General Concepts

1. What is TCGA and why is it important?

TCGA is the Cancer Genome Atlas. It is important because it provides genomics data for access by the research community.

2. What are some strengths and weaknesses of TCGA?

Strengths: Large amounts of data (of over 30 different cancer types) that can be accessed by the public; integrates DNA, RNA, and protein tumor sample data allowing for multi-omic analysis

Weaknesses: Data from untreated patients, so doesn't offer too much insight about the effects of treatment; barely any immune-oncology data

Coding Skills

1. What commands are used to save a file to your GitHub repository?

git status
git add file_name
git commit -m "informative message about the file"
git push

2. What command(s) must be run in order to use a package in R?

```
if (!require(___)){
  install.packages(___)
}
```

3. What command(s) must be run in order to use a Bioconductor package in R?

```
if(!require(___))
        BiocManager::install(____)
library(___)
```

4. What is boolean indexing?

Applying boolean variables to data using a vector in order to filter out or pinpoint data.

What are some applications of it?

It is used to apply a boolean mask on a specific row or column in a dataframe so we can see which rows/columns have certain types of data and which don't.

5. Draw a mock up (just a few rows and columns) of a sample dataframe. Show an example of the following and explain what each line of code does.

USC_students

Age	**Gender**	**Major**	**Height (cm)**
19	F	Business	170
21	NA	Quantitative Biology	181
18	M	Communications	154

a. an ifelse() statement

ifelse(USC_students$age < 20, "underclassmen, "upperclassmen")

This line of code replaces every row in the age column with a categorical variable of either being "underclassmen" or "upperclassmen" in college. If the age of the student is less than 20, then they are underclassmen, else (greater than or equal to 20) they are upperclassmen.

b. boolean indexing

na_mask <- ifelse(is.na(USC_students$gender) == "NA", F, T)

USC_students_cleaned <- USC_students[na_mask,]

These two lines of code remove the NA values from the gender column by creating a boolean mask and applying it to the column. The first line of code creates the boolean mask using an ifelse statement. The second line of code makes a new dataframe excluding these NA values by only outputting the rows that match the conditions of the mask.