Report:

For this Apriori analysis, the "NBA Players" dataset from Kaggle was used. Created by Justinas Cirtautas in 2020 and updated in 2022, this dataset includes basic box score statistics for NBA players who played between the 1996-97 and 2021-2022 seasons. These stats include points, rebounds, assists, and advanced efficiency metrics (e.g., true shooting percentage, usage rate, and net rating). The goal of this mini project was to utilize the Apriori algorithm to uncover frequently co-occurring relationships between player statistics that reflect patterns found in the real-world NBA.

Before applying the Apriori algorithm to the dataset, the dataset was cleaned and preprocessed. Players who played less than 20 games (~1/4th of the 82-game regular season) were filtered out and removed from the dataset to eliminate outlier statistics from players with small sample sizes. Additional cleaning steps were taken, including removing nulls, empty strings, and empty numbers. Post-cleaning, the dataset contained 10,679 records. The next step was to choose the essential columns. The columns chosen were team, age, height, weight, games played ("gp"), points ("pts"), rebounds ("reb"), assists ("ast"), net rating, offensive rebound percentage ("oreb_pct"), defensive rebound percentage ("dreb_pct"), usage rate/percentage ("usg_pct"), true shooting percentage ("ts_pct"), and assist percentage ("ast_pct"). These columns were chosen for their importance to the game of basketball and give the most complete picture of any given basketball player. After all of this, the data frame was saved as "bbDataImportant."

Next, continuous statistics such as true shooting percentage, usage rate, offensive rebound percentage, and net rating were binned and discretized into three categories (Low, Medium, and High). Basketball-specific thresholds were applied to each, such as a true shooting percentage above 60% being considered "elite" and a usage percentage above 25% denoted as high usage. All numeric attributes were then converted to factors, and the dataset was transformed into a "transactions" object through the arules package. Each player represented one transaction, and each categorical variable was treated as an item. This format enables Apriori to operate appropriately and analyze the co-occurring relationships among stats.

The Apriori algorithm was implemented through a series of runs, each with a summary of each and the top ten rules by confidence and lift. Starting with a support threshold of 5%, a confidence threshold of 60%, and a minimum rule length of 2, the first run generated over 26,000 rules. To improve the interpretability of the rules, the thresholds were raised to 15% support, 80% confidence, and a minimum length of 2. This resulted in fewer than 2,000 rules (1,291 rules total).

After raising the thresholds, the final Apriori model generated 1,291 rules. The top ten rules, ranked by confidence and lift, were generated for this model. The top five rules, ranked by confidence and lift, are presented in the tables below for readability and convenience, along with explanations of their meaning and application in the real-world NBA.

Top Ten Rules by Confidence:

Rule	Support	Confidence	Lift	Interpretation
{ast_pct=Low} => {ast=Low}	0.4659612	1	1.156988	Players with a low assist percentage consistently have low total assists.
{height=Tall, ast_pct=Low} => {ast=Low}	0.1951494	1	1.156988	Tall players with low assist percentages also have low assist totals. This is typical of big men throughout the last 30 years, who are more focused on rebounding and defense.
{weight=Light, dreb_pct=Low} => {reb=Low}	0.2016106	1	1.271007	Lightweight players (who are typically guards/shorter players) with low defensive rebound percentages also record low total rebounds, which is consistent with how NBA guards play.
{net_rating=Average, dreb_pct=Low} => {reb=Low}	0.1581609	1	1.271007	Players with average net ratings but low defensive rebounding rates tend to have low total rebounds. This seems to suggest that position rather than

				performance drives rebounding.
{oreb_pct=Low, dreb_pct=Low} => {reb=Low}	0.2608859	1	1.271007	Players who struggle on both offensive and defensive boards predictably have low rebound totals.

Top Ten Rules by Lift:

Rule	Support	Confidence	Lift	Interpretation
{height=Tall, weight=Medium, pts=Low, ast=Low} => {ast_pct=Low}	0.1597528	0.8245529	1.769574	Tall, medium-weight players with low scoring and assist counts almost always have low assist percentages. This is very typical of interior players, who are usually power forwards or centers.
{height=Tall, pts=Low, ast=Low} => {ast_pct=Low}	0.1652776	0.8232276	1.766730	Taller players who score and assist less frequently tend to have low assist percentages, which aligns with the above rule as well and is typical of power forwards and centers.
{height=Tall, weight=Medium, pts=Low} => {ast_pct=Low}	0.1597528	0.8213770	1.762758	Similar to the two rules above. Tall, moderately weighted players with low scoring

				numbers often show low assist percentages, consistent with the play of power forwards and centers.

{height=Tall, pts=Low}
=>
{ast_pct=Low}
	0.1652776	0.8201673	1.760162	Similar to the previous rules. Tall players with low scoring output typically have low assist percentages, which shows the limited playmaking responsibilities of frontcourt players.

{weight=Light, dreb_pct=Low}
=>
{oreb_pct=Low}
	0.1977713	0.9809568	1.709750	Lightweight players with low defensive rebounding percentages also have low offensive rebounding percentages, showing that guards typically rebound less than frontcourt players.

Given the above rules that rank in the top ten in confidence and lift, the following assessments of this dataset post-Apriori can be made:

- Guards (PG & SG), who are typically light and short players, have lower rebounding numbers but higher assist numbers.
- Big men (PF and C), who generally are heavier and taller players, demonstrate low assist rates but higher rebounding numbers.

- Rules with 100% confidence confirm consistency between the total metrics, while those with high lift highlight distinct positional roles and tendencies among players with different heights and weights.

To conclude, the Apriori algorithm successfully identified the most meaningful co-occurring relationships among NBA player stats and attributes. The discussed rules closely align with how basketball is played in the NBA today and throughout its nearly 80-year history. This mini project demonstrates the value of association rule mining, particularly the Apriori algorithm, in the context of sports analytics. These analytics provide NBA teams, players, front offices, fans, and the media with interpretable insights that complement watching basketball, offering these groups a different angle and an edge on how to approach the game. In the future, this mini project can be extended by incorporating datasets with per-minute, lineup-adjusted, and era-adjusted data to refine the discussed relationships further, as well as provide people with a bird's-eye view of how the game has evolved throughout its history.