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:\_\_\_\_\_\_\_\_\_\_\_\_\_Miesha Purcell\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ GMU G#01287718\_\_\_\_\_\_\_\_\_

Student Signature (Honor Certification): \_\_\_Miesha L. Purcell\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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.

HONOR CODE CERTIFICATION

**Your signature above declares that you have followed the conditions of this exam, and that the work is yours alone**. **Specifically:**

This must be your own work, authored and completed by you. As stated earlier, this is an “open source exam” – allowing books, notes or courseware, as well as *general* expert advice gained PRIOR to exam. YOU MAY NOT, HOWEVER, SEED OR USE ANY ADVICE ON HOW TO SOLVE THE QUESTION OR ANY CODE WRITTEN BY ANY OTHER INDIVIDUAL. *Any violation will result in an immediate failure in the exam and for the course, as well as referral to the GMU Honor Committee for determination of any other appropriate disciplinary consequences.*

*NOTE: Your* ***submission*** *of any responses, files, programs, etc. in response to the DAEN500 final exam instructions, will also be your personal certification of your full compliance with the spirit and letter of the* ***GMU Honor Code*** *standards for take home and/or in-class exams.*

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.**
* **INSERT (cut&paste) your Python code in space below and *then insert a screen shot in space below, showing code, your successful run, input and output.***

NOTE of alternative for help: To help test your code, you also may use a Python “programming window” found in the. **Zybooks Section 35 Additional Material**.

# Miesha Purcell

# Final Exam

# Question 1

print('Numbers between 2000 and 3200 disvisible by 7, but are not a multiple of 5:')

#Counter used to count number of outputs on single line.

newLineCnt = 0

#Loops through numbers 2000 through 3201 finding values

#divisible by 7 and not a multiple of 5.

for i in range(2000, 3201):

#Check i to see if divisible by 7 and not 5

if (i % 7 == 0) & (i % 5 != 0):

#if row contains 14 numbers, print 15th number and create new line

if newLineCnt < 14:

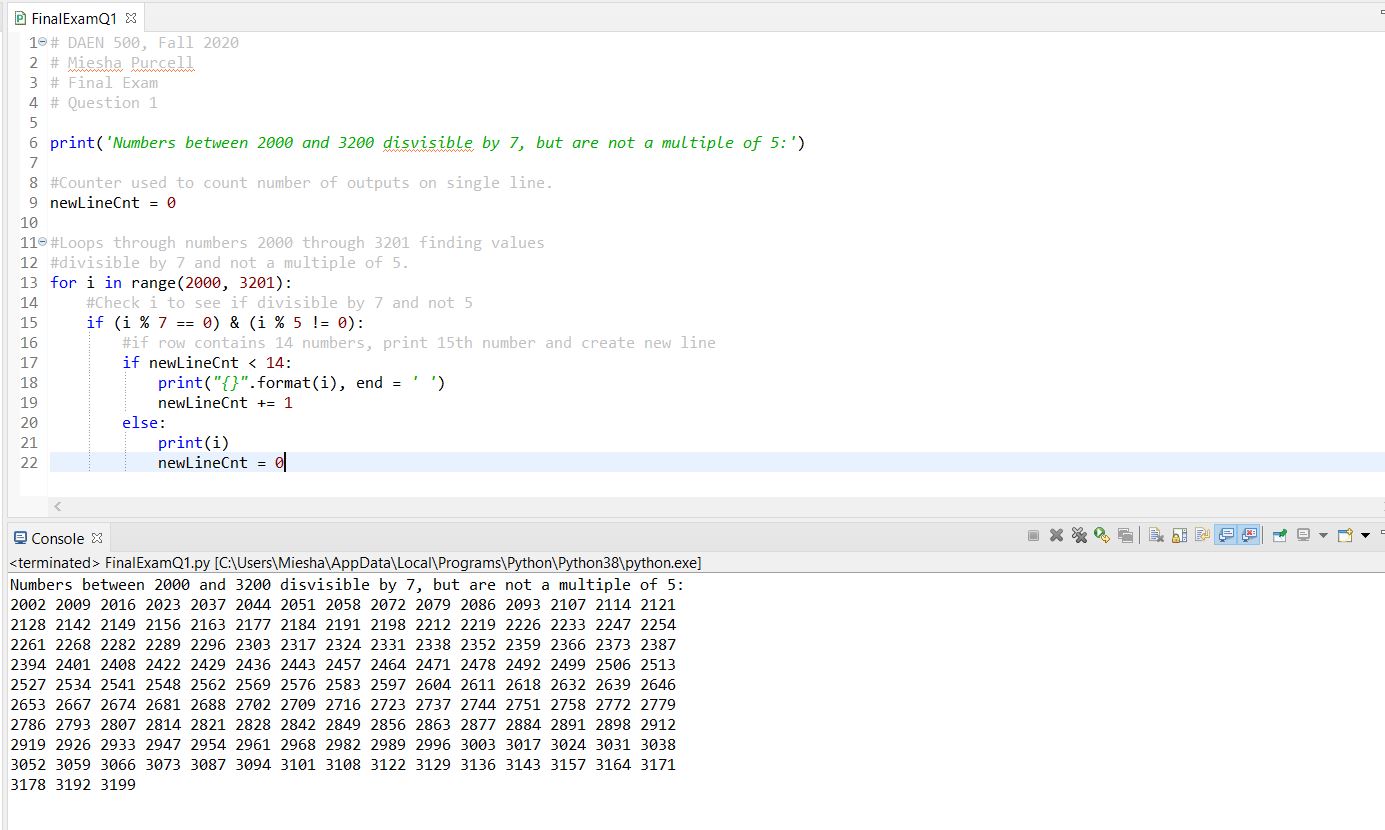
print("{}".format(i), end = ' ')

newLineCnt += 1

else:

print(i)

newLineCnt = 0



# 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)***

class StringWork():

def \_\_init\_\_(self, msg):

self.msg = msg

#self.print\_message()

def getString(self, userString=""):

userString = input("Please enter string (type end to stop):")

self.printString(userString)

def printString(self, userStr):

if(userStr != "end"):

print(userStr.upper())

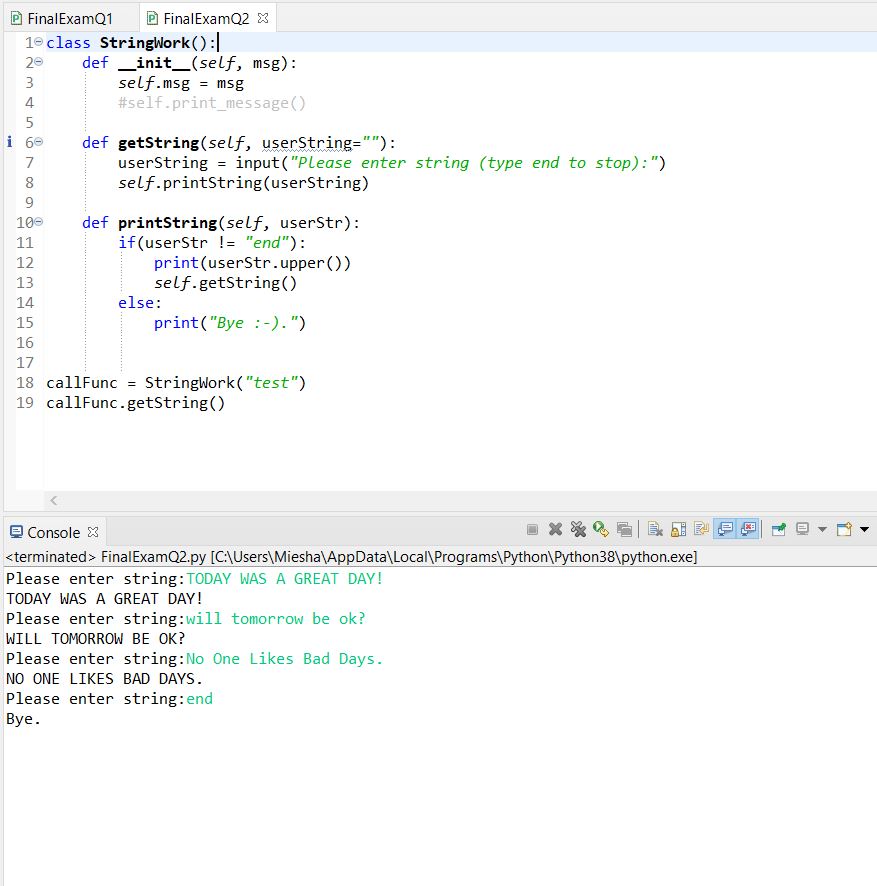
self.getString()

else:

print("Bye :-).")

callFunc = StringWork("test")

callFunc.getString()





# 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
* *INSERT* *code below* and *INSERT* a screen shot of the program and successfully run output.

# DAEN 500, Fall 2020

# Miesha Purcell

# Final Exam

# Question 3

# Problem 3: R Programming Problem

library(magrittr)

library(dplyr)

#Create myCourses vector with 8 courses.

myCourses <- c("Statistics I",

"Calculus I",

"Calculus II",

"Calculus III",

"Linear Algebra",

"DAEN 500",

"SYST 500",

"OR 541")

#Get length of myCourses vector

length(myCourses)

#Get first two courses from myCourses

myCourses[1:2]

#Get 3rd and 4th courses from myCourse

myCourses[3:4]

#Sort myCourses using a method

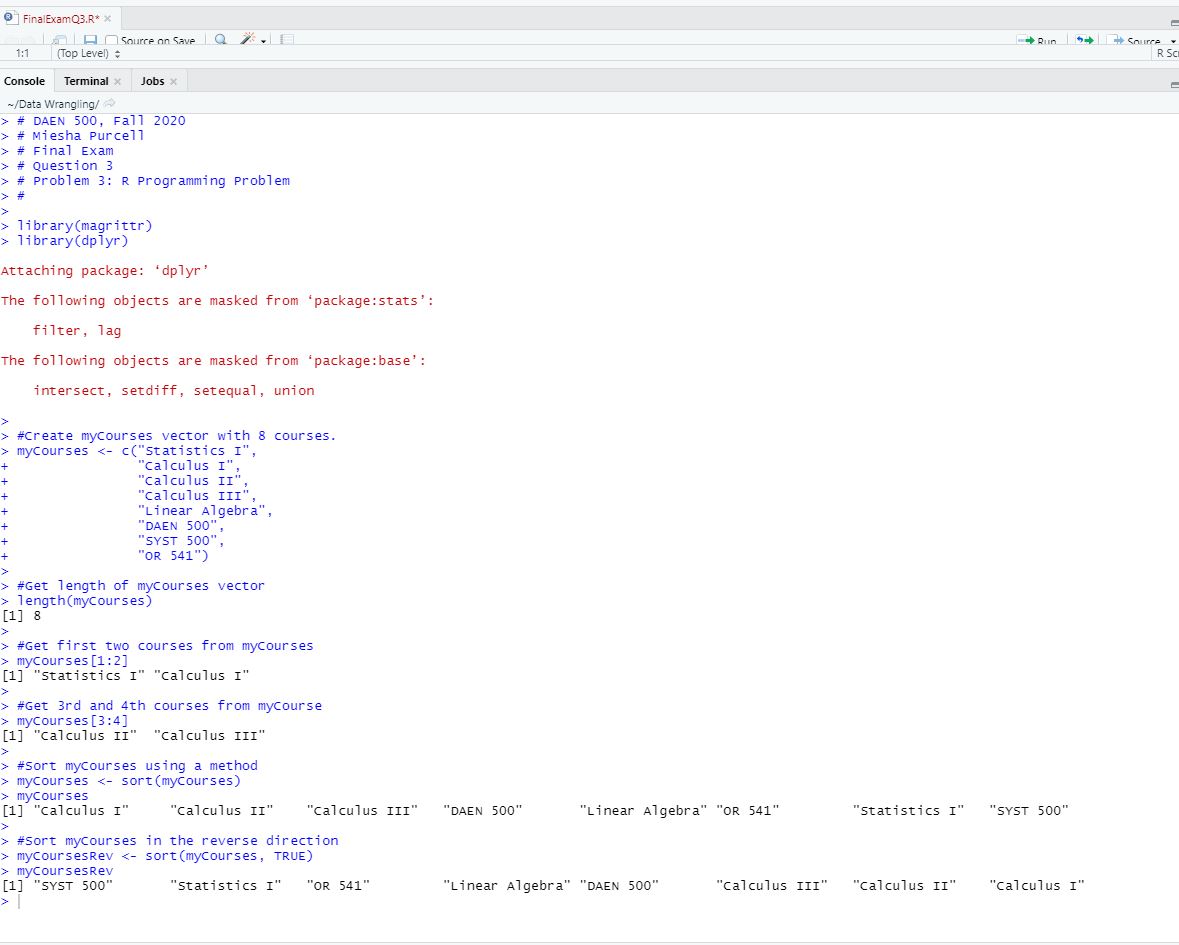
myCourses <- sort(myCourses)

myCourses

#Sort myCourses in the reverse direction

myCoursesRev <- sort(myCourses, TRUE)

myCoursesRev





# Problem 4: Principal Component Analysis

# (25 points)

**Provide a description of the following:**

1. What is a component – Provide a description (5 points)

A component is a unit eigenvector that provides the direction of the data.

1. Principal Component Analysis – Provide a description. (5 points)

Principal Component Analysis is used to analyze multivariate data. It can be used to

multiple pieces of date collected on people or objects and it can be used to reduce the

dimensions of multivariate data.

1. **Provide an specific example of Principal Component Analysis (15 points)**

An example of Principal Component Analysis, that I found in a book titled “Linear Algebra and its Applications”, is of a lab monitoring a chemical process. For the lab to monitor the process, they must take 300 samples of the material produced. Eight tests are run on each of the 300 samples, such as, melting point, density, ductility, tensile strength etc. Then a lab report is generated for each sample with 8 dimensions and forms an 8 x 300 matrix of observations.

Reference:

Lay, D. C., Lay, S. R., & McDonald, J. J. (2016). Applications to Image Processing

and Statistics. In *Linear algebra and its applications* (pp. 426-432).

Boston, MA: Pearson.

# 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)

# Multiple Regression is used when modeling linear relationships

# between one quantitative response variable and one or more

# predictor variables; whereas Logistic Regression used to model a

# binary response variable and the output is the probability of that

# response. Unlike Logistic Regression, Multiple Regression allows you

# to predict the dependent variable. You would most likely use Multiple

# Regression in cases where you have continuous dependent variables,

# such as, predicting the price of a house based on its proximity to the

# metro, its size, and square footage. Logistic Regression would be used

# when your dependent variable is binary or categorical. An example

# would be if you’re predicting the probability of a woman’s fertility

# using predictors of age, weight, pre-existing health conditions etc.

# References: Zybooks, Chapters 23 and 24

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

# The Linear Regression predicts price of house square footage based on

# the proximity to MRT, number of convenience stores in the area, and

# age of the home. The results show that house age and distance to nearest MRT station have a negative correlation to the price per square foot of the house. Meaning that when the age of the house increases, the price of the per square feet is decreased. When the distance to the nearest MRT increases, the price per square foot decreases. The number of convenience stores in the area has a positive correlation to the price per square foot. The more convenience stores in the area the price per square foot is higher. The model predicts price per square foot plus or

# minus 9.312 and it accounts for 54% of the variation.

# The Logistic Regression model below predicts the probability of normal

# or altered fertility in men. The model uses childhood diseases,

# accident/trauma, surgical intervention, high fevers in last year,

# frequency of alcohol consumption, smoking habit, and number of house

# spent sitting as predictor variables.

# 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.

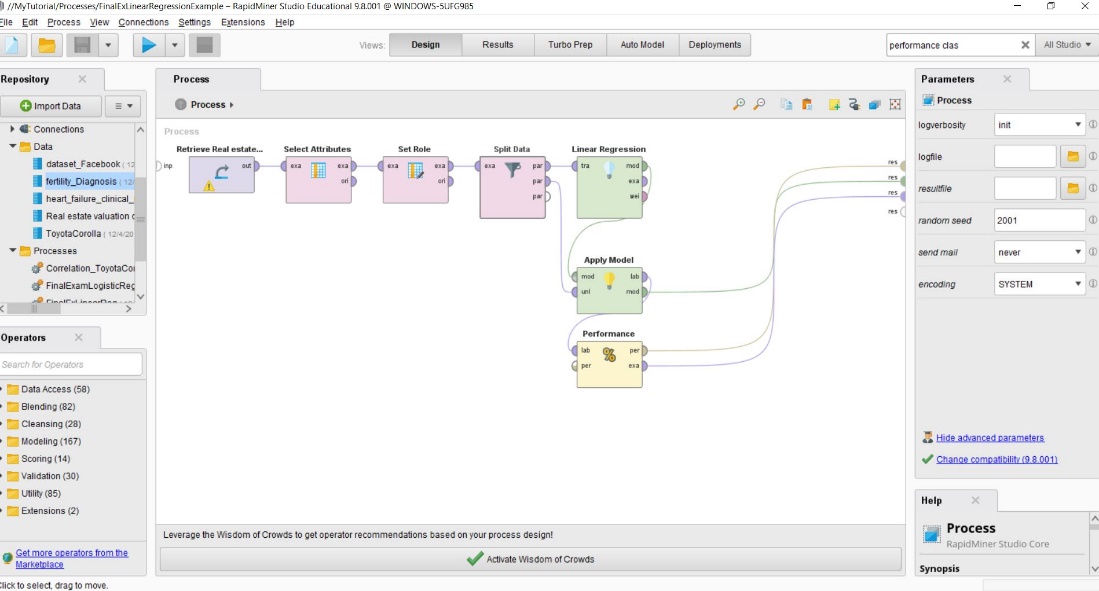
Step 3: Run regression, .and use visualizations to demonstrate the conceptual answers you provided for 5.(a).

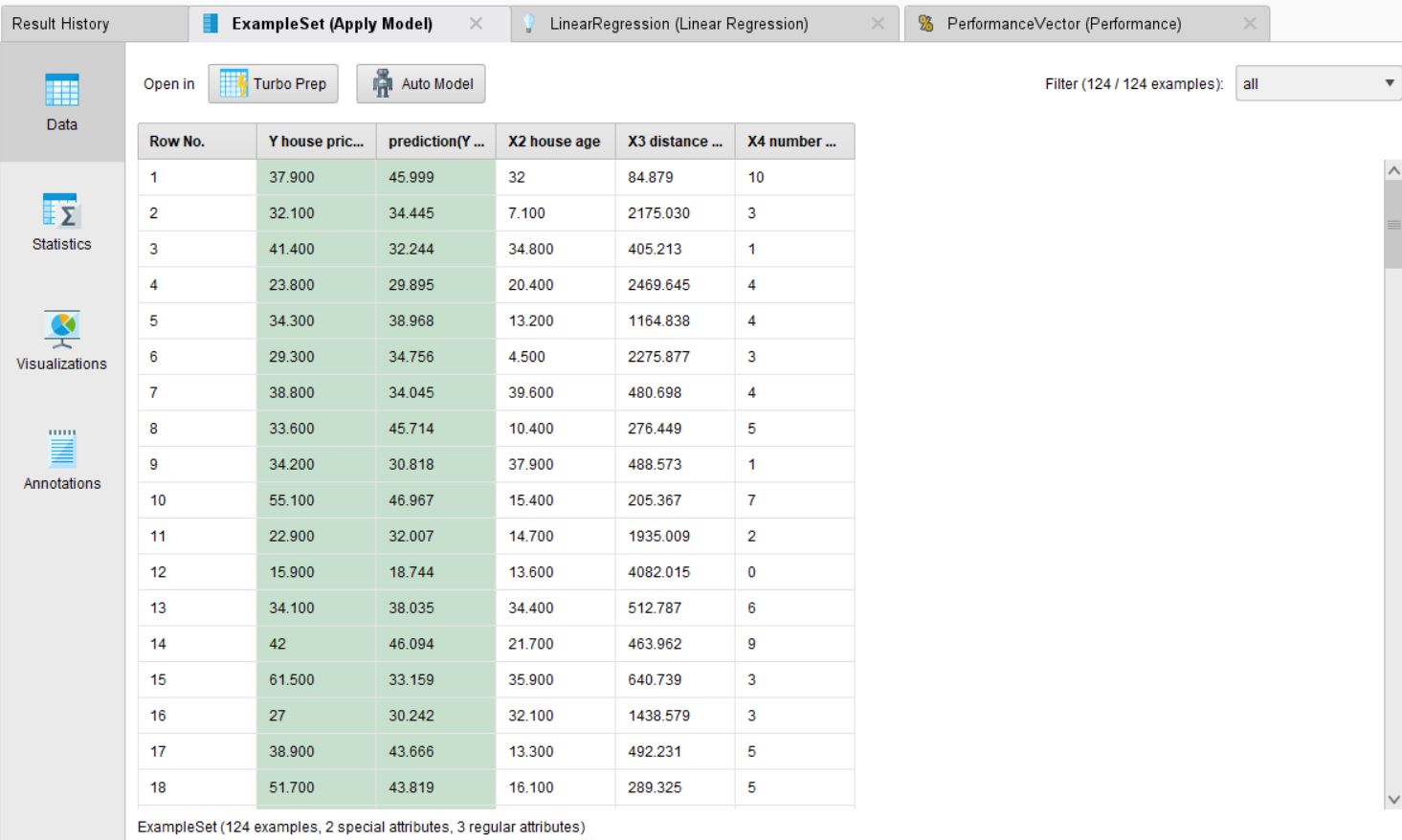
Linear Regression Model using RapidMiner

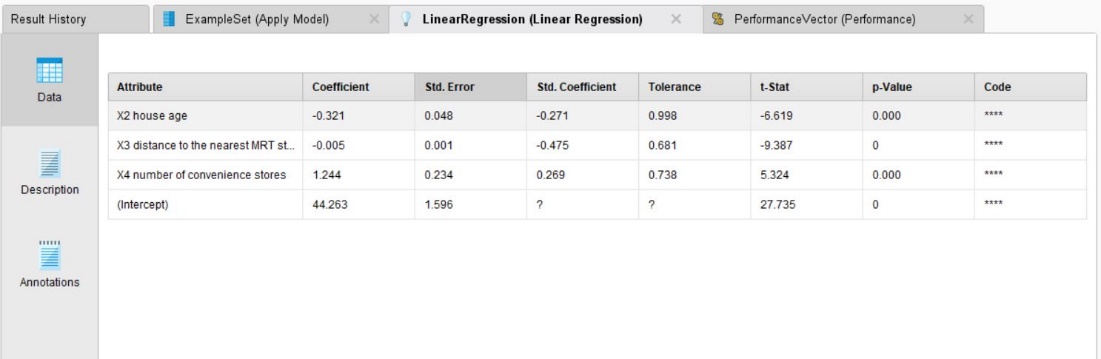
Data Source:

Original Owner and Donor  
Name: Prof. I-Cheng Yeh  
Institutions: Department of Civil Engineering, Tamkang University, Taiwan.  
Email: 140910 **'@'** mail.tku.edu.tw  
TEL: 886-2-26215656 ext. 3181  
  
Date Donated: Aug. 18, 2018

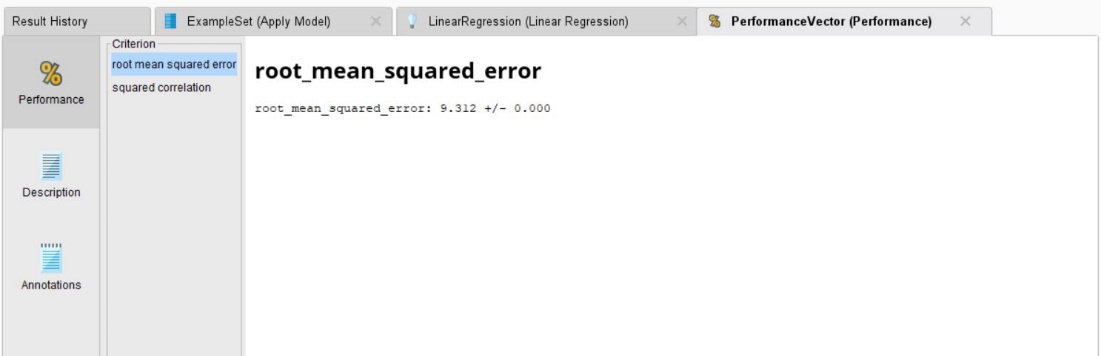
<https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set>



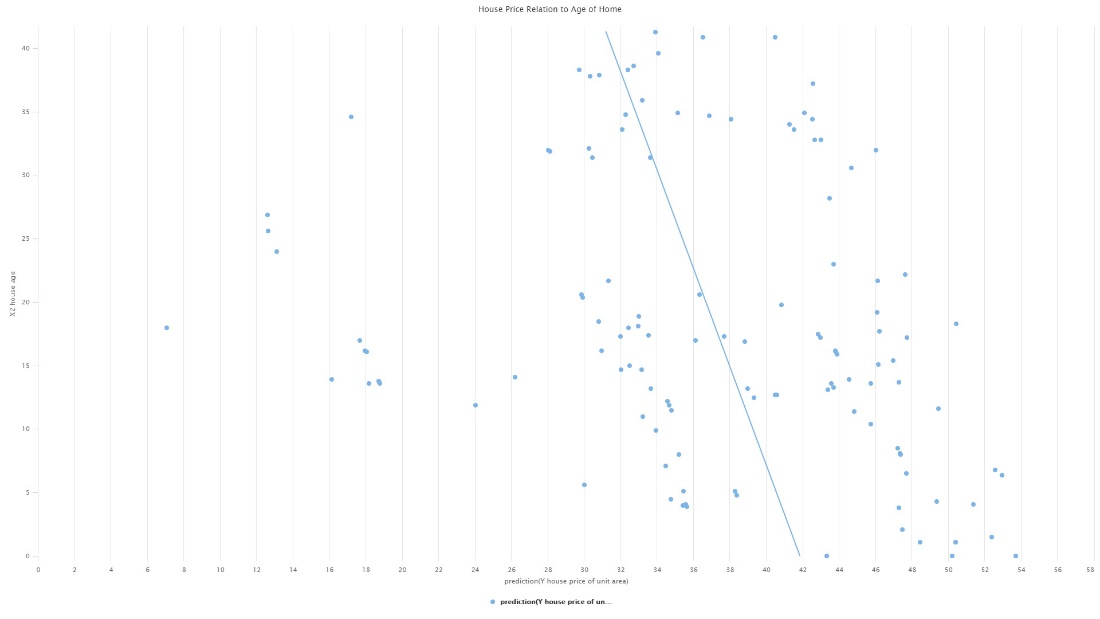


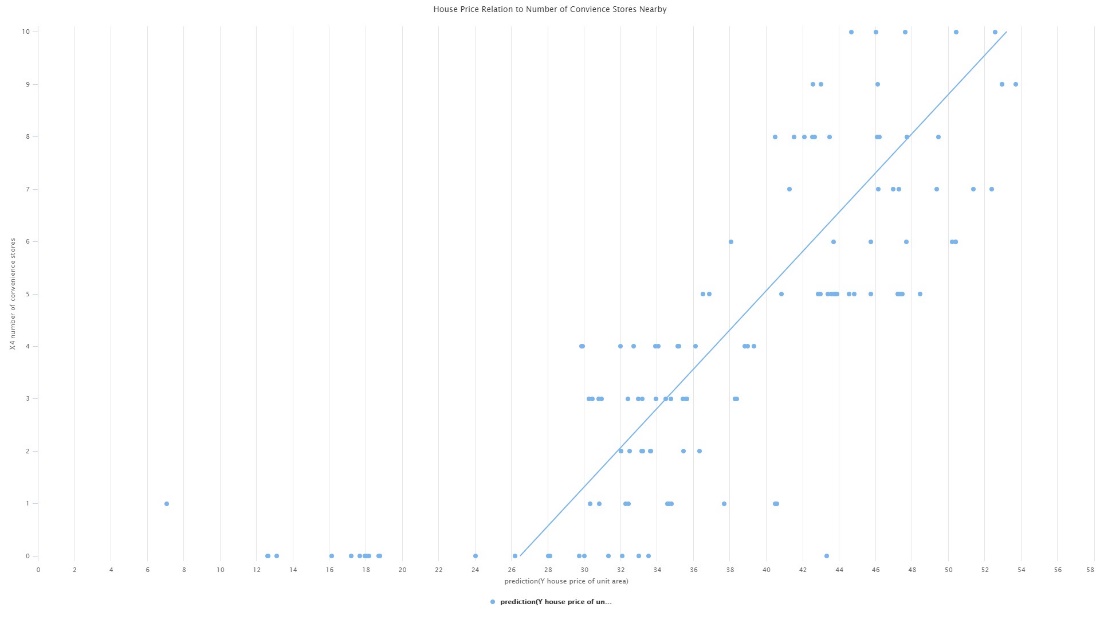


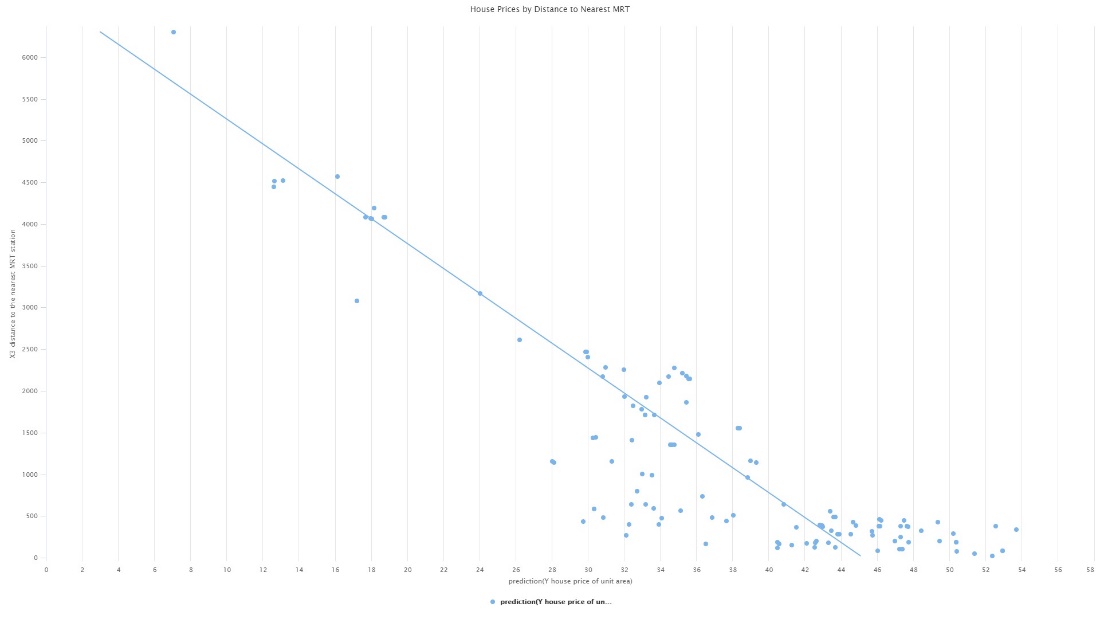


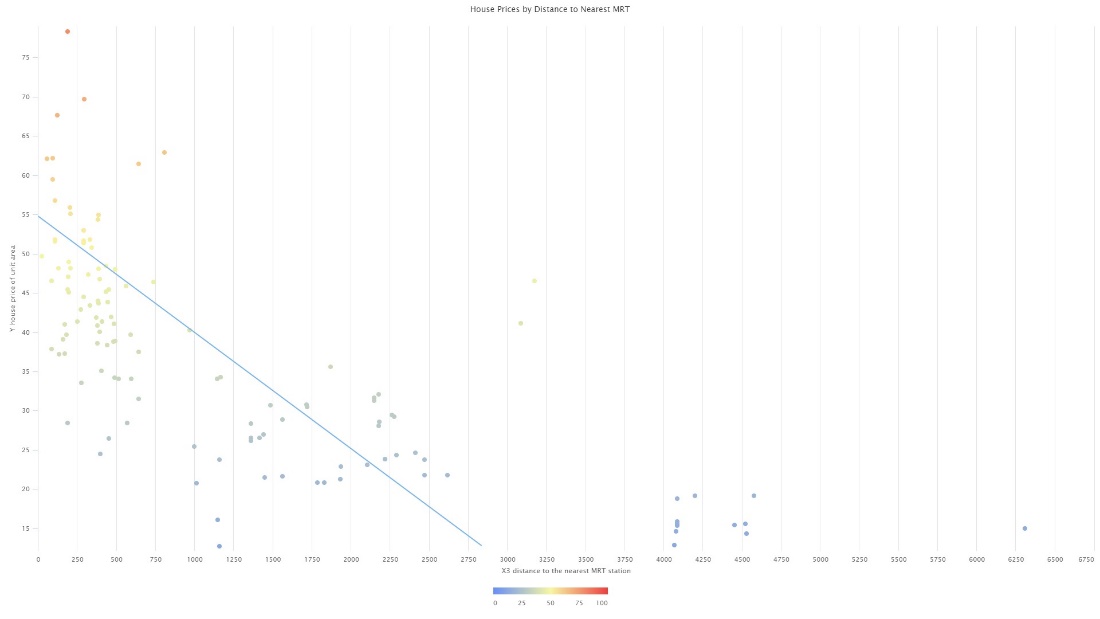












Logistic Regression using RapidMiner

Data Source:

David Gil,  
dgil '@' dtic.ua.es,  
Lucentia Research Group, Department of Computer Technology, University of Alicante  
  
Jose Luis Girela,  
girela '@' ua.es,  
Department of Biotechnology, University of Alicante

<https://archive.ics.uci.edu/ml/datasets/Fertility>

