Homework 8

For this homework you will create an R Markdown document that outputs to a PDF and you will upload both the .Rmd and .pdf files to wolfware. Be sure to put your name in the title of the document.

The purpose of this homework is to get practice with fitting and predicting with multiple linear regression models and, similarly, with logistic regression models.

Application: We will use a dataset from the UCI Machine Learning Repository. This data set is about bike sharing rentals and is available at the assignment link. You can learn more about the data here.

The data description describes the following variables:

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

We'll ignore the `Date` variable and not worry about trends from day to day. Hopefully, the `Seasons`, `Holiday`, and `Funcitonal Day` variables will allow us to account for trends over time.

To Do:

Create a document that goes through your process of reading the data, basic EDA, manipulating/creating any variables, and fitting and choosing a final MLR and logistic regression model.

Reading Data

- First read in the data
- When using `readr::read_csv()` I got an error `Error in nchar(x, "width") : invalid multibyte string, element 1`
- Google this and it is a quick fix!

Split the Data

- Split the data into a training and test set (75/25 split)
- We'll do our EDA and model fitting on the training set and see how the 'best' model does using the test set

Basic EDA

- Go through a quick EDA on the data (again, feel free to ignore the Date column)

- You may want to rename the variables (not required)
- Your focus should be on the response variable we'll use: `Rented Bike Count`
- We'll also create a binary version of that variable for use with logistic regression. Create a new variable that is `1` if the number of bikes rented is greater than or equal to 700 and `0` otherwise

Fitting MLR Models

Fit at least five different MLR models using the training data. Compare their performance on the training set using 5-fold cross-validation and RMSE as your metric.

- Normally, you might use some scientific reasoning/subject matter expertise to know what candidate models to consider.
- As we don't necessarily have that, you can figure out candidate models however you'd like. Perhaps:
  - Include some interaction terms
  - Include some quadratic terms
  - Use something like best subset selection (this would count as one model even though it fits multiple models)
- When including interactions and polynomial terms, we often standardize the variables (center and scale each observation), but that is up to you. When just exploring models you might want to standardize anyway (or maybe do more - here is an interesting article - Gelmen is a very well known statistician).
- Once you have your best model as selected by 5-fold CV, use that model to predict on the test set and evaluate the performance in terms of RMSE.

Fitting Logistic Regression Models

- Repeat the above using logistic regression models instead. Use accuracy as your performance metric.