Team members:
Rob Sisto, Annemarie Andaleon, Jacob Plaza, Melchor Ronquillo

COMP358 Project 2

Dataset description: "Student Alcohol Consumption"

Our dataset is from kaggle and uploaded by UCI Machine Learning. The data was obtained from a survey of secondary school students in Portugal. This survey primarily focuses on students and their alcohol consumption. Alcohol consumption is measured by 2 variables, WALC (Weekend Alcohol Consumption) and DALC (Workday Alcohol Consumption). Both are on a scale from 1 being very low to 5 being very high. To see what contributes to or what is affected by alcohol consumption, there are many other variables included within the dataset specific to each student. There is demographic information such as age, sex, and address, family information such as parents relationship, occupations, education, etc, and many other external information such as health condition, amount of time spent studying, traveling to school, going out, internet at home, and many others.

Link to Dataset:
https://www.kaggle.com/datasets/uciml/student-alcohol-consumption

Data Pipeline:
There are two ways to import the data into Hive in order to be able to query it. The first way is to input the table into Hive using SQL manually. The second way is to go through the Hive interface and import the csv.

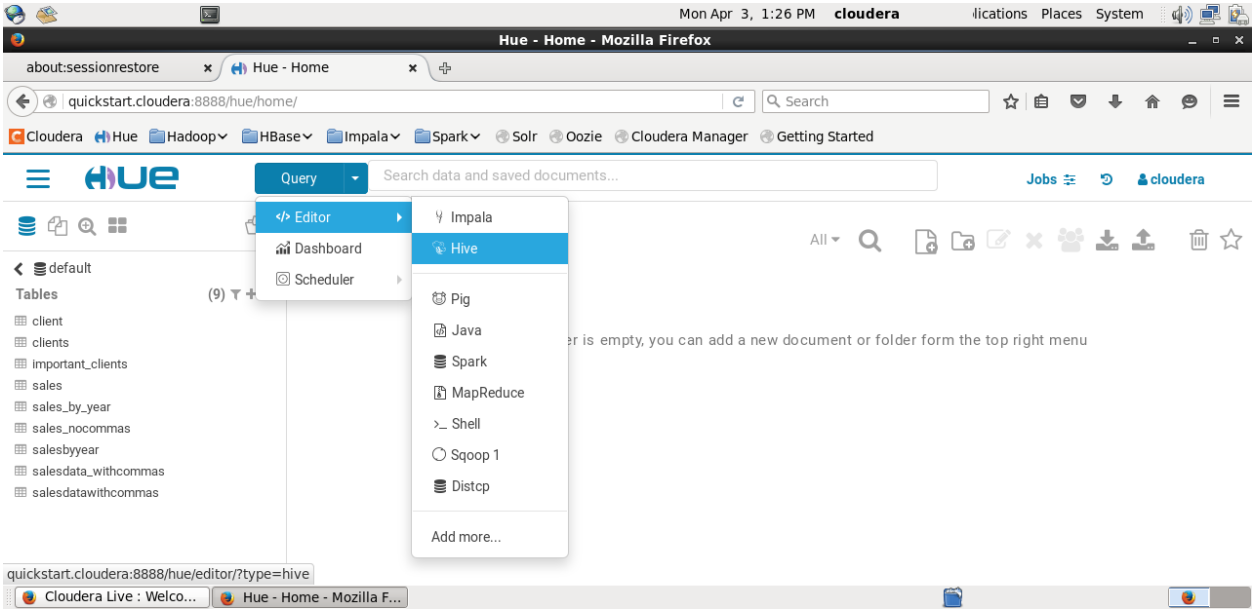
First Method:

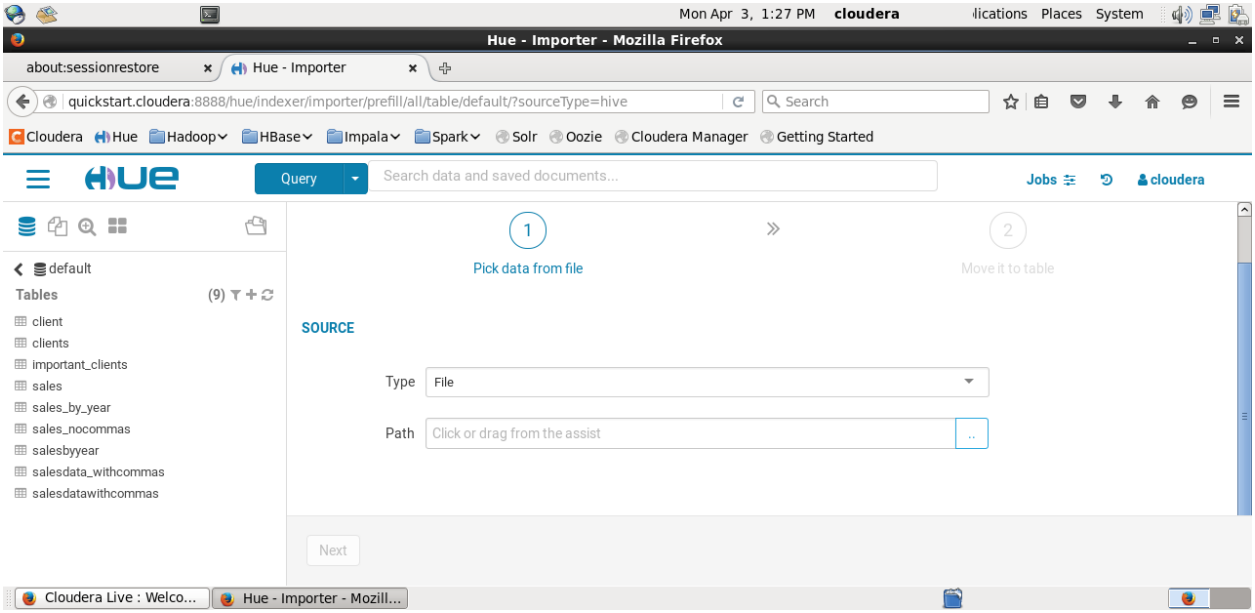
- Login to Hue from virtual machine or docker
- Upload jar through the Hue interface
    - Navigate to Manage HDFS / File Browser
    - Upload csv-serde.JAR file (from sakai) to user/cloudera (Main directory)
- Create table
    - Navigate to Query Editor → Hive
    - Make sure to go to the settings tab next to the terminal and add the csv-serde.jar as a File Resource
    - Paste the command below into the terminal:
    - create table Alcohol (RowID smallint,
        school string,
        sex string,
        age int,
        address string,
        famsize string,
        pstatus string,
        medu int,

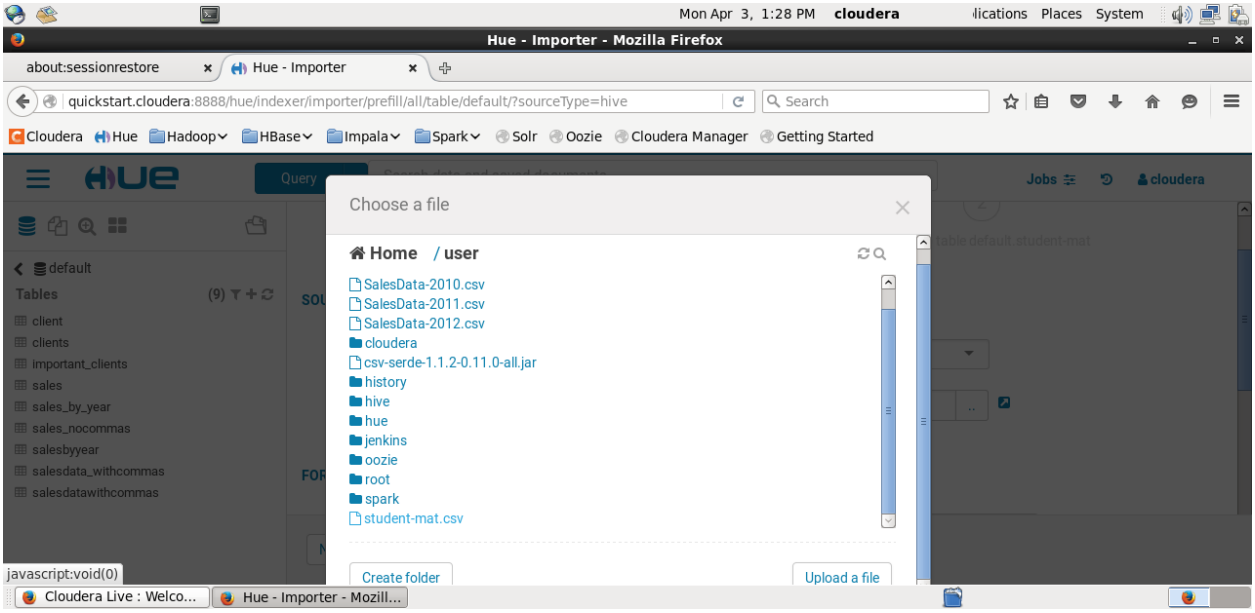
```
fedu int,
mjob string,
fjob string,
reason string,
guardian string,
traveltime int,
studytime int,
failures int,
schoolsup string,
famsup string,
paid string,
activities string,
nursery string,
higher string,
internet string,
romantic string,
famrel int,
freetime int,
goout int,
dalc int,
walc int,
health int,
absences int,
g1 int,
g2 int,
g3 int)

row format serde 'com.bizo.hive.serde.csv.CSVSerde'

stored as textfile;
```

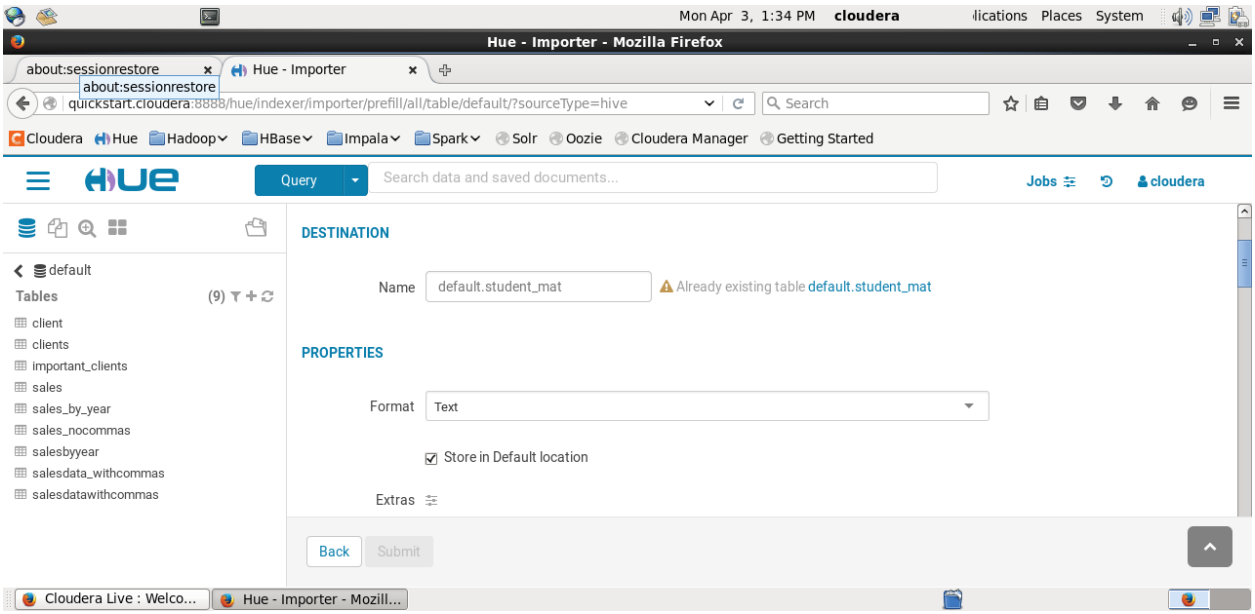
- Add csv data to table
  - Navigate to Manage HDFS / File Browser
  - Change to /user directory
    - Just click user at the top, should list more folders after
  - Navigate to HIVE → Warehouse → alcohol
  - Upload student_alcohol.csv to this folder
- Query

<u>Second Method:</u>


Mon Apr 3, 1:26 PM cloudera
Hue - Home - Mozilla Firefox
about:sessionrestore
Hue - Home
quickstart.cloudera:8888/hue/home/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query
Search data and saved documents...
Jobs
cloudera
Editor
Dashboard
Scheduler
Impala
Hive
Pig
Java
Spark
MapReduce
Shell
Sqoop 1
Distcp
Add more...
default
Tables (9)
client
clients
important_clients
sales
sales_by_year
sales_nocommas
salesbyyear
salesdata_withcommas
salesdatawithcommas
is empty, you can add a new document or folder form the top right menu
quickstart.cloudera:8888/hue/editor/?type=hive
Cloudera Live : Welco...
Hue - Home - Mozilla F...


Mon Apr 3, 1:27 PM cloudera
Hue - Importer - Mozilla Firefox
about:sessionrestore
Hue - Importer
quickstart.cloudera:8888/hue/indexer/importer/prefill/all/table/default/?sourceType=hive
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query
Search data and saved documents...
Jobs
cloudera
default
Tables (9)
client
clients
important_clients
sales
sales_by_year
sales_nocommas
salesbyyear
salesdata_withcommas
salesdatawithcommas
1
Pick data from file
2
Move it to table
SOURCE
Type
File
Path
Click or drag from the assist
..
Next
Cloudera Live : Welco...
Hue - Importer - Mozill...


Mon Apr 3, 1:28 PM cloudera
Hue - Importer - Mozilla Firefox
about:sessionrestore
Hue - Importer
quickstart.cloudera:8888/hue/indexer/importer/prefill/all/table/default/?sourceType=hive
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query
Jobs
cloudera
default
Tables (9)
client
clients
important_clients
sales
sales_by_year
sales_nocommas
salesbyyear
salesdata_withcommas
salesdatawithcommas
Choose a file
Home / user
SalesData-2010.csv
SalesData-2011.csv
SalesData-2012.csv
cloudera
csv-serde-1.1.2-0.11.0-all.jar
history
hive
hue
jenkins
oozie
root
spark
student-mat.csv
Create folder
Upload a file
javascript:void(0)
Cloudera Live : Welco...
Hue - Importer - Mozill...


Mon Apr 3, 1:34 PM cloudera
Hue - Importer - Mozilla Firefox
about:sessionrestore
Hue - Importer
quickstart.cloudera:8888/hue/indexer/importer/prefill/all/table/default/?sourceType=hive
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query
Search data and saved documents...
Jobs
cloudera
default
Tables (9)
client
clients
important_clients
sales
sales_by_year
sales_nocommas
salesbyyear
salesdata_withcommas
salesdatawithcommas
DESTINATION
Name
default.student_mat
Already existing table default.student_mat
PROPERTIES
Format
Text
Store in Default location
Extras
Back
Submit
Cloudera Live : Welco...
Hue - Importer - Mozill...

Queries:

Rob:

select cast(avg(walc + dalc) as decimal(10,2)) as avg_alc, fedu+medu as edu from alcohol group by fedu+medu;

- Average total alcohol consumption by total parent education level

	avg_alc	edu
1	NULL	NULL
2	2	1
3	4.03	2
4	3.79	3
5	3.93	4
6	3.49	5
7	3.72	6
8	3.55	7
9	3.93	8

- **select cast(avg(walc + dalc) as decimal(10,2)) as avg_alc, school, address, count(school) as num from alcohol group by school, address order by avg_alc;**
  - Average total alcohol consumption by the students school and address (either urban or rural)

	avg_alc	school	address	num
1	NULL	school	address	1
2	3.57	MS	U	21
3	3.66	GP	U	286
4	3.89	GP	R	63
5	4.88	MS	R	25

Annemarie:

- The average family relationships of families of certain parental statuses and family sizes
- SELECT cast(avg(famrel) as decimal(10,2)) as AvgFamilyRelationship, famsize, pstatus
  FROM student_mat
  GROUP BY pstatus, famsize
  ORDER BY AvgFamilyRelationship

```
SELECT cast(avg(famrel) as decimal(10,2))as AvgFamilyRelationship, count(*), famsize, pstatus FROM student_mat
GROUP BY pstatus, famsize
ORDER BY AvgFamilyRelationship
```

Query History | Saved Queries | Results (4)

	avgfamilyrelationship	count(*)	famsize	pstatus
1	3.66	21	GT3	A
2	3.87	94	LE3	T
3	3.98	260	GT3	T
4	4.09	20	LE3	A

- The failure rate of students based on their health
- SELECT health, count(*) AS `Count`, sum(failures) as TotalFailures, cast(avg(failures) as DECIMAL(10,2)) FROM student_mat
  GROUP BY health
  ORDER BY health

```
SELECT health, count(*) as `Count`, sum(failures) as TotalFailures, cast(avg(failures) as DECIMAL(10,2)) as AvgFailures FROM student_mat
GROUP BY health
ORDER BY health
```

Query History | Saved Queries | Results (5)

	health	count	totalfailures	avgfailures
1	1	47	8	0.17
2	2	45	12	0.26
3	3	91	37	0.40
4	4	66	22	0.33
5	5	146	53	0.36

Melchor:

- Find a student's average grade across 3 periods, their age, as well as their weekend and weekday alcohol consumption level given their alcohol level of 4 or higher
  - Query:
  - SELECT cast((a.g1 + a.g2 + a.g3) / 3 as decimal(10,1)) as Average_Grade, a.walc, a.dalc, a.age FROM alcohol2 a WHERE a.walc >= 4 AND a.dalc >= 4 ORDER BY a.age;

```
1 select cast(((a.g1 + a.g2 + a.g3)/3) as decimal(10,1)) as Average_Grade , a.walc, a.dalc, a.age
2 from alcohol2 a
3 where a.walc >= 4 and a.dalc >= 4;
4 order by a.age;
5
```

Next | Restart | Save as... | Format | or create a | New query

Recent queries | Query | Log | Columns | Results | Chart

	average_grade	a.walc	a.dalc	a.age
2	12	4	4	15
4	12.7	5	5	15
1	11	5	5	16
3	9.7	5	5	16
5	6.3	5	5	16
6	11.3	4	4	16
7	9.3	4	4	17
8	7.7	5	4	17
9	12.3	5	4	17
12	13.3	5	5	17
14	10.3	5	5	17
10	12.7	5	5	18
11	9	5	4	18
15	12.3	5	5	18
16	9	5	4	20
13	7.3	5	5	22

- Find the average alcohol level of students (weekend + weekday average) for every possible parent occupation combination (mother job and father job), as well as the amount of students with these parent occupation combination, order by alcohol level
    - Query:
    - SELECT cast(avg(a.walc + a.dalc) as decimal(10,1)) as Average_Alcohol, a.fjob, a.mjob, count(1) as Number_of_kids FROM alcohol2 a GROUP BY a.fjob, a.mjob ORDER BY Average_Alcohol

```
1 select cast(avg(a.walc + a.dalc) as decimal(10,1)) as average_alcohol, a.fjob, a.mjob, count(1) as Number_of_kids
2 from alcohol2 a
3 group by a.fjob, a.mjob
4 order by average_alcohol;
5
```

Execute | Save | Save as... | Explain | Format | or create a | New query

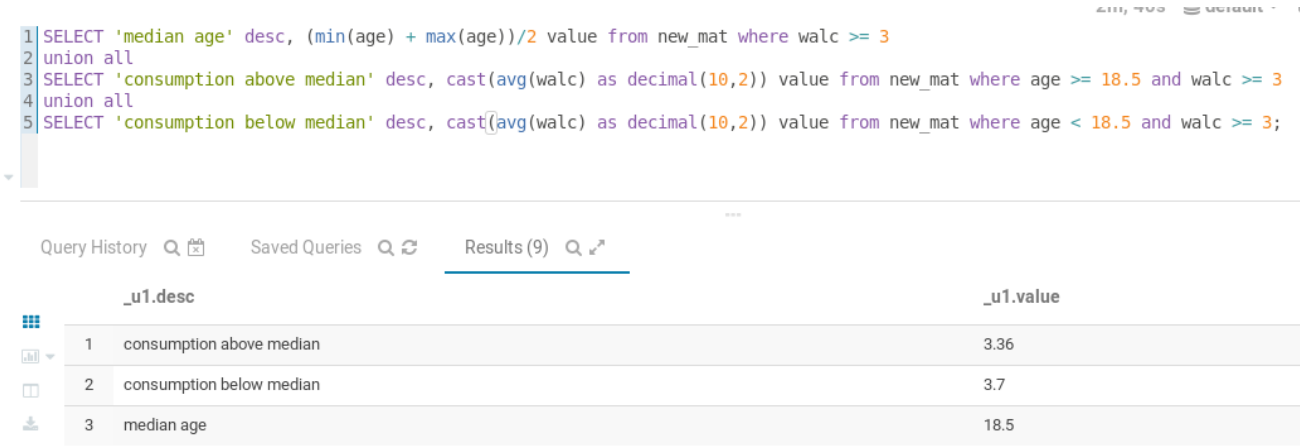
Recent queries | Query | Log | Columns | Results | Chart

	average_alcohol	a.fjob	a.mjob	number_of_kids
2	2	teacher	at_home	2
3	2	health	at_home	2
4	2.2	at_home	services	6
5	2.5	health	other	2
6	2.8	teacher	services	8
7	3	teacher	other	6
8	3	at_home	teacher	2
9	3	health	health	6
10	3	health	services	4
11	3.2	at_home	other	5
12	3.4	services	health	10
13	3.5	services	teacher	19
14	3.6	other	other	104
15	3.6	at_home	at_home	7
16	3.8	other	services	42
17	3.9	other	health	17
18	3.9	services	at_home	15
19	3.9	other	at_home	33
20	4	teacher	teacher	12
21	4.1	other	teacher	21
22	4.3	services	services	43
23	4.6	services	other	24
24	4.8	health	teacher	4
25	5	teacher	health	1

Jacob:

Query 1: Select the median age of heavy drinkers, and average alcohol consumption above and below the median

```
SELECT 'median age' desc, (min(age) + max(age))/2 value from new_mat

where walc >= 3

union all

SELECT 'consumption above median' desc, cast(avg(walc) as decimal(10,2)) value from new_mat where age >= 18.5 and
walc >= 3

union all

SELECT 'consumption below median' desc, cast(avg(walc) as decimal(10,2)) value from new_mat where age < 18.5 and
walc >= 3;
```



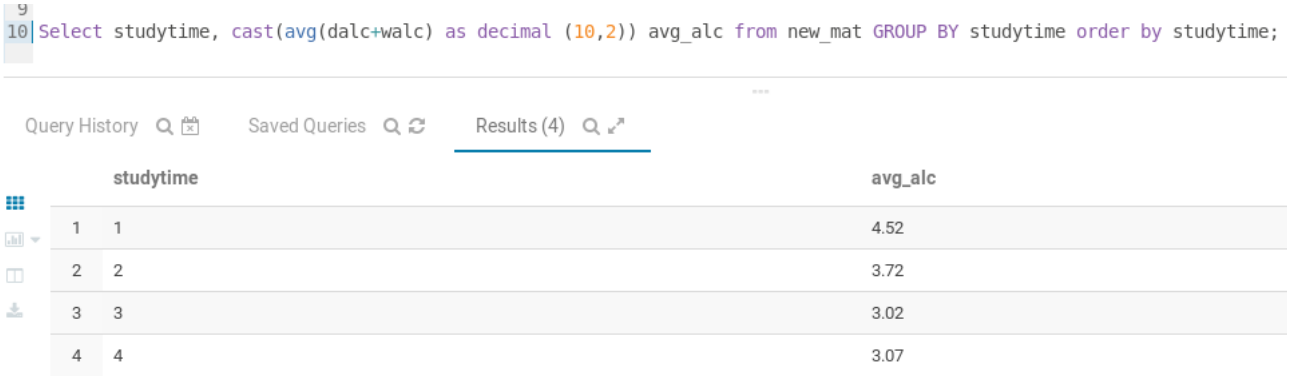
```
1 SELECT 'median age' desc, (min(age) + max(age))/2 value from new_mat where walc >= 3
2 union all
3 SELECT 'consumption above median' desc, cast(avg(walc) as decimal(10,2)) value from new_mat where age >= 18.5 and walc >= 3
4 union all
5 SELECT 'consumption below median' desc, cast(avg(walc) as decimal(10,2)) value from new_mat where age < 18.5 and walc >= 3;
```

Query History | Saved Queries | Results (9)

	_u1.desc	_u1.value
1	consumption above median	3.36
2	consumption below median	3.7
3	median age	18.5

Query 2: Average total alcohol consumption by studytime group

```
Select studytime, cast(avg(dalc+walc) as decimal (10,2)) avg_alc from new_mat group by studytime order by
studytime;
```



```
10 Select studytime, cast(avg(dalc+walc) as decimal (10,2)) avg_alc from new_mat GROUP BY studytime order by studytime;
```

Query History | Saved Queries | Results (4)

	studytime	avg_alc
1	1	4.52
2	2	3.72
3	3	3.02
4	4	3.07