

Guide d'étapes clés - Analysez l'évolution des prix de l'immobilier avec Python

Avez-vous des retours sur ce document ?

Ceci est un nouveau format que nous testons. Votre retour est précieux !

Comment utiliser ce document ?

Ce guide vous propose un découpage du projet en étapes. Vous pouvez suivre ces étapes selon vos besoins. Dans chacune, vous trouverez :

- des recommandations pour compléter la mission ;
- les points de vigilance à garder en tête ;
- une estimation de votre avancement sur l'ensemble du projet (attention, celui-ci peut varier d'un apprenant à l'autre).

Suivre ce guide vous permettra ainsi :

- d'organiser votre temps ;
- de gagner en autonomie ;
- d'utiliser les cours et ressources de façon efficace ;
- de développer une approche de résolution de projet professionnelle.

Gardez en tête que votre progression sur les étapes n'est qu'une estimation et sera différente selon votre vitesse.

Recommandations générales

L'objectif de ce projet est de vous faire découvrir et mettre en oeuvre une batterie de techniques issues de la data science que vous pourrez réutiliser dans vos projets professionnels, tels que :

- l'utilisation du langage de programmation Python ;
- l'exploration d'une série temporelle avec plusieurs facteurs ;

- les tests de corrélation ;
- la prédiction via une régression linéaire ;
- le regroupement automatique non supervisé dit “clustering”.

Pour cela, vous allez manipuler un jeu de données simplifié qui représente l'évolution des prix immobiliers sur Paris entre 2017 et 2021.

Ce projet peut représenter un niveau de difficulté important du fait du nouveau langage de programmation introduit et des notions mathématiques. Pas d'inquiétude ! Pour ces raisons, vous trouverez plus bas dans le guide hebdo les supports suivants :

- Des modèles de code partagés par Jade. Grâce à ces modèles, vous allez gagner du temps dans la rédaction du code.
- Des liens vers la documentation de certaines librairies Python (Pandas, Matplotlib, Scikit Learn).
- Des liens vers des supports d'explication des concepts mathématiques utilisés dans ce projet comme les corrélations, la régression ou le clustering.

Étape 1 : Analyse des données historiques du marché immobilier parisien

15% de progression

Avant de démarrer cette étape, je dois avoir :

Description de ce qui doit être fait avant chaque phase (cours ou parties de cours à suivre, autres étapes déjà complétées...)

- installé Python sur mon ordinateur, ou avoir accès à une console Python en ligne comme [Google Colab](#) ;
- suivi le cours d'initiation à Python ;
- suivi le cours d'introduction au Machine Learning avec Python.

Une fois cette étape terminée, je devrais avoir :

- analysé des données dans un document au format iPython Notebook. Les analyses à réaliser sont les suivantes :
 1. le type des données ;
 2. les différents types de biens immobiliers ;
 3. le nombre de transactions au total et par type de bien ;
 4. l'intervalle de l'historique des données ;
 5. l'évolution du prix au mètre carré moyen des appartements dans Paris entre 2017 et 2021 au global ;
 6. les différences de prix au mètre carré moyen des appartements par arrondissement ;
 7. les prix au mètre carré par transaction dans un arrondissement ;
 8. la corrélation entre le prix au mètre carré et la date de la transaction ;
 9. la corrélation entre le prix au mètre carré et la surface ;
 10. la différence entre les prix au mètre carré pour les appartements et pour les locaux commerciaux.

Recommandations :

- Vous pouvez vous appuyer sur le modèle de code disponible.

Points de vigilance :

- La principale difficulté est dans la compréhension du code à utiliser pour les analyses. Les opérations sur les dataframes de Pandas sont beaucoup utilisées. N'hésitez pas à lire les ressources et à rechercher sur internet, si besoin, pour comprendre chaque ligne de code que vous utilisez.

Ressources :

- **Pour la lecture des données :** la fonction de [pandas.read_excel](#).

- **Pour les opérations de filtrage sur les données :** Une ressource en français sur les [slicers](#).
- **Pour se projeter sur l'exercice qui lui est demandé :** Une visualisation des [prix immobiliers](#) sur Paris via le site [meilleursagents](#).
- **Pour construire des graphiques :** la fonction [plot de Matplotlib](#).
- **Pour le test de corrélation de Pearson :** consultez l'[explication](#).

Étape 2 : Entraînement de l'algorithme de régression linéaire

30% de progression

Avant de démarrer cette étape, je dois avoir :

- vu la ressource sur le one hot encoding pour comprendre cette technique ;
- lu la documentation sur l'échantillonnage train test pour comprendre cette technique.

Une fois cette étape terminée, je devrais avoir :

- entraîné un algorithme de régression linéaire dans un document au format iPython Notebook.

Recommandations :

- Vous pouvez vous appuyer sur le modèle de code.
- Avec un algorithme de régression linéaire simple vous devez arriver à moins de 10% d'erreur en moyenne sur la prédiction de la valeur foncière. Si vous atteignez cette cible, vous pouvez vous arrêter là.
- Votre mentor vous demandera d'avoir un avis critique sur la régression linéaire dans notre contexte et comment vous pourriez aller plus loin pour réduire ce taux d'erreur moyen.

Points de vigilance :

- Dans cette étape, la principale difficulté est dans la compréhension des nouveaux concepts à appréhender.

Ressources :

- **One hot encoding :** consultez la [vidéo explicative](#).
- **Échantillonnage train test split :** consultez la [documentation](#).
- **Régression linéaire :**
 - Consultez la [documentation sur Scikit learn](#);
 - Consultez une [explication](#) du fonctionnement en vidéo.

- **Test de l'erreur moyenne absolue** : Vous pouvez consulter la [documentation](#). À noter, dans le modèle le choix qui est fait est d'afficher cette valeur en pourcentage de la valeur foncière. Pour cette raison, la fonction de Scikit learn n'a pas été utilisée.

Étape 3 : Prédiction de la valorisation future du portefeuille et présentation des résultats

50% de progression

Avant de démarrer cette étape, je dois avoir :

- nettoyé le code de mon algorithme de prédiction pour être prêt à l'utiliser.

Une fois cette étape terminée, je devrais avoir :

- la prédiction de la valorisation future du portefeuille de l'entreprise au format csv ;
- la présentation finale des résultats à l'entreprise.

Recommandations :

- Consultez le modèle de code.
- Consultez le modèle de présentation.
- Durant votre entretien avec votre mentor, il interprétera le rôle de Maurice, le Président Directeur Général de la société "Les plus beaux logis de Paris". Vous devrez présenter :
 - l'analyse de l'évolution des prix immobiliers à Paris ;
 - la prédiction de la valorisation du portefeuille de l'entreprise.

Points de vigilance :

- Rappelez-vous que Maurice n'est pas technique, son objectif est de comprendre votre démarche en tant que non initié. Il vous challengera sur les points suivants :
 - la représentativité des données que vous avez utilisées ;
 - les limites de vos prédictions et leur fiabilité.

Étape 4 : Classification des données issues du jeu de test

90% de progression

Avant de démarrer cette étape, je dois avoir :

- récupéré les données à classer ;
- parcouru les ressources sur l'algorithme K-means pour bien les comprendre.

Une fois cette étape terminée, je devrais avoir :

- classifié de façon non supervisée des biens immobiliers du portefeuille de l'entreprise sur un document au format iPython Notebook.

Ressources :

- **Algorithme K-means :**
 - [Documentation](#)
 - [Vidéo](#) en anglais avec sous-titres français.

Étape 5 : Analyse et présentation des résultats

100% de progression

Avant de démarrer cette étape, je dois avoir :

- récupéré les labellisations correctes des biens immobiliers pour pouvoir les comparer à mes résultats.

Une fois cette étape terminée, je devrais avoir :

- analysé les résultats de l'étude pour les présenter au commanditaire au format Slides.

Recommandations :

- Reprenez le modèle de présentation.
- Votre interlocuteur n'est pas technique, votre présentation doit lui permettre de comprendre vos résultats.
- Vous serez interrogés sur les limites et précautions à prendre avec vos résultats. Préparez-vous à y répondre !

Projet terminé !