

Copy of Exploratory

November 6, 2022

```
[1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import warnings
import seaborn as sns

warnings.filterwarnings('ignore')
```

```
[2]: #df = pd.read_csv(r'/2021_OPM_FEVS_PRDF.csv')
df = pd.read_csv(r'/home/brady/dataBC/code-2022/2022Project/FEVS2021_PRDF_CSV/
↳2021_OPM_FEVS_PRDF.csv')
## Checking if there are any duplicate values
df
```

```
[2]:
```

	RandomID	agency	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	...	DRNO	DHISP	\
0	112970976817	XX	5.0	5.0	5.0	5.0	5	4	4	4	...	A	B	
1	194868625278	XX	3.0	2.0	4.0	3.0	4	2	4	2	...	NaN	A	
2	152966380283	XX	5.0	5.0	4.0	4.0	3	4	5	4	...	B	B	
3	193041162980	XX	5.0	5.0	5.0	5.0	5	5	5	5	...	B	B	
4	146655962451	XX	4.0	5.0	5.0	4.0	4	3	5	4	...	B	B	
...
292515	154057939422	ST	3.0	4.0	4.0	3.0	4	4	2	4	...	B	B	
292516	151758964104	ST	3.0	4.0	2.0	4.0	2	2	3	3	...	B	B	
292517	143492802997	ST	3.0	3.0	NaN	4.0	4	4	4	3	...	NaN	NaN	
292518	110267537558	ST	2.0	4.0	5.0	5.0	2	1	5	4	...	B	B	
292519	153195177928	ST	4.0	4.0	5.0	5.0	4	3	5	4	...	B	B	

	DDIS	DAGEGRP	DSUPER	DFEDTEN	DSEX	DMIL	DLEAVING	POSTWT
0	B	B	B	A	A	A	A	2.209652
1	B	B	B	B	A	A	C	2.209652
2	B	B	A	B	A	B	C	1.858874
3	A	B	B	B	A	A	A	1.228573
4	B	B	B	B	A	A	A	1.735842
...
292515	B	A	A	A	B	B	A	3.004992
292516	B	A	A	A	B	B	D	4.427855
292517	NaN	NaN	NaN	A	A	NaN	A	4.202227

292518	B	A	A	A	B	B	C	3.523113
292519	A	B	A	A	A	B	A	3.400784

[292520 rows x 79 columns]

```
[3]: df.drop_duplicates() ##dropping all the duplicate values
      print()
      print(df.info()) ##checking the data types of every column
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 292520 entries, 0 to 292519
Data columns (total 79 columns):
#   Column      Non-Null Count  Dtype
---  -
0   RandomID    292520 non-null  int64
1   agency      292520 non-null  object
2   Q1          291074 non-null  float64
3   Q2          288624 non-null  float64
4   Q3          290112 non-null  float64
5   Q4          291324 non-null  float64
6   Q5          290770 non-null  object
7   Q6          286449 non-null  object
8   Q7          290240 non-null  object
9   Q8          291451 non-null  object
10  Q9          291910 non-null  float64
11  Q10         292038 non-null  object
12  Q11         292018 non-null  object
13  Q12         291755 non-null  object
14  Q13         291920 non-null  object
15  Q14         291306 non-null  object
16  Q15         288257 non-null  object
17  Q16         290367 non-null  object
18  Q17         290268 non-null  object
19  Q18         291274 non-null  object
20  Q19         291402 non-null  object
21  Q20         289926 non-null  object
22  Q21         290110 non-null  object
23  Q22         290350 non-null  object
24  Q23         289942 non-null  float64
25  Q24         290169 non-null  object
26  Q25         289267 non-null  object
27  Q26         288926 non-null  object
28  Q27         289019 non-null  object
29  Q28         287561 non-null  float64
30  Q29         287462 non-null  float64
31  Q30         287369 non-null  float64
```

32	Q31	288152	non-null	float64
33	Q32	286521	non-null	object
34	Q33	285246	non-null	object
35	Q34	285497	non-null	object
36	Q35	285918	non-null	object
37	Q36	285299	non-null	object
38	Q37	285213	non-null	object
39	Q38	285251	non-null	object
40	Q39	283445	non-null	float64
41	Q40	282267	non-null	float64
42	Q41	282219	non-null	float64
43	Q42	281587	non-null	float64
44	Q43	282840	non-null	float64
45	Q44	283310	non-null	float64
46	Q45	283136	non-null	float64
47	Q46	282920	non-null	float64
48	Q47_01	279206	non-null	float64
49	Q47_02	277687	non-null	float64
50	Q47_03	276577	non-null	float64
51	Q47_04	277660	non-null	float64
52	Q47_05	279221	non-null	float64
53	Q47_06	278977	non-null	float64
54	Q47_07	279162	non-null	float64
55	Q47_08	279571	non-null	float64
56	Q47_09	279786	non-null	float64
57	Q47_10	278113	non-null	float64
58	Q47_11	277685	non-null	float64
59	Q48	280356	non-null	object
60	Q49	279335	non-null	object
61	Q50	278898	non-null	object
62	Q51	280182	non-null	object
63	Q52	279772	non-null	object
64	Q53	279702	non-null	object
65	Q54	280570	non-null	float64
66	Q55	279820	non-null	object
67	Q56	279481	non-null	object
68	Q57	279607	non-null	object
69	DRNO	240635	non-null	object
70	DHISP	253732	non-null	object
71	DDIS	258336	non-null	object
72	DAGEGRP	260285	non-null	object
73	DSUPER	272402	non-null	object
74	DFEDTEN	272112	non-null	object
75	DSEX	260998	non-null	object
76	DMIL	272565	non-null	object
77	DLEAVING	273309	non-null	object
78	POSTWT	292520	non-null	float64

dtypes: float64(31), int64(1), object(47)

memory usage: 176.3+ MB
None

```
[4]: ## There are certain variables which should be nature but they are not because  
    ↪ some of the entries have string value.  
df_c = df.copy()  
objects= [column for column, is_type in (df_c.dtypes=="object").items() if  
    ↪ is_type] ## Extracting out all the value which should be integer but is not  
## The column 'agency' holds no significance here so it is to be dropped.  
objects.remove('agency')  
for i in objects:  
    df_c[i] = df_c[i].replace(['X'], np.nan)  
    df_c[i] = df_c[i].replace(np.nan, 0)  
df_c.isnull().sum().sum()
```

[4]: 276131

```
[5]: objects[:-9]  
for i in objects[:-9]:  
    df_c[i] = df_c[i].astype(float)  
df_c.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 292520 entries, 0 to 292519  
Data columns (total 79 columns):  
#   Column      Non-Null Count  Dtype  
---  -  
0   RandomID    292520 non-null  int64  
1   agency      292520 non-null  object  
2   Q1          291074 non-null  float64  
3   Q2          288624 non-null  float64  
4   Q3          290112 non-null  float64  
5   Q4          291324 non-null  float64  
6   Q5          292520 non-null  float64  
7   Q6          292520 non-null  float64  
8   Q7          292520 non-null  float64  
9   Q8          292520 non-null  float64  
10  Q9          291910 non-null  float64  
11  Q10         292520 non-null  float64  
12  Q11         292520 non-null  float64  
13  Q12         292520 non-null  float64  
14  Q13         292520 non-null  float64  
15  Q14         292520 non-null  float64  
16  Q15         292520 non-null  float64  
17  Q16         292520 non-null  float64  
18  Q17         292520 non-null  float64  
19  Q18         292520 non-null  float64  
20  Q19         292520 non-null  float64
```

21	Q20	292520	non-null	float64
22	Q21	292520	non-null	float64
23	Q22	292520	non-null	float64
24	Q23	289942	non-null	float64
25	Q24	292520	non-null	float64
26	Q25	292520	non-null	float64
27	Q26	292520	non-null	float64
28	Q27	292520	non-null	float64
29	Q28	287561	non-null	float64
30	Q29	287462	non-null	float64
31	Q30	287369	non-null	float64
32	Q31	288152	non-null	float64
33	Q32	292520	non-null	float64
34	Q33	292520	non-null	float64
35	Q34	292520	non-null	float64
36	Q35	292520	non-null	float64
37	Q36	292520	non-null	float64
38	Q37	292520	non-null	float64
39	Q38	292520	non-null	float64
40	Q39	283445	non-null	float64
41	Q40	282267	non-null	float64
42	Q41	282219	non-null	float64
43	Q42	281587	non-null	float64
44	Q43	282840	non-null	float64
45	Q44	283310	non-null	float64
46	Q45	283136	non-null	float64
47	Q46	282920	non-null	float64
48	Q47_01	279206	non-null	float64
49	Q47_02	277687	non-null	float64
50	Q47_03	276577	non-null	float64
51	Q47_04	277660	non-null	float64
52	Q47_05	279221	non-null	float64
53	Q47_06	278977	non-null	float64
54	Q47_07	279162	non-null	float64
55	Q47_08	279571	non-null	float64
56	Q47_09	279786	non-null	float64
57	Q47_10	278113	non-null	float64
58	Q47_11	277685	non-null	float64
59	Q48	292520	non-null	float64
60	Q49	292520	non-null	float64
61	Q50	292520	non-null	float64
62	Q51	292520	non-null	float64
63	Q52	292520	non-null	float64
64	Q53	292520	non-null	float64
65	Q54	280570	non-null	float64
66	Q55	292520	non-null	float64
67	Q56	292520	non-null	float64
68	Q57	292520	non-null	float64

```

69 DRNO      292520 non-null object
70 DHISP     292520 non-null object
71 DDIS      292520 non-null object
72 DAGEGRP   292520 non-null object
73 DSUPER    292520 non-null object
74 DFEDTEN   292520 non-null object
75 DSEX      292520 non-null object
76 DMIL      292520 non-null object
77 DLEAVING  292520 non-null object
78 POSTWT    292520 non-null float64
dtypes: float64(68), int64(1), object(10)
memory usage: 176.3+ MB

```

```

[6]: df_c = df_c.dropna()
df_c.isnull().sum().sum() ##-> No missing values found in the dataset

```

```
[6]: 0
```

```

[7]: print('The number of survey entries altered:',df.shape[0]-df_c.shape[0])
print('The percentage of alterations from the original dataset is: ',round(((df.
↪shape[0]-df_c.shape[0])/(df.shape[0])),2)*100,'%')

```

The number of survey entries altered: 54401

The percentage of alterations from the original dataset is: 19.0 %

```

[8]: #check
df_ca = df_c.copy()
for i in df_ca.columns.values:
    df_ca[i] = df_ca[i].replace(0,df_ca[i].value_counts().idxmax())

df_ca.head(20)

```

```

[8]:      RandomID agency  Q1  Q2  Q3  Q4  Q5  Q6  Q7  Q8  ...  DRNO  \
0   112970976817    XX  5.0  5.0  5.0  5.0  5.0  4.0  4.0  4.0  ...    A
2   152966380283    XX  5.0  5.0  4.0  4.0  3.0  4.0  5.0  4.0  ...    B
3   193041162980    XX  5.0  5.0  5.0  5.0  5.0  5.0  5.0  5.0  ...    B
4   146655962451    XX  4.0  5.0  5.0  4.0  4.0  3.0  5.0  4.0  ...    B
5   195312241136    XX  4.0  3.0  4.0  4.0  4.0  4.0  4.0  3.0  ...    A
6   187815423862    XX  5.0  4.0  4.0  4.0  4.0  4.0  4.0  4.0  ...    B
7   174698178690    XX  1.0  1.0  2.0  1.0  3.0  1.0  2.0  2.0  ...    B
9   138139604049    XX  4.0  4.0  5.0  4.0  4.0  4.0  4.0  4.0  ...    B
10  172119452278    XX  4.0  4.0  4.0  4.0  5.0  4.0  4.0  4.0  ...    B
11  102302785265    XX  4.0  3.0  4.0  4.0  3.0  4.0  4.0  3.0  ...    B
12  135019931353    XX  4.0  4.0  5.0  5.0  2.0  2.0  4.0  4.0  ...    B
13  105078979724    XX  2.0  3.0  3.0  4.0  1.0  2.0  4.0  3.0  ...    B
14  139579921467    XX  4.0  5.0  5.0  5.0  5.0  5.0  5.0  5.0  ...    B
15  146904434378    XX  1.0  1.0  5.0  1.0  4.0  2.0  5.0  5.0  ...    B

```

16	161966059804	XX	4.0	5.0	3.0	4.0	4.0	4.0	4.0	5.0	...	B
17	143546472323	XX	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	...	B
18	186289810894	XX	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	...	B
20	104701506907	XX	5.0	5.0	5.0	5.0	2.0	5.0	5.0	4.0	...	B
21	191607232108	XX	2.0	3.0	5.0	5.0	2.0	4.0	3.0	2.0	...	B
23	102384121624	XX	3.0	4.0	4.0	4.0	5.0	4.0	4.0	2.0	...	B

	DHISP	DDIS	DAGEGRP	DSUPER	DFEDTEN	DSEX	DMIL	DLEAVING	POSTWT
0	B	B	B	B	A	A	A	A	2.209652
2	B	B	B	A	B	A	B	C	1.858874
3	B	A	B	B	B	A	A	A	1.228573
4	B	B	B	B	B	A	A	A	1.735842
5	B	B	B	A	B	B	A	A	2.076046
6	B	B	B	B	B	A	A	A	1.858874
7	B	B	B	A	B	B	B	C	1.089288
9	B	A	B	B	A	A	A	A	1.228573
10	B	B	B	B	C	A	A	A	1.735842
11	B	B	B	B	B	A	A	A	1.228573
12	B	B	B	B	B	B	B	A	1.089288
13	B	B	B	B	A	A	B	C	2.209652
14	B	B	B	A	A	A	B	A	1.735842
15	B	B	B	A	B	A	A	C	2.209652
16	B	A	A	A	A	A	A	A	2.010048
17	B	B	B	B	A	B	A	A	1.184402
18	B	B	B	B	A	B	A	A	1.255264
20	B	B	B	A	B	B	B	A	1.184402
21	B	B	B	B	B	A	A	C	1.735842
23	A	B	B	A	A	B	B	A	1.020346

[20 rows x 79 columns]

```
[9]: df_ca['DLEAVING'].value_counts()
```

```
[9]: A    160269
      C     36912
      B     29899
      D     11039
      Name: DLEAVING, dtype: int64
```

```
[10]: df_ca.loc[df_ca.DLEAVING=='C']
```

```
[10]:      RandomID agency  Q1  Q2  Q3  Q4  Q5  Q6  Q7  Q8  ...  \
2      152966380283   XX  5.0  5.0  4.0  4.0  3.0  4.0  5.0  4.0  ...
7      174698178690   XX  1.0  1.0  2.0  1.0  3.0  1.0  2.0  2.0  ...
13     105078979724   XX  2.0  3.0  3.0  4.0  1.0  2.0  4.0  3.0  ...
15     146904434378   XX  1.0  1.0  5.0  1.0  4.0  2.0  5.0  5.0  ...
21     191607232108   XX  2.0  3.0  5.0  5.0  2.0  4.0  3.0  2.0  ...
```

```

...
292401 149388396809 ST 4.0 4.0 4.0 4.0 2.0 2.0 2.0 4.0 ...
292435 183028010363 ST 2.0 2.0 2.0 2.0 2.0 2.0 4.0 3.0 ...
292440 148021669661 ST 4.0 4.0 3.0 4.0 1.0 4.0 4.0 5.0 ...
292486 185461910015 ST 1.0 2.0 1.0 3.0 4.0 1.0 5.0 1.0 ...
292518 110267537558 ST 2.0 4.0 5.0 5.0 2.0 1.0 5.0 4.0 ...

DRNO DHISP DDIS DAGEGRP DSUPER DFEDTEN DSEX DMIL DLEAVING \
2 B B B B A B A B C
7 B B B B A B B B C
13 B B B B B A A B C
15 B B B B A B A A C
21 B B B B B B A A C
...
292401 B B B A B A A B C
292435 B B B A B B B B C
292440 B A B A A A B B C
292486 A B B A B A B B C
292518 B B B A A A B B C

POSTWT
2 1.858874
7 1.089288
13 2.209652
15 2.209652
21 1.735842
...
292401 3.957027
292435 2.705716
292440 1.955644
292486 4.662723
292518 3.523113

```

[36912 rows x 79 columns]

```

[11]: df_ca_d = df_ca.loc[df_ca.DLEAVING=='D']

cols = df_ca_d.columns[2:45]
objects = [column for column, is_type in (df_ca_d.dtypes=="float").items() if
    is_type]

for i in objects:
    plt.figure(figsize=(20,40))
    sns.countplot(data=df_ca_d[cols],y=df_ca_d['agency'],hue=i)
    plt.plot()

plt.savefig("dLeavingB.png")

```



```

-----
ValueError                                Traceback (most recent call last)
Input In [11], in <cell line: 6>()
      6 for i in objects:
      7     plt.figure(figsize=(20,40))
----> 8     sns.countplot(data=df_ca_d[cols],y=df_ca_d['agency'],hue=i)
      9     plt.plot()
     11 plt.savefig("dLeavingB.png")

File ~/anaconda3/envs/my-conda-env/lib/python3.10/site-packages/seaborn/
  _decorators.py:46, in _deprecate_positional_args.<locals>.inner_f(*args,
  **kwargs)
      36     warnings.warn(
      37         "Pass the following variable{} as {}keyword arg{}: {}". "
      38         "From version 0.12, the only valid positional argument "
      (...)
      43     FutureWarning
      44 )
      45 kwargs.update({k: arg for k, arg in zip(sig.parameters, args)})
--> 46 return f(**kwargs)

File ~/anaconda3/envs/my-conda-env/lib/python3.10/site-packages/seaborn/
  categorical.py:3598, in countplot(x, y, hue, data, order, hue_order, orient,
  color, palette, saturation, dodge, ax, **kwargs)
     3595 elif x is not None and y is not None:
     3596     raise ValueError("Cannot pass values for both `x` and `y`")
-> 3598 plotter = _CountPlotter(
     3599     x, y, hue, data, order, hue_order,
     3600     estimator, ci, n_boot, units, seed,
     3601     orient, color, palette, saturation,
     3602     errcolor, errwidth, capsize, dodge
     3603 )
     3605 plotter.value_label = "count"
     3607 if ax is None:

File ~/anaconda3/envs/my-conda-env/lib/python3.10/site-packages/seaborn/
  categorical.py:1584, in _BarPlotter.__init__(self, x, y, hue, data, order,
  hue_order, estimator, ci, n_boot, units, seed, orient, color, palette,
  saturation, errcolor, errwidth, capsize, dodge)
     1579 def __init__(self, x, y, hue, data, order, hue_order,
     1580                 estimator, ci, n_boot, units, seed,
     1581                 orient, color, palette, saturation, errcolor,
     1582                 errwidth, capsize, dodge):
     1583     """Initialize the plotter."""
-> 1584     self.establish_variables(x, y, hue, data, orient,
     1585                             order, hue_order, units)
     1586     self.establish_colors(color, palette, saturation)
     1587     self.estimate_statistic(estimator, ci, n_boot, seed)

```

```

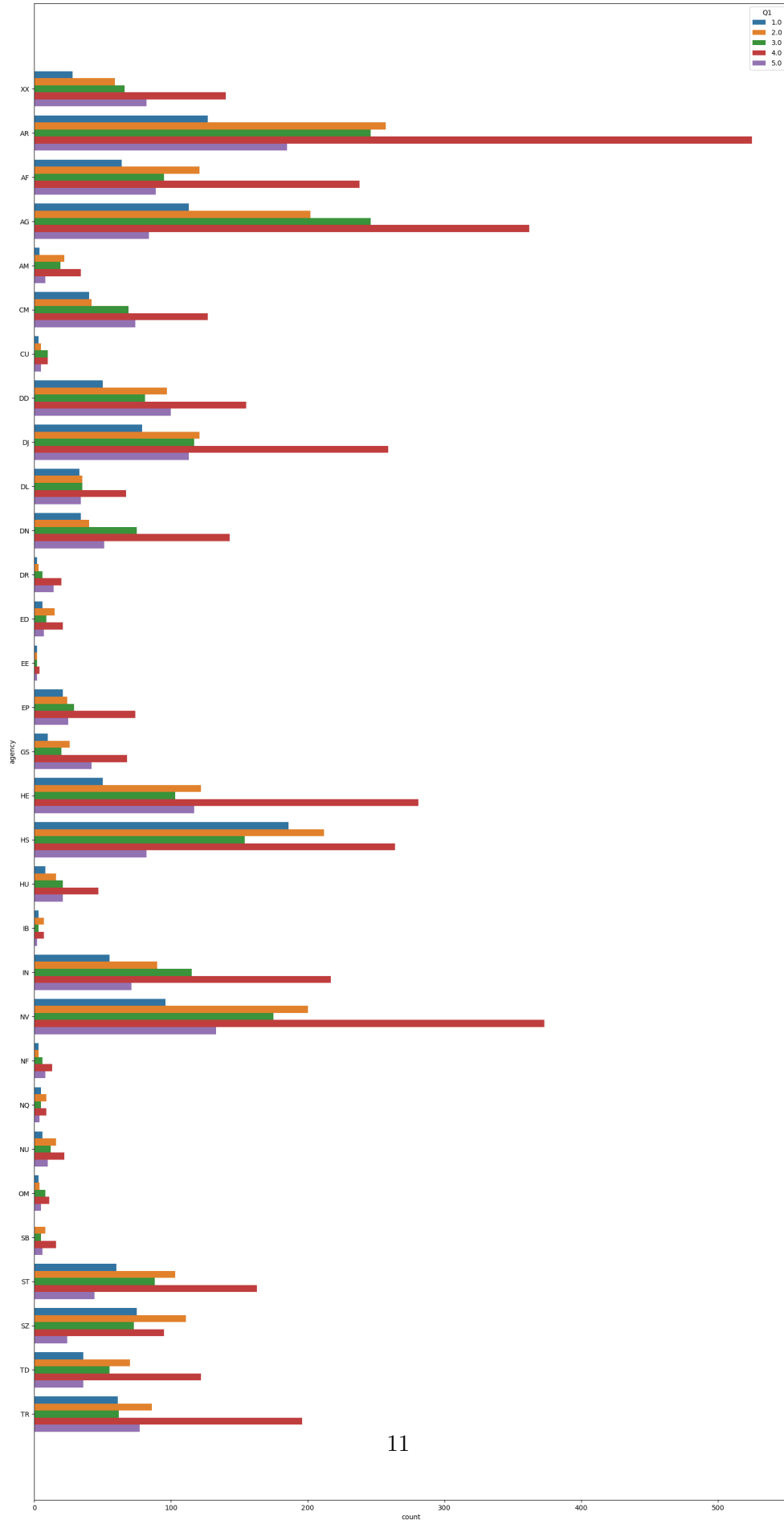
File ~/anaconda3/envs/my-conda-env/lib/python3.10/site-packages/seaborn/
↪categorical.py:153, in _CategoricalPlotter.establish_variables(self, x, y,
↪hue, data, orient, order, hue_order, units)
    151     if isinstance(var, str):
    152         err = "Could not interpret input '{}'.format(var)
--> 153         raise ValueError(err)
    155 # Figure out the plotting orientation
    156 orient = infer_orient(
    157     x, y, orient, require_numeric=self.require_numeric
    158 )

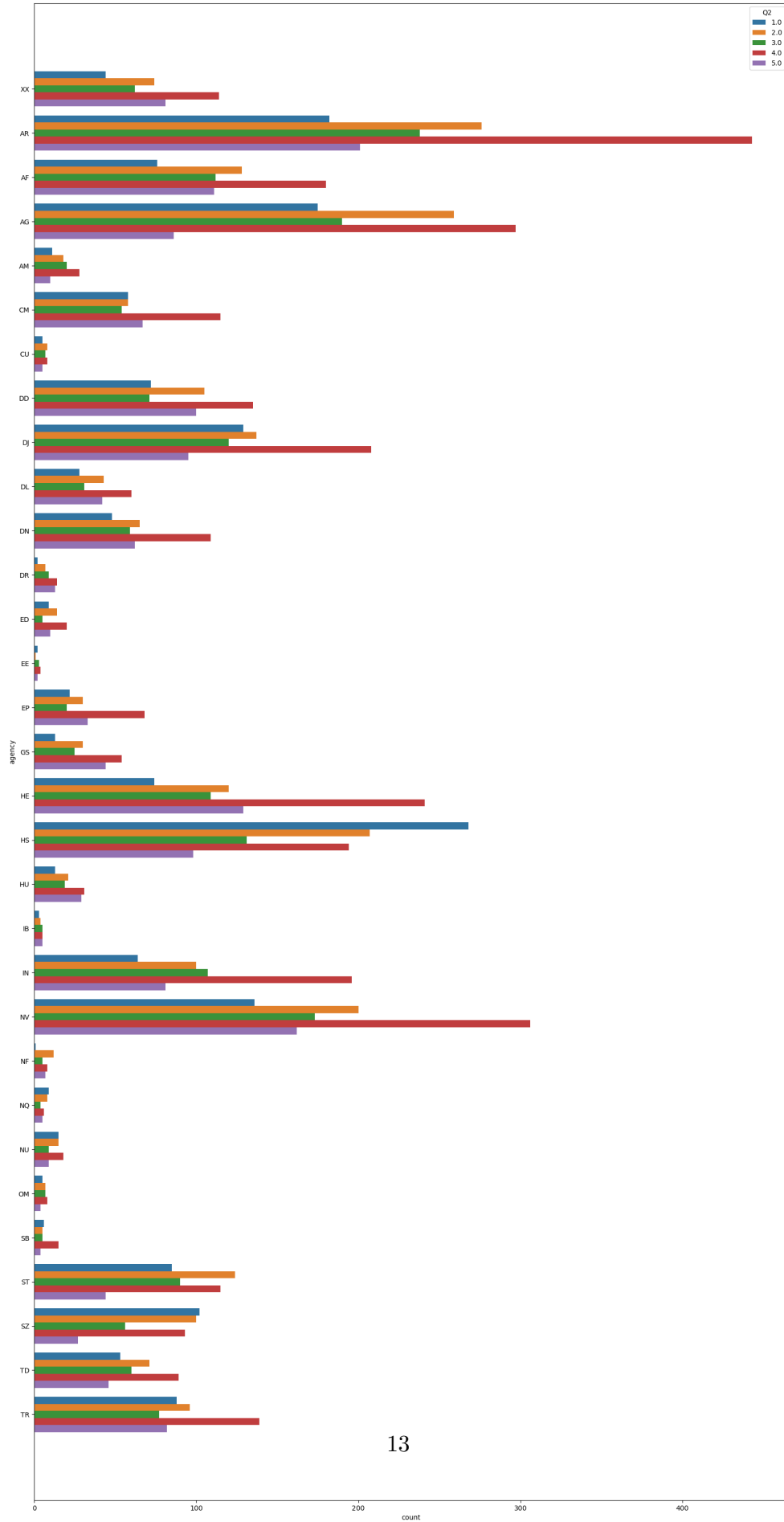
```

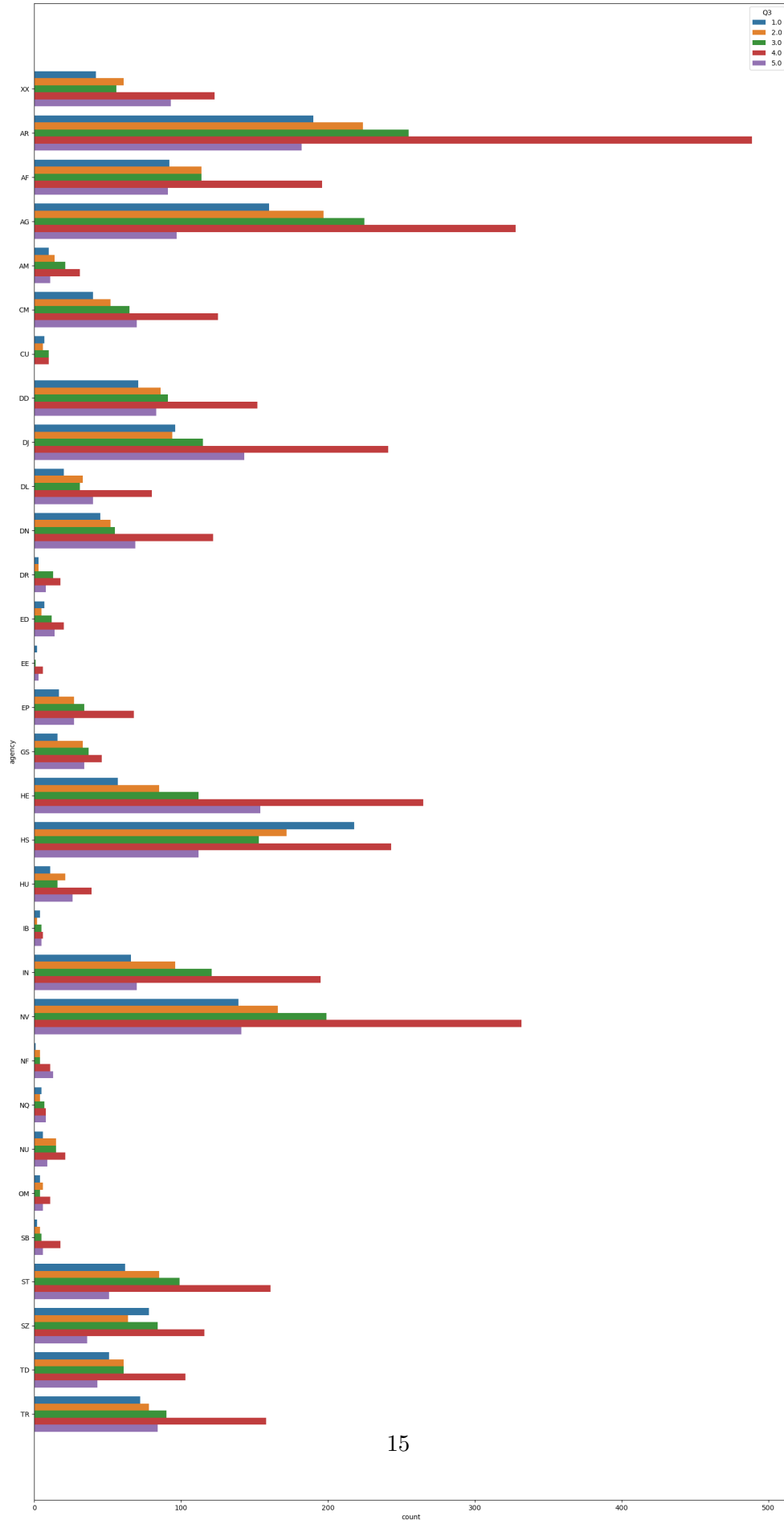
```

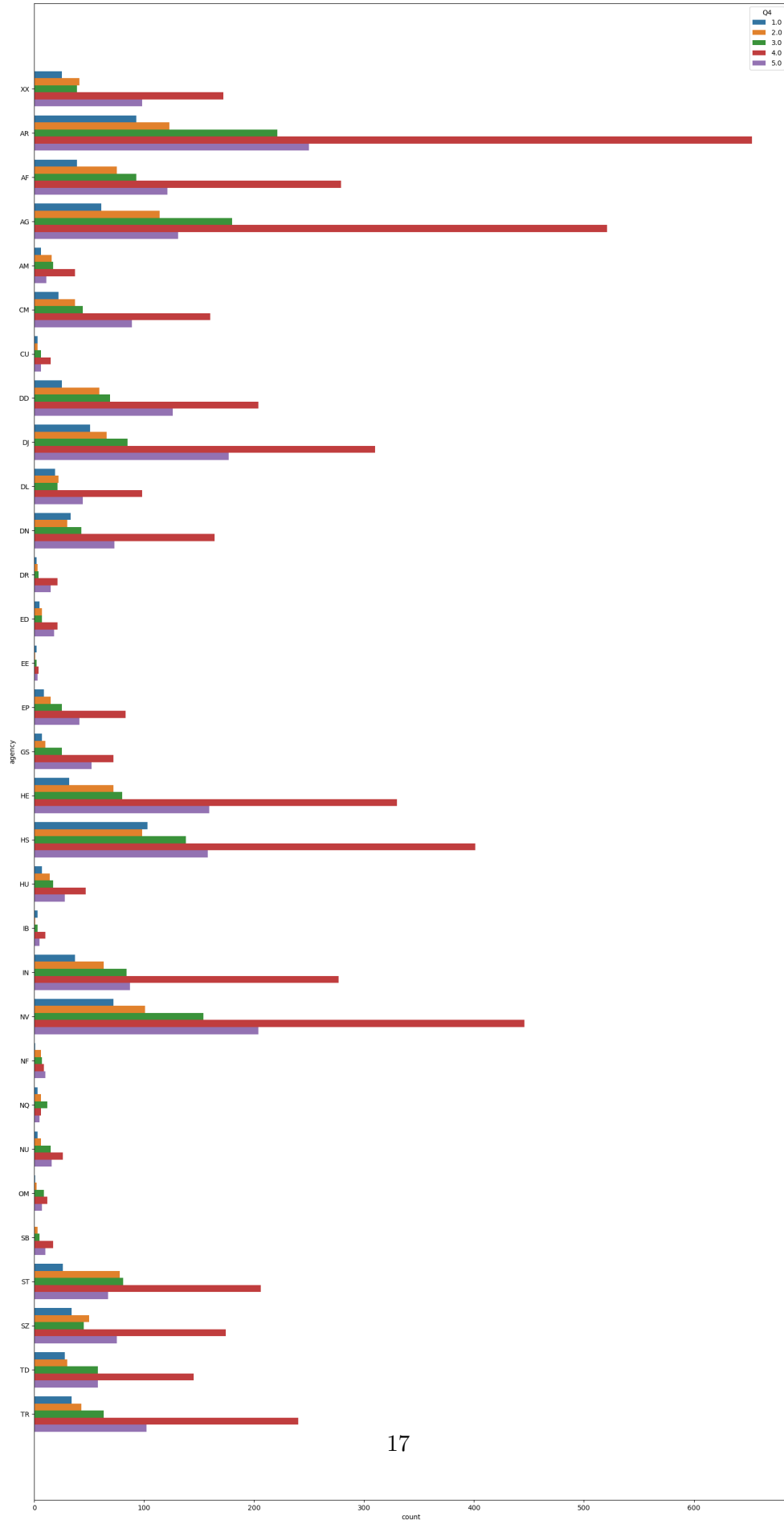
ValueError: Could not interpret input 'Q44'

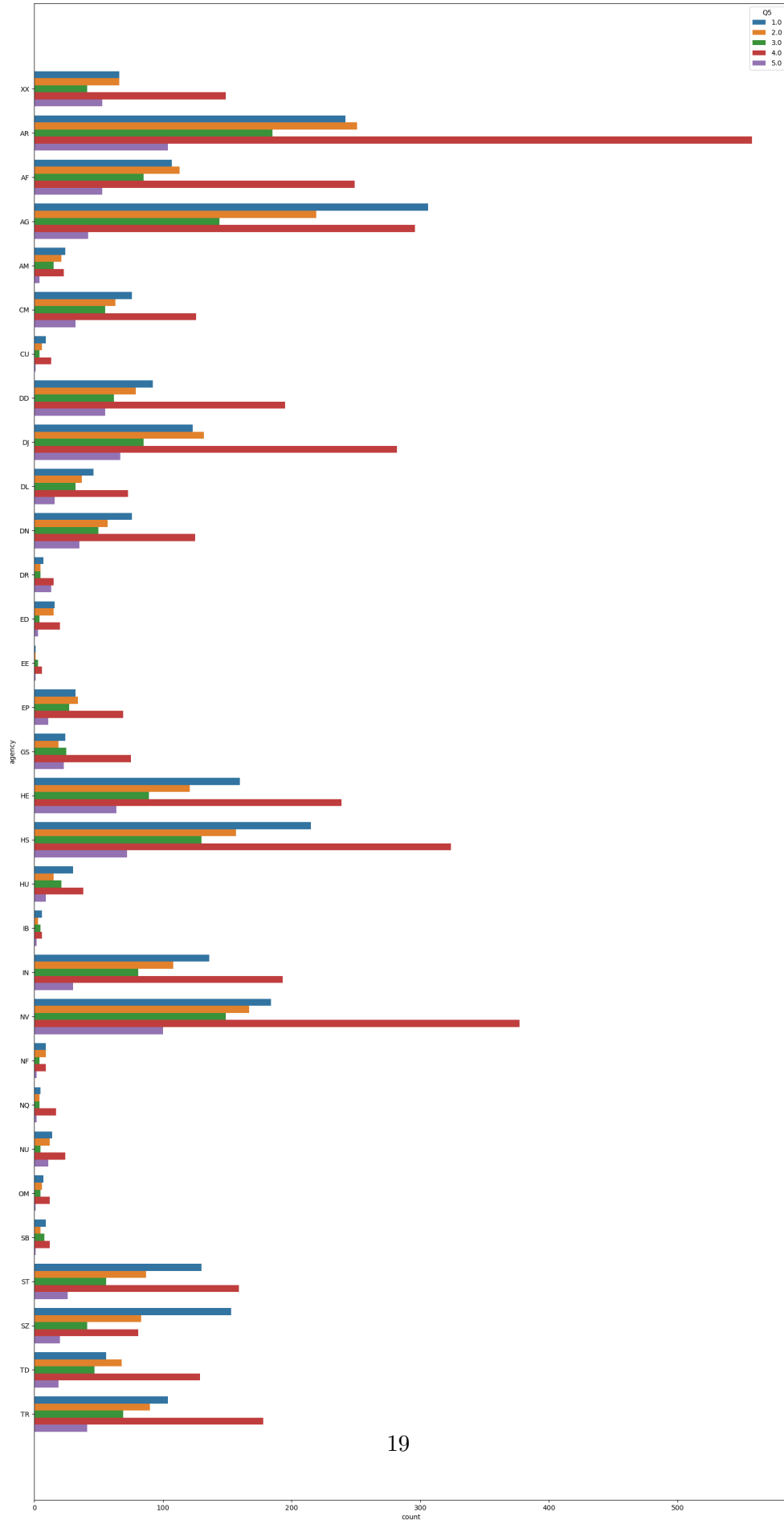
```

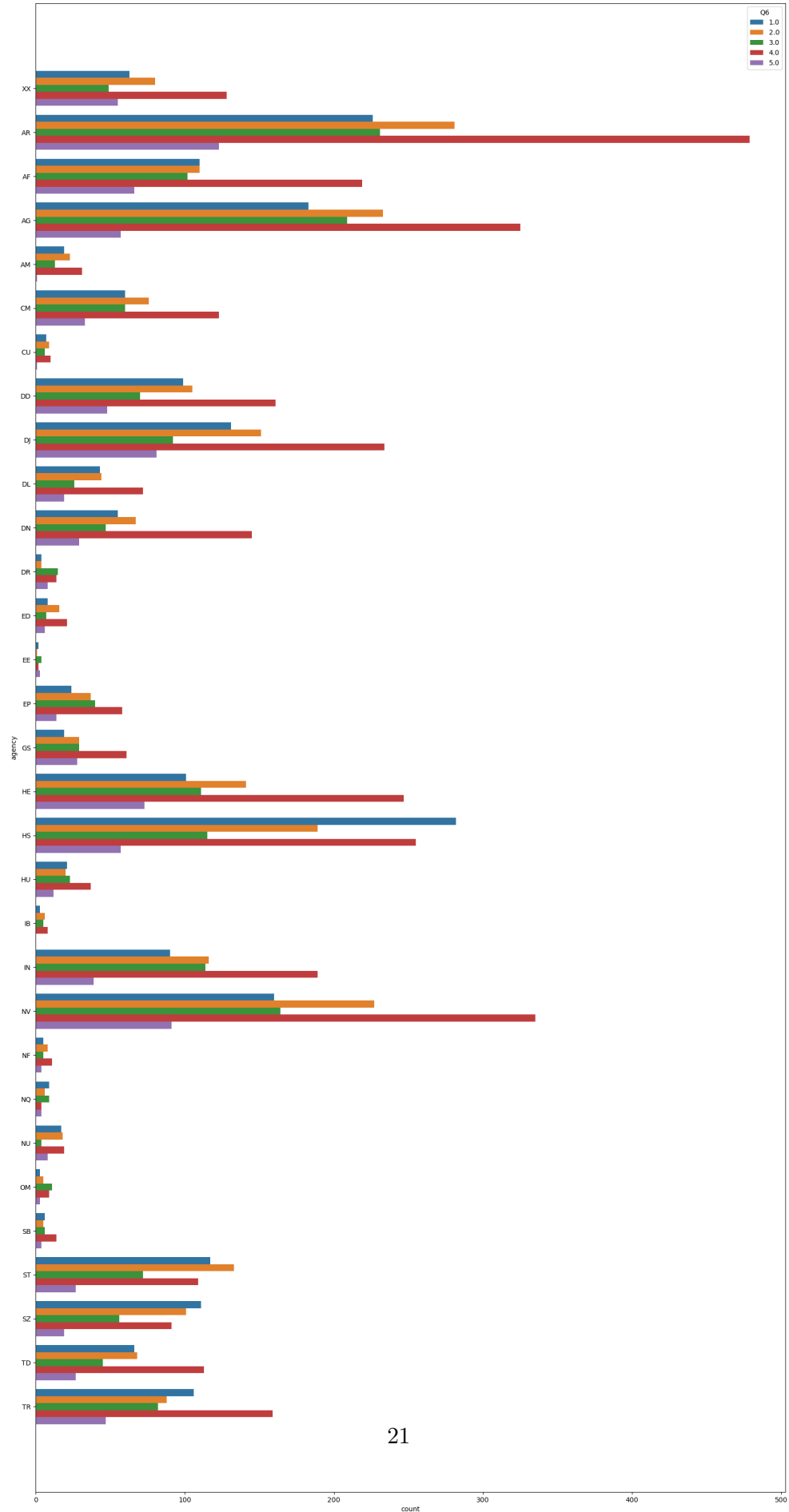


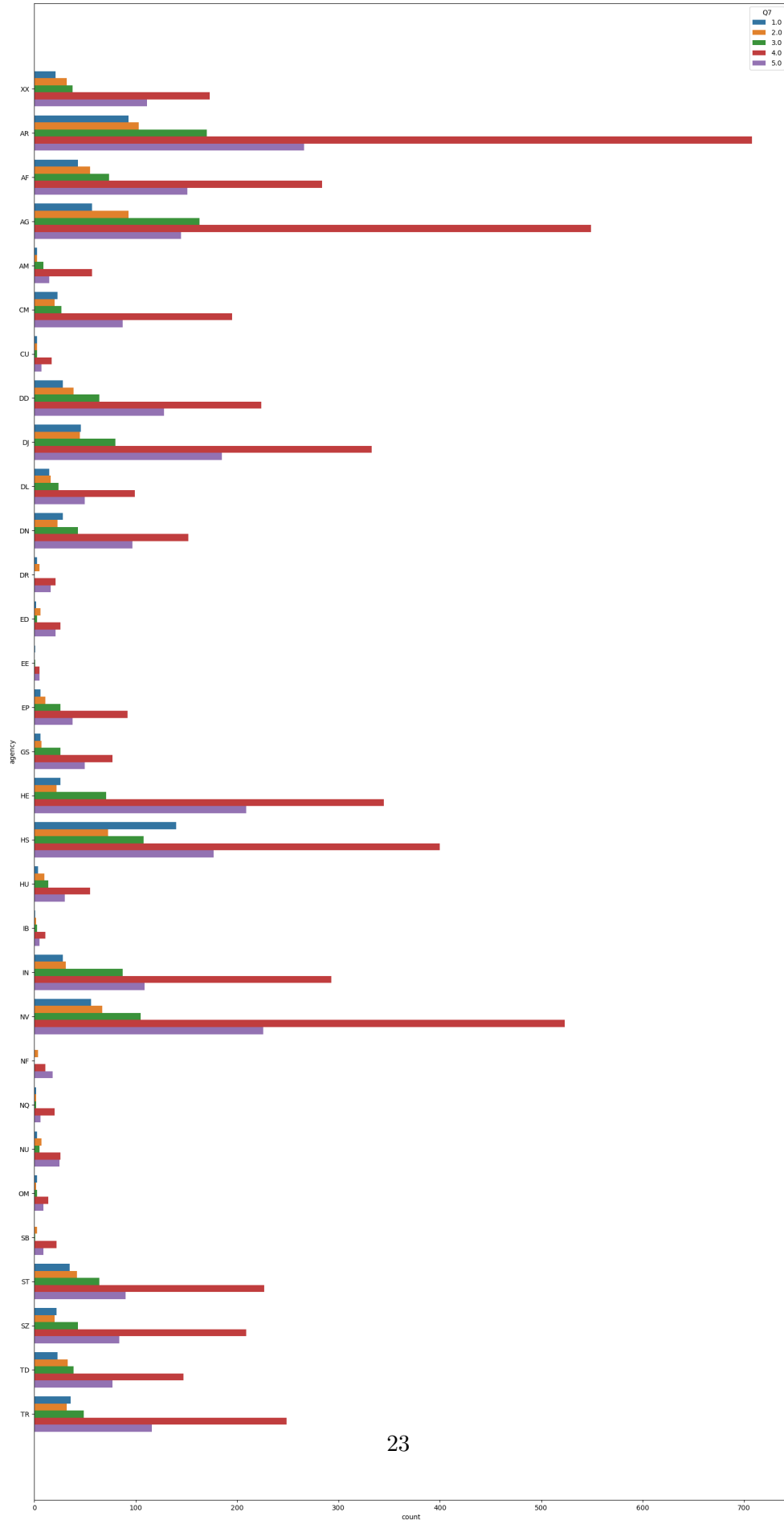


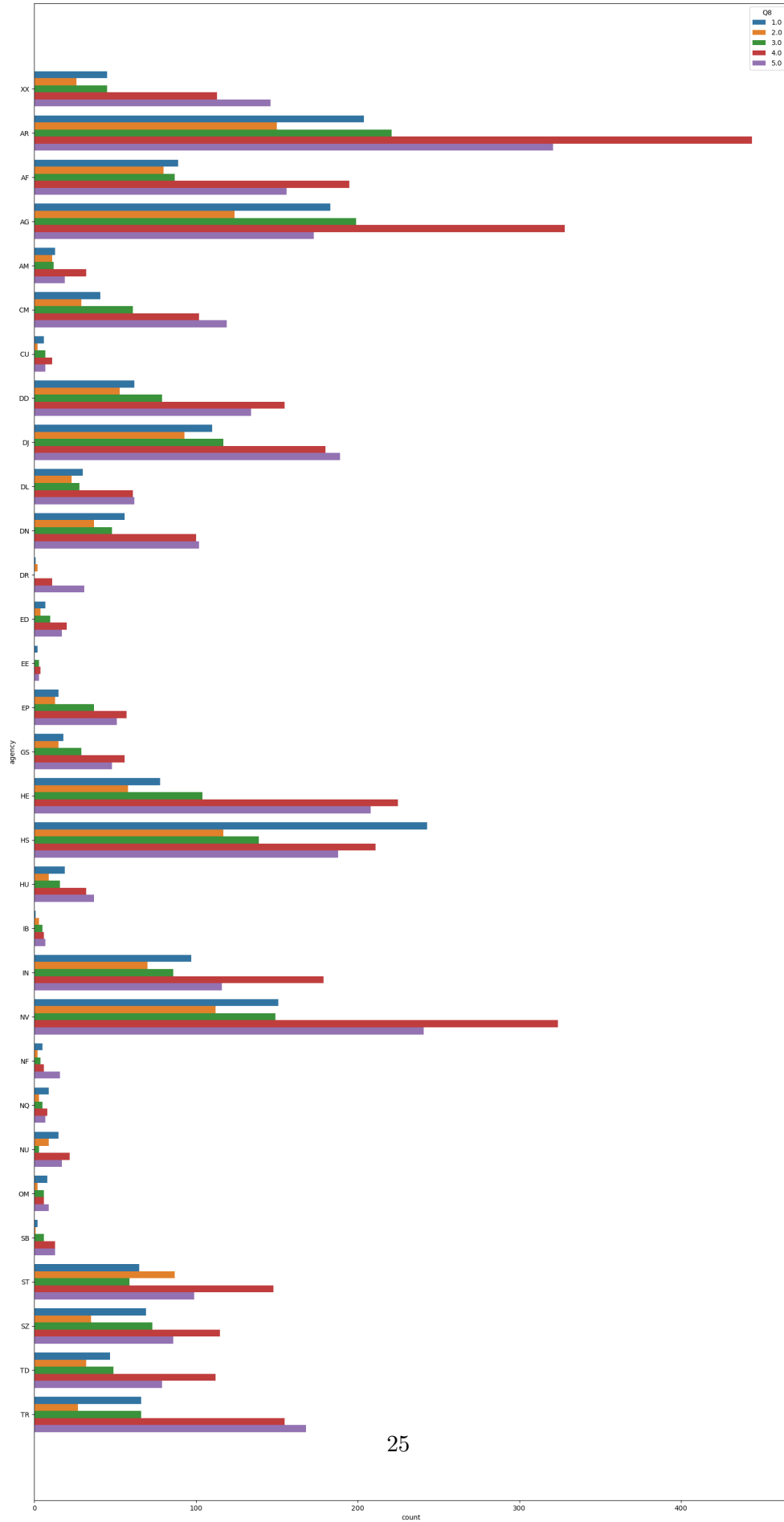


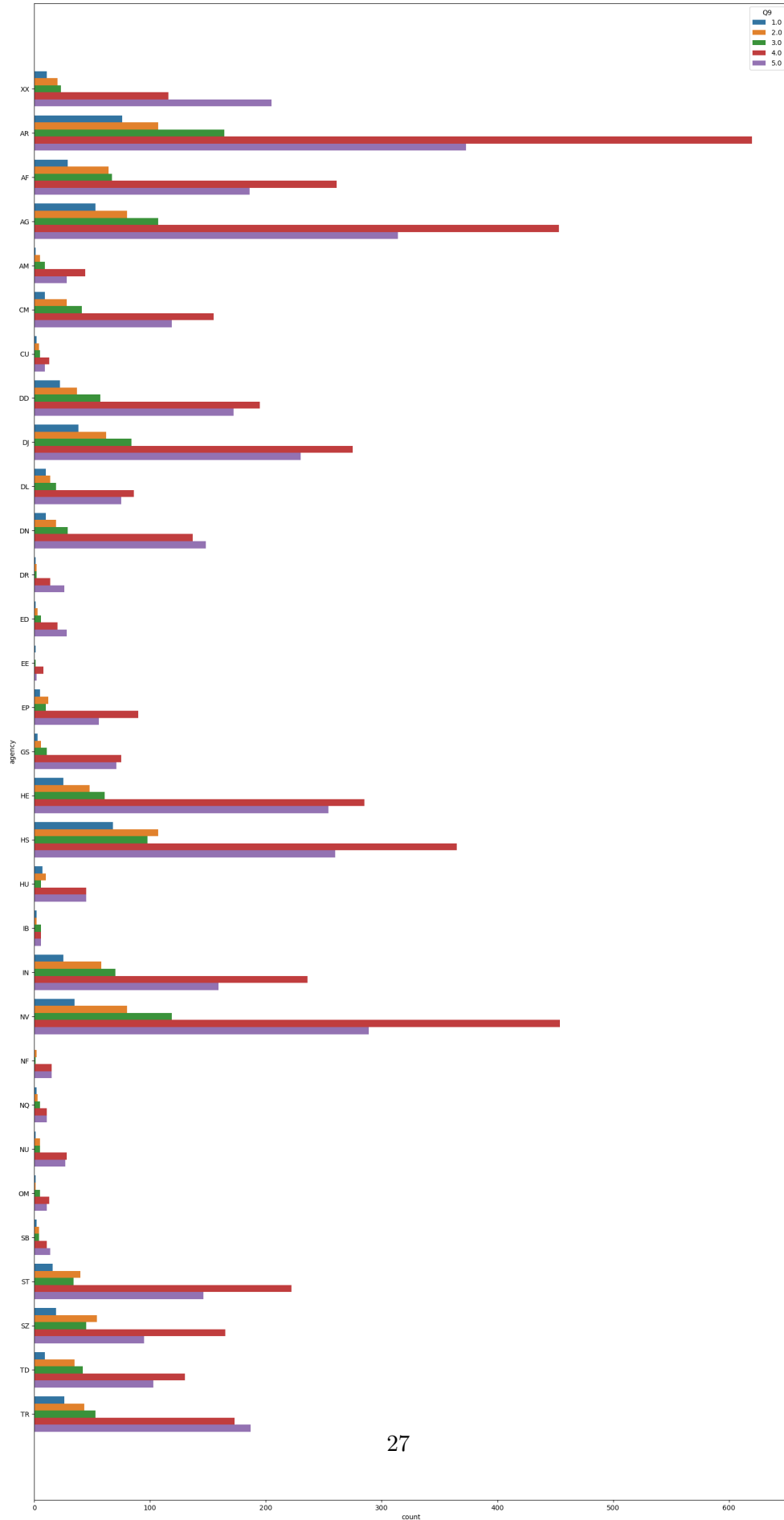


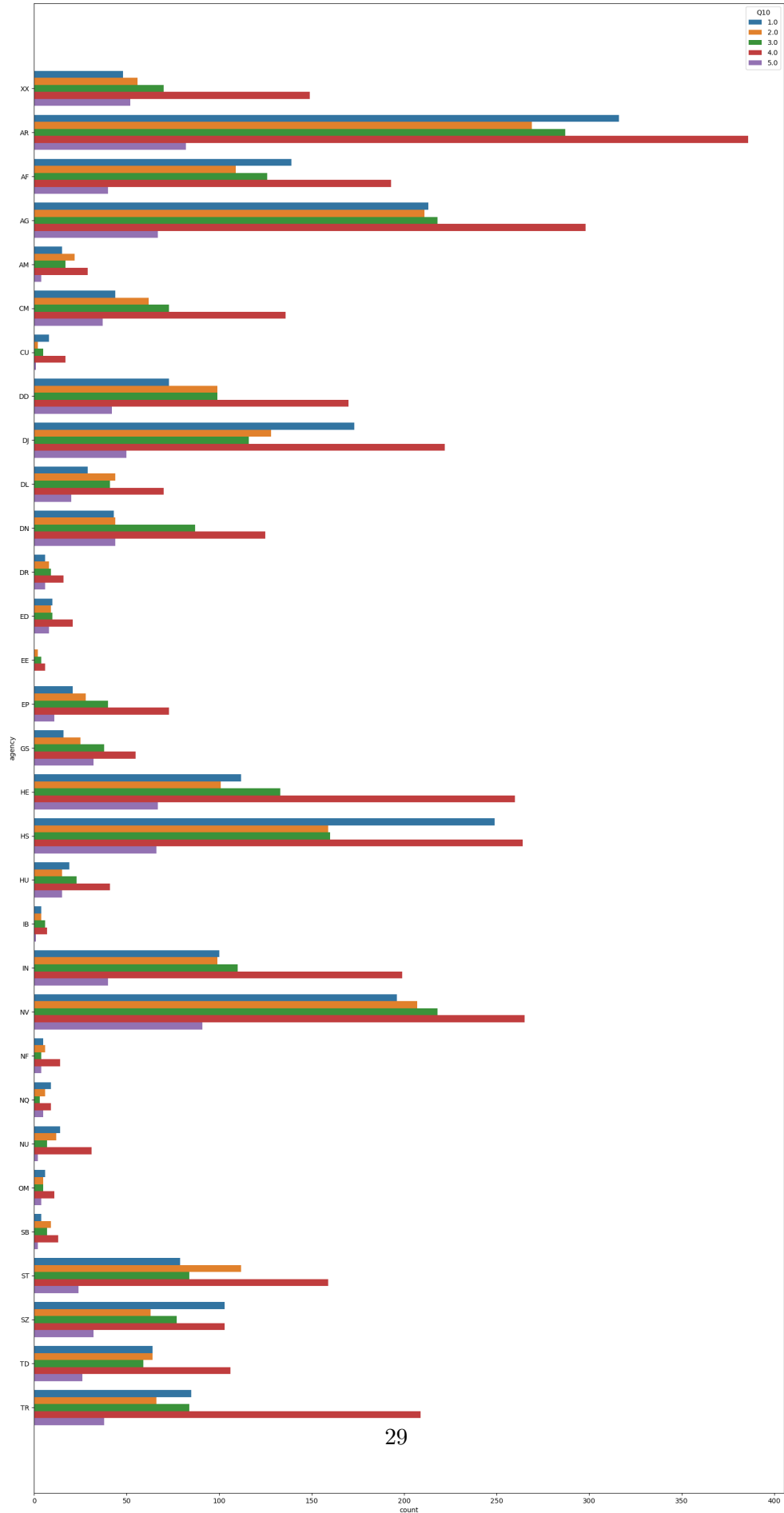


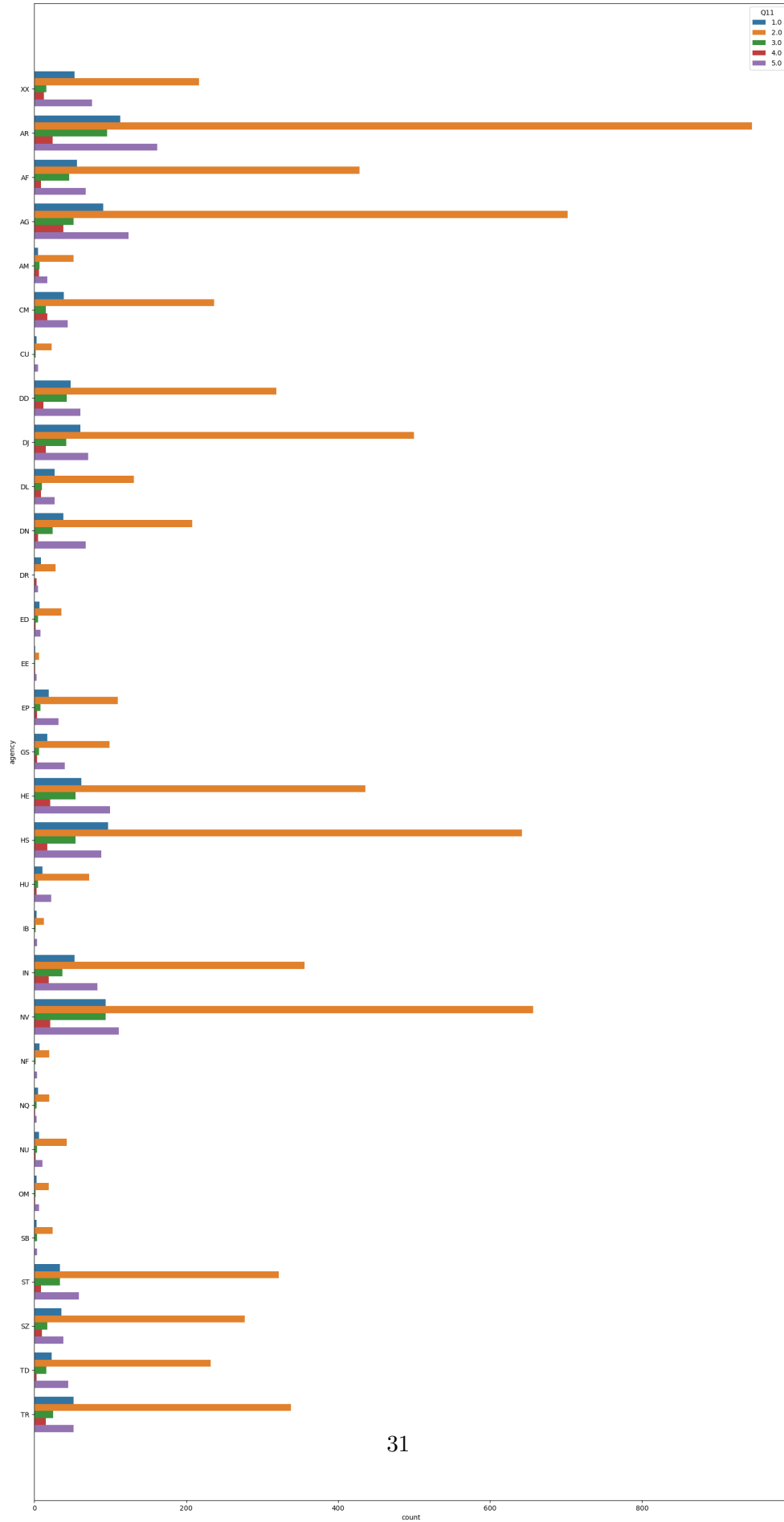


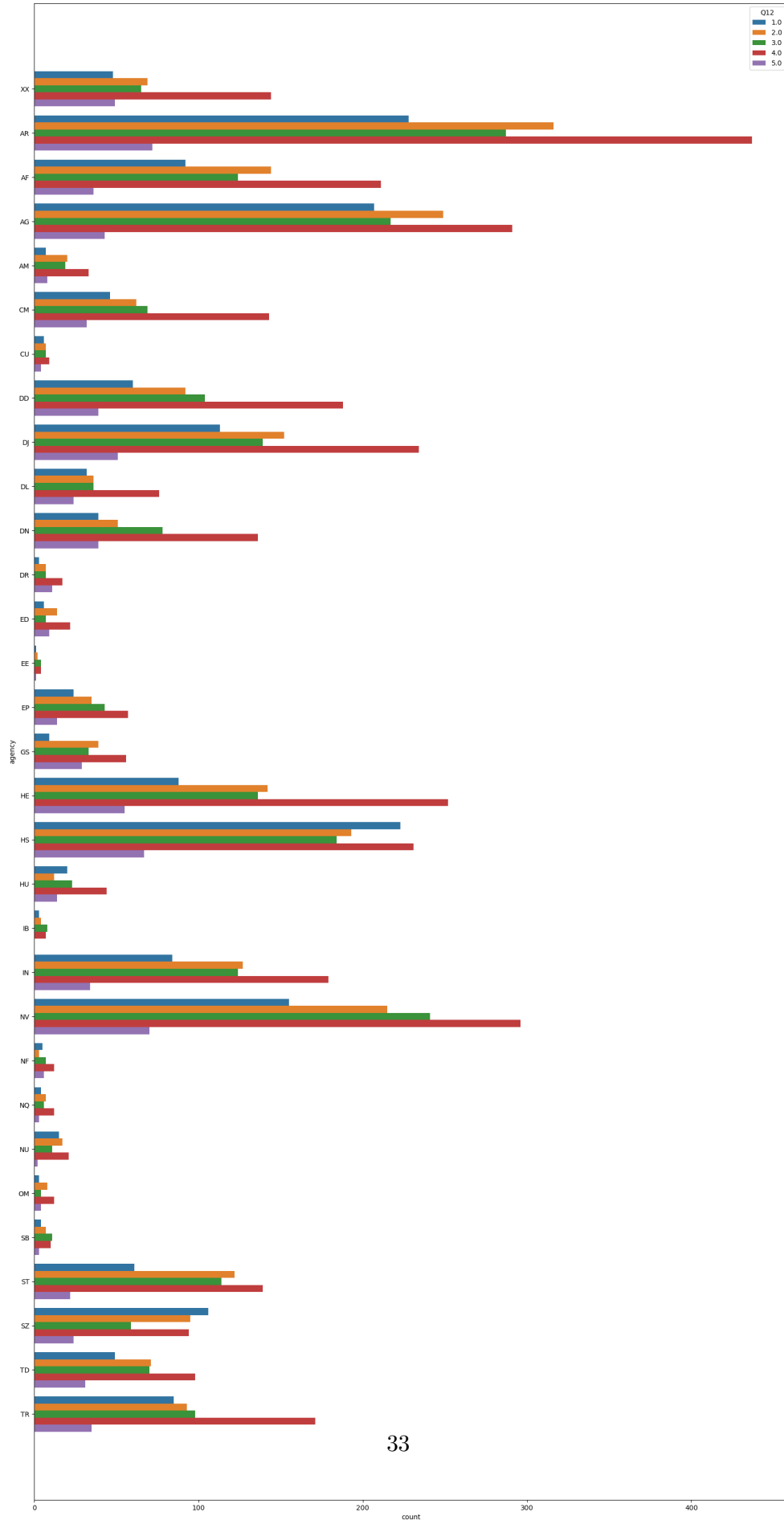


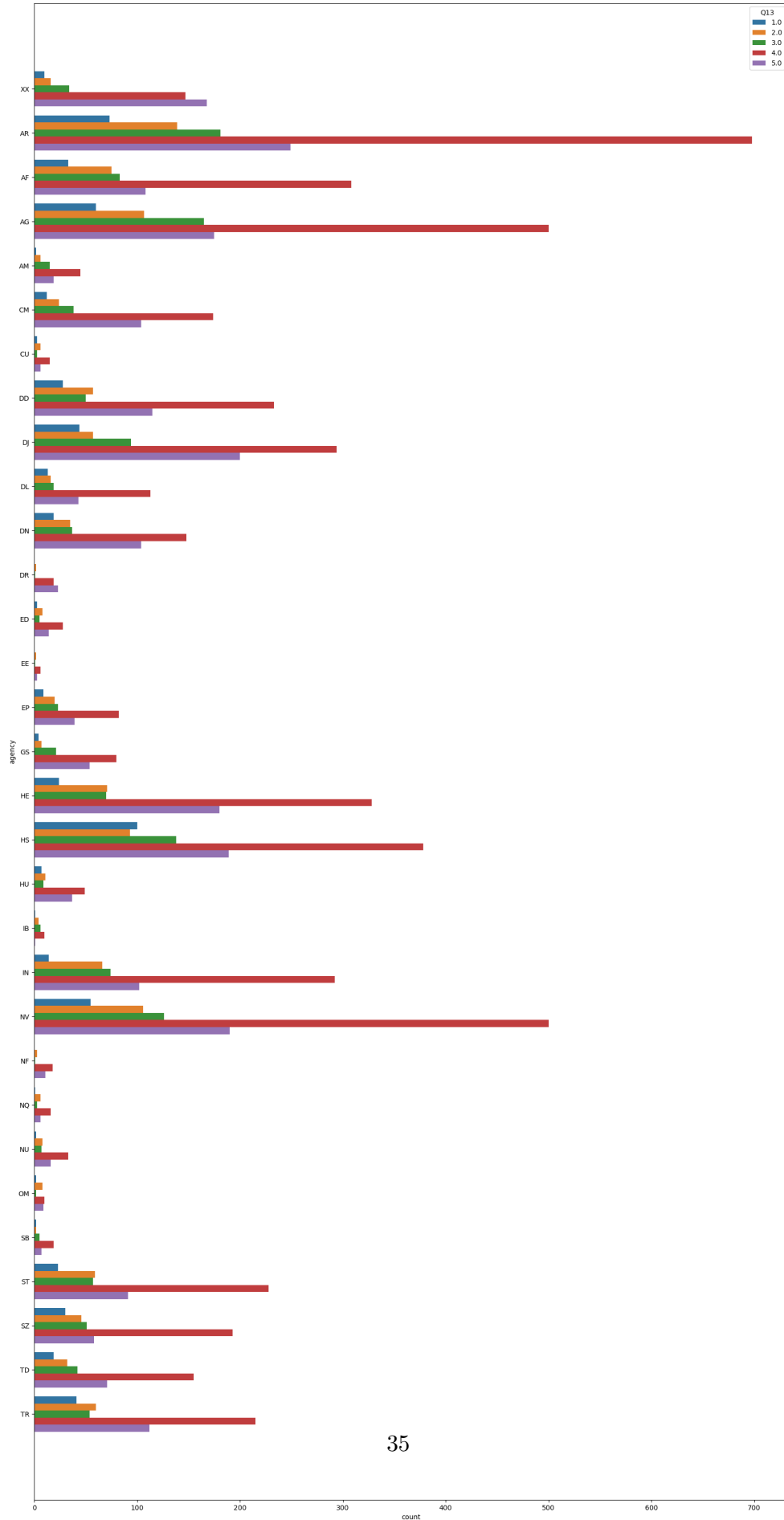


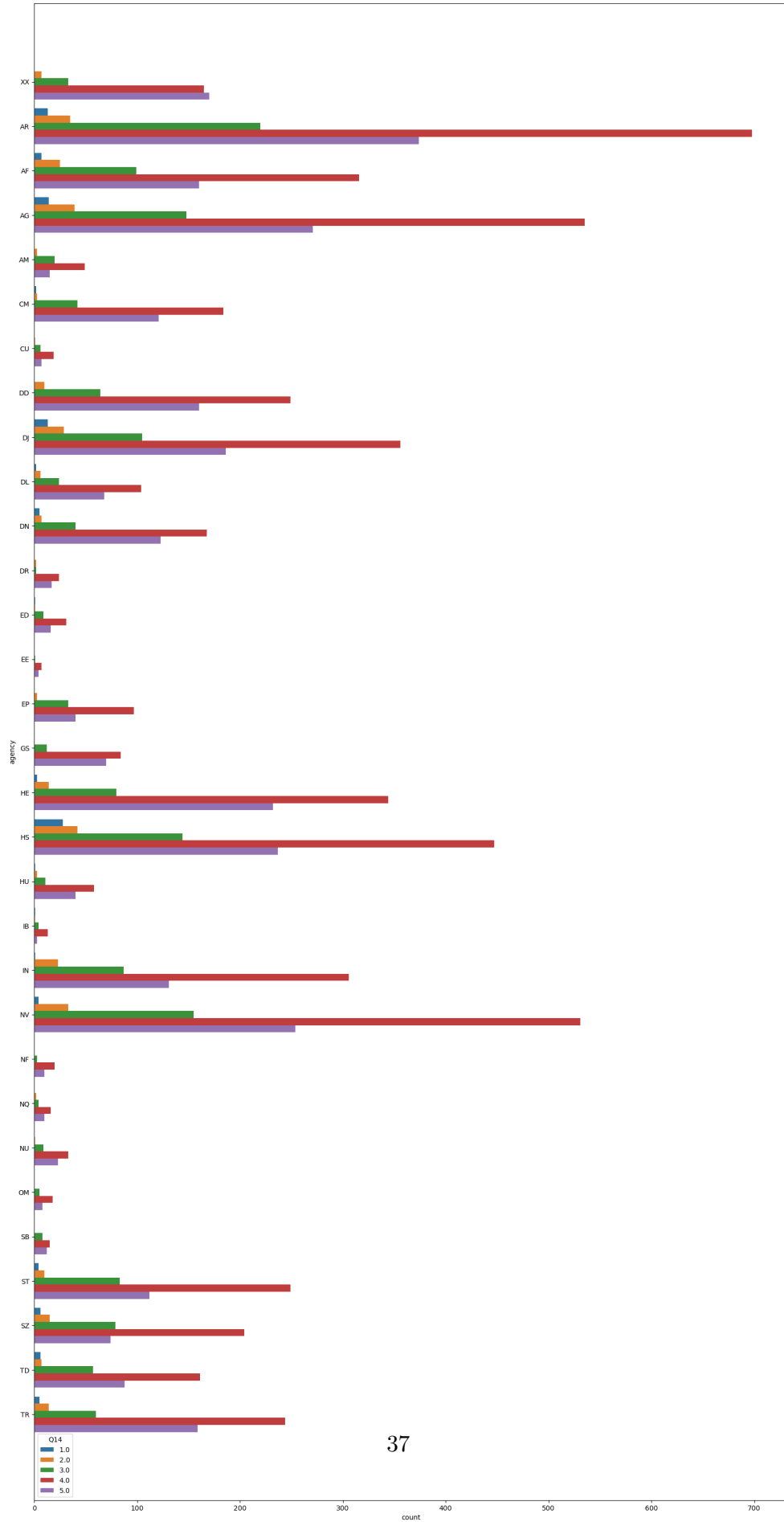


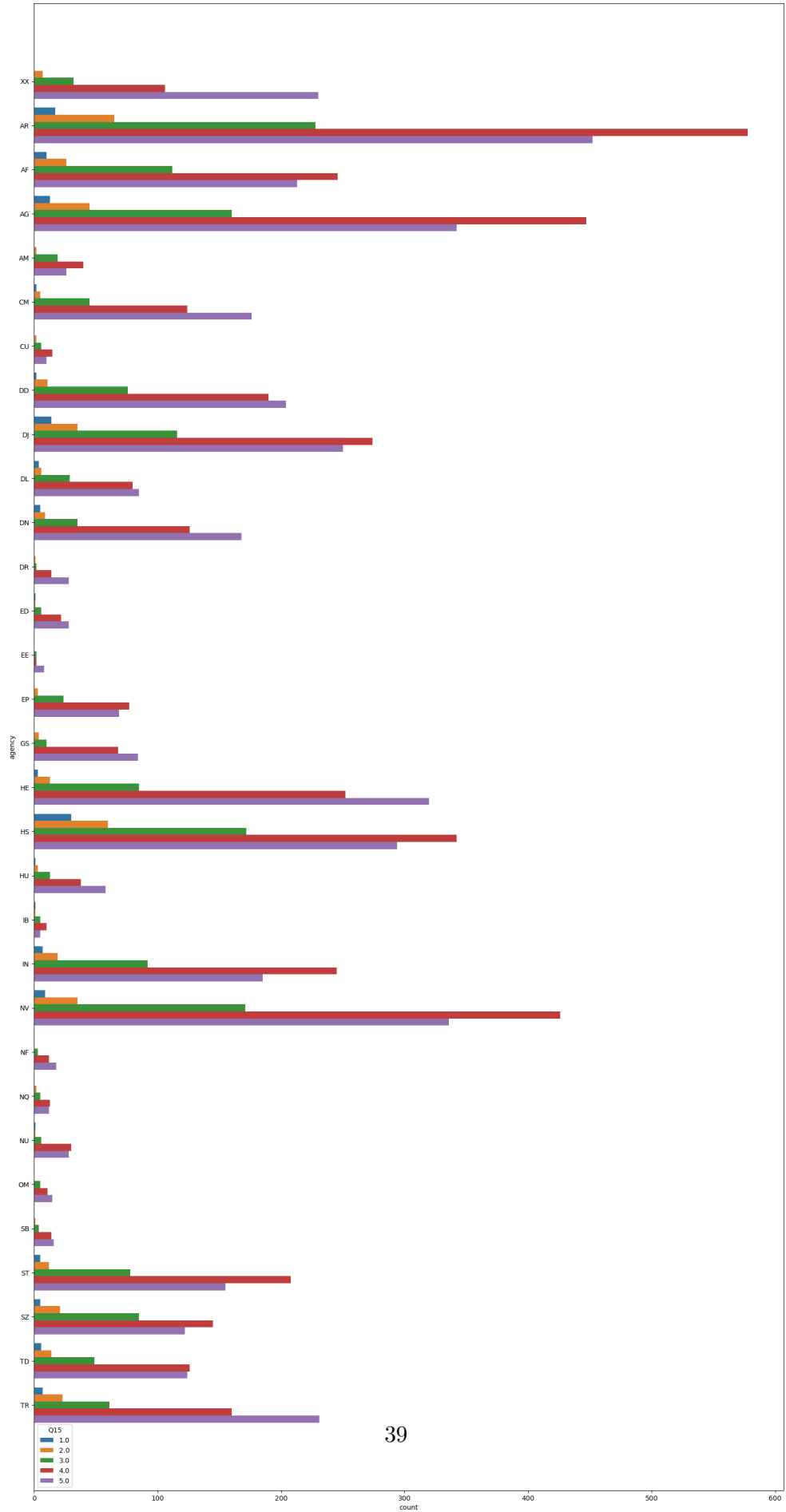


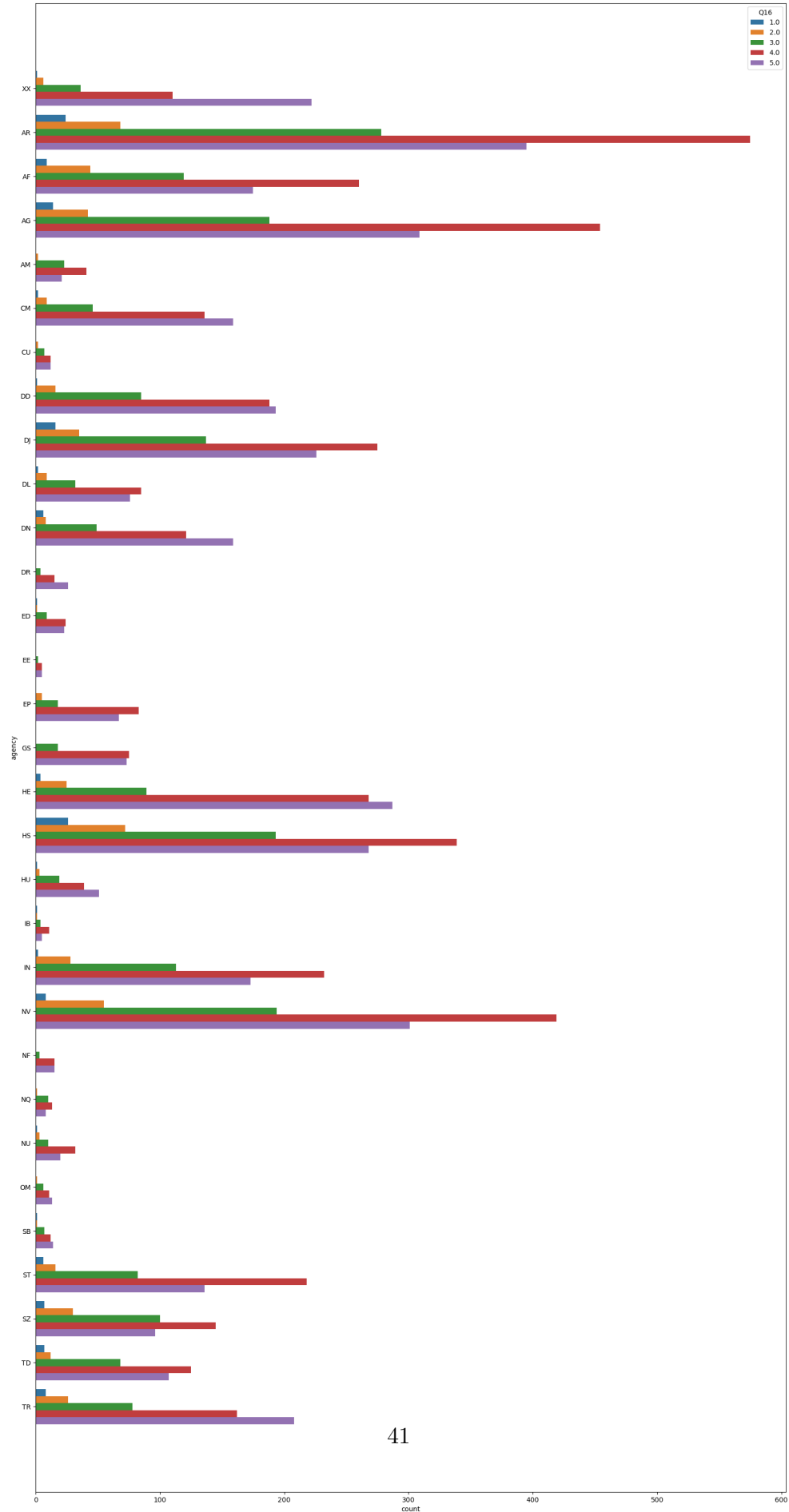


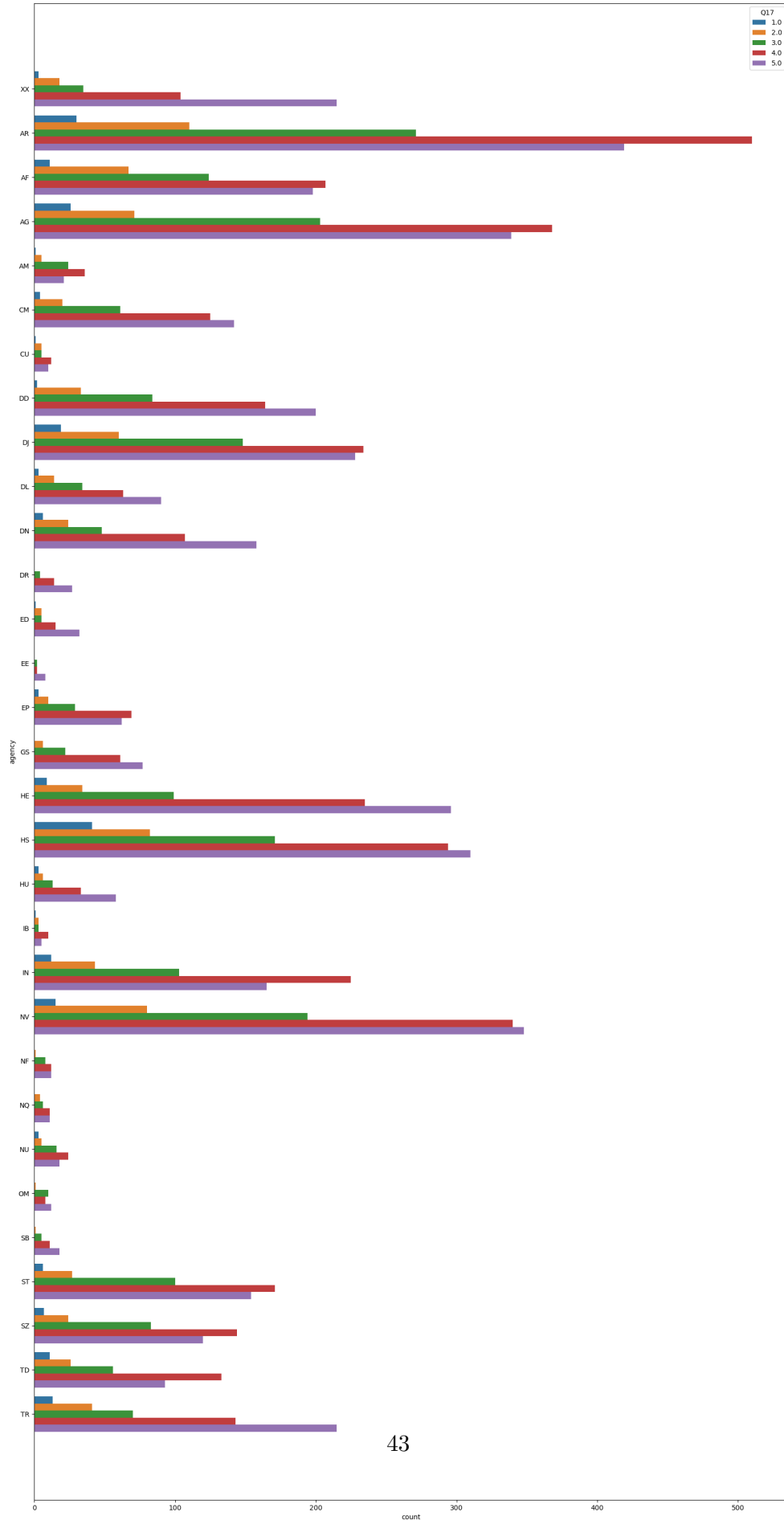


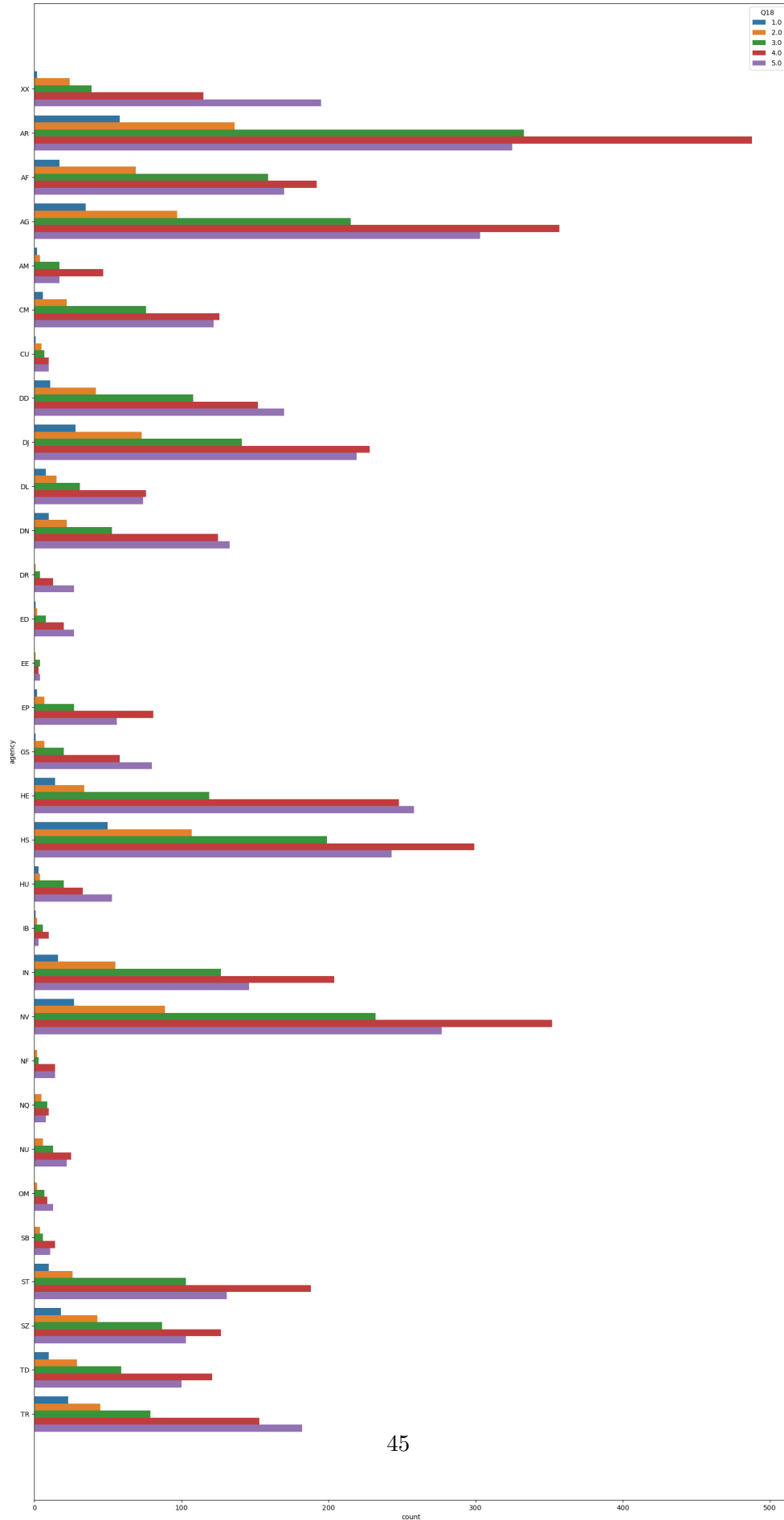


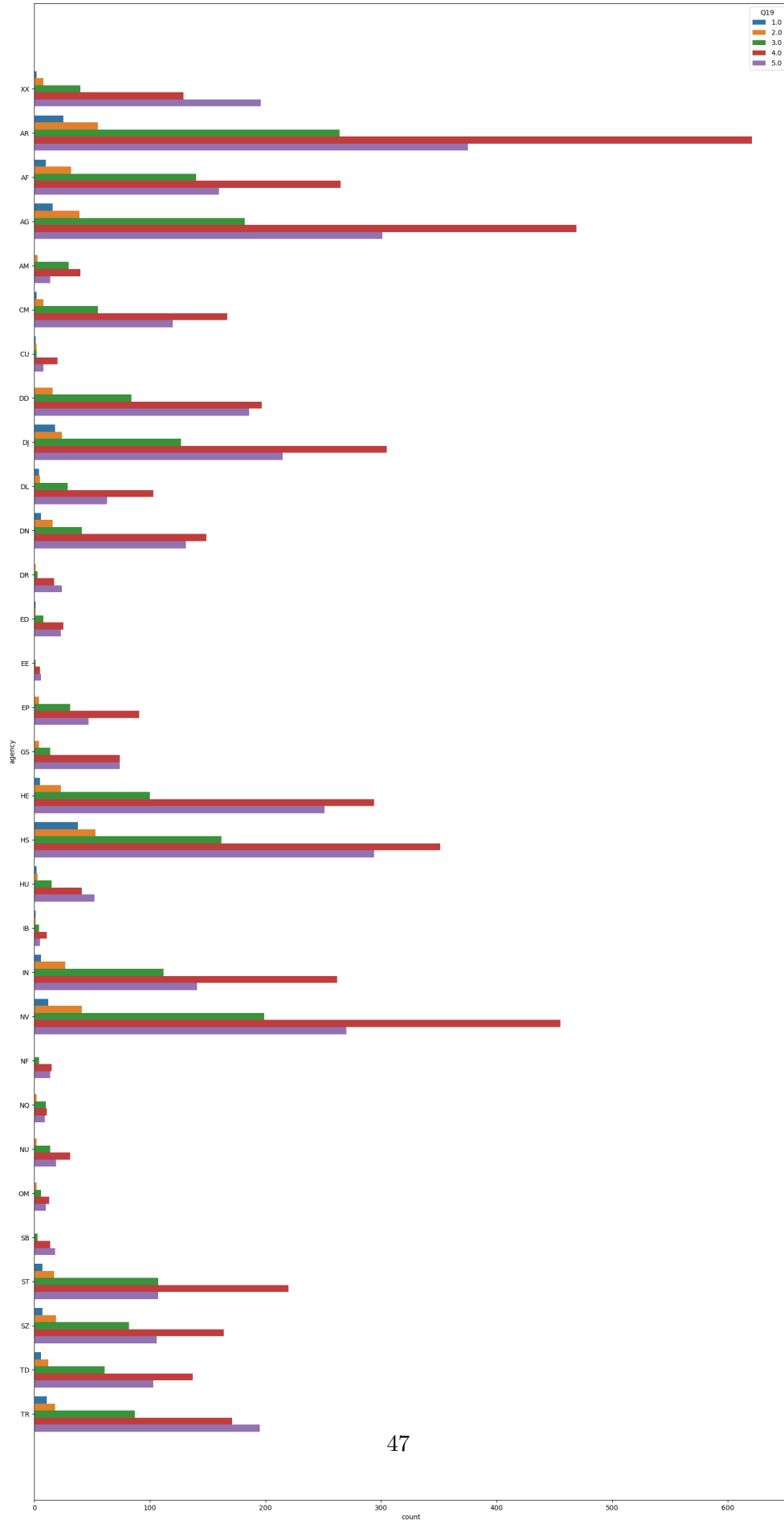


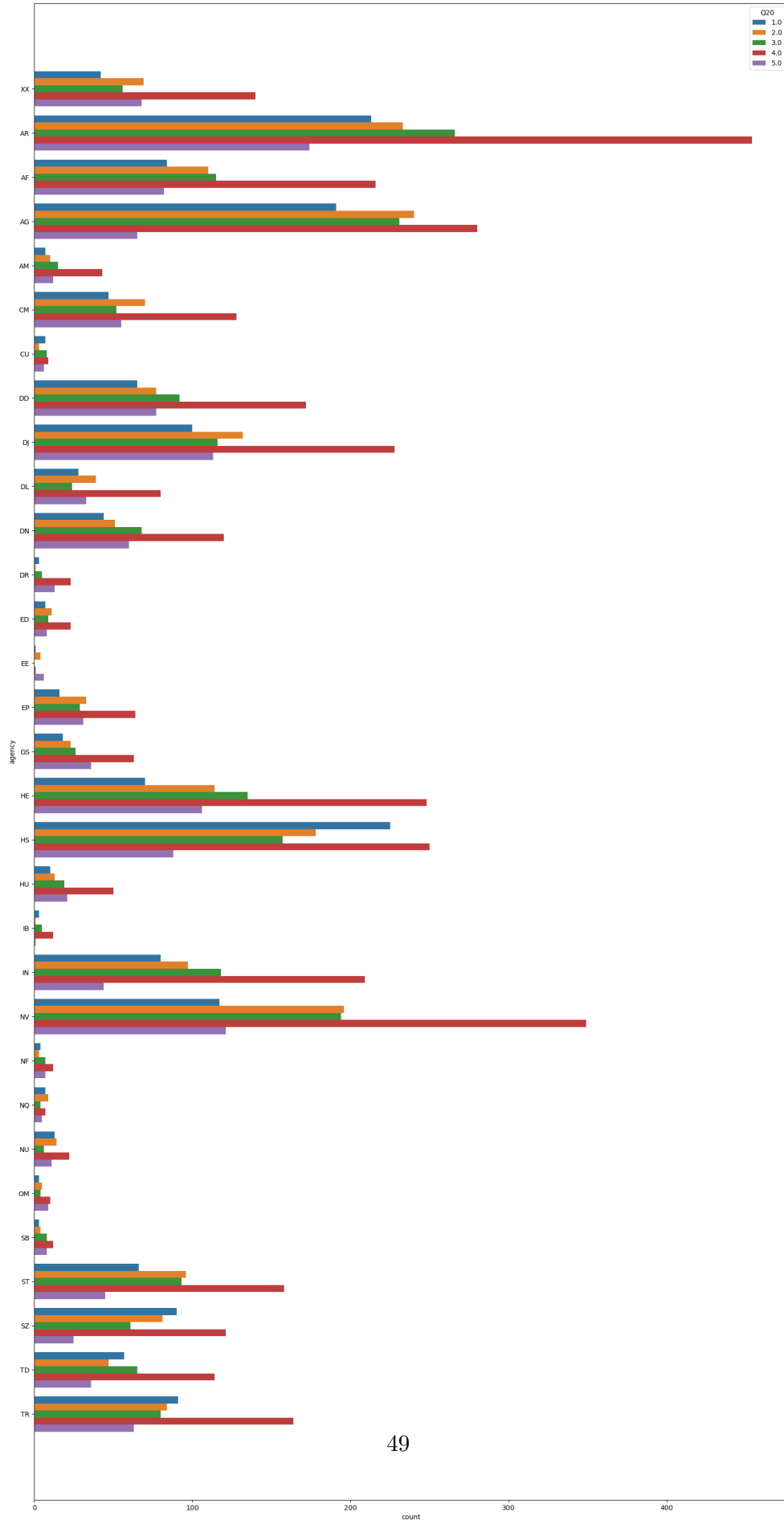


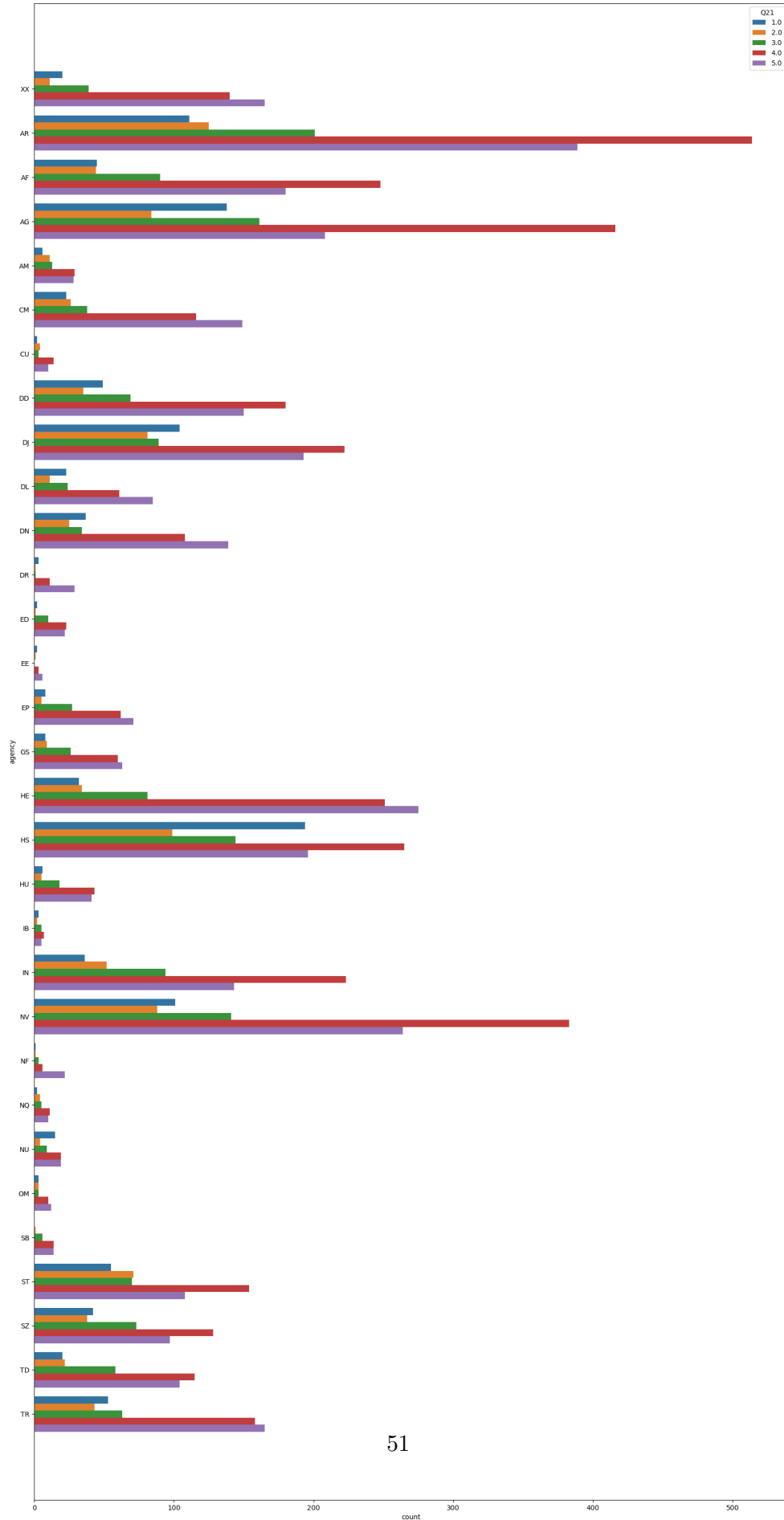


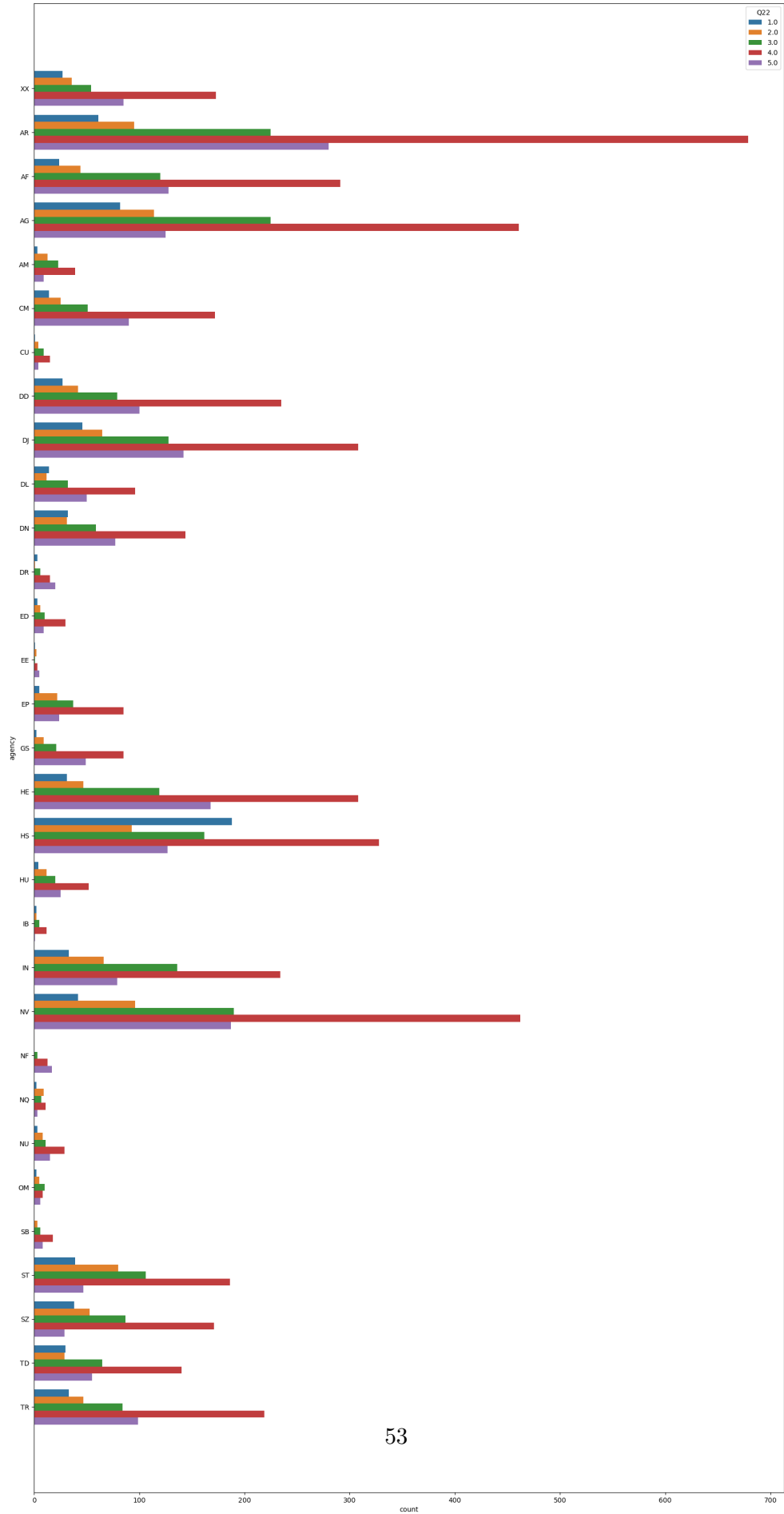


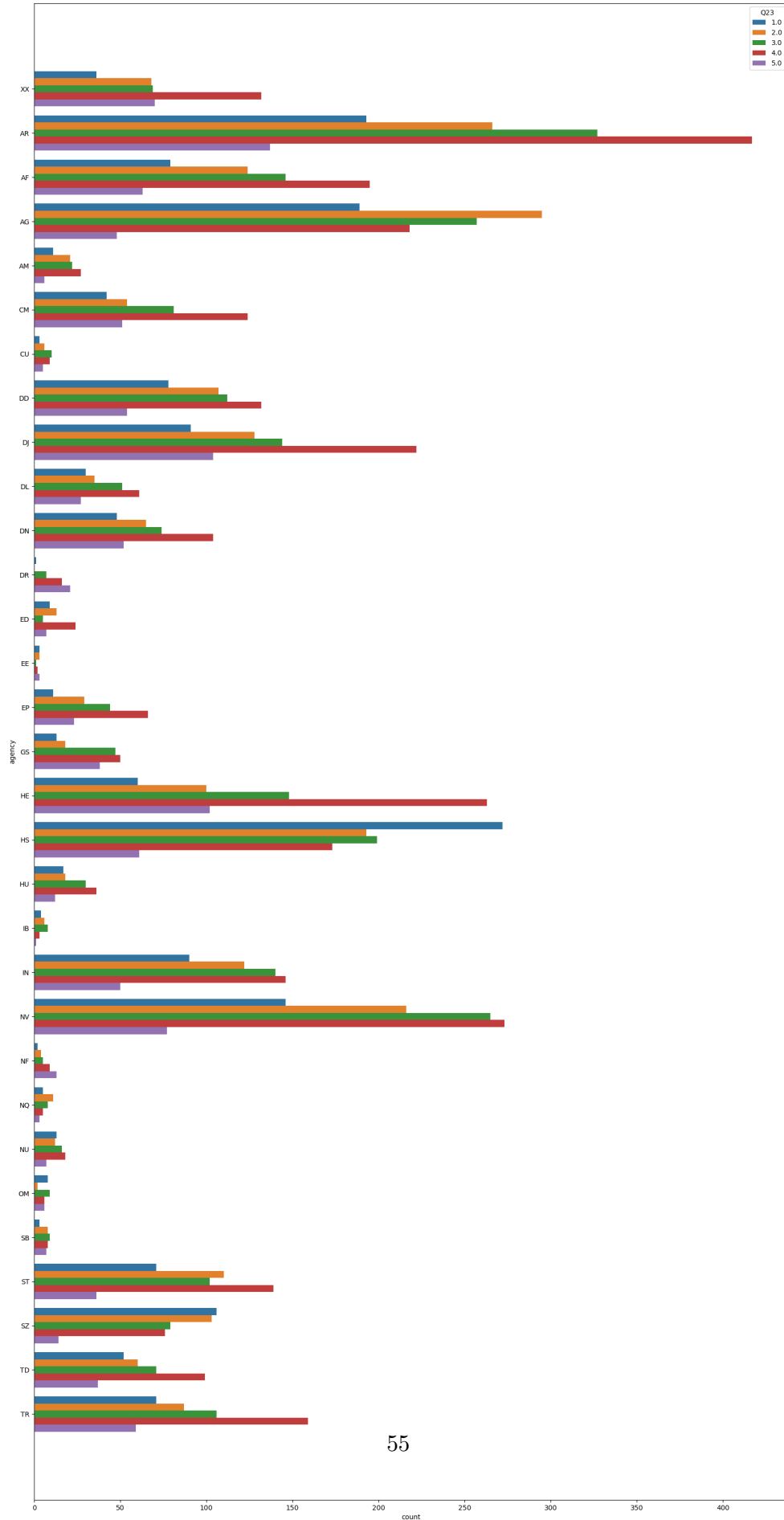


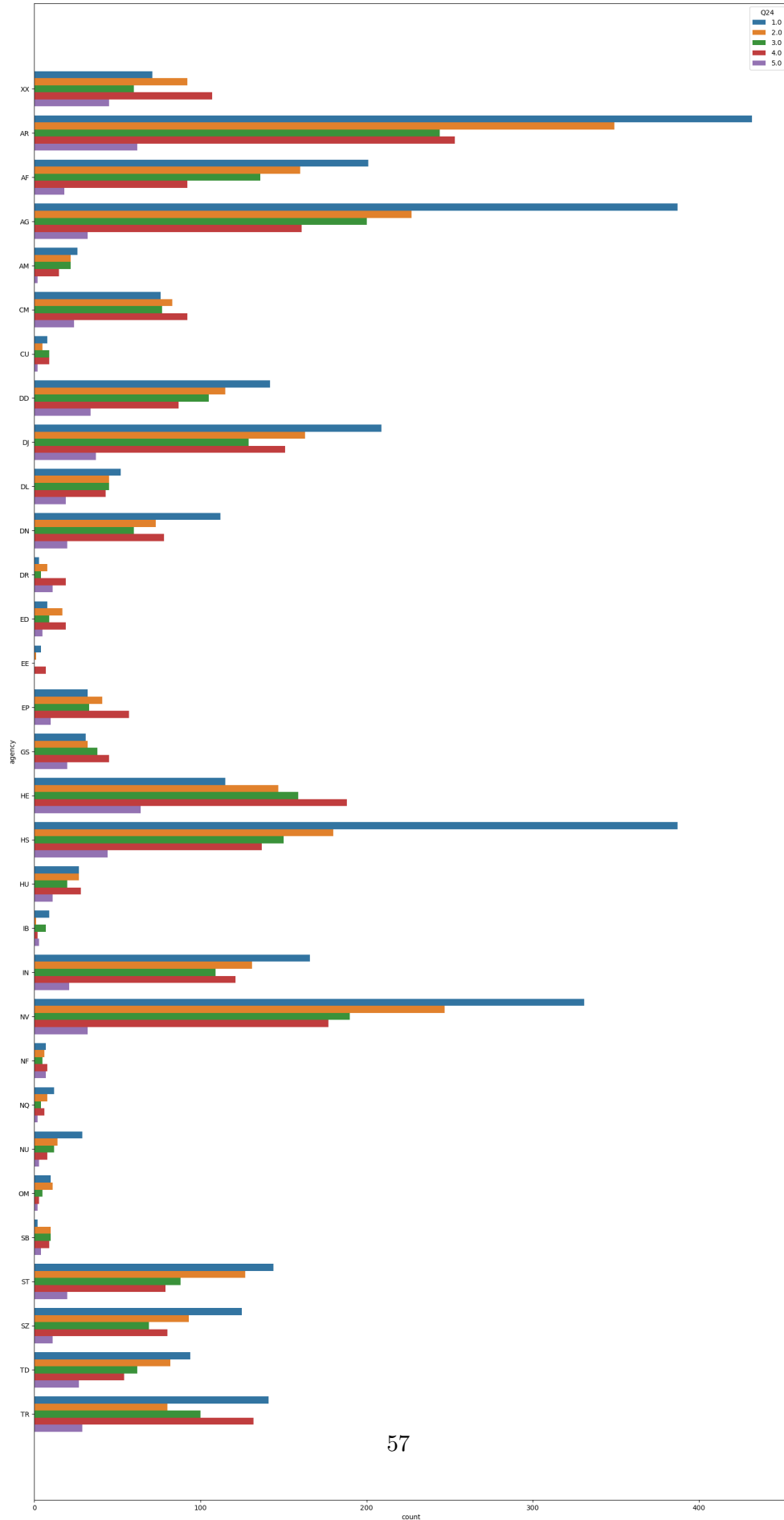


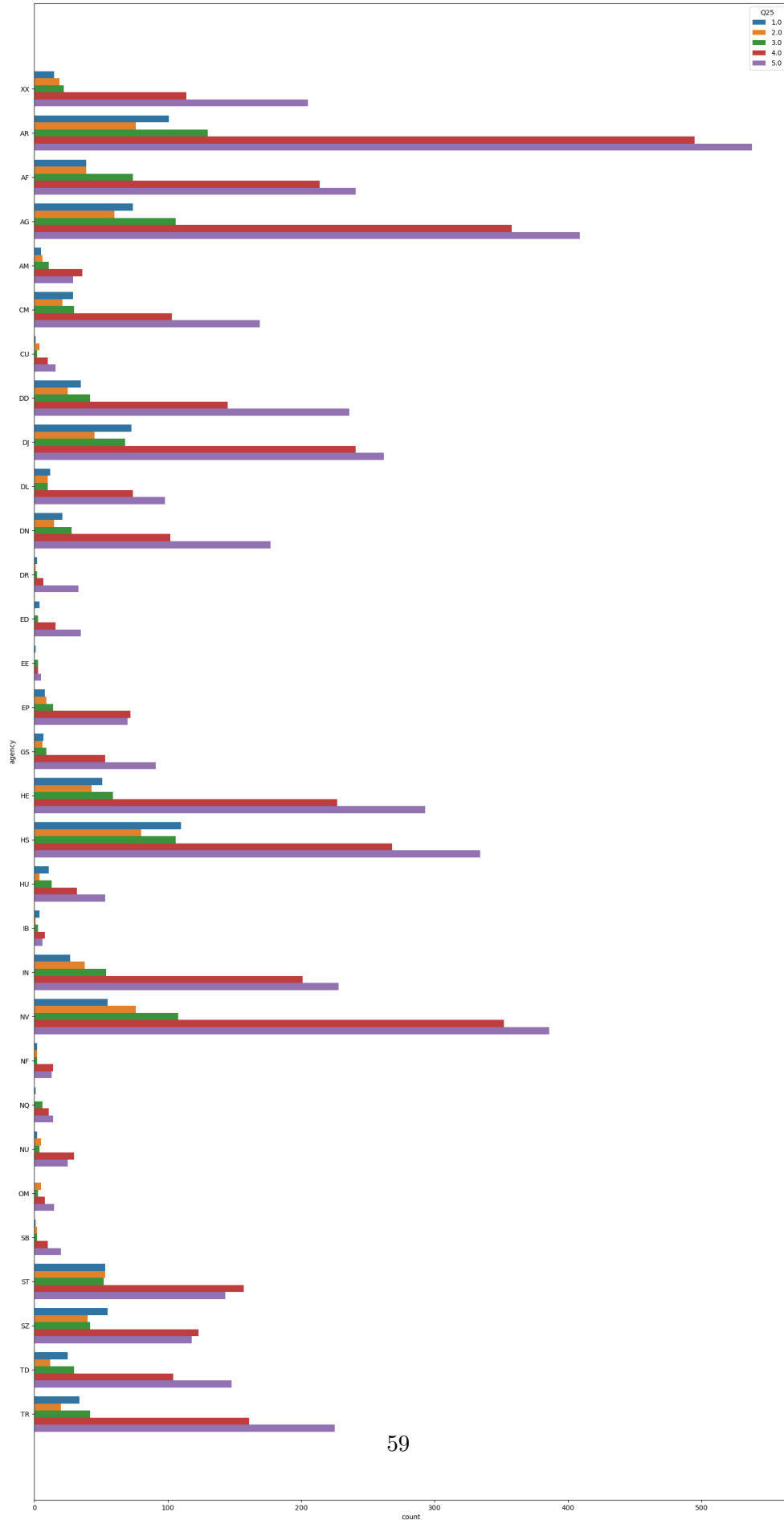


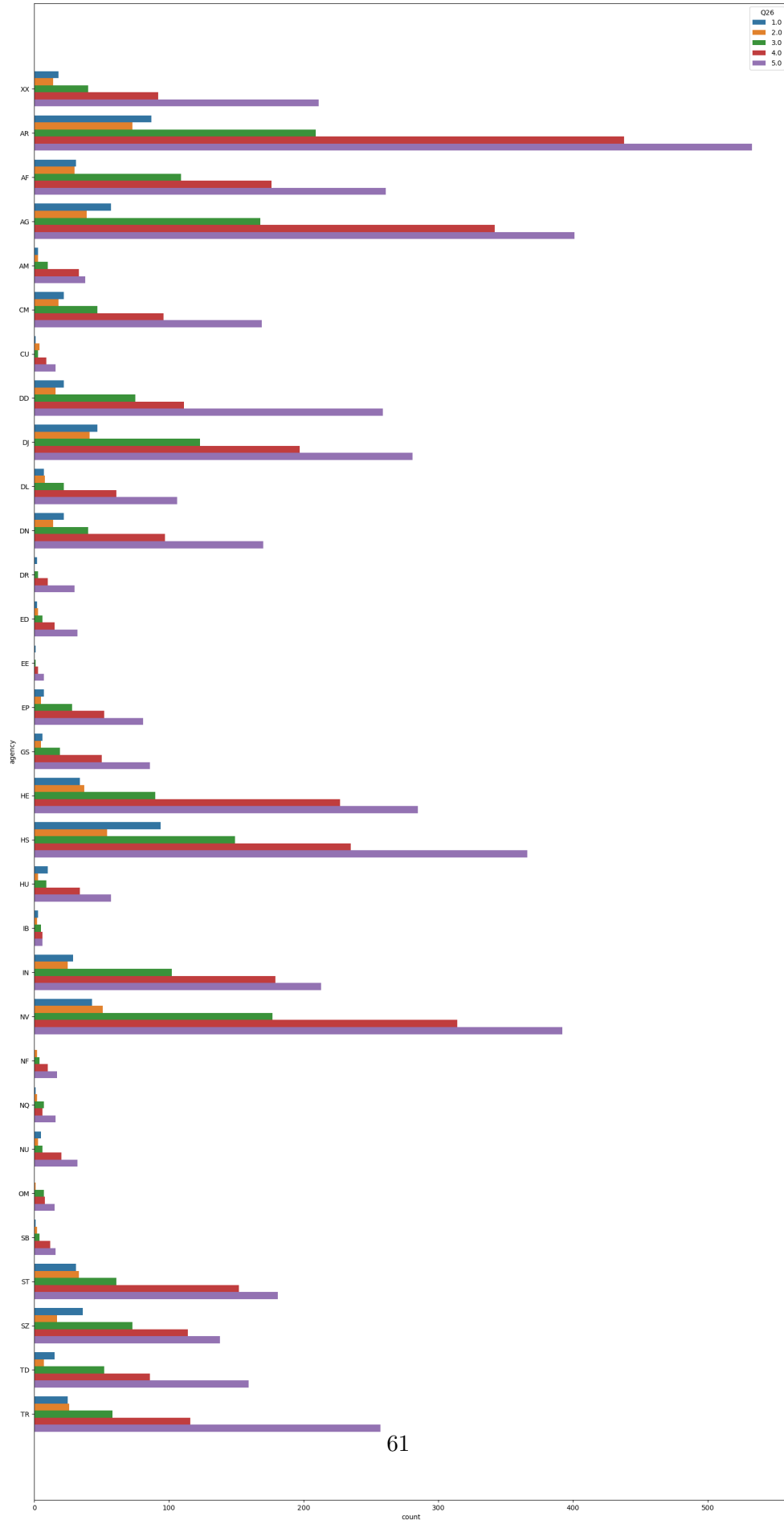


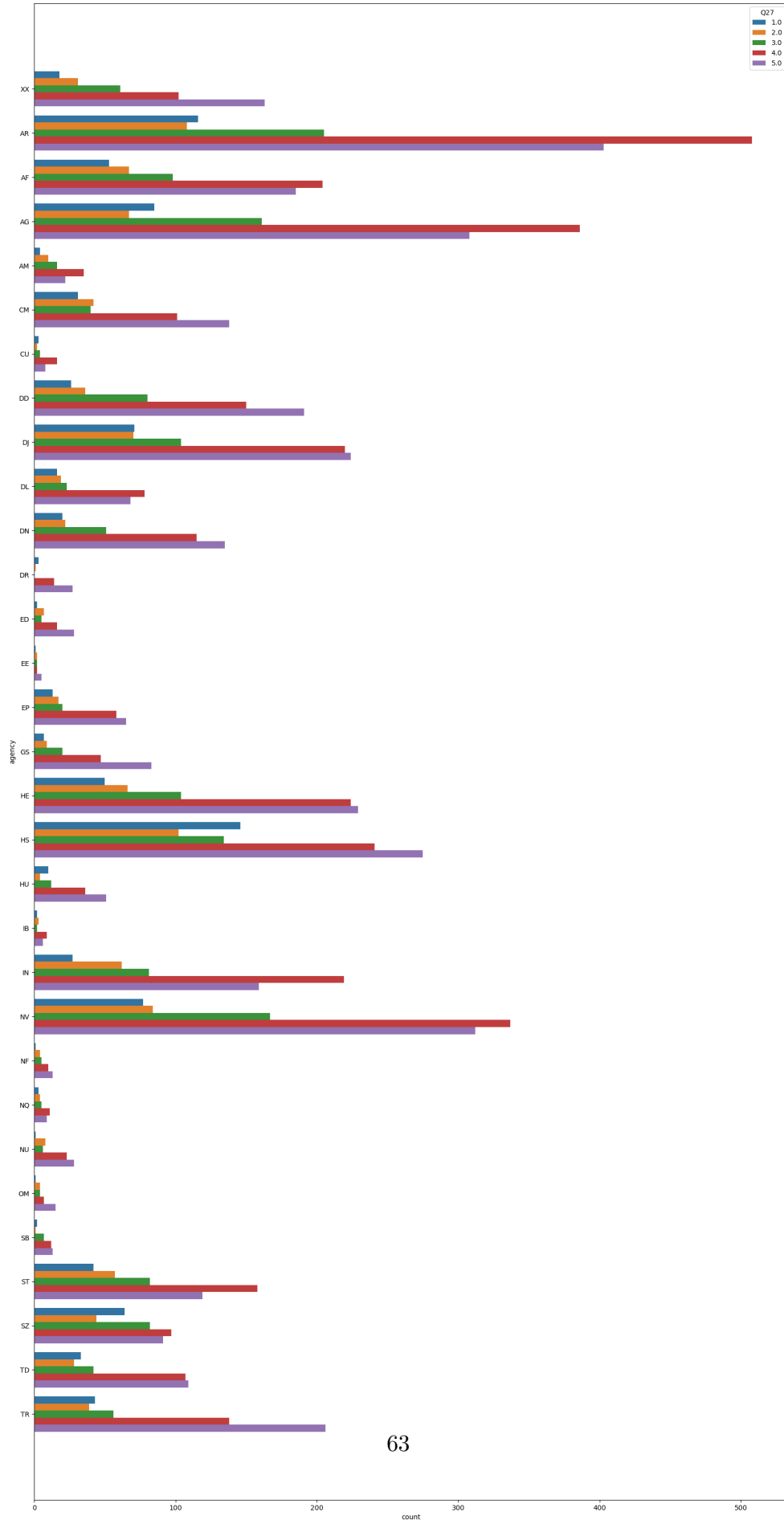


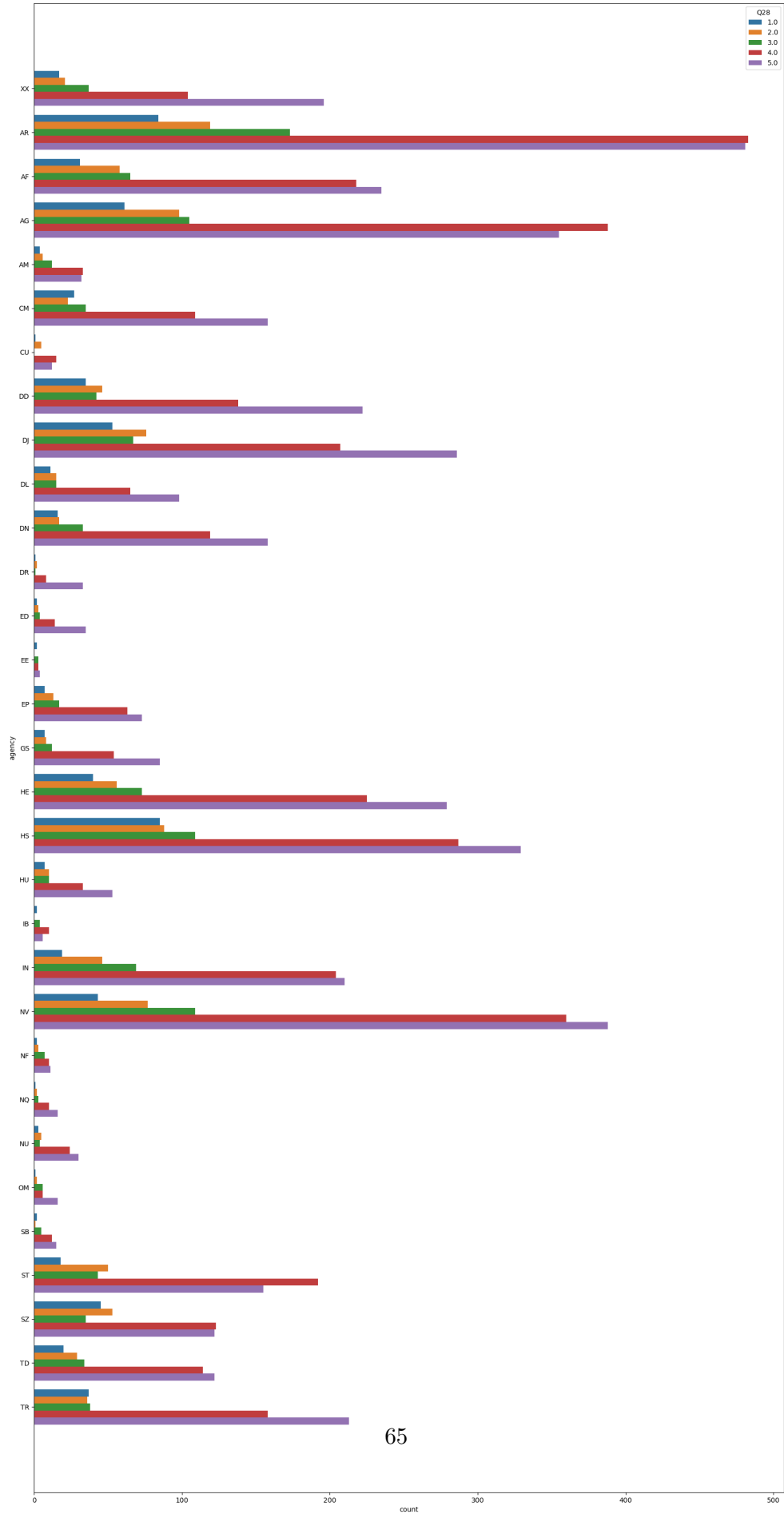


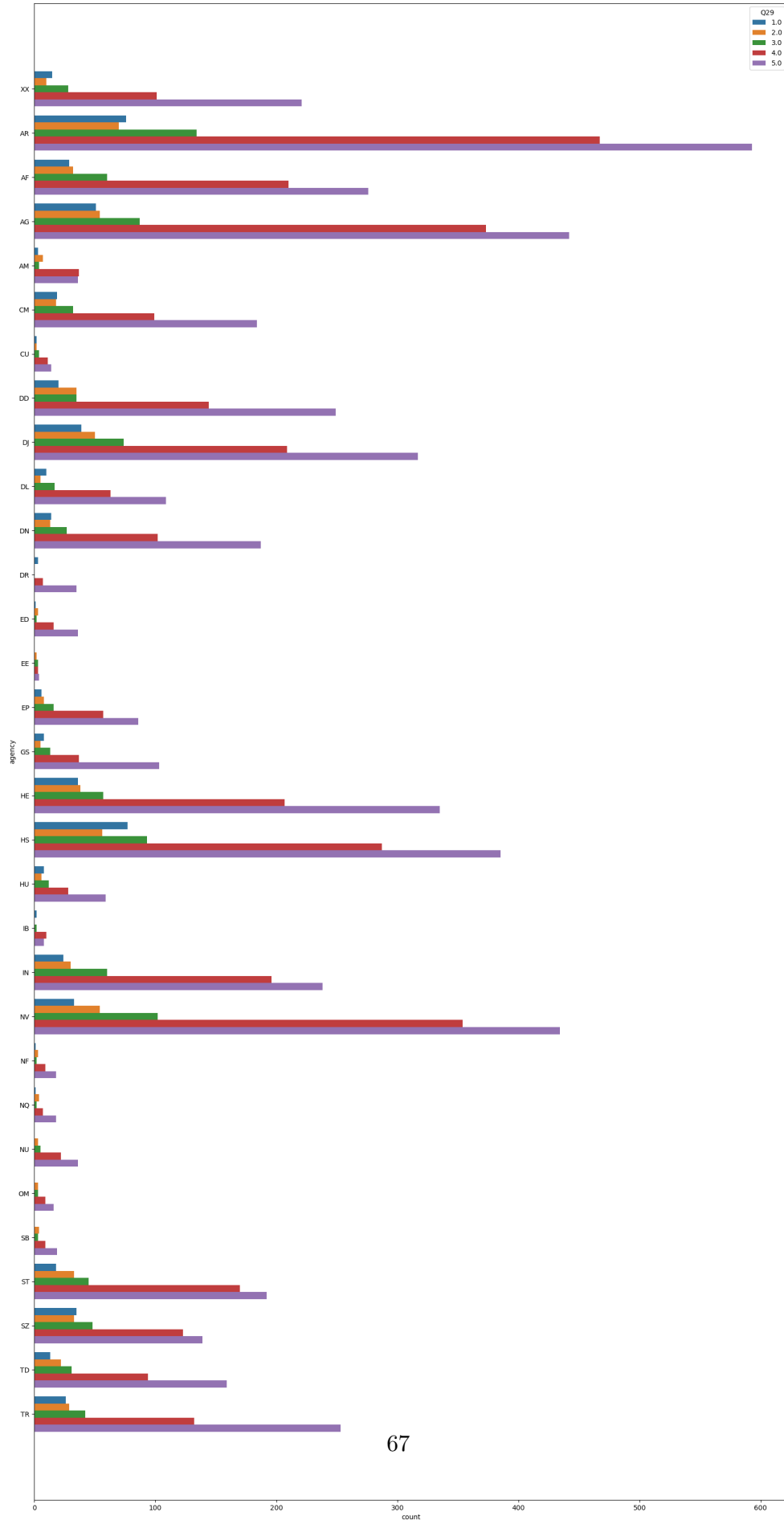


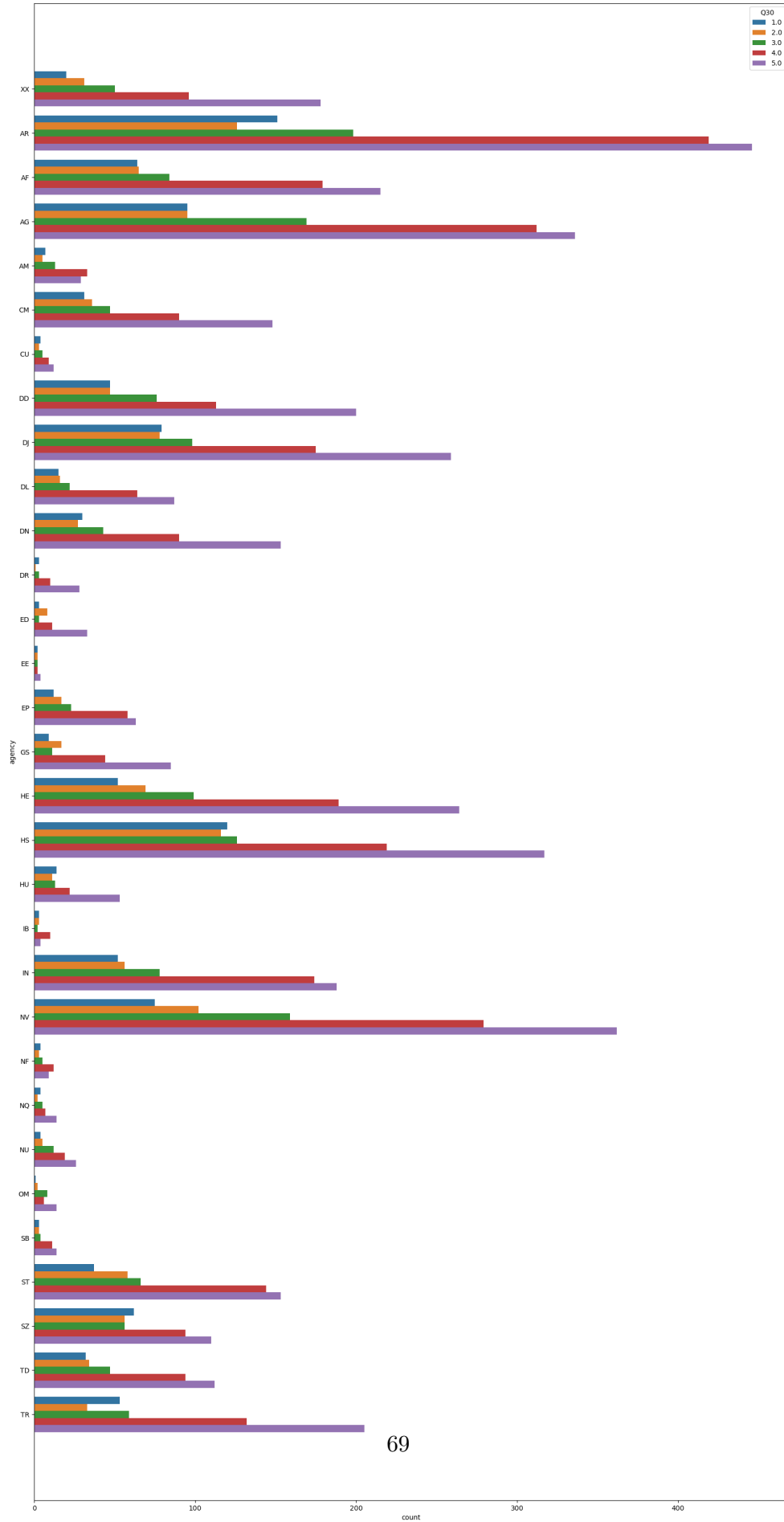


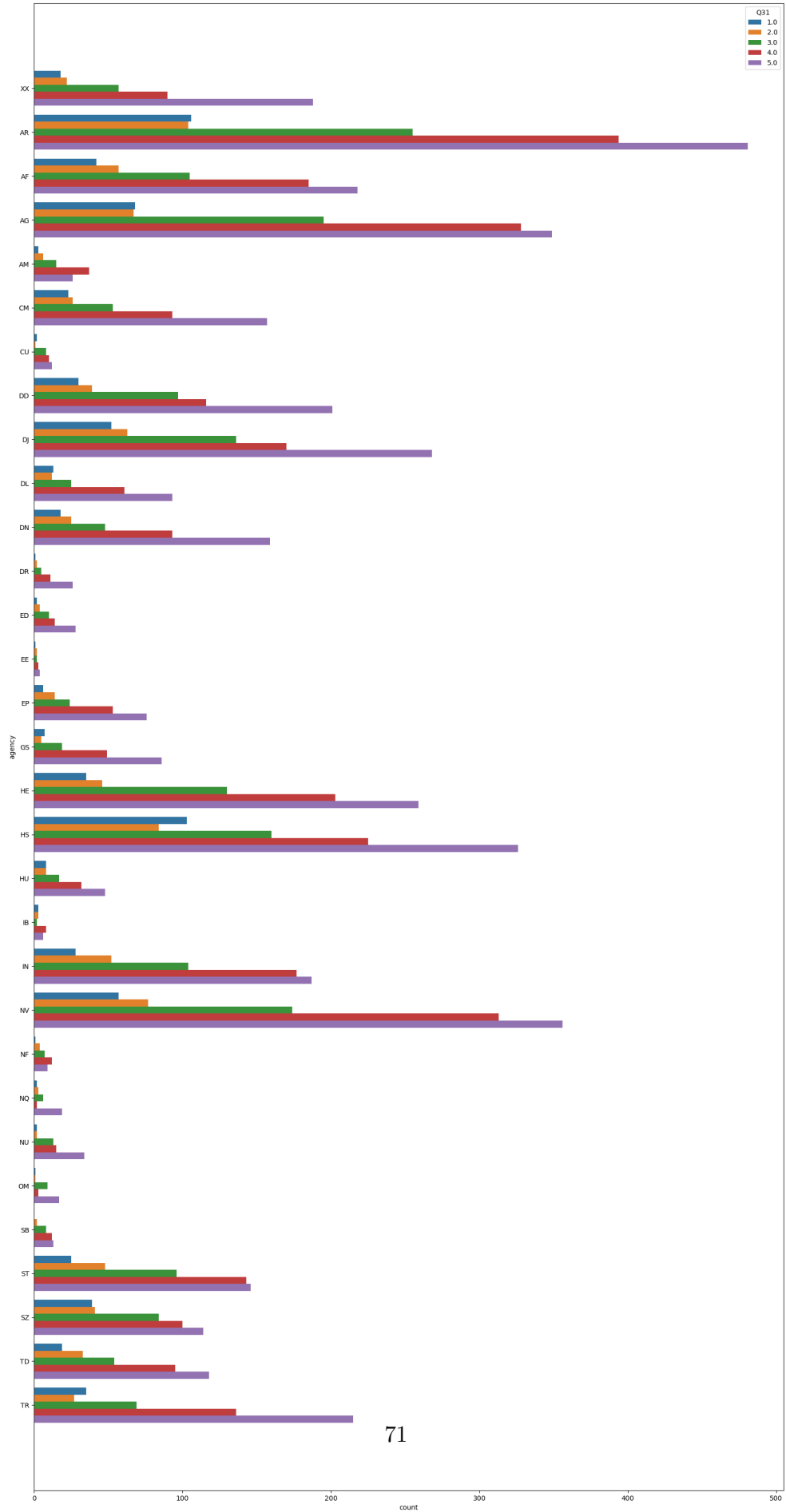


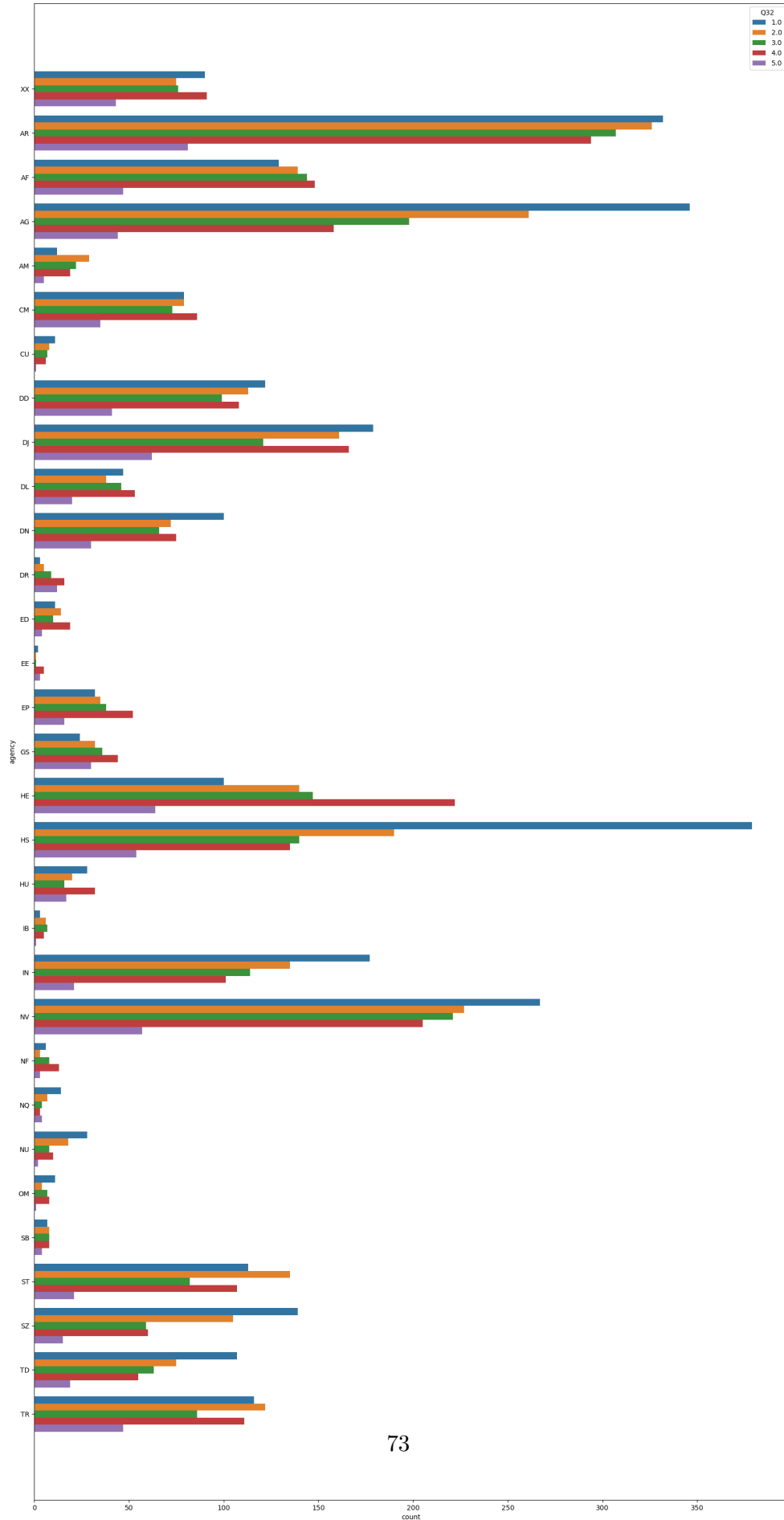


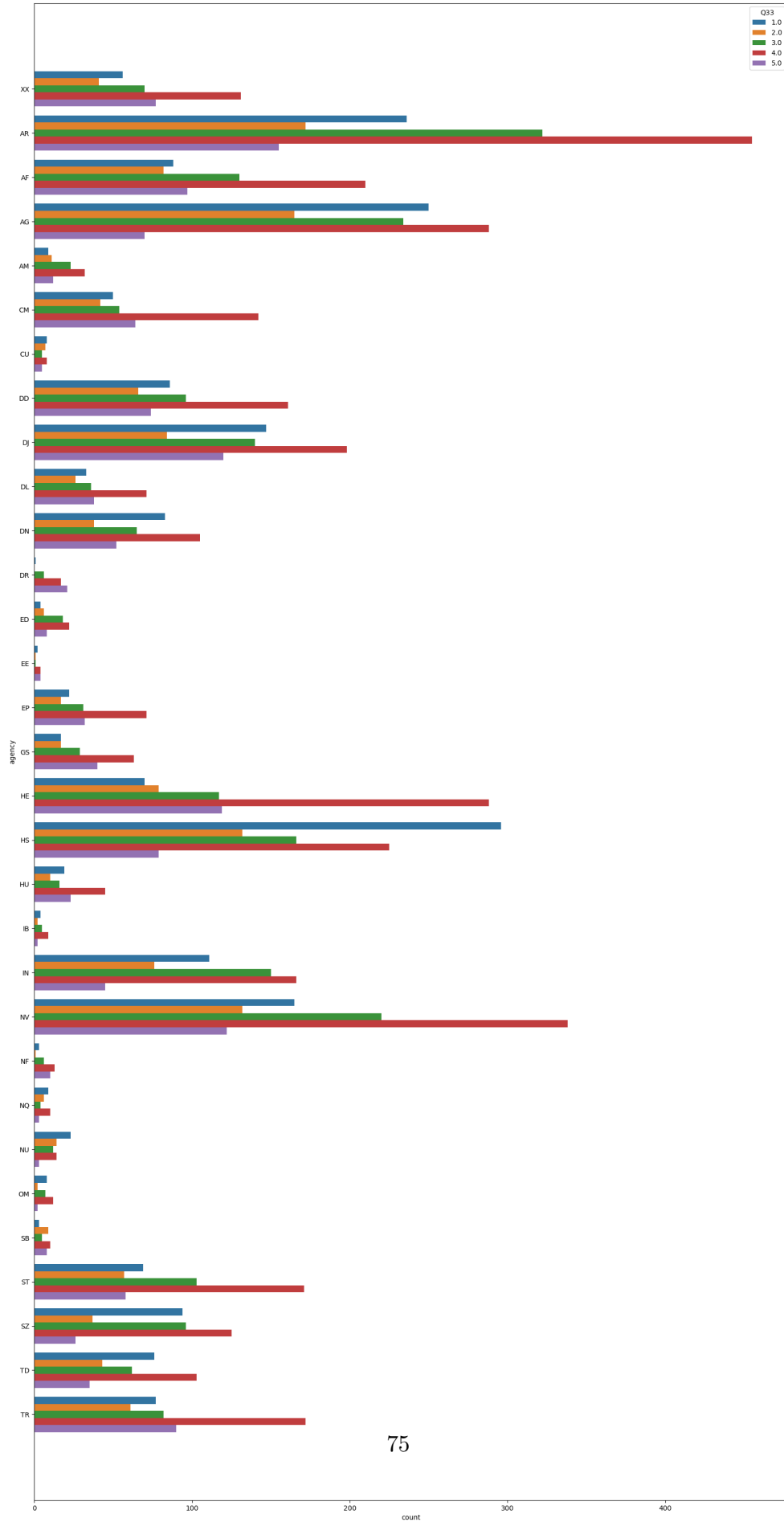


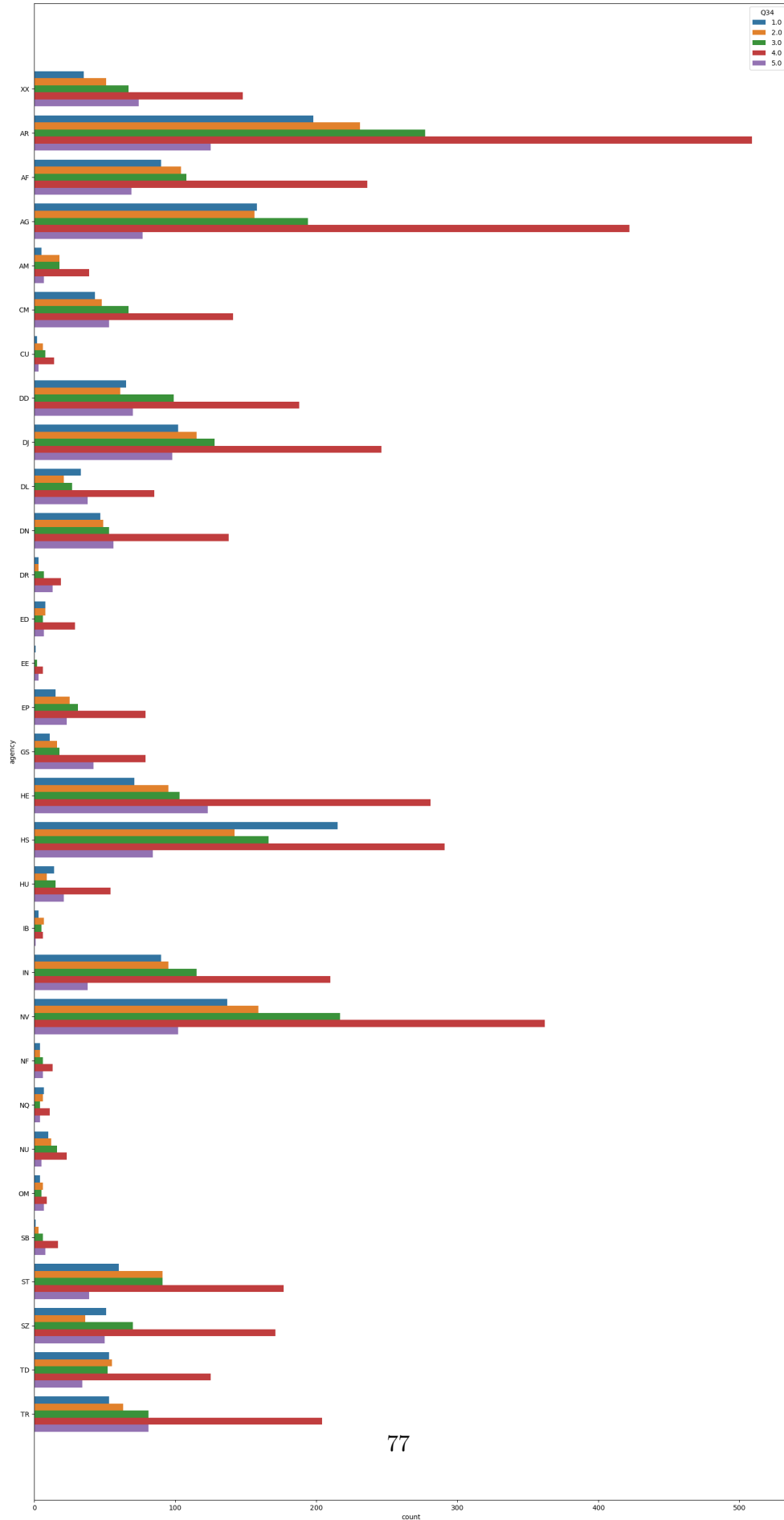


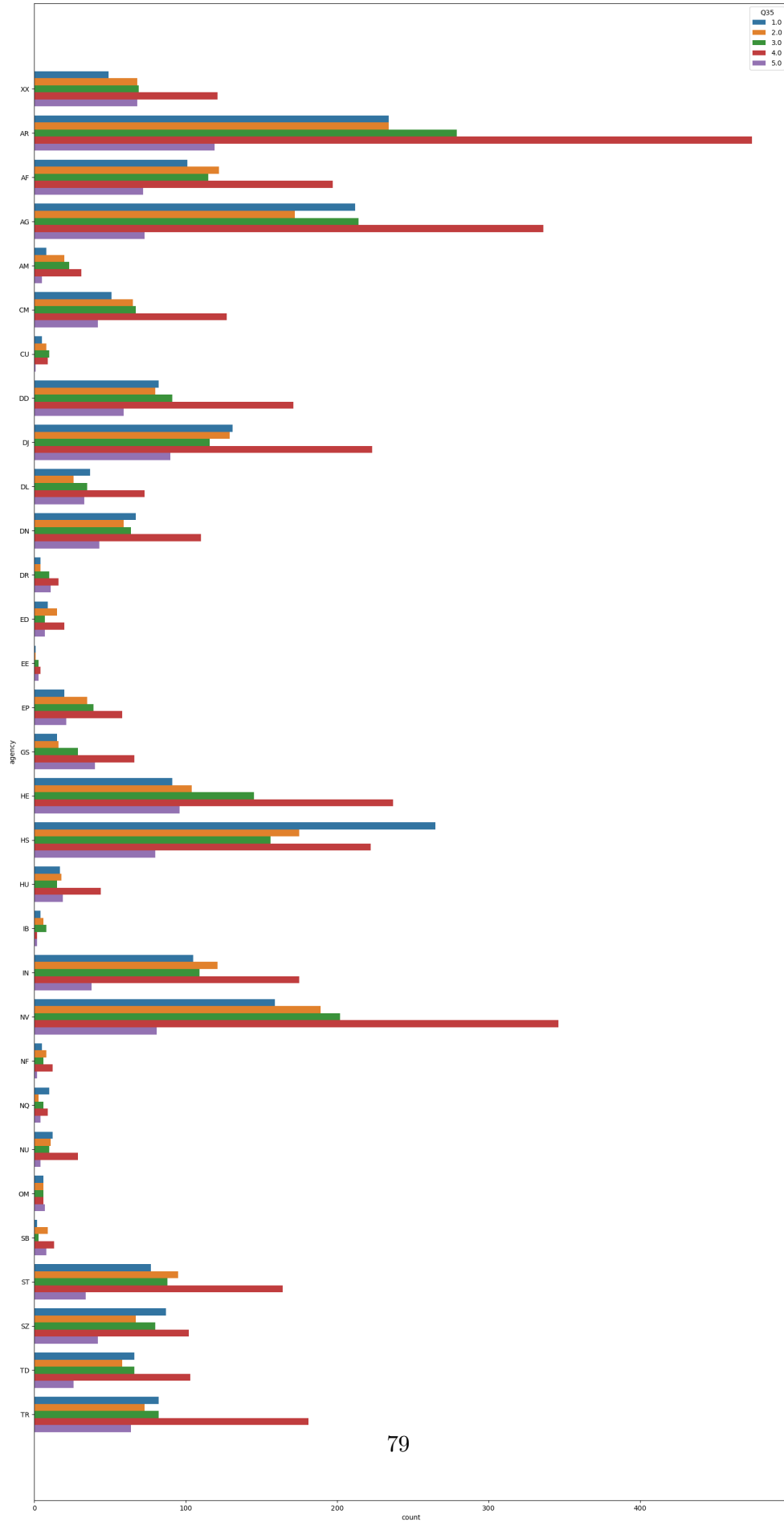


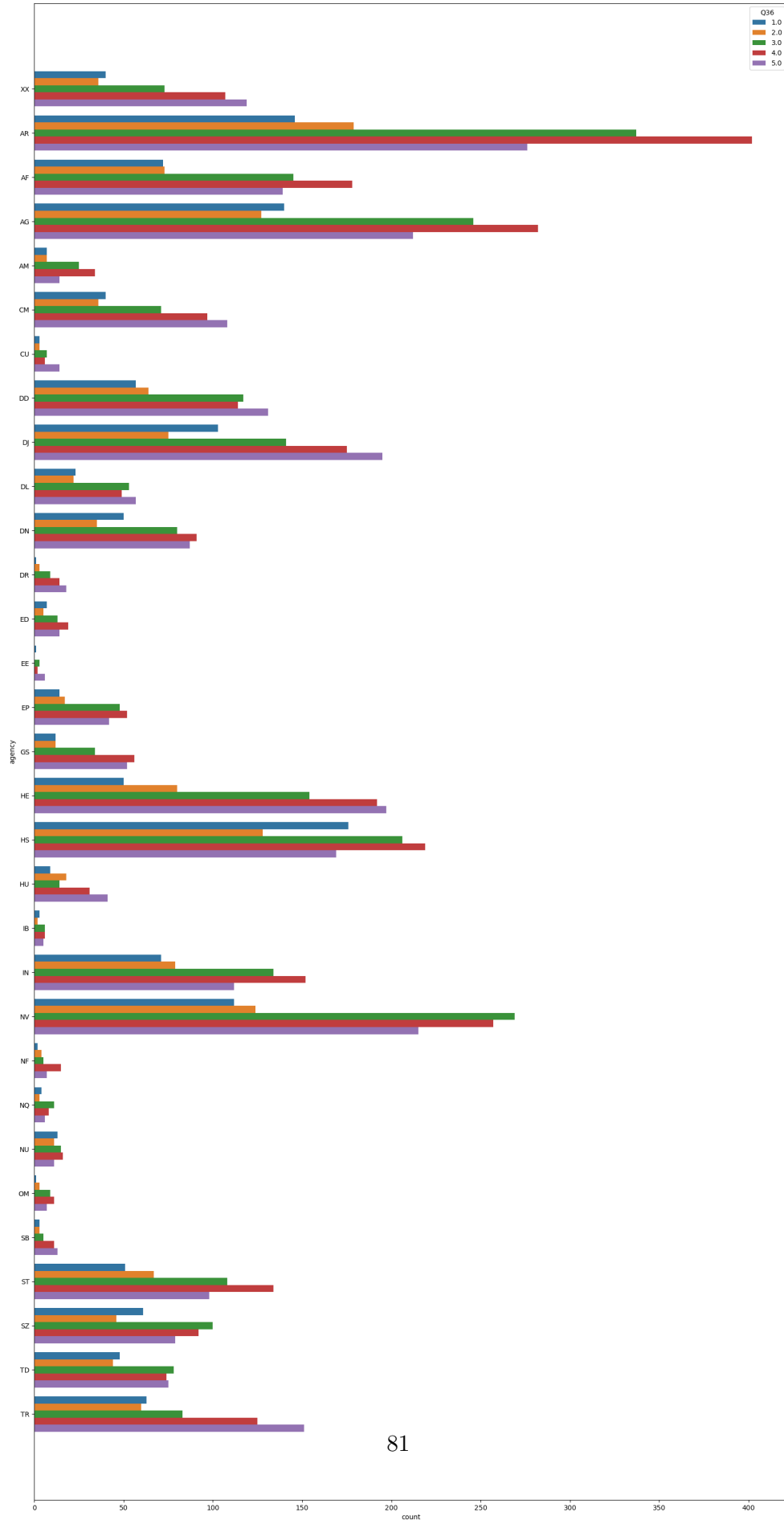


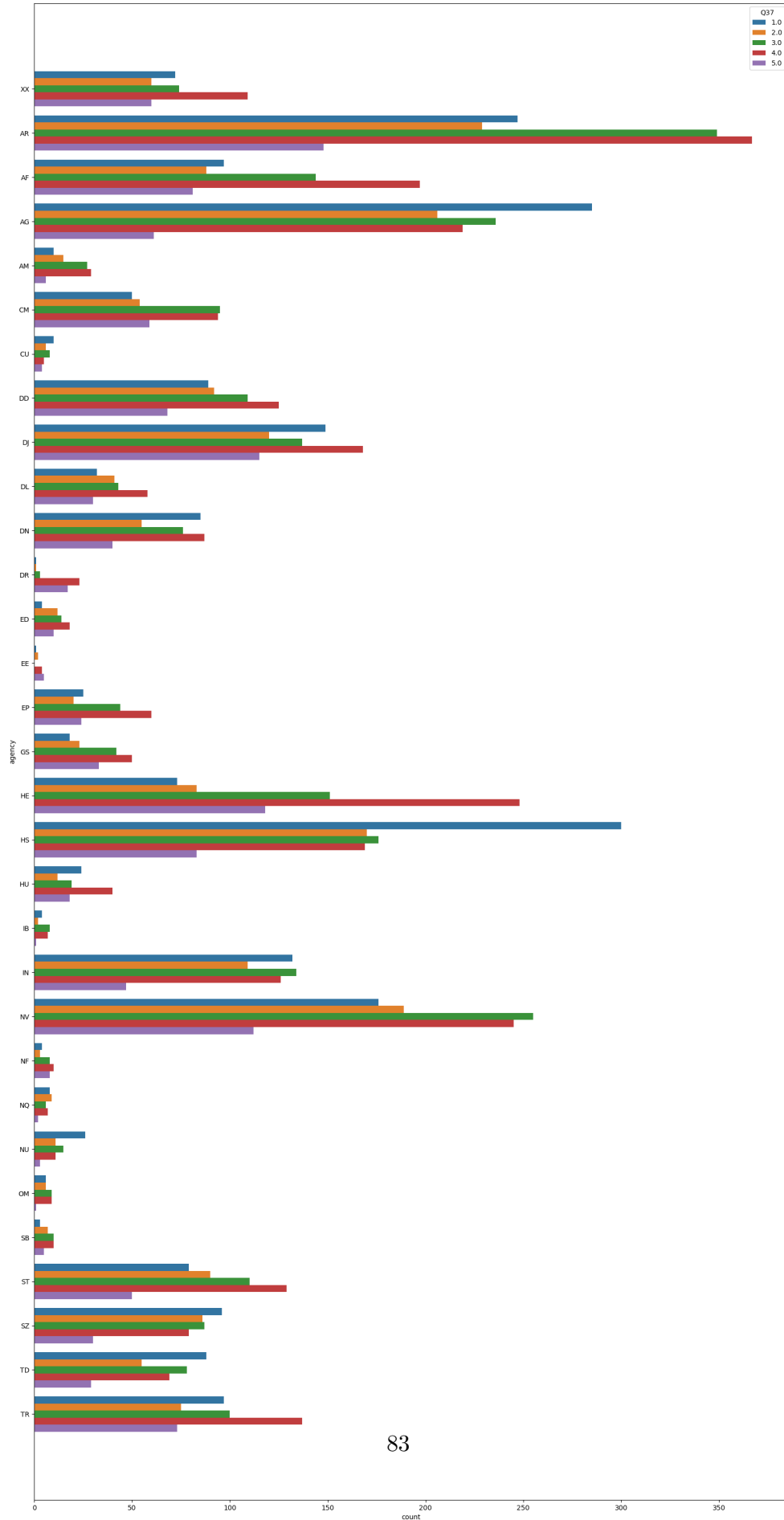


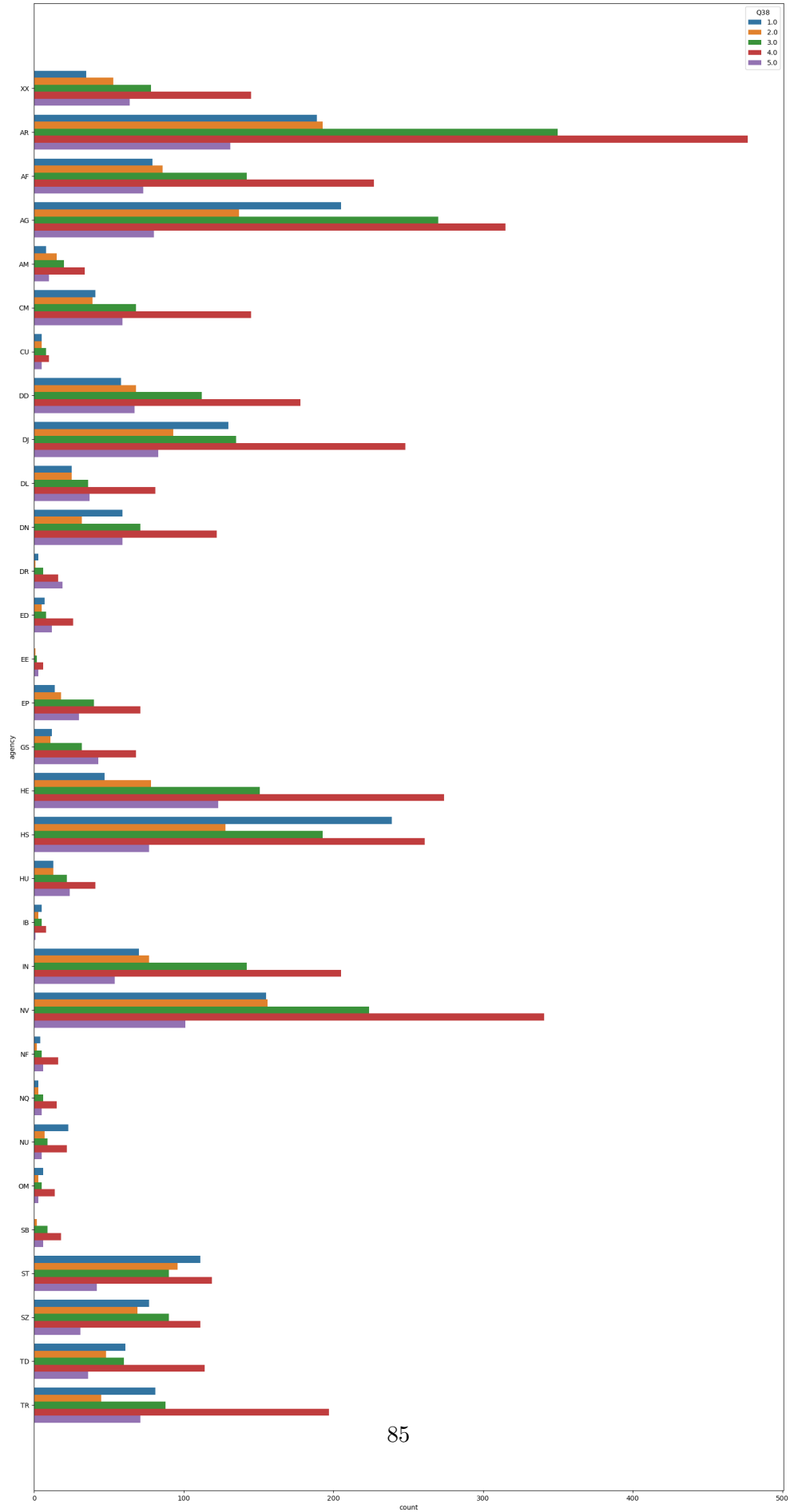


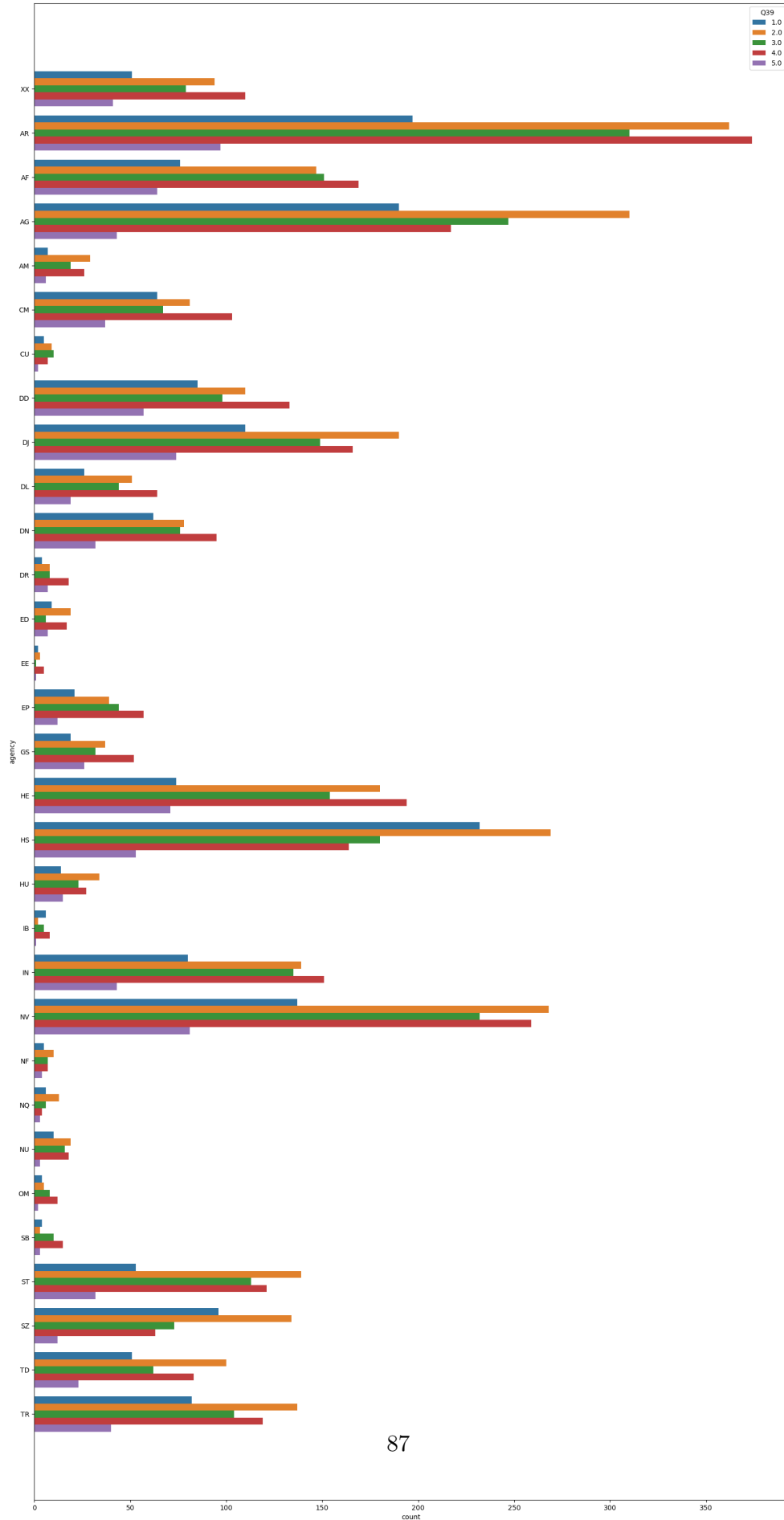


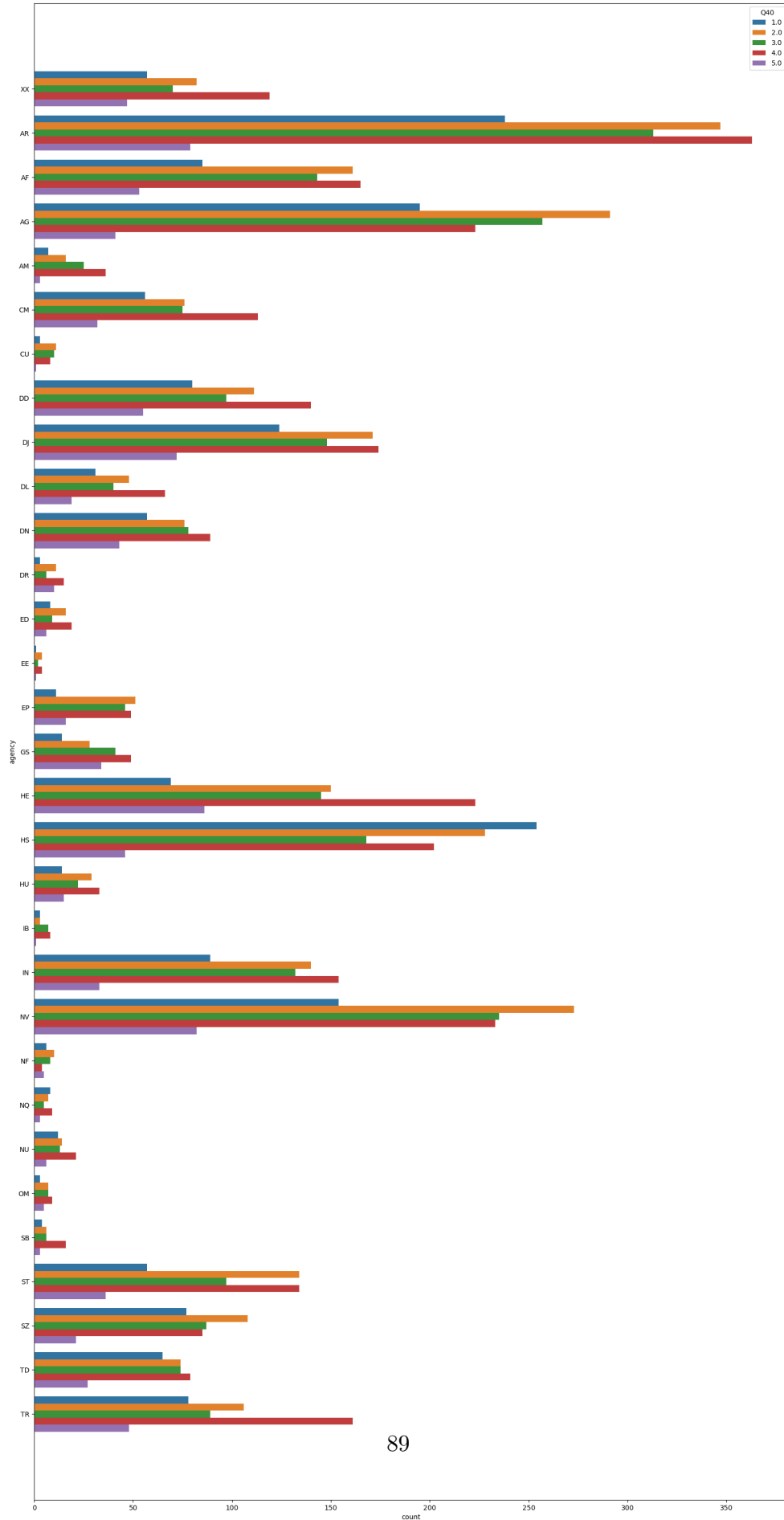


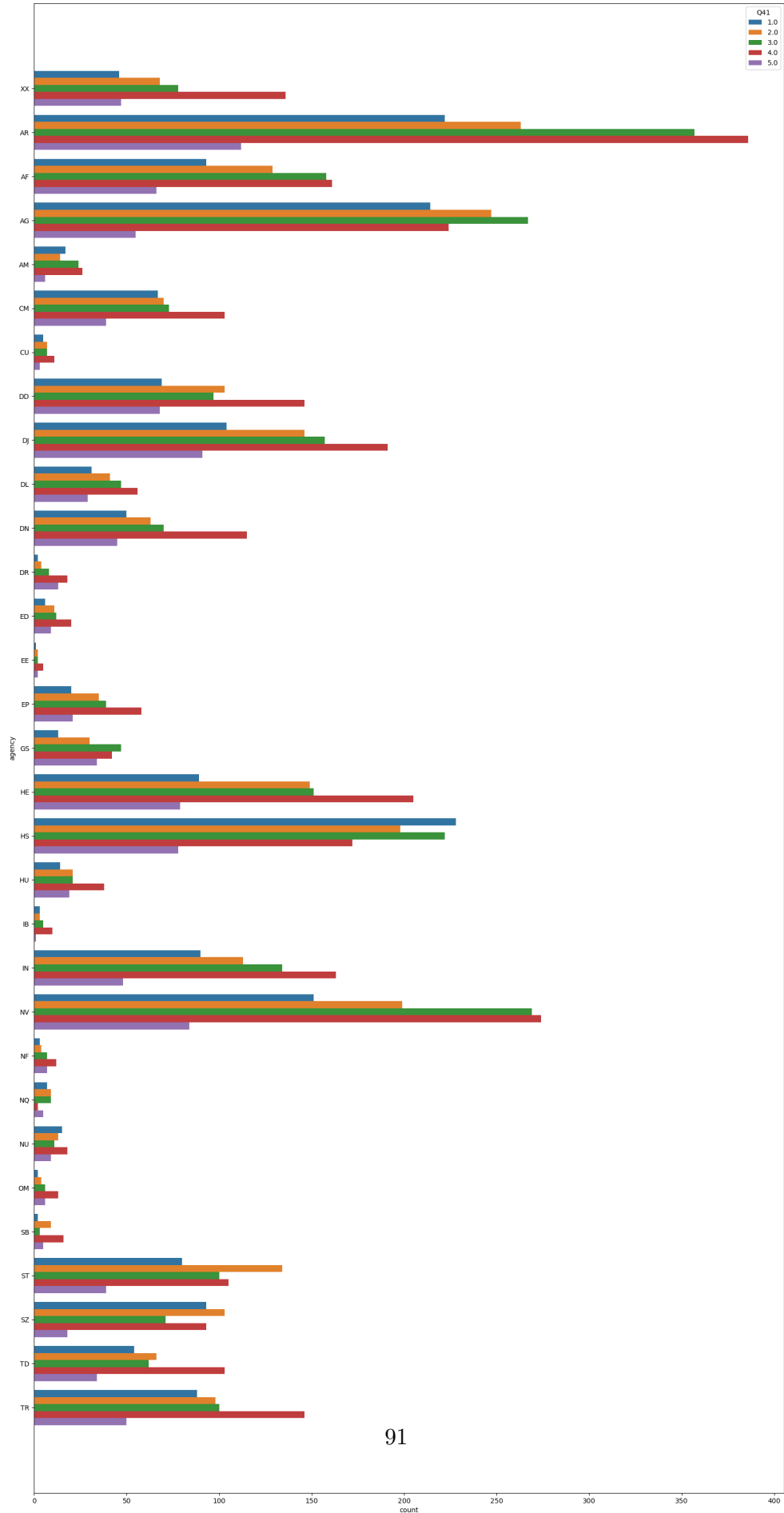


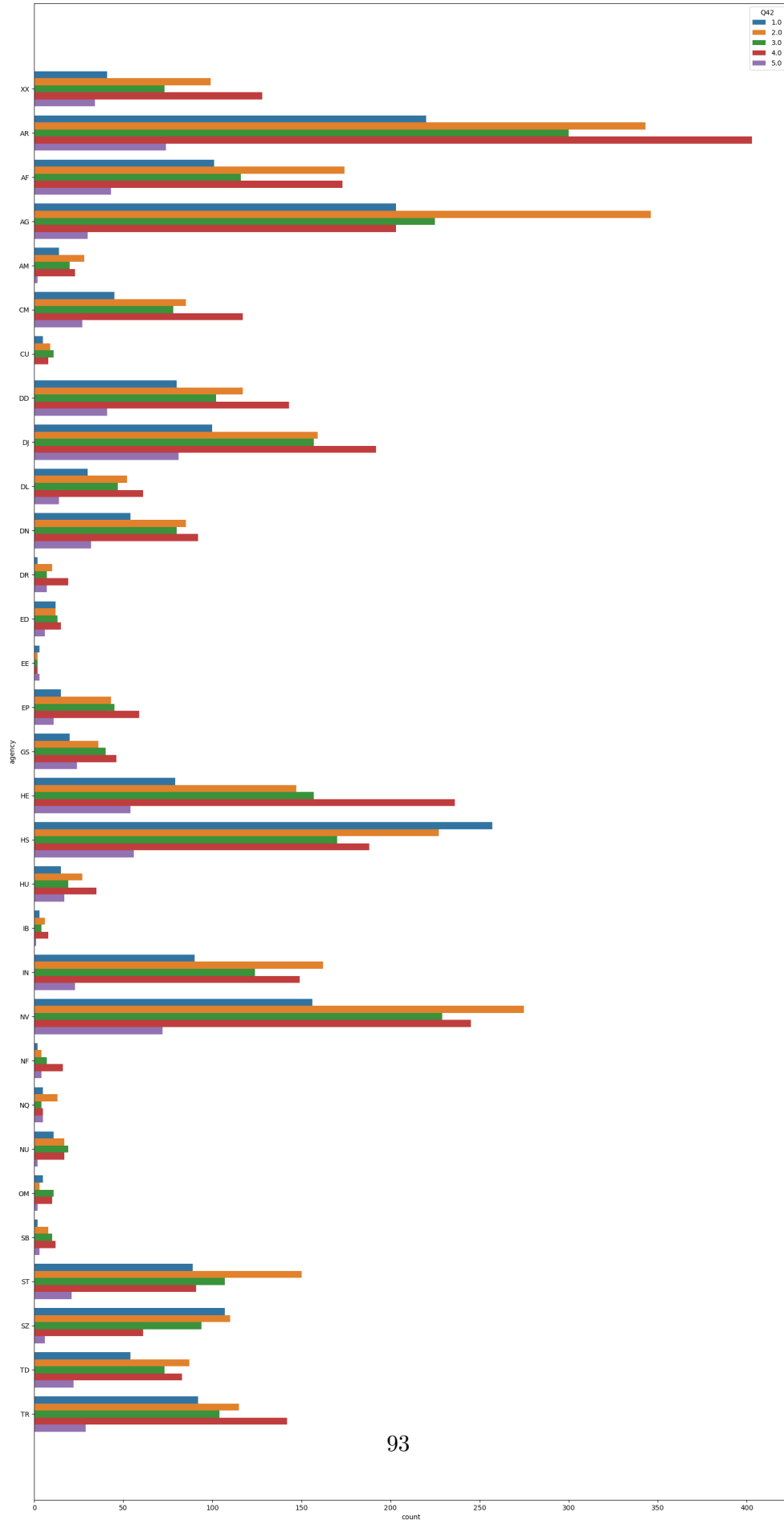


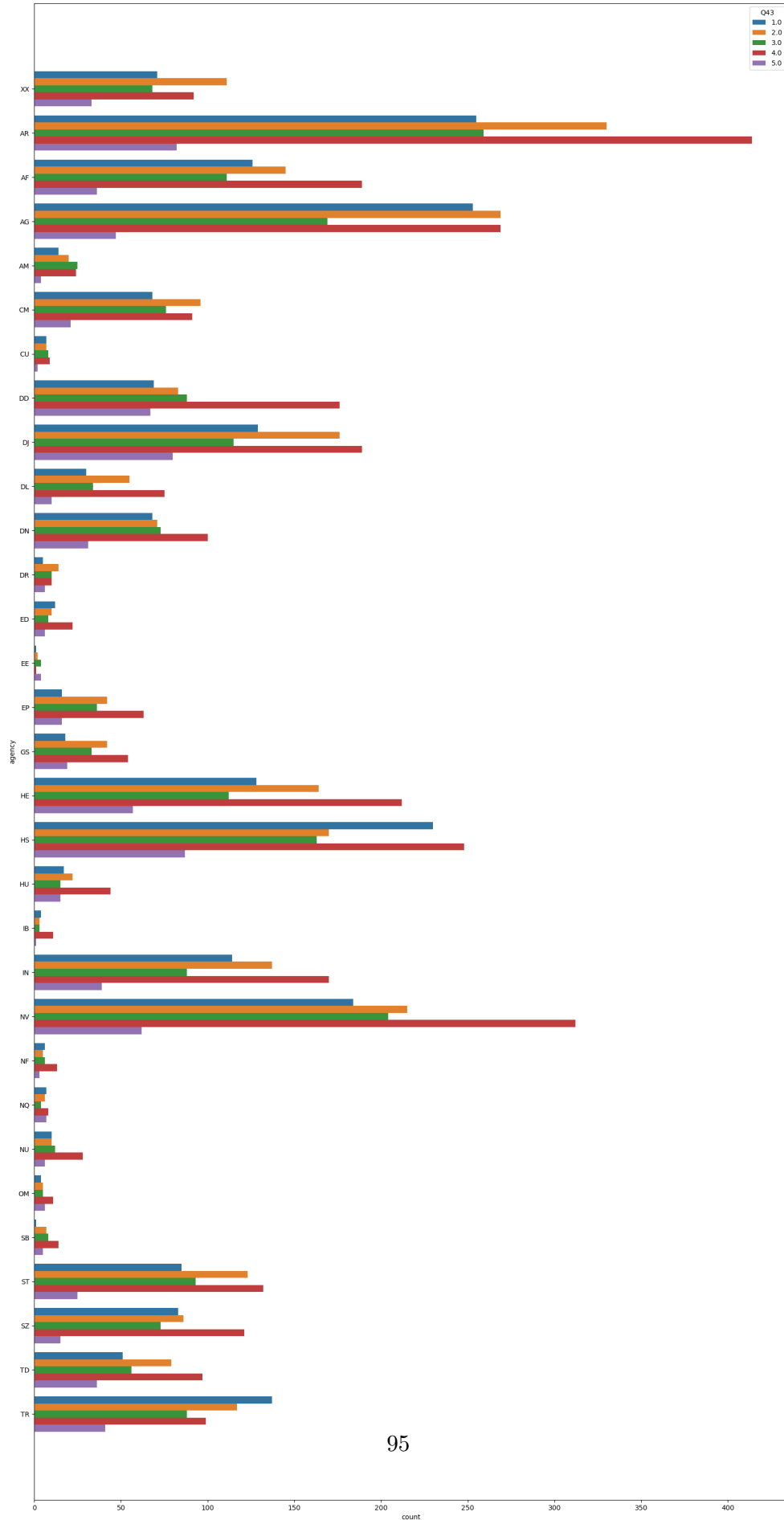












<Figure size 2000x4000 with 0 Axes>

[]:

[]: