Data analyst : Référentiel d'activités, de compétences et d'évaluation

REFERENTIEL D'ACTIVITES *décrit les situations de travail et les activités exercées, les métiers ou emplois visés*	**REFERENTIEL DE COMPETENCES** *identifie les compétences et les connaissances, y compris transversales, qui découlent du référentiel d'activités*	**REFERENTIEL D'EVALUATION** *définit les critères et les modalités d'évaluation des acquis*	
		MODALITÉS D'ÉVALUATION	**CRITÈRES D'ÉVALUATION**
Bloc n°1 Collecte des données : exploration et requêtage des différents types de bases de données, et récupération des données			
A1.1 Identification et interprétation des données disponibles A1.2 Compréhension du besoin et recherche des données correspondantes A1.3 Gestion d'opérations de collecte et de qualification de données	C1.1 Identifier les possibilités d'utilisation des données, en fonction des besoins métier, en étant force de proposition dans l'exploration, l'évaluation de la qualité et de l'interprétation de ces données C1.2 Définir une stratégie de prise de décision par les données suivant les besoins métier C1.3 Modéliser des bases de données relationnelles (notamment SQL) afin de répondre avec rigueur aux besoins des utilisateurs C1.4 Réaliser des requêtes avancées répondant à des besoins métier complexes : agrégations, jointures, vues et sous-requêtes en s'assurant de l'intégrité des données C1.5 Automatiser des collectes de données	**Mise en situation professionnelle :** Mise en place d'une collecte de données **Cadre de mise en œuvre et de déroulement de l'évaluation** Les candidats sont évalués lors de la réalisation d'une mise en situation au cours de laquelle ils doivent effectuer des recherches à partir de questions proposées, et automatiser la récupération des données depuis une base de données.	- Une exploration de la base est effectuée pour analyser la qualité de données. Les enjeux et les possibilités métiers sont clairement présentés par le candidat. - La stratégie de prise de décision par les données ("data driven") est décrite - La modélisation de la base de données correspond au besoin métier, et est optimisée pour des besoins professionnels. - Les requêtes effectuées sur la base de données relationnelle sont cohérentes avec les questions posées. La performance des requêtes effectuées et les optimisations

	afin d'exploiter les contenus et les informations récoltées sur des pages web ("web scraping") en s'assurant du respect de la réglementation en vigueur C1.6 Mettre en place une interface standard de partage automatique de données entre différentes applications et langages (par exemple avec API REST) C1.7 Contrôler les modalités de collecte et d'utilisation de données et mesurer les enjeux du RGPD afin d'estimer les risques et les responsabilités de chacun	Evaluation individuelle des candidats par un examinateur	sont présentées. - Les données sont collectées automatiquement depuis des pages web grâce au web scraping - Les requêtes API REST sont utilisées et automatisées pour compléter les données de la base de données. Le format d'échange entre langage de programmation est précisé et documenté - Le candidat décrira chaque donnée personnelle entrant dans le cadre du RGPD, explicitera les enjeux, les responsabilités et les risques, et précisera les mesures à prendre si nécessaire.

Bloc n°2 Automatisation du traitement des données : nettoyage, complétion, correction, uniformisation			
A2.1 Analyse des besoins de traitement de données A2.2 Structuration d'outils et d'algorithmes de traitements de données A2.3 Fiabilisation d'outils de traitements de bases de données et organisation de codes	C2.1 Effectuer des choix méthodologiques pour l'automatisation des traitements et les documenter avec clarté et concision C2.2 Utiliser les outils et méthodes modernes : méthodes agiles afin de permettre le travail en équipe, outils de suivi de projets, logiciel adapté à la rédaction de code C2.3 Manipuler des structures de données (chaînes de caractères, listes, dictionnaires et tuples) et utiliser l'algorithmie (variables, boucles, itérateur, conditions, fonctions) afin de traduire en script des besoins de traitements de données C2.4 Appliquer les bonnes pratiques de la programmation afin d'avoir un code organisé, réutilisable et partageable dans un cadre professionnel C2.5 Utiliser les tableaux de données (notamment les DataFrames avec Python et Pandas) afin de faciliter l'import, la manipulation et la fusion de données	**Mise en situation professionnelle** Structuration d'outils / programmes de traitement de données selon des objectifs **Cadre de mise en œuvre et de déroulement de l'évaluation** Les candidats doivent, à partir d'objectifs définis de traitement de données, structurer des outils et utiliser des algorithmes afin de manipuler des tableaux de données, et automatiser le nettoyage notamment des valeurs manquantes et aberrantes. Evaluation individuelle des candidats par un examinateur	- La documentation est présente, correspond bien aux traitements développés et explique les choix méthodologiques effectués - Les outils et méthodes de gestion de projets sont présentés par le candidat - L'algorithmie en langage python, notamment la création de fonctions avancées et optimisées, est utilisée pour l'automatisation des retraitements de données. - Les principes du "clean code" (notamment ceux décrits dans le PEP8 de la documentation officielle Python) sont respectés. - Les traitements de données et des transformations sont effectués, optimisés et

	C2.6 Nettoyer les données, retraiter les valeurs aberrantes (outliers) et les valeurs manquantes afin d'éviter les impacts sur l'exploitation et le traitement des bases de données C2.7 Utiliser les expressions régulières (RegEx) pour traiter les valeurs textuelles et permettre une anonymisation des données personnelles dans le cadre du RGPD		automatisés (notamment grâce à la bibliothèque Pandas) - Le retraitement des valeurs aberrantes et manquantes est effectué, et argumenté sur la méthodologie employée - Les expressions régulières sont correctement utilisées pour identifier des formats de valeurs textuelles (par exemple des adresses e-mail) et permettre une anonymisation

Bloc n°3 Modélisation des données structurées : identification des corrélations existantes et utilisation du Machine Learning pour établir des prévisions			
A3.1 Sélection d'informations utiles A3.2 Elaboration de structures de traitement d'informations A3.3 Modélisation de données	C3.1 Utiliser les statistiques descriptives afin de modéliser les données et en faire émerger des informations pertinentes C3.2 Maîtriser le process d'apprentissage automatique (Machine Learning) afin de permettre à des algorithmes d'apprendre automatiquement à partir de données : syntaxe, découpage jeu d'entrainement et de validation, entrainement, prédiction, mesure C3.3 Modéliser des régressions et interpréter les métriques associées afin de définir des modèles de prévisions, et de trouver des tendances futures pour des valeurs numériques C3.4 Modéliser des classifications et interpréter les métriques associées afin de catégoriser automatiquement des informations C3.5 Traiter automatiquement le langage naturel (NLP) à partir de texte brut afin d'en tirer de la valeur en fonction de classification (Analyse de sentiments)	**Mise en situation professionnelle** Modélisation de données structurées et détermination de prévisions **Descriptif de l'évaluation** Les candidats doivent présenter une proposition de modélisation de données structurées permettant de décrire les données de manière simple, d'en tirer des tendances, de prévoir des valeurs futures, et d'en interpréter les résultats. Evaluation individuelle des candidats par un examinateur	- Les statistiques descriptives (variance, quantiles, coefficients de corrélation) sont utilisées pour expliquer les données - Le process de Machine Learning est utilisé et intégré à la modélisation - Les régressions supervisées sont modélisées et les métriques associées sont interprétées correctement - Les classifications supervisées sont modélisées et les métriques associées sont interprétées correctement - Un corpus de texte est traité et catégorisé automatiquement grâce à des techniques de NLP

	C3.6 Contrôler et documenter les biais d'un modèle et des données d'entrainement afin d'estimer les risques éthiques de ce modèle, notamment dans le cas d'usages de données personnelles C3.7 Communiquer et vulgariser le fonctionnement interne d'un algorithme d'apprentissage automatique afin d'éviter le phénomène de "boite noire"		- La documentation est présente et fait bien apparaître les dimensions les plus utilisées par le modèle, les limites du modèle et les biais potentiels - L'algorithme d'apprentissage automatique est expliqué et ses résultats sont interprétés

Bloc n°4 Visualisation des données : valorisation et interprétation des données pertinentes, et mise en forme dans un tableau de bord			
A4.1 Visualisation de données A4.2 Présentation et partage de données et d'informations	C4.1 Identifier et prioriser, en fonction du besoin métier, les informations à rendre accessibles et à présenter visuellement, afin de structurer des représentations graphiques de tableaux bord C4.2 Utiliser les visualisations descriptives, notamment les nuages de points, les boîtes à moustache et les histogrammes afin de représenter graphiquement des données statistiques et des informations modélisées à destination d'analyste de données C4.3 Manipuler la Dataviz interactive et dynamique (par exemple avec Plotly ou Bokeh) afin de réaliser différents types de représentations graphiques et visuelles de données à destination d'utilisateurs opérationnels C4.4 Réaliser de la cartographie (notamment avec Folium) afin de représenter des informations	**Mise en situation professionnelle** Présentation d'un Tableau de bord réalisé à partir de besoins d'un client **Cadre de mise en œuvre et de déroulement de l'évaluation** Les candidats doivent présenter un tableau de bord réalisé à partir d'un cahier des charges détaillant, à partir de bases de données existantes, des objectifs et des attentes en termes de présentations graphiques et visuelles d'informations et de données croisées. Évaluation individuelle des candidats par un examinateur.	- Les besoins et attentes concernant le tableau de bord sont respectés. La synthèse et la maquette correspondent au besoin métier et sont présentées avec un formalisme professionnel. - Les visualisations descriptives sont correctement utilisées pour décrire les données de manière concise. Le candidat devra expliquer les choix effectués dans la sélection des indicateurs les plus pertinents. - Les visualisations interactives et dynamiques (Plotly ou Bokeh) sont utilisées à des fins de communication et de vulgarisation de l'information. - Une cartographie représente les données géographiques de manière agrégée. La cartographie est interactive afin de répondre aux besoins utilisateur. - Le fonctions avancées du tableur sont utilisées, notamment les tableaux croisés dynamiques, les recherches inter fichiers, les complétions de données et les graphiques croisés.

	géographiques C4.5 Utiliser un Tableur, et notamment les tableaux croisés dynamiques afin de proposer des croisements de variables pour obtenir des informations recherchées C4.6 Réaliser des tableaux de bord avec des outils de Business Intelligence (par exemple PowerBI ou Tableau) afin d'intégrer et de croiser des informations utiles à des approches stratégiques de problématiques C4.7 Prendre en compte les handicaps visuels afin de produire des graphiques lisibles par tous C4.8 Présenter à l'oral et à l'écrit de manière claire, concise et sans ambiguïté les informations		- Un outil de Business Intelligence (PowerBI ou Tableau) permet une utilisation stratégique du tableau de bord, afin d'orienter la prise de décision. Les commentaires rédigés sur le tableau de bord sont pertinents et correspondent aux questions métier. - Les graphiques prennent en compte les bonnes pratiques d'accessibilité pour répondre aux situations courantes de handicaps visuels (notamment le code couleur, et les explications sous les graphiques permettant une lecture vocale automatisée) - Les diapositives, le tableau de bord et la présentation orale sont clairs, homogènes et cohérents.