TRUSTING SOCIAL

DATA ANALYST SCREENING TEST

July 2019

- You may write your answers in Vietnamese or English or a mix of both languages.
- You may consult textbooks and printed and online materials.
- Discussion or sharing of test content or answers during or after the test is prohibited. Failure to comply with this rule, if uncovered by the Trusting Social, may result in voiding of your test results.
- Please show all of your work. Answers without appropriate justification will receive very little credit.
- If you scan or photograph your answers, please make sure the scans/photos are readable. On the other hand, code should be submitted as .py or .r files.

This screening test consists of:

1. Question 1: Writing SQL queries (35 points)
2. Question 2: Business case (35 points)
3. Question 3: Programming skill (30 points)

Question 1

Assuming that your company is running a loan marketplace where people who want to borrow money are matched with appropriate loan products provided by different banks, the data schemas are shown below:

Banks

Column name	Data type	Notes
bank_id	int	Primary key
bank_name	string	Examples: “HSBC”, “Ocean Bank”,...

Products

Column name	Data type	Notes
product_id	int	Primary key
loan_amount	int	Currency unit is USD. Example value: 1000
interest_rate	float	
accepted_risk_level	string	“low” / “medium” / “high”
bank_id	int	Bank that provides this product
created_date	datetime	

Customers

Column name	Data type	Notes
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	Source that brings this customer to the marketplace
created_date	datetime	

Leads

Column name	Data type	Notes
customer_id	int	
product_id	int	
apply_date	datetime	

A customer with estimated risk level X will only be matched with a product that accepts risk level X.

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

a) Show the number of products available for each accepted risk level.

b) Show the average interest rate of products provided by HSBC bank.

c) Show 2 banks that have most high risk products.

d) Show which source brings to the marketplace more low risk customers.

e) Show all months of the year 2017 that the number of customers applying for loans are 20% higher than the monthly average number of customers of the year.

f) Show the names of all leads who applied in 2017 and are older than 90% of all leads who applied in 2016

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 are extracted from a centralized database and stored in the attached file called "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. 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 value of the order 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 no 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.

- Production cost for each razor is 18$.
- Cost per one SMS is $0.050, cost per one email sent is $0.075.
- Each email or SMS will be supplied a coupon which can have value of 2$, 4$ or 6$. Coupon is valid for up to 3 razors in each order. They have the option to wrap the items as gift. Ignore wrapping and shipping costs.
- The price without coupon is 40$ / 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 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?

Suggestions

Please note that this case study is very close to what we are doing as Data Analysts at Trusting Social in spirit. We would like to know more about your problem-solving skill and the skill to gain insights from data, but we do not require a perfect answer. Do the best you can (with your critical thinking, R, Python or your programming language of choice), and Good luck!

Question 3

Please find attached the file “messy.xlsx”, use R, Python or your programming language of choice to do the following:

- Clean the names of columns to lowercase separated by “_”, remove any empty column if necessary.
- Change the date column to the same format ‘YYYY-MM-DD’.
- Change the name column to the title case (e.g: Jason Mraz).
- Make a new “email” column with form: {last_name}.{first_name}.{id}@yourcompany.com
- Change the phone number column to the format “84……”
- Find any duplicated ID and remove those who join later.
- Filter those who join since 2019 and export to a csv file, delimited by “|”, file name “emp_{report_date}.csv” with report_date = today.