Task 1: review data mining concepts and tasks
Discuss whether each of the following activities is a data mining task

1. **Dividing the customers of a company according to their gender**
   Segregating customers based on gender can be considered as **information retrieval** from the database. The task here is **to look up** individual records of the customers for gender attributes. We are not discovering any novel information or patterns in the data. Hence, this cannot be a data mining task.

2. **Dividing the customers of a company according to their profitability**
   This is not a data mining task. The profitability of the customer can be calculated by retrieving required customer information from the database and determining whether the customer is profitable based on some calculation. This is a finance or accounting task. However, predicting who are the most profitable customers can be a data mining task.

3. **Computing total sales of a company**
   This is not a data mining task. Total sales is the value of all the products sold by the company over a time period. This is a finance or accounting task to report financial performance of the company. It does not require any kind of pattern discovery or prediction.

4. **Sorting a student database based on student identification numbers**
   This is not a data mining task. Sorting student records based on identification numbers can be done using a database query. The task here is to re-arrange the records. No prediction or pattern discovery is required in this task.

5. **Predicting the outcomes of tossing a (fair) pair of dice**
   This is a probability calculation task. Rolling a pair of fair dice can have 36 possible outcomes. To find the probability; nothing but predicting the outcome of rolling the dice, we divide the event frequency (1) by the size of the sample space (36), resulting in a probability of 1/36. Hence, this cannot be a data mining task.

6. **Predicting the future stock price of a company using historical records**
   This is a data mining task that falls under the category of predictive modeling of time series data. We can build a model with various attributes that affect the stock price and then forecast the future stock price of the company.

7. **Monitoring the heart rate of a patient for abnormalities**
   This is a data mining task that falls under the **Anomaly Detection** category. Anomaly detection technique can be applied to build a model for normal heart conditions. Every heartbeat is compared against the model. If a heartbeat is very different (outlier) from the normal conditions is observed, we can notify about the abnormal heart rate of the patient.

8. **Monitoring seismic waves for earthquake activities**
   This is a data mining task that falls under predictive modeling category: classification. We can build a model with different types of seismic waves characteristics that cause the

earthquake. when one of these different types of seismic activity is observed, we can notify about the earthquake in that region.

9. **Extracting the frequencies of a sound wave**
   This is not a data mining task. Frequencies of sound waves can be queried from the device or system that is collecting the sound wave information. This is **an information retrieval** or **look-up** kind of task.

10. **Suppose that you are employed as a data mining consultant for a fintech company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.**
    Fintech companies having a large user base will have access to all the investment details, transactions, and holdings of the customers. Financial data can be utilized to understand the overall portfolio which is used by financial advisors to help the clients to improve their financial situation. Data mining techniques can be used to analyze the data more efficiently, extrapolate the missing information, predict the expenditure, spending pattern, etc.,
    Clustering -
    clustering helps in building a diversified portfolio. Stocks that exhibit high correlations fall into one cluster, those slightly less correlated in another. So, different clusters will exhibit minimal correlation from one another. This way investors can reduce the total risk.
    Classification -
    Missing information can be extrapolated using classification technique. For example - Not all the transactions that are extracted from websites will have holding information. We can make use of classification to predict the holding details which is very crucial for financial advisors.
    Association Rule Mining -
    To precisely predict the price of a share and make profits has been always a challenging task. Association Rule Mining can help find the associations, correlations among stocks. It can discover all useful patterns from stock market dataset. This helps stockbrokers and investors to maximize their profits with each trade.
    Anomaly Detection -
    Anomaly detection can be used for fraud detection in financial transactions.

For each of the following data sets, explain whether data privacy is an important issue

1. **Census data collected from 1900 – 1950**
   Yes, census data contains information like age/date of birth, gender, income, etc., which some people might be uncomfortable sharing. Although we have laws that make the census data highly confidential, in case of a data breach, data security/privacy will be an important issue.

2. **IP addresses and visit times of web users who visit your website**
   Yes, tracking IP addresses are specific to devices. So, it can be traced back to an individual. It can tell you the city, ZIP code, of internet service provider. To some

extent, it can provide the physical address of the individual subscribed to ISP. In a way, IP addresses can be counted as personal data. They hold the ability to match the address to personal details. Hence, this information could present data privacy concerns.

3. **Images from earth orbiting satellites**
   Yes, Satellite photography can raise personal privacy concerns if the images are taken at high resolution to the extent of highlighting the individual house, vehicle license plates, etc.,

4. **Names and addresses of people from the telephone book**
   No, Telephone books may be used to exist a decade back. Everything is digitized now. whitepages.com is the digital phone book. It does not pose a serious threat to data privacy. A cellular operator can easily track the address of our handset at any time. But the only issue is scam calls which are quite annoying.

5. **Names and email addresses collected from the web**
   No, Names and email addresses also do not pose a very serious threat to data privacy. The only issue is with the spam or junk emails. They are quite annoying. It might have scam offers that can cost time and money.

Task 2: practice your critical thinking and writing

An NY Times article from 2014, Google Flu Trends: The Limits of Big Data, reviews Google's web service attempt to predict influenza by aggregating Google search results. The article discusses overestimation provided by the Google service over the interval 2011- 12 flu season compared to cases reported by CDC. Cases were overestimated by 50% during 2011- 12. Even after updating the algorithm cases were overestimated by 30% during 2013 - 14 flu season.

An article from the Atlantic, In Defense of Google Flu Trends, Article discusses combining Google Flu Trends with CDC data with few tweaks could have provided better predictions. It also quotes the documents where it was mentioned about using Google Flu predictions as a stand-alone source.

Issues faced by GFT should be looked at positively. Challenges faced by GFT would probably be addressed by recent developments in machine learning and artificial intelligence. We need to appreciate the attempt made by Google for providing better global health by predicting flu trends. More attempts like GFT need to be made and such projects should be funded by the government. If we had a similar service like GFT, we could have avoided pandemic like covid, and healthcare systems would have been better prepared to manage a global crisis.