Welcome to Data Analysis using SQL

- Let's import the two libraries we need to proceed

In [1]:
```
import pandas, sqlite3
```

- Pandas is a package in python
- In this assignment, we use pandas to activate csv files and connect to sql queries

In [2]:
```
db=sqlite3.connect('math.db')
```

- Sqlite3 is a package in python that allows us to use sql queries
- First step from the above code is connecting sql to a database
- We name our database math.db because we are working with student data in terms of mathematics
- Here is a link with description of the dataset
- https://archive.ics.uci.edu/ml/datasets/student+performance

In [3]:
```
for i in range(33):
    chunks=pandas.read_csv('math.csv',chunksize=100_000)
    for chunk in chunks:
        chunk.columns = [ column.replace('','_') for column in chunk.col
umns]
        chunk.to_sql('math',db,if_exists='append')
```

- We use this database as a connection to our csv file in order to run sql queries
- In SQL, we have tables and with this loop, we call this table math
- We also have range(33) because the csv file contains 33 columns
- Let's run some queries and do a data analysis of our dataset

How many observations are there in the table?

In [4]:
```
pandas.read_sql_query('SELECT count(*) FROM math',db)
```

Out[4]:

count(*)
0 13035

- Select count() lists the total number of observations
- There are 13,035 observations in the table

In [5]:
```
pandas.read_sql_query('SELECT * FROM math limit 10',db)
```

Out[5]:

index	s_c_h_o_o_l	_s_e_x_	_a_g_e_	_a_d_d_r_e_s_s_	_f_a_m_i_z_e_	_P_s_t_a_t_u_s_	_
0	0	GP	F	18	U	GT3	A
1	1	GP	F	17	U	GT3	T
2	2	GP	F	15	U	LE3	T
3	3	GP	F	15	U	GT3	T
4	4	GP	F	16	U	GT3	T
5	5	GP	M	16	U	LE3	T
6	6	GP	M	16	U	LE3	T
7	7	GP	F	17	U	GT3	A
8	8	GP	M	15	U	LE3	A
9	9	GP	M	15	U	GT3	T

10 rows × 34 columns

- Select () from math selects the entire dataset
- Limit 10 only takes in the first ten rows, as the image displays above

In [6]:
```
pandas.read_sql_query('SELECT _G_1_, _G_2_ FROM math limit 5',db)
```

Out[6]:

_G_1_	_G_2_
0 5	6
1 5	5
2 7	8
3 15	14
4 6	10

- From this query above, you can see that we can select certain columns and still limit the number of rows
- In the table, we have students between the ages 15 to 22

How many 16 year old students are present in the table?

In [7]:
```
pandas.read_sql_query('SELECT count(*) FROM math WHERE _a_g_e_ = 16',db)
```

Out[7]:

count(*)
0 3432

- Out of all the students, there are 3432 students who are 16 years old

How many of the 16 year olds have a 1st period grade more than 10 out of 20?

In [8]:
```
pandas.read_sql_query('SELECT count(*) FROM math WHERE _a_g_e_ = 16 AND
 _G_1_ > 10',db)
```

Out[8]:

count(*)
0 1881

Of these 1881 students, how many are male and female?

In [9]:
```
pandas.read_sql_query('SELECT count(*) FROM math WHERE _a_g_e_ = 16 AND
 _G_1_ > 10 GROUP BY _s_e_x_',db)
```

Out[9]:

count(*)
0 825
1 1056

- We have binary values where 0 is male and 1 is female

What is the average of these students?

In [10]:
```
pandas.read_sql_query('SELECT avg(_G_1_) FROM math WHERE _a_g_e_ = 16 GR
OUP BY _s_e_x_',db)
```

Out[10]:

avg(_G_1_)
0 10.203704
1 11.740000

- The males have an average of 10.20
- The females have an average of 11.74
- This means of the students who are 16 years old, females have better average first period scores than males
- Let's look in a wider scale

How many males and females are there in the table?

In [11]:
```
pandas.read_sql_query('SELECT count(*) FROM math WHERE _a_g_e_ > 14 GROU
P BY _s_e_x_',db)
```

Out[11]:

count(*)
0 6864
1 6171

- There are more male students than female students in the table

Do males or females performs better overall on all exams?

In [12]:
```
pandas.read_sql_query('SELECT avg(_G_1_),avg(_G_2_),avg(_G_3_) FROM math
 WHERE _a_g_e_ > 14 GROUP BY _s_e_x_',db)
```

Out[12]:

avg(_G_1_)	avg(_G_2_)	avg(_G_3_)
0 10.620182	10.389423	9.965346
1 11.229947	11.074866	10.914439

- Females perform a little better on average than males
- Let's figure out the factors behind this analysis

In [13]:
```
pandas.read_sql_query('SELECT count(*) FROM math WHERE _s_t_u_d_y_t_i_m_
e_ > 2 GROUP BY _s_e_x_',db)
```

Out[13]:

count(*)
0 2244
1 792

- From here, we can see that there are three times as many males who study at least five hours as females do
- So, if males study more, why do females perform better?

In [14]:
```
pandas.read_sql_query('SELECT count(*) FROM math WHERE _a_b_s_e_n_c_e_s_
> 20 GROUP BY _s_e_x_',db)
```

Out[14]:

count(*)
0 396
1 99

- This query involves the number of males and females whose absences exceed more than 20
- We can see that there are four times as many males who are absent from class over females
- Usually, people who skip class are at more risk to fail exams

In [15]:
```
pandas.read_sql_query('SELECT count(*) FROM math WHERE _g_o_o_u_t_ > 3 A
ND _D_a_l_c_ > 3 AND _W_a_l_c_ > 3 GROUP BY _s_e_x_',db)
```

Out[15]:

count(*)
0 66
1 330

- This is a query that discusses friends going out and having consumption of alcohol on a workday to weekend basis
- There are five times as many women that commit to this
- It would be fair to say that these factors may not affect performance because usually the age to drink is after 18 and then on top of that, the sample of people is a very small sample

In [16]:
```
pandas.read_sql_query('SELECT count(*) FROM math WHERE _f_r_e_e_t_i_m_e_
> 2 GROUP BY _s_e_x_',db)
```

Out[16]:

count(*)
0 5115
1 5181

- Free time after school is equivalent for both groups
- That free time could be divided into extracurricular activities, sports, study time, etc.
- So usually, this factor wouldn't produce an effect to performance

In [17]:
```
pandas.read_sql_query('SELECT count(*) FROM math WHERE _a_b_s_e_n_c_e_s_
> 10 GROUP BY _a_g_e_',db)
```

Out[17]:

count(*)
0 66
1 528
2 726
3 462
4 330
5 33
6 33

- This is a prime example of the age difference in individuals who want to be absent from school
- High school sophomores to seniors and first year college freshmans are the ones expected to miss classes the most
- High school freshman and college students from second year onwards take the approach to classes seriously

In [18]:
```
pandas.read_sql_query('SELECT count(*) FROM math WHERE _t_r_a_v_e_l_t_i_
m_e_ > 3 GROUP BY _s_e_x_',db)
```

Out[18]:

count(*)
0 66
1 198

In [19]:
```
pandas.read_sql_query('SELECT count(*) FROM math WHERE _t_r_a_v_e_l_t_i_
m_e_ > 3 GROUP BY _a_g_e_',db)
```

Out[19]:

count(*)
0 99
1 33
2 66
3 66

- From the above two queries, we have a situation where we count the males and females who travel at least one hour from home along with the age groups
- Three times as many females are one hour away from school, but the age group displays an important note
- It seems that age groups between 15-18 seem to be one hour of distance from school, so that says a lot about college because usually if college is far from your home, you either dorm or buy an apartment nearby

In [20]:
```
pandas.read_sql_query('SELECT count(*) FROM math WHERE _h_e_a_l_t_h_ > 2
GROUP BY _s_e_x_',db)
```

Out[20]:

count(*)
0 5016
1 4983

- We can also confirm that the health of a student doesn't affect the performance between genders and also the absences of these students
- So here is our final conclusion on this data analysis
- The factors that don't affect performance between males and females are health, going out with friends and with consumption of alcohol, free time, and travel time
- The factors that do affect performance between males and females are the number of absences and study time
- Despite seeing that the averages of these three period grades are close between the males and females, we can fairly conclude that absence of classes deteroriate your grades and consistent study time increases your grades over a period of time with less stress