Investigating How Internet Usage Rates Vary With Socioeconomic Factors

By Kaspar Lee

Table of Contents

Introduction

In this project, I will be analysing how the proportion of a population with usage of the internet varies based on a nation's level of corruption, democracy and freedom of expression.

Datasets and Indicators

I compiled my datasets using Gapminder Tools (https://www.gapminder.org/data/), which contains data, broken down by country, on a wide range of indicators. As the indicators I have chosen are not quantitative, I have opted to use the following indices in order to quantify the data:

- **Corruption Perception Index (CPI)** - This index, calculated by Transparency International (https://www.transparency.org/research/cpi), is a measure of the level of corruption in a country. It is based on a scale of 0 to 100, with zero indicating a "Highly Corrupt" nation, and 100 indication a nation is "Very Clean".
- **Democracy Index (EIU)** - From the Economist Inteligence Nuit (http://gapm.io/ddemocrix_eiu), this is a summary measure to express the quality of a country's democratic nature, calculated using 60 indicators. Graded from 0 to 100, with 0 indicating a very low level of democracy, and 100 indicating a very high democratic nature.
- **Freedom of Expression Index (IDEA)** - Available here (http://gapm.io/ddemocrix_idea), this aggregates a set of indicators measuring media censorship and freedom of discussion and expression. Measued on a scale of 0 to 100, with 0 suggesting no freedom of expression at all, and 100 suggesting full access to freedom of expression.

The internet users dataset from The World Bank Group (https://data.worldbank.org/indicator/IT.NET.USER.ZS) contains the number of internet users as a percentage of the total population. This will allow me to compare internet users relative to the size of population of a country.

All of these datasets include historical data, however I am not interested in trends in any of these indicators so can discard all but the most recent year (that all indicators have data for). The corruption index only has data until 2017, so that will most likely be the year that I use as the most recent year.

Questions

I shall be analysing the distribution of countries' population with usage of the internet, and the relationship between the internet usage rates and the above indicators. My questions are:

- **How are internet usage rates distributed, and how do they range between different countries?**
- **Is there a correlation between the level of corruption, democracy or freedom of a country, and the number of individuals using the internet?**

Data Wrangling

Data Checking

In [1]:
```python
from functools import reduce
from IPython.display import display
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style('darkgrid')

%matplotlib inline
```

In [2]:
```python
# Load data
internet_df   = pd.read_csv('internet_usage_rate.csv')
corruption_df = pd.read_csv('corruption_index.csv')
democracy_df  = pd.read_csv('democracy_index.csv')
freedom_df    = pd.read_csv('freedom_index.csv')
```

Internet Usage Data

In [3]:
```python
display(internet_df.describe())
display(internet_df.head())
```

	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	...	2009	2010	20
count	7.0	0.0	0.0	0.0	0.0	7.0	0.0	0.0	0.0	0.0	...	189.000000	189.000000	192.0000
mean	0.0	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	NaN	...	29.371111	32.515450	35.2759
std	0.0	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	NaN	...	26.809729	27.300222	27.7310
min	0.0	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	NaN	...	0.000000	0.000000	0.0000
25%	0.0	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	NaN	...	6.150000	8.000000	9.6500
50%	0.0	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	NaN	...	22.500000	27.200000	32.0000
75%	0.0	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	NaN	...	48.800000	52.000000	55.2250
max	0.0	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	NaN	...	93.000000	93.400000	94.8000

8 rows × 59 columns

	country	1960	1961	1962	1963	1964	1965	1966	1967	1968	...	2009	2010	2011	2012	2013
0	Afghanistan	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	3.55	4.0	5.0	5.45	5.9
1	Albania	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	41.20	45.0	49.0	54.70	57.2
2	Algeria	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	11.20	12.5	14.9	18.20	22.5
3	Andorra	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	78.50	81.0	81.0	86.40	94.0
4	Angola	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	2.30	2.8	3.1	6.50	8.9

5 rows × 60 columns

The newest data is from 2018, however it seems to only have 79 countries entered, rather than the ~190 countries from the years before. Additionally, the corruption dataset only contains data until 2017. Due to this, I will take 2017 as the most recent year of data that I can analyse, which has data saved for 192 countries.

Corruption Data

In [4]:
```
display(corruption_df['2017'].describe())
display(corruption_df.head())
```

```
count    177.000000
mean      42.790960
std       18.978347
min        9.000000
25%       29.000000
50%       38.000000
75%       56.000000
max       89.000000
Name: 2017, dtype: float64
```

	country	2012	2013	2014	2015	2016	2017
0	Afghanistan	8.0	8.0	12.0	11.0	15.0	15
1	Albania	33.0	31.0	33.0	36.0	39.0	38
2	Algeria	34.0	36.0	36.0	36.0	34.0	33
3	Angola	22.0	23.0	19.0	15.0	18.0	19
4	Argentina	35.0	34.0	34.0	32.0	36.0	39

As we can see, there are 177 unique countries with data existing in the 2017 column.

Democracy Data

In [5]:
```
display(democracy_df['2017'].describe())
display(democracy_df.head())
```

```
count    164.000000
mean      54.538415
std       22.001730
min       10.800000
25%       36.100000
50%       56.600000
75%       72.000000
max       98.700000
Name: 2017, dtype: float64
```

	country	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
0	Afghanistan	30.6	30.4	30.2	27.5	24.8	24.8	24.8	24.8	27.7	27.7	25.5	25.5	29.7
1	Albania	59.1	59.1	59.1	58.9	58.6	58.1	56.7	56.7	56.7	59.1	59.1	59.8	59.8
2	Algeria	31.7	32.5	33.2	33.8	34.4	34.4	38.3	38.3	38.3	39.5	35.6	35.6	35.0
3	Angola	24.1	28.8	33.5	33.4	33.2	33.2	33.5	33.5	33.5	33.5	34.0	36.2	36.2
4	Argentina	66.3	66.3	66.3	67.3	68.4	68.4	68.4	68.4	68.4	70.2	69.6	69.6	70.2

Here we can see that the democracy data does indeed have data for 2017, however only has 164 unique countries present that year.

Freedom of Expression Data

In [6]:
```
display(freedom_df['2017'].describe())
display(freedom_df.head())
```

```
count     155.000000
mean       60.225806
std        20.713428
min         4.000000
25%        46.000000
50%        64.000000
75%        76.500000
max        92.000000
Name: 2017, dtype: float64
```

	country	1975	1976	1977	1978	1979	1980	1981	1982	1983	...	2009	2010	2011	2012	2013
0	Afghanistan	35.0	35.0	35.0	23.0	20.0	20.0	22.0	22.0	22.0	...	53.0	52.0	52	51	52
1	Albania	9.0	9.0	9.0	9.0	9.0	9.0	9.0	9.0	9.0	...	71.0	71.0	71	71	64
2	Algeria	34.0	34.0	34.0	34.0	34.0	36.0	36.0	36.0	36.0	...	58.0	57.0	57	52	57
3	Angola	19.0	19.0	19.0	19.0	19.0	19.0	19.0	19.0	20.0	...	46.0	46.0	46	46	47
4	Argentina	52.0	24.0	14.0	14.0	14.0	14.0	14.0	17.0	33.0	...	78.0	77.0	77	77	78

5 rows × 45 columns

Finally, this shows us that the freedom data only includes 155 countries for the year of 2017.

Data Cleaning

As I decided previously, I will only be working with data from 2017. I have now confirmed that all datasets have data for this year.

In order to clean the data, I need to:

- Discard the unused years in all datasets, keeping only the 2017 columns
- Discard all rows in all datasets with missing values
- Rename the 2017 columns to include the name of the data they reference (in order to distinguish once combined)
- Combine all data into one dataset using an inner merge on the `country` column

Discard Unused Columns

In [7]:
```
# Columns to keep
columns = ['country', '2017']

internet_df = internet_df.filter(columns)
corruption_df = corruption_df.filter(columns)
democracy_df = democracy_df.filter(columns)
freedom_df = freedom_df.filter(columns)
```

Discard Rows with Missing Values

In [8]:
```python
# Drop all rows with missing values
internet_df.dropna(inplace=True)
corruption_df.dropna(inplace=True)
democracy_df.dropna(inplace=True)
freedom_df.dropna(inplace=True)
```

Rename Columns

In [9]:
```python
internet_df.rename(columns={'2017': 'internet_usage'}, inplace=True)
corruption_df.rename(columns={'2017': 'corruption'}, inplace=True)
democracy_df.rename(columns={'2017': 'democracy'}, inplace=True)
freedom_df.rename(columns={'2017': 'freedom'}, inplace=True)
```

Combine Datasets

In [10]:
```python
# Inner merge two dataframes on the "country" column
def merge(left, right):
    return pd.merge(left, right, on='country', how='inner')

dataframes = [internet_df, corruption_df, democracy_df, freedom_df]
combined_df = reduce(merge, dataframes)

combined_df.head()
```

Out[10]:

	country	internet_usage	corruption	democracy	freedom
0	Afghanistan	13.5	15	25.5	51
1	Albania	71.8	38	59.8	69
2	Algeria	47.7	33	35.6	55
3	Angola	14.3	19	36.2	51
4	Argentina	74.3	39	69.6	82

In [11]:
```python
combined_df['country'].nunique()
```

Out[11]:
```
151
```

The data is now free of all missing values, and combined into one dataframe, with separate columns for the internet usage rate, level of corruption, democracy and freedom. There are 151 countries that had data saved in all four datasets, and hence those are what I am left with.

This data is now fully cleaned and ready for analysis.

Exploratory Data Analysis

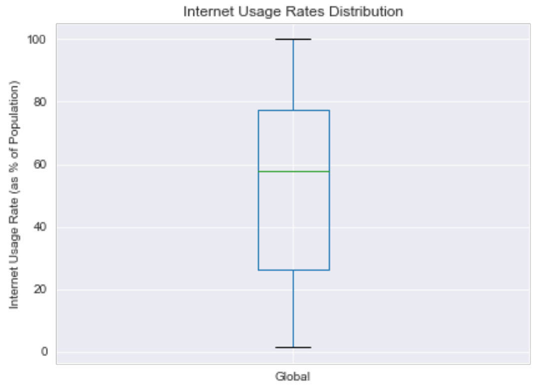
How are internet usage rates distributed, and how do they range between different countries?

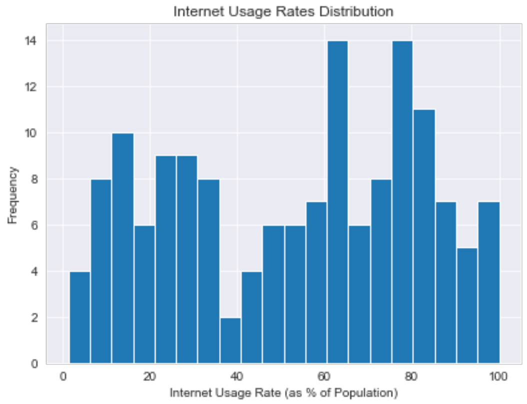
Graphs

In [12]:
```
fig, axs = plt.subplots(1,2, figsize=(15, 5))

# Create box plot
boxplot = combined_df['internet_usage'].plot.box(title='Internet Usage Rates Di
stribution', ax=axs[0])
boxplot.set_xticklabels(['Global'])
boxplot.set_ylabel('Internet Usage Rate (as % of Population)')

# Create histogram
histogram = combined_df['internet_usage'].plot.hist(bins=20, title='Internet Us
age Rates Distribution', ax=axs[1])
histogram.set_xlabel('Internet Usage Rate (as % of Population)');
```





Analysis

Looking at the distribution of internet usage rates, the median value lies just under 60%, however there is a very large range spanning from almost 0% to almost 100% of different countries' populations having usage of the internet. Additionally, there is a large interquartile range, indicating a significant spread of internet usage rates between different countries.

The histogram shows us that there are many countries with rates under 35%, and many more with rates above 60%, however it appears there are relatively few around the 40-50% mark.

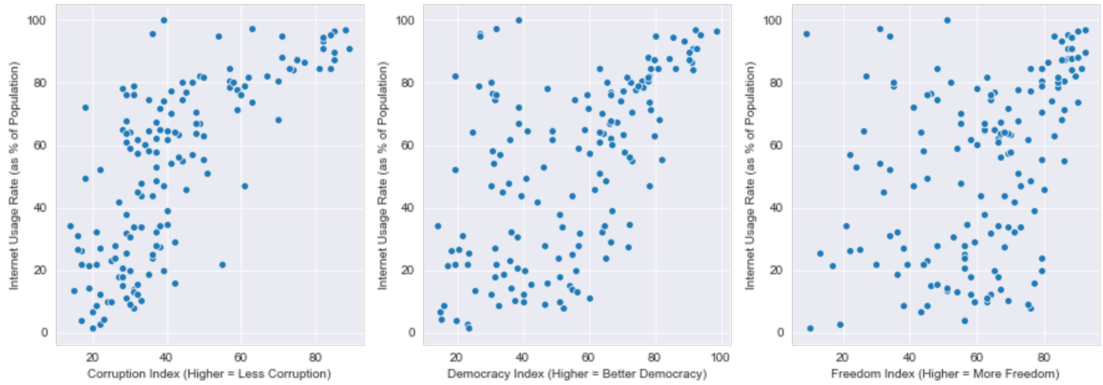
Is there a correlation between the level of corruption, democracy or freedom of a country, and the number of individuals using the internet?

Scatter Graphs

In [13]:
```
fig, axs = plt.subplots(1, 3, figsize=(15, 5))

# Draw plots
sns.scatterplot(combined_df['corruption'], combined_df['internet_usage'], ax=ax
s[0])
sns.scatterplot(combined_df['democracy'], combined_df['internet_usage'], ax=axs
[1])
sns.scatterplot(combined_df['freedom'], combined_df['internet_usage'], ax=axs
[2]);

# Add labels
axs[0].set_xlabel('Corruption Index (Higher = Less Corruption)')
axs[1].set_xlabel('Democracy Index (Higher = Better Democracy)')
axs[2].set_xlabel('Freedom Index (Higher = More Freedom)')
axs = list(map(lambda x: x.set_ylabel('Internet Usage Rate (as % of Populatio
n)'), axs))
```

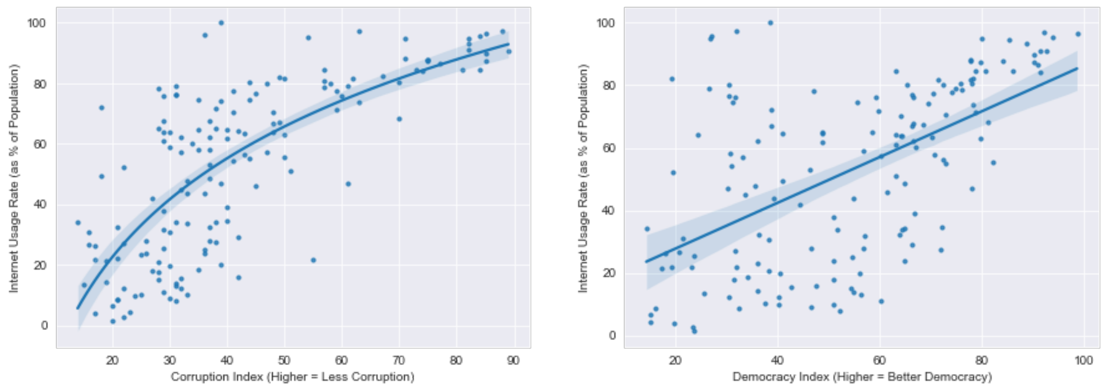


Estimated Regression Models

In [14]:
```
fig, axs = plt.subplots(1, 2, figsize=(15, 5))

# Draw plots
a = sns.regplot(combined_df['corruption'], combined_df['internet_usage'], ax=ax
s[0], scatter_kws={'s': 10}, logx=True)
b = sns.regplot(combined_df['democracy'], combined_df['internet_usage'], ax=axs
[1], scatter_kws={'s': 10});

# Add labels
a.set_xlabel('Corruption Index (Higher = Less Corruption)')
b.set_xlabel('Democracy Index (Higher = Better Democracy)')
a.set_ylabel('Internet Usage Rate (as % of Population)')
b.set_ylabel('Internet Usage Rate (as % of Population)');
```



Analysis

The first graph shows a clear positive correlation between the corruption index of a country and the proportion of the population using the internet. As the corruption index increases (level of corruption gets lower), the internet usage increases. Notably, all countries with a corruption score of over 70 have very high internet usage rates. On inspection, it looks like this relationship follows a logarithmic curve (as plotted by the first graph on the next row).

The next graph displays a weaker, but still positive, correlation between the democracy index and internet usage. More democratic countries tend to have a higher percentage of the population using the internet, while less democratic countries have lower internet usage rates.

Finally, the last graph shows a very weak positive correlation between freedom and internet usage. I appears as though countries with higher levels of freedom have slightly higher internet usage rates, however this correlation is significantly weaker than the previous two.

Conclusions

From my analysis, I have found that the distribution of internet usage rates does not follow a traditional bell curve, but rather has high frequencies of countries below 35% and above 60%, with relatively few countries around the 40-50% mark. I also found that the range was very large indeed, indicating a wide discrepancy between countries, and that the median internet usage rate is just under 60%.

Furthermore, I can conclude that their appears to a positive correlation between the corruption index of a country and the percentage of the population that use the internet. Additionally, there is a weaker positive correlation between the level of democracy and internet usage, and a very weak positive correlation between freedom and internet usage.

These results however are based on a small sample size (n=155), which means they are not the most accurate. It is almost impossible to increase the same size, as the sample includes the majority of the countries on the planet, which is a significant limitation to gaining more accurate results.

Finally, the correlation between lower corrupt and higher levels of internet usage does not necessarily mean to suggest that corruption causes lower levels of internet usage. The same applies for the level of democracy; a lower level of democracy does not mean a reduced proportion of the population accessing the internet.