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 12

Exercises: 1,2,3,4

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 *Wine* sample dataset from the UCI Machine Learning Repository (**wine.data**) into **R** using a dataframe (**Note**: The column names will need to be loaded separately). Use either the **prcomp** or **princomp** methods to perform a PCA of the wine data - justify whether scaling of the inputs should be used or not when performing the decomposition. Plot a **biplot** of the results - identify a feature which is pointed in the opposite direction of *Hue* in principal component/rotated feature space. What does this imply regarding the correlation of this feature to *Hue*? Support your result with a calcualted value. Finally, plot a **screeplot** of your results and determine the percentage of total variance explained by PC1 and PC2.

2.2 Problem 2

Load the *USArrests* sample dataset from the built-in datasets (**data(USArrests)**) into **R** using a dataframe (**Note**: Row names are states, not numerical values!). Use the **kmeans** package to perform a clustering of the data with increasing values of k from 2 to 10 - you will need to decide whether or not to center/scale the observations - justify your choice. Plot the *within-cluster sum of squares* for each value of k - what is the optimal number of clusters? Use the **tidyverse** and **fviz_cluster** plotting method from **factoextra** to plot the optimal clustering.

2.3 Problem 3

Load the *Wine Quality* sample dataset from the UCI Machine Learning Repository (**winequality-white.csv**) into **R** using a dataframe (**Note**: There is both a red and white wine file, we will use white!). Excluding the *quality* target variable, use **hclust** to perform a hierarchical clustering of the data with *single* as well as *complete* linkage. You will need to decide on whether or not to center/scale the observations - justify your choice. At what distance value are the two penultimate clusters merged? Use the **cutree** method to obtain these two clusters, and calculate their summary statistics. What feature means have the largest differences? Which linkage method produces are more balanced clustering?