AD699: Data Mining for Business Analytics
Individual Assignment #2
Spring 2021

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 four 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!"

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.

This model will be used to predict the AVG_SALARY, per year, of a National Basketball Association (NBA) player's contract. **This assignment will not require any specific domain knowledge from outside of the dataset description, dataset, and prompt**.

Main Topics: Simple Linear Regression & Multiple Linear Regression

Tasks:

- **Simple Linear Regression**:

For this assignment, we will use the dataset *nba_contracts*.csv, which can be found on our class Blackboard page.

Start by downloading this dataset.

1. Read the dataset into your environment in R.

2. Create a new variable called *ppg*. This new variable, which stands for "points per game" will be created by dividing points by games played.

3. Let's explore the relationship between points per game and average salary. Using ggplot, create a scatterplot that depicts *average salary* on the y-axis and *ppg* 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?

4. 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?

5. 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.

6. Using your training set, create a simple linear regression model, with *AVG_SALARY* as your outcome variable and *ppg* as your input variable. Use the summary() function to display the results of your model.

7. What are the minimum and maximum residual values in this model?

   a. Find the player whose salary generated the highest residual value. What was his actual salary? What did the model predict that it would be? How is the residual calculated from the two numbers that you just found?
   b. Find the player whose salary generated the lowest residual value. What was his actual salary? What did the model predict that it would be? How is the residual calculated from the two numbers that you just found?
   c. It might be unfair to say that the person in 7a is overpaid, or that the person in 7b is underpaid. Why might it be unfair to say this? (Note: You do *not* need to be a basketball fan, or to know about the NBA, in order to answer this). However, you should look at the dataset and the data description, and give this just a bit of thought before answering). You can answer this question in 2-3 sentences.

8. 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.

9. Using the accuracy() function from the forecast package, assess the accuracy of your model against both the training set and the validation set. For this answer, focus on the differences between the training and validation sets. To assess the model, focus mainly on RMSE and MAE.

10. How does your model's RMSE compare to the standard deviation of average salary in the dataset? What can such a comparison teach us about the model?

- **<u>Multiple Linear Regression</u>**:

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

1. It is certainly possible to use categorical inputs when doing linear regression, but we won't do it here. We will not use 'NAME' as an input variable in this model. Why won't NAME be a useful predictor here? To answer this, you can view the dataframe by scrolling through it, or you could use the table() function. You don't necessarily need to use any R code to answer this question.

2. Using the exhaustive subsets method shown in the book, build a multiple regression model with the data in your training set, with the goal of predicting the *Average Salary* variable. Start with ***all*** of the numeric predictors in the dataset.
    a. Using the model inputs with the highest r-squared among the options shown, generate a multiple linear regression model.

3. Call the vif() function from the car package on your model. Are there any values here for your numeric predictors that are greater than 5?

4. Now let's see where VIF really comes from. Choose any single one of the VIF values you found in the previous step, and show how VIF is determined by applying the formula for VIF to this variable (this is something that we will cover in class on 02MAR).
    a. When a variable has a very high VIF value, why might it not be very useful in a multiple regression model?

5. Next, build a correlation table that shows the correlations among all the input variables in your model, as well as the outcome variable.
    a. Now, let's examine the source(s) of trouble. Do you see any correlations *among input variables* that are higher than .70?
    b. For any highly-correlated variable pair, remove just one of the correlated variables, and briefly explain why you chose to remove the one that you took out.

6. Build a new model with the variables that are still remaining. Show the summary of this model.

7. Call the vif() function now on this new version of your model. Are all the values less than 5? If so, move to the next step. If not, create a new correlation table and remove any inputs that might be causing trouble here.

8. 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.

9. 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.** (Do not use any other definitions for SSR here -- use the one shown in bold in this step).

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

11. 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?

12. What is your model's F-Statistic? What does the F-Statistic measure? Using the formula shown in class, show how the F-Statistic for this model comes from *TSS, RSS, n, and p*.

13. Make up a fictional basketball player. What is his name? For each variable in the model, assign a value to your player (give him values that are within the dataset range). What does your model predict that this player's salary will be? To answer this, you can use a function in R or just explain it using the equation and some simple math.

14. 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 also talk about the way your MLR model differed from your SLR model in terms of accuracy.