Labs: Trees, Hierarchical Clustering, Assignment 5 Presentation Feedback (Project Proposal Feedback: One-on-One with the instructor)

Thilanka Munasinghe

Data Analytics

ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960

Group 3 Labs 7, 28st October 2022

Rpart – recursive partitioning and Conditional Inference

Reminder to go over these code snippets…

group3/lab1_rpart1.R

group3/lab1_rpart2.R

group3/lab1_rpart3.R

group3/lab1_rpart4.R

Try rpart for “Rings” on the Abalone dataset

group3/lab1_ctree1.R

group3/lab1_ctree2.R

group3/lab1_ctree3.R

Code snippets are available: https://aquarius.tw.rpi.edu/html/DA/group3/

Scripts – work through these

Reminder to finish these code examples See in folder group2/ Lab1

Go over the following scrips,

Lab1_bronx1.R.

Lab1_bronx2.R

Lab1_ctree2.R

Lab1_kknn1.R

Lab1_kknn2.R

Lab1_kknn3.R

Lab1_kmeans1.R

Lab1_nyt.R

Search before you ask! You might need to search your code errors online when you are debugging your code!.

script fragments in R available on the web site:

NOTE: you are allowed to work in small groups and discuss during this lab.

Scripts – work through these

Next…

See in folder group2/ Lab3

Go over the following scrips,

Lab3_ctree1.R

Lab3_ctree2.R

Lab3_ctree3.R

…..

And the remaining code snippets in group2/Lab 2 and Lab3

Search before you ask! You might need to search your code errors online when you are debugging your code!

script fragments in R available on the web site:

NOTE: you are allowed to work in small groups and discuss during this lab.

Reminder to finish: In-Class Work: Validation set example with Auto dataset

```
# ISLR: Introduction to Statistical Learning with R (textbook that we use in this class)
# Validation set example with Auto dataset.
library(ISLR)
library(MASS)
library(boot)
set.seed(1)
# Read the cv.glm documentation
??cv.glm

# read the documentation for sample() function
help("sample")
train = sample(392,196)
# We use the subset option in the lm() function to fit a linear regression using,
# only the observations corresponding to the training set.
lm.fit <- lm(mpg~horsepower, data = Auto, subset = train)
# Now we use predict() function to estimate the response for all 392 observations,
# and we use the mean() function to calculate the MSE of the 196 observations in the
# validation set. Note that the -train selects only the observations that are not in,
# the training set.
attach(Auto)
mean((mpg-predict(lm.fit,Auto))[-train]^2)
# Therefore, the estimated test MSE for the linear regression fit is 26.14
```

Resource/Reference: Introduction to Statistical Learning with R, 7th Edition - Chapter 5

- Make sure to finish the Cross-Validation (LOOCV, K-fold Cross validation) Labs in Group 3 folder

Reminder to finish: In-Class Work: Validation set example with Auto dataset…

```
# We can use the poly() function to estimate test error for the quadratic and cubic regression.
# Quadratic regression line
lm.fit2 <- lm(mpg~poly(horsepower,2), data = Auto, subset = train) # Quadratic
mean((mpg-predict(lm.fit2,Auto))[-train]^2)
# Cubic regression line
lm.fit3 <- lm(mpg~poly(horsepower,3), data = Auto, subset = train) # Cubic
mean((mpg-predict(lm.fit3,Auto))[-train]^2)
# The error rates are: 19.82 for quadratics and 19.78 for cubic
# If we choose different training set instead, then we will obtain somewhat different errors,
# on the validation set.
set.seed(2)
train = sample(392,196)
lm.fit <- lm(mpg~horsepower, data = Auto, subset = train)
mean((mpg-predict(lm.fit,Auto))[-train]^2)
# the error rate is 23.29
lm.fit2 <- lm(mpg~poly(horsepower,2), data = Auto, subset = train) # Quadratic
mean((mpg-predict(lm.fit2,Auto))[-train]^2)
# the error rate is 18.90
lm.fit3 <- lm(mpg~poly(horsepower,3), data = Auto, subset = train) # Cubic
mean((mpg-predict(lm.fit3,Auto))[-train]^2)
# the error rate is 19.25
# Using this split of the observations into a training set and validation set,
# we find that the validation set error rates for the models with linear, quadratic,
# and cubic terms are 23.29, 18.90 and 19.25 respectively.
# The model that predict mpg using a quadratic function of horsepower performs better,
# than a models that only involves only a linear function of horsepower, and there is a,
# little evidence in favor of a model that uses a cubic function of horsepower.
```

Resource/Reference: Introduction to Statistical Learning with R, 7th Edition - Chapter 5

Reminder to finish: In-Class Work: K-fold example with Auto dataset

```
# k-Fold Cross Validation
# The cv.glm() function can also be used to implement k-fold CV.
# We once again, set a random seed and initialize a vector in which,
# we will store the CV errors corresponding to the polynomial fits of orders one to #ten.
# here the K =10
# Read the cv.glm documentation
??cv.glm
set.seed(17)
help("rep") # read the documentation for the rep() function in R.
cv.error.10 = rep(0,10) # read documentation, help("rep")
for(i in 1:10){
  glm.fit = glm(mpg ~ poly(horsepower, i), data = Auto)
  cv.error.10[i] = cv.glm(Auto,glm.fit, K=10) $delta[1]
}
cv.error.10
# Notice the computation time is much shorter than LOOCV! :),
# This depends on your laptop performance :)
# We still see little evidence that using cubic or higher-order polynomials terms,
# leads to lower test error than simply using a quadratics fit.
```

Resource/Reference: Introduction to Statistical Learning with R, 7th Edition - Chapter 5

Trees for the Titanic

data(Titanic)

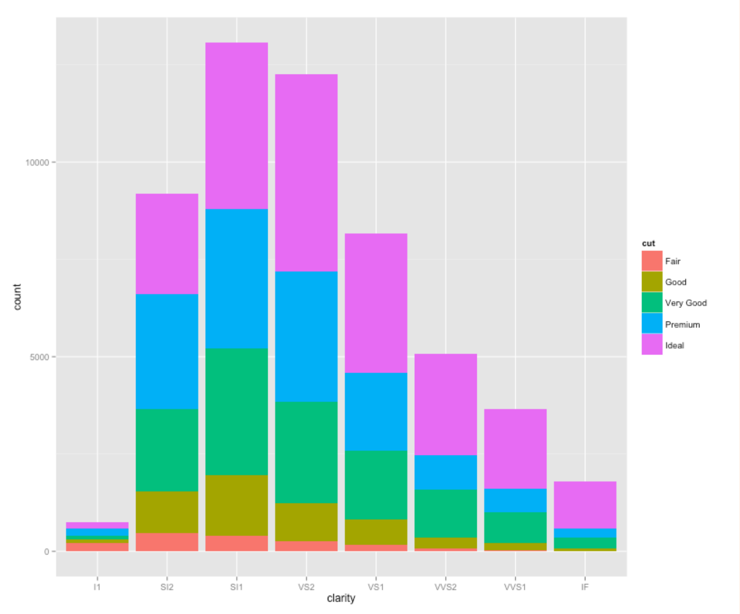
rpart, ctree, hclust, RandomForest for:

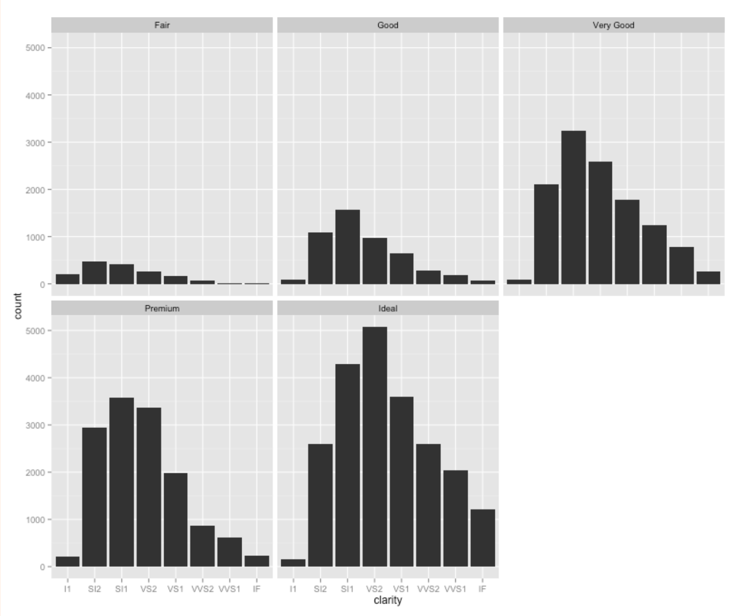
Survived ~ .

Read the titanic dataset documentation in Rdocumentation:
https://www.rdocumentation.org/packages/titanic/versions/0.1.0

Diamonds

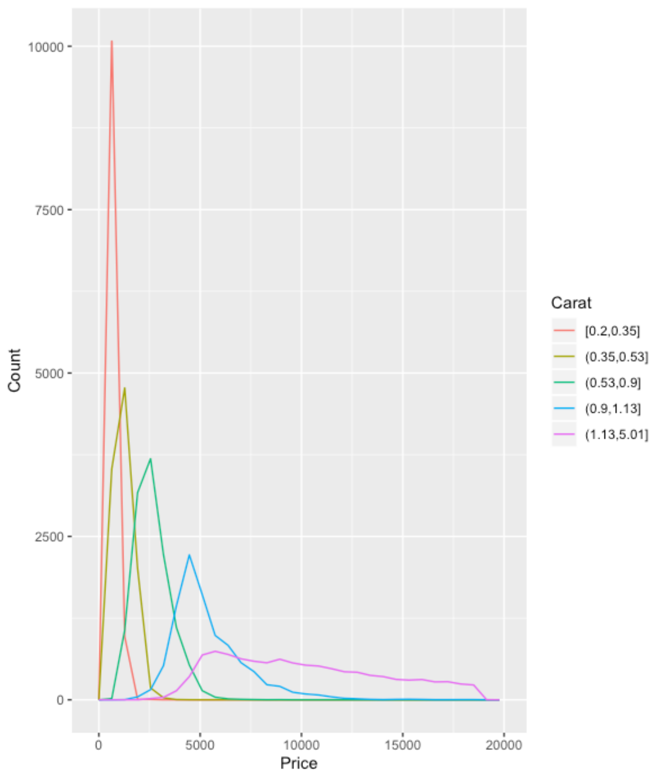
```
require(ggplot2)       # or load package first
data(diamonds)
head(diamonds)          # look at the data!
#
ggplot(diamonds, aes(clarity, fill=cut)) + geom_bar()
ggplot(diamonds, aes(clarity)) + geom_bar() + facet_wrap(~ cut)
ggplot(diamonds) + geom_histogram(aes(x=price)) + geom_vline(xintercept=12000)
```





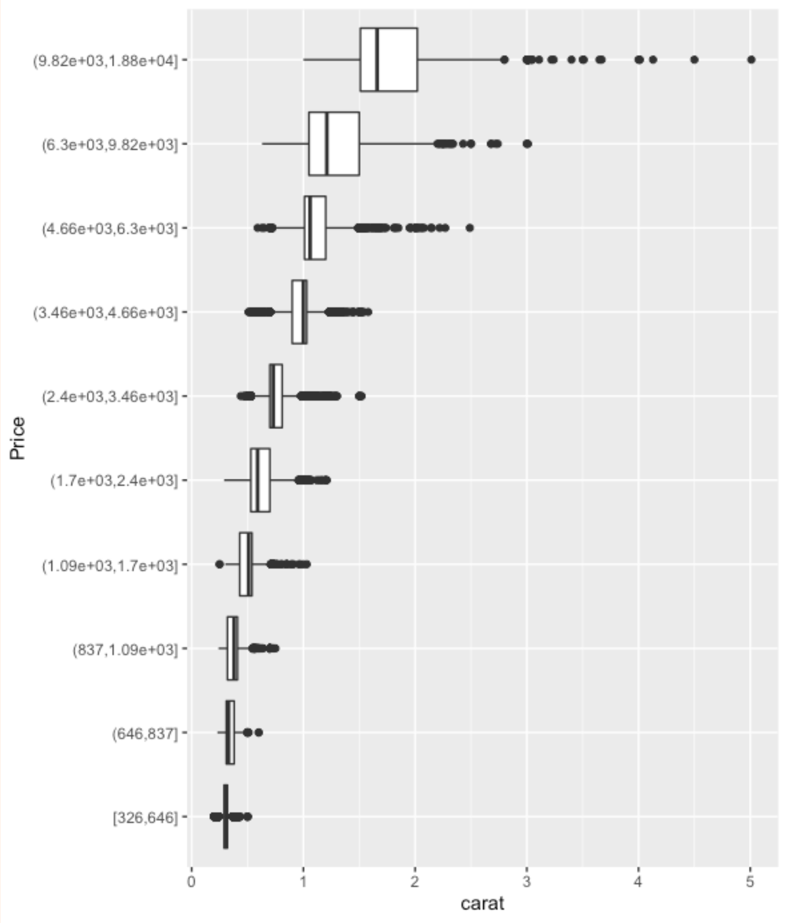
Resource/Reference: R for Data Science : https://r4ds.had.co.nz/
: https://jrnold.github.io/r4ds-exercise-solutions/exploratory-data-analysis.html

```
ggplot(
  data = diamonds,
  mapping = aes(color = cut_number(carat, 5), x = price)
) +
  geom_freqpoly() +
  labs(x = "Price", y = "Count", color = "Carat")
```



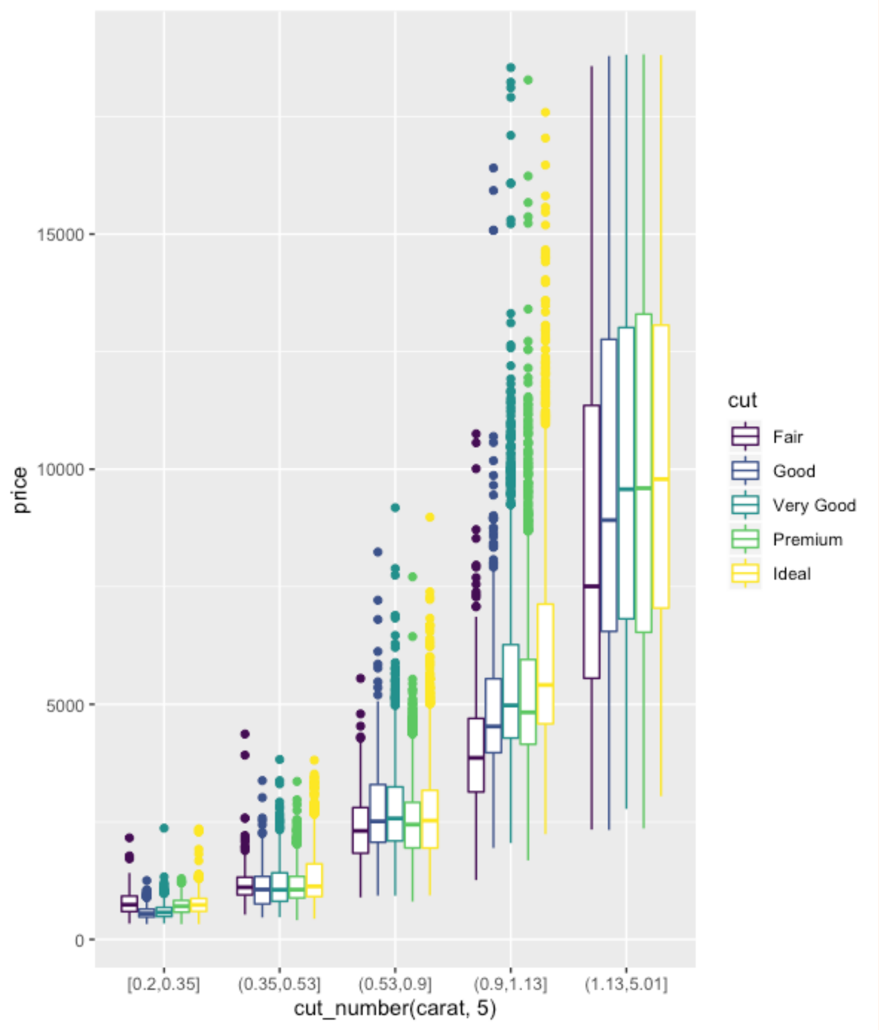
Resource/Reference: R for Data Science : https://r4ds.had.co.nz/
: https://jrnold.github.io/r4ds-exercise-solutions/exploratory-data-analysis.html

```
ggplot(diamonds, aes(x = cut_number(price, 10), y = carat)) +
  geom_boxplot() +
  coord_flip() +
  xlab("Price")
```



Resource/Reference: R for Data Science : https://r4ds.had.co.nz/
: https://jrnold.github.io/r4ds-exercise-solutions/exploratory-data-analysis.html

```
ggplot(diamonds, aes(x = cut_number(carat, 5), y =
price, colour = cut)) +
  geom_boxplot()
```



Resource/Reference: R for Data Science : https://r4ds.had.co.nz/
: https://jrnold.github.io/r4ds-exercise-solutions/exploratory-data-analysis.html

Push Your Code to Github!

- Make sure to push your Lab codes to your GitHub repository. TA will check your Lab work code on your repository.
- Reminder: Assignment 4 is **Due: October 26th, 2021** (by 11:59pm ET) Submission method: written document via LMS
- Please use the following file naming for electronic submission: **DataAnalytics_A4_YOURFIRSTNAME_YOURLASTNAME.xxx**

Assignment 5 Feedback

- Make sure to meet with the instructor to do the **Assignment 5 feedback during today's class One-on-One time**.
- If you have not met with the instructor on last Monday to do get the presentation feedback, you must meet with the instructor today to do the One-on-One session.

Meet the instructor using WebEx meeting room link:
https://rensselaer.webex.com/meet/munast