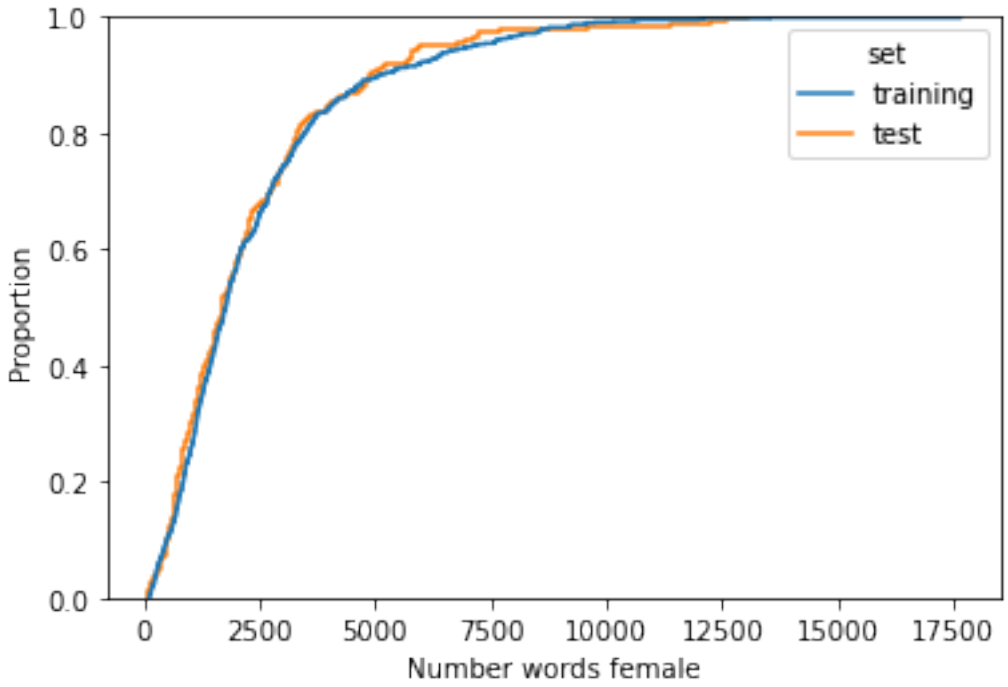
Test set slection

December 22, 2022

[1]:
```
from sklearn.model_selection import train_test_split
import pandas as pd
import numpy as np

data= pd.read_csv("train.csv")

data['Lead'].replace({'Male':1, 'Female':0}, inplace = True)

data = data[data['Number words female'] > 0]

# Separate the target variable from the dataframe as we cannot train the model␣
 ↪with the target variable.
X = data.drop(columns = ["Lead"])
y = data['Lead']

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = 2)
```

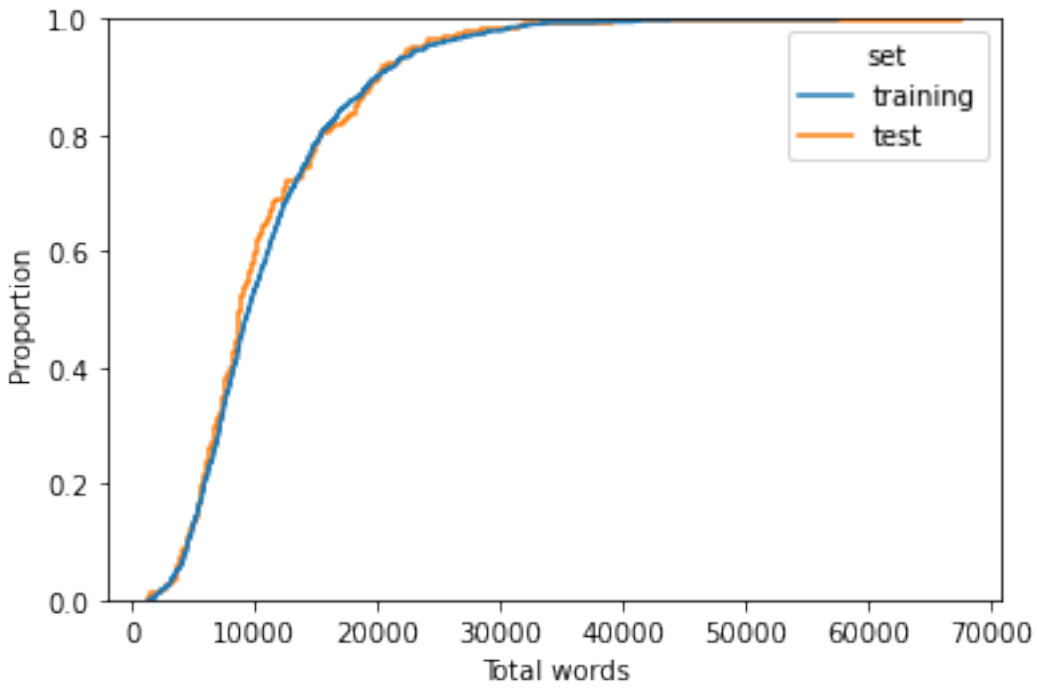
[2]:
```
import seaborn as sns

feature_name = 'Number words female'

df = pd.DataFrame({
    feature_name:np.concatenate((X_train.loc[:,feature_name],X_test.loc[:
 ↪,feature_name])),
    'set':['training']*X_train.shape[0] + ['test']*X_test.shape[0]
    })
sns.ecdfplot(data=df,x=feature_name,hue='set')
```

[2]:
```
<AxesSubplot:xlabel='Number words female', ylabel='Proportion'>
```



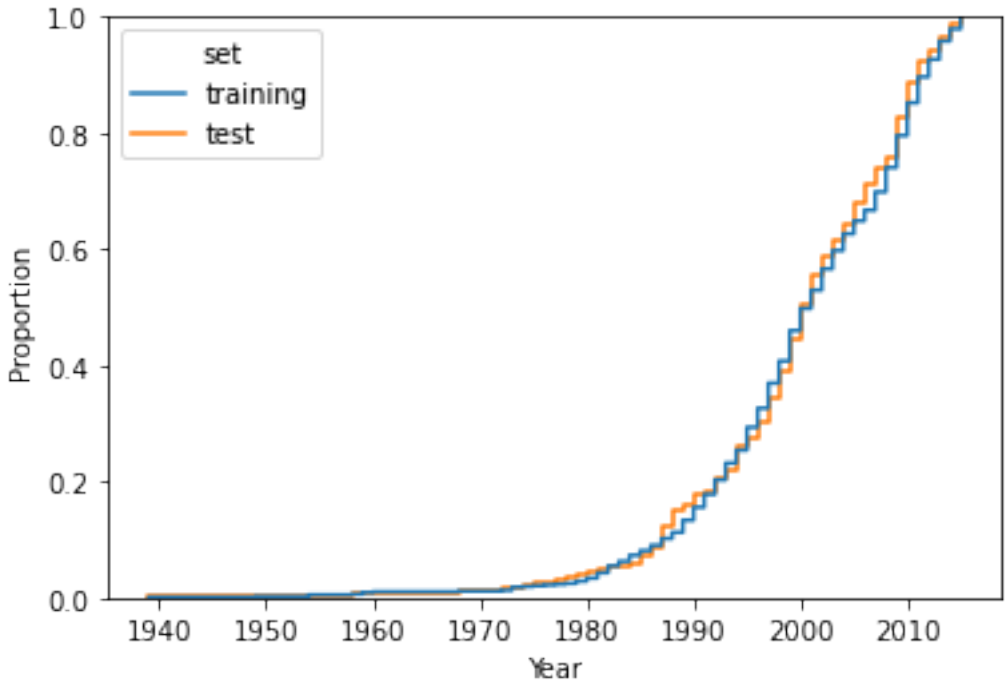
[3]:
```
feature_name = 'Total words'

df = pd.DataFrame({
    feature_name:np.concatenate((X_train.loc[:,feature_name],X_test.loc[:
 ↪,feature_name])),
    'set':['training']*X_train.shape[0] + ['test']*X_test.shape[0]
    })
sns.ecdfplot(data=df,x=feature_name,hue='set')
```

[3]:
```
<AxesSubplot:xlabel='Total words', ylabel='Proportion'>
```



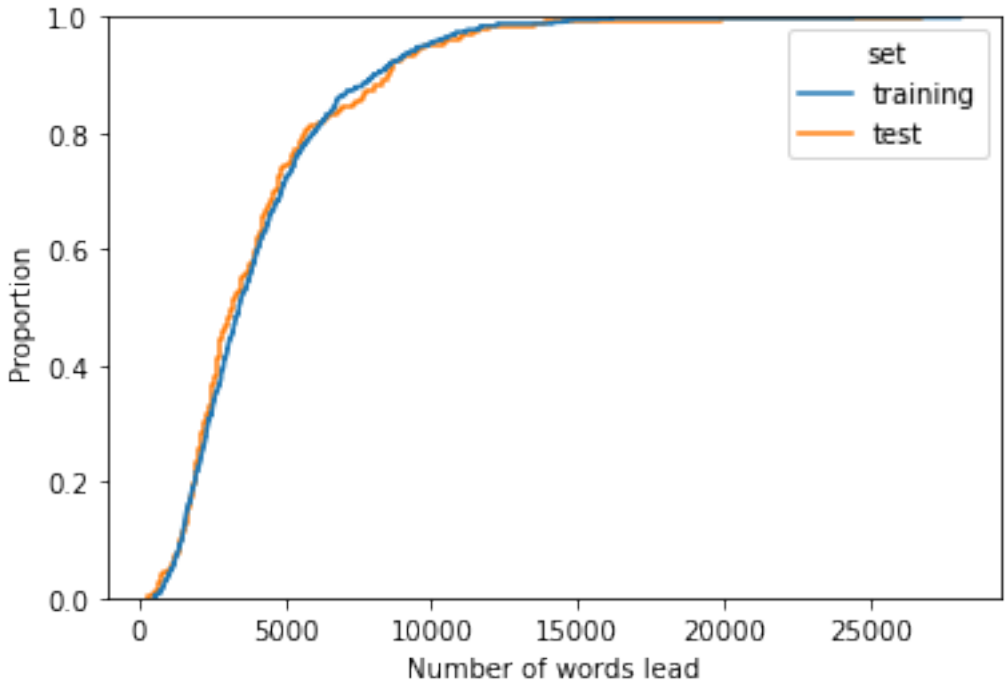
[4]:
```
feature_name = 'Year'

df = pd.DataFrame({
    feature_name:np.concatenate((X_train.loc[:,feature_name],X_test.loc[:
 ↪,feature_name])),
    'set':['training']*X_train.shape[0] + ['test']*X_test.shape[0]
    })
sns.ecdfplot(data=df,x=feature_name,hue='set')
```

[4]:
```
<AxesSubplot:xlabel='Year', ylabel='Proportion'>
```



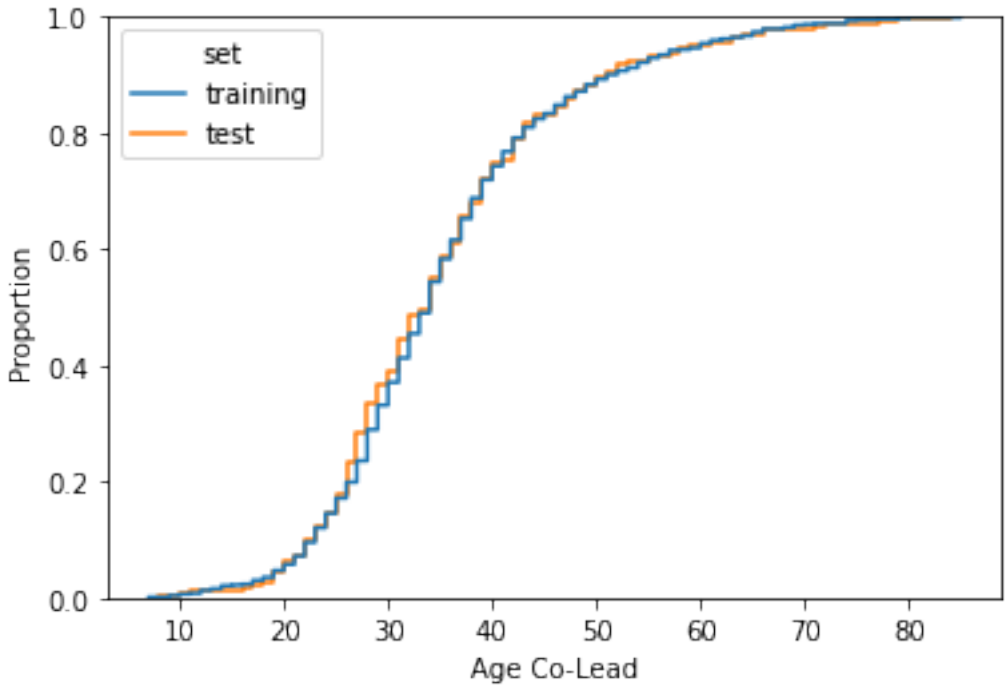
[5]:
```
feature_name = 'Number of words lead'

df = pd.DataFrame({
    feature_name:np.concatenate((X_train.loc[:,feature_name],X_test.loc[:
 ↪,feature_name])),
    'set':['training']*X_train.shape[0] + ['test']*X_test.shape[0]
    })
sns.ecdfplot(data=df,x=feature_name,hue='set')
```

[5]:
```
<AxesSubplot:xlabel='Number of words lead', ylabel='Proportion'>
```



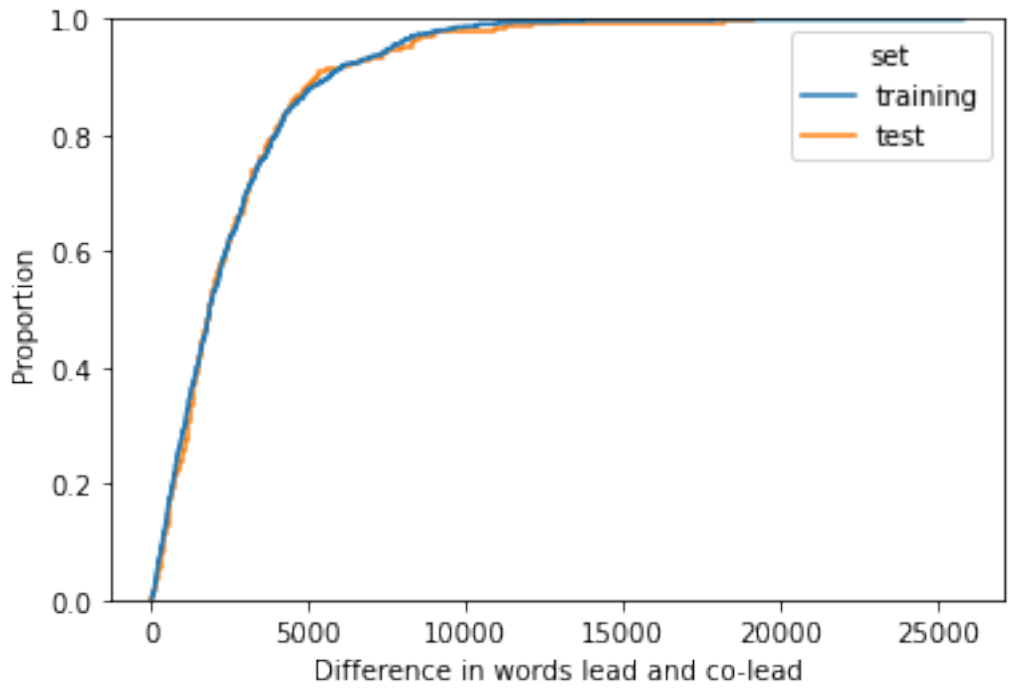
[6]:
```
feature_name = 'Age Co-Lead'
df = pd.DataFrame({
    feature_name:np.concatenate((X_train.loc[:,feature_name],X_test.loc[:
 ↪,feature_name])),
    'set':['training']*X_train.shape[0] + ['test']*X_test.shape[0]
    })
sns.ecdfplot(data=df,x=feature_name,hue='set')
```

[6]:
```
<AxesSubplot:xlabel='Age Co-Lead', ylabel='Proportion'>
```



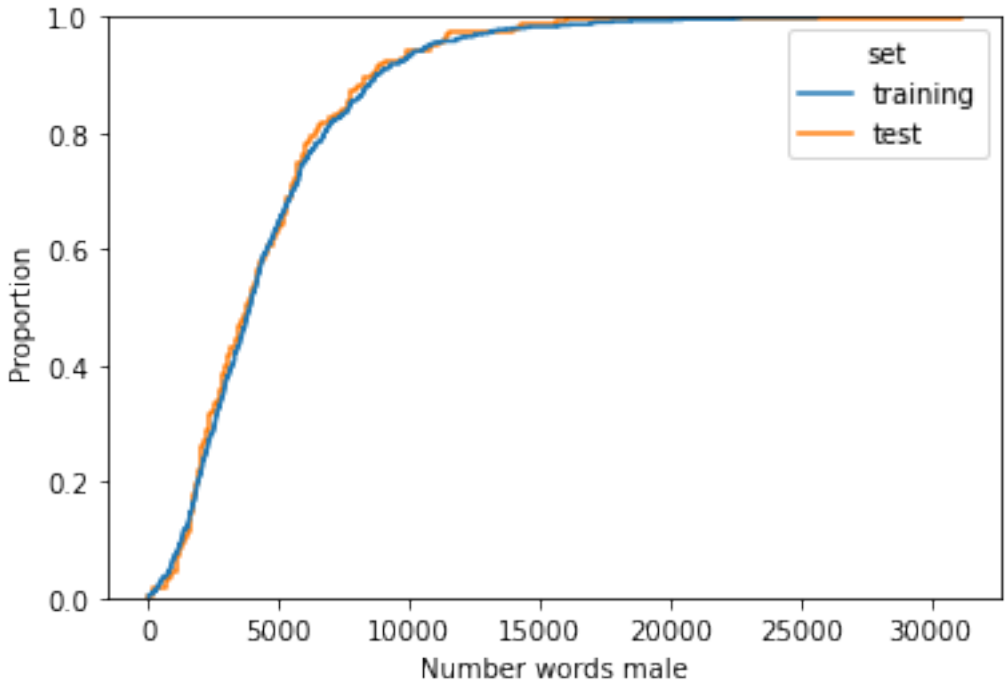
[7]:
```
feature_name = 'Difference in words lead and co-lead'

df = pd.DataFrame({
    feature_name:np.concatenate((X_train.loc[:,feature_name],X_test.loc[:
 ↪,feature_name])),
    'set':['training']*X_train.shape[0] + ['test']*X_test.shape[0]
    })
sns.ecdfplot(data=df,x=feature_name,hue='set')
```

[7]:
```
<AxesSubplot:xlabel='Difference in words lead and co-lead', ylabel='Proportion'>
```



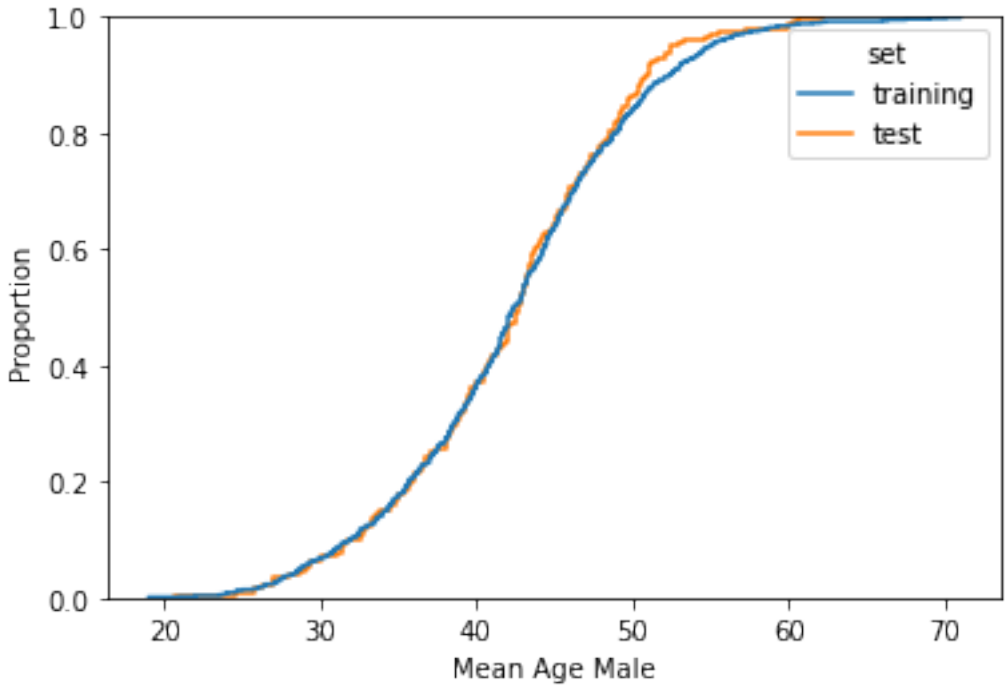
[8]:
```
feature_name = 'Number words male'

df = pd.DataFrame({
    feature_name:np.concatenate((X_train.loc[:,feature_name],X_test.loc[:
 ↪,feature_name])),
    'set':['training']*X_train.shape[0] + ['test']*X_test.shape[0]
    })
sns.ecdfplot(data=df,x=feature_name,hue='set')
```

[8]:
```
<AxesSubplot:xlabel='Number words male', ylabel='Proportion'>
```



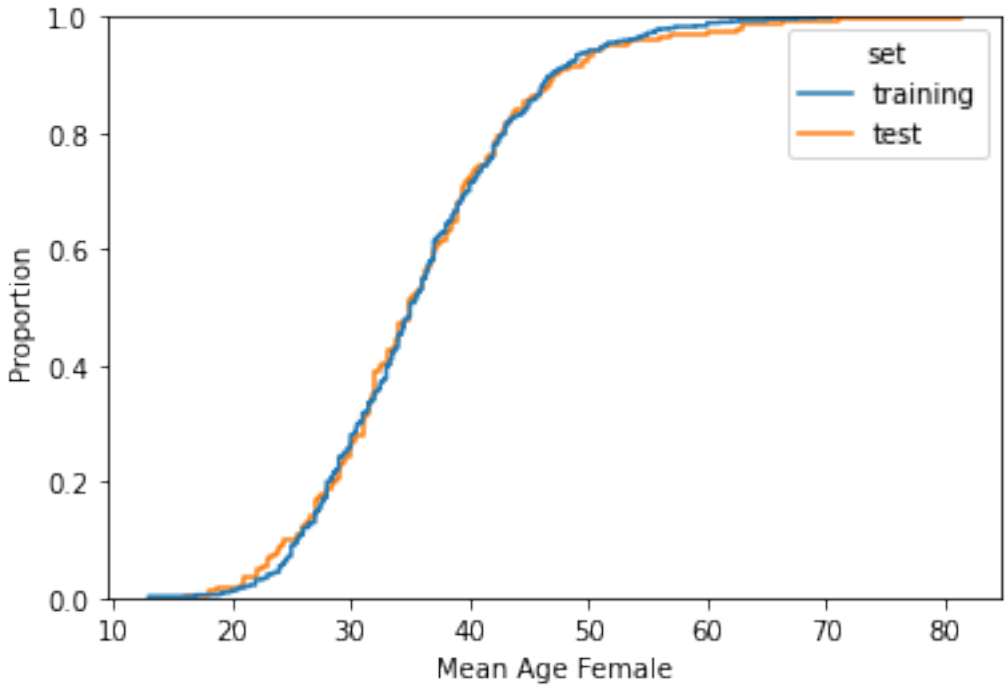
[9]:
```
feature_name = 'Mean Age Male'

df = pd.DataFrame({
    feature_name:np.concatenate((X_train.loc[:,feature_name],X_test.loc[:
 ↪,feature_name])),
    'set':['training']*X_train.shape[0] + ['test']*X_test.shape[0]
    })
sns.ecdfplot(data=df,x=feature_name,hue='set')
```

[9]:
```
<AxesSubplot:xlabel='Mean Age Male', ylabel='Proportion'>
```



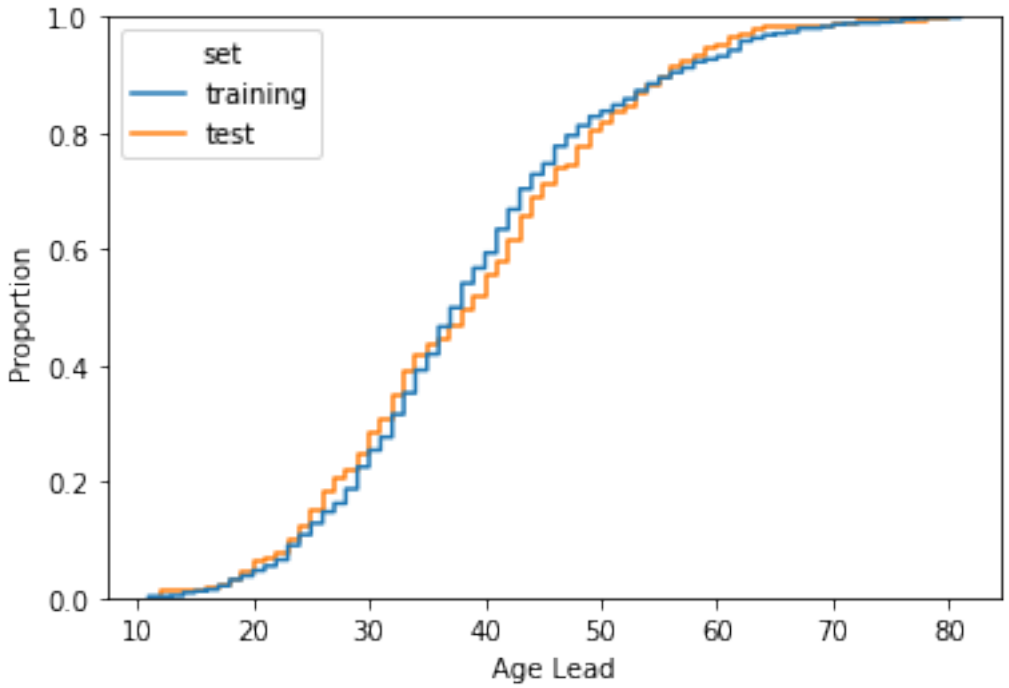
[10]:
```
feature_name = 'Mean Age Female'

df = pd.DataFrame({
    feature_name:np.concatenate((X_train.loc[:,feature_name],X_test.loc[:
 ↪,feature_name])),
    'set':['training']*X_train.shape[0] + ['test']*X_test.shape[0]
    })
sns.ecdfplot(data=df,x=feature_name,hue='set')
```

[10]:
```
<AxesSubplot:xlabel='Mean Age Female', ylabel='Proportion'>
```



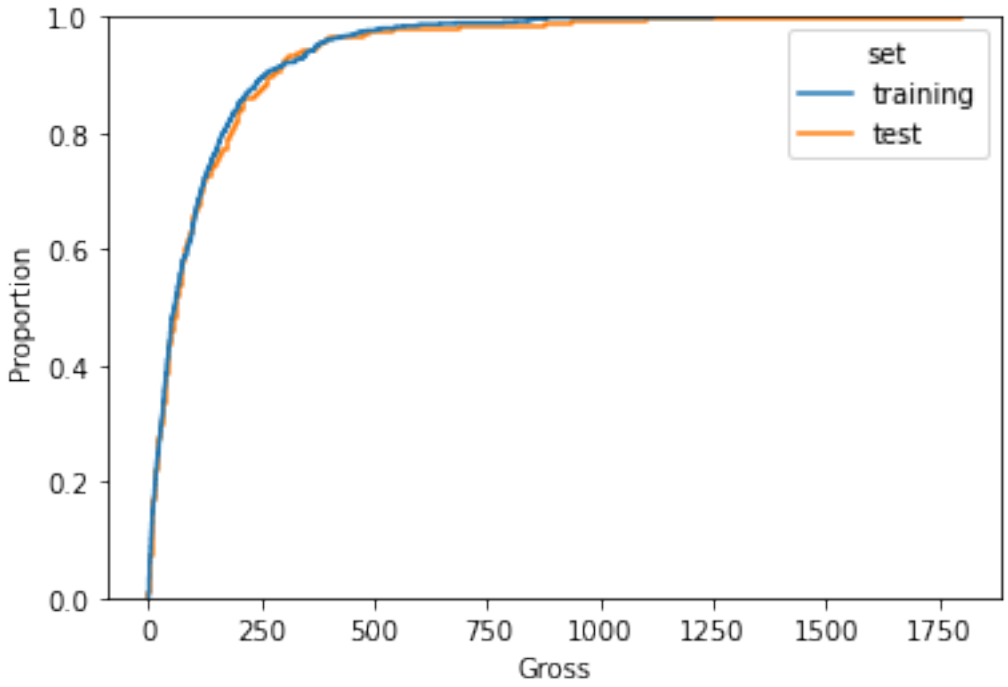
[11]:
```
feature_name = 'Age Lead'

df = pd.DataFrame({
    feature_name:np.concatenate((X_train.loc[:,feature_name],X_test.loc[:
 ↪,feature_name])),
    'set':['training']*X_train.shape[0] + ['test']*X_test.shape[0]
    })
sns.ecdfplot(data=df,x=feature_name,hue='set')
```

[11]:
```
<AxesSubplot:xlabel='Age Lead', ylabel='Proportion'>
```



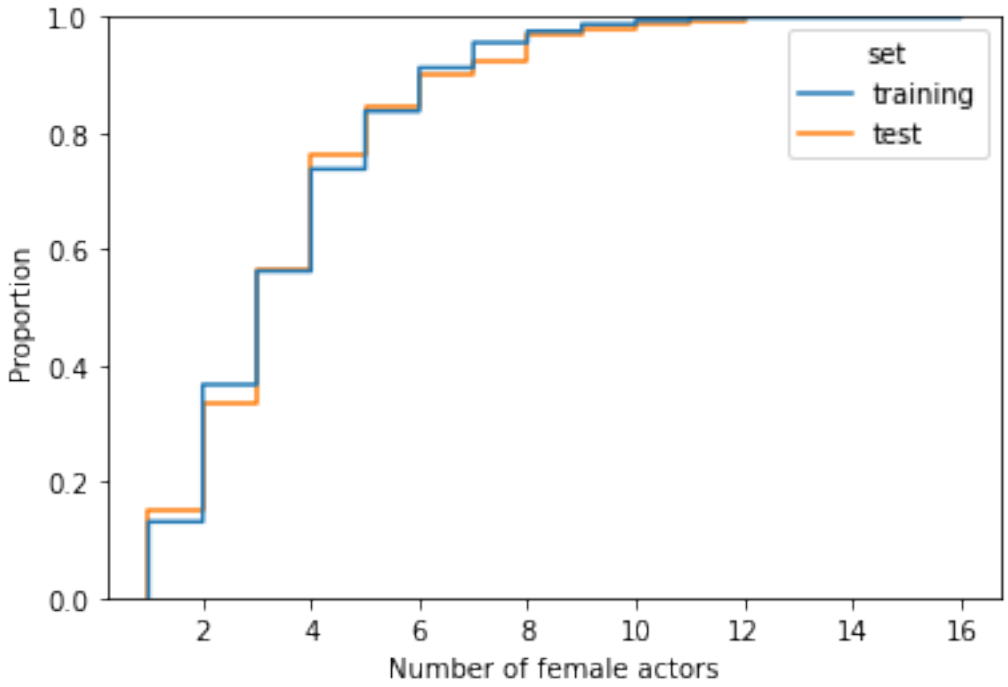
[12]:
```
feature_name = 'Gross'

df = pd.DataFrame({
    feature_name:np.concatenate((X_train.loc[:,feature_name],X_test.loc[:
 ↪,feature_name])),
    'set':['training']*X_train.shape[0] + ['test']*X_test.shape[0]
    })
sns.ecdfplot(data=df,x=feature_name,hue='set')
```

[12]:
```
<AxesSubplot:xlabel='Gross', ylabel='Proportion'>
```



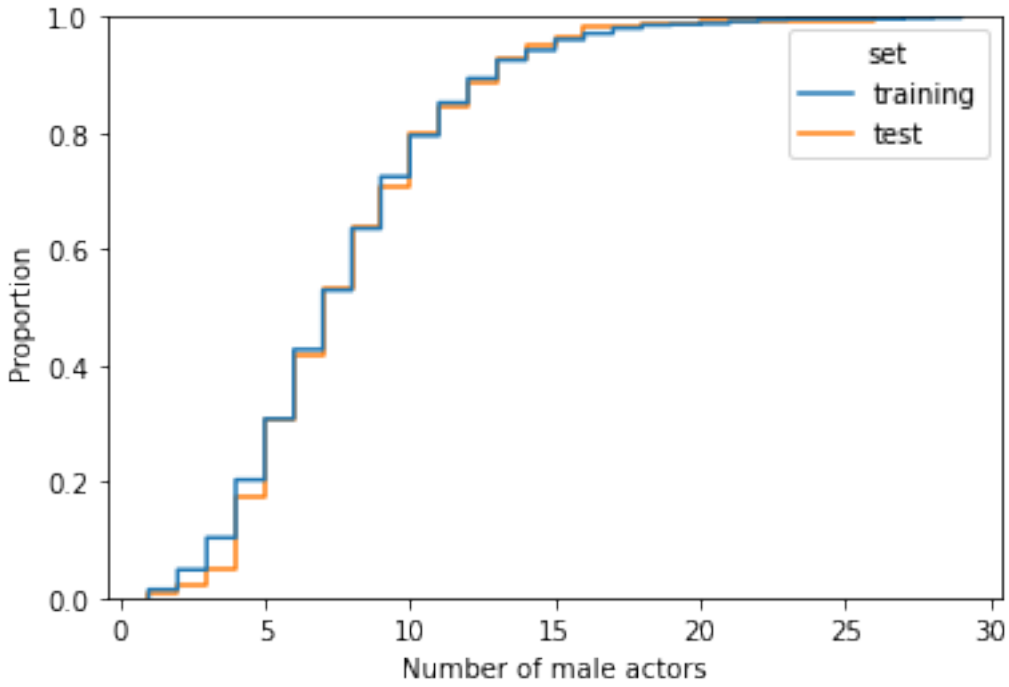
[13]:
```
feature_name = 'Number of female actors'

df = pd.DataFrame({
    feature_name:np.concatenate((X_train.loc[:,feature_name],X_test.loc[:
 ↪,feature_name])),
    'set':['training']*X_train.shape[0] + ['test']*X_test.shape[0]
    })
sns.ecdfplot(data=df,x=feature_name,hue='set')
```

[13]:
```
<AxesSubplot:xlabel='Number of female actors', ylabel='Proportion'>
```



[14]:
```
feature_name = 'Number of male actors'

df = pd.DataFrame({
    feature_name:np.concatenate((X_train.loc[:,feature_name],X_test.loc[:
 ↪,feature_name])),
    'set':['training']*X_train.shape[0] + ['test']*X_test.shape[0]
    })
sns.ecdfplot(data=df,x=feature_name,hue='set')
```

[14]:
```
<AxesSubplot:xlabel='Number of male actors', ylabel='Proportion'>
```



[15]:
```
from scipy.stats import ks_2samp

n_features = X.shape[1]

n_tries = 5000

result = []

for random_state in range(n_tries):
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3,␣
 ↪random_state = random_state)

    distances = list(map(lambda i : ks_2samp(X_train.iloc[:,i],X_test.iloc[:
 ↪,i]).statistic,range(n_features)))


    result.append((random_state, max(distances)))

result.sort(key = lambda x : x[1])
```

Obtained 4045 as the best seed