МИНОБРНАУКИ РОССИИ

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ

ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)

Кафедра МО ЭВМ

ОТЧЕТ

по практической работе №1

по дисциплине «Машинное обучение»

Студент гр. 6307	____________________	Михайлов И. Т.
Преподаватель	____________________	Жангиров Т. Р.

Санкт-Петербург

2020

Ход выполнения работы

Задание 1

Исходные данные

Предположим X и Y две случайные переменные отражающие возраст и вес, соответственно. Рассмотрим случайную выборку из 20 наблюдений
X = (69, 74, 68, 70, 72, 67, 66, 70, 76, 68, 72, 79, 74, 67, 66, 71, 74, 75, 75, 76)
Y = (153, 175, 155, 135, 172, 150, 115, 137, 200, 130, 140, 265, 185, 112, 140, 150, 165, 185, 210, 220).

A. Найдем среднее, медиану и моду величины X.

Среднее:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n).$$

```
x = pd.Series(x_vals)
x.mean()
```

```
71.45
```

Медиана:

Так как число членов четное:

$$Me = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2}$$

```
x.median()
```

```
71.5
```

Мода:

```
x.mode()[0]
```

```
74
```

B. Найдем дисперсию Y.

Пусть $X_1, \ldots, X_n$ - выборка.

Дисперсия выборки или выборочная дисперсия оценивается по формуле:

$$D = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1},$$

где $\overline{X}$ - среднее значение выборки.

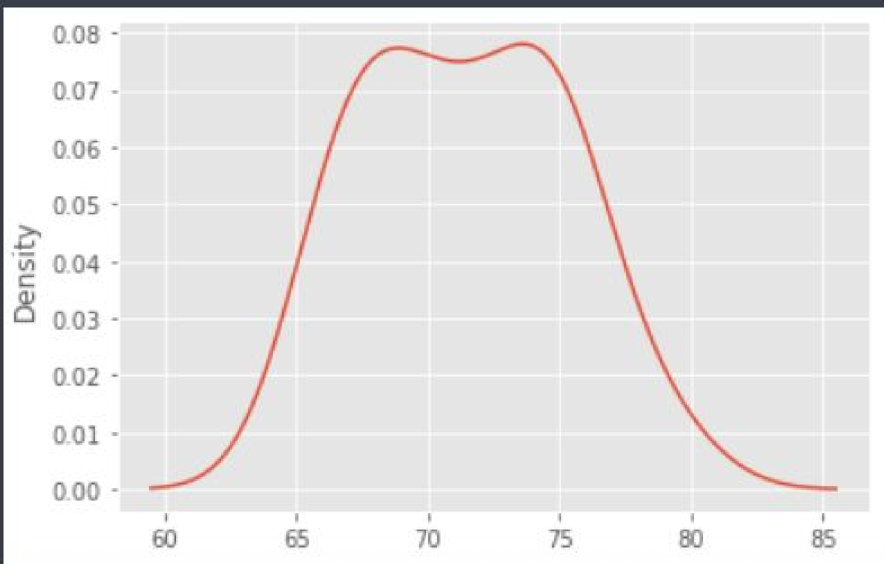
```
y = pd.Series(y_vals)
y.var(ddof=0)
```

```
1369.2099999999998
```

C. Построим график нормального распределения для X.

```
x.plot.kde()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4834ce32b0>
```



D. Найдем вероятность того, что возраст больше 80.

$$P(\alpha < X < \beta) = F_0\left(\frac{\beta - a}{\sigma}\right) - F_0\left(\frac{\alpha - a}{\sigma}\right)$$

Где $\alpha = 80$, $\beta = \infty$, a - мат. ожидание, σ - среднеквадратичное отклонение, F_0 - функция Лапласа.

```
#Подсчет вероятности
s = np.std(x)
a = x.mean()
v = (80 - a) / s
print("a =", a, "s =", s, "v =", v)
```

```
a = 71.45 s = 3.7212229172679248 v = 2.2976317705463614
```

$P(X > 80) = F_0(\infty) - F_0(2.3) = 0.5 - 0.4893 = 0.0107$

E. Найдем двумерное мат. ожидание и ковариационную матрицу для этих двух величин.

Двумерное мат. ожидание:

Пусть (ε, η) - двумерная случайная величина, тогда $M(\varepsilon, \eta) = (M\varepsilon, M\eta)$, т.е. математическое ожидание случайного вектора - это вектор из математических ожиданий компонент вектора.

```
[np.mean(x), np.mean(y)]
```

```
[71.45, 164.7]
```

Ковариационная матрица:

Ковариация двух выборок (двух случайных величин) - это мера их линейной зависимости, которая определяется следующим образом:

$$cov(X, Y) = M[(X - MX)(Y - MY)],$$

где M - математическое ожидание.

```
np.cov(x, y)
```

```
array([[  14.57631579,  128.87894737],
       [ 128.87894737, 1441.27368421]])
```

F. Определим корреляцию между X и Y

Функция corrcoef() вычисляет коэффициент корреляции Пирсона (линейный коэффициент корреляции).

Данный коэффициент вычисляется по формуле:

$$R_{XY} = \frac{C_{XY}}{\sigma_X \sigma_Y} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2 \sum(Y-\bar{Y})^2}}$$

где C_{XY} - ковариационная матрица,

$\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$ и $\overline{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t$ это средние значения выборок.

Коэффициент корреляции находится в интервале *[-1, 1]*.

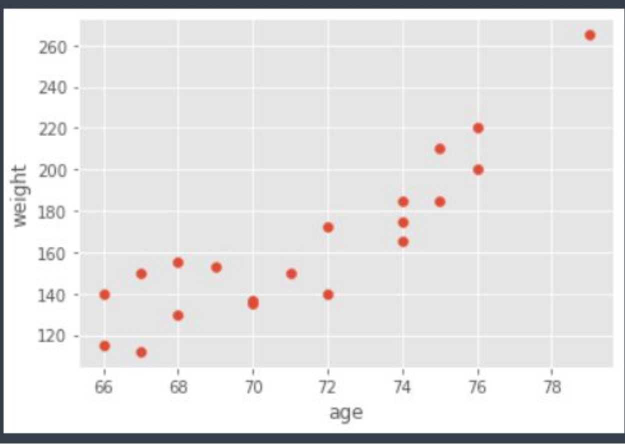
```
np.corrcoef(x, y)[0, 1]

0.8891701351748048
```

G. Построим диаграмму рассеяния, отображающая зависимость между возрастом и весом

```
pyplot.xlabel('age')
pyplot.ylabel('weight')
pyplot.scatter(x, y)

<matplotlib.collections.PathCollection at 0x7f4832eb51c0>
```



Задание 2

Для следующего набора данных

	X_1	X_2	X_3
a	17	17	12
b	11	9	13
c	11	8	19

Рассчитайте ковариационную матрицу и обобщенную дисперсию.

Ковариационная матрица:

```
x1 = [17, 11, 11]
x2 = [17, 9, 8]
x3 = [12, 13, 19]
vec = [x1, x2, x3]
np.cov(vec)
```

```
array([[ 12.        ,  17.        ,  -8.        ],
       [ 17.        ,  24.33333333, -12.83333333],
       [ -8.        , -12.83333333,  14.33333333]])
```

Обобщенная дисперсия:

```
mtrx = np.cov(vec)
round(np.linalg.det(mtrx), 2)
```

```
0.0
```

Задание 3

Даны два одномерных нормальных распределения N_a и N_b с мат. ожиданиями 4, 8 и СКО 1, 2 соответственно.

A. Для каждого из значения {5,6,7} определите какое из распределений сгенерировало значение с большей вероятностью.

B. Найди значение, которой могло быть сгенерировано обеими распределениями с равной вероятностью

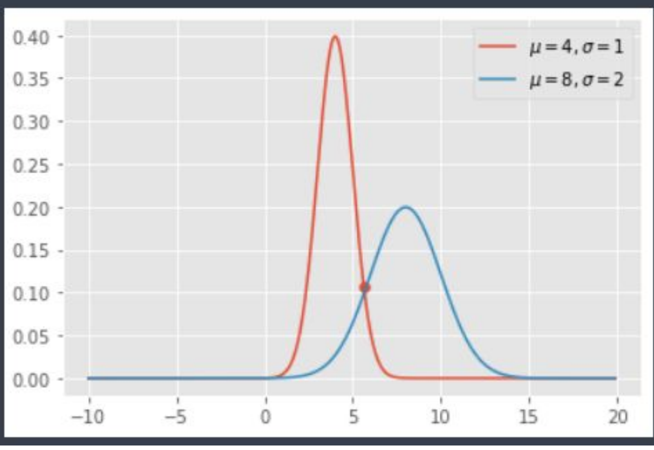
B. Значение, которое могло быть сгенерировано обеими распределениями с равной вероятностью:

```
#3
def norm(x, mu, sigma):
    return math.exp(-0.5*((x - mu) / sigma) ** 2) / (sigma * math.sqrt(2 * math.pi))

axis = np.arange(-10, 20, 0.1)
pyplot.plot(axis, [norm(x, 4, 1) for x in axis], label = '$\mu = 4, \sigma = 1$')
pyplot.plot(axis, [norm(x, 8, 2) for x in axis], label = '$\mu = 8, \sigma = 2$')
pyplot.scatter(5.62, norm(5.62, 4, 1))
pyplot.legend()

for i in range (400, 2000, 1):
    if abs(norm(i/100, 4, 1) - norm(i/100, 8, 2)) < 0.01:
        break;
print(i/100)
```

```
5.62
```



A. Для каждого из значения {5,6,7} определите какое из распределений сгенерировало значение с большей вероятностью.

Заметим, что N_a сгенерирует с большей вероятностью значения меньшие 5.62, а N_b значения большие 5.62. Таким образом 5 с большей вероятностью сгенерирует N_a, а 6 и 7 с большей вероятностью сгенерирует N_b.