**Réponses test technique Deloitte**

• Identifier les champs qui ont des données manquantes. Quelles stratégies préconisez-vous pour imputer les données manquantes.

*Réponse :* Initialement je comptais soit retirer les valeurs « NaN » soit leur imputer la valeur de la moyenne/médiane si les variables en question étaient numériques.  
 Il s’est avéré qu’elles représentaient au moins 1/3 du dataset et étaient toutes catégorielles.  
J’ai donc fait el choix de les conserver pour ne pas perdre trop d’information sur le dataset, elles ne seront pas prises en compte dans le plot des variables en question de part la propriété intrinsèque des plots sns.

• Identifier les valeurs extrêmes des variables. Que pouvez-vous faire pour normaliser les valeurs ?

*Réponse :* Initialement je comptais appliquer une méthode type score Z/ score Z modifié pour traiter les outliers. Or la distribution n’étant pas uniforme, la méthode score Z n’était pas applicable. J’ai préféré faire appel à mon sens « physique » des variables et déterminer des intervalles qui ont un sens (exemple : année du véhicule comprises entre 1950 et 2023 car aujourd’hui nous ne trouvons plus de véhicules datant d’avant 1950 donc toutes valeurs inférieures est enlevée). Bien que cette méthode soit subjective je pense qu’elle est plus pertinente lorsque l’on peut s’approprier les données.

• Pouvez-vous clusteriser les véhicules ? Quel est le profil de chaque cluster ?

*Réponse :* Initialement je pensais utiliser un algo de ML type K-mean/K-mode ou bien DBSCAN de clustering. Je n’ai pas eu le temps de travailler dessus et est simplement étudié les données par « groupby » de marque.   
Je ne suis cependant pas certain d’avoir entièrement saisi le sens de la question et d’avoir répondu à celle-ci comme attendu.

• Quelles sont les données corrélées avec la variable « price » ?

*Réponse :* Initialement je pensais calculer la matrice de corrélation des variable entre elles. En utilisant la fonction df\_train.corr() j’ai pu observer que les variables les plus corrélées étaient : year of Registration, kilometer, powerPS, vehicleType(suv, coupe, cabriole)

• Comment utilisez-vous les variables catégorielles ?

*Réponse :* Initialement je pensais utiliser le one hot encoding pour les transformer en variable bianires dites « dummy ». C’est ce que j’ai fait de part l’implémentation de la fonction « indicatrice ».

• Est-ce qu’il est possible de réduire le nombre de variable ?   
*Réponse :* Initialement je pensais utiliser des algorithmes de ML tels que PCA ou tree pour les variables catégorielles. Par manque de temps j’ai privilégié une méthode plus simpliste qui consiste à réfléchir sur le réel impact d’une variable concernant notre étude de cas (exemple : la variable « monthregistration » n’a et n’est censée avoir aucun impact sur le prix je décide donc de la retirer. Pour les variables dont ce raisonnement n’est pas évident j’ai étudié leur corrélation avec le price afin de les retirer si cela n’était pas pertinent.

• Quelles sont les variables importantes lors du pricing de la voiture ?

*Réponse* : brand, year of Registration, kilometer, powerPS, vehicleType(suv, coupe, cabriole) sont les variables les plus importantes lors du pricing de la voiture d’après leurs corrélations avec le prix.

• Est-ce qu’il existe d’autres axes d’analyse à considérer ?

*Réponse :* J’aurai aimé investiguer un peu plus sur deux autres notions :  
 le postalCode qui détermine la région de la provenance de la voiture. A savoir qu’il existe éventuellement des disparités financières dans des régions d’Allemagne qui pourraient expliquer une hausse des prix à certains endroits. Cette hausse est soit due au fait que dans la région plus fortunée les modèles de voitures sont plus luxueux ou bien que pour un même modèle, l’acheteur est prêt à payer un peu plus cher car il peut se le permettre de part son niveau de vie.

Le second axe d’étude que j’aurai aimé mener concerne la variable « days\_between » que j’ai créée qui correspond à la différence entre dateCrawled et lastSeen. Celle-ci traduit le nombre de jours de parution de l’annonce avant achat. Bien qu’apriori cette information peut sembler inutile pour un vendeur afin de déterminer le prix de sa voiture, il n’en n’est pas moins :

Si la variable possède une valeur faible, cela veut dire que la voiture en question est soit rare donc attrayante soit possède un prix en dessous du marché par rapport à ses caractéristiques. Ainsi un vendeur pourrait se baser sur cette variable pour augmenter le prix d’une voiture de même modèle partit trop vite car pas assez chère.

Modèles ML : Par manque de temps je n’ai pas pu implémenter les 3 modèles demandés. En revanche les variables X\_train et y\_train sont définies. J’aurai utilisé un modèle de régression linéaire en utilisant les tests de mse, r² et p-value pour m’assurer de la cohérence et viabilité des données.