**Executive Summary**

A retroactive analysis of the client's 2015 Q1 content and user log data suggests the client might optimize engagement by sending its newsletter less frequently and/or increasing the diversity of topics covered in the newsletter.

**Data & Methodology**

a) *Data and Tools Used*

The client seeks to optimize user engagement with their content. In order to suggest possible practices to increase the probability of users viewing content, our team conducted a retroactive descriptive analysis using the client's content SQLite database and user log data from the first quarter of 2015.

We used Python and IPython Notebook with the SQLite Python driver to interact with the client's database and analyze queried data. Occasionally, when performing computationally expensive operations, we accessed the SQLite database directly from the command line rather than through Python. We also made heavy use of the Python Matplotlib and Seaborn libraries to create the graphics used in this report.

b) *Calculating Click Rates*

We measure "engagement" as the rate at which users click on article links that are sent to them. Specifically, we measure click rate as:

$$Click\ Rate\ =\ \frac{Total\ Unique\ Clicks}{Total\ Number\ of\ Links\ Sent}$$

As the client's ultimate goal is to measure actual user engagement with content, we do not include clicks rendering an error (i.e. status code 400) in the numerator of our click rate. Likewise, we do not include duplicative clicks (i.e. clicks with the same user and article ID as a previous click), as the user had already interacted with that content previously.

To calculate click rate, we first read in the client's user log data into their main SQLite database using the command line code shown in Exhibit 1.

**Exhibit 1: Command Line Commands to Read Log Files into Content SQLite Database**

```
> .open data/content_digest.db
> CREATE TABLE rawclicks (string TEXT);
> .mode csv
> .import data/access.log rawclicks
```

The command line prompts described in Exhibit 1 read each click's timestamp, GET request, status code, and byte size as a single continuous string . Prior the analysis, we parsed the click timestamp, user ID, article ID, status code, and byte size into separate columns using the SQL query provided as an attachment to this report (*parse-log.sql*).

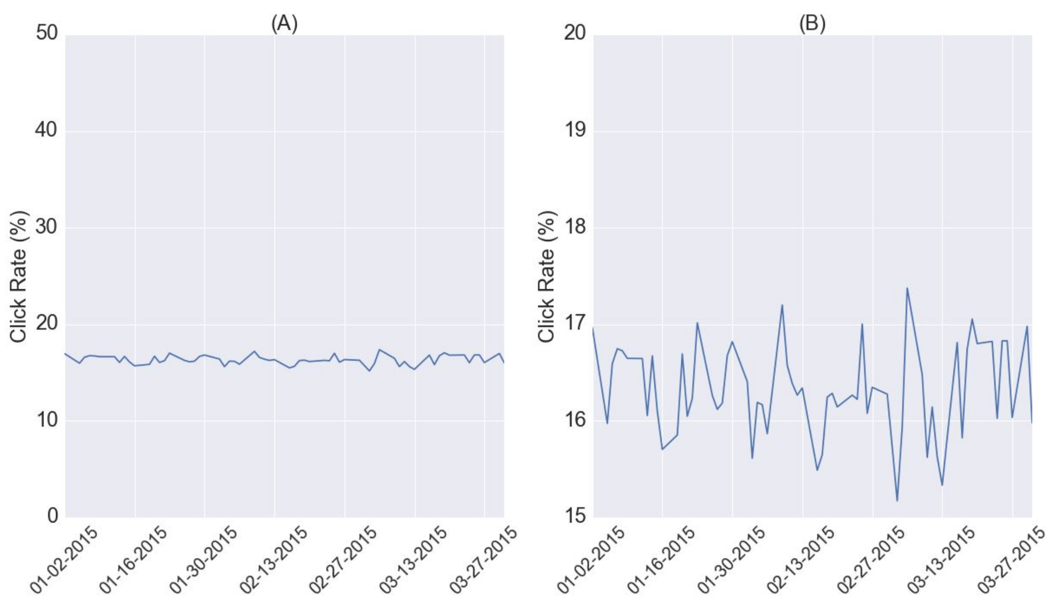
After parsing the clickstream data, we excluded clicks where status code = 400 and de-duplicated the click data . We then matched the click data to the *email_content* table of the SQLite database on user and article IDs to determine which emailed article links resulted in a successful click. Finally, we summed the total links and clicks by date and by-group of interest (such as time of day sent or article topic) and divided clicks by links to calculate click rate. An IPython Notebook (*click-rate-analysis.ipynb*) performing these analyses are provided as an attachment to this report.

**<u>Results</u>**

a) *Page View Rates Over Time*

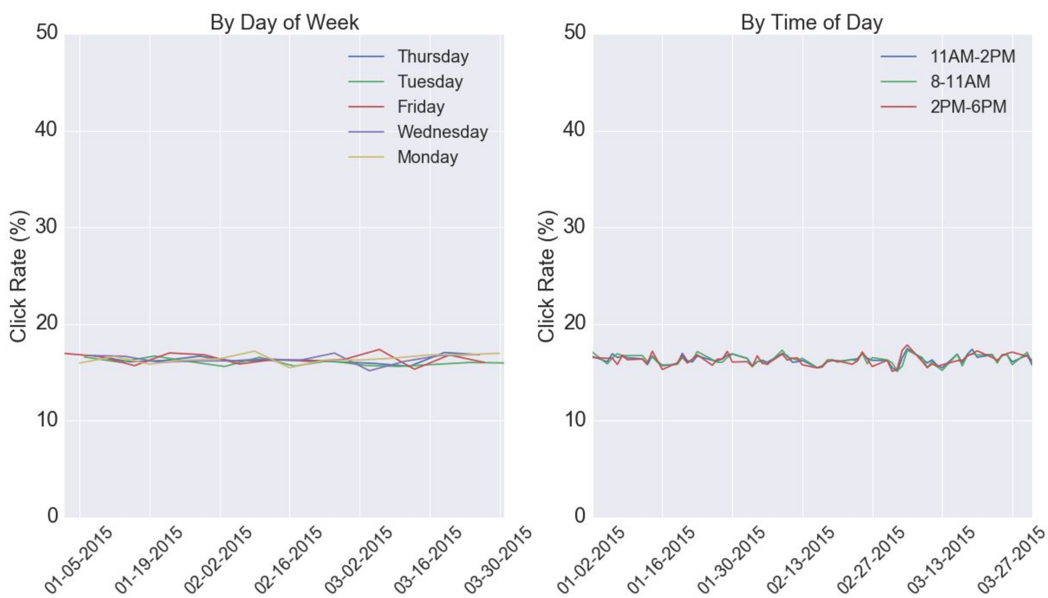
Over the three month period analyzed, 10,738,006 article links were emailed to 20,000 users. The emails resulted in 1,753,249 unique clicks, for an overall click rate of 16.3%. Click rates over time are shown in Exhibit 2. Although click rates initially appeared to remain relatively steady at around 16-17% (panel A), a finer-grained look (panel B) reveals that clicks oscillate cyclically between ~15-17%.

**Exhibit 2: Click Rates Over Time**



Interestingly, these oscillations do not appear related to the day of week or time of day links were sent to users. As shown in Exhibit 3, click rates for each day of the week and time of day remain closely clustered across the entire period; if users were more active during specific days and times, we would expect those click rates to be significantly higher across at least part of the 3-month time frame. As such, our team suspects these cycles in click rates may indicate the client is sending their newsletter too frequently, resulting in periods of relatively low user engagement following periods of higher engagement.

**Exhibit 3: Click Rates by Day of Week and Time of Day Sent**

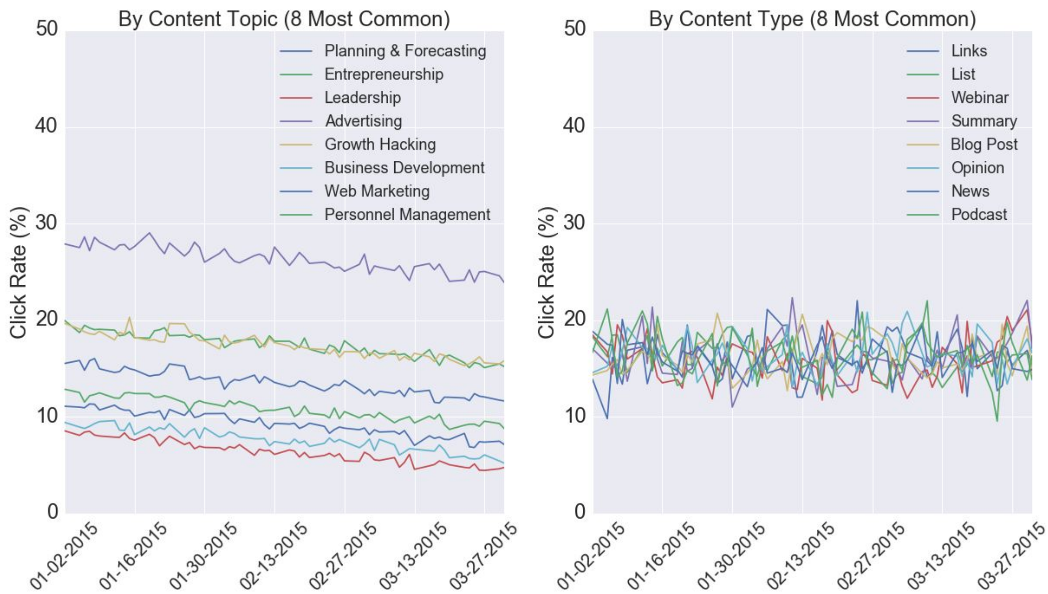



*b) Page View Rates by Topic and Type of Content*

The content the client sent to users spans 106 unique topics and 49 medium types. In order to focus our analysis, we concentrated only on the 8 article topics and types most commonly sent to users (as measured by total article links sent out). Content topic correlates strongly to user engagement, with popular topics such as advertising garnering click rates around 25-30% while the least popular topics command only 5-10% click rates. However, click rates for all 8 topics declined by ~2-5% over the three month period, which may indicate users are becoming bored with the most commonly provided content areas. See Exhibit 4 below.

In contrast to topic, medium type (e.g. webinars, blog posts, etc.) does not appear to affect user engagement. As shown in the second panel of Exhibit 4, there is no significant difference between click rates among the 8 most common types over the period analyzed.

**Exhibit 4: Click Rates by Topic and Type of Content**




## <u>Commercial Recommendations</u>

The above analysis suggests that the client's newsletter is sent too frequently, and that some content areas may be overused. However, we cannot say with certainty whether the click rate patterns observed directly resulted from these actions; there could have been other unobserved factors influencing clicks over this period. Before making any permanent changes, we would suggest running a randomized experiment to more robustly estimate the effect of altering newsletter frequency or content on user engagement.

For example, the client could randomly split it's user base into two samples of 10,000 users each, and send one group (the treatment) the newsletter less frequently and the second group (the control) the newsletter with the current standard frequency. Assuming no other changes were made to the newsletter over this period and users were randomly assigned to groups, a positive difference in click rates between the treatment and control groups would indicate decreased newsletter frequency actually increases user engagement. Similarly, the client could also test the effect of newsletter content on engagement by sending newsletters with different topic coverage to different randomly assigned user groups and measuring differences in click rates between groups.

## Further Analysis

For a future observational analysis, our team would like to analyze the following currently unavailable data:

- **Basic user demographic data:** Such as gender, age, and general geographic location. This information could help us split click rates out by user types, and determine which types of users are more likely to access different types of content.

- **More detailed email content data:** Such as the order in which links appear in the newsletter and the newsletter text describing links. This information would help us determine if link placement or description effect clicks, which would help us design experiments testing new ways to present content in order to maximize user engagement.