1. What kind of cleaning steps did you perform?

The username and password combination dataset required minimal cleaning prior to analysis in this project. In order to protect the username and password combinations from being used for illegal purposes, the data underwent substantial and careful curation prior to their release on Xato.net. These steps include:

1. Limited identifying information by removing the domain portion from email addresses.
2. Combined data samples from thousands of global incidents from the last five years with other data mixed in going back an additional ten years so the accounts cannot be tied to any one company.
3. Removed any keywords, such as company names, that might indicate the source of the login information.
4. Manually reviewed much of the data to remove information that might be particularly linked to an individual.
5. Removed information that appeared to be a credit card or financial account number.
6. Where possible, removed accounts belonging to employees of any government or military sources.

The text file consists of two data columns, usernames and passwords, and the respective columns consist of ten-million rows of string values. Additional information on the data can be found here.

2. How did you deal with missing values, if any?

The ten-million username and password combinations do not have any missing values. The username and password strings may vary in length and character composition. Character compositions of each string includes some varying combination of alphabetical (a-z, A-Z), numeric (0-9), or symbol characters found on a common keyboard.

3. Were there outliers, and how did you handle them?

Due to the nature of the dataset and its release, there are no outliers in the ten-million username and password combination data.