A Summary of Adjustments for HUMA 5630 Final Project of Group 5 (Junwei DENG)

Part 1: removed the emphasis on COVID-19 from the research question (still mentioned it when analyzing)

Part 2: data processing

* gave up making the numbers of albums for each year even and made full use of the charts I had scraped, resulting in larger data size.
* detected the English music with Python instead of handpicking.
* now there is only 1 dataset (NA-dropped) instead of 2 (NA-dropped and NA-replaced) due to the adjustments above

Part 3: words and topics

* optimized the workflow of topic modeling, including
    - switching from *nltk* to *stanza* for lyrics pre-processing (the latter one performs better)
    - using the lemmatized words instead of the original ones to avoid the situation where words with different forms are actually semantically the same
    - removing all the curse words using a GitHub list/ outliers using frequency z-scores
* the number of topics changed from 6 to 10 due to the aforementioned adjustments
* added a wordcloud of 2020s pop music lyrics

Part 4: sentiment analysis

* switched from *seaborn* to *plotly* for visualization and now the figures are interactive
* gave up asking *OpenAI* about the gender of artists and displaying it in figures, since
    - the risk of *OpenAI* making mistakes
    - that plotly still failed to achieve satisfying performance when adding the gender attribute after quite a few tries
* handpicked a new group of complete albums for the heatmaps

Part 5: about the genres

* switched from word embedding visualization to network visualization (I think the latter one makes more sense in my project)

Part 6: about the website

* divided into 4 pages / added a banner using album covers / added some widgets (toggle and selectbox) / added descriptions about my project