**Introduction to Data Science**

**Data Science for All Capstone Project Proposal**

(Three Potential Capstone Project Ideas)

**July 24, 2019**

**Submitted by:**

**Tarji R. Kinsey**

**Clark Atlanta University**

**(404) 880-8082**

**tkinsey@cau.edu**

**Approval:**

**Date:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_

**(Instructor’s Signature)**

*Note: Proposed project idea will aid in showcasing my existing and newly acquired data science skills via* ***Data Cleaning****,* ***Machine Language,*** *and* ***Exploratory Analysi****s.*

***Proposal Idea #1:*** Using one messy dataset, I will use various techniques/resources to clean, merge and explore institutional data to predict undergraduate and graduate enrollment for the Clark Atlanta University School of Business for the upcoming Fall 2019 term.

***Problem:*** Clark Atlanta University School of Business would like to have an idea of what to expect re: undergraduate and graduate enrollment for the upcoming Fall 2019 term.

***Dataset:***

* Clark Atlanta University Enrollment Data Set

***Data Wrangling Steps:*** Several data wrangling techniques were performed in R on the data frame (enroll) to transform it into a format that can be analyzed. The techniques used, but were not limited to were: Basic Data Manipulation and Dealing with Missing Values. These techniques allowed me to remove or add variables and columns, change column names, modify blank or null data fields, and include functions for term hours.

**Remove or Add Variables and Columns:**

The original data frame included 56 variables. Some of the variables were not necessary to perform enrollment prediction techniques as they did not provide relevant information for the projection model. Removing 33 variables reduced the data frame to 23 variables. Some of the variables needed partnering variables to transition easily into the model. Two new columns were added to captures partnering variables.

**Changing Columns Names:** There are five variables included in the data frame that underwent a name change. When the data was imported into R Studio, some variables names were too long, and others did not accurately describe the particular data. Therefore, the names of the variables were changed to shorter, concise names that accurately describe the data in the column.

**Add Values to Blank or “NULL” fields:** Several columns displayed data that was either blank or the word “NULL” was found. A treatment column and a column that displayed standardized variables for the final outcome was created. In the treatment column, an if-else statement was written to place a 1 for the observation that was treated and a 0 otherwise. This information helped to accurately and easily determine the control and treatment groups.

**Data Function:** One column contained a function written to normalize the data for student term hours. Adding this column increased the data frame to 23 variables. It represents the students load status (FT or PT). Values assigned to this new column are FT = Full time Student and PT = Part time Student.