CarrierFoundry
Data Immerse Course
Exercise 6.1
Adriana García Márquez

DATA SOURCE AND COLLECTION

Topic: Internet Use Across EU Countries by Demographic Group

1. Data Summary

Dataset: Individuals – Frequency of Internet Use

Data Sourcing

The data was obtained from Eurostat, the official statistical office of the European Union. As a government source, it is considered highly reliable and trustworthy.

Data Collection

The dataset is based on the 'EU survey on the use of Information and Communication Technologies (ICT) in households and by individuals, conducted annually sine 2002. The survey aims to collect comparable information on ICT usage across EU Member States. The data collection is carried out by the National Statistical Institutes.

Potential bias may arise from survey-based data collection methods.

Data Contents

The dataset contains the annual frequency of internet usage across EU countries from 2003 to 2024, expressed as percentages and segmented by individual characteristics.

Frequency of Internet Use includes the following categories:

- *Daily*
- *At least once a week*
- *Less than once a week*
- *At least once a month*
- *Less than once a month*.

Individual Types represent different demographic groups including:

- *Age groups*
- *Gender*
- *Level of education*
- *Household income quartiles*
- *Area of residence (urban, suburban, rural)*
- *Employment status (e.g. Student, Employee, Self-employed)*
- *Job sector*
- *Family status (with or without children)*

CarrierFoundry
Data Immerse Course
Exercise 6.1
Adriana García Márquez

Data Relevance

This dataset is highly relevant to the country I live in (Germany), which plays a key role in shaping the European Union's digital transformation strategy. One of the EU's objectives is to empower citizens through access to digital technologies.

I chose this topic because it provides valuable insights into how different population groups, categorized by age, gender, income and employment status, engage with internet usage across EU countries. This information is essential for understanding patterns of digital inclusion and identifying disparities within and between countries.

The dataset offers annual data on the frequency of internet use, expressed as percentages across various demographic segments, enabling a comparative analysis.

Dataset: Individuals – Internet Activities

Data Sourcing

The data was obtained from Eurostat, the official statistical office of the European Union. As a government source, it is considered highly reliable and trustworthy.

Data Collection

The dataset is based on the 'EU survey on the use of Information and Communication Technologies (ICT) in households and by individuals, conducted annually sine 2002. The survey aims to collect comparable information on ICT usage across EU Member States. The data collection is carried out by the National Statistical Institutes.

Potential bias may arise from survey-based data collection methods.

Data Contents

The dataset contains the annual frequency of internet activities across EU countries from 2003 to 2024, expressed as percentages and segmented by individual characteristics.

Internet Activities includes a wide range of categories, grouped as followed:

- *Educational purposes*
- *Selling goods or services online*
- *Consuming goods or services online (e.g. shopping, streaming)*
- *Using social media for entertainment*
- *Using social network for professional purposes*
- *Online banking and digital payments*
- *Civic or political participation*
- *Personal use (e.g. scheduling appointments, accessing health services)*

Individual Types represent different demographic groups including:

- *Age groups*
- *Gender*
- *Level of education*
- *Household income quartiles*
- *Area of residence (urban, suburban, rural)*
- *Employment status (e.g. Student, Employee, Self-employed)*
- *Job sector*
- *Family status (with or without children)*

Data Relevance

This dataset is highly relevant to the country I live in (Germany), which plays a key role in shaping the European Union's digital transformation strategy. One of the EU's objectives is to empower citizens through access to digital technologies.

I chose this topic because it provides valuable insights into how different population groups, categorized by age, gender, income and employment status, engage with internet usage across EU countries. This information is essential for understanding patterns of digital inclusion and identifying disparities within and between countries.

The dataset offers annual data on the frequency of internet activities, expressed as percentages across various demographic segments, enabling a comparative analysis.

Dataset: Purchasing Power Adjusted GDP Per Capita

Data Sourcing

The data was obtained from Eurostat, the official statistical office of the European Union. As a government source, it is considered highly reliable and trustworthy.

Data Collection

The dataset is based on the economic data reported by EU Members to the European Statistical System (ESS).

Potential bias may arise due to differences in data reporting methods or collection methods among countries.

Data Contents

The dataset contains values for GDP per capita adjusted for purchasing power from 2003 to 2024, expressed in purchasing power standard (PPS). This metric represents the total output of goods and services produces by an economy, less intermediate consumption,

plus net taxes on products and imports. GDP per capita is calculated as the ratio of GDP to the average population in an specific year. Expressing GDP in PPS eliminates the differences in price levels between countries, allowing for meaningful comparison of the economy and living standards across countries.

Data Relevance

This dataset is relevant to the project as it enables cross-country comparisons of internet usage and activities across different demographic segments in relation to living standards within EU Member States. By incorporating the GDP per capita adjusted for purchasing power, the analysis can explore the correlations between digital behavior and economic development.

2. Reason to Choose These Datasets

I chose these datasets because I believe the topic is highly relevant to the country I live in, Germany. In terms of digitalization, Germany appears to be behind some other countries. One of the key objectives of the government, both at the national and EU level, is to promote the digital inclusion and ensure that internet access and digital technologies are available to everyone.

This theme is particularly interesting to me, as it allows me to analyze the differences in the digital behavior across EU countries and highlight any digital gaps that may exist between Germany and other countries.

3. Data Cleaning Summary

Dropping Unnecessary Columns

Frequency Internet Use	Internet Activities	Purchasing Power GDP
DATAFLOW	DATAFLOW	DATAFLOW
LAST UPDATE	LAST UPDATE	LAST UPDATE
freq	freq	freq
unit	unit	na_item
CONF_STATUS	CONF_STATUS	ppp_cat
		CONF_STATUS

Renaming Columns

Frequency Internet Use	Internet Activities	Purchasing Power GDP
Indic_is – freq_internet_access	Indic_is – internet_activities	
Ind_type – demographic_group	Ind_type – demographic_group	
Geo – country	Geo – country	Geo – country
TIME PERIOD – year	TIME PERIOD – year	TIME PERIOD – year
OBS_VALUE – internet_access_rate	OBS_VALUE – internet_activities_rate	OBS_VALUE – purchasing_power_gdp
OBS_FLAG – freq_data_flag	OBS_FLAG – activities_data_flag	OBS_FLAG – pp_gdp_data_flag

Removing Prefixes

In the **Frequency of Internet Use** dataset, the values in the freq_internet_access column contained the prefix: *'Frequency of internet access: '*. This prefix was removed, leaving only the category (e.g. daily, once a week, etc.).

In the **Internet Activites** dataset, the values in the internet_activities column contained the prefix: *'Internet use: '*. This was also removed, leaving only the activity category (e.g. online banking, streaming, etc.).

Mixed Datatypes

Mixed datatypes in the three datasets where corrected.

Missing Data

Frequency Internet Use	Internet Activities
9 missing values deleted – related to the confidentiality flag.	**49 values deleted** – related to the confidentiality flag.
3621 missing values were deleted – as they were 3.9% of the total data distributed across countries, demographic groups and years.	**7662 missing values were imputed** – with the mean grouped by country and year.

Duplicated Values

There were no duplicated values in any of the datasets.

4. **Data Profile**
   To come to the final dataset a combination of concatenation and merged was done between the different datasets. Here is the information for the last dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1006934 entries, 0 to 1006933
Data columns (total 11 columns):
 #   Column                  Non-Null Count    Dtype
---  ------                  --------------    -----
 0   freq_internet_access    1006934 non-null  category
 1   demographic_group       1006934 non-null  string
 2   country                 1006934 non-null  category
 3   year                    1006934 non-null  int16
 4   internet_access_rate    1006934 non-null  float32
 5   freq_data_flag          26460 non-null    category
 6   internet_activities     1006777 non-null  string
 7   internet_activities_rate 1006777 non-null  float32
 8   activities_data_flag    35092 non-null    category
 9   purchasing_power_gdp    1000949 non-null  float32
 10  pp_gdp_data_flag        23941 non-null    category
dtypes: category(5), float32(3), int16(1), string(2)
memory usage: 33.6 MB
```

The NaN values in **flag columns** do not represent missing values, but rather indicate rows without an assigned flag.
The NaN values in purchasing_power_gdp column correspond to Kosowo, which is not included in the dataset on **Purchasing Power Adjusted GDP Per Capita**.

Descriptive Statistics:

	year	internet_access_rate	internet_activities_rate	purchasing_power_gdp
count	1.006934e+06	1.006934e+06	1.006777e+06	1.000949e+06
mean	2.014382e+03	2.424874e+01	3.664136e+01	2.834359e+04
std	5.739227e+00	3.340169e+01	2.849574e+01	1.391105e+04
min	2.003000e+03	0.000000e+00	0.000000e+00	5.900000e+03
25%	2.010000e+03	1.660000e+00	1.210000e+01	1.930000e+04
50%	2.014000e+03	5.530000e+00	2.913000e+01	2.640000e+04
75%	2.019000e+03	4.041000e+01	5.880000e+01	3.360000e+04
max	2.024000e+03	1.000000e+02	1.000000e+02	9.620000e+04

5. Data Limitations & Ethics

- Potential bias may arise from the survey-based nature of data collection.
- Additional bias may result from differences in data reporting or collection methodologies across countries.
- Some countries and years have missing or confidential data, which may affect comparability and analysis.
- Certain values are marked with observation flags, indicating issues such as breaks in time series, low reliability or estimates.
- Ethical concerns may arise from interpreting or generalizing findings across demographic groups or countries without considering cultural, social, economic and political differences.

6. Questions to explore

Demographic Trends

- Are there differences in the frequency of internet access and the type of activities performed by different demographic groups?
- Which internet activities are the most popular among demographic groups?
- Which demographic group use the internet more for professional activities, health-related topics or social networking?
- Are there notable trends or changes over time in internet use across specific demographic groups?
- Are certain age groups more active in specific internet activities?
- How does internet use vary by education level or employment status?
- Is there a gender gap in specific digital activities?
- Are older people increasing their participation in online activities over time?
- What is the relationship between frequency of internet use and types of activities by age and gender?

Geographic Patterns

- Do individuals from the same demographic group behave differently across countries?
- Are countries with higher GDP more likely to have greater internet usage?
- Which countries have the highest/lowest rates of internet access?
- Are there notable trends or changes over time in internet use across specific countries?
- How do countries with similar GDP per capita compare in internet usage patterns?
- How do internet activities vary among countries with similar internet access rate?
- Which countries show the greatest growth in internet use over time?

Germany Behavior

- How is Germany positioned compared to other EU countries in terms of internet access and digital activities?
- Which internet activities stand out in Germany compared to other EU countries?
- Which internet activities are more or less popular in Germany compared to countries with similar GDP per capita?
- Which age group uses the internet the most in Germany? Is there a significant difference compared to other EU countries?
- How has internet usage in Germany evolved over time compared to the EU average?

7. Sources consulted

- Eurostat. (n.d.). *Internet activities by individuals and frequency of use (custom extraction)*. European Commission. Retrieved July 9, 2025, from https://ec.europa.eu/eurostat/databrowser/view/isoc_ci_ifp_fu__custom_17380066/default/table?lang=en

- Eurostat. (n.d.). *Individuals using the internet for various activities (custom extraction)*. European Commission. Retrieved July 9, 2025, from https://ec.europa.eu/eurostat/databrowser/view/isoc_ci_ac_i__custom_17403281/default/table?lang=en
- Eurostat. (n.d.). *Real GDP per capita in PPS (custom extraction)*. European Commission. Retrieved July 9, 2025, from https://ec.europa.eu/eurostat/databrowser/view/sdg_10_10__custom_17380086/default/table?lang=en

- European Commission. (n.d.). *Europe's Digital Decade: Digital targets for 2030*. Retrieved July 9, 2025, from https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age_en