Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №1
по дисциплине
«Методы машинного обучения»
на тему
«Разведочный анализ данных. Исследование и визуализация данных»

Выполнил:
студент группы ИУ5-24М
Мельников К. И.

Москва — 2019 г.

```
In [1]: import pandas as pd
        import seaborn as sn

In [2]: dt = pd.read_csv('StudentsPerformance.csv', header=0)

In [5]: dt
```

Out[5]:

	gender	race/ethnicity	parental level of education	lunch	\
0	female	group B	bachelor's degree	standard	
1	female	group C	some college	standard	
2	female	group B	master's degree	standard	
3	male	group A	associate's degree	free/reduced	
4	male	group C	some college	standard	
5	female	group B	associate's degree	standard	
6	female	group B	some college	standard	
7	male	group B	some college	free/reduced	
8	male	group D	high school	free/reduced	
9	female	group B	high school	free/reduced	
10	male	group C	associate's degree	standard	
11	male	group D	associate's degree	standard	
12	female	group B	high school	standard	
13	male	group A	some college	standard	
14	female	group A	master's degree	standard	
15	female	group C	some high school	standard	
16	male	group C	high school	standard	
17	female	group B	some high school	free/reduced	
18	male	group C	master's degree	free/reduced	
19	female	group C	associate's degree	free/reduced	
20	male	group D	high school	standard	
21	female	group B	some college	free/reduced	
22	male	group D	some college	standard	
23	female	group C	some high school	standard	
24	male	group D	bachelor's degree	free/reduced	
25	male	group A	master's degree	free/reduced	
26	male	group B	some college	standard	
27	female	group C	bachelor's degree	standard	
28	male	group C	high school	standard	
29	female	group D	master's degree	standard	
..	…	…	…	…	
970	female	group D	bachelor's degree	standard	
971	male	group C	some high school	standard	
972	female	group A	high school	free/reduced	
973	female	group D	some college	free/reduced	
974	female	group A	some college	standard	
975	female	group C	some college	standard	
976	male	group B	some college	free/reduced	
977	male	group C	associate's degree	standard	
978	male	group D	high school	standard	
979	female	group C	associate's degree	standard	
980	female	group B	high school	free/reduced	
981	male	group D	some high school	standard	

982	male	group B	some high school	standard
983	female	group A	some college	standard
984	female	group C	some high school	standard
985	male	group A	high school	standard
986	female	group C	associate's degree	standard
987	male	group E	some high school	standard
988	female	group A	some high school	free/reduced
989	female	group D	some college	free/reduced
990	male	group E	high school	free/reduced
991	female	group B	some high school	standard
992	female	group D	associate's degree	free/reduced
993	female	group D	bachelor's degree	free/reduced
994	male	group A	high school	standard
995	female	group E	master's degree	standard
996	male	group C	high school	free/reduced
997	female	group C	high school	free/reduced
998	female	group D	some college	standard
999	female	group D	some college	free/reduced

	test preparation course	math score	reading score	writing score
0	none	72	72	74
1	completed	69	90	88
2	none	90	95	93
3	none	47	57	44
4	none	76	78	75
5	none	71	83	78
6	completed	88	95	92
7	none	40	43	39
8	completed	64	64	67
9	none	38	60	50
10	none	58	54	52
11	none	40	52	43
12	none	65	81	73
13	completed	78	72	70
14	none	50	53	58
15	none	69	75	78
16	none	88	89	86
17	none	18	32	28
18	completed	46	42	46
19	none	54	58	61
20	none	66	69	63
21	completed	65	75	70
22	none	44	54	53
23	none	69	73	73
24	completed	74	71	80
25	none	73	74	72
26	none	69	54	55
27	none	67	69	75
28	none	70	70	65
29	none	62	70	75

```
..                ...   ...   ...   ...
970          none    89   100   100
971     completed    78    72    69
972     completed    53    50    60
973          none    49    65    61
974          none    54    63    67
975     completed    64    82    77
976     completed    60    62    60
977          none    62    65    58
978     completed    55    41    48
979          none    91    95    94
980          none     8    24    23
981          none    81    78    78
982     completed    79    85    86
983     completed    78    87    91
984          none    74    75    82
985          none    57    51    54
986          none    40    59    51
987     completed    81    75    76
988          none    44    45    45
989     completed    67    86    83
990     completed    86    81    75
991     completed    65    82    78
992          none    55    76    76
993          none    62    72    74
994          none    63    63    62
995     completed    88    99    95
996          none    62    55    55
997     completed    59    71    65
998     completed    68    78    77
999          none    77    86    86

[1000 rows x 8 columns]
```

In [52]: dt.shape

Out[52]: (1000, 13)

In [7]: dt.dtypes

```
Out[7]: gender                         object
        race/ethnicity                 object
        parental level of education    object
        lunch                          object
        test preparation course        object
        math score                      int64
        reading score                   int64
        writing score                   int64
        dtype: object
```

In [17]: dt.describe()

Out[17]:

	math score	reading score	writing score	gender_cat
count	1000.00000	1000.000000	1000.000000	1000.000000
mean	66.08900	69.169000	68.054000	0.482000
std	15.16308	14.600192	15.195657	0.499926
min	0.00000	17.000000	10.000000	0.000000
25%	57.00000	59.000000	57.750000	0.000000
50%	66.00000	70.000000	69.000000	0.000000
75%	77.00000	79.000000	79.000000	1.000000
max	100.00000	100.000000	100.000000	1.000000

In [12]:
```
dt['gender'] = dt['gender'].astype('category')
dt['race/ethnicity'] = dt['race/ethnicity'].astype('category')
dt['parental level of education'] = dt['parental level of education'].as
dt['test preparation course'] = dt['test preparation course'].astype('ca
dt['lunch'] = dt['lunch'].astype('category')
```

In [18]:
```
dt['gender_cat'] = dt['gender'].cat.codes
dt['race/ethnicity_cat'] = dt['race/ethnicity'].cat.codes
dt['parental level of education_cat'] = dt['parental level of education'
dt['test preparation course_cat'] = dt['test preparation course'].cat.co
dt['lunch_cat'] = dt['lunch'].cat.codes
```

In [19]:
```
dt.describe()
```

Out[19]:

	math score	reading score	writing score	gender_cat	\
count	1000.00000	1000.000000	1000.000000	1000.000000	
mean	66.08900	69.169000	68.054000	0.482000	
std	15.16308	14.600192	15.195657	0.499926	
min	0.00000	17.000000	10.000000	0.000000	
25%	57.00000	59.000000	57.750000	0.000000	
50%	66.00000	70.000000	69.000000	0.000000	
75%	77.00000	79.000000	79.000000	1.000000	
max	100.00000	100.000000	100.000000	1.000000	

	race/ethnicity_cat	parental level of education_cat	\
count	1000.000000	1000.000000	
mean	2.174000	2.486000	
std	1.157179	1.829522	
min	0.000000	0.000000	
25%	1.000000	1.000000	
50%	2.000000	2.000000	
75%	3.000000	4.000000	
max	4.000000	5.000000	

	test preparation course_cat	lunch_cat
count	1000.000000	1000.000000
mean	0.642000	0.645000
std	0.479652	0.478753
min	0.000000	0.000000
25%	0.000000	0.000000
50%	1.000000	1.000000

```
75%                          1.000000     1.000000
max                          1.000000     1.000000
```

In [20]:
```
dt
```

Out[20]:

	gender	race/ethnicity	parental level of education	lunch	\
0	female	group B	bachelor's degree	standard	
1	female	group C	some college	standard	
2	female	group B	master's degree	standard	
3	male	group A	associate's degree	free/reduced	
4	male	group C	some college	standard	
5	female	group B	associate's degree	standard	
6	female	group B	some college	standard	
7	male	group B	some college	free/reduced	
8	male	group D	high school	free/reduced	
9	female	group B	high school	free/reduced	
10	male	group C	associate's degree	standard	
11	male	group D	associate's degree	standard	
12	female	group B	high school	standard	
13	male	group A	some college	standard	
14	female	group A	master's degree	standard	
15	female	group C	some high school	standard	
16	male	group C	high school	standard	
17	female	group B	some high school	free/reduced	
18	male	group C	master's degree	free/reduced	
19	female	group C	associate's degree	free/reduced	
20	male	group D	high school	standard	
21	female	group B	some college	free/reduced	
22	male	group D	some college	standard	
23	female	group C	some high school	standard	
24	male	group D	bachelor's degree	free/reduced	
25	male	group A	master's degree	free/reduced	
26	male	group B	some college	standard	
27	female	group C	bachelor's degree	standard	
28	male	group C	high school	standard	
29	female	group D	master's degree	standard	
..	…	…	…	…	
970	female	group D	bachelor's degree	standard	
971	male	group C	some high school	standard	
972	female	group A	high school	free/reduced	
973	female	group D	some college	free/reduced	
974	female	group A	some college	standard	
975	female	group C	some college	standard	
976	male	group B	some college	free/reduced	
977	male	group C	associate's degree	standard	
978	male	group D	high school	standard	
979	female	group C	associate's degree	standard	
980	female	group B	high school	free/reduced	
981	male	group D	some high school	standard	
982	male	group B	some high school	standard	
983	female	group A	some college	standard	

984	female	group C	some high school	standard	
985	male	group A	high school	standard	
986	female	group C	associate's degree	standard	
987	male	group E	some high school	standard	
988	female	group A	some high school	free/reduced	
989	female	group D	some college	free/reduced	
990	male	group E	high school	free/reduced	
991	female	group B	some high school	standard	
992	female	group D	associate's degree	free/reduced	
993	female	group D	bachelor's degree	free/reduced	
994	male	group A	high school	standard	
995	female	group E	master's degree	standard	
996	male	group C	high school	free/reduced	
997	female	group C	high school	free/reduced	
998	female	group D	some college	standard	
999	female	group D	some college	free/reduced	

	test preparation course	math score	reading score	writing score	\
0	none	72	72	74	
1	completed	69	90	88	
2	none	90	95	93	
3	none	47	57	44	
4	none	76	78	75	
5	none	71	83	78	
6	completed	88	95	92	
7	none	40	43	39	
8	completed	64	64	67	
9	none	38	60	50	
10	none	58	54	52	
11	none	40	52	43	
12	none	65	81	73	
13	completed	78	72	70	
14	none	50	53	58	
15	none	69	75	78	
16	none	88	89	86	
17	none	18	32	28	
18	completed	46	42	46	
19	none	54	58	61	
20	none	66	69	63	
21	completed	65	75	70	
22	none	44	54	53	
23	none	69	73	73	
24	completed	74	71	80	
25	none	73	74	72	
26	none	69	54	55	
27	none	67	69	75	
28	none	70	70	65	
29	none	62	70	75	
..	…	…	…	…	
970	none	89	100	100	

971	completed	78	72	69
972	completed	53	50	60
973	none	49	65	61
974	none	54	63	67
975	completed	64	82	77
976	completed	60	62	60
977	none	62	65	58
978	completed	55	41	48
979	none	91	95	94
980	none	8	24	23
981	none	81	78	78
982	completed	79	85	86
983	completed	78	87	91
984	none	74	75	82
985	none	57	51	54
986	none	40	59	51
987	completed	81	75	76
988	none	44	45	45
989	completed	67	86	83
990	completed	86	81	75
991	completed	65	82	78
992	none	55	76	76
993	none	62	72	74
994	none	63	63	62
995	completed	88	99	95
996	none	62	55	55
997	completed	59	71	65
998	completed	68	78	77
999	none	77	86	86

	gender_cat	race/ethnicity_cat	parental level of education_cat	\
0	0	1	1	
1	0	2	4	
2	0	1	3	
3	1	0	0	
4	1	2	4	
5	0	1	0	
6	0	1	4	
7	1	1	4	
8	1	3	2	
9	0	1	2	
10	1	2	0	
11	1	3	0	
12	0	1	2	
13	1	0	4	
14	0	0	3	
15	0	2	5	
16	1	2	2	
17	0	1	5	
18	1	2	3	

19	0	2	0
20	1	3	2
21	0	1	4
22	1	3	4
23	0	2	5
24	1	3	1
25	1	0	3
26	1	1	4
27	0	2	1
28	1	2	2
29	0	3	3
..	…	…	…
970	0	3	1
971	1	2	5
972	0	0	2
973	0	3	4
974	0	0	4
975	0	2	4
976	1	1	4
977	1	2	0
978	1	3	2
979	0	2	0
980	0	1	2
981	1	3	5
982	1	1	5
983	0	0	4
984	0	2	5
985	1	0	2
986	0	2	0
987	1	4	5
988	0	0	5
989	0	3	4
990	1	4	2
991	0	1	5
992	0	3	0
993	0	3	1
994	1	0	2
995	0	4	3
996	1	2	2
997	0	2	2
998	0	3	4
999	0	3	4

	test preparation	course_cat	lunch_cat
0		1	1
1		0	1
2		1	1
3		1	0
4		1	1
5		1	1

6	0	1
7	1	0
8	0	0
9	1	0
10	1	1
11	1	1
12	1	1
13	0	1
14	1	1
15	1	1
16	1	1
17	1	0
18	0	0
19	1	0
20	1	1
21	0	0
22	1	1
23	1	1
24	0	0
25	1	0
26	1	1
27	1	1
28	1	1
29	1	1
..	…	…
970	1	1
971	0	1
972	0	0
973	1	0
974	1	1
975	0	1
976	0	0
977	1	1
978	0	1
979	1	1
980	1	0
981	1	1
982	0	1
983	0	1
984	1	1
985	1	1
986	1	1
987	0	1
988	1	0
989	0	0
990	0	0
991	0	1
992	1	0
993	1	0
994	1	1

```
995                                0          1
996                                1          0
997                                0          0
998                                0          1
999                                1          0

[1000 rows x 13 columns]
```

In [22]:
```
dt.corr()
```

Out[22]:

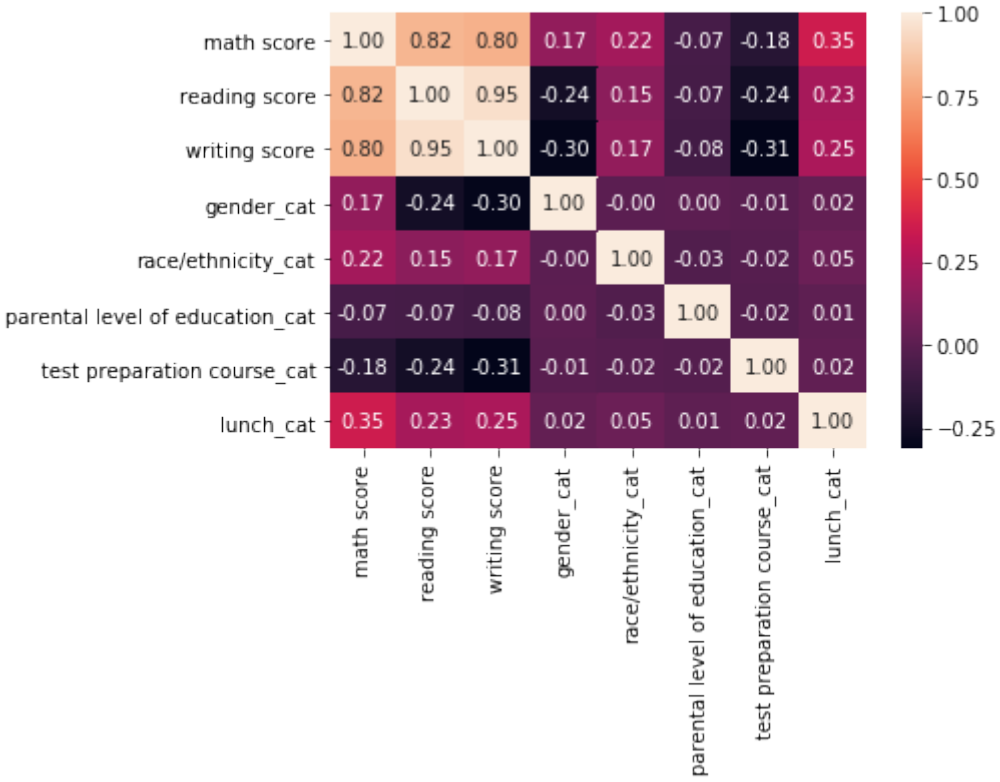
	math score	reading score	writing score
math score	1.000000	0.817580	0.802642
reading score	0.817580	1.000000	0.954598
writing score	0.802642	0.954598	1.000000
gender_cat	0.167982	-0.244313	-0.301225
race/ethnicity_cat	0.216415	0.145253	0.165691
parental level of education_cat	-0.068279	-0.072444	-0.084299
test preparation course_cat	-0.177702	-0.241780	-0.312946
lunch_cat	0.350877	0.229560	0.245769

	gender_cat	race/ethnicity_cat	\
math score	0.167982	0.216415	
reading score	-0.244313	0.145253	
writing score	-0.301225	0.165691	
gender_cat	1.000000	-0.001502	
race/ethnicity_cat	-0.001502	1.000000	
parental level of education_cat	0.001913	-0.031946	
test preparation course_cat	-0.006028	-0.017508	
lunch_cat	0.021372	0.046563	

	parental level of education_cat	\
math score	-0.068279	
reading score	-0.072444	
writing score	-0.084299	
gender_cat	0.001913	
race/ethnicity_cat	-0.031946	
parental level of education_cat	1.000000	
test preparation course_cat	-0.023968	
lunch_cat	0.006320	

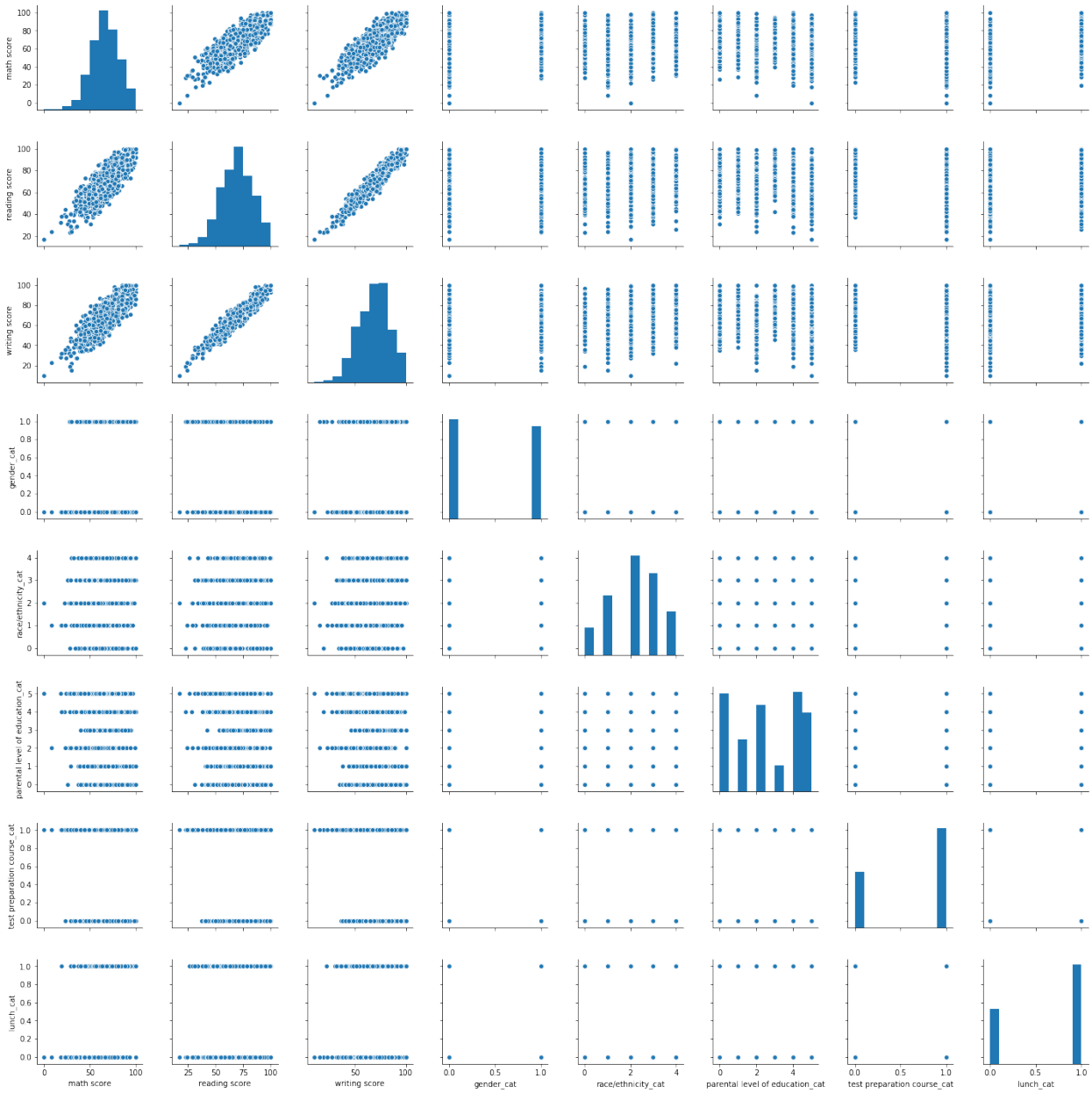
	test preparation course_cat	lunch_cat
math score	-0.177702	0.350877
reading score	-0.241780	0.229560
writing score	-0.312946	0.245769
gender_cat	-0.006028	0.021372
race/ethnicity_cat	-0.017508	0.046563
parental level of education_cat	-0.023968	0.006320
test preparation course_cat	1.000000	0.017044
lunch_cat	0.017044	1.000000

In [25]:
```
sn.heatmap(data=dt.corr(), annot=True, fmt='.2f');
```



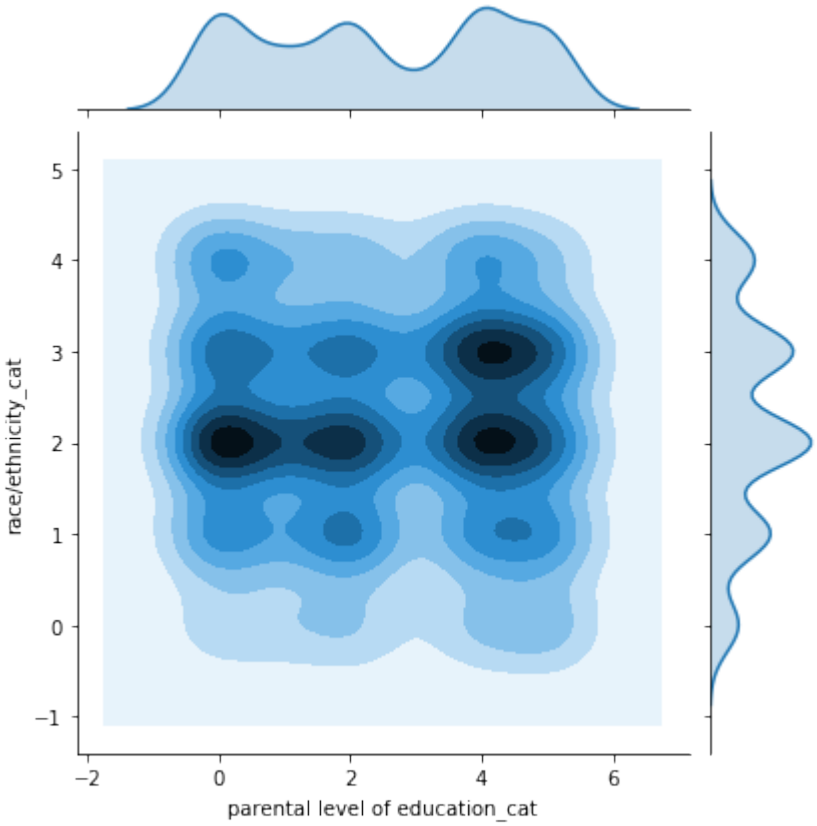
In [27]: `sn.pairplot(dt)`

Out[27]: `<seaborn.axisgrid.PairGrid at 0x7f012746ea20>`



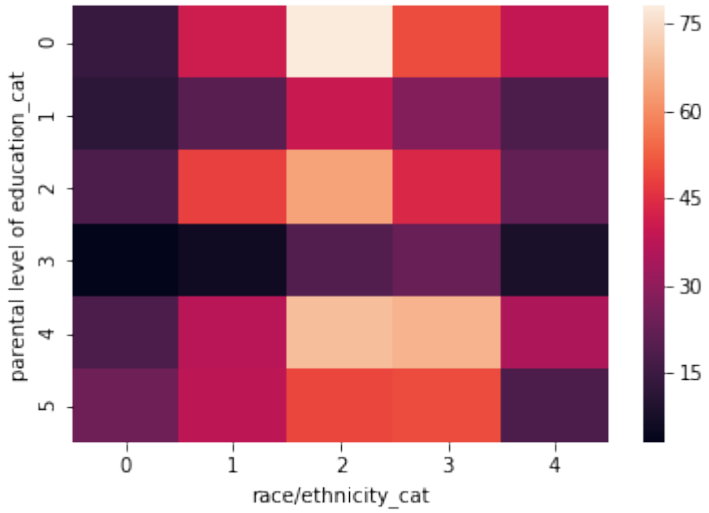

```
In [32]: sn.jointplot(data=dt, x='parental level of education_cat', y='race/ethni
```

```
Out[32]: <seaborn.axisgrid.JointGrid at 0x7f01242c69e8>
```



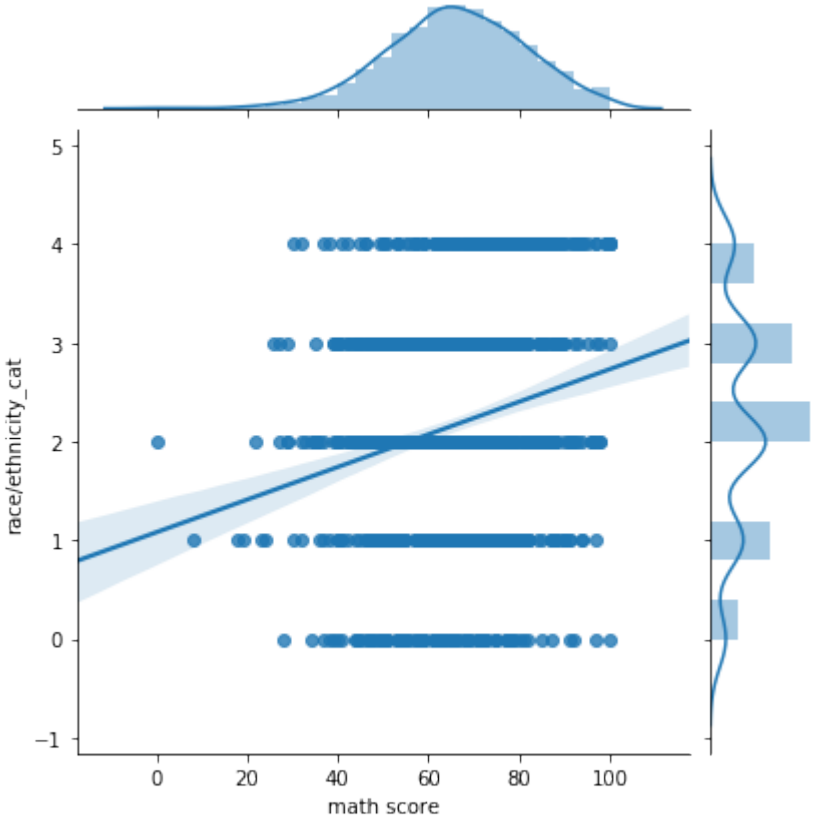
In [34]: sn.heatmap(pd.crosstab(dt['parental level of education_cat'], dt['race/et

Out[34]: <matplotlib.axes._subplots.AxesSubplot at 0x7f01241a4c50>



In [49]: sn.jointplot(data=dt, x='math score', y='race/ethnicity_cat', kind='reg'

Out[49]: <seaborn.axisgrid.JointGrid at 0x7f011c6f12e8>



In [51]: sn.lmplot(data=dt, x='math score', y='test preparation course_cat', hue=

Out[51]: <seaborn.axisgrid.FacetGrid at 0x7f011c4884a8>