Tables:

**APP\_USAGE**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Data Type** | **Description** | **Examples** |
| device\_id | integer | Identifier for a specific mobile phone | 1, 2, 3 |
| mobile\_app\_id | integer | Identifier for a specific mobile app | 1, 2, 3 |
| day | timestamp | Day of usage | 2020-08-15 |
| seconds\_of\_usage | float | Amount of time device spent using the app on the day. It will always be > 0 | 147.5 |

**DEVICES**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Data Type** | **Description** | **Examples** |
| device\_id | integer | Identifier for a specific mobile phone | 1, 2, 3 |
| manufacturer | text | The manufacturer of the phone | ‘Apple’, ‘Samsung’ |
| model | text | Model of the phone | ‘Iphone X’, ‘Galaxy S10’ |
| platform | text | IOS or Android | `iOS`, `Android` |

**MOBILE\_APPS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Data Type** | **Description** | **Examples** |
| mobile\_app\_id | integer | Identifier for a specific mobile app | 1, 2, 3 |
| title | text | Title of the app | `Instagram`, `Google Maps` |
| platform | text | IOS or Android | `iOS`, `Android` |

**Write Queries to pull the following data:**

* Get all of the app usage for device\_id = 3

 Select sum(seconds\_of\_usage)

From app\_usage

Where device\_id = 3

Select \*

From app\_usage

Where device\_id = 3

* How many devices used the app with the title ‘Instagram’ on '2020-08-15'?

Select  count(device\_id)

From app\_usage u

Join mobile\_apps a

On u.mobile\_app\_id = a.mobile\_app\_id

 Where a.title = ‘‘Instagram’ and u.day = ‘'2020-08-15'

* For how many seconds did device\_id = 2 use apps for on ‘2020-08-15'?

 Select sum(seconds\_of\_usage)

From app\_usage

Where device\_id = 2 and day = ‘‘2020-08-15'

* What is the average seconds of usage for mobile\_app\_id = 4 on ‘2020-08-15'?

  Select mean(seconds\_of\_usage)

From app\_usage

Where mobile\_app\_id = 4 and day = ‘‘2020-08-15'

* Which day had the greatest number of devices use the app 'SnapChat’?

Select  day

From app\_usage u

Join mobile\_apps a

On u.mobile\_app\_id = a.mobile\_app\_id

 Where a.title = 'SnapChat’

group by day

Order by count(device\_id)

Limit 1

* Get all of the app usage on ‘2020-08-15' for a random sample of 10,000 devices.

With devices (

Select distinct device\_id

From app\_uage

Where day = ‘2020-08-15'

Order by random()

Limit 10000

)

Select \*

From app\_usage

Join device

On app\_usage.device\_id = device.device\_id

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistics Interview questions

Consider the following scenario:

-There is large population of devices in the US (~300 million)

-Every device belongs to someone who is “conservative leaning”, or “liberal learning”

-We can obtain the political affiliation for a random sample of devices, n, where n is much less than 300 million.

-In the sample, there are k conservative leaning devices, and (n - k) liberal leaning devices.

-We want to know the proportion of the population of devices that is conservative leaning vs liberal leaning, given by the variable p.

Questions:

-What is the best estimate of the population proportion of conservative leaning devices, p?

P = k/n

-What distribution does the variable k follow?

binomial

-The “true” proportion of conservative leaning devices is given by p. What is an estimate of the 95% confidence interval for p?

p = k/n

std = sqrt( p(1-p)/n)

[k/n - 2\*std, k/n + 2\*std]

-How large does n need to be in order for pto be within 1% of p at the 95% confidence level?

//k/n - 2\*std > k/n\*0.99

2 \* std <= 0.01

2\*sqrt(k/n(1-p)/) <= 0.01

sqrt(k(n-k)/n^3) <= 0.005

-In what situation will the method for estimating the confidence interval for p work, or not work?

When n is small, or close to 0/1, which make the distribution not normal