**Brian Gordon**  
[bgordon002@regis.edu](mailto:bgordon002@regis.edu)   
352-409-4082

* Title of the project

Predicting MLB team win totals, post-season eligibility, and individual player salaries

* High level description of the project: what question or problem are you addressing?

1. Can team win totals be predicted based on team statistics?
2. Can a playoff team be predicted based on team statistics?
3. Can a player’s salary be predicted based on their individual statistics?

* What type of data science task is it? (some example answers but not limited to)
* EDA
* Visualization
* Prediction of win totals with regression ML
* Prediction of post-season status with classification ML
* Prediction of player salary with regression ML
* Data: Brief description of data. How big do you expect the data will be? Is amount of your data too big or too small? If you're web-scraping or collecting data, how long do you expect to collect the data?
* Team statistics
* Batting data
* Fielding data
* Pitching data
* Data has been obtained as CSV files from **SeanLahman.com**
* How will you analyze the data? What machine learning methods do you plan to use, and/or what business intelligence aspect do you plan on incorporating?
* Following EDA, data will be presented in text format as well as visualized to show findings (groupings, patterns, comparisons, etc.)
* Predicting win totals & salaries will be done with linear & non-linear regression algorithms
  + Players will be grouped by position
  + Additional groupings will be considered (team, division)
* Predicting playoff status will be tested with various classification algorithms
  + Target variable will be whether team made playoffs
  + Will use both supervised & unsupervised methods to identify playoff teams
* Describe any anticipated difficulties and problems. Discuss how you may overcome the problems.
* Multicollinearity will likely be present due to related statistical categories
  + Perform feature selection
* Low performance regardless of models chosen
  + Delete features
  + Scale data
* Suggest a timeline for the project.  This should be a weekly breakdown of what you plan on doing each week.
  + **Week 2** – Obtain & prepare data – Download CSVs, upload to Jupyter Notebook, concatenate where necessary, add target variable(s)
  + **Week 3** – EDA & Visualization
  + **Week 4** – Testing of ML algorithms
  + **Week 5 –** Continued ML algorithm testing; try with scaled data, feature selection
  + **Week 6 –** Implement ‘best’ algorithm from previous weeks, finalize Jupyter Notebook, writeup of results
  + **Week 7 –** Finalize Github repository & record presentation
  + **Week 8** – submit project materials
* Create GitHub repository for your Practicum project. Add this proposal, begin a ReadMe document, and begin adding your data to your repository. Add a link to your GitHub repository to this document.

<https://github.com/bgord623/Practicum1>