Eli Metzner
CPSC 446

Assignment 6 Writeup

1)

The yearly NCAA Division I Men's Basketball Tournament, also known as March Madness, is (in)famously one of the most unpredictable and exciting events in sports. Though the single-elimination tournament is consistently filled with upsets and subject to the randomness inherent in having a season decided by only one game, millions of people every year attempt to predict the outcome of each of the 68 tournament games, filling out tournament brackets and entering "bracket pools" in which the bracket with the most correctly predicted games (usually scored by weighting later tournament rounds with more points per game) wins the pool. Most pools involve a monetary buy-in, with the winner collecting the proceeds – indeed, billions of dollars each year are exchanged on March Madness sports betting, within more informal bracket pools as well as at Las Vegas sportsbooks. Luckily, basketball is a relatively data-rich sport, and readily available statistics detailing games, teams, and players can lend insight into prediction of tournament game outcomes.

Broadly, our domain situation for this visualization is understanding the predictive value of individual statistics at the game level and over time, with the goal of lending insight into prediction of future NCAA Tournament games. More concretely, we're trying to answer the following questions through our faceted visualizations:

- Which game-level statistics have the most predictive value for determining the winner of a given game?
- How have the predictive values of these statistics changed over time?
- Do correlations exist between seemingly unrelated statistics that could lend further insight into their predictive values?

2)

To answer our questions, we work with a dataset aggregated by and available from the data science website Kaggle – the dataset lists the total box-score statistics for the winning and losing teams of every NCAA Tournament game for the past 16 Tournaments (the 2003-2018 tournaments -- 2019 tournament statistics had not been aggregated in time to include). For time relevance and reduction of visual clutter, we decided to filter the dataset to only include the past 10 years of data (tournaments from 2009-2018). Since we're interested in the statistical differentials between winning and losing teams, rather than the absolute values of the statistics themselves, we needed to create derived attributes for each of the statistics by subtracting the losing team's values from the winning team's for each of the 12 statistical categories visualized. This single transformation created a dataset sufficient to answer questions 1) and 3) above, but to examine the time variation of each of these differentials, we also had to aggregate each game-level attribute to a season-level average, which we could then visualize as time series. To ensure that the data played nicely with D3, we also "melted" the data from matrix form (with each column representing an attribute, and each row a game) to key-value form (where each entry represents a specific statistic-differential pair for each game, where the statistic is the key and the differential in that game is the value). All data manipulation was done in R, using the *dplyr* and *reshape2* packages. The R script used to transform the original dataset is attached to the submission.

3)

At the encoding level, we faceted the dataset into views meant to specifically answer one of the three questions in the domain situation:

- We use the heatmap idiom to visualize the overall predictive value of each statistic at the individual game level: in the heatmap, one categorical key and one ordered key are encoded via the spatial position channel within a matrix, and one quantitative value encoded via the color channel in each cell. Individual games make up the ordered key, which is encoded by the y-position of each cell, and the statistical differentials are the categorical key, represented by the x-position of each cell. The quantitative value – the magnitude of each statistical differential – is encoded via the color of each cell based on a standard diverging colormap from colorbrewer2.org. The hue channel encodes the sign of the differential, while the luminance channel encodes its absolute value.
- To determine whether the predictive value of each statistic has changed over time, we use a series of juxtaposed area charts to visualize the change in season-level average values of game-level differentials from 2009 to 2018. Superimposing 12 line charts on the same graph was also considered, but given that this visualization is meant for the global task of comparing the slopes and values of each differential across the entire time period – and keeping in mind that there are 12 separate time series to visualize – we follow the guidelines of Javed et al. (Munzner Chapter 12) and use the juxtaposed area charts idiom instead. In this idiom, the y-position of the area encodes the average value of each statistic, and the x-position encodes time. Color and text channels are used to encode the identity of the time series displayed by each chart.
- To look at correlations between statistical differentials, we use a scatterplot matrix to plot the differential values of the selected statistic in each game against corresponding game-level values for all 12 statistics on faceted scatterplots (note that this visualization is interaction-dependent, as it requires the user to select a statistic before the plots are generated). The value of each differential is encoded by the x- and y- position of the points in each scatterplot, and the season of each game is encoded via the luminance of the point (using a single-hue colormap again from colorbrewer2.org).

We use a slider that allows the user to select and highlight individual statistics as the interaction idiom in this visualization, combined with an overview-detail navigation idiom between the faceted views. Highlighting is linked between the first and second views, and the luminance channel is used for the linked highlighting: decreasing the luminance of all items not matching the selected statistic creates significant visual popout for the statistic selected by the user for further investigation. The scatterplot matrix then generates a detailed view of the overview provided by the first two idioms after the user interaction, taking the user's selection and generating scatterplots between the selection and every other statistic for the detailed investigation of correlations and trends between statistical categories at the game level.