Girish Rajani
A20503736
CSP-571 Data Preparation and Analysis
Assignment 1

1. Recitation Exercises

1.1 Chapter 2

1.

a. **The sample size n is extremely large, and the number of predictors p is small** -
In this scenario, we can expect a flexible model to perform **better** than an inflexible model because a flexible model generally performs better when given a large number of observations. This will also reduce the chance of overfitting and provide accurate predictions when compared to using an inflexible model.

b. **The number of predictors p is extremely large, and the number of observations n is small** -
In this case, a flexible statistical learning method will perform **worse** than an inflexible model because since the number of observations n is small, the flexible model will most likely overfit.

c. **The relationship between the predictors and response is highly non-linear** -
Due to the non-linear nature of the relationship mentioned above, a flexible model will perform **better** than an inflexible model. Generally, the function of flexibility is greater in a flexible model when there is non-linearity when compared to an inflexible model.

d. **The variance of the error terms, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high.**
In this scenario, a flexible model will perform **worse** because since the variance of the error term is extremely high, the MSE increases which will result in overfitting. A simpler model using an inflexible statistical learning method will be sufficient to produce an accurate estimate of f.

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

a. This example can be looked at as a regression problem because we are interested in understanding the CEO salary. This is a quantitative (numerical) output value which means that it is a regression problem. In this case, we are most interested in inference here since we are trying to understand the factors that are affecting the CEO salary rather than actually predicting the CEO salary. n = 500, p = 3 (profit, number of employees, industry).
b. This example can be looked at as a classification problem because we are interested in predicting whether the new product will be a success or a failure which deals with a qualitative output value (categorical, belonging to two classes). In this case, we are most interested in prediction as previously mentioned. n = 20, p = 13 (price charged for the product, marketing budget, competition price, and ten other variables).

c. This example can be looked at as a regression problem because we are interested in the % change in the USD/Euro exchange rate which can be considered quantitative (numerical) output value. In this case, we are most interested in prediction since we want to predict the future % change in the exchange rate. n = 52 (52 weeks in 2012), p = 3 (the % change in the US market, the % change in the British market, and the % change in the German market).

4. You will now think of some real-life applications for statistical learning.
   a. Classification
      i. Stock Price - We collect financial data to predict whether the stock price for a given company will go up or down. The response will be qualitative, whether the stock price will go up or down while the predictors are financial data, previous stock prices, and expected growth. The goal of this scenario revolves around prediction whereby we predict whether the stock price will go up or down.

      ii. Email Spam - We can use a set of past emails to predict whether an email is classified as spam or not. The response will be whether the email is spam or not while the predictors are email header, body, images, and links. The goal of this scenario revolves around prediction whereby we predict if a certain email is a spam or not.

      iii. Job Application Acceptance - We can use job applications from previous applicants to determine whether or not we should hire a candidate. The response will be whether a candidate is accepted or rejected while the predictors are experience, education, and skills. The goal of this scenario

revolves around prediction whereby we use previous accepted job applications to predict whether a candidate is accepted or rejected.

b. Regression
    i. Income of the person - Here we would like to understand the factors contributing to a person's salary. The response will be the salary while the predictors are job title, qualification, and country. In this scenario, we are considering inference since the goal is not to predict the salary but more so to understand the relationship between the salary and the predictors to find out which factors affect salary.
    ii. Reckless Driving Accidents - Here we would like to predict the number of driver related car accidents. The response will be number of car accidents caused by a reckless driver while the predictors are location, time of day, speed, drinking, and texting. In this scenario, we are looking at the relationship between reckless driving and the number of driver related car accidents so that to predict future reckless driving related accidents.

    iii. EV Charging Time - Here we would like to predict the time taken to charge an electric vehicle from 0 to full battery. The response will be time taken for a full battery charge while the predictors are number of battery cells, battery temperature, and battery deterioration. In this scenario, we are looking to predict how long the battery would take to charge.

c. Cluster analysis
    i. Divison of household class - Here we are interested in clustering household families as either upper class, middle class, or lower class. The response is the family will fall under one of the three categories mentioned above. The variables are income before tax, location, education level, age, and martial status.

    ii. Fraudulent Uber Drivers - Here we are interested in clustering which Uber drivers are real or fake. The response will be whether the driver is real or fraudulent. The characteristics observed are GPS logs, license plate, driver name, and car description.

    iii. Division of different types of Gamers - Here we are interested in clustering people who play video games as either casual gamers or hardcore gamers based on certain observations. The response will be this gamer falls under casual or hardcore while observing characteristics such as age, time spent playing games, location, and number of tournaments won.

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

A parametric approach makes an assumption about the form of f such as assuming that f is linear or just estimating a set of parameters. However, a non parametric approach does not make such an assumption. It instead seeks an estimate of the form f that is as close to the data. A parametric approach has high interpretability and low flexibility (restrictive) while a non parametric approach has low interpretability and high flexibility (flexible).

The advantage of parametric approach to regression or classification is that since it makes that assumption, this makes the problem simpler because we don't need to estimate the entire function f but instead just estimate a parametric form of f. Additionally, parametric approach will contain fewer parameters since it would be considered a simpler model, whereas, a non parametric approach requires a large number of observations and more parameters.

The disadvantage of a parametric approach is that since we are estimating a set of parameters instead of the entire form of f, the selected model may not match the true unknown form of f. If we use a more flexible model to solve this problem, we may need to estimate more parameters which can lead to another problem, overfitting.

7.

a. Euclidean distance is defined as the length of a line segment between two points. It can be calculated using the Pythagorean Theorem.

Obs 1 = $\sqrt{3^2} = 3$

Obs 2 = $\sqrt{2^2} = 2$

Obs 3 = $\sqrt{1^2 + 3^2} = \sqrt{10} = 3.16$

Obs 4 = $\sqrt{1^2 + 2^2} = \sqrt{5} = 2.24$

Obs 5 = $\sqrt{(-1)^2 + 1^2} = \sqrt{2} = 1.41$

Obs 6 = $\sqrt{1^2 + 1^2 + 1^2} = \sqrt{3} = 1.73$

b. When we use K = 1, we first identify the observation that is nearest neighbor. Based on the Euclidean distance computed in (a) above, the nearest observation (closest to 1) is observation 5. The output of this observation is Green so since we use this observation as our KNN, we say our prediction is also Green.

c. When we use K = 3, we identify the three nearest neighbor observations which was observation 5, 6, and 2. This neighborhood consists of 2 Red outputs and 1 Green output resulting in estimated probabilities of ⅔ for the Red class and ⅓ for the Green class. Therefore, the KNN will predict Y belonging to the Red class.

d. If the Bayes decision boundary in this problem is highly non-linear, we expect the best value for K to be **small**. This is because non-linear has to do with high flexibility and lower values of K are more flexible. As K grows, the method gets less flexible and produces a decision boundary close to linear making large values of K not a good option for highly non-linear decision boundary.

1.2 Chapter 3

1.

The null hypothesis by definition mentions that there is no relationship between TV, radio and newspaper on sales.

When looking at the p-values in Table 3.4, we notice that the advertising budget for TV and radio have a significantly smaller p-value (<0.0001) when compared to that of newspaper (0.8599). With very small p-values for radio and TV, we can conclude that **there are some relationships between TV and radio on sales**. Since these relationships exists, **we can simply reject the null hypothesis** for these relationships (since their p-values are significant).

However, since the p-value for newspaper is significantly larger, we can conclude that for this large p-value, **there is no relationship between newspaper advertising and sales**, therefore, **accepting the null hypothesis** for this relationship (since the p-value of this is insignificant.

3.

(a)

Multiple Linear Rgression Model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \dots \ + \beta_p x_p$$

$$Y = 50 + 20GPA + 0.07IQ + 35Level. + 0.01(GPA * IQ) - 10(GPA * Level)$$

For college (where Level = 1):

$$Y = 50 + 20GPA + 0.07IQ + 35 + 0.01(GPA * IQ) - 10GPA$$

$$Y = 85 + 10GPA + 0.07IQ + 0.01(GPA * IQ)$$

For highschool(where Level = 0):

$$Y = 50 + 20GPA + 0.07IQ + 0.01(GPA * IQ)$$

From the above, we can see that for a fixed value of IQ and GPA, if the GPA is high enough, then the high school graduates will earn more than college graduates. We cannot just say for any fixed value of IQ and GPA, the high school graduates will earn more because if we use a low GPA in the above equations, then college graduates will earn more than high school graduates.

Therefore, iii is correct - For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

(b) Salary of a college graduate with IQ of 110 and a GPA of 4.0:

For college (where Level = 1):

$$Y = 50 + 20GPA + 0.07IQ + 35 + 0.01(GPA * IQ) - 10GPA$$
$$Y = 85 + 10(4.0) + 0.07(110) + 0.01(4.0 * 110) = \$137,100$$

(c) False - We cannot say that there is very little evidence of an interaction effect just by looking at the very small coefficient for the GPA/IQ. More information such as the p-value is required to properly justify an interaction effect.

4.

(a)
I would expect the training RSS for the cubic regression to **be lower** than the training RSS for the linear regression. This is because the model for the cubic regression would be **more flexible** than the linear regression model. As the flexibility of the model increases, the RSS would decrease, hence making the cubic regression model result in a better fit.

(b)
I would expect the testing RSS for the cubic regression to **be higher** than the testing RSS for the linear regression. This is because the true relationship between X and Y is linear so this means that using a cubic regression model in a linear relationship will cause errors and lead to overfitting on test data. Since the true relationship is linear then on test data, the linear regression model will be a better fit.

(c)
I would expect the training RSS for the cubic regression to **be lower** than the training RSS for the linear regression. Due to the increased coefficients and flexibility of the cubic regression model, the cubic regression will fit better than the linear model since there is a non-linear relationship between X and Y.

(d)
I would expect in general, that the testing RSS for the cubic regression to **be lower** than the testing RSS for the linear regression simply because the true relationship between X and Y is not linear. However, there is also **not enough information** to tell because we do not know if the relationship is closer to linear or cubic. We can say though, that if the model is closer to linear, then the cubic regression can overfit resulting in high testing RSS. Similarly, if the model is closer to cubic than that of linear then the cubic regression will fit well and have a lower RSS. Lastly, due to this uncertainty, the linearity may be in between linear and cubic making it possible for both of them to have the same/similar test RSS.