## ИУ5-61Б Дьячков М. Ю. РК №1

#### Оглавление

- 1. Задание
- 2. Описание датасета
- 3. Импорт библиотек
- 4. Загрузка и первичный анализ данных
- 5. Диаграмма рассеяния для двух стобцов
- 6. Обработка пропусков в данных

## Задание (к оглавлению)

Для заданного набора данных проведите обработку пропусков в данных для одного категориального и одного количественного признака. Какие способы обработки пропусков в данных для категориальных и количественных признаков Вы использовали? Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?

Для пары произвольных колонок данных построить график "Диаграмма рассеяния".

Необходимо подготовить отчет по рубежному контролю и разместить его в Вашем репозитории. Вы можете использовать титульный лист, или в начале ноутбука в текстовой ячейке указать Ваши Ф.И.О. и группу.

### Описание датасета (к оглавлению)

Датасет U.S. Education Datasets: Unification Project создан для образовательной аналитики. Он содержит подробные атрибуты штатов и динамику оценок обучающихся (в США в школе).

### Импорт библиотек (к оглавлению)

import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.impute import SimpleImputer
%matplotlib inline

# Загрузка и первичный анализ данных (к оглавлению)

In [30]: data = pd.read\_csv("states\_all.csv")
 data

Out[30]:		PRIMARY_KEY	STATE	YEAR	ENROLL	TOTAL_REVENUE	FEDERAL_REVENUE
	0	1992_ALABAMA	ALABAMA	1992	NaN	2678885.0	304177.0
	1	1992_ALASKA	ALASKA	1992	NaN	1049591.0	106780.0
	2	1992_ARIZONA	ARIZONA	1992	NaN	3258079.0	297888.0
	3	1992_ARKANSAS	ARKANSAS	1992	NaN	1711959.0	178571.0
	4	1992_CALIFORNIA	CALIFORNIA	1992	NaN	26260025.0	2072470.0
	•••						
	1710	2019_VIRGINIA	VIRGINIA	2019	NaN	NaN	NaN
	1711	2019_WASHINGTON	WASHINGTON	2019	NaN	NaN	NaN
	1712	2019_WEST_VIRGINIA	WEST_VIRGINIA	2019	NaN	NaN	NaN
	1713	2019_WISCONSIN	WISCONSIN	2019	NaN	NaN	NaN
	1714	2019_WYOMING	WYOMING	2019	NaN	NaN	NaN

1715 rows × 25 columns

In [31]: data.shape
Out[31]: (1715, 25)
In [32]: data.dtypes

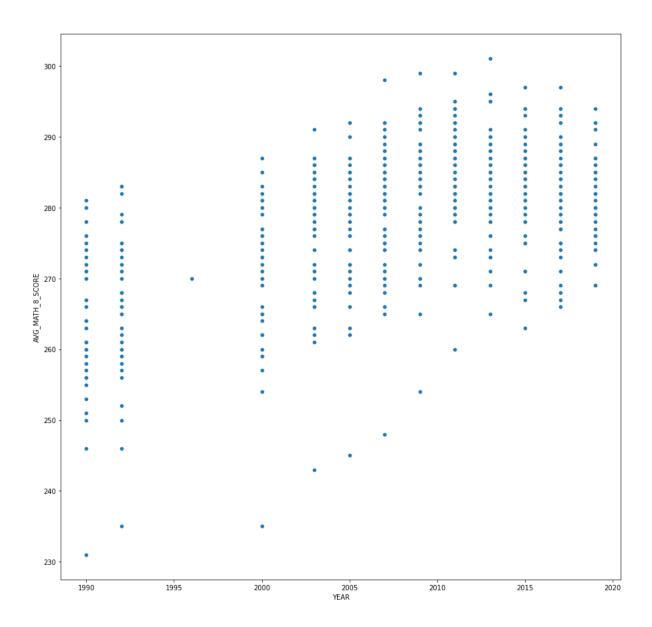
```
object
         PRIMARY_KEY
Out[32]:
         STATE
                                           object
         YEAR
                                            int64
         ENROLL
                                          float64
         TOTAL REVENUE
                                          float64
         FEDERAL_REVENUE
                                          float64
         STATE REVENUE
                                          float64
         LOCAL REVENUE
                                          float64
         TOTAL_EXPENDITURE
                                          float64
         INSTRUCTION_EXPENDITURE
                                          float64
         SUPPORT_SERVICES_EXPENDITURE
                                          float64
         OTHER_EXPENDITURE
                                          float64
         CAPITAL OUTLAY EXPENDITURE
                                          float64
         GRADES PK G
                                          float64
         GRADES_KG_G
                                          float64
         GRADES_4_G
                                          float64
         GRADES_8_G
                                          float64
                                          float64
         GRADES_12_G
         GRADES_1_8_G
                                          float64
         GRADES 9 12 G
                                          float64
                                          float64
         GRADES_ALL_G
         AVG_MATH_4_SCORE
                                          float64
         AVG MATH 8 SCORE
                                          float64
         AVG_READING_4_SCORE
                                          float64
         AVG_READING_8_SCORE
                                          float64
         dtype: object
In [33]:
         # Количество пустых значений
         total_count = data.shape[0]
         for col in data.columns:
              temp null count = data[data[col].isnull()].shape[0]
              temp_perc = round((temp_null_count / total_count) * 100.0, 2)
              print('Колонка {} - {}, {}%'.format(col, temp_null_count, temp_perc))
         Колонка PRIMARY KEY - 0, 0.0%
         Колонка STATE - 0, 0.0%
         Колонка YEAR - 0, 0.0%
         Колонка ENROLL - 491, 28.63%
         Колонка TOTAL_REVENUE - 440, 25.66%
         Колонка FEDERAL_REVENUE - 440, 25.66%
         Колонка STATE_REVENUE - 440, 25.66%
         Колонка LOCAL REVENUE - 440, 25.66%
         Колонка TOTAL EXPENDITURE - 440, 25.66%
         Колонка INSTRUCTION_EXPENDITURE - 440, 25.66%
         Колонка SUPPORT_SERVICES_EXPENDITURE - 440, 25.66%
         Колонка OTHER EXPENDITURE - 491, 28.63%
         Колонка CAPITAL OUTLAY EXPENDITURE - 440, 25.66%
         Колонка GRADES_PK_G - 173, 10.09%
         Колонка GRADES_KG_G - 83, 4.84%
         Колонка GRADES 4 G - 83, 4.84%
         Колонка GRADES 8 G - 83, 4.84%
         Колонка GRADES 12 G - 83, 4.84%
         Колонка GRADES_1_8_G - 695, 40.52%
         Колонка GRADES_9_12_G - 644, 37.55%
         Колонка GRADES_ALL_G - 83, 4.84%
         Колонка AVG_MATH_4_SCORE - 1150, 67.06%
         Колонка AVG_MATH_8_SCORE - 1113, 64.9%
         Колонка AVG_READING_4_SCORE - 1065, 62.1%
         Колонка AVG_READING_8_SCORE - 1153, 67.23%
         data.isnull().sum()
In [34]:
```

```
0
         PRIMARY_KEY
Out[34]:
          STATE
                                              0
          YEAR
                                              0
          ENROLL
                                            491
          TOTAL REVENUE
                                            440
          FEDERAL_REVENUE
                                            440
          STATE_REVENUE
                                            440
          LOCAL REVENUE
                                            440
          TOTAL_EXPENDITURE
                                            440
          INSTRUCTION_EXPENDITURE
                                            440
          SUPPORT_SERVICES_EXPENDITURE
                                            440
          OTHER_EXPENDITURE
                                            491
          CAPITAL_OUTLAY_EXPENDITURE
                                            440
          GRADES PK G
                                            173
          GRADES_KG_G
                                             83
          GRADES_4_G
                                             83
          GRADES_8_G
                                             83
          GRADES_12_G
                                             83
          GRADES_1_8_G
                                            695
          GRADES_9_12_G
                                            644
          GRADES_ALL_G
                                             83
          AVG_MATH_4_SCORE
                                           1150
          AVG_MATH_8_SCORE
                                           1113
                                           1065
          AVG_READING_4_SCORE
          AVG_READING_8_SCORE
                                           1153
          dtype: int64
In [35]: # df = data.drop('LoanedFrom', axis=1)
          # df = data.drop('Unnamed', axis=1)
```

## Диаграмма рассеяния для двух стобцов (к оглавлению)

```
In [36]: fig, ax = plt.subplots(figsize=(15, 15))
    fig.suptitle("Диаграмма рассеяния для колонок ГОД и СРОДНЯЯ ОЦЕНКА ПО МАТЕМАТИКЕ 8
    sns.scatterplot(ax=ax, x='YEAR', y='AVG_MATH_8_SCORE', data=data)

Out[36]: <a href="https://documents.com/white-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-aparts-a
```



## Обработка пропусков в данных (коглавлению)

### Количественный признак

```
In [37]: # Импьютация столбца JerseyNumber с помощью медианы

temp_col = SimpleImputer(strategy='median').fit_transform(data[['AVG_MATH_8_SCORE'
data[['AVG_MATH_8_SCORE']] = temp_col
```

### Категориальный признак

У нас отсутствуют пропуски в категориальных признаках, поэтому сделаем еще одну аимпьютацию количественного признака

```
temp_col = SimpleImputer(strategy='mean').fit_transform(data[['AVG_READING_8_SCORE
```

```
data[['AVG_READING_8_SCORE ']] = temp_col
In [38]:
                  fig, ax = plt.subplots(figsize=(15,9))
In [39]:
                   sns.heatmap(data.corr(method="pearson"), ax=ax,annot=True, fmt=".2f", center = -1)
                  <AxesSubplot:>
Out[39]:
                                           YEAR -1.00 0.03 0.24 0.28 0.22 0.23 0.23 0.23 0.25 0.20 0.16 0.23 0.04 0.05 0.07 0.12 0.01 0.04 0.06 0.64 0.22
                                        ENROLL -0.03 1.00 0.91 0.89 0.91 0.85 0.91 0.88 0.92 0.95 0.92 0.79 1.00 1.00 1.00 0.98 1.00 1.00 1.00 -0.02 -0.03 -0.07 -0.12 -0.05
                                                  024 091 100 093 097 096 100 099 099 095 092 072 089 090 091 094 092 094 091 014 005 007 001 00
                                 TOTAL REVENUE
                                                  .28 0.89 0.93 1.00 0.92 0.85 0.93 0.89 0.93 0.95 0.91 <mark>0.75</mark> 0.88 0.87 0.89 0.93 0.91 0.93 0.89 0.09 0.00 -0.03 -0.13 -0.09
                                                                                                                                                                             - 0.8
                                                  STATE REVENUE -
                                 LOCAL_REVENUE -
                                                  0.23 0.85 0.96 0.85 0.88 1.00 0.97 0.97 0.96 0.87 0.87 0.73 0.82 0.83 0.84 0.86 0.85 0.86 0.84 <mark>0.18 0.08 0.14 0.07 0.0</mark>4
                                                  0.23 0.91 1.00 0.93 0.97 0.97 1.00 0.99 0.99 0.95 0.93 0.73 0.89 0.90 0.91 0.94 0.92 0.94 0.91 0.13 0.05 0.06 4.02 4.02
                              TOTAL_EXPENDITURE
                       INSTRUCTION EXPENDITURE
                                                  0.23 0.88 0.99 0.89 0.95 0.97 0.99 1.00 0.98 0.91 0.90 0.69 0.86 0.86 0.87 0.90 0.88 0.90 0.87 0.14 0.05 0.08 0.00 0.00
                                                  0.25 <mark>0.92 0.99 0.93 0.97 0.96 0.99 0.98 1.00 0.95 0.92 0.73 0.90 0.90 0.91 0.95 0.92 0.94 0.91 <mark>0.14 0.05 0.07 -0.01 -</mark>0.01</mark>
                                                                                                                                                                              - 0.6
                             OTHER EXPENDITURE
                                                  0.00 0.95 0.95 0.95 0.95 0.87 0.95 0.91 0.95 1.00 0.92 0.76 0.95 0.94 0.95 0.97 0.96 0.97 0.95 0.04 0.01 0.02 0.10 0.00
                                                  0.16 0.92 0.92 0.91 0.91 0.87 0.93 0.90 0.92 0.92 1.00 <mark>0.78</mark> 0.89 0.90 0.91 0.92 0.94 0.94 0.91 <mark>0.07 0.01 -0.02 -0.12 -0.0</mark>
                     CAPITAL OUTLAY EXPENDITURE
                                   GRADES PK G -
                                                  0.23 0.79 0.72 0.75 0.66 0.73 0.73 0.69 0.73 0.76 0.78 1.00 0.77 0.78 0.78 0.75 0.79 0.77 0.79 0.11 0.05 0.00 0.05 0.05
                                   GRADES KG G -
                                                  0.04 1.00 0.89 0.88 0.90 0.82 0.89 0.86 0.90 0.95 0.89 0.77 1.00 1.00 0.99 0.98 1.00 0.99 1.00 0.02 0.04 0.10 0.12 0.07
                                                                                                                                                                             -04
                                    GRADES 4 G
                                                  0.05 1.00 0.90 0.87 0.90 0.83 0.90 0.86 0.90 0.94 0.90 0.78 1.00 1.00 1.00 0.98 1.00 0.99 1.00
                                                  0.07 1.00 0.91 0.89 0.91 0.84 0.91 0.87 0.91 0.95 0.91 0.78 0.99 1.00 1.00 0.98 1.00 1.00 1.00 0.01 0.03 0.08 0.11 0.0
                                    GRADES 8 G -
                                   GRADES_12_G -0.12 0.98 0.94 0.93 0.95 0.86 0.94 0.90 0.95 0.97 0.92 0.75 0.98 0.98 0.98 1.00 0.98 0.99 0.98 0.04 0.00 -0.03 -0.07 -0.04
                                  GRADES 1 8 G -
                                                  0.01 1.00 0.92 0.91 0.92 0.85 0.92 0.88 0.92 0.96 0.94 <mark>0.79</mark> 1.00 1.00 1.00 0.98 1.00 1.00 1.00 -0.02 -0.03 -0.08 -0.13 -0.08
                                                                                                                                                                              0.2
                                 GRADES 9 12 G -
                                                  0.04 1.00 0.94 0.93 0.94 0.86 0.94 0.90 0.94 0.97 0.94 <mark>0.77</mark> 0.99 0.99 1.00 0.99 1.00 1.00 1.00 -0.00 -0.02 -0.05 -0.10 -0.0
                                                 0.06 100 0.91 0.89 0.91 0.84 0.91 0.87 0.91 0.95 0.91 0.79 100 100 100 0.98 100 100 100 4.01 4.03 4.08 4.11 4.07
                                  GRADES ALL G -
                              AVG_MATH_4_SCORE -0.64 -0.02 0.14 0.09 0.10 0.18 0.13 0.14 0.04 0.07 0.11 -0.02 0.03 0.01 0.04 0.02 0.00 0.01 1.00 0.91 0.77 0.81 0
                                                  0.0
                           AVG READING 4 SCORE -
                           AVG_READING_8_SCORE - 0.17 -0.12 -0.01 -0.13 -0.07 -0.07 -0.02 -0.00 -0.01 -0.10 -0.12 -0.05 -0.12 -0.12 -0.12 -0.13 -0.10 -0.13 -0.10 -0.11 -0.81 -0.81 -0.90 -1.00 -1.00
                          AVG_READING_8_SCORE -
                                                                                                              GRADES 4 G
                                                                                                     GRADES_PK_G
                                                                                                         GRADES_KG_G
                                                                                                                        GRADES_12_G
                                                                                                                             GRADES 1 8 G
                                                                                                                                 GRADES_9_12_G
                                                                                                                                      GRADES_ALL_G
                                                           OTAL REVENUE
                                                                EDERAL_REVENUE
                                                                    STATE_REVENUE
                                                                         LOCAL_REVENUE
                                                                                   INSTRUCTION_EXPENDITURE
                                                                                       SUPPORT_SERVICES_EXPENDITURE
                                                                                            OTHER EXPENDITURE
                                                                                                 CAPITAL_OUTLAY_EXPENDITURE
                                                                                                                                           WG_MATH_4_SCORE
                                                                                                                                                AVG_MATH_8_SCORE
                                                                                                                                                    WG_READING_4_SCORE
```

#### Ответы

Для количественных признаков использовалась замена медианой.

Для дальнейшего построения моделей я бы использовал признаки, имеющие максимальную взаимную корреляцию, а некоррелирующие не использовал бы.