Homework 2 (Exploratory Data Analysis)

1 Exercises

1.1.a What is EDA?

Exploratory Data Analysis (EDA) is a statistical approach used to analyze datasets by summarizing their main characteristics, often through visual methods. It helps to uncover patterns, spot anomalies, test hypotheses, and check assumptions before applying formal modeling techniques.

1.1.b Why is it important in data science and machine learning? (List 2-3 advantages of EDA.)

- Data understanding: EDA helps data scientists understand the structure, patterns, and distribution of data, which is crucial for identifying relevant relationships.
- Data cleaning: EDA allows for the detection of missing values, outliers, and errors, making it easier to clean the data and improve the quality of the models.
- Model building: By visualizing and exploring data, EDA can help in the selection of appropriate modeling techniques, enhancing model performance.

1.2.1 Discuss whether or not each of the following activities is a data mining task.

a) Dividing the customers of a company according to their gender.

No. This is not a data mining task. It's more about basic data categorization based on predefined attributes.

b) Dividing the customers of a company according to their profitability.

No. This is basic categorization based on their "profitability". It could be considered data mining if we were to predict a new customer's profitability based on trends.

c) Computing the total sales of a company.

No. This is not a data mining task. It involves straightforward aggregation of data rather than discovering patterns or insights from the data.

d) Sorting a student database based on student identification numbers.

No. Sorting is a basic data organization operation and doesn't involve discovering patterns.

e) Predicting the outcomes of tossing a (fair) pair of dice.

No. This is not a data mining task. Predicting outcomes for fair dice involves probability theory, which has already been solved by mathematicians and does not require us to make any new discoveries.

f) Predicting the future stock price of a company using historical records.

Yes. This is a data mining task. It involves using historical data to identify patterns to forecast future stock prices, which is a common application of data mining techniques.

g) Monitoring the heart rate of a patient for abnormalities.

Yes, This can be considered a data mining task as it involves analyzing the heart rate data to identify patterns or anomalies that indicate potential health issues. The goal is to classify normal from abnormal heart behavior

h) Monitoring seismic waves for earthquake activities.

Yes. We have to analyze seismic data to detect patterns and predict earthquake activities.

i) Extracting the frequencies of a sound wave.

No. It involves signal processing to determine the frequencies in a sound wave and doesn't require data mining.

1.2.3 For each of the following data sets, explain whether or not data privacy is an important issue.

a) Census data collected from 1900–1950.

No, Privacy is less of a concern for historical data, especially if anonymized, but ethical considerations still apply regarding its use.

b) IP addresses and visit times of web users who visit your website.

Yes, Privacy is important because IP addresses and visit patterns can potentially identify individuals or their behaviors.

c) Images from Earth-orbiting satellites.

No, Privacy concerns are minimal as the images generally capture public areas, though high-resolution images could still raise some issues if they reveal sensitive information.

d) Names and addresses of people from the telephone book.

No, The data is traditionally public. But handling this data requires caution to avoid misuse or unauthorized access.

e) Names and email addresses collected from the Web.

Yes, Privacy is crucial because this data is personally identifiable and can be misused for spam or other intrusive purposes.

1.3.2 Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

a) Time in terms of AM or PM.

Binary, qualitative, ordinal

b) Brightness as measured by a light meter.

Continuous, quantitative, ratio

c) Brightness as measured by people's judgments.

Discrete, qualitative, ordinal

d) Angles as measured in degrees between 0 and 360.

Continuous, quantitative, ratio

e) Bronze, Silver, and Gold medals as awarded at the Olympics.
Discrete,qualitative, ordinal
f) Height above sea level.
Continuous, quantitative, ratio
g) Number of patients in a hospital.
Discrete, quantitative, ratio
h) ISBN numbers for books. (Look up the format on the Web.)
Discrete, qualitative, nominal
i) Ability to pass light in terms of the following values: opaque, translucent, transparent.
Discrete, qualitative, ordinal
j) Military rank.
Discrete, qualitative, ordinal
k) Distance from the center of campus.
Continuous, quantitative, ratio
l) Density of a substance in grams per cubic centimeter.
Discrete, quantitative, ratio
m) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.
Discrete, qualitative, nominal

1.3.3 You are approached by the marketing director of a local company, who believes that he has devised a foolproof way to measure customer satisfaction. He explains his scheme as follows: "It's so simple that I can't believe that no one has thought of it before. I just keep track of the number of customer complaints for each product. I read in a data mining book that counts are ratio attributes, and so, my measure of product satisfaction must be a ratio attribute. But when I rated the products based on my new customer satisfaction measure and showed them to my boss, he told me that I had overlooked the obvious, and that my measure was worthless. I think that he was just mad because our best selling product had the worst satisfaction since it had the most complaints. Could you help me set him straight?"
a) Who is right, the marketing director or his boss? If you answered his boss, what would you do to fix the measure of satisfaction?
The boss is right. We can calculate a more accurate measure of satisfaction by normalizing the number of complaints. We can do this by dividing the number of complaints for the product by the total number of sales for the product. This would give a complaint rate, which is a more meaningful measure of customer satisfaction
b) What can you say about the attribute type of the original product satisfaction attribute?
The original product satisfaction attribute is a discrete, quantitative, ratio attribute, as it just counts the number of complaints. However, in this form, it is not an appropriate measure of customer satisfaction, as it doesn't take into account the overall context such as the product's popularity or the number of sales.

1.3.7 Which of the following quantities is likely to show more temporal autocorrelation: daily rainfall or daily temperature? Why?

Daily temperature is more likely to show stronger temporal autocorrelation than daily rainfall. This is because temperature tends to change gradually from one day to the next, especially due to weather patterns and seasonal cycles, creating a stronger correlation between consecutive days. In contrast, rainfall is more sporadic and can vary significantly from one day to the next, leading to less consistent autocorrelation.

1.3.12 Distinguish between noise and outliers. Be sure to consider the following questions.

a) Is noise ever interesting or desirable? Outliers?

Noise is generally not desirable, as it can obscure patterns and reduce the accuracy of data analysis or models. Outliers, on the other hand, can be interesting and desirable because they might reveal rare insights, anomalies, or exceptions in the data.

b) Can noise objects be outliers?

Yes, noise objects can sometimes be outliers. Random noise might create extreme values that deviate from the rest of the data, appearing as outliers.

c) Are noise objects always outliers?

No, noise objects are not necessarily outliers. Noise can be more subtle, they might introduce small random variations that do not necessarily result in extreme deviations from the norm.

d) Are outliers always noise objects?

No, outliers are not always noise objects. Some outliers may represent legitimate but rare phenomena, and distinguishing between outliers from noise and meaningful outliers is crucial.

e) Can noise make a typical value into an unusual one, or vice versa

Yes, noise can distort a typical value, pushing it into an unusual range, thereby making it appear as an outlier.