Data Science

Esercizi primo compitino - serie 1 - 2019

(serie da svolgere in aula)

Nome:

Cognome:

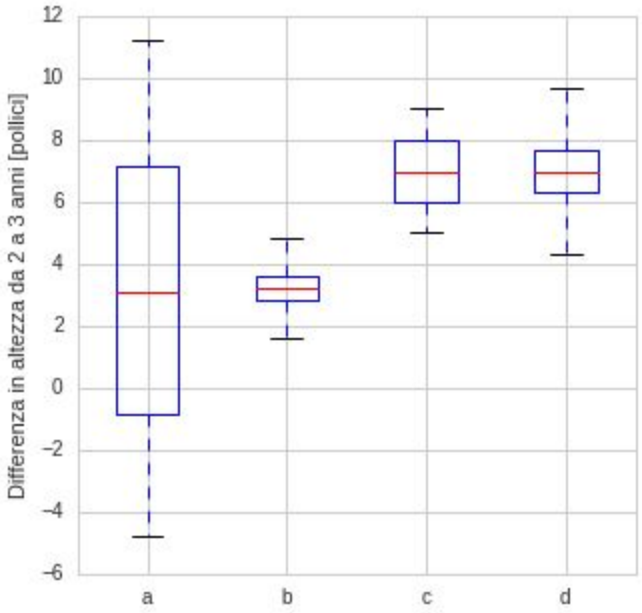
Esercizio 1

Considera il grafico “CDC Growth Charts: United States”, disponibile a questo link: https://www.cdc.gov/growthcharts/data/set1clinical/cj41l021.pdf

1.1- Ipotizza venga fornito un campione di 10000 ragazzi statunitensi di 20 anni. Disegna approssimativamente il boxplot delle loro altezze, disegnando i baffi con la convenzione di Tukey (come mostrato nelle slide), e senza disegnare gli outlier. Puoi esprimere i valori in pollici (inches) per comodita’.

1.2- Viene fornito un campione di 20000 ragazzini statunitensi: 10000 hanno appena compiuto 8 anni e 10000 ne hanno appena compiuti 12. Considera la popolazione nel suo complesso e disegna un singolo boxplot delle loro altezze.

1.3- Un campione di 1000 bambini viene seguito durante i primi 3 anni di vita. Consideriamo un dataset per cui ad ogni osservazione corrisponde uno di questi bambini: abbiamo a disposizione un attributo che corrisponde a quanto ogni bambino e' cresciuto tra il giorno in cui ha compiuto 2 anni e il giorno in cui ne ha compiuti 3, espresso in pollici. Quale dei seguenti e' un boxplot ragionevole per questo attributo? Giustifica la risposta.



Esercizio 2

In uno studio clinico si vuole analizzare l'impatto dell'attivita' fisica sulla salute della popolazione. Si considera un campione di 20 soggetti e si monitorano per un anno con un contapassi digitale. Alla fine dell'anno i pazienti vengono pesati e si ottengono i seguenti dati per i due attributi:

- Media di passi al giorno (espresso in migliaia di passi)
- Peso del paziente

ID soggetto	**2**	**13**	**3**	**11**	**0**	**1**	**15**	**17**	**14**	**5**	**12**	**7**	**18**	**8**	**6**	**9**	**19**	**16**	**4**	**10**
Passi [x1000]	0.15	0.20	2.10	2.32	2.48	2.91	3.32	3.59	3.97	4.49	4.60	5.41	7.07	8.45	9.52	9.87	13.44	14.33	15.00	24.11
Peso [kg]	88.13	82.77	81.86	85.40	91.08	84.15	83.39	90.44	81.38	84.93	82.09	81.73	87.78	81.26	78.07	78.70	73.77	74.46	67.17	66.01

Nota che nella tabella i pazienti sono stati ordinati in funzione del numero di passi.

2.1- Disegna uno scatter plot con la variabile "passi" sulla x e la variabile "peso" sulla y. Etichetta correttamente gli assi.

2.2- Stima visivamente (senza fare calcoli) il coefficiente di correlazione r tra i due attributi.

r =

2.3- Stima visivamente (senza fare calcoli) l'intercetta e la pendenza della retta, dando a entrambi i valori l'unita' di misura corretta.

Intercetta =

Pendenza =

2.4- Scegli un intervallo di discretizzazione ragionevole e rappresenta l'attributo "passi" per mezzo di un istogramma

2.5- Disegna un boxplot per l’attributo “passi”

2.6- Un giornalista scrive un articolo intitolato: “un nuovo studio dimostra che chi cammina di piu’ dimagrisce piu’ in fretta” e vi chiede un commento prima della pubblcazione. Scrivi una breve risposta al giornalista, spiegando come mai il titolo non e’ accettabile.

2.7- Ipotizza di aver calcolato il coefficiente di correlazione, l’intercetta e la pendenza della regressione lineare in modo esatto. Ci accorgiamo pero’ che alla fine dello studio il contapassi a causa di un problema hardware aveva contato solo la meta’ dei passi effettivamente svolti dalle persone: decidiamo quindi di correggere il dataset raddoppiando il valore dell’attributo “passi” per tutti i soggetti. Come cambia il valore del coefficiente di correlazione? Come cambia il valore del’intercetta? Come cambia il valore della pendenza?

2.8- L’ID di ogni soggetto e’ stato assegnato ai partecipanti in modo casuale. Che valore ti aspetti abbia il coefficiente di correlazione tra l’ID del soggetto (considerato come variabile numerica) e l’attributo peso? Giustifica la risposta

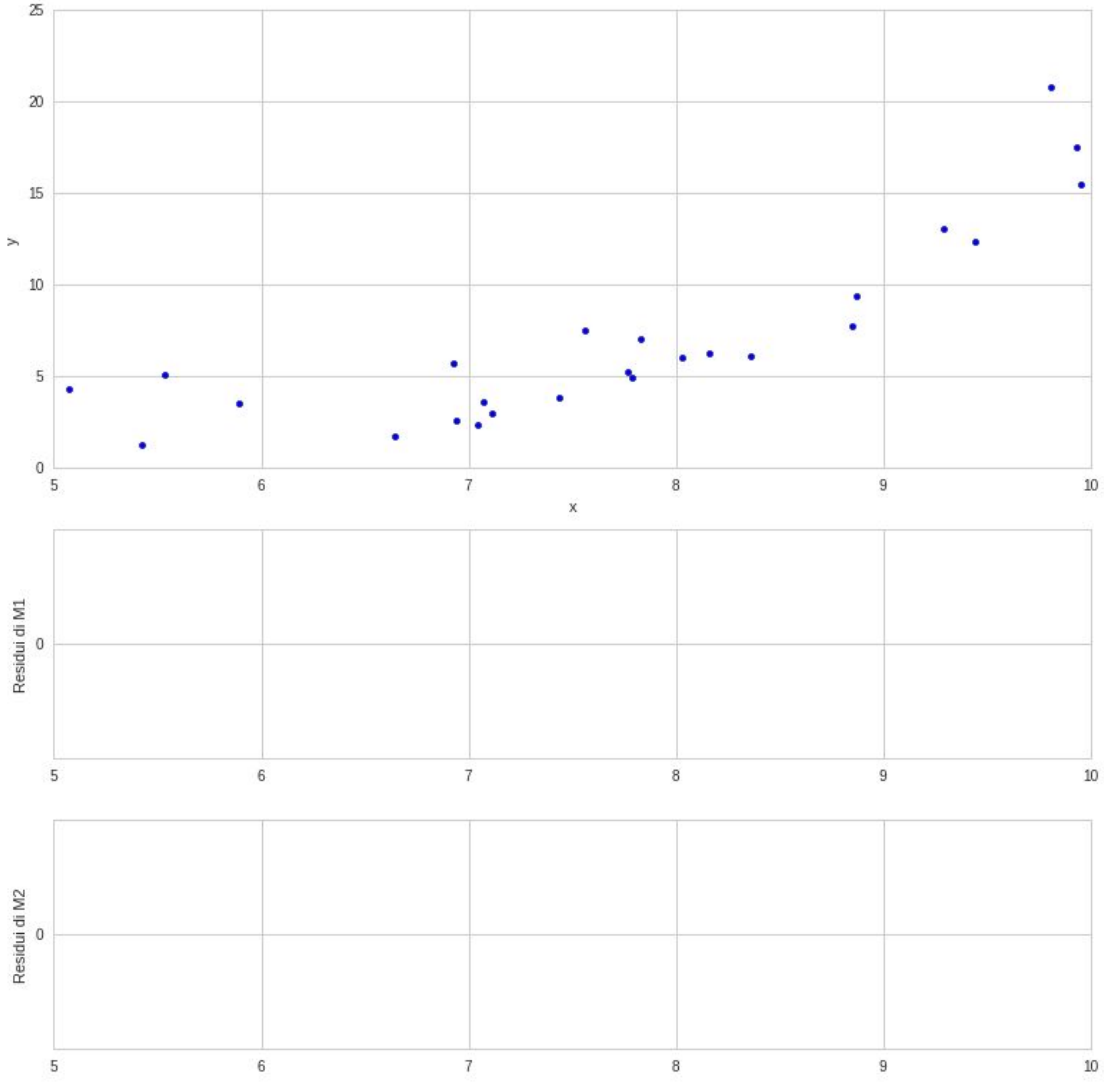
Esercizio 3

Considera lo scatter plot seguente, nel quale sono riportati i valori delle variabili x (variabile esplicativa) e y (variabile target) per un dataset di 25 osservazioni.

Considera i seguenti due modelli di regressione:

M1: y = x - 3
M2: y = (x-6)^2 + 2

3.1- Disegna (in modo approssimato) il grafico dei residui per entrambi i modelli nello spazio apposito. Etichetta correttamente l'asse delle ordinate e presta attenzione alla sua scala (e' marcata solo la posizione di y=0).



3.2- Commenta i due grafici. Quale dei due modelli consideri preferibile?

Esercizio 4

E' dato il seguente dataset con un attributo target (y) e 6 attributi esplicativi (x1,x2,x3,x4,x5,x6).

	x1	x2	x3	x4	x5	x6	y
0	9.5	1.4	8.7	2.9	3.8	7.9	22.9
1	7.8	6.6	6.0	1.3	9.5	8.0	36.4
2	7.0	8.5	8.4	9.5	1.1	8.2	11.5
3	7.2	0.5	8.8	2.2	8.1	6.7	32.0

Si propongono due modelli di regressione multipla:

M1: y = x1 + 3*x5 + 1

M2: y = x2 + 3*x5 + 1

4.1- Per ciascun modello e per ciascuna osservazione, calcola l'errore.

4.2- Per ciascun modello, calcola la somma degli errori al quadrato (SSE). Discuti quale dei due modelli e' preferibile.

Esercizio 5

L'ingegner Pippo ha un grosso allevamento di greebly da latte (una specie molto esotica). Un giorno ti dice: "*Compro tutti i miei greebly gia' adulti, e li tengo tutti per un mese per poi venderli come animali da compagnia (dopo un mese, infatti, i greebly non possono essere piu' munti). Alcuni greebly producono tanto latte, altri ne producono molto meno. Guadagnerei un sacco di soldi in piu' se potessi prevedere, prima di comprarli, quali greebly produrranno tanto latte e quanti ne produrranno meno. Purtroppo, l'unica cosa che posso sapere prima di comprare un greebly e' il suo peso. Allora, ho raccolto un dataset in cui per ciascuno dei 10000 greebly che ho comprato negli ultimi anni ho riportato: il peso comunicatomi dal venditore e la quantita' totale di latte che il greebly ha prodotto nel mese in cui l'ho tenuto. Ho calcolato il coefficiente di correlazione tra il peso del greebly e la quantita' di latte prodotta, e ho ottenuto un valore piccolissimo in valore assoluto, praticamente pari a 0. Ho anche provato a calcolare la media della produzione di latte tra i 5000 greebly con peso maggiore, e la media tra i 5000 greebly con peso minore: entrambe le medie sono praticamente uguali. Ma ho comunque incaricato un esperto Data Scientist di analizzare il mio dataset, perche' non ho perso le speranze. Ho fatto bene, secondo te?*"

5.1 Dai una risposta concisa ma giustificata all'ingegner Pippo.