

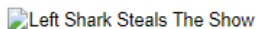
# Dev Nation

## Intro To DS lab 04

### 1. TV, halftime shows, and the Big Game

Whether or not you like football, the Super Bowl is a spectacle. There's a little something for everyone at your Super Bowl party. Drama in the form of blowouts, comebacks, and controversy for the sports fan. There are the ridiculously expensive ads, some hilarious, others gut-wrenching, thought-provoking, and weird. The half-time shows with the biggest musicians in the world, sometimes [riding giant mechanical tigers](#) or [leaping from the roof of the stadium](#). It's a show, baby. And in this notebook, we're going to find out how some of the elements of this show interact with each other. After exploring and cleaning our data a little, we're going to answer questions like:

- What are the most extreme game outcomes?
- How does the game affect television viewership?
- How have viewership, TV ratings, and ad cost evolved over time?
- Who are the most prolific musicians in terms of halftime show performances?

 Left Shark Steals The Show

[Left Shark Steals The Show](#). Katy Perry performing at halftime of Super Bowl XLIX. Photo by Huntley Paton. Attribution-ShareAlike 2.0 Generic (CC BY-SA 2.0).

The dataset we'll use was [scraped](#) and polished from Wikipedia. It is made up of three CSV files, one with [game data](#), one with [TV data](#), and one with [halftime musician data](#) for all 52 Super Bowls through 2018. Let's take a look, using `display()` instead of `print()` since its output is much prettier in Jupyter Notebooks.

### Look Into all the datasets using Display

### 2. Taking note of dataset issues

For the Super Bowl game data, we can see the dataset appears whole except for missing values in the backup quarterback columns (`qb_winner_2` and `qb_loser_2`), which make sense given most starting QBs in the Super Bowl (`qb_winner_1` and `qb_loser_1`) play the entire game.

From the visual inspection of TV and halftime musicians data, there is only one missing value displayed, but I've got a hunch there are more. The Super Bowl goes all the way back to 1967, and the more granular columns (e.g. the number of songs for halftime musicians) probably weren't tracked reliably over time. Wikipedia is great but not perfect.

An inspection of the `.info()` output for `tv` and `halftime_musicians` shows us that there are multiple columns with null values.

### 3. Combined points distribution

For the TV data, the following columns have missing values and a lot of them:

- `total_us_viewers` (amount of U.S. viewers who watched at least some part of the broadcast)
- `rating_18_49` (average % of U.S. adults 18-49 who live in a household with a TV that were watching for the entire broadcast)
- `share_18_49` (average % of U.S. adults 18-49 who live in a household with a TV *in use* that were watching for the entire broadcast)

For the halftime musician data, there are missing numbers of songs performed (`num_songs`) for about a third of the performances.

There are a lot of potential reasons for these missing values. Was the data ever tracked? Was it lost in history? Is the research effort to make this data whole worth it? Maybe. Watching every Super Bowl halftime show to get song counts would be pretty fun. But we don't have the time to do that kind of stuff now! Let's take note of where the dataset isn't perfect and start uncovering some insights.

Let's start by looking at combined points for each Super Bowl by visualizing the distribution. Let's also pinpoint the Super Bowls with the highest and lowest scores.

### 4. Point difference distribution

Most combined scores are around 40-50 points, with the extremes being roughly equal distance away in opposite directions. Going up to the highest combined scores at 74 and 75, we find two games featuring dominant quarterback performances. One even happened recently in 2018's Super Bowl LII where Tom Brady's Patriots lost to Nick Foles' underdog Eagles 41-33 for a combined score of 74.

Going down to the lowest combined scores, we have Super Bowl III and VII, which featured tough defenses that dominated. We also have Super Bowl IX in New Orleans in 1975, whose 16-6 score can be attributed to inclement weather. The field was slick from overnight rain, and it was cold at 46 °F (8 °C), making it hard for the Steelers and Vikings to do much offensively. This was the second-coldest Super Bowl ever and the last to be played in inclement weather for over 30 years. The NFL realized people like points, I guess.

*UPDATE: In Super Bowl LIII in 2019, the Patriots and Rams broke the record for the lowest-scoring Super Bowl with a combined score of 16 points (13-3 for the Patriots).*

Let's take a look at point difference now.

## 5. Do blowouts translate to lost viewers?

The vast majority of Super Bowls are close games. Makes sense. Both teams are likely to be deserving if they've made it this far. The closest game ever was when the Buffalo Bills lost to the New York Giants by 1 point in 1991, which was best remembered for Scott Norwood's last-second missed field goal attempt that went [wide right](#), kicking off four Bills Super Bowl losses in a row. Poor Scott. The biggest point discrepancy ever was 45 points (!) where Hall of Famer Joe Montana's led the San Francisco 49ers to victory in 1990, one year before the closest game ever.

I remember watching the Seahawks crush the Broncos by 35 points (43-8) in 2014, which was a boring experience in my opinion. The game was never really close. I'm pretty sure we changed the channel at the end of the third quarter. Let's combine our game data and TV to see if this is a universal phenomenon. Do large point differences translate to lost viewers? We can plot [household share](#) (average percentage of U.S. households with a TV in use that were watching for the entire broadcast) vs. point difference to find out.

## 6. Viewership and the ad industry over time

The downward sloping regression line and the 95% confidence interval for that regression *suggest* that bailing on the game if it is a blowout is common. Though it matches our intuition, we must take it with a grain of salt because the linear relationship in the data is weak due to our small sample size of 52 games.

Regardless of the score though, I bet most people stick it out for the halftime show, which is good news for the TV networks and advertisers. A 30-second spot costs a pretty [\\$5 million](#) now, but has it always been that way? And how have number of viewers and household ratings trended alongside ad cost? We can find out using line plots that share a "Super Bowl" x-axis.

## 7. Halftime shows weren't always this great

We can see viewers increased before ad costs did. Maybe the networks weren't very data savvy and were slow to react? Makes sense since DataCamp didn't exist back then.

Another hypothesis: maybe halftime shows weren't that good in the earlier years? The modern spectacle of the Super Bowl has a lot to do with the cultural prestige of big halftime acts. I went down a YouTube rabbit hole and it turns out the old ones weren't up to today's standards. Some offenders:

- [Super Bowl XXVI](#) in 1992: A Frosty The Snowman rap performed by children.
- [Super Bowl XXIII](#) in 1989: An Elvis impersonator that did magic tricks and didn't even sing one Elvis song.
- [Super Bowl XXI](#) in 1987: Tap dancing ponies. (Okay, that's pretty awesome actually.)

It turns out Michael Jackson's Super Bowl XXVII performance, one of the most watched events in American TV history, was when the NFL realized the value of Super Bowl airtime and decided they needed to sign big name acts from then on out. The halftime shows before MJ indeed weren't that impressive, which we can see by filtering our `halftime_musician` data.

## 8. Who has the most halftime show appearances? ¶

Lots of marching bands. American jazz clarinetist Pete Fountain. Miss Texas 1973 playing a violin. Nothing against those performers, they're just simply not [Beyoncé](#). To be fair, no one is.

Let's see all of the musicians that have done at least one halftime show, including their performance counts.

## 9. Who performed the most songs in a halftime show?

The world famous [Grambling State University Tiger Marching Band](#) takes the crown with six appearances. Beyoncé, Justin Timberlake, Nelly, and Bruno Mars are the only post-Y2K musicians with multiple appearances (two each).

From our previous inspections, the `num_songs` column has lots of missing values:

- A lot of the marching bands don't have `num_songs` entries.
- For non-marching bands, missing data starts occurring at Super Bowl XX.

Let's filter out marching bands by filtering out musicians with the word "Marching" in them and the word "Spirit" (a common naming convention for marching bands is "Spirit of [something]"). Then we'll filter for Super Bowls after Super Bowl XX to address the missing data issue, *then* let's see who has the most number of songs.

## 10. Conclusion

So most non-band musicians do 1-3 songs per halftime show. It's important to note that the duration of the halftime show is fixed (roughly 12 minutes) so songs per performance is more a measure of how many hit songs you have. JT went off in 2018, wow. 11 songs! Diana Ross comes in second with 10 in her medley in 1996.

In this notebook, we loaded, cleaned, then explored Super Bowl game, television, and halftime show data. We visualized the distributions of combined points, point differences, and halftime show performances using histograms. We used line plots to see how ad cost increases lagged behind viewership increases. And we discovered that blowouts do appear to lead to a drop in viewers.

This year's Big Game will be here before you know it. Who do you think will win Super Bowl LIII?