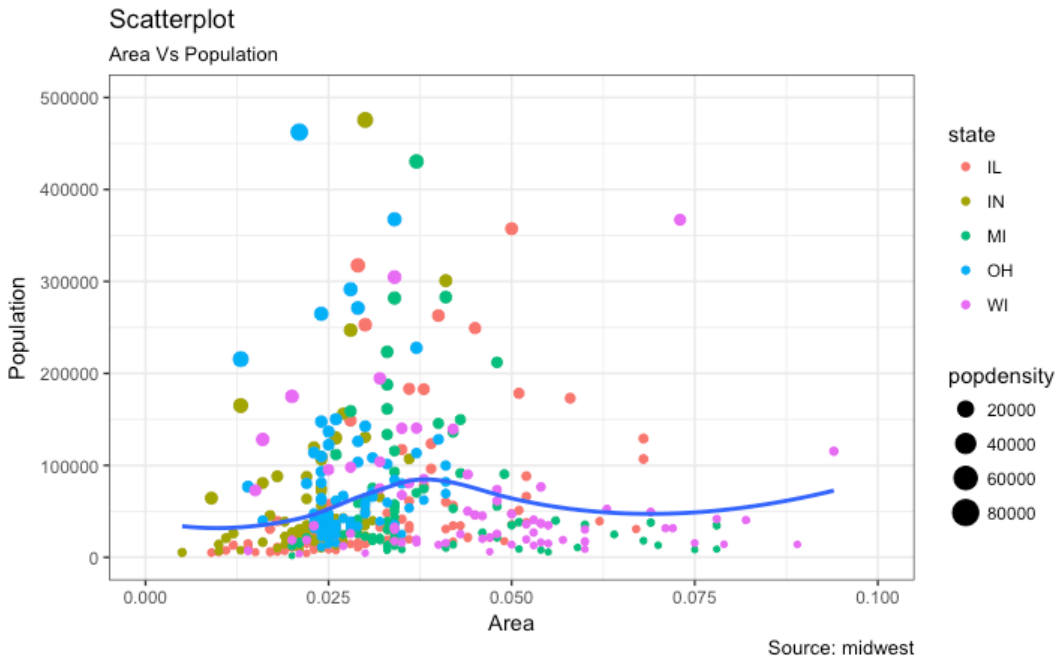
How to Transfer SQL Knowledge to R

towardsdatascience.com/how-to-transfer-sql-knowledge-to-r-e9ec951f33dc

Aaron Frederick February 20, 2019



Manipulating data with alternative syntax

Aaron Frederick

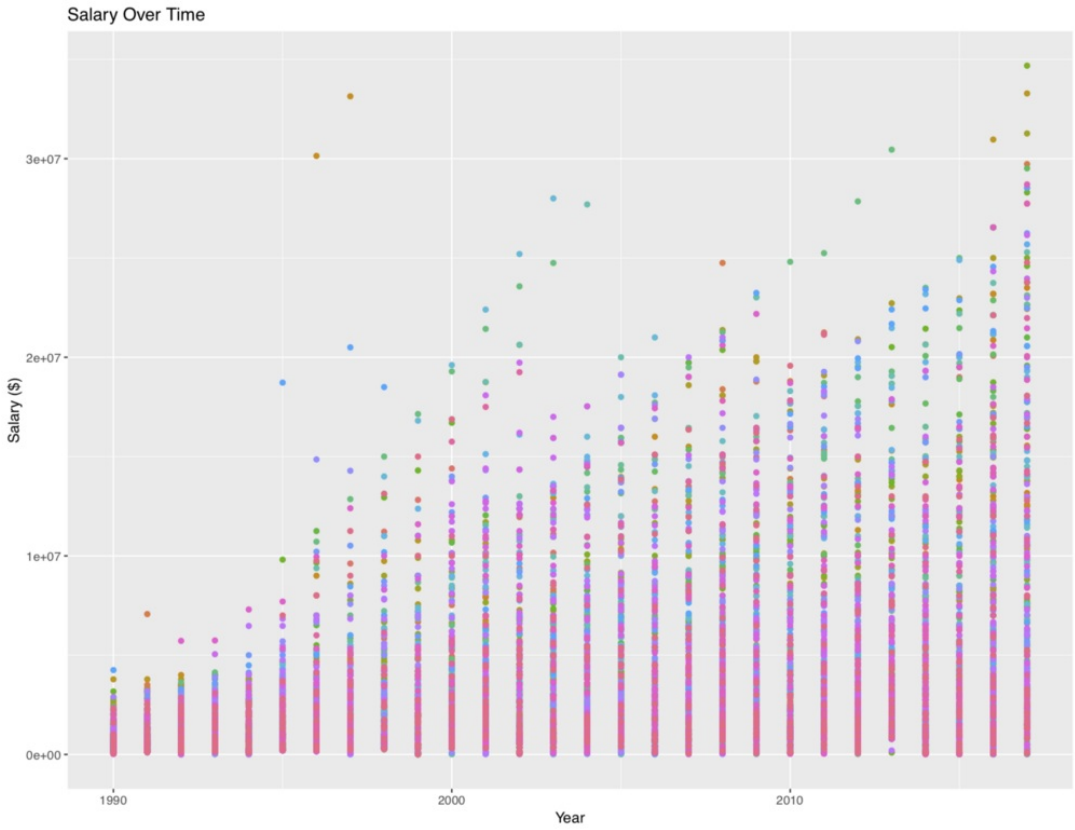
Feb 20 ★

Why would you ever use SQL syntax in the context of R? Maybe you took a few bioinformatics courses but then steered towards Python and are shaky coming back into the syntax of R. Maybe you want to expand your data science toolkit and are in desperate need of coding band-aids to keep you off Stackoverflow. Maybe you have already written the SQL query but would like to visualize the results using ggplot2. In an environment where data structures can often be confusing, SQL syntax can simplify dataframe manipulations in R so that the more time can be spent making great visualizations than on wrangling data. The code here can be found in its entirety on my GitHub.

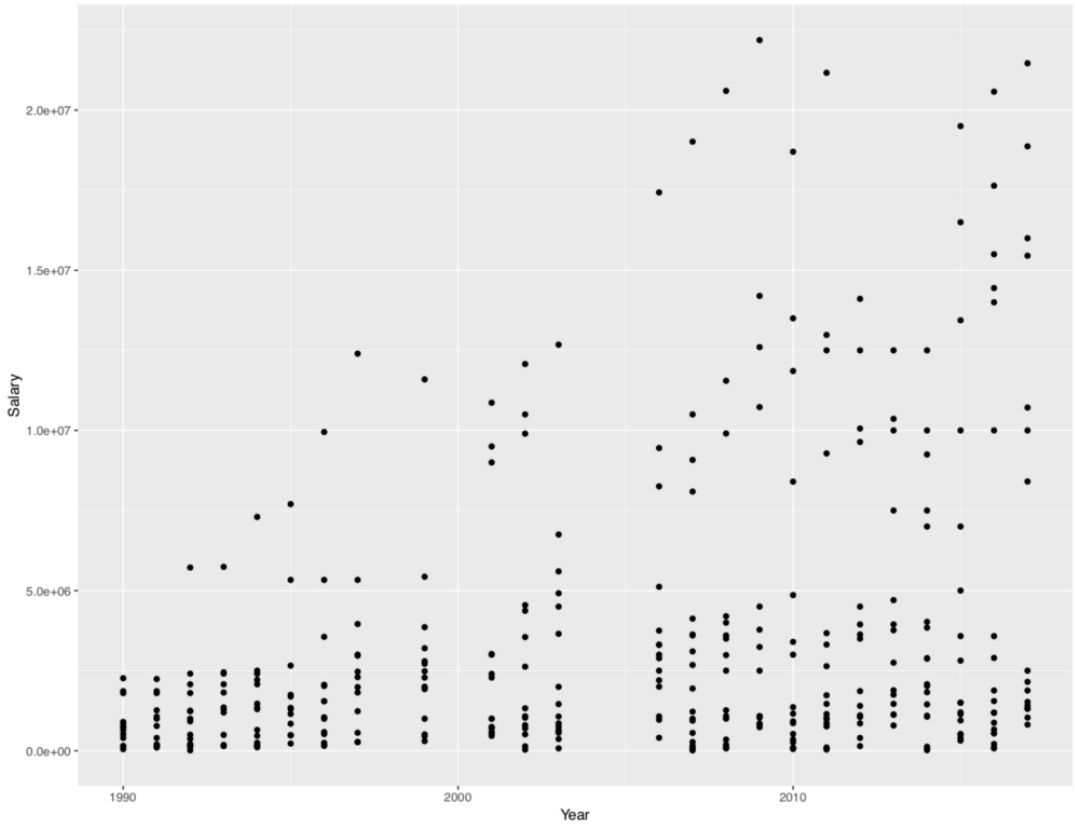
The sqldf package in R is a fantastic tool that is analyst-friendly. By simply writing a SQL query in a string, the function outputs a new dataframe according to that query. In the following blocks of code, I will share some basic restructuring tools using both the sqldf function and proper R syntax. The motivation for using R is that the ggplot2 package allows us to create visually appealing static plots with only a handful of lines of code. It is the

go-to R package for academic publishing and is very well supported and documented. I will be visualizing the results of the dataframe manipulations to show both the effectiveness of the package and the results of the queries. Below is the code I used to setup my data:

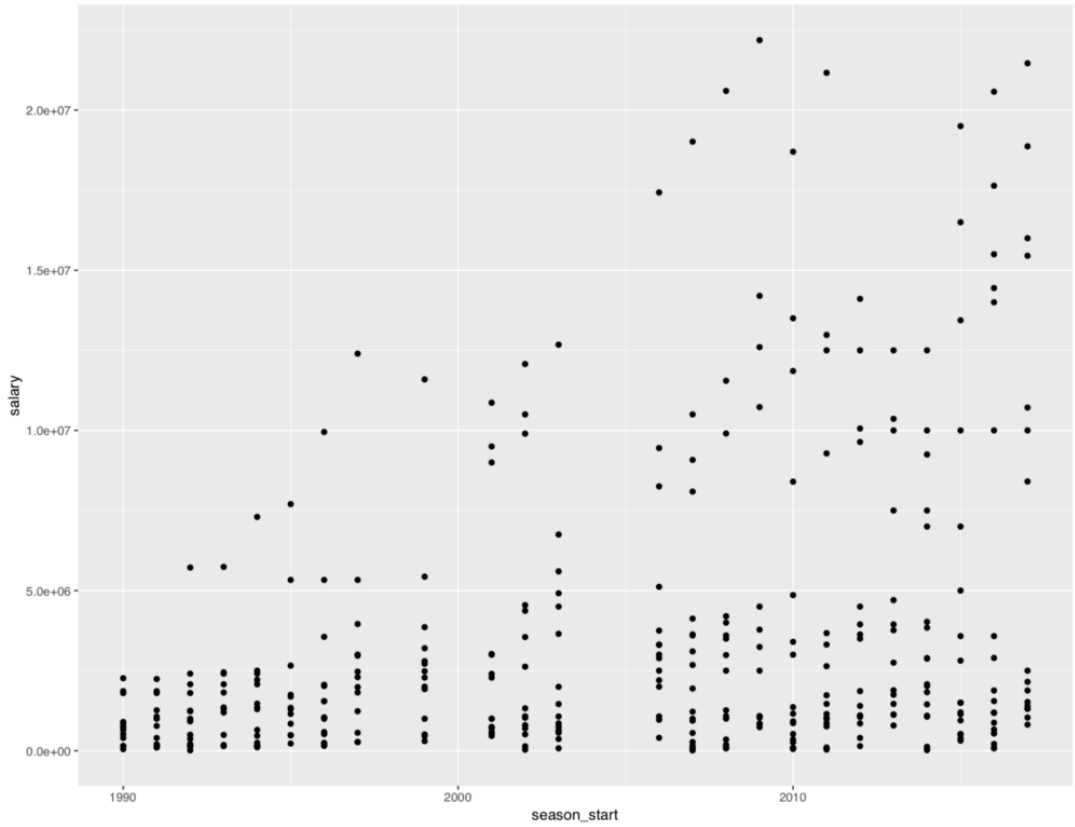
The dataset in use consists of basketball salaries from 1990–2018. The dataframe includes a player's name, their salary, the season start year, the season end year, their team's three-letter abbreviation, and full team name. The graph below shows all of the salaries color-coded by team (*because plain black scatter plots are no fun!*).



Now, my family's favorite team is the San Antonio Spurs, so it is of vital importance that I see only the salaries of the Spurs over the last 28 years. Using the SQL-styled query below, I will grab that subset and plot it.



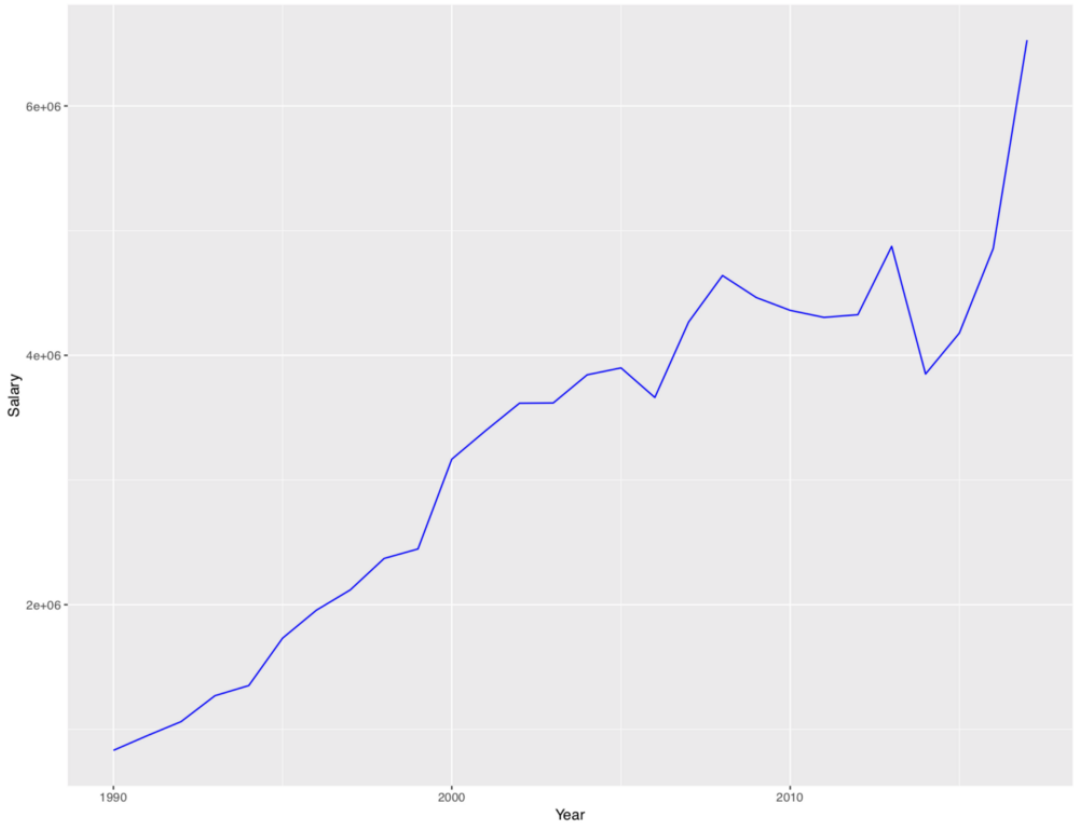
The same effect can be achieved using the subset function in R. The code and graph below show the exact same data as above.



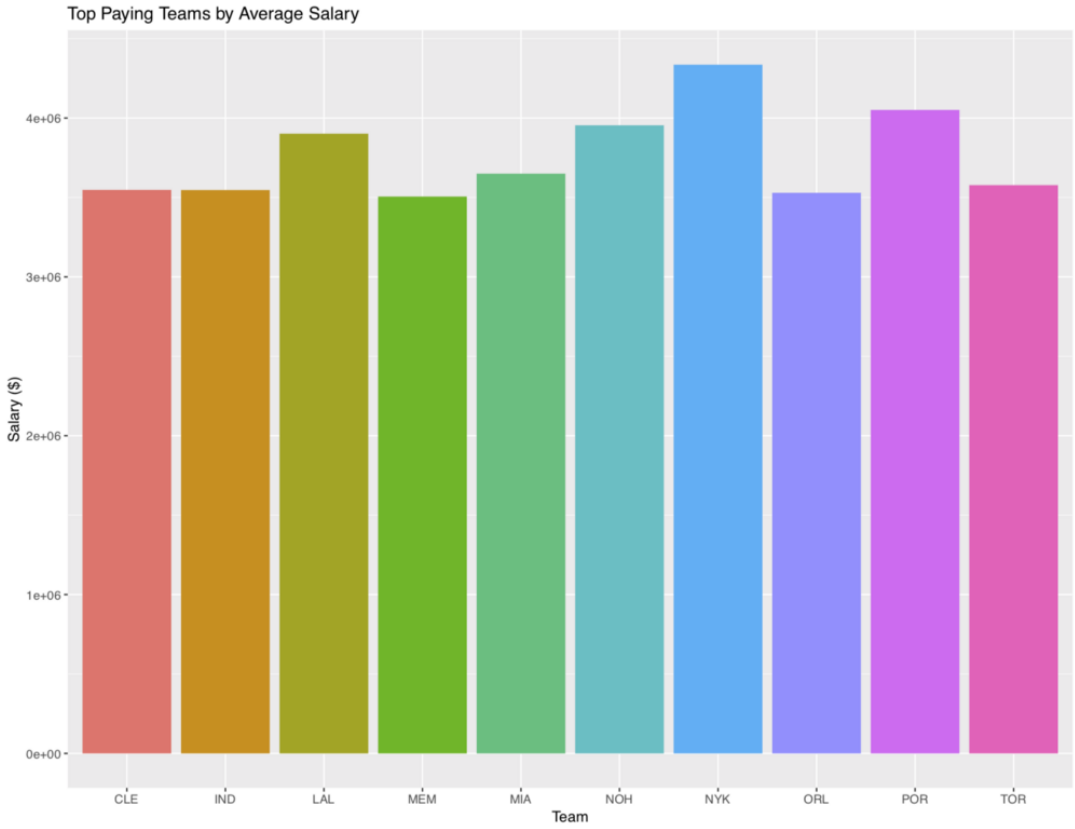
The only differences in the two plots are the axis labels. Since I renamed the column titles in the SQL query, they were automatically capitalized in the plot. Simply using the subset function would require an additional line of code in the plot function to change the axis labels.

Another use-case for SQL syntax is aggregating functions. If we wanted to

track the average salary over time, we could use either this SQL query or the aggregate function in R. The FUN parameter in the aggregate method specifies which function is to be applied when aggregating the data—in this case it averages the values. I will include both the query notation and the aggregate notation in the gist below—I promise the graphs are the same!



Finally, we'll put both of those concepts -together! To show the top 10 teams in terms of average salary over the past 28 years, we need a few lines of SQL: average the salary, group by team, then grab the top 10 in descending order. A relatively basic query, especially when compared to its counterpart with R syntax. In R, we must aggregate by team using the mean function, then reset the column names, and finally take the head of the dataframe once it is reordered by its second column. The SQL is much more readable and intuitive, especially to someone that is not an R expert. The code and plot below show the result of these equivalent manipulations.



Hopefully this is helpful for anyone with a SQL background trying to get into R for exploratory analysis; the syntax is much more straightforward and readable than the native R functions. If you enjoyed this article or any of my others, I would love to connect! I'd also like to thank Jonathan Balaban, the reason that I write about R.