
# ANLY482 – Analytics Practicum

# Final Report

## AY 2015-2016 Term 2
## Team 6: SkyTrek

Jedaiah TAN
Aseem PRABHAT
Viet Huy NGUYEN

# Table of Contents:

# 1. Sponsor and Background Information

Skyscanner is the leading global travel search site offering an unbiased, comprehensive and free flight search service as well as online comparisons for hotels and car hire.

Skyscanner's flexible search options allows users to browse prices across a whole month, or even a year; allowing users to get the best deals. When you find the perfect deal through Skyscanner, you are redirected to book direct with the airline or travel agent. This ensures customers get the lowest price, with no extra fees.

Skyscanner has been in the travel business for over 10 years and employs more than 50 different nationalities in its global offices in Edinburgh, Singapore, Beijing, Shenzhen, Miami, Barcelona, Glasgow, London, Sofia and Budapest. It has over 50 million unique visitors per month who use it to find flights, car hire and hotels in more than 30 different languages.

On its website, Skyscanner has a travel feature and news section[1]. This helps attract users to Skyscanner through its content marketing activities. The company constantly publishes news articles relating to travel trends, travel tips, top destinations, best deals and new product features in order to constantly engage its users drive more traffic to the site. The project sponsor, Ms Antoinette Tan is the content manager for APAC at the Skyscanner Singapore office. She is incharge of the Skyscanner Travel and News Site for all APAC markets including Singapore, Malaysia and Thailand.

# 2. Motivation

One of Skyscanner's goals is to acquire new users and engage its current users through content marketing on its travel features and news site. The company's goals is to drive more users to the website in order to increase its metric of unique monthly users. This metric has a large impact on revenue as well as the valuation of Skyscanner and similar internet companies. It has been growing at a high quarter on quarter growth rate over the last 2 years - a growth rate Skyscanner wishes to maintain.
As a lean organization, Skyscanner has limited resources for content marketing and hence must use resources in a way to maximize impact. This impact is measured through page views and engagement metrics. Skyscanner believes in the idea of "Build. Measure. Learn"

[1] http://www.skyscanner.com.sg/news/

and hence is constantly conducting experiments such as A/B tests in order to reevaluate and improve its processes.

The Content team has similarly been moving towards a data driven approach over the last year, but there is still a lot of room for improvement. Below is the new process flow for content creation. This process is constantly improved through experiments, feedback and learnings.


Figure 1: *Skyscanner content creation process flow*

## Analytical Problem

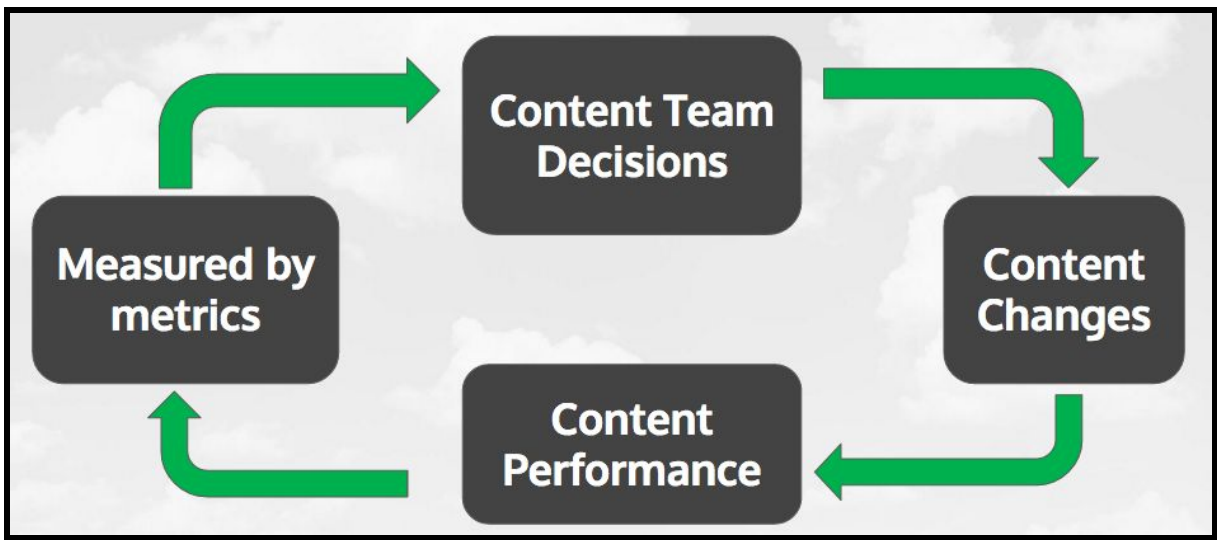



Figure 2: Skyscanner process improvement workflow

The flowchart in Figure 2 depicts the system Skyscanner operates under. The content team makes decisions that lead to content changes. This affects content performance as users react to the changes on the news articles. This is in turn measured by metrics that are interpreted by Skyscanner which leads to new decisions and so on.

# 3. Objectives

The aim of our practicum is to provide deeper insight into the performance of different content articles on the Skyscanner travel and news features site.

The client is a content manager who intends to use the results from our article performance analysis in order to make the most optimal use of resources. This will help decide what kind of content is to be created at different times of the year in order to maximize the number of visitors to the Skyscanner news site.

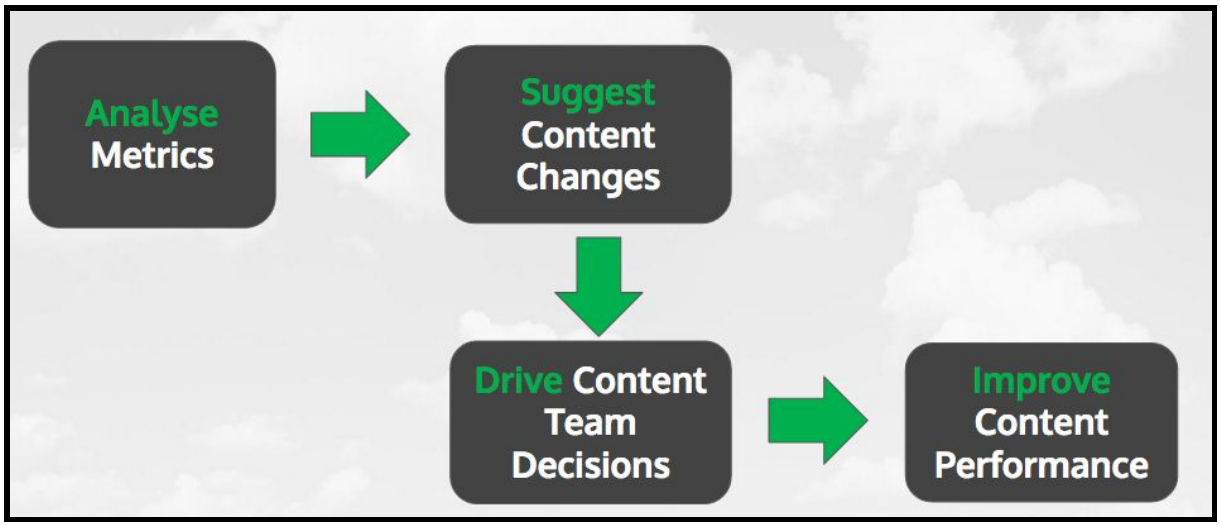



Figure 3: How this project will help Skyscanner

# 4. Project Evolution

The project started off with the primary objective of improving Skyscanner's article content planning process. With this in mind, the team went on to explore a range of related questions detailed in our proposal. While relevant and beneficial to the client, the objectives needed to be streamlined. Based on consultations with our project supervisor Prof Kam our attention was turned to the scope of learning and limited time frame available for this project. The details can be found in our mid-term proposal.

The final deliverables will aim to:

1. Increase organic growth by identifying key article attributes that draw high levels of traffic and interest
2. Validate and possibly identify new content themes (CT)

3. Validate current allocation of resources to the various CT based on evaluated performance
4. Facilitate the content planning process by way of an interactive dashboard
    a. Explore the effectiveness of advertising efforts (paid articles)
    b. Investigate the organic growth of articles (non-paid articles)
    c. Investigate performance of Content Theme Clustering

It was fortunate that the team was able to receive the dataset from the client by the end of week 2. The team conducted the necessary exploratory data analysis and assessed the need for expansion of the dataset to include web scraping of the article contents. In all, there was little fuss over data quality and availability.

# 5. Data Extraction

Our final data cube was derived from 3 main raw data sources, each requiring a different method to pull and transform the raw data.

1. **Skyscanner News Site (Scraping):**

This data is pulled from the Skyscanner news site, for each article, including attributes such as article text, number of links, images, published date etc. This data constitutes public information that is visible to the user and hence did not need to be requested for from the client. These attributes tell us about the nature of the actual content that is being seen by the user.

We noticed that Skyscanner website uses Javascript to add content to the DOM when the HTML finishes loading. Because of that, normal crawlers without ability to execute Javascript code are not be able to crawl for data within the page after the DOM is modified.

After some research, our group decided to employ the use of a headless browser, namely PhantomJS[2]. It allows for Javascript code execution, DOM access, and programmatically interaction with websites without opening any real web browser.

Looking through the DOM structure of Skyscanner article, we found that the information we need is easily accessible. For example, the main article content is nested in a div block with CSS class “main-content”, and the publication date is inside another div block with CSS class “published-date”. The repeating structure makes it easy for us to write code and scrap data from the website without too much trouble.

[2] http://phantomjs.org/


Figure 4: *DOM structure of article text and published date*

After successfully scraping the DOM data, we cleared out HTML tags using a regular expression /<(\/|).+?>/g, then proceeded to compute the necessary attributes that we want to collect.

2. **Google Analytics:**

This data was pulled for each article being tracked via the Skyscanner Google Analytics account. This contained metrics regarding the performance of each URL on the Skyscanner news site. These attributes tell us about the performance metrics of each article mainly through Unique Page Views (UPV) and Average Time on Page (ATOP). It also provides the different sources of traffic such as Facebook paid media or Google organic and the contribution of each of the different online channels.

The Skyscanner Singapore team has created an access account in Google Analytics for this project, allowing us to pull all possible combinations of data from Google Analytics relating to the Skyscanner news site. This is also summarized in the form of different dashboard views available on Google Analytics premium. The method for querying any data is through creation of custom reports.

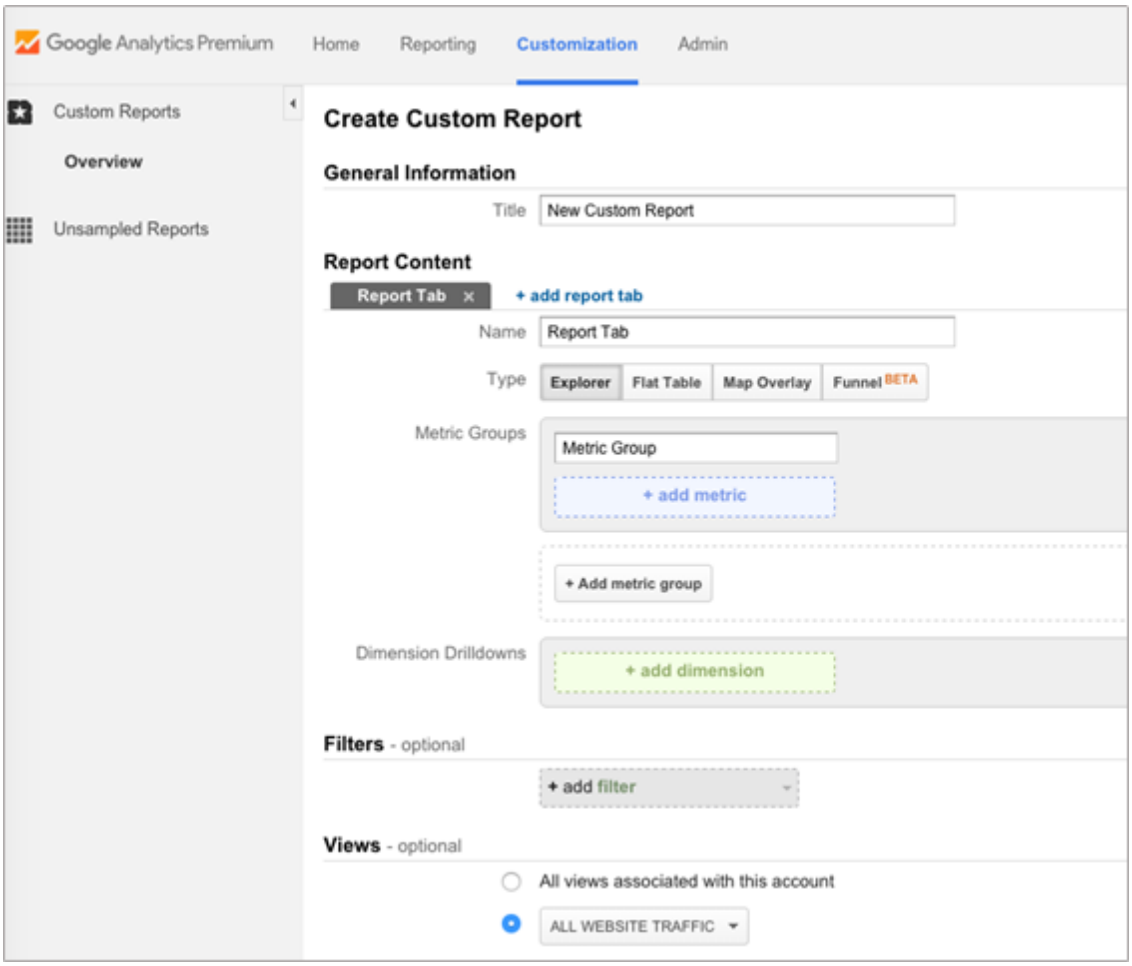



Figure 5: *Custom Report creation process on Google Analytics*

3. **Social Media Shares**:

This data was pulled via continuous scraping of a website called http://linktally.com that provides an API of social media shares for a given URL. These attributes tell us about how widely each article is shared across different social media platforms.

# 6. Exploratory Data Analysis

## 6.1. Input Variables

| Name | Type | Missing | Statistics | Min | Max | Average |
|---|---|---|---|---|---|---|
| no_of_words | Integer | 0 | | 89 | 3277 | 774.897 |
| no_of_links | Integer | 0 | | 0 | 53 | 11.461 |
| no_of_imgs | Integer | 0 | | 0 | 29 | 5.757 |
| unique_pageviews | Integer | 0 | | 0 | 14432 | 305.303 |
| organic_searchs | Integer | 0 | | 0 | 13515 | 233.509 |
| bounce_rate | Real | 0 | | 0 | 1.500 | 0.763 |
| exit_rate | Real | 0 | | 0 | 1 | 0.650 |
| avg_time_in_sec | Real | 0 | | 0 | 995.500 | 162.115 |
| facebook_shares | Integer | 0 | | 0 | 9536 | 94.281 |

Figure 6: Numerical attributes summary statistics

Figure 6 above shows the summary statistics for each of the numerical attributes that will be used in our regression analysis.

## 6.2. Numerical Attribute Distributions

### 6.2.1. Number of words

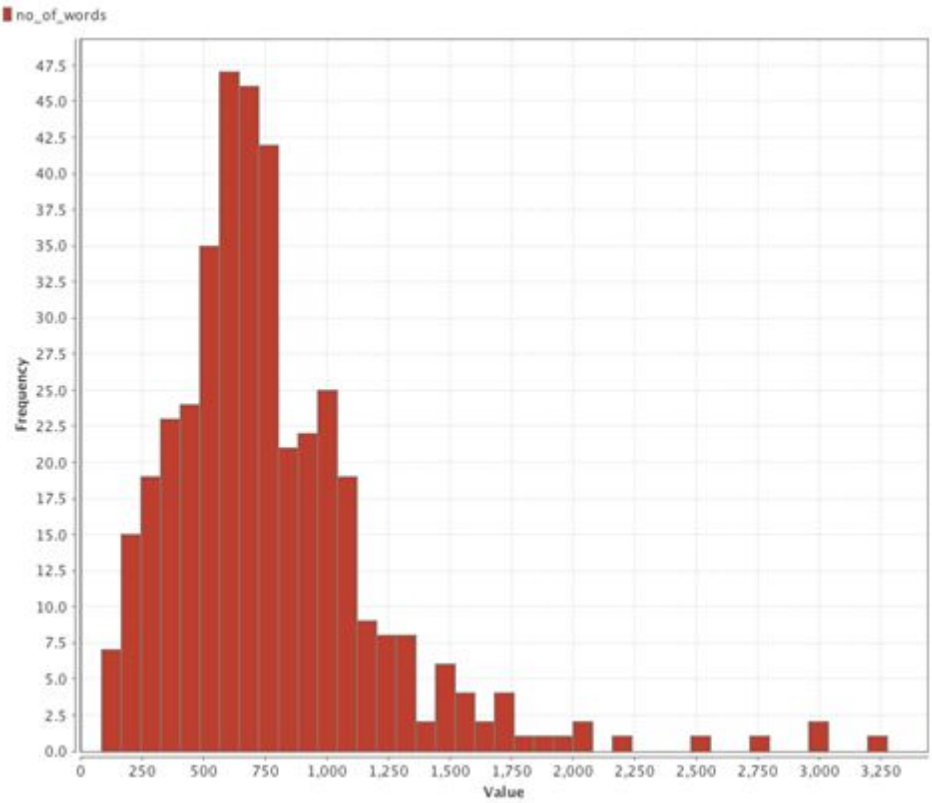



Figure 7: Histogram of number of words in article

This histogram shows that most articles generally have a length similar to the mean i.e. about 775 words. There are some articles to the right that are much larger than average. These outliers could potentially skew the data.

### 6.2.2. Number of links

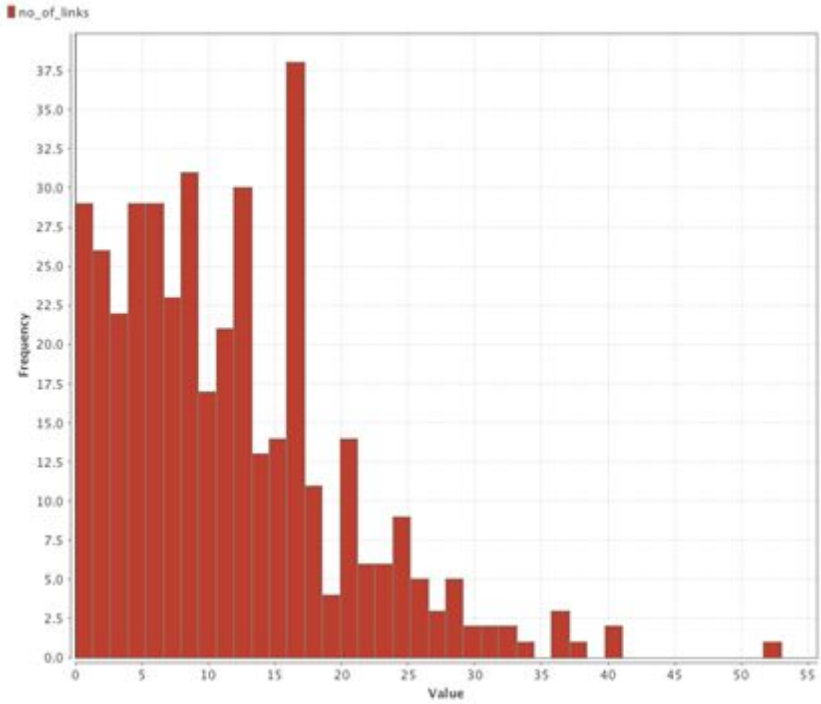



Figure 8: Histogram of number of links in article

The number of links attribute is again evenly distributed around a mean of 11 for most articles. There is one outlier in the data with about 52 links.

### 6.2.3. Number of images

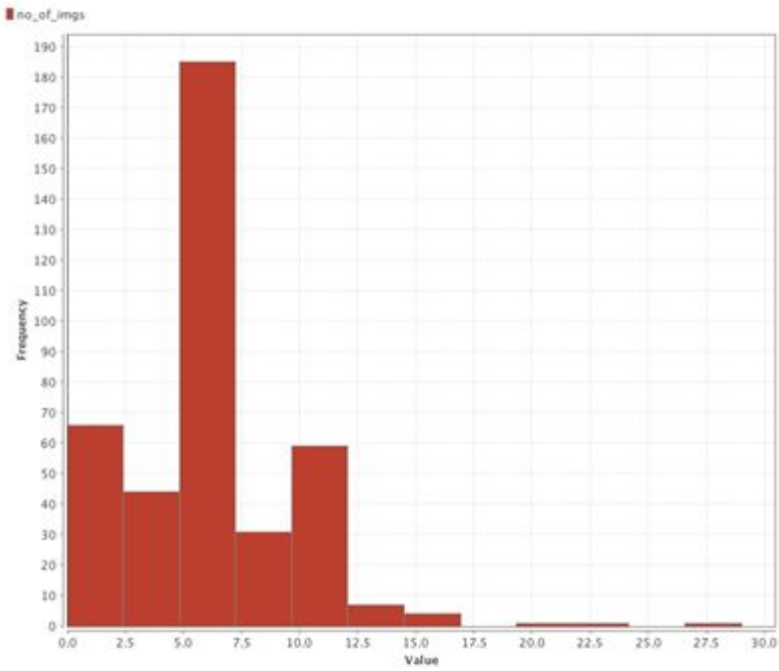



Figure 9:  Histogram of number of images in article

The number of images is again dominated by mean values, with most articles having about 6 images. The main exception here would be articles that do not have any images.

### 6.2.4. Bounce Rate

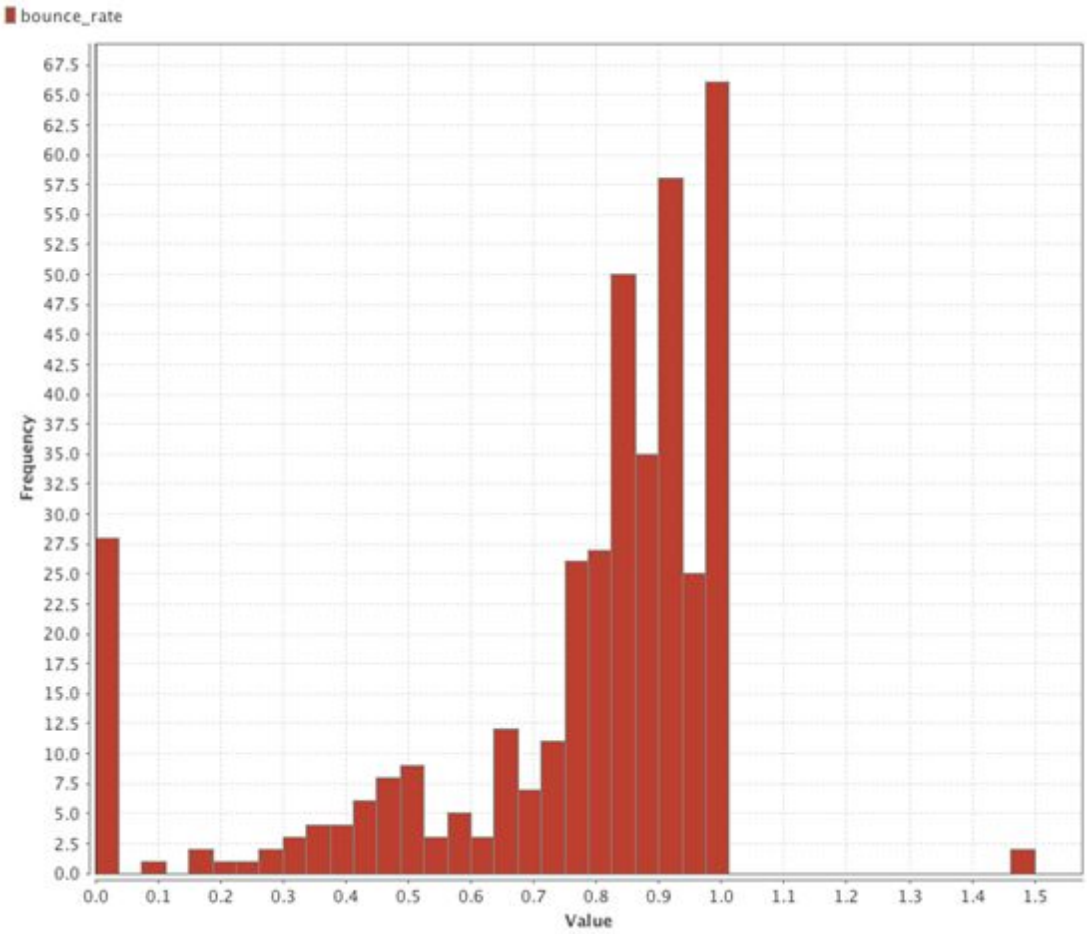



Figure 10: Histogram of article bounce rates

The bounce rate has a mean of about 75% and most articles trend to be a bit above that. This data would be skewed due the 20-30 articles that have a bounce rate of zero. It would be interesting to note that for Skyscanner, 75% would be considered a very good bounce rate. There is one outlier here where there is 150% bounce.

### 6.2.5. Exit Rate

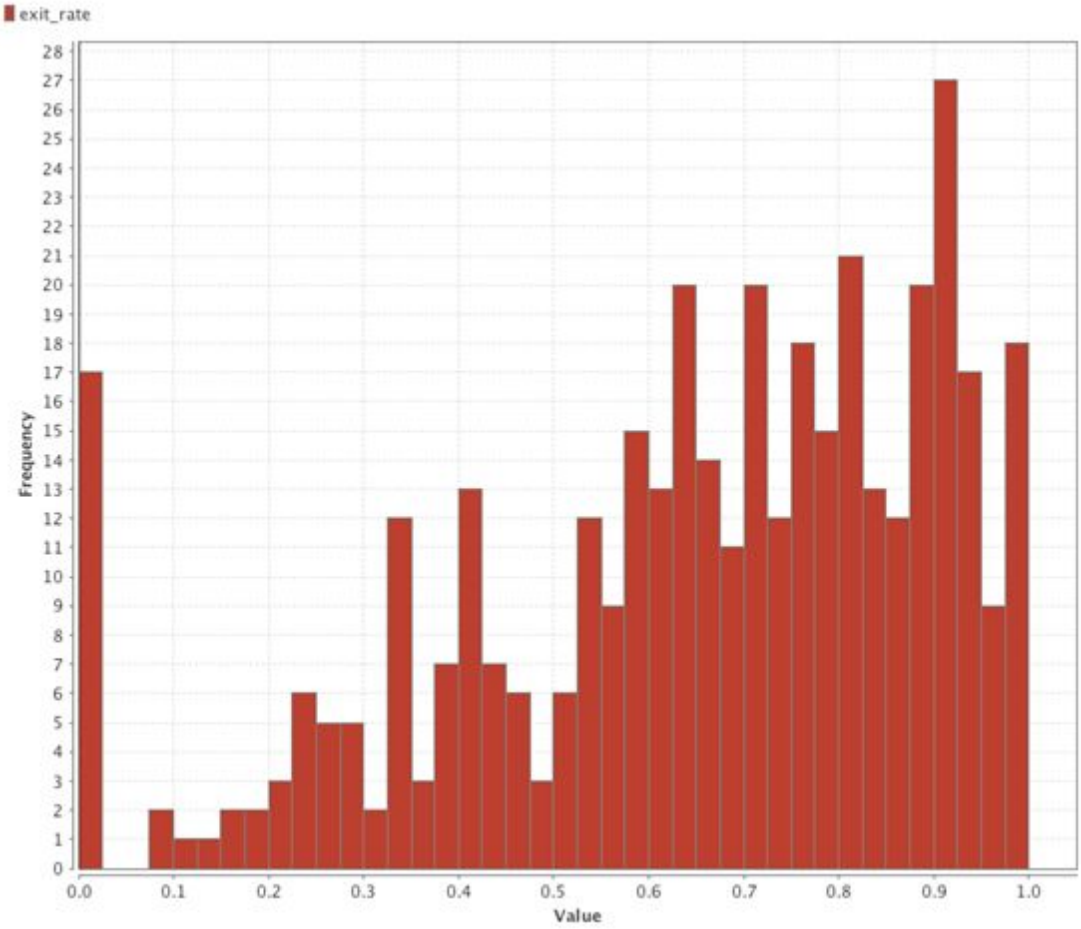



Figure 11: Histogram of article exit rate

The exit rate has a mean of about 65% and most articles trend to be slightly above that. This data would be skewed due the 17 articles that have an exit rate of zero.  It would be interesting to note that for Skyscanner, 65% would be considered a very good bounce rate.

### 6.2.6. Unique Page Views (Dependent Variable)

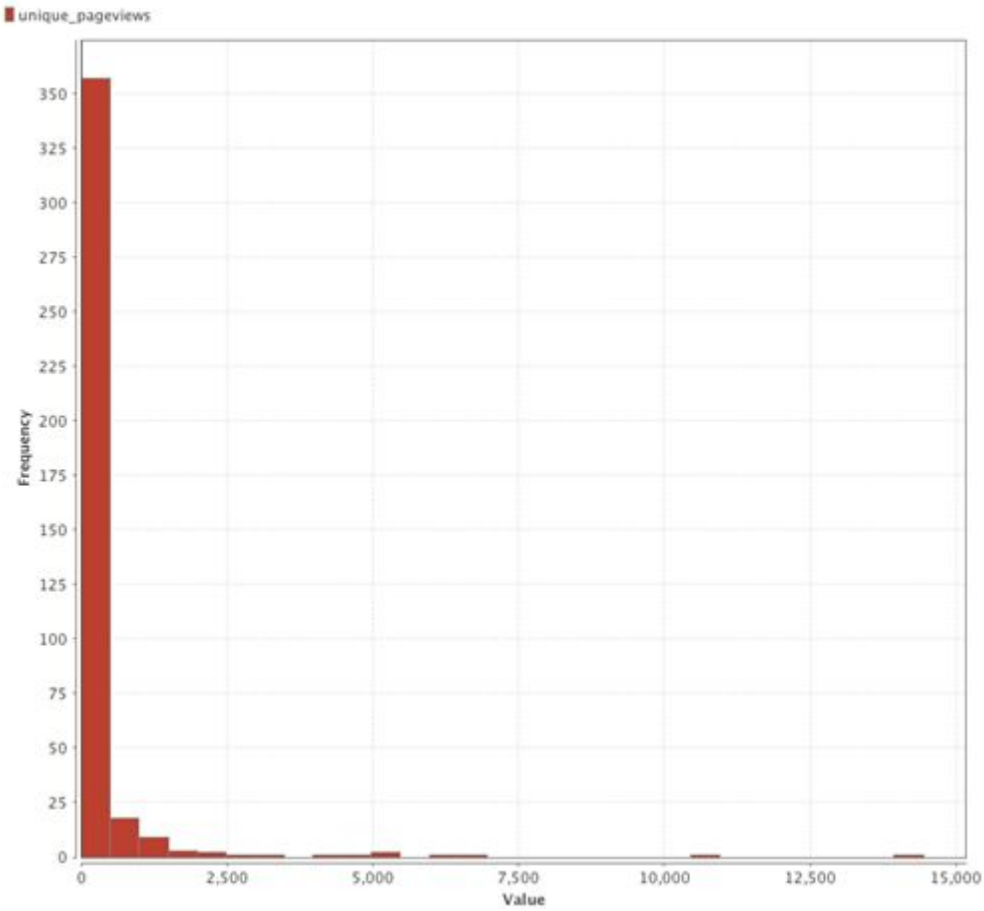



Figure 12: Histogram of article unique page views

The Unique page views attribute is the dependent variable for the regression model. It has a mean value of 305. Most articles tend to be distributed around this mean but there are some exceptionally high values that are over 5000. These articles represent high value articles and hence may need to be kept in the analysis data set.

### 6.2.7. Average Time On Page in seconds (Dependent Variable)

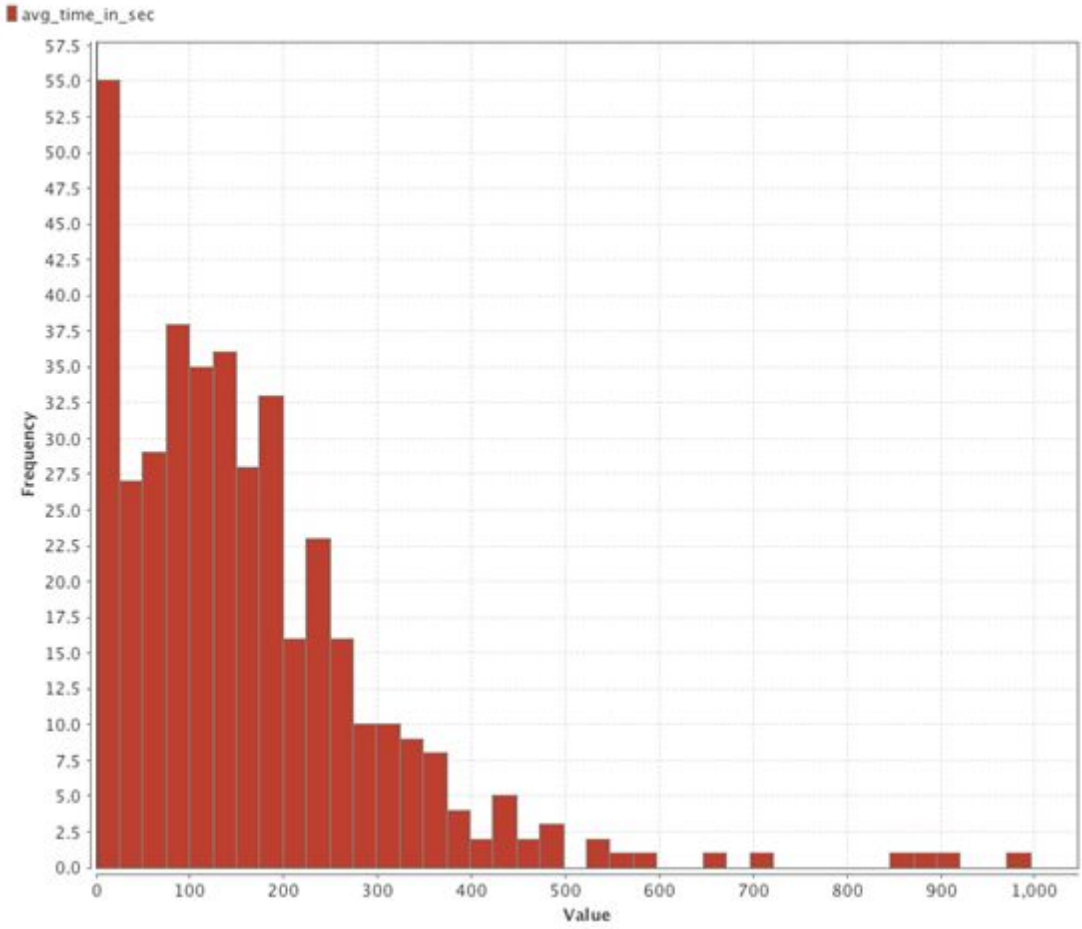



Figure 13: Histogram of article average time on page (in seconds)

The Average Time on Page attribute is another dependent variable for the regression model. It is also an independent variable when we look at the model with UPVs as the target. It has a mean value of 162 seconds. Most articles tend to be distributed around this mean but there are some exceptionally high values that are over 500. It would be interesting to understand what the reason behind these high values is.

# 7. Data Transformation

Once the three raw datasets had been extracted, there was a need to merge them so that the relationship between all of these attributes can be analysed in the next stage of our project. This requires many of the ‘dirty’ URLs to be cleaned out as part of the merging process. The article URL was used as a primary key in order to join all the attributes from the three datasets. The diagram below shows the entire ETL process showing how the three sources of data were extracted, transformed and then stored in a MySQL database.

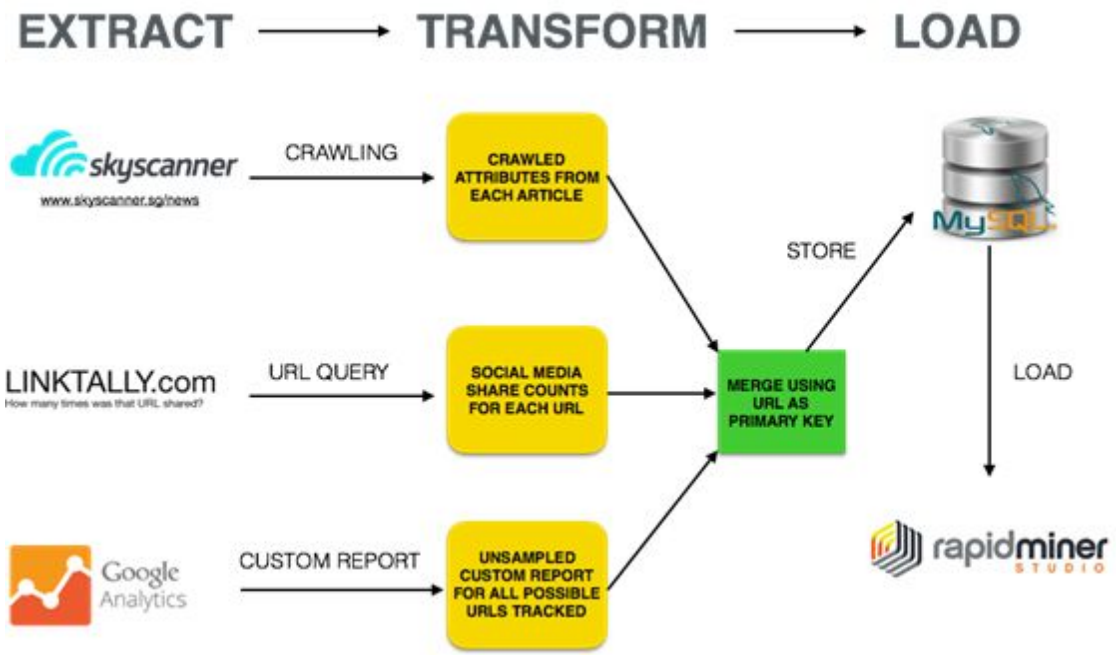



Figure 14: Data Transformation and Merging Process

## 7.1. Data Dictionary

| **Attribute Name** | **Description** | **Type** |
|---|---|---|
| URL | The original URL of the article on the Skyscanner news site | Categorical |
| Source/ Medium | The source of the traffic to the website | Categorical |
| Unique Page Views | The total number of unique views for the given article | Numerical |
| Organic Searches | The total number of organic searches for the article | Numerical |
| Average Time on Page | The average time in seconds that a user spends on the given article | Numerical |
| No. of links | The number of out-links embedded in a given article | Numerical |
| No. of images | The number of images in a given article | Numerical |
| No. of shares | The number of social media shares for a given article | Numerical |
| Published Date | The date at which the article was | Categorical |

| | | |
|---|---|---|
| | published | |
| Article Text | The body of text of the entire news article | Categorical |

## 7.2. Aggregation of data

The final dataset contains about 9621 rows at the most disaggregated level. The identifier for each row is the article URL as well as the ‘source/medium’ for each URL. In order to analyse different aspects of the business problems, this dataset has been divided into different aggregated levels. The most important levels would be based on the traffic source – mainly ‘organic’ (non-paid) VS ‘inorganic’ or ‘paid’ media. The main reason behind this division is that ‘paid’ traffic numbers tend to usually be higher than non-paid ones and hence skew the data in favour of articles that have been distributed through paid channels such as Facebook, Taboola and StumbleUpon. The diagram below shows the different levels of aggregation building up from the most disaggregated (9621 rows) to the most aggregated level (399 rows).


Figure 15: Levels of Aggregation of Data

# 8. Methodology

In response to the business questions mentioned, we have identified a 3 key approaches specified in the table below.

| Business Question of Interest | Analytics Solution Approach |
|---|---|
| To validate and possibly identify new CT | Content Theme Performance Analysis Via **Clustering** |
| To validate current allocation of resources to the various CT based on evaluated performance | |
| To increase organic growth by identifying key article attributes that draw high levels of traffic and interest | Understanding performance of news article attributes based on organic viewership via **Logistic Regression** |
| Facilitate the content planning process by way of an interactive dashboard<br>1. Explore the effectiveness of advertising efforts (paid articles)<br>2. Investigate the organic growth of articles (non-paid articles)<br>3. Investigate performance of Content Theme Clustering | Data Visualization |

## 8.1. Content Theme Performance Analysis Via Clustering

Skyscanner has classified its collection of 399 news articles into 7 CT. They are Skyscanner Product, Practical/Tips, Deals - Prices, Trending, Domestic/Local, City Guides and Inspirational. It would be useful to know which CT is most well received with readers, hence facilitating a more targeted allocation of article-writing resources. The converse also hold true. This analysis predicates on the validity of the CT provided. Hence, it would be prudent to first validate the truth of these pre-identified CT. In order to address this, we employed the use of clustering analysis.

In the clustering of articles, one would intuitively base the classification on the content covered in the article. However, given that our primary objective was to evaluate the performance of the CT discovered from this clustering, we found it more useful to perform clustering on the article titles. After all, given the swarm of information on the internet and social media, the decision to view the article or not is very much contingent on the appeal of the article title.

This then begs the question: Shouldn’t the clusters generated on the article body be similar to those on the article title? While the CT identified were similar, the examples constituting

the clusters were found to be different, leading to different conclusions on the CT performance analysis. There are a few reasons explaining this. The clustering algorithm does not directly emplace the articles into the CT. Therein lies the inherent evil of subjectivity in the profiling and assignment of clusters to new/existing CT. Last but not least, Skyscanner is known to use highly-searched keywords in the naming of their article titles in a bid to drive article performance. This might lead to the forced use of keywords which could differ from the article content. Our analysis will demonstrate the existence of this disparity between CT performance between article title and body. After considering the order of activity flow in a typical decision of whether to read an article or not, we find the the article title to take precedence. Hence, we shall base our recommendations on the article title analysis.

### 8.1.1. K means Clustering Process

In performing unsupervised clustering, we have selected the K-means algorithm. This is primarily due to its strong support on the RapidMiner platform (our analytics tool of choice). RapidMiner provides components which allow us to quickly evaluate a good value of K - a process which we will cover in detail below. One might also be interested in the fact that the K-means algorithm is widely employed for clustering purposes. The ensuing sections will demonstrate the steps involved, explaining the component functions and reasons supporting its configuration. The steps below were applied to the article content. Hence, whilst the steps involved are largely the same in clustering for both the article titles as well as the article content, some of the results discussed are based on article content.

#### 8.1.1.1. Selecting a Good K-Value for Article Content Clustering

This algorithm requires the user to specify the value of K. This value is typically identified by way of a trial-and-error process, made more difficult by the subjective nature of deciding what constitutes ‘correct’ clustering (Dimov et al, 27 Sept 2004). Can & Ozkarahan also suggests that the number of clusters in a text database can be roughly estimated by by the following formula $\frac{m*n}{t}$ where m is the number of documents, n is the number of terms and t is the number of nonzero entries in the term document matrix (Can & Ozkarahan, 1990). We will be making use of both suggestions in coming up with the value of K to select.

Figure 16 below demonstrates the components involved in this process.

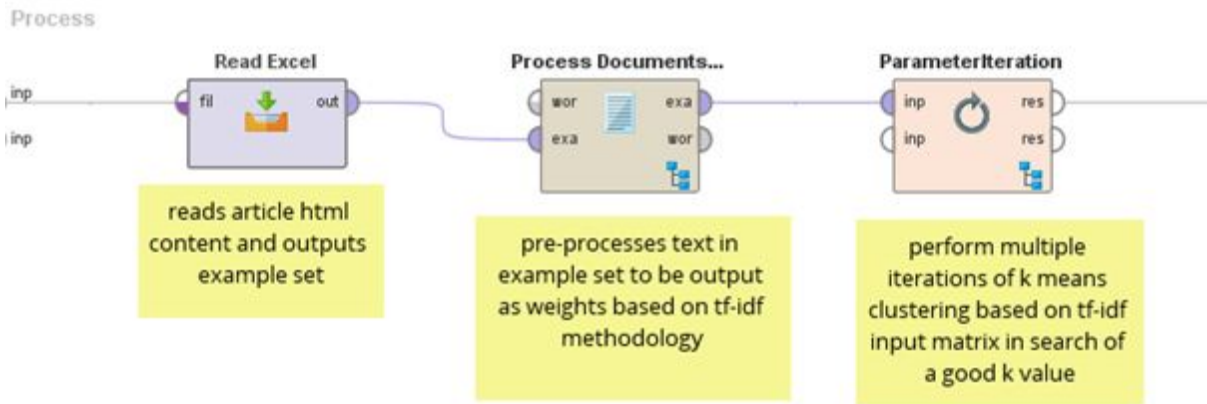
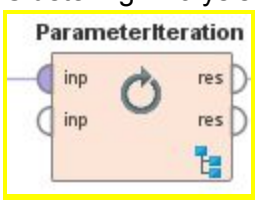



Figure 16: Determining the Optimal K Value

8.1.1.1.1. Elaboration on Components Within Process of Selecting Good Value of K


This operator reads the article html from the excel spreadsheet and outputs an example set


This operator processes the html and generates a tf-idf document matrix. We will talk more about the sub process involved later on in section 8.1.1.2.2. Data Preprocessing for Clustering Analysis.




This operator iterates over its sub process (clustering, model evaluation and logging), with a different input value of k for each iteration. The clustering model is generated, evaluated and logged on each iteration. The logged results across the iterations will allow us to make an informed selection of a good K value.

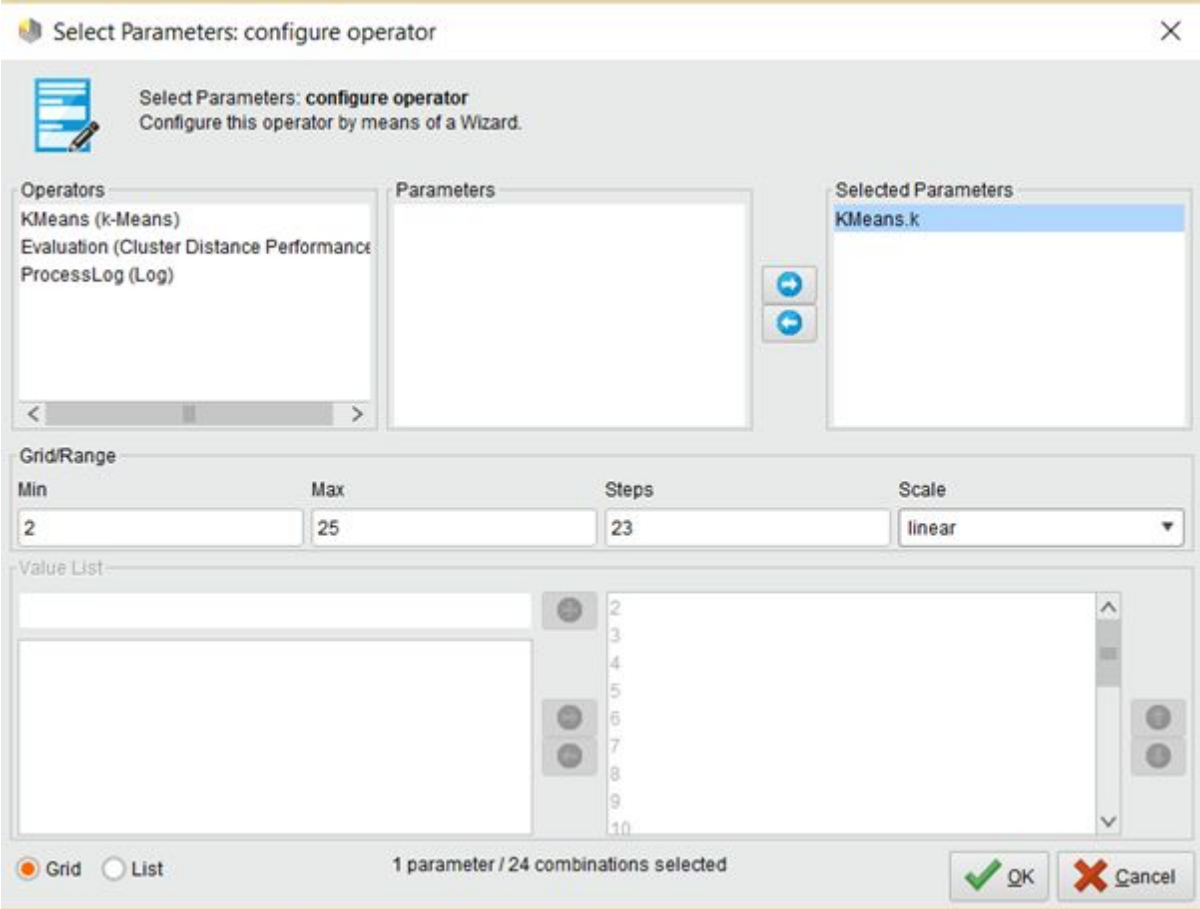



Figure 17: Wizard for configuration of parameter in parameter iteration component

Selecting the 'Edit parameter settings' button from would open the 'Select parameters: configure operator' wizard. This wizard helps us configure the iteration over the K values to explore. This being an exploratory exercise, the min and max values of 2 and 25 respectively, are first arbitrarily chosen. We will take single value increments, hence the step value of 23 with linear scaling.

Figure 18 captures the Parameter Iteration sub process mentioned earlier.

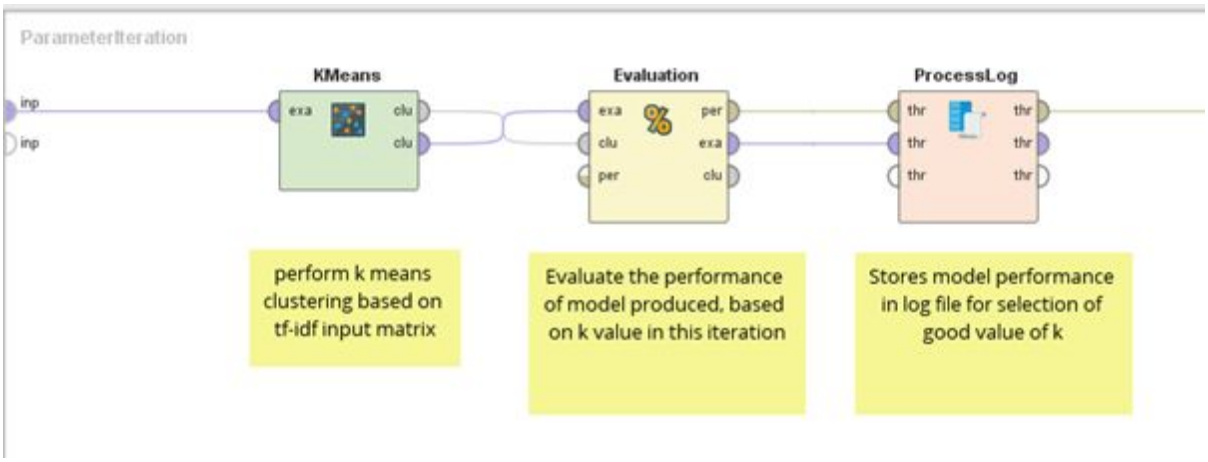
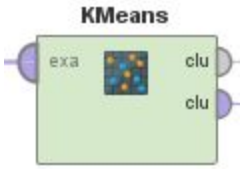
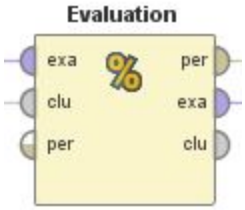


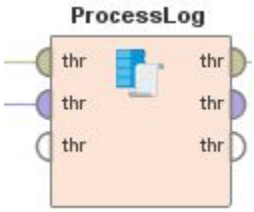

Figure 18: Parameter Iteration subprocess.Clustering| Evaluation | Logging




This is the K-Means clustering operator. This is the operator whose K value will be incremented with each iteration. Hence, there is no need to specify the K value here. Since we are processing the tf-idf weighted word matrix which is a numerical measure, we will select the corresponding option under the ‘measure types’ field. This model will also use the ‘euclidean distance’ option under the ‘numerical measure’ field. The ‘max runs’ and ‘max optimization steps’ field will be left at the default values of  10 and 100 respectively.




This operator evaluates the cluster model from each iteration, built using a different value of K.




This operator stores the results from each evaluation of the cluster model, built using a different value of K. These will then be used to chart a graph shown in Figure 19, which we will use to select a good value of K.

8.1.1.1.2. Interpreting the Results Generated From Varying K Values

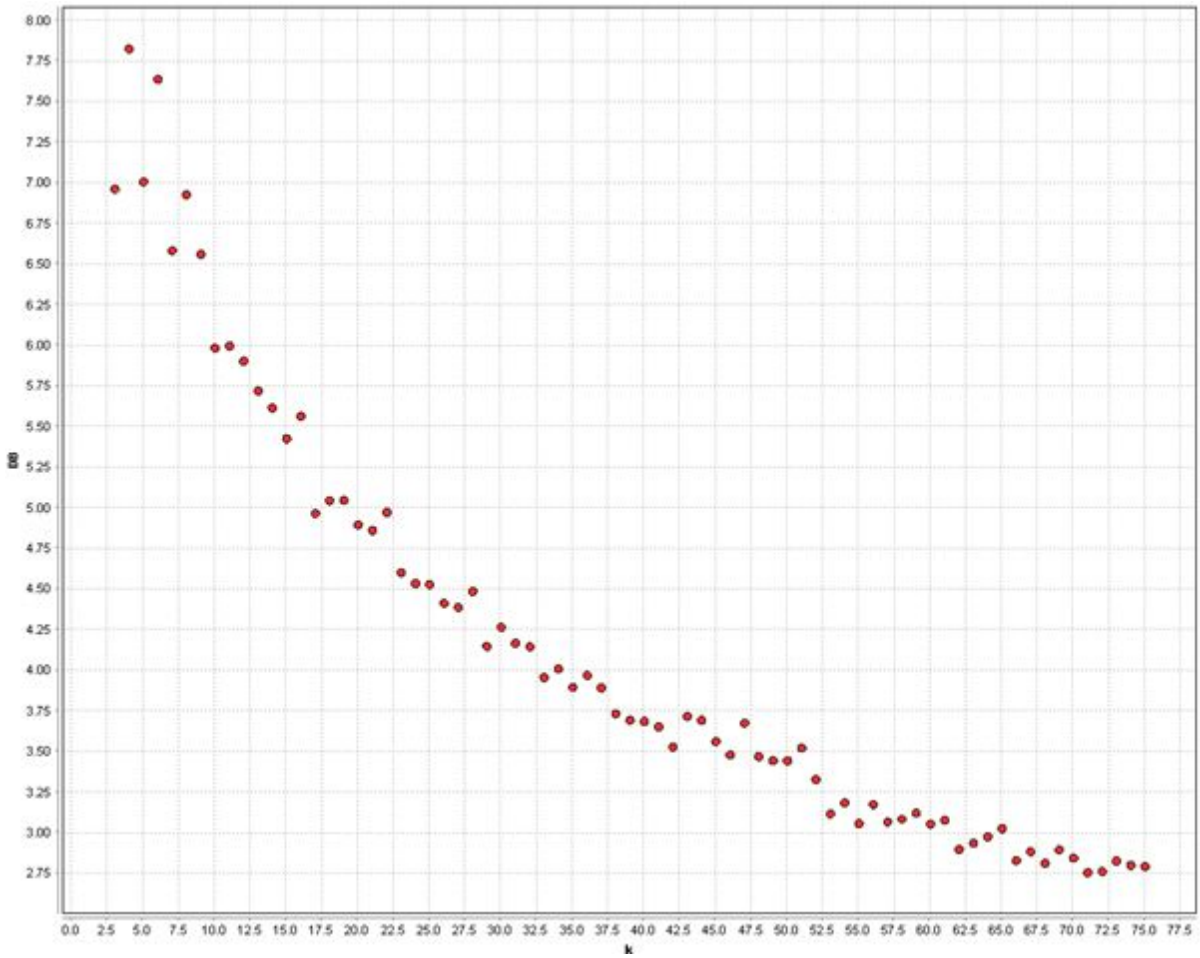



Figure 19: Davies Bouldin Index against K clusters (Article Content)

In clustering, we seek to reduce the intra-cluster distance while maximizing the inter-cluster distance. The Davies Bouldin Index (DB), captured by the the scatter plot in Figure 19 captures this information, with the ideal being a lower value. We see a general improvement in DB as the number of clusters (K) increases. From 70 clusters onwards, this improvement starts to taper off significantly. In determining the value of K to use, Dimov et al's recommends to make the K selection based on the project specific requirements (Dimov et al, 27 Sept 2004). Recalling Can & Ozkarahan's suggested formula for estimating the number of clusters, we found the following results:

m = 399
n = 17996
T = 100739
K = 399*17996/100739 = 71.3 = 71 (floored)

The value of 71 very nicely coincides with the recommendations from Figure 19 Having performed clustering with the aim of reducing manual labelling, we decided to go ahead with this value of 71 clusters, inspecting the generated clusters for meaningful profiling before considering any other action.

#### 8.1.1.2. Clustering with the Good Value of K

After having found a good value of K, we can now run the process below, with the good K value used in the 'Clustering' component.

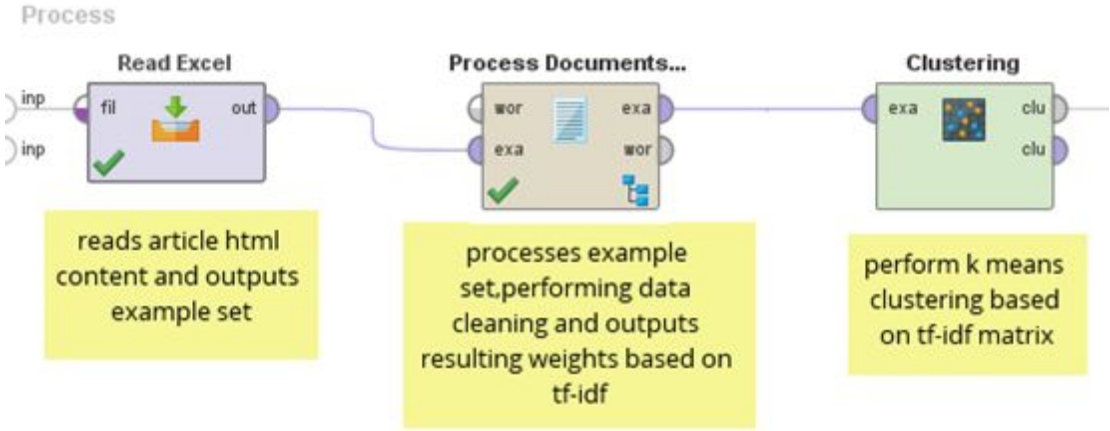



Figure 20: Clustering Process

##### 8.1.1.2.1. Elaboration of Components Within Clustering Process


This operator clusters the documents based on the input tf-idf document matrix. The good value of K found will be used here. All other settings will be similar to those defined earlier in our K-Means clustering operator.


Our clustering algorithm seeks to group document (Skyscanner articles) with similar content in the same cluster. Hence, constructing a TF-IDF matrix would be most appropriate. The term frequency (TF) measures how often a term appears in a document, giving greater weight for a higher frequency. Inverse document frequency (IDF) measures how often a term is used across documents. Terms with a lower occurrence across documents are viewed as less commonplace and hence more effective in characterising an article.

### 8.1.1.2.2. Data Preprocessing for Clustering Analysis

Figure 21 captures the Process Documents from Data subprocess involving data cleaning and generation of TF-IDF document matrix - mentioned earlier.

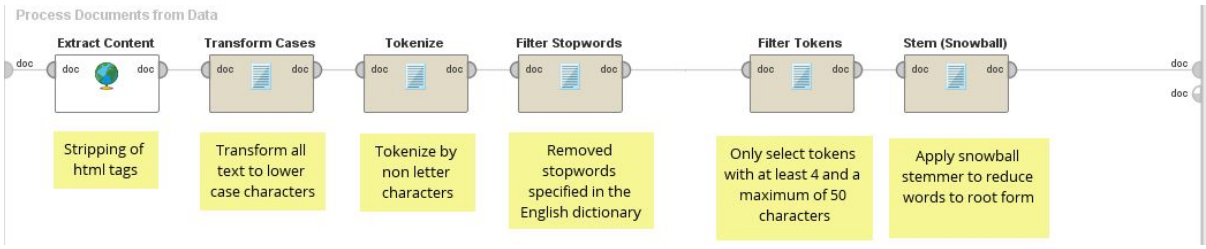
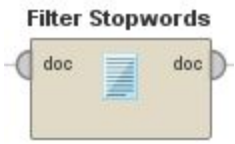



Figure 21: Process Documents from Data subprocess. Data cleaning & Generation of TF-IDF Document Matrix


The input text is html. We stripped the text of html tags such as <p>, <a>, etc. This leaves us the remaining text as the article words


In order not to count 2 words with the exact same alphabet combination but of different case as 2 separate instances, we standardized all words to be of lower case.


This operator set the delimiters to be any non-letter character, used to separate our tokens (article words)

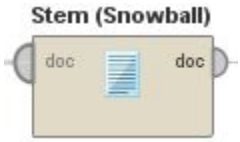



This operator uses RapidMiner's default list of English stop words[3] in identifying common recurring words like conjunctions, pronouns, prepositions, etc which have little value in identifying a document's content.

---

3 http://web.archive.org/web/20080501010608/http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_wordsa


This operator filters for tokens with a minimum of 4 characters and a maximum of 50 characters for our clustering analysis.

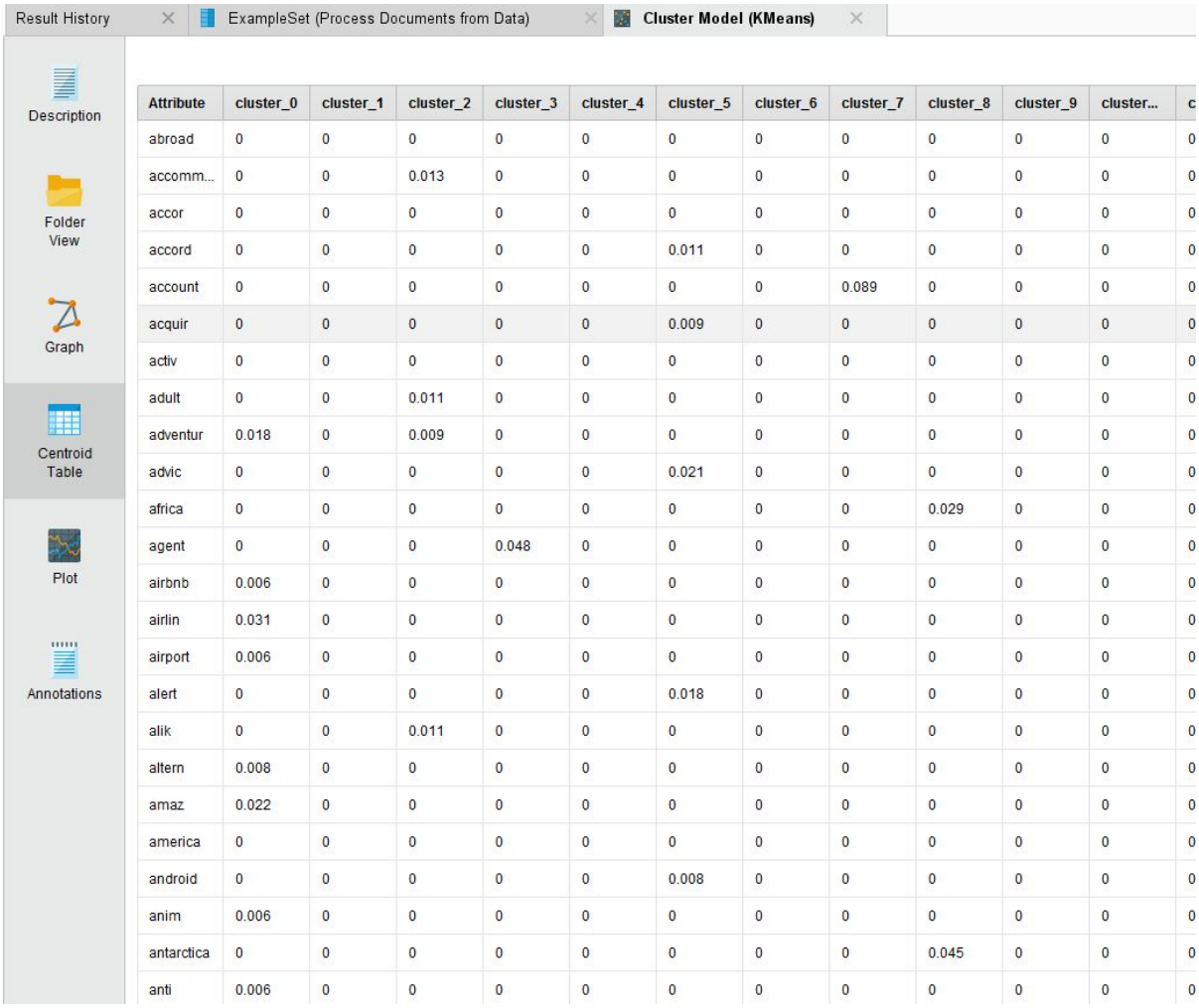



Porter's Stemmer is the most popular amongst the natural language community, known to produce the best output with the lowest error rate (Jivani, Nov 2011). Snowball was developed by Porter himself and is known to perform with a slightly faster computational time. It is for this reason we decided on the Snowball stemmer over the mainstream Porter's stemmer.

#### 8.1.1.3. Interpreting the Results

The clustering process would output the table shown in Figure 22 below. This table is then copied over to excel. For every cluster, we would sort the values in descending order to identify the features (words) which are most representative of the cluster. The profiling of the cluster would be based on the best representative features. Once each cluster has been

profiled, we binned the cluster profiles into the relevant CT.




| Attribute | cluster_0 | cluster_1 | cluster_2 | cluster_3 | cluster_4 | cluster_5 | cluster_6 | cluster_7 | cluster_8 | cluster_9 | cluster... | c |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| abroad | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| accomm... | 0 | 0 | 0.013 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| accor | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| accord | 0 | 0 | 0 | 0 | 0 | 0.011 | 0 | 0 | 0 | 0 | 0 | 0 |
| account | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.089 | 0 | 0 | 0 | 0 |
| acquir | 0 | 0 | 0 | 0 | 0 | 0.009 | 0 | 0 | 0 | 0 | 0 | 0 |
| activ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| adult | 0 | 0 | 0.011 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| adventur | 0.018 | 0 | 0.009 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| advic | 0 | 0 | 0 | 0 | 0 | 0.021 | 0 | 0 | 0 | 0 | 0 | 0 |
| africa | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.029 | 0 | 0 | 0 |
| agent | 0 | 0 | 0 | 0.048 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| airbnb | 0.006 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| airlin | 0.031 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| airport | 0.006 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| alert | 0 | 0 | 0 | 0 | 0 | 0.018 | 0 | 0 | 0 | 0 | 0 | 0 |
| alik | 0 | 0 | 0.011 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| altern | 0.008 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| amaz | 0.022 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| america | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| android | 0 | 0 | 0 | 0 | 0 | 0.008 | 0 | 0 | 0 | 0 | 0 | 0 |
| anim | 0.006 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| antarctica | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.045 | 0 | 0 | 0 |
| anti | 0.006 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure 22: Cluster centroid table

Figure 22 demonstrates the binning of the profiled clusters to the appropriate CT. The data dictionary for coded CT values are shown in Figure 24.

| cluster_assignment | Label_based_on_matrix | Label_based_on_actual_titles | Document Count | UPV | ATOP (seconds) | Highest Wordcount | CT Bucketing Assignment |
|---|---|---|---|---|---|---|---|
| cluster_0 | instagram rediscovery pictures venues garden | INSTAGRAM able venues | 11 | 7329 | 1482.48651 | instagram | 8 |
| cluster_1 | flight help booking alert price holiday value | FLIGHT related topics | 16 | 50879 | 1953.79801 | flight | 6 |
| cluster_2 | free contest festival_names | 0 | 22 | 238758 | 3781.68685 | free | 1 |
| cluster_3 | awesome part visit reason country halloween | AWESOME reasons to do something | 16 | 53618 | 2466.4229 | awesome | 1 |
| cluster_4 | try | suggestions must TRY activities(food places things_to_do) | 11 | 5606 | 1251.11227 | try | 8 |
| cluster_5 | ??? | far too varied to meaningfully classify | 60 | 112153 | 6994.41061 | 5 | 0 |
| cluster_6 | singleton_Top 5 Things to Do in Chile | singleton_Top 5 Things to Do in Chile | 1 | 53 | 67.44444 | | 2 |
| cluster_7 | new_zealand australia road_trip drive | New Zealand city guide | 8 | 22197 | 546.17442 | new zealand | 2 |
| cluster_8 | end catch concert show festivals | year_end start_year festival holiday suggestions | 5 | 6742 | 3404.72969 | 2015 | 4 |
| cluster_9 | top things to do in XXX | top X things to do in XXX | 27 | 97985 | 1442.32819 | things | 1 |
| cluster_10 | restaurant dim_sum hong_kong Singapore | 10 restaurants hong_kong | 8 | 17652 | 2430.16525 | restaurants | 9 |
| cluster_11 | hotel staycation deals | staycation | 16 | 144917 | 5646.19677 | staycation | 8 |
| cluster_12 | asia top family romantic trip destinations | ASIA top family romantic trip destinations | 37 | 2489 | 970.36013 | asia | 1 |
| cluster_13 | travel tips | travel tips | 7 | 2383 | 883.94385 | tips | 6 |
| cluster_14 | airport gifts | airport gifts | 6 | 154660 | 6197.39845 | 10 festivals gift | 1 |
| cluster_15 | singapore celebrity hipster café neighbourhood | singapore related topics (broad range) Top X \| Best XX \| 5,6,7 ways to XXX | 37 | 66603 | 1528.20482 | singapore | 3 |
| cluster_16 | spot beautiful hk couple | different types of spots in asia | 14 | 75406 | 4547.41182 | spots | 8 |
| cluster_17 | travel know advice | topics/festivals/advice to know about | 20 | 106668 | 5978.29223 | travel | 6 |
| cluster_18 | places in hk | Hong Kong <number> places to <verb> in <location> | 52 | 0 | 0 | hongkong | 2 |
| cluster_19 | skyscanner travel award deals | skyscanner topics | 25 | 0 | 0 | skyscanner | 7 |

Figure 23: Binning profiled clusters to the appropriate Content Theme

| Content Theme Description | Code |
|---|---|
| unknown | 0 |
| Inspirational- eg Top festivals, top foods, traditions etc | 1 |
| City Guides | 2 |
| Domestic/Local- Singapore related | 3 |
| Trending- Game of thrones, CNY etc | 4 |
| Deals - Prices | 5 |
| Practical/Tips | 6 |
| Product- Skyscanner feature related | 7 |
| Activity/topic discussion | 8 |
| Food | 9 |

Figure 24: Data dictionary for content theme code assignments

### 8.1.2. Validation and Identification of Possible New Content Themes

In the course of doing so, we found all 7 CT to be represented in the article. However, we felt it was appropriate to generate 2 new CT, 'Activity/topic discussion' and 'Food' since they represented 17% and 8% of articles respectively. The following pie chart shows the proportion of each CT. Once again, it should be noted that this is result of clustering on article content.

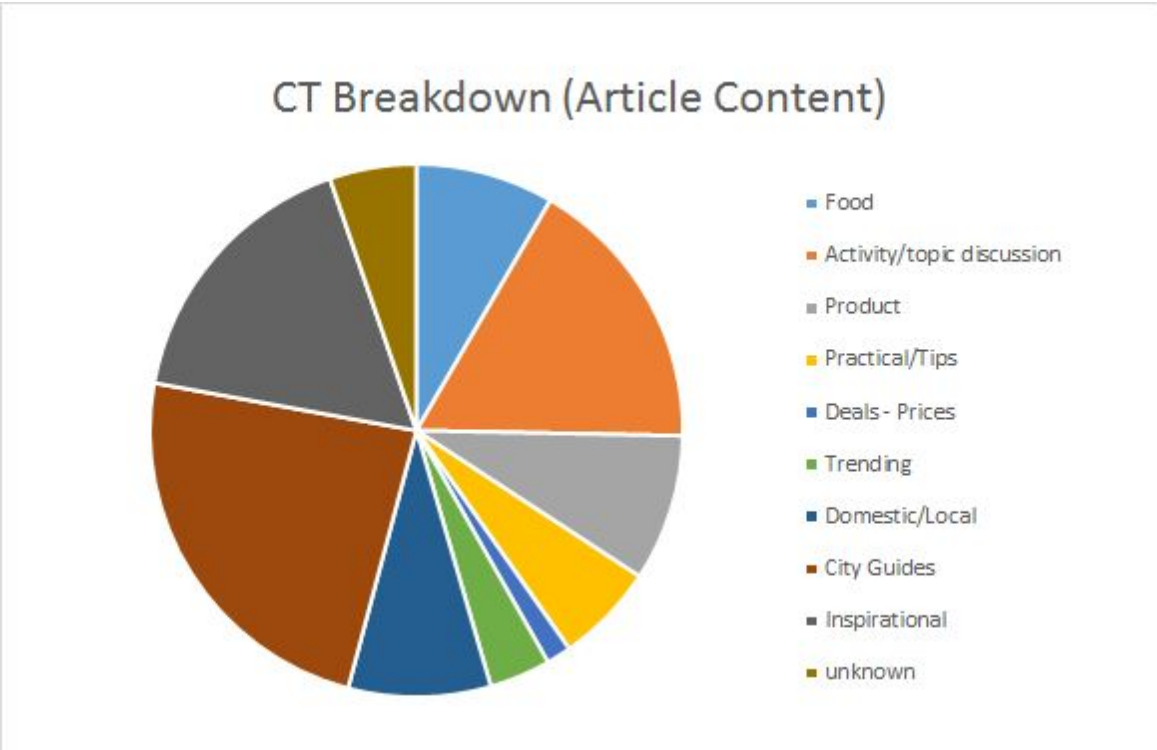



Figure 25: Proportion of Articles in each CT. Article Content Clustering

We proceeded to do the same analysis for article titles in Figure 26 below and found all but the CT Deals - Prices to be represented.

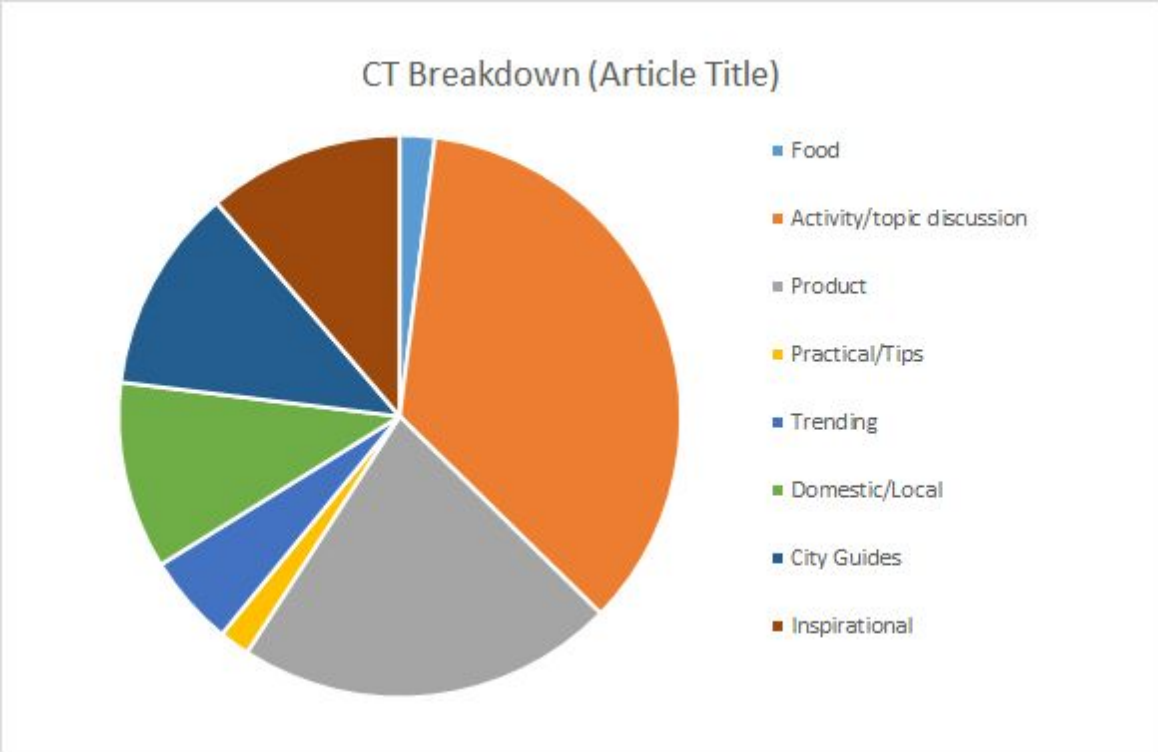



Figure 26: Proportion of Articles in each CT. Article Title Clustering

Figure 27 makes clear the number of articles binned in the 9 CT we have. The discrepancy is clearly existent. Upon review of the series of decisions a reader makes (articulated in Figure 28) in determining whether to invest time in reading the article, we have found article titles to have greater influence. The title is essentially the gatekeeper in determining whether the reader even gets to evaluate the article content or not. Needless to say, if the article does not incite the reader to reader the article, there would be no metrics to gather. Hence, we will be focusing ensuing recommendations based on the article title clustering.

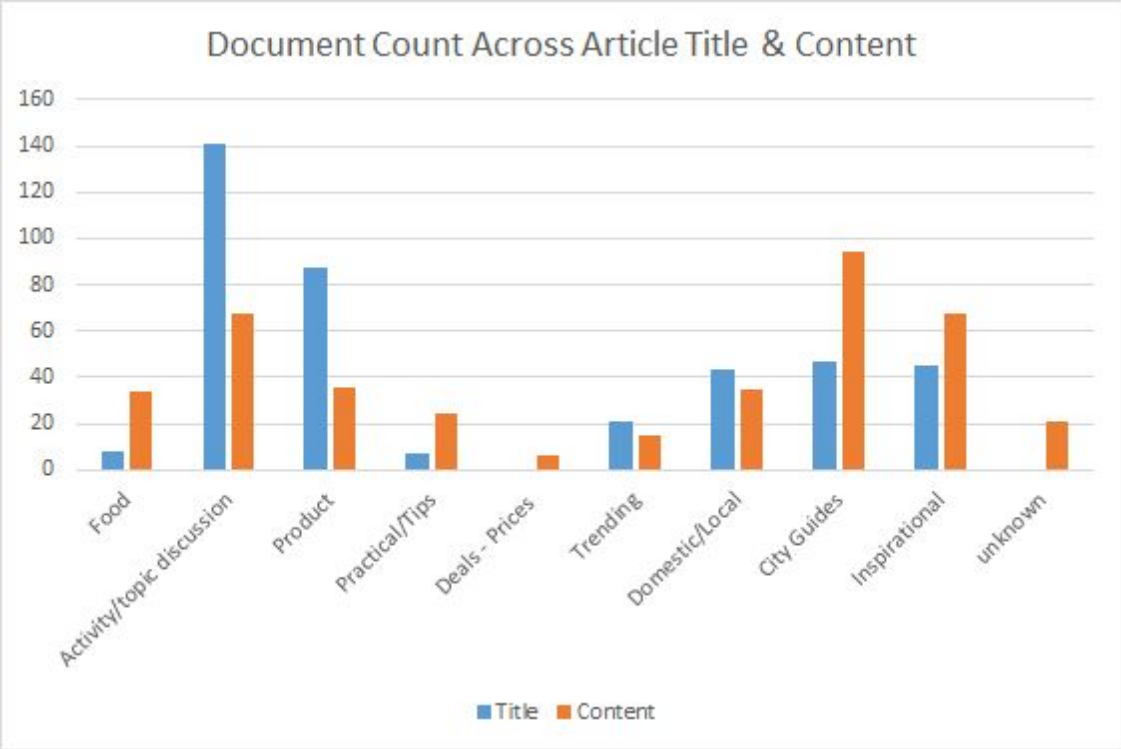



Figure 27: Document Count Across Article Title & Content

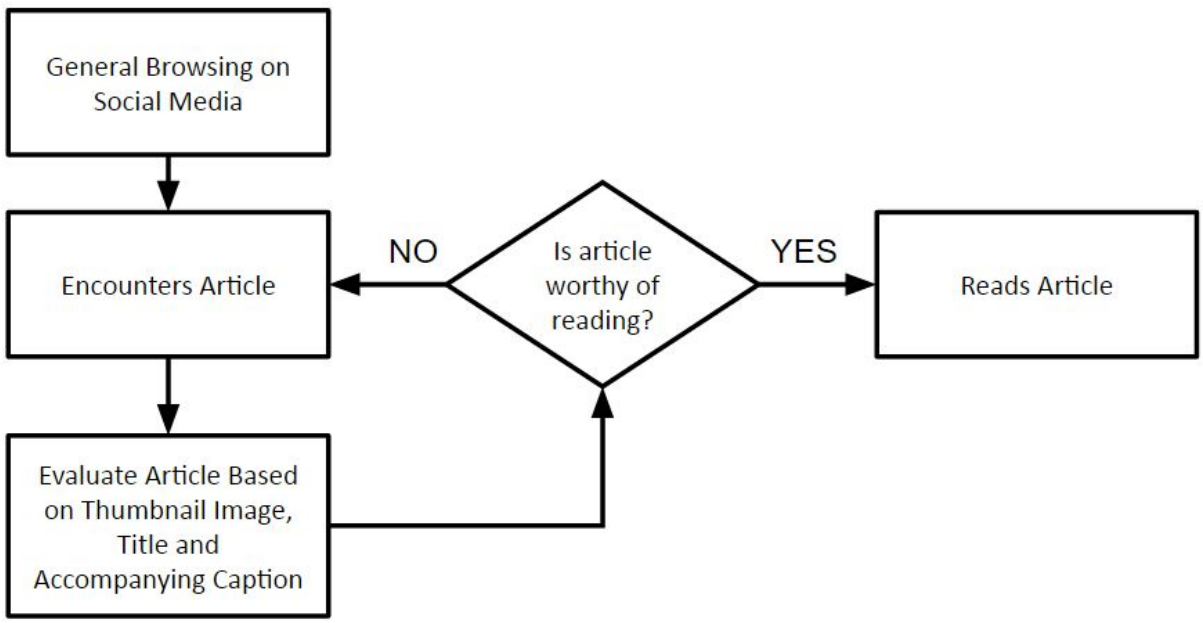



Figure 28: Decision reader makes in deciding if article is worthy of reading

### 8.1.3. Selection of Good K Value for Article Title Clustering

In clustering for the article title, we performed similar exploratory analysis in the selection of K. Figure 29 demonstrates the plot of the DB index values against various K values. We performed an evaluation for both K = 18 and K=20. In profiling the clusters for K=18, multiple clusters were found to consist of multiple CT (see Figure 30). This prompted us to further explore the K value of 20 which was eventually found to be a good value.

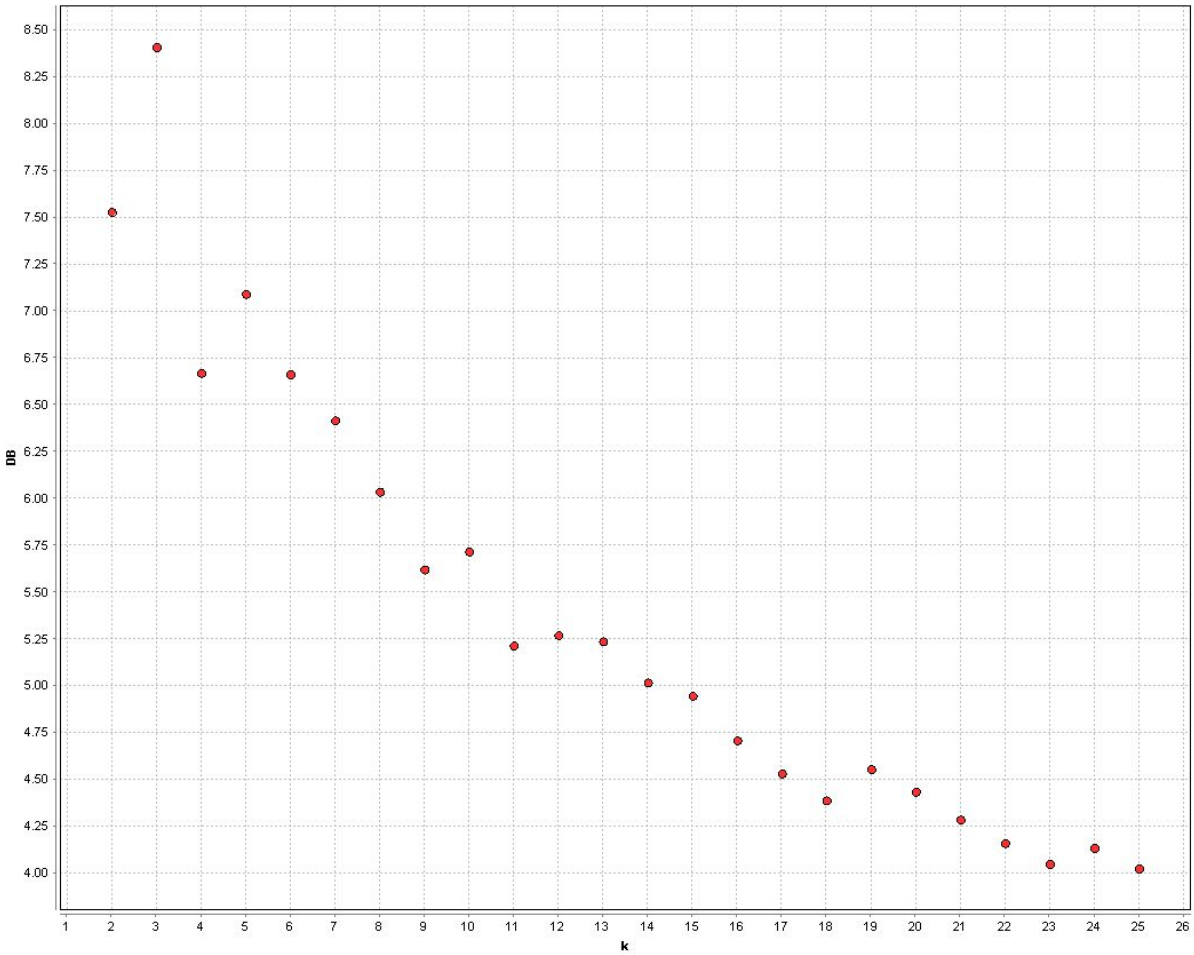



Figure 29: Davies Bouldin Index against K clusters (Article Title)

Figure 29: :

| cluster_as | Label_based_on_matrix | Label_based_on_actual_titles | Document Count |
|---|---|---|---|
| cluster_0 | instagram rediscovery pictures venues | locations to discover | 12 |
| cluster_1 | flight skyscanner book time deal get app money | skyscanner related topics | 19 |
| cluster_2 | free worth win contest <different_festival_types | guides & tips \| festival topics \| activity recommendations (very varied covereage) | 24 |
| cluster_3 | awesome visit part reason <sea_countries> | reasons to visit place / do something | 17 |
| cluster_4 | try mtr food spicy workout | food/place/activity to try around hk MTR area | 11 |
| cluster_5 | world city look adventure vacation holiday tour g | tours events from different places \| skyscanner topics | 68 |
| cluster_6 | chile | chile | 1 |
| cluster_7 | end catch concert show festival | year end recommendations | 5 |
| cluster_8 | things top <country> unique | top <number> things to do in <country> | 28 |
| cluster_9 | restaurant ty dim sum hong kong vegetarian | restaurants | 8 |
| cluster_10 | year staycation chinese celebrations | chinese_new_year recommendations \| staycation topics | 16 |
| cluster_11 | asia family trip fun top holiday | Asia recommendations | 42 |
| cluster_12 | tip travel taipei fit health safe | tips coveraging topics (travel,money,health) | 7 |
| cluster_13 | want gift airport list festival wifi miss music epic | want or unwanted things | 6 |
| cluster_14 | singapore celebrate hipster top brunch neighbou | Everything singapore | 43 |
| cluster_15 | spot escape beautiful late hk | different kind of spots around asian countries | 14 |
| cluster_16 | travel know skyscanner award | skyscanner news/generated content \| random topic discussion \| travel topics | 32 |
| cluster_17 | place hong kong | hong kong \| range of things to do at PLACES | 46 |

Figure 30: Large clusters with multiple CT associations

## 8.1.4. Article Title Clustering Results Analysis

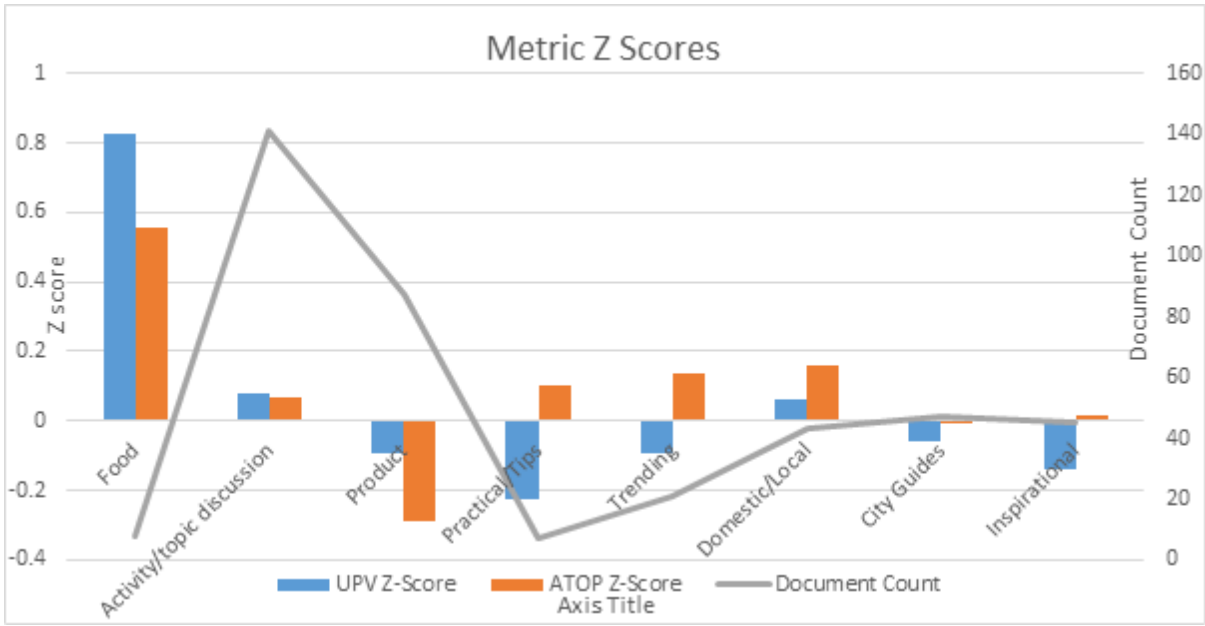



Figure 31: CT Metric Z-Scores based on article title clustering

## 8.1.5. Recommendations

#### 8.1.5.1. Resources used for writing Articles should be redirected to Food CT

The greatest number of articles are titled under the ‘Activity/topic discussion’ CT. However, the z-scores for both Unique Page Views (UPV) and Average Time on Page (ATOP) are both performing at just above average levels. This might be indicative of ineffective direction of resources. Instead, one would be inclined to shift the focus to writing food centric articles, which is currently the second least written about CT, but with the highest UPV and ATOP performance.

#### 8.1.5.2. CT to Promote and CT to Avoid

We would recommend Skyscanner direct more resources to the Food CT in view of its strong metric performance. Conversely, we the weak performance of the Product CT deems an avoidance recommendation. However, if this was purposefully done for brand product awareness, another CT to avoid would be Inspirational Topics.

# 8.2. Understanding performance of news article attributes based on organic viewership via Logistic Regression

## 8.2.1. Rationale

Logistic regression is different from linear regression and other predictive model, as it requires a binary dependent variable. This is based on an outcome of ‘Success’ and ‘Failure’ modeled as 1 and 0 respectively. The model tries to predict the probability of success or failure of the dependent variable based on the values of the independent variables inputted into the model. In our business case, the goal is to understand the factors that affect the ‘Success’ or ‘Failure’ of a given news article.
After discussions with our client at Skyscanner, we concluded that ‘Success’ of an article depends on how many unique page views it is able to get. If an article ‘Fails’ then we can conclude that it is not worth investing time and effort in creating it. Based on this business rationale, we model our analytical problem to encompass the idea of Success and Failure in Unique Page Views based on other attribute values such as No of Images, No of Links, Article Length etc.

**Dependent Variable:** Unique Page Views

The client has categorized our dependent variable to take two values:
**Success:** Over 400 Unique Views
**Failure:** Under 400 Unique Views
This implies that if an idea for an article cannot attain a minimum of 400 page views (based on the model), it may not be in the interest of the business to pursue it.

In order to understand the nature of our independent variables, we conduct some initial exploratory analysis in the form of summary statistic s on RapidMiner. The output can be seen in the diagram below:

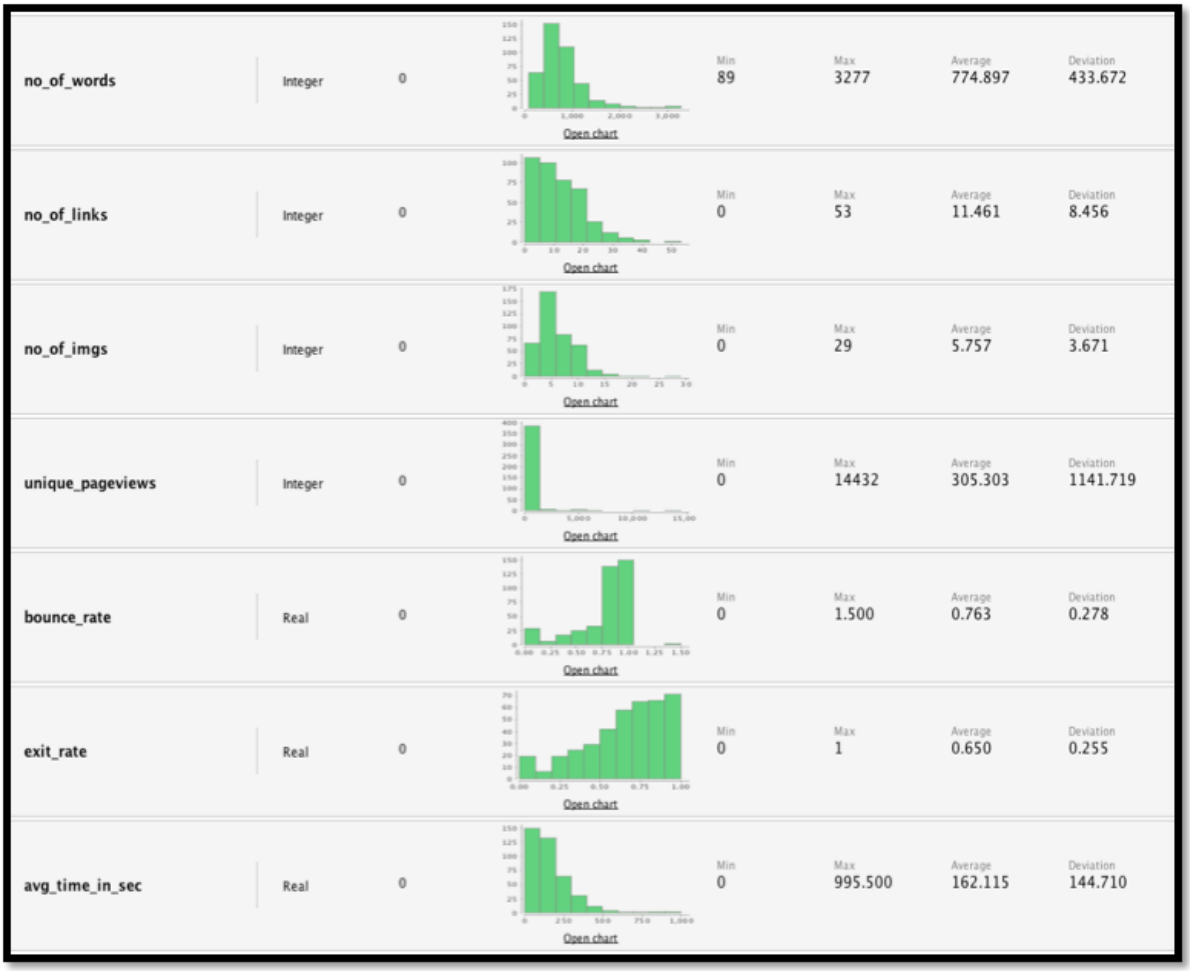



Figure 32: Summary Statistics from RapidMiner of numerical attributes used in the logistic regression

After looking at the distributions of the independent variables, we proceeded to remove outliers and also test for collinearity in order to pick the most appropriate attributes for the Logistic Regression.
In order to understand the factors that lead to high performance of this article, we used the following attributes in our model inputs as the independent variables:

**Predictors:**

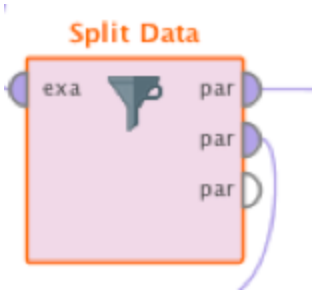
1. Average Time on Page
2. No of Images
3. No of Links
4. No of Words
5. Bounce Rate
6. Exit Rate
7. News Content Title Theme (Categorical- from Clustering)

**Removed Predictors:**

1. No. of Shares
2. Sessions
3. Pageviews
4. Organic Searches
5. Published Date

## 8.2.2. RapidMiner Logistic Regression

*Pre-processing Steps*




### 8.2.2.1. Splitting the data

One important consideration while building a predictive model such as a logistic regression model is to split your data into train and test data so that you can evaluate the model. RapidMiner provides a node to do this splitting based on the user input. The two main decisions to be made are: 1) Partition Ratios 2) Sampling Type.

#### 8.2.2.1.1. Partition Ratio

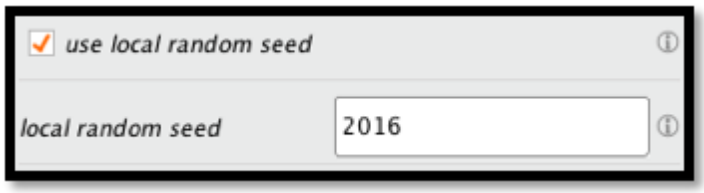
The user, after analyzing the size and nature of the dataset can decide the partition ratio. There are two competing concerns: with less training data, your parameter estimates have greater variance. With less testing data, your performance statistic will have greater variance. Broadly speaking you should be concerned with dividing data such that neither variance is too high, which is more to do with the absolute number of instances in each category rather than the percentage.
If you have a total of 100 instances, you're probably stuck with cross validation as no single split is going to give you satisfactory variance in your estimates. If you have 100,000 instances, it doesn't really matter whether you choose an 80:20 split or a 90:10 split (indeed you may choose to use less training data if your method is particularly computationally intensive). One good practice is to try a series of runs with different amounts of training data: randomly sample 20% of it, say, 10 times and observe performance on the validation data, then do the same with 40%, 60%, 80%. This should see both greater performance with more data, but also lower variance across the different random samples. This is the approach we have considered for our dataset.

#### 8.2.2.1.2. Sampling Type

RapidMiner offers three different options for sampling of the data when dividing into partitions. Each one has its own pros and cons depending on the nature of the dataset and goal of the analysis.

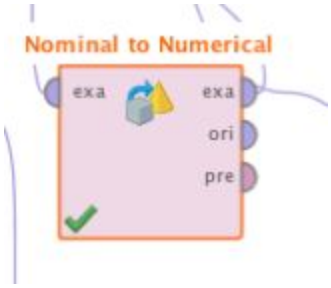
1. **Linear sampling** simply divides the data into partitions without changing the order of the examples i.e. subsets with consecutive examples are created.
2. **Shuffled sampling** builds random subsets of the data. Examples are chosen randomly for making subsets.
3. **Stratified sampling** builds random subsets and ensures that the class distribution in the subsets is the same as in the whole dataset. For example in the case of a binominal classification, Stratified sampling builds random subsets such that each subset contains roughly the same proportions of the two values of the class labels.

We proceeded to use stratified sampling with equal proportions of values of the target variable- Unique Page Views being put in both the test and train data. This ensured that both the test and train sets contain the same proportion of ‘Success’ and ‘Failure’ values.




#### 8.2.2.1.3. Use of Local Random Seed

RapidMiner also allows us to specify whether to use a local random seed. This indicates if a local random seed should be used for randomizing examples of a subset. Using the same value of local random seed will produce the same subsets. Changing the value of this parameter changes the way examples are randomized, thus subsets will have a different set of examples. This parameter is only available if Shuffled or Stratified sampling is selected. It is not available for the Linear sampling because it requires no randomization given that the examples are selected in sequence.



### 8.2.3. Converting Nominal Attributes to numerical

The Nominal to Numerical operator is used for changing the type of non-numeric attributes to a numeric type. In logistic regression, this is required to convert the categorical attributes into dummy values. In general for any attribute with ‘k’ values, there will be ‘k-1’ dummy variables. This operator provides many options for recoding for numerical variables into categorical. Below we can see all the options provided in RapidMiner:

1. **unique_integers**: If this option is selected, the values of nominal attributes can be seen as equally ranked, therefore the nominal attribute will simply be turned into a real valued attribute, the old values result in equidistant real values.
2. **dummy_coding**: If this option is selected, for all values of the nominal attribute, excluding the comparison group, a new attribute is created. The comparison group can be defined using the comparison groups’ parameter. In every example, the new attribute, which corresponds to the actual nominal value of that example, gets value 1 and all other new attributes get value 0. If the value of the nominal attribute of this example corresponds to the comparison group, all new attributes are set to 0. Note that the comparison group is an optional parameter with 'dummy coding'. If no comparison group is defined, in every example the new attribute, which corresponds to the actual nominal value of that example, gets value 1 and all other new attributes get value 0. In this case, there will be no example where all new attributes get value 0.
3. **effect_coding**: If this option is selected; for all values of the nominal attribute, excluding the comparison group, a new attribute is created. The comparison group can be defined using the comparison groups’ parameter. In every example, the new attribute, which corresponds to the actual nominal value of that example, gets value 1 and all other new attributes get value 0. If the value of the nominal attribute of this example corresponds to the comparison group, all new attributes are set to -1.

For the purposes of our analysis, we select the **‘Dummy Coding’** as it is a required input for the logistic regression. This models every variable into a column with a binary value, indicating the presence or absence of the given category.

### 8.2.4. RapidMiner Default Logistic Regression

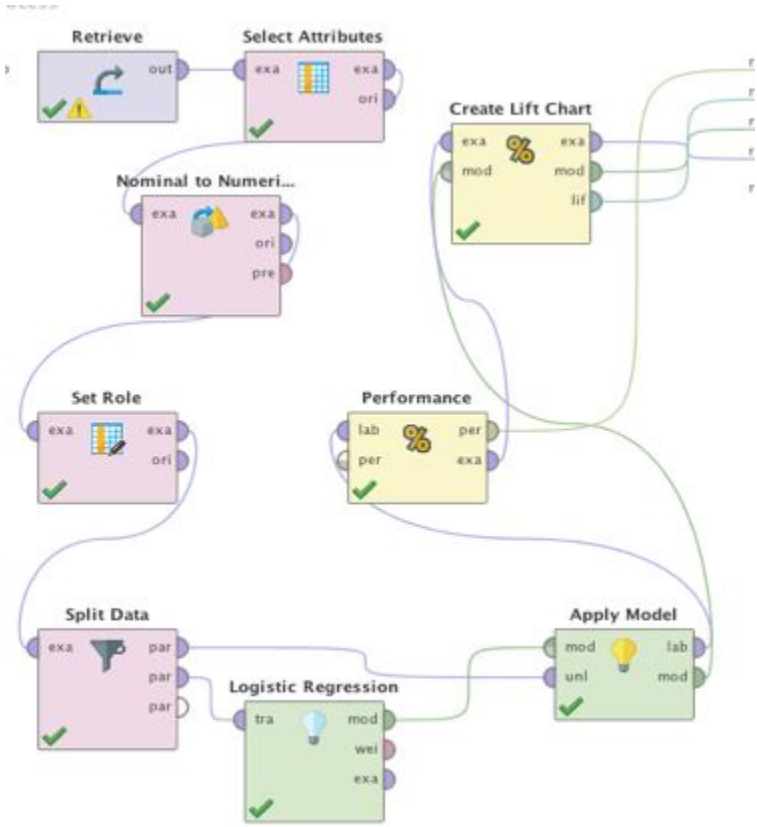



**Figure 33:** Logistic regression process in Rapidminer

In our first attempt at logistic regression, we have tried to run the logistic model with the attributes in their original form in order to explore the initial output. After this, we will iteratively refine the inputs of the model in order to come up with a model that can best help us solve our analytical problem of understanding the effect of each attribute on the dependent variable.

# Kernel Model

```
Total number of Support Vectors: 279
Bias (offset): -7.952

w[Assigned Title Label = 8] = -3.283
w[Assigned Title Label = 1] = -4.165
w[Assigned Title Label = 3] = 1.810
w[Assigned Title Label = 7] = -1.710
w[Assigned Title Label = 9] = 11.267
w[Assigned Title Label = 2] = 2.561
w[Assigned Title Label = 6] = -0.076
w[Assigned Title Label = 4] = 5.470
w[no_of_words] = 2.853
w[no_of_links] = 0.633
w[no_of_imgs] = 5.312
w[bounce_rate] = -5.457
w[exit_rate] = -3.009
w[avg_time_in_sec] = -6.970
```

**Figure 34:** Logistic regression kernel model (default) output

The default model is 82.5% accurate in predicting failure vs. success. The precision and recall can be seen in the table above.

**accuracy: 82.50%**

| | true Failure | true Success | class precision |
|---|---|---|---|
| pred. Failure | 96 | 13 | 88.07% |
| pred. Success | 8 | 3 | 27.27% |
| class recall | 92.31% | 18.75% | |

**Figure 35:** Logistic regression kernel model (default) precision and recall table

### 8.2.4.1. Problems with Default RapidMiner Operator

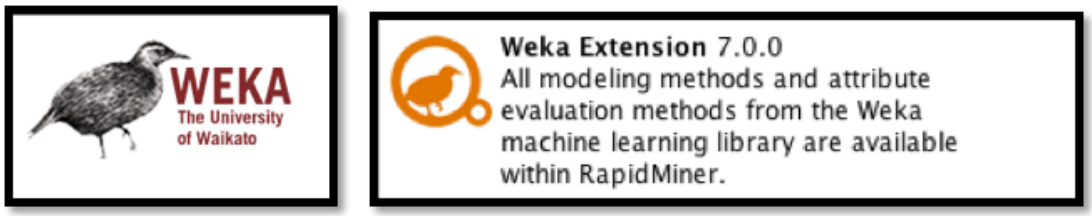
The default RapidMiner operator provides on the attribute weights for the logistic model as output. These weights, while useful in prediction, do not allow us to interpret or understand the relative effectiveness of the attributes in understanding the dependent variable. Here we have to remember that the goal of our analysis is not to create a model that will be used to as a recommendation system in the future but to understand how the attribute values such as No. of images, No of words etc. can be adjusted by the client in order to reach the ultimate goal of increasing the chance of success of the dependent variable. One way this can be done is by looking at the Odds ratio outputs for different values of the independent variables. As the default logistic operator in RapidMiner does not provide this, we have to install the Weka Package Add On for RapidMiner and Re-run our analysis.

### 8.2.4.2. Rationale behind using Weka for Logistic Regression

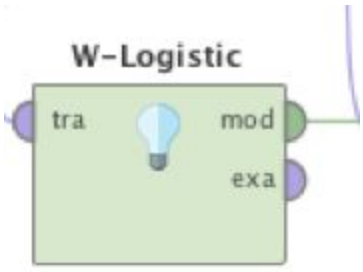
As the goal of our analysis is to better understand the effects of the different attributes on the dependent variables, we want to be able to interpret the coefficients and odds ratios of the regression. The default logistic regression operator in RapidMiner does not provide the odds ratios and probability outputs of regression and hence we must install the WEKA package add on to RapidMiner in order to run the traditional logistic regression that is similar to what is found in SAS Enterprise Miner.

### 8.2.4.3. Installing the WEKA Extension in RapidMiner

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well suited for developing new machine learning schemes. Since RapidMiner is java based, the Weka extension can easily integrate with it.




The Weka operator for Logistic Regression in RapidMiner is known as ‘W-Logistic’. This operator can be used to build a multinomial logistic regression model with a ridge estimator.



### 8.2.5. Running Logistic Regression on Original Attributes

Odds Ratios...

| Variable | Class Failure |
|---|---|
| no_of_words | 1.0004 |
| no_of_links | 0.9674 |
| no_of_imgs | 0.9316 |
| bounce_rate | 2.6499 |
| exit_rate | 0.1888 |
| avg_time_in_sec | 0.9972 |
| Assigned Title Label=8 | 0.81 |
| Assigned Title Label=1 | 3.7109 |
| Assigned Title Label=3 | 0.1168 |
| Assigned Title Label=7 | 0.8071 |
| Assigned Title Label=9 | 0.2655 |
| Assigned Title Label=2 | 0.861 |
| Assigned Title Label=6 | 1519016.0372 |
| Assigned Title Label=4 | 0.5514 |

Figure 36: Logistic regression model with weka extension

accuracy: 84.17%

| | true Failure | true Success | class precision |
|---|---|---|---|
| pred. Failure | 100 | 15 | 86.96% |
| pred. Success | 4 | 1 | 20.00% |
| class recall | 96.15% | 6.25% | |

Figure 37: Logistic regression model with weka extension recall and precision table

#### 8.2.5.1. Issue with using original attribute forms

While the model results above do solve the issue of having attribute value specific coefficients in the form of Odds Ratio, there is still a problem with our numerical attributes. The numerical values Odds Ratios do not provide us with sufficient insights that can be interpreted due to their incremental nature. In making recommendations based on article length, it would not be meaningful to say that articles with 1 word more would increase success rates by 1 percent. With no upper or lower bound, the client would be inclined to create lengthy articles. Rather, recommending that articles of lengths ranging from 900-1200 fare better than 2000-2500 words would be more meaningful. In order to do this we need to divide the numerical variable into ranges so that we can compare the performance of each range in order to provide a recommendation. In order to do this, we consider the process of discretization.

## 8.2.6. Discretization of continuous variables:

There are many different methods for binning, each one has its own advantages and disadvantages depending on the nature of the problem and the distribution of the attributes in the dataset:

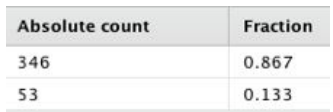
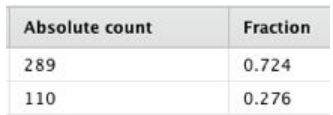
**Techniques:**

1. Discretize by Binning
2. Discretize by User Specification
3. Discretize by Size
4. Discretize by Entropy

Out of these, we have considered 2 main binning methods:

### 8.2.6.1. Discretize by Entropy

This RapidMiner operator converts the selected numerical attributes into nominal attributes. The boundaries of the bins are chosen so that the entropy is minimized based on the dependant variable

| **Attribute** | **Range** | **Failure VS Success Count** |
|---|---|---|
| **unique_pageviews** | Failure (346), Success (53) |   |
| **no_of_imgs** | range2 [4 - ∞] (289) range1 [-∞ - 4] (110) |   |
| **bounce_rate** | range1 [-∞ - 0.970] (330) range2 [0.970 - ∞] (69) |   |
| **avg_time_in_sec** | range2 [49 - ∞] (318), range1 [-∞ - 49] (81) |   |

This operator divides only 3 out of the five operators into sets of 2 ranges. In RapidMiner, we can also see the relative distribution of Success and Failure in each Range.

#### 8.2.6.2. Discretize by Frequency

In order to get useful insights for our business problem, we need to discretize the data in a meaningful way so that the results can be interpreted in the form of an action that will help solve the business problem. After discussions with the client, we have agreed that having 3 bins of equal size will help drive this decision. Comparing the probability of success in each of the 3 bins for each attribute will help understand the best range selection for a given independent variable. For example, if a the bin with high number of images is most successful out of 3 bins, then having a high number of images will produce better results for the business. In order to apply this logic, we must use the 'Discretize by Frequency' operator in RapidMiner to discretize the independent variable such that each bin has an equal number of occurrences (Sarma, n.d.).

RapidMiner provides a Discretize operator that allows us to convert numerical data into categorical based on a certain measure. As we want to have bins such that each bin has roughly the same number of values, we use the 'Discretize by Frequency' option. The boundaries of the bins are chosen so that the entropy is minimized in the induced partitions.

The table below shows the model accuracy for the different forms of discretization when with the Logistic Regression operator in RapidMiner:

| **Method** | **Bins** | **Model Accuracy** | **Squared Error** |
|---|---|---|---|
| Entropy | Variable | 82.5% | 0.130 +/- 0.340 |
| Size | 2 | 83.67% | 0.151 +/- 0.402 |
| Size | 3,4 | 83.67% | 0.128 +/- 0.241 |
| Frequency | 2 | 83% | 0.130 +/- 0.340 |
| **Frequency** | **3** | **85.8%** | **: 0.124 +/- 0.258** |
| Frequency | 4 | 84.27% | 0.124 +/- 0.258 |
| Frequency | 5 | 84.27% | 0.124 +/- 0.258 |

Given the table above, we have considered Discretization by Frequency as it provides us with the highest model accuracy and the lowest squared error. It is also a good fit for our overall business problem.
We have considered 3 partitions per numerical value with the rationale that each partition should have a minimum of over 100 data points. After running the discretization operator we get our output in the form of ranges. We have then plotted the Average Unique Page Views of each partition for a given independent variable. These values have been visualized in tableau and can be found in the image below.

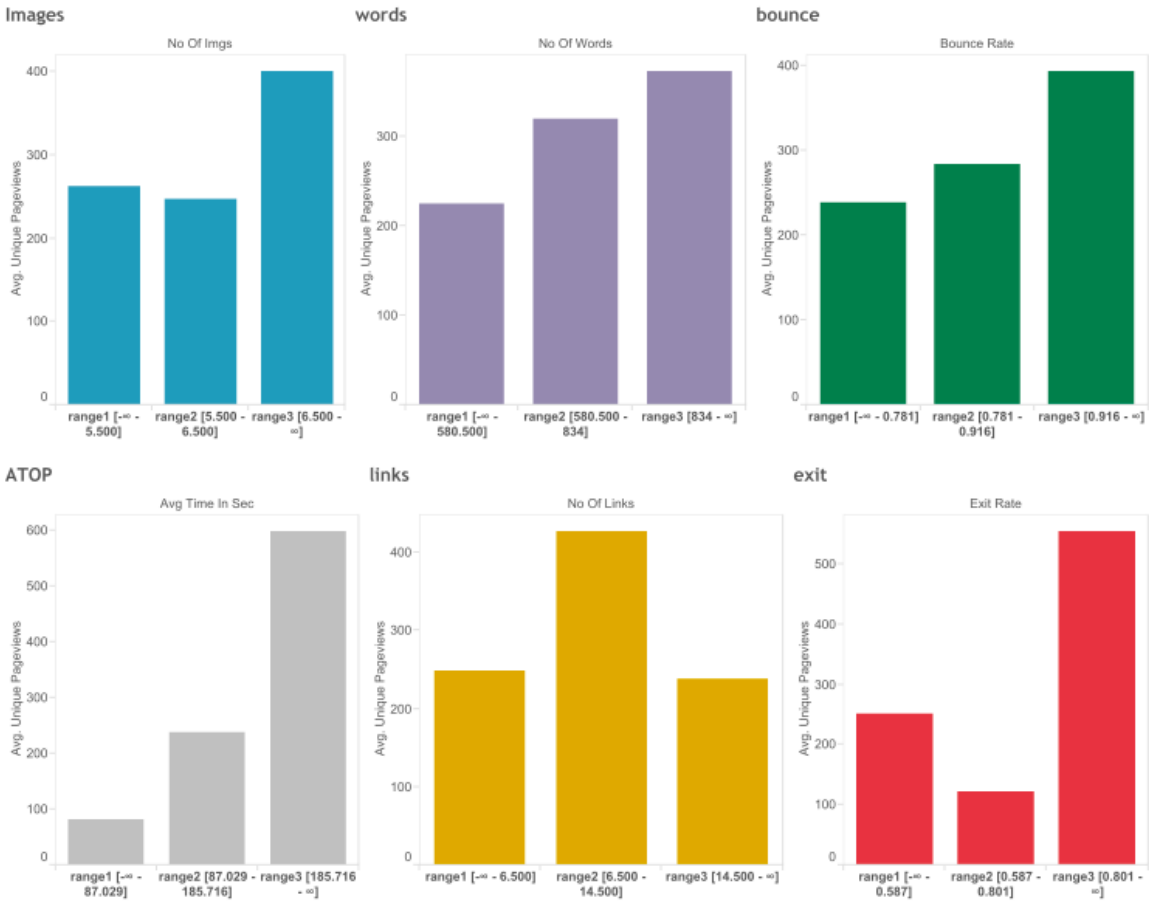



Figure 38: Distribution of UPVs across of the different bins formed by discretization

## 8.2.7. Running the W-Logistic Model on Binned Data

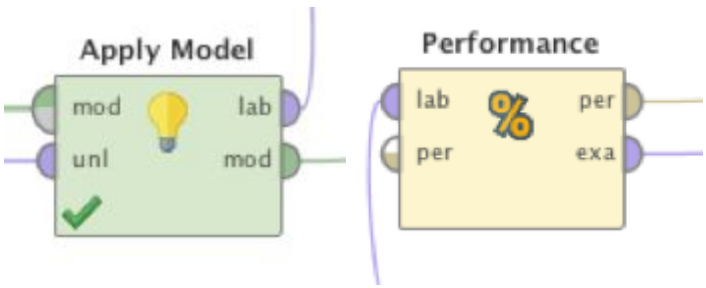
**Model Output**

Odds Ratios...

| Variable | Class Failure |
|---|---|
| no_of_words = range3 [834 - ∞] | 1.1843 |
| no_of_words = range1 [-∞ - 580.500] | 0.982 |
| no_of_words = range2 [580.500 - 834] | 0.8507 |
| no_of_links = range3 [14.500 - ∞] | 0.6925 |
| no_of_links = range2 [6.500 - 14.500] | 0.6961 |
| no_of_links = range1 [-∞ - 6.500] | 2.0277 |
| no_of_imgs = range3 [6.500 - ∞] | 0.8191 |
| no_of_imgs = range1 [-∞ - 5.500] | 0.9912 |
| no_of_imgs = range2 [5.500 - 6.500] | 1.964 |
| bounce_rate = range3 [0.916 - ∞] | 1.257 |
| bounce_rate = range2 [0.781 - 0.916] | 1.0436 |
| bounce_rate = range1 [-∞ - 0.781] | 0.7614 |
| exit_rate = range1 [-∞ - 0.587] | 1.2638 |
| exit_rate = range2 [0.587 - 0.801] | 1.2349 |
| exit_rate = range3 [0.801 - ∞] | 0.6341 |
| avg_time_in_sec = range1 [-∞ - 87.029] | 2.5599 |
| avg_time_in_sec = range3 [185.716 - ∞] | 0.4475 |
| avg_time_in_sec = range2 [87.029 - 185.716] | 0.8926 |
| Assigned Title Label=8 | 0.7781 |
| Assigned Title Label=1 | 2.5225 |
| Assigned Title Label=3 | 0.1176 |
| Assigned Title Label=7 | 0.85 |
| Assigned Title Label=9 | 0.3614 |
| Assigned Title Label=2 | 0.9578 |
| Assigned Title Label=6 | 4627474.6643 |
| Assigned Title Label=4 | 0.5829 |

**Figure 39:** Binned Data Odds Ratio Output

#### 8.2.7.1. Model Evaluation




The ‘Apply Model’ and ‘Performance’ operators in RapidMiner were used in order to evaluate the logistic model. For the performance, we looked at both the Model Accuracy and the model Squared Error.

### 8.2.7.2. Model Accuracy Table

| **Accuracy: 85.8%** | **True Failure** | **True Success** | **Class Precision** |
|---|---|---|---|
| **Pred. Failure** | 100 | 15 | **86.96%** |
| **Pred. Success** | 2 | 3 | **60%** |
| **Class Recall** | **98.04%** | **20%** | |

### 8.2.7.3. Model Interpretation

| **Attribute** | **Value** | **Label** | **Odds Ratio** | **P(Failure)** | **P(Success)** |
|---|---|---|---|---|---|
| no_of_words | range3 [834 - ∞] | High | 1.1843 | 54.22% | 45.78% |
| no_of_words | range1 [-∞ - 580.500] | Low | 0.982 | 49.55% | 50.45% |
| no_of_words | range2 [580.500 - 834] | Medium | 0.8507 | 45.97% | 54.03% |
| no_of_links | range3 [14.500 - ∞] | High | 0.6925 | 40.92% | 59.08% |
| no_of_links | range2 [6.500 - 14.500] | Medium | 0.6961 | 41.04% | 58.96% |
| no_of_links | range1 [-∞ - 6.500] | Low | 2.0277 | 66.97% | 33.03% |
| no_of_imgs | range3 [6.500 - ∞] | High | 0.8191 | 45.03% | 54.97% |
| no_of_imgs | range1 [-∞ - 5.500] | Low | 0.9912 | 49.78% | 50.22% |
| no_of_imgs | range2 [5.500 - 6.500] | Medium | 1.964 | 66.26% | 33.74% |
| bounce_rate | range3 [0.916 - ∞] | High | 1.257 | 55.69% | 44.31% |
| bounce_rate | range2 [0.781 - 0.916] | Medium | 1.0436 | 51.07% | 48.93% |
| bounce_rate | range1 [-∞ - 0.781] | Low | 0.7614 | 43.23% | 56.77% |
| exit_rate | range1 [-∞ - 0.587] | Low | 1.2638 | 55.83% | 44.17% |
| exit_rate | range2 [0.587 - 0.801] | Medium | 1.2349 | 55.26% | 44.74% |
| exit_rate | range3 [0.801 - ∞] | High | 0.6341 | 38.80% | 61.20% |
| avg_time_in_sec | range1 [-∞ - 87.029] | Low | 2.5599 | 71.91% | 28.09% |

| | | | | | |
|---|---|---|---|---|---|
| avg_time_in_sec | range3 [185.716 - ∞] | High | 0.4475 | 30.92% | 69.08% |
| avg_time_in_sec | range2 [87.029 - 185.716] | Medium | 0.8926 | 47.16% | 52.84% |
| Assigned Title Label | 8 | Activity/Topic | 0.7781 | 43.76% | 56.24% |
| Assigned Title Label | 1 | Inspirational | 2.5225 | 71.61% | 28.39% |
| Assigned Title Label | 3 | Domestic/ Local | 0.1176 | 10.52% | 89.48% |
| Assigned Title Label | 7 | Skyscanner Product | 0.85 | 45.95% | 54.05% |
| Assigned Title Label | 9 | Food | 0.3614 | 26.55% | 73.45% |
| Assigned Title Label | 2 | City Guides | 0.9578 | 48.92% | 51.08% |
| Assigned Title Label | 6 | Practical/ Tips | 4627474.664 | 100.00% | 0.00% |
| Assigned Title Label | 4 | Trending | 0.5829 | 36.82% | 63.18% |

The table above shows how certain values of each of the explanatory variables lead to a better chance of the 'successful' article as compared to others. These values can be considered in the planning stage by Skyscanner to develop more high performing articles. The table below shows which values of each of the explanatory variables will lead to more successful content articles.

### 8.2.7.4. High Performing Attribute Values

| Attribute | Value | Label | Interpretation |
|---|---|---|---|
| no_of_words | range2 [580.500 - 834] | Medium | Articles with about 580-830 words tend to do better |
| no_of_links | range3 [14.500 - ∞] | High | The more links in an article, the more likely it is to perform better |
| no_of_links | range2 [6.500 - 14.500] | Medium | |
| no_of_imgs | range3 [6.500 - ∞] | High | More image lead to higher viewership, with a minimum of 7 per article |

| | | | |
|---|---|---|---|
| bounce_rate | range1 [-∞ - 0.781] | Low | A low bounce rate is good for page views |
| exit_rate | range3 [0.801 - ∞] | High | A high exit rate boosts readership |
| avg_time_in_sec | range3 [185.716 - ∞] | High | The more time each user spends on an article, the more likely it is to be successful |
| Assigned Title Label | 8 | Activity/Topic | Articles covering topics like Activities, Local topics, Food and Trends perform better than other topics. |
| Assigned Title Label | 3 | Domestic/ Local | |
| Assigned Title Label | 9 | Food | |
| Assigned Title Label | 4 | Trending | |

#### 8.2.7.5. Low Performing Attribute Values

| Attribute | Value | Label | Interpretation |
|---|---|---|---|
| Assigned Title Label | 1 | Inspirational | Inspiration related articles do not perform as well as articles covering other topics |
| Assigned Title Label | 6 | Practical/ Tips | Practical/Tips related articles do not perform well compared to others |

### 8.2.8. Recommendations

Our findings can be summarized into the following points that would serve as recommendations for Skyscanner's content site:

- Skyscanner's content pages should have articles of moderate level length ie 500-800 words
- There should be more links (10) and images (minimum of 7) in every article.
- There should be an effort to create more content covering Topics such as Activities, Domestic, Food and Trending.
- Skyscanner should also reduce content from topics such a Tips and Inspirational.

### 8.2.9. Avenues for Further Exploration

Another important consideration for this analysis that can be explored further, is the use of a Decision Tree instead of the Logistic Model in order to better explain the relationship between the values of the independent variables in the explanatory model. This model has been run within RapidMiner and the process file, inputs and outputs can be found within our analytical data cube. While this model provides an insightful output in terms of performance, it requires multiple iterations for refinement.  Due to time and resource constraints, multiple iterations of this process could not be completed within this iteration of the project but can be further explored in future exploratory analysis efforts for Skyscanner.

## 8.3. Data Visualization

### 8.3.1. Source/Medium Dashboard

#### 8.3.1.1. Motivation

From the discussions with our sponsor, Skyscanner Content Team would like to perform exploratory analysis on the effectiveness of paid and unpaid articles, across multiple platforms that Skyscanner promotes the content. Thus we have created an interactive dashboard using Tableau software to facilitate this exploration process, based on the data that we have assimilated.

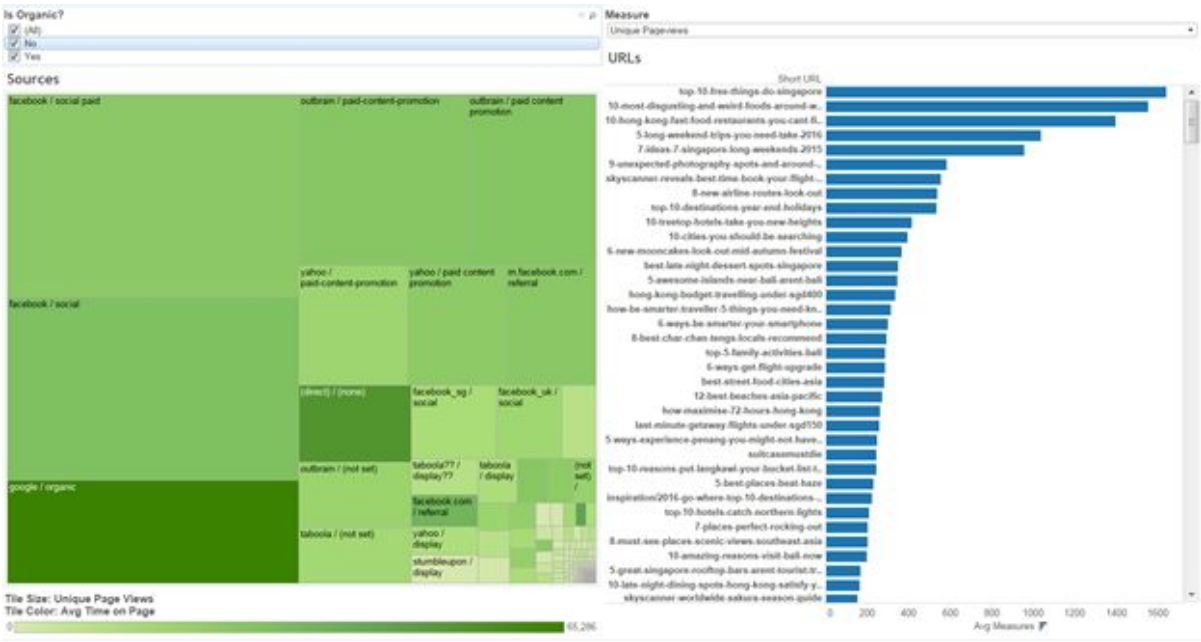



Figure 40: *Interactive Dashboard for Exploratory Data Analysis*

#### 8.3.1.2. Treemap for Visualization of UPV and ATOP

Since the attributes of interest for Skyscanner are unique page views (UPV) and average time on page (ATOP), and since Skyscanner also wants to understand which platforms or

mediums bring the most satisfactory result, we decided to represent the information using a treemap. Each tile within the map is assigned a platform/medium. The tile's size shows the relative UPV, with bigger tile meaning that particular source has higher UPV. Similarly, the tile color shows us the relative ATOP, with darker tile having higher ATOP, and vice versa.

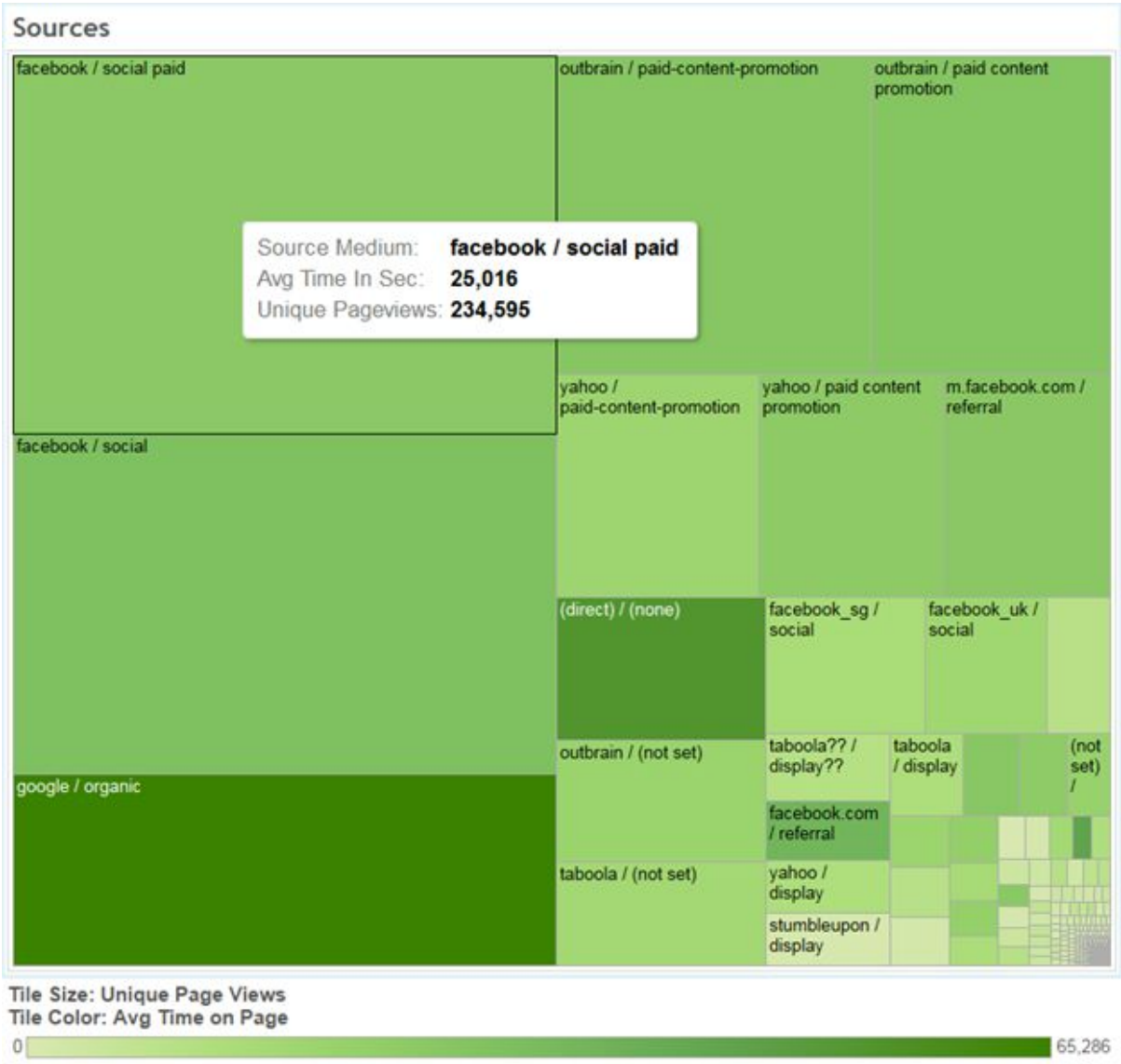



*Figure 41:* Treemap showing ATOP and UPV across multiple platforms

### 8.3.1.3. Filter for Paid and Non-paid Content

Skyscanner also wants to find out more about the organic growth of their articles (non-paid), and in term of paid content, possibly which platforms yield better UPV or ATOP, in order to focus their resources. Thus we also created a filter to separate data between organic and paid articles.

*Figure 42:* Filter for Organic and Non Organic Content

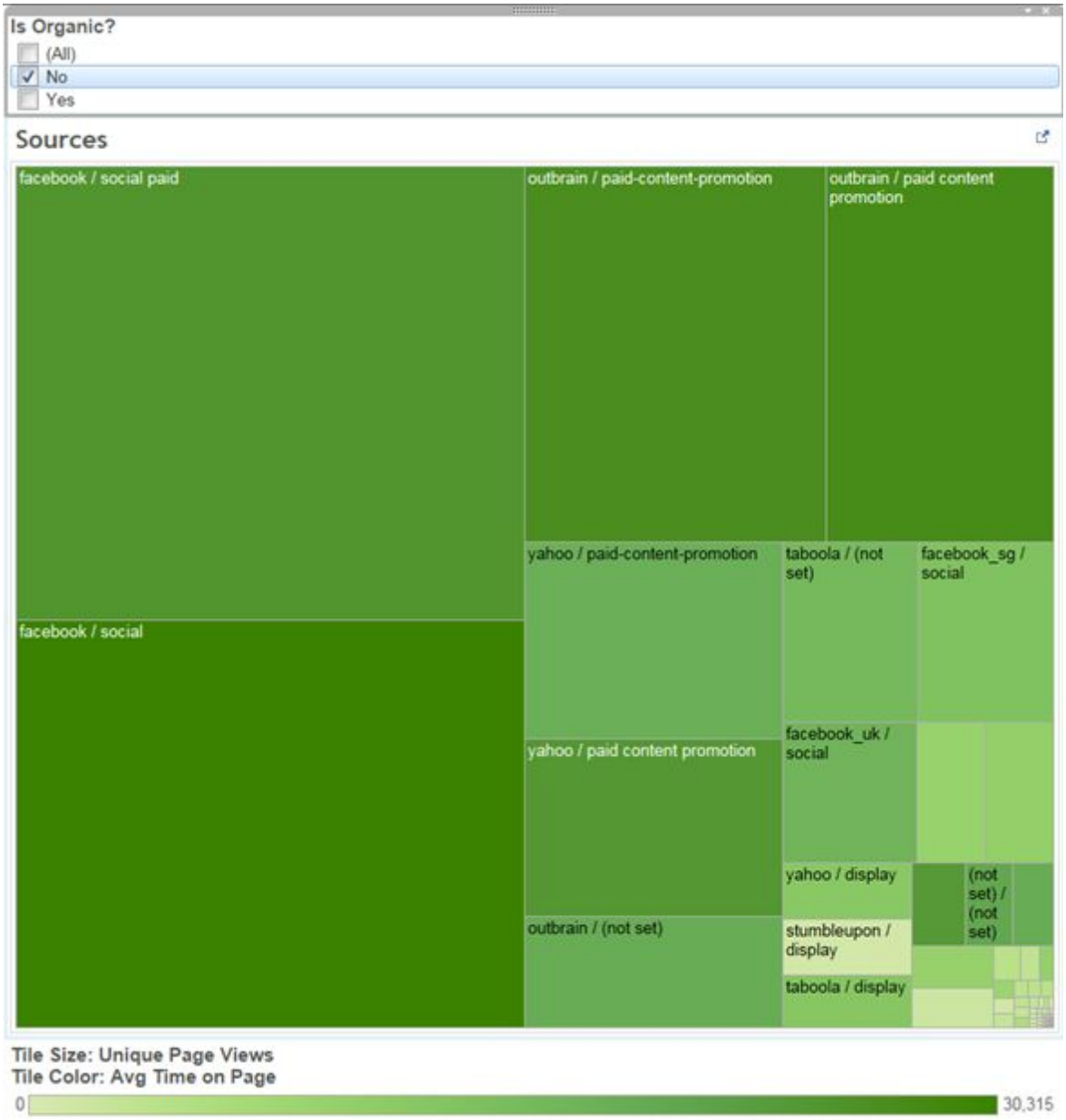



*Figure 43:* Treemap showing ATOP and UPV across multiple platforms, filtered for Non-Organic Source

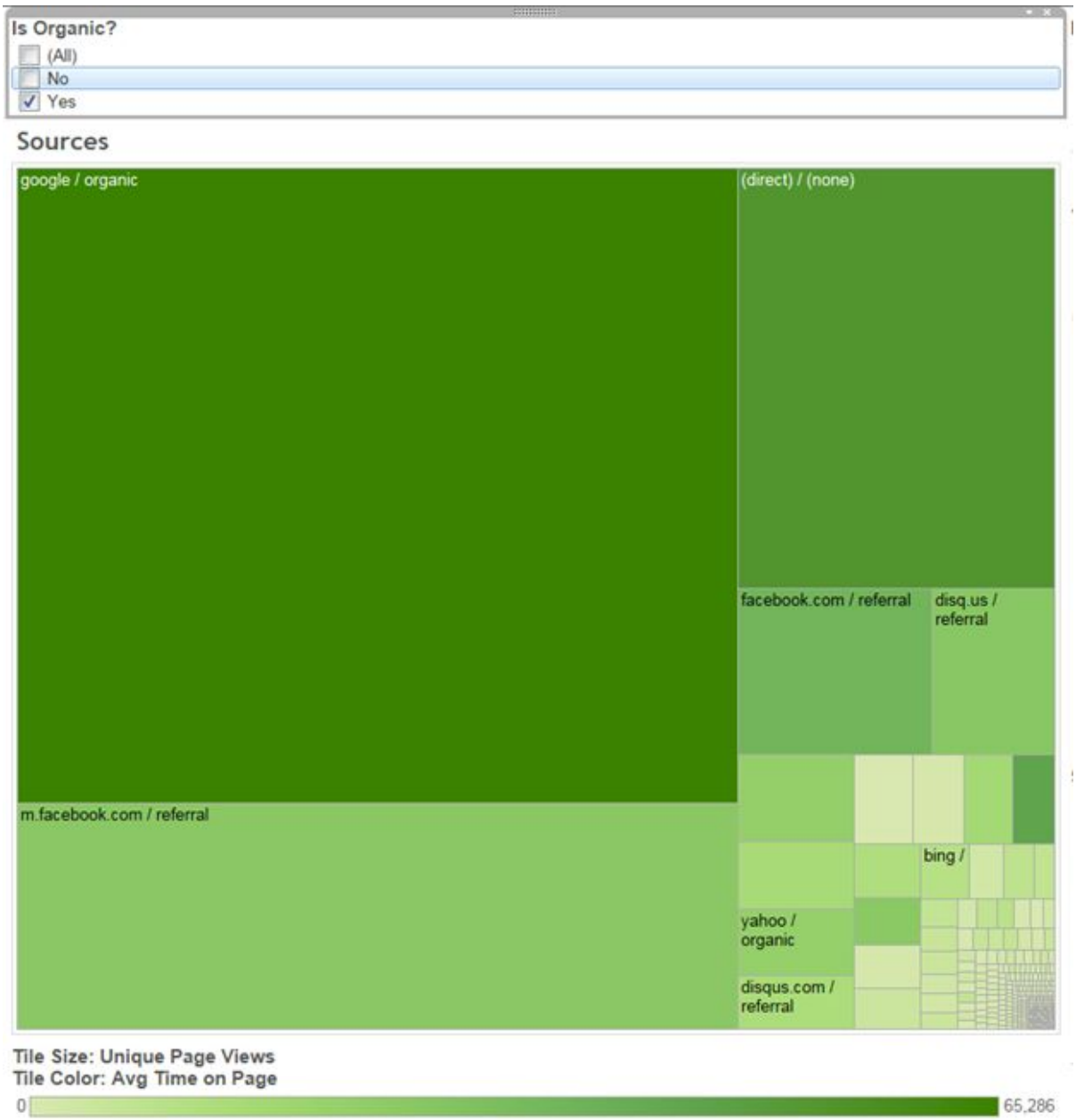



*Figure 44:* Treemap showing ATOP and UPV across multiple platforms, filtered for Organic Source

#### 8.3.1.4. Switching measures

In order to provide a better drill down view on performance of each article within a particular platform, the right-hand side of the dashboard can show the readings of the articles, based on which platform is selected from the treemap. The result can be sorted to view, for example, top 10 articles with highest UPV that were searched on Google, or top 10 articles that has the most engagement from Facebook. User can also switch between different measures (UPV, ATOP, bounce rate, exit rate, etc.) to evaluate the performance or explore various perspectives.

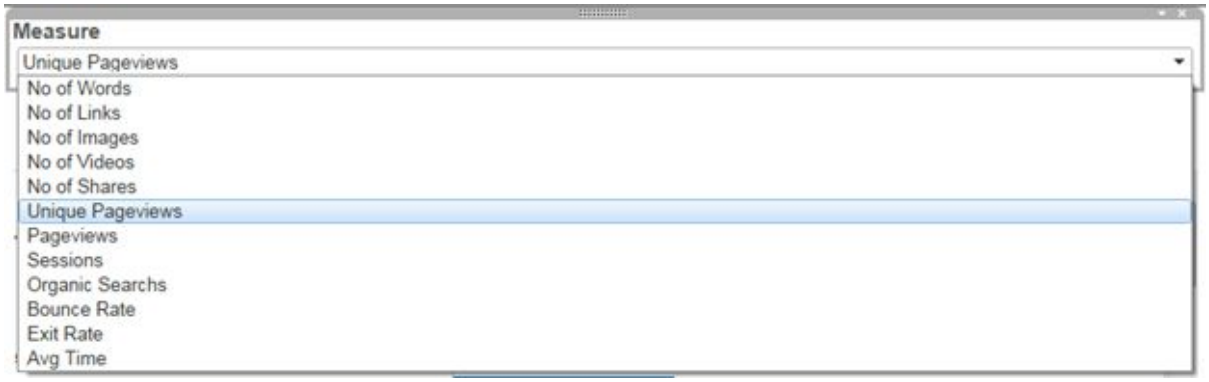



Figure 45: *Different measures can be selected to provide different perspective on the readings*

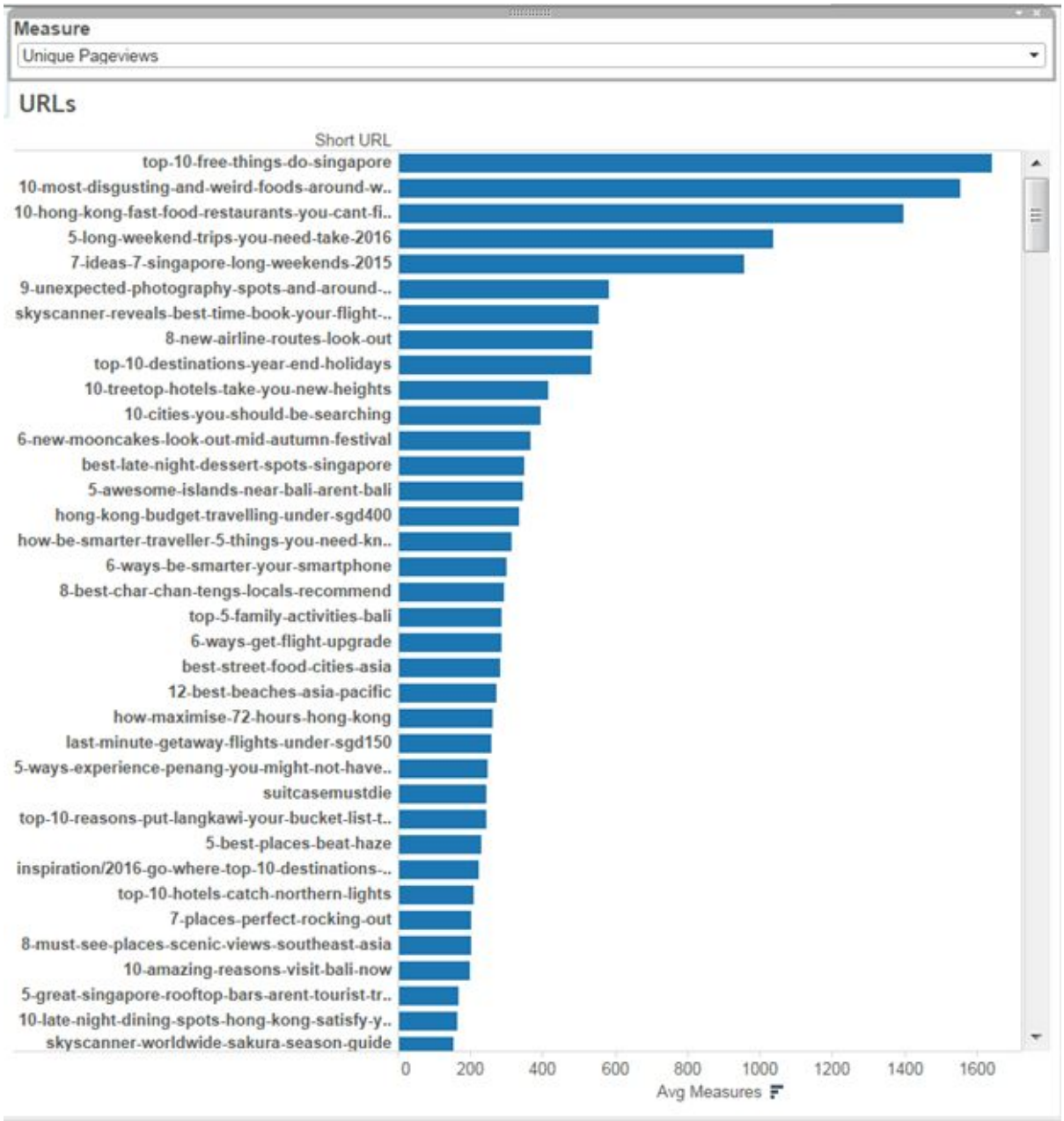


*Figure 46:* Articles with high UPV, filtered on Organic Source

## 8.3.2. Content Theme Visualization

### 8.3.2.1. Motivation

In conjunction with discovering which cluster an article belongs, it is interesting to investigate how each cluster perform against each other across multiple measures, so that Skyscanner can improve their planning process and allocate sufficient resources to the most appropriate content themes.

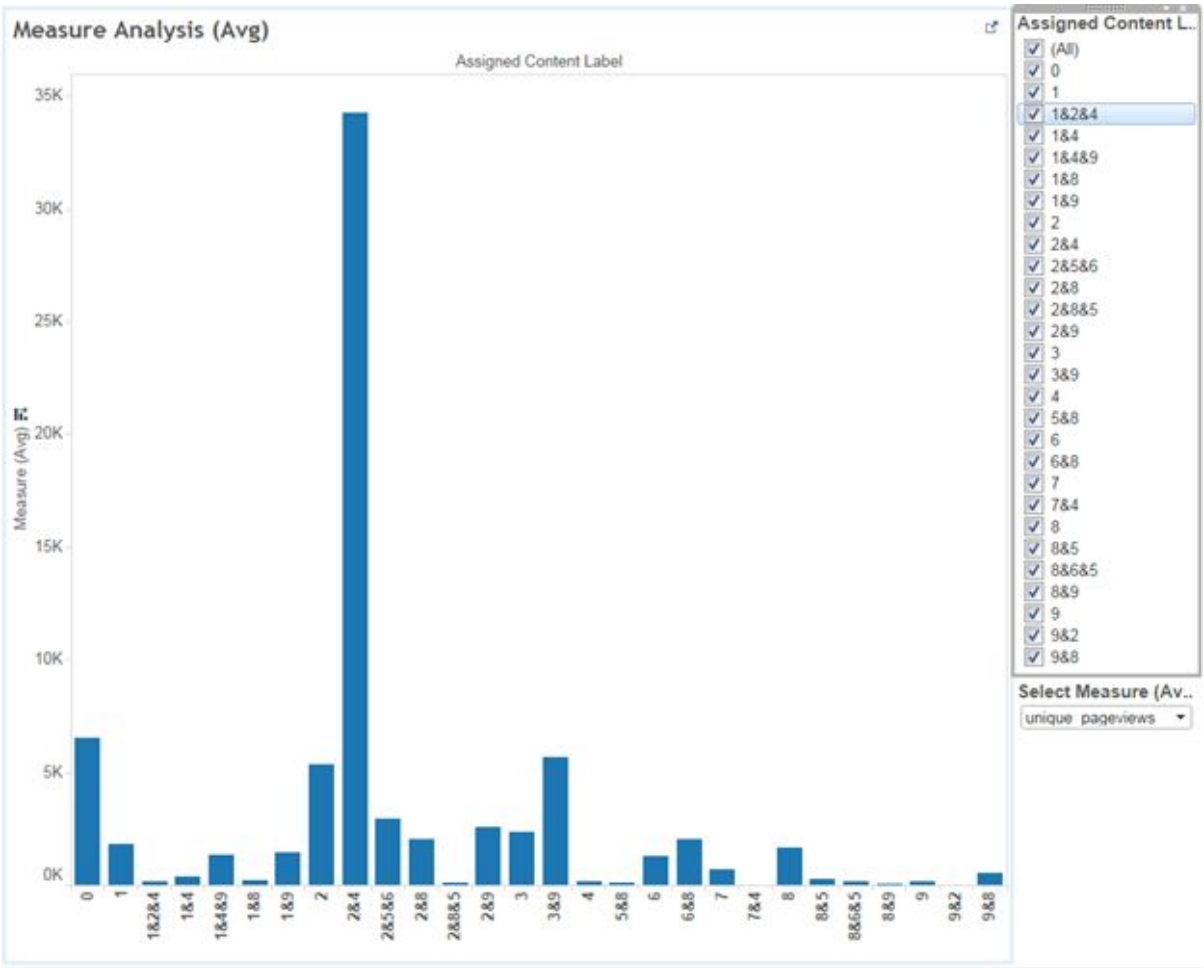


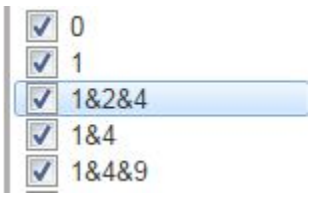

Figure 47: Dashboard for CT Clustering

### 8.3.2.2. Filter for Content Theme Classifications

| Content Theme Descriptior | Code |
|---|---|
| unknown | 0 |
| Inspirational- eg Top festiv | 1 |
| City Guides | 2 |
| Domestic/Local- Singapore | 3 |
| Trending- Game of thrones | 4 |
| Deals - Prices | 5 |
| Practical/Tips | 6 |
| Product- Skyscanner featur | 7 |
| Activity/topic discussion | 8 |
| Food | 9 |

Figure 48: Mapping of CT and Code

As mentioned above, we have identified 9 distinguish content themes, and one unclassifiable theme. However, there are some articles which could not be classified under a specific theme, and thus were assigned two, or even three themes.




Through the filter on the right hand side, Skyscanner team can easily select and compare a certain measure between chosen CT classifications, which may be hard to realize when selecting all classifications.

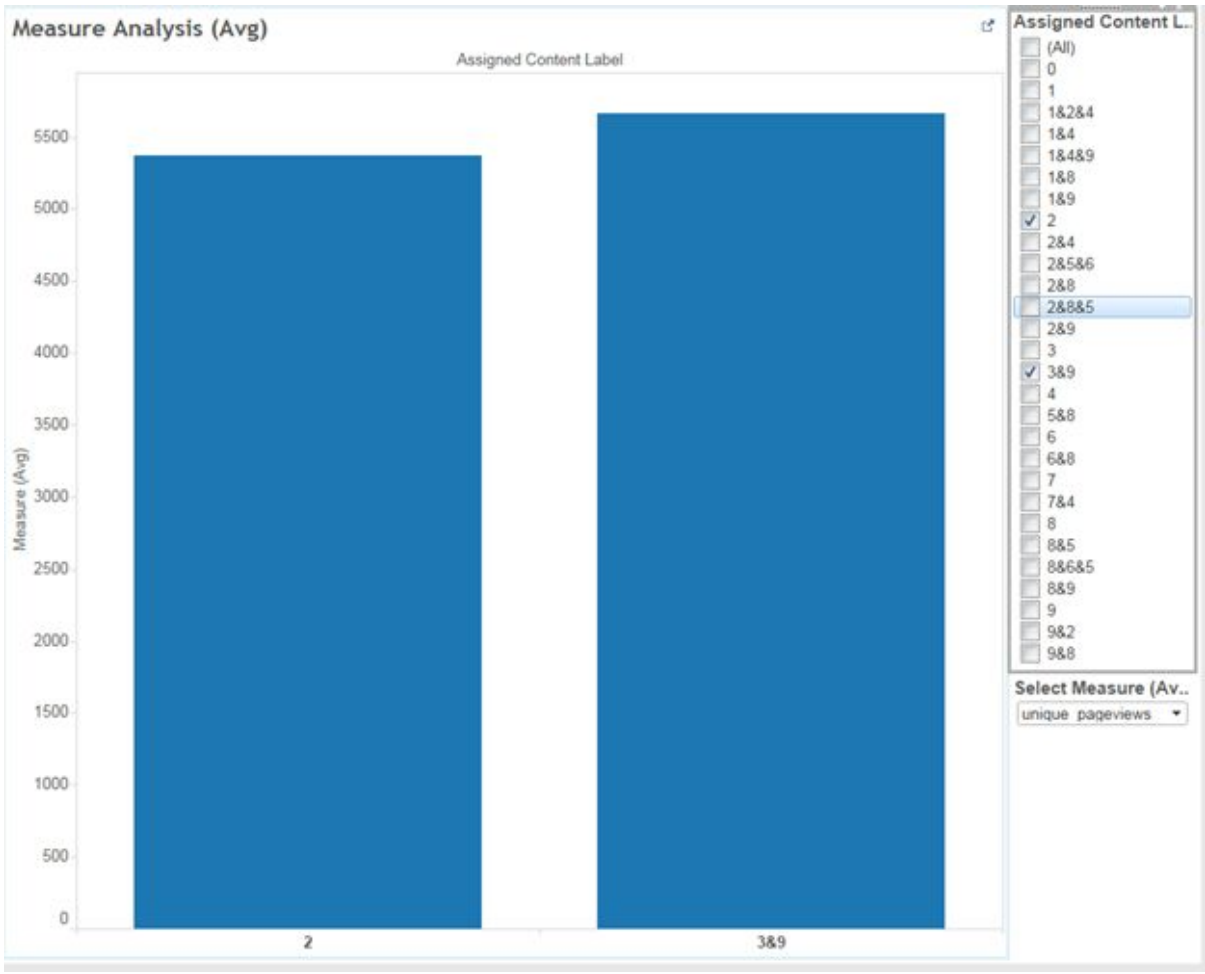



Figure 49: Comparison of UPV between articles with assigned CT 2 and articles with assigned CT 3&9

#### 8.3.2.3. Switching Measures

Our group also include the capability to switch between different measures, such as UPV, ATOP, bounce rate, and exit rate, so that Skyscanner team can dive deeper into the characteristics of each cluster. Certain clusters may drive high UPV, but low ATOP, and vice versa.

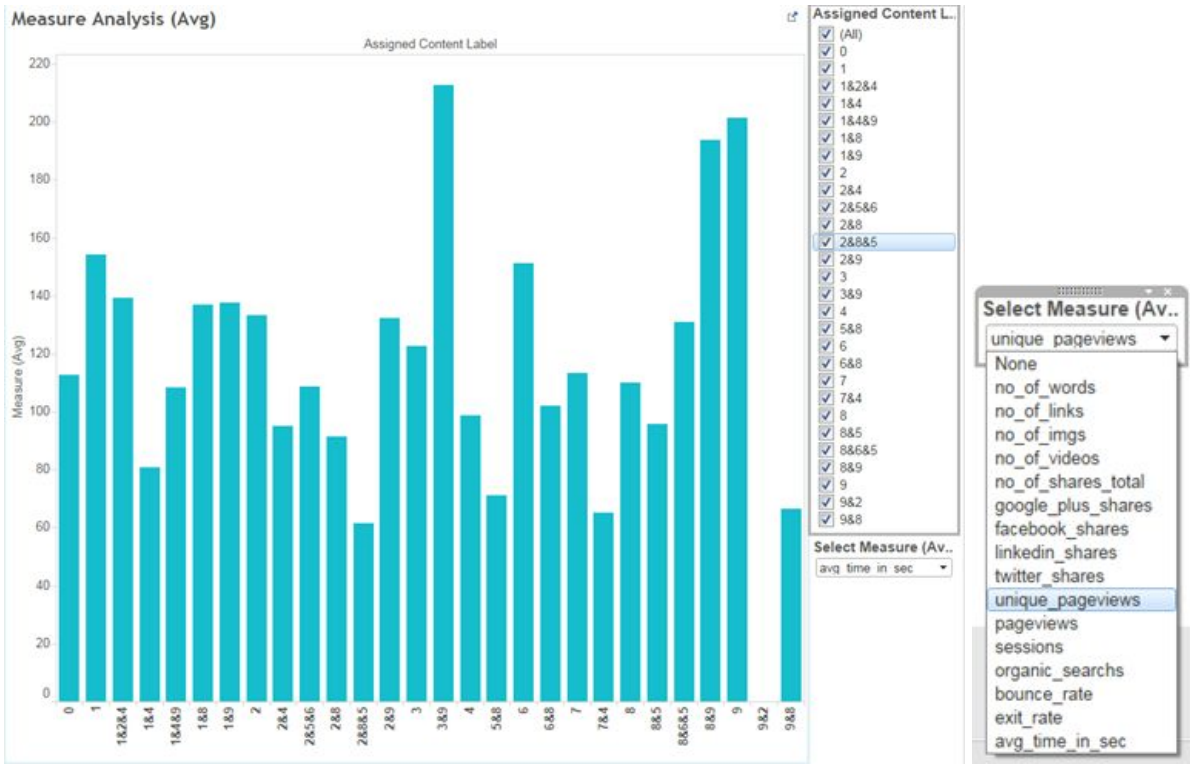



*Figure 50:* ATOP performance between different clusters

# 9. Conclusion

## 9.1. Client management

To start off, finding a client and a project for this course can be a challenging task. From the time we formally registered our team, till the start of the term, we spent most of our time contacting everyone in our professional network in order to source an appropriate analytics project. It is always a challenge for companies as you have to convince them about your value proposition and reassure them about issues such as data protection and non-disclosure. After much back and forth, we managed to get Ms Antoinette Tan at Skyscanner Singapore to be our sponsor and provide us with a business problem as well as access to Skyscanners internal data.

Secondly, we had to understand the client's business problem. During the course of the project, it was important for us to meet the client as often as possible, ideally once in two weeks in order to make sure our analysis was in line with their business goals. Meetings and discussions with the client also helped us to draw out some of the ‘tacit knowledge’ that we needed to better understand the nature of our dataset. It is often difficult for an analyst who has no knowledge of the internal workings of a company or an industry to make assumptions or interpretations from a dataset without this tacit knowledge. Getting out clients constant feedback during the analysis process helped us maximize the relevance and impact of our project deliverables.

Lastly, we had to manage expectations. Your client normally does not care about the amount of effort it takes to do a certain analysis, they care more about your final product and what it can do for them. In order to effectively manage our resources, we had to be clear to the client about what types of analysis or deliverables could be within the scope of the project. At many times we had to turn down certain additional deliverables due to infeasibility. There is a fine line between managing expectations and actually upsetting the client which needs to be managed very well for any successful project.

## 9.2. RapidMiner as a Tool for Data Analytics

With its drag and drop user interface, RapidMiner is a fuss and code free toolkit for data analytics. For the purposes of our project, it was able to adequately cover the process of exploratory analysis, to cleaning, to analytics modelling and model evaluation. In addition, there is a wide selection of extensions which companies and freelance developers have created and shared on the marketplace (a sharing platform for the RapidMiner community). This project alone has already made use of 3 such extensions[4].

This being the team's first encounter with RapidMiner, the learning curve was also found to be highly manageable. RapidMiner has a library of sample processes demonstrating the workflow of key data analytics processes. These samples ranges from the basic data preprocessing, to the advanced machine learning modelling. RapidMiner gets the user acquainted with the components involved in key workflows fairly seamlessly. Further enhancing the learning experience are the the in-app and help documentation that comes with each and every component. They were found to be concise and not superficial.

Nevertheless, one gripe we found in RapidMiner also stems from this strength in its documentation. While it does a good job most of the time, there were specific parameter optimization configuration under more advanced machine learning components that RapidMiner would have done well to include more detail.

Nevertheless, our team is of the opinion that RapidMiner is a user friendly data mining toolkit that elegantly abstracts key functionality from the complications that come with code, allowing its user to focus on data analytics.

## 9.3. Process Documentation

During the course of performing analysis, sometimes we are too engrossed in deriving insights and interesting result, that we forget to record our steps and progress. This can lead to several problems later on when we need to remember the rationale, retrace the steps, and re-perform the analysis. Documenting our analytic process and our discussions for future reference has served us well throughout the time doing this project.

[4] 3 mentioned extensions include Weka, Web Mining and Text Processing

While carrying out analytics, there are certain times when decision for choosing a specific parameter or attribute in building a model was not explicitly stated, which left some members confused. One instance was the problem of the default Logistic Regression engine in RapidMiner not providing odds ratio and probability, requiring us to install Weka extension. Without the note of the member that carried out the analysis of Logistic Regression, some of us would have be left wondering why we needed extra tools which may introduce more complexities. Another instance when documenting helped us was the decision of choosing number 71 as K for Content Theme Clustering. Without all the researches backing up the choice of good K value, other members would question if that number was an arbitrage choice, and waste more time doing trials-and-errors to confirm the goodness of K.

Documenting discussions also provided us helpful guidelines for actions needed and project management. Project requirements, model assumptions were all gathered and recorded from our meetings with Skyscanner. Possible insights and modification to project process were extracted from the minutes of supervisor meetings. The meeting minutes also capture action items, assigned to specific members, allowing us to check up on the progress and stay on schedule.

Documentation is important to manage the project, and also to make sure everyone's clear on the assumptions and goals while performing analysis. It has served us well while doing this Practicum project.

---

# 10. References


Bower, K. M. On The Use of Indicator Variables in Regression Analysis. Retrieved from https://www.minitab.com/uploadedFiles/Content/News/Published_Articles/indicator_variables_in_regression_analysis.pdf

Can, F. & Ozkarahan, E. A. (1990, Dec). Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases. ACM Transactions on Database Systems, 15, 483-517. Retrieved from http://dl.acm.org/citation.cfm?doid=99935.99938.

Dhar, S., & Cherkassky, V. Visualization and Interpretation of SVM Classifiers. Retrieved from http://www.ece.umn.edu/users/cherkass/predictive_learning/Resources/Visualization%20and%20Interpretation%20of%20SVM%20Classifiers.pdf

Dimov, S. S., Pham, D. T. & Nguyen, C. D. (2004, 27 Sept). Selection of K in K-means clustering. Proc. IMechE, 219, 103-118. Retrieved from https://www.ee.columbia.edu/~dpwe/papers/PhamDN05-kmeans.pdf

Guyon, I. , Weston, J., Barnhill S., M.D, & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.70.9598&rep=rep1&type=pdf

Hayton, T. (2015, 20 Mar). Scraping with CasperJS. Retrieved from http://toddhayton.com/2015/03/20/scraping-with-casperjs/

Hung, G. K. & Stark, L. W. (1991, Jul). The interpretation of kernels — An overview. Available from https://link.springer.com/article/10.1007%2FBF02584323#page-1

Jivani, A. G. (2011, Nov). A Comparative Study of Stemming Algorithms. I, Int. J. Comp. Tech. Appl., Vol 2 (6). Retrieved from http://www.kenbenoit.net/courses/tcd2014qta/readings/Jivani_ijcta2011020632.pdf

Kleinbaum, D. G., & Klein, M. (2006, 10 Apr )Logistic Regression: A Self-Learning Text. Retrieved from https://books.google.com.sg/books?id=HeT2BwAAQBAJ&lpg=PA385&ots=BnR4qRmt-_&dq=discretion%20for%20logistic%20regression&pg=PA379#v=onepage&q&f=false

Mitra, T. (2012, 21 Jun). Screen Scraping with Node.js. Retrieved from http://code.tutsplus.com/tutorials/screen-scraping-with-nodejs--net-25560

Perlich, C., Provost, F. , & Simonoff, J. S. (2001, Dec). Tree Induction vs Logistic Regression: Learning curve analysis. Retrieved from http://pages.stern.nyu.edu/~fprovost/Papers/logtree.pdf

Sarma, K. S. (n.d.) Theoretical Reference: Combining Decision Trees with Regression in Predictive Modeling with SAS® Enterprise Miner™. Retrieved from http://www2.sas.com/proceedings/sugi30/074-30.pdf

Logistic Regression or Decision Tree. Retrieved from http://datascience.stackexchange.com/questions/6048/decision-tree-or-logistic-regression

Advantage of Logistic over Decision Tree. Retrieved from https://www.quora.com/What-are-the-advantages-of-logistic-regression-over-decision-trees