Student Performance Data Analysis

Final Project by: Lim Swee Ming
Dataset from Kaggle

Project Overview

Student Performance Dataset

Dataset Source:
https://www.kaggle.com/larsen0966/student-performance-data-set

649 students participated in this survey and a set of data values featuring 31 columns.

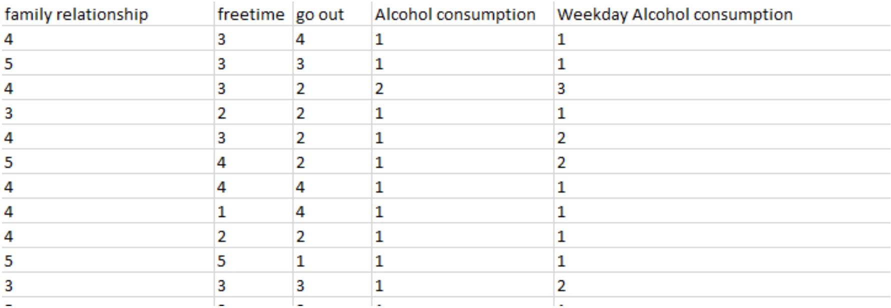
Across left to right, the data values consist of student’s parents education and occupation, student’s social life and their academic grades.

This project scope aims to reveal specific inequalities of the student’s background. Teachers may see them in a better light to guide students who performed poorly and help them succeed in academics.

Data Cleaning

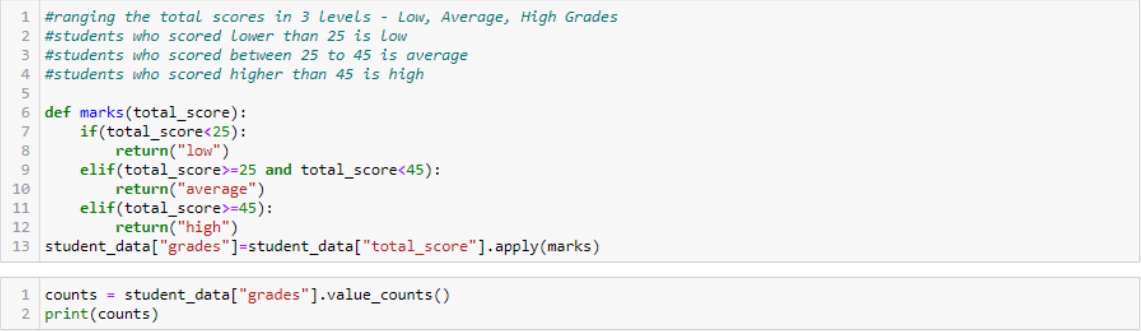
- Check for missing values
- Creating new columns and assigning new values for correlation function at later part

For instance, the dataset comes with three grade periods. Summing up altogether and categorize into low, average high score. Other variables from scale 1 – 5 is also categorize the same manner.



family relationship	freetime	go out	Alcohol consumption	Weekday Alcohol consumption
4	3	4	1	1
5	3	3	1	1
4	3	2	2	3
3	2	2	1	1
4	3	2	1	2
5	4	2	1	2
4	4	4	1	1
4	1	4	1	1
4	2	2	1	1
5	5	1	1	1
3	3	3	1	2

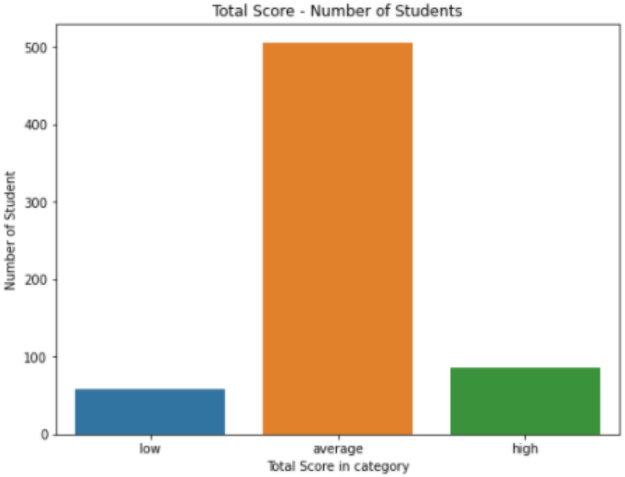
Identify the top/bottom 20 students based on their grades



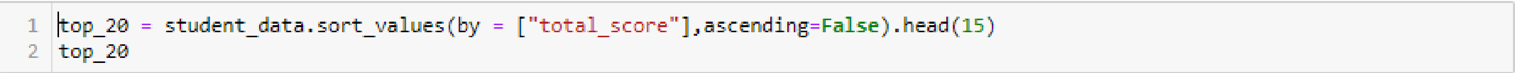
```python
#ranging the total scores in 3 levels - Low, Average, High Grades
#students who scored lower than 25 is low
#students who scored between 25 to 45 is average
#students who scored higher than 45 is high

def marks(total_score):
    if(total_score<25):
        return("low")
    elif(total_score>=25 and total_score<45):
        return("average")
    elif(total_score>=45):
        return("high")
student_data["grades"]=student_data["total_score"].apply(marks)
```

```python
counts = student_data["grades"].value_counts()
print(counts)
```

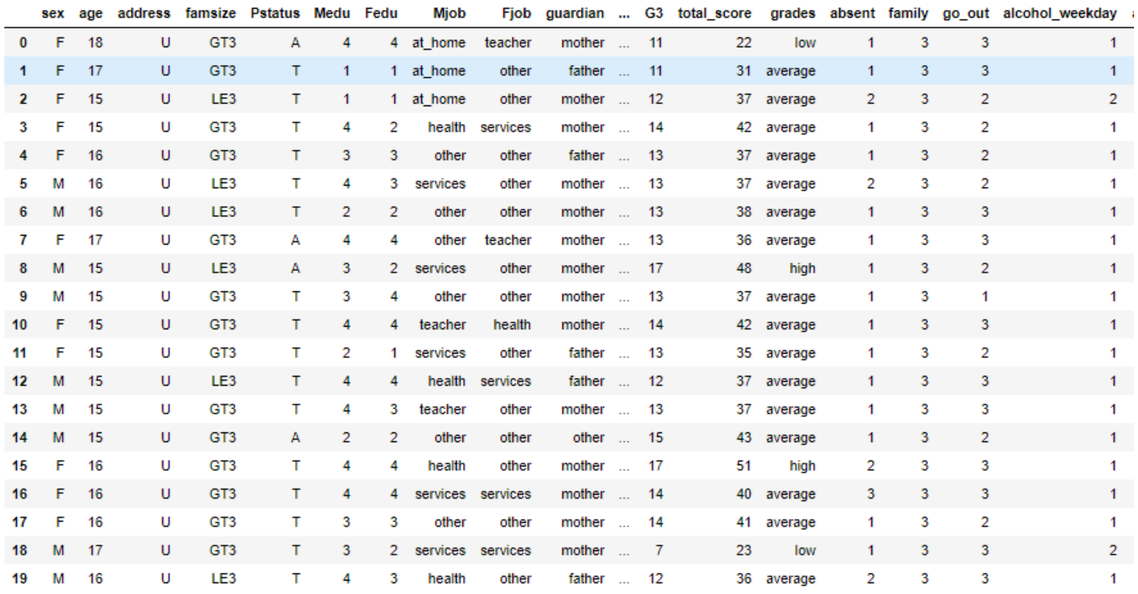
```
average    505
high        86
low         58
Name: grades, dtype: int64
```



Identify the top/bottom 20 students based on their grades

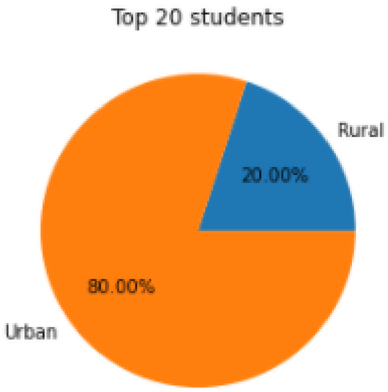


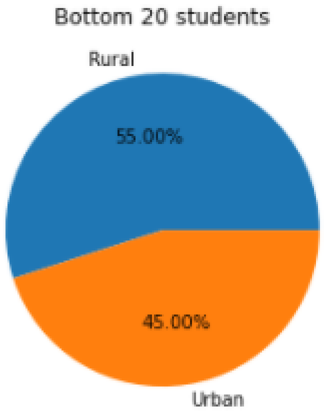
```
top_20 = student_data.sort_values(by = ["total_score"],ascending=False).head(15)
top_20
```



	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	guardian	...	G3	total_score	grades	absent	family	go_out	alcohol_weekday
0	F	18	U	GT3	A	4	4	at_home	teacher	mother	...	11	22	low	1	3	3	1
1	F	17	U	GT3	T	1	1	at_home	other	father	...	11	31	average	1	3	3	1
2	F	15	U	LE3	T	1	1	at_home	other	mother	...	12	37	average	2	3	2	2
3	F	15	U	GT3	T	4	2	health	services	mother	...	14	42	average	1	3	2	1
4	F	16	U	GT3	T	3	3	other	other	father	...	13	37	average	1	3	2	1
5	M	16	U	LE3	T	4	3	services	other	mother	...	13	37	average	2	3	2	1
6	M	16	U	LE3	T	2	2	other	other	mother	...	13	38	average	1	3	3	1
7	F	17	U	GT3	A	4	4	other	teacher	mother	...	13	36	average	1	3	3	1
8	M	15	U	LE3	A	3	2	services	other	mother	...	17	48	high	1	3	2	1
9	M	15	U	GT3	T	3	4	other	other	mother	...	13	37	average	1	3	1	1
10	F	15	U	GT3	T	4	4	teacher	health	mother	...	14	42	average	1	3	3	1
11	F	15	U	GT3	T	2	1	services	other	father	...	13	35	average	1	3	2	1
12	M	15	U	LE3	T	4	4	health	services	father	...	12	37	average	1	3	3	1
13	M	15	U	GT3	T	4	3	teacher	other	mother	...	13	37	average	1	3	3	1
14	M	15	U	GT3	A	2	2	other	other	other	...	15	43	average	1	3	2	1
15	F	16	U	GT3	T	4	4	health	other	mother	...	17	51	high	2	3	3	1
16	F	16	U	GT3	T	4	4	services	services	mother	...	14	40	average	3	3	3	1
17	F	16	U	GT3	T	3	3	other	other	mother	...	14	41	average	1	3	2	1
18	M	17	U	GT3	T	3	2	services	services	mother	...	7	23	low	1	3	3	2
19	M	16	U	LE3	T	4	3	health	other	father	...	12	36	average	2	3	3	1

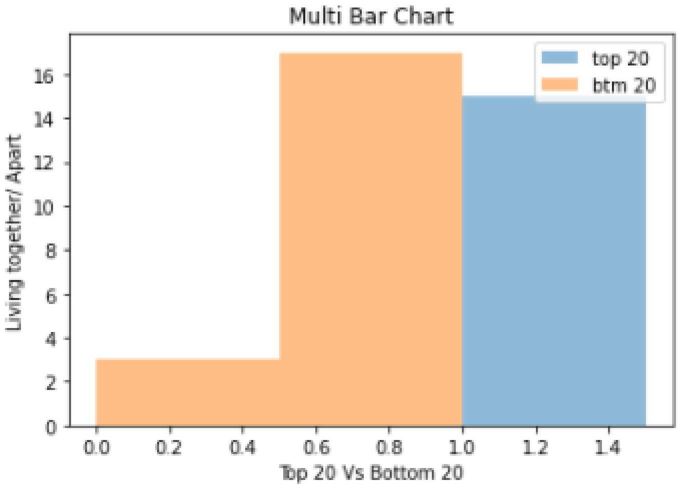
Analysis top/bottom 20 student's living in rural/urban





The unique challenge faced by rural students is the transportation and the language they use might be affecting their scores.

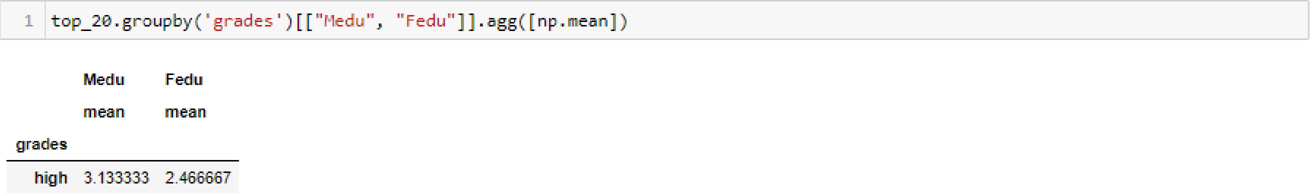
Analysis the top/bottom 20 student's living with/apart parents



Top 20 = 15 students living together with parents and 5 living apart
Btm 20 = 17 students living together with parents and 3 living apart

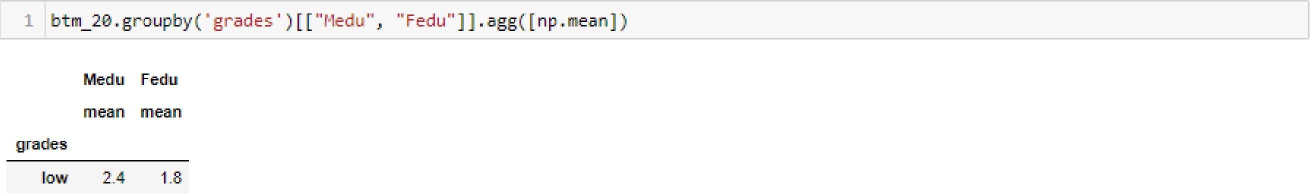
This graph shows no sign of disparity differences.

Analysis the top/bottom 20 student's parents education using mean



```python
top_20.groupby('grades')[["Medu", "Fedu"]].agg([np.mean])
```

	Medu	Fedu
	mean	mean
grades		
high	3.133333	2.466667

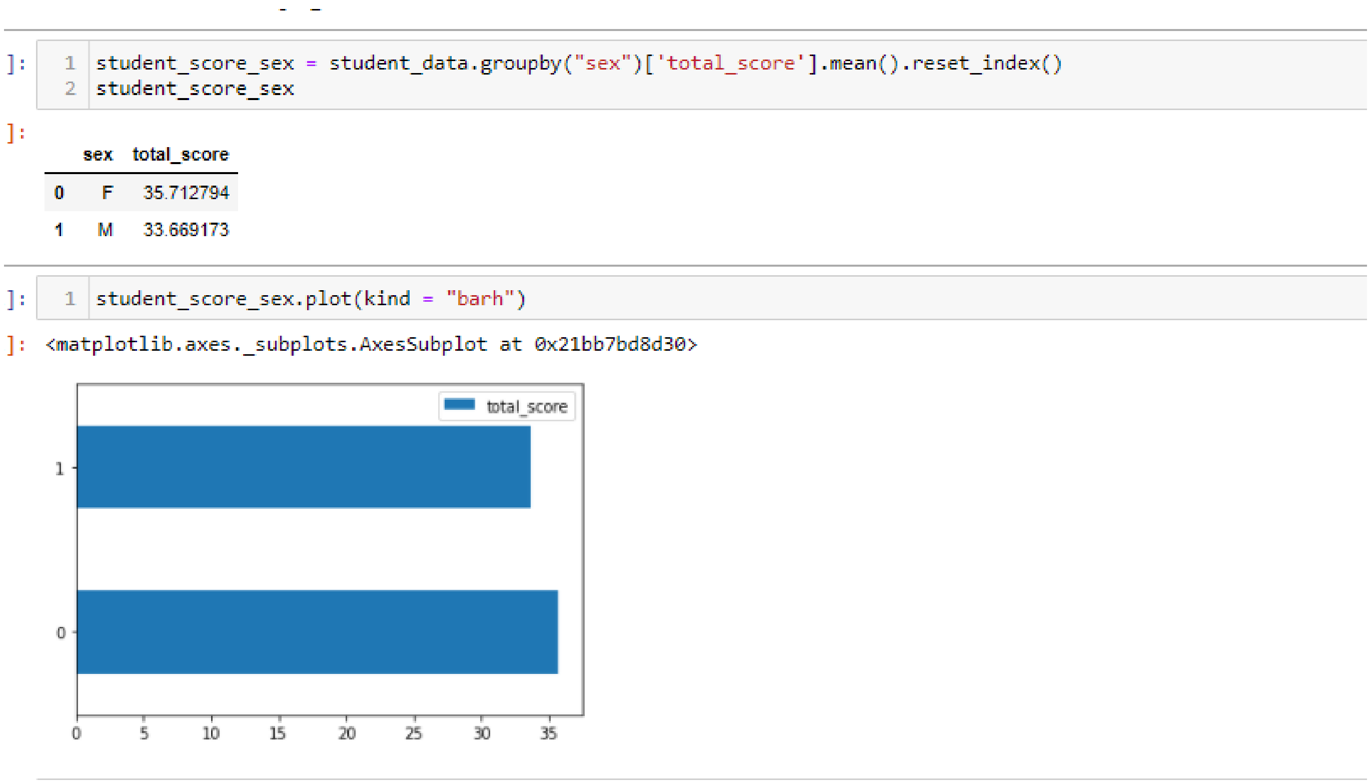


```python
btm_20.groupby('grades')[["Medu", "Fedu"]].agg([np.mean])
```

	Medu	Fedu
	mean	mean
grades		
low	2.4	1.8

From the results, this means most students who are from top 20, both their parents education level is higher than the bottom 20.

Performance by gender



```
]: student_score_sex = student_data.groupby("sex")['total_score'].mean().reset_index()
   student_score_sex
```

]:

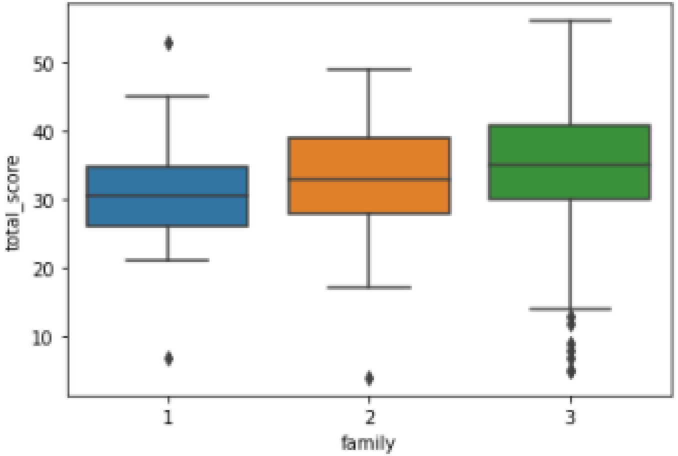
	sex	total_score
0	F	35.712794
1	M	33.669173

```
]: student_score_sex.plot(kind = "barh")
```

```
]: <matplotlib.axes._subplots.AxesSubplot at 0x21bb7bd8d30>
```

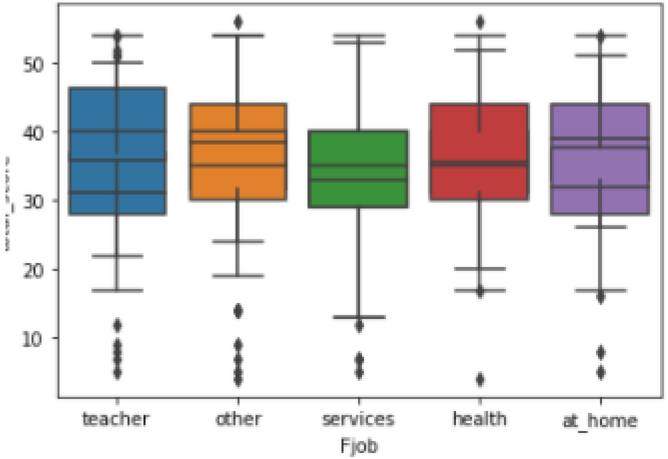
Female and male academic performance is very close. However in fact the assumption of female tend to get better grades than guys might be true in this case.

Does family support have impact towards achievements?



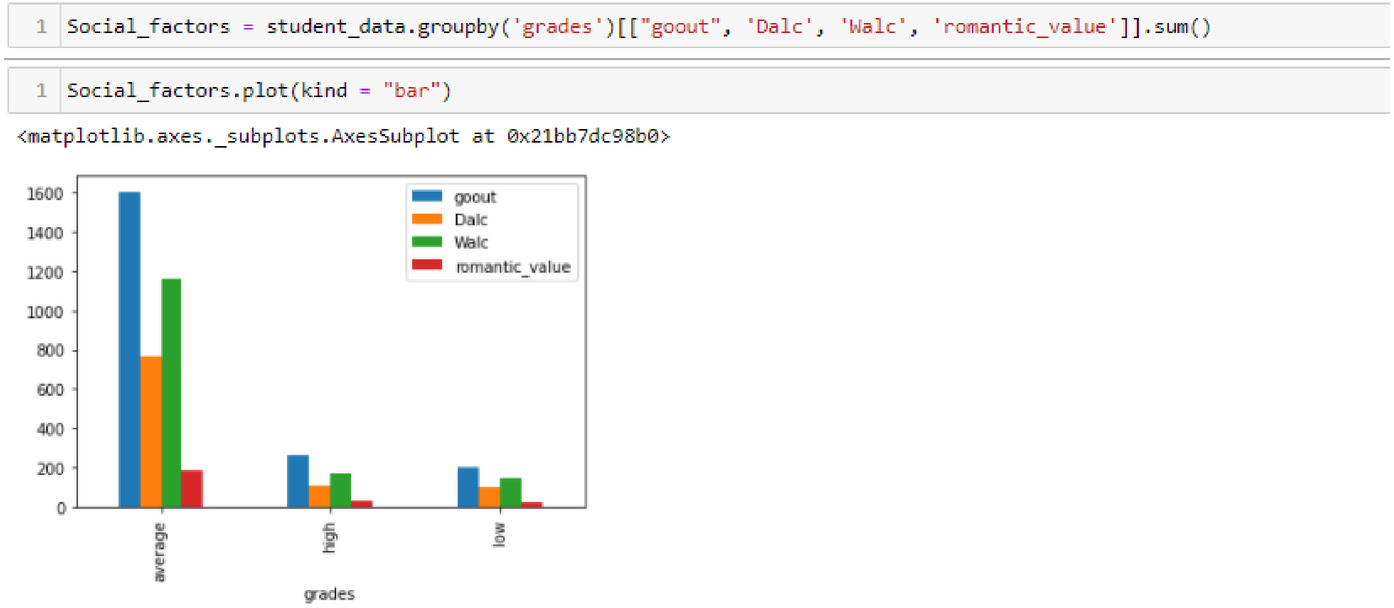
From this observation, family support does have great impact on the grades.
You can tell that students who have family support (3) scores higher. For student who have poor family relationship can be seen as positive skew. In another words, students are likely to score lower than 30.

What type of job parent's hold will have impact towards student achievements?



Parents job can be seen as very impactful towards the scores. For teacher, has both the highest median and greatest variability among all. Services job has a normal distribution. While you compared at home, it has the lowest mean. It also can be seen as positive skew since median is shown closer to the line. Meaning might not be a good sign since it is towards the low scores.

Which social factors has the most influence on academic achievement?



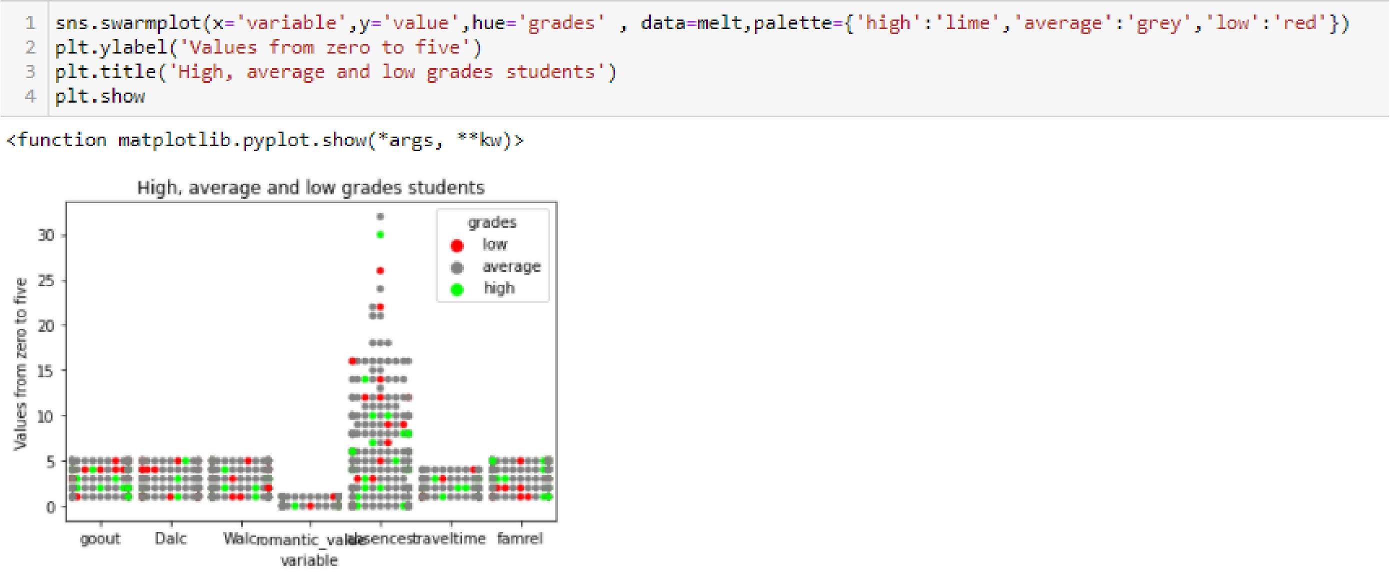
```
Social_factors = student_data.groupby('grades')[["goout", 'Dalc', 'Walc', 'romantic_value']].sum()
```

```
Social_factors.plot(kind = "bar")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x21bb7dc98b0>
```

In visualizing the social factors, what we observed that students of low and high grades have quite a similar distribution.

Which social factors has the most influence on academic achievement?



```
sns.swarmplot(x='variable',y='value',hue='grades' , data=melt,palette={'high':'lime','average':'grey','low':'red'})
plt.ylabel('Values from zero to five')
plt.title('High, average and low grades students')
plt.show
```

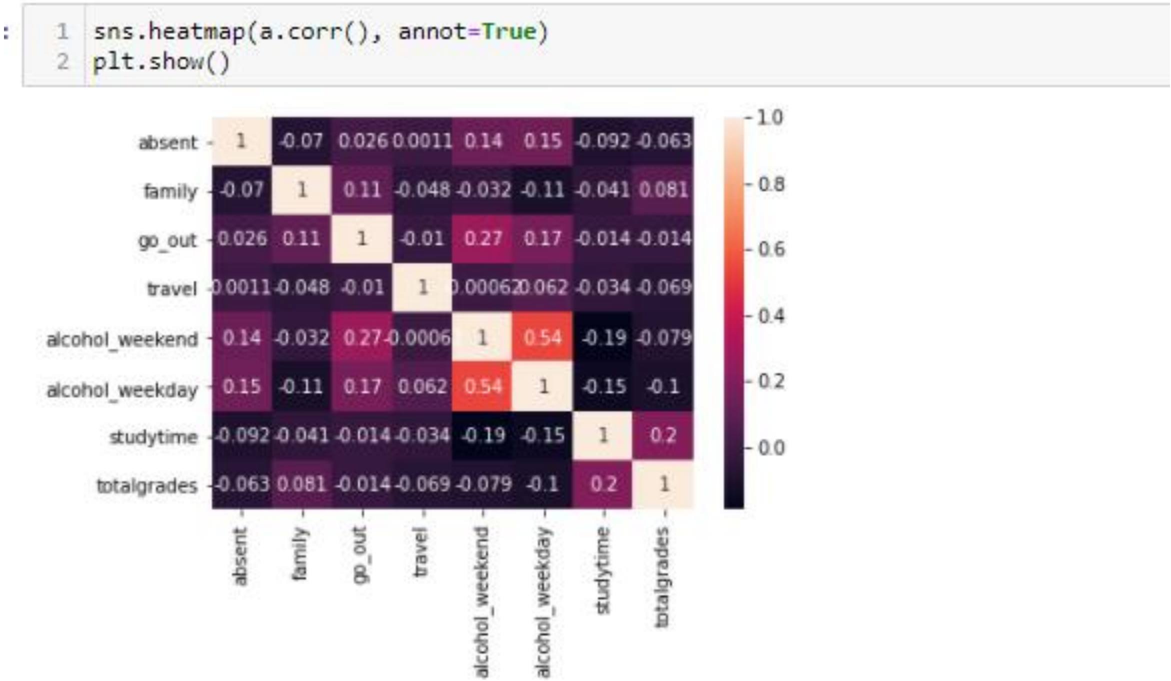
```
<function matplotlib.pyplot.show(*args, **kw)>
```

High, average and low grades students

But if we were to look at this, we can construct a better perspective.

The low scorers tends to go out more often than the high scorers. Weekend, Weekday, romantic and travel time does not constitute significant difference. For family relation, you can tell that most low scorers agree they have poor relationship with family.

Social factors correlation with student performance



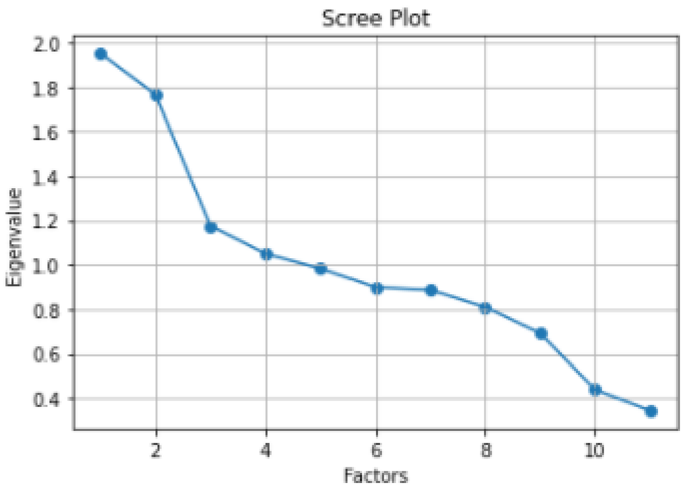
```
sns.heatmap(a.corr(), annot=True)
plt.show()
```

	absent	family	go_out	travel	alcohol_weekend	alcohol_weekday	studytime	totalgrades
absent	1	-0.07	0.026	0.0011	0.14	0.15	-0.092	-0.063
family	-0.07	1	0.11	-0.048	-0.032	-0.11	-0.041	0.081
go_out	0.026	0.11	1	-0.01	0.27	0.17	-0.014	-0.014
travel	0.0011	-0.048	-0.01	1	0.00062	0.062	-0.034	-0.069
alcohol_weekend	0.14	-0.032	0.27	-0.0006	1	0.54	-0.19	-0.079
alcohol_weekday	0.15	-0.11	0.17	0.062	0.54	1	-0.15	-0.1
studytime	-0.092	-0.041	-0.014	-0.034	-0.19	-0.15	1	0.2
totalgrades	-0.063	0.081	-0.014	-0.069	-0.079	-0.1	0.2	1

This perhaps would give an in-depth insight with the correlation method where the social lifestyle are compared with their grades.

Import Factor Analyzer – sklearn function

To identify which factor is the most important for academics to know.



List of factors

1 = total grades
2 = study time
3 = family
4 = father education
5 = mother education
6 = travel
7 = go out
8 = absent
9 = weekday alcohol consumption
10 = weekday alcohol consumption

Notices that the total grades as the highest and dropping down to the next component. To simplify, the closest to the total score means the higher factor the academics must take note of and also take note of the factor that makes students grade drop. Retain factor 2, 3, 4 and 5 which is above the 1 criteria.

Conclusion

Insights we have identified based on the report:

- Students who stay in urban areas are more likely to have higher score
- Students living together or apart with parents has no contribution to their grade
- Parents education level has effects on the student’s grade
- Female academic performance is slightly higher than male
- Higher family support can achieve greater grades
- Parents job has an impact as well to the grades
- Study, family and parent’s education factors are the most important influence for greater achievements while alcohol, absent has adverse effects. Traveling is also affecting the grades. Going out not so much of an impact.

Clearly, teachers are not able to change anything of the student’s family background and can only educate on their social activities and keep an eye on students who scored poorly. The school should promote anti consumption of alcohol and absents.

Thanks!

Any questions?