

Webscraping, présentation finale: Outil comparatif de salaires et indices de coût de la vie par régions

Eva PADRINO – DIA3

Présentation du 13/01/2023

Lien GitHub

https://github.com/eva-pa/projet_webscrapping_padrino_eva_dia3

Table des matières

1. Problématique
2. Objectifs
3. Sites à scrapper
4. Librairies
5. Avancements à la dernière séance de TD
6. Perspectives à la dernière séance de TD
7. Résultats finaux

1. Problématique

Je suis à la recherche d'un emploi et je veux savoir où m'installer selon le
salaire proposé
et l'indice du coût de la vie d'une zone géographique.

2. Objectifs

- Obtenir le **salaire** médian, minimum et maximum pour un métier, pour une **ville** recherchée ou **partout dans le monde**.
- Croiser ces données avec les indices: **du coût de la vie, de loyer, des courses, prix des restaurants, pouvoir d'achat**.
- **Comparer avec des lieux à une distance voulue**: en faisant des ratios entre le salaire et l'indice sur des maps ou graphiques.
- Chercher les meilleurs rapports salaires/indice dans le monde ou par région.

3. Sites scrappés.

- **Glassdoor.fr:** Obtenir la médiane, le minimum et le maximum d'un salaire dans une ville pour un poste.



- **fr.numbeo.com:** Indices liés au coût de la vie pour plusieurs villes dans le monde entier



Choisissez une région: Afrique Amérique Asie Europe Océanie							
Search: <input type="text"/>							
Classement	Ville	Indice du Coût de la Vie	Indice de Loyer	Indice du Coût de la Vie Plus Loyer	Indice des Courses	Indice des Prix des Restaurants	Indice du Pouvoir d'Achat Local
1	Hamilton, Bermudes	141,80	92,39	117,89	149,67	128,18	84,70
2	Bâle, Suisse	124,18	43,75	85,25	130,07	118,45	122,05
3	Lausanne, Suisse	119,46	52,74	87,17	118,44	113,54	103,69
4	Zurich, Suisse	118,65	59,53	90,04	116,01	116,93	125,01
5	Zoug, Suisse	117,17	63,59	91,24	111,97	125,44	145,57
6	Berne, Suisse	114,82	36,82	77,07	118,09	107,59	132,26
7	Santa Barbara, Californie, États Unis	111,48	91,07	101,60	119,23	115,34	71,38

3.Utilisation des données/ **Forme du résultat**



- Notebook interactif
- Entrer un métier et une ville
- Obtenir plusieurs types de visuels: cartes avec les ratios salaires/indices, statistiques de salaire
- Classements des meilleures ratios dans un périmètre choisi

4. Libraries

- Selenium
- BeautifulSoup
- Matplotlib, seaborn, plotly
- Haversine
- Folium
- Deep-translator
- Geopy
- Pandas
- Unicodedata
- Ipywidget

5. Avancements à la dernière séance de TD : Numbeo

ville	idx_cout_vie	idx_loyer	idx_cout_vie_loyer	idx_courses	idx_prix_restaurants	idx_pouvoir_achat_local	Pays	lat	lon
Johannesbourg	39,39	13,36	26,95	31,06	34,74	95,97	Afrique du Sud	-26.205000	28.049722
Pretoria	36,66	12,11	24,92	28,54	31,77	94,54	Afrique du Sud	-25.745928	28.187910
Le Cap	36,10	20,00	28,40	29,36	34,51	76,61	Afrique du Sud	-33.928992	18.417396
Durban	35,96	11,75	24,39	27,28	34,10	78,46	Afrique du Sud	-29.861825	31.009909
Munich	71,44	40,73	56,76	60,65	66,77	103,72	Allemagne	48.137108	11.575382

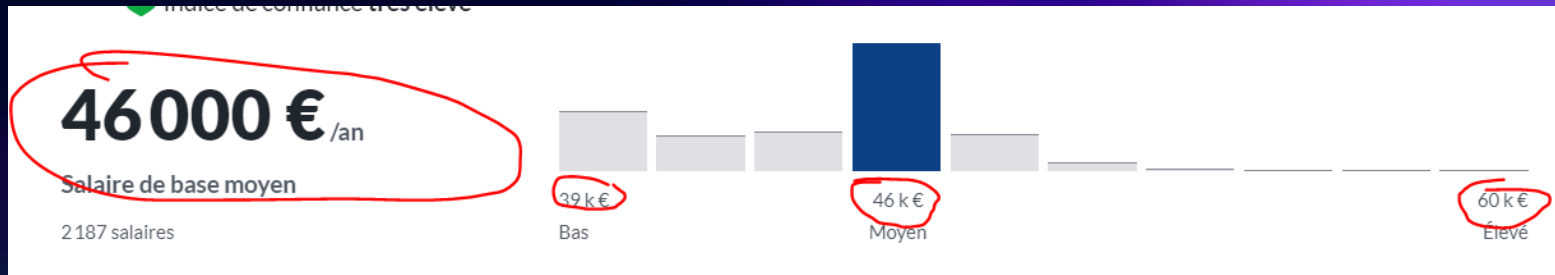
- Récupération villes sur Numbeo
- Traitements des textes des villes + création des liens
- Récupération des tableaux quand ils existaient
- Assemblage des tableaux
- Génération de la latitude et longitude avec geopy
- Vérification avec folium des coordonnées



5. Avancements à la dernière séance de TD Statistiques

Glassdoor

- En entrée: **poste, ville, pays.**
 - Assemblage de ville et pays
 - 1^{ère} **méthode:** remplissage de formulaire sur la page <https://www.glassdoor.com/Salaries/index.htm>
 - 2^{ème} **méthode:** traduction en anglais de la ville et pays et remplissage du formulaire sur **Glassdoor**
- ⇒ Ces 2 méthodes posent problème, des fois le site Glassdoor va choisir glassdoor va choisir soi-même la localisation , par exemple on entre « Johannesburg, Afrique du Sud » et choisit « bouche du rhone » alors que Johannesburg existe bien sur le site
- 3^{ème} **méthode retenue:** « glassdoor salaire + poste + ville + ',' + pays » sur moteur de recherche Ecosia car ses balises sont tout le temps les mêmes
 - On prend le 1^{er} lien s'il contient glassdoor dans l'adresse.



5. Avancements à la dernière séance de TD : Statistiques Glassdoor

On traite tous les cas : si la page est vide, s'il manque certaines valeurs....

```
ObtainResultSal('Data Scientist','Paris','France')
```

✓ 55.6s

```
{'minSal': 39000,  
'moySal': 46000,  
'maxSal': 60000,  
'titrePage': "Salaires d'un Data Scientist",  
'localisationPage': 'Paris, France'}
```

46 000 € /an

Salaire de base moyen

2187 salaires

39 k€

Bas

46 k€

Moyen

60 k€

Élevé

5. Avancements à la dernière séance de TD : Numbeo points proches.

```
res = DistanceFromPoint(df, 'lat', 'lon', 48, 2)
```

0.16

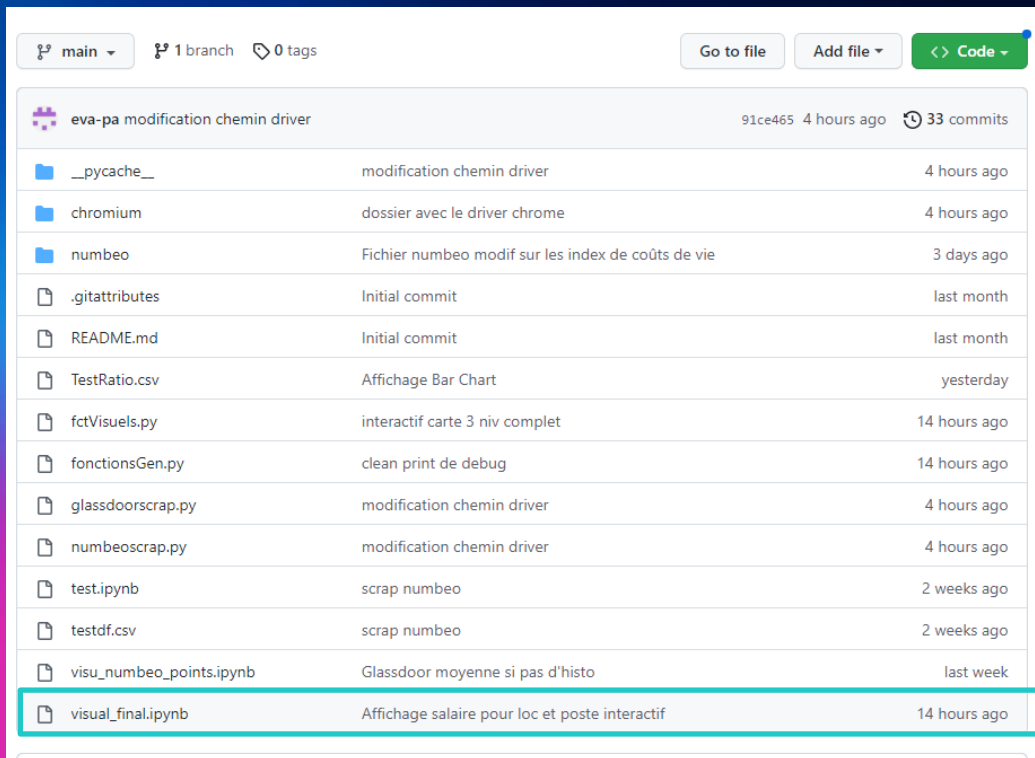
ville	idx_cout_vie	idx_loyer	idx_cout_vie_loyer	idx_courses	idx_prix_restaurants	idx_pouvoir_achat_local	Pays	lat	lon	distance_km
Johannesbourg	39,39	13,36	26,95	31,06	34,74	95,97	Afrique du Sud	-26.205000	28.049722	8651.738832
Pretoria	36,66	12,11	24,92	28,54	31,77	94,54	Afrique du Sud	-25.745928	28.187910	8607.235539
Le Cap	36,10	20,00	28,40	29,36	34,51	76,61	Afrique du Sud	-33.928992	18.417396	9255.537610
Durban	35,96	11,75	24,39	27,28	34,10	78,46	Afrique du Sud	-29.861825	31.009909	9128.987781
Munich	71,44	40,73	56,76	60,65	66,77	103,72	Allemagne	48.137108	11.575382	711.203683

6. Perspectives à la dernière séance de TD

- interactives avec ipywidget
- Afficher ratio salaire/indice dans les villes proches x km
- Afficher données coût de la ville la plus proche seulement + résultat glassdoor
- Une carte avec ratio ou salaire par ville.

7. Résultats: Mode d'emploi

- Projet disponible sur : https://github.com/eva-pa/projet_webscrapping_padrino_eva_dia3
- Après avoir téléchargé/cloné le repository, ouvrir le notebook « visual_final.ipynb »



The screenshot shows the GitHub interface for the repository 'eva-pa modification chemin driver'. At the top, there are buttons for 'Go to file', 'Add file', and 'Code'. Below this, a table lists the repository's files and folders. The file 'visual_final.ipynb' is highlighted with a red border. The table columns are: file/folder name, description, and time since last commit.

File/Folder	Description	Time
__pycache__	modification chemin driver	4 hours ago
chromium	dossier avec le driver chrome	4 hours ago
numbeo	Fichier numbeo modif sur les index de coûts de vie	3 days ago
.gitattributes	Initial commit	last month
README.md	Initial commit	last month
TestRatio.csv	Affichage Bar Chart	yesterday
fctVisuels.py	interactif carte 3 niv complet	14 hours ago
fonctionsGen.py	clean print de debug	14 hours ago
glassdoorscrap.py	modification chemin driver	4 hours ago
numbeoscrap.py	modification chemin driver	4 hours ago
test.ipynb	scrap numbeo	2 weeks ago
testdf.csv	scrap numbeo	2 weeks ago
visu_numbeo_points.ipynb	Glassdoor moyenne si pas d'histo	last week
visual_final.ipynb	Affichage salaire pour loc et poste interactif	14 hours ago

7. Résultats: Mode d'emploi

- Pip install nécessaires:
 - pip install selenium
 - pip install beautifulsoup4
 - pip install plotly
 - pip install haversine
 - pip install folium
 - pip install deep-translator
 - pip install geopy
 - pip install pandas
 - pip install unicodedata
 - pip install ipywidgets
- Avoir chrome d'installé (Version 109.0.5414.75 (Build officiel) (64 bits))
- En lançant toutes les cellules du notebook une fenêtre Chrome va s'ouvrir, ne jamais la fermer pendant l'utilisation du notebook.

7. Résultats: Mode d'emploi

- Comment utiliser les outils comparatifs?

1. Afficher les indices dans un rayon autour d'un point:

Dans le menu interactif, il faudra remplir des **champs obligatoirement (entourés ici en bleu)** pour cette option et en **rouge** à NE PAS CHANGER.

Ville	Ville
Pays	Pays
Rayon max ...	100.2
Indice coût ...	▼
Poste reche...	Poste recherché
Statistique s...	▼
Afficher ratio	False
Ordonner le...	False
Lancer la co...	False

→ Ville et pays du point désiré
Rayon maximum autour de ce point en km.
Ces options sont à taper au clavier

→ Ouvrir ce menu déroulant pour choisir parmi les indice de coûts suivants: 'idx_cout_vie', 'idx_loyer', 'idx_cout_vie_loyer', 'idx_courses', 'idx_prix_restaurants', 'idx_pouvoir_achat_local'.

→ Ouvrir ce menu déroulant et choisir True quand tous les champs obligatoires ont été remplis.

7. Résultats: Mode d'emploi

- Comment utiliser les outils comparatifs?

2. Afficher des statistiques de salaires: Salaire minimum, salaire moyen, salaire maximum:

Dans le menu interactif, il faudra remplir des **champs obligatoirement (entourés ici en bleu)** pour cette option et en **rouge** à NE PAS CHANGER.

Ville	Ville
Pays	Pays
Rayon max ...	100.2
Indice coût ...	▼
Poste reche...	Poste recherché
Statistique s...	▼
Afficher ratio	False
Ordonner le...	False
Lancer la co...	False

Ville et pays du point désiré
Rayon maximum autour de ce point en km.
Ces options sont à taper au clavier

Entrer au clavier un poste (exemple: Data Engineer)

Ouvrir ce menu déroulant pour choisir parmi les options suivantes : 'moySal'(salaire moyen), 'minSal' (salaire minimum), 'maxSal' (salaire maximum)

Ouvrir ce menu déroulant et choisir True quand tous les champs obligatoires ont été remplis.

7. Résultats: Mode d'emploi

- Comment utiliser les outils comparatifs?

3. Afficher ratio entre statistique de salaire et un indice de coût de la vie:

Dans le menu interactif, il faudra remplir des **champs obligatoirement (entourés ici en bleu)** pour cette option et en **rouge** à NE PAS CHANGER.

Ville	<input type="text" value="Ville"/>
Pays	<input type="text" value="Pays"/>
Rayon max ...	<input type="text" value="100.2"/>
Indice coût ...	<input type="text" value=""/>
Poste reche...	<input type="text" value="Poste recherché"/>
Statistique s...	<input type="text" value=""/>
Afficher ratio	<input type="text" value="False"/>
Ordonner le...	<input type="text" value="False"/>
Lancer la co...	<input type="text" value="False"/>

→ Ville et pays du point désiré

→ Rayon maximum autour de ce point en km.

→ Ces options sont à taper au clavier






→ Ouvrir ce menu déroulant pour choisir parmi les indice de coûts suivants: 'idx_cout_vie', 'idx_loyer', 'idx_cout_vie_loyer', 'idx_courses', 'idx_prix_restaurants', 'idx_pouvoir_achat_local'.

→ Ouvrir ce menu déroulant pour choisir parmi les options suivantes : 'moySal'(salaire moyen), 'minSal' (salaire minimum), 'maxSal' (salaire maximum)

→ Ouvrir ce menu déroulant et choisir True.

→ Ouvrir ce menu déroulant et choisir True quand tous les champs obligatoires ont été remplis.

7. Résultats: Mode d'emploi

Ville	<input type="text" value="Ville"/>
Pays	<input type="text" value="Pays"/>
Rayon max ...	<input type="text" value="100.2"/>
Indice coût ...	<input type="text" value=""/> 
Poste reche...	<input type="text" value="Poste recherché"/>
Statistique s...	<input type="text" value=""/> 
Afficher ratio	<input type="text" value="False"/> 
Ordonner le...	<input type="text" value="False"/> 
Lancer la co...	<input type="text" value="False"/> 

Recommandé: avant de changer un champ, mettre ce menu déroulant sur False

7. Résultats: graphiques avec des barres, exemples.

1. Afficher les indices dans un rayon autour d'un point:

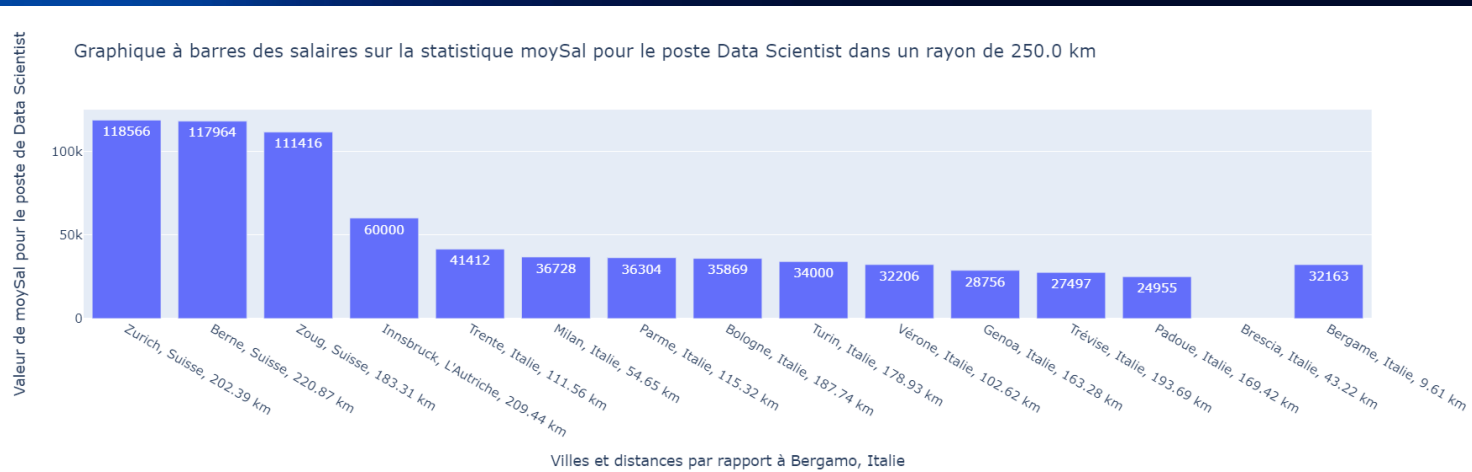
Ville	Berlin
Pays	Allemagne
Rayon max ...	300
Indice coût ...	idx_cout_vie
Poste recher...	Poste recherché
Statistique s...	
Afficher ratio	False
Ordonner le...	True
Lancer la co...	True



7. Résultats: graphiques avec des barres, exemples.

2. Afficher des statistiques de salaires: Salaire minimum, salaire moyen, salaire maximum:

Ville	Bergamo
Pays	Italie
Rayon max ...	250
Indice coût ...	
Poste reche...	Data Scientist
Statistique s...	moySal
Afficher ratio	False
Ordonner le...	True
Lancer la co...	True



Note : Bergame et Berlin existe dans la base de données Numbeo mais on peut chercher une ville qui n'en fait pas partie.

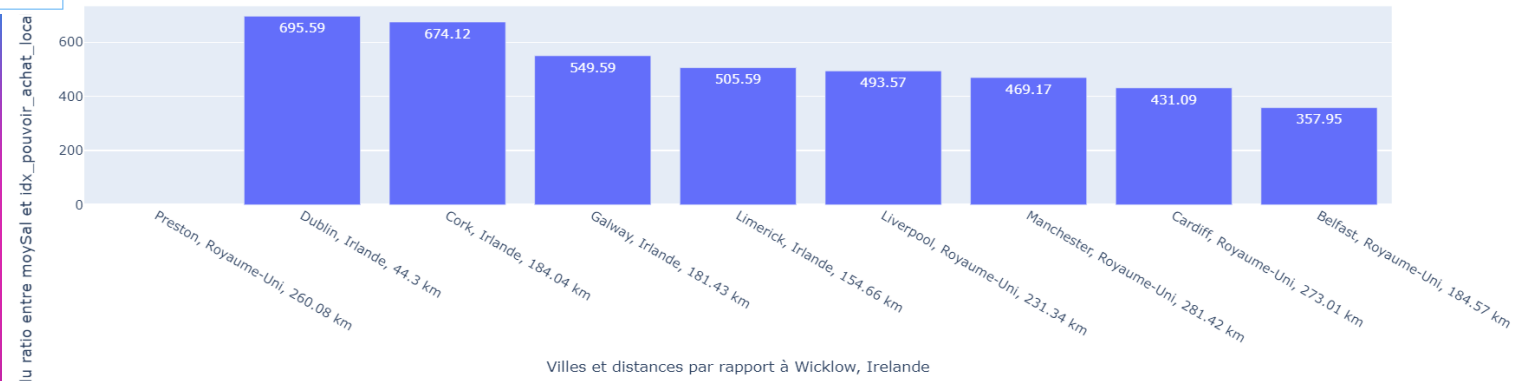
Pour Brescia, il n'existait pas de page sur Glassdoor donc il n'y a pas de barre.

7. Résultats: graphiques avec des barres, exemples.

3. Afficher ratio entre statistique de salaire et un indice de coût de la vie:

Ville	Wicklow
Pays	Irlande
Rayon max ...	300
Indice coût ...	idx_pouvoir_achat_local
Poste reche...	Data Scientist
Statistique s...	moySal
Afficher ratio	True
Ordonner le...	True
Lancer la co...	True

Graphique à barres des ratios entre la statistique de salaire moySal et l'indice idx_pouvoir_achat_local pour le poste Data Scientist dans un rayon de 300.0



7. Résultats: cartes, exemples.

1. Afficher les indices dans un rayon autour d'un point:

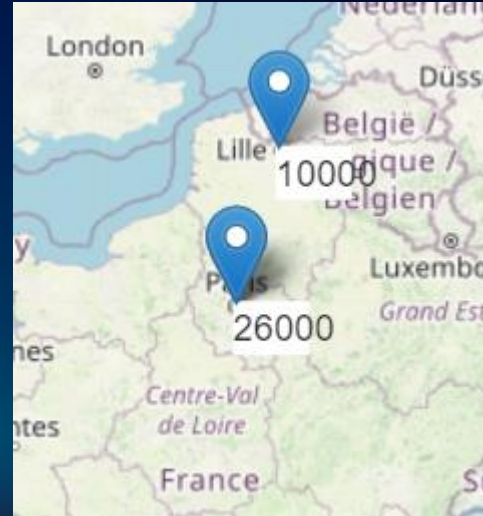
Ville	Toledo
Pays	Espagne
Rayon max ...	600
Indice coût ...	idx_cout_vie ▼
Poste reche...	Poste recherché
Statistique s...	▼
Afficher ratio	False ▼
Lancer la co...	True ▼



7. Résultats: **cartes**, exemples.

2. Afficher des statistiques de salaires: Salaire minimum, salaire moyen, salaire maximum:

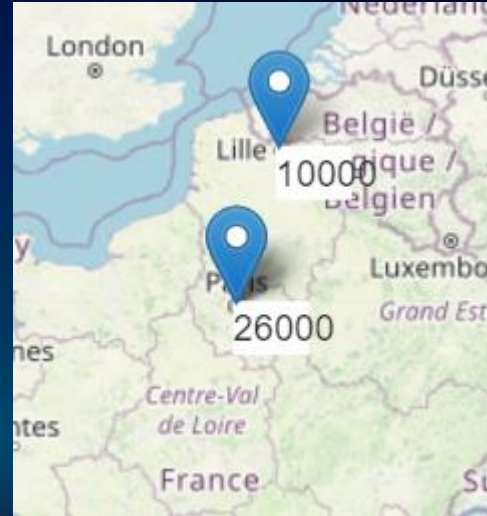
Ville	Paris
Pays	France
Rayon max ...	260
Indice coût ...	▼
Poste reche...	Community Manager
Statistique s...	minSal ▼
Afficher ratio	False ▼
Lancer la co...	True ▼



7. Résultats: **cartes**, exemples.

2. Afficher des statistiques de salaires: Salaire minimum, salaire moyen, salaire maximum:

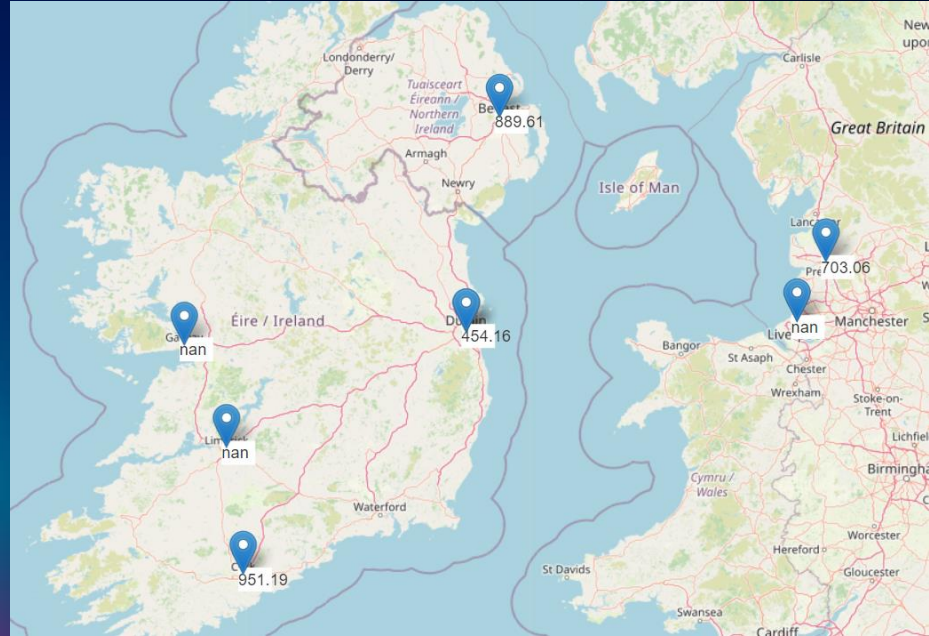
Ville	Paris
Pays	France
Rayon max ...	260
Indice coût ...	▼
Poste reche...	Community Manager
Statistique s...	minSal ▼
Afficher ratio	False ▼
Lancer la co...	True ▼



7. Résultats: cartes, exemples.

3. Afficher ratio entre statistique de salaire et un indice de coût de la vie:

Ville	Dublin
Pays	Irlande
Rayon max ...	260
Indice coût ...	idx_loyer ▼
Poste reche...	Community Manager
Statistique s...	minSal ▼
Afficher ratio	True ▼
Lancer la co...	True ▼



Note : Quand la valeur n'est pas disponible (minimum ici par exemple) nan est affiché.
La page peut exister mais n'avoir que le salaire moyen.

7. Résultats: afficher statistiques salaire, exemple

Ville	<input type="text" value="Annecy"/>
Pays	<input type="text" value="France"/>
Poste reche...	<input type="text" value="Software Engineer"/>
Lancer le pr...	<input type="button" value="True"/> ▼

Page obtenue sur Glassdoor: Software Engineer Salaries
Localisation indiquée dans la page: Annecy
Salaire minimum: None
Salaire moyen: 40922
Salaire maximum: None

Lien GitHub

https://github.com/eva-pa/projet_webscrapping_padrino_eva_dia3

Sites utilisés

fr.numbeo.com