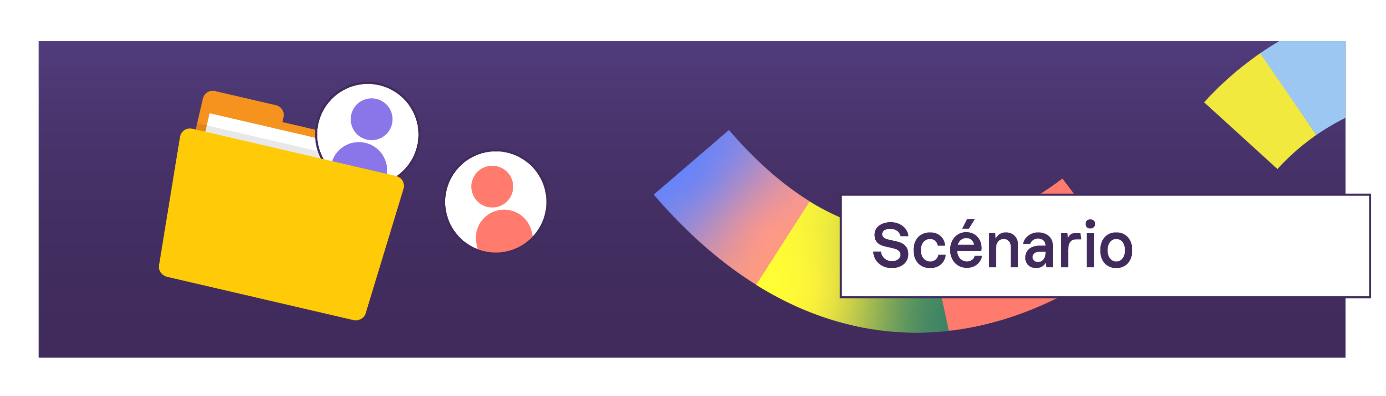
**Scénario du projet**



Vous êtes consultant Data Analyst dans une entreprise spécialisée dans la data. Votre entreprise a décroché une prestation en régie au sein de l’Organisation nationale de lutte contre le faux-monnayage (ONCFM).

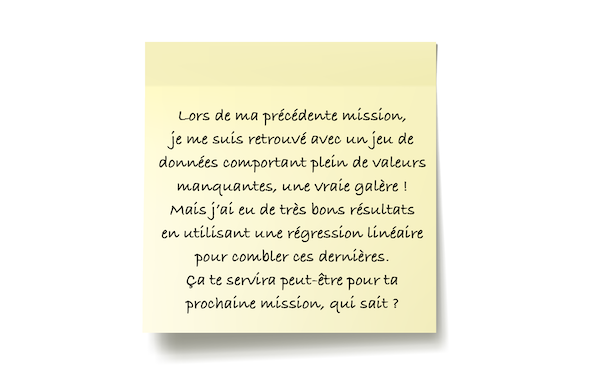
[](https://user.oc-static.com/upload/2020/11/25/16063163322759_Screen%20Shot%202020-11-05%20at%2011.15.15.png)

Cette institution a pour objectif de mettre en place des méthodes d’identification des contrefaçons des billets en euros. Ils font donc appel à vous, spécialiste de la data, pour mettre en place une modélisation qui serait capable d’identifier automatiquement les vrais des faux billets. Et ce à partir simplement de certaines dimensions du billet et des éléments qui le composent.

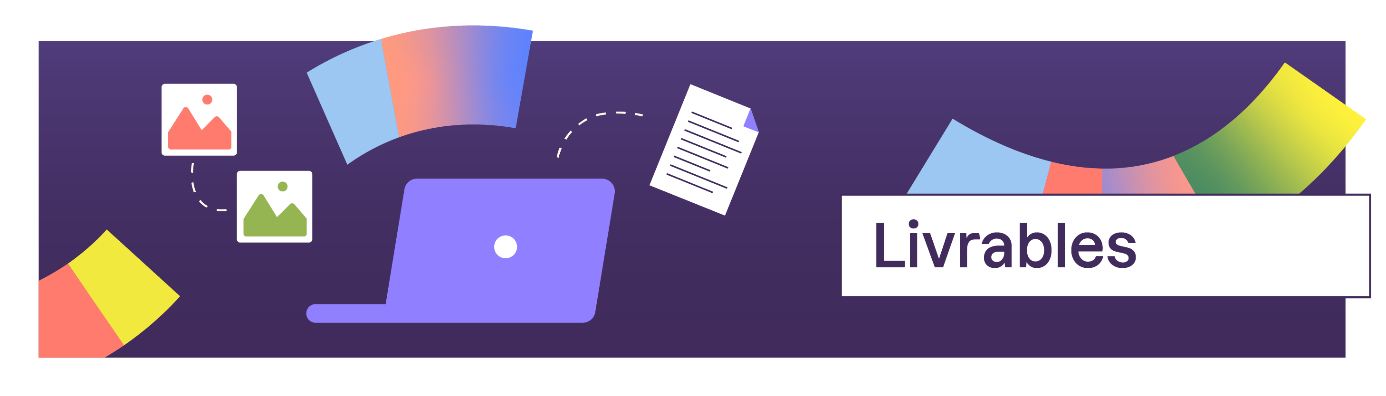
Voici [le cahier des charges de l’ONCFM](https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/parcours-data-analyst/DAN-P10-cdc-detection-faux-billets.pdf), ainsi que [le jeu de données](https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/parcours-data-analyst/billets.csv).

Le client souhaite que vous travailliez directement depuis ses locaux sous la responsabilité de Marie, responsable du projet d’analyse de données à l’ONCFM. Elle vous laissera une grande autonomie pendant votre mission, et vous demande simplement que vous lui présentiez vos résultats une fois la mission terminée. Elle souhaite voir quels sont les traitements et analyses que vous avez réalisés en amont, les différentes pistes explorées pour la construction de l’algorithme, ainsi que le modèle final retenu.

Après avoir lu en détail le cahier des charges, vous vous préparez à vous rendre à l’ONCFM pour prendre vos nouvelles fonctions. Vous notez tout de même un post-it qui se trouve sur le coin de votre bureau, laissé par un de vos collègues :

[](https://user.oc-static.com/upload/2020/11/25/16063165227448_De%CC%81tectez%20des%20faux%20billets%20-%20post%20it.png)

Vous embarquez le post-it dans vos affaires et vous dirigez donc à présent vers les bureaux de l’ONCFM pour commencer votre mission.



1. Votre **code en R ou Python** contenant :
   * l’ensemble des traitements et des tests effectués ;
   * l’application finale.

Déposez sur la plateforme, dans un dossier zip nommé “***Titre\_du\_projet\_nom\_prénom***”, le livrable du projet nommé comme suit : **Nom**\_**Prénom**\_**n° du livrable**\_**nom du livrable**\_**date de démarrage du projet**. Cela donnera :

* *Nom\_Prénom\_1\_code\_mmaaaa*

Par exemple, le livrable peut être nommé comme suit *: Dupont\_Jean\_1\_code*\_*012023.*



Durant la présentation orale, votre mentor jouera le rôle de Marie, la responsable du projet à l’ONCFM à qui vous présentez vos résultats :

* **Présentation des livrables (20 minutes)**
  + Présentation de l’ensemble de votre cheminement, des traitements et analyses réalisés en amont, des différentes pistes explorées pour la construction de l’algorithme, et du modèle final retenu (15 minutes) ;
  + Test de l’algorithme en direct avec un jeu de données qui vous sera fourni pendant la soutenance, ayant la même forme que le jeu suivant : [FICHIER billets\_production.csv](https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/parcours-data-analyst/billets_production.csv) (5 minutes).
* **Discussion (5 minutes)**
  + Échange de questions/réponses : l’évaluateur pourra revenir sur certains points pour vous questionner sur vos choix.
  + L’évaluateur vous fera un retour sur votre prestation en soutenance.
* **Debriefing (5 minutes)**
  + À la fin de la soutenance, l'évaluateur arrêtera de jouer le rôle de Marie pour vous permettre de débriefer ensemble.

Votre présentation devrait durer 20 minutes (+/- 5 minutes). Puisque le respect des durées des présentations est important en milieu professionnel, les présentations en dessous de 15 minutes ou au-dessus de 25 minutes peuvent être refusées.

 Dans ce parcours, vous vous êtes formé aux compétences de la certification [**Data Analyst PL-300 de Microsoft**](https://learn.microsoft.com/fr-fr/certifications/exams/pl-300). Si vous voulez poursuivre votre apprentissage et préparer la certification, vous pouvez consulter les ressources indiquées sur la page de la certification. Obtenir un certificat peut être un plus pour votre insertion professionnelle. Cela reste facultatif. Ce certificat n'est pas un livrable de ce projet.

## Section facultative - Pour aller plus loin

### Mission - Analysez des données de systèmes éducatifs



#### **Qu’allez-vous faire et comment ?**

Barres titres

Cette mission suit un scénario de projet professionnel. Vous pouvez suivre les étapes pour vous aider à réaliser vos livrables.

Avant de démarrer votre travail sur ce projet, nous vous conseillons de :

* lire tout le projet et les documents liés ;
* prendre des notes sur ce que vous avez compris ;
* préparer une liste de questions pour votre première session de mentorat.

#### Prêt à mener la mission ?

### Barres titres

Vous êtes Data Scientist dans une **start-up de la EdTech**, nommée **academy**, qui propose des contenus de formation en ligne pour un public de niveau lycée et université.

**Mark**, votre manager, vous a convié à une réunion pour vous présenter le projet d’**expansion à l’international** de l’entreprise. Il vous confie **une première mission d’analyse exploratoire**, pour déterminer si les données sur l’éducation de la banque mondiale permettent d’informer le projet d’expansion.

Mark aimerait explorer les pays avec un fort potentiel de clients pour les services de **academy**, et comment ce potentiel pourrait évoluer.

#### **Votre mission**Barres titres

Mark vous a donc demandé de réaliser une analyse pré-exploratoire de ce jeu de données. Il vous a transmis cet email à la suite de la réunion :

|  |
| --- |
| Hello,    Les données sont disponibles sur [ce site de la Banque mondiale](https://datacatalog.worldbank.org/dataset/education-statistics) ou en téléchargement direct à ce [lien](https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/Parcours_data_scientist/Projet+-+Donn%C3%A9es+%C3%A9ducatives/Projet+Python_Dataset_Edstats_csv.zip).    Je te laisse regarder la page d'accueil qui décrit le jeu de données. En résumé, l’organisme “EdStats All Indicator Query” de la Banque mondiale répertorie 4 000 indicateurs internationaux décrivant l’accès à l’éducation, l’obtention de diplômes et des informations relatives aux professeurs, aux dépenses liées à l’éducation... Tu trouveras plus d'info sur [ce site](http://datatopics.worldbank.org/education/).    Pour la pré-analyse, pourrais-tu :   * valider la qualité de ce jeu de données (comporte-t-il beaucoup de données manquantes, dupliquées ?) * décrire les informations contenues dans le jeu de données (nombre de colonnes ? nombre de lignes ?) * sélectionner les informations qui semblent pertinentes pour répondre à la problématique (quelles sont les colonnes contenant des informations qui peuvent être utiles pour répondre à la problématique de l’entreprise ?) * déterminer des ordres de grandeurs des indicateurs statistiques classiques pour les différentes zones géographiques et pays du monde (moyenne/médiane/écart-type par pays et par continent ou bloc géographique)     Ton travail va nous permettre de déterminer si ce jeu de données peut informer les décisions d'ouverture vers de nouveaux pays. On va partager ton analyse avec le board, alors merci de soigner la présentation et de l'illustrer avec des graphiques pertinents et lisibles !    Merci ! |

Vous avez toutes les clés en main pour accomplir votre mission !

### Livrables et session de bilan



#### **Livrables**

Barres titres

1. Un**Jupyter Notebook** comportant les analyses pré-exploratoires réalisées.
2. Un **support** de présentation pour la soutenance.

Déposez sur la plateforme, dans un dossier zip nommé “**Titre\_du\_projet\_nom\_prénom**”, votre livrable nommé comme suit : **Nom**\_**Prénom**\_**n° du livrable**\_**nom du livrable**\_**date de démarrage du projet**. Cela donnera :

* Nom\_Prénom\_1\_notebook\_mmaaaa
* Nom\_Prénom\_2\_presentation\_mmaaaa

Par exemple, votre premier livrable peut être nommé comme suit : Dupont\_Jean\_1\_notebook\_12024.

#### **Session de bilan avec votre mentor**

Barres titres

Votre mentor interprêtera le rôle de Mark. Vous lui présenterez l’ensemble de votre travail.

La session de bilan est structurée comme une soutenance pour vous permettre de vous entraîner.

* **Présentation (20 minutes)**
  + Rappel de la problématique et présentation du jeu de données (5 minutes)
  + Présentation de l’analyse pré-exploratoire du jeu de données et vos conclusions sur la pertinence de l’usage du jeu de données pour répondre aux questions stratégiques que se pose l’entreprise (15 minutes)
* **Discussion (5 minutes)**
  + Vore mentor vous challengera sur vos choix.
* **Débriefing (5 minutes)**
  + À la fin de la session, vous prendrez un moment pour débriefer ensemble.

Votre présentation devrait durer 20 minutes (+/- 5 minutes). Puisque le respect des durées des présentations est important en milieu professionnel, les présentations en dessous de 15 minutes ou au-dessus de 25 minutes peuvent être refusées.

### Étapes

**Avant de démarrer cette étape, je dois avoir :**

* suivi les différents cours Python :
  + [Initiez vous à Python pour l’analyse de données](https://openclassrooms.com/en/courses/6204541-initiez-vous-a-python-pour-lanalyse-de-donnees)
  + les deux premières parties de [Découvrez les librairies Python pour la Data Science](https://openclassrooms.com/en/courses/7771531-decouvrez-les-librairies-python-pour-la-data-science)

**Une fois cette étape terminée, je devrais avoir (au sein d’un notebook jupyter)** :

* validé la qualité du jeu de données : vérifiez s’il y a des données qui manquent ou qui sont dupliquées ;
* fait une description précise de chaque fichier de données : nombre de lignes, nombre de colonnes, type des différentes colonnes, etc.

**Recommandations :**

* Tirez partie au maximum des différentes options du jupyter notebook :
  + Assurez-vous que chaque case correspond à une tâche spécifique.
  + N’hésitez pas à ajouter des commentaires et des markdowns pour faciliter la navigation.
* Traitez chaque fichier de données l’un après l’autre.
* Appuyez-vous sur l’ensemble des fonctions d’analyses générales présentées dans le cours [Découvrez les librairies Python pour la Data Science.](https://openclassrooms.com/courses/7771531-decouvrez-les-librairies-python-pour-la-data-science/7857178-creez-votre-premier-data-frame-avec-pandas)
* Mettez en parallèle l’analyse via Python avec le répertoire des différents indicateurs pour bien comprendre quelle information se trouve dans quel fichier.

**Ressources :**

* [Initiez vous à Python pour l’analyse de données](https://openclassrooms.com/en/courses/6204541-initiez-vous-a-python-pour-lanalyse-de-donnees)
* [Découvrez les librairies Python pour la Data Science](https://openclassrooms.com/en/courses/7771531-decouvrez-les-librairies-python-pour-la-data-science)

**Avant de démarrer cette étape, je dois avoir :**

* suivi le cours [Nettoyez et analysez votre jeu de données](https://openclassrooms.com/en/courses/7410486-nettoyez-et-analysez-votre-jeu-de-donnees)

**Une fois cette étape terminée, je devrais avoir :**

* un dataframe ne contenant que les indicateurs jugés pertinents par rapport à l’analyse demandée, d’une qualité suffisante pour y mener une analyse.

**Recommandations :**

* Relisez bien le projet pour bien comprendre le but final de l’analyse : projet d’expansion à l’international.
* Réduisez la liste des indicateurs à une petite 15ène, pour rendre l’analyse plus digeste.
* Analysez la qualité des données associées à ces indicateurs.

**Ressources :**

* Chapitre du cours “Découvrez les librairies Python pour la Data Science “ sur le [filtrage des données d’un dataframe](https://openclassrooms.com/courses/7771531-decouvrez-les-librairies-python-pour-la-data-science/7857549-filtrez-les-donnees-du-data-frame).
* [Webinaire associé au projet](https://app.livestorm.co/openclassrooms-1/session-libre-explorez-et-nettoyez-un-dataset-simple-et-retirez-des-premieres-observations/live?s=eb65918b-c7f6-4477-a10a-f0ff68f55bd2#/qa) qui présente pas à pas un exemple pratique d’exploration et de nettoyage de données.
* [Le repo de la librairie missingno](https://github.com/ResidentMario/missingno) qui permet de simplifier l’analyse et la visualisation les différentes données manquantes d’un dataframe.

**Une fois cette étape terminée, je devrais avoir :**

* créé un ou plusieurs dataframes prêt à être exploités/analysés, qui ne  contiennent que les colonnes et les années pertinentes ;
* organisé le ou les dataframes de la manière suivante :
  + Chaque ligne correspond à un individu (ici un pays).
  + Chaque colonne correspond aux différents indicateurs sélectionnés lors de l’étape précédente.

**Recommandations :**

* Vu qu’il y a plusieurs années disponibles pour chaque indicateur, réfléchissez sur quelles années sont pertinentes ou non, par rapport à ce que l’on cherche et à par rapport au taux de remplissage/nombre de valeurs manquantes.
* Les fichiers originaux sont organisés de sorte à ce que tous les indicateurs sont répertoriés sur plusieurs lignes. Re-organisez ces derniers (via un pivot table par exemple) pour obtenir un dataframe dans un format plus facile à exploiter dans le cadre d’une analyse de données, avec une ligne par pays (voire pays/année) et les différentes colonnes représentant les différents indicateurs.
* Utilisez des markdowns pour bien organiser son notebook au fur et à mesure des différents travaux réalisés.

**Points de vigilance :**

* Les années jouent un rôle central dans l’organisation de nos données : il faut donc choisir comment les traiter (quelles années conserver ou non, comment gérer les valeurs manquantes, etc.) et comment les organiser.

**Ressources :**

* Chapitre du cours : "Découvrez les librairies Python pour la Data Science" sur l’utilisation du [pivot table](https://openclassrooms.com/courses/7771531-decouvrez-les-librairies-python-pour-la-data-science/7857733-agregez-des-donnees-avec-pandas).

**Avant de démarrer cette étape, je dois avoir :**

* un ou plusieurs dataframes bien organisés en ligne/individu et colonnes/indicateurs.

**Une fois cette étape terminée, je devrais avoir :**

* complété et bien organisé un notebook Jupyter qui contiendra l’ensemble des étapes :
  + de sélection des données ;
  + de nettoyage ;
  + d’analyse descriptives menées sur les différents pays ;
  + de création de graphiques univariés et bivariés.

**Recommandations :**

* Pour la partie analyse, les différentes statistiques sont un bon point d’entrée (moyenne, médiane, écart-type, etc.).
* Utilisez les différentes représentations visuelles univariées (histogrammes, boîte à moustaches) et bivariés (nuage de points, barplot, etc.) pour compléter l’analyse descriptive.
* Analysez les corrélations entre les différentes variables via des représentations graphiques adaptées et le calcul des statistiques correspondantes.
* Regardez l’évolution, en se basant sur plusieurs années (ainsi que celles à venir).
* Travaillez par zone géographique (continent, groupes de pays logiques, etc.).

**Points de vigilance :**

* L’analyse va porter sur potentiellement plusieurs années, donc le travail réalisé lors de l’étape précédente peut plus ou moins faciliter l’exploration sur ce dataframe. Les pays sont de plus mélangés avec les zones géographiques dans les différents dataframes, il faut filtrer pour ne pas fausser les différentes statistiques calculées.

**Ressources :**

* Cours [Nettoyez et analysez votre jeu de données](https://openclassrooms.com/courses/7410486-nettoyez-et-analysez-votre-jeu-de-donnees)
* [Cheat sheet sur les différentes visualisations](https://www.psiweb.org/docs/default-source/2018-psi-conference-posters/48-julie-jones.pdf)

**Avant de démarrer cette étape, je dois avoir :**

* fini l’ensemble de l’analyse de données.

**Une fois cette étape terminée, je devrais avoir :**

* préparé une présentation claire et professionnelle avec :
  + les résultats de l’analyse pré-exploratoire ;
  + mes conclusions sur la pertinence de l’usage du jeu de données pour répondre aux questions stratégiques que se pose l’entreprise.

**Recommandations :**

* Expliquez la méthodologie utilisée, en démontrant une compréhension approfondie des étapes de l’analyse.
* Prenez du recul sur les données pour apporter des pistes de réponses concrètes à la problématique d'extension initialement présentée par votre manager.
* Mettez en place une logique de **storytelling**: veillez à ne pas présenter vos résultats de façon scolaire, mais essayez de les rendre le plus intelligibles et accessibles aux personnes non techniques.
* Essayez au maximum d’adopter une posture d’un data scientist en poste qui présenterait ses résultats : soignez donc la présentation en conséquence.

Vous pouvez demander à votre mentor de vous tester en vous posant des questions afin de vous habituer à argumenter pour y répondre. Soyez prêt à justifier et à présenter votre méthodologie (étapes, outils, cohérence des résultats).

## Évaluation



Félicitations !Vos efforts ont porté leurs fruits ! Vous avez acquis toutes les compétences de ce projet.

### Évaluation : lundi 5 juin 2023

Projet validé

#### Modalités de soutenance

 [Mohamed Hassib](https://openclassrooms.com/fr/members/kf68p279dq39)

Évaluateur

* 

Jérôme Lanaro

Étudiant

lundi 5 juin 2023 à 10:00

Enregistrée

#### Livrable final

#### Enregistrement de la soutenance

* lundi 5 juin 2023 à 10:03 | 24 mins | 130 Mo

#### Remarques sur l'évaluation

**Compétences évaluées**

En vous basant sur les critères d’évaluation dans le guide mentor du projet, définissez le statut d'acquisition de chaque compétence listée ci-dessous :

Validé - Expliquez pourquoi en partageant des retours constructifs

Non validé - Expliquez de façon constructive pourquoi et comment l’étudiant peut s’améliorer, en vous appuyant sur les critères d’évaluations

**1. Réaliser une analyse prédictive**

Validé

**2. Réaliser une régression logistique**

Validé

**3. Réaliser une régression linéaire**

Validé

**4. Opérer des classifications automatiques pour partitionner les données**

Validé

**Livrable**

Points forts :

* Bonne présentation
* La regression linéaire est valide
* Les hypothèses de validité suivantes ont été vérifiées :  homoscédasticité et normalité des résidus
* L'étudiante a réalisé le k-means
* Une régression logistique a été réalisée pour prédire le type d’un billet à partir de ses données géométriques
* Une comparaison entres K-means et la regression logistique a été réalisée
* Le programme fonctionne correctement

Axes d'amélioration :

Réponse à la question : dans quel cas nous utilisons la regression logistique

**Soutenance**

Les livrables sont conformes aux attentes. Projet validé