**Beauty Lies in the Averageness: Response to Reviews**

**Author: Lam Tran**

For the final draft, I have added a figure in the Discussion section to sum up major results from the models, and added few sentences in the Discussion section about factors I could have considered (which was inspired by the questions I received after my presentation). After receiving the reviews for my first draft from Sally Hyde and Dr. Bonifonte, I have also made some changes to the wordings in my paper to make my points clearer and modified several graphs to better communicate my statements. To be more specific, my changes were as followed:

Sally’s comments:

* *I felt like some of the transitions and topic sentences in the Related Work section could be smoother, particularly the last two paragraphs of the section (pg. 5)* => Thank you for the suggestion, I have revised my wordings.
* *I think averageness needs to be operationally defined earlier in your paper. It became clearer on page 10 but prior to that I find myself continuing to ask whether ‘average’ pertained to measurements of facial features or the ratings of attractiveness.* => Very good comment, maybe I did not realize this because I am so used to knowing what “averageness” is. I have defined the concept earlier in the Introduction section. “Therefore, following this hypothesis, if a face has higher “averageness”, or if a face possesses features whose sizes are closer to the average sizes, the face should be more attractive.” I also changed some words from “facial averageness” to “feature averageness” to avoid the confusion prior to page 10 where I explicitly talked about how I calculated the averageness. I did not want to talk about the calculation earlier than page 10 because it is not at the high level that the introduction and related work sections need to be at.
* *I wonder if in your exploration of average attractiveness ratings it would be possible to look at the interactions between some of the rater demographics. I’m reminded of this article I saw and wondered if your data could, for example, compare the ratings of 45-60 y/o men to 45-60 y/o women:* [*https://www.businessinsider.com/dataclysm-shows-men-are-attracted-to-women-in-their-20s-2014-10*](https://www.businessinsider.com/dataclysm-shows-men-are-attracted-to-women-in-their-20s-2014-10)=> This was a very interesting idea and thank you for the article reference. However, I opted not to look at the interactions between rater demographics for three reasons. First, there were a lot of demographics and it is not possible to report all of them even though they were all really interesting. Second, insights from demographic interactions could largely be inferred from the information we already had (Tables 3 and 4). I have examined several interactions (including the one you mentioned) and found out they always followed the general trends: both men and women rate women as more attractive, older people got lower ratings from both people of their age and not of their age, so I did not report these interactions in the paper. Lastly, since the ratings were so different across demographics, that alone was enough to reject the hypothesis of universal beauty standard without further investigation of interactions. Nonetheless, maybe future research that focuses on attractiveness in specific demographics would use this idea, so I have mentioned this in the Discussion section. “Furthermore, studies could also investigate the interactions in ratings between demographics, such as whether older men consider younger girls as more attractive while older women prefer men their own ages”.
* *In Figure 11, does it show that Model 2 on average did a good job in predicting attractiveness as you say? Or is there a better way to represent that statement?* => This is a good point, the graph might be confusing because there were too many data points on discrete scales. I have replaced this graph (actual versus predicted ratings) by another graph (histogram of the squared errors) to better communicate my statement.
* *For your Tables (especially 5 and 6), it might be more eye-catching to make them more traditional or academic-y because right now they look less polished than some of your other work* => I have changed the color of table 3, 4, 5, 6, and 7 to gray make them look more polished and more academic, but not too colorful to distract the audience from the content
* *Spelling/grammar suggestions in google doc* => Thank you, I have revised them
* *Third person versus first person conflict* => I have paraphrased some sentences in the Result and Methodology sections to avoid speaking as myself (“I”).
* *Repetition of “Notice that” and “Note that”* => I have paraphrased several sentences that have “Notice that”, “Note that”, and “It is worth noting that”.

Dr. Bonifonte’s comments:

* *Spelling and citing errors* => Thank you for pointing out, I have made the changes
* *Not understanding “Therefore, I first explored the pairwise average ratings in each demographic group, then performed random forest models where the ratings of one pair were significantly different from the ratings of others.”* => Paraphrase to “Therefore, subsequent random forest models were only run for the demographic groups with extreme attractiveness ratings as they could most likely generate interesting results.”
* *Did you investigate max feature deviation from average vs attractiveness? Why sum? (figure 9)* => I opted not to investigate max feature deviation versus attractiveness for several reasons. First, for a face with one extreme feature, the attractiveness ratings might not be much lower because other features might be average. Second, if the feature with the max feature deviation is a small feature, it could not affect the attractiveness ratings that much. Therefore, I chose to sum up the feature deviation to take into account every facial features. However, although the trend line of sum of feature deviation and average of feature deviation are similar, I now think that average of feature deviation might be more interpretable for the audience so I have changed my x-axis to become the average value. Furthermore, to consider every feature equally, I have normalized the deviations before visualizing them in figure 9.
* *Report equation and p-value of the trend line in table (figure 9)* => Thank you for the suggestion, I have added the trend line and p-value directly to the figure instead of reporting them in a new table.
* *What does publication-friendly mean?* => I am not permitted to present any image in the 2,222 images I used for training. Instead, there was a dataset of 49 so-called “publication-friendly” images that could be presented. I have made a footnote in the paper to clarify this.
* *Result section about univariate linear models might make sense before the random forest models* => I have seriously considered this while writing up the paper, and I understand why you thought so. However, I opted not to rearrange the order of my result section because I want to point out in my linear regression section that even for variables that were very important in the random forest models, their coefficients’ signs are also positive. In order to do that, I have to mention the random forest models before the univariate linear regression models.
* ****

**Beauty Lies in the Averageness**

**Author: Lam Tran**

Denison University, Spring 2021

Senior Capstone Research – DA 401

Advisor Dr. Bonifonte

**Abstract**

With the potential applications of deepfake where fake actors, actresses, holograms, and virtual assistants can be used in commercials, the need to know what makes an attractive face has been growing in recent years. From the evolutionary point of view, there have been debates that averageness or commonness of the facial feature is one of the most important factors that affects one person’s beauty. Moreover, it is believed that there was a universal standard of beauty, or there is a set of features that strongly affect one’s attractiveness regardless of the observed or the observer’s demographic background. This study utilized univariate linear regression models to determine whether averageness positively influences attractiveness, and random forest models to determine the most important facial traits that affect attractiveness. In the results, the averageness of some facial traits, but not all, positively influenced the face’s attractiveness. Furthermore, people always consider eye, lip, jaw features, and the balance between face width and face length to be crucial for one’s beauty. However, people also consider other different facial traits when they rate other’s attractiveness, depending on both the observed and the observer’s demographic background. The results are reliable as there is a sufficient number of observations to train the models, the feature selection process is comprehensive, and data were collected to represent people of all ages, races, and genders. Future studies should be conducted to investigate which features will improve one’s attractiveness when they are average and which will improve the attractiveness when they are more extreme; and why people from different demographic backgrounds looked at different features when they assess the attractiveness of another.

**Introduction**

The question of what constitutes beauty has troubled artists and philosophers for centuries. It was not until recent years that scientists tried to figure out what makes a beautiful face, particularly because facial attractiveness has very important influences on one’s interpersonal and social life. For example, beauty is associated with upward economic ability, more professional opportunities, more and even better romantic experience (Little, Jones, & DeBruine, 2011). Attractive people are also more likely to be perceived to have likable personalities, which results in many favorable biases from other people even in professional and legal settings in a subconscious way (Little, Jones, & DeBruine, 2011). Therefore, knowing what makes a face attractive could lead to many potential applications in industries that need attractive rendered avatars such as the AI assistant, hologram, or gaming industries. Rendered attractive faces could also be utilized in place of actors and actresses in advertising. One review article mentioned that fake faces would especially be powerful if they are dynamically customized to fit in tastes of individual consumers that the brands could draw from the extraordinary rich data of the consumers’ social media profile and online activities (Kietzmann, Mills & Plangger, 2020).

However, despite many psychological studies on the social consequences of facial attractiveness and many potentials of fake attractive faces, exactly what facial features attribute to higher attractiveness remains unsolved. People usually believe that there is no universal standard of beauty, or that “beauty is in the eye of the beholder.” Darwin went further to argue that there are cross-cultural differences that exist in beauty standards, such as in skin color, body hair, body weight, or accessories (Darwin, 1896). However, empirical studies have pointed out that, although beauty standards differ culturally and individually, this difference is only to some extent as there was a high degree in cross-cultural agreement on attractiveness (Little, Jones, & DeBruine, 2011). This suggests that people are using the same or similar criteria in their judgment of beauty, which most likely stems from their biological instincts.

From an evolutionary point of view, humans always give preferences to mates who are most likely to provide the best chance of surviving to their offspring. Traits that are associated with good genes, therefore, will become attractive. Several facial traits have been proposed to advertise the good biological quality of an individual, such as symmetry and averageness/commonness (Little, Jones, & DeBruine, 2011). To elaborate, extreme and non-average features indicate that the individuals are carrying disadvantageous genes, so humans will perceive the carriers as less attractive (Koeslag, 1990). Therefore, following this hypothesis, if a face has higher “averageness”, or if a face possesses features whose sizes are closer to the average sizes, the face should be more attractive. On the same note, symmetry is one indicator of the individuals’ strong ability to maintain stable development of their morphology under environmental effects and should also be preferred in mating (Moller & Swaddle, 1997).

Based on the mentioned hypotheses, this paper will examine whether faces with average-sized features and ratios are perceived as more attractive. The paper will also explore whether people from different ethnicity, age, and gender agree on which facial features are the most important factors that affect attractiveness to determine whether a cross-cultural beauty standard exists subconsciously.

1. **Related Work**

The topic of facial attractiveness has been studied extensively in the psychology field. To confirm the correlation between facial attractiveness and feature averageness, psychological studies usually asked judges to rate both attractiveness and averageness on images of the opposite gender and then find the correlation (Grammer & Thornhill, 1994), or to rate the attractiveness of images of real people and of images that are morphed toward an average shape and then compare the two sets of ratings (Langlois & Roggman, 1990). Some studies have concluded that feature averageness has a statistically significant positive influence on facial attractiveness, which follows the evolutionary hypothesis. However, it was argued that using subjective ratings on averageness is unreliable as a person could not possibly perceive if the presented face is similar to thousands of faces he or she has seen (Komori, Kawamura, & Ishihara, 2009). Another strong critique was that digitally blending faces tend to create a more symmetrical and smoother face with bigger facial components, so it is difficult to isolate the effect of averageness on attractiveness (Komori, Kawamura, & Ishihara, 2009). Therefore, the credibility of the results of these studies is still debated.

With the rise of image processing algorithms in recent years, researchers have thought of a better way to address feature averageness. Recent studies have asked judges to rate only on attractiveness and used face recognition algorithms to computationally calculate the averageness of the features (Baudoin & Tiberghien, 2004) (Eisenthal, Dror, & Ruppin, 2006). Specifically, the algorithms would detect “facial landmarks” on the images, which represent salient points of the face components such as eyes, mouth, or nose. The models could be trained directly on the x-coordinates and y-coordinates of the landmarks, or the landmarks could be used to calculate “facial features” (such as the length of the chin or the height of the forehead) to be fed into the models. One of the drawbacks of these studies was that they used very small and not diverse training sets that most likely led to biased results. To elaborate, one research used 96 self-collected images of young Japanese female and male college students, and one used self-collected images of 62 young Caucasian adult women (Baudoin & Tiberghien, 2004) (Eisenthal, Dror, & Ruppin, 2006). This could be attributed to the lack of large, aggregated image databases back then, but better datasets are becoming publicly available and could greatly improve these studies.

Another drawback of existing studies was that they did not employ models with a good balance of accuracy and interpretability. Some employed black-box models with high accuracy (over 80%) and low interpretability, such as neural network, KNN, linear SVM, and polynomial SVM (Lal et al., 2018) (Choudhary, Agrawal, & Kaur, 2020); while some employed models with high interpretability but low accuracy. For example, a research that focused on facial attractiveness and averageness used a multivariate linear regression model to look into what facial features are the most important in attractiveness rating (Baudoin & Tiberghien, 2004). The model could give detailed insights about the effect each facial feature has on attractiveness, but it had a moderate adjusted R-squared value of 0.6, which could be perceived as not as powerful as the machine learning models, and it violated multiple assumptions about the linear regression models which most likely biased the results. Another research in facial attractiveness utilized a C4.5 decision tree model whose accuracy was as high as other black-box machine learning models, and it was moderately interpretable as it could give information about variable importance (Eisenthal, Dror, & Ruppin, 2006). Therefore, decision tree and random forest models were more suitable than aforementioned models for the research question about facial attractiveness and averageness as they had a good balance of both accuracy and interpretability.

Inspired by previous studies, this study combined univariate linear regression models and random forest models to achieve both high accuracy and high interpretability, with features that have been statistically significant in previous studies as inputs. Moreover, this study was an improvement from the dataset aspect, as the models were trained on 2222 images of US adults that have different ages, genders, and ethnicities (Bainbridge, Isola, & Oliva, 2013). This not only allowed the results to be more accurate but also more applicable to a bigger, more general pool of the population.

1. **Proposed Methodology**
   1. *Data Collection*

The dataset that was used for this study was a part of “The 10k US Adult Faces Database.” (Bainbridge, Isola, & Oliva, 2013). The original database contained 10,168 images of face photographs of the adult United States population. To collect these images, the authors used an online random name generator based on the 1990 U.S Census name distribution, then randomly sampled 25,000 first and last names and automatically download several color face photographs from Google Image Search that corresponded with random pairs of first and last names. Images of celebrities, children, faces with occluded objects or unusual makeup were filtered out by five observers to ensure consistent quality. The resulted 10k US Adult Faces Database, therefore, complied with the U.S copyright law’s fair use provision and had 10,168 individual faces that followed gender, age, and race distributions of the adult U.S Population according to the 1990 U.S Census. Although the distribution of races in the U.S population has somewhat changed in the last 30 years (Gibson & Jung, 2002) (U.S Census Bureau), this dataset is still believed to be the dataset that best represented U.S population by the time of this study as other datasets usually heavily sampled or only sampled white people.

For 2,222 randomly selected images from the 10k US Adult Faces Database, two surveys were conducted to collect demographic information and psychological attributes on each of the photos. In the first survey, 12 participants from Amazon Mechanical Turk were asked to provide their demographic information and to rate on 19 attributes of the image, such that the face’s age (<20 years old, 20-30 years old, 30-45 years old, 45-60 years old, and 60+ years old), gender (female and male), race (White, Black, East Asian, South Asian, Hispanic, Middle Eastern, and other), and attractiveness (on a scale from 1-5 with 1 being unattractive). In the second survey, there were 30 participants from Amazon Mechanical Turk, and 15 randomly selected participants among them were asked to rate on 20 psychological attributes of each face, such as attractiveness, emotion, friendliness, or trustworthiness (on a scale from 1-9 with 1 being “not at all”). Each survey included a catch question to eliminate participants who were answering at random to ensure data validity, and if the participants answered the catch question wrong, all of their answers were removed from the database. All participants consented to the data collection process following the IRB guidelines (Bainbridge, Isola, & Oliva, 2013).

The dataset that was used for this study is these 2,222 images. Rows with missing values for the attractive ratings or demographic labels were removed from the data as they are missing randomly and account for less than 5% of the total observations.

*3.2 Data Processing*

The output variable of the models was the attractiveness ratings of each image. Notice that there were two sets of ratings, one from 1-5 for 12 judges and one from 1-9 for 15 judges. Each set of ratings were min-max normalized before merging because they were on different scales. In the end, each image will have 27 attractiveness ratings from 0-1, 15 of which also had the demographic information of the raters. Averaging the ratings would make the data lose its variance and reduce the number of observations available for the models in the data analysis process, so this method was not opted for.

On the other hand, the input variables were the averageness of different facial features that were derived from the facial landmarks. First, to account for different head positions of the images in the dataset, the face in each image was re-aligned (standardized) along the x-axis (the horizontal line between two iris) and the y-axis (the vertical line from forehead to chin) using a pre-trained 3D face model (Hassner et. al, 2015). The images were also standardized to have the same width and height before each image’s facial landmarks were taken using the dlib package in Python (King, 2009). Specifically, as the dlib’s pre-trained model can best detect 80x80 pixel pictures and I wanted to retain the original width to height ratio of the images so the facial features would not be stretched out and miscalculated, all images were rescaled to 80x100 pixels.

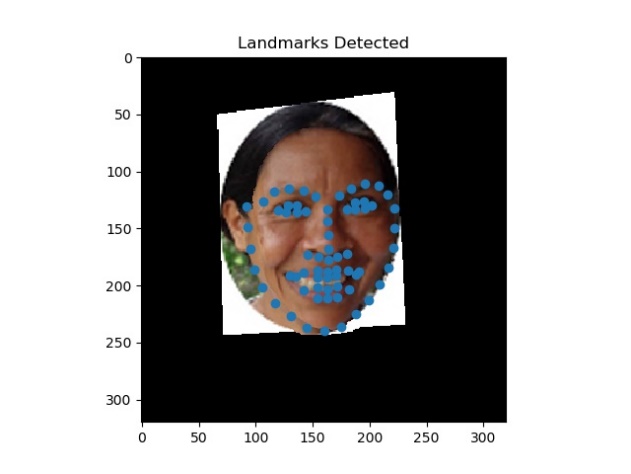
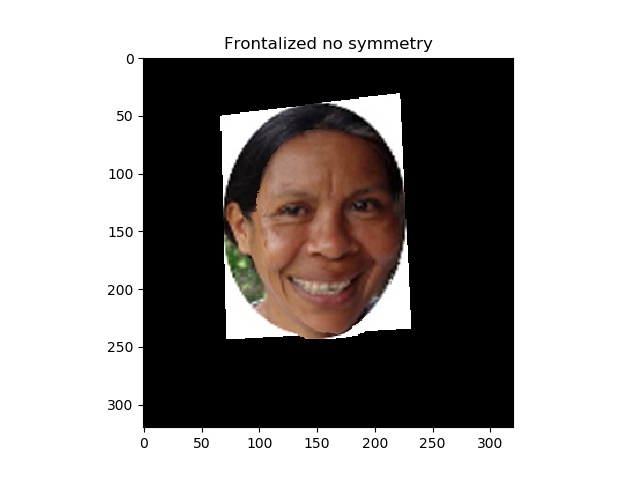
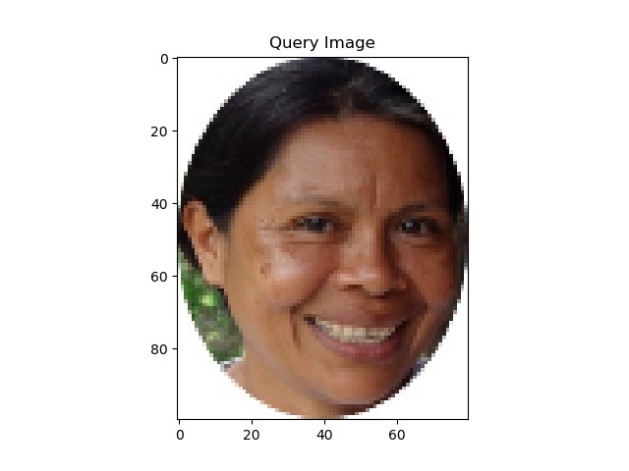


Figure 1: Example of frontalization and landmark detection process of an image

Then, 26 facial attributes were derived from the landmarks. The features that I considered for this work were 40 facial features that were statistically significant in a previous study about facial features and attractiveness (Eisenthal, Dror, & Ruppin, 2006). However, because it was strongly believed that features from the left and right side of the face (such as left eye and right eye) should have similar effects on facial attractiveness, such features were combined by taking their average value instead of separating them as in Eisenthal, Dror, & Ruppin (2006). Furthermore, the dlib package did not detect landmarks for the forehead region, thus some features could not be derived and the final number of features was 26 instead of 40. Then, the absolute value of the difference between each facial feature and its mean value was calculated. This so-called feature deviation from average was used as the inputs for the models.

Table 1: List of the measurements and names of the facial features

Graphical user interface, text, application, email

Description automatically generated

Finally, as the demographic information of the images was guessed by judges, the demographic category that had the majority votes was considered the images’ demographic category.

*3.3 Model Selection*

To examine whether faces with average-sized features and ratios are perceived as more attractive, 26 univariate linear regression models were performed for the absolute differences of each facial feature from its average value and the attractiveness ratings. If the sign of the coefficient of the variable was negative, this meant the closer the image’s facial feature got to its average value, the more attractive the face was perceived. In other words, the more average the feature was, the more attractive the face was. In contrast, if the sign of the coefficient was positive, the face was perceived as more attractive when the feature deviated further from its average value. As the hypothesis was that averageness increases attractiveness, negative signs of the coefficients would support my theory. a multivariate linear regression model with 26 independent variables was not opted for because the facial features were not independent, some were even highly correlated with each other. Therefore, in such a model, the multicollinearity assumptions for multivariate linear regression would not be satisfied and the effect of each feature on the attractiveness rating could be biased.

To further explore which facial features are the most important factors that affect the attractiveness ratings, a regression random forest model was run using 26 facial features as input. A decision tree model was tested in similar research and was proven to have both high accuracy and the ability to give insights about variable importance (Eisenthal, Dror, & Ruppin, 2006). The main problem with single decision tree models is that they tend to overfit the data as they can model highly complex, non-linear relationships between variables. Therefore, a random forest model was opted for in this study because it could prevent overfitting by aggregating and averaging the results of many decision trees.

It is worth noting that for such a regression problem, a multiple linear regression model could be employed instead because it could also give detailed insights about the effect each facial feature had on the attractiveness ratings. However, the relationship and interaction between input features in this study were highly complex and not suitable for linear models. Although polynomial terms and interaction terms could be put into a linear regression model to fit non-linear relationships, this would significantly reduce the interpretability of the model and increase the risk of overfitting, especially with 26 independent variables. Last but not least, as mentioned above, the data cannot satisfy some of the assumptions of a linear model, which would result in biased coefficient and significance terms. In conclusion, a random forest model is a better fit than linear regression models for this study.

After running a random forest model on all data, subsequent random forest models were run on different subsets of the data to confirm whether judges would subconsciously look at different facial traits when they rated faces from the same demographic background and faces from a different demographic background. To be more specific, judges and images were divided into two genders (male/female), five age groups (<20, 20-30, 30-45, 45-60, 60+ years old), and seven ethnicity groups (White, Black, East Asian, South Asian, Hispanic, Middle Eastern, and other). Though it was possible to run one random forest model for each pair of demographic backgrounds (such as males rate males, males rate females, females rate males, and females rate females for the gender group) to test the hypothesis; it was deemed unnecessary to perform and report tens of models if the results would be similar. Therefore, subsequent random forest models were only run for the demographic groups with extreme attractiveness ratings as they could most likely generate interesting results. The proposed theory that individuals have universal, cross-cultural, and unbiased beauty standards would be supported if raters considered the same facial features important no matter the demographic backgrounds where they and the images came from.

*3.4 Model Validation*

To achieve the random forest model with the highest predictive power or the lowest MSE value for all data, recursive correlation pruning (removing highly correlated variables) and recursive variable importance pruning (removing variables with the lowest importance) were performed to remove unnecessary variables from the model. To further prevent overfitting, cross-validation with a conventional split of 80% training set and 20% test set was applied to the random forest models. A scatter plot between the observed and predicted values of the attractiveness ratings was also graphed to confirm the performance of the random forest model for all data.

1. **Results**

*4.1 Data Exploration*

Because the hair and part of the forehead of the faces in the images were cropped out during the data collection process, among 2,222 images in the dataset, dlib could only detect landmarks of 2,206 images (99.28%). After examining, I concluded that these 16 images were scattered across different demographic backgrounds and attractiveness ratings, and thus could safely be ruled out of the analysis without creating any bias.

*4.1.1 Exploration of average attractiveness ratings*

Figure 2 provides an overview of the distribution of attractiveness ratings. It is worth noting that the alternate heights of the bars were not due to the way the judges rated the images but due to the way the data was normalized. To elaborate, there were two sets of attractiveness ratings: from 1-5 and 1-9. After both set of ratings were normalized, the ratings on the 1-5 scale would become 0, 0.25, 0.5, 0.75, and 1, while the ratings on the 1-9 scale would become 0, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, and 1. The common ratings of both sets after normalization were 0, 0.25, 0.5, 0.75, and 1, which explained why these ratings had much higher frequencies than the rest. After taking this into account, we could see that the ratings followed a general bell curve that indicated normal distribution, where judges were more likely to give out medium scores and were more hesitant to give the lowest and highest scores. This is similar to what has been observed in many situations involved ratings in real life and thus validated the data.

Chart

Description automatically generated

Figure 2: Frequency of the normalized attractiveness ratings

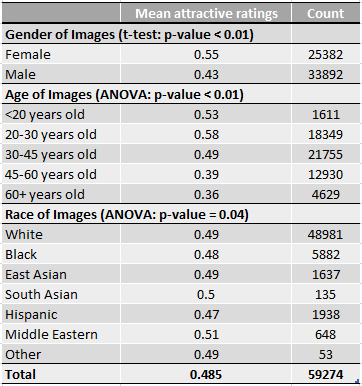
Taking a closer look at the summary statistics, we could see that the average rating was 0.485. This number was close to 0.5, supporting the claim that the ratings followed a normal distribution, but it also told us that judges tended to slightly underrate the attractiveness of the faces.

Table 2: Summary statistics of the normalized attractiveness ratings



Table 3 provides a deeper understanding of ratings that people of different demographic backgrounds received (N = 59,274). First, we see that females on average got the rating of 0.55, which was statistically significantly higher than the average rating of 0.43 that males got (two-tailed t-test: p-value < 0.01). Second, there is also a statistical difference between the ratings that people from different age groups received (ANOVA: p-value < 0.01). To elaborate, young adults that were less than 20 years old got an average rating of 0.53, which is higher than the overall average of 0.485. Then, the average rating increased to 0.58 for people from 20 to 30 years old, 10 percentage points higher than the overall average. After hitting the peak at 20 to 30 years old, the average attractiveness rating dropped to 0.49 for people from 30 to 45 years old, 0.39 for people from 45 to 60 years old, and hit the lowest of 0.36 for people older than 60. This fits in with common sense that younger adults are generally perceived as more attractive compared to middle-aged and older people.

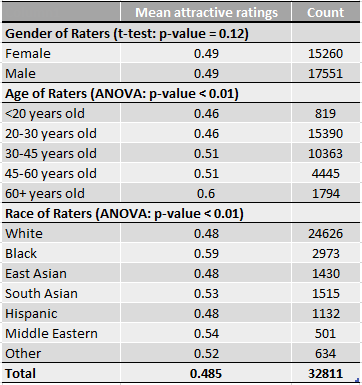
Table 3: Summary statistics of the attractiveness ratings that the images received



Last but not least, there is also a statistically significant difference between the average ratings of people with different ethnicities, with Hispanic received the lowest average attractiveness rating of 0.47 and Middle Eastern received the highest average attractiveness rating of 0.51. However, the confidence interval of this ANOVA test is narrower than the ANOVA test for age groups (p-value = 0.04 compared to p-value < 0.01). This is because 5 out of 7 ethnicity groups received an average rating of 0.48-0.50. If we removed either the Hispanic or Middle Eastern group from the ANOVA test, the p-value would surpass 0.05, meaning the average ratings of the other groups are not statistically different from each other at the 95% confidence level.

On the other hand, Table 4 provides a statistical summary of ratings that people from different demographic groups gave out. Note that only a subset of the data has the demographic information of the raters, so the number of ratings in this table will be fewer than that in Table 2 (N = 32,811 compared to N = 59,274). Interestingly, even though females received much higher ratings than males on average, actually the average ratings that females and males gave out are not statistically different at the 90% confidence level (t-test: p-value = 0.12). This could mean that both females and males systematically gave females higher attractiveness ratings and gave males lower attractiveness ratings.

Table 4: Summary statistics of the attractiveness ratings that the raters gave



Next, it is also interesting that people from different age groups rated images differently (ANOVA: p-value < 0.01). Specifically, young adults up to 30 years old gave out the average rating of 0.46 (lower than the overall average of 0.485), middle-aged adults from 30-60 gave out the average rating of 0.51, and older adults of 60+ gave out a much higher rating of 0.60 (11.5 percentage point higher than the average). Notice that the average rating increased with age, which could be attributed to the observation that older people have more open and positive attitudes toward others (Noftle & Fleeson, 2010). However, this relationship was not linear because there was not always a jump in ratings between the five age groups. For example, the age group of fewer than 20 years old gave out the same average rating as the 20 to 30 years old group did, and the age group from 30 to 45 years old gave out the same average rating as the 45 to 60 years old people did. Rather, there were two jumps around the age of 30 and 60, which might be due to life-turning events such as having a family or retirement that could greatly change a person’s perception of beauty.

Finally, there also exists a statistical difference in the average ratings that different ethnic groups gave out. To be more specific, White, Hispanic, and East Asian people gave out a rating of 0.48 on average, which is very close to the overall average rating of 0.485. Meanwhile, Middle Eastern, South Asian, and other ethnicity gave out higher average ratings (0.54, 0.53, 0.52, respectively), with Black people gave out the highest rating of 0.59, which is 10.5 percentage point higher than the overall average.

Since there were many differences between the ratings that a demographic group gave out and that they received, Figures 3, 4, and 5 provide a more detailed look at the interaction between the raters and the images. Figure 3 confirmed my hypothesis that both females and males systematically gave females higher attractiveness ratings and gave males lower attractiveness ratings. Even more interesting, both females and males gave male images an average rating of 0.43, while females and males gave female images two very close average ratings of 0.53 and 0.54. This indicated that there was no difference in the way the two genders rated images, although there existed a significant gap in the way females’ and males’ beauty were rated.

Chart, treemap chart

Description automatically generated

Figure 3: Heatmap of the average attractiveness ratings across genders

Looking at Figure 4, we see a fascinating trend that while people younger than 30 years old tended to give people from their age group the highest ratings, people older than 30 years old did not think their age group was the most attractive, and also gave 20-30 years old people the highest ratings. Overall, old people of 60+ years old gave 20 to 30 years old young adults the highest average rating of 0.71, and in return, young adults from 20 to 30 years old gave the lowest average rating of 0.35 to old adults of more than 60 years old. To sum up, it was clear that people did not have the same standard of beauty for images across different ages because they tended to consider younger adults more attractive and older adults less attractive.

Chart, treemap chart

Description automatically generated

Figure 4: Heatmap of the average attractiveness ratings across age groups

In Figure 5, the ethnicity group ‘Other’ was removed because there were too few images that belonged to the ‘Other’ category and so some interactions were not available. Again, we see a strong trend here that black people tended to give much higher ratings than any other ethnic group did. Besides, Middle Eastern rated Middle Eastern the highest average rating of 0.64, and South Asian rated Black people the lowest average rating of 0.37. Among the six races, Middle Eastern and South Asian people tended to give people from their background the highest ratings of 0.64 and 0.60, respectively. However, these numbers were biased because there were only 5 South Asians who rated South Asian images, and there were 8 Middle Eastern who rated Middle Eastern images. Therefore, there was also no clear trend that people from one ethnic background prefer one specific ethnicity’s beauty over others’.

Graphical user interface, chart

Description automatically generated

Figure 5: Heatmap of the average attractiveness ratings across races

*4.1.2 Exploration of facial features*

Figures 6, 7, and 8 are the example histograms of three facial attributes: eye size, the ratio from chin to nostril and from nostril to eyebrows, and jaw length. As the values represented here were calculated by taking the absolute difference of the original value of each facial feature from their mean, it was expected that the graph would be right-skewed. In other words, more images had their facial features closer to the average value, and fewer images had their facial features deviate further from the average value. However, the standard deviation of each facial feature was different from others, resulting in different distribution histograms. For example, we see that in Figure 6, the heights of the bars decreased gradually, indicating a larger deviation as a lot of images did not have a facial feature value that was close to the average; while in Figure 8, the heights of the bars dropped much faster, indicating smaller deviation from the mean. In Figure 7, there were certain gaps in the histogram that I could not explain. In addition, in the three graphs, we can see some outlier observations that greatly deviated from the sample’s mean. These outliers were not removed because they were the observations that could heavily affect the performance of the models, and removing them would therefore create biased results.

Chart, histogram

Description automatically generated

Figure 6: Frequency of the eye size feature

Chart, histogram

Description automatically generated

Figure 7: Frequency of the ratios of chin-nostril to nostril-eyebrows

Chart, histogram

Description automatically generated

Figure 8: Frequency of jaw length feature

*4.1.3 Interaction between facial features and attractiveness ratings*

Figure 9 plots the average of all facial features’ deviations from their mean versus the attractiveness ratings. Due to the different scales of the features, the values here were min-max normalized so all features were considered equally. In the graph, most of the points fell into the left region of the images, which means that most facial features were close to their mean, consistent with the previous findings. Recall that if averageness had a positive influence on attractiveness, the higher the facial feature deviations is, the lower the attractiveness ratings should be. Here, the trend line in the graph has a slightly negative slope, meaning that the averageness of the facial images had a slightly positive correlation with the attractiveness ratings of the face, which followed the hypothesis. However, for one unit increase in the average normalized feature deviation, the ratings only decrease by 3%. This means the influence is very weak, especially when one unit increase in the deviation is extremely large. The insignificant p-value of 0.77 for the coefficient also supported that averageness had little influence on attractiveness.

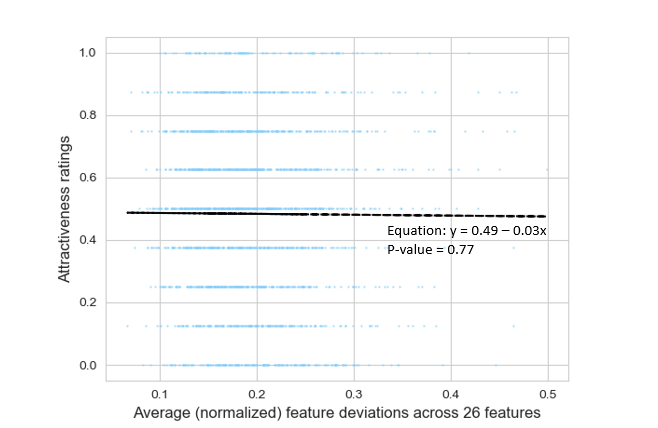


Figure 9: Plot between the average feature deviations versus the attractiveness ratings

Figure 10 shows 3 images: one was the most attractive image, one was the most average image, and one was at the top of both attractive and average images. As can be inferred, not all attractive images were average, and not all average images were attractive; however, some images were both attractive and average. Though, it should be emphasized that these images were taken from the set of 49 images that were publication-friendly[[1]](#footnote-1), which were not included in the set of 2,222 images. As 49 images were too few, the average of the facial features from these sample images did not represent the average of the whole population.

Graphical user interface, text

Description automatically generated

Figure 10: Sample of faces on the scale of attractiveness and averageness

*4.2 Random forest models on the whole dataset*

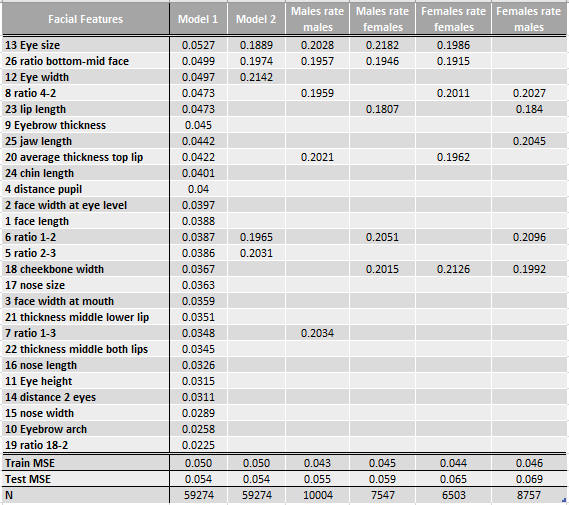
In this study, the random forest models were run with a train-test split of 80/20 and with 1000 decision trees. The first model, Model 1, was trained on the whole dataset with 26 facial attributes as input variables and the attractiveness ratings as the output variables (N = 59,274). The mean squared error (MSE) values of the train and test dataset were 0.050 and 0.054, respectively, and the importance of the variables is provided in Table 5. The MSE for the training set and test set were close to each other and were low, indicating that the model did not overfit or underfit and thus performed well. Interestingly, no feature had significantly higher importance than other features. Among the most five important variables in Model 1, two belonged to the eye regions: eye width and eye size, and one belonged to the lower part of the face: jaw length. The other two most important variables were more macro: the ratio between the distance between pupils and the face width at eye level, and the ratio between the distance from nostrils to eyebrow and the distance from chin to nostrils.

To simplify the model, the variables that were least important were sequentially removed from of the model (recursive variable importance pruning). 21 variables were therefore removed before the pruning affected the predictive power of the random forest model, resulting in Model 2. Model 2 only contained 5 variables, but it retained the same train and test MSE values from Model 1. This means that information from the 21 removed variables was also contained in the 5 most important variables, most possibly because the variables were highly correlated with each other. Similar to Model 1, because both train and test MSE of Model 2 were low, Model 2 did not overfit or underfit the data. From Figure 11, we also see that Model 2 on average did a good job in predicting the attractiveness ratings of the images as most predictions were either accurate or only off by a small margin.

Chart, histogram

Description automatically generated

Figure 11: Histogram of the squared errors between the predicted and observed ratings by Model 2

Table 5: Feature importance from Random Forest models

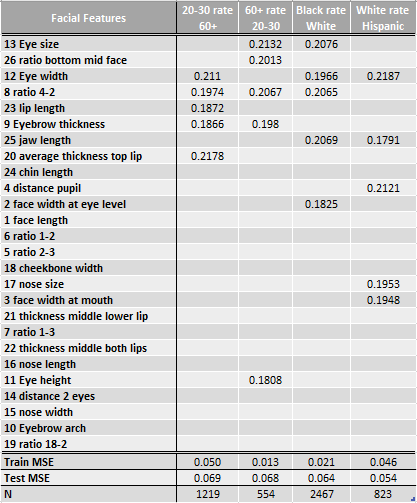
In Model 2, we again see that there was little difference between the importance of features. Compared to Model 1, the five most important variables in Model 2 also had eye width, eye size, and the ratio between the distance from nostrils to eyebrow and the distance from chin to nostrils. The other two most important variables were other ratios: the ratio between the face width at mouth level and the face width at eye level, and the ratio between the face length and the face width at eye level. In other words, Model 2 gained most of its predictive power from the interactions between the face width, face length, eye size, and eye width. As the top 5 most important variables included 3 ratios, we could conclude that humans often consider the balances between features to be important for beauty. However, it was interesting that the other two variables, including the most important variable, were smaller ones that belong specifically to the eye regions.

*4.3 Random forest models on demographic subsets of the dataset*

The subsequent random forests were trained on a subset of the dataset that had information about the demographic information of the raters (N = 32,811) to determine whether raters subconsciously looked at different facial traits when they rated images that were from the same demographic background and different demographic background as they were. Looking at the interactions between demographic groups in the data exploration process, I decided to run each random forest model for the subset of the data where males rated males, males rated females, females rated males, female rated females, 20 to 30 years old rated 60+ years old, 60+ years old rated 20 to 30 years old, Black rated White, and White rated Hispanic. Each pair of demographic backgrounds had extreme ratings of that demographic group with sufficient numbers of observations, thus they were most likely to be able to reject the null hypothesis of a universal standard of beauty.

Tables 5 and 6 provide details about 8 random forest models that were run on demographic subsets of the dataset. Comparing the similarities between the 5 most important variables of the models, we see that 7 out of 8 models had either eye width and/or eye size, 6 out of 8 models had the ratio between the distance between pupils and the face width at eye level, and 6 out of 8 models had lip fullness and/or jaw length. This combined with the finding that eye width and eye size were in the top 5 most important variables of Model 1 and Model 2 tells us that the eye feature is important for every demographic group when it comes to beauty. We also see that there was at least one ratio included in the top 5 facial features for 7 out of 8 models, and there were at least two ratios included in the top 5 features for 6 out 8 models. This was consistent with my conclusion for Model 1 and 2 that humans subconsciously think of macro facial features (such as face width and face length) as in balance with each other instead of a standalone trait when they rated other people’s attractiveness. Other than that, although lip and jaw features were not included in the top 5 features in Model 1 and Model 2, we had sufficient evidence from these 8 models to say that they have an important role in deciding facial beauty for most demographic groups.

Table 6: Feature importance from Random Forest models (continue)



However, looking at the differences between the top 5 features of each model, we see that other than the eye features and the ratios, their smaller features tend to be different from each other. First, in the gender demographic groups, when males rated males and females rated females, they considered the fullness of the lips to be important. However, when males rated females and when females rated males, they did not care about the fullness of the lips, instead, they looked for the width of the cheekbones and length of the lips. Second, in the demographic age groups, when young adults rated older adults, they looked for very specific things such as the fullness of lips, thickness of eyebrows, and length of lips. In contrast, older people when rated younger adults only looked for one specific unique feature – the eyebrow thickness – while the other 4 features were ratios or pertained to the eyes like most models. Third, while Black rated White people, they considered the jaw length and face width at eye level to be important, while White people rated Hispanic people, they looked at the jaw length, the distance between pupils, face width at mouth level, and nose size. However, it could be argued that the distance between pupils and the face width at mouth level were highly correlated with the face width at eye level in this case.

There were 4 features that only appeared in 3 models, whose importance was most likely biased: eye height for the model where 60+ years old adults rated adults from 20 to 30 years old, face width at eye level for the model where Black people rated White people, and distance between pupils and nose size for the model where White people rated Hispanic people. To elaborate, the two models where 60+ years old adults rated 20 to 30 years old adults and where White people rated Hispanic people had less than 1000 observations with only 80% of this being used in the training process, which might not be sufficient to train a complex machine learning model like random forest. Moreover, the two models where 60+ years old adults rated 20 to 30 years old adults and where Black people rated White people were overfitted, as their training MSE was significantly lower than their testing MSE, and also significantly lower than other models’ training MSE. Therefore, the results were not trustworthy enough to say that only these groups considered eye height, face width at mouth level, the distance between pupils, and nose size to be important.

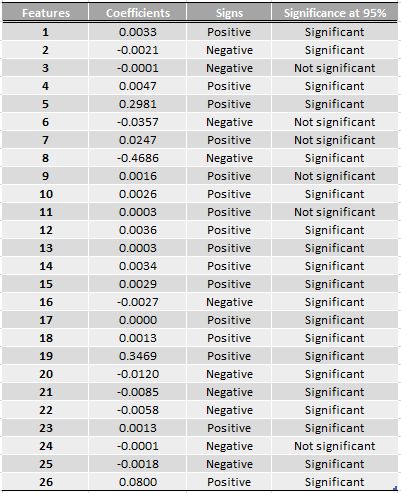
In conclusion, from the results of the models, people universally agreed that eye features and lip/jaw features were the most important facial traits, and it was crucial to have a balanced ratio between face length and face width to be considered attractive. However, people would also look at other different facial features depending on their demographic background and the other person’s demographic background.

*4.4 Univariate linear regression models*

To answer the question about averageness and attractiveness, 26 univariate linear regression models were performed, each with one facial attribute as the independent variable and the attractiveness ratings as the dependent variable. Table 7 gives details about the sign of the coefficient of the facial features in each model. As mentioned, a negative sign would support the hypothesis that the further that feature deviates from its average, the less attractive that face would be perceived. On the other hand, a positive sign would mean that the further that feature deviates from its average, the higher attractive ratings the image would receive.

From Table 7, we see that 16 among 26 facial attributes had a positive linear relationship with the attractiveness ratings. Among the 5 variables that were the most important and were included in Model 2 (Table 5), the coefficients of 4 variables had a positive sign (except for the ratio between face length and face width at the eye level). As the majority of the variables’ coefficients were positive, the result contradicted my hypothesis that averageness increased attractiveness.

Table 7: Results of 26 univariate linear regression models



In addition, among 26 variables, the coefficients of 6 were not statistically significant at the 95% level. Although it was not expected that all facial features were statistically significant as we have seen that 5 of them could have as much predictive power as 26 variables in Model 2, one important feature in Model 2 – the ratio of face length to face width at eye level – was also not statistically significant in its univariate linear regression model. However, as mentioned in the Model Selection section, the variables could be insignificant simply because they did not have a linear relationship with the attractiveness rating. The variable importance from the random forest models should have a higher priority compared to these p-values, as the random forest models could find complex, non-linear relationships between the facial features and the image’s attractiveness rating, and thus was a better fit when we considered the effect of the facial features.

1. **Discussion and Conclusion**

To sum up, in contradiction with the hypothesis, averageness is not necessarily positively correlated with attractiveness. There have been other studies that disproved the averageness theory and argued that averageness often has confounding effects such as skin texture and facial bilateral symmetry that directly affect both facial averageness and facial attractiveness (Fink, Grammer, & Thornhill, 2001) (Grammer & Thornhill, 1994) (Mealey, Bridgstock, & Townsend, 1999) (Perrett et al., 1999). This provides a possible explanation for the results of this study, as here attractiveness is not positively correlated with attractiveness once the effect of averageness was isolated from symmetry and skin texture. Additionally, some studies pointed out that for females, average faces were considered attractive, but very attractive female faces usually have extreme features such as full lips, high cheekbones, and a small face (O’Toole et al., 1998) (Perrett et al., 1994, 1998). Future research should be conducted to confirm whether extremeness instead of averageness in specific facial features can increase attractiveness.

Regarding the facial traits that are most important for facial beauty, people with different races, gender, and age universally focus on eye features, lip/jaw features, and ratios between features of the face when they decide whether another person is attractive or not. This agrees with precedent studies that the most important facial features for attractiveness included eye size, lip fullness, and jaw length. This is also consistent with the observation in real life that while drawing comic books or creating animated movies, the authors tend to emphasize on eyes and lips for attractive characters, and usually gloss over details about nose, cheekbones, or eyebrows. However, people also focus on different smaller facial features when they judge others’ attractiveness, depending on their demographic background and the other person’s demographic background.

Diagram

Description automatically generated

Figure 11: Three features that are most important to attractiveness, regardless of demographics.

There are several other interesting implications from this research. First, females systematically get higher attractiveness ratings compared to males, both when the ratings come from males and when the ratings come from females. Second, people hit their peak beauty at the age of 20 to 30 years old, and the older people get, the less attractive they are perceived. However, older people tend to give higher attractiveness ratings compared to younger people. This irony made young adults consider people of 60+ years old to be least attractive, while people of 60+ years old considered young adults most attractive. And interestingly, people tend to not consider people from their age group to be the most beautiful, but rather all agree that young adults from 20-30 years old are the most attractive.

Lastly, Black people tend to give out much higher than average attractiveness ratings (0.59 compared to 0.485). There are gaps in average attractiveness ratings that people from different ethnicities give out, which means that some races systematically rate other people with higher or lower scores. However, the scores that all ethnicities receive are fairly equal with no big gap. This could explain why previous studies have claimed that there is a universal standard of beauty regardless of racial backgrounds (Little, Jones, & DeBruine, 2011), but according to the results of my work, this could very likely be consequences of some races tend to give out higher than average ratings, which cancels out when other races tend to give out lower than average ratings.

Given the number of observations, the low MSE in both training and testing set, and the comprehensiveness in the feature selection and data processing process, the results of Model 1 and Model 2 in this research are reliable. For some of the random forest models that were performed on subsets of the data, some do not have an adequate number of observations for such complex algorithms and thus result in overfitted models. For the univariate linear regression models, it could be argued that the coefficients’ signs were biased because the models did not control for important factors that could affect attractiveness such as make-up, accessories, and facial expressions. However, these factors did not influence the facial features, and thus could not create omitted variable bias. For studies that aimed at predicting attractiveness ratings, including them in the models could improve the predicting power, but for the purpose of this study, leaving them out would not affect the conclusions drawn from the coefficients.

In conclusion, this study has filled the academic gap in knowledge about facial attractiveness and averageness with computationally calculated averageness, highly accurate and interpretable random forest models, and a large training dataset with participants and images coming from different demographic backgrounds. This study also demonstrates the difference in beauty perception by people from varied demographics to people from varied demographics. It gives insights into how a demographic group perceives the attractiveness and facial traits of others that, to my best knowledge, have never been investigated before. This has real-world applications where customized and fake attractive actors, actresses, or holograms could be created to fit in with a specific demographic’s tastes based on the traits that they consider most important.

In the future, studies should be conducted to confirm whether attractiveness increases when some facial features are more average, and when some other facial features are more extreme. Furthermore, studies could also investigate the interactions in ratings between demographics, such as whether older men consider younger girls as more attractive while older women prefer men their own ages. As several random forest models were overfitted due to lack of data (the dataset only had 27 ratings for each of 2,206 images), future research could also duplicate this study with a bigger, more comprehensive dataset to confirm the findings. Last but not least, the package that was used to detect the facial landmarks of the images was pretrained and could not detect landmarks in the forehead region, thus created limitations in the feature selection process. Even though no study has claimed that features pertained in the forehead region are important to facial attractiveness, having these features will make the analysis more comprehensive, thus future research that utilizes a better, possibly self-trained image detection model is needed.

**References**

Bainbridge, W.A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face images. Journal of Experimental Psychology: General. Journal of Experimental Psychology: General, 142(4), 1323-1334. (<http://www.wilmabainbridge.com/facememorability2.html>)

Baudoin, J. & Tiberghien, G. (2004). Symmetry, averageness, and feature size in the facial attractiveness of women. *Acta Psychologica, 117*(3), 313-332. <https://doi.org/10.1016/j.actpsy.2004.07.002>

Choudhary, P., Agrawal, P., & Kaur, G*.* (2020). Review On SVM Based Method For Identification And Recognition Of Faces By Using Feature Distances. *International Journal of Scientific and Technology Research, 9*(1)

Darwin, C. (1896). The descent of man and selection in relation to sex, Volume 2. *D. Appleton & Co.*

Eisenthal, Y., Dror, G., & Ruppin, E. (2006). Facial Attractiveness: Beauty and the Machine. *Neural Computation,* *18*(1), 119–142. <https://doi.org/10.1162/089976606774841602>

Fink, B., Grammer, K., & Thornhill, R. (2001). Human (*Homo sapiens*) facial attractiveness in relation to skin texture and color. *Journal of Comparative Psychology, 115(1)*, 92–99. [https://doi.org/10.1037/0735-7036.115.1.92](https://psycnet.apa.org/doi/10.1037/0735-7036.115.1.92)

Gibson, C. & Jung, K. (2002). Historical census statistics on population totals by race, 1970 to 1990, for the United States, regions, divisions, and states. <https://www.census.gov/content/dam/Census/library/working-papers/2002/demo/POP-twps0056.pdf>

Grammer, K., & Thornhill, R. (1994). Human (Homo sapiens) facial attractiveness and sexual selection: The role of symmetry and averageness. *Journal of Comparative Psychology, 108*(3), 233–242. <https://doi.org/10.1037/0735-7036.108.3.233>

Kietzmann, J., Mills, A. J., & Plangger, K. (2020). Deepfakes: perspectives on the future “reality” of advertising and branding. *International Journal of Advertising*. <https://doi.org/10.1080/02650487.2020.1834211>

King, D. E. (2009). Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* *(10),* 1755-1758.

Koeslag, J. H. (1990). Koinophilia groups sexual creatures into species, promotes stasis, and stabilizes social behaviour. *Journal of Theoretical Biology, 144*(1), 15-35. <https://doi.org/10.1016/S0022-5193(05)80297-8>

Komori, M., Kawamura, S. & Ishihara, S. (2009). Averageness or symmetry: Which is more important for facial attractiveness? *Acta Psychologica, 131*(2), 136-142. <https://doi.org/10.1016/j.actpsy.2009.03.008>

Lal, M., Kumar, K., Arain, R.H., Maitlo, A., Ruk, S.A., & Shaikh, H. (2018). Study of Face Recognition Techniques: A Survey. *International Journal of Advanced Computer Science and Applications, 9*. <https://doi.org/10.14569/IJACSA.2018.090606>

Langlois, J. H. & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science, 1*(2), 115-121. <https://doi.org/10.1111/j.1467-9280.1990.tb00079.x>

Little, A. C., Jones, B. C., & DeBruine, L. M. (2011). Facial attractiveness: evolutionary based research. *Phil. Trans. R. Soc. B*, *366*, 1638–1659. <https://doi.org/10.1098/rstb.2010.0404>

Mealey, L., Bridgstock, R., & Townsend, G. C. (1999). Symmetry and perceived facial attractiveness: A monozygotic co-twin comparison. *Journal of Personality and Social*

*Psychology, 76(1)*, 151–158.  <https://doi.org/10.1037/0022-3514.76.1.151>

Moller, A. P. & Swaddle, J. P. (1997) Asymmetry, developmental stability, and evolution. *Oxford University Press*

Noftle, E. E., & Fleeson, W. (2010). Age differences in big five behavior averages and variabilities across the adult life span: Moving beyond retrospective, global summary accounts of personality. *Psychology and Aging, 25(1)*, 95–107. <https://doi.org/10.1037/a0018199>

O’Toole, A. J., Deffenbacher, K. A., Valentin, D., McKee, K., Huff, D., & Abdi, H. (1998). The perception of face gender: The role of stimulus structure in recognition and classification. *Memory and Cognition, 26*, 146–160. <https://doi.org/10.3758/BF03211378>

Perrett, D. I., Burt, D. M., Penton-Voak, I. S., Lee, K. J., Rowland, D. A., & Edwards, R. (1999). Symmetry and human facial attractiveness. *Evolution and Human Behavior, 20*, 295–307. <https://doi.org/10.1016/S1090-5138(99)00014-8>

Perrett, D. I., Lee, K. J., Penton-Voak, I., Rowland, D. A., Yoshikawa, S., Burt, D. M., Henzi, S. P., Castles, D. L., & Akamatsu, S. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature, 394*, 826–827. <https://doi.org/10.1038/29772>

Perrett, D. I., May, K. A., & Yoshikawa, S. (1994). Facial shape and judgments of female attractiveness. *Nature, 368*, 239–242. <https://doi.org/10.1038/368239a0>

U.S Census Bureau. Quick Facts. <https://www.census.gov/quickfacts/fact/table/US/PST045219>

**Appendix**

The codes are hosted on the author’s Github: [https://github.com/lamtran21/Data-Analytics-Senior Research](https://github.com/lamtran21/Data-Analytics-Senior%20Research). The data is available by request on the website <http://www.wilmabainbridge.com/facememorability2.html>

1. Due to the data privacy agreement, the 2,222 images used in modeling were not permitted to be published or presented in papers and posters. Instead, 49 other images were permitted for such purposes. [↑](#footnote-ref-1)