# Upsets in Tennis

DSC 530 Term Project
Christopher Briggs
May 30, 2020

**Statistical/hypothetical questions.**
The statistical question to be answered was: are upsets more common in women's tennis than men's? Do upset rates vary by surface?

**Outcome of the EDA.**
I observed that based on the scatterplots, the distributions for men and women, and across surfaces, were basically indistinguishable. Carpet has disappeared from use, while the other three surfaces are still in active use. The outcome of the study is that gender and surface were significant in predicting the outcome of a match based on rank difference. After controlling for tournament level (at least Masters), the mean rank difference among men was 31.25 and among women was 31.77. This difference of means was found to be insignificant with $p = 0.568$. The mean rank difference on hard courts was 32.2 and on clay courts was 28.8, and this difference was significant with $p=0.0$. Similar differences were found with other surfaces.

**What was missed during the analysis?**
I identify below that the analysis was limited as described by the variable and assumption limitations. So whatever improved resolution could have been gained from that further detail was missed.

**Were there any variables that could have helped the analysis?**
The courts have speed ratings. The surface type is only a rough approximation of the speed (Jonathan, 2020). Even the weather conditions can have a large impact on the court behavior; for example, the Madrid clay court is one of the fastest clay courts because of the altitude which causes low air pressure and humidity. So having the individual court CPIs and the weather conditions would have helped the analysis.

**Were there any assumptions made which were maybe incorrect?**
Rank difference as a computed variable does address the differences on the courts or between the genders, but is unlikely to capture the whole story. For example, if #1 plays #41, the rank difference is 40; this pairing has a more certain outcome than if #201 plays #241. In truth the proper variable is probably somewhere between the difference and the ratio.

**What challenges did you face, and what did you not fully understand?**
The biggest challenge was how to fit the list of requirements into the question I was trying to answer. I'm used to using whichever tools are necessary to answer a question, rather than to forcing the use of a tool into a solution in order to demonstrate my proficiency with the tool. So this project had us reconciling the dual goal of exploring the hypothesis while trying to fit every needed graph into the analysis. For the analysis, I was grateful to have the book as a reference. The code in thinkstats.py and thinkplot.py felt loosely organized to me as I went through the exercises, but upon revisiting the code I could really appreciate the naming conventions and organization.