* + 1. Exploratory Data Analysis (EDA)
       1. Objectif :
* Classer les observations comme étant une galaxie, un trou noir ou une étoile.
  + - 1. Checklist de base
         1. Analyse de forme :
* **Variable Target :** class
* **Lignes et colonnes :** 100 000 lignes, 18 colonnes
* **Types de variables :** 17 Quantitatives **(**10 float, 7 Int), 1 Qualitative (Object 🡪Target)
* **Analyse des valeurs manquantes :** Aucune valeurs manquantes
  + - * 1. Analyse de fond :
* **Visualisation de la Target : «**GALAXY » : 59.445%, « STAR » : 21.594%, « QSO » 18.961%
* **Signification des Features/Elimination des variables inutiles :**

Pour les variables de type **float** :

* Distribution symétrique pour « alpha »,
* Distribution asymétrique pour « obj\_ID », « delta », « u », « g », « r », « i », « z », « spec\_obj\_ID », « redshift »
* Il semblerait que des **outliers** soient présents pour les variables « u », « g », et « z » (on affiche les valeurs logarithmiques pour mieux voir)

Pour les variables de type **int** :

* Distribution symétrique pour « cam\_col »
* Distribution asymétrique de « field\_ID », « run\_ID », « plate » et « MJD », « fiber\_ID »
* « Rerun\_id » n’a qu’une seule valeur, donc inutile

Variables susceptibles d’être inutiles : “obj\_ID”, “run\_ID”, “rerun\_ID”, “cam\_col”, “field\_ID”, “spec\_obj\_ID”, “fiber\_ID” car ce sont des identifiants.

* **Relation Features/Target :**

Pas de grandes différences des taux sur delta, alpha, field\_ID, et fiber ID

« rerun\_ID » à exclure car une seule valeur donc inutile

Légères variations pour obj\_ID, cam\_col, run\_ID

Variations tres importantes pour u, g, r, i, z, spec\_obj\_ID, redshift, plate et MJD

* **Relation entre les variables**

**Les outliers de u et g ne nous permettent pas d’avoir une matrice de corrélation claire**

« r », « i » et « g » semblent très corrélées

MJD et spec\_obj\_ID semblent très corrélées

MJD et plate semblent très corrélées

Obj\_id et run\_ID sont parfaitement corrélées (donc pas bon)

Spec\_obj\_ID et plate sont parfaitement corrélées (donc pas bon)

« cam\_col », « rerun\_ID » et « fiber\_ID » n’apportent aucune informations intéressantes

* + - 1. Conclusion Initiale

Aucunes données manquantes

Exclusion de données aberrantes (33,5%)

Il semblerait que plusieurs features soient corrélées donc certaines seront exclues pour ne pas influencer la prédiction de par leur lien

Pas mal de variables pourrait être discriminante de la Target

* + 1. Analyse plus détaillée
       1. Relation Variables / Variables

Relation entre r, i et z supérieure à 0.9

Relation entre MJD et plate supérieur à 0.9

* + - 1. NaN Analyse

Pas de NaN

* + - 1. Hypothèses nulles (H0) :
* Les taux différents de redshift sont significativement différents pour les galaxy, star et qso
  + 1. Preprocesing