Question 1:

Identify two data sets that are publicly available online and answer the following questions for each of the two data sets:

(a) URL of the webpage to access the data set.

(b) Brief description of the data set, including number of objects, number of attributes, attribute types, and other relevant information.

(c) What knowledge may be mined from this data set?

(d) How would the knowledge be useful in some applications?

Please make sure your answers are in the same order as the questions with at least one newline after each answer.

Your Answer:

Data set 1: a. https://www.kaggle.com/datasets/drgilermo/nba-players-stats

Links to an external site.

b. The data set has 3 csv files. One has a collection of the 3922 NBA players since 1950 with eight total attributes with nominal categorical variables such as college attended and continuous numeric ratio-scale variables such as weight. Another has the stats for each player in each season they played in the NBA containing 24,700 total player seasons and 53 total attributes, with name being the only categorical variable and some cumulative numeric discrete ratio-scale variables like points and some calculated rate statistics that are continuous numeric and interval-scale like Player Efficiency Rating. The third csv file contains one entry per player and eight total attributes like the first csv file but contains more players (4500) and additional variables such as year start and year end which are discrete, numeric and interval scale along with position which is nominal and categorical.

c. Trends in player performance may be mined from this data set such as which positions tend to score the most points or grab the most rebounds and how those associations have changed over time.

d. The knowledge that could be mined from this data set could help teams make better judgments of a player's value or help fans gain a better understanding of the game.

Dataset 2: a. https://www.commerce.alaska.gov/web/aogcc/Data.aspx#dataminer

Links to an external site.

(Third link under Data Miner")

b. There are 12,894,352 total rows in this data set, each of which represent a month of production for a well in this data set. There are 16 total attributes. Wells are uniquely identified by their API number. The data set contains nominal categorical variables such as well name and operator name that describe the well and continuous numeric ratio-scale variables that represent the total oil, gas and water produced in a given month.

c. Trends such as which areas or fields tend to have the highest producing wells could be mined from this data set.

d. The knowledge that could be m mined from this data set could be useful for an oil company in determining the best sites in which to drill a new well.