Tonal

**Technical Take-home**

# Applicant: [Edima Udo]

## Goals

The goal of this technical take-home includes:

1. use of a web-based business intelligence tool
2. investigating a new dataset with SQL
3. communication of your results via written and visual replies

## Context

As part of our work as analysts, we're often required to pick up new tools and generate insights quickly. As part of our interview process, we'd like you to try (what may be) a new tool to answer some business questions. Mode Analytics is a business intelligence and visualization to be used in this take-home. ***Instructions to create a new Mode Analytics account can be found at the end of this document.*** Additionally, the schema containing the relevant tables are included as part of Mode's Public warehouse.

## Instructions

**Please work in and edit this document directly to answer all questions and provide access to any linked documentation.**

We anticipate this exercise to take approx 2-3 hours. Email any questions and your final version to [michelle@tonal.com](mailto:michelle@tonal.com).

To answer each question, we ask that you include two things:

* a brief answer to the questions; please clearly state your assumptions and the logic of your solution
* your SQL query/code; you may include an external link (Google Doc or Mode Report) to reference your source code
* You may also create a notebook in Mode Analytics to complete this exercise if you would like to, but please include SQL code for every answer
* Keep in mind code format, comments etc. – make the code as legible as possible

Yammer is a social network for communicating with coworkers. Individuals share documents, updates, and ideas by posting them in groups. Please use the Yammer data to answer the following questions. The four tables available in mode are:

tutorial.yammer\_emails

tutorial.yammer\_events

tutorial.yammer\_experiments

tutorial.yammer\_users

## Part 1

Please complete the following questions based upon the data referenced above. This section focuses on your coding skills.

Before you begin, draw out a simple database schema diagram on how you think these datasets are linked together and share this in your answer key.

1. How many users are in the yammer dataset? What percentage of users are active?
2. Which acquisition month drove the highest # of active users? And how many users in that month?
3. What percentage of active users have opened emails (any emails) from Yammer?
4. Provide the 5 most active users logging in to Yammer from the United States for company\_id 1, 2 & 4 (total of 15 user\_ids).
5. For the Weekly Digest emails, which weeks had the highest email open rate and click through rate and what were the rates during those weeks?
6. For the companies which have more than 100 English users, what is the percentage of users who have both ‘liked’ a message AND done a ‘search autocomplete’ event for each company?

## Part 2

This section focuses on your ability to do an analysis.

1. Looking at all of the yammer datasets, what conclusions can you draw about the types of users who have the highest engagements with Yammer?
2. Looking at the yammer ‘publisher update’ experiment, would you say that the experiment was successful? Did it work better in some locations vs. others? Would you recommend rolling it out?

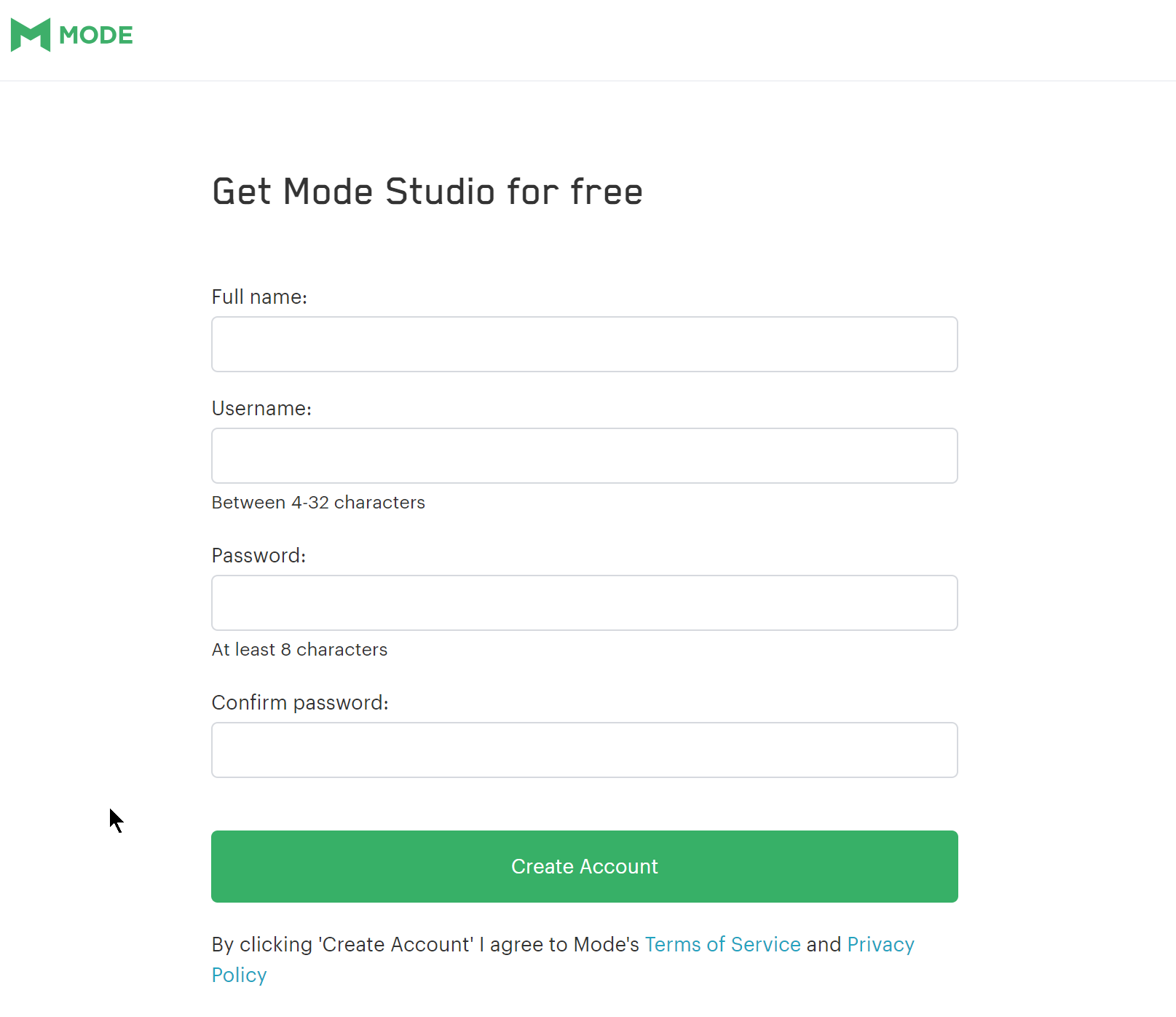
Support your answer with data and/or visualizations.

*While you may use Excel or non-SQL tools to answer this question, we prefer you use Mode Dashboards/Reports and be prepared to present them to us.*

**PLEASE NOTE:** Some free Mode accounts no longer support “public” views or they have made it harder to share than in the past. As such, if you can’t share your dashboard then please export your dashboard to PDF and email to us or attach a link to it below.

## Appendix: How to set up a free Mode Account

1. Visit <https://modeanalytics.com/>
2. Enter your email to sign up for a free account for Mode Studio, and enter your details on the following page.



1. Click "Start using Mode with Sample data" to open the SQL editor.

Graphical user interface, text, application

Description automatically generated

4. Enter the following code to confirm you have access to the data:

select \* from tutorial.yammer\_emails

A screenshot of a computer

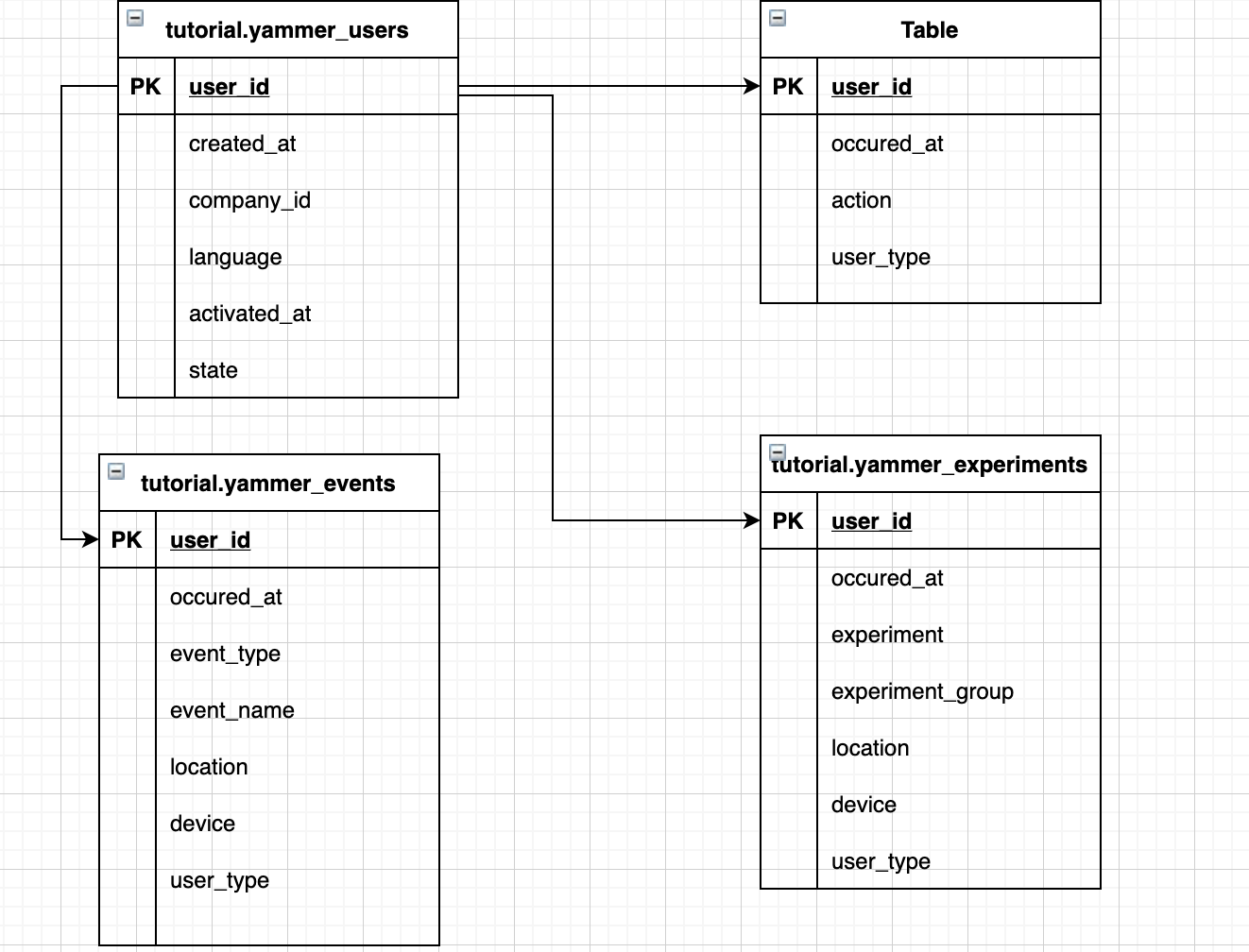
Description automatically generated

## Your Answers

Use the space below - **formatted as you see fit** - to answer the test questions.

**Part 1**

1. **Design a simple database schema diagram on how you think these datasets are linked together**



1. **How many users are in the yammer dataset? What percentage of users are active?**

The number of users in the dataset is 19066. The logic is to use a count of user id.

SQL Query

select count(distinct user\_id) from tutorial.yammer\_users

The percentage of active users are: 49.2%. The logic is to use a case statement and average.

SQL Query

Select AVG(CASE WHEN state = 'active' THEN 1.0 ELSE 0 END)

From tutorial.yammer\_users

1. **Which acquisition month drove the highest # of active users? And how many users in that month?**

Looking at Year and month

The acquisition month that drove the highest # of active user was: August

And the number of users are: 1990. The logic behind this is to get date level information using built in functions and a count using case statement. The same applies when looking at month only.

SQL Query

Select DATE\_PART('year',created\_at) as YearInfo , DATE\_PART('month',created\_at) as MonthInfo ,COUNT(CASE WHEN state = 'active' THEN 1.0 ELSE 0 END) AS activeUsers

from tutorial.yammer\_users

group by DATE\_PART('year',created\_at), DATE\_PART('month',created\_at)

order by 3 DESC

Looking at month only

The acquisition month that drove the highest # of active user was: August

And the number of users are: 2626

SQL Query

Select DATE\_PART('month',created\_at) as MonthInfo ,COUNT(CASE WHEN state = 'active' THEN 1.0 ELSE 0 END) AS activeUsers

from tutorial.yammer\_users

group by DATE\_PART('month',created\_at)

order by 2 DESC

1. **What percentage of active users have opened emails (any emails) from Yammer?**

Assumption: opened email means open email and email click through

The percentage of active users that have open emails from yammer are: 32.6%. The logic behind this is to use case and average to get the percentage.

SQL Query

Select AVG(CASE WHEN tu.state = 'active' and te.action in ('email\_open', 'email\_clickthrough') THEN 1.0 ELSE 0 END) as ActiveEmailCount

From tutorial.yammer\_users tu inner join tutorial.yammer\_emails as te on tu.user\_id = te.user\_id

1. **Provide the 5 most active users logging in to Yammer from the United States for company\_id 1, 2 & 4 (total of 15 user\_ids).**

Assumption logged in only factors in the event name login.

Output

|  |  |  |
| --- | --- | --- |
| user\_id | company\_id | activecount |
| 2940 | 1 | 39 |
| 2990 | 1 | 27 |
| 9479 | 1 | 24 |
| 5134 | 1 | 22 |
| 14018 | 1 | 19 |
| 10067 | 2 | 35 |
| 4442 | 2 | 24 |
| 4248 | 2 | 18 |
| 5168 | 2 | 18 |
| 1650 | 2 | 16 |
| 2474 | 4 | 22 |
| 6372 | 4 | 20 |
| 12937 | 4 | 17 |
| 12178 | 4 | 9 |
| 13671 | 4 | 5 |

SQL Query

(select tu.user\_id, company\_id, count(\*) as activeCount

from tutorial.yammer\_users tu inner join tutorial.yammer\_events te on tu.user\_id = te.user\_id

where tu.company\_id = 1

and te.location = 'United States'

and tu.state = 'active'

and te.event\_name = 'login'

group by tu.user\_id, company\_id

order by 3 DESC

LIMIT 5)

union

(select tu.user\_id, company\_id, count(\*) as activeCount

from tutorial.yammer\_users tu inner join tutorial.yammer\_events te on tu.user\_id = te.user\_id

where tu.company\_id = 2

and te.location = 'United States'

and tu.state = 'active'

and te.event\_name = 'login'

group by tu.user\_id, company\_id

order by 3 DESC

limit 5)

UNION

(select tu.user\_id, company\_id, count(\*) as activeCount

from tutorial.yammer\_users tu inner join tutorial.yammer\_events te on tu.user\_id = te.user\_id

where tu.company\_id = 4

and te.location = 'United States'

and tu.state = 'active'

and te.event\_name = 'login'

group by tu.user\_id, company\_id

order by 3 DESC

limit 5)

ORDER BY company\_id ASC, activeCount DESC

1. **For the Weekly Digest emails, which weeks had the highest email open rate and click through rate and what were the rates during those weeks?**

Week with the highest click through rate is week 18 with 12.2%. The logic behind this is to arrange the information by week and get the weekly rate using case and average plus built in date functionality.

SQL Query

select DATE\_PART('week',te.occurred\_at) as WeekInfo,

AVG(CASE WHEN te.action = 'email\_clickthrough' THEN 1.0 ELSE 0 END) as emailClickThrough

From tutorial.yammer\_emails as te

group by 1

order by 2 DESC

limit 5

Week with the highest email open rate is week 35 with 23.95%

select DATE\_PART('week',te.occurred\_at) as WeekInfo,

AVG(CASE WHEN te.action = 'email\_open' THEN 1.0 ELSE 0 END) as emailOpenRate

From tutorial.yammer\_emails as te

group by 1

order by 2 DESC

limit 5

1. **For the companies which have more than 100 English users, what is the percentage of users who have both ‘liked’ a message AND done a ‘search autocomplete’ event for each company?**

The percentage is 22.85%. The logic behind this is to get the percentage of user. It involves using a sub query and case statement.

SQL Query

select AVG(CASE WHEN te.event\_name in ('like\_message','search\_autocomplete') THEN 1.0 ELSE 0 END) as englishSearch

From tutorial.yammer\_users as tu inner join tutorial.yammer\_events as te on tu.user\_id = te.user\_id

where tu.company\_id in (select tu.company\_id

From tutorial.yammer\_users as tu

where tu.language = 'english'

group by 1

having count(\*) > 100)

Support your answer with data and/or visualizations.

## Part 2

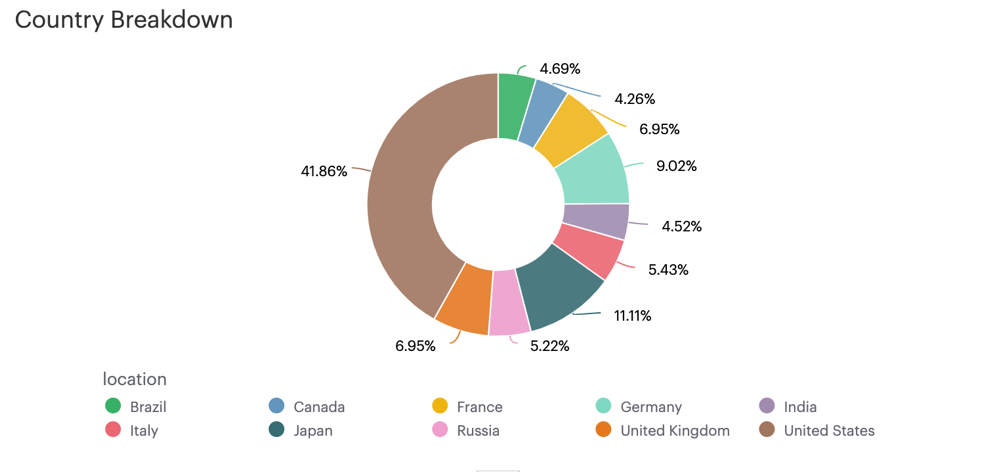
1. **Looking at all of the yammer datasets, what conclusions can you draw about the types of users who have the highest engagements with Yammer?**

I’ll look at the top 10 by location, device and language

1. The top 10 countries that had the most engaged users were

|  |  |
| --- | --- |
| **Count** | **Country** |
| 1741 | United States |
| 462 | Japan |
| 375 | Germany |
| 289 | United Kingdom |
| 289 | France |
| 226 | Italy |
| 217 | Russia |
| 195 | Brazil |
| 188 | India |

Using a donut chart it looks like below



SQL Queries for country

select count(\*) as userCount, te.location

From tutorial.yammer\_users as tu inner join tutorial.yammer\_events as te on tu.user\_id = te.user\_id

where te.event\_type = 'engagement' and tu.state = 'active'

group by te.location

order by 1 DESC

limit 10

1. The top 10 devices were

|  |  |
| --- | --- |
| **Count** | **Devices** |
| 1952 | macbook pro |
| 1309 | lenovo thinkpad |
| 1025 | iphone 5 |
| 950 | macbook air |
| 803 | samsung galaxy s4 |
| 677 | dell inspiron notebook |
| 626 | iphone 5s |
| 621 | nexus 5 |
| 478 | ipad air |
| 409 | iphone 4s |

Apple devices took about 61% of the top 10 values.

SQL Query for device

select count( distinct tu.user\_id) as userCount, te.device

From tutorial.yammer\_users as tu inner join tutorial.yammer\_events as te on tu.user\_id = te.user\_id

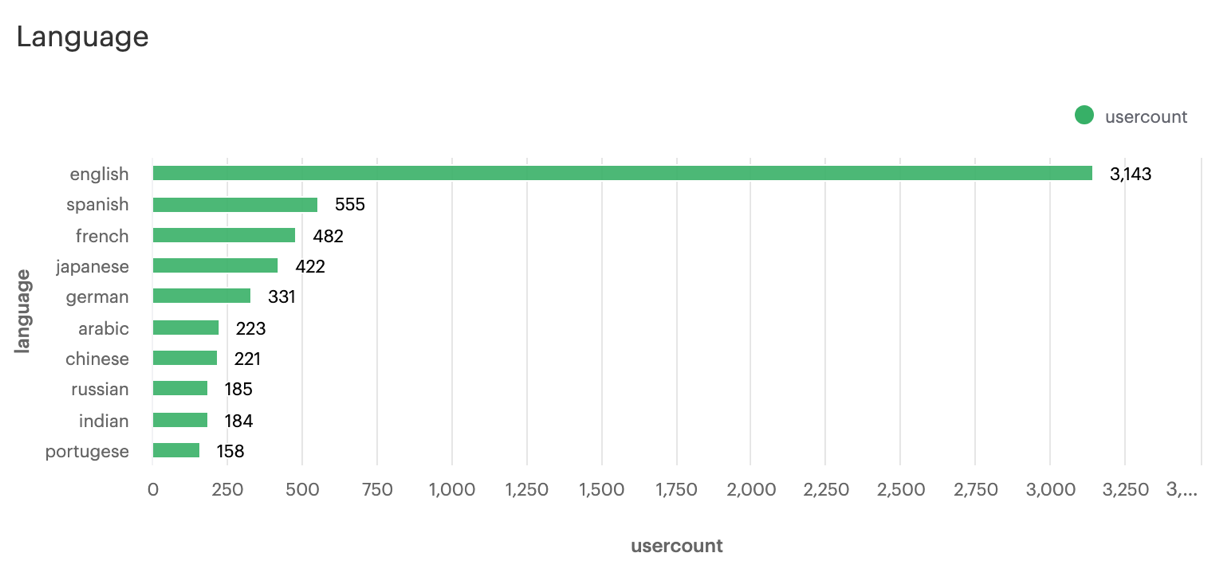
where te.event\_type = 'engagement' and tu.state = 'active'

group by te.device

order by 1 DESC

limit 10

1. The top 10 language are shown below. English takes up about 50% of the top 10.



SQL Query for language

select count( distinct tu.user\_id) as userCount, tu.language

From tutorial.yammer\_users as tu inner join tutorial.yammer\_events as te on tu.user\_id = te.user\_id

where te.event\_type = 'engagement' and tu.state = 'active'

group by tu.language

order by 1 DESC

limit 10

1. **Looking at the yammer ‘publisher update’ experiment, would you say that the experiment was successful? Did it work better in some locations vs. others? Would you recommend rolling it out?**

Need more context on what the experiment is. I will design a query that would showcase the statistics of the experiment.

1. From the simple hypothesis analysis it seems the analysis looks sound but a deep dive is needed.



SQL Query

SELECT c.experiment,

c.experiment\_group,

c.users,

c.total\_treated\_users,

ROUND(c.users/c.total\_treated\_users,4) AS treatment\_percent,

c.total,

ROUND(c.average,4)::FLOAT AS average,

ROUND(c.average - c.control\_average,4) AS rate\_difference,

ROUND((c.average - c.control\_average)/c.control\_average,4) AS rate\_lift,

ROUND(c.stdev,4) AS stdev,

ROUND((c.average - c.control\_average) /

SQRT((c.variance/c.users) + (c.control\_variance/c.control\_users))

,4) AS t\_stat,

(1 - COALESCE(nd.value,1))\*2 AS p\_value

FROM (

SELECT \*,

MAX(CASE WHEN b.experiment\_group = 'control\_group' THEN b.users ELSE NULL END) OVER () AS control\_users,

MAX(CASE WHEN b.experiment\_group = 'control\_group' THEN b.average ELSE NULL END) OVER () AS control\_average,

MAX(CASE WHEN b.experiment\_group = 'control\_group' THEN b.total ELSE NULL END) OVER () AS control\_total,

MAX(CASE WHEN b.experiment\_group = 'control\_group' THEN b.variance ELSE NULL END) OVER () AS control\_variance,

MAX(CASE WHEN b.experiment\_group = 'control\_group' THEN b.stdev ELSE NULL END) OVER () AS control\_stdev,

SUM(b.users) OVER () AS total\_treated\_users

FROM (

SELECT a.experiment,

a.experiment\_group,

COUNT(a.user\_id) AS users,

AVG(a.metric) AS average,

SUM(a.metric) AS total,

STDDEV(a.metric) AS stdev,

VARIANCE(a.metric) AS variance

FROM (

SELECT ex.experiment,

ex.experiment\_group,

ex.occurred\_at AS treatment\_start,

u.user\_id,

u.activated\_at,

COUNT(CASE WHEN e.event\_name = 'send\_message' THEN e.user\_id ELSE NULL END) AS metric

FROM (SELECT user\_id,

experiment,

experiment\_group,

occurred\_at

FROM tutorial.yammer\_experiments

WHERE experiment = 'publisher\_update'

) ex

JOIN tutorial.yammer\_users u

ON u.user\_id = ex.user\_id

JOIN tutorial.yammer\_events e

ON e.user\_id = ex.user\_id

AND e.occurred\_at >= ex.occurred\_at

AND e.occurred\_at < '2014-07-01'

AND e.event\_type = 'engagement'

GROUP BY 1,2,3,4,5

) a

GROUP BY 1,2

) b

) c

LEFT JOIN benn.normal\_distribution nd

ON nd.score = ABS(ROUND((c.average - c.control\_average)/SQRT((c.variance/c.users) + (c.control\_variance/c.control\_users)),3))

b) It looks like it did but it still needs more analysis

|  |  |  |  |
| --- | --- | --- | --- |
|  | experiment\_group | |  |
| location | control\_group | test\_group | Totals |
| Argentina | 15 | 4 | **19** |
| Australia | 30 | 22 | **52** |
| Austria | 11 | 6 | **17** |
| Belgium | 19 | 4 | **23** |
| Brazil | 54 | 29 | **83** |
| Canada | 65 | 15 | **80** |
| Chile | 4 | 3 | **7** |
| Colombia | 8 | 6 | **14** |
| Denmark | 14 | 3 | **17** |
| Egypt | 10 | 4 | **14** |
| Finland | 11 | 3 | **14** |
| France | 89 | 41 | **130** |
| Germany | 108 | 56 | **164** |
| Greece | 6 | 1 | **7** |
| Hong Kong | 7 | 8 | **15** |
| India | 56 | 21 | **77** |
| Indonesia | 23 | 16 | **39** |
| Iran | 13 | 5 | **18** |
| Iraq | 6 | 3 | **9** |
| Ireland | 3 | 4 | **7** |
| Israel | 13 | 5 | **18** |
| Italy | 69 | 23 | **92** |
| Japan | 122 | 65 | **187** |
| Korea | 41 | 25 | **66** |
| Malaysia | 10 | 5 | **15** |
| Mexico | 38 | 29 | **67** |
| Netherlands | 26 | 9 | **35** |
| Nigeria | 7 | 5 | **12** |
| Norway | 11 | 5 | **16** |
| Pakistan | 4 | 1 | **5** |
| Philippines | 13 | 4 | **17** |
| Poland | 18 | 6 | **24** |
| Portugal | 6 | 3 | **9** |
| Russia | 53 | 30 | **83** |
| Saudi Arabia | 21 | 14 | **35** |
| Singapore | 9 | 2 | **11** |
| South Africa | 11 | 3 | **14** |
| Spain | 32 | 17 | **49** |
| Sweden | 25 | 9 | **34** |
| Switzerland | 22 | 5 | **27** |
| Taiwan | 12 | 12 | **24** |
| Thailand | 10 | 6 | **16** |
| Turkey | 17 | 6 | **23** |
| United Arab Emirates | 14 | 10 | **24** |
| United Kingdom | 80 | 43 | **123** |
| United States | 502 | 247 | **749** |
| Venezuela | 8 | 6 | **14** |
| Totals | **1746** | **849** | **2595** |

1. Based on the result I would hold off on recommendation. I would still need to dig deeper into investigate user treatments (or splitting users out into new and existing cohorts). I’ll also need to look at any other confounding or interaction effects.