\_Fondamentaux de Python

Projet

# House Market

Allons un peu plus loin dans python

1. Téléchargez les trois fichiers “superficie.csv” / “nombre\_de\_sdb.csv” / “nombre\_de\_chambres.csv”
2. Importez les trois datasets

|  |
| --- |
| superficie = pd.read\_csv("datasets/superficie.csv", header=None) nombre\_de\_sdb = pd.read\_csv("datasets/nombre\_de\_sdb.csv", header=None) nombre\_de\_chambres = pd.read\_csv("datasets/nombre\_de\_chambres.csv",header=None) |

1. Donnez un nom de colonne à ces trois datasets

|  |
| --- |
| superficie = superficie.rename(columns={0:"id",1:"supercificie"}) nombre\_de\_sdb = nombre\_de\_sdb.rename(columns={0:"id","1":"nombre\_de\_sdb"}) Nombre\_de\_chambres = Nombre\_de\_chambres.rename(columns={0:"id","1":"nombre\_de\_chambres"}) |

1. Fusionner ces trois datasets un seul dataframe (on fera une fusion simplement par index) qu’on nommera *dataset*

|  |
| --- |
| dataset = pd.concat([nombre\_de\_sdb, nombre\_de\_chambres, superficie], axis=1) |

1. Quelle est la superficie moyenne des maisons de notre dataset ?

|  |
| --- |
| dataset.superficie.mean() |

1. Quelle est le nombre de chambre médian et moyen ?

|  |
| --- |
| dataset.describe() |

1. Nous un nouveau dataset “house\_price.csv” que nous voudrions intégrer à notre dataset
   1. Cette fois nous voudrions fusionner les index avec les id du dataset.
      1. Importez house\_price.csv dans une variable qu’on nommera *house\_price*
      2. Créez un id pour *dataset* qui sera égal à l’index de *dataset*
      3. Fusionnez *house\_price* à *dataset*

|  |
| --- |
| house\_pricing = pd.read\_csv("house\_price.csv") dataset["id"] = dataset.index dataset.merge(house\_pricing, on = "id") |

1. Quel est le coût moyen d’une maison ?

|  |
| --- |
| dataset.house\_price.mean() |

1. Quel est le coût moyen par chambre ?

|  |
| --- |
| dataset.groupby("nb\_de\_chambres").mean() |

1. Le coût moyen par chambre ne nous apprend pas grand chose.
   1. Tentons de faire des catégories de superficie
      1. Créez une nouvelle colonne dans votre dataset qu’on appellera *taille\_maison*
      2. Créez trois catégories “très grande” / “grande” / “moyenne” / “petite” / “très petite” qui respectivement correspondront à :
         1. “une maison dont la taille est supérieure à 25 000 sqrt\_feet”
         2. “une maison dont la taille est comprise entre 20 000 et 25 000 sqrt\_feet”
         3. “une maison dont la taille est comprise entre 15 000 et 20 000 sqrt\_feet”
         4. “une maison dont la taille est comprise entre 10 000 et 15 000 sqrt\_feet”
         5. “une maison dont la taille est inférieure à 10 000 sqrt\_feet”
      3. Remplissez les lignes correspondantes dans la colonne *taille\_maison*

|  |
| --- |
| *dataset["taille\_maison"] = 0 for i in range(0, len(dataset)):  if dataset.iloc[i,2] > 25000:  dataset.iloc[i, -1] = "Très Grande"  elif dataset.iloc[i,2] > 20000:  dataset.iloc[i, -1] = "Grande"  elif dataset.iloc[i, 2] > 15000:  dataset.iloc[i, -1] = "Moyenne"  elif dataset.iloc[i, 2] > 10000:  dataset.iloc[i, -1] = "Petite"* |

1. Quel est le coût moyen d’une maison en fonction de sa catégorie de superficie ?

|  |
| --- |
| dataset.groupby("taille\_maison").mean() |

1. Appliquez le code ci-dessous pour visualiser votre résultat. Que pouvez vous conclure ?

|  |
| --- |
| import seaborn as sns sns.lmplot(x = "sqrt\_meters", y="house\_price", data = dataset) |

On peut voir sur le graphique que plus la superficie d’une maison monte, plus son prix monte. Il y a donc une corrélation linéaire assez forte entre ces deux variables.