# Prep Exercise (PE10) Support Vector Machines (SVM)

### General Instructions

1. For this exercise you will answer all of the questions in this document and turn it in to Blackboard.
2. Before you get started make sure to read Chapter 18 of *An Introduction to Data Science* and execute the code throughout the chapter to gain familiarity.
3. Getting Started:
   1. Further exploring data mining techniques, this week we branch off from unsupervised learning techniques such as associated rules mining where there were no particular criteria we tried to predict, to a supervised learning technique called Support Vector Machines (SVM). Support vector machines (SVM) are a highly flexible and powerful method of doing supervised machine learning. Supervised learning means that there is a criterion one is trying to predict. The typical strategy is to divide data into a training set and a test set (for example, two-thirds training and one-third test), train the model on the training set, and then see how well the model does on the test set. For this week’s prep ex and homework, we will be using a chunk of the “diamonds” dataset from the ggplot2 package to do some classification with an SVM. For the sake of simplicity, we will see if we can correctly classify the “cut” of a diamond into one of the two categories, “premium” or “ideal”.
   2. As usual we will use this Prep Ex to set you up for the homework exercises and test your knowledge of materials within the chapter reading. Let’s begin…

# IST 687, Standard Homework Heading

#

# Student name: Thadhani Hitesh Chandrakumar

# Homework number: PE10

# Date due: 6th Nov 2019 11:59 PM

#

# Attribution statement: (choose the statements that are true)

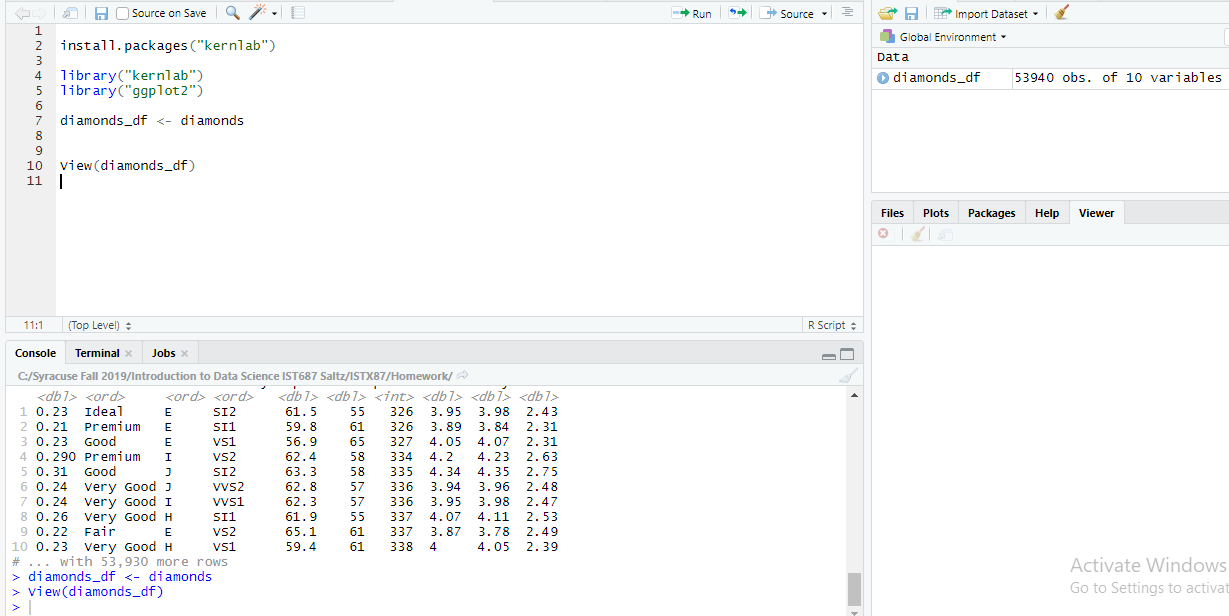
# 1. I did this work by myself, with help from the book and the professor

# 2. I did this work with help from the book and the professor and these Internet sources: https://www.r-bloggers.com, https://towardsdatascience.com

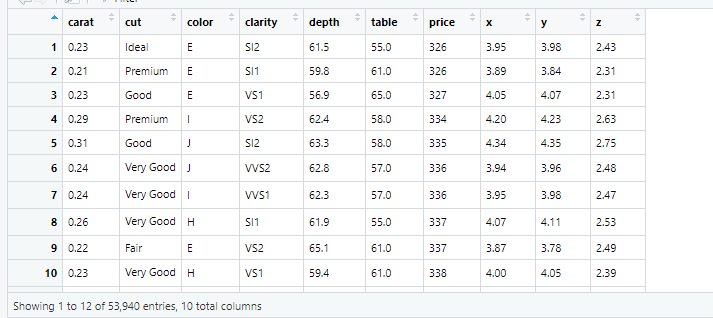
# 3. I did this work with coaching from <Name of another student> but did not cut and paste any code

### Prep Exercise

1. **Getting Ready: Loading and Verifying the Diamonds Dataset.**
   1. This week’s “diamonds” dataset comes from the ggplot2 package while the SVM function are located within a package called kernlab. Therefore, you will need to install and library ggplot2 and kernlab.
   2. Assign the “diamonds” dataset to a dataframe and view the dataset. You will notice that there are five cuts of diamonds: Fair, Good, Very Good, Premium and Ideal.

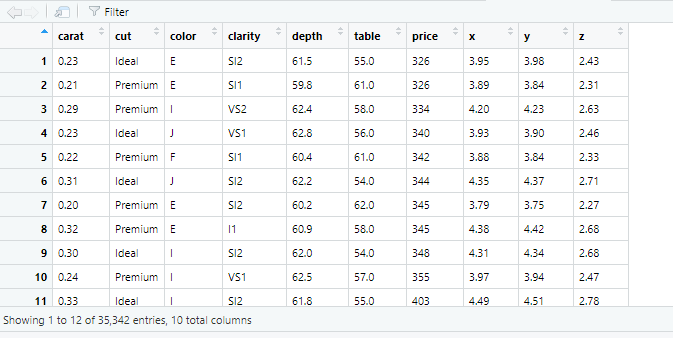


* 1. View the dataframe and place a screenshot below. It is not important to fit the entire dataset in the screenshot, just enough to verify that you have completed the above steps.



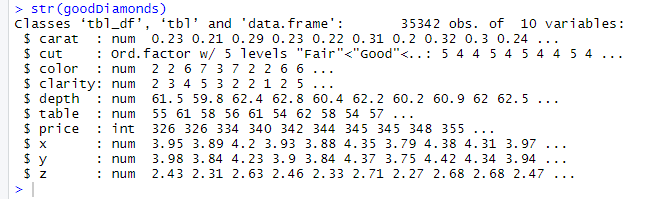
1. **Cleaning the data.**
   1. To simplify our classification task, we will focus only on Premium and Ideal cuts of diamonds. Below, there are two different ways to create the subset of the dataframe that fits the two “cut” categories. Create a new dataframe, called ‘goodDiamonds’ that only has the “Primium” and “Ideal” cut of diamons.

goodDiamonds <- diamonds\_df[which(diamonds\_df$cut %in% c('Premium','Ideal')),]



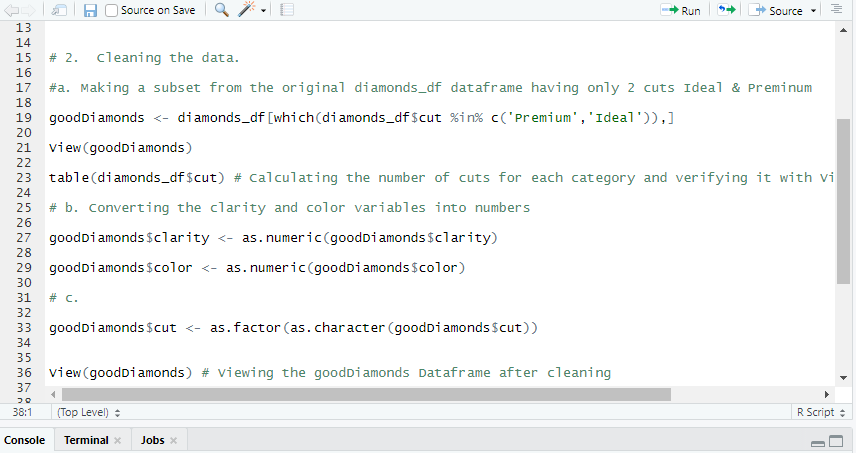
* 1. The clarity and color variables in the dataframe are “ordered factors.” This means that for analytical purposes such as this, you can convert the factor level directly into a number and it will make sense. Convert the clarity and color variables into numbers.

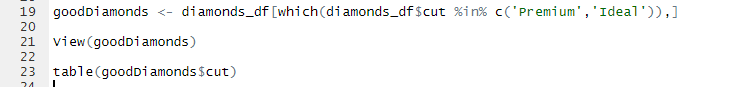
(**Hint:** use the *as.numeric()* command to accomplish this)

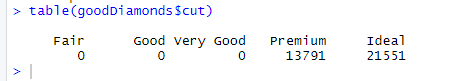


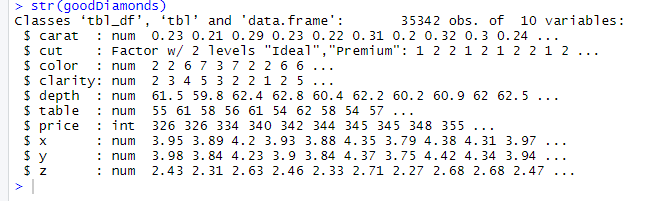
* 1. The cut attribute now has two level (just premium and idea), but the dataframe still has the initial five factors. You can fix this by using the following code:

as.factor(as.character(goodDiamonds$cut))

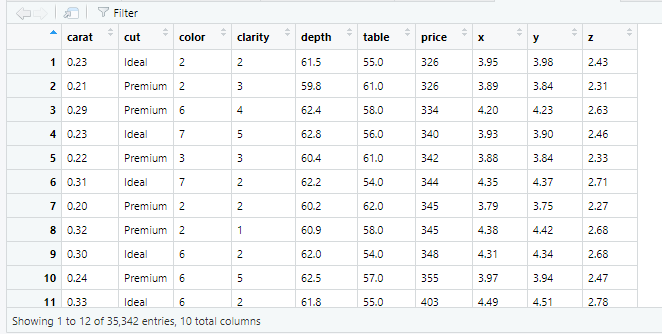








* 1. View the adjusted dataframe and place a screenshot below.



* 1. In a block comment, describe the meaning of each variable within the dataframe.

There are total 10 variables in the dataframe. Carat refers to weight of the diamond between 0.2 and 5.01, cut refers to quality of cut which are of 5 types Ideal, Premium, Good, Very Good & Fair. Color is now made a ordinal variable after converting it into numeric variable from categorical variable which takes values from 1 to 7 for best to worst (D to J now 1 to 7) . Similarly clarity variable is also converted into numeric variable from categorical variable taking values from 1 to 8(1 being worst to 8 being best) for how clear the diamond is. Depth is the total depth percentage and table is the width of top of diamond relative to widest point values ranging from 43 to 95. Price is the price in USD lowest being 326 and highest being 18823. x,y,z are the length, width and depth in mm respectively for the diamond.

1. **Understanding Terminology that will be used in this PE and HW.**
   1. In a paragraph or two explain the concept of a confusion matrix and the theoretical process behind creating one in RStudio.

Confusion matrix is used to measure the performance of a machine learning classification. We can classify the input into 2 or more classes using the classification model but to measure how accurate it is in terms of predicting we use confusion matrix. It is a table of 4 different combinations of predicted & actual values such as True Positive, False Positive, True Negative and False Negative.

After dividing data into training and test data, training the model using the training data and kvsm function generates predictions based on all attributes in the training data. We use predict() function on test data which gives two-dimensional list giving two lists 1 for one category of test data (ex. Ideal Diamonds) and 0 for another category of test data(ex. Premium Diamonds) each list is of same length as the test Data. We create a new dataframe with 1st variable to be type of cut (i.e. a factor indicating Ideal & Premium) from the test Data and 2nd variable to be 1’s and 0’s for each category (Ideal corresponds to 1 and Premium to 0) and use table() on this newly created dataframe to view the output of confusion matrix which contains 4 values.

1. **List any additional resources that you used here.**

https://www.r-bloggers.com, https://towardsdatascience.com

1. **Be sure to save your R file as this will become the starting code for your homework.**

***You must submit all Prep Exercises to blackboard prior to the deadline specified for each assignment.*** PE assignments are due on the evening prior to the lecture class. Late PE assignments will not be accepted for credit.

*Materials provided by the instructors of this course are copyrighted 2018 by Jeffrey Saltz and Jeffrey Stanton, or by the respective instructor that produced them. These materials are provided to enrolled students for their exclusive use during the semester of instruction. Uploading instructor-produced materials to Internet sites (e.g., Coursehero) without permission is considered a breach of professional ethics and will be reported as a possible academic integrity violation.*