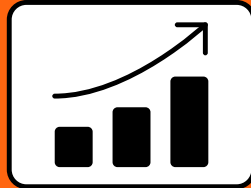


I CHOOSE YOU...

# Pokemon study



Dzmitry Kurch, DSE



# Meeting Agenda

Discussion topics for today

01

**Exploratory  
Data Analysis**

02

**Classification**  
Logistic  
regression

03

**Dimensional  
Reduction and  
clustering:**  
PCA (classic/  
non-linear),  
DBSCAN,  
Agglomerative

04

**Regression**  
Generalized  
Linear Model



# Exploratory data analysis



# Dataset

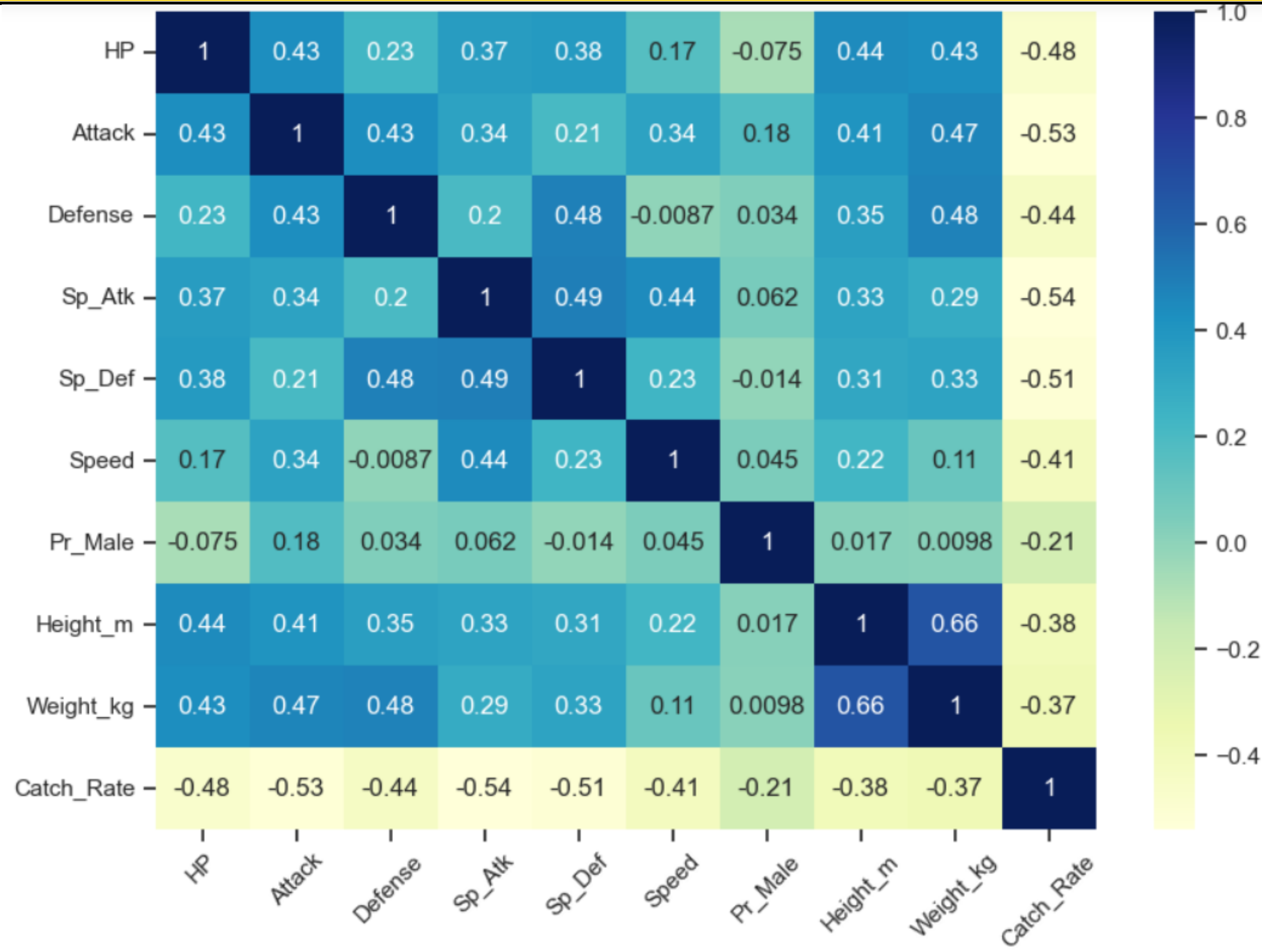
## Dataset sample

	Number	Name	Type_1	Type_2	Total	HP	Attack	Defense	Sp_Atk	Sp_Def	Speed	Generation	isLegendary	Color	hasGender	Pr_Male	Egg_Group
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False	Green	True	0.875	Mon
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False	Green	True	0.875	Mon
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False	Green	True	0.875	Mon
3	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False	Red	True	0.875	Mon
4	5	Charmeleon	Fire	NaN	405	58	64	58	80	65	80	1	False	Red	True	0.875	Mon

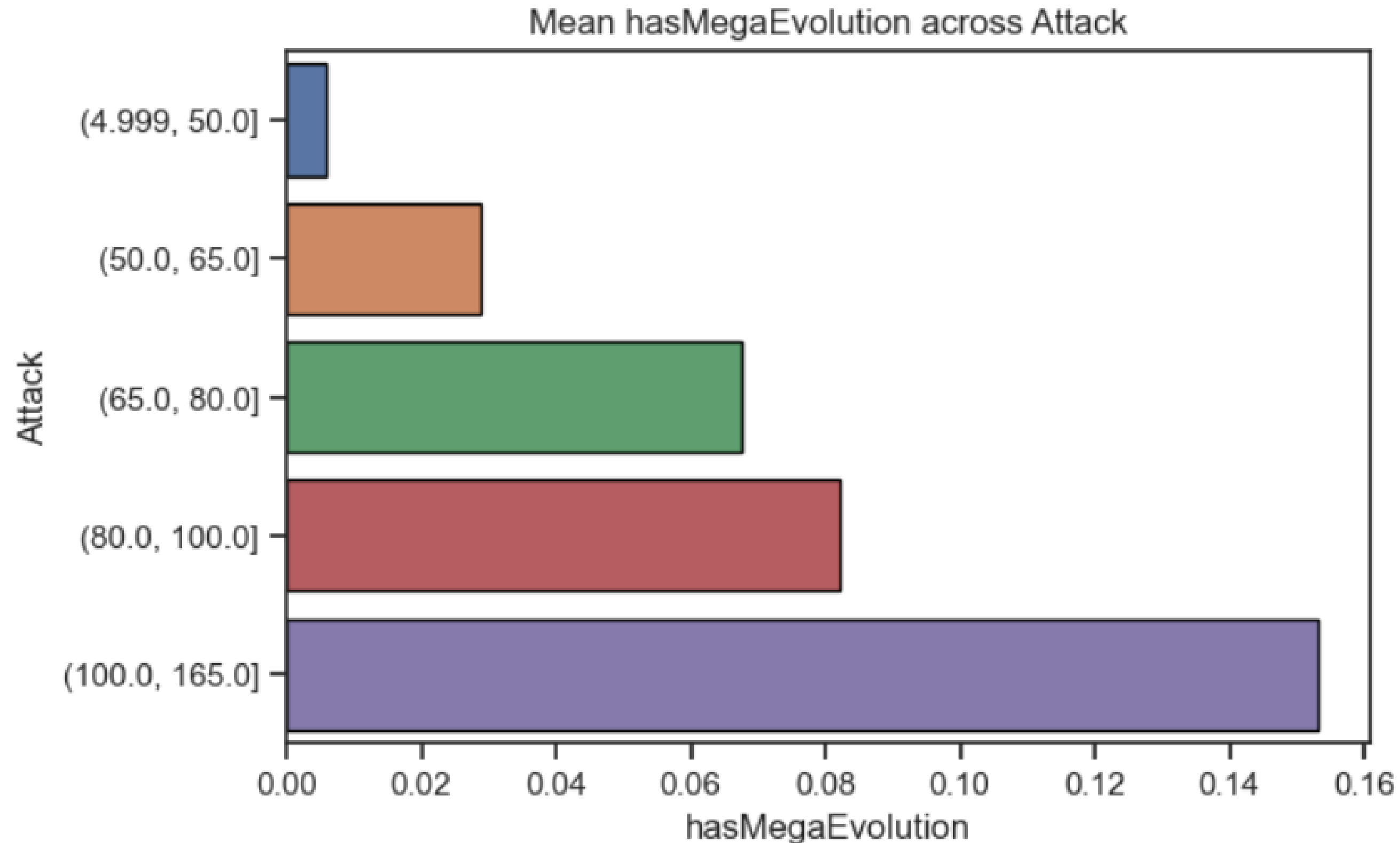
Dataset consists of the Pokemons and their characteristics.

Binary field **hasMegaEvolution** is an important feature in the Pokemon world which indicates its ability to have a temporary *superpower*. We have only **6.4%** samples with that attribute activated.

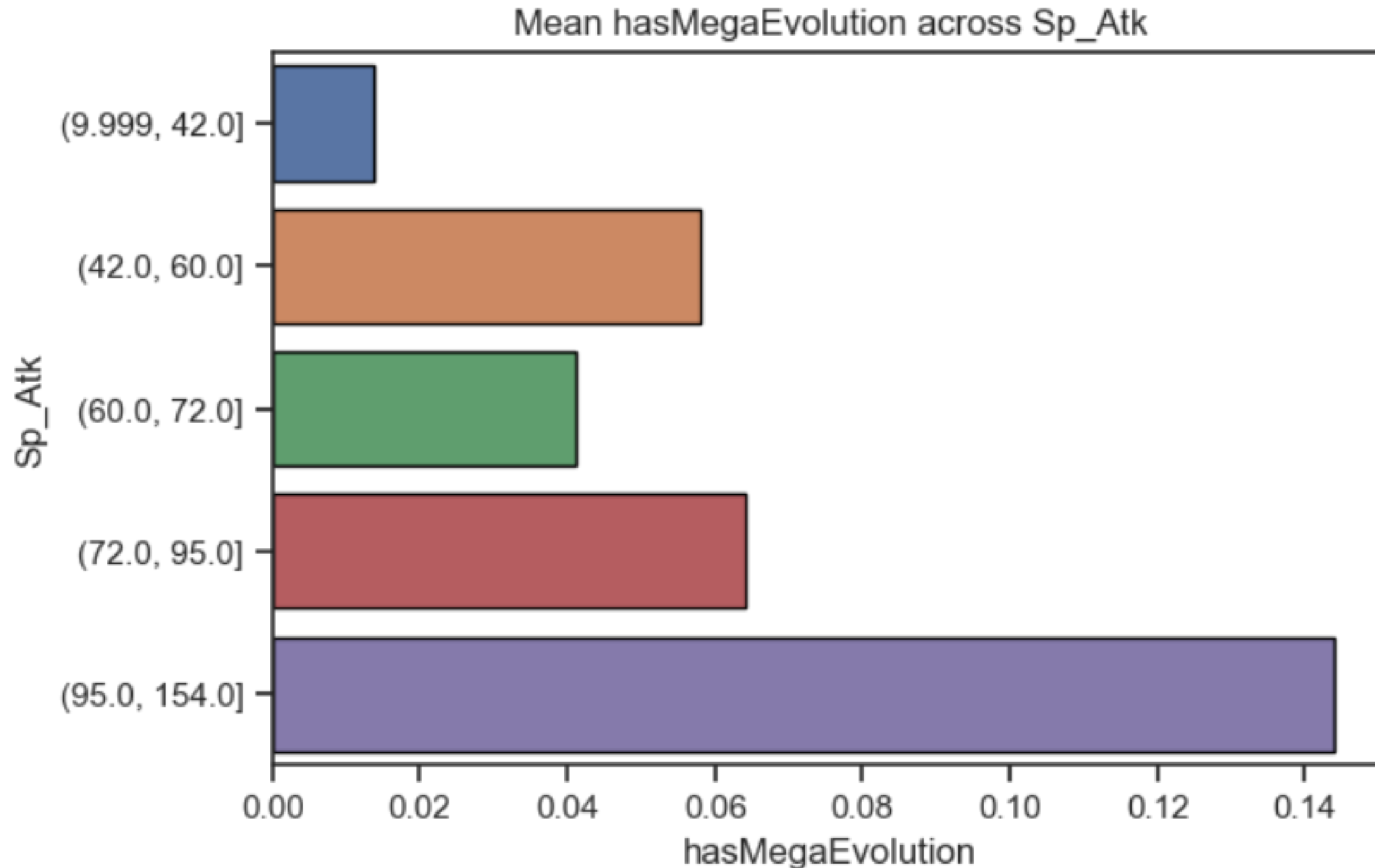
# Correlation matrix



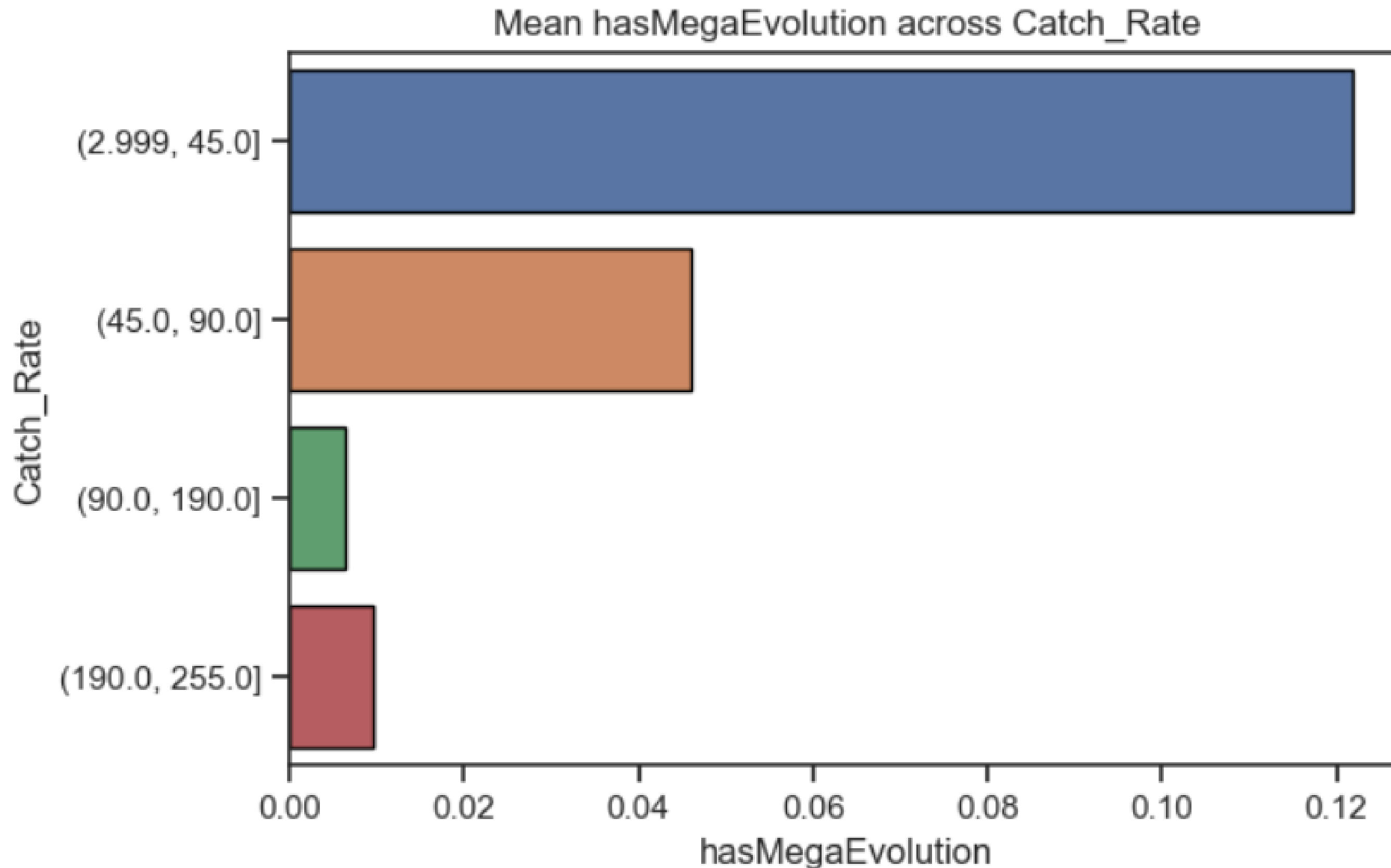
# Target density across features



# Target density across features

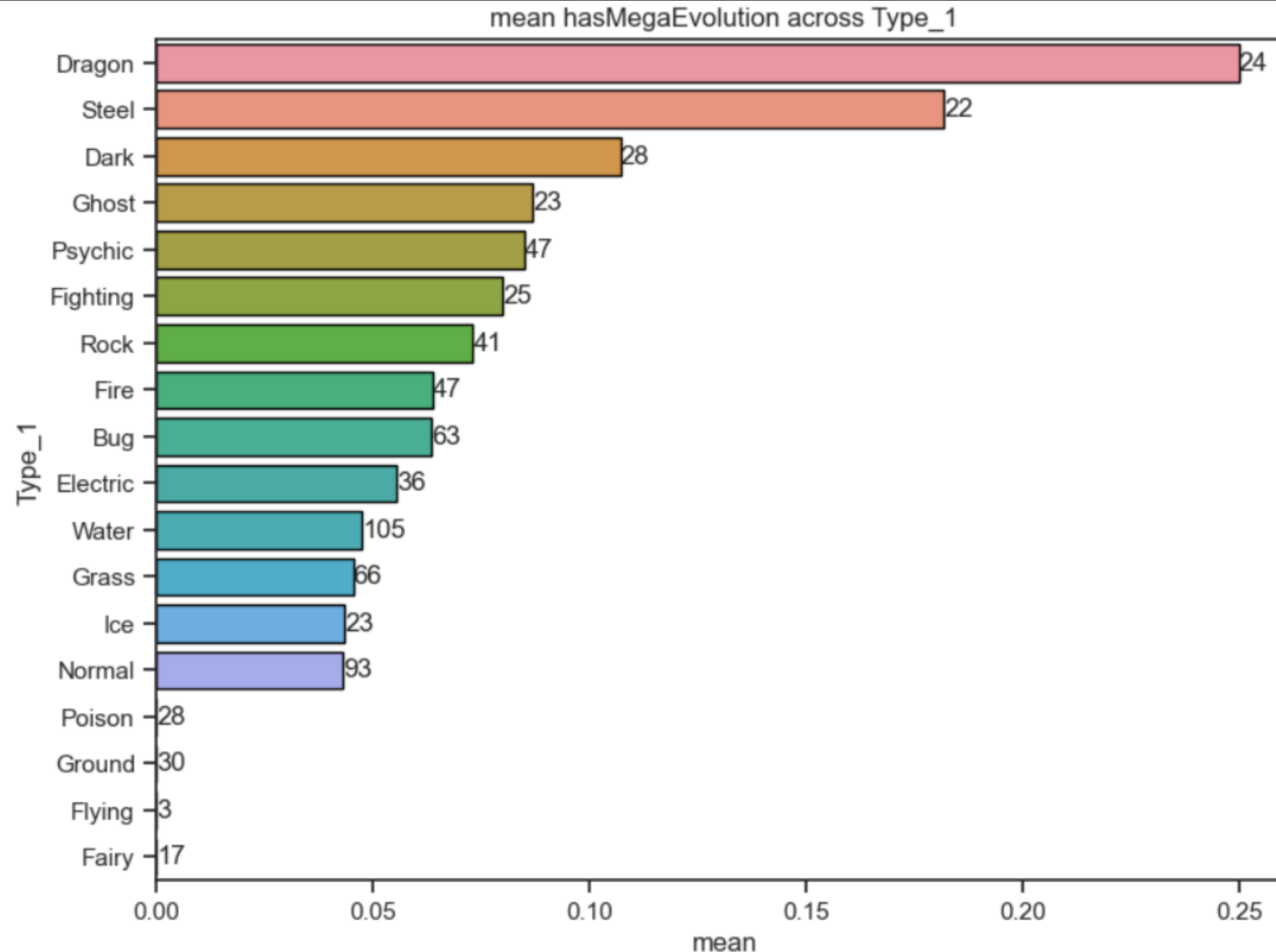


# Target density across features





# Target density across features





# Logistic regression

# Data preprocessing

## Dataset sample

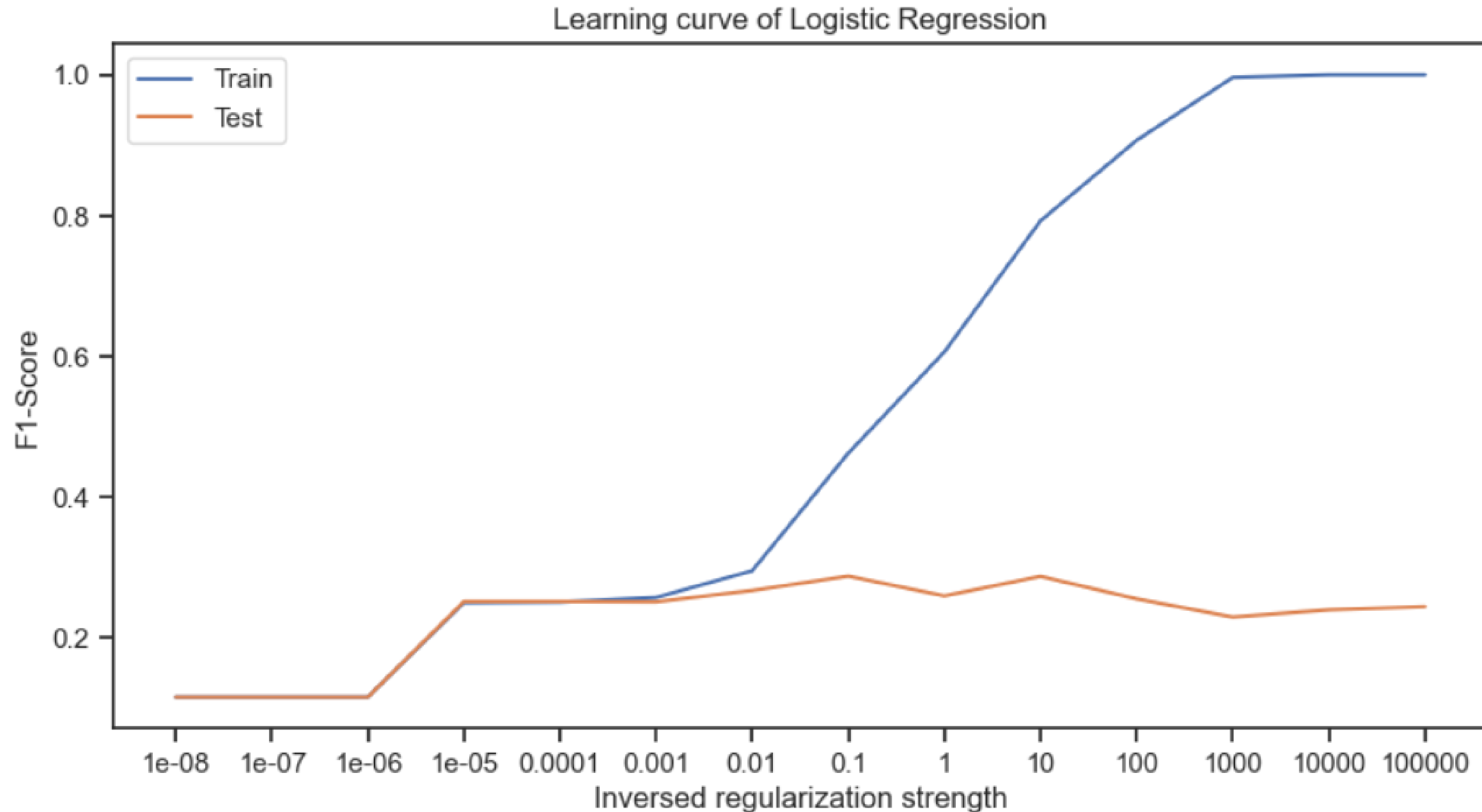
---

	Number	Name	Type_1	Type_2	Total	HP	Attack	Defense	Sp_Atk	Sp_Def	Speed	Generation	isLegendary	Color	hasGender	Pr_Male	Egg_Group
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False	Green	True	0.875	Mon
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False	Green	True	0.875	Mon
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False	Green	True	0.875	Mon
3	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False	Red	True	0.875	Mon
4	5	Charmeleon	Fire	NaN	405	58	64	58	80	65	80	1	False	Red	True	0.875	Mon

## Preprocessing steps

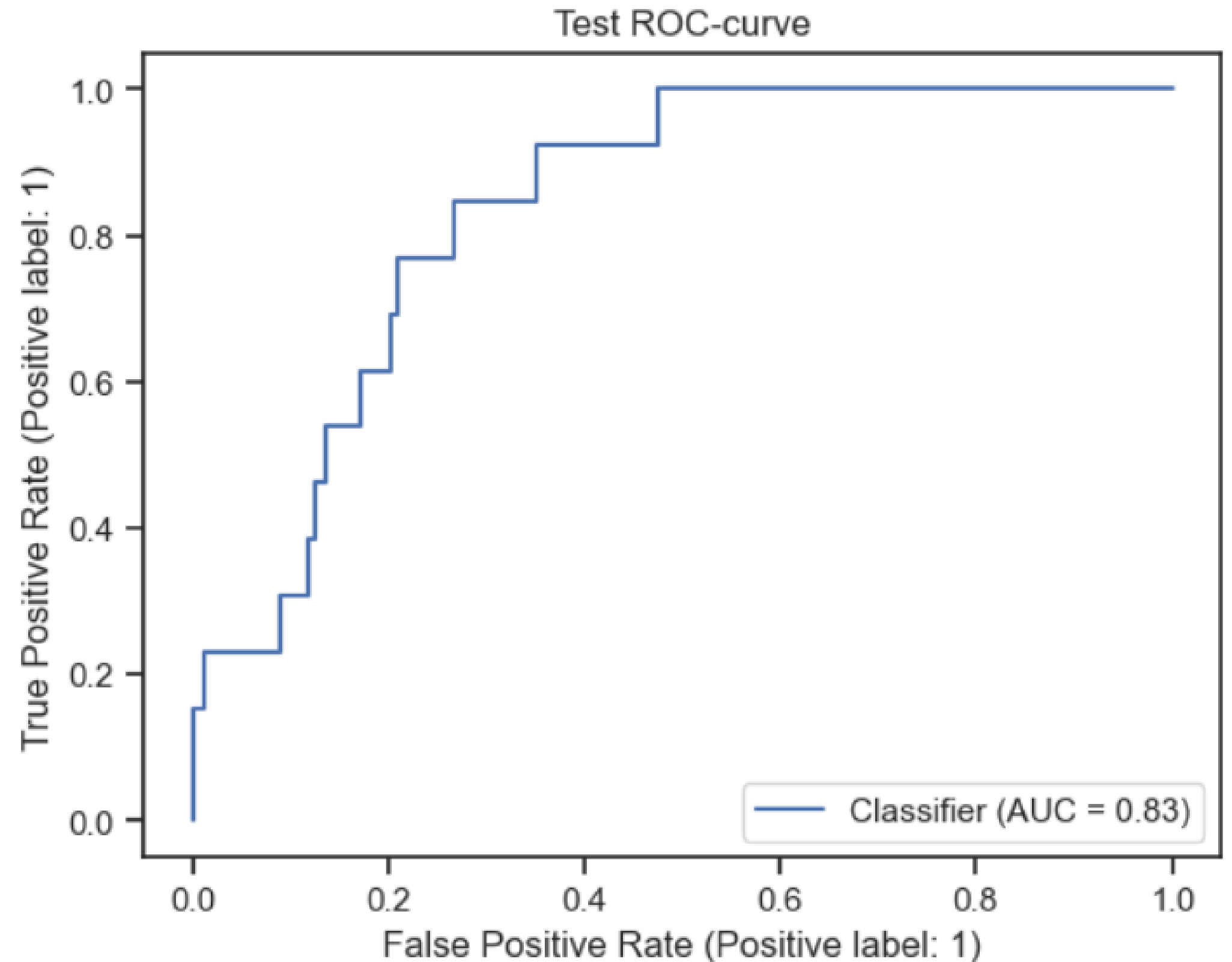
- Train-Validation-Test split
- OHE categorical features
- Scaling numerical features

# Hyperparameter tuning

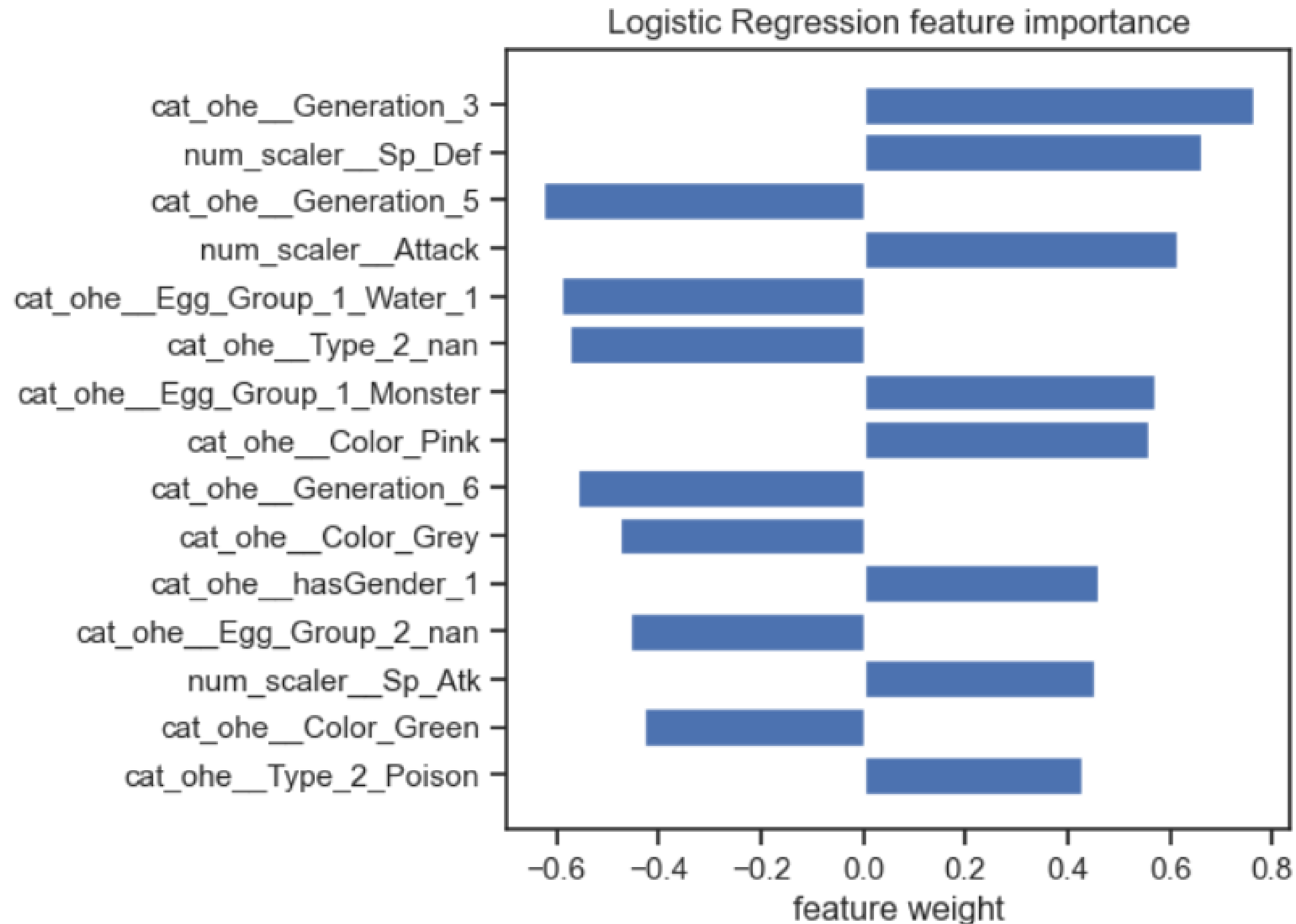


# Test results

	precision	recall	f1-score	support
0	0.95	0.88	0.91	168
1	0.19	0.38	0.26	13



# Feature importance

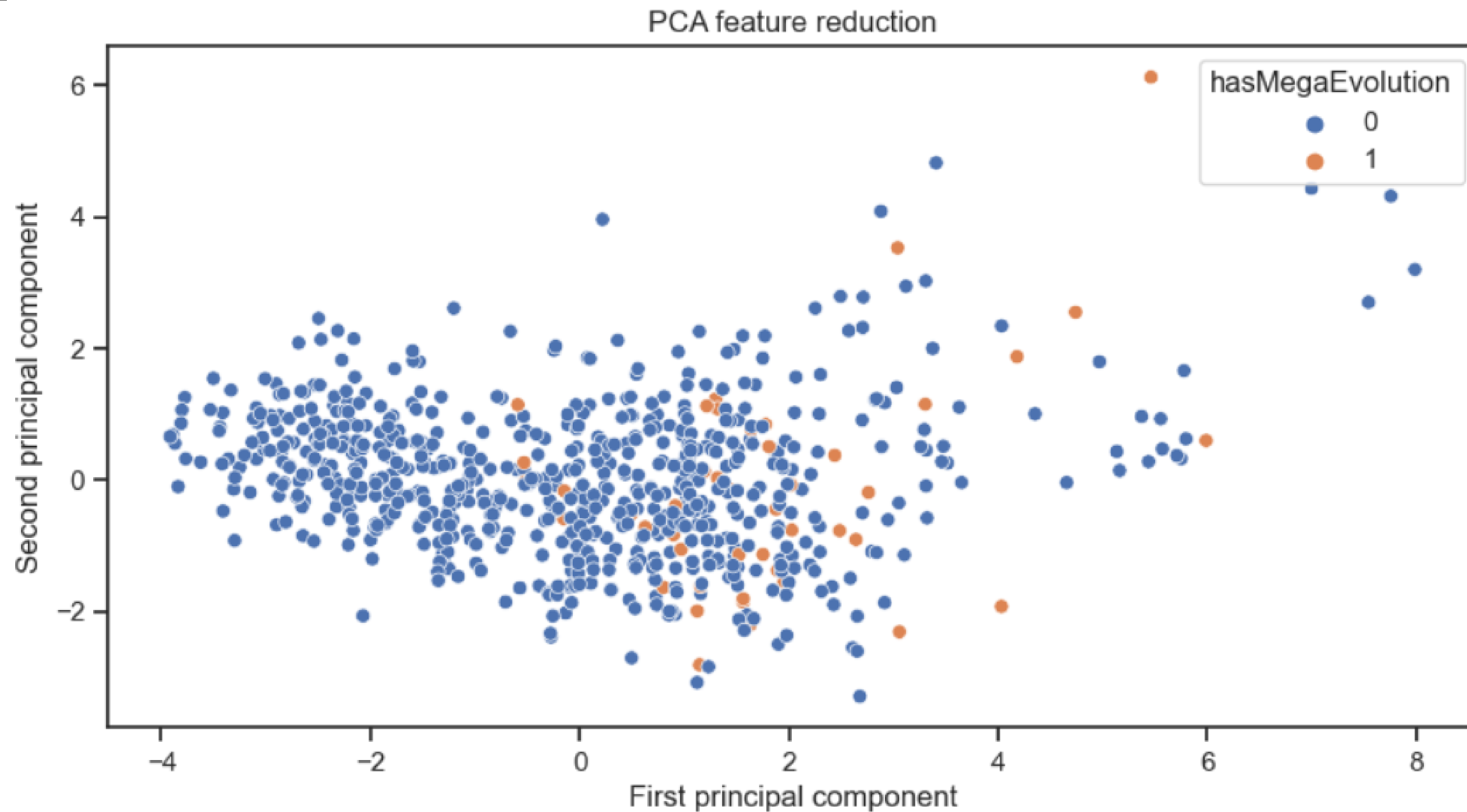




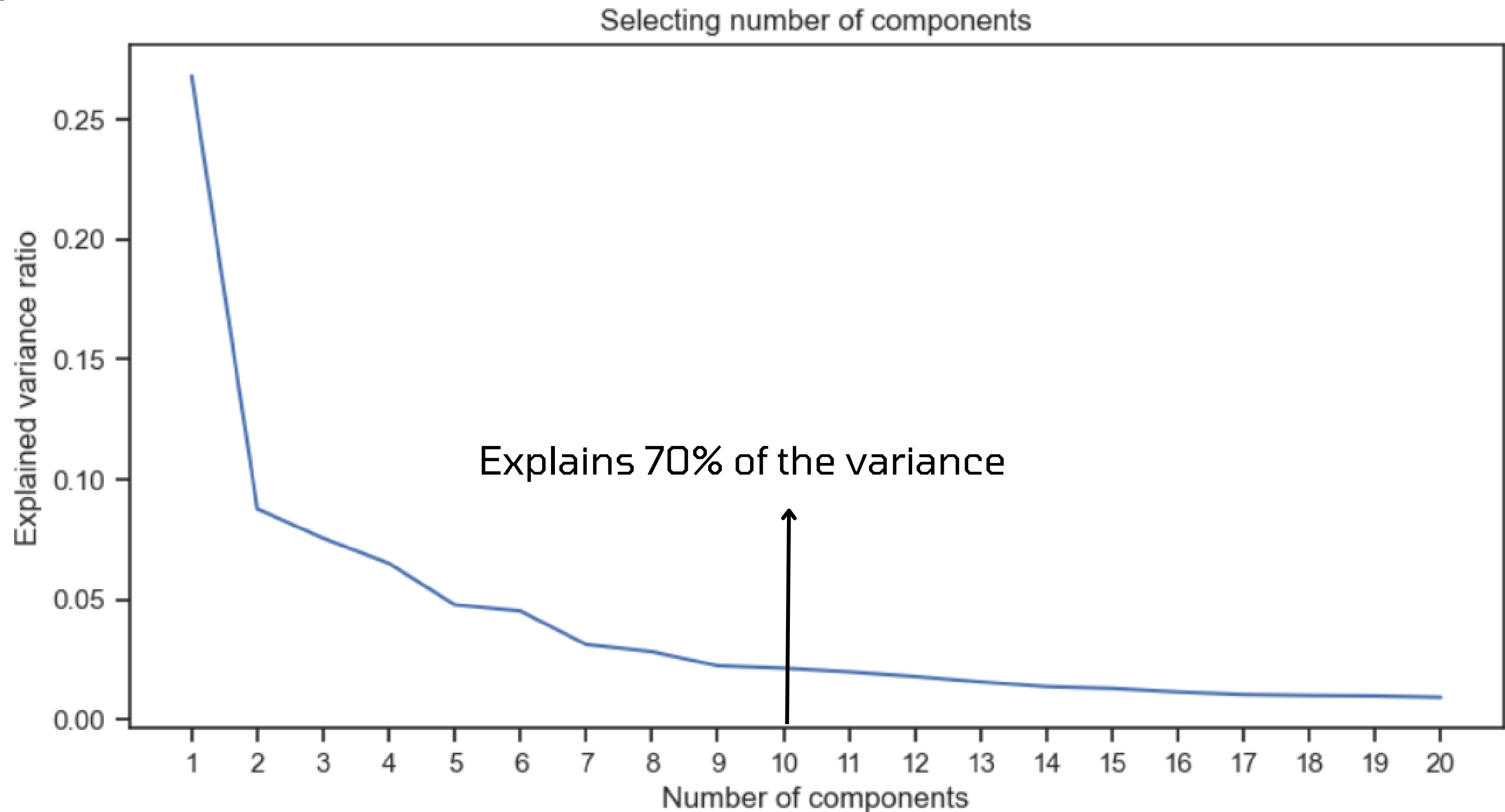


# Dimensional reduction

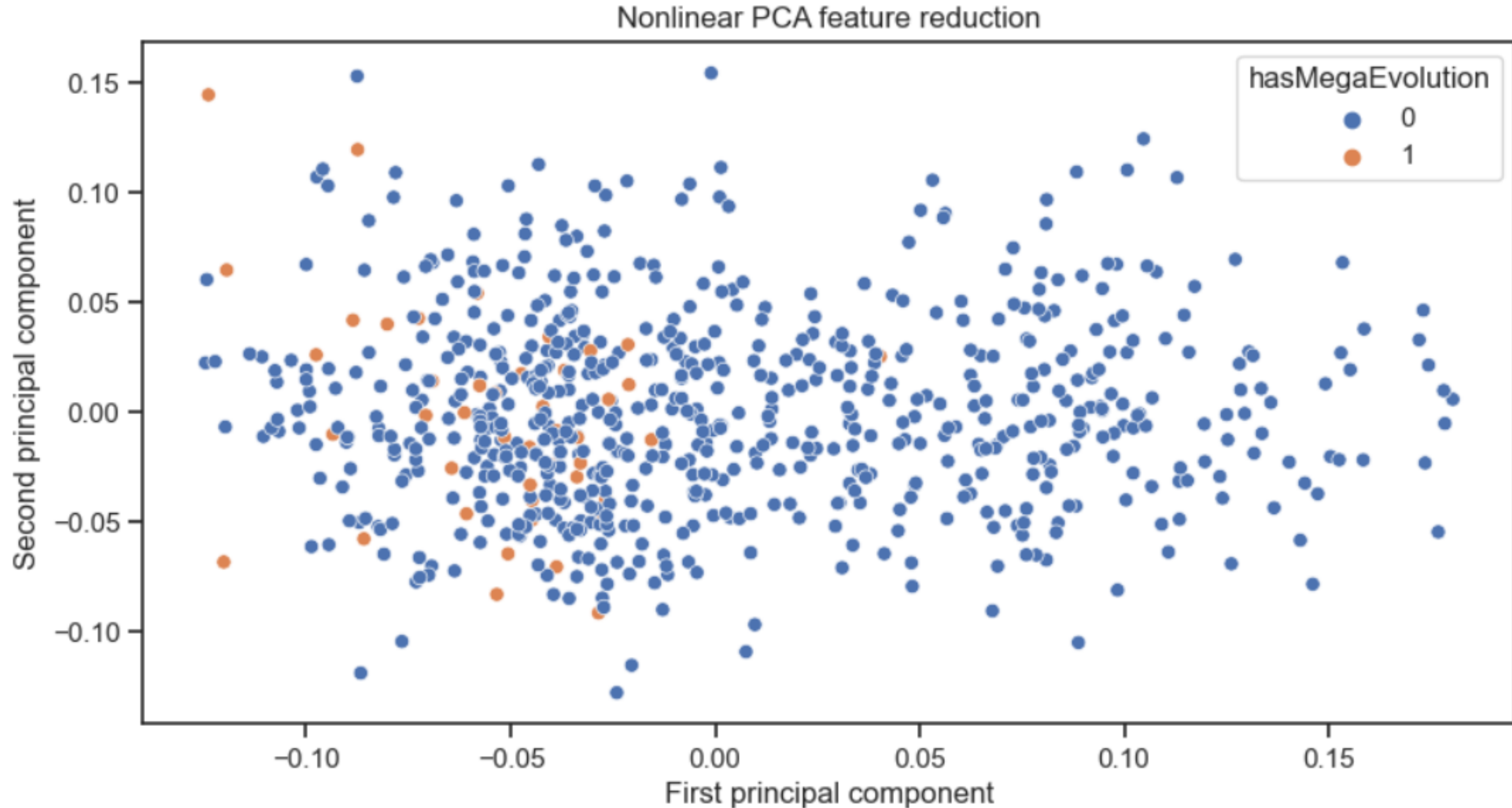
# PCA target



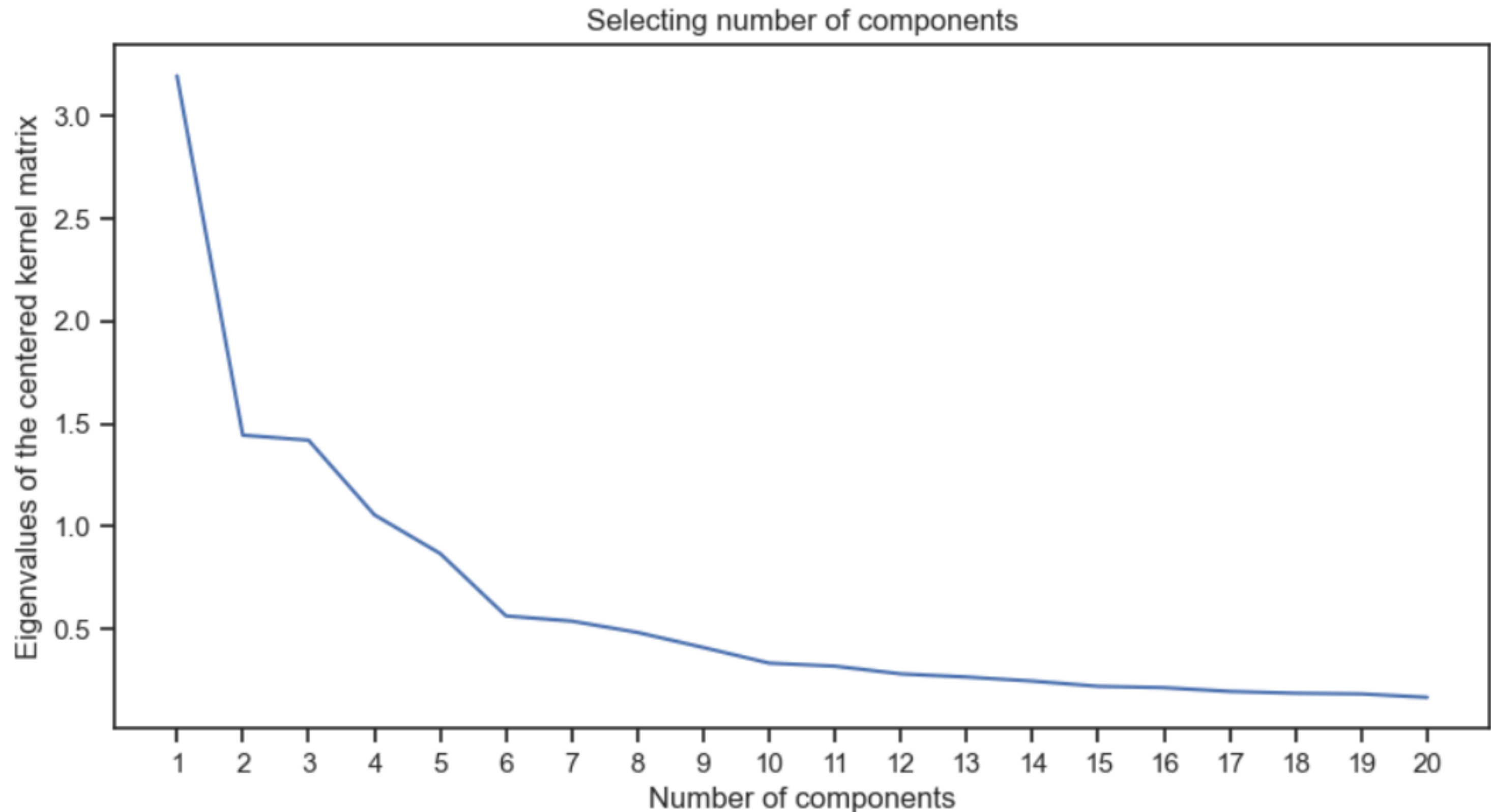
# PCA selecting number of comp



# Non-linear PCA (sigmoid kernel) target



# Non-linear PCA: Selecting number of comp

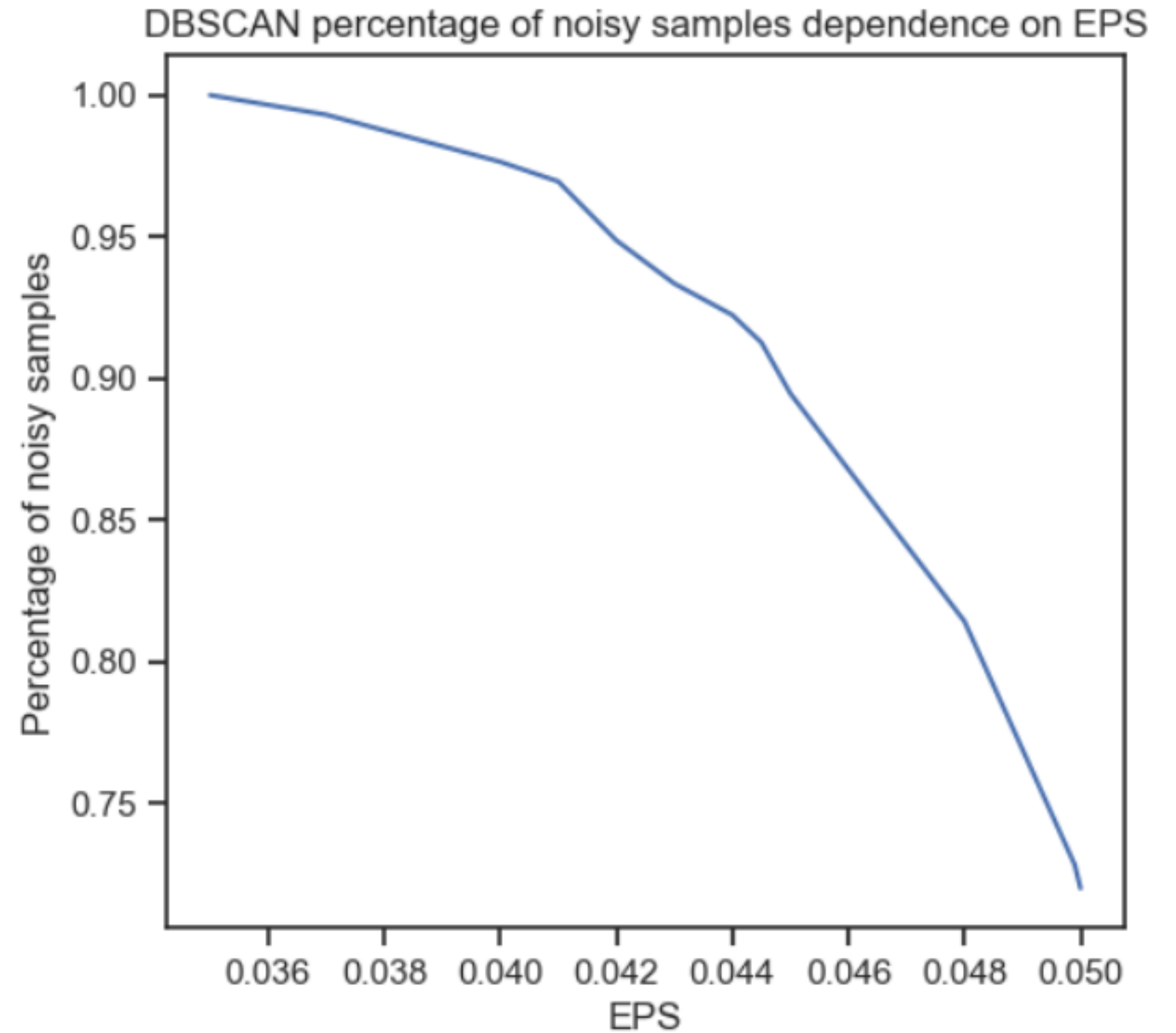
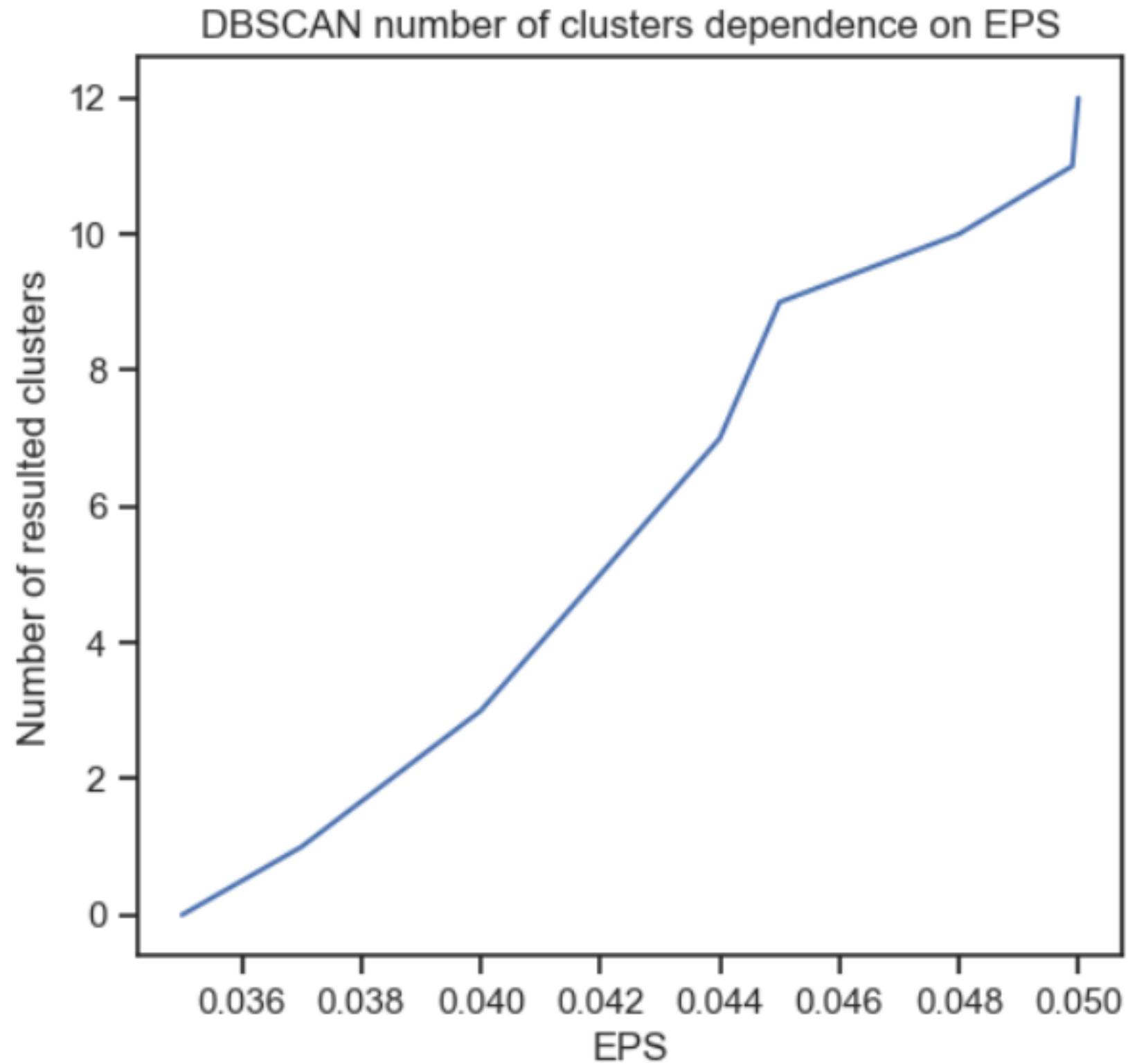




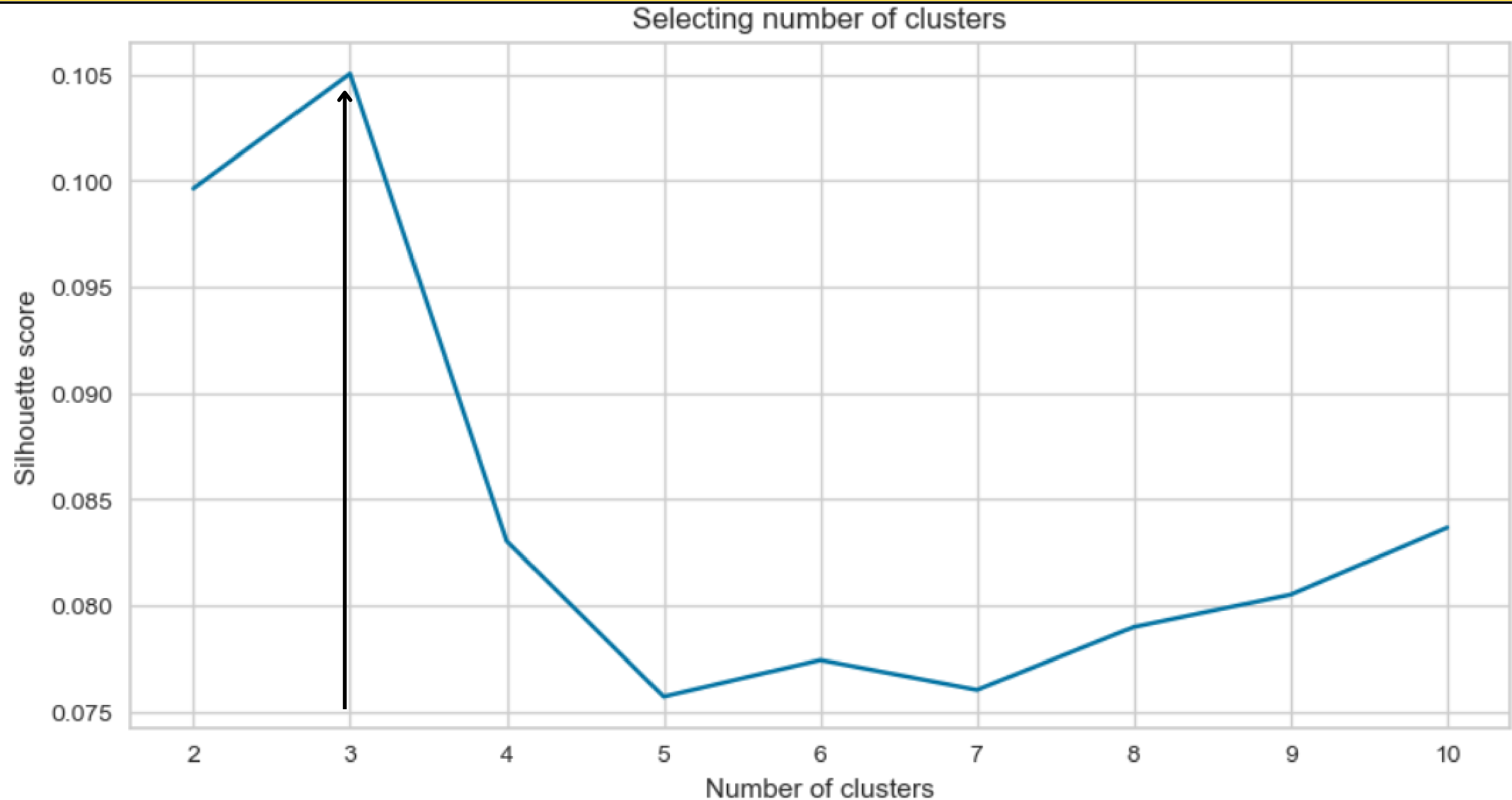
# Clustering



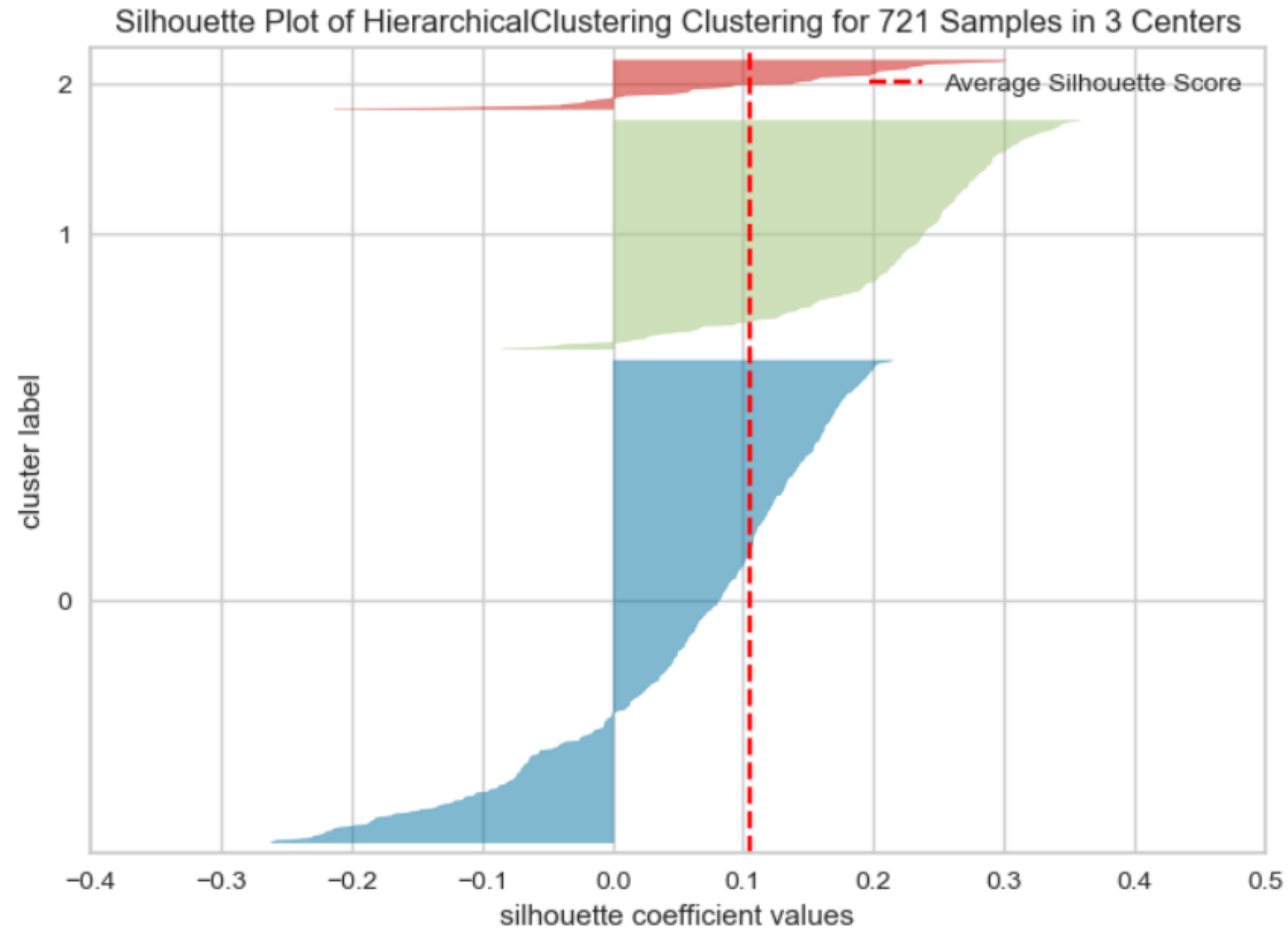
# DBSCAN eps tuning



# Hierarchical clustering

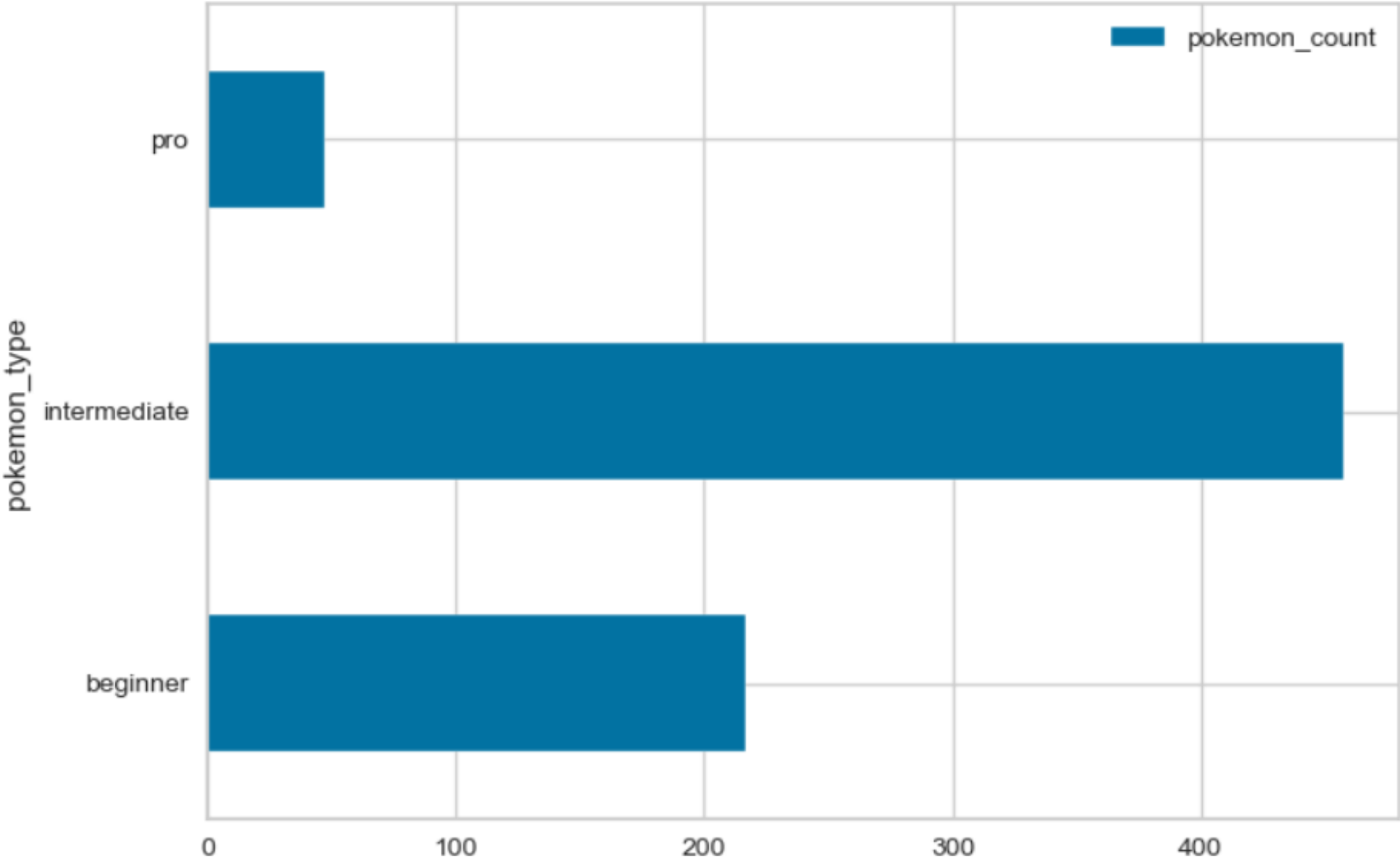


# Silhouette plot

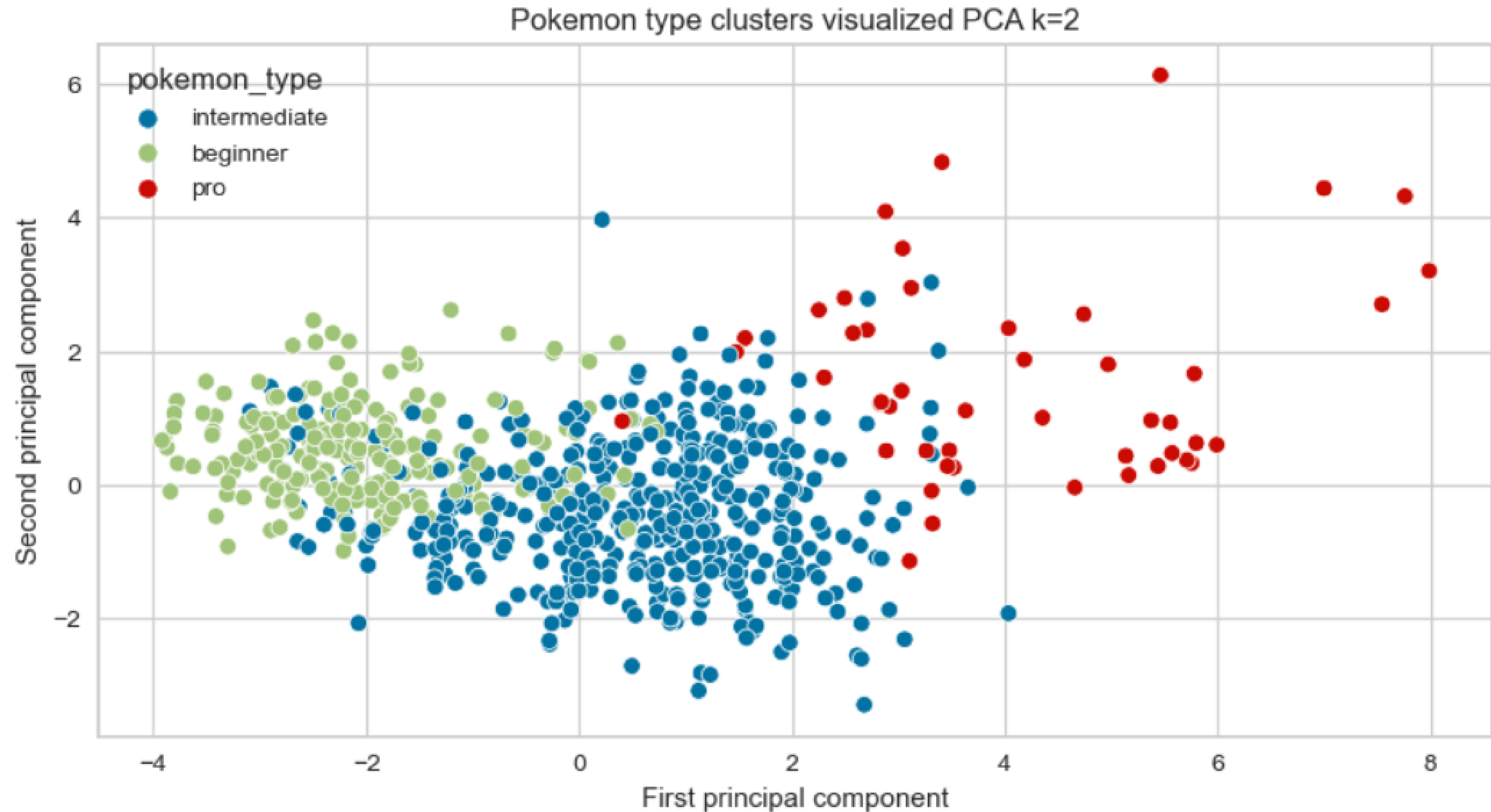


# Clustering results

	Number	Total	HP	Attack	Defense	Sp_Atk	Sp_Def	Speed	Pr_Male	hasMegaEvolution	Height_m	Weight_kg	Catch_
label													
0	354.969365	454.124726	73.129103	79.730853	73.986871	77.234136	76.485777	73.557987	0.574945	0.080963	1.144289	47.991028	68.86
1	356.686636	307.170507	50.645161	56.990783	56.552995	46.124424	49.539171	47.317972	0.490783	0.018433	0.616083	19.541935	181.45
2	439.553191	577.617021	104.085106	112.361702	105.723404	90.531915	90.531915	74.382979	0.545213	0.106383	3.593617	314.065957	30.44



# PCA clusters

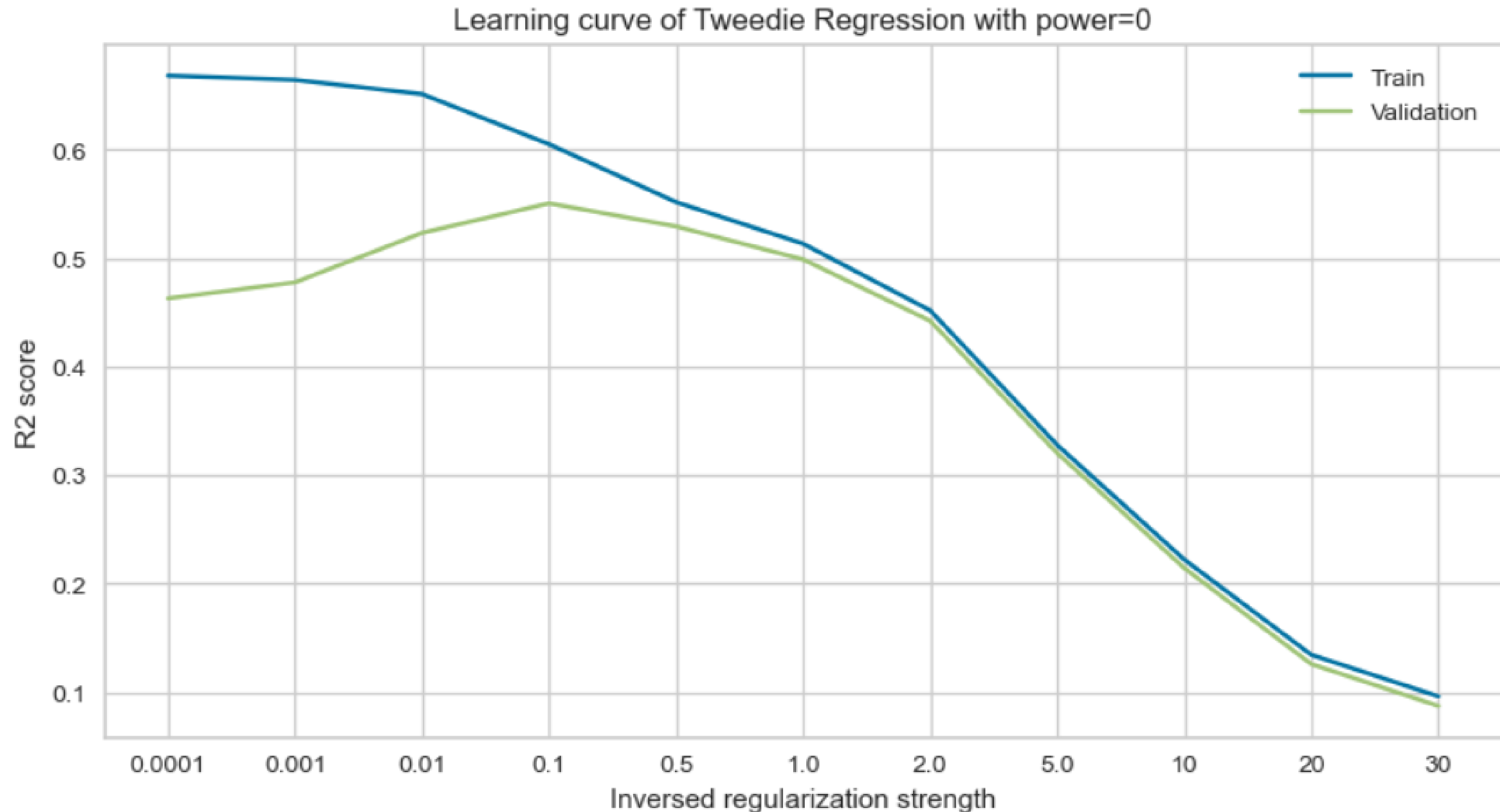




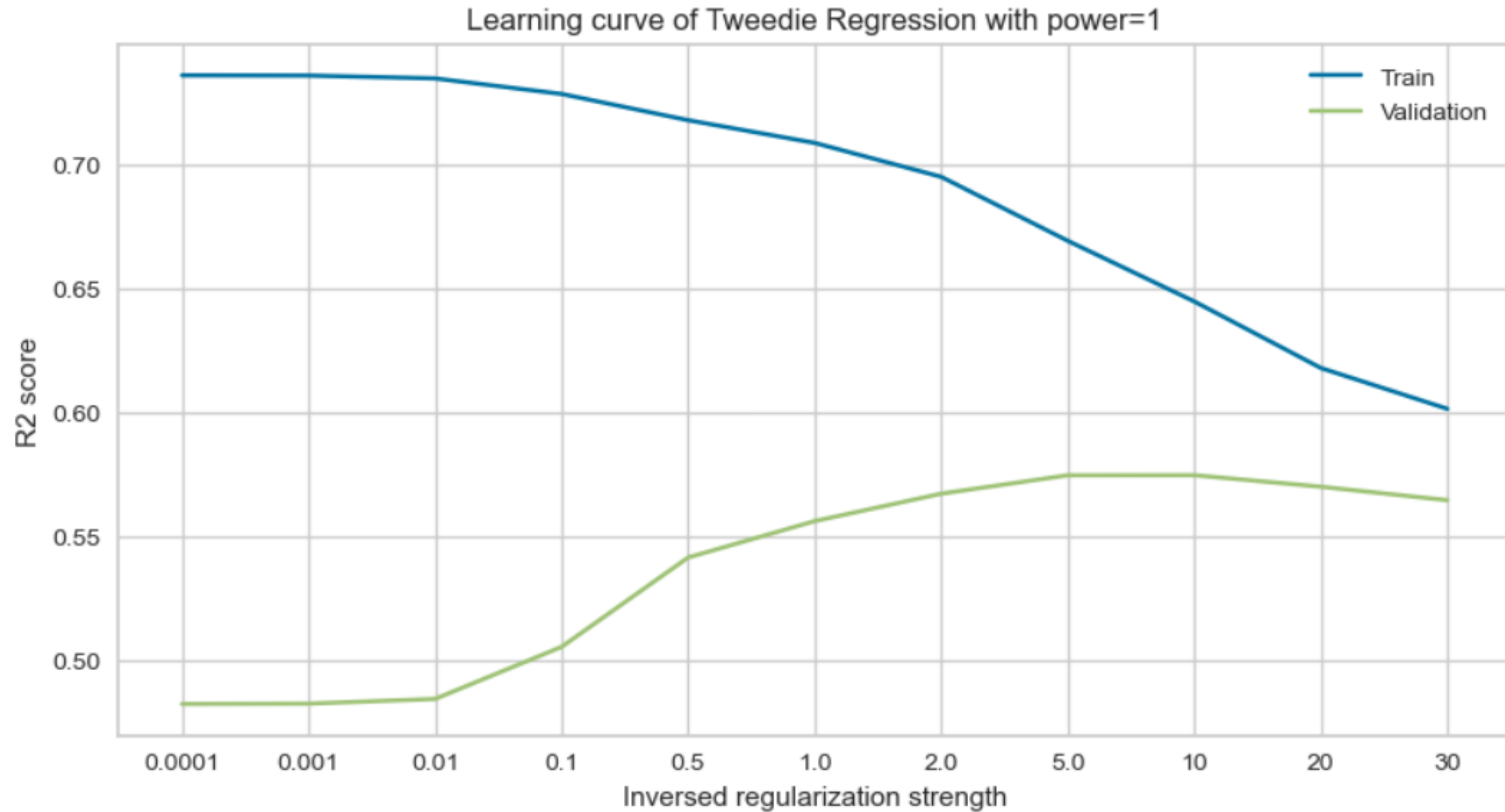
# Generalized Linear Models



# Normal target assumption



# Poisson target assumption



# Gamma target assumption

