Point de contrôle qualité	Problème observé et commentaire	Correction : comment ? importance ?
1. le jeu de données est difficilement accessible (format “image”, PDF, HTML...), le fichier est mal formé	Le jeu de données est facilement accessible, puis qu’il est au format csv, format non-propriétaire.	Rien à signaler.
2. la licence est absente ou inhabituelle, le jeu de données n’est pas “open data”	Le jeu de données est open data : on n’a pas besoin de s’inscrire à Github pour accéder au fichier contestants.csv. Cependant, il n’y a pas de licence puisque c’est une initiative de fans de l’Eurovision. En revanche, dans le readme du repository, il est notifié la source des données scrapées (« The metadata and voting data are provided by the EurovisionWorld fansite. ») et il est également possible de télécharger les fichiers python pour reproduire l’ensemble des jeux de données (« You can download the entire dataset using the scraping code included in this repository. »). Par contre, il est demandé de citer les auteurs du jeu de données.	Ce n’est pas une plateforme d’open data, mais on voit qu’il y a une volonté de sérieux de la part des personnes à l’initiative du projet.
3. le fichier fait peu appel aux standards répandu et aux données pivots	Ici on a du csv, ce qui correspond à ce qui est répandu et recommandé (format non-propriétaire).	Rien à signaler.
4. le fichier est mal documenté	Le fichier est documenté : en effet, il est possible d'avoir une explication sur les champs du fichier dans le read me du repository.	Dans le cadre du projet, on remarque qu'il y a des colonnes qui seront à supprimer (par exemples, celles concernant les demi-finales puisque les pays comme les Big Five et le pays organisateur ne participent pas aux demi-finales).
5. il existe des problèmes de syntaxe	Il existe quelques problèmes de cohérence : tout d’abord, dans la colonne « to_country_id », on remarque qu’il y a parfois l’id (qui lui n’est pas documenté) du pays pour lequel le représentant concoure, ou	On supprimera cette colonne et on se basera sur la colonne « to_country ». On prendra des valeurs

	le nom écrit directement. Au niveau des colonnes concernant les points obtenus et la place finale (que cela soit en demi-finale ou en finale), il y a aussi un souci de cohérence : parfois on a des nombres décimaux (15.0 pour 15ème place) et parfois on a des nombres entiers (15).	décimales pour les points, mais des valeurs entières pour la place obtenue.

Point de contrôle qualité	Problème observé et commentaire	Correction : comment ? importance ?
6. Valeurs aberrantes, suspectes, inexplicables, pas crédibles	A part le fait qu'on aurait dû gagner avec Barbara Pravi, non rien d'aberrant. La seule colonne qui m'a fait tiquée est « sf_num » car je voyais pleins de 1.0 / 2.0 mais c'est en fait le numéro de la demi-finale à laquelle le participant a pris part et c'est expliqué dans la documentation. De plus, il y a eu parfois au début plusieurs gagnants et des ex-aequo donc il est normal d'avoir parfois plusieurs gagnants pour les premières éditions du concours.	Rien à signaler.
7. Il manque des données et cela n'est pas documenté (trous, données tronquées, valeurs vides, granularité / fréquence / maillage / fraîcheur)	Oui, il y a beaucoup de trous et ce n'est pas expliqué. Cependant, si on connait un peu le principe de l'Eurovisions ces trous sont explicables. Pour la colonne « sf_num », il y a des trous car pendant très longtemps il n'y a pas eu de demi-finales car il y avait très peu de pays qui participaient (surtout dans les premières années du concours). Pour « running_sf » et « running_final », il y a des trous car tous les pays ne prennent pas part aux demi-finales (c'est le cas des cinq pays fondateurs de l'Eurovision, les Big Five, dont fait partie la France mais aussi du pays organisateur) ni aux finales (si le pays perd en demi-finale, il ne peut pas participer en finale). Pour les colonnes liées aux points attribués par le jury et le télévote, les trous s'expliquent de différentes manières : le site qui sert pour la captation des données n'a pas toutes les données, notamment des premières années. Des plus, le télévote n'existait pas au départ. On voit aussi des écarts dans le nombre de points qu'obtiennent par exemple les gagnants : c'est parce qu'il y a entre autres maintenant beaucoup plus de pays participants qu'au début. Les données sont fraîches car contiennent la dernière édition de mai 2023.	Ne pas se baser sur les colonnes relatives aux nombres de points et aux demi-finales.
8. Trop de données : doublons, inutilement vieilles, précision /	Le fichier contient 1735 lignes. Il n'y a pas de doublons : par exemple, il y a des chanteurs (Mahmood, Sergey Lazarev etc.) qui ont participé à plusieurs éditions. Par contre il y a des données inutiles, type	Supprimer la colonne lyrics et youtube.

fréquence / maillage / fraîcheur	lyrics (paroles) : à part si on étudie les chansons gagnantes et qu'elles seraient dans la même langue (ce qui n'est pas le cas, donc il faudrait les traduire), les paroles sont difficilement exploitables. Idem pour les liens youtube.	
9. Données posant problème avec la réglementation (données perso, relatives à la santé, la religion..., propriété littéraire et artistique, etc.)	Les informations dans ce jeu de données sont publiques : le nom des chanteurs, de la chanson, les paroles, les données liées à l'ordre de passage, les points etc… sont fournis par le concours de l'Eurovision lui-même.	Rien à signaler.
10. Les contenus posent problèmes : synonymies, non traduits (USA), cryptique (DAECPP), utilisation du 0 au lieu du "null"...	Le seul souci repose sur les id des pays qui n'est pas défini, mais on peut se passer de cette colonne grâce à la colonne « to_country ».	Utiliser la colonne « to_country ».