Course One
Foundations of Data Science

Instructions

Utilisez ce document de stratégie **PACE** pour enregistrer vos décisions et réflexions tout au long de ce projet de fin de cours.
Vous pouvez vous en servir comme **guide** afin d'orienter vos réponses et réflexions à différentes étapes du **processus d'analyse de données**.
De plus, les documents de stratégie PACE peuvent être utilisés comme **ressource de référence** lors de vos **projets futurs**.

Course Project Recap

Quel que soit le parcours que vous avez choisi, vos objectifs pour ce projet sont les suivants :

- ☐ **Compléter le document de stratégie PACE** afin de planifier votre projet tout en prenant en compte vos **audiences**, vos **coéquipiers**, les **jalons clés**, et l'**objectif global** du projet.

- ☐ **Créer une proposition de projet** destinée à l'**équipe de données** (*data team*).

Questions d'entretien pertinentes

La réalisation de ce projet de fin de cours vous permettra de répondre efficacement aux sujets d'entretien suivants :

- En tant que nouveau membre d'une équipe d'analyse de données, **quelles étapes entreprendriez-vous pour vous mettre à jour** sur un projet en cours ? Quelles actions meneriez-vous ? Avec qui souhaiteriez-vous vous entretenir ?

- **Comment planifieriez-vous un projet d'analyse de données ?**

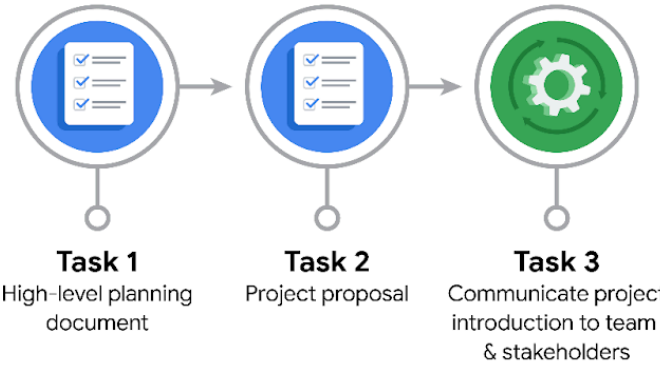
- **Quelles étapes suivriez-vous pour traduire une question métier en une solution analytique ?**

- **Pourquoi la gestion active des données est-elle une responsabilité importante** pour une équipe d'analyse de données ?

- **Quels éléments devez-vous garder à l'esprit lors de la présentation des résultats ?**

Guide de référence

Ce projet comporte **trois tâches** ; le visuel suivant montre **comment les différentes étapes de la méthode PACE** sont intégrées à travers ces tâches.



Data Project Questions & Considerations

PACE: Plan Stage

- Who is your audience for this project?

L'audience principale de ce projet est **la Commission des taxis et limousines de la ville de New York (TLC)**, c'est-à-dire **le client** qui a mandaté Automatidata pour développer le modèle de régression prédictif des tarifs.

- What are you trying to solve or accomplish? And, what do you anticipate the impact of this work will be on the larger needs of the client?

L'objectif de ce projet est de **concevoir un modèle de régression capable de prédire avec précision le tarif d'une course en taxi à New York avant son départ**

En développant ce modèle prédictif, Automatidata vise à fournir à la **Taxi and Limousine Commission (TLC)** un **outil d'aide à la décision basé sur les données**, permettant d'améliorer la transparence tarifaire, la satisfaction client et la planification opérationnelle des chauffeurs et exploitants.

- What questions need to be asked or answered?

Données et sources

- Quelles sont les **sources de données** disponibles auprès de la TLC (courses en taxi, covoiturage, capteurs GPS, conditions météo, etc.) ?
- Les données sont-elles **complètes, récentes et exemptes de biais** ?
- Quelles variables doivent être **prises en compte** dans le modèle (distance, durée, heure, trafic, conditions météorologiques, localisation de départ/arrivée, etc.) ?
- Comment les données **manquantes ou aberrantes** seront-elles traitées ?

Objectifs et métriques

- Quelle est la **marge d'erreur acceptable** pour les prédictions de tarifs ?
- Quel niveau de **précision minimale** (ex. 90 %) est attendu par la TLC ?
- Comment sera **mesuré le succès** du modèle (RMSE, MAE, R^2, etc.) ?

Contraintes et déploiement

- Quelles sont les **contraintes techniques** (infrastructure, format, compatibilité avec les systèmes existants) ?
- Le modèle doit-il être **mis à jour en temps réel** ou de manière périodique ?
- Quelles sont les **attentes de la TLC** en matière de **visualisation et de reporting** des résultats ?
- Comment les résultats seront-ils **présentés aux parties prenantes non techniques** (tableaux de bord, graphiques, rapports synthétiques) ?

- What resources are required to complete this project?

1. Ressources humaines

- **Équipe d'analystes de données** pour le nettoyage, l'exploration et la modélisation des données.
- **Scientifiques des données (Data Scientists)** pour concevoir, entraîner et évaluer le modèle de régression.
- **Ingénieurs en données (Data Engineers)** pour la gestion, la structuration et l'intégration des données issues des différentes sources (TLC, GPS, météo, etc.).
- **Chef de projet** pour coordonner les étapes, assurer la communication entre l'équipe et la TLC, et garantir le respect des délais.
- **Concepteurs de visualisation (Data Visualization Specialists)** pour présenter les résultats de manière claire et accessible aux parties prenantes non techniques.

2. Ressources techniques

- **Environnements de développement** : Python (Pandas, Scikit-learn), R, Jupyter Notebook.
- **Outils de gestion et de stockage des données** : bases SQL, systèmes cloud (ex. Google BigQuery, AWS ou Azure).
- **Outils de visualisation** : Tableau, Power BI ou Google Looker Studio.
- **Infrastructure informatique** : serveurs capables de traiter des volumes importants de données historiques et en temps réel.

3. Ressources en données

- **Jeux de données historiques** de la TLC (courses, tarifs, localisation, durée, météo, etc.).
- **Flux de données en temps réel** provenant des capteurs GPS des véhicules.
- **Sources externes** : données météorologiques, horaires de pointe, événements spéciaux pouvant influencer les trajets.

4. Ressources organisationnelles

- **Budget** pour la location de serveurs, les licences logicielles et les frais de personnel.

- **Soutien de la TLC** pour accéder aux données confidentielles et valider les livrables.
- **Canaux de communication** efficaces (réunions hebdomadaires, outils collaboratifs comme Slack, Trello ou Jira).

- What are the deliverables that will need to be created over the course of this project?

1. Phase de planification (Plan)

- **Proposition de projet détaillée** décrivant l'objectif, la portée, les étapes clés et le calendrier du projet.
- **Document de cadrage des besoins** précisant les variables à étudier, les sources de données à utiliser et les contraintes techniques.
- **Plan de gestion des données** incluant les protocoles de sécurité, de confidentialité et de validation des données fournies par la TLC.

2. Phase d'analyse (Analyze)

- **Rapport d'exploration des données (EDA)** présentant les tendances, anomalies et corrélations initiales observées dans le jeu de données.
- **Jeu de données nettoyé et préparé**, prêt pour la modélisation.
- **Rapport d'évaluation de la qualité des données**, incluant les méthodes de traitement des valeurs manquantes, doublons ou données aberrantes.

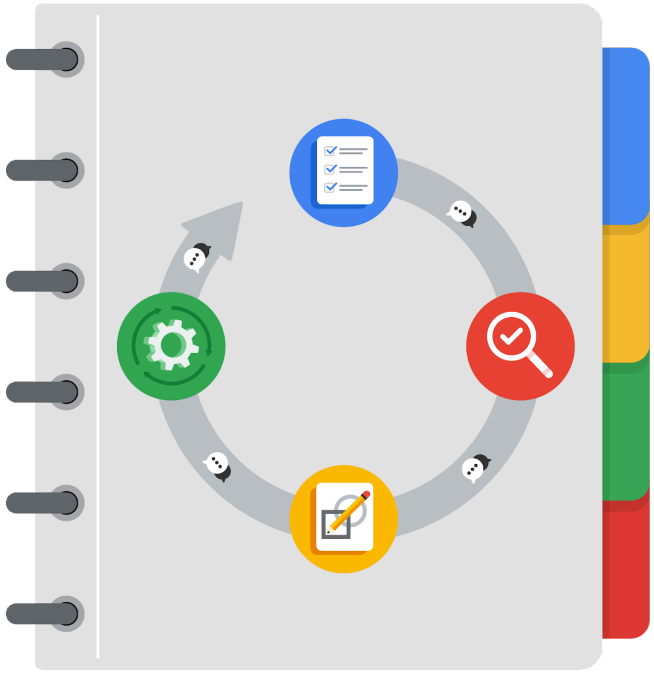
3. Phase de construction (Construct)

- **Modèle de régression prédictive** permettant d'estimer le tarif des trajets à partir de variables telles que la distance, l'heure, le trafic, ou les conditions météo.
- **Documentation technique du modèle**, décrivant les algorithmes utilisés, les métriques de performance (R^2, RMSE, etc.) et les limites identifiées.
- **Rapport de test du modèle** présentant les résultats de la phase de validation et les ajustements apportés pour atteindre la précision cible (≥ 90 %).

4. Phase d'exécution (Exécute)

- **Tableau de bord interactif** (Tableau, Power BI ou Looker Studio) permettant à la TLC de visualiser les prévisions tarifaires et les performances du modèle.
- **Présentation finale aux parties prenantes** résumant la méthodologie, les résultats, et les recommandations d'optimisation.
- **Guide d'implémentation** pour l'intégration du modèle dans les systèmes internes de la TLC, incluant des instructions pour la maintenance et les mises à jour futures.

THE PACE WORKFLOW



[Texte alternatif : Le flux de travail PACE avec ses quatre étapes en cercle : planifier, analyser, construire et exécuter.]

On vous a demandé de montrer à l'équipe de données de l'entreprise comment vous utiliseriez le flux de travail PACE pour organiser et classer les tâches du projet à venir.

Sélectionnez une étape du processus PACE à partir des menus déroulants.
Certaines tâches concernent plus d'une étape du flux de travail PACE.
De plus, toutes les situations professionnelles ne nécessitent pas l'exécution de chaque tâche.

Si vous avez besoin de plus d'informations sur les tâches incluses dans le projet, consultez le document "aperçu du projet de portfolio de fin de cours 1".

Tâches du projet

Voici un ensemble de tâches que l'équipe de données de votre entreprise a déterminé comme nécessaires à la réalisation de ce projet.

Le **responsable de l'analyse des données** vous a demandé d'**organiser ces tâches** en vue de la rédaction du **document de proposition de projet**.

1. Commencez par **identifier à quelle étape du flux de travail PACE** chaque tâche correspond le mieux, en utilisant le menu déroulant.

2. Ensuite, **expliquez pourquoi** vous avez choisi cette étape pour chaque tâche.

Consultez les lectures suivantes pour vous aider dans vos choix et vos explications :

- **Les étapes du processus PACE**

- **Communiquer les objectifs à l'aide d'une proposition de projet**

Plus tard, vous **réorganiserez ces tâches** dans le cadre de la **proposition de projet**.

1. **Evaluating the model:** Analyze

   Pourquoi avez-vous choisi cette étape pour cette tâche ?

Parce que l'évaluation du modèle implique d'analyser sa performance et de vérifier s'il répond aux critères de précision et aux objectifs du projet. Cela fait partie de la phase *Analyser*, où les résultats sont interprétés et les ajustements nécessaires sont identifiés.

2. **Conduct hypothesis testing:** Analyze **and** Construct

   Why did you select these stages for this task?

Parce que les tests d'hypothèse nécessitent d'abord une analyse des données pour formuler des suppositions, puis la construction de tests statistiques afin de confirmer ou d'infirmer ces hypothèses avant la phase de modélisation.

3. **Begin exploring the data:** Analyze

Why did you select this stage for this task?

L'exploration initiale des données fait partie de la phase **Analyze**, car elle vise à comprendre la structure du jeu de données, identifier les variables pertinentes, repérer les valeurs manquantes et détecter d'éventuelles anomalies. Cette étape permet de formuler des premières hypothèses sur les relations entre les variables.

4. **Data exploration and cleaning:** Analyze **and** Construct

Why did you select these stages for this task?

Le **nettoyage** des données commence dans la phase **Analyze**, lorsque les incohérences et les erreurs sont détectées. Il se poursuit dans la phase **Construct**, où les transformations et corrections sont appliquées pour préparer les données à la modélisation. Ensemble, ces étapes garantissent un ensemble de données fiable et exploitable.

5. **Establish structure for project workflow (PACE):** Plan

Why did you select this stage for this task?

L'établissement de la structure du flux de travail appartient à la phase **Plan**, car c'est à ce stade que les objectifs, les livrables, le calendrier, les rôles et les ressources du projet sont définis. Cela permet de créer une feuille de route claire pour l'ensemble de l'équipe avant toute analyse ou modélisation.

6. **Communicate final insights with stakeholders:** Execute

Why did you select this stage for this task?

La communication des résultats finaux se situe dans la phase **Execute**, car c'est à ce moment que les conclusions et recommandations sont présentées aux parties prenantes. L'objectif est de traduire les résultats techniques du modèle en informations exploitables pour la prise de décision stratégique.

7. **Compute descriptive statistics:** Analyze

Why did you select this stage for this task?

Le calcul des statistiques descriptives fait partie de la phase **Analyze**, car il s'agit de comprendre la distribution, les tendances et la variabilité des données. Ces mesures (moyenne, médiane, écart-type, etc.) permettent de dégager les caractéristiques principales du jeu de données avant de passer à la modélisation.

8. **Visualization building:** Analyze **and** Execute

Why did you select these stages for this task?

Les visualisations sont d'abord créées pendant la phase **Analyze** pour explorer les relations entre les variables et détecter des tendances. Elles sont ensuite affinées et utilisées dans la phase **Execute** pour communiquer les résultats et les insights finaux aux parties prenantes de manière claire et percutante.

9. **Write a project proposal:** Plan

Why did you select this stage for this task?

La rédaction de la proposition de projet relève de la phase **Plan**, car elle consiste à définir les objectifs, la portée, les ressources nécessaires, les délais et les livrables du projet. C'est une étape stratégique qui fixe la direction à suivre avant toute analyse ou construction de modèle.

10. **Build a regression model:** Construct **and** Analyze

Why did you select this stage for this task?

La construction du modèle de régression se situe dans la phase Construct, car elle implique la création et le paramétrage du modèle à partir des données nettoyées. Cependant, l'évaluation de la performance du modèle et l'interprétation des résultats relèvent également de la phase Analyze, afin de valider la pertinence du modèle et son adéquation aux objectifs du projet.

11. **Compile summary information about the data:** Analyze

Why did you select this stage for this task?

Compiler des informations résumées sur les données relève de la phase **Analyze**, car il s'agit d'examiner et de synthétiser les caractéristiques principales du jeu de données (taille, types de variables, valeurs manquantes, moyennes, distributions, corrélations, etc.). Cette étape permet de mieux comprendre la structure et la qualité des données avant de construire des modèles.

12. **Build machine learning model:** Construct

Why did you select this stage for this task?

La construction d'un modèle de machine learning appartient à la phase Construct, car elle consiste à développer, entraîner et ajuster le modèle à partir des données disponibles. C'est l'étape où les algorithmes sont sélectionnés et testés afin de produire des prédictions fiables, en lien direct avec les objectifs définis dans la phase de planification.