Rapport de Projet:
Analyse Exploratoire et Visualisation de Données

Etude de l'influence des traits du visage sur l'attractivité chez les célébrités

Introduction:

Nous avons choisis de travailler sur le dataset CelebA, classifiant les traits du visage des photos de célébrité de 1,45 Go datant de 2018. De très nombreux paramètres sont pris en compte, comme par exemple la taille du nez, la présence de maquillage, le genre, etc. pour un total de 41 attributs sur les 202599 photos de célébrités. Tous ces attributs sont renseignés de manière binaire, ils sont soit vrais ***(1)*** soit faux ***(-1)***. Un paramètre nous intéresse particulièrement pour notre traitement: "*Attractive*" renseigne si oui ou non la personne est considérée comme attirante. Bien que très subjectif et dépendant de la personne qui a classé les photos, nous concentrerons nos recherches sur la corrélation entre ce paramètre et tous les autres, dans le but d'analyser les traits les plus attrayants. A noter que ce dataset contient également des *bounding box* permettant de connaître où se trouve un certain attribut sur la photo, mais étant donné que les photos ne sont pas toutes prises de la même manière (orientation et cadrage du visage) nous avons choisis de ne pas aborder ces informations dans ce rapport.

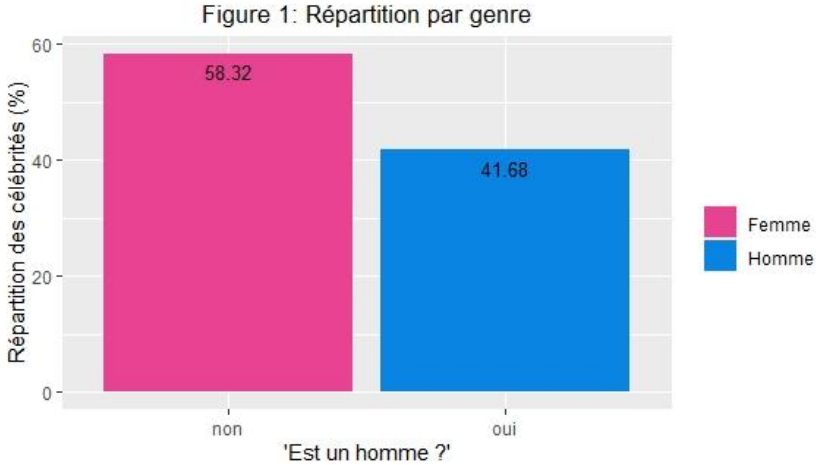
Voici des exemples d'attributs et de photos parmis les deux cent mille étudiées:

Note: Dans cette analyse, nous emploierons le raccourci "un attribut est plus ou moins attrayant", signifiant en réalité qu'il est plus ou moins fortement corrélé à l'attribut "Attractive" dans le dataset.

Analyse 1: Répartition du dataset par genre et selon l'attirance

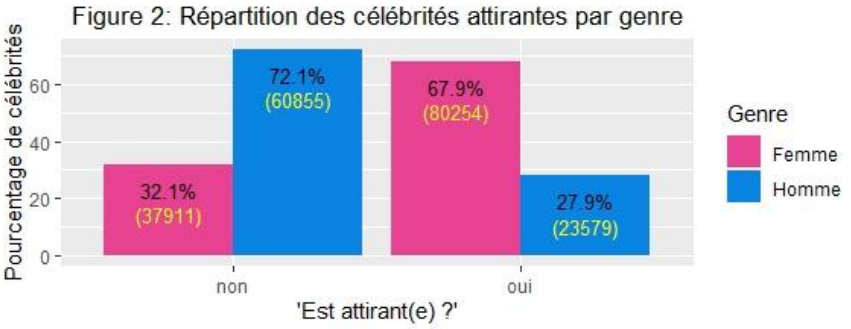
Pour pouvoir étudier de manière plus intéressante notre dataset, nous avons besoin de pouvoir différencier le genre des célébrités pour éviter les biais dû au fait que certains attributs typiquement féminins sont aussi renseignés pour les hommes et inversement, comme le maquillage. Aussi étant donné que nous réaliserons des études statistiques sur l'ensemble des données, nous avons besoin de connaître le nombre de personnes de chaque genre, pour savoir si l'un à plus de poids que l'autre dans nos statistiques.

La figure 1 montre la répartition par genre du dataset en pourcentage à gauche. On observe un plus grand nombre de photos de femme dans le dataset, ce qui pourra impacter certains résultats.



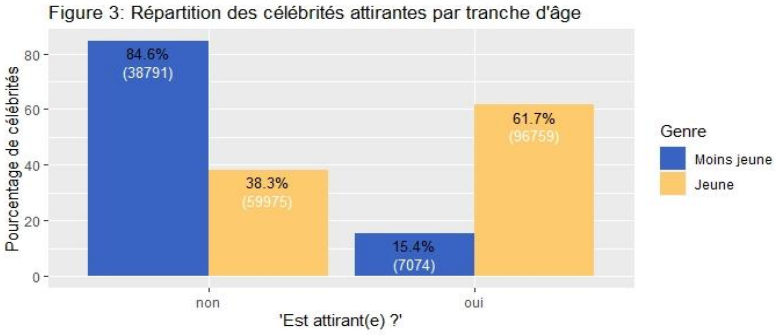
Ces écarts pourraient biaiser certains diagrammes réalisés sans distinction de genre et sont donc importants à garder en tête. Par exemple, nous verrons que les cheveux blonds sont généralement plus corrélés à l'attirance que les bruns, mais que cela est dû au fait qu'il y a plus de femmes. Si la population était égale, les cheveux bruns seraient probablement plus appréciés de manière générale.

La figure 2 nous permet d'observer la répartition de l'attirance en fonction du genre. On voit que généralement les femmes sont beaucoup plus souvent considérées comme attirantes que les hommes. Cela corrèle avec un fait de société observé, mais c'est également impacté par le fait que nous étudions des visages de célébrité.



Analyse 2: Impacte de l'âge sur l'attirance

L'âge des célébrités n'est pas directement renseigné dans notre dataset. Nous avons simplement un attribut binaire, et subjectif, qui permet de savoir si la personne a une apparence jeune ou non sur sa photo. La figure 3 montre si une personne est considérée comme attirante en fonction de sa catégorie d'âge.

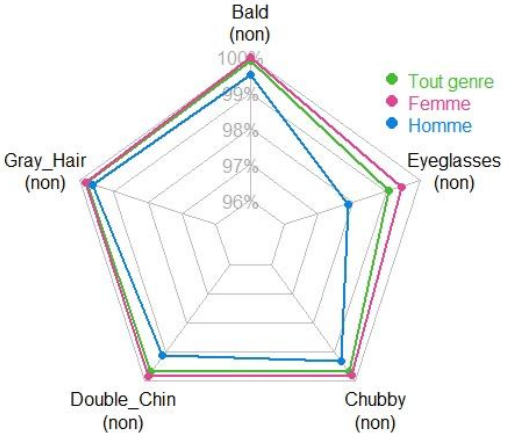


L'âge est un facteur très important, en effet près de 85% des personnes âgées ne sont pas marquées comme attirantes. A noter qu'il y a beaucoup plus de personnes jeunes que de personnes âgées dans le dataset.

Analyse 3: Traits du visage les plus corrélés à la beautée (première approche)

Dans cette partie nous analyserons quels attributs rendent une personne belle ou non. Nous ne nous intéressons pas aux caractéristiques qui ne sont pas des traits du visage, comme par exemple si la photo est floue ou non, ou si la personne sourit ou non.
Notre **première approche** à été de compter le nombre d'occurrence de chaque trait chez les personnes attirantes.

Figure 4: Pourcentage des célébrités attirantes par genre, parmis le top 5 des traits attirants (Tout genre)



La figure 4 montre les résultats donnés par cette méthode pour les 5 traits qui sont le plus courants chez les personnes attirantes. Les attributs les plus importants semblent en fait être ceux que l'on ne doit **pas** avoir. Être chauve, être en surpoids, etc. Près de 100% des personnes qui sont attirantes n'ont pas ces attributs, excepté pour les lunettes qui sont légèrement plus acceptées chez les hommes.

Cependant cette méthode d'analyse est biaisée. En effet, dans la base de données, la très grande majorité des personnes ne présentent pas ces caractéristiques (>95%) qu'ils soient attirants ou non. C'est donc normal que les personnes attirantes ne les présentent pas non plus, mais elles ne sont pas forcément corrélées. Mais étant donné que toutes les photos sont des célébrités, on peut aller plus loin et supposer qu'une sélection est faite dans la vraie vie, empêchant les personnes présentant ces traits de devenir célèbres et donc de figurer dans le dataset.

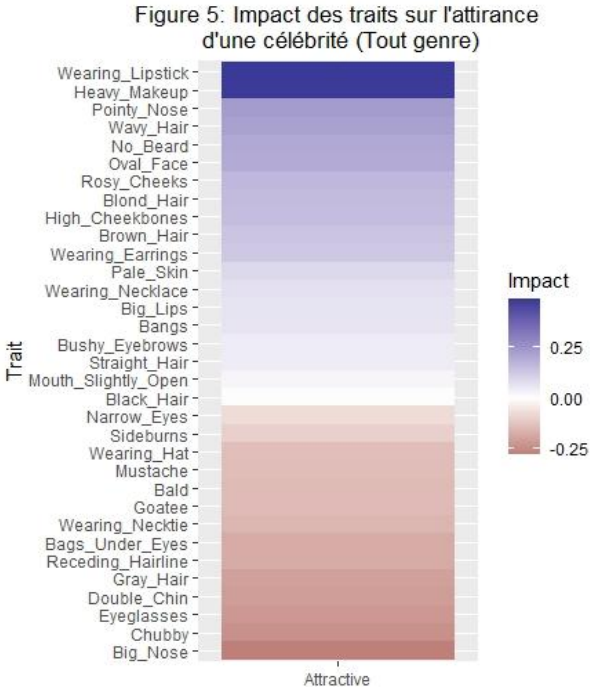
En analysant les traits les plus attractifs de chaque genre, on s'aperçoit mieux de la limite de cette approche. Par exemple ici chez les femmes, on remarque que les traits qui apparaissent le plus souvent chez les personnes attirantes sont l'absence de moustache, de calvitie, de barbichette etc., ce qui est évident puisque ces traits n'apparaissent jamais sur aucune des photos de femme.

	cols	value	count	percentage
	<chr>	<chr>	<int>	<dbl>
1	Mustache	non	80254	100
2	Bald	non	80253	100.
3	Sideburns	non	80251	100.
4	Goatee	non	80248	100.
5	Wearing_Necktie	non	80240	100.

Cette méthode permet donc d'apprendre quels traits sont peu présents dans le dataset, et d'avoir une idée des caractéristiques à éviter pour être attirant. Nous nous intéressons maintenant à une vraie méthode de corrélation entre chaque trait et l'attirance.

Analyse 4: Traits du visage les plus corrélés à la beautée (seconde approche)

Dans cette **seconde approche** nous nous intéressons à l'influence réelle de chaque trait sur l'attribut "*Attractive*", grâce à un indice de corrélation: les indices positifs (resp. négatifs) signifiant un impact positif (resp. négatif) sur l'attirance de la célébrité.



La figure 5 nous permet d'observer plus clairement quels attributs sont bons ou mauvais pour l'apparence d'une personne. Nous avons donc via cette approche un nouveau top 5 des traits influençant positivement l'attirance. De même, on retrouve dans les coefficients très bas les traits vus dans l'approche précédente.

Telle que la méthodologie employée précédemment, il serait judicieux ici d'appliquer cette approche à chaque genre, les traits n'étant pas communs à tous. Ainsi, nous pouvons observer dans la figure 6 les 5 traits ayant un plus grand impact sur l'attractivité d'une célébrité pour chaque genre. Le tout appliqué par genre afin de visualiser les différences d'influences.

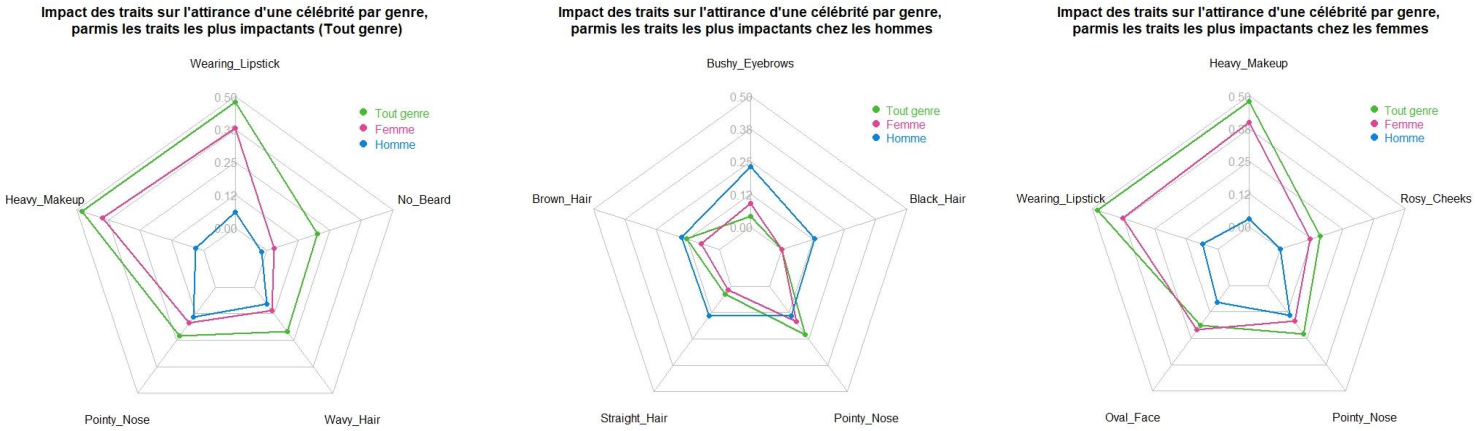


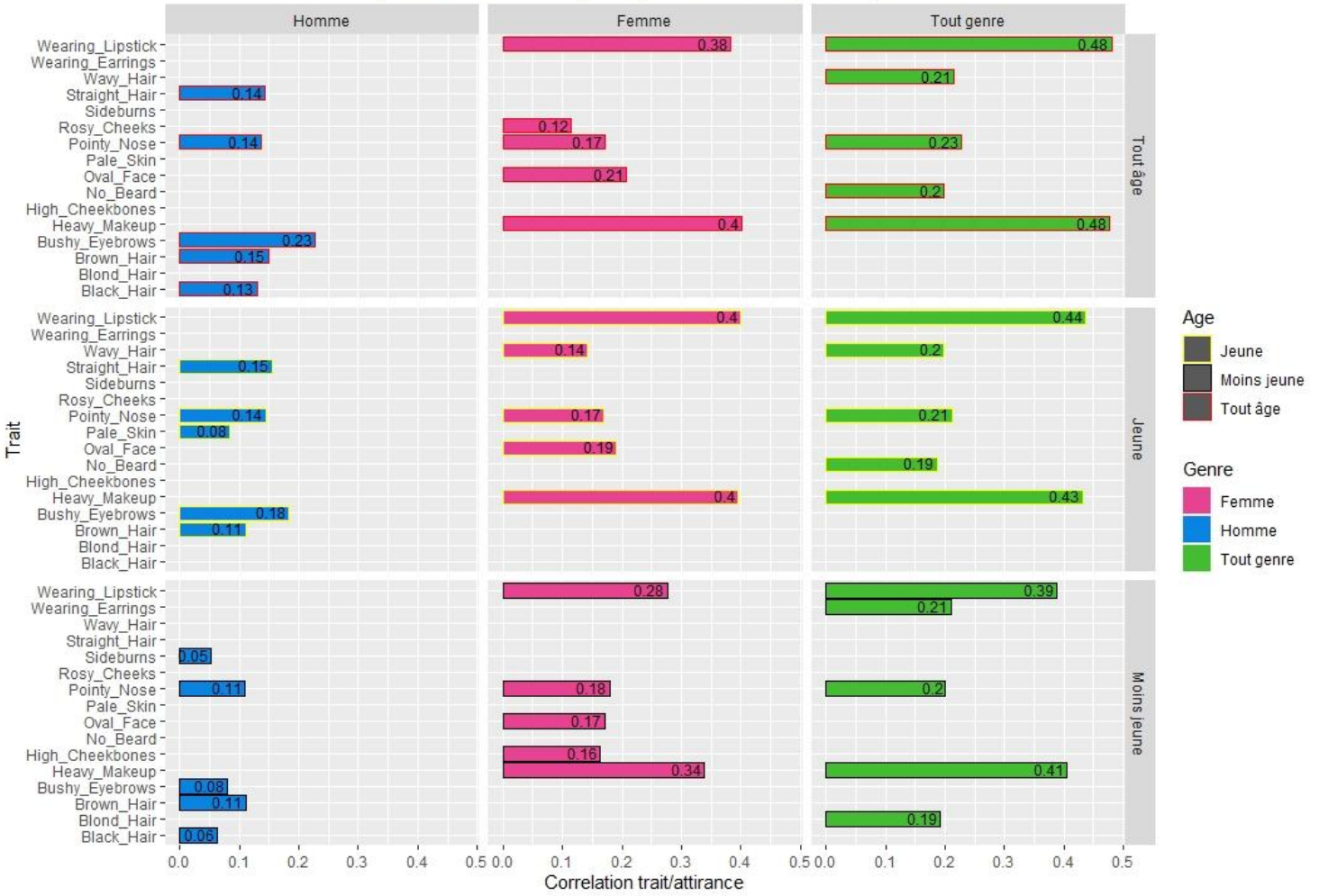
Figure 6: Top 5 des traits les plus attrayants pour chaque genre

Il est intéressant de noter ici que les meilleurs critères de beauté chez les femmes ont un impact considérable contrairement à ceux des hommes dont les traits sont moins corrélés à l'attirance. Cela signifie que peu de traits définissent la beauté chez les femmes tandis que les hommes peuvent être considérés comme attirants de bien plus des manières, comme le confirme l'étude Rating Attractiveness: Consensus Among Men, Not Women de 2009.

Le raisonnement sur le partitionnement des catégories afin de posséder plusieurs modèles (top 5) différents peut être poussé d'une dimension supplémentaire: les tranches d'âge. Jusqu'à maintenant, nous avons considéré les différents genres grâce à l'attribut "*Male*" (Tout genre/Homme/Femme). Il est donc possible de considérer les différentes tranches d'âge grâce à l'attribut "*Young*", nous donnant les catégories supplémentaires: Tout âge, Jeune, Moins jeune.

Ce raisonnement visible dans la figure 7 nous permet d'avoir 9 classements différents en fonction de groupes d'appartenance plus précis, permettant de mieux apprécier la fonction déterminant le taux d'attirance.

Figure 7: Impact des traits sur l'attirance d'une célébrité, selon son genre et sa tranche d'âge, parmis les 5 traits les plus impactants de chaque catégorie



Conclusion:

Grâce à cette analyse, nous avons pu déterminer quels sont les traits décisifs dans le caractère attrayant d'une célébrité, selon son genre et sa tranche d'âge. En effet, il a été montré lors de cette étude que les traits attrayants varient grandement selon le genre, tandis que de plus faibles variations ont été observées pour des tranches d'âge différentes. Cependant, les résultats obtenus dépendent fortement des données utilisées. Données pouvant être biaisées lors de la création du dataset. En effet, nous ne savons pas dans quelles circonstances l'attribut "*Attractive*" a été défini, celui-ci a pu être influencé par la carrière de la célébrité en question dans le cas où l'étape de labellisation a été effectuée manuellement, soit de manière subjective. De même pour l'attribut "*Young*" sur lequel nous n'avons pas d'information: âge fixe ? / défini par une ou plusieurs personnes ? Beaucoup de questions sur ces données restent sans réponse.