SQL Question 1:
Please write SQL that would provide results for each question. Give meaningful names to output columns.
a) For each employee, find the name of their department, including employees that do not belong to any department.

```
SELECT "Employees"."id" AS "Employee Id", "Employees"."name" AS "Employee name", "Departments"."name" AS "Department Name" FROM { OJ "Departments" RIGHT OUTER JOIN "Employees" ON "Departments"."id" = "Employees"."Dep_id" } ORDER BY "Employee Id" ASC
```

Result

Employee Id	Employee name	Department Name
1	John Smith	Finance
2	Jack Smith	
3	Becky Smith	Finance
4	Rebecca Smith	Operations
5	Sonny Smith	Deployment

b) Find the highest salary per department, and order the departments by this (highest first).

```
SELECT "Departments"."name" AS "Department Name", MAX( "Employees"."Salary" ) AS "Highest Salary" FROM "Employees", "Departments" WHERE "Employees"."Dep_id" = "Departments"."id" GROUP BY "Departments"."name", "Departments"."id"
```

Result

Department Name	Highest Salary
Finance	2000
Operations	700
Deployment	3000

c) Calculate the average team salary per manager. Note that Employees.Manager_id is a foreign key to Employees.id

```
SELECT "Manager_id", AVG( "Salary" ) FROM "Employees" GROUP BY "Manager_id" HAVING ( ( "Manager_id" IS NOT NULL ) )
```

Result

Manager_id	Average Team Salary
1	2250
2	1350

SQL Question 2:

Write SQL queries for the following questions:
a) Revenue for a specific purchase is: items ×× price. Calculate the sum of revenue for female users.

SELECT SUM("items" * "price") AS "revenue for female users" FROM "Purchases" WHERE LEFT(LCASE("user_gender"), 1) = 'f' GROUP BY LEFT(LCASE("user_gender"), 1)

Result

revenue for female users
500

b) Calculate the sum of items, grouped by gender (female vs. male).

In 2 successive queries:
;WITH “Query 2 b 1” AS (
SELECT CASE WHEN LEFT(LCASE("user_gender"), 1) = 'f' THEN 'female' ELSE 'male' END AS "Gender", "items" FROM "Purchases")

SELECT "Gender", SUM("items") AS "sum of items" FROM "Query 2 b 1" GROUP BY "Gender"

Result

Gender	sum of items
female	5
male	2

c) Count the number of (distinct) female users who have purchased more than 3 items in total.

In 2 successive queries:
; WITH “Query 2 c 1” AS (
SELECT "user_id", CASE WHEN LEFT(LCASE("user_gender"), 1) = 'f' THEN 'female' ELSE 'male' END AS "Gender", SUM("items") AS "sum of items" FROM "Purchases" GROUP BY "user_id")

SELECT COUNT("user_id") AS "number of female users with 4 items purchased mini" FROM "Query 2 c 1" GROUP BY "Gender", "sum of items" HAVING (("Gender" = 'female' AND "sum of items" > 3))

Result

number of female users with 4 items purchased mini
1

d) Find the ids of the top 3 highest male spenders.

```
SELECT TOP 3 "user_id", SUM( "items" * "price" ) AS "Revenue" FROM "Purchases" WHERE
CASE WHEN LEFT( LCASE( "user_gender" ), 1 ) = 'f' THEN 'female' ELSE 'male' END = 'male'
GROUP BY "user_id" ORDER BY "Revenue" DESC
```

Result

user_id	Revenue
105	200
103	0

SQL Question 3:

Each user (user_id) has a set of transactions (transaction_id) with associated time stamps (transaction_ts). Please write the most optimum SQL statement to:

a) Select the first item each user purchased.

```
; WITH “Query 3 a 1” AS (
SELECT "user_id", MIN( "transaction_ts" ) AS "earliest ts" FROM "Transactions" GROUP BY "user_id")

SELECT "Transactions"."user_id", "Transactions"."item" AS "First Item Purchased" FROM "Transactions",
"Query 3 a 1" WHERE "Transactions"."user_id" = "Query 3 a 1"."user_id" AND
"Transactions"."transaction_ts" = "Query 3 a 1"."earliest ts"
```

Result

user_id	First Item Purchased
13811335	glove
3378024101	dress

b) Calculate how many transactions a user has made 72 hours since their first transaction.

```
; WITH “Query 3 a 1” AS (
SELECT "user_id", MIN( "transaction_ts" ) AS "earliest ts" FROM "Transactions" GROUP BY "user_id")

SELECT "Transactions"."user_id" AS "User", COUNT( "Transactions"."transaction_id" ) AS "Number of
transaction  within 3 days" FROM "Transactions", "Query 3 a 1" WHERE "Transactions"."user_id" = "Query
3 a 1"."user_id" AND DATEDIFF( 'hh', "Query 3 a 1"."earliest ts", "Transactions"."transaction_ts" ) < 72
GROUP BY "Transactions"."user_id"
```

Result

User	Number of transaction within 3 days
13811335	6
3378024101	3

Case Study Question 1

You are in charge of a weekly data blog for Whatsapp. Where do you look for interesting facts and how would you go about it?

I would start with a weekly dashboard containing the following metrics and their latest variations:

– volume of data exchanged overall and by country
– volumes of text, picture & video data
– call volumes & average quality associated
– video volumes & average quality associated

Any significant changes would be investigated and the cause isolated and reported in the blog. Seasonalities would be interesting to highlight: hourly & daily variations for example.

Beyond this, I would try to measure the commitment of users to the app:

- lag between outgoing messages (average and distribution)
- lag between outgoing messages with no answers from counterparty (average and distribution)
- reply lag to receiving a message (average and distribution)

I would also try to measure the intensity of interaction of the user with people (number of messages per unit of time (min, max, average, daily, hourly, weekly and media used) and determine if there are defined clusters of friends in those.

I would be interested in understanding the mechanics behind the splitting of long messages: do users split long messages systematically or just with specific people. I would measure message length, cumulative message length (unanswered successive outgoing messages) and determine if there are significant differences amongst the aforementioned clusters of friends.

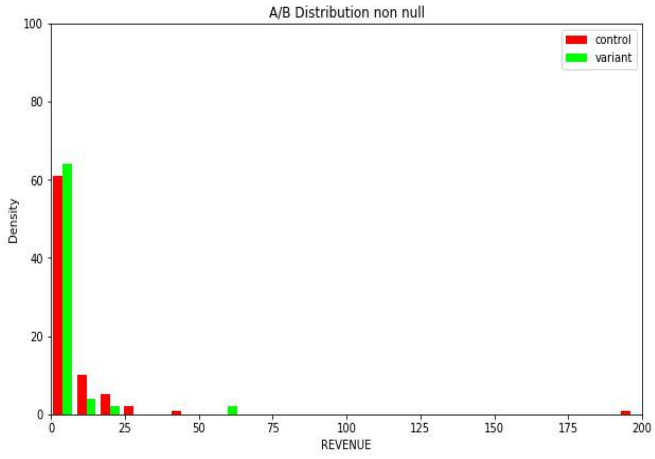
Case Study Question 2

An A/B test was run on a website with a goal to increase revenue generated. The attached Excel file shows raw data on user_id, test variant and revenue generated during the test. You have been asked to analyse the results and provide an update and recommendations to the Product Manager.
Please do so whilst explaining any techniques and tools you would use in order to approach this. Please provide working in Python.

We want to test the difference in revenue per visitor. The usual Student Test is not appropriate and its main assumption violated: the distribution of revenue per visitor does not follow a normal distribution.
The dataset is disproportionally populated by visitors who do not order and contribute a large percentage of zero values to the overall distribution of revenue per visitor.

Basic stats for the dataset					
	mean	std	**count**	min	max
control	0.13	3	**4984**	0	196
variant	0.07	1.3	**5016**	0	58.63
Basic stats for the dataset without null revenue					
	mean	std	**count**	min	max
control	8.03	22.5	**80**	0.02	196
variant	4.88	9.9	**72**	0.02	58.63

Also, revenue values have a lower bound of zero as orders can't be made for less. But revenues are unbounded above. The distribution of revenue values is right-skewed, making t-tests a sub-optimal choice for evaluating differences in revenue between a control and variant.



We shouldn't use a non-linear transformation to normalise the data as whilst such a transformation can normalise right-skewed data, it cannot normalise data with a large number of zero values.

My approach is to employ a non-parametric technique to test for differences in revenue: the Mann-Whitney-U test. This is a rank-based approach that takes all revenue values of the control and variant, combines them into one vector, orders them, separating the vectors apart again and testing the difference of the ordered values.

As the distribution of the revenue graph shows: we are in the presence of an outlier and the analysis will be performed in parallel excluding the data point to compare results

I have decided to test the null hypothesis Ho = the control ranks are lower than the variant ranks.

The p-value calculated on the Mann-Whitney-U statistic on the full data set is 0.76
Excluding the outliers it only marginally lowers to 0.73
As a check, excluding the non purchasing website visitors produces higher p-values

Clearly there is weak evidence against the null hypothesis and we fail to reject it.

The recommendation to the Product Manager is therefore to keep the control version of the website.

Mann Whitnet Wilcoxon
Full DataSet

U Statistic	p-Value
12521564	0.76

Mann Whitnet Wilcoxon
DataSet non null revenue

U Statistic	p-Value
3356	0.96

Mann Whitnet Wilcoxon
Full DataSet excluding outliers

U Statistic	p-Value
12516548	0.74

Mann Whitnet Wilcoxon
DataSet non null revenue excluding outliers

U Statistic	p-Value
3284	0.95

Case Study Question 3

Badoo users are prompted to rate the app at random times during a session. Event level data on user activity is available. You have been asked to review and improve this project. Please provide a few key points on your plan to solve this challenge.
You may choose your own focus area but sample topics include data preparation, any machine learning techniques and packages in Python.

Preliminary Analysis

I see 2 interlinked projects in this:

1. How do you make more users rate the app? (binary target) that is target those users before others and in certain conditions (I imagine just after a "like" or a message or a match).
2. How do you make users rate the app higher? (numerical target) that is when and under what conditions can I be sure to get 5 stars.

I am leaving out the wider project of making the users more satisfied with the app.

Available data

I have event level data available, some examples that could be tracked:

- clicks on profiles (numerical)
- time lags (numerical)
- first-time user (binary)
- demographics (categorical)
- rating (pseudo numerical)
- acceptance to rate (binary)

Data preparation

a. Data extraction
For unstructured data (profile texts for example)I will extract words used or using regular expression I will try to identify other indicators.

b. Missing data:
Depending on the variable's nature and the distribution of this variable I will either leave empty or null or create a "general" class if categorical or replace by the mean if numerical.

c. Excluding non relevant data :
I will review all variable and exclude any data that shouldn't not have an impact or that will not be known in advance rendering the model not useable.

d. Checks and modifications
I will review the distributions of all variables.

I would transform 2 linked variable into ratios or differences to extract the inner information: for example 2 dates become a lag in number of days.

e. Categorical data treatment
Some variables have category labels as string and I need to convert all of them to integer labels for the Scikit RandomTree

f. Data sets for training, cross validation and testing
I would shuffle the dataset and depending on the size of it sample some of it.
Then I would divided the sample set in 3 parts: the training set (to fix the model), the validation set (to tweak the parameters of the model) and the test set (to do a final review of its capacity at predicting the pledged amount for new data.

Machine Learning

In terms of machine learning techniques that we could envisage, most supervised techniques are possible but I would start with Random Tree Classification (1.) /Regression (2.)
As it is a robust, low computation solution. It can also deal with numerical and categorical data altogether.
Also Random Trees offers a clear path of action to obtain the rating and high grade.
Scikit (sclearn) has sklearn.ensemble.RandomForestRegressor and RandomForestClassifier

The input data is very likely to be the same of very similar for the Regressor and Classification. Hence a gain in terms of space and time.

I would evaluation the quality of my model using R Square as my main score for the Regression on rating. The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by the model. For the classication a confusion matrix will provide a score. In both cases we score the training ,validation and test sets being careful to avoid data mining.

Further Developments
The second model to consider would be Neural Networks as there could be some conditional and logical mechanisms. Neural Networks could possibly model better the relationship between data and rating but it is also more difficult to infer solution to our improvement project.