PRAEKELT
.ORG

BRIEF

Data Scientist Challenge

AUTHOR.............................. Colette Mathiesen
DATE.................................. November 2022

INTRODUCTIONS

Hello Potential Data Scientist!

We have so enjoyed speaking to you so far but for this stage of the interview process we would like to ask you to complete a short challenge that will not only give us an idea of how you think and tackle challenges but also give you an idea of some of the work that would come across your desk in this role.

THE CHALLENGE

QUESTION 1:

Developers created a Whatsapp chatbot service that sends important health alerts about infectious diseases. These alerts contain additional options that users can select to get more information.

The data is stored in two data tables:

1. messages.csv
2. chats.csv.

The data dictionary for these tables is as follows:

Chats

id	Identifier for a chat. Can be used to join messages and chats tables.
inserted_at	The timestamp when the user engaged with the service for the first time (i.e. time of first message)

Messages

message_id	Identifier for a message
chat_id	Identifier for a chat. Can be used to join messages and chats tables.
direction	Whether the message was sent (outbound) or received (inbound)
inserted_at	The timestamp that the message was sent or received

You have been asked to create a dashboard showing the following metrics:

1. Total number of unique users
2. Total number of inbound messages
3. Number of inbound messages over time compared to number of outbound messages over time.
4. Challenge: Number of returning users over time, compared with total number of users
5. Optional: Any other metrics that would be interesting to show

You can decide on the graph types, layout, and colours.

You can use any dashboard platform you like. If you want to use Tableau, you can create a free Tableau Public profile and then submit the .twbx file or the link to your viz.If you choose to use an alternative platform, please send through a link if possible, otherwise please send a pdf export or screenshots of your dashboard.

QUESTION 2 (TO BE COMPLETED ON THE TEST DOME TOOL):

You will receive an email from Test Dome with the second part of this challenge. The expected solving time is 40 min with a maximum solving time of 1 hour for this question.

In summary the questions are as follows:

2.1 Given the following data definition:

```
TABLE roads
  name VARCHAR(20) NOT NULL PRIMARY KEY
  length INTEGER NOT NULL
```

Create a view, named longRoads that selects the road's name and length, which have a greater than or equal to average length.

2.2. An educational magazine publishes rankings of students and their colleges in a competition in building unmanned aerial vehicles (drones). Students who participated in the competition in different years, their ranking in the competition, and their colleges are contained in the following tables:

```
TABLE colleges
  id INTEGER PRIMARY KEY
  name VARCHAR(50) NOT NULL

TABLE students
  id INTEGER PRIMARY KEY
  name VARCHAR(50) NOT NULL
  collegeId INTEGER
  FOREIGN KEY (collegeId) REFERENCES colleges(id)

TABLE rankings
  studentId INTEGER
  ranking INTEGER NOT NULL
  year INTEGER NOT NULL
  FOREIGN KEY (studentId) REFERENCES students(id)
```

Write a query that lists all colleges that have at least one student with a ranking between 1 and 3 (both inclusive), for the year 2015. The query should return:

- The college name.
- The rank of their best ranking student for 2015.
- The number of students who had rankings between 1 and 3 (both inclusive) for the year 2015.

Rank 1 is the best rank, rank 2 the second-best, and so on. More than one student can tie for a rank in a year.

2.3 Implement the function class_grades, which accepts all students in a school and returns all classes in the school, together with their median grades.
Students are stored in a two-dimensional list where each element of the list is another list with the following elements: Student name, class name, student grade.

The function should return a two-dimensional list where each element is a list with the following elements: Class name, median grade for students in the class. The order of returned classes in not important.

TIMELINES AND SUBMISSION DETAILS

You will have until **Monday, 28 November 2022 at 12pm** to complete your challenge.

RESOURCES

In order to complete this challenge you have been provided with the following:

Links to the follow data sets:

- Messages https://drive.google.com/drive/folders/180HlxxXafP6IeIOJh2VjnB3I42e7cSIJ?usp=sharing
- Chats https://drive.google.com/drive/folders/180HlxxXafP6IeIOJh2VjnB3I42e7cSIJ?usp=sharing

Please create copies of each of the documents in order to create your own documents for submission.

QUESTIONS

Please note that we do understand that there may be information that you would like to have available to complete this that you do not have. It's completely understandable that you may feel a little lost, but this is not a test of exactly what you know, but rather how you think and use the skills at your disposal to solve a problem. Please feel free to make assumptions where you don't have all the information you need but just document this for us. However, if you have any burning questions that you would like answered please feel free to mail marcelle@praekelt.org with your questions and we will get back to you.

CONFIDENTIALITY

In completing this challenge we would like to ask you to keep the information and templates that we provide you confidential.