**DAEN 500- DL1 – Data Analytics Fundamentals**

**Fall 2020 Final Examination Exercise**

**11/24 – 12/05/2020**

**Final Submission Deadline: NLT 11:59PM (EST). Saturday, Dec 5, 2010**

***Failure to submit ON TIME will result in DAEN COURSE FAILURE***

**Name: Brady Wilkinson GMU G# G01014006**

**Student Signature (Honor Certification): Brady Wilkinson**

This exam is **OPEN BOOK/OPEN NOTES**. You may consult any of the course texts, and the various reference materials recommended in the syllabus. ***The exam of course IS NOT “Open Web”,*** especially in that you may NOT utilize expert “help” sites such as Stack Overflow, or other programming help or collaboration sites. 

Additionally, you are restricted from discussing the substance of the questions on this exam with any other individual, until after you have submitted your final response for grading. The completed exam -- with your answers embedded in this docx document (add extra pages as necessary) should be submitted following instructions contained in the Final Exam Instructions BB site. If you have any trouble submitting and have extra parts of the answers you have trouble appending to this document, you may simply submit additional pages separately (the exam submission site is set for multiple submissions, just in case). Make certain all are submitted PRIOR TO THE DEADLINE!

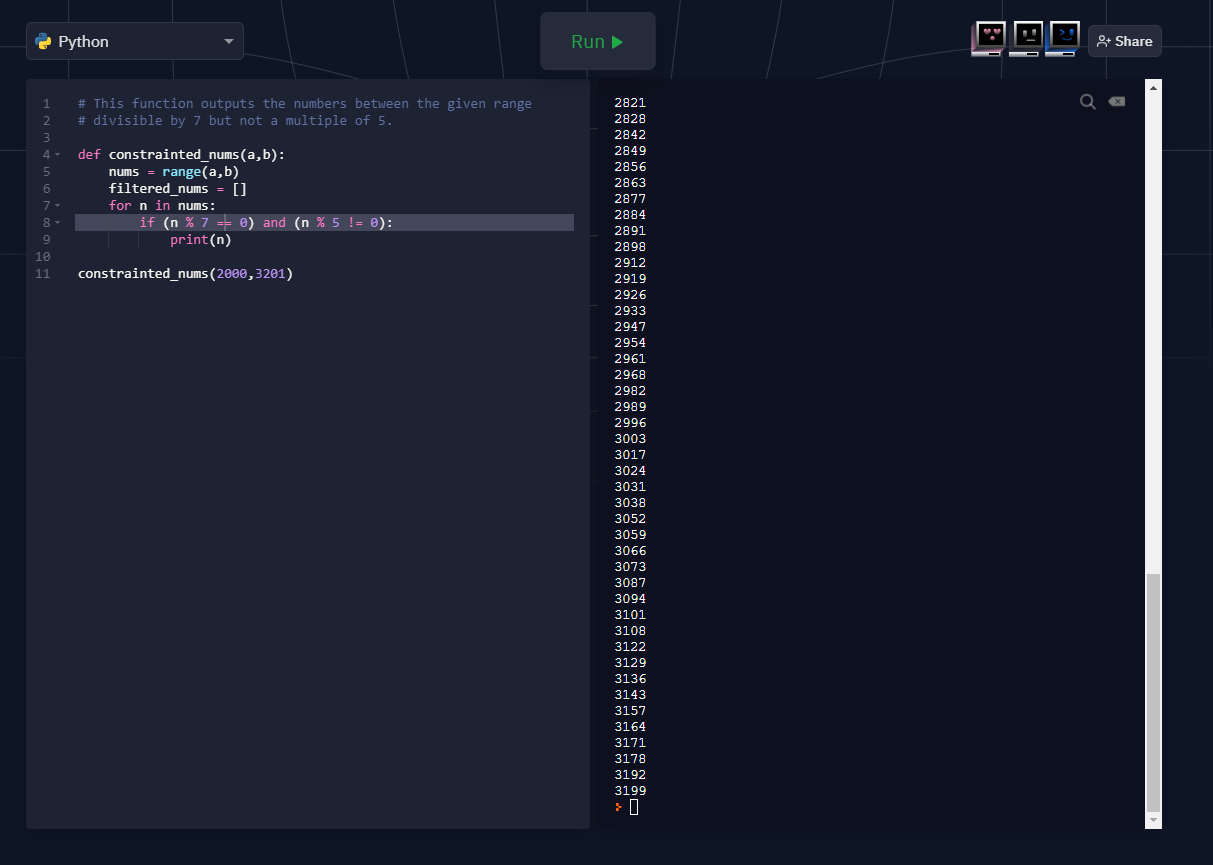
**FINAL EXAM PROBLEMS**

**COMPLETE ALL & INSERT ANSWERS BELOW QUESTIONS**

# Problem 1: Python Programming Problem (15 Points Total)

* **Design and implement a Python program that is based on the following requirements: a) program will find all numbers which are divisible by 7 but are not a multiple of 5; and b) numbers between 2000 and 3200.**

**notepad**



# Problem 2: Python Programming Problem

# (15 Points Total)

* **Design and implement a Python program that is based on the following requirements:**

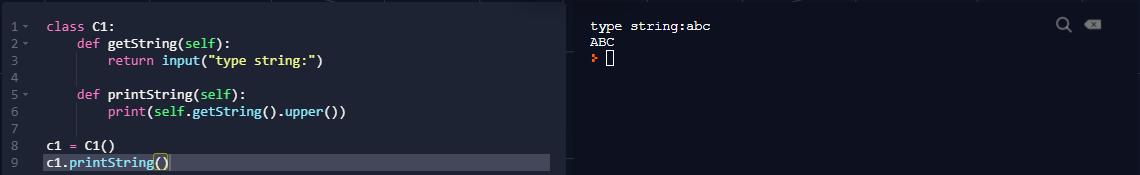
**a) define a class which has *at least two* methods**

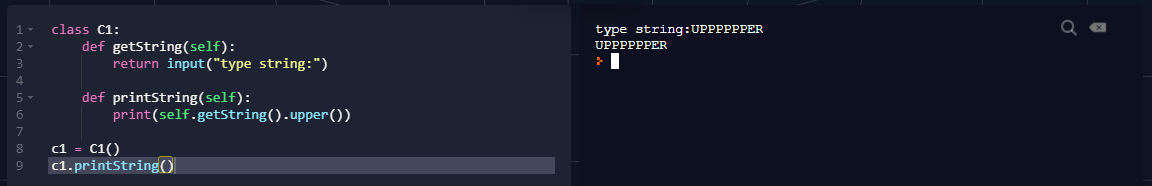
* + **Method 1 – getString: to get a string from console input; and,**
  + **Method 2 - printString: to print the string in upper case.**

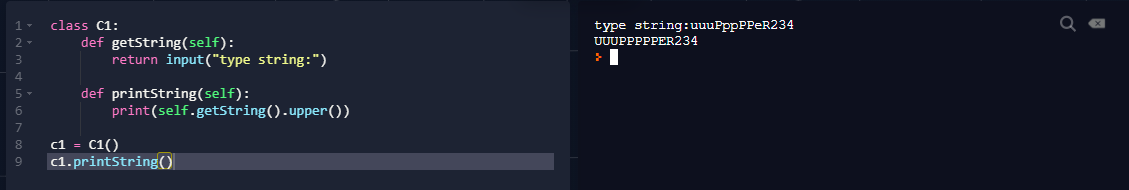
**b) demonstrate code works using three different test input strings**

* ***INSERT* *code below* and *INSERT* a screen shot of the program and successfully run output that *includes test input for input strings (test strings must include (a) all upper case, (b) all lower case, and (c) mix of upper and lower case).***

notepad









# Problem 3: R Programming Problem

# (20 Points Total)

* **Perform the following problems using R:**
  + Create a vector of courses (e.g., MATH 101) you have taken previously. Make sure you have at least 8 courses. Name the vector myCourses
  + Get the length of the vector myCourses
  + Get the first two courses from myCourses
  + Get the 3rd and 4th courses from myCourses
  + Sort myCourses using a method
  + Sort myCourse in the reverse direction

my.courses <- c('CMDA2005', 'CMDA2006', 'CMDA3634', 'CMDA4653', 'MATH4445', 'GER4504', 'GER4105', 'GER4106')

length(my.courses)

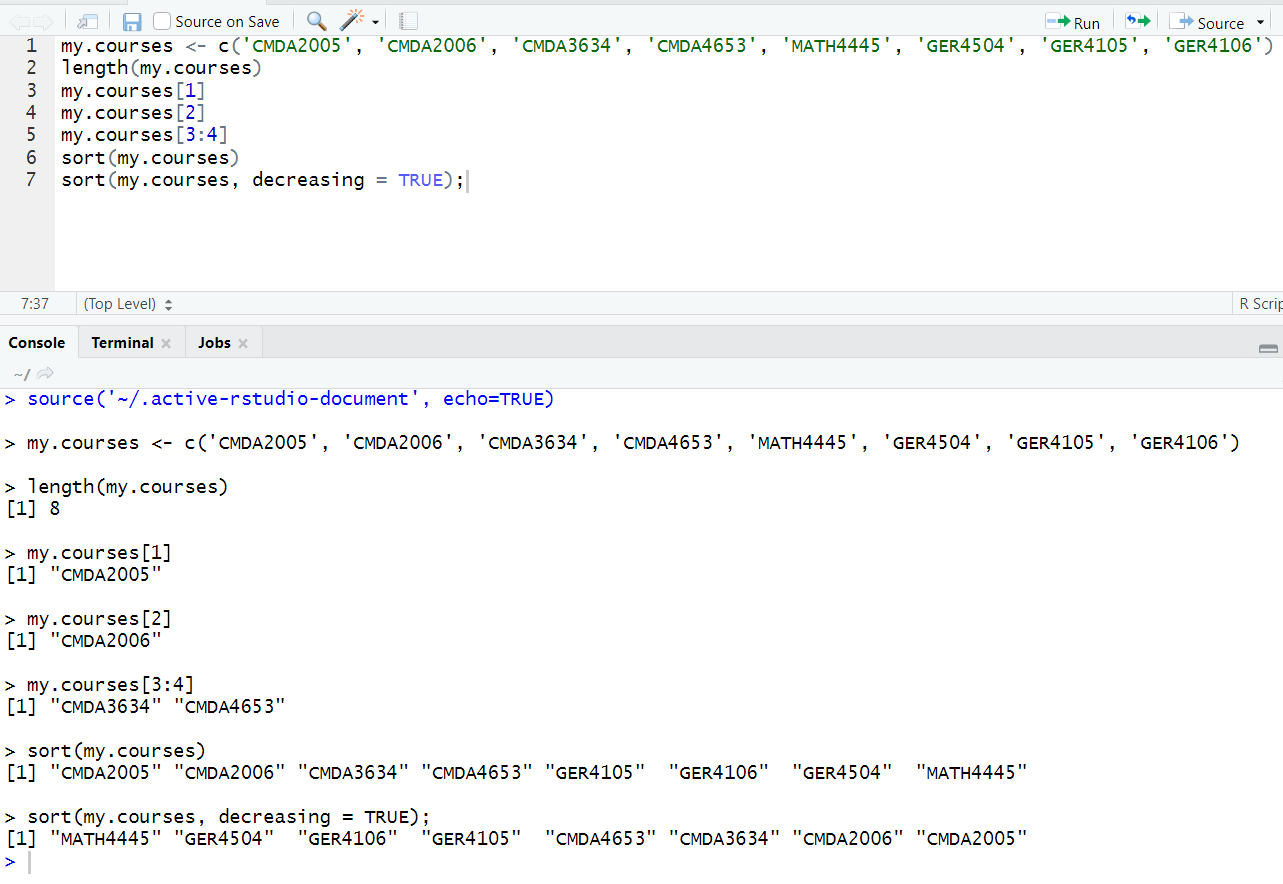
my.courses[1]

my.courses[2]

my.courses[3:4]

sort(my.courses)

sort(my.courses, decreasing = TRUE)





# Problem 4: Principal Component Analysis

# (25 points)

**Provide a description of the following:**

1. What is a component – Provide a description (5 points)
2. Principal Component Analysis – Provide a description.(5 points)
3. **Provide an specific example of Principal Component Analysis(15 points)**
4. **A component is a distinct piece of information that adds unique value. In the context of PCA, a component is a variable, or combination of variables.**
5. **PCA aims to reduce the amount of variables in a dataset while still preserving the key dimensions of information. Combining variables that are linear combinations of each other reduces the number of dimensions while retaining the principal component. This can be useful in creating a machine learning model, as fewer variables to analyze can be simpler.**
6. **zyBooks provides an example given a dataset with academic success as the response and GPA, test score, clubs, and volunteering as predictors. Although GPA and test score are two supposedly different fields, they highly correlate with each other, meaning they share some principal component. Creating a principal component out of these two features may yield in a successful reduction, however there are other fields that correlate greater: clubs and volunteering. This potential principal component must be considered first, as they correlate more.**

# Problem 5: Multiple vs. Logistic

# (30 points)

# Describe: What is difference between Multiple Regression and Logistic Regression? What circumstances might determine which to use? (10 points)

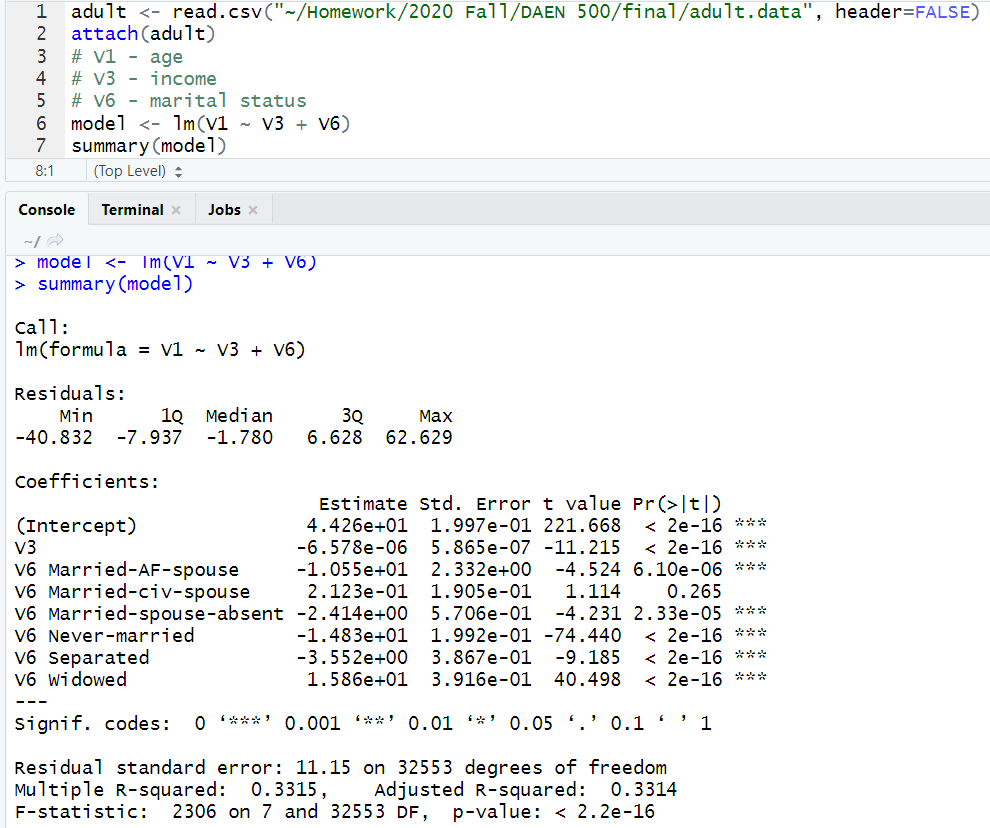
# Demonstrate: Using any data, and any tool set you’ve learned about, show differences (20 points)

# SUGGESTION: may be solved using RapidMiner, or other toolsets, BOTH TO ANALYZE AND TO VISUALIZE REGRESSION DIFFERENCES.

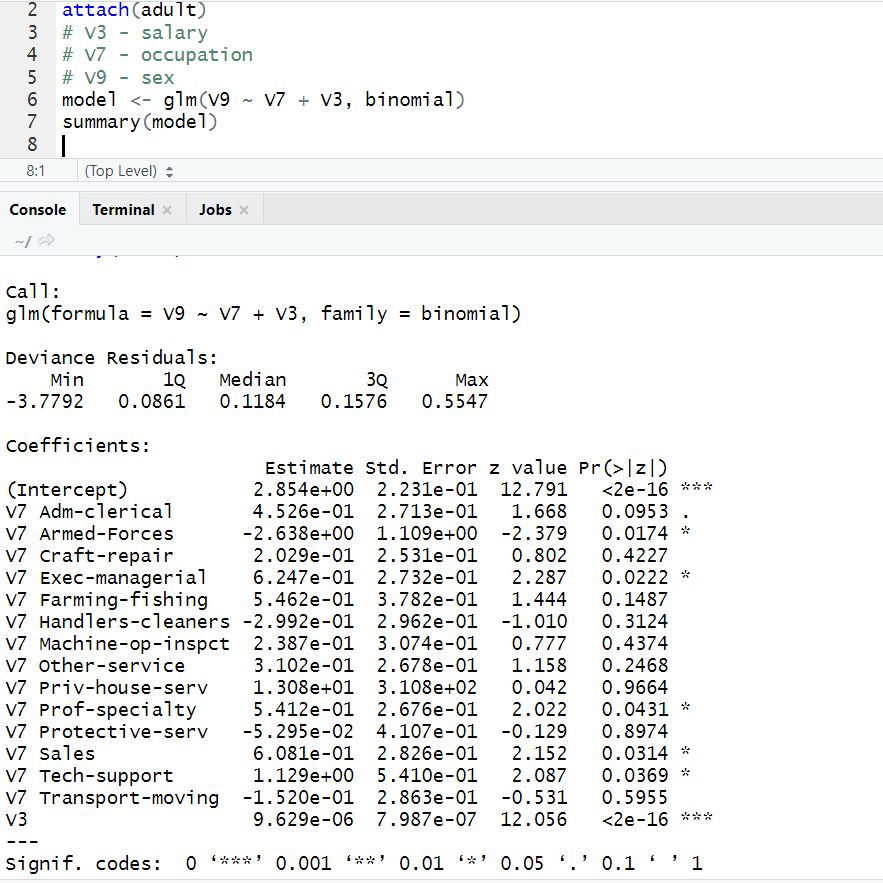
**Step 1: Perform a quick search of the** [**UCIS public data archive**](https://archive.ics.uci.edu/)**, a well-curated site which you already have seen as part of your introductory RapidMiner training.**

**Step 2: Pick a dataset you find interesting, input dataset into regression tools you’ve chosen.**

**Step 3: Run regression, .and use visualizations to demonstrate the conceptual answers you provided for 5.(a).**

1. Multiple regression and logistic regression differ in several ways, one of them being the difference in output variable types. When one is looking at a binary output (yes or no, good or bad, car or not a car, etc.), logistic regression is preferred, as the output will be a number between 0 and 1 correlating to the probability of the preferred outcome. Multiple regression on the other hand observes a continuous (numerical) output, where the output may be some dollar amount, grams of something, etc. When modeling an outcome with this type of variable, a multiple regression is desired.
2. 

above is a screenshot from a multiple linear regression model with predictors of marital status and income, as well as an outcome of age.



above is a logistic regression with an outcome of sex and predictors of salary and occupation.