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…