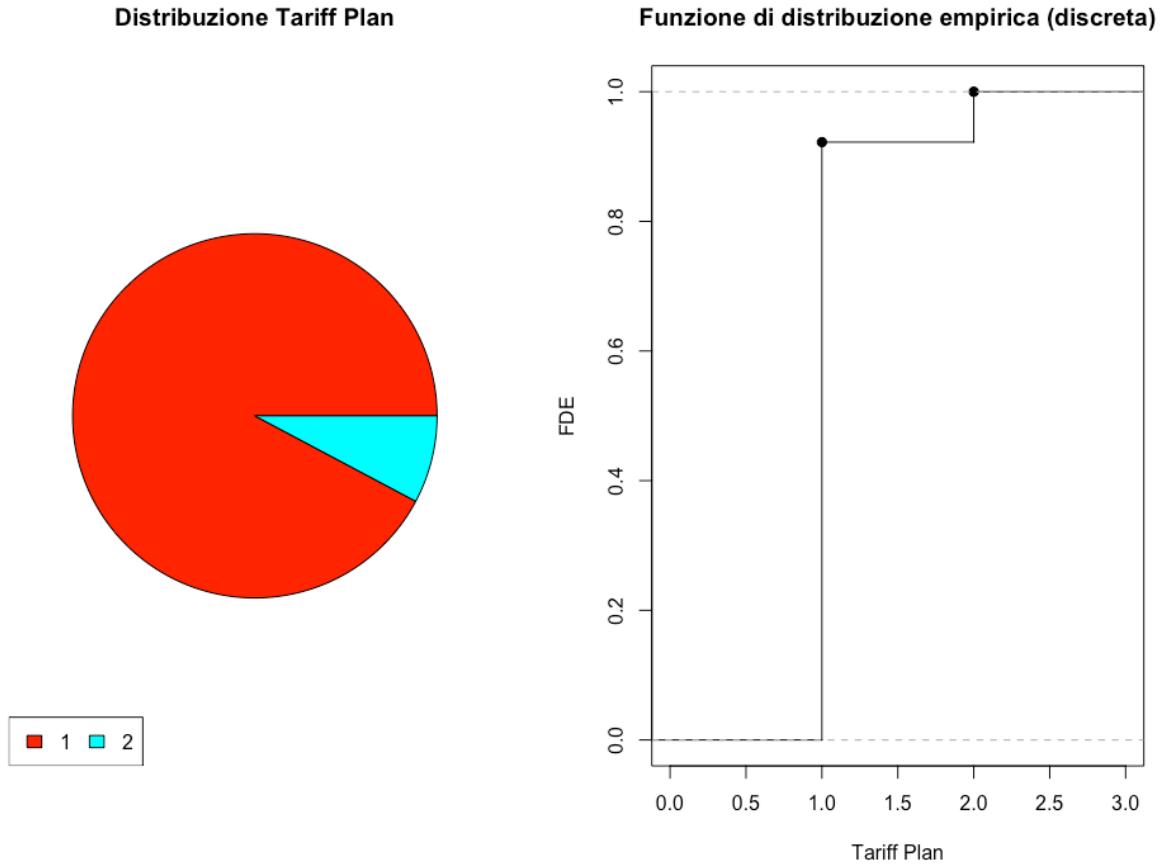


Figura 48 Diagramma a torta e FDE Tariff plan

## 2.1.12. Status

La feature “Status” rappresenta una variabile binaria che indica se il servizio è ancora attivo (**1: Attivo, 2: Non attivo**).

Data la natura qualitativa della variabile, l’analisi procederà tramite lo studio delle frequenze e delle distribuzioni.

Analizziamo quindi le **frequenze assolute** [9] dei valori assunti dalla variabile

Status:

Valore	Frequenza
1: Attivo	<u>2368</u>
2: Non Attivo	<u>782</u>

Andiamo inoltre a vedere le **frequenze relative** [10]:

Valore	Frequenza
1: Attivo	<u>0.75</u>
2: Non Attivo	<u>0.25</u>

Possiamo quindi notare che il 75.17% dei servizi rimasti attivi.

Mentre il restante 24.83% dei servizi è disattivato.

Per avere un’idea più chiara possiamo osservare il diagramma a torta e il diagramma rappresentante la funzione di distribuzione empirica (discreta) sottostanti:

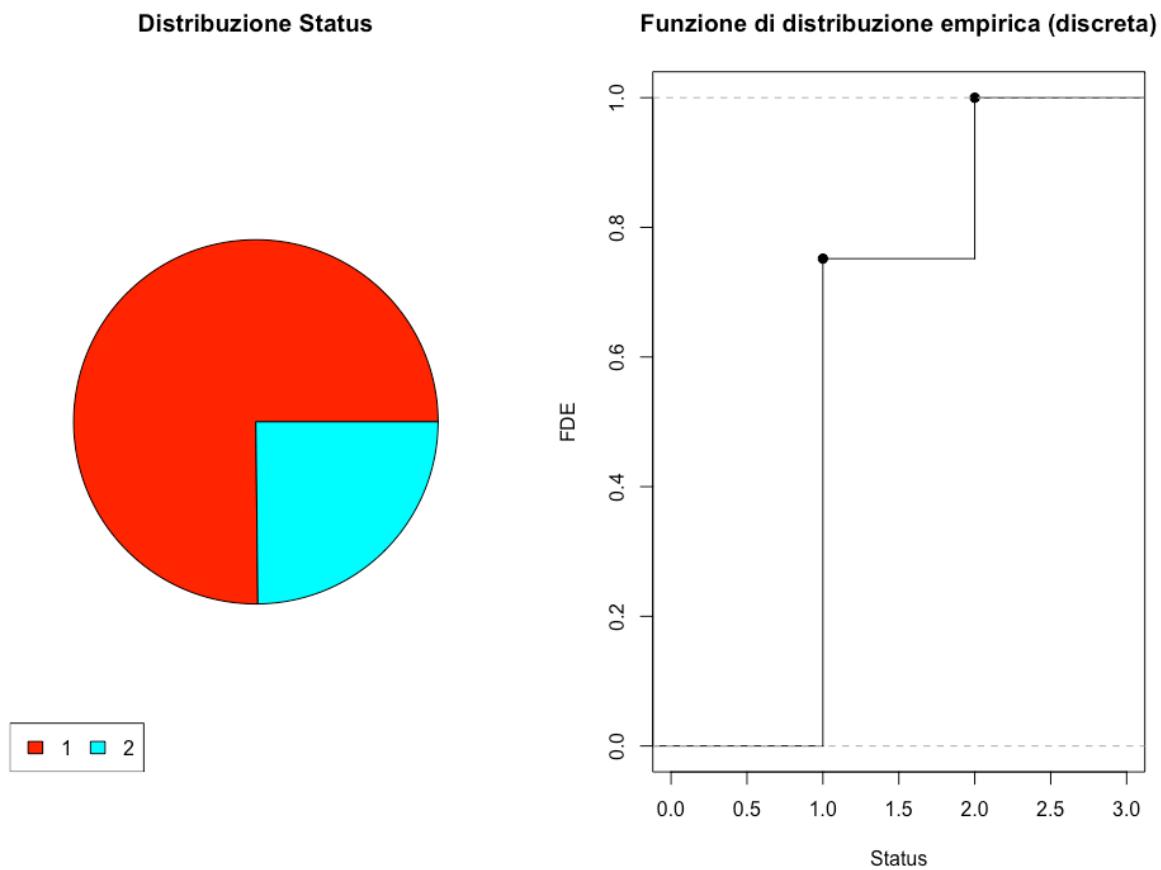


Figura 49 Diagramma a torta e FDE Status

### 2.1.13. Age (Feature Rimossa)

La feature “Age” è una variabile quantitativa discreta espressa in numeri interi, rappresentante l’età di ogni fruitore del servizio. Per una completa caratterizzazione statistica della variabile, si procederà con un’analisi delle sue misure di centralità e dispersione, seguita da un’analisi grafica.

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Age” risulta pari a **30.99**.
- **Mediana campionaria [2]**: La mediana è pari a **30**.
- **Moda campionaria [3]**: La moda invece risulta essere **30**.

La predominanza della moda pari a 30 indica una distribuzione unimodale, con un picco concentrato sul valore 30. Dalle misure di media e mediana possiamo desumere che la distribuzione sia quasi del tutto simmetrica ma con una piccola asimmetria a destra.

Un boxplot della variabile *Age* permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.

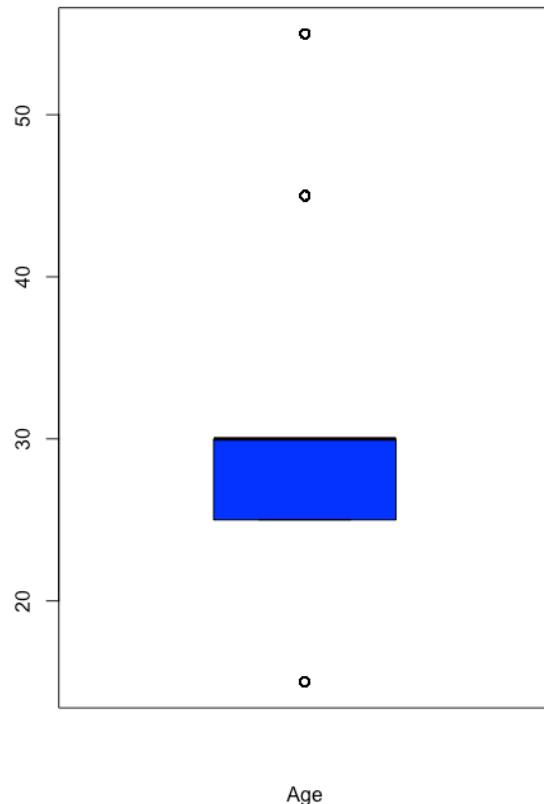


Figura 50 Boxplot Age

Possiamo notare dall’immagine che abbiamo alcuni outliers, alcuni sono superiori alla mediana e altri inferiori.

Utilizzando lo scarto interquartile, abbiamo rilevato i seguenti outliers: **15, 45, 55**.

Tramite poi una funzione apposita confermiamo che il **primo quartile** è **25** mentre il **terzo quartile** è **30**.

Inoltre, abbiamo il **minimo** uguale a **15** ed un **massimo** uguale a **55**.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] delle età dei fruitori.

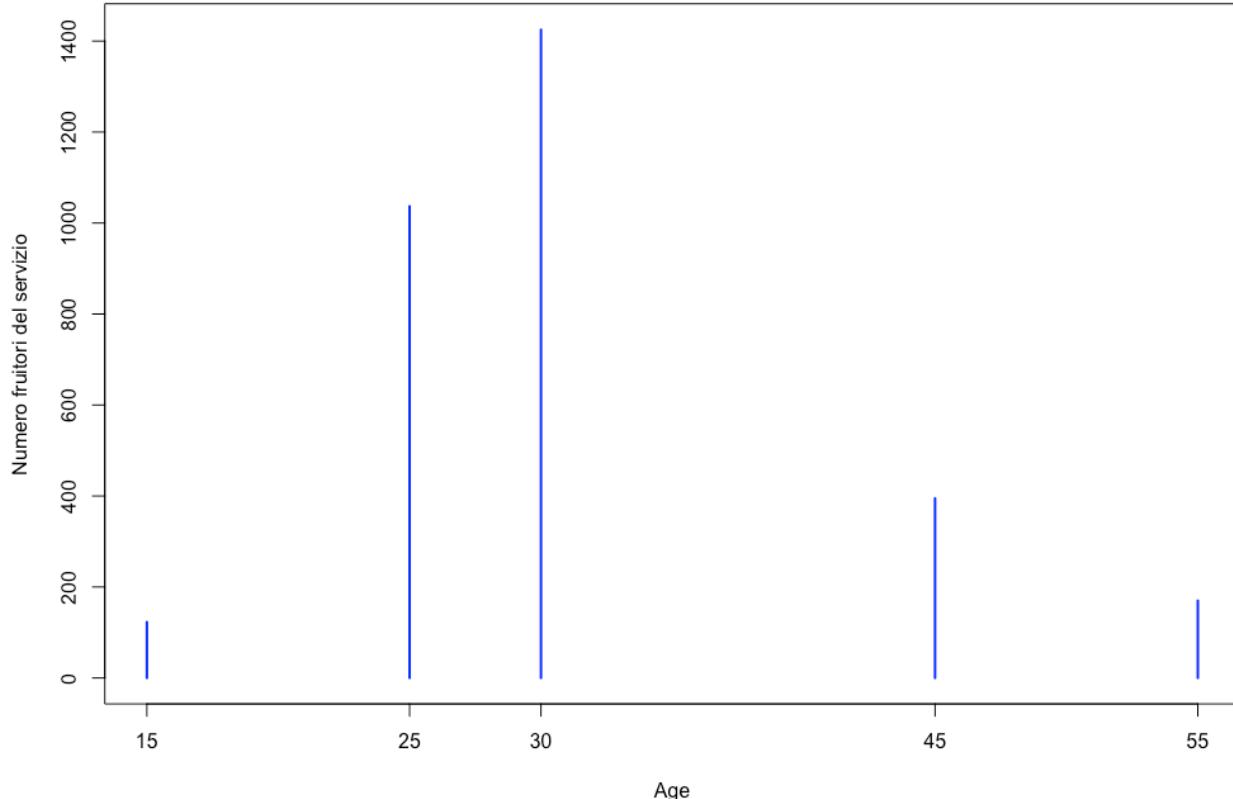


Figura 51 Istogramma Age

Un istogramma della variabile *Age* mostra la frequenza assoluta [9] delle età di tutti i fruitori. Le ascisse rappresentano le età, mentre le ordinate indicano la quantità di utenti corrispondenti.

La cosa interessante è che su tutto il campione abbiamo solo utenti/fruitori di età ben precise, ovvero: **15, 25, 30, 45, 55**.

Il grafico conferma una distribuzione asimmetrica, con una concentrazione di osservazioni attorno al valore **30** ed una asimmetria a destra.

Un'analisi delle frequenze relative [10] tramite **Funzione di Distribuzione Empirica (discreta)** evidenzia ulteriormente come una larga porzione degli utenti presenti valori prossimi al valore 30.

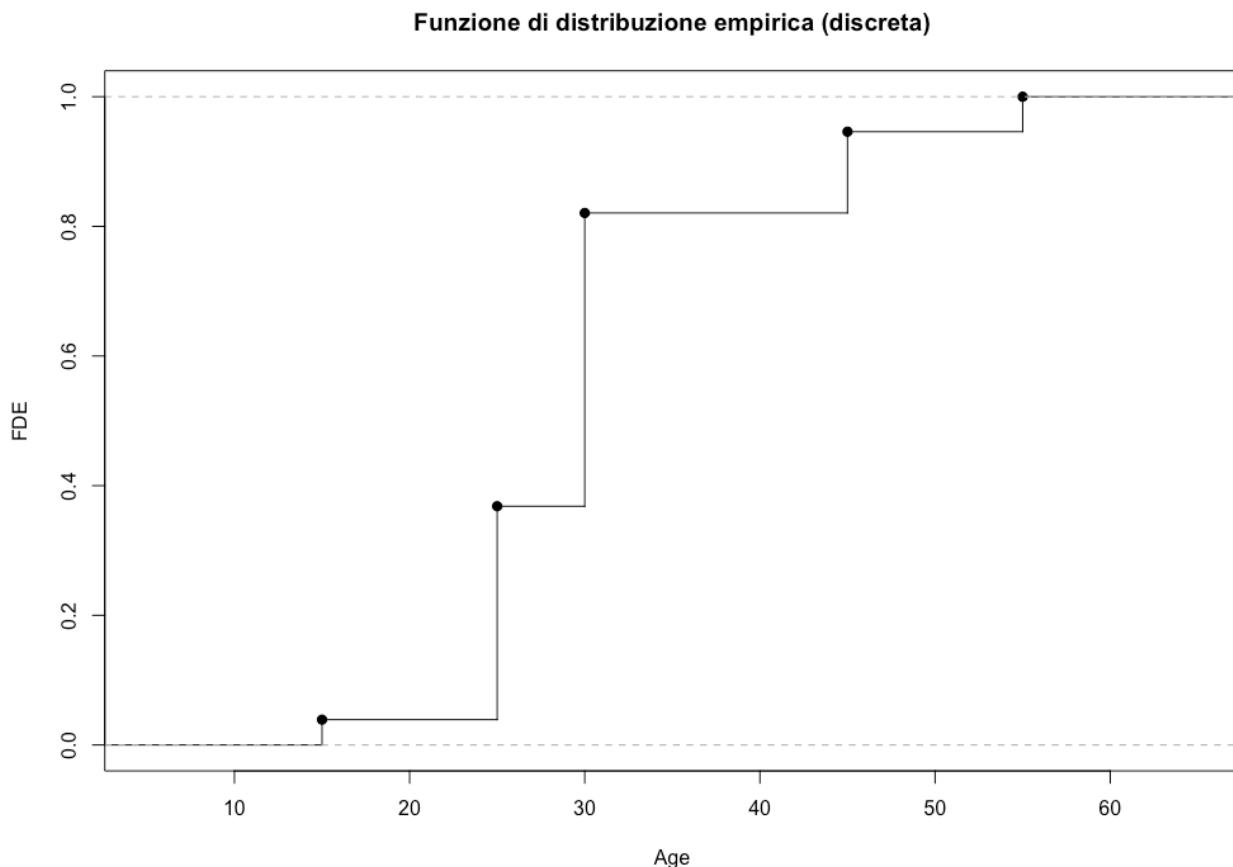
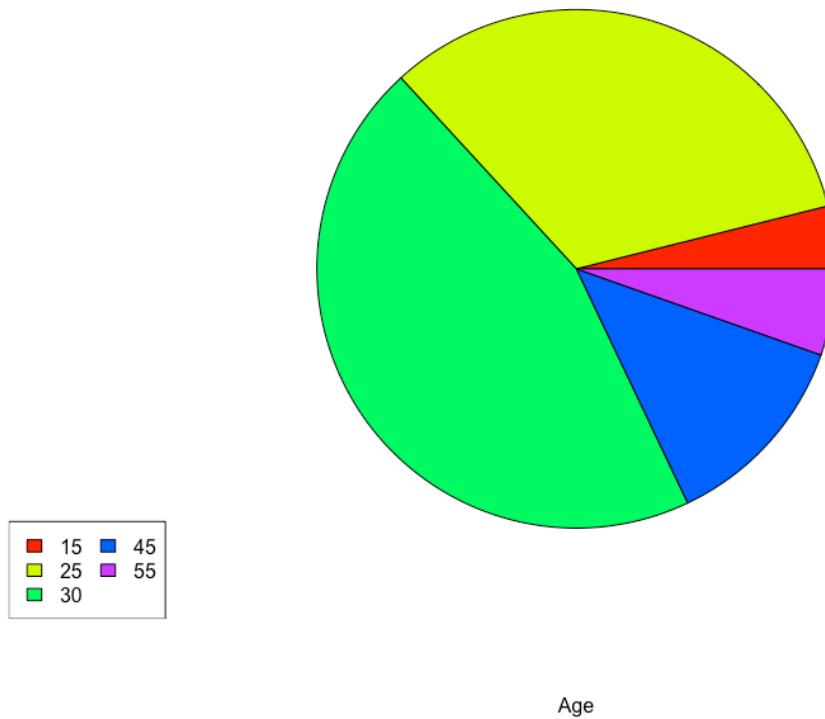


Figura 52 Funzione di distribuzione empirica (discreta) Age

L'analisi della **Funzione di Distribuzione Empirica (FDE)** conferma ulteriormente l'asimmetria menzionata precedentemente: mostra infatti la maggioranza dei valori che converge al valore 30.



*Figura 53 Diagramma a torta Age*

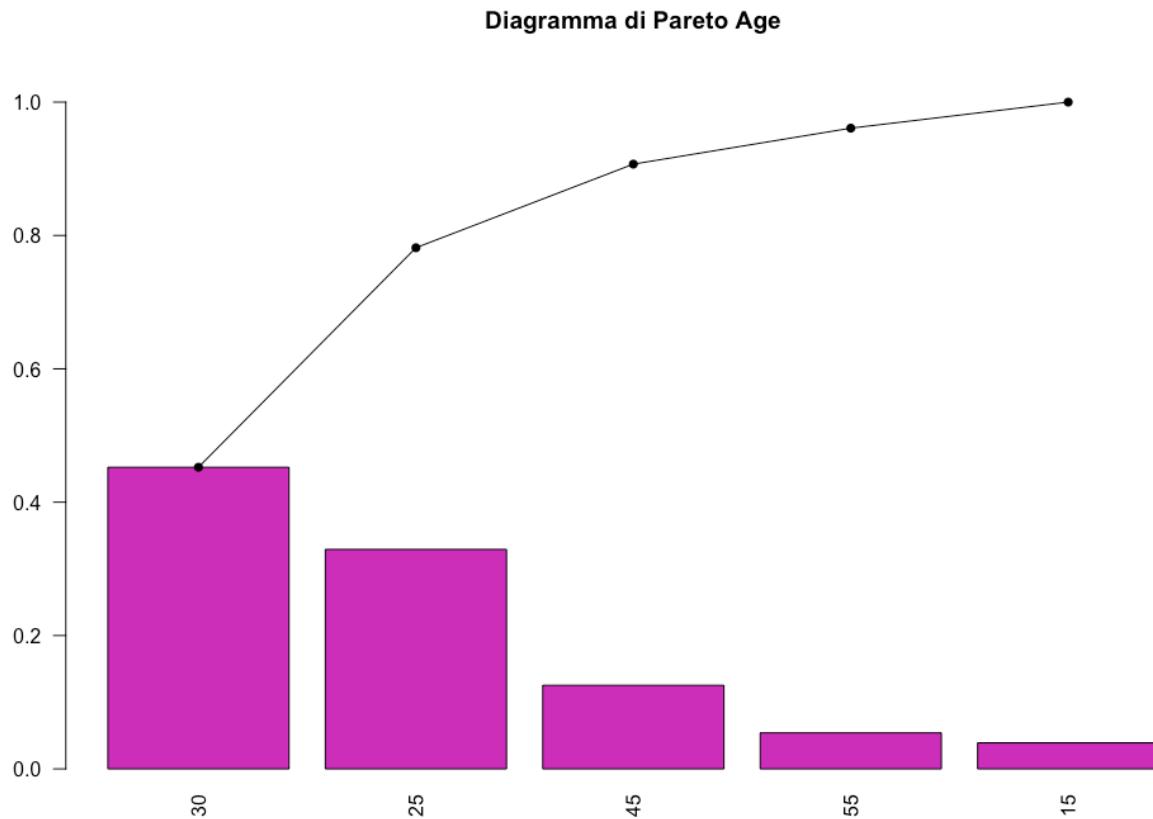
Una rappresentazione tramite **diagramma a torta** illustra che la maggior parte degli utenti ha 30 anni.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza [4]: 77.99**
- **Deviazione standard [5]: 8.83**
- **Coefficiente di variazione [6]: 28.49%**

Il basso coefficiente di variazione ci fa comprendere che tra tutti gli utenti non c'è un eccessivo discostamento riguardo l'età.

**Distribuzione di Frequenza tramite Diagramma di Pareto:** L'analisi tramite diagramma di Pareto permette di visualizzare come le frequenze assolute [9] siano associate alla frequenza relativa cumulativa, sottolineando la predominanza di utenti che hanno 30 anni e il peso cumulativo delle età.



*Figura 54 Diagramma di Pareto Age*

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 1.25**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 4.23**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza delle età, confermando le caratteristiche sopra descritte.

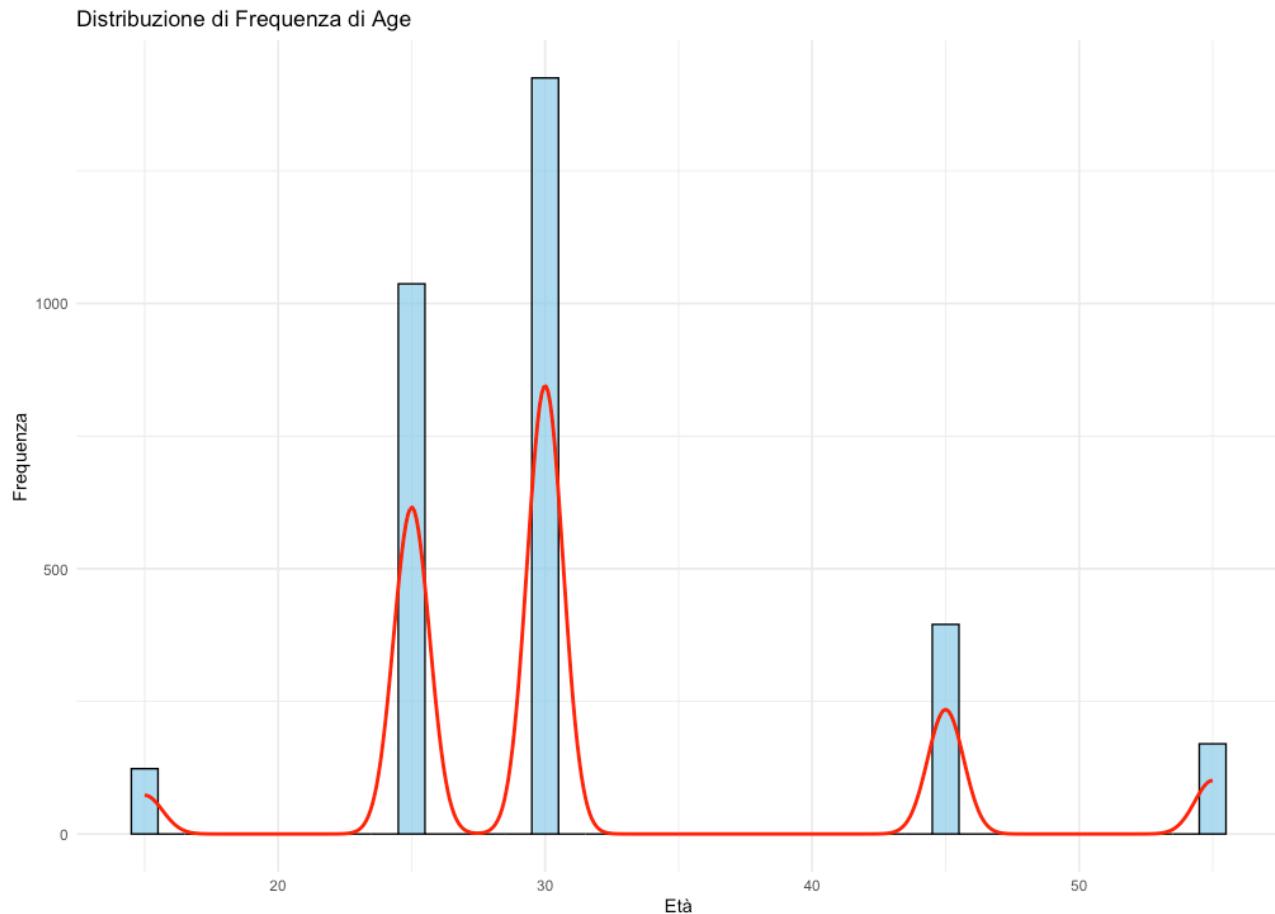


Figura 55 Distribuzione di frequenza Age

**In ogni caso, poiché l'età dei fruitori dei servizi può essere desunta dalla fascia d'età, provvederemo a rimuovere la variabile *Age* dal dataset, in quanto rappresenta un'informazione ridondante dato che c'è già la presenza della feature *Age Group* la quale già suddivide le età in 5 insiemi.**

## 2.1.14. Churn

La feature “[Churn](#)” rappresenta una variabile binaria che indica se il cliente ha abbandonato o meno il servizio (**0: Non abbandonato il servizio, 1: Abbandonato il servizio**).

Data la natura qualitativa della variabile, l’analisi procederà tramite lo studio delle frequenze e delle distribuzioni.

Analizziamo quindi le **frequenze assolute** [9] dei valori assunti dalla variabile

Status:

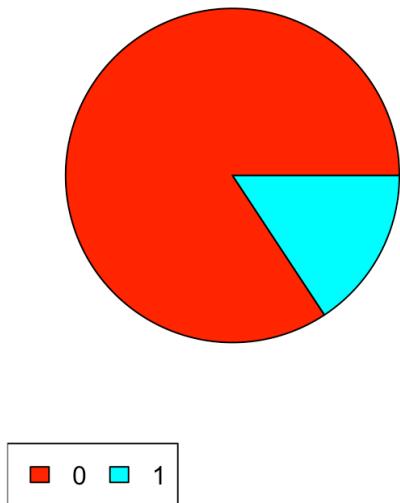
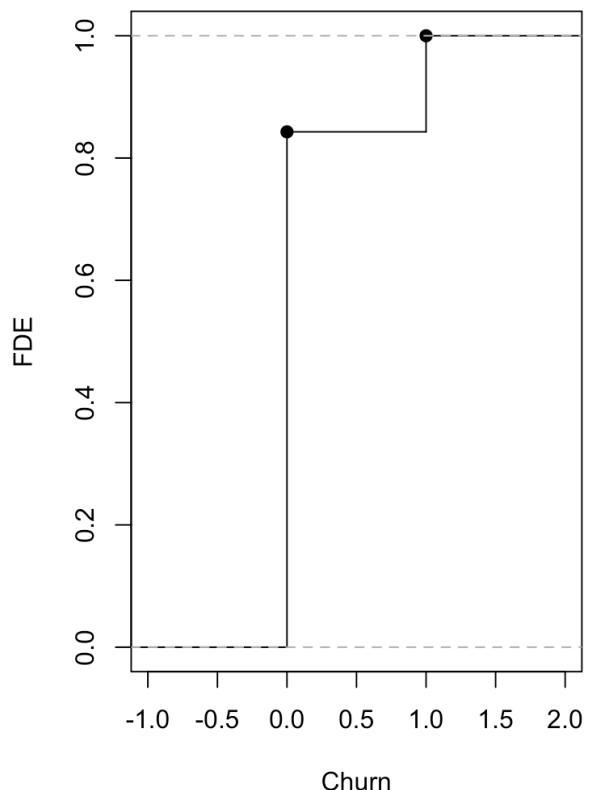
Valore	Frequenza
0: Non abbandonato	<u>2665</u>
2: Abbandonato	<u>485</u>

Andiamo inoltre a vedere le **frequenze relative** [10]:

Valore	Frequenza
0: Non abbandonato	<u>0.84</u>
2: Abbandonato	<u>0.16</u>

Possiamo quindi notare che l’**84.28%** dei servizi non è stato annullato ed è ancora attivo. Mentre il restante **15,72%** dei servizi è stato disattivato.

Per avere un’idea più chiara possiamo osservare il diagramma a torta e il diagramma rappresentante la funzione di distribuzione empirica (discreta) sottostanti:

**Distribuzione Churn****Funzione di distribuzione empirica (discreta)***Figura 56 Diagramma a torta e FDE Churn*

## 2.1.15. Customer Value

La feature “Customer value” è una variabile quantitativa discreta espressa in numeri interi, rappresentante la valutazione per ogni fruitore del servizio. Per una completa caratterizzazione statistica della variabile, si procederà con un’analisi delle sue misure di centralità e dispersione, seguita da un’analisi grafica.

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Customer Valuerisulta pari a 70.97.
- **Mediana campionaria [2]**: La mediana è pari a 228.48.
- **Moda campionaria [3]**: La moda invece risulta essere 0.

La predominanza della moda pari a 0 indica una distribuzione unimodale, con un picco concentrato sul valore zero. Dalle misure di media e mediana possiamo desumere che la distribuzione sia asimmetrica positiva (sbilanciata a destra).

In altre parole:

- **Asimmetria verso destra**: La distribuzione presenta una "coda" estesa a destra a causa della presenza di valori più elevati di Customer value.
- **Concentrazione attorno allo zero**: La maggior parte degli utenti riporta valori bassi/assenti, con valori di media spostati a destra rispetto alla mediana e alla moda.

Un boxplot della variabile *Customer Value* permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.

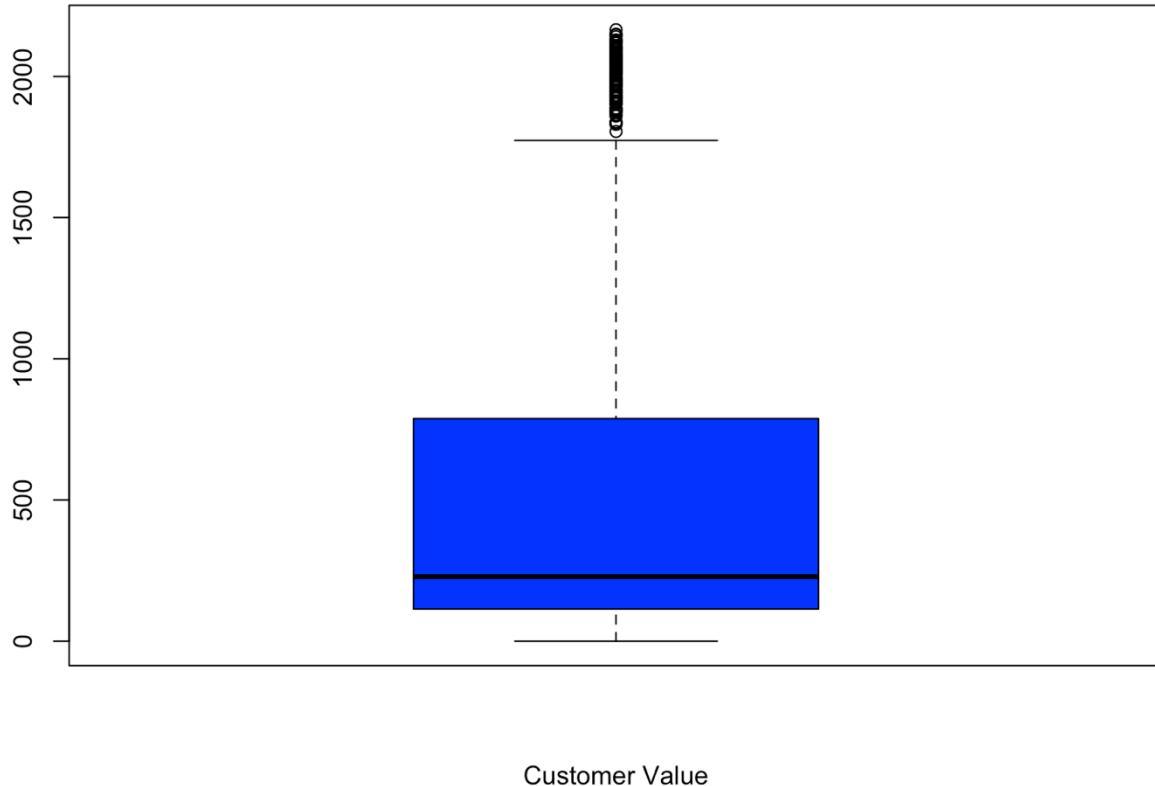


Figura 57 Boxplot Customer Value

Possiamo notare dall'immagine che abbiamo moltissimi outliers.

Utilizzando lo scarto interquartile, abbiamo rilevato i seguenti outliers, che per semplicità verranno espressi sottoforma di intervallo **[1805.04, 2165.28]**.

Tramite poi una funzione apposita confermiamo che il **primo quartile** è **113.8** mentre il **terzo quartile** è **788.4**.

Inoltre, abbiamo il **minimo** uguale a **0.0** ed un **massimo** uguale a **2165.3**.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] delle valutazioni dei fruitori.

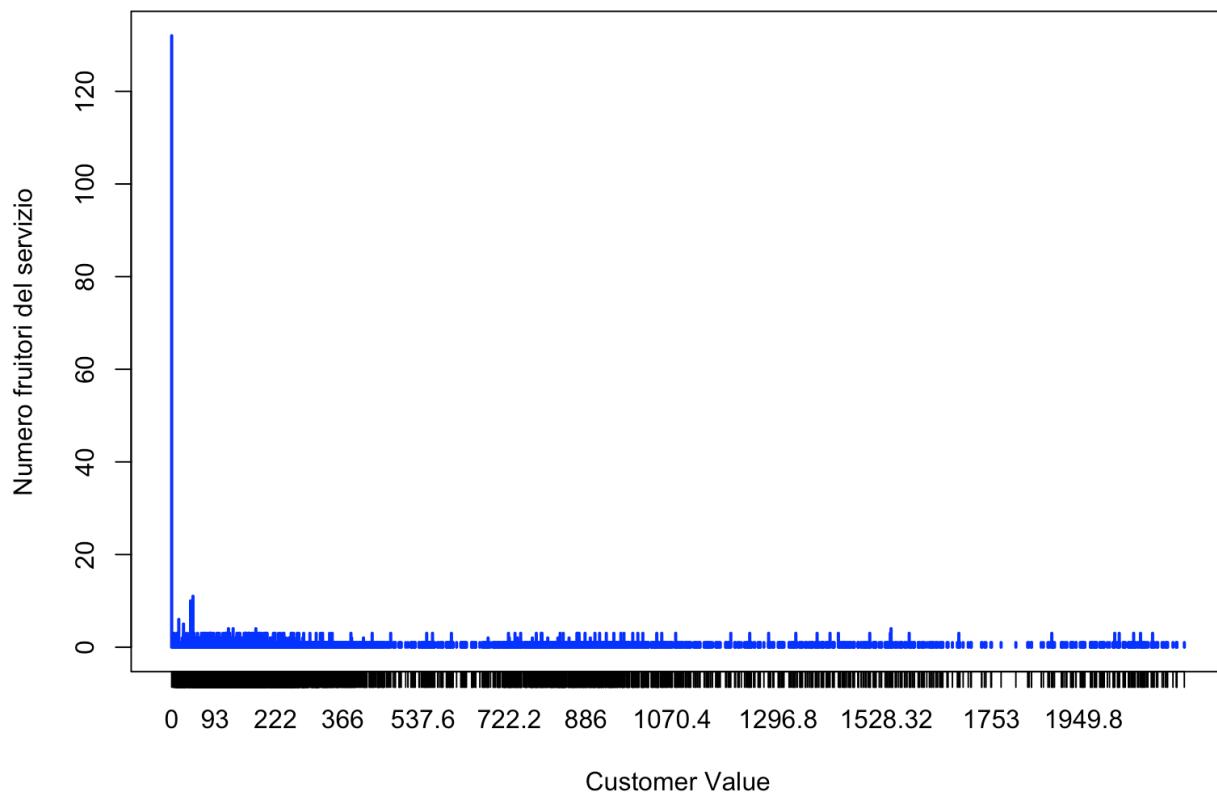


Figura 58 Istogramma Customer Value

Un istogramma della variabile *Customer Value* mostra la frequenza assoluta [9] dei valori associati ai fruitori del servizio. Le ascisse rappresentano il valore associato, mentre le ordinate indicano la quantità di utenti corrispondenti.  
Il grafico conferma una distribuzione asimmetrica, con una concentrazione di osservazioni attorno a valori bassi e una coda verso destra.

Un'analisi delle frequenze relative [10] tramite **Funzione di Distribuzione Empirica (discreta)** evidenzia ulteriormente come una larga porzione degli utenti presenti valori prossimi allo zero.

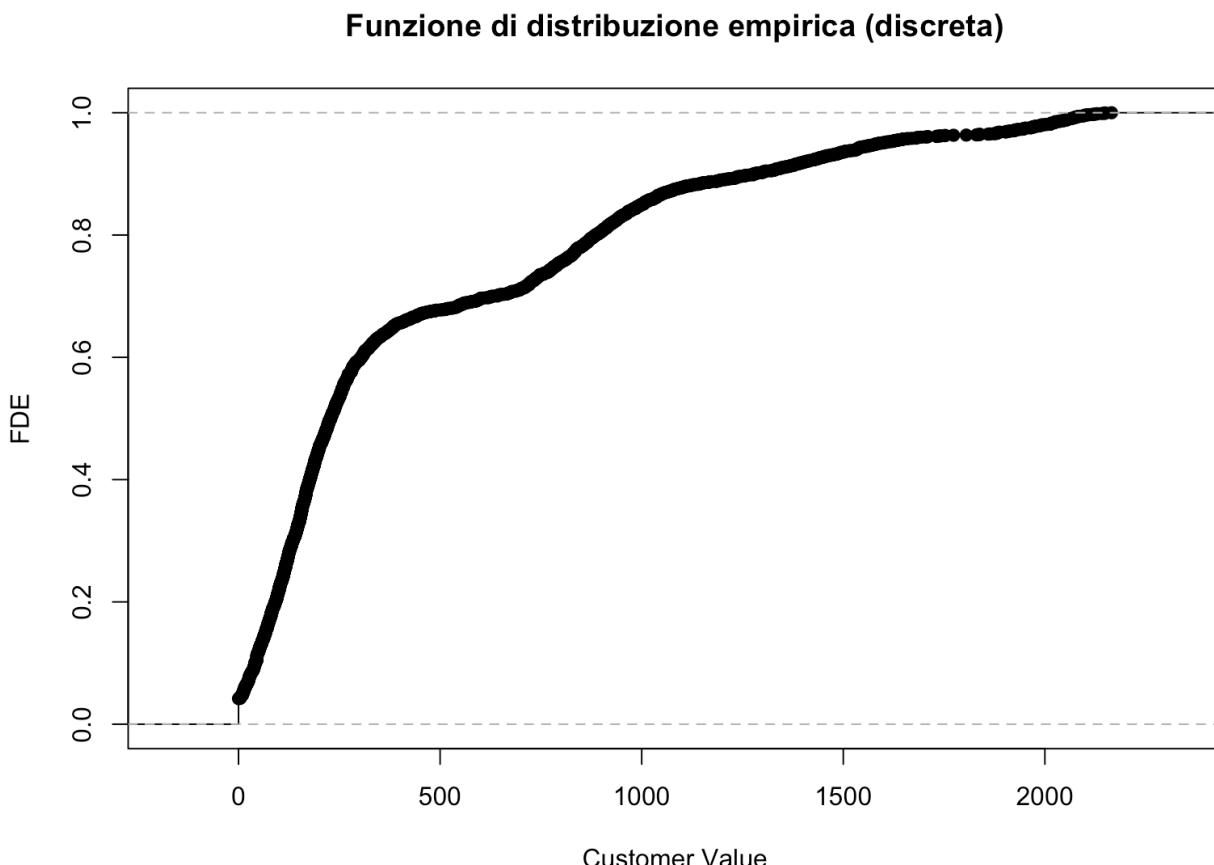


Figura 59 Funzione di distribuzione empirica (discreta) Customer Value

L'analisi della **Funzione di Distribuzione Empirica (FDE)** conferma ulteriormente l'asimmetria menzionata precedentemente: mostra infatti una rapida crescita iniziale (data dalla frequenza elevata di valori prossimi allo zero), seguita da un incremento più graduale in corrispondenza dei valori più elevati.

Inoltre, possiamo notare l'alta densità di valori e quindi di diversità di valori assunti dalla variabile.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza [4]: 267305**
- **Deviazione standard [5]: 517.02**
- **Coefficiente di variazione [6]: 109.78%**

L'elevato coefficiente di variazione indica una **forte dispersione dei valori** rispetto alla media, segnalando una significativa variabilità nelle valutazioni tra gli utenti.

**Distribuzione di Frequenza tramite Diagramma di Pareto:** L'analisi tramite diagramma di Pareto permette di visualizzare come le frequenze assolute [9] siano associate alla frequenza relativa cumulativa, sottolineando la predominanza di utenti con basse valutazioni e il peso cumulativo degli utenti con più fallimenti.

Diagramma di pareto Customer Value

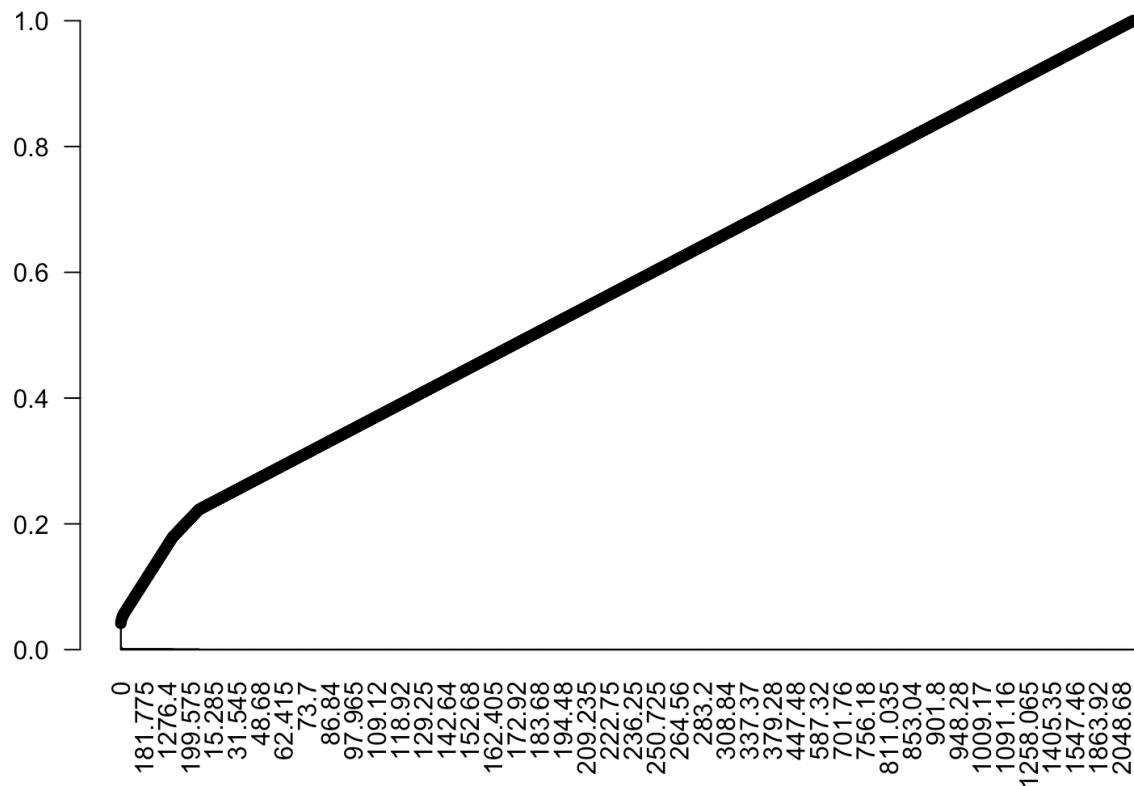


Figura 60 Diagramma di Pareto Customer Value

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 1.43**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 4.22**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza delle valutazioni dei fruitori, confermando le caratteristiche sopra descritte.

Distribuzione di Frequenza di Customer Value

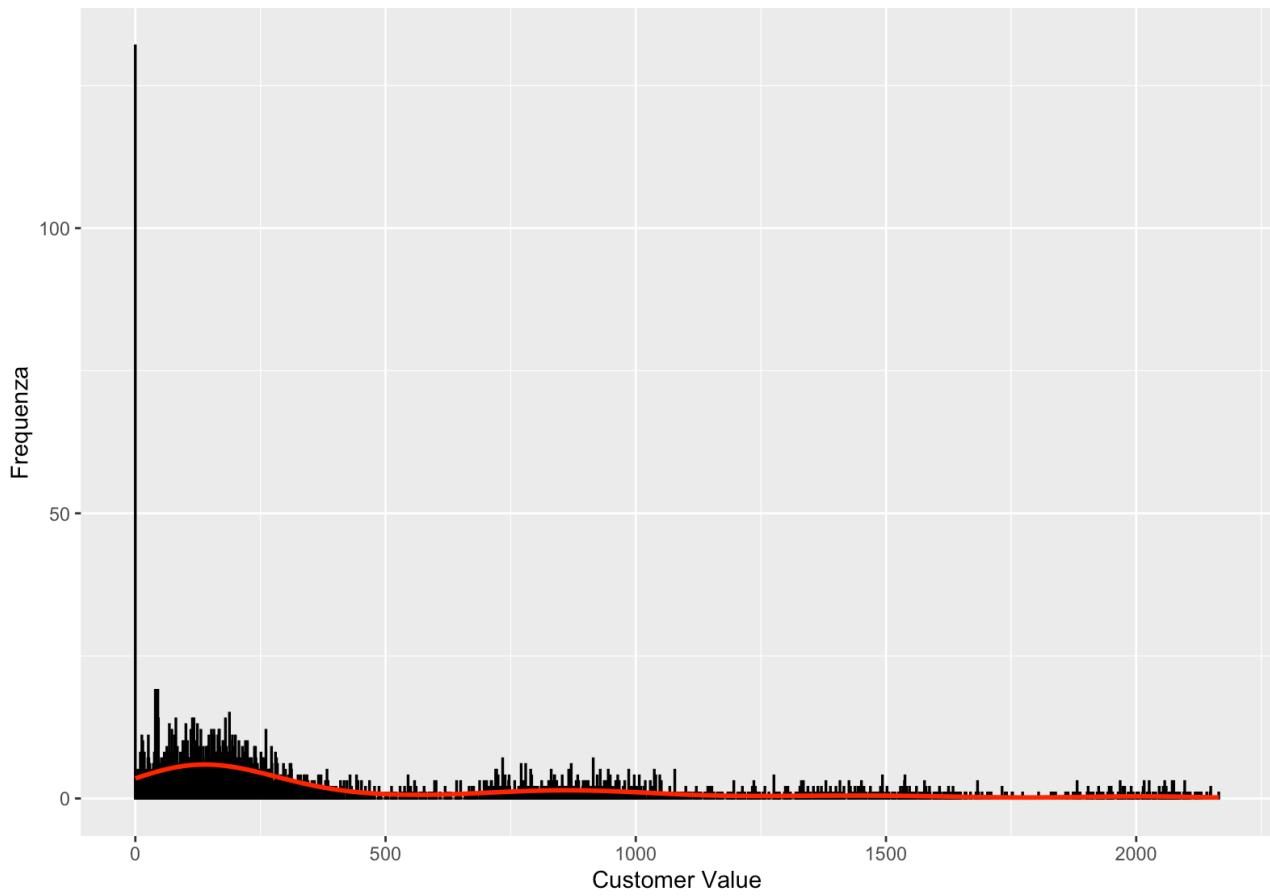


Figura 61 Distribuzione di frequenza Customer value

## 2.2. Analisi Bivariata

Un'analisi approfondita dei dati permette di esplorare le interazioni tra le diverse variabili e di identificare potenziali relazioni di dipendenza.

In questo capitolo esamineremo le relazioni bivariate e multivariate tra le variabili del dataset in analisi.

### 2.2.1. Customer Value VS. Frequency of SMS

In fase preliminare di analisi sorge spontanea la domanda: “**Come viene calcolato il Customer Value per ciascun fruitore?**”

Tramite l'analisi bivariata possiamo verificare se il [Customer Value](#) è influenzato da specifiche variabili o combinazioni di variabili del dataset.

Pertanto, iniziamo identificando le variabili che potrebbero influenzare il Customer Value. Dall'analisi dei grafici di correlazione tra il Customer Value e altre variabili, emerge una forte correlazione lineare con la variabile [Frequency of SMS](#).

Di seguito il plot che traccia questa correlazione:

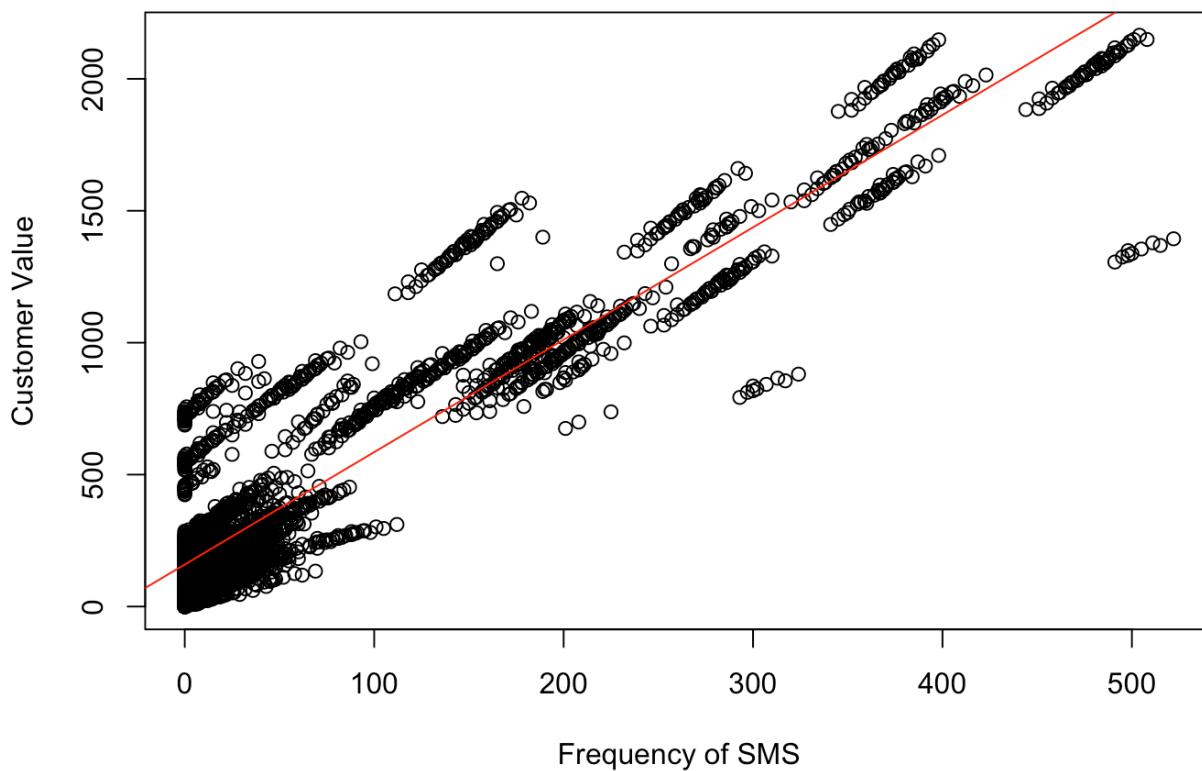


Figura 62 Correlazione Customer Value & Frequency of SMS

Questa correlazione può essere verificata anche tramite la **covarianza campionaria** [14], che misura quanto le due variabili variano insieme.

Calcolando la covarianza, otteniamo un valore pari a **53669.30**, il quale indica una relazione positiva tra le due variabili. Questo risultato è confermato anche dal grafico, dove osserviamo una retta di interpolazione crescente.

Determinando il **coefficiente di correlazione campionario** [15], si ottiene un valore pari a **0.92**, che indica una forte correlazione positiva tra le due variabili.

Un coefficiente di correlazione pari a **0.92** suggerisce una relazione lineare forte e positiva tra le due variabili. Nel contesto del dataset in esame:

Quando la **frequenza degli SMS** aumenta, il **valore del cliente** tende ad aumentare in modo consistente.

Andando poi a calcolare il **modello di regressione lineare semplice** [16] di questa relazione tra variabili troviamo che correlazione ha un **coefficiente di determinazione** [18] (che in questo caso di regressione lineare semplice, il coefficiente di determinazione coincide con il quadrato del coefficiente di correlazione) pari a **0.92**, il che indica che il **92% della variabilità del Customer Value può essere spiegata dalla variabile Frequency of SMS** in un modello di regressione lineare.

Questo valore suggerisce che la regressione è altamente attendibile per spiegare la relazione tra queste due variabili.

Il coefficiente di regressione per *Frequency of SMS* è pari a **4.26**.

Questo valore indica che, in media, ogni unità aggiuntiva di *Frequency of SMS* si associa a un incremento di **4.26** unità nel *Customer Value*.

Calcolando poi la media campionaria dei **residui** [17] abbiamo un valore pari a **-2.323813e-14**.

Un valore quindi molto vicino allo 0 che ci dice che il modello non ha una tendenza a sovrastimare o sottostimare i valori reali. In pratica, i residui oscillano attorno a zero in modo equilibrato.

## 2.2.2. Customer Value VS. Frequency of use

Customer Value, come indicato, mostra una dipendenza del **92%** dalla variabile Frequency of SMS.

Tuttavia, è possibile ottenere una comprensione più approfondita delle dinamiche che determinano il **Customer Value** aggiungendo ulteriori variabili dipendenti al modello.

In questa analisi, prendiamo in considerazione anche la variabile Frequency of Use.

Anche questa variabile sembra mostrare una correlazione diretta con il Customer Value.

L'osservazione del plot ottenuto tra **Customer Value** e **Frequency of Use** evidenzia anche qui una relazione lineare non troppo significativa tra le due variabili, suggerendo che *Frequency of Use* potrebbe contribuire in modo rilevante alla spiegazione della variabilità del Customer Value.

Di seguito il plot di correlazione:

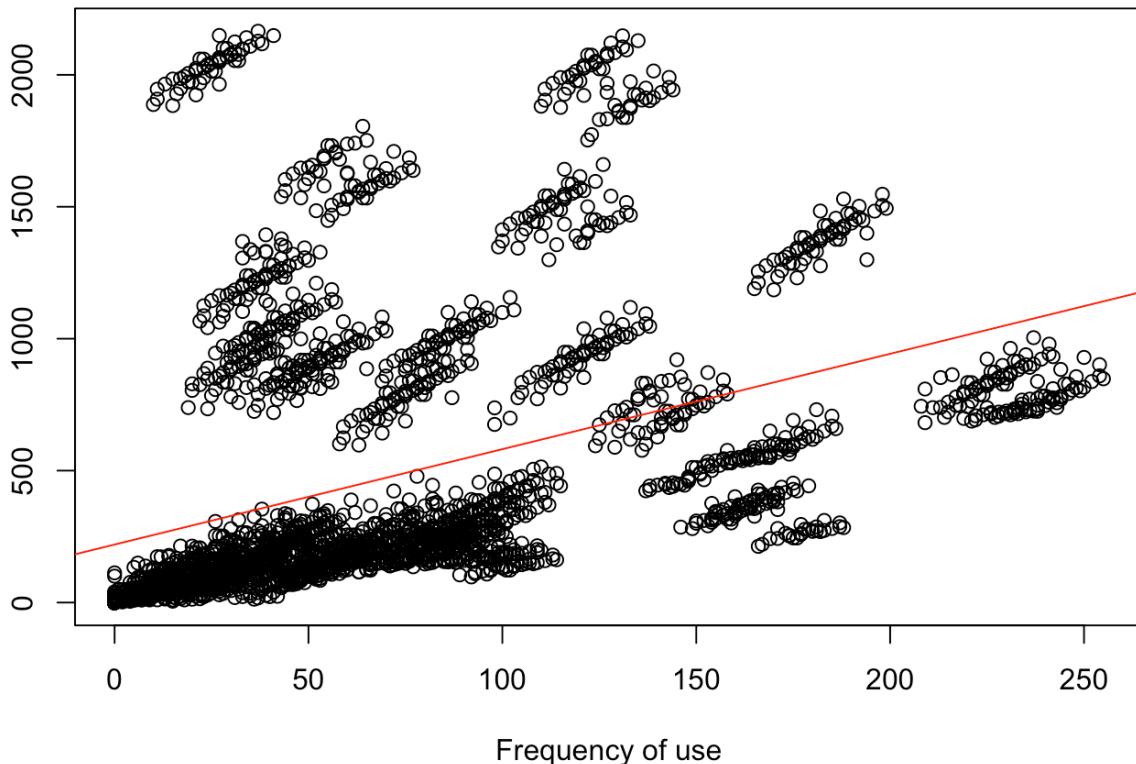


Figura 63 Correlazione Customer Value & Frequency of use

Questa correlazione può essere verificata anche tramite la **covarianza campionaria** [14], che misura quanto le due variabili variano insieme.

Calcolando la covarianza, otteniamo un valore pari a **11920.1**, il quale indica una relazione positiva tra le due variabili.

Questo risultato è confermato anche dal grafico, dove osserviamo una retta di interpolazione crescente.

Determinando il **coefficiente di correlazione campionario** [15], si ottiene un valore pari a **0.92**, che indica una forte correlazione positiva tra le due variabili.

Un coefficiente di correlazione pari a **0.40** non suggerisce una relazione lineare molto forte ma sicuramente ci dice che sono correlate positivamente.

Andando poi a calcolare il **modello di regressione lineare semplice** [16] di questa relazione tra variabili troviamo che correlazione ha un **coefficiente di determinazione** [18] (che in questo caso di regressione lineare semplice, il coefficiente di determinazione coincide con il quadrato del coefficiente di correlazione) pari a **0.161**.

Questo valore suggerisce una correlazione molto debole, quasi assente, tra le due variabili in un contesto di regressione lineare semplice.

La situazione cambia significativamente quando si applica un modello di regressione lineare multipla con *Customer Value* come variabile dipendente e includendo sia **Frequency of SMS** che **Frequency of Use** come variabili indipendenti.

### 2.2.3. Customer Value VS. Frequency of SMS & Frequency of use

Utilizzando la variabile [Customer Value](#) come variabile dipendente e le variabili [Frequency of Use](#) e [Frequency of SMS](#) come variabili indipendenti, il modello di **regressione lineare multiplo** [19] calcolato mostra un **coefficiente di determinazione** [18] pari a **0.95**. Questo indica che il **95.20%** della variabilità di Customer Value può essere spiegata dalle variabili Frequency of SMS e Frequency of Use.

Di seguito, il grafico 3D risultante che visualizza questa relazione:

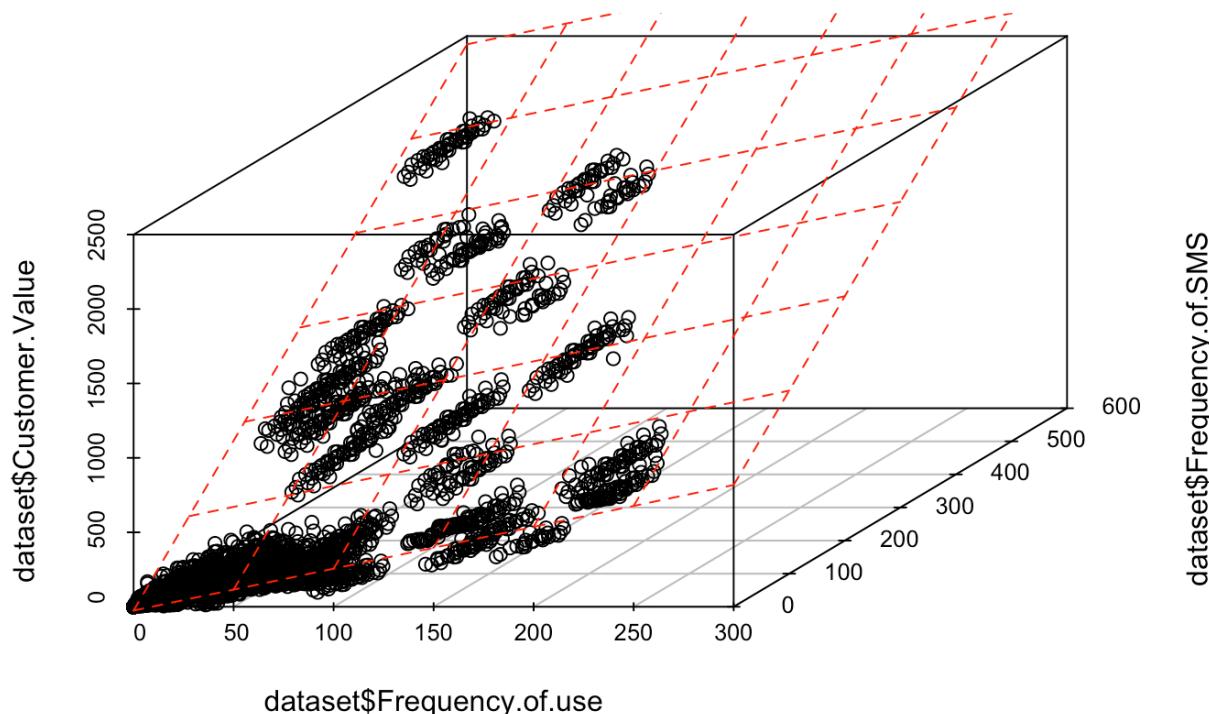


Figura 64 Customer Value in funzione di Frequency of use e Frequency of SMS

Calcolando poi la media campionaria dei residui abbiamo un valore pari a **-3.185262e-14**. Un valore quindi molto vicino allo 0 che ci dice che il modello non ha una tendenza a sovrastimare o sottostimare i valori reali. In pratica, i residui oscillano attorno a zero in modo equilibrato.

**Possiamo quindi dire per certo che la variabile customer value è calcolata utilizzando le variabili Frequency of Use e Frequency of SMS.**

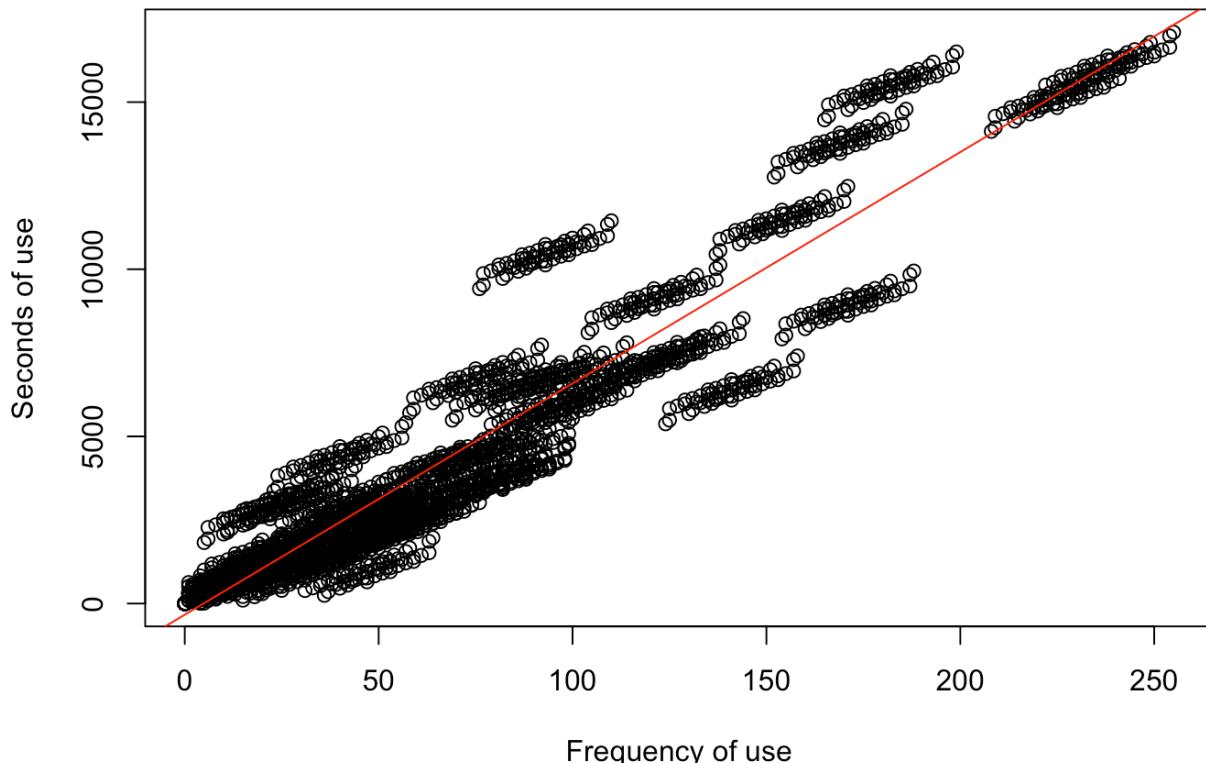
## 2.2.4. Frequency of use VS. Seconds of use

Nel [capitolo 2.1.7](#) abbiamo ipotizzato che la feature *Frequency of Use* rappresenti la frequenza delle chiamate distinte per ciascun fruitore del servizio.

Per confermare questa ipotesi, o almeno dimostrare che *Frequency of Use* è correlata al numero di chiamate effettuate, possiamo esaminare la sua relazione con la variabile *Seconds of Use*.

In teoria, se un utente effettua un numero  $n$  di chiamate per una durata complessiva  $t$ , ci si aspetterebbe una correlazione tra le due variabili.

Andando quindi a piazzare le due variabili tracciando la retta ottenuta dal modello lineare dato dalla correlazione delle due variabili come vediamo nel seguente plot, possiamo notare che c'è una forte correlazione fra le due.



*Figura 65 Correlazione Frquency of use & Seconds of use*

Questa correlazione può essere verificata anche tramite la **covarianza campionaria** [\[14\]](#), che misura quanto le due variabili variano insieme.

Calcolando la covarianza, otteniamo un valore pari a **228118.9**, il quale indica una relazione positiva tra le due variabili.

Questo risultato è confermato anche dal grafico, dove osserviamo una retta di interpolazione crescente.

Determinando il **coefficiente di correlazione campionario** [\[15\]](#), si ottiene un valore pari a **0.95**, che indica una forte correlazione positiva tra le due variabili.

Un coefficiente di correlazione pari a **0.95** suggerisce una relazione lineare forte e positiva tra le due variabili.

Nel contesto del dataset in esame:

Quando il numero di **secondi di utilizzo** aumenta, il **frequenza di utilizzo** tende ad aumentare in modo consistente.

Andando poi a calcolare il **modello di regressione lineare semplice** [16] di questa relazione tra variabili troviamo che correlazione ha un **coefficiente di determinazione** [18] (che in questo caso di regressione lineare semplice, il coefficiente di determinazione coincide con il quadrato del coefficiente di correlazione) pari a **0.95**, il che indica che il **95%** della variabilità del campo **Frequency of use** può essere spiegata dalla variabile **Seconds of use** in un modello di regressione lineare.

Questo valore suggerisce che la regressione è altamente attendibile per spiegare la relazione tra queste due variabili.

Il coefficiente di regressione per *Frequency of SMS* è pari a **69.20**. Questo valore indica che, in media, ogni unità aggiuntiva di *Frequency of use* si associa a un incremento di **69.21** unità nel *Seconds of use*.

Calcolando la media campionaria dei residui, otteniamo un valore pari a **1.031122e-13**.

Un valore così prossimo allo zero indica che il modello non mostra una tendenza sistematica a sovrastimare o sottostimare i valori reali, in quanto i residui oscillano attorno allo zero in modo bilanciato.

**Possiamo dire quindi che l'assunzione fatta nel capitolo 2.1.7 associando il Frequency of use al numero di chiamate distinte può essere attendibile.**

### 3. Clustering

Per proseguire l'analisi del dataset, risulta fondamentale indagare le ragioni che determinano l'abbandono del servizio ([Churn](#) = 1) o la permanenza ([Churn](#) = 0) dei clienti. A tal fine, possiamo applicare una tecnica di **clustering**, per poi esaminare i dati suddivisi in cluster, al fine di ottenere una visione più chiara delle dinamiche sottostanti al comportamento di Churn rate e fedeltà.

Bisogna prima di tutto andare a comprendere quale tipo di clustering utilizzare se un algoritmo di clustering gerarchico oppure uno non gerarchico.

Per comprendere meglio andiamo ad analizzare pro e contro di entrambe le soluzioni:

- **Clustering gerarchico**

**Pro:**

1. Non richiede di specificare il numero di cluster a priori:  
Il clustering gerarchico permette di esplorare i dati e scegliere il numero di cluster ottimale osservando il dendrogramma.
2. Rappresentazione visiva:  
Il dendrogramma consente di visualizzare le relazioni tra i dati e i cluster, offrendo un'interpretazione più intuitiva.
3. Adatto per cluster non sferici:  
Funziona bene anche se i cluster non sono simmetrici o hanno forme irregolari.
4. Supporta varie metriche di distanza:  
È possibile utilizzare distanze diverse (Euclidea, Manhattan, ecc.), adattandosi meglio a dataset di natura mista o complessa.

**Contro:**

1. Limitato a dataset di dimensioni piccole o medie:  
La complessità computazionale cresce rapidamente con il numero di punti, rendendolo impraticabile per dataset molto grandi.
2. Non rivedibile:  
Una volta che unione o divisione tra cluster è stata effettuata, non può essere corretta. Questo lo rende meno flessibile rispetto a metodi iterativi.
3. Difficoltà nel gestire rumore e outlier:  
Gli outlier possono distorcere significativamente la struttura gerarchica.

- **Clustering non gerarchico**

**Pro:**

1. Computazionalmente efficiente:  
K-means è più rapido e scalabile rispetto al clustering gerarchico, ideale per dataset di grandi dimensioni.
2. Facilmente implementabile:  
È semplice da configurare e ampiamente supportato da software statistici.
3. Adatto a cluster sferici:  
Funziona molto bene quando i cluster sono simmetrici e ben separati.
4. Iterativo e flessibile:  
Può riassegnare i punti ai cluster durante ogni iterazione, migliorando progressivamente l'accuratezza.

**Contro:**

1. Richiede di specificare  $k$  a priori:  
Il numero di cluster deve essere definito prima di eseguire l'algoritmo, il che richiede una fase esplorativa (es. metodo dell'Elbow).
2. Sensibile ai valori iniziali:  
La qualità del risultato può dipendere dalla scelta iniziale dei centroidi.
3. Cluster sferici richiesti:  
Non funziona bene con cluster di forme irregolari o con densità variabile.
4. Meno robusto agli outlier:  
Gli outlier possono influenzare notevolmente i centroidi, distorcendo i risultati.

Per questa analisi, abbiamo scelto di utilizzare l'algoritmo di clustering **K-means** (non gerarchico), principalmente per la sua maggiore efficienza computazionale rispetto al clustering gerarchico.

Questo è particolarmente vantaggioso nel caso di dataset con un numero elevato di osservazioni, come il nostro.

Inoltre, le caratteristiche delle variabili da analizzare supportano l'uso di **K-means**:

- Due variabili qualitative: Come verrà approfondito in seguito tramite l'analisi con Random Forest, le variabili qualitative coinvolte non presentano outlier, poiché i valori possibili sono categoriali e ben definiti.
- Una variabile quantitativa: Questa variabile, pur essendo soggetta alla presenza di outlier, può essere utilizzata per esplorare le differenze tra il clustering effettuato con e senza outlier.

In particolare, sarà interessante osservare come la rimozione degli outlier possa o meno influenzare il clustering.

Come algoritmo per la suddivisione in cluster, quindi, è stato scelto **K-means** [11], che, per questo specifico problema, presenta i seguenti vantaggi:

1. **Scalabilità ed efficienza:** Il dataset in esame contiene numerose istanze, rendendo necessaria l'adozione di un algoritmo efficiente e scalabile, in grado di gestire grandi quantità di dati senza compromettere significativamente le performance.

- 
2. **Robustezza agli outlier:** Sebbene **K-means** possa essere sensibile alla presenza di outlier, l'algoritmo risulta relativamente robusto. In presenza di outlier, è possibile applicare tecniche di **pre-processing** per mitigare l'impatto di questi valori anomali sul clustering.  
Inoltre le variabili che verranno scelte più avanti saranno due di tipo qualitativo (non possono avere outliers) ed una di tipo quantitativo (può avere outliers ma verrà fatto anche uno studio di come si comporta il clustering con k-means in presenza e non di outliers).
  3. **Raggruppamento significativo:** **K-means** permette di raggruppare i clienti in cluster distinti, che possono essere successivamente analizzati per comprendere meglio i fattori che influenzano il comportamento di **Churn rate** (abbandono) e **permanenza**.

Il primo passo consiste nell'identificare il numero ottimale di cluster, denotato come **k**, in cui suddividere i dati.

Sebbene, in linea teorica, un clustering basato su una variabile binaria (come **Churn**, che assume i valori 0 o 1) possa essere eseguito con **k = 2**, le analisi preliminari non hanno prodotto risultati soddisfacenti con tale configurazione.

Una volta definito il numero ottimale di cluster, sarà possibile esaminare come i comportamenti dei clienti (ad esempio, frequenza d'uso, durata dell'abbonamento, numero di reclami) differiscano tra i vari gruppi, e correlare queste informazioni con il **Churn rate**, facilitando così una comprensione dettagliata delle cause che influenzano l'abbandono o la fedeltà dei clienti.

Per determinare il numero ottimale di cluster, è stato scelto l'**Elbow Method**. Questo metodo è stato preferito in quanto fornisce una tecnica semplice e visiva per determinare il numero ottimale di cluster, riducendo il rischio di **overfitting** e garantendo un clustering più significativo e generalizzabile, inoltre, per avere una maggiore sicurezza sul numero **k** scelto verrà fatta anche una verifica con il metodo **Silhouette**.

L'**Elbow Method** è particolarmente utile in contesti in cui l'algoritmo di clustering, come **K-means** [11], richiede la definizione a priori del numero di cluster.

Esso consente di identificare il numero **k** ottimale basato su un'analisi chiara della **varianza intra-cluster** (ovvero la dispersione all'interno di ciascun cluster), permettendo di bilanciare la complessità del modello con la qualità del clustering.

**Prima di procedere al calcolo del valore ottimale di k** per il clustering, è necessario selezionare le features su cui basare l'analisi.

Utilizzando l'algoritmo **Random Forest**, è possibile identificare le variabili più significative che influenzano maggiormente la variabile target **Churn**. Questo algoritmo, tramite la creazione di alberi decisionali, ci consente di determinare quali caratteristiche dei dati impattano di più sulla variabile di interesse, facilitando così la scelta delle features più rilevanti per il clustering.

Di seguito il grafico che descrive le variabili quanto impattano sulla variabile churn:

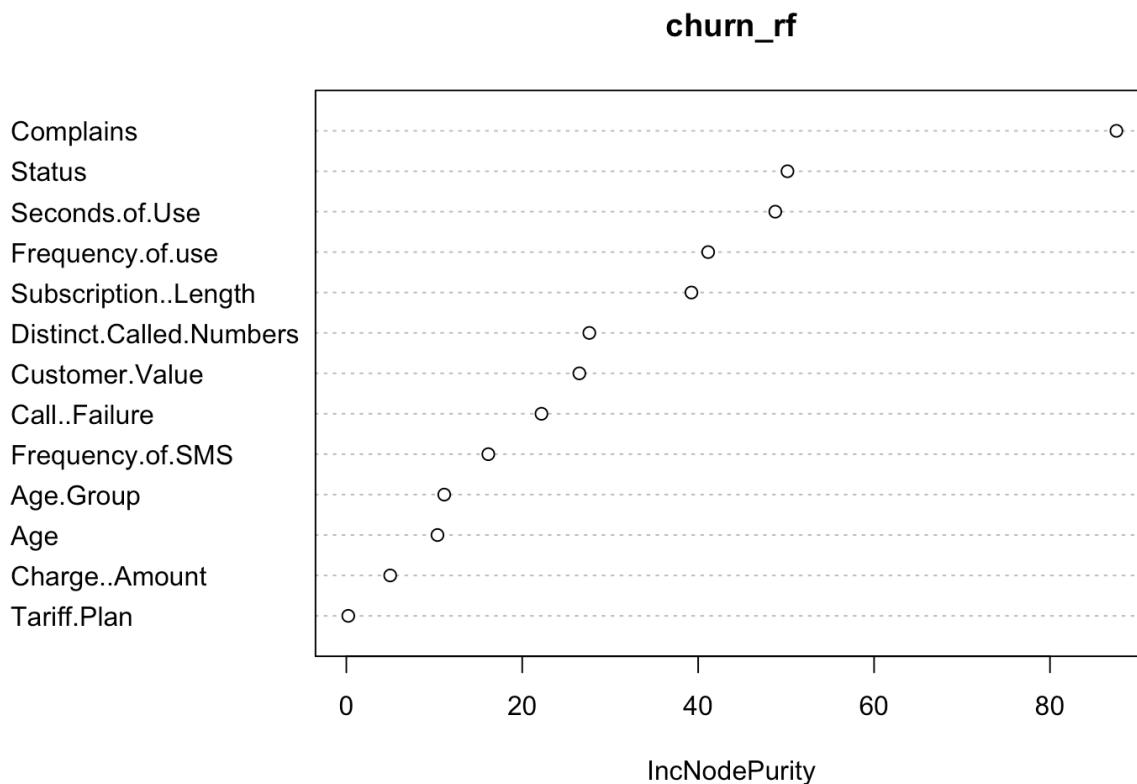


Figura 66 Impatto variabili indipendenti su variabile Churn

Le variabili quindi che sembrano avere maggiore impatto sono:

- Complains
- Status
- Seconds of use

I dati ovviamente sono stati opportunamente scalati (standardizzati sottraendo la media della variabile e dividendo per la deviazione standard) dato che sappiamo che il **k-means si basa sulla distanza euclidea** e questo tipo di metrica è soggetta ad essere fortemente collegata alle unità di misura e essendo la variabile **seconds of use è strettamente legata ai secondi** è stato opportuno scalarla.

Utilizziamo poi il metodo dell'Elbow per determinare il numero ottimale di cluster **k** a partire dalle variabili selezionate.

In questo caso, per applicare il metodo dell'Elbow, è stato calcolato il **Within-Cluster Sum of Squares** (WSS) [12] per un intervallo di valori di k compreso tra 1 e 10. Il valore di k ottimale corrisponde al punto in cui si osserva un cambiamento significativo nella pendenza della curva, ovvero il punto in cui il WSS inizia a diminuire più lentamente.

Ora possiamo procedere al kmeans.

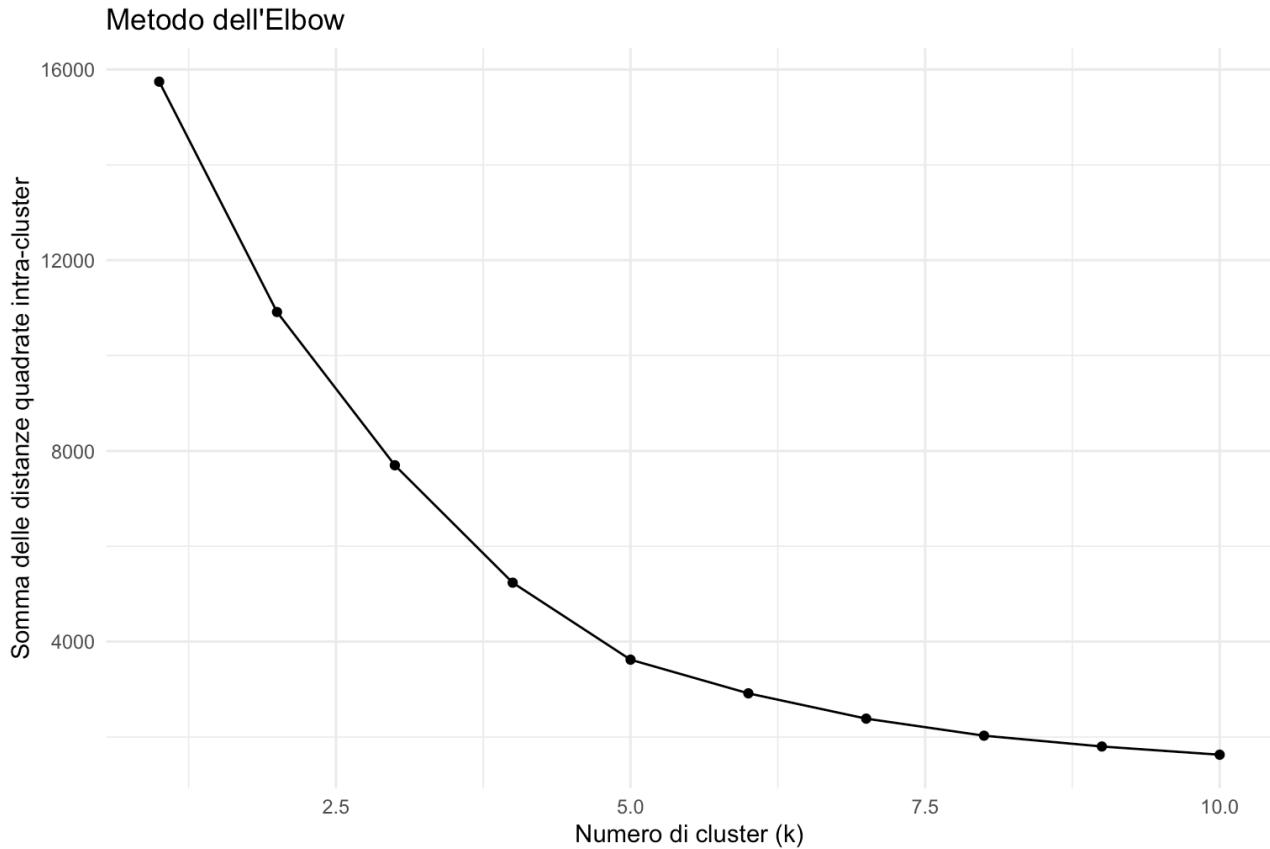


Figura 67 Risultato Elbow method

Analizzando il grafico risultante abbiamo che il  $k$  ottimo è 4.

Andiamo a confermare questa teoria sfruttando il metodo **Silhouette**.

Il metodo **Silhouette** trova il  $k$  ottimo sulla base ottimo analizzando la coerenza interna dei cluster e la separazione tra di essi, scegliendo  $k$  che massimizza il valore medio della silhouette.

Il seguente grafico mappa il numero di cluster con il valore medio della **silhouette**:

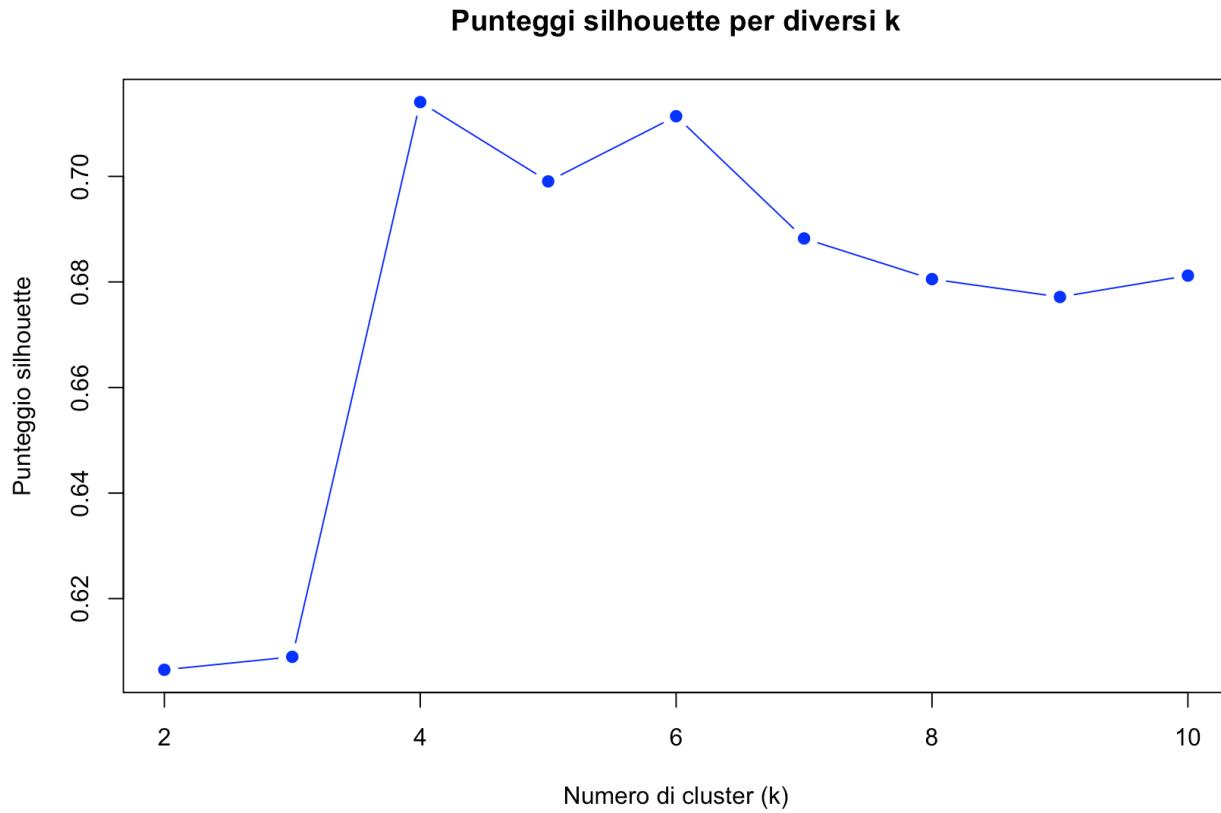


Figura 68 Silhouette per calcolo k

Dal grafico della silhouette possiamo confermare che il **k** ottimo è proprio 4 dato che scegliamo quello con il punteggio più alto.

Abbiamo quindi effettuato il clustering utilizzando l'algoritmo **k-means** [11], ottenendo cinque cluster. Di seguito è riportata la distribuzione del numero di utenti che hanno effettuato il churn e di quelli che non lo hanno fatto all'interno di ciascun cluster:

Churn rate	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Non abbandonato	41	487	1732	395
Abbandonato	200	0	66	229

Inoltre, il clustering ha fornito i seguenti valori di bontà:

- **Indice di Calinski-Harabasz** [13]: 7800.86 (un valore piuttosto alto, indicativo di una buona separazione tra i cluster).
- **WSS (Within-Cluster Sum of Squares)** [12]: 1119.47 (non particolarmente basso, ma accettabile considerando la possibile presenza di outlier nel capitolo successivo analizzeremo il clustering rimuovendo gli outliers dall'unica variabile quantitativa coinvolta ovvero Seconds of use).
- **BSS (Between Cluster Sum of Squares)**: 8327.53.

Per ciascun cluster, possiamo analizzare come le variabili selezionate siano state aggregate, al fine di comprendere meglio le caratteristiche che accomunano i dati raggruppati in ogni cluster.

### 3.1.1. Complains

Andando a tracciare quanti Utenti si sono lamentati e quanti utenti non si sono lamentati per ogni cluster otteniamo:

Complains	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Non lamentato	0	487	1798	624
Lamentato	241	0	0	0

Analizzando cluster per cluster abbiamo che:

- 1) Per il cluster 1 che la lamentela è stato un fattore scatenante dell'abbandono degli utenti, di fatti notiamo che nel cluster 1 l'**83%** dei clienti che si sono lamentati hanno abbandonato il servizio.
- 2) Nel cluster 2 tutti i clienti che si sono lamentati del servizio sono rimasti abbonati.
- 3) Il cluster 3 ha lo stesso comportamento del cluster 2.
- 4) Nel cluster 4 invece vediamo che la lamentela non ha scatenato l'abbandono dei 229 che hanno effettivamente abbandonato il servizio.

### 3.1.2. Status

Andiamo ora a studiare i valori di status che ricordiamo possono essere (**1: Attivo, 2: Non attivo**).

Tecnicamente qui andiamo a tracciare che se un utente ha abbandonato gli è stato disattivato il servizio.

Status	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Attivo	83	487	1798	0
Disattivo	158	0	0	624

Analizzando cluster per cluster abbiamo che:

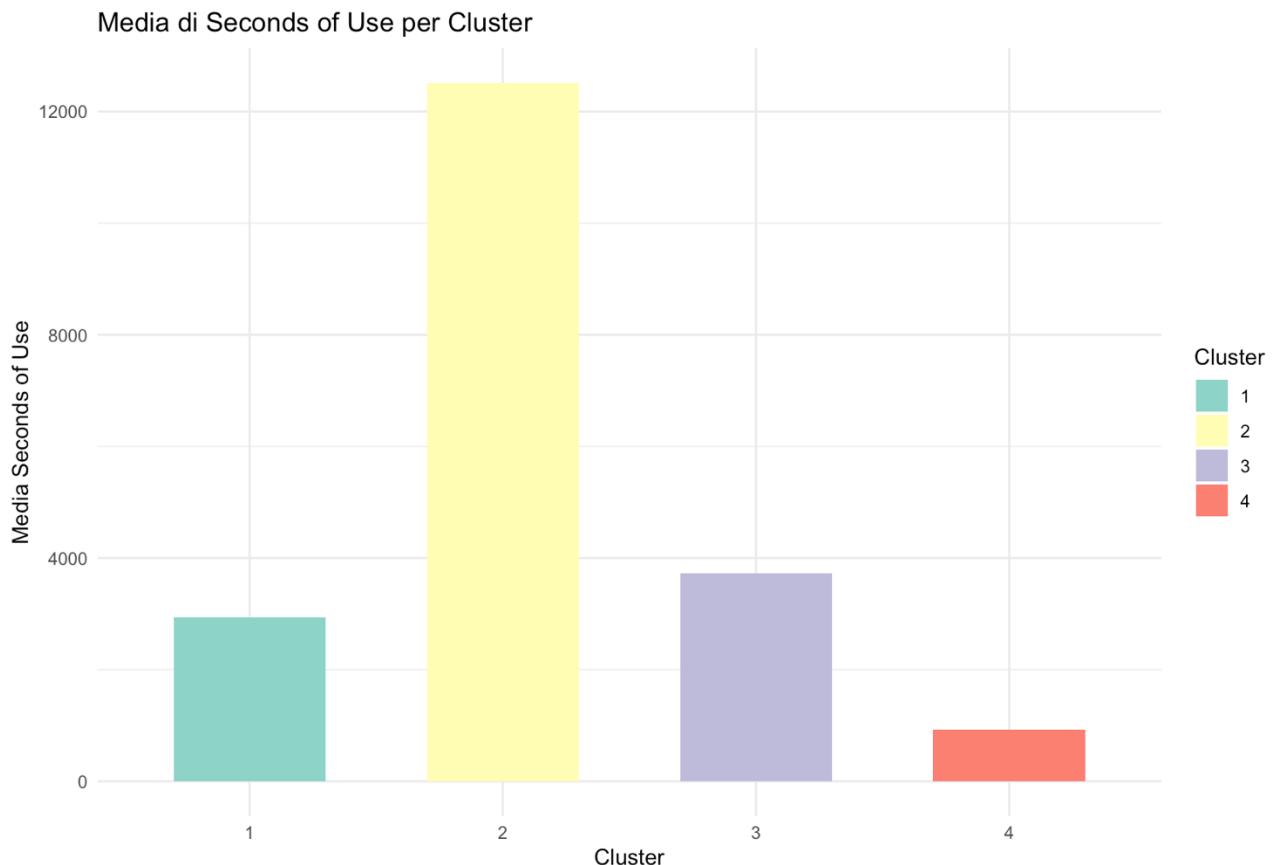
- 1) Per il cluster 1 al **79%** degli abbonati che hanno abbandonato il servizio è stato disattivato l'abbonamento.
- 2) Nel cluster 2 tutti i clienti che non hanno abbandonato il servizio hanno ancora il servizio attivo.

- 3) Il cluster 3 ha lo stesso comportamento del cluster 2.
- 4) Nel cluster 4 invece vediamo che seppure il **65.30%** non ha abbandonato il servizio ha comunque l'abbonamento non più attivo.

### 3.1.3. Seconds of use

Per la variabile seconds of use invece andiamo a fare un discorso diverso.

Andremo a verificare la media di secondi di chiamata per ogni cluster per evidenziare come un'alta concentrazione di utenti che hanno abbandonato il servizio implica un basso utilizzo di un servizio. Andremo a confrontare le medie della variabile Seconds of use di ogni cluster tramite un barplot:



*Figura 68 Risultato del clustering su seconds of use*

Come vediamo nel grafico è abbastanza evidente che gruppi di utenti(cluster) che hanno la maggioranza di utenti che hanno abbandonato il servizio hanno in media un numero di secondi di utilizzo inferiore ai gruppi che invece hanno quasi la totalità di utenti che hanno mantenuto il servizio.

Diremo quindi che gli utenti che abbandonano il servizio tendenzialmente lo utilizzano meno.

## 3.2. Clustering con rimozione degli outliers

Come discusso nel capitolo precedente, abbiamo identificato  $k = 4$  come il valore ottimale per il clustering e individuato le variabili che maggiormente influenzano la variabile target Churn nel dataset.

Le variabili utilizzate per il clustering includono due qualitative (Complains e Status), che per loro natura non possono presentare outlier, e una quantitativa (Seconds of Use), che invece può contenere outlier.

Procediamo quindi a verificare se la rimozione degli outlier nella variabile quantitativa produce risultati coerenti rispetto al clustering originale.

### 3.2.1. Rimozione degli outliers dalla variabile Seconds of use

La rimozione degli outliers da features di un dataset tipicamente consiste in 3 passaggi:

- Trovare gli outliers della variabile che come abbiamo visto in precedenza sono:  
15140, 15485, 16075, 15200, 15545, 16135, 15080, 15425, 16015, 15220, 15565, 16155, 15060, 15405, 15995, 15400, 15745, 16335, 14880, 15225, 15815, 15320, 15665, 16255, 14960, 15305, 15895, 14373, 15740, 16085, 16675, 14540, 14885, 15475, 15255, 15600, 16190, 15025, 15370, 15960, 15330, 15675, 16265, 14950, 15295, 15885, 14483, 15850, 16195, 16785, 14430, 14775, 15365, 14835, 15180, 15770, 14895, 15240, 15830, 15120, 15710, 14915, 15260, 15690, 15095, 15440, 16030, 14575, 14920, 15510, 15015, 15360, 15950, 14655, 15000, 15590, 15435, 15780, 16370, 14235, 14580, 15170, 14720, 15065, 15655, 14990, 15580, 14178, 16480, 14125, 14470, 15445, 15790, 16380, 14138, 15505, 16440, 15385, 15730, 16320, 14158, 15525, 15870, 16460, 14338, 15705, 16050, 16640, 15185, 15530, 16120, 14258, 15625, 15970, 16560, 15265, 15610, 16200, 14678, 16045, 16390, 16980, 14845, 15190, 15905, 16495, 16570, 14788, 16500, 17090, 14735, 15670.
- Trovare la mediana del dato che, come visto in precedenza, è 2990.
- Sostituire i valori degli outliers con la mediana e verificare se c'è un cambiamento nel clustering.

Anche in questo caso è stato utilizzato l'algoritmo di clustering **kmeans** con **k** pari a **4**. Di seguito è riportata la distribuzione del numero di utenti che hanno effettuato il churn e di quelli che non lo hanno fatto all'interno di ciascun cluster:

Churn rate	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Non abbandonato	1469	395	41	750
Abbandonato	48	229	200	18

Inoltre, il clustering ha fornito i seguenti valori di bontà:

- **Indice di Calinski-Harabasz:** **8007.03** (un valore più alto rispetto a quello con la presenza di outliers).
- **WSS** (Within-Cluster Sum of Squares) [12]: **1093.98** (un valore ancora non basso ma migliorato rispetto a quello con la presenza di outliers).
- **BSS** (Between Cluster Sum of Squares): **8353.02**.

Diciamo quindi che analizzando solo gli indici vediamo che la situazione del clustering è migliorata ma non eccessivamente, di fatti possiamo dedurre che gli outliers non hanno in questo caso un così grande impatto sulla clusterizzazione dei dati.

Possiamo comunque per ciascun cluster, analizzare come le variabili selezionate siano state aggregate come fatto per il caso con outliers.

### 3.2.2. Complains

Andando a tracciare quanti Utenti si sono lamentati e quanti utenti non si sono lamentati per ogni cluster otteniamo:

Complains	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Non lamentato	1517	624	0	768
Lamentato	0	0	241	0

Analizzando cluster per cluster abbiamo che:

- 1) Per il cluster 1 che la lamentela non è stato un fattore scatenante dell'abbandono degli utenti, anche per i 48 che hanno poi abbandonato il servizio.
- 2) Nel cluster 2 invece, abbiamo che anche se il **36.70%** ha abbandonato il servizio questo non ha presentato lamentele.
- 3) Il cluster 3 invece esprime che la lamentela è stato il fattore scatenante dell'abbandono di fatti il ben **83%** degli utenti che si è lamentato ha poi abbandonato il servizio.

- 
- 4) Nel cluster 4 vediamo che, molto similmente al cluster 1, seppure nessuno degli utenti si è lamentato, 18 hanno comunque abbandonato il servizio.

### 3.2.3. Status

Andiamo ora a studiare i valori di status che ricordiamo possono essere (**1: Attivo, 2: Non attivo**).

Status	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Attivo	1517	0	83	768
Disattivo	0	624	158	0

Analizzando cluster per cluster abbiamo che:

1. Per il cluster 1 al vediamo che seppure 48 utenti di questo cluster abbiano abbandonato il servizio il loro servizio ancora non è stato disattivato.
2. Nel cluster 2 abbiamo che seppure 395 non abbiano abbandonato il servizio, comunque il servizio è stato disattivato per queste utenze.
3. Il cluster 3 ci dice che seppure 200 utenti appartenenti a questo cluster abbiano abbandonato il servizio, solo al 79% degli utenti è stato effettivamente disattivato il servizio.
4. Nel cluster 4 invece vediamo che tutti gli utenti appartenenti hanno ancora il servizio attivo anche se di questi 768 18 hanno abbandonato il servizio.

### 3.3. Clustering con rimozione degli outliers ed utilizzo k-means++

Precedentemente, abbiamo utilizzato l'algoritmo **k-means** per effettuare il clustering basandoci sulle variabili Complains, Status e Seconds of Use, dopo aver rimosso gli outlier relativi alla feature Seconds of Use.

Un'analisi ancora più approfondita consiste nell'utilizzare l'algoritmo **k-means++** anziché k-means, per evitare di scegliere i k centroidi iniziali in modo casuale. **K-means++** seleziona i centroidi iniziali massimizzando la distanza tra loro, aumentando così il **BCSS** (Between Cluster Sum of Squares) e, di conseguenza, migliorando la qualità del clustering. Questo approccio porta a un **incremento dell'Indice di Calinski-Harabasz**, che misura la coesione e la separazione tra i cluster.

Effettuiamo ora clustering **kmeans++** con k pari a **4**.

Di seguito è riportata la distribuzione del numero di utenti che hanno effettuato il churn e di quelli che non lo hanno fatto all'interno di ciascun cluster:

Churn rate	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Non abbandonato	41	395	750	1469
Abbandonato	200	229	18	48

Inoltre, il clustering ha fornito i seguenti valori di bontà:

- **Indice di Calinski-Harabasz:** **8007.03**.
- **WSS** (Within-Cluster Sum of Squares) [[12](#)]: **1093.98**.
- **BSS** (Between Cluster Sum of Squares): **8353.02**.

Notiamo dai dati ottenuti dal clustering effettuato con k-means++ che i risultati sono pressoché identici a quelli ottenuti con il k-means classico. Ciò significa che, nel caso di questo dataset, i centroidi iniziali scelti casualmente dal k-means erano già abbastanza ideali.

## 3.4. Informazioni ottenute dal clustering

Per concludere l'analisi del clustering, abbiamo identificato come le variabili [Complains](#) (presenza di una lamentela da parte di un utente), [Status](#) (indicazione se il servizio è attivo o disattivo) e [Seconds of Use](#) (tempo di utilizzo del servizio da parte di un utente, espresso in secondi) abbiano un impatto significativo sull'abbandono del servizio ([churn](#)) da parte degli utenti.

In particolare, l'analisi dei cluster ha rivelato che:

- Gli utenti che hanno **presentato una lamentela (Complains)** mostrano una **maggior probabilità di abbandonare il servizio**, evidenziando una correlazione diretta tra insoddisfazione e churn.
- Gli utenti che hanno **abbandonato il servizio presentano sistematicamente uno Status disattivo**, confermando che il churn è associato alla cessazione dell'attività del servizio.
- Gli utenti con un **basso utilizzo del servizio** (minori valori di Seconds of Use) **tendono maggiormente ad abbandonare**, suggerendo che una scarsa frequenza o durata di utilizzo potrebbe essere un indicatore precoce di disinteresse o insoddisfazione.

Questi risultati forniscono una visione chiara delle dinamiche sottostanti al churn, evidenziando come l'insoddisfazione, la disattivazione del servizio e un utilizzo limitato rappresentino fattori critici nell'abbandono da parte degli utenti. Queste informazioni possono essere utilizzate per identificare clienti a rischio e adottare strategie mirate per aumentarne la fidelizzazione.

Possiamo quindi dire di aver trovato una risposta alla [Research question 2](#).

## 4. Studio proiezione popolazione di Churn su scala globale

Per rispondere alla [Research Question 3](#), analizziamo il comportamento degli utenti su scala mondiale, proiettando i dati forniti dal dataset.

Ci concentreremo sulla variabile [Churn](#), che indica l'abbandono del servizio (**o: l'utente ha abbandonato il servizio, 1: l'utente non ha abbandonato il servizio**).

Come evidenziato nel capitolo dedicato all'analisi univariata della variabile Churn, il dataset contiene informazioni su **3.150** utenti, di cui **485** hanno abbandonato il servizio (il **15,7%**).

Questo valore indica un tasso di abbandono relativamente basso.

Per portare avanti questo studio, ci avvarremo degli intervalli di confidenza.

Prima di iniziare, eseguiremo un test del chi-quadro per comprendere quale distribuzione rappresenti meglio la variabile churn.

Poiché vogliamo contare il totale dei casi di successo (1: abbandono), possiamo provare ad utilizzare una distribuzione binomiale.

Per iniziare, definiamo le ipotesi:

- **H<sub>0</sub>**: La distribuzione non è binomiale
- **H<sub>1</sub>**: La distribuzione è binomiale

Effettueremo il test del chi-quadro sui seguenti valori:

- **Numero di successi: 485**
- **Numero di insuccessi: 2655**
- **Probabilità di churn: 0.157**

Effettuando il test, otteniamo un **p-value pari a 1**, il che significa che la **variabile churn è descrivibile con una distribuzione binomiale [20]**.

Per continuare poi proveremo tramite il **teorema del limite centrale [21]** ad approssimare la nostra distribuzione **binomiale in una distribuzione normale**, di fatti come sappiamo, con un campione sufficientemente grande ( $n > 30$  e nel nostro caso **3150**) la distribuzione binomiale può essere approssimata in una normale.

Le distribuzioni normali hanno bisogno di due valori principali:

- **$\mu$  (il valore medio):** per il metodo dei momenti e per il metodo della massima verosimiglianza sappiamo che il valore medio della distribuzione è uno stimatore non distorto per ottenere  $\mu$ . Andando a calcolare il valore atteso (medio) della distribuzione, otteniamo un valore pari a **485**.

Il valore atteso di una distribuzione binomiale se noto il parametro **p** darebbe  **$n * p$**  ma in questi casi non lo abbiamo.

Andiamo quindi a calcolare quello che è il nostro  $\hat{p}$  empirico che otteniamo effettuando l'operazione **k/n** ovvero **485/2655**. Effettuando le operazioni otteniamo che  **$\mu = n * \hat{p}$**  il che vuol dire che il nostro  **$\mu$  è proprio uguale a k**.

- **$\sigma$  (deviazione standard):** per il metodo dei momenti e della massima verosimiglianza sappiamo che uno stimatore non distorto di  $\sigma$  è la deviazione

standard della distribuzione binomiale. Sappiamo che la deviazione standard per una distribuzione binomiale è  $\sigma = \sqrt{np(1-p)}$  ovviamente anche in questo caso non siamo in possesso del valore teorico di  $p$ , proprio per questo anche qui andiamo ad utilizzare il valore  $\hat{p}$  empirico che otteniamo effettuando l'operazione  $k/n$ . Effettuando i calcoli otteniamo che il valore di  $\sigma$  è **20.82**.

Otteniamo quindi una distribuzione Normale  $N(\mu, \sigma^2)$  ovvero **N(485, 483.47)**.

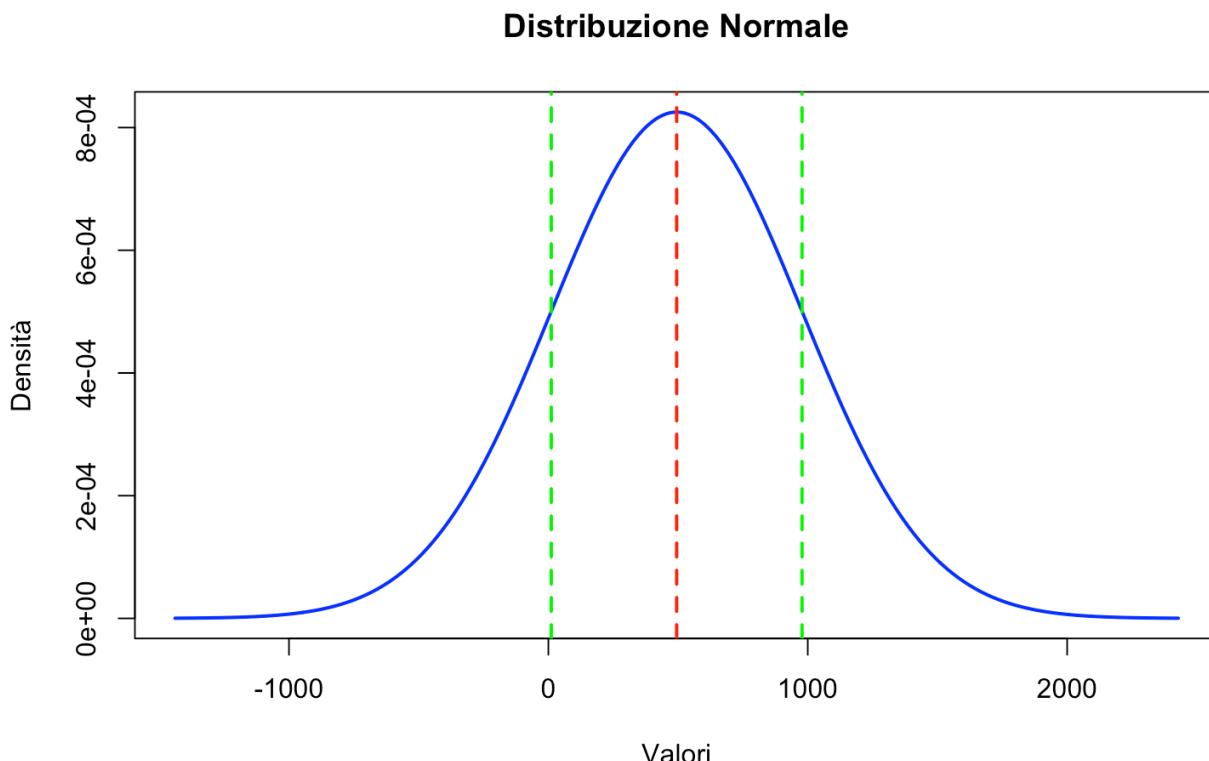


Figura 69 Rappresentazione normale della variabile di Churn

Ora, quindi, ci troviamo in una situazione ottimale perché avendo a disposizione una distribuzione normale, ci sarà più semplice ottenere il valore  $\mu$  per la popolazione globale. Vogliamo quindi trovare un intervallo di confidenza che andrà a identificare il nostro  $\mu$  che sarà compreso tra un certo quantile **-z1** e un altro quantile **z1**.

Andiamo quindi a scegliere un  $\alpha$  che ci dirà quanto è affidabile la nostra previsione. In questo caso scegliamo un  $\alpha$  pari a **0.1** per ottenere un'affidabilità del **90%**.

Per ottenere il valore di  $\mu$  utilizziamo **il metodo pivotale [23]**, ovvero, dopo aver **ottenuto una distribuzione Normale standard** standardizzando la media con **il teorema centrale della convergenza [22]** sarà più facile elaborare il valore  $\mu$  incognito da  **$N(0, 1)$** .

Utilizziamo il metodo pivotale dato che questo metodo rende la distribuzione più facile da elaborare e, allo stesso tempo, non dipende dal valore di  $\mu$  che vogliamo calcolare.

Abbiamo quindi che:

$$P\left(-z_1 < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_1\right) = 0.9$$

Andando poi a isolare  $\mu$  otteniamo:

$$P\left(\bar{X} - z_1 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_1 \frac{\sigma}{\sqrt{n}}\right) = 0.9$$

Avvalendoci poi della **tabella della distribuzione normale** scopriamo che i percentili  $-z_1$  e  $z_1$  sono rispettivamente: **-1.645** e **1.645**.

Di seguito la rappresentazione della **normale standard**:

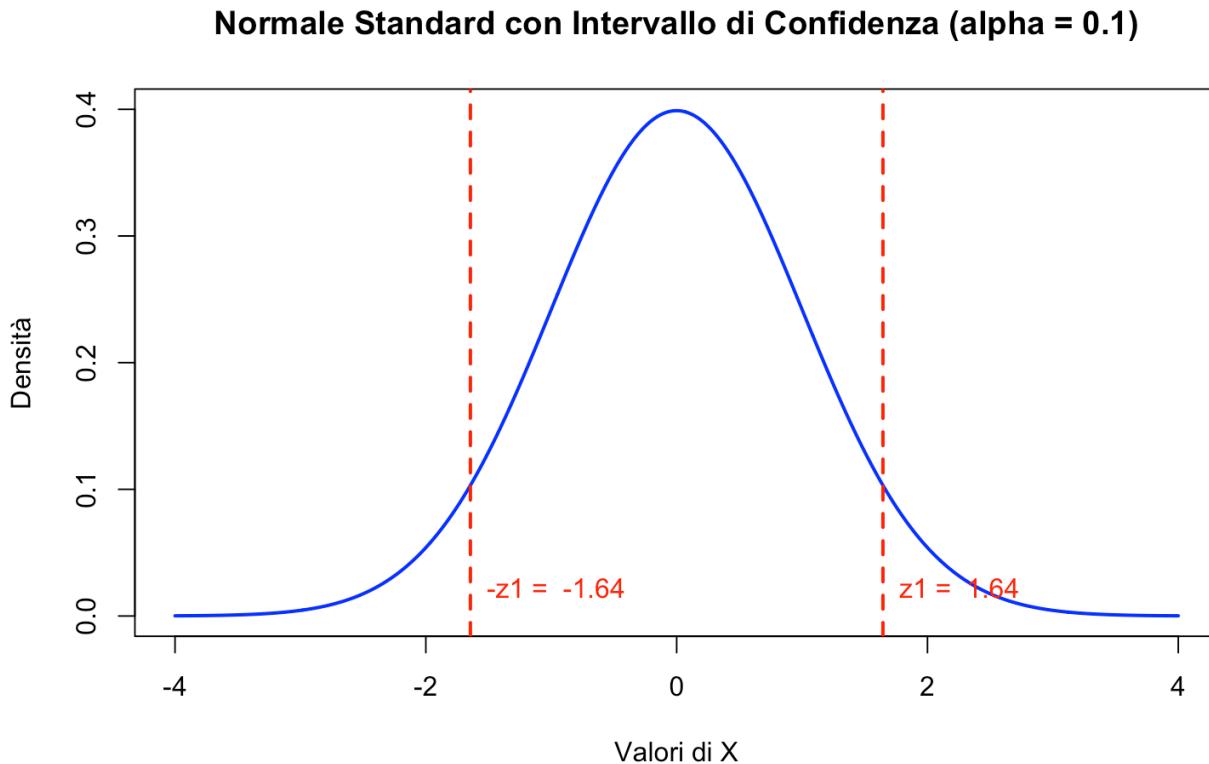


Figura 70 Rappresentazione normale standard

Andando quindi a calcolare l'intervallo di confidenza per mu.  
Questo risulta essere:

- **Limite inferiore: 493.94**
- **Limite superiore: 495.16**

Quindi rapportandolo in percentuale (dividendo di nuovo per n) otteniamo:

- **Limite percentuale inferiore: 0.1568063 => 15.68%**
- **Limite percentuale superiore: 0.1571937 => 15.71%**

Quindi molto molto simile a quella che era la percentuale di partenza che rimane comunque molto bassa.

## 5. Studio popolazioni divise per le variabili utilizzate nel clustering

Nel capitolo precedente abbiamo analizzato il comportamento degli utenti nel caso in cui il provider del servizio si espandesse su scala globale.

Un'analisi interessante potrebbe riguardare il comportamento della variabile [churn](#) suddividendo le popolazioni in base alle variabili utilizzate nel [clustering](#) e confrontandole tra loro.

Questo permetterebbe di confermare ulteriormente l'impatto di queste variabili sulla probabilità di abbandono.

Le feature considerate per l'analisi sono:

1. [Status](#)
2. [Complains](#)
3. [Seconds of use](#)

Tuttavia, lo studio delle popolazioni sulla base della variabile Status risulterebbe poco utile. Inoltre, la variabile Seconds of use non segue nessuna distribuzione.

Per questo motivo ci concentreremo sulla seguente variabile:

**Complains:** confronto tra le popolazioni generate in base al tasso di abbandono (churn) degli utenti che hanno espresso una lamentela rispetto a quelli che non l'hanno fatto.

### 5.1.1. Confronto delle popolazioni sulla base della feature Complains

Per proseguire con l'analisi, suddividiamo i dati in base alla variabile Complains:

- **0: l'utente non ha espresso lamentele.**
- **1: l'utente ha espresso una lamentela.**

Poiché Complains è una variabile binaria, possiamo considerare di modellarla utilizzando una **distribuzione binomiale [20]**, considerando il valore **1 (utente che ha espresso una lamentela)** come “**successo**” e il valore **0 (utente che non ha espresso lamentele)** come “**insuccesso**”.

Per scoprire se la variabile **complains** sia o meno distribuibile secondo una distribuzione binomiale, andiamo a verificare tramite il test del chi-quadro le seguenti ipotesi:

- **H<sub>0</sub>:** La distribuzione non è binomiale
- **H<sub>1</sub>:** La distribuzione è binomiale

Effettueremo il test del chi-quadro sui seguenti valori:

- **Numero di successi: 241**
- **Numero di insuccessi: 2909**
- **Probabilità di lamentela: 7.65%**

Effettuando il test, otteniamo un **p-value pari a 1**, il che significa che la **variabile complains è descrivibile con una distribuzione binomiale**.

Creiamo ora i due sottoinsiemi di dati, in tal modo potremmo definire due popolazioni distinte e analizzare, su scala globale, quale sarebbe la percentuale di abbandono in:

- 1. Una popolazione composta esclusivamente da utenti che esprimono lamentele.**
- 2. Una popolazione composta esclusivamente da utenti che non esprimono lamentele.**

Se dall'analisi emergesse che il tasso di churn è significativamente più alto nella popolazione che ha espresso lamentele rispetto a quella che non lo ha fatto, questo confermerebbe i risultati ottenuti nel capitolo dedicato al **clustering**, ovvero che le lamentele influenzano significativamente il fenomeno di abbandono.

Come analizzato in precedenza, sappiamo che:

Complains	Numero di utenti
Lamentato	<u>241</u>
Non lamentato	<u>2909</u>

Abbiamo quindi diviso il dataset in due set delle dimensioni sopra menzionate.  
Andiamo ora a presentare come si mostrano le popolazioni per ogni dataset.

## **Popolazione di churn con lamentele**

Scopriamo che la porzione di **dataset di utenti che ha espresso lamentele** ha cardinalità **241**.

Andiamo ora a verificare come su questi 241 utenti sono distribuiti gli abbandoni:

Churn	Utenti con lamentele
<b>1: abbandonato</b>	<b>200</b>
<b>0: non abbandonato</b>	<b>41</b>

Sapendo che la variabile di **churn** come visto nel capitolo precedente si può distribuire secondo una distribuzione binomiale, grazie al **teorema del limite centrale [21]** andiamo a creare una distribuzione normale con i seguenti valori:

- **$\mu$  (il valore medio):** per il metodo dei momenti e per il metodo della massima verosimiglianza sappiamo che il valore medio della distribuzione è uno stimatore non distorto per ottenere  $\mu$ . Andando a calcolare il valore atteso (medio) della distribuzione, otteniamo un valore pari a **200**. Il valore atteso di una distribuzione binomiale se noto il parametro **p** darebbe  **$n*p$**  ma in questi casi non lo abbiamo. Andiamo quindi a calcolare quello che è il nostro  $\hat{p}$  empirico che otteniamo effettuando l'operazione **k/n** ovvero **200/241**. Anche in questo caso come nel capitolo precedente sappiamo che  $\mu = n * \hat{p} = k = 200$ .
- **$\sigma$  (deviazione standard):** per il metodo dei momenti e della massima verosimiglianza sappiamo che uno stimatore non distorto di  $\sigma$  è la deviazione standard della distribuzione binomiale. Sappiamo che la deviazione standard per una distribuzione binomiale è  $\sigma = \sqrt{np(1-p)}$  ovviamente anche in questo caso non siamo in possesso del valore teorico di  $p$ , proprio per questo anche qui andiamo ad utilizzare il valore  $\hat{p}$  empirico che otteniamo effettuando l'operazione **k/n**. Effettuando i calcoli otteniamo che il valore di  $\sigma$  è **1,41**.
- **$\alpha: 0,1$  (per un'accuratezza del 90%)**

Otteniamo quindi una distribuzione Normale  **$N(\mu, \sigma^2)$**  ovvero  **$N(200, 1,41)$** .

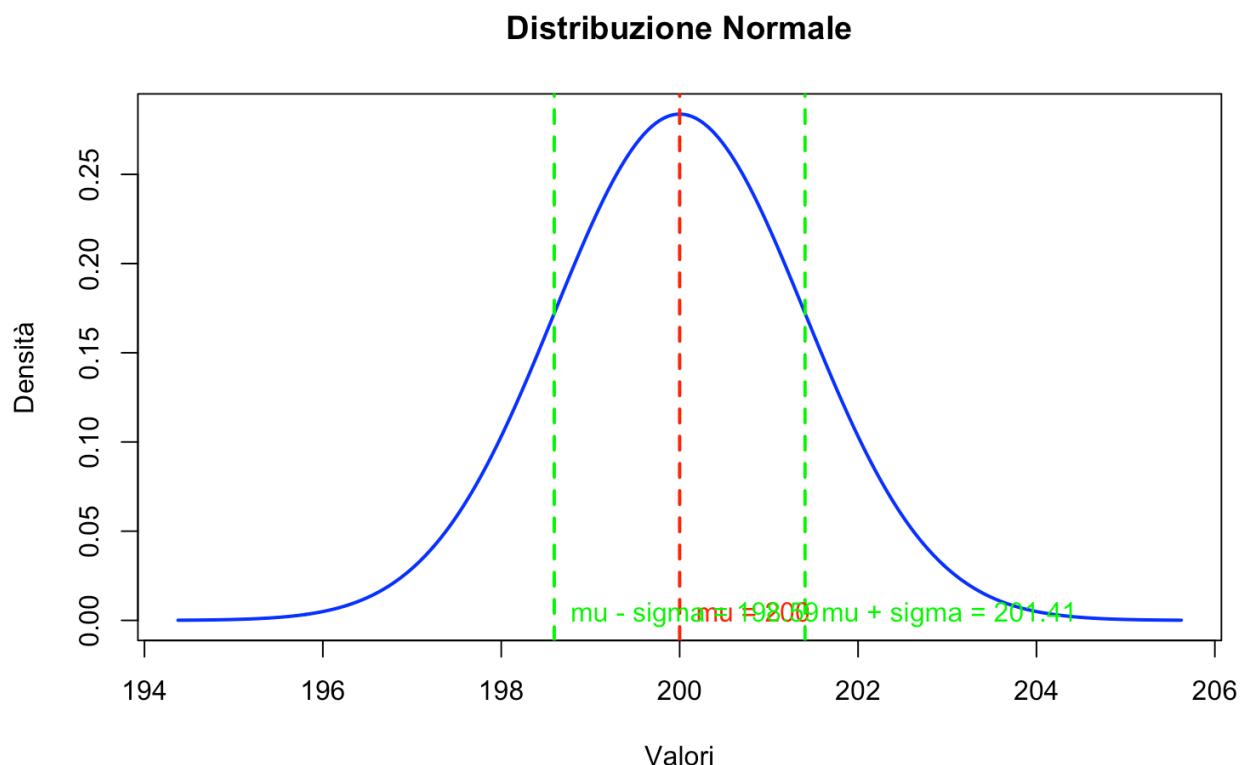


Figura 71 Popolazione di churn su dataset con lamentele

## Popolazione senza lamentele

Scopriamo che la porzione di **dataset di utenti che ha espresso lamentele** ha cardinalità **2909**.

Andiamo ora a verificare come su questi 2909 utenti sono distribuiti gli abbandoni:

Churn	Utenti con lamentele
<b>1: abbandonato</b>	<b><u>295</u></b>
<b>0: non abbandonato</b>	<b><u>2614</u></b>

Sapendo che la variabile di churn come visto nel capitolo precedente si può distribuire secondo una distribuzione binomiale, grazie al **teorema del limite centrale [21]** andiamo a creare una distribuzione normale con i seguenti valori:

- **$\mu$  (il valore medio):** per il metodo dei momenti e per il metodo della massima verosimiglianza sappiamo che il valore medio della distribuzione è uno stimatore non distorto per ottenere  $\mu$ . Andando a calcolare il valore atteso (medio) della distribuzione, otteniamo un valore pari a **295**. Il valore atteso di una distribuzione binomiale se noto il parametro **p** darebbe  **$n*p$**  ma in questo caso non lo abbiamo. Andiamo quindi a calcolare quello che è il nostro  $\hat{p}$  empirico che otteniamo effettuando l'operazione **k/n** ovvero **295/2909**. Anche in questo caso come nel capitolo precedente sappiamo che  $\mu = n * \hat{p} = k = 295$ .
- **$\sigma$  (deviazione standard):** per il metodo dei momenti e della massima verosimiglianza sappiamo che uno stimatore non distorto di  $\sigma$  è la deviazione standard della distribuzione binomiale. Sappiamo che la deviazione standard per una distribuzione binomiale è  $\sigma = \sqrt{np(1-p)}$  ovviamente anche in questo caso non siamo in possesso del valore teorico di  $p$ , proprio per questo anche qui andiamo ad utilizzare il valore  $\hat{p}$  empirico che otteniamo effettuando l'operazione **k/n**. Effettuando i calcoli otteniamo che il valore di  $\sigma$  è **1.13**.
- **$\alpha: 0.1$  (per un'accuratezza del 90%)**

Otteniamo quindi una distribuzione Normale  **$N(\mu, \sigma^2)$**  ovvero  **$N(295, 1.13)$** .

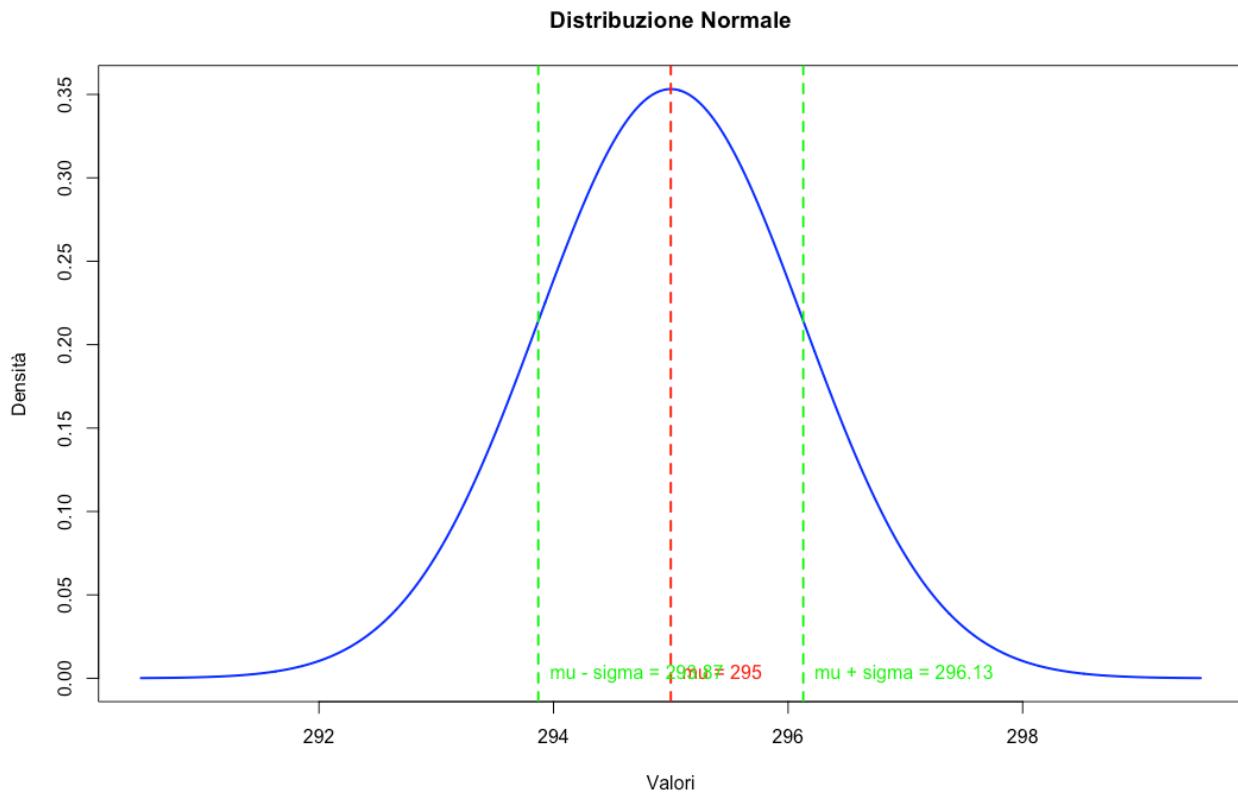


Figura 72 Popolazione di churn su dataset senza lamentele

## Confronto tra le popolazioni

Prima di rappresentare il confronto tra le popolazioni sembra abbastanza evidente che dato che  $\hat{p}$ (tasso di abbandono nella popolazione con lamentele) = 0.83 è strettamente maggiore di  $\hat{p}$ (tasso di abbandono nella popolazione senza lamentele) 0.11.

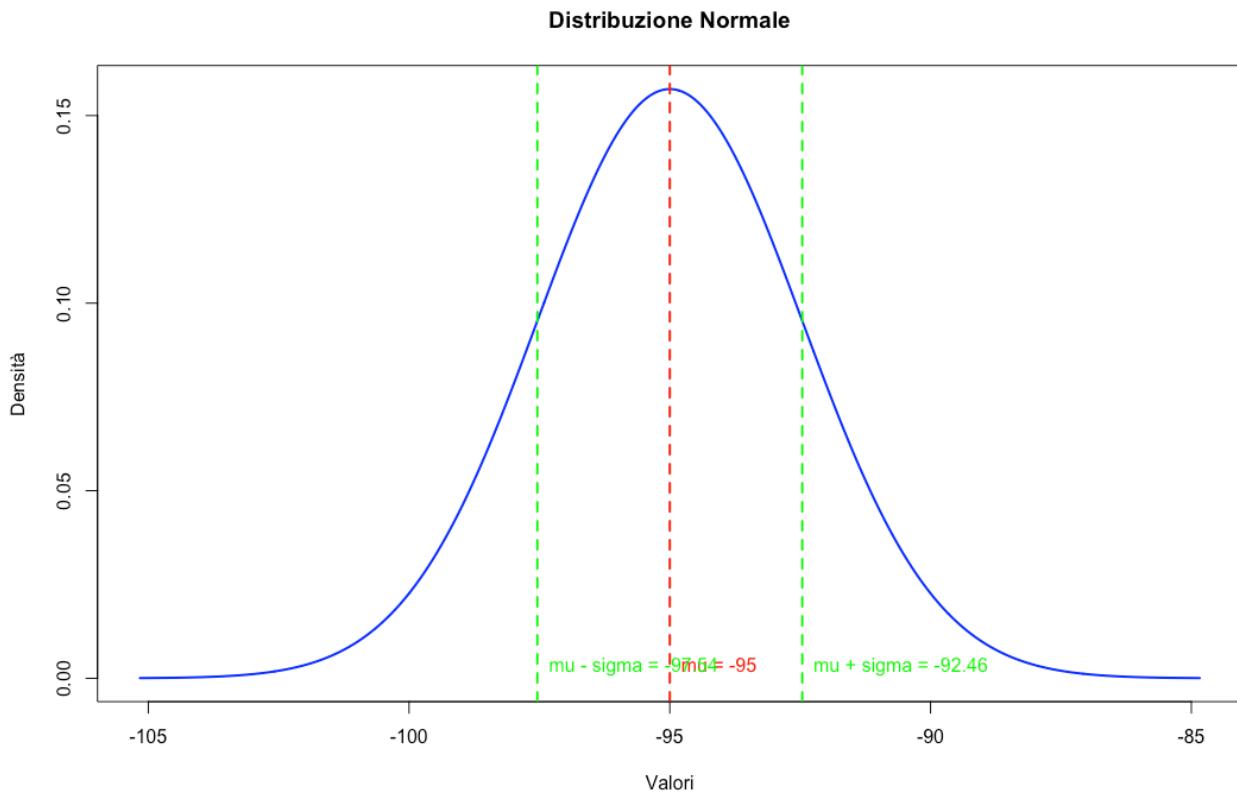
Ora sappiamo che le due popolazioni siano distribuite secondo una Binomiale e calcoliamo l'intervallo di confidenza per  $\hat{p}_1 - \hat{p}_2$ , dove:

- $\hat{p}_1$  è la probabilità che un utente che ha espresso una lamentela abbandoni il servizio (**0.83**).
- $\hat{p}_2$  è la probabilità che un utente che non ha espresso una lamentela abbandoni il servizio (**0.11**).

Abbiamo determinato:

- **$\mu$  (il valore medio):** per il metodo dei momenti e per il metodo della massima verosimiglianza sappiamo che il valore medio della distribuzione è uno stimatore non distorto per ottenere  $\mu$ . Dato che stiamo parlando di un confronto tra popolazioni abbiamo che  $\mu_D = n_1 p_1 - n_2 p_2$ . Di conseguenza  $\mu = k_1 - k_2$ . Abbiamo quindi che  $\mu = -95$ .
- **$\sigma$  (deviazione standard):** per il metodo dei momenti e della massima verosimiglianza sappiamo che uno stimatore non distorto di  $\sigma$  è la deviazione standard della distribuzione binomiale. Sappiamo che la deviazione standard in questo caso è  $\sigma_D = \sqrt{n_1 p_1 (1 - p_1) + n_2 p_2 (1 - p_2)}$  e quindi uguale a **2.54**.
- **$\alpha: 0.1$  (per un'accuratezza del 90%)**

Otteniamo la seguente normale:



*Figura 73 Rappresentazione normale ottenuta con confronto popolazioni*

## 6. Generazione dei Dati Sintetici con LLM con scripting Few-shot con anomalie

Per generare i dati sintetici, è stato utilizzato un Large Language Model (specificare quale modello, ad esempio GPT-4). Il prompt engineering è stato sviluppato per ottenere dati sintetici con caratteristiche simili a quelle del dataset reale. Le feature principali sono state mantenute per garantire coerenza tra i due set di dati.

Questo è stato il messaggio inviato a **ChatGpt4-o** per generare il dataset:

“

Ciao,

*ti chiedo di generare un dataset sintetico simile al dataset scelto per un progetto di statistica e analisi dei dati. Il mio progetto si basa sul dataset Iranian Churn. Il dataset Iranian Churn è stato raccolto casualmente dal database di una compagnia telefonica iraniana nell'arco di 12 mesi. Contiene 3.150 righe di dati, ciascuna rappresentante un cliente, con informazioni distribuite su 14 colonne.*

- **Call Failures (Fallimenti di Chiamata)**: numero di fallimenti di chiamata del fruitore
- **Complains (Lamentela)**: Lamentele riportate dal fruitore
- **Subscription Length (Durata della sottoscrizione)**: totale mesi di fruizione del servizio
- **Charge Amount (Importo addebitato)**: fascia di prezzo del servizio scelto
- **Seconds of Use (Secondi di Utilizzo)**: totale secondi di chiamate
- **Frequency of use (Frequenza di Utilizzo)**: numero totale di chiamate da parte del fruitore del servizio
- **Frequency of SMS (Frequenza di SMS)**: numero totale di messaggi di testo da parte del fruitore del servizio
- **Distinct Called Numbers (Numeri Chiamati Distinti)**: numero totale di chiamate distinte da parte del fruitore del servizio
- **Age Group (Gruppo di Età)**: gruppo d'età a cui appartiene il fruitore del servizio
- **Tariff Plan (Piano Tariffario)**: piano tariffario del servizio
- **Status (Stato)**: stato dell'attivazione del servizio
- **Age (Età)**: Età del fruitore del servizio
- **Churn (Abbandono)**: Abbandono del servizio
- **Customer Value (Valore del Cliente)**: il valore calcolato del cliente

*Di seguito ti invio le prime 20 righe del dataset per farti un'idea di come è strutturato [... prime 20 righe del dataset...]*

*Inoltre, ecco la suddivisione delle variabili in quantitative e qualitative:*

*Le variabili quantitative di questo dataset sono:*

- **Call Failures:** variabile numerica intera
- **Subscription Length:** variabile numerica intera
- **Charge Amount:** variabile ordinale (quantitativa discreta) (0: importo più basso, 9: importo più alto)
- **Seconds of Use:** variabile numerica intera
- **Frequency of use:** variabile numerica intera
- **Frequency of SMS:** variabile numerica intera
- **Distinct Called Numbers:** variabile numerica intera
- **Age:** variabile numerica intera
- **Age Group:** quantitativa discreta, rappresenta gruppi di età (1: età più giovane, 5: età più anziana)

Variabili qualitative del dataset

Le variabili qualitative di questo dataset vengono espresse tramite dati binari e sono le seguenti:

- **Complains:** variabile binaria (0: Nessuna lamentela, 1: lamentela)
- **Tariff Plan:** variabile binaria (1: Pay to go, 2: Pagamento contrattuale)
- **Status:** variabile binaria (1: Attivo, 2: Non attivo)
- **Churn:** variabile binaria (0: Non abbandonato il servizio, 1: Abbandonato il servizio)

Ti fornisco inoltre i nomi delle colonne:

"Call..Failure" "Complains" "Subscription..Length" "Charge..Amount" "Seconds.of.Use"  
 "Frequency.of.use" "Frequency.of.SMS" "Distinct.Called.Numbers" "Age.Group"  
 "Tariff.Plan" "Status" "Age" "Customer.Value" "Churn".

In fine ti chiedo di inserire qualche anomalia(outlier) in modo che sia più reale possibile.

”

Una volta generato il dataset potremmo andare ad effettuare un'analisi univariata per capire se effettivamente i dati sintetici mantengono dei comportamenti simili a quelli non sintetici.

## 6.1. Analisi univariata Dataset Sintetico

### 6.1.1. Call Failures

La feature “[Call Failures](#)” del dataset generato sinteticamente ha prodotto i seguenti risultati:

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Call Failures” risulta pari a **4.55** (nel dataset reale aveva un valore di **7.63** ciò dimostra che abbiamo all’effettivo valori molto diversi tra le due variabili).
- **Mediana campionaria [2]**: La mediana è pari a **4**(nel dataset reale aveva un valore di **7**).
- **Moda campionaria [3]**: La moda è pari a **4** (nel dataset reale aveva un valore di **0**).

Il fatto che media, moda e mediana siano relativamente vicine nei dati sintetici suggerisce una possibile simmetria nella distribuzione. Tuttavia, analizzando più a fondo, emerge che la **media campionaria** è leggermente maggiore rispetto alla mediana e alla moda. Questo potrebbe indicare una leggera asimmetria positiva.

Di seguito un boxplot della variabile *Call Failures Sintetica* ci permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.

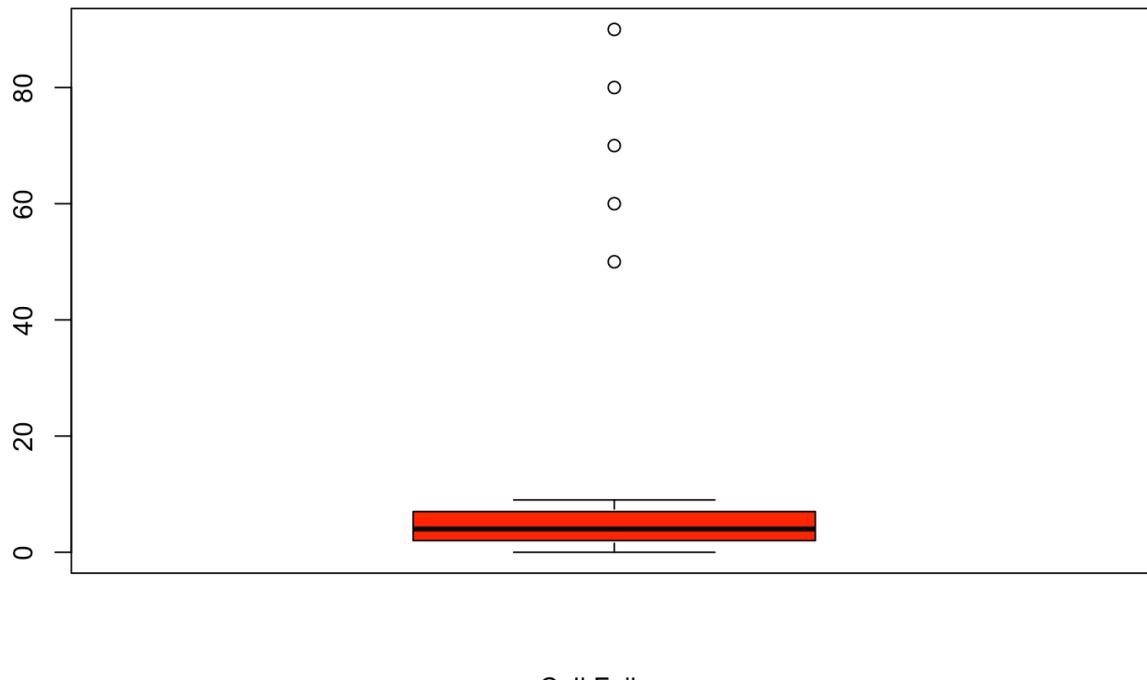


Figura 74 Boxplot Call Failures Sintetico

Possiamo notare dall'immagine che ci sono solo 5 outliers, di fatti, seppur abbiamo chiesto all'AI di inserire delle anomalie per rendere i dati più simili ad un caso reale, constatiamo che comunque il dataset generato presenta dei valori in un range sicuro inserendo qualche anomalia.

Di seguito l'elenco degli outliers del dataset: **50, 60, 70, 80, 90**.

Tramite poi una funzione apposita confermiamo che il **primo quartile** è **2** mentre il **terzo quartile** è **7**.

Inoltre, abbiamo il **minimo** uguale a **0.00** ed un **massimo** uguale a **90.00**.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] delle chiamate fallite dei fruitori generati sinteticamente.

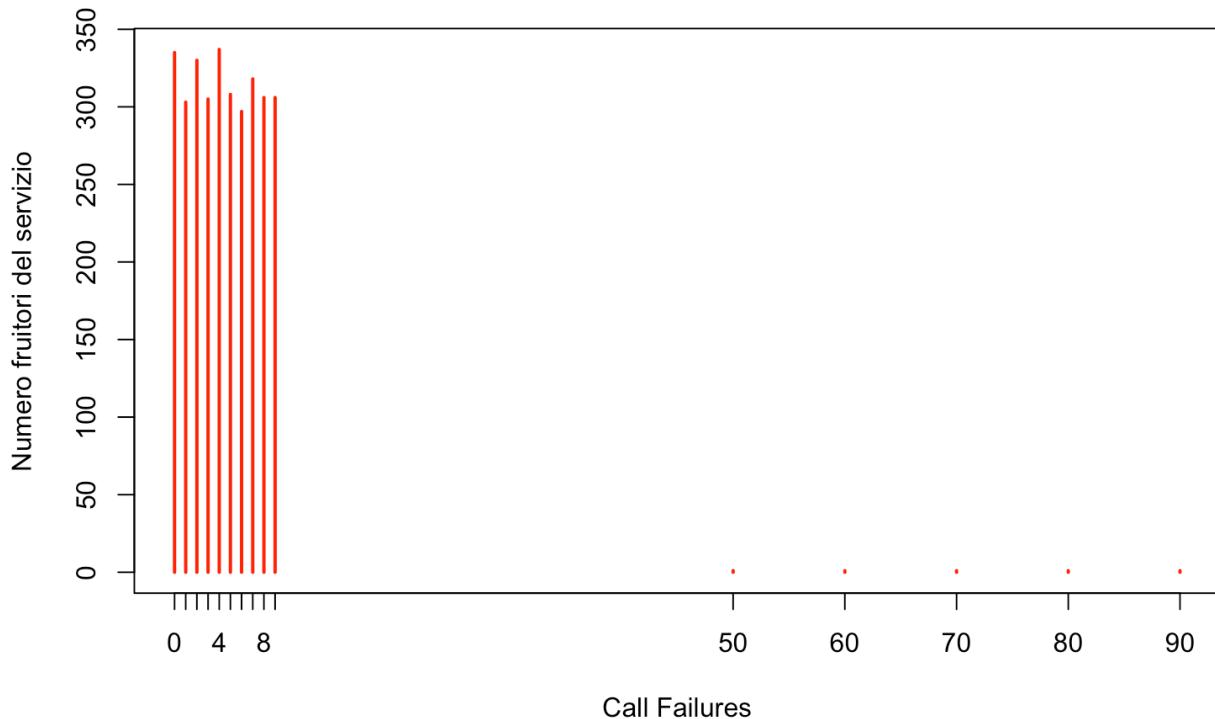


Figura 75 Istogramma Call Failures Sintetico

Un istogramma della variabile *Call Failures* mostra la frequenza assoluta [9] dei fallimenti di chiamata per ciascun valore osservato. Le ascisse rappresentano il numero di fallimenti di chiamata, mentre le ordinate indicano la quantità di utenti corrispondenti.

Il grafico conferma una asimmetria di distribuzione ed inoltre notiamo come l'intelligenza artificiale abbia aggiunto degli outliers molto alti che vanno a creare la coda di destra della distribuzione.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza [4]: 15.39**
- **Deviazione standard [5]: 3.92**
- **Coefficiente di variazione [6]: 86.25%**

L'elevato coefficiente di variazione indica una **forte dispersione dei valori** rispetto alla media, segnalando una significativa variabilità nel numero di fallimenti di chiamata tra gli utenti.

Di seguito per avere un maggiore impatto visivo andiamo a vedere come i valori sono distribuiti in un diagramma a torta.

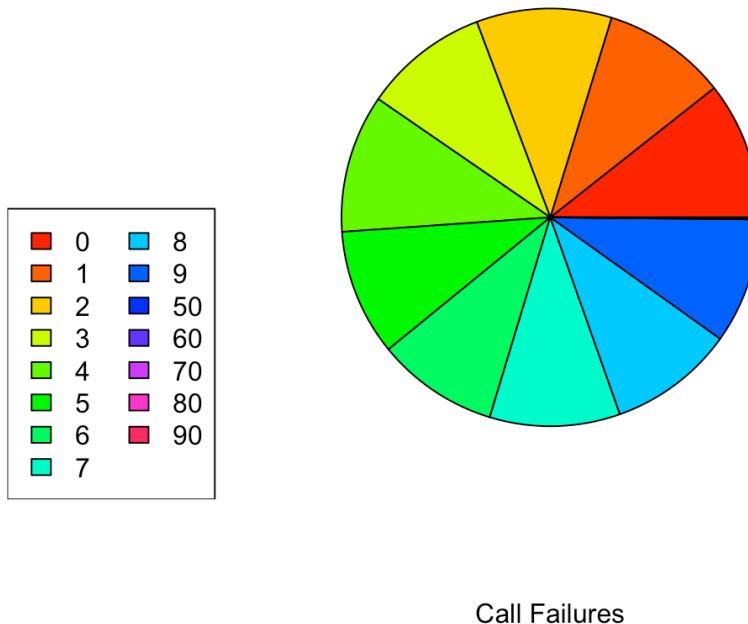


Figura 76 Pie chart Call Failures Sintetico

Notiamo quindi che l'AI durante la creazione del dato sintetico ha equamente distribuito anche i valori outlier che ricordiamo essere 59, 60, 70, 80, 90.

Ciò ci porta a dire che l'AI non è stata in grado di creare per questa feature una variabile simile a quella reale seppure mediante l'utilizzo del few shot sarebbe avrebbe dovuto apprendere in che modo gli outlier si comportassero mediamente nel dataset iniziale.

Per concludere il discorso andiamo a studiare la distribuzione di frequenza.

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 8.38**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 158.60**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza dei fallimenti di chiamata, confermando le caratteristiche sopra descritte.

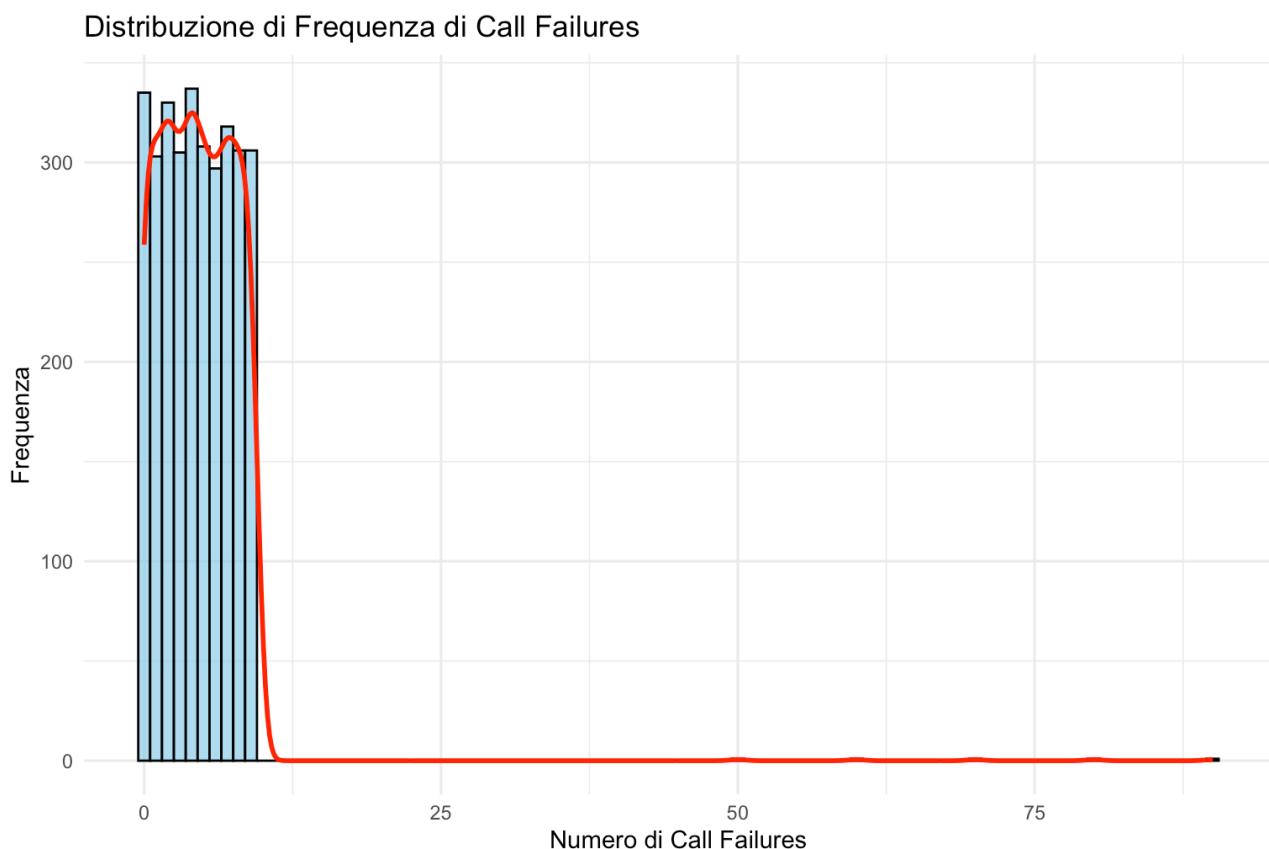


Figura 77 FDE Call of failures Sintetico

## 6.1.2. Complaints

La feature “Complains” (**0: Nessuna lamentela, 1: Lamentela**) generato sinteticamente una volta analizzato ha prodotto i seguenti risultati:

Analizziamo quindi le **frequenze assolute** [9] dei valori assunti dalla variabile Complains:

Valore	Frequenza
0: Nessuna lamentela	<u>1589</u>
1: Lamentela	<u>1561</u>

Andiamo inoltre a vedere le **frequenze relative** [10]:

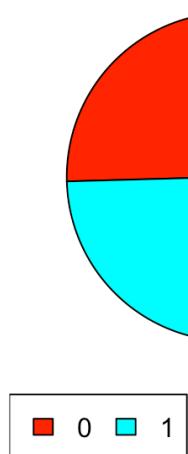
Valore	Frequenza
0: Nessuna lamentela	0.50
1: Lamentela	0.50

Possiamo quindi notare che il **50.44%** dei fruitori non ha espresso alcuna lamentela riguardante il servizio.

Mentre il restante **49.56%** ha espresso una lamentela.

Per avere un’idea più chiara possiamo osservare il diagramma a torta e il diagramma rappresentante la funzione di distribuzione empirica (discreta) sottostanti:

Distribuzione Complains



Funzione di distribuzione empirica (discr.)

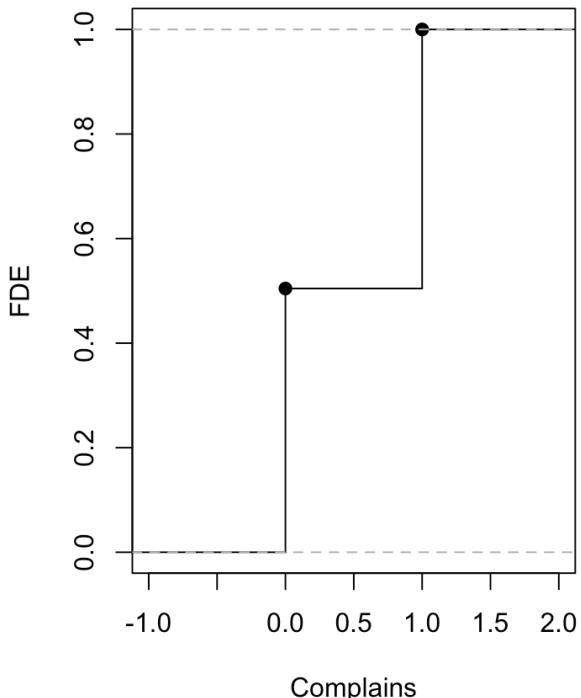


Figura 78 Diagramma a torta e FDE Complains Sintetico

Diremo quindi che anche qui l'intelligenza artificiale durante la creazione del dato sintetico abbia osato poco, di fatti, notiamo che la distribuzione dei valori (**0: Nessuna lamentela, 1: Lamentela**) sia quasi del tutto uniforme.

Anche per questa variabile quindi possiamo dire che questo è un caso irrealistico dato che si trova estremamente lontano dal caso reale.

### 6.1.3. Subscription Length

La feature “Subscription Length” del dataset generato sinteticamente ha prodotto i seguenti risultati:

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Subscription Length” risulta pari a **31.39** (nel dataset reale aveva un valore di **32.5419** ciò dimostra che abbiamo all’effettivo valori molto diversi tra le due variabili).
- **Mediana campionaria [2]**: La mediana è pari a **30** (nel dataset reale aveva un valore di **35**).
- **Moda campionaria [3]**: La moda è pari a **54** (nel dataset reale aveva un valore di **0**).

Analizzando le misure di centralità notiamo che la **media campionaria** è maggiore rispetto alla mediana e alla moda. Questo potrebbe indicare una asimmetria positiva.

Di seguito un boxplot della variabile *Subscription Length Sintetica* ci permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.

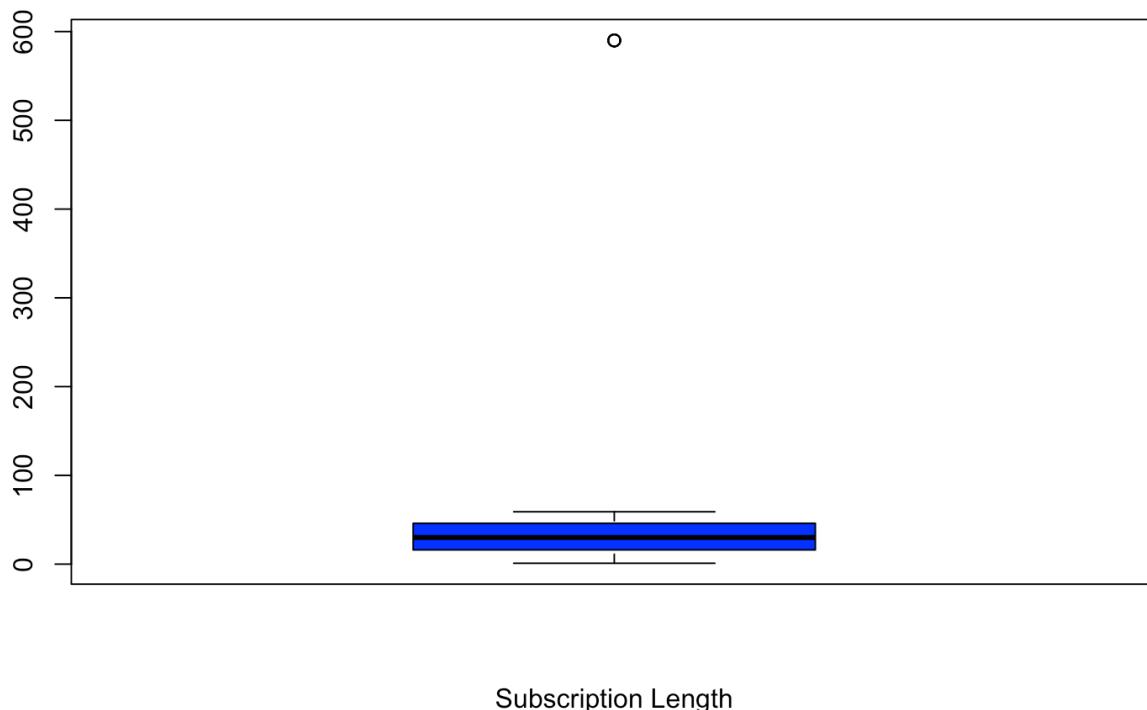


Figura 79 Boxplot Subscription Length Sintetico

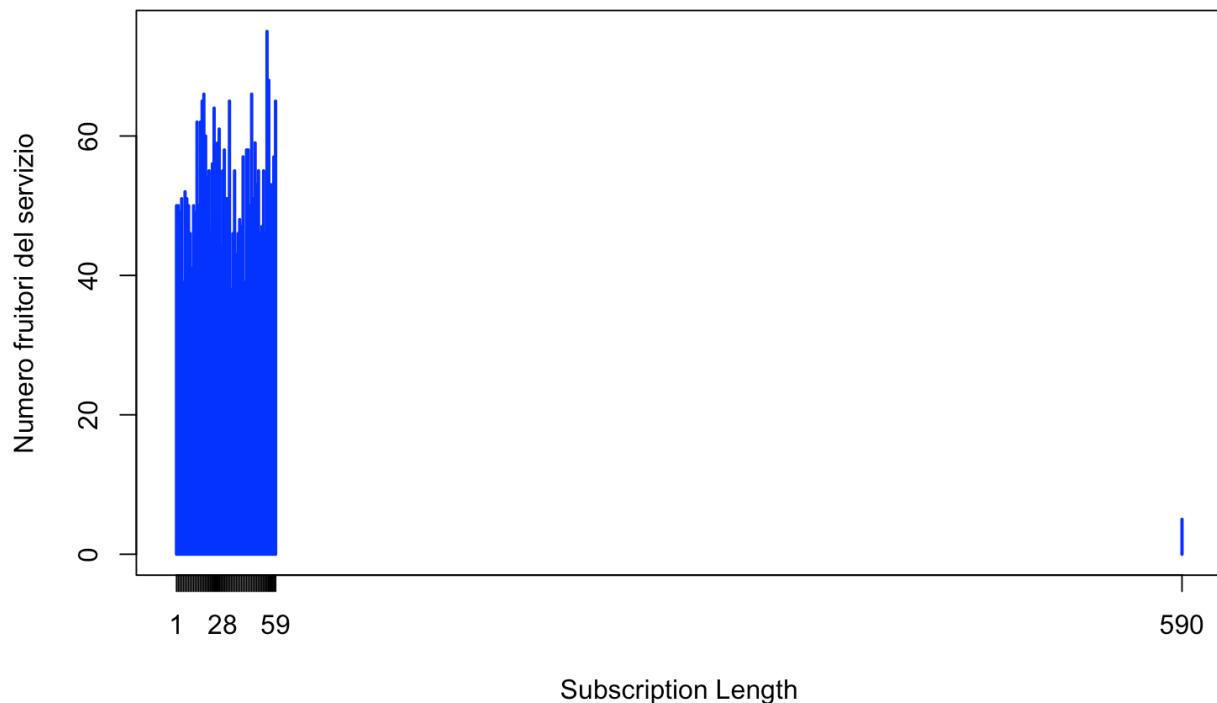
Possiamo notare dall'immagine che c'è un unico outlier, di fatti, seppur abbiamo chiesto all'AI di inserire delle anomalie per rendere i dati più simili ad un caso reale, constatiamo che comunque il dataset generato presenta dei valori in un range sicuro inserendo un'unica anomalia

Di seguito l'elenco degli outliers del dataset: **590**.

Tramite poi una funzione apposita confermiamo che il **primo quartile** è **16** mentre il **terzo quartile** è **45.75**.

Inoltre, abbiamo il **minimo** uguale a **1.00** ed un **massimo** uguale a **590.00**.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] delle lunghezze delle sottoscrizioni al servizio dei fruitori generati sinteticamente.



*Figura 80 Istogramma Subscription Length Sintetico*

Un istogramma della variabile *Subscription Length* mostra la frequenza assoluta [9] delle lunghezze di sottoscrizione per ciascun valore osservato. Le ascisse rappresentano il numero mesi di sottoscrizione, mentre le ordinate indicano la quantità di utenti corrispondenti.

Il grafico conferma una asimmetria di distribuzione ed inoltre notiamo come l'intelligenza artificiale abbia aggiunto un outlier molto alto che va a creare la coda di destra della distribuzione.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza** [4]: **786.73**
- **Deviazione standard** [5]: **28.05**
- **Coefficiente di variazione** [6]: **89.34%**

L'elevato coefficiente di variazione indica una **forte dispersione dei valori** rispetto alla media, segnalando una significativa variabilità nel numero di mesi di sottoscrizione tra gli utenti.

Di seguito per avere un maggiore impatto visivo andiamo a vedere come i valori sono distribuiti in un diagramma a torta.

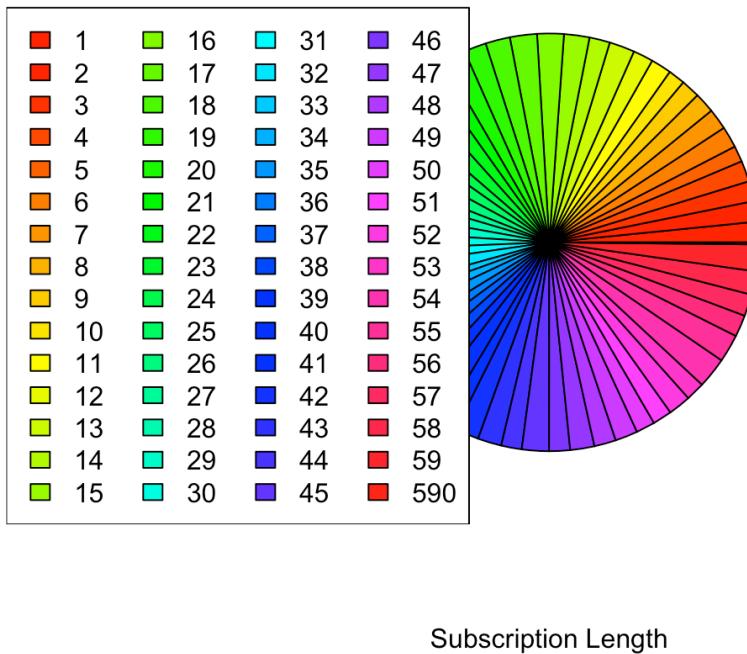


Figura 81 Pie chart Subscription Length Sintetico

Notiamo quindi che l'AI durante la creazione del dato sintetico ha equamente distribuito il numero di mesi partendo dal numero 1 fino al 59 saltando poi direttamente al 590.

Ciò ci porta a dire che l'AI non è stata in grado di creare per questa feature una variabile simile a quella reale seppure mediante l'utilizzo del few shot sarebbe avrebbe dovuto apprendere in che modo gli outlier si comportassero mediamente nel dataset iniziale.

Per concludere il discorso andiamo a studiare la distribuzione di frequenza.

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 12.51**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 250.11**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza dei fallimenti di chiamata, confermando le caratteristiche sopra descritte.

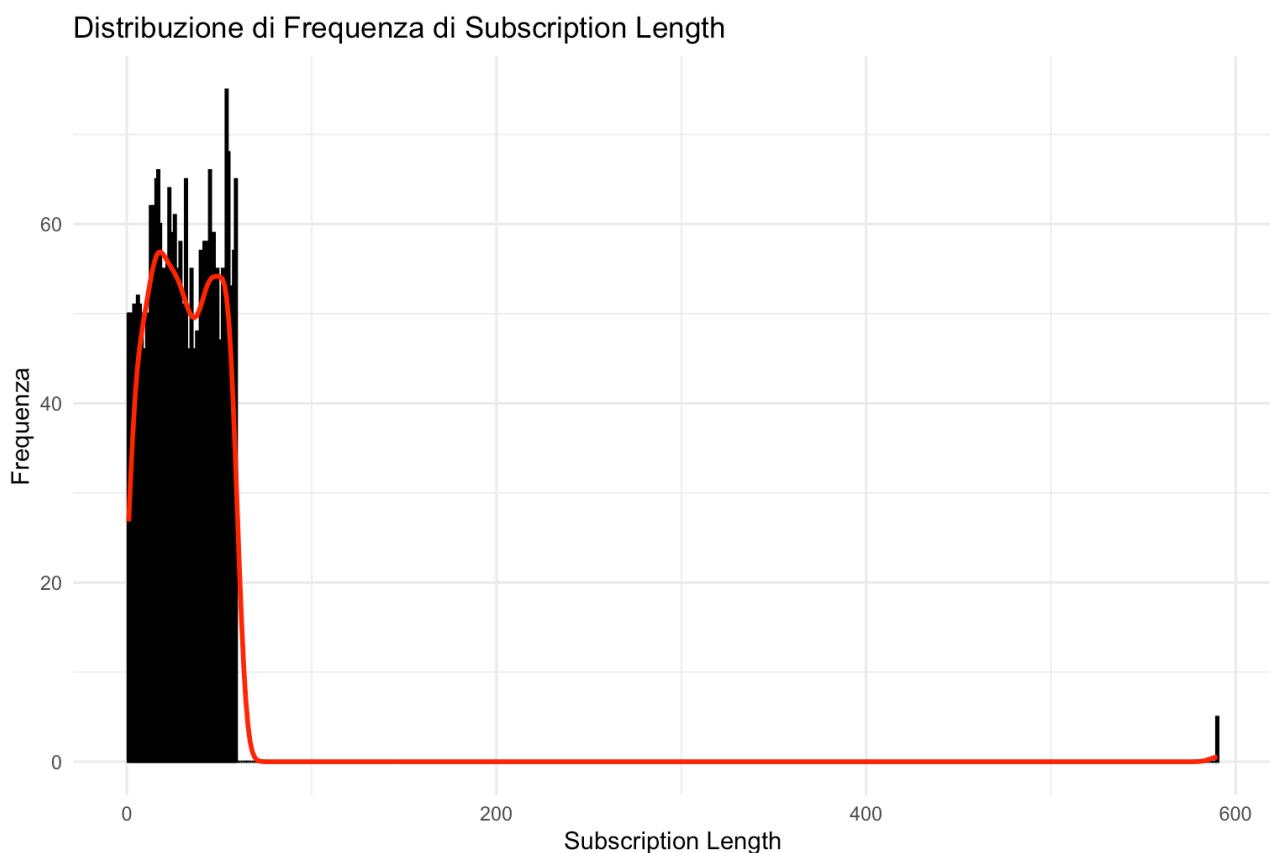


Figura 82 FDE Subscription length Sintetico

### 6.1.4. Charge Amount

La feature “Charge Amount” del dataset generato sinteticamente ha prodotto i seguenti risultati:

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Charge amount” risulta pari a **4.68** (nel dataset reale aveva un valore di **0.94** ciò dimostra che abbiamo all’effettivo valori molto diversi tra le due variabili).
- **Mediana campionaria [2]**: La mediana è pari a **5** (nel dataset reale aveva un valore di **0**).
- **Moda campionaria [3]**: La moda è pari a **9** (nel dataset reale aveva un valore di **0**).

Guardando le misure di centralità notiamo che la **media campionaria** è maggiore rispetto alla mediana e alla moda. Questo potrebbe indicare una asimmetria positiva.

Di seguito un boxplot della variabile *Charge Amount Sintetica* ci permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.

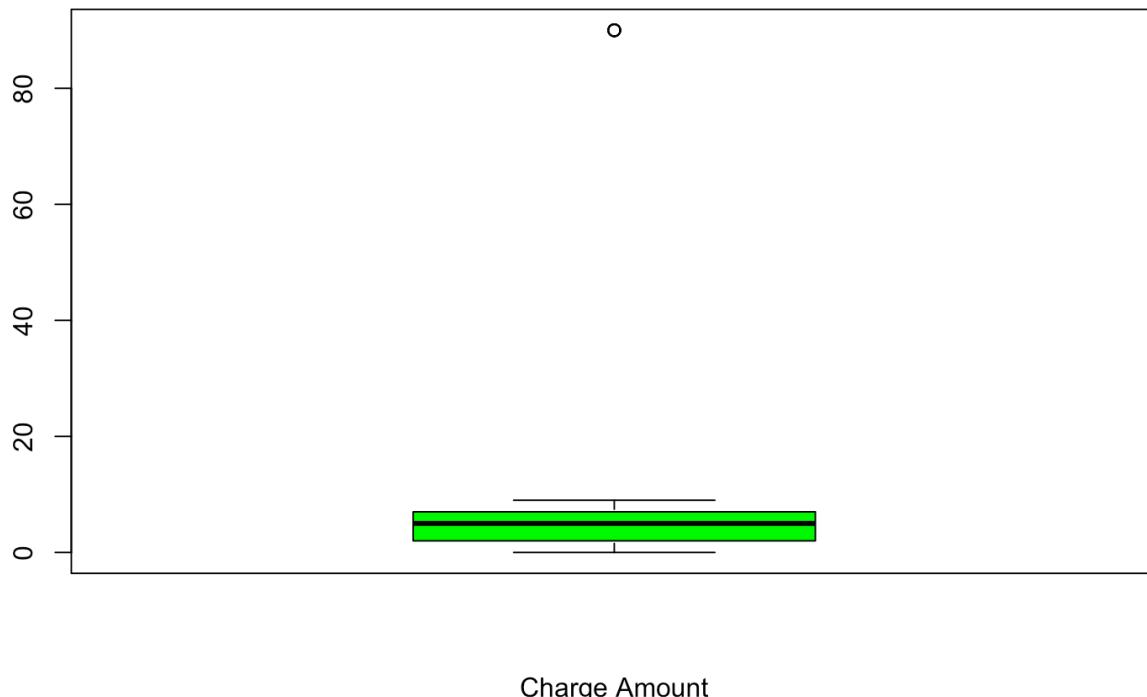


Figura 83 Boxplot Charge Amount Sintetico

Possiamo notare dall'immagine che c'è un unico outlier, di fatti, seppur abbiamo chiesto all'AI di inserire delle anomalie per rendere i dati più simili ad un caso reale, constatiamo che comunque il dataset generato presenta dei valori in un range sicuro inserendo un'unica anomalia

Di seguito l'elenco degli outliers del dataset: **90**.

Tramite poi una funzione apposita confermiamo che il **primo quartile** è **2** mentre il **terzo quartile** è **7**.

Inoltre, abbiamo il **minimo** uguale a **0.00** ed un **massimo** uguale a **90.00**.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] delle lunghezze delle sottoscrizioni al servizio dei fruitori generati sinteticamente.

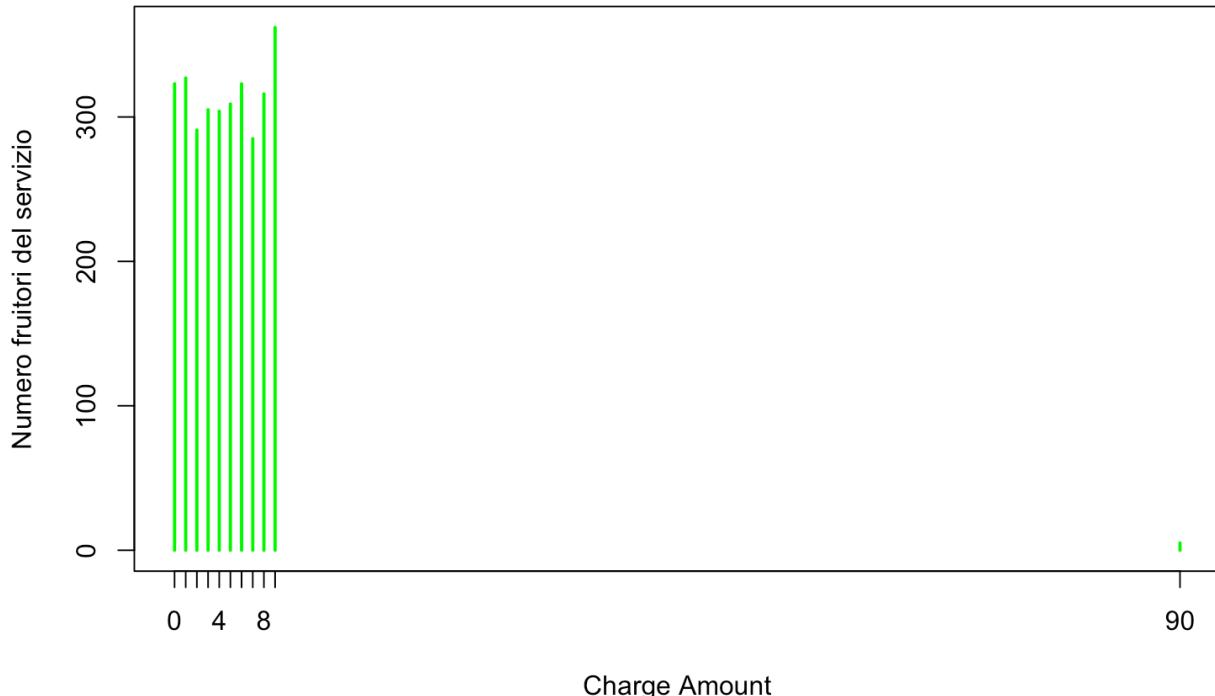


Figura 84 Istogramma Charge amount Sintetico

Un istogramma della variabile *Charge amount* mostra la frequenza assoluta [9] delle spese per ciascun valore osservato. Le ascisse rappresentano il quantitativo speso, mentre le ordinate indicano la quantità di utenti corrispondenti.

Il grafico conferma una asimmetria di distribuzione ed inoltre notiamo come l'intelligenza artificiale abbia aggiunto un outlier molto alto che va a creare la coda di destra della distribuzione. Caso molto simile alla variabile *Subscription Length*, di fatti, notiamo come abbia aggiunto l'unico outlier come 90 che di fatto è un valore, il quale neanche dovrebbe essere considerato nella fascia di spese effettuate dall'utente.

A quanto pare anche passando il dominio di ogni variabile all'AI non è stata in grado di inserire correttamente questo dato. Probabilmente la richiesta fatta di ottenere un dataset con delle anomalie ha portato l'AI a non considerare più il dominio della variabile ma ad andare addirittura molto oltre.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza [4]: 20.12**
- **Deviazione standard [5]: 4.49**
- **Coefficiente di variazione [6]: 95.58%**

L'elevato coefficiente di variazione indica una **forte dispersione dei valori** rispetto alla media, segnalando una significativa variabilità nelle fasce di spese tra gli utenti.

Di seguito per avere un maggiore impatto visivo andiamo a vedere come i valori sono distribuiti in un diagramma a torta.

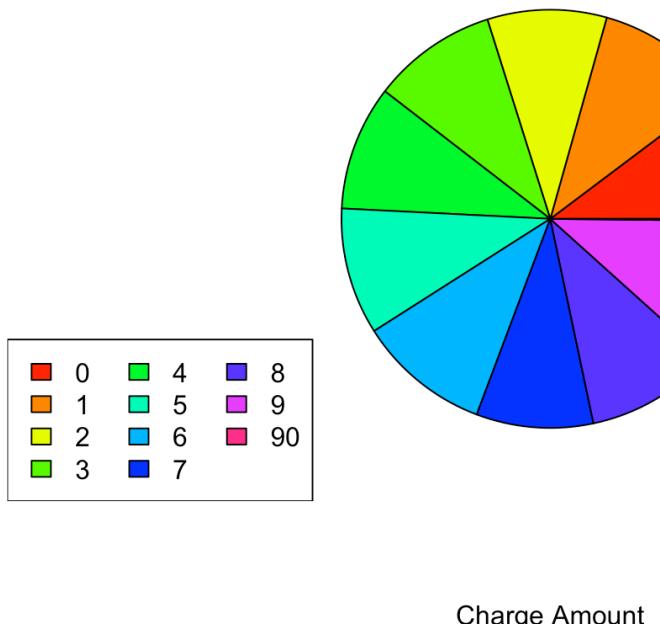


Figura 85 Pie chart Charge amount Sintetico

Notiamo quindi che l'AI durante la creazione del dato sintetico ha equamente distribuito il numero di mesi partendo dal numero 1 fino al 9 (dominio quindi corretto) saltando poi direttamente al 90 il quale valore è totalmente fuori dominio.

Ciò ci porta a dire che l'AI non è stata in grado di creare per questa feature una variabile simile a quella reale seppure mediante l'utilizzo del few shot sarebbe avrebbe dovuto apprendere in che modo gli outlier si comportassero mediamente nel dataset iniziale.

Per concludere il discorso andiamo a studiare la distribuzione di frequenza.

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 10.89**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 208.19**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

---

Il seguente grafico riassume la distribuzione di frequenza dei pagamenti dei fruitori, confermando le caratteristiche sopra descritte.

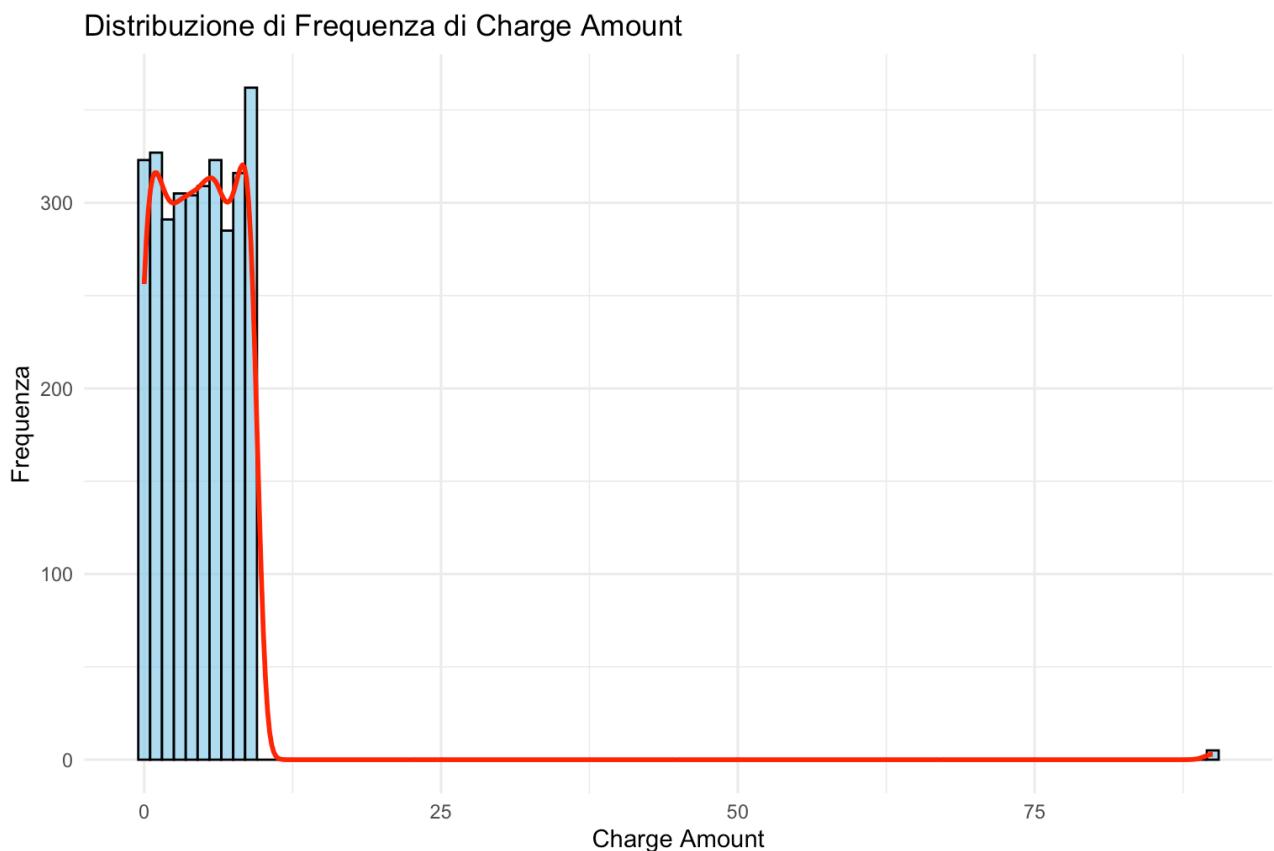


Figura 86 FDE Charge amount Sintetico

### 6.1.5. Second of use

La feature “Seconds of use” del dataset generato sinteticamente, che ricordiamo che per il dataset reale è stato fondamentale per il clustering, ha prodotto i seguenti risultati:

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Seconds of use” risulta pari a **13719.49** (nel dataset reale aveva un valore di **4472.46** ciò dimostra che abbiamo all’effettivo valori molto diversi tra le due variabili).
- **Mediana campionaria [2]**: La mediana è pari a **9917.5** (nel dataset reale aveva un valore di **2990**).
- **Moda campionaria [3]**: La moda è pari a **199960** (nel dataset reale aveva un valore di **0**).

Guardando le misure di centralità notiamo che la **media campionaria** è maggiore rispetto alla mediana. Questo potrebbe indicare una asimmetria positiva.

Di seguito un boxplot della variabile Seconds of use *Sintetica* ci permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.

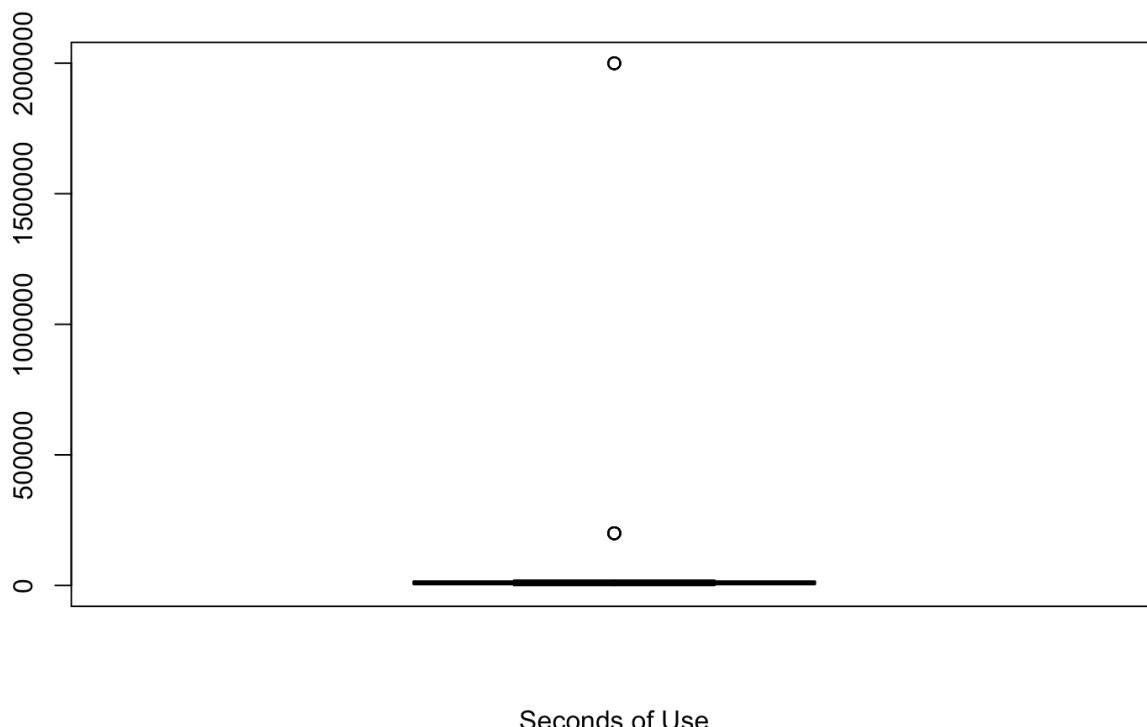


Figura 87 Boxplot Seconds of use Sintetico

Possiamo notare dall'immagine che ci sono due diversi outliers, di fatti, seppur abbiamo chiesto all'AI di inserire delle anomalie per rendere i dati più simili ad un caso reale, constatiamo che comunque il dataset generato presenta dei valori in un range sicuro inserendo un'unica anomalia

Di seguito l'elenco degli outliers del dataset: **199960, 1999600**.

Tramite poi una funzione apposita confermiamo che il **primo quartile** è **4868** mentre il **terzo quartile** è **15256**.

Inoltre, abbiamo il **minimo** uguale a **11** ed un **massimo** uguale a **1999600**.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] delle lunghezze delle sottoscrizioni al servizio dei fruitori generati sinteticamente.

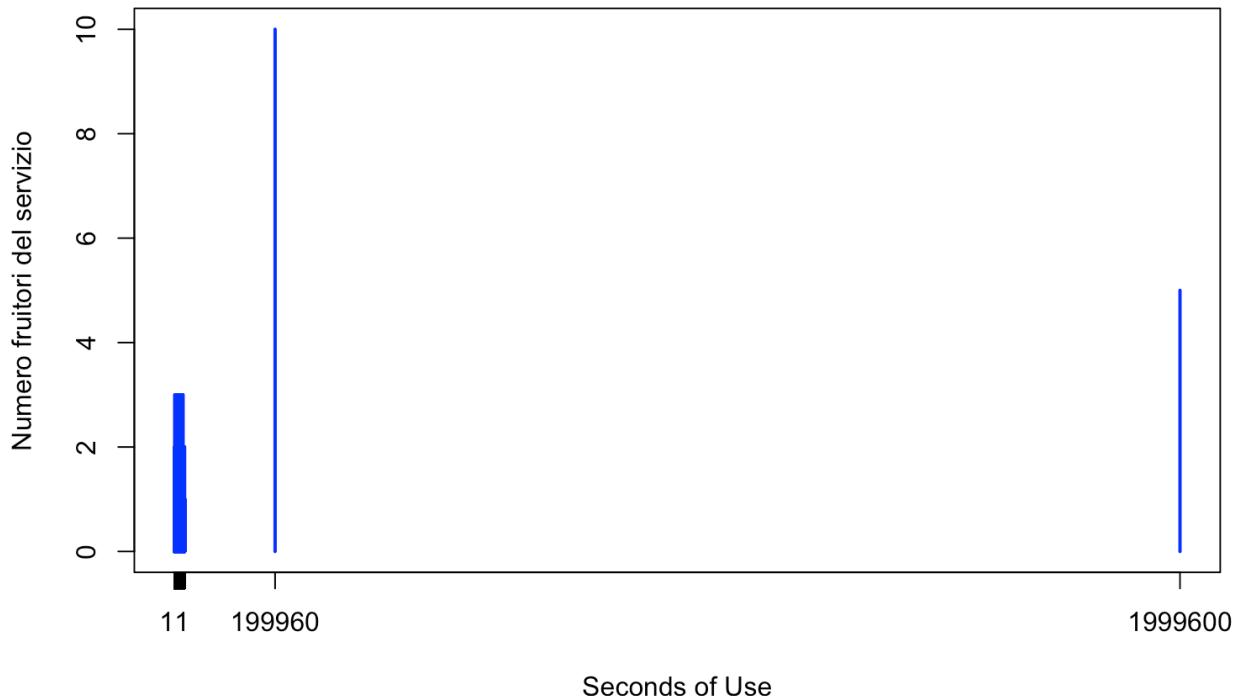


Figura 88 Istogramma Seconds of use Sintetico

Un istogramma della variabile *Seconds of use* mostra la frequenza assoluta [9] dei secondi di utilizzo ciascun valore osservato. Le ascisse rappresentano il numero di secondi, mentre le ordinate indicano la quantità di utenti corrispondenti.

Il grafico conferma una asimmetria di distribuzione ed inoltre notiamo come l'intelligenza artificiale abbia aggiunto un outlier molto alto che va a creare la coda di destra della distribuzione.

Notiamo che ha aggiunto un outlier veramente molto alto con il valore **1999600** che trasformato in giorni sarebbero **23.14** giorni di utilizzo, sicuramente parliamo di un valore eccessivamente alto per essere considerato ragionevole.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza** [4]: **6419750521**
- **Deviazione standard** [5]: **80123.35**

- **Coefficiente di variazione [6]: 584.01%**

L'elevato coefficiente di variazione indica una **forte dispersione dei valori** rispetto alla media, segnalando una significativa variabilità nelle fasce di spese tra gli utenti. Inoltre, il valore così alto dei valori di dispersione è sicuramente dato dall'outlier estremamente grande inserito dall'AI.

Ciò ci porta a dire che l'AI non è stata in grado di creare per questa feature una variabile simile a quella reale seppure mediante l'utilizzo del few shot sarebbe avrebbe dovuto apprendere in che modo gli outlier si comportassero mediamente nel dataset iniziale.

Per concludere il discorso andiamo a studiare la distribuzione di frequenza. I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 24.21861**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 599.4855**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza dei secondi di utilizzo, confermando le caratteristiche sopra descritte.

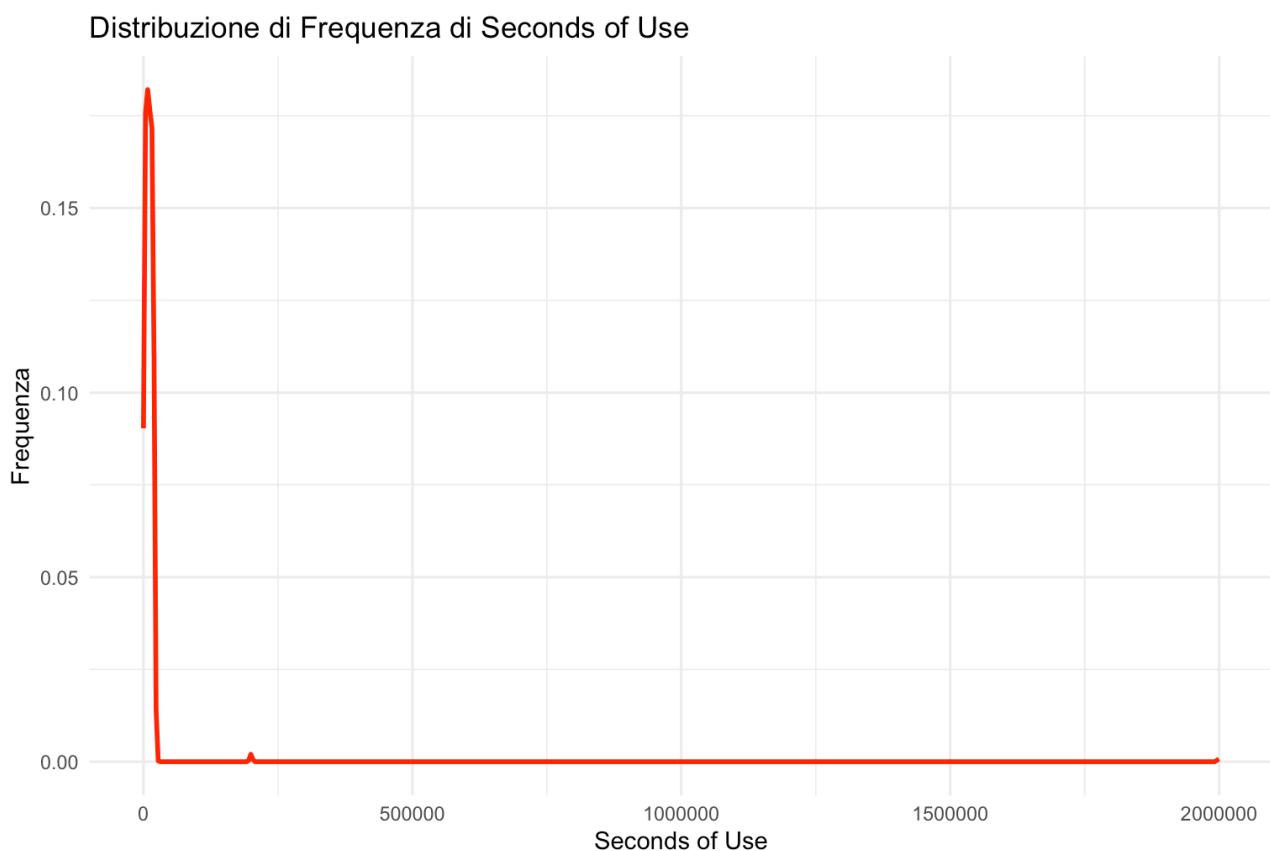


Figura 89 FDE Seconds of use Sintetico

### 6.1.6. Seconds of use intervals

Seconds of use intervals è una Feature aggiunta nel dataset reale per studiare meglio l'enorme quantità di valori espressa dalla variabile Seconds of use.  
 Questo lavoro di suddivisione in 50 intervalli è stato fatto anche per questo dataset sintetico e ha prodotto i seguenti risultati:

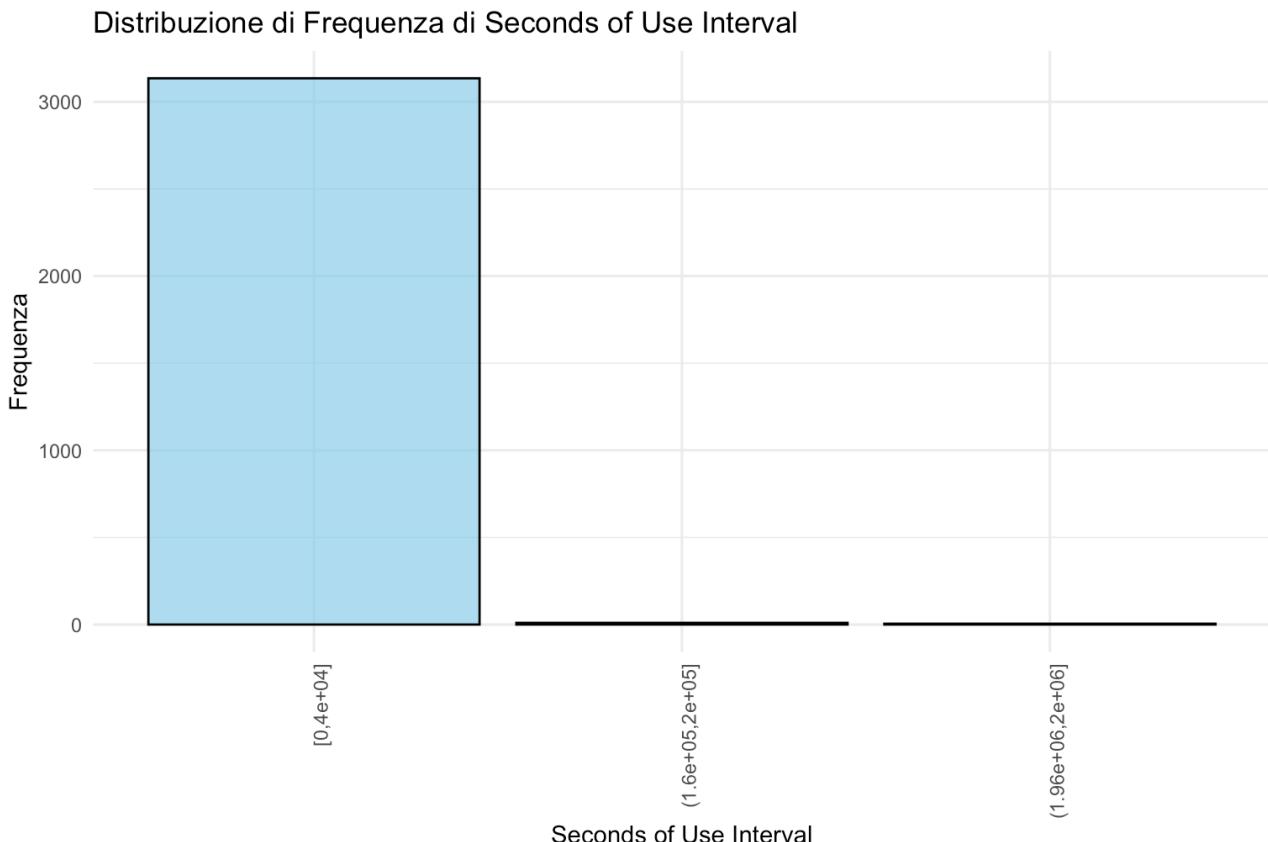


Figura 90 Distribuzione di frequenza Seconds of use Interval

Già dalla sola distribuzione di frequenza notiamo che la quasi totalità di valori creati dall'AI si concentra sull'intervallo  $[0, 4e+04]$ .

Gli altri due intervalli rappresentano gli outliers inseriti dall'AI su richiesta.

Possiamo anche qui dire che la variabile creata non è soddisfacente per simulare un caso reale.

## 6.1.7. Frequency of use

La feature “Frequency of use” del dataset generato sinteticamente ha prodotto i seguenti risultati:

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Frequency of userisulta pari a 155.02 (nel dataset reale aveva un valore di 69.46063 ciò dimostra che abbiamo all’effettivo valori molto diversi tra le due variabili).
- **Mediana campionaria [2]**: La mediana è pari a 153 (nel dataset reale aveva un valore di 53).
- **Moda campionaria [3]**: La moda è pari a 35 (nel dataset reale aveva un valore di 0).

Guardando le misure di centralità notiamo che la **media campionaria** è maggiore rispetto alla mediana e alla moda. Questo potrebbe indicare una asimmetria positiva.

Di seguito un boxplot della variabile *Frequency of use Sintetica* ci permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.

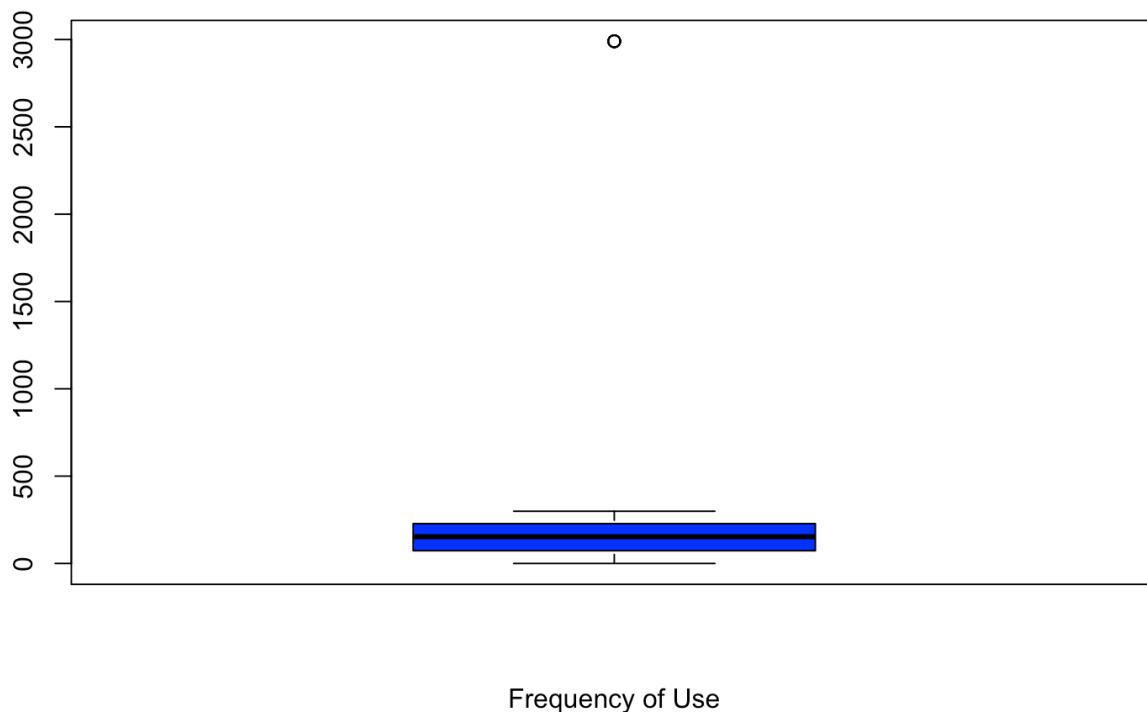


Figura 91 Boxplot Frequency of use Sintetico

Possiamo notare dall'immagine che c'è un unico outlier, di fatti, seppur abbiamo chiesto all'AI di inserire delle anomalie per rendere i dati più simili ad un caso reale, constatiamo che comunque il dato generato presenta dei valori in un range sicuro inserendo un'unica anomalia

Di seguito l'elenco degli outliers del dataset: **2990**.

Tramite poi una funzione apposita confermiamo che il **primo quartile** è **73** mentre il **terzo quartile** è **228**.

Inoltre, abbiamo il **minimo** uguale a **0** ed un **massimo** uguale a **2990**.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] delle lunghezze delle sottoscrizioni al servizio dei fruitori generati sinteticamente.

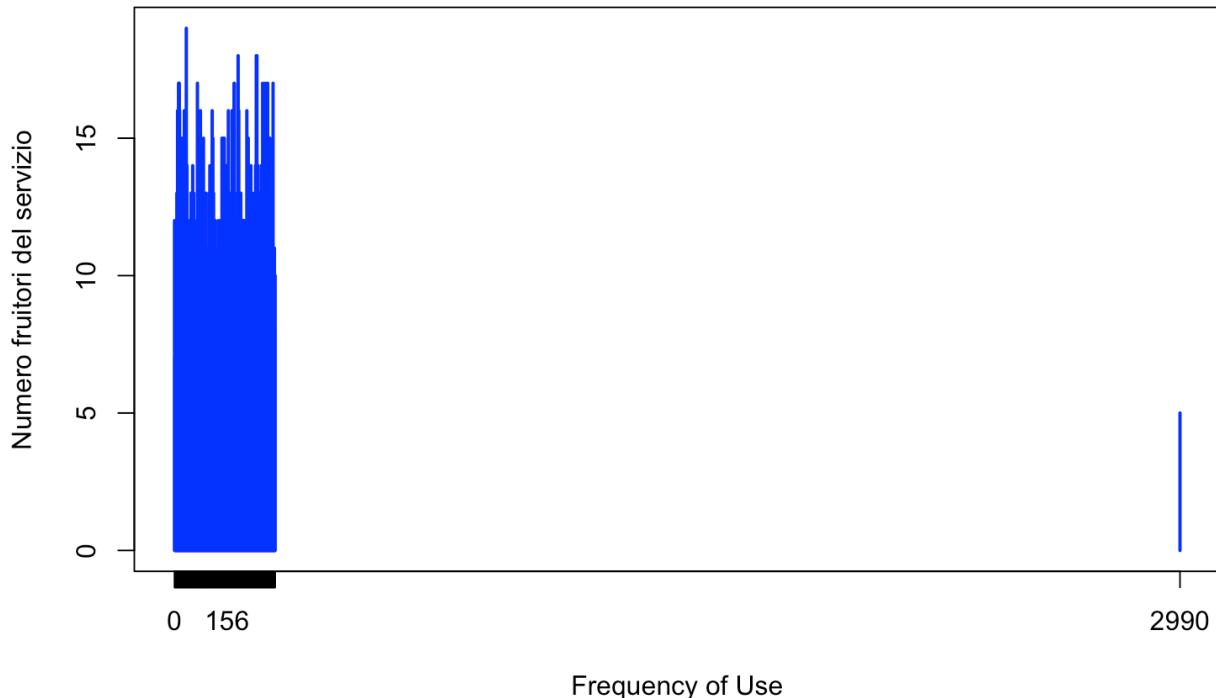


Figura 92 Istogramma Frequency of use Sintetico

Un istogramma della variabile *Frequency of use* mostra la frequenza delle chiamate ciascun valore osservato. Le ascisse rappresentano il quantitativo speso, mentre le ordinate indicano la quantità di utenti corrispondenti.

Il grafico conferma una asimmetria di distribuzione ed inoltre notiamo come l'intelligenza artificiale abbia aggiunto un outlier molto alto che va a creare la coda di destra della distribuzione.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza [4]: 20490.5**

- **Deviazione standard [5]: 143.15**
- **Coefficiente di variazione [6]: 92.34%**

Per concludere il discorso andiamo a studiare la distribuzione di frequenza.

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 12.29**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 244.61**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza di chiamate per ogni fruitore, confermando le caratteristiche sopra descritte.

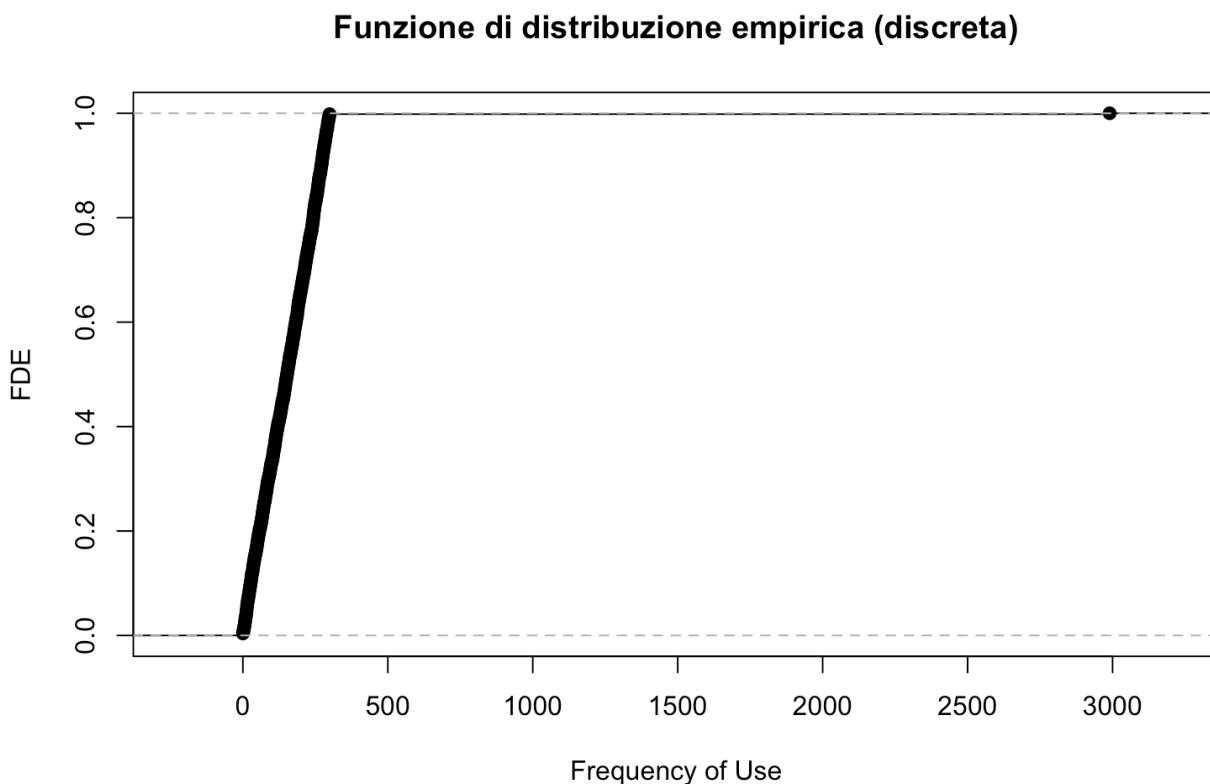


Figura 93 FDE Frequency of use Sintetico

## 6.1.8. Frequency of sms

La feature “[Frequency of sms](#)” del dataset generato sinteticamente ha prodotto i seguenti risultati:

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Frequency of sms” risulta pari a **254.94**(nel dataset reale aveva un valore di **73.18** ciò dimostra che abbiamo all’effettivo valori molto diversi tra le due variabili).
- **Mediana campionaria [2]**: La mediana è pari a **245** (nel dataset reale aveva un valore di **21**).
- **Moda campionaria [3]**: La moda è pari a **111** (nel dataset reale aveva un valore di **0**).

Guardando le misure di centralità notiamo che la **media campionaria** è maggiore rispetto alla mediana e alla moda. Questo potrebbe indicare una asimmetria positiva.

Di seguito un boxplot della variabile *Frequency of sms Sintetica* ci permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.

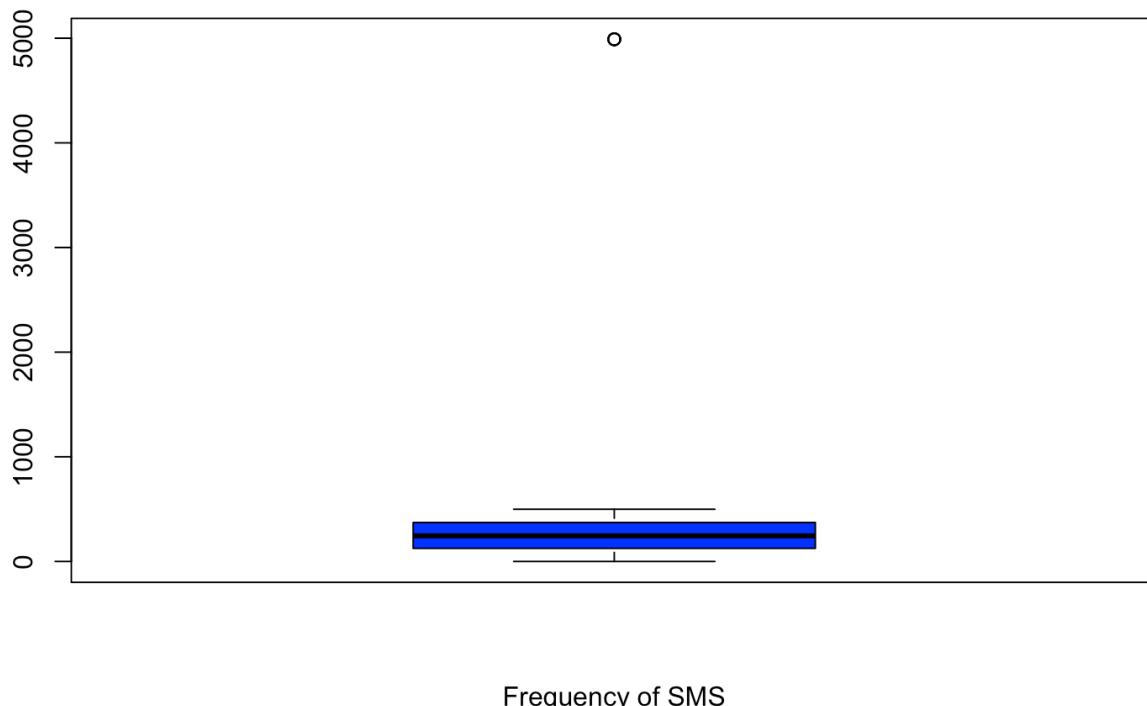


Figura 94 Boxplot Frequency of sms Sintetico

Possiamo notare dall'immagine che c'è un unico outlier, di fatti, seppur abbiamo chiesto all'AI di inserire delle anomalie per rendere i dati più simili ad un caso reale, constatiamo che comunque il dataset generato presenta dei valori in un range sicuro inserendo un'unica anomalia

Di seguito l'elenco degli outliers del dataset: **4990**.

Tramite poi una funzione apposita confermiamo che il **primo quartile** è **125.0** mentre il **terzo quartile** è **371.0**.

Inoltre, abbiamo il **minimo** uguale a **0.0** ed un **massimo** uguale a **4990.0**.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] delle lunghezze delle sottoscrizioni al servizio dei fruitori generati sinteticamente.

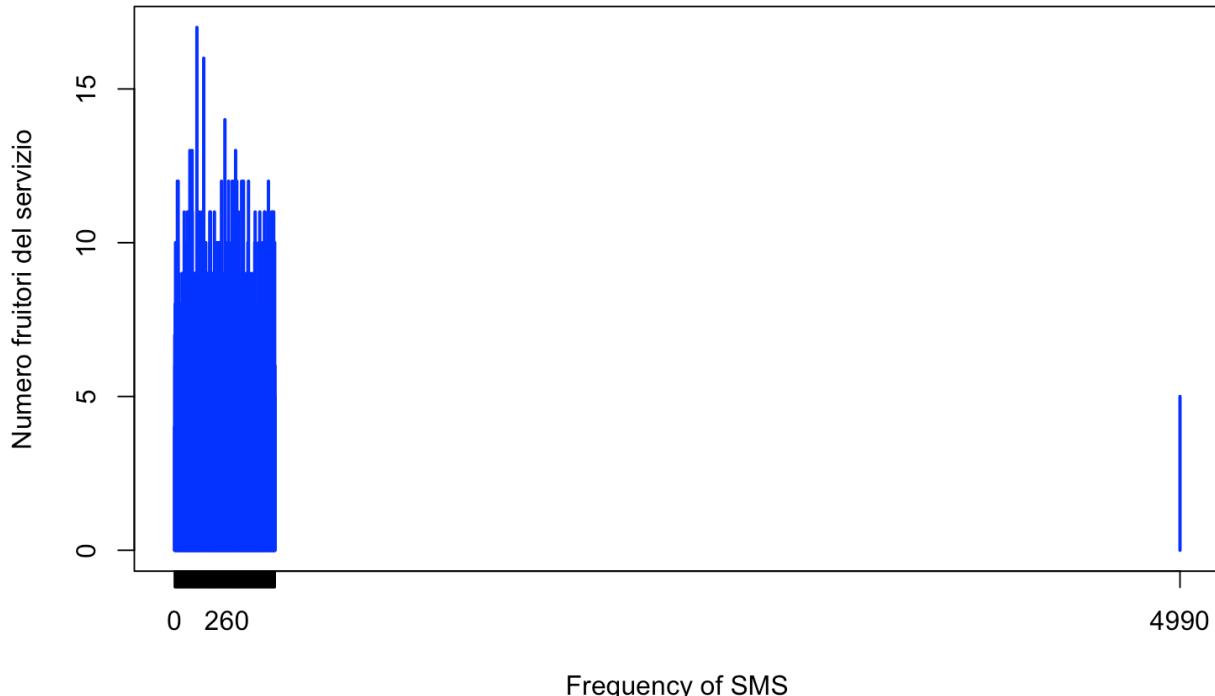


Figura 95 Istogramma Frequency of sms Sintetico

Un istogramma della variabile *Frequency of sms* mostra la frequenza assoluta [9] degli sms inviati per ciascun valore osservato. Le ascisse rappresentano il numero di sms, mentre le ordinate indicano la quantità di utenti corrispondenti.

Il grafico conferma una asimmetria di distribuzione ed inoltre notiamo come l'intelligenza artificiale abbia aggiunto un outlier molto alto che va a creare la coda di destra della distribuzione

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza [4]: 55913.23**
- **Deviazione standard [5]: 236.46**
- **Coefficiente di variazione [6]: 92.75%**

L'elevato coefficiente di variazione indica una **forte dispersione dei valori** rispetto alla media, segnalando una significativa variabilità nelle fasce di spese tra gli utenti.

Notiamo quindi che l'AI durante la creazione del dato sintetico è distribuito in un range che va circa da 0 a 260 per poi fare un salto a 4990.

Per concludere il discorso andiamo a studiare la distribuzione di frequenza.

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness** [7]: 12.72, che conferma l'asimmetria verso destra.
- **Curtosi** [8]: 255.63, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza dei pagamenti dei fruitori, confermando le caratteristiche sopra descritte.

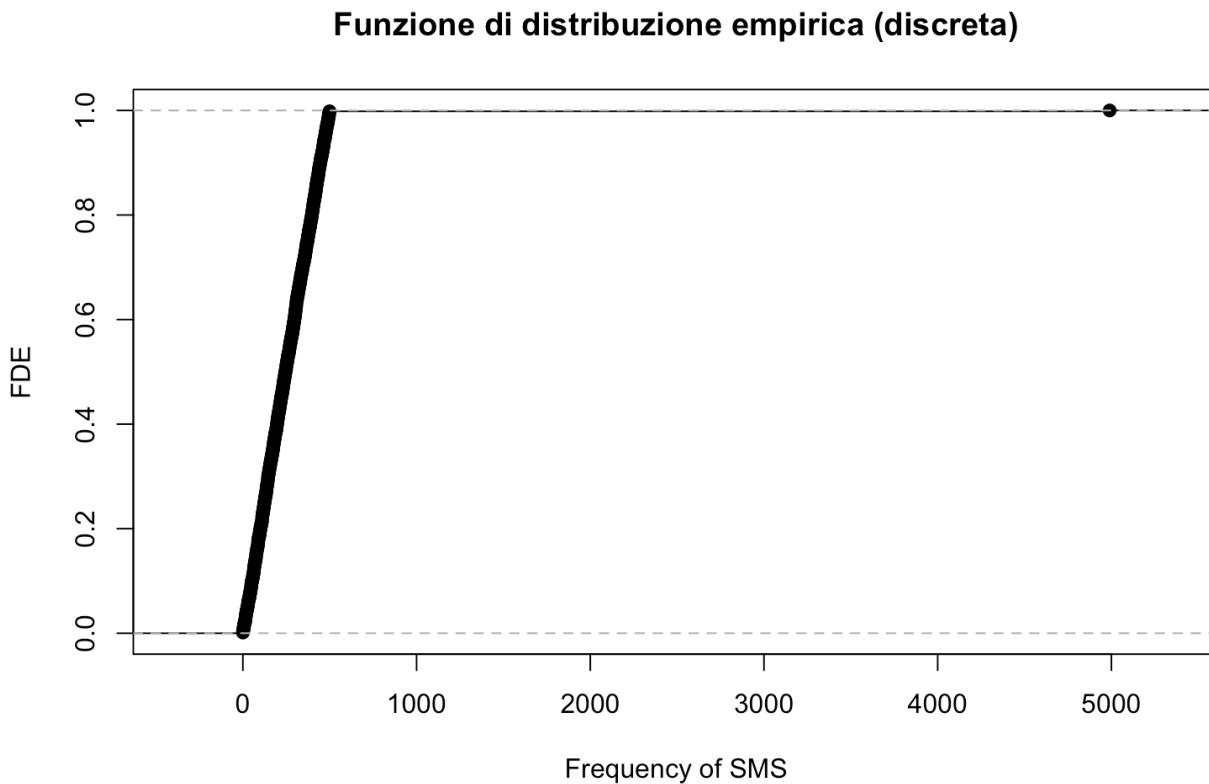


Figura 96 FDE Frequency of sms Sintetico

### 6.1.9. Distinct Call Numbers

La feature “[Distinct Call Numbers](#)” del dataset generato sinteticamente ha prodotto i seguenti risultati:

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Distinct Call Numbers” risulta pari a **24.93238** (nel dataset reale aveva un valore di **23.51** ciò dimostra che abbiamo all’effettivo valori molto diversi tra le due variabili).
- **Mediana campionaria [2]**: La mediana è pari a **24** (nel dataset reale aveva un valore di **21**).
- **Moda campionaria [3]**: La moda è pari a **47** (nel dataset reale aveva un valore di **0**).

Guardando le misure di centralità notiamo che la **media campionaria** è maggiore rispetto alla mediana e alla moda. Questo potrebbe indicare una asimmetria positiva.

Di seguito un boxplot della variabile *Distinct call numbers* ci permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.

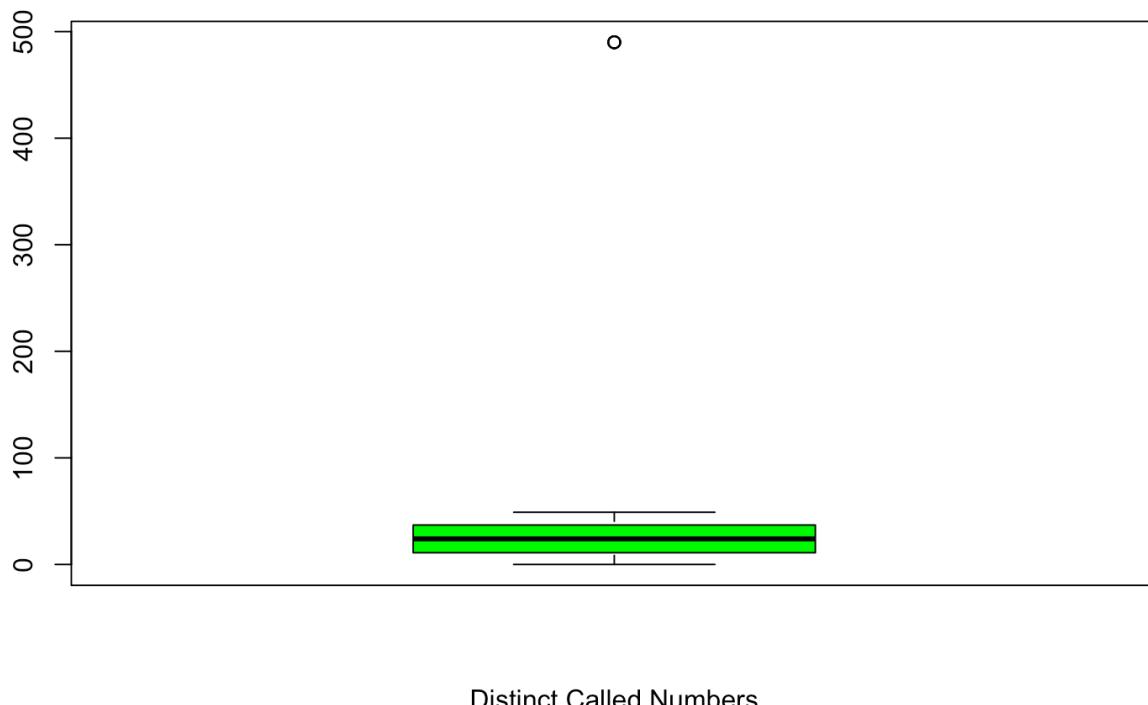


Figura 97 Boxplot Distinct Call Numbers Sintetico

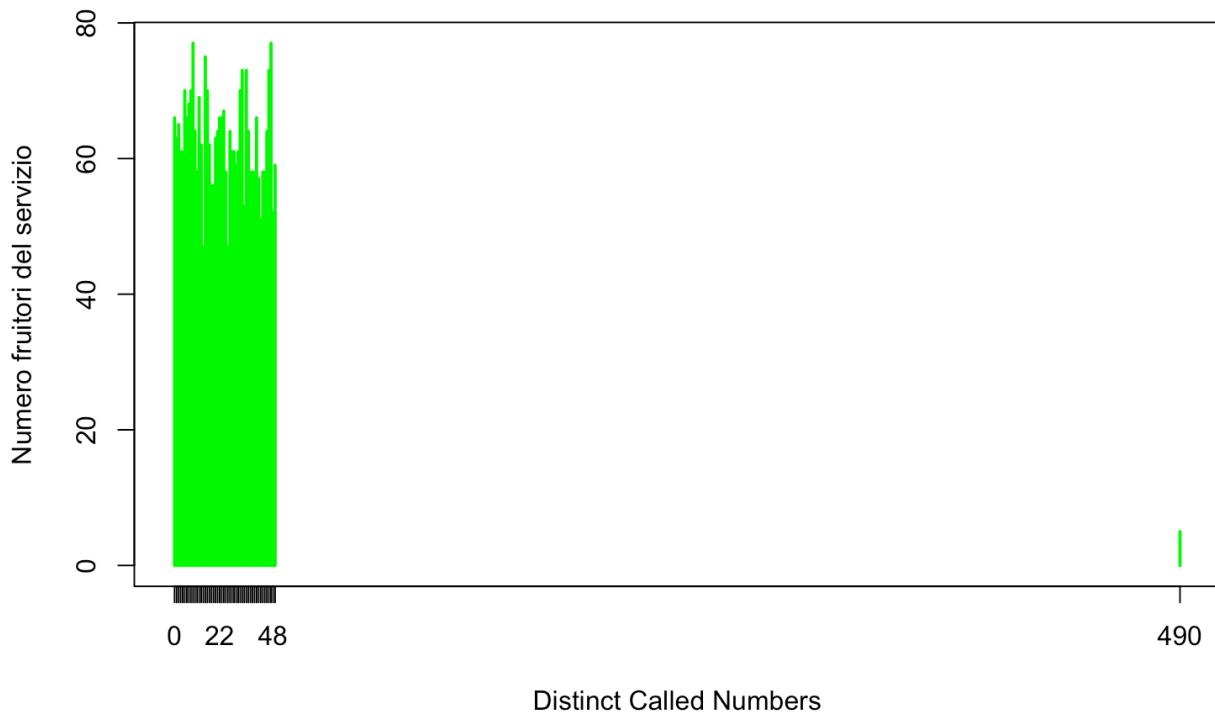
Possiamo notare dall'immagine che c'è un unico outlier, di fatti, seppur abbiamo chiesto all'AI di inserire delle anomalie per rendere i dati più simili ad un caso reale, constatiamo che comunque il dataset generato presenta dei valori in un range sicuro inserendo un'unica anomalia

Di seguito l'elenco degli outliers del dataset: **490**.

Tramite poi una funzione apposita confermiamo che il **primo quartile** è **11.00** mentre il **terzo quartile** è **37.00**.

Inoltre, abbiamo il **minimo** uguale a **0.00** ed un **massimo** uguale a **490.00**.

Tramite l'istogramma poi possiamo andare a verificare le chiamate a numeri distinti dei fruitori generati sinteticamente.



*Figura 98 Istogramma Distinct call numbers Sintetico*

Un istogramma della variabile *Distinct call numbers* mostra la frequenza assoluta [q] degli numeri chiamati per ciascun valore osservato. Le ascisse rappresentano il numero numeri chiamati, mentre le ordinate indicano la quantità di utenti corrispondenti.

Il grafico conferma una asimmetria di distribuzione ed inoltre notiamo come l'intelligenza artificiale abbia aggiunto un outlier molto alto che va a creare la coda di destra della distribuzione

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza** [4]: **553.72**
- **Deviazione standard** [5]: **23.53**

- **Coefficiente di variazione [6]: 94.38%**

L'elevato coefficiente di variazione indica una **forte dispersione dei valori** rispetto alla media, segnalando una significativa variabilità nelle fasce di spese tra gli utenti.

Notiamo quindi che l'AI durante la creazione del dato sintetico è distribuito in un range che va circa da 0 a 48 per poi fare un salto a 490.

Per concludere il discorso andiamo a studiare la distribuzione di frequenza. I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 12.23**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 242.60**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza numeri distinti chiamati, confermando le caratteristiche sopra descritte.

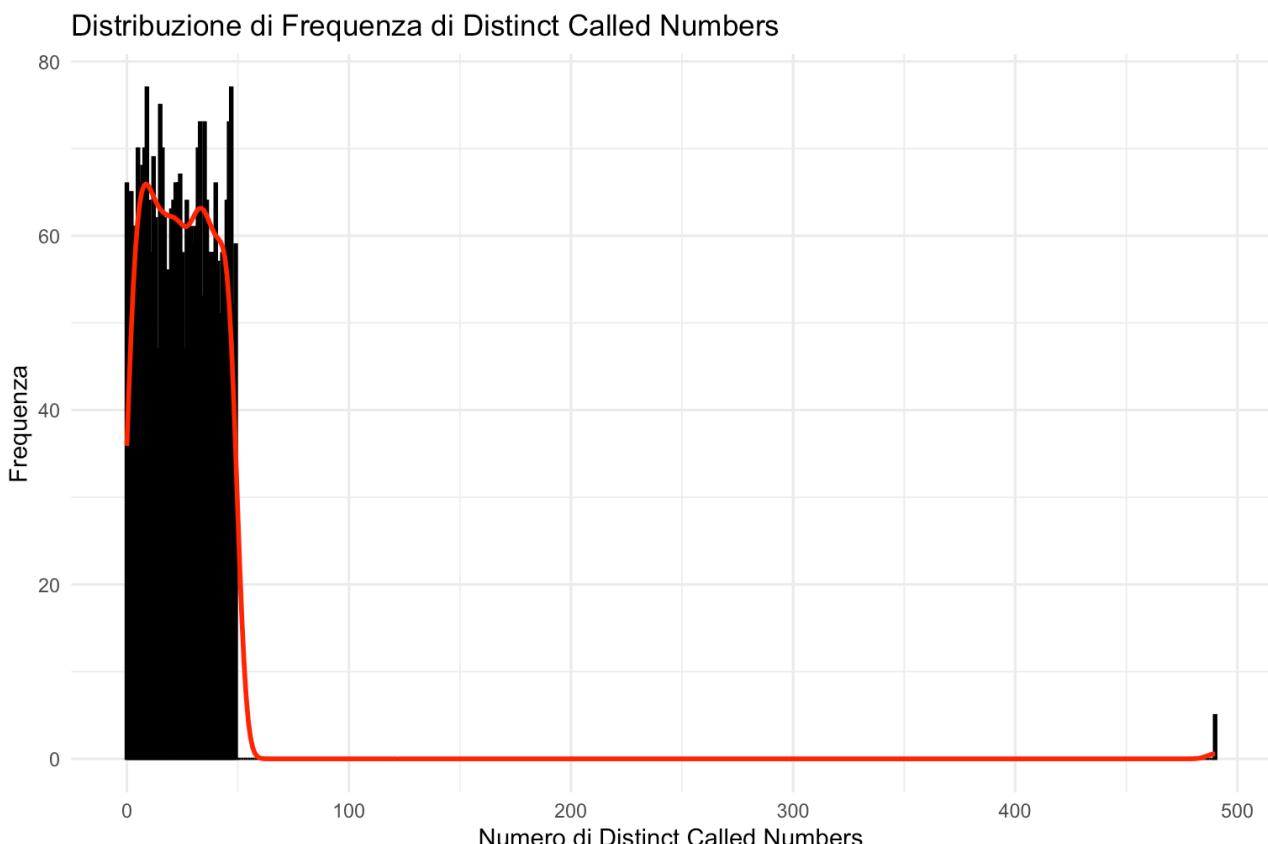


Figura 99 Distribuzione di frequenza Distinct call numbers Sintetico

### 6.1.10. Age Group

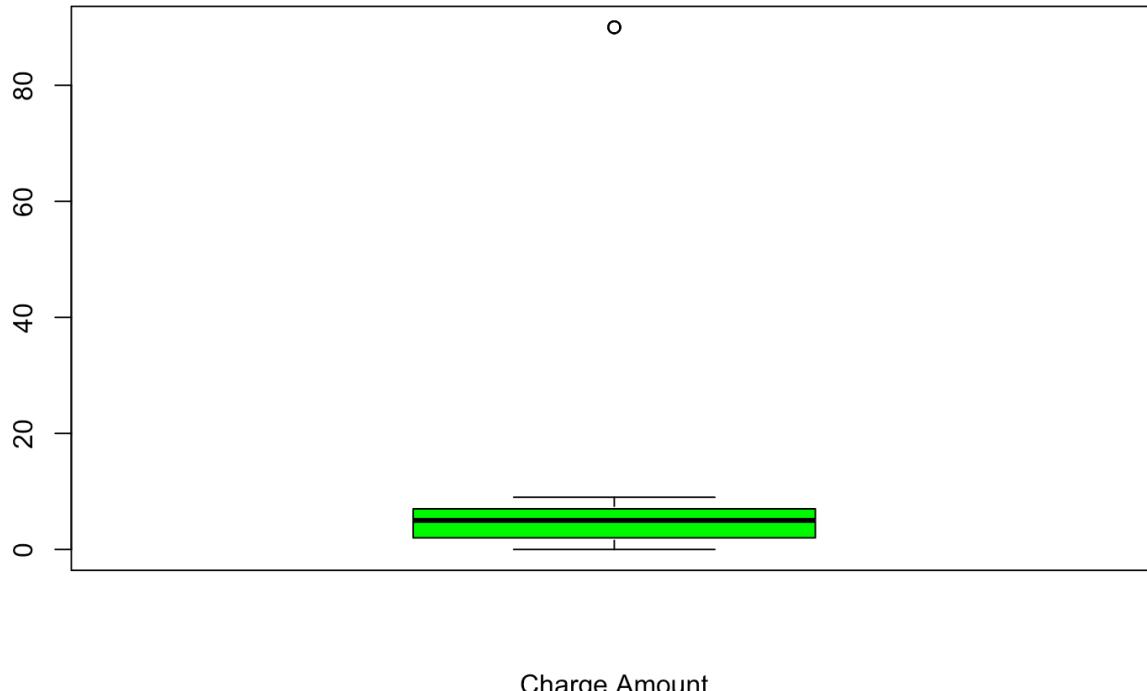
La feature “[Age Group](#)” del dataset generato sinteticamente ha prodotto i seguenti risultati:

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Age Group” risulta pari a **4.68** (nel dataset reale aveva un valore di **0.94** ciò dimostra che abbiamo all’effettivo valori molto diversi tra le due variabili).
- **Mediana campionaria [2]**: La mediana è pari a **5** (nel dataset reale aveva un valore di **0**).
- **Moda campionaria [3]**: La moda è pari a **9** (nel dataset reale aveva un valore di **0**).

Guardando le misure di centralità notiamo che la **media campionaria** è maggiore rispetto alla mediana e alla moda. Questo potrebbe indicare una asimmetria positiva.

Di seguito un boxplot della variabile Age Group *Sintetica* ci permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.



*Figura 100 Boxplot Age Group Sintetico*

Possiamo notare dall'immagine che c'è un unico outlier, di fatti, seppur abbiamo chiesto all'AI di inserire delle anomalie per rendere i dati più simili ad un caso reale, constatiamo

che comunque il dataset generato presenta dei valori in un range sicuro inserendo un'unica anomalia

Di seguito l'elenco degli outliers del dataset: **90**.

Tramite poi una funzione apposita confermiamo che il **primo quartile** è **2** mentre il **terzo quartile** è **7**.

Inoltre, abbiamo il **minimo** uguale a **0.00** ed un **massimo** uguale a **90.00**.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] delle lunghezze delle sottoscrizioni al servizio dei fruitori generati sinteticamente.

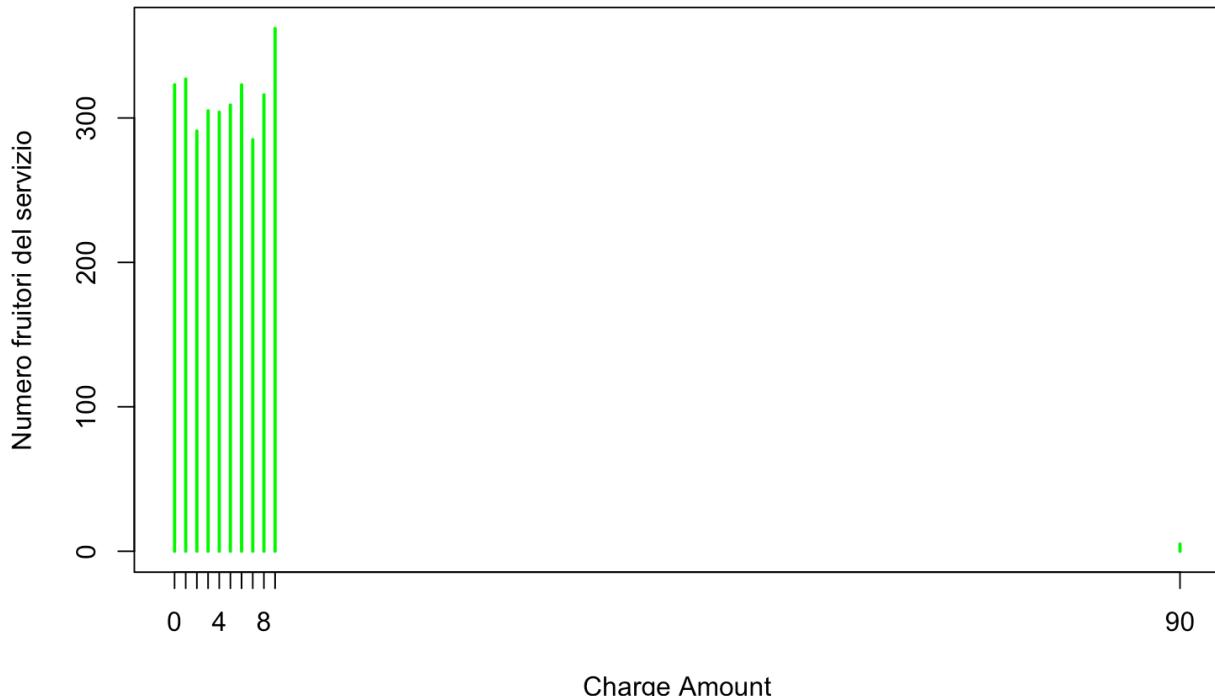


Figura 101 Istogramma Age Group Sintetico

Un istogramma della variabile *Age Group* mostra la frequenza assoluta [9] delle spese per ciascun valore osservato. Le ascisse rappresentano il quantitativo speso, mentre le ordinate indicano la quantità di utenti corrispondenti.

Il grafico conferma una asimmetria di distribuzione ed inoltre notiamo come l'intelligenza artificiale abbia aggiunto un outlier molto alto che va a creare la coda di destra della distribuzione. Caso molto simile alla variabile *Subscription Length*, di fatti, notiamo come abbia aggiunto l'unico outlier come 90 che di fatto è un valore, il quale neanche dovrebbe essere considerato nella fascia di spese effettuate dall'utente.

A quanto pare anche passando il dominio di ogni variabile all'AI non è stata in grado di inserire correttamente questo dato. Probabilmente la richiesta fatta di ottenere un dataset con delle anomalie ha portato l'AI a non considerare più il dominio della variabile ma ad andare addirittura molto oltre.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza [4]: **20.12****

- **Deviazione standard [5]: 4.49**
- **Coefficiente di variazione [6]: 95.58%**

L'elevato coefficiente di variazione indica una **forte dispersione dei valori** rispetto alla media, segnalando una significativa variabilità nelle fasce di spese tra gli utenti.

Di seguito per avere un maggiore impatto visivo andiamo a vedere come i valori sono distribuiti in un diagramma a torta.

Charge Amount

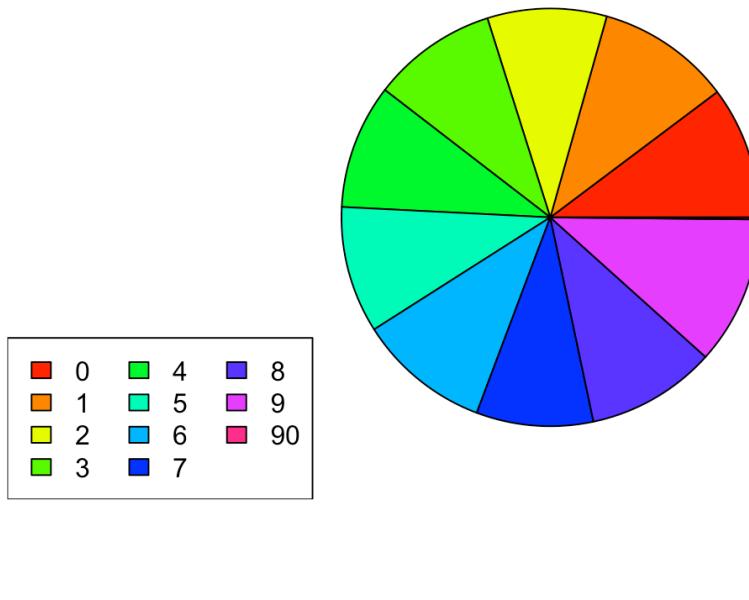


Figura 102 Pie chart Age Group Sintetico

Notiamo quindi che l'AI durante la creazione del dato sintetico ha equamente distribuito il numero di mesi partendo dal numero 1 fino al 9 (dominio quindi corretto) saltando poi direttamente al 90 il quale valore è totalmente fuori dominio.

Ciò ci porta a dire che l'AI non è stata in grado di creare per questa feature una variabile simile a quella reale seppure mediante l'utilizzo del few shot sarebbe avrebbe dovuto apprendere in che modo gli outlier si comportassero mediamente nel dataset iniziale.

Per concludere il discorso andiamo a studiare la distribuzione di frequenza.

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 10.89**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 208.19**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

---

Il seguente grafico riassume la distribuzione di frequenza dei pagamenti dei fruitori, confermando le caratteristiche sopra descritte.

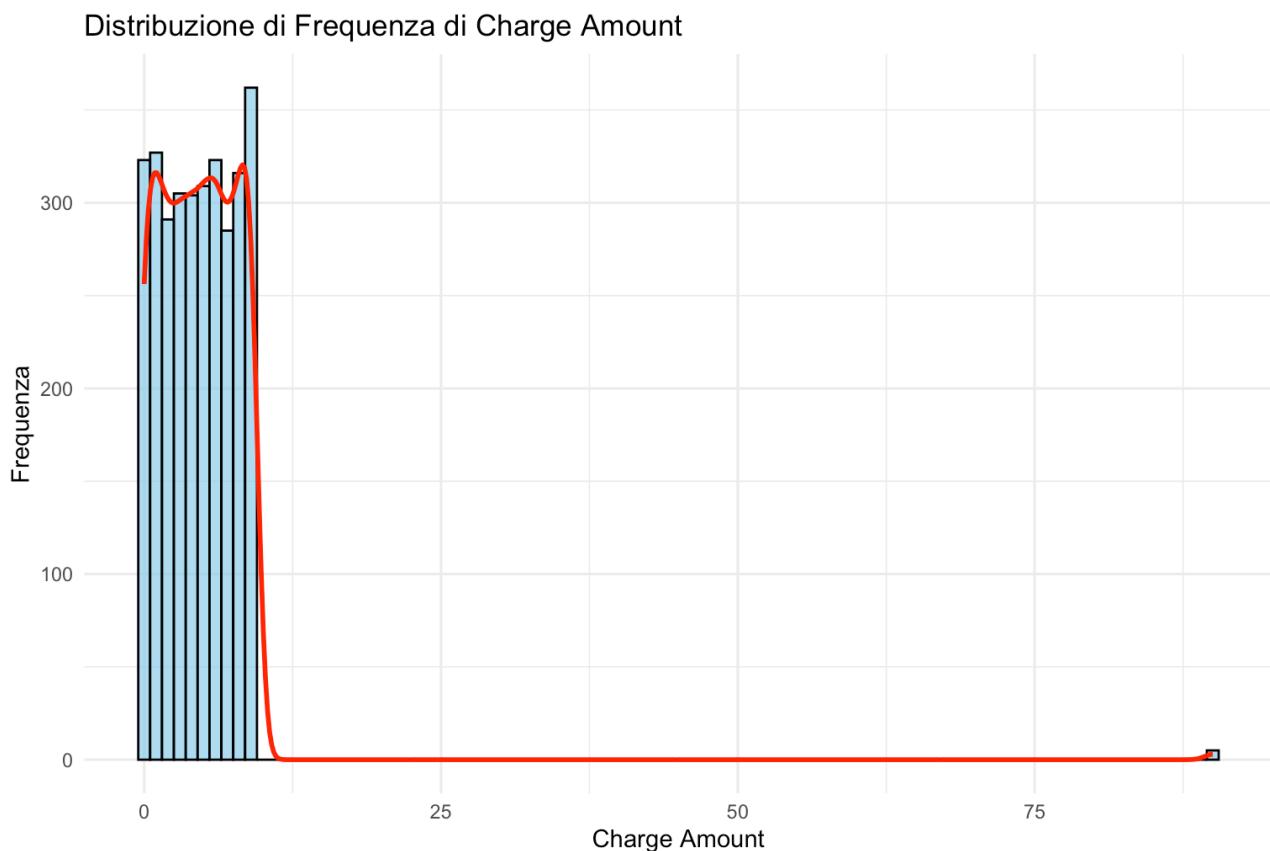


Figura 103 FDE Age Group Sintetico

### 6.1.11. Tariff plan

La feature “[Tariff plan](#)” (**1: Pay to go, 2: Pagamento contrattuale**) generato sinteticamente una volta analizzato ha prodotto i seguenti risultati:

Analizziamo quindi le **frequenze assolute** [9] dei valori assunti dalla variabile Tariff plan:

Valore	Frequenza
1: Pay to go	<u>1579</u>
2: Pagamento contrattuale	<u>1571</u>

Andiamo inoltre ad analizzare le **frequenze relative** [10]:

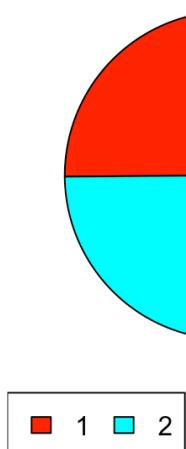
Valore	Frequenza
1: Pay to go	0.50
2: Pagamento contrattuale	0.50

Possiamo quindi notare che il **50.13%** dei fruitori ha un contratto pay to go.

Mentre il restante **49.87%** ha un contratto con Pagamento contrattuale.

Per avere un’idea più chiara possiamo osservare il diagramma a torta e il diagramma rappresentante la funzione di distribuzione empirica (discreta) sottostanti:

Distribuzione Tariff Plan



Funzione di distribuzione empirica (discr.)

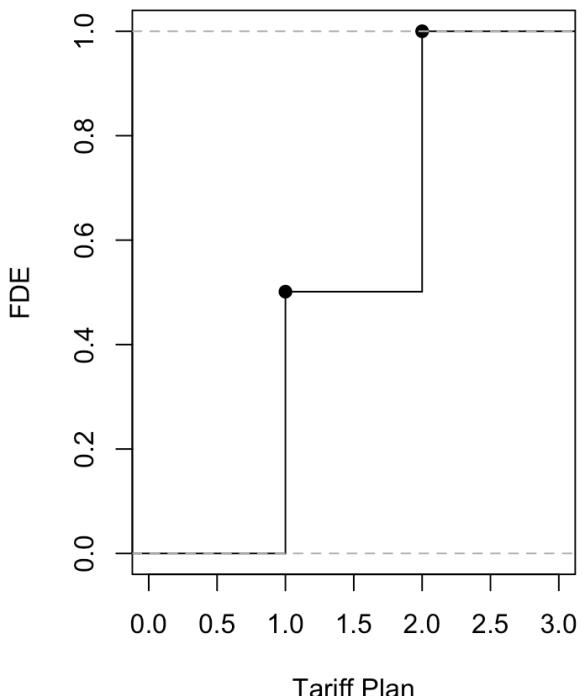


Figura 104 Diagramma a torta e FDE Tariff plan Sintetico

Possiamo dire che anche qui l'intelligenza artificiale durante la creazione del dato sintetico abbia osato poco, di fatti, notiamo che la distribuzione dei valori (**1: Pay to go, 2: Pagamento contrattuale**) sia quasi del tutto uniforme quasi un 50 e 50.

Anche per questa variabile quindi possiamo dire che questo è un caso irrealistico dato che si trova estremamente lontano dal caso reale.

### 6.1.12. Status

La feature “Status” (**1: Attivo, 2: Non attivo**) generato sinteticamente una volta analizzato ha prodotto i seguenti risultati:

Analizziamo quindi le **frequenze assolute** [9] dei valori assunti dalla variabile Tariff plan:

Valore	Frequenza
1: Attivo	<u>1593</u>
2: Non attivo	<u>1557</u>

Andiamo inoltre ad analizzare le **frequenze relative** [10]:

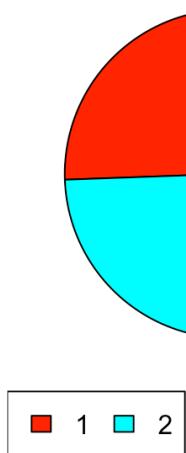
Valore	Frequenza
1: Attivo	0.50
2: Non attivo	0.50

Possiamo quindi notare che il **50.57%** dei fruitori ha un contratto attivo.

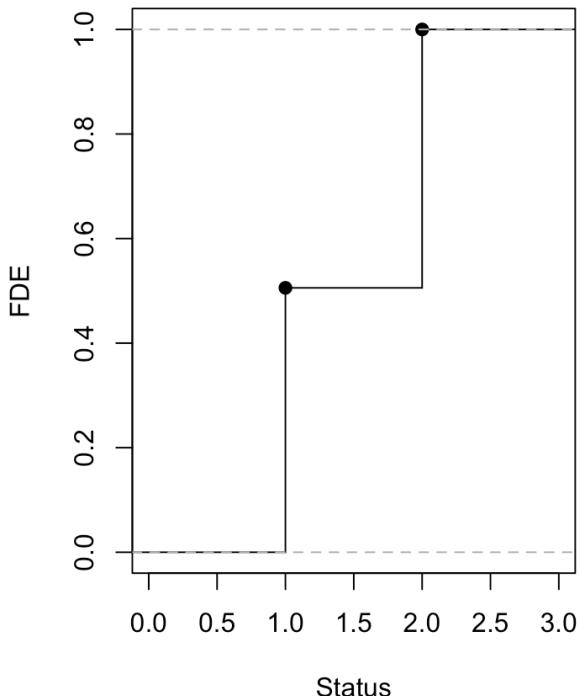
Mentre il restante **49.43%** ha un contratto non più attivo.

Per avere un’idea più chiara possiamo osservare il diagramma a torta e il diagramma rappresentante la funzione di distribuzione empirica (discreta) sottostanti:

Distribuzione Status



Funzione di distribuzione empirica (discr.)



Possiamo dire che anche qui l'intelligenza artificiale durante la creazione del dato sintetico abbia disperso poco i valori, di fatti, notiamo che la distribuzione dei valori (**1: Attivo, 2: Non attivo**) sia quasi del tutto uniforme quasi un 50 e 50.

Anche per questa variabile quindi possiamo dire che questo è un caso irrealistico dato che si trova estremamente lontano dal caso reale.

### 6.1.13. Churn

La feature “Churn” (**0: Non abbandonato il servizio, 1: Abbandonato il servizio**) generato sinteticamente una volta analizzato ha prodotto i seguenti risultati:

Analizziamo quindi le **frequenze assolute** [9] dei valori assunti dalla variabile Tariff plan:

Valore	Frequenza
0: Non abbandonato	<u>1532</u>
1: Abbandonato	<u>1618</u>

Andiamo inoltre ad analizzare le **frequenze relative** [10]:

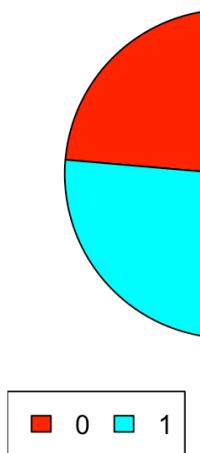
Valore	Frequenza
0: Non abbandonato	0.49
1: Abbandonato	0.51

Possiamo quindi notare che il **48.63%** dei fruitori non ha abbandonato il servizio.

Mentre il restante **51.37%** ha abbandonato il servizio

Per avere un’idea più chiara possiamo osservare il diagramma a torta e il diagramma rappresentante la funzione di distribuzione empirica (discreta) sottostanti:

Distribuzione Churn



Funzione di distribuzione empirica (discr.)

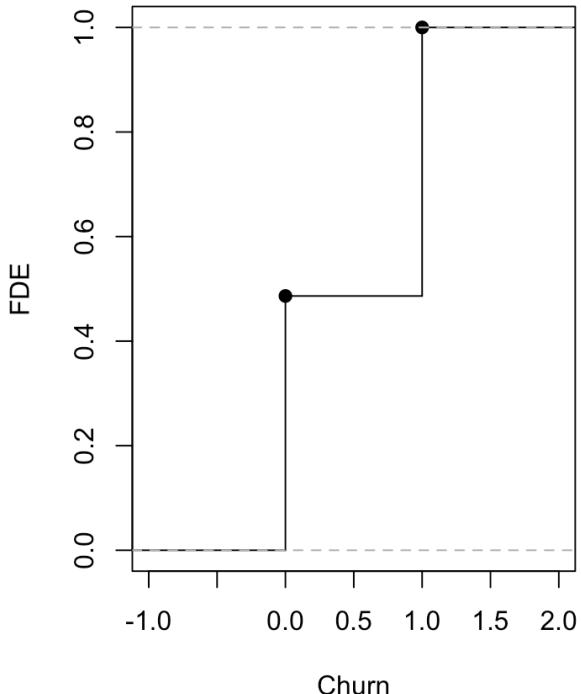


Figura 106 Diagramma a torta e FDE Churn Sintetico

Possiamo dire che anche qui l'intelligenza artificiale durante la creazione del dato sintetico abbia disperso poco i valori, di fatti, notiamo che la distribuzione dei (**0: Non abbandonato il servizio, 1: Abbandonato il servizio**) sia quasi del tutto uniforme quasi un 50 e 50.

Anche per questa variabile quindi possiamo dire che questo è un caso irrealistico dato che si trova estremamente lontano dal caso reale.

Ovviamente in un caso realistico nel caso in cui più della metà dei fruitori abbandona il servizio non permetterebbe ad un servizio di rimanere attivo.

L'AI non può pensare anche a questi dettagli ma viene da pensare che con una variabile così importante come il churn non sia corretto avere dei valori così poco attenzionati.

### 6.1.14. Customer value

La feature “[Customer value](#)” del dataset generato sinteticamente ha prodotto i seguenti risultati:

Prima di tutto procediamo con il verificare quelle che sono le misure di centralità:

- **Media campionaria [1]**: La media del campo “Customer value” risulta pari a **3456.74** (nel dataset reale aveva un valore di **70.97** ciò dimostra che abbiamo all’effettivo valori molto diversi tra le due variabili).
- **Mediana campionaria [2]**: La mediana è pari a **2583.20** (nel dataset reale aveva un valore di **228.48**).
- **Moda campionaria [3]**: La moda è pari a **49994.55** (nel dataset reale aveva un valore di **0**).

Guardando le misure di centralità notiamo che la **media campionaria** è maggiore rispetto alla mediana e alla moda. Questo potrebbe indicare una asimmetria positiva.

Di seguito un boxplot della variabile Customer value *Sintetica ci* permette di individuare visivamente il minimo, il massimo, il primo e il terzo quartile, oltre a segnalare i valori outliers.

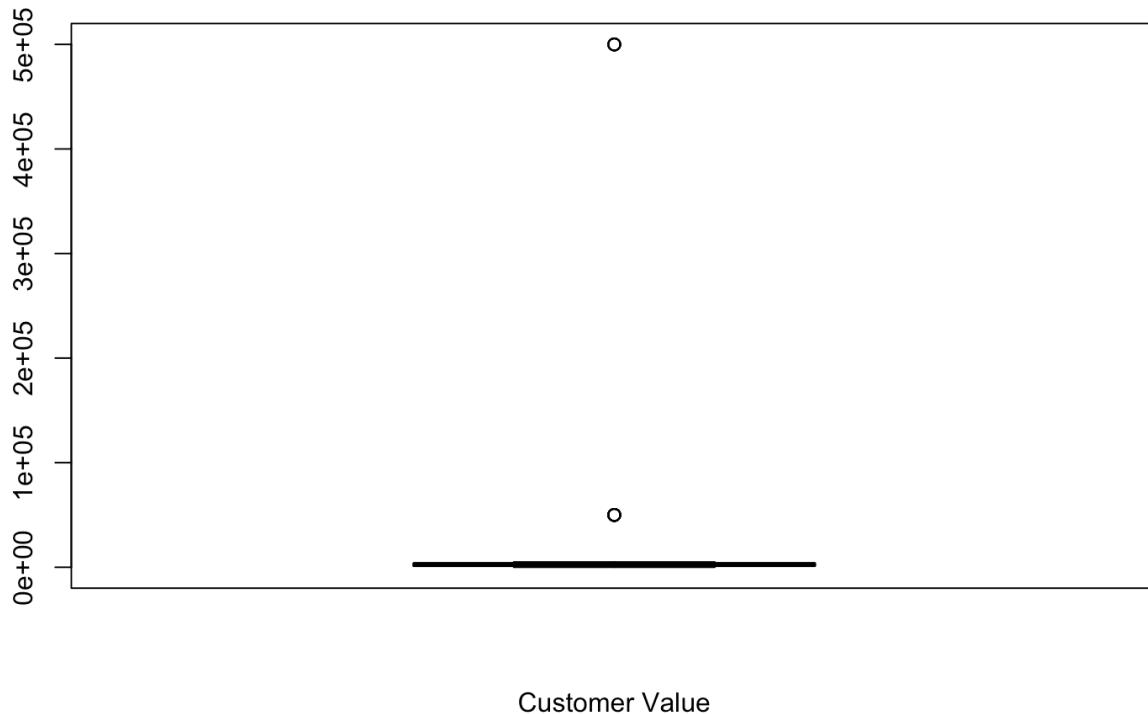


Figura 107 Boxplot Customer Value Sintetico

Possiamo notare dall'immagine che ci sono ben due outliers, di fatti, seppur abbiamo chiesto all'AI di inserire delle anomalie per rendere i dati più simili ad un caso reale, constatiamo che comunque il dataset generato presenta dei valori in un range sicuro inserendo solo due anomalie.x

Di seguito l'elenco degli outliers del dataset: **499945.53, 49994.55**.

Tramite poi una funzione apposita confermiamo che il **primo quartile** è **1295.1** mentre il **terzo quartile** è **3791.6**.

Inoltre, abbiamo il **minimo** uguale a **0.1** ed un **massimo** uguale a **499945.53**.

Tramite l'istogramma poi possiamo andare a verificare le frequenze assolute [9] delle lunghezze delle sottoscrizioni al servizio dei fruitori generati sinteticamente.



*Figura 108 Istogramma Customer Value Sintetico*

Un istogramma della variabile *Customer value* mostra la frequenza delle valutazioni per ciascun valore osservato. Le ascisse rappresentano la valutazione di un utente, mentre le ordinate indicano la quantità di utenti corrispondenti.

Il grafico conferma una asimmetria di distribuzione ed inoltre notiamo come l'intelligenza artificiale abbia aggiunto due outliers molto alto che va a creare la coda di destra della distribuzione. Prevediamo comunque un'altra forte dispersione dei valori.

Andiamo ora a verificare come i dati sono dispersi calcolando gli indici di dispersione:

- **Varianza** [4]: **400528354**
- **Deviazione standard** [5]: **20013.2**
- **Coefficiente di variazione** [6]: **578.96%**

Abbiamo qui registrato il più elevato coefficiente di variazione il che implica che c'è una **forte dispersione dei valori** rispetto alla media.

I risultati dei calcoli di skewness e curtosi forniscono una descrizione dettagliata della forma della distribuzione:

- **Skewness [7]: 24.28**, che conferma l'asimmetria verso destra.
- **Curtosi [8]: 601.68**, indicando una distribuzione leptocurtica, caratterizzata da un picco elevato.

Il seguente grafico riassume la distribuzione di frequenza dei pagamenti dei fruitori, confermando le caratteristiche sopra descritte.

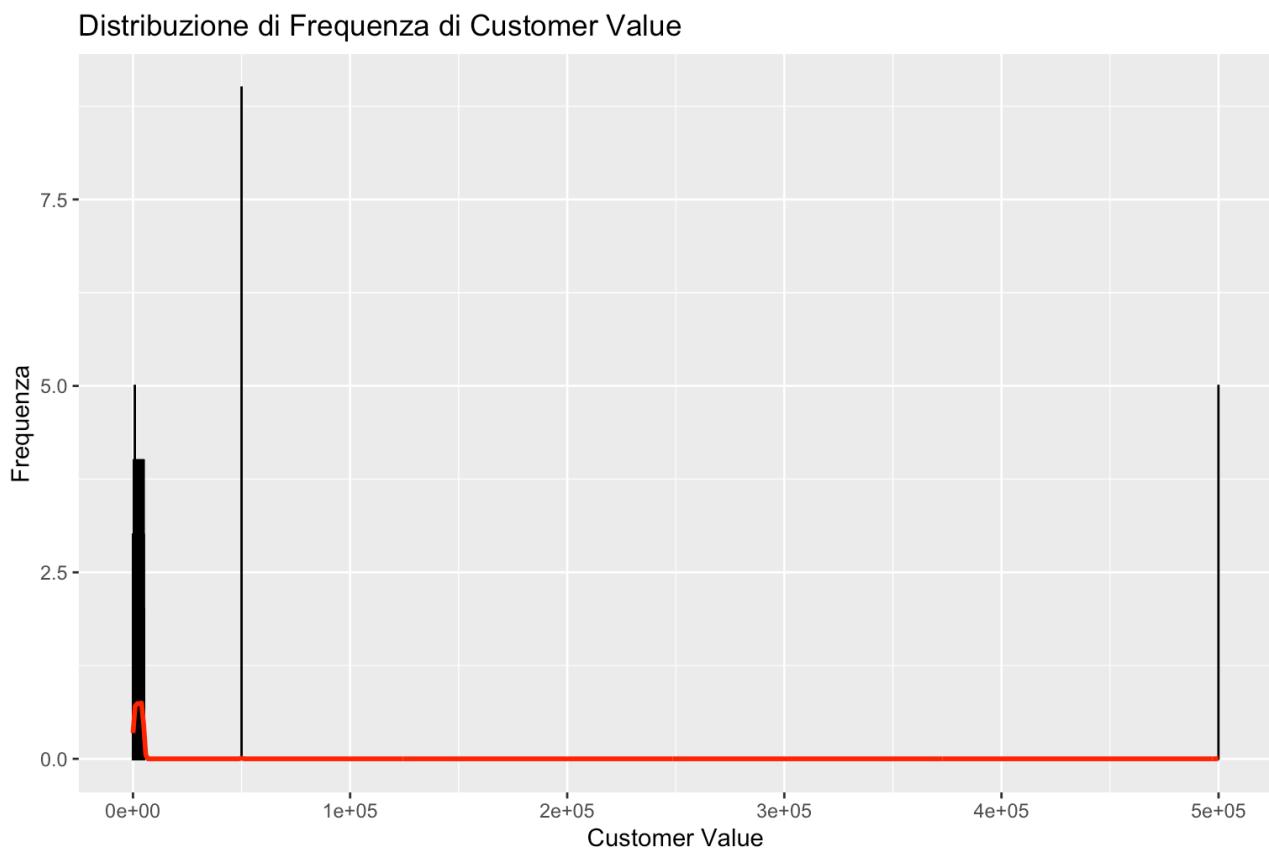


Figura 109 FDE Customer Value Sintetico

# 7. Analisi dei Risultati

## 7.1. Confronto tra Dati Reali e Sintetici

Nel [capitolo precedente](#) è stato evidenziato come i dati generati dall'intelligenza artificiale differiscano significativamente da quelli reali. In particolare, è emerso che le variabili quantitative sono state gestite in modo inadeguato, con valori generati quasi completamente uniformi, privi della variabilità osservata nei dati reali.

L'analisi univariata ha inoltre mostrato che le correlazioni tra le variabili presenti nel dataset reale non sono state rispettate nei dati sintetici.

Questo aspetto è comprensibile, poiché l'intelligenza artificiale non ha accesso alle informazioni sulla struttura di dipendenza tra le variabili del dataset originale.

Tuttavia, come evidenziato dal grafico sottostante, nei dati sintetici non è stata aggiunta alcuna correlazione significativa tra le variabili, confermando l'assenza di una struttura di dipendenza simile a quella reale.

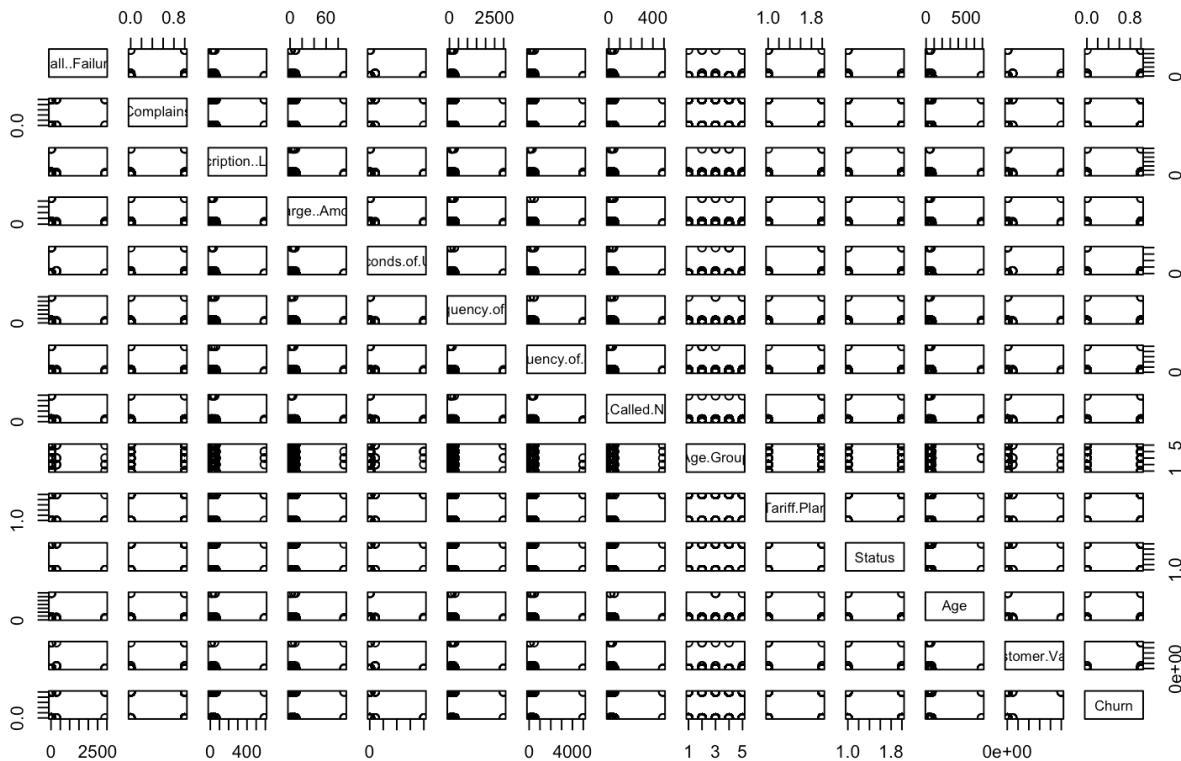


Figura 110 Pairs dataset sintetico

Magari fornendo all'AI l'intero dataset avrebbe potuto creare meglio dei dati sintetici. Inoltre, la richiesta esplicita di includere outlier per ogni variabile ha prodotto risultati incoerenti: i valori generati erano estremamente elevati e fuori scala, privi di una logica coerente con la distribuzione dei dati reali. Questo comportamento si è verificato nonostante l'utilizzo di una tecnica few-shot, in cui erano state fornite all'algoritmo 20 righe del dataset originale per facilitarne la comprensione delle caratteristiche dei dati.

# 8. Conclusioni

L'analisi condotta ha evidenziato che i dati sintetici generati dai Large Language Model mostrano alcune somiglianze con i dati reali in termini di proprietà statistiche di base. Tuttavia, la gestione delle variabili quantitative si è rivelata problematica, con una tendenza alla generazione di valori uniformi e all'assenza di correlazioni tra variabili. Inoltre, i dati sintetici non hanno replicato le relazioni strutturali presenti nel dataset reale, come evidenziato sia dall'analisi univariata che dall'osservazione delle correlazioni.

## 8.1. Research Question 1

Per quanto riguarda la [Research Question 1](#) (“I dati sintetici generati dai Large Language Model mantengono le stesse proprietà statistiche dei dati reali?”), è emerso che, sebbene alcune proprietà statistiche siano state mantenute per alcune variabili, la mancanza di correlazioni tra le variabili evidenzia una limitata capacità dell'AI di replicare completamente le caratteristiche statistiche dei dati reali.

Inoltre, è importante sottolineare che i dati sono stati generati in maniera quasi uniforme e, sebbene fosse stato richiesto l'inserimento degli outlier, questi sono stati effettivamente introdotti, ma in quantità molto ridotta (1-2 per variabile) e in proporzione simile ai valori non outlier creando così una situazione alquanto artificiale e molto lontana dalla casistica reale.

## 8.2. Research Question 2

Relativamente alla [Research Question 2](#) (“Perché gli utenti abbandonano il servizio?”), l'analisi effettuata tramite il clustering ha evidenziato che le variabili che influenzano maggiormente l'abbandono del servizio da parte degli utenti sono:

- [Complains](#)
- [Status](#)
- [Seconds of use](#)

In particolare, è stato osservato che gli utenti tendono ad abbandonare il servizio in presenza di lamentele (“complains”), di uno stato non più attivo (“status”) e di un basso numero di secondi di utilizzo (“seconds of use”).

Queste condizioni combinano una mancanza di coinvolgimento con esperienze potenzialmente negative, rendendole fattori determinanti per il churn.

## 8.3. Research Question 3

Per la [Research question 3](#) la domanda era la seguente “Con i dati forniti dal dataset Iranian Churn e proiettando il servizio su scala mondiale, quale sarebbe il tasso di abbandono degli utenti entro 9 mesi?”. Abbiamo dato una risposta nel [capitolo 5](#).

Abbiamo scoperto che proiettando il servizio su scala globale, abbiamo potuto osservare che con una probabilità del **90%** (il nostro  $\alpha$  è 0.1) il tasso di abbandono sarebbe compreso tra il **15.68%** ed il **15.72%**.

Il che significa che il tasso di abbandono nel giro di 9 mesi sarebbe accettabile e molto simile a quello del nostro dataset che è del **15.7%**.

Possiamo quindi dire che il provider del servizio Iranian Churn può estendere il proprio servizio su scala mondiale non rischiando eccessivamente dato che l'abbandono degli utenti non sarebbe molto alto.

# 9. Appendice

In questo formulario si farà riferimento ad un'unica variabile “Call Failures” del dataset per evitare ridondanza.

## 1. Media campionaria:

La media campionaria è la somma di tutti i valori di un campione divisa per il numero totale di osservazioni. Rappresenta il valore medio dei dati e fornisce una misura della tendenza centrale del campione.

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

Codice R utilizzato per calcolare la media campionaria è il seguente:

```
# Calcolo della media per la colonna Call Failures
mean(dataset$Call..Failure)
```

## 2. Mediana campionaria:

La mediana campionaria è il valore centrale di un insieme di dati ordinati, appartenenti a un campione. Divide il campione in due parti uguali, tali che metà dei valori sia minore o uguale alla mediana e l'altra metà sia maggiore o uguale.

Codice R utilizzato per calcolare la mediana campionaria è il seguente:

```
# Calcolo della mediana per la colonna Call Failures
median(dataset$Call..Failure)
```

## 3. Moda campionaria:

La moda campionaria è il valore che compare con maggiore frequenza in un campione. Rappresenta la modalità più comune tra i dati ed è una misura della tendenza centrale

Codice R utilizzato per calcolare la moda campionaria è il seguente:

```
# Definizione funzione per calcolare la moda
moda <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

# Calcolo della moda per la colonna Call Failures
moda(dataset$Call..Failure)
```

In questo caso la moda è stata utilizzata anche per comprendere al meglio anche la distribuzione dei valori (unimodale/bimodale/multimodale).

#### **4. Varianza campionaria:**

La varianza campionaria è una misura della dispersione dei dati rispetto alla media campionaria.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (n = 2, 3, \dots),$$

Codice R utilizzato per calcolare la varianza campionaria è il seguente:

```
# Calcolo varianza per la colonna Call Failures
var(dataset$Call..Failure, na.rm = TRUE)
```

La varianza è molto sensibile agli outliers data l'elevazione al quadrato.  
Viene espressa con l'unità di misura dei dati analizzati elevata al quadrato.

#### **5. Deviazione standard:**

La deviazione standard campionaria è una misura della dispersione dei dati rispetto alla media campionaria, espressa nella stessa unità di misura dei dati.

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (n = 2, 3, \dots).$$

Codice R utilizzato per calcolare la deviazione standard è il seguente:

```
# Calcolo deviazione standard per la colonna Call Failures
sd(dataset$Call..Failure, na.rm = TRUE)
```

È la radice quadrata della varianza campionaria (quindi è meno sensibile agli outliers) e viene espressa nella stessa unità di misura dei dati analizzati.

#### **6. Coefficiente di variazione:**

Il coefficiente di variazione è un indice di dispersione che ha senso soltanto per campioni aventi la media campionaria non nulla.

Il CV è particolarmente utile per confrontare la dispersione di dati che hanno unità o scale diverse, poiché normalizza la deviazione standard rispetto alla media.

$$CV = \frac{s}{\bar{x}}$$

→ Deviazione Standard  
→ Media Campionaria

Il coefficiente di variazione è adimensionale ed è utile per calcolare la percentuale di dispersione delle variabili rispetto alla media.

- CV < 10%: Tipicamente indica una bassa variabilità rispetto alla media. I dati sono relativamente stabili.
- CV tra 10% e 30%: Indica una moderata variabilità
- CV > 30%: Indica un'elevata variabilità, suggerendo che i dati hanno una notevole dispersione rispetto alla media

Codice R utilizzato per calcolare il coefficiente di variazione è il seguente:

```
# Calcolo coefficiente di variazione per la colonna Call Failures
sd(dataset$Call..Failure, na.rm = TRUE)/mean(dataset$Call..Failure)*100
```

## 7. Skewness:

La skewness misura il grado di asimmetria della distribuzione dei dati rispetto a una distribuzione simmetrica.

$$\text{Skewness} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$$

- Skewness > 0: Distribuzione asimmetrica a destra (coda più lunga a destra).
- Skewness < 0: Distribuzione asimmetrica a sinistra (coda più lunga a sinistra).
- Skewness = 0: Distribuzione simmetrica.

Codice R utilizzato per calcolare la skewness è il seguente:

```
# Calcolo skewness per la colonna Call Failures
skewness_value <- skewness(dataset$Call..Failure)
```

## 8. Kurtosis:

La curtosi misura la “piattezza” o la “concentrazione” della distribuzione rispetto a una normale.

$$\text{Curtosi} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{s^4} - 3$$

- Curtosi > 0: Distribuzione leptocurtica (più concentrata con code più lunghe).

- Curtosi < 0: Distribuzione platicurtica (più piatta con code più corte).
- Curtosi = 0: Distribuzione mesocurtica (simile alla normale).

Codice R utilizzato per calcolare la curtosi è il seguente:

```
# Calcolo curtosi per la colonna Call Failures
kurtosis_value <- kurtosis(dataset$Call..Failure)
```

## **9. Frequenza assoluta:**

La frequenza assoluta è una misura che indica quante volte un certo valore o categoria si presenta in un insieme di dati. È rappresentata come un numero intero e non tiene conto della proporzione rispetto al totale.

$$\sum f_i = n$$

Codice R utilizzato per calcolare la frequenza assoluta è il seguente:

```
# Calcolo frequenze assolute per colonna Complains
freq_assolute <- table(dataset$Complains)
```

## **10. Frequenza relativa:**

La frequenza relativa misura la proporzione di un valore o categoria rispetto al totale delle osservazioni in un insieme di dati. È espressa come un numero compreso tra 0 e 1 (o in percentuale, moltiplicandolo per 100).

$$fr_i = \frac{f_i}{n}$$

Codice R utilizzato per calcolare la frequenza relativa è il seguente:

```
# Calcolo frequenze relative per colonna Complains
freq_relative <- prop.table(freq_assolute)
```

## 11. K-means

Il metodo K-means è un algoritmo di clustering non supervisionato che suddivide un insieme di dati in k gruppi (o cluster) sulla base della similarità dei dati. Ogni cluster è rappresentato dal suo centroide, che è il punto medio dei dati appartenenti a quel cluster. L'algoritmo è sensibile alla presenza di outlier, che possono influenzare significativamente la posizione dei centroidi e la qualità del clustering.

Fasi principali del K-means:

1. Inizializzazione: Seleziona casualmente k centroidi iniziali, dove k è il numero di cluster desiderati.
2. Assegnazione: Ogni punto del dataset viene assegnato al cluster il cui centroide è il più vicino, utilizzando una metrica di distanza (di solito la distanza euclidea).
3. Aggiornamento: Calcola un nuovo centroide per ciascun cluster come la media dei punti assegnati a quel cluster.
4. Iterazione: Ripeti i passaggi di assegnazione e aggiornamento finché i centroidi non cambiano significativamente o viene raggiunto un numero massimo di iterazioni.

Codice R utilizzato per effettuare il clustering è il seguente:

```
# Subset dei dati e scaling
data_subset <- dataset[, c("Complains", "Status", "Seconds.of.Use")]
data_scaled <- scale(data_subset)

# Clustering con k = 4
kmeans_model <- kmeans(data_scaled, centers = 4, nstart = 10)
dataset$Cluster <- as.factor(kmeans_model$cluster)
```

## 12. Within-Cluster Sum of Squares (WSS)

Il WSS (Within-Cluster Sum of Squares) è una metrica utilizzata per valutare la qualità del clustering in algoritmi come il K-means. Esso misura la somma delle distanze quadratiche tra i punti di un cluster e il loro centroide. Un valore di WSS più basso indica che i punti sono più vicini al centroide, suggerendo che il cluster è più compatto.

$$WSS_k = \sum_{x \in C_k} \|x - \mu_k\|^2$$

Codice R utilizzato per calcolare il WSS è il seguente:

```
wss <- sum(kmeans_model$withinss)
```

### **13. Indice di Calinski-Harabasz**

L'indice di Calinski-Harabasz (o criterio di varianza inter-cluster) è una metrica utilizzata per valutare la qualità di un clustering. Questo indice misura il rapporto tra la dispersione totale tra i cluster (inter-cluster) e la dispersione all'interno dei cluster (intra-cluster). Valori più alti dell'indice indicano un clustering migliore.

$$CH = \frac{SSB/(k-1)}{SSW/(n-k)}$$

Codice R utilizzato per calcolare il CH è il seguente:

```
wss <- sum(kmeans_model$withinss)
bss <- sum(kmeans_model$betweenss)

n <- nrow(data_scaled)
k <- 4
s_t <- (1 / (n - 1)) * sum((data_scaled - colMeans(data_scaled))^2)
ch_index <- (bss / (k - 1)) / (wss / (n - k))
cat("Calinski-Harabasz Index:", ch_index, "\n")
```

### **14. Covarianza campionaria**

Covarianza campionaria: è una misura del grado di correlazione lineare tra le due variabili, ed è definita come:

$$c_{xy} := \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y}).$$

In base al valore di  $c_{xy}$ , si pu`o definire il tipo di relazione che lega le variabili:

- $c_{xy} < 0 \Rightarrow X$  ed  $Y$  sono linearmente correlate negativamente;
- $c_{xy} = 0 \Rightarrow X$  ed  $Y$  sono linearmente non correlate;
- $c_{xy} > 0 \Rightarrow X$  ed  $Y$  sono linearmente correlate positivamente;

Codice R utilizzato per calcolare la covarianza campionaria è il seguente:

```
#Calcolo covarianza campionaria della relazione tra Customer Value e Frequency of SMS
cov(dataset$Frequency.of.SMS, dataset$Customer.Value)
```

## 15. Coefficiente di correlazione campionario

Un indice strettamente legato alla covarianza campionaria è il coefficiente di correlazione campionario, una misura normalizzata del grado di relazione lineare tra le variabili X ed Y.

Esso è definito come:

$$r_{xy} := \frac{c_{xy}}{s_x s_y},$$

ed assume sempre valori compresi tra -1 ed 1.

Il valore del coefficiente di correlazione campionario indica il tipo e l'intensità della relazione lineare presente tra le variabili X ed Y:

- $r_{xy} = -1 \Rightarrow$  X ed Y sono in correlazione lineare negativa perfetta;
- $-1 < r_{xy} < 0 \Rightarrow$  X ed Y sono linearmente correlate negativamente;
- $r_{xy} = 0 \Rightarrow$  X ed Y sono linearmente non correlate;
- $0 < r_{xy} < 1 \Rightarrow$  X ed Y sono linearmente correlate positivamente;
- $r_{xy} = 1 \Rightarrow$  X ed Y sono in correlazione lineare positiva perfetta;

Codice R utilizzato per calcolare il coefficiente di correlazione campionario è il seguente:

```
#Calcolo Calcola il coefficiente di correlazione della relazione tra Customer Value e Frequency of SMS
cor(dataset$Frequency.of.SMS, dataset$Customer.Value)
```

## 16. Regressione lineare semplice

Siano X ed Y due variabili, e siano  $(x_1, x_2, \dots, x_n)$  e  $(y_1, y_2, \dots, y_n)$  campioni con n osservazioni relativi ad esse.

Un modello di regressione lineare semplice `e una funzione lineare della variabile X che approssima Y:

$$\hat{Y} = \beta \cdot X + \alpha,$$

dove  $\alpha$  e  $\beta$  sono detti rispettivamente intercetta e coefficiente angolare di  $X$ .

Codice R utilizzato per calcolare il modello di regressione lineare semplice e è il seguente:

```
# Calcolo della regressione lineare
model <- lm(Customer.Value ~ Frequency.of.SMS, data = dataset)
```

## 17. Residui

Siano  $(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$  i valori predetti dal modello di regressione lineare  $\hat{Y} = \beta \cdot X + \alpha$ , definiamo allora i-esimo residuo il valore:

$$E_i := y_i - \hat{y}$$

## 18. Coefficiente di determinazione

Una metrica dell'efficienza di un modello di regressione è il coefficiente di determinazione, che possiede 3 diverse definizioni:

- $D_1^2 := 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2};$
- $D_2^2 := \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2};$
- $D_3^2 := \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$

Per i modelli di regressione lineare, risulta che:

$$D_1^2 = D_2^2 = D_3^2 \text{ e che:}$$

$$0 \leq D_1^2 = D_2^2 = D_3^2 \leq 1,$$

Mentre, in generale, si ha che:

$$0 \leq D_1^2 \leq 1;$$

Codice R utilizzato per calcolare il coefficiente di determinazione è il seguente:

```
# Calcolo coefficiente di determinazione
summary(model)$r.squared
```

## 19. Regressione lineare multipla

Siano  $X_1, X_2, \dots, X_p$  ed  $Y$  variabili, siano  $(x_{1,j}, x_{2,j}, \dots, x_{n,j}), j = 1, 2, \dots, p$  e  $(y_1, y_2, \dots, y_n)$  campioni con  $n$  osservazioni relativi ad esse.

Un modello di regressione lineare multipla è una funzione lineare delle variabili  $X_1, X_2, \dots, X_n$  che approssima  $Y$ :

$$\hat{Y} = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \dots + \beta_p \cdot X_p + \alpha,$$

dove  $\alpha$  e  $\beta_j$  sono detti rispettivamente intercetta e coefficiente angolare di  $X_j$  ( $\forall j = 1, 2, \dots, p$ ).

Codice R utilizzato per calcolare il modello di regressione lineare multiplo è il seguente:

```
# Calcolo della regressione lineare
model <- lm(dataset$Customer.Value ~ dataset$Frequency.of.use + dataset$Frequency.of.SMS)
```

## 20. Distribuzione Binomiale

$$X \sim \mathcal{B}(n, p)$$

La distribuzione Binomiale è utile per modellare tutti quegli eventi che possono avere esito 0 o 1 (successo o fallimento) e per costruire una variabile aleatoria  $X$  che rappresenta il numero di successi su un certo numero di prove indipendenti, ciascuna con probabilità  $p$  di successo.

La probabilità che  $X$  assuma un valore  $k$  (cioè  $k$  successi su  $n$  prove) è data da:

$$\binom{n}{x} p^x (1-p)^{n-x},$$

E avrà una funzione di probabilità:

$$p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, \dots, n \\ 0, & \text{altrimenti} \end{cases}$$

$$(n = 1, 2, \dots; 0 < p < 1)$$

Inoltre, abbiamo:

- Valore atteso:  $\frac{np}{n}$
- Varianza:  $\frac{np(1-p)}{n}$
- Coefficiente di Variazione:  $\sqrt{\frac{1-p}{np}}$

## **21. Teorema del limite centrale**

Il Teorema del Limite Centrale afferma che, se si prende un numero sufficientemente grande di osservazioni indipendenti da una stessa distribuzione qualsiasi (con media finita e varianza finita), la loro media campionaria seguirà approssimativamente una distribuzione normale.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

## **22. Teorema centrale di convergenza**

Il Teorema Centrale di Convergenza afferma che, dato un insieme di variabili casuali indipendenti con distribuzione identica e con una media finita  $\mu$  e una varianza finita  $\sigma^2$ , la distribuzione della loro media campionaria converge, all'aumentare della dimensione del campione  $n$ , ad una distribuzione normale.

In altre parole, se prendiamo ripetutamente campioni dalla stessa popolazione e calcoliamo la media per ogni campione, allora man mano che il numero di osservazioni ( $n$ ) aumenta, la distribuzione delle medie campionarie diventa sempre più simile ad una distribuzione normale, anche se la distribuzione originale dei dati non è normale.

## **23. Metodo pivotale**

Il metodo pivotale è un modo per costruire intervalli di confidenza senza dover conoscere esattamente la distribuzione dei dati campionari.

L'idea è di usare una variabile pivotale, cioè una funzione dei dati campionari e del parametro da stimare, che però ha una distribuzione nota e indipendente dal parametro incognito.

$$P(\alpha_1 < \gamma(X_1, X_2, \dots, X_n; \vartheta) < \alpha_2) = 1 - \alpha$$

Il metodo pivotale sfrutta una variabile pivotale per trasformare una distribuzione che dipende dai parametri incogniti in una forma in cui possiamo stimare questi parametri.