Data analysis and visualization on Global Internet Usage

This notebook will essentially be about analyzing the data about Global Internet Usage while getting insights on the same both visually and statistically. The aim of this notebook is to do a deep dive into this collection of data about Global Internet Usage.

How to run the code

This is an executable *Jupyter notebook* hosted on Jovian.ml, a platform for sharing data science projects. You can run and experiment with the code in a couple of ways: *using free online resources* (recommended) or *on your own computer*.

Option 1: Running using free online resources (1-click, recommended)

The easiest way to start executing this notebook is to click the "Run" button at the top of this page, and select "Run on Binder". This will run the notebook on mybinder.org, a free online service for running Jupyter notebooks. You can also select "Run on Colab" or "Run on Kaggle".

Option 2: Running on your computer locally

1. Install Conda by following these instructions. Add Conda binaries to your system `PATH`, so you can use the `conda` command on your terminal.
2. Create a Conda environment and install the required libraries by running these commands on the terminal:

```
conda create -n zerotopandas -y python=3.8
conda activate zerotopandas
pip install jovian jupyter numpy pandas matplotlib seaborn opendatasets --upgrade
```

3. Press the "Clone" button above to copy the command for downloading the notebook, and run it on the terminal. This will create a new directory and download the notebook. The command will look something like this:

```
jovian clone notebook-owner/notebook-id
```

4. Enter the newly created directory using `cd directory-name` and start the Jupyter notebook.

```
jupyter notebook
```

You can now access Jupyter's web interface by clicking the link that shows up on the terminal or by visiting http://localhost:8888 on your browser. Click on the notebook file (it has a `.ipynb` extension) to open it.

Downloading the Dataset

```
!pip install jovian opendatasets --upgrade --quiet
```

Let's begin by downloading the data, and listing the files within the dataset.

```
dataset_url = 'https://www.kaggle.com/datasets/sansuthi/gapminder-internet'
```

```
import opendatasets as od
od.download('https://www.kaggle.com/datasets/sansuthi/gapminder-internet')
```

```
Please provide your Kaggle credentials to download this dataset. Learn more:
http://bit.ly/kaggle-creds
Your Kaggle username: roctivmaina001
Your Kaggle Key: ·····
Downloading gapminder-internet.zip to ./gapminder-internet
100%|██████████| 4.65k/4.65k [00:00<00:00, 2.98MB/s]
```

The dataset has been downloaded and extracted.

```
data_dir = './gapminder-internet'
```

```
import os
os.listdir(data_dir)
```

```
['gapminder_internet.csv']
```

```
project_name =( "data analysis and visualization on Global Internet Usage ")
```

```
!pip install jovian --upgrade -q
```

```
import jovian
```

```
jovian.commit(project=project_name)
```

```
[jovian] Updating notebook "mainavictor004/data-analysis-and-visualization-on-global-internet-usage" on https://jovian.ai
[jovian] Committed successfully! https://jovian.ai/mainavictor004/data-analysis-and-visualization-on-global-internet-usage
```

```
'https://jovian.ai/mainavictor004/data-analysis-and-visualization-on-global-internet-usage'
```

Data Preparation and Cleaning

At this point in order to perform more operations and get insight into the data the data will be cleaned and restructured to meet the projects' needs.

```
import pandas as pd
```

```
usage_df=pd.read_csv('./gapminder-internet/gapminder_internet.csv')
```

```
usage_df
```

	country	incomeperperson	internetuserate	urbanrate
0	Afghanistan	NaN	3.654122	24.04
1	Albania	1914.996551	44.989947	46.72
2	Algeria	2231.993335	12.500073	65.22
3	Andorra	21943.339900	81.000000	88.92
4	Angola	1381.004268	9.999954	56.70
...	...	...	...	...
208	Vietnam	722.807559	27.851822	27.84
209	West Bank and Gaza	NaN	36.422772	71.90
210	Yemen, Rep.	610.357367	12.349750	30.64
211	Zambia	432.226337	10.124986	35.42
212	Zimbabwe	320.771890	11.500415	37.34

213 rows × 4 columns

```
usage_df= usage_df.fillna(0)
```

```
type(usage_df)
```

```
pandas.core.frame.DataFrame
```

```
usage_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 213 entries, 0 to 212
Data columns (total 4 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   country          213 non-null    object 
 1   incomeperperson  213 non-null    float64
 2   internetuserate  213 non-null    float64
 3   urbanrate        213 non-null    float64
dtypes: float64(3), object(1)
memory usage: 6.8+ KB
```

```
usage_df.shape
```

```
(213, 4)
```

```
usage_df.describe()
```

	incomeperperson	internetuserate	urbanrate
count	213.000000	213.000000	213.000000
mean	7797.105890	32.119631	54.104131
std	13738.698119	28.437108	26.203836
min	0.000000	0.000000	0.000000
25%	456.385712	6.497924	33.960000
50%	2161.546510	26.740025	56.700000
75%	8445.526689	51.958038	73.500000
max	105147.437700	95.638113	100.000000

```
usage_df.columns
```

```
Index(['country', 'incomeperperson', 'internetuserate', 'urbanrate'], dtype='object')
```

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "mainavictor004/data-analysis-and-visualization-on-global-internet-usage" on https://jovian.ai
[jovian] Committed successfully! https://jovian.ai/mainavictor004/data-analysis-and-visualization-on-global-internet-usage
```

```
'https://jovian.ai/mainavictor004/data-analysis-and-visualization-on-global-internet-usage'
```

Exploratory Analysis and Visualization

In this section a more indepth analaysis of the student performance will be evaluated.

```
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline

sns.set_style('darkgrid')
matplotlib.rcParams['font.size'] = 14
matplotlib.rcParams['figure.figsize'] = (9, 5)
matplotlib.rcParams['figure.facecolor'] = '#00000000'
```

```
usage_df
```

	country	incomeperperson	internetuserate	urbanrate
0	Afghanistan	0.000000	3.654122	24.04
1	Albania	1914.996551	44.989947	46.72
2	Algeria	2231.993335	12.500073	65.22
3	Andorra	21943.339900	81.000000	88.92
4	Angola	1381.004268	9.999954	56.70
...	...	...	...	...
208	Vietnam	722.807559	27.851822	27.84
209	West Bank and Gaza	0.000000	36.422772	71.90
210	Yemen, Rep.	610.357367	12.349750	30.64
211	Zambia	432.226337	10.124986	35.42
212	Zimbabwe	320.771890	11.500415	37.34

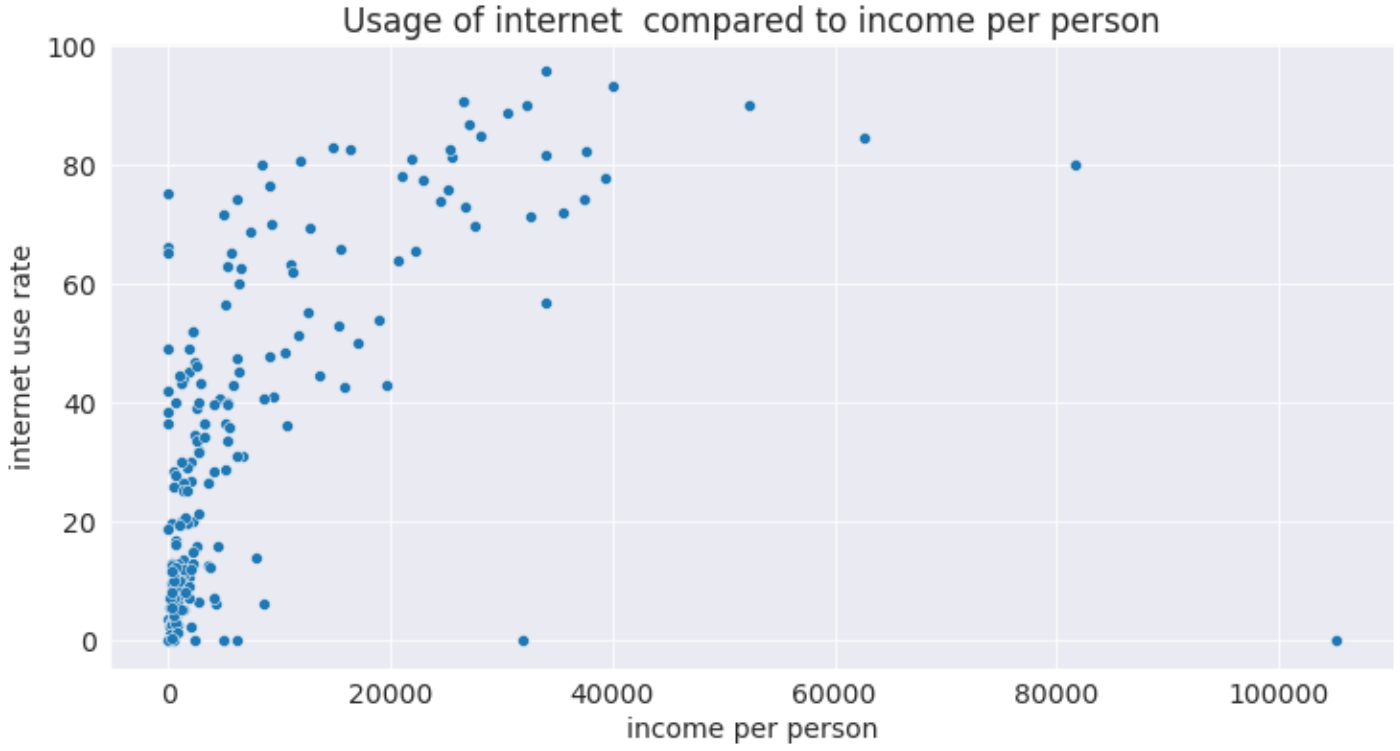
213 rows × 4 columns

```
usage_df.country.unique()
```

```
array(['Afghanistan', 'Albania', 'Algeria', 'Andorra', 'Angola',
       'Antigua and Barbuda', 'Argentina', 'Armenia', 'Aruba',
       'Australia', 'Austria', 'Azerbaijan', 'Bahamas', 'Bahrain',
       'Bangladesh', 'Barbados', 'Belarus', 'Belgium', 'Belize', 'Benin',
       'Bermuda', 'Bhutan', 'Bolivia', 'Bosnia and Herzegovina',
       'Botswana', 'Brazil', 'Brunei', 'Bulgaria', 'Burkina Faso',
       'Burundi', 'Cambodia', 'Cameroon', 'Canada', 'Cape Verde',
       'Cayman Islands', 'Central African Rep.', 'Chad', 'Chile', 'China',
       'Colombia', 'Comoros', 'Congo, Dem. Rep.', 'Congo, Rep.',
       'Cook Islands', 'Costa Rica', "Cote d'Ivoire", 'Croatia', 'Cuba',
       'Cyprus', 'Czech Rep.', 'Denmark', 'Djibouti', 'Dominica',
       'Dominican Rep.', 'Ecuador', 'Egypt', 'El Salvador',
       'Equatorial Guinea', 'Eritrea', 'Estonia', 'Ethiopia',
       'Faeroe Islands', 'Fiji', 'Finland', 'France', 'French Polynesia',
       'Gabon', 'Gambia', 'Georgia', 'Germany', 'Ghana', 'Gibraltar',
       'Greece', 'Greenland', 'Grenada', 'Guadeloupe', 'Guam',
       'Guatemala', 'Guinea', 'Guinea-Bissau', 'Guyana', 'Haiti',
       'Honduras', 'Hong Kong, China', 'Hungary', 'Iceland', 'India',
       'Indonesia', 'Iran', 'Iraq', 'Ireland', 'Israel', 'Italy',
       'Jamaica', 'Japan', 'Jordan', 'Kazakhstan', 'Kenya', 'Kiribati',
       'Korea, Dem. Rep.', 'Korea, Rep.', 'Kuwait', 'Kyrgyzstan', 'Laos',
       'Latvia', 'Lebanon', 'Lesotho', 'Liberia', 'Libya',
       'Liechtenstein', 'Lithuania', 'Luxembourg', 'Macao, China',
       'Macedonia, FYR', 'Madagascar', 'Malawi', 'Malaysia', 'Maldives',
       'Mali', 'Malta', 'Marshall Islands', 'Martinique', 'Mauritania',
       'Mauritius', 'Mexico', 'Micronesia, Fed. Sts.', 'Moldova',
       'Monaco', 'Mongolia', 'Montenegro', 'Morocco', 'Mozambique',
```

```
 'Myanmar', 'Namibia', 'Nauru', 'Nepal', 'Netherlands',
 'Netherlands Antilles', 'New Caledonia', 'New Zealand',
 'Nicaragua', 'Niger', 'Nigeria', 'Niue', 'Norway', 'Oman',
 'Pakistan', 'Palau', 'Panama', 'Papua New Guinea', 'Paraguay',
 'Peru', 'Philippines', 'Poland', 'Portugal', 'Puerto Rico',
 'Qatar', 'Reunion', 'Romania', 'Russia', 'Rwanda',
 'Saint Kitts and Nevis', 'Saint Lucia',
 'Saint Vincent and the Grenadines', 'Samoa', 'San Marino',
 'Sao Tome and Principe', 'Saudi Arabia', 'Senegal', 'Serbia',
 'Serbia and Montenegro', 'Seychelles', 'Sierra Leone', 'Singapore',
 'Slovak Republic', 'Slovenia', 'Solomon Islands', 'Somalia',
 'South Africa', 'Spain', 'Sri Lanka', 'Sudan', 'Suriname',
 'Swaziland', 'Sweden', 'Switzerland', 'Syria', 'Taiwan',
 'Tajikistan', 'Tanzania', 'Thailand', 'Timor-Leste', 'Togo',
 'Tonga', 'Trinidad and Tobago', 'Tunisia', 'Turkey',
 'Turkmenistan', 'Tuvalu', 'Uganda', 'Ukraine',
 'United Arab Emirates', 'United Kingdom', 'United States',
 'Uruguay', 'Uzbekistan', 'Vanuatu', 'Venezuela', 'Vietnam',
 'West Bank and Gaza', 'Yemen, Rep.', 'Zambia', 'Zimbabwe'],
dtype=object)
```

```
plt.figure(figsize=(12, 6))
plt.xlabel("income per person")
plt.ylabel("internet use rate")
plt.title("Usage of internet  compared to income per person")
sns.scatterplot(x=usage_df.incomeperperson,y=usage_df.internetuserate);
```



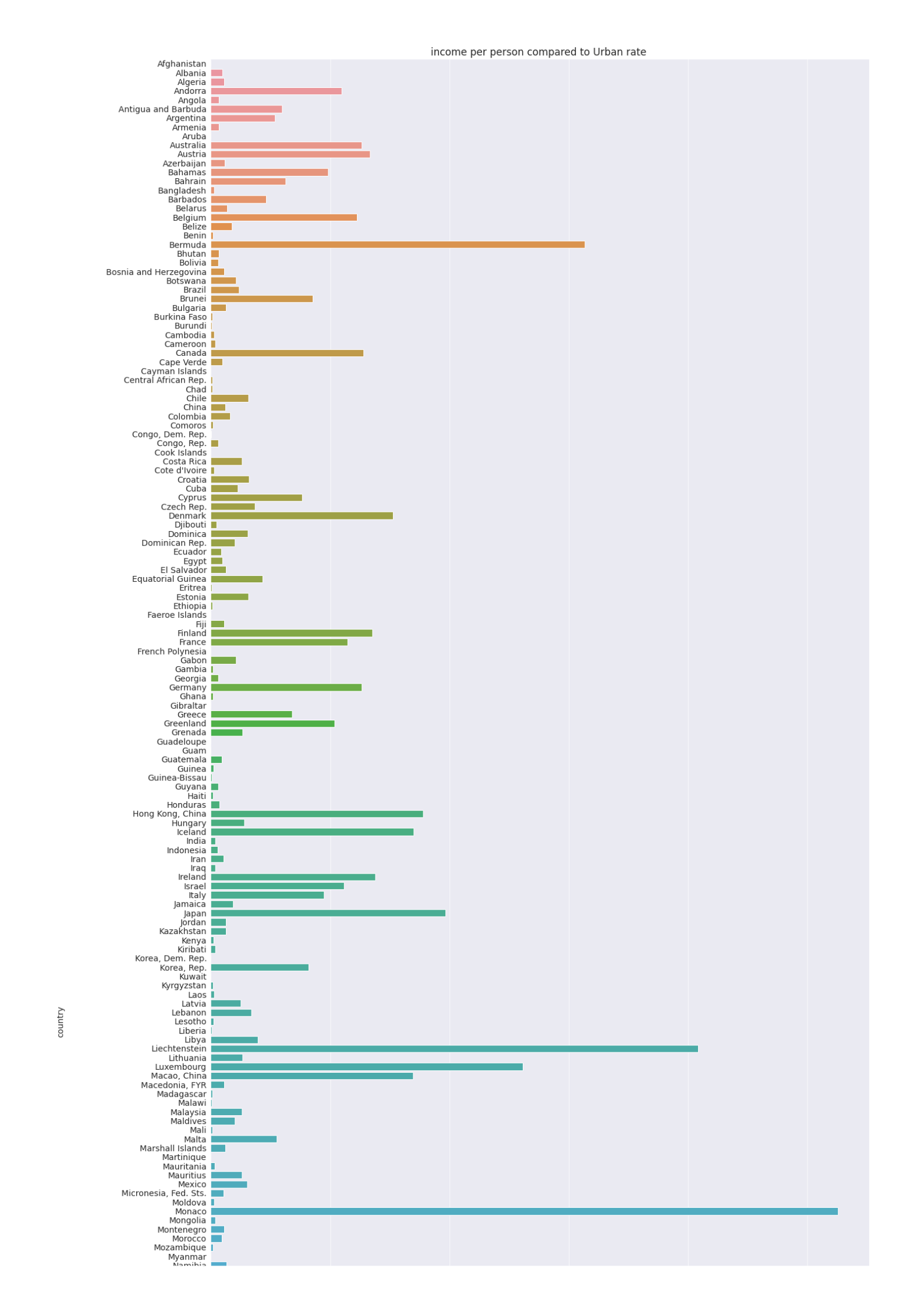
The above plot compares the age difference of student who attend the school GP against those who attend the school MS.from the above plot it seems like students tend to join the MS school from ages of of 17 and above while those of GP tend to be from ages of 15 and going to highs of up to 22 yrs old.

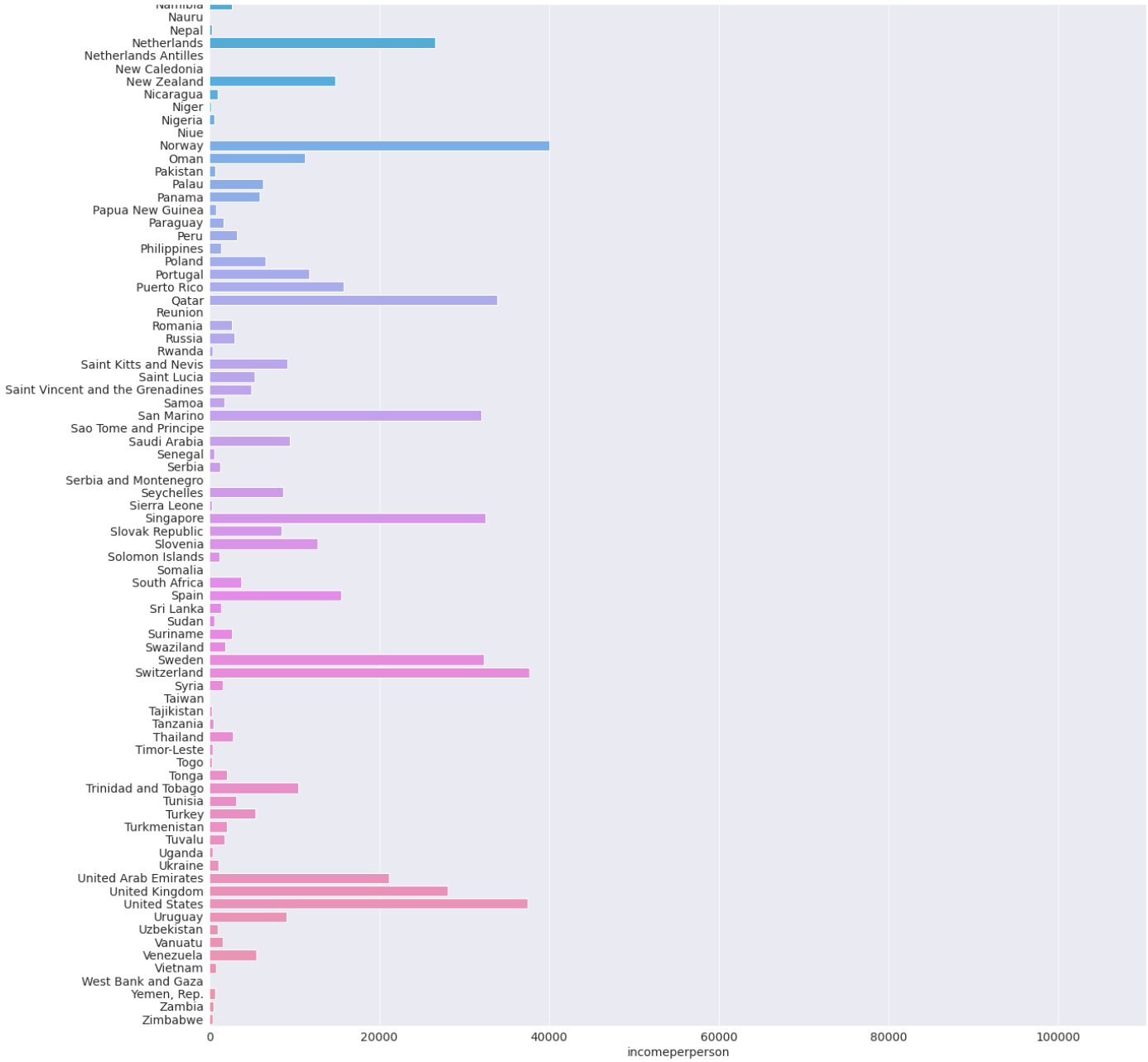
```
usage_df
```

	country	incomeperperson	internetuserate	urbanrate
0	Afghanistan	0.000000	3.654122	24.04
1	Albania	1914.996551	44.989947	46.72
2	Algeria	2231.993335	12.500073	65.22
3	Andorra	21943.339900	81.000000	88.92
4	Angola	1381.004268	9.999954	56.70
...	...	...	...	...
208	Vietnam	722.807559	27.851822	27.84
209	West Bank and Gaza	0.000000	36.422772	71.90
210	Yemen, Rep.	610.357367	12.349750	30.64
211	Zambia	432.226337	10.124986	35.42
212	Zimbabwe	320.771890	11.500415	37.34

213 rows × 4 columns

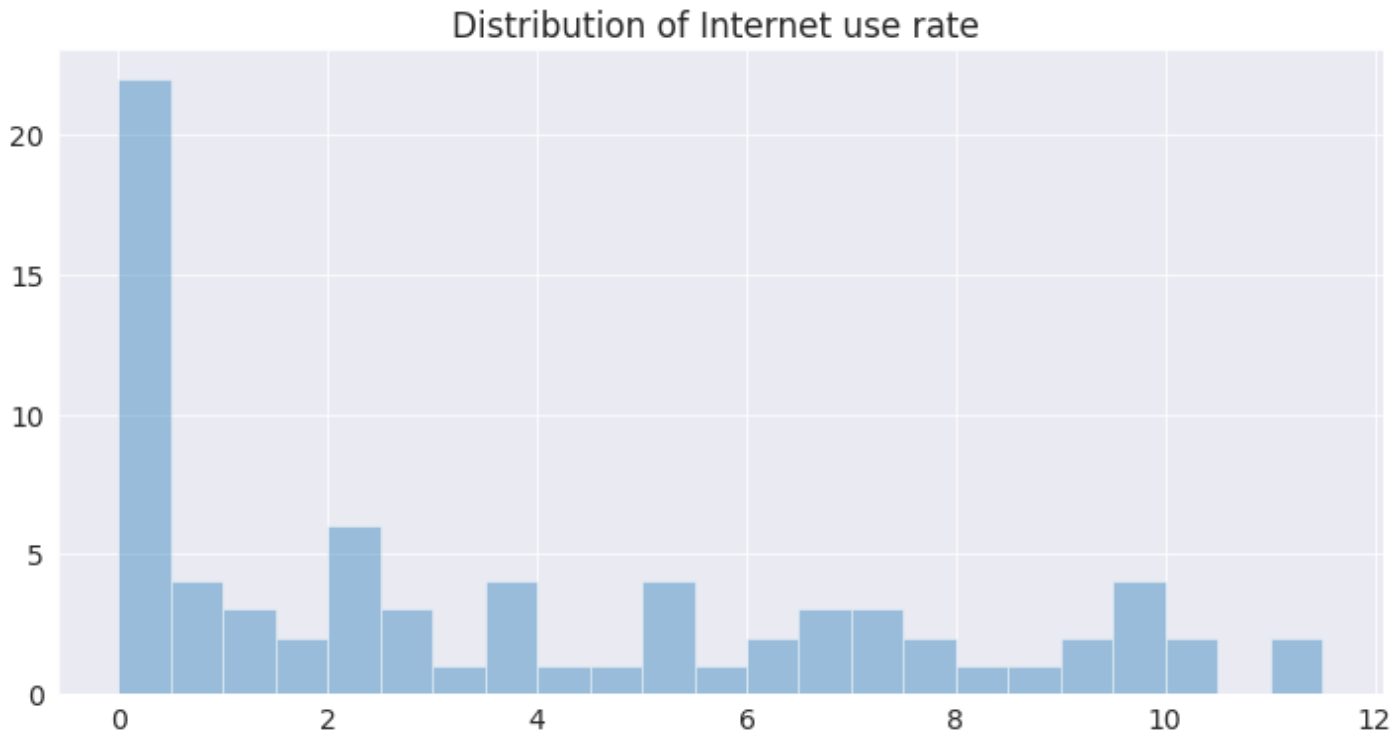
```
plt.figure(figsize=(20, 60))
plt.xlabel("income per person")
plt.ylabel("Urban rate")
plt.title("income per person compared to Urban rate ")
sns.barplot(x=usage_df.incomeperperson,y= usage_df.country);
```


income per person compared to Urban rate
country
Afghanistan
Albania
Algeria
Andorra
Angola
Antigua and Barbuda
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Bangladesh
Barbados
Belarus
Belgium
Belize
Benin
Bermuda
Bhutan
Bolivia
Bosnia and Herzegovina
Botswana
Brazil
Brunei
Bulgaria
Burkina Faso
Burundi
Cambodia
Cameroon
Canada
Cape Verde
Cayman Islands
Central African Rep.
Chad
Chile
China
Colombia
Comoros
Congo, Dem. Rep.
Congo, Rep.
Cook Islands
Costa Rica
Cote d'Ivoire
Croatia
Cuba
Cyprus
Czech Rep.
Denmark
Djibouti
Dominica
Dominican Rep.
Ecuador
Egypt
El Salvador
Equatorial Guinea
Eritrea
Estonia
Ethiopia
Faeroe Islands
Fiji
Finland
France
French Polynesia
Gabon
Gambia
Georgia
Germany
Ghana
Gibraltar
Greece
Greenland
Grenada
Guadeloupe
Guam
Guatemala
Guinea
Guinea-Bissau
Guyana
Haiti
Honduras
Hong Kong, China
Hungary
Iceland
India
Indonesia
Iran
Iraq
Ireland
Israel
Italy
Jamaica
Japan
Jordan
Kazakhstan
Kenya
Kiribati
Korea, Dem. Rep.
Korea, Rep.
Kuwait
Kyrgyzstan
Laos
Latvia
Lebanon
Lesotho
Liberia
Libya
Liechtenstein
Lithuania
Luxembourg
Macao, China
Macedonia, FYR
Madagascar
Malawi
Malaysia
Maldives
Mali
Malta
Marshall Islands
Martinique
Mauritania
Mauritius
Mexico
Micronesia, Fed. Sts.
Moldova
Monaco
Mongolia
Montenegro
Morocco
Mozambique
Myanmar
Namibia



from the above scatter plot it indicates that the higher the income per person the higher their likelihood to live in an urban rate.

```
import numpy as np
plt.figure(figsize=(12, 6))
plt.title("Distribution of Internet use rate ")
plt.hist(usage_df.internetuserate,alpha=0.4,bins=np.arange(0,12,0.5));
```




```
sample_df= usage_df.sort_values(by=['incomeperperson','internetuserate'],ascending=(Tru
pie_df=
plot = sample_df ('urbanrate').plot.pie(y='urbanrate', figsize=(5, 5))
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
/tmp/ipykernel_38/2276888109.py in <module>
      1 sample_df= usage_df.sort_values(by=['incomeperperson','internetuserate'],ascendi
(True,False)).head(10)
----> 2 pie_df=sample_df ('urbanrate')
      3 plot = series.plot.pie(y='urbanrate', figsize=(5, 5))

TypeError: 'DataFrame' object is not callable
```

```

```

Let us save and upload our work to Jovian before continuing

```
import jovian
```

```
jovian.commit()
```

Asking and Answering Questions

In this section 5 questions will be asked about the global internet usage data and inferences and attempts to answer this questions as accurately as possible will be provided.

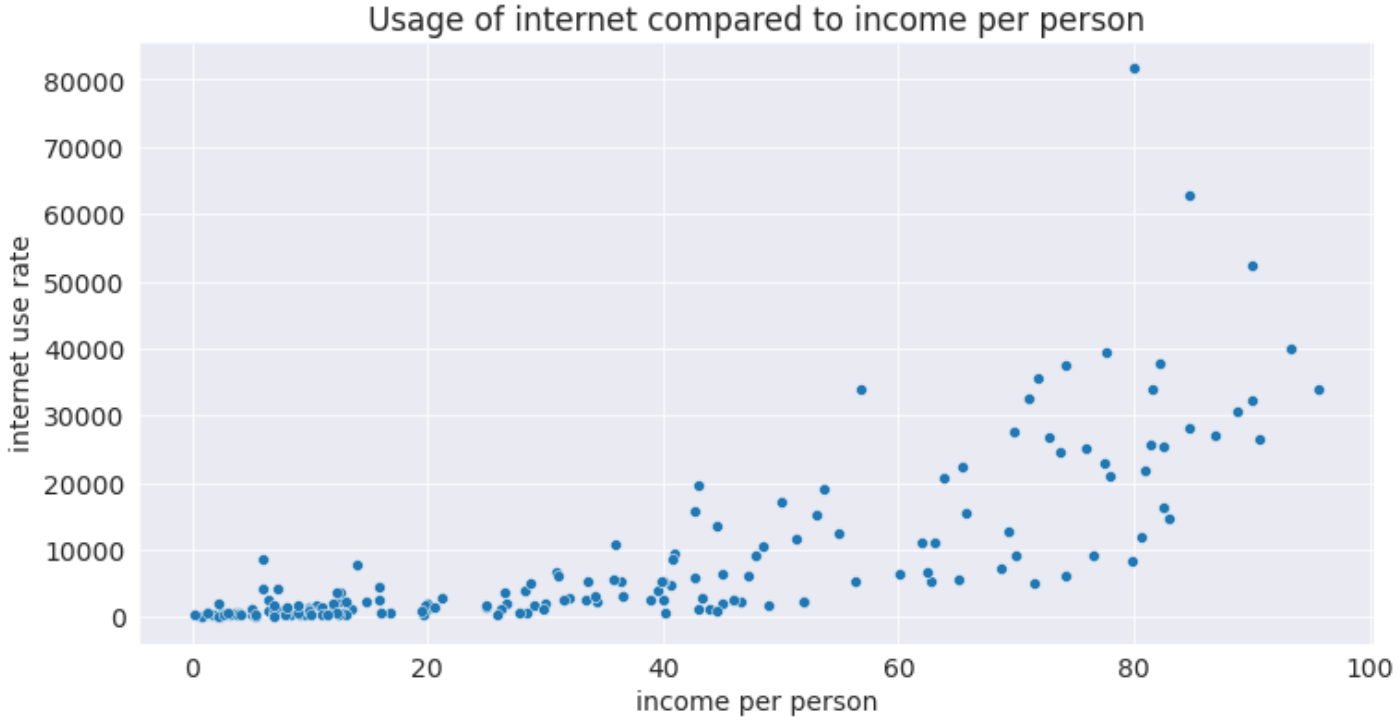
Q1: How does the Usage of internet compare to income per person ?

```
usage_income=usage_df[["incomeperperson","internetuserate"]]
usage_income
```

	incomeperperson	internetuserate
0	NaN	3.654122
1	1914.996551	44.989947
2	2231.993335	12.500073
3	21943.339900	81.000000
4	1381.004268	9.999954
...	...	...
208	722.807559	27.851822
209	NaN	36.422772
210	610.357367	12.349750
211	432.226337	10.124986
212	320.771890	11.500415

213 rows × 2 columns

```
plt.figure(figsize=(12, 6))
plt.xlabel("income per person")
plt.ylabel("internet use rate")
plt.title("Usage of internet compared to income per person")
sns.scatterplot(x=usage_df.internetuserate,y=usage_df.incomeperperson);
```



From the above plot it can be concluded that the higher a persons' income is the higher their internet usage, this can be beacuse with higher income on is able to purchase data connectivity either through wifi or mobile subscriptions.Also higher income might lead to one having more time in their hands hence spend more time on the internet as compared to lower income earners who might spend more time working.

Q2: How does the income per person compare to Urban rate?

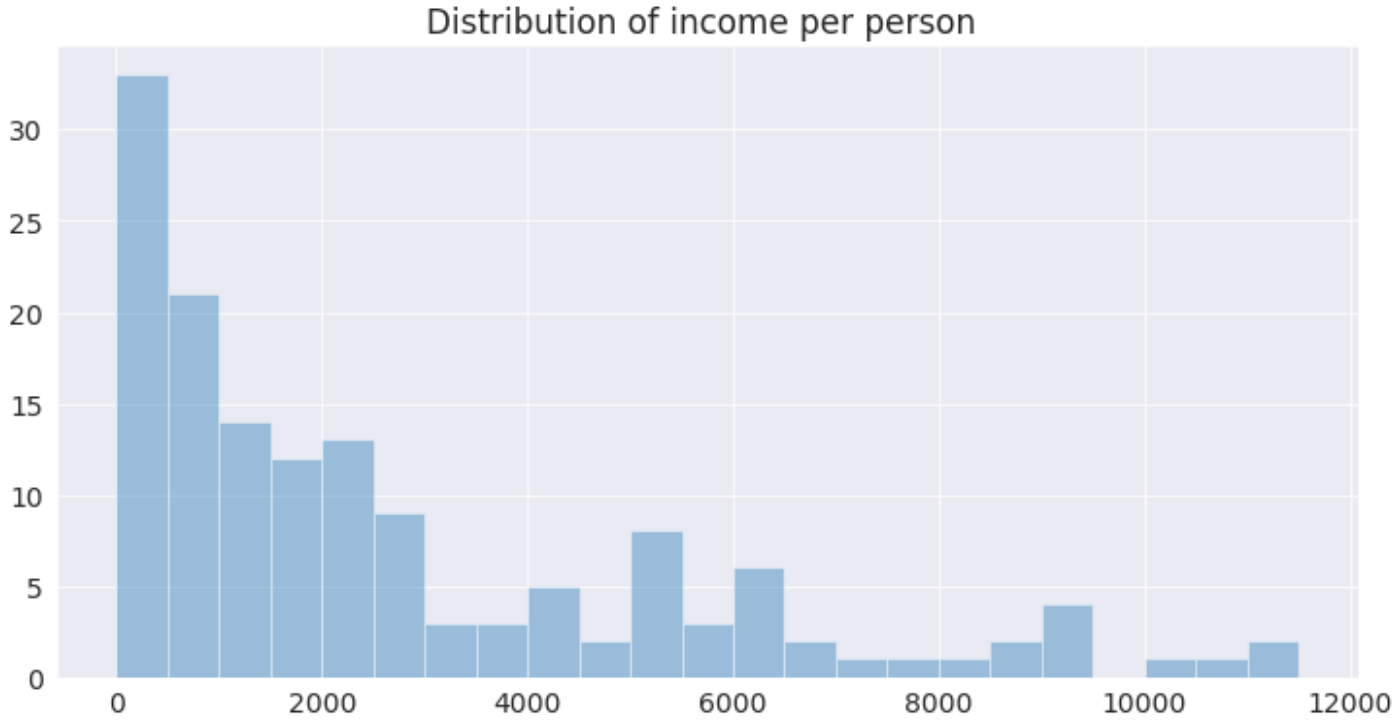
```
income_urban=usage_df[['urbanrate']]
income_urban
```

	urbanrate
0	24.04
1	46.72
2	65.22
3	88.92
4	56.70
...	...
208	27.84
209	71.90
210	30.64
211	35.42
212	37.34

213 rows × 1 columns

Q3: What is the income distribution per person?

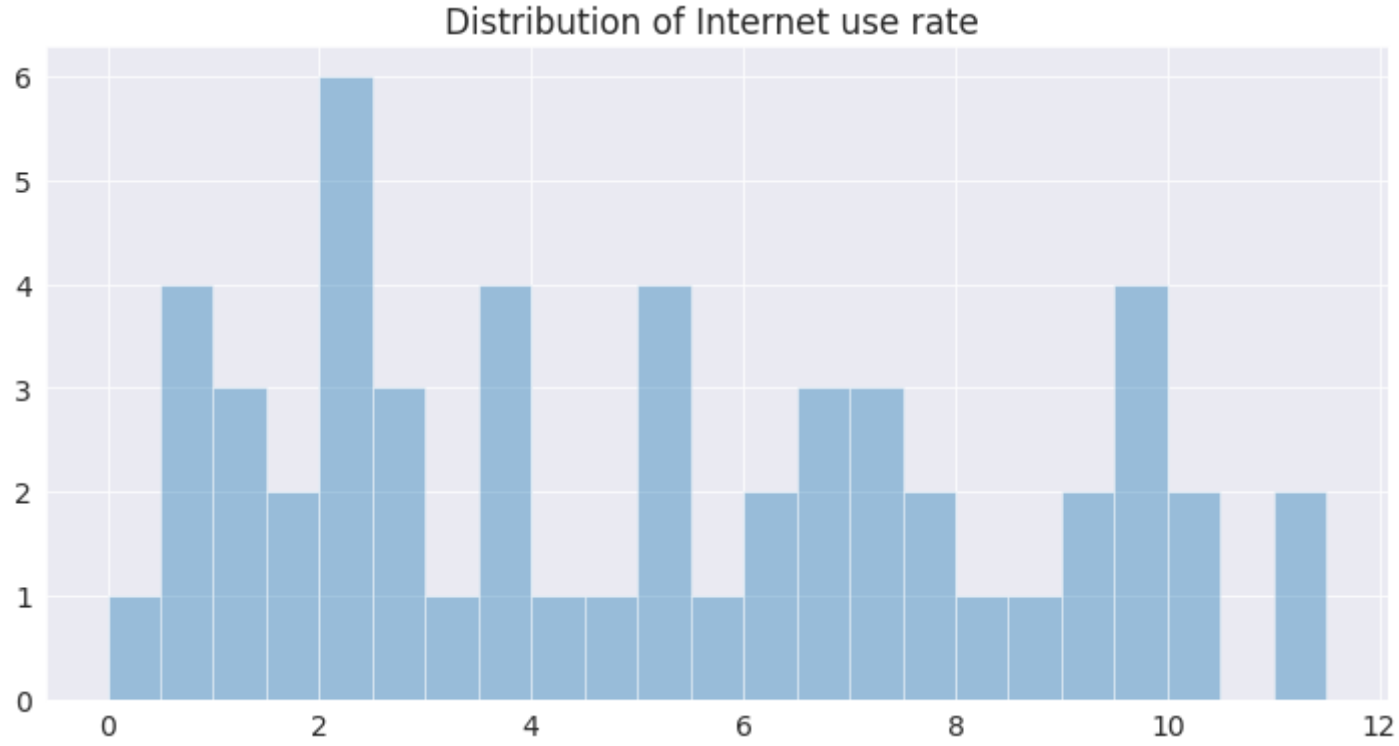
```
import numpy as np
plt.figure(figsize=(12, 6))
plt.title("Distribution of income per person ")
plt.hist(usage_df.incomeperperson,alpha=0.4,bins=np.arange(0, 12000, 500));
```



from the above plot it can be inferred that the higher the income one earns the higher the chances the are living in an urban area,this could be directly as a result of higher access to finances an affordance of urban living.

Q4: What is the Distribution of Internet use rate?

```
import numpy as np
plt.figure(figsize=(12, 6))
plt.title("Distribution of Internet use rate ")
plt.hist(usage_df.internetuserate,alpha=0.4,bins=np.arange(0,12,0.5));
```




```
usage_df.sort_values(by=['incomeperperson','internetuserate'],ascending=(True,False)).h
```

	country	incomeperperson	internetuserate	urbanrate
41	Congo, Dem. Rep.	103.775857	0.720009	33.96
29	Burundi	115.305996	2.100213	10.40
58	Eritrea	131.796207	5.399667	20.72
107	Liberia	155.033231	7.000214	60.14
79	Guinea-Bissau	161.317137	2.450362	29.84
141	Niger	180.083376	0.829997	16.54
115	Malawi	184.141797	2.259976	18.80
60	Ethiopia	220.891248	0.749996	17.00
35	Central African Rep.	239.518749	2.300027	38.58
114	Madagascar	242.677534	1.699985	29.52

Q5: How does the internet use rate,urban rate and income per person compare in a country?

```
usage_df.sort_values(by=['incomeperperson','internetuserate'],ascending=(True,False)).h
```

	country	incomeperperson	internetuserate	urbanrate

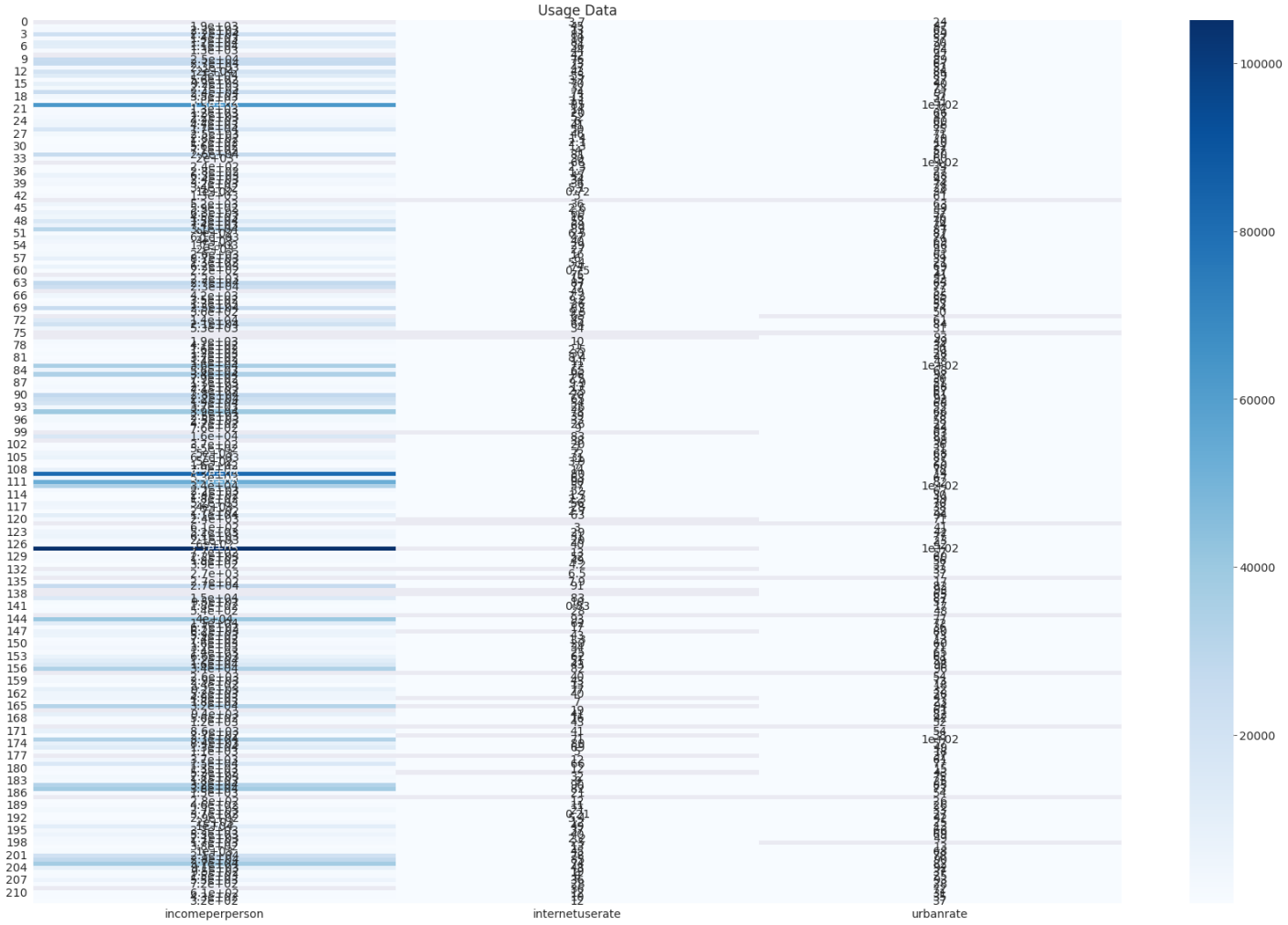
	country	incomeperperson	internetuserate	urbanrate
41	Congo, Dem. Rep.	103.775857	0.720009	33.96
29	Burundi	115.305996	2.100213	10.40
58	Eritrea	131.796207	5.399667	20.72
107	Liberia	155.033231	7.000214	60.14
79	Guinea-Bissau	161.317137	2.450362	29.84
141	Niger	180.083376	0.829997	16.54
115	Malawi	184.141797	2.259976	18.80
60	Ethiopia	220.891248	0.749996	17.00
35	Central African Rep.	239.518749	2.300027	38.58
114	Madagascar	242.677534	1.699985	29.52

```
data_df=usage_df[['incomeperperson','internetuserate','urbanrate']]
data_df
```

	incomeperperson	internetuserate	urbanrate
0	NaN	3.654122	24.04
1	1914.996551	44.989947	46.72
2	2231.993335	12.500073	65.22
3	21943.339900	81.000000	88.92
4	1381.004268	9.999954	56.70
...	...	...	...
208	722.807559	27.851822	27.84
209	NaN	36.422772	71.90
210	610.357367	12.349750	30.64
211	432.226337	10.124986	35.42
212	320.771890	11.500415	37.34

213 rows × 3 columns

```
plt.figure(figsize=(30,20))
plt.title("Usage Data")
sns.heatmap(data_df, annot=True, cmap='Blues');
```



Let us save and upload our work to Jovian before continuing.

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "mainavictor004/data-analysis-and-visualization-on-global-internet-usage" on https://jovian.ai
[jovian] Committed successfully! https://jovian.ai/mainavictor004/data-analysis-and-visualization-on-global-internet-usage
```

```
'https://jovian.ai/mainavictor004/data-analysis-and-visualization-on-global-internet-usage'
```

Inferences and Conclusion

Generally from the above analysis of the Global Internet Usage data, it can be inferred that people who earn high incomes in the world will tend to have the most access to beast internet speeds and providence(hence high internet usage among high income earners),also high income earners tend to graviate towards more developed areas for living(hence the higher demographic of high income earners in urban rates). Also it can be inferred that countries with higher rates of high income earners have higher urban rates and vice versa.

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "mainavictor004/data-analysis-and-visualization-on-global-internet-usage" on https://jovian.ai
[jovian] Committed successfully! https://jovian.ai/mainavictor004/data-analysis-and-visualization-on-global-internet-usage
```

```
'https://jovian.ai/mainavictor004/data-analysis-and-visualization-on-global-internet-usage'
```

References and Future Work

In the Global Internet Usage data, countries Can also in future be narrowed into continents and compare internet use rates,urban rates and income per person within continents and also between continents.

links: https://www.kaggle.com/code/arvindkale/notebook5a3321ea4c/data

https://matplotlib.org/3.1.1/api/

https://pandas.pydata.org/docs/user_guide/index.html

https://www.kaggle.com/datasets

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "mainavictor004/data-analysis-and-visualization-on-global-internet-usage" on https://jovian.ai
[jovian] Committed successfully! https://jovian.ai/mainavictor004/data-analysis-and-visualization-on-global-internet-usage
```

```
'https://jovian.ai/mainavictor004/data-analysis-and-visualization-on-global-internet-usage'
```

```
```