**Fundamentals of Data Science: Homework 2**

**Author: Shivani Grover**

**In preparation for homework I took the following steps:**

* Unix: practiced creating RMD files using UNIX, R studio and JYPITER.
* Installed Jypiter notebook to save code
* Downloaded code file from jypiter
* Code files are available for review in the compressed folder

1. **Practiced suggested Unix Commands.**

**Answer:** Code is in the unixusecase.txt file in the uploaded folder**.**

**2. How effective is the compression for homework file tarball?**

**Answer:** The code used for extraction is below. The tar file 1 MB and the extracted folder is 2.7 MB. That is a compression ratio of 44.44%. Very useful to learn this to manage big genomic and transcriptomic data files later on.

sdm0f00h3f22140:FDS apple$ tar -xzvf homework2.tar.gz

x homework/homework1/

x homework/homework1/.DS\_Store

x homework/homework1/homework1\_html.html

x homework/homework1/nanostring\_annotation.txt

x homework/homework1/TB\_microbiome\_annotation.txt

x homework/homework1/viral.fasta

x homework/homework1/TB\_microbiome\_data.txt

x homework/homework1/TB\_nanostring.txt

sdm0f00h3f22140:FDS apple$ ls -lh homework2.tar.gz

-rw-r--r--@ 1 apple staff 1.0M Jan 23 11:02 homework2.tar.gz

sdm0f00h3f22140:FDS apple$ du -sh extracted\_folder

du: extracted\_folder: No such file or directory

sdm0f00h3f22140:FDS apple$ ls -lh homework2.tar.gz

-rw-r--r--@ 1 apple staff 1.0M Jan 23 11:02 homework2.tar.gz

sdm0f00h3f22140:FDS apple$ du -sh extracted\_folder

du: extracted\_folder: No such file or directory

sdm0f00h3f22140:FDS apple$ du -sh homework2

du: homework2: No such file or directory

sdm0f00h3f22140:FDS apple$ ls -lh homework2

ls: homework2: No such file or directory

sdm0f00h3f22140:FDS apple$ du -sh homework

2.7M homework

sdm0f00h3f22140:FDS apple$

**3. Research the chmod function**

**Answer:** The chmod function helps define the permissions of a directory. I created a folder “practice” and “TB\_microbiome\_data.txt” to learn the command’s usage.

chmod function works with ls -l and cd commands. We can go to any directory with cd and check read, write, and execute permissions using the ls -l command. Then, I can change that for specific users. This will be very useful when I for example want to save my original script files for a particular data or when learning I want to not interfere with certain files. This will keep any code errors form changing certain files.

Code:

sdm0f00h3f22140:FDS apple$ mkdir practice

sdm0f00h3f22140:FDS apple$ echo "The chmod command is used to change the permissions of a file or directory in Unix/Linux systems. Permissions control who can read, write, or execute a file."

The chmod command is used to change the permissions of a file or directory in Unix/Linux systems. Permissions control who can read, write, or execute a file.

sdm0f00h3f22140:FDS apple$ cd homework

sdm0f00h3f22140:homework apple$ ls -l

total 0

drwxr-xr-x@ 9 apple staff 288 Apr 30 2024 homework1

sdm0f00h3f22140:homework apple$ cd homework1

sdm0f00h3f22140:homework1 apple$ ls -l

total 5560

-rw-r--r--@ 1 apple staff 1245 Apr 29 2024 TB\_microbiome\_annotation.txt

-rw-r--r--@ 1 apple staff 22503 Apr 29 2024 TB\_microbiome\_data.txt

-rw-r--r--@ 1 apple staff 524314 Apr 30 2024 TB\_nanostring.txt

-rw-r--r--@ 1 apple staff 1459935 Apr 30 2024 homework1\_html.html

-rw-r--r--@ 1 apple staff 9198 Apr 30 2024 nanostring\_annotation.txt

-rw-r--r--@ 1 apple staff 813612 Apr 29 2024 viral.fasta

sdm0f00h3f22140:homework1 apple$ ls -l TB\_microbiome\_data.txt

-rw-r--r--@ 1 apple staff 22503 Apr 29 2024 TB\_microbiome\_data.txt

sdm0f00h3f22140:homework1 apple$ chmod 600 TB\_microbiome\_data.txt

sdm0f00h3f22140:homework1 apple$ ls -l TB\_microbiome\_data.txt

-rw-------@ 1 apple staff 22503 Apr 29 2024 TB\_microbiome\_data.txt

sdm0f00h3f22140:homework1 apple$ chmod 644 TB\_microbiome\_data.txt

sdm0f00h3f22140:homework1 apple$ ls -l TB\_microbiome\_data.txt

-rw-r--r--@ 1 apple staff 22503 Apr 29 2024 TB\_microbiome\_data.txt

sdm0f00h3f22140:homework1 apple$ chmod u+x TB\_microbiome\_data.txt

sdm0f00h3f22140:homework1 apple$ ls -l TB\_microbiome\_data.txt

-rwxr--r--@ 1 apple staff 22503 Apr 29 2024 TB\_microbiome\_data.txt

sdm0f00h3f22140:homework1 apple$

**4. a) How many FC receptor genes are present in the ’TB\_nanostring.txt’ file?**

**Answer: 4**

**b)** How many samples (rows) in the ’nanostring\_annotation.txt’ do not have a co-morbid condition

or other risk factor?(i.e., inverse search – how many rows do not have a "Yes")

Answer: 86

c) How many coronavirus genomes are present in the ’viral.fasta’ file? How many of these are

SARS-COV-2?

Answer: coronavirus genomes: 3; SARS-COV-2: 0

d) How many times does the letter ’A’ (capital or lowercase) appear in all the files from the homework1

tar file? (i.e., ignore case).

Answer: 373144

e) What Staphylococcus species are present in the ’TB\_microbiome\_data.txt’ file? (hint: each

separate microbe has its own row in the file). Print out the counts for Mycobacterium tuberculosis.

How many Streptococcus species are present?

Answer: Staphylococcus species: 4 [Staphylococcus aureus, Staphylococcus epidermidis, Staphylococcus haemolyticus, Staphylococcus saprophyticus]

Mycobacterium tuberculosis: 1 count

Streptococcus: 16 count

**Code is in the file in the tar folder**

**5. Answer: less -SN +/coronavirus viral.fasta**

Less is advantageous over MS word because Office tends to crash when opening multiple big files. It can be used without a graphical interface and has more navigation commands like grep and cat.

**6. Answer:** Beginning of a line press: 0; end: $

Insert: press I or i

Paste: P or p

Delete: x, dw or dd

I used :q! to exit vim without saving my result. I used :wq to save my file and quit.

Less is better for viewing while Vim is better for editing when we don’t have a graphical interface. Less is easier to use than vim. I believe mistakes can happen in vim more often than less.

**7. Answer:**

**Learn about pipes and redirects in Unix. In which scenarios would you use them, and why are they helpful?**

A pipe (>) seems similar to a cat command, except it doesn’t append files but appends commands. This is useful to write a string of code. A Redirect sends instructions of commands to specific files. This is useful to run the code. It is written by using >>.

I will use these commands to extract and clean my data for ex: look for specific gene sequence files in whole genome data. It will be very helpful to not go through the entire genome to find one gene variant sequence.

**Describe what the following commands do:**

**(a) ls -l | less:** Listed details of current directory

**(b) ls -l > directory\_contents.txt:** Redirected the output of the command to a new file.

**(c) ls -l » directory\_contents.txt:** Error in above command. Need to use single >

**(d) cat directory\_contents.txt | head -3 | tail -2:** Used to search a file. Head selects first two lines and tail outputs the last two lines of the first three lines.

**(e) ls | grep -c html:** Output for how many files has html in their name. Answer: 1

**(f) ls | wc -l:** Counted files in the directory. Answer: 10

**(g) cat file1.txt file2.txt > file3.txt:** combined file 1 and 2 created by echo command; contents into 3

**You can also us pipes in R! Investigate how to do this and give the code for a great example.**

**Code:**

sdm0f00h3f22140:homework1 apple$ vim viral.fasta

sdm0f00h3f22140:homework1 apple$ ls -l less viral.fasta

ls: less: No such file or directory

-rw-r--r--@ 1 apple staff 813612 Apr 29 2024 viral.fasta

sdm0f00h3f22140:homework1 apple$ ls l >less viral.fasta

ls: l: No such file or directory

sdm0f00h3f22140:homework1 apple$ ls -l>less viral.fasta

sdm0f00h3f22140:homework1 apple$ ls -l>grep "Staphylococcus" TB\_microbiome\_data.txt

ls: Staphylococcus: No such file or directory

sdm0f00h3f22140:homework1 apple$ ls/grep "Staphylococcus" TB\_microbiome\_data.txt

bash: ls/grep: No such file or directory

sdm0f00h3f22140:homework1 apple$ ls|grep "Staphylococcus" TB\_microbiome\_data.txt

Staphylococcus aureus 0 0 0 0 730 0 0 0 00 0 0 0 0 43 0 125 534 13860 5 2748 0 3722 0 6512 0 402 0 0 0

Staphylococcus epidermidis 13 21356 41 18 725 63 32 25958 16 161 28 62 26 21140 0 12 86 46 00 3018 34 5708 0 900 31 18146 0 1212 0

Staphylococcus haemolyticus 0 0 0 0 0 0 0 00 0 0 0 0 0 0 0 4 0 0 0138 0 0 0 0 0 0 3 0 0

Staphylococcus saprophyticus 0 0 0 0 0 0 0 17 0 0 0 0 0 0 0 0 86 0 00 0 0 0 0 0 0 0 0 0 0

sdm0f00h3f22140:homework1 apple$ cat file1.txt file2.txt > file3.txt

cat: file1.txt: No such file or directory

cat: file2.txt: No such file or directory

sdm0f00h3f22140:homework1 apple$ ls| grep viral

viral.fasta

sdm0f00h3f22140:homework1 apple$ ls -lh/less

ls: invalid option -- /

usage: ls [-@ABCFGHILOPRSTUWXabcdefghiklmnopqrstuvwxy1%,] [--color=when] [-D format] [file ...]

sdm0f00h3f22140:homework1 apple$ les -lh|less

bash: les: command not found

sdm0f00h3f22140:homework1 apple$ echo "New line" >> files\_list.txt

sdm0f00h3f22140:homework1 apple$ ls -l | less

sdm0f00h3f22140:homework1 apple$ ls -l > directory\_contents.txt

sdm0f00h3f22140:homework1 apple$ ls -l » directory\_contents.txt

ls: »: No such file or directory

-rw-r--r--@ 1 apple staff 682 Jan 24 16:33 directory\_contents.txt

sdm0f00h3f22140:homework1 apple$ cat directory\_contents.txt | head -3 | tail -2

-rw-r--r--@ 1 apple staff 1245 Apr 29 2024 TB\_microbiome\_annotation.txt

-rwxr--r--@ 1 apple staff 22503 Apr 29 2024 TB\_microbiome\_data.txt

sdm0f00h3f22140:homework1 apple$ ls | grep -c html

1

sdm0f00h3f22140:homework1 apple$ ls | wc -l

10

sdm0f00h3f22140:homework1 apple$ cat file1.txt file2.txt > file3.txt

cat: file1.txt: No such file or directory

cat: file2.txt: No such file or directory

sdm0f00h3f22140:homework1 apple$ echo "file 1">file1.txt

sdm0f00h3f22140:homework1 apple$ echo "file 2">file.txt

sdm0f00h3f22140:homework1 apple$ cat file1.txt file2.txt > file3.txt

cat: file2.txt: No such file or directory

sdm0f00h3f22140:homework1 apple$ echo "file 2">file2.txt

sdm0f00h3f22140:homework1 apple$ cat file1.txt file2.txt > file3.txt

sdm0f00h3f22140:homework1 apple$

**8. Answer:** I learned the Find command. I used to find all files with a particular name. It is very useful when no GUI is available.

Below is the syntax I used:

sdm0f00h3f22140:homework1 apple$ find . -name "file1.txt"

./file1.txt

sdm0f00h3f22140:homework1 apple$