Exploratory Data Analysis with SQL

Using SQL queries to analyze my friends' NBA bets

Gollnick's NBA Playoffs Predictions Game was a betting game that I developed from 2018 to 2022 for the sake of having a little bit of fun with my friends during the NBA playoffs. Betting was actually free and there was no prize money.

Data was split into five tables:

- *rounds*
- *teams*
- *people*
- *series*
- *predictions*

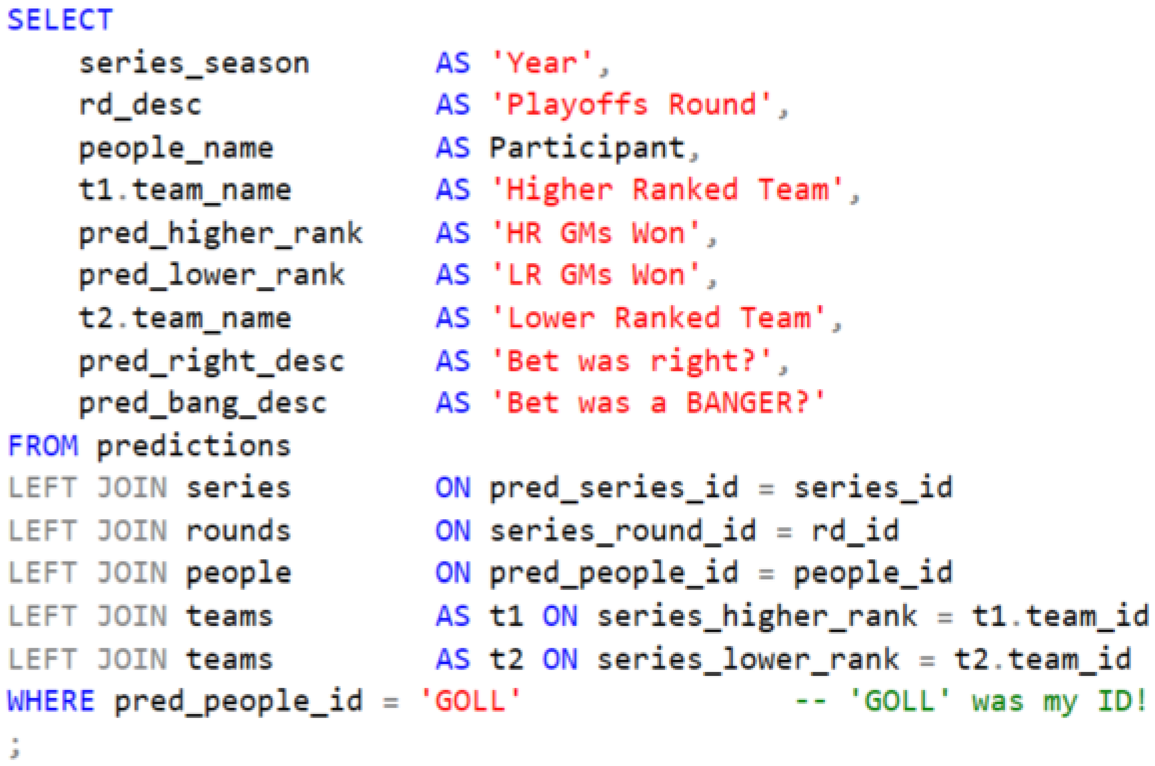
The last one could be considered as a fact table since it contains data from all the bets made by all participants through all the five years of Predictions Game.

Before querying, of course I need to create the tables and insert data on them. For this, I used **CREATE TABLE** and **BULK INSERT** commands. You may check them on the SQL file that is also available in this repository.

****For this study, please note that I am considering just the final score of games won in each series (as in 4–1), not the score for each individual game (as in, say 121–109).****

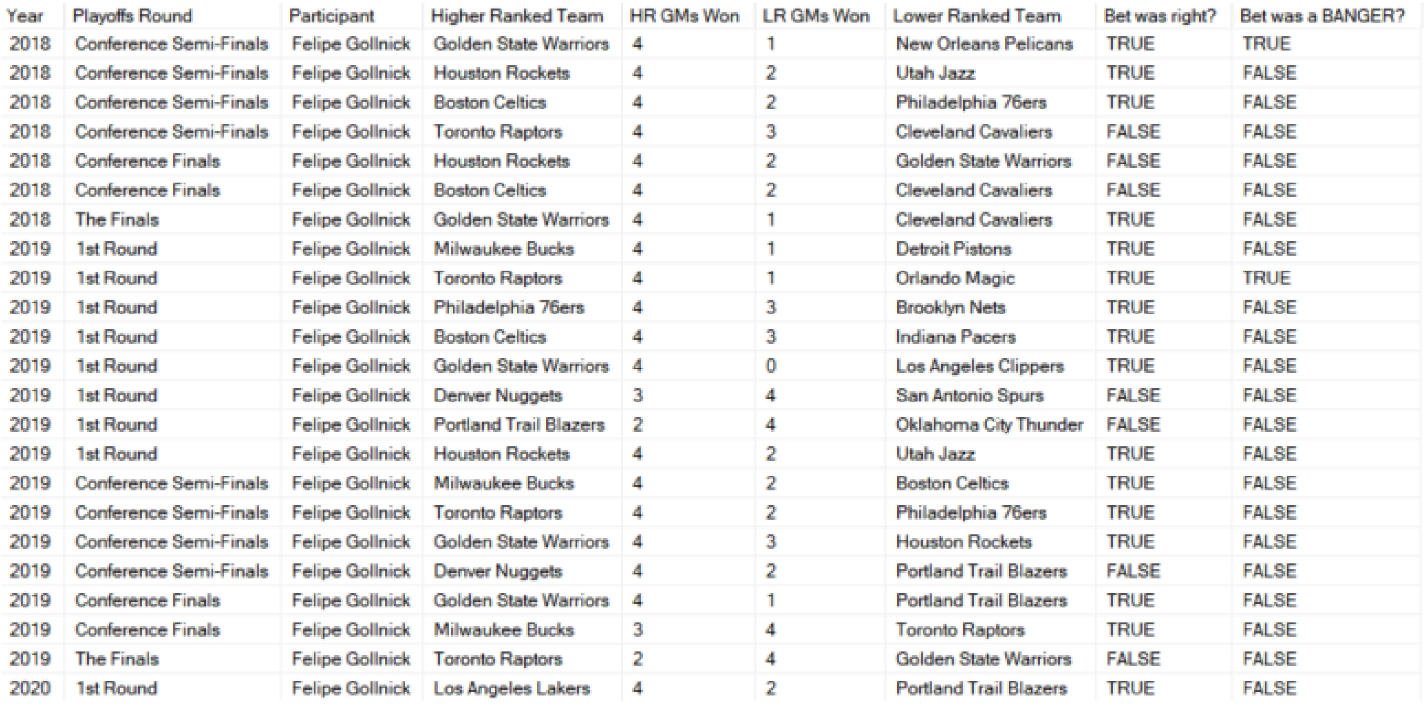
1st Query

Let's start finding out all the bets I made on all those years:



```
SELECT
    series_season       AS 'Year',
    rd_desc             AS 'Playoffs Round',
    people_name         AS Participant,
    t1.team_name        AS 'Higher Ranked Team',
    pred_higher_rank    AS 'HR GMs Won',
    pred_lower_rank     AS 'LR GMs Won',
    t2.team_name        AS 'Lower Ranked Team',
    pred_right_desc     AS 'Bet was right?',
    pred_bang_desc      AS 'Bet was a BANGER?'
FROM predictions
LEFT JOIN series        ON pred_series_id = series_id
LEFT JOIN rounds        ON series_round_id = rd_id
LEFT JOIN people        ON pred_people_id = people_id
LEFT JOIN teams         AS t1 ON series_higher_rank = t1.team_id
LEFT JOIN teams         AS t2 ON series_lower_rank = t2.team_id
WHERE pred_people_id = 'GOLL'               -- 'GOLL' was my ID!
;
```

And here's the output:



Year	Playoffs Round	Participant	Higher Ranked Team	HR GMs Won	LR GMs Won	Lower Ranked Team	Bet was right?	Bet was a BANGER?
2018	Conference Semi-Finals	Felipe Gollnick	Golden State Warriors	4	1	New Orleans Pelicans	TRUE	TRUE
2018	Conference Semi-Finals	Felipe Gollnick	Houston Rockets	4	2	Utah Jazz	TRUE	FALSE
2018	Conference Semi-Finals	Felipe Gollnick	Boston Celtics	4	2	Philadelphia 76ers	TRUE	FALSE
2018	Conference Semi-Finals	Felipe Gollnick	Toronto Raptors	4	3	Cleveland Cavaliers	FALSE	FALSE
2018	Conference Finals	Felipe Gollnick	Houston Rockets	4	2	Golden State Warriors	FALSE	FALSE
2018	Conference Finals	Felipe Gollnick	Boston Celtics	4	2	Cleveland Cavaliers	FALSE	FALSE
2018	The Finals	Felipe Gollnick	Golden State Warriors	4	1	Cleveland Cavaliers	TRUE	FALSE
2019	1st Round	Felipe Gollnick	Milwaukee Bucks	4	1	Detroit Pistons	TRUE	FALSE
2019	1st Round	Felipe Gollnick	Toronto Raptors	4	1	Orlando Magic	TRUE	TRUE
2019	1st Round	Felipe Gollnick	Philadelphia 76ers	4	3	Brooklyn Nets	TRUE	FALSE
2019	1st Round	Felipe Gollnick	Boston Celtics	4	3	Indiana Pacers	TRUE	FALSE
2019	1st Round	Felipe Gollnick	Golden State Warriors	4	0	Los Angeles Clippers	TRUE	FALSE
2019	1st Round	Felipe Gollnick	Denver Nuggets	3	4	San Antonio Spurs	FALSE	FALSE
2019	1st Round	Felipe Gollnick	Portland Trail Blazers	2	4	Oklahoma City Thunder	FALSE	FALSE
2019	1st Round	Felipe Gollnick	Houston Rockets	4	2	Utah Jazz	TRUE	FALSE
2019	Conference Semi-Finals	Felipe Gollnick	Milwaukee Bucks	4	2	Boston Celtics	TRUE	FALSE
2019	Conference Semi-Finals	Felipe Gollnick	Toronto Raptors	4	2	Philadelphia 76ers	TRUE	FALSE
2019	Conference Semi-Finals	Felipe Gollnick	Golden State Warriors	4	3	Houston Rockets	TRUE	FALSE
2019	Conference Semi-Finals	Felipe Gollnick	Denver Nuggets	4	2	Portland Trail Blazers	FALSE	FALSE
2019	Conference Finals	Felipe Gollnick	Golden State Warriors	4	1	Portland Trail Blazers	TRUE	FALSE
2019	Conference Finals	Felipe Gollnick	Milwaukee Bucks	3	4	Toronto Raptors	TRUE	FALSE
2019	The Finals	Felipe Gollnick	Toronto Raptors	2	4	Golden State Warriors	FALSE	FALSE
2020	1st Round	Felipe Gollnick	Los Angeles Lakers	4	2	Portland Trail Blazers	TRUE	FALSE

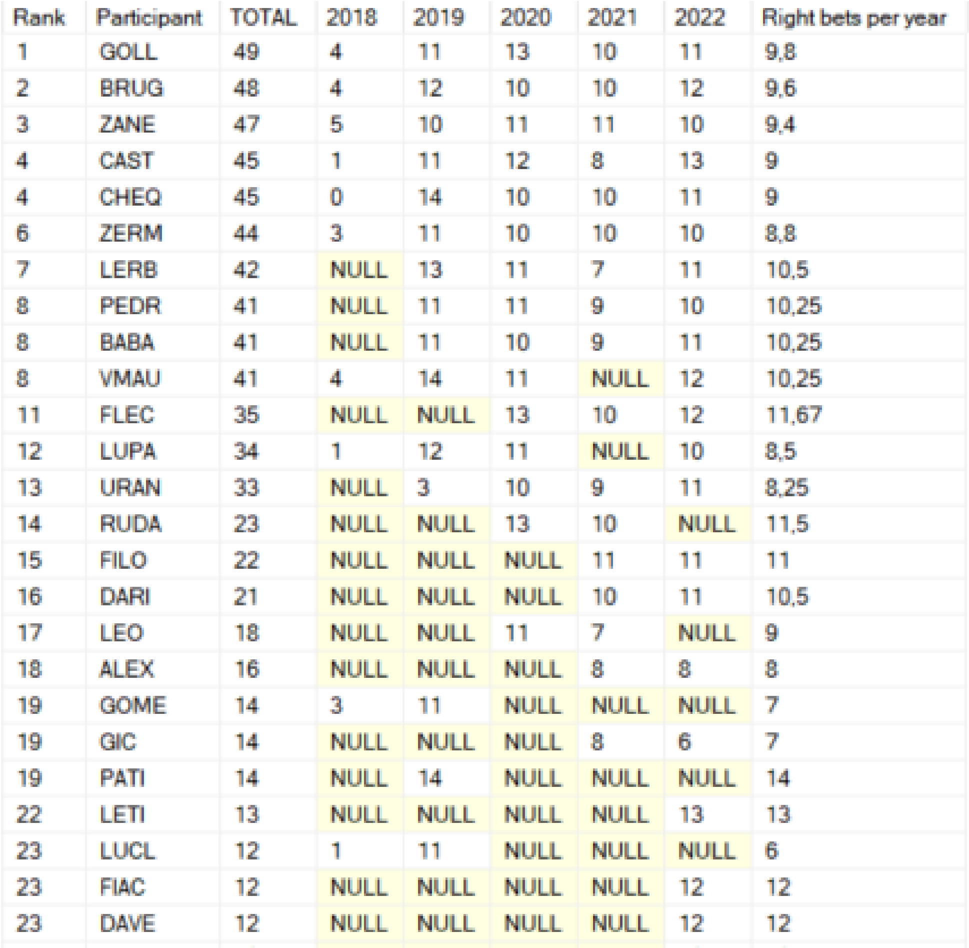
** **HR GMs Won** = my bet on how many games the higher ranked team would win in that series;*
** **LR GMs Won** = how many games the lower ranked team would win;*
** **BANGER** = Correctly guessing the final result of the series.*

2nd Query

Let's rank the participants by the number of times they correctly predicted the team that won the series:

```
SELECT
    RANK() OVER(ORDER BY SUM(pred_right) DESC)  AS 'Rank',
    pred_people_id                              AS Participant,
    SUM(pred_right)                             AS TOTAL,
    SUM(CASE WHEN series_season = 2018 THEN pred_right ELSE NULL END) AS '2018',
    SUM(CASE WHEN series_season = 2019 THEN pred_right ELSE NULL END) AS '2019',
    SUM(CASE WHEN series_season = 2020 THEN pred_right ELSE NULL END) AS '2020',
    SUM(CASE WHEN series_season = 2021 THEN pred_right ELSE NULL END) AS '2021',
    SUM(CASE WHEN series_season = 2022 THEN pred_right ELSE NULL END) AS '2022',
    ROUND((SUM(CAST(pred_right AS float)) / COUNT(DISTINCT series_season)), 2) AS 'Right bets per year'
FROM predictions
LEFT JOIN series ON pred_series_id = series_id
WHERE pred_people_id <> 'REAL' -- excluding the rows that contains the actual final result from the series
GROUP BY pred_people_id
ORDER BY Rank ASC;
```

The output:

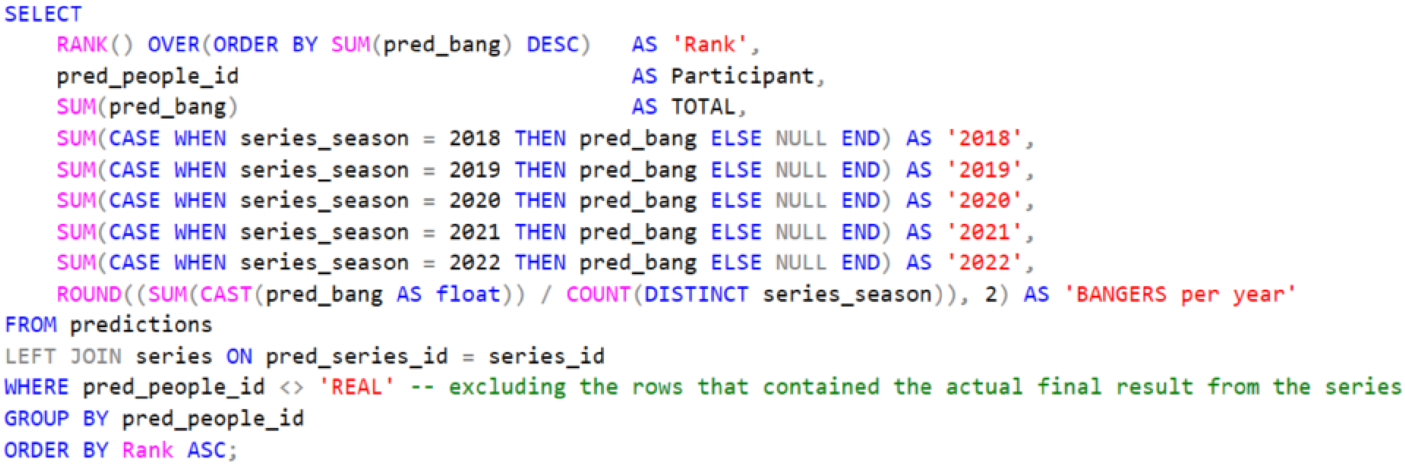


Rank	Participant	TOTAL	2018	2019	2020	2021	2022	Right bets per year
1	GOLL	49	4	11	13	10	11	9,8
2	BRUG	48	4	12	10	10	12	9,6
3	ZANE	47	5	10	11	11	10	9,4
4	CAST	45	1	11	12	8	13	9
4	CHEQ	45	0	14	10	10	11	9
6	ZERM	44	3	11	10	10	10	8,8
7	LERB	42	NULL	13	11	7	11	10,5
8	PEDR	41	NULL	11	11	9	10	10,25
8	BABA	41	NULL	11	10	9	11	10,25
8	VMAU	41	4	14	11	NULL	12	10,25
11	FLEC	35	NULL	NULL	13	10	12	11,67
12	LUPA	34	1	12	11	NULL	10	8,5
13	URAN	33	NULL	3	10	9	11	8,25
14	RUDA	23	NULL	NULL	13	10	NULL	11,5
15	FILO	22	NULL	NULL	NULL	11	11	11
16	DARI	21	NULL	NULL	NULL	10	11	10,5
17	LEO	18	NULL	NULL	11	7	NULL	9
18	ALEX	16	NULL	NULL	NULL	8	8	8
19	GOME	14	3	11	NULL	NULL	NULL	7
19	GIC	14	NULL	NULL	NULL	8	6	7
19	PATI	14	NULL	14	NULL	NULL	NULL	14
22	LETI	13	NULL	NULL	NULL	NULL	13	13
23	LUCL	12	1	11	NULL	NULL	NULL	6
23	FIAC	12	NULL	NULL	NULL	NULL	12	12
23	DAVE	12	NULL	NULL	NULL	NULL	12	12

** A **NULL** in this case means that the participant didn't take part of the game on that year.*

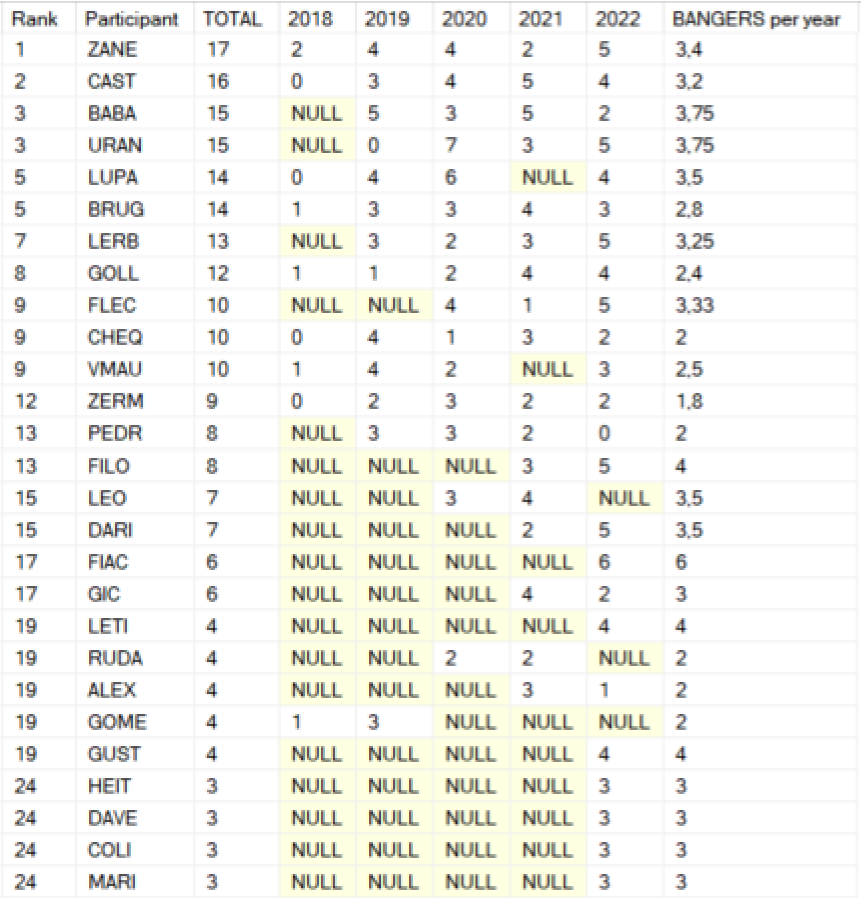
3rd Query

Let's remake that query, but this time to find out who had the most BANGERS, correctly guessing the final result of the series:



```
SELECT
    RANK() OVER(ORDER BY SUM(pred_bang) DESC)    AS 'Rank',
    pred_people_id                               AS Participant,
    SUM(pred_bang)                               AS TOTAL,
    SUM(CASE WHEN series_season = 2018 THEN pred_bang ELSE NULL END) AS '2018',
    SUM(CASE WHEN series_season = 2019 THEN pred_bang ELSE NULL END) AS '2019',
    SUM(CASE WHEN series_season = 2020 THEN pred_bang ELSE NULL END) AS '2020',
    SUM(CASE WHEN series_season = 2021 THEN pred_bang ELSE NULL END) AS '2021',
    SUM(CASE WHEN series_season = 2022 THEN pred_bang ELSE NULL END) AS '2022',
    ROUND((SUM(CAST(pred_bang AS float)) / COUNT(DISTINCT series_season)), 2) AS 'BANGERS per year'
FROM predictions
LEFT JOIN series ON pred_series_id = series_id
WHERE pred_people_id <> 'REAL' -- excluding the rows that contained the actual final result from the series
GROUP BY pred_people_id
ORDER BY Rank ASC;
```

The output:



Rank	Participant	TOTAL	2018	2019	2020	2021	2022	BANGERS per year
1	ZANE	17	2	4	4	2	5	3,4
2	CAST	16	0	3	4	5	4	3,2
3	BABA	15	NULL	5	3	5	2	3,75
3	URAN	15	NULL	0	7	3	5	3,75
5	LUPA	14	0	4	6	NULL	4	3,5
5	BRUG	14	1	3	3	4	3	2,8
7	LERB	13	NULL	3	2	3	5	3,25
8	GOLL	12	1	1	2	4	4	2,4
9	FLEC	10	NULL	NULL	4	1	5	3,33
9	CHEQ	10	0	4	1	3	2	2
9	VMAU	10	1	4	2	NULL	3	2,5
12	ZERM	9	0	2	3	2	2	1,8
13	PEDR	8	NULL	3	3	2	0	2
13	FILO	8	NULL	NULL	NULL	3	5	4
15	LEO	7	NULL	NULL	3	4	NULL	3,5
15	DARI	7	NULL	NULL	NULL	2	5	3,5
17	FIAC	6	NULL	NULL	NULL	NULL	6	6
17	GIC	6	NULL	NULL	NULL	4	2	3
19	LETI	4	NULL	NULL	NULL	NULL	4	4
19	RUDA	4	NULL	NULL	2	2	NULL	2
19	ALEX	4	NULL	NULL	NULL	3	1	2
19	GOME	4	1	3	NULL	NULL	NULL	2
19	GUST	4	NULL	NULL	NULL	NULL	4	4
24	HEIT	3	NULL	NULL	NULL	NULL	3	3
24	DAVE	3	NULL	NULL	NULL	NULL	3	3
24	COLI	3	NULL	NULL	NULL	NULL	3	3
24	MARI	3	NULL	NULL	NULL	NULL	3	3

See what happened with participant **GOLL** (which happens to be myself)?
In the previous query, I was in first place, but in the second one, I'm eighth.

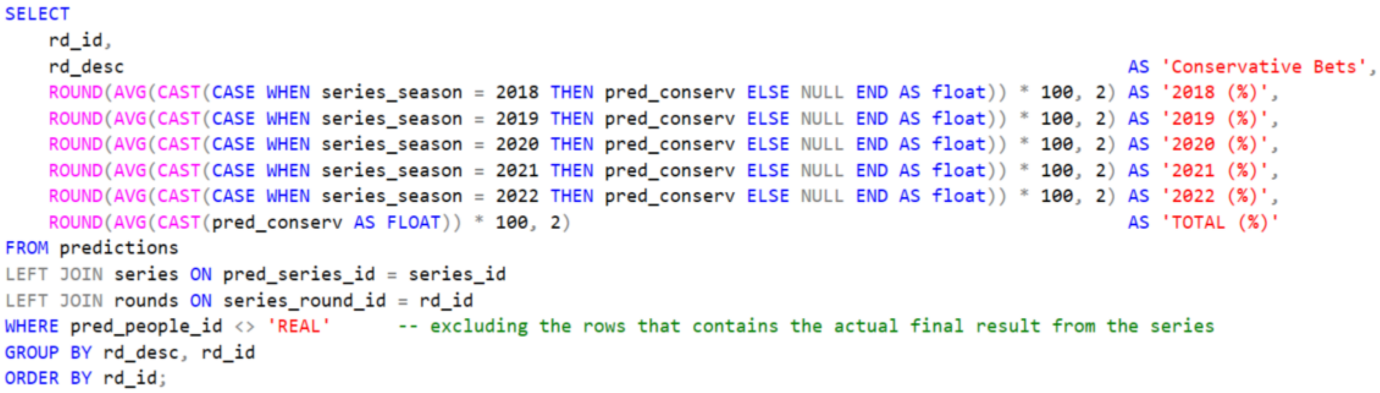
That means that I can get the winning teams right, but not the final scores. That also means that I was probably had the most **conservative bets**.

And by that, I mean that I thought that the **higher-ranked team would beat the lower-ranked team in a series.**

I had a feeling that, in this game's five-year history, most of the bets made by all the participants would be conservative too.

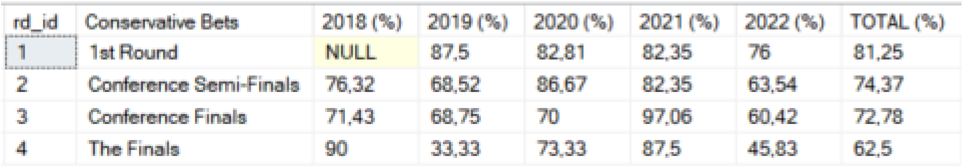
4th Query

Finding out the % of all the bets that were conservative, by year and round:



```
SELECT
    rd_id,
    rd_desc                                                                                                   AS 'Conservative Bets',
    ROUND(AVG(CAST(CASE WHEN series_season = 2018 THEN pred_conserv ELSE NULL END AS float)) * 100, 2) AS '2018 (%)',
    ROUND(AVG(CAST(CASE WHEN series_season = 2019 THEN pred_conserv ELSE NULL END AS float)) * 100, 2) AS '2019 (%)',
    ROUND(AVG(CAST(CASE WHEN series_season = 2020 THEN pred_conserv ELSE NULL END AS float)) * 100, 2) AS '2020 (%)',
    ROUND(AVG(CAST(CASE WHEN series_season = 2021 THEN pred_conserv ELSE NULL END AS float)) * 100, 2) AS '2021 (%)',
    ROUND(AVG(CAST(CASE WHEN series_season = 2022 THEN pred_conserv ELSE NULL END AS float)) * 100, 2) AS '2022 (%)',
    ROUND(AVG(CAST(pred_conserv AS FLOAT)) * 100, 2)                                                   AS 'TOTAL (%)'
FROM predictions
LEFT JOIN series ON pred_series_id = series_id
LEFT JOIN rounds ON series_round_id = rd_id
WHERE pred_people_id <> 'REAL'      -- excluding the rows that contains the actual final result from the series
GROUP BY rd_desc, rd_id
ORDER BY rd_id;
```

The output:



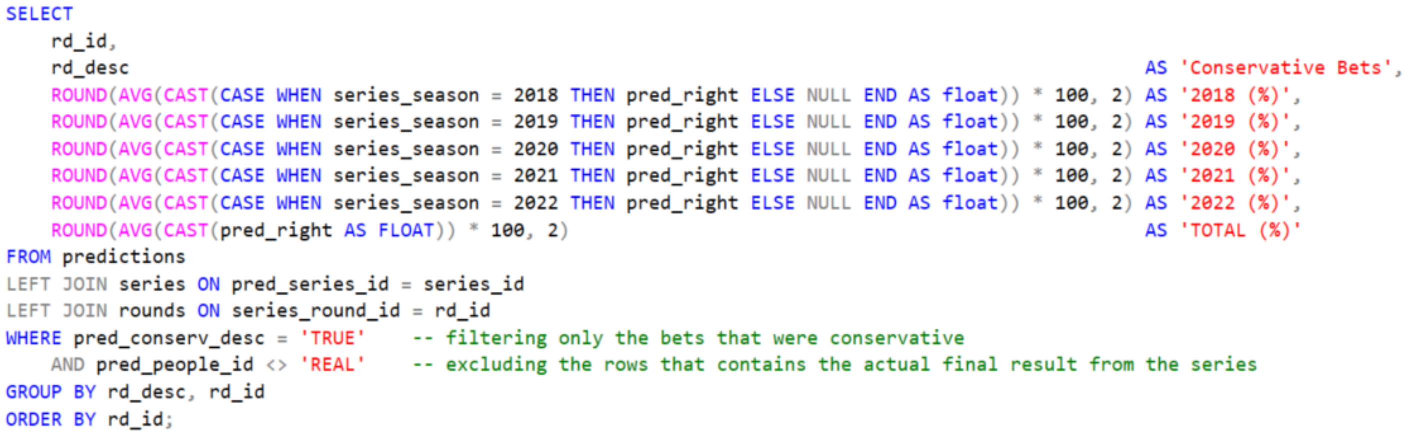
rd_id	Conservative Bets	2018 (%)	2019 (%)	2020 (%)	2021 (%)	2022 (%)	TOTAL (%)
1	1st Round	NULL	87,5	82,81	82,35	76	81,25
2	Conference Semi-Finals	76,32	68,52	86,67	82,35	63,54	74,37
3	Conference Finals	71,43	68,75	70	97,06	60,42	72,78
4	The Finals	90	33,33	73,33	87,5	45,83	62,5

**There was no betting in the first round of the 2018 Playoffs.*

Yes, most of the bets were conservative. That outlier on the 2019 Finals was that clash between the Toronto Raptors and the Golden State Warriors, where the Raptors were the higher ranked team.

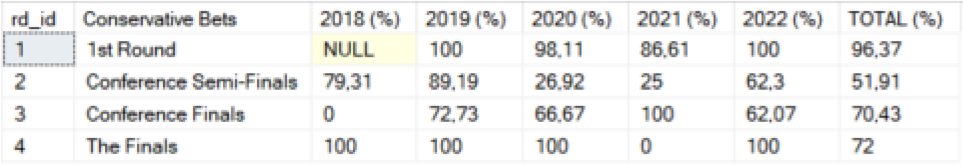
5th Query

And was it worth, making conservative bets? Let's find the percentage of all the conservative bets that were right, by year and round:



```
SELECT
    rd_id,
    rd_desc                                                                                   AS 'Conservative Bets',
    ROUND(AVG(CAST(CASE WHEN series_season = 2018 THEN pred_right ELSE NULL END AS float)) * 100, 2) AS '2018 (%)',
    ROUND(AVG(CAST(CASE WHEN series_season = 2019 THEN pred_right ELSE NULL END AS float)) * 100, 2) AS '2019 (%)',
    ROUND(AVG(CAST(CASE WHEN series_season = 2020 THEN pred_right ELSE NULL END AS float)) * 100, 2) AS '2020 (%)',
    ROUND(AVG(CAST(CASE WHEN series_season = 2021 THEN pred_right ELSE NULL END AS float)) * 100, 2) AS '2021 (%)',
    ROUND(AVG(CAST(CASE WHEN series_season = 2022 THEN pred_right ELSE NULL END AS float)) * 100, 2) AS '2022 (%)',
    ROUND(AVG(CAST(pred_right AS FLOAT)) * 100, 2)                                            AS 'TOTAL (%)'
FROM predictions
LEFT JOIN series ON pred_series_id = series_id
LEFT JOIN rounds ON series_round_id = rd_id
WHERE pred_conserv_desc = 'TRUE'    -- filtering only the bets that were conservative
    AND pred_people_id <> 'REAL'    -- excluding the rows that contains the actual final result from the series
GROUP BY rd_desc, rd_id
ORDER BY rd_id;
```

The output:



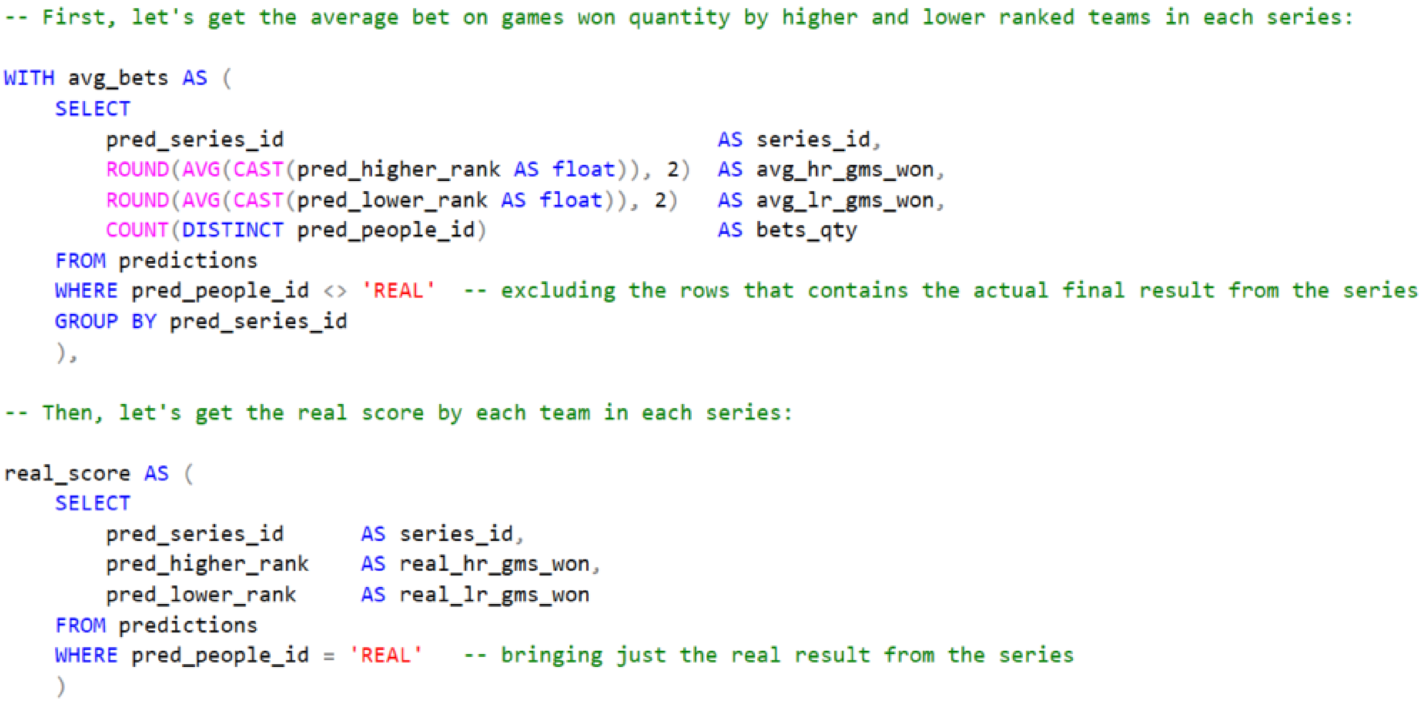
rd_id	Conservative Bets	2018 (%)	2019 (%)	2020 (%)	2021 (%)	2022 (%)	TOTAL (%)
1	1st Round	NULL	100	98,11	86,61	100	96,37
2	Conference Semi-Finals	79,31	89,19	26,92	25	62,3	51,91
3	Conference Finals	0	72,73	66,67	100	62,07	70,43
4	The Finals	100	100	100	0	100	72

**There was no betting in the first round of the 2018 Playoffs.*

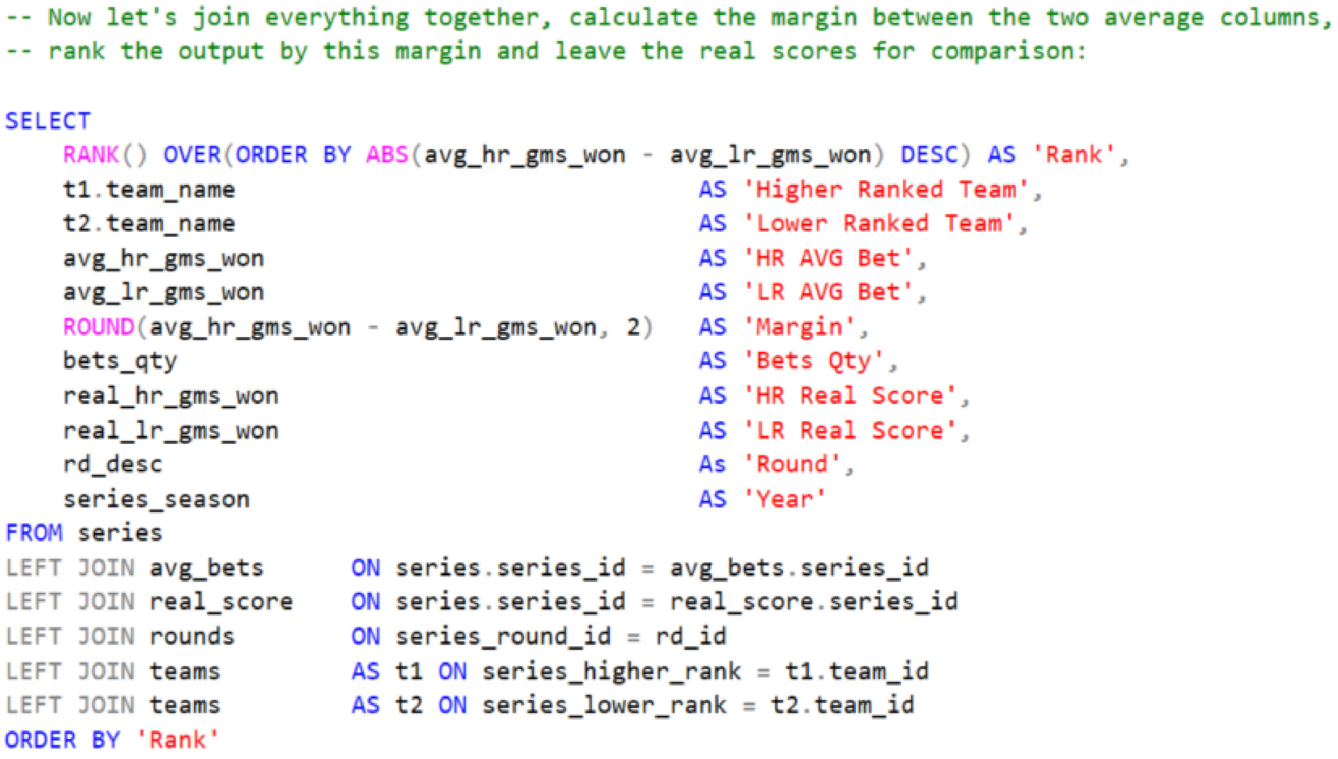
Yes, most of the times, being conservative was worth it. Except when it wasn't.

6th Query

Talking about series results, is it possible to discover on which ones the participants thought one team would really blast the other?
And the answer is: yes! For this, let's play around with a couple of CTEs:

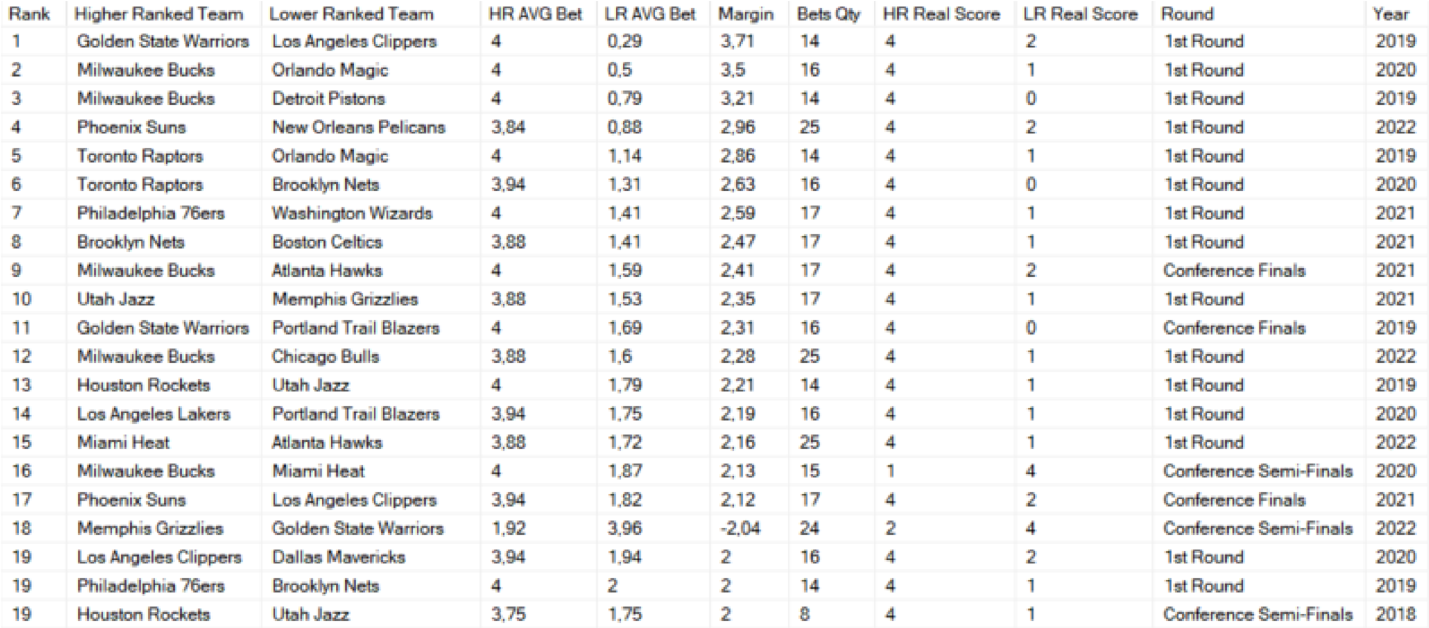


```
-- First, let's get the average bet on games won quantity by higher and lower ranked teams in each series:

WITH avg_bets AS (
    SELECT
        pred_series_id                                  AS series_id,
        ROUND(AVG(CAST(pred_higher_rank AS float)), 2)  AS avg_hr_gms_won,
        ROUND(AVG(CAST(pred_lower_rank AS float)), 2)   AS avg_lr_gms_won,
        COUNT(DISTINCT pred_people_id)                  AS bets_qty
    FROM predictions
    WHERE pred_people_id <> 'REAL'  -- excluding the rows that contains the actual final result from the series
    GROUP BY pred_series_id
    ),

-- Then, let's get the real score by each team in each series:

real_score AS (
    SELECT
        pred_series_id      AS series_id,
        pred_higher_rank    AS real_hr_gms_won,
        pred_lower_rank     AS real_lr_gms_won
    FROM predictions
    WHERE pred_people_id = 'REAL'   -- bringing just the real result from the series
    )
```



```
-- Now let's join everything together, calculate the margin between the two average columns,
-- rank the output by this margin and leave the real scores for comparison:

SELECT
    RANK() OVER(ORDER BY ABS(avg_hr_gms_won - avg_lr_gms_won) DESC) AS 'Rank',
    t1.team_name                                AS 'Higher Ranked Team',
    t2.team_name                                AS 'Lower Ranked Team',
    avg_hr_gms_won                              AS 'HR AVG Bet',
    avg_lr_gms_won                              AS 'LR AVG Bet',
    ROUND(avg_hr_gms_won - avg_lr_gms_won, 2)   AS 'Margin',
    bets_qty                                    AS 'Bets Qty',
    real_hr_gms_won                             AS 'HR Real Score',
    real_lr_gms_won                             AS 'LR Real Score',
    rd_desc                                     As 'Round',
    series_season                               AS 'Year'
FROM series
LEFT JOIN avg_bets      ON series.series_id = avg_bets.series_id
LEFT JOIN real_score    ON series.series_id = real_score.series_id
LEFT JOIN rounds        ON series_round_id = rd_id
LEFT JOIN teams         AS t1 ON series_higher_rank = t1.team_id
LEFT JOIN teams         AS t2 ON series_lower_rank = t2.team_id
ORDER BY 'Rank'
```

The output:



Rank	Higher Ranked Team	Lower Ranked Team	HR AVG Bet	LR AVG Bet	Margin	Bets Qty	HR Real Score	LR Real Score	Round	Year
1	Golden State Warriors	Los Angeles Clippers	4	0,29	3,71	14	4	2	1st Round	2019
2	Milwaukee Bucks	Orlando Magic	4	0,5	3,5	16	4	1	1st Round	2020
3	Milwaukee Bucks	Detroit Pistons	4	0,79	3,21	14	4	0	1st Round	2019
4	Phoenix Suns	New Orleans Pelicans	3,84	0,88	2,96	25	4	2	1st Round	2022
5	Toronto Raptors	Orlando Magic	4	1,14	2,86	14	4	1	1st Round	2019
6	Toronto Raptors	Brooklyn Nets	3,94	1,31	2,63	16	4	0	1st Round	2020
7	Philadelphia 76ers	Washington Wizards	4	1,41	2,59	17	4	1	1st Round	2021
8	Brooklyn Nets	Boston Celtics	3,88	1,41	2,47	17	4	1	1st Round	2021
9	Milwaukee Bucks	Atlanta Hawks	4	1,59	2,41	17	4	2	Conference Finals	2021
10	Utah Jazz	Memphis Grizzlies	3,88	1,53	2,35	17	4	1	1st Round	2021
11	Golden State Warriors	Portland Trail Blazers	4	1,69	2,31	16	4	0	Conference Finals	2019
12	Milwaukee Bucks	Chicago Bulls	3,88	1,6	2,28	25	4	1	1st Round	2022
13	Houston Rockets	Utah Jazz	4	1,79	2,21	14	4	1	1st Round	2019
14	Los Angeles Lakers	Portland Trail Blazers	3,94	1,75	2,19	16	4	1	1st Round	2020
15	Miami Heat	Atlanta Hawks	3,88	1,72	2,16	25	4	1	1st Round	2022
16	Milwaukee Bucks	Miami Heat	4	1,87	2,13	15	1	4	Conference Semi-Finals	2020
17	Phoenix Suns	Los Angeles Clippers	3,94	1,82	2,12	17	4	2	Conference Finals	2021
18	Memphis Grizzlies	Golden State Warriors	1,92	3,96	-2,04	24	2	4	Conference Semi-Finals	2022
19	Los Angeles Clippers	Dallas Mavericks	3,94	1,94	2	16	4	2	1st Round	2020
19	Philadelphia 76ers	Brooklyn Nets	4	2	2	14	4	1	1st Round	2019
19	Houston Rockets	Utah Jazz	3,75	1,75	2	8	4	1	Conference Semi-Finals	2018