# HW1

## 1. Exercises

## 1.1

## Q1)

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

A) No, that's not a data mining task. We can easily divide the customers with a SQL query or a conditional statement without using any techniques of data mining.

b) Dividing the customers of a company according to their prof itability?

A) No, that's not an data mining task. We can easily get the customer who is most profitable by sorting in the database, because we already have the data in the database. So, no need for using datamining techniques.

c) Computing the total sales of a company.

A) No, this doesn't come under data mining task. We doesn't use datamining techniques for computational purpose or problems.

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

A) No, this doesn't comes under datamining tasks. We can use an simple searching tool in the student data base to find the student details.

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

A) No, these kind of probabilistic problems are derived by mathematicians already. So, it doesn't comes under datamining task. If it wasn't an fair die then we may use datamining techniques to predict the outcome.

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

A) Yes, this is an data mining task. This comes under Predictive Modeling. Here we use the independent variables or historical data of a company to find the target variable or future stock. Here we uses the techniques for datamining.

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

A) Yes, this is an data mining task. This comes under anomaly analysis which is a field of datamining, where it monitors patient hart rate for abnormalities to give an alert, if patient is in severe condition it informs the person who is in charge.

h) Monitoring seismic waves for earthquake activities?

A) Yes, we can create an dataset with various kinds of seismic wave behaviors to know earthquake activities. This is an classification problem, we will raise an alarm whenever those seismic activities were observed, this is an anomaly detection which is a field of datamining.

i) Extracting the frequencies of a sound wave?

A) No, this doesn't comes under datamining tasks. We can seperate the sounds in one audio, but we can't extract them by using datamining technologies.

**Q3**)

a) Census data collected from 1900–1950?

A) No, these kind of data does not need privacy, this data can be used to see where the government services can be located. These data does not contain any individual private data, so there is no privacy issue.

b) IP addresses and visit times of web users who visit your website?
A) Yes, the IP addresses is an individual private data. So, this can used seen publicly by the other people. And tracking the user's web browsing timeline is an privacy issue.

c) Images from Earth-orbiting satellites?

A) No, these pictures should be available for everyone, then they get to know how they are revolving around the space and how the planets look like. And this also used as educational purpose for students.

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

A) No, this is not an issue of privacy. If we look for a person in telephone book, that means we already know his details. Otherwise, If a stranger finds your address also he can't do anything with that data.

e) Names and email addresses collected from the Web?

A) No, just by knowing someone's email and name, we cant get to know all their details. With this few details, someone cant bother them. If he try's to do fishing attacks, the person should be carefully while opening mails and put them spam if harmful.

**1.2**

**2**)

a) Time in terms of AM or PM?

A) Binary, Qualitative, Ordinal.

b) Brightness as measured by a light meter?

A) Continuous, Quantitative, Ratio.

c) Brightness as measured by people's judgments?

A) Discrete, Qualitative, Ordinal.

d) Angles as measured in degrees between $0^\circ$ and $360^\circ$?

A) Continuous, Quantitative, Ratio.

e) Bronze, Silver, and Gold medals as awarded at the Olympics?

A) Discrete, Qualitative, Ordinal.

f) Height above sea level?

A) Continuous, Quantitative, Interval/Ratio: This is based or depends on the sea level they had considered on that specific region, here we need more details to get to better understanding and conclusion.

g) Number of patients in a hospital?

A) Discrete, Quantitative, Ratio.

h) ISBN numbers for books?

A) Discrete, Qualitative, Nominal.

i) Ability to pass light in terms of the following values: opaque, translucent, transparent?

A) Discrete, Qualitative, Ordinal.

j) Military rank?

A) Discrete, Qualitative, Ordinal.

k) Distance from the center of campus?

A) Continuous, Quantitative, Interval/Ratio.

(It's based on the university. And by time being the university may expand its size by constructing or including more departments, then the geographical area increases. So, then the ratio/length of the university also increases.

l) Density of a substance in grams per cubic centimeter?

A) Discrete, Quantitative, Ratio.

m) Coat check number?

A) Discrete, Qualitative, nominal.

**3**)

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?

A) Here the Boss is correct.

Before Fixing:

Example1:

Consider, there are two products. If the two products was been purchased by 100 customers and they both got equal no of complaints like product1 got 10 and product2 got 10, does that mean they both are considered not satisfied

Example2:

Consider, there are two products. If they two products were been purchased by customers and they got unequal complaints like, if product1 had got 10 and the product 2 got 4 complaints. Does that mean product1 is not satisfied.

After Fixing:

Purchase= Total no of complaints / Total no of purchases

Example1: Product1: Phone

Total Purchases: 100

Total Complaints: 10

Product2: Laptop

Total Purchases: 20

Total Complaints: 10

Satisfaction1= 10/100

Satisfaction2 =10/20

Here, we got 90% satisfaction for the product1 and the product2 got 50% satisfaction.

Example 2: Product: Phone1

Total Purchases: 100

Total Complaints:10

Satisfaction1=10/100

Product: Phone2

Total purchases:20

Total Complaints:5

Satisfaction2=5/20

Here, we got 90% percent satisfaction for the product1 and the 25% Satisfaction for the product2 in the purchases.

Therefore, the ratios are the better way to calculate these kind of problems.

b) What can you say about the attribute type of the original product satisfaction attribute?

A) The original product satisfaction attribute is based on the situation\

Example1:

Product : P1

Total products purchased : 2

Total no.of Complaints : 1

Satisfaction = 1/2

Here, the 50% satisfaction of the customer for P1.

Example2:

Product :P2

Total products purchased: 100

Total no.of complaints :50.

Satisfaction = 50/100

Here, is also the same 50% satisfaction.

Therefore, the product with same satisfaction rate may differ in the no.of complaints raised and purchase rate and vice-versa.

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

Temporal autocorrelation refers to the relationship between consecutive values of the same attribute. The daily temperature shows more auto correlation than the daily rainfall. It will have more chances for same temperature with the physically close areas, than places farther away. The rainfall will not be the same in different areas, the rainfall will not be same as temperature, and like it doesn't cover much geographical area it had been restricted to some portion of area. So, the daily temperature has the more auto correlation than the daily rainfall.

**12**.

a) Is noise ever interesting or desirable? Outliers?

A) No, the noise is not the original data. It distracts the linear line by pretending as an original data and disturbs the original series of data. The outliner is an desirable legitimate value of data, but it's hard to find those data values.

b) Can noise objects be outliers?

A) Yes, the noise is which disturbs the series of data, the outliers are will be present little far from the decision boundary. Sometimes the outliers looks like the noise, they difficult to know is the value is legitimate.

c) Are noise objects always outliers?

A) No, the values of the dataset are included with the noise. The noise will be present near to the decision boundary, so it also touches the noise. But the outliers are the usual values far from boundary, this is the task of datamining to figure it out.

d) Are outliers always noise objects?

A) No, the outliers are the legitimate values of dataset which are not noisy data. If we try to fit the data which contains outliers with linear regression then it will not fit properly, so we use polynomial regression to fit the values of the dataset.

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

A) Yes, as we known that the noise is an unusual data. By this noise the data will be disturbed and some values seem as an unusual data values.

**2**. Programming Problem

2.1-iv-a

We cannot plot the data in that format, it should be in the numerical form. So, it shows the warning if we doesn’t remove those both columns, then its difficult to plot. But those are the columns are unique, which are independent columns.

2.1-iv-b

The highest correlation found is 0.97, it is between the "total_cases" and the "total_deaths".That means, the maximum people are died who are reported as covid.Here there is hight relationship between those both..

2.1-iv-c

The lowest correlation found is 0.0029, it is between the "cases_last_7_days" and "cases_ratio_per_100k".Here, this means the people who were reported as having covid in past 7 days are very less compared to the ratio of 100k.

2.1-v-b

The anomaly is the data which doesnt comes under the normal pattern. these kind of anomaly data also known as outliers in some cases, the data shown in the above graph shows some anomalies, which are located far from the normal data and looks unusual.The main focus of anomaly detection is to discover these kind of anomaly data, to know either they are actual data or not.When we calculate the Threshold value of the data,then we will calculate the distance of each data, now if we found a data which is grater than the threshold then it is an outlier or anomaly.