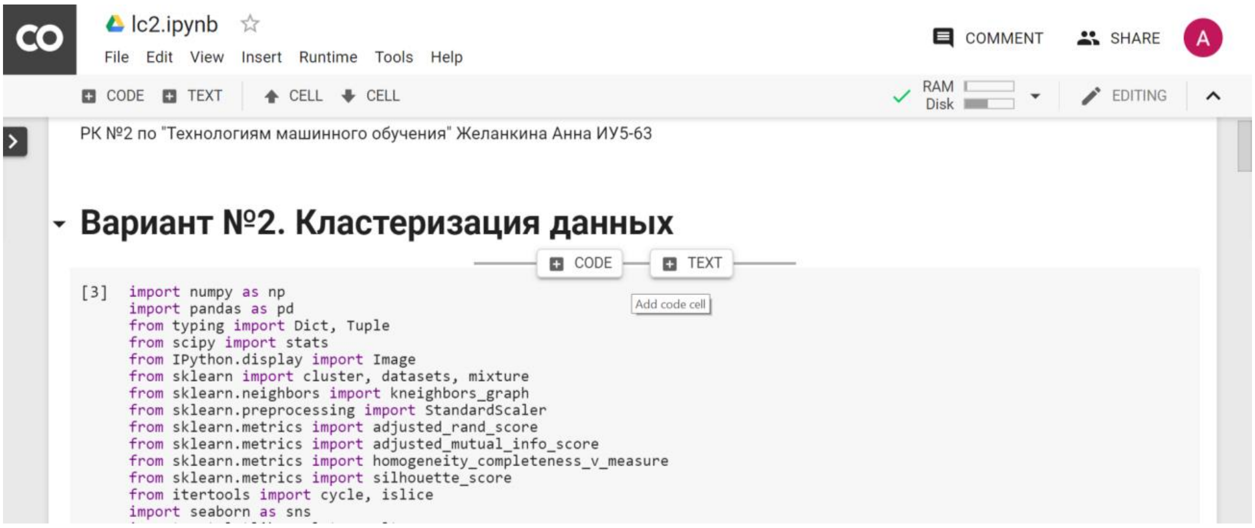
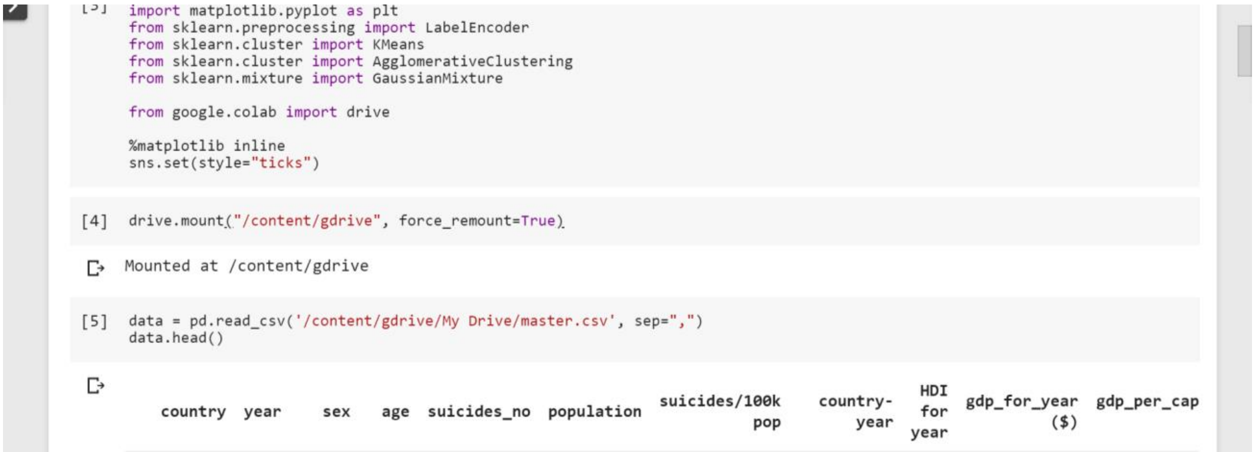


РК №2 по "Технологиям машинного обучения" Желанкина Анна ИУ5-63

Вариант №2. Кластеризация данных

```
[3]  import numpy as np
     import pandas as pd
     from typing import Dict, Tuple
     from scipy import stats
     from IPython.display import Image
     from sklearn import cluster, datasets, mixture
     from sklearn.neighbors import kneighbors_graph
     from sklearn.preprocessing import StandardScaler
     from sklearn.metrics import adjusted_rand_score
     from sklearn.metrics import adjusted_mutual_info_score
     from sklearn.metrics import homogeneity_completeness_v_measure
     from sklearn.metrics import silhouette_score
     from itertools import cycle, islice
     import seaborn as sns
```

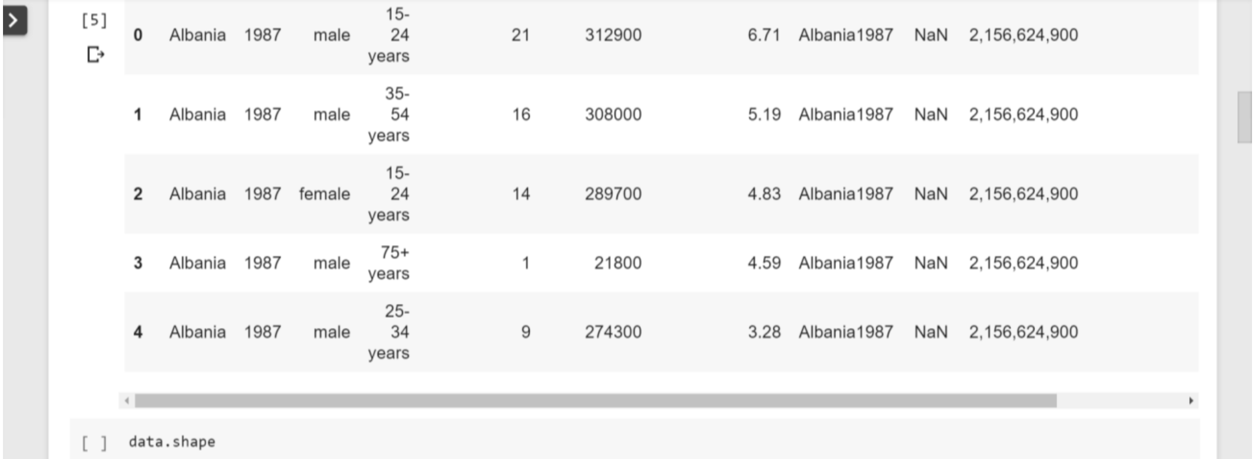


```
[3]  import matplotlib.pyplot as plt
     from sklearn.preprocessing import LabelEncoder
     from sklearn.cluster import KMeans
     from sklearn.cluster import AgglomerativeClustering
     from sklearn.mixture import GaussianMixture

     from google.colab import drive

     %matplotlib inline
     sns.set(style="ticks")
```

```
[4]  drive.mount("/content/gdrive", force_remount=True)
```

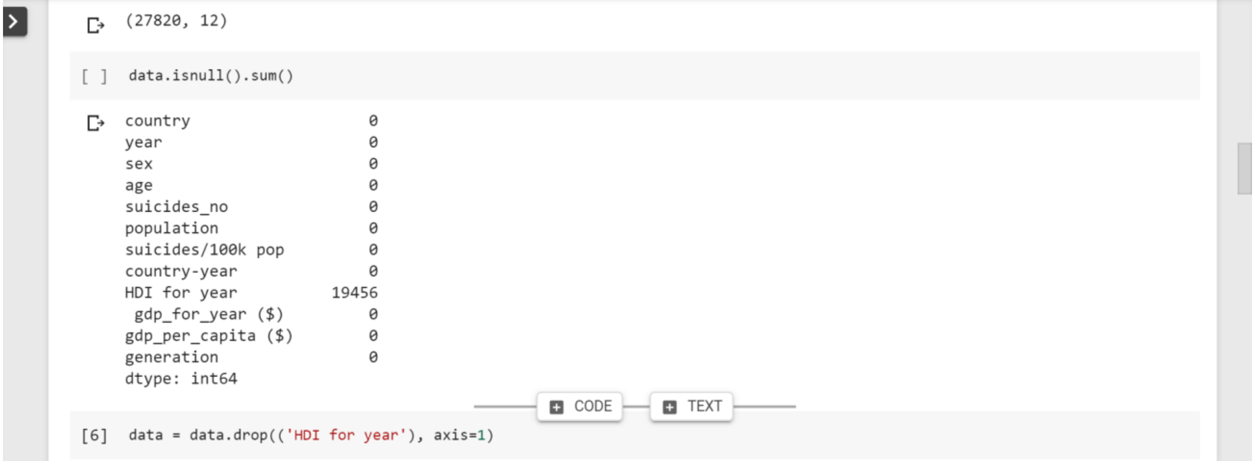
```
Mounted at /content/gdrive
```

```
[5]  data = pd.read_csv('/content/gdrive/My Drive/master.csv', sep=",")
     data.head()
```



	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_cap
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NaN	2,156,624,900	
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NaN	2,156,624,900	
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NaN	2,156,624,900	
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2,156,624,900	
4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NaN	2,156,624,900	

```
[ ]  data.shape
```

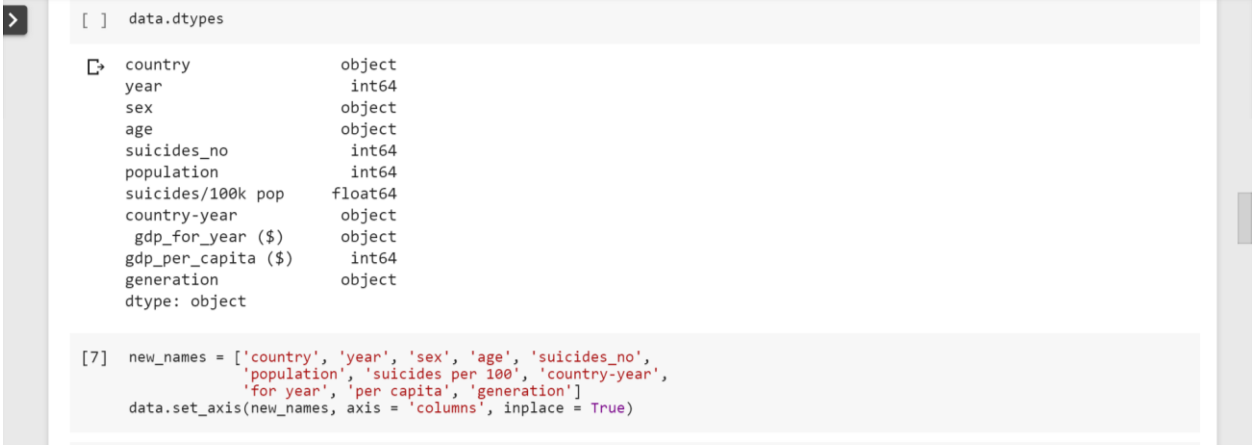


```
(27820, 12)
```

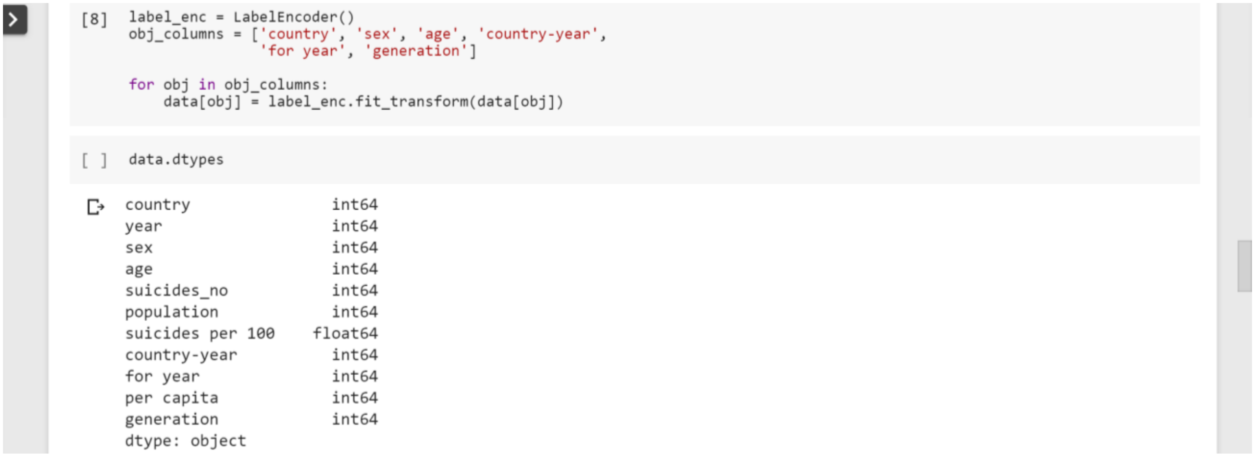
```
[ ]  data.isnull().sum()
```

```
country                   0
year                      0
sex                       0
age                       0
suicides_no               0
population                0
suicides/100k pop         0
country-year              0
HDI for year          19456
 gdp_for_year ($)         0
gdp_per_capita ($)        0
generation                0
dtype: int64
```

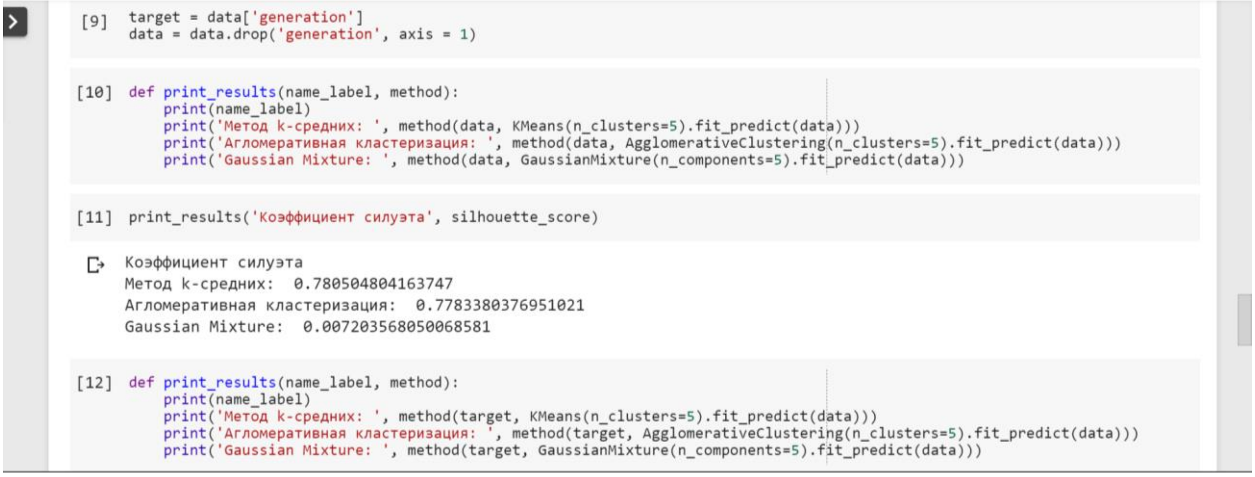
```
[6]  data = data.drop(('HDI for year'), axis=1)
```



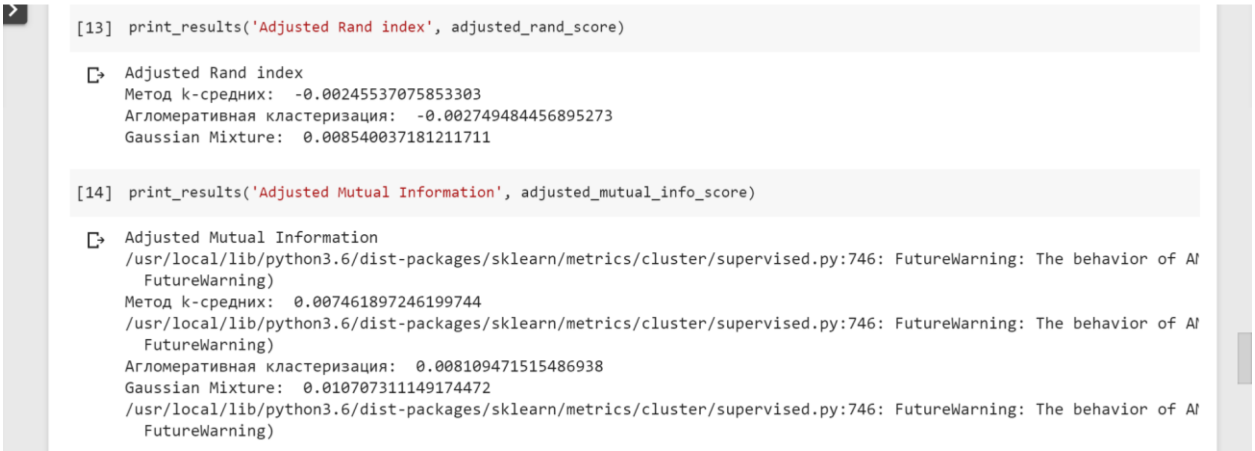
```
[ ]  data.dtypes

     country                 object
     year                     int64
     sex                     object
     age                     object
     suicides_no              int64
     population               int64
     suicides/100k pop      float64
     country-year            object
      gdp_for_year ($)       object
     gdp_per_capita ($)       int64
     generation              object
     dtype: object

[7]  new_names = ['country', 'year', 'sex', 'age', 'suicides_no',
                  'population', 'suicides per 100', 'country-year',
                  'for year', 'per capita', 'generation']
     data.set_axis(new_names, axis = 'columns', inplace = True)
```



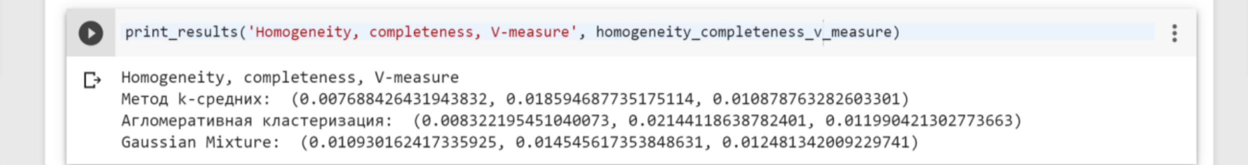
```
[8]  label_enc = LabelEncoder()
     obj_columns = ['country', 'sex', 'age', 'country-year',
                    'for year', 'generation']

     for obj in obj_columns:
         data[obj] = label_enc.fit_transform(data[obj])

[ ]  data.dtypes

     country              int64
     year                 int64
     sex                  int64
     age                  int64
     suicides_no          int64
     population           int64
     suicides per 100   float64
     country-year         int64
     for year             int64
     per capita           int64
     generation           int64
     dtype: object
```



```
[9]  target = data['generation']
     data = data.drop('generation', axis = 1)

[10] def print_results(name_label, method):
         print(name_label)
         print('Метод k-средних: ', method(data, KMeans(n_clusters=5).fit_predict(data)))
         print('Агломеративная кластеризация: ', method(data, AgglomerativeClustering(n_clusters=5).fit_predict(data)))
         print('Gaussian Mixture: ', method(data, GaussianMixture(n_components=5).fit_predict(data)))

[11] print_results('Коэффициент силуэта', silhouette_score)

     Коэффициент силуэта
     Метод k-средних:  0.780504804163747
     Агломеративная кластеризация:  0.7783380376951021
     Gaussian Mixture:  0.007203568050068581

[12] def print_results(name_label, method):
         print(name_label)
         print('Метод k-средних: ', method(target, KMeans(n_clusters=5).fit_predict(data)))
         print('Агломеративная кластеризация: ', method(target, AgglomerativeClustering(n_clusters=5).fit_predict(data)))
         print('Gaussian Mixture: ', method(target, GaussianMixture(n_components=5).fit_predict(data)))
```



```
[13] print_results('Adjusted Rand index', adjusted_rand_score)

     Adjusted Rand index
     Метод k-средних:  -0.00245537075853303
     Агломеративная кластеризация:  -0.002749484456895273
     Gaussian Mixture:  0.008540037181211711

[14] print_results('Adjusted Mutual Information', adjusted_mutual_info_score)

     Adjusted Mutual Information
     /usr/local/lib/python3.6/dist-packages/sklearn/metrics/cluster/supervised.py:746: FutureWarning: The behavior of AI
       FutureWarning)
     Метод k-средних:  0.007461897246199744
     /usr/local/lib/python3.6/dist-packages/sklearn/metrics/cluster/supervised.py:746: FutureWarning: The behavior of AI
       FutureWarning)
     Агломеративная кластеризация:  0.008109471515486938
     Gaussian Mixture:  0.010707311149174472
     /usr/local/lib/python3.6/dist-packages/sklearn/metrics/cluster/supervised.py:746: FutureWarning: The behavior of AI
       FutureWarning)
```



```
     print_results('Homogeneity, completeness, V-measure', homogeneity_completeness_v_measure)

     Homogeneity, completeness, V-measure
     Метод k-средних:  (0.007688426431943832, 0.018594687735175114, 0.010878763282603301)
     Агломеративная кластеризация:  (0.008322195451040073, 0.02144118638782401, 0.011990421302773663)
     Gaussian Mixture:  (0.010930162417335925, 0.014545617353848631, 0.012481342009229741)
```