Phil Robinson
618 Final Project Proposal

The dataset that I have chosen to work with the final project is about the 2005 college football season, as supplied to cfbstats.com and coachbythenumbers.com, with the direct link to the url at the bottom of this file. Although more recent data would be more fun and interesting to look at, even their most basic package of recent data is priced at $350.

The dataset is a relational database of CSV files, all coming in to populate 17 tables. While some of the tables are straight forward (the game-statistics table gives the game_code, attendance, and duration), others are much more complex (Player-game-statistics contains more than 50 columns) or contain a significant number of rows (the Play table, which details every play of every game contains more than 137,000 instances). The central table of the dataset tables is a play, with branching out relation sets diverging from there. In all, this is a very rich dataset, filled with lots of answers to questions.

Questions

1. How does rate of fumbles by freshman running backs change in comparison to sophomores, juniors and seniors from week to week or month to month. This is almost a proxy question of: do freshmen get “better” more quickly than more seasoned players. I also propose looking at the same trend, but for freshman through senior quarterbacks and their interception rates, but freshman are unlikely to have a starting quarterback position, this insight might yield less dependable results (low population size), but could still be interesting to look at.
2. How does the Big Ten compare to other conferences in terms of defensive statistics, such as forced fumbles, sacks, limiting offensive time of possession, etc. The Big Ten is generally known as a strong defense conference, so this would be interesting to see the actual statistics.
3. In the 4th quarter, is there a linear relationship between how many points a team is down compared to the number of pass plays that the team runs when they are on offense. I believe a team down by a lot would run a more risky offense (more passes). In addition, at what point do we see a point differential as “insurmountable”, and the offense stops running a pass offense. These would show up as outliers.
4. How do teams change the way they play when they are facing a rival as opposed to a program that isn’t a rival. Do they try more rushes or passes? How does the defense change? Rivalries are not defined in this dataset, so I will have to self-define rivalries for this dataset. For particularly interesting examples, such as Michigan and ohio state (not capitalized on purpose), I would want to look at, for example, if Michigan is running the ball more because the data from ohio state’s games show they have a weak run defense.

Descriptions

1. A seaborn line graph of the data points would be best for this question. Because we are asking about data that changes over time, we would want to to know how are points are changing, thus line graphs would be optimal, grouped by class and time period, allowing for slope comparison.
2. Comparison between conferences could be best visualized via boxplot in seaborn. Comparison of the defense categories would contain a box plot for each conference.
3. This data would most likely best be described by a scatter plot of each offensive drive by the losing team, with each point being a representation of the drive. Each point is the amount of passing plays and points the team is down by, and then have a linear regression from the points, including the pearson's correlation, for seaborn
4. This data would again be best visualized via a box plot in seaborn. Because we want to know if these data are really different, a box plot (and ANOVA analysis as well), would tell us how different a team might play a rival as opposed to a non-rival.

http://coachesbythenumbers.com/wp-content/custom-php/sportsource-data-sample-advanced.zip

Whats your favorite college football rivarly, outside of The Game?iron bowl