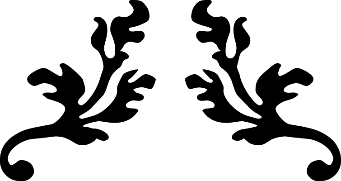
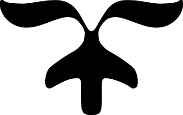
|  |  |
| --- | --- |
| Cairo University Faculty of Engineering Computer Department |  |



BIG DATA AND ANALYTICS PROJECT

|  |  |
| --- | --- |
| **Students’ Names:**  Mohamed Osama Hussein | 1142032 |
| Nourhan ElSayed Attya | 1142061 |
| Salma Ibrahim Ahmed | 1144055 |
| Yousra Hussein Ibrahim | 1142072 |

**Submitted to:** Eng. Hussein Fadl

APRIL 22, 2019

Table of Contents

1. [Introduction 3](#_bookmark0)
2. [Problem Description 3](#_bookmark1)
3. [Project Pipeline 3](#_bookmark2)
4. [Analysis and solution of the problem 5](#_bookmark4)
   1. [Data Sources 5](#_bookmark5)
   2. [Data preprocessing 6](#_bookmark7)
   3. [Data Visualization 8](#_bookmark11)
   4. [Model Building 12](#_bookmark20)
      1. [Segmentation Model 12](#_bookmark21)
      2. [Association Rules 13](#_bookmark22)
5. [Unsuccessful Trials 15](#_bookmark24)
6. [Results and Evaluation 17](#_bookmark27)
   1. [FCN Model Accuracy 17](#_bookmark28)
   2. [Analysis of Colored Pieces Matching Results 17](#_bookmark29)
   3. [Analysis of Trending Colors Matching Results 22](#_bookmark35)
   4. [Analysis of Trending Colors According to Skin Color 27](#_bookmark43)
7. [Enhancements and Future Work 28](#_bookmark46)
8. [References 29](#_bookmark47)

Table of Figures

[Figure 1 - Project Pipeline 3](#_bookmark3)

[Figure 2 - CFPD and CCP dataset Samples 5](#_bookmark6)

[Figure 3 – The Input Image 6](#_bookmark8)

[Figure 4 – Segmented Pieces 6](#_bookmark9)

[Figure 5 - Images in the CSV File 7](#_bookmark10)

[Figure 6 - CSV Files Sample 8](#_bookmark12)

[Figure 7 - Most frequent colors found in the dataset 8](#_bookmark13)

[Figure 8 - Most frequent color-piece pairs found in the dataset 9](#_bookmark14)

[Figure 9 - Most frequent colors in jackets 10](#_bookmark15)

[Figure 10 - Most frequent colors in tops 10](#_bookmark16)

[Figure 11 - Most frequent colors in bags 11](#_bookmark17)

[Figure 12 - Most frequent colors in shoes 11](#_bookmark18)

[Figure 13- Most frequent colors in sweaters 12](#_bookmark19)

[Figure 14 - Comparison between Apriori, Eclat and FP-Growth Algorithms 13](#_bookmark23)

[Figure 15 - Unsuccessful Matching Rules for Colors 15](#_bookmark25)

[Figure 16 - Rules Generated with High Support 16](#_bookmark26)

[Figure 17 - Color-Piece Matching Rules 17](#_bookmark30)

Figure 18 - Lift of Matching Rules 18

Figure 19 - RPF of Matching Rules 18

[Figure 20 - Conviction of Matching Rules 19](#_bookmark31)

[Figure 21 - Scatter plot 19](#_bookmark32)

[Figure 22 - Grouped Matrix 20](#_bookmark33)

[Figure 23 – Matching Rules as Graphs 21](#_bookmark34)

[Figure 24 - Trending Colors Matching Rules 22](#_bookmark36)

[Figure 25 - Lift of Colors Matching Rules 23](#_bookmark37)

[Figure 26 - RPF of Colors Matching Rules 24](#_bookmark38)

[Figure 27 - Conviction of Colors Matching Rules 24](#_bookmark39)

[Figure 28 - Scatter Plot of Colors Matching Rules 25](#_bookmark40)

[Figure 29 - Matrix Plot of Colors Matching Ru 25](#_bookmark41)

[Figure 30 - Colors Matching Rules as Graphs 26](#_bookmark42)

[Figure 31 - Blonde People Color Trends 27](#_bookmark44)

[Figure 32 - Brunette People Color Trends 27](#_bookmark45)

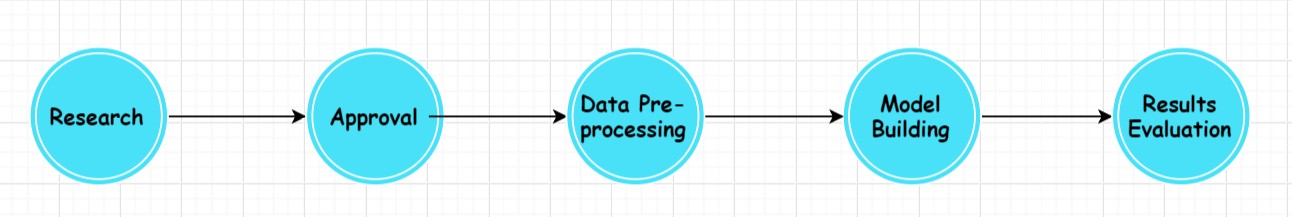
# Introduction

This document provides an overview about a Clothes’ Matcher System that help in having insights about the fashion trends in clothes and what pieces match with each other. Throughout the document, we will discuss the project pipeline, the analysis of the problem and the proposed solution, the results we have and their evaluation, the unsuccessful trials that we faced and our future work on the problem.

# Problem Description

Shopping nowadays is a part of every person’s life, but somehow not every person knows how to match his or her outfits according to the fashion trends. Fashion trends keep changing nearly all the time. When a girl wants to buy new clothes, she needs to know what the current trends are and how to match her outfits together. Instead of surfing Instagram pages and fashion websites, she just needs a program that tells her what the fashion trends are now and how to match the pieces together.

# Project Pipeline



*Figure 1 - Project Pipeline*

* 1. **Research:** The first phase of the project was to conduct a research to choose an idea that best fits our scope and is innovative at the same time. We decided to do a Clothes’ Matcher system and we searched for datasets that are related to fashion and clothing in order to apply our model to. We found many datasets that can be used in the project; we examined them carefully and decided to use two of them together. We will talk more about these two datasets later in this document.

(**Time Spent:** 1 week)

* 1. **Approval:** we made the project proposal and submitted it for approval. This phase ended once we had received the e-mail of approval.
  2. **Data Pre-Processing:** In this phase, we prepared the two datasets to use in the model; we also extracted features from the clothing items that would help in the matching process, for example: the color of each piece. This phase will be discussed in detail in a separate section later in the document. After the pre- processing of each dataset on its own, we merged the two ready datasets together in order to have larger amount of data to do the analysis on. (**Time Spent:** 4 days)
  3. **Model Building:** we started brainstorming the different analytical techniques that we can use and which will help us achieve our objectives. We decided to use “Association Rules Mining”. We searched for the different algorithms to implement it. We compared between the algorithms and chose “Eclat” algorithm. This phase will also be discussed in detail in a separate section.

(**Time Spent:** 3 days)

* 1. **Results Evaluation:** This was the final phase in which we have set some criteria upon which we can evaluate the performance of the model. We analyzed the results and calculated the measures to finally produce a model that can find the fashion trends in the datasets and can match the different clothing pieces together. (**Time Spent:** 2 days)

# Analysis and solution of the problem

## Data Sources

Our data was collected from two datasets of images. Each one of them has a pixel- annotated file of categories for each image.

1. [Colorful Fashion Parsing Data (CFPD)](https://drive.google.com/file/d/0BwIcx4kBjPrcTzlQYi1VbmthVVU/edit?usp=sharing) Dataset

<https://sites.google.com/site/fashionparsing/dataset>

1. Clothing Co-Parsing (CCP) Dataset

[https://github.com/bearpaw/clothing-co-](https://github.com/bearpaw/clothing-co-parsing/blob/master/README.md?fbclid=IwAR04Waf5ACIOdGdcw_kh62CsXhHQ0hJK2aunCPcvh_9MSgX1bLXvJYA5QN8) [parsing/blob/master/README.md?fbclid=IwAR04Waf5ACIOdGdcw\_kh62CsXhHQ0hJK2aunCPcvh\_9MS](https://github.com/bearpaw/clothing-co-parsing/blob/master/README.md?fbclid=IwAR04Waf5ACIOdGdcw_kh62CsXhHQ0hJK2aunCPcvh_9MSgX1bLXvJYA5QN8) [gX1bLXvJYA5QN8](https://github.com/bearpaw/clothing-co-parsing/blob/master/README.md?fbclid=IwAR04Waf5ACIOdGdcw_kh62CsXhHQ0hJK2aunCPcvh_9MSgX1bLXvJYA5QN8)

* + - CFPD has 23 categories of clothes, shoes, bags and accessories.
    - CCP has 59 categories of clothes, shoes, bags and accessories.



*Figure 2 - CFPD and CCP dataset Samples*

## Data preprocessing

**Step 1:** As mentioned in the previous section, the data sets are originally images. The first step in the pre-processing phase was segmenting each image into pieces. For each piece, we extracted the colors in that piece. We used HSL color model where we defined the ranges of the Hue, Saturation and Lightness manually in order to produce 11 main colors: Black, White, Gray, Red, Orange, Yellow, Green, Cyan, Blue, Purple, and Pink. The following figures show an input image and how it is segmented into pieces.

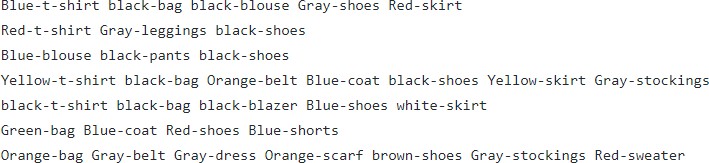


*Figure 3 – The Input Image*



*Figure 4 – Segmented Pieces*

**Step 2:** For each image in the data set, we have put a row in the csv file corresponding to this image where it includes all the pieces in that image, each piece with its colors. The following figure shows some of the rows in the csv file.

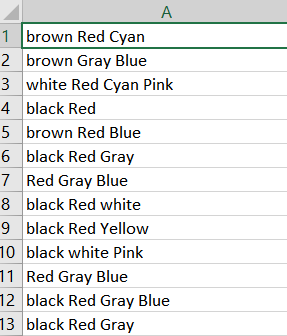
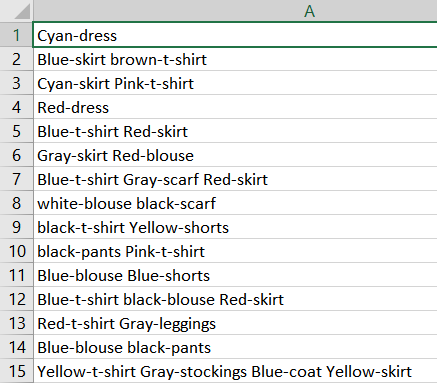


*Figure 5 - Images in the CSV File*

**Step 3:** The last step in the pre-processing is merging the CSV files of the two datasets to produce a CSV file of 3686 rows corresponding to 3686 images. This CSV file is used for generating the matching rules in R Studio. An important thing to note is that the CFPD dataset already contained the colors of pieces but we did not use those colors and extracted the colors from the images ourselves in order to have the same list of colors for the two datasets, so that we can unify them into one dataset.

## Data Visualization

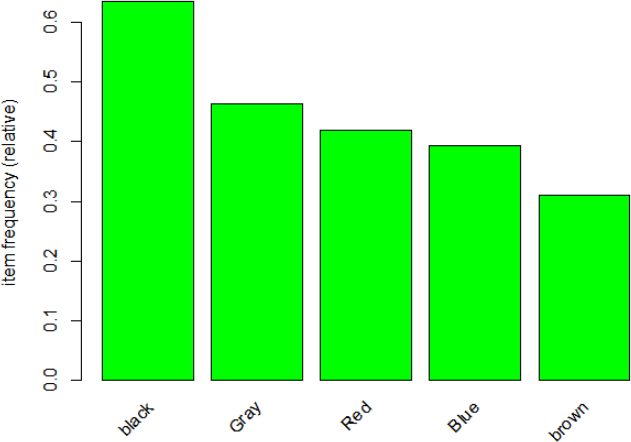
As mentioned before, each row represents one image data. The following figures show the input data to R Studio.



*Figure 6 - CSV Files Sample*

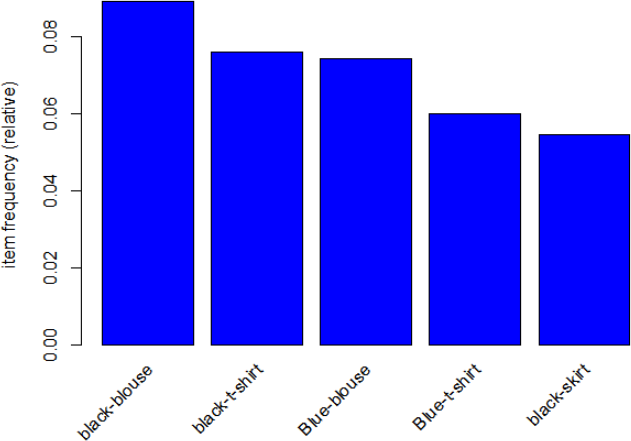
We used the CSV files in R to visualize the data and get:

1. The most frequent colors found in the dataset (Trending Colors)



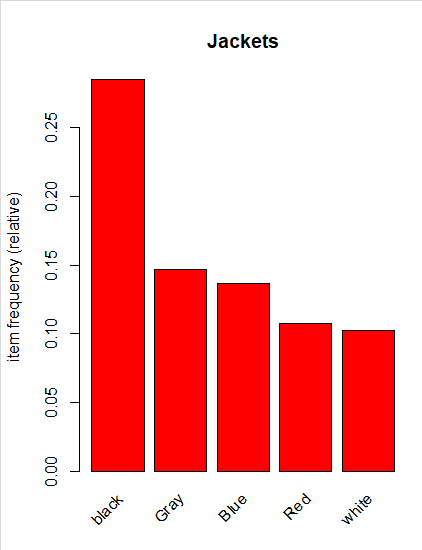
*Figure 7 - Most frequent colors found in the dataset*

1. The most frequent color-piece combinations found in the dataset

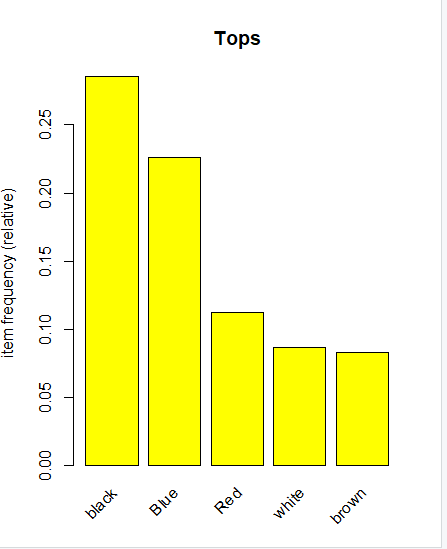


*Figure 8 - Most frequent color-piece pairs found in the dataset*

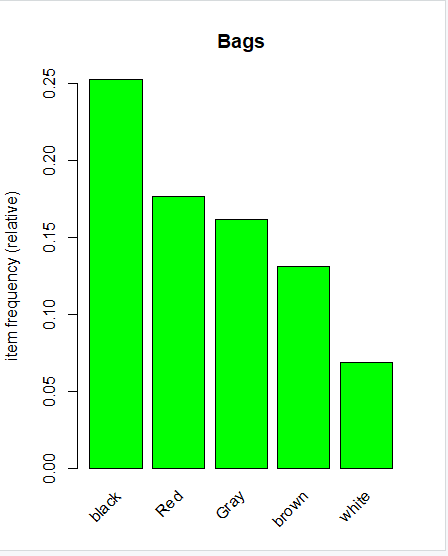
1. The most frequent colors per each category (Trending colors per Category)
   * Trending colors in Jackets



*Figure 9 - Most frequent colors in jackets*

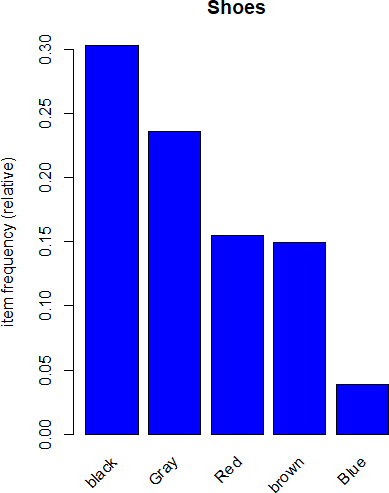
* + Trending colors in Tops

*Figure 10 - Most frequent colors in tops*

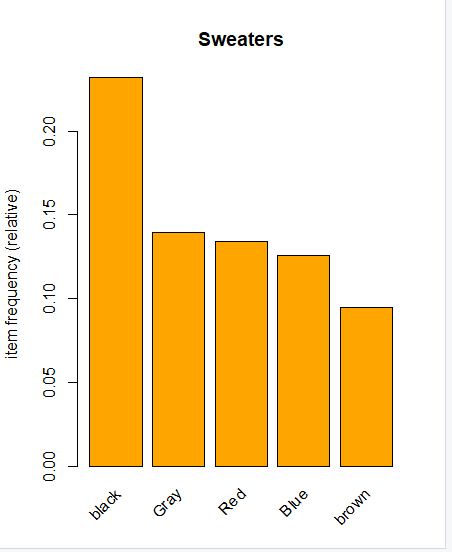
* + Trending colors in Bags

*Figure 11 - Most frequent colors in bags*

* + Trending colors in Shoes



*Figure 12 - Most frequent colors in shoes*

* + Trending colors in Sweaters

*Figure 13- Most frequent colors in sweaters*

## Model Building

### Segmentation Model

We have two approaches in order to extract the pieces from the images in order to get their colors and categories. The first one was building an FCN (Fully Convolutional Network) Model. This FCN segments the image into pieces and give each piece a category label (blouse, pants, etc.). The second approach was using the Pixel-Annotated files of the images in order to identify different pieces from it. Using these files, we can segment the image piece by piece based on the annotations; and for each piece, we extract its colors. We used the second approach to generate the CSV files as described in the pre-processing section. The reason why we used the Pixel-Annotated files is that they contain the ground-truth values for the categories (blouse, pants, etc.), so we are certain that the categories will be correct so we can get meaningful analysis. The FCN approach to get the pieces categories resulted in having 99% accuracy on the training set and 90% on the test set. It will be used in the future work to segment images from the websites and Instagram crawling of fashion bloggers profiles.

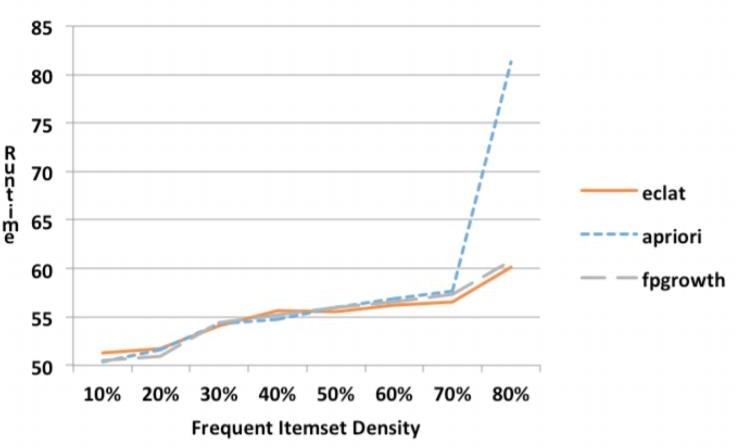
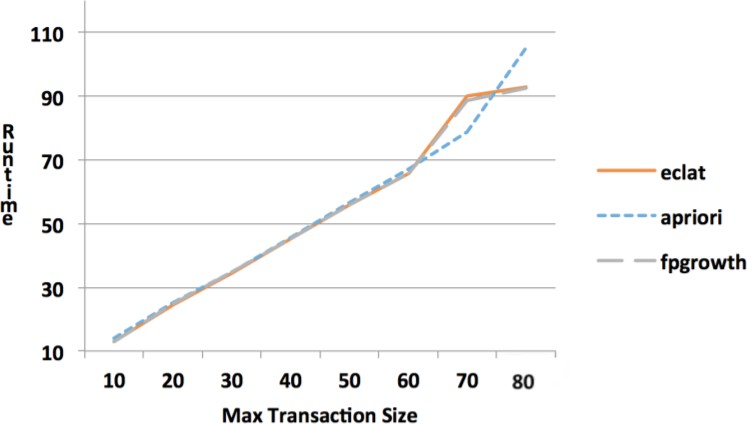
### Association Rules

We used the association rules mining in order to get the matching rules for the categories along with the colors, for example (black shirt matches with blue jeans and white sneakers). We also used it to get the trend in the matching of colors from the dataset, which means which colors match with each other; as an example (blue matches with black and white all together).

#### Algorithms Used for Generating the Rules:

* + - * 1. “Apriori” Algorithm
        2. “Eclat” Algorithm

Both algorithms result in the same matching rules. They only differ in the runtime. Apriori is an easily understandable frequent item set mining algorithm. Because of this, Apriori is a popular starting point for frequent item set study. However, Apriori has serious scalability issues and exhausts available memory much faster than Eclat and FP- Growth algorithms. Because of this, Apriori should not be used for large datasets. Most frequent item set applications should consider using either FP-Growth or Eclat. The following figure shows a comparison between the three algorithms regarding the runtime of each when the dataset’s size increases.



*Figure 14 - Comparison between Apriori, Eclat and FP-Growth Algorithms*

#### Metrics Used in Generating the Rules

1. **Support:** Is an indication of how frequently the item set appears in the dataset. Rule: NBoth/NTotal
2. **Confidence:** Is an indication of how often the rule has been found to be true. Rule: NBoth/NLeft
3. **Lift:** The ratio of the observed support to that expected if X and Y were [independent](https://en.wikipedia.org/wiki/Independence_(probability_theory))



1. **Conviction:** A high conviction value means that the consequent is highly depending on the antecedent. It can be used to ensure the results of the lift



1. **Rule Power Factor (RPF):** Rule Power Factor is an indication of how intense a rule’s items are associated with each other in terms of positive relationship. Rule Power Factor is defined as:

**=**



* A case that shows the importance of RPF:
  1. If “item A” appeared in 20 transactions and B in 50 out of total 100 transactions and item A and B both together appear 15 transactions. Then conf(A->B) = 0.15/0.2=0.75 = 75%.
  2. If “item A” appeared in 30 transactions and B in 60 out of total 100 transactions and item A and B both together appear 20 transactions. Then conf=0.2/0.3=0.66 = 66%.

However, in case (b), both antecedent and consequent item’s occurrences increased individually and also increased in association (both A and B items together purchased) occurrences. While interest measure confidence says surprisingly that case (a) is more important (75%) than case (b) 66%.

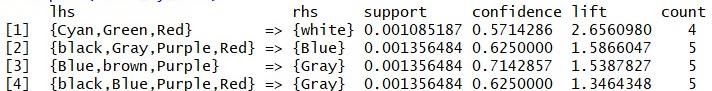
If we take the help of Rule Power Factor (RPF): (a) 0.75\*0.15= 0.11

(b) 0.66\*0.2= 0.13

RPF, correctly judge that case (b) is more important.

# Unsuccessful Trials

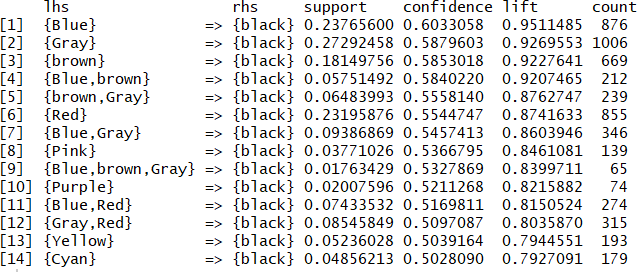
1. As we are extracting the colors found in each piece and adding them in the CSV files, we first tried to include all the colors per piece. However, since each piece can contain many different colors with different percentages for each color, this makes the rules very random and containing many colors that could not be worn together if these percentages increase. For example: if a jumper is blue and contains red, yellow and green stripes, this will make the red, yellow and green percentages small in comparison with the blue percentage, and thus this can match with blue pants. However, if the red, yellow and green colors increase, the above matching rule will no longer have a meaning and it will result in wearing things that cannot match. The following figure shows the unsuccessful matching rules for trending colors (which colors match together).



*Figure 15 - Unsuccessful Matching Rules for Colors*

**Solution:** To avoid this problem, we considered only the dominant color for each piece (the one with the highest percentage); we no longer took all the colors that are found in a piece.

1. Using a high support in the “Apriori” or “Eclat” algorithms results in almost all the rules implying to the color that is found in most of the data (which is the black color), and as a result of this, these rules have a lift which is less than 1 and this makes sense. As these rules are really generated by coincidence because of having the black color very frequent in the dataset. The following figure shows some of the rules generated when using high support



*Figure 16 - Rules Generated with High Support*

**Solution:** Decreasing the support.

1. Using the dataset with common pieces that are found in each image/outfit makes them appear in all the rules. For example: bags category is found in all images so it appears in all the rules and doesn’t give the chance to the clothing pieces that should actually appear to appear in the rules which makes the rules meaningless because they do not contain actual clothing pieces. This problem arises because the dataset has for each type of clothes almost five categories; jackets for example are divided into: blazers, capes, coats, cardigans, vests and jackets, so the number of times each of these categories appear is far less than the number of times “bags” category appear, that is why bags and shoes overtake almost all rules.

**Solution:** Removing the bags, shoes and accessories from the dataset leaving only the actual clothing pieces to apply matching rules on.

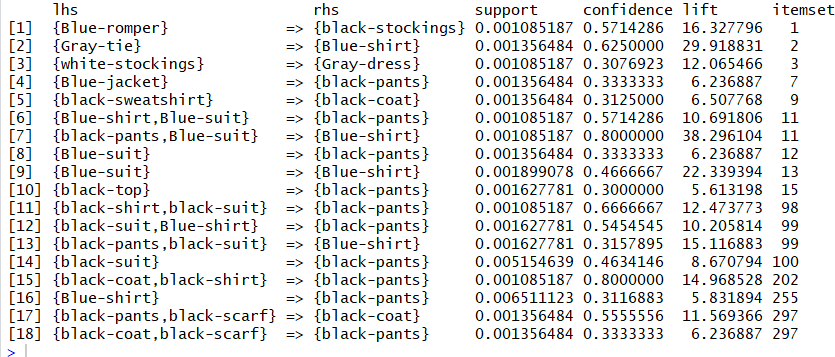
# Results and Evaluation:

## FCN Model Accuracy:

FCN resulted in 99 percent accuracy on the training set and 90 percent accuracy on the test set.

## Analysis of Colored Pieces Matching Results:

1. Rules generated using Eclat Algorithm with minimum support = 0.001 and minimum confidence = 0.3

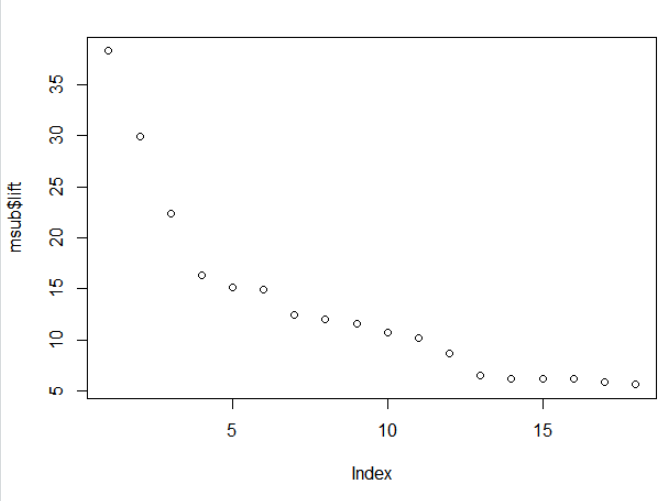


*Figure 17 - Color-Piece Matching Rules*

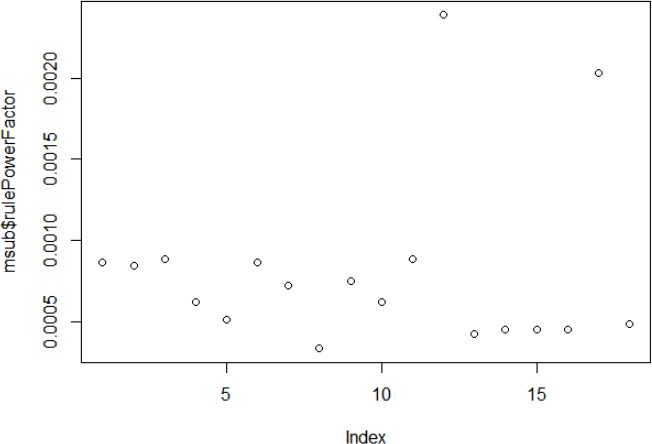
* + - Note that no rules are generated with lift near to one (rules by coincidence). That is after removing the bags, shoes and other similar items that were always appearing in data causing coincident rules.

1. Plotting rules with Lift, RPF and Conviction measures:

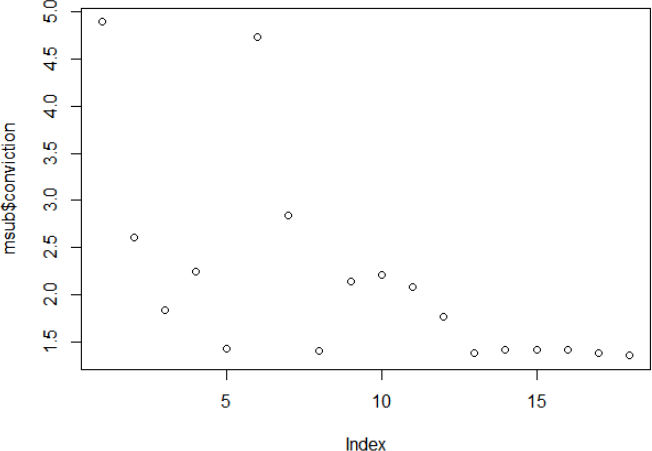
Note that: 1) all the Lift values are greater than one indicating that the rules did not come by coincidence. 2) The higher the RPF, the higher the positive relationship between the consequent and the antecedent. 3) Conviction values greater than one indicates that the rule did not come by coincidence and that the consequent is highly depending on the antecedent



*Figure 18 - Lift of Matching Rules*

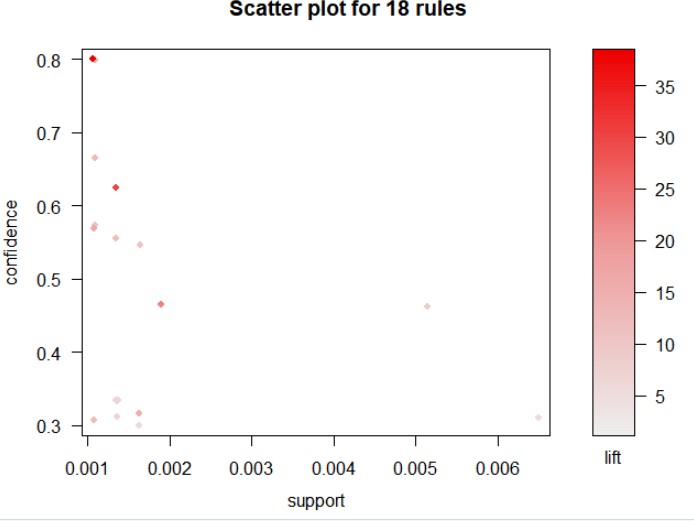


*Figure 19 - RPF of Matching Rules*



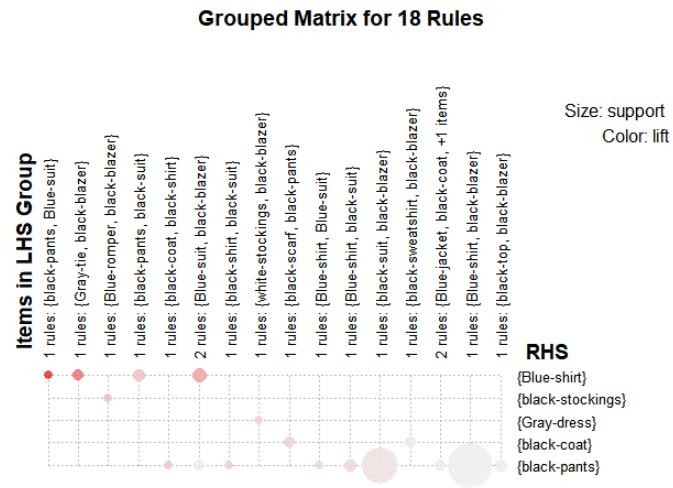
*Figure 20 - Conviction of Matching Rules*

1. Scatter Plot showing the Support, Confidence and Lift of the Rules



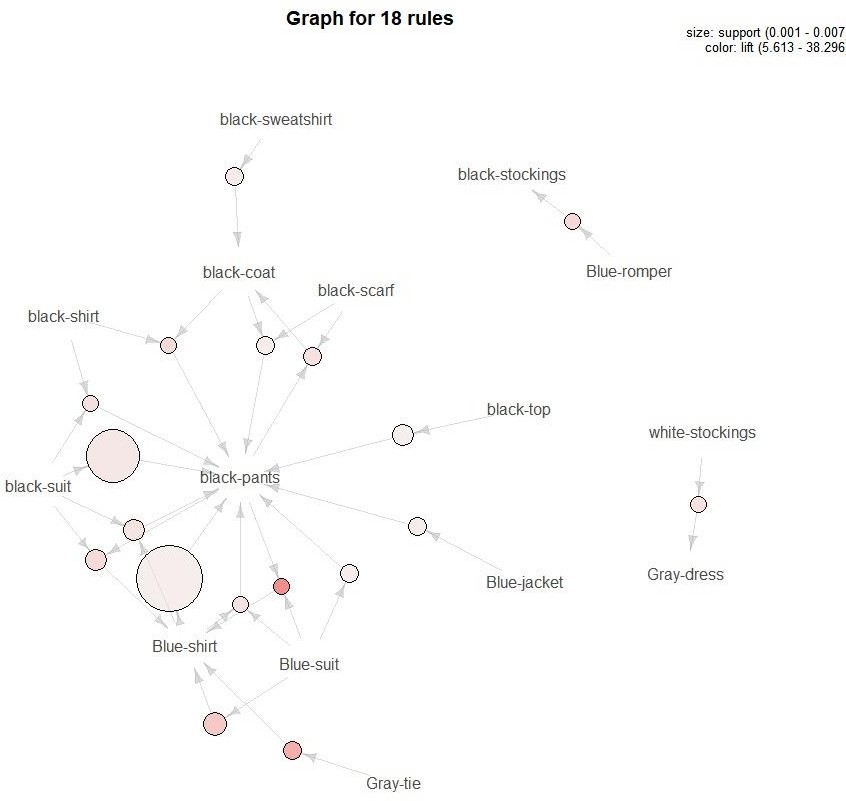
*Figure 21 - Scatter plot*

1. Grouped Matrix plot which shows the rules themselves with the lift and support to make it more clear



*Figure 22 - Grouped Matrix*

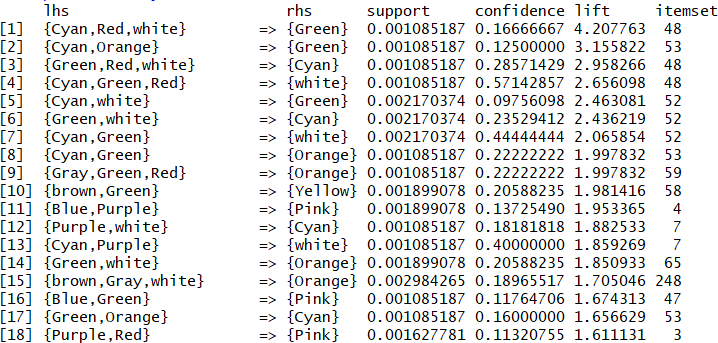
1. Graph for the color-piece matching generated rules



*Figure 23 – Matching Rules as Graphs*

## Analysis of Trending Colors Matching Results:

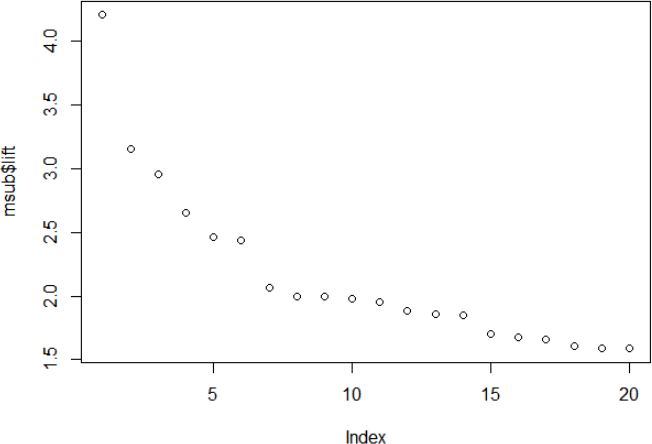
* + 1. Some of the trending colors matching rules that are having lift more than 1.5



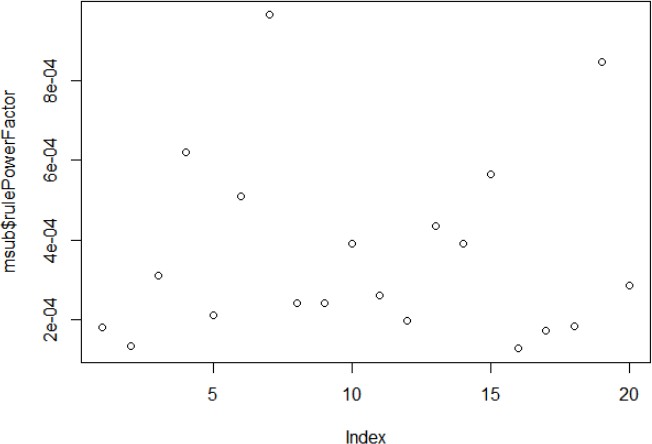
*Figure 24 - Trending Colors Matching Rules*

* + 1. Plotting rules with Lift, RPF and Conviction measures:

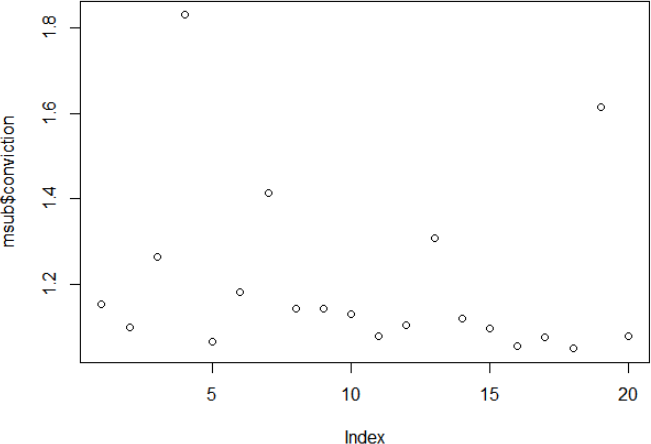
Note that: 1) all the Lift values are greater than one. 2) The higher the RPF value, the better the rule. 3) We can use Conviction value as a threshold such that we take rules that has high lift value but also has a conviction value greater than one.



*Figure 25 - Lift of Colors Matching Rules*

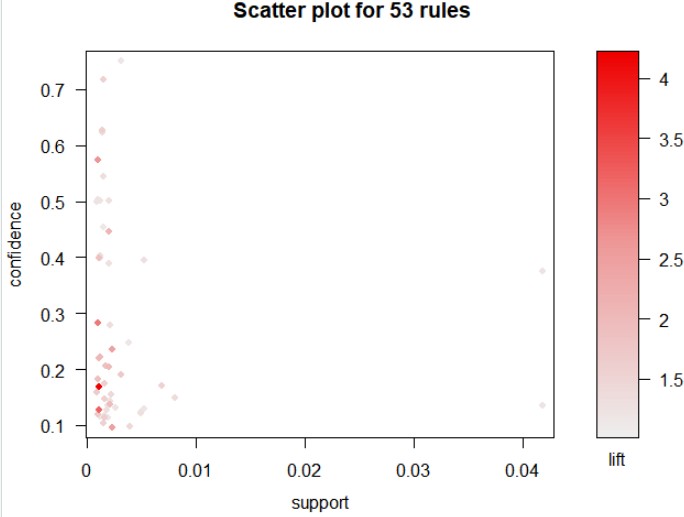


*Figure 26 - RPF of Colors Matching Rules*



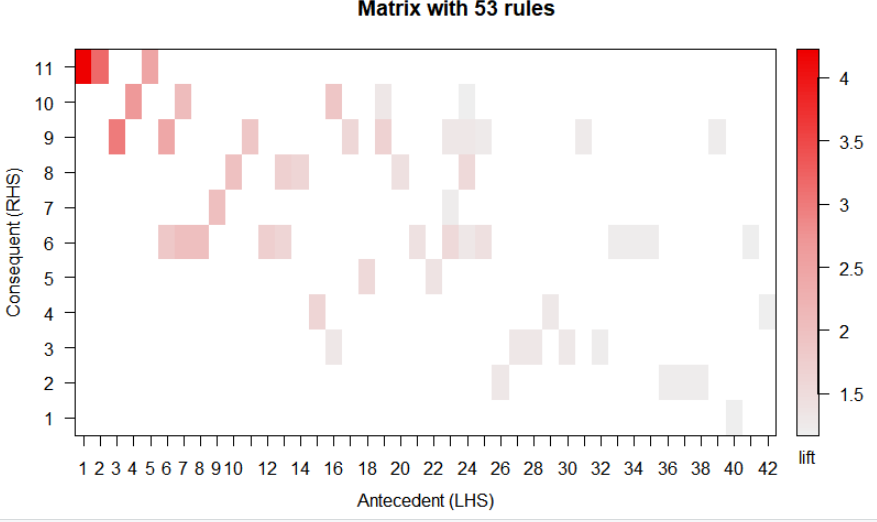
*Figure 27 - Conviction of Colors Matching Rules*

* + 1. Scatter Plot showing the Support, Confidence and Lift of the Rules



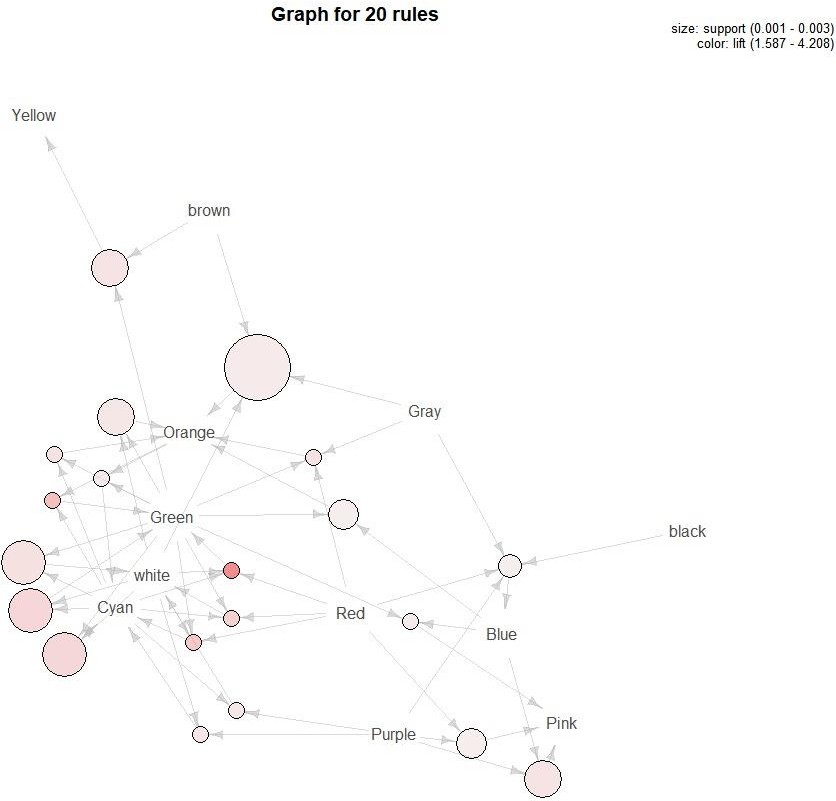
*Figure 28 - Scatter Plot of Colors Matching Rules*

* + 1. Matrix Plot that shows the Antecedent, Consequent and the Lift of each rule.



*Figure 29 - Matrix Plot of Colors Matching Ru*

* + 1. Graph for the trending colors matching generated rules



*Figure 30 - Colors Matching Rules as Graphs*

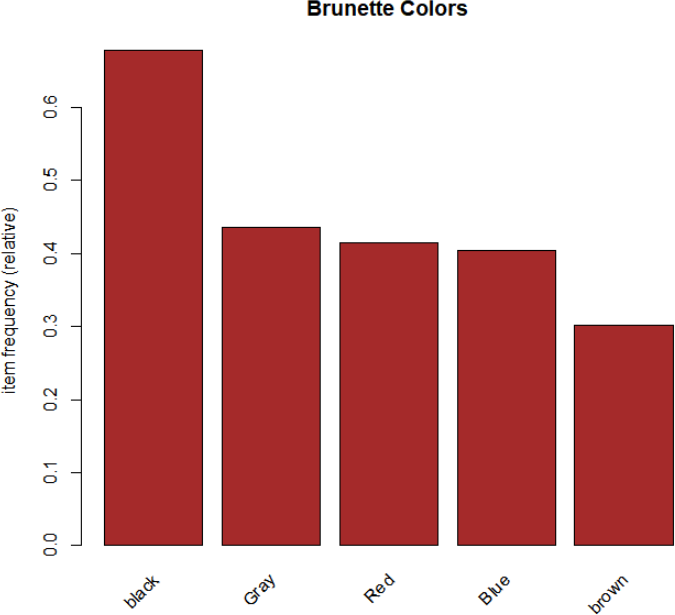
## Analysis of Trending Colors According to Skin Color:

* + 1. Most frequent colors that blonde people wear



*Figure 31 - Blonde People Color Trends*

* + 1. Most frequent colors that brunette people wear



*Figure 32 - Brunette People Color Trends*

Note that from the graphs, we can see that the blondes and brunettes wear nearly the same colors with slight differences and that makes sense. As their skin color is nearly the same, the only difference is in their hair color, so it matches with nearly the same colors. What was going to make difference is the most frequent colors worn by dark skinned people. However, this does not appear in our case, as our dataset does not contain dark skinned people.

# Enhancements and Future Work:

Enhancements will be made on the current model to include some future work that is of great value. The future work includes:

* 1. Improving the FCN model to obtain higher accuracy to be able to crawl fashion websites and Instagram blogs, extract the pieces from them and find the matching rules based on the current trend.
  2. Find the trending outfits and colors for each season by saving the date at which the image was taken. Therefore, we find this way: summer outfits, winter outfits and so on.
  3. Use Gabor Feature Extractor along with the colors extractor in order to find both textures and colors in each image; this will help in generating rules that are more specific.
  4. Usage of the rules obtained in shopping stores for Pieces Placement/Organization, such that pieces that match together would be placed together in stores.
  5. Usage of the rules obtained in online shopping such that when a person is about to buy a certain clothing piece, the website recommends to him other pieces to buy that matches with this piece (cross-sell)

# References

[https://en.wikipedia.org/wiki/Association\_rule\_learning](https://en.wikipedia.org/wiki/Association_rule_learning#Conviction) <https://en.wikipedia.org/wiki/Unsupervised_learning>

<https://cran.r-project.org/web/packages/arulesViz/vignettes/arulesViz.pdf> <https://rdrr.io/cran/arules/man/interestMeasure.html> <https://arxiv.org/pdf/1701.09042.pdf>