Laboratory of Data Science

Erica Cau, Andrea Failla, Federico Mazzoni

A.A. 2021-22

1 Parte 1

1.1 Task 0: Modellazione della base di dati

Partendo dallo schema fornito, è stato creata la base di dati contenente al suo interno cinque diverse tabelle: Match, Tournement, Player, Date e Geography, come riportato nella figura seguente.

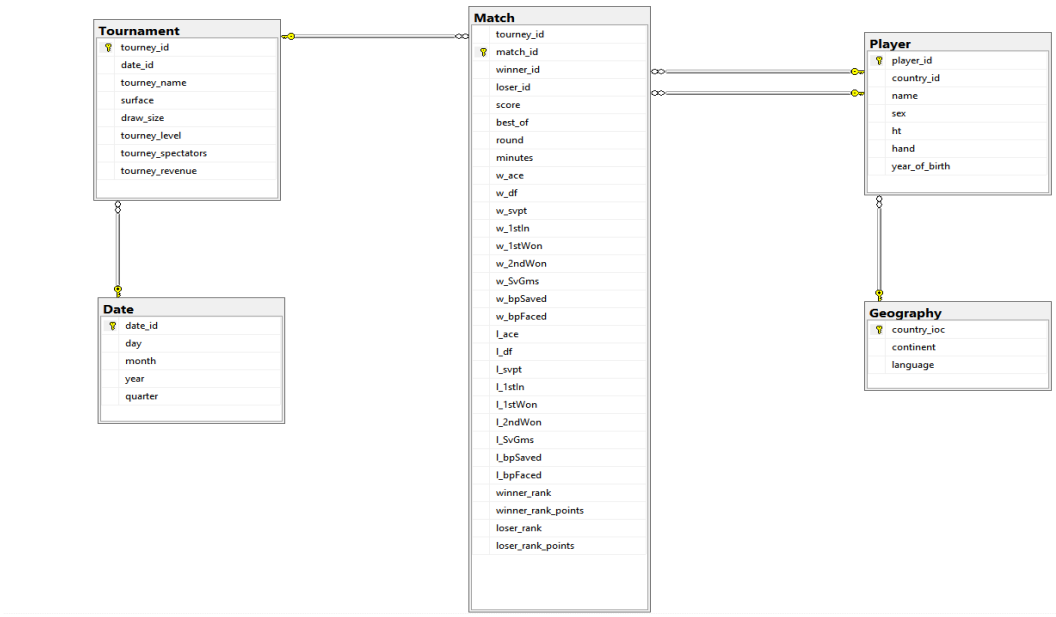

Figura 1: Schema del database

Per la natura stessa dei dati sono stati scelti i datatype così come riportati in tabella.

Tabella	Colonne
Date	date_id (int), day (int), month (int), year (int), quarter (int)
Geography	country_ioc (nchar), continent (varchar), language (varchar)
Tournament	tourney_id (varchar), date_id (int), tourney_name (varchar), surface (varchar), draw_size (int),

	tourney_level (varchar), tourney_spectators (int), tourney_revenue (float)
Player	player_id (int), country_id (nchar), name (varchar), sex (nchar), ht (nchar), hand (nchar), year_of_birth (nchar)
Match	tourney_id (varchar), match_id (varchar), winner_id (int), loser_id (int), score (varchar), best_of (int), round (varchar), minutes (int), w_ace (int), w_df (int), w_svpt (int), w_1stIn (int), w_1stWon (int), w_2ndWon (int), w_SvGms (int), w_bpSaved (int), w_bpFaced (int), l_ace (int), l_df (int), l_svpt (int), l_1stIn (int), 1_1stWon (int), l_SvGms (int), l_bpSaved (int), l_bpSaved (int), l_bpFaced (int), winner_rank (int), winner_rank_points (int), loser_rank (int), loser_rank_points (int)

Tabella 1: Attributi delle tabelle nella base di dati

Per ciascuna tabella è stata scelta una chiave primaria, che ne identificasse univocamente ogni singola istanza: match_id - ottenuto dall'unione delle stringhe di tourney_id e match_num (entrambe feature presenti in *tennis.csv*) - per match, tourney_id per tourney, country_ioc per Geography, player_id per Player (ottenuto dai valori unici di winner_id e loser_id) e date_id per date, assegnato arbitrariamente in modo incrementale partendo da 1.

Successivamente, sfruttando le rispettive chiavi primarie, sono state definite delle relazioni fra le varie tabelle. Si noti che nell'ultimo caso è stato necessario relazionare la primary key con due foreign key.

- PK: Date.date_id; FK: Tournament.date_id;
- PK: Tournament.tourney_id; FK: Match.tourney_id;
- PK: Geography.country_ioc; FK: Player.country_id;
- PK: Player.player.id; FK: Match.winner_id, Match.loser_id.

1.2 Task 1: Data Understanding

L'analisi è partita dalla *data understanding* di quattro diversi dataset, originariamente in formato .csv:

- *tennis.csv*, contenente informazioni su vari incontri di tennis;
- *male_player.csv* e *female_player.csv*, contenente nomi e cognomi di tutti i giocatori del rispettivo sesso;
- *country_codes.csv*, contenente i codici IOC di vari Paesi e informazioni collaterali (fra cui lingua e continente).

L'ultimo dataset, resosi necessario per ottenere alcune informazioni per popolare la tabella geography, è stato scaricato dal link che segue: https://github.com/datasets/country-codes/blob/master/data/country-codes.csv

Il dataset principale, tennis.csv, conteneva informazioni anche sui giocatori (come ad esempio ace o hand), identificati come winner o loser di un particolare incontro 1vs1.

Per effettuare l'operazione di popolamento della base di dati a partire da tennis.csv, sono stati realizzati i seguenti dataset:

- *match.csv*, contenente esattamente informazioni già incluse nell'originario *tennis.csv;*
- *player.csv,* contenente le informazioni sui singoli giocatori, siano essi *winner* o *loser* nel dataset originario. Il sesso è stato ricavato dal dataset *male_player.csv,* postulando che i giocatori non inclusi in tale dataset fossero di sesso femminile;
- *geography.csv,* contenente tutti i codice *IOC* dei paesi dei giocatori, le rispettive lingue e il rispettivo continente, ricavati dal sopracitato *country_code.csv.* 17 codici IOC dei giocatori, non presenti nel dataset *country_codes.csv,* non presentano informazioni relative a lingua e continente;
- *tournement.csv,* contenente informazioni sui tornei ricavate direttamente da *tennis.csv;*

- *date.csv,* contenente giorno, mese, anno e trimestre dei tornei di *tournement.csv* e un codice univoco identificativo.

Per quanto concerne il popolamento della base di dati, è stato fatto uso della librerie *pyodbc*. Tutti i dati sono stati in un primo momento letti come *string*, presentando quindi problemi durante l'operazione di popolamento. Si è dunque reso necessario un cast nell'adeguato datatype, in particolare per gli *int* e tourney_revenue, unico caso di *float*.

2 Parte 2

2.1 Task 0: For every tournament, the players ordered by number of matches won

Per svolgere questo *task*, è stato effettuato l'accesso alle tabelle di *match* e *player* utilizzando le credenziali assegnate al gruppo. Successivamente, si è calcolato il numero di incontri vinti da ciascun giocatore per ogni torneo tramite un nodo di aggregazione. Infine, dopo aver effettuato una *star join* tra le tabelle, i dati sono stati ordinati in base al *tourney_id* e al numero di incontri vinti. L'output, infine, è stato salvato su un file di testo.

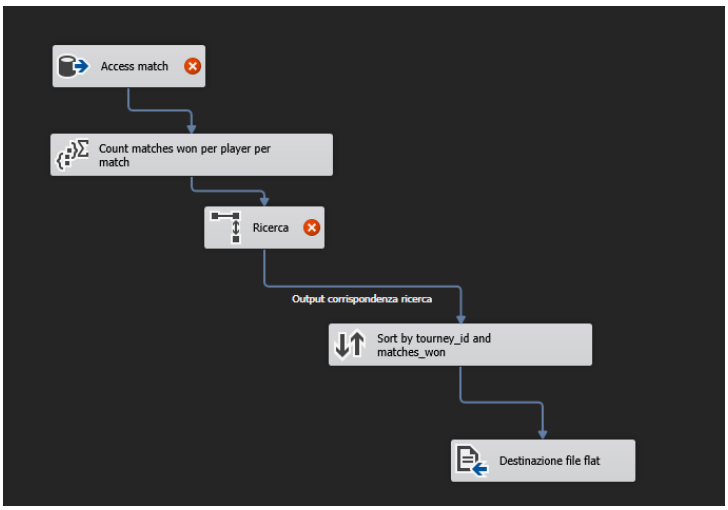

Figura 2: Soluzione SSIS per il *task 0*

2.2 Task 1: A tournament is said to be ”worldwide” if no more than 30 percent of the participants come from the same continent. List all the worldwide tournaments

Il secondo *task* ha richiesto operazioni nettamente più sofisticate. In primo luogo, si è acceduto alla tabella *geography*, per recuperare i dati sui *country_id* e integrato questi con i *player_id* corrispondenti mediante una *star join*. Tramite il nodo *Unpivot* sono stati aggregati i dati dei vincitori e dei perdenti in un'unica colonna dopo aver effettuato l'accesso alla tabella *Match*. Ciò è servito per calcolare il numero di giocatori in ciascun torneo per ciascun continente, così come il numero totale di giocatori unici per ciascun torneo.

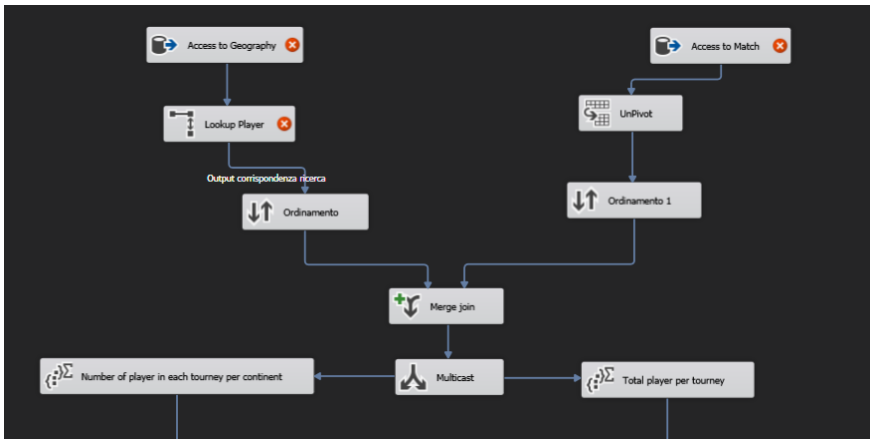

Figura 3: Soluzione SSIS per il *task 1* (prima parte)

Infine, è stata creata una colonna derivata per calcolare il valore in percentuale dei partecipanti provenienti dallo stesso continente (chiamata is_worldwide). Si è infine usato un nodo di Suddivisione condizionale per filtrare tra tutti i tornei solo quelli che avessero valori di *is_wordwide* minori o uguali a 30, che sono stati poi salvati su file di testo.

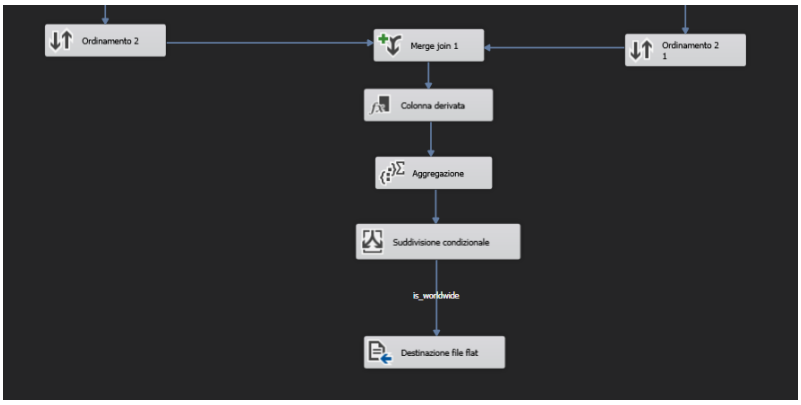

Figura 4: Soluzione SSIS per il *task 1* (seconda parte)

2.3 Task 2: For each country, list all the players that won more matches than the average number of won matches for all players of the same country

Per l'ultimo *task*, si è acceduto alla tabella *match*, combinata con *player* mediante *star join*. Attraverso un'operazione di aggregazione (*group by*) sono stati contate le vittorie di ciascun giocatore. Ciò ha permesso, con un'altra aggregazione, di individuare il numero medio di vittore per Paese.

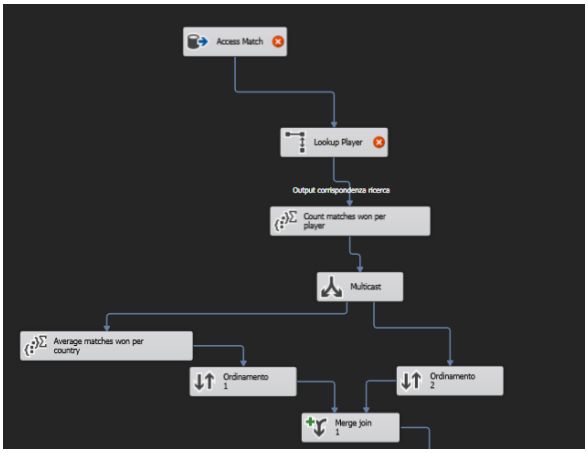

Figura 5: Soluzione SSIS per il *task 2* (prima parte)

In seguito, tramite suddivisione condizionale, sono stati mantenuti nel flusso solo i giocatori che hanno vinto più incontri rispetto alla media del proprio Paese. Infine, per facilitarne l'interpretazione, i dati sono stati ordinati per numero di vittorie e integrati con i nomi dei giocatori.

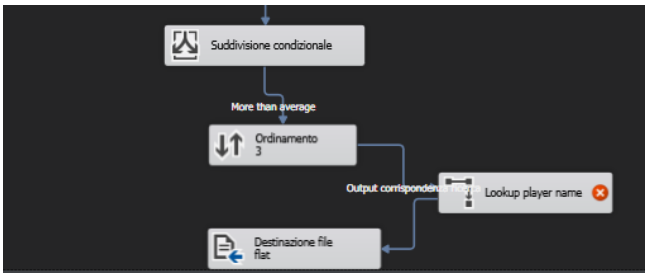

Figura 6: Soluzione SSIS per il *task 2*
(seconda parte)

3 Parte 3

Per la creazione del cubo, sono state selezionate come dimensioni le tabelle Match, Player e Tournament, utilizzando come chiavi primarie rispettivamente *match_id*, *player_id* e *tournament_id*. A *player_id* e *tournament_id* sono stati inoltre associati i rispettivi name come valori. Il nome assegnato è *524324_R2*.

Figura 7: Dimensioni *data cube*

Si è scelto di non selezionare Date e Geography come dimensioni, inserendo i rispettivi attributi - e realizzando le appropriate gerarchie - rispettivamente all'interno delle dimensioni Player e Tournament.

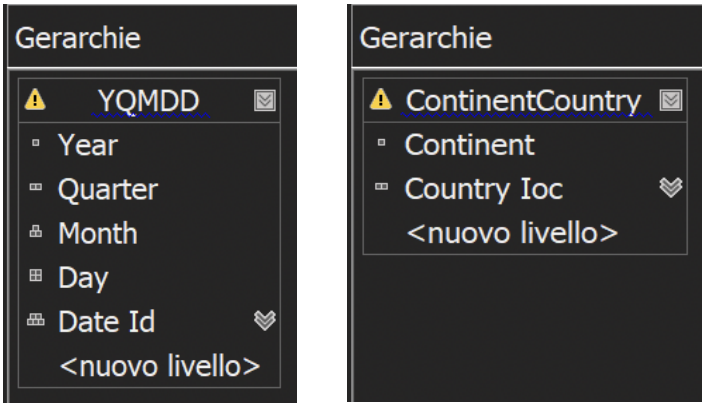

Figura 8: Gerarchie *data cube*

Come misure sono stati selezionati gli attributi della tabella Match: Visual Studio ha poi creato *Conteggio di Match*, *Conteggio di Player* e *Conteggio di Tournament*.

Si è ottenuto così un cubo avente le seguenti dimensioni:

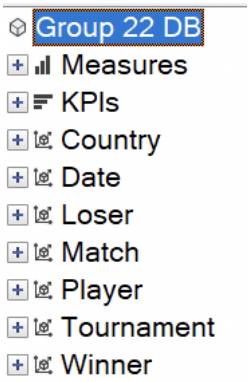

Figura 9: Dimensioni *data cube*

Si noti che le dimensioni *Country* e *Date* sono state inferite in automatico rispettivamente dalle tabelle *Player* e *Tournament*. Poiché *Player* aveva due chiavi primarie, la rispettiva dimensione è stata "divisa" in due dimensioni: *Winner* e *Loser*.

3.1 Query

Come già presentato nella prima sezione di questo elaborato, nel dataset erano presenti 19 *Country IOC* senza un rispettivo *Continent*. Per lo svolgimento della seconda e della terza query, si è quindi deciso di non visualizzarle tra i risultati. Poiché formalmente nel dataset appartenevano a un Continent con valore "" (un *blank value*, piuttosto che un *nan*), non è stato possibile utilizzare le funzioni di Visual Studio per nascondere i valori mancanti. Si è allora operato a livello di MDX.

3.1.1 Query 1

```
with MEMBER past as
```

```
([Tournament].[Year].currentmember.lag(1),
[Measures].[Winner Rank Points])

MEMBER curr as
[Measures].[Winner Rank Points]

MEMBER perc_incr as
iif(past = 0, "-", (curr - past)/past),

format_string = "percent"
select ([Tournament].[Year].[Year], perc_incr) on
columns,
NONEMPTY(([Winner].[Player Id].[Player Id])) on rows
from [Group 22 DB]
```

	2016	2017	2018	2019	2020	2021
	perc_incr	perc_incr	perc_incr	perc_incr	perc_incr	perc_incr
Aada Inna	-	-	-	-	-	-
Aalisha Alexis	-	-	-	-	-	-
Aaliya Ebrahim	-	-	-	-	-	-
Aaliyah Hohmann	-	-	-	-	-	-
Aalyka Ebrahim	-	-	-	-	-	-
Aanisha Rahul Shewate	-	-	-	-	-	-
Aanisha Rahul Shewate Aanisha Rahul Shewate	-	-	-	-	-	-
Aanu Ayegbusi	-	-	-	-	-	-
Aareyalee Amrutsinh Chavan	-	-	-	-	-	-
Aaro Pollanen	-	-	-	-	-	-
Aaron Addison	-	-100.00%	-	-	-	-

Figura 10: Risultato della prima query

3.1.2 Query 2

```
with MEMBER country_rank as
([Winner].[ContinentCountry].currentmember,
[Measures].[Winner Rank Points])

MEMBER continent_rank as
([Winner].[ContinentCountry].currentmember.parent,
[Measures].[Winner Rank Points])

MEMBER country_perc as
iif(country_rank = null, 0, country_rank /
continent_rank),
format_string = "percent"

select (country_perc) on columns,
filter(([Winner].[Continent].[Continent],
[Winner].[ContinentCountry].[Country Ioc]),
[Winner].[ContinentCountry].parent.MEMBERVALUE <> '')
on rows
from [Group 22 DB]
```

		country_perc
AF	ALG	0.32%
AF	BDI	0.53%
AF	BEN	0.00%
AF	BOT	0.00%
AF	CMR	0.00%
AF	COD	0.00%
AF	EGY	13.75%
AF	ERI	0.00%
AF	GAB	0.00%
AF	GHA	0.00%
AF	KEN	0.00%

Figura 11: Risultato della seconda query

3.1.3 Query 3

```
with member continent_rank as
([Loser].[ContinentCountry].currentmember.parent,
[Measures].[Loser Rank Points])

member ratio as
[Measures].[Loser Rank Points]/continent_rank,
format_string = 'percent'

set namedContinent AS
FILTER([Loser].[ContinentCountry].children,
[Loser].[ContinentCountry].currentmember.MEMBERVALUE <>
'')

select {[Measures].[Loser Rank Points], ratio,
continent_rank} on columns,
nonempty(filter(([Tournament].[YQMDD].[Year],
namedContinent,
[Loser].[Player Id].[Player Id]), ratio > 0.1)) on rows

from [Group 22 DB]
```

			Loser Rank Points	ratio	continent_rank
2016	AF	Kevin Anderson	18830	32.85%	57329
2016	AF	Malek Jaziri	14648	25.55%	57329
2016	OC	Daria Gavrilova	24328	10.68%	227829
2016	OC	Samantha Stosur	47260	20.74%	227829
2017	AF	Kevin Anderson	18910	27.54%	68652
2017	AF	Malek Jaziri	16633	24.23%	68652
2017	AF	Ons Jabeur	11187	16.30%	68652
2017	OC	Daria Gavrilova	35570	13.89%	256076
2017	OC	Samantha Stosur	32081	12.53%	256076
2018	AF	Kevin Anderson	69120	50.46%	136968
2018	AF	Malek Jaziri	25603	18.69%	136968

Figura 12: Risultato della terza query

3.1.4 Task 4

In questo penultimo *task* si è passati alla fase di visualizzazione dei dati contenuti all'interno del *data cube* e, in particolare, alla realizzazione di una dashboard che rappresentasse la distribuzione geografica dei *Winner Rank Points* e dei *Loser Rank Points*. Per lo svolgimento di tale lavoro è stato utilizzato il software Microsoft Power BI.

Prima di tutto è stato necessario accedere al cubo 524324_R2 per mezzo del comando *Recupera dati – Analysis Services.* Il passo seguente è stato procedere con la visualizzazione dei dati.

Nella dashboard, visibile in Figura 13, sono stati inclusi i seguenti grafici:

- Due grafici della tipologia *Indicatore KPI* per *Winner* e *Loser Rank Points;*
- Due grafici a torta che mostrassero, per ciascun continente, il numero totale di *Winner* e *Loser Rank Points*;
- Due mappe per mostrare la distribuzione dei due punteggi per ogni Paese e Continente;
- Un grafico ad area in pila per visualizzare i valori *Winner e Loser Rank Points* per ogni Paese;
- Un grafico a barre in pila per visualizzare i valori di *Winner e Loser Rank Points* per ogni continente.

Inoltre, sono state incluse due tabelle mostranti i valori usati per i diversi *plot*, includendo quindi gli attributi *Player Id, Continent, Country IOC, Winner Rank Point* e *Loser Rank Points.*

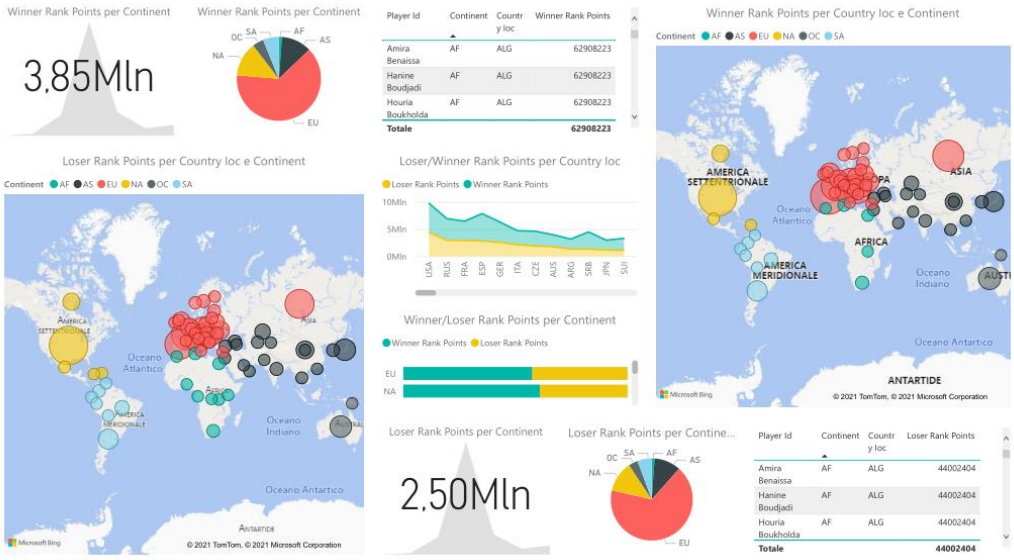

Figura 13: Dashboard per la visualizzazione della distribuzione geografica di *Winner/Loser Rank Points*

3.1.5 Task 5

Dato che in questo ultimo task è stata richiesta la realizzazione di una dashboard a discrezione del gruppo, si è deciso di inserire dei grafici che permettessero di analizzare i match e i giocatori.

Come nel caso del task precedente, i dati utilizzati sono quelli contenuti nel data cube.

In questo caso, si è fatto uso di:

- Una mappa che mostrasse la distribuzione dei match per Paese;
- Un istogramma a colonne raggruppate, con il quale si è potuta osservare la distribuzione dei match per ogni anno (2016-2021) e per ogni quadrimestre;
- Un grafico a linea in cui è incluso un istogramma, che mostrasse i diversi andamenti delle variabili *Tourney Revenue* e *Number of Spectators* in base all'anno di svolgimento dei tornei;

- Un grafico a barre orizzontale (realizzato con questo orientamento per favorire la visualizzazione) che mostrasse i *Winner Rank Points* per player;
- Tre indicatori KPI per numero di *Match*, *Player* e *Torneo*;
- Un grafico a barre per mostrare la distribuzione di *Winner Rank Points* per sesso dei giocatori;
- Un grafico ad anello per mostrare la distribuzione dei tornei per ciascun anno.

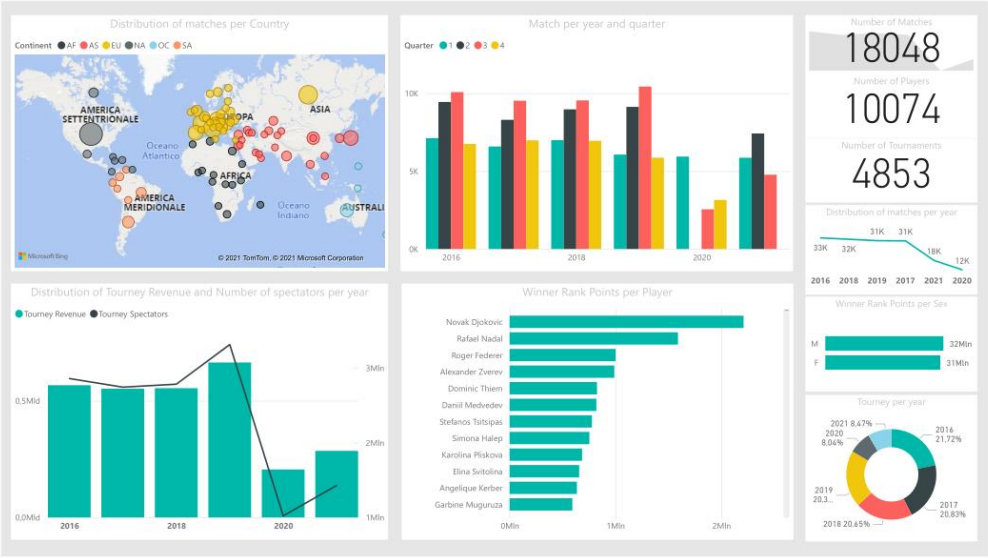

Figura 14: Dashboard per l'analisi dei match e dei player del dataset