Mark Gameng

HW 1

Part 1.1

1) Data mining task or not:
    a. No, dividing the customers of a company according to the gender is not a data mining task. It's simply going through customers and separating by gender.
    b. No, with similar reasoning to part a, but more complex as profitability might include some calculations of what the customer bought and the prices. But, it is still not a data mining task
    c. No, computing the total sales of a company is just a computation, not data mining.
    d. No, sorting a student database based on student id is just sorting or database query, not a data mining task.
    e. No, predicting the outcome of tossing a fair die is just a calculation using probability, not a data mining task.
    f. Yes, predicting the future stock price of a company using historical records is a data mining task. Can use a simple or complex model or techniques to predict future stock price. Regression.
    g. Yes, monitoring the heart rate of a patient for abnormalities is a data mining task. Can build a model of the behavior of a normal heart rate and raise an alarm if it doesn't follow the model, which is the abnormality. Anomaly detection.
    h. Yes, monitoring seismic waves for earthquake activities is a data mining task. Can build models of seismic waves behaviors which are connected to earthquakes and if current seismic waves are found to be following one of the models, then likely to be an earthquake incoming. Classification.
    i. No, extracting the frequencies of a sound wave is just using a tool and math to process and get the frequencies of the sound wave, not a data mining task.
3) Is data privacy an issue:
    a. Yes, data privacy is an issue for census data collected from 1900-1950. Census data holds information of where you are living and of the people who are living with you. However, there is a 72-year rule where records from the census are publicly accessible after 72 years. So currently, the latest public census records are from 1940.
    b. Yes, data privacy is an issue for IP addresses and visit times of users from your website. IP address can allow a person to know your ISP, and possible exploit or DDOS you. The visit times can also be used to learn about the user's behavior online and relate it to offline.
    c. No, data privacy is not an issue for images from satellites as they are not that high definition and doesn't really provide that much information that invades people's privacy, such as seeing them inside the house. Also, as far as I know, satellite images of earth are publicly accessible, at least some of them.
    d. No, data privacy is not an issue for name and addresses from telephone book, because anyone can get a telephone book.

e. Usually no, data privacy is not an issue for name and email addresses collected from the web. If it's collected from the web ethically, and were given permission or the user themselves made that information public, then data privacy is not an issue.

Part 1.2

2) Binary, discrete, or continuous. Qualitative or quantitative.
    a. Continuous and quantitative, interval.
    b. Continuous and quantitative, ratio.
    c. Discrete and qualitative, ordinal
    d. Continuous and quantitative, ratio
    e. Discrete and qualitative, ordinal
    f. Continuous and quantitative, ratio
    g. Discrete and quantitative, ratio
    h. Discrete and qualitative, nominal
    i. Discrete and qualitative, ordinal
    j. Discrete and qualitative, ordinal
    k. Continuous and quantitative, ratio
    l. Continuous and quantitative, ratio
    m. Discrete and qualitative, nominal.
3) Marketing story questions:
    a. The boss is correct, just counting the # of complaints per product is not a valid way of measuring customer satisfaction. Instead of just comparing the counts of customer complaints, it should instead be ratio of complaints and number of products sold. The less the ratio is, meaning less complaints in relation to # products sold, the higher the customer satisfaction for that product.
    b. The attribute type of the original product satisfaction and count of customer complaints is quantitative and ratio.
7) Daily temperature is more likely to show temporal autocorrelation because the amount of rain depends on the season and the location while temperature is relatively similar day to day wherever. You usually don't see temperature changing from really cold to really hot within days.
12) Noise and outliers questions
    a. Noise usually distorts values so its usually not interesting or desirable. However, outliers can be interesting or desirable as it can be a valid data that we are interested in.
    b. Noise objects can be outliers because noise can make the data seem random and thus there is a chance a noisy data appears as outliers.
    c. Noise objects are not always outliers, as they can appear as normal data.
    d. Outliers are not always noise objects; they can be objects of interests depending on the situation
    e. Yes, noise can make a typical value into an unusual one or vice versa.