

# Progetto Statistica Numerica

Dataset: Brain Weight in Humans (Anubhab Swain) Hotel Bookings Analysis (The Devastator)

Federico Malservigi

0001115961

2024-2025

# I Dataset selezionati

Hotel Bookings Analysis (The Devastator)

**Descrizione:**

Il dataset contiene informazioni dettagliate sulle prenotazioni alberghiere, inclusi dati sui clienti, durata del soggiorno, richieste speciali, tariffe, cancellazioni, e altro.

Per l'analisi sono state selezionate solo alcune features, ritenute rilevanti:

**Colonne utilizzate:**

- lead\_time, stays\_in\_weekend\_nights, stays\_in\_week\_nights, adults, children babies, adr, total\_of\_special\_requests (numeriche)
- is\_canceled (categorica)

**Dimensione finale del dataset:**

118.896 righe × 9 colonne

**Dataset per la Regressione**

Brain Weight in Humans (Anubhab Swain)

**Descrizione:**

Dataset derivato da uno studio medico contenente dati su dimensioni della testa e peso del cervello, oltre a genere e fascia d'età.

**Variabili utilizzate per la regressione:**

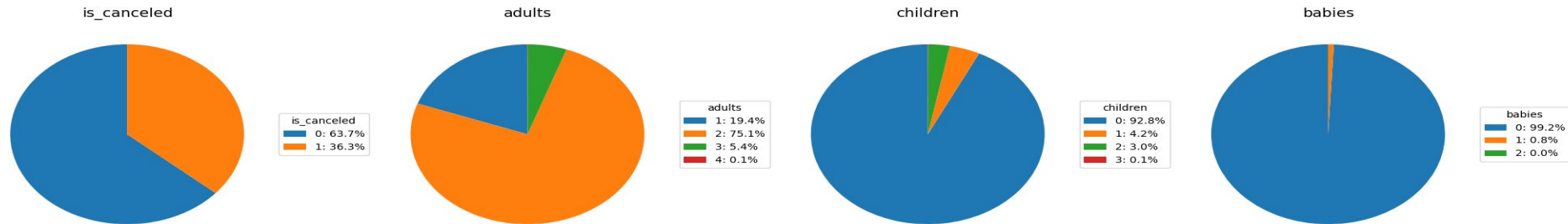
- Head Size (cm<sup>3</sup>) (numerica continua)
- Brain Weight (grams) (numerica continua)

**Dimensione finale del dataset:**

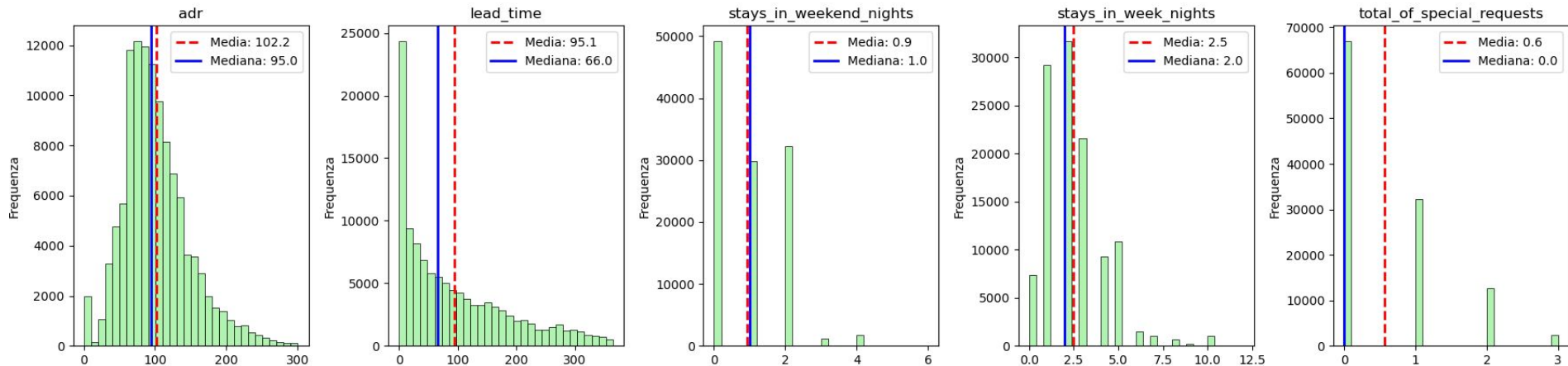
237 righe × 2 colonne

# EDA Univariata (Hotel Bookings Analysis)

Diagrammi a torta



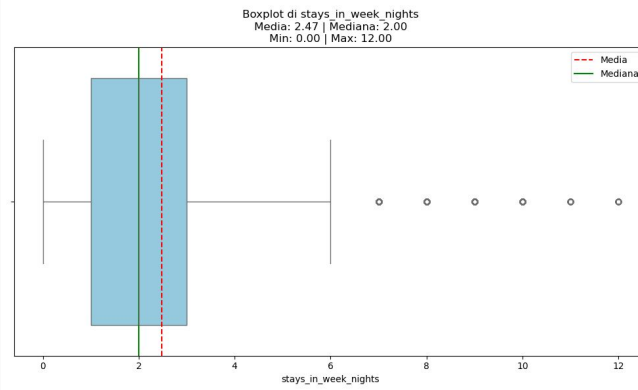
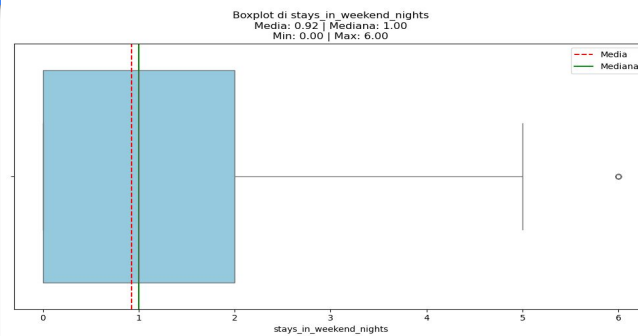
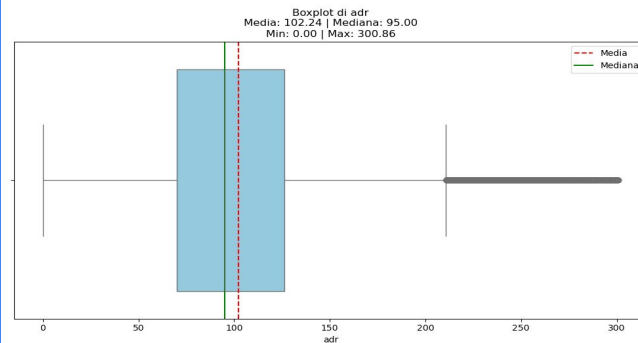
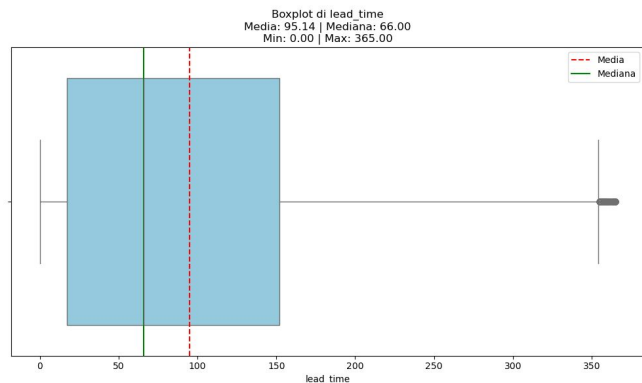
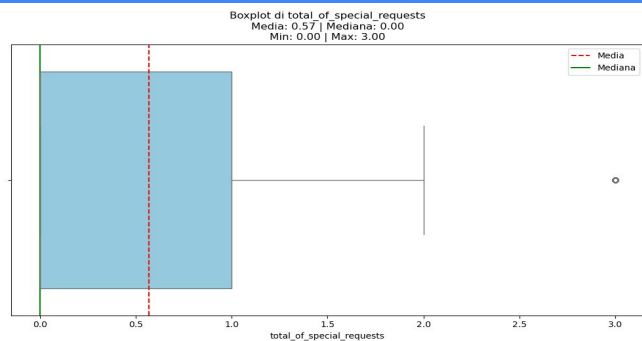
Istogrammi



# EDA Univariata

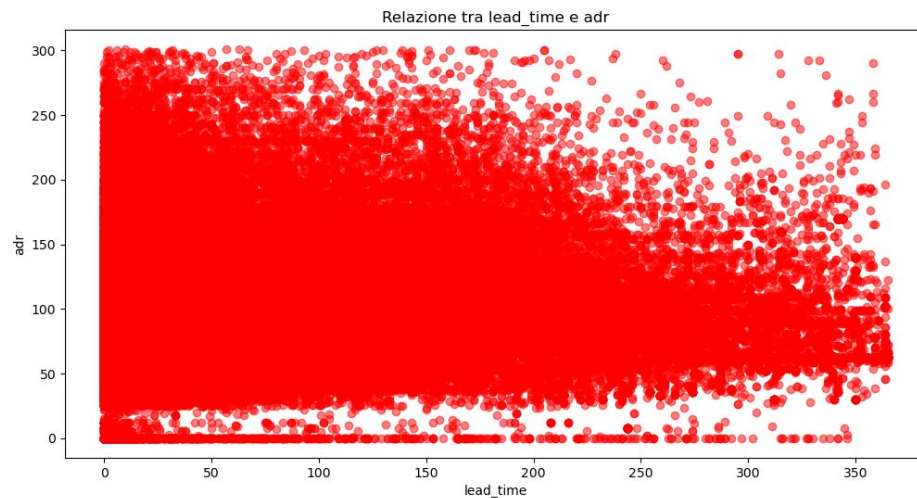
## Boxplot

(Hotel Bookings Analysis)



# EDA Bivariata (Hotel Bookings Analysis)

Dallo scatter plot emerge che, in media, quando il lead\_time è più breve, l'ADR tende ad essere più alto. Questo suggerisce una relazione negativa tra le due variabili, anche se non perfettamente lineare.

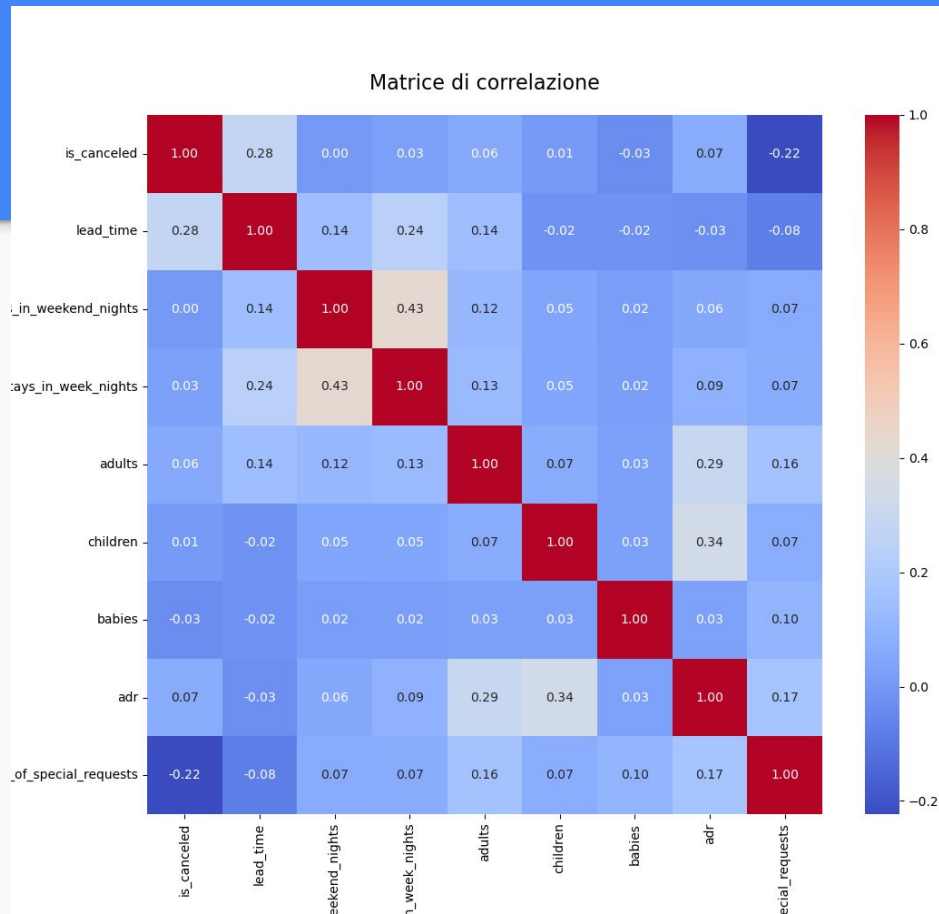


# EDA Multivariata

## (Hotel Bookings Analysis)

Le correlazioni più forti confermano che:

- **Lead time** influisce positivamente sul rischio di cancellazione.
- **Richieste speciali** influiscono negativamente sul rischio di cancellazione.
- **Prezzo medio giornaliero della camera** è influenzato positivamente soprattutto dal numero di adulti e bambini



# Classificazione (Hotel Bookings Analysis)

Dataset ridotto al 5% (5766 righe) per motivi computazionali, mantenendo la rappresentatività.

Target: **is\_canceled**

Suddivisione ottimale:

**70% training – 15% validation – 15% test**

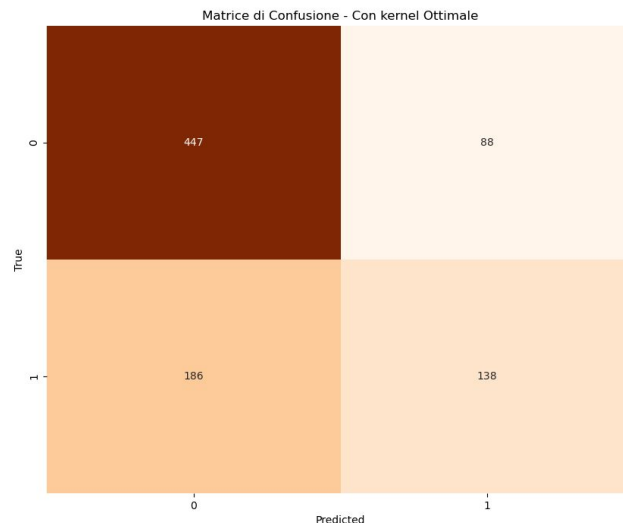
Modello scelto: **SVM** lineare, configurazione ottimale:  
**{'kernel': 'linear', 'C': 10}**

**Accuratezza test set: 68.10%**

Circa 2 prenotazioni su 3 sono classificate correttamente.

**PPV** - Prenotazioni non cancellate:  $TP/(TP+FP)=83.5\%$

**NPV** - Prenotazioni cancellate:  $TN/(TN+FN)=42.6\%$



# Classificazione

## Tuning degli Iperparametri (Hotel Bookings Analysis)

Kernel	Degree	C	Gamma	Accuratezza Media sulla Validation Set
SVM Linear	N/A	10	N/A	0.7051
SVM Poly	2	10	N/A	0.6856
SVM Poly	3	10	N/A	0.6844
SVM Rbf	N/A	10	'scale'	0.6696



# Studio statistico sui risultati della valutazione

## (Hotel Bookings Analysis)

Per dare una corretta valutazione del modello, vengono ripetuto le fasi di addestramento e testing  $k=20$ . Il `random_state` viene iterato  $k$  volte per individuare il valore migliore.

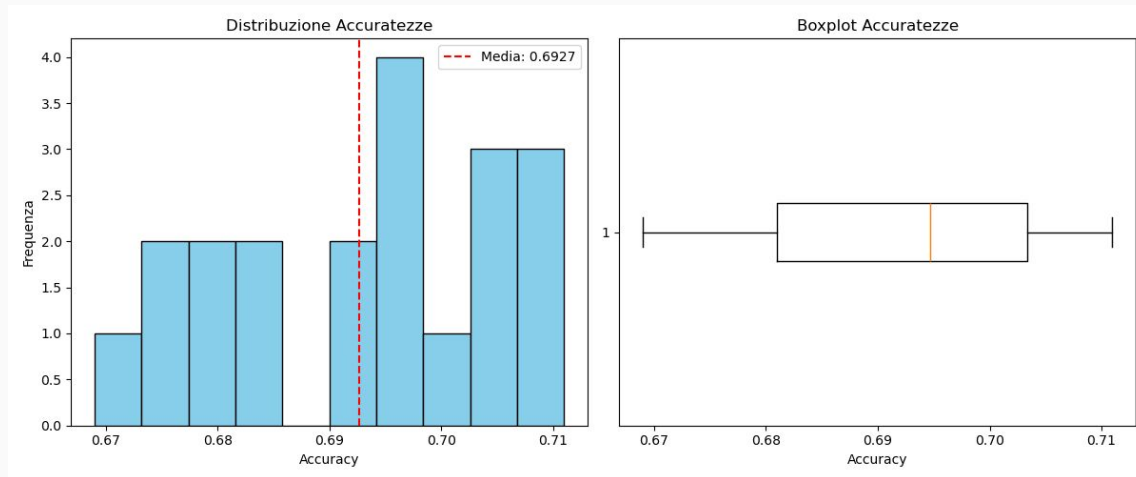
### Statistiche delle accurtezze:

- ❑ Media: **0.6927**
- ❑ Deviazione standard: **0.0125**
- ❑ Minimo: **0.6690**
- ❑ Massimo: **0.7110**
- ❑ Mediana: **0.6946**

== INFERENZA STATISTICA ==

Intervallo di confidenza al 95.0%:

(0.6867, 0.6987)



# Regressione Lineare Semplice

## (Brain Weight in Humans)

Variable indipendente -> Head Size(cm<sup>3</sup>)

Variable dipendente -> Brain Weight (grams)

Intercetta ( $\beta_0$ ): 319.83 grams

Pendenza ( $\beta_1$ ): 0.2648 grams

Coefficiente di semplice di determinazione ( $R^2$ )(test set): 0.70

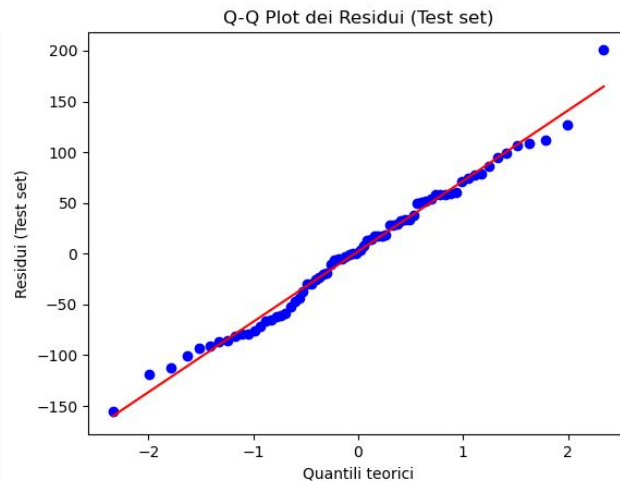
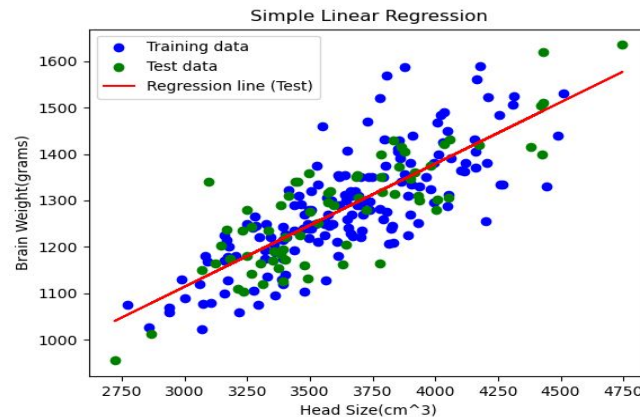
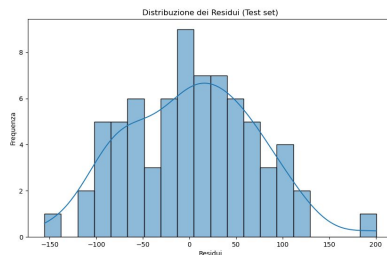
MSE: 4617.92

Per ogni aumento di 1 cm<sup>3</sup> nella dimensione cranica, il peso del cervello aumenta in media di 0.2648 grammi.

Analisi di Normalità dei residui (Test set)

Shapiro-Wilk p-value: 0.8087

I residui seguono una distribuzione normale (non rifiutiamo H0)



$$\varepsilon_i = y_i - (\beta_1 x_i + \beta_0)$$