Visualisation du dataset Speed Dating

Cédric Bousquet

Semaine 2 – mini projet – Formation Fullstack Jedha Bootcamp

Objectif

- Visualiser le jeu de données "Speed Dating"
  - Constitué par les professeurs Ray Fisman et Sheena Iyengar de la Columbia business
  - Disponible sur Kaggle
- Identifier les variables qui permettent de prédire un rendez-vous suite à la rencontre entre deux participants
  - Par exemple, identifier les caractéristiques du partenaire qui le rendent attirant, et savoir si des centres d’intérêts en commun peuvent faciliter le RDV

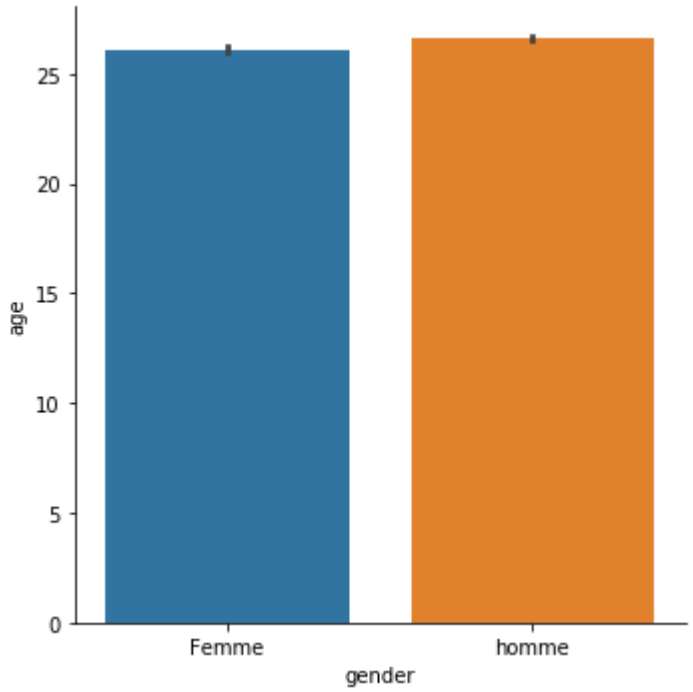
Matériel

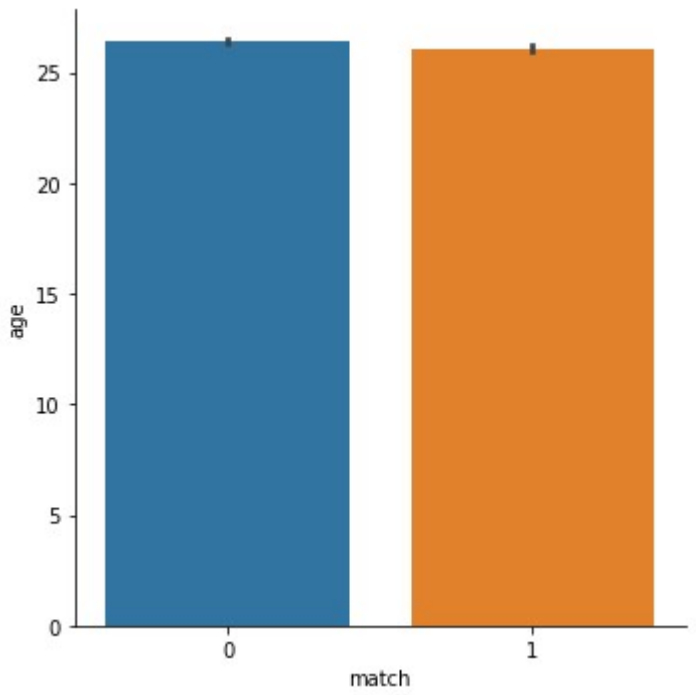
- 8378 lignes, 195 colonnes
- 551 participants (277 hommes et 274 femmes)

- Difficile de visualiser des relations entre toutes les variables
  - Identification des variables explicatives les plus corrélées avec la variable à prédire

Age et sexe des participants

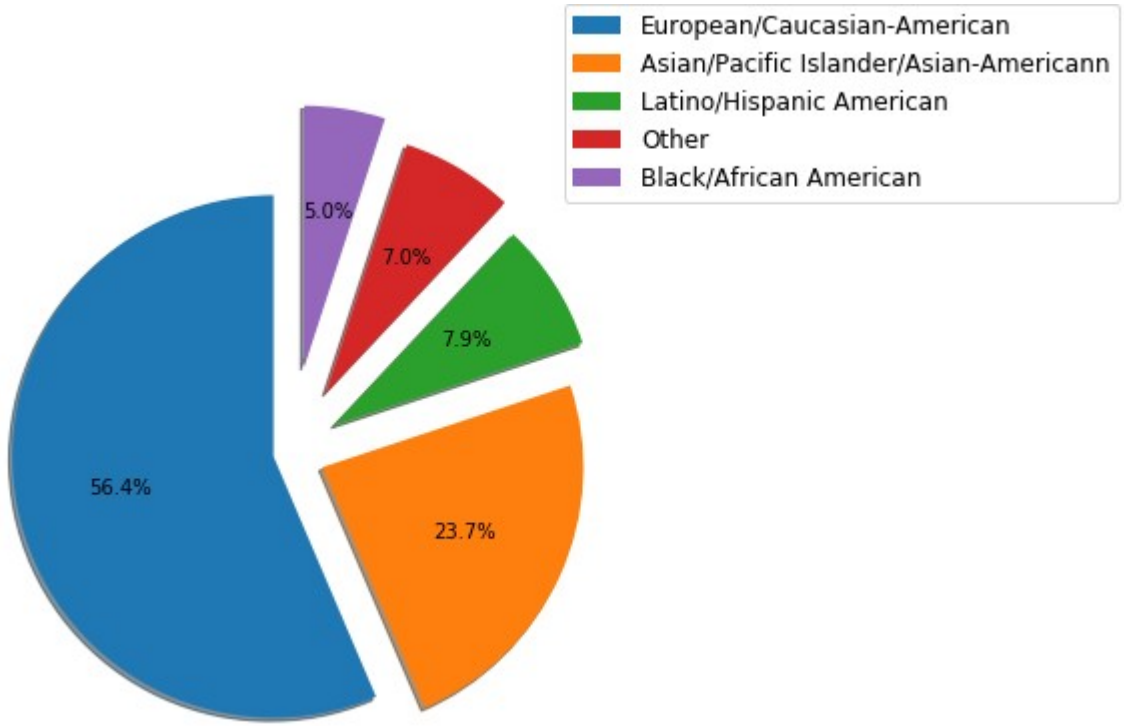
- Peu d'effet du sexe ou de l'âge du participant sur la décision conjointe d'un nouveau RDV



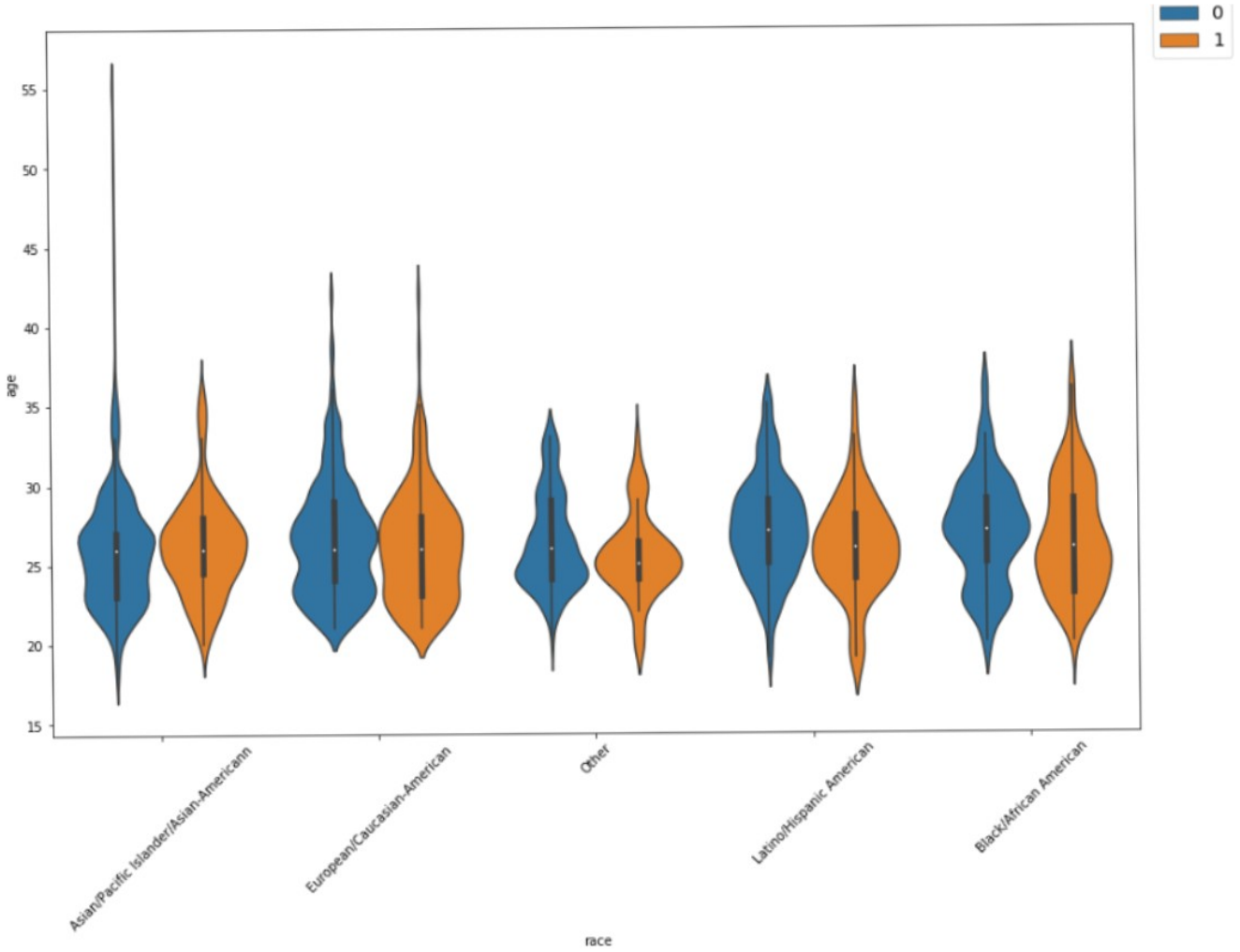


Race des participants

- On note une grand prédominance des types européens/caucasiens-américains



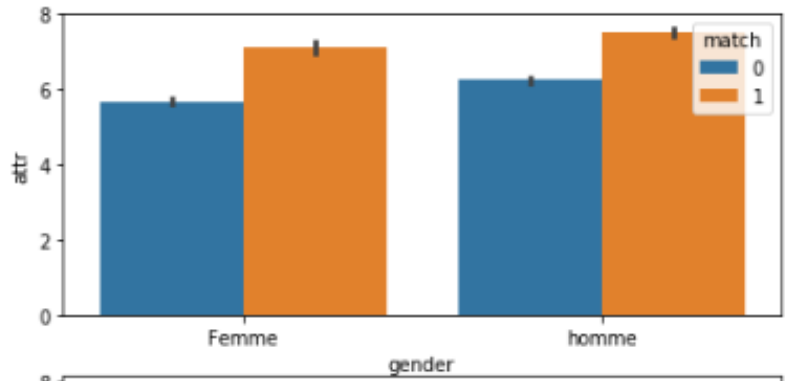
Pas d'effet explicatif de la race des participants

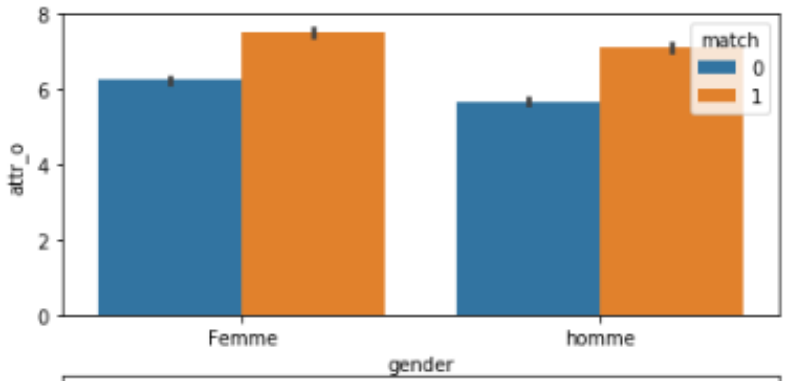


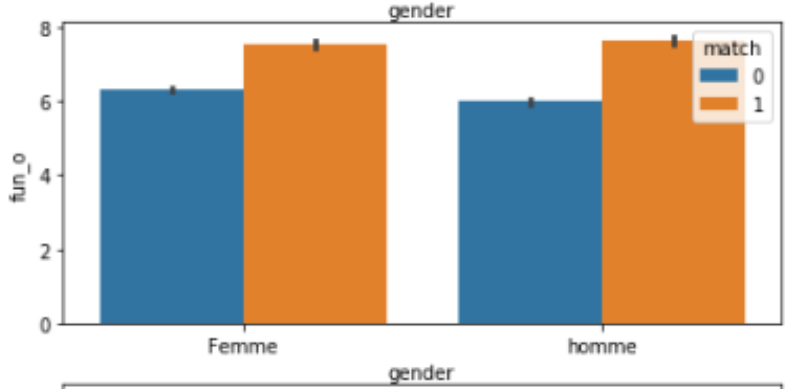
Heatmap des variables les plus corrélées

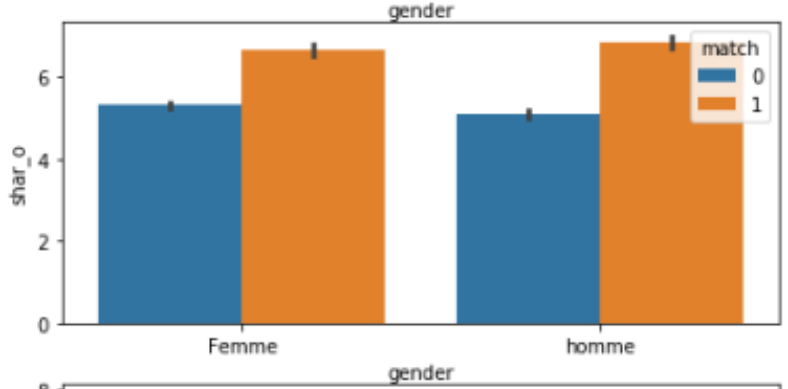
- Sélection des variables avec coefficient de corrélation > à 0,15 avec match

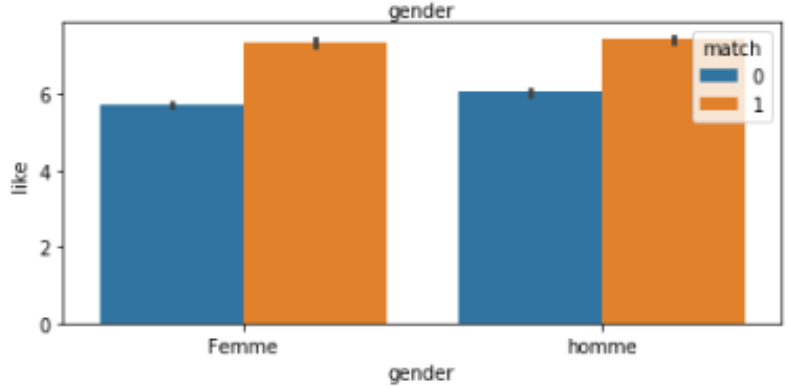
Effets différents des variables en fonction du sexe des participants

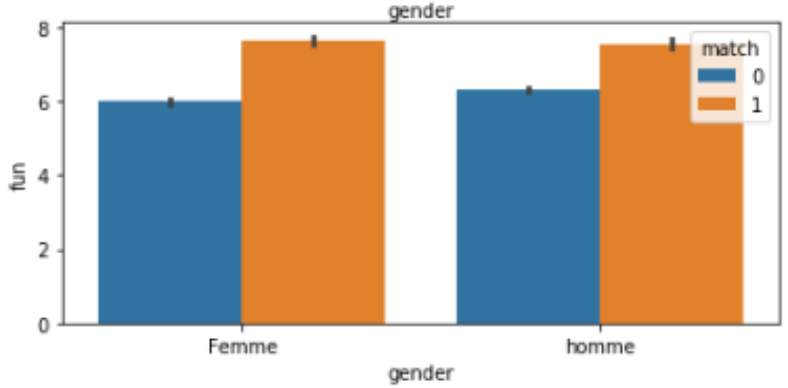












Conclusion

- Les variables attr, attr_o, fun_o, shar_o, like, fun, dec et dec_o sont les plus corrélées avec la variable à prédire
- Les effets de ces variables sont différents en fonction du sexe des participants
- L'effet du sexe, de l'age ou de la race des participants semble peu significatif