Deliverable 2

Data Understanding

Github link: https://github.com/Vishnu-Vardhan1809/Big-Data-Project.git

a)Exploratory data analysis

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df=pd.read_csv('survey.csv')
df
```

```
C:\Users\bgompa\AppData\Local\Temp\1\ipykernel_10088\450253707.py:1: DtypeWarning: Columns (0,195,201,285,286,287,288,289,290,291,292) have mixed types. Specify dtype option on import or set low_memory=False.
  df=pd.read_csv('survey.csv')
```

	Time from Start to Finish (seconds)	Q1	Q2	Q3	Q4	Q5	Q6	Q7_Part_1	Q7_Part_2	Q7_Part_3	...	Q38_B_Part_3	Q38_B_Part_4	Q38_
0	Duration (in seconds)	What is your age (# years)?	What is your gender? - Selected Choice	In which country do you currently reside?	What is the highest level of formal education ...	Select the title most similar to your current ...	For how many years have you been writing code ...	What programming languages do you use on a reg...	What programming languages do you use on a reg...	What programming languages do you use on a reg...	...	In the next 2 years, do you hope to become mor...	In the next 2 years, do you hope to become mor...	In yea bec
1	910	50-54	Man	India	Bachelor's degree	Other	5-10 years	Python	R	NaN	...	NaN	NaN	
2	784	50-54	Man	Indonesia	Master's degree	Program/Project Manager	20+ years	NaN	NaN	SQL	...	NaN	NaN	

```
In [3]: df.shape
Out[3]: (25974, 369)

In [4]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25974 entries, 0 to 25973
Columns: 369 entries, Time from Start to Finish (seconds) to Q38_B_OTHER
dtypes: object(369)
memory usage: 73.1+ MB

In [6]: df['Time from Start to Finish (seconds)'].dtype
Out[6]: dtype('O')

In [7]: df.columns
Out[7]: Index(['Time from Start to Finish (seconds)', 'Q1', 'Q2', 'Q3', 'Q4', 'Q5',
               'Q6', 'Q7_Part_1', 'Q7_Part_2', 'Q7_Part_3',
               ...
               'Q38_B_Part_3', 'Q38_B_Part_4', 'Q38_B_Part_5', 'Q38_B_Part_6',
               'Q38_B_Part_7', 'Q38_B_Part_8', 'Q38_B_Part_9', 'Q38_B_Part_10',
               'Q38_B_Part_11', 'Q38_B_OTHER'],
              dtype='object', length=369)
```

1.visulising the degrees and their job roles based on the survey results

In [10]:
```
set(df['Q4'])
```

Out[10]:
```
{'Bachelor's degree',
 'Doctoral degree',
 'I prefer not to answer',
 'Master's degree',
 'No formal education past high school',
 'Professional doctorate',
 'Some college/university study without earning a bachelor's degree',
 'What is the highest level of formal education that you have attained or plan to attain within the next 2 years?'}
```

In [11]:
```
set(df['Q5'])
```

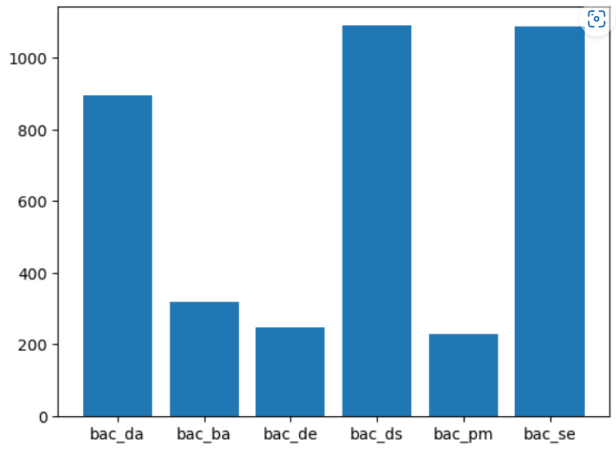
Out[11]:
```
{'Business Analyst',
 'Currently not employed',
 'DBA/Database Engineer',
 'Data Analyst',
 'Data Engineer',
 'Data Scientist',
 'Developer Relations/Advocacy',
 'Machine Learning Engineer',
 'Other',
 'Product Manager',
 'Program/Project Manager',
 'Research Scientist',
 'Select the title most similar to your current role (or most recent title if retired): - Selected Choice',
 'Software Engineer',
 'Statistician',
 'Student'}
```

In [24]:
```
bac_da=len(df[(df['Q4']=='Bachelor's degree') & (df['Q5']=='Data Analyst')])
bac_ba=len(df[(df['Q4']=='Bachelor's degree') & (df['Q5']=='Business Analyst')])
bac_de=len(df[(df['Q4']=='Bachelor's degree') & (df['Q5']=='Data Engineer')])
bac_ds=len(df[(df['Q4']=='Bachelor's degree') & (df['Q5']=='Data Scientist')])
bac_pm=len(df[(df['Q4']=='Bachelor's degree') & (df['Q5']=='Program/Project Manager')])
bac_se=len(df[(df['Q4']=='Bachelor's degree') & (df['Q5']=='Software Engineer')])

mas_da=len(df[(df['Q4']=='Master's degree') & (df['Q5']=='Data Analyst')])
mas_ba=len(df[(df['Q4']=='Master's degree') & (df['Q5']=='Business Analyst')])
mas_de=len(df[(df['Q4']=='Master's degree') & (df['Q5']=='Data Engineer')])
mas_ds=len(df[(df['Q4']=='Master's degree') & (df['Q5']=='Data Scientist')])
mas_pm=len(df[(df['Q4']=='Master's degree') & (df['Q5']=='Program/Project Manager')])
mas_se=len(df[(df['Q4']=='Master's degree') & (df['Q5']=='Software Engineer')])


doc_da=len(df[(df['Q4']=='Doctoral degree') & (df['Q5']=='Data Analyst')])
doc_ba=len(df[(df['Q4']=='Doctoral degree') & (df['Q5']=='Business Analyst')])
doc_de=len(df[(df['Q4']=='Doctoral degree') & (df['Q5']=='Data Engineer')])
doc_ds=len(df[(df['Q4']=='Doctoral degree') & (df['Q5']=='Data Scientist')])
doc_pm=len(df[(df['Q4']=='Doctoral degree') & (df['Q5']=='Program/Project Manager')])
doc_se=len(df[(df['Q4']=='Doctoral degree') & (df['Q5']=='Software Engineer')])
```

In [31]:
```
b=[bac_da,bac_ba,bac_de,bac_ds,bac_pm,bac_se]
m=[mas_da,mas_ba,mas_de,mas_ds,mas_pm,mas_se]
d=[doc_da,doc_ba,doc_de,doc_ds,doc_pm,doc_se]
b1=['bac_da','bac_ba','bac_de','bac_ds','bac_pm','bac_se']
m1=['mas_da','mas_ba','mas_de','mas_ds','mas_pm','mas_se']
d1=['doc_da','doc_ba','doc_de','doc_ds','doc_pm','doc_se']
```

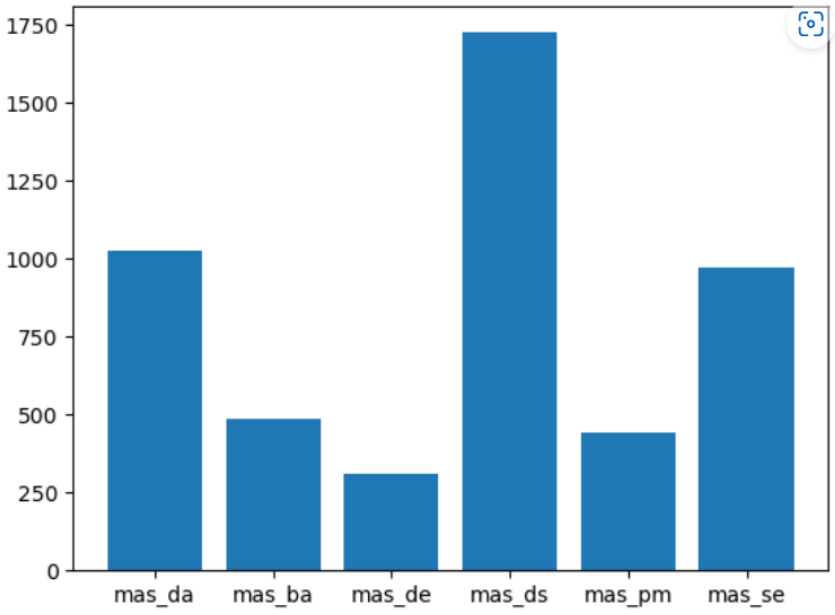
In [32]:
```
plt.bar(b1,b)
```

Out[32]: <BarContainer object of 6 artists>



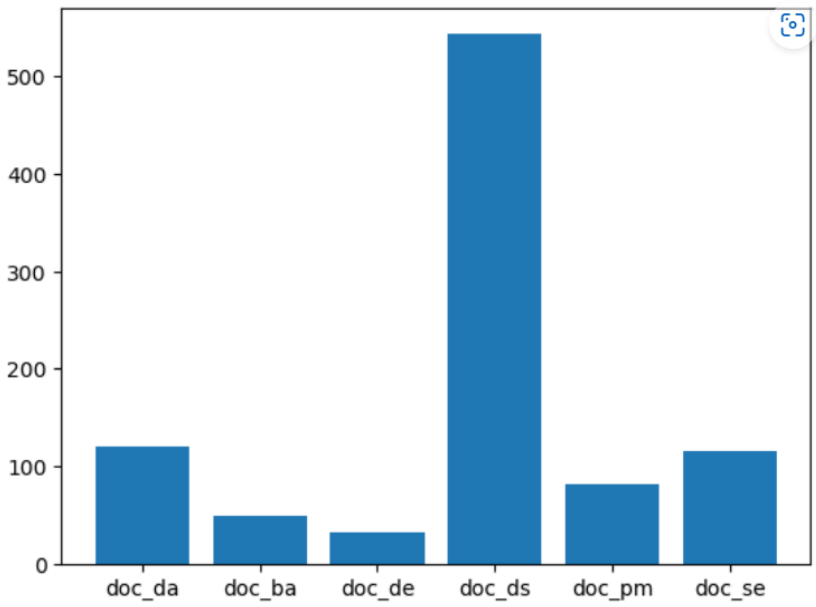
In [33]:
```
plt.bar(m1,m)
```

Out[33]: <BarContainer object of 6 artists>



In [34]:
```
plt.bar(d1,d)
```

Out[34]:
```
<BarContainer object of 6 artists>
```



Country wise count of people involved in data science job roles

In [43]:
```
j=list(set(df['Q3']))
```

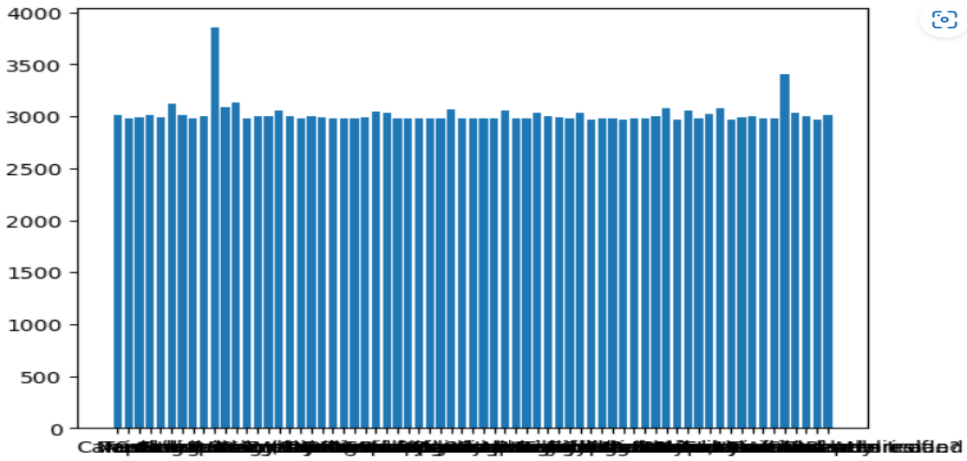
In [45]:
```
country=[]
count1=[]
for i in j:
    ds_role=len(df[(df['Q5']=='Data Engineer') | (df['Q5']=='Data Analyst') | (df['Q5']=='Data Scientist') & (df['Q3']==i)])
    country.append(i)
    count1.append(ds_role)
```

In [49]:
```
c=list(zip(country,count1))
c
```

Out[49]:
```
[('Canada', 3014),
 ('Nepal', 2977),
 ('Taiwan', 2994),
 ('Australia', 3017),
 ('Chile', 2987),
 ('Brazil', 3124),
 ('South Korea', 3011),
 ('Singapore', 2985),
 ('Ukraine', 3002),
 ('India', 3852),
 ('Russia', 3088),
 ('Other', 3128),
 ('Hong Kong (S.A.R.)', 2978),
 ('Kenya', 3001),
 ('Colombia', 3006),
 ('Turkey', 3054),
```

In [50]:
```
plt.bar(country,count1)
```

Out[50]:
```
<BarContainer object of 67 artists>
```



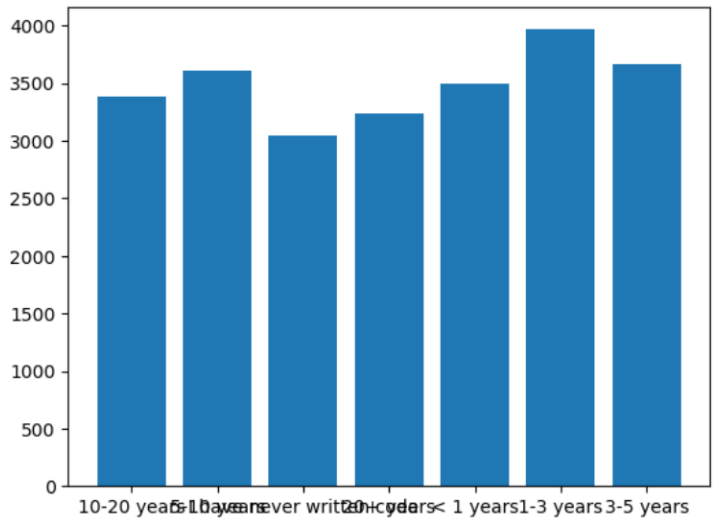
In [62]:
```
g=list(set(df['Q6']))
g.pop()
g
```

Out[62]:
```
['10-20 years',
 '5-10 years',
 'I have never written code',
 '20+ years',
 '< 1 years',
 '1-3 years',
 '3-5 years']
```

In [63]:
```
exp1=[]
count2=[]
for i in g:
    exp=len(df[(df['Q5']=='Data Engineer') | (df['Q5']=='Data Analyst') | (df['Q5']=='Data Scientist') & (df['Q6']==i)])
    exp1.append(i)
    count2.append(exp)
```

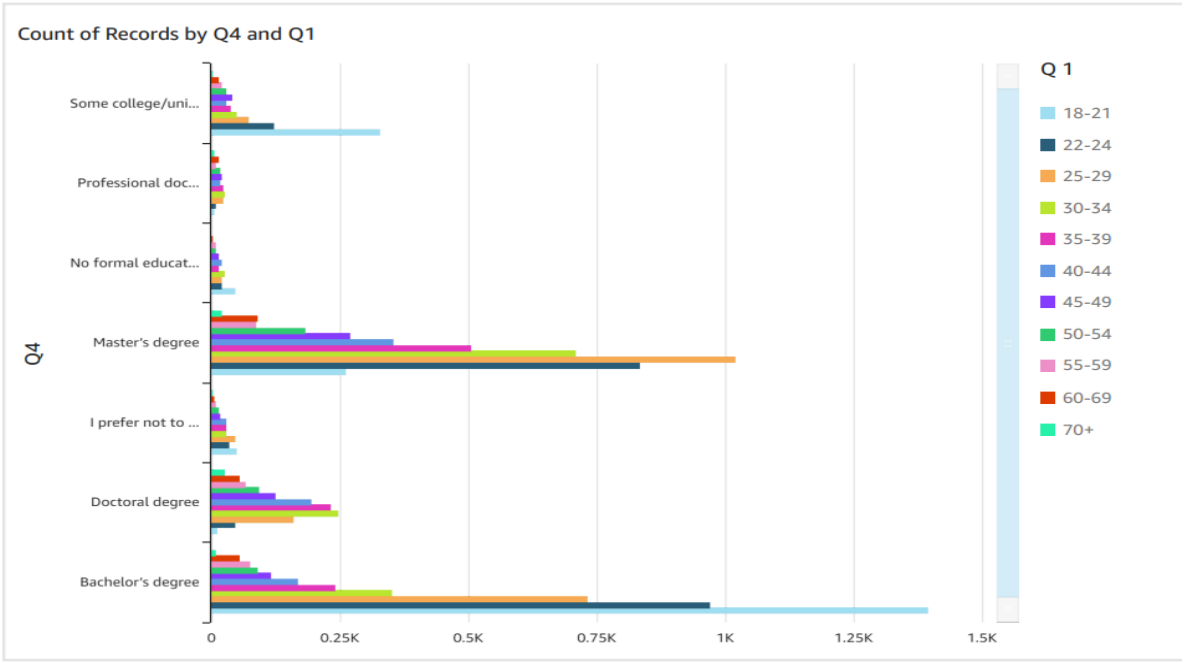
In [64]:
```
plt.bar(exp1,count2)
```

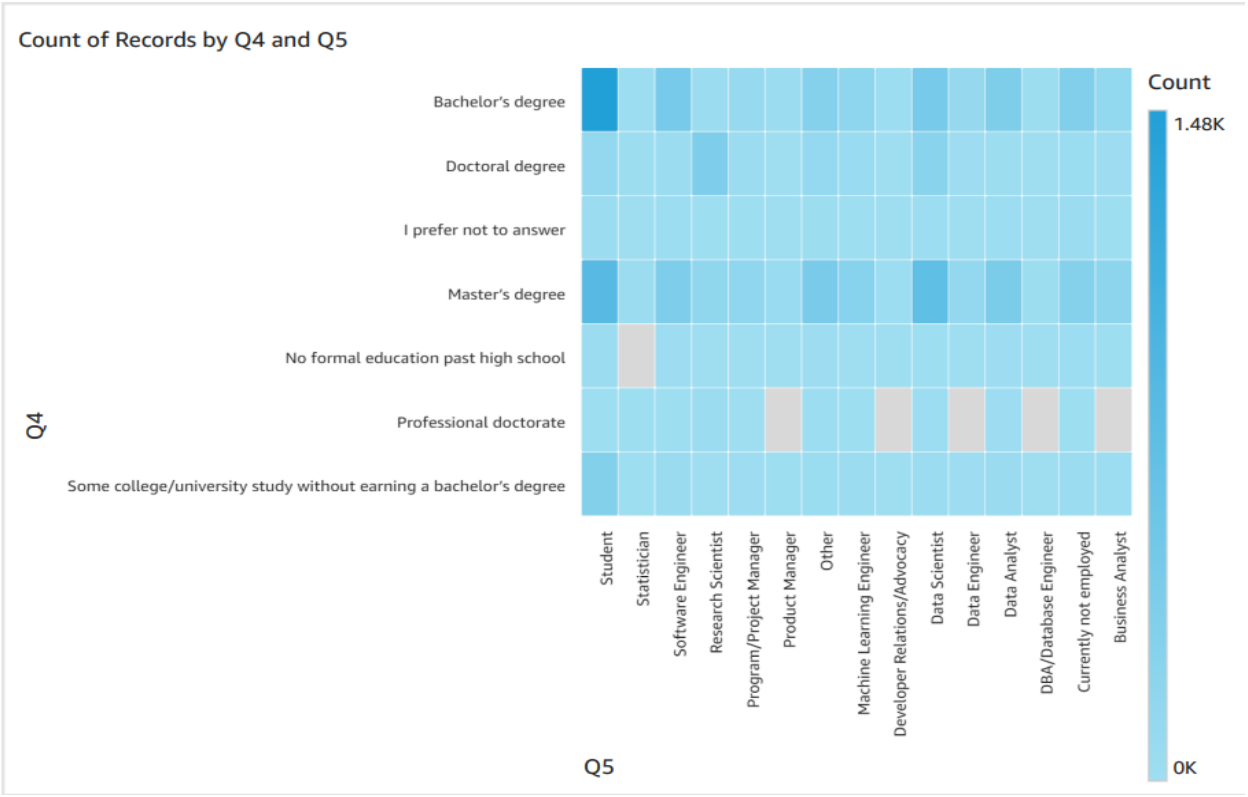
Out[64]:
```
<BarContainer object of 7 artists>
```



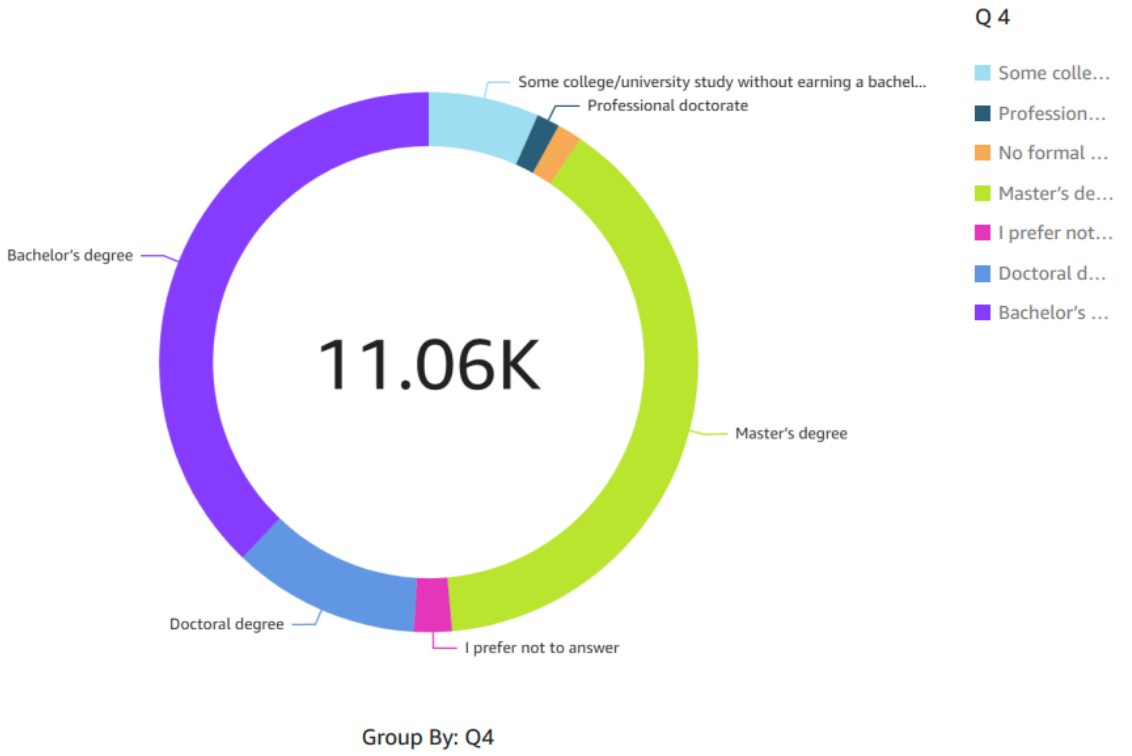
b) Dashboard

4/24/23, 10:39 PM



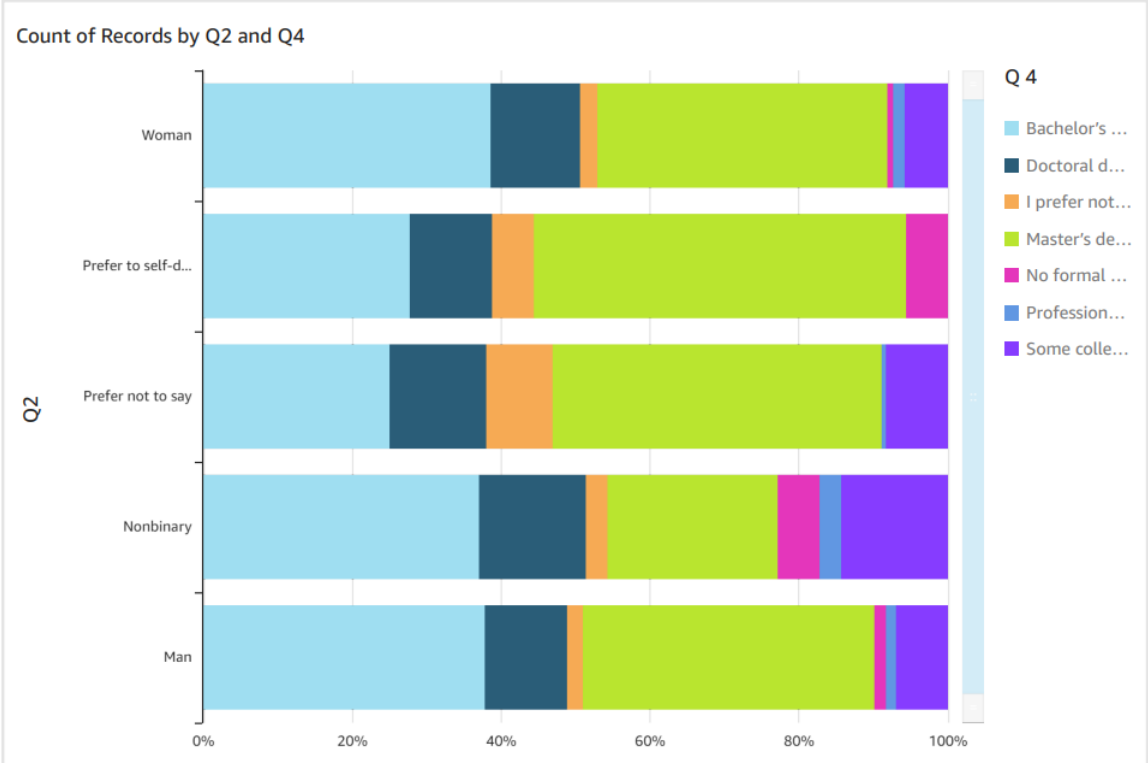


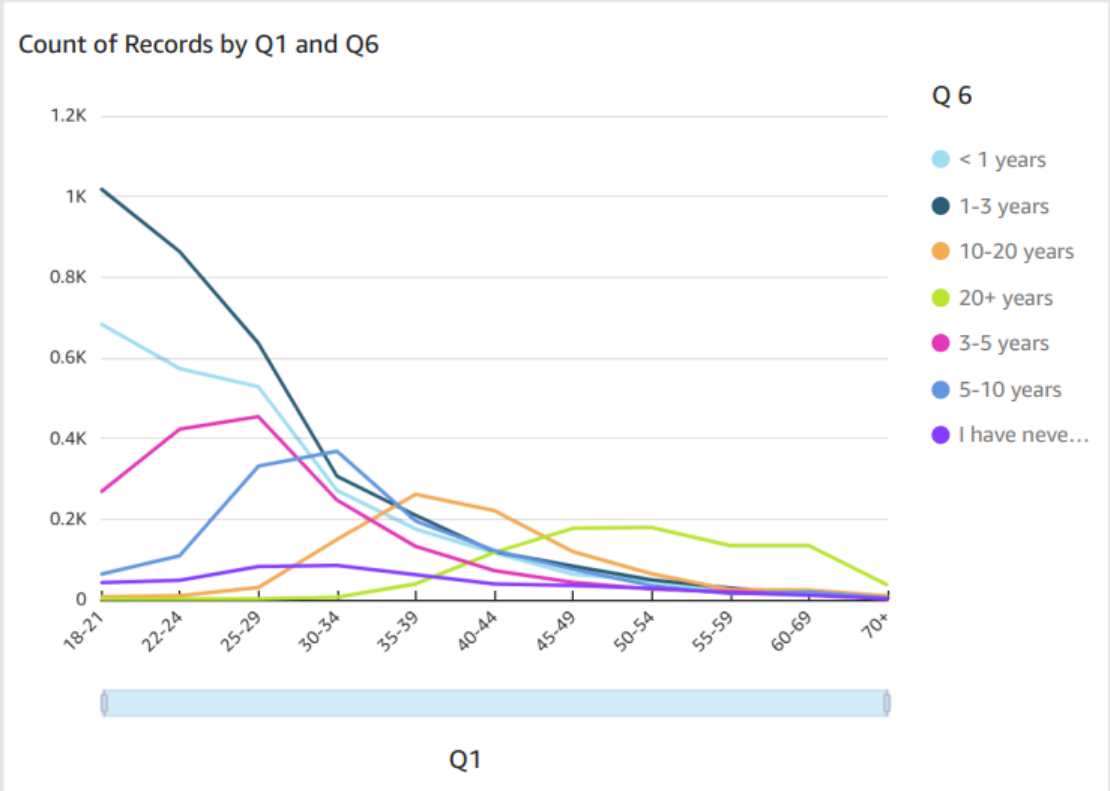
Count of Records by Q4



Group By: Q4

Count of Records by Q1 and Q5


Count of Records by Q2 and Q4
Q 4
Bachelor's ...
Doctoral d...
I prefer not...
Master's de...
No formal ...
Profession...
Some colle...
Q2
Woman
Prefer to self-d...
Prefer not to say
Nonbinary
Man
0%
20%
40%
60%
80%
100%


Count of Records by Q1 and Q6
Q 6
< 1 years
1-3 years
10-20 years
20+ years
3-5 years
5-10 years
I have neve...
1.2K
1K
0.8K
0.6K
0.4K
0.2K
0
18-21
22-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-69
70+
Q1

Data Preparation

Claening the Data

In [106]:
```
df.head(25)
```

18	18	479	18-21	Man	Pakistan	degree	Data Scientist	years	Python	NaN	...	NaN	NaN
19	19	249	22-24	Man	Japan	Master's degree	Software Engineer	3-5 years	Python	NaN	...	NaN	NaN
20	20	650	30-34	Man	Egypt	Bachelor's degree	Other	< 1 years	NaN	NaN	...	NaN	NaN
21	21	1461	70+	Man	Singapore	Bachelor's degree	Other	< 1 years	Python	NaN	...	NaN	NaN
22	22	551	25-29	Woman	Turkey	Bachelor's degree	Data Scientist	3-5 years	Python	R	...	NaN	NaN
23	23	258	30-34	Man	Indonesia	Master's degree	Student	1-3 years	NaN	R	...	NaN	NaN
24	24	773	35-39	Man	Brazil	Master's degree	Machine Learning Engineer	20+ years	Python	NaN	...	NaN	NaN

25 rows × 370 columns

In [112]:
```
p=list(df.apply(lambda x: x.count(), axis=1))
len(p)
```

Out[112]: 25974

In [113]:
```
g=[]
for i in range(len(p)):
    if p[i]>45:
        g.append(i+1)
g
```

```
47,
48,
49,
50,
51,
52,
53,
54,
55,
59,
60,
64,
66,
67,
69,
70,
71,
73,
74,
75,
```

In [114]:
```
df.drop(g,inplace=True)
df
```

Out[114]:

	index	Time from Start to Finish (seconds)	Q1	Q2	Q3	Q4	Q5	Q6	Q7_Part_1	Q7_Part_2	...	Q38_B_Part_3	Q38_B_Part_4	Q38_B_Part_5	Q38_I
0	0	Duration (in seconds)	What is your age (# years)?	What is your gender? - Selected Choice	In which country do you currently reside?	What is the highest level of formal education ...	Select the title most similar to your current ...	For how many years have you been writing code ...	What programming languages do you use on a reg...	What programming languages do you use on a reg...	...	In the next 2 years, do you hope to become mor...	In the next 2 years, do you hope to become mor...	In the next 2 years, do you hope to become mor...	In t year beco
8	8	484	30-34	Man	India	Bachelor's degree	Data Scientist	5-10 years	Python	NaN	...	NaN	NaN	NaN	
10	10	655	30-34	Man	Turkey	I prefer not to answer	Other	1-3 years	Python	NaN	...	NaN	NaN	NaN	
11	11	1777	40-44	Man	Australia	Doctoral degree	Other	1-3 years	Python	R	...	NaN	NaN	NaN	
12	12	3081	18-21	Woman	India	Master's degree	Student	< 1 years	Python	R	...	NaN	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
i956	25956	397	35-39	Man	India	Bachelor's degree	Product Manager	3-5 years	Python	NaN	...	NaN	NaN	NaN	
						Bachelor's	Software	3-5							

In [125]:
```
j=list(df.isna().sum())
j
```

```
10723,
10628,
9338,
10903,
8447,
9670,
9775,
10085,
9217,
10592,
10557,
9714,
9901,
10400,
10551,
10573,
10420,
10075,
9843,
9697,
```

In [127]:
```
df.drop(df.columns[u], axis=1,inplace=True)
```

In [128]:
```
df
```

Out[128]:

	index	Time from Start to Finish (seconds)	Q1	Q2	Q3	Q4	Q5	Q6	Q7_Part_1	Q7_Part_2	...	Q36_B_Part_3	Q36_B_Part_5	Q36_B_Part_6	Q
0	0	Duration (in seconds)	What is your age (# years)?	What is your gender? - Selected Choice	In which country do you currently reside?	What is the highest level of formal education ...	Select the title most similar to your current ...	For how many years have you been writing code ...	What programming languages do you use on a reg...	What programming languages do you use on a reg...	...	Which categories of automated machine learning...	Which categories of automated machine learning...	Which categories of automated machine learning...	
8	8	484	30-34	Man	India	Bachelor's degree	Data Scientist	5-10 years	Python	NaN	...	NaN	NaN	NaN	
10	10	655	30-34	Man	Turkey	I prefer not to answer	Other	1-3 years	Python	NaN	...	Automated model selection (e.g. auto-sklearn, ...	NaN	NaN	
11	11	1777	40-44	Man	Australia	Doctoral degree	Other	1-3 years	Python	R	...	NaN	NaN	NaN	
12	12	3081	18-21	Woman	India	Master's degree	Student	< 1 years	Python	R	...	NaN	NaN	Automation of full ML pipelines (e.g. Google C...	

In [140]:
```
k=list(df.columns)
k.pop(0)
k
```

```
'Q7_Part_6',
'Q7_Part_7',
'Q7_Part_11',
'Q7_OTHER',
'Q8',
'Q9_Part_1',
'Q9_Part_2',
'Q9_Part_3',
'Q9_Part_4',
'Q9_Part_5',
'Q9_Part_6',
'Q9_Part_7',
'Q9_Part_8',
'Q9_Part_11',
'Q10_Part_1',
'Q10_Part_2',
'Q10_Part_16',
'Q11',
'Q12_Part_1',
'Q12_Part_2',
```

In [136]:
```
for column in df[k]:
    mode = df[column].mode()
    df[column] = df[column].fillna(mode)
```

In [141]:
```
for i in k:
    df[i] = df[i].fillna(df[i].mode()[0])
```

In [142]:
```
df
```

Out[142]:

	index	Time from Start to Finish (seconds)	Q1	Q2	Q3	Q4	Q5	Q6	Q7_Part_1	Q7_Part_2	...	Q36_B_Part_3	Q36_B_Part_5	Q36_B_Part_6	C
0	0	Duration (in seconds)	What is your age (# years)?	What is your gender? - Selected Choice	In which country do you currently reside?	What is the highest level of formal education ...	Select the title most similar to your current ...	For how many years have you been writing code ...	What programming languages do you use on a reg...	What programming languages do you use on a reg...	...	Which categories of automated machine learning...	Which categories of automated machine learning...	Which categories of automated machine learning...	
8	8	484	30-34	Man	India	Bachelor's degree	Data Scientist	5-10 years	Python	R	...	Automated model selection (e.g. auto-sklearn, ...	Automated hyperparameter tuning (e.g. hyperopt...	Automation of full ML pipelines (e.g. Google C...	
10	10	655	30-34	Man	Turkey	I prefer not to answer	Other	1-3 years	Python	R	...	Automated model selection (e.g. auto-sklearn, ...	Automated hyperparameter tuning (e.g. hyperopt...	Automation of full ML pipelines (e.g. Google C...	
												Automated	Automated	Automation of	

In [144]:
```
df.reset_index(inplace=True)
```

```
C:\Users\bgompa\AppData\Local\Temp\1\ipykernel_10088\136303746.py:1: PerformanceWarning: DataFrame is highly fragmented.  This is usually the result of calling `frame.insert` many times, which has poor performance.  Consider joining all columns at once using pd.concat(axis=1) instead. To get a de-fragmented frame, use `newframe = frame.copy()`
  df.reset_index(inplace=True)
```

In [145]:
```
df
```

Out[145]:

	level_0	index	Time from Start to Finish (seconds)	Q1	Q2	Q3	Q4	Q5	Q6	Q7_Part_1	...	Q36_B_Part_3	Q36_B_Part_5	Q36_B_Part_6	Q36_B_Pa
0	0	0	Duration (in seconds)	What is your age (# years)?	What is your gender? - Selected Choice	In which country do you currently reside?	What is the highest level of formal education ...	Select the title most similar to your current ...	For how many years have you been writing code ...	What programming languages do you use on a reg...	...	Which categories of automated machine learning...	Which categories of automated machine learning...	Which categories of automated machine learning...	W categori autom mac learni
1	8	8	484	30-34	Man	India	Bachelor's degree	Data Scientist	5-10 years	Python	...	Automated model selection (e.g. auto-sklearn, ...	Automated hyperparameter tuning (e.g. hyperopt...	Automation of full ML pipelines (e.g. Google C...	N
2	10	10	655	30-34	Man	Turkey	I prefer not to answer	Other	1-3 years	Python	...	Automated model selection (e.g. auto-sklearn,	Automated hyperparameter tuning (e.g.	Automation of full ML pipelines (e.g.	N

In [146]:
```
df.drop(["index", "level_0"], axis = 1, inplace = True)
```

In [147]:
```
df
```

Out[147]:

	Time from Start to Finish (seconds)	Q1	Q2	Q3	Q4	Q5	Q6	Q7_Part_1	Q7_Part_2	Q7_Part_3	...	Q36_B_Part_3	Q36_B_Part_5	Q36_B_Par
0	Duration (in seconds)	What is your age (# years)?	What is your gender? - Selected Choice	In which country do you currently reside?	What is the highest level of formal education ...	Select the title most similar to your current ...	For how many years have you been writing code ...	What programming languages do you use on a reg...	What programming languages do you use on a reg...	What programming languages do you use on a reg...	...	Which categories of automated machine learning...	Which categories of automated machine learning...	Wh categorie automa mach learnin
1	484	30-34	Man	India	Bachelor's degree	Data Scientist	5-10 years	Python	R	SQL	...	Automated model selection (e.g. auto-sklearn, ...	Automated hyperparameter tuning (e.g. hyperopt...	Automatio full pipelines (Google
2	655	30-34	Man	Turkey	I prefer not to answer	Other	1-3 years	Python	R	SQL	...	Automated model selection (e.g. auto-sklearn, ...	Automated hyperparameter tuning (e.g. hyperopt...	Automatio full pipelines (Google
3	1777	40-44	Man	Australia	Doctoral degree	Other	1-3 years	Python	R	SQL	...	Automated model selection (e.g. auto-sklearn	Automated hyperparameter tuning (e.g.	Automatio full pipelines (