Milestone 5 | NBA Statistics

INTRODUCTION: Stats for players and teams have long been a part of professional sports, but since the 2000s, data analytics has become an increasingly important part of developing and running a successful sports team. This revolution in data has also resulted in new ways of measuring what it means for a player or team to be effective.

If you're feeling a little rusty on the details of professional basketball, here's how it's played:

A game of basketball is played between two teams, each with five players. The objective is to score more points than the opposing team by shooting a ball through a hoop/basket. Players can score for their team in a variety of ways – point values are assigned to the location of the shot.

A basket made from inside the "three-point line" is worth two points, while a shot made from beyond the line is worth three points. "Free throws" can also be awarded to a player or a team when the opposing team commits a foul or breaks a rule. These are worth one point each.

HOW IT WORKS: Follow the prompts in the questions below to investigate your data. Post your answers in the provided boxes: the **yellow boxes** for the queries you write and **blue boxes** for text-based answers. When you're done, export your document as a pdf file and submit it on the Milestone page – see instructions for creating a PDF at the end of the Milestone.

RESOURCES: If you need hints on the Milestone or are feeling stuck, there are multiple ways of getting help. Attend Drop-In Hours to work on these problems with your peers, or reach out to the HelpHub if you have questions. Good luck!

PROMPT: In this Milestone, you'll be looking at the way that professional basketball in the NBA has changed over seventeen recent seasons. If you were a coach in the league, what could you say about how the game is being played, and what are the most successful teams doing to be successful?

SQL App: Here's that link to our specialized SQL app, where you'll write your SQL queries and interact with the data.

— Data Set **Description**

The NBA games dataset (`nba.games`) contains information about 23 335 games played from the 2004 season through the 2020 season. There are eighteen columns in the dataset, of which the following will be used in the Milestone:

- **`season`** - Starting year for the season the game was played. For example, games that are part of the 2010–11 season will have a season value of 2010, even if they are played in 2021.
- **`team_home`**, **`team_away`** - Full name of the home and visiting teams, respectively. Names will always reflect their current franchise names, even if they were known by a different name in prior years.
- **`pts_home`**, **`pts_away`** - Number of points scored by the home and visiting teams, respectively, in each game.
- **`home_team_win`** - Flag indicating whether the home team won (1) or the visiting team won (0).
- **`pct_3p_home, pct_3p_away`** - Percentage of 3 point shots made by the home team and away team, respectively.

— **Task 1:** Game Statistics Trends Over Time

A. Start by calculating the total number of rows and the first & last seasons in the dataset. This should be done in one query. If done correctly, the number of games is 23 335, the first season represented is 2004, and the last season represented is 2020.

(paste your query below 👇)

```
SELECT
  season
FROM
  nba.games
```

B. Write a query that returns the average score from the home team, away team, and the average of the home_team_win column. The average of the home_team_win column can be interpreted as the win rate for the home team. What do these values tell you about what you can expect from the result of a random NBA game?

(paste your query below 👇)

```
SELECT
  AVG(pts_home) AS average_home_points,
  AVG(pts_away) AS average_away_points,
  AVG(home_team_win) AS home_win_pct
FROM
  nba.games
```

(write your **answer** below 👇)

On average we can grow to expect the home team to win most games at a 58% chance, roughly speaking. Beyond just that, the games can be expected to go over 200 total points and with the win percentage, home is obviously more likely to score majority of those points at roughly 103 as compared to the away, who scores roughly 100 points on average. Beyond just the SQL data, this can be reflected in historical betting odds as teams that are evenly matched with talent the home team is usually given a -2.5 to -3.5 cushion which those numbers reside around the 3 point average difference of the years in the set. This number has shifted in recent years to even odds to -1.5 so I expect that change to be reflected in the data.

C. Modify your query from part B, so that the average scores from the home team, away team, and the home team win rate are grouped by each NBA season. Sort your output so that the seasons are ordered chronologically. What can you say about the trend in these values over the years?

(paste your query below 👇)

```
SELECT
  season,
  AVG(pts_home) AS average_home_points,
  AVG(pts_away) AS average_away_points,
  AVG(home_team_win) AS home_win_pct
FROM
  nba.games
GROUP BY
  season
ORDER BY
  season DESC
```

(write your **answer** below 👇)

As I pointed out in my answer above the betting odds tend to reflect the change in data which is very cool. From 2004–2019 the point average has stayed the difference between the home and away teams, average points has likewise also gone up from the 90s to north of 110 points, probably from more three point shot attempts in the 'Steph Era'. Interestingly the win percentage of the home team drastically changed (possibly from covid?) and the score difference got closer to match the difference of home win percentage. Theoretically covid couldve dampened home advantage with less fans, as well as teams naturally getting more competitive. I would need more recent data to conclude but my guess is covid.

D. Add two more summaries to your query from part C, to get the average 3-point shot rate for both away and home teams. Do these values change over time?

(paste your query below 👇)

```
SELECT
  season,
  AVG(pts_home) AS average_home_points,
  AVG(pts_away) AS average_away_points,
  AVG(home_team_win) AS home_win_pct,
  AVG(pct_3p_home) AS avg_home_3pt_pct,
  AVG(pct_3p_away) AS avg_away_3pt_pct
FROM
  nba.games
GROUP BY
  season
ORDER BY
  season DESC
```

(write your **answer** below 👇)

Interestingly it does change by roughly a percent for home from 2004–2021 and by just under 2% for away in the same time frame. I would've expected a larger jump in percentage, but it is important to keep in mind more shots are being taken so we will probably not see huge jumps in the data. Although 2% is a very big jump for the away teams and more than likely is the reason for the scores being closer as the three point shot is the most valuable shot- you make more you tend to win more.

— **Task 2:** Investigating 3-point Shooting

The average three-point shot rate is about 35.4% over the entire dataset. Let's write some queries to investigate just how important a high three-point shot rate is in terms of winning games.

A. Write a query that returns the average home team win rate and average three-point percentage at home grouped by home team name and season. **Note:** you will not be looking at the away team in this analysis.

(You should get a table with 510 rows.)

(paste your query below 👇)

```
SELECT
  team_home,
  season,
  AVG(pts_home) AS average_home_points,
  AVG(home_team_win) AS home_win_pct,
```

```
  AVG(pct_3p_home) AS avg_home_3pt_pct
FROM
  nba.games
GROUP BY
  team_home,
  season
ORDER BY
  avg_home_3pt_pct DESC,
  home_win_pct DESC
```

B. Modify your query so we are only looking at results from 2018 or later. Remember, the season column is a text field - don't forget your quotes! (This should reduce your results down to 90 rows.)

(paste your query below 👇)

```
SELECT
  team_home,
  season,
  AVG(pts_home) AS average_home_points,
  AVG(home_team_win) AS home_win_pct,
  AVG(pct_3p_home) AS avg_home_3pt_pct
FROM
  nba.games
WHERE
  season >= '2018'
GROUP BY
  team_home,
  season
ORDER BY
  avg_home_3pt_pct DESC,
  home_win_pct DESC
```

C. Add another expression to your query to answer the following question: How many teams had a three-point shot rate of at least 37% (i.e. 0.37)? (You'll get this from the output of the SQL app interface, rather than directly from the query.)

(paste your query below 👇)

```
SELECT
  team_home,
  season,
  AVG(pts_home) AS average_home_points,
  AVG(home_team_win) AS home_win_pct,
  AVG(pct_3p_home) AS avg_home_3pt_pct
FROM
  nba.games
WHERE
  season >= '2018'
GROUP BY
  team_home,
  season
HAVING
  AVG(pct_3p_home) >= 0.37
ORDER BY
  avg_home_3pt_pct DESC,
  home_win_pct DESC
```

(write your **answer** below 👇)

25 different NBA teams from 2018–2021 shot north of 37% from the three point line at home.

D. Add an additional condition to your query to filter to teams with a losing record (win rate < 0.5), in addition to having a high three-point shot rate. How many teams had a losing record while having a high 3-point shot

percentage? (As with the previous part, you'll read this from the SQL app interface instead of directly from the query.)

(paste your query below 👇)

```
SELECT
  team_home,
  season,
  AVG(pts_home) AS average_home_points,
  AVG(home_team_win) AS home_win_pct,
  AVG(pct_3p_home) AS avg_home_3pt_pct
FROM
  nba.games
WHERE
  season >= '2018'
GROUP BY
  team_home,
  season
HAVING
  AVG(pct_3p_home) >= 0.37
  AND AVG(home_team_win) < 0.50
ORDER BY
  avg_home_3pt_pct DESC,
  home_win_pct DESC
```

(write your **answer** below 👇)

Only two teams have shot above 37% from three while having a losing record between the years of 2018–2021. Those teams are the 2019–2020 New Orleans Pelicans and the 2020–2021 Sacramento KIngs.

E. Repeat parts C and D, but this time filtering to teams that had a low 3-point shooting rate of 34% (0.34) or less. How many teams had this low of a

3-point accuracy, and how many of these teams had a losing record? (Paste only the query that answers the last question into the query box.)

(paste your query below 👇)

```
SELECT
  team_home,
  season,
  AVG(pts_home) AS average_home_points,
  AVG(home_team_win) AS home_win_pct,
  AVG(pct_3p_home) AS avg_home_3pt_pct
FROM
  nba.games
WHERE
  season >= '2018'
GROUP BY
  team_home,
  season
HAVING
  AVG(pct_3p_home) <= 0.34
  AND AVG(home_team_win) < 0.50
ORDER BY
  avg_home_3pt_pct DESC,
  home_win_pct DESC
```

(write your **answer** below 👇)

7 different teams from 2018-2021 have had a losing record while simultaneously shooting below 34% from the three point line as the home team.

F. What conclusions can you draw from your analysis in the previous Parts regarding the relationship between 3-point shot rates and team success?

(write your **answer** below 👇)

Very good teams tend to shoot above 37% from the free throw line with some notable outliers, while worse teams tend to shoot low percentages from the three point line and simultaneously losing. This means the more threes the team makes the more likely they are to win, at least at home.

— **LevelUp:** Building the Team Stats table

Most of the time when working with data, you will have to build summary tables yourself. This is done to alleviate storage costs, especially since these tables have to constantly be updated via a SQL query. In this Level Up you'll create two summary tables, one for the home team and one for the away team. Although you don't have the ability to join these tables (yet!), you can come back to this LevelUp and create the full table once you have learned how to join data.

A. Write a query that returns the average number of home points scored, average 3 point percentage for the home team, and the number of wins for each team and season combination in the `nba.games` table. (You should get a table with 510 rows.)

(paste your query below 👇)

```
SELECT
  team_home,
  season,
  AVG(pts_home) AS average_home_points,
  AVG(pct_3p_home) AS avg_home_3pt_pct,
  SUM(home_team_win) AS n_of_home_wins
FROM
  nba.games
GROUP BY
  team_home,
  season
```

B. Repeat part A but now do the same thing for the **away** team. Note that you will have to get a little creative to calculate the number of away wins since the table only tells you whether or not the home team won the game. (You should get a table with 510 rows.)

(paste your query below 👇)

```
SELECT
  team_away,
  season,
  AVG(pts_away) AS average_away_points,
  AVG(pct_3p_away) AS avg_away_3pt_pct,
  SUM(CASE WHEN home_team_win = 0 THEN 1 ELSE 0 END) AS
n_of_away_wins
FROM
  nba.games
GROUP BY
  team_away,
  season
```

— Submission

Great work completing this Milestone! To submit your completed Milestone, you will need to download / export this document as a PDF and then upload it to the Milestone submission page. You can find the option to download as a PDF from the File menu in the upper-left corner of the Google Doc interface.