Демонстрация возможностей нотбука с Кластеризацией

Первоначальный отбор признаков для предобработки

Необходимо ли отобрать признаки из всех? (y/n)

y

Признаки, участвующие в кластеризации:

- [x] Client ID
- [x] Genre
- [x] Age
- [x] Annual Income (k$)
- [x] Spending Score (1-100)
- [ ] Unknown feature
- [ ] Unknown feature 2
- [x] category

Выбор индекса

Установите название столбца с индексами (id)

Column: Client ID

Обработка дат при наличии

Обработка дат

```
cb = date_preproc_widget(df)
```

```
В признаках присутствуют даты?(y/n)
y
Укажите столбцы с датой:
```

- [ ] Genre
- [x] date_test

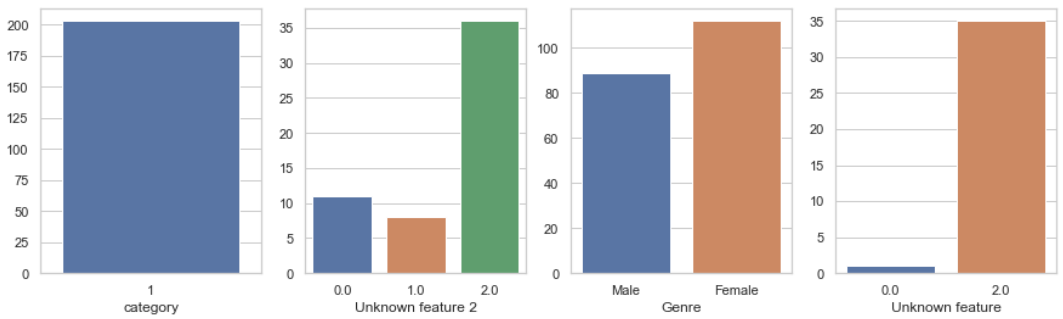
```
df = date_preproc(df, cb, dayfirst =True)
```

```
Столбец  date_test

Что делать с датой?
1.Преобразовать
2.Удалить
1

1.Количество дней до текущей даты
2.Количество дней до другой даты
1
```

Вывод распределений признаков

```
cat_plots(df)
```

Распределение категориальных признаков



Распределение численных признаков

```
df.describe()
```

	Age	Annual Income (k$)	Spending Score (1-100)	Unknown feature	Unknown feature 2	category
count	202.000000	202.000000	202.000000	36.000000	55.000000	203.0
mean	39.475248	61.445545	50.495050	1.944444	1.454545	1.0
std	18.111921	28.047392	28.500021	0.333333	0.812362	0.0
min	4.000000	15.000000	-40.000000	0.000000	0.000000	1.0
25%	28.250000	42.000000	34.250000	2.000000	1.000000	1.0
50%	36.000000	62.000000	50.000000	2.000000	2.000000	1.0
75%	49.000000	78.000000	73.000000	2.000000	2.000000	1.0
max	200.000000	200.000000	200.000000	2.000000	2.000000	1.0

Удаление объектов с большим числом пропусков

Пропуски

```
df = drop_nan_rows(df)
```

```
Количество пустых значений в объектах:
```

	Число NA	Число объектов
0	6	1
1	2	147
2	1	20
3	0	35

```
Удалить объекты с большим числом NA?(y/n) y
```

Выберите отсечение, по которому удалить объекты

```
Удалить объекты, у которых число признаков с NA>=6

Количество оставшихся объектов: 202
```

Удаление признаков с низкой вариативностью

Удаление низковариативных признаков

```
df = drop_unvariative_cols(df)
```

Признаки с низким коэффициентом вариативности $\frac{std}{mean}$:

category : 0.000

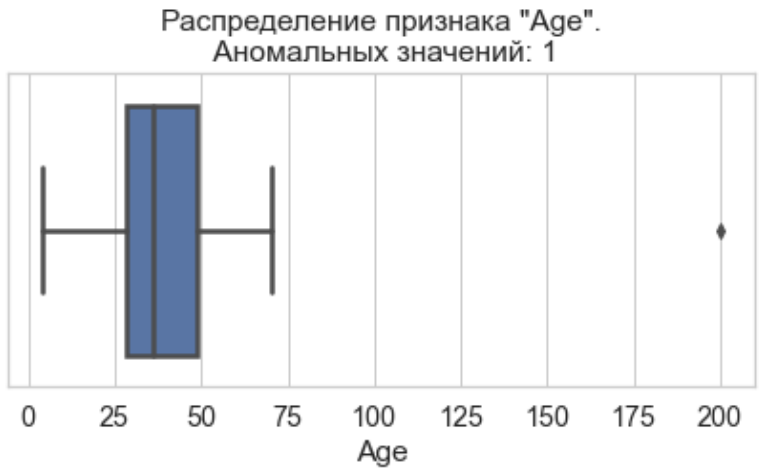
Удалить признаки с низкой вариативностью? (y/n):

```
y
Удалить признаки, с коэфф. вариативности не больше: 0
```

Обработка аномалий (для каждого признака)

Выберите метод обработки аномалий для признаков

Age




```
Метод обработки аномальных значений:
1.Удалить
2.Ограничить
```

Удаление признаков с большим числом пропусков

Количество пропусков в признаках:

	%NA	Тип
Unknown feature	82.09%	float64
Unknown feature 2	72.64%	float64
Genre	0.5%	object

Удалить признаки с большим процентом NA? (y/n) y

Удалить признаки, с % NA>= 50

Выбор заполнения пропусков для остальных признаков

Выберите метод заполнения пропусков для каждого признака

Признак **Genre**

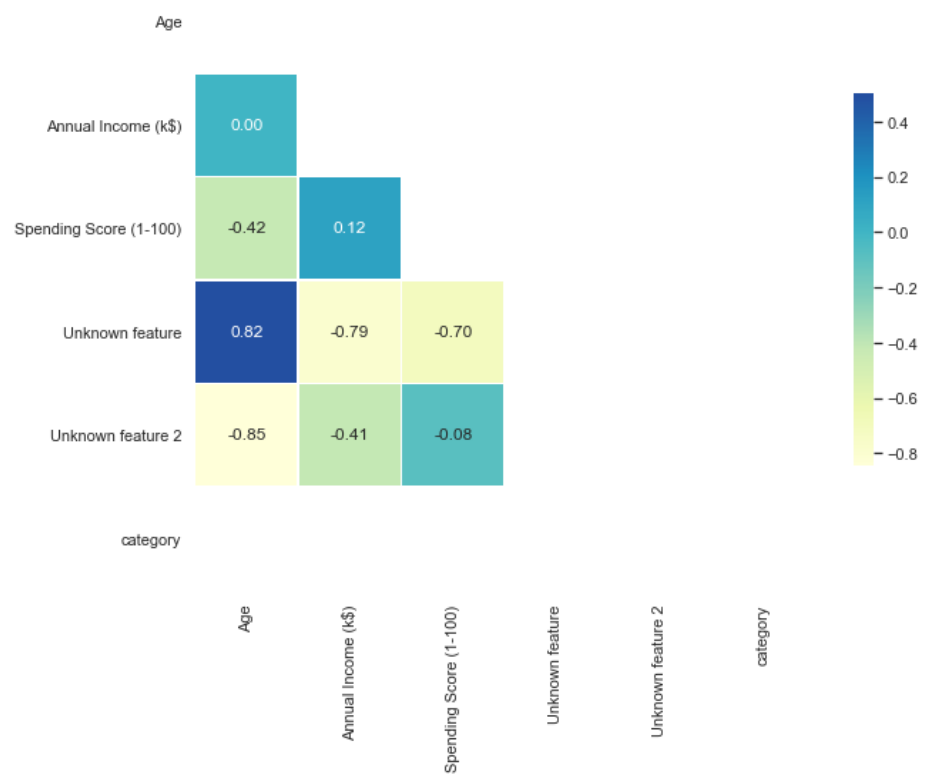
Распределение:

```
Female    0.56
Male      0.44
Name: Genre, dtype: float64
```

Заполнить NA:
1.Female (Мода)
2.Male
3.Случайно (с весами из распределения)
4.Ручной ввод

Построение матрицы корреляции

Матрица корреляций признаков



Признаки с высокой корреляцией:

Unknown feature и **Age** = 0.81917

Unknown feature 2 и **Age** = -0.84671

Кодирование и масштабирование признаков

Кодирование признаков

```
df = encoding(df)
```

Genre, уникальных значений: 2

```
Способ кодирования:
1.LabelEncoder
2.OneHot
3.Удалить этот признак
Выбор: 1
```

Масштабирование

```
dfc = df.copy()
df = scaling(df)
```

```
Способ масштабирования данных:
1.Стандартизация
2.Нормализация[0,1]
3.Не масштабировать
Выбор: 2
```

Выбор метода для рассмотрения

Кластеризация

Выбор метода

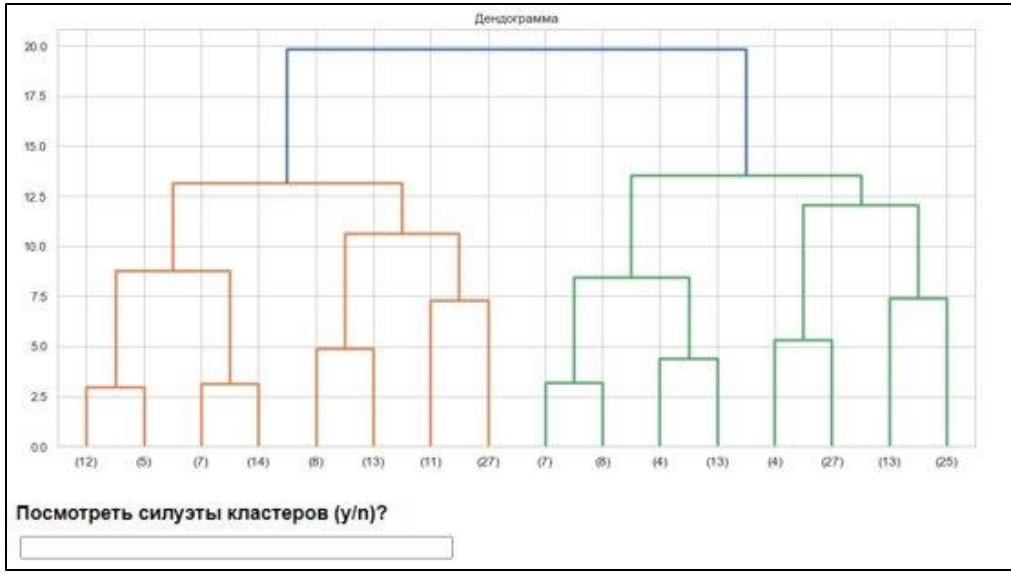
```
method = choice_method_widget()
```

Выберите метод кластеризации

Метод: K-means / Agglomerative / DBSCAN

Подбор параметров для методов кластеризации

agglomerative

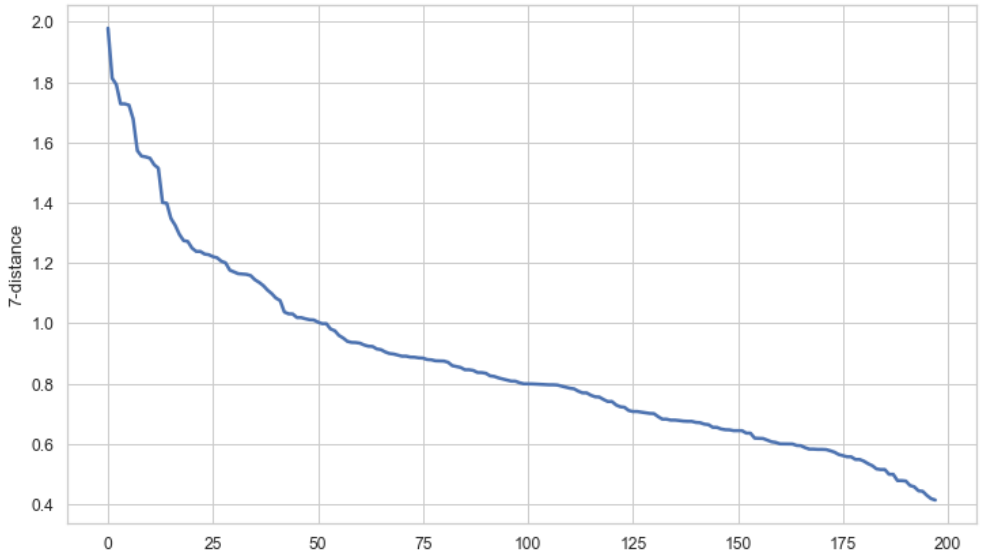


Посмотреть силуэты кластеров (y/n)?

dbscan

Построить график расстояний до k-ого соседа для оценки параметра *eps*?(y/n)

y



Параметр *eps* нужно определить как дистанцию до k-ого ($k = dim * 2 - 1$) соседа, при которой возникает первая "впадина" на графике

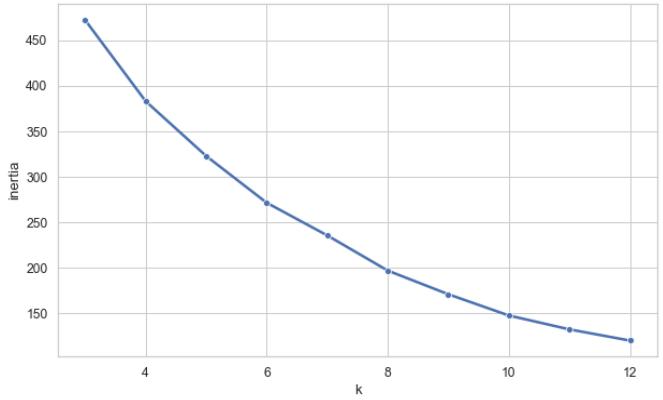
Параметр minPt предлагается устанавливать как $2 \cdot dim$

Ссылка на статью

kmeans

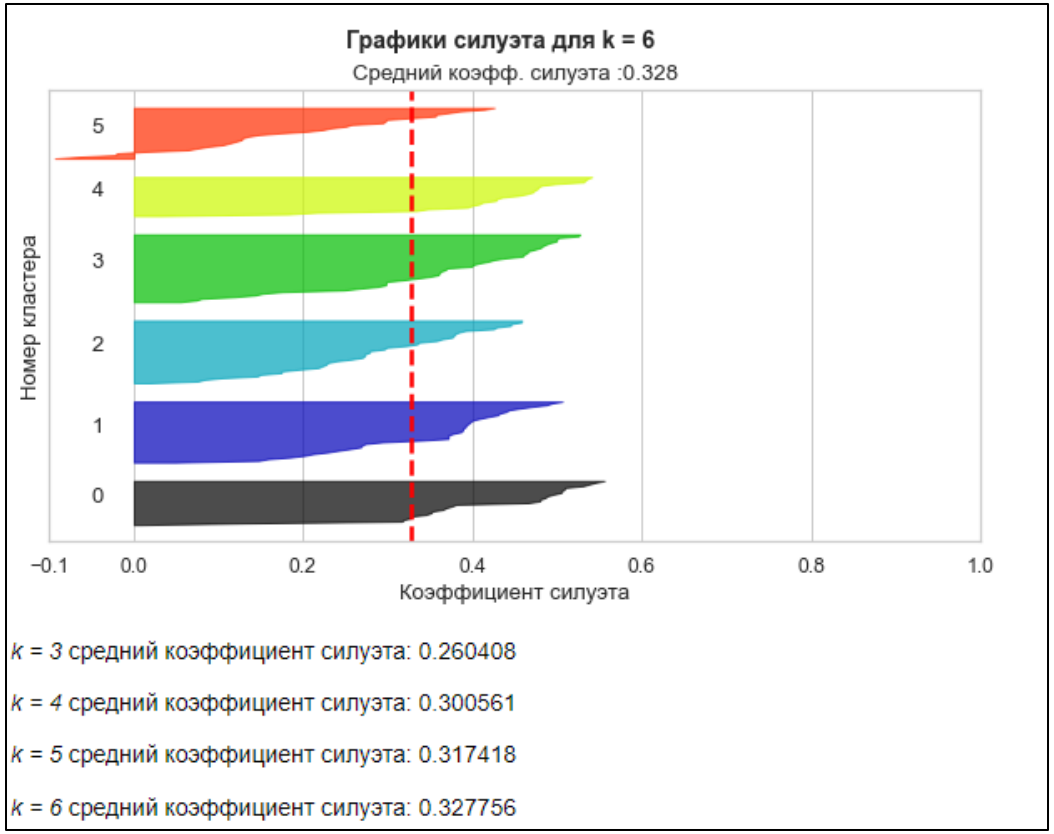
12/12 кластер

График изменения суммы квадратов расстояний в кластерах



Посмотреть силуэты кластеров (y/n)?

Построение силуэтов для различного количества кластеров



k = *3* средний коэффициент силуэта: 0.260408

k = *4* средний коэффициент силуэта: 0.300561

k = *5* средний коэффициент силуэта: 0.317418

k = *6* средний коэффициент силуэта: 0.327756

Построение характеристик кластеров

Средние значения характеристик в кластерах

cluster	Genre	Age	Annual Income (k$)	Spending Score (1-100)
0	0.391753	28.577320	60.185567	68.000000
1	0.449275	52.173913	46.333333	40.971014
2	0.531250	41.750000	87.593750	16.125000