

**ANLY 6100**

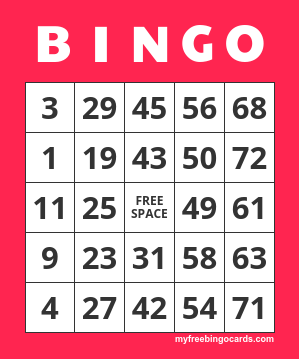
**assignment #2: 100 PTS**

This assignment is designed to get you started working with (a) some advanced functions in R, and (b) linear regression modeling. For the first part of the assignment, you will submit an R file on Canvas that I can run in its entirety. For the second part of the assignment, please submit a Word or PDF file on Canvas.

For the first part of this assignment, the “R Advanced Functions” file on Canvas contains some helpful code that could be useful. In addition, the DataCamp course “Intermediate R” is also likely to be helpful if you aren’t familiar with R or general programming of loops and functions. I will “assign” Intermediate R to you on DataCamp, but **Intermediate R is NOT due** like a regular DataCamp assignment. It is purely optional. If you complete it before the deadline for Assignment 2, you will receive 5 extra credit points for this assignment.

**Part I (50 points): Playing Bingo with R!**

If you have ever been to a student event or a retirement home in the United States, you may have played the game of bingo before. 24 numbers from 1 to 75 are chosen at random and placed in a grid under the letters B, I, N, G, and O. Bingo cards look like this:



To construct the card, 5 numbers from 1 to 15 are randomly chosen for the B column, 5 numbers from 16 to 30 are randomly chosen for the I column, 4 numbers from 31 to 45 are randomly chosen for the N column, 5 numbers from 46 to 60 are randomly chosen for the G column, and 5 numbers from 61 to 75 are randomly chosen for the O column. Then a caller selects numbers one at a time (for example, “G-47”), and if the number appears on the card, a player marks it off.

If you mark all the numbers on your card, you have achieved a blackout (also called a coverall) bingo. For this exercise in R, we want to program R to play a game of bingo and see how long it takes for a player to achieve a blackout bingo by covering all the numbers on their card.

The rules of the game are:

1. A player must receive **one valid, randomly-generated** bingo card (in other words, pay attention to the column values!)
2. The player then plays one game of bingo
   1. If the player achieves a blackout in 50 numbers or fewer, the player receives $100
   2. If the player achieves a blackout in 55 numbers or fewer, the player receives $75
   3. If the player achieves a blackout in 60 numbers or fewer, the player receives $50
   4. If the player achieves a blackout in 65 numbers or fewer, the player receives $25
   5. If the player achieves a blackout in 70 numbers or fewer, the player receives $10
   6. If the player takes 71 numbers or more to achieve their blackout, the player loses $10

The basic outline of the game with some tips is already created for you, in Bingo\_Outline.R (in the Assignment Files folder on Canvas). Your task is to complete the file to play a full game of Bingo, then find the expected return of playing 100 Bingo games. Note that you are not required to complete the outline as written, if you have a different/better way to do it! However, at minimum your code should include:

* A function
* A loop
* A seed set to 24646 at the beginning of your code file

Make sure your code is able to be run completely in R from start to finish. We will be running your file to get your answer! You should be able to submit your R simulation file through Canvas once it’s complete.

**Part II (50 points): Introduction to/Review of Linear Regression**

One of the most famous (and slightly morbid) data sets out there is the collection of data from passengers on the Titanic, including whether or not a given passenger survived. While this may seem a little dark, this is actually the goal of the data. Because the Titanic sank in the middle of the Atlantic Ocean, many bodies of people who didn’t survive were never recovered, and because the Titanic sailed in 1912, many records even for survivors are incomplete today. To address this, researchers collected data on passengers of the Titanic who were known to have survived or not survived, as well as characteristics of those passengers’ trip. Using this data, researchers would be better able to predict if someone had/had not survived based on their characteristics. If you’ve seen this data before, it’s probably because now the data set is famous as an introduction to data analytics!

The data in the accompanying file **Titanic\_sample.csv** (posted on Canvas) is publicly available online, e.g. as part of the beginner Kaggle challenge [here](https://www.kaggle.com/c/titanic). The variables in the data set are:

1. PassengerID: a unique ID number to identify the passengers in the data set
2. Survived: Did the passenger survive the sinking of the Titanic? (0 = No; 1 = Yes)
3. Class: Passenger Class (X, Y, or Z)
4. Sex: Gender (female or male)
5. Age
6. SibSp: The number of siblings/spouses the passenger was traveling with
7. Parch: The number of parents/children the passenger was traveling with
8. Fare: The amount the passenger paid for his/her ticket (in pounds)
9. Embarked: The port where the passenger boarded the Titanic (C = Cherbourg; Q = Queenstown; S = Southampton)

There are two goals. The first is to predict the variable Fare as a function of the other variables. The second is to determine the effect of survival on the fare paid.

**Assignment**

Please answer all questions in the dedicated space and upload on Canvas. Please ensure that your numbering of questions matches those below. Include any R code you used to answer each question with your response. If you are asked to include output, you’re welcome to put it at the end of your assignment in an appendix, but make sure it matches both your code and your answers. Remember: you are allowed to work with others in the class on this assignment, but don’t forget to include their names in the last question!

1. **(5 points) Before the regression:** Conventional wisdom says we should consider removing a variable from a data set if it has 40% or more missing values. Do we need to consider removing any of the variables in this data set? Explain.  
     
   No, none of our columns have 40% or more missing values.

sapply(titanic,function(x) mean(is.na(x)))

1. **(25 points) Basic regression in R**: Start by running a multiple regression using Fare as your Y variable and all other variables as X variables. Be sure to include the output of your model (using **summary()**) with your assignment.
   1. Do you consider this a good model? Why or why not?
   2. If you were to remove a variable from this model, what variable would you suggest removing? Why?
   3. According to your model, what is the effect on the fare a passenger paid when they survived vs. not survived? (Make sure to give this effect in terms of the problem!)
   4. Does your answer to part (c) make sense to you? Explain.
   5. One of the researchers on the Titanic team knows the passengers were actually classified as first class (the highest class), second class, and third class (the lowest class) rather than using letters like X, Y, and Z. Based on your regression output, which letter likely corresponds to first class? Explain.
2. **(10 points) More regression in R**: Run another linear regression for Fare using the variables Survived, Class, Sex, Age, SibSp, and Parch. Be sure to include the output of your model (using **summary()**) with your assignment.
   1. What is the resulting R2? Is this better or worse than your model in Question (1)? Explain your answer.
   2. What is the predicted Fare for a 60-year-old female passenger who survived, was traveling with 2 siblings/spouses, and was traveling with 2 parents/children? (In other words, Survived = 1, Sex = female, Class = X, Age = 60, SibSp = 2, Parch = 2)
3. **(10 points) Even more regression:** One of the senior Titanic researchers includes an interaction term in the model from Question 2 between Sex and Age.
   1. What is this researcher saying about the effects of Sex and Age on Fare?
   2. Add this interaction to your model from Question 2. Do you think the researcher was correct to include the interaction? Explain.