In [29]:
```python
# Import packages
import pandas as pd
from scipy.stats import chi2_contingency

# Import data
data = pd.read_csv('candidates_original_preprocessed.csv')
```

Chi-squared tests for Hired vs. Not hired

Chi-squared tests to test whether sex/age/region are related to being hired.

In [30]:
```python
# Debugging code removed for clarity
```

```
--- Analysis for Sex ---
Contingency Table:
 Candidate State  Hired  Other    All
Sex
Female             130   2694   2824
Male               338   9070   9408
All                468  11764  12232
Chi-Square Test Statistic: 6.029950636346644
p-value: 0.19692267929101923
Degrees of Freedom: 4
Conclusion: The variable 'Sex' and a person being hired or not, are likely to be i
ndependent (no significant association).

--- Analysis for Age Range ---
Contingency Table:
 Candidate State  Hired  Other    All
Age Range
20 - 25 years       30   3284   3314
26 - 30 years      168   5550   5718
31 - 35 years       90   1021   1111
36 - 40 years       49    384    433
40 - 45 years       33    200    233
< 20 years          45    877    922
> 45 years          53    448    501
All                468  11764  12232
Chi-Square Test Statistic: 342.81644326571313
p-value: 1.3200582257954743e-64
Degrees of Freedom: 14
Conclusion: The variable 'Age Range' and a person being hired or not, are likely t
o be dependent (there is a significant association).

--- Analysis for Region ---
Contingency Table:
 Candidate State        Hired  Other    All
Region
(FOREIGN)                 1     32     33
(ITALY)                   2     19     21
(OVERSEAS)                7    256    263
Abruzzo                  14    211    225
Aosta Valley              0     37     37
```

```
Basilicata                 7    245    252
Calabria                   6    161    167
Campania                  30    603    633
Emilia Romagna            38    533    571
Friuli Venezia Giulia      7     88     95
Lazio                     41    605    646
Liguria                    2    278    280
Lombardy                  84    967   1051
Marche                     7    146    153
Molise                     1     73     74
Piedmont                  96   3834   3930
Puglia                    58   2117   2175
Sardinia                   3    145    148
Sicily                    27    771    798
Trentino Alto Adige        0     51     51
Tuscany                   31    329    360
Umbria                     1     64     65
Veneto                     5    195    200
All                      468  11760  12228
Chi-Square Test Statistic: 151.67180144148736
p-value: 3.2888394346071096e-13
Degrees of Freedom: 46
Conclusion: The variable 'Region' and a person being hired or not, are likely to b
e dependent (there is a significant association).
```

For 'Region', the #observations is <5 in for some. This does not adhere with the rules of thumb of the Chi-squared test, this needs to be accounted for still.

Chi-squared tests for various candidates states

Candidate state vs. Sex

In [26]:
```python
# Order of selection
selection_order = ['Imported', 'In selection', 'First contact', 'QM', 'Vi

# Initialize a dictionary
contingency_tables_sex = {}

# Loop through each state in the process order
for i, state in enumerate(selection_order):
    post_states = selection_order[i+1:]

    if post_states:   # Skips hired as it has no states after
        contingency_table_sex = pd.DataFrame({
            f'Post {state}': data[data['Candidate State'].isin(post_state
            state: data[data['Candidate State'] == state].groupby('Sex').
        })
        contingency_tables_sex[state] = contingency_table_sex
```

In [20]:
```python
# Debugging code removed for clarity
```

```
--- Analysis for Imported ---
Chi-Square Test Statistic: 63.66134178926101
p-value: 1.477558472913691e-15
Degrees of Freedom: 1
Conclusion: For 'Imported', the variables 'Sex' and 'Candidate State' are dependen
t (significant association).

--- Analysis for In selection ---
Chi-Square Test Statistic: 4.409181101166797
p-value: 0.03574599691749415
Degrees of Freedom: 1
Conclusion: For 'In selection', the variables 'Sex' and 'Candidate State' are depe
ndent (significant association).

--- Analysis for First contact ---
Chi-Square Test Statistic: 4.834975066237263
p-value: 0.02788804410371374
Degrees of Freedom: 1
Conclusion: For 'First contact', the variables 'Sex' and 'Candidate State' are dep
endent (significant association).

--- Analysis for QM ---
Chi-Square Test Statistic: 10.565197261191749
p-value: 0.0011523673663921489
Degrees of Freedom: 1
Conclusion: For 'QM', the variables 'Sex' and 'Candidate State' are dependent (sig
nificant association).

--- Analysis for Vivier ---
Chi-Square Test Statistic: 0.8770501768346592
p-value: 0.3490108870009826
Degrees of Freedom: 1
Conclusion: For 'Vivier', the variables 'Sex' and 'Candidate State' are independen
t (no significant association).

--- Analysis for Economic proposal ---
Chi-Square Test Statistic: 0.8325878370068877
p-value: 0.3615252948108346
Degrees of Freedom: 1
Conclusion: For 'Economic proposal', the variables 'Sex' and 'Candidate State' are
independent (no significant association).
```

Candidate state vs. Age Range

In [22]:

```python
# Order of selection
selection_order = ['Imported', 'In selection', 'First contact', 'QM', 'Vi

# Initialize a dictionary
contingency_tables_age = {}

# Loop through each state in the process order
for i, state in enumerate(selection_order):
    post_states = selection_order[i+1:]

    if post_states:   # Skips hired as it has no states after
        contingency_table_age = pd.DataFrame({
```

```
            f'Post {state}': data[data['Candidate State'].isin(post_state
            state: data[data['Candidate State'] == state].groupby('Age Ra
        })
        contingency_tables_age[state] = contingency_table_age
```

In [23]:
```
# Debugging code removed for clarity
```

```
--- Analysis for Imported ---
Chi-Square Test Statistic: 3259.7363634241924
p-value: 0.0
Degrees of Freedom: 6
Conclusion: For 'Imported', the variables 'Age Range' and 'Candidate State' are de
pendent (significant association).

--- Analysis for In selection ---
Chi-Square Test Statistic: 16.396566428031583
p-value: 0.011776528708639615
Degrees of Freedom: 6
Conclusion: For 'In selection', the variables 'Age Range' and 'Candidate State' ar
e dependent (significant association).

--- Analysis for First contact ---
Chi-Square Test Statistic: 92.78203768768347
p-value: 8.001878080863548e-18
Degrees of Freedom: 6
Conclusion: For 'First contact', the variables 'Age Range' and 'Candidate State' a
re dependent (significant association).

--- Analysis for QM ---
Chi-Square Test Statistic: 15.833494766518513
p-value: 0.014676125266923598
Degrees of Freedom: 6
Conclusion: For 'QM', the variables 'Age Range' and 'Candidate State' are dependen
t (significant association).

--- Analysis for Vivier ---
Chi-Square Test Statistic: 5.891409692097254
p-value: 0.4354638520055768
Degrees of Freedom: 6
Conclusion: For 'Vivier', the variables 'Age Range' and 'Candidate State' are inde
pendent (no significant association).

--- Analysis for Economic proposal ---
Chi-Square Test Statistic: 6.650736751992762
p-value: 0.35435723648351186
Degrees of Freedom: 6
Conclusion: For 'Economic proposal', the variables 'Age Range' and 'Candidate Stat
e' are independent (no significant association).
```

Candidate state vs. Region

In [24]:
```
# Order of selection
selection_order = ['Imported', 'In selection', 'First contact', 'QM', 'Vi

# Initialize a dictionary
```

```
contingency_tables_region = {}

# Loop through each state in the process order
for i, state in enumerate(selection_order):
    post_states = selection_order[i+1:]

    if post_states:   # Skips hired as it has no states after
        contingency_table_region = pd.DataFrame({
            f'Post {state}': data[data['Candidate State'].isin(post_state
            state: data[data['Candidate State'] == state].groupby('Region
        })
        contingency_tables_region[state] = contingency_table_region
```

In [25]:
```
# Debugging code removed for clarity
```

```
--- Analysis for Imported ---
Chi-Square Test Statistic: 3155.019021680986
p-value: 0.0
Degrees of Freedom: 22
Conclusion: For 'Imported', the variables 'Region' and 'Candidate State' are depen
dent (significant association).

--- Analysis for In selection ---
Chi-Square Test Statistic: 58.60831947120769
p-value: 3.5816827670575614e-05
Degrees of Freedom: 22
Conclusion: For 'In selection', the variables 'Region' and 'Candidate State' are d
ependent (significant association).

--- Analysis for First contact ---
Chi-Square Test Statistic: 53.10577079139694
p-value: 0.00021953020654831283
Degrees of Freedom: 22
Conclusion: For 'First contact', the variables 'Region' and 'Candidate State' are
dependent (significant association).

--- Analysis for QM ---
Chi-Square Test Statistic: 30.40062341794713
p-value: 0.08422824252642458
Degrees of Freedom: 21
Conclusion: For 'QM', the variables 'Region' and 'Candidate State' are independent
(no significant association).

--- Analysis for Vivier ---
Chi-Square Test Statistic: 33.34887180215482
p-value: 0.04248897706394278
Degrees of Freedom: 21
Conclusion: For 'Vivier', the variables 'Region' and 'Candidate State' are depende
nt (significant association).

--- Analysis for Economic proposal ---
Chi-Square Test Statistic: 56.19826716362774
p-value: 4.7191881513890646e-05
Degrees of Freedom: 21
Conclusion: For 'Economic proposal', the variables 'Region' and 'Candidate State'
are dependent (significant association).
```

For both 'Age Range' and 'Region' in some cells in the contingency table for various candidate states, the #observations is <5. This does not adhere with the rules of thumb of the Chi-squared test, this needs to be accounted for still.

In [ ]:

In [ ]:

In [ ]: