**Overview**

In this project you will be applying the skills you have learned in the class so far. More specifically, you will be asked to apply an clustering algorithm to a dataset of your own choice.

As you are choosing your data set, make sure that you think about the following important items:

1. What is the business problem you are trying to solve? In one or two sentences explain the problem you are trying to solve. Make sure that this problem is specific enough! For example, if you are working with baseball MLB data:
   * **Bad Problem Statement:** Understanding baseball players.
   * **Better Problem Statement:** From demographic info and statistics of MLB batters, can we detect hidden groups that would allow an MLB team to diversify their roster and cancel the contracts of players already redundant?
2. What is the data available to solve this business problem? Why do you think that this dataset can answer the problem at hand? Have you made enough research on the problem and the dataset?
3. Double check the source of the dataset. Make sure that you have done enough data exploration so there are not wrong entries, lots of duplicates, mysterious columns, nonsense statistics, etc. Don't get unpleasantly surprised after you have already invested 3 weeks in modeling. Invest in exploration of your data in the beginning of the project. Make sure that you didn't set yourself up for the [garbage in garbage out](https://en.wikipedia.org/wiki/Garbage_in,_garbage_out) problem
4. Prepare your data and make sure that you have created interesting and insightful new features that will improve the machine learning algorithm's performance. (Check [Feature Engineering](https://en.wikipedia.org/wiki/Feature_engineering))
5. Decide which clustering algorithm you would like to use. Make sure to choose this algorithm because it is the best fit for this project but not because it sounds like the fanciest algorithm (e.g [KMeans](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html) if you want to use [DBSCAN](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html#sklearn.cluster.DBSCAN) then, for example,you should know what the epsilon parameter is and why DBSCAN is preferable to KMeans in your problem.) What is the metric you will be focusing on to measure this algorithm's performance?(e.g. [silhouette\_score](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html)) How can you communicate this technical metric with your business partners?
6. Suppose you settled on a particular clustering with a certain number of clusters. Now go back and explain what these clusters correspond with in terms of business metrics. You can create plots and visualizations to support your findings.
7. Finally, summarize your findings and the limitations of your findings in a conclusion. Also mention how you would proceed if you had more time and resources on this project.

**Deliverables**

* A Jupyter Notebook (I will refer it as technical notebook or report). This notebook should be in a Github Repo with a ReadMe. As previous homeworks, you will be submitting the link of this notebook.
* This time I will not be grading the ReadMe. I assume that you have learned how to submit a good project repo with a good ReadMe in your previous projects. Please practice these learned skills.

**Deadlines**

Submit the completed deliverable by November 15th at 11:59 PM. Use the link below for the submissions.