Introduzione al dataset

I dati presenti all’interno del dataset provengono dal sito World Population Review, che a sua volta fa uso di dati provenienti dalle Nazioni Unite e dagli istituti di statistica di ciascuna nazione. Il dataset contiene principalmente informazioni relative ai valori della popolazione negli stati e nei territori del mondo, ed alcune informazioni di categoria più generale come il continente, l’area, la capitale e la densità abitativa nella nazione. Il dataset è quindi composto da record ciascuno descritto da 17 attributi, ed è raggiungibile nel [sito web](https://www.kaggle.com/datasets/iamsouravbanerjee/world-population-dataset?resource=download).

Analisi dei dati

Questo capitolo si occupa dell’analisi dei dati del dataset e della valutazione della loro qualità

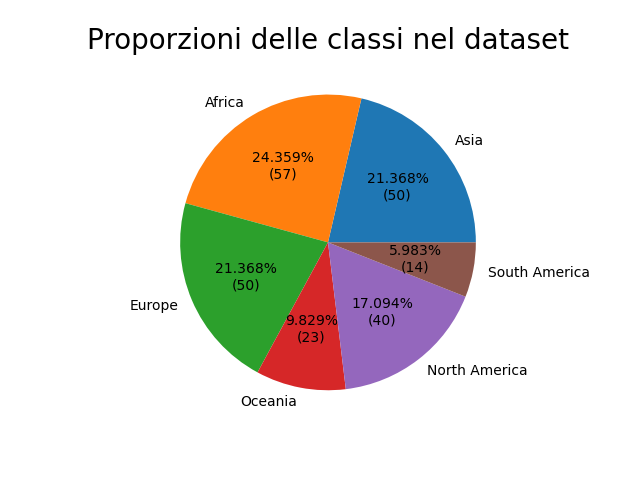
**Descrizione**

Il dataset è una matrice con 234 righe, una per record, e 17 colonne. 16 di queste colonne sono dedicate agli attributi, dei quali: 3 sono categorici (“*CCA3”*, “*Country/Territory”*, “*Capital”*), 3 sono *float* (“*Density”*, “*Growth Rate”*, “*World Population Percentage”*), mentre gli altri sono interi. Si hanno quindi solo 2 attributi continui, al contrario della maggioranza discreta. Inoltre, non sono presenti dati mancanti.

|  |  |  |  |
| --- | --- | --- | --- |
| **Nome dell’attributo** | **Tipo** | **Descrizione** | **Unità** |
| Rank | Discreto | Classifica sulla base della popolazione |  |
| CCA3 | Nominale | Codifica a 3 lettere dei nomi delle nazioni e dei territori |  |
| Country/Territories | Nominale | Nome delle nazioni e dei territori |  |
| Capital | Nominale | Nome della capitale |  |
| 2022 Population | Discreto | Popolazione nell’anno 2022 |  |
| 2020 Population | Discreto | Popolazione nell’anno 2020 |  |
| 2015 Population | Discreto | Popolazione nell’anno 2015 |  |
| 2010 Population | Discreto | Popolazione nell’anno 2010 |  |
| 2000 Population | Discreto | Popolazione nell’anno 2000 |  |
| 1990 Population | Discreto | Popolazione nell’anno 1990 |  |
| 1980 Population | Discreto | Popolazione nell’anno 1980 |  |
| 1970 Population | Discreto | Popolazione nell’anno 1970 |  |
| Area | Discreto | Superficie | Km2 |
| Density | Continuo | Densita della popolazione | Ab./Km2 |
| Growth Rate | Continuo | Tasso di crescita |  |
| World Population Percentage | Continuo | Percentuale della popolazione mondiale | % |

Gli attributi numerici sono di tipo ratio, mentre i categorici sono di tipo nominale – non è presente un ordine intrinseco nei nomi delle capitali e delle nazioni né negli acronimi usati per rappresentarle. Il vettore delle classi (“*Continent”*) è composto da 6 classi, ciascuna rappresentante il continente di appartenenza della nazione o del territorio.

Il significato di ogni attributo è descritto nella tabella riportata

**Risultati**

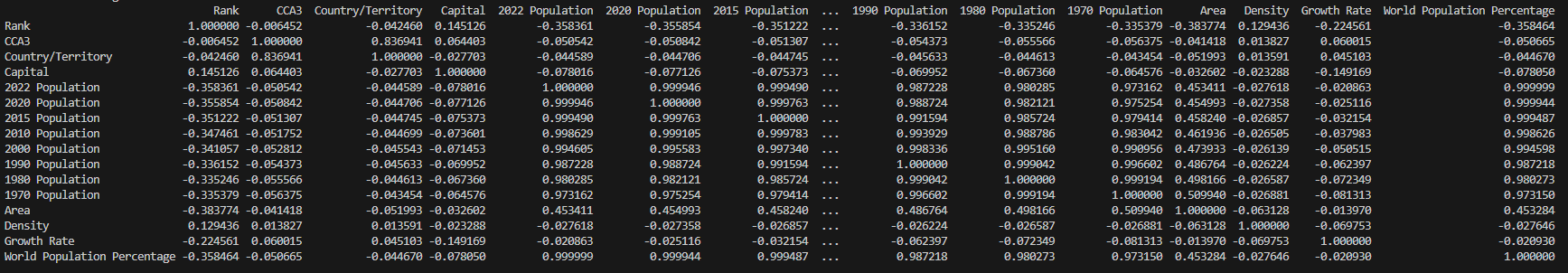
Da una prima analisi emerge che si tratta di un dataset sbilanciato, in quanto si hanno: 57 record appartenenti alla classe “*Africa”* (24.36%), 50 alla classe “*Asia”* (21.37%), 50 alla classe “*Europe”* (21.37%), 40 alla “*North America”* (17.09%), 23 alla “*Oceania”* (9.83%) e 14 alla “*South America”* (5.98%). Si nota infatti una differenza significativa nel numero di record appartenenti alle varie classi – ma questo squilibrio non è causato dal metodo di raccolta dati, in quanto è noto che siano queste le suddivisioni geografiche delle nazioni.

L’analisi dei dati è stata eseguita tramite la funzione *data\_analysis()* tratta dal file dataAnalysis.py, costruita appositamente per questo dataset. Questa lo prende in ingresso e, per ogni attributo numerico calcola media, moda, mediana, deviazione standard, varianza, range, ed i percentili del 10%, 25%, 50%, 75%, 90%. Gli attributi categorici sono stati inclusi nell’analisi a seguito di una conversione a valori numerici – necessaria per poterne analizzare la correlazione con gli altri. Questi dati vengono poi stampati sul terminale dalla funzione *print\_analysis()*.

Nel file sono presenti anche altre tre funzioni, relative alla creazione dei grafici utili all’analisi del dataset. *boxplot()* crea una serie di box plot dei 13 attributi numerici. Per consentire una lettura più agevole dei grafici la funzione li stampa 2 volte – nella prima non considera gli outlier, consentendo di visualizzare il centro e la distribuzione dei dati; nella seconda vengono rappresentati anche i valori estremi, per identificarli e rendere evidenti gli effetti. La funzione *histogram()* genera un subplot con 13 istogrammi, che consentono di analizzare l’andamento di ciascun attributo in base alla classe di appartenenza – fondamentale per individuare gli attributi che possono contribuire di più nei modelli. La terza, *pie\_chart()* genera un grafico a torta per rappresentare la percentuale delle classi presenti nel database, e si appoggia sulla funzione *pct\_chart()* per una stampa migliore delle stesse.

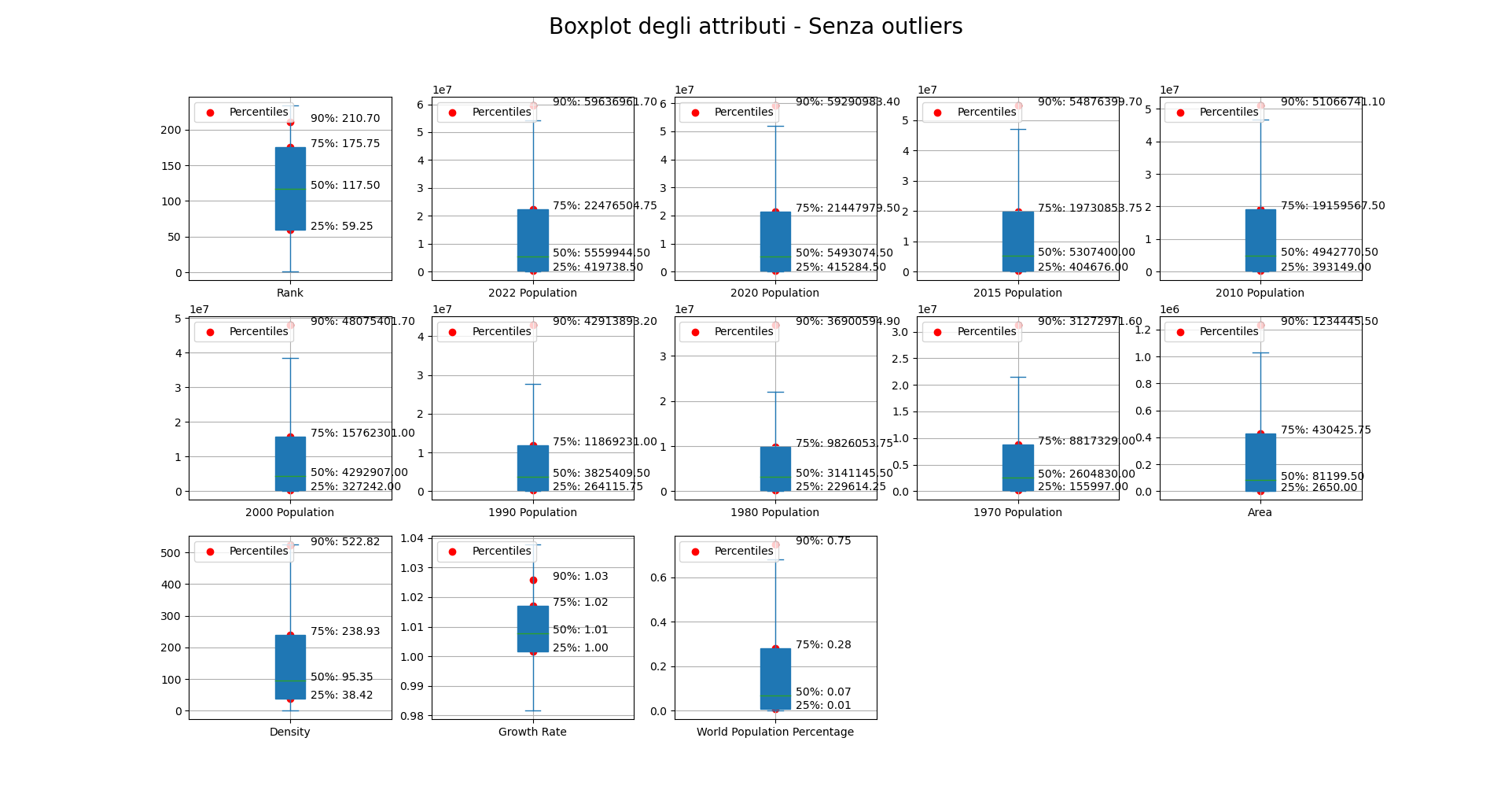
A black screen with white text

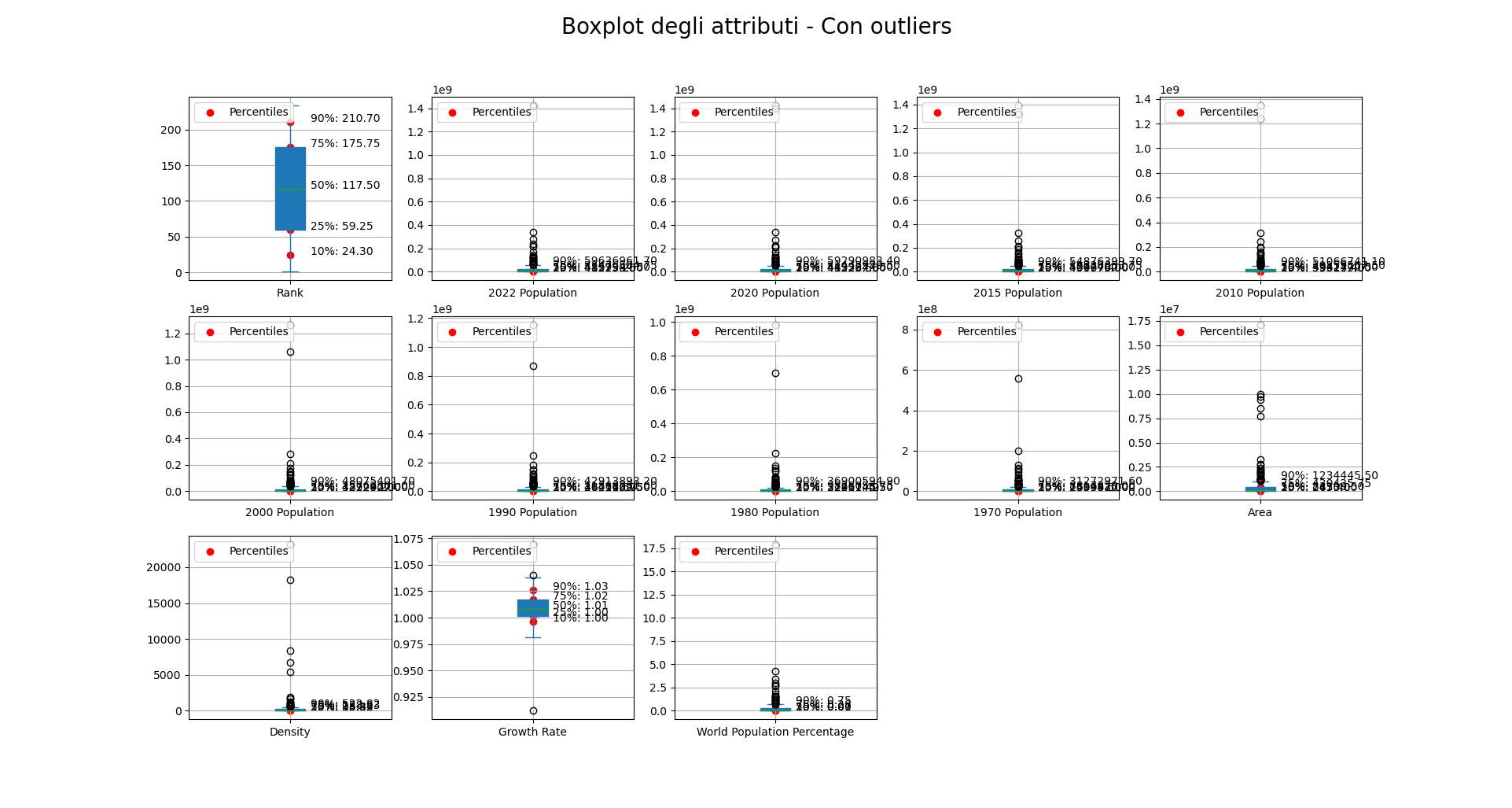
AI-generated content may be incorrect.I risultati della funzione *data\_analysis()* sono disponibili nella tabella seguente.

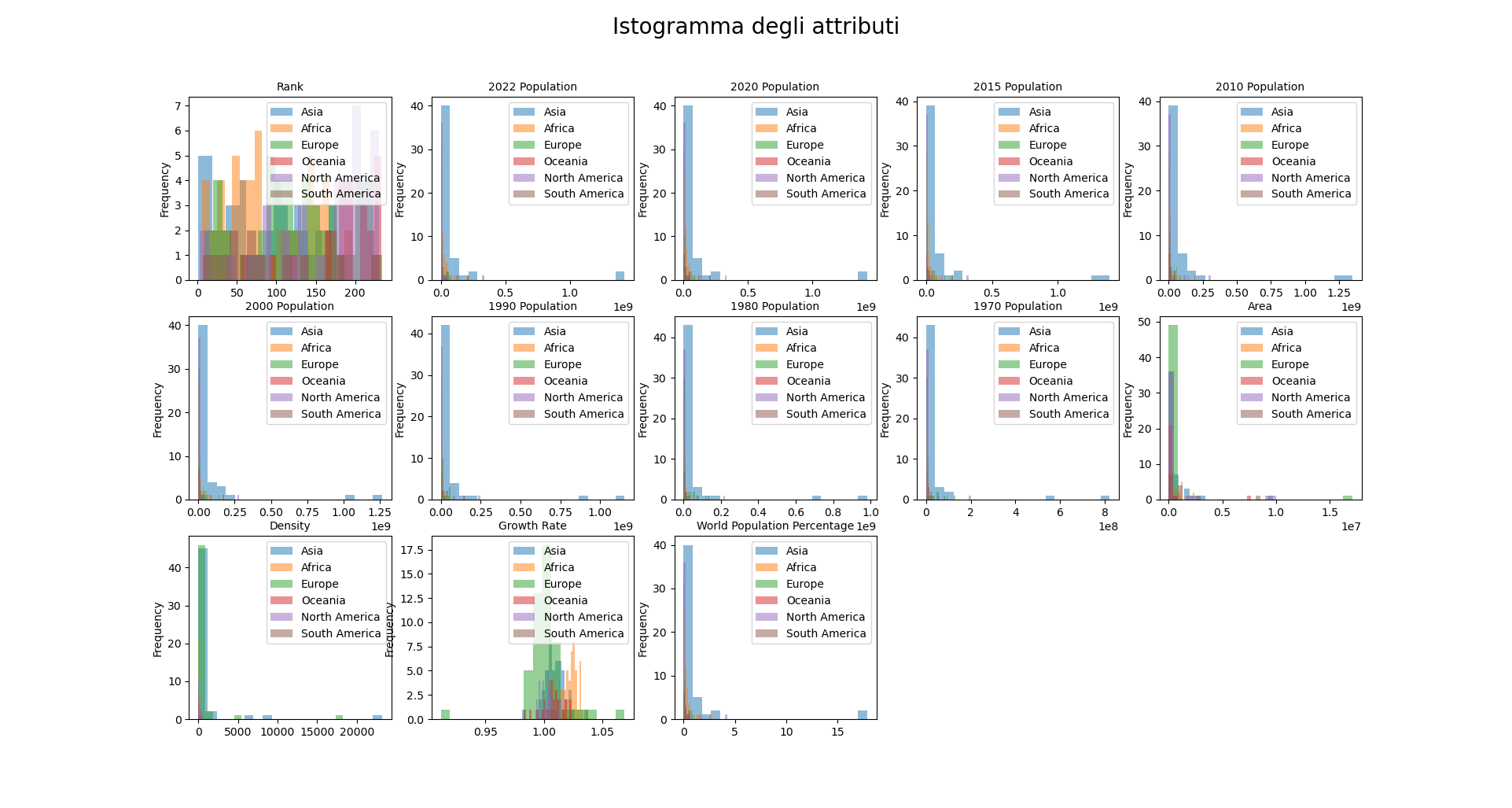
Mentre la correlazione lineare (calcolata secondo il metodo di Pearson) è in quest’altra tabella.

Gli output della funzione *boxplot()*, risultano utili per visualizzare graficamente la distribuzione dei dati. Consento infatti di verificare in maniera rapida la normalità dei dati e la presenza degli outlier. La linea centrale viene usata per rappresentare la mediana, mentre la parte inferiore e superiore della scatola mostrano il 25esimo e 75esimo percentile. La posizione della mediana all’interno della scatola consente di verificare la simmetria dei dati: se è significativamente più vicina ad uno degli estremi della scatola, la distribuzione dell’attributo è asimmetrica. I baffi che partono dalla scatola si estendono per 1.5 volte la lunghezza della scatola, che rappresenta il range interquartile, ed i record che si trovano all’esterno dei baffi vengono detti outlier – in quanto sono al di fuori della variazione attesa. Inoltre nei grafici viene anche rappresentato il 90esimo percentile, identificato come un pallino rosso.

Questa prima figura mostra i box plot senza gli outlier.



Questa seconda figura mostra invece l’effetto degli outlier sugli attributi.

Nella figura seguente si vede l’output della funzione *histogram()*, che consente di ottenere una idea della distribuzione dei dati.

**Discussione**

Fra gli attributi si possono identificare alcuni gruppi con caratteristiche simili.

Vi sono gli attributi categorici (“*CCA3”*, “*Capital”*, “*Country/Territory”*) che insieme con Rank come si può vedere dalla mappa delle correlazioni, sono poco legati agli altri. Le eccezioni sono “*CCA3”* e “*Country/Territory”*, che mostrano una forte correlazione tra di loro. Tutti gli attributi in argomento presentano un’alta varianza, dovuta al fatto che non sono presenti valori ripetuti.

Un secondo gruppo è quello composto dagli attributi legati alla popolazione nel corso degli anni. Questi mostrano una correlazione molto forte tra di loro e con la “*World Population Percentage”*, ed una correlazione elevata, anche se meno forte rispetto alle precedenti, con l’attributo “*Area”*. Risulta interessante notare come quest’ultima correlazione diminuisca negli anni, partendo da uno 0.51 con la popolazione degli anni 70 fino allo 0.45 con la popolazione del 2022. Analizzando i box plot senza outlier relativi agli attributi si nota anche una forte asimmetria nella distribuzione, con la mediana fortemente spostata verso il basso, quindi di tipo positivo. Dal secondo set di box plot si nota anche come gli outlier dominino tutte le misure non robuste alla loro presenza – due valori infatti svettano molto lontani da tutti gli altri, schiacciando i grafici verso il basso. Si nota inoltre come il minore dei due raggiunga l’altro con il tempo. Anche gli istogrammi mostrano questa forte asimmetria a destra e consentono di identificare i valori anomali come appartenenti alla classe “*Asia”*. I valori alti di varianza e deviazione standard mostrano il risultato dell’estrema variazione dei valori all’interno delle classi – come anche il range che supera 1.4 miliardi. Anche i percentili calcolati mostrano questa tendenza, con quartili che nel 2020 vanno dal 415,284.5 al 25% a 21,447,979.5 al 75%

L’attributo “*Area”* si comporta in maniera simile al gruppo delle popolazioni, con una forte asimmetria positiva e un outlier che domina tutti gli altri. Al contrario della popolazione si può distinguere anche un secondo gruppo di valori molto più alti della norma, che formano un gruppo a metà strada tra l’estremo ed il resto dei valori. In questo caso l’istogramma mostra che il massimo appartiene alla classe “*Europe*”, mentre il secondo gruppo è composto da record appartenenti ad altre 3 classi: “*North America”*, “*South America”* ed “*Oceania”*. Come per il gruppo delle popolazioni la varianza e la deviazione standard sono molto alte, ma analizzando i percentili si nota che gli outlier hanno una forte influenza su questi valori.

La feature “*Density”* si dimostra poco correlata con gli altri attributi, e mostra anch’essa una forte asimmetria positiva con la presenza di due gruppi di valori estremi che si distaccano nettamente dagli altri. In entrambi i gruppi i valori appartengono alle classi di “*Europe”* ed “*Asia”*. I dati numerici mostrano una deviazione standard bassa rispetto a quella degli altri attributi, risultato dovuto al minore range dei valori – in particolare quelli che appartengono allo scarto interquartile. Questa minore variabilità si nota anche nel box plot senza i valori estremi dove, al contrario degli altri attributi discussi in precedenza, il 90esimo percentile si trova sul baffo, seppure al limite.

*“Growth Rate”* mostra correlazione simile all’attributo precedente, ma si ha solo una leggera asimmetria positiva, con outlier che, pur avendo una forte influenza sui risultati di molti dei valori dell’analisi, non si mostrano capaci di dominarla del tutto come nel caso tutti gli attributi precedenti. Il box plot mostra infatti solo 3 valori estremi, ed in questo caso si trovano oltre entrambi i lati dei baffi – distribuzione che può diminuirne l’impatto su alcune misure di posizione. L’istogramma consente inoltre di vedere una distribuzione gaussiana dei valori, sebbene con una “campana” molto stretta. Una minore variabilità che si rispecchia anche in valori bassi sia per la deviazione standard che per la varianza, con uno scarto interquartile molto basso.

Infine, “*World Population Percentage”* si comporta in maniera simile alle popolazioni: fortemente correlato alle popolazioni, una buona correlazione con l’area, asimmetria fortemente positiva, con due outlier che dominano il resto dei valori – entrambi appartenenti alla classe “*Asia”*. Si differenzia dalle popolazioni nel caso dei valori assoluti per la natura stessa dell’attributo, ma mostra comunque una alta deviazione standard considerando i valori di partenza ed un range interquartile con 0.01 al 25% e 0.28 al 75%.

In sintesi, si osserva che il dataset è sbilanciato, con una sottorappresentazione di “*North America”*, “*South America”* ed “*Oceania”*. Un gruppo di attributi è di scarso interesse in quanto composto da valori unici. Gli attributi legati alla popolazione nel corso degli anni ed il “*World Population Percentage”* presentano informazioni con correlazioni così alte da risultare poco significativi. Le caratteristiche degli attributi si riflettono anche sulle distribuzioni: quasi tutte non gaussiane, fortemente asimmetriche e con la presenza di outlier che non possono essere ignorati dallo studio. Si notano anche l’assenza del rumore e la scarsa correlazione lineare tra la maggior parte degli attributi.