CIC: School of Witchcraft And Wizardry

Which CIC House Do You Belong In?

Emanuel Becerra Soto

Octubre, 2019

Introduction

In this document the relationships between the results of the internet based quiz: *What is your Hogwarts house percentage?* would be explored. Initially an exploratory data analysis would be performed. Next, an attempt to develop a methodology to assign people to their corresponding Hogwarts Houses will be made.

Two perspectives would be analyzed, should a new student be assigned to the House to which her has the most affinity or the one with more people like her.

The data comes from a cohort of 39 students from the Computer Research Center of the National Polytechnic Institute (CIC-IPN).

The complete quiz could be found at: https://www.buzzfeed.com/eleanorbate/accurate-af-sorting-quiz

---

Gryffindor *Values courage, bravery, nerve, and chivalry. Gryffindor's mascot is the lion, and its colors are scarlet and gold.*

Hufflepuff *Values hard work, patience, justice, and loyalty. The house mascot is the badger, and canary yellow and black are its colors.*

Ravenclaw *Values intelligence, creativity, learning, and wit. The house mascot is an eagle and the house colours are blue and bronze (blue and grey in the films).*

Slytherin *Values ambition, cunning, leadership, and resourcefulness. The house mascot of Slytherin is the serpent, and the house colours are green and silver.*

The Sorting Hat *Is a sapient artefact used at Hogwarts, which uses Legilimency (the ability to read minds) to determine which of the four school houses (Gryffindor, Hufflepuff, Ravenclaw or Slytherin) each new student is to be assigned for their years at Hogwarts. The hat resembles a dilapidated conical leather wide-brimmed wizard's hat, with folds and tears that make it appear to have eyes and a mouth. During the opening banquet at the beginning of each school year, the Hat is placed on every first-year student's head. The Hat announces its choice aloud, and the student joins the selected house. The Hat speaks to the student while they're being sorted and is willing to take the student's preferences into account when it makes its decision, still sometimes it does not have the need to do so.*

Summary Statistics

Data

Table 1: People Preferences sorted by Gini Index

Nombre	Gini	Gryffindor	Slytherin	Ravenclaw	Hufflepuff
Ale	0.325	0.25	0.04	0.24	0.47
Joshua	0.320	0.23	0.43	0.31	0.03
Mike	0.270	0.21	0.49	0.15	0.15
Emanuel	0.270	0.21	0.27	0.43	0.09
Viri	0.255	0.48	0.15	0.20	0.17
Eli	0.250	0.41	0.27	0.23	0.09
LuisRicardo	0.235	0.42	0.27	0.17	0.14
Diana	0.230	0.39	0.13	0.31	0.17
Ismael	0.220	0.25	0.21	0.41	0.13
David	0.210	0.25	0.37	0.28	0.10
Mauro	0.210	0.23	0.35	0.32	0.10
Vita	0.210	0.36	0.09	0.29	0.26
Karina	0.200	0.42	0.20	0.22	0.16
Patty	0.195	0.23	0.23	0.40	0.14
Alexis	0.190	0.40	0.17	0.18	0.25
OmarH	0.190	0.32	0.32	0.28	0.08
Abraham	0.185	0.42	0.21	0.19	0.18
Hector	0.185	0.29	0.14	0.36	0.21
Jessica	0.180	0.34	0.24	0.30	0.12
Bruno	0.175	0.39	0.17	0.24	0.20
Johnny	0.175	0.34	0.30	0.23	0.13
Raul	0.170	0.18	0.19	0.39	0.24
Ray	0.160	0.22	0.17	0.38	0.23
DavidH	0.155	0.24	0.15	0.26	0.35
Kike	0.155	0.22	0.33	0.30	0.15
Miroslava	0.150	0.35	0.25	0.25	0.15
Franz	0.150	0.16	0.34	0.28	0.22
Itzel	0.145	0.32	0.15	0.30	0.23
LuisOctavio	0.140	0.30	0.14	0.31	0.25
Samantha	0.140	0.20	0.22	0.38	0.20
Dani	0.135	0.35	0.20	0.26	0.19
Andrea	0.130	0.31	0.14	0.27	0.28
Carlos	0.120	0.33	0.18	0.26	0.23
HectorCar	0.120	0.26	0.15	0.29	0.30
LuisAntonio	0.110	0.32	0.26	0.24	0.18
PamelaH	0.110	0.18	0.26	0.32	0.24
Byron	0.100	0.30	0.30	0.20	0.20
KarlaH	0.085	0.26	0.18	0.29	0.27
Isaac	0.060	0.31	0.23	0.23	0.23

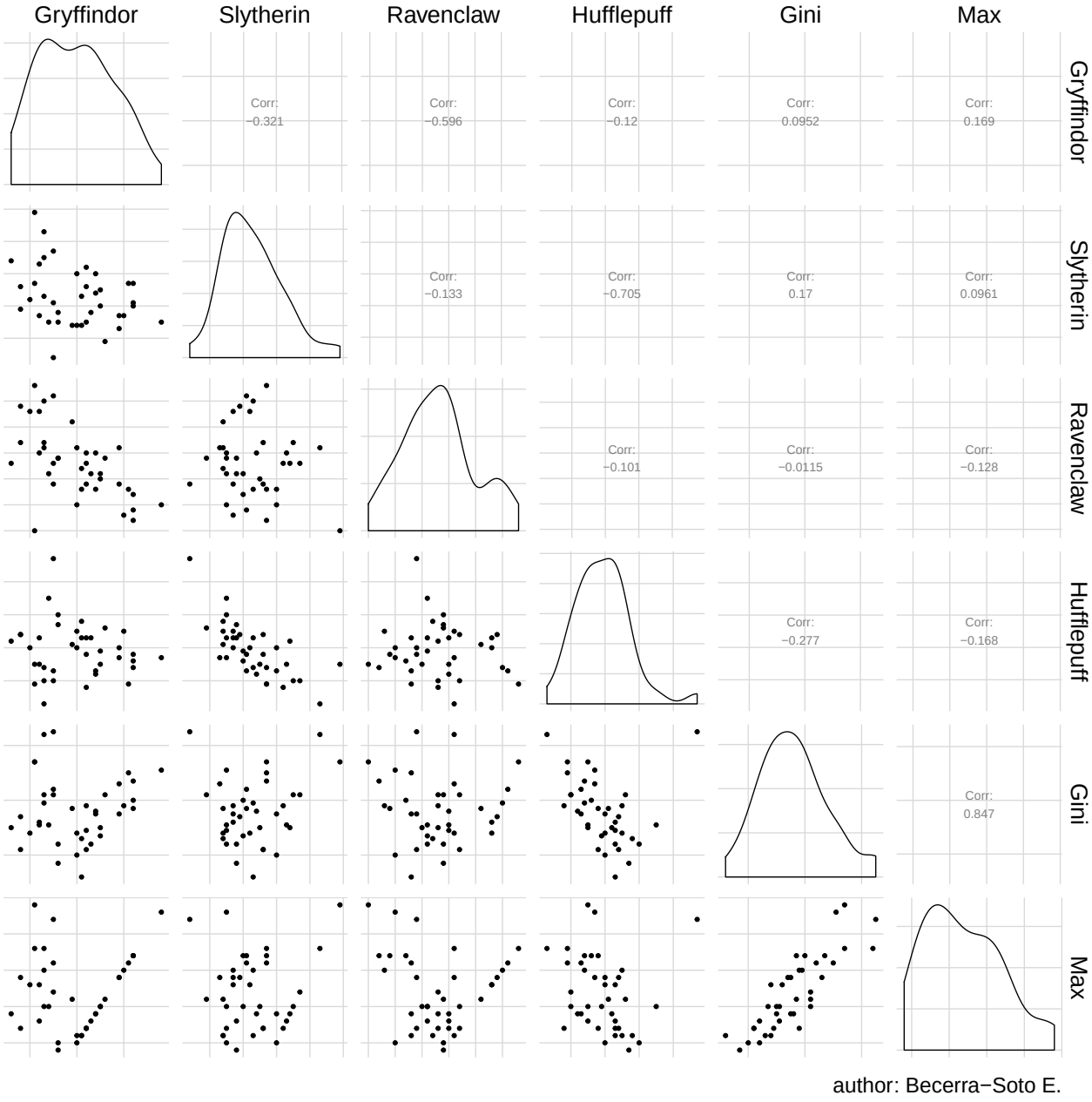
Summary Table

variable	mean	sd	minimum	q1	med	q3	maximum

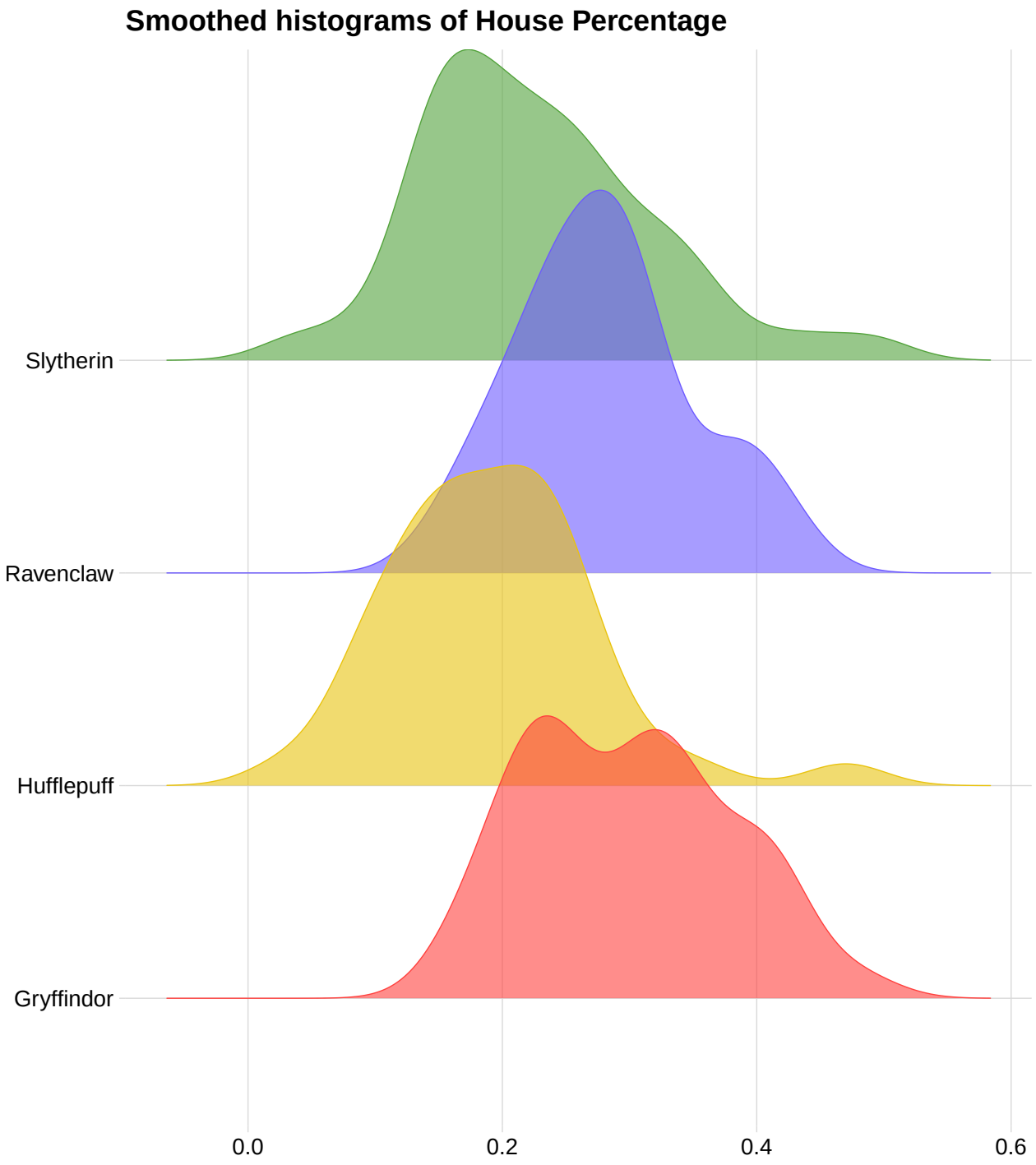
Table 2: Summary Statistics

variable	mean	sd	minimum	q1	med	q3	maximum
Gryffindor	0.299	0.080	0.16	0.230	0.300	0.350	0.480
Slytherin	0.228	0.092	0.04	0.160	0.210	0.270	0.490
Ravenclaw	0.281	0.068	0.15	0.235	0.280	0.310	0.430
Hufflepuff	0.193	0.082	0.03	0.140	0.190	0.235	0.470
Gini	0.180	0.060	0.06	0.140	0.175	0.210	0.325
Max	0.369	0.052	0.29	0.325	0.360	0.405	0.490

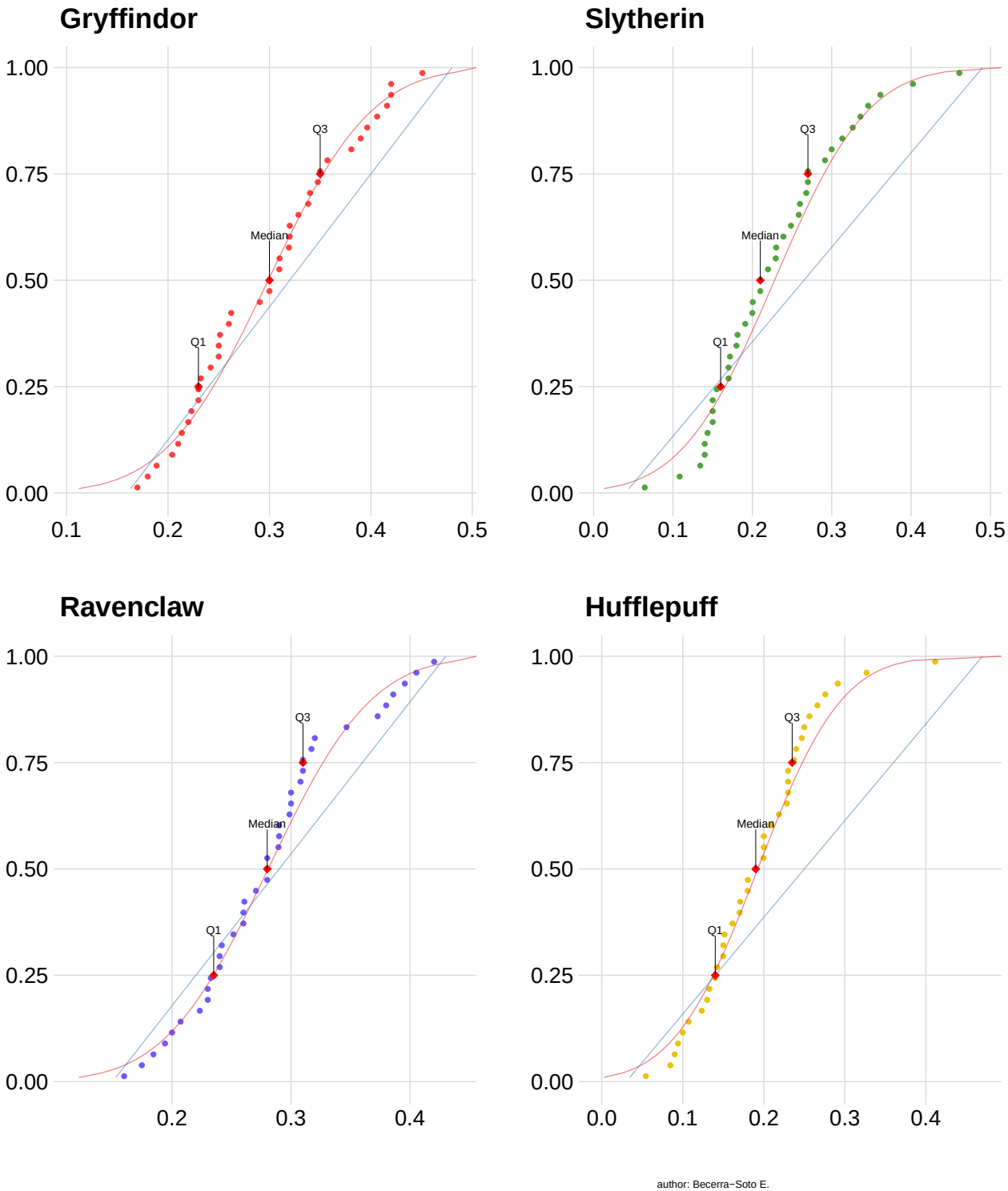
Scatter Plot Matrix



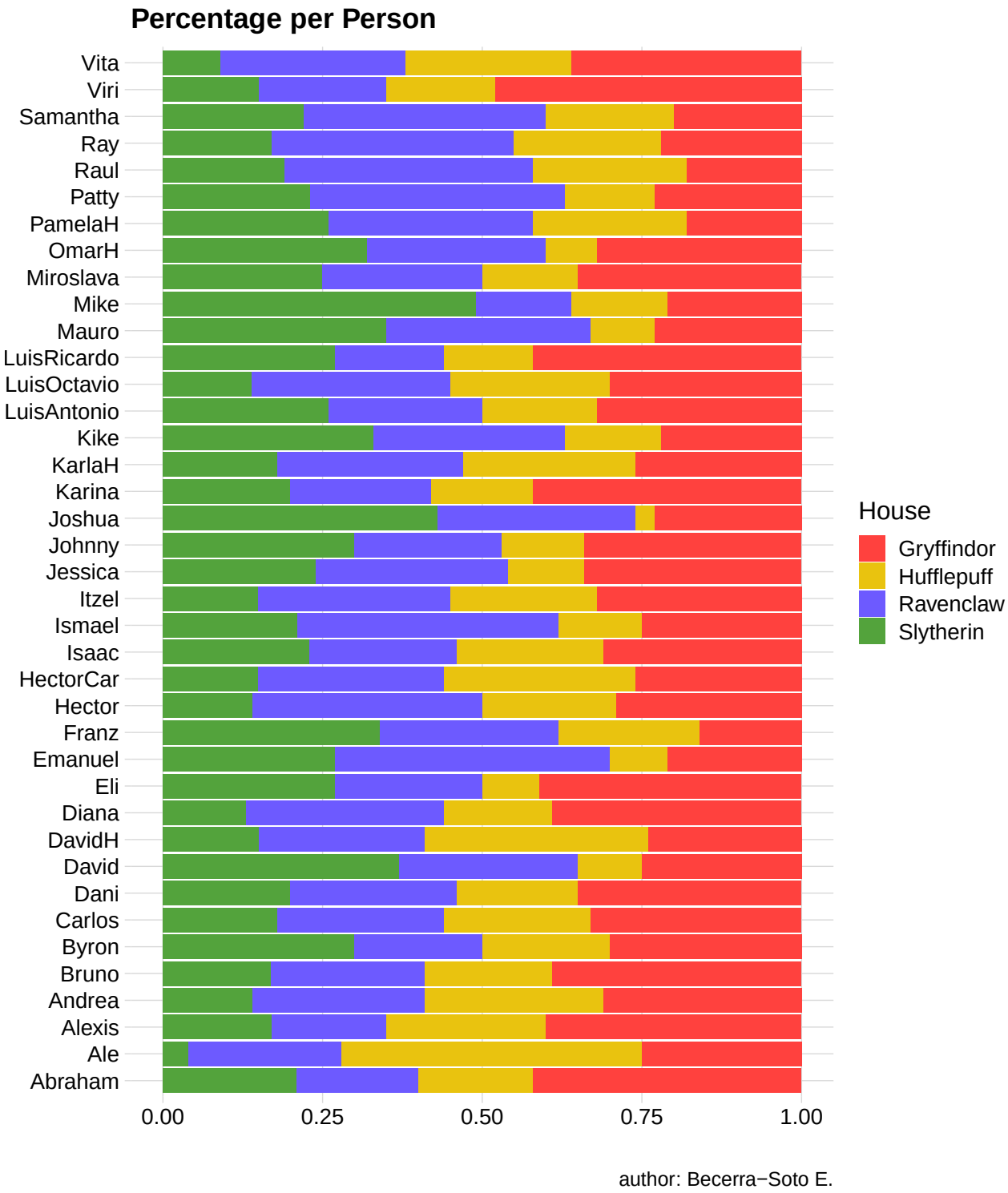
Smoothed histograms of House percentage



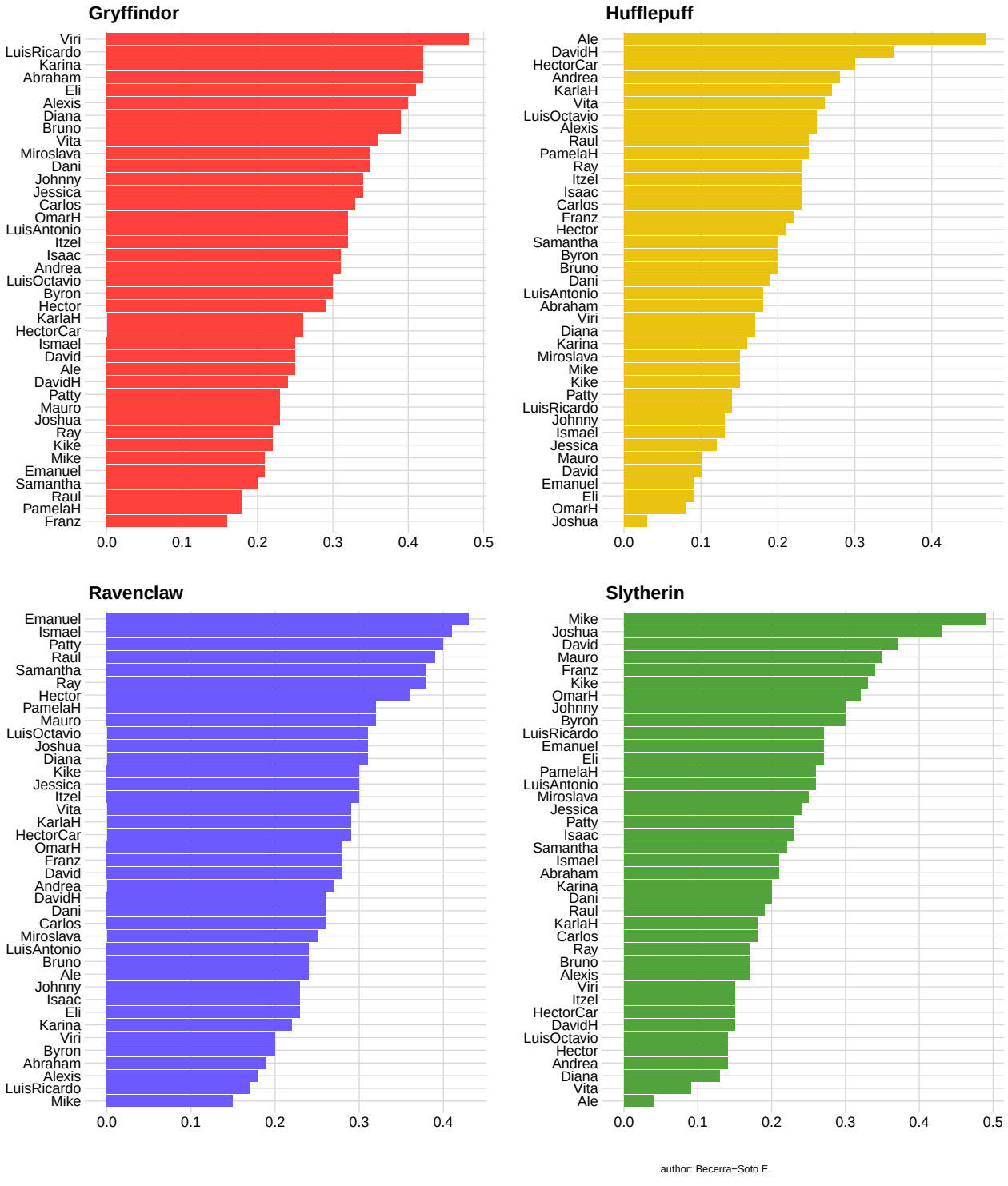
Empirical Cumulative Distributions



Bar Plot by Person



Bar Plot by House



author: Becerra−Soto E.

How to assign Houses?

The point of the quiz is to choose a House for the person taking it.
The output of the quiz are four different numbers that are to be interpreted as (I believe) the percentage of the answers that agree to each House. In particular, for person x, 35%, 12%, 24%, 29% could be a valid output.

The quiz assigns a House just by taking the one with the maximum percentage. So a natural question arises, is this a sensible method to do it?

It seems like so, but also the House's Percentages could be not far away from each other making the selection non-obvious. Not to mention that by taking the maximum percentage we are neglecting all the other information provided by the remaining categories.

So first let's find out if the selection criteria used by the quiz does relatively well. To do that we are going to label the data and plot it using dimensional reduction techniques, to "look-up" for patterns, this is by no means a formal technique as we are just assessing by "eye" our data.

In order to perform something more elaborated we would need to define evaluation metrics, but this is difficult as we don't know the true Houses for the people, as we, sadly, have never been in Hogwarts School or the Wizarding World.

The dimensional reduction techniques that would be used are: T-distributed Stochastic Neighbor Embedding (t-SNE) and Principal Component Analysis (PCA).

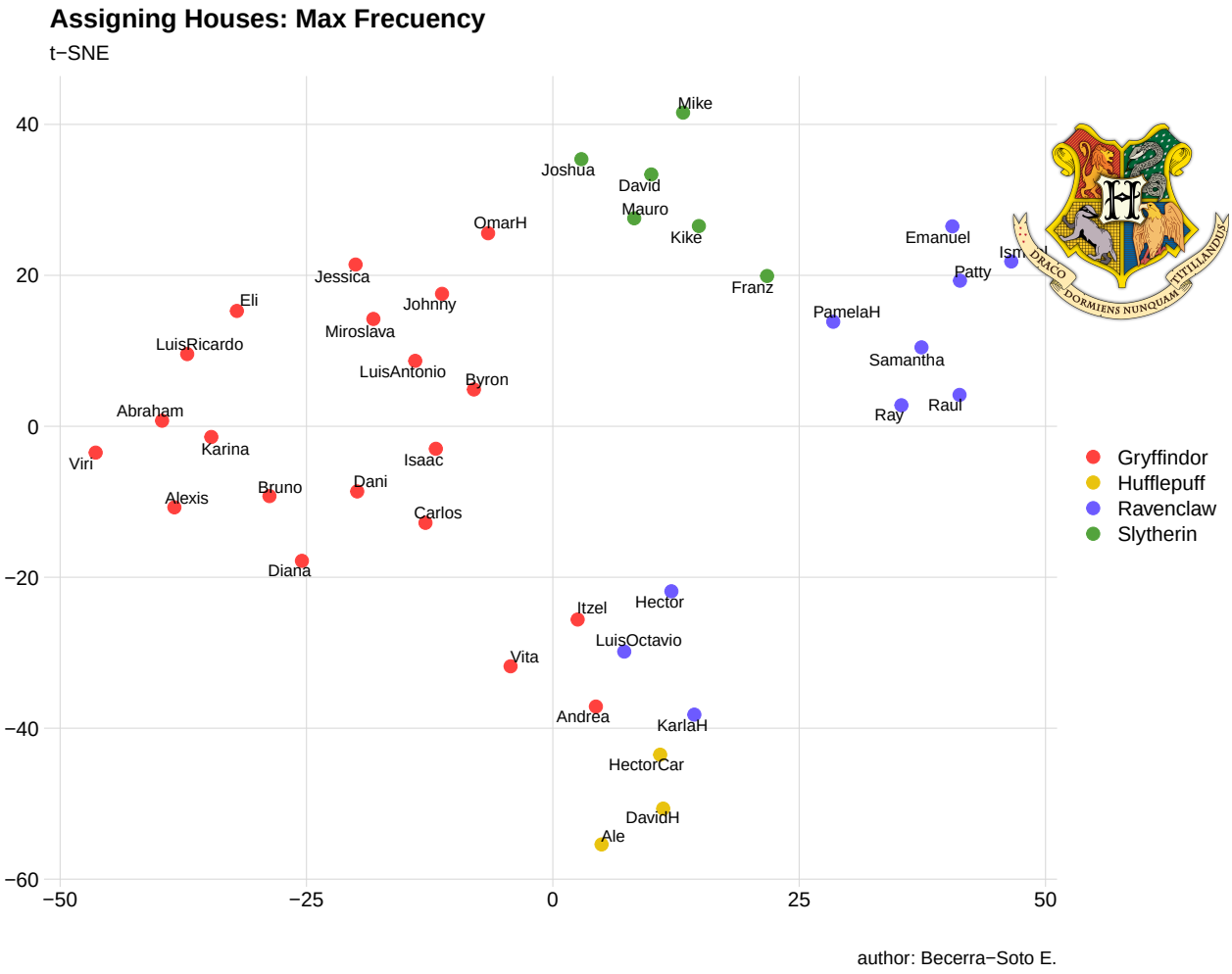
Basically four dimensions (one per house) would be represented just by two, in favor of data visualization and spotting patterns by "eye".

Direct Approach

In this approach we just take the maximum percentage and assign that House to that person.

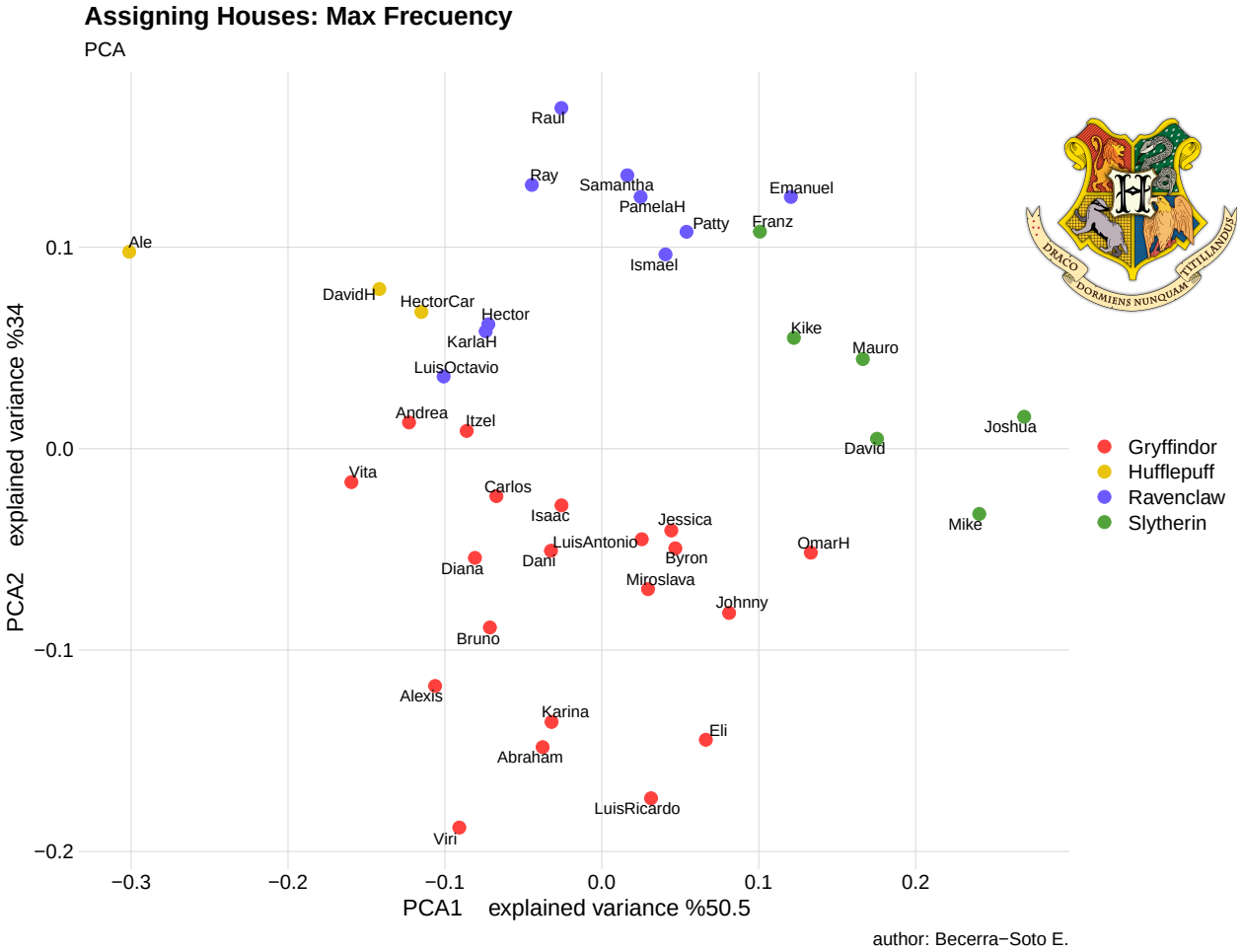
Direct Approach t-SNE

Plotting each person onto the new t-SNE axes gives the following plot:



Direct Approach PCA

Plotting each person onto the new PCA axes gives the following plot:



It seems that the Direct Approach meets the "eye test" as labels with the same House lie close to each other. Let us try something else.

Purity Approach (Semi-supervised)

One of the criticism of the Direct Approach was that it takes the House with the maximum percentage even if there are others with a close percentage to the winner case. For instance one extreme case is that each House gets 33%. How to determine a House on such a case?

One idea of assigning Houses is as follows:

1. Assume that each person has a variable component of how appropriate is to be assigned onto a given House. So there isn't a definitive correct House to belong to.
2. Assume that we don't know the true components of each person, but the test is a good proxy for it.
3. Somehow calculating the "purity" of each person, where purity is seen as a score of how much a given person has a strong component in one House and low components in the remaining ones. For example a person with 100% in one house and 0% in the others should have a huge purity score.
4. Then establish a purity threshold and assign the House with the maximum percentage to the persons that meet the threshold. So only the most "pure" people get a House, because there are the ones with the clearer signal on which House to belong to. For the sake of clarity let us call these people "seeds".
5. Run any clustering algorithm over the data and then start labeling non-seed people. An unlabeled person gets the label of the most common House (seeds labels) in their same cluster. In case of a tie leave the given person unlabeled. In case of a cluster without any seed, leave everyone in the cluster unlabeled.
6. Furthermore a label could be forced onto all persons if we iteratively start reducing the number of clusters, as when the number of clusters is only one it is guaranteed that any data point is in the same cluster as a seed or a previously labeled person. So we iteratively start reducing the number of clusters and taking a vote at each iteration assigning to any unlabeled point the most common House within the cluster. Then stop when all the points have been labeled. Breaking ties arbitrarily when the number of clusters is one (the last iteration).

So, to implement the former idea of assigning houses it is necessary to pinpoint some details. First a purity score needs to be chosen, a threshold established and a clustering algorithm selected.

For our luck the Gini Index used to measure income inequality in economics meets the criteria for a purity score. The Gini Index is 0 at equality between features and 1 at maximum inequality. As we are interested in maximum inequality, for our case maximum purity, the highest the Gini Index the highest the purity.

For a threshold, we want to choose one that retrieves the top of the observed purity distribution (Gini). So we are going to take some k standard deviations from the mean as a threshold. Empirically, as we only have a few data points, we notice that k = 0.8 works well.

As we only have a few data points the hierarchical clustering techniques seem appropriate. For the sake of comparison we are going to use two different clustering algorithms, that assume relatively different things.

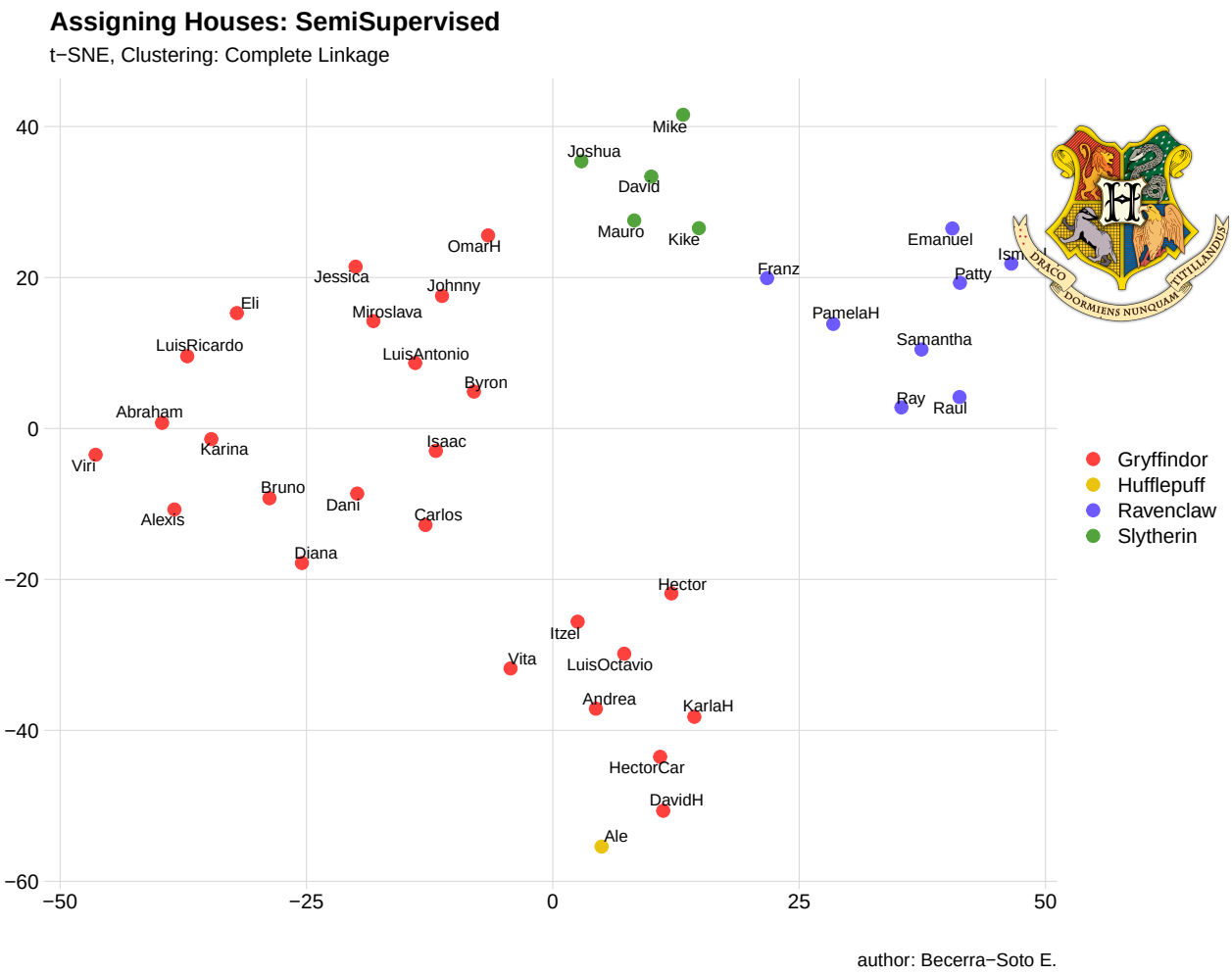
- The complete linkage method finds similar clusters. This method is also called the diameter or maximum method. In this method, we consider similarity of the furthest pair. That is, the distance between one cluster and another cluster is taken to be equal to the longest distance from any member of one cluster to any member of the other cluster. It tends to produce more compact clusters. One drawback of this method is that outliers can cause close groups to be merged later than what is optimal.
- Ward's minimum variance method aims at finding compact, spherical clusters. Ward's method aims to minimize the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged. In other words, it forms clusters in a manner that minimizes the loss associated with each cluster. At each step, the union of every possible cluster pair is considered and the two clusters whose merger results in minimum increase in information loss are combined. Here, information loss is defined by Ward in terms of an error sum-of-squares criterion (ESS).

So let us implement the previously stated idea.

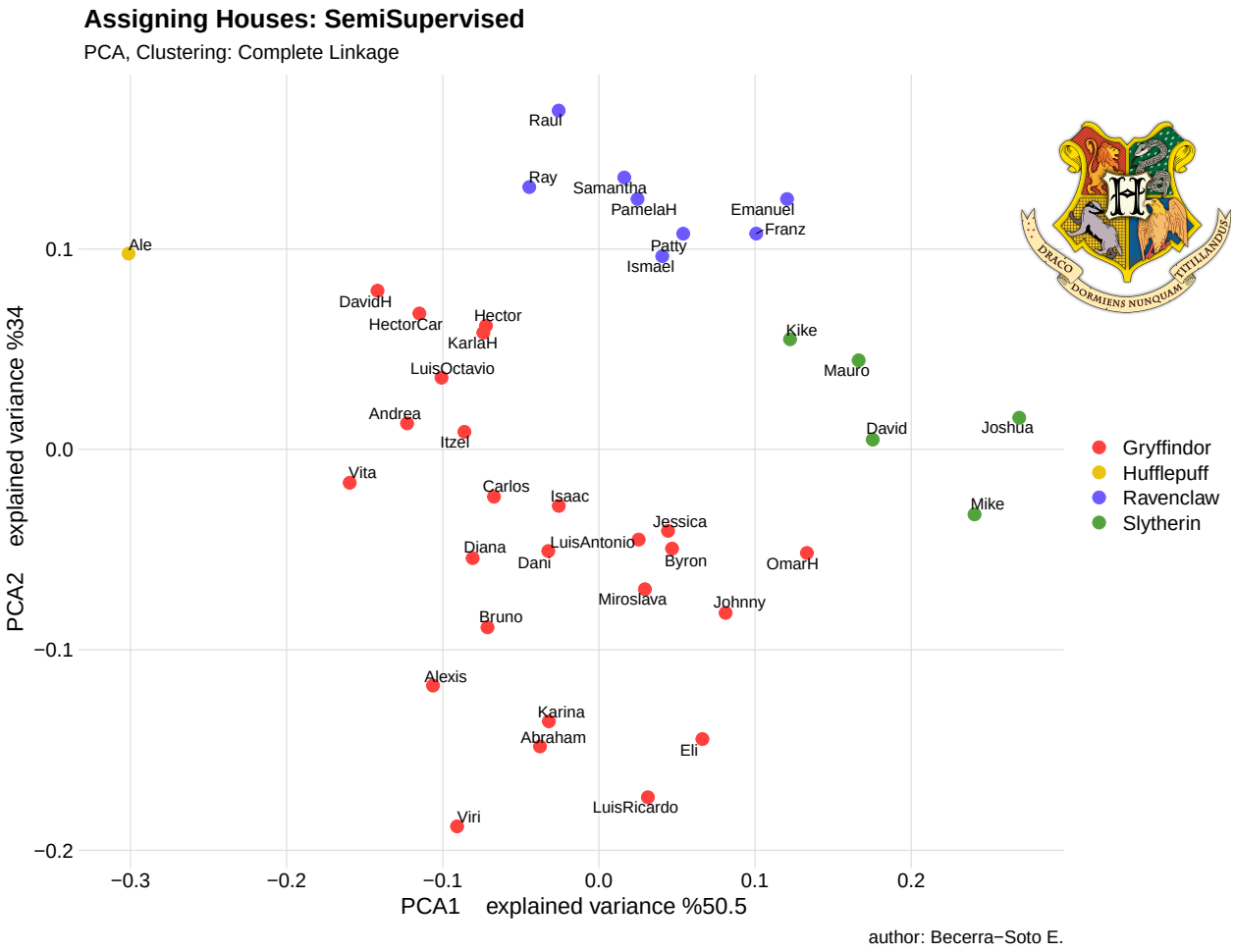
Complete Linkage Clustering

Using Complete Linkage Clustering we get the following results:

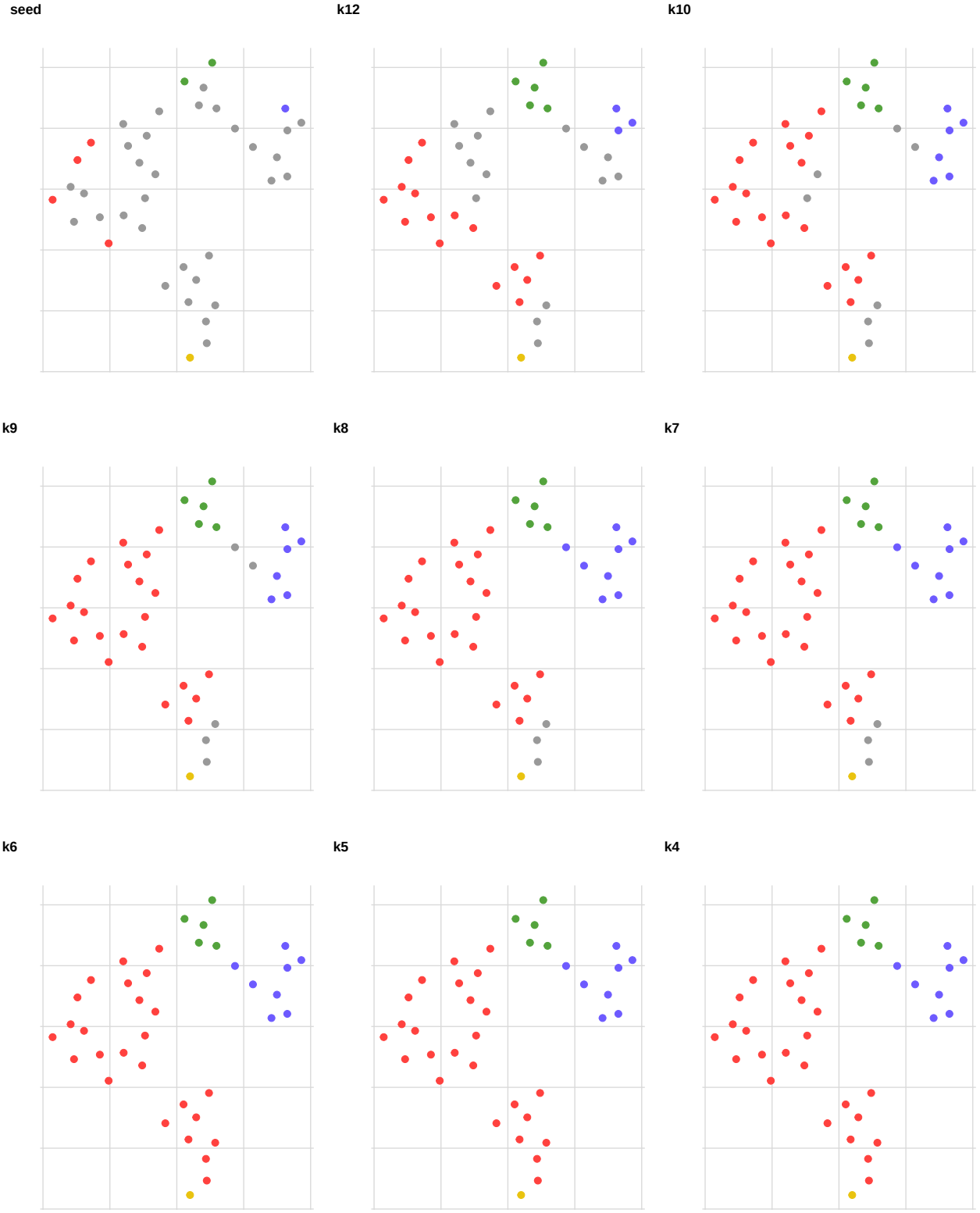
Complete Linkage t-SNE



Complete Linkage PCA



Epochs Complete Linkage t-SNE

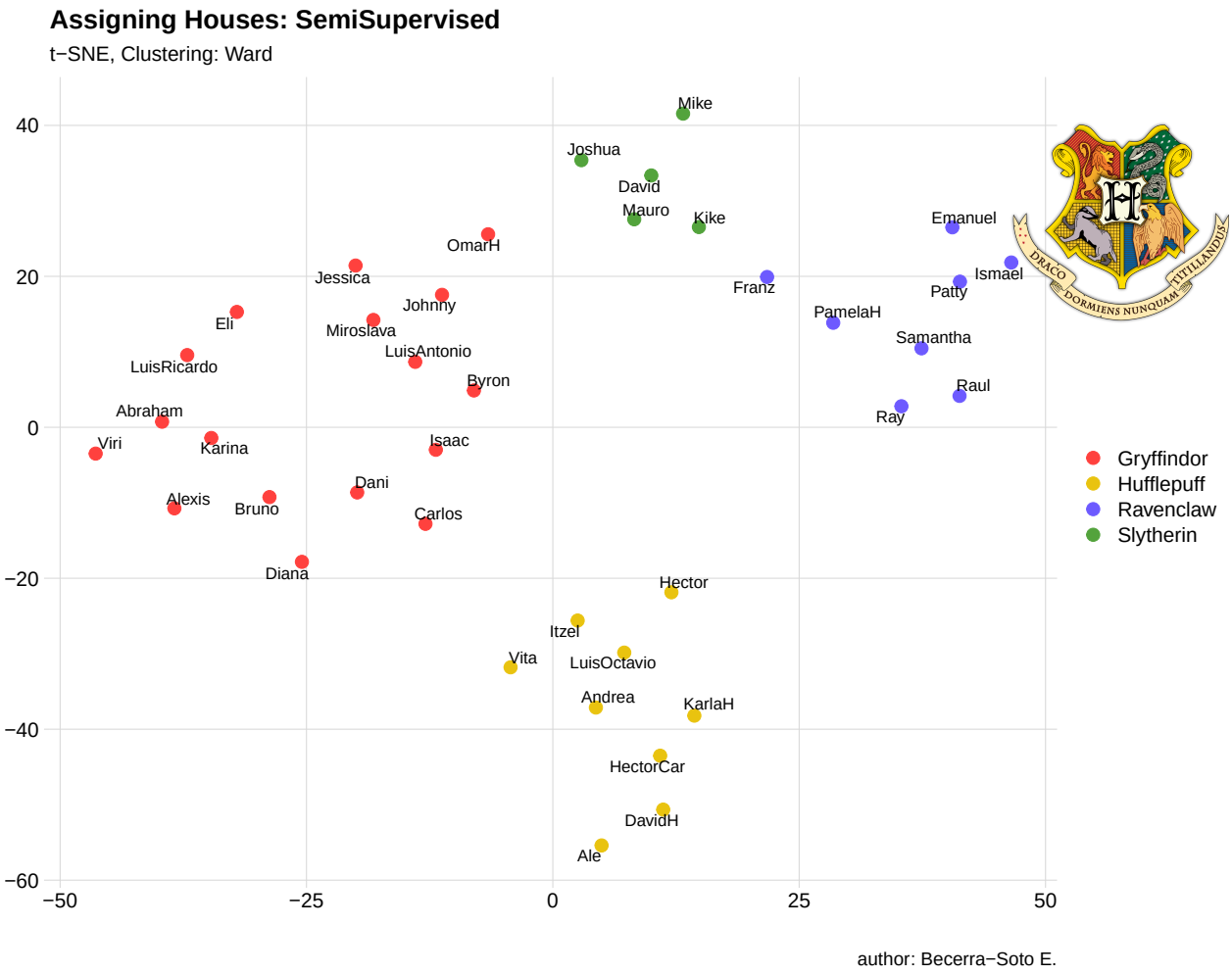


author: Becerra−Soto E.

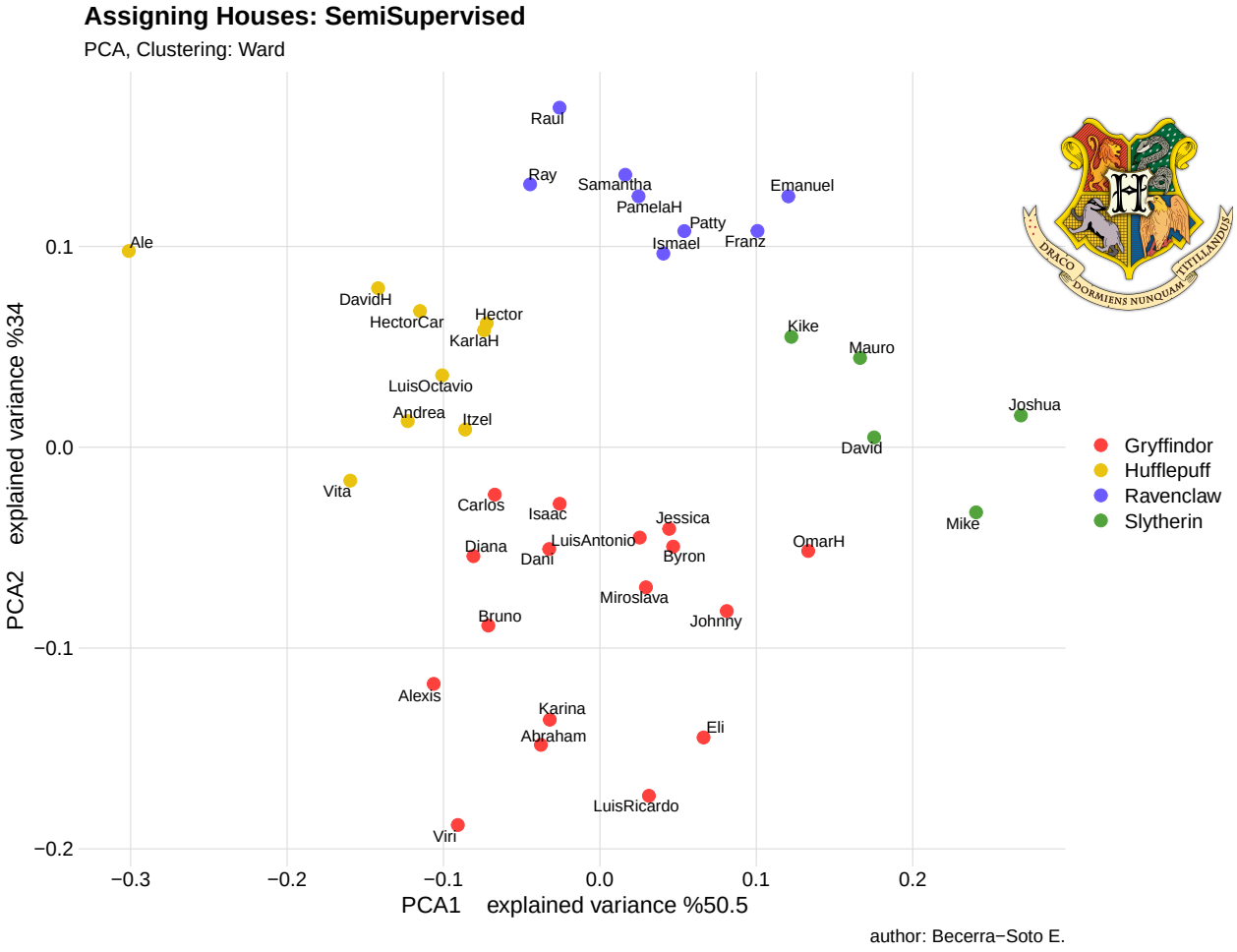
Ward Clustering

Using Ward Clustering we get the following results:

Ward t-SNE

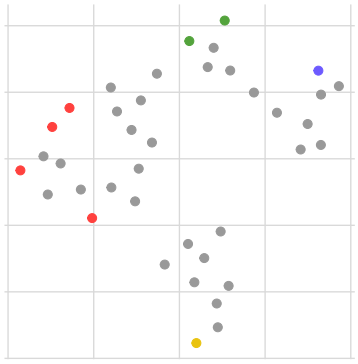


Ward PCA

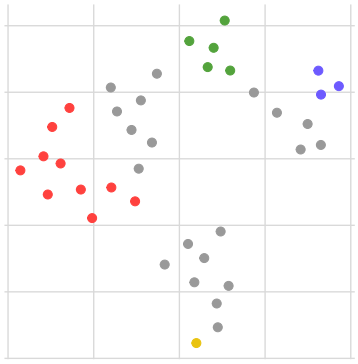

Assigning Houses: SemiSupervised
PCA, Clustering: Ward
Raul
Ray
Samantha
Emanuel
PamelaH
Patty
Ismael
Franz
Ale
DavidH
HectorCar
Hector
KarlaH
Kike
Mauro
LuisOctavio
Andrea
Itzel
Joshua
David
Vita
Carlos
Isaac
Jessica
Mike
Diana
LuisAntonio
Dani
Byron
OmarH
Miroslava
Bruno
Johnny
Alexis
Karina
Abraham
Eli
LuisRicardo
Viri
Gryffindor
Hufflepuff
Ravenclaw
Slytherin
PCA2 explained variance %34
PCA1 explained variance %50.5
0.1
0.0
−0.1
−0.2
−0.3
−0.2
−0.1
0.0
0.1
0.2
author: Becerra−Soto E.

Epochs Ward t-SNE

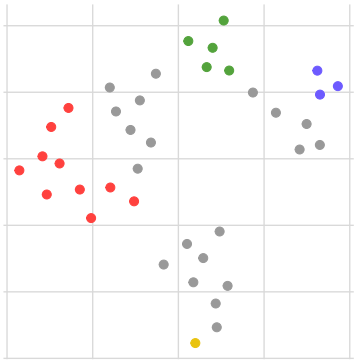
seed



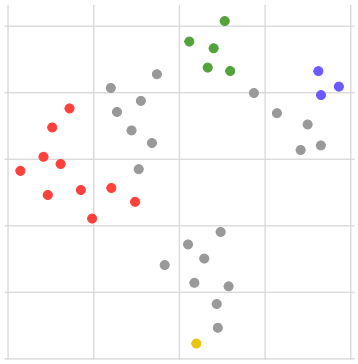
k12



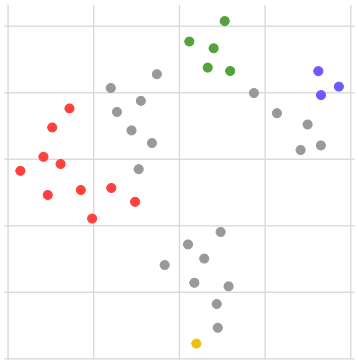
k10



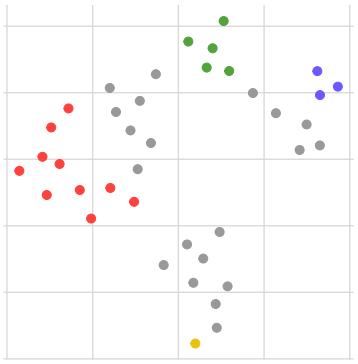
k9



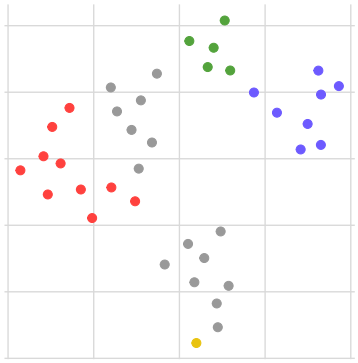
k8



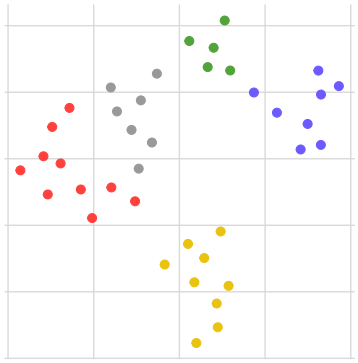
k7



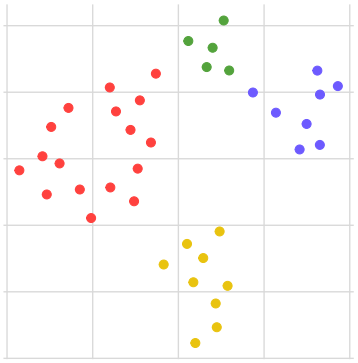
k6



k5

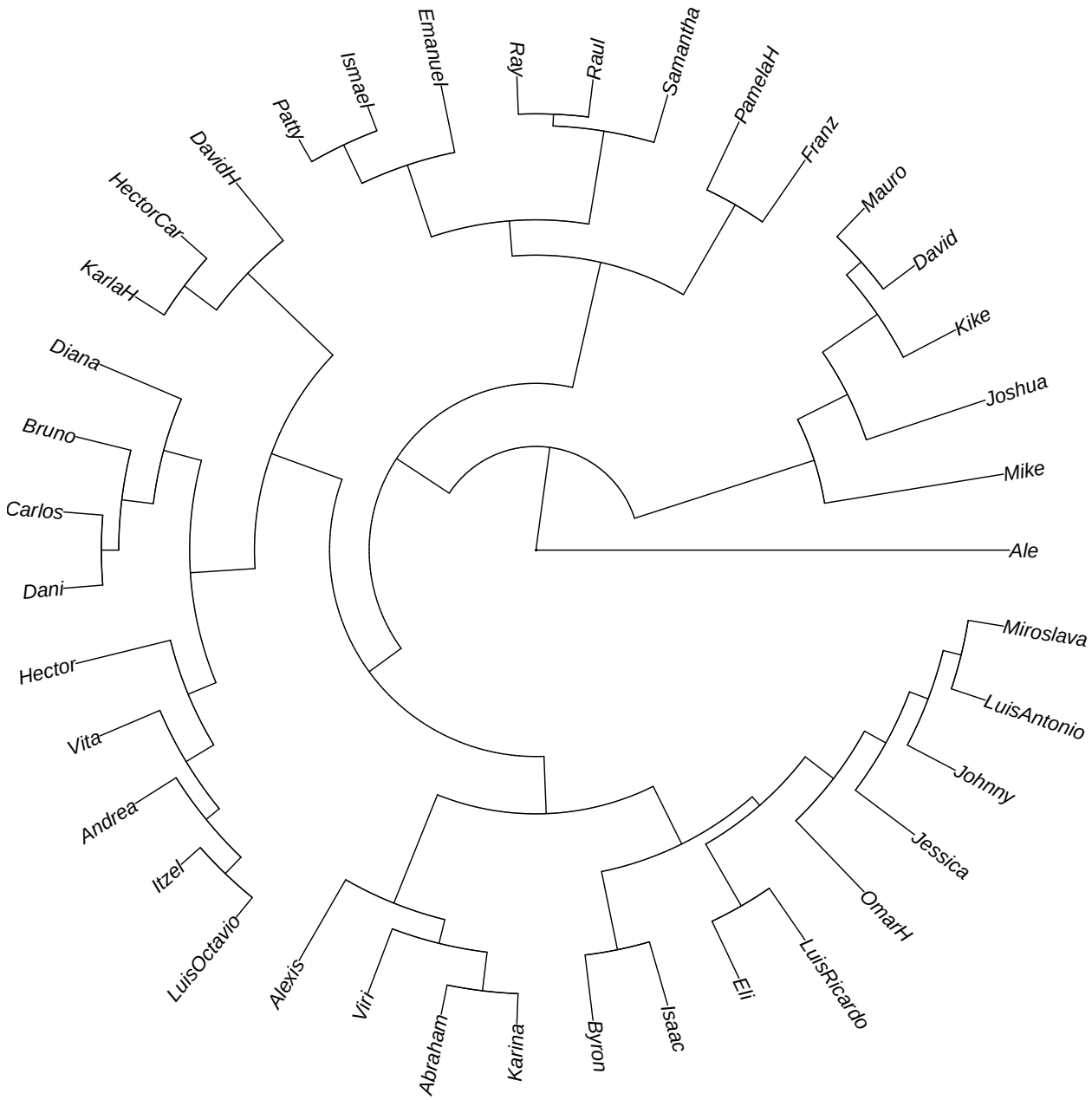


k4



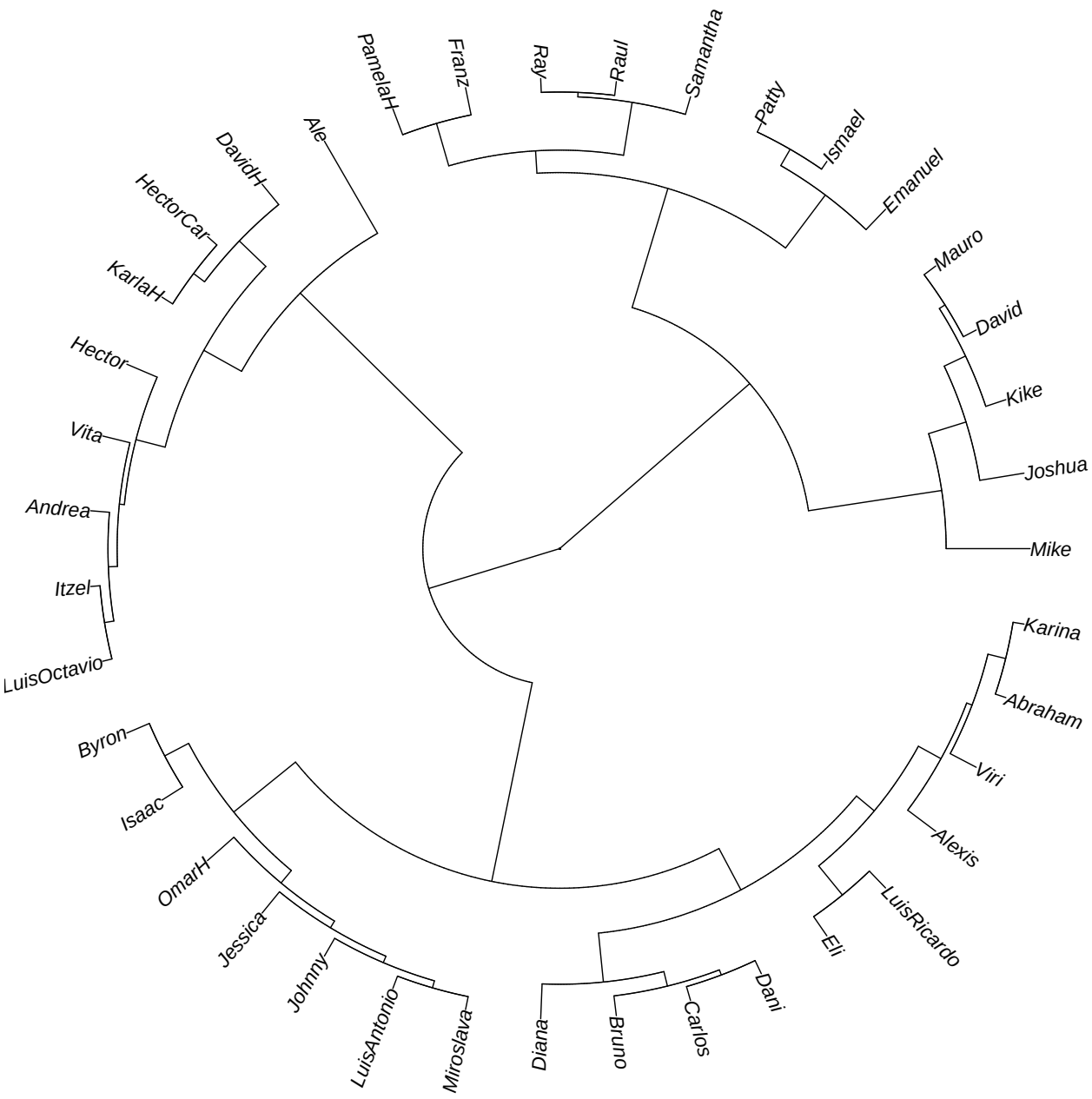
author: Becerra−Soto E.

Complete Linkage dendrogram



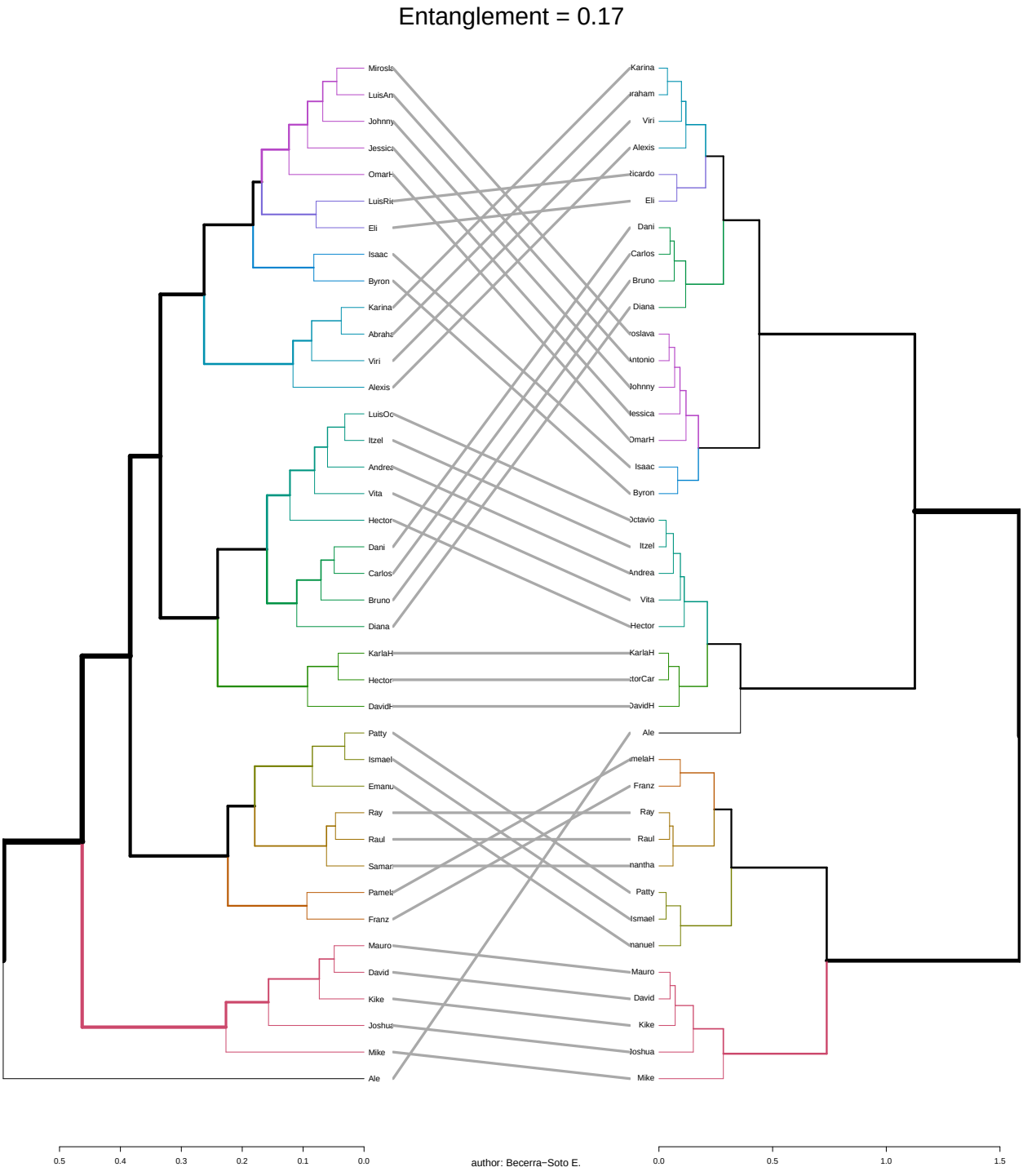
author: Becerra−Soto E.

Ward dendrogram



author: Becerra−Soto E.

Comparision of the two dendrograms



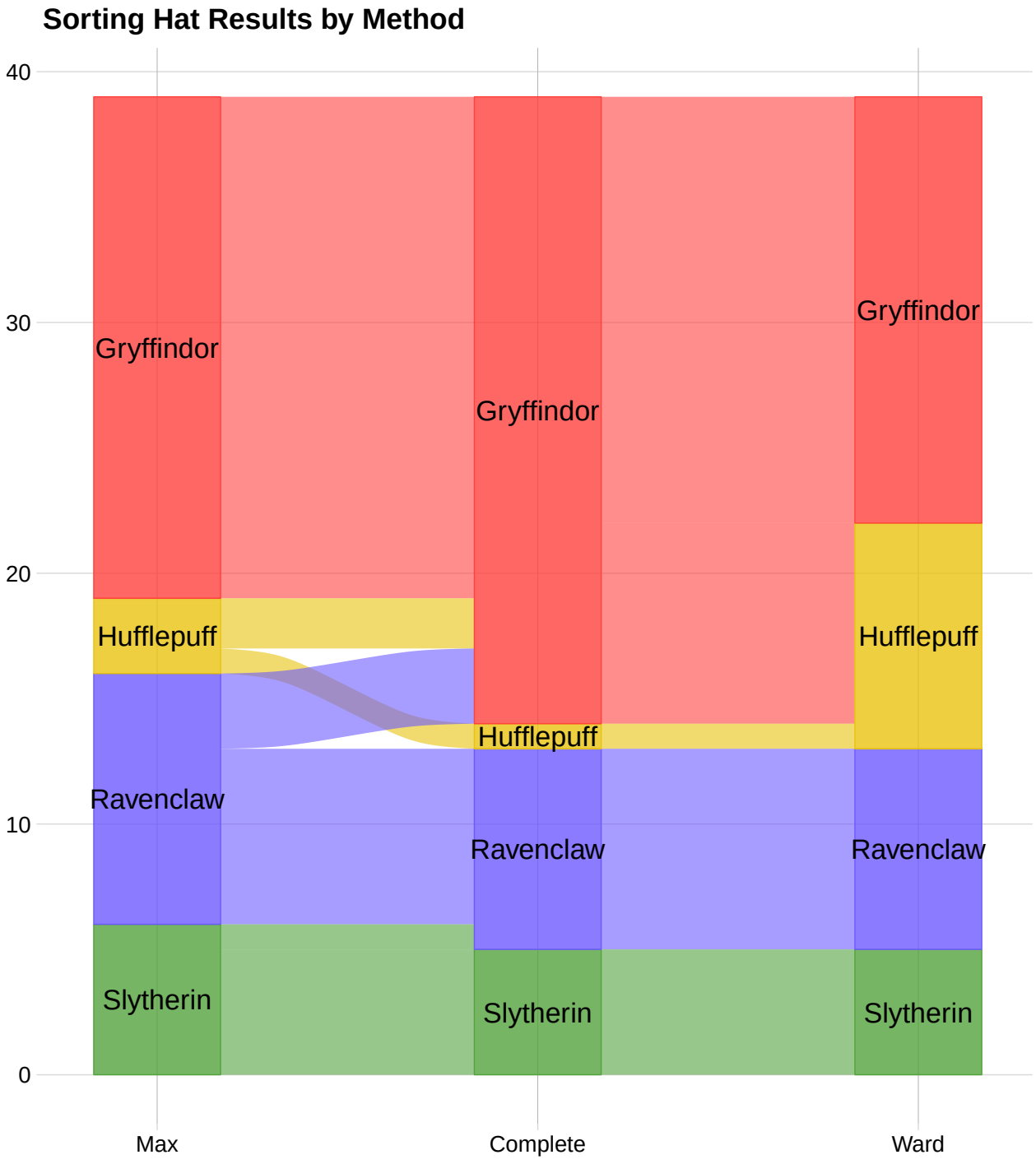
To the left is the complete linkage dendrogram and to the right is the ward dendrogram. Entanglement is a measure of dissimilarity between dendrograms, where 0 means that they are equal, up to 1 where they are totally different.

Results

Table 3: Different Assignation Methodologies

person	gini	max	complete	ward
Ale	0.325	Hufflepuff	Hufflepuff	Hufflepuff
Joshua	0.320	Slytherin	Slytherin	Slytherin
Mike	0.270	Slytherin	Slytherin	Slytherin
Emanuel	0.270	Ravenclaw	Ravenclaw	Ravenclaw
Viri	0.255	Gryffindor	Gryffindor	Gryffindor
Eli	0.250	Gryffindor	Gryffindor	Gryffindor
LuisRicardo	0.235	Gryffindor	Gryffindor	Gryffindor
Diana	0.230	Gryffindor	Gryffindor	Gryffindor
Ismael	0.220	Ravenclaw	Ravenclaw	Ravenclaw
David	0.210	Slytherin	Slytherin	Slytherin
Mauro	0.210	Slytherin	Slytherin	Slytherin
Vita	0.210	Gryffindor	Gryffindor	Hufflepuff
Karina	0.200	Gryffindor	Gryffindor	Gryffindor
Patty	0.195	Ravenclaw	Ravenclaw	Ravenclaw
Alexis	0.190	Gryffindor	Gryffindor	Gryffindor
OmarH	0.190	Gryffindor	Gryffindor	Gryffindor
Abraham	0.185	Gryffindor	Gryffindor	Gryffindor
Hector	0.185	Ravenclaw	Gryffindor	Hufflepuff
Jessica	0.180	Gryffindor	Gryffindor	Gryffindor
Bruno	0.175	Gryffindor	Gryffindor	Gryffindor
Johnny	0.175	Gryffindor	Gryffindor	Gryffindor
Raul	0.170	Ravenclaw	Ravenclaw	Ravenclaw
Ray	0.160	Ravenclaw	Ravenclaw	Ravenclaw
DavidH	0.155	Hufflepuff	Gryffindor	Hufflepuff
Kike	0.155	Slytherin	Slytherin	Slytherin
Miroslava	0.150	Gryffindor	Gryffindor	Gryffindor
Franz	0.150	Slytherin	Ravenclaw	Ravenclaw
Itzel	0.145	Gryffindor	Gryffindor	Hufflepuff
LuisOctavio	0.140	Ravenclaw	Gryffindor	Hufflepuff
Samantha	0.140	Ravenclaw	Ravenclaw	Ravenclaw
Dani	0.135	Gryffindor	Gryffindor	Gryffindor
Andrea	0.130	Gryffindor	Gryffindor	Hufflepuff
Carlos	0.120	Gryffindor	Gryffindor	Gryffindor
HectorCar	0.120	Hufflepuff	Gryffindor	Hufflepuff
LuisAntonio	0.110	Gryffindor	Gryffindor	Gryffindor
PamelaH	0.110	Ravenclaw	Ravenclaw	Ravenclaw
Byron	0.100	Gryffindor	Gryffindor	Gryffindor
KarlaH	0.085	Ravenclaw	Gryffindor	Hufflepuff
Isaac	0.060	Gryffindor	Gryffindor	Gryffindor

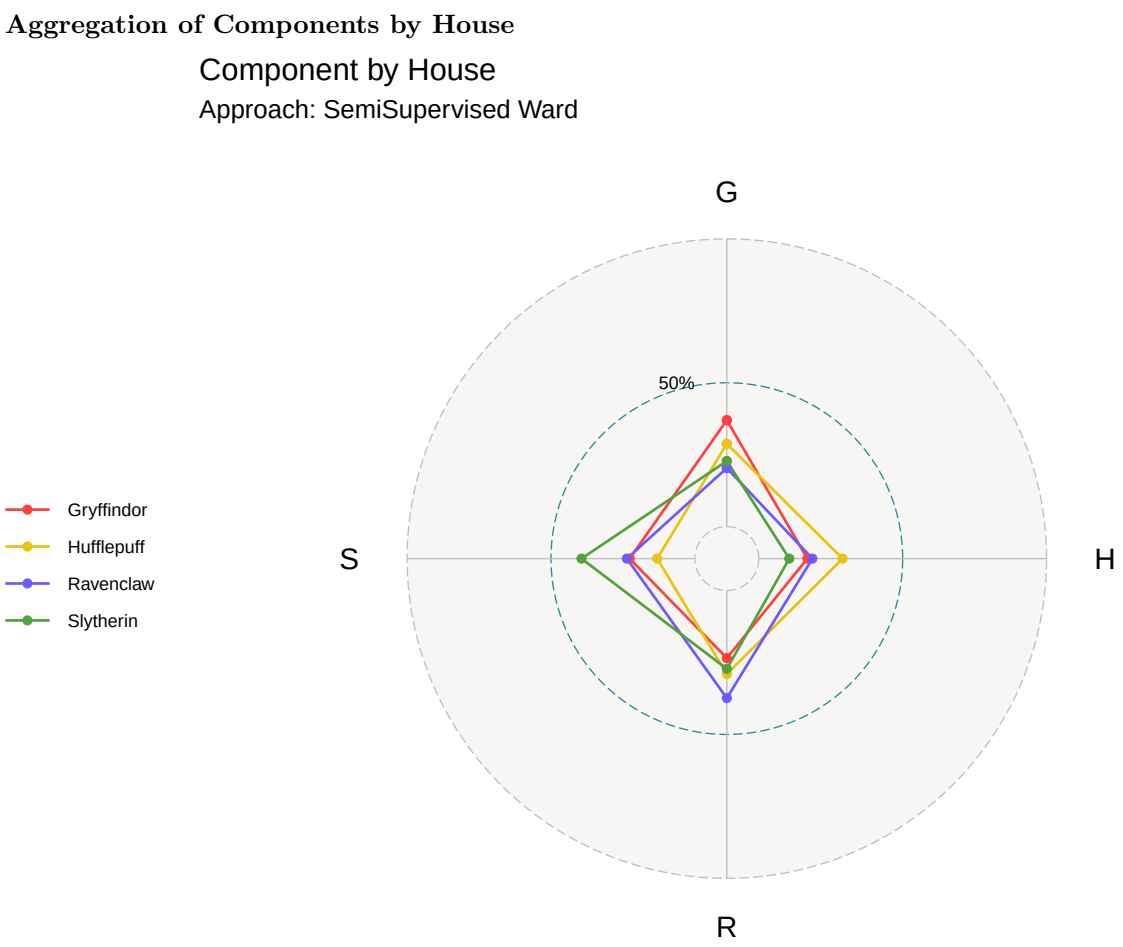
Alluvial Diagram of Results



author: Becerra−Soto E.

Plotting Results for Ward Clustering

Aggregation of Components by House

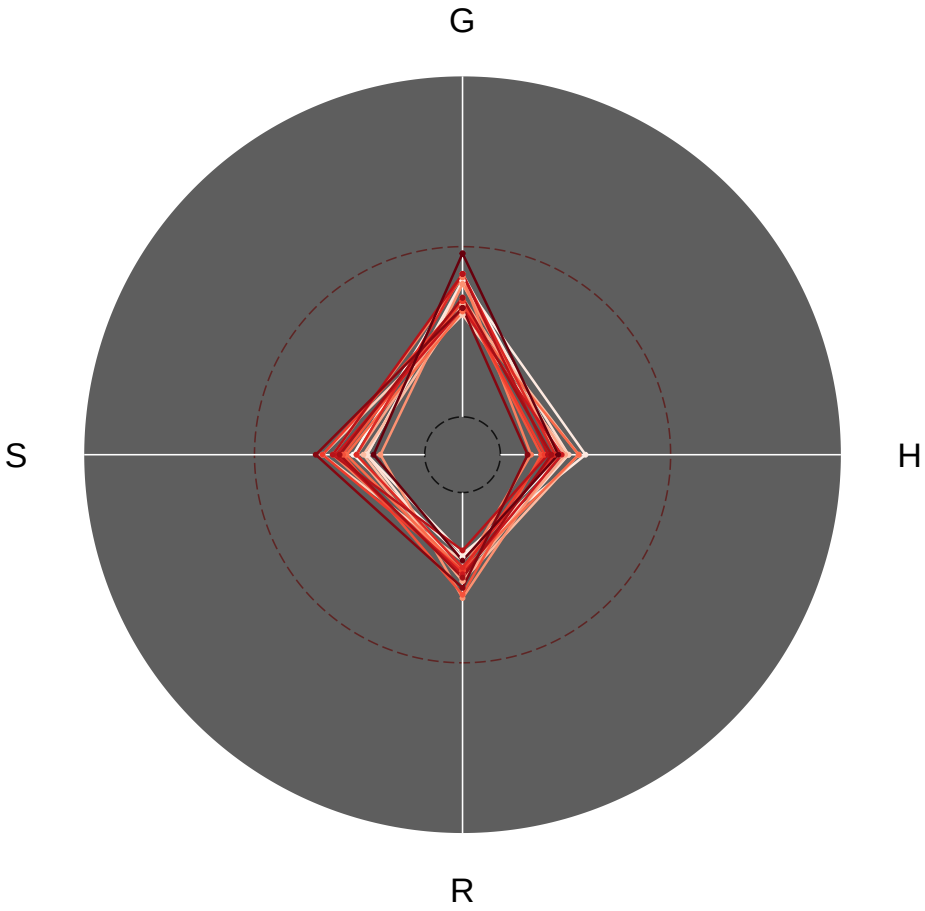


author: Becerra−Soto E.

Gryffindor Radar

Gryffindor's Students Profile

Approach: SemiSupervised Ward

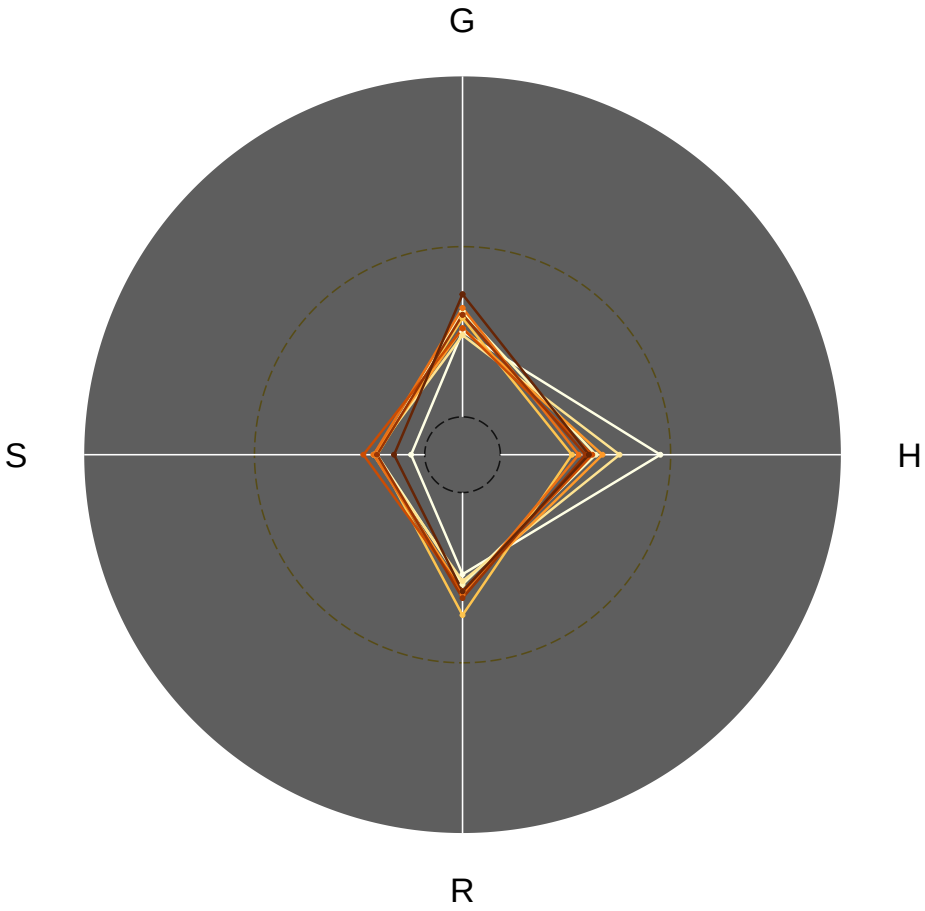


author: Becerra−Soto E.

Radar Hufflepuff

Hufflepuf's Students Profile

Approach: SemiSupervised Ward

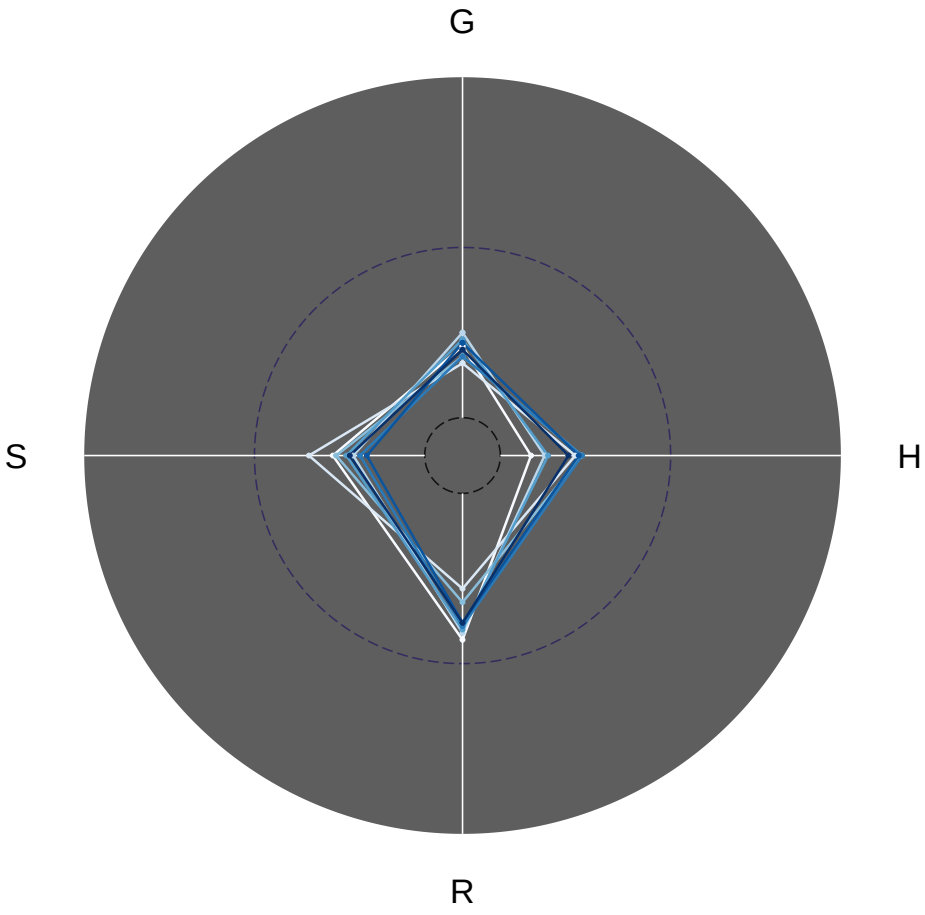


author: Becerra−Soto E.

Radar Ravenclaw

Ravenclaw's Students Profile

Approach: SemiSupervised Ward

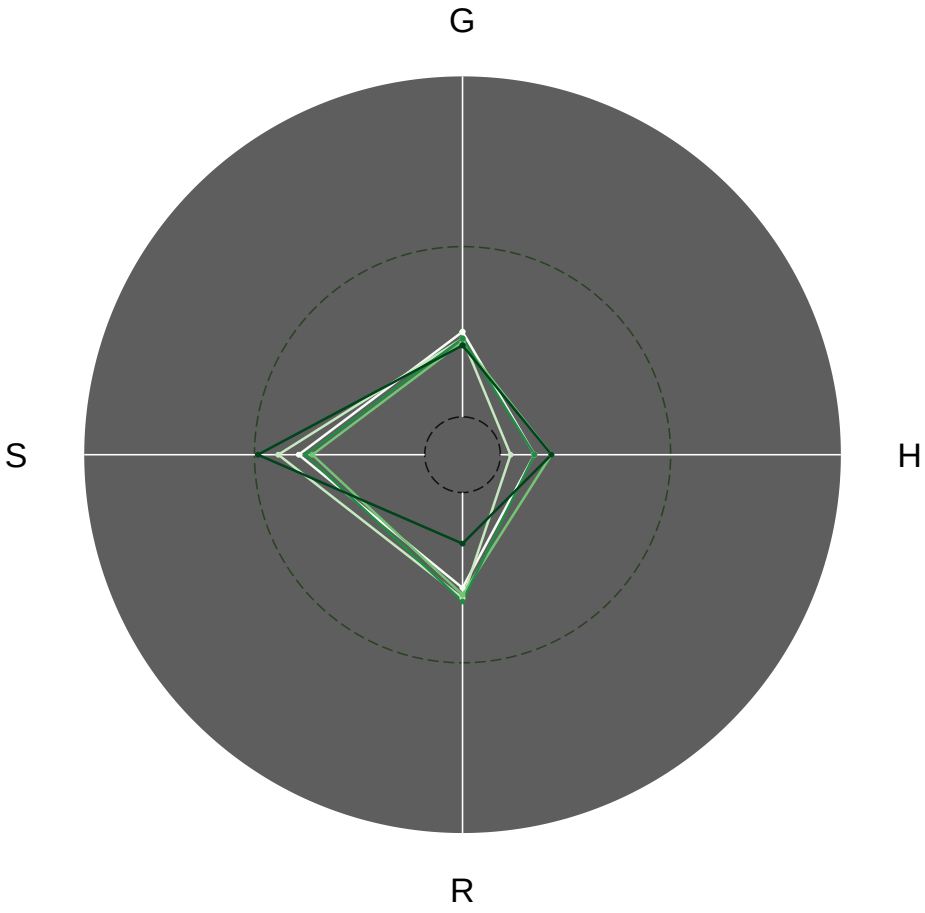


author: Becerra−Soto E.

Radar Slytherin

Slytherin's Students Profile

Approach: SemiSupervised Ward

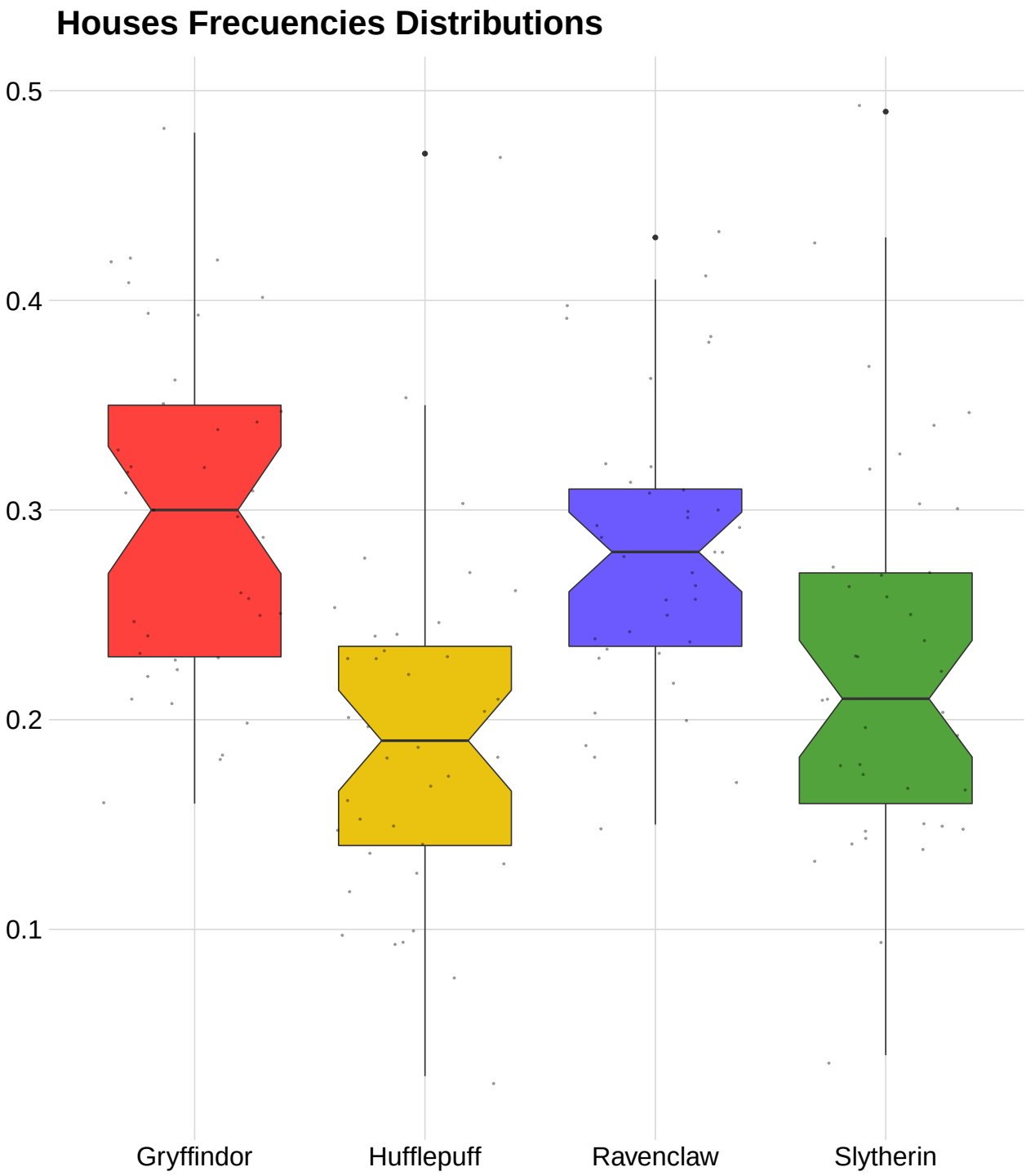


author: Becerra−Soto E.

Extras

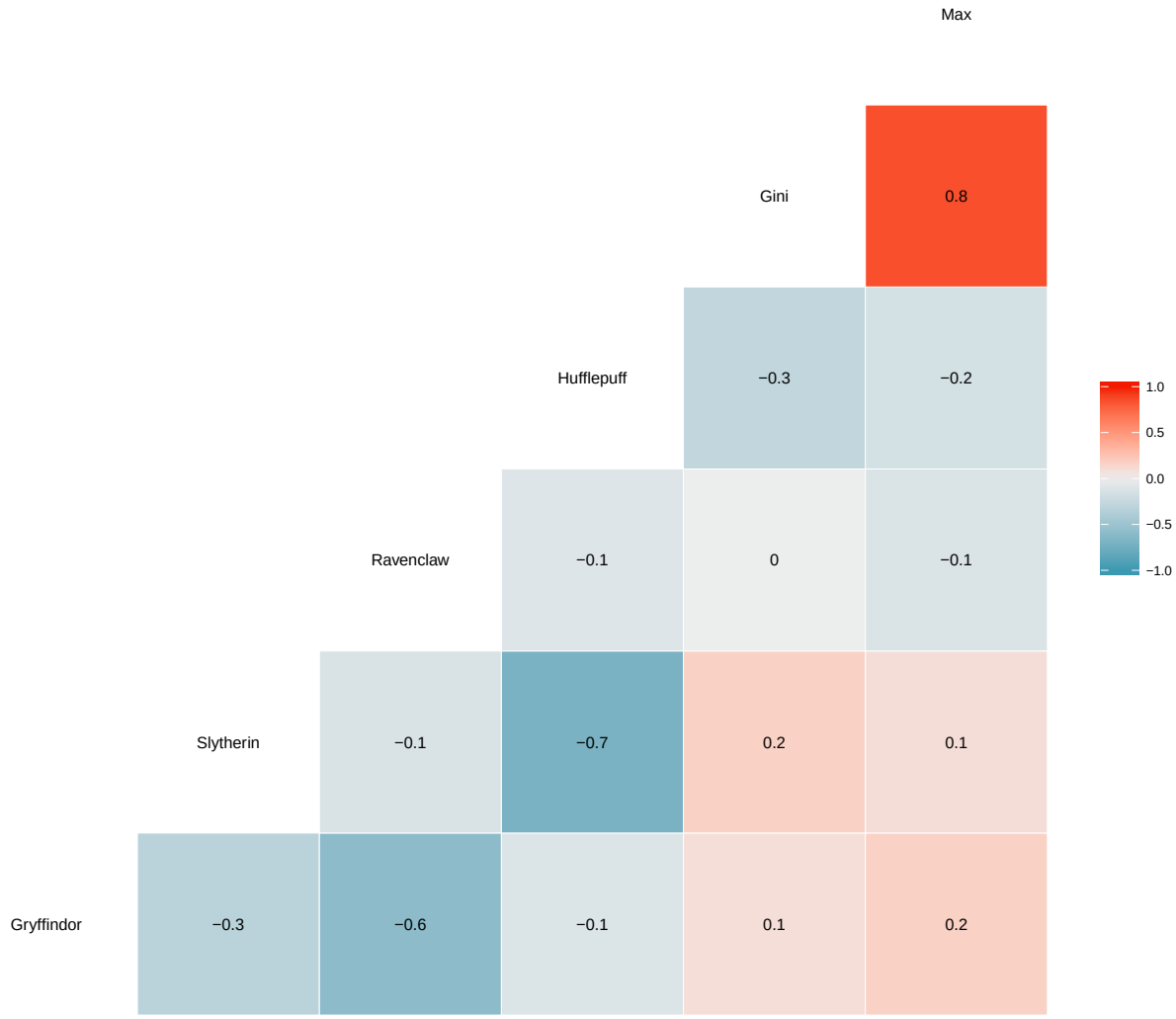
Plots

Box Plot of Houses



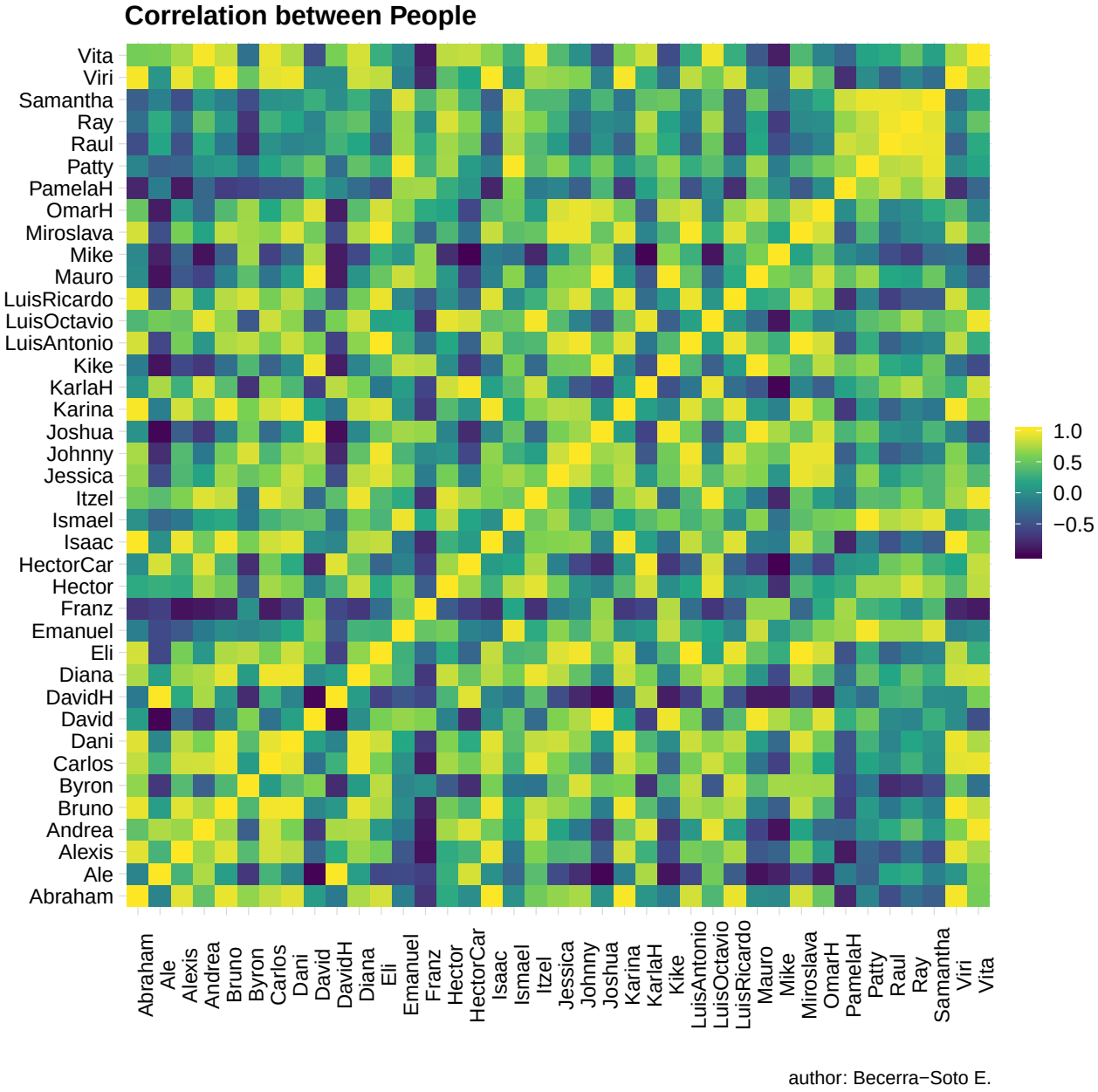
Correlation of Houses

Correlation between Houses



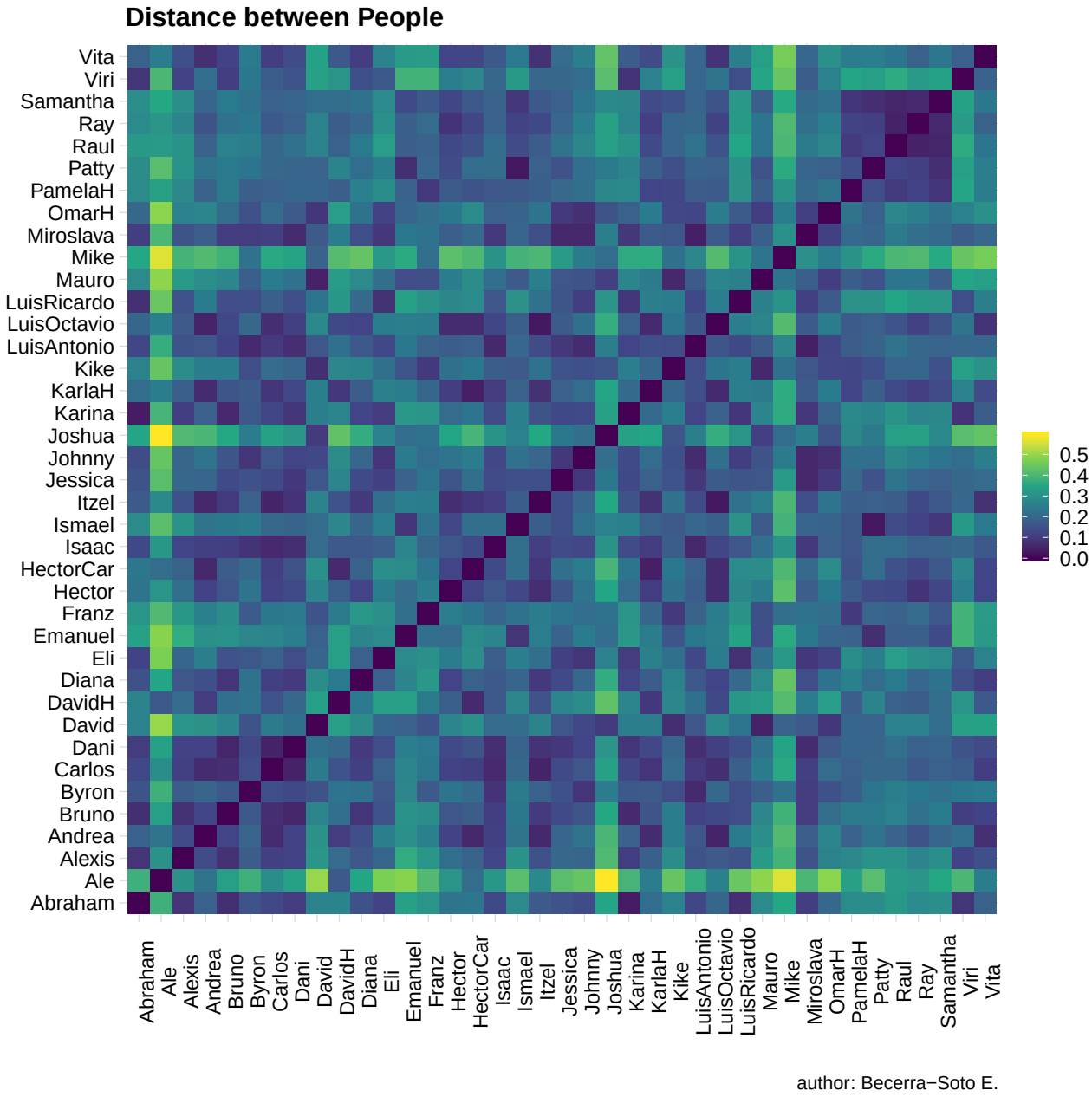
author: Becerra−Soto E.

Correlation Matrix of People



author: Becerra−Soto E.

Distance Matrix of People



author: Becerra−Soto E.

Combinatorics

How many different results are in the quiz?

Just to get an approximation we are going to suppose that each question has four answers and that each answer contributes towards only one House percentage. (We know that our assumptions are wrong, as the questions 1. and 13. only have three answers).

If each question maps to just one House we could represent an instance of the quiz as a string of length 15 where each position marks the House that is supported by the answer with the same number as the position in the string. For instance the string: GHRSRSGG... would represent a quiz that the answer number 1 contributes to Gryffindor, the answer number 2 contributes to Hufflepuff. the number 3 to Ravenclaw, and so on.

Because we are interested in the total contribution to each House, the order does not matter and also we are allowed to repeat the the contribution to the same House between different questions.

So the problem gets reduced to count the different ways to fill 15 positions (answers), from a pool of 4 characters (Houses), with repetition and where order does not matter.

Using separators it could be easily solved.

$\binom{15+4-1}{4-1} = 816$

So there are, an approximation of, 816 unique ways to answer the quiz.

References

https://www.buzzfeed.com/eleanorbate/accurate-af-sorting-quiz

https://en.wikipedia.org/wiki/Hogwarts

https://en.wikipedia.org/wiki/Magical_objects_in_Harry_Potter

https://www.r-bloggers.com/how-to-perform-hierarchical-clustering-using-r/

Quiz

1. **You've made it to Hogwarts, which means you've already bought a wand from Ollivander's. What material is at its core?**

- Phoenix Feather
- Dragon Heartstring
- Unicorn Hair

2. **During the end-of-year exams, you notice that one of your classmates was using an enchanted quill. You come top of the class anyway, but they are second. What do you do?**

- Tell the professor immediately – cheating is wrong, no matter what.
- Nothing, but if I hadn't come top of the class, I'd definitely tell the professor.
- Encourage the other student to admit what they'd done to the professor.
- Give them a high five for managing to sneak the quill into the exam.

3. **You would be most hurt if a person called you...**

- Weak
- Ignorant
- Unkind
- Boring

4. **You're locked in a duel with a skilled opponent. They fire an unknown spell at you, and you shout...**

- Expelliarmus!
- Protego!
- Stupefy!
- Crucio!

5. **It's your fifth year at Hogwarts, and you've just received a Howler from your parents. What for?**

- Sneaking into the Forbidden Forest at night on a dare.
- Getting caught cheating in my Divination O.W.L.
- Being put in detention after I was caught in the library after hours.
- Nothing! I'd never do anything to warrant a Howler.

6. **Which of these Dumbledore quotations speaks to you?**

- "Pity the living, and above all, those who live without love."
- "Words are, in my not-so-humble opinion, our most inexhaustible source of magic."
- "It matters not what someone is born, but what they grow to be."
- "It does not do to dwell on dreams and forget to live."

7. **Which of these most accurately describes your relationship with your closest friends?**

- I love surrounding myself with people – the more friends I have, the better!

- I have a few very close friends that I would trust with my life.
- I tend to be wary around new people, so don't make new friends often.
- I find myself becoming friends with people who can help me to succeed.

8. **Which of your skills are you most proud of?**

- My ability to absorb new information.
- My ability to make new friends.
- My ability to get what I want.
- My ability to keep secrets.

9. **The first Quidditch match of the season is approaching, and you can't wait to get** involved. What role are you playing?

- Seeker. I want the glory!
- Chaser. I like to be involved, and work as part of the team.
- Beater. I like having all that power.
- I'll be in the crowd, making sure supporter morale is high!

10. **You're allowed a pet at Hogwarts: an owl, a cat, or a toad. Which do you bring?**

- Owl
- Cat
- Toad
- Nothing. I can't be trusted to look after a pet!

11. **It's Saturday, you've finished your homework, and you have some free time. You decide to spend some time away from your common room. Where do you go?**

- The Forbidden Forest
- The library
- The kitchens
- The Room of Requirement

12. **What would you see in the Mirror of Erised?**

- Myself, surrounded by riches.
- Myself, surrounded by my loving family and friends.
- Myself, knowledgeable above all.
- Myself, experiencing a marvelous adventure.

13. **Choose a Deathly Hallow.**

- The Elder Wand
- The Resurrection Stone
- The Cloak of Invisibility

14. **Which path do you intend to follow after leaving Hogwarts?**

- I'd join the Ministry – I want to make a difference in the world.
- I think I'd travel for a while before committing to a career.
- I'd settle down and start a family as soon as possible!
- I'd continue to work hard in order to achieve as much success as possible.

15. **And finally: We know that the Sorting Hat takes into account your preferences. So which Hogwarts house do you feel you identify with most closely?**

- Gryffindor
- Hufflepuff
- Ravenclaw
- Slytherin