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

Лабораторная работа №5
по дисциплине
«Методы машинного обучения»
на тему
«Линейные модели, SVM и деревья решений»

Выполнил:
студент группы ИУ5-22М
Ромичева Е.

Москва — 2019 г.

Рубежный контроль №1

Ромичева Е.В., ИУ5-22М

0.1.1. Задание:

Для заданного набора данных постройте основные графики, входящие в этап разведочного анализа данных. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Какие графики Вы построили и почему? Какие выводы о наборе данных Вы можете сделать на основании построенных графиков?

In [1]:
```
import numpy as np
import pandas as pd
import sklearn
import seaborn as sns
import matplotlib.pyplot as plt
```

In [7]:
```
data = pd.read_csv(r'Admission_Predict.csv', sep=",")
data.head()
```

Out[7]:

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	C
0	1	337	118	4	4.5	4.5	9
1	2	324	107	4	4.0	4.5	8
2	3	316	104	3	3.0	3.5	8
3	4	322	110	3	3.5	2.5	8
4	5	314	103	2	2.0	3.0	8

	Research	Chance of Admit
0	1	0.92
1	1	0.76
2	1	0.72
3	1	0.80
4	0	0.65

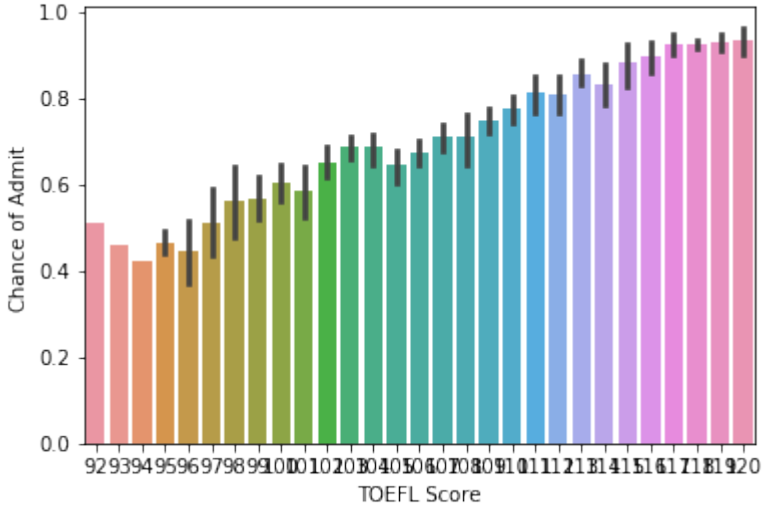
In [8]:
```
data.shape
```

Out[8]:
```
(400, 9)
```

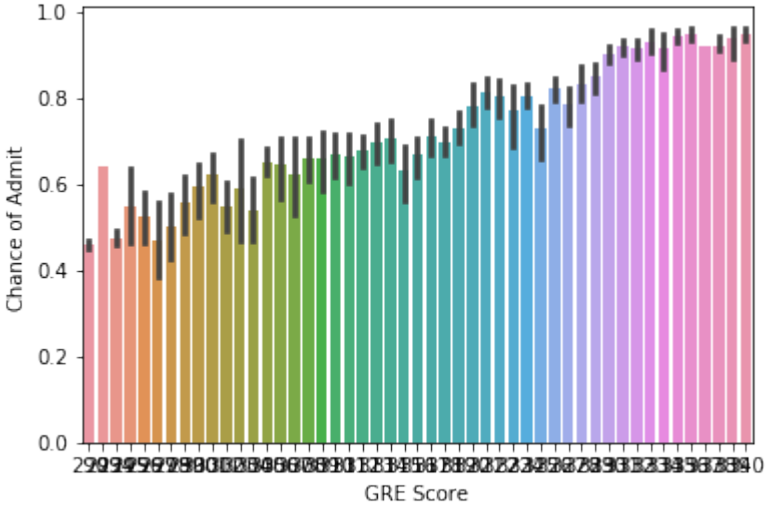
In [9]:
```
for col in data.columns:
    # Количество пустых значений - все значения заполнены
    temp_null_count = data[data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

```
Serial No. - 0
GRE Score - 0
TOEFL Score - 0
University Rating - 0
SOP - 0
LOR  - 0
CGPA - 0
Research - 0
Chance of Admit  - 0
```

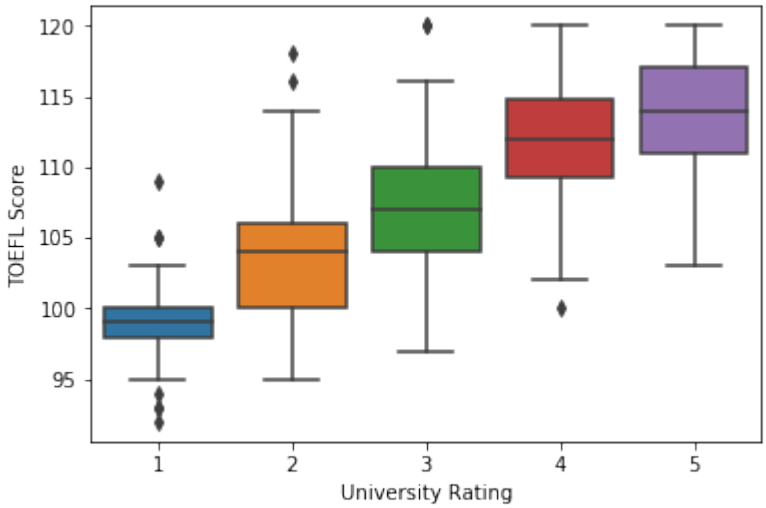
```
In [11]: ax = sns.barplot(x="TOEFL Score", y="Chance of Admit ", data=data)
```




```
In [12]: ax = sns.barplot(x="GRE Score", y="Chance of Admit ", data=data)
```




```
In [36]: sns.boxplot('University Rating','TOEFL Score', data = data)

Out[36]: <matplotlib.axes._subplots.AxesSubplot at 0x1473e4400b8>
```



In [29]:
```
corrmat = data.corr()
fig,ax = plt.subplots(figsize = (12,9))
sns.heatmap(corrmat, annot=True, vmax=.8, square=True)
corrmat
```

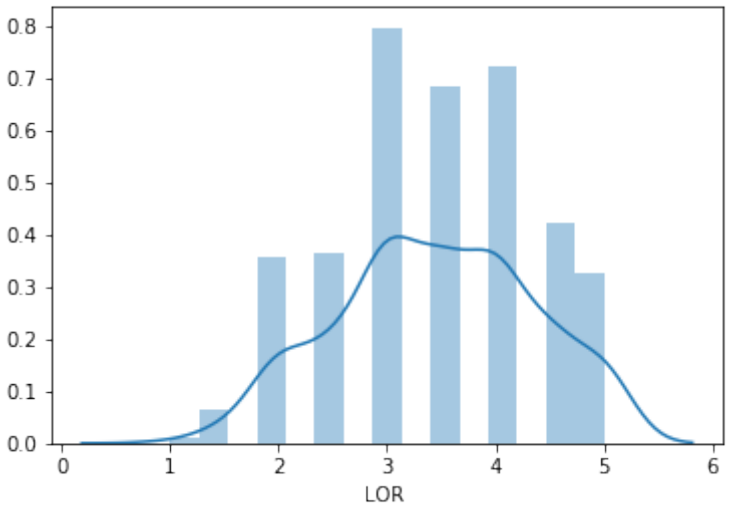
Out[29]:

	Serial No.	GRE Score	TOEFL Score	University Rat
Serial No.	1.000000	-0.097526	-0.147932	-0.169
GRE Score	-0.097526	1.000000	0.835977	0.668
TOEFL Score	-0.147932	0.835977	1.000000	0.695
University Rating	-0.169948	0.668976	0.695590	1.000
SOP	-0.166932	0.612831	0.657981	0.734
LOR	-0.088221	0.557555	0.567721	0.660
CGPA	-0.045608	0.833060	0.828417	0.746
Research	-0.063138	0.580391	0.489858	0.447
Chance of Admit	0.042336	0.802610	0.791594	0.711

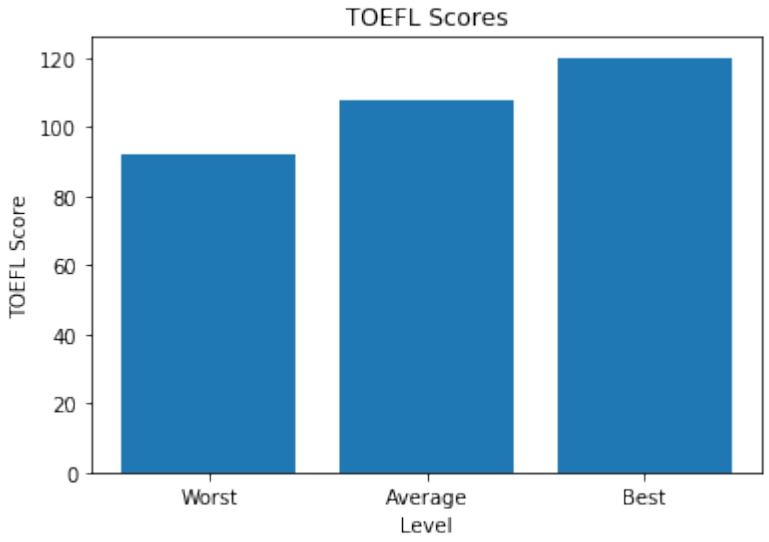
	SOP	LOR	CGPA	Research	Chance of
Serial No.	-0.166932	-0.088221	-0.045608	-0.063138	0.
GRE Score	0.612831	0.557555	0.833060	0.580391	0.
TOEFL Score	0.657981	0.567721	0.828417	0.489858	0.
University Rating	0.734523	0.660123	0.746479	0.447783	0.
SOP	1.000000	0.729593	0.718144	0.444029	0.
LOR	0.729593	1.000000	0.670211	0.396859	0.
CGPA	0.718144	0.670211	1.000000	0.521654	0.
Research	0.444029	0.396859	0.521654	1.000000	0.
Chance of Admit	0.675732	0.669889	0.873289	0.553202	1.

In [19]: sns.distplot(data['LOR '])

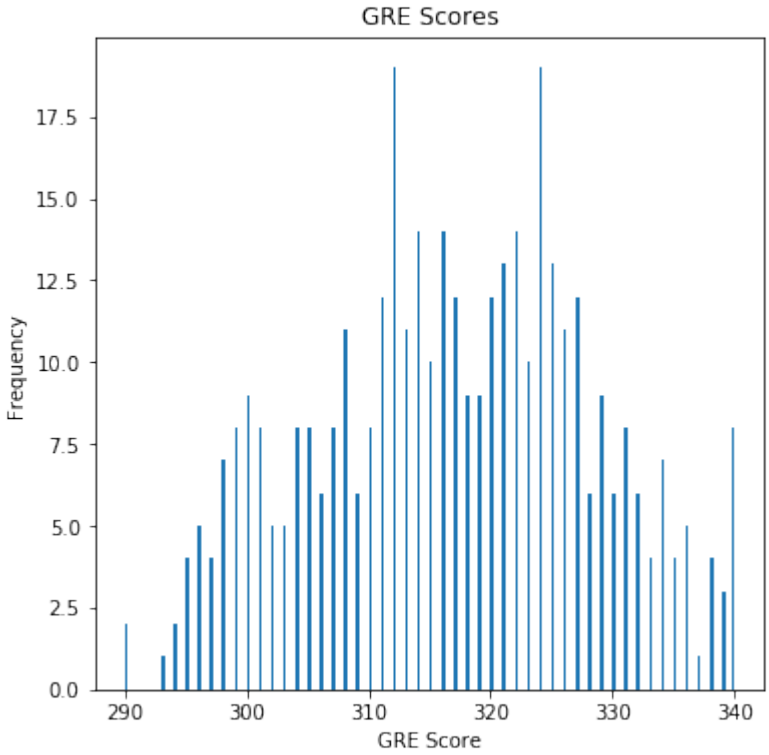
Out[19]: <matplotlib.axes._subplots.AxesSubplot at 0x1473fbefa20>



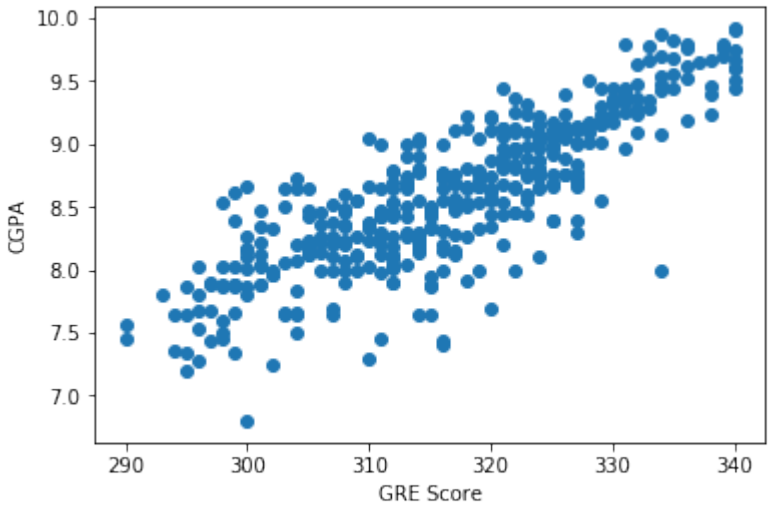
```
In [21]: y = np.array([data["TOEFL Score"].min(),data["TOEFL Score"].mean(),da
         x = ["Worst","Average","Best"]
         plt.bar(x,y)
         plt.title("TOEFL Scores")
         plt.xlabel("Level")
         plt.ylabel("TOEFL Score")
         plt.show()
```



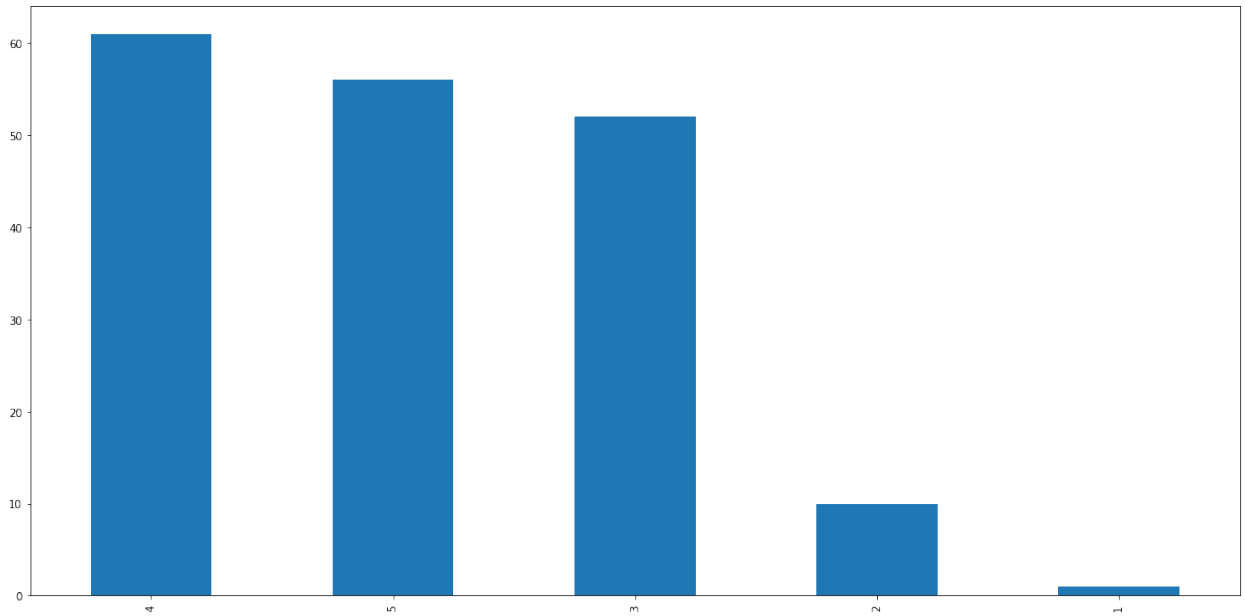

```
In [23]: data["GRE Score"].plot(kind = 'hist',bins = 200,figsize = (6,6))
         plt.xlabel("GRE Score")
         plt.ylabel("Frequency")
         plt.show()
```



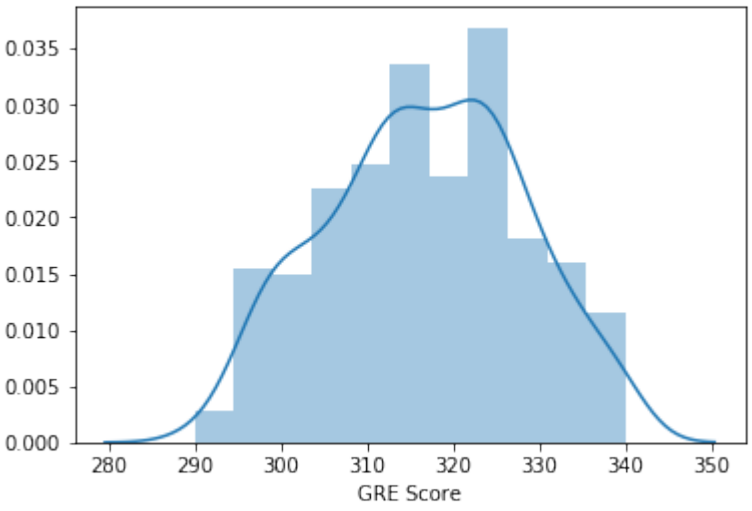

```
In [31]: plt.scatter(data["GRE Score"],data.CGPA)
         plt.xlabel("GRE Score")
         plt.ylabel("CGPA")
         plt.show()
```

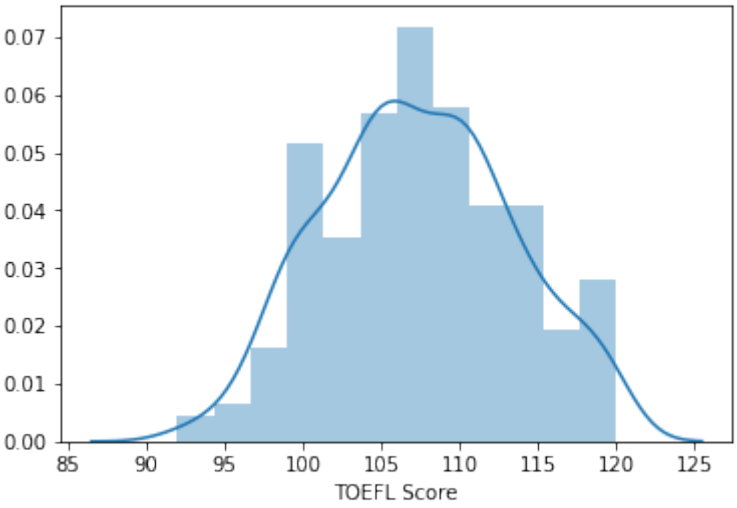


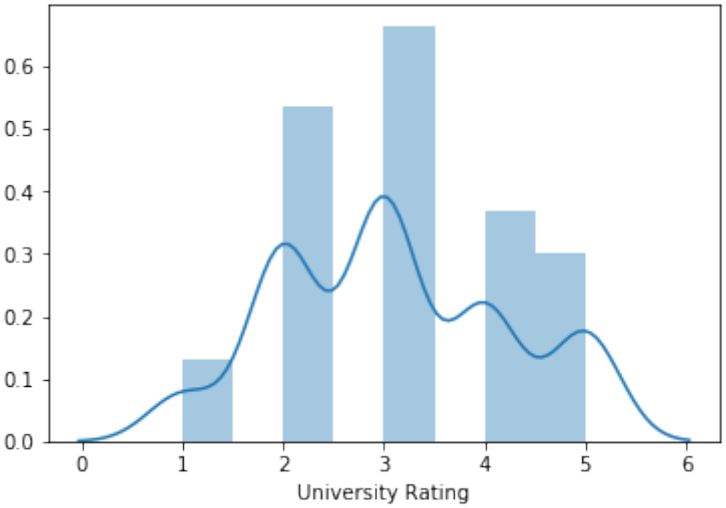

```
In [28]: s = data[data["Chance of Admit "] >= 0.75]["University Rating"].value
         s.plot(kind='bar',figsize=(20, 10))
         plt.show()
```

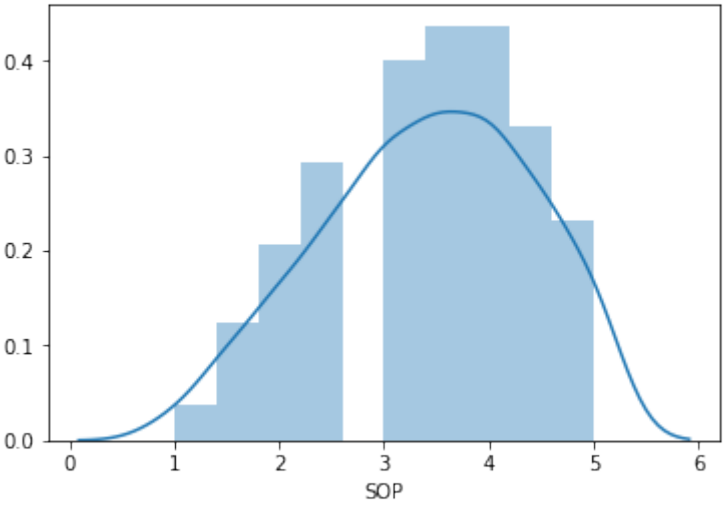


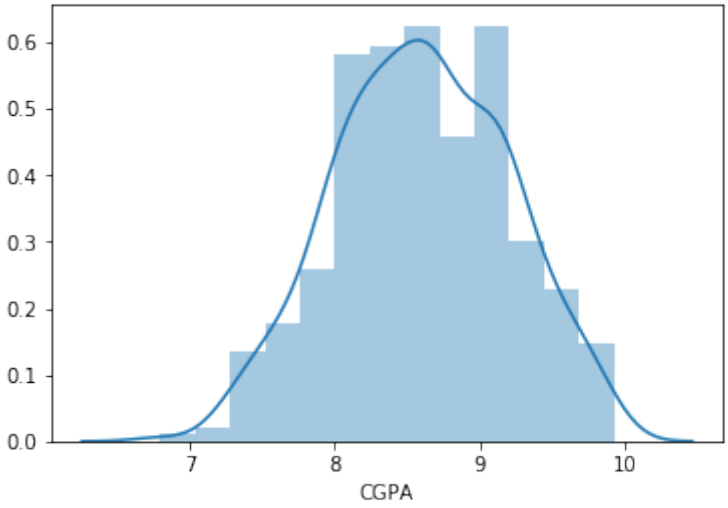

```
In [33]: fig = sns.distplot(data['GRE Score'], kde=True)
         plt.show()

         fig = sns.distplot(data['TOEFL Score'], kde=True)
         plt.show()

         fig = sns.distplot(data['University Rating'], kde=True)
         plt.show()

         fig = sns.distplot(data['SOP'], kde=True)
         plt.show()

         fig = sns.distplot(data['CGPA'], kde=True)
         plt.show()

         plt.show()
```




0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
85
90
95
100
105
110
115
120
125
TOEFL Score


0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
University Rating


0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
SOP



In [35]: `data.describe()`

Out[35]:

	Serial No.	GRE Score	TOEFL Score	University Rating	
count	400.000000	400.000000	400.000000	400.000000	400.00
mean	200.500000	316.807500	107.410000	3.087500	3.40
std	115.614301	11.473646	6.069514	1.143728	1.00
min	1.000000	290.000000	92.000000	1.000000	1.00
25%	100.750000	308.000000	103.000000	2.000000	2.50
50%	200.500000	317.000000	107.000000	3.000000	3.50
75%	300.250000	325.000000	112.000000	4.000000	4.00
max	400.000000	340.000000	120.000000	5.000000	5.00

	LOR	CGPA	Research	Chance of Admit
count	400.000000	400.000000	400.000000	400.000000
mean	3.452500	8.598925	0.547500	0.724350
std	0.898478	0.596317	0.498362	0.142609
min	1.000000	6.800000	0.000000	0.340000
25%	3.000000	8.170000	0.000000	0.640000
50%	3.500000	8.610000	1.000000	0.730000
75%	4.000000	9.062500	1.000000	0.830000
max	5.000000	9.920000	1.000000	0.970000

In [ ]: