This looks good, but you should start looking at the data in detail very soon, as what you find may impact the feasibility of what you have proposed. I am also concerned that your project does not have enough depth. You propose a number of possible queries, but you need to do some analysis beyond queries - something that requires writing imperative SQL. Plan on taking a look at the data and submitting a revised Part 2 based on what you learn by 10/17. I will award you 10 points towards this grade when that is successfully completed. In particular, I am looking for some ideas about more in depth analysis that you will be doing.  
  
Each task is required to have a primary owner, even tasks that the team will work on. You may break up the ERD and Data exploration tasks into components (e.g. ZY will lead the modeling and exploration with a reviewer focus. YX will lead the modeling and exploration with a restaurant focus, etc.)

Our project has four goals:

* Initial data explorations (XY)
  + What are the top 10 cities which have the most restaurants?
  + What are the restaurant categories in each city?
  + What is the highest rated restaurant in each city? The result might be visualized to be more vivid. The rating is taken directly from the business profile for now.
  + What is the average ratings of all restaurants in each city?
  + What is the restaurant with most reviews of each category in each city?
  + Compare the rating of each restaurant by the data from its own profile and calculation result from reviews. This cross validation will check the yelp rating strategy to some degree.
  + What latitude ranges (no more a certain degree, say, 20 degrees) have the most restaurants? If there are multiple ranges with the same restaurant number, find the one with the highest average restaurant rating. (YZ)
* Restaurant recommendation
  + Identify personal preferred restaurant categories (say 3), called set U (XY)
    - This can be identified by counting the high rating reviews a user gives for each category. The more high ratings a category receives, the more preferred a user is.
  + Identify recommended restaurants by neighborhood (LW)
    - Identify user neighborhood
      * The neighborhood information is provided in the business table only. We find the user neighborhood by referring to the restaurant neighborhoods frequented by this user.
    - Calculate the *category popularity index* for each restaurant category, pick top K categories as neighborhood N’s most popular restaurant categories, called set P
      * Restaurant Popularity index = stars \* weightR,
      * Where weightR = f(review\_count)
      * Category Popularity index = ,
      * Where weightC = f(number of restaurants in this category)
    - Calculate *similarity index* (preference overlap rate) between a user’s preference set U and neighborhood recommendation set P
      * Similarity index = |P∩U|
    - Aggregate similarity index over n randomly chosen users to compute the final *robustness index* for recommendation by neighborhood
      * Robustness index =
      * The higher this index is, the more accurate recommendation by neighborhood is.
  + by friend (YZ)
    - Find friend cycle (connected component)
    - Find top K preferences in a user’s friend cycle
    - Calculate *similarity index* (preference overlap rate) between a user’s preference with his friend cycle’s top k preferences
    - Aggregate similarity index over n randomly chosen users to compute the final robustness index for recommendation by friend
  + Compare similarity index through the two methods
    - understanding how the genders of the customers that affect their reviews and tips.
* fraud users identification (XY)

1. **To identify the fraud users**

Before we start analysing the data, we need to decide the restaurant of interest. We will first focus on the most popular restaurant obtained from section 3-1-d) because the popularity enable it has more user data points that can provide more subjective results.

* 1. **How many review that the restaurant have each day?**
  2. **What is the averaged review times that restaurant have each day?**
  3. **How many checkin that the restaurant have each day?**
  4. **What is the averaged checkin times that restaurant have each day?**
     1. We need to know the averaged checkin time and review time to identify the dates of interest
     2. We may not consider the dates if they have few review/checkin times
  5. **What is the averaged stars of a restaurant in each date (in the dates of interest)?** 
     1. The averaged stars are calculated based on the given star in each date from ‘review.json’
  6. **How does the averaged dately stars of a particular business change with time?**
     1. By joining ‘review.json’ with ‘business,json’, we can average starts of a restaurant in each date since the opening date
     2. The stars-dating plot shows how consistent of the product quality in the restaurant is
  7. **Which customer (user\_id) gives the star that is 1 point higher or lower than the average star at the date of interest?**
     1. We can compare star given by each user with the average with the averaged star in that particular date
     2. If the difference between given star and the average star out of 1, we believe the user is suspicious - giving abnormal ratings

We will later on employ our analysis method to the restaurants of interest.

1. **Understanding how the genders of the customers that affect their reviews and tips**
   1. **What is the gender distribution of customers?**
      1. We will start with the restaurants in 3-1-d) which may have the most users
   2. **How many male customers give tips?**
   3. **How many female customers give reviews?**
      1. We are interested about whether male customers prefer give tips or reviews
   4. **How many compliments (including hot, cute, plain, cool, and funny) that given by each male/female customer to the restaurants they have visited?**
   5. **What is the averaged compliments of a male/female customer?**