Prédiction des Revenus dans le SecteurTech et Analyse de ses Facteurs Déterminants (Données Stack Overflow 2023)

Barkad Nour Hassan

Plan

- Note de concept et plan de mise en œuvre:
  - Contexte
  - Objectifs
  - Relation aux ODD
- Données
  - Collecte de données
  - Analyse exploratoire des données (AED) et ingénierie des caractéristiques
- Sélection et entraînement du modèle
  - Évaluation du modèle et ajustement des hyperparamètres
  - Affinage du modèle et tests
- Résultats
- Déploiement
- Travaux futurs

Note de concept et plan de mise en œuvre

Contexte

- Le projet vise à réaliser une analyse prédictive des données de l'enquête annuelle des développeurs Stack Overflow 2023.
- En 2023, l'enquête s'est déroulée du 8 au 19 mai et a attiré une participation massive de développeurs professionnels.
- L'accent de cette édition est mis sur l'intelligence artificielle (IA) et l'apprentissage automatique (ML).

Objectifs du Projet

1. Identifier les variables clés influençant le revenu des développeurs à temps plein dans le secteur technologique.
2. Mettre en évidence les facteurs déterminants du revenu dans ce secteur.
3. Élaborer des recommandations pour orienter :
   - ❖ Les développeurs dans le choix de leurs compétences et parcours professionnels.
   - ❖ Les universitaires dans la mise à jour des programmes éducatifs.
   - ❖ Les personnes en réorientation de carrière vers des opportunités en IT.
   - ❖ Les décideurs politiques dans l'élaboration de politiques éducatives et de formation.
   - ❖ Les organisations de développement des compétences pour concevoir des programmes de formation adaptés aux besoins du marché.

Relation aux ODD

ODD 8 : Travail décent et croissance économique

- Identifier les facteurs influençant les salaires pour conseiller **les étudiants et les personnes en réorientation professionnelle** sur les compétences et parcours en IT afin d'augmenter leurs revenus .

Données

Collecte de données

Les données utilisées pour ce projet proviennent de l'enquête annuelle des développeurs de Stack Overflow 2023.

Notre base d'étude se compose de 89 184 observations (individus) et de 84 variables (80 variables qualitatives et 4 variables quantitatives).

les principales variables de la base de notre étude :

1. Basic Information (Age, type d'emploi, ..)

2. Education, Work, and Career (Niveau d'éducation, nombre d'années en codage, taille de l'organisation, …)

3. Technology and Tech Culture (Langage de programmation, BDD utilisées, plateformes cloud, ….)

4. Artificial Intelligence (Utilisation de l'IA, opinion sur l'IA, Confiance dans l'IA, …)

La variable d'intérêt est ConvertedCompYearly (salaires annuels en USD).

Population élligible

- **Je** vais limiter mon analyse aux employés à temps plein (**Employed, full-time**). Suite à cette sélection, il reste 53 748 observations dans notre base de données.

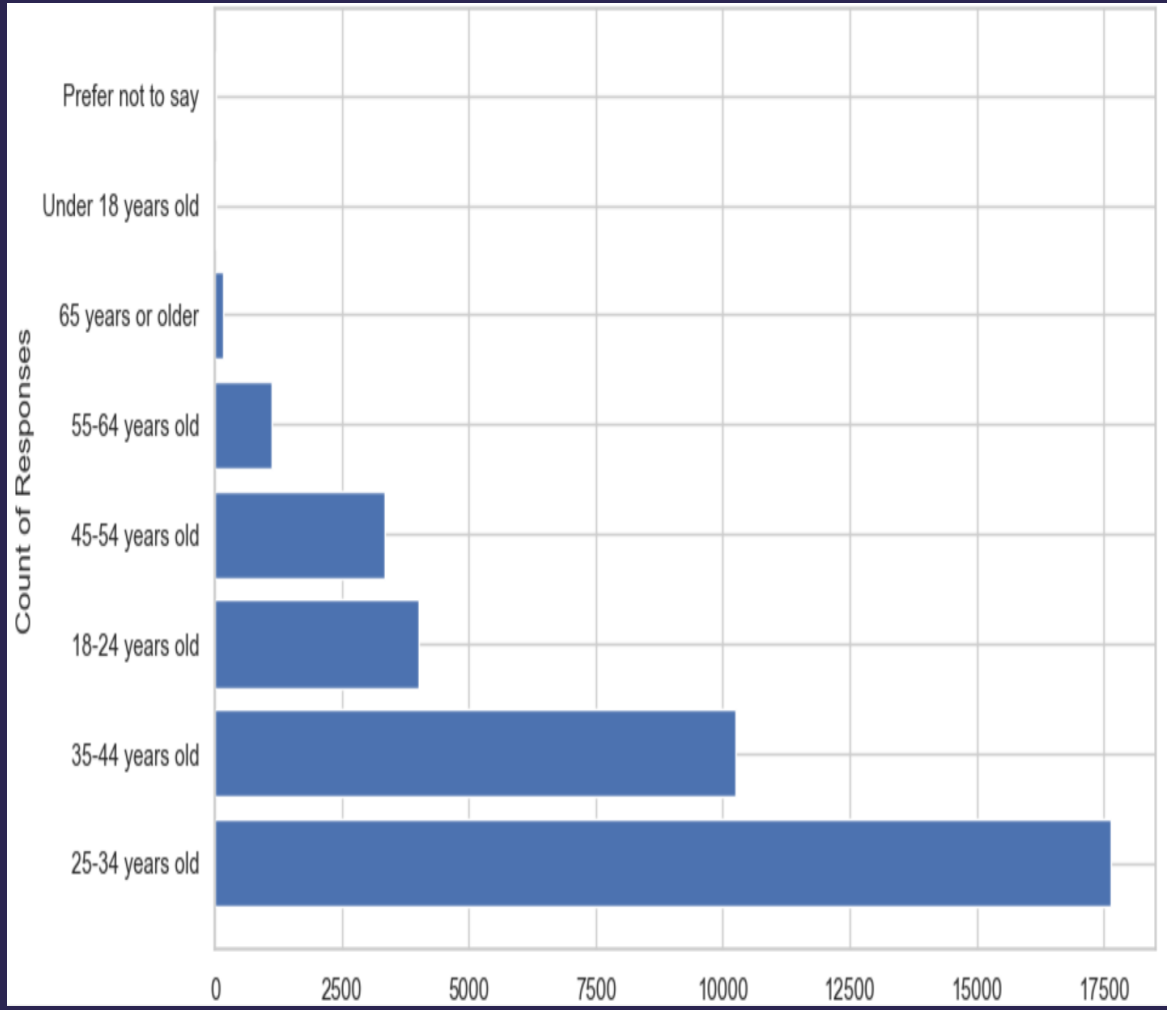
- De plus, 3 087 individus qui n'ont pas répondu à **50 % des questions posées seront exclus** de la base de données. Ils seront supprimés de la base.

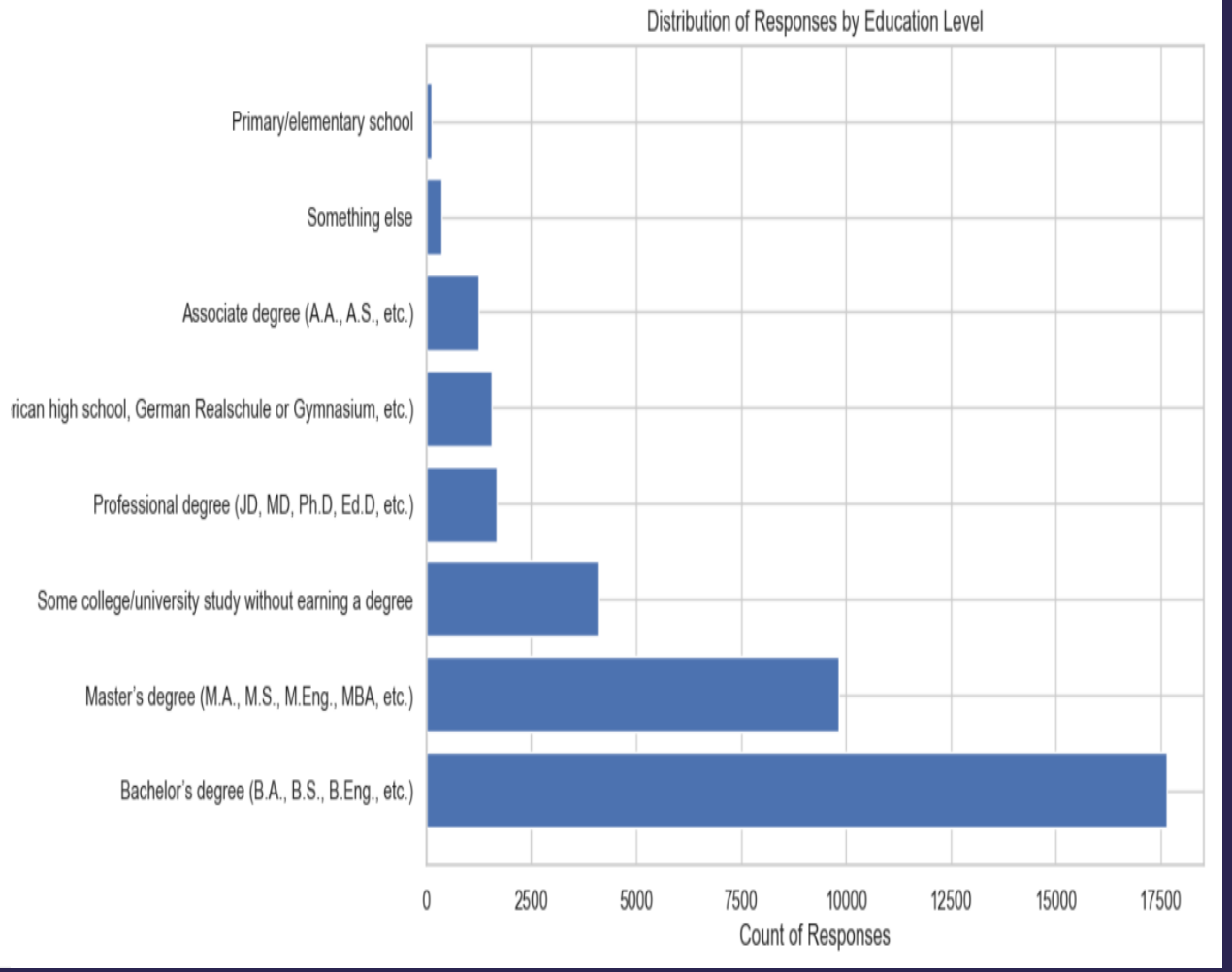
- La variable ConvertedCompYearly, qui est notre variable d'interet, **comporte 12 477 valeurs manquantes.** Ces observations seront éliminées.

- Enfin, nous procéderons à la suppression des valeurs aberrantes, de la variable d'interet totalisant exactement 1 640 observations exclues. Après ces ajustements, notre population éligible se compose de **36 544 observations.**

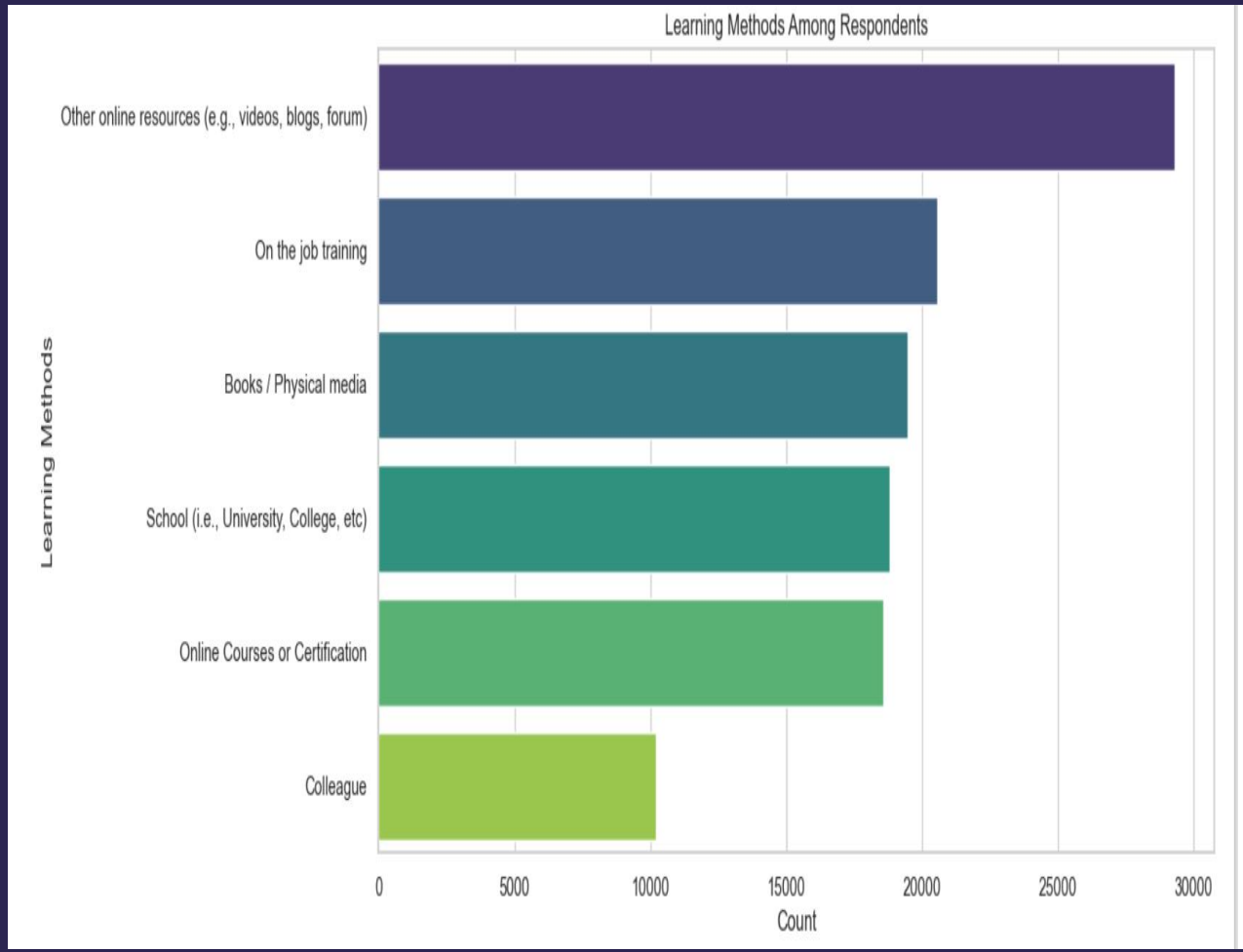
Valeurs manquantes:

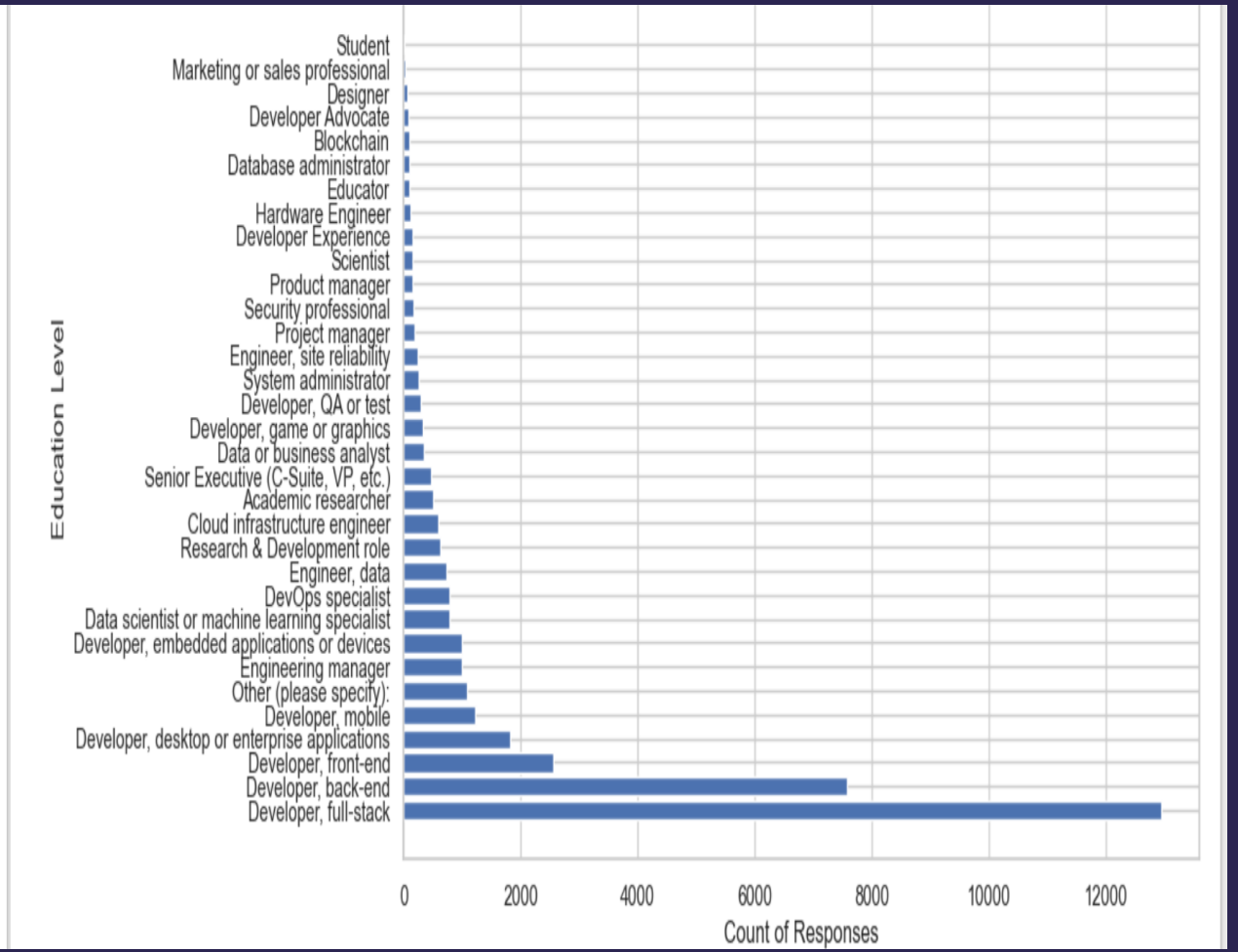
Âge et Diplômes des Développeurs



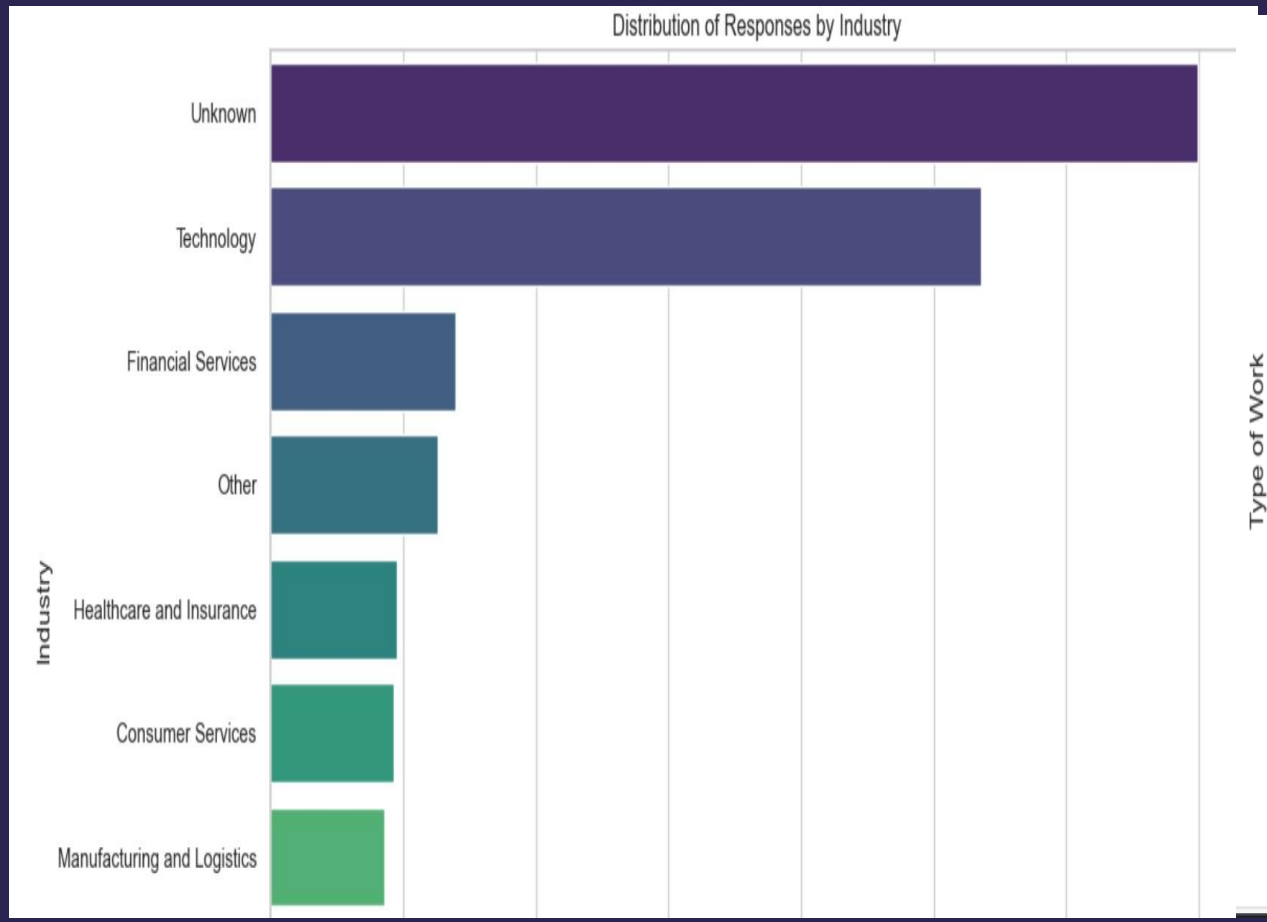


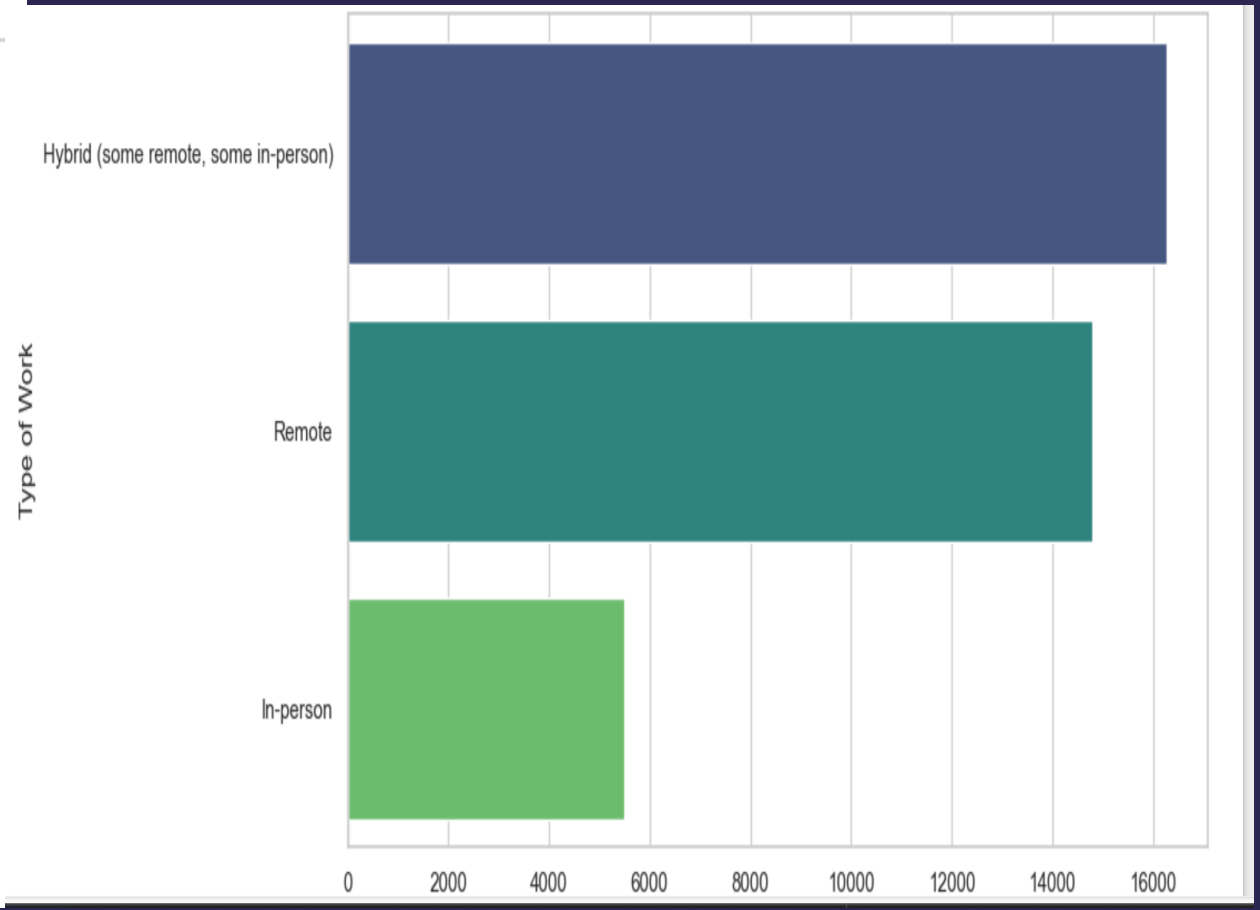
Analyse exploratoire des données (AED) et ingénierie des caractéristiques



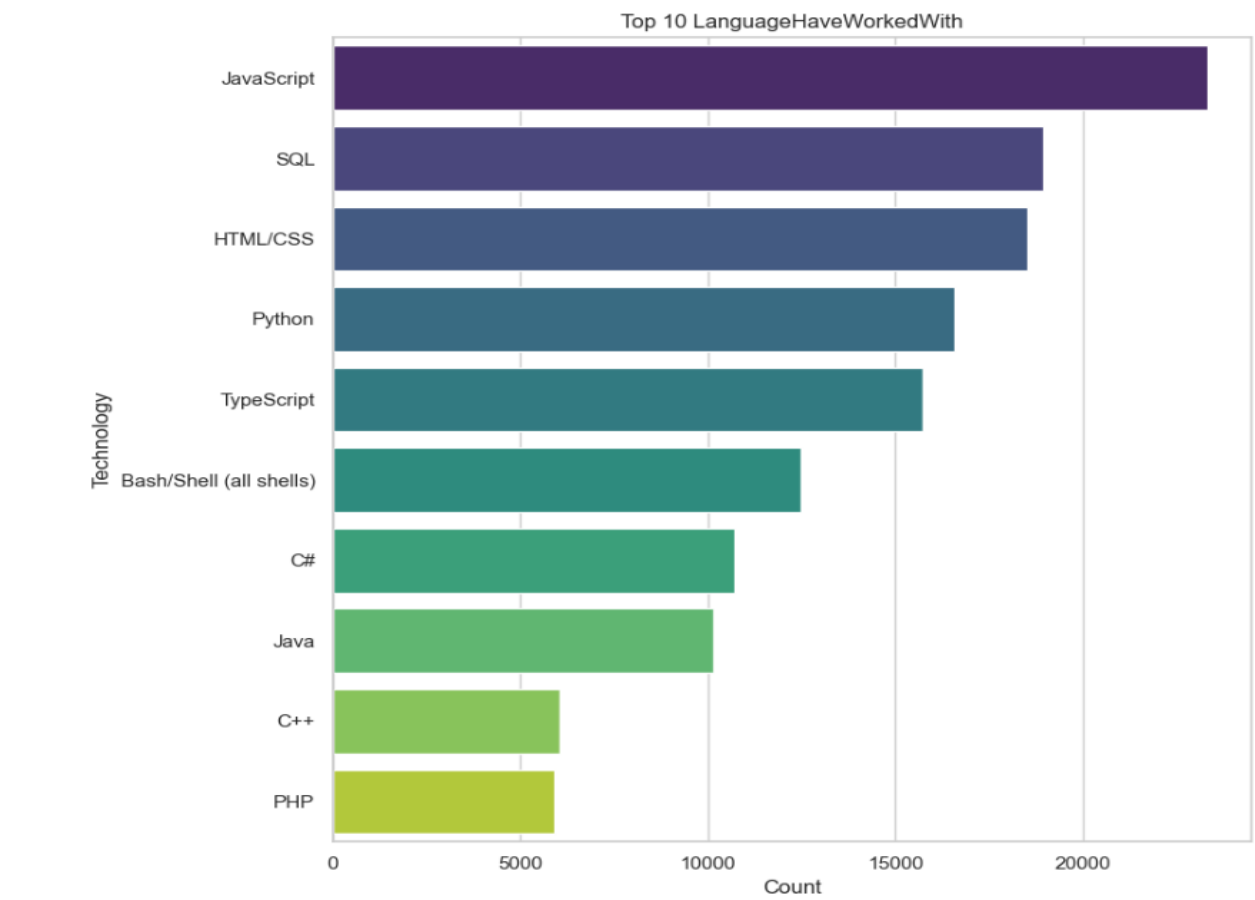


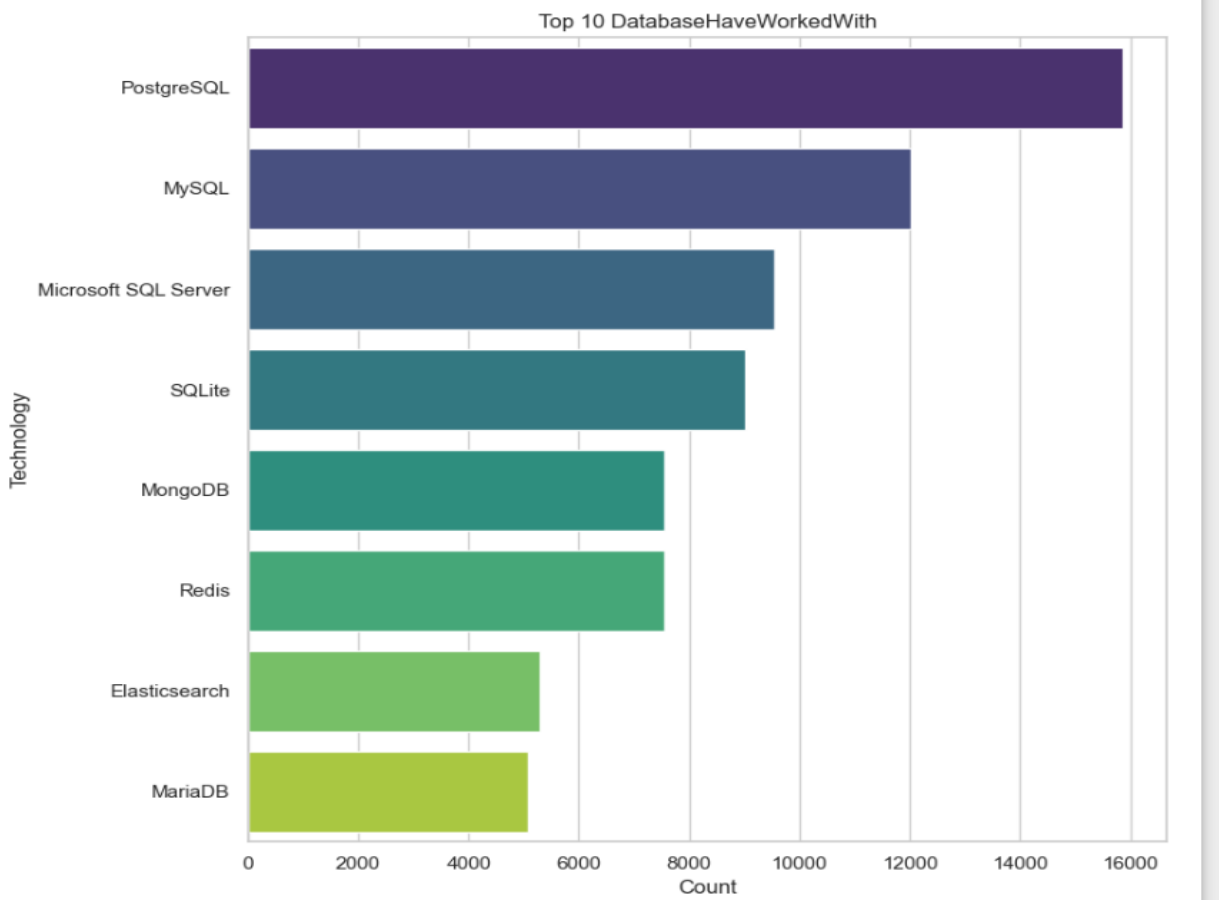
Analyse exploratoire des données (AED) et ingénierie des caractéristiques



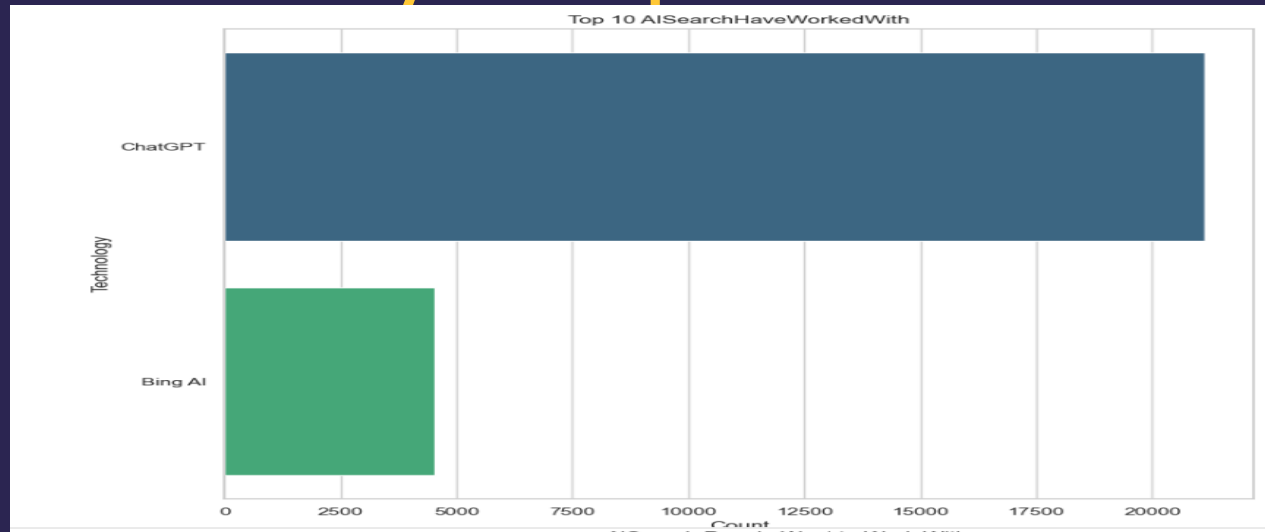


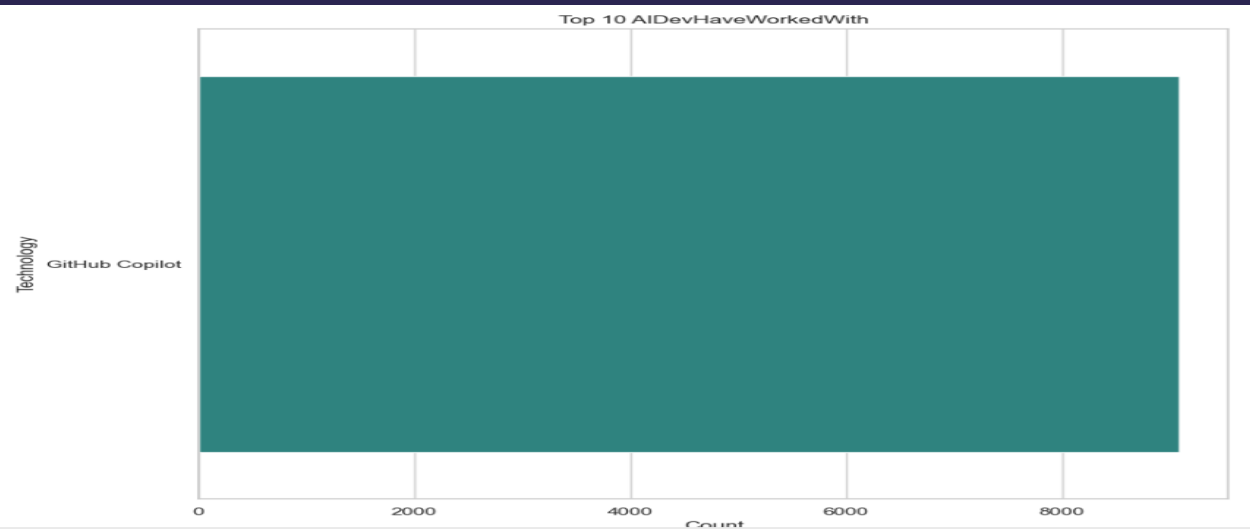
Analyse exploratoire des données (AED)

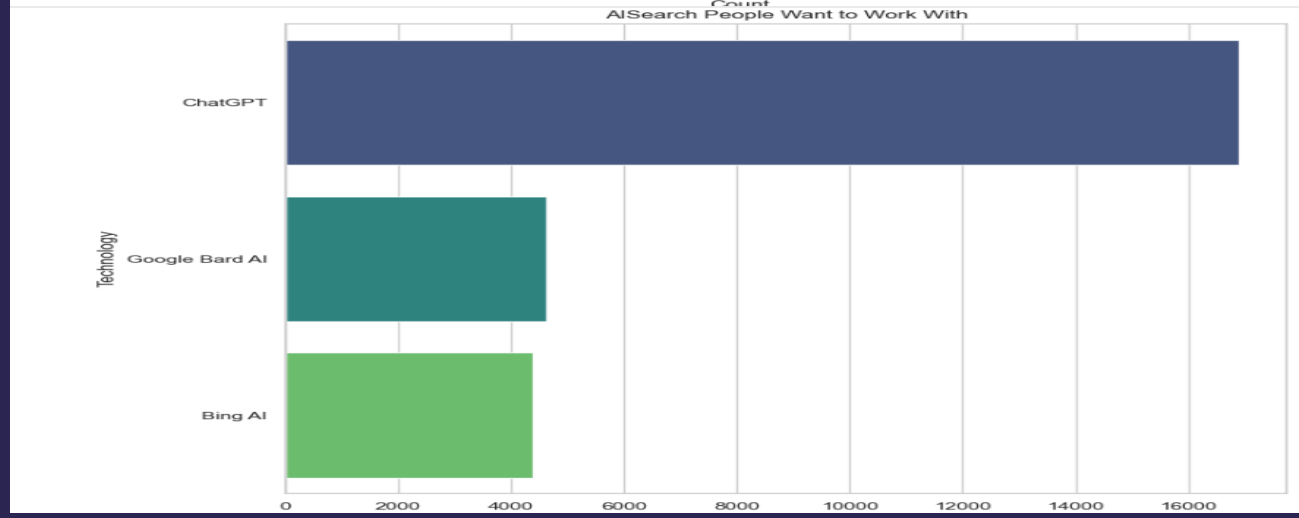


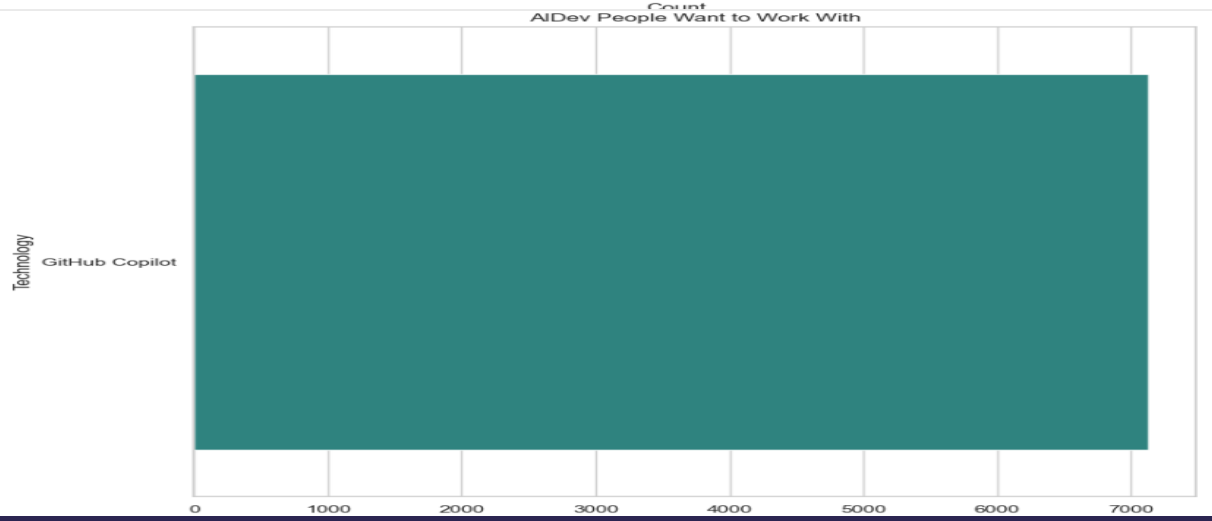


Analyse exploratoire des données

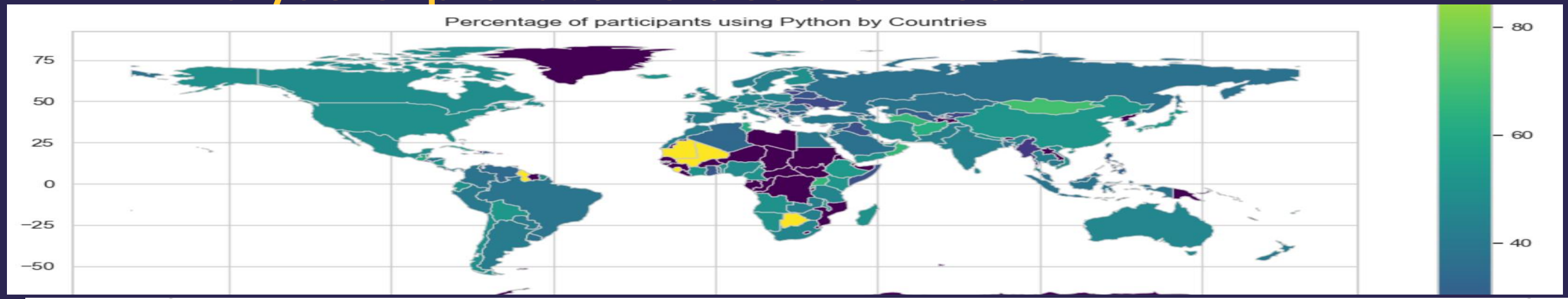


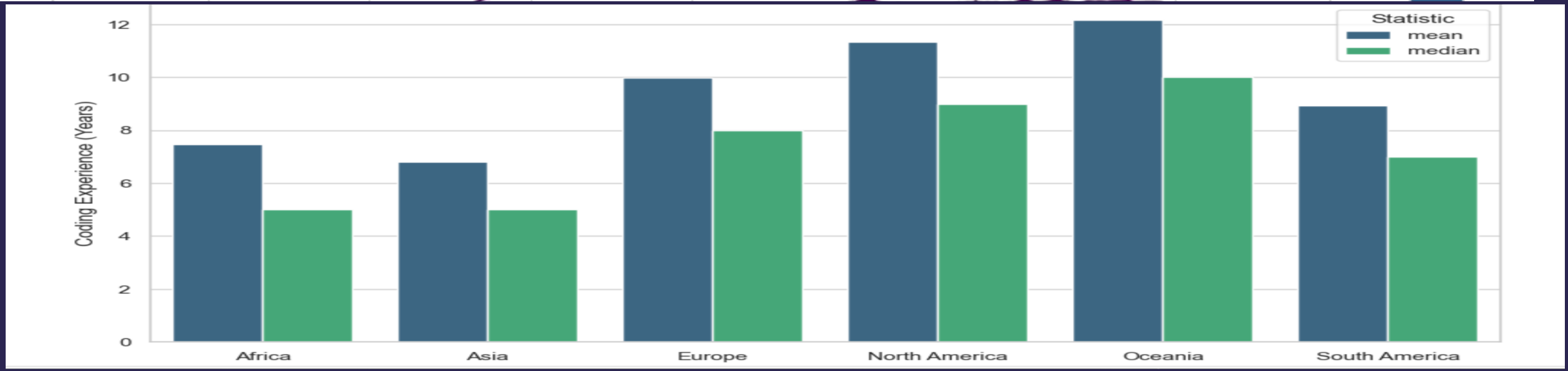




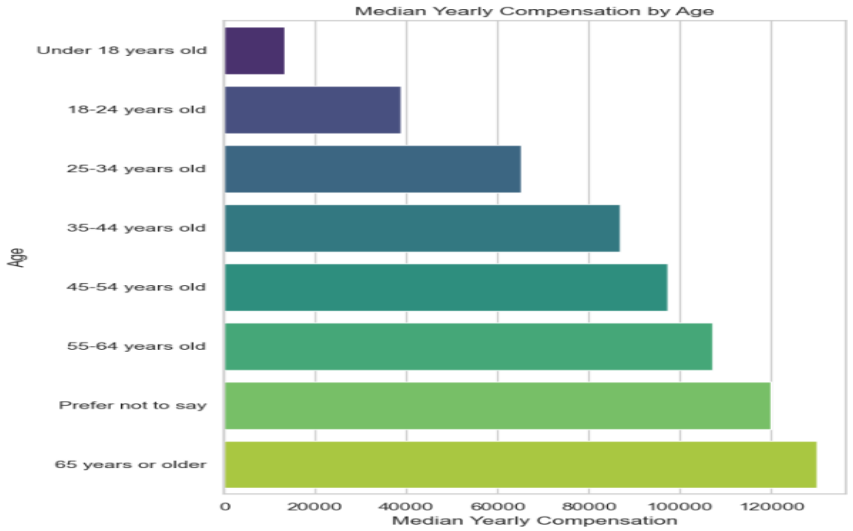


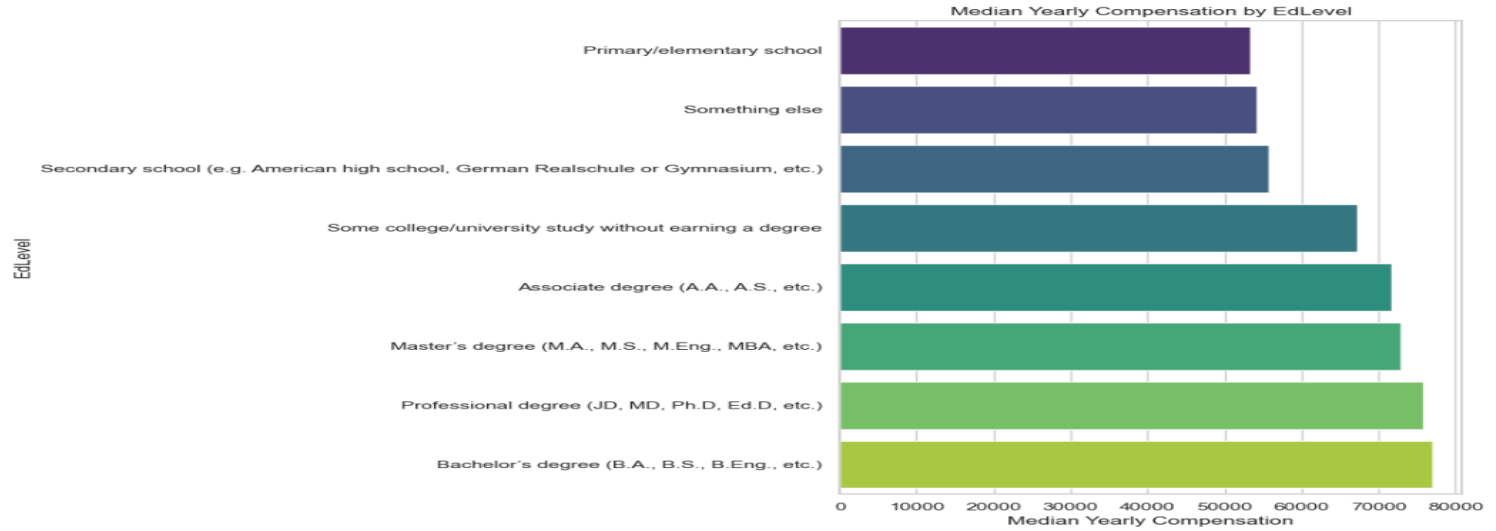
Analyse exploratoire des données

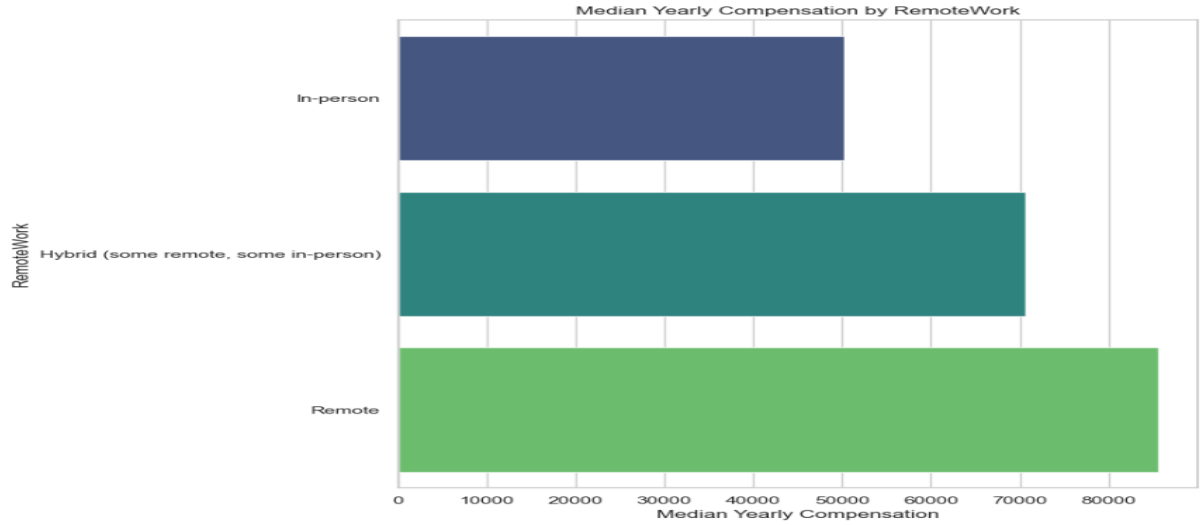


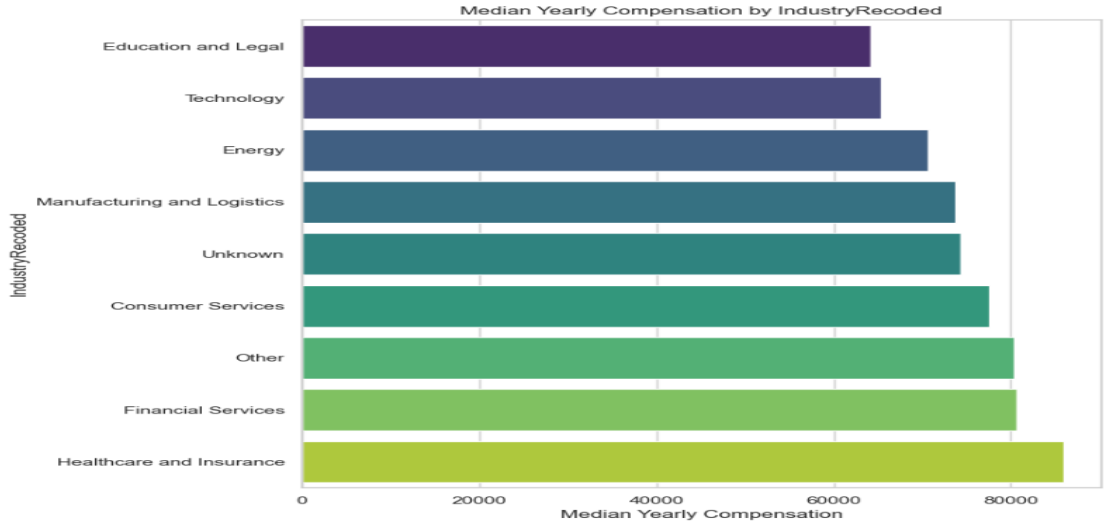


Analyse exploratoire des données

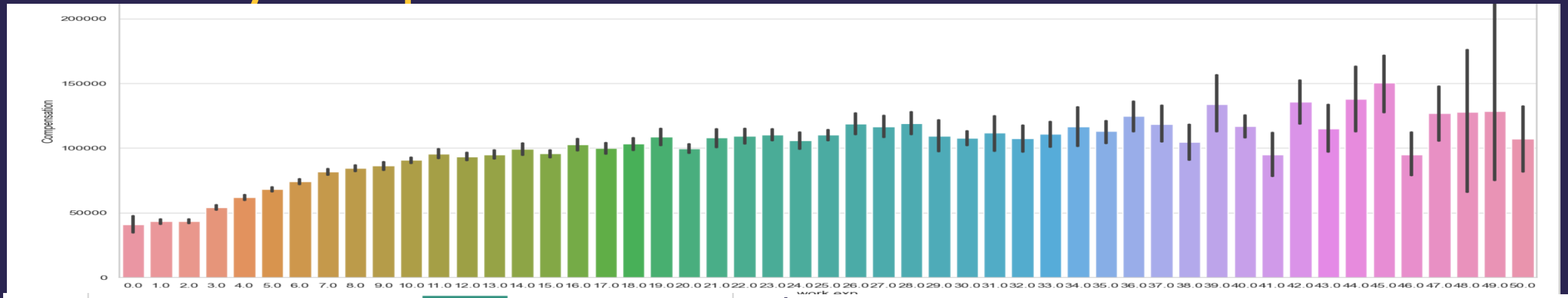


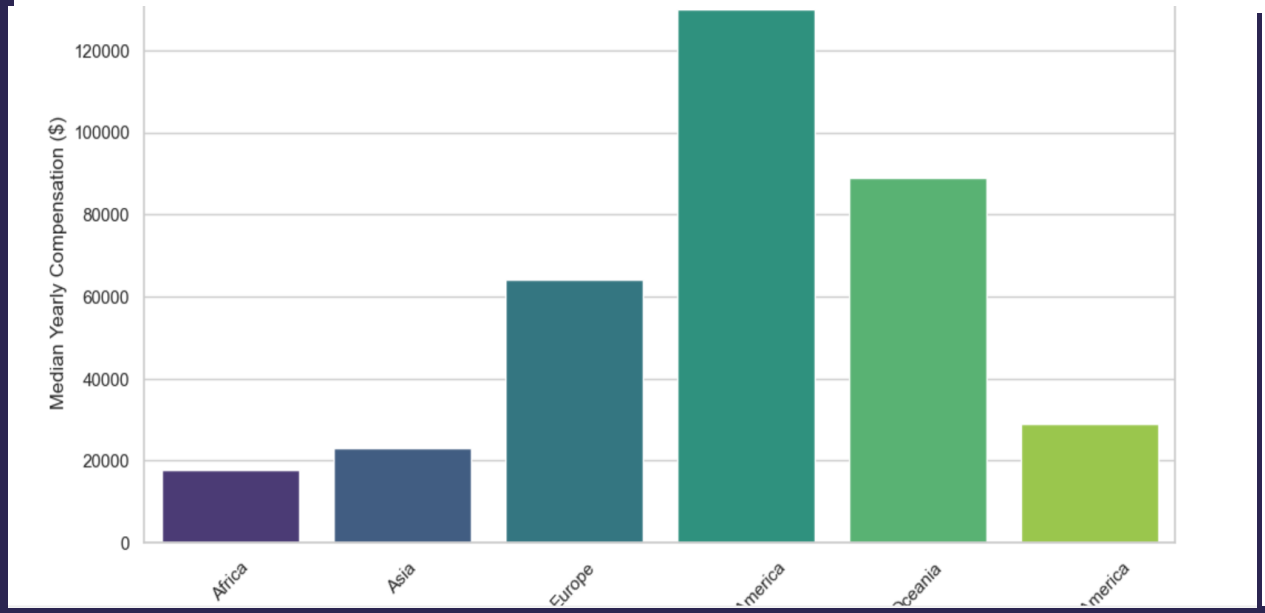


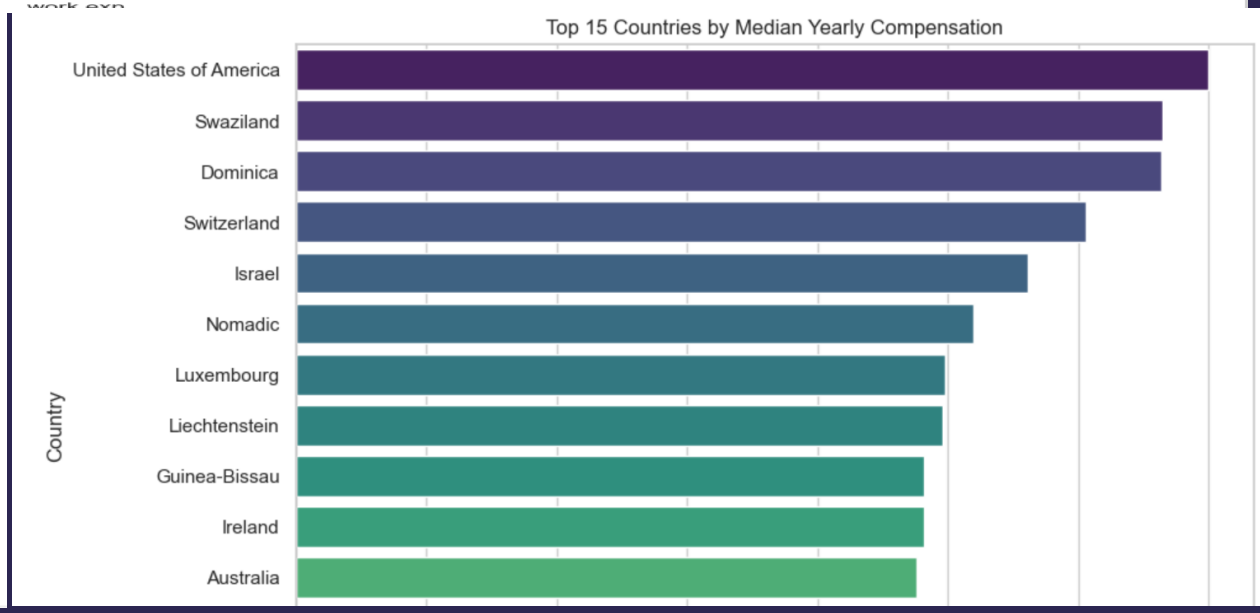




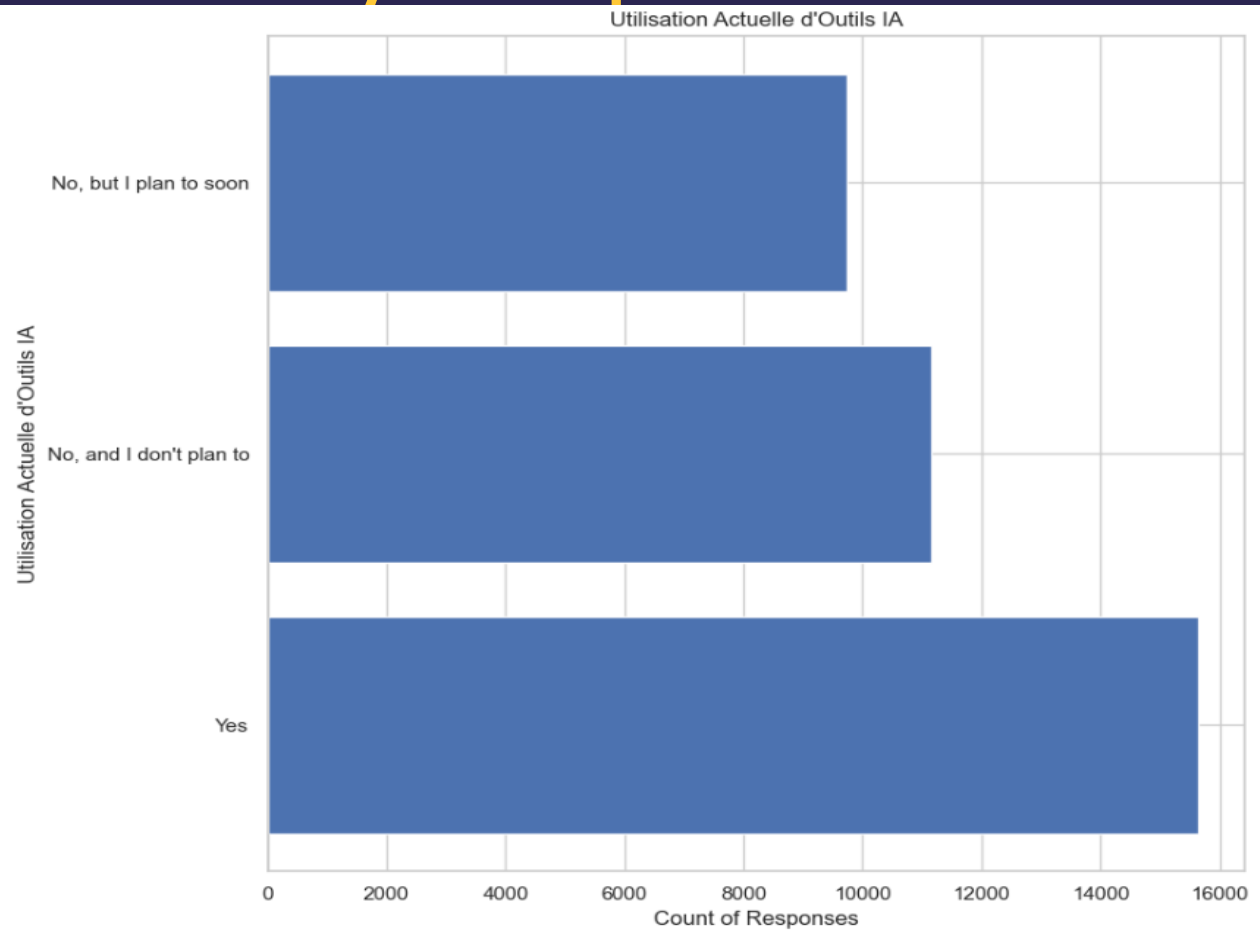
Analyse exploratoire des données

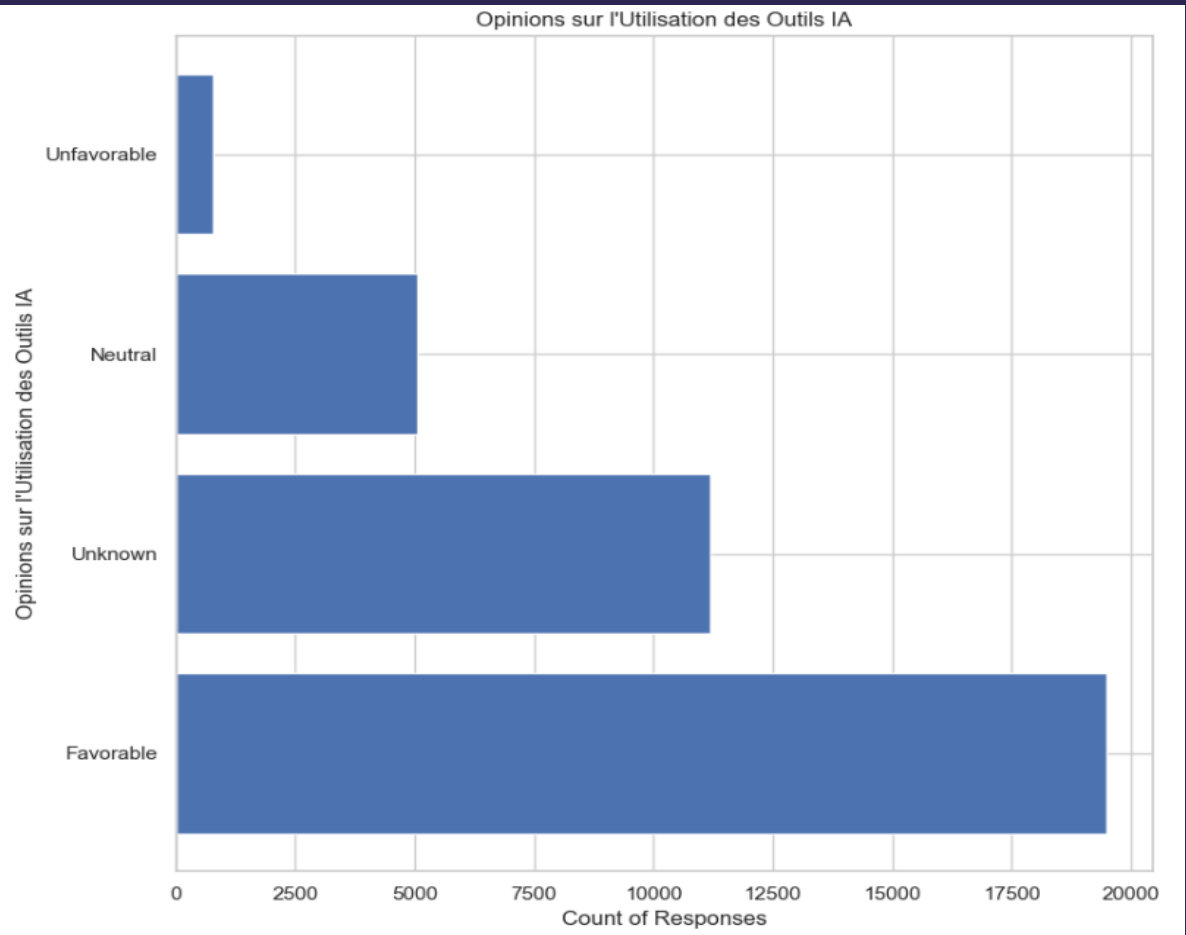






Analyse exploratoire des données





Ingénierie des caractéristiques

```
num_pipeline = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='median')),
    ('poly', PolynomialFeatures(degree=2, interaction_only=True, include_bias=False)),
    ('scaler', MinMaxScaler())
])

cat_pipeline = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('onehot', OneHotEncoder(handle_unknown='ignore'))
])
```

- **Caractéristiques Numériques**
-
- **Imputation** : Les valeurs manquantes dans les caractéristiques numériques sont imputées en utilisant la médiane des valeurs existantes.
- **PolynomialFeatures** : Des termes d'interaction de degré 2 sont ajoutés aux variables numériques. Cela inclut des produits croisés des variables. Cette étape crée de nouvelles caractéristiques qui sont des produits de chaque paire de caractéristiques initiales, ce qui peut aider à modéliser des relations non linéaires entre les caractéristiques et la cible.
- **MinMaxScaler** : Normalisation des caractéristiques numériques pour que leurs valeurs soient transformées dans l'intervalle [0, 1]. Cette mise à l'échelle peut être particulièrement utile avec des algorithmes qui sont sensibles aux échelles des entrées, tels que les algorithmes basés sur les gradients.

Analyse exploratoire des données (AED) et ingénierie des caractéristiques

- **<u>Caractéristiques Catégorielles:</u>**
- **Imputation** : Les valeurs manquantes dans les caractéristiques catégorielles sont imputées en utilisant le mode (la valeur la plus fréquente). Cela garantit que les valeurs imputées sont représentatives des données existantes.
- **OneHotEncoder** : Encodage one-hot des variables catégorielles. Cette méthode transforme chaque valeur catégorielle en une nouvelle caractéristique binaire (0 ou 1), permettant aux modèles de machine learning de traiter efficacement les données catégorielles.
- Nous avons crée une nouvelle variables categorielle ‘Continent’ à partir de la variable ‘Country’.

Modèle

Sélection et entrainement du modèle

```
models = {
    'Linear Regression': LinearRegression(),
    'Decision Tree': DecisionTreeRegressor(),
    'Random Forest': RandomForestRegressor(n_estimators=100),
    'Gradient Boosting': GradientBoostingRegressor(),
    'AdaBoost': AdaBoostRegressor(),
    'Support Vector Regressor': SVR(),
    'K-Neighbors Regressor': KNeighborsRegressor()
}
```

Etape 1 : Utilisation de plusieurs modelés

Choix d'un modèle spécifique :

- Le projet vise à prédire une variable quantitative à partir de diverses caractéristiques.
- Plusieurs modèles de régression, notamment la Régression Linéaire, l'Arbre de Décision, la Forêt Aléatoire, le Gradient Boosting, l'AdaBoost, le SVR, et le KNN seront testées.
- Après, nous comparerons ces modèles en utilisant des mesures telles que le RMSE, le MAE et le R^2 pour évaluer et comparer objectivement leurs performances.
- Après entrainement, la régression linéaire, les forets aléatoires et le gradient boosting se démarquent pour leur équilibre entre erreur et variance. Comme le modèle linéaire est simple, on va continuer avec le gradient boosting et les forets aléatoires.

	LRegression	DecisionTree	RandomForest	GradientBoo	AdaBoost	Support Vect	K-Neighbors
R2	0,64	0,32	0,61	0,62	0,47	-0,01	0,52
MAE		31076	23507	23587	30622	41796	26902
RMSE		44096	33055	32643	38774	53991	36778

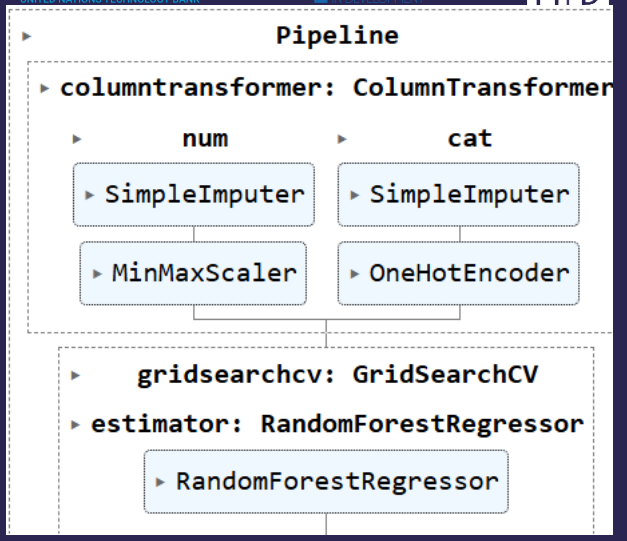
Affinage du modèle et tests

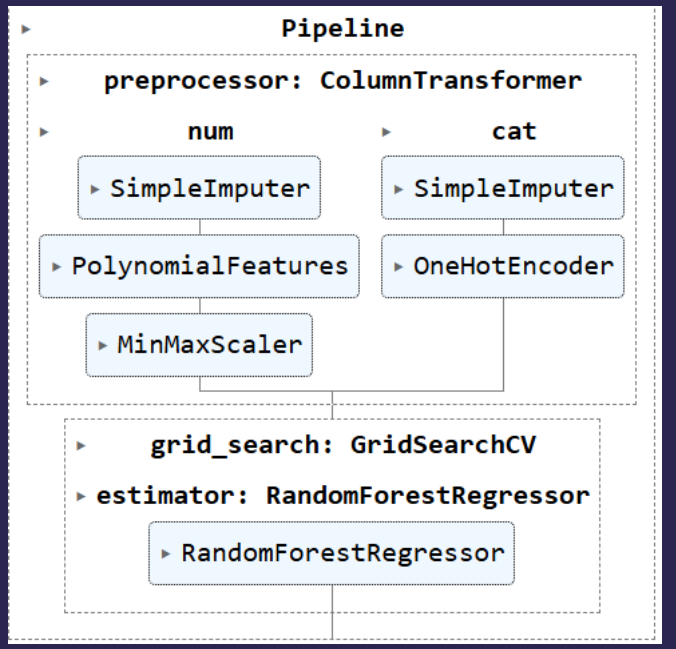
- ## Etape 2 : Utilisation de GridSearchCV

- Le modèle Gradient Boosting affiche des métriques de performance assez robustes avec un **R2 Score**: 0.662 (alors qu'il était de 0.62 avant GridSearchCV).
- Pour le Random Forest, les performances sont légèrement inférieures à celles du Gradient Boosting avec un **R2 Score**: 0.639 (alors qu'il était de 0.617).

- ## Etape 3 : Ajout de PolynomialFeatures(degree2)

- Le modèle Gradient Boosting des performances relativement élevées avec un score **R^2 de 0.671** (alors qu'il était de 0.66 avec le GridSearch et 0.62 sans).
- Le modèle Random Forest montre un score R^2 0.6468 légèrement inférieur à celui du Gradient Boosting.

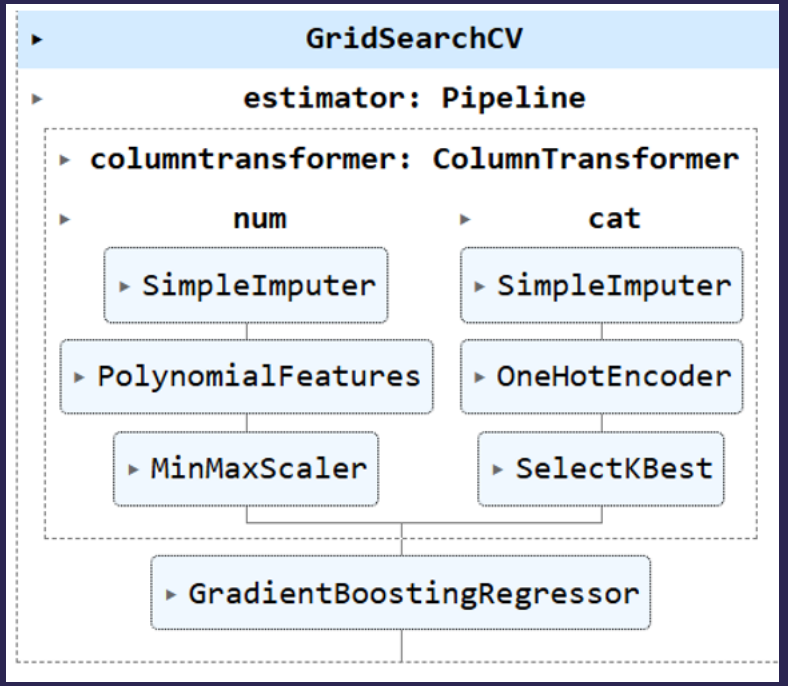




Affinage du modèle et tests

- Justification du choix d'un modèle spécifique
- la profondeur maximale et du nombre d'arbres peut augmenter le risque de surajustement donc pour le reste de nos tentations d'amélioration, nous allons nous concentrer sur le **Gradient Boosting**.
- Etape 4 : Ajout de SelectKBest = 15
- Comme nous avions choisi manuellement les variables qui nous semblait important 'à nos yeux', nous allons faire une analyse plus poussée pour identifier les variables les plus informatives pourrait être bénéfique. Disons pour commencer les 15 variables !
- La performance de notre modèle a chuté à 0.59, ce qui indique que ce n'était pas une bonne stratégie.

Choix final: Notre modèle final sera donc le Gradient Boosting avec un score **R^2 de 0.671.**



Résultats

Recommandations

1. **Développeurs :**
   - ❖ Jeunes : Se spécialiser en développement web et Python.
   - ❖ Data Science/Machine Learning : Envisager ces domaines pour des opportunités élitistes.
   - ❖ Apprentissage : Utiliser les ressources en ligne pour l'apprentissage continu.
2. **Universitaires :**
   - ❖ Programmes : Intégrer plus de cours pratiques en développement web et Python.
   - ❖ Spécialisation : Proposer des modules en data science et machine learning.
3. **Personnes en Réorientation :**
   - ❖ Spécialisation : Choisir le développement web et Python.
   - ❖ Formation : Utiliser des ressources en ligne et des formations sur le poste de travail.

Recommandations

4. Décideurs Politiques :

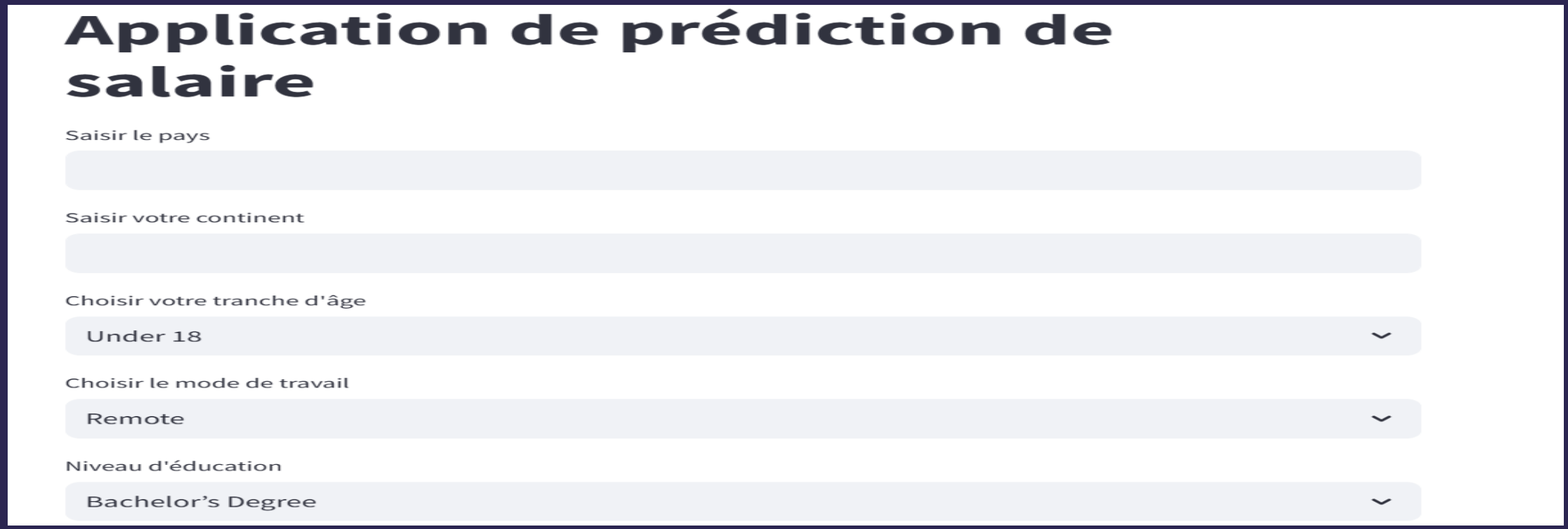
1. Éducation/Formation : Promouvoir des politiques de soutien à la formation en technologies demandées.
2. Télétravail : Faciliter le travail hybride.

5**. Organisations de Développement :**

1. Programmes de Formation : Cibler le développement web, Python, et data science/machine learning.
2. Apprentissage Autonome : Encourager l'utilisation de ressources en ligne.

Déploiement

- Merci de cliquer ICI



Travaux futurs

- Inclure les consultants et indépendants pour offrir des perspectives sur les opportunités freelance
- Comprendre les Préférences des Développeurs en Matière d'Environnement de Travail (préférences pour le travail à distance, sur site ou hybride), ce qui contribue à l'amélioration du bien-être au travail en alignement avec l’ODD 3 (Bonne santé et bien-être) et aura un impact sur les politiques de planification urbaine pour soutenir l'ODD 11 (Villes et communautés durables)

References

- https://scikit-learn.org/
- https://www.kaggle.com/

Thank you!


ISTANBUL INTERNATIONAL
CENTER FOR PRIVATE SECTOR
IN DEVELOPMENT


S D G
A I
L A B


UN
TECHNOLOGY
UNITED NATIONS TECHNOLOGY BANK
FOR LEAST DEVELOPED COUNTRIES