Progetto finale

Laboratorio di Big Data Analytics

Juri Cicalè, Marco Antonio Lepore, Manuel Manzo

Università Politecnica delle Marche

Laboratorio di Big Data Analytics

Prof. Luca Virgili

Indice

Qlik

Descrizione del dataset

Il dataset utilizzato raccoglie informazioni relative all'attività di Airbnb nel comune di Roma, nello specifico i dati relativi alle caratteristiche del singolo alloggio (Host, quartiere, coordinate, tipologia alloggio, livello recensioni ecc..).

I dati di questo dataset sono estati estratti dal portale in data 9 dicembre 2022

Fonte: Airbnb in Italy | Kaggle

Approfondimento delle variabili:

Poiché il dataset originario presentava 75 variabili si è deciso di utilizzare su qlik solamente una parte di queste, che vengono di seguito indicate:

id	Identificativo dell'annuncio
host_id	Identificativo dell'host
host_since	Data in cui l'host ha pubblicato il primo annuncio
host_response_time	Tempo medio di risposta dell'host
host_response_rate	Tasso medio di risposta dell'host
host_acceptance_rate	Tasto medio di accettazione dell'host
host_is_superhost	L'host è un superhost?
neighbourhood	Quartiere in cui si trova l'alloggio oggetto dell'annuncio
latitude	Latitudine
longitude	Longitudine
bedroom	numero stanze
beds	Numero dei letti
price	Prezzo per notte

minimum_nights	Notti minime per soggiornare nell'alloggio
maximum_nights	Notti massime per soggiornare nell'alloggio
availability_30	Giorni in cui l'alloggio è stato disponibile per essere affittato negli ultimi 30 giorni
availability_90	Giorni in cui l'alloggio è stato disponibile per essere affittato negli ultimi 90 giorni
availability_365	Giorni in cui l'alloggio è stato disponibile per essere affittato negli ultimi 365 giorni
first_review	Data in cui è stata inserita la prima recensione
review_score_rating	Punteggio della recensione riguardante l'esperienza complessiva
review_score_accuracy	Punteggio della recensione riguardante l'accuratezza dell'annuncio
review_score_cleanliness	Punteggio della recensione riguardante il livello di pulizia
review_score_ckeckin	Punteggio della recensione riguardante la procedura di Check-in
review_score_communication	Punteggio della recensione riguardante la puntualità dell'*host* nel rispondere ai messaggi
review_score_location	Punteggio della recensione riguardante la posizione dell'alloggio
review_score_value	Punteggio della recensione riguardante il rapporto qualità prezzo
Revier_per_month	Numero di recensioni in media per mese

Trasformazione dei dati

In una fase precedente al caricamento dei dati su Qlik Cloud sono state effettuate le seguenti modifiche sui dati:

- Eliminazione delle osservazioni che non presentavano valori nella colonna del quartiere;
- Eliminazione delle osservazioni associate a quartieri con meno di 2 annunci;
- Eliminazione di alcune osservazioni con prezzo errato (ad esempio 91.000 euro per notte per un monolocale);

Obiettivo dell'analisi

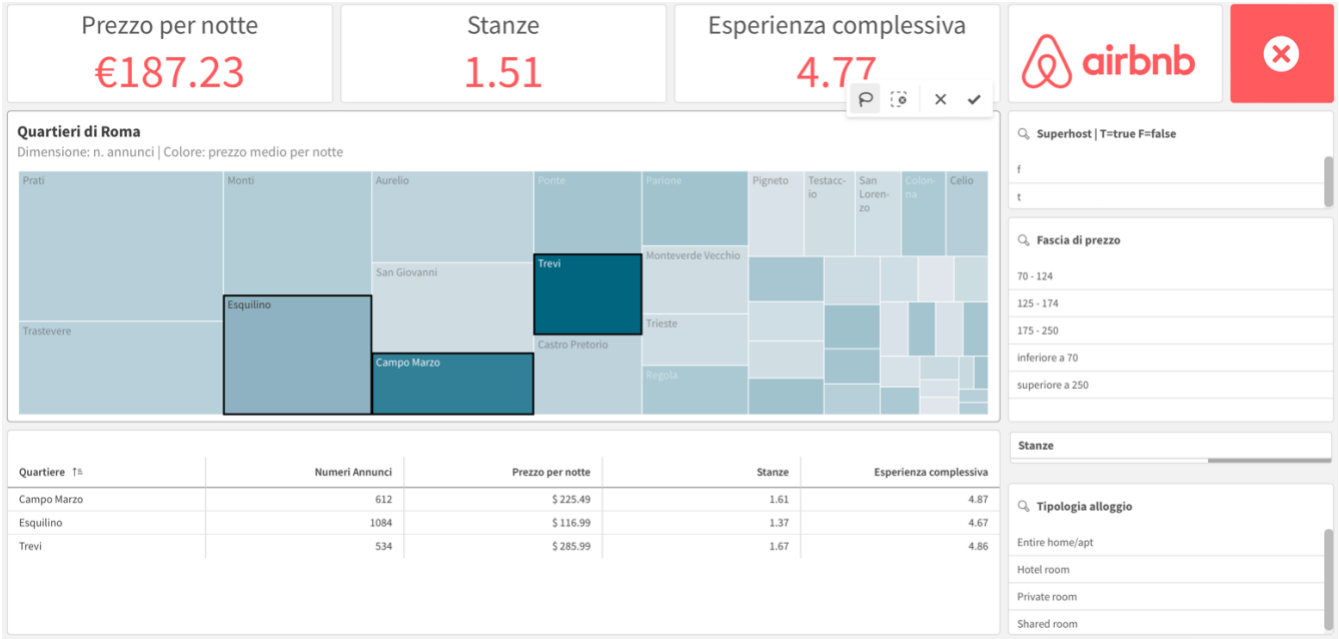
Obiettivo del presente lavoro è quello di implementare uno strumento di analisi relativamente ai dati operativi del portale Airbnb.

Attualmente l'azienda non è dotata di tale tool analitico e l'esigenza sembra palesarsi con evidenza osservando il fiorire di portali terzi che mettono a disposizione questo tipo di servizi a pagamento.

Si è valutata come ideale, per il raggiungimento dell'obiettivo sopraindicato, l'articolazione del lavoro in 4 *Dashboard* distinte, ognuna con un focus analitico specifico ben definito.

Descrizioni delle dashboard:

Offerta per quartiere



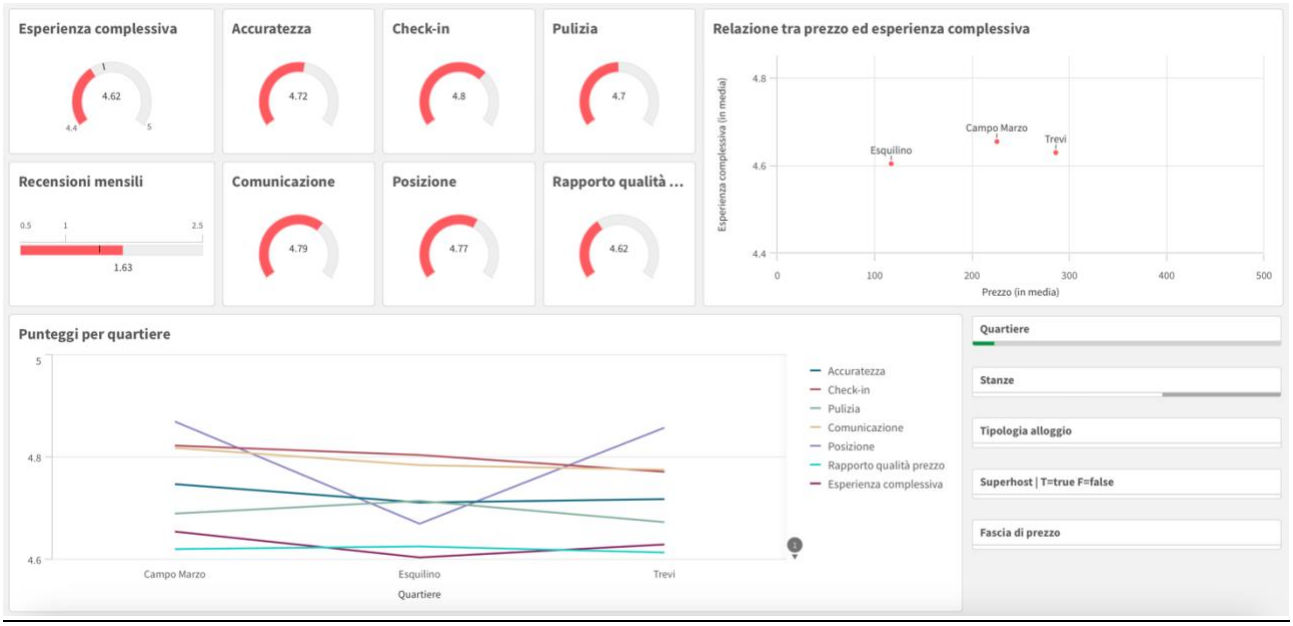
La prima dashboard si concentra nel fornire una visione specifica dell'offerta per quartiere, possiamo vedere quindi i dati generali relativi agli alloggi ed alle esperienze degli ospiti; nello specifico si esplicitano i seguenti relativi KPI: prezzo medio per notte, numero stanze, valutazione recensione.

Per la visualizzazione immediata dei singoli quartieri si è scelto di utilizzare lo strumento della mappa ad albero che attraverso la dimensione dei quadranti (numero annunci) e la sfumatura del colore (livello di prezzo) ne caratterizza la specifica offerta sul portale.

La stessa mappa ad albero ed una serie di filtri permettono di *customizzare* l'analisi in base al quartiere, tipologia di alloggio, numero di stanze, *range* di prezzo e presenza o meno di un *host* esperto

Per approfondire in maniera dettagliata le informazioni meno immediate è stata creata *Ad Hoc* una tabella personalizzabile nei campi (chiamata da qlik "*straight table"*) e facilmente consultabile.

Esperienza degli ospiti



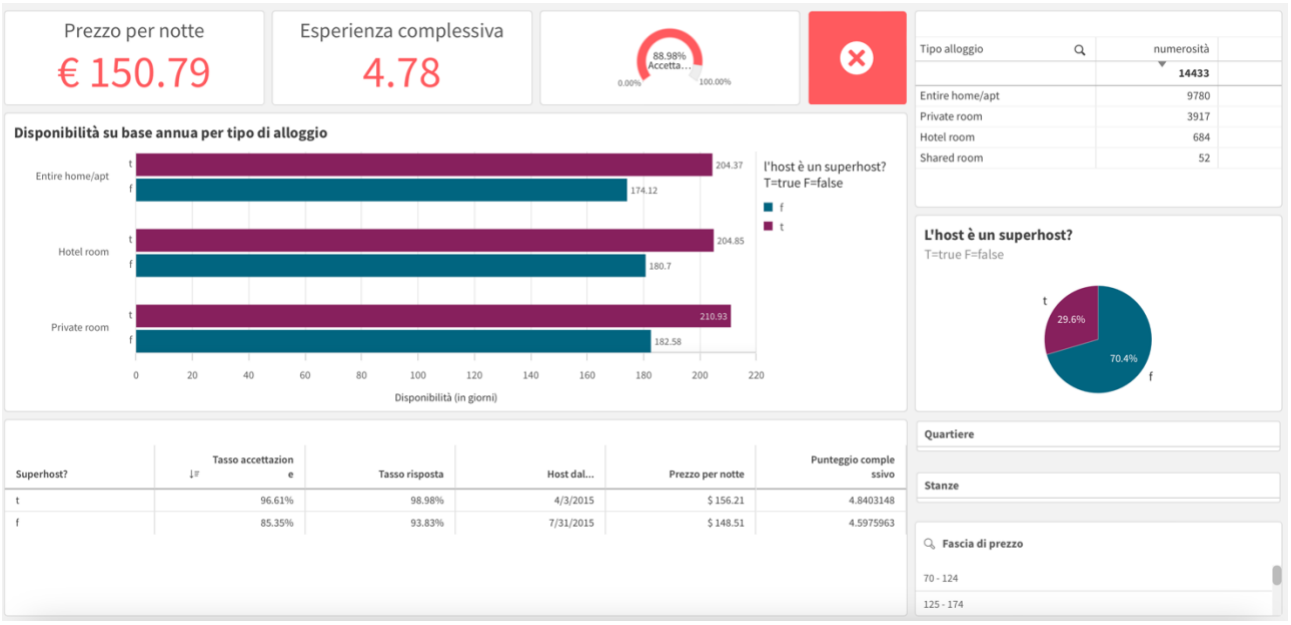
Nella seconda Dashboard si è approfondito il tema dell'esperienza degli ospiti attraverso la visualizzazione immediata delle valutazioni specifiche espresse nelle recensioni (accuratezza, check-in, pulizia, comunicazione, posizione e rapporto qualità-prezzo). Si è scelto di utilizzare l'oggetto qlik "misuratore" per ottenere dinamica ed immediatezza, in relazione al livello dello score ottenuto.

Per confrontare rapidamente i valori dei rispettivi quartieri, gli score sono stati utilizzati anche per generare un grafico lineare posto nella parte inferiore della visualizzazione.

Il rapporto qualità-prezzo è stato esplicitato attraverso la costruzione di uno Scatterplot con le variabili prezzo e score dell'esperienza complessiva.

Anche in questo caso, una serie di filtri permette di personalizzare ed articolare in modo più specifico la ricerca.

Informazioni sugli alloggi e gli host



Nella terza dashboard si vanno ad approfondire le informazioni relative agli alloggi, nello specifico si è provveduto ad inserire nella visualizzazione: le caratteristiche degli host, la disponibilità media annua ed il *room type*.

Come KPI si è messo in evidenza attraverso delle card: il prezzo medio per notte, esperienza complessiva ed il tasso di risposta dell'*host*.

Per quanto riguarda le caratteristiche degli *host*, attraverso un grafico a torta interattivo, è possibile notare quanti di questi vengano riconosciuti dal portale come di "livello superiore" (i cosiddetti *superhost*) ed è possibile selezionarli per concentrare l'analisi su di loro.

Analogamente a quanto proposto nella dashboard 1 anche qui è stata inserita una tabella personalizzabile di tipo *straight table* per approfondire in maniera dettagliata le informazioni meno immediate

Simulatore di rendimento

Nell'ottica di fornire un servizio innovativo, nella dashboard 4 si fornisce uno strumento che consente a potenziali investitori e a coloro che sono interessati all'acquisto di un immobile di considerare le prospettive di rendimento dello stesso nel mondo degli *short term rentals*.

Attraverso una serie di caselle di testo compilabili dall'utente, quali: prezzo al mq dell'immobile, tasso di occupazione e prezzo medio a notte viene simulata un'ipotesi di guadagno che può invogliare il potenziale cliente a prendere in considerazione l'iscrizione al portale Airbnb.

Affinché questo tool funzionasse è stato necessario attuare una serie di ipotesi iniziali attinenti la dimensione dell'immobile (in linea con la dimensione media degli immobili di airbnb), le commissioni a carico dell'host per l'utilizzo della piattaforma, la tassazione e le spese di gestione.

Qualora Airbnb decidesse di mettere in produzione uno strumento simile a quello qui proposto potrebbe fornire confronti sulla redditività dei diversi quartieri, aiutando così potenziali investitori a identificare quelli più promettenti e redditizi. Ciò sarebbe realizzabile sfruttando una serie di variabili e risorse molto più ampie rispetto a quelle disponibili per questo progetto.

PowerBI

Descrizione del dataset

Il dataset utilizzato raccoglie informazioni relative allo sviluppo socioeconomico delle nazioni e, nello specifico, cerca di riassumerle in un indice detto HDI (*human development index*) basato sulle tre dimensioni ritenute fondamentali: istruzione, ricchezza e salute.

Le suddette dimensioni sono rappresentate nel dataset attraverso le seguenti variabili:

mean years of schooling	Numero di anni di scolarizzazione che un bambino in età di ingresso nella scuola può aspettarsi di ricevere
gni	Gross National Income pro capite: reddito aggregato di un'economia generato dalla sua produzione e dalla sua proprietà di fattori di produzione, meno i redditi pagati per l'utilizzo di fattori di produzione di proprietà del resto del mondo
life expectancy at birth	Numero di anni che un neonato potrebbe aspettarsi di vivere
human development index	Un indice composito che misura il rendimento medio in tre dimensioni fondamentali dell'essere umano sviluppo: una vita lunga e sana, conoscenza e un tenore di vita dignitoso
human development level	Livello dell'indice *human development index*
human development ranking	Posizione per *human development index*
gni rank minus hdi rank	Differenza nella classifica per RNL pro capite e per valore HDI.
country	Paese

Per completare in maniera adeguata i dati relativamente alle nazioni in gioco, attraverso il *web scraping* (nello specifico Fonte **GeoNames**) abbiamo integrato un nuovo dataset avente i campi di Ns interesse:

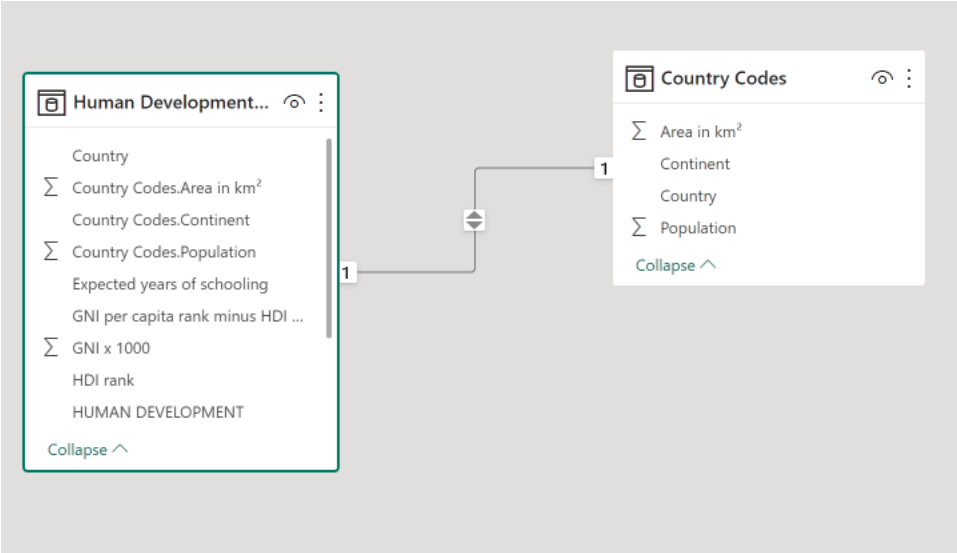
continente	Nome del continente
popolazione	Abitanti del continente
area	Estensione geografica del continente

Le due tabelle sono state messe in relazione attraverso il campo comune *"Country"* ed è stata

successivamente effettuata un'operazione di *merging* vista la dimensione esigua delle variabili in gioco nella seconda tabella.

Fonte: Human Development Index and Components | Kaggle

Relazioni tra le tabelle



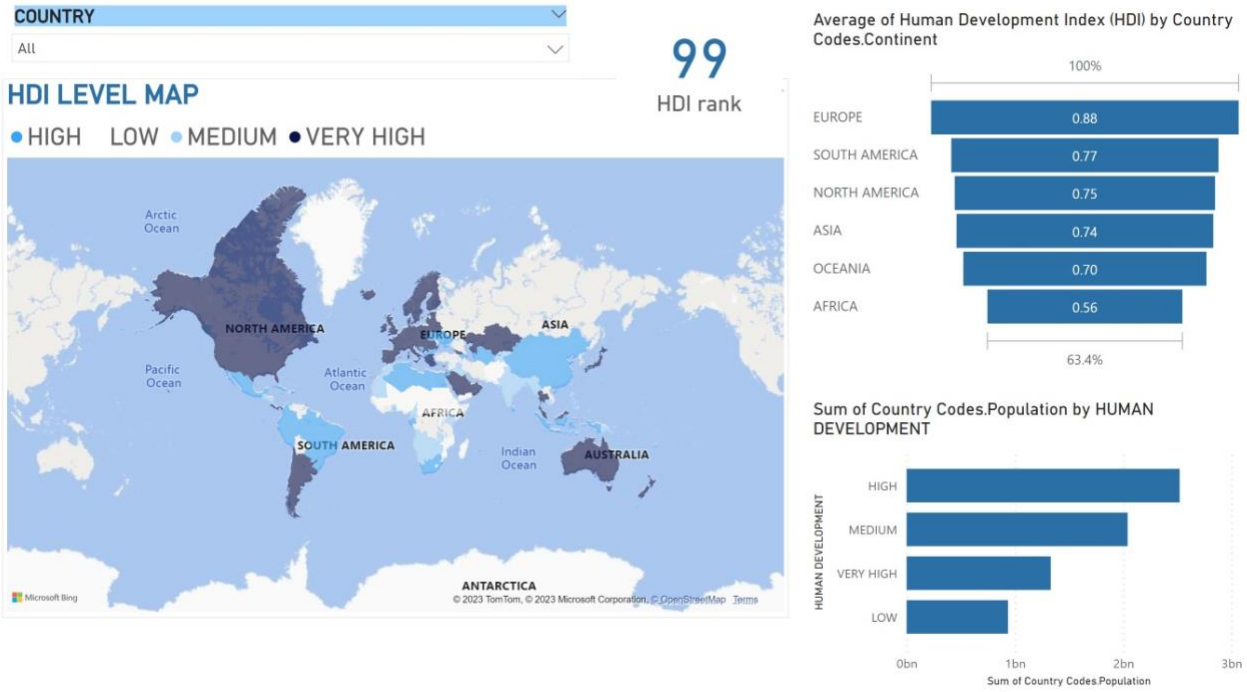
Obiettivo dell'analisi

Obiettivo dell'analisi è quello di visualizzare la situazione mondiale a livello di sviluppo umano (con un focus particolare sulle differenze territoriali), caratterizzarne le determinanti principali cogliendo spunti potenzialmente interessanti per una *call to action* volta a migliorare la situazione globale, indicando una via da seguire.

A livello operativo si è scelto di dare priorità assoluta all'immediatezza nella comprensione e nella visualizzazione dei punti chiave, impostando il lavoro con l'obiettivo primario di presentare i dati ad un pubblico terzo.

Descrizioni delle dashboard:

HDI Index nel mondo



In questa prima dashboard l'obiettivo è dare una visione immediata ed interattiva dell'HDI Index nel mondo.

Attraverso una mappa interattiva colorata (il colore identifica i 4 livelli di sviluppo umano: *low*, *medium*, *high* e *very high*) andiamo ad individuare i vari stati in relazione al loro HDI index, per lo scopo possiamo anche utilizzare un filtro a tendina che ci permette di individuare anche gli stati dei quali non conosciamo la posizione geografica.

Con una *card* molto evidente vediamo subito il ranking a livello di HDI Index dello stato selezionato.

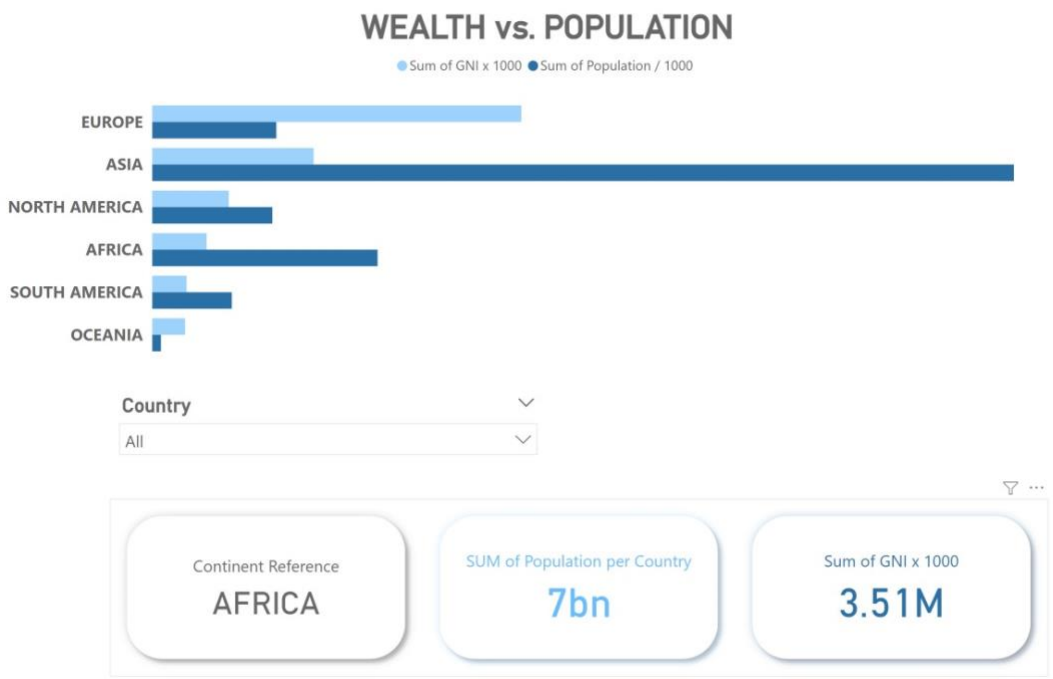
Attraverso un grafico ad imbuto abbiamo una visualizzazione immediata di quella che è la situazione HDI rispetto ai vari continenti e con un grafico a barre vediamo la popolazione mondiale suddivisa per le categorie di HDI, questo per renderci conto, a colpo d'occhio, della condizione nella quale vivono le persone nel mondo.

Andamento dell'HDI index

Nella seconda dashboard vogliamo visualizzare le determinanti dell'HDI cercando di coglierne gli andamenti. Per fare ciò si è ritenuto ideale l'utilizzo di uno *scatterplot* dove vediamo come l'HDI abbia una relazione lineare con il livello medio di educazione e con l'aspettativa di vita mentre fortemente non lineare con il livello di ricchezza, che incide molto nell'immediato ma sempre meno da un certo livello in poi.

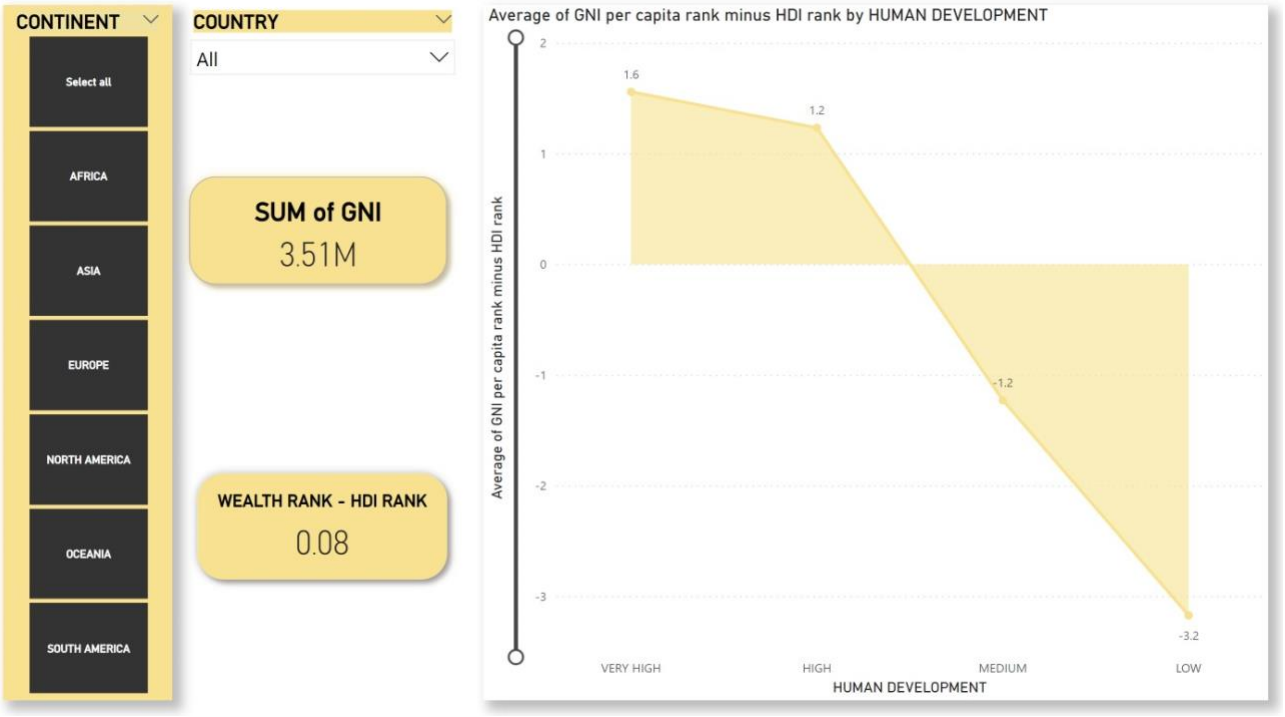
Per riassumere a livello analitico quanto immediatamente proposto negli *scatterplot* abbiamo inserito un oggetto *Powerbi* denominato "Fattori di influenza chiave".

Distribuzione della ricchezza



Nella terza dashboard abbiamo voluto spostare il focus sull'annoso tema delle sperequazioni di ricchezza a livello mondiale andando a rendere evidente il rapporto tra ricchezza e popolazione con un grafico a barre e delle card che ci restituiscono immediatamente il dato specifico di ogni nazione selezionata facilmente attraverso un apposito filtro.

Sviluppo vs ricchezza



Per concludere il lavoro vogliamo evidenziare un dato estremamente interessante e per certi versi controintuitivo: lo sviluppo umano non segue di pari passo la ricchezza di una nazione, è possibile notare, attraverso l'indicatore differenziale tra rating della ricchezza e rating HDI, che spesso la situazione economica non comporta una equivalente situazione di sviluppo umano (differenziale negativo), possiamo vederlo opportunamente su un grafico ad aree ed identificare il valore specifico su una card dedicata.

Questo ci porta a concludere con un richiamo specifico ed inderogabile: occuparsi dell'istruzione e della salute, in ottica sviluppo umano è di fondamentale importanza, non c'è sviluppo umano che possa prescinderne indipendentemente dai livelli di ricchezza raggiunti.

Tableau

Descrizione del dataset

Il dataset utilizzato raccoglie informazioni relative al livello di inquinamento dell'aria rilevato da analisi effettuate in circa 14000 città nel mondo.

Sono presenti, nello specifico, oltre al livello di inquinamento complessivo, anche quello causato dai principali inquinanti conosciuti (*PM 2.5, ozono, diossido di nitrogeno e monossido di carbonio*).

Fonte: World Air Quality Index by City and Coordinates | Kaggle

Approfondimento delle variabili:

Country	*Nazione di appartenenza della Città*
City	*Città della rilevazione*
AQI Value	*Valore dell' "AIR QUALITY INDEX" (per città)*
AQI Category	*Categoria di AQI (good, moderate, unealthy, unealthy for sensitive groups, very unealthy, hazardous)*
CO AQI Value	*Valore di CO (Carbon Oxide) (per città)*
CO AQI Category	*Categoria di CO (good, moderate, unealthy, unealthy for sensitive groups, very unealthy, hazardous)*
Ozone AQI Value	*Valore di Ozono (per città)*
Ozone AQI Category	*Categoria di Ozono(good, moderate, unealthy, unealthy for sensitive groups, very unealthy, hazardous)*
NO2 AQI Value	*Valore di NO2 (Nitrogen Dioxide) (per città)*
NO2 AQI Category	*Categoria di NO2 (good, moderate, unealthy, unealthy for sensitive groups, very unealthy, hazardous)*
PM2.5 AQI Value	*Valore di PM2.5 (polveri sottili) (per città)*
PM2.5 AQI Category	*Categoria di PM2.5(good, moderate, unealthy, unealthy for sensitive groups, very unealthy, hazardous)*
lat	*Latitudine della città della rilevazione*
lng	*Longitudine della città della rilevazione*

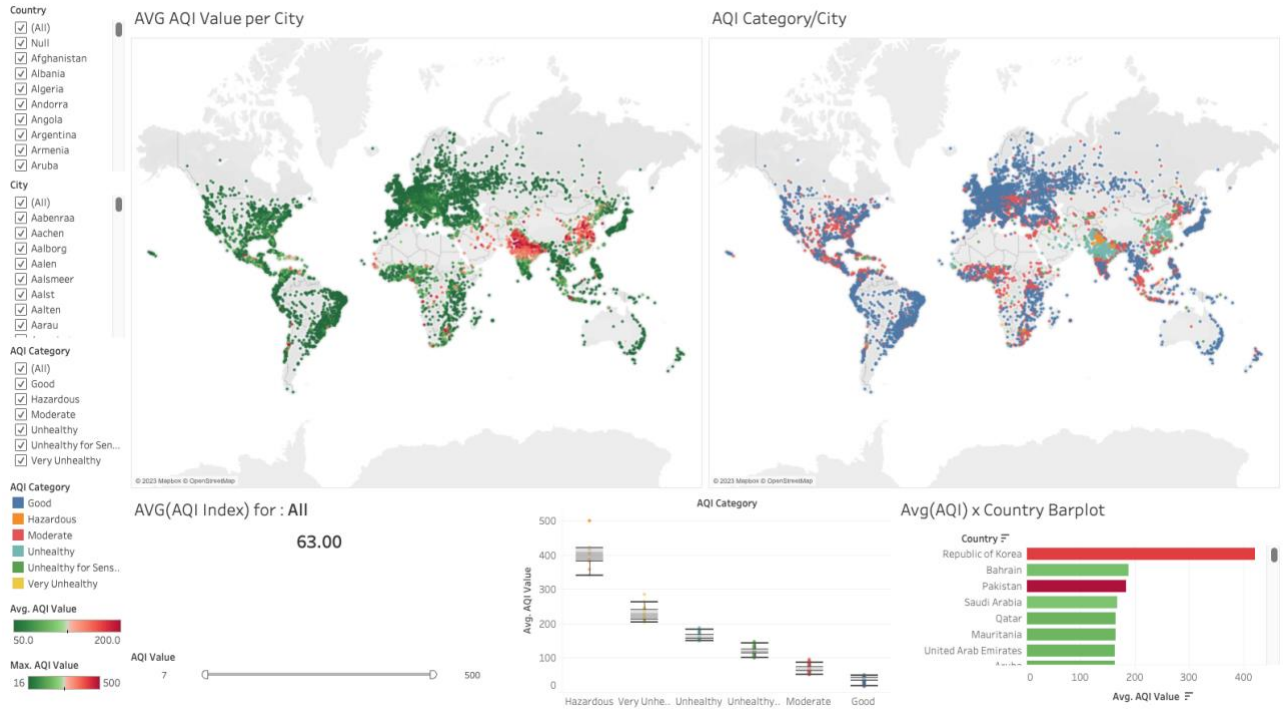
Obiettivo dell'analisi

Obiettivo dell'analisi è quello di visualizzare la situazione mondiale a livello di qualità dell'aria (con un focus particolare sulle differenze territoriali), caratterizzarne le determinanti principali ed eventuali relazioni rilevanti, così da avere una sorta di cartina tornasole ed una quanto più fedele istantanea della situazione globale.

Si è scelto di articolare la visualizzazione in 4 dashboard distinte, ognuna delle quali dedicata ad uno specifico approfondimento analitico.

Descrizioni delle dashboard:

Mappa della qualità dell'aria



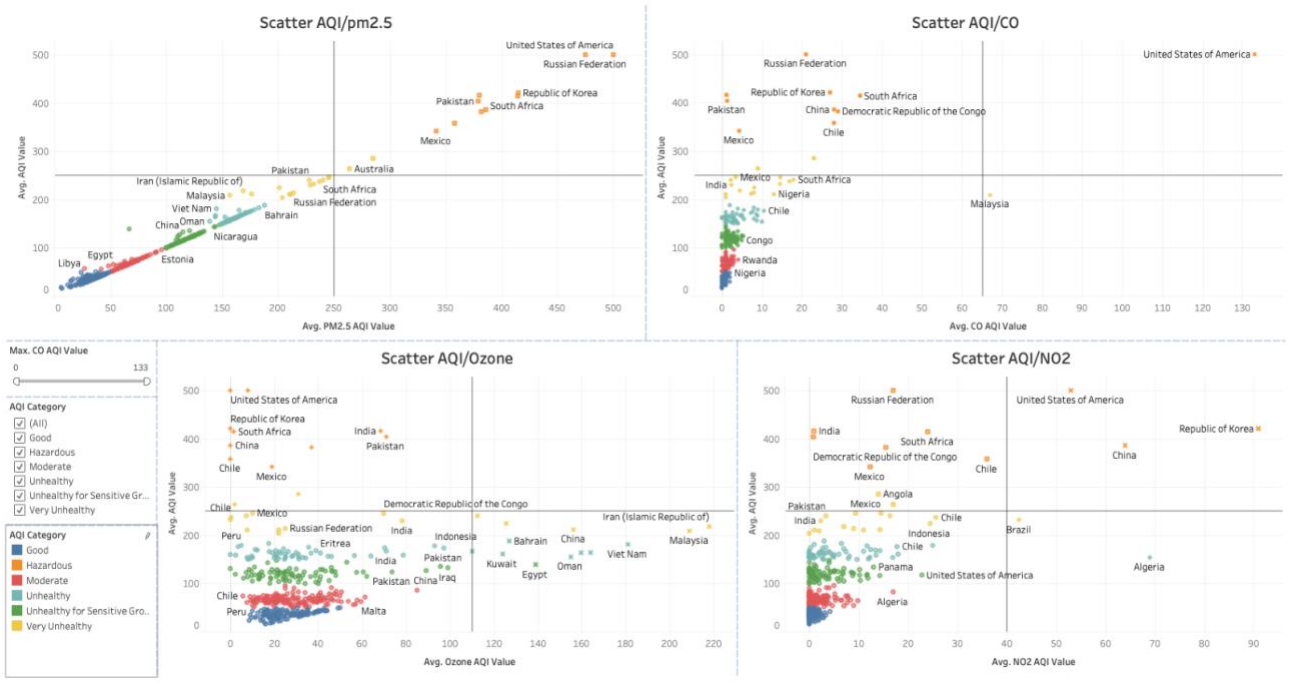
Nella prima dashboard si è scelto di ottenere una visione generale ed immediata della situazione attraverso lo strumento della mappa.

Nello specifico ne sono state articolate 2 distinte: una per livello di inquinamento medio ed un'altra per

evidenziare, a livello geografico, la categoria delle città per quanto riguarda, ancora una volta, il livello di inquinamento (*good, moderate, unhealthy for sensitive group, unhealthy, very unhealthy e hazardous*).

Attraverso una card interattiva, visualizziamo il valore specifico dell'inquinamento complessivo, presente in una singola unità geografica (***AQI Value***). Tramite un *boxplot* ed un *barplot*, possiamo renderci conto, rispettivamente, della distribuzione dei valori all'interno di ogni gruppo e del livello di inquinamento dei singoli stati, con la possibilità di ottenere un confronto semplice ed immediato.

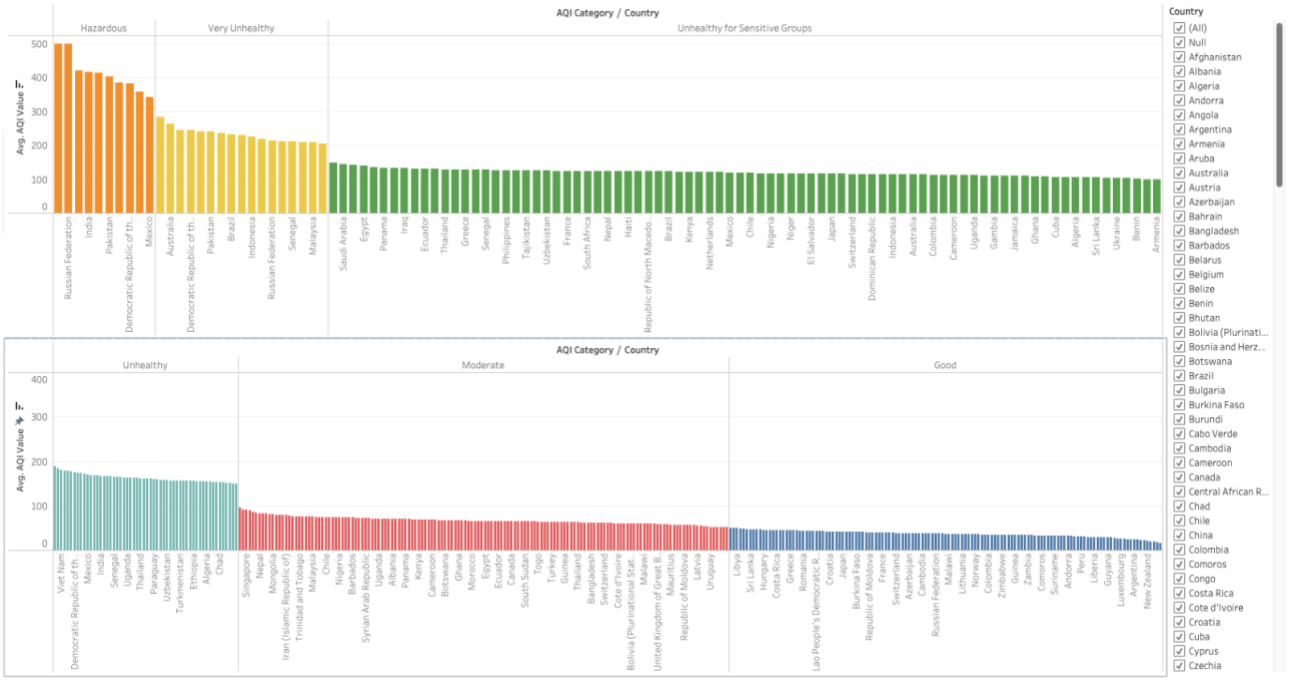
Inquinanti



Nella seconda Dashboard il focus si è spostato sui singoli inquinanti, al fine di coglierne la relazione con il livello di inquinamento complessivo.

La visualizzazione è stata esplicitata attraverso lo strumento dello *Scatterplot* con quadranti, attraverso il quale è possibile cogliere immediatamente le tendenze tra le due variabili chiamate in causa. Nello specifico l'unico inquinante che è in relazione lineare con l'**AQI** complessivo e il **PM2.5**; **monossido di carbonio (CO)** e **diossido di azoto (NO2)** hanno effetti estremamente nefasti (ad un loro minimo aumento corrisponde un aumento enorme dell'AQI complessivo) mentre per l'**Ozono** non sembra rilevabile una correlazione netta ed evidente.

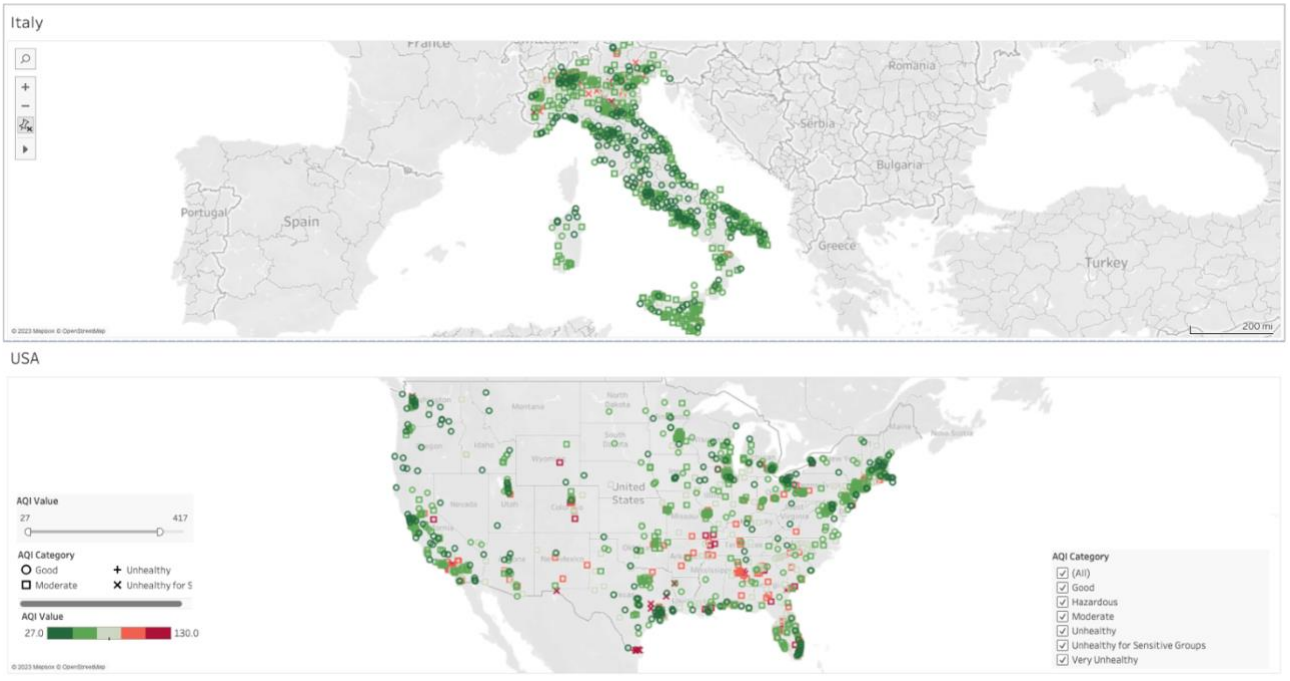
Dimensione dell'inquinamento



Nella terza Dashboard si è voluto porre l'accento sulla dimensione dell'inquinamento, attraverso l'utilizzo di un *barplot*, all'interno del quale ogni colore corrisponde ad un livello di inquinamento medio complessivo (relativo alla categoria **AQI Value**).

La presenza di un filtro interattivo, ci permette di navigare la dimensione del livello di inquinamento da un punto di vista geografico, con immediatezza e consapevolezza.

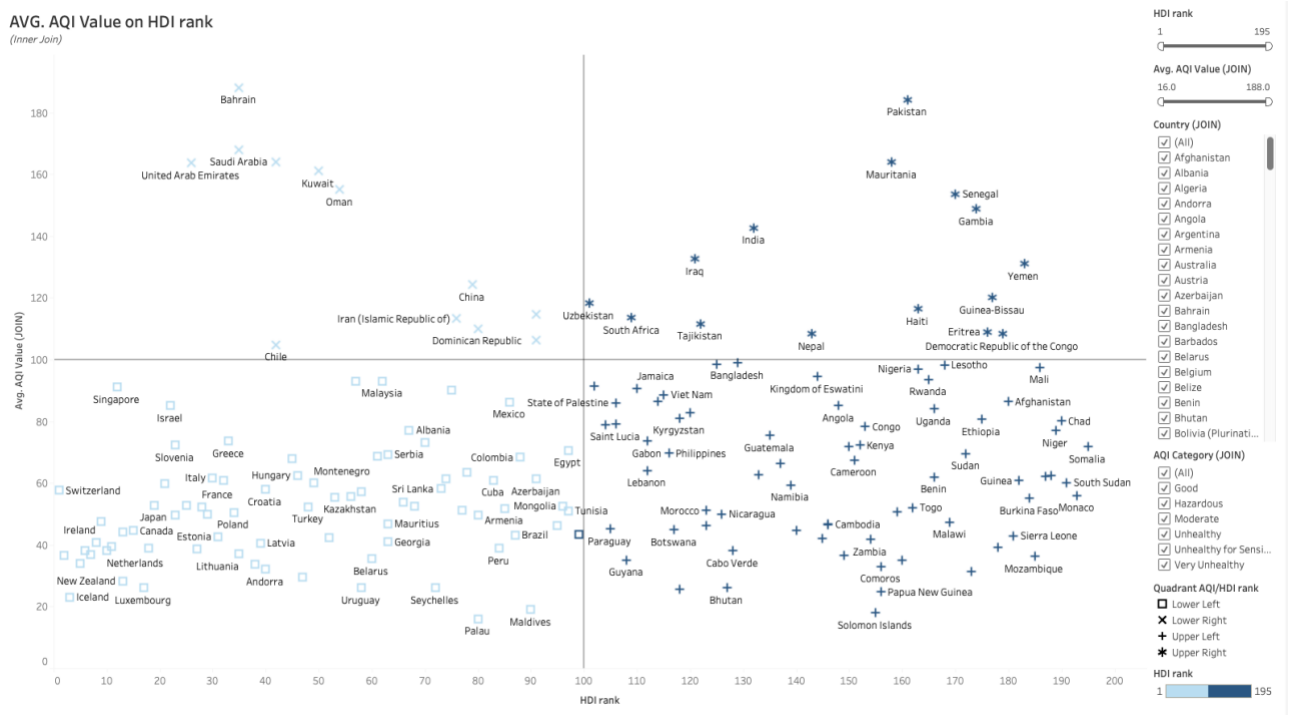
Confronto tra Paesi



Una Dashboard ulteriore potrebbe essere adattata all'occorrenza, per confrontare 2 o più paesi contemporaneamente, tenendo conto delle metriche scelte.

Nel caso specifico della Dashboard di cui sopra, si è deciso di confrontare il livello di inquinamento delle città italiane con quello delle città americane, osservando in questo caso le categorie (***AQI Value***) alle quali le singole città appartengono.

Inquinamento e Sviluppo Umano: una visione di insieme



Per concludere, sulla base delle precedenti analisi, è stato elaborato uno *Scatterplot* con quadranti al fine di visualizzare l'eventuale relazione esistente tra lo sviluppo umano e l'inquinamento dell'aria; tematica di grande attualità ed estrema rilevanza pubblica.

Per raggiungere tale obiettivo è stato effettuato un Inner Join tra il dataset della Qualità dell'aria e quello dello Sviluppo Umano