What kind of cleaning steps did you perform?

1. The column headers in the data files are different. Hence the same column names were used to identify the dataset
2. In the 'ALL_RATE' column there are lot of Non Integer values used for privacy protection. Hence a round up strategy was used where the values given in a range are remapped to a value which is the higher range
3. A "PS" notation in the data file identifies when the number of COHORT is within 1-5 students. Hence the 'ALL_RATE' column is substituted with 0

How did you deal with missing values, if any?

1. The missing values are replaced with 0
2. The cells with a value as '.' Are also replaced with 0

Were there outliers, and how did you decide to handle them?

1. The records with 'ALL_RATE' as 0 are ignored during the analysis

During data cleanup, I have used a random generator to replace the suppressed and missing values in the adjusted cohort graduation rates dataframe. This was done by generating a random number between the range given in the original dataframe and replace the corresponding value. In the final step of data wrangling I check the following to verify the data integrity:

1. If there are records with more than 100% graduation rate
2. If there are records with less than 0% graduation rate
3. If there are records with a non-integer graduation rate

I also checked to see if there were any record with adjusted cohort graduation rate as 0 and decided to drop these rows from my analysis. To do this, I dropped all the columns in the dataframes and checked to count after dropping the rows with values. After a comparison, I found that 9 & 27 records are dropped from the adjusted cohort graduation rates dataframe of 2011 and 2014 respectively.