KNN

Steps:

1. Data pre-processing
2. Normalize dataset
3. Feature selection
4. Choose hypervariables
5. Estimate Enew and performance

### Feature selection

To begin with, I import the data and look at it by using the following commands:

film\_data.describe()

film\_data.head()

film\_data.info()

film\_data.skew(axis=0)

from this I draw the conclusions that the dataset is skew

display the data

• Do men or women dominate speaking roles in Hollywood movies?

• Has gender balance in speaking roles changed over time (i.e. years)?

• Do films in which men do more speaking make a lot more money than films in which women

speak more?

**How I will do the estimation and calculations:**

I will split away maybe 15% of the date to estimate Enew. This is because I can’t use cross validation for this if I use cross validation for feature selection…

The analysis of the effectiveness to use kNN to predict category “Male” or “Female in the given dataset was done in three steps. The first step included analysis of the structure of the data, in the second step the hypervariable “k” was chosen and estimation of E[error\_new] using cross validation, in the third and final step further evaluation terms were found, also using cross validation. The result of the analysis can be seen in TABLE REF and PICTURE REF.