## Technical

1. Given a database table with the following fields: raw\_credit\_card\_transaction, transaction\_date, transaction\_amount, zipcode, merchant.

Question:

1. Write a query that returns transaction revenue and transactions counts by day, by merchant.

**select** transaction\_date**,**

merchant**,**

**sum(**transaction\_amount**)** daily\_transaction\_revenue**,**

**count(**raw\_credit\_card\_transaction**)** transactions\_cnt

**from** table

**group** **by** transaction\_date**,** merchant

**order** **by** 1, 2**;**

1. Write a function (language of choice) that identifies anomalies in daily transaction revenue by merchant.

**def** detect\_anomalies**(**df**,** target **=** 'transaction\_amount'**,** group\_var **=** 'transaction\_date'**,** group\_var2 **=** 'merchant'**):**

# Group the daily total spending of each merchant

new\_df **=** df**.**groupby**([**group\_var**,** group\_var2**])[**target**].**agg**([**'sum'**]).**reset\_index**()**

/\* Use z-score to determine outliers by identifying the daily spending that are 3 standard deviations away from the average daily spending for each merchant\*/

stds **=** 3.0

z **=** new\_df**.**groupby**(**group\_var2**).**transform**(**

**lambda** x**:** **(**x **-** x**.**mean**()).**div**(**x**.**std**()))**

outliers **=** z**.**abs**()** **>** stds

**return(**new\_df**[**outliers**.**any**(**axis **=** 1**)])**

2. Given a dataset/dataframe with the following schema:

Date, Country, App\_name, User\_id

Data dictionary:

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Format |
| Date | The date on which a user was active on an application (phone app). Date ranges from 2014 – present. | YYYY-mm-dd |
| Country | The country from which a user was active on an application. | 2-digit ISO country code. |
| App\_name | the application that the user was active on. | The data only tracks 10 apps with the following names: a1, a2, a3, a4, … a10. |
| User\_id | The unique ID assigned to a user to track history | U1, U2… etc. |

A few sample rows of data:

2014-07-11, US, a10, u100

2016-01-22, FR, a8, u46

.

.

.

2014-09-24, IT, a3, u56

Provide a solution (query or code) to find the top 3 other apps used by users of app a08 in 2017.

--extract the top 3 app, excluding app 'a08', of each user of app 'a08' in 2017

**select** rk,

year\_**,**

user\_id**,**

app\_name**,**

app\_cnt

**from**

**(**

--extract and rank the usage counts of the all apps, excluding 'a08', of each user of 'a08' in 2017

**select** **dense\_rank()**over**(partition** **by** user\_id **order** **by** **count**(app\_name)**desc)** rk**,**

**to\_char(**date\_**,** 'YYYY'**)** year\_**,**

user\_id**,**

app\_name**,**

**count(**app\_name**)** app\_cnt

**from** apptable

**where** user\_id **IN** **(**

--extract the users who used the app 'a08' in 2017

**select** **distinct** user\_id

**from** apptable

**where** app\_name **=** 'a08'

**and** **to\_char(**date\_**,** 'YYYY'**)** **=** '2017'

**)**

**and** app\_name **!=** 'a08'

**and** **to\_char(**date\_**,** 'YYYY'**)** **=** '2017'

**group** **by** **to\_char(**date\_**,** 'YYYY'**),** user\_id**,** app\_name

**)** x

**where** rk **<=** 3**;**

Bonus question - if we only focused on active users of app a08 (those who use the app frequently), how would your answer change?

Hint: How would you define an active user?

Let’s define an active user as a user who uses an app at least twice a week in a year.

Therefore, to be considered as an active user of app ‘a08’, the user must have used the app at least twice a week in a year.

We first need to find the users of app ‘a08’ who used the app at least twice in any given week of 2017.

--extract the top 3 app usage, excluding app 'a08', of each **active** user of app 'a08' in 2017

**select** rk,

year\_**,**

user\_id**,**

app\_name**,**

app\_cnt

**from**

**(**

--extract the usage counts of all the apps, excluding 'a08', of each **active** user of 'a08' in 2017

**select dense\_rank()**over**(partition** **by** user\_id **order** **by** **count**(app\_name)**desc)** rk**,**

**to\_char(**date\_**,** 'YYYY'**)** year\_**,**

user\_id**,**

app\_name**,**

**count(**app\_name**)** app\_cnt

**from** apptable

**where** user\_id **IN** **(**

--extract the users who used the app 'a08' at least twice in every week of 2017

**select** user\_id

**from**

**(**

--for each user, extract the weeks of 2017 in which users used the app 'a08' at least twice

**select** **to\_char(**date\_**,** 'WW'**)** week**,**

user\_id**,**

**count(**app\_name**)** app\_cnt

**from** apptable

**where** app\_name **=** 'a08'

**and** **to\_char(**date\_**,** 'YYYY'**)** **=** '2017'

**group** **by** **to\_char(**date\_**,** 'WW'**),** user\_id

**having** **count(**app\_name**)** **>=** 2

**)**

**group** **by** user\_id

**having** **count(**week**)** >**=** 52

**)**

**and** app\_name **!=** 'a08'

**and** **to\_char(**date\_**,** 'YYYY'**)** **=** '2017'

**group** **by** **to\_char(**date\_**,** 'YYYY'**),** user\_id**,** app\_name

**)** x

**where** rk **<=** 3**;**

For app a07 in 2016 what are the 5 most active countries (by user count) adjusting for relative user representation by country in the data?

Hint: A large country would (probably) have more users but that does not mean the app is frequently used there

Let’s consider an app as an active app for a country if the users of the app in that country use the app at least twice every week in a year

We need to find out the active users of app ‘a07’ who used the app at least twice every week in 2016.

Then we find the countries with the most active users of app ‘a07’.

--extract the top 5 most active countries based on their active user counts for app 'a07' in 2016

**select** rk**,**

year\_**,**

country**,**

app\_name**,**

user\_cnt

**from**

**(**

--rank the countries based on the number of the active users of the app 'a07' in 2016

**select** **dense\_rank()**over**(order** **by** **count(distinct** user\_id**)desc)** rk**,**

**to\_char(**date\_**,** 'YYYY'**)** year\_**,**

country**,**

app\_name**,**

**count(distinct** user\_id**)** user\_cnt

**from** apptable

**where** app\_name **=** 'a07'

**and** **to\_char(**date\_**,** 'YYYY'**)** **=** '2016'

**and** user\_id **IN** **(**

--extract the users who used the app 'a07' at least twice in every week of 2016

**select** user\_id

**from**

**(**

--for each user, extract the weeks of 2016 in which users used the app 'a07' at least twice

**select** **to\_char(**date\_**,** 'WW'**)** week**,**

user\_id**,**

**count(**app\_name**)** app\_cnt

**from** apptable

**where** app\_name **=** 'a07'

**and** **to\_char(**date\_**,** 'YYYY'**)** **=** '2016'

**group** **by** **to\_char(**date\_**,** 'WW'**),** user\_id

**having** **count(**app\_name**)** **>=** 2

**)**

**group** **by** country

**having** **count(**week**)** **>=** 52

**)**

**group** **by** **to\_char(**date\_**,** 'YYYY'**),** country**,** app\_name

**)**

**where** rk **<=** 5**;**