1. **Automobile Insurance Claims**. (Frees 1.3 page 16) As an actuarial analyst, you are working with a large insurance company to help it understand claims distribution for private passenger automobile policies. You hae available claims data for recent year, consisting of

   - STATE CODE (state): codes 01 through 17 used, with each code randomly assigned to an actual individual state
   - CLASS (class): rating class of operator, based on age, sex, marital status, and use of vehicle
   - GENDER (gender): operator sex AGE (age): operator age
   - PAID (paid): amount paid to settle and close a claim.

   You are focus on older drivers, 50 and older, for which there are $n = 6,773$ claims available. For the graphing, please use package ggplot2 for histogram and scaterplot in (b) - (e) below.

   (a) Compute descriptive statistics for the amount paid (PAID). What is the typical paid (mean, median)?

   (b) Graph a histogram and (normal) qq plot for PAID. Comment on the shape of the distribute.

   (c) Graph again the histogram and (normal) qq plot for LNPAID (logpaid). Comment on the shape of the distribute.

   (d) Scatterplot LNPAID against age. Comment on if there is any relationship between the payment (LNPAID) and age.

   (e) add color by GENDER to (d) above. Comment on relationship between the payment and gender.

2. James page 56 exercise 10. This exercise involves the Boston housing data set.
   To begin, load in the Boston data set. The Boston data set is part of the MASS library in R.

   (a) Make some pairwise scatterplots of the predictors in this data set. Describe your findings.

   (b) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

   (c) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

   (d) How many of the suburbs in this data set bound the Charles river?

   (e) What is the median pupil-reacher ratio among the towns in this data set?

   (f) Which suburb of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

(g) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than either rooms per dwelling.

3. James 2.7 Page 53

The table below provides a training data set containing data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	1	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red
7	-1	2	-1	Green

Table 1: K-nearest neighbors application

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compaute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

(b) What is our prediction with $K = 1$? Why?

(c) What is our prediction with $K = 3$? Why?

(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the *best* value for $K = 1$ to be large or small? Why?