Qingboyuan Wang (NetID: qw120)
Keqin Liu (NetID: kl972)
Jia Lyu (NetID: jl2749)

1. Basic info

Figure Skating: How have technical components progressed in figure skating over the years? And are there any biases in scoring these components based on the nationality of the judges and athletes?

2. Overview

Figure skating is a popular and highly technical sport which originated in the later part of the 19th century in Europe. It then began to take place in world-class competitions. The audiences, including me, were often fascinated by the beautiful movements and difficult jumps the athletes made. Another beauty of this sport is that no one would know who will be the winner since the highly technical movements and the enormous pressure one has to compete under. Then, our team kept wondering from Torino 2006 to Beijing 2022 Winter Olympics, how the technical components have progressed in figure skating, and if there are any biases in scoring these components based on the nationality of the judges and athletes.

3. Description of the data

1) https://github.com/BuzzFeedNews/figure-skating-scores

There are four CSV files (judge-scores, judged-aspects, performances, programs) and their raw JSON versions in this data set. The “judge-scores” contains three columns: “aspect_id”, “judge” and “score”. And it will be merged (left-join) with “judged-aspects” by “aspect_id” to form a comprehensive data set containing twelve columns representing different compositions of the final score (scores_of_panel). Also, “performances” and “programs” will be left-joined together by “performance_id”, which will contain the nations represented, total segment and element scores, total deductions and the total component score.

We would need to clean the “judged-aspects” dataset since there are sparse NULL values in it. We are going to either drop them or fill them with imputed values. We will be using Python to implement the processing pipeline.

Qingboyuan Wang (NetID: qw120)
Keqin Liu (NetID: kl972)
Jia Lyu (NetID: jl2749)

2) **https://www.kaggle.com/the-guardian/olympic-games?+select=winter.csv&select=dictionary.csv**

There are three CSV files in this data source: dictionary, summer and winter. Both summer and winter dataset describes the Olympic sports and medals at summer and winter games from 1896 to 2014. It contains: Year, City, Sport, Discipline, Athlete, Country, Gender, Event, Medal. And the dictionary dataset contains the IOC country codes and population/GDP estimates and it will be merged with both summer and winter datasets. These dataset do not require much cleaning.

3) **https://www.kaggle.com/katiewerner/us-figure-skating-data-2020-regionals-int-ladies?select=Scraped_Component_Scores.csv**

These datasets are focusing on the 2020 Regional Intermediate Ladies events and it contains five CSV files: Event Details (event index, round, program type), Official Details (judge and technical official names, cities), Skater Details (skater placement, skate order, club, overall scores and deductions), Technical Score Details (base values, GOEs), Component Score Details (component scores). And the event index and skater placement columns are keys that can be used to join these datasets together, as appropriate. These dataset do not require much cleaning.

4) **https://skatingscores.com/**

These are official real-time score reports for skating athletes who participated in past Olympic Winter Games. And we will be focusing on Men and Women Single Skating score reports. So that will be four PDF files that are copies of the official score reports containing: Rank, Name, NOC Code, Starting Number, Tota; Segment Score, Total Element Score, Total Program Component Score and Total Deductions. And each of these scores is computed by taking the average of the nine judges' scores.

These dataset would require an extensive data cleaning process since we would need to convert PDF files into CSV and that would cause

Qingboyuan Wang (NetID: qw120)
Keqin Liu (NetID: kl972)
Jia Lyu (NetID: jl2749)

a lot of formatting issues. But we will be developing an automated pipeline to clean these files more efficiently.

4. Usage scenario & tasks

Mr. Smith is the head judge appointed by the Olympic committee. During the 2026 Olympic Winter Games, he received a series of complaints regarding the unfairness of scores for figure skating. Mr. Smith would need to explore the possible causes or reasons behind it. For the first step, he needs to identify which score reports are being considered unfair and then compare the potential unfair scores with scores from past years. He would learn from the visualizations from previous years and see if the score trends for an athlete are similar to previous games. But he then thinks that the score trends would not be a good reference since assuming the athletes' skillfulness stays the same level is inappropriate. So, he decided to use the model we developed to actually predict a player's performance/scores and compare them to the scores the judges gave. Additionally, he figured that it would be better if we can visualize the trends of scores given by each judge to an athlete so it would be easier to determine if the judges are making mistakes. Hence, he needs to conduct a follow-on study.

Qingboyuan Wang (NetID: qw120)
Keqin Liu (NetID: kl972)
Jia Lyu (NetID: jl2749)

5. Description of visualization & initial sketch

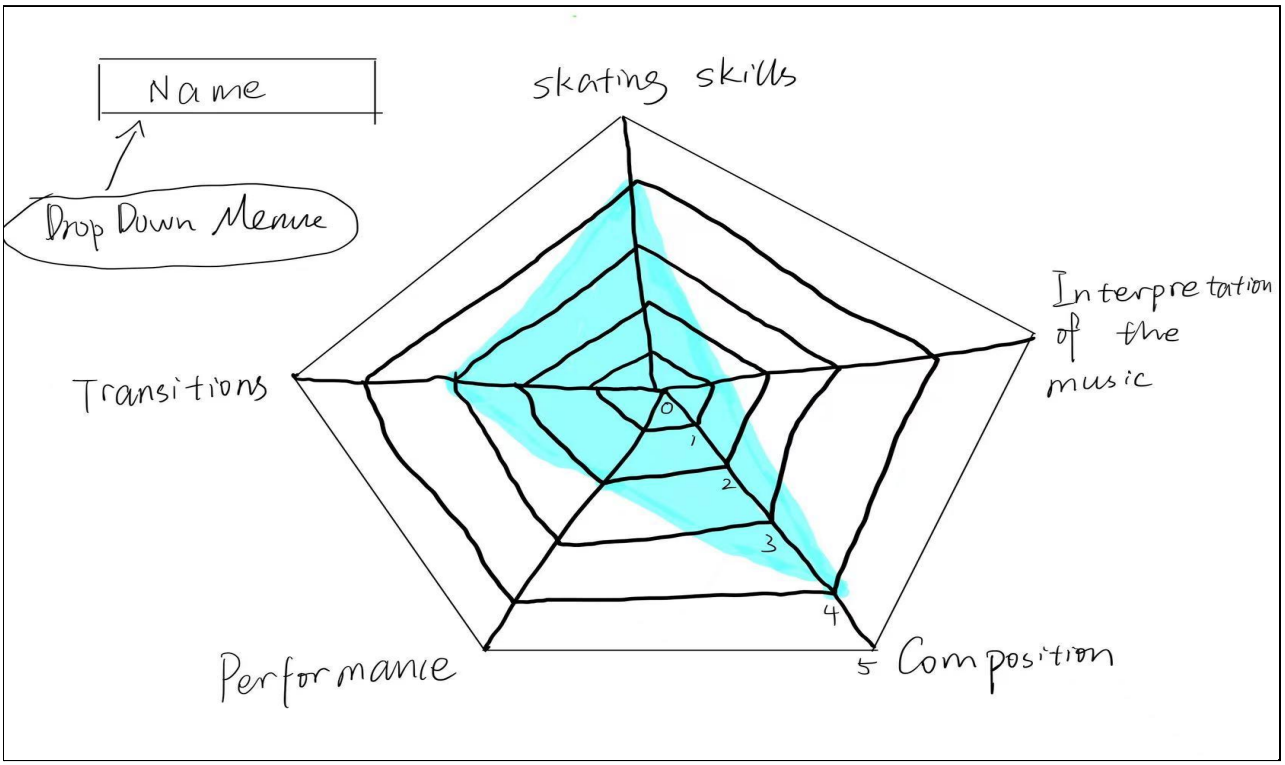


Figure 1

Figure 1 High-Level Description (Spider Plot):

Figure 1 gives us a detailed breakdown of each skater/athlete's figure skating scores and it would enable Mr. Smith, the head judge, to clearly see the past scores decompositions for a specific athlete. Thus, he would make an informed decision on whether the scores are fair or not. He would also be able to select different names to see different scores for different athletes.

Qingboyuan Wang (NetID: qw120)
Keqin Liu (NetID: kl972)
Jia Lyu (NetID: jl2749)

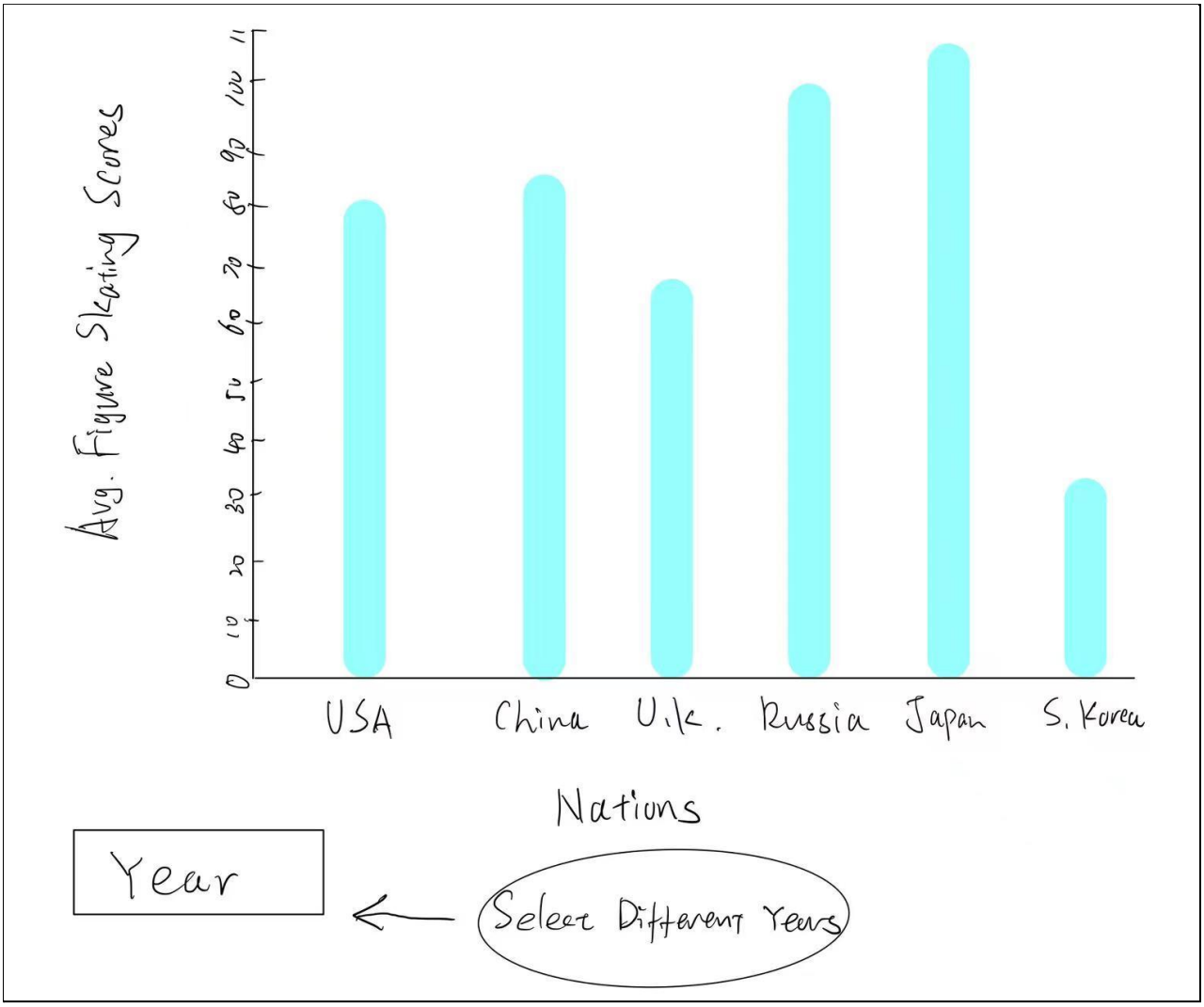


Figure 2

Figure 2 High-Level Description (Bar Plot):

Figure 2 gives a high-level summary of the average figure skating scores for each nation represented. Mr. Smith would be able to see the average score changes throughout different years by selecting the “Year” dropdown menu. It would be more convenient to observe the changes of scores. And Mr. Smith will be able to see if a specific athlete’s score (from Figure 1) is similar to the average score of the country he/she represented. Thus, he would make an informed decision on whether the scores are fair or not.

Qingboyuan Wang (NetID: qw120)
Keqin Liu (NetID: kl972)
Jia Lyu (NetID: jl2749)

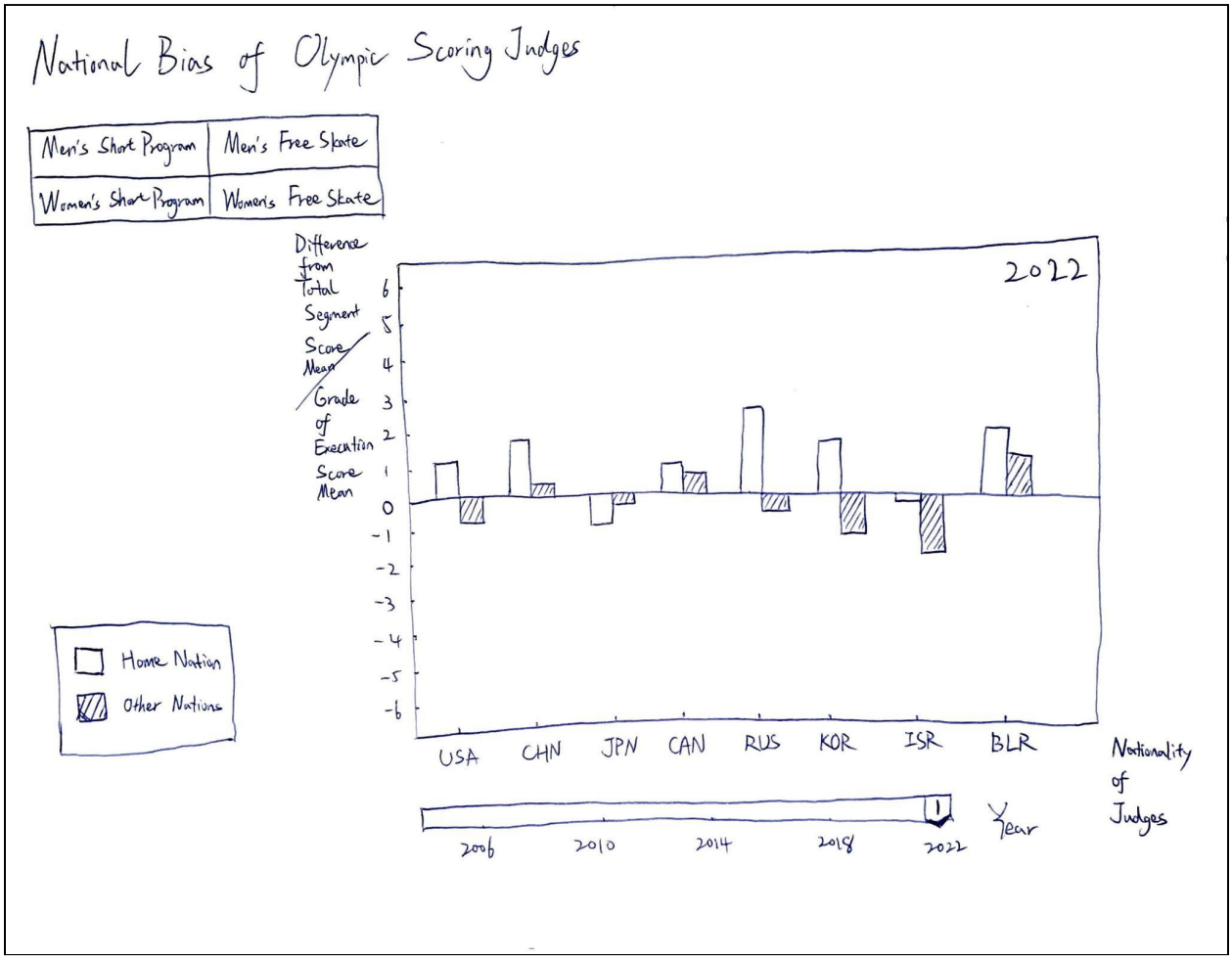


Figure 3

Figure 3 High-Level Description (Side-by-Side Bar Chart):

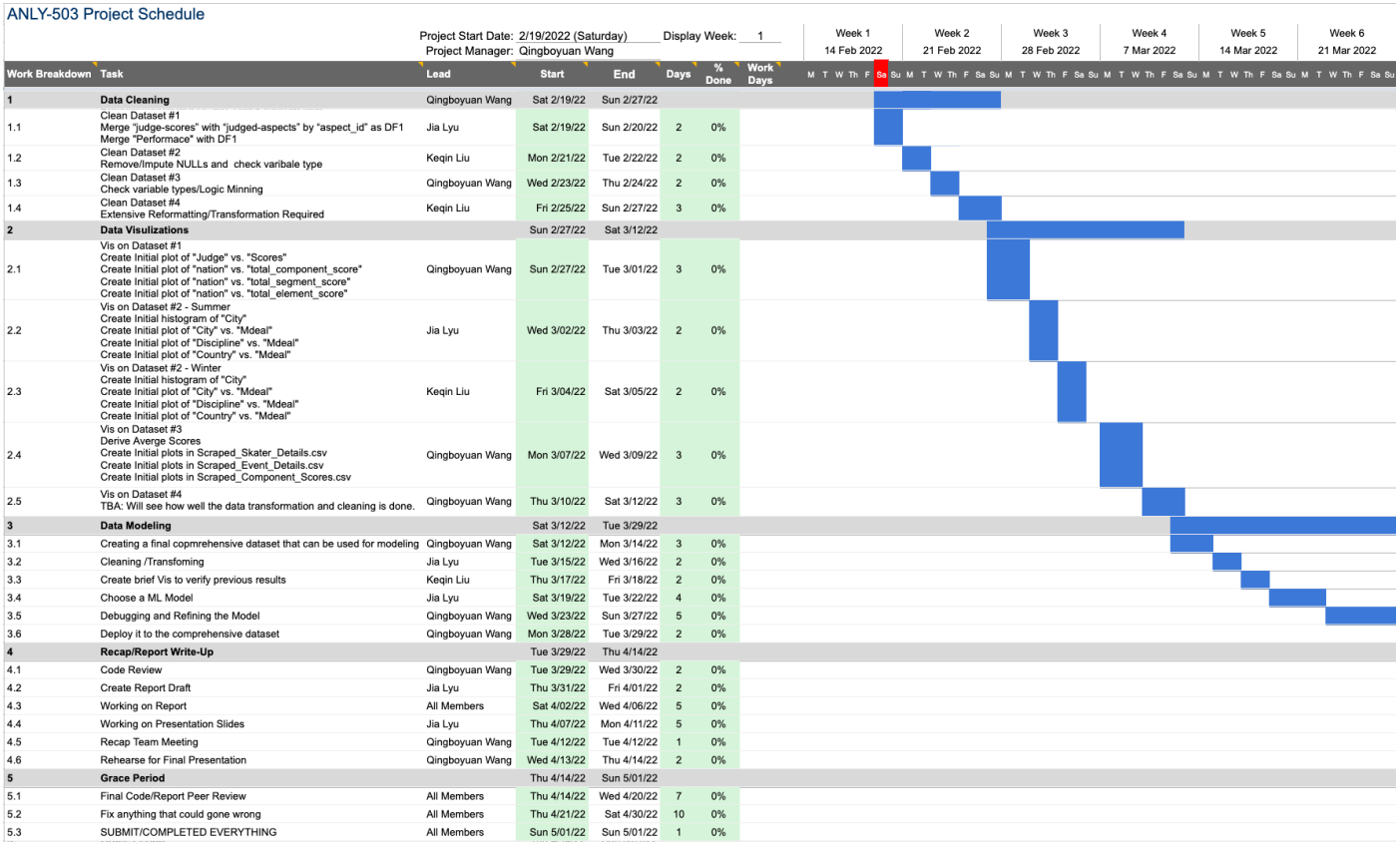
Figure 3 gives a high-level summary of the scoring differences for judges of each nationality from the mean figure skating scores. Mr. Smith would be able to see the difference throughout different years by sliding the “Year” slider. It would be more convenient to observe the biases of each judge when scoring athletes of their home nations versus other nations. And Mr. Smith will be able to see if a judge has a tendency to score higher for athletes of their home nation origin versus other origins. Thus, he would make an informed decision on whether the scores are fair or not.

Qingboyuan Wang (NetID: qw120)
Keqin Liu (NetID: kl972)
Jia Lyu (NetID: jl2749)

6. Work breakdown and schedule

We are using a Gantt Chart on Google SpreadSheet for our group's work breakdown and schedule. A brief screenshot is attached below and it can be accessed using this link: https://docs.google.com/spreadsheets/d/1M3ZgbYWwrO9kRF7wwRixxqP-OSj_TvTDxyTjswh04G4/edit?usp=sharing

Adjust the "Display Week" parameter in cell "H3" to see graphical schedules for each week.



ANLY-503 Project Schedule

Project Start Date: 2/19/2022 (Saturday) Display Week: 1
Project Manager: Qingboyuan Wang

Work Breakdown	Task	Lead	Start	End	Days	% Done	Work Days
1	**Data Cleaning**	Qingboyuan Wang	Sat 2/19/22	Sun 2/27/22			
1.1	Clean Dataset #1 Merge "judge-scores" with "judged-aspects" by "aspect_id" as DF1 Merge "Performace" with DF1	Jia Lyu	Sat 2/19/22	Sun 2/20/22	2	0%	
1.2	Clean Dataset #2 Remove/Impute NULLs and check varibale type	Keqin Liu	Mon 2/21/22	Tue 2/22/22	2	0%	
1.3	Clean Dataset #3 Check variable types/Logic Minning	Qingboyuan Wang	Wed 2/23/22	Thu 2/24/22	2	0%	
1.4	Clean Dataset #4 Extensive Reformatting/Transformation Required	Keqin Liu	Fri 2/25/22	Sun 2/27/22	3	0%	
2	**Data Visulizations**		Sun 2/27/22	Sat 3/12/22			
2.1	Vis on Dataset #1 Create Initial plot of "Judge" vs. "Scores" Create Initial plot of "nation" vs. "total_component_score" Create Initial plot of "nation" vs. "total_segment_score" Create Initial plot of "nation" vs. "total_element_score"	Qingboyuan Wang	Sun 2/27/22	Tue 3/01/22	3	0%	
2.2	Vis on Dataset #2 - Summer Create Initial histogram of "City" Create Initial plot of "City" vs. "Mdeal" Create Initial plot of "Discipline" vs. "Mdeal" Create Initial plot of "Country" vs. "Mdeal"	Jia Lyu	Wed 3/02/22	Thu 3/03/22	2	0%	
2.3	Vis on Dataset #2 - Winter Create Initial histogram of "City" Create Initial plot of "City" vs. "Mdeal" Create Initial plot of "Discipline" vs. "Mdeal" Create Initial plot of "Country" vs. "Mdeal"	Keqin Liu	Fri 3/04/22	Sat 3/05/22	2	0%	
2.4	Vis on Dataset #3 Derive Averge Scores Create Initial plots in Scraped_Skater_Details.csv Create Initial plots in Scraped_Event_Details.csv Create Initial plots in Scraped_Component_Scores.csv	Qingboyuan Wang	Mon 3/07/22	Wed 3/09/22	3	0%	
2.5	Vis on Dataset #4 TBA: Will see how well the data transformation and cleaning is done.	Qingboyuan Wang	Thu 3/10/22	Sat 3/12/22	3	0%	
3	**Data Modeling**		Sat 3/12/22	Tue 3/29/22			
3.1	Creating a final copmrehensive dataset that can be used for modeling	Qingboyuan Wang	Sat 3/12/22	Mon 3/14/22	3	0%	
3.2	Cleaning /Transfoming	Jia Lyu	Tue 3/15/22	Wed 3/16/22	2	0%	
3.3	Create brief Vis to verify previous results	Keqin Liu	Thu 3/17/22	Fri 3/18/22	2	0%	
3.4	Choose a ML Model	Jia Lyu	Sat 3/19/22	Tue 3/22/22	4	0%	
3.5	Debugging and Refining the Model	Qingboyuan Wang	Wed 3/23/22	Sun 3/27/22	5	0%	
3.6	Deploy it to the comprehensive dataset	Qingboyuan Wang	Mon 3/28/22	Tue 3/29/22	2	0%	
4	**Recap/Report Write-Up**		Tue 3/29/22	Thu 4/14/22			
4.1	Code Review	Qingboyuan Wang	Tue 3/29/22	Wed 3/30/22	2	0%	
4.2	Create Report Draft	Jia Lyu	Thu 3/31/22	Fri 4/01/22	2	0%	
4.3	Working on Report	All Members	Sat 4/02/22	Wed 4/06/22	5	0%	
4.4	Working on Presentation Slides	Jia Lyu	Thu 4/07/22	Mon 4/11/22	5	0%	
4.5	Recap Team Meeting	Qingboyuan Wang	Tue 4/12/22	Tue 4/12/22	1	0%	
4.6	Rehearse for Final Presentation	Qingboyuan Wang	Wed 4/13/22	Thu 4/14/22	2	0%	
5	**Grace Period**		Thu 4/14/22	Sun 5/01/22			
5.1	Final Code/Report Peer Review	All Members	Thu 4/14/22	Wed 4/20/22	7	0%	
5.2	Fix anything that could gone wrong	All Members	Thu 4/21/22	Sat 4/30/22	10	0%	
5.3	SUBMIT/COMPLETED EVERYTHING	All Members	Sun 5/01/22	Sun 5/01/22	1	0%	