AD699: Data Mining for Business Analytics
Individual Assignment #2
Summer 2020

You will submit two files via Blackboard:

(1) **Your write-up**. This should be a PDF that includes your written answers to any questions that ask for written answers, along with the other things asked for in the prompt.
(2) **Your R Script**. This is the script that you will use to write your assignment. If you use Markdown, you'll submit an .RMD rather than a .R file.

As always, remember to take advantage of your available resources: We'll have five live Q&A sessions next week, in addition to unlimited opportunities to schedule a Zoom session on any other day or time. ***For this assignment in particular, the video library can be quite helpful***. As the course slogan says, "Get After It!"

A dataset description file will be posted to Blackboard along with the dataset and prompt. **That document explains what all the variables mean.**

For each step, your write-up should clearly display your code and your results. For any step in the prompt that includes a question, the question should be answered in written sentences.

Main Topics: Simple Linear Regression & Multiple Linear Regression

Tasks:

- **Simple Linear Regression**:

1. For this assignment, we will use the dataset *Carseats*, which comes from the ISLR package. After you have installed ISLR, and used the library() function to bring this

package into your environment, you can bring this dataset into your environment with the name carseats, in the following way:

```
> carseats <- data(Carseats)
```

In this section, we will explore the relationship between the carseat sales at particular sites and the price charged by retailers at those sites.

2. Let's explore the relationship between these variables in a visual way. Using ggplot, create a scatterplot that depicts the Sales variable on the y-axis and the Price variable on the x-axis. Add a best-fit line to this scatterplot.

   What does this plot suggest about the relationship between these variables? Does this make intuitive sense to you? Why or why not?

3. Now, find the correlation between these variables. Then, use cor.test() to see whether this correlation is significant.

   What is this correlation? Is it a strong one? Is the correlation significant?

4. Using your assigned seed value, create a data partition. Assign approximately 60% of the records to your training set, and the other 40% to your validation set. Keep in mind that a seed value has no relationship to the data itself -- it's just an arbitrary number.

5. Using your training set, create a simple linear regression model, with *Sales* as your outcome variable and *Price* as your input variable. Include a screenshot of the summary of your model.

6. If your r-squared value seems low, do not assume that you did something wrong, or that there is something wrong with the assignment. Instead, think about what r-squared means. Why might more variables lead to a higher r-squared?

7. What is the regression equation generated by your model? Make up a hypothetical input value and explain what it would predict as an outcome. To show the predicted outcome value, you can either use a function in R, or just explain what the predicted outcome would be, based on the regression equation and some simple math.

8. Using the accuracy() function from the forecast package, assess the accuracy of your model against both the training set and the validation set. What do you notice about these results? Describe your findings in a couple of sentences.

- **Multiple Linear Regression**:

For this part of the assignment, use the same training set and the same validation set that you used in Part I.

1. Build a correlation table in R that depicts the correlations among all of the numerical variables that you might use as predictors (use your training set to build this). Are there any variable relationships that suggest that multicollinearity could be an issue here? If so, for any strongly correlated variable pair, remove any variables that should be taken out of the model. If you removed any, how did you decide which ones to remove? If not, why did you keep the ones that you have left?

2. What are dummy variables? In a couple of sentences, describe what they are and explain their purpose. (We won't create dummy variables here, because they'll be automatically generated in R when we call the lm() function).

3. Using backward elimination, build a multiple regression model with the data in your training set, with the goal of predicting the *Sales* variable. Start with all of the potential predictors that you have left (if you eliminated any in Step 1, don't bring them back...they're gone!)

4. Based in part on what was recommended by the backward elimination process, and in part on your judgement, which variables will you keep? (Note: This is not a trick question. Part of making a multiple linear regression model involves subjective judgement). No R code is required for this step.

4b. Using the variables that you will keep, build a multiple linear regression model. Show a summary of your multiple regression model.

5a. What is the total sum of squares for your model? (SST). This can be found by summing all of the squared differences from the mean for your outcome variable.

5b. What is the total sum of squares due to regression for your model? (SSR). This can be found by summing all the squared differences between the fitted values and the mean for your outcome variable.

5c. What is your SSR / SST? Where can you also see this value in the summary of your regression model?

6. Getting from a t-value to a p-value. Choose one of the predictors from your model. What is the t-value for that predictor? Using the visualize.t() function from the visualize package, create a plot of the t-distribution that shows the distribution for that t-value and the number of degrees of freedom in your model. What percent of the curve is shaded? How does this relate to the p-value for that predictor?

7. Make up a fictional retail location, and assign attributes to it for each of the predictors in your model. What does your model predict that this location's sales will be? To answer this, you can use a function in R or just explain it using the equation and some simple math.

8. Using the accuracy() function from the forecast package, assess the accuracy of your model against both the training set and the validation set. What do you notice about these results? Describe your findings in a couple of sentences. In this section, you should talk about the overfitting risk and also about the way your MLR model differed from your SLR model in terms of accuracy.