GROUPE ISI

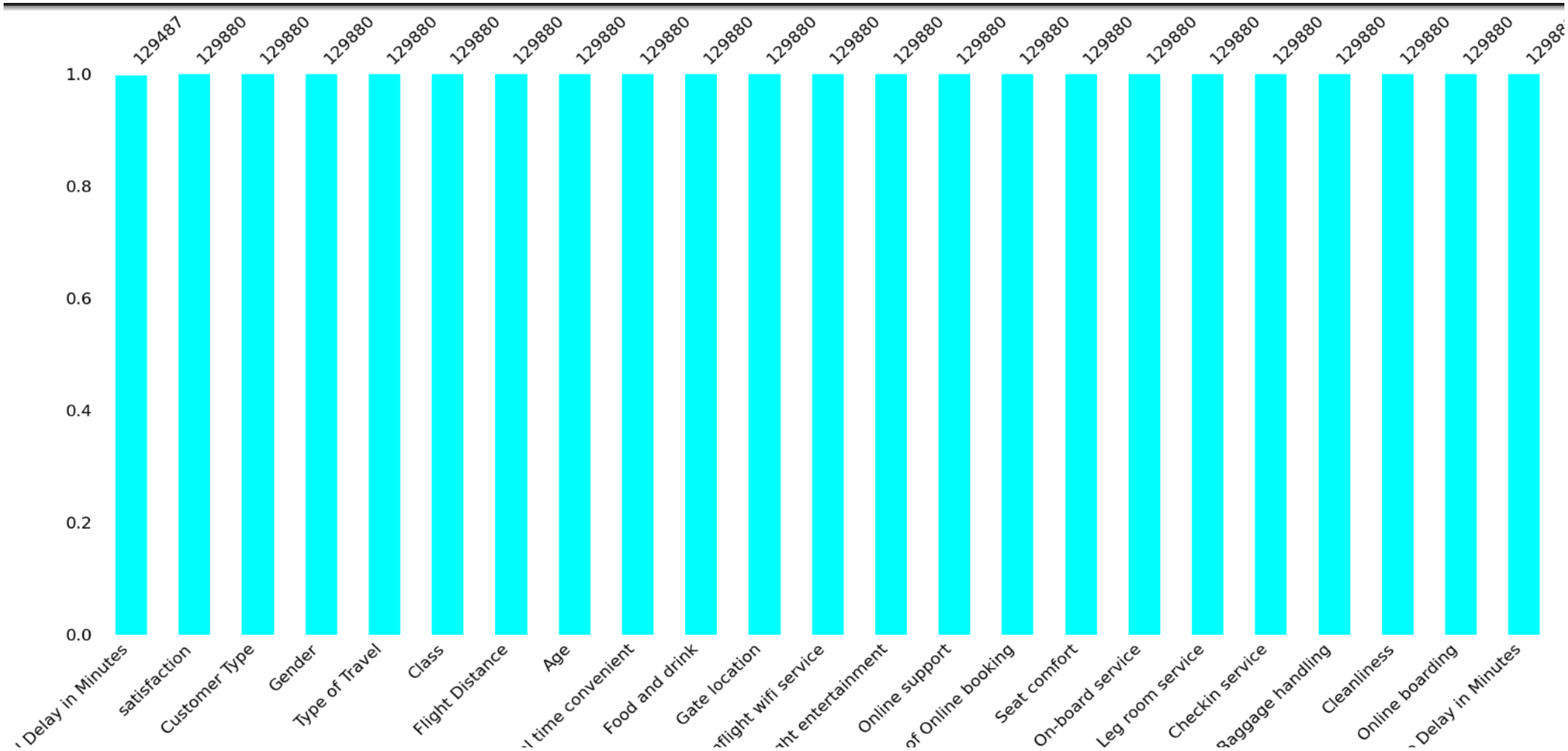
Institut Supérieur Informatique

Master I en Data Science & IA

Projet de Machine Learning

Analyse Prédictive et Visualisation des Données sur la Satisfaction Client

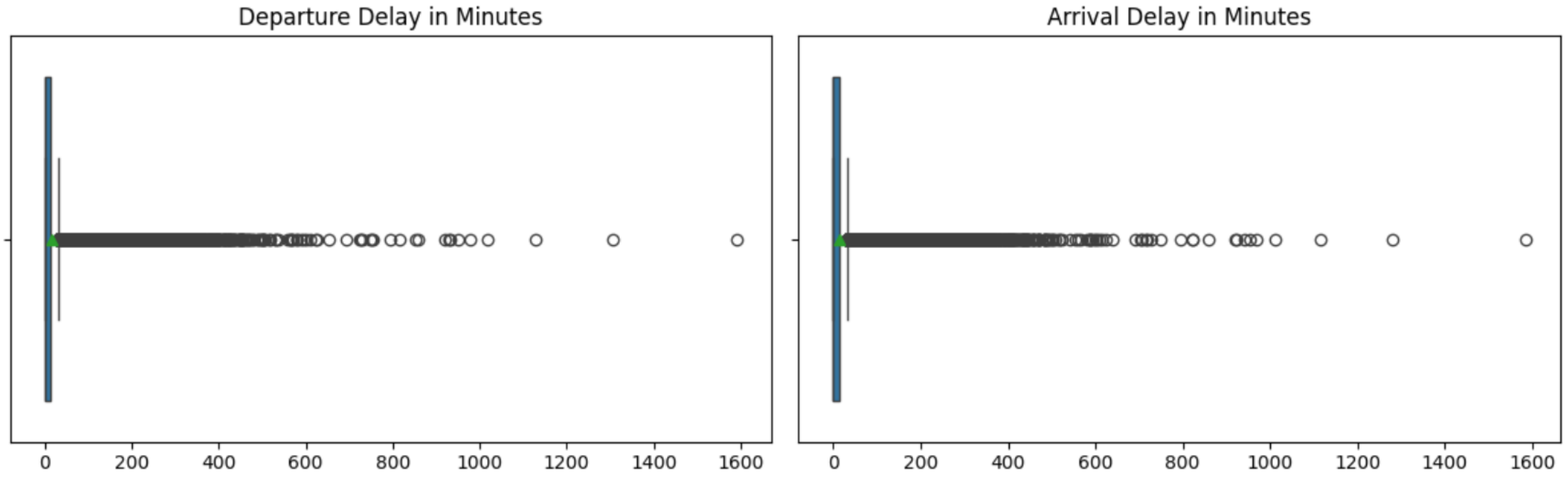
Gestion des valeurs manquantes



Interpretation →

Cette figure nous permet de voir les variables qui ont des valeurs manquantes, et seule la variable Arrival Delay in Minutes contient 393 valeurs manquantes

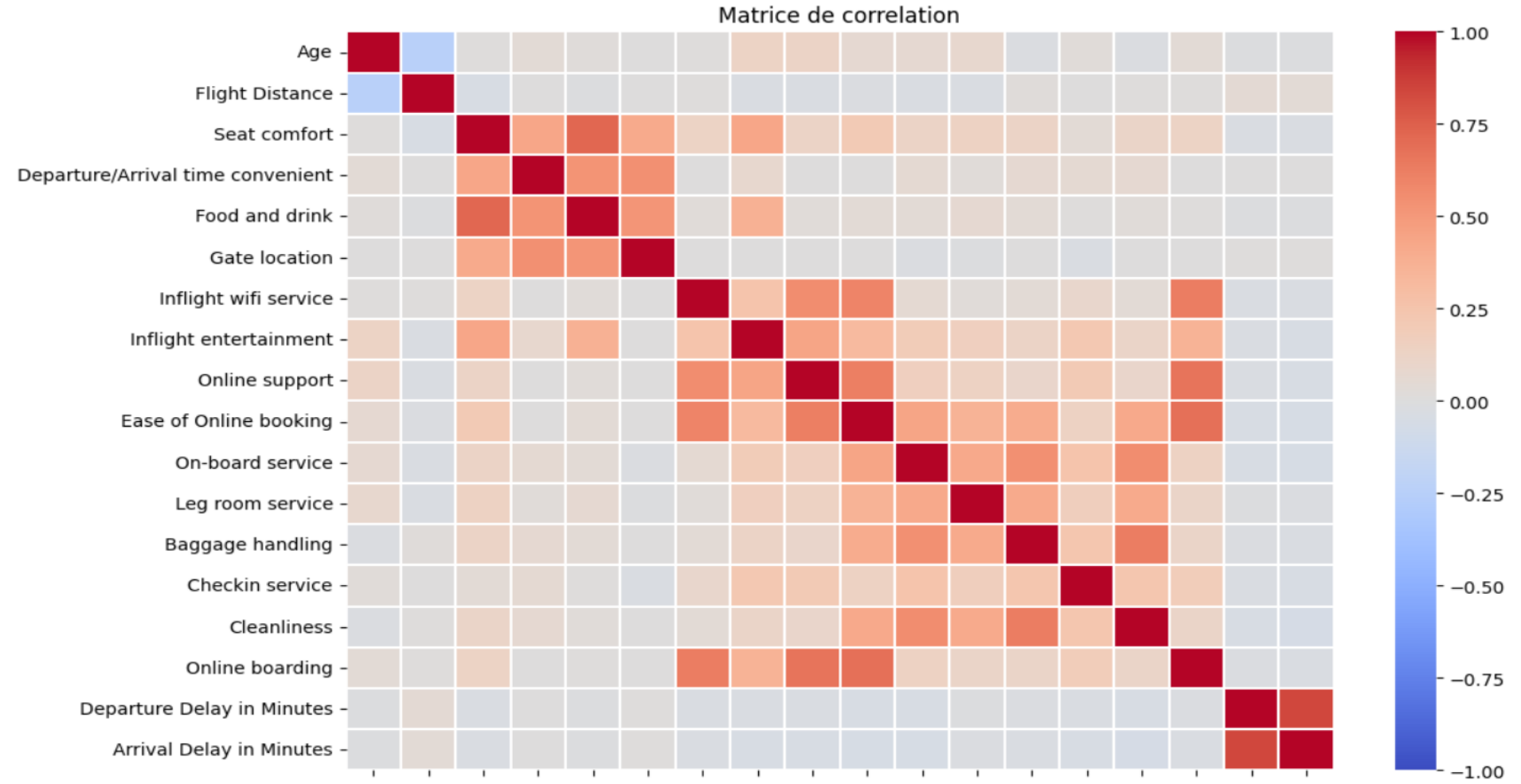
Gestion des valeurs aberrantes



Interpretation →

D’apres les graphique qu’on on constate qu’il y a pleusieurs variables qui ont des valeurs aberrantes. Mais on peut tous les representer ici, on a choisi juste celles qui trop des valeurs aberrantes.

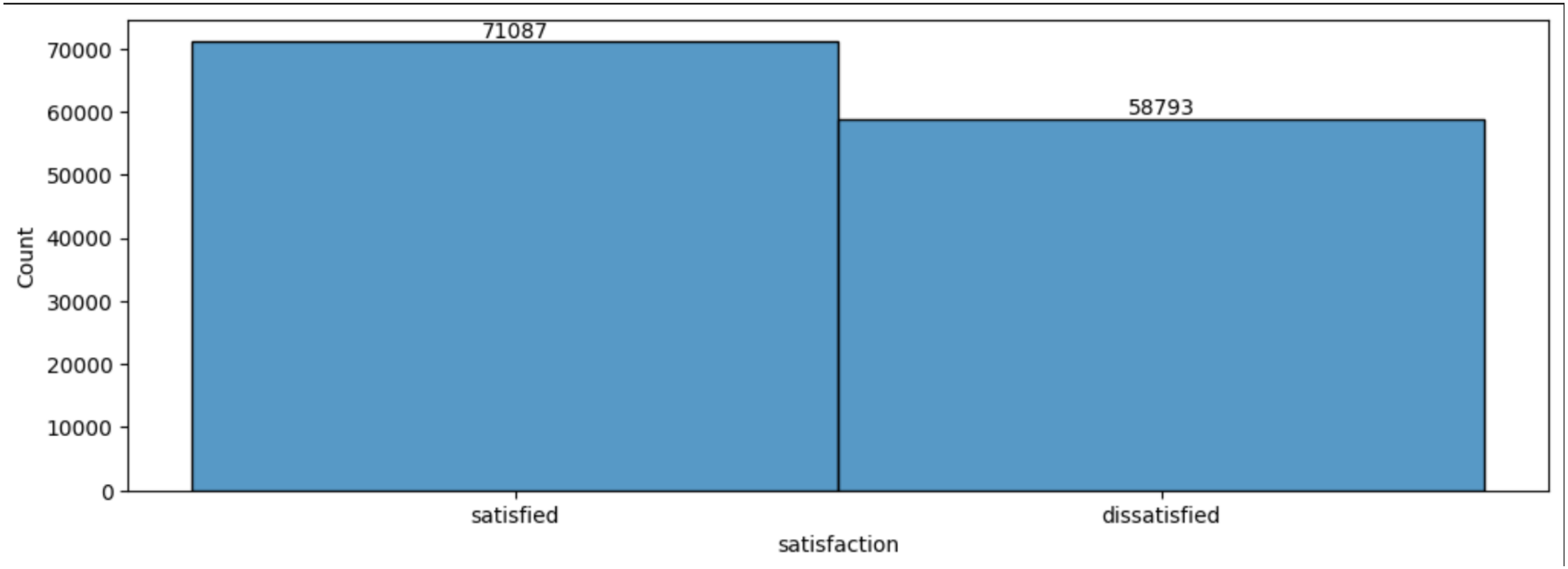
corrélations entre les variables



Interpretation

- Corrélations positives fortes : Les services en vol (comme le confort des sièges et la gestion des bagages) sont fortement corrélés entre eux, indiquant que les améliorations dans un domaine sont souvent associées à des améliorations dans d'autres.
- Corrélations faibles ou neutres : L'âge des passagers et la distance du vol n'ont que peu ou pas de corrélation avec d'autres aspects du service, suggérant un impact limité sur l'expérience globale
- Corrélations négatives ou absentes : Très peu de corrélations négatives sont observées, ce qui implique une absence de conflits notables entre les différents aspects du service.

Distribution des données pour les différentes classes



Interpretation →

On a constate que 55% des clients sont satisfaits et 45% le contraire.

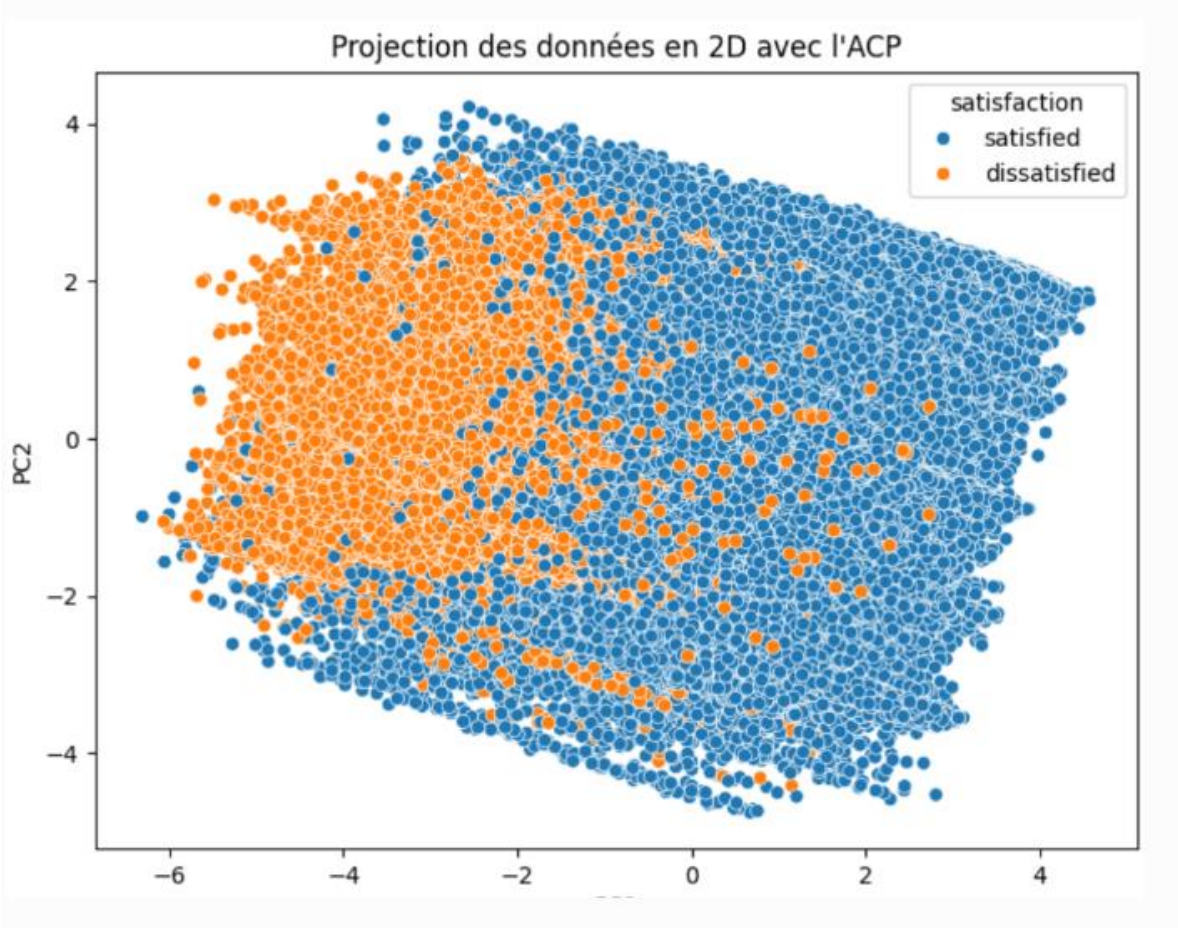
Documentation du pipelines du prétraitement des donnees

```
def preprocess_data(df):
"""
Prétraitement d'un DataFrame pour la modélisation.
    Cette fonction effectue les étapes suivantes :
    1. Gestion des valeurs aberrantes :
        - Identification des valeurs aberrantes à l'aide de l'intervalle interquartile.
        - Remplacement des valeurs aberrantes par les bornes de l'intervalle.
    2. Transformation des variables numériques :
        - Imputation des valeurs manquantes par la médiane.
        - Standardisation des variables pour une moyenne nulle et une variance unitaire.
    3. Transformation des variables catégorielles :
        - Encodage one-hot pour représenter les catégories sous forme de variables binaires.
    4. Création d'un nouveau DataFrame avec les données prétraitées.

    Args:
        df (pd.DataFrame): DataFrame à prétraiter.

    Returns:
        pd.DataFrame: DataFrame prétraité. """
```

Reduction de dimensionnalité avec l’ACP



Interpretation →

Les points représentant les clients satisfaits (bleu) et insatisfaits (orange) se chevauchent beaucoup sur le graphique. Cela signifie qu'il n'y a pas de frontière nette entre les deux groupes. Pas de cluster nette.

Rapport de comparaison des performances de chaque modèle

	KNN	NAÏVE DE BAYES
F1 Score	0.92	0.83
Precision	0.92	0.82
Rappel	0.91	0.83
Accuracy	0.94	0.84

	MSE	RMSE	R2
Regression Lineaire	0.12	0.35	0.48

Vu les resultats qu'on a dans les tableaux ci-dessus , KNN est le modèle le plus performant avec une précision, un rappel et un F1-score plus élevés que Naïve Bayes, ce qui en fait le modèle idéal pour ce problème de classification. La Régression Linéaire semble moins adaptée si la satisfaction client n'est pas directement liée de manière linéaire aux variables explicatives.

Evaluation, Régularisation et Optimisation du modèle

	KNN	NAÏVE DE BAYES
F1 Score	0.92	0.83
Precision	0.92	0.82
Rappel	0.91	0.83
Accuracy	0.94	0.84

	MSE	RMSE	R2
Regression Lineaire	0.12	0.35	0.48

Après avoir optimise et regularise les models, on constate qu'on a les meme évaluation avant et après optimisation.