Project presentation

Data mining presentation Group 14

Dalla Noce Niko, Lombardi Giuseppe, Ristori Alessandro

17 dicembre 2021

Main sections

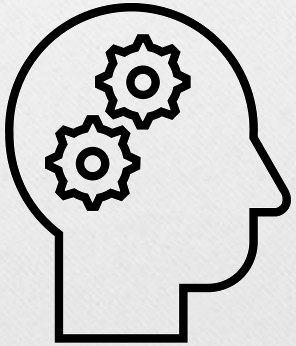
1) Data understanding and preparation

2) Clustering

3) Predictive analysis

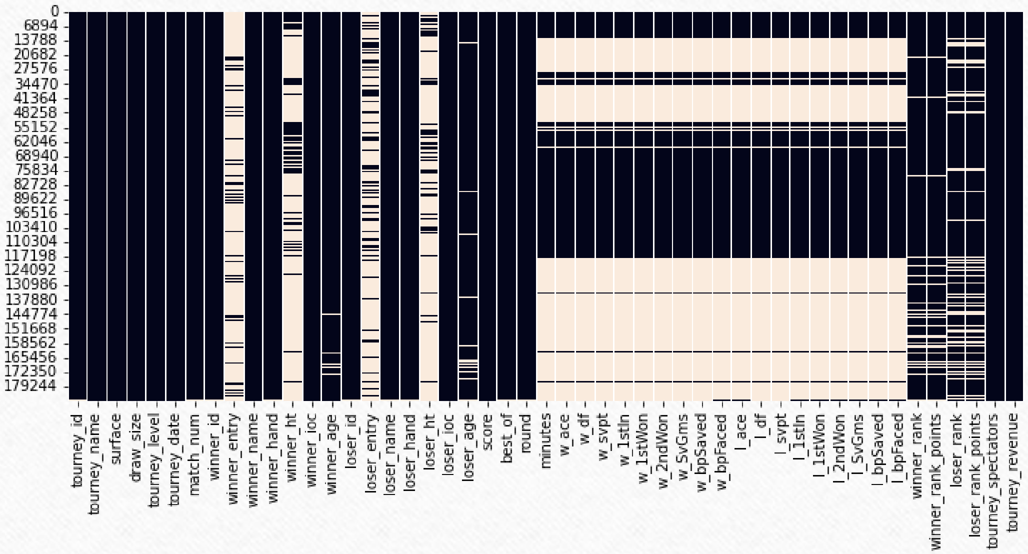
4) Time series

Data Understanding



Data understanding:
data analysis and integration

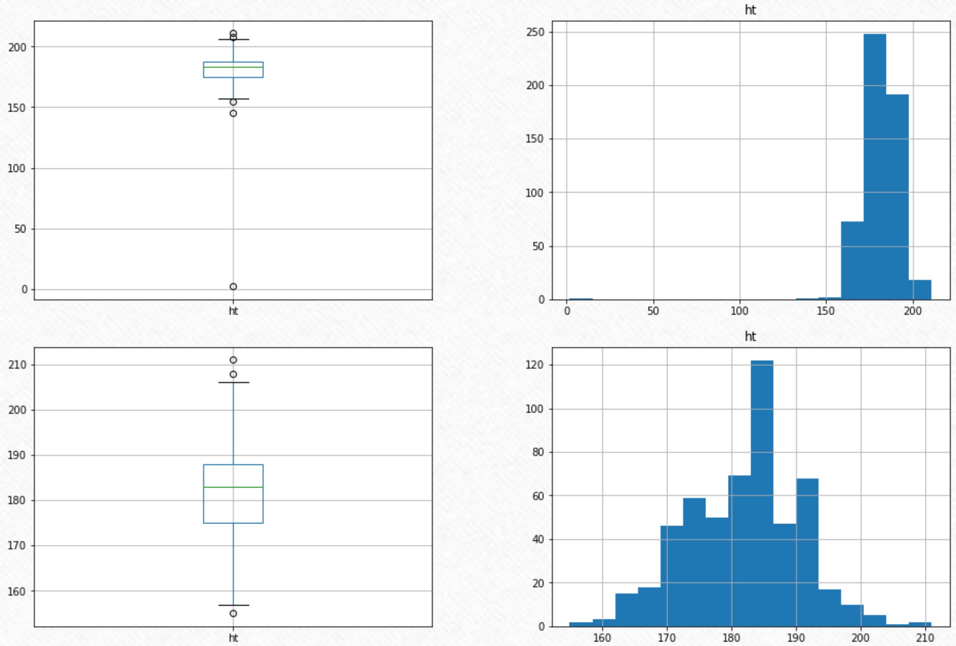
- **Understanding** the **meaning** of each **feature** of the original dataset of the **tennis matches**
- **Dropping irrelevant matches** (without statistics) and **useless features** for our purpose as the tournay name, the draw size, the winnet entry…
- **Dropping** 302 **duplicated records**
- **Integrating** data by **filling** the **missing values** and **fixing** issues with the **ambiguous values**
  - Fix association player id / name
  - Retrieve some missing values (when it is possible)
  - Assign special character to some missing values



The missing values heatmap

Data understanding:
outlier detection

- All the feature are **Gaussian** or **half normal**
- We computed the **median M**, the **first quartile Q1**, the **third quartile Q3** and the **interquartile IQR** for all variables
- We computed the **upper bound U = Q3 + 1.5 * IQR** and the **lower bound L = Q1 – 1.5 * IQR**
- **Outliers** are **≥ L** for the **non-negative variables** or **out** of the **range (L, U)** for the **other variables**
- **Outliers** are **substituted** by **M, L** or **U** based on the **semantics** of the feature
- **Some outliers** managed by heuristic approach based on **external knowledge**



An example of a feature's distribution after dealing with outliers

Data Preparation

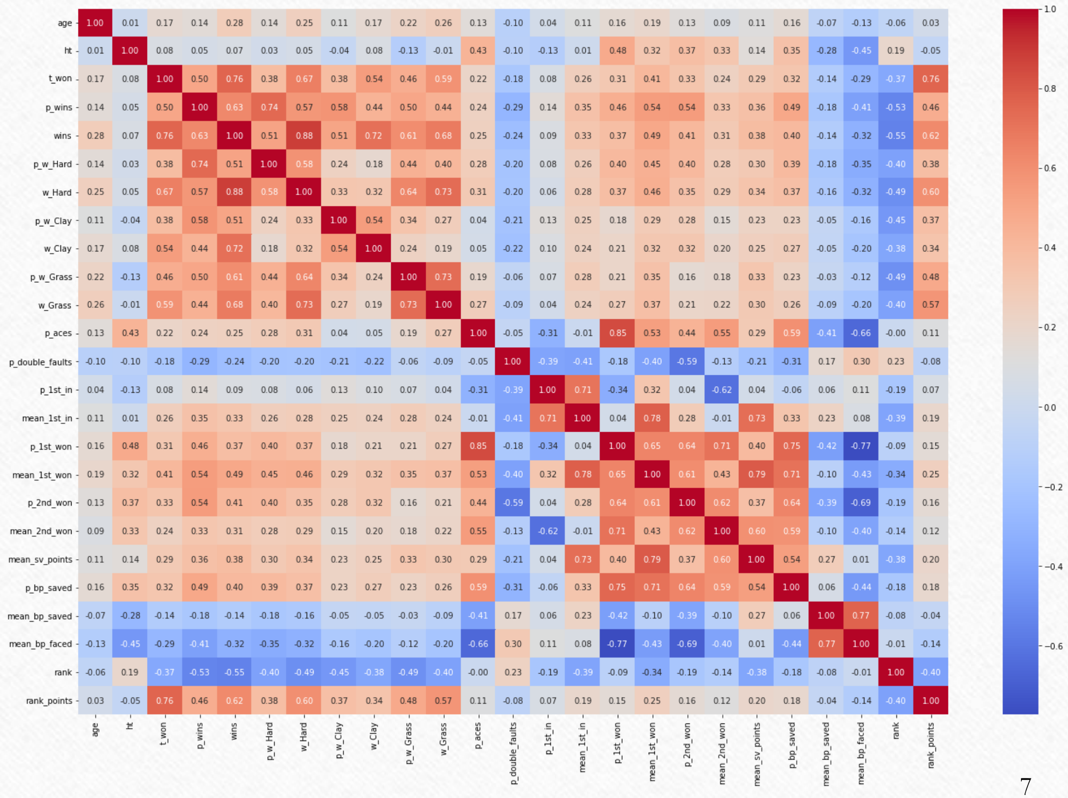
Data preparation:
building the player's profile

- From the matches dataset we **build the players' profile**, using several strategies to avoid the not retrievable missing data
- The players' attributes chosen are **representative** and **non-redundant**

Players' profile	
Categorical	**Numerical**
Sex	Wins and Losses
Age	Tournaments won
Ioc	Surfaces
Height	Statistics
Hand	Rank and rank points

- We created new categorical features that **split** the dataset by age, height and rank **range**, to better analyze the future results

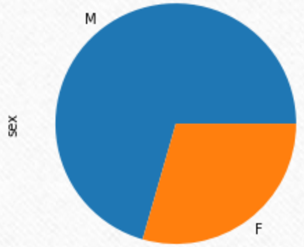
Correlation matrix of the new dataset with the features we added

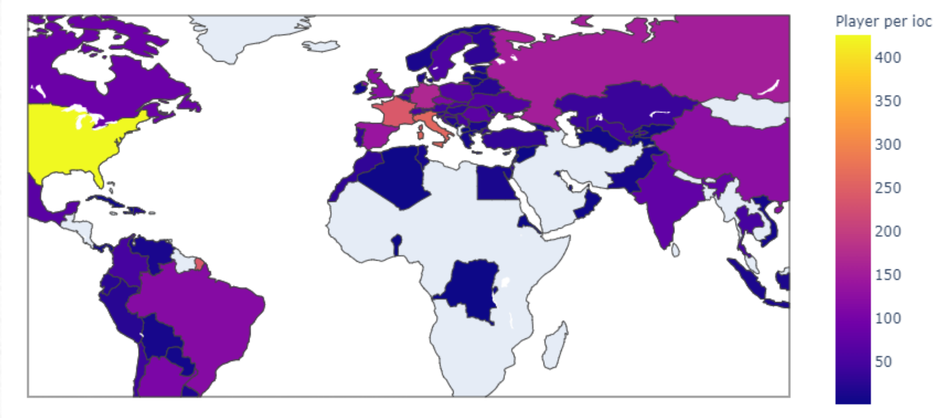


Data preparation:
player's profile

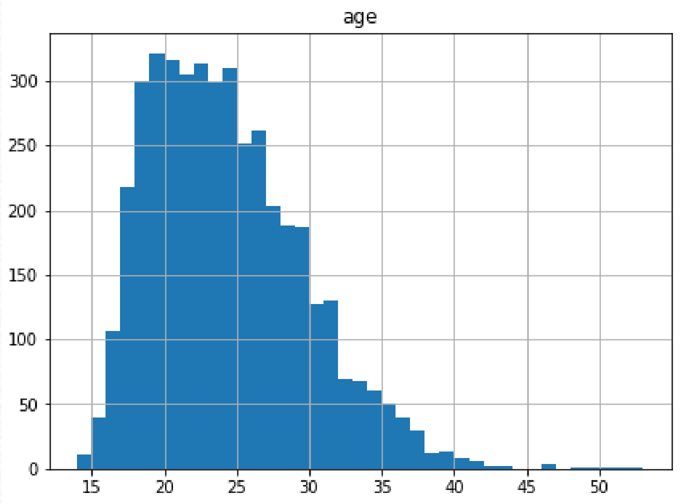
- **Before starting** the clustering and predictive analysis of the players, we observed the distribution of the our hand-engineered attributes for a **preliminary analysis**

Sex players distribution



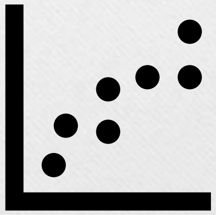


Ioc players distribution



Ioc players distribution

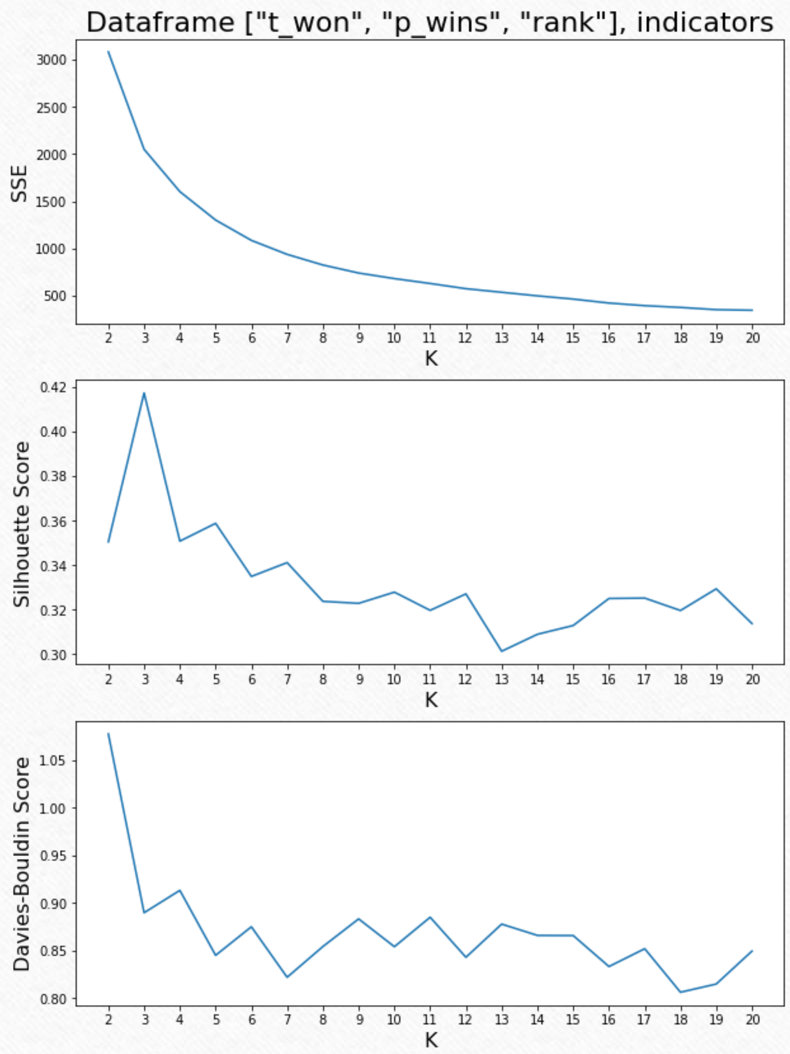
Clustering Analysis



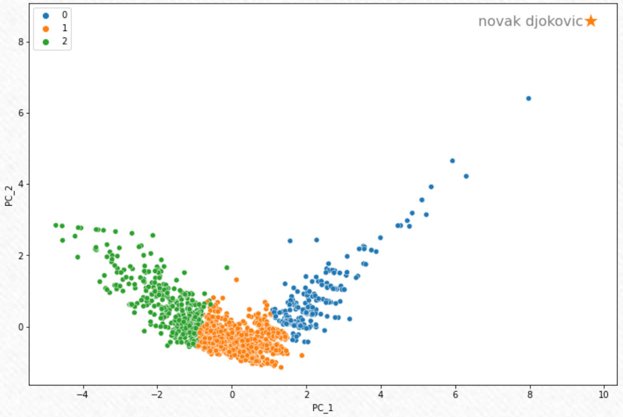
Clustering analysis:

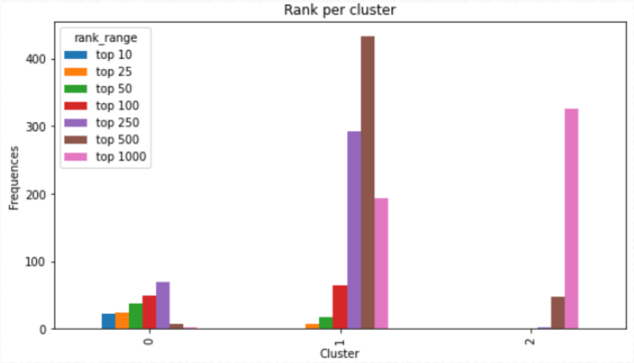
K-Means

- We **normalize the numerical features by using their z-score** by the *StandardScaler* from scikit-learn.
- We then tried **k-means on three different sets of features**.
- For each of one the **sets we found the best k value** (number of clusters) by using the **elbow rule on the SSE** and by looking at the **Silhouette and Davies-Bouldin scores**.
- We then **chose the set of features** which we thought had clustering results that could be **easier to understand** for those that don't follow the sport.

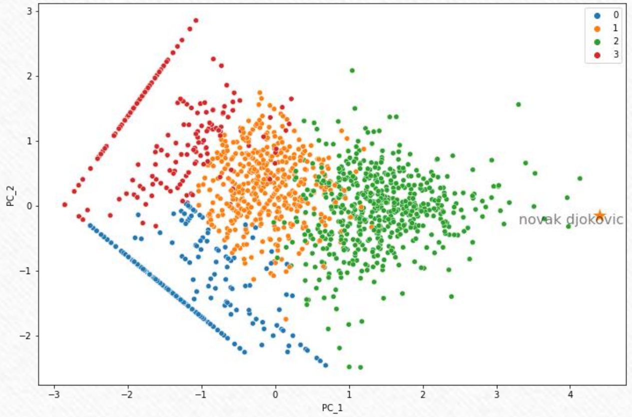


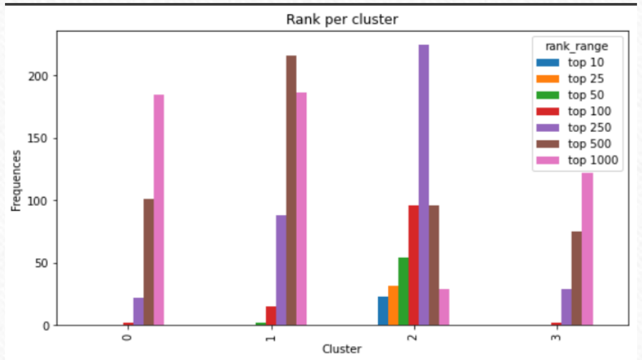
Clustering analysis:
K-Means



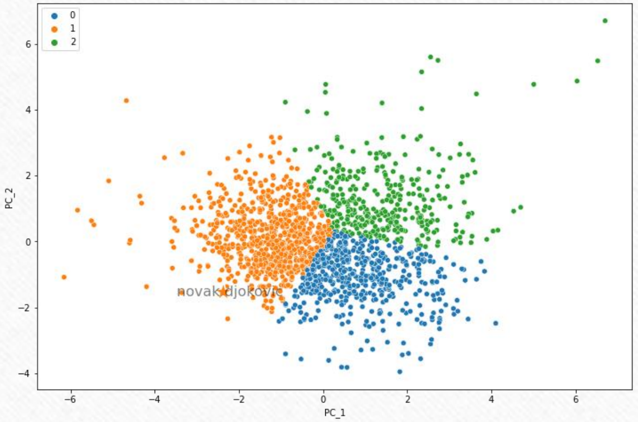


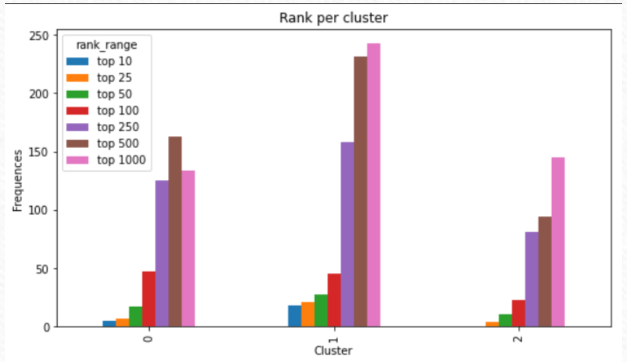
[t_won, p_wins, rank]





[p_w_Hard, p_w_Clay, p_w_Grass]



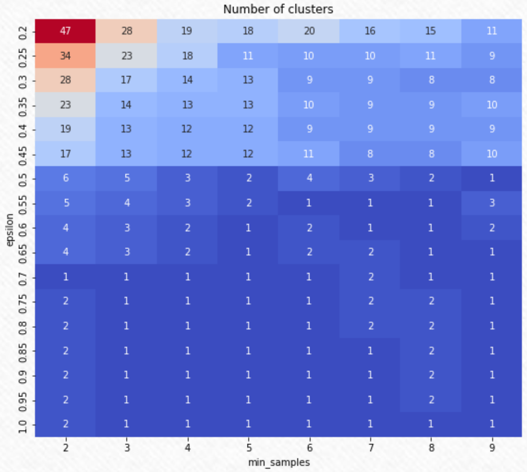


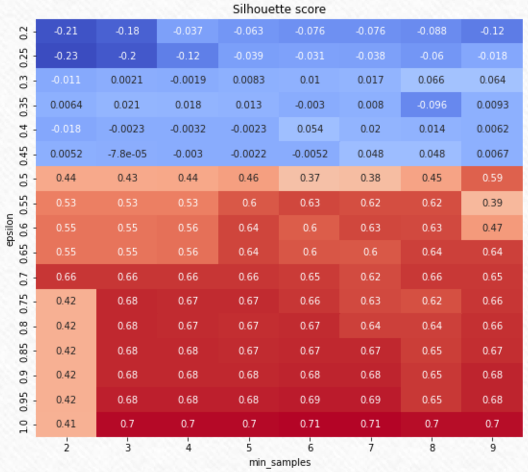
[p_aces, p_df, p_1st_in, p_1st_won, p_2nd_won]

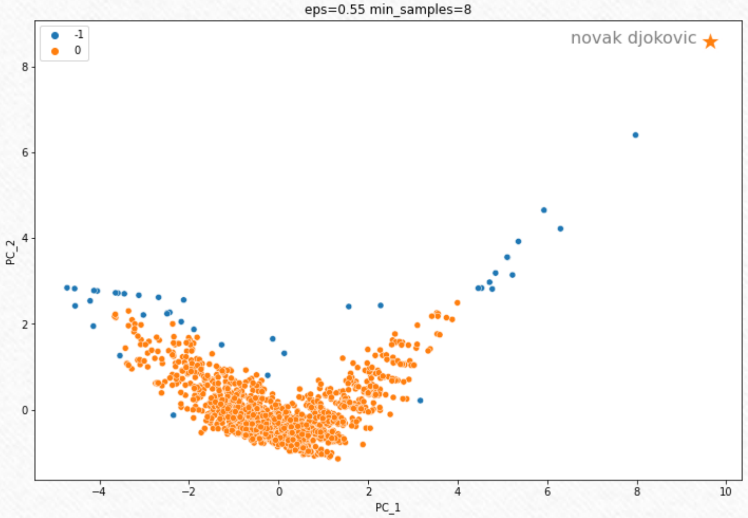
Clustering analysis:

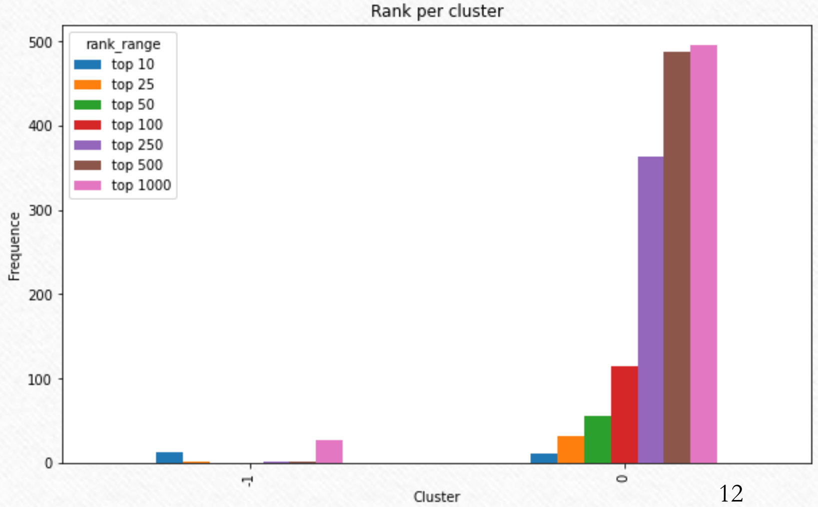
DBScan

- We did a **grid search** for finding the **best value of eps and min_samples**.
- We then **applied DBScan on the dataframe previously chosen** during K-Means.
- The **results are really underwhelming**, all the players fall in the **same cluster.**
- **The noise points are the best and worst players plus some players that have done some exploits** (like winning minor tournaments.



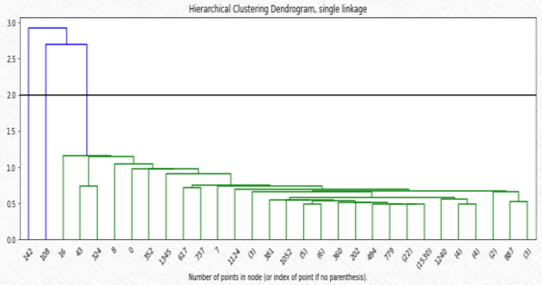


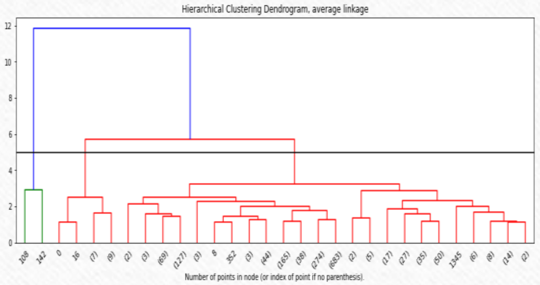


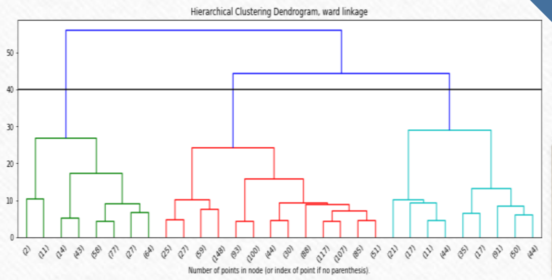


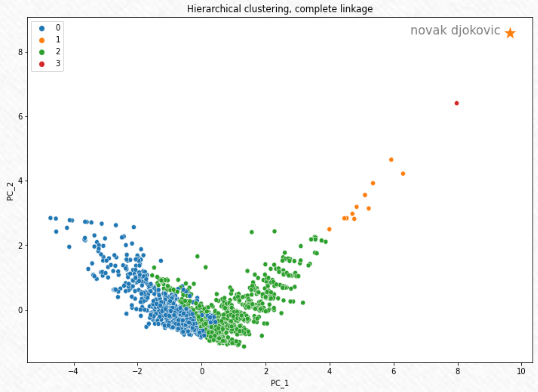
Clustering analysis:
Hierarchical clustering

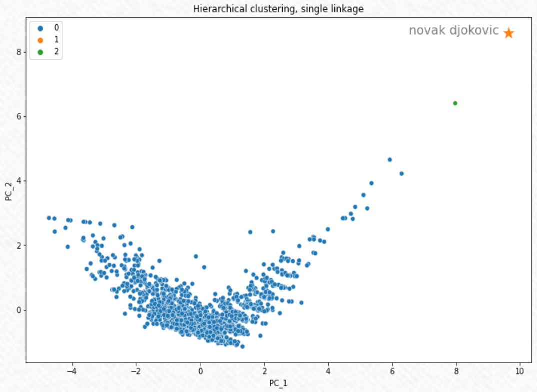
Method	cluster id : its dimension	Silhouette
Complete	0: 759, 1: 12, 2: 827, 3: 2	0.3082
Single	0: 1598, 1: 1, 2: 1	0.8013
Average	0: 1580, 1: 2, 2: 18	0.6485
Ward	0: 330, 1: 296, 2: 974	0.3756

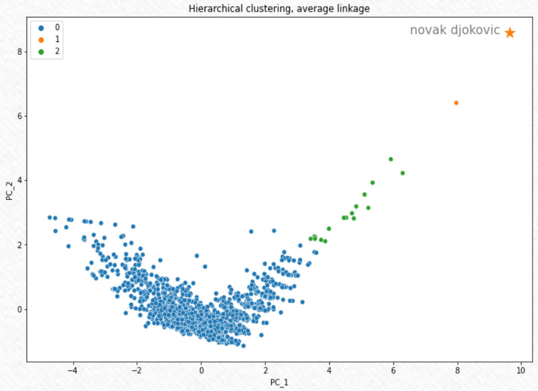


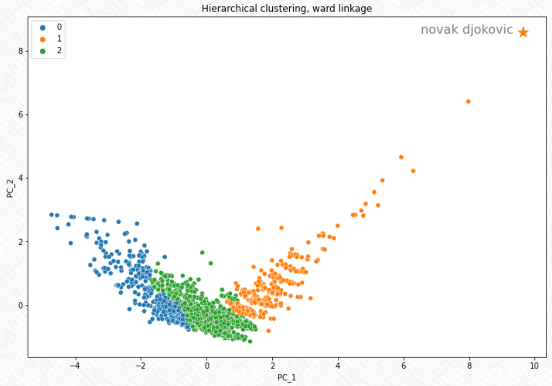










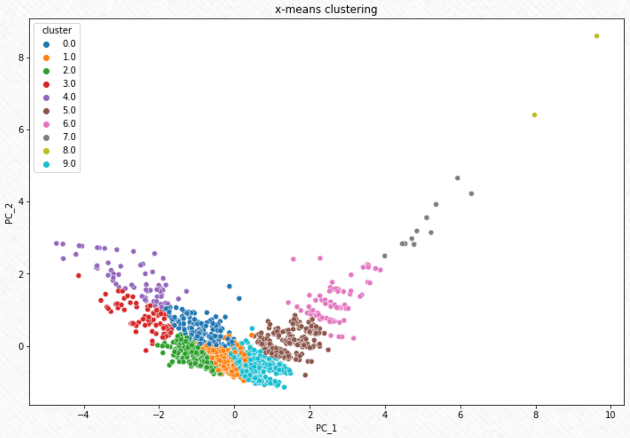


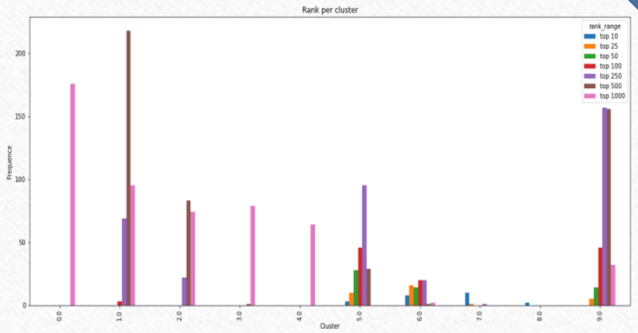
Some **similar results to K-Means** and some **extremely strange ones**. We've **learnt to not trust the Silhouette** alone.

Clustering analysis:
X-Means and SOM

X-Means

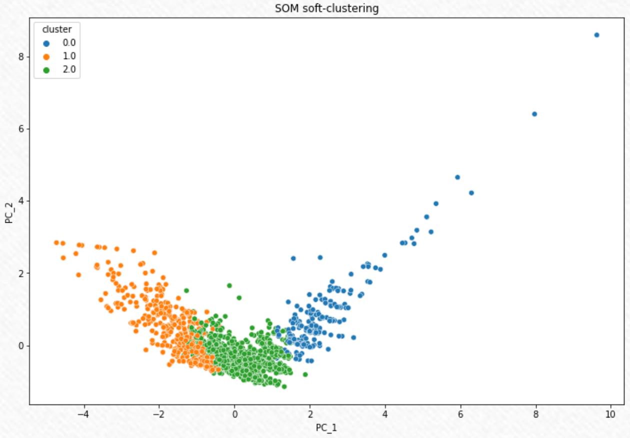
- **Starts with one cluster** and then **creates more by splitting at each epoch**.
- **Extremely different results** between each run.
- The players' are split into **more "gradual" clusters.**

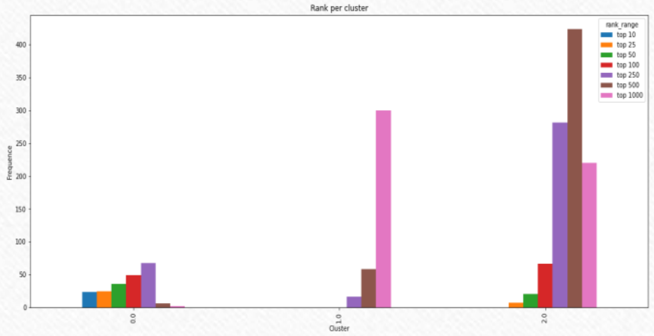




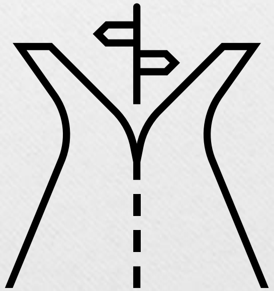
SOM Soft Clustering

- Uses **Self-Organizing Maps** as means to create clusters.
- The results are **somewhat similar to those obtained by K-Means**.





Predictive Analysis



Predictive analysis:
dataset initialization

Dataset derived from the data preparation phase, totalling 1600 players

Female players: 32%

Male players: 68%

Weak players 74%

Strong players 26%

24%

Weak players 76%

Players	Male	Female
Strong	Top 5, 10, 25, 50, 100	Top 5, 10, 25, 50, 100, 250
Weak	Top 250, 500, 1000	Top 500, 1000

Over-sampling by using SMOTE and under-sampling in order to get a 55/45 distribution for the weak and the strong class, respectively.

Predictive analysis:
model selection

- Having split the dataset into male and female players, we developed some models that can classify only the male players and others that work exclusively on the female players.
- This means that at inference time, the players' sex has to be known a priori.
- For those models that need hyperparameters tuning, we use the following workflow:

/

Development set: 85%

Test set: 15%

↓

K-Fold cross validation, k=4.
We used the accuracy as a target metric.

→ Best hyperparameters for each model. →

Models for **male** players

Models for **female** players

Predictive analysis:
model assessment

Female players classification								
Model	Training				Test			
	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.
Decision Tree	95/89	91/94	93/91	92	96/62	82/89	89/73	84
Naive Bayes	91/84	86/90	89/87	88	96/62	82/89	89/73	84
Random Forest	100/97	98/100	99/98	99	95/75	91/83	93/79	89
AdaBoost	100/100	100/100	100/100	100	95/83	95/83	95/83	92
Rule Based	90/93	95/87	92/90	91	91/48	74/78	82/60	75
KNN	95/86	88/95	91/90	91	94/68	88/83	91/75	87
SVM	100/95	96/100	98/97	97	95/75	91/83	93/79	89
Neural Network	92/87	89/91	91/89	90	98/71	88/94	93/81	89
TabNet	97/85	86/97	91/91	91	93/70	89/78	91/74	87

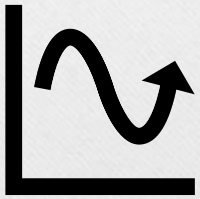
A comparison between the performance of the algorithms employed to classify the **female** players.

Male players classification								
Model	Training				Test			
	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.
Decision Tree	97/95	96/96	96/96	96	99/81	92/98	95/88	93
Naive Bayes	94/88	89/92	92/90	91	97/76	90/91	93/83	90
Random Forest	100/97	98/100	99/99	99	99/84	93/98	96/90	95
AdaBoost	95/91	93/94	94/93	93	99/79	91/98	95/88	93
Rule Based	95/93	95/94	95/94	94	92/65	85/79	89/72	84
KNN	96/94	95/95	96/95	95	92/82	93/84	94/83	91
SVM	100/97	98/99	99/98	98	95/88	96/86	96/87	93
Neural Network	93/92	94/91	93/92	93	98/85	94/95	96/90	95
TabNet	94/88	90/93	92/90	91	97/85	94/93	96/89	94

A comparison between the performance of the algorithms employed to classify the **male** players.

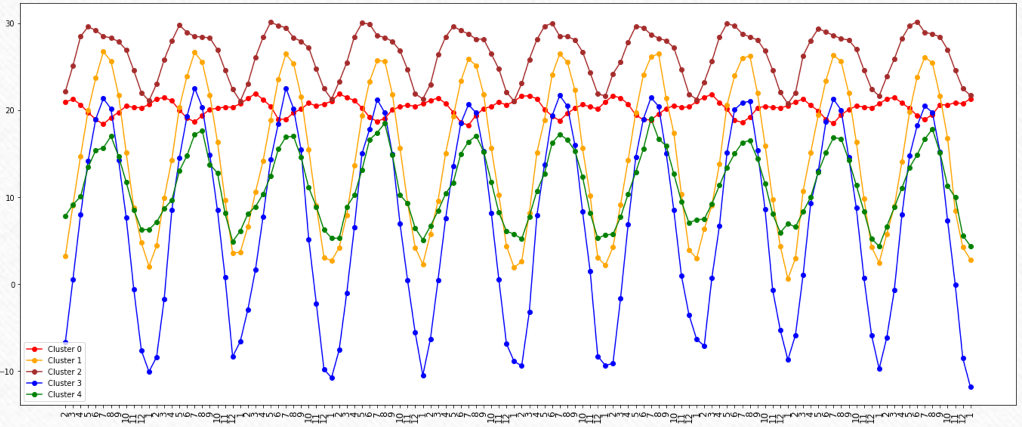
- We summarized the performance of the models by distinguishing those that work with the female data to the ones that work with the male data.
- As described before, female players were only 32% of the entire dataset while male players were 68%. As a consequence, the models fit on the male players dataset carried out much better results than the one fit on the female players dataset.
- AdaBoost is the best model to classify female players, whereas Random Forest works better for the male players.

Time Series Analysis



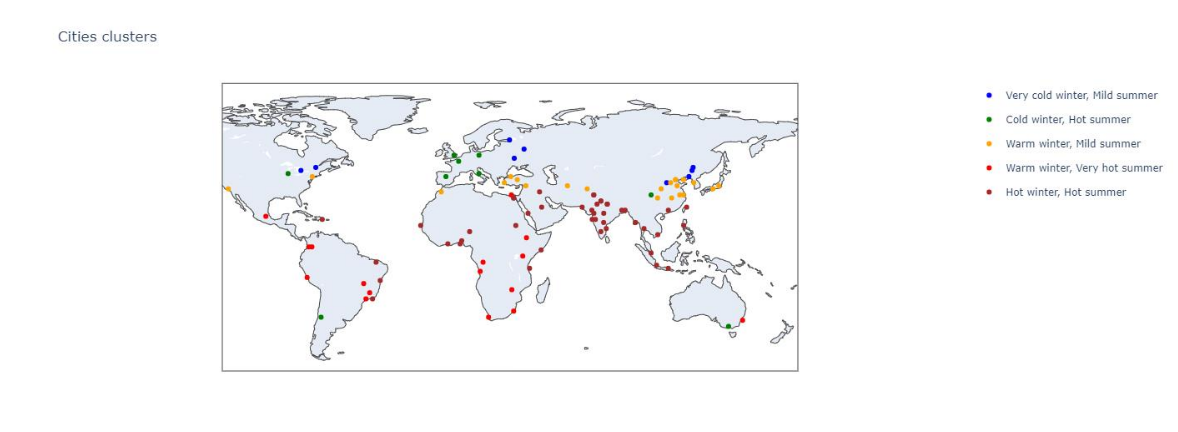
Time Series analysis:

Shape-based clustering



time City	2000-02-01	2000-03-01	2000-04-01	2000-05-01	2000-06-01	2000-07-01	2000-08-01	2000-09-01	2000-10-01	2000-11-01	2000-12-01	2001-01-01	2001-02-01	2001-03-01	2001-04-01	2001-05-01	2001-06-01	2001-07-01	2001-08-01	2001-09-01	2001-10-01
Abidjan	27.685	29.061	28.162	27.547	25.812	24.870	24.884	25.405	26.074	27.315	26.929	26.920	28.234	28.706	27.702	27.653	25.940	24.841	24.280	24.797	26.278
Addis Abeba	19.183	20.230	20.398	19.977	18.254	17.109	16.944	17.542	17.113	17.741	17.013	17.454	18.864	20.043	20.233	19.908	17.978	17.011	17.152	17.867	18.047
Ahmadabad	21.246	26.565	32.275	32.847	32.490	28.678	28.616	29.087	29.285	25.577	21.785	19.770	22.438	27.198	31.034	33.358	30.717	27.730	27.893	29.490	29.073
Aleppo	6.832	10.421	17.743	22.240	27.781	31.957	29.873	25.877	18.846	13.380	7.636	7.306	8.787	14.904	17.666	21.191	27.922	30.727	30.330	26.419	20.153
Alexandria	14.300	15.266	19.556	21.828	24.619	27.091	26.930	25.939	22.842	20.524	16.460	15.312	15.403	18.834	19.709	22.533	24.652	27.094	28.246	26.963	23.599
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Tokyo	1.765	5.709	11.155	17.261	20.396	25.214	25.843	22.005	15.197	9.572	4.245	0.573	2.519	6.038	12.373	17.246	20.659	26.326	24.092	20.184	14.846
Toronto	-4.177	2.675	4.968	13.068	17.095	18.296	18.522	14.296	9.723	1.849	-8.854	-5.783	-5.604	-2.592	6.434	13.708	17.972	19.042	20.842	14.547	8.854
Umm Durman	24.991	27.506	32.774	34.786	34.795	32.279	31.233	31.790	30.641	27.746	23.539	22.502	24.995	29.504	33.552	34.729	33.332	31.398	30.242	32.027	31.831
Wuhan	5.842	13.016	17.898	23.914	26.511	30.222	28.404	24.146	17.993	10.449	7.588	5.061	7.204	12.875	16.768	23.488	26.109	30.729	27.718	25.359	19.310
Xian	0.943	8.997	13.714	20.568	22.387	25.578	22.517	17.823	11.116	4.095	1.265	-1.249	2.649	8.873	12.072	19.679	23.561	26.094	23.148	17.400	12.659

- The dataset was **built by using the method pivot()** from pandas.
- **The cities were split into five clusters** in accordance to their **temperature trends**.
- **DTW performed better than euclidean distance since it takes into account the shift** caused by the seasons in different emispheres.

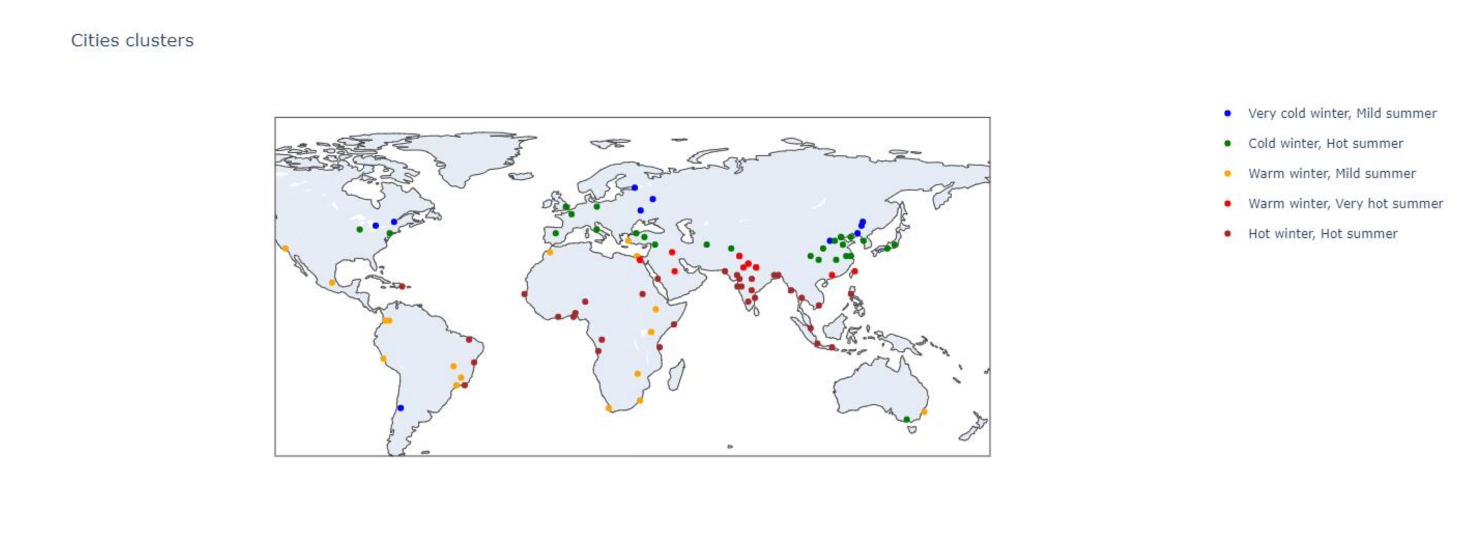


Time Series analysis:

Feature-based clustering

- The features we built are emisphere agnostic.
- The **results are similar to those in shape-based clustering**.
- **No city had a major shift** (cold city now classified as hot or viceversa).
- **We tried using the altitude** to see if it has a major influence, but there's only one city with an high value (Mexico City) and it's on the tropic so it's still an hot city.

City	Mean_t	Max_t	Min_t	Mean_t_spring	Mean_t_summer	Mean_t_autumn	Mean_t_winter
Abidjan	26.930375	28.9006	24.8005	27.288200	25.163233	27.032300	28.237767
Addis Abeba	18.351717	20.1868	17.0864	19.386967	17.278933	17.448633	19.292333
Ahmadabad	27.416742	33.5029	20.2431	32.368867	28.550233	25.237767	23.510100
Aleppo	18.345783	30.6218	6.0204	22.338700	28.934033	13.338167	8.772233
Alexandria	21.331192	27.8062	14.8878	22.221067	27.251033	20.196133	15.656533
...	...	...	...	...	...	...	...
Tokyo	13.370042	24.9984	2.0084	16.403667	23.487200	9.685000	3.904300
Toronto	6.934975	19.6951	-6.8175	11.824533	18.299633	2.579400	-4.963667
Umm Durman	29.882492	34.5343	23.1234	33.874233	31.481200	28.315100	25.859433
Wuhan	17.779067	29.7552	4.5549	22.614300	27.624333	12.683967	8.193667
Xian	12.469417	25.0434	-1.3598	18.846433	22.008100	5.877267	3.145867



рахмат
danke
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
спасибо
faafetai lava
vinaka
mersi
kia ora
barka
welalin
tack
spas
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
blagodaram
спасибі
kiitos
dankie
nanni
nandri
dhanyavad
maururu
thank you
gracias
хвала
asante
manana
obrigada
tenki
bedankt
bayarlalaa
gràçie
hvala
köszönöm
dankon
ačiū
akun
enkosi
dziękuję
djiere dieuf
tau
mochchakkeram
murakoze
дякую
mamnun
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
chokrane
sobodi
děkuji
sagolun
sukriya
kop khun krap
obrigado
mèsi
ありがとう
tanemirt
grazie
arigatô
takk
dakujem
trugarez
didi madloba
najis tuke
terima kasih
rahmet
diolch
dhanyavadagalu
shukriya
mercé
мерси
kam sah hamnida
rahmat
ευχαριστώ
xièxie
merci
তোমাকে ধন্যবাদ
감사합니다