Assignment 2

Risk Factor for Cardiovascular Heart Disease

This CSV data set consists of 70000 patients medical history which could be used to determine the risk factors for heart disease

id	Consists of unique identification number given to each individual (int)
age	Age of patients in days (int)
gender	Gender of patients (string)
height	Height of patients in centimeters (int)
weight	Weight of individuals in kilogram (float)
ap_hi	Systolic blood pressure of patients (int)
ap_lo	Diastolic blood pressure of patients (int)
cholesterol	Total cholesterol level read as mg/dl (int)
gluc	Glucose level read as mmol/l (int)
smoke	represented in the form of 1 if the individual smoke and 0 if they do not smoke(int)
alco	represented in the form of 1 if the individual drink alcohol and 0 if they do not drink alcohol (int)
active	consists of 0 if the individual is not physically active else 1 (int)
cardio	consists of 0 if the individual is not suffering from cardiovascular disease else 1 (int)

This data set still has further modifications to be done to interpret data correctly. Age column is given in days but the age of a person is easier to convey when referred to in years by dividing it by 365. Age of patients has continuous data from 40 – 64 years hence it would be better to drop rows that are beyond this range as the amount of data related to people less than 40 years is very less. Weight of patients has been recorded in kilograms which needs to be converted to lbs or pounds i.e. measuring metric has to be changed for weight column. The given dataset does not consist of any null values in any of the cells hence there is no necessity for dropping the rows. Cholesterol has a total cholesterol level of patients on a scale of 0-5 where there is 20 mg/dl increase or decrease for each unit. Gluc has glucose level of patients on a scale of 0-16 where there is 2 mmol/l increase or decrease for each unit. Smoke, alco, active and cardio indicates 1 if they are actively performing the action else 0 if they are not.

Three different side by side graph can be drawn to know the average cholesterol level of male and female of different age groups who actively smoke, drink and those who are physically active. This gives a better comparison to check on different factors and making note of peaks in each graph and analyzing those groups could provide more valuable information. A pie chart can be drawn to state how many of the 70000 patients have suffered from cardiovascular disease so that viewers can understand the risk and factors can be considered. A box lot could be generated

for different age groups using the systolic and diastolic blood pressure to know their regular blood pressure.
https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas

Hotel Reservations

Booking_ID	Unique number for identifying bookings (int)
no_of_adults	No of adults who would be staying at the hotel (int)
no_of_children	No of children who would be staying at the hotel (int)
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the room has been booked (int)
no_of_week_nights	Number of week nights (Monday to Friday) the room has been booked (int)
type_of_meal_plan	Meal plan chosen by the customer (string)
required_car_parking_space	1 if customer requires parking space else 0 (int)
room_type_reserved	room type reserved by the customer (int)
lead_time	Date difference between booking date and customer arrival date (int)
arrival_year	Year of customer arrival (int)
arrival_month	Month of customer arrival (int)
arrival_date	Date of customer arrival (int)
market_segment_type	Mode of making the booking (string)
repeated_guest	1 if customer has already visited the hotel else 0 (int)
no_of_previous_cancellations	No of cancellations done by the customer before confirming his choice (int)
no_of_previous_bookings_not_canceled	No of cancellations not done by the customer before confirming his choice (int)
avg_price_per_room	Price per room is varying for seasons hence average price has considered (float)
no_of_special_requests	Special requests made by the customers (int)
booking_status	Cancelled if bookings have cancelled or else not cancelled (string)

The CSV dataset consists of 36276 online reservations made by customers from July' 2017 – December' 2018. As no blanks have been reported in any rows there are no bad records that should be removed from the dataset. Average price per room is recorded in Euros which needs to be converted to dollars.

A line graph can be drawn to see the average bookings done in each month and also another line in different colour considering the cancelled rows in booking status. This graph would show two

different lines one indicating bookings and the other cancellations. Bar can also be drawn. A pie chart can be drawn for the market segment to know which segment has more bookings such that the hotel can look forward to those segments. The peaks in line graph indicate high bookings interpreting that there could be high demand for the hotel in those months. A visual representation can be made to check the common number of adults and children booking the hotel such that special offers can be made for attraction. A heat map could be drawn to check in which months the hotel has received high guests.
https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset

Video Game Sales

Rank	Rank on basis of sales (int)
Name	Name of the game (string)
Platform	Platform in which the game has been released (string)
Year	Year in which the game has been released (int)
Genre	Genre of the game (string)
Publisher	Publisher of the game (string)
NA_Sales	Sales in North America (float)
EU_Sales	Sales in Europe (float)
JP_Sales	Sales in Japan (float)
Other_Sales	Sales in the parts of the world (float)
Global_Sales	Worldwide sales (float)

The CSV dataset consists of sales of video games in different parts of the world from 1980 to 2020. The dataset has a total of 16599 records of which 272 records do not have year mentioned to them. It is best to delete these records from the dataset as the year is not confirmed. It could be difficult when analysis is being made as these set of records form a different group. The EU sales and JP sales consist of sales in euros and yen respectively. To maintain all the data in the same format EU and JP sales need to be converted to Dollars.

The given dataset consists of past 40 years data so plotting could be quite difficult while viewing. So by considering past 20 years data instead 0f 40 years and grouping every 2 years the information can be categorised and would be much easier while depicting the visuals when considering yearly analysis. Bar graphs can be plotted to check the video game which had the highest sales over every 2 years. Bar graphs can be plotted to see which genre has been sold out the most with genre on x axis and highest total sale on y axis. This helps us to know the genre that people are most interested in or to which genre people have been shifting to frequently. A pie chart can be drawn to check which part of the world has made high sales, determining the high chance of sales in that region. A stacked chart can be drawn by mentioning the sales each genre has created in 2 years. This shows the increase in popularity among players by year. A heat map could also be generated to check which area has had the highest sales for the past 20 years.

https://www.kaggle.com/datasets/gregorut/videogamesales