**Proposal Rubric**

| Proposal Rubric | | |
| --- | --- | --- |
| **Criteria** | **Ratings** | **Pts** |
| Summarize and motivate your proposed project: What is your proposed project and why are you proposing it? What are the question(s) you want to answer, or goals to achieve? | I wonder if you can account for changes over time.  From the time 1879 to 2014 can be accounted as the data sets contain the information for that time range only. | 8 / 10.0 pts |
| Choose and describe two different data sources you plan to access and manipulate. The data sources should require different access mechanisms and/or different data formats. (For example, you might pick one data source that uses a JSON API, and the second might use SQL.) You should give details on the size, location, important variables, format and other key aspects of each dataset. | It would be helpful to have the URL for the data sources. I am assuming that you can get the SQLite databases directly from the data.gov site, but I'm not sure exactly from where.  <https://www.kaggle.com/kaggle/us-baby-names>  https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-level-data | 16 / 20.0 pts |
| Describe how you'll need manipulate the datasets: what initial processing will have to be done on each? And once processed, how will you combine the datasets, and what will be produced as output? What new information will result from combining them? | As you'll learn in this week's lecture, you don't need to convert SQL to CSV -- you can access the variables directly! How will you handle missing data? What about names that are minor spelling variants of others (e.g. Chris/Kris/Christopher/ Kristoffer)?   * Extract required data from sqlite3 database using methods learnt in class. * Extract data from csv file using csv package in python * The data set is already clean hence there are no missing data rows. * Names with different spellings would be treated as a different name altogether (even if it is a variant of others). | 15 / 20.0 pts |
| Describe one interesting visualization you might make based on your final output dataset (that would not be possible with either of the original datasets alone). | How would you choose the names to represent? There must be thousands of names -- will you generate only the most popular? The most popular in a state? The ones that are "significantly" more prominent (i.e. occur more often than expected) in a particular state?  -Plot graph of the most frequent boy /girl name in the last ten years at the national level i.e. from 2004 to 2014 | 8 / 10.0 pts |
| Total Points: 47 | | |