Author: https://github.com/randomoi/

Student Number: XXXXXXXX

Created on Mac via Anaconda

WHAT FACTORS INFLUENCE INTERNET PRICES AROUND THE WORLD?

Table of Contents

Abstract

Taking into account affordability and currency weighting, Ukraine has the least expensive internet globally, while the United Arab Emirates has the most expensive. Global internet prices fluctuate wildly from country to country, a phenomenon that appears to be influenced partly by a consumer's ability to pay, partly by the complexity of technology deployment (for example, running fiber to small islands), regulatory burdens (the price of spectrum licenses, etc.) and currency fluctuations / Purchasing Power Parity. The main goal of this paper is to study internet prices and the factors behind them in the following countries: Argentina, Australia, Canada, France, India, Peru, Sri Lanka, Ukraine, United Arab Emirates, and the United States. These countries were selected to represent the variation in income levels across the world.

Additionally, the research will place particular emphasis on Ukraine within the context of the global scale of internet prices.

The emphasis was placed on Ukraine, because the author is Ukrainian.

Back to Table of Contents

Terminology

Terms used in the project:

- Internet stands for landline/broadband internet
- Internet price stands for the approximate average of the cost of broadband internet. Mobile or satellite internet connection is excluded from the analysis.
- Prices are in the United States Dollars (USD)
- PPP stands for Purchasing Power Parity
- Wages stand for average monthly net wages
- Numbeo is a crowd-sourced global database
- Kaggle a user-published dataset platform
- World Bank is a platform for free/open access data

Back to Table of Contents

Introduction

Internet prices vary drastically around the world. Research indicates that, of the countries studied, the least expensive internet is in Ukraine USD4.65/per month of “60 mbps or more, unlimited data, cable/adsl” (numbeo.com, n.d.)., while the most expensive internet is found in the United Arab Emirates USD99.10/per month of “60 mbps or more, unlimited data, cable/adsl” (numbeo.com, n.d.). Internet cost appears to be impacted by several critical factors including a) wealth of consumers (UAE, etc.), b) distortions in purchasing power parity influenced by currency fluctuations, c) high regulatory costs (USA, etc.), and d) difficulty of infrastructure (Caribbean nations, landlocked countries, etc.).

Why are internet prices drastically different around the world?

The research will focus on determining if the internet adoption rate, wages, and GDP per capita, PPP are the factors that impact internet costs for Argentina, Australia, Canada, France, India, Peru, Sri Lanka, Ukraine, United Arab Emirates, and the United States.

Back to Table of Contents

Project Background and Aims

The report aims to answer the following questions.

1. What are the internet prices around the world?
2. Are internet prices influenced by the internet usage rate?
3. Are internet prices influenced by GDP indicators?
4. Are internet prices influenced by wages?

The report will explore three variables that could impact internet prices globally: one of two GDP indicators, internet usage rate, and wages. Additionally, it will reflect on the placement of Ukraine on a global scale of internet prices.

Back to Table of Contents

Literature

The studies conducted on internet prices are primarily focused on fluctuation within a specific country or a small group of countries and not from a global perspective. For example, Calzada & FernandoMartínez-Santos (2014) (Joan Calzada, 2014) analyzed internet prices for 15 EU countries based on the effect of bundled services (TV, phone, and internet) and the type of technology offered. Kim & Moon & Yang (2004) (Heekyung Hellen Kim, 2004) analyzed internet penetration together with low-cost prices in South Korea and the effects it had on the political process. Lastly, Liu & Prince & Wallsten (2018) examined consumer willingness to pay for the internet based on the various download and upload speeds, data caps, etc. That being said, there have been a few attempts, according to the research, to identify what influences internet prices across the globe.

The report was inspired by a study conducted by Grechyn & McShane (Viktor Grechyn, 2016).

Back to Table of Contents

Scope of Work

Inclusions

- The report will focus primarily on 10 countries: Australia, Canada, France, United Arab Emirates, and the United States with high GDP per capita, and Argentina, Peru, India, Sri Lanka, and Ukraine with lower GDP per capita
- The report will overview the internet usage rate.
- The report will compare GDP per capita and GDP per capita, PPP with the purpose of determining which indicator to use for the analysis. It will then create a ratio between the internet prices and the selected indicator.
- The report will overview wages for a 5-year period. It will then create a ratio between internet prices and wages.
- The report will analyze 2021 for internet prices, GDP indicators, and wages.
- The report will analyze 2020 for internet usage.
- The report will analyze 6 datasets:
  - two datasets for the internet prices
  - one dataset for GDP per capita
  - one dataset for GDP per capita, PPP
  - one dataset for internet usage rate
  - one dataset for wages

Exclusions

- The report will not analyze every country in the world.
- The report will not analyze gross average wages due to taxation inequalities.
- The report will not analyze government policies and regulations such as if internet providers are regulated.
- The report will not analyze infrastructure.

Back to Table of Contents

Challenges

In the research process, it was discovered that internet price comparison among countries is not straightforward. Some available datasets with open licenses don't explicitly indicate the speed and type of technology (fiber optics, ADSL, and cable) used in the data. Initially, it was decided to use the OECD dataset from an intergovernmental organization and the Numbeo.com dataset from a crowdsourced platform to analyze internet prices. However, after further data exploration, it became apparent that the OECD dataset only contained 38 wealthy countries, which are part of the OECD organization for economic cooperation. Therefore, the OECD dataset was removed from the analyses and replaced with a dataset from Kaggle.com, which had a larger data sample aggregated from multiple sources, to understand internet prices in various economies (poor and wealthy).

Back to Table of Contents

Report Structure

1. Internet prices
    - What are internet prices worldwide?
    - What are internet prices in 10 selected countries?
2. Internet usage rate
    - What is the internet usage rate worldwide?
    - What is the internet usage rate in 10 selected countries?
    - What is the ratio between internet prices and internet usage rates?
3. GDP indicators
    - What are GDP indicators?
    - Which GDP indicator to use for analysis?
    - What is the ratio between internet prices and one GDP indicator?
4. Wages
    - What are wages worldwide?
    - What are wages in 10 selected countries?
    - What is the ratio between internet prices and wages?

Back to Table of Contents

Data Acquisition Techniques

Internet prices

- Web scraping from numbeo.com
- The download of the dataset from kaggle.com

Internet usage rate

- The download of the dataset from worldbank.org

GDP

- API request from WBGAPI
  - The World Bank: GDP per capita
  - The World Bank: GDP per capita, PPP

Wages

- The download of the dataset from kaggle.com (see Data Explorer in the side bar)

Back to Table of Contents

Techniques

- Web scraping Numbeo.com
- CVS import Kaggle.com
- API WorldBank.org
- Data Plotting
- Conversion from Excel format to CSV format
- Regular expression to float value conversion
- Word Cloud
- World map visualization

Back to Table of Contents

Libraries

Why were these libraries chosen?

- Pandas library was chosen because it has basic tools for cleaning and analyzing data. Furthermore, this library has many online resources for learning about various techniques.
- BeautifulSoup library was chosen because of its simple set of tools to create Python objects.
- Matplotlib.pyplot was chosen to plot various datasets because it has a beautiful presentation of plots.
- Requests.exceptions/ HTTPError library was chosen as it allows checking for errors when handling data requests during web scraping.
- Numpy library was used as it provides much better tools to handle arrays, in comparison to Python lists.
- OS library was used to work with the computer directory to check if the file exists. This was used based on the recommendations from online resources.
- Pathlib/Path library was used to check if the file already exists or not.
- WBGAPI library was used to request the World Bank data, this was done strictly to demonstrate the ability to use API. My preferred method is downloading and reading CSV file.
- Plotly.express was used to visualize prices and usage rates on the map. This was the only known library that can visualize data in the shape of a map.
- WordCloud/STOPWORDS were used to visualize words. It was used strictly to show the ability to use World Cloud as per the rubric. Frankly, it offered zero impact on the report.
- matplotlib.style.use('ggplot') was used to simulate ggplot, which is a simple plotting package for R.

Back to Table of Contents

Internet Prices

What are internet prices around the world?

Before jumping to the main question of the report “What factors influence internet prices around the world?”, it’s important to know the baseline of what internet prices are like around the world. Multiple datasets were considered to verify the data's integrity.

The author chose to work with the following resources.

- Numbeo.com is a crowd-sourced global database, which is free to use for personal and academic purposes. The website publishes the most recent statistics, but it is not from official governmental resources, therefore, another source was considered for validation purposes. A web scraping technique was used to retrieve the data and a regular expression to float value conversion technique was used to format the data.
- OECD, an inter-organization that collects data from 38 member countries, was considered as a second source, but after further data exploration, it was removed from the analysis as the dataset only contained wealthy countries, which did not fit the diverse sample objective. For example, one of the countries of interest in the analysis is Ukraine, which doesn’t fall into the wealthy category. Therefore, OECD data was excluded from the report.
- Kaggle.com is a user-published dataset platform, which has access to data with various licenses. The selected dataset has CC BY-NC-SA 4.0 license, which allows sharing and adaptation for non-commercial purposes with proper attribution. The dataset included over 80 countries from wealthy and poor countries, which fit the objective of the project. An import from CSV file to Jupyter Notebooks technique was used to retrieve the data after the dataset was downloaded from Kaggle.com.

Techniques

- web scraping
- regular expression to float value conversion
- download from the source website
- CSV Import to Jupyter Notebooks

Data Source

- Numbeo: Price Rankings by Country of Internet
- Kaggle: Average Monthly Internet
- License (Kaggle): CC BY-NC-SA 4.0

Steps: Internet Prices

WEB SCRAPING (NUMBEO.COM)

The request to web scrape Numbeo.com includes a check for errors using the Request/HTTPError library, in the event, the page cannot be reached, a failure message will be displayed.

In [70]:
```
import requests # import requests library
from requests.exceptions import HTTPError # HTTP error library
# try to request data from numbeo.com
try:
    r = requests.get('https://www.numbeo.com/cost-of-living/country_price_ra
    r.raise_for_status() # returns an HTTP Error object if an error happenec
except HTTPError: # raise exception
    print ('Could not download', r.url) # if data can not be downloaded prin
else:
    print (r.url, 'downloaded successfully') # otherwise, print success mess
    data = r.text # save request in a data variable
```

```
https://www.numbeo.com/cost-of-living/country_price_rankings?itemId=33 downl
oaded successfully
```

Beautiful Soup library was used to parse the data from Numbeo.com for the 2021 internet prices. The library was chosen because of its simple set of tools to create Python objects. One table containing 3 columns and 104 countries was web scraped from Numbeo.com. The headers were assigned as Rank, Country, and 2021 using a for loop, and the index was assigned to column "Rank".

In [71]:
```
from bs4 import BeautifulSoup # import beautiful soup library
soup = BeautifulSoup(data, "html.parser") # parse the HTML
table = soup.find_all("table") # get all the table elements
rows = table[1].find_all("tr") # extract all the the table row elements and
headers = ['Rank', 'Country', '2021'] # assign headers
```

In [72]:

```
# reference: https://medium.com/analytics-vidhya/how-to-web-scrape-tables-on
import pandas as pd # import panads library
numbeo_internet_prices = pd.DataFrame(columns = headers) # create dataframe
# loop over the table rows and extract table data
for j in table[1].find_all('tr')[0:]:
    row_data = j.find_all('td')
    row = [i.text.strip() for i in row_data]
    del row[2]
    length = len(numbeo_internet_prices)
    numbeo_internet_prices.loc[length] = row

numbeo_internet_prices.set_index('Rank', inplace=True) # setting index
save_web_scraped_internet_price_data = numbeo_internet_prices
```

saving web scraped data

The data in the original format was saved to prevent changes that may break further data manipulations. Pathlib library was used to check if the file already exists, then print out a message, if the file does not exist then create the file and print the message saying that.

In [73]:

```
# check if the file exists, if does not than save the file
# reference: https://www.pythontutorial.net/python-basics/python-check-if-fi
from pathlib import Path
path_to_file = './data/web_scraped_data/original/web_scrapped_data_numbeo_or

path = Path(path_to_file)

if path.is_file():
    print(f'The file {path_to_file} exists')
else:
    save_web_scraped_internet_price_data.to_csv("./data/web_scraped_data/ori
    print(f'The file {path_to_file} does not exist. The file wil be saved to
```

```
The file ./data/web_scraped_data/original/web_scrapped_data_numbeo_original.
csv exists
```

challenges

Although the web scraped data is already sorted from highest to lowest price, it was decided to print the highest and lowest-ranked internet prices for clarity of the analysis. To print out the highest and lowest numbers a simple minimum and maximum function was used. This is where the unexpected error occurred. The data from web scraping was not saved with the correct values and required further manipulation. The data was saved as a regular expression therefore the 2021 column had to be converted into a float value. See the next section for the conversion procedure.

Back to Table of Contents

REGULAR EXPRESSION TO FLOAT VALUE CONVERSION

The data was scraped in a string format, therefore, the 2021 column required conversion from regular expression to float.

reading the file

The original CSV file needs to be read before conversion from regular expression to float. It's unwise to perform conversion from live website data due to changes in the website data and its unavailability. The original data contains a Rank column which is unnecessary in further analysis, therefore, only the Country and 2021 columns were read. The index was set to Country.

In [74]:

```
numbeo_internet_prices_ww = pd.read_csv('./data/web_scraped_data/original/we
numbeo_internet_prices_ww.set_index("Country", inplace=True) # assign Countr
numbeo_internet_prices_ww.dtypes # checking data types
```

Out[74]:

```
2021    object
dtype: object
```

conversion

The ".replace" function was used to remove the "$" sign, spaces, and comma from 2021 and ".astype" function was used to convert the cleaned data from regular expression to float.

In [75]:

```
# converting values in the 2021 column from regular expression to float
numbeo_internet_prices_ww['2021'] = numbeo_internet_prices_ww['2021'].replac
save_web_scraped_internet_price_data_float = numbeo_internet_prices_ww # ass
save_web_scraped_internet_price_data_float # table view of the data
```

Out[75]:

Country	2021
United Arab Emirates	99.10
Qatar	90.13
Saudi Arabia	72.10
United States	68.64
Iceland	66.68
...	...
India	8.90
Russia	8.38
Romania	8.37
Turkey	8.36
Ukraine	4.65

104 rows × 1 columns

TABLE 1: CONVERTED DATA, SOURCE: NUMBEO.COM

saving web scraped data with float values for 2021 to a CSV file

The converted data is to be saved in a new CSV file for the purpose of further use in determining the minimum and maximum internet prices and to visualize the data on the map. Pathlib library was used to check if the file already exists, then print out a message, if the file does not exist then create the file and print the message saying that.

In [76]:

```
# check if the file exists, if does not than save the file
# reference: https://www.pythontutorial.net/python-basics/python-check-if-fi
from pathlib import Path
path_to_file = './data/web_scraped_data/converted_to_float/web_scrapped_data

path = Path(path_to_file)

if path.is_file():
    print(f'The file {path_to_file} exists')
else:
    save_web_scraped_internet_price_data_float.to_csv("./data/web_scraped_da
    print(f'The file {path_to_file} does not exist. The file wil be saved to
```

```
The file ./data/web_scraped_data/converted_to_float/web_scrapped_data_numbeo
_float.csv exists
```

Back to Table of Contents

INTERACTIVE MAP: INTERNET PRICES

Pandas library was used for data processing to read the CSV file from the data folder, rename column names, check for null values, and to find the minimum and maximum values in the dataset. In the author's opinion, it was important to use cleansed and validated data, therefore, the newly saved file was read for the interactive map visualization.

In [77]:

```
import pandas as pd # data processing
internet_prices_worldwide = pd.read_csv('./data/web_scraped_data/converted_t
internet_prices_worldwide.rename(columns = {'Country Name':'Country', '2021'
internet_prices_worldwide # display table
```

Out[77]:

	Country	2021 Internet Price ($)
0	United Arab Emirates	99.10
1	Qatar	90.13
2	Saudi Arabia	72.10
3	United States	68.64
4	Iceland	66.68
...	...	...
99	India	8.90
100	Russia	8.38
101	Romania	8.37
102	Turkey	8.36
103	Ukraine	4.65

104 rows × 2 columns

TABLE 2: CONVERTED DATA, SOURCE: NUMBEO.COM

check for null values

The data was validated before further analysis. The ".notnull and .isNull" functions were used to check for NULL values. It was decided to do it in this section and not before saving the data in the previous step to be sure the data did not get corrupted.

In [78]:

```
# check for null values
internet_prices_worldwide = internet_prices_worldwide[pd.notnull(internet_pr
internet_prices_worldwide.isnull().sum()
```

Out[78]:

```
Country                    0
2021 Internet Price ($)    0
dtype: int64
```

The data does not have NULL values, so the analysis proceeded.

maximum and minimum values

This section will use the maximum and minimum functions to determine the highest and lowest internet prices.

In [79]:

```
numbeo_max = internet_prices_worldwide[internet_prices_worldwide['2021 Inter
numbeo_max # display result
```

Out[79]:

	Country	2021 Internet Price ($)
0	United Arab Emirates	99.1

TABLE 3: MAXIMUM INTERNET PRICE, SOURCE: NUMBEO.COM

In [80]:

```
numbeo_min = internet_prices_worldwide[internet_prices_worldwide['2021 Inter
numbeo_min # display result
```

Out[80]:

	Country	2021 Internet Price ($)
103	Ukraine	4.65

TABLE 4: MINIMUM INTERNET PRICE, SOURCE: NUMBEO.COM

visualizing internet prices on the map

Plotly Express library was used to visualize global internet prices on the map. Choropleth Map style was used in the report. The representation of prices on the map provides a better understanding of the disparity on a global scale. The map shows that the dataset has some gaps, for example, a large territory in Africa is not covered and there are some gaps in South America and Asia.

Map Colors: The colors on the map represent internet prices, the darker the color, the higher the price of internet in that country.

The code was adopted from Shawkat Sujon.

In [81]:

```
# reference: https://plotly.com/python/choropleth-maps/
import plotly.express as px

fig = px.choropleth(internet_prices_worldwide, locations='Country', location
fig.update_layout(margin={'r':0,'t':0,'l':0,'b':0}, coloraxis_colorbar=dict(
    title = 'Price $ (2021)',
    ticks = 'outside',
    tickvals = [10,20, 30, 40, 50, 60, 70, 80, 90, 100],
    dtick = 7
))
fig.show()
```

FIGURE 1: GLOBAL MAP OF INTERNET PRICES, SOURCE: NUMBEO.COM, Code Source: SHAWKAT SUJON

filtering data for 10 countries

Panadas library was used to find 10 countries in the dataset (Australia, Canada, France, United States, United Arab Emirates, Argentina, Peru, India, Sri Lanka, and Ukraine) in preparation to save the findings as a new CSV file. A new column was inserted and country codes were added to the dataset. The code is manually assigning values to each row in the Country Code column without checking for correctness in the event the file is not filtered in the same order. Due to timing constraints, the code was not refactored. This could be done in the future.

In [82]:

```
import pandas as pd # import pandas library
# read csv
ws_internet_prices=pd.read_csv("./data/web_scraped_data/converted_to_float/w
# assign Country column as index
ws_internet_prices.set_index("Country", inplace=True)
# look up 10 selected countries
lookup_internet_prices_10 = ws_internet_prices.loc[ws_internet_prices.index.
lookup_internet_prices_10.insert(0, "Country Code", ['ARE', 'USA', 'CAN', 'A
# save findings in a separate csv file
save_internet_prices_10 = lookup_internet_prices_10
```

saving data for 10 countries

The dataset for 10 countries will be saved for further analysis to compare internet prices against internet usage rate, GDP per Capita, PPP, and wages. Pathlib library was used to check if the file already exists, then print out a message, if the file does not exist then create the file and print the message saying that.

In [83]:

```
# check if the file exists, if does not than save the file
# reference: https://www.pythontutorial.net/python-basics/python-check-if-fi
from pathlib import Path
path_to_file = './data/10_selected_countries/numbeo_internet_prices_10_count

path = Path(path_to_file)

if path.is_file():
    print(f'The file {path_to_file} exists')
else:
    save_internet_prices_10.to_csv("./data/10_selected_countries/numbeo_inte
    print(f'The file {path_to_file} does not exist. The file wil be saved to
```

```
The file ./data/10_selected_countries/numbeo_internet_prices_10_countries.cs
v exists
```

findings

Table 3 and **Table 4** show that the highest-ranking country is the United Arab Emirates with a price of USD99.10 and the lowest-ranking country is Ukraine with a price of internet of USD4.65.

Back to Table of Contents

CSV DATA (KAGGLE)

Pandas library was used for data processing to read the CSV file from the data folder, set the index, and find the minimum and maximum values in the dataset.

In [84]:

```
import pandas as pd # import pandas libary
kaggle_avg_internet_price = pd.read_csv("./data/kaggle_data/kaggle_average_m
kaggle_avg_internet_price.rename(columns = {'Country Name':'Country', '2021'
kaggle_avg_internet_price # display
```

Out[84]:

	Country	2021 Internet Price ($)
0	Albania	20.52
1	Algeria	40.05
2	Argentina	23.88
3	Australia	58.17
4	Austria	43.05
...	...	...
82	United Arab Emirates	100.09
83	United Kingdom	42.77
84	United States	68.55
85	Uruguay	38.31
86	Vietnam	10.92

87 rows × 2 columns

TABLE 5: INTERNET PRICES, SOURCE: KAGGLE.COM

world cloud

The word cloud was created with matplotlib.pyplot, WordCloud, and STOPWORDS libraries to visualize internet prices for the Kaggle dataset.

In [85]:

```
# reference: https://www.javatpoint.com/wordcloud-package-in-python
!pip install wordcloud

import matplotlib.pyplot as plt
from wordcloud import WordCloud, STOPWORDS
# text variable
text = " ".join(cat for cat in kaggle_avg_internet_price.Country)
# create word cloud
word_cloud = WordCloud(
        width=2000,
        height=2000,
        random_state=1,
        background_color="green",
        colormap="Set3",
        collocations=False,
        stopwords=STOPWORDS,
        ).generate(text)
# plot the word cloud
plt.imshow(word_cloud)
plt.axis("off")
plt.show()
```

```
Requirement already satisfied: wordcloud in /Users/tech/anaconda3/lib/python
3.10/site-packages (1.9.3)
Requirement already satisfied: numpy>=1.6.1 in /Users/tech/anaconda3/lib/pyt
hon3.10/site-packages (from wordcloud) (1.23.5)
Requirement already satisfied: pillow in /Users/tech/anaconda3/lib/python3.1
0/site-packages (from wordcloud) (9.4.0)
Requirement already satisfied: matplotlib in /Users/tech/anaconda3/lib/pytho
n3.10/site-packages (from wordcloud) (3.7.0)
Requirement already satisfied: contourpy>=1.0.1 in /Users/tech/anaconda3/li
b/python3.10/site-packages (from matplotlib->wordcloud) (1.0.5)
Requirement already satisfied: cycler>=0.10 in /Users/tech/anaconda3/lib/pyt
hon3.10/site-packages (from matplotlib->wordcloud) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in /Users/tech/anaconda3/li
b/python3.10/site-packages (from matplotlib->wordcloud) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /Users/tech/anaconda3/li
b/python3.10/site-packages (from matplotlib->wordcloud) (1.4.4)
Requirement already satisfied: packaging>=20.0 in /Users/tech/anaconda3/lib/
python3.10/site-packages (from matplotlib->wordcloud) (22.0)
Requirement already satisfied: pyparsing>=2.3.1 in /Users/tech/anaconda3/li
b/python3.10/site-packages (from matplotlib->wordcloud) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in /Users/tech/anaconda
3/lib/python3.10/site-packages (from matplotlib->wordcloud) (2.8.2)
Requirement already satisfied: six>=1.5 in /Users/tech/anaconda3/lib/python
3.10/site-packages (from python-dateutil>=2.7->matplotlib->wordcloud) (1.16.
0)

[notice] A new release of pip is available: 23.2.1 -> 24.0
[notice] To update, run: pip install --upgrade pip
```

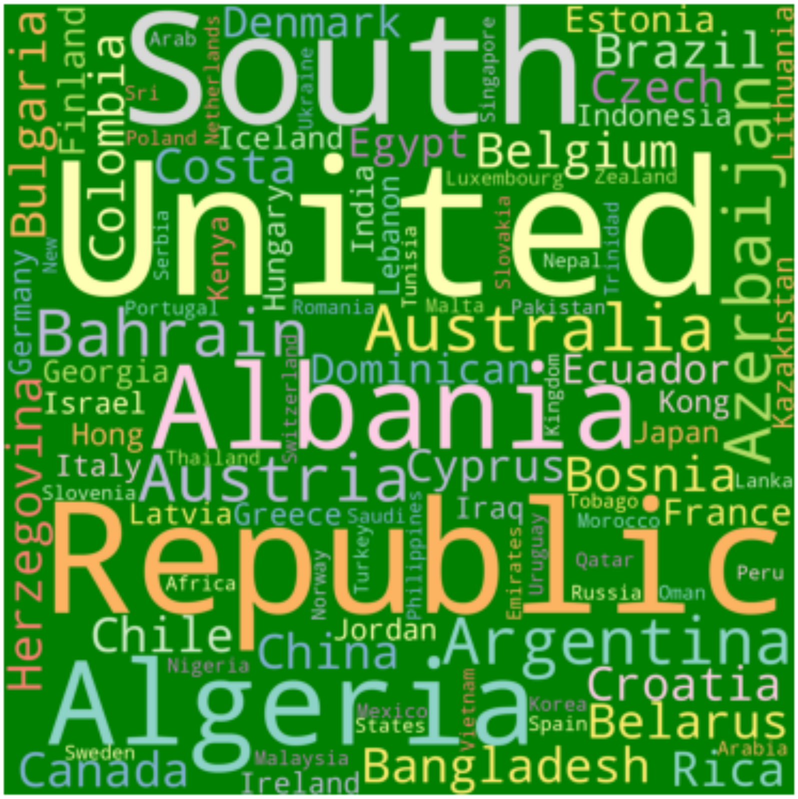


FIGURE 2: WORD CLOUD, SOURCE: KAGGLE.COM

maximum and minimum values

In [86]:

```
kaggle_max = kaggle_avg_internet_price[kaggle_avg_internet_price['2021 Inter
kaggle_max # display result
```

Out[86]:

	Country	2021 Internet Price ($)
82	United Arab Emirates	100.09

TABLE 6: MAXIMUM INTERNET PRICE, SOURCE: KAGGLE.COM

In [87]:

```
kaggle_min = kaggle_avg_internet_price[kaggle_avg_internet_price['2021 Inter
kaggle_min # display result
```

Out[87]:

	Country	2021 Internet Price ($)
81	Ukraine	5.99

TABLE 7: MINIMUM INTERNET PRICE, SOURCE: KAGGLE.COM

findings

Similarly, to web scraped data from Numbeo.com, **Table 6** and **Table 7** show that the highest-ranking country is the United Arab Emirates with a price of internet of USD100.09 and the lowest-ranking country is Ukraine with a price of internet of USD5.99.

CONCLUSION

Data from Numbeo.com and Kaggle.com shows that both the United Arab Emirates and Ukraine are the highest and lowest-ranked countries in internet prices, respectively.

The comparison of prices from both sources shows a marginal difference of 0.99 and 1.34 difference between the maximum and minimum prices of the internet. It's possible that Numbeo.com and Kaggle.com aggregated data from different providers, therefore, the price average calculation from each source varies slightly. It's important to note that data from Kaggle.com is static and data from Numbeo.com is dynamic, meaning periodically updated. After reflection, it was decided that in spite of the marginal difference between the two sources the general conclusion of the highest and lowest internet price is consistent therefore it's reasonable to use the data knowing that there is a slight price variation. For the purposes of the report, the data collected from Numbeo.com will be considered the correct data and will be used in further analysis.

The data from the internet price analysis posed a question; why are the maximum and minimum values so far apart? In other words, what influences such a big price difference? Could it be due to the internet adoption technology being on two different spectrums for the United Arab Emirates and Ukraine?

Back to Table of Contents

Internet Usage Rate

What is the internet usage rate around the world?

This section will analyze the internet usage rate with the purpose to understand if the rate of internet adoption influences internet prices.

The data integrity was validated based on the following reasoning.

- The World Bank is a financial institution that provides financial services to governments. It provides access to an “Open Knowledge Repository”. (Wikipedia, World Bank, n.d.) The selected dataset has a CC BY 4.0 license, which allows the copying, sharing, remixing, and transformation of the data for any purpose with proper attribution and indication if any changes have been made. The World Bank is a legitimate institution that provides access to data, collected from 189 countries it cooperates. The “Individuals using the Internet (% of population)” contains data from International Telecommunication Union (ITU) World Telecommunication, governmentally aggregated data. Therefore, it’s unnecessary to use any additional resources for the validation of data for correctness. The dataset includes over 250 countries from wealthy and poor countries, which fits the objective of the project.

Could internet prices be affected by how technologically developed the infrastructure in the country is? If internet usage is low, could that drive the prices up? And if the usage is high, could that drive the prices down?

Techniques

- conversion of the dataset from Excel to CSV file
- statistics analysis with the pandas library
- saving findings as a separate CSV file
- perform data visualization with the pandas library
- perform data visualization with the mathplot library
- merge two datasets

Data Sources

Data source was referenced as per The World Bank requirements.

Individuals using the Internet (% of population)

The World Bank (Bank, GDP per capita, PPP (current international $), n.d.)

Dataset name: Individuals using the Internet (% of population)

Data source: International Telecommunication Union (ITU) World Telecommunication/ICT Indicators Database

License: CC BY-4.0

Steps: Internet Usage Rate

CHALLENGES

The dataset for 2021 was not available, therefore, the analysis is based on 2020 data. Considering that the internet usage rates don't change drastically, it is reasonable to use the available 2020 dataset for the analysis.

CONVERTING DATA FROM EXCEL TO CSV

Pandas library was used for data processing such as reading Excel file and converting it to CSV format. The OS library was used to check if the file exists in the specified directory, if yes, then print the message, if no, save the file and print a message.

In [88]:

```
!pip install xlrd

import pandas as pd
import os # to access operating system functionality
# read excel file
read_worldbank_file = pd.read_excel('./data/worldbank_data/import_internet_u

file_path = r'./data/worldbank_data/import_internet_usage/worldbank_API_IT.N
if os.path.exists(file_path):
    print('file already exists')
else:
    # convert excel to csv format
    read_worldbank_file.to_csv ('./data/worldbank_data/import_internet_usage
    print(f'The file {path_to_file} does not exist. The file wil be saved to
Requirement already satisfied: xlrd in /Users/tech/anaconda3/lib/python3.10/
site-packages (2.0.1)

[notice] A new release of pip is available: 23.2.1 -> 24.0
[notice] To update, run: pip install --upgrade pip
file already exists
```

working with data

Pandas library was used to read the CSV file from the specified directory, set index, convert the 2020 column to float and remove NULL values. Numpy library was used to round the float value to 2 decimals.

The 2020 column represents data in percentage values. The only way it was possible to add a percentage sign to the code was to first convert the 2020 column to a string and then concatenate the % sign. The code would have to be reused further in the analysis and requires float values to perform data visualization, which means the operation would have to be reversed. This was a lengthy process for very little reward, therefore, the data is represented in float values herein. It may be possible to present the 2020 column in percentages and use it in Bar Graph visualization, but due to timing constraints, this part of the code was not researched further.

In [89]:

```
import numpy as np
# read file
worldbank_internet_usage=pd.read_csv('./data/worldbank_data/import_internet_
worldbank_internet_usage.set_index("Country Name", inplace=True)
# # convert columns from string to float
worldbank_internet_usage[['2020']].astype(float)
# # round value to 2 decimal places
worldbank_internet_usage[['2020']] = np.round(worldbank_internet_usage[['202
# # remove NAN values
df_withoutnan=worldbank_internet_usage.dropna(subset=['2020'])
# # show first 10 rows
internet_usage_worldwide = df_withoutnan
```

In [90]:

```
df_withoutnan.head()
```

Out[90]:

	Country Code	2020
Country Name		
Africa Eastern and Southern	AFE	27.35
Afghanistan	AFG	18.40
Africa Western and Central	AFW	34.13
Angola	AGO	36.00
Albania	ALB	72.24

TABLE 8: FIRST 5 COUNTRIES IN DATASET, SOURCE: WORLDBANK.ORG

In [91]:

```
df_withoutnan.tail()
```

Out[91]:

	Country Code	2020
Country Name		
Vietnam	VNM	70.30
World	WLD	59.94
South Africa	ZAF	70.00
Zambia	ZMB	19.80
Zimbabwe	ZWE	29.30

TABLE 9: LAST 5 COUNTRIES IN DATASET, SOURCE: WORLDBANK.ORG

In [92]:

```
df_withoutnan.describe()
```

Out[92]:

	2020
count	197.000000
mean	63.046041
std	25.840064
min	6.500000
25%	37.800000
50%	70.400000
75%	84.990000
max	100.000000

TABLE 10: STATISTICS, SOURCE: WORLDBANK.ORG

The lowest usage rate is 6.5% and the highest is 100%.

saving cleansed data in a separate file

The cleansed dataset for worldwide countries will be saved for further analysis to visualize the data on a map. Pathlib library was used to check if the file already exists, then print out a message, if the file does not exist then create the file and print the message saying that.

In [93]:

```python
# check if the file exists, if does not than save the file
# reference: https://www.pythontutorial.net/python-basics/python-check-if-fi
from pathlib import Path

path_to_file = './data/worldbank_data/import_internet_usage/cleaned/worldban
path = Path(path_to_file)

if path.is_file():
    print(f'The file {path_to_file} exists')
else:
    internet_usage_worldwide.to_csv("./data/worldbank_data/import_internet_u
    print(f'The file {path_to_file} does not exist. The file wil be saved to
```

```
The file ./data/worldbank_data/import_internet_usage/cleaned/worldbank_inter
net_usage_worldwide.csv exists
```

visualizing internet usage on the map

Pandas library was used to read cleansed data from CSV and rename columns.

Plotly Express library was used to visualize the global internet usage rate on the map. Choropleth Map style was used in the report. The representation of the internet usage rate on the map provides a better understanding of the disparity on a global scale. The map shows that the dataset has some gaps, for example, some territories in Africa and South America are not covered.

Map Colors: The colors on the map represent internet usage, the darker the color, the higher the internet usage in that country.

The code was adopted from Shawkat Sujon.

```
In [94]: internet_usage_worldwide = pd.read_csv('./data/worldbank_data/import_interne
         internet_usage_worldwide.rename(columns = {'Country Name':'Country', '2020':
         # reference: https://plotly.com/python/choropleth-maps/
         import plotly.express as px
         # map figure
         fig = px.choropleth(internet_usage_worldwide, locations='Country', locationm
         fig.update_layout(margin={'r':0,'t':0,'l':0,'b':0}, coloraxis_colorbar=dict(
             title = 'Usage % (2020)', # title of the ticks section
             ticks = 'outside', # location of ticks
             tickvals = [10, 20,30, 40, 50,60, 70, 80, 90, 100], # ticks based on val
             dtick = 5 # tick step
         ))
         fig.show() # display figure
```

FIGURE 3: GLOBAL MAP OF INTERNET USAGE RATE (% OF POPULATION), SOURCE: WORLDBANK.ORG, CODE SOURCE: SHAWKAT SUJON

filtering data for 10 countries

Panadas library was used to find 10 countries in the dataset (Australia, Canada, France, United States, United Arab Emirates, Argentina, Peru, India, Sri Lanka, and Ukraine) in preparation to save the findings as a new CSV file.

In [95]:

```
# locate 10 countries
lookup_internet_usage_10 = df_withoutnan.loc[df_withoutnan.index.str.contain
save_internet_usage_10 = lookup_internet_usage_10 # new variable
save_internet_usage_10 # display result
```

Out[95]:

	Country Code	2020
Country Name		
United Arab Emirates	ARE	100.00
Argentina	ARG	85.50
Australia	AUS	89.60
Canada	CAN	96.97
France	FRA	84.80
India	IND	43.00
Sri Lanka	LKA	35.00
Peru	PER	65.25
Ukraine	UKR	75.04
United States	USA	90.90

TABLE 11: 2020 INTERNET USAGE RATE IN %, SOURCE: THE WORLD BANK

findings

The United Arab Emirates has 100% internet usage. In contrast, Sri Lanka has a 35% internet usage rate, showing a large difference between the two countries.

saving data for 10 countries

The dataset for 10 countries will be saved for further analysis to compare internet prices against internet usage rates. Pathlib library was used to check if the file already exists, then print out a message, if the file does not exist then create the file and print the message saying that.

In [96]:

```
# check if the file exists, if does not than save the file
# reference: https://www.pythontutorial.net/python-basics/python-check-if-fi
from pathlib import Path

path_to_file = './data/10_selected_countries/worldbank_internet_usage_rate_1
path = Path(path_to_file)

if path.is_file():
    print(f'The file {path_to_file} exists')
else:
    save_internet_usage_10.to_csv("./data/10_selected_countries/worldbank_in
    print(f'The file {path_to_file} does not exist. The file wil be saved to
```

```
The file ./data/10_selected_countries/worldbank_internet_usage_rate_10_count
ries.csv exists
```

Back to Table of Contents

MERGING TWO DATASETS: INTERNET PRICES AND INTERNET USAGE RATES

Pandas library was used to read datasets, merge them together, set the index to start from 1, remove unnecessary columns after the merge operation and rename columns to clarify the presented data.

In [97]:

```
import pandas as pd # import pandas libary
numbeo_internet_price_10 = pd.read_csv("./data/10_selected_countries/numbeo_
internet_usage_rate_10 = pd.read_csv("./data/10_selected_countries/worldbank
price_vs_usage = pd.merge(numbeo_internet_price_10, internet_usage_rate_10,
price_vs_usage.index = price_vs_usage.index + 1 # set index to start from 1
price_vs_usage.drop(price_vs_usage.columns[-2],axis=1,inplace=True) # remove
price_vs_usage.rename(columns = {'2021':'Price $ (2021)', '2020':'Usage % (2
price_vs_usage
```

Out[97]:

	Country	Country Code	Price $ (2021)	Usage % (2020)
1	United Arab Emirates	ARE	99.64	100.00
2	United States	USA	68.64	90.90
3	Canada	CAN	61.70	96.97
4	Australia	AUS	52.12	89.60
5	Peru	FRA	31.53	84.80
6	France	PER	31.20	65.25
7	Argentina	ARG	21.77	85.50
8	Sri Lanka	LKA	9.50	35.00
9	India	IND	8.93	43.00
10	Ukraine	UKR	4.65	75.04

TABLE 12: INTERNET PRICES AND INTERNET USAGE OF 10 COUNTRIES, SOURCE: NUMBEO.COM AND WORLDBANK.ORG

data visualization: internet prices vs internet usage rate

The merged data contained USD and percentage values. The bar chart was used to visualize the comparison between internet prices and internet usage rates for 10 countries. Although this comparison is not mathematically accurate, it shows a general understanding of the difference between price and usage.

Matplotlib.pyplot library was used to create the comparison bar chart.

In [98]:

```
import matplotlib.pyplot as plt # import library
price_vs_usage.plot(x="Country", y=["Price $ (2021)", "Usage % (2020)"], kin
plt.title("Price vs Usage",fontsize=14 ) # title of the plot
plt.show()
```

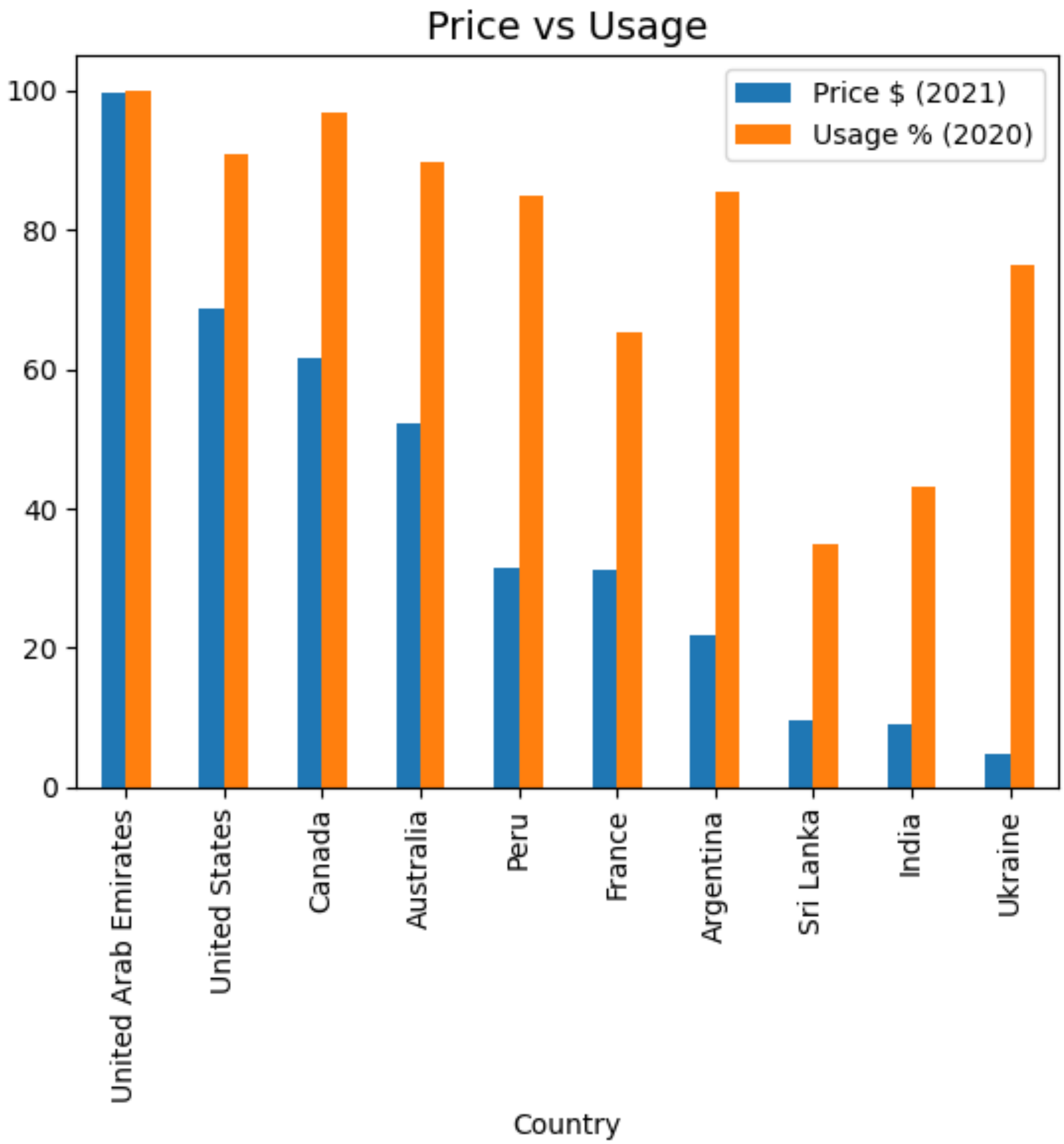


FIGURE 4: INTERNET PRICE VS INTERNET USAGE RATE, SOURCE: NUMBEO.COM AND WORLDBANK.ORG

findings

Ukraine has the biggest difference between internet price and internet usage. What's interesting is that internet usage and internet price in the United Arab Emirates are both high.

CONCLUSION

The hypothesis was that if the internet usage rate was low the price would be high and vice versa. This proved not to be the case, for example, in **Figure 4** the data shows that UAE has 100% internet adoption, yet it has the highest price. In other words, price and usage alone do not explain global internet prices. Other factors are at work.

Back to Table of Contents

GDP per capita versus GDP per capita PPP

What is the GDP per capita?

What is the GDP per capita, PPP?

What is the GDP per capita indicator for Ukraine?

What is the GDP per capita indicator for countries worldwide?

What is the GDP per capita, PPP indicator for Ukraine?

What is the GDP per capita, PPP indicator for countries worldwide?

Which indicator to use to calculate the ratio between indicator and internet price?

What is the Penn Effect?

This section will describe the meaning of GDP per capita, GDP per capita, PPP indicators, and the Penn Effect. It will look at the indicators for 10 countries, determine which indicator to use in further analysis, and calculate the ratio between internet prices and the chosen indicator.

GDP per capita: “Growth is calculated from constant price GDP data in local currency. Sustained economic growth increases average incomes and is strongly linked to poverty reduction. GDP per capita provides a basic measure of the value of output per person, which is an indirect indicator of per capita income. Growth in GDP and GDP per capita is considered broad measures of economic growth.” (Bank, Metadata Glossary, n.d.)

GDP per capita, PPP: “GDP per capita based on purchasing power parity (PPP). PPP GDP is gross domestic product converted to international dollars using purchasing power parity rates. An international dollar has the same purchasing power over GDP as the U.S. dollar has in the United States. GDP at purchaser's prices is the sum of gross value added by all resident producers in the country plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in constant 2017 international dollars." (Bank, Metadata Glossary, n.d.)

Penn Effect: "The overstatement of the economic size of high-income countries and the understatement of the economic size of low-income countries that results when exchange rate converted GDPs (Gross domestic product) is used to establish the relative sizes of economies. The Penn effect arises because price levels are usually higher in high-income countries than they are in low-income countries and exchange rates do not take account of price level differences between countries when used to convert their GDPs to a common currency." (Eurostat, 2013-2019)

The data integrity was validated based on the following reasoning.

- The World Bank is a financial institution that provides financial services to governments. It provides access to an "Open Knowledge Repository". (Wikipedia, World Bank, n.d.) The selected dataset has a CC BY 4.0 license, which allows the copying, sharing, remixing, and transformation of the data for any purpose with proper attribution and indication if any changes have been made. The World Bank is a legitimate institution that provides access to data, collected from 189 countries it cooperates. The GDP repository contains data from International Comparison Program, World Development Indicators database, Eurostat-OECD PPP Program, World Bank national accounts data, and OECD National Accounts data files, all governmentally aggregated data. Therefore, it's unnecessary to use any additional resources for the validation of data for correctness. The dataset includes over 189 countries from wealthy and poor countries, which fits the objective of the project.

Then, to perform an analysis it's necessary to evaluate if GDP per capita and GDP per capita, PPP are the same. And lastly, based on the findings, use one indicator to calculate the ratio between internet prices and the indicator to understand if internet prices are influenced by GDP.

Techniques

- request data through the use of World Bank API (wbgapi)
- perform data visualization with pandas library
- perform data visualization with mathplot library
- merge two datasets

Data Sources

Data source was referenced as per The World Bank requirements.

GDP per capita, PPP

The World Bank (Bank, GDP per capita, PPP (current international $), n.d.)

Dataset name: GDP per capita, PPP (current international $)

Data source: International Comparison Program, World Bank | World Development Indicators database, World Bank | Eurostat-OECD PPP Programme

License: CC BY-4.0

GDP per capita

The World Bank (Bank, GDP per capita, PPP (current international $), n.d.)

Dataset name: GDP per capita (current US$)

Data source: World Bank national accounts data, and OECD National Accounts data files.

License: CC BY-4.0

Steps: GDP per capita versus GDP per capita PPP

CHALLENGES

The use of wbgapi required knowledge of World Bank acronyms to request and manipulate the data. This section will include World Bank API helper functions, which can be uncommented to validate the following acronyms.

"**LIC:** low-income countries", "**LMC:** lower-middle-income countries", "**UMC:** upper-middle-income countries", "**HIC:** high-income countries", source: The World Bank.

"HIC": "AUS = Australia", "CAN = Canada", "FRA = France", "LUX = United States", "ARE = United Arab Emirates", source: The World Bank.

"UMC": "ARG = Argentina", "PER = Peru", source: The World Bank.

"LMC": "IND = India", "LKA = Sri Lanka", "UKR = Ukraine", source: The World Bank.

REQUEST DATA FROM WORLD BANK TO ANALYZE GDP PER CAPITA AND GDP PER CAPITA, PPP INDICATORS FOR UKRAINE

This section will make a request to the World Bank API to download the GDP per capita and GDP per capita, PPP datasets for Ukraine. Then visualize data to perform a comparison of two indicators from 2013 to 2021 in two year-increment. The 2-year increment was chosen to reduce the density of the chart.

During market research for another project, it was discovered that GDP per capita and GDP per capita, PPP for Ukraine are different. This suggested that it's important to conduct a comparison of indicators to understand if there is an overestimation for high-performing countries and an underestimation for low-performing countries related to the currency exchange conversion. The analysis will aim to determine if Penn Effect is applicable to GDP per capita and GDP per capita, PPP indicators. This is an important step in the further analysis as it will determine if the GDP per capita or GDP per capita, PPP indicator will be used to calculate the ratio between an indicator and internet prices.

data retrieval

The World Bank API was used to retrieve data for GDP per capita and GDP per capita, PPP. The used data was also downloaded in the CSV format in the event API request takes a long time or is unavailable during the grading process.

In [99]:

```
!pip install wbgapi

import wbgapi as wb # importing world bank library
```

```
Requirement already satisfied: wbgapi in /Users/tech/anaconda3/lib/python3.1
0/site-packages (1.0.12)
Requirement already satisfied: requests in /Users/tech/anaconda3/lib/python
3.10/site-packages (from wbgapi) (2.31.0)
Requirement already satisfied: PyYAML in /Users/tech/anaconda3/lib/python3.1
0/site-packages (from wbgapi) (6.0)
Requirement already satisfied: tabulate in /Users/tech/anaconda3/lib/python
3.10/site-packages (from wbgapi) (0.8.10)
Requirement already satisfied: charset-normalizer<4,>=2 in /Users/tech/anaco
nda3/lib/python3.10/site-packages (from requests->wbgapi) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /Users/tech/anaconda3/lib/pyt
hon3.10/site-packages (from requests->wbgapi) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /Users/tech/anaconda3/l
ib/python3.10/site-packages (from requests->wbgapi) (1.26.14)
Requirement already satisfied: certifi>=2017.4.17 in /Users/tech/anaconda3/l
ib/python3.10/site-packages (from requests->wbgapi) (2022.12.7)

[notice] A new release of pip is available: 23.2.1 -> 24.0
[notice] To update, run: pip install --upgrade pip
```

helper functions

The following functions check for indicators, country id, country name, region, and income level. The code is commented out so as not to make the report longer.

In [100…

```
# help(wb) # help on wbgapi
# wb.source.info() # World Bank indicators
# wb.economy.info() # country id, country name, region and income level
```

Back to Table of Contents

GDP PER CAPITA VS GDP PER CAPITA, PPP FOR UKRAINE

Pandas library was used to merge GDP per capita and GDP per capita, PPP datasets for Ukraine, then columns were renamed and the index was reset.

Matplotlib.pyplot was used to create a comparison bar chart between two indicators for Ukraine.

In [101…

```
import pandas as pd
import matplotlib.pyplot as plt

# get data from Word Bank API
gdppercap_ukr=wb.data.DataFrame('NY.GDP.PCAP.CD',
                      ['UKR'],
                      time=range(2013,2022,2))

ppp_ukr_index = gdppercap_ukr.reset_index() # reset index

# get data from Word Bank API
gdppercap_ppp_ukr=wb.data.DataFrame('NY.GDP.PCAP.PP.CD',
                      ['UKR'],
                      time=range(2013,2022,2))

gdppercap_ukr_index = gdppercap_ppp_ukr.reset_index() # reset index
gdp_vs_ppp = pd.merge(ppp_ukr_index, gdppercap_ukr_index, on='economy', how=
gdp_vs_ppp.rename(columns = {"YR2021_x":"GDP 2021",'YR2021_y':'PPP 2021', 'e
gdp_vs_ppp.plot(x="Country Code", y=[ "GDP 2021",'PPP 2021' ] , kind="bar")
plt.title("GDP per capita vs GDP per capita, PPP for Ukraine",fontsize=14 )
plt.show()
```

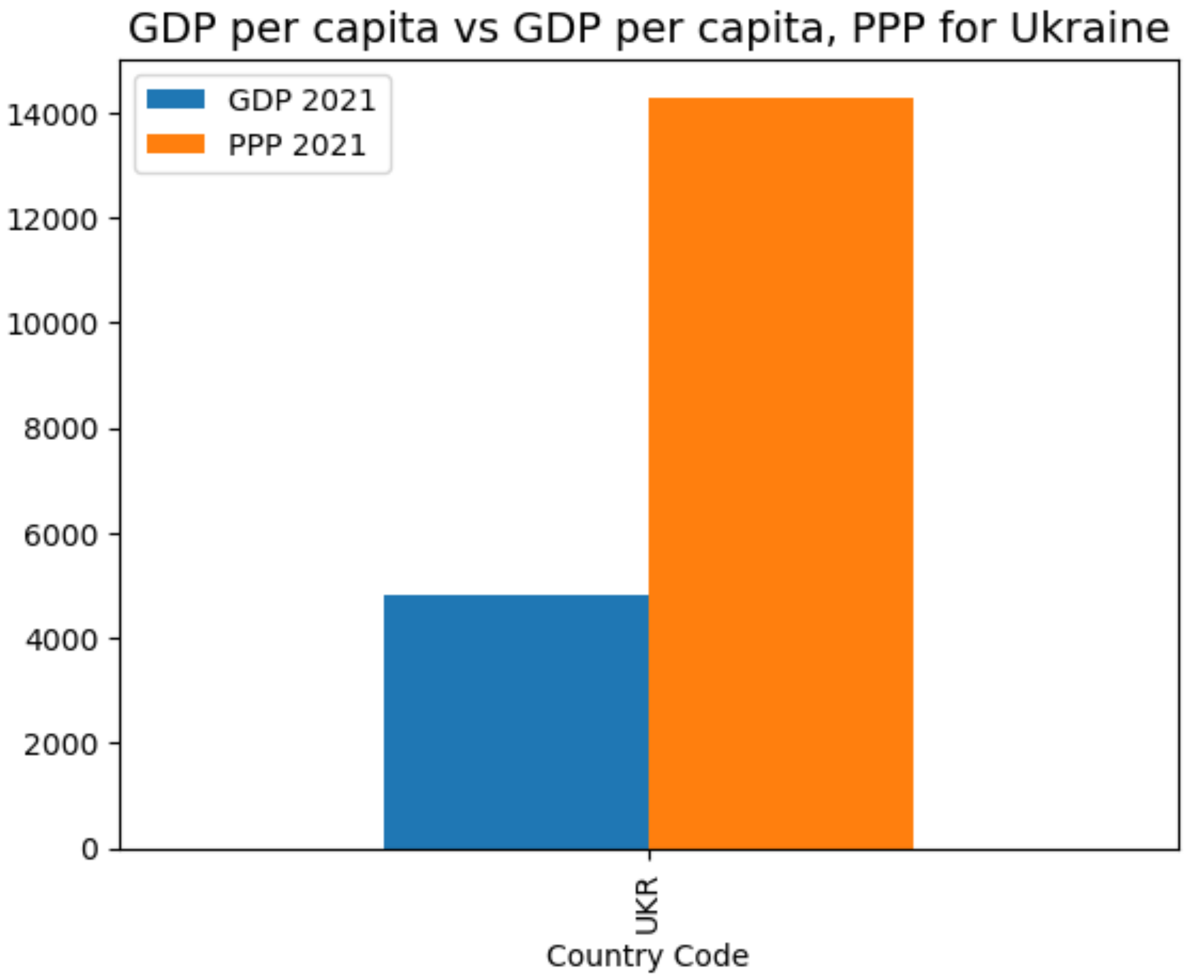


FIGURE 5: GDP PER CAPITA VS GDP PER CAPITA, PPP FOR UKRAINE, SOURCE: WORLDBANK.ORG

findings

The comparison analysis shows that GDP per capita and GDP per capita, PPP for Ukraine are different. To make a proper analysis it's necessary to look at a larger sample.

Back to Table of Contents

COMPARISON OF GDP PER CAPITA AND GDP PER CAPITA, PPP INDICATORS FOR 10 COUNTRIES

Pandas library was used to select 10 countries from GDP per capita and GDP per capita, PPP datasets.

Matplotlib.pyplot was used to create a comparison bar chart between 10 countries for each indicator. The 2013-2021 period was used to have a larger sample to understand if Penn Effect is present in the datasets.

The purpose is to display both indicators in one bar chart to simplify the analysis. The visualization will display GDP per capita and GDP per capita, PPP indicators from 2013 to 2021 in two year-increment. The 2-year increment was chosen to reduce the density of the chart. To make an educated analysis, it was necessary to assess more than one year to better understand the trend.

In [102…

```
import pandas as pd
import matplotlib.pyplot as plt
# GDP per capita indicator
gdppercap=wb.data.DataFrame('NY.GDP.PCAP.CD',
                      ['USA','ARE', 'LKA', 'PER','UKR', 'ARG', 'AUS', 'CAN',
                      time=range(2013, 2022,2))
# GDP per capita PPP indicator
gdppercap_ppp=wb.data.DataFrame('NY.GDP.PCAP.PP.CD',
                      ['USA','ARE', 'LKA','PER','UKR', 'ARG', 'AUS', 'CAN',
                      time=range(2013,2022,2))

# plot comparing two indicators
plt.figure()
fig,ax = plt.subplots()
ax1=gdppercap.plot(kind='bar', color= 'blue', ax=ax) # blue is for GDP per c
ax2=gdppercap_ppp.plot( kind='bar',  color= 'orange', ax=ax) # orange is for

plt.title("GDP per capita vs GDP per capita, PPP", fontsize=14 ) # title of
plt.show()
```

```
<Figure size 640x480 with 0 Axes>
```

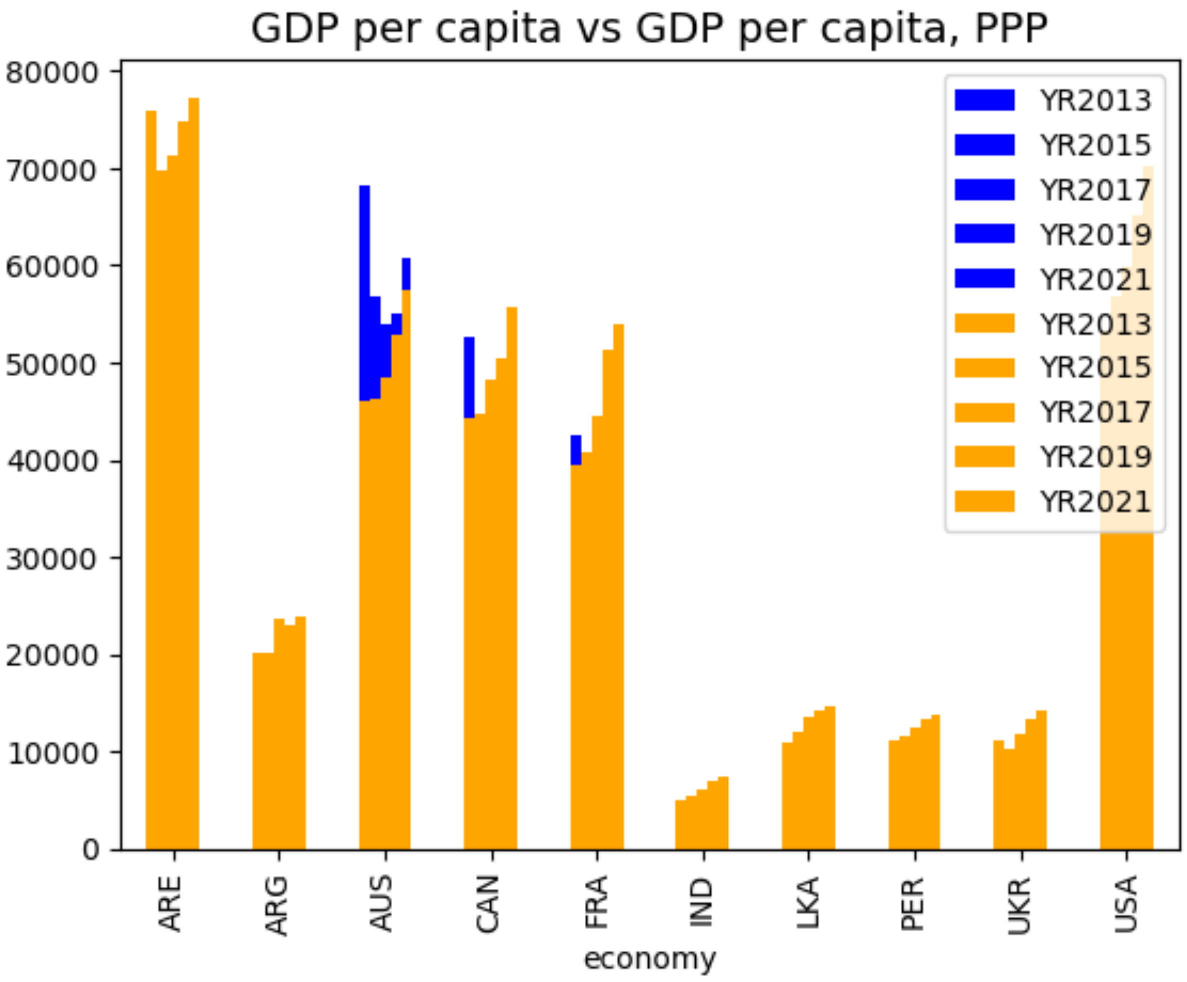


FIGURE 6: GDP PER CAPITA AND GDP PER CAPITA, PPP FOR 10 COUNTRIES, SOURCE: WORLDBANK.ORG

findings

Figure 6 demonstrates that in high-income countries such as Australia, Canada, and France the GDP per capita is greater than the GDP per capita, PPP. In low-income countries such as India, Sri Lanka, and Ukraine, the GDP per capita is lesser than the GDP per capita, PPP. The analysis also shows that there is no difference in the GDP per capita and the GDP per capita, PPP for the United Arab Emirates.

Considering that four out of three high-income countries showed differences in indexes, it will be considered that the Penn Effect is applicable to this dataset. Therefore, the GDP per capita, PPP will be used to calculate the ratio between the indicator and internet prices.

prepare for saving data

The data needed a small adjustment before saving it. The further analysis only requires GDP per capita, PPP data for the 2021 year, and the index to be set to the Country Code.

In [103…

```
ppp=gdppercap_ppp[['YR2021']] # only save index and 2021 column
ppp.index.names = ['Country Code']
save_ppp_10 = ppp # new variable
```

saving data for 10 countries

The dataset for 10 countries will be saved for further analysis to compare internet prices against GDP per Capita, PPP. Pathlib library was used to check if the file already exists, then print out a message, if the file does not exist then create the file and print the message saying that.

In [104…

```
# check if the file exists, if does not than save the file
# reference: https://www.pythontutorial.net/python-basics/python-check-if-fi
from pathlib import Path

path_to_file = './data/10_selected_countries/gdp_ppp_10.csv'
path = Path(path_to_file)

if path.is_file():
    print(f'The file {path_to_file} exists')
else:
    save_ppp_10.to_csv("./data/10_selected_countries/gdp_ppp_10.csv")
    print(f'The file {path_to_file} does not exist. The file wil be saved to
```

```
The file ./data/10_selected_countries/gdp_ppp_10.csv exists
```

Back to Table of Contents

MERGING DATA: INTERNET PRICES AND GDP PER CAPITA, PPP

Pandas library was used to read and merge two datasets, then the index was set to Country and columns were renamed. The library was also used to insert a column and calculate the ratio between internet prices and GDP per capita, PPP, saved under the Ratio name. GDP per capita, PPP data is not available for the United Arab Emirates. Nonetheless, it was decided to proceed with the analysis. The United Arab Emirates will not be discussed in the conclusion of this section.

In [105…

```
# prepare for merging datasets
numbeo_internet_price_10 = pd.read_csv("./data/10_selected_countries/numbeo_
gdp_ppp_10 = pd.read_csv("./data/10_selected_countries/gdp_ppp_10.csv") # in
price_vs_gdp_ppp = pd.merge(numbeo_internet_price_10, gdp_ppp_10, on='Countr
price_vs_gdp_ppp.set_index("Country", inplace=True) # assign Country column 
price_vs_gdp_ppp.rename(columns = {'2021':'Price $ (2021)','YR2021':'PPP $ (
price_vs_gdp_ppp['Ratio'] = price_vs_gdp_ppp['Price $ (2021)'] / price_vs_gd
price_vs_gdp_ppp
```

Out[105]:

	Country Code	Price $ (2021)	PPP $ (2021)	Ratio
Country				
United Arab Emirates	ARE	99.64	NaN	NaN
United States	USA	68.64	69287.536588	0.000991
Canada	CAN	61.70	52085.035685	0.001185
Australia	AUS	52.12	55807.444025	0.000934
Peru	FRA	31.53	50728.667429	0.000622
France	PER	31.20	13895.275816	0.002245
Argentina	ARG	21.77	23627.394294	0.000921
Sri Lanka	LKA	9.50	14127.211279	0.000672
India	IND	8.93	7333.505612	0.001218
Ukraine	UKR	4.65	14219.790039	0.000327

TABLE 13: INTERNET PRICES AND GDP PER CAPITA, PPP MERGED DATA AND RATIO CALCULATIONS BY AUTHOR, SOURCE: NUMBEO.COM AND WORLDBANK.ORG

ratio: visualizing data

Matplotlib.pyplot library was used to create a bar chart visualizing the ratio between the GDP per capita, PPP, and internet prices for 10 countries in the form of a horizontal bar chart.

In [106…

```
import matplotlib.pyplot as plt
ratio_df = pd.DataFrame(price_vs_gdp_ppp, columns=["Ratio"]) # get ratio dat
ratio_df.plot.barh() # plot the dataframe
plt.title("Ratio", fontsize=12 ) # title of the plot
plt.show() # display bar graph
```

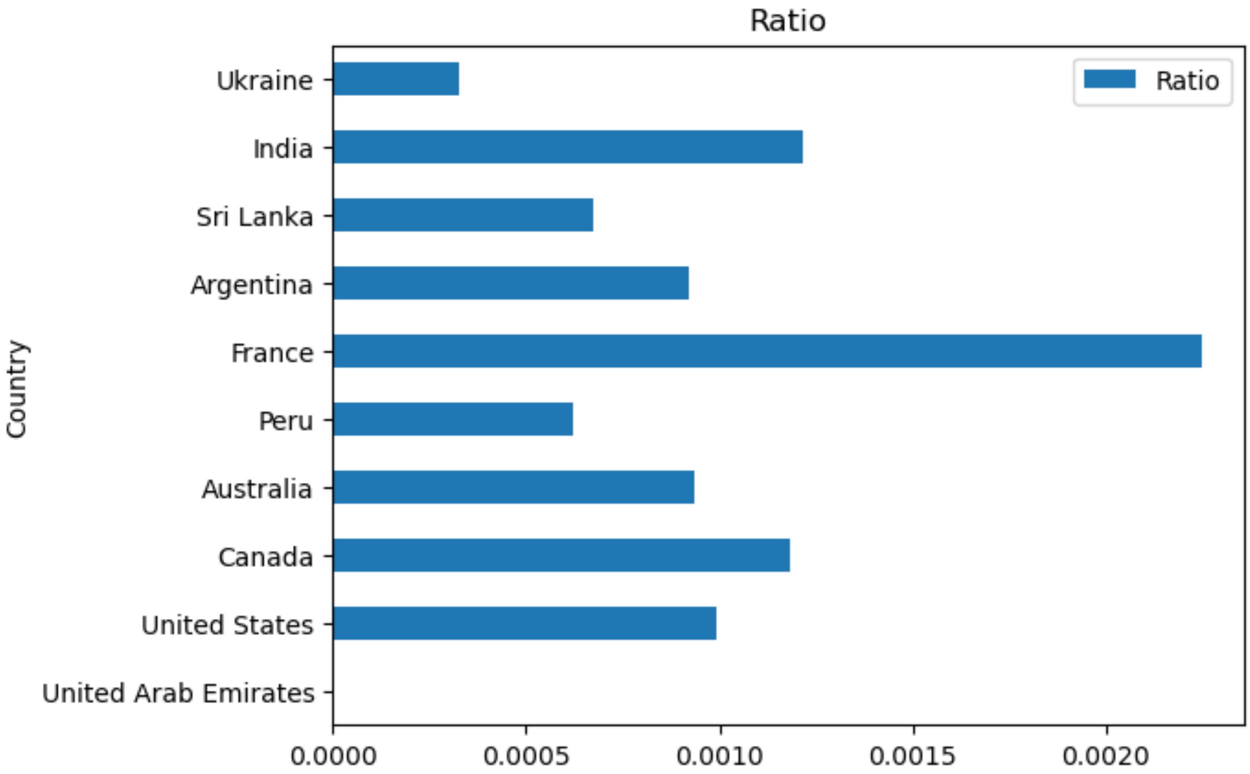


FIGURE 7: INTERNET PRICES VS GDP PER CAPITA, PPP RATIO, SOURCE: NUMBEO.COM, WORLDBANK.ORG AND AUTHOR'S CALCULATIONS

CONCLUSION

Figure 7 shows the disparity among countries. For example, Ukraine and Sri Lanka have similar GDP per capita, PPP, but the internet price in Sri Lanka is doubled compared to Ukraine. The same is happening between France and Canada. Peru is the most interesting of all. **Figure 7** shows the biggest disparity in the sample. GDP per capita, PPP in Peru is less than USD14,000 and internet price is USD33.10. Is the internet price influenced by GDP per capita, PPP? It seems that this conclusion is only applicable to Ukraine out of all countries, meaning high GDP per capita, PPP translates to lower internet prices. Perhaps that is not the right conclusion on a global scale. Another interesting aspect of this analysis is that the World Bank classified Peru as an upper-middle-income country (UMC) with an income below 14K and Ukraine was classified as a low-middle-income country (LMC) with an income above 14K. Shouldn't Ukraine be considered as UMC or Peru as LMC? An error in classification is highly unlikely considering the World Bank is a reputable financial institution, perhaps, another set of parameters is used to form an indicator. This is out of the scope of this research.

Back to Table of Contents

Income

What is the average income worldwide?

This section will look at the average monthly net wages for countries worldwide. The report will refer to “average monthly net wages” in a shortened form “wages”. Then it will narrow down the focus to 10 countries and will calculate a ratio between internet prices and wages.

The data integrity was validated based on the following reasoning.

- Kaggle.com is a user-published dataset platform, which has access to data with various licenses. The selected dataset has CC BY-NC-SA 4.0 license, which allows sharing and adaptation for non-commercial purposes with proper attribution. The dataset downloaded from Kaggle.com specified that it was aggregated from World Bank and Numbeo.com. From previous sections, it was established that the World Bank dataset consists of governmentally aggregated data and the Numbeo.com dataset consists of crowd-sourced data, the most up-to-date data, therefore, it’s unnecessary to use any additional resources for the validation of data for correctness. The Kaggle.com dataset includes over 80 countries from wealthy and poor countries, which fits the objective of the project. An import from CSV file to Jupyter Notebooks technique was used to retrieve the data after the dataset was downloaded from Kaggle.com. The integrity of the data will be analyzed by evaluating 5 years of wages.

Techniques

- import dataset from CSV file into Jupyter Notebooks
- perform data visualization with the pandas library
- perform data visualization with the mathplot library
- merge two datasets

Data Source

Average wages after tax

Kaggle.com (MKHURANA000, Internet Prices Datasets for Analysis, n.d.)

Dataset name: Average wages after tax (see data explorer section)

Data source: WorldBank.org and Numbeo.com

License: CC BY-NC-SA 4.0

Steps: Income

CHALLENGES

The dataset downloaded from Kaggle.com did not include World Bank indicators for various income levels. Therefore, the author had to filter the data to put selected countries into the categories assigned by the World Bank (LMC, UMC, HIC) to better understand the dataset. After reviewing the first sample, the author discovered discrepancies in the World Bank's assigned income categories. For example, Argentina and Peru are considered to be upper-middle-income countries, and Ukraine and India are considered to be low-middle-income countries, yet wages in Ukraine and India are higher than in Argentina and Peru. The only reasonable explanation for this is that Argentina and Peru have higher income taxes than Ukraine and India, or the World Bank did not classify countries properly, which is highly unlikely considering it’s a reputable financial institution.

Example:

- Sri Lanka $258.91
- Argentina $390.07
- Peru $469.35
- Ukraine $524.60
- India 579.72

WAGES FOR 2017-2021

Pandas library was used to read the file, display statistics, locate the range of wages, reset the index, select a 5-year period and start the index from 1 for 88 countries from 2013 to 2021. The data was presented for a 5-year period to understand if the 2021 column has accurate values. The data were further sorted into three categories: low-income, mid-income, and high-income countries to better visualize the disparities in the bar charts created with Matplotlib.pyplot library.

In [107…

```
# import libraries
import pandas as pd
import matplotlib.pyplot as plt
```

In [108…

```
df_kaggle_average_wages = pd.read_csv("./data/kaggle_data/kaggle_average_aft
```

In [109…

```
df_kaggle_average_wages.describe()
```

Out[109]:

	2021	2020	2019	2018	2017
count	87.000000	87.000000	87.000000	87.000000	87.000000
mean	1578.715402	1400.959655	1358.948046	1343.145402	1324.202184
std	1406.426012	1224.061436	1133.412642	1118.679599	1111.119190
min	231.260000	191.820000	207.760000	175.520000	155.520000
25%	482.640000	448.055000	472.655000	450.950000	448.260000
50%	1003.850000	907.000000	914.440000	925.680000	896.860000
75%	2873.185000	2445.165000	2340.085000	2294.830000	2335.260000
max	7023.070000	6276.630000	5613.730000	5423.690000	5381.410000

TABLE 14: STATISTICS FOR COUNTRIES WORLDWIDE, SOURCE: KAGGLE.COM (SEE DATA EXPLORER SIDE BAR)

findings

Table 14 shows that the highest wage is USD7,023 and the lowest wage is USD231.26, a significant difference between the maximum and minimum.

DEMONSTRATE REASONING AND CRITERIA FOR DATA SPLIT INTO SUB-CATEGORIES

The dataset contains 87 countries, although it is unclear whether countries fall within which income bracket. To simplify the analysis, it was decided to categorize countries based on the following criteria.

- low-income countries (0>1450)
- mid-income countries (1450>2900)
- high-income countries (2900>)

The above brackets don't represent an official threshold for income classification from the World Bank. The World Bank classification includes an additional income classification and different thresholds, which don't correspond to the income categories reviewed earlier (see Challenges).

Therefore, the classifications used in this section are only for the purpose of making sense of the available dataset and should not be used for any other reasons.

2021 The World Bank Income Classification (METREAU, 2021)

- Low income < 1,046
- Lower-middle income 1,046-4,095
- Upper-middle income 4,096 - 12,695
- High income > 12,695

Back to Table of Contents

LOW-INCOME COUNTRIES (0>1450)

This section will present **Figure 10** with wages below USD1,450. The low-income category has 50 countries with the lowest wage being in Nepal at USD252.19 and the highest wage being in Bahrain at USD1,361.76 in 2021. This section will visualize wages for 5 years, from 2017 to 2021. This period was taken to see how low-income countries responded to the COVID-19 lockdown period. It was particularly interesting to see the sample for 3 years prior to COVID-19 to get a baseline, 2020 during the shutdown, and all the way through 2021 to see how each country bounced back.

Note: Other alternatives were considered and tested to present this graph in a more readable format; however, these alternatives did not provide a complete picture of the low-income countries. Therefore, it was decided to keep it unchanged.

In [110…

```
# # uncomment for validation of amounts (done to reduce number of pages)
# low_income_df
```

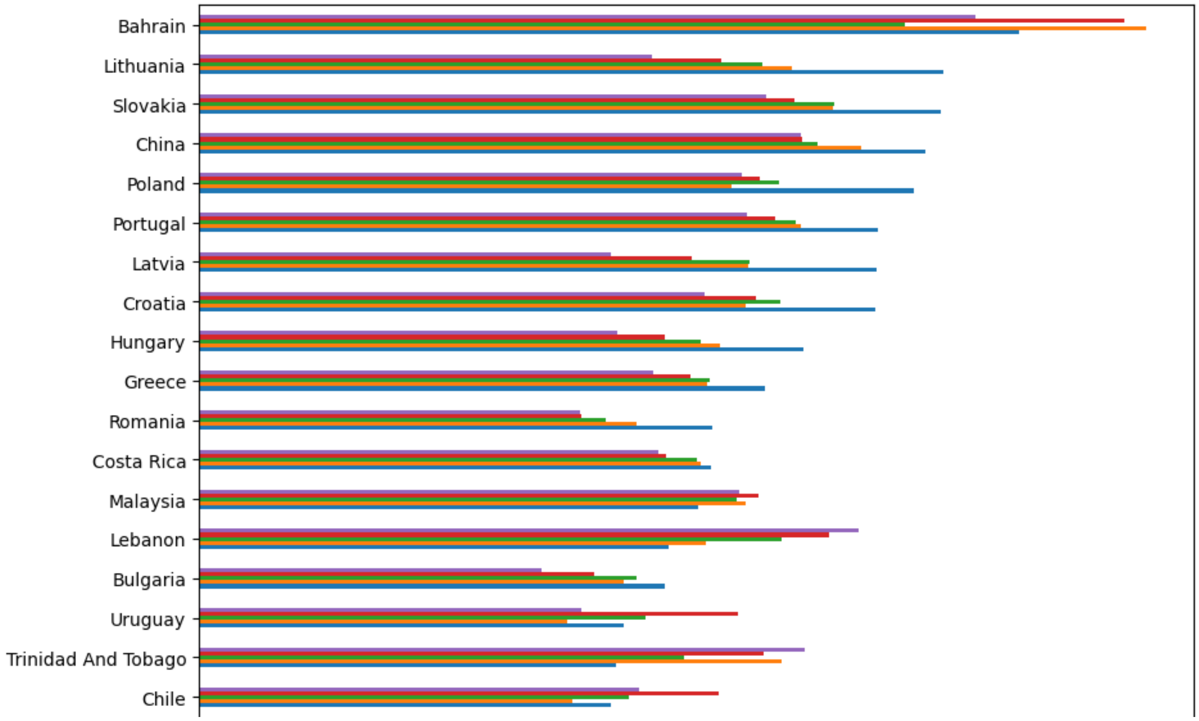
In [111…

```
low_income = df_kaggle_average_wages.loc[(df_kaggle_average_wages['2021'] >2
low_income.reset_index(inplace=True) # reset index
low_income_df = low_income[['Country', '2021', '2020', '2019', '2018', '2017
low_income_df.index =low_income_df.index+1 # set index to start from 1

low_income_df.plot(kind = 'barh', x = 'Country', y = ['2021','2020','2019',
plt.show # display plot
```

Out[111]:

```
<function matplotlib.pyplot.show(close=None, block=None)>
```



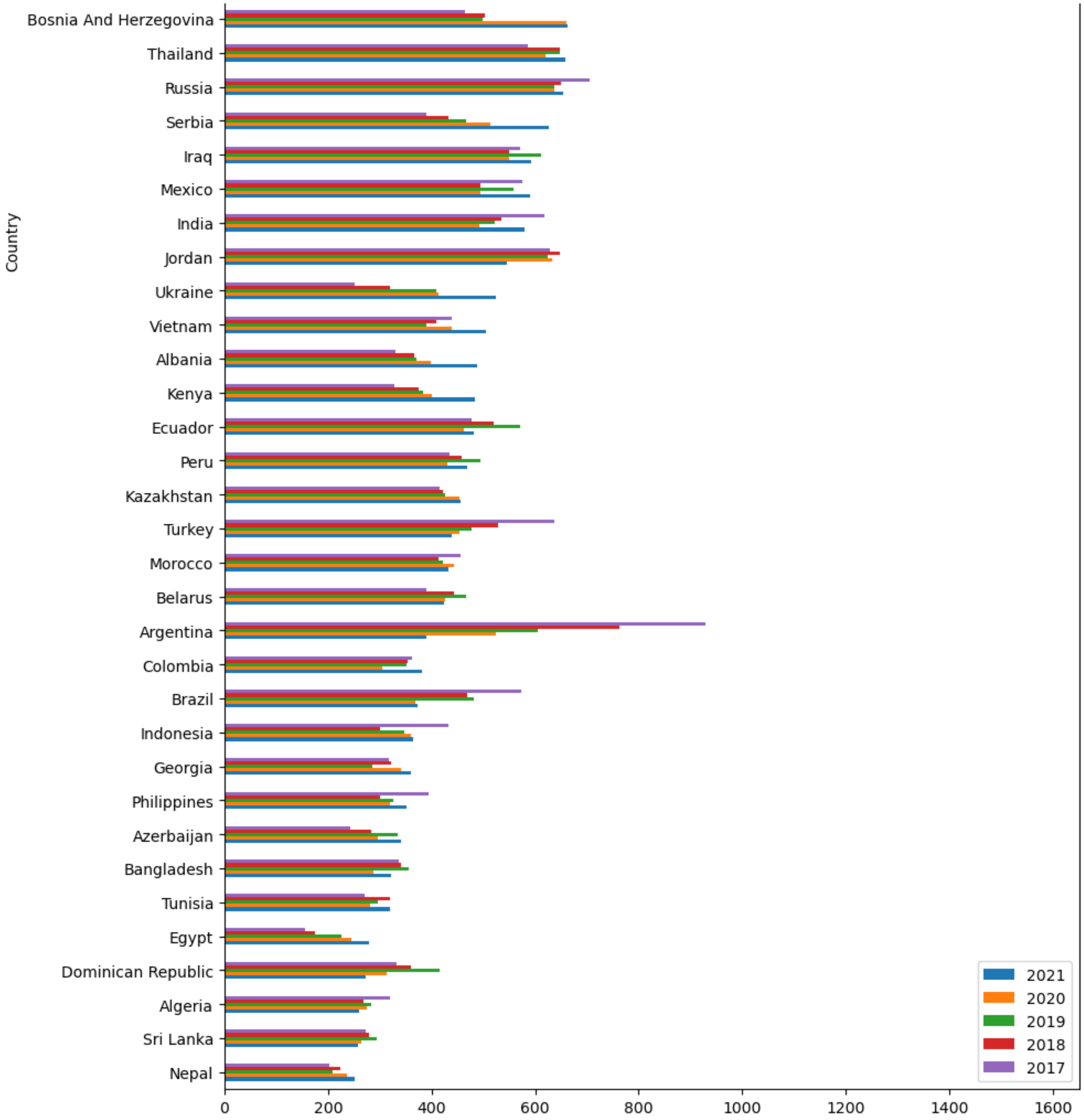


FIGURE 10: WAGES FOR LOW-INCOME COUNTRIES 2017 - 2021, SOURCE: KAGGLE.COM (SEE DATA EXPLORER SIDE BAR)

Back to Table of Contents

MID-INCOME COUNTRIES (1450>2900)

This section will present **Figure 11** with wages above USD1,450 and below USD2,900. The mid-income category has 13 countries with the lowest wage being in Cyprus at USD1,514.82 and the highest wages being in Belgium at USD2,827.45 in 2021. This section will visualize wages for 5 years, from 2017 to 2021. This period was taken to see how middle-income countries responded to the COVID-19 lockdown period. It was particularly interesting to see the sample for 3 years prior to COVID-19 to get a baseline, 2020 during the shutdown, and all the way through 2021 to see how each country bounced back.

In [112…

```
# # uncomment for validation of amounts (done to reduce number of pages)
# mid_income_df
```

In [113…

```
mid_income = df_kaggle_average_wages.loc[(df_kaggle_average_wages['2021'] >=
mid_income.reset_index(inplace=True) # reset index
mid_income_df = mid_income[['Country', '2021', '2020', '2019', '2018', '2017
mid_income_df.index =mid_income_df.index+1 # set index to start from 1

mid_income_df.plot(kind = 'bar', x = 'Country', y = ['2021','2020','2019', '
plt.show # display plot
```

Out[113]:

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

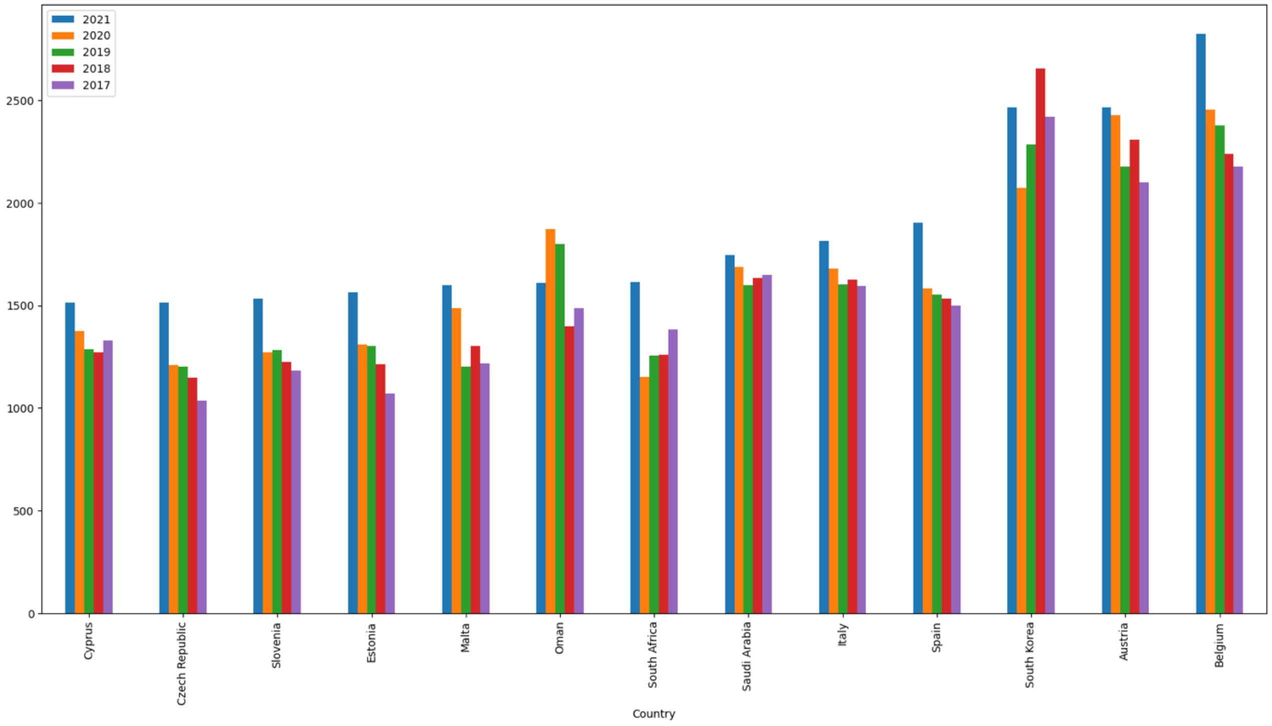


FIGURE 11: WAGES FOR MID-INCOME COUNTRIES 2017 - 2021, SOURCE: KAGGLE.COM (SEE DATA EXPLORER SIDE BAR)

Back to Table of Contents

HIGH-INCOME COUNTRIES (2900<)

This section will present **Figure 12** with wages above USD2,900. The high-income category has 22 countries with the lowest wage being in France at USD2,918.92 and the highest wages being in Switzerland at USD7,023.07 in 2021. This section will visualize wages for 5 years, from 2017 to 2021. This period was taken to see how high-income countries responded to the COVID-19 lockdown period. It was particularly interesting to see the sample for 3 years prior to COVID-19 to get a baseline, 2020 during the shutdown, and all the way through 2021 to see how each country bounced back.

In [114…

```
# # uncomment for validation of amounts (done to reduce number of pages)
# high_income_df
```

In [115…

```
high_income = df_kaggle_average_wages.loc[(df_kaggle_average_wages['2021'] >
high_income.reset_index(inplace=True) # reset index
high_income_df = high_income[['Country', '2021', '2020', '2019', '2018', '20
high_income_df.index = high_income_df.index+1 # set index to start from 1

high_income_df.plot(kind = 'bar', x = 'Country', y = ['2021','2020','2019',
plt.show # display plot
```

Out[115]:

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

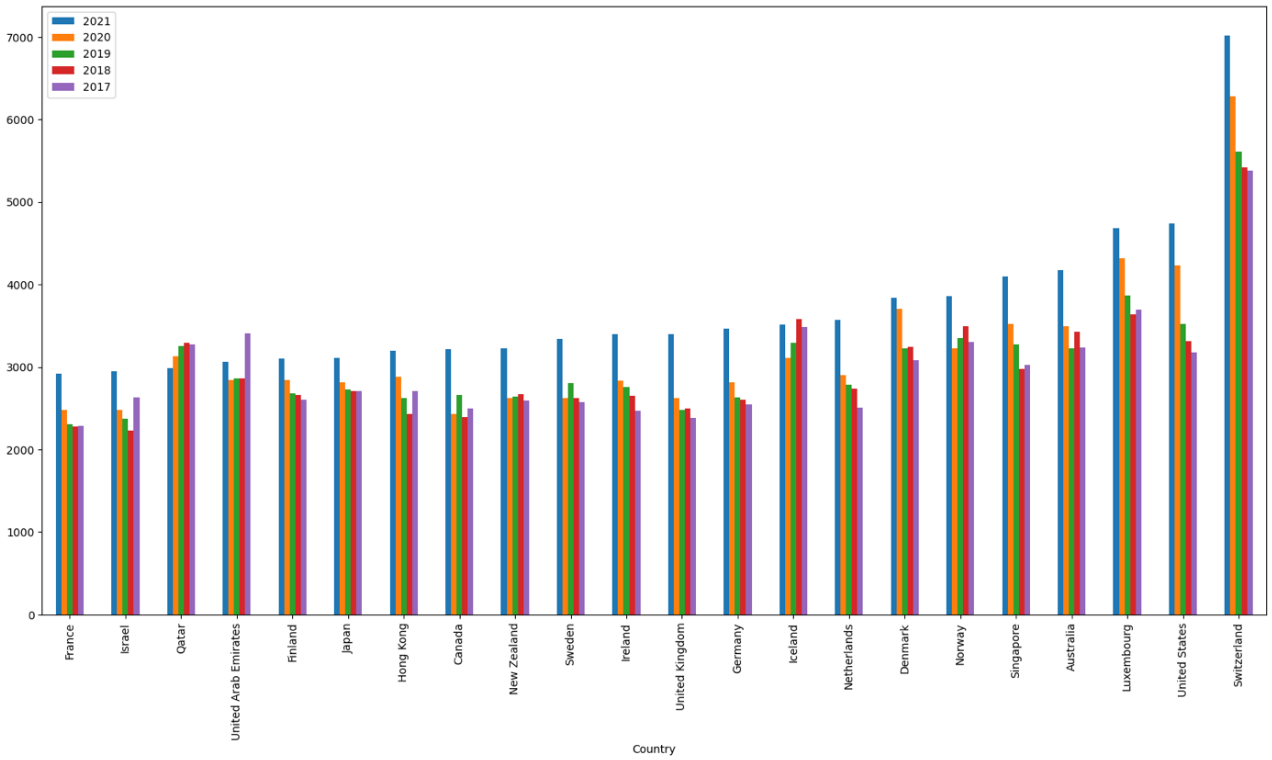


FIGURE 12: WAGES FOR HIGH-INCOME COUNTRIES 2017 - 2021, SOURCE: KAGGLE.COM (SEE DATA EXPLORER SIDE BAR)

filtering data for 10 countries

Panadas library was used to find 10 countries in the dataset (Australia, Canada, France, United States, United Arab Emirates, Argentina, Peru, India, Sri Lanka and Ukraine) in preparation to save the findings as a new CSV file. A new column was inserted and country codes were added to the dataset. The code is manually assigning values to each row in the Country Code column without checking for correctness in the event that the file is not filtered in the same order. Due to timing constraints, the code was not refactored. This could be done in the future.

In [116…

```
wages = pd.read_csv("./data/kaggle_data/kaggle_average_after_tax_wages.csv")
# # assign Country column as index
wages.set_index("Country", inplace=True)
lookup_wages_10 = wages.loc[wages.index.str.contains('Australia|Canada|Franc
kaggle_wages_10_2021 = lookup_wages_10[["2021"]] # look up 2021 column
kaggle_wages_10_2021.insert(0, "Country Code", ['LKA', 'ARG', 'PER', 'UKR',
kaggle_wages_10 = kaggle_wages_10_2021.copy # make a copy to work with datas
kaggle_wages_10=pd.DataFrame(kaggle_wages_10_2021)
kaggle_wages_10.rename(columns = {'2021':'Wages $ (2021)'}, inplace = True)
save_kaggle_wages_10=kaggle_wages_10
save_kaggle_wages_10
```

Out[116]:

	Country Code	Wages $ (2021)
Country		
Sri Lanka	LKA	258.91
Argentina	ARG	390.07
Peru	PER	469.35
Ukraine	UKR	524.60
India	IND	579.72
France	FRA	2918.92
United Arab Emirates	ARE	3065.62
Canada	CAN	3212.58
Australia	AUS	4171.74
United States	USA	4734.67

TABLE 15: WAGES FOR 2021, SOURCE: KAGGLE.COM (SEE DATA EXPLORER SIDE BAR)

saving data for 10 countries

The dataset for 10 countries will be saved for further analysis to compare internet prices against wages. Pathlib library was used to check if the file already exists, then print out a message, if the file does not exist then create the file and print the message saying that.

In [117…

```
# check if the file exists, if does not than save the file
# reference: https://www.pythontutorial.net/python-basics/python-check-if-fi
from pathlib import Path
path_to_file = './data/10_selected_countries/kaggle_wages_10.csv'
path = Path(path_to_file)

if path.is_file():
    print(f'The file {path_to_file} exists')
else:
    save_kaggle_wages_10.to_csv("./data/10_selected_countries/kaggle_wages_1
    print(f'The file {path_to_file} does not exist. The file wil be saved to
```

```
The file ./data/10_selected_countries/kaggle_wages_10.csv exists
```

Back to Table of Contents

MERGING DATA: INTERNET PRICES AND WAGES

Pandas library was used to read and merge two datasets, a duplicate column was removed, then the index was set to Country, and columns were renamed. The library was also used to insert a column and calculate the ratio between internet prices and wages, saved under the Ratio name.

In [118…

```
# prepare for merging datasets
numbeo_internet_price_10 = pd.read_csv("./data/10_selected_countries/numbeo_
kaggle_wages_10 = pd.read_csv("./data/10_selected_countries/kaggle_wages_10.

# merge two datasets
price_vs_wages = pd.merge(numbeo_internet_price_10, kaggle_wages_10, on='Cou
price_vs_wages.drop('Country_y', inplace=True, axis=1) # remove duplicate co
price_vs_wages.rename(columns = {'2021':'Price $ (2021)','Country_x':'Countr
price_vs_wages.set_index("Country", inplace=True) # assign Country column as
price_vs_wages['Ratio'] = price_vs_wages['Price $ (2021)'] / price_vs_wages[
price_vs_wages
```

Out[118]:

	Country Code	Price $ (2021)	Wages $ (2021)	Ratio
Country				
United Arab Emirates	ARE	99.64	3065.62	0.032502
United States	USA	68.64	4734.67	0.014497
Canada	CAN	61.70	3212.58	0.019206
Australia	AUS	52.12	4171.74	0.012494
Peru	FRA	31.53	2918.92	0.010802
France	PER	31.20	469.35	0.066475
Argentina	ARG	21.77	390.07	0.055810
Sri Lanka	LKA	9.50	258.91	0.036692
India	IND	8.93	579.72	0.015404
Ukraine	UKR	4.65	524.60	0.008864

TABLE 16: INTERNET PRICES, WAGES, AND AUTHOR'S RATIO CALCULATIONS FOR 2021, SOURCE: NUMBEO.COM AND SOURCE: KAGGLE.COM (SEE DATA EXPLORER SIDE BAR)

findings

Australia and the United States are high-income countries and have very close wages and internet prices. Canada and the United Arab Emirates are also high-income countries but have a very big difference in internet prices. Peru, Argentina, and Ukraine have similar trends.

ratio: visualizing data

Matplotlib.pyplot library was used to create a bar chart visualizing the ratio between 10 countries.

In [119…

```
import matplotlib.pyplot as plt
price_vs_wages = pd.DataFrame(price_vs_wages, columns=["Ratio"]) # get data
price_vs_wages.plot.barh() # plot the dataframe
plt.title("Ratio", fontsize=12 ) # title of the plot
plt.show() # display bar graph
```

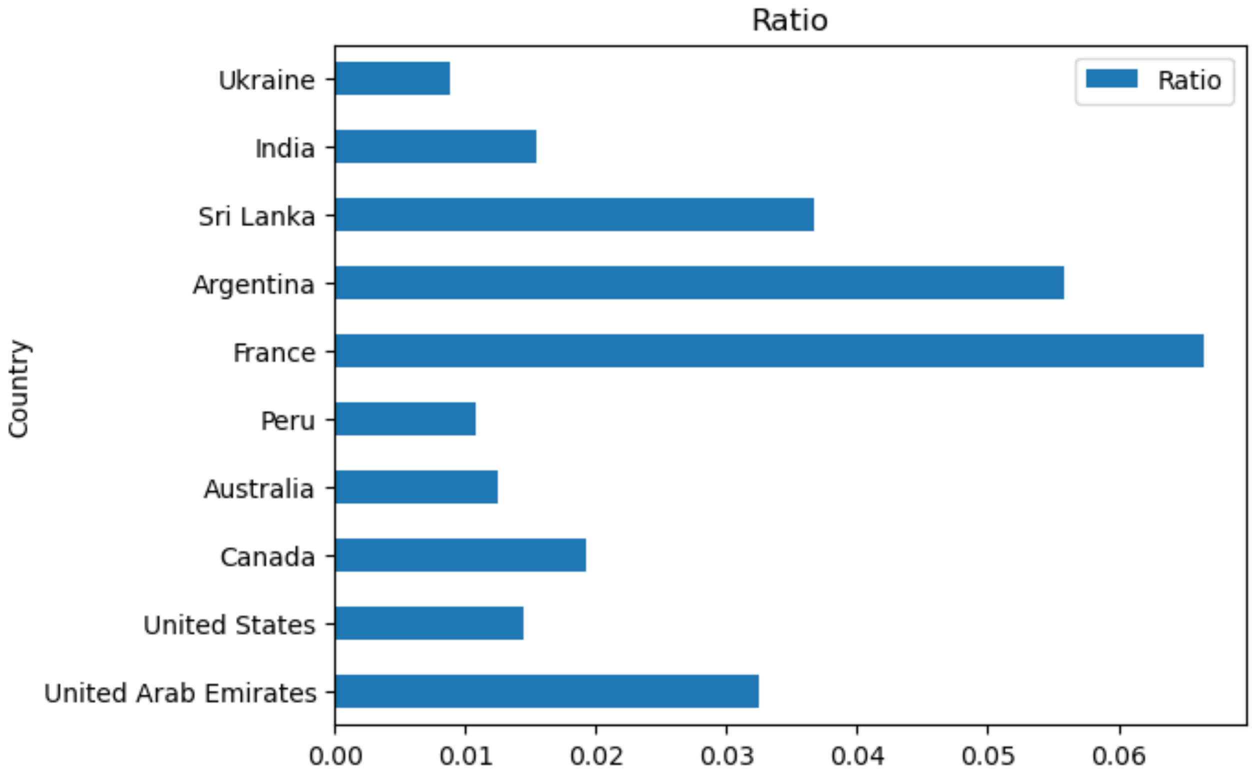


FIGURE 13: AUTHOR'S CALCULATIONS OF RATIO, SOURCE: NUMBEO.COM AND SOURCE: KAGGLE.COM (SEE DATA EXPLORER SIDE BAR)

CONCLUSION

Figure 13 shows that internet prices vs wages are not equally distributed. Peru has a very high internet price compared to wages and France has a very low internet price compared to wages. Based on the selected sample of 10 countries, it’s difficult to be certain if internet prices are affected by wages. Perhaps, a correlation can be drawn for some countries, but not others.

Back to Table of Contents

DATA DISCREPANCY

The analysis of the wages between 2017 and 2021 shows a surprising discrepancy, wages skyrocketed in 2021 after the COVID-19 economic standstill. The author had a theory that the wages will be lower during the economic reboot as the employment pool should be bigger after massive layoffs during the pandemic. Further investigation into this phenomenon should be done, but due to time limitations in preparation for this project, Forbes.com was reviewed to understand if the data is compromised or if there is a true basis for this drastic jump. According to Forbes, the high unemployment rate creates an environment for companies to increase wages due to the smaller labor market, which makes it harder to find qualified employees. For example, the entertainment industry is experiencing a shortage of qualified workers, which is something that hasn't happened before. Furthermore, a lot of workers are not re-entering the labor market due to financial assistance from governments. It's suggested by Forbes.com that high demand and low availability create a need for employers to raise wages. Based on this, it will be considered reasonable to accept the dataset as correct data. (Levanon, 2021)

In [120…

```
import matplotlib # import library
matplotlib.style.use('ggplot')  # adjusts the style to simulate ggplot, whic
# display for 5 year-period
df_kaggle_average_wages.sort_values(by='Country', ascending=False).plot(kind
```

Out[120]:

```
<Axes: xlabel='Country'>
```

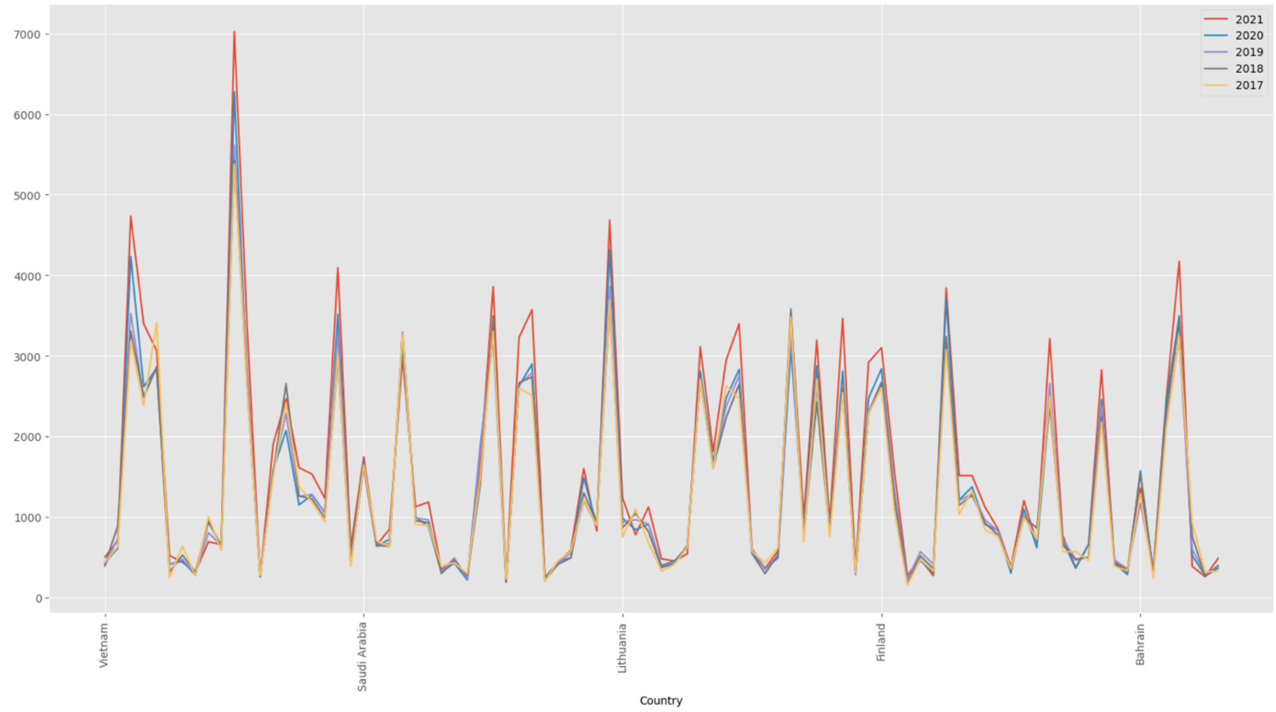


FIGURE 14: DISCREPANCIES OF WAGES 2017 - 2021, SOURCE: KAGGLE.COM (SEE DATA EXPLORER SIDE BAR)

Back to Table of Contents

Ethics of Data Use

Ethics of data use played a major role in the resources selected to collect data for the purpose of the research. The report contains data from Numbeo.com, Kaggle.com, and World Bank for the following reasons.

Numbeo.com

Numbeo.com, a crowd-sourced global database, allows the free use of data for personal and academic purposes with appropriate credit. Although Numbeo's reputation is well-regarded and used by some academics, its data is questionable due to the design of the website. According to its general disclaimer, the website is designed to permit anyone to make changes and no information is validated by experts. Furthermore, the website doesn't take responsibility in case of damage or loss caused by the use of its website. (Terms of Use - Numbeo.com, n.d.) In spite of thereof, the dataset was used in the report as it contained the most up-to-date information on internet prices and the author relied on data integrity as it's trusted by academics such as Viktor Grechyn and Ian McShane from Centre for Urban Research, RMIT University.

A web scraping technique was used to collect the data from Numbeo.com on internet prices. In the author's opinion, this technique was appropriate as the data scraped from the website did not include personal identifiers and the website did not explicitly prohibit web scraping.

Kaggle.com

Kaggle.com, a user-published dataset platform, provides access to data with various licenses. Datasets used in the report have CC BY-NC-SA 4.0 and CC BY 4.0 licenses, which allow sharing and adaptation for non-commercial purposes with proper attribution. The CC BY 4.0 license specifically allows the copying, sharing, remixing, and transformation of the data for any purpose with proper attribution and indication if any changes have been made. Kaggle.com is a popular platform for data science projects. Although it is a subsidiary of Google and described by Wikipedia as "...an online community of data scientists and machine learning practitioners..." (Wikipedia, Kaggle, n.d.), the reliability and correctness of the data are questionable. In the author's opinion, the warranty disclaimer by Kaggle.com raises concerns about data validity as the website states that the company is not responsible for "...the accuracy, copyright compliance, legality..." (Terms, n.d.) of its contents.

Kaggle.com explicitly prohibits the use of web scraping on its website, therefore, the dataset was downloaded after the user account was set up and later imported to Jupyter Notebooks in the CSV format.

World Bank

The World Bank is a financial institution that provides financial services to governments. It provides access to data, collected from 189 countries it cooperates. (Wikipedia, World Bank, n.d.) The selected datasets have the CC BY 4.0 license, which allows the free use, copying, sharing, remixing, and transformation of the data for any purpose with proper attribution and indication if any changes have been made. The repository specific to the analysis contains data from International Comparison Program, World Development Indicators database, Eurostat-OECD PPP Program, World Bank national accounts data, and OECD National Accounts data files, all governmentally aggregated data. As per the terms of use, the author must disclose that the World Bank did not endorse the project or the data used in the project. (Summary Terms of Use, n.d.) The exclusion of liability section indicates that the World Bank "...shall not be responsible or liable for the accuracy, usefulness or availability of any data in the Datasets..." (Bank, Terms of Use for Datasets, 2018), which raises concerns about the data validity. The World Bank provides access to its data through downloads in various formats or APIs. The author chose to obtain data through a download and APIs to showcase different skills for the project.

Potential Biases

The data used in the report does not contain personal identifiers and the country-specific datasets do not appear to target or discriminate against any particular group of people.

Dataset Modifications

Parts of some datasets were reformatted to be combined with other datasets to perform calculations, described in the report as “author calculations”. The substance of the data was not altered. Nonetheless, the data was manipulated for presentation in an easy-to-understand format such as tables and figures.

Legal Considerations

The terms of use pages from Numbeo.com, Kaggle.com, and WorldBank.org are alarming, but each appears to send a consistent message about companies' legal exposure. The author is not an attorney or legal professional and cannot properly validate the following statement, but it appears that the Terms of Use display standard legal disclaimers to limit companies’ exposure to lawsuits. The data in the report was used with an assumption that thereof, is true and data is reasonably correct and valid.

The data and the author’s conclusions should only be used for a general understanding of the topic. The author is not an expert in the field, and therefore, cannot be personally liable for the correctness of conclusions made in the report. Nonetheless, the information presented in the report should not be reused or used to make any type of decision.

Intellectual Property

It is doubtful that the findings can have the potential to create intellectual property due to the author’s lack of expertise, however, the author reserves the right to make this report their intellectual property now or at any time in the future.

Back to Table of Contents

Final Remarks

The report proposed that 3 factors influence internet prices worldwide: internet usage rate, GDP per capita, PPP, and wages. In spite of a very lengthy analysis, it is unclear if each proposed factor is applicable to every country worldwide. It would be reasonable to suggest that an in-depth study should be done to understand if the proposed factors are in fact what drives prices up or down. Perhaps, there are other factors that influence the prices of the internet worldwide such as the competitiveness of the internet industry, and monopolization by one or a few firms, given the huge investments required. The author would like to further explore this topic in the future and create an indicator, similar to GDP, to better understand where each country lies within internet accessibility (price, internet usage rate, topography, access to education, and demographics).

It is essential to note that the majority of datasets utilized a single snapshot in time. If ten years of data were used for each analysis, it may have been possible to detect a trend and better explain internet prices. This can be done in the future.

And finally, Ukraine's position in the world is clear, it has the cheapest internet. The report allows drawing the following conclusion that low wages, high internet usage rate, and low GDP per capita, PPP do influence the low cost of internet in the country.

Personal Note

The report ended up being incredibly time-consuming for two reasons: i) a lack of Data Science experience and ii) a lack of familiarity with the chosen topic. Due to time limits, it was not possible to complete the report by adding links to all associated citations and tables. Thereof, future improvements are possible. Nonetheless, the Bibliography section that follows covers the sources used in this research.

All citations were saved in a Microsoft Word file and then transferred to Jupyter Notebooks as-is.

Final Challenges

1. The pip freeze command created a file with a strange format. Therefore, two versions of the requirements were saved for the project.

- pip list --format=freeze > requirements.txt (adamgy, 2020)
- pip freeze > requirements_unformated.txt

1. Images and interactive maps were not displaying in "PDF via LaTeX (.pdf)" conversion. The issue was troubleshooted for several hours without success. Consequently, the image (with reference to (METREAU, 2021)) was recreated as text. Unfortunately, no workarounds for interactive maps could be identified. Please review the Python file to display internet prices on the map and visualizing internet usage on the map.

Back to Table of Contents

Bibliography

adamgy. (2020, July 14). pip freeze creates some weird path instead of the package version. Retrieved December 2022, from stackoverflow.com: https://stackoverflow.com/questions/62885911/pip-freeze-creates-some-weird-path-instead-of-the-package-version

Attribution 4.0 International (CC BY 4.0). (n.d.). Retrieved November 2022, from creativecommons.org: https://creativecommons.org/licenses/by/4.0/

Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0). (n.d.). Retrieved November 2022, from creativecommons.org: https://creativecommons.org/licenses/by-nc-sa/4.0/

Bank, T. W. (2018, March 23). Terms of Use for Datasets. Retrieved November 2022, from worldbank.org: https://www.worldbank.org/en/about/legal/terms-of-use-for-datasets

Bank, T. W. (n.d.). GDP per capita, PPP (current international $). Retrieved November 2022, from worldbank.org: https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD

Bank, T. W. (n.d.). Individuals using the Internet (% of population). Retrieved November 2022, from worldbank.org: https://data.worldbank.org/indicator/IT.NET.USER.ZS

Bank, T. W. (n.d.). Individuals using the Internet (% of population). Retrieved November 2022, from worldbank.org: https://data.worldbank.org/indicator/IT.NET.USER.ZS

Bank, T. W. (n.d.). Metadata Glossary. Retrieved November 2022, from worldbank.org: https://databank.worldbank.org/metadataglossary/statistical-capacity-indicators/series/5.51.01.10.gdp

Bank, T. W. (n.d.). Metadata Glossary. Retrieved November 2022, from worldbank.org: https://databank.worldbank.org/metadataglossary/world-development-indicators/series/NY.GDP.PCAP.PP.KD

datatofish.com/. (2021, July 3). How to Convert Strings to Floats in Pandas DataFrame. Retrieved November 2022, from datatofish.com/: https://datatofish.com/convert-string-to-float-dataframe/

Definition: Penn effect. (n.d.). Retrieved from tariffnumber.com: https://www.tariffnumber.com/info/abbreviations/16227

educba.com. (n.d.). BeautifulSoup Table. Retrieved November 2022, from educba.com: https://www.educba.com/beautifulsoup-table/

Eurostat, O. f.-O. (2013-2019). Definition: Penn effect. Retrieved November 2022, from tariffnumber.com: https://www.tariffnumber.com/info/abbreviations/16227

geeksforgeeks.org. (2020, December 4). numpy.round*() in Python. Retrieved November 2022, from geeksforgeeks.org: https://www.geeksforgeeks.org/numpy-round*-python/

Heekyung Hellen Kim, J. Y. (2004). Broadband penetration and participatory politics: South Korea case. Proceedings of the 37th Hawaii International Conference on System Sciences - 2004, 1-10.

javatpoint.com. (n.d.). Wordcloud Package in Python. Retrieved November 2022, from javatpoint.com: https://www.javatpoint.com/wordcloud-package-in-python

Joan Calzada, F. M.-S. (2014). Broadband prices in the European Union: Competition and commercial strategies. Information Economics and Policy, Volume 27, 24-38.

Jones, A. (2021, August 13). Discover a World of Data with WBGAPI. Retrieved November 2022, from towardsdatascience.com: https://towardsdatascience.com/access-a-world-of-data-with-wbgapi-61849354f769

jupyterbook.org. (n.d.). Images and figures. Retrieved November 2022, from jupyterbook.org: https://jupyterbook.org/en/stable/content/figures.html

Levanon, G. (2021, July 26). Why Wages Are Growing Rapidly—Both Now And In The Future. Retrieved November 2022, from forbes.com: https://www.forbes.com/sites/gadlevanon/2021/07/26/why-wages-are-growing-rapidly-both-now-and-in-the-future/?sh=27ceec12cfe9

Luvsandorj, Z. (2020, June 20). Simple word cloud in Python. Retrieved November 2022, from towardsdatascience.com: https://towardsdatascience.com/simple-wordcloud-in-python-2ae54a9f58e5

matplotlib.org. (n.d.). ggplot style sheet. Retrieved November 2022, from matplotlib.org: https://matplotlib.org/stable/gallery/style_sheets/ggplot.html

matplotlib.org. (n.d.). Pyplot tutorial. Retrieved November 2022, from matplotlib.org: https://matplotlib.org/2.0.2/users/pyplot_tutorial.html

METREAU, N. H. (2021, July 1). New World Bank country classifications by income level: 2021-2022. Retrieved November 2022, from https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2021-2022

MKHURANA000. (n.d.). Internet Prices Datasets for Analysis. Retrieved November 2022, from kaggle.com: https://www.kaggle.com/datasets/mkhurana000/internet-prices-datasets-for-analysis?select=average_after_tax_wages.csv

MKHURANA000. (n.d.). Internet Prices Datasets for Analysis/Average after-tax wages. Retrieved November 2022, from kaggle.com: https://www.kaggle.com/datasets/mkhurana000/internet-prices-datasets-for-analysis?select=average_after_tax_wages.csv

Numbeo. (n.d.). Retrieved from wikipedia.org: https://en.wikipedia.org/wiki/Numbeo

numbeo.com. (n.d.). Retrieved from Price Rankings by Country of Internet (60 Mbps or More, Unlimited Data, Cable/ADSL) (Utilities (Monthly)): https://www.numbeo.com/cost-of-living/country_price_rankings?itemId=33

pandas.pydata.org. (n.d.). How do I create plots in pandas? Retrieved November 2022, from pandas.pydata.org: https://pandas.pydata.org/docs/getting_started/intro_tutorials/04_plotting.html

pandas.pydata.org. (n.d.). Intro to data structures. Retrieved November 2022, from pandas.pydata.org: https://pandas.pydata.org/docs/user_guide/dsintro.html

pandas.pydata.org. (n.d.). pandas.DataFrame. Retrieved November 2022, from pandas.pydata.org: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

pandas.pydata.org. (n.d.). pandas.DataFrame.dropna. Retrieved November 2022, from pandas.pydata.org: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.dropna.html

pandas.pydata.org. (n.d.). pandas.DataFrame.dtypes. Retrieved November 2022, from pandas.pydata.org: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.dtypes.html

pandas.pydata.org. (n.d.). pandas.DataFrame.loc. Retrieved November 2022, from pandas.pydata.org: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.loc.html

pandas.pydata.org. (n.d.). pandas.DataFrame.rename. Retrieved November 2022, from pandas.pydata.org: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rename.html

pandas.pydata.org. (n.d.). pandas.DataFrame.set_index. Retrieved November 2022, from pandas.pydata.org: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.set_index.html

pandas.pydata.org. (n.d.). pandas.Index.names. Retrieved November 2022, from pandas.pydata.org: https://pandas.pydata.org/docs/reference/api/pandas.Index.names.html

pandas.pydata.org. (n.d.). pandas.notnull. Retrieved November 2022, from pandas.pydata.org: https://pandas.pydata.org/docs/reference/api/pandas.notnull.html

pandas.pydata.org. (n.d.). pandas.read_table. Retrieved November 2022, from pandas.pydata.org: https://pandas.pydata.org/docs/reference/api/pandas.read_table.html

projectpro.io. (2022, June 9). How to create a word cloud in Python? Retrieved November 2022, from projectpro.io: https://www.projectpro.io/recipes/create-word-cloud-python

pydata.org. (n.d.). How do I select a subset of a DataFrame? Retrieved November 2022, from pydata.org: https://pandas.pydata.org/docs/getting_started/intro_tutorials/03_subset_data.html pythontutorial.net. (n.d.). Python Check If File Exists. Retrieved November 2022, from pythontutorial.net: https://www.pythontutorial.net/python-basics/python-check-if-file-exists/

Reinhart, J. (2013, May 12). Correct way to try/except using Python requests module? Retrieved November 2022, from stackoverflow.com: https://stackoverflow.com/questions/16511337/correct-way-to-try-except-using-python-requests-module

requests.readthedocs.io. (n.d.). Errors and Exceptions. Retrieved November 2022, from requests.readthedocs.io: https://requests.readthedocs.io/en/latest/user/quickstart/#errors-and-exceptions

Summary Terms of Use. (n.d.). Retrieved November 2022, from worldbank.org: https://data.worldbank.org/summary-terms-of-use

Terms. (n.d.). Retrieved November 2022, from kaggle.com: https://www.kaggle.com/terms Terms of Use - Numbeo.com. (n.d.). Retrieved November 2022, from numbeo.com: https://www.numbeo.com/common/terms_of_use.jsp

tutorialspoint.com. (n.d.). How to extract floating number from text using Python regular expression? Retrieved November 2022, from tutorialspoint.com: https://www.tutorialspoint.com/How-to-extract-floating-number-from-text-using-Python-regular-expression

Viktor Grechyn, I. M. (2016, December). What Influences International Differences in Broadband Prices? Australian Journal of Telecommunications and the Digital Economy, 4, 89 - 105. Retrieved from https://telsoc.org/journal/ajtde-v4-n4/a67

Vu, D. (2019, November). Generating WordClouds in Python Tutorial. Retrieved November 2022, from datacamp.com: https://www.datacamp.com/tutorial/wordcloud-python

w3resource.com. (2022, August 19). Pandas DataFrame: plot.barh() function. Retrieved November 2022, from w3resource.com: https://www.w3resource.com/pandas/dataframe/dataframe-plot-barh.php

w3resource.com. (2022, August 19). Pandas DataFrame: tail() function. Retrieved November 2022, from w3resource.com: https://www.w3resource.com/pandas/dataframe/dataframe-tail.php

w3schools.com. (n.d.). Pandas DataFrame describe() Method. Retrieved November

2022, from w3schools.com: https://www.w3schools.com/python/pandas/ref_df_describe.asp

w3schools.com. (n.d.). Pandas DataFrame head() Method. Retrieved November 2022, from w3schools.com: https://www.w3schools.com/python/pandas/ref_df_head.asp

w3schools.com. (n.d.). Pandas DataFrame merge() Method. Retrieved November 2022, from w3schools.com: https://www.w3schools.com/python/pandas/ref_df_merge.asp

w3schools.com. (n.d.). Pandas Read CSV. Retrieved November 2022, from w3schools.com: https://www.w3schools.com/python/pandas/pandas_csv.asp

Wikipedia, t. f. (n.d.). Kaggle. Retrieved November 2022, from wikipedia.org: https://en.wikipedia.org/wiki/Kaggle

Wikipedia, t. f. (n.d.). World Bank. Retrieved November 2022, from wikipedia.org: https://en.wikipedia.org/wiki/World_Bank

Zaire, C. (2021, January 4). How to Web Scrape Tables Online, Using Python and BeautifulSoup. Retrieved November 2022, from medium.com: https://medium.com/analytics-vidhya/how-to-web-scrape-tables-online-using-python-and-beautifulsoup-36d5bafeb982

Back to Table of Contents