DATA 210

Shefali Emmanuel

DATA 210 – Dataset Organization and Management
Spring 2020 - Syllabus

Instructor: Kebin Xu ***E-mail:*** xuk@cofc.edu ***Office:*** HWEA 306
Class Webpage: http://xuk.people.cofc.edu/DATA210/www/data210.htm
Office Hours: T 9:30 am – 12:30 pm, F 2:30 pm – 3:30 pm, other times by appointment

Required Reading:

Our main resource will be the *notes* in class. *Data Science at the Command Line: Facing the Future with Time-Tested Tools* by Jeroen Janssens which we will use sometimes can be accessed through the *Link*

Recommended Reading:

Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement by Eric Redmond and Jim Wilson. *Link*

Class Meeting Times:

S1: MWF 11:30 am– 12:20 pm; S2: MWF 12:30 pm – 1:20 pm; S3: MWF 1:30 pm – 2:20 pm
Class Meeting Location: HWEA 300

Course Description

A course to introduce the structure of databases and the management of datasets for information extraction. Concepts include the relational and entity relationship models, and local and distributed storage and access. The preparation and management of datasets for analysis is covered, and includes data cleaning, reorganization and security.

Pre-requisites: None

Learning Outcomes:

- To understand the difference between the different types of data (structured, semi- structured, unstructured, and metadata) and how this impact underlying storage mechanisms.
- To understand the key differences between data management systems, such as relational database systems, key-value, document-oriented, big data driven alternatives (e.g., NoSQL), etc.
- Be able to select and implement the optimal data management system and associated workflow as a function of the data's volume, velocity, variety, and veracity.
- To understand basic cloud computing concepts, to include infrastructure as a service, platform as a service, software as a service, and managed service.
- To understand the fundamental organization and use of semi-structured data (e.g., XML, JSON)
- To construct and perform queries to select data for assembly into an analysis workflow for data science.
- To understand the need for and to use an available tool to carry out data cleaning and other data pre-processing activities in preparing data resources integrated from various sources.
- To understand common patterns of data flow through data science analysis workflows, such as big data processing patterns using a parallel processing paradigm such as map-reduce.

Course Policies:

- **Attendance**:
  Attendance is mandatory. Within the first 10 minutes of each class attendance will be taken (which will be used to calculate the attendance portion of your final grade). If you walk into class after attendance has been taken or leave before the class is finished (for any reason other than an emergency), both will be recorded as an unexcused absence – no exceptions. If you have more than 6 excused or unexcused absences (2 weeks' worth of class), regardless of the reason, a WA grade will be given – no exceptions. If you do miss class,

you're responsible for announcements made in class, assignment due dates, etc. If you miss class, you must get an absence memo from the Absence Memo Office (http://studentaffairs.cofc.edu/about/services/absence.php).

- **Disability Accommodation:**
  The College will make reasonable accommodations for persons with documented disabilities. Students should apply at the Center for Disability Services / SNAP http://disabilityservices.cofc.edu/ , located on the first floor of the Lightsey Center, Suite 104. Students approved for accommodations are responsibility for notifying me, during my office hours, as soon as possible and for contacting me one week before accommodation is needed.

- **Homework**:
  Each assignment is due by the date and time that will be stated on the assignment. Assignments will be accepted only via OAKS. No assignments will be accepted late. Do NOT submit assignments to me for grading via email. Students are expected to abide by the **Honor System** of the College of Charleston and the **Student Code of Conduct** (http://deanofstudents.cofc.edu/honor-system/studenthandbook/student-handbook-2019-2020.pdf), especially sections on Cheating, Plagiarism (pp. 40)).

- **Hands-on Exercise**:
  During class meeting, time will be devoted to hands-on demonstrations and exercises. Each student will be responsible for completing the hands-on exercises before the end of the class meeting. Typically, a hands-on exercise will involve completing a partially developed program, entering it into the computer, and executing it.
- **Quizzes**:
  In-class quizzes (total of 20%) will be given to cover the material from the previous week. Quizzes are usually held at the beginning of class (first 10 to 15 minutes) once of every 1 or 2 weeks.
- **Electronics Devices**:
  You are highly recommended to bring your laptop to the class for hands-on exercises. Be respectful about unnecessary distractions to you and to others seated around you.

Grade Calculation:

- **Grading Policy**
  Exams – 65% (Tests 1, 2 and 3, 15% each; Quizzes: 20%), Homework – 10%, Final Exams – 25%
- **Scale:**
  A: 93—100; A-: 90—92; B+: 87—89; B: 83—86; B-: 80—82; C+:77—79; C: 73—76; C-: 70—72; D: 60—69; F: below 60

Important Dates:

Monday, January 20th: Martin Luther King, Jr. Holiday, observed. No classes. College closed.
Wednesday, February 12th: Test 1(tentative)
Wednesday, March 4th: Test 2 (tentative)
Friday, March 13: Last day to drop with grade of "W"
Sunday, Mar 15 — Saturday, March 21: Spring Break
Wednesday, April 8: Test 3 (tentative)
Wednesday, April 22: Last day of classes

Final Exam Date/Time:

Section 01 (MWF11:30 am– 12:20 pm): Wednesday, May 1: From 08:00 to 11:00am
Section 02 (MWF 12:30 pm – 1:20 pm): Monday, April 29: From 8:00 to 11:00am
Section 03 (MWF 1:30 pm – 2:20 pm): Saturday, April 27: From 8:00 to 11:00am

which — location of app

cat — display file content

head -n 1 f1.txt

wc lines words byte
-l -w -c
-L

touch — create empty file

cp myfile1.txt myfile2.txt — copy contents of 1 into 2

mv file1.txt out.txt — rename

mv file1.txt ../.. — moves file to parent directory

rm myfile.txt — remove file

rm out.txt deleteThis.txt

rm myfile*.txt] remove myfile1 myfile2 myfile3

rm *.txt] remove with same extension

rm folder (empty)

rm -r folder

seq 5 3 20
increment by 3

seq 10 > allNum.txt

num/count = wc -l

echo -n 'text'
no new line

echo 'word' >> out.txt
append

echo -e 'hello\nthere\t'

grep 'word' txtfile
-i ignore case
-w whole word
-wi

Quiz 1

which python = location of app

cat = display file content

head -n 2 f1.txt

touch = empty file

cp file1.txt file2.txt] copy contents of 1 to 2

mv file1.txt file.txt] RENAME

mv .. /..] to parent directory

rm file1.txt] remove

rm out.txt trash.txt] bulk delete

rm myfile*.txt] all myfile_.txt

rm *.txt] all txt

rmdir folder] empty

rmdir -r folder] non empty

seq 5 3 20
increment by

seq 10 > allNum.txt
rewrite

echo -n "text"
no new line

echo "word" >> out.txt
append

echo -e 'hello\nthere\t'
escape

grep -i -w 'word' txtfile
ignore case
word

count
Num

WC Lines Words Bytes
-l -w -c

-L Longest Line

wc -L movies.txt

mv
rm
echo

Name: Shefali E

DATA 210: Quiz 01
Session # 1

-7.5 92.5%

Write the commands that will:

each q is 4pt

1) Print the current directory

pwd

2) List the files and folders in current directory

ls

3) Create a folder with the name 'my_folder' in current directory

mkdir my-folder

4) Change current directory so that you're in 'my_folder'

cd my-folder

5) Write the command that appends the word 'Work' to the file 'movies.txt'

echo 'work' >> movies.txt

6) Create a file (even.txt) that has only one line of text: 'Even numbers from 0 to 100'

touch even.txt
echo 'Even numbers from 0 to 100' > even.txt
~~cat 'even numbers from 0 to 100' even.txt~~

7) Append to the file (even.txt) the even numbers from 0 to 100

seq 0 2 100 >> even.txt

8) Create a second file (odd.txt) that has one line of text: 'Odd numbers from 100 to 200)

~~cat 'odd numbers from 100 to 200' odd.txt~~
touch odd.txt
echo 'odd numbers from 100 to 200' > odd.txt

9) Append to the file (odd.txt) the odd numbers from 100 to 200

seq 101 2 200 >> odd.txt

10) Create a third file ('all.txt') that contains the content of the first file plus the content of the second file

cat even.txt odd.txt > all.txt

11) Delete the three files 'even.txt', 'odd.txt', and 'all.txt' (you can use m ultiple commands)

rm even.txt odd.txt all.txt

12) Which command will print the location/path of an application (or program) if it's installed on your machine?

which

13) Display the number of multiples of threes from 1 to 100

-1.5 seq 1 3 100 | wc -l

14) Display the first 20 lines of 'all_numbered.txt'

head -n 20 all_numbered.txt

15) Display the last 20 lines of 'all_numbered.txt'

tail -n 20 all_numbered.txt

16) Write the command that will delete the folder: Work (assuming the folder IS empty)

-1 rmdir Work or rm -d work

17) Assume that in the current directory, we have the following folders:
'Gym' 'Knowledge' 'Charleston' 'Work'
How do you create a new folder inside 'Work' with the name 'Origami'?

Mkdir work/origami

18) List three of the five tasks of a data scientist (must be in order)

-1 because order

① collect data
Sort data
③ analysis / ④ interpret data
② clean data

obtain / collect
clean / scrub
Explore
Model
understand

19) How do you display the number of characters in the longest line in some file (say 'songs.txt)

wc -L songs.txt

20) Write the command that will rename the file 'todo_list.txt' to 'songs.txt'

mv todo_list.txt songs.txt

21) Write the command that will delete the folder: School (assuming the folder is NOT empty)

rm -r School

22) Write the command that displays the number/count of numbers that satisfy the following conditions:

1) Between 6 and 400 2) Multiples of 2 3) Contain the number 3

seq 6 2 400 | grep '3' | wc -l

23) From the first 5 lines in the file: hobbies.txt, print the lines that contain the word: Gym (as a whole word)

head -n 5 hobbies.txt | grep -w 'Gym'

24) Given that the content of the file 'file03.txt' is the following:

Hilo is the rainiest city in the US
Charleston
Charlotte

What's the second line that will be printed if we run this command: sort file03.txt (this command will not modify the file)

Charlotte

Charleston
Charlotte
Hilo

25) What's the output of the following command (use file content from previous question):

```
head -n 1 file03.txt | wc -w
```

lines -l words -w bytes -c

-4

1

not after the sorted but before

man (ual)

Quiz 2

-help -h

unpack

gzip

cp -a or -r L04 L05

find -name *.db

find L04 -type d] all of the files inside

in2csv imdb.xlsx > imdb.csv

in2csv imdb.xlsx -- sheet 2 >> imdb.csv

csvcut -d, -C 'j' imdb.csv > movies.csv

remove column named j

csvcut -c 'year'

Cut -c5-12 file.txt] Char from 5 to 12 including both

csvlook

ls -l long

ls -lS sort by size descending

ls -lSr reverse

Control + D save file

MYSQL > SHOW DATABASES;
CREATE DATABASE data;
USE data,
SHOW TABLES;
CREATE TABLE clown(id INT);
Structure [DESC clown;
DROP TABLE IF EXISTS clown;
INSERT INTO clown(id) VALUES (21);
SELECT * FROM clown;
RENAME TABLE clown TO goons;
UPDATE clown SET name='Andan' WHERE id=21;

ALTER TABLE customer ADD id VARCHAR(22);
DROP COLUMN age;
MODIFY age varchar(22);

DELETE FROM customer WHERE rums=4;

Primary Key : can not hv 2 rows w/ same values || be null

U S W
AI A
DC
M
DF W

U S W
AI A
DC
M
DF W

U S W
AI A
DC
M
DF W

Quiz 2

man(ual)

if not → help -h --help

unpack file.txt
compress [gzip file.txt

cp -a or -r L04 L05

find -name *.db

find L04 -type d] all of the folders in L04

in2csv imdb-2.xlsx > imdb.csv

in2csv imdb-2.xlsx --sheet 2010-09-26 > imdb.csv

csvcut -c 'Year'

* csvcut -d, -C 'j' imdb.csv > movies.csv
if piped

removes the last column named j

cut -c5-10 file.txt] chs from 5 to 10 including both

head -n 5 movies.csv | csvcut -c 'Title,title,trim,Year' | csvlook

ls -l long
ls -lS sort by size descending order
ls -lSr reverse order

control+D to save the file

```
MySql> SHOW DATABASES;
       CREATE DATABASE data210;
       USE data210;
       SHOW TABLES;
       CREATE TABLE customer (id INT);
       DESC customer;
       DROP TABLE IF EXISTS customer;
       INSERT INTO customer(id)VALUES(22);
       SELECT * FROM customer;
       RENAME TABLE customer TO not,
```

show structure

ALTER TABLE customer ADD address VARCHAR(22) DEFAULT '20'
DROP COLUMN id
MODIFY age VARCHAR(2)

Name: Shefali Emmanuel DATA 210: Quiz 02 91 % Page 1 of 3

UPDATE SET WHERE Session # 1

-9 points

1. Write the command that will display the columns from 3 to 9 for the Excel file file05.xlsx (do not use 'csvlook' to format output). Remember that you will need to convert the Excel file to csv first.

in2csv file05.xlsx | csvcut -c3-9

2. Which command will list the files and folders in a long format and a descending order (with respect to size)

ls -lS

3. Write the command that will decompress the file file03.txt.gz

unpack file03.txt.gz

4. Write the command that will compress the file file02.txt

gzip file02.txt

5. Write the command that will display the column(s) Year, Rating, Release Date, Title for the Excel file file03.xlsx (use csvlook to format output). Remember that you will need to convert the Excel file to csv first.

in2csv file03.xlsx | | csvlook

└ csvcut -c 'Year, Rating, Release Date, Title

6. What's the command that will display the characters from 4 to 10 for the file file05.txt

cut -c4-10 file05.txt

7. Which command will list the files and folders in a long format and an ascending order (with respect to size)

ls -lSr

8. Write the command that will display the columns from 2 to 9 for the Excel file file03.xlsx (use csvlook to format output). Remember that you will need to convert the Excel file to csv first.

in2csv file03.xlsx | csvcut -c2-9 | csvlook

9. Write the command that will display the column(s) Year, Release Date for the Excel file file06.xlsx (do not use csvlook to format output). Remember that you will need to convert the Excel file to csv first.

in2csv file06.xlsx | csvcut -c 'Year, ReleaseDate'

10. What's the command that will change the first name (f_name) for all the records in table employee to Jim that satisfy the following criterion: age > 23

UPDATE table SET f_name = 'Jim' WHERE age>23;

11. Write the command that will create a table with the name student with the following columns:
id VARCHAR(10) PRIMARY KEY
full_name VARCHAR(30)

CREATE TABLE student(id VARCHAR(10) PRIMARY KEY, full_name VARCHAR(30));

12. What's the command that will rename the table student to faculty

RENAME TABLE student TO faculty;

13. Write the command that will create a table with the name employee with the following columns:
id INT
full_name VARCHAR(30)

CREATE TABLE employee(id INT, full_name VARCHAR(30));

14. What's the command that will set the database hospital_db to be the default(current) database for subsequent statements?

USE hospital_db;

15. What's the command that will INSERT into the table employee the value 'Steve' to column 'f_name' (VARCHAR(30)) and the value 20 to column 'age' (INT)

INSERT INTO employee(f_name, age) VALUES('Steve', 20);

16. What's the command that will show the structure of the table staff?

DESC staff;

17. What's the command that will show the tables in your current database?

SHOW TABLES;

18. What's the command that will add a column (num_of_baths — int) to the table 'house'

-1 ~~INSERT INTO~~ ALTER TABLE house ✓ ADD ✓ COLUMN ✓ num_of_baths ✓ INT;

19. What's the command that will delete the column 'num_of_baths' from the table 'house'

-1 ALTER TABLE DROP num_of_baths ~~FROM~~ house;

20. What's the command that will update the column 'age' and set its data type to INT from table 'student'

-2 ~~UPDATE~~ ALTER TABLE MODIFY student(age INT);

21. What's the command that will delete all rows in the table 'my_table'

-1 ~~DROP~~ DELETE * FROM my-table;

22. What's the command that will delete the table 'my_table'

-1 DROP TABLE my-table;

23. What's the command that will delete all rows in the table 'my_table' that have attribute/column 'age' greater than 30

-2 ~~DROP~~ DELETE FROM my-table WHERE age > 30;

What is command line?

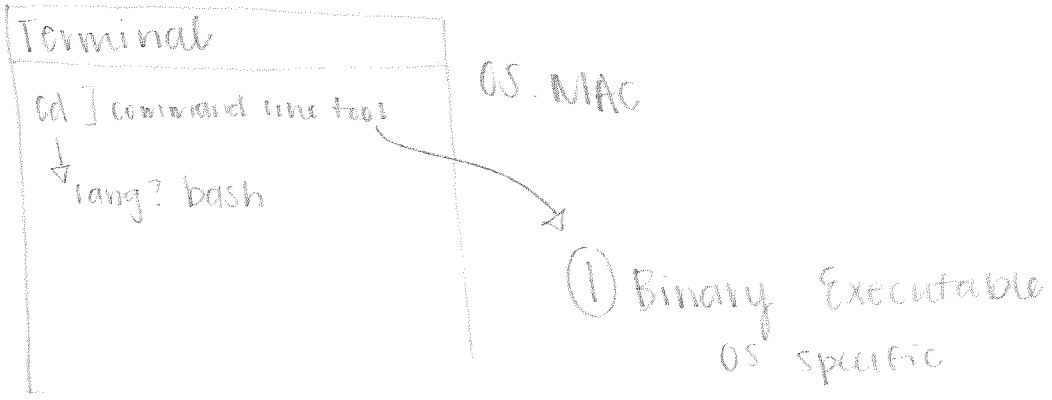
why?

Agile (close to data✓)
Augmenting (integration✓)
Scalable (automation✓)
Extensible (easy to create new tools)
Ubiquitous (it is everywhere✓)

5 Tasks of Data Science

Collect
Clean
Explore
Model
Understand

ENVIRONMENT



Absolute

from the root folder

VS

Relative

where is the file from HERE

DATABASE collection of data
- many tables
- collection of unprocessed items

TABLE Relational
- SQLang
- easy structured

Record

Field
- name
- size
- datatype

FILES
- RECORDS
  - Fields
    - Characters (1 byte)

Maintainance

Add Delete Modify

Validation (check w/ rules)

File Processing VS DB

- ↑ redundancy
- ↑ inconsistency

DBMS

Query Lang

Query by Example (gui assist)

data entry Form

Report Generator

Name: Shatan E DATA 210: Exam 01 92
Session # 1

1. Data science is OSEMN. List the five stages/steps (in order) that data science consists of; and provide an example for each stage/step.

Collect/Observe ex: gain or collect data
Clean/Select ex: clean it up / get rid of outliers
Evaluate ex: make sure everything is valid
Model ex: create Random Tree forest from data
Interpret/N ex: what did you learn or abstract from the model

-05

WRITE THE COMMAND THAT WILL

2. Print the current directory

```
pwd
```

3. Create a folder with the name 'my_folder' in current directory

```
mkdir my-folder
```

4. Change current directory so that you're in 'my_folder'

```
cd my-folder
```

5. List the files and folders in current directory

```
ls
```

6. Create a file (even.txt) that has only one line of text: 'Even numbers from 20 to 60'

```
echo 'Even numbers from 20 to 60' > even.txt
```

7. Append to the file (even.txt) the even numbers from 20 to 60

```
seq 20 2 60 >> even.txt
```

8. Create a second file (multiples.txt) that has one line of text: 'Multiple of fives from 200 to 400'

```
echo 'Multiples of fives from 200 to 400' > multiples
```

9. Append to the file (multiples.txt) with the multiple of fives from 200 to 400 .

seq 200 5 400 >> multiples.txt

10. Create a third file ('all.txt') that contains the content of the first file plus the content of the second file

cat even.txt multipler.txt > all.txt

* 11. Create a fourth file ('all_numbered.txt') contains the lines in 'all.txt' but numbered

-2 ~~wc~~ -l all.txt > all_numbered.txt all.txt > all_numbered.txt cat -n all.txt > all_nun.txt

12. Delete the three files 'even.txt', 'multiples.txt', and 'all.txt' (you can use multiple commands)

rm even.txt
rm multiples.txt
rm all.txt

13. Which command will print the location/path of an application (or program) if it's installed on your machine?

which

14. Display the header (the first line) of 'iris.csv'

head -n 1 iris.csv

15. Write the command that will delete the folder: University (assuming the folder is NOT empty)

rm -r University

-1 ~~rm~~ ir -r University

16. Write the command that will delete the folder: University (assuming the folder is empty)

rm -r University

rmdir University

17. Write the command that will rename the file 'todo_list.txt' to stories.txt'

mv todo_list.txt stories.txt

18. How do you display the number of characters in the longest line in some file (say 'stories.txt)

wc -L stories.txt

cat stories.txt | wc -L

19. From the first 3 lines in the file: todo_list.txt, print the number of lines that contain the word: 'Charleston' (could be part of another word but should be case insensitive)

head -n 3 todo_list.txt | grep -i 'Charleston' | wc -l

20. Write the command that lists the numbers between 30 and 500 that are multiples of 4 and that also contain the number 5

-0.5 seq 30 4 500 | grep '5'
32

21. Write the command that displays the number/count of numbers that satisfy the following conditions:

1) Between 2 and 200 2) Multiples of 3 3) Contain the number 5

-0.5 seq 2 3 200 | grep '5' | wc -l
3

22. From the first 5 lines in the file: songs.txt, print the lines that contain the word: Love (as a whole word)

head -n 5 songs.txt | grep -w 'Love'

23. Which command will list the files and folders in a long format and a descending order (with respect to size) descending

ls -lS

24. Write the command that will decompress the file file03.gz

unpack file03.gz

25. Write the command that will compress the file file02.txt

gzip file02.txt

26. Which command will list the files and folders in a long format and an ascending order (with respect to size)

ls -lSr

uni code A < a

27. Given that the content of the file 'file03.txt' is the following:

A cheerful heart is good medicine
but a crushed spirit dies up the bone
apple

letter takes precidence over white space

What's the second line that will be printed if we run this command: sort -r file03.txt (this command will not modify the file)

A cheerful heart is good medicine apple

28. What's the output of the following command (use file content from previous question):

head -n 2 file03.txt | wc -l

-l -w -c
line word byte

2

29. Write the command that will display the columns from 3 to 9 for the Excel file: file04.xlsx (do not use 'csvlook' to format output). Remember that you will need to convert the Excel file to csv first.

in2csv file04.xlsx | csvcut -c3-9

30. Write the command that will display the column(s) Year, Rating, Release Date, Title for the Excel file file04.xlsx (use csvlook to format output). Remember that you will need to convert the Excel file to csv first

in2csv file04.xlsx | csvcut -c 'Year, Rating, ReleaseDate, Title' | csvlook

31. What's the command that will display the characters from 4 to 10 for the file: file05.txt

cut -c4-10 file.txt

32. Write the command that will display the columns 2, 3, and 9 for the Excel file: file06.xlsx (use csvlook to format output). Remember that you will need to convert the Excel file to csv first.

in2csv file06.xlsx | csvcut -c 2,3,9 | csvlook

33. Write the command that will display the column(s) Year, Release Date for the Excel file file06.xlsx (do not use csvlook to format output). Remember that you will need to convert the Excel file to csv first.

in2csv file06.xlsx | csvcut -c 'Year, ReleaseDate'

SQL

34. Write the command that will create a table with the name employee with the following columns:
id VARCHAR(10) PRIMARY KEY
full_name VARCHAR(30) NOT NULL

CREATE TABLE employee (id VARCHAR(10) PRIMARYKEY, full_name VARCHAR(30) NOT NULL);

35. What's the command that will INSERT into the table student the value 'Steve' to column 'f_name' (VARCHAR(20)) and the value 20 to column 'age' (INT)

INSERT INTO student(f_name, age) VALUES ('steve', 20);

36. What's the command that will change the first name (f_name) for all the records in table employee to 'Bob' that satisfy the following criterion:
age < 25

UPDATE employee SET f_name='Bob' WHERE age<25;

37. What's the command that will rename the table faculty to student

RENAME TABLE faculty TO student;

38. What's the command that will show the structure of the table student?

DESC student;

39. Write the command that will create a table with the name patient with the following columns:
id INT PRIMARY KEY
full_name VARCHAR(30) NOT NULL
age SMALLINT DEFAULT 20

CREATE TABLE patient(id INT PRIMARY KEY,
full_name VARCHAR(30) NOT NULL
age SMALLINT DEFAULT 20);

40. What's the command that will set the database hospital_db to be the default(current) database for subsequent statements?

USE hospital_db;

41. What's the command that will show the tables in your current database?

SHOW TABLES;

42. What's the command that will add a column (num_of_baths — int) to the table 'house'

ALTER TABLE house ADD num_of_baths INT;

43. What's the command that will delete the column 'num_of_baths' from the table 'house'

DROP
ALTER TABLE house ~~DELETE~~ COLUMN num_of_baths;

44. What's the command that will add the primary key constraint to the table 'visitors' — column id (INT). The column 'id' already exists, and the table is empty.

ALTER TABLE visitors MODIFY id INT PRIMARY KEY;

45. Display the content of the table after the two following commands; given that 'id' is a primary key, and given that the table was originally empty):
INSERT INTO staff (id, first_name, last_name) VALUES(1, 'Adam', 'S');
INSERT INTO staff (id, first_name, last_name) VALUES(1, 'Jack', 'P');

1st Insert

id	first_name	last_name
1	Adam	S

can not add bc only 1 row can have primary key 1 rejects 2nd into statement

2nd Insert — ERROR

46. What would happen if you try to insert a record/instance/row without specifying a value for a NOT NULL column, given that the data type of that column is VARCHAR?

if it is Null there will be an error - the command will not be accepted

-1 ✓+

47. What's the command that will create a database named 'data210'

```
CREATE DATABASE data210;
```

48. What's the command that will delete all records/instances/rows in the table 'employee'

```
DELETE * FROM employee;
```

49. What's the command that will delete the table 'employee'?

```
DROP TABLE employee;
```

50. What does it mean for a column to be primary key?

① unique
② can not be null

-1

can not be duplicated

only one record can have the value

A primary key is a unique identifier ✓+

```
grep --color 'He\|She' movies.txt
     -w      whole word
     -o      each occurance on a seperate line
     -c      count # of lines
     -i      case insensitive

uniq -c      # of lines for similar consecutive lines

sort -n   file.txt
          numerically
     -nr  normal reverse
     -r

curl -s   www.google.com > out.txt        source code from web
          with out showing errors
  └── transfer & display data

tr '[:lower:]' '[:upper:]' < cns.txt > out.txt
```

- l link
- d directory
- \- file

Current: R W X (read, write, execute) | users in same group | outside group

Show content: echo $PATH | tr ':' '\n' — each folder on a seperate line

chmod u+r file.txt (u = current user, r = read)

```
grep -oE '\w+' Charleston.txt | sort | uniq -c | sort -n | head -n 12
```

(sort | uniq -c = frequency)

sort -n
normal

seq 8 11 | sort -r
reverse
10
11
8
9

sort

grep --color 'He\|She' movies.csv

grep -w 'She' movies.csv | sort
match the whole word ascending
-i print lines containing
-o every match on a separate line
-c count the # of lines

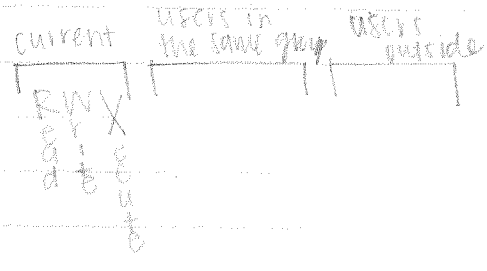
sort uniq -c # of occurances for similar consecutive lines

sort -nr file.txt (numerically)
reverse

curl -s retrieve source code from web | tail -n 7
sort

tr list source code as input sc → lowercase
replace Upper with
tr '[:lower:]' '[:upper:]' < characters.txt > out.txt

L link
d directory
- file



chmod u+w file01.txt
adding permission to current user

remove r * go inside cd but not ls
X not cd

grep -o 'awesell wine' sc.txt | sort | uniq -c

grep --color 'He\|She' movies.txt

- -w whole word
- -o each occurance on seperate lines
- -c count # of lines
- -i case insensitive

uniq
- -c # of lines consecutive

sort (uniq)
- -n numerically
- -nr normal reverse
- -r reverse

curl
- -s w/out showing errors

www.google.com > out.txt

└ transfer data

tr '[:lower:]' '[:upper:]' < chs.txt > out.txt

---

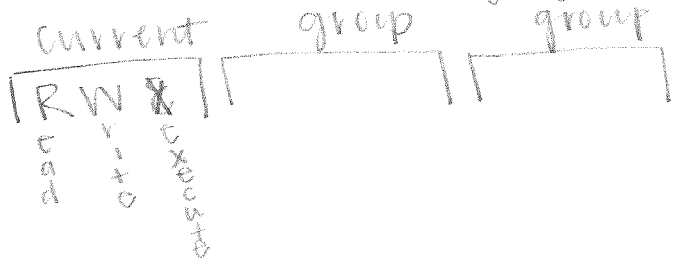
5 most frequent

grep -oE '\w+' top.txt | sort | uniq -c | sort -nr | head -n 5

l link

d directory

— file



chmod u+r, file.txt

frequency

sort | uniq -c

Name: Shefali E Session #1

DATA 210: Quiz 03

-1

99%

Q1) What's the command that will print the lines that contain the word "shop" or "cheese" (case sensitive) in the file sc.txt

```
grep 'shop\|cheese' sc.txt
```

```
grep -E 'shop | cheese'
sc.txt
```

Q2) What's the command that will change all lower-case letters to upper- case for the file hm.txt and saves the result in sc_out.txt

```
tr '[:lower:]' '[:upper:]' <hm.txt > sc_out.txt
cat hm.txt | tr '[:lower:]' '[:upper:]' > sc_out.txt
```

Q3) What's the output of the following command: seq 8 11 | sort -n numerically

```
8
9
10
11
```

Q4) Write the command that will transfer data (source-code) from the website: http://www.kdnuggets.com, and display the entire content of the source-code (while showing progress and error messages)

```
curl http://www.kdnuggets.com
```

Q5) Write the command that will transfer data (source-code) from the website: http://www.kdnuggets.com without showing progress, and saves the data into the file: top.txt

```
curl -s http://www.kdnuggets.com > top.txt
```

Q6) What's the command that will display the file sc.txt but "numerically" sorted

```
sort -n sc.txt
```

Q7) What's the command that will display the frequencies of the two words: 'cliff' and 'beach' (both case sensitive) in the file sc.txt

```
grep -o 'cliff\|beach' sc.txt | sort | uniq -c
grep -oE 'cliff|beach' sc.txt | sort | uniq -c
```

Q8) What's the command that will print the lines that contain the word beach or cliff (case insensitive) in the file hm.txt

```
grep -i 'beach\|cliff' hm.txt
grep -iE 'beach|cliff' hm.txt
```

Duration: 25 minutes

Q9) What's the command that will print each occurrence of the word shop or cheese (case sensitive) in the file hm.txt

grep -o 'shop\|cheese' hm.txt
-oE

Q10) What's the command that will replace each 'v' with 'b' for the file sc.txt and store the result in sc_out.txt

cat sc.txt | tr 'v' 'b' >sc_out.txt

tr 'v' 'b' < sc.txt > sc_out.txt

Q11) What's the output of the following command: seq 8 11 | sort -nr

numerical reverse

11
10
9
8

Q12) What's the command that will display the content of the file 'hm.txt' but sorted

sort hm.txt

Q13) What's the output of the following command:

seq 8 11 | sort -r

treat each line as a string

9
8
11
10

Q14) What's the command that will change all upper-case letters to lower- case for the file Charleston.txt and saves the result in hm_out.txt

tr '[:upper:]' '[:lower:]' <charleston.txt >hm_out.txt

Q15) What's the output of the following command:

seq 8 11 | sort

10
11
9
8

Q16) Write the command that will transfer data (source-code) from the website: http://www.kdnuggets.com without showing progress, and display only the last 20 lines from the source-code

curl -s http://www.kdnuggets.com | tail -n 20

Duration: 25 minutes

Q17) What's the command that will print the frequencies of each word in the file sc.txt

```
grep -oE '\w+' sc.txt | sort | uniq -c
```

Q18) What's the command that will print the 20 most frequent words (case sensitive) in the file hm.txt:

Part1) use grep & head commands

```
grep -oE '\w+' hm.txt | sort | uniq -c | sort -nr | head -n 20
```

Part 2) use grep & tail commands

```
grep -oE '\w+' hm.txt | sort | uniq -c | sort -n | tail -n 20
```

Q19) What's the command that will print the 6 least frequent words (case sensitive) in the file Charleston.txt

Part1) use grep & head commands

```
grep -oE '\w+' Charleston.txt | sort | uniq -c | sort -n | head -n 6
```

Part 2) use grep & tail commands

```
grep -oE '\w+' Charleston.txt | sort | uniq -c | sort -nr | tail -n 6
```

Q20) What's the command that will remove the execute permission for the file hm.txt, for the current user

```
chmod u-x hm.txt
```

Q21) What's the command that will add the read permission for the file Charleston.txt, for the current user

```
chmod u+r Charleston.txt
```

Q22) What's the command that will add the write permission for the file hm.txt, for the current user

```
chmod u+w hm.txt
```

Q23) What's the command that will display the content of (or list of folders/directories in) the $PATH variable (using 'echo')

```
echo $PATH
```

Duration: 25 minutes

Q24) What's the command that will display the content of (or list of folders/directories in) the $PATH variable; displaying each folder on a separate line

echo $PATH | tr ':' '\n'

tr ':' '\n'

Q25) (True/False)

Suppose file file01.txt and folder d01 has read, write and execute permission. file01.txt d01

F 1. After removing read permission of file01.txt, you can still open and edit it.

T 2. After removing read permission of file01.txt, you can still append other file to it.

T 3. After removing read permission of d01, you can still access (using cd) to it

T F 4. After removing execute permission of d01, you cannot access (using cd) to it.

after removing read ✓ go inside CD
X LS

removing execute X CD

removing -r ✓ CD
X LS

removing -X X CD

-r ✓ CD

-X X CD

-r ✓ CD

-X X CD

Duration: 25 minutes

awk → grabs data from txt files

awk 'NR<=3' my.txt // print the line
(# of current lines)

awk 'NR<=3 {print}' mydata.txt

awk 'NR<=3 {print $0}' mydaat.txt
(every line)

awk 'NR<=3 {print "Hello"}' mydata.txt → Hello Hello Hello

awk '{print length($1)}' mydata.txt

awk '2 {print}' hm.txt // EVERYLINE

awk '0 {print}' hm.txt //none

awk '{print $1}' hm.txt // 1st COL

awk -F' ' '{print $0}' hm.txt
(delimiter)

awk
NR % 2 == 0 // even
% 2 == 1 — odd
% 2

% 3 == 0 every 3rd line

awk '{print $2 $1}' hm.txt

between lines 10 & 12

awk 'NR == 10, NR == 12 {print $2 $1}' hm.txt
(: ; ")

(NR >= 10) & (NR <= 12)

awk '(NR >= 10) & (NR <= 12) {print $2 $1}' hm.txt

awk '{if (length($1) <= length($2)) print $0}'

awk '{if (length($1) <= length($2)) print $0}'

NR % 2 == 0 // even
NR % 2 == 1 // odd
NR % 2

% 3 == 0 print every 3rd line

Q09) What's the output of the following: awk 'NR%3 {print}' mydata.txt

Printing lines 1, 2, 4, 5, 7, 8, ...

```
cat -n mydata.txt|awk 'NR%3 {print}'
```

Q10) What's the output of the following: awk 'NR%3==0 {print}' data.txt

Printing lines 3, 6, 9, 12, 15 ...

```
cat -n mydata.txt|awk 'NR%3==0 {print}'
```

$$H(C=\text{Sail}) = -\tfrac{7}{10} * \log(\tfrac{7}{10}) - \tfrac{3}{10} * \log(\tfrac{3}{10})$$

$$= .88$$

$$H(\text{Sail} \mid O=\text{Sunny}) = -\tfrac{5}{5} * \log(\tfrac{5}{5}) - \tfrac{0}{5} * \log(\tfrac{0}{5})$$

$$= 0$$

$$H(\text{Sail} \mid O=\text{Rainy}) = -\tfrac{2}{5} * \log(\tfrac{2}{5}) - \tfrac{3}{5} * \log(\tfrac{3}{5})$$

$$= .97$$

$$H(\text{Sail} \mid \text{Outlook}) = \tfrac{5}{10} * 0 + \tfrac{5}{10} * .97$$

$$= .485$$

$$IG(C \mid \text{Outlook}) = H(C=\text{Sail}) - \mathcal{E}H(\text{Outlook})$$

$$= .88 - .485$$

$$= .395$$

$$H(\text{Sail} \mid C=\text{Big}) = -\tfrac{3}{3} * \log(\tfrac{3}{3}) - \tfrac{0}{3} * \log(\tfrac{0}{3})$$

$$= 0$$

$$H(\text{Sail} \mid C=\text{Med}) = -\tfrac{3}{4} * \log(\tfrac{3}{4}) - \tfrac{1}{4} * \log(\tfrac{1}{4})$$

$$= .811$$

$$H(\text{Sail} \mid C=\text{NO}) = -\tfrac{1}{3} * \log(\tfrac{1}{3}) - \tfrac{2}{3} * \log(\tfrac{2}{3})$$

$$= .92$$

$$H(\text{Sail} \mid \text{Company}) = \tfrac{3}{10} * 0 + \tfrac{4}{10} * .811 + \tfrac{3}{10} * .92$$

$$= .60$$

$$IG(C \mid \text{Company}) = H(C=\text{Sail}) - \mathcal{E}H(\text{Company})$$

$$= .88 - .60$$

$$= .28$$

awk → grabs data from txt files

awk 'NR<=3' mydata.txt // print the line
# of current lines

awk 'NR<=3 {print}' mydata.txt

awk 'NR<=3 {print $0}' mydata.txt
entire line

awk 'NR<=3 {print "Hello"}' mydata.txt → "Hello"
"Hello"
"Hello"

awk '{print length($0)}'

awk '2 {print}' hm.txt // everyline

awk '0 {print}' hm.txt // none

awk '{print $1}' hm.txt // print 1st column

awk -F, '{print $1}' hm.txt
delimiters

awk '{print $2 $1}' hm.txt
2nd 1st word

between lines 10 & 20
awk 'NR==10, NR==20 {print $2 $1}' hm.txt

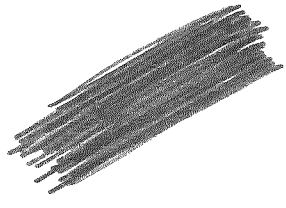
{print $2,$1} hm.txt // seperate by space
" "
":"

awk '(NR>=10) && (NR<=12) {print length($2)}' mydata.txt
len of word 2 between lines 10 & 12

awk '{if (length($1) == length($2) print $0}' hm.out

NR % 2 == 0 // even
NR % 2 == 1 // odd
NR % 2

NR % 3 == 0 print every 3rd line



French Italian

$H(x) = -(1/2) * \log_2(1/2) - (1/2) * \log(1/2) = 1$

Thai Burger

$H(y) = -(2/4) * \log(2/4) - (2/4) * \log(2/4) = 1$

$H(C \mid Type = French) = 1$
$H(C \mid Type = Italian) = 1$
$H(C \mid Type = Thai) = 1$
$H(C \mid Type = Burger) = 1$

$H(C \mid Type) = 1 - 1 = 0$

$H(Sail \mid Outlook = Sunny) = -(5/5) * \log(5/5) - (0/5) * \log(0/5)$
$= -1 * 0 - 0 * 0$
$= 0$

$H(Sail \mid Outlook = Rainy) = -(2/5) * \log(2/5) - (3/5) * \log(3/5)$
$= .97$

$H(Sail \mid Outlook) = 5/10 * 0 + 5/10 * .97$
$= .485$

$IG(C \mid Outlook) = H(C = Sail) - EH(Outlook)$
$= .88 - .485$
$= .395$

Name: Shefali Emmanuel DATA 210: Quiz 04 awk score 52/52 Page 1 of 7 d Tree 37.5/50

Session # 1

Problem A

- 1/2 * log 1/2

1. (Multiple choice) What's the output of the following: awk 'NR%2 {print}' hm_out.txt

a) Printing odd lines
b) Printing even lines
c) Printing nothing
d) Error

== 0

2. (Multiple choice) What's the output of the following: awk '2 {print}' hm_out.txt

a) Printing the first 2 lines
b) Printing every line
c) Printing nothing
d) Error

3. (Multiple choice) What's the output of the following: awk '0 {print}' hm_out.txt

a) Printing the first 2 lines
b) Printing every lines
c) Printing nothing
d) Error

4. (Multiple choice) What's the output of the following: awk '(NR%2==1) {print}' hm_out.txt

a) Printing odd lines
b) Printing even lines
c) Printing nothing
d) Error

print $5 ':' $8

5. For the file: file04.txt, write the command that will print the 5th and 8th words (separated by a colon only), for every line where the length of the first word is more than 4 characters (use the command 'awk', and don't use an if statement)

awk 'length($1)> 4 {print $5 ':' $8}' file04.txt

6. How do you print the lines starting at line 8 (using $0) from the file file01.txt using the command 'awk' and NR; and using greater than or equal to (>=). Use an if statement.

awk '{if (NR>=8) print $0}' file01.txt

7. For the file: hm_out.txt, write the command that will print the length of each line on a separate line (use the command awk)

awk '{print length($0)}' hm_out.txt

8. How do you print the first 9 lines (using $0) from the file file04.txt using the command 'awk' and NR, and using less than or equal to (<=). Use an if statement.

awk '{if (NR<=9) print $0}' file04.txt

9. How do you print the lines (using $0) starting at line 10 and ending at line 14 from the file file02.txt using the command 'awk' and NR; using greater than or equal to (>=) and less than or equal to (<=). Don't use an if statement.

awk '(NR>=10)(NR<=14) {print $0}' file02.txt

10. For the file: file03.txt, write the command that will print the length of each line on a separate line (use the command awk)

awk '{print length($0)}' file03.txt

11. For the file: file03.txt, write the command to print the 4th word if the 7th word has same length as the 8 (don't use an if statement)

awk 'length($7) == length($8){print $4}' file03.txt

12. How do you print the lines (using $0) starting at line 8 and ending at line 15 from the file file01.txt using the command 'awk' and NR; using greater than or equal to (>=) and less than or equal to (<=). Use an if statement.

awk '{if (NR>=8) && (NR<=15) print $0}' file01.txt

13. How do you print the lines starting at line 9 (using $0) from the file file03.txt using the command 'awk' and NR; and using greater than or equal to (>=). Don't use an if statement.

awk 'NR>=9 {print $0}' file03.txt

Problem B: (50pts) Build the decision tree to classify 'Sail' for the table below (show your work). -12.5 / 37.5

	Sailboat	Outlook	Sail
1	Big	Rainy	Yes
2	Big	Sunny	Yes
3	Big	Rainy	No
4	Big	Sunny	No
5	Small	Rainy	No
6	Small	Sunny	No
7	Small	Rainy	No
8	Small	Sunny	No

Knowing $log_2 \frac{1}{x} = -log_2 x$, $\log_2 3 = 1.58$, $\frac{3}{4} * \log_2 3 = 1.189$

$$H(\text{table}) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) = .811$$

2^1 2 2^2 2^1 2^0 4 2 0

1. (5pts) what is the entropy for this table. Entropy $H(C) = -p_1 * \log_2(p_1) - p_2 * \log_2(p_2)$

   where p_1 is the probability of Sail = Yes, and where p_2 is the probability of Sail = No

$P1 = H(\text{Sail} \mid C=\text{yes}) = -\frac{2}{2} * \log(\frac{2}{2}) - \frac{0}{2} * \log(\frac{0}{2}) = 1$ (* log 1, 0)

$P2 = H(\text{Sail} \mid C=\text{NO}) = -\frac{0}{6} * \log(\frac{0}{6}) - \frac{6}{6} * \log(\frac{6}{6}) = 1$

2. Split the table by one attribute at a time, and calculate the weighted/expected entropy EH for ~~class~~ Sail with respect to attribute A for the resulting subtables after the split.

$EH(\text{Sail}) = \frac{2}{8} * [\ 1\ ] + \frac{6}{8} * [\ 1\ ] = .811$

PartA.(15pts)Let's try to split by 'Sailboat' first;

Sailboat has two different values; 'Big' and 'Small'.

a) (5pts) Draw the subtable when 'Sailboat' is 'Big' and calculate the entropy $H(C|_{\text{Sailboat = Big}})$

Sailboat	Outlook	Sail
Big	Rainy	yes
Big	Sunny	Yes
Big	Rainy	No
Big	Sunny	NO

$H(C|\text{Sailboat = Big})$

$= -2/4 * \log(2/4) - 2/4 * \log(2/4)$

$= 1$

b) (5pts) Draw the subtable when 'Sailboat' is 'Small' and calculate the $H(C|_{\text{Sailboat = small}})$

Sailboat	Outlook	Sail
Small	Rainy	NO
Small	Sunny	NO
Small	Rainy	NO
Small	Sunny	NO

$H(C|\text{Sailboat = small})$

$= -0/4 * \log(0/4) - 4/4 * \log(4/4)$

$0 * \log_2 0 = 2^0 \quad 1 \quad 2^1 \quad 1 \quad 2$

$= 0$

c) (5pts) Calculate expected/weighted entropy EH

$EH(\text{Sailboat}) = 4/8 * \boxed{1} + 4/8 * \boxed{0} = .5$

PartB (15pts) Now let's try to split our table by 'Outlook' using the same way as partA, Draw the subtables and calculate $H(C|_{\text{Outlook= Rainy}})$, $H(C|_{\text{Outlook= Sunny}})$, and EH

a)(3pts) Draw the subtable when 'Outlook' is 'Rainy', and calculate $H(C|_{\text{Outlook= Rainy}})$

Sailboat	Outlook	Sail
Big	Rainy	yes
Big	Rainy	No
Small	Rainy	NO
Small	Rainy	NO

$H(C|\text{Outlook = Rainy})$

$= -1/4 * \log(1/4) - 3/4 * \log(3/4)$

$= .811$

b) (3pts) Draw the subtable when 'Outlook' is 'Sunny', and calculate H(C| Outlook= Rainy)

Sailboat	Outlook	Sail
Big	Sunny	Yes
Big	Sunny	No
Small	Sunny	No
Small	Sunny	No

$H(C \mid \text{Outlook} = \text{Rainy})$

$= -1/4 * \log(1/4) - 3/4 * \log(3/4)$

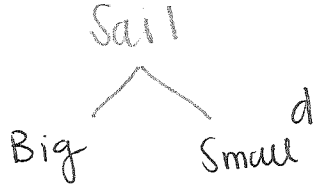
$= .811$

c)(4pts) Calculate the expected/weighted entropy EH

$EH = \underline{4/8} * \underline{.811} + \underline{4/8} * \underline{.811} = .811$

-2.5 ~~8~~ **(5pts) Based on the answers above, what is the first feature you should split on, why? Draw the first split.**

Sailboat as it splits and gives a different variant



draw full table

PartC (15pts)

If you answer of Question 3 is "Outlook", please do the following, if not, go to Q6

4. (5pts) What is Entropy of the right subtalbe (Outlook = Rainy)? Will you further split it?

5. (5pts) What is Entropy of the left subtalbe (Outlook = Sunny)? Will you further split it? If so, what is the next feature you will split on? Go to question8.

-2.5

6. (5pts) What is Entropy of the right subtalbe (Sailboat = Small)? Will you further split it?

NO

$EH = \frac{Sail}{.811} - (Sailboat = small)$

-2.5

7. (5pts) What is Entropy of the left subtalbe (Sailboat = Big)? Will you further split it? If so, what is the next feature you will split on?

yes

$EH = .811 - (Sailboat = Big)$

~~8.~~ (5pts) Draw the final tree

Let's try to split by 'Outlook', which is the only feature we're left with.

When Outlook is Rainy;

	Sailboat	Outlook	Sail
1	Big	Rainy	Yes
3	Big	Rainy	No

Entropy here is still 1

When Outlook is Sunny;

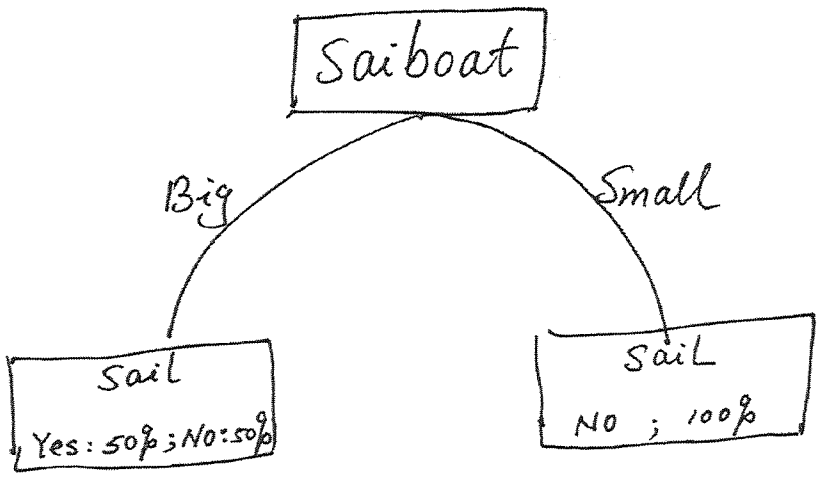
	Sailboat	Outlook	Sail
2	Big	Sunny	Yes
4	Big	Sunny	No

Entropy here is still 1

The weighted/expected entropy of feature Outlook EH(Outlook) is 1. Notice here that before this split, the entropy of the subtable was 1, and after the split the weighted/expected entropy is still 1. IG(Outlook) = H(C) – EH(Outlook) = 0 That means that asking the question about 'Outlook' won't help us make a better prediction, and for that reason, we should not split our table any further.

8. (5pts) Draw the final tree.

Our final tree is;



$$Entropy = -p_1 * \log_2(p_1) - p_2 * \log_2(p_2)$$

$$EH = \frac{sail}{yes} * entropy\left(\frac{sail}{yes}\right) + \frac{sail}{no} * entropy\left(\frac{sail}{no}\right)$$

$$IG = HC - Sail$$

vs

$$HC - Outlook$$

Entropy = $-p1 * \log_2(p1) - p2 * \log_2(p2)$

$= -2/8 * \log_2 2/8 - 6/8 * \log(6/8)$

$= .811$

$H(C \mid \text{Sailboat} = \text{Big}) = -2/4 * \log 2/4 - 2/4 * \log 2/4$

$= -1/2 * \log 1/2 - 1/2 * \log 1/2$

$= 1$

$H(C \mid \text{Sailboat} = \text{small}) = 0/4 * \log 0/4 - 4/4 * \log 4/4$

$= 0 * \log 0 - 1 * \log 1$

$= 0$

$EH = \underline{4/8} * \underline{1} + \underline{4/8} * \underline{0}$

$.5 + 0$

$= .5$ ↓ the better

$H(C \mid \text{outlook} = \text{Rainy}) = -1/4 * \log_2(1/4) - 3/4 * \log_2(3/4)$

$= .811$

$H(C \mid \text{outlook} = \text{Sunny}) = .811$

$EH = 4/8 * \underline{.811} + 4/8 * \underline{.811}$

$= .811$

To print the first "column" from each line, we can do the following:

```
[/data/L05]$ awk '{print $1}' hm.txt
```

By default, each column is separated by a space; for example,
Then executing this command: awk '{print $1}' hm.txt will output:

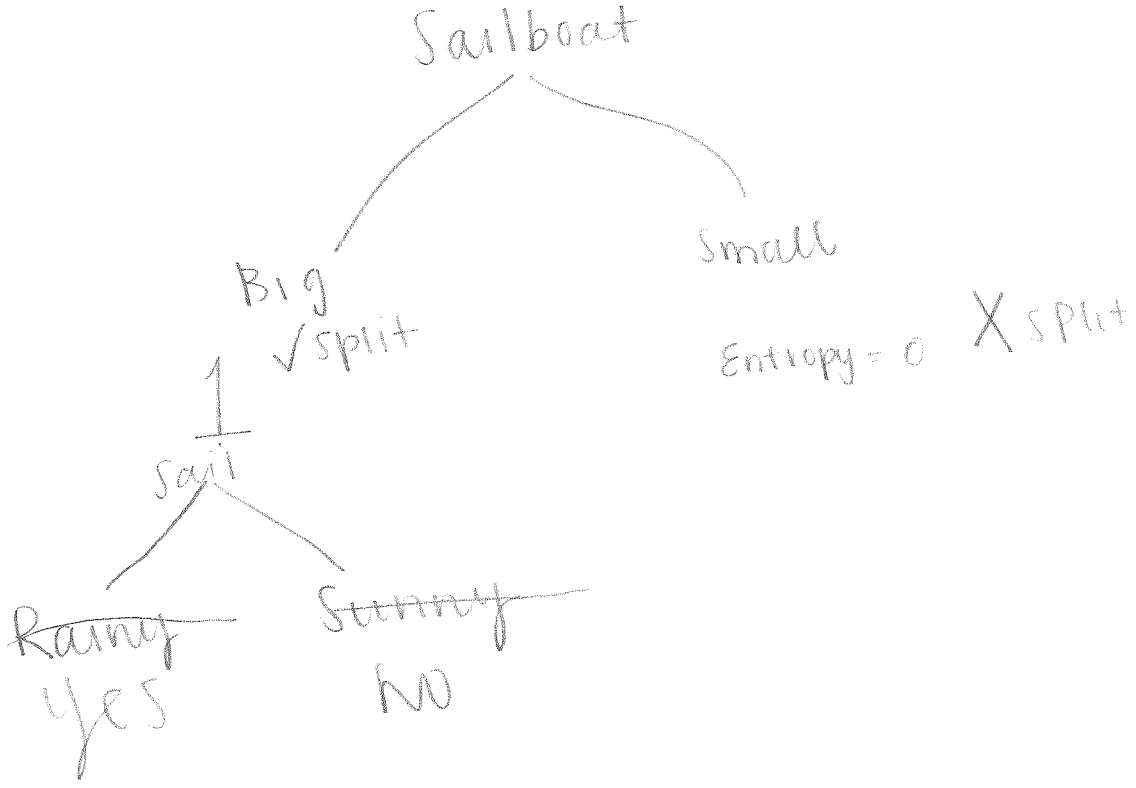
```
This
LetMeIgnoreSome
89Hello
bang
```

$2 corresponds to the 2nd word, so on and so forth

IG = .811 − .5 = .311 ✓

H(c) | Cutlass

.811 − .811 = 0



If we do the following:

```
[/data/L05]$ awk 'NR<=3 {print "Hello"}' mydata.txt
```

Then this would mean: for each line that satisfies the criterion NR<=3 (which is essentially the first three lines), print the word "Hello"

So, the output of that command would essentially be:

```
"Hello"
"Hello"
"Hello"
```

If we remove the condition, and just simply do the following:

```
[/data/L05]$ awk '{print "Hello"}' mydata.txt
```

Then the terminal will print the word "Hello" for every single line.

If we get rid of "Hello", then every single line will be printed:

```
[/data/L05]$ awk '{print}' mydata.txt
```

The command above will print every single line in "mydata.txt". So awk '{print $0}' mydata.txt does.

If we type a number before {print} then:

- The condition will evaluate to true if it's positive
- The condition will evaluate to false if it's 0
- Negative numbers won't be allowed.

For examples:

We have another file "hm.txt" contains the following:

```
This is the first line
LetMeIgnoreSome spaces here
89Hello There...
bang bang da
```

```
[/data/L05]$ awk '2 {print}' hm.txt // This will print every single line
[/data/L05]$ awk '0 {print}' hm.txt // This won't print anything
```

French Italian

$H(x) = -(1/2) * \log_2(1/2) - (1/2) * \log(1/2) = 1$

Thai Burger

$H(y) = -(2/4) * \log(2/4) - (2/4) * \log(2/4) = 1$

$H(C \mid Type = French) = 1$
$H(C \mid Type = Italian) = 1$
$H(C \mid Type = Thai) = 1$
$H(C \mid Type = Burger) = 1$

$H(C \mid Type) = 1 - 1 = 0$

$H(Sail \mid Outlook = Sunny) = -(5/5) * \log(5/5) - (0/5) * \log(0/5)$
$= -1 * 0 - 0 * 0$
$= 0$

$H(Sail \mid Outlook = Rainy) = -(2/5) * \log(2/5) - (3/5) * \log(3/5)$
$= .97$

$H(Sail \mid Outlook) = 5/10 * 0 + 5/10 * .97$
$= .485$

$IG(C \mid Outlook) = H(C = Sail) - EH(Outlook)$
$= .88 - .485$
$= .395$

Name: Shefali Emmanuel DATA 210: Quiz 04 awk score 52/52 Page 1 of 7 d Tree 37.5/50

Session # 1

Problem A

- 1/2 * log 1/2

1. (Multiple choice) What's the output of the following: awk 'NR%2 {print}' hm_out.txt

a) Printing odd lines
b) Printing even lines
c) Printing nothing
d) Error

== 0

2. (Multiple choice) What's the output of the following: awk '2 {print}' hm_out.txt

a) Printing the first 2 lines
b) Printing every line
c) Printing nothing
d) Error

3. (Multiple choice) What's the output of the following: awk '0 {print}' hm_out.txt

a) Printing the first 2 lines
b) Printing every lines
c) Printing nothing
d) Error

4. (Multiple choice) What's the output of the following: awk '(NR%2==1) {print}' hm_out.txt

a) Printing odd lines
b) Printing even lines
c) Printing nothing
d) Error

print $5 ':' $8

5. For the file: file04.txt, write the command that will print the 5th and 8th words (separated by a colon only), for every line where the length of the first word is more than 4 characters (use the command 'awk', and don't use an if statement)

awk 'length($1)> 4 {print $5 ':' $8}' file04.txt

6. How do you print the lines starting at line 8 (using $0) from the file file01.txt using the command 'awk' and NR; and using greater than or equal to (>=). Use an if statement.

awk '{if (NR>=8) print $0}' file01.txt

7. For the file: hm_out.txt, write the command that will print the length of each line on a separate line (use the command awk)

awk '{print length($0)}' hm_out.txt

8. How do you print the first 9 lines (using $0) from the file file04.txt using the command 'awk' and NR, and using less than or equal to (<=). Use an if statement.

awk '{if (NR<=9) print $0}' file04.txt

9. How do you print the lines (using $0) starting at line 10 and ending at line 14 from the file file02.txt using the command 'awk' and NR; using greater than or equal to (>=) and less than or equal to (<=). Don't use an if statement.

awk '(NR>=10)(NR<=14) {print $0}' file02.txt

10. For the file: file03.txt, write the command that will print the length of each line on a separate line (use the command awk)

awk '{print length($0)}' file03.txt

11. For the file: file03.txt, write the command to print the 4th word if the 7th word has same length as the 8 (don't use an if statement)

awk 'length($7) == length($8){print $4}' file03.txt

12. How do you print the lines (using $0) starting at line 8 and ending at line 15 from the file file01.txt using the command 'awk' and NR; using greater than or equal to (>=) and less than or equal to (<=). Use an if statement.

awk '{if (NR>=8) && (NR<=15) print $0}' file01.txt

13. How do you print the lines starting at line 9 (using $0) from the file file03.txt using the command 'awk' and NR; and using greater than or equal to (>=). Don't use an if statement.

awk 'NR>=9 {print $0}' file03.txt

Problem B: (50pts) Build the decision tree to classify 'Sail' for the table below (show your work). -12.5 / 37.5

	Sailboat	Outlook	Sail
1	Big	Rainy	Yes
2	Big	Sunny	Yes
3	Big	Rainy	No
4	Big	Sunny	No
5	Small	Rainy	No
6	Small	Sunny	No
7	Small	Rainy	No
8	Small	Sunny	No

Knowing $log_2 \frac{1}{x} = -log_2 x$, $\log_2 3 = 1.58$, $\frac{3}{4} * \log_2 3 = 1.189$

$$H(\text{table}) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) = .811$$

2^1 2 2^2 2^1 2^0 4 2 0

1. (5pts) what is the entropy for this table. Entropy $H(C) = -p_1 * \log_2(p_1) - p_2 * \log_2(p_2)$

   where p_1 is the probability of Sail = Yes, and where p_2 is the probability of Sail = No

$P1 = H(\text{Sail} \mid C=\text{yes}) = -\frac{2}{2} * \log(\frac{2}{2}) - \frac{0}{2} * \log(\frac{0}{2}) = 1$ (* log 1, 0)

$P2 = H(\text{Sail} \mid C=\text{NO}) = -\frac{0}{6} * \log(\frac{0}{6}) - \frac{6}{6} * \log(\frac{6}{6}) = 1$

2. Split the table by one attribute at a time, and calculate the weighted/expected entropy EH for ~~class~~ Sail with respect to attribute A for the resulting subtables after the split.

$EH(\text{Sail}) = \frac{2}{8} * [\ 1\ ] + \frac{6}{8} * [\ 1\ ] = .811$

PartA.(15pts)Let's try to split by 'Sailboat' first;

Sailboat has two different values; 'Big' and 'Small'.

a) (5pts) Draw the subtable when 'Sailboat' is 'Big' and calculate the entropy $H(C|_{\text{Sailboat = Big}})$

Sailboat	Outlook	Sail
Big	Rainy	yes
Big	Sunny	Yes
Big	Rainy	No
Big	Sunny	NO

$H(C|\text{Sailboat = Big})$

$= -2/4 * \log(2/4) - 2/4 * \log(2/4)$

$= 1$

b) (5pts) Draw the subtable when 'Sailboat' is 'Small' and calculate the $H(C|_{\text{Sailboat = small}})$

Sailboat	Outlook	Sail
Small	Rainy	NO
Small	Sunny	NO
Small	Rainy	NO
Small	Sunny	NO

$H(C|\text{Sailboat = small})$

$= -0/4 * \log(0/4) - 4/4 * \log(4/4)$

$0 * \log_2 0 = 2^0 \quad 1 \quad 2^1 \quad 1 \quad 2$

$= 0$

c) (5pts) Calculate expected/weighted entropy EH

$EH(\text{Sailboat}) = 4/8 * \boxed{1} + 4/8 * \boxed{0} = .5$

PartB (15pts) Now let's try to split our table by 'Outlook' using the same way as partA, Draw the subtables and calculate $H(C|_{\text{Outlook= Rainy}})$, $H(C|_{\text{Outlook= Sunny}})$, and EH

a)(3pts) Draw the subtable when 'Outlook' is 'Rainy', and calculate $H(C|_{\text{Outlook= Rainy}})$

Sailboat	Outlook	Sail
Big	Rainy	yes
Big	Rainy	No
Small	Rainy	NO
Small	Rainy	NO

$H(C|\text{Outlook = Rainy})$

$= -1/4 * \log(1/4) - 3/4 * \log(3/4)$

$= .811$

b) (3pts) Draw the subtable when 'Outlook' is 'Sunny', and calculate H(C| Outlook= Rainy)

Sailboat	Outlook	Sail
Big	Sunny	Yes
Big	Sunny	No
Small	Sunny	No
Small	Sunny	No

$H(C \mid \text{Outlook} = \text{Rainy})$

$= -1/4 * \log(1/4) - 3/4 * \log(3/4)$

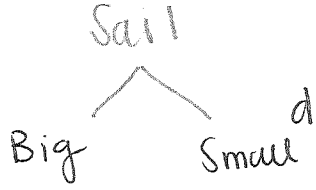
$= .811$

c)(4pts) Calculate the expected/weighted entropy EH

$EH = \underline{4/8} * \underline{.811} + \underline{4/8} * \underline{.811} = .811$

-2.5 ~~8~~ **(5pts) Based on the answers above, what is the first feature you should split on, why? Draw the first split.**

Sailboat as it splits and gives a different variant



draw full table

PartC (15pts)

If you answer of Question 3 is "Outlook", please do the following, if not, go to Q6

4. (5pts) What is Entropy of the right subtalbe (Outlook = Rainy)? Will you further split it?

5. (5pts) What is Entropy of the left subtalbe (Outlook = Sunny)? Will you further split it? If so, what is the next feature you will split on? Go to question8.

-2.5

6. (5pts) What is Entropy of the right subtalbe (Sailboat = Small)? Will you further split it?

NO

$EH = \frac{Sail}{.811} - (Sailboat = small)$

-2.5

7. (5pts) What is Entropy of the left subtalbe (Sailboat = Big)? Will you further split it? If so, what is the next feature you will split on?

yes

$EH = .811 - (Sailboat = Big)$

~~8.~~ (5pts) Draw the final tree

Let's try to split by 'Outlook', which is the only feature we're left with.

When Outlook is Rainy;

	Sailboat	Outlook	Sail
1	Big	Rainy	Yes
3	Big	Rainy	No

Entropy here is still 1

When Outlook is Sunny;

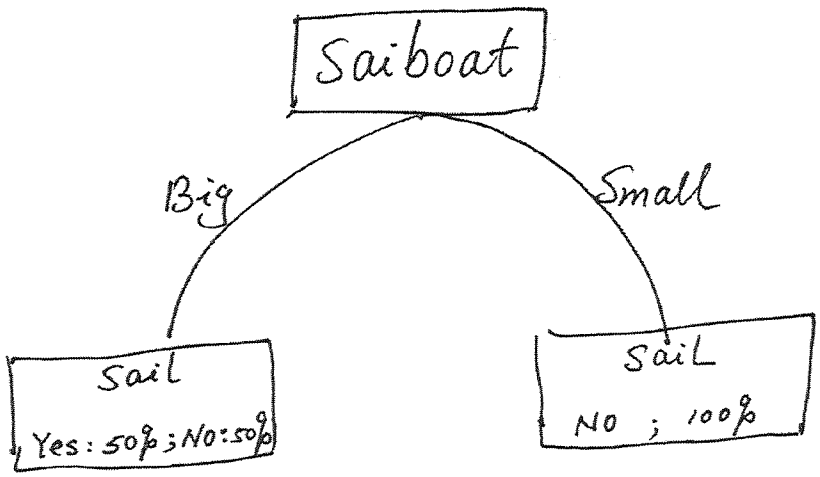
	Sailboat	Outlook	Sail
2	Big	Sunny	Yes
4	Big	Sunny	No

Entropy here is still 1

The weighted/expected entropy of feature Outlook EH(Outlook) is 1. Notice here that before this split, the entropy of the subtable was 1, and after the split the weighted/expected entropy is still 1. IG(Outlook) = H(C) – EH(Outlook) = 0 That means that asking the question about 'Outlook' won't help us make a better prediction, and for that reason, we should not split our table any further.

8. (5pts) Draw the final tree.

Our final tree is;



$$Entropy = -p_1 * \log_2(p_1) - p_2 * \log_2(p_2)$$

$$EH = \frac{sail}{yes} * entropy\left(\frac{sail}{yes}\right) + \frac{sail}{no} * entropy\left(\frac{sail}{no}\right)$$

$$IG = HC - Sail$$

vs

$$HC - Outlook$$

Entropy = $-p1 * \log_2(p1) - p2 * \log_2(p2)$

$= -2/8 * \log_2 2/8 - 6/8 * \log(6/8)$

$= .811$

$H(C \mid \text{Sailboat} = \text{Big}) = -2/4 * \log 2/4 - 2/4 * \log 2/4$

$= -1/2 * \log 1/2 - 1/2 * \log 1/2$

$= 1$

$H(C \mid \text{Sailboat} = \text{small}) = 0/4 * \log 0/4 - 4/4 * \log 4/4$

$= 0 * \log 0 - 1 * \log 1$

$= 0$

$EH = \underline{4/8} * \underline{1} + \underline{4/8} * \underline{0}$

$.5 + 0$

$= .5$ ↓ the better

$H(C \mid \text{outlook} = \text{Rainy}) = -1/4 * \log_2(1/4) - 3/4 * \log_2(3/4)$

$= .811$

$H(C \mid \text{outlook} = \text{Sunny}) = .811$

$EH = 4/8 * \underline{.811} + 4/8 * \underline{.811}$

$= .811$

To print the first "column" from each line, we can do the following:

```
[/data/L05]$ awk '{print $1}' hm.txt
```

By default, each column is separated by a space; for example,
Then executing this command: awk '{print $1}' hm.txt will output:

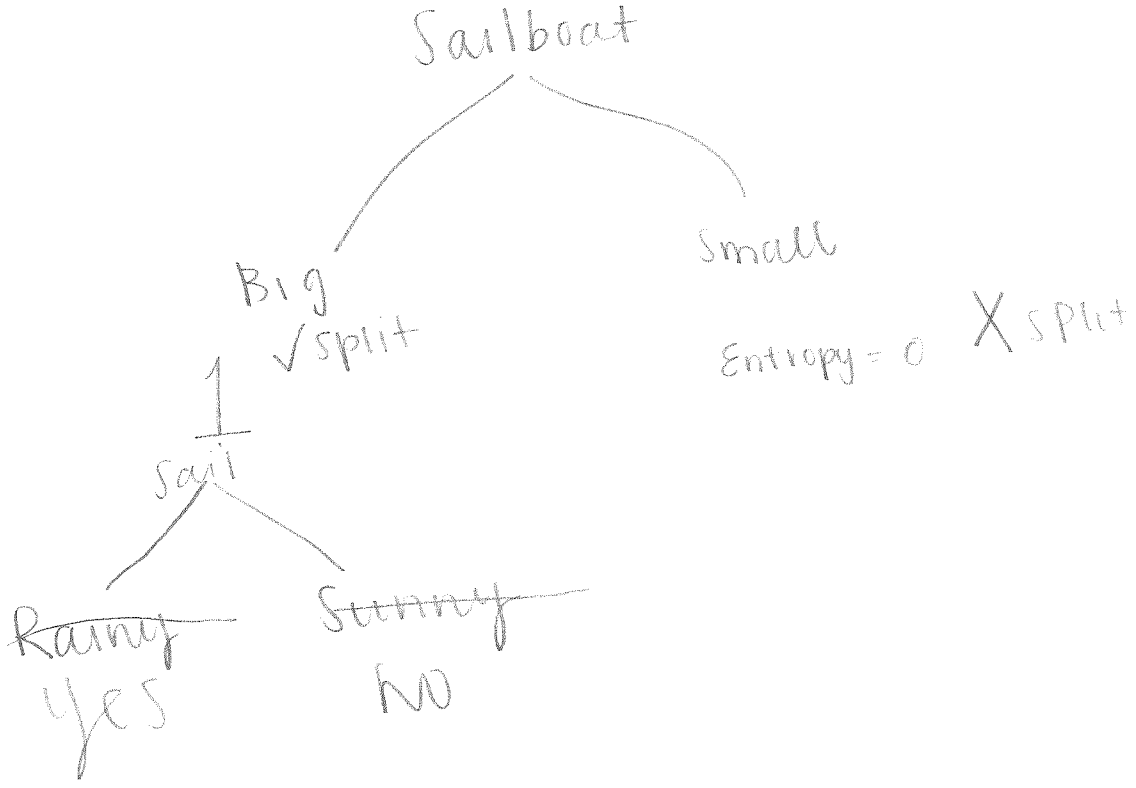
```
This
LetMeIgnoreSome
89Hello
bang
```

$2 corresponds to the 2nd word, so on and so forth

IG = .811 − .5 = .311 ✓

H(c) | Cutlass

.811 − .811 = 0



If we do the following:

```
[/data/L05]$ awk 'NR<=3 {print "Hello"}' mydata.txt
```

Then this would mean: for each line that satisfies the criterion NR<=3 (which is essentially the first three lines), print the word "Hello"

So, the output of that command would essentially be:

```
"Hello"
"Hello"
"Hello"
```

If we remove the condition, and just simply do the following:

```
[/data/L05]$ awk '{print "Hello"}' mydata.txt
```

Then the terminal will print the word "Hello" for every single line.

If we get rid of "Hello", then every single line will be printed:

```
[/data/L05]$ awk '{print}' mydata.txt
```

The command above will print every single line in "mydata.txt". So awk '{print $0}' mydata.txt does.

If we type a number before {print} then:

- The condition will evaluate to true if it's positive
- The condition will evaluate to false if it's 0
- Negative numbers won't be allowed.

For examples:

We have another file "hm.txt" contains the following:

```
This is the first line
LetMeIgnoreSome spaces here
89Hello There...
bang bang da
```

```
[/data/L05]$ awk '2 {print}' hm.txt // This will print every single line
[/data/L05]$ awk '0 {print}' hm.txt // This won't print anything
```

grep 'shop\|cheese' sc.txt

tr '[:lower:]' '[:upper:]' < hm.txt > sc_out.txt

-r
nine
eight

seq 8 11 | sort -n

8
9
10
11

sort -n sc.txt

curl -s http://w > tip.txt

grep -oE 'cliff.|beach' sc.txt | sort | uniq -c

and
\|
or

grep -i 'beach\|cliff' hm.txt

grep -oE 'shop\|cheese' hm.txt

tr -c 'v' 'b' < sc.txt > sc_out.txt

\n

echo $PATH | tr ':' '\n'

after removing read you CANNOT open

```
lata/book/ch03/data]$ grep -o '1990\|1991' movies.csv | uniq -c
      3 1991
      3 1990
      6 1991
```

That's because there were 3 consecutive occurrences of 1991, then 3 occurrences of 1990, and finally 6 occurrences of 1991. Output of 'grep -o '1990\|1991' movies.csv' displayed below:

```
[/data/book/ch03/data]$ grep -o '1990\|1991' movies.csv
1991
1991
1991
1990
1990
1990
1991
1991
1991
1991
1991
1991
```

More comments about the 'sort' command. Say the file 'file.txt' contains the following text:

```
19
201
32
241
```

grep --color -E 'Th|ou' fi.txt

grep -oE '\w {2}' file.txt
└ {2,4} match word of length 2
min MAX

grep -E '[aeiou] {2,}' myfile.txt
these letters ↑ 2+ of

-E 'th.s' file.txt
↓ this or thus or th_s

grep -oE 's....' file01.txt
1 2 3 4
S, Sat
1 2 3 4

-E 'th.*s' file.txt
↓ any # of characters

'^The'
└ beginning of the line

'^[A-K]'

'us$'
└ end of the line

'work*sub'
for astric only

'work.sub'
for dot

'a{3}' txt
└ 3 consecutive a's

'a{2,}' txt

'[a-pz0-9]{2}at' file.txt
or or

print any substring that starts with 'Co...n'

grep -E 'Ch...h' file02.txt

'\w{3}' Female/Yes/Sat/Lady

cut off word & add ending

grep -oE 'L.n' file.txt | tr '\n' 'd'

LundLind

grep -E 'L.*n' file.txt | tr '\n' 'e'

LuncheLince

6 letter ...

7 letter ....

quiz 5 part 1

tr -d ' ' < .txt
└ delete a space

tr -d -c '[aio]' < hm.txt
: compliment
delete all char except

tr -d -c 'c' < file.txt > f1.txt
tr -d -c 'b' < file04.txt > f

seq -w 8 11
08
09
10
11

seq -f "This is line number: %g" 8 11 > lines.txt
8
9
10
11

sed = to filter lines

sed '1,3p' lines.txt p means duplicate
-n
\ won't print anything else unless explicitly asked to print

sed '1,3d' lines.txt
└ deletes lines that match restriction
'1,3!p'

prints starting @ line 4

tail -n +4
\ removes first 3 lines

grep -E '[cbht]{2}at' file.txt → chat that

'[cbht]{2}at' file.txt → bat hat

'[a-p]{2}at' → aat bat hat

'[a-pz0-9]{2}at'
or or

'[0-9]{3}-[0-9]{3}-[0-9]{4}' phone #

. , * , @ is treated like a delimilator

echo 'hello world' | tr ' ' '_' → hello_world

tr -d ' ' → helloworld

tr -d 'i' '[iao]' < hm.txt] delete char from file

-c
└ delete all char except

seq -w 8 11
08
09
10
11

seq -f 'This is line #: %g' 8 11
→ This is line #: 8
9
10
11

sed '1,3p' lines.txt
└ duplicate printing line

sed -n '1,3p' lines.txt
└ print lines 1 to 3

< lines.txt sed '1,3d'
└ deletes every line that matches restrictions

* tail is faster tail -n +4 to remove first 3 lines

quiz 5 part 2

sed -n '1~2p'
start | print
increment

display all the lines starting at 31
sed '1,30d' file01.txt

sed -n '1,3p ; 5,19p'] display lines 1 to 3 & 5 to 19

sed -n '1,3p' file.txt

header -a output

grep -i chapter alice.txt

grep -E '^Chapter.*The' alice.txt

grep -E '[a-zA-Z][a-zA-Z0-9]{1,}@[a-zA-Z]{1,}\.'
1 or more

---

csvsql --query "SELECT * FROM tips" tips.csv | csvlook | head -n 5

ORDER BY day (col name) DESC

"SELECT SUM(bill) FROM tips" tips.csv

AVG
MAX
MIN
COUNT(DISTINCT)
# of

# of Rows Count(*)

display all rows but Sex Smoker & Tip

csvsql --query "SELECT sex, tip FROM

Name: Shefali E DATA 210: Quiz 05
Session # __1 Page 1 of

Q01) Assume that the content of the file 'file03.txt' is the following:

Lunch
Line
Day
Dinner
Laugh

What's the output of this statement:

grep -E '^L' file03.txt | tr '\n' '.'

Ans: Lunch.Line.Laugh.

Q02) Assume that the content of the file 'file04.txt' is the

following: Nobody likes rats

I have a cat
A lot of people think that bats are birds
That's great!

What's the output of this statement:

grep -oE '[hcbr]at' file04.txt | tr -d '\n'

Ans: ratcathatbathat

Q03) What's the command that will print all the lines that start with 'Thi' AND end with 'line'; from the file 'file01.txt

Ans: grep -E '^Thi.*line' file01.txt grep -E '^Thi.*line$' file01.txt

Q04) What's the 'grep' command that will print the lines that contain any 5-letter substring that starts with the two letters 'Ki' and end with the letter 'e' from the file file04.txt (use the option for the extended regular expressions)

print any lines that contain 'Ki..e'

Ans: grep -E 'Ki..e' file04.txt

.. 2 letters

Q05) What's the command that will print all the lines that start with 'Thi' OR end with 'line'; from the file 'file01.txt'

Ans: grep -E '^Thi | \nS' file01.txt

line

Q06) Assume that the content of the file file03.txt is the

following: 27.18,2.0,Female,Yes,Sat,Dinner,2

What's the output of this statement:

grep -oE 'e....' file03.txt

Ans: emale or es,Sa

Q07) Assume that the content of the file 'file02.txt' is the following: Lunch
Line
Day
Dinner
Laugh

What's the output of this statement:

grep -oE 'L.*n' file02.txt | tr '\n' 'j'

Ans: LunjLinj

Q08) Assume that the content of the file file03.txt is the following:

27. 18 2.0 Female Yes Sat Lunch 2

What's the output of this

statement: grep -oE '\w{3}'

file03.txt | tr -d '\n'

Ans: FemaleYesSatLun

Q09) What's the command that will print all the lines that contain the text 'Th*s'; from the file 'file01.txt'?

Ans: grep -oE 'Th*s' file01.txt

Q10) Assume that the content of the file 'file02.txt' is the following:

27.18 2.0 Female Yes Sat

Lunch 2 What's the output of this

statement:

grep -oE '\w{2,4}' file02.txt | tr '\n' '.'

Ans: 27.18.Fema.le.Yes.Sat.Lunc.

Q11) What's the 'csvstat' command that will show the sum for the columns 'day,sex,size' from the file 'file01.csv'

Ans: csvstat -c day, sex, size file01.csv —sum

Q12) What's the 'csvstat' command that show all the statistics for all the columns from the file 'file01.csv'

Ans: csvstat file01.csv

Q13) What's the 'csvsql' command that will display the number of rows in the file 'file03.csv'

Ans: csvsql —query "SELECT COUNT(*) FROM file03" file03.csv

Q14) What's the single 'sed' command that will display all the lines starting at line number 47; for the file file01.txt

Ans: sed '1,46d' file01.txt

Q15) What's the single 'sed' command that will only display the lines from 7 to 11 for the file file02.txt

Ans: sed -n '7,11p' file02.txt

Q16) What's the 'csvsql' command that will display all the 'number of unique/ distinct' values for the column 'size' from the file 'file01.csv'

Ans: csvsql --query "SELECT COUNT(DISTINCT(size)) FROM file01" file01.csv

Q17) What's the 'csvsql' command that will display all the rows but only the columns 'day' and 'sex' from the file 'file02.csv'

Ans: csvsql --query "SELECT day, sex FROM file02" file02.csv

Q18) Given that our file 'file01.txt' contains the following

text: mmm mm mmmm m

What's the output of the following commands:

a) grep -oE 'm{2,5}' file01.txt | wc -l
Ans: mmm mm mmmm
3

b) grep -oE 'm{2,}' file01.txt | wc -l 2 or more
Ans: mmm mm mmmm
3

c) grep -oE 'm{2}' file01.txt | wc -l match exactly 2
Ans: mm
4

mmm mm, mm,mm, m
1 2 3 4

Q19) What's the 'csvsql' command that will display the maximum value for the column 'bill' from the file 'file02.csv'

Ans: csvsql —query "SELECT MAX(bill) FROM file02" file02.csv

Q20) What's the command that will delete all occurrences of the letters 'i', 'o', and 'a'; from the file 'file01.txt'

Ans: tr -d '[ioa]' < file01.txt

tr -d 'i', 'a', 'o' < file01.txt

Q21) What's the 'seq' command that will display the following:

12
11
10
09
08

because the digits lined up

Ans: seq -rw 8 12 seq [-w] 12 -1 8

Q22) What's the output of the following command: seq 3 | header -a Numbers

Ans: Numbers
1
2
3

Q23) Write the command that will match and print all phone number that have the following format: xxx-xxx-xxxx, where x could be any number; from the file 'file01.txt'

Ans: grep ~~-E~~ -E '[0-9][0-9]{2}-[0-9]{3}-[0-9]{4}' file01.txt

-oE

'[0-9]{3}-[0-9]{3}-[0-9]{4}' file01.txt

Q24) What's the 'csvsql' command that will display the content of the file 'file01.csv' sorted in an ascending order based on the column 'size'

Ans: csvsql —query "SELECT * FROM file01 ORDER BY size DESC"

Q25) What's the command that will delete all the letters except the letter 'b' from the file 'file02.txt' and saves the result in the file 'file01.txt'.

Ans: tr -d -c 'b' < file02.txt > file01.txt

quiz 6 part 1

csvstat tips.csv --unique

-c time, col, date] column titles

# of unique values for each column

-- min
-- max
-- sum
-- mean
-- median
-- stdev
-- freq
-- len

csvstat file01.csv] all the stats for all the col

\list or \l] show the database
^
all

CREATE DATABASE data210; } CREATE TABLE count (name TEXT NOT

\connect data210] change a database

\dt] show the current tables

\d employee] description of table

INSERT INTO employee VALUES(4, 'Ben', 'Smith');

SELECT * FROM count;

!= IS NOT NULL

\h] to show all commands

INSERT INTO cities VALUES('', 90, 'US');

UPDATE
SET
WHERE;

UPDATE db3 SET size = 5 WHERE grade = 4;

DROP TABLE user;

DELETE FROM comp;] delete all rows

WHERE age = 5

name TEXT NOT NULL CHECK(name <> '')

not null

empty string

Primary Key
- unique
- not null
- a table = 1 PK

Foreign Key Constraint
- enforce between 2 col in 2 tables
- link is created between 2 tables when [PK] is referenced by [FK] in other
- Save Space } accessing/ table querying is faster
- must NOT be unique
- 1 table = ∞ FK

Join
combines 2 tables to 1

ex1
CROSS JOIN
+ generate test data
+ look for missing val
+ every row is joined regardless of match

```
ALTER TABLE countries RENAME county_code TO code;

SELECT * FROM company CROSS JOIN countries
    ORDER BY company.name DESC;
```

companies.name

user base = 1000 * (# of col) = 5000 dp

1000 * # of col in 1st + # of unique # of col in 2nd

Name: Session # __

Shefali Emmanuel True or False (16pt)

1. Foreign keys must be unique. Ans: FALSE
2. A single table can contain only one primary key constraint. Ans: TRUE
3. A single table can contain only one foreign key constraint. Ans: FALSE
4. Primary keys must be unique. Ans: TRUE

Would the following PostgreSQL code (question 5 to 11) cause an error? If yes, please explain why (28pt)

5. CREATE TABLE company (code CHAR(5) PRIMARY KEY, city VARCHAR(30), state CHAR(2));
   CREATE TABLE employee (f_name VARCHAR(30), l_name VARCHAR(30), zip_code CHAR(5) REFERENCES company);
   INSERT INTO company VALUES('0035', 'Hattiesburg', 'MS') ;
   INSERT INTO company VALUES('0035', 'Charleston' , 'SC');

Ans: ERROR because 0035 is used 2x which violates that there can only be 1 PK per Table

6. CREATE TABLE cities (code CHAR(5) PRIMARY KEY, city VARCHAR(30), state CHAR(2));
   CREATE TABLE employee (f_name VARCHAR(30), l_name VARCHAR(30), zip_code CHAR(5) REFERENCES cites);
   INSERT INTO employee VALUES('Michael', 'Lewis', '29492') ;

Ans: ERROR because you must first INSERT INTO cities before INSERTING INTO employee

7. CREATE TABLE store (code CHAR(5) PRIMARY KEY, city VARCHAR(30), state CHAR(2));
   CREATE TABLE shopper (f_name VARCHAR(30), l_name VARCHAR(30), zip_code CHAR(5) REFERENCES place);
   INSERT INTO store VALUES('22180', 'Vienna', 'AT');
   INSERT INTO shopper VALUES('Malcolm', 'Lewis', '29492') ;

Ans: ERROR because 22180 and 29492 do not match so they can not be mapped together

8. CREATE TABLE countries (country_code char(2) PRIMARY KEY, country_name TEXT UNIQUE);
   INSERT INTO countries VALUES ('ES', 'Netherlands');
   INSERT INTO countries VALUES ('AT', 'Netherlands');

Ans: ERROR because Netherlands is used 2x

Duration: 25 minutes

9. CREATE TABLE company (name TEXT NOT NULL, employees INT, country_code CHAR(2));
   INSERT INTO company VALUES('', 9000, 'US'); // '' is two single quotations

Ans: NO ERROR

10. CREATE TABLE employee (f_name VARCHAR(30), l_name VARCHAR(30), is_manager CHAR(1) CHECK (is_manager IN ('Y', 'N')));
    INSERT INTO employee VALUES('Malcolm', 'Gladwell', 'H');

Ans: ERROR because H IS NOT 'Y' or 'N'

11. CREATE TABLE hotel (name TEXT NOT NULL CHECK (name <> 'DK'), rooms INT, country_code CHAR(2));
    INSERT INTO hotel VALUES('DK', 2000, 'AU');

Ans: ERROR because DK is not allowed

12. What's the PostgreSQL command that will delete the database 'data210' .

Ans: DROP DATABASE data210;

13. What's the PostgreSQL command that will change the column 'rooms' to the value 3, for all the rows that have the value 10 for the column 'size' in table 'db3'?

Ans: UPDATE db3 SET rooms = 3 WHERE size = 10;

14. What's the PostgreSQL command that will delete the rows in table 'tips' where the value of the column 'size' is equal to the number 3?

Ans. DELETE FROM tips WHERE size = 3;

15. What's the PostgreSQL command that will delete all the rows in table employee?

Ans: DELETE FROM employee;

16. What's the PostgreSQL command that will delete the table user?

Ans: DROP TABLE user;

17. What's the command that will describe the table 'faculty' in PostgreSQL?

Ans: \d faculty

18. What's the command that will list all the databases in PostgreSQL?

Ans: \list

Duration: 25 minutes

Name: Session #__

19. What's the command that will make the current database 'my_database' in PostgreSQL?

Ans: \connect my_database

Assume the following: (8pt)

SELECT * FROM countries;

code	name
FR	France
DE	Germany

SELECT * FROM company;

name	age	salary
John	28	80000
Grace	26	120000

What's the output of the following commands (question 20 and 21) if there is no error? If there is an error, explain why.

20. SELECT code, age FROM countries CROSS JOIN company;

all other from company

Ans:

CODE	AGE
FR	28
FR	26
DE	28
DE	26

list for each

21. SELECT name, age FROM countries CROSS JOIN company;

Ans: ERROR because with name you have to do countries.name
ALSO countries doesn't contain an age column

countries.name

Name: Session # __

22. Given that we have a table 'users' with the following columns: 'f_name', 'l_name', 'city', 'state', and 'zip_code', filled with 400 rows. How many data points would we save if we split our table into two tables 'users' and 'zip_codes', where 'zip_codes' has the columns 'zip_code', 'city', and 'state'. Assume that we have 30 unique zip codes.

Ans:
(user base) * (# of column in 1st) + (# of unique) * (# of column in 2nd)
(400 * 5)-((400*3) + (30 * 3)) = 710 data points

Given that we have a table 'instructor' with the following columns: 'f_name' VARCHAR(30), 'l_name' VARCHAR(30),, 'department_code' CHAR(3), 'department_name' VARCHAR(30), and 'department_address VARCHAR(100), filled with 300 rows. We want to split the table into two tables: 'instructors' table and 'department' table to remove the data redundancy.

23. Which table should be created first, instructor or department, why? (2pts)

Ans: department since 'department_code needs to be a primary key

24. Please write the statement to create the two tables and using foreign key to build up the relationship between the two tables. (6pts)

```
Ans: CREATE TABLE department_code(
        code CHAR(3) PRIMARY KEY,
        department_name VARCHAR(30),
        department_address VARCHAR(100));

CREATE TABLE instructor(
        f_name VARCHAR(30),
        l_name' VARCHAR(30),
        department_code CHAR(3) REFERENCES department));
```

25. After to split the original instructor table, which has 300 rows, into the two tables shown above, how many data points would we save? Assume that we have 20 unique department code.

Ans:
(user base) * (# of column in 1st) + (# of unique) * (# of column in 2nd)
300 * 5 - ((300 * 3) + (20 * 3)) = 540 data points

Duration: 25 minutes

Quiz 6 Grading

4pt each

8. Error because country_name is unique (-4)

20.

code	age
FR	28
FR	26
DE	28
DE	26

(-4)

(-2)

22. $(400 * 5) - ((400 * 3) + (30 * 3)) = 710$

25. $300 * 5 - ((300 * 3) + (20 * 3)) = 1500 - 960$

$= 540$

(-2)

import csv to table in PostgreSQL

① extract CREATE TABLE syntax

```
csvsql tips.csv
CREATE TABLE tips(
    bill DECIMAL NOT NULL;
    sex VARCHAR(6) NOT NULL
```

② create db csv

```
CREATE DATABASE csv
\connect
```

③ copy tips.csv into postgres container's data210 folder

```
docker cp tips.csv model:/data210/.
```

④ import csv file to postgreSQL table

```
postgres@c7b82e820761:/$ psql -d csv -c "\copy tips FROM '/data210/tips.csv'
                                        delimiter ',' csv header
```

⑤ show the first 10 rows of those smoker is 'yes'

```
CSV=# SELECT * FROM tips WHERE smoker='yes' LIMIT 10;
```

count(age) = # of times

INNER JOIN

only rows that satisfy the condition in the ON clause (matched rows) are returned in the result set

```
csv=# CREATE TABLE A (a INT);
                     (b INT);

INSERT INTO A VALUES (102), (104), (106), (107);

SELECT A.a, B.b FROM A, B WHERE A.a = B.b;
SELECT * FROM A INNER JOIN B on A.a = B.b;
```

for each customer → total # of items purchased & price

```
SELECT
    customers.name AS customer,
    count(items.name) AS item,
    sum(cartitems.qty * items.price) AS total

FROM customers
        INNER JOIN carts ON
          carts.customer_id = customers.id
        INNER JOIN cartitems.cart ON
          cartitems.cart_id = carts.id
        INNER JOIN items ON
          items.id = cartitems.item_id
GROUP BY customers.name
```

ex: csv=# SELECT sex, COUNT(sex) FROM tips
GROUP BY sex;

Views
└ pseudo tables

CREATE VIEW company-view AS SELECT id, name, age FROM company;

list all views [\dv

#1 DROP TABLE company CASCADE;] deletes the table & dependencies

#2 DROP VIEW company_view;

DROP TABLE company;

Indexes / Indices
└ when PK are created = btree
└ makes queries much faster
✓ create Index for non primary key column ? NOT unique
✓ more than 1 index constraint per table

How 2 create index constraint on col

① NOT specifying type ② EXPLICITLY specifying type

csv=# CREATE INDEX my-i ON company(name)

\d my-i

CREATE INDEX my-2nd [illegible]
ON company USING hash(nam
(or)
ON company USING btree(nam

DROP INDEX my-i;

look into putting links on webside
↳ put in future hours

what are the impact on my income taxes → you hv to file in both countries

NC 1
SC 11
GA
MS

Mike $
Daisy $$
Joseph

23 [13 + 16 + 19] 48
25 [15 + 17 + 18] 50

NC 13
SC 16 + 15 + 17
GA 19
MS 18