NANCELLE Alexis

5ieme année

Valenciennes

Documentation technique

5VPO – Rattrapage

2023-2024

1. Collecte des données (5BDA - 4pts)

Pour la partie collecte des données, j'ai réalisé un script permettant déjà de faire un cleaning des data.

Vous le trouverez dans le github : https://github.com/XeilaS/5VPO

Il permet dans un premier temps de charger les dataframes comme ici :

```
df0 = pd.read_csv('tennis_dataset/atp_matches_2010.csv')
df1 = pd.read_csv('tennis_dataset/atp_matches_2011.csv')
df2 = pd.read_csv('tennis_dataset/atp_matches_2012.csv')
df3 = pd.read_csv('tennis_dataset/atp_matches_2013.csv')
df4 = pd.read_csv('tennis_dataset/atp_matches_2014.csv')
df5 = pd.read_csv('tennis_dataset/atp_matches_2015.csv')
df6 = pd.read_csv('tennis_dataset/atp_matches_2016.csv')
df7 = pd.read_csv('tennis_dataset/atp_matches_2017.csv')
df8 = pd.read_csv('tennis_dataset/atp_matches_2018.csv')
df9 = pd.read_csv('tennis_dataset/atp_matches_2019.csv')
df10 = pd.read_csv('tennis_dataset/atp_matches_2020.csv')
df11 = pd.read_csv('tennis_dataset/atp_matches_2021.csv')
df12 = pd.read_csv('tennis_dataset/atp_matches_2022.csv')
df13 = pd.read_csv('tennis_dataset/atp_matches_2023.csv')
df14 = pd.read_csv('tennis_dataset/atp_matches_2024.csv')
```

Et ensuite de les concaténer avec cette ligne :

```
# Concaténation des dataframes
newDf = pd.concat([df0,df1,df2,df3,df4,df5,df6,df7,df8,df9,df10,df11,df12,df13,df14])
```

Pour la suite que je vais éviter de mettre ici en entier, j'ai typé les données des colonnes pour que l'on soit avec les bons typages comme toutes les données de nombres entier.

```
newDf['winner_ht'] = newDf['winner_ht'].fillna(0)
newDf['winner_ht'] = newDf['winner_ht'].astype(int)

newDf['winner_age'] = newDf['winner_age'].fillna(0)
newDf['winner_age'] = newDf['winner_age'].astype(int)
```

Ensuite je vérifie les données en plus de leur type et ensuite je sauvegarde.

```
print(newDf.head(10))
print(newDf.dtypes)

# Sauvegarde du nouveau dataset
newDf.to_csv('atp_matches_2010_to_2024.csv', index=False)
```

Ensuite pour le data processing des données, j'ai donc fait un autre fichier python s'appelant 'data_processing.py' qui est et qui va pour grandement résumer crée un nouveau dataframe.

Ce dataframe contiendra les "id" des joueurs, leur nom/prénoms, (etc. pour les caractéristiques des joueurs), une statistique sur le pourcentage points marqué aux premiers services (FirstSvcMade/FirstSvcWon) avec les colonnes de totales (1stWon + 1stIn), une statistique de victoire/défaite avec le nombre de matchs perdu ou gagné, une statistique du nombre de matchs, de participation à des tournois mais aussi des rangs aux cours des années de tennis et le nombre de points par année et pour finir la métrique qui est pour détermine le meilleur joueur : le total de points carriere.

Pour pourrez retrouver le fichier python au même endroit que le fichier de cleaning.

2. Lister les meilleurs de tous les temps (5BDA - 8pts)

Tout comme les deux autres scripts python, le fichier est retrouvable dans le repo github.

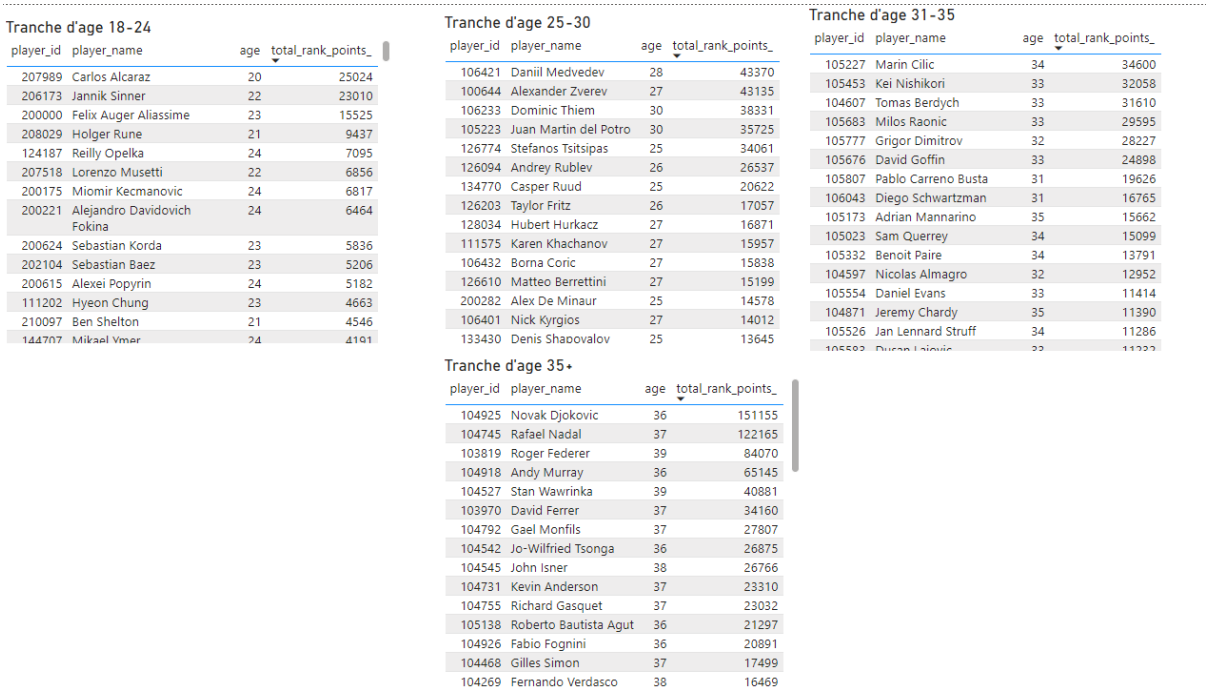
Comme dit précédemment, toutes les listes ici que je devais faire se base sur le nombre de points global et totale qu'un joueur a pu obtenir chaque année.

3. Data Viz (5PBI - 20pts)

Ici, je ne vais mettre que la consigne et un screen de la visualisation des données.

Puisque contrairement à l'exercice d'avant qui est à afficher en mode console, j'ai vraiment ici un visuel détaille.

- Une page qui liste les meilleurs joueurs par tranche d'âge : [18-24], [25-30], [3135],]35, +[⇒ 3 pts (ordonné dans l'ordre croissant)



Tranche d'age 18-24

player_id	player_name	age	total_rank_points_
207989	Carlos Alcaraz	20	25024
206173	Jannik Sinner	22	23010
200000	Felix Auger Aliassime	23	15525
208029	Holger Rune	21	9437
124187	Reilly Opelka	24	7095
207518	Lorenzo Musetti	22	6856
200175	Miomir Kecmanovic	24	6817
200221	Alejandro Davidovich Fokina	24	6464
200624	Sebastian Korda	23	5836
202104	Sebastian Baez	23	5206
200615	Alexei Popyrin	24	5182
111202	Hyeon Chung	23	4663
210097	Ben Shelton	21	4546
144707	Mikael Ymer	24	4191

Tranche d'age 25-30

player_id	player_name	age	total_rank_points_
106421	Daniil Medvedev	28	43370
100644	Alexander Zverev	27	43135
106233	Dominic Thiem	30	38331
105223	Juan Martin del Potro	30	35725
126774	Stefanos Tsitsipas	25	34061
126094	Andrey Rublev	26	26537
134770	Casper Ruud	25	20622
126203	Taylor Fritz	26	17057
128034	Hubert Hurkacz	27	16871
111575	Karen Khachanov	27	15957
106432	Borna Coric	27	15838
126610	Matteo Berrettini	27	15199
200282	Alex De Minaur	25	14578
106401	Nick Kyrgios	27	14012
133430	Denis Shapovalov	25	13645

Tranche d'age 31-35

player_id	player_name	age	total_rank_points_
105227	Marin Cilic	34	34600
105453	Kei Nishikori	33	32058
104607	Tomas Berdych	33	31610
105683	Milos Raonic	33	29595
105777	Grigor Dimitrov	32	28227
105676	David Goffin	33	24898
105807	Pablo Carreno Busta	31	19626
106043	Diego Schwartzman	31	16765
105173	Adrian Mannarino	35	15662
105023	Sam Querrey	34	15099
105332	Benoit Paire	34	13791
104597	Nicolas Almagro	32	12952
105554	Daniel Evans	33	11414
104871	Jeremy Chardy	35	11390
105526	Jan Lennard Struff	34	11286
105583	Dusan Lajovic	33	11232

Tranche d'age 35+

player_id	player_name	age	total_rank_points_
104925	Novak Djokovic	36	151155
104745	Rafael Nadal	37	122165
103819	Roger Federer	39	84070
104918	Andy Murray	36	65145
104527	Stan Wawrinka	39	40881
103970	David Ferrer	37	34160
104792	Gael Monfils	37	27807
104542	Jo-Wilfried Tsonga	36	26875
104545	John Isner	38	26766
104731	Kevin Anderson	37	23310
104755	Richard Gasquet	37	23032
105138	Roberto Bautista Agut	36	21297
104926	Fabio Fognini	36	20891
104468	Gilles Simon	37	17499
104269	Fernando Verdasco	38	16469

- Une page représentant les nationalités les plus représentées dans le tennis (top 25), ⇒ 3 pts (ordonné dans l'ordre croissant)

Nationalité les plus réprésenté

ioc	Nombre de joueurs réprésenté par Nationalité
USA	68
RUS	23
SUI	16
SWE	15
SRB	14
SLO	13
TPE	13
RSA	12
SVK	12
THA	12
UKR	11
URU	11
ROU	9
VEN	9
UZB	8
PUR	7
TUR	6
SRI	4
TUN	4
VIE	4
SYR	3
ZIM	3
TOG	2
SOL	1
UNK	1

- Une page représentant confrontations les plus longues (top 10), ⇒ 3 pts (Ordonné dans l'ordre croissant)

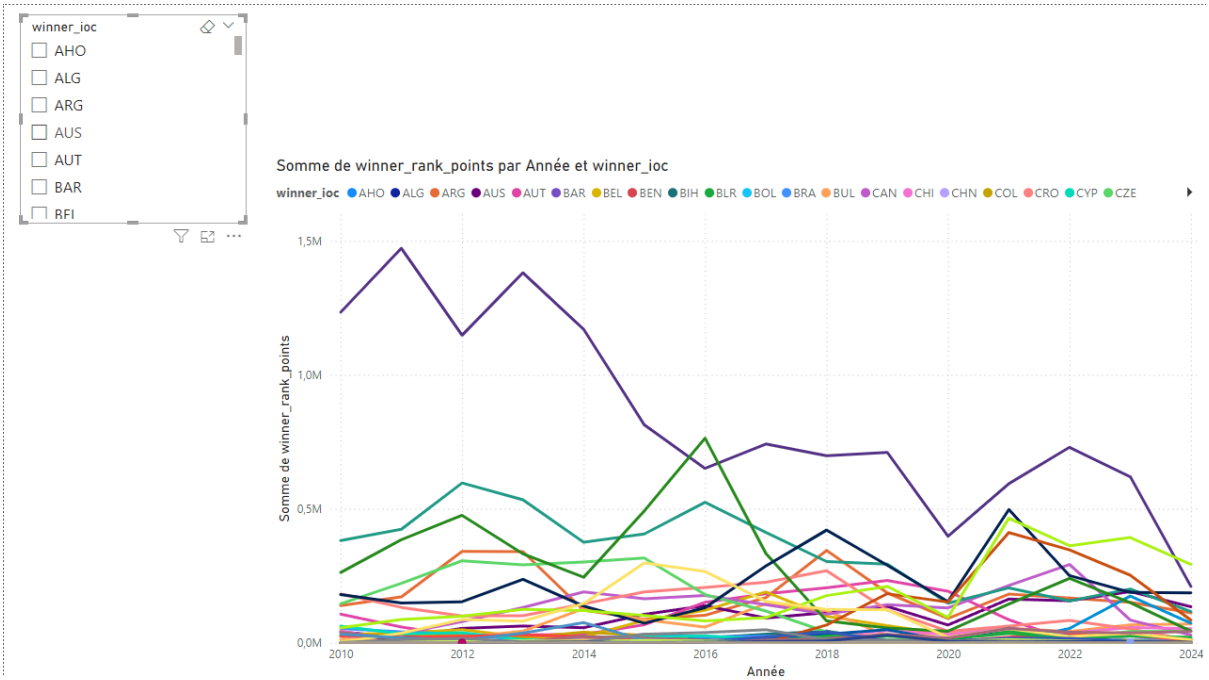
minutes	winner_id	winner_ioc	winner_name	loser_id	loser_ioc	loser_name
1146	104180	LUX	Gilles Muller	104871	FRA	Jeremy Chardy
987	111202	KOR	Hyeon Chung	105373	SVK	Martin Klizan
665	104545	USA	John Isner	103917	FRA	Nicolas Mahut
396	104731	RSA	Kevin Anderson	104545	USA	John Isner
365	105841	ITA	Lorenzo Giustino	144895	FRA	Corentin Moutet
356	104797	UZB	Denis Istomin	103529	PAK	Aisam Ul Haq Qureshi
353	104925	SRB	Novak Djokovic	104745	ESP	Rafael Nadal
345	104918	GBR	Andy Murray	106423	AUS	Thanasi Kokkinakis
344	106175	VEN	Ricardo Rodriguez	122082	PAR	Francisco Yim Kim
341	103908	FRA	Paul Henri Mathieu	104545	USA	John Isner

- Une page reprenant les statistiques joueurs par origine (NB de compet, NB de victoire, NB d'echec, Total de points, Meilleur rang, ...) ⇒ 4pts
  (ordonné sur les nationalités si origine = nationalité)

Id du joueurs	Noms joueurs	age	ioc	Nombre de compétitions	Nombres de matchs	Matchs gagnés	Matchs perdus	Ratio Victoire défaite	Points de rang carriere	rank_2024	rank_2023	rank_2022	rank_2021	rank_2020	rank_2019	rank_2018	r
103630	Martijn Van Haasteren	29	AHO	1	4	1	3	0.33	0	99999	99999	99999	99999	99999	99999	99999	99999
104996	Alexander Blom	23	AHO	2	7	4	3	1.33	1	99999	99999	99999	99999	99999	99999	99999	99999
104467	Lamine Ouahab	35	ALG	4	21	11	10	1.1	433	99999	99999	99999	99999	664	440	617	99999
103428	Juan Ignacio Chela	32	ARG	2	123	63	60	1.05	3765	99999	99999	99999	99999	99999	99999	99999	99999
103675	Diego Junqueira	31	ARG	1	15	4	11	0.36	1015	99999	99999	99999	99999	99999	99999	99999	99999
103900	David Nalbandian	31	ARG	5	123	77	46	1.67	3815	99999	99999	99999	99999	99999	99999	99999	99999
103976	Juan Pablo Brzezicki	28	ARG	0	2	1	1	1.0	275	99999	99999	99999	99999	99999	99999	99999	99999
104076	Jose Acasuso	28	ARG	0	7	3	4	0.75	1064	99999	99999	99999	99999	99999	99999	99999	99999
104122	Carlos Berlocq	35	ARG	8	262	117	145	0.81	5726	99999	99999	99999	99999	99999	99999	131	96
104216	Maximo Gonzalez	33	ARG	1	53	13	40	0.32	2315	99999	99999	99999	99999	99999	99999	99999	163
104314	Brian Dabul	26	ARG	1	19	7	12	0.58	1042	99999	99999	99999	99999	99999	99999	99999	99999
104338	Juan Monaco	32	ARG	9	333	186	147	1.27	9195	99999	99999	99999	99999	99999	99999	99999	99999
104547	Horacio Zeballos	33	ARG	2	214	86	128	0.67	5304	99999	99999	99999	99999	99999	99999	152	58
104651	Martin Alund	28	ARG	1	17	7	10	0.7	916	99999	99999	99999	99999	99999	99999	99999	99999
104724	Eduardo Schwank	26	ARG	2	57	22	35	0.63	1558	99999	99999	99999	99999	99999	99999	99999	99999
104919	Leonardo Mayer	32	ARG	9	340	161	179	0.9	8244	99999	99999	99999	99999	99999	60	41	52
105223	Juan Martin del Potro	30	ARG	8	412	304	108	2.81	35725	99999	99999	99999	99999	99999	12	4	17
105477	Marco Trungelliti	34	ARG	1	36	15	21	0.71	2533	197	99999	99999	198	99999	121	188	162
105487	Facundo Bagnis	34	ARG	4	131	44	87	0.51	5671	136	136	113	80	134	152	177	113
105550	Guido Pella	33	ARG	13	281	129	152	0.85	7856	99999	228	99999	77	37	25	58	72
105643	Federico Delbonis	32	ARG	10	364	165	199	0.83	9267	99999	131	125	46	78	67	87	81
105819	Guido Andreozzi	31	ARG	0	39	12	27	0.44	1896	99999	240	99999	99999	99999	113	118	142
105948	Federico Coria	32	ARG	1	121	53	68	0.78	3746	83	108	78	68	99	99999	335	99999
105952	Renzo Olivo	27	ARG	2	56	23	33	0.7	1582	99999	99999	99999	99999	297	99999	99999	118
106032	Facundo Arguello	22	ARG	2	17	1	16	0.06	363	99999	99999	99999	99999	99999	99999	99999	99999
106043	Diego Schwartzman	31	ARG	15	475	250	225	1.11	16765	99999	113	16	15	9	14	19	25
106044	Nicolas Kicker	25	ARG	2	48	19	29	0.66	1595	99999	99999	99999	99999	99999	99999	93	96
106228	Juan Ignacio Londero	28	ARG	3	85	35	50	0.7	2286	99999	99999	139	99999	69	56	99999	99999
106398	Pedro Cachin	29	ARG	2	70	25	45	0.56	2637	91	54	57	99999	354	280	99999	99999

Je rappelle ici que les valeurs avec 999999 sont les valeurs qui étaient vides et donc j'ai décidé de mettre la valeur la plus hautes pour éviter tout problème dans les classements par la suite.

Une page représentant les évolutions de points par année par nationalité (graphique).
⇒ 4pts



Ici, j'ai réussi à mettre "un segment" pour pouvoir sélectionner quel pays l'on veut voire. (Il faut CTRL + Clic gauche pour en sélectionner plusieurs)