Learning goals

1) Appreciate what the different tools do, and choose which ones I might like to use or learn more about.
2) Have the confidence to find and start using any tool suited to a particular application, and be limited only by my algorithms / logic, and not the usage of tools.

Two aspects that are crucial to performing a good analysis:

1) Dataset
    1) Characteristics of a good dataset:
        1) Has potential for real world applications / impact, or provides useful insight
        2) Large “enough”
        3) Complete and consistent
        4) Unbiased, random sample
    2) Example: temperatures of cities dataset (from Data Analysis Techniques class)
2) Plan of what you will do with, or want from your data
    1) Hypotheses / stories you might want to tell (may evolve as you analyze the data)
        1) Places get colder from south to north.
        2) Regions play an important role in determining temperature.
    2) Questions to ask (either independently, or to validate / invalidate your hypotheses)
        1) Does longitude affect temperature significantly?
        2) What are the coldest and warmest cities in the US?
        3) Does being coastal or non-coastal affect the temperature?

Dataset used in class

1) Information available (fields)
    1) Home team, visiting team, home score, visitor score, year, week, point spread.
2) Applications, hypotheses, other interesting questions to ask
    1) Predicting outcome of matches (betting)
    2) Writing a sports article - say the story of a team that outperformed all expectations
    3) Is the point spread accurate?
    4) Do teams perform better in home matches than away matches?
    5) Are some teams more dominant than others? What makes them so?
    6) Does the (year, week) of the match matter - do teams get better or worse with time?

Simple statistics (all database type questions) computed in class

1) Number of matches played.
2) Number of distinct teams.
3) Number of matches played by different teams.
4) Average home scores of teams.
5) Number of matches played by particular home - visitor team pairs.

Other things you could try analyzing

1) Point difference - Point spread
2) Given a pair of teams, predict their outcome?
    1) Using only bookies’ point spread
    2) Using past history of matches
    3) Using past point spreads + past history + current point spread
    4) What if a pair of teams don’t have any shared history? Use common teams they have played against?