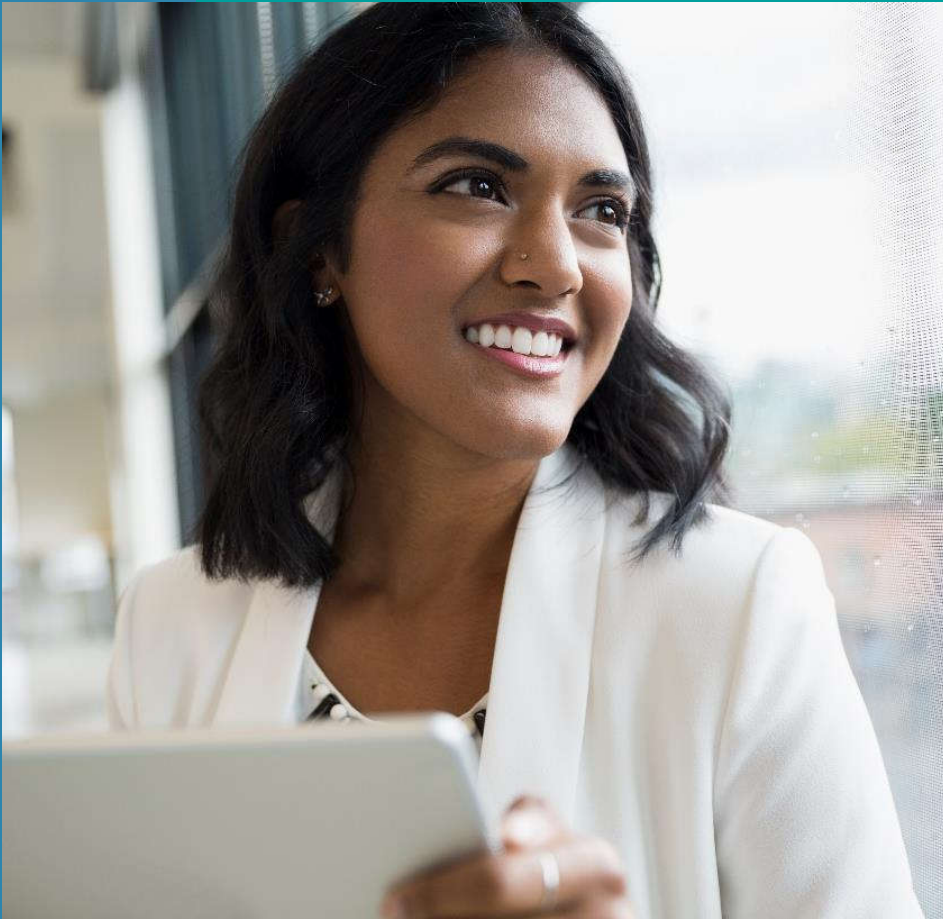




Coders Wanted 2022 Hackathon

Category 4: Open



Content

- Introduction & Objective
- Data Cleaning
 - Assumptions made on the data
- 1st level analysis
- 2nd level analysis
- Take away from the Data

Introduction & Objective

- This is a challenge by Coders Wanted 2022 Hackathon
- I will be doing Category 4: Open.
- The dataset provided seems to be information of users particularly on their occupation level as majority of the columns describe job related info.
- **Objective:** I will try to uncover valuable insights from the dataset which may provide potential impactful changes to the data owner.



Data Cleaning

Using excel conditional formatting, and some matching, able to identify 1782 duplicates based on user_id column. In addition there are also difference between some of the duplicates for columns:

- Country (e.g. user_id: 2206, all columns are similar except for country – one showing Japan, the other showing South Korea)
- gender
- avg_apply_pass
- current_inactive_days

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	
1	user_id	country	gender	years	no_of	no_of	no_of	avg_ap	no_of	signu	current	educal	job_role	skills		user_id	country	gender	years	no_of	no_of	no_of	avg_ap	avg_pi	no_of	signu	current	educal	job_role	skills					
2	19	South Kor	female		0	4	3	0	0.2	0	0	2015	844	Graduate Salary, Comper HRM	HRC	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	
3	19	South Kor	female		0	4	3	0	0.2	0	0	2015	844	Graduate Salary, Comper HRM	HRC	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
4	44	South Kor	male		0	0	0	0	0	0	0	2015	1048	Graduate Machine Learn++	CUDA	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	
5	44	South Kor	male		0	0	0	0	0	0	0	2015	1048	Graduate Machine Learn++	CUDA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
6	45	South Kor	female		0	3	0	0	0	0	0	2015	878	Graduate Lecturer	Compens	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	
7	45	South Kor	male		0	3	0	0	0	0	0	2015	878	Graduate Lecturer	Compens	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
8	68	South Kor	female		0	0	0	0	0	0	0	2015	66	Graduate Service Planne	Consultin	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	
9	68	South Kor	female		0	0	0	0	0	0	0	2015	44	Graduate Service Planne	Consultin	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
10	69	South Kor	female		0	1	0	0	0	0	0	2015	765	Graduate Marketer	Advertisi	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	
11	69	South Kor	female		0	1	0	0	0	0	0	2015	765	Graduate Marketer	Advertisi	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
12	71	South Kor	male		0	17	0	0	0	0	0	2015	84	Graduate Strategic Planr	Consultin	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	
13	71	South Kor	male		0	17	0	0	0	0	0	2015	71	Graduate Strategic Planr	Consultin	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
14	119	South Kor	male		0	0	0	0	0.4	0	0	2016	42	No Colleg Web Develop	AWS Typr	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	
15	119	South Kor	male		0	0	0	0	0.4	0	0	2016	32	No Colleg Web Develop	AWS Typr	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
16	120	South Kor	male		0	0	0	0	0	0	0	2016	1016	No Colleg Web Designer	Adobe III	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	
17	120	South Kor	male		0	0	0	0	0	0	0	2016	1016	No Colleg Web Designer	Adobe III	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
18	122	South Kor	male		0	0	0	0	0	0	0	2016	84	No Colleg QA, Test Engin	RESTful a	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	
19	122	South Kor	male		0	0	0	0	0	0	0	2016	77	No Colleg QA, Test Engin	RESTful a	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
20	125	South Kor	male		0	243	0	0	0	0	0	2016	794	Graduate Data Scientist	AMPL Tak	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	
21	125	South Kor	male		0	243	0	0	0	0	0	2016	794	Graduate Data Scientist	AMPL Tak	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
22	126	South Kor	male		0	19	0	0	0	0	0	2016	973	Graduate Video Editing	Adobe Pr	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	
23	126	South Kor	male		0	19	0	0	0	0	0	2016	973	Graduate Video Editing	Adobe Pr	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
24	128	South Kor	female		0	2	0	0	0	0	0	2016	786	Graduate Product Desig	Adobe Ph	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	
25	128	South Kor	female		0	2	0	0	0	0	0	2016	786	Graduate Product Desig	Adobe Ph	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
26	129	South Kor	female		0	0	0	0	0	0	0	2016	790	Graduate UX Designer	Axure Go	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
27	129	South Kor	female		0	0	0	0	0	0	0	2016	790	Graduate UX Designer	Axure Go	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
28	131	South Kor	male		0	11	0	0	0	0	0	2016	71	Graduate Web Develop	Kotlin	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	
29	131	South Kor	male		0	11	0	0	0	0	0	2016	34	Graduate Web Develop	Kotlin	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
30	223	South Kor	male		0	9	0	0	0	0	0	2016	884	Graduate Social Markete	Figma	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	
31	223	South Kor	male		0	9	0	0	0	0	0	2016	884	Graduate Social Markete	Figma	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
32	230	South Kor	male		0	0	0	0	0	0	0	2016	910	Graduate Sales Speciali	Compens	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	
33	230	South Kor	male		0	0	0	0	0	0	0	2016	910	Graduate Sales Speciali	Compens	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
34	237	South Kor	male		0	39	0	0	0	0	0	2016	89	Graduate Partnership Sp	Content C	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	
35	237	South Kor	male		0	39	0	0	0	0	0	2016	76	Graduate Partnership Sp	Content C	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
36	238	South Kor	male		0	2	0	0	0	0	0	2016	855	Graduate Writer	Marketin	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	
37	238	South Kor	male		0	2	0	0	0	0	0	2016	855	Graduate Writer	Marketin	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
38	242	South Kor	male		0	8	0	0	0	0	0	2016	63	No Colleg Web Develop	Go Node	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	
39	242	South Kor	male		0	8	0	0	0	0	0	2016	77	No Colleg Web Develop	Go Node	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
40	243	South Kor	male		0	0	0	0	0	0	0	2016	32	Graduate Python Develop	Vue S Ty	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	
41	243	South Kor	male		0	0	0	0	0	0	0	2016	13	Graduate Python Develop	Vue S Ty	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
42	245	South Kor	female		0	0	0	0	0	0	0	2016	1005	Graduate Service Planne	Kotlin No	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	
43	245	South Kor	female		0	0	0	0	0	0	0	2016	1005	Graduate Service Planne	Kotlin No	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	

There are also 15 rows with null values in the dataset

```
1 # we have 15rows of null values
2 df.isnull().sum().sum()
[289] ✓ 0.6s
... 15
```

Data Cleaning

Given there are difference with some of the duplicate rows, I do not know which rows contain the accurate information. So I will be dropping all the duplicates and the null rows to be more accurate in the analysis.

As there is no metadata provided on the data set, assumptions will be made before making the analysis.

```
1 new_df = df.drop_duplicates(subset=['user_id'], keep=False).dropna()
2 new_df = new_df.reset_index(drop=True)
3 print(new_df.info())
```

[284] ✓ 0.4s

... <class 'pandas.core.frame.DataFrame'>
RangeIndex: 18170 entries, 0 to 18169
Data columns (total 15 columns):

#	Column	Non-Null Count	Dtype
0	user_id	18170 non-null	int64
1	country	18170 non-null	object
2	gender	18170 non-null	object
3	years_of_experience	18170 non-null	int64
4	no_of_apply	18170 non-null	int64
5	no_of_pass	18170 non-null	int64
6	no_of_hire	18170 non-null	int64
7	avg_apply_pass	18170 non-null	float64
8	avg_pass_hire	18170 non-null	float64
9	no_of_event_reg	18170 non-null	int64
10	signup_year	18170 non-null	int64
11	current_inactive_days	18170 non-null	float64
12	education_level	18170 non-null	object
13	job_role	18170 non-null	object
14	skills	18170 non-null	object

dtypes: float64(3), int64(7), object(5)

Assumptions

- years_of_experience: the number of years of experience the users have
- no_of_apply: the number of job applications the users have made
- no_of_hire: times of being hired from their job applications
- signup_year: number of new users sign up for that year
- current_inactive_days: number of days users has been inactive for
- job_role: job that the user is currently holding
- skills: skills that the user currently have

```
1 new_df = df.drop_duplicates(subset=['user_id'], keep=False).dropna()
2 new_df = new_df.reset_index(drop=True)
3 print(new_df.info())
```

[284] ✓ 0.4s

... <class 'pandas.core.frame.DataFrame'>
RangeIndex: 18170 entries, 0 to 18169
Data columns (total 15 columns):

#	Column	Non-Null Count	Dtype
0	user_id	18170 non-null	int64
1	country	18170 non-null	object
2	gender	18170 non-null	object
3	years_of_experience	18170 non-null	int64
4	no_of_apply	18170 non-null	int64
5	no_of_pass	18170 non-null	int64
6	no_of_hire	18170 non-null	int64
7	avg_apply_pass	18170 non-null	float64
8	avg_pass_hire	18170 non-null	float64
9	no_of_event_reg	18170 non-null	int64
10	signup_year	18170 non-null	int64
11	current_inactive_days	18170 non-null	float64
12	education_level	18170 non-null	object
13	job_role	18170 non-null	object
14	skills	18170 non-null	object

dtypes: float64(3), int64(7), object(5)

1st level analysis

A simple surface analysis tell us likely the data-owner is a recruitment firm, with strong presence in South Korea.

The company has been seeing a stable increase in sign up rates on their platform since 2015. (year 2022 not ended so not to take into account).

Average years of work experience of the users using the platform is at 2.6 years. Each users averagely applied 30 jobs, however the average hired rate is less than 1.

top 10 countries	
South Korea	93.1%
United States	3.2%
Japan	1.7%
United Kingdom	0.4%
Vietnam	0.3%
Hong Kong	0.3%
India	0.1%
Taiwan	0.1%
China	0.1%
Thailand	0.1%

signup rate each year	
2015	612
2016	1328
2017	1985
2018	2138
2019	3407
2020	2820
2021	4957
2022	923

years_of_experience	
count	18170.000000
mean	2.621354
std	3.553147
min	0.000000
25%	0.000000
50%	1.000000
75%	4.000000
max	131.000000

no_of_apply	
count	18170.000000
mean	30.265768
std	68.225869
min	0.000000
25%	4.000000
50%	13.000000
75%	32.000000
max	3133.000000

no_of_hire	
count	18170.000000
mean	0.689653
std	0.489583
min	0.000000
25%	0.000000
50%	1.000000
75%	1.000000
max	6.000000

1st level analysis

The top few user's job roles are mainly in the IT industry – shows that this recruitment company mainly focuses on IT sector job placements.

Lastly the average user's inactive period is around 116 days (~ 4months).

56% of the users is only inactive for less than a month, followed by 29% for between a month to half a year, 7.8% more than half a year to a year and lastly 7.1% has been inactive for more than a year.

top 10 job roles	
Web Developer	12.6%
Web Designer	5.1%
Strategic Planner	4.2%
Front-end Engineer	4.0%
Service Planner	3.8%
Social Marketer	3.3%
UX Designer	2.8%
iOS Developer	2.5%
Java Developer	2.0%
Web Publisher	1.7%

inactive_days	
count	18170.000000
mean	116.866924
std	251.732912
min	0.000000
25%	4.000000
50%	22.000000
75%	85.000000
max	1357.000000

<= a month	56.0%
a month to half a year	29.1%
half a year to a year	7.8%
more than a year	7.1%

2nd level analysis

Next, I will take a deeper dive into the data.

I broken down the average application rate & hire rate of the users from the top 10 countries.

Even though users' were mainly from South Korea, UK has the highest average application rate of ~70 per user, however its unfortunately that none of the country has even average of 1 for their average hired rate.

	no_of_apply	no_of_hire
country		
United Kingdom	70.9	0.8
Taiwan	69.8	0.8
United States	47.2	0.9
China	41.2	0.9
Vietnam	31.3	0.7
South Korea	29.9	0.7
Hong Kong	29.2	0.9
Thailand	23.2	0.8
Japan	7.9	0.1
India	7.5	0.0

Next its encouraging to see that the ratio of female to male in the top few tech roles is roughly balanced out as its often reported that females are being underrepresented in this industry.

gender	female	male
job_role		
Web Designer	52.4%	47.6%
Java Developer	51.7%	48.3%
Front-end Engineer	50.3%	49.7%
Strategic Planner	50.0%	50.0%
Web Developer	49.6%	50.4%
iOS Developer	49.3%	50.7%
UX Designer	49.2%	50.8%
Service Planner	47.5%	52.5%
Web Publisher	47.2%	52.8%
Social Marketer	47.1%	52.9%

2nd level analysis

Lastly, I split up the user's skills set of the top 10 roles to see what is the 10 most common skill that users in those roles have. It is represented by

- ('skill', 'number of users who have this skill')

Immediately the top most common skill that majority of the users had are AWS (amazon web service), goes to show that AWS is a valuable skill to have to be in these roles.

```
WebDeveloper_topskills
[('AWS', 1967), ('Git', 757), ('Java', 672), ('MySQL', 616), ('Python', 580), ('JavaScript', 558), ('Docker', 520), ('React', 481), ('TypeScript', 457), ('GitHub', 419)]
WebDesigner_topskills
[('Adobe Illustrator', 733), ('Adobe Photoshop', 704), ('UI Design', 237), ('Graphic Design', 207), ('Figma', 170), ('Sketch', 164), ('3D', 160), ('Branding', 150), ('Web Design', 141), ('Adobe XD', 117)]
StrategicPlanner_topskills
[('AWS', 124), ('Accounting', 112), ('JIRA', 104), ('SQL', 103), ('Excel', 95), ('Google Analytics', 92), ('Tableau', 65), ('Python', 58), ('Axure', 54), ('IFRS', 52)]
Frontend_Engineer_topskills
[('AWS', 611), ('Git', 253), ('React', 203), ('JavaScript', 186), ('TypeScript', 185), ('Docker', 181), ('Python', 178), ('MySQL', 167), ('Java', 152), ('APIs', 129)]
ServicePlanner_topskills
[('AWS', 133), ('Accounting', 85), ('JIRA', 77), ('Excel', 77), ('Adobe Photoshop', 73), ('SQL', 70), ('MySQL', 69), ('Google Analytics', 65), ('Java', 64), ('JavaScript', 63)]
SocialMarketer_topskills
[('Tableau', 165), ('AMPL', 161), ('Adobe Photoshop', 103), ('Google Analytics', 100), ('Adobe Illustrator', 64), ('Marketing Strategy', 52), ('Sketch', 52), ('Figma', 48), ('Marketing Operations', 43), ('Content Creation', 43)]
UXDesigner_topskills
[('Adobe Photoshop', 389), ('Adobe Illustrator', 389), ('UI Design', 166), ('Figma', 129), ('Graphic Design', 125), ('Sketch', 103), ('Branding', 94), ('3D', 89), ('Zeplin', 87), ('Service Design', 85)]
iOSDeveloper_topskills
[('AWS', 405), ('Python', 170), ('Java', 151), ('React', 148), ('iOS', 143), ('Swift', 133), ('Git', 126), ('TypeScript', 111), ('JavaScript', 109), ('MySQL', 108)]
JavaDeveloper_topskills
[('AWS', 294), ('Java', 133), ('Python', 124), ('Git', 118), ('MySQL', 114), ('iOS', 90), ('React', 90), ('JavaScript', 86), ('Docker', 85), ('Spring Framework', 85)]
WebPublisher_topskills
[('AWS', 225), ('Git', 98), ('JavaScript', 94), ('MySQL', 76), ('Java', 75), ('Python', 63), ('React', 63), ('Docker', 52), ('Node.js', 51), ('TypeScript', 51)]
```

Take away from the Data

From a business perspective, its positive sign to see the steady increase in sign up rate since 2015.

The average users using the platform are early in their careers (as shown from their year of experience), also signs that the average users base are quite young. Something the company can consider doing is organizing a “platform member only” career talks from veteran in the industry. This aim to reduce the inactive period of the users and attract more young talents onto the platform.

We could make use of the data on user's skill to assist users themselves with job search. The company can consider building a skill match algorithm based on the jobs the user are interested in and the user's skill to match them together. Or if the skills does not match, the algorithm will propose what skills the applicant is lacking so the applicant is aware and can work towards that area. This aims to improve user satisfactory level by increasing their chance of being hired through the platform.