Please complete the assigned problems to the best of your abilities. Ensure that the work you do is entirely your own, external resources are only used as permitted by the instructor, and all allowed sources are given proper credit for non-original content.

1 Recitation Exercises

These exercises are to be found in: **Introduction to Statistical Learning, 2nd Edition (Online Edition)** by *Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani.*

1.1 Chapter 2

Exercises: 1,2,4,6,7

1.2 Chapter 3

Exercises: 1,3,4

2 Practicum Problems

These problems will primarily reference the *lecture materials and the examples given in class* using **R** and **CRAN**. It is suggested that a *RStudio* session be used for the programmatic components.

2.1 Problem 1

Load the *iris* sample dataset into **R** using a dataframe (it is a built-in dataset). Create a boxplot of each of the 4 features, and highlight the feature with the largest empirical *IQR*. Calculate the parametric standard deviation for each feature - do your results agree with the empirical values? Use the *ggplot2* library from **CRAN** to create a colored boxplot for each feature, with a box-whisker per flower species. Which flower type exhibits a significantly different *Petal Length/Width* once it is separated from the other classes?

2.2 Problem 2

Load the *trees* sample dataset into **R** using a dataframe (it is a built-in dataset), and produce a 5-number summary of each feature. Create a histogram of each variable - which variables appear to be normally distributed based on visual inspection? Do any variables exhibit positive or negative skewness? Install the *moments* library from **CRAN** use the *skewness* function to calculate the skewness of each variable. Do the values agree with the visual inspection?

2.3 Problem 3

Load the *auto-mpg* sample dataset from the UCI Machine Learning Repository (**auto-mpg.data**) into **R** using a dataframe (**Hint**: You will need to use *read.csv* with *url*, and set the appropriate values for **header**,**as.is**, and **sep**). The *horsepower* feature has a few missing values with a **?** - and will be treated as a string. Use the *as.numeric* casting function to obtain the column as a numeric vector, and replace all NA values with the median. How does this affect the value obtained for the mean vs the original mean when the records were ignored?

2.4 Problem 4

Load the *Boston* sample dataset into **R** using a dataframe (it is part of the **MASS** package). Use **lm** to fit a regression between *medv* and *lstat* - plot the resulting fit and show a plot of fitted values vs. residuals. Is there a possible non-linear relationship between the predictor and response? Use the **predict** function to calculate values response values for *lstat* of 5, 10, and 15 - obtain confidence intervals as well as prediction intervals for the results - are they the same? Why or why not? Modify the regression to include $lstat^2$ (as well *lstat* itself) and compare the R^2 between the linear and non-linear fit - use **ggplot2** and *stat_smooth* to plot the relationship.