For your final project, essentially you would submit a Jupyter notebook that contains an application of **clustering analysis** / time-series forecasting / supervised machine learning using a real-world dataset. If you are doing an analytics related project (e.g. demand forecasting) for your capstone/thesis, you may use that as a submission for this class. The only main requirement is that you would use machine learning for your project, and it would essentially follow the same case study format that you guys have done previously. Likewise, your project must be different from the case studies that you have done previously.

Topic: [Clickbait News Articles](https://www.kaggle.com/datasets/amananandrai/clickbait-dataset?fbclid=IwAR0WeS73HK8AkFKSW2IQY_QN3RCVQOYPUmK2fqA110oeDCXF3jmTwgT1E8c) for Clustering Analysis

1. Introduction/Motivation
   1. After nearly three decades since the dawn of the internet, the world has seen exponential growth in the advancement of information technology, which led to improvements in quality of life that we enjoy today such as online communications, access to information, and cashless transactions. While it definitely has amplified the convenience of many facets of life, what inevitably followed is the danger of social engineering. Although it may be less evident, clickbaiting may arguably be considered as one of these given that it lures people into consuming misleading content. As such this study aims to investigate the nature of this phenomenon, specifically by analyzing news headlines and delving into the common words used as well as the subject matter of the article. With this, the proponents hope to shed light on the tendencies of such a phenomenon so that the public can discern and be more critical of the content that they consume.
2. Problem Statement/s

The advancements in technology have led society towards a digital space of information overload and noise, generating a wave of clickbait and misleading information even in journalism. Today, internet consumers have become the victims of misinformation or “fake news” as an increasing number of sites use exaggerated titles or “emotional traps” to stand out and gain foot traffic in a world of digital oversaturation. Overall, information literacy has been negatively impacted by such practices from these sites. In order to elaborate on these concerns, the group seeks to answer the following questions:

* + 1. What words are commonly used in clickbait and non-clickbait headlines?
    2. What topics are commonly talked about in clickbait and non-clickbait articles?
    3. What distinguishes clickbait from non-clickbait? How to predict if a headline is clickbait or non-clickbait?
    4. What can be common signs of clickbait and how can we avoid them?

1. Data Description
   1. The clickbait headlines are collected from sites such as ‘BuzzFeed’, ‘Upworthy’, ‘ViralNova’, ‘Thatscoop’, ‘Scoopwhoop’ and ‘ViralStories’.
   2. The relevant or non-clickbait headlines are collected from many trustworthy news sites such as ‘WikiNews’, ’New York Times’, ‘The Guardian’, and ‘The Hindu’.
2. Methodology
   1. ETL
   2. K-means clustering
3. Exploratory Data Analysis
   1. Word cloud
4. Clustering Analysis
5. Word Cloud Analysis
   1. Clickbait headlines use directive and sensational language to create claims that peak consumer interest. In the first cluster, the most evident words seen are “zodiac”, “sign” and “base” followed by “character”, “know” and “favorite”. This combination suggests that astrology and other personalized tests are highly engaging as people seek to increase their self-awareness or validate their self-concept. In the second cluster, the most evident words seen are “thing”, “people” and “time” followed by “know”, “like”, “need”, “new”, “actually”, “life”, “good” and “love”. This combination can suggest that people are interested in learning something new or personal about life, themselves or others. In the third cluster, the most evident words seen are “base”, “know” and “question” followed by “harry”, “favorite”, “celebrity”, “need”, “potter” and “guess”. This combination suggests that people are interested in shocking and superficial content. Overall, the clickbait clusters use words that sensationalize or personalize information to incite emotion and attachment towards topics. This attracts consumers as it increases curiosity and interest, thereby increasing conversion to these sites as people click to “know more”.
   2. On the other hand, non-clickbait headlines use more specific, neutral and informative language. In the first cluster, the most evident words revolve around death, aging, and public figures such as singers, actors and authors. In the second cluster, the most evident words are “new”, “win”, “kill” and “find” followed by “president”, “obama”, “china”, “dead”, “year”, “australian”, “plan” and “court”. In the third cluster, the most evident words are “kill”, “crash”, “pakistan” and “bomb” followed by “attack”, “afghanistan”, “blast”, “bombing”, “dead'', “suicide” and “plane”. Overall, the non-clickbait clusters include specific key details and information that can precisely reflect relevant topics and content delivered in a period of time. These are highly informative as they get direct to the point and express the value and pertinence of information being shared.
6. Conclusion
7. Recommendations