General Concepts

1. What is TCGA and why is it important?

TCGA, also known as The Cancer Genome Atlas, is a project that has made genomic, transcriptomic, epigenomic and proteomic cancer data publicly accessible to researchers around the world. It is important because researchers have used TCGA data to gain a better understanding of the genetic basis of cancer through conducting statistical analyses.

2. What are some strengths and weaknesses of TCGA?

Some strengths of TCGA is that it contains patient data from over 30 different types of cancer and it has clinical, genomic and transcriptomic data that lines up with each patient. However, one weakness of TCGA is that many of the clinical data columns are empty and/or inconsistently inputted and do not provide much additional value.

Coding Skills

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

cd into local repository, git status to check files that have local changes, git add, git commit -m “informative message”, git push

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

install.packages() and library() functions

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

BiocManager::install() and library() functions

4. What is boolean indexing? What are some applications of it?

Boolean indexing is the process of subsetting a dataframe based on a condition using a boolean vector. One way to do this is by creating a boolean mask that defines the type of data you want to keep (give it a T value) or remove (give it a F value) and then applying the mask on either the rows or columns of the dataframe. This will create a new data frame that is a subset of the old one, but only has the rows or the columns that you want to use in further analysis. Some applications of boolean indexing are removing NA or empty values, creating a clinical data frame with a specific demographic of patients (i.e. females or cancer patients treated with chemotherapy).

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. Sample dataframe called world

	Country	Capital	National Plant	Population
1	USA	Washington DC	Rose	331,002,651
2	Canada	Ottawa	Maple	37,742,154
3	Mexico	Mexico City	Dahlia	128,932,753
4	Ireland	Dublin	Shamrock	4,937,786
5	Argentina	Buenos Aires	Cockspur Coral Tree	45,195,774
6	Japan	Tokyo	Cherry Blossom	126,476,461

a. an ifelse() statement

```
pop_mask <- ifelse(world$Population < 40,000,000, F, T)
```

- This mask uses an ifelse statement to create a vector of TRUE and FALSE values. If the population in a specific row is less than 40,000,000, the vector will add a FALSE value, else it will add a TRUE value (in which case the population is greater than or equal to 40,000,000).

```
TRUE FALSE TRUE FALSE TRUE TRUE
```

b. boolean indexing

```
world_cleaned <- world[pop_mask, ]
```

- This line uses the mask and boolean vector created in part a to subset the world dataframe by selecting only the rows where the Population column is greater than 40,000,000 (TRUE in pop_mask vector). The new dataframe, world_cleaned, only contains these rows.

	Country	Capital	National Plant	Population
1	USA	Washington DC	Rose	331,002,651
3	Mexico	Mexico City	Dahlia	128,932,753
5	Argentina	Buenos Aires	Cockspur Coral Tree	45,195,774
6	Japan	Tokyo	Cherry Blossom	126,476,461