COMP 370 Exam 1

Answer the following questions. Do your work and write your answers in the space provided on these exam sheets.

Question 1: Data file exploration (10 pts)

You're given the data file grades.csv and have no idea what's in it. You want to pop it onto your UNIX dev server and use the command line to conduct a few analysis. Give the commands you'd use to do the following.

a) Download the file from the URL you've been given for it: http://files.mcgill.ca/grades.csv (2 pts)

b) Print out the number of lines in the file. (2 pts)

c) Print out the top 30 lines in the file (2 pts)

d) Extract the first 3000 records of the file (*without the header line*) into a file named grades.raw.3000.csv (4 pts)

```
head -n 3001 grades.csv | tail -n 3000 > grades.raw.3000.csv
```

Question 2: Cloud dev machine (10 pts)

a) Explain using a diagram and text how a cloud-based data science machine can provide better performance through proximity compared to a laptop. (4 pts)

b) Following the approach taken in class, give the steps required to create and configure a new user account on your cloud dev machine. (3 pts)

1. Log in as a super user (e.g., “ubuntu”)
2. Run the adduser command as the super user
3. Add the user to the sudo group (not necessary)
4. Install the user’s public key into their user’s .ssh directory in the authorized_keys file

c) What is sudo for? Give three examples of situations where you’ve had to use it within the context of homeworks thus far. (3 pts)

Question 3: Data science details (10 pts)

a) Diagram the steps in the data science process. (2 pts)

b) In what two steps does machine learning tool use/development fit most often? Why? (4 pts)

c) Explain the concept of loops in the data science process. Use your experience from the My Little Pony analysis to give a concrete example of how a loop might occur in the process. Be specific and indicate the steps the loop involves. (4 pts)

Question 4: Data Science Tooling (10 pts)

a) What are the 3 primary reasons for developing code discussed in class? (3 pts)

b) Provide the high-level structure for a data science project outlined in class. Give an example file that would sit in each part. (5 pts)

c) What are the three principles of UNIX CLI tool design that carry over to data science CLI design? (2 pts)