1. Introduction to Data Mining: Exercises 1.7 – Page 21: Question #3  
   **Explain whether data privacy is an issue for:**
   * Census Data collected from 1900-1950: No, because census data is public record and voluntarily disclosed to the census bureau, which then publishes it.
   * IP Addresses and visit times of users who visit your website: Yes, because it can reveal intimate details about the user and their location.
   * Images from earth-orbiting satellites – no, generally speaking satellite images don’t constitute a privacy breach because most of the images taken are public property and legally considered “fair game”.
   * Names and addresses of people from the telephone book – No; on one hand, the telephone book is public record, and anybody can view the contact information for both businesses and individuals in it, but people may still feel uncomfortable being contacted based on that information. Skip tracing in the debt collection industry to verify the place of employment of an individual often uses public access telephone book data to find ways to contact a debtor.
   * Names and email addresses collected on the web: **No –** generally speaking email addresses are either purchased from the email provider or given voluntarily through newsletters, signup forms, etc.
2. Introduction to Data Mining: Exercises 2.6 – Page 105-112: Questions #2 & #16

**1. Question 2**

a. Time in terms of AM or PM. – **Discrete, Quantitative, Interval**

b. Brightness as measured by a light meter. – **Continuous, Quantitative, Ratio**

c. Brightness as measured by people’s judgments.- **Discrete, Qualitative, Nominal**

d. Angles as measured in degrees between 0 and 360. – **Continuous, Qualitative, 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. - **Binary, Qualitative, Nominal**

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. – **Continuous, Quantitative, Ratio  
*This is a ratio because a ratio has a meaningful zero point by definition.  
Density = M/V, and research has confirmed the existence of a massless particle, making zero-density substances possible.***

m. Coat check number – **Discrete, Qualitative, Nominal**

**2. Question 16**

**Part a:** If the term occurs in one document only, we would see a high inverse frequency. Similarly, if the term occurred in each document, the inverse frequency converges on zero.

**Part b:** the inverse frequency is useful in identifying documents that focus on a specific term or concept – when the term occurs in one document, a high inverse frequency suggests that a term might be associated with a concentrated subject matter. The converse of this is when the inverse frequency is zero, so a term is non-unique in helping us narrow our scope of search for a given term of interest.

1. **Data Science Using Python and R – Chapter 12 – Questions 1,2,3,4,10**

**1.** What do we mean by high dimensionality in data science?

High dimensionality refers to the phenomenon of a dataset having many different variables (or columns) for any given row.

**2.** Why do we need dimension reduction methods?

Dimension reduction methods help increase the accuracy of predictive analytical models by getting rid of collinearity, reducing the amount of processing power needed to create computations, and preventing the problem of overfitting.

**3.** What does principal components replace the original set of *m* predictors with?

PCA replaces the original set *m* with a linear uncorrelated combination of variables called components that account for some of the variance of *m*, with each subsequent component accounting for less of the variation.

**4.** Which principal component accounts for the most variability?

The first component.

**10.** When we use the principal components as predictors in our model, what value do the VIFs take? What does this mean?

VIFs for principal components converge onto 1, the lowest possible value, meaning that there is no multicollinearity in our predictors/model.