DS 210 final_project report

Dataset:
In the Project, I used the dataset of NBA 2022-2023. This dataset records all the players of 2022-2023 in the NBA, there are records of their age, name, team , and basketball ability. In this project, I will mainly use nba players for the nodes.
Here is the link of the data:
https://www.kaggle.com/code/osniteodoro/nba-regular-season-and-22-23-playoffs-analysis

Purpose
The main purpose of this project :
1. Get the node graph of all players between the team(use graphviz) so that we can directly see the relationship of players.
2.Get the average distance of the nodes based the graph,
3. Find the Graph clustering and partitioning, setting the k value as 5 and find the 5 best representatives in the graph for partitioning.
4. Find the most similar nodes and most dissimilar nodes, and find their similarity percentage.

For the Rust code,

Regarding the explanation of equations in Rust code, I will use comments in rust code to explain in detail. However, I will briefly introduce it here.
In the model module, I mainly read the two columns of player and team based on the csv.
In the graph module, what I mainly calculate here is the average distance of nodes and generate a dot file based on the data given by csv, and then use graphviz to generate pictures.
The main thing we are looking for in the "similarity module" is to use the jaccard similarity calculation method to find the percentage of similarity between the most similar node and the least similar node.
In the cluster module, I set the k value to 5 and use this to find the 5 most representative nba players
The io module is for reading csv files

Finally, I wrote a test under each module. I assumed that some of the values were 0 or something else (number), and then saw if it could pass or if it did not match the correct answer. You can check this in the rust code.

Here is the output result, The average distance is 0.96
That said “Graph has been exported to nba_graph.dot” , that mean you have to use graphviz create the graph by yourself, but don’t worry, I already put the output graph on the file
(The dot file almost 10000 lines)
If I set the K value as 5, then showing the player are “ Jose Alvarado”, “Ryan Arcidiacono”, “RJ Barrett”, “Jalen Brunson”, and “ Dyson Daniels”
The most similarly player, their similarity percentage is 97% and the dissimilar is 0 %
Output