NETS 213 - Crowdsourcing and Human Computation
Fabrec Final Report

Nayeong Kim, Rose Kong, Shannie Cheng, Jina Lo

I. Project Overview

1. Introduction

Lots of people have a desire to try out a variety of styles, but sometimes, we don't have enough resource. So, we often ask our friends for second opinions, but this may not be enough. Therefore we present to you Fabrec, a crowdsourcing based fashion recommendation service. Fabrec is a business idea based on crowdsourcing that would get real-time opinions from crowd to show the top rated recommendation for a given piece of clothes - top, bottom, or dress.

There exists a similar clothes recommendation website called JustFab, which recommend clothes for customers based on a specific style they choose. JustFab uses machine learning algorithms to recommend clothes suited for customers of their choice of style, but these recommendations are not generated based on the specific user input. At Fabrec, we ask each of our customer to give a image of a piece of clothes that they want a recommendation for, and we then use crowdsourcing website Amazon Mechanical Turk to find the top 3 complementary pieces of clothes as recommendations. While Fabrec is customer-centric to give our customers best recommendations of piece of clothes, we have also conducted some analysis to identify the trend of fashion and to obtain meaningful insights.

Our service is following the pipeline shown in Figure 1, where we have 5 essential components to provide recommendations to our customers, including Data Collection, Receive Input, Aggregate Data, Quality Control & Ranking and Return Output.

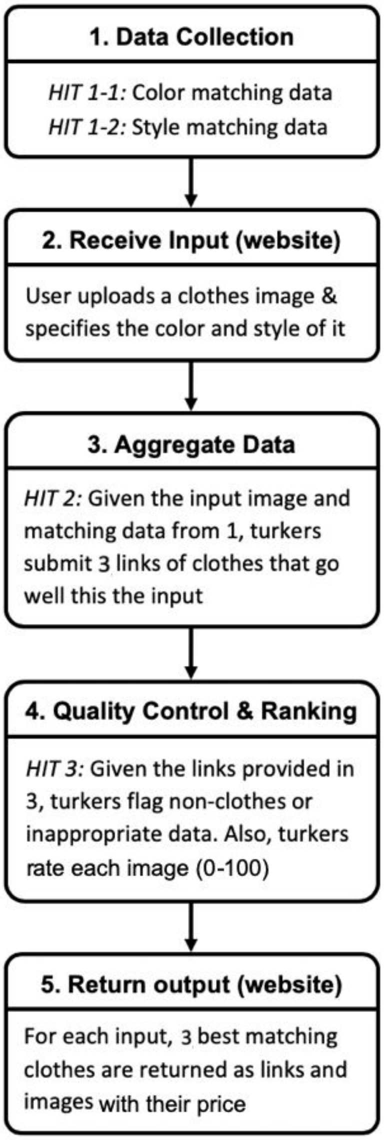

1. Data Collection
HIT 1-1: Color matching data
HIT 1-2: Style matching data
2. Receive Input (website)
User uploads a clothes image & specifies the color and style of it
3. Aggregate Data
HIT 2: Given the input image and matching data from 1, turkers submit 3 links of clothes that go well this the input
4. Quality Control & Ranking
HIT 3: Given the links provided in 3, turkers flag non-clothes or inappropriate data. Also, turkers rate each image (0-100)
5. Return output (website)
For each input, 3 best matching clothes are returned as links and images with their price

Figure 1: Fabrec Pipeline

2. Methodology

1) Data Collection (Mturk)

Data collection is essential for providing guidelines for recommendation. To receive inputs from a large majority of people, we used Amazon Mechanical Turk. This task is separated into 2 sections: HIT 1-1 (color matching) and HIT 1-2 (style matching).

1) HIT 1-1 (Color Matching): For each of the 18 frequently chosen colors, we asked 3 turkers to select 5 colors that go well with the that color. For each task, we provided a

$0.01 incentive. Then, we aggregated the results to find the top 3 color matchings for each colors. (Results: HIT1-1Processed.csv)

1) HIT 1-2: We have chosen 9 styles of tops, 18 styles of bottoms, shoes and dresses each that were frequently worn. Then, we created a HIT where given a certain style, turkers select 3 complementary pieces (top was matched with bottom, bottom was matched with top, and dress was matched with shoes). The incentive was $0.02 per task. Afterwards, we aggregated the results to find the top 3 complementary style matchings for each style. (Results: HIT1-2Processed.csv)

2) Receive Input (Website)

On our website (password: fabrec), customers were asked to provide the link of clothes they would like to receive a recommendation for, and specify the color and style of that piece (or the most similar one among the choices provided). The input links were collected and uploaded to HIT 2, explained in the next section.

3) Aggregation of Recommendations (Mturk Sandbox)

For each link we received from the previous section, we used the color and style input to find each of the top 3 color matchings and style matchings. Then, we uploaded the image of the input piece in HIT 2, along with the top 3 color matching and style matching data on Amazon Mechanical Turk Sandbox. Turkers had to submit 3 complementary clothes links from Amazon that followed one of the colors and one of the styles provided.

4) Quality Control & Ranking (Mturk Sandbox)

Aggregating the links provided in the previous section, we created HIT 3 for quality control and ranking on Mechanical Turk Sandbox. For each input item, we have collected 9 links (images) of matching items. We first sorted out non Amazon links. Then, In HIT3, given the input item and 9 images of recommendation, Turkers had rate each image in a range of 0-100 according to how well it went with the input item. For quality control, they were also told to report the image if it was inappropriate, wasn't the correct type of clothes, or was already provided the same image beforehand. Based on the total ratings of each image, we found the top 3 recommendations for each input item.

5) Return Output (Website)

After aggregation and quality control, we posted the input image along with the top 3 recommendations that the crowd provided in our website. The price was each item was also written, and if one clicked on the image, it was redirected to the Amazon link so the user could purchase the item any time.

II. The Crowd & Skills

Since we are a crowdsourcing business, we have used crowd workers in several components of our pipeline. Members of the crowd that participated in our project come from two sources: one is from NETS213 class and the other is Amazon Mechanical Turk. For (1) Data Collection, it was posted on Amazon Mechanical Turk with actual payment, and for (2) Receive Input, only the students from NETS213 class participated. (3) Aggregation of Recommendation and (4) Quality Control & Ranking were both posted in Mechanical Turk Sandbox for student and sandbox user participation. The number of people who participated in each HIT are as follows:

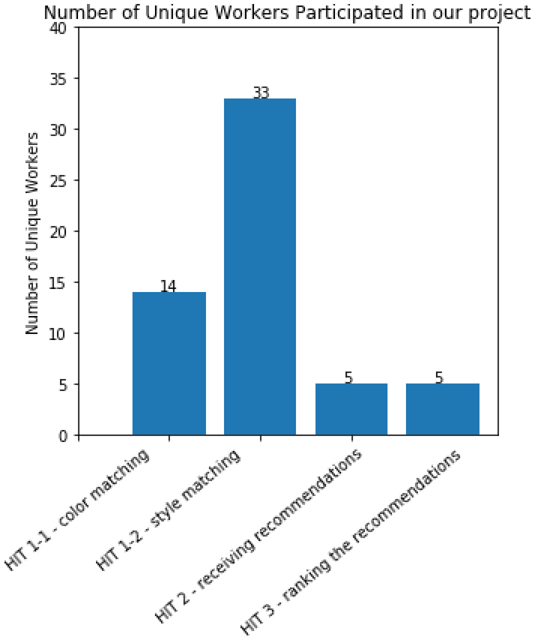


Figure 2: Number of Unique Workers who Participated in each HIT

In terms of skills that workers need in order to participate in our project, our crowd workers do not need specialized skills, since they are doing a basic matching based on their tastes of fashion. We would want crowds with high fashion sense, but since fashion is a subjective concept, we can't put a label to define skilled and unskilled workers. Also, since fashion is extremely difficult to quantify, we will say that all workers are equally qualified. But we do need our crowdworkers to be following the instruction, especially when they are asked to give recommendations based on the color matching and style matching result as reference.

III. Incentives

Among all the participants of our project, we have found that Turkers consist of majority of the participants; students only represent a small portion of the participant population. Upon analyzing participation of the crowd, we have found that students participated in our project have either submitted the item they wanted to get recommendation for or completed HITs for aggregation and quality control & ranking. Their incentive of participating in the project is mainly because of the Be the Crowd assignment where they were asked to complete at least 2 hours on helping other teams' projects.

On the other hand, Turkers who participated in our project, specifically for data collection, are mainly incentivized by the pay for each HIT they complete. We have paid them $0.01-0.02 per assignment. Since the HITs are relatively easy tasks that did not required any professional skills, the pay was sufficient to get enough workers to complete the tasks within a day.

Based on the input we received, we only had a limited number of customers who submitted an item they would like to get recommendations for. We think that this is because only the NETS 213 students were asked to submit the forms. If we were to expand our service to a wider range of customers, their main incentive to use our service is to get a better recommendation for what to wear with the item they already have.

IV. Information Received - what the crowd gives us

We can receive three types of information from our crowd, including color matching and style matching, items that would go well with the given input item, and rankings/evaluation for the suggested items. We have thought about automating part of our service, but we thought that (1) Data Collection, would have been impossible to be automated since they functioned as a baseline for the “good” recommendations. Yet giving recommendations would have been possible if enough data are collected for matching styles and colors, just like machine learning movie or TV show recommendation models used these days.

However, we then should have collected more labels other than color and style for each item - we have simplified the style to neckline for tops, for example, but even if two tops have same color and neckline, they could still be very different. This would have make the submission must trickier for customers since they would have to specify more information about their item. Also, this kind of automation would have required us to have a database of links to thousands of fashion items beforehand, along with their specific details. We think that such automation would be achievable if the items were limited to a single brand’s website. We could have used data base of clothes images instead, but they would have been less meaningful recommendation for customers because we cannot provide exact link where to get them.

Other than automating the service, we have also thought about integrating machine learning components into our project. But we realized that fashion is a subjective concept and we would probably not get enough data to perform machine learning on it.

V. Quality Control

To clean up our data and only receive valid responses, we have implemented quality control measures. For quality control, they were also told to report the image if it was inappropriate, wasn’t the correct type of clothes, or was already provided the same image beforehand. We have checked how many pictures are reported so differentiate the qualified and unqualified works. In addition, we have checked whether all the input links are Amazon links. We want to know if the workers are actually following the instructions of our HITs and how well they are following these instructions. Most of the recommendations are qualified.

VI. Aggregation

We submitted our color and style matches from previous HITs and received links for complementary pieces, and our aggregated results show the top 3 recommended pieces from MTurk. Analyses we did on the aggregated results were how well the quality control from the previous step worked and the adherence of the workers to the color-style guide. The aggregated responses were more helpful than individual responses, due to aggregated responses being ranked for quality.

From the result from each HIT, we have computed the average work time, which is the time it takes for each worker to complete the HIT (Figure 3). HIT1-1, HIT1-2 are just selecting good matches for a given color/style, therefore it would not take long. HIT2 involves giving an Amazon link to a clothes that fits a specific color/style, which would take the longest time since turkers have to do more research before submitting the answers. HIT3 would also take long time since each turker have to click on the Amazon link to view the content before ranking.

Since we gave turkers 17 choices of colors and 1 input color to choose the best matching color, we want to see the number of times each of color is selected. Since this data is pretty subjective, we did not have an expected result of the most frequently selected color. The top 3 selected colors were white, black, and dark blue, which are very universal colors. Black and white are the most popular choices since it would go well with other any colors. The color frequencies from figure 4 show that the most neutral colors were the most frequently selected, followed by more colorful or loud palettes.

Additionally, we want to see that to see what is the most popular style of top/bottom/dress from our data. Analyzing turkers' preference of different styles of bottoms given a top piece, top given a bottom piece, and shoes given a dress. We can see that the most popular top is the *gypsy* style, the most popular bottom is the *pegged* style and the most popular shoe style is *kitten*.

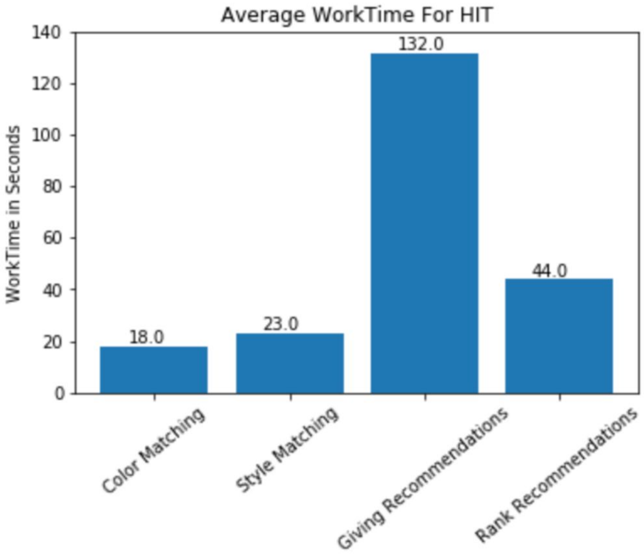


Figure 3: Average worktime for HIT 1-1, 1-2, 2, and 3.

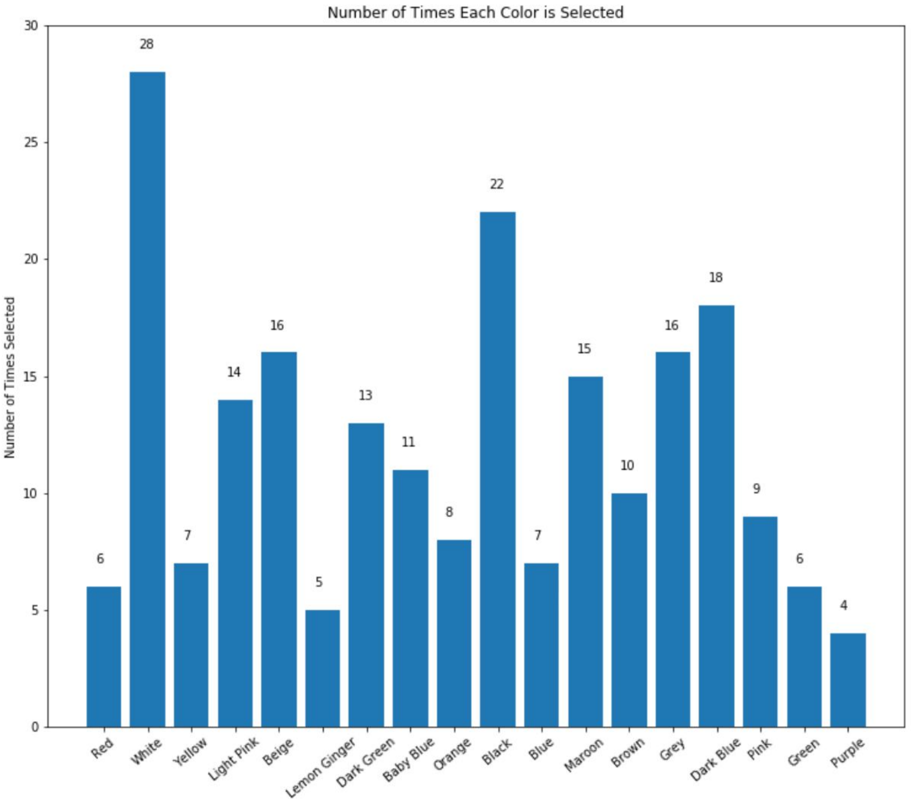


Figure 4: Number of times each color was selected for.

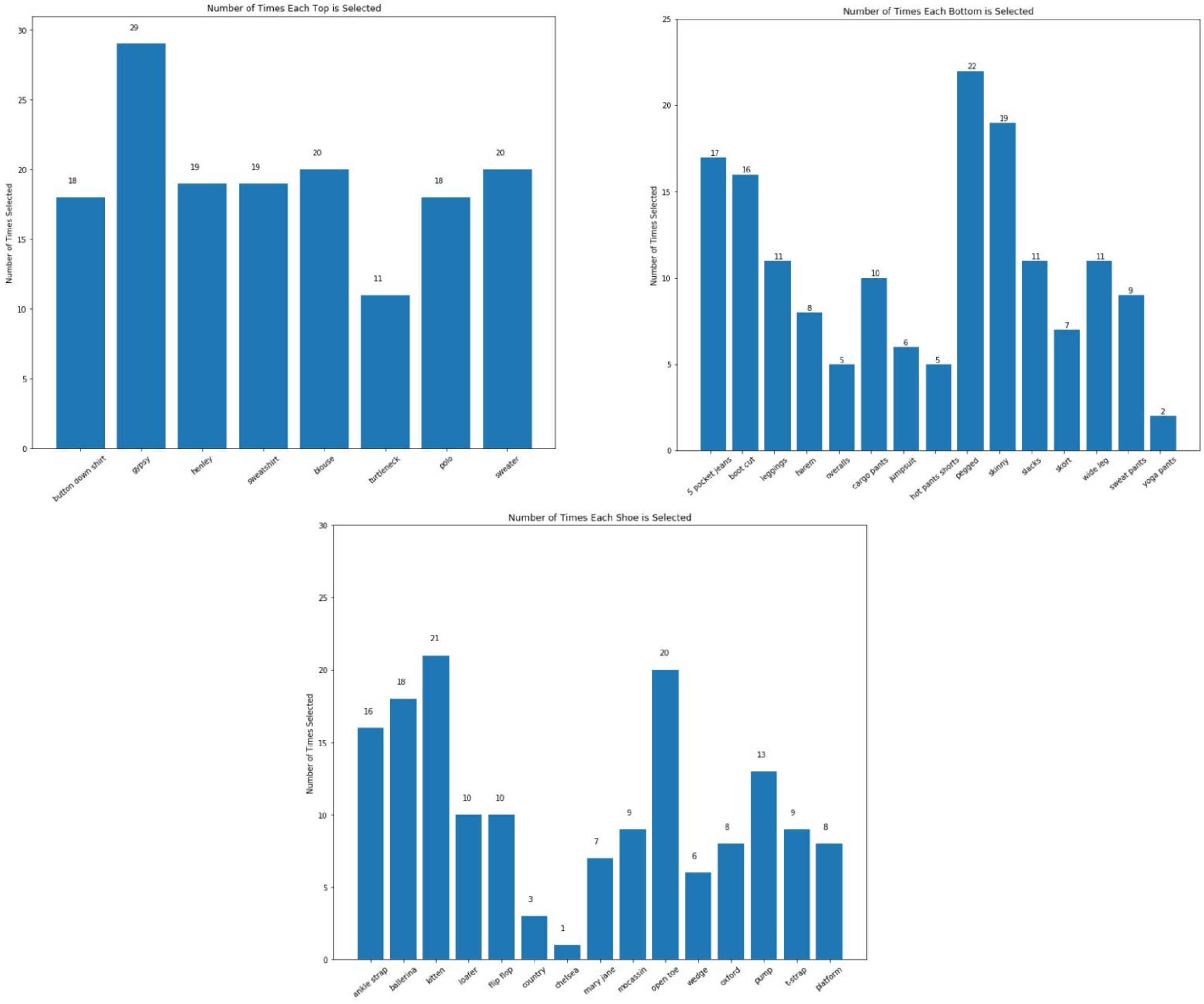
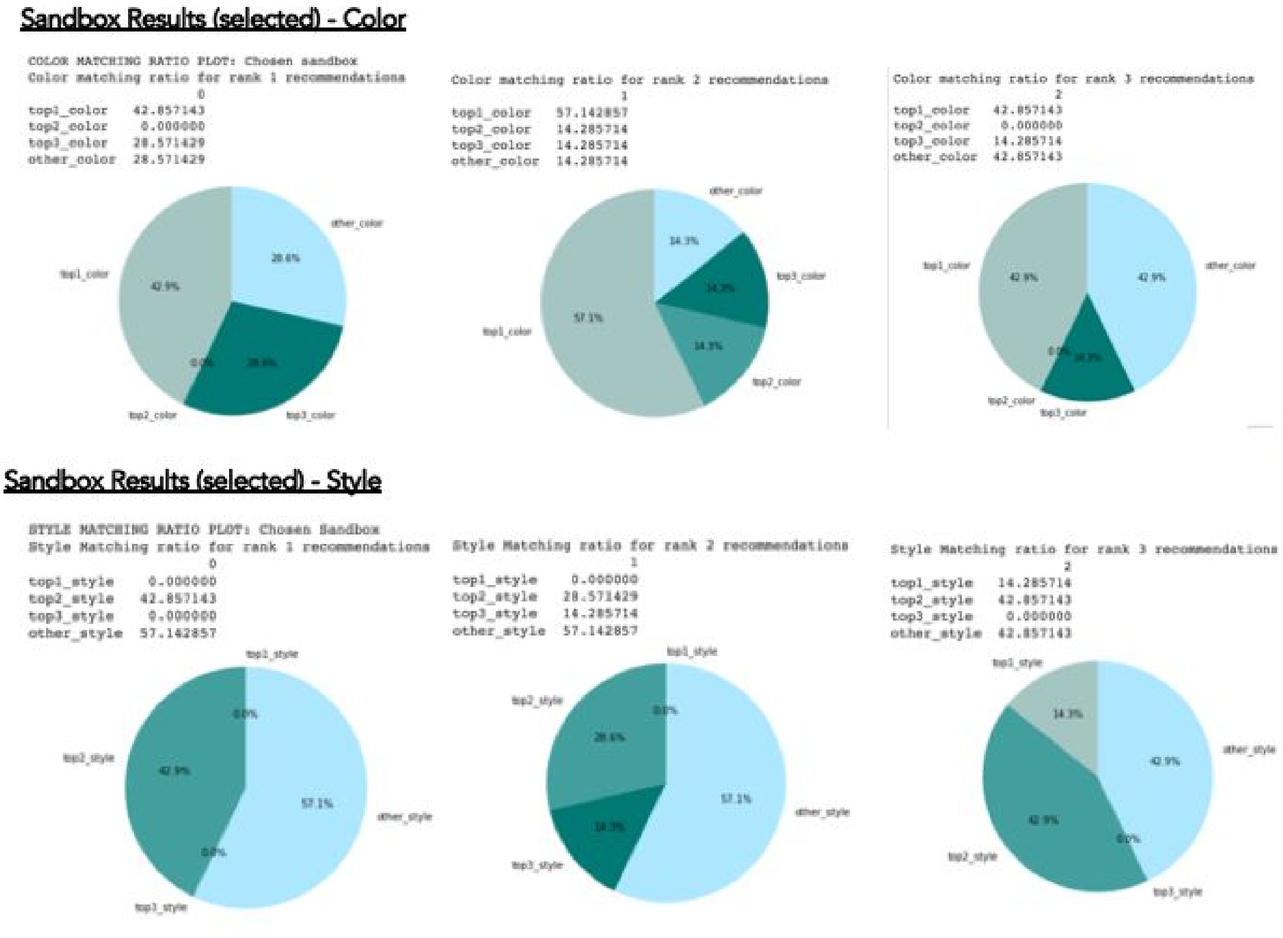


Figure 5, 6, 7. The frequency of selected styles of tops, bottoms and shoes.

From the data collected from HIT3, we analyzed how the chosen top 3 recommendations had colors and styles included in the top 3 colors and styles obtained from HIT1-1 and HIT 1-2. We can see that for colors, about 70% of top 3 recommendations combined had colors included in the top 3 choices; on the other hand for style, only 47% among top 3 recommendations had style included in the top 3 styles. This results implies that color matching information had greater consensus, while style matching seems to be more subjective. Unlike what we have expected, however, there was no significant difference between rank 1, 2, 3 recommendations, and often recommendations at lower ranks had higher adherence to the top 3 color and style information.


Sandbox Results (selected) - Color

```
COLOR MATCHING RATIO PLOT: Chosen sandbox
Color matching ratio for rank 1 recommendations
                        0
top1_color   42.857143
top2_color    0.000000
top3_color   28.571429
other_color  28.571429
```

```
Color matching ratio for rank 2 recommendations
                        1
top1_color   57.142857
top2_color   14.285714
top3_color   14.285714
other_color  14.285714
```

```
Color matching ratio for rank 3 recommendations
                        2
top1_color   42.857143
top2_color    0.000000
top3_color   14.285714
other_color  42.857143
```

Sandbox Results (selected) - Style

```
STYLE MATCHING RATIO PLOT: Chosen Sandbox
Style Matching ratio for rank 1 recommendations
                        0
top1_style    0.000000
top2_style   42.857143
top3_style    0.000000
other_style  57.142857
```

```
Style Matching ratio for rank 2 recommendations
                        1
top1_style    0.000000
top2_style   28.571429
top3_style   14.285714
other_style  57.142857
```

```
Style Matching ratio for rank 3 recommendations
                        2
top1_style   14.285714
top2_style   42.857143
top3_style    0.000000
other_style  42.857143
```

Figure 8: Colors and Styles of the Top 3 Sandbox recommendations

VII. Scaling Up

Fabrec is focused on giving individual fashion recommendations, and we don’t believe getting thousands of contributions would necessary benefit our product. Having more contributors may be beneficial to some extent, as different people have different preference of fashion styles, and the more input we have, the wider the range of recommendations could be. However, over a certain amount of people, the diversity of recommendation would not increase that much, since there would be a large amount of overlap of style and color.

However, we would like to increase the amount of recommendation links we receive per item (currently is 9). This would help us give better recommendations, although it would require more Mturk batches to rank the top 3 within these recommendations.

We’ve tried scaling up our project by uploading the same item for recommendation in Mechanical Turk Sandbox (small crowd - mostly classmates) and Mechanical Turk (larger crowd). We compared the diversity of styles and colors, and average work time.

The first analysis was done to compare the colors Mturk Sandbox and Mturk workers have chose, while the top 3 color matching to the input was given.

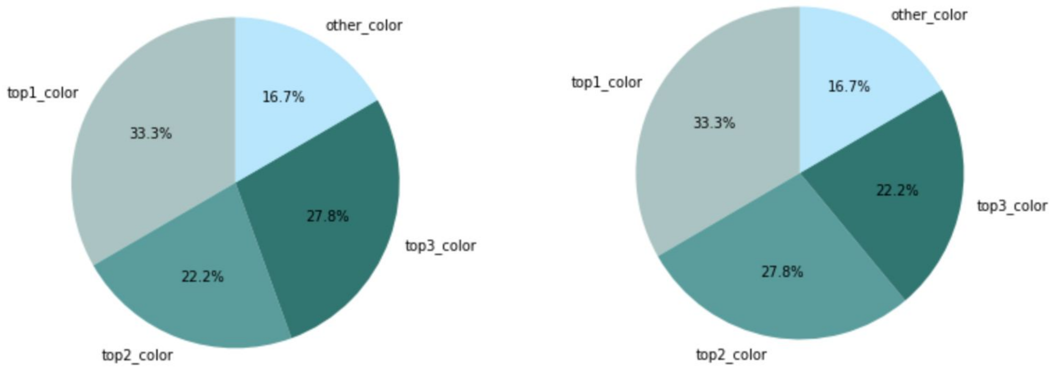
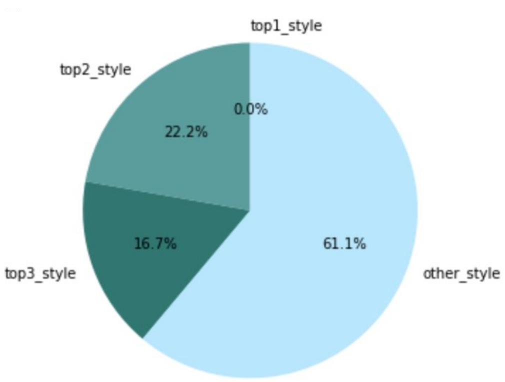


Figure 9: Colors of the recommendations from Mturk Sandbox (left) and Mturk (right)

Although there were some proportion of workers that chose colors that were not the 3 color matchings given, the overall distribution of color choices the workers have done were very similar.

The second analysis compares the styles Mturk Sandbox and Mturk workers have chose, while the top 3 style matching was given.



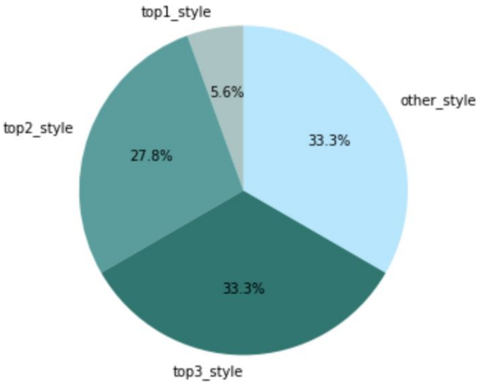


Figure 10: Styles of the recommendations from Mturk Sandbox (left) and Mturk (right)

We discovered that a lot of workers have selected a recommendation that had a style that wasn't given (didn't follow). However, compared to Sandbox workers, actual Mturk workers followed the instructions better and had more diversity of style.

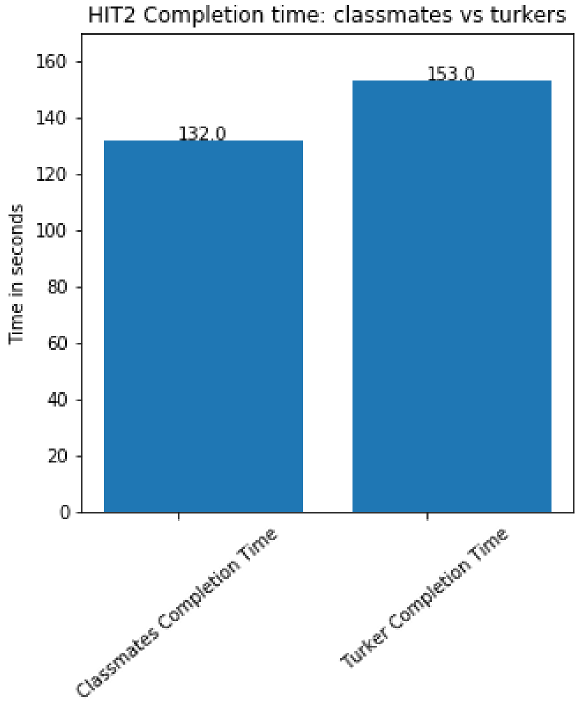


Figure 11: HIT 2 Average Completion Time of Mturk Sandbox and Mturk

It took a slightly longer time for workers in MTurk to complete our task. However, the difference wasn't that significant. However, this result may be because they are professional workers and they pay more attention to the instructions, and spend more time to complete the task.

IX. Project Analysis

Overall, our project did succeed and we were able to receive quality recommendations. From the results we obtained from HIT3, we have received our top recommendations for each piece. We can only analyze our results qualitatively, as how well our recommendations actually suited our clothing pieces. From looking at our final batches of data, we can confirm (as 4 authorities of style, obviously) that the crowd workers have chosen great outfits, albeit safe choices.

For more detailed understanding of our performance, we compared the top 3 sandbox recommendations from non-top 3 sandbox recommendations. From Figure 12, we can see that on average, about 70% and 47% of top 3 chosen recommendations had colors and styles included in the top 3 colors and styles recommendations, respectively. Compared to that result, as demonstrated in Figure 13, only 27 % of non-selected recommendations had top 3 colors and only 19% of them had styles included in the top 3 recommendations. From these results, we can conclude that the chosen top 3 recommendations for items better adhered to guideline, or consensus, of what colors or styles go well together when paired together, which signifies that the chosen top 3 recommendations are actually "good" ones in terms of the crowd's judgements.

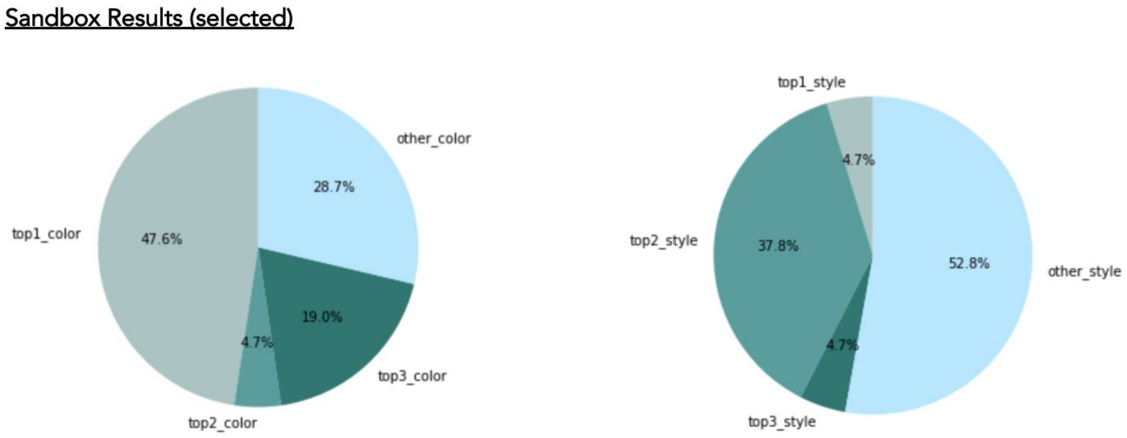


Figure 12: Adherence to Top 3 Colors(left) and Styles (Right) for Top 3 Recommendations

Sandbox Results (not selected)

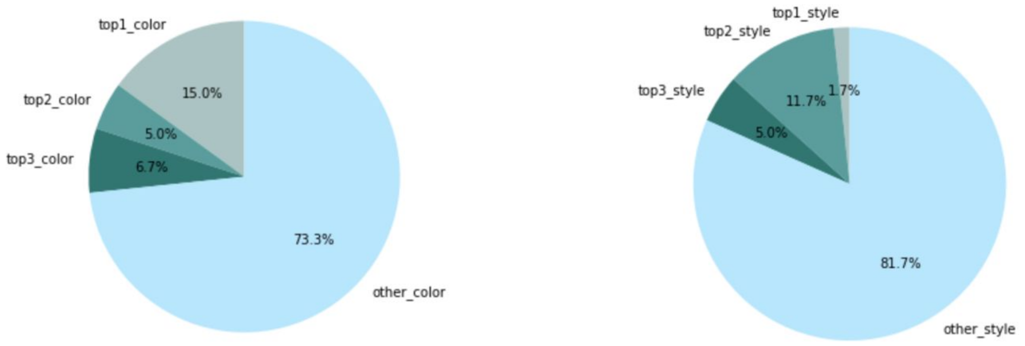


Figure 13: Adherence to Top 3 Colors(left) and Styles (Right) for non-Top3 Recommendations

X. Technical Challenges and Limitations

Between the initial proposal for the project and the final product, there was no change or adjustment made. Yet there were some limitations that leave a room for future improvements. The main limitation of our project was the amount of data that we could accumulate due to the class size and amount of money we could spend to use MTurk. This might have caused less accurate analysis, as one or two outliers might have influenced the analysis results significantly. With more data, we would like to identify the reason why MTurk responses and sandbox responses differ significantly only in the style matching ratio.

Also, in the future, we could improve on the labels for styles. We have minimized the number of styles and colors used, which made some categorizations for items rather ambiguous. For example of tops, instead of including mainly necklines, we can divide up the categories into sleeves, necklines and length to make them more specific and accurate. On the other hand, in case of shoes, there were too many categories that were unfamiliar to the crowd, which made it harder for workers to come up with an item that corresponds to each category. For the future direction for analysis, it would be also meaningful to upload a HIT without color or style matching guidelines to see whether top 3 recommendations actually follow those matchings, which will provide us information about the best matching combinations of fashion items.

The biggest challenges we had were manual uploading and sequential HITs. We had to update all HITs manually by downloading results, downloading images, and uploading AWS, then sending it to MTurk. Additionally, we needed to process the data of one HIT before the next, sequentially due to the results being used in the next HIT.

There were tremendous substantial technical components to the project. However, there was a technical difficulty we ran into - since we had many HITs that were sequential, we wanted to write a script so that all the data (links, images, etc) from the previous HIT would be aggregated and "automatically" uploaded to a new HIT.

With respect to the technical difficulty, we decided that the amount of data we had was small enough to be manually downloaded and put into a .csv or on AWS rather than writing a script. Things such as image results from HIT2 were downloaded from their respective website links, uploaded to AWS, put into a .csv file, and uploaded to HIT3. In the future, to further our project, we must write a script that can automate some of our in-between processes.