NORTH ANALYSIS

In [1]:
```
import pandas as pd

# Load dataset
north_path = "North.xlsx"
north_df = pd.read_excel(north_path)

# Display the first few rows
print(north_df.head())

# Show all column names
print("\nColumn Names in Dataset:")
print(north_df.columns)
```

```
   Dissemination area  Total private dwellings  \
0         35190274.0                   184.0
1         35190273.0                   185.0
2         35190357.0                   146.0
3         35190359.0                   105.0
4         35190348.0                    99.0

   Total - Age groups of the population - 100% data      0 to 4 years  \
0                                             545.0              10.0
1                                             530.0              20.0
2                                             455.0              15.0
3                                             300.0              10.0
4                                             305.0              15.0

      5 to 9 years      10 to 14 years      15 to 19 years  \
0             20.0                40.0                25.0
1             20.0                25.0                25.0
2             30.0                25.0                25.0
3             10.0                15.0                20.0
4             15.0                15.0                15.0

      20 to 24 years      25 to 29 years      30 to 34 years  ...  \
0               40.0                35.0                30.0  ...
1               25.0                35.0                30.0  ...
2               15.0                20.0                35.0  ...
3               10.0                20.0                20.0  ...
4               15.0                15.0                10.0  ...

     30 to 44 minutes    45 to 59 minutes    60 minutes and over  \
0               50.0                15.0                   30.0
1               55.0                15.0                   35.0
2               45.0                10.0                   10.0
3               35.0                10.0                   10.0
4               25.0                25.0                   15.0

   Total - Time leaving for work for the employed labour force aged 15 years and o
ver with a usual place of work or no fixed workplace address - 25% sample data  \
0                                                 165.0
1                                                 245.0
2                                                 130.0
3                                                 100.0
4                                                 105.0

     Between 5 a.m. and 5:59 a.m.    Between 6 a.m. and 6:59 a.m.  \
0                            20.0                            20.0
1                            10.0                            10.0
2                             0.0                            25.0
3                             0.0                            20.0
4                             0.0                            10.0

     Between 7 a.m. and 7:59 a.m.    Between 8 a.m. and 8:59 a.m.  \
0                            40.0                            60.0
1                            60.0                            75.0
2                            25.0                            40.0
3                            40.0                            20.0
4                            35.0                            25.0

     Between 9 a.m. and 11:59 a.m.    Between 12 p.m. and 4:59 a.m.
0                             25.0                              0.0
1                             65.0                             30.0
2                             15.0                             20.0
3                             15.0                              0.0
4                             30.0                             10.0
```

```
[5 rows x 56 columns]

Column Names in Dataset:
Index(['Dissemination area', 'Total private dwellings',
       'Total - Age groups of the population - 100% data', '    0 to 4 years',
       '    5 to 9 years', '    10 to 14 years', '    15 to 19 years',
       '    20 to 24 years', '    25 to 29 years', '    30 to 34 years',
       '    35 to 39 years', '    40 to 44 years', '    45 to 49 years',
       '    50 to 54 years', '    55 to 59 years',
       'Average age of the population', 'Median age of the population',
       'Total - Census families in private households by family size - 100% data',
       '  2 persons', '  3 persons', '  4 persons', '  5 or more persons',
       'Average size of census families',
       'Average number of children in census families with children',
       '  Total - Persons not in census families in private households - 100% dat
a',
       '    Living alone', 'Total - Household type - 100% data',
       '    Couple-family households', '      With children',
       '      Without children',
       '  Median total income of couple-with-children economic families  in 2020
($)',
       '  Median after-tax income of couple-with-children economic families in 202
0 ($)',
       '  Average family size of couple-with-children economic families',
       '  Average total income of couple-with-children economic families in 2020
($)',
       '  Average after-tax income of couple-with-children economic families in 20
20 ($)',
       'Participation rate', 'Employment rate', 'Unemployment rate',
       'Total - Place of work status for the employed labour force aged 15 years a
nd over - 25% sample data',
       '  Worked at home', '  No fixed workplace address',
       '  Usual place of work', '  Car, truck or van', '  Public transit',
       '  Less than 15 minutes', '  15 to 29 minutes', '  30 to 44 minutes',
       '  45 to 59 minutes', '  60 minutes and over',
       'Total - Time leaving for work for the employed labour force aged 15 years
and over with a usual place of work or no fixed workplace address - 25% sample dat
a',
       '  Between 5 a.m. and 5:59 a.m.', '  Between 6 a.m. and 6:59 a.m.',
       '  Between 7 a.m. and 7:59 a.m.', '  Between 8 a.m. and 8:59 a.m.',
       '  Between 9 a.m. and 11:59 a.m.', '  Between 12 p.m. and 4:59 a.m.'],
      dtype='object')
```

In [3]:

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

# STEP 1: LOAD THE NORTH DATASET
north_path = "North.xlsx"
north_df = pd.read_excel(north_path)
print("NORTH DATASET INFO")
north_df.info()
```

```
NORTH DATASET INFO
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 61 entries, 0 to 60
Data columns (total 56 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   Dissemination area
60 non-null     float64
 1   Total private dwellings
60 non-null     float64
 2   Total - Age groups of the population - 100% data
60 non-null     float64
 3       0 to 4 years
60 non-null     float64
 4       5 to 9 years
60 non-null     float64
 5       10 to 14 years
60 non-null     float64
 6       15 to 19 years
60 non-null     float64
 7       20 to 24 years
60 non-null     float64
 8       25 to 29 years
60 non-null     float64
 9       30 to 34 years
60 non-null     float64
 10      35 to 39 years
60 non-null     float64
 11      40 to 44 years
60 non-null     float64
 12      45 to 49 years
60 non-null     float64
 13      50 to 54 years
60 non-null     float64
 14      55 to 59 years
60 non-null     float64
 15  Average age of the population
60 non-null     float64
 16  Median age of the population
60 non-null     float64
 17  Total - Census families in private households by family size - 100% data
60 non-null     float64
 18    2 persons
60 non-null     float64
 19    3 persons
60 non-null     float64
 20    4 persons
60 non-null     float64
 21    5 or more persons
60 non-null     float64
 22  Average size of census families
60 non-null     float64
 23  Average number of children in census families with children
60 non-null     float64
 24    Total - Persons not in census families in private households - 100% data
60 non-null     float64
 25      Living alone
60 non-null     float64
 26  Total - Household type - 100% data
60 non-null     float64
 27      Couple-family households
60 non-null     float64
```

```
 28        With children
60 non-null     float64
 29        Without children
60 non-null     float64
 30    Median total income of couple-with-children economic families  in 2020 ($)
56 non-null     float64
 31    Median after-tax income of couple-with-children economic families in 2020
($)
57 non-null     float64
 32    Average family size of couple-with-children economic families
60 non-null     float64
 33    Average total income of couple-with-children economic families in 2020 ($)
55 non-null     float64
 34    Average after-tax income of couple-with-children economic families in 2020
($)
55 non-null     float64
 35  Participation rate
60 non-null     float64
 36  Employment rate
60 non-null     float64
 37  Unemployment rate
60 non-null     float64
 38  Total - Place of work status for the employed labour force aged 15 years and
over - 25% sample data
60 non-null     float64
 39    Worked at home
60 non-null     float64
 40    No fixed workplace address
60 non-null     float64
 41    Usual place of work
60 non-null     float64
 42    Car, truck or van
60 non-null     float64
 43    Public transit
60 non-null     float64
 44    Less than 15 minutes
60 non-null     float64
 45    15 to 29 minutes
60 non-null     float64
 46    30 to 44 minutes
60 non-null     float64
 47    45 to 59 minutes
60 non-null     float64
 48    60 minutes and over
60 non-null     float64
 49  Total - Time leaving for work for the employed labour force aged 15 years and
over with a usual place of work or no fixed workplace address - 25% sample data  6
0 non-null     float64
 50    Between 5 a.m. and 5:59 a.m.
60 non-null     float64
 51    Between 6 a.m. and 6:59 a.m.
60 non-null     float64
 52    Between 7 a.m. and 7:59 a.m.
60 non-null     float64
 53    Between 8 a.m. and 8:59 a.m.
60 non-null     float64
 54    Between 9 a.m. and 11:59 a.m.
60 non-null     float64
 55    Between 12 p.m. and 4:59 a.m.
60 non-null     float64
dtypes: float64(56)
memory usage: 26.8 KB
```

In [4]:
```
# STEP 2: HANDLE MISSING VALUES
numeric_cols = north_df.select_dtypes(include=['number']).columns
north_df[numeric_cols] = north_df[numeric_cols].fillna(north_df[numeric_cols].media
```

In [5]:
```
# STEP 3: DISPLAY BASIC STATISTICS
print("\nNorth Summary Statistics")
print(north_df.describe())
```

North Summary Statistics

	Dissemination area	Total private dwellings
count	6.100000e+01	61.000000
mean	3.519041e+07	225.991803
std	2.342197e+02	180.938076
min	3.519027e+07	64.000000
25%	3.519030e+07	126.000000
50%	3.519034e+07	160.500000
75%	3.519036e+07	216.000000
max	3.519106e+07	856.000000

	Total - Age groups of the population - 100% data	0 to 4 years
count	61.000000	61.000000
mean	626.475410	24.549180
std	391.017204	18.970966
min	175.000000	5.000000
25%	405.000000	10.000000
50%	490.000000	17.500000
75%	770.000000	30.000000
max	2220.000000	90.000000

	5 to 9 years	10 to 14 years	15 to 19 years
count	61.000000	61.000000	61.000000
mean	30.901639	32.622951	35.819672
std	22.721286	22.575919	22.916157
min	10.000000	5.000000	5.000000
25%	20.000000	15.000000	20.000000
50%	20.000000	25.000000	30.000000
75%	30.000000	40.000000	45.000000
max	105.000000	115.000000	120.000000

	20 to 24 years	25 to 29 years	30 to 34 years	...
count	61.000000	61.000000	61.000000	...
mean	36.803279	32.459016	32.295082	...
std	20.372219	22.278969	24.521653	...
min	10.000000	5.000000	5.000000	...
25%	25.000000	20.000000	20.000000	...
50%	35.000000	30.000000	25.000000	...
75%	40.000000	35.000000	35.000000	...
max	100.000000	145.000000	145.000000	...

	30 to 44 minutes	45 to 59 minutes	60 minutes and over
count	61.000000	61.000000	61.000000
mean	42.622951	17.049180	16.639344
std	26.608372	14.814887	16.776007
min	0.000000	0.000000	0.000000
25%	25.000000	10.000000	0.000000
50%	40.000000	15.000000	15.000000
75%	55.000000	20.000000	20.000000
max	130.000000	80.000000	90.000000

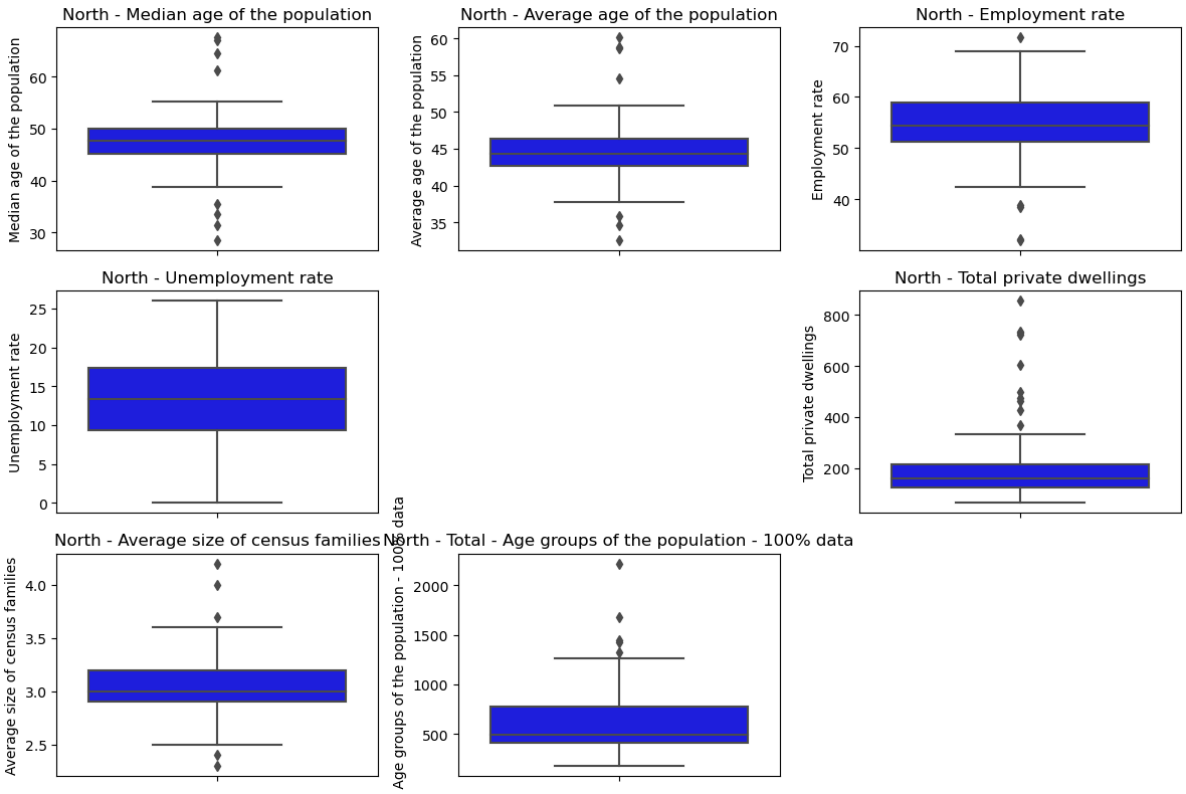
	Total - Time leaving for work for the employed labour force aged 15 years and over with a usual place of work or no fixed workplace address - 25% sample data
count	61.000000
mean	163.811475
std	95.805365
min	45.000000
25%	100.000000
50%	142.500000
75%	205.000000
max	565.000000

	Between 5 a.m. and 5:59 a.m.	Between 6 a.m. and 6:59 a.m.

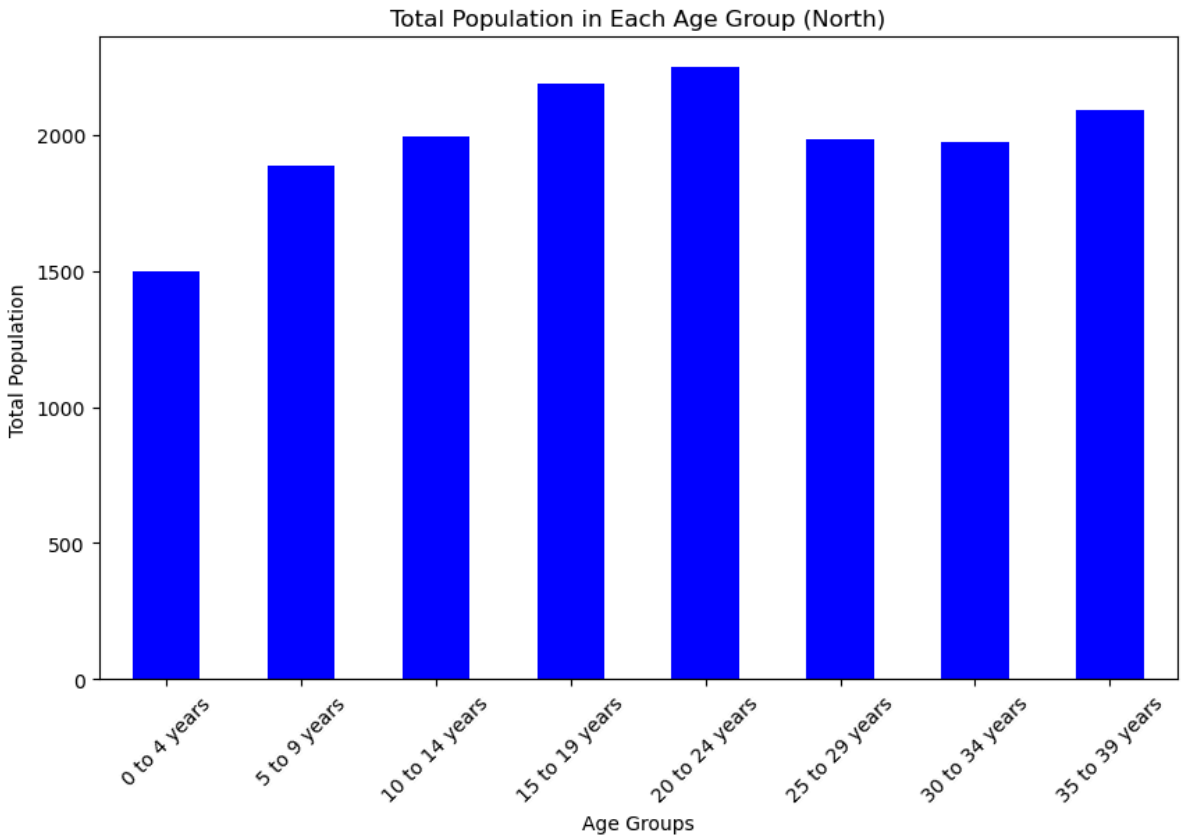
```
count                          61.000000                          61.000000
mean                            6.065574                          17.213115
std                             9.921137                          14.986788
min                             0.000000                           0.000000
25%                             0.000000                          10.000000
50%                             0.000000                          15.000000
75%                            10.000000                          25.000000
max                            35.000000                          60.000000

         Between 7 a.m. and 7:59 a.m.    Between 8 a.m. and 8:59 a.m.  \
count                         61.000000                          61.000000
mean                          33.032787                          42.581967
std                           24.803876                          27.049211
min                            0.000000                           0.000000
25%                           20.000000                          20.000000
50%                           30.000000                          37.500000
75%                           40.000000                          55.000000
max                          160.000000                         140.000000

         Between 9 a.m. and 11:59 a.m.    Between 12 p.m. and 4:59 a.m.
count                         61.000000                          61.000000
mean                          41.557377                          19.426230
std                           31.497897                          16.152873
min                            0.000000                           0.000000
25%                           25.000000                          10.000000
50%                           30.000000                          15.000000
75%                           55.000000                          30.000000
max                          165.000000                          65.000000

[8 rows x 56 columns]
```

In [6]:
```python
# STEP 4: ANALYZE KEY METRICS
columns_to_analyze = ['Median age of the population', 'Average age of the populatio
                      'Employment rate', 'Unemployment rate', 'Public transit',
                      'Total private dwellings', 'Average size of census families',
                      'Total - Age groups of the population - 100% data']
```

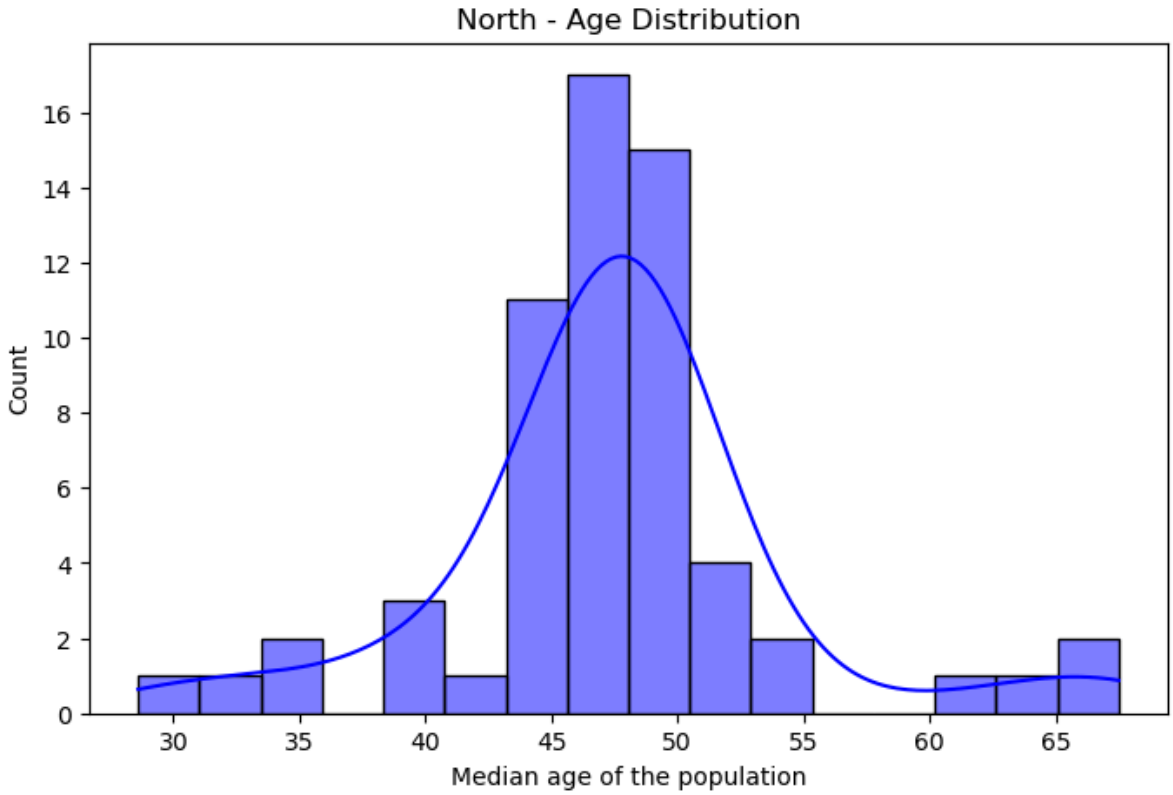
In [8]:
```python
# STEP 5: VISUALIZE DATA
plt.figure(figsize=(12, 8))
for i, col in enumerate(columns_to_analyze, 1):
    col = col.strip()  # Ensure no extra spaces in column names
    if col in north_df.columns:
        plt.subplot(3, 3, i)
        sns.boxplot(y=north_df[col], color='blue')
        plt.title(f'North - {col}')
plt.tight_layout()
plt.show()
```



In [10]:
```
# Remove leading/trailing spaces from column names
north_df.columns = north_df.columns.str.strip()

# Now try summing the values across the age groups
age_groups = ['0 to 4 years', '5 to 9 years', '10 to 14 years', '15 to 19 years',
              '20 to 24 years', '25 to 29 years', '30 to 34 years', '35 to 39 years

# Sum of people in each age group across the regions
age_group_sums = north_df[age_groups].sum()

# Plotting the bar plot
plt.figure(figsize=(10, 6))
age_group_sums.plot(kind='bar', color='blue')
plt.title('Total Population in Each Age Group (North)')
plt.xlabel('Age Groups')
plt.ylabel('Total Population')
plt.xticks(rotation=45)
plt.show()
```

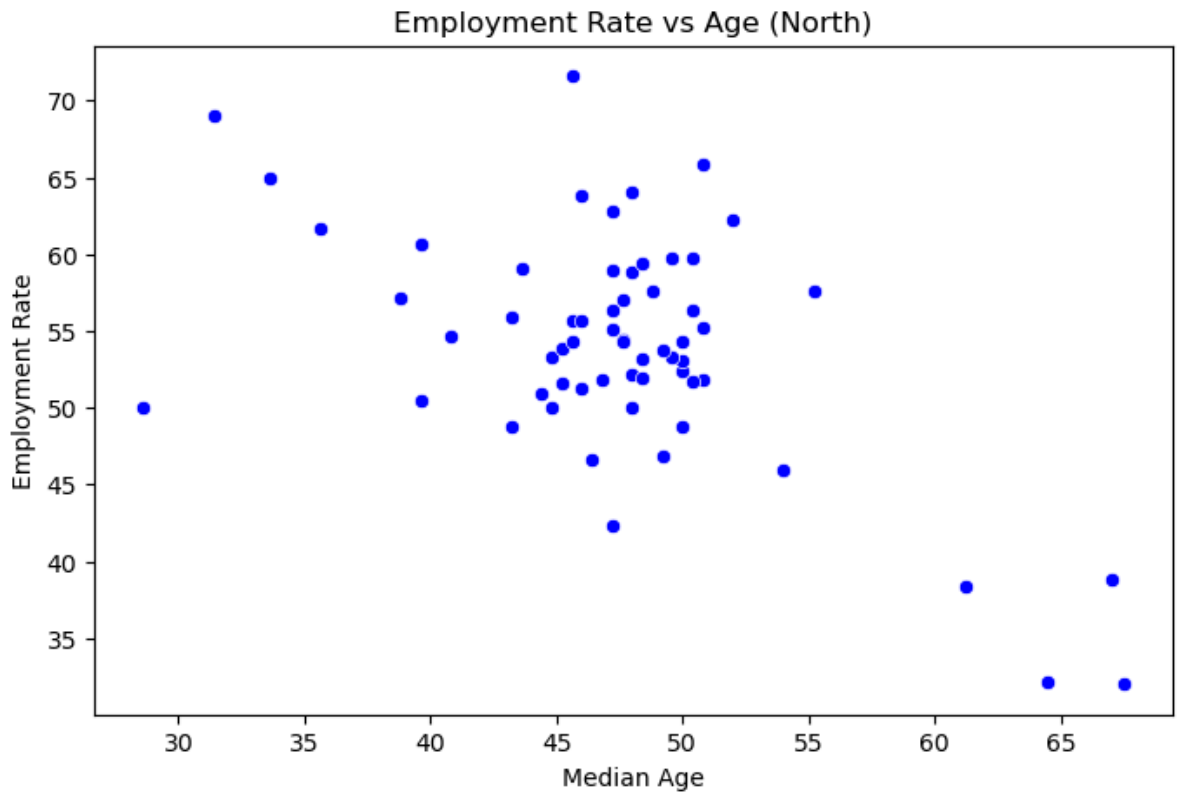


In [12]:
```
# STEP 6: PLOT AGE DISTRIBUTION
plt.figure(figsize=(8, 5))
sns.histplot(north_df['Median age of the population'], color='blue', kde=True)
plt.title("North - Age Distribution")
plt.show()
```

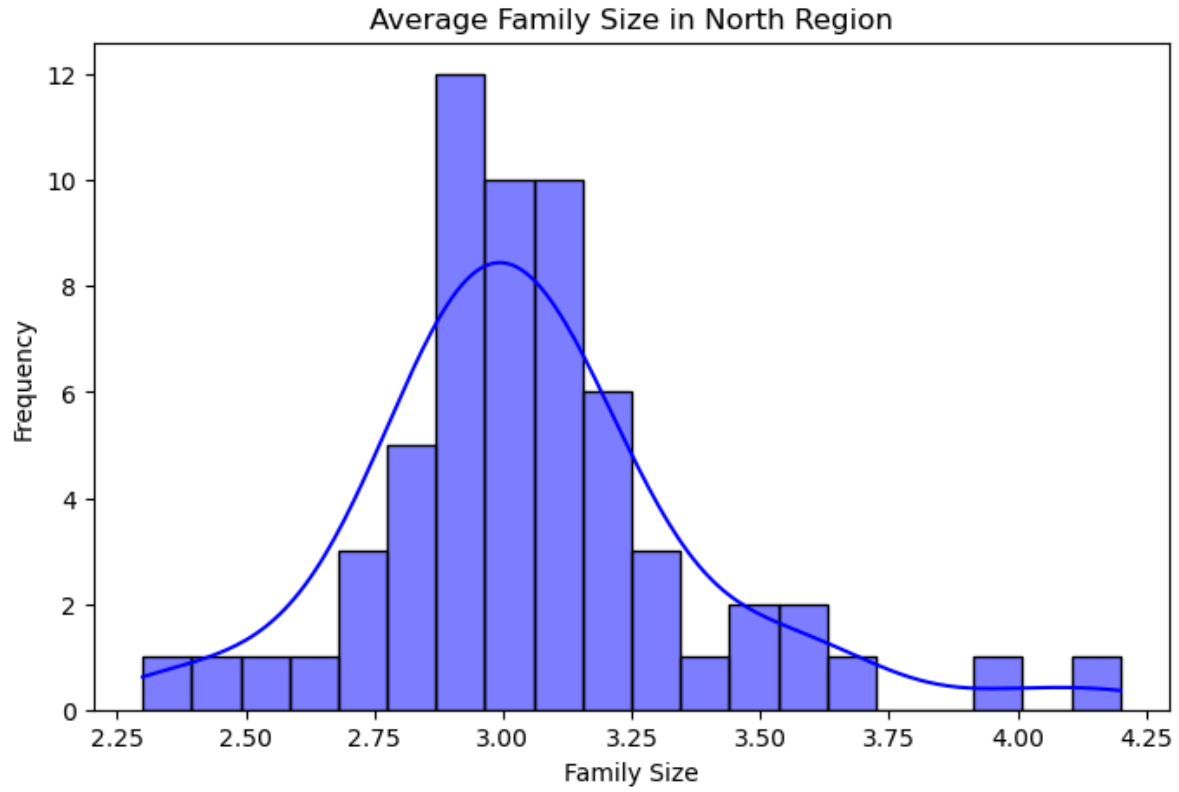


In [13]:
```
# STEP 7: SCATTERPLOT - AGE VS EMPLOYMENT RATE
plt.figure(figsize=(8, 5))
sns.scatterplot(x=north_df['Median age of the population'], y=north_df['Employment
plt.title("Employment Rate vs Age (North)")
plt.xlabel("Median Age")
```

```
plt.ylabel("Employment Rate")
plt.show()
```



In [16]:
```
# Plotting average family size across the North region
plt.figure(figsize=(8, 5))
sns.histplot(north_df['Average size of census families'], bins=20, kde=True, color=
plt.title("Average Family Size in North Region")
plt.xlabel("Family Size")
plt.ylabel("Frequency")
plt.show()
```



In [30]:
```
print(repr(north_df.columns.tolist()))
```

```
['Dissemination area', 'Total private dwellings', 'Total - Age groups of the popul
ation - 100% data', '0 to 4 years', '5 to 9 years', '10 to 14 years', '15 to 19 ye
ars', '20 to 24 years', '25 to 29 years', '30 to 34 years', '35 to 39 years', '40
to 44 years', '45 to 49 years', '50 to 54 years', '55 to 59 years', 'Average age o
f the population', 'Median age of the population', 'Total - Census families in pri
vate households by family size - 100% data', '2 persons', '3 persons', '4 person
s', '5 or more persons', 'Average size of census families', 'Average number of chi
ldren in census families with children', 'Total - Persons not in census families i
n private households - 100% data', 'Living alone', 'Total - Household type - 100%
data', 'Couple-family households', 'With children', 'Without children', 'Median to
tal income of couple-with-children economic families  in 2020 ($)', 'Median after-
tax income of couple-with-children economic families in 2020 ($)', 'Average family
size of couple-with-children economic families', 'Average total income of couple-w
ith-children economic families in 2020 ($)', 'Average after-tax income of couple-w
ith-children economic families in 2020 ($)', 'Participation rate', 'Employment rat
e', 'Unemployment rate', 'Total - Place of work status for the employed labour for
ce aged 15 years and over - 25% sample data', 'Worked at home', 'No fixed workplac
e address', 'Usual place of work', 'Car, truck or van', 'Public transit', 'Less th
an 15 minutes', '15 to 29 minutes', '30 to 44 minutes', '45 to 59 minutes', '60 mi
nutes and over', 'Total - Time leaving for work for the employed labour force aged
15 years and over with a usual place of work or no fixed workplace address - 25% s
ample data', 'Between 5 a.m. and 5:59 a.m.', 'Between 6 a.m. and 6:59 a.m.', 'Betw
een 7 a.m. and 7:59 a.m.', 'Between 8 a.m. and 8:59 a.m.', 'Between 9 a.m. and 11:
59 a.m.', 'Between 12 p.m. and 4:59 a.m.']
```

In [31]:
```
north_df.rename(columns={
    'Median total income of couple-with-children economic families  in 2020 ($)': '
    'Median after-tax income of couple-with-children economic families in 2020 ($)'
    'Average total income of couple-with-children economic families in 2020 ($)': '
    'Average after-tax income of couple-with-children economic families in 2020 ($)
}, inplace=True)
```

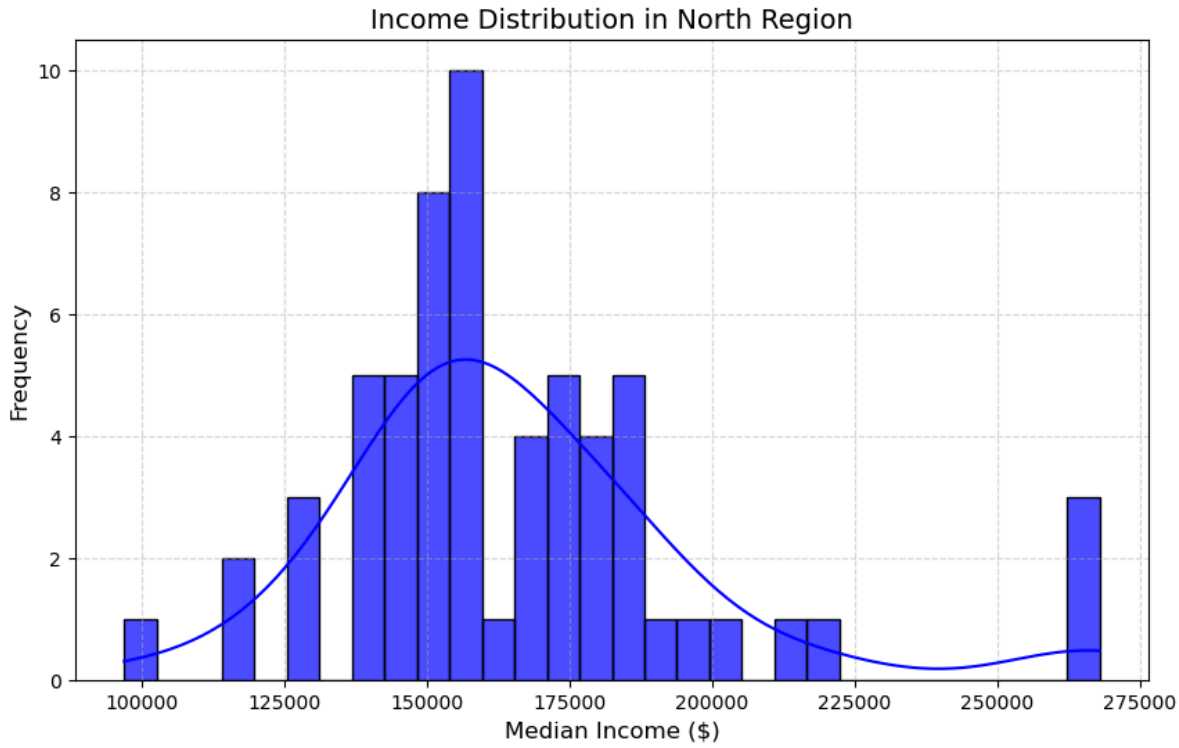
In [32]:
```
for col in north_df.columns:
    print(f"'{col}'")
```

```
'Dissemination area'
'Total private dwellings'
'Total - Age groups of the population - 100% data'
'0 to 4 years'
'5 to 9 years'
'10 to 14 years'
'15 to 19 years'
'20 to 24 years'
'25 to 29 years'
'30 to 34 years'
'35 to 39 years'
'40 to 44 years'
'45 to 49 years'
'50 to 54 years'
'55 to 59 years'
'Average age of the population'
'Median age of the population'
'Total - Census families in private households by family size - 100% data'
'2 persons'
'3 persons'
'4 persons'
'5 or more persons'
'Average size of census families'
'Average number of children in census families with children'
'Total - Persons not in census families in private households - 100% data'
'Living alone'
'Total - Household type - 100% data'
'Couple-family households'
'With children'
'Without children'
'Median_Income'
'Median_After_Tax_Income'
'Average family size of couple-with-children economic families'
'Avg_Income'
'Avg_After_Tax_Income'
'Participation rate'
'Employment rate'
'Unemployment rate'
'Total - Place of work status for the employed labour force aged 15 years and over
- 25% sample data'
'Worked at home'
'No fixed workplace address'
'Usual place of work'
'Car, truck or van'
'Public transit'
'Less than 15 minutes'
'15 to 29 minutes'
'30 to 44 minutes'
'45 to 59 minutes'
'60 minutes and over'
'Total - Time leaving for work for the employed labour force aged 15 years and ove
r with a usual place of work or no fixed workplace address - 25% sample data'
'Between 5 a.m. and 5:59 a.m.'
'Between 6 a.m. and 6:59 a.m.'
'Between 7 a.m. and 7:59 a.m.'
'Between 8 a.m. and 8:59 a.m.'
'Between 9 a.m. and 11:59 a.m.'
'Between 12 p.m. and 4:59 a.m.'
```

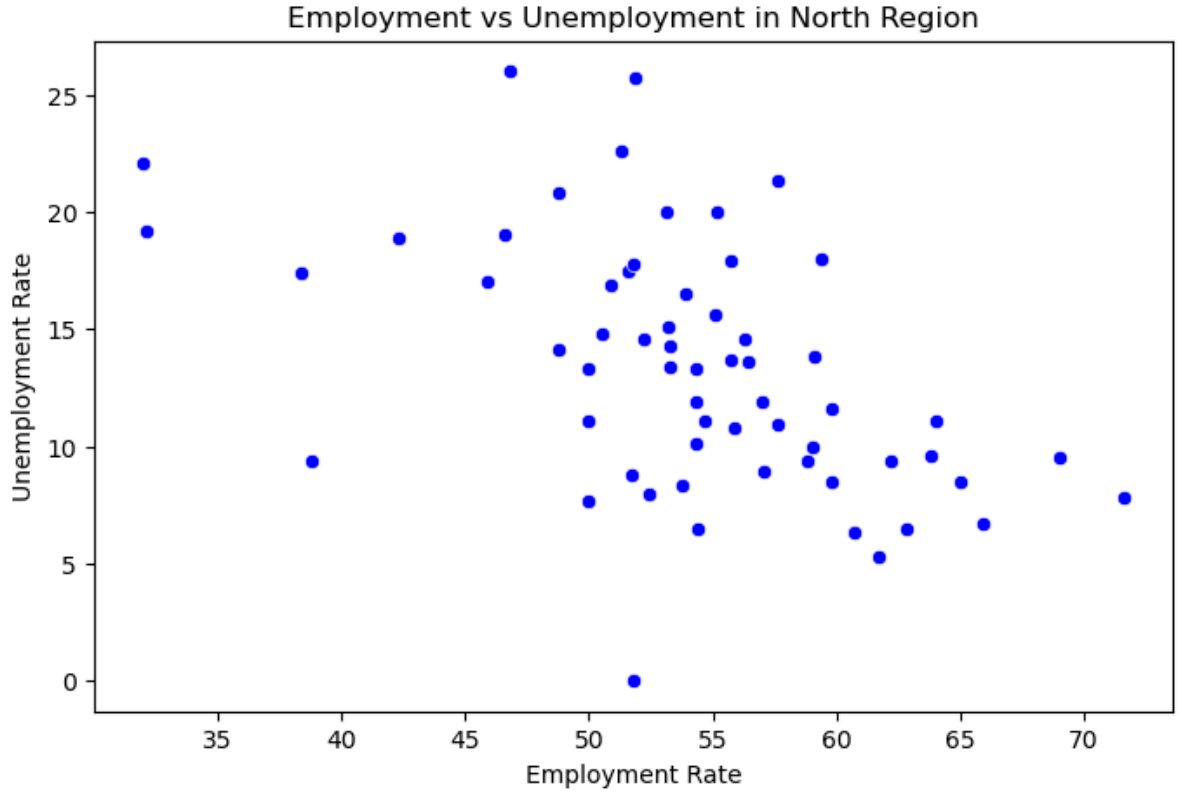
In [34]:

```
plt.figure(figsize=(10, 6))  # Increase figure size
sns.histplot(north_df['Median_Income'].dropna(), bins=30, kde=True, color='blue', a

plt.xlabel("Median Income ($)", fontsize=12)
plt.ylabel("Frequency", fontsize=12)
```

```
plt.title("Income Distribution in North Region", fontsize=14)
plt.grid(True, linestyle="--", alpha=0.5)  # Add a subtle grid for clarity
plt.show()
```



In [35]:
```
# Plotting employment rate vs. unemployment rate for the North region
plt.figure(figsize=(8, 5))
sns.scatterplot(x=north_df['Employment rate'], y=north_df['Unemployment rate'], col
plt.title("Employment vs Unemployment in North Region")
plt.xlabel("Employment Rate")
plt.ylabel("Unemployment Rate")
plt.show()
```



In [39]:
```
north_df[['Employment rate', 'Unemployment rate']].corr()
```

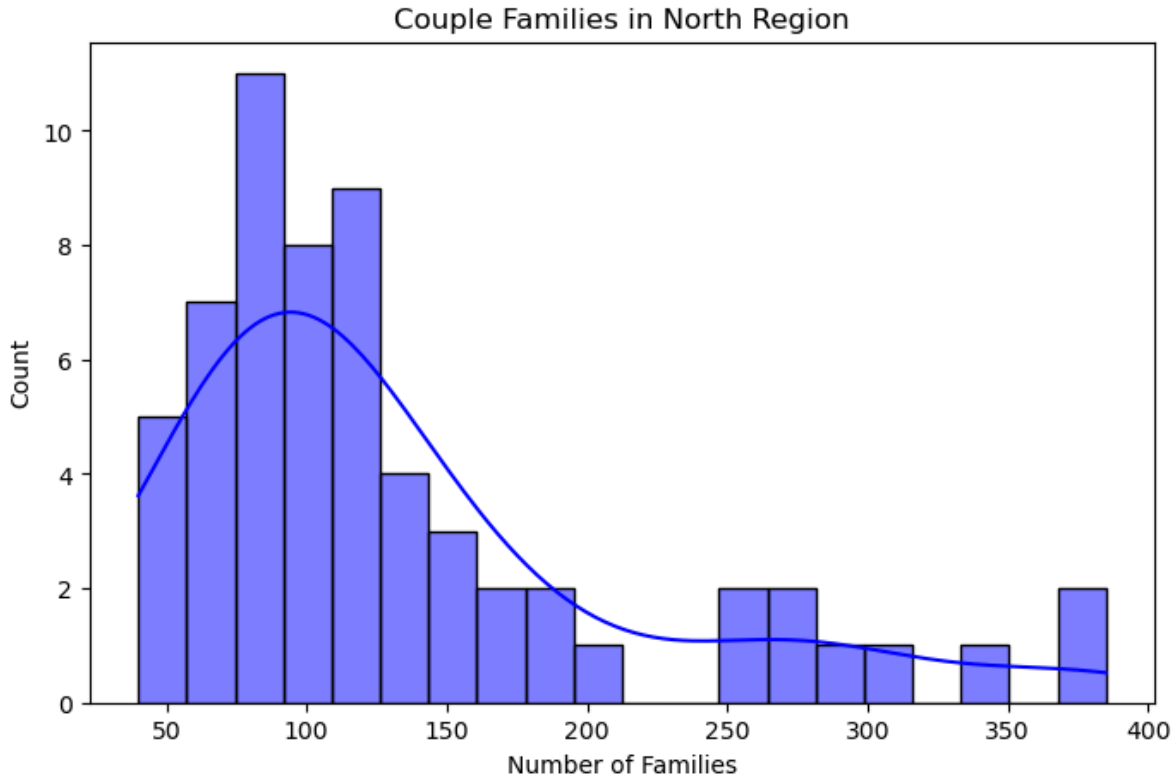
Out[39]:

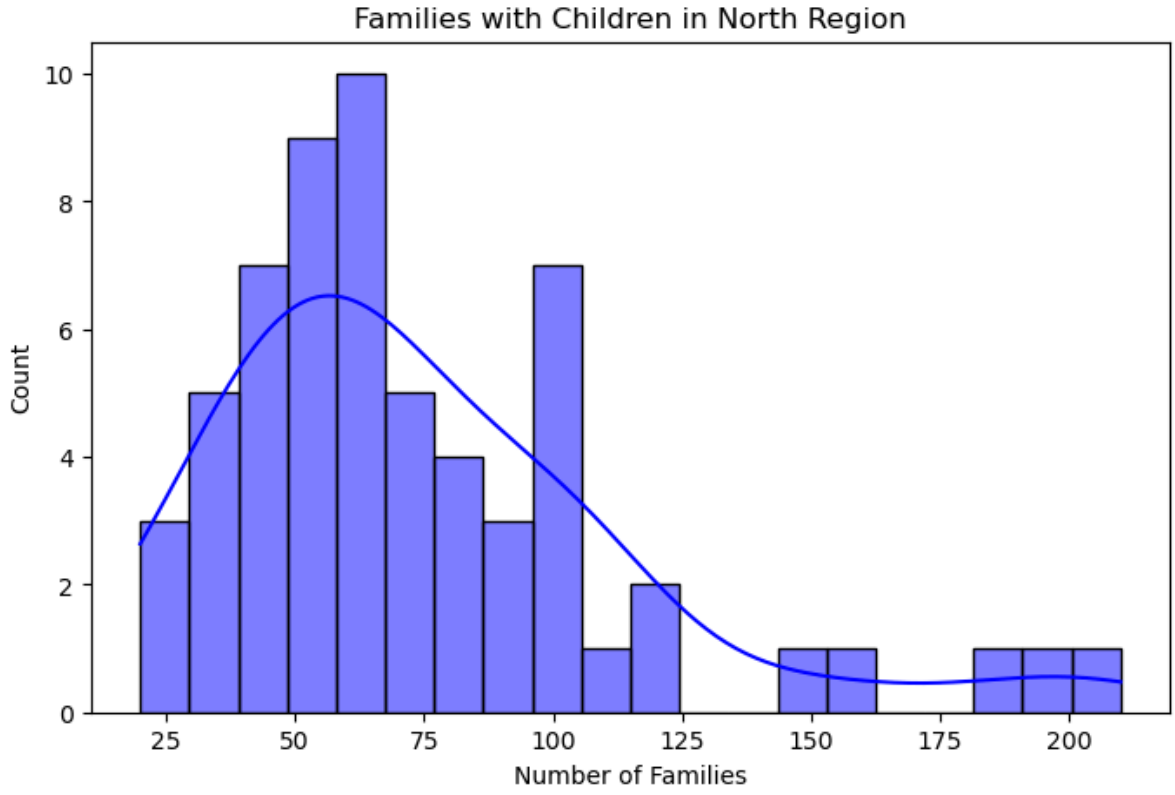
	Employment rate	Unemployment rate
Employment rate	1.000000	-0.497489
Unemployment rate	-0.497489	1.000000

In [41]:
```
print(north_df.columns)
```

```
Index(['Dissemination area', 'Total private dwellings',
       'Total - Age groups of the population - 100% data', '0 to 4 years',
       '5 to 9 years', '10 to 14 years', '15 to 19 years', '20 to 24 years',
       '25 to 29 years', '30 to 34 years', '35 to 39 years', '40 to 44 years',
       '45 to 49 years', '50 to 54 years', '55 to 59 years',
       'Average age of the population', 'Median age of the population',
       'Total - Census families in private households by family size - 100% data',
       '2 persons', '3 persons', '4 persons', '5 or more persons',
       'Average size of census families',
       'Average number of children in census families with children',
       'Total - Persons not in census families in private households - 100% data',
       'Living alone', 'Total - Household type - 100% data',
       'Couple-family households', 'With children', 'Without children',
       'Median_Income', 'Median_After_Tax_Income',
       'Average family size of couple-with-children economic families',
       'Avg_Income', 'Avg_After_Tax_Income', 'Participation rate',
       'Employment rate', 'Unemployment rate',
       'Total - Place of work status for the employed labour force aged 15 years a
nd over - 25% sample data',
       'Worked at home', 'No fixed workplace address', 'Usual place of work',
       'Car, truck or van', 'Public transit', 'Less than 15 minutes',
       '15 to 29 minutes', '30 to 44 minutes', '45 to 59 minutes',
       '60 minutes and over',
       'Total - Time leaving for work for the employed labour force aged 15 years
and over with a usual place of work or no fixed workplace address - 25% sample dat
a',
       'Between 5 a.m. and 5:59 a.m.', 'Between 6 a.m. and 6:59 a.m.',
       'Between 7 a.m. and 7:59 a.m.', 'Between 8 a.m. and 8:59 a.m.',
       'Between 9 a.m. and 11:59 a.m.', 'Between 12 p.m. and 4:59 a.m.'],
      dtype='object')
```

In [44]:
```
plt.figure(figsize=(8, 5))
sns.histplot(north_df['Couple-family households'], bins=20, kde=True, color='blue')
plt.title("Couple Families in North Region")
plt.xlabel("Number of Families")
plt.show()

# Or

plt.figure(figsize=(8, 5))
sns.histplot(north_df['With children'], bins=20, kde=True, color='blue')
plt.title("Families with Children in North Region")
plt.xlabel("Number of Families")
plt.show()
```


Couple Families in North Region
Count
Number of Families
0
2
4
6
8
10
50
100
150
200
250
300
350
400


Families with Children in North Region
Count
Number of Families
0
2
4
6
8
10
25
50
75
100
125
150
175
200

In [ ]: