Dan Fanelli	Prof. Andy Catlin
CUNY DA 607 - Final Project Proposal	4/16/2016

Performa Data Analysis of Historical NFL Point Spreads

1. **Motivation:** I'm the administrator of an NFL "Pick Em" every year. I always tend to place in the top 3, year after year, using a simple formula: For the sake of curiosity, I would like to know if my "system" is a signal, or if it is noise. Would the strategy hold up historically? Extra motivation: There "might" be some monetary incentive for someone capable of picking NFL games consistently against the spread :) The main feeling is this: Bookies create spreads NOT based upon what they think the score will actually be. The create spreads based upon what will cause 50% of the betting population to bet on team A, while the other 50% bet on team B, guaranteeing them a "take" each week. If the general public is "betting with their hearts", as man do, then gaining a sustainable advantage through analytics should be possible.
2. **ESEMN Workflow:**
    a. **Obtain:** websites such as http://www.footballlocks.com provide historical scores. I also plan to retrieve information about the host cities themselves, and perhaps some type of sales information to score a team's "popularity" - theory being that teams that sell more jerseys, for example, would have more "heart betters"
    b. **Scrub:** Since the data will be coming from a couple data sources, it will need to be collected and joined into a single data frame.
    c. **Explore:** I will brainstorm and see what data is available that could possibly contribute to bias in betters choices
    d. **Models:** I will explore multiple models to see which seems to fit best
    e. **iNterpret:** If a general strategy becomes apparent, I will look at outliers and see if there are tangible reasons why these games did not hold to the pattern.
3. **At least 2 Data Sources:**
    a. Scrubbing http://www.footballlocks.com/ for historical spreads and scores, or perhaps some other data source
    b. At least 1 CSV or relational data source (perhaps for city or sales information)
4. **Data Transformation:** The data will have to at least be merged into a single data set to do the analysis and graphs
5. **Statistical Analysis** will be performed to rank which spreads seem to be the most biased. **Project will use graphics** to describe or validate the data.
6. Project will include graphics that show how some attributes affect the spread bias. (or how none really do)
7. **A graphic** will be used to support the conclusion of whether a reliable strategy was found.
8. A statistical analysis will be used to support the above conclusion (that there are or there are not attributes that can be used to gain an advantage in NFL spread picks)
9. A feature not covered in class will be used (**TBD**)

…
(The rest of the requirements dealing with presentation, etc.)