Lecture 2: Text Processing and Useful R Tools

Problem Set for Lecture 2

These questions will assist in bolstering your understanding of the material in Lecture 2. Emphasis should be placed on having the correct code/output as well as communication. The deliverable should be a knitted RMarkdown document to pdf without any code running off the page. You will be able to present these in office hours as an oral assessment interview along with the problem set for Lecture 3 and Lecture 4.

***** First, make sure you have read the Lecture Material's writeup *****

1. Answer all of the following questions after running the following code:

```
> library(openintro)
> players <- mlb_players_18$name
```

   (a) If you look at the "players" variable you will notice that there is a space at the front of all of the strings. Remove this space and save the corrected variable.

   (b) Display the observations (and just the first 5 columns) for people with the last name "Smith". How many people match this description?

   (c) What proportion of the last names end in "son"?

   (d) How many name "repeats" are there (that is how many names appear more than once)? Can you determine/display which names are repeats?

   (e) Using the `any()` function, determine if any players contain a double z ("zz") within their name.

2. Answer all of the following questions after running the following code:

```
> athletes <- mlb_players_18[ , c("name", "games", "AB")]
# Note: to access certain columns use athletes$column_name
```

   (a) Display the athletes who played more than 160 games. How many people match this description?

   (b) Determine the average number of games played for people with a first name starting with the letter "M"

   (c) Display the players whose "AB" total is one more than a multiple of 7 (that is 1,8,15,22,etc.). How many people match this description?

   (d) Determine the index values for the athletes who played exactly 0,50,75,100, or 150 games

3. Answer all parts below using the built-in dataset `islands`:

   (a) Display the islands whose values are between 100 and 1000. Do this with two different methods.

   (b) Use `set.seed(42)` and then use `sample()` to randomly select 10 island names from `island_names` without replacement. Save the result as a vector named `pick`

   (c) Use a compound logical statement to display the island names in `pick` that start with a "T"/"H" or contains an area less than 15

   (d) Remove the first letter in each island name and then save the result as `pick_edit`

   (e) Using the `pick_edit` variable, determine the proportion of observations which now start with a vowel.