

Question 1

Assume that your company operates a loan marketplace where individuals seeking to borrow money are matched with suitable loan products offered by various banks. The data schemas are displayed below:

BANKS

Column name	Data type	Description
bank_id	int	Primary key
bank_name	string	Examples: "HSBC", "Ocean Bank",...

PRODUCTS

Column name	Data type	Description
product_id	int	Primary key
product_name	string	The name of loan product
loan_amount	int	The currency unit is USD.
interest_rate	float	
accepted_risk_level	string	"low" / "medium" / "high"
bank_id	int	The bank that provides this product
created_date	DateTime	

CUSTOMERS

Column name	Data type	Description
customer_id	int	Primary key
customer_name	string	Example value: "Morgan Freeman"
customer_age	int	Example value: 45
estimated_risk_level	string	"low" / "medium" / "high"
source	string	The source that brings this customer to the marketplace
created_date	DateTime	

LEADS

Column name	Data type	Description
customer_id	int	
product_id	int	
apply_date	DateTime	
latest_status	string	"pending" / "approved" / "rejected"

A customer can apply for no products, one product, or multiple products, with each application considered a lead. A customer with an estimated risk level X can be matched with products that accept risk level X or higher.

Based on the above data tables, please write SQL queries to:

- a) **Portfolio management:** Display the cumulative approval rate by month for each source in 2023.
- b) **Marketing:** Display the names of customers who applied in 2023 and are older than 92.5% of all leads who applied in 2022.
- c) **Rollover detection:** Display the names of customers who applied 5 times consecutively, with a 25-35 days gap between any two continuous application dates, the loan amount is increasing (each subsequent application being larger than the previous one), and all 5 banks applied being different.
- d) **Create a user-defined SQL function** that takes a customer_id as input. The output should be a message congratulating the customer if they can apply for the largest possible loan amount from a bank name, or a message expressing regret if no suitable loan product is found. Ensure that A customer with a certain risk level cannot be matched with products that accept a lower risk level.

Output example:

"Congratulations, You can apply for a loan with the amount of 10,000 USD from HSBC bank."

Or

"Sorry, we cannot find any suitable loan products for you at this time."

- e) **The race of products:** For each month in the year 2023, identify and display the following information:
 - The month (formatted as YYYY-MM).
 - The number of unique loan products that had at least 100,000 applications in that month and in every previous month of 2023 ("good" product).
 - The name of the "best" loan product for the month.
 - The number of applications for the "best" loan product.

The "best" product must be a "good" product and have the highest number of applications for the month. If there are more than one product has maximum number of applications, the product with the lower proportion of estimated 'high' risk level applicants is considered better. Output example:

Month	No. good products	Best product	Maximum applications
2023-01	20	TPB01	565,048
2023-02	18	VIB02	532,647
2023-03	10	TPB01	416,990
2023-04	9	HSBC03	388,115
2023-05	4	TPB01	625,759
...
2023-12	1	TPB01	798,202

Question 2

You are a Data Analyst for a company which produces a new generation of electric men razor. Your company registered an e-commerce site at www.Coolmen-Coolrazors.com 1 month ago to sell its product online instead of the traditional supermarket channel. During the last month, it piloted advertising on 2 channels:

- Email Channel
- SMS Channel

Data is extracted from a centralized database and stored in the attached file called "Q2_mkt_data.csv".

Dataset

The schema for this dataset is as follow:

id	Format: Integer, representing each message
send_date	Format: data, date when SMS/Email was sent
estimated_age	Format: Integer, ranging from 0 to 100
age_range	Format: string. The audience is divided into 4 age ranges
channel	Format: string, either SMS or Email
coupon	Format: float, the value of coupon expressed in each message, valid for up to 3 units for each order
clicked	Format: binary, either 0 (customer doesn't click on the link in SMS/Email) or 1 (they clicked)
last_step	Format: string. It can have one of the following values: "received", "bounced", "saw review", "added to cart", "payment page", "purchased"
nb_units	Format: integer, representing the number of units of customers' order.
order_value	Format: float, representing the value of the order the customer made. Already minus the coupon applied.

The column "last_step" is the final point of contact with customers before they leave our website. Its values are explained below:

- Received: sms/email sent successfully, but not clicked.
- Bounced: they clicked but exited immediately.
- Saw review: scroll down and read the review and information of the product
- Added to cart: customers added the product to cart to check out

- Payment page: They stopped at payment without finishing it
- Purchased: They made an order

Financial Information

Together with the data above, you have additional information about the production cost and the marketing campaigns.

- The production cost for each razor is 18\$.
- Cost per one SMS is \$0.050, cost per one email sent is \$0.075.
- Each email or SMS will be supplied with a coupon that can have a value of 2\$, 4\$ or 6\$. The coupon is valid for up to 3 razors per order. They have the option to wrap the items as a gift. Ignore wrapping and shipping costs.
- The price without coupons is 40\$ / razor.
- From experience (and some models), potential customers are divided into 4 age groups:
 - 18 - 30
 - 31 - 45
 - 46 - 60
 - 60 +

Question

2.a. For the next quarter, your marketing department has a budget of \$60,000 to spend on online campaigns. How would you allocate it between SMS and Email? Assume that we have a potential customer pool for each age group as below:

Age Group	Pool size
18 - 30	300,000
31 - 45	350,000
46 - 60	500,000
60+	200,000

2.b. Now assume that you are also responsible for the operation of the company's website. Do you have any comments or suggestions so that we can improve the website's performance in order to maximize net profit?

Question 3 [Optional]

As a Data Analyst for a company specializing in electric men's razors, your task is to analyze a week's worth of logs from conversations between the AI chatbot and customers. The chatbot is responsible for onboarding customers, selling products, and managing relationships.

The provided data ("Q3_data.csv" included in this test kit) contains a timestamp column and the messy raw logs of conversations between the chatbot and customers. The timestamp column has several different formats, and the `customer_satisfaction` column is missing values in many rows.

Your task is to write functions (Python/Excel is preferred) to reformat the timestamp column and complete the `customer_satisfaction` column. Then, create a visualization that displays the Satisfaction Rate by date.