1 Recitation problems

1.1 Tan, Chapter 1

1.

(a) No. It is a database query question.

(b) No. It is an accounting calculation by calculate the balance of user’s profitability.

(c) No. It is an accounting calculation.

(d) No. It is a database query question.

(e) No. It is an accounting calculation by calculate the possibility of the result. Because each result of dice is fair.

(f) Yes. We need to collect and analyze the data of previous stock price, then create a model to predict the future price, which is the task of predictive.

(g) Yes. We need to collect and analyze the data of normal heart rate behavior then create a model to detect the abnormal situation to raise an alarm, which is the task of anomaly detection.

(h) Yes. Similar to the situation in (g). We need to use normal seismic waves data to create the model and raise an alarm if different types of seismic waves which indicates the earthquake activities occur.

(i) No. It is a signal processing problem.

3.

(a) No. Because it is necessary for social management.

(b) Yes. Because it is user’s privacy problem which may collect to user’s address.

(c) No.

(d) No. Because if we could get this kind of data, it means the data is open to us.

(e) No. Similar to the problem in (d). But if we could collect it from web, maybe it means open to the public.

1.2 Tan, Chapter 2

2.

(a) Binary, qualitative, ordinal.

(b) Continuous, quantitative, ratio.

(c) Discrete, qualitative, ordinal.

(d) Continuous, quantitative, ratio.

(e) Discrete, qualitative, ordinal.

(f) Continuous, quantitative, ratio.

(g) Discrete, quantitative, ratio.

(h) Discrete, qualitative, nominal.

(i) Discrete, qualitative, ordinal.

(j) Discrete, qualitative, ordinal

(k) Continuous, quantitative, ratio.

(l) Discrete, quantitative, ratio.

(m) Discrete, qualitative, nominal.

3.

(a) In my opinion, the boss is right. If I were the boss, I would give a better measure:

$$satisfaction\ of\ the\ product = \frac{the\ number\ of\ complaints}{the\ total\ numver\ of\ sales}$$

(b) I think there are nothing about the attribute type of the original product satisfaction. Because the same satisfaction may have different number of complaints. What’s more, satisfaction is not any type of attributes since it more about personal experiences and it varies from person to person. And there is no way satisfaction can be quantified.

7.

Temperature shows more temporal autocorrelation, rainfall is hard to predict, but temperature won’t change rapidly.

12.

(a)No. Just generated by error.
(b)Yes. If standard deviation from normal is large enough.
(c)No. Again noise is not same as outliers.
(d)No, outliers sometimes represent a bigger cluster and we can actually find it if we have enough data.
(e)Yes.

1.3 ISLR 7e (Gareth James, et al.)

1.

The null hypotheses advertise budgets of “TV”, “radio” or “newspaper” which don’t have an effect on sales.

Assume $H_0^{(1)}: \beta_1 = 0$ $H_0^{(2)}: \beta_2 = 0$ $H_0^{(3)}: \beta_3 = 0$

The corresponding p-values are significant to “TV” and “radio” and not significant to “newspaper”, so we reject $H_0^{(1)}$and $H_0^{(2)}$ but $H_0^{(3)}$.

We could conclude that newspaper advertising budget do not affect sales.

3.

(a) iii.

The least square line is:

$$y^{\wedge} = 50 + 20GPA + 0.07IQ + 35Gender + 0.01GPA * IQ - 10GPA * Gender$$

For males:

$$y^{\wedge} = 50 + 20GPA + 0.07IQ + 0.01GPA * IQ$$

For females:

$$y^{\wedge} = 85 + 10GPA + 0.07IQ + 0.01GPA * IQ$$

Therefore, the condition of males earn more on average than females is $GPA \geq 3.5$. Which means the answer iii is correct.

(b) According to the least square line in (a), we can get:

$$y^{\wedge} = 85 + 10 * 4.0 + 0.07 * 110 + 0.01 * 4.0 * 110 = 137.1$$

So the predicted salary is 137100.

(c) False. We need to test the hypothesis $H_0: \beta_4 = 0$ and look at the p-value associated with the t or the F statistic to draw a conclusion.

4-a.

There is not enough information to tell which training RSS is lower between linear or cubic. However, because the X and Y are linear relationship, we expect the least squares line get close to the true regression line. Therefore, the RSS for the linear regression may be lower than the cubic regression.