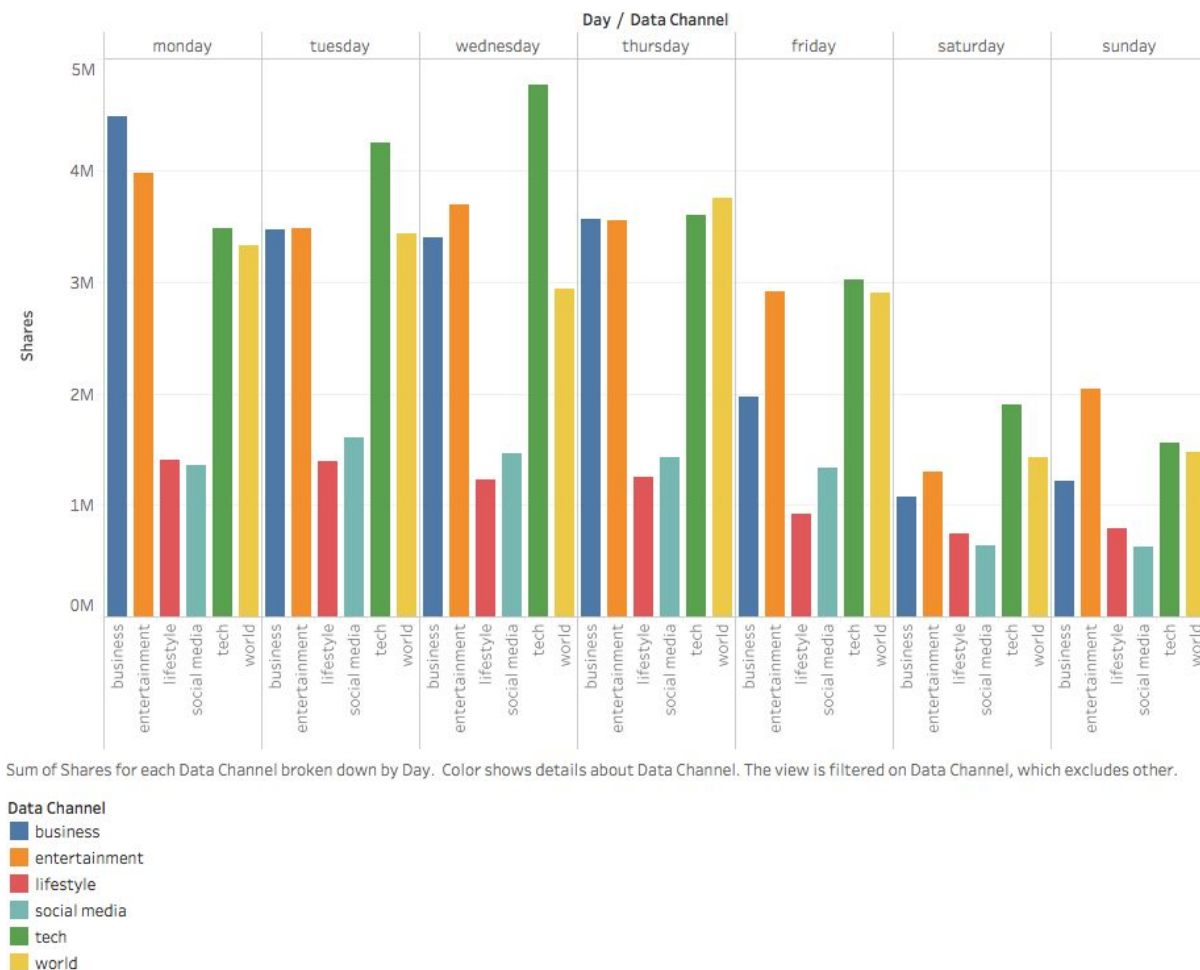


NewsIt: Data Exploration and Visualization

We had some hypothesis in mind and wanted to confirm them from the data so we decided to play with the data a little, and created some visualizations with Tableau. We also found some interesting patterns as below.

Hypothesis 1: The articles related to fashion and/or travel published on the weekends might get more shares as people are free on weekends and spend more time on leisure activities.

Number of Shares varying with the Channel and Day of the week

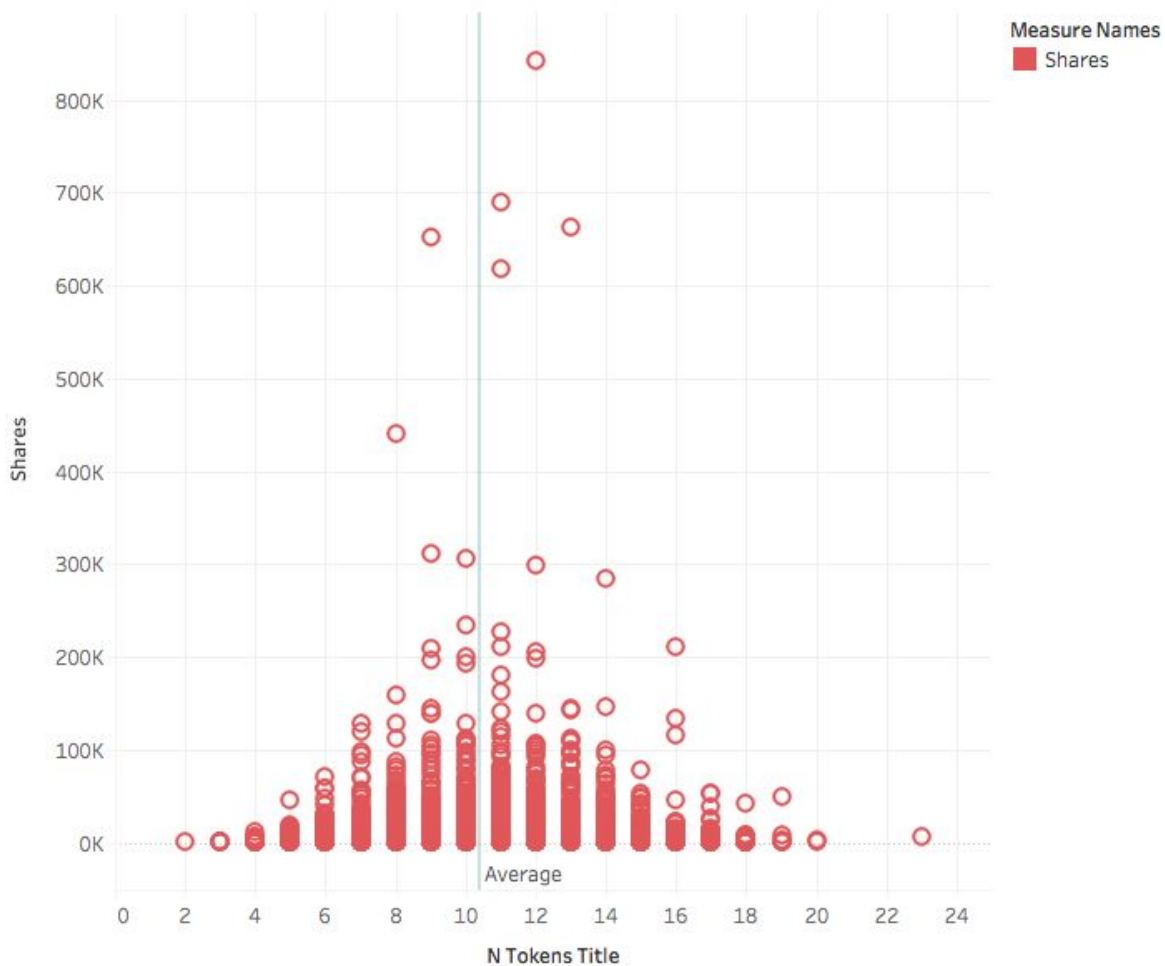


From the bar chart above, we can see that business articles get max shares when shared on Monday, as compared to other days. Also, Tech articles receive maximum shares on Wednesday,

and are mostly in top2 shared articles on each day. On Sundays, entertainment articles receive the most number of shares as compared to any other channel.

Hypothesis 2: The number of words in title affect the number of shares an article will receive.

Number of Shares with respect to words in content and words in title



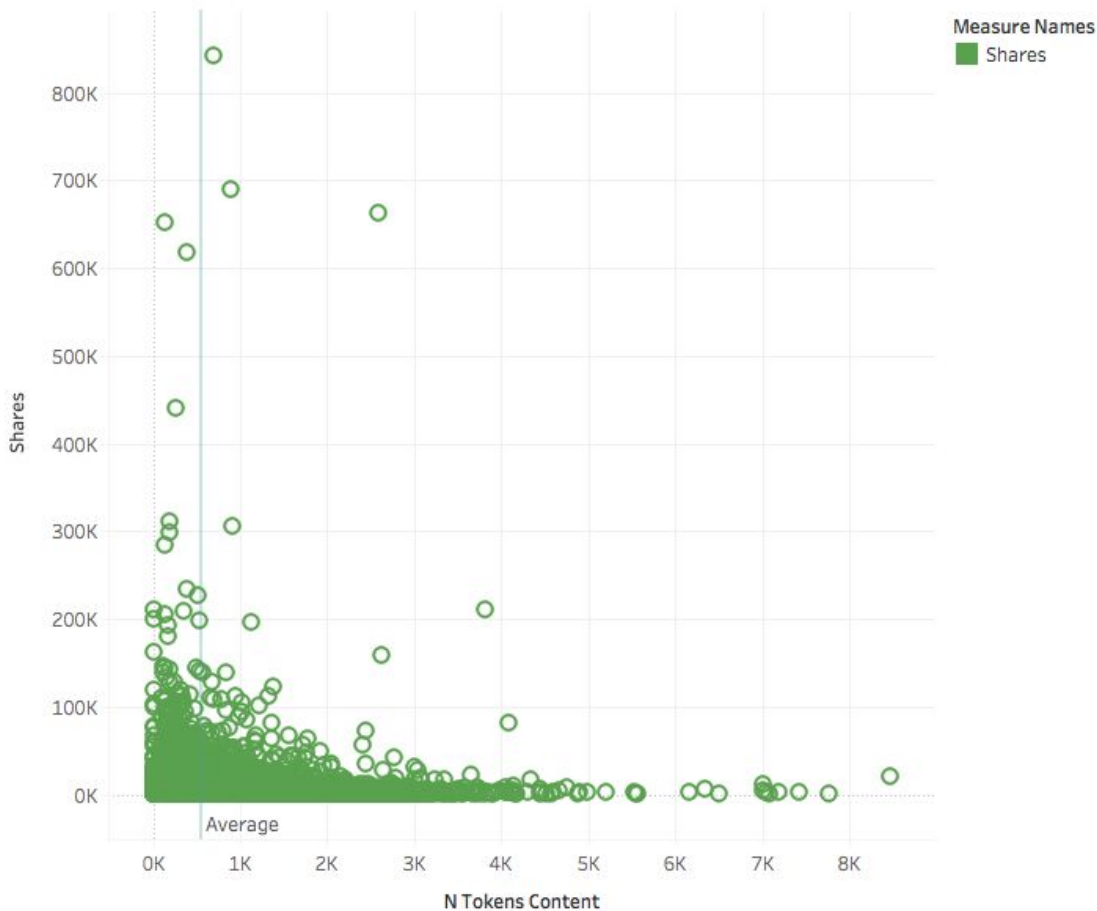
Sum of N Tokens Title vs. Shares. Color shows details about Shares. Details are shown for Url.

It can be seen that number of tokens in title are normally distributed with number of shares. Hence articles with ~10 words in title receive more shares. This makes sense because if there are too many words in a title, the reader might get distracted and not want to read further. Whereas

too less words in the title might not give out information on what the article is actually about and hence not catch user's attention.

Hypothesis 3: The more the number of words in the content, the less shares an article will receive.

Number of Shares with respect to words in content and words in title

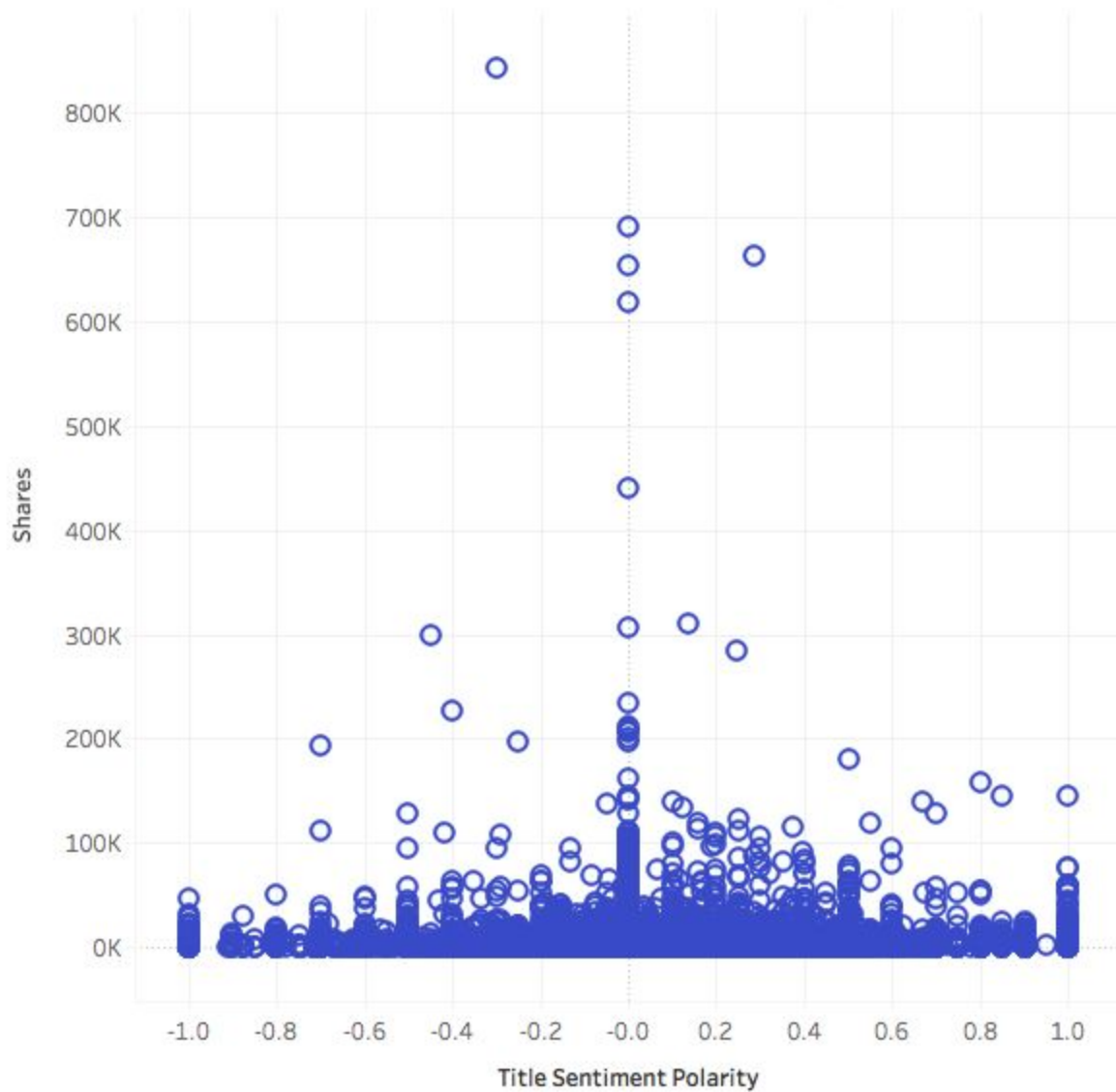


Sum of N Tokens Content vs. Shares. Color shows details about Shares. Details are shown for Url.

The visualization above supports our hypothesis, the more the number of words in the article, the more likely it is to achieve less shares. This makes sense because in general people don't really like to read long articles.

We see (from the graph below) that title sentiment polarity plays an important role in number of shares, which is intuitive, the articles which are more neutral tend to get more shares.

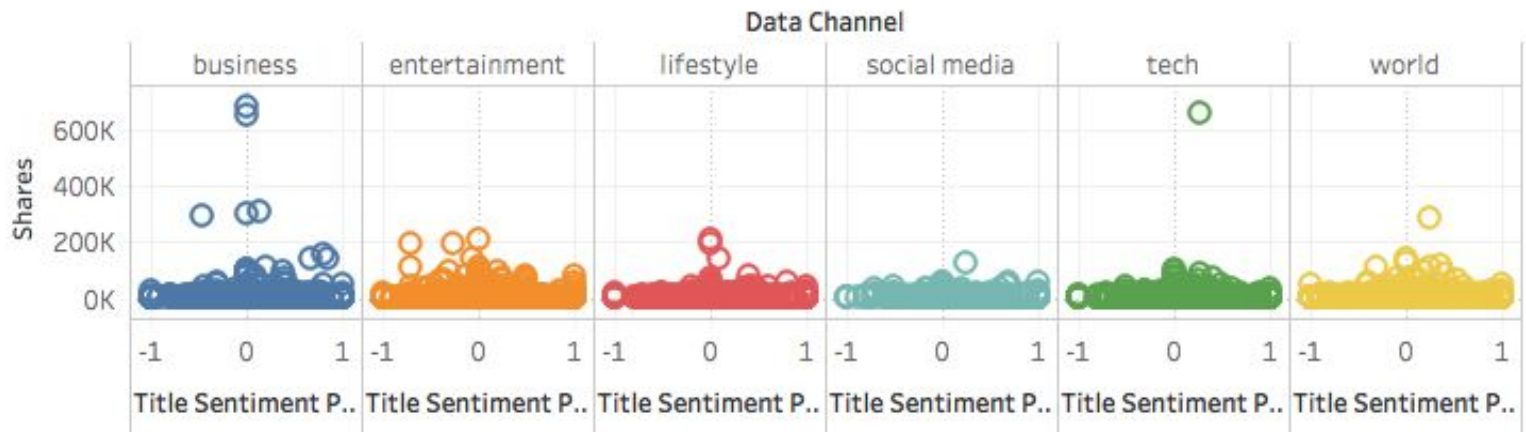
Shares as a function of Title Sentiment Polarity



Sum of Title Sentiment Polarity vs. sum of Shares. Details are shown for Url.

But, another interesting thing is that the polarity makes a difference only for some of the channels. From the figure below, we can see that polarity in business and world channel articles is more important than other data channels.

Shares as a function of Title Sentiment Polarity for different Data Channels



Sum of Title Sentiment Polarity vs. sum of Shares broken down by Data Channel. Color shows details about Data Channel. Details are shown for Url. The view is filtered on Data Channel, which excludes other.

Therefore, depending on the data channel the writers might want to use different words in order to increase the number of shares an article might receive.

One thing to note is that the insights gained above might be only related to how mashable writes articles, since our dataset contains information only about Mashable articles.