1. What are the potential target variable(s) for supervised learning in the LPGA data set?

Based on my brief understanding of supervised learning and golfing ([Golf ExplanationLinks to an external site.](https://www.youtube.com/watch?v=MSzJDo9WJlE&ab_channel=HandicapGolf)), I believe the target variables to understand a player's performance are:

* average drive (yards)  - to understand their ability to take the first shot
* percent of greens reached in regulation - Since greens are the place of the final goal, knowing how often they can reach that place can help
* average putts per round - how many shots per round to analyze their level as a player
* percent of sand saves - ability to get a ball out of a bunker
* prize winnings - by prize calculations can be done
* average percentile in tournaments - performance across tournaments

2. How would you use unsupervised learning to create groups of golfers that are different as groups but similar to each other?

I would use one of the clustering algorithms like K-means, Hierarchical or mixed Gaussians. I would start off with k-means as it is easy to implement.

3. What are some interesting data visualizations on this data set?

As mentioned by others, would probably use the following:

* Scatter plot
* Pair plots
* Histograms
* Box plots