World Bank Science and Technology Data Analysis

Importing the Data

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

```
st_data_2018 = pd.read_csv("/content/science_tech_2018.csv")
st_data_2018.head()
```

	Country Name	High-technology exports (% of manufactured exports)	High-technology exports (current US$)	Trademark applications, total	Trademark applications, direct resident	Trademark applications, direct nonresident	Patent applications, residents	Patent applications, nonresidents	Scientific and technical journal articles
0	Afghanistan	NaN	NaN	NaN	NaN	NaN	NaN	NaN	111.72
1	Albania	0.049514	591717.0	3713.0	917.0	2796.0	15.0	3.0	180.36
2	Algeria	NaN	NaN	9490.0	5469.0	4021.0	152.0	521.0	5231.44
3	American	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

```
st_data_2018.info()
st_data_2018.shape
st_data_2018.size
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 219 entries, 0 to 218
Data columns (total 10 columns):
 #   Column                                               Non-Null Count  Dtype  
---  ------                                               --------------  -----  
 0   Country Name                                         219 non-null    object 
 1   High-technology exports (% of manufactured exports)  132 non-null    float64
 2   High-technology exports (current US$)                131 non-null    float64
 3   Trademark applications, total                        133 non-null    float64
 4   Trademark applications, direct resident              128 non-null    float64
 5   Trademark applications, direct nonresident           129 non-null    float64
 6   Patent applications, residents                       118 non-null    float64
 7   Patent applications, nonresidents                    119 non-null    float64
 8   Scientific and technical journal articles            199 non-null    float64
 9   Research and development expenditure (% of GDP)      73 non-null     float64
dtypes: float64(9), object(1)
memory usage: 17.2+ KB
2190
```

```
st_data_2018.columns
```

```
Index(['Country Name', 'High-technology exports (% of manufactured exports)',
       'High-technology exports (current US$)',
       'Trademark applications, total',
       'Trademark applications, direct resident',
       'Trademark applications, direct nonresident',
       'Patent applications, residents', 'Patent applications, nonresidents',
       'Scientific and technical journal articles',
       'Research and development expenditure (% of GDP)'],
      dtype='object')
```

```
st_data_2009 = pd.read_csv("/content/science_tech_2009.csv")
st_data_2009.head()
```

	Country Name	High-technology exports (% of manufactured exports)	High-technology exports (current US$)	Trademark applications, total	Trademark applications, direct resident	Trademark applications, direct nonresident	Patent applications, residents	Patent applications, nonresidents	Scientific and technical journal articles
0	Afghanistan	NaN	NaN	NaN	NaN	NaN	NaN	NaN	26.30
1	Albania	1.328024	10128935.0	3456.0	213.0	3243.0	NaN	361.0	70.35
2	Algeria	0.653887	4616076.0	5345.0	NaN	NaN	NaN	NaN	2135.32
3	American	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

```
st_data_2009.info()
st_data_2009.shape
st_data_2009.size
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 219 entries, 0 to 218
Data columns (total 10 columns):
 #   Column                                           Non-Null Count  Dtype  
---  ------                                           --------------  -----  
 0   Country Name                                     219 non-null    object 
 1   High-technology exports (% of manufactured exports)  131 non-null    float64
 2   High-technology exports (current US$)            131 non-null    float64
 3   Trademark applications, total                    111 non-null    float64
 4   Trademark applications, direct resident          101 non-null    float64
 5   Trademark applications, direct nonresident       101 non-null    float64
 6   Patent applications, residents                   93 non-null     float64
 7   Patent applications, nonresidents                101 non-null    float64
 8   Scientific and technical journal articles        199 non-null    float64
 9   Research and development expenditure (% of GDP)  97 non-null     float64
dtypes: float64(9), object(1)
memory usage: 17.2+ KB
2190
```

```
st_data_2009.columns
```

```
Index(['Country Name', 'High-technology exports (% of manufactured exports)',
       'High-technology exports (current US$)',
       'Trademark applications, total',
       'Trademark applications, direct resident',
       'Trademark applications, direct nonresident',
       'Patent applications, residents', 'Patent applications, nonresidents',
       'Scientific and technical journal articles',
       'Research and development expenditure (% of GDP)'],
      dtype='object')
```

Cleaning the Data

Missing Data

```
st_data_2018.isnull().sum()
```

```
                                                        0
                                     Country Name       0
  High-technology exports (% of manufactured exports)  87
               High-technology exports (current US$)   88
                      Trademark applications, total    86
               Trademark applications, direct resident 91
            Trademark applications, direct nonresident 90
                      Patent applications, residents  101
                   Patent applications, nonresidents  100
            Scientific and technical journal articles  20
      Research and development expenditure (% of GDP) 146

dtype: int64
```

```
st_data_2018_clean = st_data_2018.dropna()
st_data_2018_clean.isnull().sum()
```

```
                                                       0
                                     Country Name      0
  High-technology exports (% of manufactured exports)  0
               High-technology exports (current US$)   0
                      Trademark applications, total    0
               Trademark applications, direct resident 0
            Trademark applications, direct nonresident 0
                      Patent applications, residents   0
                   Patent applications, nonresidents   0
            Scientific and technical journal articles  0
      Research and development expenditure (% of GDP)  0

dtype: int64
```

```
st_data_2009.isnull().sum()
```

```
                                                        0
                                     Country Name       0
  High-technology exports (% of manufactured exports)  88
               High-technology exports (current US$)   88
                      Trademark applications, total   108
               Trademark applications, direct resident 118
            Trademark applications, direct nonresident 118
                      Patent applications, residents  126
                   Patent applications, nonresidents  118
            Scientific and technical journal articles  20
      Research and development expenditure (% of GDP) 122

dtype: int64
```

```
st_data_2018_clean.shape
```

```
(51, 10)
```

```
st_data_2009_clean = st_data_2009.dropna()
st_data_2009_clean.isnull().sum()
```

	0
Country Name	0
High-technology exports (% of manufactured exports)	0
High-technology exports (current US$)	0
Trademark applications, total	0
Trademark applications, direct resident	0
Trademark applications, direct nonresident	0
Patent applications, residents	0
Patent applications, nonresidents	0
Scientific and technical journal articles	0
Research and development expenditure (% of GDP)	0

dtype: int64

```
st_data_2009_clean.shape
```

```
(56, 10)
```

Exploratory Data Analysis

Descriptive Statistics

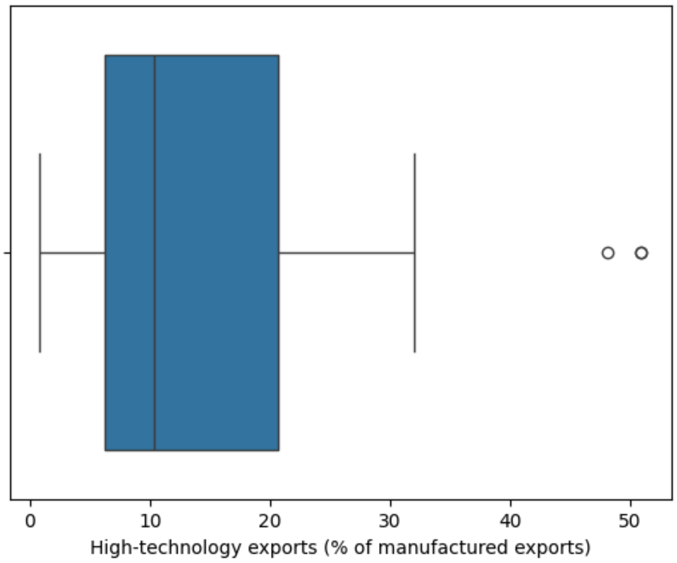
```
# Display descriptive statistics for the cleaned 2018 data
#pd.set_option('display.float.format', lambda x:'%.3f'%x)
st_data_2018_clean.describe().style.format("{:,.0f}")
```

	High-technology exports (% of manufactured exports)	High-technology exports (current US$)	Trademark applications, total	Trademark applications, direct resident	Trademark applications, direct nonresident	Patent applications, residents	Patent applications, nonresidents	Scientific and technical journal articles	Re devel expen (% c
count	51	51	51	51	51	51	51	51	
mean	12	25,697,867,097	44,515	33,921	10,594	16,828	11,187	30,499	
std	8	48,690,230,336	85,394	69,219	18,797	56,775	44,725	65,091	
min	0	115,051	1,974	141	1,008	1	4	127	
25%	5	433,839,196	4,203	1,858	1,937	96	26	1,416	
50%	10	4,294,542,879	10,025	4,711	3,417	678	168	10,345	

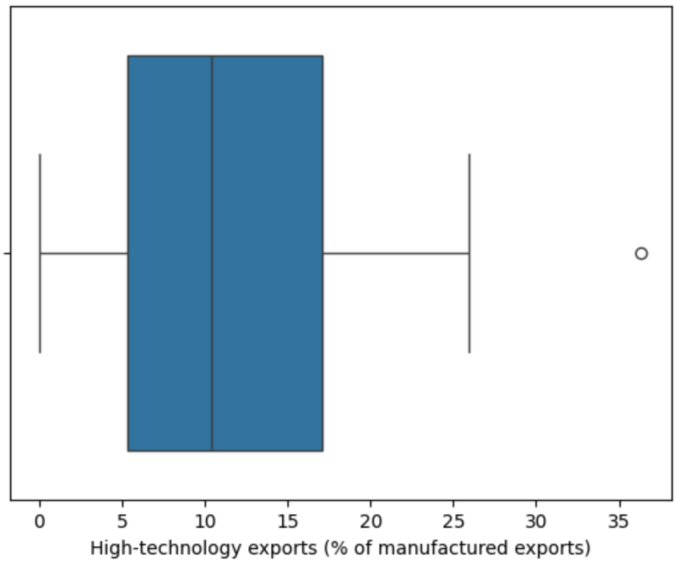
```
#pd.set_option('display.float.format', lambda x:'%.3f'%x)
st_data_2009_clean.describe().style.format("{:,.0f}")
```

	High-technology exports (% of manufactured exports)	High-technology exports (current US$)	Trademark applications, total	Trademark applications, direct resident	Trademark applications, direct nonresident	Patent applications, residents	Patent applications, nonresidents	Scientific and technical journal articles	Re devel expen (% c
count	56	56	56	56	56	56	56	56	
mean	15	26,983,628,831	49,843	40,669	9,173	18,743	10,847	30,421	
std	12	58,687,616,358	114,676	105,257	10,702	58,855	33,561	67,078	
min	1	4,537,422	911	446	465	1	3	54	
25%	6	514,534,600	6,095	2,598	2,936	242	45	2,020	
50%	10	4,063,042,924	14,560	8,744	5,674	924	392	8,399	

```
sns.boxplot(x=st_data_2009_clean["High-technology exports (% of manufactured exports)"]);
```

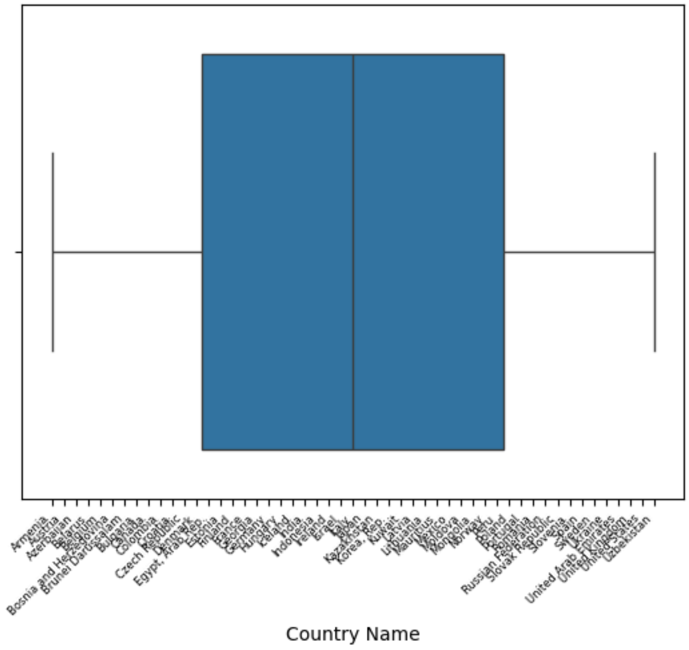


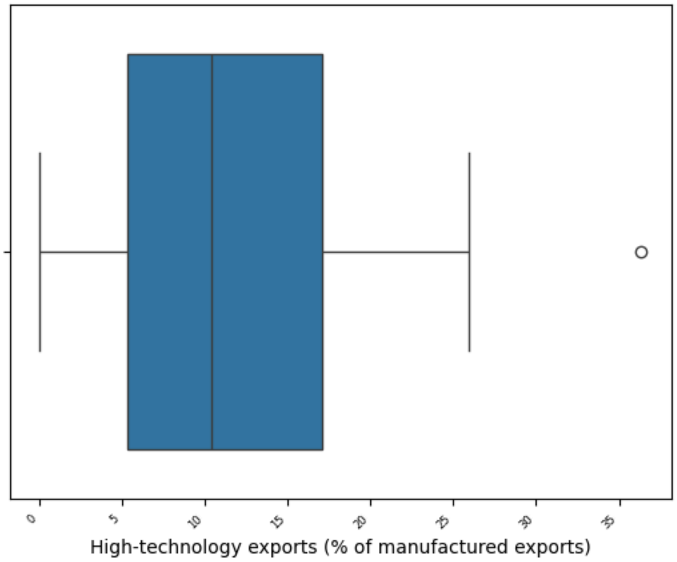
```
sns.boxplot(x=st_data_2018_clean["High-technology exports (% of manufactured exports)"]);
```

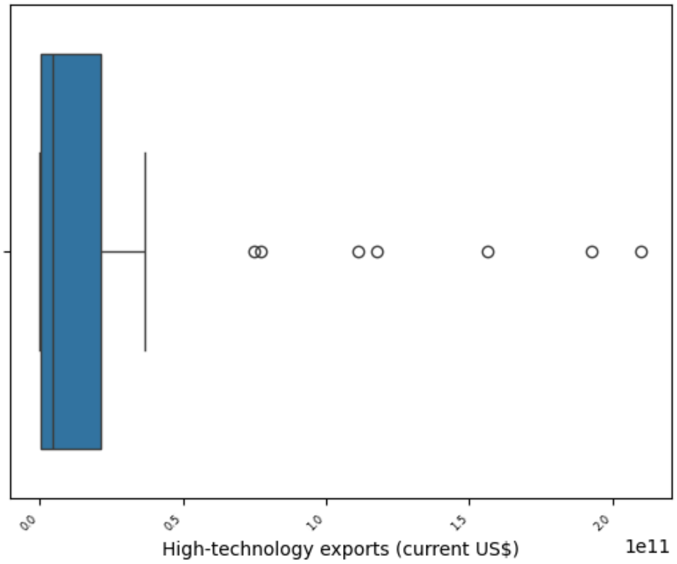


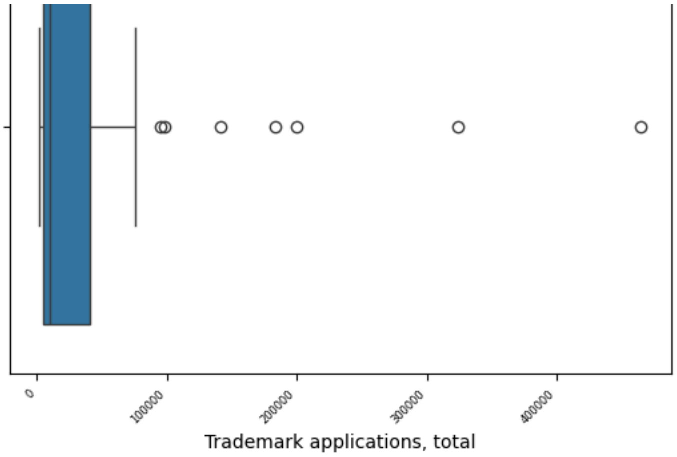
Outliers

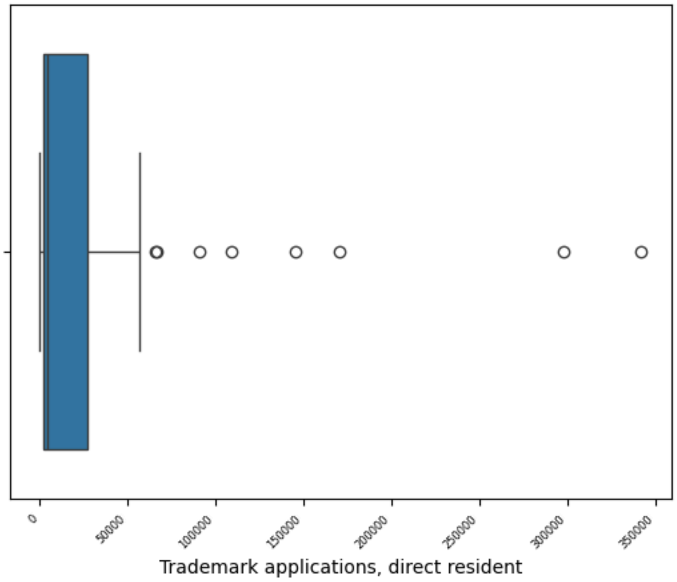
```
columns = st_data_2018_clean.columns

for col in columns:
  sns.boxplot(x=st_data_2018_clean[col])
  plt.xticks(rotation=45, ha='right', fontsize=6)  # Added fontsize=8
  plt.show()
```

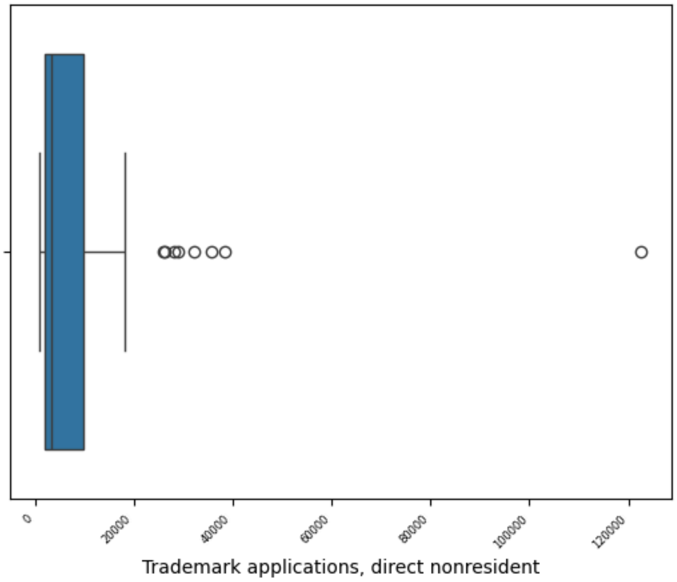

Armenia
Austria
Azerbaijan
Belarus
Belgium
Bosnia and Herzegovina
Brunei Darussalam
Bulgaria
Canada
Colombia
Croatia
Czech Republic
Denmark
Egypt, Arab Rep.
Estonia
Finland
France
Georgia
Germany
Hungary
Iceland
India
Indonesia
Ireland
Israel
Italy
Japan
Kazakhstan
Korea, Rep.
Kuwait
Latvia
Lithuania
Mauritius
Mexico
Moldova
Mongolia
Norway
Peru
Poland
Portugal
Romania
Russian Federation
Slovak Republic
Slovenia
Spain
Sweden
Ukraine
United Arab Emirates
United Kingdom
United States
Uzbekistan
Country Name

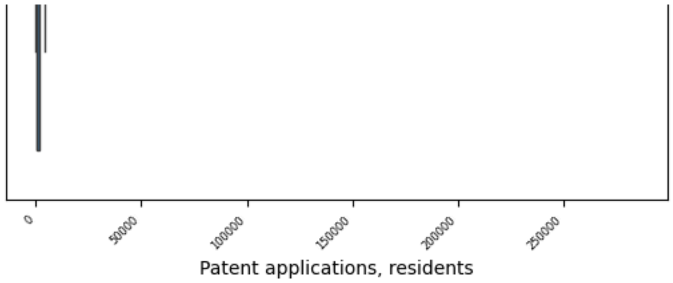

0
5
10
15
20
25
30
35
High-technology exports (% of manufactured exports)

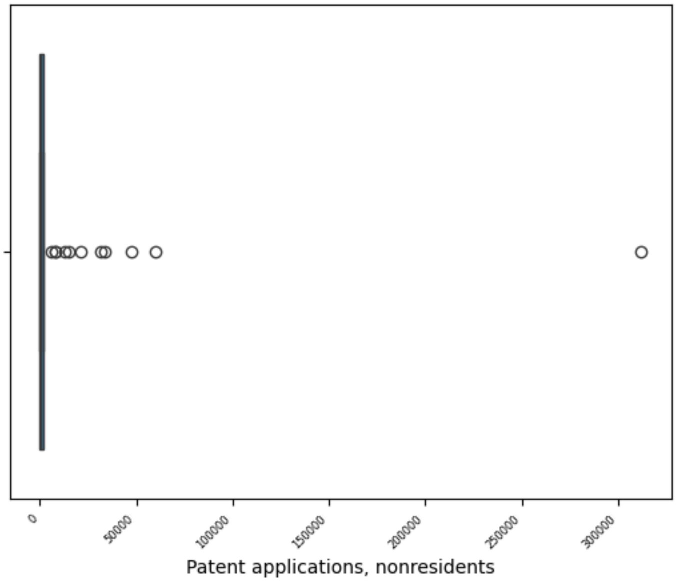

0.0
0.5
1.0
1.5
2.0
1e11
High-technology exports (current US$)

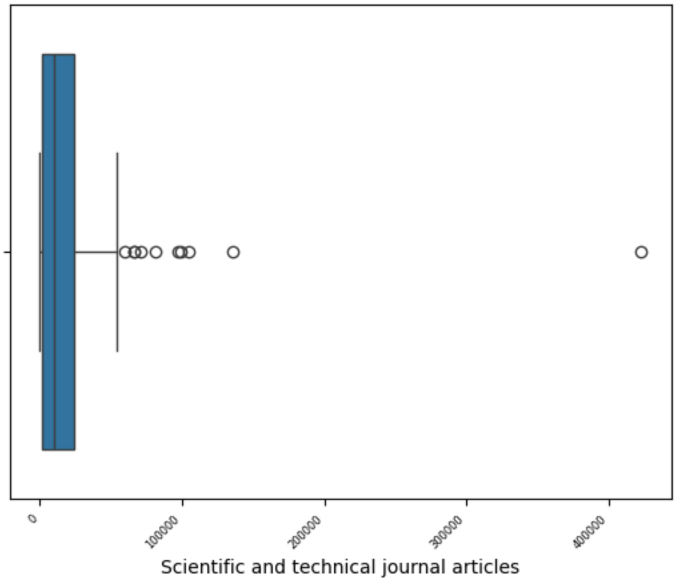

0
100000
200000
300000
400000
Trademark applications, total

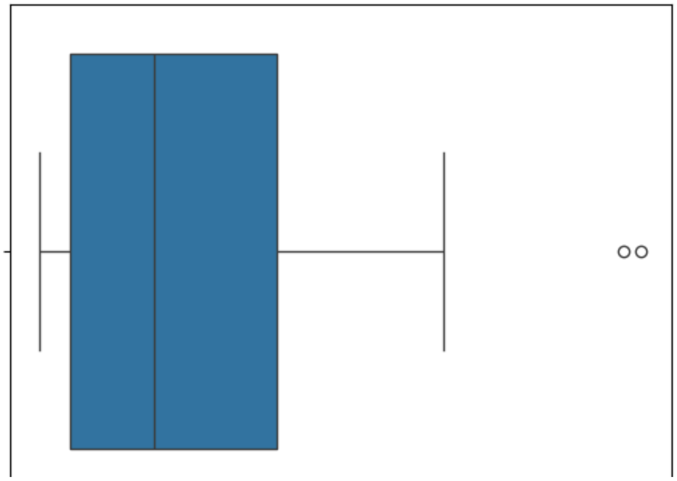

0
50000
100000
150000
200000
250000
300000
350000
Trademark applications, direct resident

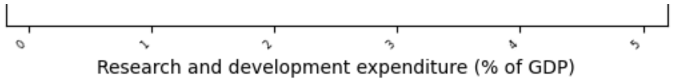

0
20000
40000
60000
80000
100000
120000
Trademark applications, direct nonresident


0
50000
100000
150000
200000
250000
Patent applications, residents


0
50000
100000
150000
200000
250000
300000
Patent applications, nonresidents


0
100000
200000
300000
400000
Scientific and technical journal articles




0
1
2
3
4
5
Research and development expenditure (% of GDP)

```
article_max = st_data_2018_clean["Scientific and technical journal articles"].max()
article_max
```

```
422807.71
```

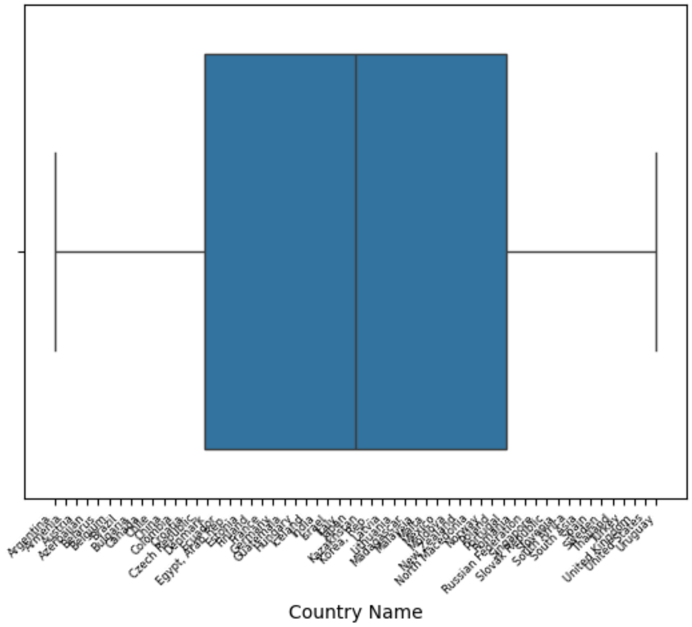
```
st_data_2018_clean[st_data_2018_clean["Scientific and technical journal articles"] == article_max]
```

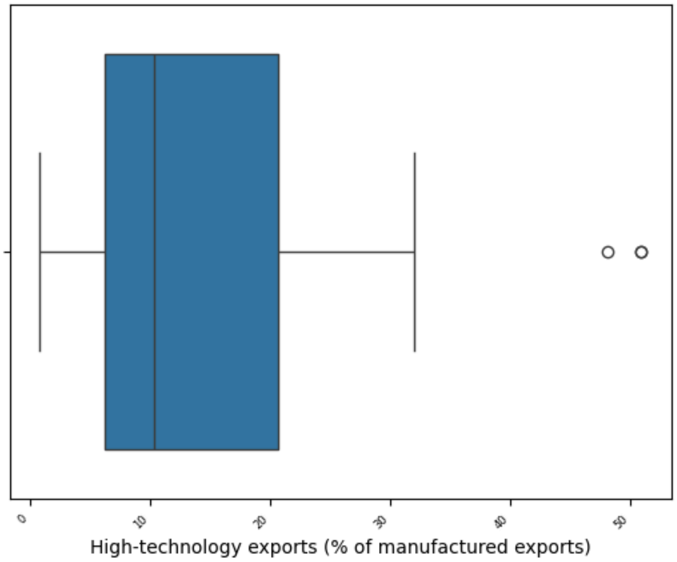
	Country Name	High-technology exports (% of manufactured exports)	High-technology exports (current US$)	Trademark applications, total	Trademark applications, direct resident	Trademark applications, direct nonresident	Patent applications, residents	Patent applications, nonresidents	Scienti techni jour artic

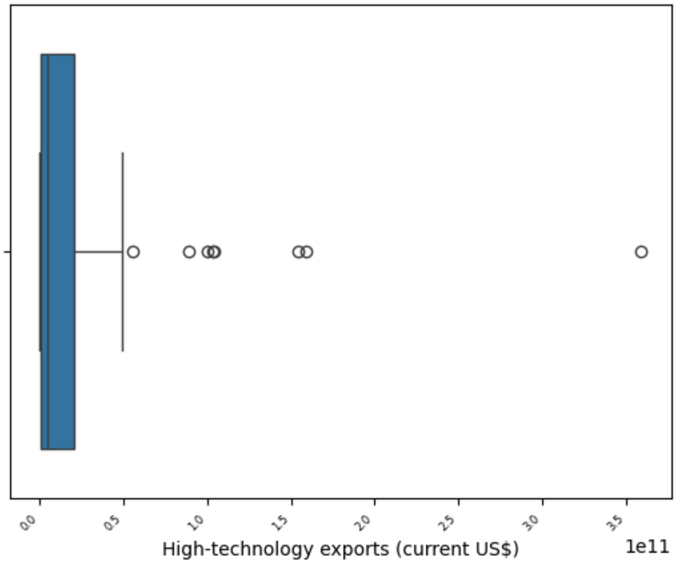
```
st_data_2018_clean[st_data_2018_clean["Scientific and technical journal articles"] > 60000]
```

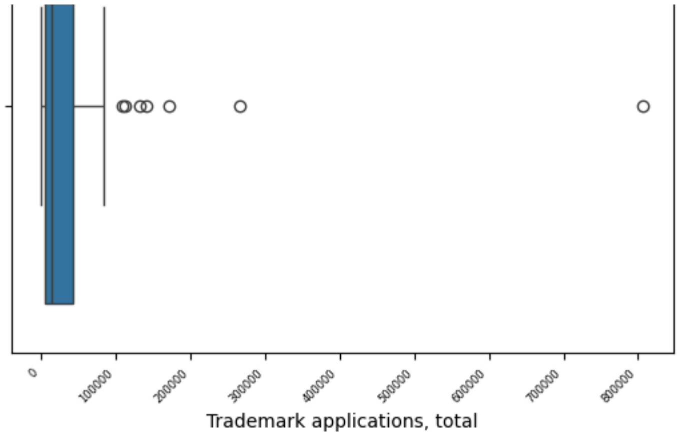
	Country Name	High-technology exports (% of manufactured exports)	High-technology exports (current US$)	Trademark applications, total	Trademark applications, direct resident	Trademark applications, direct nonresident	Patent applications, residents	Patent applications, nonresidents	Scient techn jou arti
68	France	25.920	117814412441.000	98279.000	90581.000	7698.000	14303.000	1919.000	6635
73	Germany	15.778	210082307180.000	75236.000	65686.000	9550.000	46617.000	21281.000	10439
89	India	9.008	20273090235.000	323970.000	297750.000	26220.000	16289.000	33766.000	13578
96	Italy	7.505	32581025234.000	42580.000	37320.000	5260.000	8921.000	900.000	7124
98	Japan	17.268	111020443595.000	183657.000	145269.000	38388.000	253630.000	59937.000	9879
104	Korea, Rep.	36.347	192789656676.000	199476.000	170541.000	28935.000	162561.000	47431.000	6637
161	Russian Federation	10.963	10183007833.000	75081.000	49132.000	25949.000	24926.000	13031.000	8157
206	United	22.643	76926541023.000	94915.000	66833.000	28082.000	12865.000	8076.000	9768

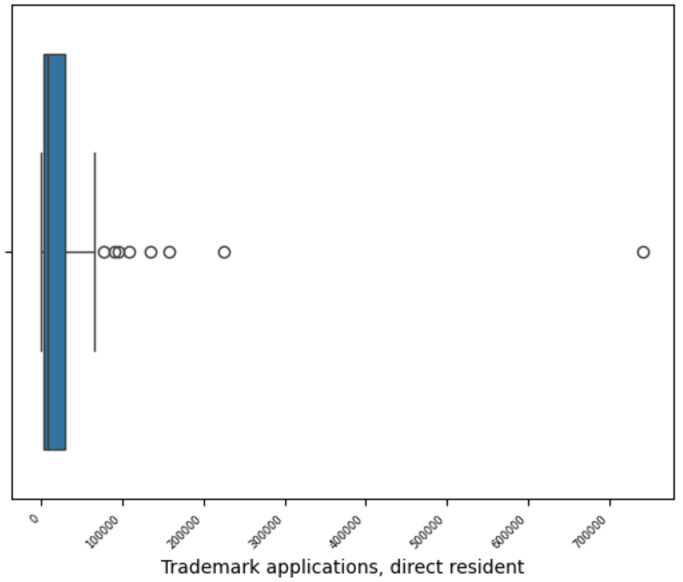
```
columns = st_data_2009_clean.columns

for col in columns:
  sns.boxplot(x=st_data_2009_clean[col])
  plt.xticks(rotation=45, ha='right', fontsize=6)  # Added fontsize=8
  plt.show()
```

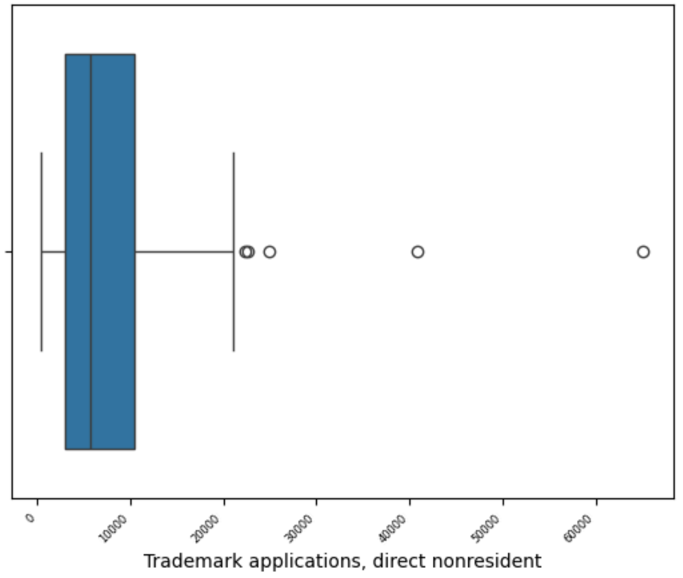

Argentina
Armenia
Austria
Azerbaijan
Belarus
Belgium
Brazil
Bulgaria
Canada
Chile
China
Colombia
Croatia
Czech Republic
Denmark
Egypt, Arab Rep.
Estonia
Finland
France
Germany
Guatemala
Hungary
Iceland
India
Israel
Italy
Japan
Kazakhstan
Korea, Rep.
Latvia
Lithuania
Madagascar
Malaysia
Malta
Mexico
Moldova
New Zealand
North Macedonia
Norway
Poland
Portugal
Romania
Russian Federation
Singapore
Slovak Republic
Slovenia
South Africa
South Asia
Spain
Sweden
Thailand
Turkey
United Kingdom
United States
Uruguay
Country Name


0
10
20
30
40
50
High-technology exports (% of manufactured exports)


0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1e11
High-technology exports (current US$)


0
100000
200000
300000
400000
500000
600000
700000
800000
Trademark applications, total


0
100000
200000
300000
400000
500000
600000
700000
Trademark applications, direct resident


0
10000
20000
30000
40000
50000
60000
Trademark applications, direct nonresident