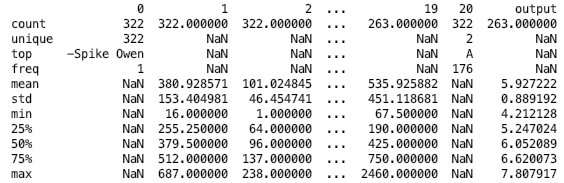
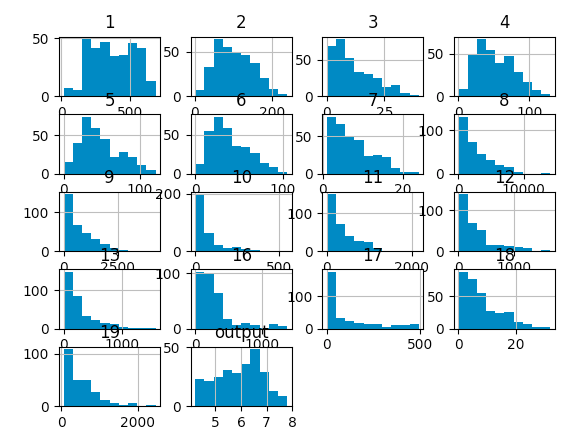
Exploratory Data Analysis using scikit-learn

1. **Use the exploratory data analysis scheme to determine visually which are the candidate features for the model (correlation and scatter plot)**The candidate features for the model can be home runs, at bat, years, put outs, and errors. These are the features that can be used to see if there is any correlation with the salary of a baseball player. We can analyze how the amount of home runs or errors a player makes influences his/her salary.
2. **Calculate the descriptive statistics for the dataset. What is this information good for?**  
   These are the observations of the sample based on pandas .describe() function. The function provides information about the sample such as the distribution for the data. I converted the data to type float, so the .describe() function in Pandas will provide the frequency, mean, standard deviation, minimum and maximum of the dataset.
3. **Plot the histogram of each column. How do you interpret the distribution plots?**  
   

The ‘AtBat’ column has a bimodal distribution. The other columns have a right-skewed distribution. The log of the salary column, which is the output, is normally distributed. In order to fix the right-skewed distribution, we can take the log of the values to make a normal distribution.