# APANPS5335 3: Machine Learning Concept & Applications, Columbia University

*Professor Siddhartha Dalal*

*Friday September 6th, 2019*

**Due : Wednesday, September 11th 2019 before 02:00PM EST (total points: 100)**

**Directions: Please submit the RMD and "knitted" PDF files, as well as the UNI_session_info.txt and background_info.RDS files on the Canvas class website**

In this class we will be using R to implement the algorithms that we cover. The goal of this assignment is to ensure that you are proficient in basic data manipulation and exploratory data analysis using R.

If you need to review R basics, here are a few suggested resources:

- R Basics Class: https://campus.datacamp.com/courses/free-introduction-to-r/
- R Cheat Sheets: https://www.rstudio.com/resources/cheatsheets/
- Data Mining with R by Luis Torgo
- R in a Nutshell by Joseph Adler
- **RMarkdown:** https://rmarkdown.rstudio.com/lesson-15.html

Most questions have been asked before, and stackoverflow is an invaluable resource:

- https://stats.stackexchange.com/?tags= and https://stackoverflow.com/questions/tagged/r

Finally, for a deeper understanding of R (and its eccentricities), these two books are good:

- Advanced R by Hadley Wickham http://adv-r.had.co.nz/
- The R Inferno http://www.burns-stat.com/pages/Tutor/R_inferno.pdf

### 0. Environment setup (1 point)

Install R and RStudio, and the Rmarkdown, ggplot2, tidyverse, and ISLR packages

For background on R package organization, see this article https://www.datacamp.com/community/tutorials/r-packages-guide

In the command window, type sessioninfo::session_info() and hit enter

Save the console output of session_info() to a text file called UNI_session_info.txt (where UNI is your University Network ID) attach this with the answers to the homework.

### 1. About You (4 points)

We would like to get an idea of your background, interests, and experience in Machine Learning.

(3 points) a. Construct a dataframe called background_info with exactly one row and columns with the following names:

```
c("UNI", "first_name", "last_name", "preferred_name", "highest_degree", "undergrad_major",
"graduation_year", "alma_mater", "calculus_level", "linear_algebra_level", "probability_level",
```

`"regression_level","r_level", "fave_hobby", "fave_book", "fave_musician", "fave_movie", "fave_brand", "fave_nyc_restaurant")`

- UNI
- First Name
- Last Name
- The name you prefer to be called
- Highest degree program (B.A., M.A., E.D., Ph.D., etc)?
- Undergrad major
- Undergrad graduation year
- Undergrad institution
- Mathematical proficiency (Calculus) **Scale 1(Poor) - 5(Excellent)**
- Mathematical proficiency (Linear Algebra) **Scale 1(Poor) - 5(Excellent)**
- Statistical proficiency (Probability, Statistical Distributions) **Scale 1(Poor) - 5(Excellent)**
- Statistical inference proficiency (Regression, Logistic Regression) **Scale 1(Poor) - 5(Excellent)**
- R programming proficiency (Plots, Functions, Rmarkdown) **Scale 1(Poor) - 5(Excellent)**
- Favorite hobby
- Favorite book
- Favorite musician
- Favorite movie
- Favorite brand
- Favorite restaurant in NYC

b. (1 point) Verify that the dataframe that you made has one row and 19 columns. Set the rowname of that single row to be your UNI.

c. (1 point) Save your dataframe to an RDS file, named UNI____background_info.RDS (where UNI is your Columbia network ID) and include this file with your submission

**2. R basics / warm up (15 points)**

a. (2 points) Write a function called `roll_die()` that returns an integer between 1 and 6, distributed uniformly.

b. (2 points) Write a function called `roll_two_dice()` that returns a vector representing a roll of two dice.

c. (3 points) Write a function `roll_loaded_die(weighted_value, loaded_weight)` that simulates the roll of a loaded die, where loaded_weight is the probability mass that is placed on on the specified weighted value. Your function should validate that weighted_value $\in \{1, 2, 3, 4, 5, 6\}$ and that loaded_weight doesn't violate any rules about probailities.

d. (2 points) Generate a sample of 10000 rolls for each of your fair and loaded dice, using the functions that you defined above (choose any loading and weighted value you wish).

e. (1 point) Generate a sample of 10000 rolls of a fair die using only the built-in R `sample` function.

f. (1 point) Generate a sample of 10000 rolls of your loaded die using only the R `sample` function

g. (2 points) Plot histograms showing the distributions of the samples from your fair and loaded dice rolls. (Either show them on the same plot, or stacked one above the other so that they are easy to compare). Use basic R plotting functions.

**2. Kaggle DontGetKicked Dataset (40 points)**

Go to https://www.kaggle.com/c/DontGetKicked/ and read about this machine learning competition.

Download all the data for the DontGetKicked competition from https://www.kaggle.com/c/DontGetKicked/data. (The simplest way to do this is to just download the zip file of data from the website; if you wish, you can also set up the Kaggle API client.)

a. (5 points) If you don't already have a kaggle account, sign yourself up and provide your kaggle user name. Store your kaggle user name in a variable called `kaggle_user_name`.

b. (2 points) Read the DontGetKicked training data (training.csv) into a dataframe.

c. (3 points) What are the types (classes) of each column in the your dataframe? Print this out as a table. Do you notice anything strange? If so, briefly explain what is going on, and then fix it.

d. (2.5 points) Explore and briefly describe the dataframe you have loaded. What are its dimensions? Use the summary function to get a quick descriptive summary of all the variables.

e. (2.5 points) What is the range of the vehicles' year of manufacture?

f. Use ggplot2 to graph a histogram of cars by year. Make sure that your plot has a title and sensible axis labels and ticks. (5 points)

g. (2.5 points) Generate a panel of these histograms for each Make of car

h. (2.5 points) How many values are missing (NA) from each column?

i. (2.5 points) Remove rows that are missing values in any of the vehicle price variables and provide the dimensions of the resulting dataset.

j. (5 points) How many different manufacturers, models, and sizes are in the dataset?

k. (5 points) Summarize the acqusition price (VehBCost) per vehicle by fabrication year, model and color. Which model has the highest average cost?

l. (2.5 points) Which vehicle model has the greatest variability in acquisition price?

### 3. Iris Dataset (40 points)

The *iris* dataset is a popular toy dataset used to illustrate machine learning algorithms, and is provided with R.

a. (2.5 points) How many observations are in the iris dataframe? What are the variables, and what do they represent?

b. (5 points) Using the graphics package of your choice, generate a grid of scatterplots comparing each variable against the others, and color the points by species. Describe what you observe.

c. (2.5 points) Divide the iris data into a testing and training set (use 120 observations in your training set).

d. (15 points) Train a linear regression model to predict Sepal.Width from Sepal.Length (and other variables, if you wish), using your training dataset. Based on the summary statistics from the model, briefly discuss whether you think your model is a good one (and if not, improve it). Apply your model to predict Sepal.Width on the observations in your test set, and make a plot that summarizes the accuracy of your predictions.

e. (15 points) Train a K-nearest neighbours model to classify the Species using the quantitative variables (one of the easiest ways to do this is with the caret package). Summarize the performance of your KNN model (on the test data) with a confusion matrix.

For additional guidance on this exercise, you may work through
https://www.datacamp.com/community/tutorials/machine-learning-in-r