Projet Maths Data Science

Equipe Les Anglais : Jad Salloum, Nabil Hatri, Monali Patel, Yassine Essamadi

Sommaire :

1. La base de données utilisée
2. Data cleaning
3. Problématiques posées
4. Annexe

1. La base de données

La base de données étudiée est le résultat d'un formulaire complété par 64461 personnes qui utilisent StackOverflow.

Elle est comprise de 61 colonnes représentant l'ID généré automatiquement par le formulaire, ainsi que toutes questions du formulaire.

Les questions du formulaire se portent sur des informations sur la personne qui l'a complété, leur travail avec des outils et langage informatique, leur niveau d'étude et la situation de leur travail ou recherche de travail.

Les valeurs des colonnes sont présentées en Integer, Float, String, avec 23,49 % de valeurs Not as Number (NaN) [1], soit un total de 908664 valeurs.

2. Data cleaning

Le data cleaning est un processus très important dans l'analyse d'une base de données et sa restitution dans une dataframe.

En effet, les données peuvent rapidement se trouver mal étiquetées ou dupliquées au sein d'un même ensemble. Les données corrompues peuvent aussi se joindre au reste et peut souvent modifier voire falsifier les résultats obtenus.

C'est pour cela qu'avant de commencer les analyses et de les modéliser sur des graphiques, nous faisons un nettoyage des données.

Premièrement, nous avons cherché un moyen d'identifier plus facilement les valeurs de la base de données, et trouvé que la colonne « Respondent » pouvait servir d'ID pour notre base.

En effet, celle-ci contient :

- Des valeurs Int chaque valeur étant unique
- Ne possède pas de valeur NaN

Nous avons donc importé la base de données dans notre dataframe en prenant « Respondent » comme ID. [2]

Les colonnes SurveyEase et SurveyLenght ont été supprimées du dataframe : Elles représentent respectivement la facilité dans laquelle le formulaire a été rempli, ainsi que la durée de la complétion du formulaire.

Nous avons jugé ces deux colonnes comme non importantes pour notre étude, et qu'il était plus pertinent de les retirer.

Sur les colonnes restantes, nous avons jugé nécessaire de remplacer les valeurs NaN en fonction de la colonne par une valeur expliquant le manque de valeur : [3]

Nous supposons ici que le formulaire possède une option de sélection pour les personnes ne souhaitant pas répondre. Nous supposons aussi que certaines de ces valeurs ont pu être mal entrées dans le fichier, soit par mauvais format de valeur, ou autre.

3. Problématiques posées

La base de données employée contient une multitude de colonnes, c'est pour cela que nous avons trouvé pertinent de recenser quelques problématiques et de répondre à ces dernières, ce qui est susceptible de nous fournir des éléments de réponse à celles-ci.

Certaines de ces problématiques sont à intérêt réel, et aident à s'adapter aux besoins du monde du développement.

1 : Quelles sont les caractéristiques des développeurs qui gagnent le plus ?

Quels facteurs peuvent influencer le salaire des développeurs ?

Nous avons trouvé sur la base de données que le salaire avait tendance à beaucoup varier d'un développeur à un autre et nous voulions connaitre si cela est dû à des facteurs spécifiques ou non.

Afin de rechercher une réponse à cette problématique, nous avons tout d'abord recherché les colonnes d'intérêt parmi le dataframe : les données liées aux développeurs, leur salaire et les facteurs pouvant influencer.

Les colonnes « ConvertedComp » et « CompFreq » nous informent du salaire en US dollars et de sa fréquence (weekly, monthly, yearly) [4], et la colonne « MainBranch » permet de séparer les développeurs des autres professions.

Cependant, nous rencontrons un problème : les salaires sont séparés par leur fréquence. Pour les uniformiser, nous avons converti Monthly en Yearly (Yearly = Monthly * 11) et Weekly en Yearly (Yearly = Weekly * 4 * 11). Nous avons supposé que les développeurs prennent un mois de congé pour le calcul du salaire.

Afin de déterminer au mieux les colonnes qui peuvent influencer de façon signifiante les salaires, nous allons rechercher des problématiques intermédiaires :

1.a. Est-ce que le fait de commencer le développement plus tôt influence le salaire des développeurs ?

Pour cette partie, nous allons nous intéresser aux colonnes « Age1stCode », « ConvertedComp » et « MainBranch ».

« Age1stCode » représente l'âge auquel la personne ayant rempli le formulaire a commencé à s'exercer au développement.

Après avoir tracé les graphiques (lineplot) [5], nous avons remarqué un manque de volume de données sur « Age1stCode » à partir de 45 ans et plus. Nous n'avons donc pas pris en compte cette partie.

Après analyse des graphs, nous pouvons conclure que les personnes ayant commencé le développement le plus tôt tendent généralement à recevoir un plus grand salaire. On remarque cependant une exception : la tendance remonte à partir de 20 ans. En effet, On peut supposer qu'une personne aurait déjà atteint un salaire plus élevé par son expérience, contrairement aux moins de 20 ans, toujours en période d'apprentissage. On peut donc dire qu'un facteur qui influe sur le salaire est l'âge du premier code.

2 : Quels sont les web frameworks les plus recherchés en développement

Dans cette partie nous cherchons à définir les web frameworks les plus utilisés, et les plus demandés afin d'en conclure une différence si elle existe.

2.a : Les web frameworks les plus utilisés aujourd'hui :

Nous cherchons d'abord à définir quels sont les web frameworks les plus utilisés, pour cela nous avons travaillé sur la colonne «WebframeWorkedWith», et nous avons choisi de modéliser le résultat obtenu dans un histogramme [6]. On remarquera que les trois frameworks web les plus utilisés sont : JQuery, React.js et Angular.

2.b : Les web frameworks les plus demandés aujourd'hui :

Le formulaire de StackOverflow contient une question sur les web frameworks désirés pour l'an prochain par les développeurs, nous avons donc travaillé sur la colonne «WebframeDesireNextYear». Comme pour **2.a** nous avons choisi d'afficher les résultats obtenus dans un histogramme [7].

Nous remarquons que les trois frameworks web les plus demandés sont : React.js, Vue.js et Angular.js.

En comparant **2.a** et **2.b** on remarque que les web frameworks les plus utilisés ne correspondent pas forcément à ce que les développeurs recherchent le plus. De ce fait nous pouvons conclure que les développeurs sont en perpétuel recherche d'évolution et veulent toujours travailler avec les meilleures technologies du moment.

Nous remarquons aussi que les frameworks web front-end du langage de programmation **JavaScript** sont les plus représentatifs des applications internet d'aujourd'hui.

3 : Combien de développeurs continuent à apprendre de nouvelles technologies?

Nous avons ensuite cherché à nous intéresser à la répartition des développeurs par rapport à leur fréquence d'apprentissage de nouvelles technologies. Pour cela nous avons travaillé sur la colonne «NEWLearn» de notre DataFrame, et nous avons choisi d'afficher les résultats obtenus dans une pie chart [8]. On peut ainsi mieux faire la distinction entre les résultats obtenus.

Avec le graphe obtenu, on peut observer que la plupart des développeurs apprennent une nouvelle technologie entre quelques mois et une fois par an.

On peut en conclure que le monde du développement informatique est toujours en évolution, ce qui pousse les développeurs à s'adapter en apprenant des nouvelles technologies.

4 : Quelle est la proportion des Systèmes d'exploitation les plus utilisés par les développeurs.

Les outils de développement sont variés, et permettent une flexibilité importante en fonction de leur utilisation. Cependant chaque développeur a sa propre manière de travailler ainsi que ses outils de prédilection. Un de ces outils est le Système d'exploitation dont les principaux sont : Windows, Linux, MacOS.

Nous avons étudié la colonne «OpSys», et nous avons choisi d'afficher les résultats dans un histogramme [9].

Nous remarquons que l'OS le plus utilisé est Windows, suivi par Linux et MacOS. Le système le moins utilisé est BSD.

5 : Répartition des développeurs par rapport à leur niveau d'éducation.

Nous avons trouvé pertinent de s'intéresser à la répartition générale du niveau d'éducation des développeurs, et pour cela nous avons utilisé la colonne : «EdLevel», et nous avons choisi d'afficher les résultats obtenus dans un pie chart [10].

Le graphique nous montre que la plupart des développeurs possèdent un «Bachelor's degree» (équivalent à un bac+3). Une grande partie des autres développeurs ont un «Master's degree» (équivalent à un bac+5).

On conclue donc qu'une bonne partie des développeurs atteignent un diplôme de bac+3 ou équivalent.

6 : Répartition de la fréquence des salaires des développeurs.

La répartition des salaires peut se montrer avec la colonne : «CompFreq».

Cette information reflète la situation de travail des développeurs.

Nous avons choisi de la représenter dans un pie chart [11].

Nous remarquons que 49 % des développeurs reçoivent un salaire annuel et 47 % d'entre eux un salaire mensuel, et 3 % un salaire hebdomadaire.

7 : Les principaux facteurs qui donneraient l'avantage à une offre de travail par rapport à une autre.

Les développeurs informatiques possèdent une grande variété d'emplois parmi lesquels choisir. Il y a en général plus d'offres que de demandes. C'est pour cela que nous avons jugé pertinent de définir les principaux facteurs qui tendraient un développeur à choisir une offre plutôt qu'une autre.

Nous avons travaillé sur la colonne : «JobFactors» et nous avons choisi de modéliser le résultat avec un WordCloud [12].

De par notre graphique nous pouvons conclure que les principaux facteurs sont : Les technologies, Langages et frameworks, culture et environnement de travail.

8 : Répartition des DevSecOps dans les entreprises.

Le métier de DevOps n'a cessé d'accroître ces dernières années, et a pris une place importante dans les grandes entreprises. C'est pour cela que nous avons trouvé intéressant de répertorier la présence ou absence de ces derniers dans les entreprises.

Nous avons travaillé sur la colonne : «NEWDevOps», et nous avons affiché les résultats dans une pie chart [13].

L'analyse de celle-ci nous informe qu'il y a autant de développeurs qui ont reporté une présence, que de développeurs qui ont reporté une absence.

9 : Répartition des développeurs par rapport à leur situation professionnelle.

Dans notre DataFrame, on remarque qu'une démographie variée se trouve parmi les développeurs. En effet on peut le voir très clairement dans la colonne que l'on explore ici : «Employment».

On cherche à savoir quelle est la répartition des développeurs selon leur situation professionnelle. Pour cela nous avons utilisé un histogramme [14].

Après analyse de notre graphique, nous remarquons les développeurs employés à temps plein sont de loin les plus nombreux dans la base de données (près de 40.000), tandis que les indépendants constituent une infime partie de la population de notre DataFrame.

9.a Répartition des développeurs ayant une Licence par rapport à leur situation professionnelle.

Après avoir étudié la répartition des développeurs par rapport à leur situation professionnelle en général, nous avons porté intérêt à l'impact des diplômes sur la situation professionnelle, en commençant par la licence.

Nous avons affiché les résultats obtenus dans un histogramme [15].

Nous pouvons voir un résultat similaire à l'analyse générale, cependant avec une plus grande répartition des développeurs indépendants.

9.b Répartition des développeurs ayant un Master par rapport à leur situation professionnelle.

Nous passons maintenant à l'analyse de la situation professionnelle des développeurs ayant un Master ou diplôme équivalent.

Nous avons affiché les résultats obtenus dans un histogramme similaire [16].

Le résultat de celui-ci nous montre une même tendance que la comparaison entre le graphique générale et celui de la licence : on retrouve une plus grande proportion de développeurs indépendants comparé aux développeurs ayant une Licence.

On peut donc en conclure que plus le diplôme obtenu est important plus les développeurs se tournent vers l'auto-entreprenariat : celui-ci leur permet plus de flexibilité et fait en sorte qu'ils aient moins besoin du support d'une entreprise pour se lancer.

10. Y a-t-il une relation entre l'expérience professionnelle et le niveau d'éducation ?

Nous avons remarqué lors de notre étude, et de la résolution de nos problématiques précédentes que le niveau d'éducation des développeurs influé souvent sur plusieurs autres paramètres, c'est donc pour cela que nous avons voulu voir s'il y avait une relation entre l'expérience professionnelle et le niveau d'éducation.

Nous avons travaillé sur deux colonnes différentes : «YearsCodePro» et «EdLevel».

Nous avons choisi un histogramme [17] pour modéliser les résultats obtenus.

4. Annexe

1.

```
[291]: nb_cells = np.product(df.shape)
       nb_na = df.isna().sum().sum()
       nb_na / nb_cells * 100
```

```
[291]: 23.493895533733575
```

2.

```
df = pd.read_csv("developer_survey_2020/survey_results_public.csv", index_col="Respondent")

df
```

Respondent	MainBranch	Hobbyist	Age	Age1stCode	CompFreq	CompTotal	ConvertedComp	Country	CurrencyDesc	CurrencySymbol	...	SurveyEase	SurveyLeng
1	I am a developer by profession	Yes	NaN	13	Monthly	NaN	NaN	Germany	European Euro	EUR	...	Neither easy nor difficult	Appropriate leng
2	I am a developer by profession	No	NaN	19	NaN	NaN	NaN	United Kingdom	Pound sterling	GBP	...	NaN	Na
3	I code primarily as a hobby	Yes	NaN	15	NaN	NaN	NaN	Russian Federation	NaN	NaN	...	Neither easy nor difficult	Appropriate leng
4	I am a developer by profession	Yes	25.0	18	NaN	NaN	NaN	Albania	Albanian lek	ALL	...	NaN	Na
5	I used to be a developer by profession, but no...	Yes	31.0	16	NaN	NaN	NaN	United States	NaN	NaN	...	Easy	Too sh
...	...	...	...	...	...	...	...	...	...	...	...	...	

3.

```
# Les différentes valeurs présentes sur la colonne 'Trans' avant modifs

pd.unique(df['Trans'])
```

```
array(['No', nan, 'Yes'], dtype=object)
```

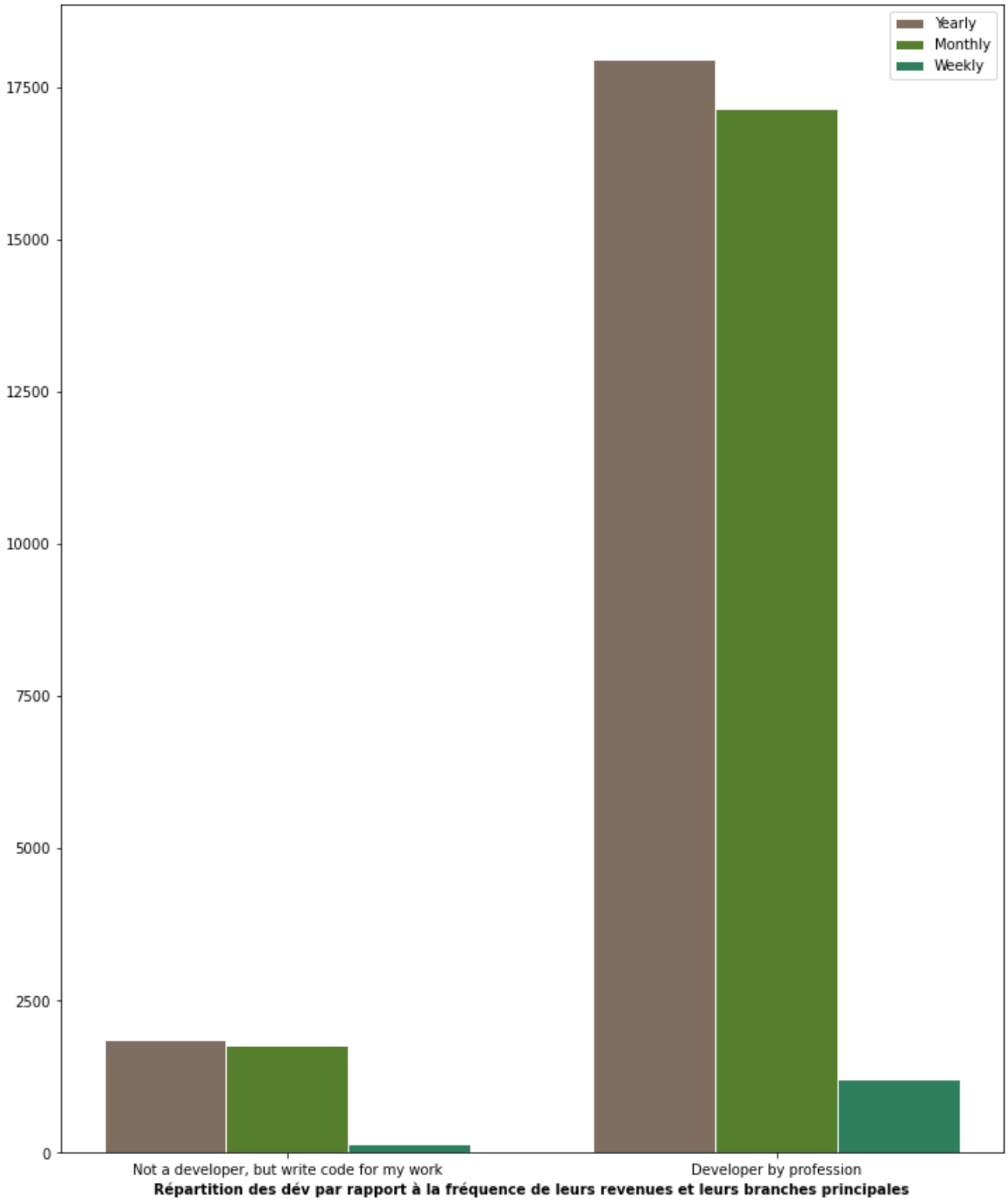
```
df['Trans'].fillna('Not mentionned' , inplace= True)
```

```
# Les différentes valeurs présentes sur la colonne 'Trans' après modifs

pd.unique(df['Trans'])
```

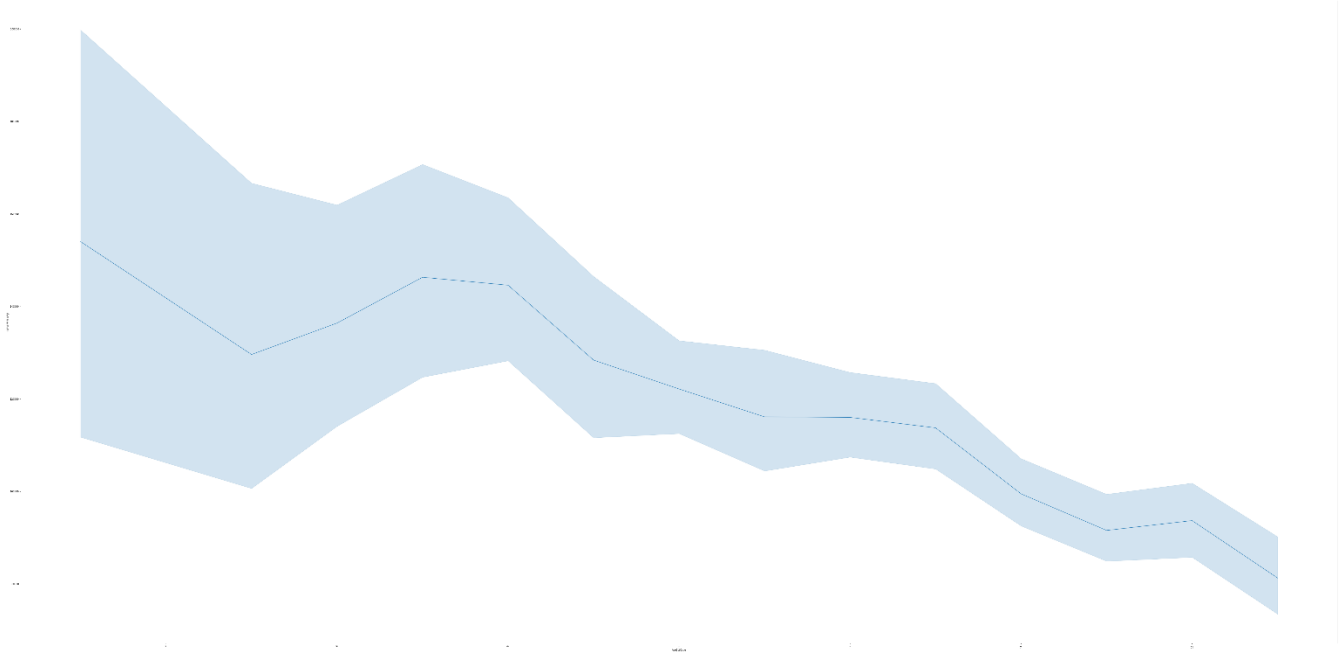
```
array(['No', 'Not mentionned', 'Yes'], dtype=object)
```

4.

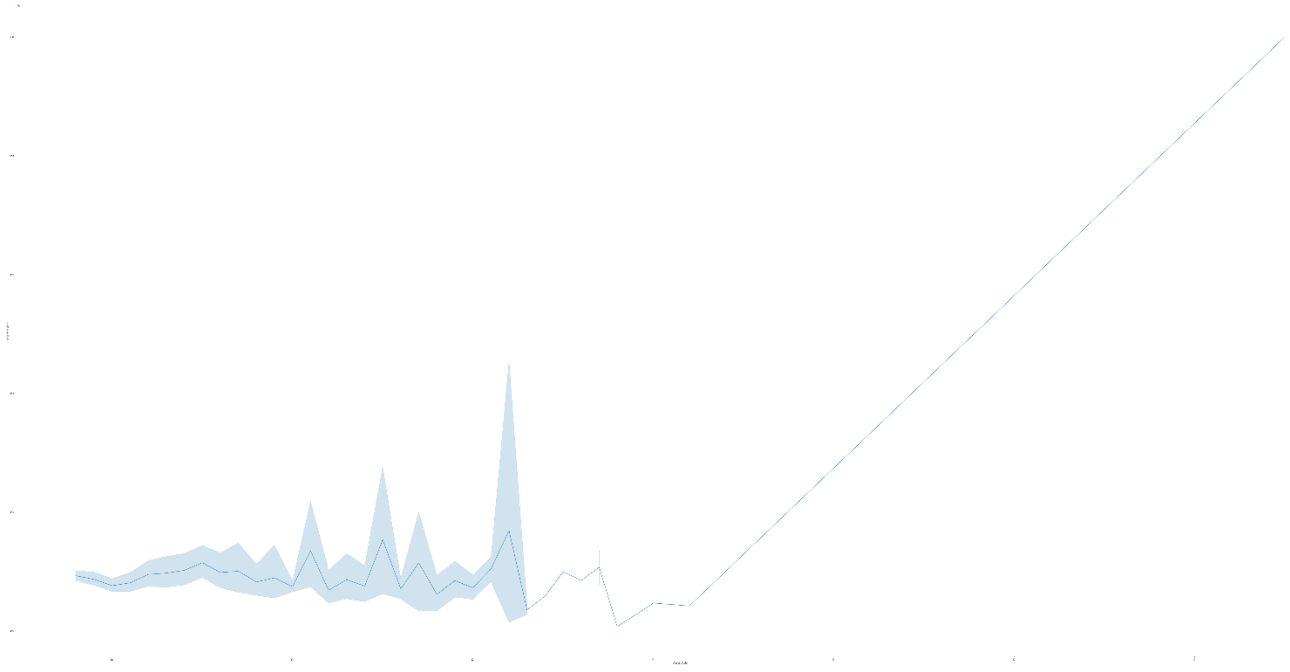

Yearly
Monthly
Weekly
17500
15000
12500
10000
7500
5000
2500
0
Not a developer, but write code for my work
Developer by profession
Répartition des dév par rapport à la fréquence de leurs revenues et leurs branches principales

5.

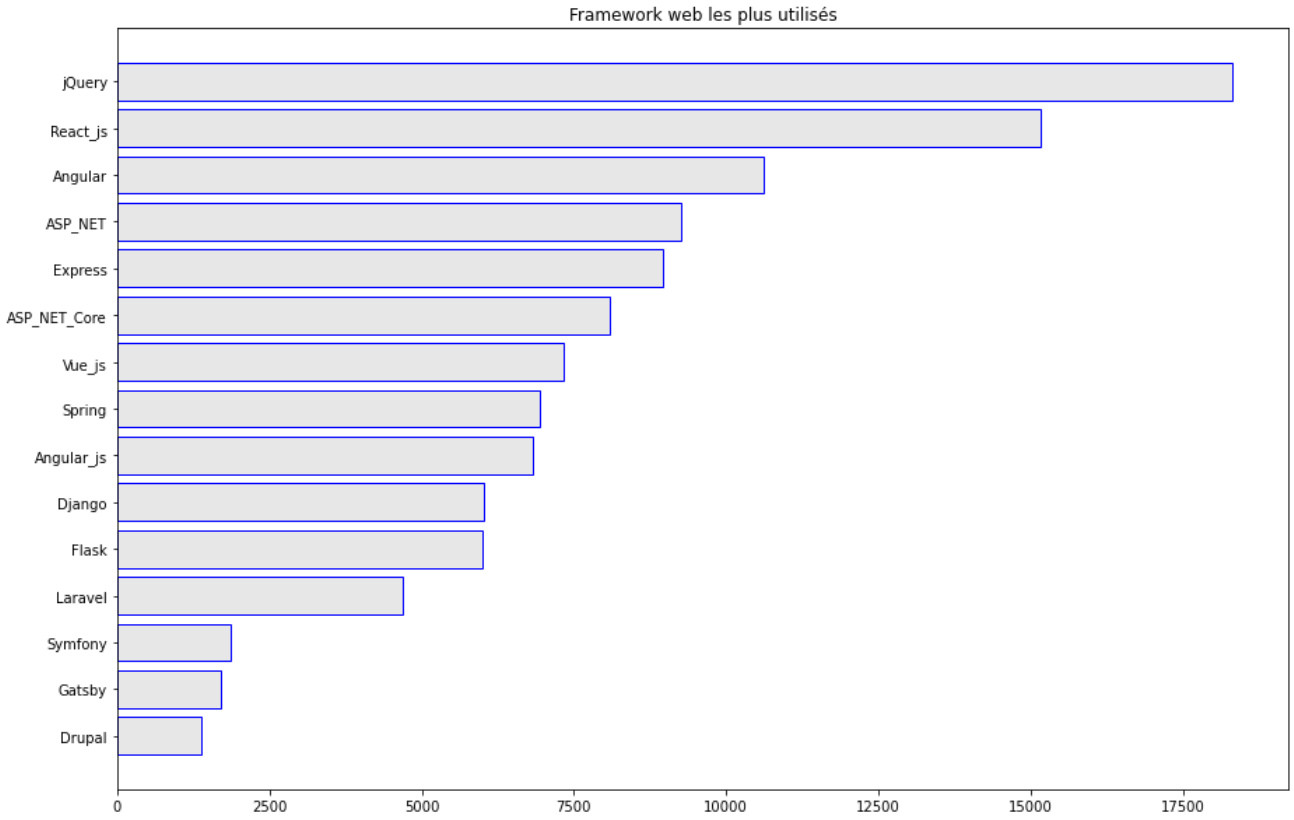
Salaire des développeurs par rapport à l'âge de leur premier code (moins de 18 ans)



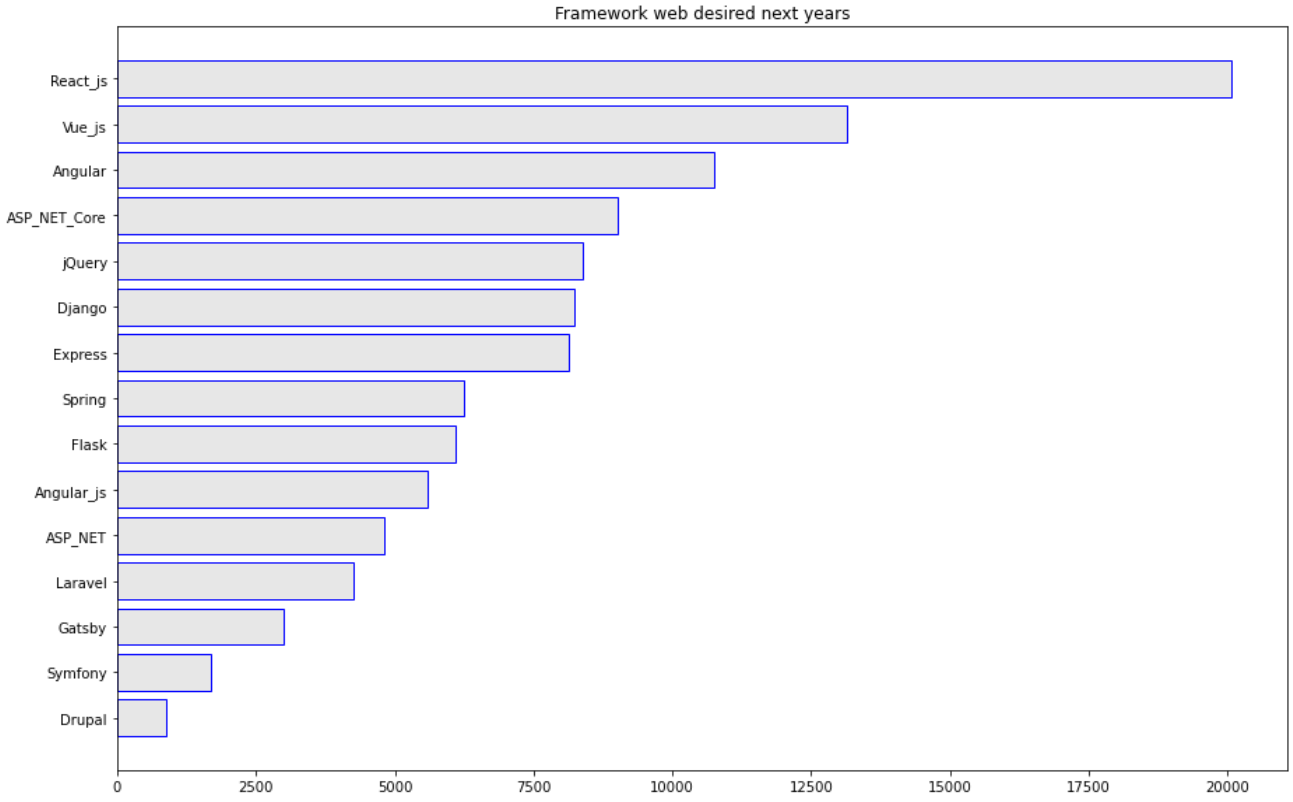
Salaire des développeurs par rapport à l'âge de leur premier code (18 ans et plus)



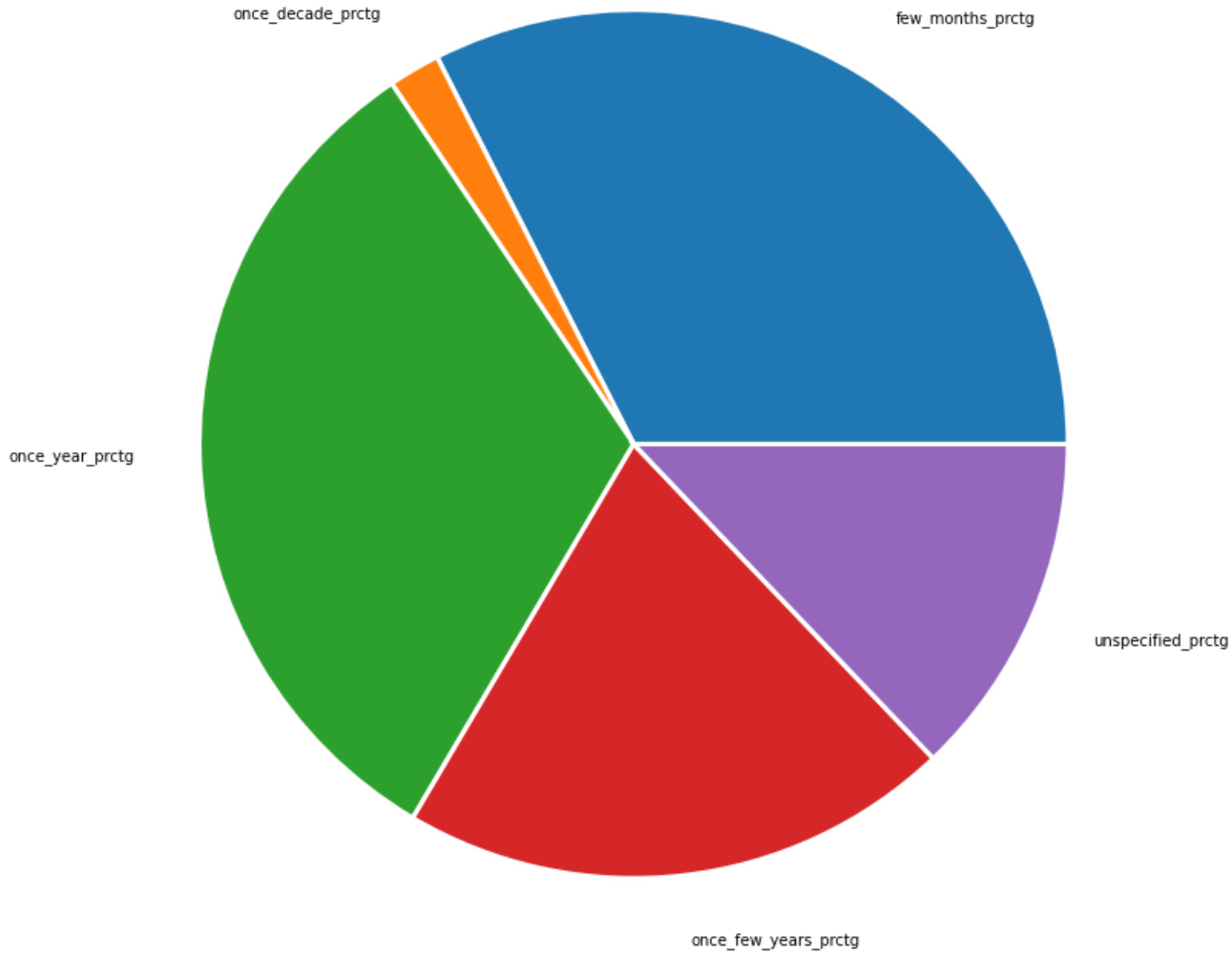
6.


Framework web les plus utilisés
jQuery
React_js
Angular
ASP_NET
Express
ASP_NET_Core
Vue_js
Spring
Angular_js
Django
Flask
Laravel
Symfony
Gatsby
Drupal
0
2500
5000
7500
10000
12500
15000
17500

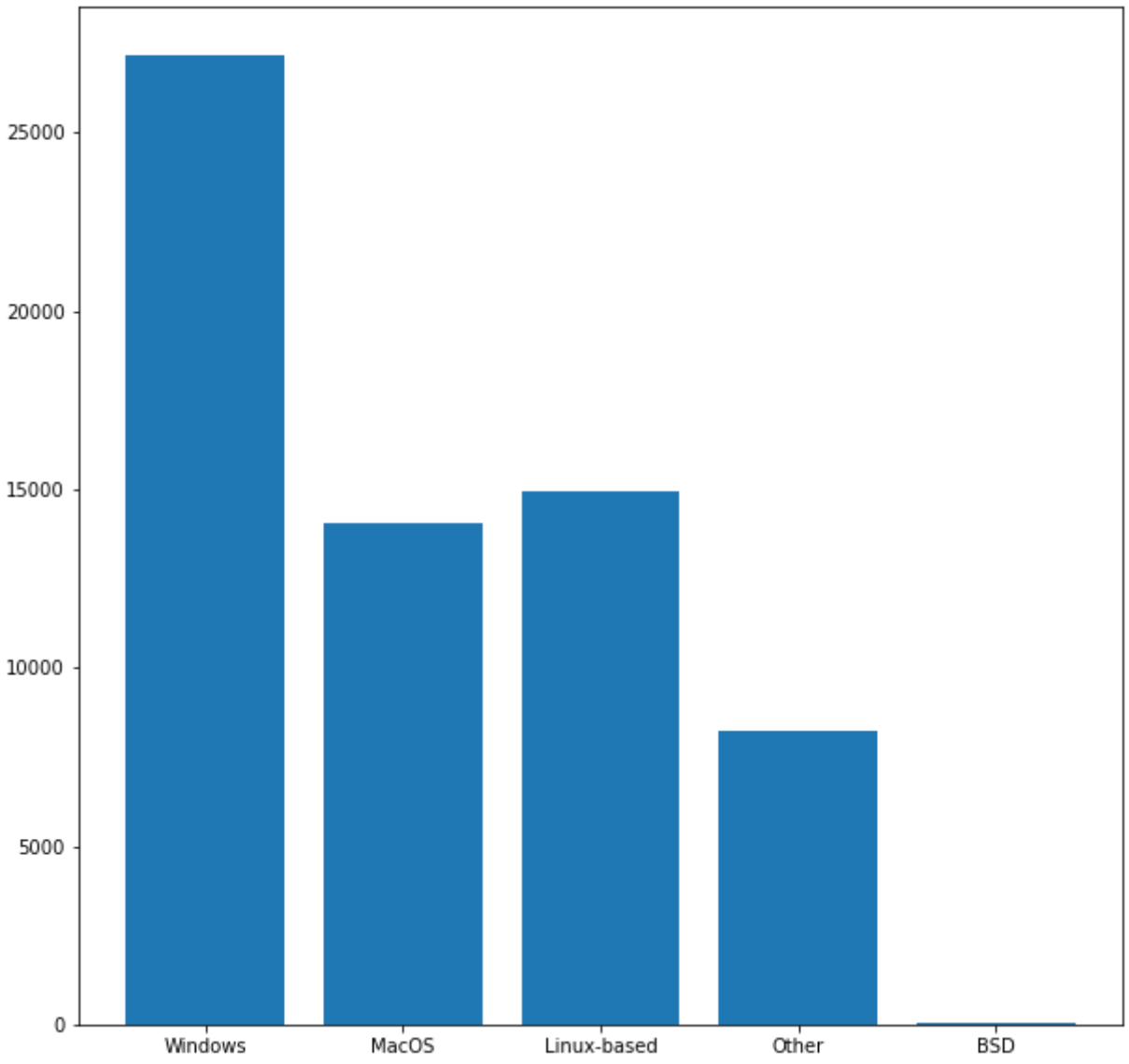
7 .


Framework web desired next years
React_js
Vue_js
Angular
ASP_NET_Core
jQuery
Django
Express
Spring
Flask
Angular_js
ASP_NET
Laravel
Gatsby
Symfony
Drupal
0
2500
5000
7500
10000
12500
15000
17500
20000

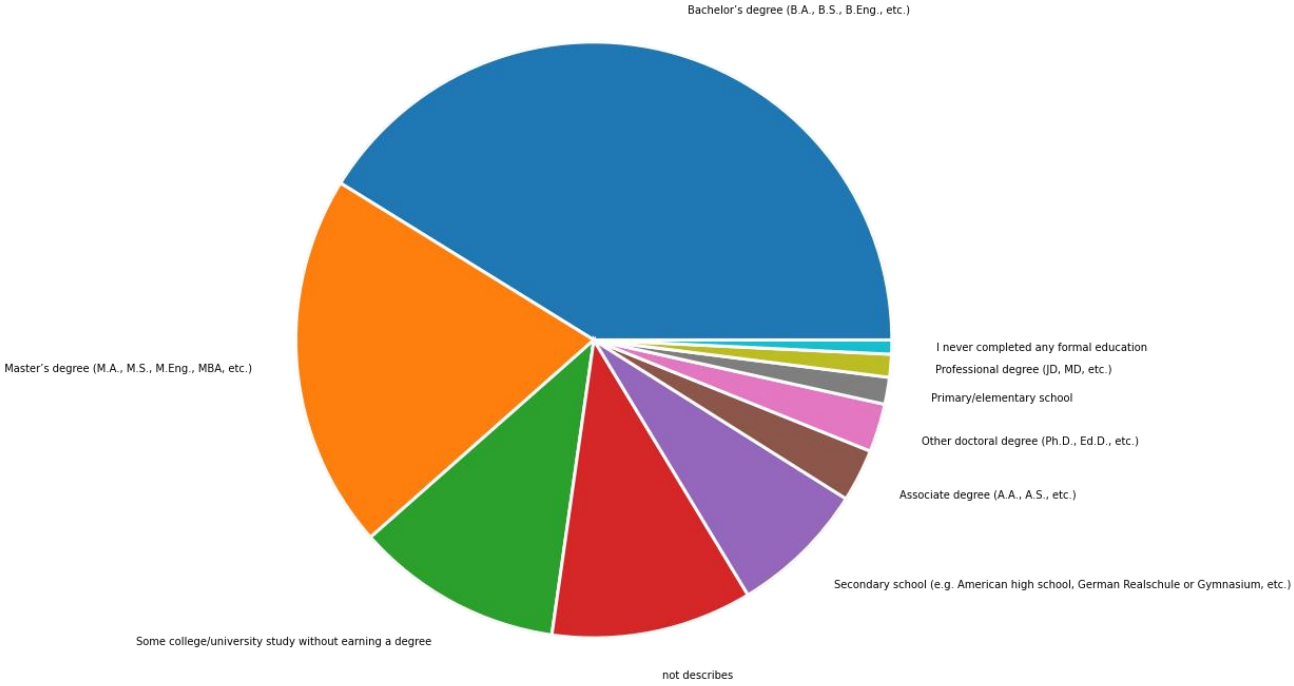
8.


once_decade_prctg
few_months_prctg
once_year_prctg
unspecified_prctg
once_few_years_prctg

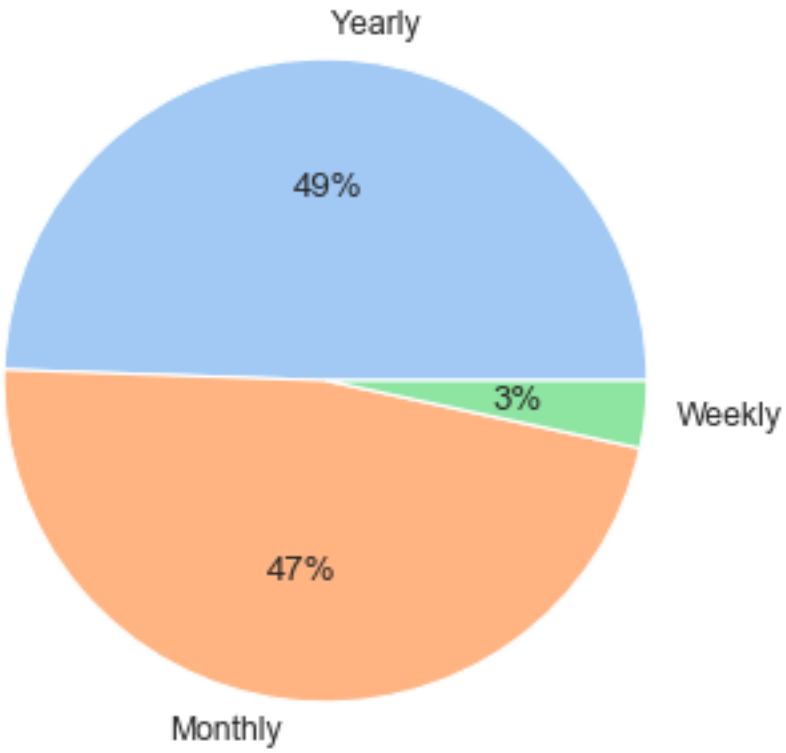
9.


25000
20000
15000
10000
5000
0
Windows
MacOS
Linux-based
Other
BSD

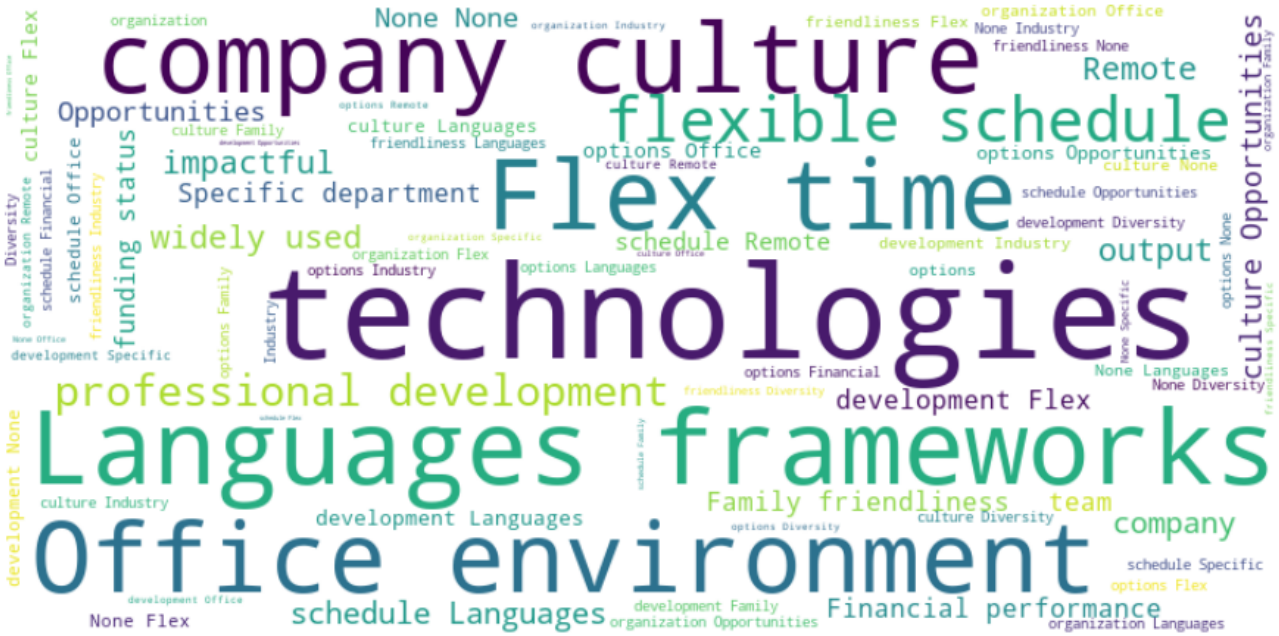
10.


Bachelor's degree (B.A., B.S., B.Eng., etc.)
Master's degree (M.A., M.S., M.Eng., MBA, etc.)
I never completed any formal education
Professional degree (JD, MD, etc.)
Primary/elementary school
Other doctoral degree (Ph.D., Ed.D., etc.)
Associate degree (A.A., A.S., etc.)
Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.)
Some college/university study without earning a degree
not describes

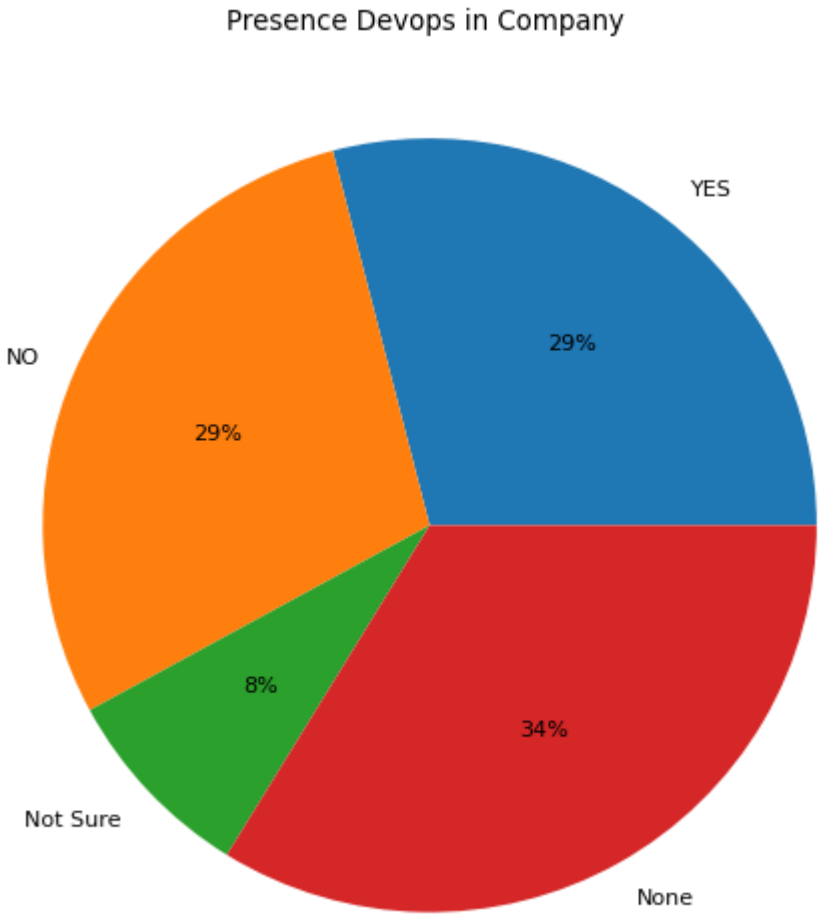
11.


Yearly
49%
3%
Weekly
47%
Monthly

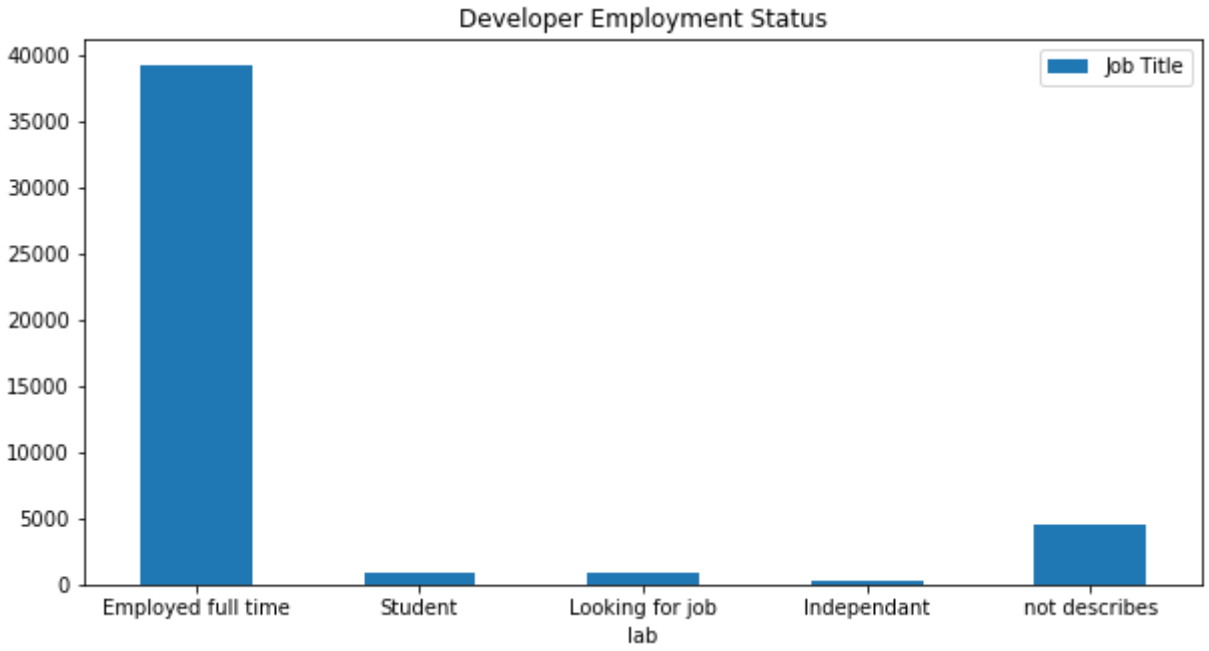
12.


company culture
flexible schedule
Flex time
technologies
Languages frameworks
Office environment
professional development
Opportunities
Remote
impactful
Specific department
widely used
schedule Remote
output
culture Opportunities
funding status
culture Flex
schedule Office
development Flex
Family friendliness
team
company
development Languages
schedule Languages
Financial performance
None Flex
None None
options Office
options Opportunities
culture Languages
friendliness Languages
culture Family
organization
organization Office
None Industry
friendliness None
friendliness Flex
schedule Opportunities
development Diversity
development Industry
options Industry
options Languages
organization Flex
options Financial
None Languages
None Diversity
culture Industry
culture Diversity
development Specific
development None
schedule Financial
Diversity
schedule Specific
options Flex
development Family
organization Opportunities
organization Languages

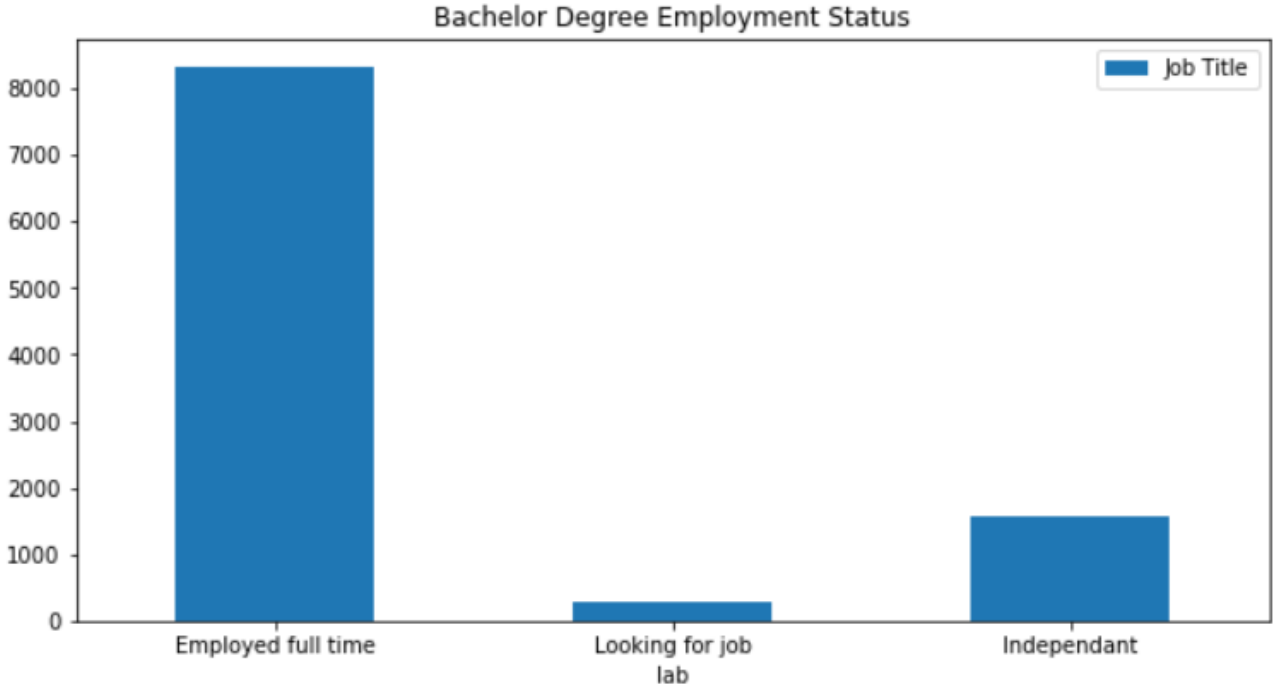
13.


Presence Devops in Company
YES
29%
NO
29%
8%
34%
Not Sure
None

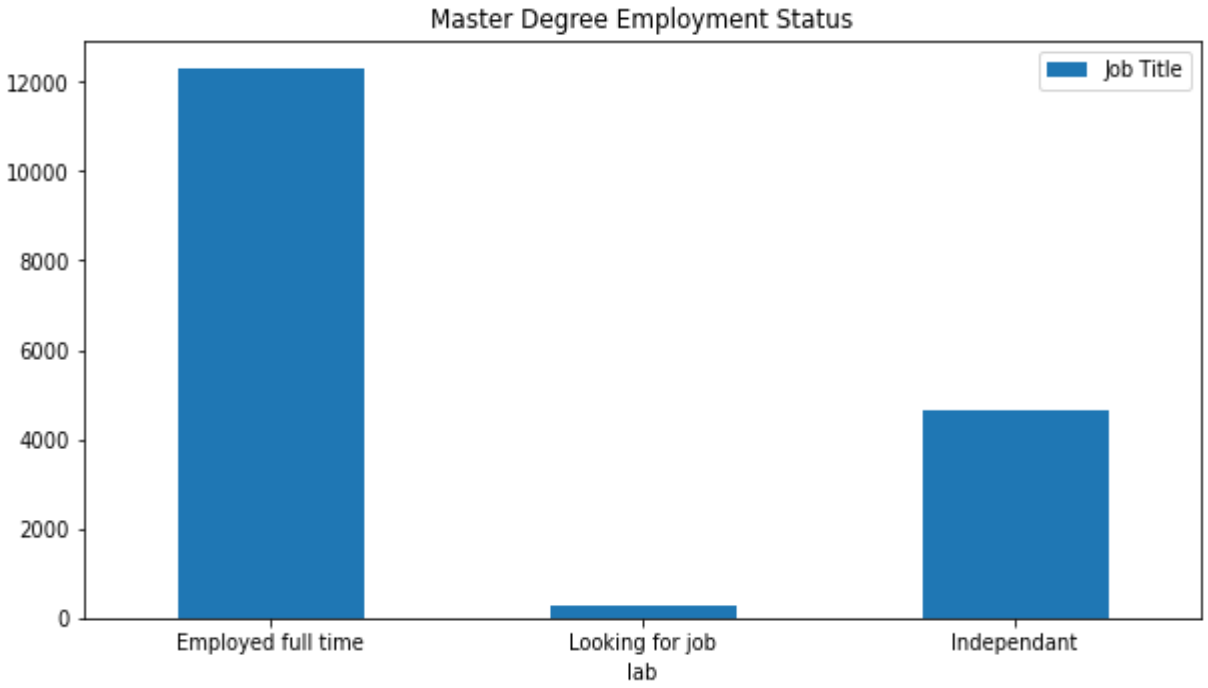
14.


Developer Employment Status
Job Title
40000
35000
30000
25000
20000
15000
10000
5000
0
Employed full time
Student
Looking for job lab
Independant
not describes

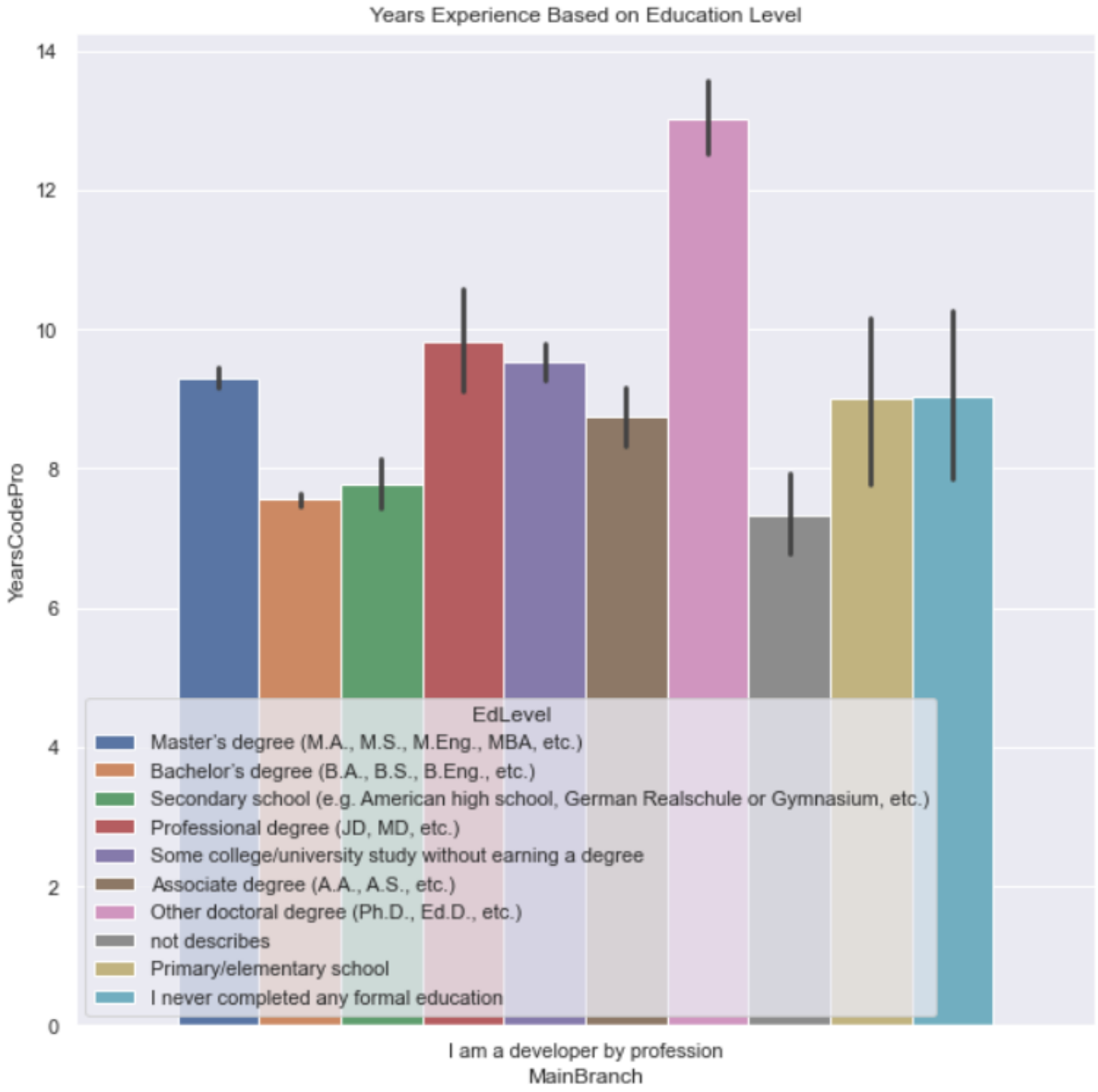
15.


Bachelor Degree Employment Status
Job Title
8000
7000
6000
5000
4000
3000
2000
1000
0
Employed full time
Looking for job lab
Independant

16.


Master Degree Employment Status
Job Title
12000
10000
8000
6000
4000
2000
0
Employed full time
Looking for job
lab
Independant

17.


Years Experience Based on Education Level
YearsCodePro
14
12
10
8
6
4
2
0
EdLevel
Master's degree (M.A., M.S., M.Eng., MBA, etc.)
Bachelor's degree (B.A., B.S., B.Eng., etc.)
Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.)
Professional degree (JD, MD, etc.)
Some college/university study without earning a degree
Associate degree (A.A., A.S., etc.)
Other doctoral degree (Ph.D., Ed.D., etc.)
not describes
Primary/elementary school
I never completed any formal education
I am a developer by profession
MainBranch