Bar graph shows that the average ratings fall between 3.2 and 4.0 meaning we expect that our setiment analysis to show fairly positive reviews.

We need to change the “body” column to a string. The following steps are lowercasing all reviews (this helps in the process of normalization which is an important step to keep the words in a uniform manner), removing punctuation and special characters (not helpful and create their own tokens), stop-words (have no predictive power (I, me, they, etc.), and stemming (removes suffixes like –ing, ses, etc.

We create two new columns to calculate the polarity (senti\_score\_polarity) and subjectivity (senti\_score\_sujectivity) of our sentiment score, and create a new dataframe.

Create a box polt to give a visual representation of the minimum, maximum, and quartiles. We see that the reviews are fairly positive. Polarity shows most reviews are between 0 and .5 (.24 is the mean) while subjectivity falls between .24 and .6 (.49 is the mean).

Create a scatter plot to find the covariance and correlation between the two vairiables.

We then calculated the covariance and correlation

* Covariance between the two variables is 0.12822921. This positive numbers suggests the variable change in the same direction as we expect.
* The correlation is 0.875 which suggests a high level of correlation (any value above 0.5 and close to 1.0)

We created a polarity distribution and density curve, which describes all the values of the variable polarity within a given range. By looking at the distribution we can identify how the polarity is distributed among clients and key factors as mean, median and standard deviation. We want to see that the curve is concentrated in the center and decreases on either side. Ours has results similar to what we wanted.

Frequently we want to know which words are the most common from survey s since we are looking for some patterns.

We then graphed these words by using a pie graph.