# CMPT 353 Project Report

**Ike Chan 301396674 ikec@sfu.ca**

**Lucas Lee 301437248 ldl5@sfu.ca**

https://github.sfu.ca/ldl5/CMPT-353-Final-Project

## Overview

For the final project, our group plans to use the Reddit Submissions and Comments cluster to answer the question:

***"Are manga subreddits more active when anime adaptations are made?"***

To answer this, we plan on incorporating the methods we used in Exercise 5 to analyze a shift in p-values from the number of posts and comments by identifying if certain time frames match up with a provided announcement.

For example, an anime adaptation of a popular manga known as Oshi no Ko was announced on June 9, 2022. Our group would like to analyze a change in the number of posts and comments a week before and after the announcement was made, including the day of the announcement, and follow this same example with other announcements of other series. Since we are comparing p-values, our hypotheses are:

**Null Hypothesis:** There is no difference in the amount of activity in a subreddit when an anime adaption based on a manga is released

**Alternate Hypothesis:** There is a difference in the amount of activity in a subreddit when an anime adaption based on a manga is released

However, determining “activity” is a relative measurement. A subreddit with only one person who posts all the time may be considered more active than a subreddit with thousands of people who only post once in a blue moon. As such, our team has decided to identify the total number of comments posted in a month, as well as the number of comments per submission, to determine how “active” a subreddit is.

## Data Used

Before we started our project, we needed some items to ensure our code was working . One outside library we used is known as **JikanPy**, a Python wrapper for the Jikan API which gets anime and manga data from MyAnimeList. MyAnimeList is a popular anime website to record anime and manga data so being able to use this API helps grab a given anime’s data without having to record everything manually. Furthermore, we retrieved Reddit data from SFU’s cluster by filtering for specific time frames; in this case, retrieving submissions before and after an anime adaptation was released.

- JikanPy - Python API for anime releases: https://github.com/abhinavk99/jikanpy
- Cluster Data from SFU: http://cluster.cs.sfu.ca/
- Cluster Files for our Project: https://github.sfu.ca/ldl5/Reddit-Cluster-Files

To select manga that have anime adaptations and have subreddits on Reddit, we chose anime from MyAnimeList that followed 3 criteria:

1. Each anime adaptation has a corresponding subreddit with at least 100k members
2. The anime adaptation was in the top 5 of popularity in its respective release season
3. A season was released before August 2022 as that is when SFU's Reddit data stops

As such, we chose the following four anime/manga subreddits that fit the criteria. Furthermore, each anime that we retrieved was the first adaptation of its respective subreddit (e.g. JuJutsu Kaisen's first season was in fall 2020, Kimetsu No Yaiba's first season was in spring 2019). The following subreddits were chosen like so:

**Format**: *SubredditName (Release Season)*

1. JuJutsuKaisen (Fall 2020)
2. SpyxFamily (Spring 2022)
3. Komi_san (Fall 2021)

Once we imported the anime data using JikanPy, we then got the previous, and next, months depending on the day of the month the anime was released. If an anime was released in the first 10 days of the month, we recorded the previous month, and if it was released in the last 10 days of the month, we recorded the next month. If an anime was released in between those first and last 10 days, we would get the previous and next month. The reason for this is that we could only get monthly data, and needed to draw the line as to which other month we would compare the data we got from the month the anime was released.

We could then filter and clean our data by ensuring our cluster data matched with the provided data we retrieved from JikanPy. For example, if JuJutsuKaisen was released on October 3, 2020 and SpyxFamily was released on April 9, 2022, we would only get data from JuJutsuKaisen in September and October of 2020, while also getting data from SpyxFamily in March and April of 2022. The subreddits and corresponding dates would not overlap and thus we would not need to read extra data from the cluster.

## Data Analyzation Techniques

Using the SFU Reddit Cluster, we decided to utilize the number of comments in each submission to calculate our p-values to determine which hypothesis is reached. We will reject our null hypothesis if $p < 0.05$. To analyze our data and get our p-values, we employed a couple of statistical tests namely a Mann-Whitney U and chi-squared test.

The Mann-Whitney U test was to identify if there was any lopsided data meaning that one month would have more submissions or comments than the previous or next month. If we found p to be less than 0.05, we would then reject the null hypothesis for this given scenario.

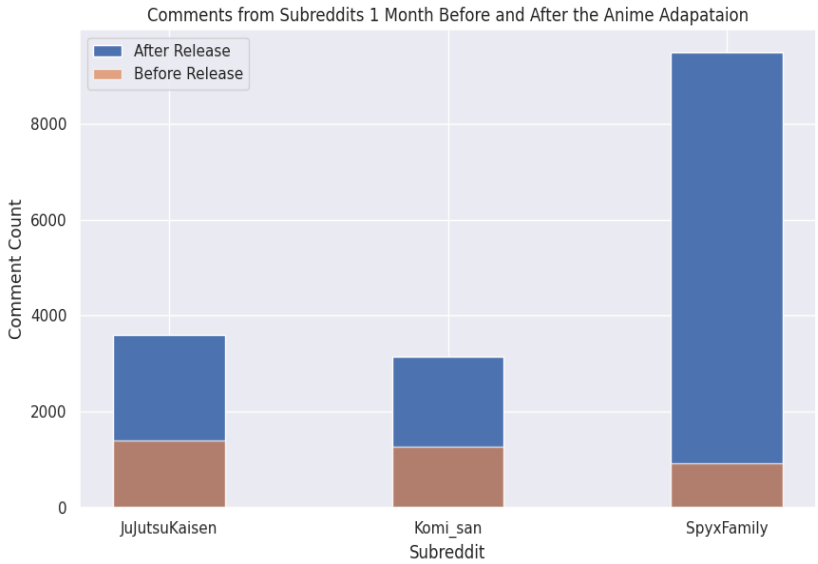
For the chi-squared test, we would identify if releasing an anime had an effect on the activity of a subreddit by identifying the release month's submissions and comments to the previous, or next, month's submissions and comments. Just like our Mann-Whitney U test, we would reject our null hypothesis if $p < 0.05$.

## Results

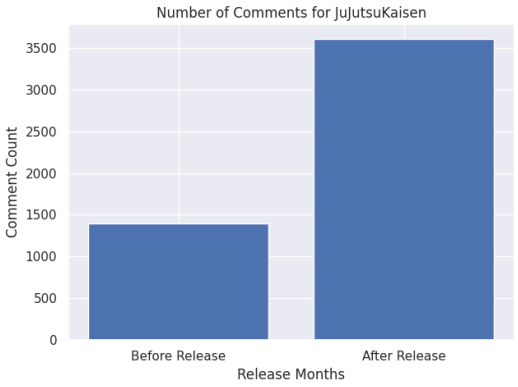
After running our code, we observed p-values using all three subreddits to be less than 0.05. This indicates that we should reject our null hypothesis and accept our alternative hypothesis. In the Mann-Whitney U tests, JuJutsuKaisen obtained a p-value of 0.0003770739, SpyxFamily obtained a p-value of 0.0010319917, and Komi_san obtained a p-value of 0.00274725274. As for the chi-squared test, we obtained a statistic of 15,385 leaving us with a p-value of 0. As such, this means that there is a difference in the amount of activity once an anime adaptation is released, presumably because there is more to discuss about a given manga/anime series.

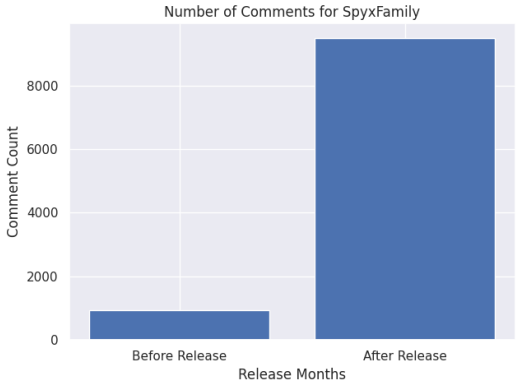
## Visualization of Data

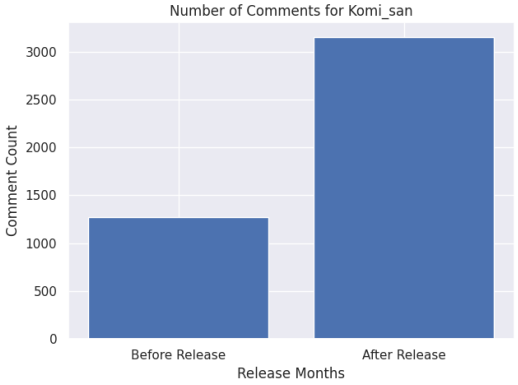
The visualizations we have are indicative of the results we have obtained that back up our alternative hypothesis. The bar charts shown here indicate that there is a higher proportion of comments after an anime was released.



We can also break down the data into individual charts to further see the distribution in comments. It is easy to notice the difference in comments before and after an anime is released, with the timeframe after a release having significantly more comments than before it was released. This helps further prove our alternate hypothesis to be correct by visually identifying a change in the number of comments.




## Limitations

When thinking of the problem we wanted to solve, our initial idea was to calculate the activity one week before and after an anime was released. However, we quickly noticed that the Reddit data did not record individual dates and had to shift our plan to record the month before/after and during the release. This made our data more inaccurate than we would have liked it to be as we had to calculate whether to get data from the previous, or next month, depending on when the anime was released. As such, our p-values may not be as accurate as we would like it to be. If given more time, we would possibly look for an alternate source of data that included individual days in both Reddit submissions and comments.

Furthermore, we would also have liked to analyze more subreddits to provide a more accurate overview of our data. However, we noticed that some subreddits had an absurd amount of data that was not feasible for our study. For example, the subreddit known as ShingekiNoKyojin with 1.2 million users added 10,000 tasks when reading the data from the cluster and would have significantly slowed down our research. If provided with more time and better hardware and software, we could analyze more subreddits to get a more accurate overview.

Finally, although not exactly a limitation, the subreddits we chose all seemed to be in the fall or spring season. This was unintentional and does not give a full scope of activity found in each season. If we were to continue this project, we would retrieve subreddits from every possible release season.

# Project Experience Summaries

## Ike Chan

Entering this project, I was under the impression that this would be easy. I admit that I believed our thesis to be simple. However, when actually implementing the methods to achieve its proof, it began to wreck my brain. I found myself struggling to imagine the solutions. Even when I did manage to get a clue as to what I should be doing, actually implementing the methodologies proved to be even more difficult.

At the beginning, we wanted to measure the activity of a subreddit each day, assigning a score to each activity: post a submission, a comment, an upvote, etc. Then, we realized that the dataset doesn't have the day locked down. We ditched that idea and decided to measure by month instead. I had a suspicion that the "score" property is the upvote property, but due to uncertainty, we ditched the idea of measuring activities using upvotes as well. At the end, we decided to only measure activity using the number of submissions and comments individually.

We had a simple idea: t-test the datasets to make sure the means are different, so to disprove the null hypothesis. Troubles introduce themselves when I realized that the dataset doesn't even meet the requirements for t-tests.

We solved all those problems, but the end results are completely different from what we had imagined at the beginning-- really proved the randomness of data science and the virtues of a data scientist to be able to improvise on the spot.

## Lucas Lee

To complete the project, I took initiative by actively communicating with my teammate, who I had not met before, and devising ideas related to data science. Once we had agreed on the idea that we wanted to solve, it was my job to filter and clean the data from the cluster and develop appropriate visualizations related to our data.

To achieve this, I first searched for ways to retrieve anime statistics into Python and stumbled upon Jikan, and JikanPy, to retrieve anime statistics from MyAnimeList. With the help of the API, I manually picked manga subreddits that had corresponding anime adaptations before August 2022. From there, I used Spark Dataframes to read from SFU's Reddit cluster and then filtered the data by ensuring I downloaded certain months and years that corresponded to a given subreddit. Once the data was collected, I decided to visualize our data using bar charts to indicate a numerical difference in the number of comments before and after an anime release.

As a result, the data I cleaned and collected ensured we could obtain our p-values to be as accurate as possible, as well as provide a brief visual overview through our visualizations. While a statistician may understand how to use p-values, an average person may not, which is why the visualizations were needed in further helping our reader understand our result.