Regression_Students

April 11, 2023

[ ]:
```
!pip install seaborn
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: seaborn in
/home/huynh/.local/lib/python3.10/site-packages (0.12.2)
Requirement already satisfied: numpy!=1.24.0,>=1.17 in /usr/lib/python3/dist-
packages (from seaborn) (1.21.5)
Requirement already satisfied: matplotlib!=3.6.1,>=3.1 in /usr/lib/python3/dist-
packages (from seaborn) (3.5.1)
Requirement already satisfied: pandas>=0.25 in
/home/huynh/.local/lib/python3.10/site-packages (from seaborn) (2.0.0)
Requirement already satisfied: tzdata>=2022.1 in
/home/huynh/.local/lib/python3.10/site-packages (from pandas>=0.25->seaborn)
(2023.3)
Requirement already satisfied: pytz>=2020.1 in /usr/lib/python3/dist-packages
(from pandas>=0.25->seaborn) (2022.1)
Requirement already satisfied: python-dateutil>=2.8.2 in
/home/huynh/.local/lib/python3.10/site-packages (from pandas>=0.25->seaborn)
(2.8.2)
Requirement already satisfied: six>=1.5 in /usr/lib/python3/dist-packages (from
python-dateutil>=2.8.2->pandas>=0.25->seaborn) (1.16.0)
```

[ ]:
```
import pandas as pd
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.decomposition import PCA
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
import numpy as np
```

159 . (baseline).
, 119 .

1 0.

— , . ,
- , . , , , .

train_test_split() , .
Species. ( )
.

[ ]:
```
def preprocessing(data):
    dummies_data = pd.get_dummies(data['Species'], dtype=int)
    X = data.drop(['Weight'], axis=1)
    X = pd.merge(X, dummies_data, left_index=True, right_index=True)
    Y = data['Weight']
    return X, Y
```

[ ]:
```
data = pd.read_csv('fish_train.csv')
X, Y = preprocessing(data)
```

[ ]:
```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2,␣
↪random_state=21, stratify=data['Species'])
```

[ ]:
```
X_train, y_train = X, Y
```

[ ]:
```
X_train.drop(['Species'], axis=1, inplace=True)
X_test.drop(['Species'], axis=1, inplace=True)
```

[ ]:
```
X_train.head()
```

[ ]:

	Length1	Length2	Length3	Height	Width	Bream	Parkki	Perch	Pike	\
0	20.4	22.0	24.7	5.8045	3.7544	0	0	0	0	\
1	25.4	27.5	28.9	7.2828	4.5662	0	0	1	0	
2	26.5	29.0	34.0	12.4440	5.1340	1	0	0	0	
3	36.2	39.5	45.3	18.7542	6.7497	1	0	0	0	
4	19.0	21.0	22.5	5.6925	3.5550	0	0	1	0	

	Roach	Smelt	Whitefish
0	1	0	0
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0

[ ]:
```
X_test.head()
```

[ ]:

	Length1	Length2	Length3	Height	Width	Bream	Parkki	Perch	Pike	\
31	33.7	36.4	39.6	11.7612	6.5736	0	0	0	0	\
42	20.7	22.7	24.2	5.9532	3.6300	0	0	1	0	

```
54      28.5      30.7      36.2   14.2266   4.9594       1       0       0       0
95      14.3      15.5      17.4    6.5772   2.3142       0       1       0       0
11      35.0      38.5      44.1   18.0369   6.3063       1       0       0       0
```

```
    Roach  Smelt  Whitefish
31      0      0          1
42      0      0          0
54      0      0          0
95      0      0          0
11      0      0          0
```

Width .

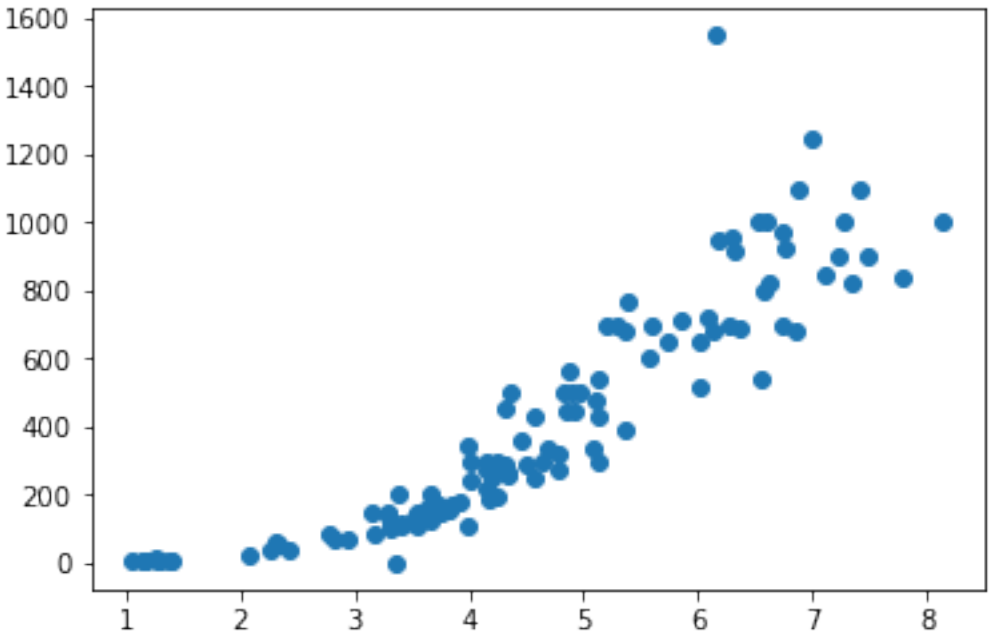
[ ]:
```
# < ENTER YOUR CODE HERE >
X_train['Width'].mean()
```

[ ]:
```
4.507406722689075
```

[ ]:
```
import matplotlib.pyplot as plt

plt.scatter(X['Width'], Y)
```

[ ]:
```
<matplotlib.collections.PathCollection at 0x7fe71ab65720>
```



2 1.

(LinearRegression()) .
. r2_score().

[ ]:
```
def train_test_get_r2score(model , X_train, X_test, y_train, y_test):
    # model = LinearRegression()
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    return r2_score(y_test, y_pred)
```

[ ]:
```
model = LinearRegression()
train_test_get_r2score(model, X_train, X_test, y_train, y_test)
```

[ ]:
```
0.8948321458297644
```

3 2. .

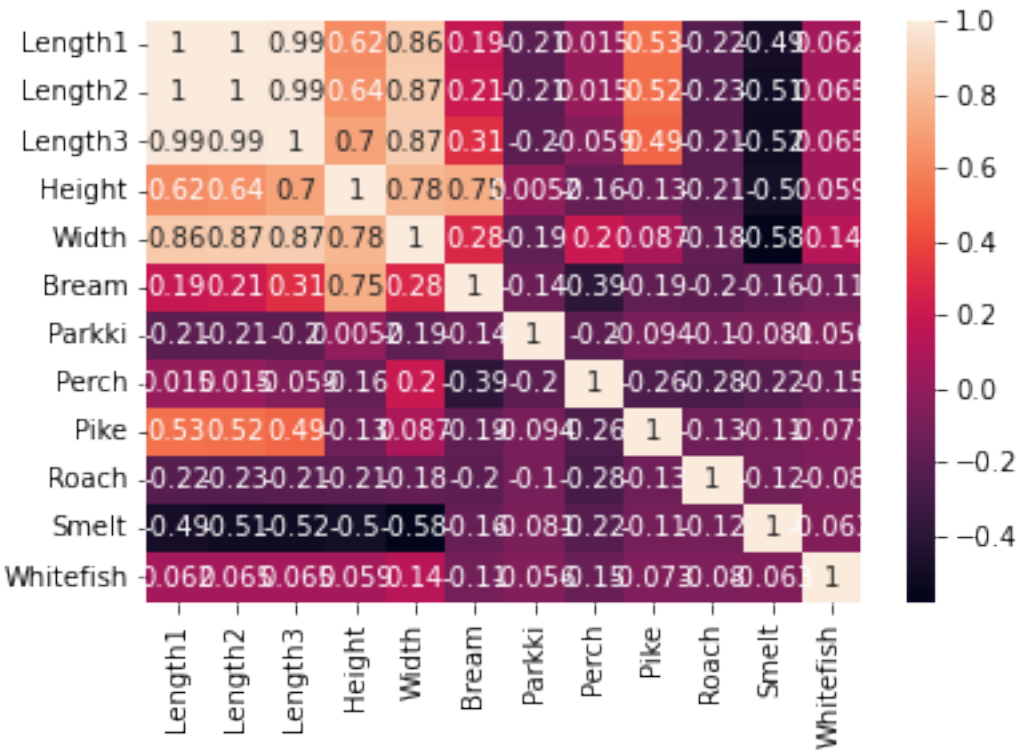
3.1 PCA

, , sns.heatmap(),
.

[ ]:
```
corr_matrix = X_train.corr()

sns.heatmap(corr_matrix, annot=True)
```

[ ]:
```
<AxesSubplot:>
```



,

.

.

1) (PCA(n_components=3, svd_solver='full'))

.

2) .

3) Lengths,

.

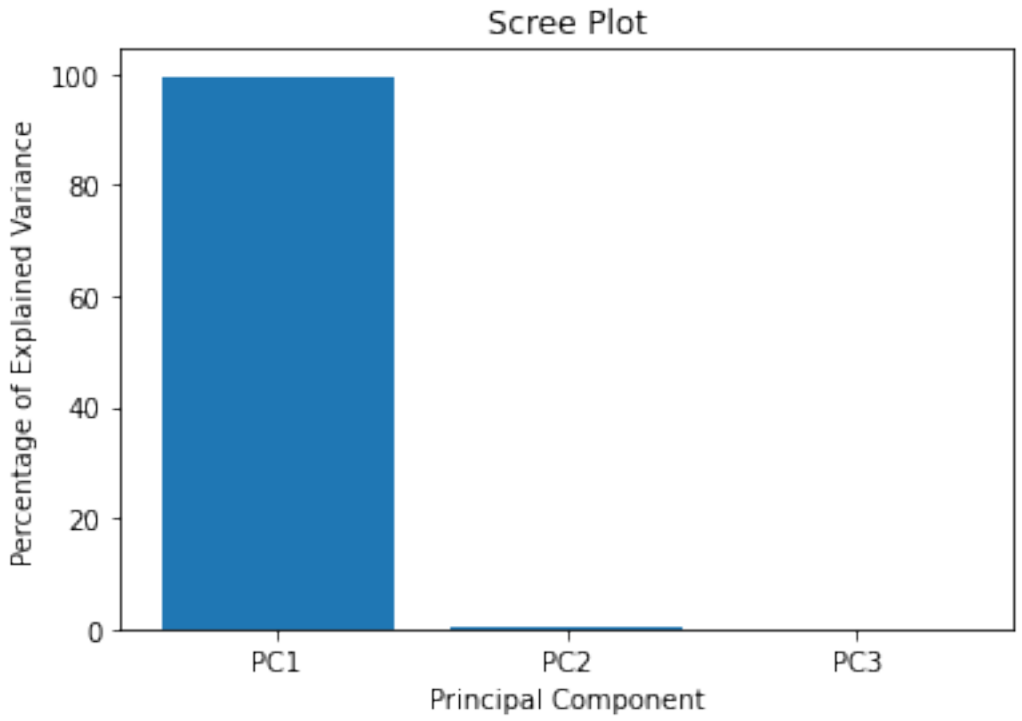
[ ]:
```
pca = PCA(n_components=3, svd_solver='full')
# new_X_train = X_train.drop(['Height', 'Width'], axis=1)
new_X_train = X_train.loc[:, ['Length1', 'Length2', 'Length3']]
pca.fit(new_X_train)
```

[ ]:
```
PCA(n_components=3, svd_solver='full')
```

[ ]:
```
per_var = np.round(pca.explained_variance_ratio_*100, decimals=1)
labels = ['PC'+str(x) for x in range(1, len(per_var)+1)]

plt.bar(x=range(1, len(per_var)+1), height=per_var, tick_label=labels)
plt.ylabel('Percentage of Explained Variance')
plt.xlabel('Principal Component')
```

```
plt.title('Scree Plot')
plt.show()
```



[ ]:
```
pca.explained_variance_ratio_[0]
```

[ ]: 0.9962761645834379

[ ]:
```
def transform_data_and_replace_column(X, pca):
    #  new_X = X.drop(['Height', 'Width', 'Bream',
       #  'Parkki', 'Perch', 'Pike', 'Roach', 'Smelt', 'Whitefish'], axis=1)
    new_X = X.loc[:, ['Length1', 'Length2', 'Length3']]
    X_pca = pca.transform(new_X)
    X_pca_df = pd.DataFrame(X_pca)
    # print(X_pca_df.head())
    X_Lengths = X_pca_df.iloc[:, 0]
    X.drop(['Length1', 'Length2', 'Length3'], axis=1, inplace=True)
    X = X.set_index(X_Lengths.index)
    X.insert(2, 'Lengths', X_Lengths)
    return X
```

. , PCA .

.

[ ]:
```
X_train.head()
```

[ ]:

	Length1	Length2	Length3	Height	Width	Bream	Parkki	Perch	Pike	\
0	20.4	22.0	24.7	5.8045	3.7544	0	0	0	0	\
1	25.4	27.5	28.9	7.2828	4.5662	0	0	1	0	
2	26.5	29.0	34.0	12.4440	5.1340	1	0	0	0	
3	36.2	39.5	45.3	18.7542	6.7497	1	0	0	0	
4	19.0	21.0	22.5	5.6925	3.5550	0	0	1	0	

	Roach	Smelt	Whitefish
0	1	0	0
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0

[ ]:
```
X_train = transform_data_and_replace_column(X_train, pca)
X_test = transform_data_and_replace_column(X_test, pca)
```

[ ]:
```
X_train.head()
```

[ ]:

	Height	Width	Lengths	Bream	Parkki	Perch	Pike	Roach	Smelt	\
0	5.8045	3.7544	-11.486348	0	0	0	0	1	0	\
1	7.2828	4.5662	-3.048526	0	0	1	0	0	0	
2	12.4440	5.1340	1.557884	1	0	0	0	0	0	
3	18.7542	6.7497	19.779440	1	0	0	0	0	0	
4	5.6925	3.5550	-14.171988	0	0	1	0	0	0	

	Whitefish
0	0
1	0
2	0
3	0
4	0

[ ]:
```
X_test.head()
```

[ ]:

	Height	Width	Lengths	Bream	Parkki	Perch	Pike	Roach	Smelt	\
0	11.7612	6.5736	13.132277	0	0	0	0	0	0	\
1	5.9532	3.6300	-11.232588	0	0	1	0	0	0	
2	14.2266	4.9594	4.967269	1	0	0	0	0	0	
3	6.5772	2.3142	-23.005900	0	1	0	0	0	0	
4	18.0369	6.3063	17.819541	1	0	0	0	0	0	

	Whitefish
0	1
1	0
2	0
3	0

4 0

, .

, r2_score().

```
[ ]: train_test_get_r2score(model, X_train, X_test, y_train, y_test)
```

```
[ ]: 0.8947758151607493
```

, .

3.2

, , sns.pairplot().

```
[ ]: def merge_data(X_train, y_train):
         temp_df = pd.DataFrame(y_train)
         new_df = pd.merge(X_train, temp_df, left_index=True, right_index=True)
         return new_df
```

```
[ ]: new_df = merge_data(X_train, y_train)
```

```
[ ]: # sns.pairplot(new_df)
```

```
[ ]: # sns.pairplot(data)
```

, , , . , ,
, . ?

, - , .
, — ,

$$m = \rho \cdot V.$$

, , , ,

$$m \sim V \sim d^3$$

, (Height, Width, Lengths),
, , .

```
[ ]: def cube_features(X):
         X['Height'] = X['Height']**3
         X['Width'] = X['Width']**3
         X['Lengths'] = X['Lengths']**3
         return X
```

```
[ ]: X_train.head()
```

[ ]:

	Height	Width	Lengths	Bream	Parkki	Perch	Pike	Roach	Smelt	\
0	5.8045	3.7544	-11.486348	0	0	0	0	1	0	\
1	7.2828	4.5662	-3.048526	0	0	1	0	0	0	
2	12.4440	5.1340	1.557884	1	0	0	0	0	0	
3	18.7542	6.7497	19.779440	1	0	0	0	0	0	
4	5.6925	3.5550	-14.171988	0	0	1	0	0	0	

	Whitefish
0	0
1	0
2	0
3	0
4	0

[ ]:
```
X_train = cube_features(X_train)
X_test = cube_features(X_test)
```

[ ]:
```
X_train.head()
```

[ ]:

	Height	Width	Lengths	Bream	Parkki	Perch	Pike	Roach	\
0	195.566492	52.920218	-1515.464970	0	0	0	0	1	\
1	386.273710	95.206103	-28.331510	0	0	1	0	0	
2	1926.992424	135.321746	3.780985	1	0	0	0	0	
3	6596.227555	307.505871	7738.235923	1	0	0	0	0	
4	184.462936	44.928179	-2846.376115	0	0	1	0	0	

	Smelt	Whitefish
0	0	0
1	0	0
2	0	0
3	0	0
4	0	0

[ ]:
```
X_test.head()
```

[ ]:

	Height	Width	Lengths	Bream	Parkki	Perch	Pike	Roach	\
0	1626.877698	284.059829	2264.749149	0	0	0	0	0	\
1	210.984922	47.832147	-1417.227131	0	0	1	0	0	
2	2879.409033	121.979658	122.561198	1	0	0	0	0	
3	284.526777	12.393748	-12176.365210	0	1	0	0	0	
4	5867.940377	250.797891	5658.346418	1	0	0	0	0	

	Smelt	Whitefish
0	0	1
1	0	0
2	0	0
3	0	0

4 0 0

Width .

```
[ ]: X_train['Width'].mean()
```

```
[ ]: 128.92916356818105
```

, Weight Width .

```
[ ]: # sns.pairplot(merge_data(X_train, y_train))
```

, .

, r2_score().

```
[ ]: train_test_get_r2score(model, X_train, X_test, y_train, y_test)
```

```
[ ]: 0.9591276154981445
```

, !

3.3

one-hot Species, , pd.get_dummies().

. ,

r2_score().

: , Species

. . , , , .

```
[ ]: # < ENTER YOUR CODE HERE >
```

.

, one-hot . , ,

drop_first=True. . ,

r2_score().

```
[ ]: # < ENTER YOUR CODE HERE >
```

, , .

```
[ ]: reserved_data = pd.read_csv('fish_reserved.csv')
dummies_data = pd.get_dummies(data['Species'], dtype=int)
X_test = reserved_data.drop(['Species'], axis=1)
X_test = pd.merge(X_test, dummies_data, left_index=True, right_index=True)
X_test = transform_data_and_replace_column(X_test, pca)
X_test = cube_features(X_test)
X_test.head()
```

[ ]:

	Height	Width	Lengths	Bream	Parkki	Perch	Pike	Roach	\
0	228.614702	35.751102	-2308.030768	0	0	0	0	1	
1	260.182831	44.136677	-2545.130350	0	0	1	0	0	
2	179.005641	28.378207	-5553.298626	1	0	0	0	0	
3	591.054429	24.848519	-4798.944450	1	0	0	0	0	
4	406.078856	19.098395	-7061.710296	0	0	1	0	0	

	Smelt	Whitefish
0	0	0
1	0	0
2	0	0
3	0	0
4	0	0

[ ]:
```
y_pred = model.predict(X_test)
```

[ ]:
```
y_pred
```

[ ]:
```
array([ 115.09542494,  167.16936251,   99.12886442,  126.24659395,
        112.1744372 ,  198.10125498,   67.21469221,  308.06001752,
        467.44675927,  142.92956477, 2002.3220395 ,  168.35915909,
        623.32251851, 1098.71038117,    5.13384353,  564.36182331,
        222.73154997,  833.7057696 , 1358.23047718,  100.3565067 ,
        564.71154429,  230.76690949,  174.8736096 ,  158.93253763,
        230.48964244,  572.19576307,  285.79290417,  -33.10226129,
         26.1652189 ,  -16.79036008,  938.68254721,  597.00063586,
        516.57723253,  302.5850343 ,  -20.92020551,  219.54135221,
        225.09507122,  942.16843229,  751.45974073,  173.97427872])
```