Investigate a Dataset [Gapminder World]

August 13, 2020

1 Project: Dataset Investigation - Gapminder World

1.1 Table of Contents

Introduction

The Gapminder Foundation is a non-profit that promotes sustainable development by increased use of, and understanding of statistics. The organisation gathers information about how people live in different countries, tracked across the years, and on a number of different indicators.

For this project four variables are investigated, namely income per person (GDP/capita, PPP$ inflation-adjusted), fixed line subsribers (per 100 people) , cell phone (per 100 people) and broadband subscribers (per 100 people). An additional dataset is used to supplement country level geographical data. Further details on the aforementioned metrics, and how they were collected can be found in the links to the references sections of this report.

To improve readability the variables are abbreviated as follows:

income per person (GDP/capita, PPP$ inflation-adjusted) : **income**

fixed line subsribers (per 100 people) : **fixed**

cell phone (per 100 people) : **phone**

broadband subscribers (per 100 people) : **broadband**

A baseline understanding is first drawn by asking how each of the variables has changed for the period recorded. Further granularity is added to the analysis by grouping countries by continent. In addition to scrutinizing each variable indivually, the variables are evaluated in relation to each other.

[37]:
```
# import packages
import numpy as np
import pandas as pd
```

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

[38]:
```python
# plot visualizations in notebook environment
%matplotlib inline
```

Data Wrangling

1.1.1 General Properties

[39]:
```python
# Load data with four variables, plus supplementary dataset
df_income = pd.read_csv(r"C:\Users\noama\OneDrive\My␣
↪Documents\OneDrive\Udacity\Project␣
↪2\income_per_person_gdppercapita_ppp_inflation_adjusted.csv")
df_fixed = pd.read_csv(r"C:\Users\noama\OneDrive\My␣
↪Documents\OneDrive\Udacity\Project 2\fixed_line_subscribers_per_100_people.
↪csv")
df_phone = pd.read_csv(r"C:\Users\noama\OneDrive\My␣
↪Documents\OneDrive\Udacity\Project 2\cell_phones_per_100_people.csv")
df_broadband = pd.read_csv(r"C:\Users\noama\OneDrive\My␣
↪Documents\OneDrive\Udacity\Project 2\broadband_subscribers_per_100_people.
↪csv")
df_continent = pd.read_csv(r"C:\Users\noama\OneDrive\My␣
↪Documents\OneDrive\Udacity\Project 2\datasets_14947_19943_countryContinent.
↪csv", encoding="ISO-8859-1")
```

[40]:
```python
#validate income dataset with inspection of first five rows
df_income.head()
```

[40]:
```
       country  1800  1801  1802  1803  1804  1805  1806  1807  1808  …  \
0  Afghanistan   603   603   603   603   603   603   603   603   603  …
1      Albania   667   667   667   667   667   668   668   668   668  …
2      Algeria   715   716   717   718   719   720   721   722   723  …
3      Andorra  1200  1200  1200  1200  1210  1210  1210  1210  1220  …
4       Angola   618   620   623   626   628   631   634   637   640  …

    2031   2032   2033   2034   2035   2036   2037   2038   2039   2040
0   2550   2600   2660   2710   2770   2820   2880   2940   3000   3060
1  19400  19800  20200  20600  21000  21500  21900  22300  22800  23300
2  14300  14600  14900  15200  15500  15800  16100  16500  16800  17100
3  73600  75100  76700  78300  79900  81500  83100  84800  86500  88300
4   6110   6230   6350   6480   6610   6750   6880   7020   7170   7310

[5 rows x 242 columns]
```

[41]:
```python
#validate fixed dataset with inspection of first five rows
df_fixed.head()
```

[41]:
```
       country    1960  1961  1962  1963  1964    1965  1966  1967  1968  …  \
0  Afghanistan  0.0856   NaN   NaN   NaN   NaN  0.0934   NaN   NaN   NaN  …
1      Albania  0.4180   NaN   NaN   NaN   NaN  0.7380   NaN   NaN   NaN  …
2      Algeria     NaN   NaN   NaN   NaN   NaN  0.5790   NaN   NaN   NaN  …
3      Andorra     NaN   NaN   NaN   NaN   NaN  2.7000   NaN   NaN   NaN  …
4       Angola  0.1220   NaN   NaN   NaN   NaN  0.1730   NaN   NaN   NaN  …

      2009    2010     2011    2012    2013    2014   2015    2016    2017  \
0   0.0181   0.057   0.0449   0.289   0.297   0.305   0.32   0.323   0.327
1  12.2000  11.300  11.6000  10.700   9.680   8.140   7.84   8.610   8.550
2   7.2900   8.120   8.3400   8.800   8.210   7.960   8.22   8.400   9.910
3  44.9000  45.200  45.9000  46.500  47.800  48.300  49.80  50.100  49.900
4   1.3500   1.200   0.6580   0.830   0.826   1.070   1.02   1.060   0.540

     2018
0   0.344
1   8.620
2   9.950
3  51.100
4   0.558

[5 rows x 60 columns]
```

[42]:
```
#validate phone dataset with inspection of first five rows
df_phone.head()
```

[42]:
```
       country  1960  1961  1962  1963  1964  1965  1966  1967  1968  …  \
0  Afghanistan   0.0   NaN   NaN   NaN   NaN   0.0   NaN   NaN   NaN  …
1      Albania   0.0   NaN   NaN   NaN   NaN   0.0   NaN   NaN   NaN  …
2      Algeria   0.0   NaN   NaN   NaN   NaN   0.0   NaN   NaN   NaN  …
3      Andorra   0.0   NaN   NaN   NaN   NaN   0.0   NaN   NaN   NaN  …
4       Angola   0.0   NaN   NaN   NaN   NaN   0.0   NaN   NaN   NaN  …

   2009  2010   2011   2012   2013   2014   2015   2016   2017   2018
0  37.0  35.0   45.8   49.2   52.1   55.2   57.3   61.1   65.9   59.1
1  82.9  91.3  106.0  120.0  127.0  116.0  118.0  117.0  126.0   94.2
2  92.6  91.1   97.1  100.0  104.0  111.0  109.0  116.0  111.0  112.0
3  76.4  77.6   77.7   77.5   79.1   83.6   91.4   98.5  104.0  107.0
4  36.0  40.3   49.8   50.9   51.1   52.2   49.8   45.1   44.7   43.1

[5 rows x 60 columns]
```

[43]:
```
#validate broadband dataset with inspection of first five row
df_broadband.head()
```

[43]:
```
       country  1998  1999  2000  2001  2002    2003     2004     2005  \
0  Afghanistan   NaN   NaN   NaN   NaN   NaN     NaN  0.00081  0.00086
```

```
1      Albania   NaN   NaN   NaN   NaN   NaN     NaN      NaN    0.00881
2      Algeria   NaN   NaN   NaN   NaN   NaN  0.0558  0.11000    0.40700
3      Andorra   NaN   NaN   NaN   NaN  1.64  4.9200  8.24000   13.10000
4       Angola   NaN   NaN   NaN   NaN   NaN     NaN      NaN        NaN

       2006  …      2009      2010     2011      2012      2013      2014  \
0   0.00189  …   0.00352   0.00514      NaN   0.00481   0.00465   0.00449
1       NaN  …   3.09000   3.58000   4.3800   5.49000   6.29000   7.18000
2   0.50500  …   2.32000   2.50000   2.6800   3.09000   3.36000   4.11000
3  18.00000  …  27.20000  29.00000  30.8000  32.60000  34.30000  36.30000
4   0.03700  …   0.06660   0.06420   0.0653   0.08170   0.08570   0.32600

      2015     2016     2017    2018
0   0.0205   0.0249   0.0463   0.043
1   8.4000   9.2300  10.5000  12.600
2   5.7100   7.0500   7.7600   7.260
3  39.3000  42.0000  44.5000  46.300
4   0.5510   0.2940   0.3250   0.356

[5 rows x 22 columns]
```

[44]:
```
#validate continent dataset with inspection of first five rows
df_continent.head()
```

[44]:
```
          country code_2 code_3  country_code     iso_3166_2 continent  \
0     Afghanistan     AF    AFG             4  ISO 3166-2:AF      Asia
1   Åland Islands     AX    ALA           248  ISO 3166-2:AX    Europe
2         Albania     AL    ALB             8  ISO 3166-2:AL    Europe
3         Algeria     DZ    DZA            12  ISO 3166-2:DZ    Africa
4  American Samoa     AS    ASM            16  ISO 3166-2:AS   Oceania

         sub_region  region_code  sub_region_code
0     Southern Asia        142.0             34.0
1   Northern Europe        150.0            154.0
2   Southern Europe        150.0             39.0
3   Northern Africa          2.0             15.0
4         Polynesia          9.0             61.0
```

1.1.2 Observations:

-income, fixed, phone and broadband datasets need to be tidied so that each row is an observation and each column a variable.

[45]:
```
#reshape income dataframe from wide to long format
df_income = pd.melt(df_income, id_vars='country', var_name='year',␣
↪value_name='income')
df_income.head()
```

[45]:

	country	year	income
0	Afghanistan	1800	603
1	Albania	1800	667
2	Algeria	1800	715
3	Andorra	1800	1200
4	Angola	1800	618

[46]:
```
#reshape fixed dataframe from wide to long format
df_fixed = pd.melt(df_fixed, id_vars='country', var_name='year',␣
↪value_name='fixed')
df_fixed.head()
```

[46]:

	country	year	fixed
0	Afghanistan	1960	0.0856
1	Albania	1960	0.4180
2	Algeria	1960	NaN
3	Andorra	1960	NaN
4	Angola	1960	0.1220

[47]:
```
#reshape phone dataframe from wide to long format
df_phone = pd.melt(df_phone, id_vars='country', var_name='year',␣
↪value_name='phone')
df_phone.head()
```

[47]:

	country	year	phone
0	Afghanistan	1960	0.0
1	Albania	1960	0.0
2	Algeria	1960	0.0
3	Andorra	1960	0.0
4	Angola	1960	0.0

[48]:
```
#reshape broadband dataframe from wide to long format
df_broadband = pd.melt(df_broadband, id_vars='country', var_name='year',␣
↪value_name='broadband')
df_broadband.head()
```

[48]:

	country	year	broadband
0	Afghanistan	1998	NaN
1	Albania	1998	NaN
2	Algeria	1998	NaN
3	Andorra	1998	NaN
4	Angola	1998	NaN

[49]:
```
#diplay number of columns and rows in each dataset
print("The income dataset contains " + str(df_income.shape[0]) + " rows and " +␣
↪str(df_income.shape[1]) + " columns")
```

```
print("The fixed dataset contains " + str(df_fixed.shape[0]) + " rows and " +␣
↪str(df_fixed.shape[1]) + " columns")
print("The phone dataset contains " + str(df_phone.shape[0]) + " rows and " +␣
↪str(df_phone.shape[1]) + " columns")
print("The broadband dataset contains " + str(df_broadband.shape[0]) + " rows␣
↪and " + str(df_broadband.shape[1]) + " columns")
print("The continent dataset contains " + str(df_continent.shape[0]) + " rows␣
↪and " + str(df_continent.shape[1]) + " columns")
```

```
The income dataset contains 46513 rows and 3 columns
The fixed dataset contains 11446 rows and 3 columns
The phone dataset contains 11446 rows and 3 columns
The broadband dataset contains 4032 rows and 3 columns
The continent dataset contains 249 rows and 9 columns
```

1.1.3 Observations:

-The income dataset contains the highest number of recorded observations, while the reverse is true for the broadband dataset (excluding the supplementary dataset). Said otherwise, income has the longest recorded history of obervations of the four variables. To facilitate a direct comparison between variables the merged dataset is trimmed to the number of observations recored in the smallest dataset (broadband).

[50]:
```
#join income dataset & fixed datasets together
income_fixed = df_income.merge(df_fixed, on=['country', 'year'])
income_fixed.head(1)
```

[50]:

	country	year	income	fixed
0	Afghanistan	1960	2740	0.0856

[51]:
```
#add phone variable to dataset containing income & fixed variables
fixed_phone = income_fixed.merge(df_phone, on=['country', 'year'])
fixed_phone.head(1)
```

[51]:

	country	year	income	fixed	phone
0	Afghanistan	1960	2740	0.0856	0.0

[52]:
```
#add broadband variable to dataset containing income, fixed and phone variables␣
↪
phone_broadband = fixed_phone.merge(df_broadband, on=['country', 'year'])
phone_broadband.head(1)
```

[52]:

	country	year	income	fixed	phone	broadband
0	Afghanistan	1998	800	0.147	0.0	NaN

[53]:

```
#add continent data to dataset containing income, fixed, phone and broadband␣
↪variables
df_all = phone_broadband.merge(df_continent, on='country')
df_all.head(1)
```

[53]:
```
       country  year  income  fixed  phone  broadband code_2 code_3  \
0  Afghanistan  1998     800  0.147    0.0        NaN     AF    AFG

   country_code     iso_3166_2 continent     sub_region  region_code  \
0             4  ISO 3166-2:AF      Asia  Southern Asia        142.0

   sub_region_code
0             34.0
```

[54]:
```
#reorder columns and retain only those of interest
df = df_all.loc[:,['country', 'continent', 'year', 'income', 'fixed', 'phone',␣
↪'broadband']]
df.head(1)
```

[54]:
```
       country continent  year  income  fixed  phone  broadband
0  Afghanistan      Asia  1998     800  0.147    0.0        NaN
```

[55]:
```
#display column names, dtype and number of missing values
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3486 entries, 0 to 3485
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   country    3486 non-null   object
 1   continent  3486 non-null   object
 2   year       3486 non-null   object
 3   income     3486 non-null   int64
 4   fixed      3394 non-null   float64
 5   phone      3406 non-null   float64
 6   broadband  2586 non-null   float64
dtypes: float64(3), int64(1), object(3)
memory usage: 217.9+ KB
```

1.1.4 Observations:

-The merged dataset contains 3486 rows (observations) and 7 columns (variables).

-The year variable can be converted to a data time object.

-fixed, phone and broadband variables contain missing values.

-income, fixed, phone, and broadband are quantitative variables that can be numerically analysed.

[56]:
```
#count number of missing values for fixed, phone and broadband columns
df.isnull().sum()
```

[56]:
```
country        0
continent      0
year           0
income         0
fixed         92
phone         80
broadband    900
dtype: int64
```

1.1.5 Data Cleaning

[57]:
```
# convert year column from object to date time object
df['year'] = pd.to_datetime(df.year)
```

[58]:
```
#drop observatitions with missing values across multiple variables
df.dropna(how='all', subset=['fixed', 'phone', 'broadband'], inplace=True)
df.isnull().sum()
```

[58]:
```
country        0
continent      0
year           0
income         0
fixed         30
phone         18
broadband    838
dtype: int64
```

1.1.6 Obervations:

-the variables fixed, phone and broadband now contain fewer variables that can be replaced.

[59]:
```
#fill missing values using interpolation
df.interpolate(method='linear', axis=0, inplace=True)
df.tail()
```

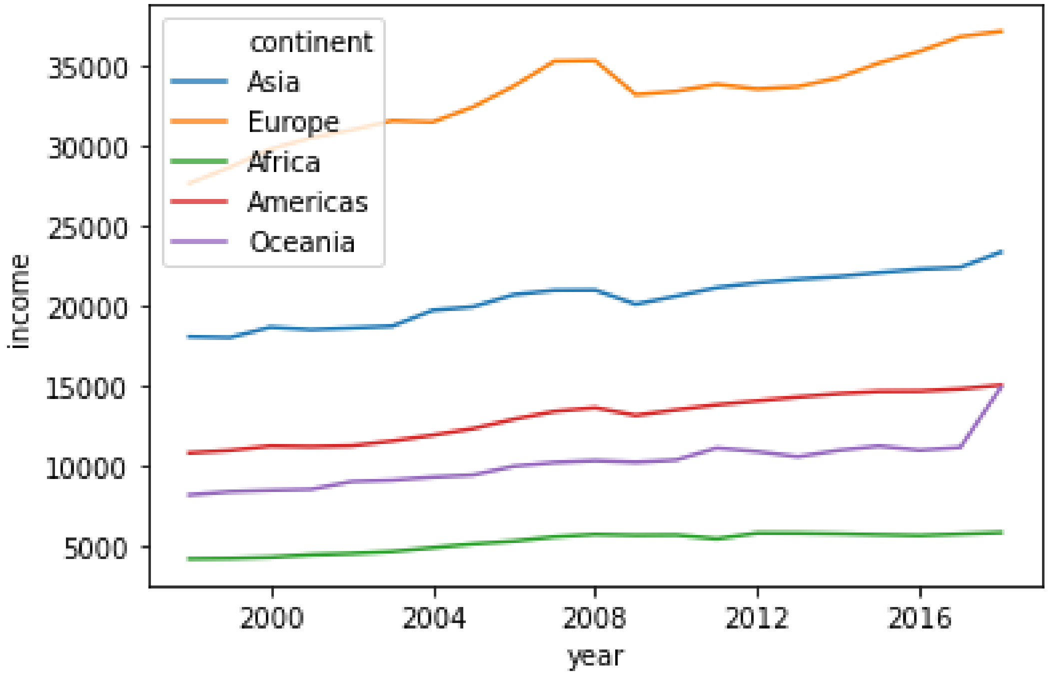
[59]:

	country	continent	year	income	fixed	phone	broadband
3481	Zimbabwe	Africa	2014-01-01	2510	2.42	86.8	1.12
3482	Zimbabwe	Africa	2015-01-01	2510	2.42	92.3	1.19
3483	Zimbabwe	Africa	2016-01-01	2490	2.18	91.8	1.22
3484	Zimbabwe	Africa	2017-01-01	2570	1.86	99.0	1.32
3485	Zimbabwe	Africa	2018-01-01	2620	1.86	89.4	1.41

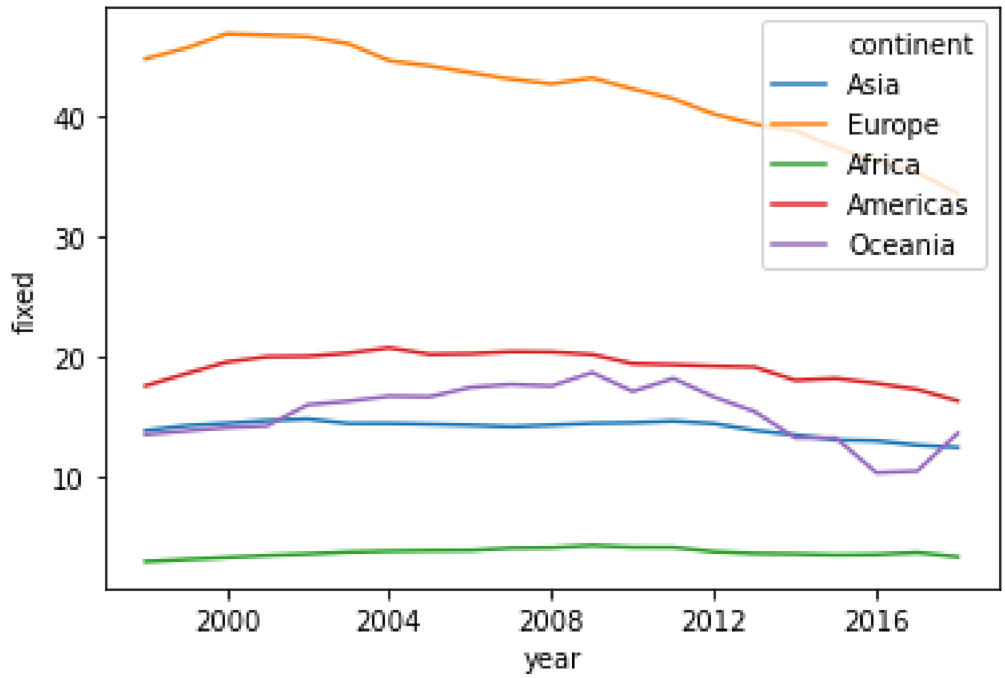
Exploratory Data Analysis

1.1.7 Research Question 1: How have income, fixed, phone and broadband changed over time?

[60]:
```
#line plot of the variable income by continent
sns.lineplot('year', 'income', data=df, hue='continent', ci=None);
```

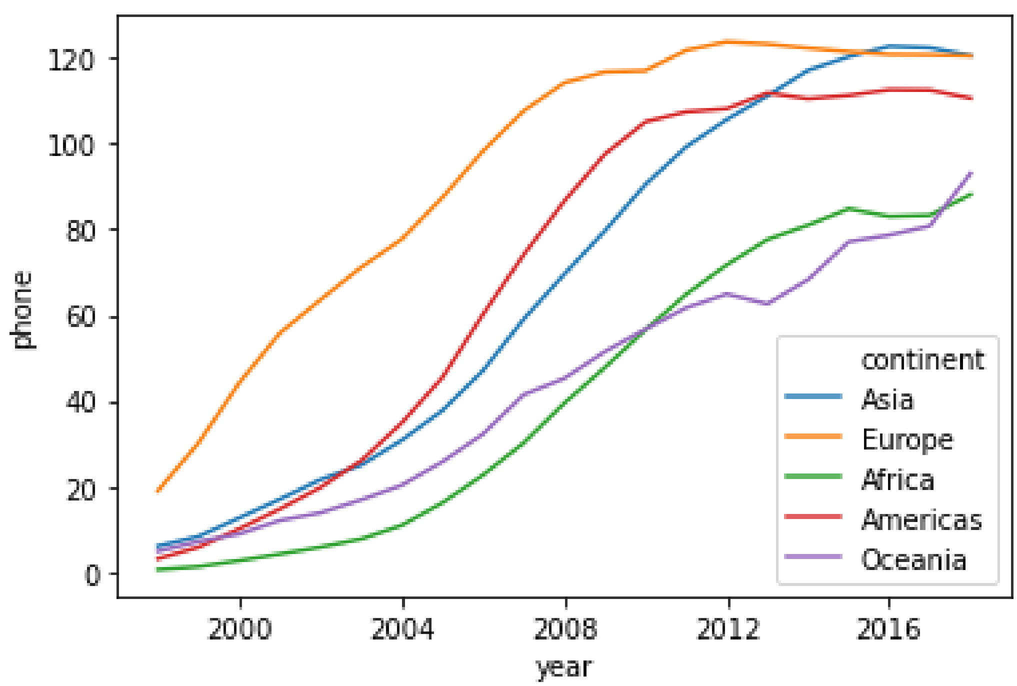


[61]:
```
#line plot of the variable fixed by continent
sns.lineplot('year', 'fixed', data=df, hue='continent', ci=None);
```

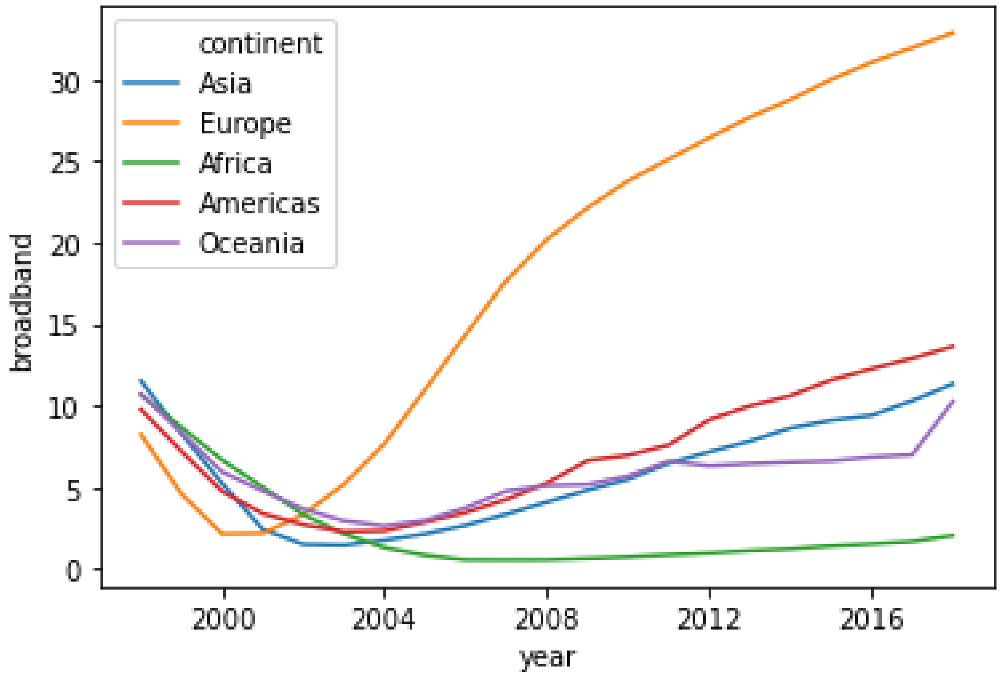


[62]:

```
#line plot of the variable phone by continent
sns.lineplot('year', 'phone', data=df, hue='continent', ci=None);
```



[63]:
```
#line plot of the variable broadband by continent
sns.lineplot('year', 'broadband', data=df, hue='continent', ci=None);
```



1.1.8 Observations:

-income has grown consistently from the turn of the century for all continents, albeit at different rates

-fixed line connections have either remained steady or declined as a channel of communcation for the time period recorded

-phone connections saw explosive growth over the same time period, with some markets beginning to show evidence of saturation

-broadband connections declined at the beginning of the century, before recovering. Europe, again is at the forefront of this uptick

1.1.9 Research Question 2: What is the shape of the distribution for the latest year for which data is available?

[64]:
```
#subset dataframe for year of interest (2018)
df_2018 = df.loc[df.year == '2018', :]
df_2018.head()
```

[64]:

	country	continent	year	income	fixed	phone	broadband
20	Afghanistan	Asia	2018-01-01	1740	0.344	59.1	0.043

```
41        Albania    Europe 2018-01-01   12300   8.620   94.2     12.600
62        Algeria    Africa 2018-01-01   13900   9.950  112.0      7.260
83        Andorra    Europe 2018-01-01   51500  51.100  107.0     46.300
104        Angola    Africa 2018-01-01    5730   0.558   43.1      0.356
```

[65]:
```
#summary stats on numerical columns for 2018
df_2018.describe()
```

[65]:

	income	fixed	phone	broadband
count	147.000000	147.000000	147.000000	147.000000
mean	19908.462585	15.693578	108.234354	14.314086
std	20422.548500	17.212123	34.232195	14.602554
min	629.000000	0.000000	27.400000	0.001820
25%	4335.000000	1.520000	87.250000	0.836500
50%	12800.000000	12.000000	113.000000	9.660000
75%	28750.000000	23.450000	132.000000	27.650000
max	113000.000000	112.000000	209.000000	51.200000

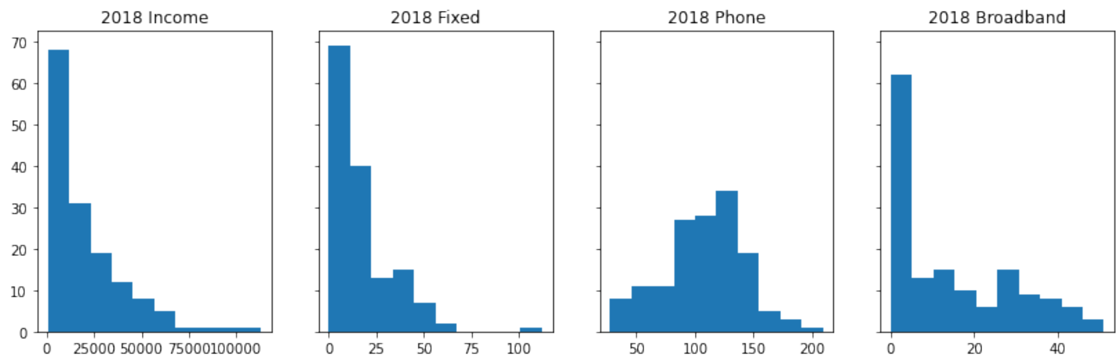
[66]:
```
# create array of values and assign to variable
w = df_2018.income.values
x = df_2018.fixed.values
y = df_2018.phone.values
z = df_2018.broadband.values

#create space for figure and subplots
fig, axs = plt.subplots(nrows=1, ncols=4, figsize= (14,4), sharey= True)

# draw histogram for each of the variables
axs[0].hist(w);
axs[1].hist(x);
axs[2].hist(y);
axs[3].hist(z);

#label title for each plot
axs[0].set_title('2018 Income')
axs[1].set_title('2018 Fixed')
axs[2].set_title('2018 Phone')
axs[3].set_title('2018 Broadband')
```

[66]:
```
Text(0.5, 1.0, '2018 Broadband')
```

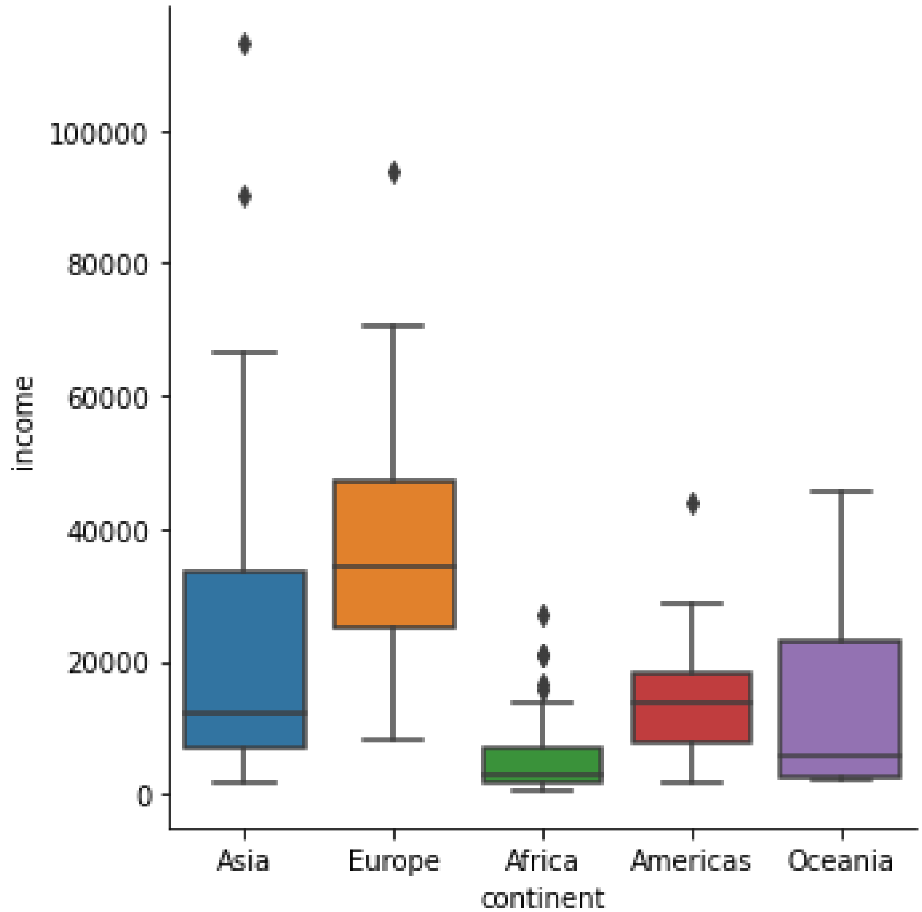


1.1.10 Observations:

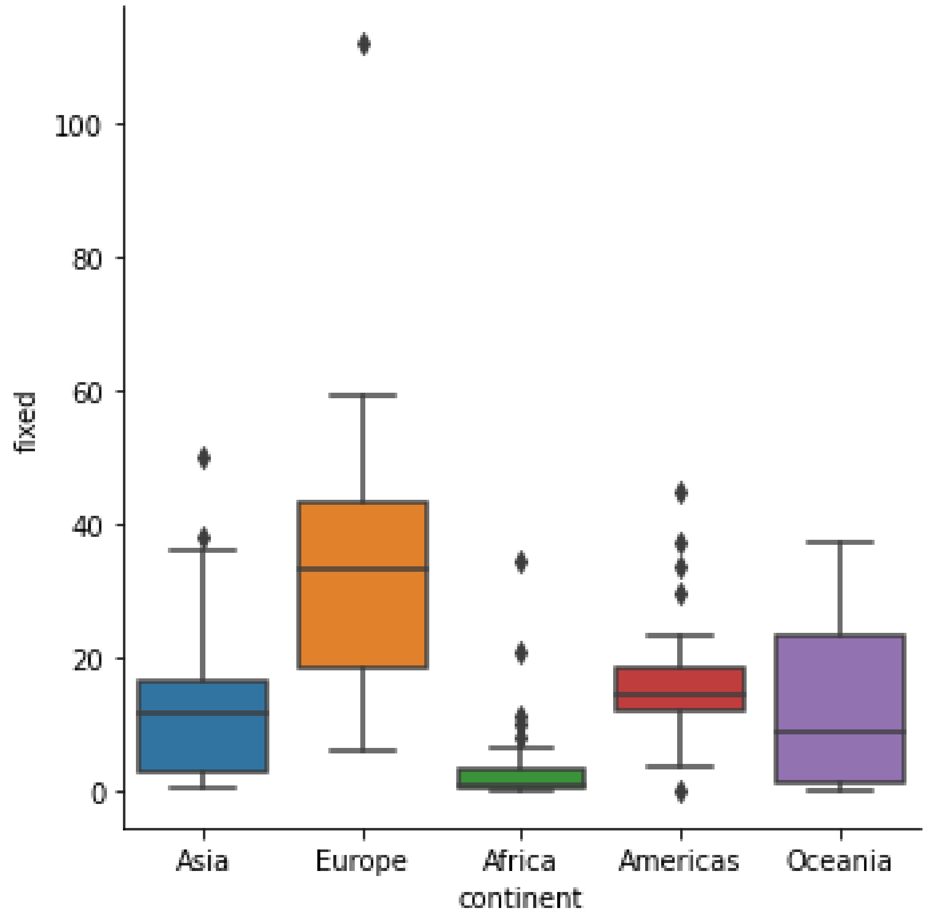
-The variables Income, Fixed and Broadband are skewed to the right, while the variable phone comes closest to resembling a normal distrubtion. Where skew is present, the median may be a more accurate measure of central tendency (most common value) than the mean.

1.1.11 Research Question 3: How does the shape of the distribution differ across geographical regions for 2018?

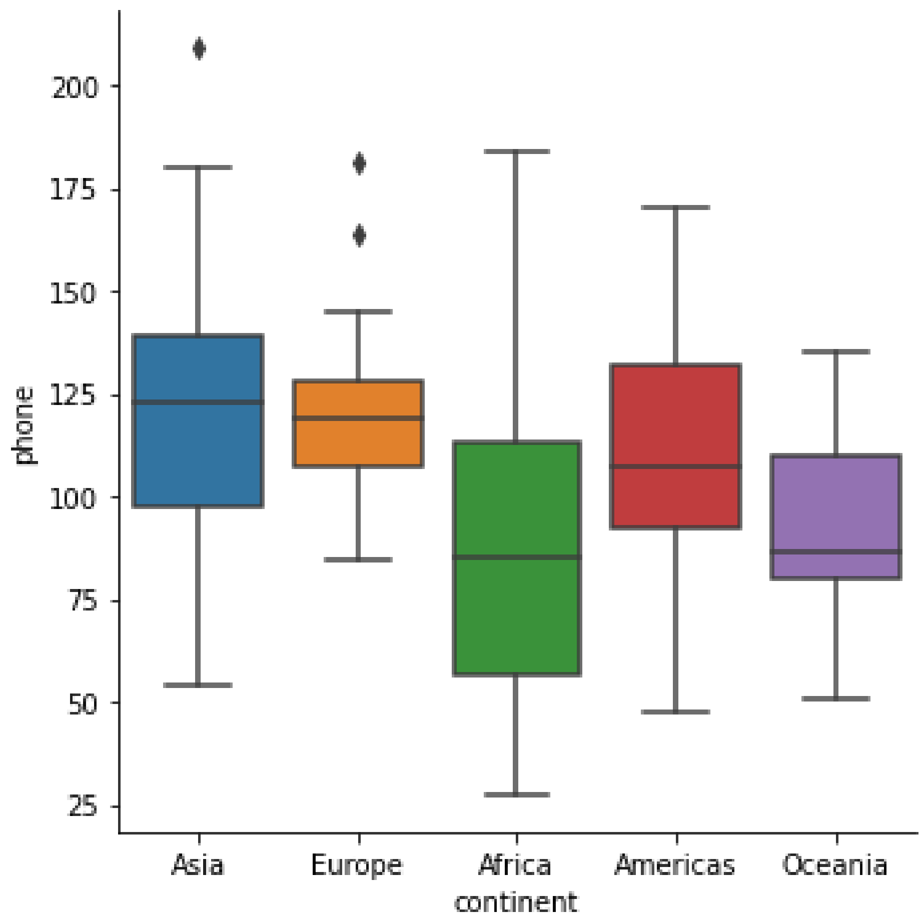
[67]:
```
#box plot of the variable income by continent
sns.catplot(x="continent", y="income", kind="box", data=df_2018);
```



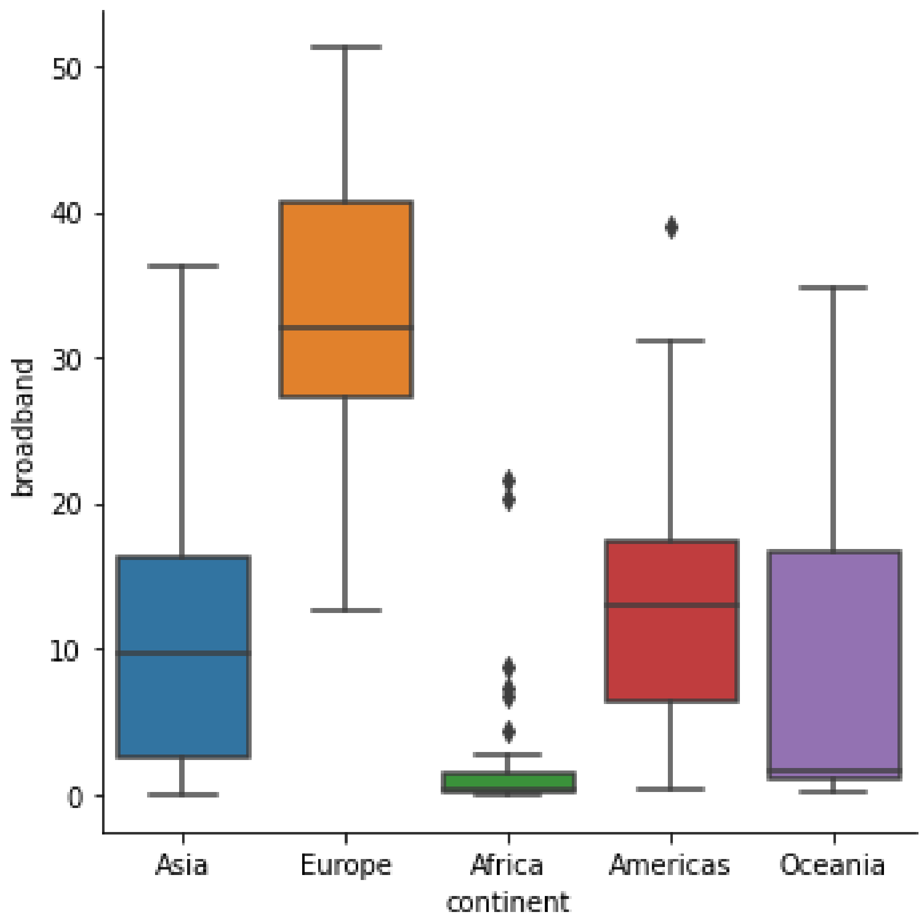
[68]:
```
#box plot of the variable fixed by continent
sns.catplot(x="continent", y="fixed", kind="box", data=df_2018);
```



[69]:
```
#box plot of the varible phone by continent
sns.catplot(x="continent", y="phone", kind="box", data=df_2018);
```




```
[70]: #box plot of the variable broadband by continent
sns.catplot(x="continent", y="broadband", kind="box", data=df_2018);
```



1.1.12 Observations:

-On average (median), Europe has the highest median income, rates of fixed, phone and braodband connections

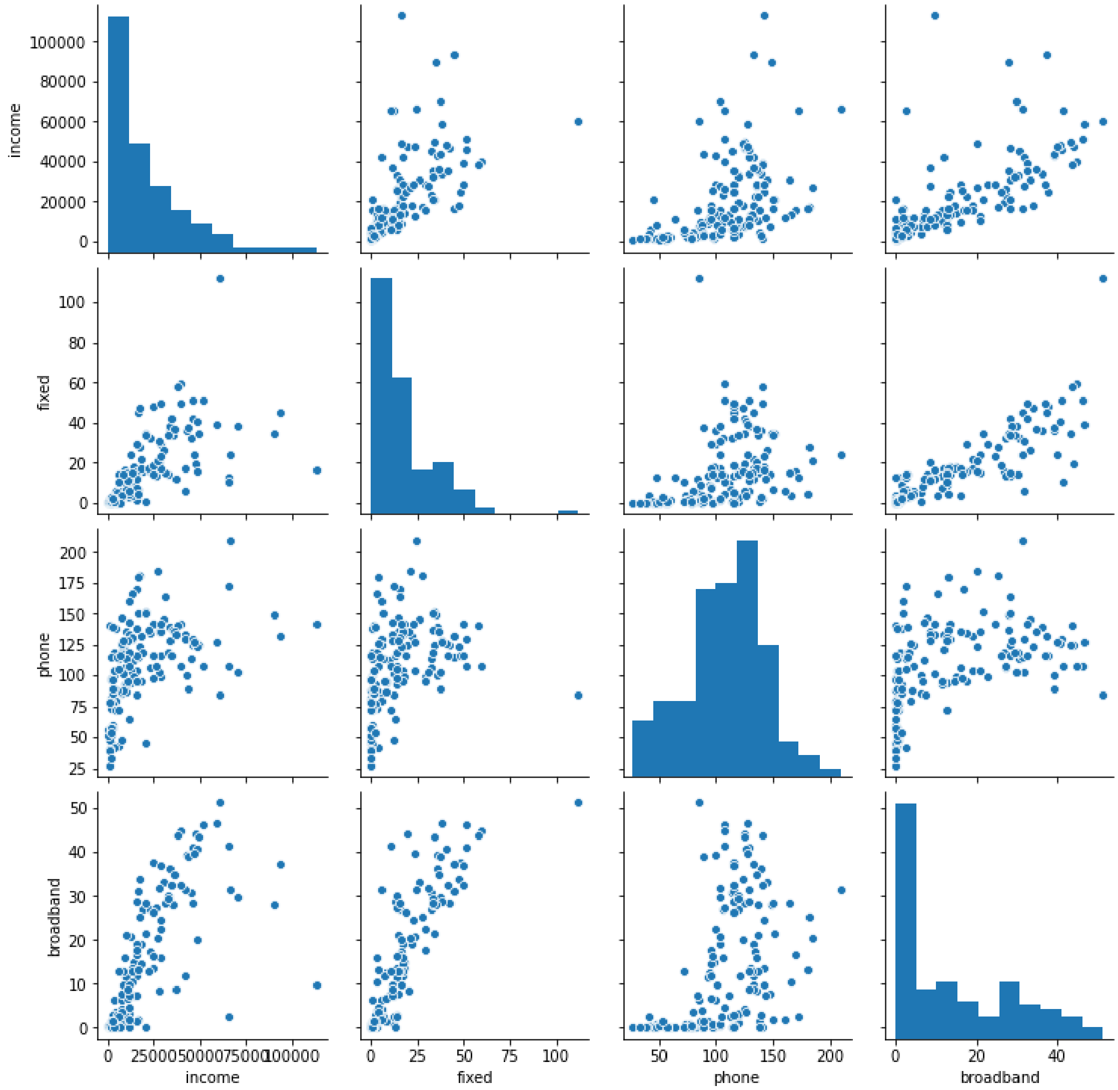
-The IQR (middle 50 of the distribution) for broadband and phone is wider than it is for fixed and income

-The variable Phone demonstrates the greatest uniformity across continents

-The presence of outliers

1.1.13 Research Question 4: What relationship, if any, is there between income, fixed, phone and broadband?

[71]:
```
sns.pairplot(df_2018);
```

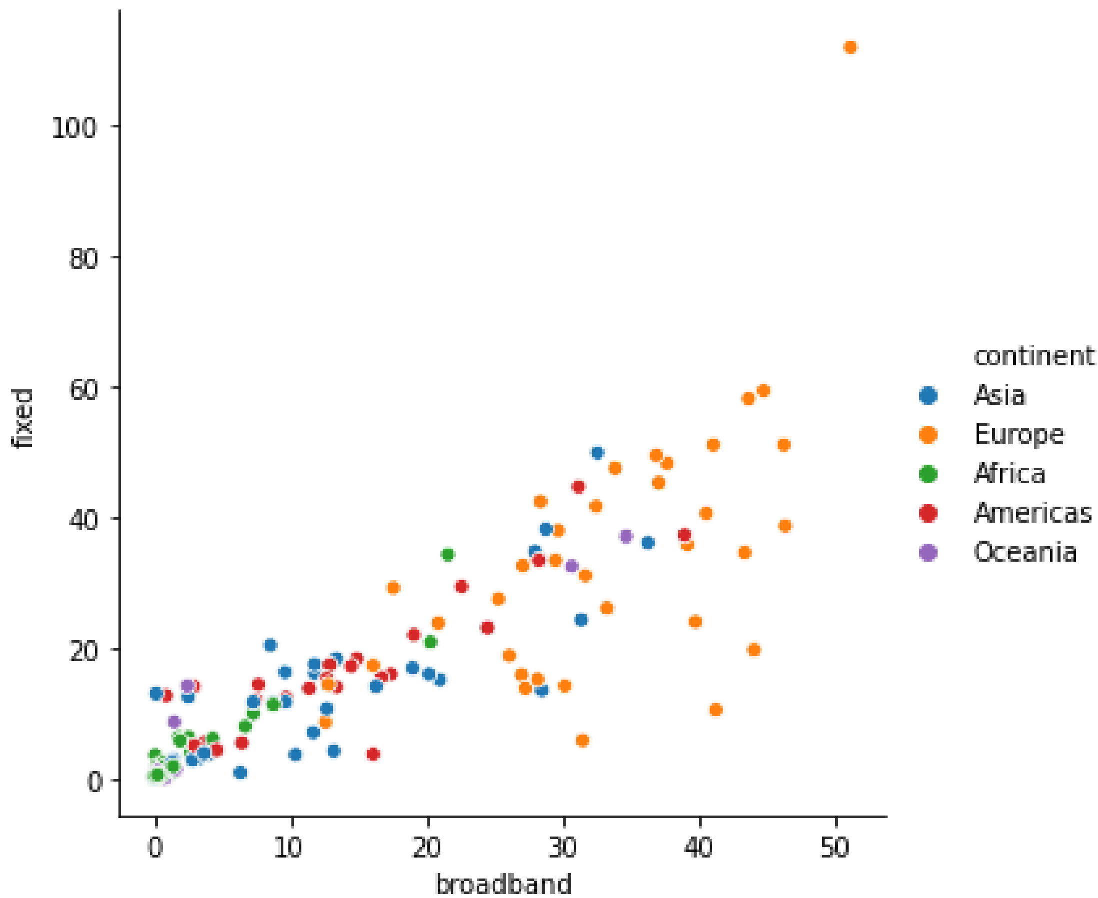


1.1.14 Observations:

-weak positive association betwen income and fixed and income and broadband

-moderate positive correlation between fixed and broadband

[72]:
```
#relationship plot for broadband and fixed by continent
sns.relplot(x="broadband", y="fixed", data=df_2018, hue='continent');
```



Conclusions

Summary

RQ1: 1. Income has grown steadily since across the world since the turn of the century. 2. Fixed line connections as a mode of communucation has declined across the world since the turn of the century. 3. After explosive growth early in the century, phone line connections have begun to show signs of saturation. 4. Broadband connections contiunue to grow as a mode of communcation across the globe.

RQ2: 1. The distribution of Income, fixed, and broadband connections is skewed to the right, implying the presence of a few countries significantly different from the rest of the world. 2. The distribution of phone line connections is most equitable for phone line connections.

RQ3: 1. Europe is a leader among the continents across all of the variables measured. 2. Africa is the only continent to display a decling trend in the number of broadband connections.

RQ4: 1. There is evidence for a positive correlation between the level of income a country possess and the number of fixed and broadband connections. 2. There is also evidence to suggest the number of fixed line connections a country possess is positively correlated with the number of number of broad band connections.

Limitations **Treatment of Missing Values**: Missing values were interpolated under the assumption that the realtionship is linear. Formal statistical techniques can be applied to assess the validity of such a claim.

Better yet, an investigation into the causes of the missing values may reveal systematic bias. In other words, an assessment could be made to to evaluate whether the values are missing at random. It may be that missing values are a placeholder for the value zero.For example it is etirely plausible that the missing value for Afganisain in 1998 under broadband connectivity is another way of stating that broadband was absent from the country at that moment in time. Replacing the missing value with the numeric zero would therefore be an accurate representation of reality.

Outlier Treatment: The numerical summary as well as the plot of distriubtions revealed outliers. Suffice it to say here that an entire literature has developed around investigating the cause, and proper treatment outliers.

References

1. https://www.gapminder.org/data/
2. https://www.gapminder.org/data/documentation/gd001/
3. https://data.worldbank.org/indicator/IT.NET.BBND.P2
4. https://data.worldbank.org/indicator/IT.CEL.SETS.P2
5. https://data.worldbank.org/indicator/IT.NET.USER.ZS