**AltSchool of Data Science – Tinyuka 2024 Project Report** **Name:** Mariam Olanipekun  
 **Track:** Data Science  
 **Project Title:** **“Which Headline Works Better? A/B Testing with Real API Data”**

**Project Focus & Research Question**

In this project, I set out to simulate an A/B test using real-world news data to investigate whether certain headline features,specifically **headline length, influence** the depth of article content in health-related news.

Originally, my idea was to test whether the **time of publication** (morning vs. evening) influenced article characteristics, simulating how timing might affect readership or content quality. However, due to constraints in the News API data (explained below), I revised my approach to focus on **headline length** instead.

### **Final Research Questions:**

1. Do shorter or longer health-related headlines tend to produce deeper or more detailed articles?
2. What is the overall emotional tone of health-related headlines, and how might that affect how they are perceived or how deep the article becomes?

**These are highly relevant in health communication, where headlines shape first impressions and can influence trust, engagement, and even behavioral change.**

### **API Data Retrieval: Setup and Challenges**

I began by registering on the **NewsAPI platform** to access live news articles through a REST API. My focus was on articles in English related to the topic “health.”

I initially attempted to pull articles from **April 16, 2024**, but encountered this error:

*“You are trying to request results too far in the past. Your plan permits you to request articles as far back as 2025-05-14.”*

This error occurred because the free NewsAPI tier only permits access to news from the past **30 days**. I corrected this by updating the date range to start from **May 14, 2025**, which successfully returned results.

After fetching the data via the /everything endpoint, I:

* Converted the response into a DataFrame.
* Saved the raw dataset locally.
* Uploaded and saved the file permanently in **my Google Drive** via Google Colab for reproducibility.

### **Static Data for Reproducibility**

Since news data can change with every API call, I chose to **work with a static dataset** to ensure that my analysis remains consistent and shareable. The raw dataset, named health\_raw\_articles.csv, was saved in my Google Drive and forms the basis for the entire project analysis.

### **Initial A/B Test Plan: Time-Based Grouping**

My initial A/B test simulation involved separating articles into two groups based on the time they were published:

* **Group A** – Articles published between **6 AM–12 PM** (morning)
* **Group B** – Articles published between **6 PM–12 AM** (evening)

However, this approach faced several challenges:

* **Inconsistent time formats** (mostly in UTC)
* **Sparse and uneven distribution** of articles across those time windows
* A **limited sample size** (about 90 articles total)

Given these issues, this method did not provide meaningful or reliable groupings. I decided to pivot to a more stable alternative.

### **Final A/B Test: Based on Headline Length**

I transitioned to a simpler and more structured A/B test by using **headline length** as my grouping variable. This method had several advantages:

* Every article had a headline (title)
* Headline length was easy to quantify
* Avoided complications related to time zones or timestamp parsing

#### **➤ Grouping Strategy:**

* I computed the **character length** of each headline.
* I used the **median length** as a cutoff point:  
  + **Group A:** Headlines **shorter than or equal to** the median
  + **Group B:** Headlines **longer than** the median

Using the **median** ensured that the grouping wasn’t skewed by extremely short or long headlines.

### **Exploratory Data Cleaning and Feature Engineering**

To prepare the dataset:

* I **checked for missing values**, most of which were in non-critical fields like author and urlToImage.
* I **dropped columns** that weren’t essential for this analysis: author, url, and urlToImage.
* I added a new column for **headline length**, used for grouping.
* I calculated **article length** by counting the number of characters in the article content.

To improve the reliability of analysis:

* I **removed outliers** in article length using the **Interquartile Range (IQR)** method.

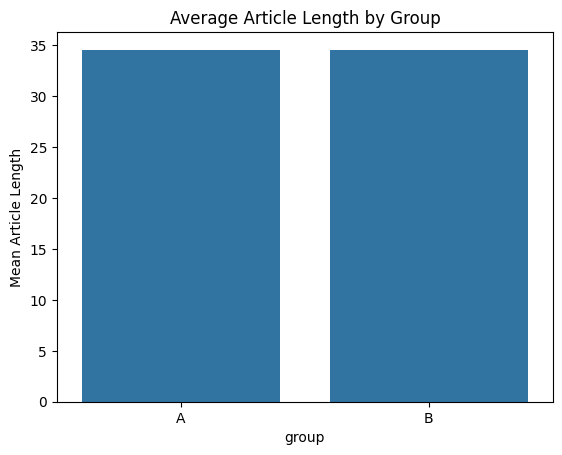
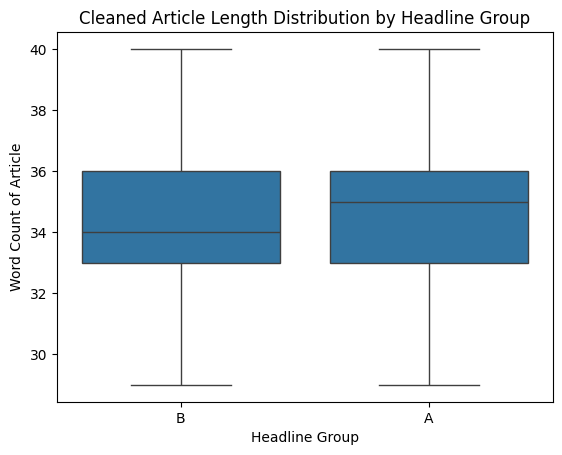
### **Analysis & Visualization**

#### **Headline Length vs. Article Depth**

Using a **two-sample t-test**, I compared the average article length between the two headline groups.

**T-statistic:** -0.067  
 **P-value:** 0.946

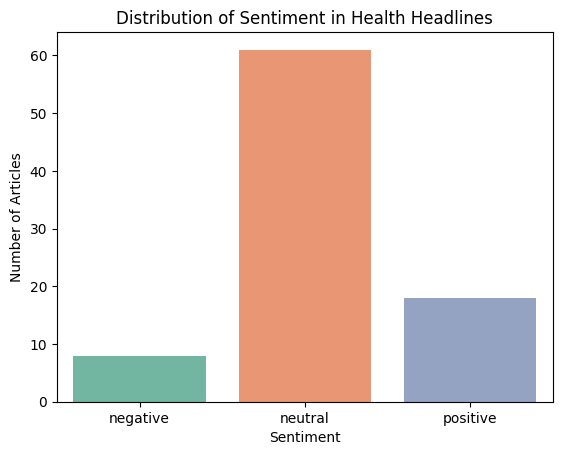
**Interpretation:** The p-value indicates **no statistically significant difference** in article length between shorter and longer headline groups. In other words, **headline length does not appear to impact the depth of the article** based on the available data.



### **Bonus Analysis: Emotional Tone (Sentiment)**

To dig deeper, I conducted a **bonus sentiment analysis** of the headlines using **TextBlob** to classify them into:

* **Positive**
* **Negative**
* **Neutral**

****

Then I examined:

1. Whether articles with **positive vs. negative** headlines were longer.
2. Whether **Group A vs. Group B** differed in emotional tone.

#### 

#### 

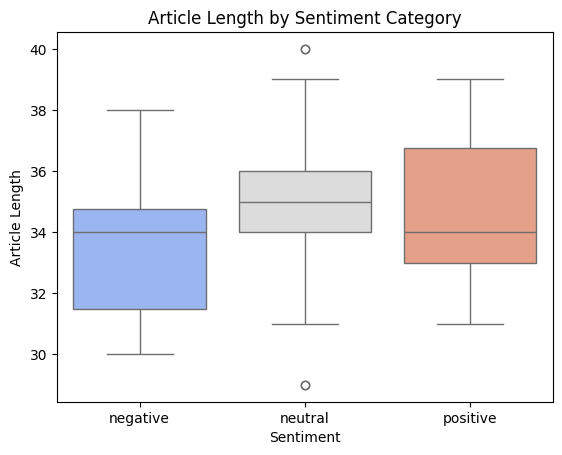
#### 

#### 

#### 

#### 

#### **➤ Sentiment vs. Article Length**



**T-statistic:** 0.623  
 **P-value:** 0.539

#### 

#### 

#### 

#### 

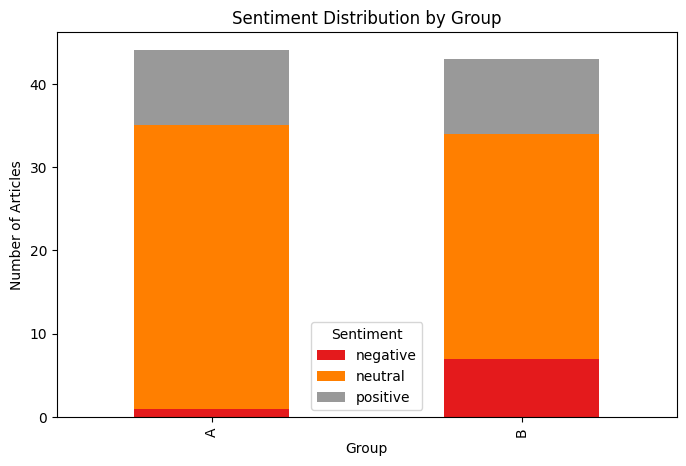
#### 

#### 

#### 

#### 

#### **➤ Headline Length Group vs. Sentiment Polarity**



**T-statistic:** -1.361  
 **P-value:** 0.186

Interpreting Statistical Results

Across the analyses conducted, none of the comparisons showed statistically significant differences. Here's why:

1. Headline Length vs. Article Depth

* t-statistic = -0.067
* p-value = 0.946

A very small t-statistic indicates that the difference between the means of the two groups (short vs. long headlines) is almost negligible. The p-value is much greater than the common significance threshold of 0.05, meaning we fail to reject the null hypothesis. In simple terms, there's no evidence that headline length affects article length in this sample.

2. Sentiment vs. Article Length

* t-statistic = 0.623
* p-value = 0.539

Again, the t-statistic is small, suggesting the groups (positive vs. negative/neutral headlines) do not differ much in article length. The p-value is above 0.05, so the result is not statistically significant.

3. Headline Group (A/B) vs. Sentiment Score

* t-statistic = -1.361
* p-value = 0.186

This test explored whether longer or shorter headlines were more emotionally charged. The p-value here is also above 0.05, and the t-statistic isn't large enough to suggest a strong difference between groups.

### Why These Results Are Not Statistically Significant

A high p-value (> 0.05) means that any observed differences could easily have occurred by chance given the sample size and variation in the data. Similarly, small t-statistics suggest that the means of the groups are close ,too close to infer any real-world effect with confidence.

## Key Insights Summary

* There was no significant difference in article length between shorter and longer headlines.
* Emotional tone (sentiment) of headlines also did not significantly affect article depth.
* The small dataset and lack of direct engagement metrics (e.g., clicks, read time) are likely reasons why strong relationships were not observed.
* Future work could incorporate larger datasets, reader engagement data, and advanced NLP tools for deeper insight.

### Key Insights & Recommendations

* No significant relationship was found between headline length and article depth in health-related articles.
* Sentiment tone also showed no significant effect on article length, though further exploration with a larger dataset could yield more insight.
* Working with real API data introduced real-world challenges (rate limits, missing data), which made the analysis more practical and grounded.
* Simulated engagement via article length is a proxy and doesn’t fully capture how readers interact with headlines. Future projects could use metrics like clicks, shares, or read time for richer insights.

### **CONCLUSION**

This project used a mix of API data retrieval, A/B testing principles, data cleaning, visualization, and statistical inference to explore whether **headline features** (length and sentiment) influence health article depth. While no clear performance difference was found between groups, the process built critical skills in real-world data handling, hypothesis testing, and storytelling through data.